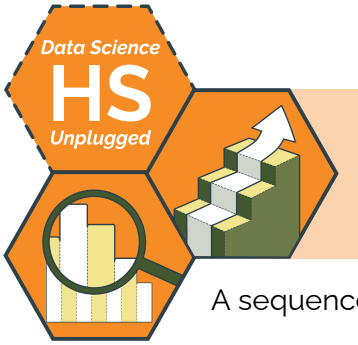


Data Science "Unplugged"

A sequence about *Unplugged* data science 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 low technology skills through a series of "unplugged" activities, where students engage in data science learning *without* using computational tools. 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 coding 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 manipulatives 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.

Data Science Standards Alignment

The chart below provides a summary of how the lessons in this sequence align to the Virginia Department of Education's Data Science standards of learning:

Lesson Name	***Data Science Standard*** *(see VDOE site 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 Finding & Collecting Data								✓		✓			
04 Preparing Data				✓				✓				✓	
05 Power of Visualizations						✓				✓			
06 Choosing Visualizations	✓					✓				✓			
07 Descriptive Statistics	✓									✓		✓	✓
08 Creating Visualizations	✓					✓				✓			
09 Creating Models									✓		✓		
10 Making Predictions									✓		✓	✓	✓
11 Overfitting & Noise							✓		✓				
12 Developing Questions	✓	✓											
13 Understanding Research			✓		✓								
14 Project Practice	✓	✓	✓			✓			✓				✓
14 Summative Project Frame	✓	✓			✓				✓				

All of the lessons in this sequence are **"unplugged"**, meaning that they do not involve coding or using a computational aid to calculate values, produce visualizations, or complete other data science analytical tasks. CodeVA's "plugged" data science sequences are on GoOpenVA: https://goopenva.org/profile/18746.

Materials

- Access to the Desmos mathematics education platform & the Kaggle data science web resource
- 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

Student Prerequisite Knowledge & Skills

This lesson sequence assumes a level of 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 3 decimal places
- 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 them to find 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.

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 the following pedagogical and content skills:

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 and modify curriculum to meet student needs
- Can guide students as they navigate open-ended, self-directed questions & learning experiences

Content Area Skills/Knowledge

- Download, upload, and manipulate CSVs and spreadsheet files (e.g., Excel, Google Sheets)
- 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

It is possible for a motivated educator to learn these prerequisite skills by studying the lesson plans and materials in this curriculum. 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).

Scope & Sequence

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

Lesson Name	*Summary*	*DS Standards*
01 What is Data?	This two-day lesson introduces students to different ways of expressing multivariate data, especially in non-computational formats.	DS. 6, DS.10
02 Types of Data	In this activity, students will learn different types of data, including quantitative, categorical, ordinal, and unstructured (i.e., qualitative) data.	DS. 1, DS.4, DS.6
03 Finding & Collecting Data	In this lesson, students are introduced to multiple methods of collecting and finding data and how to describe that data using its attributes.	DS.8, DS.10
04 Preparing Data	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.11
05 Power of Visualizations	In this lesson, students will explore the power of visualizations in making a point or communicating information about data.	DS.6, DS.10
06 Choosing Visualizations	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
07 Descriptive Statistics	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
08 Creating Visualizations	In this lesson, students will learn how to choose 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
09 Creating Models	In this lesson, students will start using a by eye technique to create models based on scatter plots. Then, students will categorize patterns to create predictive regression models based on data sets.	DS.9, DS.11
10 Making Predictions	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, DS.12, DS.13
11 Overfitting & Noise	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
12 Understanding Research	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.4
Summative Project	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:

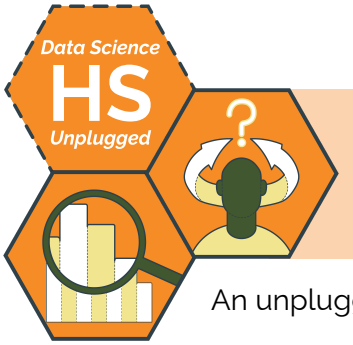
Lesson Name	*Summary*	*DS Standards*
14 Developing Questions	In this lesson, students will develop questions to answer with data. This activity is designed to provide a foundation for student-driven project-based learning experiences.	DS.1, DS.2
15 Project Practice	In this lesson, students will complete 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
16 Summative Project Frame	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.9

These activities are very open-ended, and educators have a lot of flexibility in how they facilitate them. Educators might have students complete the ***12 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 Unplugged Project Examples

To help educators envision what sorts of projects students might complete, this curriculum includes example projects, each with detailed documentation of the planning and execution processes. 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:

Project Name	*Summary*
Required Reading	Analyze the contents of the school library to investigate diversity
Lunar Myth	Analyze news data to assess the impact of the full moon



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 Eurpoeans. 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: Appears in CodeVA's Data Science with Python *&* Data Science with CODAP *sequences as* 01 Unplugged: What is Data?

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

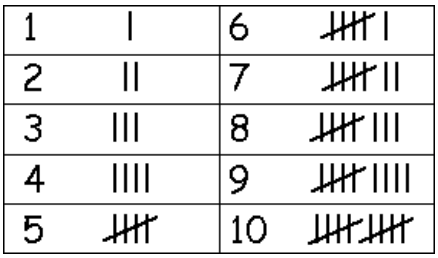
Standards Alignment

- **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	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. 
Matrix	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.
Table	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.
Quipu	A quipu is a recording device historically used by cultures in South America, including the Inca. The knots in the cords represented numeric values. 
Data Representation	A data representation is a way to visualize and organize collected information

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.

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

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 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?

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.

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

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)

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

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.

   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

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.

   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:

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.

	Proficiency	***Yes***	***No***	***Notes***
Completeness	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			
Representation	The representation that was selected was appropriate for the data collected. Data was accurately placed into representation.			
Summary	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").

	Proficiency	***Yes***	***No***	***Notes***
Concept	The concept of the postcard is correctly interpreted and explained			
Representation	Students accurately explain the meanings of *all* aspects of the representation, using the representation keys as a guide.			
Coherence	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

	Proficiency	***Yes***	***No***	***Notes***
Data	Student includes accurate data about their life			
Traditional Representation	Student produces a "traditional" data representation that is appropriate and accurate for their data.			
Creative Representation	Student creates a "non-traditional" representation to accurately represent a "data story" in their lives.			
Representation Utility	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.

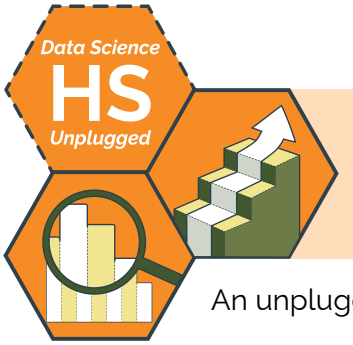
Provide a vocabulary sheet, like the one above for students who are learning English (suggested for WIDA levels 4 and below).

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

An unplugged introduction to data representations by Sara Fergus

Summary

In this 1-day activity, 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 03 Types of Data lesson plan from the CodeVA Python Data Science sequence, which is available here. The Python sequence includes a coding practice activity, while this one focuses on discussion.

Objectives

The students will be able to . . .

- Students will classify data as quantitative, categorical, ordinal or qualitative/unstructured.
- Students will generate data questions that can be answered given different types of data.
- Students will 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

Materials

- Google Survey for Warm-Up (view Google Form or make a copy)
- 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".

Vocabulary

Term	Definition
Data	"Data" is recorded information describing an event, person, place, or phenomenon
Quantitative Data	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 *Note: not all data using numbers is quantitative. For example "tv channel" or "ID number" would not be quantitative.*
Qualitative Data	Qualitative data is typically words and descriptions. These are used for open-ended questions. For example: "what was your favorite part of this week"? 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.
Ordinal Data	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… • Strongly disagree, disagree, agree, strongly agree • Poor, good, great • On a scale of 1 to 10, how much does the injury hurt?
Categorical Data	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". *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.*
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 *broad question* or a *subjective question*. Answering the question allows further questions to arise. Answering the question should contribute to a larger understanding of the world or an overarching question.

Outline

Formative Assessment Notes

1. **Warm-Up:** have students fill out this Google survey (be sure to make a copy so you can see the data), 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?

Students should notice, without vocabulary, that:

- **Qualitative data is displayed as text**
- **Categorical data uses pie charts**
- **Quantitative data uses histograms.**

2. **Sorting Data Types:** 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

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.

Recategorize: 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.

3. **Discussion.** Then, ask where students put "race". Many will have placed it in categorical. Discuss advantages, for example:

    - Identifying & addressing issues of descrimination by answering questions like "is race related to GPA", which reveals grade discrimination against non-white students
    - May reveal covert racism by revealing trends
    - etc.

    . . . as well as problems and limitations:

    - Leaving out people of mixed race
    - Not being able to include all racial/ethnic identities in the data.
    - etc.

    Note: Be sure to discuss this topic sensitively, paying special attention to questioning stereotypes and avoiding microaggressions toward marginalized students.

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. **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. Consider having students work in groups.

Stress that they are looking for patterns in questions for the exit ticket, not using specific vocabulary words for "types of questions"

6 **Exit Ticket:** Have students find patterns in data: what kinds of questions can be asked about categorical / ordinal / quantitative / qualitative data?

    Provide students with the *Types of Data Cheat Sheet.*

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:

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

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.

Python Extension: You can add in a Python coding extension by having students complete the workbook from *03 Types of Data* from the CodeVA *Python Data Science* sequence.

Types of Data Cheat Sheet

	Definition	***Example***	***Questions to Ask***	***Notes***
Quantitative / Numeric	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 *spread of the data*	Not all data using numbers is quantitative. For example "tv channel" or "ID number" would not be quantitative.
Ordinal	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... • Strongly disagree, disagree, agree, strongly agree • Poor, good, great • On a scale of 1 to 10, how much does the injury hurt? 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 *spread of the data*	
Categorical	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 *composition of the data*	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.
Qualitative	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 *patterns and descriptions in the data*	

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?

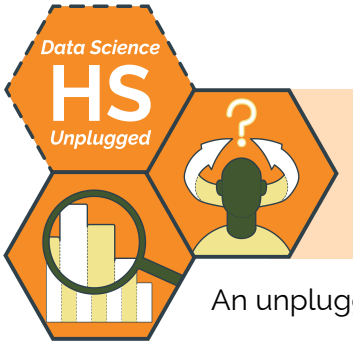
Name: _____ Date: _____

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?



Finding and Collecting Data

An unplugged introduction to acquiring, collecting, and preparing data by Christa VanOlst

Summary

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

Note: This lesson is similar to the 04 Finding & Collecting Data lesson plan from the CodeVA Data Science with CODAP sequence. This lesson includes CODAP activities, while the unplugged one focuses on "by hand" activities.

Objectives

The students will be able to . . .

- Find & acquire data from multiple online sources.
- Organize data points/elements in a structured data table.
- Define attributes that describe the collected data points/elements.

Standards Alignment

- **DS.8:** The student will be able to acquire and prepare big data sets for modeling and analysis.
- **DS.10:** The student will summarize and interpret data represented in conventional 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)
- 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 one or more attributes, Each row of a table is known as a case or 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.

Before the Lesson

This lesson requires a good amount of preparation and planning. On *Day 3*, you'll need to distribute a printout (the *Finding Data in Nature Activity Guide*, 1 per student) and take the class outside to a place where they can collect objects from the environment. A park or outdoor space on campus should work.

Day 1 Outline

Formative Assessment Notes

1. **Warm-Up 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 1900s employers used to keep time sheets.

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

2. **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 based on US Census data.

   Pose the question to students: *"Where do you think this data came from? How did they collect it? Who collected it?"*.

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

3. **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.

   In their journals have students categorize the following as a Data Table/Data 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

4. **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 (The HDI is a numeric summary of each state's average in life expectancy, education, and Income per Capita) surprises you the most?"

Students may need help navigating the site, consider projecting to the board and exploring on your own to demonstrate. Students should be looking. See *Assessment Strategies* below to assess student findings.

   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!

5. **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.

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

6. **Collecting Crowdsourced Data:**

   Facilitate the Crowdsourcing a Data Table activity, where students answer a few questions on a cut out sliver of paper and then place their answers together as a class on the board to create a data table and demonstrate crowdsourcing in real time.

   Once the data table is completed, 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.

Set a "checkpoint" where students should check in and receive feedback. Look for descriptive names, accurate results in calculation, if applicable, and creativity in portraying their findings. Provide feedback before moving on!

7.. **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.

   Have students find a data set and save it in their journal:

   - Digital Journal - Students should download the dataset, save it to their google drive, and upload a shared link into their journal.
   - Physical Journal - log the path where their data set can be found

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

Make sure everyone has logged at least one data set for future use in upcoming lessons.

Day 3 Outline

Formative Assessment Notes

8. **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. *Feel free to add as many attributes that the class can brainstorm.*

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

9. **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).

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

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

See *Assessment Strategies* below.

Assessment Strategies

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

Conclusion Activity

Give students editing rights to the responses spreadsheet and have students respond to the following questions, or have them complete this google form and show students the responses spreadsheet,

- *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
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/Modifications

In the *Collecting Data in Nature* (see step #8) activity, a collection of artifacts could be presented to students instead of having them go outside to collect their own. Students could also choose to collect artifacts in the room instead of outside.

For ESL students 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:

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

Explore and find 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 yourself. 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, or whole class discussion.

Activity 2 Guide - Crowdsourcing a Data Table

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

Activity Instructions for *(Teacher use only)*:

1. Print the empty table on the following page out on paper and cut the paper into rows and give each student a random row. (Add more rows if needed)
2. Ask the following questions aloud and have students write their answers to the following questions in each cell from left to right:
    - Question 1: What grade are you in?
    - Question 2: What is your favorite subject?
    - Question 3: What is your favorite social media platform?
    - Question 4: What is your favorite genre of music?
    - Question 5: Do you have a job?
3. Once students have completed answering the questions, have students come together and tape all of their rows onto the board in the front of the room *(they could also lay them out on a table or desk)*.
4. Discuss with students how we just collected and organized our very first data set!
5. Pose the question 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 having descriptive attributes.*
6. Explain how the column headers of a table are called **attributes** of data points and how they can be useful in analyzation to find trends, patterns, or summaries of our data points/elements.
7. Have the students as a class name each attribute with a meaningful *descriptive* and *short* name.
8. Once the data table is complete, break the students into small groups and assign them each an attribute.
9. Have each group analyze their attributes for information that can be generalized about their classmates.
10. Using colored markers or pencils and a piece of construction paper, each group will then write a one to two sentence summary of the data analyzed in their attribute.
    - For example if a group was assigned the favorite job attribute their conclusion may be - *"We found that 65% of our classmates hold down jobs outside of school."*
    - Encourage students to be artistic when writing on the construction paper. They could add in any images or drawing they may find relevant to their attribute to make it aesthetically pleasing at a glance.
11. Once each group is finished, have all groups glue or tape their construction paper to one large sticky or poster paper - the end result will be an **About Us Collage** to summarize the lifestyles of our class.

1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					
11.					
12.					
13.					
14.					
15.					
16.					
17.					
18.					
19.					
20.					
21.					
22.					
23.					
24.					
25.					
26.					
27.					
28.					

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: _____

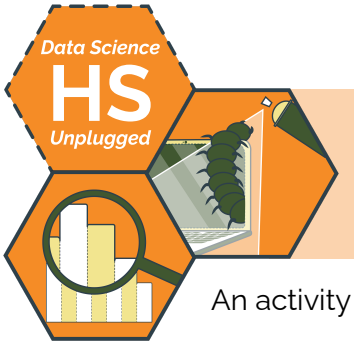
DIRECTIONS: Consider the image following images:

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 plan is similar to the Preparing Data *lessons from the CodeVA* Python Data Science *&* Data Science with CODAP *sequences. This lesson includes an "unplugged" data cleaning assessment, 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 (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) (PDF), print 1 per student

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 *not* at random if there *is* 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.

   See discussion essential understandings below for assessment information

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

   See discussion essential understandings below for assessment information

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

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

   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.

5. **Exit Ticket Cleaning Data Activity:** Give students this larger subset of the Roller Coaster (messy) data set and have students use a highlighter to markup any cleaning techniques learned that could be applied to this data set. Have students complete the exit ticket,

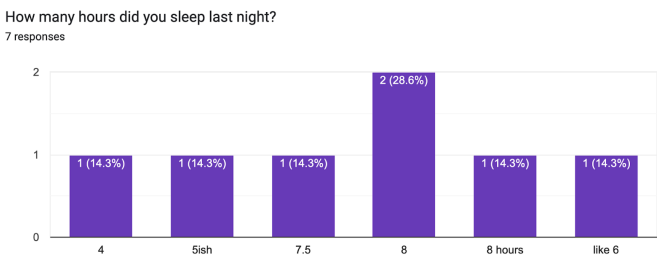
   Make sure that students are consistently considering what bias they may be introducing with their data cleaning. See Exit Ticket below.

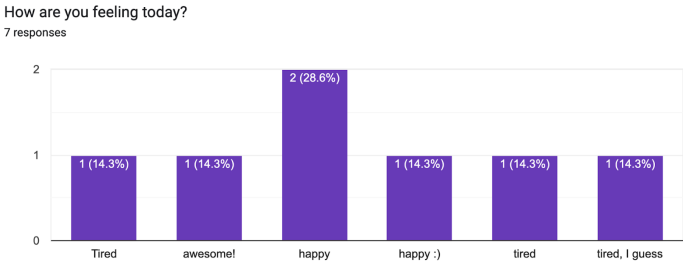
Assessment Strategies

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:





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: • Missing values • Different values that mean the same thing • Mismatched data types • Duplicate cases • Answers that don't make sense	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

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

Name	Score
John	230
Danny	240
Danny	240
Kyra	423

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.

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

Date	Temperature (F)
Jan 14	30
Jan 15	32
Jan 16	60
Jan 17	31
Jan 16	300
Jan 16	32

Definitely an error:

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.

   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

   b. 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

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.

5. *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.

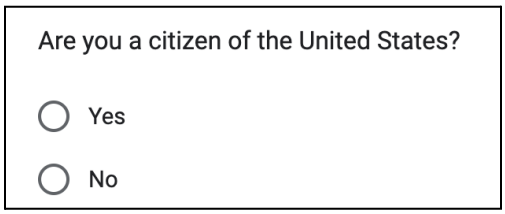
When we analyze missing values, we see two main types of missing data:

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.

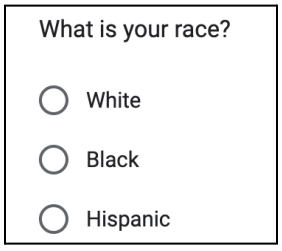
Type 2: Missing not at Random

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.



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.



<u>You can only remove cases for missing data if the data is missing at random</u>.

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

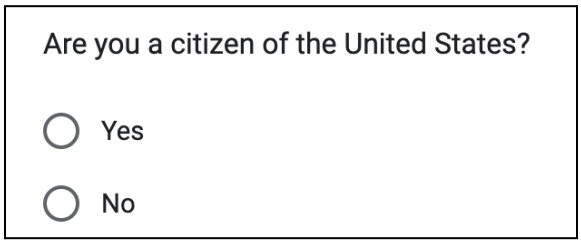
5. *Something else.* There are lots of methods out there!

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.



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?***

Unplugged: The Power of Visualizations

An introduction to interpreting visualizations by Christa VanOlst

Summary

In this lesson, students will explore the power of visualizations in making a point, supporting an argument, or communicating 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 create their own storybook to analyze data that was previously collected in prior lessons.

Note: A variation of this lesson is in the Data Science with Python & CODAP *sequences* as The Power of Visualizations (06 & 04).

Objectives

The students will be able to . . .

- Compare and contrast different visualizations of the same data set.
- Interpret different types of charts and diagrams used for data visualization.
- Discuss the impact visualizations have in supporting statements about the meaning of data.
- Create rough sketches of visualizations.

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 visualizations

Materials

- Craft supplies, including dot stickers, poster paper (large sticky), sticky notes, colored markers/pencils, scissors, rulers, colored string, yarn, white boards and markers
- *The Power of Visualizations* Slide Deck (direct link or make a copy)
- E-Cigarettes Line Graphs Data Talk (Desmos)

Vocabulary

Term	Definition
Data Representation	A data representation is a way to visualize and organize information
Visualization	A way to represent information in the form of a chart, diagram, picture, etc.

Day 1 Outline

Formative Assessment Notes

1. **Warm Up:** Give each student a sticky dot upon entry into the room

   Display or have students analyze the three images, and evaluate the claim below:

   - Visualization A - Cell Tower Data
   - Visualization B - Cell Tower Data
   - Visualization C - Cell Tower Data

   "At&T has the best cellular data coverage in the United States."

   Each student should place their dot in the table below in the row to represent their vote on which visualization supports the claim.

	Results for Vote
A	
B	
C	

   Then, have students use white boards to write about their choice.

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

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.

2. **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?***

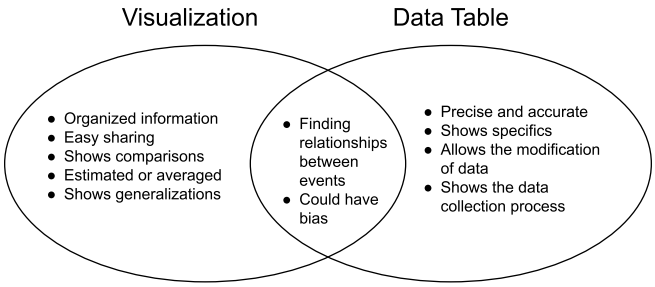
   During the activity, students work in groups to sort & analyze data. Have students justify why their ideas fit within the categories, and how the categories relate to or differ from one another.

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.

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

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.

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.)..

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

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?

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.

   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:

   a. 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.
   b. 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).
   c. 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

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.

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

10. **Mini-Project Creating Your Own Story:** Have students complete the Mini-Project: Creating Your Own Story.

    Summary: Students should create either a poster or a digital story using data of their choice.

Assess the students data visualization choice, accuracy, and design (see Assessment Strategies below for details & rubric)

11. Have students complete the *Exit Ticket* where they can practice interpreting visualizations.

See Assessment Strategies below for details & rubric

Assessment Strategies

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

Mini-Project: Creating a Story

This assessment can be completed in class time, or you could encourage students to complete outside of class time to present later in the week. The goal for this mini-project is to create a "story" using data of their choice (or, you could provide students with a particular dataset). Students should create either a poster or a digital story.

The story should include:

- A description of how the data was collected.
- An insight into how the data was held/organized.
- At least 3 visuals to show findings, trends, or patterns in their data.
- A coherent story.

Milestones:

1. Students should have data collected from the previous lesson and a data table representation of that data already completed.
2. Brainstorm the use of an effective visualization using appropriate chart type, accurate data points and effective design.
3. Fabricate a story where these visualizations will be useful when describing your objects at a glance.
4. Create a poster to tell your story.

Sample: Example Story using clovers collected from outside.

Mini-Project: Creating a Story Rubric

	Proficiency	***Yes***	***No***	***Notes***
Data	The students' choice of data attributes and organization is **mostly insightful** to the collected object itself.			
Narrative	The student created story connects to their collected data using a **mostly coherent** narrative.			
Visual	The student used **suitable visuals** that support their data **AND** used those visuals to drive the narrative.			

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.

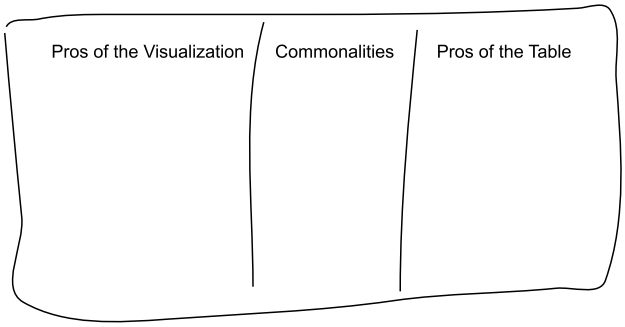
Extensions

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.
    - 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.
    - *What does the table accomplish better than the visualization?*
    - *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.
    - *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.
    - 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.
    - 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?

Unplugged: 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 also appears in CodeVA's Data Science with Python & Data Science with CODAP sequences with additions & extensions.

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 be able to summarize and interpret data represented in both conventional and emerging 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)
- *Visualizations in the News* Handout (see below or make a copy of the Google Doc version)

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!

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.

   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.

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.

