

# Data Science with CODAP

A sequence about data science with CODAP by Sara Fergus, Christa VanOlst, & Jon Stapleton

## Lesson Sequence Summary

This lesson sequence offers students and teachers a way to develop data science skills using the CODAP "low-code" data analysis tool. Through a mix of "plugged" and "unplugged" activities, students engaging with the lessons in this sequence will learn about different types of data, create and evaluate data visualizations, and identify patterns in data sets using descriptive statistics, modeling, and other techniques. At the end of the sequence, students generate an original research question based on their individual goals, interests, needs, and desires and address it using the data science skills they have developed over the course of the sequence.

Throughout the sequence, students engage in many different knowledge-making activities, including small and large group discussions, guided and independent analysis activities, and journaling. The sequence also provides many opportunities for guided and independent **project-based learning**, where students either a.) engage in open-ended problem solving to address a question or problem, or b.) generate original, expressive, creative work using the skills they've developed during instruction.

## Lesson Sequence Objectives

*The students will be able to . . .*

- Use a wide range of data collection techniques (e.g., surveys, public data sets, crowdsourcing) to generate useful information to address data questions
- Interpret, create, and evaluate data visualizations using CODAP to explore patterns in data sets and to communicate trends and patterns to non-technical audiences
- Identify patterns, relationships, and trends in data using a variety of data science techniques (e.g., descriptive statistics, visualizations, modeling)
- Use predictive models (including linear regression) to make predictions given a related dataset
- Evaluate predictive models, assess how to use them appropriately to inform human decision-making, and identify problems with models which make them unreliable (e.g., bias, overfitting)
- Identify the role data science and statistics play in society at large, including the impact of misleading data visualizations, bias in data collection and analysis, and other issues at the intersection of data science and human experience.



CS Lesson Sequence

**View online!**  
[codeva.info/ds-codap](https://codeva.info/ds-codap)



# Data Science Standards Alignment

The Virginia Department of Education has provided Data Science standards to help guide teachers in facilitating high-quality data science instruction at the high school level. The chart below provides a summary of how the lessons in this sequence align to the various data science standards.

Lesson Name	Data Science Standard (see <a href="#">VDOE site</a> for full text)												
	1	2	3	4	5	6	7	8	9	10	11	12	13
01 What Is Data?*						✓				✓			
02 Types of Data	✓			✓		✓				✓			
03 Power of Visualizations						✓				✓			
04 Finding & Collecting Data		✓						✓		✓			
05 Preparing Data				✓				✓				✓	
06 Choosing Visualizations	✓					✓				✓			
07 Creating Visualizations	✓					✓				✓			
08 Descriptive Statistics	✓									✓		✓	✓
09 Creating Models									✓		✓	✓	✓
10 Making Predictions									✓		✓		
11 Overfitting & Noise							✓		✓				
12 Understanding Research*			✓		✓								
13 Developing Questions*	✓	✓											
14 Project Practice	✓	✓	✓			✓			✓				✓
15 Summative Project	✓	✓			✓			✓	✓			✓	✓

\*Denotes an “unplugged” lesson

Some of the lessons in this sequence are **“unplugged”**, meaning that they do not involve using CODAP or another computational aid to calculate values, produce visualizations, or complete other data science analytical tasks. These “unplugged” lessons help students develop **conceptual understanding** which they then apply during the “plugged” lessons.

The “unplugged” lessons in this sequence also appear in a fully “unplugged” version of this sequence, where students engage in data science learning *without* using computational tools. You can find CodeVA's unplugged data science sequence on CodeVA's GoOpenVA profile: <https://goopenva.org/profile/18746>.



## Materials

- Access to the Desmos mathematics education platform
- Access to the [Resources](#) linked in the CodeVA Curriculum Google Drive
- A printer (color is best, but black and white will work)
- Handheld whiteboards, or a similar tool for students to respond to questions from their seats
- Various craft supplies, including sticky notes & poster board; see individual lessons for itemized lists
- Access to the [CODAP](#) web-based data analysis tool
- Access to the [Kaggle data science web resource](#)

## Student Prerequisite Knowledge & Skills

This lesson sequence assumes a level of computer science and mathematics prior knowledge consistent with a 9th or 10th grade high school student who is “on grade level”. You can read about the grade-level standards on the Virginia Department of Education's [website](#).

### Mathematics Prerequisite Knowledge

We recommend Algebra I as a prerequisite course for the content in this curriculum. Specifically, students should be able demonstrate mastery of the following skills, knowledge, and competencies:

- Can perform algebraic operations on equations with multiple terms and variables, including equations with decimal values up to 8 decimal places using computational aids
- Can calculate the mean, median, and mode of a list of values
- Can identify the maximum and minimum values of a list of values and use the max and min to calculate the range.
- Can investigate and analyze linear function families algebraically, graphically & verbally, and can write the equation of a line when given the graph of the linear function

The lesson sequence is very flexible, so you should feel free to incorporate lessons and instruction to address gaps in students' prior knowledge as you go. Consider adding lessons or adjusting the pacing of the curriculum to suit your students' needs.

## Teacher Prerequisite Knowledge & Skills

This curriculum includes activities where students collaboratively and independently practice various algebraic, statistical, and programming tasks using scaffolds provided in the materials for each lesson. Educators facilitating these learning experiences will need an appropriate set of pedagogical skills and content area expertise in order to successfully facilitate the activities in this curriculum.

### Pedagogical Skills/Knowledge

- Can skillfully facilitate whole class and small group discussion around a wide range of topics, including potentially sensitive topics like racism, bias, justice, and equity
- Can perform formative, informal assessment of student skills while students work independently
- Can adjust instruction to meet student needs, and modify curricular materials to accommodate those needs
- Can guide students as they navigate open-ended, self-directed questions & learning experiences
- Can perform “lab lecture”-style instruction, where the teacher uses CODAP “live” in front of students and engages them in analysis and discussion

### Content Area Skills/Knowledge

- Can teach students mathematical concepts including: linear, quadratic, and polynomial functions; mean, median, mode, & standard deviation; data visualizations including scatter plots, box plots, bar charts, pie charts, etc.; regression modeling & related concepts
- Download, upload, and manipulate CSVs and spreadsheet files (e.g., Excel, Google Sheets)
- Can connect data science concepts & topics to their impact on social issues, including systemic bias, communication, propaganda, and injustice.

It is possible for a motivated educator to learn these prerequisite skills by studying the lesson plans and materials in this curriculum, but facilitating the activities this way adds a significant amount of preparation time to each of the lessons. The teachers who will be most successful facilitating this lesson sequence are those who have some amount of professional experience or training in high school mathematics education (Algebra I and/or statistics) *and* computer science education (Computer Science Principles and/or Programming).





## Scope & Sequence

Below, you'll find a list of the lessons in this sequence along with links to the standalone documents.

<i>Lesson Name</i>	<i>Summary</i>	<i>DS Standards</i>
<b><u>01 What Is Data?*</u></b>	This two-day lesson introduces students to different ways of expressing multivariate data, especially in non-computational formats.	DS.6, DS.10
<b><u>02 Types of Data</u></b>	In this activity, students will learn different types of data, including quantitative, categorical, ordinal, and unstructured (i.e., qualitative) data and how to store this data in CODAP.	DS.1, DS.4, DS.6, DS.10
<b><u>03 Power of Visualizations</u></b>	In this lesson, students will explore the power of visualizations in making a point or communicating information about data.	DS.6, DS.10
<b><u>04 Finding &amp; Collecting Data</u></b>	In this lesson, students are introduced to multiple methods of collecting and finding data and how to describe that data using its attributes.	DS.2, DS.8, DS.12
<b><u>05 Preparing Data</u></b>	In this lesson, students will explore data cleaning techniques. In their explorations, students consider how data cleaning can introduce bias.	DS.4, DS.8, DS.12
<b><u>06 Choosing Visualizations</u></b>	In this lesson students will explore how visualizations can serve a variety of purposes in communicating data. Throughout the lesson, students defend designs and unpack the communicative power of visualizations.	DS.1, DS.6, DS.10
<b><u>07 Creating Visualizations</u></b>	In this lesson, students will learn how to create a visualization for a given dataset and data question. They will create and modify a variety of visualizations, and practice generating visualizations.	DS.1, DS.6, DS.10
<b><u>08 Descriptive Statistics</u></b>	In this three day lesson, students learn how to analyze datasets by calculating descriptive statistics. At the end, students will complete a project where they will find data and transform it into a short news article.	DS.1, DS.10, DS.12, DS.13
<b><u>09 Creating Models</u></b>	In this lesson, students will start using a by eye technique to create models based on scatter plots. Then, students will learn one-click regression in CODAP in order to categorize patterns to create predictive regression models based on data sets.	DS.9, DS.11, DS.12, DS.13
<b><u>10 Making Predictions</u></b>	In this lesson, students explore datasets throughout the lesson by creating quick scatter plots and models to predict outcomes. Then, students collect data, analyze it for correlation (positive, negative, null).	DS.9, DS.11
<b><u>11 Overfitting &amp; Noise*</u></b>	In this lesson, students learn the concept of "noise" in data science, and how it relates to the overfitting (or underfitting) of predictive models.	DS.7, DS.9
<b><u>12 Understanding Research*</u></b>	In this lesson, students will explore the source of a data-based news report to assess the report, and then record a small "news clip" describing the results of a detailed data report in layman's terms.	DS.3, DS.5
<b>Summative Project</b>	See below	N/A

## Summative Data Science Project

This lesson sequence also includes materials for students to complete a **summative data science project**, where they complete an exploratory or research data project addressing a question of their choice. The materials for this summative project are linked in the table below:

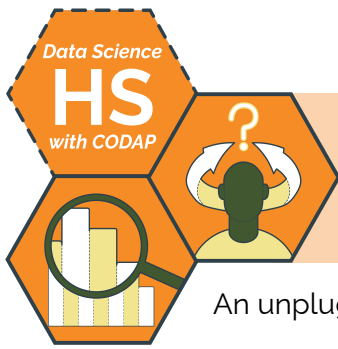
<i>Lesson Name</i>	<i>Summary</i>	<i>DS Standards</i>
<b><u>13 Developing Questions*</u></b>	In this lesson, students will develop questions to answer with data. This activity is designed to provide a foundation for student-driven project-based learning experiences.	DS.1, DS.2
<b><u>14 Project Practice</u></b>	In this lesson, students will complete a full iteration of the data cycle by modeling question formulation, data collection, analysis, visualization and the modeling processes through research & exploratory analysis projects.	DS.1, DS.2, DS.3, DS.6, DS.9, DS.13
<b><u>15 Summative Project</u></b>	In this project, you choose or collect data to engage with and explore a topic that interests you. You will run a data analysis and then present your findings in a meaningful deliverable that can inspire deep thought or action in your viewers.	DS.1, DS.2, DS.5, DS.8, DS.9, DS.12, DS.13

These activities are very open-ended, and educators have a lot of flexibility in how they facilitate them. Educators might have students complete the **13 Developing Questions** lesson, complete the **14 Project Practice** to give students a chance to work on a project with a little more structure, and then have them complete the **15 Summative Project Frame**. This entire process will likely take between 4 and 8 weeks of instruction, depending on how big students' questions end up being. Alternatively, educators might just have students complete a **14 Project Practice** assignment, omitting the open-ended, larger-scale **15 Summative Project Frame**. Likewise, they might skip **14 Project Practice** and go straight to the project frame if students are ready for a more open-ended, self-directed project. Educators could even re-sequence these three lessons, completing the project practice before developing their questions.

## Data Science with CODAP Project Examples

To help educators envision what sorts of projects students might complete, the **15 Summative Project Frame** includes example projects, each with detailed documentation of the project work. Educators implementing this curriculum can use these example projects as a way to teach themselves the details of completing the summative project, as a resource for students who might benefit from examples as they plan their individual projects, or as something to compare students' projects to during the summative assessment stage. The table below contains information about each of the examples, and links to the example data science project materials:

<i>Project Name</i>	<i>Summary</i>
<a href="#"><u>Unions in the United States</u></a>	Investigating relationships among union membership, race, & income
<a href="#"><u>Standardized Testing</u></a>	Comparing test scores by state & other measures of success.



# Unplugged: What is Data?

An unplugged introduction to data representations for high school students by Sara Fergus

## Summary

This two-day lesson introduces students to different ways of expressing multivariate data, especially in non-computational contexts (e.g., textiles, tables, collections, etc.). Students will explore “traditional” data representations (matrices and tables) as well as some “non-traditional” representations (e.g., bubble charts and heat maps, quipus and physical data, unnamed data representations). Students will discuss features of “traditional” representations, and create their own “non-traditional” representation (see *Day 2 Outline* below) to tell a story about a past experience, or another aspect of their lives.

*Note: The second day is focused on what we are calling “non-traditional” representations. However, these are only “non-traditional” in a culture which is dominated by white Americans and Europeans. After completing the quipu notice-and-wonder, consider assigning students the [Even Graphics Can Speak With a Foreign Accent](#) article, and facilitate a discussion about why we represent the data the way we do, and what may be lost when some cultures dominate over others.*

*Note: This is identical to 01 Unplugged: What Is Data? from the [Data Science Unplugged](#) & [Data Science with Python](#) sequences.*

## Objectives

The students will be able to . . .

- Create and interpret matrices and tables
- Compare and contrast matrices and tables
- Create and interpret “non-traditional” data representations


## Standards Alignment

- **DS.6:** Students will justify the design, use and effectiveness of different forms of data
- **DS.10:** The student will be able to summarize and interpret data represented in both conventional and emerging visualizations

## Materials

- White board or giant sticky, postcards, and colored writing supplies (ex. colored pencils, markers)
- [Survey](#) & [Dear Data Project](#)
- Student journals

# Vocabulary

Term	Definition																				
Tally Marks	<p>Tally marks help to keep count of data as you collect it. To use tallies, you will draw one line for each count, and every fifth will cross the previous four.</p> <table><tr><td>1</td><td>I</td><td>6</td><td>    I</td></tr><tr><td>2</td><td>II</td><td>7</td><td>    II</td></tr><tr><td>3</td><td>III</td><td>8</td><td>    III</td></tr><tr><td>4</td><td>IIII</td><td>9</td><td>    IIII</td></tr><tr><td>5</td><td>    </td><td>10</td><td>        </td></tr></table>	1	I	6	I	2	II	7	II	3	III	8	III	4	IIII	9	IIII	5		10	
1	I	6	I																		
2	II	7	II																		
3	III	8	III																		
4	IIII	9	IIII																		
5		10																			
Matrix	<p>A basic two-way matrix shows counts of intersecting attributes. Each box represents the number of data points that have the attribute of the corresponding row and column.</p>																				
Table	<p>A way to represent data points with more than two attributes. Each row of a table is a data point / element, and each column is an attribute.</p>																				
<a href="#">Quipu</a>	<p>A quipu is a recording device historically used by cultures in South America, including the Inca. The knots in the cords represented numeric values.</p> 																				
Data Representation	<p>A data representation is a way to visualize and organize collected information</p>																				

# Day 1 Outline

Formative Assessment Notes

1. As students enter the classroom, give each student a ten block or other linear object (e.s. straw, toothpick, pencil). On a table in the room, set up two columns like so:

Likes to spend time outdoors	Likes to spend time indoors

Have students place their ten blocks in the appropriate column. Once everyone has placed their blocks, have students journal about what information they could draw from this table\*. Then, discuss what students wrote.

2. On another table, set up a matrix as shown below. Have students take their block back and place it in the appropriate place on the matrix. You may want to have the rows labeled ahead of time and covered until this point.

	Favorite Season is Spring or Summer	Favorite Season is Fall or Winter
Likes being inside		
Likes being outside		

Lead a discussion about what information can be gathered from this representation. Compare and contrast this data representation with the two column table in step #1.

Monitor this short class discussion. Mention that the ten blocks could be tally marks, which would make it easier to read.

Monitor class discussion. Encourage students to compare all four boxes (which combination is most common?) as well as rows and columns (what is most common, liking to be inside or outside?)

Tell students that there are many different ways to represent data. Tallies or tables are one way, but there are lots of other ways too!

3. Draw a second table on the board, or have students fill out [this survey](#) and display the Google Sheets results. As a group:
  1. Determine what insights we might glean from this table
  2. Compare and contrast this table with the matrix in step 2.

Season	Inside / Outside	Number of Siblings	Favorite Holiday	Free time activity

This is a good time to explore what questions could be answered with the data. For example, do most people's favorite holidays fall in their favorite season?

4. **Make student-created surveys.** Have students write their own 2-4 question survey on a piece of paper. Once their surveys are written, instruct students to have 4-5 peers fill out their survey.

Once students have results, they should:

1. Draw a table or matrix (whichever is appropriate) to represent their data.
2. Write a brief summary (1-2 sentences) describing what is represented in their table/matrix and at least one interesting thing

Guide students to the conclusion that the second table allows you to keep someone's information together, ask more questions, and ask different kinds of questions. The matrix is easier to interpret and there is less room for error.

Monitor students as they create surveys and interpret results.

See [Assessment Strategies](#) below for details & rubric

## Day 2 Outline

### Formative Assessment Strategies

5. **Warm-Up:** Show students an image of a [quipu](#) (see vocabulary section), and either in pairs or on paper write what they notice about it, and what they wonder about it.

Have students share their "notice" and "wonder". Then, describe what a quipu is (see [Vocabulary](#) section for details)

6. **Practice reading non-traditional data representation:** Split students into 3 - 6 groups. Assign each group to be an "expert" on one of [these "Dear Data" representations](#). Once they have an understanding of their visualization, create groups including one expert from each of the initial groups and have them share how to interpret the visualizations with their peers. (Or, have each group present their representation to the class).

If groups present to the class, display their representation for the class to see. Otherwise, make sure each person brings their representation with them to their expert group.

7. **All together**, analyze [a week in our past](#). Then, complete the [Dear Data Assessment](#). Assessment strategies include two versions: one extended which includes data collection, one brief to be completed in class time.

**Notice & Wonder:** have students share, especially those who noticed or wondered something data-related.

Observe students while they present their interpretations to each other. Correct any misunderstandings and provide feedback on their explanations.

See [Assessment Strategies](#) below for details & rubric

See [Assessment Strategies](#) below for details & rubric

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Student-Created Surveys & Data Representations

Use this rubric to assess the sStudent-created surveys, traditional representations, and summaries (see [Day 1](#)). Students should have fill in the matrices or tables with data in appropriate locations.

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Completeness</b>	Student... 1. Created a 2-4 question survey 2. Had 4-5 peers complete the survey 3. Visually represented the data 4. Wrote a summary			
<b>Representation</b>	The representation that was selected was appropriate for the data collected.  Data was accurately placed into representation.			
<b>Summary</b>	Brief summary describes what is represented in their table/matrix			

### Non-Traditional “Presentations”

Students' interpretations should be accurate and make use of the keys on the back side of the card. All aspects of the data representation should be included (e.g. the location, color, and order in the complaint representation). An excellent presentation would also include some interpretation beyond simple data representation (“You can see that she complained to others more than she complained privately”).

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Concept</b>	The concept of the postcard is correctly interpreted and explained			
<b>Representation</b>	Students accurately explain the meanings of <i>all</i> aspects of the representation, using the representation keys as a guide.			
<b>Coherence</b>	Overall presentation is clear and coherent			



## Dear Data Assessment: Brief Version

After analyzing a week of their past (see [Day 2](#)), have students choose to either represent their past in a different way, or represent a different data set of their choice. Each student should create some way to represent their data traditionally (table, tally marks, etc.) and non-traditionally, as in the “dear data” exercise (see step #6). The same rubric as the extended option can be used.

## Dear Data Assessment: Extended Version

*You can complete this assessment during class time, or you could encourage students to collect data throughout the week and turn in a larger project (as in the Dear Data Project)*

1. Choose a topic to collect and represent data on. Use what we saw in the dear data groups as inspiration
2. Collect data throughout the week
3. Create a creative representation of your data. Your representation should show all of your data without being cluttered and hard to read.
4. Make sure that your representation has a key, as necessary
5. Create a traditional representation of some part of your data

## Dear Data Assessment: Rubric

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Data</b>	Student includes accurate data about their life			
<b>Traditional Representation</b>	Student produces a “traditional” data representation that is appropriate and accurate for their data.			
<b>Creative Representation</b>	Student creates a “non-traditional” representation to accurately represent a “data story” in their lives.			
<b>Representation Utility</b>	Student’s creative representation incorporates and effectively communicates multiple attributes of their data story.			

## Some Accommodations & Extensions

Consider breaking the lesson down into more days to adjust the pace, if needed.

Design groups intentionally to meet student needs (e.g., peer collaboration, group students with similar instructional needs together, etc.)

Encourage students to put as much information as possible into their final data representation. This will allow students who work faster to “opt-in” to a challenge, while allowing those who work slower to still meet the requirements in the time allotted.

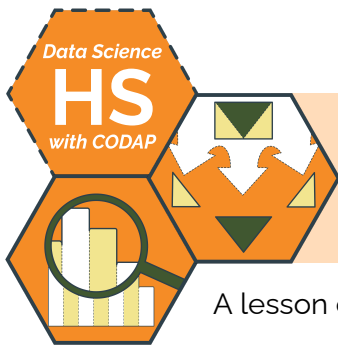
Provide a vocabulary sheet, like the one above for students who are learning English (suggested for WIDA levels 4 and below).

In the final assessment, you may choose to differentiate requirements. For example, some students may be allowed to use the “week in our past” concept (with their own data and representation), while others may be required to come up with their own topic of data collection.

## Other Resources

Here are some resources from other lessons that might be helpful here...

- [Reference Sheet](#): A resource from *Lesson 01: What is Data?* for the CODAP sequence for inputting data by hand, spreadsheets, CSVs, and json tables.



# Types of Data

A lesson on differentiating and cleaning data with CODAP by Sara Fergus

## Summary

In this activity, you will first introduce students to different ways of expressing multivariate data in computational contexts, including human-readable data formats (e.g., JSON) and spreadsheets. Students will read these different data formats, evaluate their affordances and constraints, and “teach back” procedures for uploading and viewing the data in CODAP. Then, students will learn different types of data, including quantitative, categorical, ordinal, and unstructured (i.e., qualitative) data. They will learn the kinds of questions these types of data are suited to address. Students will also consider the limitations of particular types of data, for example the restricting nature of categorical data.

*Note: This lesson is similar to the 02 Types of Data lesson plans from the CodeVA [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes several CODAP practice activities, which the others omit.*

## Objectives

*The students will be able to . . .*

- Interpret different data formats , including Excel files, CSVs, and JSON tables.
- Upload and view data in various formats within CODAP
- Identify and define parts of data tables (“case”, “attribute”, etc.)
- Classify data as quantitative, categorical, ordinal or qualitative/unstructured.
- Generate data questions that can be answered given different types of data.
- Develop guidelines for questioning data.

## Standards Alignment

- **DS.1:** The student will identify specific examples of real-world problems that can be effectively addressed using data science.
- **DS.4:** The student will be able to identify biases in the data collection process, and understand the basic ethical implications and privacy issues surrounding data collection.
- **DS.6:** Students will justify the design, use and effectiveness of different forms of data
- **DS.10:** The student will be able to summarize and interpret data represented in both conventional and emerging visualizations

## Materials

- Google Survey (What Is Data?) for Day #1 Warm-Up (view [Google Form](#) or [make a copy](#))
- Google Survey for Day #2 Warm-Up (view [Google Form](#) or [make a copy](#))
- *Data Tables in CODAP* slideshow for students to edit (view [Google Slides](#) or [make a copy](#))
- Reference Sheet: Inputting Data by Hand (see [below](#))
- Reference Sheet: Spreadsheets (see [below](#))
- Reference Sheet: CSV Files (see [below](#))
- Reference Sheet: Human-Readable Formats (see [below](#))
- Data Attribute Cards ([single-sided PDF](#), [double-sided PDF](#)): Each slide has a description and example, where students will categorize and group by "what kind of data it is".
- *Types of Data* practice quiz ([Desmos](#), [PDF](#))

## Vocabulary

Term	Definition
Spreadsheet	A spreadsheet is a table: data can be entered into rows and columns. Often, software allows you to manipulate this data within a spreadsheet. Below are two examples of spreadsheet software.
Google Sheets	Google sheets is an online spreadsheet tool where data can be stored in rows and columns, as a table. Google sheets links directly with data gathered by Google Forms.
Microsoft Excel	Microsoft Excel is a spreadsheet software that can store and manipulate data in table form.
Table	A way to represent data points with more than two attributes. Each row of a table is a data case/record, and each column is an attribute.
Case / Record	A "case" or "record" is one row of a table, which represents one entry. All of the attributes in that row belong to the same "case".
Attribute	An "attribute" is a column of a table. It is a piece of information that describes each case. Most of the time, each case will have multiple attributes.
CSV File	A CSV (comma-separated values) file is a way to store data in a computer. In a CSV, each row is a row in a table/ a case, and each attribute of that case is separated by a comma.
JSON Table	A JSON table is a specific type of human-readable table that is coded.
Data	"Data" is recorded information describing an event, person, place, or phenomenon

## Vocabulary (continued)

Term	Definition
Quantitative Data	<p>Quantitative data uses numbers to describe an amount of something. Measures like mean and median would make sense with this data. For example: age, year, number of pets, height</p> <p><i>Note: not all data using numbers is quantitative. For example "tv channel" or "ID number" would not be quantitative.</i></p>
Qualitative Data	<p>Qualitative data is typically words and descriptions. These are used for open-ended questions. For example: "what was your favorite part of this week"?</p> <p>Qualitative data can be hard to do traditional data analysis with. However, emerging visualizations like word-clouds and tools like sentiment analysis have begun to make qualitative data analysis more common.</p>
Ordinal Data	<p>Ordinal data is data that can be put in an order. Quantitative data is a type of ordinal data, but ordinal data does not need to be numeric. Ordinal data often has to do with 'rating'. For example...</p> <ul style="list-style-type: none"> <li>• Strongly disagree, disagree, agree, strongly agree</li> <li>• Poor, good, great</li> <li>• On a scale of 1 to 10, how much does the injury hurt?</li> </ul>
Categorical Data	<p>Categorical data puts respondents into groups. Categorical data is often collected using a multiple choice question. For example, favorite season breaks respondents into "spring", "summer", "fall", and "winter".</p> <p><i>Note: some 'categories' would require an 'other' in order to categorize. This is particularly true of categories like 'race', where people differ a lot. It is important to consider how many people would fall into the 'other' category. If it would be a large number of respondents, consider collecting qualitative data instead.</i></p>
string	A string is a piece of data that is a word
int	An "int" is a piece of data that is an integer (number without a decimal point)
float	A float is a piece of data that is numeric and has a decimal point
Data Question	<p>A Data Question is a question that can be answered with data and facilitate a quality data analysis. A data question might arise from a <i>broad question</i> or a <i>subjective question</i>. Answering the question allows further questions to arise. Answering the question should contribute to a larger understanding of the world or an overarching question.</p>

## Day 1 Outline

### Formative Assessment Notes

*Note: To save time, day 1 can be condensed into brief direct instruction about the different ways to upload data into CODAP.*

1. **Warm-Up:** Have students fill out [this survey](#) (be sure to have them fill out your copy so you can see the responses), which asks basic questions about the students (favorite season, what they like to do in their free time, etc.). It includes a variety of types of questions (multiple choice, numeric, open-ended, etc.). This is a brief activity to create data for later in the lesson
2. **Uploading Data into CODAP:** In this activity, students will be exploring different ways to upload data into CODAP. During this activity, students will do the following:
  - “By Hand” Group: Read 1 article, upload data to CODAP
  - “Spreadsheets” Group: Read 2 articles, upload data to CODAP
  - “CSV” Group: Download and open as CSV, watch video, read bulleted list
  - “Human-Readable” Group: View websites, copy & paste data into CODAP

Break students into four groups (one for each data format).

- Give students access to [this slideshow](#). Each group has a few dedicated slides where they will input vocabulary and other notes about their method. They will use their slides at the end of the class to present.
- Students will analyze their data type and upload their data to CODAP. Give each group their respective [reference sheet](#), which will guide them through their data format upload method and vocabulary related to their data format.

Encourage students who need an extension to add more detail about their data format to their presentations—what industries use their assigned format? Can they think of reasons different industries use different data formats?

3. **Exit Ticket:** Have students complete an *Exit Ticket* (see [Assessment Strategies](#) below) on their way out.

**Monitor as students work—either by walking around the room or by monitoring what students add to the group slideshow**

**You may choose to have students write what they learn on their reference sheet as they go**

## Day 2 Outline

### Formative Assessment Notes

1. **Warm-Up:** have students fill out [this Google survey](#), which asks a variety of questions to collect data for the students to consider. Once all students have filled out the survey, show the results to the whole class. Point out that Google used different methods to show the results for different questions.

Have students do a think-pair-share: What patterns do you notice in how Google shows results? What kind of data did Google represent in each way?

2. Hand out the [Data Attribute Cards](#) to students in small groups. Instruct them to sort data into 3-4 categories based on "what kind of data it is". They can choose what those categories should be.

Once they have sorted, have each group share their categories.

As they share, write their categories on the board. Have students group related responses together. Then have students give headers to each column. You may want to include a fifth column to hold words that don't fit well anywhere.

??	??	??	??
Better or worse	Numbers	groups	descriptions
In order	Greater or less than	types	words
More or less		categories	longer

3. Once all groups have shared, write the vocabulary words into the table. In the example above, you would replace the "??"s with ordinal, quantitative, categorical, and qualitative in order. Have students re-categorize as needed to fit those titles.

**Discussion.** Then, ask where students put "race". Many will have placed it in categorical. Discuss the advantages (identifying discrimination by answering questions like "is race related to GPA") and limitations (leaving out people of mixed race, not being able to include all races) of this. Be sure to discuss this topic sensitively, paying special attention to questioning stereotypes and avoiding microaggressions toward marginalized students.

Students should notice, without vocabulary, that:

- Qualitative data is displayed as text
- Categorical data is displayed as a pie chart
- Quantitative data is displayed as a histogram.

While students sort, make sure that they are sorting by type of data, not topic of data.

For example, you don't want them to put together "number of pets" and "favorite animal"; these are different data types, despite the fact they both have to do with animals.

Check students' recategorizations. You may ask students to defend their choice; then, you can either question to point them in the right direction, or point out that some data makes sense in multiple types.

4. **Generating Data Questions:** Using the same attribute cards from step 2, come up with questions for a few examples together. Then, have students come up with their own questions for the remaining cards with their group, recording their questions in their journals. You may choose to have students write their questions on the board to get them moving.
5. **Desmos Activity:** Assess student understanding of data types using this [Desmos activity](#) ([PDF Version](#))
6. **Exit Ticket:** Have students find patterns: what kinds of questions can be asked about categorical / ordinal / quantitative / qualitative data?

Provide students with this [Types of Data Cheat Sheet](#).

Stress that they are looking for patterns in questions for the exit ticket, not using specific vocabulary words for “types of questions”. Consider having students work in groups.

See [Assessment Strategies](#) below for questions and sample responses.



## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Day 1 Exit Ticket (see [below](#) for printable copies)

Name: \_\_\_\_\_ Date: \_\_\_\_\_

*What are the ways to store data that we talked about today?*

*Which do you think is the most useful and why?*

*Which do you feel the most comfortable teaching to someone else?*

### Slideshow & Presentations Rubric

	<i>Proficiency</i>	<i>Yes</i>	<i>No</i>	<i>Notes</i>
<b>Vocabulary</b>	Student(s) use relevant terms correctly and provide accurate definitions/explanations for terms when appropriate.			
<b>Pros and Cons (for groups 1-3)</b>	Student(s) provide reasoned, accurate explanations of the pros and cons of their method			
<b>Using the Method</b>	Students accurately demonstrate how to upload data to CODAP via their method.			

**Day 2 Exit Ticket** (See [here](#) for printable copies)

Name: \_\_\_\_\_

Date: \_\_\_\_\_

*What kinds of questions can be asked about qualitative data?***Possible Answer:** Qualitative data can help you find patterns and gain understanding.*What kinds of questions can be asked about quantitative data?***Possible Answer:** Qualitative data can help you find patterns and gain understanding.*What kinds of questions can be asked about ordinal data?***Possible Answer:** Ordinal data can answer questions about the scale of the data.*What kinds of questions can be asked about categorical data?***Possible Answer:** Categorical data can answer questions about the makeup of the data.

## Some Accommodations & Extensions

You may want to place English Language Learners in the Human-Readable or CSV group, since it does not involve reading an article, or distribute materials in advance (a couple days ahead of time) for students who will need extra preparation to contribute to the group work.

When groups present their findings, some students may benefit from a copy of all four reference sheets to follow along with during presentations. You can have students record notes on a collaborative copy (e.g., Google Doc) of the reference sheet so everyone can follow the other groups' work.

For students with impaired vision, consider encouraging students to view the slideshow on their personal device while their peers are presenting

Some students may be given the data examples ahead of time so that they can better participate in group work.

You may choose to have students write categorizations or questions on the board to get them moving, however for students with mobility challenges you may choose to have them simply share out or write on paper as a group. You could also use an online platform like jamboard.

Students who may need work in smaller chunks may be given only a subset of the attribute cards.

Extension attribute cards facilitate the beginning of bivariate thinking. You may choose to not use these at all, use them with the whole class, or use them only with the students who work more quickly.

At the end of the class, you may provide some/all students with [this cheat sheet](#).

## Other Resources

These resources accompany other lessons, but could be helpful here...

- Warm-Up comparing tables, CSVs, aggregate tables, and aggregate CSVs ([slides version](#), [Desmos version](#))
- [Unplugged Dear Data project](#) to practice representing data
- [Kaggle Datasets](#): a resource for students to find data, usually as a CSV file

# Inputting Data by Hand

A guide to data in CODAP by Sara Fergus

## Vocabulary

**Table:** A *table* is a way to organize data. It is made up of rows and columns. Putting data into a particular cell of the table tells us that the piece of data comes from the row that it is in and represents the column it is in. See the image (right) of rows and columns.

	Column 1	Column 2	Column 3
Row 1		X	
Row 2	X		
Row 3			
Row 4	X		
Row 5		X	X
Row 6			X

**Case / Record:** A *case* or *record* is one row of the data. You can think of this as meaning where the data in that row comes from. For example, one case might represent a single person. Everything in that row then is information about that one person. The next row would be a different person.

**Attribute:** An *attribute* is one column of the data. This is one measurement that belongs to the row. For example, in a table whose rows are people, there may be attributes like height, weight, age, and eye color. These are all situated on the same row to represent that they are attributes of that specific case.

Assignment: Put these definitions in your own words into slide 3

## Pros and Cons

Read [this article](https://bizfluent.com/info-8585434-advantages-disadvantages-manual-data-entry.html) (https://bizfluent.com/info-8585434-advantages-disadvantages-manual-data-entry.html) about manual data entry. Assignment: Describe in your own words the biggest advantages and disadvantages of this form of data entry in slide 4.

## CODAP

The most straightforward way to input data into CODAP is by hand. [There are instructions on how to do this in CODAP](https://codap.concord.org/help/work-data/how-to-input-data-hand) (https://codap.concord.org/help/work-data/how-to-input-data-hand). First, try inputting the data that you collected as a class.

## Univariate Data

On CODAP, create a visual representation of each column of the table by pressing "graph" and then dragging the column name to the x-axis. The table allows us to easily make sense of one attribute at a time. What conclusions about the data can you draw from the visualization that CODAP creates?

Assignment: When you share with the class, you will show them:

1. how to input the data by hand.
2. what each column looks like when graphed and
3. what conclusions you can draw about the data.

# Spreadsheets

A guide to data in CODAP by Sara Fergus

## Overview

You will be looking at spreadsheets. Before we get started, read [this article about spreadsheets](https://www.techtarget.com/whatis/definition/spreadsheet) (<https://www.techtarget.com/whatis/definition/spreadsheet>). As a class, you generated a spreadsheet with Google Sheets. You will be working with this now.

## Vocabulary

**Spreadsheet:** A *spreadsheet* is a way that computers store tables. It is made up of rows and columns, representing different parts of the data. Spreadsheets on the computer allow you to *manipulate data* to better understand or to make new calculations. For example, a column of a spreadsheet may represent the measure of a circle's diameter. Using the spreadsheet software, you could make a new column that represents the circle's area, by multiplying the diameter by pi.

**Google Sheets:** Google sheets is a spreadsheet software. An advantage of Google Sheets is that it can get information from Google Forms, as you saw as a class. You can download Google Sheet spreadsheets in many different forms.

**Microsoft Excel:** Microsoft Excel is a downloadable software that allows you to manipulate spreadsheets. It is almost the same as Google Sheets, except that it has to be downloaded and that it has additional features. When you save a spreadsheet in excel, it will usually have the extension .xlsx to represent that that data is stored in a spreadsheet.

Assignment: Put these definitions in your own words into slide 7.

## Pros and Cons

Read [this article](https://www.theguardian.com/politics/2020/oct/05/how-excel-may-have-caused-loss-of-16000-covid-tests-in-england) (<https://www.theguardian.com/politics/2020/oct/05/how-excel-may-have-caused-loss-of-16000-covid-tests-in-england>). Using what you have learned from this article and the previous one, compile a list of pros and cons for spreadsheets.

Assignment: Put this list into slide 8

## CODAP

CODAP allows you to easily upload information that is in spreadsheet form. First, download the Google Sheets data that you collected as a class as an excel spreadsheet by clicking *file > download > microsoft excel (.xlsx)*. Now, upload the data to CODAP using [these instructions](https://codap.concord.org/help/work-data/how-to-import-data-csv-json-codap-or-txt-file) (<https://codap.concord.org/help/work-data/how-to-import-data-csv-json-codap-or-txt-file>). What happened when you did that?

Assignment: You will be sharing with the class how to download the Google Sheet and upload it into CODAP.

# CSV Files

A guide to data in CODAP by Sara Fergus

## Overview

You will be manipulating data using a CSV file. Assignment: First, download the data you collected as a class from Google Sheets as a csv by clicking [file > download > Comma Separated Values \(.csv\)](#). Now, find the file on your computer and right-click on it. Choose [open with > TextEdit](#) or [Notebook](#). TextEdit or Notebook may not be on your machine. If it is not, open with any text editor.

How is a row of a table represented in a CSV file?

How is a column of a table represented in a CSV file?

## Vocabulary

CSV stands for Comma Separated values. Watch [this video](https://youtu.be/OGOD2Fqy5k8) (<https://youtu.be/OGOD2Fqy5k8>) to get an idea of how a CSV file works.

Assignment: Using the questions from the overview section (above) and what you learned in the video, create a definition of a CSV file and put it into slide 11

## Pros and Cons

Read [this article](https://www.shopping-cart-migration.com/must-know-tips/5985-csv-what-why-and-how) (<https://www.shopping-cart-migration.com/must-know-tips/5985-csv-what-why-and-how>). Choose the three most important pros and the three most important cons of a CSV file. Assignment: put them in your own words on slide 12.

## CODAP

CODAP allows you to easily upload information that is in a csv file. You should already have the data that you class collected as a csv file. If you do not, download the data as a csv file now. Then, upload the data to CODAP using [these instructions](https://codap.concord.org/help/work-data/how-to-import-data-csv-json-codap-or-txt-file) (<https://codap.concord.org/help/work-data/how-to-import-data-csv-json-codap-or-txt-file>). What happened when you did that?

Assignment: You will be sharing with the class what a csv file looks like in a text editor, how to read it, and how to upload it to CODAP.

Once you are prepared to present, spend some time exploring the data on CODAP.

# Human-Readable Formats

A guide to data in CODAP by Sara Fergus

## Overview

Many times, something that makes sense to a computer does not make sense to a human. It is your job as a data scientist to translate between the computer and the human. When you do that, you are putting your data into *human readable formats*.

Take a look at the tables on [this website \(https://educationdata.org/student-loan-debt-statistics\)](https://educationdata.org/student-loan-debt-statistics). You will notice that the tables look different than the Google Sheets file your class created. That is because they were translated into a human-readable format. The programmers of this website translated using a language called CSS.

There are lots of other ways to translate between what the computer understands and a human-readable format. Another common method is called JSON, which looks like this (right).

```
SELECT *
FROM
  JSON_TABLE('{ "id": 1, "first_name": "Rob", "last_name":
    "Hedgpeth", "email": "robh@mariadb.com" }',
    '$' COLUMNS (
      id INT PATH '$.id',
      first_name VARCHAR(25) PATH '$.first_name',
      last_name VARCHAR(25) PATH '$.last_name',
      email VARCHAR(50) PATH '$.email')
    ) AS person;
```

You can see that JSON helped convert information from code (top of image) to human-readable table (bottom of image).

id	first_name	last_name	email
1	Rob	Hedgpeth	robh@mariadb.com

a

## Vocabulary

Assignment: On slide 15, define *human readable format* and *JSON* in your own words.

## CODAP

CODAP does a great job of using human-readable formats. Go back to [the website we were looking at before \(https://educationdata.org/student-loan-debt-statistics\)](https://educationdata.org/student-loan-debt-statistics). Choose one of the tables and do the following:

1. Highlight the whole table, *including the headers!*
2. Copy the table by pressing ctrl+C or right clicking and choosing "copy"
3. Go to CODAP
4. Press *tables > new from clipboard*

What happened?

Assignment: You will be showing your peers how to input data into CODAP this way.

