Data Science HS with Python

Data Science with Python

A sequence about Python 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 the Python programming language. Through a mix of "plugged" and "unplugged" activities, students engaging with the lessons in this sequence will learn about different types of data, create and evaluate data visualizations, and identify patterns in data sets using descriptive statistics, modeling, and other techniques. At the end of the sequence, students generate an original research question based on their individual goals, interests, needs, and desires and address it using the data science skills they have developed over the course of the sequence.

Throughout the sequence, students engage in many different knowledge-making activities, including small and large group discussions, guided and independent 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, sensors, public data sets, crowdsourcing) to generate useful information to address data questions
- Interpret, create, and evaluate data visualizations using Python libraries & frameworks (e.g., matplotlib) 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 and decision tree models) 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.

CodeVA CS Lesson Sequence

View online!
codeva.info/ds-python

Data Science Standards Alignment

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

Lesson Name	***Data Science Standard*** *(see VDOE site for full text)*												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
01 What is Data?*						✓				✓			
02 Data in Python						✓				✓			
03 Types of Data	✓			✓		✓							
04 Collecting Data	✓	✓		✓									
05 Preparing Data				✓				✓				✓	
06 Power of Visualizations*						✓				✓			
07 Choosing Visualizations*	✓					✓				✓			
08 Creating Visualizations						✓							✓
09 Descriptive Statistics										✓		✓	✓
10 Creating Models									✓		✓	✓	✓
11 Making Predictions									✓		✓		
12 Overfitting & Noise							✓		✓				
13 Decision Tree Models						✓		✓	✓			✓	✓
14 Understanding Research*			✓		✓								
15 Developing Questions*	✓	✓			✓								
16 Project Practice	✓	✓	✓			✓			✓				✓
17 Summative Project	✓	✓			✓			✓	✓			✓	✓

**Denotes an "unplugged" lesson*

Many 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. These "unplugged" lessons help students develop **conceptual understanding** which they then apply during the "plugged" lessons. The "unplugged" lessons in this sequence also appear in a fully "unplugged" version of this sequence; you can find CodeVA's unplugged data science sequence on CodeVA's GoOpenVA profile: https://goopenva.org/profile/18746.

Computer Science Standards Crosswalk

This lesson sequence makes use of **Python** as its primary data science analysis tool. As such, this lesson provides some opportunities for students to engage in computer science learning as they build their data science skills. The chart below provides a crosswalk which teachers can use to identify points of integration between data science and computer science throughout the sequence.

Note that this curriculum does not cover all of the Computer Science Fundamentals Virginia Standards of Learning.

Lesson Name	*DS Standards*	*CSF Standards*	*Rationale*
01 What is Data?*	DS. 6, DS.10	CSF.9-11	Students analyze data visualizations and consider the role of data in daily life
02 Data in Python	DS. 6, DS.10	CSF.12-13, 15a, 16-17, 19, 21a	Students write programs in Python using variables & libraries
03 Types of Data	DS.1, DS.4, DS.6	CSF.9, 12-13, 15a, 16-17, 21a	Students write programs in Python & discuss data representation issues
04 Collecting Data	DS.1, DS.2, DS.4	CSF.1a, 12-13, 15a, 16-19, 21	Students write code that stores data from sensors in variables & files using libraries
05 Preparing Data	DS.4, DS.8, DS.12	CSF.9, 12-13, 15a, 20, 21a	Students write programs in Python & discuss data representation issues
06 Power of Visualizations*	DS.6, DS.10	CSF.10-11, 23, 25	Students analyze data visualizations for patterns in real-life data
07 Choosing Visualizations*	DS.1, DS.6, DS.10	CSF.10-11, 23, 25	Students analyze data visualizations for patterns in real-life data
08 Creating Visualizations	DS.6, DS.13	CSF.10-13, 15a, 16, 19-21a	Students use Python libraries to create data visualizations
09 Descriptive Statistics	DS.10, DS.12, DS.13	CSF.10-13, 15a, 16, 18-21a	Students use Python libraries to calculate descriptive statistics of data sets
10 Creating Models	DS.9, DS.11, DS.12, DS.13	CSF.10-13, 15a, 16, 18-21a	Students use Python libraries to create regression models
11 Making Predictions	DS.9, DS.11	CSF.10-13, 15a, 16, 18-21a	Students use Python libraries to create regression models & make predictions
12 Overfitting & Noise*	DS.7, DS.9	CSF.11, 23, 25	Students analyze overfit models and how they may amplify/reinforce human bias
13 Decision Tree Models	DS.6, DS.8, DS.9, DS.12, DS.13	CSF.10-13, 15a, 16, 18-21a	Students use Python libraries to create decision-tree predictive models
14 Developing Questions*	DS.1, DS.2	N/A	N/A
15 Understanding Research*	DS.3, DS.5	N/A	N/A

Materials

- Access to the Desmos mathematics education platform
- Access to the *Resources* linked in the CodeVA Curriculum Google Drive
- A printer (color is best, but black and white will work)
- Handheld whiteboards, or a similar tool for students to respond to questions from their seats
- Various craft supplies, including sticky notes & poster board; see individual lessons for itemized lists
- An integrated development environment that supports running & editing Python programs and notebook (.ipynb) files (see *A Note on Python IDEs*)
- Access to the Kaggle data science web resource

Student Prerequisite Knowledge & Skills

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

Computer Science Prerequisite Knowledge

- Can read, write, trace, and debug programs written in a text or block-based programming language which include a.) variables of different types (numbers, strings, objects), b.) arithmetic operations on those variables, c.) calls methods on objects based on a library API or framework, and d.) generates text output via a "print" command or similar function/method (e.g., programming competencies aligned to the 5th grade Virginia Standards of Learning for Computer Science)
- Can explain, in an abstract way, how computers store data via sensors and/or human input.
- Can explain, in an abstract way, how computers can automate decision-making processes or data analysis processes

Mathematics Prerequisite Knowledge

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

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

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

Teacher Prerequisite Knowledge & Skills

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

Pedagogical Skills/Knowledge

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

Content Area Skills/Knowledge

- Can read, write, trace, and debug Python code using libraries including pandas, scipy, matplotlib, and other data science Python tooling
- Can use a local or cloud-based Python IDE to edit, modify, run, and create Python notebooks
- Can teach students the basics of program execution, digital data storage, filesystems, and other computing concepts
- 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
- Can connect data science concepts & topics to their impact on social issues, including systemic bias, communication, propaganda, and injustice.

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

Scope & Sequence

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

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 Data in Python	This hands-on guide explores different computational expressions of data, including human-readable formats, data frames, and tables.	DS. 6, DS.10
03 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
04 Collecting Data	In this lesson, students will learn the basics of collecting data using a variety of Python tools, including user input, APIs, and text scraping.	DS.1, DS.2, DS.4
05 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.12
06 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
07 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
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 in Python, and practice generating visualizations.	DS.6, DS.13
09 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.10, DS.12, DS.13
10 Creating Models	In this 4-day 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, DS.12, DS.13
11 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
12 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
13 Decision Tree Models	In this lesson, students read and create decision trees. Students will have to identify what sort of datasets are effectively modeled using decision trees, and will have to clean datasets to prepare for modeling.	DS.6, DS.8, DS.9, DS.12, DS.13
14 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.5
Summative Project	See below.	See below

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*
15 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
16 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
17 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.8, DS.9, DS.12, DS.13

These activities are very open-ended, and educators have a lot of flexibility in how they facilitate them. Educators might have students complete the ***15 Developing Questions*** lesson, complete the ***16 Project Practice*** to give students a chance to work on a project with a little more structure, and then have them complete the ***17 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 ***16 Project Practice*** assignment, omitting the open-ended, larger-scale ***17 Summative Project Frame***. Likewise, they might skip ***16 Project Practice*** and go straight to the project frame if students are ready for a more open-ended, self-directed project. Educators could even re-sequence these three lessons, completing the project practice before developing their questions.

Data Science with Python Project Examples

To help educators envision what sorts of projects students might complete, the ***17 Summative Project Frame*** includes example projects, each with detailed documentation of the project work. Educators implementing this curriculum can use these example projects as a way to teach themselves the details of completing the summative project, as a resource for students who might benefit from examples as they plan their individual projects, or as something to compare students' projects to during the summative assessment stage. The table below contains information about each of the examples, and links to the example data science project materials:

Project Name	*Summary*
Political Tweets	Analyze the language politicians use on Twitter to investigate polarization

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: This is identical to 01 Unplugged: What Is Data? *from the* Data Science Unplugged *&* Data Science with CODAP *sequences.*

Objectives

The students will be able to . . .

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

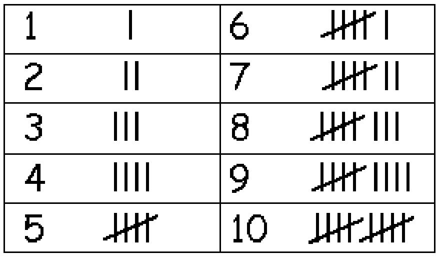
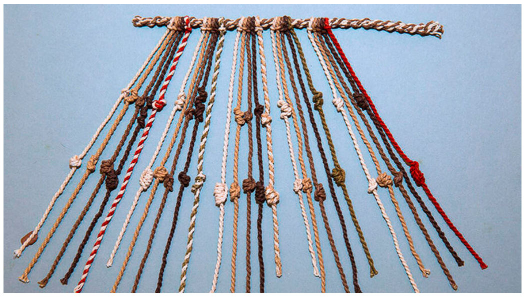
Standards Alignment

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

Materials

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

Vocabulary

<table>
<tr><th>Term</th><th>Definition</th></tr>
<tr><td>Tally Marks</td><td>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.
<table><tr><td>1</td><td>𝍷</td><td>6</td><td>𝍸𝍷</td></tr><tr><td>2</td><td>𝍷𝍷</td><td>7</td><td>𝍸𝍷𝍷</td></tr><tr><td>3</td><td>𝍷𝍷𝍷</td><td>8</td><td>𝍸𝍷𝍷𝍷</td></tr><tr><td>4</td><td>𝍷𝍷𝍷𝍷</td><td>9</td><td>𝍸𝍷𝍷𝍷𝍷</td></tr><tr><td>5</td><td>𝍸</td><td>10</td><td>𝍸𝍸</td></tr></table></td></tr>
<tr><td>Matrix</td><td>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.</td></tr>
<tr><td>Table</td><td>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.</td></tr>
<tr><td>Quipu</td><td>A quipu is a recording device historically used by cultures in South America, including the Inca. The knots in the cords represented numeric values.
</td></tr>
<tr><td>Data Representation</td><td>A data representation is a way to visualize and organize collected information</td></tr>
</table>

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!

CodeVA CS Lesson Plan

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.

Data in Python

An introduction to data representations in a coding context by Sara Fergus

Summary

This lesson is an introduction to computational data representation as preparation for data analysis using the Python coding language. This hands-on guide explores different computational expressions of data, including raw and human-readable formats, data frames, and aggregated/filtered tables. Students will be able to use vocabulary to talk about data in the context of the computer and will be able to search for, upload, and store data effectively.

Objectives

The students will be able to . . .

- Classify the structures and uses of various data structures and files, including spreadsheets, CSV files, and Dataframes
- Use relevant vocabulary to communicate about aspects of data structures using the words "aggregate", "case" or "record", and "attribute".
- Read, debug, and write code using the Python Pandas library to read data from a CSV file to a Pandas Dataframe

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

- See a note on Python IDEs
- Introduction to Pandas (ipynb)
- Roller Coaster Data (CSV)
- Kaggle Datasets (If your school blocks Kaggle, you can have them search through another method (for example Google Dataset search), or provide them with a class data folder.)
- What is Data? Warm-Up (Desmos, Slides)

Vocabulary

Term	Definition
Spreadsheet	A spreadsheet is a way to represent a table, usually online. Data can be entered into rows and columns. Often, software allows you to manipulate this data within the spreadsheet. Below are two examples of spreadsheet software.
Google Sheets	Google sheets is an online spreadsheet tool where data can be stored in rows and columns, as a table. Google sheet links directly with data gathered by Google Forms.
Microsoft Excel	Microsoft Excel is a spreadsheet software that can store and manipulate data in table form.
Table	A way to represent data points with more than two attributes. Each row of a table is known as a case or record, and each column is an attribute.
Aggregate Table	A way to represent data that combines information from multiple cases based on a common attribute. It is usually "adding up" one of the attributes
Case / Record	A "case" or "record" is one row of a table, which represents one entry. All of the attributes in that row belong to the same "case".
Attribute	An "attribute" is a column of a table. It is a piece of information that describes each case. Most of the time, each case will have multiple attributes.
CSV File	A CSV (comma-separated values) file is a way to store data in a computer. In a CSV, each row is a row in a table (a case), and each attribute of that case is separated by a comma.
Pandas	PANDAS in a Python library created to help with data science by allowing easy data storage and manipulation. A library is a collection of pre-made functions that can be used in coding. When using a library, it is important to refer directly to the documentation. Here is the pandas library documentation.
Pandas dataframe	A pandas dataframe is a data structure that stores information in a format similar to a table. It works similarly to a Python dictionary (*don't worry if you don't know what that means*).

Outline

Formative Assessment Notes

1. **Warm Up:** Have students respond to these two slides in their journal or on paper. Then, have them share out loud. Slides have students compare/contrast and list advantages/disadvantages of representing data using a CSV file, a table, an aggregate CSV file, or an aggregate table.

   At the end, define CSV, aggregate CSV, table, and aggregate table.

Students should notice the following…

- **The table and csv have more information**
- **The tables are easier for a human to read**

2. **Research:** Have students explore Kaggle Datasets for data that they are interested in.