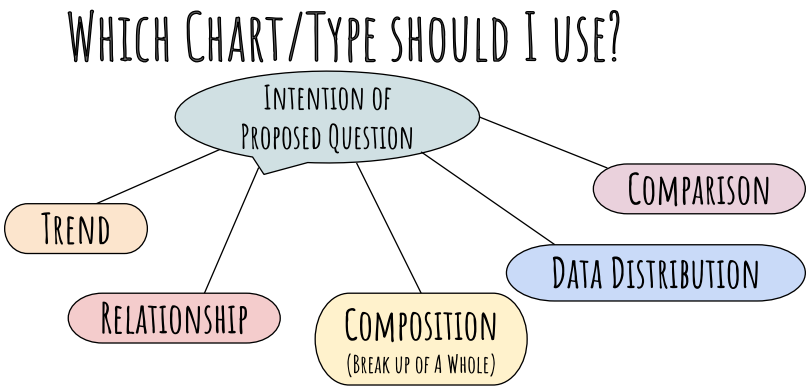
   Monitor student responses for the ability to categorize the use of visualization types.

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 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

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.

	Proficiency	***Yes***	***No***	***Notes***
Article Choice	Students chosen article is local on a state or community level **AND** the article presents a missed opportunity for the use of a visualization.			
Visualization Sketch	Visualization sketch depicts the data described in the article **AND** includes scaled axes.			
Visualization Style and Choice	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.)

Teacher Directions - Jigsaw Exploratory Analysis (Extension)

A matching activity for students to identifying the goal beneath a question and connect it to a visualization.

Consider using a jigsaw approach to conduct the following activity.

Give each group a topic and have them use half of a large sticky and colored marker/pencils to sketch a visualization to support the given question and a data set. Each visualization should be complete with a title, labeled and scale axis, and accurate data points.

Group 1: Using the Richmond Weather Data Set *(28 Data points)*

- **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.
- **Question:** I wonder what direction the wind tends to blow in my town.
  - Possible Outcome: Create a histogram

Group 2: Using the Most Followed Instagram Accounts Data Set *(50 Data points)*

- **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.

Group 3: Use the Dogs Data Set *(106 Data points)*

- **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.

Group 4: Use the US Cities Data Set & US Map *(Use a sample - total 1,000 data points)*

- **Question/Thought**: "The most populated cities in the US are on the borders."
  - Possible Outcome: Create a heat map overlaid on a map of the United States

Regroup after the exploration session to discuss/share possible outcomes.

- Once each group has created a possible visualization, have each group choose an expert to travel and show their visualization to the other groups. Other members of the group will question and give feedback to the experts of each group as they rotate throughout the room.

Steph Curry Is One of The Best (Data Sheet)

Records of his Shots during the 2015-2016 regular season

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

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

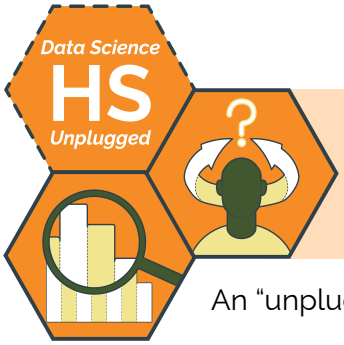
(copy and paste web page url here)

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



Calculating Basic Statistics

An "unplugged" introduction to descriptive statistics by Christa Van Olst & Sara Fergus

Summary

In this three 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* Data Science with CODAP *&* Data Science with Python *sequences. This lesson focuses on manual calculation, while the others include practice with Python & CODAP tooling.*

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 to analyze data.

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?
- M&M Activity Materials, including *Teacher Directions* (see below) & M&M Color Distribution Article
- *Philosopher's Chair* activity materials (see step #12), including *Teacher Directions* & *Statements*
- *Descriptive Statistics Checklist* (see below, 1 per student) & data sets (Salaries by College Type, Salaries by Region, Salaries by Degree, Starbucks Drinks, Starbucks Food)
- Day 2 *Exit Ticket* (see below), print 1 per student

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

Day 1 Outline

Formative Assessment Notes

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

   - ***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.

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.

2. **Generating Questions:** Have students individually access this Desmos: New York Times Data Talk. Use the Desmos to have students address the following:

   - 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?

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

For example, students may ask *"Why do older people make more money?"* You might respond: *"To start to answer that question, 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.

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

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.

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.

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.

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.

7. **Calculating Descriptive Statistics by Hand:** Place students into pairs. Have them collaboratively complete the Worksheet - Calculating Descriptive Statistics.

   Consider completing question 6 as a class, as some students may need direct instruction in order to calculate standard deviation.

8. **Exit Ticket:** Use the Data Talk: Interpreting Descriptive Statistics to facilitate a closing data talk using the same techniques as step 2.

Day 2 Outline

Formative Assessment Notes

9. **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.

10. **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.

11. **Statistics Practice:** Use the Teacher Directions - M&M Activity to have students investigate the accuracy of inferential statistics.

    Summary: In this activity, students calculate statistics then make a hypothesis using a small sample of M&Ms. As a class, students combine their results, recalculate, and assess the accuracy of their statements. After students read this article on the actual M&M Color Distribution throughout the years.

Use this as an opportunity to check with students regarding their calculation skills and their ability to interpret the statistical data

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 (view Google Sheet)
    - College Salaries by Region (view Google Sheet)
    - College Salaries-by-degree (view Google Sheet)
    - Starbucks Drinks Data (view Google Sheet)
    - Starbucks Food Data (view Google Sheet)

    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

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?

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

Extensions

Have students create deliverable visualization for the Activity 2 Checklist - Calculating Descriptive Statistics in Data Sets using the guidelines from lesson 6.

Predicting and Calculating Mean and Median Worksheet

Vocabulary	Definition	How to find
Mean	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.
Median	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	
Predict: 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	
Predict: 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	
Predict: 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.

Calculating Descriptive Statistics by Hand

An activity to recall the steps to calculate mean, median, mode, range, and standard deviation by hand.

Using the following small data set complete the following calculations: 32, 20, 24, 20, 25, 25, 36, 25, 32, 23, 28.

<table>
<tr><td>1. Identify the sample size.</td><td colspan="3"></td></tr>
<tr><td>2. Calculate the mean.</td><td colspan="3"></td></tr>
<tr><td>3. Calculate the median.</td><td colspan="3"></td></tr>
<tr><td>4. Calculate the mode.
How is the mode often displayed in visualizations?</td><td colspan="3"></td></tr>
<tr><td>5. Calculate the range.</td><td colspan="3"></td></tr>
<tr><td rowspan="2">6. Calculate the standard deviation.
(Standard deviation measures the variability of the data set. Like range, a smaller standard deviation indicates less variability in the data)</td><td>Identify the distance from mean (calculate for each value)</td><td>Variance
(average the distances found)</td><td>Standard Deviation
($\sqrt{}$ the variance)</td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

7. Suppose this dataset was collected by asking 11 people: *"What age did you move out on your own and get your first apartment/house?"*. Use the descriptive statistics you calculated and write a paragraph to summarize your findings using your answers from 1 - 6.

Teacher Directions - M&M Activity

How accurate are inferential statistics?

1. Distribute or have students each grab two random handfuls of m&ms
2. Have students arrange m&ms into a bar chart sorted by colors to complete step 3.
3. Have students identify their sample size and sketch the following table to complete

Color	Amount	Percent of Handful

4. Have students make conclusions about the **most common color in the entire bag**.
    - For example: *"Blue must be the color of most M&Ms because 58% of mine are blue."*
5. Have students come up to the board to create a cohesive data table of all student values values, for example:

Color	Amount	Percent of Handful	
Blue	5	32%	Student A Responses
Red	9	44%	
Brown	4	12%	
Green	7	25%	
Blue	3	11%	Student B Responses
Red	2	5%	
Orange	11	47%	

6. Identify the class sample size on the board and tally up the totals for each color, then have students compare their assumptions made.
7. Pose the following questions:
    - *What was the average amount of yellows in each handful?*
    - *What was the median percent for red M&Ms?*
    - *What would the population of this sample be?*
8. In conclusion, have students read the following article explore some data science that has already been studied on this topic - M&M Color Distribution Research.

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.	*This is misleading since it was a multiple select survey, dentists could have also recommended other brands before colgate.*
In 2021, a school summarized that 63% of students who are late to school have jobs.	*Are the students working during the week?*
In 1973, UC Berkeley's graduate school admitted 44% of male applicants and 35 % of the female applicants and was sued for discrimination.	*Consider sorting the data into subgroups and analyzing each department that students applied to then calculate these averages.*
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.	*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?*
The average depth of the Potomac River is 10'3 feet.	*This is misleading because some parts of the River get up to 33'5 feet deep. This could lead to a dangerous assumption.*
The average number of feet for a U.S. Senator is 1.98.	*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.*
The average temperature in Virginia is 39.8°F per year so it is not a vacationing state.	*Many other factors to consider [seasons, location, time of day, etc]*
In the middle ages the average life span was 40 years, so most people probably lived to see their hair turn white.	*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.*

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 data cycle from the beginning given only a data set.

- ☐ I can identify the sample size of this data set
  - Sample Size = _____
  - Describe the Sample: _____
- ☐ 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 calculate the mean of a data set
  - Mean = _____
  - What does this mean represent in the context of your data? _____
- ☐ I can calculate the median of a data set
  - Median = _____
  - What does this median represent in the context of your data? _____
- ☐ I can calculate the mode of a data set
  - mode = _____
  - What does this mode represent in the context of your data? _____
- ☐ 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 two different, relevant and meaningful visualization of my data
  - Sketch of visualizations below:

- ☐ I can identify outliers and skews in the data and communicate these findings
- ☐ I can summarize attributes by using statistics in writing

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: _____

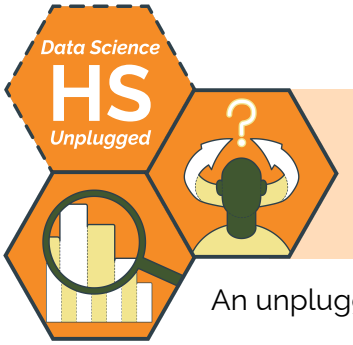
1. Brainstorm and explain what limitations would exist in society if all descriptive statistics were calculated by hand.

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 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 by hand. 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: Variations on this lesson appear in the Data Science with CODAP *&* Data Science with Python *sequences.*

Objectives

The students will be able to . . .

- Quickly sort and subsort 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 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
- Manipulatives, include raw fettuccine noodles, books, toothpicks and popsicle Sticks
- One of the following manipulatives: popcorn kernels, stickers, chocolate chips, round tokens
- Using Basic Visualizations activity materials (see step #2), including *Teacher Directions*, *Student Guide*, and *Basic Favorites* data
- Visualization Station materials, (see step #5) including *Teacher Directions*, *Peer Feedback Forms*, & *Station Guides* (1: Histograms, 2: Scatter Plots, 3: Heat Map, 4: Box Plot, 5: Pie Chart, 6: Combining Visualizations)
- Mini-Project materials, including *Directions & Student Planner* & *Reflection*
- Datasets (Dogs, US States, Summer Paralympics Multi-Medalists, Favorite Classes, Birth Stats, Internet Users, Prices & Speeds)

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.

On Day 1, be sure to review the *Use of Basic Visualizations* teacher directions (see step #2) and prepare to distribute the *Student Guides* for that activity according to the instructions.

On Day 2, review the *Stations Workshop* activity in detail (see step #5). There are six stations, and each has a printout and several other materials you'll need to prepare.

On Day 3, be sure to prepare the mini-project materials for the activity you or your students select (see step #6 and #7 below)

Day 1 Outline

Formative Assessment Notes

1. **Journal Entry and Advantages of Sorted Data:** Compare and Contrast the following data sets: Basic Favorites *(use the tabs on the bottom to see the separate sets)*

   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 relationship can you see about the students who like the color light blue?***
   4. ***What is the advantage of sorting data?***

   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.

2. **Using Basic Visualizations:** Use the Teacher Directions - The Use of Basic Visualizations while students use the Visualizations Guide.

   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 Google Sheet).

   - *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.

   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?***
   3. ***For example like the website does – Cinema: Explaining a movie plot through data visualization***
   4. ***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

Formative Assessment Notes

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:

   - 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?

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

5. **Stations: Creating Basic Data Visualizations:** Use the Teacher Directions - Workshop Stations to have students rotate to each station and create different types of visualizations.

   Have each student use the activity guides for each station to access data sets and directions to:

   - Create histogram, scatter, box, pie and heat map visualizations
   - Answer questions about each data set
   - Identify opportunities to explore deeper
   - Reflect on the data cycle

   This activity requires a good amount of setup, so be sure to prepare accordingly. You may want to have students "reset" the stations for the next class and/or keep the station materials on a surface that you can move around easily (e.g., poster board)

Consider collecting station 6 activity guides as a quick completion assignment.

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

Day 3 Outline

Formative Assessment Notes

6. **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 exploratory 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 then conclude their project with a driving question that arose from their results and identify any possible solutions or recommendations to address their issue.

   OR

   Mini-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 details

See *Assessment Strategies* below for Teacher Directions.

6. **Mini Project Reflection:** Have students complete the Reflection piece of the Mini Project

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 to complete the project. Collect and analyze student reflections for genuinity. Use the rubric below to assess student presentations.

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

Mini-Project: News Article

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

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 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

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

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.

Extensions

Have students explore the following site: Data Visualization Tips For More Effective And Engaging Design

Teacher Directions - The Use of Basic Visualizations

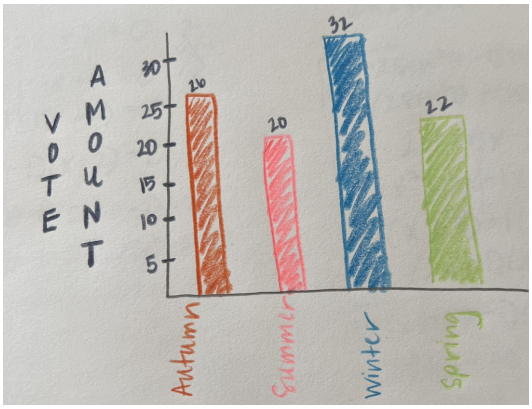
An activity for a class to explore how a basic visualization can transform into a creative freestyle representation.

Step 1: Bar Chart

1. Distribute the following: Basic Favorites Data Set
2. Have each student choose an attribute.
3. Have students use the Student Bar Chart Guide to quickly create a bar chart of their categorical data by hand.
    - *Seasons attribute Example Table*

Attribute Name: _____	
Outcome	Total
summer	*20*
spring	*22*
winter	*32*
autumn	*26*

Seasons attribute Example Bar Chart



Step 2: Collage

4. Have students choose a data representation style and create their collage. Consider the following Suggestions:
    - A google drawing using images from the internet
    - A paper/pencil one-sheet poster using colored utensils and drawings by hand
    - A construction paper one-sheet poster and clippings from magazines
    - A one-sheet paper using glue/tape and resources from outdoors or in the classroom
5. Give students 15 minutes to create their collage.
    - Example Collage 1 (Seasons Google Drawing)
    - Example Collage 2 (Seasons by Hand)
6. Once complete, have students clear their desks to display only their collage, their bar chart, and a blank peer feedback form their desk.
7. 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.

Student Basic Visualizations Guide

An activity guide for students to analyze data in order to create a bar chart.

Given the *Basic Favorites* data set, answer the following:

1. Choose an attribute to focus on (circle one): *Favorite Color - Favorite Season - Favorite Class*

2. In your attribute column, use the table below to keep track of each choice and their totals.

Attribute Name:	
Outcome	*Total*

3. Sketch a bar chart below of your results above.

Teacher Directions - Workshop Stations

A guide to creating basic visualizations by hand using manipulatives

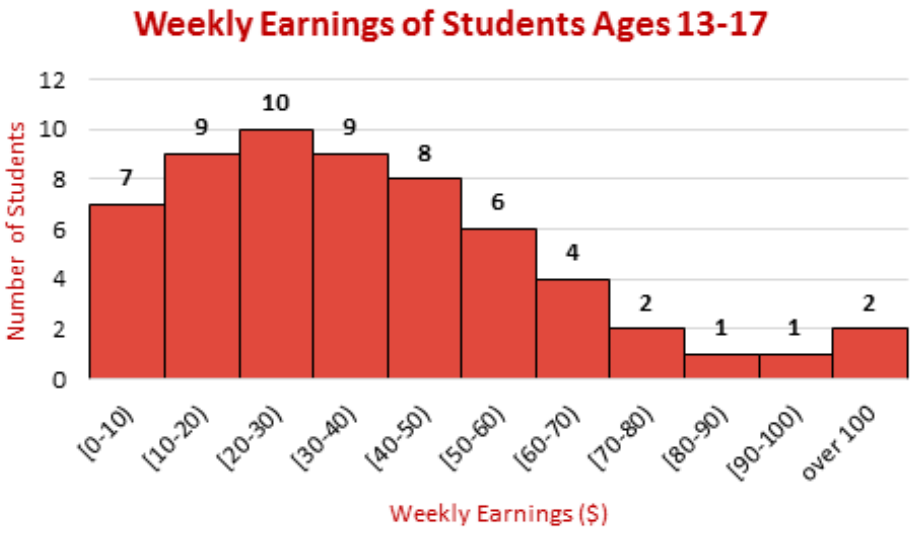
1. Arrange the desks into 5 stations and set up materials for each station (print class copies of each guide).
2. Create and assign 5 groups or randomly place students at stations 1-5
   - Materials for Station 6 will be distributed to each group once they have completed all stations.
3. Discuss/facilitate by walking and engaging with different groups or consider staying at one station that students may need extra help with, this means all students will eventually check in with you.

Data Set	Materials	Possible Outcome
Station 1 Histogram		
Dogs	Student Guide - What is a Histogram? Raw Wide Fettuccine Noodles - Books - Rulers - Construction Paper - Scissors - Glue	Max Life Span for Dog Breeds (Histogram)
Station 2 Scatter Plot		
US States	Student Guide (Scatter Plot) Glue & One of the following: Uncooked Popcorn, Kernels, Stickers, Mini Chocolate Chips, Round Tokens	US State Median Income vs. Percent of College Graduates
Station 3 Heat Map & Geospatial		
US States	Student Guide (Heat Map & Geospatial) Colored Pencils	US State Population Heat Map
Station 4 Box Plot		
Summer Paralympics Multi-medalists Results	Student Guide (Box Plot) Toothpicks - Popsicle Sticks - Markers/Colored Pencils	Comparative Box Plot for Paralympic Athlete Gold Medals
Station 5 Pie Charts		
Favorite Classes	Student Guide (Pie Charts) Scissors - Protractors	Pie Chart for Grade Level Pie Chart for Favorite Class
Station 6 Combining Visualizations		
Data Set	**Materials**	
Birth Stats Reduced	Student Guide (Combining Visualizations)	

Station 1 Guide - What is a Histogram?

A **histogram** is used to show the pattern or the distribution of numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".

Example

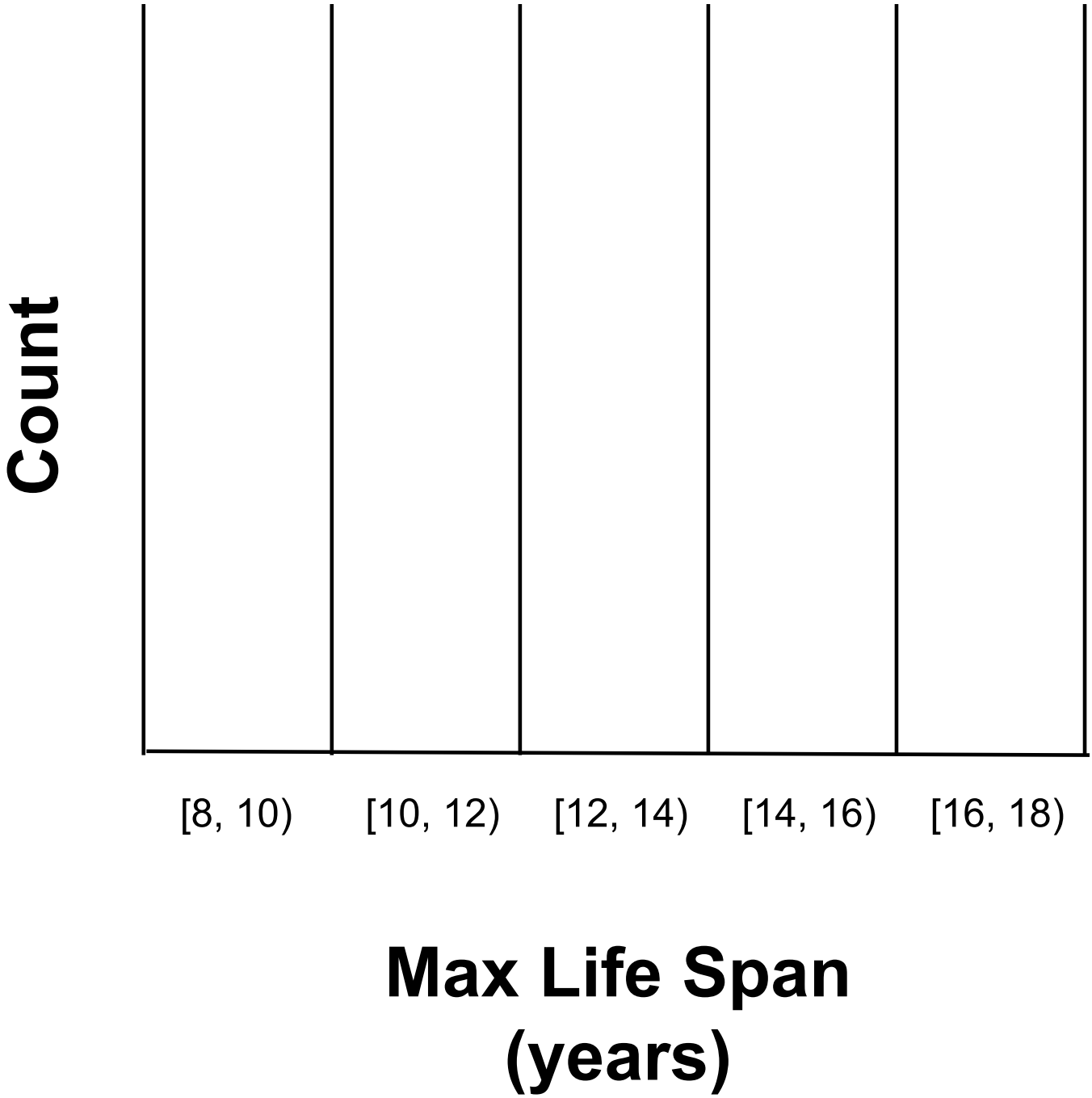


Given the histogram above, which bucket would you fall into? How do you earn your money?	

Create the histogram on the back of this paper, and then complete the Reflection questions below.

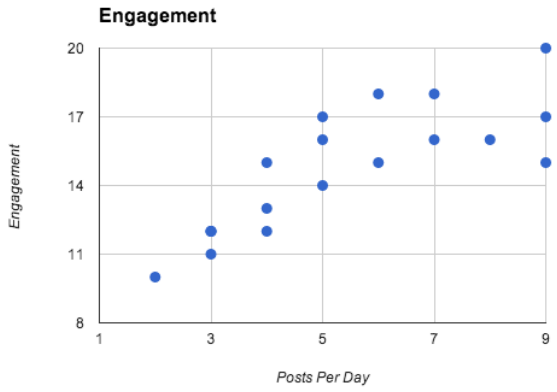
Reflection Questions (Answer after creating the histogram on the back)	
What range of values are most common for this attribute?	
What range of values are least common for this attribute?	
What ranges of values do or do not appear in this attribute and why?	
Exploring Further: Which other attributes could you test to create relevant histograms?	

DIRECTIONS: Make a histogram below for the maximum life span column of the Dogs data set. Choose one of the objects on the table to manipulate and represent the height of each bar to accurately represent the count for the data that falls within each bin.



Station 2 Guide - What is a Scatter plot?

A **scatter plot** is a visualization in which the values of two numeric variables are plotted along two axes. The pattern of the resulting points reveals any correlation present.

Example
Anicka is an analyst for Instagram. She created this scatter plot to test for a correlation between the amount of posts per day a user has and the engagement traffic of their profile per hour. 

Based on this scatterplot, a user with 6 posts a day should get approximately how much profile engagement per hour?	