## Types of Data Cheat Sheet

	<b>Definition</b>	<b>Example</b>	<b>Questions to Ask</b>	<b>Notes</b>
<b>Quantitative / Numeric</b>	Quantitative data uses numbers to describe an amount of something. Measures like mean and median would make sense with this data. Basic arithmetic would also make sense with this type of data	age, year, number of pets, height	What is "normal"? What is the range of the data? How "different" is one data point?  These questions ask about the <i>spread of the data</i>	Not all data using numbers is quantitative. For example "tv channel" or "ID number" would not be quantitative.
<b>Ordinal</b>	Ordinal data is data that can be put in an order. Quantitative data is a type of ordinal data, but ordinal data does not need to be numeric.	Ordinal data often has to do with 'rating'. For example... <ul style="list-style-type: none"> <li>Strongly disagree, disagree, agree, strongly agree</li> <li>Poor, good, great</li> <li>On a scale of 1 to 10, how much does the injury hurt?</li> </ul> Dates may also be considered ordinal	What is "normal"? What is the range of the data? How "different" is one data point?  These questions ask about the <i>spread of the data</i>	
<b>Categorical</b>	Categorical data puts respondents into groups.	Categorical data is often collected using a multiple choice or multiple answer question. It cannot be ordered. For example, favorite season breaks respondents into "spring", "summer", "fall", and "winter".	What is most common? What is the makeup of the data?  These questions ask about the <i>composition of the data</i>	Some 'categories' would require an 'other' in order to categorize. This is particularly true of categories like 'race', where people differ a lot. It is important to consider how many people would fall into the 'other' category. If it would be a large number of respondents, consider collecting qualitative data instead.
<b>Qualitative</b>	Qualitative data is typically words and descriptions. These types of questions are useful when you can't clearly categorize questions.	These are used for open-ended questions. For example: "what was your favorite part of this week"? Or "If you could have any superpower, what would it be?"	How do people feel about this? Are there patterns in the data?  These questions ask about <i>patterns and descriptions in the data</i>	



## Day 1 Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

*What are the ways to store data that we talked about today?*

*Which do you think is the most useful and why?*

*Which do you feel the most comfortable teaching to someone else?*

Name: \_\_\_\_\_

Date: \_\_\_\_\_

*What are the ways to store data that we talked about today?*

*Which do you think is the most useful and why?*

*Which do you feel the most comfortable teaching to someone else?*

## Day 2 Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

*What kinds of questions can be asked about qualitative data?*

*What kinds of questions can be asked about quantitative data?*

*What kinds of questions can be asked about ordinal data?*

*What kinds of questions can be asked about categorical data?*

Name: \_\_\_\_\_

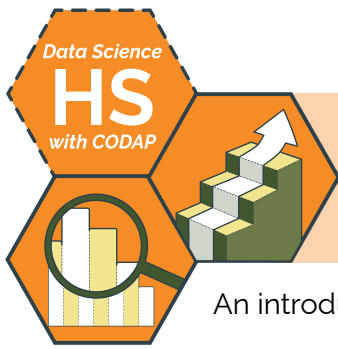
Date: \_\_\_\_\_

*What kinds of questions can be asked about qualitative data?*

*What kinds of questions can be asked about quantitative data?*

*What kinds of questions can be asked about ordinal data?*

*What kinds of questions can be asked about categorical data?*



# The Power of Visualizations

An introduction to interpreting visualizations with CODAP by Christa VanOlst

## Summary

In this lesson, students will explore the power of visualizations in making a point, supporting an argument, or to communicate information about data. Students will interpret visualizations, justify the use of visualizations to tell a story about data, and create visual narratives using speculative data. At the end, students will connect and justify the use of visualizations in their local news.

*Note: This lesson is similar to the The Power of Visualizations lesson plans from the CodeVA [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes a CODAP activity, which the others omit.*

## Objectives

*The students will be able to . . .*

- Identify the importance of effective visualizations types based on given data.
- Interpret the meaning of visualizations in published works.
- Analyze key characteristics of visualizations that can lead to curiosity.
- Create rough sketches of data visualizations based on research.

## Standards Alignment

- **DS.6:** Students will justify the design, use and effectiveness of different forms of data.
- **DS.10:** The student will be able to summarize and interpret data represented in both conventional and emerging visualizations.

## Materials

- Craft supplies, including colored markers/pencils, rulers, string or yarn, white boards and markers
- CODAP File: Which Visualization is the best? ([codap](#))
- Teacher Directions: [Affinity Mapping](#)
- Slides: The Power of Visualizations Class Deck (view [Google Slides](#) or [make a copy](#))
- E-Cigarettes Line Graphs Data Talk ([Desmos](#))
- Website: [Google Trends](#)

## Vocabulary

Term	Definition
Visualization	The art of representing information in the form of a chart, diagram, picture, infographic, etc. for an audience.
Effectiveness	Used to describe a visualization that includes all of the information that is needed to produce the intended level of understanding at a glance.
Design Elements	The formatting, headings, colors and style of a data visualization.

## Day 1 Outline

### Formative Assessment Notes

- Warm Up - Which Visualization is the best?** Have each student vote for which visualization is most effective at presenting the information from the dataset.

While students walk into the room, have the [CODAP \(Which Visualization is Best?\)](#) file projected on the screen. When students arrive, have them quickly vote by entering their choice as a case in the table named "Votes" on your computer. Observe the graph named "Results" to watch the votes create a visualization

Then, in their journals, have students write about their choice.

Consider discussing the efficiency of the "Results" visualization for interpreting our class's choice versus having to count and tally up each student's choice one by one.

- Affinity Mapping:** Use the [Teacher Directions](#) to facilitate an affinity mapping activity to answer the following:

- 1. What does a visualization accomplish better than the table?**
- 2. What does the table accomplish better than the visualization?**
- 3. What do both representations have in common?**

During the activity, students work in groups to sort & analyze data. Have students justify why their ideas fit within the categories, and how the categories relate to or differ from one another.

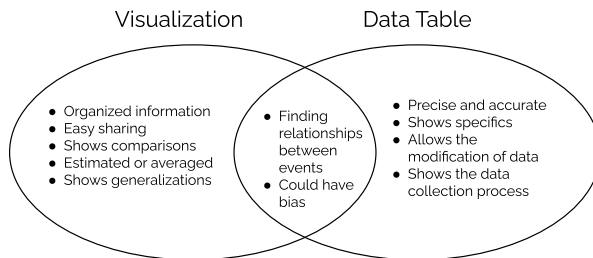
**Assess students' ability to choose the most relevant visualization.**

**Assess students' rationales on their white boards, or provide formative feedback using a think-pair-share strategy to engage in conversation with students.**

**Guide students to the conclusion that the data itself is powerful and visualizations are an effective aid in communicating findings, identifying patterns or trends, and interpreting data at glance.**

3. **Reflection:** In their journals, have students draw a Venn diagram to compare & contrast visualizations vs data tables. Have them share their ideas with the class.

### Example Result



4. **Communicating Interpretation:** Break students into groups of 2–3 (enough to evenly split the 11 examples in the activity).

Assign each group a slide number and use the directions in the following resource: [The Power of Visualizations Class Deck](#) to complete the activity.

Give students time to explore groups' visualizations and findings on the other student produced slides.

5. **Revisit and Interpret a Dear Data Representation:** Have students choose one day (data point) and identify all the attributes represented in the visualization. Have students write one sentence to “decode” a single data point, describing the phenomenon expressed by the visualization of that data point.

For example, for this data point I might write:



She must have purposely (the long pink symbol indicated this) smelled a beauty product (the color purple indicated this) that was mildly intense (medium sized symbol indicated this) and lasted only a second (the gray duration symbol indicated this). Since this was her first smell of the week, maybe this was a perfume or

cosmetic product in her morning routine, especially since this symbol occurs periodically in the week of smells.

- [A Week of Smells](#)
- [How to Read](#)

Have students add one smell from their day and then have a classmate interpret each other's additions.

**Guide students toward the conclusion that there will always be trade offs when using visualizations to portray information.**

**Ensure that students identify at least one smell with the following attributes: What is the smell? How long did the smell last? Did the smell give her deja vu? Did she enjoy the smell? Did the smell require proximity?**

## Day 2 Outline

### Formative Assessment Notes

6. **Warm Up Data Talk:** Use the following data talk to introduce interpreting line graphs - [NYT E-cigarettes \(Line Graph\)](#)
- Be sure to engage students in using the appropriate vocabulary (x-axis, y-axis, title, slope, maximum, minimum, line style, etc.).

7. **Interpreting Google Trends:** Show the students the following visualization: [Google Search for Data Science](#) (or choose your own from [Google Trends](#)), & explain that it is a visualization of Google search terms.

Have students discuss the following prompt in pairs:

*Why do you think the graph looks this way? Do you have any theories about the “spikes” in Google searches?*

Have students share their theories with another group. Then, repeat the activity with the larger student groups and a new Google Trends graph.

8. **Predicting Google Trends:** Model creating a chart where you predict the shape of the [Search for Fortnite](#) trend:
- Draw the x-axis & label starting & ending dates relevant to the search. For this example, the x-axis should start at 2017 (when the game was created) and end in the present day.
  - Demonstrate using your string to display a line graph showing a spike every fall/early winter (when they release new seasons). Over the years, the peaks of each spike decrease (due to waning popularity).
  - Show the actual visualization, & compare it to your prediction

Have students use an 18-inch string (or a hand-drawn line) & a whiteboard to create graphs that visualize what they *think* some of the following Google Trends will show:

**Google Trends for Analysis:** [Search for Data Science](#), [Search for Motivational Quotes](#), [Search for Funny Vines](#), [Search for Blue Light Glasses](#), [Search for Jobs Near Me](#), [Search for Super Bowl](#)

Take time to teach back relevant terms if students have trouble using them in context.

Throughout the activity monitor the students ability to visualize and justify their prediction of the “spike”. The main focus is for students to defend their reasoning.

Check in as needed during group discussions, and repeat the activity as needed to make sure everyone is on track.

You may find that students need some additional context to successfully reason about trends. Feel free to provide a narrative for them to relate to their data. For example, ask:

- Where do you see a “spike”?
- Do you know any big events that happened then?

If students have trouble with these questions, provide additional context

9. **"Telling a Data Story" Journal Entry:** Display all three (or one) following visualization: *The Fried Ratio* ([web/Drive](#)), *Is there Life on Mars* ([web/Drive](#)), *Electricity Prices* ([web/Drive](#))

In their journals, have students pick one visualization and write a fictional short story (8-10 sentences) identifying the information:

- Which visualization did you choose?
- Who would have collected and visualized this data?
- Why did they collect this data in the first place?
- How did they collect the data?
- Why did they visualize the data?
- Who is the audience of the data visualization?

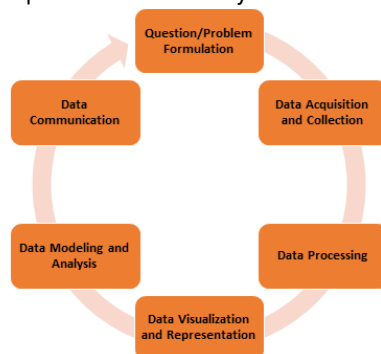
Have each group share their short story with the adjacent group.

10. Have students use a chosen data set from the previous lessons (if desired, they may do a google search to find a table or look at some websites that have a lot of data, like [kaggle](#) - students will explore these more in future lessons).

Have students open their CODAP from Lesson 1 where they chose their own data set. Have students add a text element to their file labeled Data Journal - Opportunities to Explore.

Have the students answer the following questions in the text box.

1. What do you wonder about this data set given the attributes identified?
2. What relationships could you explore based on certain attributes or your entire data set?
3. How do you predict that using a visualization will aid in making your point based on your answer to number 2?



11. Have students complete the [Exit Ticket](#) where they can practice interpreting visualizations.

**Observe students while they share their interpretations with each other. Correct any misunderstandings and provide feedback through discussions on their explanations.**

**Students can turn in their CODAP so that the teacher can assess for an opportunity to progress with this data set and an opportunity for feedback.**

**Some students may have chosen data sets from last class that could be restricting exploration, this is an opportunity for students to begin self assessing using the data cycle.**

**See [Assessment Strategies](#) below for details & rubric**



## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few opportunities for students to show their learning by creating artifacts:

**Exit Ticket** (Google Form [Exit Ticket: Is this a data visualization?](#) or see [here](#) for printable copies)

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Consider the image by Sara Weber where the scarf represents the length of daily delays on one woman's 40-minute commute to work. The commuter knitted two rows each day.

- Gray for delays under five minutes
- Pink for up to 30 minutes
- Red for a delay of more than a half-hour or delays in both directions.



Do you consider the following image above a data visualization? What would be considered the data? What other information could be helpful when interpreting this visualization? What story does this tell?

**Possible Answer:** Yes this is considered a data visualization because it is a graphical representation of information. The data would be vertically sewn lines and the color of each line. If we knew the time of year this was sewn then maybe we could pinpoint why there was consecutive red towards the right end, was it holiday traffic or construction maybe?



## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

For students with vision impairment, consider encouraging students to view all external resources on their personal device while displaying or interacting as a class.

The teacher can intentionally assign groups based on student levels including but not limited to performance or age.

The teacher could provide a vocabulary sheet with correlating images of each word listed above for ELL students to annotate/revisit throughout the duration of the lesson.

Bullet points could be provided for the reflection of Activity 1, then students could be tasked with sorting them correctly in the venn diagram.

Paragraphs could be already written in the slides for Activity 2 to focus mainly on interpreting visualizations, not communicating and interpreting visualizations. Students could match the correct visualization to each paragraph.

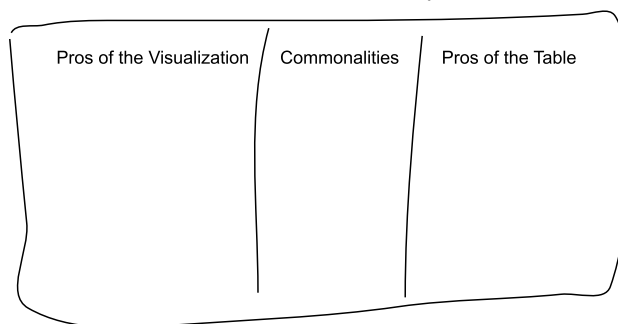
### Extensions

Have students explore the following site: <https://pudding.cool/2017/03/film-dialogue/> and reflect on the impact of having multiple visualizations to aid in making a point.

## Affinity Mapping - Teacher Directions

**Optional:** Before administering this activity watch this short video: [What is Affinity Mapping?](#)

1. Distribute or display the following resources: [Visualization](#) vs. [Data Table](#).
2. Split students into groups of about 4 (by teacher discretion).
3. Each group should have a marker, one large sticky poster, stickers (optional) and a pile of sticky notes.
  - o Have students use the marker to draw a map of the following categories on the large sticky.



4. Display the question "What does the visualization accomplish better than the table?" on the board.
5. Have students write their ideas on sticky notes (one idea per note).
6. Students should place these stickies in no particular order under the Pros of Visualization column.
7. Repeat steps 4-6 with the following questions, students will place their stickies on the corresponding columns.
  - o What does the table accomplish better than the visualization?
  - o What do both representations have in common?
8. Once all of the ideas have been generated, in their groups starting with the Pros of Visualizations column, have students begin grouping their ideas into similar categories.
  - o *Assessment Strategy:* Have students justify why these ideas fit within the categories and how the categories relate to or differ from one another.
9. Once distinct categories have formed, have students give each category a label or title.
  - o Some ideas may result in their own category.
10. Repeat steps 8 and 9 with the Commonalities and Pros of Table columns.
11. Display the posters around the room and give each student a set of stickers. Have the students gallery walk to read each group's ideas.
12. Students should place stickers next to ideas that matched their groups.
  - o This can also be done by placing check marks or stars next to ideas with a writing utensil.

## Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Consider the image by Sara Weber where the scarf represents the length of daily delays on one woman's 40-minute commute to work. The commuter knitted two rows each day.

- Gray for delays under five minutes
- Pink for up to 30 minutes
- Red for a delay of more than a half-hour or delays in both directions.



Do you consider the following image above a data visualization? What would be considered the data? What other information could be helpful when interpreting this visualization? What story does this tell?

Name: \_\_\_\_\_

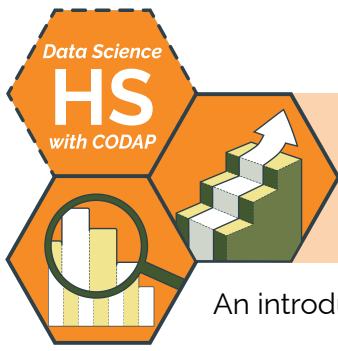
Date: \_\_\_\_\_

Consider the image by Sara Weber where the scarf represents the length of daily delays on one woman's 40-minute commute to work. The commuter knitted two rows each day.

- Gray for delays under five minutes
- Pink for up to 30 minutes
- Red for a delay of more than a half-hour or delays in both directions.



Do you consider the following image above a data visualization? What would be considered the data? What other information could be helpful when interpreting this visualization? What story does this tell?



# Finding & Collecting Data

An introduction to acquiring, collecting, and preparing data for future use by Christa VanOlst

## Summary

This lesson introduces students to several methods of collecting and finding data. Students focus on how to source data and how to describe data using attributes. Throughout the lesson, students provide information to create a crowdsourced data set in CODAP, explore existing data sets on the internet, and then individually create their own data set to describe collected artifacts from nature.

*Note: This lesson is similar to the [03 Finding & Collecting Data](#) lesson plan from the CodeVA [Unplugged Data Science](#) sequence. This lesson includes CODAP activities, while the unplugged one focuses on "by hand" activities.*

## Objectives

*The students will be able to . . .*

- Explore methods of finding data including research and collection.
- Organize data points/elements in a structured data table.
- Define attributes that will describe the collected data points/elements.

## Standards Alignment

- **DS.2:** The student will be able to formulate a top-down plan for data collection and analysis, with quantifiable results, based on the context of a problem.
- **DS.8:** The student will be able to acquire and prepare big data sets for modeling and analysis
- **DS.10:** The student will be able to summarize and interpret data represented in both conventional and emerging visualizations.

## Materials

- Craft, supplies, including large sticky notes, poster paper, construction paper, tape/glue, stickers
- *Punching into a Time Clock* ([reference image](#))
- *Loyalty Card* ([reference image](#))
- Finding Data in Nature Activity Guide ([see below](#); print one per student)
- CODAP Warm Up ([codap](#))
- Finding Data in Nature ([codap](#))
- *Measure of America* form (view [Google Form](#) or [make a copy](#))

## Online Resources

Be sure that these websites & resources are not blocked by your school's filter

- Video: [What is a Punch Clock?](#)
- [Youth Disconnection Site](#)
- [Exploring Census Data](#)
- [Measure of America Map](#)
- [Kaggle All Data Sets](#) & [Kaggle Small Data Sets](#)
- [Datasets by Year](#)
- [Google Dataset Search](#)
- [Word Cloud Generator](#)

## Vocabulary

Term	Definition
Raw Data	Data that has been collected but has not yet been organized
Data Source	The location where the data points/elements were originated from or collected
Crowdsourced Data	Data that is provided from a crowd, usually collected in real-time and obtained through surveys
Table	A way to represent data points with more than two attributes. Each row of a table is a data point / element.
Attributes	A quality or characteristic given to a data point/element, usually represented as a column in a data table.
Records/Cases	Records/cases in a database or spreadsheet are usually represented by "rows"
Data Representation	A data representation is a way to visualize and organize collected information

## Day 1 Outline

### Formative Assessment Notes

1. **Warm Up:** Before students arrive, have the following [CODAP \(Warm Up\)](#) loaded onto your computer. Consider projecting this page, or have a computer open on the front desk. Have each student add a row and type their responses to each question into the table from left to right as they arrive. *You'll revisit this on Day 2.*
2. **Discussion:** Display the [Punching into a Time Clock](#) and [Loyalty Card](#) images as students arrive. Have students answer the following question in their journals:

***"How are the devices in the images collecting data? What data are they collecting? Why are they collecting it?"***

If students are unfamiliar with what a time clock is and how it functions, consider showing this [short video](#) and then conducting a short discussion about how employers used to keep time sheets.

3. **Data Snowball Discussion and Research:** Have students explore the following site: [Youth Disconnection Site](#) - A tool to understand the trends in work life of 16-24 year olds in America. Ask:

***"Where do you think this data came from? How did they collect it? Who collected it?"***

4. **Optional Extension:** Show the students the [Exploring Census Data](#) for future reference in the course, consider having the students search *Virginia* to quickly explore the data sets that are available. Have students categorize the following as a table or visualization strength:

- Identifying a Case/Record
- Identifying Attributes of data
- Identifying Precise Calculations
- Identifying Generalizations of attributes
- Comparing/Contrasting data attributes
- Finding a trend or pattern in data

**Collecting the data at the beginning of class while students are exploring in steps 2 & 3**

**Skim answers for interpretations of the data being stored, and provide feedback & as necessary.**

**Consider suggesting to students they could bookmark the Exploring Census Data site. (See [Extensions](#) below for details).**

5. **Discussion:** Have students explore the [Measure of America](#) site (data about well-being in America). Pose the following question:

***“Which state’s Human Development Index (a numeric summary of each state’s average life expectancy, education, & Income per Capita) surprises you the most?”***

In pairs, have students turn to one another and share one piece of information they found surprising about the well-being of America by completing the sentence: “I wonder why *[insert state]* has such a high or such a low *[insert attribute]* rate?”

- **Example:** “I wonder why Alaska has one of the highest income rates given the small population in the state?”
- **Optional:** You can allow time here for students to research their questions with their partner. Some questions may be answerable through articles and research, others may require deeper research and data science!

6. **Digging Deeper:** Keep each pair together, and have them select “counties” in the “where” column on the site on their own devices. Have each pair relate two statistics about their county that are relevant to them, a family member, friend, or their community.

Have each pair turn to another pair (creating a group of 4) to share their findings.

Complete the [Conclusion Activity](#) to document student findings, and create a word cloud of the results.

Students may need help navigating the site, consider projecting to the board and exploring on your own to demonstrate. See [Assessment Strategies](#) below to assess student findings in step 6.

Some students may get stuck on finding “relevant” statistics. Float around and model how to find information on the site for students who need more support.



## Day 2 Outline

### Formative Assessment Notes

7. **Collecting Crowdsourced Data:** Facilitate the [Crowdsourcing a Data Table](#) activity (see [below](#)).

*Summary:* After completing the warm up from Day 1 and collecting all student answers, share the [CODAP \(Warm Up\)](#) with students so that they all have their own copy.

Once shared, break students into small groups and assign them an attribute to analyze. Students will:

- Brainstorm descriptive names for the attributes (columns) of the collected data.
- In small groups summarize the data in each attribute.
- As a class, create an **About Us Collage** to summarize the lifestyles and interests of the class.

8. **Research Finding Data Sets on the Internet:** Research in the suggested sites below or choose your own:

- [Kaggle All Data Sets](#)
- [Kaggle Small Data Sets](#)
- [Datasets by Year](#)
- [Google Dataset Search](#)

Students may not have access to Kaggle - consider downloading or compiling a shared folder of [data sets](#) or simply have students explore using google to find small data sets. Students should follow the steps below to find and upload their data:

- Each student should find a CSV data set of their choice, download and save it to their computer.
- Students should then create a new CODAP file and upload their CSV file into the document.
- Students should save this CODAP for future use in upcoming lessons.

Assess by having each group reach a checkpoint with you before they begin making their infographic. Look for accurate results in calculation, if applicable, and creativity in portraying their findings.

Consider suggesting to students they could bookmark these sites. (See [Extensions](#) below).

Consider having students turn in their CODAP and grading as a checkpoint, logging at least one data set for future use, for an ongoing activity in upcoming lessons.



## Day 3 Outline

### Formative Assessment Notes

9. **Class Demonstration:** Have each student bring their writing utensil for the day to the front of the room and set them next to each other.



Once all writing utensils are present, draw a table on the board and have students begin to identify attributes that could be useful when organizing data that describes writing utensils.

End result could produce something similar to table below:

	Type	Writing Color	Exterior Color	Functionality	Erasable
1.	pen	blue	transparent	cap	yes
2.	pencil	gray	purple	click	yes
3.	marker	orange	orange	cap	no
4.	pencil	gray	green	click	yes
5.	pen	red	red	click	no
6.	highlighter	yellow	yellow	cap	no

10. **Finding Data in Nature:** After the demonstration, have students explore outside and collect 15 - 20 artifacts (data elements) . These could be rocks, leaves, weeds, flowers, litter, etc..

Have students complete the [Finding Data in Nature Activity Guide](#) to create a data set (you will use the data set in upcoming lessons).

11. After returning inside, have them create a copy of the following [CODAP \(Finding Data in Nature\)](#) to complete the activity. The outline of the CODAP document matches [this worksheet](#). You may choose to provide students both for support while outside.

*Summary:* Students will create a cohesive data table with at least four attributes that describe each of their collected items. Students should save this CODAP for future use in upcoming lessons.

12. **Exit Ticket:** Have students complete the *Exit Ticket*

Assess students' data sets to make sure they have chosen fields that make sense for their object categories.

Consider having students submit their CODAP so that you can provide individual feedback.

See [Assessment Strategies](#) below.

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Conclusion Activity

Have students complete this Google form ([view form](#) or [make a copy](#)), and show students the responses,

- *What is one question you can think of that could be answered using local census data?*
- *What are the two statistics about your county you found relevant? (enter as separate responses)*

Once the class has entered and collected their responses, create a word cloud to interpret any trends or commonalities in student responses about the community.

Demonstrate using the collected data to create a word cloud visualization by following the steps below:

1. Highlight and copy (ctrl + C) the entirety of column A in the responses spreadsheet
2. Go to the following website: [Word Cloud Generator](#) (or a site of your choice)
3. Paste the data (ctrl + V) into the *Paste/TypeText* textbox.
4. Change any of the custom options if desired.
5. Click visualize to produce a word cloud output below the text.
6. If desired, you can save the word cloud by clicking download as PNG.

**Exit Ticket** (See [here](#) for printable copies)

Name: \_\_\_\_\_

Date: \_\_\_\_\_

**DIRECTIONS:** Consider the image following images:



*What attributes could be collected if we stored each image as a data element?*

**Possible Answer:** *Some attributes that could prove useful in data analyzation could be [season, terrain, people, body of water, buildings or architecture, foliage, etc].*

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

*Activity 2 - Collecting Data in Nature:* A collection of artifacts could be presented to students instead of having them go outside to collect their own.

- Students could choose to collect artifacts in the room instead of outside.

*ELL Accommodation for Finding Datasets:* Have a collection of datasets printed in english or other desired language and stored in your classroom

### Extensions

*Bookmark Data Sets Folder Demo:* Students create a folder on the google chrome toolbar to bookmark and store common sites used to find data sets. Throughout the course students will need to find data sets, so organizing their sites could prove useful. Follow the steps below to demonstrate for students:

1. On the google chrome home page - right click on the toolbar under the address bar.
2. Select the Add Folder.. Option
3. Create a folder named Data Set Sites
4. Go to the site you wish to bookmark and select the three dots on the upper right hand corner.
5. Choose Bookmarks > Bookmark this tab..
6. Name the website and choose the folder you created in step 3.
7. Press Done.

*Explore Finding Data Sets in the Library:*

- Instead of having students search the internet for a complete data set, reserve the school library for the period and have students explore datasets in books or documentation.
- Your school librarian should be able to give you more information on the resources they have access to.


*Explore - Finding Data about You:*

- Have students simply google themselves, parents/guardians, or friends.
- Students should conduct their research with the intentions to answer the following questions:
  - "How much data can I find about myself?"
  - "How did this data become available? Did I provide the data myself?"
  - "How comfortable am I with this data being easily accessible?"
- Consider having students share their findings to classmates in pairs, groups, to the whole class

## Teacher Directions - Crowdsourcing a Data Table

*An activity for a class to collect data and create a data table by Christa VanOlst.*

Upon completion of collecting all student answers into the [CODAP \(Warm Up\)](#) from step 1 of Day 1. Discuss with students how we just collected and organized our very first data set!

1. Share the [CODAP \(Warm Up\)](#) with students so that they all have their own copy.
  - Consider posting the link on your school platform.
2. Pose the questions to students ["Would this be considered a fully complete data table?", "Is there anything missing from our data set?"]
  - The conversation should produce the importance of descriptively named attributes.
3. Explain how the column headers of a table are called **attributes** of data points and how they can be useful in analysis used to find trends, patterns, or summaries of our data points/elements.
4. Have the students, in their own CODAP document, name each attribute with a meaningful, descriptive, and short name.
5. Break the students into small groups and assign each group an attribute.
6. Have each group analyze their attributes for information about their classmates by creating a visualization in CODAP.
  - Jobs Attribute: [example here](#)
7. Each group will then export their visualization from CODAP using the camera button on the bottom of the graph  toolbar.
8. Groups will then create a half page infographic on their findings:
  - Students can create a google doc, google drawing or use an external site like [venngage](#), [Canva](#), etc.
  - Encourage students to be artistic when creating their infographic, they should add in any images or drawings they find relevant to their attribute to make it aesthetically pleasing at a glance.
9. Once each group is finished, have all groups submit their infographics to you.
  - Combine the results digitally or print and glue them to a poster or large sticky.
  - The end result will be an **About Us Collage** to summarize the lifestyles and interests of our class - [example here](#).

## Worksheet - Finding Data in Nature

---

Use this worksheet to help you analyze the 15-20 artifacts that you collected outside. In this assignment, we are paying special attention to defining attributes and exploring the ability to recognize, store, and organize data.

1. What was the object you collected? How many did you collect (this is known as sample size)?
2. Why did you collect this object? At what location did you find these items?
3. What attributes do all of your items possess that could lead to analysis after organizing?  
*Must identify at least four attributes.*
4. Sketch a Data Table below using the attributes and start tracking your artifacts!!
5. After filling in your data table: What are a few questions that you could explore with this data set?



# Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

**DIRECTIONS:** Consider the image following images:



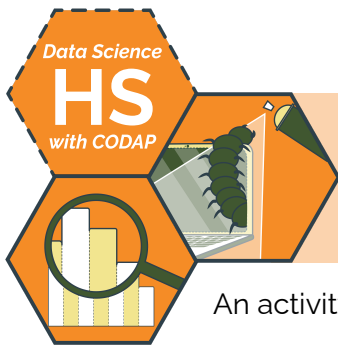
*What attributes could be collected if we stored each image as a data element?*

Name: \_\_\_\_\_

**DIRECTIONS:** Consider the image following images:



*What attributes could be collected if we stored each image as a data element?*



# Preparing Data

An activity guide covering the basics of data cleaning by Sara Fergus & Christa VanOlst

## Summary

In this lesson, students will explore data cleaning techniques including removing unwanted outliers, handling missing data, formatting data, and removing irrelevant data so that visualizations and models can be successful. In their explorations, students will learn to consider how data cleaning decisions could introduce bias, and how to make strong data cleaning decisions.

*Note: This lesson is similar to the Preparing Data lessons the CodeVA [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes a CODAP activity, which is omitted from the other versions.*

## Objectives

*The students will be able to . . .*

- Articulate the importance of data cleaning
- Employ basic data cleaning techniques in Python

## Standards Alignment

- **DS.4:** The student will be able to identify biases in the data collection process, and understand the basic ethical implications and privacy issues surrounding data collection.
- **DS.8:** The student will be able to acquire and prepare big data sets for modeling and analysis.
- **DS.12:** The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.

## Materials

- Warm Up Sleep Survey ([view](#) or [make a copy](#))
- *Data Cleaning Considerations* Worksheet (see [below](#), 1 per student)
- Video: [Coded Bias](#) (make sure this resource isn't blocked)
- Extension: *How is Face Recognition Surveillance Technology Racist?* ([web](#) or PDF)
- *Data Cleaning Scenarios* Cards ([printable PDF](#), printed & cut out, 1 per student group)
- Example Data (Messy) ([CSV](#))
- CODAP Assignment: Data Cleaning ([codap](#))

## Vocabulary

Term	Definition
Data Cleaning	Data cleaning is the process of preparing data for analysis. Often, there are mistakes in datasets that can skew the results of your analysis or even prevent the computer from properly running an analysis at all. Data cleaning finds errors and fixes them.
Messy Data	Messy data is a data set that has not been cleaned/prepared.
Bias	In this context, bias refers to anything that shifts the data analysis further from the truth. Commonly, bias is introduced when all records of a certain group are systematically excluded or misinterpreted.
Missing Values	Missing values are attributes in data that are not filled. Depending on the data set, they may be indicated with N/A, NaN, 0, -1, -, a blank space, or something else.
Missing at Random (MAR)	Data is missing at random if there is no pattern in data that is missing. This type of missing data reflects unintentional human errors. Cases with values missing at random could be dropped without introducing bias.
Missing not at Random (MNAR)	Data is missing <i>not</i> at random if there <i>is</i> a pattern in the data that is missing. This may indicate that a particular group of people were not able or did not want to provide a certain piece of data, or some other systematic data missingness. Removing these records would introduce bias.
Duplicate Cases	Duplicate cases are when there are two cases that are identical in every field. Sometimes, this can be valid. Other times, it may indicate a human error.
Mismatched Data Types	Data types are mismatched when the computer is interpreting attributes in one way, but the data is actually a different type. This often happens when numbers are spelled out, and so the computer interprets the attribute to be descriptive/strings or objects when it should be numeric/floats or ints.



# Outline

## Formative Assessment Notes

1. **Journal Warm-Up:** have students take [this survey](#). The data collected from this survey will be used to start the lesson. Ahead of time, put in bad responses that point to data cleaning problems (responding 8 with one, but 8 hours with another). See the [example](#) below.

Show students the results. What do they notice? What do they wonder? Have them respond in their journals and then share with a peer.

2. **Reading:** Have students read and annotate the [Data Cleaning Considerations Worksheet](#) (Parts 1–2), or read it all together. It discusses common issues in messy data, and what to consider when cleaning data.

Facilitate a discussion on data cleaning:

- Encourage students to share errors that might need to be fixed or considerations that might need to happen that weren't included in the worksheet.
- Prompt students to consider the effects of data bias

3. **Video:** Watch [Coded Bias](#) all together, which talks about racial bias in machine learning. Have a discussion with your students.

Once students have understood how the bias exists, prompt them to consider what effects this might have on technology.

*Extension: Read [this article about facial recognition & racial bias](#)*

4. **Discussion:** Split students into groups and give them each one [data cleaning scenario](#), which describes a piece of messy data and how a data scientist fixed it. The scenario card then asks questions about whether the right decision was made. Have students answer the questions on the card. Then, have each group share with the class and discuss.

Consider having students write answers to the scenario questions on paper or poster board before presenting.

5. **Exit Ticket Cleaning Data Activity:** Give students [this CODAP file](#) and have students apply the data cleaning techniques learned today. Then, in the CODAP worksheet (or [printed version](#)) have students describe the changes they made and consider any bias they may have introduced.

See [discussion essential understandings](#) below for assessment information

See [discussion essential understandings](#) below for assessment information

Guide students to understanding that leaving out certain faces in training data amplifies racial bias.

During the discussion, guide students to the conclusion that there can be lots of different errors in one data set, but what the error is and what the goal is can change your decision.

Make sure that students are consistently considering what bias they may be introducing with their data cleaning. See [Exit Ticket](#) below.

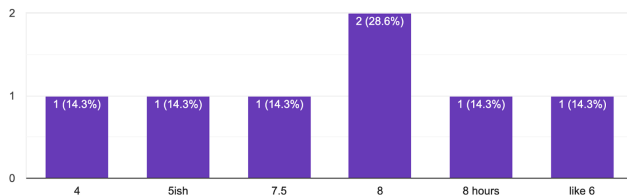
## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

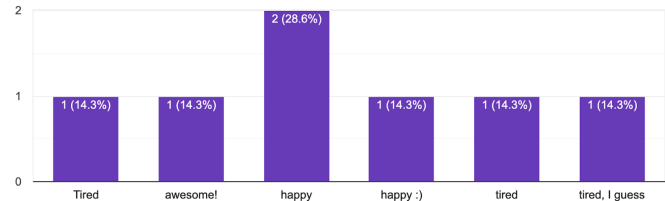
### Warm Up Example

The goal is for the results to look something like this:

How many hours did you sleep last night?  
7 responses



How are you feeling today?  
7 responses



You can see that a few answers were repeated (you will need to have repeated answers in order to get this bar chart), but some should have been put together and weren't. You can also point out that the numbers are not in order, because the computer thinks that they are words.

### Discussions Throughout the Lesson

This lesson includes a lot of discussion. Throughout the lesson, guide students to these essential understandings:

Essential Understanding	Discussion Number(s)
Surveys should be created to keep data clean. You could achieve this with response validation (the answer must be a number) or suggested responses in the form of multiple choice questions	1
There are a lot of mistakes that can be in data sets. For example: <ul style="list-style-type: none"> <li>Missing values</li> <li>Different values that mean the same thing</li> <li>Mismatched data types</li> <li>Duplicate cases</li> <li>Answers that don't make sense</li> </ul>	1, 2
Data cleaning, if not done carefully, can introduce bias and silence the voices of specific groups. One example of when this is done is when data is "missing not at random."	2, 3

**Exit Ticket (Key)** – Print the [blank version below](#) & distribute to students

Name: \_\_\_\_\_

Date: \_\_\_\_\_

After cleaning the [Roller Coaster \(messy\) data set](#), list 5 suggestions you have for cleaning this data**Student answers may vary**

<b>Data Cleaning suggestion</b>	<b>Can you think of any bias this may present? If so, What impacts does this bias have?</b>
Delete the duplicated cases on rows 13 and 14	Would need to confirm these are actual duplicates and not two roller coasters that are named the same.
Replace the Seventy and Ninety-Eight with their numerical values on rows 5 and 34	
Replace the WOODEN typo with Wooden on rows 5, 20, 66, 77, and 92, so the values will be grouped together during analysis.	
Delete the cases that are missing values for the Duration attribute during analysis.	This could exclude a certain park entirely making our analysis less inclusive, I would need to double check that this would actually be necessary, meaning I may not even be using this attribute in my analysis.
Look up the Roller Coasters that are missing cities.	I would need to make sure my research is accurate.

## Some Accommodations & Extensions

*Note: All students benefit from accommodations; consider implementing the accommodations below for everyone*

### Accommodations

Some students may benefit from receiving some of the extra resources below, like the [data cleaning guidebook](#), to help them draw conclusions.

In classes with a large number of students who have small group accommodations, all discussions can be done within a small group of students rather than all together. This could also allow students to work at different paces, and to discuss the information at the level of rigor that makes sense for them.

Some students may benefit from having the data cleaning scenarios or the Python worksheet ahead of time to prepare.

You could support students who are learning English by providing them with the vocabulary table above.

Students with cognitive disabilities or students who are learning English could benefit from the [adapted version of the resources](#).

### Extensions

One extension is included within the plan: students may read [this article](#) to get a better understanding of the effects of bias in technology. In addition, Kaggle has an advanced [data cleaning tutorial](#) that could be used as an extension.

## Other Resources

- Consider using the [Data Cleaning Guidebook](#), which is an online PDF book that dives deeper into data cleaning and the errors that could arise
- For more advanced data cleaning activities consider exploring the [Kaggle data cleaning activities](#)
- Consider using this article ([Data Cleaning Article](#)) to reiterate how AI can be unintentionally biased and how data cleaning and awareness can help prevent the problem
- Here is a list of the Virginia Department of Education [Data Cleaning Resources](#)

# Data Cleaning

A description of data cleaning considerations (Part 1) by Sara Fergus

## What makes Data “Messy”?

Take a look at this “messy” data set. List as many things as you can think of that make this data set “messy”. Then, describe how you might fix the problem. One has been completed for you. Come up with at least 3.

Coaster	Park	tsoc	Max_Height	Drop	Length	Duration	Type	Design	Year_Opened	Age_Group	Inversions	Num_of_Inversions
Rampage	VisionLand	56.0	120	102.0	3500.0	NaN	Wooden	Sit Down	2003	newest	N	0.0
Arkansas Twister	Magic Springs and Crystal Falls	NaN	95	92.0	3340.0	NaN	Wooden	Sit Down	2000	newest	N	0.0
Big Bad John	Magic Springs and Crystal Falls	37.0	32	41.0	2349.0	180.0	Steel	Sit Down	2002	newest	N	0.0
X	Six Flags Magic Mountain	76.0	175	215.0	3610.0	NaN	steel	4th Dimension	2002	newest	Y	2.0
Giant Dipper	Santa Cruz Beach Boardwalk	55.0	Seventy	65.0	2640.0	112.0	Wooden	Sit Down	1924	older	N	0.0

What makes it messy	How I could fix it
I am not sure what tsoc means	Figure out what it means and rename that column to make more sense.

# Data Cleaning

A description of data cleaning considerations (Part 2) by Sara Fergus

## Preparing Data Considerations

It is very important to “clean up” messy data, so that your analysis can be accurate. However, you could accidentally change the outcomes of your analysis by cleaning your data incorrectly. So, it is important to make the best “data cleaning decisions” that you can. Read through this list of considerations. Annotate as you read by writing ideas and questions, highlighting important points, and underlining vocabulary.

### Consideration #1: Is it an error?

Before doing any data cleaning, it is important to consider whether the changes you are making are actually cleaning an error.