   Once they pick a dataset, have them write in their journal what file type their data is saved as (Excel, CSV, etc.) and a description of their data set.

   Point out that most data sets are available in a CSV format.

Have students share their descriptions. In describing their data sets, they will likely talk about "rows" and "columns" or other descriptors.

3. On the board, write what each person shares in two unlabeled columns. At the end of student sharing, label the columns "case/ record" and "attributes" to introduce vocabulary.

??	??
people	Weight, height, hair color, gender
pokemon	Name, attack strength, defensive strength
Country	Happiness rating, health care accessibility
Day	Temperature, season, weather

   Once you have built the table as a class, discuss:

   Why do you think I have sorted your descriptor words into two categories like this? What is similar among things in each category? What is different between them?

Make sure that students are mentioning cases and attributes (although not necessarily using those words). If they are not, prompt them to take a closer look at the data and describe in more detail.

4. Have students work through the introduction to Pandas worksheet (ipynb), which will show them how to store data in a dataframe. You may choose to work through the first part of the worksheet as a full class, in small groups, or individually Make sure that the worksheet as well as the example data, Roller Coasters.csv, is saved in the same folder of whatever IDE your class is using. (See a note on IDEs here)*

4. **Exit Ticket:** Ask the students to respond to this question:

Think about the dataset you stored into Python. What questions could this data set answer?

Assess exit tickets to make sure students understand cases and attributes

See *Assessment Strategies* below for sample responses

Assessment Strategies

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

Python Worksheet

You may choose to work through the first part of the worksheet as a full class, in small groups, or individually. If you would like for students to turn in this worksheet, they can turn it in as

1. A link to the file *(works for cloud-based IDEs)*
2. Screenshots on a Google Doc or word document *(any IDE)*
3. A screencast of the program running *(any IDE)*
4. A .ipynb file *(Software IDEs, "download as" for cloud-based IDEs)*

You may also have them submit their data file or a link to their data file. Make sure students imported Pandas, correctly used read_csv and correctly extracted one column of the data. Using exploratory commands (e.g. info, describe, head) is not necessary.

Have students use these simple "I can" statements to self-assess the coding commands learned today.

I can . . .

- □ import pandas
- □ import a CSV file
- □ read a CSV file and store it in a descriptively named data frame variable
- □ extract one column of the data and store it in a descriptively named variable

Exit Ticket *(See here for printable copies)*

Name: ____ Date: ________________

Think about the dataset you stored into Python. What questions could this data set answer?

Possible Answers: Responses may vary according to each student's unique data set chosen. For example, if a student chose this data set ______________, some questions could be ____________.

Some Accommodations & Extensions

Students with vision impairments may benefit from having access to the warm up slides on their personal computer. If using desmos, point out the "make full screen" button on the visual.

For students who work more slowly or may become overwhelmed, consider choosing a few Kaggle Datasets ahead of time that they can pick from.

Some IDEs are better for accommodations than others. For example, if using Kaggle you may have the csv file already uploaded to the same folder as the code.

Printable Exit Tickets

Name: ____ Date: ________________

Think about the dataset you stored into Python. What questions could this data set answer?

Name: ____

Think about the dataset you stored into Python. What questions could this data set answer?

Name: ____

Think about the dataset you stored into Python. What questions could this data set answer?

Types of Data

An introduction to data representations in Python & data science 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 02 Types of Data *lesson plans from the CodeVA* Unplugged Data Science *&* Data Science with CODAP *sequences. This lesson includes a Python coding practice activity, which the others omit.*

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"
- Python Worksheet: Types of Data (.ipynb)
- *Types of Data Cheat Sheet* (one per student)

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.*
string	A string is a piece of data that is a word
int	An "int" is a piece of data that is an integer (number without a decimal point)
float	A float is a piece of data that is numeric and has a decimal point
Data Question	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

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?***

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

   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.

Formative Assessment Notes

Students should notice, without vocabulary, that:

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

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.

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"

5. Have students complete the *Python Worksheet: Types of Data*, which relates these types of data to "string", "int", and "float" data structures.

See *Vocabulary* for details

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

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

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

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

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

At the end of the class, you may provide some/all students with this cheat sheet.

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?

Collecting Data with Python

An introduction to generating data using computational tools by Jon Stapleton

Summary

In this optional lesson, students will learn the basics of collecting data using a computer, from collecting keystrokes & user input to downloading data from servers. They will analyze different types of data, consider the role of computing and automation in data collection, and discuss the ethical elements of different data collection practices.

Objectives

The students will be able to . . .

- Read, trace, debug, and write Python programs that a) collect data & store it in variables, b) add that data to a pandas DataFrame, and c) write that data to plaintext file formats (i.e., CSV)
- Explain the difference between different kinds of data using the terms "qualitative", "quantitative", "ordinal", and "categorical".
- Evaluate the potential benefits and harms of using computers to automate data collection.

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.4** The student will be able to identify data biases in the data collection process, and
- understand the implications and privacy issues surrounding data collection and processing.

Materials

- Student journals
- *Python DataFrames Notebook* (.ipynb, digital resource, one per student)
- *Favorite Classes* data (CSV) & *UFO Sightings* data (CSV)
- *Data Collection Station Files* (download once for every collection station, see *Day 2*), or each individually: Station #1 (folder), Station #2 (folder), Station #3 (folder), Station #4 (folder)
- *Data Collection Stations Activity Guide*, one per small student group (2-3 members)

Vocabulary

Term	Definition
pandas	PANDAS is a Python library created to help with data science by allowing easy data storage and manipulation. A library is a collection of pre-made functions that can be used in coding. When using a library, it is important to refer directly to the documentation. Here is the pandas library documentation.
Pandas dataframe	A pandas dataframe is a data structure that stores information in a format similar to a table. It works similarly to a Python dictionary (*don't worry if you don't know what that means*).
CSV File	A CSV (comma-separated values) file is a way to store data in a computer. In a CSV, each row is a row in a table (a case), and each attribute of that case is separated by a comma.

Before the Lesson

Prerequisite Knowledge: You may find that students are most successful with the material in this lesson after a good amount of general Python instruction. Students with some programming experience will be very successful, while students who are not yet proficient with a text-based language will need more support. You can skip this lesson if the students in your class have not taken a programming class yet, or you feel that the open-ended stations activity on Day 2 is too challenging.

Setup: This lesson requires a good amount of setup, especially for *Day 2*. Before starting the lesson, be sure to print out all of the printable materials (or prepare to distribute them digitally). Then, before *Day 2*, download the *Data Collection Station Files* (see *Materials*) to each of the station computers.

Day 1 Outline

		Formative Assessment Notes
1.	**Unplugged Activity:** Prompt students to think about what it would be like to collect data about student life, specifically about the mental health of students who have jobs compared to those who do not—as much data as possible over the course of 3 days. Have students respond to the following questions in their journals, and then share with a peer: ● What data do you need to collect to learn about this issue? ● Where should collectors go to find this data? ● How many collectors would you need to collect data about the whole school? Your entire town?	**Consider facilitating a short class discussion where students share their responses. Make sure students are thinking through the problems of scale when you collect data without using computational tools.**
2.	**Discussion:** Then, discuss the following questions: ***1. What are some problems with having people collect data manually?*** ***2. What tasks would a computer need to do to replace people in your data collection plan?***	
3.	Call attention to the fact that people often use computers to automate the data collection process—computers can write things down in files, and can observe using sensors or user input. **Brainstorm** a list of computing devices that record data in school, community, etc., and record student responses in a public place (e.g., sticky notes on a physical or virtual whiteboard).	**Students might bring up ethical issues during this discussion—encourage them to think about these topics, as they'll be engaging with them later.**
4.	**Coding Practice:** Have students work in pairs to complete the *Python DataFrames Notebook*, which covers the following: ● Review reading CSV data into pandas DataFrames, ● Modify those DataFrames ● Write DataFrames to a CSV file to record data. After some practice, students will analyze the UFO dataset for data types (qualitative, quantitative, ordinal, and/or categorical).	**Consider having students periodically share their data type labels with the class to help students who are stuck on those questions. The set contains several kinds of data!**
5.	**Exit Ticket:** Ask students to complete the *Day 1 Exit Ticket* (see *Assessment Strategies* below). Their responses will serve as a starting point for discussions in *Day 2*'s lesson.	**See *Assessment Strategies* below for details.**

Day 2 Outline

Formative Assessment Notes

6. **Before Class:** Open the appropriate notebook files from the *[Data Collection Station Files]* on each of four station computers. Install the following Python packages on each station using `pip` (if needed, depending on your IDE):

- `pip install pandas`
- `pip install requests`
- `pip install keyboard`

If you'd like to have smaller groups and more than four stations, feel free to create duplicates. The computers will stay at the stations as students rotate between them.

7. **Ethics Discussion Warm-Up:** Have students randomly select (anonymized) *Day 1* exit tickets, or display several exit tickets to the class. Facilitate a discussion about students' responses and how some data collection practices are ethical, and some may not be.

A good question to ask to get at students' reasoning is: "Why do you think this person said this method is ethical/unethical?"

If you have trouble finding good examples from student responses, here are some options to get the conversation going:

- Monitoring student internet use to learn their interests
- Using cameras to learn where students go during school
- Hiding microphones around the school to monitor bullying

8. **Data Collection Stations:** Put students into small groups and distribute the *[Data Collection Station Activity Guide]*. For each station, students will:

This is a great moment to float around and re-teach Python concepts from *Day 1* to smaller groups of students.

- Identify the type of data the scripts in the station collect
- Write a new Python script to collect a different kind of data & write it to a new file
- Come up with one good use for their collection strategy (i.e., what question their collection strategy could address)

Consider setting up a "parking lot" for Python or data science questions that come up during the activity so you can address common issues with groups or the whole class, if needed.

By the end, each group should have developed scripts to collect qualitative, quantitative, ordinal, and categorical data. Here are the topics provisioned in the *[Data Collection Station Files]*:

- **[Station 1]**: Record user input data, write to file (starts qualitative)
- **[Station 2]**: Record keystrokes, write to file (starts ordinal)
- **[Station 3]**: Text analysis (starts quantitative)
- **[Station 4]**: Record data from an API (starts categorical)

Note: Some of these are really difficult to make qualitative data with!

9.	**Assessment:** Complete one or both of the assessments below: the *Data Collection System Design Discussion*, and the *Automated Data Collection Project* as an optional extension. If you choose to facilitate the project, you may need to extend it into a third day to give students enough time.	**See *Assessment Strategies* below for details.**

Assessment Strategies

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

Day 1 Exit Ticket *(See here for printable copies)*

Name: ______________________________ Date: ______________

Using computers allows people to collect lots of data at once, but it's not always right to do it. Sometimes, collecting data from people in an automated way violates their privacy or causes other problems data scientists should avoid.

1. **What is one useful & ethical example of automated data collection?**

2. **What is one *unethical* example of automated data collection?**

Data Collection System Design Discussion

Have students respond to the following prompts in their journals. Then, have students share their reflections with a partner. Finally, facilitate a class discussion about the topic.

Think of some data that is collected from you and stored using a computing system:

- *Where do you think that data is stored?*
- *Who owns the data?*
- *Is collecting & storing the data ethical, or unethical?*

Automated Data Collection Project

Complete the *Data Collection System Design Discussion*, then ask students to prototype their system using the station code as a model.

1. Put students into small groups, or have them continue with their station group
2. Have students come up with a new data question based on the collection techniques available
3. Have students modify lesson code to create a computing system that collects the data
4. Have students make predictions about what patterns they think the data will contain
5. *Optional*: have students actually collect data using their system over an appropriate period of time

If you complete step #5, consider having students use these data sets for future projects and lesson activities (visualizations, descriptive statistics, modeling, etc). This is kind of an advanced project, and may require more time and attention than you have to spare this early in the sequence. Feel free to skip this project, as students will have plenty of opportunities for self-directed data science projects later on when they have developed more skills.

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 walking students through some of the coding in a "lab lecture" format to introduce them to the programming concepts, especially if some of the students are new to Python.

Assigning parts of the lesson as homework prior to or following the lesson can help students who need additional time to complete the work.

Extensions

Consider starting a conversation about the ethics of data collection by having students discuss and reflect on one of the talks in this TED playlist: https://www.ted.com/playlists/130/the_dark_side_of_data

Consider adapting the *04 Finding & Collecting Data* lesson from the *Data Science with CODAP* sequence so students have another opportunity to gather data for their Python analysis using forms & observations.

Day 1 Exit Ticket

Printable

Name: ________________________________ Date: ______________

Using computers allows people to collect lots of data at once, but it's not always right to do it. Sometimes, collecting data from people in an automated way violates their privacy or causes other problems data scientists should avoid.

3. **What is one useful & ethical example of automated data collection?**

4. **What is one *un*-ethical example of automated data collection?**

Name: ________________________________ Date: ______________

Using computers allows people to collect lots of data at once, but it's not always right to do it. Sometimes, collecting data from people in an automated way violates their privacy or causes other problems data scientists should avoid.

5. **What is one useful & ethical example of automated data collection?**

6. **What is one *un*-ethical example of automated data collection?**

Data Collection Stations Activity Guide (Student)

Right now, there are **four stations** in the classroom, each one using a different data collection method. Your goal is to figure out which type of data each station is collecting (quantitative, qualitative, ordinal, categorical), and add a new script to the notebook collecting a different data type.

1. Run the Scripts in the Notebook

Your notebook contains at least one script that collects data. Run each script in the station's notebook, and follow the instructions the computer provides to you.