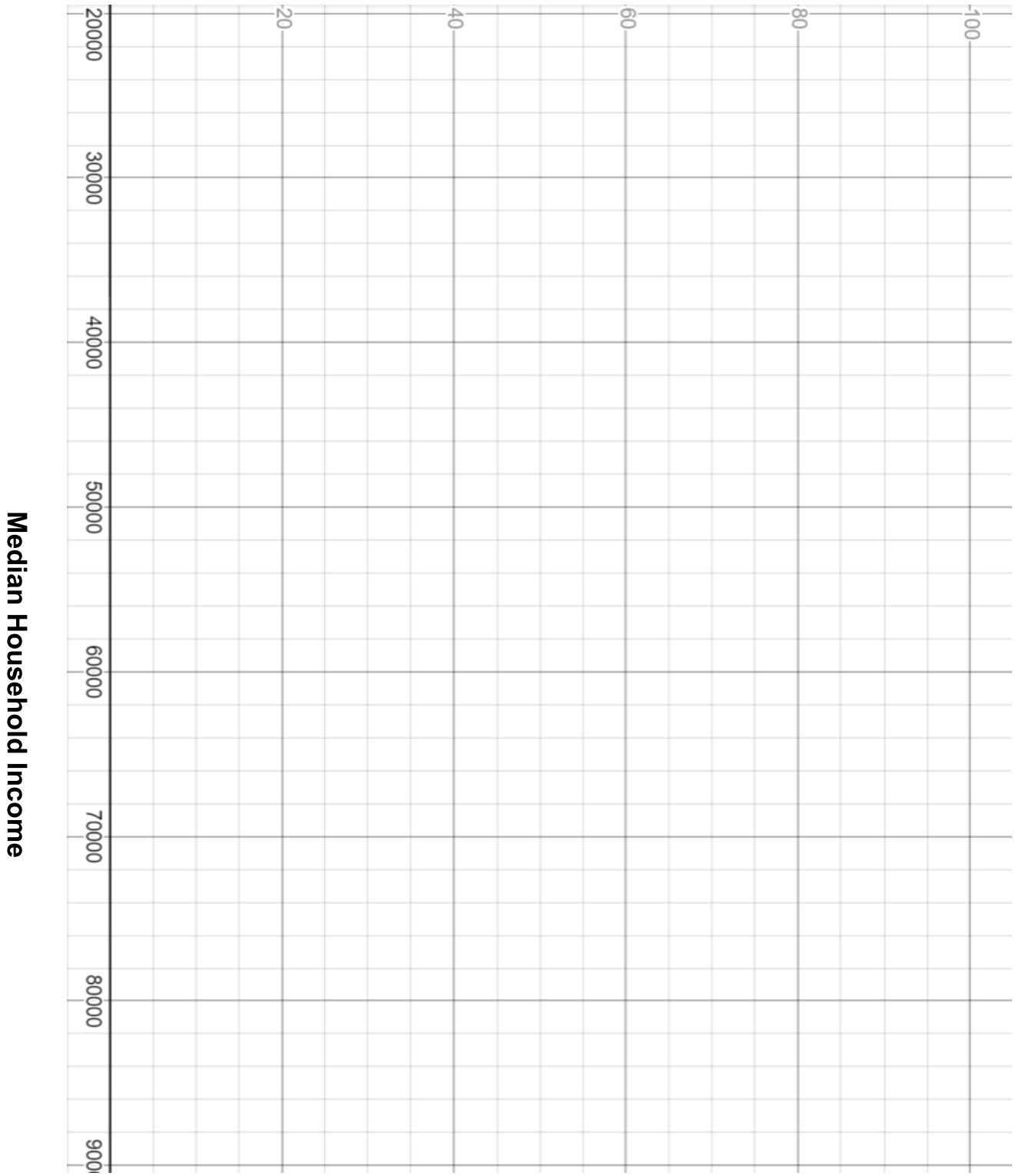
Create the scatter plot on the back of this paper, and then complete the Reflection questions below.

Reflection Questions (Answer after creating the scatter plot on the back)	
What is the range of median household incomes on this chart?	
What is the range of percentages of adult college graduates?	
Is there a relationship between attributes? If so, how would you describe the relationship?	
Exploring Further: List one way you might do more research to investigate a relationship you found.	
Exploring Further: Which other attributes could you test to create relevant scatterplots?	

DIRECTIONS: Turn your paper landscape to make a scatterplot using the US States dataset. Choose

one of the objects on the table to plot on the axes provided to compare the Median Household Income & Percent Adult College Graduates attributes to analyze the results for any correlation.

Percent of College Graduates



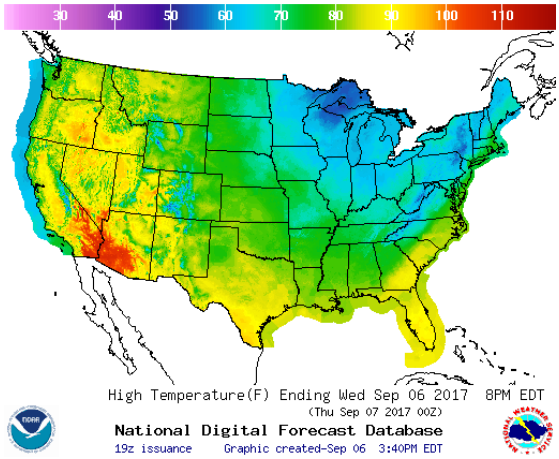
Station 3 Guide - What is a Heat Map?

Print in color, one per station 3

A **heatmap** is a graphical representation which shows data in the form of a map or diagram in which results are represented as colors varying in intensity.

Example

Heat maps are often used in the news during the weather segments

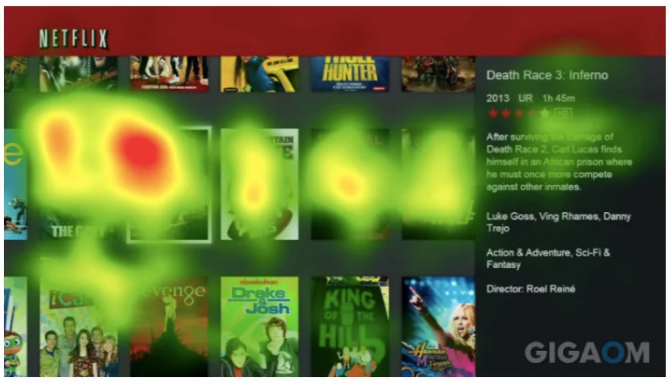


Based on the heatmap above, summarize the weather in your area on this day in 2017. Does this hold true to what you would expect, do you think this is generally accurate?	

Create a heatmap on the back of this paper, and then complete the Reflection questions below.

Reflection Questions (Answer after creating the heat map on the back)

Netflix is perhaps one of the best examples of a digital business that uses heatmaps to gain insights on user behavior and improve user experiences. **What information does the heat map to the right convey?**

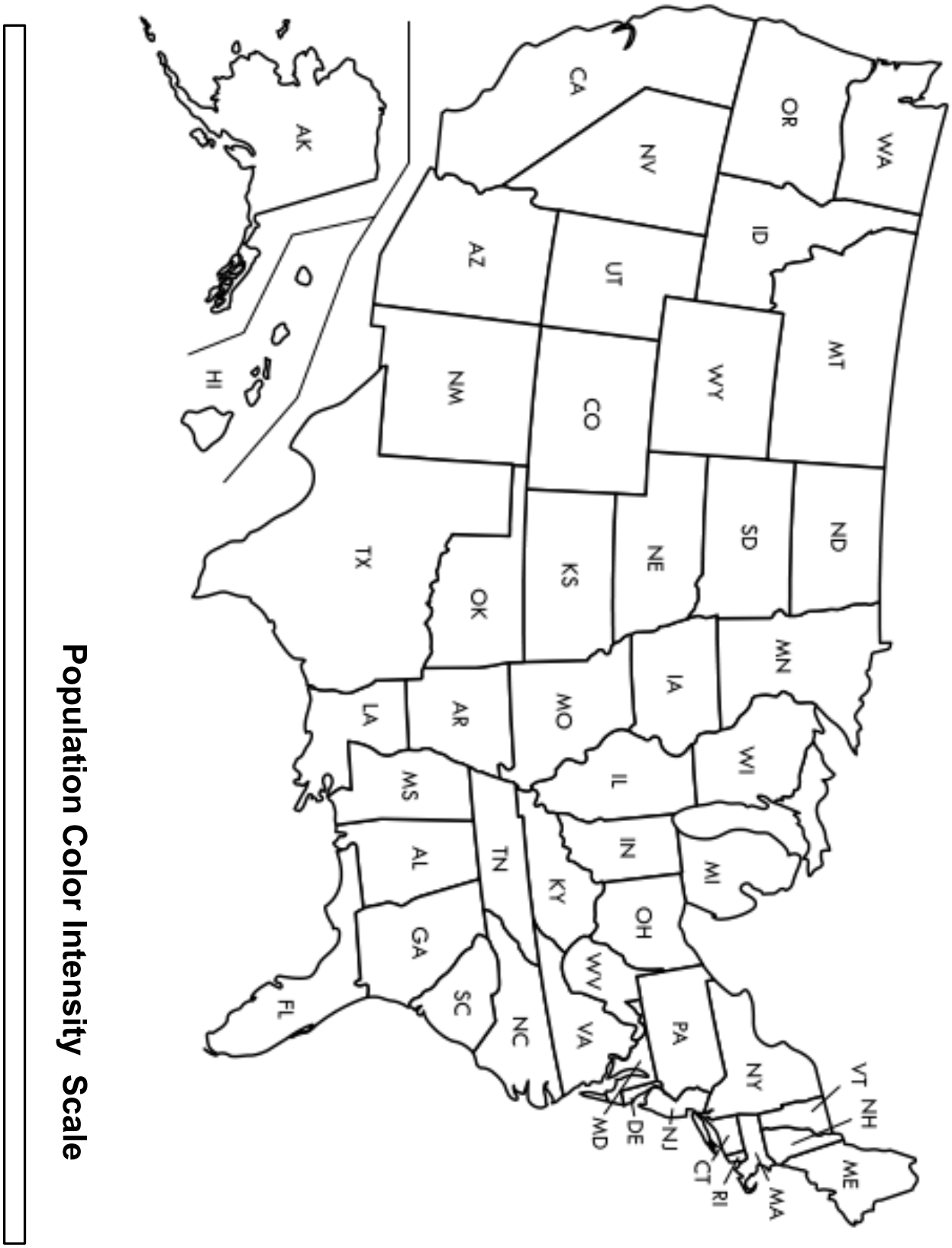


Netflix's old TV UI forced viewers to constantly look back and forth, as seen in this eyetracking study result.

High Eye Traffic — Low Eye Traffic

DIRECTIONS: Use the US States dataset and the map below to create a heat map using the

population attribute. Be sure to include a legend and to blend colors as smoothly as possible.



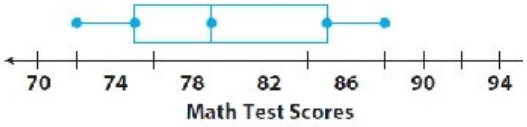
Station 4 Guide - What is a Box plot?

A **box and whisker plot**, or boxplot for short, is generally used to summarize the distribution of a data sample. Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

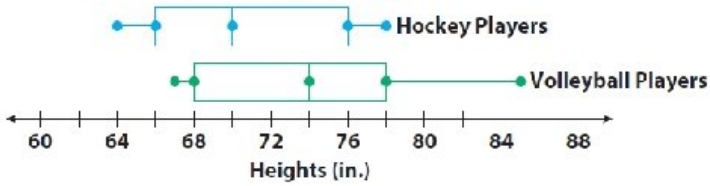
Example

For 1–3, use the box plot Terrence created for his math test scores. Find each value. (Explore Activity)

1. Minimum = Maximum =
2. Median = _____
3. Range = _____ IQR = _____



For 4–7, use the box plots showing the distribution of the heights of hockey and volleyball players. (Examples 1 and 2)



4. Which group has a greater median height? _____
5. Which group has the shortest player?
6. Which group has an interquartile range of about 10? _____

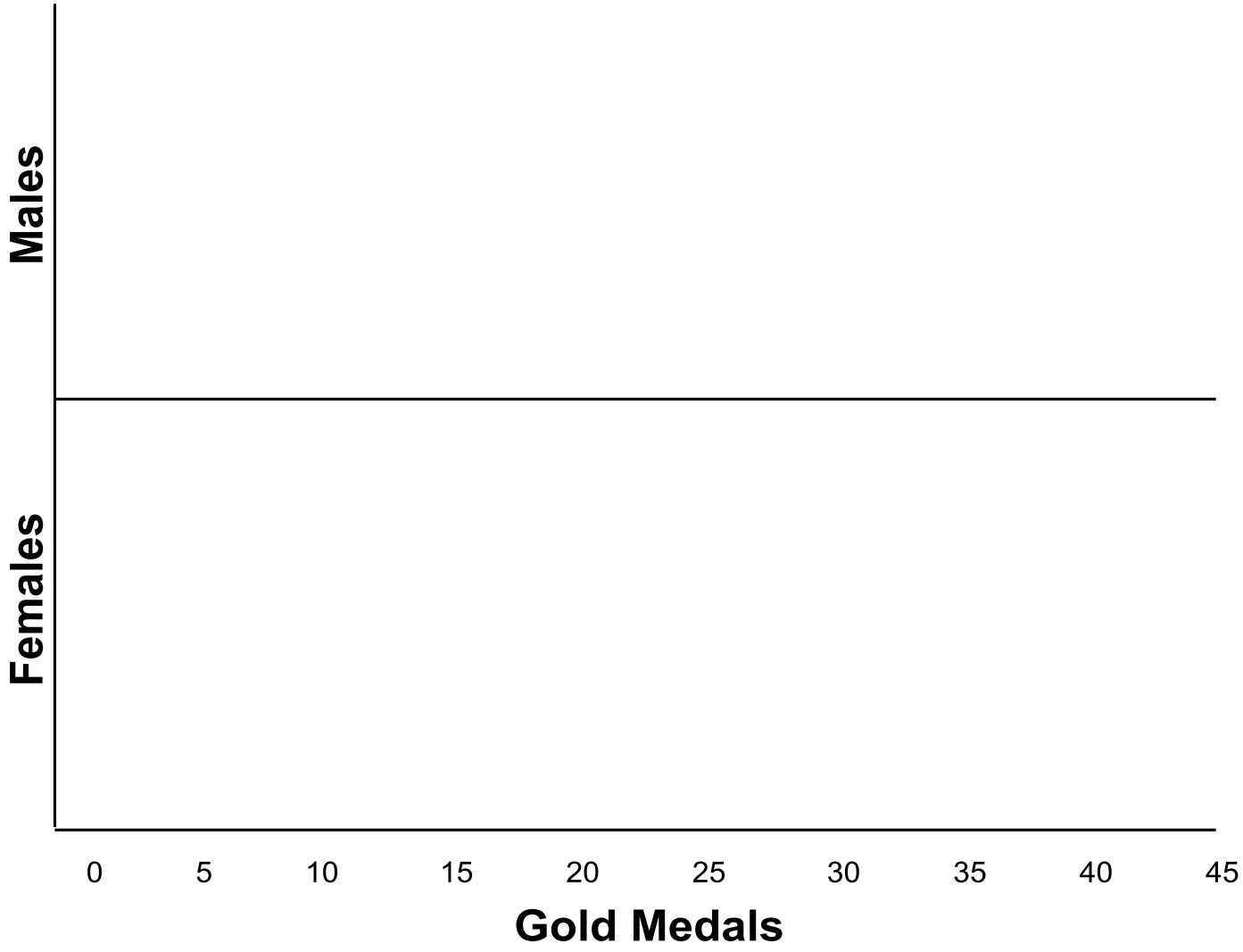
Create a boxplot on the back of this paper, and then complete the Reflection questions below.

Reflection Questions (Answer after creating the boxplot on the back)	
Which gender has a higher median?	
Which gender has a smaller spread?	
How does the skewness compare for males and females?	

DIRECTIONS: Make a comparative boxplot for the Gender of Paralympics Athletes by cutting strips of

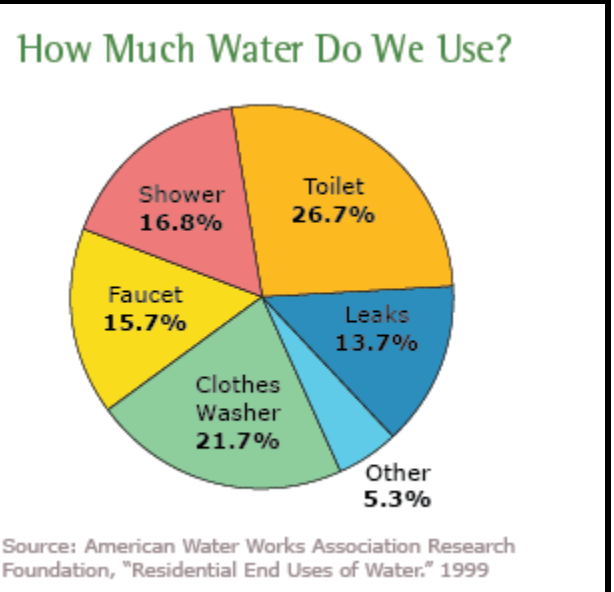
paper and placing them on the graph below. Use the gold medals attribute in the data set by calculating and plotting the following for each gender.

Male Gold Medals	Female Gold Medals
Minimum = _____	Minimum = _____
Maximum = _____	Maximum = _____
Median = _____	Median = _____
Q1 = _____	Q1 = _____
Q3 = _____	Q3 = _____



Station 5 Guide - What is a Pie Chart?

A **pie chart** is a type of graph in which a circle is divided into sectors that each represent a proportion of the whole.

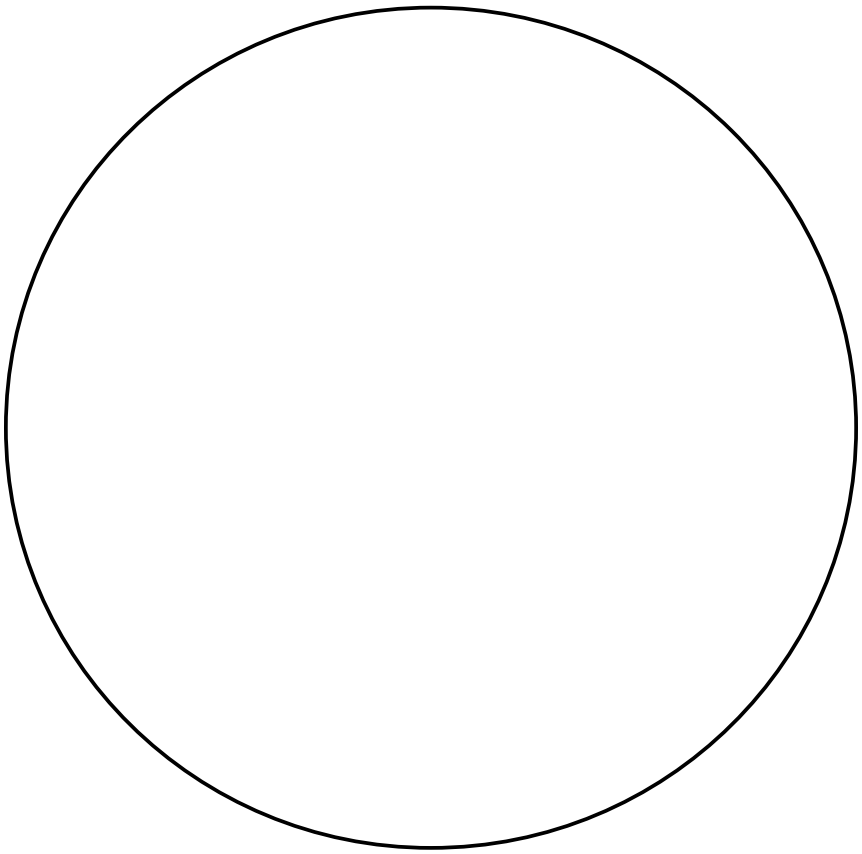
Example	
After interpreting the pie chart to the right, does your water usage match the pie chart?	
How does the leak piece compare with the shower piece? What does this mean?	
What do you think "other" could be?	

Create a pie chart on the back of this paper, and then complete the Reflection questions below.

Reflection Questions (Answer after creating the boxplot on the back)	
What is the overall favorite subject?	
Which two subjects closely compare?	
Exploring Further: Of the students who liked Art the most, what percentage of them were sophomores?	

DIRECTIONS: Make a pie chart for Favorite Classes by calculating the percentages, using the protractor and the formula below to draw on the circle to create a pie chart that is proportional to the data. Be sure to make a legend.

Subject	**Percentage** (Count/Total)	**Central Angle** (Percent * 360°)



Station 6 Guide - Combining Visualizations

Use multiple visualizations to compare attributes of Birth Stats Reduced. This dataset contains information on new born babies and their parents. Perform an exploratory analysis to tell a data story with some of the questions below.

Instructions: Create three supporting visualizations that show a relationship between newborn babies and their parents. Sketch a draft of your visualizations below and then make a conclusion statement. Use a poster board, glue, and the materials of your choice from the previous stations to create a poster of your visualizations.

Driving Questions (choose at least 2 to explore, or come up with your own)

1. Is there a relationship between maternal height and baby length?
2. Can mother's height predict baby length?
3. Can you predict a low birth weight (< 6lbs) using any of the other attributes?
4. Is there a relationship between smoking and low baby weight?
5. Is there a relationship between mothers over 35 years old and low baby weight?

Visualization 1	Visualization 2	Visualization 3

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

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 guide and large graph paper to access the datasets and explore relationships between the 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 Set: Worldwide Internet Users, Prices, and Speeds		
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)	*Paste your link to your presentation here once you have begun to build it*
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	☐ Uses 2 different visualizations ☐ Describe what you thought your visualizations would look like when you chose your attributes from your data set, but before you created the visualizations? ☐ Explain what patterns, trends, or information the visualizations convey ☐ Identifies and describes a problem or area or concern expressed by the visualizations and identifies possible solutions or recommendations to address the issue ☐ Poses 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.

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 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 - Think of something that you are interested in that data could help to explore. 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 survey questions and the data type of the responses

Question	***Data Type***

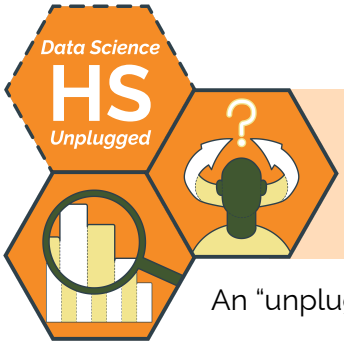
Part 4: Create Visualizations - Create at least two visualizations and sketch them below

Visualization 1	Visualization 2

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

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

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.



Creating Simple Models

An "unplugged" guide to regression models using by eye techniques by Christa VanOlst

Summary

In this lesson, students will use a by-eye technique to create linear and polynomial regression models over scatter plots. As a class, students will compete to make a life-size scale scatter plot using string to model their line of best fit for the data set. As an extension, students may also explore logistic and multivariate models.

Note: This is similar to Creating Simple Models *from the* Data Science with CODAP *&* Data Science with Python *sequence. This lesson focuses on calculating residuals, while the others include opportunities for students to practice with Python & CODAP tooling..*

Objectives

The students will be able to . . .

- Create linear and polynomial models through sketching by eye over a scatter plot
- Make predictions given a linear and polynomial models
- Categorize model types

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

- Warm Up: Optimism Regression & Plot Cards (PDF)
- Correlation Investigation (Desmos) & *Correlation vs Regression* guide (see below)
- Day 2 Warm Up Plots: *Years Experience vs. Salary* (.png), *Prisoner Count vs Population* (.png), *College GPA vs Change of Admission* (.png)
- Cars Complete Data Set (CSV), & Incomplete Data Set (CSV)
- *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)

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
Logistic Regression	Logistic regression approximates binary relationships (True/False, Yes/No, 0/1) 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***, you'll need to **find an outdoor space** to do the regression and residual activity (see steps #8-12). You may find this slideshow, which explains the activity, helpful. You'll also need to print the following materials:

- The regression visualization sheet (see below), one per student
- The Cars Incomplete Data Set (CSV), enough copies for each student to get one row of data

On ***Day 3***, be sure to print the *Calculating & Assessing Residuals* 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.

   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.

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.

   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.

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.

   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.

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

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

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

Day 2 Outline

Formative Assessment Notes

6. **Linear Regression Warm-Up:** Give students this sheet which has the following three visualizations:

   - Years Experience vs. Salary
   - Prisoner Count vs. Population
   - College GPA vs. Chance of Admission

   Then, have them address the the following questions in pairs or small groups:

   - ***How does experience relate to salary?***
   - ***How do incarceration rates compare to the population?***
   - ***Can your GPA impact your chances of admission to Graduate School?***

   Have students briefly share their reflections, & discuss.