Name	Age
John	13
Danny	-10
Xavier	32
Kyra	100
Alysia	45
Carl	150

For example, you may have a data set with people's ages (left). This data set is messy because it contains data that doesn't make sense, like someone being -10 years old. Probably, someone typed a minus by accident. You would want to clean that issue. 150 years old also doesn't make sense. Maybe someone typed a 0 at the end by accident, and are actually 15. However, you have to pick a cutoff for what *does* make sense. One cutoff option is 100. While it is possible to be 100 years old, based on the rest of the data it seems unlikely. Maybe they typed an extra zero and are actually 10 years old. In this case, you should go see what the data represents to decide whether 100 years old is an error.

Duplicate values (right) is another possible error. Here, Danny is listed twice, earning 240 points both times. It is possible that this is an error – maybe the computer reloaded and resubmitted. However, maybe Danny actually did earn 240 points twice in a row, or maybe there are multiple people named Danny who scored 240 points. In this case, you need to decide whether this is an error. If you have more information, for example when the data was collected, that would be helpful. You could also consider things like how likely it is for someone to get the same score twice.

Name	Score
John	230
Danny	240
Danny	240
Kyra	423

**What in this dataset is definitely an error? What might be an error? What would help you decide?**

Date	Temperature (F)	Definitely an error:
Jan 14	30	
Jan 15	32	
Jan 16	60	
Jan 17	31	
Jan 16	300	
Jan 16	32	
		Maybe an error:

## Consideration #2: How should I clean it?

There are a lot of decisions you could make about how to clean certain data. Here are some methods:

### 1. Fix the error by hand.

This works if there is not a lot of data, the errors are easy to see, and what the survey taker meant is obvious. For example, in the data set to the right, the Giant Dipper has a Max Height of "Seventy". The computer is going to interpret that to be different than the number 70, but we know that they are the same so we could make the fix on our own.

Coaster	Park	tsoc	Max_Height
Rampage	VisionLand	56.0	120
Arkansas Twister	Magic Springs and Crystal Falls	NaN	95
Big Bad John	Magic Springs and Crystal Falls	37.0	32
X	Six Flags Magic Mountain	76.0	175
Giant Dipper	Santa Cruz Beach Boardwalk	55.0	Seventy

Sometimes, this may be a harder decision. For example, in the age data set, it is likely that the person who put in -10 meant to put in 10. However, maybe they didn't. It is up to you to decide how likely it is that that mistake was made and how you should clean it.

### 2. Replace messy data with 'N/A'. Often, missing data is represented by 'NaN' or 'NA'. There are two times to replace messy data with N/A.

- Some data sets use other things to indicate that data is missing. They may put a blank space, a dash, a zero, a negative one, or something else. Go to your data source and determine how missing data was indicated and consider replacing with N/A
- If there is an error, but you want to include the case in general, you could replace the error with N/A. For example, if you are not confident in *why* someone put -10, you could replace -10 with NA. Then, it won't be considered in your calculations

### 3. Use other columns. Sometimes you can deduce what a value should be based on other columns. For example, the data set to the right says that the area of one of the squares is "banana". However, since we know the side length of a square, and we know that the area of the square is side multiplied by side, we can calculate and replace "banana" with 16.

Side Length of Square	Area
2	4
4	banana
3	9
2	4

- Drop the case.** The most common thing to do is to drop the case. This means that the row with the error will be completely removed from the data set. This is common, but takes a lot more consideration, which we will talk about in considerations 3 and 4.
- Something else.** There are a lot of other methods you can use. You could do some research to find the correct information, or re-collect data. You may choose to not consider a column with a lot of errors at all.

## What Data Cleaning Decisions would you make?

Date	Temp (F)	Snow?
Jan 14	30	Yes
Jan 15		Yes
Jan 16	60	No
Jan 17	31	No
Jan 17	31	No
Jan 16	300	Yes
Jan 16	32	87

If you are interested in graphing the temperature over time, what data cleaning decisions would you make with this dataset?

**Consideration #3: Am I introducing any bias?**

There are lots of ways that you could introduce bias in your data cleaning. For example, if you decide that an age over 100 must be an error, and it is in fact *not* an error, you could be introducing bias against the extremely elderly. The most common way to introduce bias is by dropping cases with missing values.

When we analyze missing values, we see two main types of missing data:

**Type 1: Missing at Random**

Sometimes, people skip over questions for no reason at all. For example, in the roller coaster data set, *tsoc* stands for "Top Speed of Coaster". This is missing for the Arkansas Twister. Probably, someone just forgot to fill that out.

**Type 2: Missing not at Random**

Sometimes, data is missing not-at-random. This could be because people are afraid or uncomfortable, they don't feel that they can accurately answer the question, or something else.

Are you a citizen of the United States?

- ☐ Yes
- ☐ No

For example, this question (left) might be missing not-at-random. Someone who is not a citizen of the United States may worry that answering this question could get them in trouble and might not fill this out. By dropping missing data in this column, the voices of a specific group of people are being ignored.

This question (right) might go unanswered because someone does not feel that they can accurately answer this question. This could be someone of mixed race or someone of a race that is not listed. By dropping missing data in this column, your data will only include people who are purely white, black, or hispanic.

What is your race?

- ☐ White
- ☐ Black
- ☐ Hispanic

You can only remove cases for missing data if the data is missing at random.



There have been many cases of missing data leading to biased results. One example, from stem equity, is quoted below:

*For example, if a researcher follows the recommendation of Coletta & Steinert (2020) and removes the data for students who have pretest scores over 80%, then they are selectively removing data from students with the strongest physics backgrounds. As Van Dusen & Nissen (2019a) showed, these students are most likely to be white men. In high performing classes, this data cleaning technique will likely make differences in performance across groups appear artificially small.*

In education, data is most often missing from students with lower grades (Nissen, Donatello, & Van Dusen, 2019). Another example could be a temperature sensor breaking down and not providing data. While this might happen randomly, it could also be because the sensor does not work at a certain temperature (for example, if it is over 100°F), and so the missing data actually leaves out important patterns.

**What bias could be introduced if this data were improperly cleaned?**

### Consideration #4: Do I have enough information left?

If the data is too messy, you may not want to use it at all. One issue with data that is too messy is that the decision to drop messy cases could remove a large portion of the data, and so you don't have enough data to analyze anymore. To avoid this, you could remove only the cases that are missing data in the columns you are analyzing at the time. For example, if you are looking at the relationship between a person's height and their weight and their "name" is missing, you could choose to not remove that case for that part of your analysis. Always check to make sure that the majority of the data is usable!

### Works Cited

- Coletta, V. P., & Steinert, J. J. (2020). Why normalized gain should continue to be used in analyzing pre-instruction and post-instruction scores on concept inventories. *Physical Review Physics Education Research*, 16(1). <https://doi.org/10.1103/physrevphyseducres.16.010108>
- Nissen, J., Donatello, R., & Van Dusen, B. (2019). Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research*, 15(2). <https://doi.org/10.1103/physrevphyseducres.15.020106>
- NSF. (2021, August 6). *Data Cleaning - stem equity - empowering diversity of research in STEM Education*. STEM Equity. Retrieved July 11, 2022, from <https://stemequity.net/data-cleaning/>
- Pereira, T. (2020, February 2). *The problem of missing data*. Medium. Retrieved July 11, 2022, from <https://towardsdatascience.com/the-problem-of-missing-data-9e16e37ef9fc>
- Van Dusen, B., & Nissen, J. (2019). Equity in college physics student learning: A critical quantitative intersectionality investigation. *Journal of Research in Science Teaching*, 57(1), 33–57. <https://doi.org/10.1002/tea.21584>

# Data Cleaning

A description of data cleaning considerations (Part 2 ADAPTED) by Sara Fergus

## Preparing Data Considerations

It is very important to “clean up” messy data, so that your analysis can be accurate. It is also important to make the best “data cleaning decisions” that you can. Read through this list of considerations. Take notes as you read by writing ideas and questions, highlighting important points, and underlining vocabulary.

### Consideration #1: Is it an error?

Before doing any data cleaning, it is important to consider whether the changes you are making are actually cleaning an error.

Name	Age
John	13
Danny	-10
Xavier	32
Kyra	100
Alysia	45
Carl	150

For example, a data scientist should decide if the three unusual ages (-10, 100, and 150) are mistakes or not.

### Consideration #2: How should I clean it?

There are a lot of decisions you could make about how to clean certain data. Here are some methods:

1. *Fix the error by hand.*

This works if there is not a lot of data, the errors are easy to see, and what they meant is obvious.

Coaster	Park	tsoc	Max_Height
Rampage	VisionLand	56.0	120
Arkansas Twister	Magic Springs and Crystal Falls	NaN	95
Big Bad John	Magic Springs and Crystal Falls	37.0	32
X	Six Flags Magic Mountain	76.0	175
Giant Dipper	Santa Cruz Beach Boardwalk	55.0	Seventy

For example, in the data set above, the Giant Dipper has a Max Height of “Seventy”. You might want to change it to 70.

2. *Replace messy data with 'N/A'.* Often, missing data is represented by 'NaN' or 'NA'. If there is an error, but you want to include the case in general, you could replace the error with N/A.
3. *Use other columns.* Sometimes you can figure out what a value should be based on other columns.

Side Length of Square	Area
2	4
4	banana
3	9
2	4

For example, the data set to the right says that the area of one of the squares is "banana". But since we know the side length, we know it is actually 16 and can replace it.

4. *Get rid of the row.* This is common, but could introduce bias
5. *Something else.* There are lots of methods out there!

### Consideration #3: Am I introducing any bias?

Bias is when you get rid of data that is important. When you get rid of specific data, you might be ignoring a specific group of people. But, you might want to get rid of certain rows if there is a lot missing.

**They might be "missing at random".** Sometimes, people skip over questions for no reason at all. You can drop these.

**Or, they could be Missing not at Random.** Other Times, data is missing not-at-random. This could be because people are afraid or uncomfortable, they don't feel that they can accurately answer the question, or something else.

Are you a citizen of the United States?

☐ Yes

☐ No

For example, This question might be missing not-at-random. Someone who is not a citizen of the United States may worry that answering this question could get them in trouble might not fill this out. By dropping missing data in this column, the voices of a specific group of people are being ignored.

***Consideration #4: Do I have enough information left?***

When you get rid of data, be sure to only get rid of the data that you need to. Sometimes, if you aren't answering a question about a certain attribute, you don't need to get rid of things just because that attribute is missing.

Make sure that you check to see if you dropped so much of the data that it is not useful anymore!

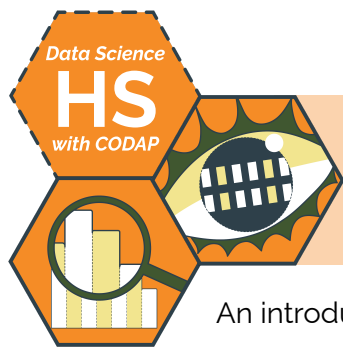
# Printable Exit Tickets (Print this page & distribute to students)

Name: \_\_\_\_\_

Date: \_\_\_\_\_

After cleaning the [Roller Coaster \(messy\) data set](#), list 5 suggestions you have for cleaning this data set.

Data Cleaning suggestion	Can you think of any bias this may present? If so, What impacts does this bias have?



# Choosing Visualizations

An introduction to exploring and selecting types of visualizations by Christa VanOlst

## Summary

In this lesson students will explore how visualizations can serve a variety of purposes in communicating data. Throughout the lesson students defend style and chart type to emphasize the power of a visualization over a data table. Students then discover and categorize chart strengths and weaknesses in order to support a question statement. Using local news articles, students will then justify the missed opportunity of a powerful visualization and then analyze the article to propose and sketch a visualization. In conclusion, students practice exploratory data analysis to create effective visualizations.

*Note: This lesson is similar to the Choosing Visualizations lesson plans from the CodeVA [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes an extra day for CODAP activities.*

## Objectives

- Students will defend the use of chart types and styles in visualizations.
- Students will interpret the strengths of visualization types.
- Students will create emerging visualizations using their data collected from Lesson 3.

## Standards Alignment

- **DS.1:** The student will identify specific examples of real-world problems that can be effectively addressed using data science.
- **DS.6:** The student will justify the design, use, and effectiveness of different forms of data visualizations.
- **DS.10:** The student will summarize and interpret data represented in conventional visualizations.

## Materials

- Large Stickys or Poster Paper, Construction Paper, Tape/Glue, Stickers
- Student Whiteboards
- Steph Curry Shooting Stats ([table](#) & [heat map](#))
- Steph Curry Visualizations Slides ([view](#) or [make a copy](#))
- *Desmos Interactive Notes: Choosing Good Visualizations* (view on [Desmos](#))
- Google Questions & Goals Cutout (see [below](#)) & Google Visualizations Slides ([view](#) or [make a copy](#))
- Day 2 Warm-Up Slide (view [Google Slide](#) or [make a copy](#))
- *Visualizations in the News* Handout (see [below](#) or [make a copy](#) of the [Google Doc version](#))
- CODAP Exploratory Analysis (Teacher Directions, see [below](#))
- CODAP Files ([Weather](#), [Instagram](#), [Dogs](#), [US Cities Sample Data Set](#), [Finding Data in Nature](#))



## Vocabulary

Term	Definition
Data Representation	A data representation is a way to visualize and organize collected information
Visualization	A representation of information in the form of a chart, diagram, picture, infographic, etc. for an audience.
Scatter Plot	Graphical representation of the relationship between two numerical sets of data.
Bar Chart	Graphical representation of categorical data created by grouping data into rectangular bars, usually color coded, to represent the frequency of the categories. The bars can be horizontal or vertical.
Histogram	Graphical representation of numerical data created by grouping it into “bins” to show frequency within a range of values.
Box Plot	Graphical representation of the median value, spread and skewness of data through their quartiles.
Line Plot	Graphical representation which portrays data as a continuous series of data points connected by straight line segments.
Pie Chart	Graphical representation which shows comparative data including parts of a data set vs. the entirety of a data set.
Heat Map	Graphical representation which shows data in the form of a map or diagram in which results are represented as colors varying in intensity.

## Before the Lesson

This lesson requires a fair amount of printing and preparation be sure to prepare the following print materials in advance of class time:

- The [Google Questions & Goals](#) cutouts—print 1 per small, student group, trim along the dotted lines
- The [Google Visualizations](#) slide deck—make a copy for students to view during class, or print the images; 1 per small student group)
- The [Visualizations in the News](#) handout, which students must fill out using Google Docs—be sure to have a copy ready for students to use as a template.

There are also many different visualizations that serve as discussion points throughout the lesson; be sure to review them and be prepared to prompt student inquiry regarding what they represent!

## Day 1 Outline

### Formative Assessment Notes

1. **Communicating with Data Warm-Up:** Show/distribute [this chart](#), which shows data about Stephen Curry's basketball shooting stats. Have students share one piece of information from the data table that communicates something to them.

Next, display this [heat map](#) visualization (see [Vocabulary](#)). Ask students what they notice about the visualization, and then have the students identify & discuss one piece of information the visualization quickly communicates.

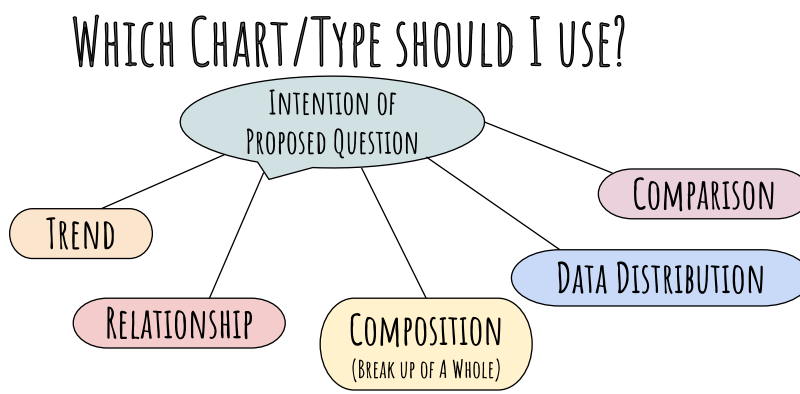
Finally, display [these visualizations](#) and have students respond to the following question in their journals:

***Which visualization best defends the statement “Steph Curry is the best shooter in the league”? Explain your answer.***

2. **Desmos Discussion:** Use the [Desmos Interactive Notes: Choosing Good Visualizations](#) to have students learn the different types and utilities of visualizations including scatter plots, histograms, pie charts, box plots, line plots, and heat maps.
3. **Supporting the Question:** Distribute the cut-out [Google Questions and Goals](#) slips and the [Google Visualizations](#) images.

Have students work in small groups (3-4) to match the questions/goals to the visualizations

Then, have students check their work using the [Google Analytics Documentation](#) site and match each question/goal & visualization pair to one of the visualization types *below*:



The emphasis of activity should be how visualizations can serve multiple purposes, however the important part is to choose the best one to support the given statement.

Monitor student responses for the ability to categorize the use of visualization types.

The intention of this activity is to have students identify the real world/industry need for visualizations.



4. **Designing Visualizations:** Have students complete the [Visualizations in the News](#) mini-project, where they analyze a news article and design a visualization that reinforces it.
5. **Optional Extension:** Use the [Teacher Directions - Jigsaw Exploratory Analysis](#) to demonstrate and explore with students how to create relevant visualizations given a data set.
6. **Conclusion Research:** Have students complete “rapid research” (research that takes under 10 minutes using, e.g., Google) to find real life examples of bad or misleading visualizations.

Consider showing students this [Ted Talk](#) (4 mins) and/or [these examples](#) to get the conversation started.

See [Assessment Strategies](#) below.

Have students share their findings with peers and monitor discussion to make sure students are successful in identifying misleading visualizations

## Day 2 Outline

Formative Assessment Notes

7. **Warm Up:** Display [this slide](#) on the board. Have students write in their journals what they notice about these graphs and what they wonder. They will be revisiting the graphs in the next step.
8. **Exploratory Analysis:** Use the [Teacher Directions - CODAP Exploratory Analysis](#) to demonstrate and explore with students to create relevant visualizations given a data set.

*Summary:* Students are presented with several data sets. Within each CODAP, there will be a data set and a question. Students will explore using relevant attributes to create emerging visualizations to support the question. Students will then generate their own questions.

9. **Visualizing your Data from Nature:** Have students revisit their [CODAP \(Finding Data in Nature\)](#) assignment from Lesson 5 step 11.

Students should use their set from last class to revisit their recorded questions and use visualizations that are relevant to support these thoughts.

See [directions](#) for assessment strategy

Consider having students submit their CODAP so that you can provide individual feedback on their visualizations. (see [Assessment Strategies](#) below)

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Visualizations in the News

In this activity, students find a new article that could benefit from a visualization, and create one that supports the content of the article. Have students use the [Visualizations in the News Guide](#) to complete the assignment in Google Docs.

Consider having students switch assignments and verbally give each other feedback about their peer's visualizations. Student visualizations may not be entirely accurate, but should loosely support the information in their chosen artifact/article.

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Article Choice</b>	Students chosen article is local on a state or community level <b>AND</b> the article presents a missed opportunity for the use of a visualization.			
<b>Visualization Sketch</b>	Visualization sketch depicts the data described in the article <b>AND</b> includes scaled axes.			
<b>Visualization Style and Choice</b>	Visualization choice and style is appropriate for the data described in the article. (eg. Line Graph, Scatter Plot, Histogram, ... )			

## Some Accommodations & Extensions

**Optional Pre-Assessment to Presentation:** Previously print and cut out all of the visualizations from the presentation. Give each student one image and have them place their visualization in the category they think it belongs. *Most students should have some prior knowledge of some types of charts but not all.*

Types of Visualizations						
Scatter Plot	Histogram	Pie chart	Bar Chart	Box Plot	Line Plot	Heat Map

**Extension Activity:** Once you finish the lecture portion of the lesson, have students create visualizations based on made up data using manipulatives. Then tape or use magnets to display their creation on the walls around the room. Use a gallery walk approach to have students explore each other's visualizations and come up with interpretations. Prepare multiples of each chart type (depending on class size):

- Scatter Plot Material (Give students a small data set and use glue to paste data points)
- Histogram Material (Give students a small data set and use scissors to cut appropriately sized rectangles for bins. Then glue to paste bars on the chart.)
- Box Plot Material (Give students a small data set and use popsicle sticks to glue key data points such as min value,  $Q_1$ , median,  $Q_3$ , and max values. Use glue to paste popsicle sticks)
- Pie Chart Material (Give students a small data set and use scissors to cut appropriately proportions. Use glue to paste into a circle)
- Heat Map (Give students a completed heat map but in black and white. Have students use colored pencils to create a color intensity scale and have students color each data point appropriately.)

## Steph Curry Is One of The Best (Data Sheet)

Records of his Shots during the 2015-2016 regular season

<i>SHOT DISTANCE (5FT)</i>	<i>FGM</i>	<i>FGA</i>	<i>FG%</i>	<i>3PM_</i>	<i>3PA</i>	<i>3P%</i>	<i>EFG%</i>	<i>BLKA</i>	<i>FGM (%AST)</i>	<i>FGM (%UAST)</i>
<b>2015-16</b>	805	1598	50.4	402	886	45.4	63	52	46.6	53.4
<b>Less Than 5 ft.</b>	272	422	64.5	0	0	0	64.5	31	40.4	59.6
<b>5-9 ft.</b>	35	72	48.6	0	0	0	48.6	10	22.9	77.1
<b>10-14 ft.</b>	29	57	50.9	0	0	0	50.9	1	31	69
<b>15-19 ft.</b>	38	102	37.3	0	0	0	37.3	4	28.9	71.1
<b>20-24 ft.</b>	158	335	47.2	129	276	46.7	66.4	4	66.5	33.5
<b>25-29 ft.</b>	251	563	44.6	251	563	44.6	66.9	1	51	49
<b>30-34 ft.</b>	15	26	57.7	15	26	57.7	86.5	0	20	80
<b>35-39 ft.</b>	2	5	40	2	5	40	60	0	0	100
<b>40+ ft.</b>	4	14	28.6	4	14	28.6	42.9	1	0	100

<i>SHOT AREA</i>	<i>FGM</i>	<i>FGA</i>	<i>FG%</i>	<i>3PM_</i>	<i>3PA</i>	<i>3P%</i>	<i>EFG%</i>	<i>BLKA</i>	<i>FGM (%AST)</i>	<i>FGM (%UAST)</i>
<b>Restricted Area</b>	263	399	65.9	0	0	0	65.9	29	39.5	60.5
<b>In The Paint (Non-RA)</b>	55	113	48.7	0	0	0	48.7	12	32.7	67.3
<b>Mid-Range</b>	85	200	42.5	0	0	0	42.5	7	32.9	67.1
<b>Left Corner 3</b>	30	63	47.6	30	63	47.6	71.4	0	96.7	3.3
<b>Right Corner 3</b>	27	53	50.9	27	53	50.9	76.4	1	85.2	14.8
<b>Above the Break 3</b>	342	757	45.2	342	757	45.2	67.8	2	50.3	49.7
<b>Backcourt</b>	2	11	18.2	2	11	18.2	27.3	1	0	100

## Google Questions & Goals Cutouts (Sourced from [Google Analytics Documentation](#))

**Question 1:** How many new users are we (Google) acquiring every day?

**Goal:** Compare values (number of users) over time (days)

**Question 2:** What channels (mediums) are these new users coming from?

**Goal:** Display the composition of the data (which source users came from) over time (comparing the number of new users across days).

**Question 3:** Which referrers (other websites) are driving the most traffic to our website?

**Goal:** Compare values (number of sessions) across categories (other websites).

**Question 4:** Which referrers (other websites) tend to drive more traffic to our website from desktops, and which ones tend to drive more traffic from mobile devices?

**Goal:** Comparing values (number of sessions) across categories (other websites) and looking at composition within each bar (mobile vs. web traffic).

**Question 5:** How does the traffic from mobile and desktop stack up across referrers (other websites)?

**Goal:** Comparing values (number of sessions) across categories (other websites) in multiple dimensions (mobile and desktop).

**Question 6:** What time of day sees the highest number of users on our website?

**Goal:** Comparing values (number of sessions) over time (hours) across multiple dimensions (days).

**Question 7:** Which pages are driving the most engagement by channel (mediums)?

**Goal:** Look at the relationship between channels (mediums) and pages to see how the different combinations influence average session duration

**Question 8:** Where do we have opportunities to drive more traffic to high-performing web pages?

**Goal:** Show the relationship between values (conversion rates and number of sessions) to help pinpoint pages with high conversion rates that could be better promoted.

## Visualizations in the News Guide (Handout)

### Directions:

1. Research local or state news segments or articles on your school website, local paper site, google, youtube, or other appropriate sites.
2. Find a news video clip/article/post/blog about a local issue that does NOT include a data visualization, but would benefit from including one to help make the story easier to understand
3. Create a google drawing to sketch the layout and prediction of what you think a data visualization could look like if included in the article.
4. Complete the Assignment attached below.
5. Use the example [here](#) as a guideline.
6. Consider having students use the template below.

---

### Creative Title

*(copy and paste web page url here)*

*A short summary of what the article is describing. (2 - 3 sentences)*

*What kind of chart will you use?*

*What type of labels must be included with this chart?*

*Double click to  
sketch your  
google drawing*

## Teacher Directions - CODAP Exploratory Analysis

*An activity for students to identifying the goal beneath a question and connect it to a visualization.*

Consider limiting students to one CODAP and one posing each question at a time. Regroup after the exploration session to discuss/share possible outcomes after the exploration session.

Using the NOAA Weather Data in your area: [CODAP \(Weather\)](#)

- **Question:** How did the temperature change throughout the day yesterday?
  - **Possible Outcome:** Create a line graph using the when attribute on the x-axis and the Temp(F) attribute on the y-axis. Select by highlighting all the cases from a particular day in the dataset, choose the eye tool on the graph and select 'Hide Unselected Cases' (this will show the data for one day).
- **Question:** I wonder what direction the wind mostly blows in my town.
  - **Possible Outcome:** Create a histogram using the WDir attribute on the x-axis, naturally the data points will create a bar shape. In the configuration tool, check the fuse dots into bars button and change the bin size to 60 or 90 (think of degrees in a circle).

Using the Most Followed Instagram Accounts: [CODAP \(Instagram\)](#)

- **Question:** I wonder which profession has the most followers?
  - **Possible Outcome:** Create a stacked bar chart using the Professions attribute on the x-axis and the Followers In Millions attribute dropped within the graph to create a legend using color intensity.

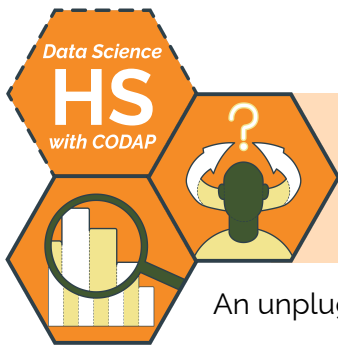
Use the Dogs Data: [CODAP \(Dogs\)](#)

- **Question/Thought:** "I bet there is a correlation between a dog's weight and lifespan."
  - **Possible Outcome:** Create a scatter plot by dragging the minimum weight attribute to the y-axis and the maximum life span attribute to the y-axis.

Use the World Cities Data: [CODAP \(US Cities Sample Data Set\)](#)

- **Question/Thought:** "The most populated cities in the US are on the borders."
  - **Possible Outcome:** Use the Map tool (this automatically plots latitude and longitude) and drag the population attribute to the middle of the map.





# Creating Visualizations

An unplugged introduction to creating data visualizations by Christa Van Olst & Sara Fergus

## Summary

In this lesson, students will learn techniques to prepare data for analysis and create basic visualizations in CODAP. Students will use their skills to transform a basic visualization into a non-traditional visualization. The lesson concludes with a mini-project where students show how using basic visualizations can support deeper analysis in later stages of the data cycle.

*Note: This lesson is similar to the Creating Visualizations lesson plans from the CodeVA [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes a CODAP activity, which the others omit.*

## Objectives

*The students will be able to . . .*

- Quickly sort and sub-sort a dataset.
- Transform a basic visualization into a freestyle representation of the same data.
- Create basic visualizations of provided data sets.
- Interpret data visualizations, and identify follow-up questions based on them

## Standards Alignment

- **DS.1** The student will identify specific examples of real-world problems that can be effectively addressed using Data Science.
- **DS.6** The student will justify the design, use, and effectiveness of different forms of data visualizations.
- **DS.10** The student will be able to summarize and interpret data represented in visualizations.

## Materials

- Craft supplies, include large sticky notes, poster paper, construction paper, tape/glue, stickers, markers/colored pencils, scissors, protractors, rulers, and large graph paper
- Using Basic Visualizations activity materials (see step #2), including [Teacher Directions](#) & [.codap](#) file
- Visualization Practice Station materials, (see step #5) including the student guide (see [below](#), view [Google Doc](#) or [make a copy](#))
- Mini-Project materials, including *Directions & Student Planner* (see [below](#), view [Google Doc](#) or [make a copy](#))



## Vocabulary

Term	Definition
Data Representation	A data representation is a way to visualize and organize collected information
Visualization	The art of representing information in the form of a chart, diagram, picture, infographic, etc. for an audience.
Categorical Data	Data that can be divided into groups or categories.
Quantitative Data	Data that can be tracked numerically.
Scatter Plot	Graphical representation of the relationship between two numerical sets of data.
Bar Chart	Graphical representation of categorical data created by grouping data into rectangular bars, usually color coded, to represent the frequency of the categories. The bars can be horizontal or vertical.
Histogram	Graphical representation of numerical data created by grouping it into "bins" to show frequency within a range of values.
Box Plot	Graphical representation of the median value, spread and skewness of data through their quartiles.
Line Plot	Graphical representation which portrays data as a continuous series of data points connected by straight line segments.
Pie Chart	Graphical representation which shows comparative data including parts of a data set vs. the entirety of a data set.
Heat Map	Graphical representation which shows data in the form of a map or diagram in which results are represented as colors varying in intensity.

## Before the Lesson

This lesson has some pretty labor-intensive parts, so it's important to plan ahead so you have time to set everything up. Review the *Materials* list and the *Outline* thoroughly ahead of time to make sure you have prepared everything you will need.

## Day 1 Outline

### Formative Assessment Notes

1. **The Advantages of Sorted Data:** Distribute this [CODAP: Warm Up - Quickly Sorting Data](#) to students. (*This CODAP continued in step 2*)

*Summary:* Given a large data set students discover that directly clicking on a single attribute name will open a drop-down menu. Students explore using the **Sort Ascending (A→Z, o→g)** option to quickly sort the cases alphabetically by their chosen attribute. Students then discover how the order in which you sort each column creates opportunities to produce a subsort of the cases in the table.

Have students answer the following questions in their journals titled **Why Sort Data?**:

1. *What is the difference between the multiple data sets?*
2. *In the season sorted sheet, how is it easier to identify the favorite season when sorted?*
3. *In the color sorted sheet, what relationships can you find between the students who like the color light blue?*
4. *What is the advantage of sorting data?*

Consider having students share their responses with a partner, and then facilitate a discussion with the class.

2. **Using Basic Visualizations:** Use the [Teacher Directions - The Use of Basic Visualizations](#) while students use the CODAP [file](#) from step 1.

*Summary:* During this activity, students create a basic bar chart to create a quick visual representation of their chosen categorical attribute from the *Basic Favorites* data set (view [CODAP](#)).

- *For example, if a student chose the season attribute, creating a bar chart can aid in quickly analyzing which season is favored.*

Afterwards students will do a gallery walk and give [feedback](#).

The main goal of this activity is to challenge and demonstrate to students the power of exploring the CODAP tool and its included features.

Students should be able to articulate the following in their journal entries: Data sorting is a process that involves arranging data into a meaningful order to make it easier to understand, analyze or visualize.

See [Assessment Strategies](#) below.

3. **Explore and Journal:** Have students explore the following site - [Other Data Visualizations](#), and then respond in their journals:

1. **Brainstorm an area, field or industry that you would be most interested in creating a purposeful data visualization.**
2. **Write a one line description of what the outcome of the data visualization would look like?**
  - **For example like the website does – Cinema: Explaining a movie plot through data visualization**
3. **Use a bulleted list or sketch a design to describe how you imagine the visualization would appear.**

You could use this time to assess students' work from step 3.

## Day 2 Outline

4. **Box Plot Warm-Up:** Before students come into the classroom, prepare a "number line" by marking 4, 5, and 6 feet equal distance from each other on the wall (or shorter for younger students). When students come into the room, have them place themselves on the number line based on their height. Using masking tape on the ground between them, create a box plot around your students:

- Tape a straight line between the shortest and the bottom of the fourth quartile
- Tape a box around the IQR
- Tape a straight line at the median
- Tape a straight line between the top of the IQR and the max

Alternatively, if you line students up in front of a white board you can sketch the box plot behind them.

Ask students to hypothesize what the meaning of the masking tape is, and lead them to the basic elements of a box plot (quartiles, median, max, min, etc).

Ask students some basic box plot questions, for example: "if there were one or two very tall people and everyone else were short, what would the boxplot look like?"

Formative Assessment Notes

If you have multiple classes, you may consider leaving the previous classes box plot in order for students to compare.

5. Have each student use the *Student Activity Guide - Making Basic Visualizations* (see [below](#), view [Google Doc](#) or [make a copy](#)) to complete the assignment.

*Summary:* Students will use the student activity guide to access CODAP templates to independently do the following:

- Create basic visualizations
- Store their CODAP links
- Answer questions about each data set
- Identify opportunities to explore deeper
- Reflect on the data cycle

## Day 3 Outline

1. **Mini-Project: Combining Visualizations:** Put students in pairs or small groups, and have them use the [Mini Project Directions & Student Planner](#) to complete the project.

*Summary:* In this mini-project, students will perform an exploratory data analysis to uncover relationships between worldwide internet speeds and prices. Students will then communicate their findings using multiple visualizations in a short presentation. They will conclude their project with a driving question that arose from their results and identify any possible solutions or recommendations to address their issue.

2. **Mini Project Reflection:** Have students complete the [Reflection](#) piece of the Mini Project

Consider collecting student activity guides as a quick completion assignment.

Use this opportunity to scan student answers to drive further instruction.

### Formative Assessment Notes

See [Assessment Strategies](#) below for details

See [Assessment Strategies](#) below for details

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Using Basic Visualizations Peer Review Forms

Once students have completed the gallery walk in step 2, collect the peer reviews for each student to gauge where students strengths and weaknesses are, see [here](#) for printable copies.

### Mini-Project Combining Visualizations

Have students work with the Mini Project Directions & Student Planner (see [below](#), view [Google Doc](#) or [make a copy](#)) to complete the project. Collect and analyze [student reflections](#) for genuinity. Use the rubric below to assess student presentations.

	<i>Proficiency</i>	<i>Yes</i>	<i>No</i>	<i>Notes</i>
<b>Visualizations</b>	Presentation includes <b>at least two relevant and accurate visualizations</b> using the manipulatives in the room.			
<b>Description of Visualization</b>	Student(s) describe each visualization using vocabulary from the course and include axis labels, legends, etc.			
<b>Concern/Solution</b>	Student identifies and describes <b>at least one outcome as a problem</b> or area of concern and <b>identifies possible solutions</b> or recommendations to address the issue			
<b>Rising Questions</b>	Rising question is <b>mostly relative</b> to the outcome of the visualization			

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

Consider spending a whole day on the stations for students who may need more time to plot and discover relationships.

Consider making the stations digital using interactive slides like google slides or desmos for students to complete virtually.

Consider pairing students with high and low tech skills to complete the [Student Activity Guide - Creating Visualizations](#).

Consider having students create the visualizations in a stand alone day, and then complete the diving deeper and exploratory analysis questions on a second day two.

Consider completing the stations from the [unplugged version](#) of this lesson.

### Extensions

Have students explore the following site: [Data Visualization Tips For More Effective And Engaging Design](#)

Consider using the Mini Project as a performance task to prepare them for a larger project where students research and use their own data set to complete the project in a similar fashion.




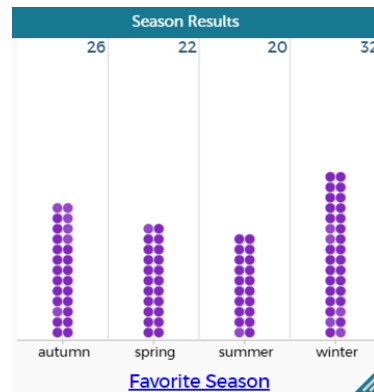
# Teacher Directions - The Use of Basic Visualizations

An activity for a class to explore how a basic visualization can transform into a creative freestyle representation.

Prior to completing this activity students should have completed the [warm up](#) in step 1. All students should be completing the following changes on their own CODAP copy.

## Step 1: Bar Chart

1. Have students create new graph element.
2. Have students create a histogram of their categorical data by dragging their chosen attribute the x-axis.
3. Have students click the  **Measure** button and turn on the count for each category.
  - o The outcome will result in a histogram that quickly represents their data.
  - o This allows students to begin to analyze the results of the data.
  - o Example outcome of the season attribute: *here you can see the most favorable seasons in order are winter→autumn→spring→summer*



## Step 2: Collage

1. Have students choose a data representation style and create their collage. Propose the following suggestions:
  - o A google drawing using images from the internet
  - o A paper/pencil one-sheet poster using colored utensils and drawings by hand
  - o A construction paper one-sheet poster and clippings from magazines
  - o A one-sheet paper using glue/tape and resources from outdoors or in the classroom
2. Give students 15 minutes to create their collage.
  - o Example Collage 1 (Seasons Google Drawing, [PDF](#))
  - o Example Collage 2 (Seasons by Hand, [PDF](#))
3. Once complete, have students clear their desks to display only their collage, their bar chart, and a blank [peer feedback form](#) their desk.
4. Use a gallery walk to have students explore their peers' results and give feedback by using a colored utensil to assess their peers on the scales provided in the [peer feedback form](#).

## Printable Peer Reviews

### Using Basic Visualizations (Peer Review)

While assessing your peer's work, draw a ✓ on the scale to indicate how much you agree with the statement.

Peer Review Statements:

1. The collage uses images/drawings that are relevant to the chosen attribute:

Disagree ←————→ Agree

2. The collage aids in interpreting the data set and attribute findings at a glance

Disagree ←————→ Agree

3. The proportions of the collage correlate the bar chart results:

Disagree ←————→ Agree

4. The collage is aesthetically pleasing:

Disagree ←————→ Agree

### Using Basic Visualizations (Peer Review)

While assessing your peer's work, draw a ✓ on the scale to indicate how much you agree with the statement.

Peer Review Statements:

1. The collage uses images/drawings that are relevant to the chosen attribute:

Disagree ←————→ Agree

2. The collage aids in interpreting the data set and attribute findings at a glance

Disagree ←————→ Agree

3. The proportions of the collage correlate the bar chart results:

Disagree ←————→ Agree

4. The collage is aesthetically pleasing:

Disagree ←————→ Agree

# Student Activity Guide - Making Basic Visualizations

A guide to creating basic visualization in the CODAP environment. View [Google Doc](#) or [make a copy](#).

**DIRECTIONS:** Use the CODAP template files to explore, create and analyze multiple data sets. Complete all corresponding sections for each chart type.

Scatter Plots	
Template: <a href="#">Practice - Scatter Plots</a>	Link to your CODAP: <i>use the share button in the top right corner of the file</i>
<b>Exploring Further:</b> List one way you might do more research to investigate a relationship you found.	
<b>Exploring Deeper:</b> Which attributes did you test to create relevant scatterplots?	
Attributes Tested	Outcome

Histograms	
Template: <a href="#">Practice - Histograms</a>	Link to your CODAP: <i>use the share button in the top right corner of the file</i>
<b>Exploring Deeper:</b> Using the US National Parks dataset, which attributes did you test to create relevant histograms?	
Attributes Tested	Outcome

Comparative Box Plots	
Template: <a href="#">Practice - Comparative Box Plots</a>	Link to your CODAP: <i>use the share button in the top right corner of the file</i>
<b>Exploring Deeper:</b> Using the Microdata tool, explore relationships between about 100 randomly selected people. What two conclusions did you discover after analyzing other available attributes?	
<b>Outcome</b>	

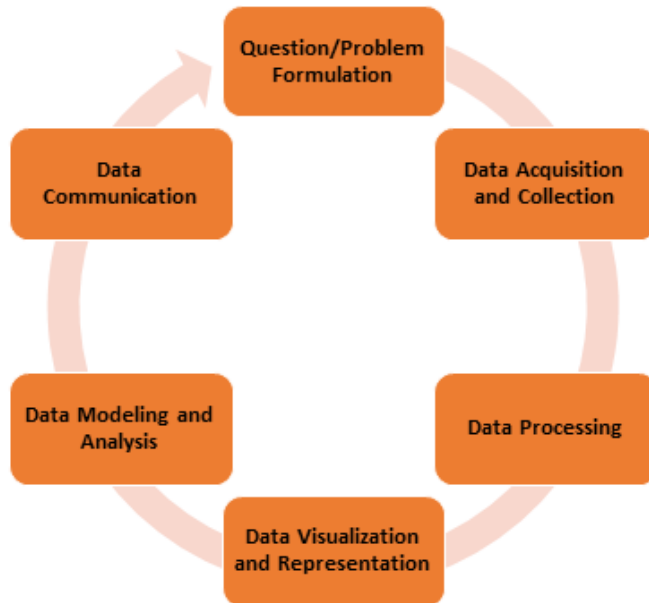
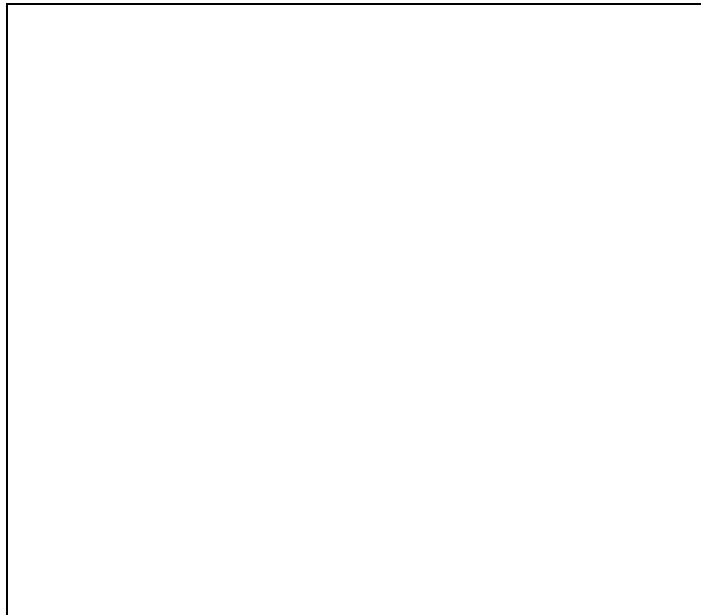
Stacked Bar	
Template: Practice - <a href="#">Stacked Bar</a>	Link to your CODAP: <i>use the share button in the top right corner of the file</i>
<b>Exploring Deeper:</b> Using the Words data set, which attributes did you test to create relevant stacked bar charts?	
<b>Attributes Tested</b>	<b>Outcome</b>
<b>Exploring More:</b> Use the Words data set to answer the following questions.	
Which part of speech shows up most often in this data set?	
Which part of speech seems to be longest, on average?	
Which part of speech seems to be shortest, on average?	
How long is the "typical" noun? How can you tell?	