2. Determine the Type of Data

The scripts on your station collect quantitative, qualitative, categorical, or ordinal data. Work with your group to figure out which kind of data your station collects by looking at the script and the CSV file! **Record your answer in the table on the other side of this page.**

3. Add a New Script

Using the other scripts in the station's notebook as an example, **add a new script** to the notebook that collects a different kind of data:

1. Create a new (blank) CSV file
2. Write code that a.) loads the blank CSV into a DataFrame, b.) adds new data to the DataFrame, and c.) writes the DataFrame data back to the CSV file at the end of the program
3. Work with your group to fill in the last two columns of the table on the other side of this page for your station
4. Then, wait until the teacher tells you to move on to the next station!

Make sure by the end you've written **one script for each type:** ordinal, categorical, quantitative, qualitative

04 Collecting Data with Python

Data Science with Python

Group Members' Names: ____

	The old scripts collect….	***My new script collects…***	***People could use my script to find out…***
Station #1	*Check each that apply:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	*Choose one:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	
Station #2	*Check each that apply:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	*Choose one:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	
Station #3	*Check each that apply:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	*Choose one:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	
Station #4	*Check each that apply:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	*Choose one:* ☐ Quantitative ☐ Qualitative ☐ Ordinal ☐ Categorical	

Data Science HS with Python

Preparing Data

An activity guide covering data cleaning with Python by Sara Fergus

Summary

In this lesson, students will explore data cleaning techniques including removing unwanted outliers, handling missing data, converting data types, 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 Preparing Data *lessons from the CodeVA* Unplugged Data Science *&* Data Science with CODAP *sequences. This lesson includes data cleaning with Python activities, which are 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, see below)
- Video: Coded Bias (make sure this resource isn't blocked)
- Extension: *How is Face Recognition Surveillance Technology Racist?* (web)
- *Data Cleaning Scenarios* Cards (printable PDF, printed & cut out, 1 per student group)
- Example Data (Messy) (CSV)
- Python Assignment: Data Cleaning (.ipynb)

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. **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 write what they notice and wonder into 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. **Cleaning with Python:** Show tricks in pandas using this Data Cleaning In Python Assignment. Have students follow along.

   At the end, have students upload and clean their own data set (they may use the one from lesson 1).

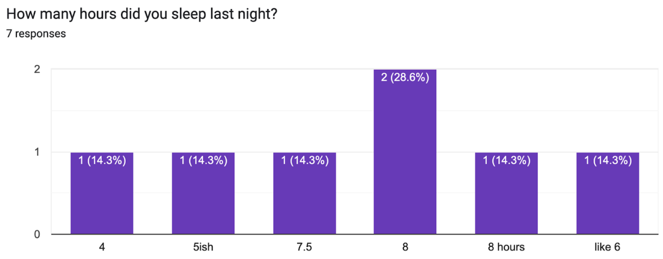
Make sure that students are consistently considering what bias they may be introducing with their data cleaning. See Exit Ticket below.

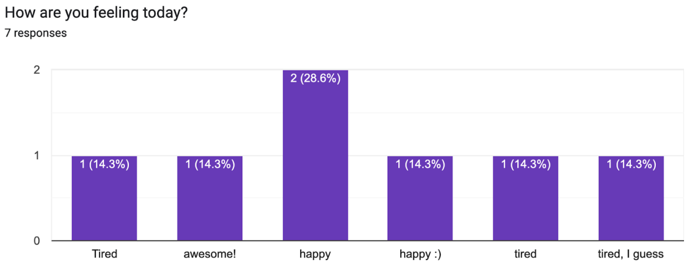
Assessment Strategies

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 *(Print this page & distribute to students)*

Name: ____ Date: ________________

Using the statements below, rate your skills by marking an X on the scaled below:

- I can rename columns ....

  Uneasy Neutral Confident

- I can replace values ....

  Uneasy Neutral Confident

- I can change data to a numeric type ....

  Uneasy Neutral Confident

- I can drop missing values ....

  Uneasy Neutral Confident

- I can drop duplicates ....

  Uneasy Neutral Confident

- I can reformat strings ....

  Uneasy Neutral Confident

Can you think of some bias issues that you or someone else might encounter when they work with data sets like the one you cleaned?

What impacts does this bias have?

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

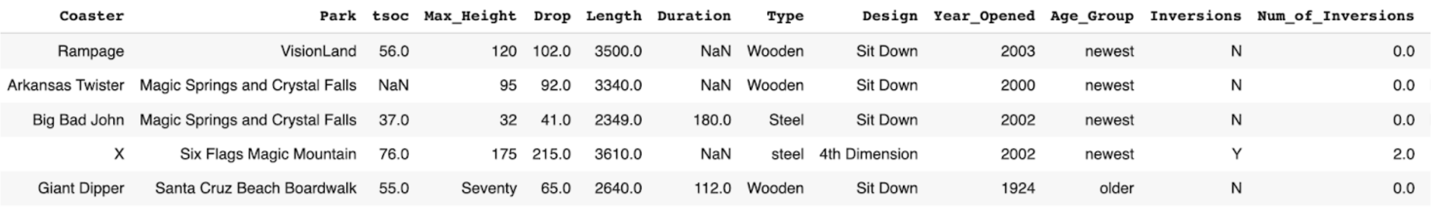
- Consider using the Data Cleaning Guidebook, which is an online PDF book that dives deeper into data cleaning and the errors that could arise
- Consider reading this Note on Python IDEs, a guide to selecting and using various IDEs, and an introduction for how these will be used throughout the CodeVA modules
- 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.

Later, you will see how to clean this data set.

Data Cleaning

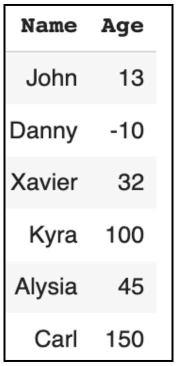
A description of data cleaning considerations (Part 2) by Sara Fergus

Preparing Data Considerations

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

Consideration #1: Is it an error?

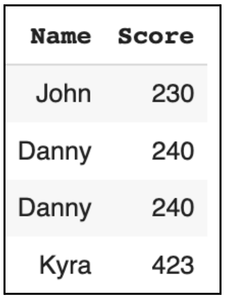
Before doing any data cleaning, it is important to consider whether the changes you are making are actually cleaning an error.



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

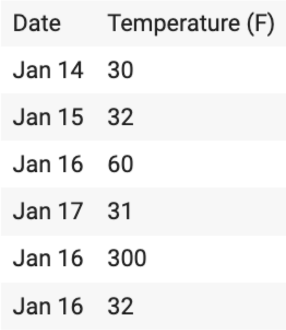
For example, you may have a data set with people's ages (left). This data set is messy because it contains data that doesn't make sense, like someone being -10 years old. Probably, someone typed a minus by accident. You would want to clean that issue. 150 years old also doesn't make sense. Maybe someone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

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



Name	Score
John	230
Danny	240
Danny	240
Kyra	423

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



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

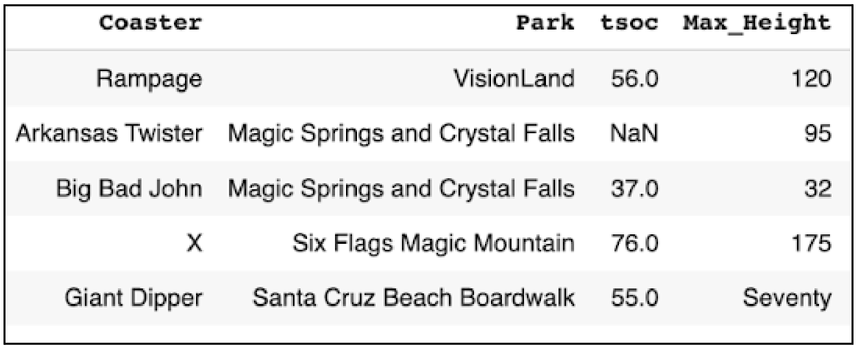
Definitely an error:

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.*



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

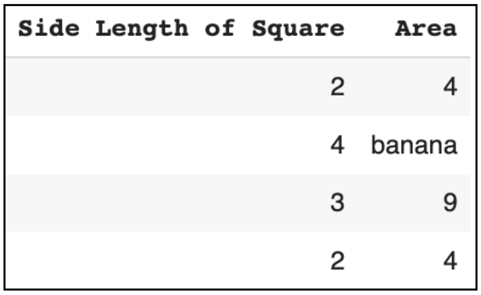
   This works if there is not a lot of data, the errors are easy to see, and what the survey taker meant is obvious. For example, in the data set below, the Giant Dipper has a Max Height of “Seventy”. The computer is going to interpret that to be different than the number 70, but we know that they are the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

   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



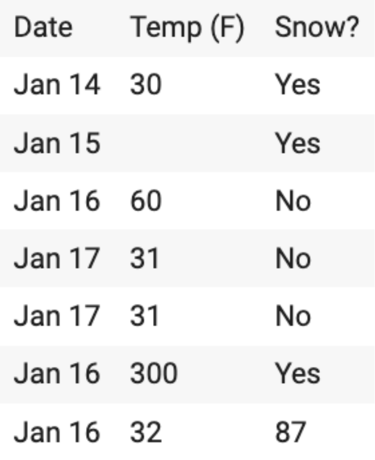
Side Length of Square	Area
2	4
4	banana
3	9
2	4

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 times side, we can calculate and replace “banana” with 16.

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.

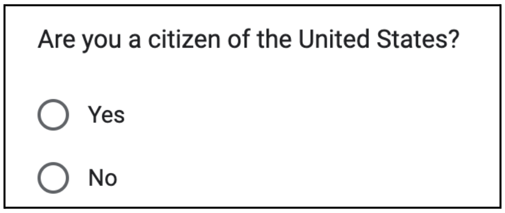
When we analyze missing values, we see two main types of missing data:

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,



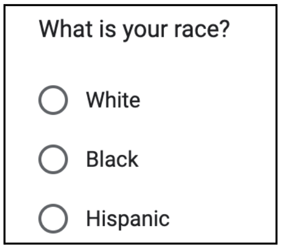
Are you a citizen of the United States?

○ Yes

○ No

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

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



What is your race?

○ White

○ Black

○ Hispanic

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

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

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

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

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

Consideration #4: Do I have enough information left?

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

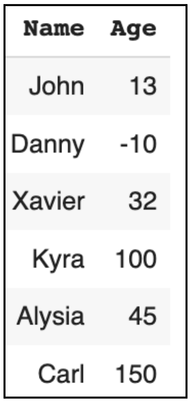
A description of data cleaning considerations (Part 2 ADAPTED) by Sara Fergus

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

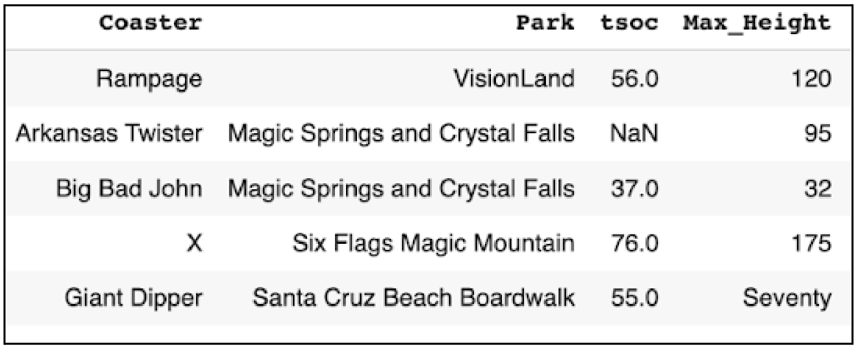
For example, a data scientist should decide if the three unusual ages (-10, 100, and 150) are mistakes or not.

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.

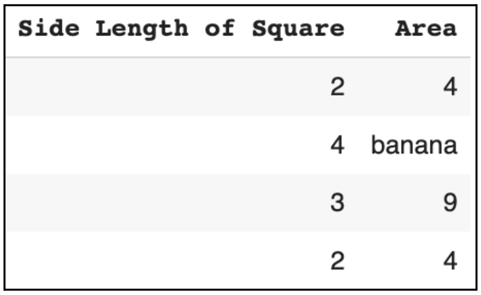


Coaster	Park	tsoc	Max_Height
Rampage	VisionLand	56.0	120
Arkansas Twister	Magic Springs and Crystal Falls	NaN	95
Big Bad John	Magic Springs and Crystal Falls	37.0	32
X	Six Flags Magic Mountain	76.0	175
Giant Dipper	Santa Cruz Beach Boardwalk	55.0	Seventy

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

2. *Replace messy data with 'N/A'.* Often, missing data is represented by 'NaN' or 'NA'. If there is an error, but you want to include the case in general, you could replace the error with N/A.

3. *Use other columns.* Sometimes you can figure out what a value should be based on other columns.



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

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

4. *Get rid of the row.* This is common, but could introduce bias

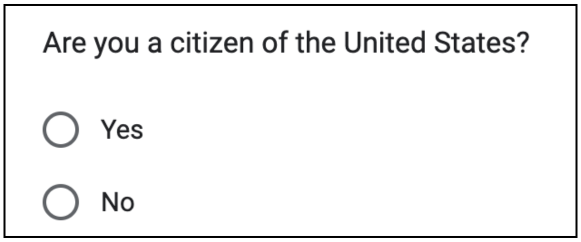
5. *Something else.* There are lots of methods out there!

Consideration #3: Am I introducing any bias?

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

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

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



Are you a citizen of the United States?

○ Yes

○ No

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

Consideration #4: Do I have enough information left?

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

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

Unplugged: The Power of Visualizations

An unplugged 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

Note: This lesson plan is identical to 05 The Power of Visualizations *from CodeVA's* Unplugged Data Science *sequence, and is similar to* 04 The Power of Visualizations *from the* Data Science with CODAP *sequence.*

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 (view Google Slides 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

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 US."