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

7. **By-Eye Modeling:** On each graph have students use a pencil and a ruler to sketch a line of best fit by eye. Challenge the students to get as close to the points as possible while maintaining a straight line. (aka student should be adjusting and comparing the slopes of the lines)

   Have students compare their lines of best fit with a partner and discuss whose line fits the data best. Then, discuss:

   .

   How do you know which line fits the data best? Do you think there is a better way to assess the fit of the line?

   After discussion, let students know that in the next activity, they'll learn a way to calculate the quality of a line of best fit.

Not all students will want to share their graphs with the class, so consider anonymizing them or just taking a few volunteers for each graph.

8. **Creating Models Outdoor Activity Setup:** Split the class into 3 or 4 groups. Find a field or a large area outside or clear a space in the room and cut up the rows from this data set.

   a. Create an X and Y axis on the field/floor and give each student a row from the data upon arrival outside.
   b. Each group should find the "complete" rows in their groups' data, and plot the case on the coordinate plane.

9. **Creating Models Line Plotting:** Assign 2 people within each group to be the "line team", and one person to be the "recorder". Have each group calculate a linear model using string:

   a. Have 2 members of the "line team" hold each end of the string and "plot" a line through the flags using the string with input from the rest of the group.
   b. After the "line team" decides on a position, have the other members of the group begin to use their tape measures to find the distance between the flag they placed and the line of best fit. This distance will represent a residual value.

   Explain to students that in a perfect model we would want every flag's residual value to be 0, however that usually isn't possible—just have them get as close as they can.

Use this part of the activity to reinforce "residual" as a vocabulary word—there will be ample opportunities for you and the students to use it in context to describe what they are doing.

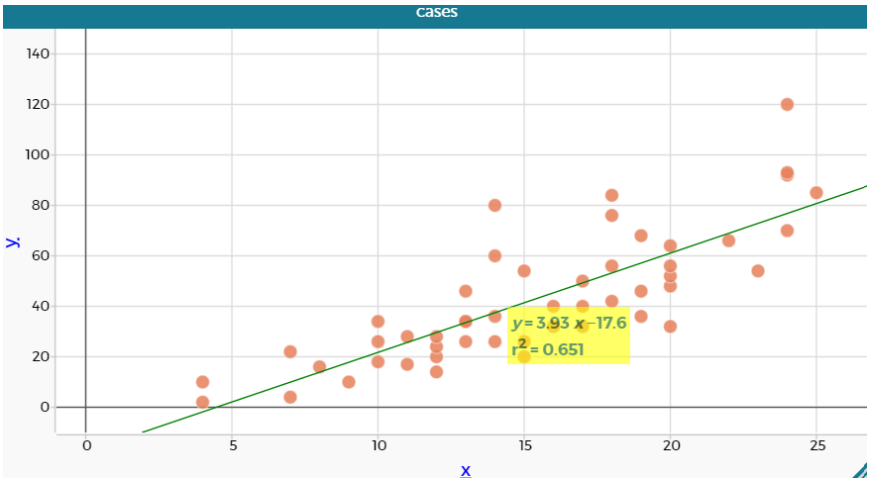
10. **Creating Models & Making Predictions:** Give students time to debate, discuss, and manipulate the line to settle with what they think is the best model with the lowest residual values.

    Then, have students plot their "incomplete" cases on their prediction line and estimate the missing value based on their position. Have students write down their predicted value; they'll be comparing it to a "real" regression line soon.

The takeaway here is that students will discover the tradeoffs of the line of best fit. Moving the line closer to one point may impact the distance of multiple other points.

11. **Assessing the Model:** After students place their predicted values, show students the linear regression equation and the mathematical line of best fit: $y = 3.93x - 17.6$

This activity can be conducted outside using flags or inside on a flat surface using sticky notes/dots.



Have students assess their predicted flags using the actual linear regression line below by calculating the residual between their placed flag and predicted outcome. Have students compare their residuals and evaluate which group's model was the closest to the regression line.

12. **Discussion:** Then have students speculate about the data:

The purpose of this discussion is for students to start to get a sense for the limitations of a model like this, where extreme cases often don't pass the "common sense" test.

If the x-axis represents UFO sightings, and the y-axis is cows per square mile:

- ***How many cows would there be if there are 40 sightings?***
- ***How many UFO sightings do you expect to see if there are 300 cows on a 5 square mile local farm?***

Feel free to change the question from UFOs/cows to a different correlation. The data is for car speeds vs stopping distances.

Day 3 Outline

Formative Assessment Notes

13. **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.

14. **Regression Practice:** Use the Calculating & Assessing using Residuals resource to calculate the residuals and assess the model using the data from the warm-up.

15. **More Practice:** Complete the Desmos: Calculating and Plotting Residuals (with Card sort) and/or the Residuals Review Quizizz activities with students.

16. **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?"

17. **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.

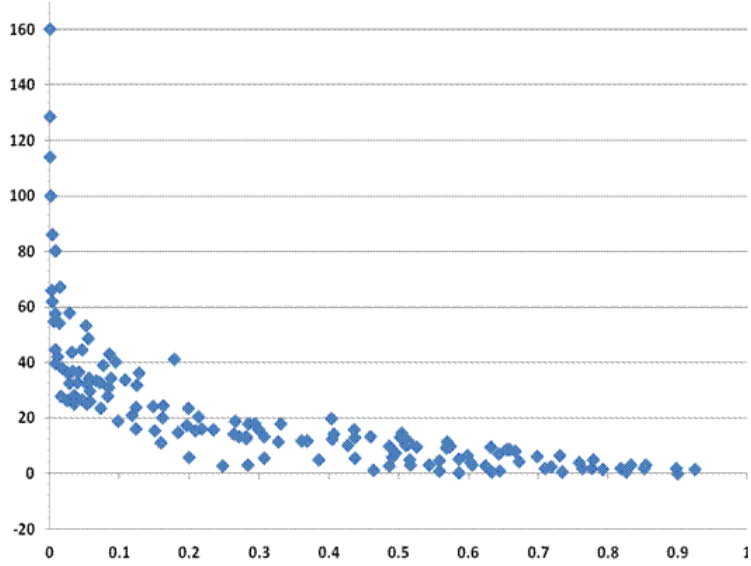
Assessment Strategies

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

Day 3 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.

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.

Encourage students to have definitions from the previous classes available for reference, or provide students with vocabulary sheets/sentence frames to help with discussion sections of the lesson.

Extensions

This lesson is similar to the ***10 Creating Models*** lesson in the ***Python Sequence***. You could extend this lesson by having students complete the Python notebook activities from that lesson after Day 2.

You can also extend this lesson by having students use CODAP to automate the regression & visualization process for larger data sets. Check out the lessons below from the ***CODAP Sequence*** for some ideas:

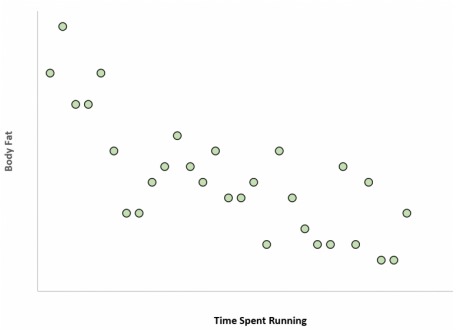
- 08 Creating Basic Visualizations
- 09 Visualizations with Descriptive Statistics
- 10 Creating Simple Models

Student Activity Guide - Correlation vs. Regression

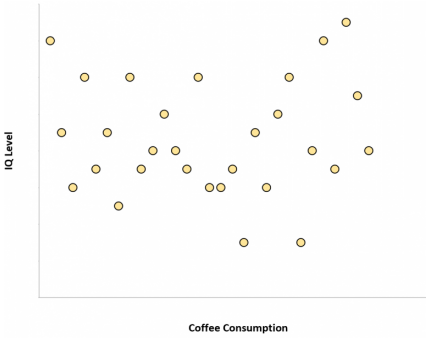
What is a Correlation?

A correlation measures the relationship between two variables.

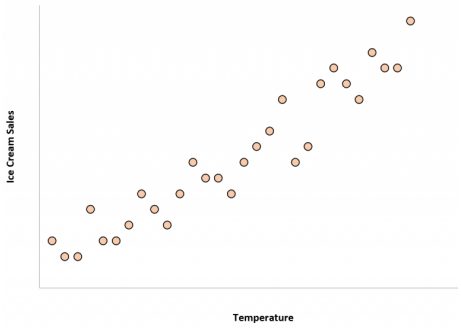
Negative



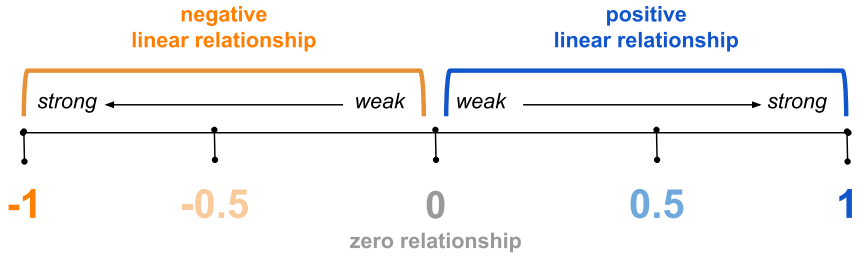
Zero



Positive



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.

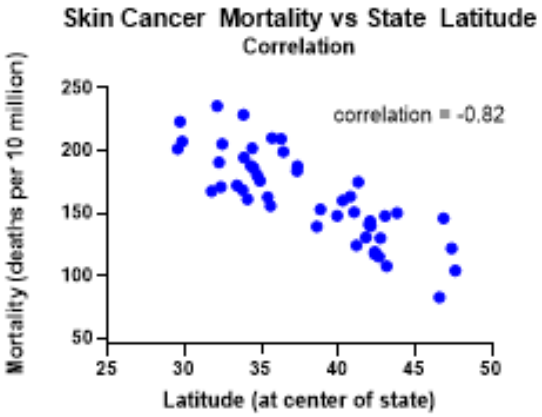


Figure A

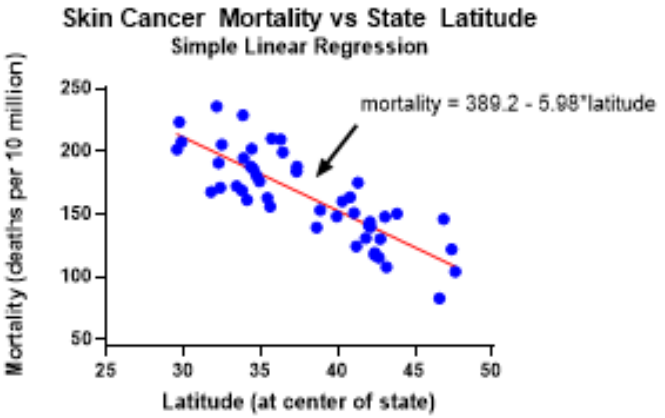
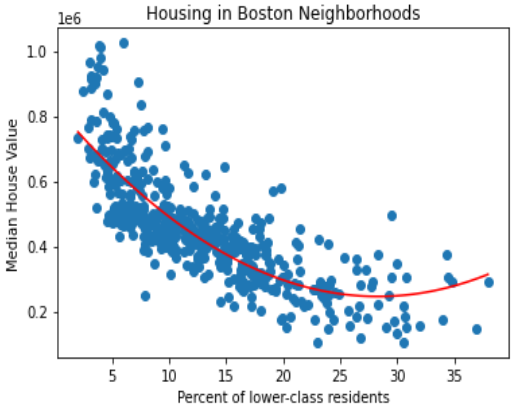


Figure B Citation: GraphPad

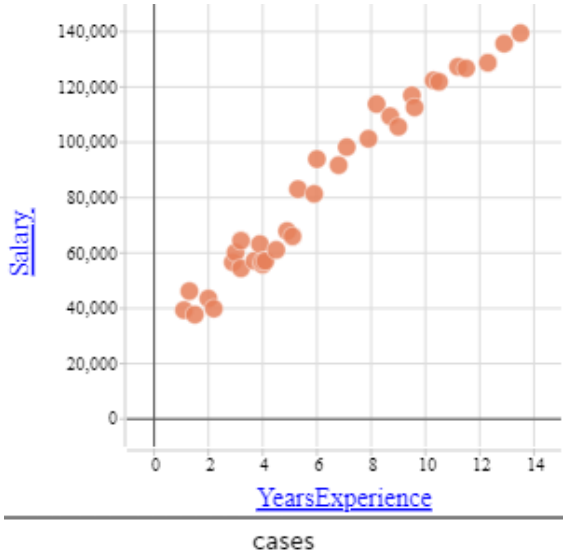
DIRECTIONS: Use the figures above to make predictions by eye and compare them to the calculated regression using the equation.

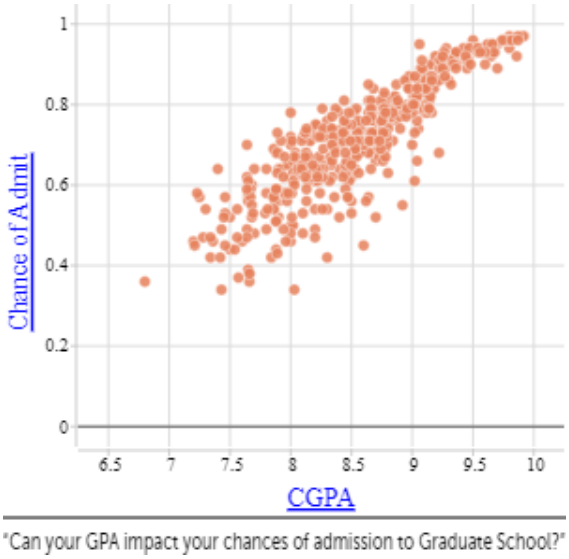
Latitude	By Eye Prediction	Calculated Prediction `-5.98(latitude) + 389.2`
45	*140*	*`-5.98(45) + 389.2 = 120.1`*
36		
25		
50		

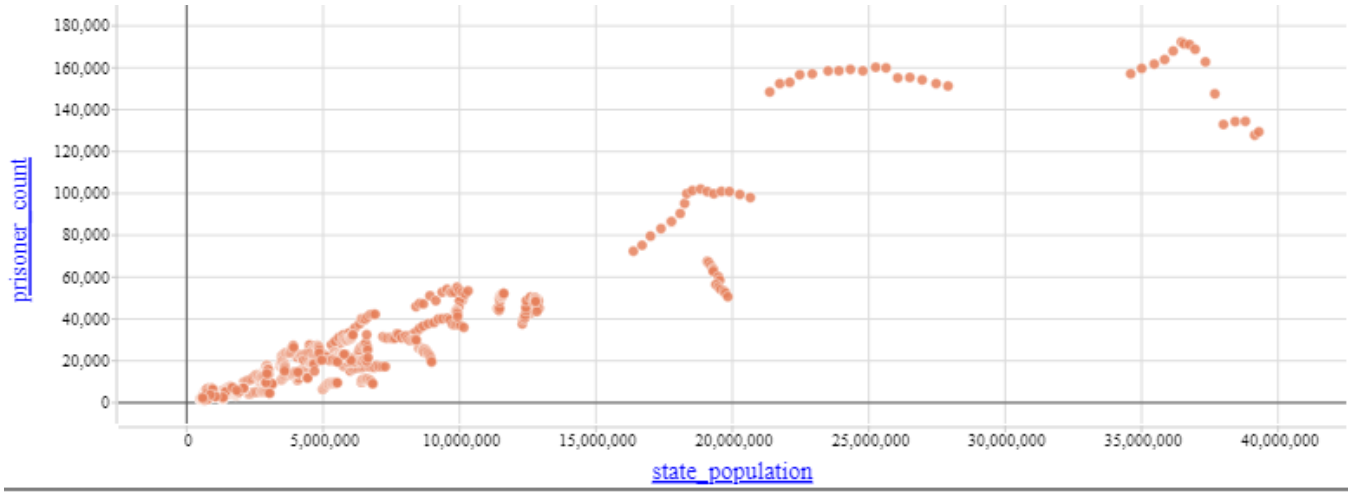
Write a sentence or two to interpret the regression model to the right. What type of regression would you consider this? What questions arise?



Linear Regression Scatter Plots







Using the scatter plots above, answer the following questions:

1. How does experience relate to salary?

2. How do incarceration rates compare to the population?

3. Can your GPA impact your chances of admission to Graduate School?

Calculating & Assessing using Residuals

Vocabulary

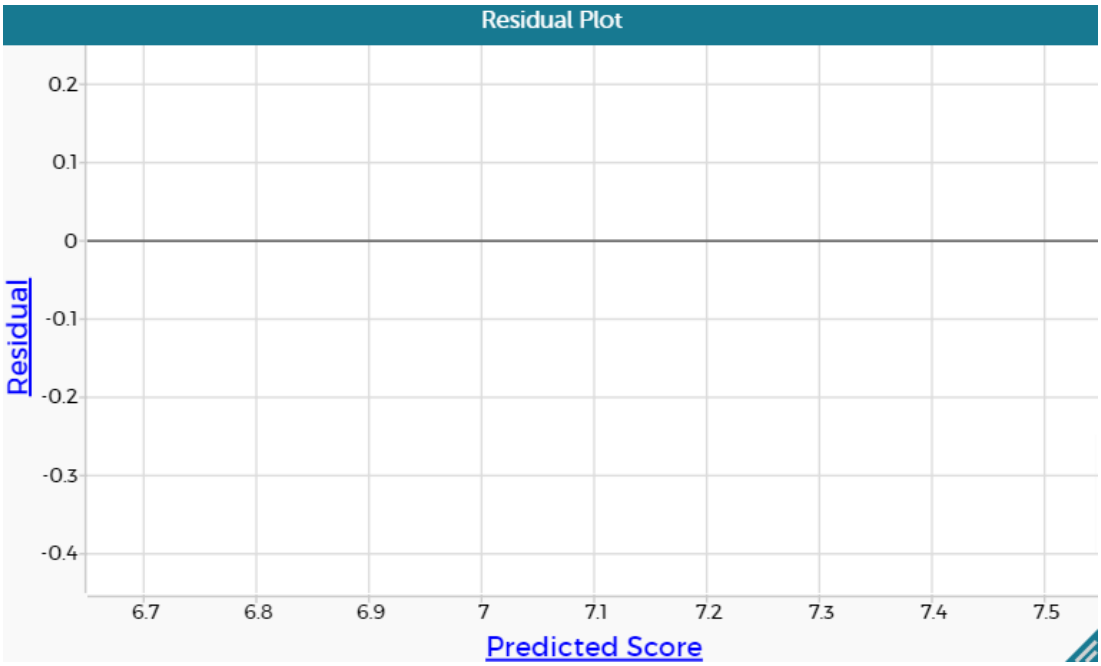
Residual	The difference between the observed value and the model predicted value.
Residual Plot	A residual plot graphs the residuals of a line of a best fit on the vertical axis and the independent variable on the horizontal axis. Residual plots can be used to determine whether a linear model is appropriate for the data.

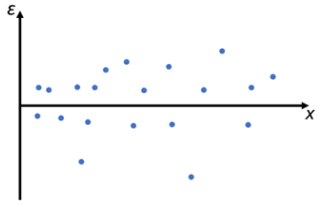
DIRECTIONS: Calculate the following values to complete the table.

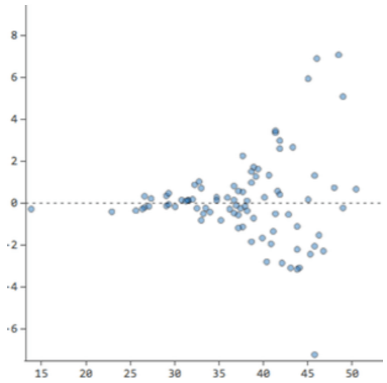
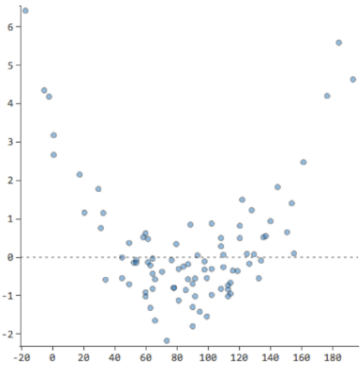
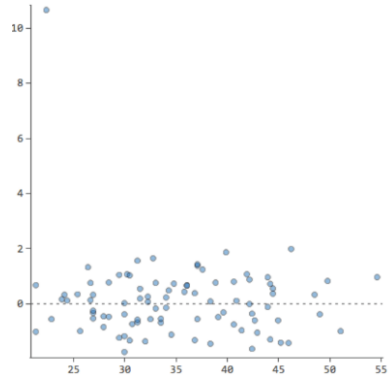
Overall rank	Country	Freedom to make life choices	Score	Predicted Score 3.09*Freedom + 5.6	Residual PredictedScore - Score
1	Finland	0.596	7.769		
2	Denmark	0.592	7.6		
3	Norway	0.603	7.554		
4	Iceland	0.591	7.494		
5	Netherlands	0.557	7.488		
6	Switzerland	0.572	7.48		
7	Sweden	0.574	7.343		
8	New Zealand	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
Random Distribution - the residuals are approximately distributed in the same manner. In other words, we do not see any patterns in the value of the residuals as we move along the x-axis.	

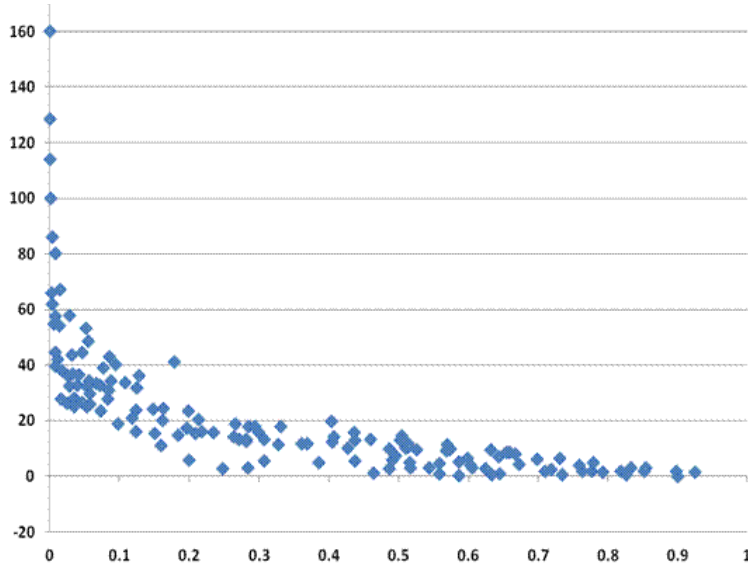
Bad Residual Plots		
Uneven spread - the model does not fit consistently across all x-values.	*Curved* - If there are patterns or curves in the residual plot then a nonlinear model may be more appropriate (quadratic, polynomial, etc.)	*Outlier* - There may be an underlying data recording error. Remove to see what the effect is whether it is influential or not.
		

If you can detect a clear pattern or trend in your residuals, then your model has room for improvement

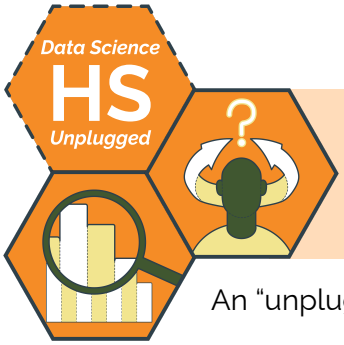
Day 3 Exit Ticket *printable*

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.



Using Models to Predict

An "unplugged" guide to using linear and polynomial regression models by Christa VanOlst

Summary

In this three day lesson, students use two techniques to make predictions about missing or future data: a by eye technique and evaluating functions using a given mathematical model. Students explore datasets throughout the lesson by creating scatter plots and models to predict outcomes. At the end, students discuss the tradeoffs and limitations of certain models, collect data, and analyze it for correlation.

Note: This lesson is very similar to Making Predictions *from the* Data Science with CODAP *&* Data Science with Python *sequences. In this lesson, students primarily compare & contrast "by-eye" predictions with models, rather than Python or CODAP for analysis.*

Objectives

The students will be able to . . .

- 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.
- **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 tools to analyze data effectively.