Geographical	
Templates: <a href="#">Introduction to Geospatial Data</a> & <a href="#">Explore - AirBNB Washington DC</a>	Link to your CODAP(S): <i>use the share button in the top right corner of the file</i>
<b>Recap:</b> In order to use the map feature in CODAP what attributes must be present in the data set?	

Combining Visualizations	
Template: Practice - <a href="#">Combining Visualizations</a>	Link to your CODAP: <i>use the share button in the top right corner of the file</i>
<b>Exploring Deeper:</b> Create three supporting visualizations that show a relationship between newborn babies and their parents. <b>Make a conclusion statement</b> and <b>export each visualization</b> as an image to <b>paste in the table</b> .	
Conclusion	Supporting Visualizations

Word Cloud	
Data set: <a href="#">AirBNB Data - Washington DC</a> Tool: <a href="#">Word Cloud Maker</a>	Link to your Word Cloud: <i>use the share button in the top right corner of the file</i>
<b>Conclusion:</b> What can you conclude people are looking for in DC Air BNBS based on the results of the word cloud?	

**Reflection** - Which steps of the Data Cycle did you see in today's activity? How proficient do you feel with this step of the cycle and what questions do you have pertaining to this step of the cycle?



## Mini Project Directions & Student Planner - Combining

Throughout this mini project you will perform exploratory analysis to uncover relationships between internet users, speeds, and prices in different countries. You will then communicate your findings using multiple visualizations in a short presentation and conclude with a driving question that arose from your results.

**DIRECTIONS:** Use the following [CODAP: Internet Data](#) to access the datasets and explore relationships between the tables and their attributes. Complete the following student planner to produce a presentation that uses at least 2 different visualizations to pose a concluding driving question.

Exploratory Analysis Guide		
Data Sets in CODAP: <i>Worldwide Internet Users, Worldwide Internet Speeds, Worldwide Internet Prices</i>		
Attributes Tested	What patterns did you see?	Sketch of Visualization
Presentation Guide		
Choose Presentation Format (Circle One)	Slides - Poster - Movie - Essay - Song - Poem - Other: _____	
Link to Presentation (if it is digital)	<i>Paste your link to your presentation here once you have begun to build it</i>	
Rising Question(s) Conclusion As a result of your analysis, what question(s) arose to provide the opportunity for further exploration?		
Write a short paragraph Identify and describe at least one outcome as a problem or area or concern. Identify possible solutions or recommendations to address the issue.		
Presentation Checklist	<input type="checkbox"/> Uses 2 different visualizations <input type="checkbox"/> Describe what you thought your visualizations would look like when you chose your attributes from your data set, but before you created the visualizations? <input type="checkbox"/> Explain what patterns, trends, or information the visualizations convey <input type="checkbox"/> Identify and describe a problem or area or concern expressed by the visualizations and identifies possible solutions or recommendations to address the issue <input type="checkbox"/> Pose a new question about the data or the topic based on the visualizations	

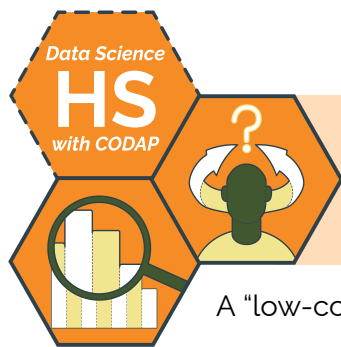


## Mini Project Reflection

---

*Reflect on your experience throughout the completion of the Mini Project.*

1. Describe what went well.
2. Describe what you struggled with.
3. Describe one way you would improve on your project.
4. Explain how you demonstrated mastery of creating basic visualizations. Be sure to cite specific evidence from your project.
5. Describe how you might apply what you learned from this experience to your next project.



# Calculating Basic Statistics

A “low-code” introduction to descriptive statistics by Christa Van Olst & Sara Fergus

## Summary

In this four day lesson, students will analyze datasets by calculating descriptive statistics. They will learn to distinguish between descriptive and inferential statistics, calculate statistics by hand, interpret statements of conclusion for bias, then apply these analysis practices to existing datasets. At the end, students will complete a project where they will find a data set and transform it into a short news article.

*Note: This lesson is similar to the Descriptive Statistics lessons from the [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes a CODAP activity, which the other lessons omit.*

## Objectives

*The students will be able to . . .*

- Calculate descriptive statistics including mean, median, mode, range, and standard deviation.
- Analyze and interpret calculations for tendencies in the data.
- Use statistical calculations to summarize a dataset.

## Standards Alignment

- **DS.1** The student will identify examples of real-world problems to be addressed using data science
- **DS.10** The student will be able to summarize and interpret data represented in visualizations.
- **DS.12** The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.
- **DS.13** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

## Materials

- *Calculating Mean & Median* Worksheet (see [below](#)), print 1 per student/student group
- *New York Times* Data Talk ([Desmos](#))
- [Salaries](#) & [Situations](#) Resources (see below, print according to steps #3 & #5)
- *Interpreting Descriptive Statistics* Data Talk ([Desmos](#))
- *Calculating Descriptive Statistics* worksheet (see [below](#)), print 1 per student group of 2-3
- Youtube Video - [Descriptive vs. Inferential Statistics](#)
- Exploratory Website - [Where does the day go?](#)
- CODAP Files, including *Teacher Directions* ([One Click Statistics](#), [Other Descriptive Statistics](#))
- Online Article - [M&M Color Distribution Article](#)
- *Philosopher's Chair* activity materials (see step #12), including [Teacher Directions](#) & [Statements](#)

# Vocabulary

Term	Definition
Descriptive Statistics	A brief summary using the methods described below to depict any tendencies of a data set
Mean (Average)	The numeric sum, divided by the total amount of values in a set
Median	The middle element in a sorted set of values
Mode	The most frequently repeated element in a set of values.
Standard Deviation	The measure of how far each observed value is from the mean
<a href="#">CODAP Function Library</a>	Documentation of the functions included in the CODAP Environment

## Day 1 Outline

### Formative Assessment Notes

1. **Recalling Mean and Median:** Have students discuss the following:

- *When was the last time you calculated these stats?*
- *Can you think of a real world scenario that uses these stats to describe data?*
- *What limitations are there when calculating by hand?*

Use the [Calculating Mean and Median](#) worksheet to review how students can calculate these statistical measures by hand.

2. **Generating Questions:** Have students individually access this [Desmos: New York Times Data Talk](#). Ask:

- What do you notice? What do you wonder?
- What patterns stand out to you in this data?
- What do you think leads to the patterns in this data?
- What conclusions could we draw using this data?
- Come up with a catchy headline to summarize this data

After some discussion, pose the following question:

*What is a “normal” salary, according to the data?*

Consider having students work pairs and reflect to promote team building.

Consider having students bring some examples of mean & median with them before the lesson as homework.

Respond to student's questions by relating them to “normalcy”, so that most questions have students wondering “What is a normal salary?”

For example, students may ask “Why do older people make more money?” You might respond: “Let’s figure out if they do make more. What is a normal salary for a young person? What is a normal salary for an older person?”

3. **Practice Activity:** Using the [Salaries Resource](#) - cut out and give each pair/student 20 salaries. One list includes Jeff Bezos' income (a significant outlier).

Write two columns on the board, one labeled mean, the other labeled median:

Mean	Median

Have students calculate the mean and median of their salaries (by hand, using a calculator). Once students have calculated, have them write their mean and median on the board in the appropriate column.

4. **Discussion:** Have students analyze the table and discuss how and why one sample's mean is substantially different than its median. Facilitate the discussion using the suggestions below:
1. Have students develop questions about what they see.
  2. Have them share their questions with a peer and then with the class.
  3. Have students theorize, either in groups or as a class, how and why the unusual mean occurred.

Students should come to the conclusion that the median is robust to outliers (though they may not use these words, exactly), while the mean can be deceptive. Discuss the following prompt as a class to reinforce this idea:

*If you were curious about what a "normal" salary is, which would you rather use: mean, or median?*

After the discussion, show the list that includes the outlier so students can see why the statistics were so skewed.

5. **Check for Understanding:** Using the [Situations Resource](#), split the students into groups and provide each group with a cut up list of real-world situations.

Have students classify each situation into a sorted table using the headers "mean", "median", and "mode" to describe which statistic would be most appropriate.

Assess calculations for accuracy as students add them to the chart.

This step should provide students with opportunities to share their ideas and summarize their thinking.

Students should notice an unusual mean - this is not a miscalculation.

Use this as an opportunity to check in with individuals to make sure they understand the vocabulary

6. **Reflection:** In their journals, have students reflect by answering the prompt below:

*If I were to give you a data set of student grades . . .*

1. *What would the median tell you?*
2. *What would the mean tell you?*
3. *What would it mean if the mean and median are not close to each other?*

Have students share their answers with a peer or as a whole group. Be sure to check in with students to make sure they understand how the median is robust to outliers compared to the mean, and what this tells them about datasets.

## Day 2 Outline

### Formative Assessment Notes

7. **Calculating Descriptive Statistics using CODAP:** Use the [Teacher Directions - One Click Statistics using CODAP](#) to demonstrate live how to calculate statistics in the CODAP environment.

*Summary:* This activity will introduce students to the Sampler tool in CODAP. The Sampler allows students to collect data in real time. Students will simulate choosing 30 M&Ms from a bag to quickly calculate the mode color of their selection. Students will then simulate collecting random numeric data to represent screen time hours per day, where they will then quickly analyze the set using mean, median, and a boxplot.

8. Use the [Teacher Directions - Other Descriptive Statistics using CODAP](#) to demonstrate calculating other statistics in CODAP.
9. **Exit Ticket:** Use the [Data Talk: Interpreting Descriptive Statistics](#) to facilitate a closing data talk using the same techniques as step 2.

Consider beginning with a blank file during the demo. Have students do the same to take notes alongside.

## Day 3 Outline

### Formative Assessment Notes

10. **Warm Up: Is this Data Science?** Have students read the "[Where does the day go?](#)" website with the intention of discussing the following questions as a class:

- *What is the collected data?*
- *What is the impact of changing the visualization simultaneously throughout the page?*
- *What are the descriptive statistics here?*
- *Is this data science?*

The data collected is the students input to questions; these are numerical values of how long it takes to complete tasks.

Assess students' answers while floating.

11. **Descriptive vs. Inferential Statistics:** Have students watch this video: [Descriptive vs. Inferential Statistics](#).

In their journals, have students describe a specific social situation (if there were no limitations) in which they would be interested in collecting and calculating descriptive statistics.

Facilitate a brief discussion where students share their journal responses.

12. **Assessing Statistical Statements:** Use the [Teacher Directions - Philosophers Chair Activity Guide](#) to facilitate this activity, where students are given [statements](#) that include descriptive statistics. Students will evaluate the statements, deciding whether or not they are misleading. Then, they will suggest questions and data that should be collected to improve the statement.

13. **Calculating Descriptive Statistics in Data Sets:** Give students the following data sets (or data sets of your choice):

- [College Salaries-by-college-type](#)
- [College Salaries-by-region](#)
- [College Salaries-by-degree](#)
- [Starbucks Drinks Data](#)
- [Starbucks Food Data](#)

Students should explore their data set, using this [checklist](#) to assess their ability to use the statistical skills they learned.

Skim student worksheets for accuracy throughout the activity.

14. **Exit Ticket:** See [Assessment Strategies](#) for details

## Day 4 Outline

Formative Assessment Notes

15. **Project Data Set → News Article:** Have students use the [Student Guide - Project News Article](#) to complete the project.

*Summary:* In this activity, students choose their own dataset to create visualizations and calculate descriptive statistics. Students will then use their findings as artifacts to aid in writing a short news article using the *Newspaper Template* ([view](#) or [make a copy](#)) or creating their own.

See [Assessment Strategies](#) below for Teacher Directions.



# Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

## Day 2 Exit Ticket

Have students complete a google form of the questions below or simply print the [printable copies](#):

Name: \_\_\_\_\_

Date: \_\_\_\_\_

1. Brainstorm and explain what limitations would exist in society if all descriptive statistics were calculated by hand.

**Possible Answer:** *If all descriptive statistics were calculated by hand society would be quite limited. Think about school/education, if our grades were calculated by hand it would take teachers much more time to calculate our overall grades and would probably also limit the amount of assignments teachers grade.*

**Possible Answer:** *In sports, if all calculations were calculated by hand then players and teams would have less insight into their skills, because it would be hard to store and calculate these stats quickly. Playoffs and other tournaments may also be impacted by human error or difficult to calculate in a pinch.*

**Possible Answer:** *In production factories, if statistics like standard deviation were calculated by hand they may be less accurate which could cost a company money due to the inefficiency and inconsistency of products/machines.*

Describe how the inferential statistics applied in the following scenarios could be misleading. What other questions should be asked of the sample?

2. Inference: 64% of the US population owns a winter coat. Sample Size: 1,000 people.

**Possible Answer:** *This implies the statement 64% of the US population owns a winter coat when in reality we don't know where this small sample was polled. Were they spread across the us? Were they in the same state/location? Was the survey online or paper? What time of year was the data collected? etc.*

3. Inference: The average American throws away 4.9 pounds of trash daily. Sample Size: 2,500 high school students.

**Possible Answer:** *High school students may not be the most accurate when predicting pounds of trash accumulated per day. A highschool tends to produce an excess of trash due to students packing lunches/snacks and the amount of students in a building at the same time.*

4. Inference: 7% of Americans believe the moon landing was faked. Sample Size: 100 Americans

**Possible Answer:** *100 is a small sample. What were the ages of the surveyed participants? What state are they from? Where do these people find sources for this claim? What news channel do they watch? What political party are they? Would these 100 people identify as conspiracy theorists?*

## Project - News Article

In this project students will start the data cycle from the beginning, where they will summarize a dataset using visualization(s) and descriptive statistics to create an old school news article. Consider having students work in pairs or individually.

### Student will be required to:

- ☐ Pose a Question/Problem
- ☐ Collect/Find Data
- ☐ Process/Store their Data
- ☐ Visualize Data
- ☐ Calculate Statistics
- ☐ Communicate Outcomes

**Before the Project:** Have students annotate the statements below by using the guiding questions:

- What other information would be insightful?
- What are the similarities/differences in the statements?
- Which one is the best?

The mean of exam two is 77.7. The median is 75, and the mode is 79. Exam two had a standard deviation of 11.6.	Overall the company had another excellent year. We shipped 14.3 tons of fertilizer for the year, and averaged 1.7 tons of fertilizer during the summer months. This is an increase over last year, where we shipped only 13.1 tons of fertilizer, and averaged only 1.4 tons during the summer months. (Standard deviations were as follows: this summer .3 tons, last summer .4 tons).
Group A (87.5) scored higher than group B (77.9) while both had similar standard deviations (8.3 and 7.9 respectively).	After sampling 53 classmates we found that the average student's family has been within the same 10 mile radius for over 100 years. Of those 53 students 23% do not have any siblings.

### During the Project:

1. Have students read this article: [Statistics and Visuals](#)
  - Discuss as a class how descriptive Statistics is the least amount of information that one needs to paint a picture of the distribution of your data, the amount of additional information lies solely on you
  - You don't have to include irrelevant information in your article
  - Your main focus should be on the statistics that will help your reader understand your argument and not ones that are going to mislead them

2. Have students brainstorm a hypothesis/question/problem
  - To streamline the project consider pulling a few datasets and competition prompts from [Kaggle Competitions](#)
3. Have students collect their own data using techniques from the course or choosing a preexisting one
4. Have students create at least two visualizations
5. Have students calculate descriptive statistics by answering the following:
  - Describe the size of your sample
  - Describe the center of your data
  - Describe the spread of your data
  - Assess the shape and spread of your data distribution
  - Compare data from different attributes
6. Students will then use their findings as artifacts to aid in writing a short news article using this *Newspaper Template* ([view](#) or [make a copy](#)) or creating their own

### After the Project: Rubric Project News Article

	<i>Proficiency</i>	<i>Yes</i>	<i>No</i>	<i>Notes</i>
<b>Dataset</b>	Students' chosen dataset depicts students' interest and the <b>data attributes present the opportunity for data analysis</b> using descriptive statistics AND the student used at <b>least one function to create a new attribute</b> .			
<b>Calculations</b>	Students calculate the following: sample size, mean, median, mode, standard deviation. Student <b>calculations are accurate AND used in the students' news article</b> .			
<b>Writing</b>	Students' summary uses effective communication skills by writing their descriptive statistical <b>findings in context of the data attributes</b> AND student identifies any areas needing more <b>research or any questions that could arise</b> .			
<b>Visualization(s)</b>	Students' choice of <b>visualizations are appropriate</b> for the data attributes and provide insight.			

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

Move through the CODAP tutorials one step at a time, all together, instead of allowing students to go through it on their own. Alternatively, live code the tutorial and have students follow along and ask questions.

When exploring the descriptive statistics for the salary, give a relatively confident student the set with Jeff Bezos's salary, since they will need to be confident in their vastly different result.

You may choose to have only some students read the articles, or give articles to some students ahead of time

### Extensions

Have students create deliverable visualization for the [Checklist - Calculating Descriptive Statistics in Data Sets](#) using the guidelines from lesson 7 & 8.

Explore the [unplugged version](#) of this Lesson

Add a fourth day to include a functions only lesson where you can complete the parts of the lesson that include the CODAP function library. You could then complete the project on Day 5.

Consider having student practice and explore more using these CODAP files:

- CODAP: Tallest Buildings ([.codap](#))
- CODAP: Youtube Videos ([.codap](#))

Have students dig deeper into the CODAP function documentation. Have students explore, learn, and show a partner how to use another function from the CODAP library.

## Predicting and Calculating Mean and Median Worksheet

Vocabulary	Definition	How to find
<b>Mean</b>	The numeric sum, divided by the total amount of values in a set, used to show an averaged "center" of a data set	Add up all the numbers, then divide by how many numbers there are.
<b>Median</b>	The middle element in a sorted set of values	Place the numbers in value order and find the middle number.

Set 1: weights of personal transportation devices	
34, 28, 34, 900, 50, 36, 39, 28, 35, 33, 260, 19, 15, 38, 19, 42, 15, 45, 44, 20	
<b>Predict:</b> Which will be larger, mean or median? Why?	
Calculate Mean = _____	
Calculate Median = _____	

Set 2: money spent/earned per day	
17, 30, -42, 26, 25, 24, 27, 30, 34, 37, 24, 24, 0, 19, 23, 13, 39, 34, -100, 24	
<b>Predict:</b> Which will be larger, mean or median? Why?	
Calculate Mean = _____	
Calculate Median = _____	

Set 3: ages of first promotion at work	
17, 20, 32, 38, 38, 38, 15, 27, 27, 40, 36, 28, 29, 46, 36, 39, 14, 21, 30, 35	
<b>Predict:</b> Which will be larger, mean or median? Why?	
Calculate Mean = _____	
Calculate Median = _____	

## Salaries Resource

Cut along dotted lines

72473.42	45492.58	47709.68	38396.51	56461.1	38924.48
56567.53	50365.12	66696.95	50966.82	37605.01	54854.53
61892.41	40268.95	45464.96	45941.33	38827.1	62130.77
36075.44	52598.24	47989.65	52028.3	58630.5	54645.56
30530.18	53392.39	60382.45	57205.42	46541.32	60174.61
53114.76	49740.98	48465.9	61362.43	33842.95	53651.12
20175.14	64335.89	54367.46	42378.18	63042.56	64956.14
50788.98	52337.08	55607.48	42961.2	51920.14	58219.59
64281.35	47294.45	55054.76	50551.77	49606.21	65675.77
31726.21	30083.55	36261.94	66388.89	57549.9	44249.76
42693.46	40190.93	63195.26	49140.95	49078.49	42882.34
42567.3	47658.44	49784.71	37360.48	55476.39	53221.15
26769.04	54676.45	50499.14	44054.46	38507.15	34030.09
70651.29	50113.73	49105.21	68925.86	35998.33	58216.39
57885.24	61691.99	44510.07	56604.94	57181.46	52227.46
42513.66	56469.21	60033.19	37355.49	38037.74	67628.41
51594.32	46461.78	62115.02	58608.68	48066.4	50647.23
41148.3	30587.33	53000.76	51145.07	44604.52	52190.42
45548.06	50873.41	49281	75865.02	45182.34	50875.01
52665.37	40925.21	41658.22	40713.31	46525.34	58270.03
61512.32	47887.99	49069.85	46964.35	53668.17	52803.49
48844.95	58418.47	37813.25	46891.65	71370.82	45968.11
50206.27	49846.72	60317.8	52075.27	31890.28	47461.16
52590.21	36601.71	42793.16	36669.87	41397.16	59108.77
59259.52	33616.8	53741.51	50698.13	60854.14	46131.61
29635.19	33157.27	51509.74	40047.25	53177.45	56536.71
46662.97	66096.91	46860.95	52847.81	50360.7	41282.99
57699.09	53984.83	55792.33	61354.35	39452.92	78500000000.00
58986.88	58017	39393.81	47748.44	42139.33	33867.64
59485.24	26273.94	51878.15	58281.55	44086.95	39988.9
41051.98	37116.63	44437.73	44486.16	38991.69	45404.4
59076.25	43902.45	56323.79	48392.17	69688.13	48370.68
44975.97	38263.84	52770.4	61187.38	44524.45	44494.77
48496.08	42965.81	50414.46	49284.89	44793.8	48292.49
71327.35	48798.73	56928.5	52009.65	42610.84	62513.67
45365.8	53310.00	46490.07	61945.78	43226.71	52972.99
65247.14	50628.11	58940.07	65541.18	54348.76	38437.32
55567.3	37610.39	38559.4	51369.21	46913.28	46270.97
51184.38	51656.67	50211.4	48568.16	47689.78	39791.37
52084.86	58925.68	40899.32	59053.33	39739.65	61051.36

## Situations Resource *Cut along the dotted lines.*

Mean	Median	Mode
A real estate agent wants to calculate the _____ price of houses in a particular area so they can inform their clients of what to expect to spend on a house.	An insurance agent wants to calculate the _____ amount spent on healthcare each year by individuals so they can know how much insurance they need to be able to provide.	An insurance analyst wants to calculate the _____ age of the individuals they provide insurance for so the marketing team can pinpoint advertisements to this age group.
A human resource manager wants to calculate the _____ salary of individuals in a certain field so that they can know what type of salary to offer to new employees.	A real estate agent wants to calculate the _____ price of houses in a particular area so they can inform their clients of the "typical" home price.	A real estate agent wants to calculate the _____ number of bedrooms per house so they can inform their clients on the amount of bedrooms to expect to have in houses in a particular area.
A marketer wants to calculate the _____ revenue earned per advertisement so they can understand how much money their company is making on each one minute ad.	A human resource manager wants to calculate the _____ salary of individuals in the entire company so that they can know what type of salary the job offers.	A human resource manager wants to calculate the _____ of different positions in the company so that they can be aware of the most common position at their company.
A school truancy officer wants to calculate the _____ amount of absences for a single student.	A social worker is collecting national monthly incomes to calculate the _____ so that they can depict the US poverty line.	A marketer wants to calculate the _____ type of ad used (TV, radio, digital) so they can know which type of ads their company uses.
An insurance agent wants to calculate the _____ amount spent on healthcare each year by healthy 22 year olds so that they can use this to inform college graduates at a college career fair.	A K-12 school truancy officer wants to calculate the _____ amount of absences from all of the students in a community.	A K-12 school truancy officer wants to calculate the _____ to categorize the grade levels that are impacted most by absences.



## Teacher Directions - One Click Statistics using CODAP

An activity to introduce students to the Sampler tool in CODAP where students will collect data in real time and quickly analyze the median and mean with one click.

1. Use this Template: M&M Sampler ([codap](#))
2. Click the Start button to simulate randomly choosing 30 M&Ms and storing their colors as data points in a table.
  - o Change the slider to the right from medium to fast to speed up the sampler
  - o Rename the value attribute to color.
3. Have students brainstorm: "Can you calculate mean and median?"
  - o Conclude we would need to analyze the data further (break into categories) or we need more data (the average amount of total blues per student) to calculate relevant statistics.
4. Give students the opportunity to conclude that, in addition to the mean and median, calculating the mode could be easily done.
5. Have students make a bar chart using the Color attribute to calculate their color most chosen.
  - o For a possible outcome see figure 1.
6. In the sampler tool, click CLEAR DATA and the - button to clear the spinner. Minimize the bar chart.
7. Choose the ... button and enter the range 1-10.
  - o Discuss with students that to give context to the data these values will represent a random sample of users daily phone screen-time measured in hours.
8. Change the sample size to select 50 items and then run 2 samples.
  - o Rename the attribute Daily Screen Time (in hours)
9. Have students calculate the one-click statistics mean and median by using the ruler button on the toolbar. Be sure to change the column attributes so that the type is numeric.
10. Have students make a bar chart using the Daily Screen Time attribute to display their data.
  - o In a text box element have students interpret the meaning of the mean and median in context of the data.
    - For example, a mean of 5.35 and a median of 5.4 shows the data was relatively symmetrical.
  - o Consider having the students add more samples by clicking the start button again.
11. In the toolbar have students check the box plot function.
12. In a text box element have students answer the following questions:
  - o What percent of people spent under 3 hours on a device?
  - o What percent of people spent under 5 hours on a device?
  - o What percent of people spent under 8 hours on a device?
  - o What percent of the people spent between 3 and 8 hours on a device?
  - o For a possible outcome see figure 2.
13. Consider showing the [M&M Color Distribution Research](#) as a wrap up.

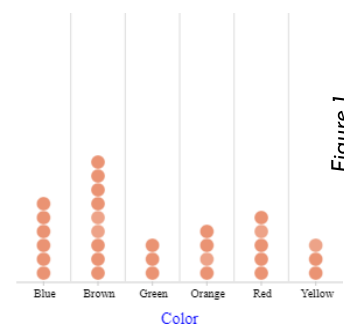


Figure 1

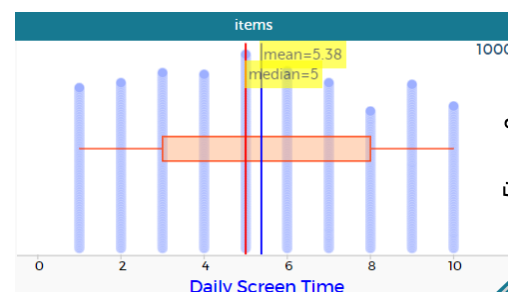


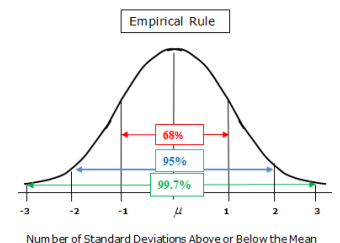
Figure 2



# Teacher Directions - Other Descriptive Statistics in CODAP

An activity to demonstrate how to calculate standard deviation and how to use the CODAP function library.

- Use this *Template: Other Statistics in CODAP* ([.codap](#)). Start by dragging the Legendary attribute to the first column of the data table (to the left of index), the words **drop attribute to create a new collection** will appear with a yellow background.
  - The result should be a collection of tables that are linked together and sorted based on a desired attribute.
- Have students answer the following questions: 1) What's the sample size of our data set? [800 total Pokemon] 2) What is the sample size of just the pokemon who are Legendary? [65 cases]
- Use the `count()` function to show how we can store the count of the values beside their attributes.
  - In the Legendaries table, Add an attribute named Count
  - Once named, click on the Count attribute and choose the edit formula option
  - Type the command `count()` and press enter - This will display the count of each item from the large data set
- Drag the Attack 1 attribute between the two tiers to create a three tiered table. [\[Example here\]](#)
- Have students use the skill from step 3 to answer the following:
  - What is the sample size of just the pokemon who are not Legendary and use a Fairy Attack 1? [16 cases]
  - What is the sample size of just the pokemon who are Legendary and use a Grass Attack 1? [16 cases]
- Create a new graph element and calculate the one click statistics on the Hit Points attribute.
  - Discuss these results in context of the data, what do they mean? [\[Example Graph\]](#)
- In the toolbar choose mean and standard deviation.
  - In a text box element, give students the definition of standard deviation, have students complete this along with you in their own CODAP files.
  - Describe how *The Empirical Rule* applies to normally distributed data:
    - Approximately 68% of the data will fall within 1 standard deviation of the mean.
    - Approximately 95% will fall within 2 standard deviations of the mean.
    - Approximately 99.7% will fall within 3 standard deviations of the mean.
  - Suppose the Pokemon data was normally distributed, use the standard deviation estimate to answer by hand, using the empirical rule: How many of the pokemon have speeds between 39.2 and 97.4? [544 Pokemon]
- Use the `max()` function to show how we can store the fastest speed of each Attack 1 type.
  - In the Attack 1 tier of the table create a new attribute called MaxSpeed
  - Once named, click on the MaxSpeed attribute and choose the edit formula option
  - Type the command `max()` and then press insert value button to choose the Speed attribute
  - This will display the fastest speed of each Attack 1 type
  - What is the fastest psychic attack Pokemon? [Deoxys Speed Forme]
- Have students use the [CODAP Function Library](#) to explore and test one new function on their own or with a partner. Consider posing the students with challenges or showing them the following examples:
  - What if we added an attribute to format each pokemon's name and speed for easy copy and paste output?
    - Add an attribute to the third tier of the table named Formatted Summary and edit the formula to be `concat(Name, ": ", Speed, " mph")` note the spaces inside of the quotations for formatting.
  - What if I knew nothing about Pokemon? So instead of rating the Defense attribute numerically I wanted to summarize it so that any defense over 100 would be labeled good and anything below is considered bad.
    - Add an attribute to the third tier of the table named Defense Rating and edit the formula to be



`if(`Defense Points`>100,"Good", "Bad")` note the commas between each part

## Teacher Directions - Philosophers Chair Activity Guide

Statements	Possible Questions/Thoughts/Outcomes
In 2007, Colgate claimed "More than 80% of Dentists recommend Colgate." which was based on surveys of dentists and hygienists that allowed the participants to select one or more toothpaste brands.	<i>This is misleading since it was a multiple select survey, dentists could have also recommended other brands before colgate.</i>
In 2021, a school summarized that 63% of students who are late to school have jobs.	<i>Are the students working during the week?</i>
In 1973, UC Berkeley's graduate school admitted 44% of male applicants and 35 % of the female applicants and was sued for discrimination.	<i>Consider sorting the data into subgroups and analyzing each department that students applied to then calculate these averages.</i>
A software company is working on creating two different interfaces for their app. Using simple A or B surveying, they reported that 60% of survey respondents prefer Version A over Version B.	<i>At the least collecting attributes of sampled respondents should be considered: Who was surveyed? When were they surveyed? Where were they surveyed? How was the survey conducted?</i>
The average depth of the Potomac River is 10'3 feet.	<i>This is misleading because some parts of the River get up to 33'5 feet deep. This could lead to a dangerous assumption.</i>
The average number of feet for a U.S. Senator is 1.98.	<i>At first glance this is an interesting thought but true due to the United States senator, Ladda Tammy Duckworth, from Illinois who is a retired Army National Guard lieutenant colonel.</i>
The average temperature in Virginia is 39.8°F per year so it is not a vacationing state.	<i>Many other factors to consider [seasons, location, time of day, etc]</i>
In the middle ages the average life span was 40 years, so most people probably lived to see their hair turn white.	<i>It should be considered that many children did not survive as babies back then due to lack of access to medical assistance, the infant mortality rate was incredibly high.</i>

1. Present the following statements one at a time for the class to consider (see [below](#) for printable version, the table above is a key)
2. Have each student decide a position they'll take on the statement and why. Ask:
  - *Is the statement accurate? Misleading?*
  - *Is there enough information? If not, what other data should be considered?*
  - *What questions could be formulated? How could the statement be interpreted?*
3. Have students spend 1 minute writing their ideas about the statement on their white boards and pose questions they may have about the statement. Then have students turn to a partner to discuss

their ideas and positions for about 2-3 minutes.

4. In their journals, after each statement or at the end of the activity, have students write a reflection:
  - *A comment/perspective that challenged their thinking*
  - *Whether or not their mind was changed at any point*
  - *How open-minded they were at the start and end of the conversation*
5. Extension: Have students read this article: [Simpson's Paradox using US Presidential Elections](#)
6. Conclusion: As a class have students reflect and discuss the following quote

*"When researching and collecting data, we must decide whether to break the data into separate distributions, or to keep the data combined. The correct decision is entirely situational and this is part of the reason why data science exists at the intersection of mathematics/statistics, computer science and business/domain knowledge: We need to know our data, and more importantly, what we want out of our data, in order to choose which approach to take. We need to know what we are looking for, and to choose the best data-viewpoint giving a fair representation of the truth."*

- "Tom Grigg" ([The challenge of finding the right view through data](#))

## Worksheet - Philosophers Chair Statements

**DIRECTIONS:** Decide a position you will take each statement and why. Consider the following thoughts::

- Is the statement accurate? Misleading?
- Is there enough information? If not, what other data should be considered?
- What questions could be formulated? How could the statement be interpreted?

Statements	Possible Questions/Thoughts/Outcomes
In 2007, Colgate claimed "More than 80% of Dentists recommend Colgate." which was based on surveys of dentists and hygienists that allowed the participants to select one or more toothpaste brands.	
In 2021, a school summarized that 63% of students who are late to school have jobs.	
In 1973, UC Berkeley's graduate school admitted 44% of male applicants and 35 % of the female applicants and was sued for discrimination.	
A software company is working on creating two different interfaces for their app. Using simple A or B surveying, they reported that 60% of survey respondents prefer Version A over Version B.	
The average depth of the Potomac River is 10.3 feet.	
The average number of feet for a U.S. Senator is 1.98.	
Virginia is not a state that a lot of people vacation in, because the average temperature is 39.8 degrees	
In the middle ages the average life span was 40 years, so most people probably lived to see their hair turn white.	

# Calculating Descriptive Statistics in Data Sets Checklist

Use this checklist to self assess your skills learned thus far and your ability to start the process from the beginning given only a data set.

- ☐ I can upload this dataset into a new CODAP file
- ☐ I can identify the sample size of this data set
  - Sample Size = \_\_\_\_\_
- ☐ I can identify a subset of the data table by sorting the data table into multiple tiers
  - Identify your subset: \_\_\_\_\_
  - Sample size of the subset = \_\_\_\_\_
- ☐ I can use at least 2 different CODAP functions to calculate and add new attributes to my data set
  - CODAP function used: \_\_\_\_\_
  - CODAP function used: \_\_\_\_\_
  - Other functions: \_\_\_\_\_
- ☐ I can calculate the mean of a data set
  - Mean = \_\_\_\_\_
  - What does this mean represent?  
\_\_\_\_\_
- ☐ I can calculate the median of a data set
  - Median = \_\_\_\_\_
  - What does this median represent?  
\_\_\_\_\_
- ☐ I can calculate the mode of a data set
  - mode = \_\_\_\_\_
  - What does this mode represent? \_\_\_\_\_
- ☐ I can calculate the standard deviation of a data set
  - Standard Deviation = \_\_\_\_\_
  - How many cases are within 1 standard deviation of the mean? \_\_\_\_\_
- ☐ I can create at least three different visualizations with my data.
- ☐ I can summarize my findings in context of the data by writing in a text box element
- ☐ I can identify outliers and skews in the data by writing my thoughts in a text box element
- ☐ I can save and store my CODAP file in an organized manner

## Student Guide - Project News Article

In this project you will start the data cycle from the beginning, where you will summarize a dataset using visualization(s) and descriptive statistics to create an old school news article.

### Project Checklist:

- ☐ Pose a Question/Problem
- ☐ Collect/Find Data
- ☐ Process/Store their Data
- ☐ Visualize Data
- ☐ Calculate Statistics
- ☐ Communicate Outcomes

**Part 1: Reading** - Read the following article [Statistics and Visuals](#)

**Part 2: Brainstorm** - Jot down 3 ideas for a hypothesis/question/problem, and then narrow it down to one that would be the best answered with statistics, and is the most interesting to you.

Idea 1	
Idea 2	
Idea 3	

**Part 3: Collect Data** - Collect data to support your hypothesis/question/problem using techniques from the course. List the questions and the data type of the responses

Question	Data Type

**Part 4: Create Visualizations** - Create at least two visualizations and sketch them below

Visualization 1	Visualization 2

**Part 5: Statistics** - Calculate descriptive statistics by answering the following:

<i>Describe the size of your sample</i>	
<i>Describe the center of your data</i>	
<i>What makes the most sense for your data and why? Mean, Median, Mode, Range</i>	
<i>Describe and assess the shape and spread of your data distribution.</i>	
<i>Compare the descriptive statistics from different attributes.</i>	

**Part 6: Communicate Outcomes** - Use your findings as artifacts to aid in writing a short news article using this [Newspaper Template](#) or creating your own.

## Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

1. Brainstorm and explain what limitations would exist in society if all descriptive statistics were calculated by hand.

Describe how the inferential statistics applied in the following scenarios could be misleading. What other questions should be asked of the sample?

2. Inference: 64% of the US population owns a winter coat. Sample Size: 1,000 people.
3. Inference: The average American throws away 4.9 pounds of trash daily. Sample Size: 2,500 high school students.
4. Inference: 7% of Americans believe the moon landing was faked. Sample Size: 100 Americans

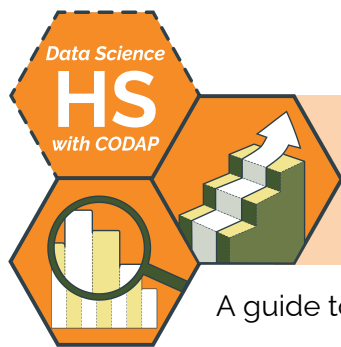
Name: \_\_\_\_\_

1. Brainstorm and explain what limitations would exist in society if all descriptive statistics were calculated by hand.

Describe how the inferential statistics applied in the following scenarios could be misleading. What other questions should be asked of the sample?

2. Inference: 64% of the US population owns a winter coat. Sample Size: 1,000 people.
3. Inference: The average American throws away 4.9 pounds of trash daily. Sample Size: 2,500 high school students.
4. Inference: 7% of Americans believe the moon landing was faked. Sample Size: 100 Americans





# Creating Regression Models

A guide to creating linear regression models in CODAP by Christa VanOlst and Sara Fergus

## Summary

In this lesson, students will start by categorizing scatter plot relationships and exploring correlation coefficients to assess models. On day two, students will use CODAP to create predictive models and decide whether those models are appropriate for the data given. In conclusion, after demonstrating and exploring with students using CODAP, students will create their own linear models in a mini project.

*Note: This lesson is similar to Creating Models lessons from the [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes a CODAP activity, which the others omit.*

## Objectives

- Students will Identify and classify relationships in scatter plots including: Linear, Polynomial, Logistic, and Clustered
- Students will create basic by-eye models to represent data.
- Students will create linear models in CODAP and plot them over a scatter plot.