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

	Results for Vote
A	
B	
C	

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

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.

Formative Assessment Notes

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.

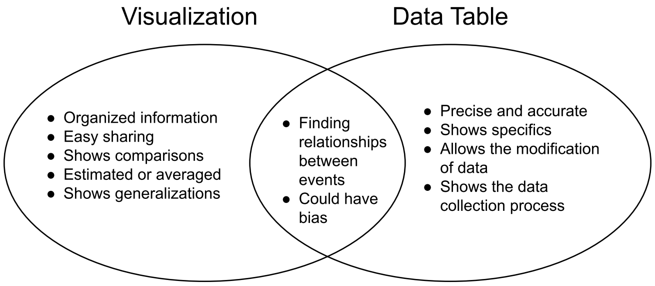
Assess students' ability to choose the most relevant visualization.

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

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

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

Example Result



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

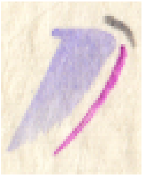
4. **Communicating Interpretation:** Break students into groups of 2–3 (enough to evenly split the 11 examples in the activity).

   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

6. **Warm Up Data Talk:** Use the following data talk to introduce interpreting line graphs - NYT E-cigarettes (Line Graph)

   Be sure to engage students in using the appropriate vocabulary (x-axis, y-axis, title, slope, maximum, minimum, line style, etc.)..

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

   Have students discuss the following prompt in pairs:

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

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

8. **Predicting Google Trends:** Model creating a chart where you predict the shape of the Search for Fortnite trend:

   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

Formative Assessment Notes

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

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

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

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

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

If students have trouble with these questions, provide additional context

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

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

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

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

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

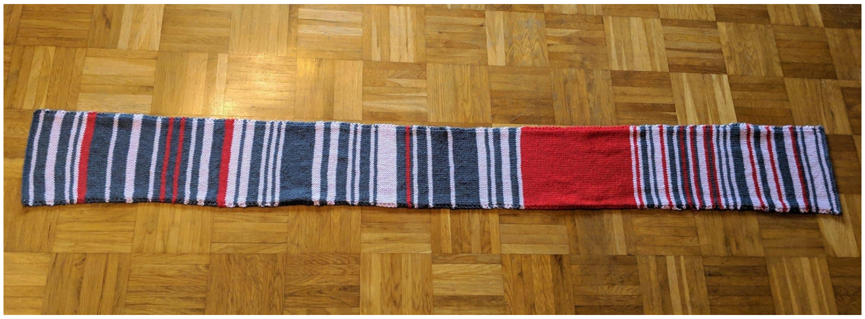
	Proficiency	*Yes*	*No*	*Notes*
Data	The students' choice of data attributes and organization is **mostly insightful** to the collected object itself.			
Narrative	The student created story connects to their collected data using a **mostly coherent** narrative.			
Visual	The student used **suitable visuals**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

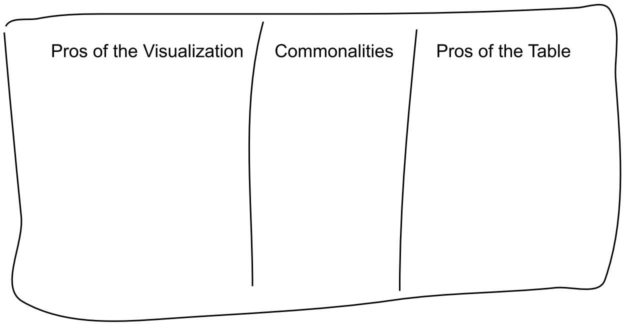
Extensions

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.



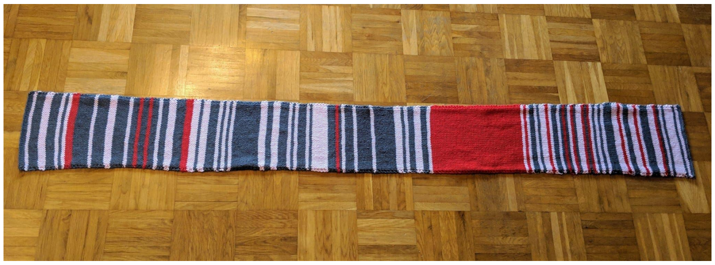
4. Display the question "*What does the visualization accomplish better than the table?"* on the board.
5. Have students write their ideas on sticky notes (one idea per note).
6. Students should place these stickies in no particular order under the Pros of Visualization column.
7. Repeat steps 4-6 with 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 is identical to 06 Unplugged: Choosing Visualizations *from CodeVA's* Data Science Unplugged *sequence, and similar to the* 07 Choosing Visualizations *lesson from the* Data Science with CODAP *sequence.*

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.

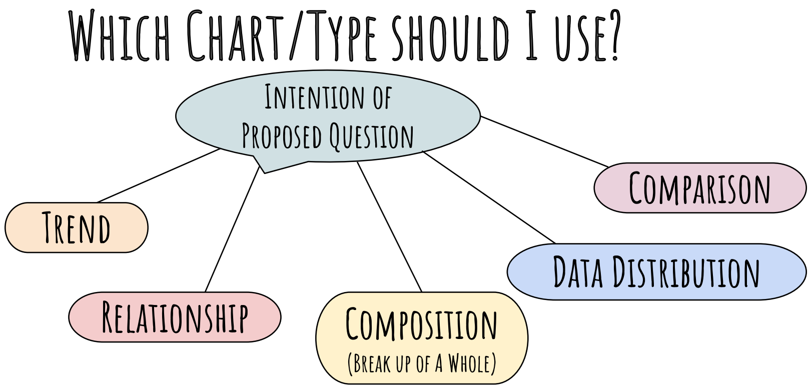
Monitor student responses for the ability to categorize the use of visualization types.

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.


WHICH CHART/TYPE SHOULD I USE?
INTENTION OF PROPOSED QUESTION
TREND
RELATIONSHIP
COMPOSITION (BREAK UP OF A WHOLE)
DATA DISTRIBUTION
COMPARISON

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.

See Assessment Strategies below.

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.

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

Data Science HS with Python

Creating Visualizations

A guide to creating visualizations in Python by Sara Fergus

Summary

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 in Python, and practice generating visualizations using data of their choice. Students will also practice reading Python library documentation.

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

Objectives

The students will . . .

- Students will create line graphs, histograms, bar graphs, pie charts, and scatter plots using matplotlib
- Students will utilize matplotlib documentation to help them customize their graphs
- Students will understand how to utilize other python documentation (e.g. seaborns, numpy) to create graphs not easily created in matplotlib

Standards Alignment

- **DS.6:** The student will justify the design, use and effectiveness of different forms of data visualizations.
- **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

- Creating Visualizations in Python Worksheet (collated)
  - Part 1: Inputting Data (.ipynb)
  - Part 2: Creating Plots (.ipynb)
  - Part 3: Make them look nice (.ipynb)
  - Part 4: Using Subsets (.ipynb)
  - Extension: Line Graphs (.ipynb)
- Visualization Cards (PDF, one per group of 5-6 students)
- Matplotlib.pyplot documentation (check to make sure this isn't blocked)
- Kaggle Datasets (check to make sure this isn't blocked)
- Python Graph Gallery (check to make sure this isn't blocked)
- Roller Coasters Data (CSV)

Vocabulary

Term	Definition
Matplotlib.pyplot	A python library commonly used to create simple data visualizations
Python Library	A set of functions that can be imported into Python and used. Most libraries have functions that all achieve a certain task (statistics, machine learning, data analysis, etc)
Documentation	Documentation is the list of each function in a given library and their descriptions (most often including their return values and parameters)
Parameter or argument	A parameter or argument is passed into a function to tell the function how to behave. For example, in print("hello world"), "hello world" is a parameter telling the print function what to display on the console.
Default parameter or argument	A default parameter is one that can be changed, but is not necessary. These are indicated with an equal sign. For example, plot(x, color = 'red') will plot the values of x, but change the color from the default of blue to red.

Outline

Formative Assessment Notes

1. **Warm Up:** Arrange your classroom into 6 groups, and give each group one of these *Visualization Cards*, which show a survey and data question, and ask students to choose the best visualization.

   Instruct students to work together to choose the best visualization(s), and write in their journal which visualization(s) their group chose and why.

   Once all groups have chosen a visualization, have them hand their card to a different group (alternatively, have the students move to a different table) and repeat with the new card.

   Repeat this process 6 times. Then go through each card together and ask groups to share which visualization they chose. Consider facilitating a discussion about their choices. Some of the data could be represented with different representations (see key).

If there is discrepancy in students' answers, have them share their reasoning, and have the whole class vote. If they select one that doesn't , provide feedback.

If students all give the same answer when there are multiple good answers, prompt them to brainstorm whether any other visualizations would be appropriate.

2. Have students complete the creating visualizations in Python worksheets. The worksheets show them examples of creating visualizations, and then ask them to try it. There is not a lot of direct instruction in the worksheet, encouraging students to make connections between the examples and their data without help.

   After each worksheet, have students a.) share their plots with their peers and help each other, and then b.), have one or two students share their plots with the class.

   - Part 1: Inputting Data (no plots to share after this one)
   - Part 2: Creating Plots
   - Part 3: Make them look nice
   - Part 4: Using Subsets

   After they've finished, have students present their completed work (see *Assessment Strategies* below).

3. **Exit Ticket:** Have students explore the matplotlib.pyplot documentation. Then, ask them to:

   1. Write at least one function or attribute that you didn't use, but would like to try.
   2. Describe a scenario where that attribute or function might be helpful.

Allow time for productive struggle instead of walking them through the steps in the Python worksheet. When having students share plots with the class, have them share what kind of plot it is, what attributes are being represented, and what conclusions they can draw. For parts 3 and 4, have students highlight what they added to their plots.

See *Assessment Strategies* for details.

See *Assessment Strategies* below for details & printable sheets

Assessment Strategies

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

Exit Ticket *(See here for printable copies)*

Name: ______________________________________ Date: ________________

DIRECTIONS: Explore the matplotlib.pyplot documentation.

1. Write at least one function (like title) or parameter (like c=) that you didn't use today, but might want to try in the future.

2. Describe a hypothetical dataset or data question where that parameter or function might be helpful.

Graphical Presentations

After students have completed the four-part visualization work, you can have them present their visualizations to provide a summative assessment opportunity. You may choose to have students share their final graphs with either a peer or the class, or create informal video presentations for you to review later. Use this rubric to guide instructions and expectations for these presentations:

	Proficiency	***Yes***	***No***	***Notes***
Data	The student describes what data they are representing with their visualization. They should mention both the overall dataset and the attributes in particular that they are using. e.g. "I used a dataset that had a lot of different information about pokemon. Each row was a different pokemon. In the visualization I am going to show you, I focused on the attack strength and hit points of the Pokemon."			
Visualization Selection	The student selects a visualization choice that is appropriate for the attributes they are representing.			
Visualization Creation	The visualization that the student shares is properly modified from the default. Students should include a title for any visualization, and then different modifications based on what they are trying to show.			
Visualization Explanation	The student explains what their visualization is showing accurately. e.g. "You can see that the dots are all over the place, with no pattern. That tells me that hit points are unrelated to attack strength"			

Some Accommodations & Extensions

Note: all students benefit from accommodations. Consider providing all students with accommodations from the list below.

Accommodations

During the warm up, you may choose to have all students move to a different table, or to have them pass the data card to the next table. For students with mobility difficulties, you may choose to have students remain seated.

If you choose to have students remain seated during the warm-up, you could assign students who benefit from frequent movement to be the "runner" and bring the card to the next table.

You may choose to provide students with a dataset to follow along with, rather than allowing them to choose their own.

For students who need instructions given in smaller chunks, make sure that the cells in the Python notebook are collapsed, so the student is only looking at one at a time.

For a simpler exit ticket, you may have students explore the matplotlib.pyplot documentation and write what they notice and what they wonder about it.

Extensions

The Python notebook includes an extension about line graphs. Line graphs are more difficult to create on Python, so you may choose to not cover them with the full class. The most straightforward application is time-series data, so encourage students who would benefit from the extension to choose a dataset with a 'year', or other time, attribute.

Printable Exit Tickets

Name: ____ Date: ________________

DIRECTIONS: Explore the matplotlib.pyplot documentation.

1. Write at least one function (like title) or parameter (like c=) that you didn't use today, but might want to try in the future.

2. Describe a hypothetical dataset or data question where that parameter or function might be helpful.

Name: ____ Date: ________________

DIRECTIONS: Explore the matplotlib.pyplot documentation.

1. Write at least one function (like title) or parameter (like c=) that you didn't use today, but might want to try in the future.

2. Describe a hypothetical dataset or data question where that parameter or function might be helpful.

Data Science HS with Python

Calculating Descriptive Statistics

An introduction to descriptive statistics by Christa VanOlst and Sara Fergus

Summary

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

Note: This lesson is similar to the Descriptive Statistics *lessons from the* Data Science Unplugged *&* Unplugged Data Science *sequences. This lesson includes several opportunities to calculate statistics for datasets using Python, which the other lessons omit.*

Objectives

The students will . . .

- Students will calculate descriptive statistics (mean, median, mode, range, & standard deviation).
- Students will analyze and interpret calculations for tendencies in the data.
- Students will use statistical calculations to summarize a dataset.