Materials

- Health Foods Data Talk (Desmos)
- Data Sets: Experience & Salary (Google Sheet or make a copy), Fuel Consumption since 1990 (Google Sheet or make a copy), Crime Data (Google Sheet or make a copy), College Admissions (Google Sheet or make a copy), Position and Salaries (Google Sheet or make a copy), Fish Data Set (Google Sheet or make a copy), Stopping Distance & Speed (CSV)
- Student Guide - Sketching Models & Making Predictions (see below)
- Student Guide - Exploring Regression (see below)
- Whiteboard Activity Slideshow (Google Slides or make a copy)
- Project: Prove a Relationship (see below)

Vocabulary

Term	Definition
Line of Best Fit	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 nonlinear relationships (quadratic, cubic, etc) between two variables to make predictions and describe situations

Day 1 Outline

Formative Assessment Notes

1. **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. Either based on student responses or after responses, be sure to discuss these ideas:

   a. 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
   b. Ask students how "good" the relationship is, and why. Use this discussion to transition into the next activity, which introduces a correlation coefficient.

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.

2. **Reviewing Regression Lines:** Give the students the: Experience & Salary Data & the Experience vs Salary scatterplot from the previous lesson, where they sketched a regression line before learning about residuals. Have students review briefly in pairs:

   - ***Describe how experience relates to salary.***
   - ***Is the correlation strong or weak?***
   - ***What questions would you ask to learn about the reasons for this pattern?***

If students need to review unit conversions, you might ask students – "If someone has 3.25 years of experience how does that correlate to months? They would have 3 years and how many months experience?"

4. **Making Predictions:** Model creating a by eye sketch of a linear model on the Years Experience vs. Salary scatter plot. [*example model sketch*]

   Have students use the model to predict the outcomes for these scenarios:

   - 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]

Students need to create models more or less independently in step #7. If your students need more review, use this time to do it.

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?***

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.

6. Repeat step 2 using the Fuel Consumption since 1990 Data Set

   Discuss with students that without technology /programming we can calculate these equations by hand, but with larger data sets it is helpful to have computers do these calculations automatically.

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

7. **Regression Practice:** Have students create their own by-eye regression models using the Student Guide - Exploring Regression.

   Summary: In pairs or groups, students will explore relationships between attributes in other data sets.

Consider having students write their findings in their journals.

Day 2 Outline

Formative Assessment Notes

8. **Warm Up:** 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)

9. **Using Mathematical Models to Make Predictions:** Have students revisit their by eye predictions from the scenarios in step #4.

   Give them the following linear regression model:

   - Salary = 8732*YearsExperience + 28860

   Review slope-intercept form from Algebra 1:

   $$y = m(x) + b$$
   $$\downarrow$$
   $$Salary = 8732(YearsExperience) + 28860$$

   Demonstrate how to use the function to predict an outcome:

   - If, Years Experience = 1.5
   - Then, Salary → 8732*1.5 + 28860 = $41,958

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.

10. **Comparing Predictions:** Have students return to their "by eye" predictions from step #4, but this time use the linear model to calculate the expected outcomes, Have students compare their by eye predictions to the model predictions:

    - A doctor with 7 years experience
      - By eye ~90K → Calculated = $89,984
    - A data scientist with 3 years experience
      - By eye prediction ~55K → Calculated = $55,056
    - A voice-over artist with 12 years experience
      - By eye prediction ~134K → Calculated = $133,644

Check in with students to make sure they understand how to plug values into the model and derive predicted values.

11. **Predictions Practice:** In pairs, have students repeat this step using the Fuel Consumption since 1990 Data Set, by making predictions and then using the linear regression model to assess their predictions:

    - Fuel Use = 2.575(Year Since 1990) + 132.7

12. Using the table at the end of the Student Guide - Sketching Models & Making Predictions, give students the following mathematical models for each scatter plot:

Models	
Linear	prisoner_count = 0.004(state_population) - 434
Linear	chance_of_admission = 0.18(cumulative_gpa) - 0.85
Exponential	salary = $23695(1.4)^{level}$
Quadratic	weight = $0.62(diag\_length)^2$ - 50.4(diag_length) + 1240.26

    Have students calculate predictions by evaluating the function using a calculator and then reflect on how these values compare to their previous by eye predictions.

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(speed) - 17.6

    Have students use the equation to predict the stopping distance for cars traveling: 4 mph, 15 mph, 25 mph, 75 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 3 Outline

Formative Assessment Notes

14. **Project: Prove a Relationship:** In this project, students hypothesize a correlation that can be supported or disproved by collecting data from their classmates.

See *Assessment Strategies* below for details

Assessment Strategies

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

Mini Project: Prove a Relationship

Each will come up with their own hypothesis to test through surveying classmates. They will then use these responses to create a scatter plot to assess the relationship. Once plotted, students use a model to predict future values and reflect on the applicability and limitations of their findings.

Milestones for Students:

1. Make a hypothesis statement
2. Survey and collect data from peers
3. Plot the data on a scatter plot
4. Sketch a model of best fit and make predictions by eye
5. Express their model as a function (teacher discretion)
6. Conclusion to reflect on their hypothesis statement and the applicability and limitations of their findings. Prompt students to answer the questions:
    - Was your hypothesis correct? Were you surprised by the outcome?
    - What are the limitations of your model? Does it make sense for all x-values?
    - What further questions arise when you analyze your results?
7. Create a presentation to include: hypothesis statement, a summary of data collection, visualization, regression model, at least one prediction, and conclusion

Mini-Project Rubric

	Proficiency	***Yes***	***No***	***Notes***
Hypothesis	Student-created hypothesis is a **tangible statement** that can **prove or disprove a correlation** between tested attributes			
Survey	Student-created survey is **relevant to their hypothesis** AND **appropriate data is collected** from their peers, stored, and organized			
Data Visual	Students' scatter plot is **appropriate for the data attributes** AND **provides insight** to sketch a model			
Model	Students' sketch of their **regression model is appropriate** and relatively accurate for the data AND the student **makes a valid prediction**			

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

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 to help students organize into smaller chunks.

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 CODAP and Python Versions of this lesson. Making regression equations and predicting values based on those models is much more powerful using large datasets, and calculating these equations by hand doesn't make much sense when students can use tools like CODAP and Python.

Student Guide - Sketching Models & Making Predictions

Vocabulary

Line of Best Fit	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 sets: Crime Data, College Admissions Data, Position and Salaries Data, and Fish Data Set

Create the following scatter plots and sketch an estimated line of best fit, including a scaled and labeled axis for each plot.

State Population vs. Prisoner Count	**Cumulative GPA vs. Chance of Admission**
Position Level vs. Salary	**Diagonal Length vs. Weight** *plot each type of fish in a different color*

Answer the following quesitons:

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 relationships positive/negative? Strong/weak?

Complete the following by eye predictions using your models:

Scenario	By 👁 Prediction	Mathematical Model *(Your teacher should give this to you)*	Calculated Prediction	Comparison
California has a population of 39.35 million, what is their predicted prisoner amount?				
A student has a gpa of 8.3 (out of 10), what is their predicted chance of admission?				
A 4.5 level manager should expect what salary?				
If a perch fish has a diagonal length of 20 cm, what is the expected weight?				

Student Guide - Regression

Data Set 1: _____

1. What attributes are being tested?
2. What is the question you are exploring?
3. What is the shape of the line of best fit? (linear or polynomial)
4. Is the relationship strong/weak? Positive/negative?
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.

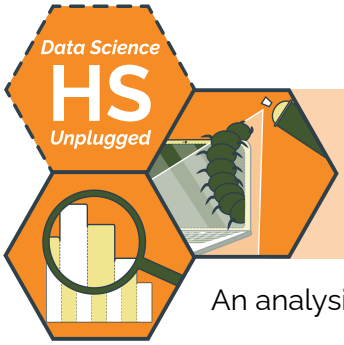
Input Value	***By Eye Prediction***

Data Set 2: _____

7. What attributes are being tested?
8. What is the question you are exploring?
9. What is the shape of the line of best fit? (linear or polynomial)
10. Is the relationship strong/weak? Positive/negative?
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.

Input Value	***By Eye Prediction***



Overfitting and Noise

An analysis of overfitting models by Sara Fergus

Summary

In this lesson, students learn 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 Data Science with CODAP & 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 Science Scenario – Communication (see below)

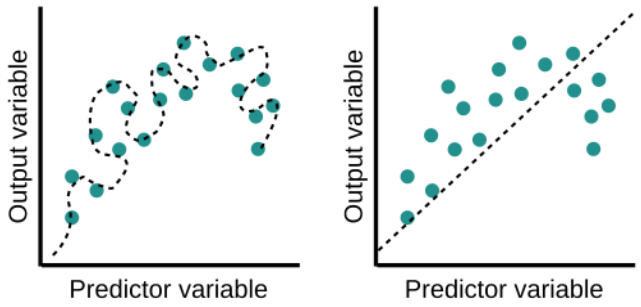
Vocabulary

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

1. **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.

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.

3. Have students read and annotate this worksheet: Model Limitations: Noise and Overfitting, focusing on finding definitions for the terms "noise", "underfitting", and "overfitting".

Float around to check for understanding.

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?

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 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?

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?***

When reviewing answers as a class, circle around the room all together and focus on the "how is this an example of overfitting?" question.

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.

See *Assessment Strategies* for details & a rubric.

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:

If reading the entire machine bais 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"

- ***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.

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 R2].

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

	Proficiency	*Yes*	*No*	*Notes*
Example	The example used **is an example of overfitting** in that it either is too closely based on past instances and/or the model is overly complicated, thus modeling noise more than pattern.			
Presentation	The presentation is **clear and understandable.**			
Workshop	The student workshopped a peer's project and **accurately advised on whether the example is overfitting.** The student then **thoughtfully answered all questions.**			

Some Accommodations & Extensions

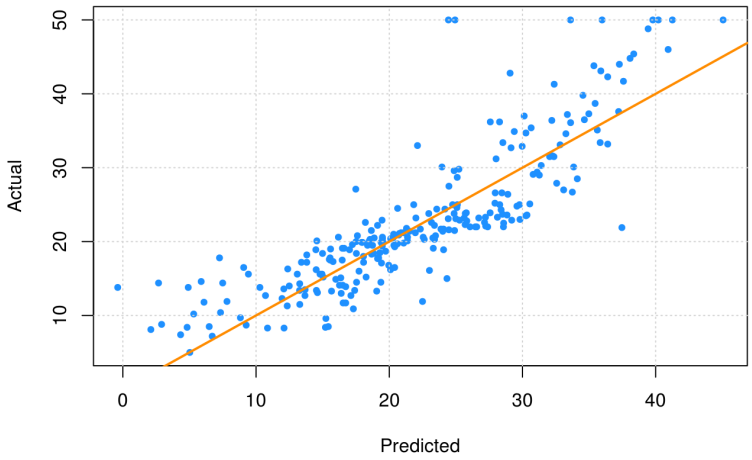
Students who need additional time reading may benefit from getting the overfitting worksheet ahead of time. The worksheet could also be annotated as a class or in a small group. For students with small group accommodations, consider pulling aside a few students and helping them to complete the worksheet, 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.

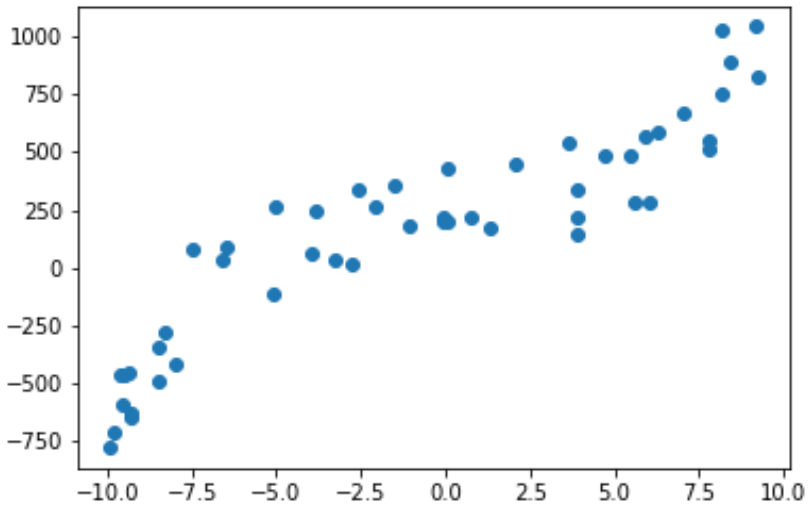
Model Limitations: Noise and Overfitting

A guide to noise, overfitting, and bias by Sara Fergus

Noise

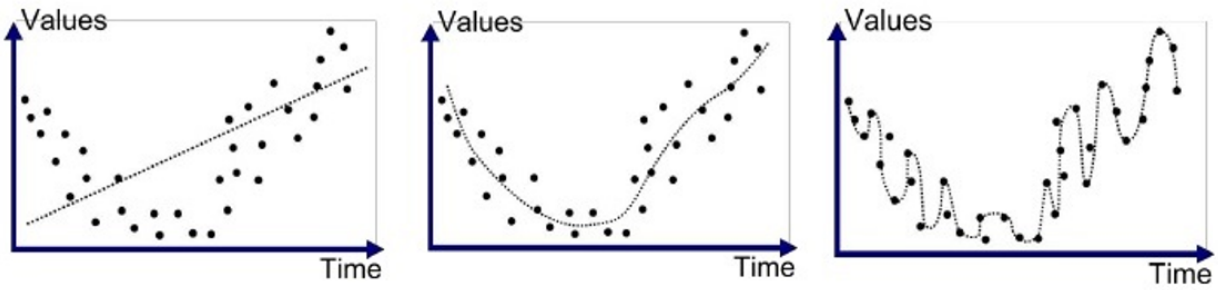


In Data Science, **noise** is a word used to describe random pieces of data that make the underlying pattern less clear. This comes from the fact that neither humans nor nature are perfect. For example, noise is introduced by measurement and rounding errors in data collection. Noise is also introduced 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.	

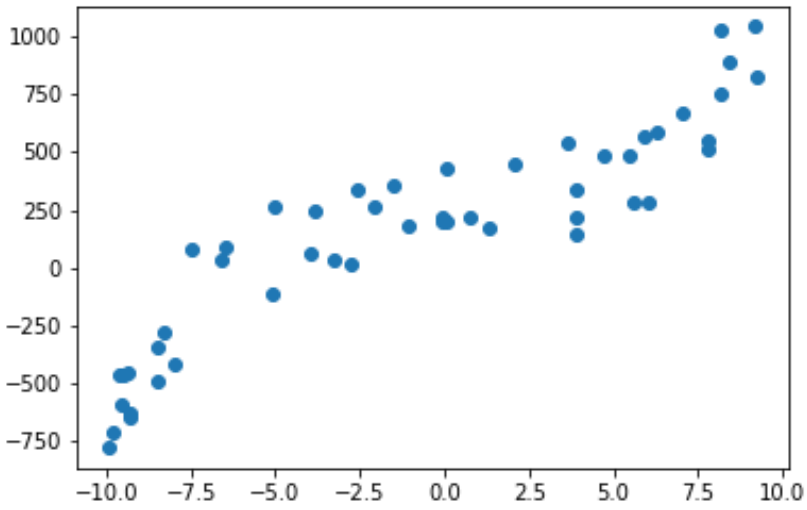
Underfitting and Overfitting

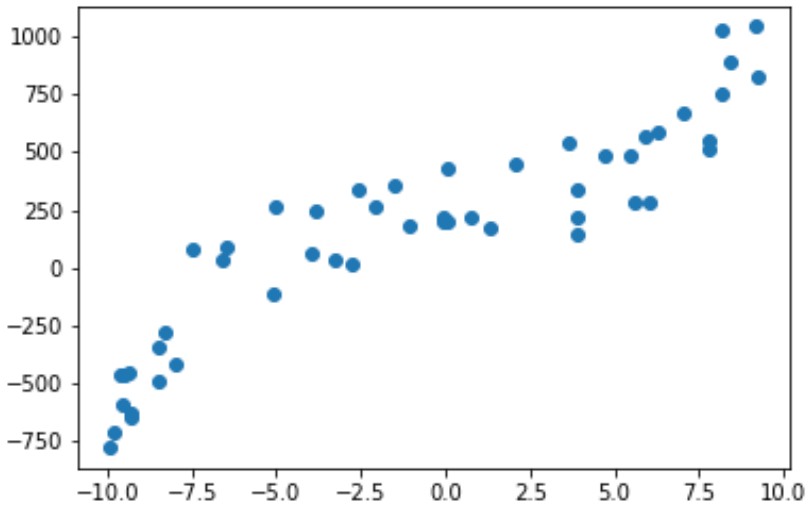
Underfitting is when a model is too simple and leaves out some important information. **Overfitting** is when useless details (i.e. noise) have too much of an effect on the model. These graphs show an example of each. You can see that, in general, the values decrease and then increase. A linear model 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.**

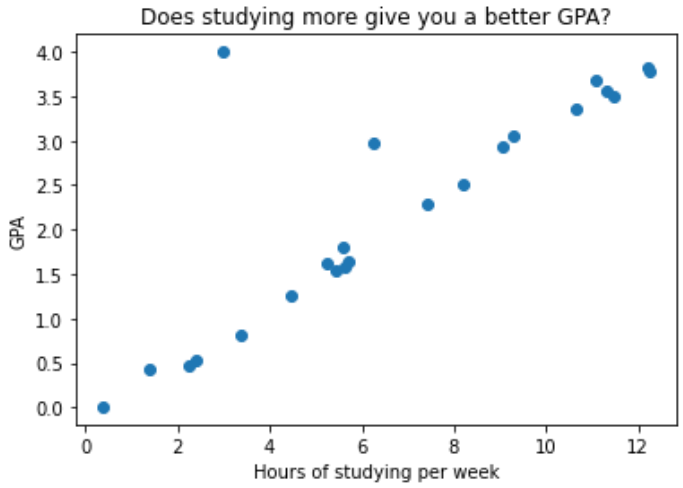
Sketch a by eye model that would be underfitted	

Sketch a by eye model that would be overfitted	
Define overfitting in your own words	
Define underfitting in your own words	

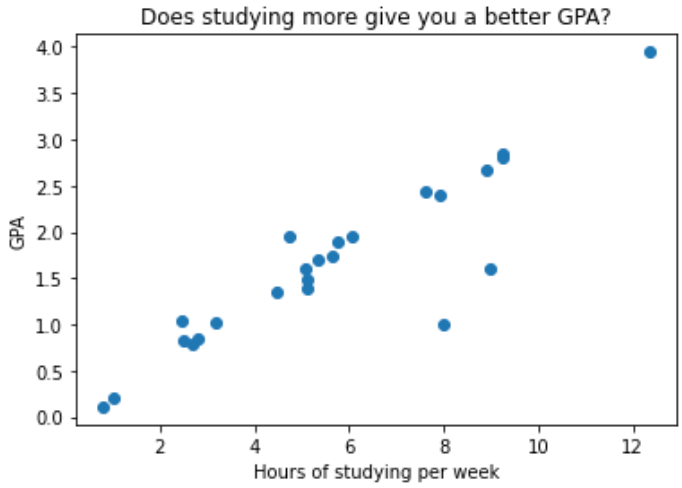
Training and Testing

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.



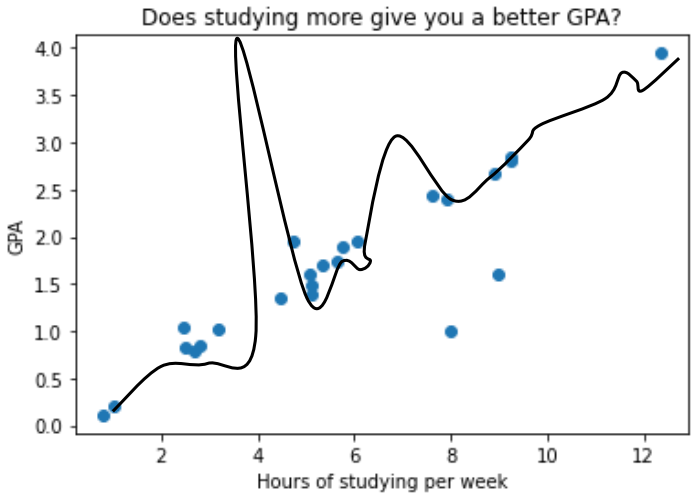
Training Data



Testing Data

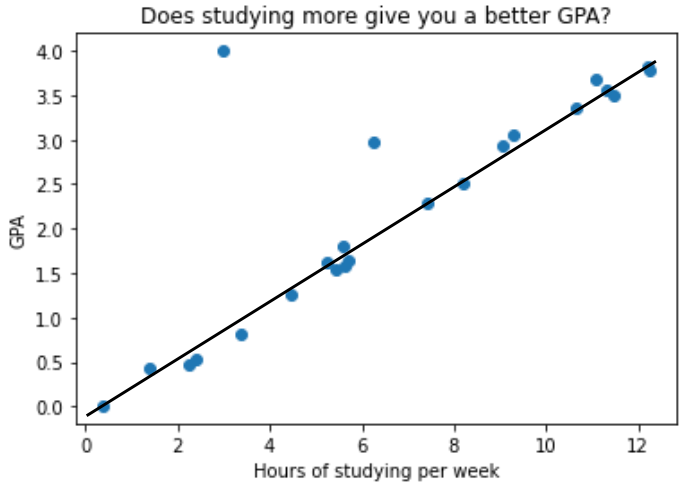
Let's say that the Data Scientist overfits their training data. It may look like the training model below. This fits the data *really well*. However, it is an overfitted model. One good way to tell that the model is overfitted is that the increases and decreases are meaningless. For example, this model 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

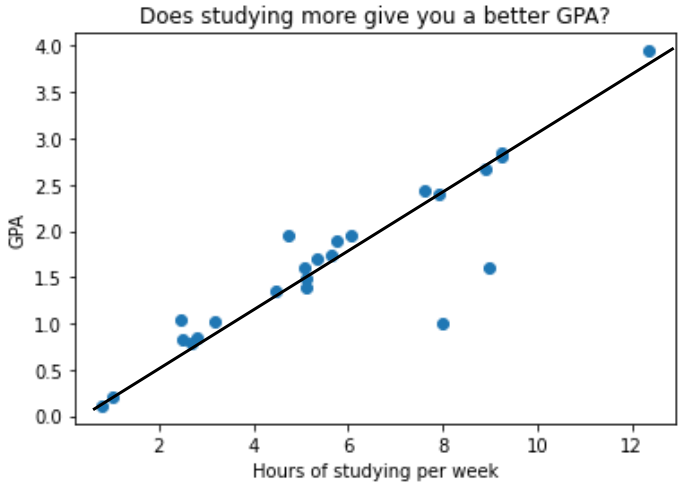


Testing Data

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?"

indeed®
one search. all jobs.

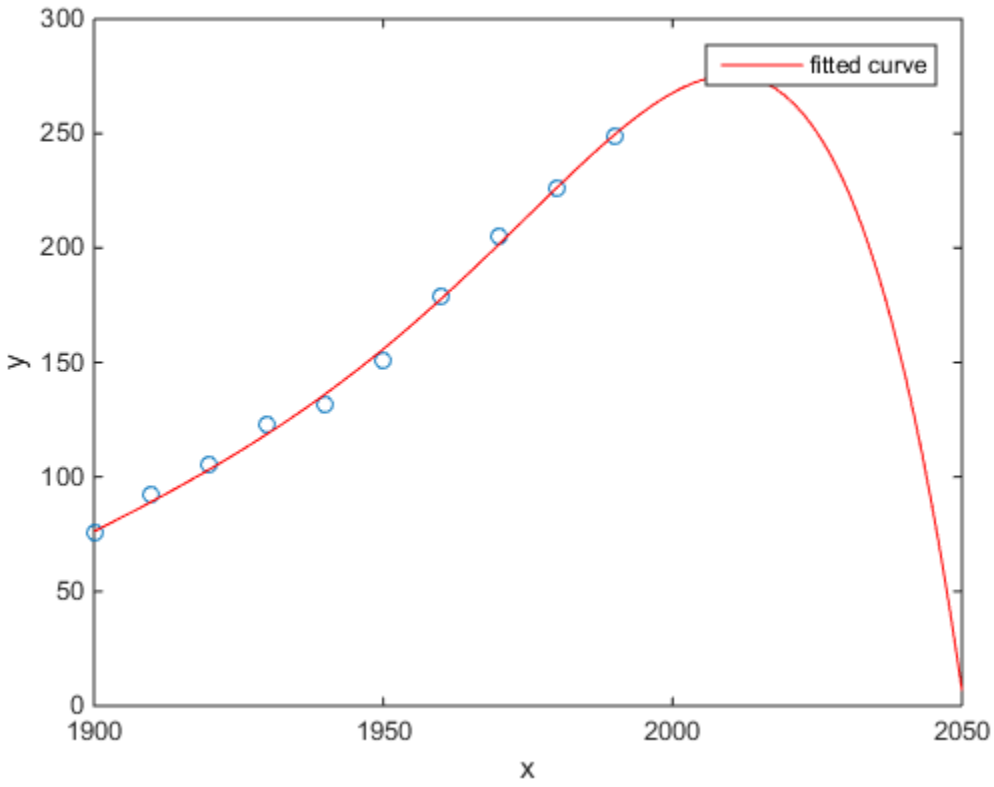
Wanted: Electrical Engineer. 42 year old androgynous person with degrees in Electrical Engineering, mathematics, and animal husbandry. Must be 68 inches tall with brown hair, a mole over the left eye, and prone to long winded diatribes against geese and misuse of the word 'counsel'.