## Standards Alignment

- **D.S. 9:** The student will select and analyze data models to make predictions
- **DS.11:** The student will select statistical models and use goodness of fit testing to extract actionable knowledge directly from data.
- **DS.12:** The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.
- **DS.13:** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

## Materials

- Warm Up: [Optimism Regression](#) & Plot Cards ([PDF](#))
- Correlation Investigation ([Desmos](#)) & *Correlation vs Regression* guide (see [below](#))
- *Years Experience vs. Salary* ([.png](#)), *Prisoner Count vs Population* ([.png](#)), *GPA vs Admission* ([.png](#))
- *2019 World Happiness* (view [Google Sheet](#) or [make a copy](#)) & *The Impact of Freedom* ([.png](#))
- *Calculating & Assessing using Residuals* worksheet (see [below](#))
- Calculating and Plotting Residuals w/ card sort ([Desmos](#)) & Residuals Review ([Quizizz](#))
- *Regression Guide* worksheet (see [below](#)) & *Student Guide Regression Mini-Project* (see [below](#))
- CODAP: Linear Regression ([.codap](#))
- Data Sets: Fuel Consumption since 1990 ([CSV](#)), Position & Salaries ([CSV](#)), Fish ([CSV](#))

## Vocabulary

Term	Definition
Model	A model is a framework for making predictions or describing situations using mathematical equations (i.e. regression) or other algorithms (i.e. decision tree).
Linear Regression	Linear regression is a statistical technique used to approximate a linear relationship of two variables to make predictions and describe situations.
Line of Best Fit	A line of best fit is a straight line through a scatter plot that represents the linear regression.
Polynomial Regression	Polynomial regression approximates nonlinear polynomial relationships (quadratic, cubic, etc) of two variables to make predictions and describe situations
Clustering	Clusters are a type of relationship that is discrete (unlike regression relationships). Data falls into specific "clusters". This pattern can be used to make predictions and describe situations.
Residual	The difference between an observed value and the associated model predicted value.
Residual Plot	A residual plot graphs the residuals of a line of a best fit on the vertical axis and the independent variable on the horizontal axis. Residual plots can be used to determine whether a linear model is appropriate for the data.

## Before the Lesson

This lesson requires a good amount of logistical forethought and planning. For **Day 1**, print out the following materials and be ready to distribute them to students:

- The [Plot Cards](#), one per small group of students. You can trim them yourself ahead of time, or have students cut them out themselves during the activity (see step #2)
- The *Correlation vs Regression* student guide (see [below](#)), one per student.

On **Day 2**, be sure to print the *Calculating & Assessing Residuals* guide (see [below](#)), one per student.

On **Day 3**, be sure to print the *Regression* guide (see [below](#)), one per student.

## Day 1 Outline

### Formative Assessment Notes

1. **Analysis & Discussion:** Show students [Optimism Regression Visualization](#), which shows the percent of optimistic younger people compared to the percent of optimistic older people by country.

Ask students to write in their journals what they notice and what they wonder. Facilitate a short discussion where students share some of their reflections.

*Optional:* Ask students– Are you surprised by where the United States is? Does the trend work well for the United States, or is it an outlier? "Do you fit in the trend, or would you be an outlier?" to connect their analysis to prior lessons.

2. **Sorting Relationships:** Split students into groups of 2-4 and give each group the [Plot Cards](#).

Have them sort cards into 3-5 categories. Allow them to sort into whatever categories they come up with. When they are finished, have them write category titles on post-it notes.

Select and sequence a few of the categories. Point out categories that group related correlation patterns together:

- No relationship
- Linear Relationship
- Polynomial Relationship
- Logistic Relationship
- Clustering / Patterns

Consider having students move around the room to view & comment on one another's categories.

3. **Exploration of Correlation Coefficients:** Have students complete [this Desmos: Correlation Investigation](#), which introduces the correlation coefficient.

Use the Desmos pacing feature to restrict students to screens 1-6. When students finish with screen 6, pause for a class-wide discussion on their findings. Highlight student responses that include keywords like "accuracy", "strength", and "predict".

Then, allow students to finish the activity. The remaining slides practice identifying the correlation coefficient.

Have a few students share what they wrote. Make sure to have students who commented on the trend of the data, or who commented on predictions to share.

Students may not have used that language exactly, introduce the language if not. You may also choose to point out categories like negative relationships v. positive relationships or strong relationships v. weak relationships.

Listen for student use of vocabulary, and reinforce/re-teach as appropriate.

4. **Correlation/Regression Coefficients Activity:** Have students use the [Student Activity Guide - Correlation vs. Regression](#) to explore models and answer reflection questions based on interpreting correlations.

*Summary:* Students categorize scenarios by predicting a correlation coefficient. They then compare correlation to regression by predicting outcome using a by eye approach and using a linear regression equation.

5. **Correlation/Causation Data Talk:** Conclude this topic using the [Data Talk: Correlation and Causation](#) to facilitate a data talk using the sequencing tool on Desmos.

*Summary:* In the data talk students will justify their thoughts in a *Which One Doesn't Belong* slide, given a group of models. Then they interpret multiple "off-the-wall" correlation models to support that correlation does not mean causation.

**Extension:** Feel free to use other nonsensical correlations from the following resource: [Spurious Correlations](#)

You may choose to pace the students to debrief after the first and second pages, or have them complete the full worksheet.

When sharing student answers, look for vocabulary like "correlation", "linear", "causation", "model shape" or "relationship". Reinforce/reteach as appropriate.

## Day 2 Outline

6. **Warm-Up:** Print and give students the following resources:

- [2019 World Happiness - Top 20](#)
- Scatter plot - [The Impact of Freedom](#)

Pose the question:

***Is this model appropriate for the data?***

Facilitate a discussion with students, reviewing the concepts from the discussion at the end of the previous day's instruction.

7. **Linear Regression in CODAP:** Using the [CODAP: Linear Regression](#) file pose the following questions to students"
1. How does experience relate to salary?
  2. How do the incarceration rates in each state compare to the population?
  3. Can your GPA impact your chances of admission to Graduate School?

### Formative Assessment Notes

Have students turn in their "by eye" models and labels as a quick check for understanding. Follow up with students who are having trouble.

Have a few students share their notice-and-wonders with the class. Students should be finding that they all have a positive linear relationship.

Give students some time to use exploratory analysis on the data sets to answer the question above by expanding the boxes and dragging in attributes..

Then, as a class, regroup to demonstrate creating linear relationship models by creating the following scatter plots:

- [Years Experience vs. Salary](#)
- [Prisoner Count vs. Population](#)
- [College GPA vs. Chance of Admission](#)

Have students discuss in small groups:

- What do you notice?
- What do you wonder?
- What correlations do you see? Are they strong or weak?
- What questions arise?

8. **Pose the question:** *"Is this model appropriate for the data?"*

Demonstrate calculating the residuals by using one of the techniques below:

Using the CODAP Function Library:

- In the existing table, next to Predicted Salary, defined add a new attribute named 'Residual Salary'
- Edit the 'Residual Salary' to have the following function:
  - `linRegrResidual (YearsExperience, Salary)`
- This should calculate the residuals for each case.

**Optional:** Using a user defined function:

- In the existing table, next to Predicted Salary, add a new attribute named 'Residual Salary'
- Edit the 'Residual Salary' to have the following formula:
  - `'Salary' - 'Predicted Salary'`
- This should calculate the residuals for each case.

As a class, analyze the residual plot:

- Create a new graph element using the Salary on the x-axis and the Residual Salary on the y-axis to create a scatter plot.
- Residuals close to zero are desirable as they indicate smaller differences between the observed and predicted value.
- Residuals can also show patterns in the data.

9. **Residual Practice:** Give students the [Worksheet - Calculating & Assessing using Residuals](#) resource to calculate the residuals and assess the model using the data from the warm-up.

**They should be able to use vocabulary to describe this relationship by now—be sure to model & review if needed.**

**The slight difference in the numerical results of the residual techniques are due to rounding in the linear regression. Consider showing students both to explore this topic further.**

**Complete the optional activity to explore this topic further with students.**

10. **More Practice:** Complete the [Desmos: Calculating and Plotting Residuals \(with Card sort\)](#) and/or the [Residuals Review Quizizz](#) activities with students.
11. **Curve Sketching:** Have students go back to the [Plot Cards](#) from the sort and, for each plot, sketch a curve that describes the relationship. The curve should be as simple as possible, as accurate as possible, and continuous. Find and share one example of:
- A good line of best fit
  - A simple polynomial regression
  - A logistic curve / close to a logistic curve.
  - An unusual/creative one for slide 26 or 33.

Ask students to discuss the following prompt:

***"How would you figure out if your sketch was a good model compared to someone else's?"***

12. **Exit Ticket:** See [Assessment Strategies](#) below

Have students turn in their "by eye" models and labels as a quick check for understanding. Follow up with students who are having trouble.

Students should be able to identify calculating residuals as a good way of assessing a model for accuracy.

Collect exit tickets. Make sure that students have a good understanding of the meaning of models

## Day 3 Outline

13. **Warm-Up:** Display the graphs on [page 2 of this data talk](#), which both show goals made by women's soccer players. The first is a heat map of the field, the second shows the likelihood of scoring based on distance from goal.

Have students write what they notice about the graphs and what they wonder about them in their journals. While they are writing, ask a few students to share specific thoughts.

Either using a student's thought or after students have shared, ask students to make predictions using the graph on the right. You may have them write their prediction in their journal before sharing.

### Formative Assessment Notes

When asking students to share their thoughts, highlight answers that mention prediction or important changes/ events that lead to change.

14. **Polynomial Regression in CODAP:** Give students the following data sets and have them explore the scatter plots:

1. [Fuel Consumption since 1990 Data Set](#) (Quadratic)
2. [Position and Salaries Data](#) (Exponential)
3. [Fish Data Set](#) (Quadratic)

*\*\*Note CODAP only has a built-in linear regression tool. If desired other types of regression can be calculated in Python or Desmos, and then CODAP can be used for prediction\*\**

Give students the following mathematical models for each set:

1. Quadratic:  $\text{FuelConsumption} = -0.11(\text{year})^2 + 4.62(\text{year}) + 127.6$
2. Exponential:  $\text{salary} = 23695(1.4)^{\text{level}}$
3. Quadratic:  $\text{weight} = 0.62(\text{diag\_length})^2 - 50.4(\text{diag\_length}) + 1240.26$

Demonstrate by using the plotted function tool to demonstrate plotting these models over their scatter plots.

1. [Plotted Quadratic Function 1](#) & [Example Graph 1](#)
2. [Plotted Exponential Function 2](#) & [Example Graph 2](#)
3. [Plotted Quadratic Function 3](#) & [Example Graph 3](#)

15. Have students create their own regression models.

In pairs or groups, students will explore relationships between other attributes in the data sets already distributed throughout this lesson.

Have students use the [Student Guide - Regression](#).

16. **Exit Ticket:** See [Assessment Strategies](#) below.

Through the exploratory analysis, listen for vocab like “curve”, “nonlinear”, etc.

Students should conclude nonlinear relationships between most tested attributes.

Discuss with students that CODAP will not calculate these equations for you, but can be helpful when wanting to plot a function on top of their data.

Consider having students share their findings in a discussion post on your school platform (google classroom, canvas, etc.)

Collect exit tickets. Make sure that students have a good understanding of the meaning of models

## Day 4 Outline

### Formative Assessment Notes

17. **Mini Project:** Have students complete the [Regression Mini-Project](#) (see [Assessment Strategies](#) below) to practice creating models.

Check in with students along the way.



## Assessment Strategies

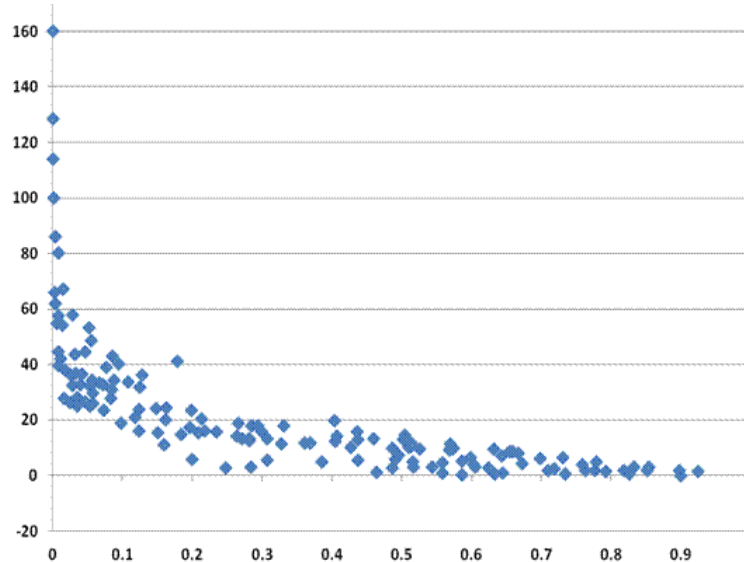
In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Day 2 Exit Ticket (see [below](#) for printable copies)

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Given the following scatter plot, sketch the line/curve of best fit.



Create a scenario using real world context that this model could represent. Your scenario does not have to be realistic but should describe two attributes that would have this type of impact on each other. Write your summary as if your audience has no knowledge of Data Science.



### Day 3 Exit Ticket *(see [here](#) for printable copies)*

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Choose any one model that you saw or created in class today. Write a summary of the meaning of that model. What does the model tell you about the situation being analyzed? Are there any limitations in this model? What are the bias in this model? Write your summary as if your audience has no knowledge of Data Science.

### Rubric for Day 4 Exploratory Analysis

	<i>Exemplary</i>	<i>Proficient</i>	<i>Developing</i>
<b>Questions</b>		All questions 1-8 were addressed and communicated in the summary.	
<b>Data Visual</b>		An appropriate scatter plot is created that has minimal errors, shows a connection to the model, and captures the insights of the patterns or trends.	
<b>Model</b>		A model is accurately created using the appropriate commands and attributes relative to their visualization.	
<b>Summary &amp; Communication</b>		<p>Summary incorporates data science vocabulary within the context of the investigation.</p> <p><b>AND</b></p> <p>Summary communicates student understanding of the data cycle through:</p> <ul style="list-style-type: none"> <li>descriptions of the attributes tested and how their behavior is impacted by each other</li> <li>describing any patterns discovered and any questions that arose.</li> <li>explaining the visualization(s) in detail and how it connects to the model.</li> <li>exploring and communicating the validity and limitations of the model</li> </ul>	

## Regression Mini-Project

Have students create a new notebook. Using either Kaggle or a class set of datasets, instruct students to find some data that they are interested in that has at least 2 numeric attributes.

In their notebook, students should

1. Upload and save the data set
2. Clean the data as necessary
3. Create a scatter plot
4. Determine the relationship
5. Create a linear or logistic regression model
6. Make at least one prediction based on your model

Have students use the [student guide](#) to guide their exploration.

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Concept</b>	The student accurately identifies the relationship between their two numeric variables.			
<b>Representation</b>	The student generates, expresses, and assesses a model to best fit their data.			
<b>Coherence</b>	The student uses their model to make a prediction and expresses the prediction in the context of the data.			

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

Some students may even benefit from creating their own document instead of using the worksheet, and keeping everything in one place.

To avoid having to upload a lot of different datasets, you could have students only upload the incarceration OR admission data for the linear regression and only the titanic OR framingham data for logistic regression. If students do this, make sure they run only their code block, instead of the whole worksheet. This could be helpful with students who may become overwhelmed with a lot of files open.

### Extensions

Consider completing the Day 2 (Building a Physical Model Activity) from the [unplugged version](#) of this lesson to get students up and moving.

**Self-Paced Exploratory Analysis Independent Practice Day:** Give students 5-10 minutes to explore the [CODAP: Recalling Regression \(Insurance Data\)](#) to answer the following questions in a text box element in their file:

- Which region has the highest average charge for insurance?
- Which factors influence the price of health insurance most?
- How do we calculate R if they only have access to  $R^2$ ?

Have students answer the following questions in a text box element in their CODAP file:

- What are the relationships are they testing?
- Which attributes have the highest correlation?
- Which attributes have the impact on the dependent variable?
- Are there any negative correlation relationships?
- Are the correlation(s) you found strong or weak?
- What questions arise?
- What research question arose to provide the opportunity to explore the relationship deeper?

**More Practice:** Give students the following data sets to choose from (or have them choose their own and have them create a new CODAP file:

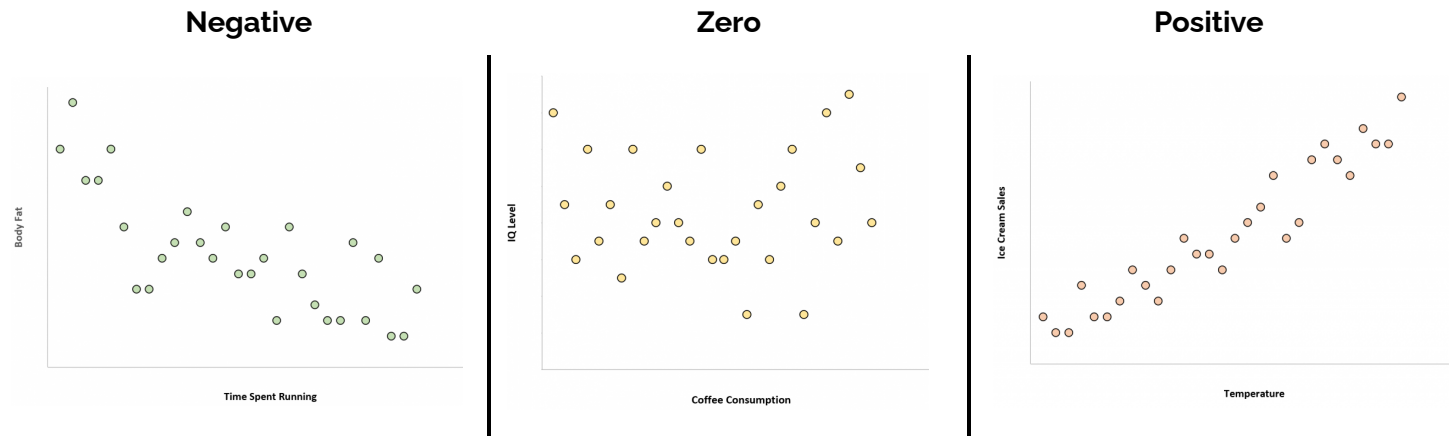
- [5-minute\\_crafts\\_data](#)
- [Hotel\\_data](#)
- [pirates\\_and\\_global\\_warming\\_data](#)

Have students create scatterplots with the data. Have each group run a regression on their scatter plot. Then, have students complete the [Worksheet - Regression Analysis](#) in a small group. Have each group present their summary. Their summary should include references to all of the questions.

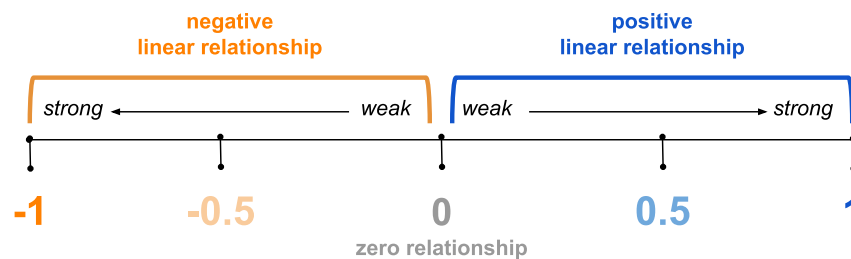
# Student Activity Guide - Correlation vs. Regression

## What is a Correlation?

A correlation measures the relationship between two variables.



Recall the spectrum of Correlation Coefficients



**DIRECTIONS:** Categorize the scenarios below by predicting their correlation coefficient value using a number between -1 and 1 inclusive.

- \_\_\_\_\_1. The height of a person and the salary they earn.
- \_\_\_\_\_2. The shoe size of a person and the number of movies they watched.
- \_\_\_\_\_3. The height and the weight of a person.
- \_\_\_\_\_4. The quote "With Age Comes Wisdom".
- \_\_\_\_\_5. The amount of time you spend in water (swimming/bathing) and the wrinkles in your skin.
- \_\_\_\_\_6. The speed of a wind turbine and the amount of electricity that is generated.
- \_\_\_\_\_7. The amount of moisture in an environment and the growth of mold spores.
- \_\_\_\_\_8. A student's screen time and their grades.
- \_\_\_\_\_9. A person's pizza consumption and their zodiac sign.
- \_\_\_\_\_10. A person's average pulse rate and the calories they are burning.
- \_\_\_\_\_11. The temperature it is outside and the amount of layers of clothing a person wears.
- \_\_\_\_\_12. The size of a herd of animals and the amount of food to go around.

# Correlation vs Regression

When studying the relationship between numeric variables, it is important to know the difference between correlation and regression.

## What is Regression?

Regression is a statistical technique used to approximate a linear relationship of two variables to make predictions.

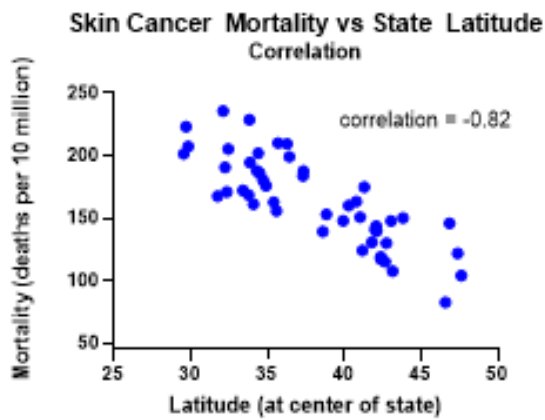


Figure A

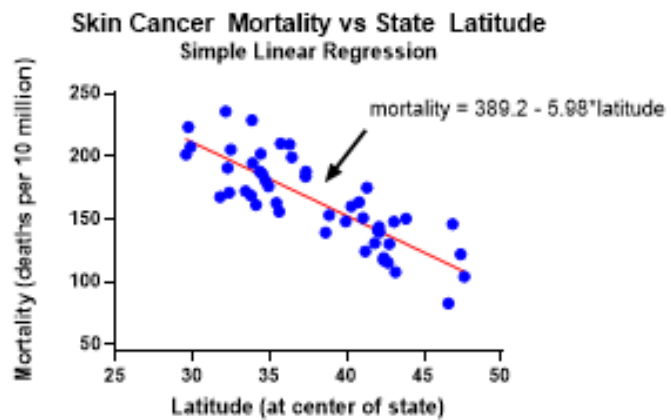


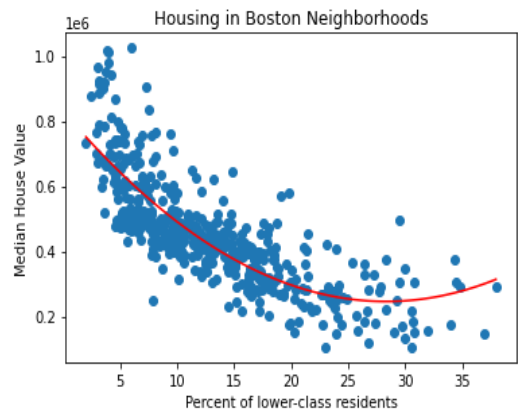
Figure B

Citation: [GraphPad](#)

**DIRECTIONS:** Use the figures above to make predictions by eye and compare them to the calculated regression using the equation.

Latitude	By Eye Prediction	Calculated Prediction $-5.98(\text{latitude}) + 389.2$
45	140	$-5.98(45) + 389.2 = 120.1$
36		
25		
50		

Write a sentence or two to interpret the regression model to the right. What type of regression would you consider this? What questions arise?



# Worksheet - Calculating & Assessing using Residuals

## Vocabulary

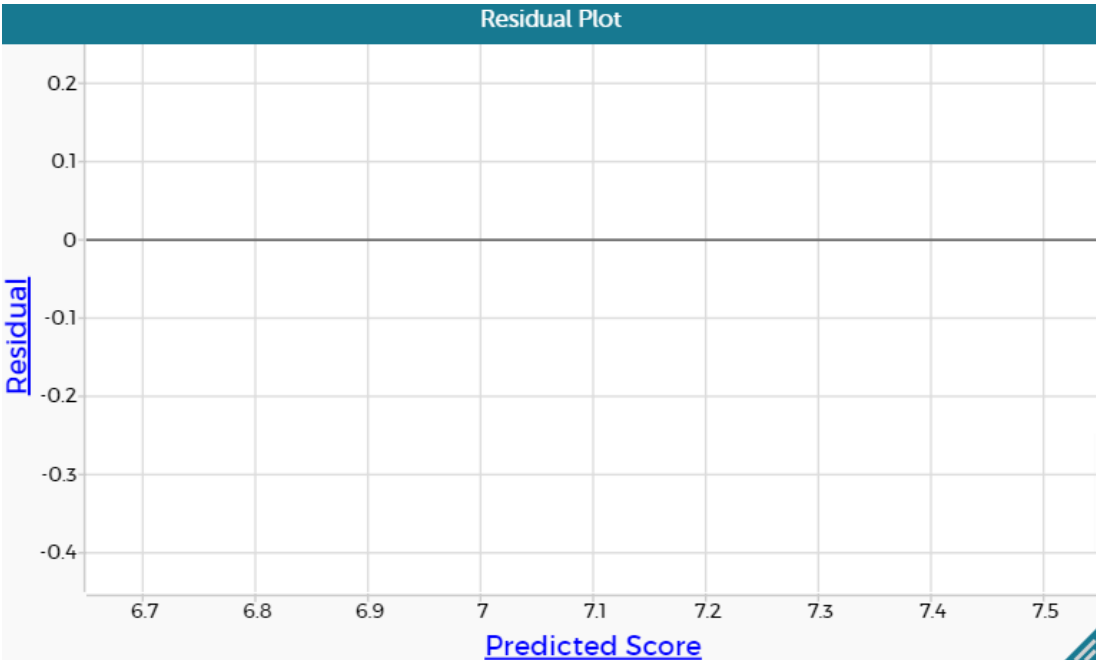
Residual	The difference between the observed value and the model's predicted value.
Residual Plot	A residual plot graphs the residuals of a line of a best fit on the vertical axis and the independent variable on the horizontal axis. Residual plots can be used to determine whether a linear model is appropriate for the data.

**DIRECTIONS:** Calculate the following values to complete the table.

Overall rank	Country	Freedom to make life choices	Score	Predicted Score $3.09 \times \text{Freedom} + 5.6$	Residual $\text{Predicted Score} - \text{Score}$
1	Finland	0.596	7.769		
2	Denmark	0.592	7.6		
3	Norway	0.603	7.554		
4	Iceland	0.591	7.494		
5	Netherlands	0.557	7.488		
6	Switzerland	0.572	7.48		
7	Sweden	0.574	7.343		
8	New Zealand	0.585	7.307		
9	Canada	0.584	7.278		
10	Austria	0.532	7.246		
11	Australia	0.557	7.228		
12	Costa Rica	0.558	7.167		
13	Israel	0.371	7.139		
14	Luxembourg	0.526	7.09		
15	United Kingdom	0.45	7.054		
16	Ireland	0.516	7.021		
17	Germany	0.495	6.985		
18	Belgium	0.473	6.923		
19	United States	0.454	6.892		
20	Czech Republic	0.457	6.852		

**Sketch** a residual plot on the axis to the right and **interpret** the plot using the examples from below.

Is the model a good fit?



Good Residual Plots		Example
<p>Random Distribution - the residuals are approximately distributed in the same manner.</p> <p>In other words, we do not see any patterns in the value of the residuals as we move along the x-axis.</p>		
Bad Residual Plots		
<p><i>Uneven spread</i> - the model does not fit consistently across all x-values.</p>	<p><i>Curved</i> - If there are patterns or curves in the residual plot then a nonlinear model may be more appropriate (quadratic, polynomial, etc.)</p>	<p><i>Outlier</i> - There may be an underlying data recording error. Remove to see what the effect is whether it is influential or not.</p>
If you can detect a clear pattern or trend in your residuals, then your model has room for improvement		

Student Guide - Regression

Data Set 1: \_\_\_\_\_

1. What attributes are you testing?
2. What is the question you are exploring?
3. What is the equation of the Least Squares Line?
4. What is the  $r^2$ -value? What is the  $r$ -value? What does this value mean for your data?
5. Describe these relationships in context of the data.
6. Complete the table below by choosing three values to predict. List the value in the first column, then make a prediction by eye and then use a formula to compare your guess to the calculated prediction.

Making Predictions		
<i>Input Value</i>	<i>By Eye Prediction</i>	<i>Calculated Prediction</i>

Data Set 2: \_\_\_\_\_

7. What attributes are being tested?
8. What is the question you are exploring?
9. What is the equation of the Least Squares Line?
10. What is the  $r^2$ -value? What is the  $r$ -value?
11. Describe these relationships in context of the data.
12. Complete the table below by choosing three values to predict. List the value in the first column, then make a prediction by eye and then use a formula to compare your guess to the calculated prediction.

Making Predictions		
<i>Input Value</i>	<i>By Eye Prediction</i>	<i>Calculated Prediction</i>



## Worksheet - Regression Analysis

---

1. Which attributes have the highest correlation?
2. Which attributes have the impact on the dependent variable?
3. Are there any negative correlation relationships?
4. Are the correlation(s) you found strong or weak?
5. What questions arise?
6. What research question arose to provide the opportunity to explore the relationship deeper?
7. What predictions can be made?
8. What are the limitations in the model?
9. Write a summary of your findings, appropriate to an audience with no experience in Data Science

## Student Guide: Regression Mini-Project

---

In this project, you will practice creating models with data. Follow these steps.

- ☐ Find a data set that interests you and has at least 2 numeric columns.

Numeric Column 1:

Numeric Column 2:

- ☐ Upload and store the data into CODAP (Conduct any data cleaning necessary)
- ☐ Create a scatterplot with your data
- ☐ Do you think that this is...
- ☐ A positive linear relationship
  - ☐ A negative linear relationship
  - ☐ A null relationship
  - ☐ Something else: \_\_\_\_\_
- ☐ In the context of the data, why do you think that these attributes have this type of relationship?

- ☐ If you have a relationship that is not linear, test different attributes until you have a linear relationship. Once you have found one, sketch it here and draw a "by eye" model. Be sure to label your axes.



- ☐ Using CODAP, create a linear regression model (if applicable, if not continue to explore attributes to find a linear relationship)

Equation:

- ☐ How good is your model? How do you know?
- ☐ Make at least one prediction with your data. Be sure to write the results in the context of your data.  
*Example: When somebody is 4 years old, you can expect them to be about 40.2 inches tall.*

My Prediction:

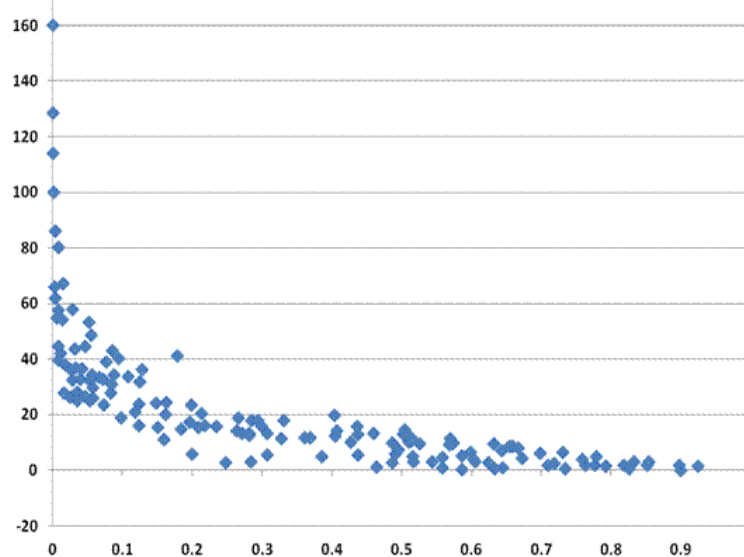
- ☐ Summarize what you found, in the context of your data. *Example: There is a strong linear relationship between a person's age and their height (r-value .87). This makes sense, since people grow taller as they get older. This relationship is strongest from the ages 0 to 16. At about 16 years old, the relationship is not as strong, since people tend to stop growing around then.*

## Printable Day 2 Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Given the following scatter plot, sketch the line/curve of best fit.



Create a scenario using real world context that this model could represent. Your scenario does not have to be realistic but should describe two attributes that would have this type of impact on each other. Write your summary as if your audience has no knowledge of Data Science.

## PrintableDay 3 Exit Tickets

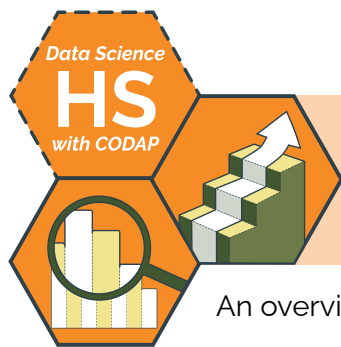
Name: \_\_\_\_\_

Date: \_\_\_\_\_

*Choose any one model that you saw or created in class today. Write a summary of the meaning of that model. What does the model tell you about the situation being analyzed? Are there any limitations in this model? What are the bias in this model? Write your summary as if your audience has no knowledge of Data Science.*

Name: \_\_\_\_\_

*Choose any one model that you saw or created in class today. Write a summary of the meaning of that model. What does the model tell you about the situation being analyzed? Are there any limitations in this model? What are the bias in this model? Write your summary as if your audience has no knowledge of Data Science.*



# Making Predictions

An overview of making predictions using models by Sara Fergus & Christa VanOlst

## Summary

In this two day lesson, students use two techniques to make predictions about missing or future data: by eye and through evaluating functions using a given mathematical model. Students explore datasets throughout the lesson by creating quick scatter plots and models to predict outcomes. In conclusion, students then discuss the tradeoffs and limitations of certain models. Then, students complete a mini-project where they collect data, analyze it for correlation (positive, negative, null), and use modeling to support to disprove their claim.

*Note: This lesson is similar to the Making Predictions lessons from the [Unplugged Data Science](#) & [Data Science with Python](#) sequences. This lesson includes a CODAP activity, which the others omit.*

## Objectives

*The students will be able to . . .*

- Create linear and polynomial models sketching by eye predictions over a scatter plot
- Make predictions given a linear and polynomial models
- Compare and contrast by eye and mathematical model predictions
- Create and test hypothesis statements through visualizing and modeling collected data

## Standards Alignment

- **D.S. 9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.11:** The student will select statistical models and use goodness of fit testing to extract actionable knowledge directly from data

## Materials

- Health Foods Data Talk ([Desmos](#))
- Student Guide: Sketching Models & Making Predictions (see [below](#))
- Review Correlation Vocabulary ([Quizzizz](#))
- Mini Project: Prove a Relationship (see [below](#))
- Additional Data Sets: Crime Data Subset ([CSV](#)), College Admissions Subset ([CSV](#)), Position and Salaries Data ([CSV](#)), Fish Data Subset ([CSV](#))

## Vocabulary

Term	Definition
Line of Best Fit	A line of best fit is a straight line through a scatter plot that represents the linear regression.
Linear Regression	A linear function used to approximate a relationship between two variables used to make predictions and describe situations
Polynomial Regression	A polynomial function used to approximate a nonlinear relationships (quadratic, cubic, etc) between two variables to make predictions and describe situations
Model	A model is a framework for making predictions or describing situations using mathematical equations (i.e. regression) or other algorithms (e.g. decision tree).

## Day 1 Outline

### Formative Assessment Notes

- Warm-Up:** Facilitate this [Health Foods data talk](#), which puts foods on a scatter plot based on health and perceived health. Share student responses using Desmos “select” and “sequence” features. Be sure to discuss these ideas:
  - Have students make a prediction based on the linear relationship. *For example, you may ask students: “if 30% of Americans say a food is healthy, what portion of nutritionists would say that that food is healthy?”. Use this to introduce the concept of prediction with linear regression*
  - Ask students how “good” the relationship is, and why. Use this discussion to transition into the next activity, which introduces a correlation coefficient.
- Using Regression to make Predictions:** Give the students the [Experience & Salary Data \(CSV\)](#). Using CODAP, have students create a scatter plot of the two data attributes. (Years Experience on the x-axis and Salary on the y-axis) (see an [example scatter plot](#))
- Discussion:** Ask students to discuss the following in small groups:
  - How does experience relate to salary? Describe the relationship
  - What do you notice?
  - What do you wonder?
  - What correlation do you see? Are they strong or weak?
  - What questions arise?

When sharing a student's answers, look for vocabulary like “scatterplot”, “linear”, “outlier”, or “relationship”. If these words aren't used, direct students' attention to the linear relationship and model how to use the vocabulary to describe the data.

Check in and make sure students can use CODAP to create scatter plots.

Students should be pretty good at this after the previous lesson—use this as an opportunity to check in

4. **Part 1 Making Predictions:** Have students create a regression model on their scatter plots. [\[example model\]](#)

Have students use the model to predict the outcomes for these scenarios, have them type their results in a text box element:

- A doctor with 7 years experience [predictions ~90K]
- A data scientist with 3 years experience [predictions ~55K]
- A voice-over artist with 12 years experience [predictions ~134K]

5. **Discussion:** Have students research actual average salaries for different levels of experience within these fields to compare to their model. Discuss with students the applicability and limitations of the model in these three cases.

- Are these predictions realistic?
- Who would you ask to help validate these conclusions?
- What other factors lead to difference in salary *besides* work experience?

6. Repeat step 2 using the [Fuel Consumption since 1990 Data Set](#), have students make three predictions in a text box element.

Discuss with students that without technology /programming we could calculate these equations by hand but that would be redundant and has room for error.

7. **Creating Models:** In pairs, give students the following data tables:

- [Crime Data Subset \(CSV\)](#)
- [College Admissions Subset \(CSV\)](#)
- [Position and Salaries Data \(CSV\)](#)
- [Fish Data Set \(CSV\)](#)

Have students complete the [Student Guide - Creating Models & Making Predictions](#) to practice developing linear models.

*Some parts of this activity will be completed below in part 2, step 11.*

8. **Comparing Linear & Quadratic:** Have students work in pairs to compare a fuel consumption prediction from step #6 to a prediction using this quadratic model:

- $\text{FuelConsumption} = -0.113624(\text{year})^2 + 4.620214(\text{year}) + 127.598$

Compare the model in step 6 (linear) and discuss which is better.

Students need to create models more or less independently in step #7. If your students need more review, use this time to do it.

If you needed to do a lot of review in step #4, consider ending here or on step #6 so students have enough time to learn the basics and perform well in step #7.

Through the exploratory analysis, listen for vocab like "curve", "nonlinear", etc.

Check in with students as they complete the worksheet. They may need a review of how to create models using CODAP (a recently developed skill)—re-teach/review as needed.



9. **Part 2 Using Mathematical Models to Make Predictions:** Have students revisit their by eye predictions textbox from step 2. Give them the following linear regression model:

- $\text{Salary} = 8732 * \text{YearsExperience} + 28860$
- *This may denote a quick discussion on how calculating models by hand is not reliable, this is where programming skills for calculations are important for data science models.*

Review slope-intercept form from Algebra 1:

$$y = m(x) + b$$

↓

$$\text{Salary} = 8732(\text{YearsExperience}) + 28860$$

Create a new attribute in the CODAP table and insert a new function using the model above. Demonstrate how to evaluate using a function to predict an outcome:

- This will calculate the predictions for every case
- If, Years Experience = 1.5, then, Salary  $\rightarrow 8732 * 1.5 + 28860 = 41,958$

Have students return to their by eye predictions from step 2 and have them compare to the predictions:

- A doctor with 7 years experience
    - By eye ~90K  $\rightarrow$  Calculated = \$89,984
  - A data scientist with 3 years experience
    - By eye prediction ~55K  $\rightarrow$  Calculated = \$55,056
  - A voice-over artist with 12 years experience
    - By eye prediction ~134K  $\rightarrow$  Calculated = \$133,644
10. Have students work in pairs to repeat step 9 (creating a new attribute as a function) using the [Fuel Consumption since 1990 Data Set](#) and the following quadratic regression model:

- $\text{FuelConsumption} = -0.113624(\text{year})^2 + 4.620214(\text{year}) + 127.598$

Be sure to check in with students during their pair discussion to make sure they can see how the linear models fail to “fit” some of the data sets.

Make sure that students notice that the function shows the slope and the intercept of the regression line, which gives us enough information to make an equation, graph and predict outcomes.

The primary goal of this activity is to have students compare the linear predictions they made with CODAP to “better” predictions using different regression equations.

11. Have students go to the table at the end of the [Student Guide - Creating Models & Making Predictions](#) and pose the question:

***“Do you see any linear models that don’t fit the data well?”***

Give students the following mathematical models for each plot:

Models	
Linear	$\text{prisoner\_count} = 0.004(\text{state\_population}) - 434$
Linear	$\text{chance\_of\_admission} = 0.18(\text{cumulative\_gpa}) - 0.85$
Exponential	$\text{salary} = 23695(1.4)^{\text{level}}$
Quadratic	$\text{weight} = 0.62(\text{diag\_length})^2 - 50.4(\text{diag\_length}) + 1240.26$

Have students calculate predictions by evaluating the function using CODAP or a calculator and then reflect on how these values compare to their previous by eye predictions.

12. **Optional Extra Practice:** Give each group a white board and a dry erase marker. Go through [this slideshow](#), stopping after each slide to show responses. Ask students to share why they picked the value that they did.

When asking students to share, be sure to point out that students may not be choosing the actual value at that point, but predicted (for example, in slide 2, they should choose something near 150, even though the actual point at  $x = 50$  is 10)

13. **Wrap-Up: Limitations of Modeling:** Show students with the [cars.csv](#) data set from the previous lesson, or simply describe the data set to students. It has two numeric columns: car speed(in mph) and stopping distance(in feet).

Give students the regression line for predicting stopping distance from speed:

- Predicted Distance =  $3.93(\text{speed}) - 17.6$

Have students use the equation to predict the stopping distance for cars traveling: 4 mph, 15 mph, 25 mph, 75 mph, 1202 mph and discuss their results.