Standards Alignment

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

Materials

- *Calculating Mean & Median* Worksheet (see below), 1 per student
- *New York Times* Data Talk (Desmos)
- Salaries Resource & Situations Resource, 1 per student group; printed & cut along dotted lines
- Python checklist & datasets (salaries by college type, region, & degree; Starbucks drinks & food)
- *Python Worksheet: Basic Statistics* (.ipynb)
- *Interpreting Descriptive Statistics* Data Talk (Desmos)
- Youtube Video - Descriptive vs. Inferential Statistics
- Exploratory Website - Where does the day go?
- Newspaper Project Template (view or make a copy)

Vocabulary

Term	Definition
Descriptive Statistics	Statistical characteristics of a data set used to identify trends, including mean, median, mode, and standard deviation.
Mean (Average)	The numeric sum, divided by the total number of values in a set
Median	The middle element in a sorted set of values (numeric or ordinal)
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 (numeric)

Day 1 Outline

Formative Assessment Notes

1. **Warm Up Recalling Mean & Median:** Use the Calculating Mean and Median Worksheet to review calculating these statistical measures by hand.

   Summary: Students calculate the mean & median for several sets of numbers

Consider having students work pairs and discuss the last time they calculated these stats to promote team building.

2. **Generating Questions:** Have students individually access this Desmos: New York Times Data Talk by assigning to your class and then using the Desmos built-in select-and-sequence tool.

   Use the Desmos to have students answer the following questions:

   - 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

   As students generate, select as many different noticings/wonders, and the most clear headlines (include a funny one, if there is one) to get them thinking about the initial/guiding questions. After some discussion, pose the following question:

   What is a "normal" salary, according to the data?

Students will be generating their own questions to lead into the lesson. 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?"*

CodeVA CS Lesson Plan

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 using Python:** Use the Python Worksheet: Basic Statistics to learn to calculate and visualize basic statistics in Python.

8. **Guided Practice:** Practice calculating statistics in Python, all together, using the Teacher Directions - Other Descriptive Statistics, where you will model making inferences about a data set using mean, median, and standard deviation.

If students need more time on step #7, you can do this for the *Day 2* warm-up (step #10)

9. **Data Talk:** 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

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

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

OR

Python Skills Check: Review the Python Worksheet: Basic Statistics from the previous lesson, reviewing the skills needed for step #11.

11. **Starting from Scratch in Python:** Give students the following data sets (or data sets of your choice):

    - College Salaries-by-college-type
    - College Salaries-by-region
    - College Salaries-by-degree
    - Starbucks Drinks Data
    - Starbucks Food Data

    Have the students use this checklist (see *Assessment Strategies*) to self assess their ability to put it all together in Python. This activity combines skills students' learned throughout the course and their ability to start the process from the beginning given only a data set.

Skim student worksheets for accuracy throughout the activity.

Engage with students in their findings by relating their results to peers that choose the same or how they could relate to peers that chose different attributes.

12. **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.

13. **Exit Ticket:** Have students complete the ***Exit Ticket***, which asks them to evaluate inferential statistics for misleading statistics.

    OR

    Extension: Use the Teacher Directions - Philosophers Chair Activity Guide to facilitate this activity, where students evaluate whether or not statements using descriptive statistics are misleading or not.

See Assessment Strategies below

Day 3 Outline

Formative Assessment Notes

14. **Project Data Set → News Article:** Have students use the Student Guide - Project News Article to complete the project.

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

See *Assessment Strategies* below for Teacher Directions.

Assessment Strategies

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

Starting from Scratch Python Skills Checklist *(See printable version below)*

Use this checklist to have students self assess their skills learned thus far and their ability to start the analysis process from the beginning given only a data set:

- □ I can upload this dataset into a new Python file
- □ I can clean the file by dropping missing values
  - **Bonus** I can drop only the cases that have missing values in my column of interest
- □ I can identify the sample size of this data set
  - Sample Size = _______
- □ I can access a subset of the data
  - Identify your subset: ____________________
  - Sample size of the subset = _______
- □ I can find the min and max of the column I am using
  - Minimum: ____
  - Maximum: ____
- □ I can calculate the mean of an attribute in the data set
  - Mean = _______
  - What does this mean represent?
    ____
- □ I can calculate the median of an attribute in the data set
  - Median = ______
  - What does this median represent?
    ____
- □ I can calculate the mode of an attribute in the data set
  - mode = _______
  - What does this mode represent? ____
- □ I can calculate the standard deviation of an attribute in the data set
  - Standard Deviation = _______
  - How many cases are within 1 standard deviation of the mean? ________________
- □ I can create at least three different visualizations with my data.
- □ I can summarize my findings in context of the data
  - Summary:
- □ I can save and store my ipynb file in an organized way

Day 2 Exit Ticket *(See here for printable copies)* print the following:

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 scenario 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 a full 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

Project - News Article

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

Student will be required to:

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

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

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

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

During the Project:

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

2. Have students brainstorm a hypothesis/question/problem
   - To streamline the project consider pulling a few datasets and competition prompts from Kaggle Competitions

3. Have students collect their own data using techniques from the course or choosing a preexisting one

4. Have students create at least two visualizations

5. Have students calculate descriptive statistics by answering the following:
   - Describe the size of your sample
   - Describe the center of your data
   - Describe the spread of your data
   - Assess the shape and spread of your data distribution
   - Compare data from different attributes

6. Students will then use their findings as artifacts to aid in writing a short news article using this *Newspaper Template* (view or make a copy) or creating their own

After the Project: Rubric Project News Article

	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

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

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

Extensions

Have students explore the Python statistics package documentation. Have students explore, learn, and show a partner how to use a function from this package.

Worksheet - Predicting and Calculating Mean and Median

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.

Data Set 1: 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 = ______	

Data Set 2: 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 = ______	

Data Set 3: 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.

Teacher Directions - Other Descriptive Statistics

An activity to demonstrate how to calculate measures of central tendency, standard deviation, and covariance/correlation. This activity also shows students how to use functions line max(), min(), and count().

1. Have students create a new Python notebook and upload this dataset. Have students read the data into Python. Then, have them create two data frames– one with only legendary pokemon and one with only pokemon that are not legendary.
   - Provide suggested names (ex. pokemon_df, legendary_df, non_legendary_df)

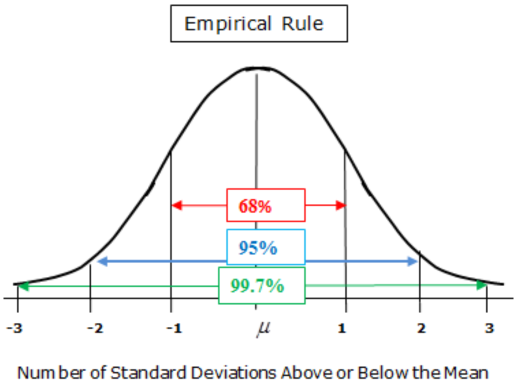
2. Have students answer the following questions:
   - What's the sample size of our data set? [800 total Pokemon]
   - What is the sample size of just the pokemon who are Legendary? [65 cases]

3. Have students brainstorm how to find the solutions to the following questions. [suggested option: create subcategories of legendary_df and non_legendary_df]
   - What is the sample size of just the pokemon who are not Legendary and use a Fairy Attack 1? [16 cases]
   - What is the sample size of just the pokemon who are Legendary and use a Grass Attack 1? [16 cases]

4. Calculate basic statistics (mean, median, standard deviation) on the Hit Points attribute.
   - Discuss these results in context of the data, what do they mean?
   - You may choose to have students also create a box-plot or a histogram

5. Create a graph using the Speed attribute (box-plot or histogram)

6. Calculate the mean and standard deviation of the Speed attribute
   - On the board, give students the definition of standard deviation.
   - Describe how *The Empirical Rule* applies to normally distributed data:
     - Approximately 68% of the data will fall within 1 standard deviation of the mean.
     - Approximately 95% will fall within 2 standard deviations of the mean.
     - Approximately 99.7% will fall within 3 standard deviations of the mean.
   - Suppose the Pokemon data was normally distributed, use the standard deviation calculation to answer: How many of the pokemon have speeds between 39.2 and 97.4? [544 Pokemon]



7. Have students brainstorm how to use the max() function to find the fastest speed of each Attack 1 type.
   - What is the fastest psychic attack Pokemon? [Deoxys Speed Forme]

8. Have students brainstorm the following question: What if I knew nothing about Pokemon? So instead of rating the Defense attribute numerically I wanted to summarize it so that any defense over 100 would be labeled good and anything below is considered bad.

     - Use the replace function from the data cleaning lesson.

Starting from Scratch Python Skills Checklist

□ I can upload this dataset into a new Python file

□ I can clean the file by dropping missing values
- **Bonus** I can drop only the cases that have missing values in my column of interest

□ I can identify the sample size of this data set
- Sample Size = _______

□ I can access a subset of the data
- Identify your subset: ____________________
- Sample size of the subset = _______

□ I can find the min and max of the column I am using
- Minimum: ____
- Maximum: ____

□ I can calculate the mean of an attribute in the data set
- Mean = _______
- What does this mean represent?

  ____

□ I can calculate the median of an attribute in the data set
- Median = _______
- What does this median represent?

  ____

□ I can calculate the mode of an attribute in the data set
- mode = _______
- What does this mode represent? ____

□ I can calculate the standard deviation of an attribute in the data set
- Standard Deviation = _______
- How many cases are within 1 standard deviation of the mean? ________________

□ I can create at least three different visualizations with my data.

□ I can summarize my findings in context of the data
- Summary:

□ I can save and store my ipynb file in an organized way

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 it is true due to the United States senator, Ladda Tammy Duckworth, from Illinois who is a retired Army National Guard lieutenant colonel, who lost a limb in active duty.*
Virginia is not a state that a lot of people vacation in, because the average temperature is 39.8 degrees	*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. Give students the following Worksheet - Philosophers Chair Statements.
2. Have each student decide a position they'll take on the statement and why. Consider posing the prompts:
   - 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 about:
    - 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. Have students read this article: Simpson's Paradox using US Presidential Elections

    - Have students reflect on the articles in their journal by identifying what consequences this could have in society.

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.	

Student Guide - Project News Article

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

Project Checklist:

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

Part 1: Reading - Read the following article Statistics and Visuals

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

Idea 1	
Idea 2	
Idea 3	

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

Question	***Data Type***

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

Visualization 1	Visualization 2

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

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.

Printable Exit Ticket

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 scenario 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 a full 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

Data Science HS with Python

Creating Models

A 4-day guide to model creation and relationship finding by Sara Fergus & Christa VanOlst

Summary

In this 4-day lesson, students will use Python to create linear and polynomial regression models over scatter plots. On days 2 & 3, students will categorize patterns & create predictive regression models.

Note: This is similar to the Creating Simple Models *from the* Data Science Unplugged *&* Data Science with CODAP *sequences. This lesson provides opportunities for students to use Python to create models on days 2 & 3, which the other lessons omit..*

Objectives

The students will be able to . . .

- Identify relationships in scatter plots including: Linear, Polynomial, Logistic, and Clustered
- Create linear models in Python (using scipy) and plot them over a scatter plot.
- Create logistic models in Python (using sklearn) and plot them over a scatter plot.
- Make predictions using linear and logistic models that they created in Python.

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 technological tools and functions within those tools to analyze and communicate data effectively.

Materials

- Warm Up: Optimism Regression & Plot Cards (PDF)
- Correlation Investigation (Desmos) & *Correlation Student Activity Guide* (see below)
- Correlation vs Causation Data Talk (Desmos)
- Python Worksheet: Linear Regression (.ipynb, key)
- Python Worksheet: Logistic Regression (.ipynb, key)
- Data Sets: Salary.csv, crime_and_incarceration_by_state.csv, admission_data.csv, housing.csv, Fish.csv, Social_Network_Ads.csv, titanic.csv, framingham.csv
- Extensions; Polynomial Regression (.ipynb) & Logistic w/ Multiple Predictors (.ipynb)

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.
`scipy`	`Scipy` is a python library often used for modeling in Data Science. The documentation can be found here. Used in this lesson: ● `linregress` ● `expit`
`numpy`	`Numpy` is a python library known for its handling of arrays. It can also fit polynomial regression. The documentation can be found here, Used in this lesson: ● `to_numpy` ● `reshape` ● `linspace` ● `poly1d` (Extension only) ● `polyfit` (Extension only)
`Scikit-learn (sklearn)`	`Sklearn` is a Python library often used for machine learning, which in this case includes logistic regression and decision trees. The documentation can be found here. Used in this lesson: ● `LogisticRegression`

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. **Regression with Python:** Run each code block in the "running a linear regression" section of the Linear Regression worksheet one at a time, asking students afterward what that block did, and what in the code made it happen.

   Consider asking students to write out the steps in pseudocode or add additional comments to the worksheet to help them identify each functional component of the code.

Make sure that students notice that the code found the slope and the intercept of the regression line, which is enough information to make a graph.

7. **Linear Regression Practice:** Have students run their own linear regression with the incarceration or college admission data based on the example. When students are stuck, prompt them with questions:

   - What should you copy from the example?
   - What should you change?
   - Which lines are necessary? Which are the most important lines? (for example, rounding the slope and intercept are not strictly necessary)

   Students are working through a number of steps. Make sure that, as they work, they know which step that they are on.

Before they begin, ask: "How will you know you were successful?"

Students will need to remember how to upload data and access columns. If they are stuck on this, point them to their work from previous lessons. See the KEY for more notes on students' potential responses.

8. **Extension:** Have students that finish well before other students look at the polynomial regression extension worksheet. They will need housing.csv and Fish.csv.

   If students finish a little bit before the class and don't have time for the extension, have them work to improve the style of their plots.

Day 3 Outline

Formative Assessment Notes

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

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

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

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

10. **Logistic Regression Activity:** *Again, you can have students work through the worksheet on their own, in groups, or as a class using the following steps.*

   On the board, run the first code block in this worksheet. Make sure to upload Social_Network_Ads.csv, titanic.csv, and framingham.csv before running.