Sports Resource

Population Resource

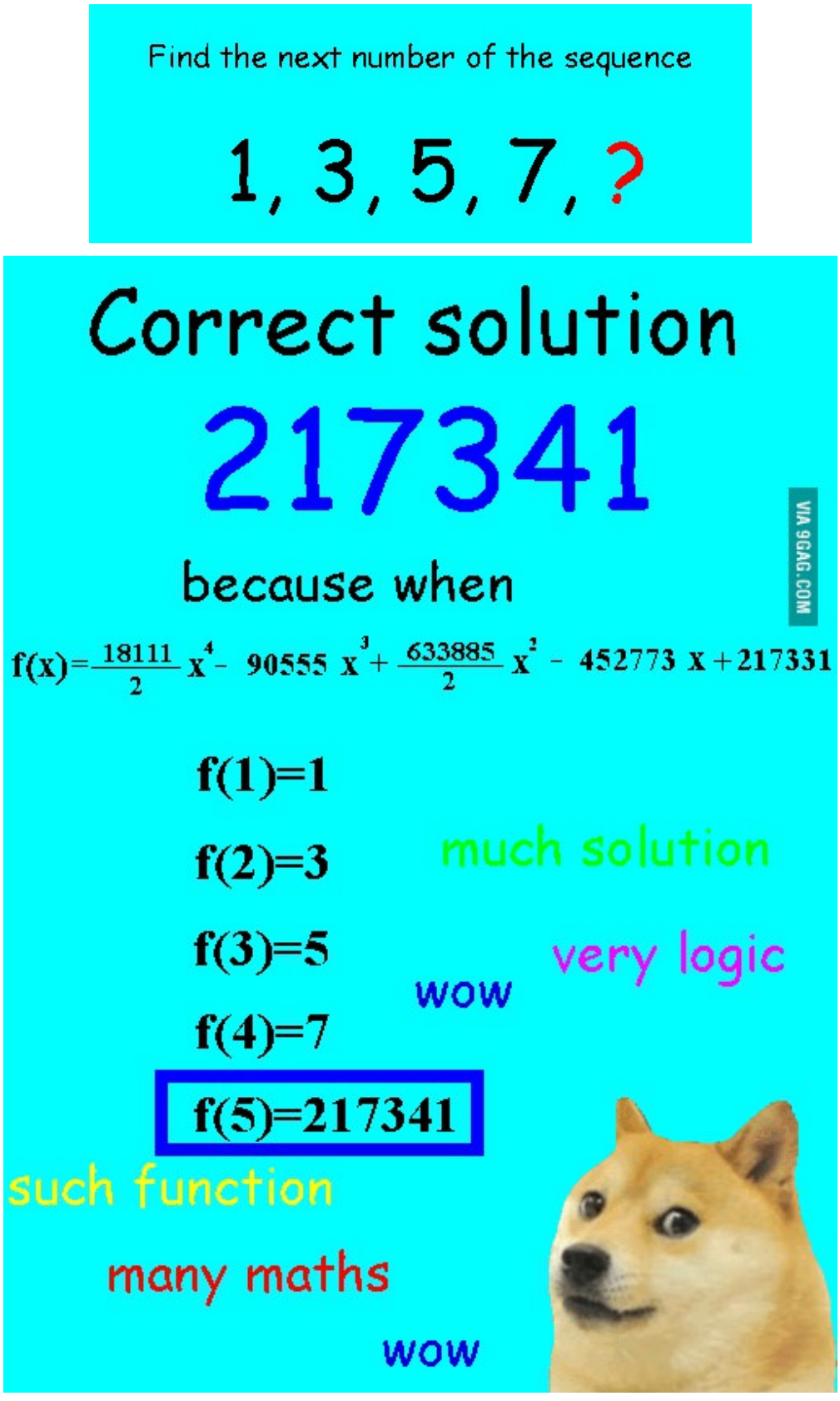
The Apocalypse is Coming!

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.

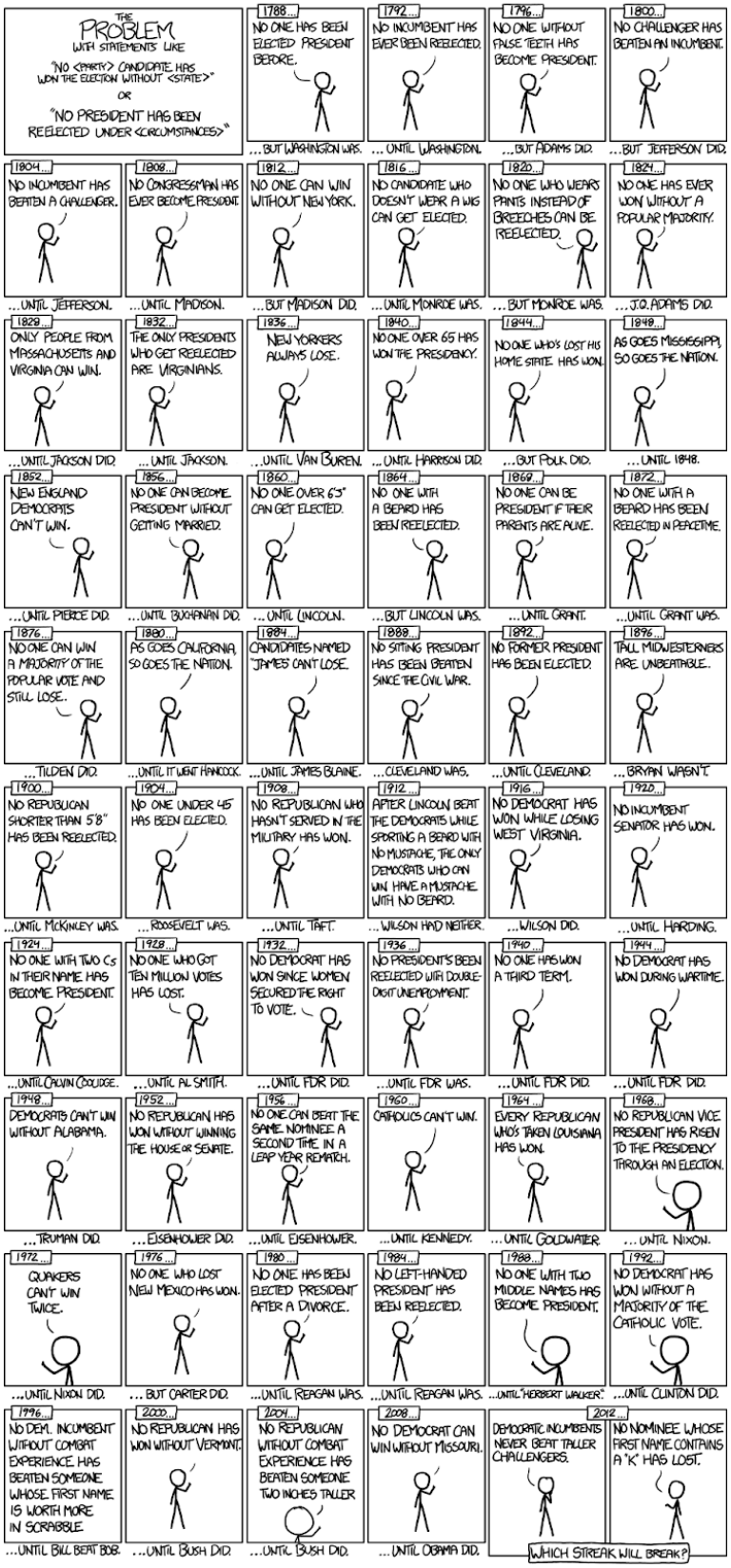


We suggest moving out of the country as soon as possible. Researchers are still working on modeling populations in other countries. It is possible that this event will not only occur in the United States.

Pattern Resource



XKCD President Comic Resource



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

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

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 measurement 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.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]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(hair_length) - 6.56`
2. R = `0.961`
3. R^2 = `0.924`

↓

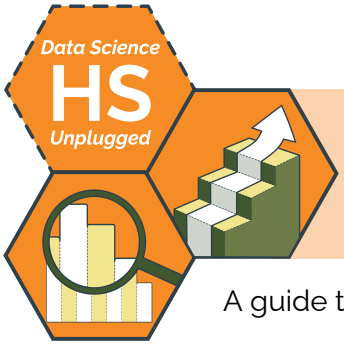
Data Communication:

(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. At the end of the lesson, they will be better prepared to write and summarize their own data reports.

Note: This lesson also appears in CodeVA's Data Science with CODAP & 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 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.

Add notes as appropriate for assessing student learning in this step of the lesson

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 Rish
   - 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)

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.

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

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.

4. **Mini Project:** Have students completed the *Data Journalist* mini-project (see *Assessment Strategies* below).

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

	Proficiency	***Yes***	***No***	***Notes***
Completeness	You accurately included the most important points from your report in your news clip.			
Representation	You included at least two visualizations that help viewers better understand your report.			
Summary	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

	Exemplary	***Proficient***	***Developing***
Important Points		You accurately included the most important points from your report in your news clip.	
Visualizations		You included at least two visualizations that help viewers better understand your report.	
Report Style		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.

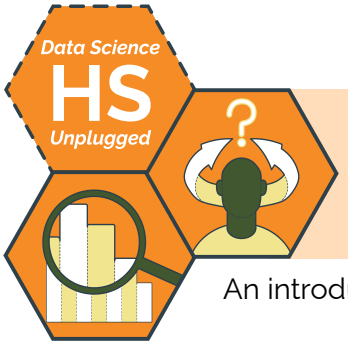
Accommodations

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: Developing 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 Data Science with CODAP *&* 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

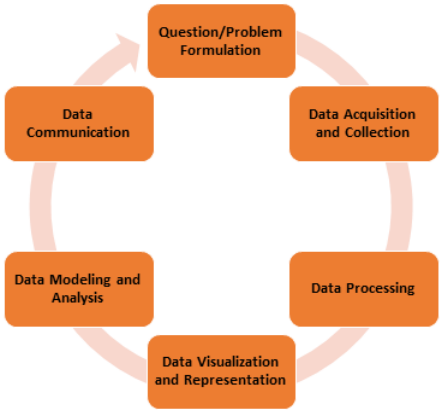
Standards Alignment

- **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	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. 
Research Project	A research project follows the traditional data cycle: 1. Choose a topic you are interested in 2. Ask a specific question 3. Collect or acquire data to answer the question 4. Perform a data analysis 5. Draw a conclusion 6. Ask a new question
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, 1. Acquire or Collect Data 2. Explore the Data 3. Ask a specific question Once a specific question is asked, an exploratory data analysis becomes a research project at step 4 4. Perform a data analysis 5. Draw a conclusion 6. Ask a new question

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 *broad question* or a *subjective question*. 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	This is the starting point A *broad question* is one that cannot be answered on its own because it is unclear and/or undefined. For example: • What makes you a good athlete? • Are girls more successful than boys? A vague question can often break down into good Data Questions.
Subjective Questions	A *subjective question* 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. • What is the best book ever written? → What is a common favorite book? • Who is the best leader in history? → What traits do people look for in a leader? A subjective question can often break down into good Data Questions.
"Dead-End" Questions	A *dead-end question* 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. • How many people live in the United States? • How tall is the tallest person in the world? • Who was Alexander the Great?
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.

   Consider providing students with examples from previous classes if they are stuck

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.

   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.

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.

   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.

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.

5. **Research Question Exit Ticket:** Have students draft a question they might investigate during their final projects

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.

Check in with students about their chosen data

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).

Students should be able to identify the attributes in the data set and pose questions based on them without much guidance at this point.

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.

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.

	Proficiency	***Yes***	***No***	***Notes***
Data Question	Can be answered with data			
Specific	The research question is not a "broad question" and has been distilled.			
Objective	The research question is objective. If the topic of research is subjective, the research question itself is about people's perception of the topic			
Fruitful	The question cannot be answered with a simple Google Search. Answering the data question would lead to more questioning and future projects.			
Data Collection	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?

	Proficiency	***Yes***	***No***	***Notes***
Project Type	Student correctly identifies whether their project is a research project or an exploratory data analysis			
Data Question	Can be answered with data			
Specific	The research question is not a "broad question" and has been distilled.			
Objective	The research question is objective. If the topic of research is subjective, the research question itself is about people's perception of the topic			
Fruitful	The question cannot be answered with a simple Google Search. Answering the data question would lead to more questioning and future projects.			
Data Collection	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?

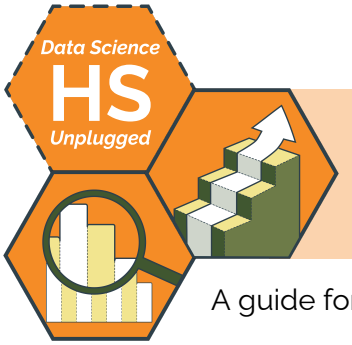
5. Did the article answer their research question? If they did, what was their answer?

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:

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

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. 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 Data Science with CODAP & 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)
- Diamonds Data (view Google Sheet or make a copy)
- Grocery Store Marketing Analytics (view Google Sheet or make a copy)

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: 1. Choose a topic you are interested in 2. Ask a specific question 3. Collect or acquire data to answer the question 4. Perform a data analysis 5. Draw a conclusion 6. Ask a new question
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, 1. Acquire or Collect Data 2. Explore the Data 3. Ask a specific question Once the data scientists asks a specific question, an exploratory data analysis becomes a research project starting at step #4 (see above).
Metadata	Metadata is information about your data set. This usually includes brief descriptions of each attribute. It may also include how the data was collected, data cleaning notes, and other important notes on the data.

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?

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.

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.

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:

   a. Have students complete Part 1, where they design their project using the DS Project Menu & the DS Designing Scaffold.
   b. Have students complete Part 2, where they will plan and implement creating their visualizations, modeling, and analyze their findings.
   c. Have students complete Part 3, where they will share their findings
   d. Have students complete Part 4 (Reflection)

See *Assessment Strategies* below.

Day 2 Outline

Formative Assessment Notes

4. **Warm Up:** Given the following Diamonds Data Set, use the 11X17 graph paper and exploratory analysis to support or disprove:

   "The bigger the diamond the better it is."

Students should identify attributes including clarity, cut, and color.

5. **Class Exploratory Data Analysis Project:** Group students into pairs or groups and provide students the Grocery Store Marketing Analytics Data Set and the Metadata.

   Sample the data set so that each group gets 50 cases. Have each group choose attributes to explore patterns in buyer trends. Have them use the 11X17 Graph Paper to create visualization sketches.

   Using the attributes explored, have students fill out the research question flowchart and distiller in their groups to narrow down their research question.

Have students do a quick check in with you to assess their understanding on their distiller.

6. **Exploratory Project Practice:** Have students complete the steps below to complete an exploratory data science project:

   a. Have students complete Part 1, where they design their project using the DS Project Menu & the DS Designing Scaffold.
   b. Have students complete Part 2, where they will plan and implement creating their visualizations, modeling, and analyze their findings.
   c. Have students complete Part 3, where they will share findings
   d. Have students complete Part 4 (Reflection)

See *Assessment Strategies* below.

Practice Project Rubric

	Exemplary	***Proficient***	***Developing***
Question Formulation and Project Design	The research question is one that can be thoroughly answered with data Research question is relevant with real-world applications Question is clearly communicated	Research question is well communicated but cannot be properly answered with data science OR Research question is well communicated and can be answered, but is irrelevant to the real world OR Question is relevant, but is not clearly communicated	The question is communicated
Data Selection and Preparation	Substantial data is selected from a reputable source Selected data corresponds with the question You have vetted the data set to avoid issues	Appropriate data is selected, but the data set is not large enough to reliably answer your research question OR Appropriate data is selected, but is unreliable OR The data selected is reliable and substantial, but irrelevant.	Data is selected.
Visualizations	Multiple visualizations communicate project findings Visualizations are clear, concise, and well explained Visualizations are appropriate for the data	Exactly one visualization communicates project findings OR One or more visualizations are unclear or poorly labeled, but are present and appropriate OR Choice of one or more visualizations are not suited to the data, but other visualizations demonstrate findings	Visualizations are missing or are invalid.
Models	An accurate, predictive mathematical model is created to help answer the research question. The model is accurately interpreted	A mathematical model is created with small errors OR A mathematical model is created, but cannot be used to answer the research equation	A mathematical model with substantial errors in accuracy, applicability, and explanation is created.
Communication	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.	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 A substantial portion of the write up is unclear	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 small data sets or 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 in telling your data 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 meets the goals you set out above.
- **Create Models:** Create your model using graph paper or another appropriate tool.
- **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 the visualizations you created by hand, 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.

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 some exploration in CODAP, possibly using an extended version of the data.

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.

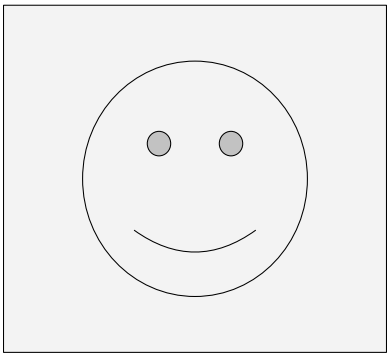


Figure 1. A smiley face

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.

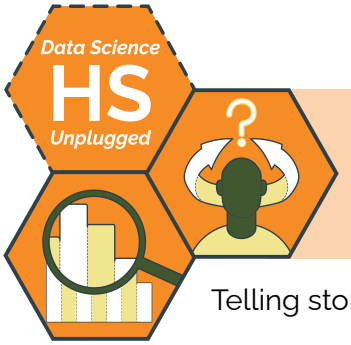
Metadata

Data Set: [Grocery Store Marketing Analytics Data Set]

Attribute Descriptors

Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5th campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on *gold* products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase

Table 1: Meta-data table



Unplugged Data Science

Telling stories with real-world data by Sara Fergus

Summary

In this project, students will apply what they have learned into a real-world example. They will use data to uncover something that people may not be aware of, and come up with a solution or a way of telling people about what you discovered. Students will create artifacts to demonstrate their skills, like data analysis write-ups, visualizations and models, or "fix" proposals describing what they uncovered and suggesting solutions.

Objectives

The students will be able to . . .

- Construct a good Data Question for a research project or an exploratory data analysis
- Collect or acquire data to help answer the Data Question
- Select and create appropriate visualizations to communicate patterns in the data
- Select and analyze models to understand trends in data and make predictions
- Summarize a data analysis and communicate findings

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.9** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.

Materials

- Supplies for creating creative visualizations, e.g. markers, construction paper, poster board, etc.
- The student-facing *Project Frame* document (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)
- *Literature* (see description & write-up below) & *Lunar Myth* (see description & write-up below) 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 see what a successful project looks 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.

Unplugged Data Science Project Prompt

Explore a topic or phenomenon that people in your school or community experience. For example, you may choose something that you see as a problem within your school, or in a public place in your community. It could be a topic that your community experiences together, or an issue that you think people in your school or community should know more about. Use the idea worksheet to help you come up with an idea. Conduct a data analysis to make visible aspects which were previously invisible. Once your data analysis is complete, create a presentation that communicates your work to an audience. If your topic highlights a problem, suggest a "fix" to help make the structure better (more fun, more accessible, more equitable, etc.).

Explore the examples to get a better idea of what you could do!

- Unplugged Project: Literature
- Unplugged Project: Lunar Moods

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

☐ **Collect the Data**: Come up with a method to collect your data. Be sure to think about the ethics of your data collection process, and whether or not your data might be biased.

☐ **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 whether a descriptive or predictive model (e.g., line of best fit, description of patterns in data) can help you in telling 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 mode. 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 during the project, but you may choose to add to the overall infographic using sites like Canva.

- ☐ **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

	Exemplary	***Proficient***	***Developing***
Question Formulation & Project Design	The research question is one that can be thoroughly answered with data Research question is relevant with real-world applications Question is clearly communicated with background information	Research question is well communicated but cannot be properly answered with data science OR Research question is well communicated and can be answered, but is irrelevant to the real world OR Question is not clearly communicated, but is relevant and can be answered with data.	The question is communicated
Data Acquisition & Preparation	Substantial data is selected from a reputable source Selected data corresponds with the research question You have vetted the data set to avoid any "glaring" issues.	Appropriate data is selected, but the data set is not large enough to reliably answer your research question OR Appropriate data is selected, but the data is not taken from a reputable source or has "glaring" issues/ OR The data selected is reliable and substantial, but does not relate to the research question.	Data is acquired.
Visualizations	Multiple visualizations communicate project findings Visualizations are clear, concise, and well explained Visualizations are appropriate for the data	Exactly one visualization communicates project findings OR One or more visualizations are unclear or poorly labeled, but are present and appropriate OR Choice of one or more visualizations are invalid for the data being represented, but multiple visualizations demonstrate findings and are clear	Visualizations are missing or are invalid.

Models	An accurate, predictive or descriptive mathematical model is created to help answer the research question. The model is accurately interpreted in the write up	A mathematical model is created with small errors OR A mathematical model is created, but cannot be used to answer the research equation OR The model is inaccurately or not interpreted in the write up	A mathematical model with substantial errors in accuracy, applicability, and explanation is created.
Communication	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 aesthetic deliverable successfully communicates the question and the findings.	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 More than half of of the pieces of the write up are unclear	Writeup is unclear or is missing a significant amount of essential information
General Project Feedback			

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.

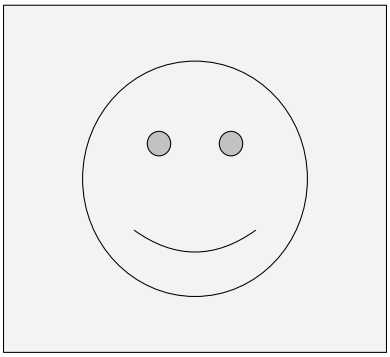


Figure 1. A smiley face

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.

Required Reading Lists Example Project 📖

This project is an unplugged basic exploratory data analysis.