**Students should learn that mathematical models of data sets have limited ranges of applicability and using a model outside its range can lead to poor predictions.**

## Day 2 Outline

Formative Assessment Notes

14. Have students complete the [Quizizz - Review Correlation Vocabulary](#)

*Summary:* The quizizz allows students to work at their own pace formally assessing through questions on types of correlation and review of categorizing strong, weak, positive, and negative relationships.

**Assess which questions were the most missed, discuss these as a class.**

15. **Prove a Relationship Mini-Project:** In [this mini-project](#), students hypothesize a correlation that can be supported by collecting data from their classmates. Students scatter, plot and model their findings to predict future values and reflect on the limitations of their findings.

See [Assessment Strategies](#) below

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few opportunities for students to show their learning by creating artifacts:

### Mini-Project: Prove a Relationship

In this project, students will come up with a hypothesis to test through surveying classmates. Students will then use these responses to create a scatter plot and assess the relationship. Once plotted students use a model to predict future values and reflect on the applicability and limitations of their findings.

Most students will likely choose something with a null relationship. After the project, take a moment to discuss this with the class.

#### Project Milestones:

- ☐ Make a hypothesis statement
- ☐ Survey and collect data from peers
- ☐ Plot the data on a scatter plot using Python (matplotlib)
- ☐ Sketched a model of best fit and then run a regression model using Python, then compare your by eye predictions
- ☐ Create a presentation to include: hypothesis statement, a summary of data collection, visualization, regression model, and at least one prediction
- ☐ Reflect on their hypothesis statement and the applicability/limitations of their findings.

The next page contains a rubric for assessing student work:

## Mini-Project Rubric

	<i>Proficiency</i>	<i>Yes</i>	<i>No</i>	<i>Notes</i>
<b>Hypothesis</b>	Student created hypothesis is a <b>tangible statement</b> that can <b>prove or disprove a correlation</b> between tested attributes			
<b>Survey</b>	Student created survey is <b>relevant to their hypothesis</b> AND <b>appropriate data is collected</b> , stored, and organized from their peers			
<b>Data Visual</b>	Students' choice of visualizations is <b>appropriate for the data attributes</b> AND <b>provides insight</b> to sketch a model			
<b>Model</b>	Students' sketch of their <b>regression model is appropriate</b> and accurate for the data AND the student <b>makes a valid prediction</b>			
<b>Presentation</b>	Students' presentation includes <b>ALL of the requirements in milestone 5</b> .			
<b>Reflection</b>	Students' <b>reflection is thoughtful and relevant</b> when describing the applicability and limitations of their findings			

## Some Accommodations & Extensions

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Accommodations

Encourage students to have their definitions from the previous classes available for reference, or provide vocabulary sheets for students to use throughout the lesson.

The student guides could be broken into smaller chunks: sketching the lines, answering questions, and then making predictions. In the mini-project, the 7 steps could be converted into a checklist..

Consider recording short videos where you model the following:

- Drawing a regression line "by eye" based on a scatter plot
- Derive a linear equation from a "by eye" regression equation
- Use a regression equation to predict a value given

### Extensions

Consider spending a day or two on the Python Version of this lesson. Making regression equations and predicting values using multiple function family models.

## Student Guide - Using Models & Making Predictions

### Vocabulary

Line of Best Fit	A line of best fit is a straight line through a scatter plot that represents the linear regression.
Linear Regression	A linear function used to approximate a relationship between two variables used to make predictions and describe situations
Polynomial Regression	A polynomial function used to approximate a nonlinear relationships (quadratic, cubic, etc) between two variables to make predictions and describe situations

Using the following data tables: [Crime Data](#), [College Admissions Data](#), [Position and Salaries Data](#), & [Fish](#)


In CODAP, create the following scatter plots and regression models. Sketch or copy/paste your results here.

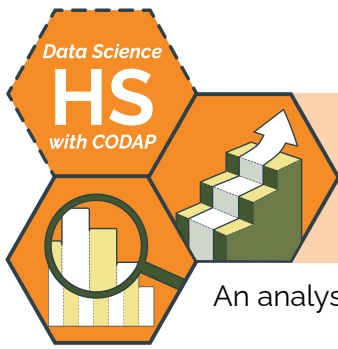
State Population vs. Prisoner Count	Cumulative GPA vs. Chance of Admission
Position Level vs. Salary	Diagonal Length vs. Weight
	Hint: plot each type of fish in a different color

Answer the following questions:

1. How does the incarceration rate in each state compare to the population?
2. Can your GPA impact your chances of admission to Graduate School?
3. How does position level compare to salary? How has your sketch changed? How is this different from the years experience example from before?
4. How does the diagonal length impact the weight of the perch fish? What about the whitetail?
5. Are these correlations positive/negative? Strong/weak?

Complete the following by eye predictions using your models:

Scenario	By  Prediction	Mathematical Model (given in part 2)	Calculated Prediction (completed in part 2)	Comparison (completed in part 2)
California has a population of 39.35 million, what is their predicted prisoner amount?				
A student has a gpa of 8.3, what is their predicted chance of admission?				
A 4.5 level manager should expect a salary of what?				
If a perch fish has a diagonal length of 20 cm, what is the expected weight?				



# Overfitting and Noise

An analysis of overfitting models by Sara Fergus

## Summary

In this lesson, students will be introduced to the concept of "noise" in data science, and how it relates to the overfitting (or underfitting) of predictive models. They will explore the concept of overfitting in a non-computing context to understand its drawbacks in order to be prepared to consider the concept in mathematical modeling.

*Note: This lesson also appears in the [Unplugged Data Science](#) & [Data Science with Python](#) sequences.*

## Objectives

*The students will be able to . . .*

- Assess the strength of a model, taking overfitting and underfitting into account
- Differentiate important underlying patterns in data from noise

## Standards Alignment

- **DS. 9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS. 7:** The student will be able to assess reliability and validity of source data in preparation for mathematical modeling.

## Materials

- Reading: [Model Limitations: Noise and Overfitting](#) (1 per student/reading group)
- Examples of overfitting materials ([Job Posting](#), [Sports](#), [Population](#), [Pattern](#), [President](#), [Flu](#))
- Video: "What is Machine Learning?" ([YouTube](#))
- Article: *Machine Bias* ([PDF](#))
- Data Cycle Scenario – Communication (see [below](#))

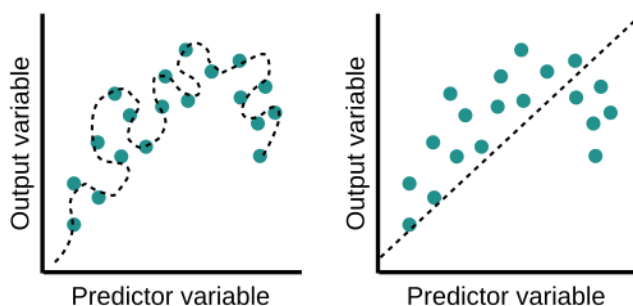
## Vocabulary

Term	Definition
Error	Error is a measure of how inaccurate your model is. Error could refer to training data, testing data, or a combination of both.
Noise	Noise is variation due to natural imperfection or measurement error. "Noisy" data has a lot of variation that is unrelated to the underlying relationship.
Overfitting	A model is overfitted if the noise of the data has a large effect on the model. An overfitted model represents meaningless variation rather than an overall pattern.
Training Data / Training Set	Training data is the data that is used to create a model
Testing Data / Testing Set	Testing data measures the utility of a model by testing to see if it holds for general data that was not necessarily used to create the model. For example, a model created to represent the heights and weights of 8-year olds should be tested using the next year's 8-year olds, and still be a strong model.
Underfitting	A model is underfitted if it fails to demonstrate important patterns in the underlying relationship. For example, using a linear regression to show a quadratic relationship would be underfitting.

## Outline

### Formative Assessment Notes

- Warm Up:** Show the students these scatterplots. Ask them to respond to the following prompt:



**What is wrong with each of these models?**

Students should be able to tell that the left-hand model is overfitted and the right-hand one is underfitted, but they are unlikely to use those words. Try to get them to identify the *concept* so you can apply vocabulary later on in the lesson.



2. **Discussion and Artifact Analysis:** Have students evaluate the accuracy of the statement below, and recording their reasoning in their journals:

***"If one line/model touches more data points than a different line/model, it is a better model"***

Then, have students discuss their thoughts in small groups. As a whole class, have students vote ("true" or "false") and describe examples and counterexamples.

3. Have students read and annotate this worksheet: [Model Limitations: Noise and Overfitting](#), focusing on finding definitions for the terms "noise", "underfitting", and "overfitting".
4. **Station Part 1:** Post each of these five resources at stations around the room along with a posterboard or large sticky note:
1. [Job Posting Resource](#)
  2. [Sports Resource](#)
  3. [Population Resource](#)
  4. [Pattern Resource](#)
  5. [President Resource](#)
  6. [Google Flu Resource](#)

Groups and instruct each group to go to one station. After they read the example, ask them to add this question and a response:

***What is the question that the researcher was trying to answer?***

**Stations Part 2:** Then, have them rotate to the next station. Instruct students to read the resource and review the previous answer. Put a "smile" if you agree, or fix it if you disagree.

Then, have them add another question & response:

***Imagine what data they may have used: What would the cases have been? What would the attributes have been?***

Students should start to consider the idea of noise and overfitting, which they will formalize in the next activity.

Examples / counterexamples should be scatterplots that are overfitted.

Float around to check for understanding.

In general, students should find that the prediction is based too closely on specific instances of the past and overly complicated models.

While reviewing answers, connect student responses to vocabulary like "noise", "bias", "prediction", "training set", and "test set".

**Stations Part 3:** Repeat the cycle, answering the following:

*What would the “training set” be in this example? What would the “testing set” be?*

*How is this an example of overfitting?*

*What would you suggest the researcher do in order to answer their question without overfitting the data?*

On the last rotation, have students check the work of all previous groups. Then, as a whole class, review the answers to each resource.

5. **Overfitting Mini Project:** Complete the [Overfitting Mini-Project](#) below, where students create an artifact that demonstrates their understanding of overfitting.
6. **Modeling & Machine Learning:** Discuss with students that data modeling is heavily used in machine learning to create computers that can identify patterns based on data.

If time allows, consider showing the following video: [What is machine learning?](#) Then, Have students read [this article about machine learning bias](#) and respond to the following in their journals:

- How is machine learning bias related to overfitting or underfitting?
- Is it possible to create an unbiased risk assessment system to help in criminal justice?
- Why do you think companies are not sharing the data that goes into a risk assessment calculation? Do you think that this is appropriate?
- Why do you think the risk assessment system is biased against certain people?
- What are some strategies data scientists should practice to mitigate bias?

Connect the ideas in this article to the *Coded Bias* Ted Talk from earlier in the sequence.

When reviewing answers as a class, circle around the room all together and focus on the “how is this an example of overfitting?” question.

See [Assessment Strategies](#) for details & a rubric.

If reading the entire machine bias article does not make sense for your students or time window, students can read only until “Sometimes, the scores make little sense even to defendants”

7. **Conclusion:** In pairs, have students complete the [Data Cycle Scenario - Communication](#) half-sheet.

*Summary:* Students identify bias in the data collection phase and complete the communication phase of the data cycle given a scenario, synthesizing the information about modeling they have studied over the past several lessons.

Through rapport with students, as you monitor their progress, encourage them to dive deep to describe a clear distinction of the two values [R and R<sup>2</sup>].

Possible outcomes are described in the resource.

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Overfitting Mini-Project Guidelines

**Part 1: Create an example of overfitting:** In this activity, you will create your own example of overfitting. You may either:

- Create a comic that shows overfitting
- Find an example of research that was overfitted and write a brief summary on it, either as a warning or other report.
- Create an overfitted mathematical model to data of your choice
- Come up with another creative example of overfitting, similar to the resources you saw.

Make sure that the presentation is similar to something you would see in the real world, like in the resources we looked at in class. Give students ample time to think of an example before getting started, since the thinking of the example is the important part.

**Part 2: Workshop:** In pairs, workshop your peer's example. Make sure that their example shows overfitting, and then answer the following questions:

1. What is the question that the researcher was trying to answer?
2. Imagine what data they may have used: What would the cases have been? What would the attributes have been?
3. What would the "training set" be in this example? What would the "testing set" be?
4. How is this an example of overfitting?
5. What would you suggest the researcher do in order to answer their question without overfitting the data?

## Overfitting Mini-Project Rubric

	<i>Proficiency</i>	<i>Yes</i>	<i>No</i>	<i>Notes</i>
<b>Example</b>	The example used <b>is an example of overfitting</b> in that it either is too closely based on past instances and/or the model is overly complicated, thus modeling noise more than pattern.			
<b>Presentation</b>	The presentation is <b>clear and understandable</b> .			
<b>Workshop</b>	The student workshopped a peer's project and <b>accurately advised on whether the example is overfitting</b> . The student then <b>thoughtfully answered all questions</b> .			

## Some Accommodations & Extensions

Students who need additional time reading may benefit from getting the overfitting worksheet ahead of time. The worksheet could also be annotated as a class or in a small group. For students with small group accommodations, consider pulling aside a few students and helping them to complete the worksheet, while other students complete the assignment on their own. This could also be helpful for students learning English.

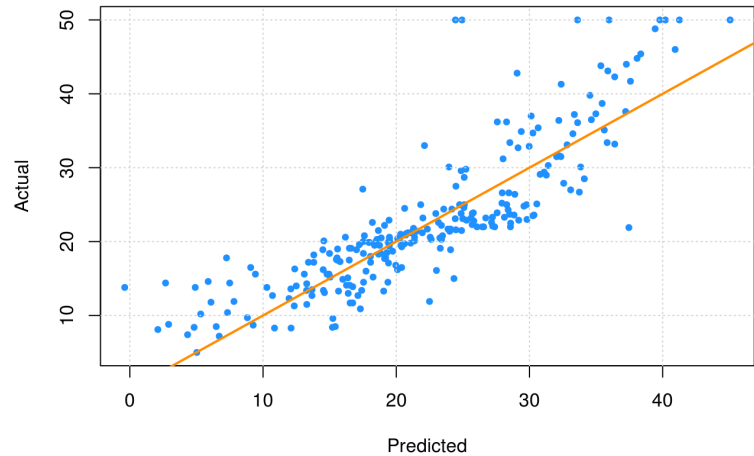
You may provide the [vocabulary list](#) for students learning English.

# Model Limitations: Noise and Overfitting

A guide to noise, overfitting, and bias by Sara Fergus

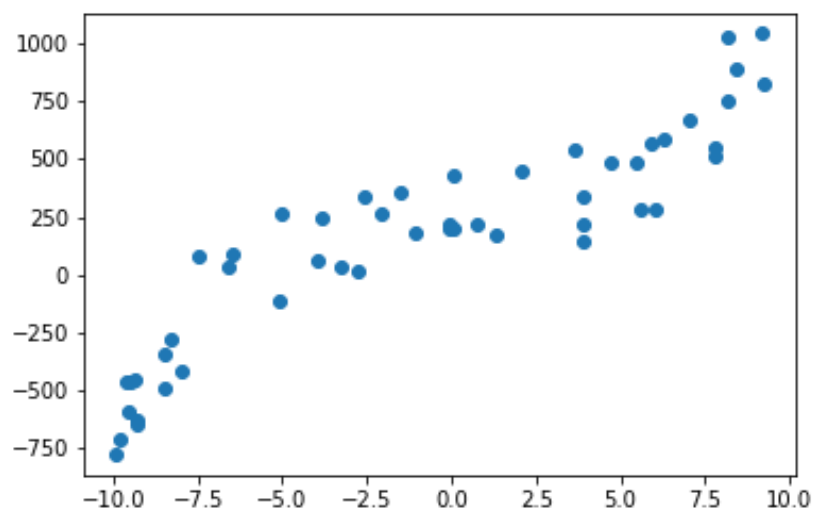
## Noise

In Data Science, **noise** is a word used to describe random pieces of data that make the underlying pattern less clear. This comes from the fact that neither humans nor nature are perfect. For example, noise is introduced by measurement and rounding errors in data collection. Noise is also introduced when there is small variation in a relationship. For example, the stem of a particular flower may be 3 times the length of its petal, but for one flower it is actually 3.2 times the length. The imperfection does not disprove the underlying pattern. This graph shows some noise. You can see that, in general, the data is pretty linear– as the predicted value increases, the actual value increases. However, not all data points fall exactly on that line.



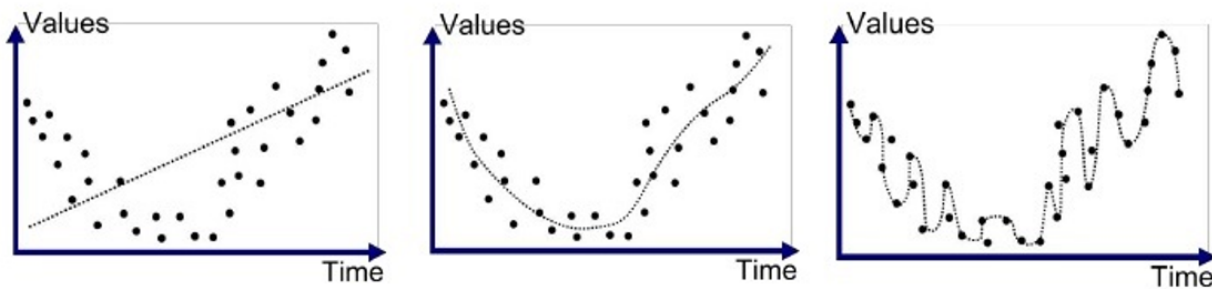
Define “noise” in your own words

Draw a by eye model for this data. Be sure to model the underlying pattern instead of the noise.



## Underfitting and Overfitting

**Underfitting** is when a model is too simple and leaves out some important information. **Overfitting** is when useless details (i.e. noise) have too much of an effect on the model. These graphs show an example of each. You can see that, in general, the values decrease and then increase. A linear model wouldn't be specific enough, because it misses an important aspect of the pattern– the decrease in this case. However, the third graph is overfitted. It takes the individual data points too much into account.



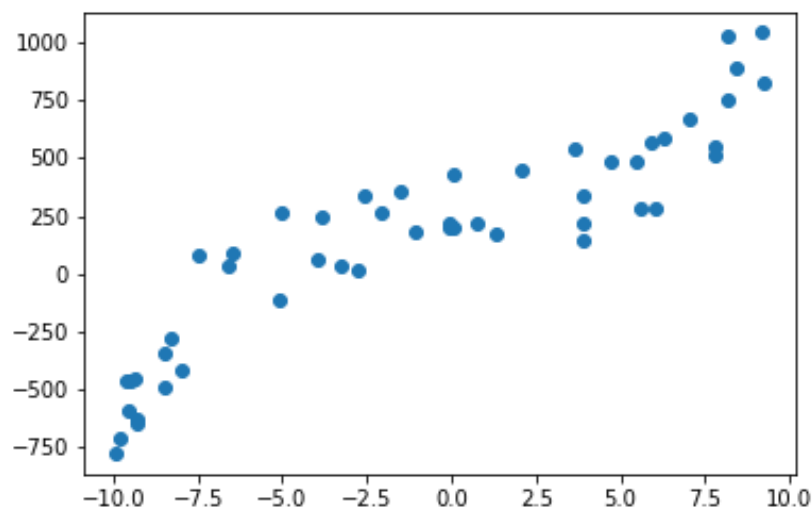
Underfitted

Good Fit/Robust

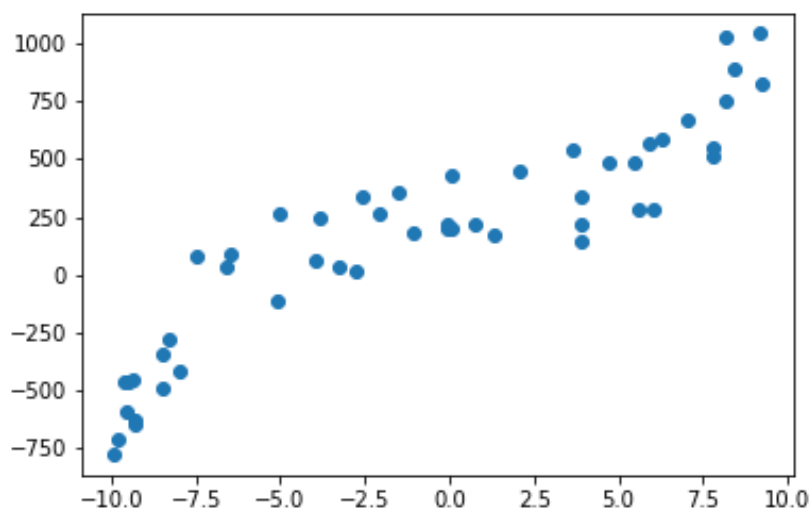
Overfitted

The biggest reason that overfitting is bad is that it will not be accurate on any **test set**.

Sketch a by eye model that would be underfitted



Sketch a by eye model that would be overfitted



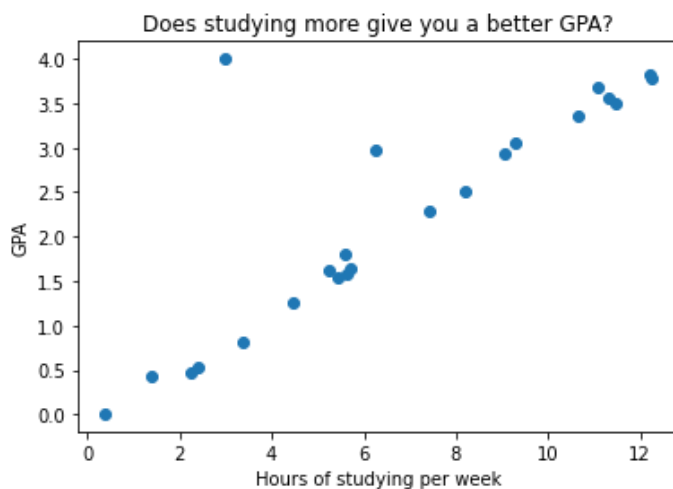
Define overfitting in your own words

Define underfitting in your own words

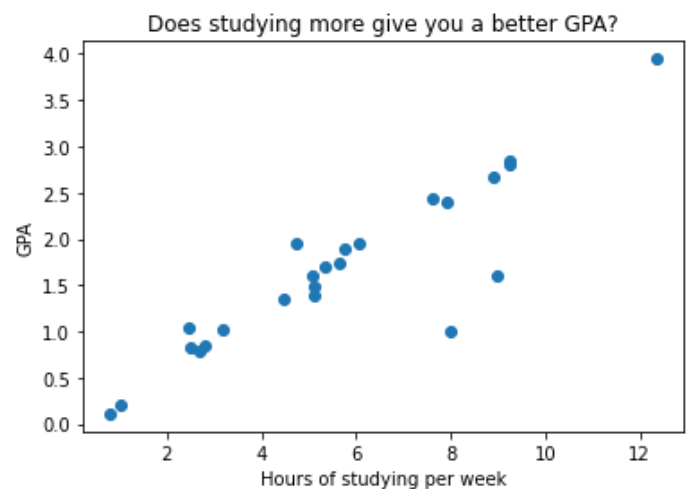
## Training and Testing

When a Data Scientist creates a model, their goal is usually to be able to predict something in the future. For example, a Data Scientist might ask 30 students how many hours they study per week, and what their GPA is. The goal of the study would be to be able to predict the future, to give a guideline like “if you study for at least 5 hours a week, you are more likely to do well in school!” or “If you want a 3.0 GPA, you should probably study for at least 10 hours a week”.

The 30 students in this study are the **training data set**. Their information is being used to create a model. After a model is created, the data scientist would test their model on 30 more students. The 30 additional students would be the **testing data set**. Ideally, the accuracy of the model is pretty similar for both the training and testing. Let's say that these are the results for testing and training. It is pretty clear that if you study more, you will have a higher GPA. Both scatter plots have some noise and outliers, but the trend is pretty clear.



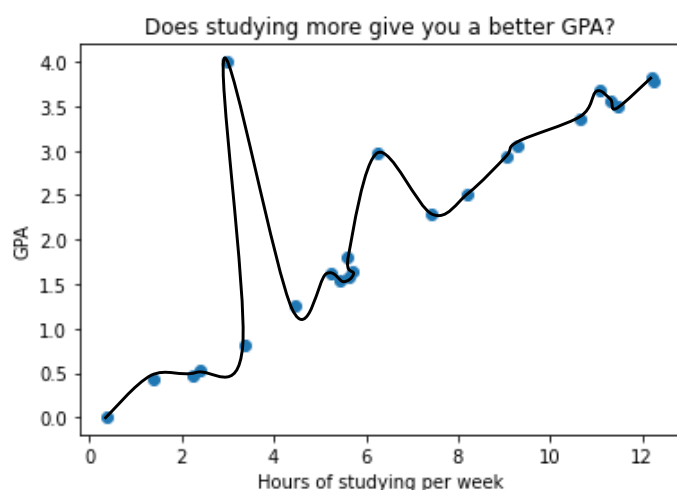
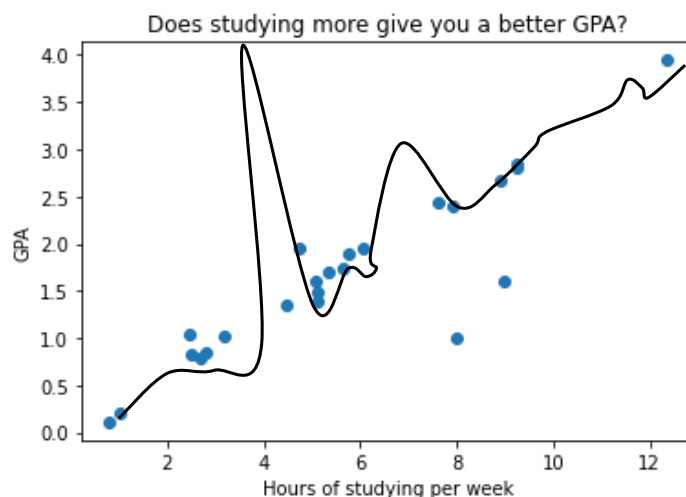
**Training Data**



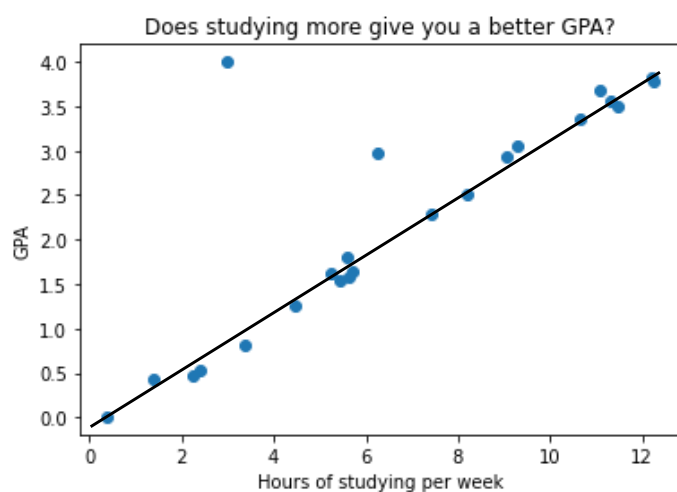
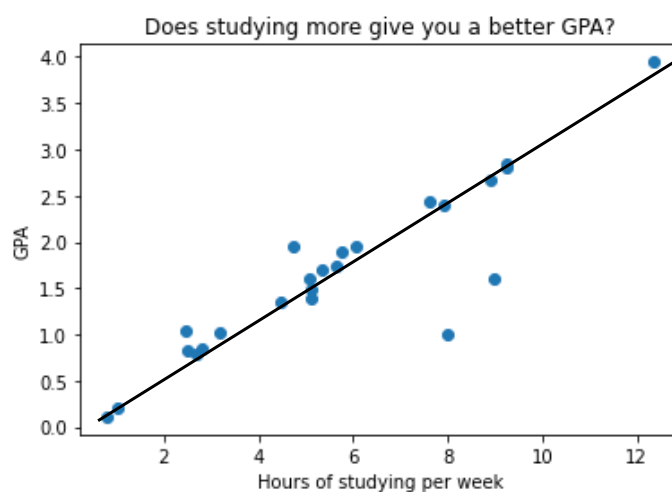
**Testing Data**

Let's say that the Data Scientist overfits their training data. It may look like the training model below. This fits the data *really well*. However, it is an overfitted model. One good way to tell that the model is overfitted is that the increases and decreases are meaningless. For example, this model suggests that something changes after 6 hours of studying that starts to make studying worse. Looking at the bigger picture, however, we can see that that is not true, so the decrease is meaningless. Another hint is that this pattern of increasing and decreasing doesn't hold in the test data (see below).



**Training Data****Testing Data**

A better way to predict this data would be with a linear model. It won't be so exact with the training data, but it will be much better with the testing data.

**Training Data****Testing Data**

Overall, if the data is **overfitted** to the training data, then it is letting the **noise** in that data take over, and will make the testing data less accurate.

Suppose you have a model that shows the height of 8 year olds. Read each scenario. For each, answer the question: **Is this a good model? Why or why not?** Be sure to use the vocabulary from today.

**Scenario 1:** In 2022, the model accurately predicts 95% of 8 year old's heights.

In 2023, the model accurately predicts 23% of 8 year old's heights.

**Scenario 2:** In 2022, the model accurately predicts 45% of 8 year old's heights.

In 2023, the model accurately predicts 51% of 8 year old's heights.

**Scenario 3:** In 2022, the model accurately predicts 87% of 8 year old's heights.

In 2023, the model accurately predicts 91% of 8 year old's heights.

In these scenarios, what is the **training** data set?

In these scenarios, what is the **testing** data set?

A lot of times it is not reasonable to collect data twice. A common thing that data scientists do to make sure they are not overfitting is to break their data into two groups (a training set and a testing set) and make their model using just the first set. Then, they see how well the model fits with the second set. If the model fits both sets pretty well, they know that they most likely have not over or underfitted the data.

## Bias

It is important to not overfit or underfit your data so that your predictions in the future are more accurate. Overfitting can also introduce bias from outliers. For example, the study may have been conducted in a school with a large amount of socioeconomic diversity. Over or underfitting can hide underlying patterns in the data, which gives people opportunities to make decisions that could introduce bias or push a personal belief or agenda.

## Overfitting Practice Answer Key

Suppose you have a model that shows the height of 8 year olds. Read each scenario. For each, answer the question: **Is this a good model? Why or why not?** Be sure to use the vocabulary from today.

**Scenario 1:** In 2022, the model accurately predicts 95% of 8 year old's heights.

In 2023, the model accurately predicts 23% of 8 year old's heights.

*This is not a good model. This model is overfitted. In 2022, the Data Scientist modeled the noise of the data, which was different in 2023.*

**Scenario 2:** In 2022, the model accurately predicts 45% of 8 year old's heights.

In 2023, the model accurately predicts 51% of 8 year old's heights.

*This is not a good model. This model is underfitted. In 2022, the Data Scientist did not account for some major underlying patterns, which caused a poor model both in 2022 and 2023.*

**Scenario 3:** In 2022, the model accurately predicts 87% of 8 year old's heights.

In 2023, the model accurately predicted 91% of 8 year old's heights.

*This model is a good model. Its accuracy is not dependent on noise in a particular year.*

In these scenarios, what is the **training** data set?

*Heights of 8 year olds in 2022.*

In these scenarios, what is the **testing** data set?

*Heights of 8 year olds in 2023.*

## Job Posting Resource

"Agh! Pat is leaving the company. How are we ever going to find a replacement?"



Wanted: Electrical Engineer. 42 year old androgynous person with degrees in Electrical Engineering, mathematics, and animal husbandry. Must be 68 inches tall with brown hair, a mole over the left eye, and prone to long winded diatribes against geese and misuse of the word 'counsel'.

## Sports Resource

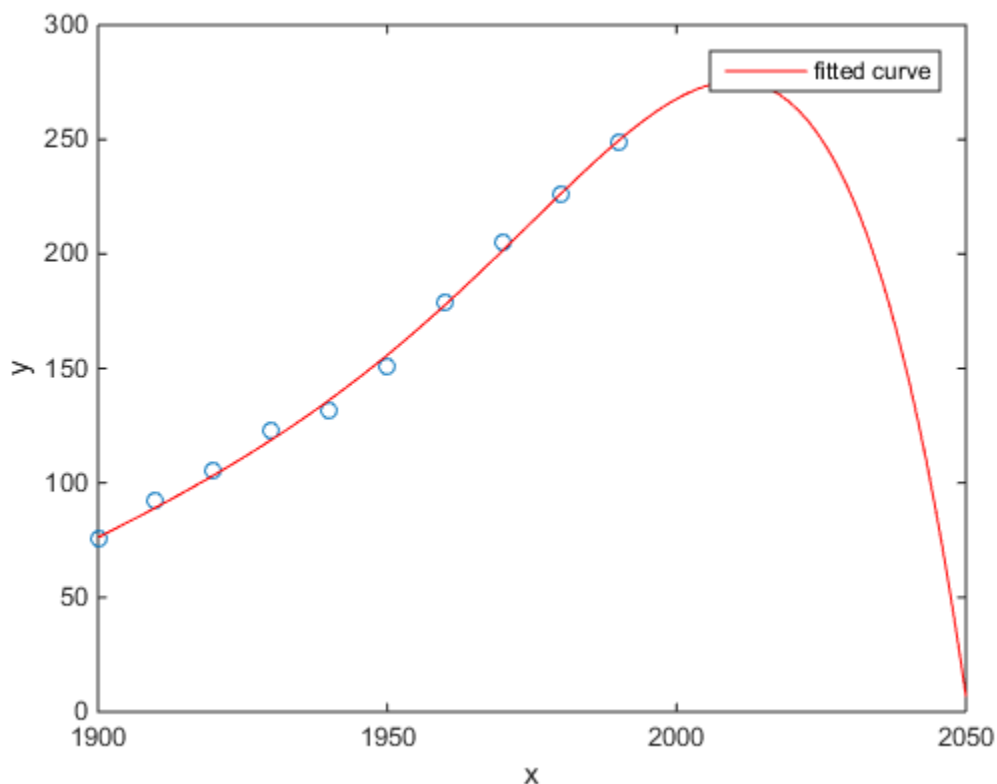


## Population Resource



### *The Apocalypse is Coming!*

According to recent research, we can predict that there will be no living people remaining in the United States by the year 2050. This decline will be beginning in 2010, although the cause is still unknown. This prediction fits previous data very well, so we know that our model is strong.



We suggest moving out of the country as soon as possible. Researchers are still working on modeling populations in other countries. It is possible that this event will not only occur in the United States.

## Pattern Resource

Find the next number of the sequence

1, 3, 5, 7, ?

Correct solution

217341

because when

VIA 9GAG.COM

$$f(x) = \frac{18111}{2}x^4 - 90555x^3 + \frac{633885}{2}x^2 - 452773x + 217331$$

$$f(1)=1$$

$$f(2)=3$$

much solution

$$f(3)=5$$

very logic

wow

$$f(4)=7$$

$$f(5)=217341$$

such function

many maths

wow





**THE PROBLEM WITH STATEMENTS LIKE**  
 "NO <PARTY> CANDIDATE HAS WON THE ELECTION WITHOUT <STATE>"  
 OR  
 "NO PRESIDENT HAS BEEN REELECTED UNDER <CIRCUMSTANCES>"

1788... NO ONE HAS BEEN ELECTED PRESIDENT BEFORE. ...BUT WASHINGTON WAS.	1792... NO INCUMBENT HAS EVER BEEN REELECTED. ...UNTIL WASHINGTON.	1796... NO ONE WITHOUT FALSE TEETH HAS BECOME PRESIDENT. ...BUT ADAMS DID.	1800... NO CHALLENGER HAS BEATEN AN INCUMBENT. ...BUT JEFFERSON DID.
1804... NO INCUMBENT HAS BEATEN A CHALLENGER. ...UNTIL JEFFERSON.	1808... NO CONGRESSMAN HAS EVER BECOME PRESIDENT. ...UNTIL MADISON.	1812... NO ONE CAN WIN WITHOUT NEW YORK. ...BUT MADISON DID.	1816... NO CANDIDATE WHO DOESN'T WEAR A WIG CAN GET ELECTED. ...UNTIL MONROE WAS.
1820... ONLY PEOPLE FROM MASSACHUSETTS AND VIRGINIA CAN WIN. ...UNTIL JACKSON.	1824... THE ONLY PRESIDENTS WHO GET REELECTED ARE VIRGINIANS. ...UNTIL JACKSON.	1828... NEW YORKERS ALWAYS LOSE. ...UNTIL VAN BUREN.	1832... NO ONE OVER 65 HAS WON THE PRESIDENCY. ...UNTIL HARRISON DID.
1836... NO ONE WHO'S LOST HIS HOME STATE HAS WON. ...BUT POLK DID.	1840... NO ONE WHO'S LOST HIS HOME STATE HAS WON. ...BUT POLK DID.	1844... NO ONE CAN BE PRESIDENT IF THEIR PARENTS ARE ALIVE. ...UNTIL GRANT.	1848... AS GOES MISSISSIPPI, SO GOES THE NATION. ...UNTIL 1848.
1852... NEW ENGLAND DEMOCRATS CAN'T WIN. ...UNTIL PIERCE DID.	1856... NO ONE CAN BECOME PRESIDENT WITHOUT GETTING MARRIED. ...UNTIL BUCHANAN DID.	1860... NO ONE OVER 65 CAN GET ELECTED. ...UNTIL LINCOLN.	1864... NO ONE WITH A BEARD HAS BEEN REELECTED. ...BUT LINCOLN WAS.
1876... NO ONE CAN WIN A MAJORITY OF THE POPULAR VOTE AND STILL LOSE. ...TILDEN DID.	1880... AS GOES CALIFORNIA, SO GOES THE NATION. ...UNTIL TILGHMAN HANCOCK.	1884... CANDIDATES NAMED JAMES CAN'T LOSE. ...UNTIL JAMES BLAINE.	1888... NO SITTING PRESIDENT HAS BEEN BEATEN SINCE THE CIVIL WAR. ...CLEVELAND WAS.
1892... NO FORMER PRESIDENT HAS BEEN ELECTED. ...UNTIL CLEVELAND.	1896... TAIL MIDWESTERNERS ARE UNBREATHABLE. ...BRYAN WAGNET.	1900... NO REPUBLICAN SHORTER THAN 5'8" HAS BEEN REELECTED. ...UNTIL MC KINLEY WAS.	1904... NO ONE UNDER 45 HAS BEEN ELECTED. ...ROOSEVELT WAS.
1908... NO REPUBLICAN WHO HASN'T SERVED IN THE MILITARY HAS WON. ...UNTIL TAFT.	1912... AFTER LINCOLN BEAT THE DEMOCRATS WHILE SPORTING A BEARD WITH NO MUSTACHE, THE ONLY DEMOCRATS WHO CAN WIN HAVE A MUSTACHE WITH NO BEARD. ...WILSON HAD NEITHER.	1916... NO DEMOCRAT HAS WON WHILE LOSING WEST VIRGINIA. ...WILSON DID.	1920... NO INCUMBENT SENATOR HAS WON. ...UNTIL HARDING.
1924... NO ONE WITH TWO CS IN THEIR NAME HAS BECOME PRESIDENT. ...UNTIL CROWN COBBLER.	1928... NO ONE WHO GOT TEN MILLION VOTES HAS LOST. ...UNTIL AL SMITH.	1932... NO DEMOCRAT HAS WON SINCE WOMEN SECURED THE RIGHT TO VOTE. ...UNTIL FDR DID.	1936... NO PRESIDENT'S BEEN REELECTED WITH DOUBLE DIGIT UNEMPLOYMENT. ...UNTIL FDR WAS.
1940... NO ONE HAS WON A THIRD TERM. ...UNTIL FDR DID.	1944... NO DEMOCRAT HAS WON DURING WARTIME. ...UNTIL FDR DID.	1948... DEMOCRATS CAN'T WIN WITHOUT ALABAMA. ...TRUMAN DID.	1952... NO REPUBLICAN HAS WON WITHOUT WINNING THE HOUSE OR SENATE. ...EISENHOWER DID.
1956... NO ONE CAN BEAT THE SAME NOMINEE A SECOND TIME IN A LEAF-YEAR REPEAT. ...UNTIL FDR DID.	1960... CATHOLICS CAN'T WIN. ...UNTIL FDR WAS.	1964... EVERY REPUBLICAN WHO'S TAKEN LOUISIANA HAS WON. ...UNTIL FDR DID.	1968... NO REPUBLICAN VICE PRESIDENT HAS RISEN TO THE PRESIDENCY THROUGH AN ELECTION. ...UNTIL FDR DID.
1972... QUAKERS CAN'T WIN TWICE. ...UNTIL NIXON DID.	1976... NO ONE WHO LOST NEW MEXICO HAS WON. ...BUT CARTER DID.	1980... NO ONE HAS BEEN ELECTED PRESIDENT AFTER A DIVORCE. ...UNTIL REAGAN WAS.	1984... NO LEFT-HANDED PRESIDENT HAS BEEN REELECTED. ...UNTIL REAGAN WAS.
1988... NO ONE WITH TWO MIDDLE NAMES HAS BECOME PRESIDENT. ...UNTIL GOLDWATER.	1992... NO DEMOCRAT HAS WON WITHOUT A MAJORITY OF THE CATHOLIC VOTE. ...UNTIL NIXON.	1996... DEMOCRATIC NOMINEES NEVER BEAT TALLER CHALLENGERS. ...UNTIL HERBERT WALKER.	2000... NO NOMINEE WHOSE FIRST NAME CONTAINS A 'K' HAS LOST. ...UNTIL CLINTON DID.
2004... NO DEM. INCUMBENT WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE WHOSE FIRST NAME IS WORTH MORE IN SCRAMBLE. ...UNTIL BUSH DID.	2008... NO REPUBLICAN HAS WON WITHOUT VERMONT. ...UNTIL BUSH DID.	2012... NO REPUBLICAN WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE TWO INCHES TALLER. ...UNTIL BUSH DID.	2016... NO DEMOCRAT CAN WIN WITHOUT MISSOURI. ...UNTIL OBAMA DID.