   Ask students to write in their journals what they notice and what they wonder about the three plots. Encourage them to consider the similarities and differences, recalling the patterns they explored in previous lessons.

Have a few students share their notice-and-wonders with the class. Students should be finding that they all can be modeled with a logistic regression curve.

11. Once students have compared, define the similarities and differences of logistic and linear regression on the board. To get them started, remind them of the examples they have seen:

   Linear: Years of experience and salary, prisoner count and state population, GPA and school admission

   Logistic: Age and responding to an ad, ticket fare and whether or not you survived the Titanic crash, number of cigarettes and whether you develop heart disease.

Make it clear to students that they should not only combine lists, but check the other group to make sure that their combined list is accurate.

Make sure also that students are writing situations that are actually unique from previous.

12. **Logistic & Linear Regression Brainstorm:** Then, break students into small groups of 2 or 3 students. Have students collaborate to write down as many examples of when you would use logistic regression as they can think of. After one minute, have them write as many examples as they can think of for linear regression.

    Once students have examples written, have groups "pair up" into bigger groups of 4 to 6. Have them combine and critique their list.

    Optionally, you could continue to combine groups until you have a class-wide list of examples that the class has decided are all accurate.

    As a class, revisit the original comparison and incorporate similarities and differences in use-cases.

After revisiting similarities and differences, students should come to the conclusion that linear (and polynomial) regression is used for a relationship between two continuous variables, and a logistic regression is a relationship between one continuous variable and a boolean (True or False) variable. As they share, define "continuous" and "boolean".

13. **Logistic Regression Demo:** Run each code block in the "running a logistic regression" section of the logistic regression notebook one at a time, asking students afterward what that block did, and what in the code made it happen.

    Have students run their own logistic regression with the Titanic data or the Framingham (heart study) data.

This code involves a little bit more set-up. Additional comments are included in the worksheet

Check out the KEY for additional details

14. **Extension:** Have students that finish well before other students look at the extension worksheet: logistic regression with multiple predictors. They will need titanic.csv and framingham.csv.

    If students finish only a little bit before the class, you could have them improve the style of their plots.

Day 4 Outline

Formative Assessment Notes

15. **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.

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

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

Check in with students along the way.

Assessment Strategies

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

Regression Mini-Project

Have students create a new notebook. Using either Kaggle or a class set of datasets, instruct students to find some data that they are interested in that has at least 2 numeric attributes.
In their notebook, students should

1. Upload and save the data set
2. Clean the data as necessary
3. Create a scatterplot
4. Determine the relationship
5. Create a linear or logistic regression model
6. Make at least one prediction based on your model

Have students use the student guide to guide their exploration.

	Proficiency	*Yes*	*No*	*Notes*
Concept	The student accurately identifies the relationship between their two numeric variables.			
Representation	The student generates, expresses, and assesses a model to best fit their data.			
Coherence	The student uses their model to make a prediction and expresses the prediction in the context of the data.			

Some Accommodations & Extensions

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

Accommodations

Some students may even benefit from creating their own document instead of using the worksheet, and keeping everything in one place.

To avoid having to upload a lot of different datasets, you could have students only upload the incarceration OR admission data for the linear regression and only the Titanic OR Framingham data for logistic regression. If students do this, make sure they run only their code block, instead of the whole worksheet. This could be helpful with students who may become overwhelmed with a lot of files open.

Extensions

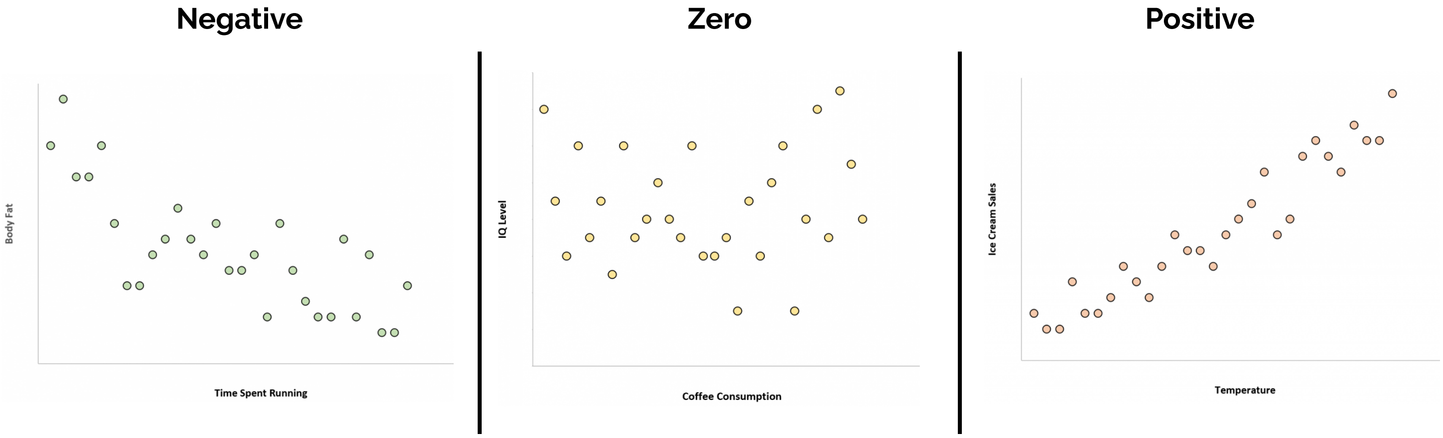
We've included a few extensions—polynomial regression and logistic regression with multiple predictors—in the outline. Students should be able to work on these worksheets on their own while the rest of the class finishes their models. If students finish with only a little bit of extra time, you can encourage them to improve the style of their plot (ex. Add color). You could even challenge them to use .text() to print the model's equation on their plot.

To avoid having to upload a lot of different datasets, you could have students only upload the incarceration OR admission data for the linear regression and only the Titanic OR framingham data for logistic regression. If students do this, make sure they run only their code block, instead of the whole worksheet. This could be helpful with students who may become overwhelmed with a lot of files open.

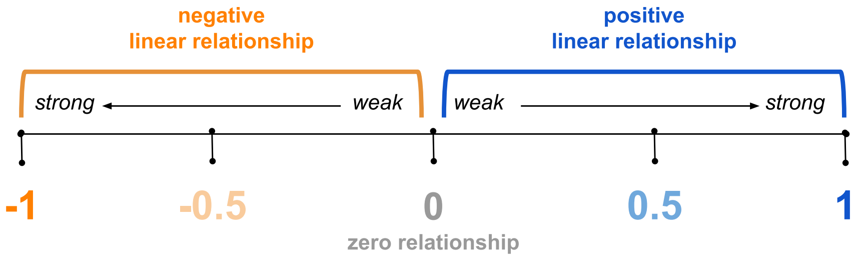
Student Activity Guide - Correlation vs. Regression

What is a Correlation?

A correlation measures the relationship between two variables.



Recall the spectrum of Correlation Coefficients



DIRECTIONS: Categorize the scenarios below by predicting their correlation coefficient value using a number between -1 and 1 inclusive.

______1. The height of a person and the salary they earn.

______2. The shoe size of a person and the number of movies they watched.

______3. The height and the weight of a person.

______4. The quote "With Age Comes Wisdom".

______5. The amount of time you spend in water (swimming/bathing) and the wrinkles in your skin.

______6. The speed of a wind turbine and the amount of electricity that is generated.

______7. The amount of moisture in an environment and the growth of mold spores.

______8. A student's screen time and their grades.

______9. A person's pizza consumption and their zodiac sign.

______10. A person's average pulse rate and the calories they are burning.

______11. The temperature it is outside and the amount of layers of clothing a person wears.

______12. The size of a herd of animals and the amount of food to go around.

Correlation vs Regression

When studying the relationship between numeric variables, it is important to know the difference between correlation and regression.

What is Regression?

Regression is a statistical technique used to approximate a linear relationship of two variables to make predictions.

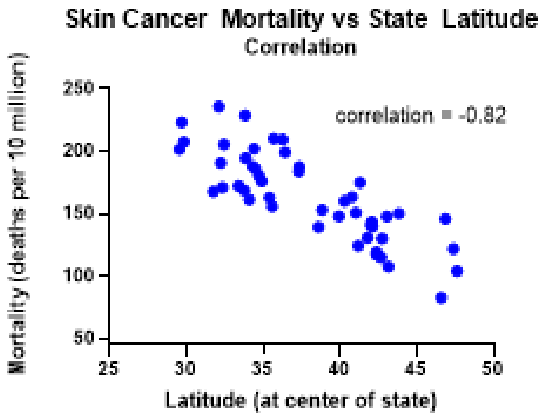


Figure A

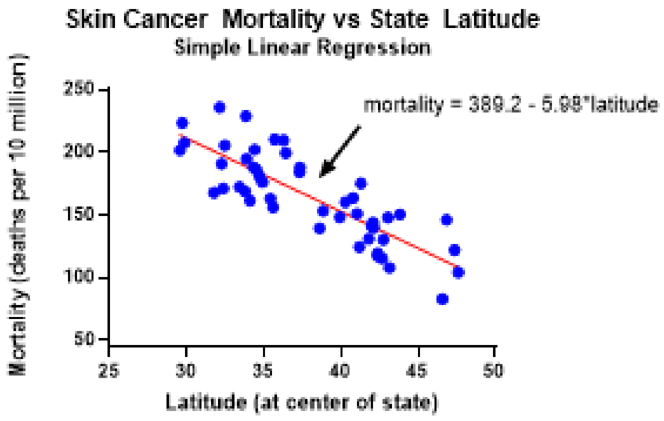


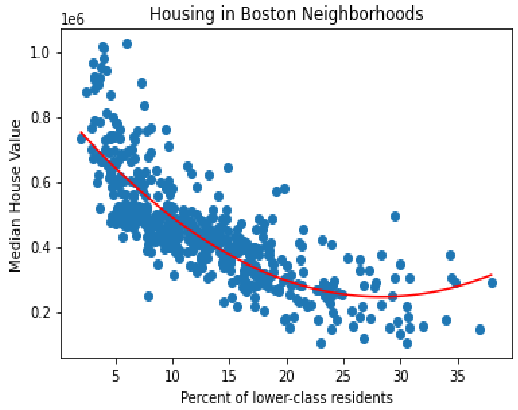
Figure B

Citation: [GraphPad]

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

Latitude	By Eye Prediction	Calculated Prediction `-5.98(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?



KEY - Creating a Linear Regression

- `*****` represent where students need to make changes
- *italics* represent lines that could be changed, but don't need to be

`from scipy.stats import linregress`
`import pandas as pd`
`import matplotlib.pyplot as plt`
	Make sure students include the import statements

`#Read in the Data`
`df_name = pd.read_csv('/content/******.csv')`
`x = df_name['*****']`
`y = df_name['******']`
	Students need to change:
	1. The filename
	2. The column names
	Students should change:
	1. The data frame variable name
	Notes:
	- If students are not sure which columns to use, point them to the scatterplots created at the beginning of the worksheet.
	- When using the incarceration data, make sure to drop missing values.
	- When using the admission data, the column header for `'Chance of Admit '` has a space at the end.

`#Run a Linear Regression`
`result = linregress(x, y)`
`slope = result.slope`
`intercept = result.intercept`
 	
`rounded_slope = round(slope,2)`
`rounded_intercept = round(intercept,2)`
`print("Linear Regression Line: y =",`
`rounded_slope, 'x +', rounded_intercept)`
	Notes:
	- Rounding is good, but not strictly necessary
	- Make sure students understand that the highlighted line is where the linear regression is actually calculated

`#Analyze Linear Regression Accuracy`
`print("R-value: ", result.rvalue)`
`print("This r-value suggests *******.")`
`print()`
	Students need to change:
	1. The interpretation of the r value
	Note:
	- Students will have to run the script without the interpretation before they can add that part.

`#Plot Linear Regression`
`plt.scatter(x, y)`
	Students need to change:
	- The title, xlabel, and ylabel of the graph.

`plt.plot(x, slope*x+intercept)`
`plt.title("*****")`
`plt.xlabel("*****")`
`plt.ylabel("*****")`
`plt.show()`	

`#Make a prediction`
`print("Prediction Example")`
`est_sal = slope*(*****)+intercept`
`rounded_salary=(round(est_sal,2))`
`print("******", rounded_salary)`
	Students need to change:
	● ***** with an x value that makes sense in their data set. They can look at the original scatterplot to get an idea of what that might be
	● The interpretation of the estimation (******)
	Students should change:
	● ***est_sal*** and ***rounded_salary***, to reflect the dataset
	Notes:
	● Rounding is not strictly necessary
	● Students will have to run the script without the interpretation before they can add that part.