Classroom Highlights

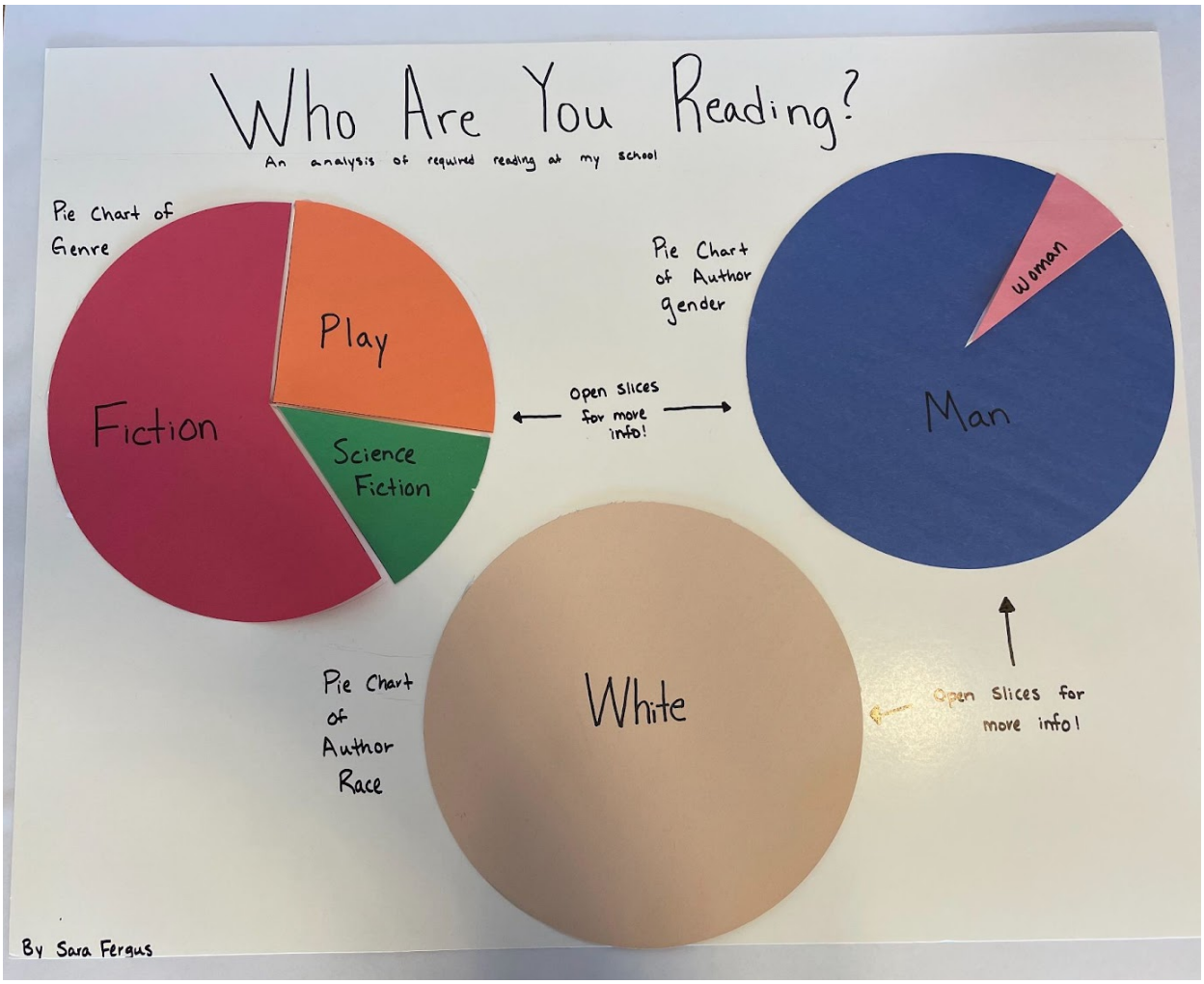
This project example demonstrates:

1. Unplugged Data Science
2. Interactive visualization
3. Interactive *and* adaptive visualization
4. Emerging visualization (map)
5. Data Science in a local context
6. A call-to-action in response to findings

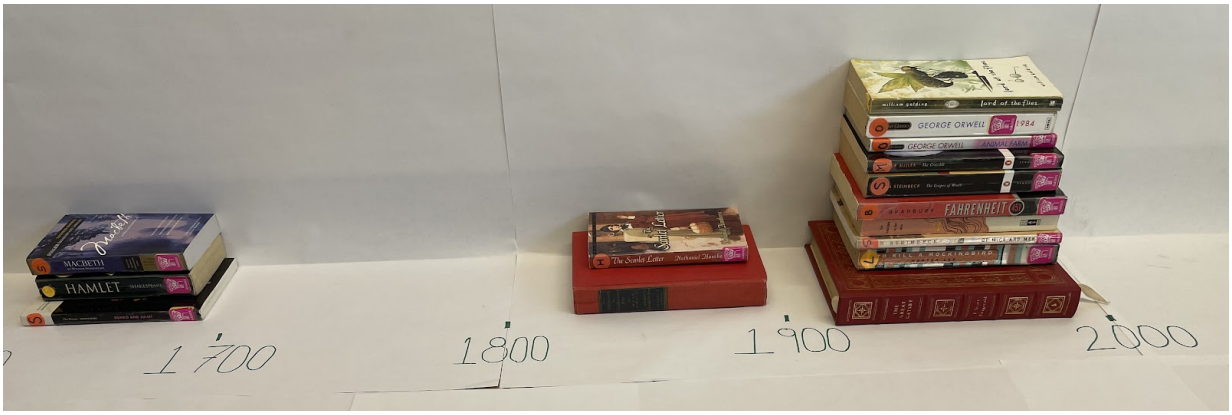
The Project

In this project, I explored the demographic data associated with common required reading books and their authors. A student might choose to do a project like this using their school's required reading lists. Placing data science in a local context helps students to engage with the material through place-based learning, grounding their learning in a real situation with real potential for impact. A student could use a project like this to examine the common narratives, themes, and topics that learners in their school are asked to engage with, and advocate for broadening required reading lists or including important new topics that seem to be missing. When I created this example, I didn't use books from a particular school; instead, I just used a contrived list of common required reading (They were books shelved as "required reading" on GoodReads).

I looked at 5 dimensions: genre and year of publication of the book, and race, gender, and nationality of the author. To visualize author race, author gender, and book genre, I created interactive pie charts. When my viewer lifts a slice of the pie they will find the books/authors in that slice. I was lucky to be able to use diverse skin-color construction paper, since it really helps demonstrate the findings.

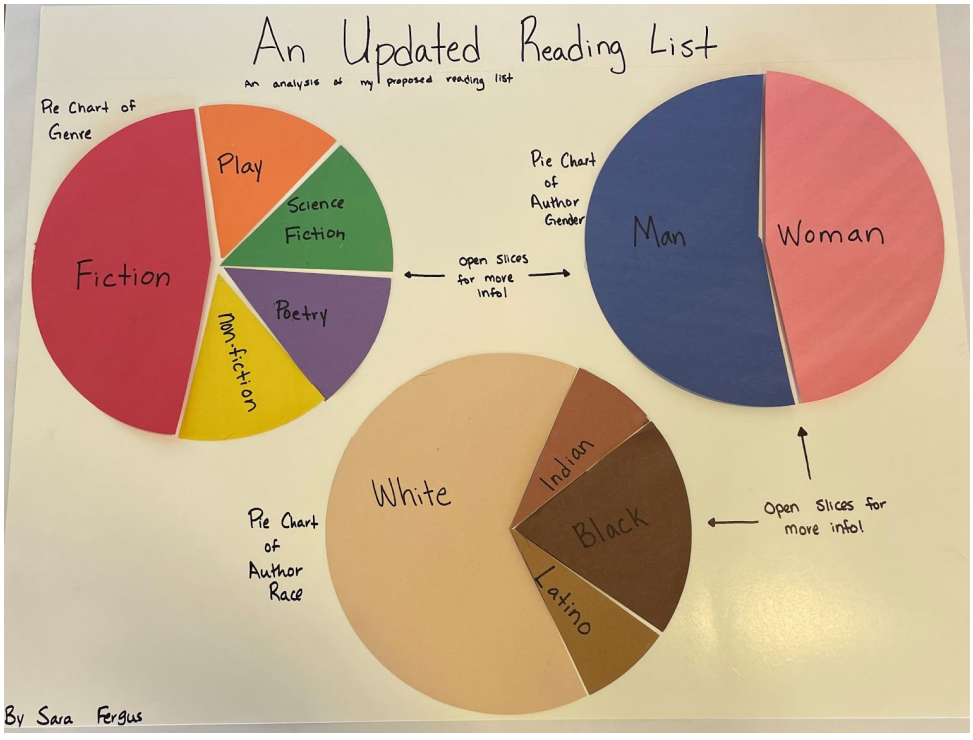


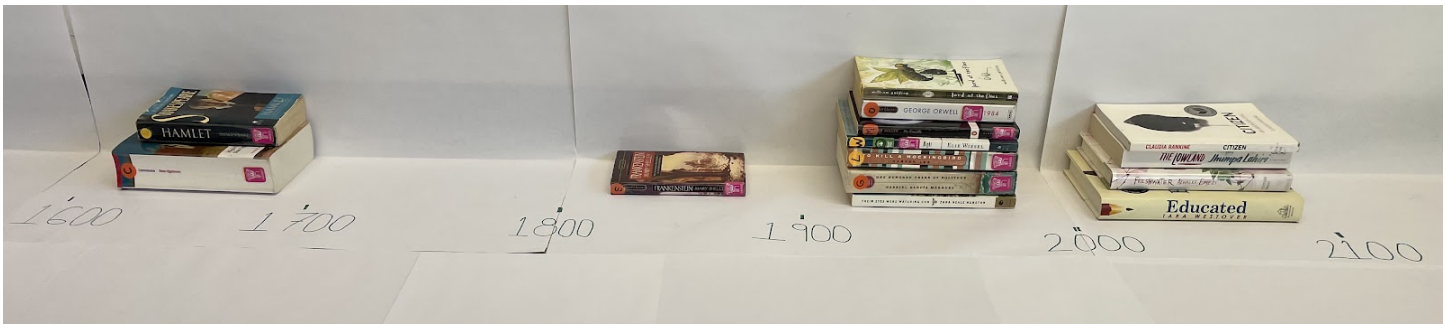
To visualize the year of publication for each book, I created an interactive histogram by stacking books according to century of publication. This allowed the visualization to change as curriculum changes--the histogram is simply a stack of required books! It also allows viewers to physically see and manipulate the histogram. It's not *totally* accurate, since the books are different sizes, but I found that the benefits of learning with the manipulatives outweigh the costs. Plus, the physical visualization has a high degree of coherence; it really shows the data *using* the data.



Finally, I created a map to show the authors' nationality:

I found that the required reading list I created was missing a lot of diversity, and could use an update. If a student were to do this project, they may stop here with the analysis and proceed to the follow-up. However, if they are "above-and-beyond" students, they may continue trying to dig deeper to provide some recommendations to fix issues they found in their data. After I explored the data, I created a new reading list in an attempt to diversify the curriculum. To ensure that the new list was balanced, I analyzed the new list in exactly the same way:





The key to exploratory data analysis is the follow-up (this is especially true with controversial topics). My follow up included a report and a call to action. In my report, I explained the reasons for the problems with the reading list in the first place, and argued that the new list still holds the same important lessons that the "traditional" one did while also being a much richer representation of the diversity of authorship in the literary canon. I also explored what it might mean to change a reading curriculum, and found that hurdles for teachers are a huge issue (write a whole new curriculum?? No way). I addressed these issues with some suggestions and resources to help with diversifying required reading lists. You can check out the whole report here.

The Power of Words Example Write-Up 📖

An analysis of my school's required reading books and their authors

Research Question and Background

While classic literature is studied for a variety of reasons, and English curriculums are built on thoughtful decision making, it is common for the diversity in these books to be lacking. This is something that has been found all across the country and has been reported on many times, like by Insider and Harvard Education. It can be difficult, however, to see diversity (or lack of diversity) at first glance. So, this is something I would like to take a look at. This analysis addresses the question: "What is the demographic breakdown of authors of required reading books in my school?". I am defining demographics on a few measures: the gender, race, and nationality of the author, the genre of the book, and the year of publication.

The lack of diversity in high school reading lists is something that many people have already considered. In fact, there are groups, such as Diversify Our Narrative and the AdLit Diverse Books Project, who are committed to updating reading lists to be more racially and otherwise diverse. Diversify Your Narrative even provides curriculum resources to help teachers make the change. I hope to use my data analysis to pinpoint where our curriculum can be diversified, and then work with these groups to suggest change.

Data and Data Collection

I used my school's required reading list as the basis of my data. The list is included in Table 1. I obtained this list from the Department Head of the English Department.

- To Kill a Mockingbird, Harper Lee
- The Great Gatsby, F. Scott Fitzgerald
- Romeo and Juliet, William Shakespeare
- Lord of the Flies, William Golding
- The Scarlet Letter, Nathaniel Hawthorne
- The Catcher in the Rye, J.D. Salinger
- Of Mice and Men, John Steinbeck
- The Grapes of Wrath, John Steinbeck
- Hamlet, William Shakespeare
- Animal Farm, George Orwell
- The Crucible, Arthur Miller
- Fahrenheit 451, Ray Bradbury
- Nineteen Eighty-Four, George Orwell
- Macbeth, William Shakespeare
- The Adventures of Huckleberry Finn, Mark Twain

Information about each book is included within the book and is public knowledge. Here's the data:

Title	*Author*	*Author Nationality*	*Author Race*	*Author Gender*	*Genre*	*Year of Publication*
To Kill a Mockingbird	Harper Lee	United States	White	Woman	Literary Fiction	1960
The Great Gatsby	F. Scott Fitzgerald	United States	White	Man	Literary Fiction	1925
Romeo and Juliet	William Shakespeare	United Kingdom	White	Man	Play	1597
Lord of the Flies	William Golding	United Kingdom	White	Man	Literary Fiction	1954
The Scarlet Letter	Nathaniel Hawthorne	United States	White	Man	Literary Fiction	1850
The Catcher in the Rye	J.D. Salinger	United States	White	Man	Literary Fiction	1951
Of Mice and Men	John Steinbeck	United States	White	Man	Literary Fiction	1937
Hamlet	William Shakespeare	United States	White	Man	Play	1600
Animal Farm	George Orwell	United Kingdom	White	Man	Literary Fiction	1945
The Crucible	Arthur Miller	United States	White	Man	Play	1953
Fahrenheit 541	Ray Bradbury	United States	White	Man	Science Fiction	1953
Nineteen Eighty-Four	George Orwell	United Kingdom	White	Man	Science Fiction	1949
Macbeth	William Shakespeare	United Kingdom	White	Man	Play	1623
Huckleberry Finn	Mark Twain	United States	White	Man	Literary Fiction	1884
The Grapes of Wrath	John Steinbeck	United States	White	Man	Literary Fiction	1939

Data Analysis

To analyze my data, I visualized each country of origin on a map, I created a histogram for the year of publication, and I created pie charts to show the breakdown of other attributes of the data. From there, I was able to make some adjustments and then analyze again.

The Current Reading List

My data analysis focused on five aspects of the books: Country of Publication, Race of Author, Gender of Author, Genre, and Year of Publication. Figure 1 shows a geographic representation of the Country of Publication. Here, we see that all required reading novels were published in either the United States or the United Kingdom.

Figure 1. Map of Author's Country of Origin

Figure 2b (bottom of figure) shows a pie chart which represents the author's race. All authors are white. Figure 2c (Top Right) shows a pie chart of the author's gender. Only one author, Harper Lee (To Kill a Mockingbird), is not a man. Figure 2a (Top Left) is a pie chart representing genre. Over half of the books are Literary Fiction, there are four plays, and two Science Fiction novels.

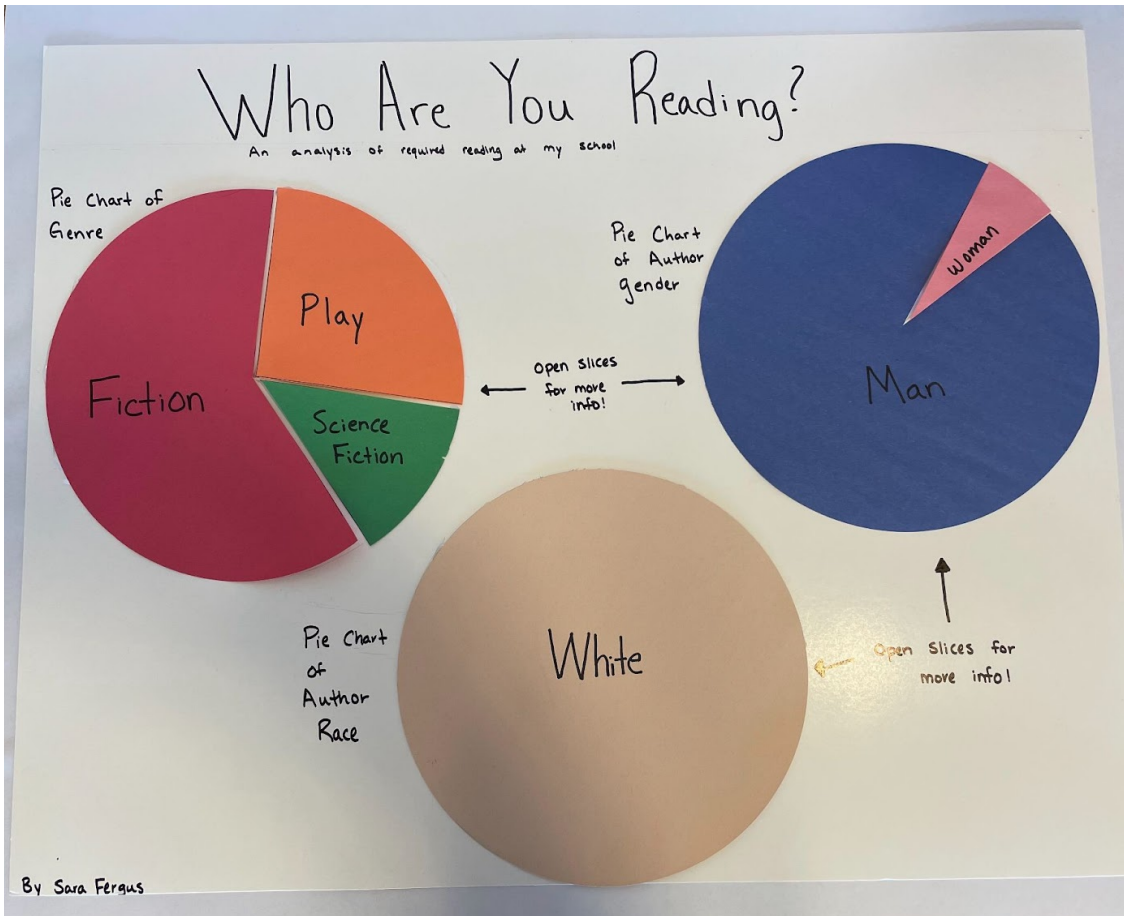


Figure 2. (a) Pie Chart of Book Genre (b) Pie Chart of Author Race (c) Pie Chart of Author Gender

Finally, Figure 3 shows a histogram of the year of publication of these books. There is a spread of about 360 years, with the majority of books falling between 1900 and 2000 (C.E.).

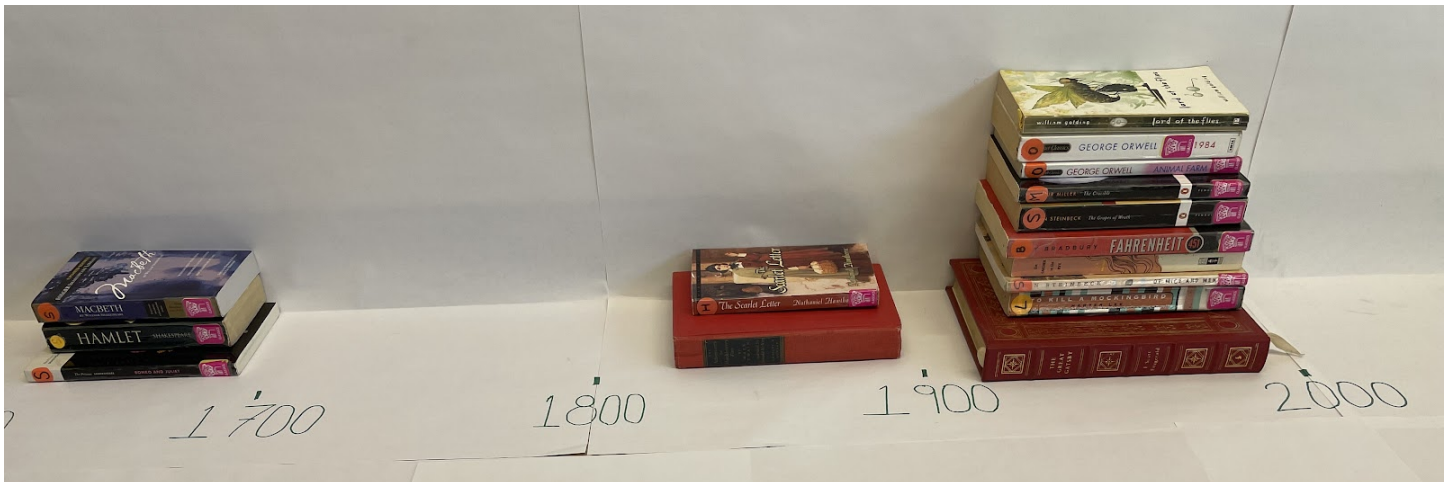


Figure 3. Histogram of Publication Year

My Proposed Changes

After conducting my data analysis, I created an updated required reading list to address some of the underrepresentation in the curriculum. After some research, I decided on the list in Table 3.

- To Kill a Mockingbird, Harper Lee
- Lord of the Flies, William Golding
- Hamlet, William Shakespeare
- The Crucible, Arthur Miller
- Nineteen Eighty-Four, George Orwell
- The Odyssey, Homer
- Night, Elie Wiesel
- Their Eyes were Watching God, Zora Neale Hurston
- Lowland, Jhumpa Lahiri
- Educated, Tara Westover
- Citizen, Claudia Rankine
- Freshwater, Akwaeka Emezi
- One Hundred Years of Solitude, Gabriel Garcia Marquez
- Don Quixote, Miguel de Cervantes
- Frankenstein, Mary Shelley

Table 3. Updated Reading List

I collected information in the same way I did for the initial reading list. Here is the updated info:

Title	Author	Author Nationality	Author Race	Author Gender	Genre	Year
To Kill a Mockingbird	Harper Lee	United States	White	Woman	Literary Fiction	1960
Lord of the Flies	William Golding	United Kingdom	White	Man	Literary Fiction	1954
Hamlet	William Shakespeare	United States	White	Man	Play	1600
The Crucible	Arthur Miller	United States	White	Man	Play	1953
Nineteen Eighty-Four	George Orwell	United Kingdom	White	Man	Science Fiction	1949
The Odyssey	Homer	Greece	White	Man	Poetry	-800
Night	Elie Wiesel	Romania	White	Man	Nonfiction	1956
Their Eyes Were Watching God	Zora Neale Hurston	United States	Black	Woman	Literary Fiction	1937
The Lowland	Jhumpa Lahiri	India	Indian	Woman	Literary Fiction	2013
Educated	Tara Westover	United States	White	Woman	Nonfiction	2018
Citizen	Claudia Rankine	United States	Black	Woman	Poetry	2014
Freshwater	Akwaeka Emezi	Nigeria	Black	Woman	Literary Fiction	2018
One Hundred Years of Solitude	Gabriel Garcia Marquez	Colombia	Latino	Man	Literary Fiction	1967
Don Quixote	Miguel de Cervantes	Spain	White	Man	Literary Fiction	1605
Frankenstein	Mary Shelley	United Kingdom	White	Woman	Science Fiction	1818

Figure 4 shows the Countries of Publication after my proposed changes. As you can see, I added books from Colombia, Spain, Nigeria, Romania, Greece, and India. There is still a heavy emphasis on literature from the United States.

Figure 4. Map of Author's Country of Origin, Proposed List

Figure 5b (Bottom Center) shows a pie chart representing the author's race. I add some books by Black American and African authors, as well as one book by an Indian author (Jhumpa Lahiri, *The Lowland*) and one book by a Hispanic author (Gabriel Garcia Marquez, *One Hundred Years of Solitude*). For this analysis, I categorized Homer (Greek, *The Odyssey*), Miguel De Cervantes (Spanish, *Don Quixote*), and Elie Wiesel (Romanian, *Night*) as "white". In the original reading list, 100% of authors were white. In the new list, only 66.7% of authors are white.

Figure 5c (Top Right) shows a pie chart of author gender. After changes, about 47% of authors are women; there is only one more male author (8 authors) than female (7 authors). Figure 5a (Top Left) shows genre. I added a few genres: "Nonfiction" (*Educated* and *Night*) and "Poetry" (*The Odyssey* and *Citizen*). I halved the proportion of works that are plays from 26.7% to 13.3% by removing multiple plays by William Shakespeare. The proportion of literary fiction in the list decreased from 60.0% to 46.7%.