WHICH STREAK WILL BREAK?

# Google Flu Resource: Read just the highlighted paragraphs of this article.

## FINAL FINAL

## POLICYFORUM

### BIG DATA

# The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,<sup>1,2\*</sup> Ryan Kennedy,<sup>1,3,4</sup> Gary King,<sup>3</sup> Alessandro Vespignani<sup>3,5,6</sup>

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3, 4), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict  $x$  has become commonplace (5–7) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (8). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

### Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (9–11). However, quantity of data does not mean that one can ignore foundational issues of mea-



surement and construct validity and reliability and dependencies among data (12). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (13). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (13). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (2, 14). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (4). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (16–19), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near-real-time health data (2, 20). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

<sup>1</sup>Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. <sup>2</sup>Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. <sup>3</sup>Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. <sup>4</sup>University of Houston, Houston, TX 77204, USA. <sup>5</sup>Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. <sup>6</sup>Institute for Scientific Interchange Foundation, Turin, Italy.

\*Corresponding author. E-mail: d.lazer@neu.edu.



## Data Cycle Scenario - Communication

**DIRECTIONS:** Complete the bias reflection and Communication portion of the data cycle in this scenario.

**Question/Problem Formulation:** The longer your hair grows, the more shampoo you will need.



**Data Acquisition and Collection:** At a local salon I surveyed all of the clients for the day (22 people). I collect each client's hair length in inches and their average amount of shampoo measured in teaspoons.

- Identify any bias in the data collection process:



**Data Processing:** I created a table using my 22 cases. Each case has two attributes: `hair_length` and `shampoo_amount` to begin visualization and analysis.



**Data Visualization and Representation:** I created a scatter plot using the two variables collected.



**Data Modeling and Analysis:** When plotted there seemed to be a very strong positive correlation. When calculating the linear regression line I discovered the following outcomes:

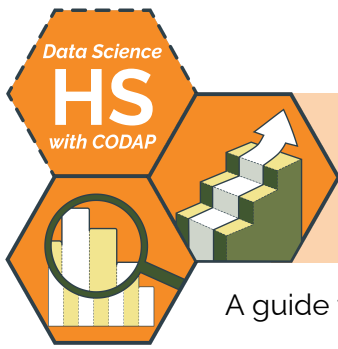
1. Linear Regression:  $0.644(\text{hair\_length}) - 6.56$
2.  $R = 0.961$
3.  $R^2 = 0.924$



**Data Communication:**

(For Teacher Only) **Possible Outcomes**

4. Sampling Bias: Since the data was collected in a salon during one work day a sampling bias most likely occurred due to only getting feedback from a specific portion of the overall audience (a totally random sample). The sampled surveyees may lack diversity in terms of gender, varying hair length, race, ethnicity, age, etc. Also the ability for clients to accurately estimate accurate teaspoons should be considered.
5.  $R = 0.961$  (The correlation between the actual amount of shampoo used and the predicted amount by the model is 0.961, a very strong positive correlation)
6.  $R^2 = 0.924$  (The R-squared for this regression model is 0.924. This tells us that 92.4% of the variation in the shampoo amount can be explained by the length of someone's hair)



# Unplugged: Understanding Research

A guide to interpreting data science research articles by Sara Fergus

## Summary

In this lesson, students will explore the source of a data-based news report to assess the report and to practice reading traditional data reports. They will then use their skills of breaking down and summarizing data-heavy reports to record a small "news clip" describing the results of a detailed data report in layman's terms. This lesson will improve their data literacy skills, as well as their data-scientist skills of translating between data and the general public.

*Note: This lesson also appears in CodeVA's [Unplugged Data Science](#) & [Data Science with Python](#) sequence.*

## Objectives

*The students will be able to . . .*

- Identify the important points in a data-heavy report
- Summarize a data-heavy report into everyday language
- Represent a data story using a mix of verbal language and data visualizations.

## Standards Alignment

- **DS.3:** The student will recognize the importance of data literacy and develop an awareness of how the analysis of data can be used in problem solving to effect change and create innovative solutions.
- **DS.5:** The student will use storytelling as a strategy to effectively communicate with data

## Materials

- [Warm Up Video Clip](#)
- Practice Data Report ([PDF](#))
- [Student Guide: A Data Journalist](#) (1 copy per student)

## Vocabulary

Term	Definition
Data Report	A data report is a report written directly from data. Many articles and newscasts are written from data reports rather than from the data correctly.



## Outline

### Formative Assessment Notes

1. **Warm Up:** Show [this clip](#), where journalists discuss findings about rising temperatures around the world.

Have students write what they think the most important 3 points are. Instruct them to be specific, (not “the world is getting hotter”)

Have students share their points. Organize them on the board.

2. **Exploring the Source:** Dive Deeper: Break students into 5 groups. Give each group a section of [this report](#) to become an expert on:

- Abstract and Introduction
- Future Facing Risk
- Dangerous Days
- Heat Waves
- AC Consumption, Costs, and Emissions

Give students time to read the section on their own. Give each student 3 index cards. On each card, have them write:

- An important point
- How they knew it was an important point
- Where it came from in the article (visualization, specific sentence, specific paragraph, etc)

(These points may or may not match the points from the clip)

3. **Discussion:** In their groups, have students categorize their important points by how they knew it was important (affinity mapping). Once students have sorted their index cards, have each group share their category titles. These should be tips and tricks for reading the report.

As students share, write their tricks on the board. Here are some examples of what they might identify:

- Visualizations usually show important information
- Numbers show important information, but the specific numbers may not be important
- Abstracts are usually pretty good summaries
- Section headers can help you identify main ideas

4. **Mini Project:** Have students completed the *Data Journalist* mini-project (see [Assessment Strategies](#) below).

**Add notes as appropriate for assessing student learning in this step of the lesson**

**Depending on your class environment, you may choose to first hand out one section to each student, and then have them find their group, rather than putting them into groups before starting.**

**If students get stuck here and have trouble identifying important points, go through one of the sections as a class as you model the reading strategies they can be employing as they analyze the reading.**

**See [Assessment Strategies](#) for details & a rubric.**

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Mini-Project: A Data Journalist

In small groups (3-4) have students choose one of these reports:

- [Women's Insurance Coverage](#)
- [Maternity Care](#) (any one section)
- [Gun Violence](#)
- [Partisanship](#) (any one section)
- [Gender Identity](#) (2-3 sections)
- [Social Media and Technology](#)
- Any other article approved by the teacher. One good source is [Pew Research Center filtered as "reports"](#). In approving reports, make sure that the report is lengthy enough to need summarization, includes a substantial number of different visualizations, and allows room for interpretation.

They will create a 2-3 minute news report on the findings using [this guide](#).

### Rubric

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Completeness</b>	You accurately included the most important points from your report in your news clip.			
<b>Representation</b>	You included at least two visualizations that help viewers better understand your report.			
<b>Summary</b>	Your report is two to three minutes long. It is engaging and well-produced.			

## Student Guide - A Data Journalist

In this project, you will be taking the role of a journalist. Your job is to translate between a data-heavy report and something the general public would understand, in the form of a news report. You will be creating a two to three minute clip reporting on the findings from a specific data report.

### Step 1: Choose a Report

Choose a report to learn about:

- [Women's Insurance Coverage](#)
- [Maternity Care](#) (any one section)
- [Gun Violence](#)
- [Partisanship](#) (any one section)
- [Gender Identity](#) (2-3 sections)
- [Social Media and Technology](#)
- Any other article approved by the teacher. One good source is [Pew Research Center filtered as "reports"](#)

### Step 2: Read the article

As you read, keep in mind the tips and tricks for finding important information. Annotate as you go so that you can remember what you would like to report on.

### Step 3: Create a Script

After you have read the article, work with your group to determine what should be included in your news clip. Then, write a script for the reporter to use. Make sure that you cite your source somewhere in the script.

### Step 4: Report the News

Choose one or two group members to be the reporters and record them sharing your script. Your news clip should be 2-3 minutes long and include at least two visualizations (you may show them in your recording, or edit them in after). Use [this clip about climate change](#) as a guide.

### Rubric

	<i>Exemplary</i>	<i>Proficient</i>	<i>Developing</i>
<b>Important Points</b>		You accurately included the most important points from your report in your news clip.	
<b>Visualizations</b>		You included at least two visualizations that help viewers better understand your report.	
<b>Report Style</b>		Your report is two to three minutes long. It is engaging and well-produced.	

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

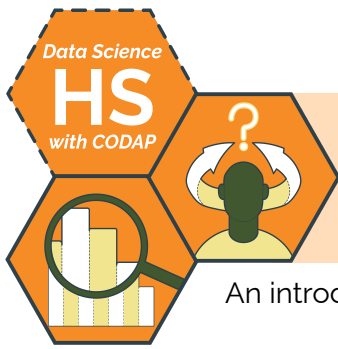
### Accommodations

Students with lower levels of reading or slower processing speeds may choose a smaller report to do their project on. For example, students are generally encouraged to analyze two or three sections of the [Gender Identity](#) report, but some students may be allowed to choose just one section.

Some students may benefit from receiving their portion of the report early, so they can prepare to participate in group work.

### Extensions

Encourage students who need an extension to find a report on a topic that they are interested in, even if it is long or dense.



# Unplugged: Formulating Research Questions

An introduction to project brainstorming by Sara Fergus

## Summary

In this lesson, students will develop questions to answer with data ("Data Questions"). This activity is designed to provide a foundation for student-driven project-based learning, where students find or produce data, generate questions, and make a plan to address those questions using data science skills and practices. Students will use the data cycle to develop questions for research projects and exploratory data analyses.

*Note: This lesson also appears in the CodeVA [Unplugged Data Science](#) & [Data Science with Python](#) sequences.*

## Objectives

*The students will be able to . . .*

- Compare and contrast a research project and an exploratory data analysis
- Ask a relevant question that can be answered with data, including a.) Identifying questions that can or cannot be answered with data, and b.) crafting data-based research questions
- Plan a data science project

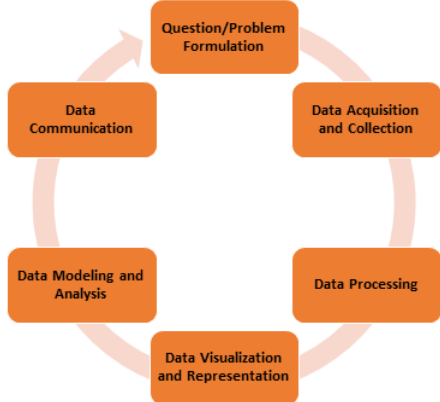
## Standards Alignment

- **DS.1:** The student will identify specific examples of societal problems that can be effectively addressed using data science
- **DS. 2:** The student will be able to formulate a top down plan for data collection and analysis based on the context of a problem.

## Materials

- Research question flowchart ([view the Google Doc](#) or [make a copy, example](#))
- Data Question Worksheet ([view the Google Doc](#) or [make a copy](#))
- Day 1 Exit Ticket, print 1 per student (see [below](#))
- Day 2 Exit Ticket, print 1 per student (see [below](#))

## Vocabulary

Term	Definition
The Data Cycle	<p>In the data cycle, a data scientist will ask a question, collect or acquire the data needed to answer that question, perform a data analysis (including pre-processing and processing, visualizations, models, and general analysis) and then communicate their findings. Once they have communicated their findings, they will notice that new questions have been brought to light. These may come in the form of a further question, the need for an additional attribute of the data, the need for different data (different people, different place, different time, etc), or a critique of the data analysis process. Once that question is created, the data cycle is repeated.</p>  <pre> graph TD     A[Question/Problem Formulation] --&gt; B[Data Acquisition and Collection]     B --&gt; C[Data Processing]     C --&gt; D[Data Visualization and Representation]     D --&gt; E[Data Modeling and Analysis]     E --&gt; F[Data Communication]     F --&gt; A   </pre>
Research Project	<p>A research project follows the traditional data cycle:</p> <ol style="list-style-type: none"> <li>1. Choose a topic you are interested in</li> <li>2. Ask a specific question</li> <li>3. Collect or acquire data to answer the question</li> <li>4. Perform a data analysis</li> <li>5. Draw a conclusion</li> <li>6. Ask a new question</li> </ol>
Exploratory Data Analysis	<p>In industry, you will often see an “exploratory data analysis.” In this case, the data scientist is given data and asked to “make sense of it”. This results in a slightly different interpretation of the data cycle. In the first part,</p> <ol style="list-style-type: none"> <li>1. Acquire or Collect Data</li> <li>2. Explore the Data</li> <li>3. Ask a specific question</li> </ol> <p>Once a specific question is asked, an exploratory data analysis becomes a research project at step 4</p> <ol style="list-style-type: none"> <li>4. Perform a data analysis</li> <li>5. Draw a conclusion</li> <li>6. Ask a new question</li> </ol>



## Vocabulary (continued)

Term	Definition
Data Question	A Data Question is a question that can be answered with data and facilitate a quality data analysis. A data question might arise from a <i>broad question</i> or a <i>subjective question</i> . Answering the question allows further questions to arise. Answering the question should contribute to a larger understanding of the world or an overarching question.
Broad Questions	<p>This is the starting point</p> <p>A <i>broad question</i> is one that cannot be answered on its own because it is unclear and/or undefined. For example:</p> <ul style="list-style-type: none"> <li>• What makes you a good athlete?</li> <li>• Are girls more successful than boys?</li> </ul> <p>A vague question can often break down into good Data Questions.</p>
Subjective Questions	<p>A <i>subjective question</i> is one that cannot be answered as written because it is an opinion. It should be rewritten to focus on public perception of the question.</p> <ul style="list-style-type: none"> <li>• What is the best book ever written? → What is a common favorite book?</li> <li>• Who is the best leader in history? → What traits do people look for in a leader?</li> </ul> <p>A subjective question can often break down into good Data Questions.</p>
"Dead-End" Questions	<p>A <i>dead-end question</i> is one that can be answered with data, but does not lend itself to a data analysis. This is usually because there is only one variable or consideration. It has one simple answer/explanation and can be looked up. It is a fact or figure.</p> <ul style="list-style-type: none"> <li>• How many people live in the United States?</li> <li>• How tall is the tallest person in the world?</li> <li>• Who was Alexander the Great?</li> </ul>
Unethical Questions	An unethical question would require unethical data collection in order to be answered by infringing on privacy or otherwise causing harm.

## Day 1 Outline

### Formative Assessment Notes

1. **Warm Up:** Tell students that during today's class, they'll come up with a plan for a Data Science project about a topic of their choice. Then, have students write in their journal one topic they are interested in, one issue they are passionate about, and one topic they would like to know more about.

Give students the opportunity to share out if they choose to.

2. **Part 1: Can the question be answered with data?** Display each of the following questions (or any questions you would like):

- What is the best book ever written?
- How do I make more friends?
- Who is the greatest athlete of all time?
- Does a person's height help them play basketball?
- How can I save the environment?
- Is the Earth's temperature increasing?
- Who was Alexander the Great?
- What is the most popular clothing brand?
- Are girls more successful in school than boys are?

Give each student a pile of red, green, and yellow sticky notes or dot stickers (any three colors work). Have students read the problems and put a green sticky note if they feel that the question can be answered with data. Put a yellow if it may be able to be answered with data, or parts of the question could be, and put a red if the question cannot be answered with data.

**Discussion:** Place students in small groups. Have each group choose one question that students marked as "green" and discuss what the data for this question might look like. Write what the students share next to the question on the board.

3. **Part 2: Building a Research Question:** On a different board, create a chart with headings "too broad" and "cannot be answered". As a class, sort the questions that were marked yellow or red in step #2 into columns. Consider providing an example with a question or two before having the students sort.

Choose one question in the too broad category and fill out the [project idea flowchart](#) worksheet together as a class

Use the [Examples of questions and categorization](#) resource below.

Consider providing students with examples from previous classes if they are stuck

Pay attention to where students place their sticky notes. If a student is consistently mis-categorizing, check in with them during think-pair-share

It might be beneficial to have students write questions on sticky notes and use dot stickers or to pre-print the questions. This will allow for tactile sorting in step 3.

**Optional Discussion:** How does filling out the worksheet for questions that are too broad help prepare for the project, more than having an already-green question?

4. **Part 3: Your Research Question:** Have students return to what they wrote for their warm-ups. Break students into groups (you can do this randomly, or based on the warm-up)
  - a. Have each group choose one group member's topic and fill out the question flowchart as a group.
  - b. Once they have a question, have groups brainstorm what the data would be. Would they acquire it or collect it? Would it be a survey or observation? What would the cases be? What would the attributes be?
  - c. Have students repeat the process with the other group members' topics.
  - d. Have students share their starting point, their final question, and their data ideas.
5. **Research Question Exit Ticket:** Have students draft a question they might investigate during their final projects

Collect a completed flowchart from the group to assess understanding

Float around during group work to make sure everyone has picked a topic and no one is stuck trying to identify data to use.

See the [Assessment Strategies](#) below.

## Day 2 Outline

### Formative Assessment Notes

1. **Warm-Up:** Have students explore [Kaggle Datasets](#) for data they are interested in. Have them write what the data set is, what the cases are, and what the attributes are.
2. **Exploratory Data Analysis:** Present students with [this data set](#) (World Happiness Report), or a data set of your choice. Have students develop questions they think the data might answer on the board. At this stage, they may be broad questions, like:
  - Are richer countries happier?
  - What makes a country happy?
  - Do countries with more freedom trust their government more?

Then, fill out the flow chart and distilling sheet all together to narrow down their questions.

Once they have finished, tell them that there are two types of Data Science projects. There are research projects (Day 1) and exploratory data analyses (Day 2).

4. **Question Choice Exit Ticket:** Have students write down a question that they might like to investigate for their final project, defining their question, their hypothesis, and the type of data they will need to generate.

Check in with students about their chosen data

Students should be able to identify the attributes in the data set and pose questions based on them without much guidance at this point.

See the [Assessment Strategies](#) below.

## Assessment Strategies

In addition to formative assessments (see *Outline* above), here are a few summative opportunities:

### Day 1 Exit Ticket *See printable version [below](#).*

Have students complete a google form of the questions below or simply print the following:

Name: \_\_\_\_\_ Date: \_\_\_\_\_

1. What is your research question?
  
  
  
2. Why is it a good question to investigate?

Use this as an opportunity to get a sense of what students are interested in studying for their project, and what sorts of data they may need to collect or acquire for it. They'll do something very similar on *Day 2*, which provides you with an additional opportunity to provide feedback and support.

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Data Question</b>	Can be answered with data			
<b>Specific</b>	The research question is not a "broad question" and has been distilled.			
<b>Objective</b>	The research question is objective. If the topic of research is subjective, the research question itself is about people's perception of the topic			
<b>Fruitful</b>	The question cannot be answered with a simple Google Search. Answering the data question would lead to more questioning and future projects.			
<b>Data Collection</b>	Data collection/acquisition is feasible; data accurately describes cases and attributes; cases and attributes contain enough information			

## Day 2 Exit Ticket

See printable version [below](#).

During this lesson, students have worked to identify research and exploratory questions that interest them, and have practiced refining questions into "data questions" that serve as fuel for a project. At the end of the lesson, students will choose a research question for an exploratory data analysis or a research project.

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Is your project a **research project**, or an **exploratory data analysis**? (Circle your choice)

What question will you investigate?

What data will you use?

	<b>Proficiency</b>	<b>Yes</b>	<b>No</b>	<b>Notes</b>
<b>Project Type</b>	Student correctly identifies whether their project is a research project or an exploratory data analysis			
<b>Data Question</b>	Can be answered with data			
<b>Specific</b>	The research question is not a "broad question" and has been distilled.			
<b>Objective</b>	The research question is objective. If the topic of research is subjective, the research question itself is about people's perception of the topic			
<b>Fruitful</b>	The question cannot be answered with a simple Google Search. Answering the data question would lead to more questioning and future projects.			
<b>Data Collection</b>	Data collection/acquisition is feasible; data accurately describes cases and attributes; cases and attributes contain enough information			

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

In the reading activity, articles are rated by difficulty. Newsela articles allow students to change the reading level and the language of the article. Choose a reading level appropriate for your students.

Consider providing reading materials & guiding questions to students in advance.

Activities using the board and sticky notes could be adapted to be online (using tools like jamboard) or you can place sticky notes at student suggestion.

### Extensions

**Reading Assignment:** Break students into groups. Give each student one of [these articles](#). Have them fill out [this worksheet](#) to help them analyze the article. At the end of the worksheet, have students write a summary based on their findings, and share their group summaries with the class.

## Day 1 Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

3. What is your research question?

4. Why is it a good question to investigate?

Name: \_\_\_\_\_

Date: \_\_\_\_\_

5. What is your research question?

6. Why is it a good question to investigate?

Name: \_\_\_\_\_

Date: \_\_\_\_\_

7. What is your research question?

8. Why is it a good question to investigate?

## Day 2 Printable Exit Tickets

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Is your project a **research project**, or an **exploratory data analysis**? (Circle your choice)

What question will you investigate?

What data will you use?

Name: \_\_\_\_\_

Date: \_\_\_\_\_

Is your project a **research project**, or an **exploratory data analysis**? (Circle your choice)

What question will you investigate?

What data will you use?



## Examples

### What is the best book ever written?

This question cannot be answered with data as written. It is *subjective*. Before doing a data analysis, "best" needs to be defined, and there will need to be parameters on time and place. As written, it is also a "dead-end". Once defined, one needs only to find the maximum of a list. Some examples of questions that can be answered with data:

- Which of the current top 20 books gained the most popularity this year?
- How does the popularity of the 20 most popular books vary by country?
- Which genres are most popular by country?
- How long does a typical book take to go from publication to being featured in the New York Times?
- What characteristics make a book a common favorite?
- How does the general popularity of a book relate to it being taught in schools?

### How do I make more friends?

This question cannot be answered with data as written. It is *subjective*. Instead, consider popular belief. Some examples of questions that can be answered with data:

- What characteristics do people look for in friends?
- What do people think the best way to make friends is?
- How many close friends do people have throughout their lives?

### Who is the greatest athlete of all time?

This question cannot be answered with data as written. It is *vague*. What does it mean to be the "greatest athlete"? It is a "dead-end". Once "greatest athlete" is defined, no deep data analysis is required. Instead:

- Who do people consider to be the greatest athlete of all time?
- How many points have each of these three basketball players scored over the years, and how has their ranking changed?
- Are people's opinion of the "greatest athlete of all time" influenced by their favorite sport?

### How can I save the environment?

This question cannot be answered with data as written. It is *vague*. What does it mean to "save the environment"?

- How does an individual's pollution compare to a corporation's? Do individuals have the power to change the rate of pollution without regulating corporations?
- How much trash is in the ocean, and has the rate of ocean litter increased over time? How much of the ocean's trash is in parts of the ocean which are "highly populated" by wildlife?

### Is the Earth's temperature increasing?

This question cannot be answered with data as written. It is a *dead-end* question. One could Google the answer. Instead:

- How has the rate of global warming changed over time, and does that relate to the worldwide human population?
- What actions have the biggest impact on global warming?
- How has the temperature of the Earth changed over time? How does that vary based on geographic location? How has the rate of increase changed over time?

### Who was Alexander the Great?

This question cannot be answered with data as written. It is a *dead-end* question. One could Google "Who was Alexander the Great" and get a simple description of who he was. Instead:

- How much do people today know about Alexander the Great?
- How did Alexander the Great's rule change the economy of ancient Greece?
- Is there a pattern in when and where Alexander the Great's many invasions were successful?
- What were the common ways for ancient Greek kings to come to power, and did those methods change over time?

### What is the most popular clothing brand?

This question cannot be answered with data as written. It is a *dead-end* question.

- What are people's perceptions of popular clothing brands?
- What makes a clothing brand popular? What events lead to popularity? For example, does a celebrity endorsement increase sales of a clothing brand?
- What is the most popular clothing brand right now, and how has that changed over time? How much more popular is the most popular brand than the second-most?

### Are girls more successful in school than boys are?

This question cannot be answered with data as written. It is a *vague* question. What do you mean by "successful"? Simple fixes to the question could make it *dead-end* (are there more girls or boys in college?)

- How has the gender breakdown of college enrollment changed over the years?
- How happy are girls in comparison to boys, and how does that vary country-to-country?

## Worksheet: Data Science in the World

A reading guide by Sara Fergus

Use this worksheet to help you analyze a data science article. In this assignment, we are paying special attention to research questions and data collection.

1. What is the title of your article?
2. Was this an exploratory data analysis or a research question?
3. What was the research question the author was trying to answer? *Note: they may not have written it exactly!*
4. Imagine what data they may have used:
  - a. What would the cases have been?
  - b. What would the attributes have been?
5. Did the article answer their research question? If they did, what was their answer?
6. Did the article suggest new questions or changes at the end of their article? If they did, what questions or changes did they suggest?

Using your answers to the questions above, write a summary of the article to share with your class:

Articles to choose from:



CS Lesson Plan

This work is licensed under a [CC-BY-SA-NC 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/)  
Attribute to "CodeVA 2022" or "Sara Fergus for CodeVA 2022"



[What the DNA of Ancient Humans Reveals about Pandemics](#) (Hard, Wired)

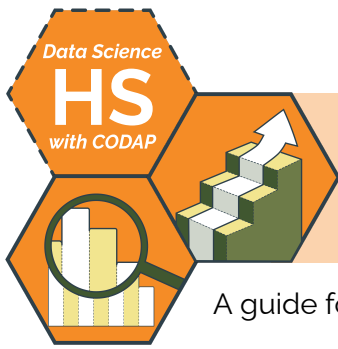
[The US Can Halve its Emissions by 2030– if It Wants To](#) (Hard, Wired)

[Can Disgusting Images Motivate Good Public Health Behavior?](#) (Medium, Wired)

[One-fifth of Reptiles Worldwide Face Risk of Extinction, Study Finds](#) (Easy, Newsela)

[Study Topic Influences Funding Disparity for Black Scientists](#) (Easy, The Scientist)

[More than a Million Reasons for Hope: Youth Disconnection in America Today](#) (Medium, Measure of America)



# Project Practice

A guide for students to model the application of the data cycle skills by Christa VanOlst

## Summary

Throughout the lesson, students will complete a full iteration of the data cycle by modeling question formulation, data collection, analysis, visualization and the modeling processes. This 2 day lesson includes a guide for students to use on a class Data Science Research Project. On day 2, students will be given local data sets to conduct an Exploratory Analysis Data Science Project to reiterate the data cycle skills.

*Note: Variations on this lesson appear in the [Unplugged Data Science](#) & [Data Science with Python](#) sequences.*

## Objectives

*The students will be able to . . .*

- Complete at least two iterations of the data cycle
- Employ data cycle skills developed throughout the course, including:
  - a. constructing a strong data question
  - b. collecting/acquiring reliable and useful data,
  - c. processing data effectively, controlling for bias,
  - d. creating visualizations and models to understand data,
  - e. presenting findings as a data story or a data science write-up
- Conduct a Research Project
- Conduct an Exploratory Analysis Project

## Standards Alignment

- **DS.1:** The student will identify specific examples of real-world problems that can be effectively addressed using data science.
- **DS.2:** The student will be able to formulate a top-down plan for data collection and analysis, with quantifiable results, based on the context of a problem.
- **DS.3:** The student will recognize the importance of data literacy and develop an awareness of how the analysis of data can be used in problem solving to effect change and create innovative solutions.
- **DS.6:** The student will justify the design, use and effectiveness of different forms of data visualizations.
- **DS.9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.13:** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

## Materials

- [Data Science Project Design Menu](#) & Design Scaffold (view [Google Drawing](#) or [make a copy](#))
- Project Write-Up Template (see [below](#))
- Data Science Design Flowchart (view [Google Doc](#) or [make a copy](#)) & [example](#)
- Data Science Research Question Distiller (view [Google Doc](#) or [make a copy](#))
- 11X17 Graph Paper (view [Google Doc printable](#) or [make a copy](#))
- Grocery Store Marketing Analytics (view [Google Sheet](#) or [make a copy](#))
- Data Sets Websites: [Data World \(Virginia\)](#), [Open Data Network \(Virginia\)](#), [Kaggle Datasets](#)
- Data Sets: Diamond Prices ([CSV](#)), Netflix Daily Top 10 ([Kaggle](#)), U.S. International Air Traffic Data (1990-2020, [Kaggle](#)), Forbes Highest Paid Athletes (1990-2020, [Kaggle](#)), Antarctica Penguin Data ([Kaggle](#)), Climate Change: Earth Surface Temperature Data ([Kaggle](#)), Real / Fake Job Posting Prediction ([Kaggle](#)), Bigfoot Sightings ([data.world](#)), Harry Potter Data Sets ([Kaggle](#))

## Vocabulary

Term	Definition
The Data Cycle	The Data Cycle is a framework of data science. In the data cycle, a data scientist will ask a question, collect or acquire the data needed to answer that question, perform a data analysis (including pre-processing and processing, visualizations, models, and general analysis) and then communicate their findings. Once they have communicated their findings, they will notice that new questions have been brought to light. These may come from a further question, the need for an additional attribute of the data, the need for different data (different people, different place, different time, etc), or a critique of the data analysis process. Once that question is created, the data cycle is repeated.
Research Project	A research project follows the traditional data cycle: <ol style="list-style-type: none"> <li>1. Choose a topic you are interested in</li> <li>2. Ask a specific question</li> <li>3. Collect or acquire data to answer the question</li> <li>4. Perform a data analysis</li> <li>5. Draw a conclusion</li> <li>6. Ask a new question</li> </ol>
Exploratory Data Analysis	In industry, you will often see an "exploratory data analysis" in this case, the data scientist is given data and asked to "make sense of it". This results in a slightly different interpretation of the data cycle. In the first part, <ol style="list-style-type: none"> <li>1. Acquire or Collect Data</li> <li>2. Explore the Data</li> <li>3. Ask a specific question</li> </ol> <p>Once a specific question is asked, an exploratory data analysis becomes a research project at step 4.</p>

## Day 1 Outline

### Formative Assessment Notes

1. **Warm Up:** Given the following bad research questions have students annotate to rewrite them in their journals:
  - Which national park is the best?
  - What are the advantages and disadvantages of cell phone use in schools?
  - Are gray cats better than orange cats?
  - Has the population of the world increased in the past century?

Have students share their updates with a peer. Here are some possible re-written versions of the questions above:

- What features do the most popular national parks have in common?
- How does restricting cell phone use in school affect student social interaction?
- When tested for intelligence and longevity, how do gray cats and orange cats compare?
- What factors have influenced population growth in the fastest growing countries?

2. **Class Research Project:** Group students in pairs (or individually if desired). Introduce the following research prompts to students or have students add to the list by creating their own:
  - a. In what ways does having a pet at home require responsibility from a child?
  - b. What features do the best colleges have?
  - c. How do government regulations impact the pollution produced per state?
  - d. In what ways do students in different grade levels deal with stress throughout the four quarters?
  - e. What activities are included in an enjoyable first date?
  - f. What social media apps produce the most screen time?
  - g. How does time on social media impact the amount of impulse buyers?
  - h. How does the role of fitness ads affect young adult exercising practices?
  - i. What would the world economy be like without wars?
  - j. What characteristics did the world's most successful leaders have?

Have students choose one of the questions above and use the research question [flowchart](#) and [distiller](#) to narrow down their question.

**Pay attention to how students are rewording the questions. If a student isn't making effective changes check in with them during the share out.**

**Have students do a quick check in with you to assess their understanding on their distiller.**

3. **Project Practice:** Have students follow the steps below to complete their practice research project:
- Have students complete [Part 1](#), where they design their project using the DS Project Menu & the DS Designing Scaffold.
  - Have students complete [Part 2](#), where they will plan and implement creating their visualizations, modeling, and analyze their findings.
  - Have students complete [Part 3](#), where they will share their findings
  - Have students complete [Part 4 \(Reflection\)](#)

See [Assessment Strategies](#) below for a rubric.

## Day 2 Outline

### Formative Assessment Notes

1. **Warm Up:** Given the following [Diamonds Data Set](#), have students use CODAP and exploratory analysis to support or disprove:

***"The bigger the diamond the better it is."***

**Students should identify attributes that are impacted by the size of a diamond including clarity, cut, and color.**

2. **Exploratory Data Analysis Project:** Group students in pairs (or individually if desired). Provide the students with 2-3 data sets. Use the following sites to explore local data sets:

- [Data World \(Virginia\)](#)
- [Open Data Network \(Virginia\)](#)
- [Kaggle Datasets](#)

**Have students do a quick check in with you to assess their understanding on their distiller.**

Or have students choose from the following:

- [Netflix Daily Top 10](#) (March 2020 - March 2022)
- [U.S. International Air Traffic Data](#) (1990-2020)
- [Forbes Highest Paid Athletes](#) (1990-2020)
- [Antarctica Penguin Data](#)
- [Climate Change: Earth Surface Temperature Data](#)
- [Real / Fake Job Posting Prediction](#)
- [Bigfoot Sightings](#)
- [Harry Potter Data Sets](#)

Using their chosen data, have students fill out the research question [flowchart](#) and [distiller](#) in their groups to narrow down their research question.



6. **Exploratory Project Practice:** Have students complete the steps below to complete an exploratory data science project:
- Have students complete [Part 1](#), where they design their project using the DS Project Menu & the DS Designing Scaffold.
  - Have students complete [Part 2](#), where they will plan and implement creating their visualizations, modeling, and analyze their findings.
  - Have students complete [Part 3](#), where they will share findings
  - Have students complete [Part 4 \(Reflection\)](#)

See [Assessment Strategies](#) below for a rubric.

## Some Accommodations & Extensions

Consider implementing the accommodations & extensions below as needed. Remember that all students may benefit from accommodations, even if they are designed to meet a particular student's needs.

### Accommodations

You may choose to create groups strategically in order to balance student's strengths and weaknesses, or in order to create groups that you intend to spend more time supporting.

Some students may benefit from an abbreviated version of the write up template, that includes the title, research question, data analysis and findings (together), and conclusion.

### Extensions

For students who finish early, you may encourage them to create a tactile model (see project examples). You could also choose to have them supplement their project with an analysis of an additional data set, or ask them to collect supplemental data to address questions that arise in their initial analysis.

## Practice Project Rubric

	<i>Exemplary</i>	<i>Proficient</i>	<i>Developing</i>
<b>Question Formulation and Project Design</b>	<p>The research question is one that can be thoroughly answered with data</p> <p>Research question is relevant with real-world applications</p> <p>Question is clearly communicated</p>	<p>Research question is well communicated but cannot be properly answered with data science OR</p> <p>Research question is well communicated and can be answered, but is irrelevant to the real world OR</p> <p>Question is relevant, but is not clearly communicated</p>	The question is communicated
<b>Data Selection and Preparation</b>	<p>Substantial data is selected from a reputable source</p> <p>Selected data corresponds with the question</p> <p>You have vetted the data set to avoid issues</p>	<p>Appropriate data is selected, but the data set is not large enough to reliably answer your research question OR</p> <p>Appropriate data is selected, but is unreliable OR</p> <p>The data selected is reliable and substantial, but irrelevant.</p>	Data is selected.
<b>Visualizations</b>	<p>Multiple visualizations communicate project findings</p> <p>Visualizations are clear, concise, and well explained</p> <p>Visualizations are appropriate for the data</p>	<p>Exactly one visualization communicates project findings OR</p> <p>One or more visualizations are unclear or poorly labeled, but are present and appropriate OR</p> <p>Choice of one or more visualizations are not suited to the data, but other visualizations demonstrate findings</p>	Visualizations are missing or are invalid.
<b>Models</b>	<p>An accurate, predictive mathematical model is created to help answer the research question.</p> <p>The model is accurately interpreted</p>	<p>A mathematical model is created with small errors OR</p> <p>A mathematical model is created, but cannot be used to answer the research equation</p>	A mathematical model with substantial errors in accuracy, applicability, and explanation is created.
<b>Communication</b>	Write up successfully communicates the question with background information, data collection techniques and decisions, data cleaning techniques and decisions, modeling and visualization decisions, model limitations, and findings. Final deliverable successfully communicates the question and the findings.	<p>Two or more pieces of the write-up (communicate the question with background information, describe: data collection techniques and decisions, data cleaning techniques and decisions, modeling and visualization decisions, model limitations, and findings) are missing OR</p> <p>A substantial portion of the write up is unclear</p>	Writeup is unclear or is missing a significant amount of essential information

## Part 1: Design the Project

Use the template linked [here](#) (make a copy by clicking [here](#)) to create a Project Plan, where you define your research question, set goals for what data science skills you will use, and define who your audience will be when you present your work at the end of the project.

## Part 2: Complete Analysis

- **Locate Data:** Collect Data or explore the resources provided (or another resource, like [kaggle](#)) for a data set that can be used for your project
- **Plan Visualizations:** Based on the types of data you collected, what sorts of visualizations make sense? What pieces of the data relate to your research question, and how can you represent them? Write some ideas here:
- **Plan Models:** Determine whether a descriptive or predictive model can help you tell your story.

Data to Represent	Possible Visualizations and/or models

- **Create Visualizations:** Make sure they are accurate, clear and clearly labeled, and presented in a way that uses one of the [aesthetic perspectives](#). *Disruption* is one that may make sense for this project.
- **Create Models:** Create your model using CODAP.
- **Answer your Question:** Using the [Write Up Template](#), draw connections between your data and multiple representations of your data, and how they answer your research question. Make sure your findings are clear and directly related to the research question. Make sure your final argument is clear.
- **Reflect on the Data:** Consider your findings and how they relate to the real world. Share your reflection through a solution or call to action, an infographic, or a reflective portion of your write-up..

## Part 3: Share your Findings

Share your project with your community. If you created an infographic or video, you could share on your personal social media account, or ask to share on your school's social media account.

- Create an artifact to communicate your findings. You should use visualizations created with CODAP, but you may choose to add to the overall infographic using sites like [canva](#).

**Part 4: Reflection****Student Reflection**

Describe what went well.

Describe what you struggled with.

Describe one way you would improve on your project.

Describe a future step for data collection or analysis.

Share your personal progress throughout the project.

Reflect on your management of this project.

- Did you meet most deadlines?
- Did you use your class time wisely?

Describe your overall experience with this project.

## Template: A Write-Up

[Title of Write-Up]

[Subtitle]

**The Title:** The most important thing about your title is that it communicates what the paper is about. Be creative! If you have a creative title that does not fully communicate the topic of the paper, add a subtitle.

### Research Question and Background

Here, clearly state your research question. Make sure you take time before you start to [develop a strong research question](#). Briefly explain any context that is necessary for understanding your research question. Be sure to explain *why* your question is important, why should the reader care about your question?

Then, provide some background research on the question.

**Tone:** The tone of your paper should be relatively scientific. Avoid "talking to your reader" ("I bet you are wondering..."). However, it is not necessarily bad to describe your personal interest.

### Data and Data Collection

First, describe where and how you collected / found your data. Then, describe the data itself— for example say how many entries there are, or how many questions were asked. If you had to do any data cleaning, describe the decisions you made and why you made those decisions.

### Data Analysis

Now you get to share your findings! This section is where you put your well-labeled visualizations and models. In text, make connections between the visualizations and the models. Briefly discuss what each visualization or model shows. Make sure any visualizations are referenced by number and labeled. For example, in Figure 1 you see a brief description of what is being shown.



Figure 1. A smiley face

### Findings

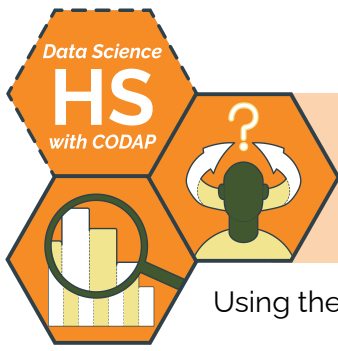
Now that you have run your data analysis, answer your research question. In answering your research question, directly reference research that you conducted and trends/patterns shown in the visualizations and models you created. Make sure your question is clearly answered.

### Conclusion

Here, discuss any problems in your analysis. For example, how might different data cleaning decisions have affected your findings? Or different data collection techniques?

Then, discuss any questions that arise from your analysis. Maybe there is a part of your research question you were not able to fully answer. Maybe there is an interesting follow-up question you thought of while conducting your data analysis.

This is also where you can give recommendations on how to enact positive change based on the findings in your data analysis.



# CODAP Summative Project

Using the computer to tell an engaging data story by Sara Fergus

## Summary

Throughout this unit, students have learned to use CODAP to tell a data story. In this project, students will use or collect data to explore and analyze a topic that interests them. They will run a data analysis, and then present their findings in a meaningful deliverable that can inspire deep thought or action. To tell their story, they will draw on the data questioning, visualization, and modeling skills they have learned.