KEY - Creating a Logistic Regression

- ***** represent where students need to make changes
- *italics* represent lines that could be changed, but don't need to be

```
from scipy.stats import linregress
import pandas as pd
import matplotlib.pyplot as plt
```

Make sure students include the import statements

```
#Read in the Data
df_name = pd.read_csv('/content/******.csv')
df_name = df_name.dropna()
x = df_name['*****'].to_numpy().reshape(-1,1)
y = df_name['******']
```

Students need to change:

3. The filename
4. The column names

Students should change:

2. The data frame variable name

Notes:

- If students are not sure which columns to use, point them to the scatterplots created at the beginning of the worksheet.

```
#Run a Logistic Regression
result = LogisticRegression(random_state=0).fit(x, y)
print("Logistic Equation: p(x) = 1/(1 +
exp(-(",round(result.intercept_[0],2),"+",round(result.coef_[0][0],2), "x)")

#Analyze Logistic Regression Accuracy
print("Accuracy: ", result.score(x, y)*100, "% (for the given data)")
```

Notes:

- Rounding is good, but not strictly necessary
- Make sure students understand that the highlighted line is where the logistic regression is actually calculated
- Nothing in this part needs to be changed.

```
#Plot Logistic Regression
x_plot = np.linspace(x.min(), x.max(), 300)
to_plot = expit(result.coef_*x_plot + result.intercept_).ravel() plt.scatter(x,y)
plt.plot(x_plot, to_plot)
plt.title("*****")
plt.xlabel("***")
plt.ylabel("***")
plt.show()
```

Students need to change:

- The title, xlabel, and ylabel of the graph.

```
#Make a prediction
print("Prediction Example")
prob = expit(result.coef_******+result.intercept_)
print("******", round(prob[0][0]*100,2), "% chance of *******")
```

Students need to change:

- ***** with an x value that makes sense in their data set. They can look at the original scatterplot to get an idea of what that might be
- The interpretation of the prediction (******)

Notes:

- Rounding is not strictly necessary
- Students will have to run the script without the interpretation before they can add that part.

Student Guide: Regression Mini-Project

In this project, you will practice creating models with data. Follow these steps.

- □ Find a data set that interests you and has at least 2 numeric columns.

  Numeric Column 1:

  Numeric Column 2:

- □ Upload and store the data into a new Python notebook. Conduct any data cleaning necessary
- □ Create a scatterplot with your data
- □ Do you think that this is…
  - □ A logistic relationship
  - □ A positive linear relationship
  - □ A negative linear relationship
  - □ A null relationship
  - □ Something else: _______________
- □ In the context of the data, why do you think that these attributes have this type of relationship?

- □ If you have a relationship that is not linear or logistic, keep looking until you have a logistic or linear relationship. Once you have found one, sketch it here and draw a "by eye" model. Be sure to label your axes.

□ Using Python, create a logistic or linear regression model, depending on the relationship.

Equation:

□ Add the Python result to your by-eye prediction on this worksheet

□ How good is your model? How do you know?

□ Make at least one prediction with your data. Be sure to write the results in the context of your data. *Example: When somebody is 4 years old, you can expect them to be about 40.2 inches tall.*

My Prediction:

□ Summarize what you found, in the context of your data. *Example: There is a strong linear relationship between a person's age and their height (r-value .87). This makes sense, since people grow taller as they get older. This relationship is strongest from the ages 0 to 16. At about 16 years old, the relationship is not as strong, since people tend to stop growing around then.*

Worksheet - Calculating & Assessing using Residuals

Vocabulary

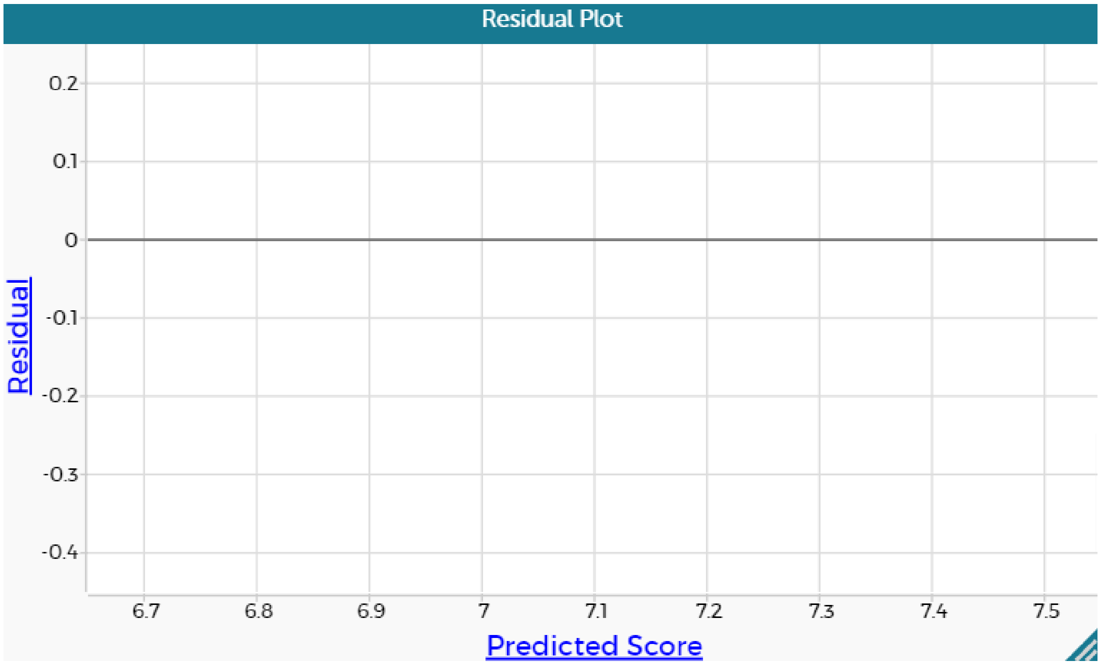
Residual	The difference between the observed value and the model's predicted value.
Residual Plot	A residual plot graphs the residuals of a line of a best fit on the vertical axis and the independent variable on the horizontal axis. Residual plots can be used to determine whether a linear model is appropriate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

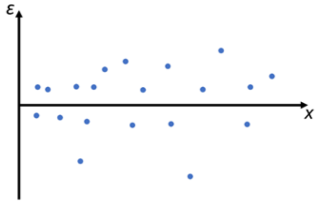
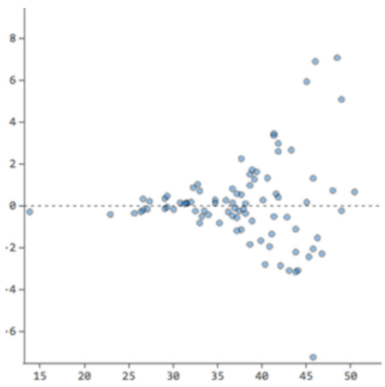
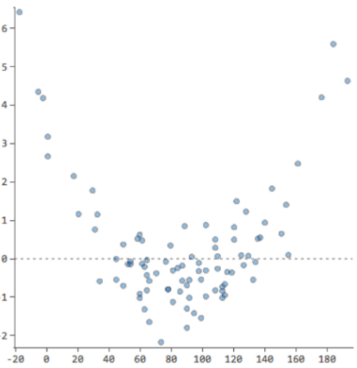
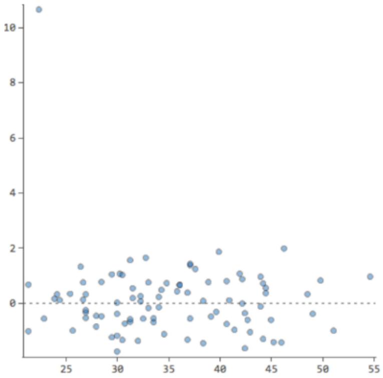
DIRECTIONS: Calculate the following values to complete the table.

Overall rank	Country	Freedom to make life choices	Score	Predicted Score 3.09*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?



<table>
<tr><th colspan="2">Good Residual Plots</th><th>Example</th></tr>
<tr><td colspan="2">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.</td><td> </td></tr>
<tr><th colspan="3">Bad Residual Plots</th></tr>
<tr><td><i>Uneven spread</i> - the model does not fit consistently across all x-values.</td><td><i>Curved</i> - If there are patterns or curves in the residual plot then a nonlinear model may be more appropriate (quadratic, polynomial, etc.)</td><td><i>Outlier</i> - There may be an underlying data recording error. Remove to see what the effect is whether it is influential or not.</td></tr>
<tr><td> </td><td> </td><td> </td></tr>
<tr><td colspan="3"><i>If you can detect a clear pattern or trend in your residuals, then your model has room for improvement</i></td></tr>
</table>

Data Science HS with Python

Making Predictions

An overview of making predictions using models by Sara Fergus & Christa VanOlst

Summary

In this two day lesson, students use two techniques to make predictions about missing or future data: by eye and through evaluating functions using a given mathematical model. Students explore datasets throughout the lesson by creating quick scatter plots and models to predict outcomes. In conclusion, students then discuss the tradeoffs and limitations of certain models. Then, students complete a mini-project where they collect data, analyze it for correlation (positive, negative, null), and use modeling to support their claim. As an extension, students may also explore logistic and multivariate models.

Note: This lesson is very similar to Making Predictions *from the* Data Science Unplugged *&* Data Science with CODAP *sequences. In this lesson, students primarily compare & contrast linear and nonlinear models, and perform analysis using Python tooling.*

Objectives

The students will be able to . . .

- Create linear and polynomial models sketching by eye predictions over a scatter plot
- Make predictions given a linear and polynomial models
- Compare and contrast by eye and mathematical model predictions
- Create and test hypothesis statements through visualizing and modeling collected data

Standards Alignment

- **D.S. 9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.11:** The student will select statistical models and use goodness of fit testing to extract actionable knowledge directly from data

Materials

- Warm Up: Health Foods (Desmos)
- Experience vs Salary (CSV, example scatterplot, & example regression plot)
- Fuel Consumption Since 1990 Data Set (CSV)
- Student Guide: Sketching Models & Making Predictions
- Predictions Slideshow (view in Google Drive or make a copy)
- Mini Project: Prove a Relationship
- Additional Data Sets: Crime Data Subset (CSV), College Admissions Subset (CSV), Position and Salaries Data (CSV), Fish Data Subset (CSV)

Vocabulary

Term	Definition
Line of Best Fit	A line of best fit is a straight line through a scatter plot that represents the linear regression.
Linear Regression	A linear function used to approximate a relationship between two variables used to make predictions and describe situations
Polynomial Regression	A polynomial function used to approximate a nonlinear relationships (quadratic, cubic, etc) between two variables to make predictions and describe situations
Model	A model is a framework for making predictions or describing situations using mathematical equations (i.e. regression) or other algorithms (e.g. decision tree).

Day 1 Outline

Formative Assessment Notes

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. **Sketching Regression to make Predictions:** Give the students the following data: Experience & Salary Data (CSV)

   Using Python & matplotlib, have students create a scatter plot of the two data attributes.

   - Years Experience on the x-axis and Salary on the y-axis
   - Example scatter plot

Use this as an opportunity to check in and make sure students are comfortable using Python to create scatter plots.

3. **Discussion:** Ask students to discuss the following in small groups:

- How does experience relate to salary? Describe the relationship
- What do you notice?
- What do you wonder?
- What correlation do you see? Are they strong or weak?
- What questions arise?

Students should be pretty good at this after the previous lesson—use this as an opportunity to shore up any misunderstandings.

4. **Making Predictions:** Model creating a by eye sketch of a linear model on the Years Experience vs. Salary scatter plot. [*example model sketch*] ***OR*** review creating linear regression models with Python using the data set.

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 could calculate these equations by hand but that would be redundant and has room for error.

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

7. **Creating Models:** In pairs, give students the following data tables:

- Crime Data Subset (CSV)
- College Admissions Subset (CSV)
- Position and Salaries Data (CSV)
- Fish Data Set (CSV)

Have students complete the Student Guide - Sketching Models & Making Predictions to practice developing linear models.

Check in with students as they complete the worksheet. They may need a review of how to create models using Python (a recently developed skill)—re-teach/review as needed.

7. **Comparing Linear & Quadratic:** Have students work in pairs to compare a fuel consumption prediction from step #6 to a prediction using this quadratic model:

   - FuelConsumption = -0.113624(year)^2 + 4.620214(year) + 127.598

   Compare the model in step 3 (linear) and discuss which is better.

8. have students evaluate the models they created in the Student Guide - Sketching Models & Making Predictions worksheet. Ask:

   > ***"Do you see any linear models that don't fit the data well?"***

   Have students discuss in pairs, then show them the models below:

   - 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 using the models above, and discuss how these values compare to their previous predictions.

The primary goal of this activity is to have students compare the linear predictions they made with Python to "better" predictions using different regression equations. Be sure to check in with students during their pair discussion to make sure they can see how the linear models fail to "fit" some of the data sets.

9. **Optional Extra Practice:** Give each group a white board and a dry erase marker. Go through this slideshow, stopping after each slide to show responses. Be sure to point out that students should come up with a predicted value, not the actual value at that point.

10. **The Limitations of Modeling:** Show students with the cars.csv data

    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. Discuss the limitations of the model given these conditions.

Students should learn that mathematical models of data sets have limited ranges of applicability and using a model outside its range can lead to poor predictions.

Day 2 Outline

11. **Prove a Relationship Mini-Project:** In this mini-project, students hypothesize a correlation that can be supported by collecting data from their classmates. Students scatter, plot and model their findings to predict future values and reflect on the limitations of their findings.

See *Assessment Strategies* below

Assessment Strategies

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

Mini-Project: Prove a Relationship

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

Most students will likely choose something with a null relationship. After the project, take a moment to discuss this with the class.

Project Milestones:

- ☐ Make a hypothesis statement
- ☐ Survey and collect data from peers
- ☐ Plot the data on a scatter plot using Python (matplotlib)
- ☐ Sketched a model of best fit and then run a regression model using Python, then compare your by eye predictions
- ☐ Create a presentation to include: hypothesis statement, a summary of data collection, visualization, regression model, and at least one prediction
- ☐ Reflect on their hypothesis statement and the applicability/limitations of their findings.

The next page contains a rubric for assessing student work:

Mini-Project Rubric

	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**, stored, and organized from their peers			
Data Visual	Students' choice of visualizations is **appropriate for the data attributes** AND **provides insight** to sketch a model			
Model	Students' sketch of their **regression model is appropriate** and accurate for the data AND the student **makes a valid prediction**			
Presentation	Students' presentation includes **ALL of the requirements in milestone 5.**			
Reflection	Students' **reflection is thoughtful and relevant** when describing the applicability and limitations of their findings			

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

The student guide 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.