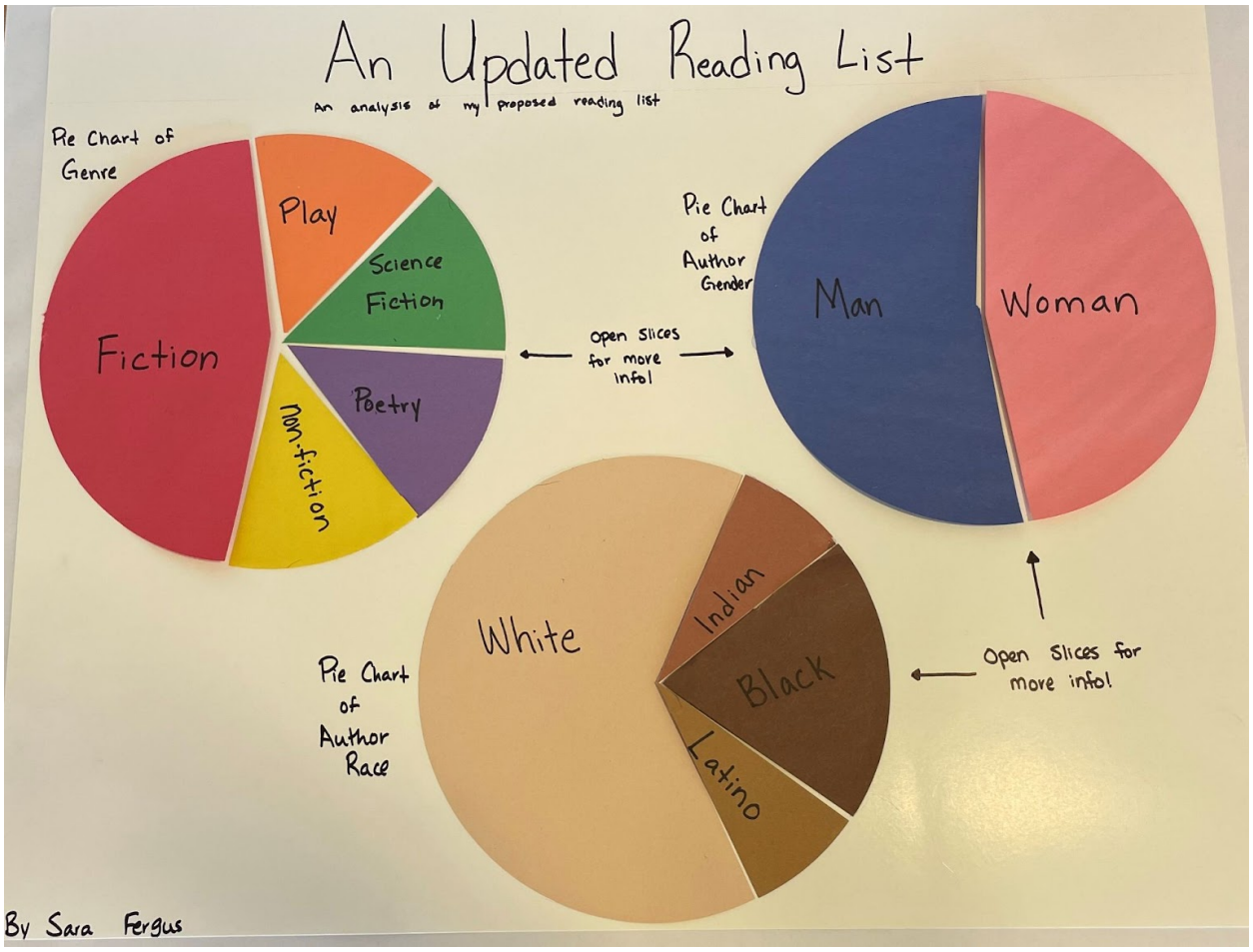


Figure 5. Proposed List

Figure 6 shows a histogram of the year of publication for each book. I omitted one outlier, *The Odyssey*, from the histogram. It is believed to have been first written around the 8th century BC--well before what would be mathematically considered an outlier in this data set. The range of publication among the remaining texts is bigger than it was before (whether or not you include the outlier). I added four modern books published after 2000 (*Educated*, *The Lowland*, *Citizen*, and *Freshwater*). If you include *The Odyssey*, the spread is now over 2800 years. Without *The Odyssey*, *Hamlet* and *Don Quixote* are the oldest books, putting the range at a little over 400 years.

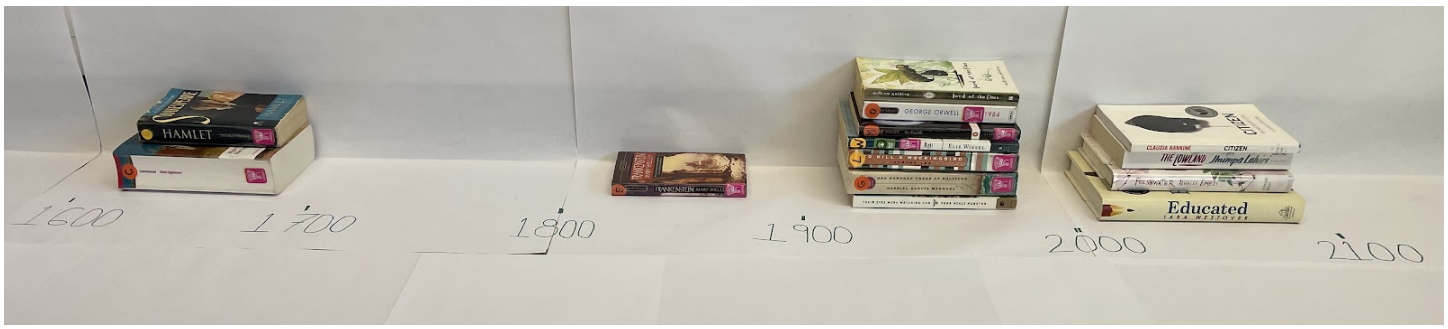


Figure 6. Year of Publication, Proposed List

Findings

My school would definitely benefit from diversifying our reading curriculum. There are so many wonderful authors in the world, and yet we are reading multiple books from the same ones (William Shakespeare, John Steinbeck, George Orwell)! Before my changes, there was not a single non-white author, and only one female. All of the authors were from The United States and the United Kingdom. While this makes sense considering we are limited to books either written in or translated to English, there are far more English-speaking countries that could be represented. There was a relatively large range of publication dates, from Shakespeare in the 15th century to Harper Lee in the 20th. However, there were no books from the 21st century, a time that my peers and I relate to the most. Finally, over half of the books are "Literary/Realistic Fiction". Outside of that genre, there are a few plays and a few *Science Fiction* novels. This leaves out authors that speak in different ways, like through poetry.

After my changes, there is a lot more diverse representation. There are now nearly as many women authors as there are men, and there are a number of non-white authors (Black, Indian, and Latino) which now make up a third of the curriculum. I added examples of books written in Spain, Colombia, Nigeria, Romania, Greece, and India. There is still an emphasis on books from the United States, but I find that this makes sense, since these books will be the most relatable to me and many of my classmates. I added three books from the 21st century, without eliminating the historical perspective we get by reading books from other time periods. Finally, while "Literary Fiction" still makes up almost half of the curriculum, and there are still examples of "Science Fiction" and plays, I added genres like "Poetry" and "Non-Fiction".

Conclusion

The demographic breakdown of required reading in my school is lacking diversity in race, gender, nationality, year of publication, and genre. It is important to see ourselves reflected in our literature, but as written, the while, male students see this a lot more than the rest of the student body. It is also harder to see yourself reflected in time periods before your own--there are no books in the curriculum written in the 21st century, the time period my peers and I have spent our whole lives in. It is important to use literature to learn about different times and different people, but right now this is only true for learning about people from the United States or the United Kingdom. To follow my research question, I have made some suggestions. I have written a "Call to Action Letter" to the English Department to let them know about these findings and provide suggestions.

Call to Action

Dear English Department,

I have been taking a look at the required reading selections in our school through data analysis. While I know that the books that have been chosen are chosen for good reasons, I noticed that there is a bit of a lack of diversity in authors that we are reading, in genre, and in author nationality. Further, there is a lack of literature from the generation of me and my peers. In my data analysis, I found that the majority of the authors we read are white men from the United Kingdom or United States. While I understand that we need books in English so the United Kingdom and United States make sense, I know that other English-speaking countries, like India and Nigeria, are underrepresented. While I understand that there are more white male authors to choose from, particularly in books from a long time ago, I believe that there is a way to increase diversity. I am proposing a number of changes to the required reading curriculum:

- *To Kill a Mockingbird*, Harper Lee
- *Lord of the Flies*, William Golding
- *Hamlet*, William Shakespeare
- *The Crucible*, Arthur Miller
- *Nineteen Eighty-Four*, George Orwell
- *The Odyssey*, Homer
- *Night*, Elie Wiesel
- *Their Eyes were Watching God*, Zora Neale Hurston
- *Lowland,* Jhumpa Lahiri
- *Educated*, Tara Westover
- *Citizen,* Claudia Rankine
- *Freshwater*, Akwaeka Emezi
- *One Hundred Years of Solitude*, Gabriel Garcia Marquez
- *Don Quixote*, Miguel de Cervantes
- *Frankenstein*, Mary Shelley

My data analysis shows that this new list greatly increases diversity of race, gender, nationality, and genre. It also adds more recent works without compromising the range and diversity of publication date that was already in place.

As you can see, the first five were already included in the school's curriculum. *The Odyssey*, *Night*, *Their Eyes Were Watching God*, *One Hundred Years of Solitude*, *Don Quixote*, and *Frankenstien* are all very common required reading books. To me, that means that people with expertise in literature have deemed them good to study, and there will also be helpful resources for teachers! I added a few beyond these: *Educated* is a memoir of an American woman who grew up in a family denying her an education. *Citizen* is a collection of modern poems by the American poet Claudia Rankine. *Freshwater* is a modern novel written by a Nigerian woman, Akwaeka Emezi.

To make space for these new books, I did need to cut a few from the current curriculum. I felt that only one full Shakespeare was necessary, so I cut *Romeo and Juliet* and *Macbeth*. Similarly, I removed *Animal Farm* since *Nineteen-Eighty Four*, also by George Orwell, was already in the curriculum. The other books I cut were *The Scarlet Letter*, *The Catcher in the Rye*, *Of Mice and Men*, *Fahrenheit 451*, *The Adventures of Huckleberry Finn*, and *The Grapes of Wrath*. I felt comfortable removing *Fahrenheit 451*, since the dystopian theme is covered in *Nineteen Eighty Four*. *The Adventures of Huckleberry Finn* and *The Scarlet Letter* were both 19th century American Fiction Novels, a time period which is covered with *The Crucible*. *Frankenstein* is also from that time period. *Of Mice and Men* and *The Grapes of Wrath* are both 20th century American novels, which are covered with *To Kill a Mockingbird* and *Their Eyes Were Watching God*. *One Hundred Years of Solitude* is also from that time period.

While my list is not perfect, I do feel that it adds diversity to what we learn by maintaining most of the important things covered in the books we read now. I hope that you take my new list into consideration, though I understand that I am not an expert in High School Literature, so my list may not be adopted exactly. There are a lot of supports for this sort of change, such as Diversify Our Narrative, that could help teachers with the curriculum change. Thank you for your consideration!

The Lunar Myth Example Project 🌙

This project is an unplugged basic exploratory data analysis.

Classroom Highlights

This exemplar demonstrates:

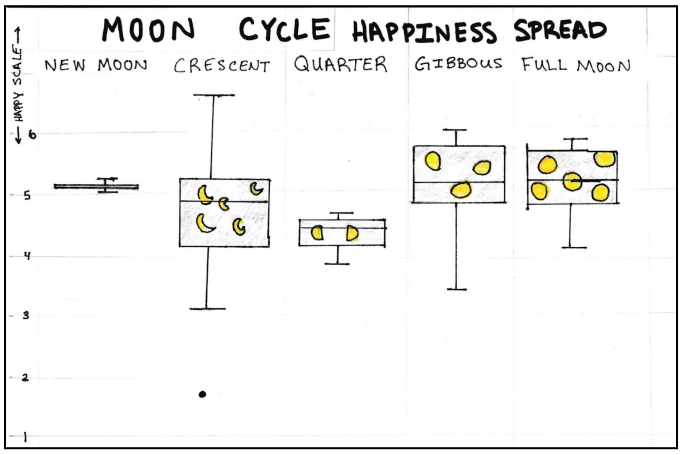
1. Unplugged Data Science
2. Finding no relationship
3. Data Tools like the hedonometer
4. A physical visualization
5. The opportunity for student generated questioning of outliers and of the size of data
6. Representations for categorical and numerical data.

The Project

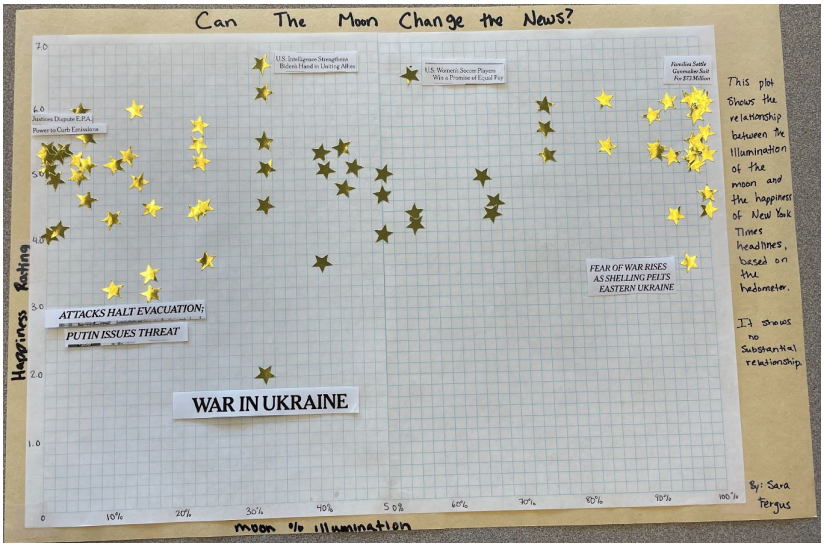
In this example project, I looked into the lunar effect, which is the idea that human behavior changes when the moon is full. A lot of people believe that this is the case (at least jokingly). In fact, teachers in my own school often contribute outlandish student behavior to the full moon. So, over the course of about a month, I noted the phases of the moon, the percent illumination, and the "happiness" of New York Times headlines for the day. I measured happiness using the average hedonometer score for each word in the title.

To communicate my findings, I created two visualizations. These two visualizations contrast numerical and categorical data by presenting what is essentially the same information in different ways. I created a comparative box-and-whisker plot relating numerical happiness to the categorical phase of the moon, and then a scatterplot relating the percent illumination of the moon and numerical happiness.

My first categorical representation was a physical box and whisker plot showing the spread of happiness for each phase of the moon. It appears that there is no substantial relationship between moon phase and "happiness" in the newspaper headlines. There was not a lot of data in general; this was particularly true for full and new moons. This lack of data for an important measurement is something that I explored in the write-up. The most extreme spreads are in the categories with the least data. Prompting students to think about why that is can lead them to a stronger understanding of ideas like outliers, significance, and sample size.



The second visualization I created was a scatterplot showing the (lack of) relationship between illumination percent and happiness. Again, we can see that that the box-and-whisker was created for moon phase, a categorical variable, and a scatterplot was created for a continuous numerical variable, percent illumination. I labeled some major points (for example, *War in Ukraine* had a much lower happiness score than any of the other headlines) in the scatterplot to add another dimension to the data story. In the end, there appears to be no relationship.



In the write up, we conducted some follow-up research to answer the question of why so many people believe in the "lunar effect", even though we didn't find that it had any effect.

Teaching Notes

A Null 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 interesting sites of inquiry, and that they should not be afraid to investigate ideas they do not already know about. Importantly, you should encourage students to investigate and identify these null relationships; not every project needs to uncover a new phenomenon, and finding *no* pattern can be just as interesting and surprising as finding a relationship.

Numerical or Categorical?

Students sometimes struggle with determining what visualizations to make based on the type of data that they have. Even more so, they struggle with the idea of *cleaning* data to afford them flexibility in data type. For example, I had a student who was working with salary data and wanted to get an idea of what ranges of salaries are most common. She felt that a pie chart would help her in visualizing this. However, her data was numerical salaries. Together, we were able to group her data to allow her to explore categorical representations by adding a column for salary range which, based on the person's salary, could be categories like "40k - 60k" or "Under 20k". This allowed her to get a good understanding of what sorts of data are required for certain visualizations (She first tried a pie chart with the numerical data. It did not look so great) and explore ways she can manipulate the data she has to answer different kinds of questions.

The Lunar Myth Example Write-Up

An analysis of the phases of the moon and its effect on news.

Research Question and Background

Werewolves are not the only ones keeping track of the moon. Many people think that the phases of the moon affect our behavior. Medical professionals will tell you that more people hurt themselves when the moon is full. Teachers will tell you that students act out more with a full moon. This idea is called the "lunar effect." Scientific American describes the lunar effect as the persistent idea that human and animal behavior is affected by the moon. In this data analysis, I will be exploring the idea of the lunar effect by determining whether the phase of the moon has an effect on human behavior, as measured by the news reported in the New York Times.

Data and Data Collection

I chose to measure the moon in two ways: first by the percent illumination and then by the phase of the moon. This information was provided each day by "Moon Giant" and confirmed with The Weather Channel. I chose to consider both measurements since the lunar effect is mostly linked to the "phase" of the moon, but in terms of the portion of moon showing, there is variation within a phase, and repeats between phases (a waning gibbous and a waxing gibbous may have the same portion of the moon showing).

The behavior of people around the world was measured by the "happiness" of the New York Times headlines. Three headlines were used from each day, all taken from the front page. The "happiness" was measured using a "hedonometer", a measure created by the University of Vermont.

Most of the data included in this study were collected between February 14 and March 10, 2022. One additional date, January 17 2022, was included as well to provide more data on the days with full moons. It is important to note that there were some important events in this range that would have a major effect on the news. For example, the Russian-Ukrainian War began in this time period.

Data Analysis

The data was analyzed in two ways. First, the relationship between headline happiness and the percent illumination of the moon was analyzed. This did not take phase specifically into account. Figure 1 shows a scatterplot of this data. A few notable points are marked. For example, the New York Times announced war in Ukraine, a particularly unhappy title, on February 25, 2022, when the moon illumination was 32%. That same day, the Times wrote a particularly positive article, *U.S. Intelligence Strengthens Biden's Hand in Uniting Allies*. There was also a particularly happy announcement on February 23, when the moon was illuminated 55%, that *U.S. Women's soccer players win a promise of equal pay*. A few more notable points are labeled as well.

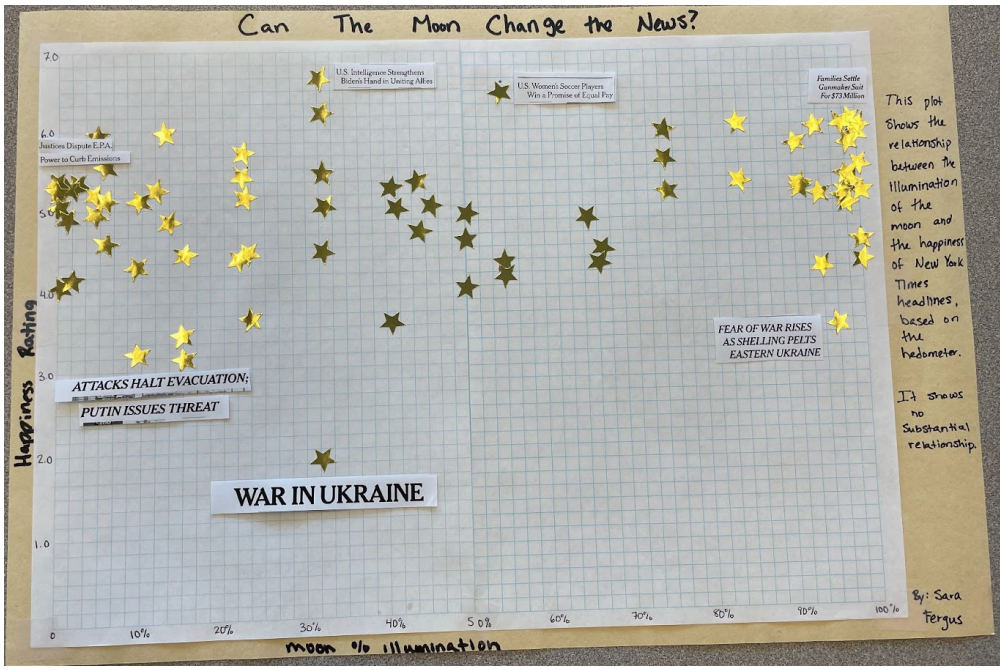


Figure 1. Happiness Scores and Moon Illumination

The second perspective considered the average and the spread of happiness by moon phase. In this case, phases were separate despite their illumination (Waxing Gibbous and Waning Gibbous are different, even though the portion of the moon showing is similar). Findings are shown in Figure 2.

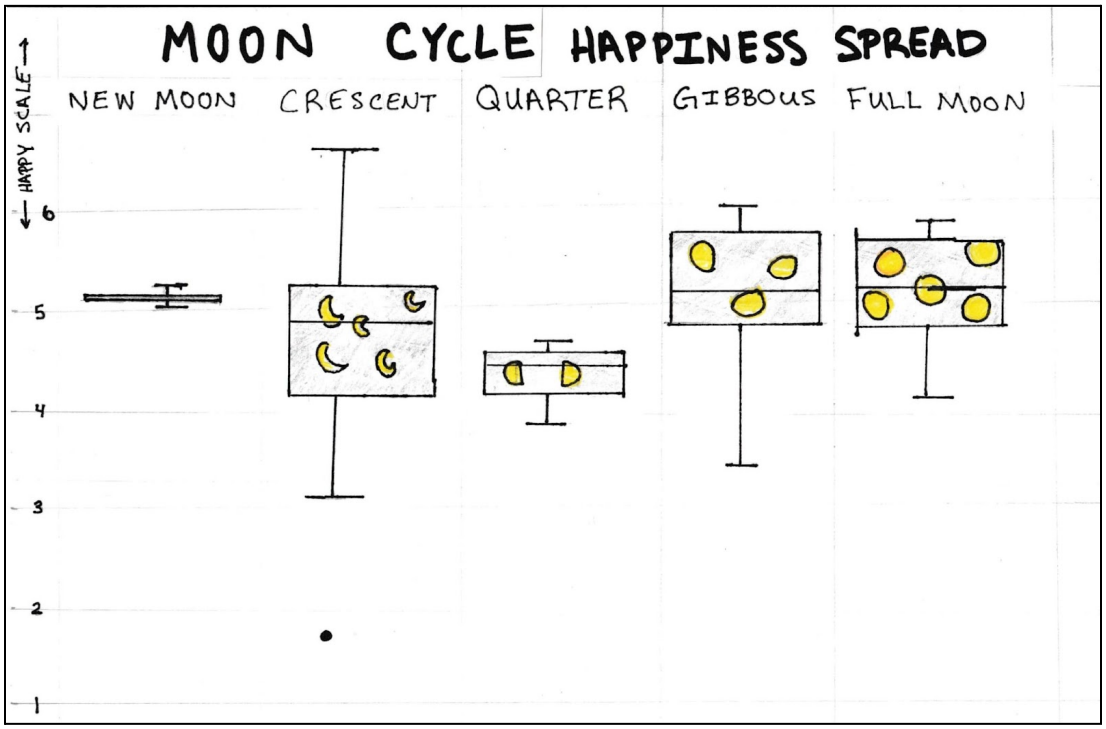


Figure 2. Spread of Happiness by Phase

Here, we see that on a Full Moon, the distribution is relatively normal. Additionally, the average, interquartile range, and range are all similar to what is seen in most other moon phases. The spread of data for a New Moon is notably small, this is likely because the least data was collected during a New Moon phase. The Waning Crescent has a particularly large range, this is because on February 25, 2022, the day that both the happiest article (*U.S. Intelligence Strengthens Biden's Hand in Uniting Allies*) and the least happy (*War in Ukraine*) fell on a Waning Crescent.

Findings

This data analysis shows no real relationship between the happiness of the New York Times and either the phase or percent illumination of the moon. No notable relationship is found either in the analysis of the phase or of the percent illumination. Further, no unusual patterns were found for *any* phase of the moon or percent illumination. This is with the exception of a Waning Crescent, which has a particularly large spread because of a singular day, February 25, 2022.

Most scientific communities do not believe in the lunar effect. The Scientific American article previously mentioned continues to describe the idea as an "illusory correlation", or an association we make that doesn't actually exist. They credit this particular illusory correlation to the human tendency to notice events. Then something weird happens during a full moon, we notice the connection. This skews our perception, since we fail to consider the full moon periods where nothing strange does happen, and when there is no full moon we don't think to connect an odd occurrence to the moon. This idea is confirmed with my data analysis.

Conclusion

This data analysis could be improved in a few ways, the most impactful of which would be collecting more data. With only about one month of data, 79 data points, outliers had a strong pull on the results and relationships may not appear as clear as they would with more data. Additionally, only the titles of the articles were analyzed. This analysis could be strengthened by analyzing all of the words in the article, to balance a particularly "sensationalist" title.

The "lunar effect" could also be explored in other capacities. For example, a local newspaper could be used, or people could rate their moods on a given day. However, even with improvements or changes, I believe that it is unlikely that the idea of the lunar effect, the moon changing human behavior, is a reality. While it would be great to have a reason for the wonderful, or the terrible, or the unique events in our lives, the phase of the moon is very likely not the answer.