## Objectives

*The students will be able to . . .*

- Ask a relevant question that can be answered with data
- Conduct an exploratory data analysis
- Produce high quality and relevant visualizations to communicate their findings
- Produce a model to represent their data or make a prediction
- Summarize a data analysis by connecting the question to findings and visualizations to next steps and proposals
- Propose solutions to structural problems to representatives from the community

## Standards Alignment

- **DS.1** The student will identify specific examples of real-world problems that can be effectively addressed using data science.
- **DS.2** The student will be able to formulate a top-down plan for data collection and analysis, with quantifiable results, based on the context of a problem.
- **DS.5** The student will use storytelling as a strategy to effectively communicate with data.
- **DS.8** The student will be able to acquire and prepare big data sets for modeling & analysis.
- **DS.9** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.12** The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.
- **DS.13** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

## Materials

- The [CODAP](#) browser-based data analysis tool
- The student-facing *Project Frame* document (see below, view [Google Doc](#), or make a copy)
- The *Data Science Project Design Menu* (view [Google Drawing](#) or [make a copy](#))
- The *Project Write-Up Template* (see [below](#), view [Google Doc](#), or [make a copy](#))
- *Unions* (see [below](#), [.codap](#)) and *Standardized Testing* ([.codap](#)) example projects

## Before the Lesson

This summative project is very open-ended, and requires a high degree of independence on the part of students. They are expected to choose a dataset that addresses a question of interest to them (or collect relevant data themselves), perform analysis on that data, and draw conclusions without much scaffolding from the teacher. In order for students to be successful, you will likely need to do some preparatory work before having them start working on the project frame. Here are some suggestions:

- **Develop Questions In Advance:** Consider facilitating the [13 Developing Questions](#) lesson plan from this sequence to help students figure out what they will investigate during their project.
- **Practice the Project:** Consider facilitating the [14 Project Practice](#) lesson plan from this sequence to help students go through the entire project process in a less open-ended way so they can see what sort of work they should plan to do during their self-directed project.
- **Analyze Project Examples:** Have students analyze the examples linked in the *Materials* section (and in the project frame document) as a group to develop a sense of what a successful project might look like.

## Using This Document

The following pages in this document are intended to be filled out by the *student* as they work through their summative project. It takes the form of a checklist, showing the different steps students should go through as they plan, execute, and share their project. You can distribute this document by printing it (leaving off the first two pages), or digitally by [making a copy](#) of this [Google Doc](#).

You can structure students' engagement with the project in several ways. We do not provide explicit guidance because your scaffolding and guideline should be responsive to your students (who we will never know). Here are some suggestions:

- **Pacing:** If you want to provide a pacing guide, set deadlines for each step in the checklist based on how much time you think students should take to complete them.
- **Scaffolding Choice:** Sometimes, students will have a difficult time finding a data set relevant to their interests. Consider providing options for them to choose from (see [14 Project Practice](#)) if they are not successful in choosing their own.
- **Modifying the Design Menu:** The *Project Design Menu* is available as an [editable Google Drawing](#); replace the options we have provided with options that you think are better suited to your students' inquiries. Consider filling out the design menu in collaboration with students so you can guide their project goal-setting process.

# Data Science with CODAP Project Prompt

Relationships in data help us write a data story, and our data stories help us make meaning of the world around us. In this project, you will explore a dataset to highlight a meaningful relationship. You could look for a simple relationship, like correlation, or a different pattern of data relationships. Then, organize your findings into a simple visual that you can share with the community, so that they can be aware of the relationship you found. This may be an infographic, a video, a physical representation, or anything else.

Explore the examples to get a better idea of what you could do!

- CODAP Project: Unions (see [below](#), [.codap](#))
- CODAP Project: Standardized Tests ([.codap](#))

## 1. Design the Project

☐ **Brainstorm Ideas:** Consider the following questions as you plan your data science project:

- Where in your community might something be underrepresented or hidden?
- How could your data analysis show something that might be underrepresented or hidden?
- How could it contribute to a cause that you are passionate about?
- How does it change with time or interpretation?
- How does it show an experience that many people in your community have?

Write some ideas here:

☐ **Choose One Idea:** Base your idea on interest and the feasibility of data collection. Turn your idea into a well-constructed *data question*. Write your data question here:

☐ **Plan Your Project:** Use the *Project Design Menu* (view [Google Drawing](#), or [make a copy](#)) to create a project plan, where you define your data question, set goals for what data science skills you will use, and define who your audience will be when you present your work at the end of the project.



## 2. Complete the Project

- ☐ **Acquire the Data:** Explore the resources provided throughout the module (or another resource) for a data set that can be used for your project
- ☐ **Plan Visualizations:** Based on the types of data you collected, what sorts of visualizations make sense? What pieces of the data relate to your research question, and how can you represent them? Write some ideas here:

Data to Represent	Possible Visualizations

- ☐ **Plan Models:** Determine what kinds of modeling can help you tell your data story.

Possible Predictor(s)	Possible Prediction	Appropriate Model

- ☐ **Create Visualizations:** Make sure they are accurate, clear and clearly labeled, and presented in a way that meets the goals you've set above.
- ☐ **Build Models:** If a predictive or descriptive model makes sense for your data question, build and analyze the model. If a predictive or descriptive model would not make sense with your data question, explain why here:

- ☐ **Answer Your Question:** In a write-up, draw connections between your data and multiple representations of your data, and how they answer your research question. Make sure your findings are clear and directly related to the research question. Make sure your final argument is clear.
- ☐ **Communicate Your Findings:** Create an artifact to communicate your findings with the general public. You should use visualizations created with CODAP, but you may choose to add to the overall infographic using sites like [Canva](https://www.canva.com/).
- ☐ **Reflect On the Data:** Consider your findings and how they relate to the real world. Share your reflection through a solution or call to action, an infographic, or a reflective portion of your write-up.

### 3. Share Your Work

Choose the best option from the choices below to share your work with the wider world.

- **Advocate for Change:** Present your solution to a community member who would be able to implement the changes you've outlined. This may be a teacher or administrator, a member of a local community group, a local government official, or anyone else who would be interested. In your reflection (which may be a write up, a video, a conversation, or another method), include a discussion of the issue you chose. If appropriate, explain what changes the community should make to address the topic you've found.
- **Communicate Online:** Use your project to engage and educate via an online platform. If you created an infographic or video, you could share on your personal social media account, or ask to share on your school's social media account. In either case, ask an engaging question and keep track of how people respond to your work.
- **Give a Community Presentation / Lecture:** Prepare an informational presentation for members of your community about your project. Advertise your lecture to any groups that may be interested in your topic.
- **Communicate Offline:** If your community has a public posting board, create a one page summary of your findings to post on the community board. Depending on your project, it could be purely informational, encourage personal change in the members of your community, or advertise another event that shares your project. If you feel comfortable, you can add contact slips to the bottom of your flier for people who want to know more.
- **Create a Fundraiser:** If it makes sense with your project, create a fundraising event to donate to a local charity or group related to your project. This could be a small event, like an online fund, or a bigger event, like a benefit concert.

## Assessment

	<i>Exemplary</i>	<i>Proficient</i>	<i>Developing</i>
<b>Question Formulation &amp; Project Design</b>	<p>The research question is one that can be thoroughly answered with data</p> <p>Research question is relevant with real-world applications</p> <p>Question is clearly communicated with background information</p>	<p>Research question is well communicated but cannot be properly answered with data science</p> <p>OR</p> <p>Research question is well communicated and can be answered, but is irrelevant to the real world</p> <p>OR</p> <p>Question is not clearly communicated, but is relevant and can be answered with data.</p>	<p>The question is communicated</p>
<b>Data Acquisition &amp; Preparation</b>	<p>Substantial data is selected from a reputable source</p> <p>Selected data corresponds with the research question</p> <p>You have vetted the data set to avoid any "glaring" issues.</p>	<p>Appropriate data is selected, but the data set is not large enough to reliably answer your research question</p> <p>OR</p> <p>Appropriate data is selected, but the data is not taken from a reputable source or has "glaring" issues/</p> <p>OR</p> <p>The data selected is reliable and substantial, but does not relate to the research question.</p>	<p>Data is acquired.</p>
<b>Visualizations</b>	<p>Multiple visualizations communicate project findings</p> <p>Visualizations are clear, concise, and well explained</p> <p>Visualizations are appropriate for the data</p>	<p>Exactly one visualization communicates project findings</p> <p>OR</p> <p>One or more visualizations are unclear or poorly labeled, but are present and appropriate</p> <p>OR</p> <p>Choice of one or more visualizations are invalid for the data being represented, but multiple visualizations demonstrate findings and are clear</p>	<p>Visualizations are missing or are invalid.</p>

<b>Models</b>	<p>An accurate, predictive or descriptive mathematical model is created to help answer the research question.</p> <p>The model is accurately interpreted in the write up</p>	<p>A mathematical model is created with small errors</p> <p>OR</p> <p>A mathematical model is created, but cannot be used to answer the research equation</p> <p>OR</p> <p>The model is inaccurately or not interpreted in the write up</p>	<p>A mathematical model with substantial errors in accuracy, applicability, and explanation is created.</p>
<b>Communication</b>	<p>Write up successfully communicates the question with background information, data collection techniques and decisions, data cleaning techniques and decisions, modeling and visualization decisions, model limitations, and findings.</p> <p>Final aesthetic deliverable successfully communicates the question and the findings.</p>	<p>Two or more pieces of the write-up (communicate the question with background information, describe: data collection techniques and decisions, data cleaning techniques and decisions, modeling and visualization decisions, model limitations, and findings) are missing</p> <p>OR</p> <p>More than half of of the pieces of the write up are unclear</p>	<p>Writeup is unclear or is missing a significant amount of essential information</p>
<b>General Project Feedback</b>			

## Template: A Write-Up

[Title of Write-Up]

[Subtitle]

**The Title:** The most important thing about your title is that it communicates what the paper is about. Be creative! If you have a creative title that does not fully communicate the topic of the paper, add a subtitle.

### Research Question and Background

Here, clearly state your research question. Make sure you take time before you start to [develop a strong research question](#). Briefly explain any context that is necessary for understanding your research question. Be sure to explain *why* your question is important, why should the reader care about your question?

Then, provide some background research on the question.

**Tone:** The tone of your paper should be relatively scientific. Avoid "talking to your reader" ("I bet you are wondering..."). However, it is not necessarily bad to describe your personal interest.

### Data and Data Collection

First, describe where and how you collected / found your data. Then, describe the data itself— for example say how many entries there are, or how many questions were asked. If you had to do any data cleaning, describe the decisions you made and why you made those decisions.

### Data Analysis

Now you get to share your findings! This section is where you put your well-labeled visualizations and models. In text, make connections between the visualizations and the models. Briefly discuss what each visualization or model shows. Make sure any visualizations are referenced by number and labeled. For example, in Figure 1 you see a brief description of what is being shown.

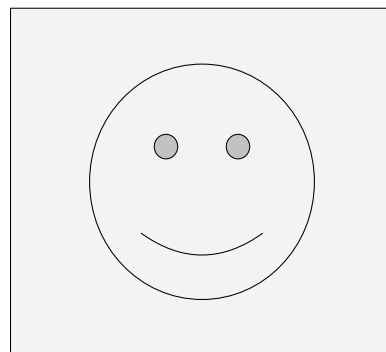


Figure 1. A smiley face

### Findings

Now that you have run your data analysis, answer your research question. In answering your research question, directly reference research that you conducted and trends/patterns shown in the visualizations and models you created. Make sure your question is clearly answered.

### Conclusion

Here, discuss any problems in your analysis. For example, how might different data cleaning decisions have affected your findings? Or different data collection techniques?

Then, discuss any questions that arise from your analysis. Maybe there is a part of your research question you were not able to fully answer. Maybe there is an interesting follow-up question you thought of while conducting your data analysis.

This is also where you can give recommendations on how to enact positive change based on the findings in your data analysis.

# Unions in the United State Example Project



This project is a basic exploratory data analysis in CODAP

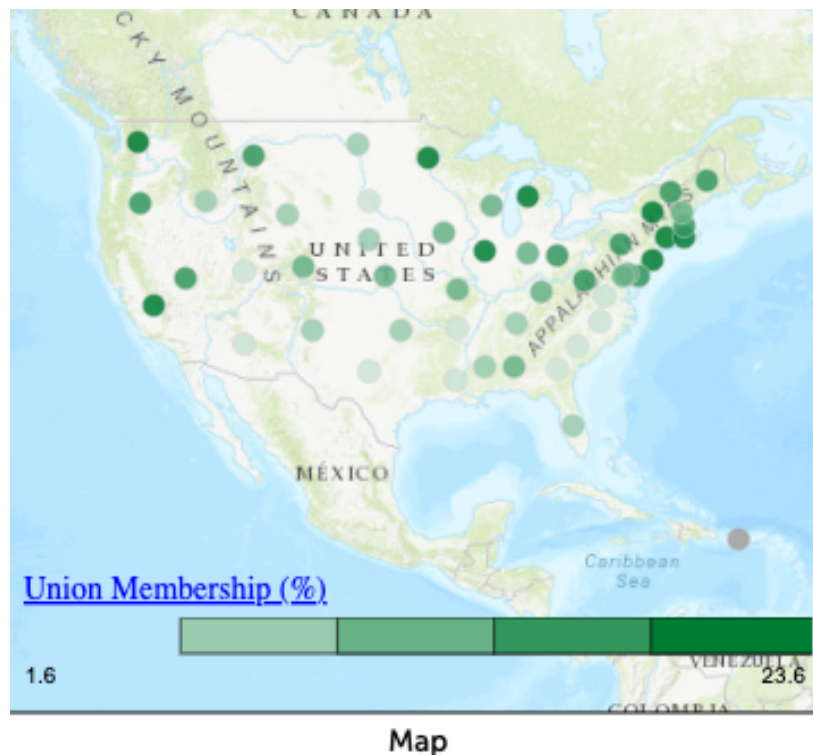
## Classroom Highlights

This exemplar demonstrates:

1. Basic visualizations (scatterplot, histogram, box-and-whisker),
2. Modeling with linear regression,
3. An example of an "emerging visualization" (a geographic map)
4. An example of "follow-up thinking",
5. The use of [Measure of America](#) as a source of data,
6. A null result
7. A controversial topic

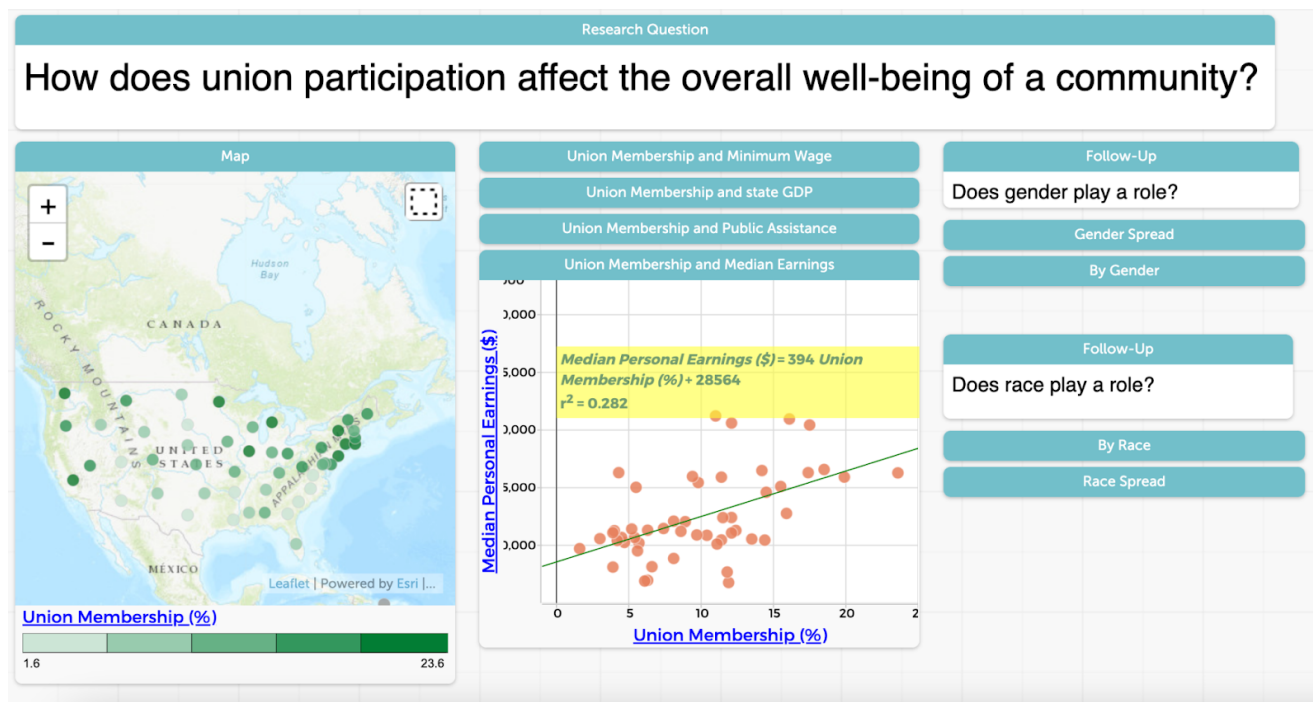
## The Project

In this project, I explore labor unions using data from the wage data of [Measure of America](#). I explored the relationships between a state's percent of union participation and other measures of well-being. To start, I visualized unions in the country using an *emerging visualization*, a geographic map.



[Measure of America](#) has a *lot* of data. An advantage of using technology like [CODAP](#) is that I could look at a lot of different measures fairly quickly and easily. So, a student would be able to start from a place where they know nothing about the topic. With some extra exploration facilitated by technology, they can narrow down their questions as they go.

I found that, at an initial exploratory analysis phase, weak relationships ( $r^2$  between 0.2 and 0.4) exist between union participation and four measures: minimum wage, state GDP, public assistance expenditure, and median earnings.



At this point, I practiced *follow-up thinking*. I explored two follow up questions: the role of race and the role of gender in the relationship between unions and median earnings. When broken down by either variable, the  $r^2$  values fall below 0.2.

In my write up, I attempted to reconcile the lack of concrete findings in this report with the overwhelming research pointing to a relationship between unions and well-being. This required follow-up research, as well as multiple forms of questioning:

- Is there something flawed in my data that leads me to find a result different from what the research says?
- Is there something flawed in my data analysis that leads me to find a result different from what the research says?
- Is there another factor that can be affecting my findings?
- Is the question one that can be fully answered with Data Science?



Answering "no" to any of those questions doesn't make this a bad project, it just teaches the process of data analysis and encourages students to think critically. Whether the answer to those questions are yes or no, I always encourage students to ask questions based on their results and consider where differences may come from, both on their own and formally in a write-up.

You can see the whole CODAP document [here](#). You can check out the whole report [below](#).

## A Null/Weak Result

Many students will choose to explore relationships that do not actually exist (sometimes, this will be painfully clear to you before your student even starts). This example helps to communicate to students that weak relationships still provide insight and inquiry, and that they should not be afraid to investigate ideas they do not already know about.

## Controversial Issues & Facilitation Tips

Officially, unions are illegal in Virginia. So, the topic of unions is very political and relatively controversial. Should I let my students explore a controversial topic like this? *Absolutely*. I encourage students to study anything they are interested in, and in my experience, they are often interested in political/controversial ideas (racism, environmental issues, LGBTQ+ issues, ...). This requires some closer facilitation and some more discussion, but is absolutely possible and, personally, encouraged.

Many data science teachers (many of whom were formerly, or still are, math teachers) have only had to discuss and facilitate conversations on controversial issues in informal ways that are not necessarily related to their curriculum. In my math classrooms, I have considered controversial issues of identity when I think about access to resources to succeed in my class, or when I overhear students discussing something (likely off-task) that I feel needs to be addressed. It has never been difficult, however, to keep politics out of the Mathematics classroom, and so many data science teachers might be used to leaving controversial topics at the door.

However, the idea of discussing controversial topics and facilitating discovery of deeply personal topics, like ethics, is not a new thing. English teachers teach books like *A Scarlet Letter* and *Huckleberry Finn* as discuss misogyny and racism, History teachers teach about the horrors of slavery, and Government classes regularly debate hot political topics as they discuss the American political system. My partner, a history teacher, regularly considers the ethical and political implications of every aspect of his classroom. He considers scope and sequence ("Would teaching about the Mayans earlier in the year help students understand the gravity of the Spanish colonial takeover?"), specific concepts ("How much can I discuss Selma riots with my middle schoolers?"), and even the movies he shows his class ("Should I show a historically accurate and enlightening movie, even though it uses offensive or potentially alienating language?"). Many

classroom discussions and student projects touch on issues of race, gender, immigration, colonization, and more.

Because of this, there are lots of resources to help teachers facilitate controversial topics in the classroom. I highly recommend taking a look at some of them, and having a conversation on the topic with your colleagues in the humanities. When I asked my partner how he learned to facilitate controversial topics, he answered that he learned from the stories of more experienced teachers around him. Here are some resources to get you started:

- [Common Sense](#): Discussing Controversial Issues in the Classroom
- [PBL Works](#): Exploring Controversial Issues in Project Based Learning
- [Edutopia](#): Social Justice Projects in the Classroom

There is a lack, however, of discussion of how to approach controversial issues in the data science classroom. For this, I will speak on my experience as a Data Science teacher and knowledge as a Data Scientist.

First and foremost, the part of Data Science that people think of (making graphs and models, calculating statistics), in itself is unbiased. After conducting a data analysis, the data says what it says. The scientist can present their analysis in an unbiased way.

That is *not* to say that Data Science as a whole is unbiased. Bias arises in a few steps of the data cycle. Many people think that the whole cycle of Data Science is unbiased. "Show me the numbers" is often used as an end-all in political discussion. This is far from true.

To help educate good Data Scientists, teach students to ask questions along the way:

- **In developing a research question:** what questions can be explored? What questions are worth exploring? What associations will I explore, and why do I think those associations might exist?
- **In collecting data:** who am I surveying? How am I asking my questions? What, exactly, am I using for measurement? Or, who am I getting the data from? (Garbage in, Garbage Out)
- **In cleaning data:** what mistakes in data input are enough to throw out a row? (For example, if someone leaves "race" unanswered, should I drop that item, or would that be leaving out a specific group of individuals?)
- **In determining what visualizations to create:** what associations should I look for? What visualization best answers my research question? Could this visualization be somehow misleading?

These questions and considerations are by no means exhaustive.

To facilitate good data science education, it is important to get your students thinking about those questions from the start. Once your student chooses to conduct their project on a controversial issue:

1. **Don't Discourage Them!** It is important that students study what they are interested in, and we do not want to communicate that controversial issues cannot be helped with good Data Science.
2. **Make sure the student is aware that the issue is controversial.** Sometimes they just don't know! That should not discourage them to do the analysis, but should prepare them for scrutiny and get them thinking about the questions above.
3. **Stress the importance of background research.** Make sure the student conducts research before choosing exactly what to measure and what associations to explore.
4. **Stress the importance of the write up.** Make sure that, in their write up, the student discusses research before and after the analysis. Make sure that in the write up the student addresses all data collection, data cleaning, and data representation decisions they made along the way.
5. **Check in with the student along the way.** Of course, you should be doing this with all of your students, but it might be a good idea to check in with a controversial topic a bit more frequently.
6. **Be ready to defend yourself and your student.** Especially if the project is shown to the general public, people may become unhappy. Be prepared to defend your student and yourself. In the end, this is a learning experience, and you are teaching students to address important issues with a curious mind.

## Off-Task or Critical Thinking?

My data questioning led me to an interesting side question: what is the spread of earnings by race and gender? [CODAP](#) allowed me to briefly explore that idea.



While I wouldn't include it in the final analysis, asking questions and searching for answers is something that we want to encourage in our students. When my students are working on projects, I keep this in mind when I am circling around. In a 90 minute class period of work time, I need to be prepared for students to spend some time academically off-task (asking side questions) and some time non-academically off-task (brain breaks). Part of project-based learning is teaching students the benefits of this off-task time, but helping them to balance.

# Unions in America Project Write-Up

*An analysis of the effects Unions have on the lives of workers in the United States*

## Research Question and Background

Labor Unions in the United States have a long history of advocating for workers. They have [contributed to fair income and safe working conditions](#) for Americans in a wide range of industries. On top of these benefits, they have [fought against income inequality](#), [increased job security](#). The history of labor unions, however, has been a source of contention in the United States for many years. In 2022, according to Gallup, [about 70% of Americans approve of labor unions](#). This approval rate has fluctuated between a low of 48% approval in the early 2000s and highs of about 70% in the 1930s, 1950s, and today.

This debate has led to labor union laws that vary from state-to-state. At a high level, some states are considered ["right to work states"](#), meaning that laws do not allow labor unions to be required for a given job. Today, Virginia is among the right to work states. [Labor unions oppose these right-to-work laws](#), arguing that they need as much support as possible to meaningfully support workers. There is evidence to show that these right-to-work laws do, in fact, have an effect on the success of labor unions. According to the [AFL-CIO](#), average workers in states without right-to-work laws earn over \$6,000 more than workers in states with right-to-work laws.

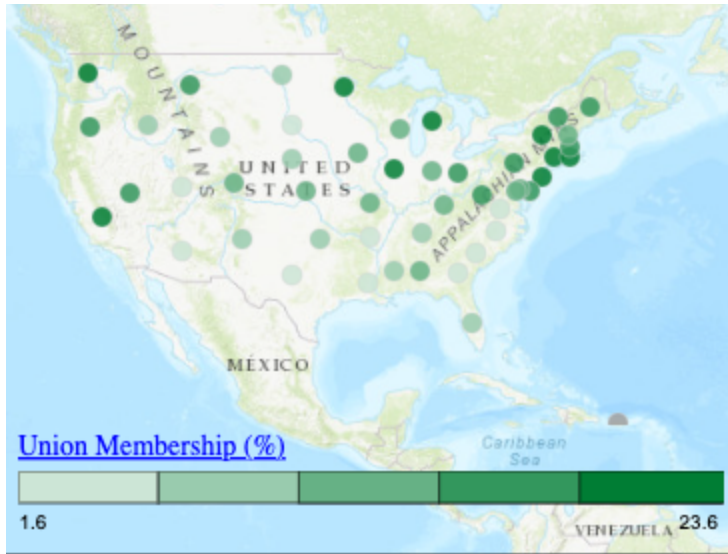
In this ongoing debate, it is essential to gain as much understanding as possible about the effects that unions have on workers in the United States, as well as on the economy and industry as a whole. In this project, I will explore these effects. Using data from Measure of America, I will compare union membership percent to various measures of economic prosperity including individual measures, such as the average employee's salary, and also wider measures, like state GDP. This research can help in informing the debate on worker's rights and unions in order to best support American workers.

## Data and Data Collection

Data for this analysis came from [Measure of America](#), who collects data related to the United Nations Sustainable Development Goals. The attribute I was most interested in was Union Percent, which describes the percentage of workers in the state who are members of a labor union. Since Virginia has laws limiting unions, I used the data at the state level in order to compare with places with very high labor union participation. I compared unions to most of the economic measures in the dataset, largely revolving around employment, poverty, and income.

## Data Analysis

Before diving into my analysis, I wanted to get an idea of where, geographically, unions were the most and least present in the United States. My results are shown in Figure 1. You can see that union membership is high in New England and the North East, and much lower in the Deep South and Southwest.



Map

Figure 1. Union Membership in the United States

I ran my analysis at the state level. After testing a number of different combinations, I found relationships between union membership and four different statewide measures: minimum wage, median earnings, state GDP, and public assistance (listed in order of relationship strength). These relationships are shown in Figures 2(a) - 2(d). R-squared values and linear regression models are listed in Table 1.

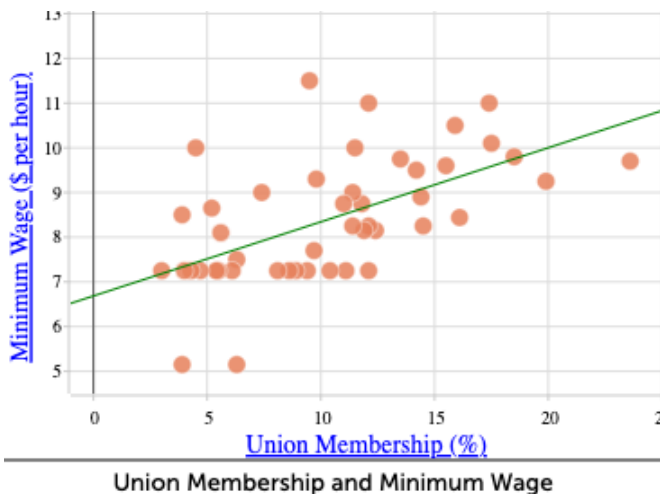


Figure 2a. Minimum Wage

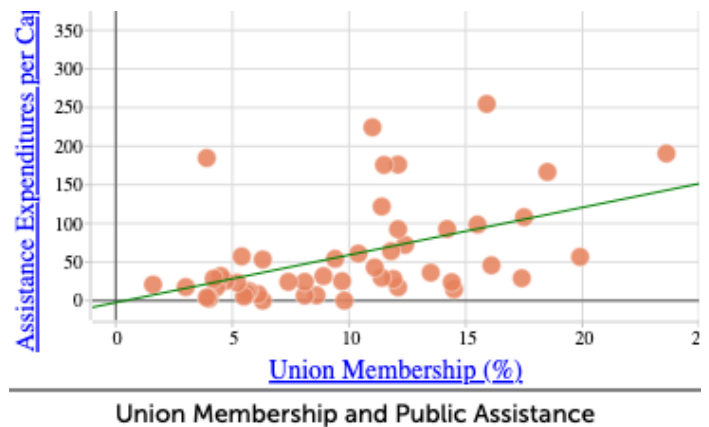


Figure 2b. Public Assistance



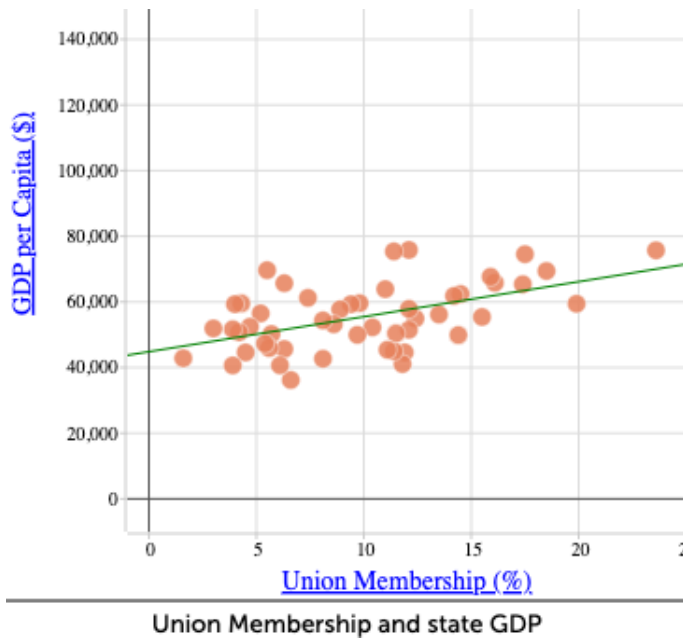


Figure 2c. State GDP

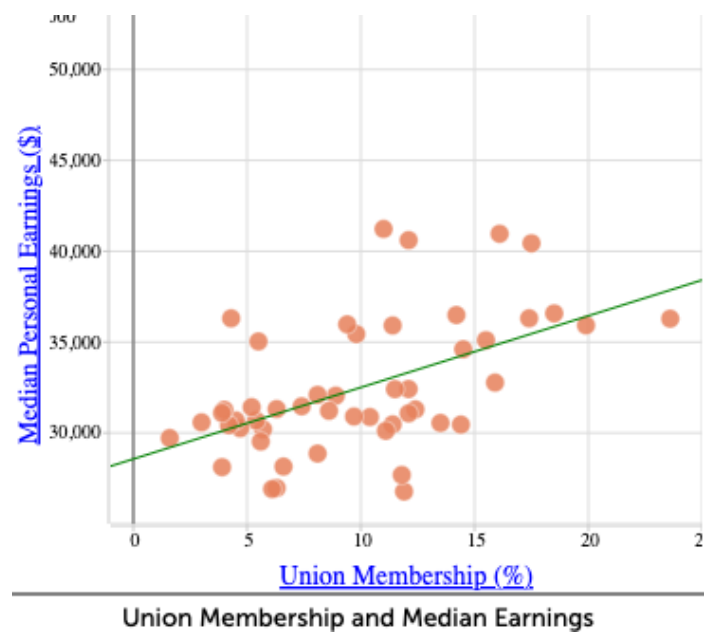


Figure 2d. Median Earnings

Measure	Model	Strength (R-squared)
Minimum Wage	Minimum Wage = $0.166(\text{Union Membership}) + 6.7$	Moderate Positive (0.33)
Median Earnings	Median Earnings = $394(\text{Union Membership}) + 28,564$	Moderate Positive (0.282)
State GDP	State GDP = $1066(\text{Union Membership}) + 44,827$	Low Positive (0.275)
Public Assistance	Public Assistance = $6.16(\text{Union Membership}) - 2.5$	Low Positive (0.227)

Table 1. Models of Attributes Related to Union Membership

## Follow Up

After exploring basic relationships, I conducted a follow-up analysis to determine how race and gender interacted with the results, if they did. These results could tell us whether union membership has a disproportionate effect on a group of people. I completed an analysis for this using only median salary (since this is the only of the four attributes that can be broken individually). Figure 3(a) shows the spread of salaries by race, unrelated to union percent. This information can help us to inform our interpretation of Figure 3(b), which shows the relationship between union percent and median salary when separated by race.



Figure 3a. Median Salary Based on Race

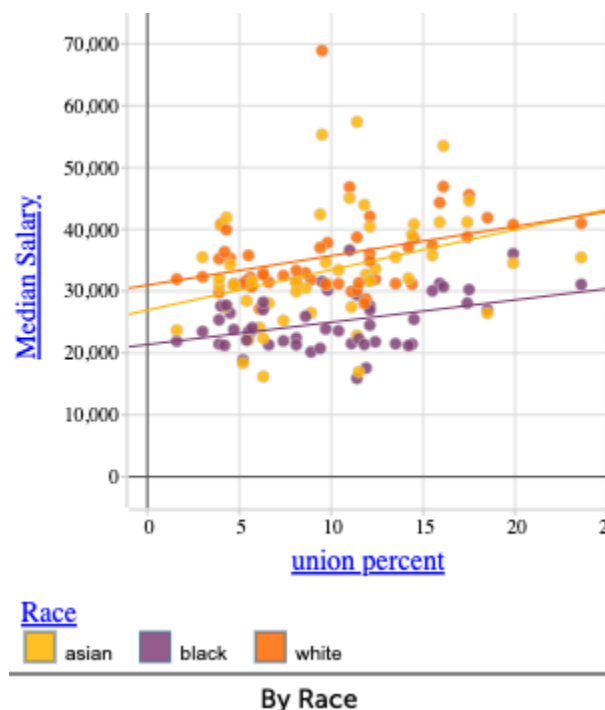


Figure 3b. Union Effect and Median Salary by Race

When separated by race, there was no relationship between union membership and median salary for any individual racial group (all r-squared values less than 0.13). I repeated this process broken down by gender. Figures 4a and 4b show the results.



Figure 4a. Median Salaries based on Gender

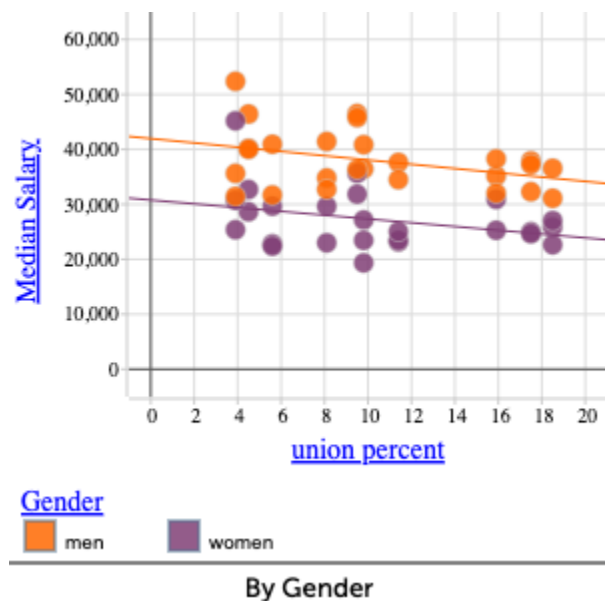


Figure 4b. Union Effect on Median Salary by Gender.

Both individual groups did not have a strong relationship (r-squared less than 0.15 for both).



## Findings

In my analysis, I explored unions in the United States. Union membership varies from state to state, roughly corresponding to [right to work laws](#), which dictate rules about joining unions. Using state-wide data, I compared union membership with a number of different wellness attributes, and found relationships between the percent of citizens in a union by state and the state's GDP, median salary, minimum wage, and public assistance.

The strongest relationship was between union membership and minimum wage. This is consistent with what we see historically, that minimum wage was originally [brought about by labor unions](#). In my data analysis, I found that states with higher percentages of union membership had higher minimum wages. Specifically, each additional percent of membership increases the state's minimum wage by about 17 cents. This is notable, a difference of only 6 percent of membership could mean a full dollar increase in minimum wage (in a 40 hour work week, 50 weeks of the year, this is an annual increase of \$2,000).

Union membership also had a moderate positive relationship with a state's median earnings. Given that union membership can mean an increase in minimum wage, it follows that median earnings would increase. Earnings may also increase for careers making above minimum wage. According to my model, each additional percent increase in union membership corresponds to an increase in median salary by about \$394. When membership increases substantially, for example by 5% or 10%, this indicates a noticeable difference in the salary of an average American, for example \$1,970 with a 5% increase, and \$3,940 with a 10% increase. Again, this makes sense when you consider the history and purpose of unions, [increasing worker pay is a common and central goal of labor unions](#).

Weaker, but still present, relationships were found between unions and state GDP and unions and public assistance. It makes sense that these relationships are weaker than minimum wage and median salary, since neither state GDP nor public assistance align with the goals of a typical labor union. However, my analysis does point to a possible positive relationship between the percent of workers and a labor union and state GDP or public assistance. This could be due to a direct relationship, or due to a confounding variable.

After exploring overall union membership by state, I was curious about discrepancies within a state. The only measure with a relationship with labor union participation that could vary between different groups within a state was median earnings, so I conducted a follow up investigation where I considered the effects broken down by race, and then by gender. Interestingly, when broken by either race or gender, the relationship weakens to non-existence. This is a difficult fact to interpret. At the very least, it indicates that the benefits of labor unions are consistent among workers, regardless of race or gender.

## Conclusion

In this project, I found relationships between union membership in a state and various measures of individual and statewide economic prosperity. Consistent with the research, I found that higher union membership results in higher wages for workers (in the form of a higher median salary and a higher minimum wage), as well as higher state GDP, and more public assistance. Interestingly, however, these patterns weakened substantially when the scope of the study was narrowed based on race or gender. Given that existing research usually shows a positive relationship between unions and economic success, this discrepancy points to the need for further study. We can conclude that union membership likely does not benefit any one group of people more than another group, and there is likely at least a small positive relationship between union membership and economic success for all.

An interesting further study would be to look more locally, for example comparing counties of one state with each other rather than full states. It would also be useful to further account for the legal intricacies of union membership. For example, it would be interesting to explore whether a highly enrolled optional union in a right-to-work state has as much, more, or less influence than a mandatory enrollment in a union state. It would also be interesting to incorporate more measures of job-related success, like job safety and job satisfaction. Overall, since the economy and the lives of American workers are such complex structures, more research could help to pinpoint the exact benefits of unions in terms of what exactly their benefit is, and to what extent that benefit reaches.

## Sharing Findings: A Letter to my Senator

Dear Virginia Senator,

I am a high school student in the state of Virginia studying Data Science. In school, we have learned the importance of exploring data and using our findings to make informed decisions. To practice these skills, I have conducted a research project on the effects of labor unions in the United States. I compared union membership percentages in different states to measures of economic prosperity like median income, minimum wage, and state GDP.

In my study, I have found some evidence that higher levels of union membership can be greatly beneficial to the average American worker. I believe that it is important, as a democratic government, to do everything we can to support the average American. Therefore, I think that more should be done to help support labor unions in their goals of helping workers. For your reference, I have attached my research to this letter.

Through my background research, I have learned that Virginia is a right-to-work state. While I certainly see the argument to not require union membership, to allow for more individual decision making, I also see that labor unions in states without right to work laws have a lot more influence and can make a lot more change than in right-to-work states like Virginia. I understand that I am by no means an expert in economics or workers' rights in Virginia. However, I do believe, based on the data, that further supporting labor unions can be beneficial to Virginians. I would like to request that you please consider the power of labor unions this year as you serve and support the average worker in our state. I believe that with further research and debate, our state can come to an agreement that balances the views of our people with the demonstrated benefits of labor unions.

I appreciate your help and ask that you please send me a response about what you can do to help support labor unions in our state, and what we can do together to help enact positive change in Virginia. Thank you for your consideration.

Sincerely,

Sara Fergus