Student Guide - Creating Models & Making Predictions

Vocabulary

Line of Best Fit	A line of best fit is a straight line through a scatter plot that represents the linear regression.
Linear Regression	A linear function used to approximate a relationship between two variables used to make predictions and describe situations
Polynomial Regression	A polynomial function used to approximate a nonlinear relationships (quadratic, cubic, etc) between two variables to make predictions and describe situations

Using the following data tables: Crime Data, College Admissions Data, Position and Salaries Data, & Fish

In Python, create the following scatter plots and linear regression lines. Paste your results here.

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

Answer the following questions:

1. How does the incarceration rate in each state compare to the population?

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

3. How does position level compare to salary? How has your sketch changed? How is this different from the years experience example from before?

4. How does the diagonal length impact the weight of the perch fish? What about the whitetail?

5. Are these correlations positive/negative? Strong/weak?

Complete the following predictions using your models:

Scenario	By 👁 Prediction	Mathematical Model (from Python regression)	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 a salary of what?				
If a perch fish has a diagonal length of 20 cm, what is the expected weight?				

CodeVA CS Lesson Plan

Data Science HS with Python

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 Unplugged Data Science *&* Data Science with CODAP *sequences.*

Objectives

The students will be able to . . .

- Assess the strength of a model, taking overfitting and underfitting into account
- Differentiate important underlying patterns in data from noise

Standards Alignment

- **DS. 9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS. 7:** The student will be able to assess reliability and validity of source data in preparation for mathematical modeling.

Materials

- Reading: Model Limitations: Noise and Overfitting (1 per student/reading group)
- Examples of overfitting materials (Job Posting, Sports, Population, Pattern, President, Flu)
- Video: “What is Machine Learning?” (YouTube)
- Article: *Machine Bias* (PDF)
- Data Cycle Scenario – Communication (see below)

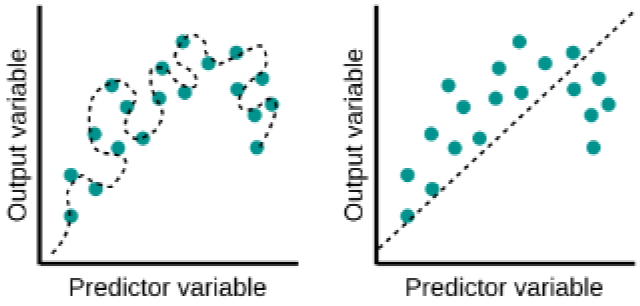
Vocabulary

Term	Definition
Error	Error is a measure of how inaccurate your model is. Error could refer to training data, testing data, or a combination of both.
Noise	Noise is variation due to natural imperfection or measurement error. "Noisy" data has a lot of variation that is unrelated to the underlying relationship.
Overfitting	A model is overfitted if the noise of the data has a large effect on the model. An overfitted model represents meaningless variation rather than an overall pattern.
Training Data / Training Set	Training data is the data that is used to create a model
Testing Data / Testing Set	Testing data measures the utility of a model by testing to see if it holds for general data that was not necessarily used to create the model. For example, a model created to represent the heights and weights of 8-year olds should be tested using the next year's 8-year olds, and still be a strong model.
Underfitting	A model is underfitted if it fails to demonstrate important patterns in the underlying relationship. For example, using a linear regression to show a quadratic relationship would be underfitting.

Outline

Formative Assessment Notes

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?

   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?

In general, students should find that the prediction is based too closely on specific instances of the past and overly complicated models.

While reviewing answers, connect student responses to vocabulary like "noise", "bias", "prediction", "training set", and "test set".

Stations Part 3: Repeat the cycle, answering the following:

- ***What would the "training set" be in this example? What would the "testing set" be?***
- ***How is this an example of overfitting?***
- ***What would you suggest the researcher do in order to answer their question without overfitting the data?***

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:

- ***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.

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"

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

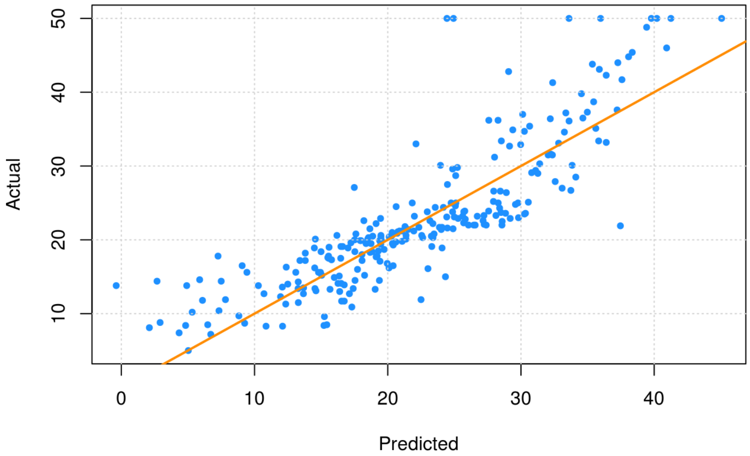
Students who need additional time reading may benefit from getting the overfitting worksheet ahead of time. The worksheet could also be annotated as a class or in a small group. For students with small group accommodations, consider pulling aside a few students and helping them to complete the worksheet,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

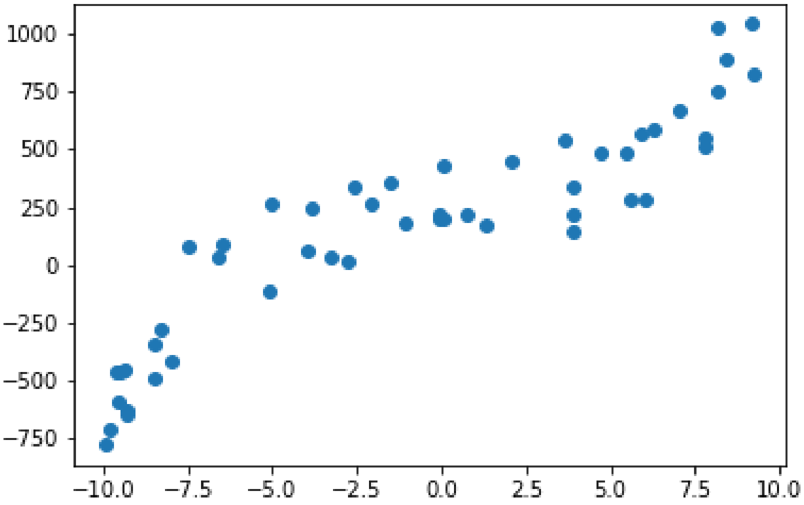
Model Limitations: Noise and Overfitting

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

Noise

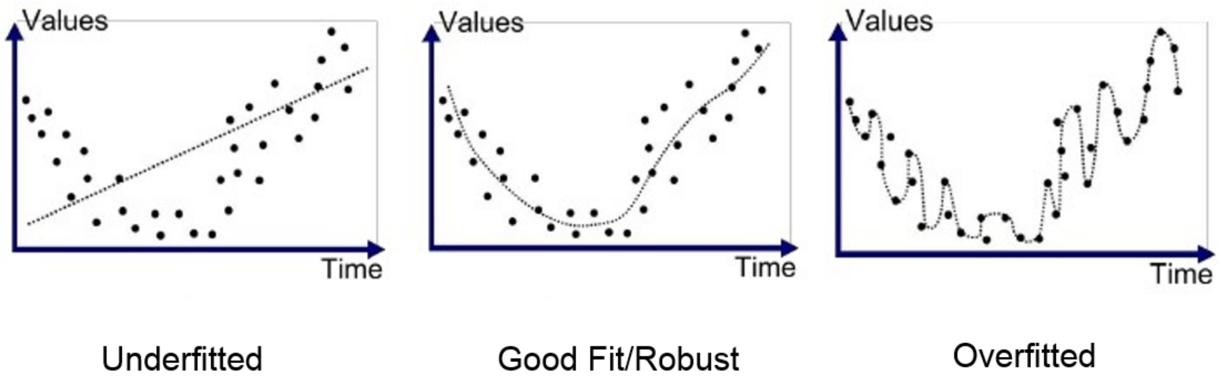


In Data Science, **noise** is a word used to describe random pieces of data that make the underlying pattern less clear. This comes from the fact that neither humans nor nature are perfect. For example, noise is introduced by measurement and rounding errors in data collection. Noise is also introduced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



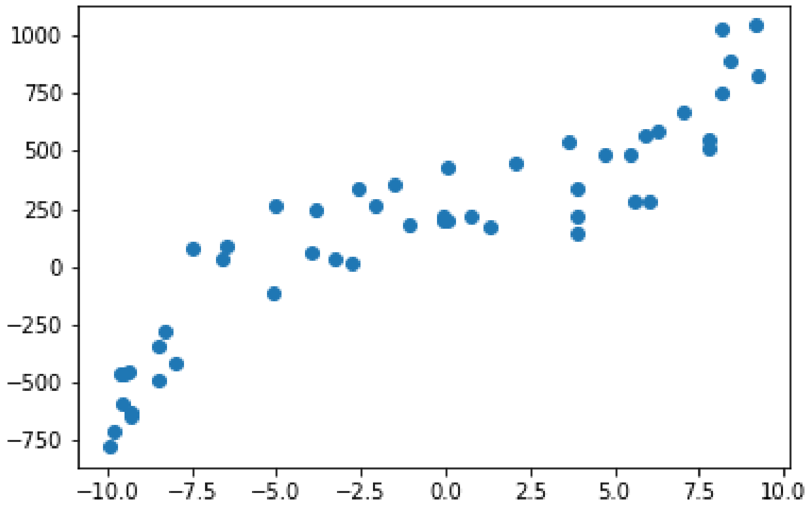
Underfitted

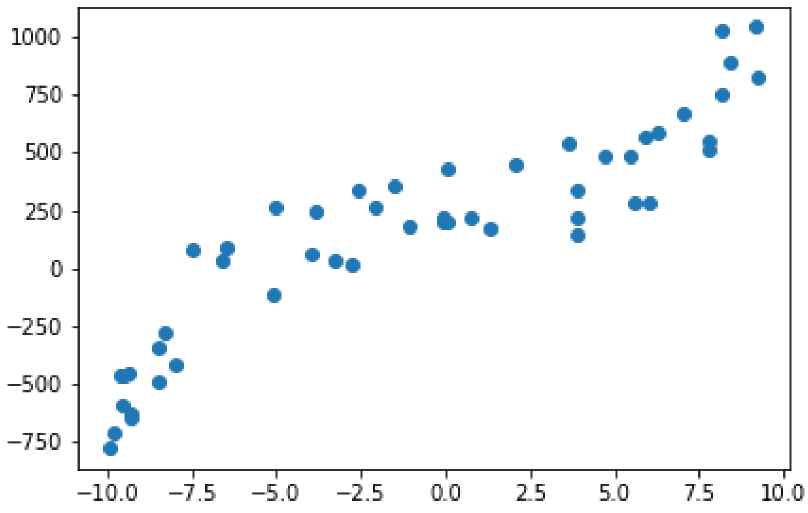
Good Fit/Robust

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

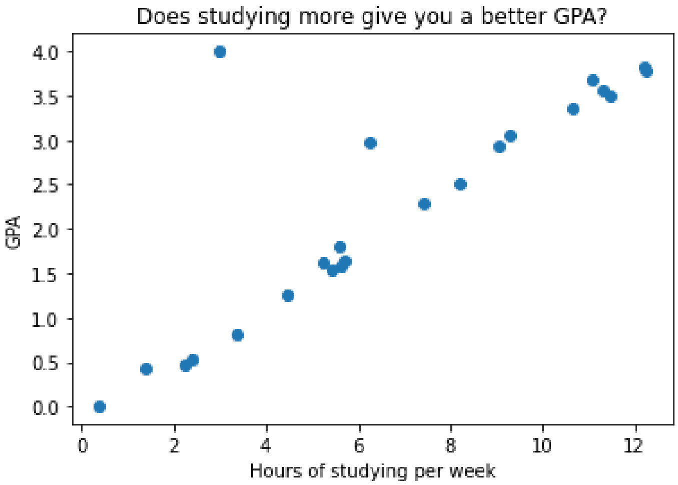


Sketch a by eye model that would be overfitted	
Define overfitting in your own words	
Define underfitting in your own words	

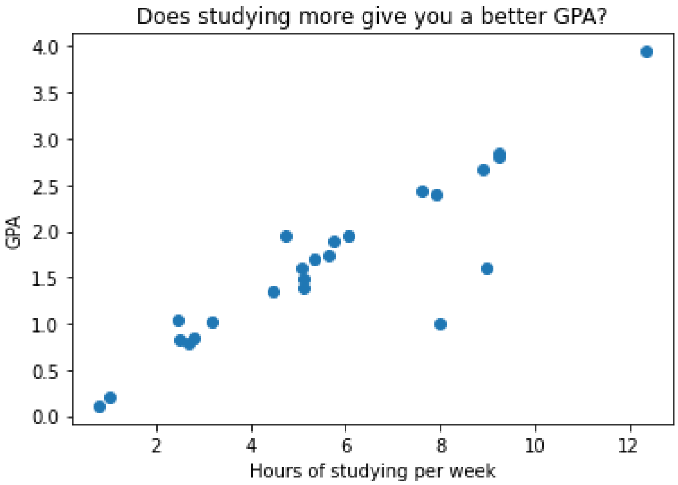
Training and Testing

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.

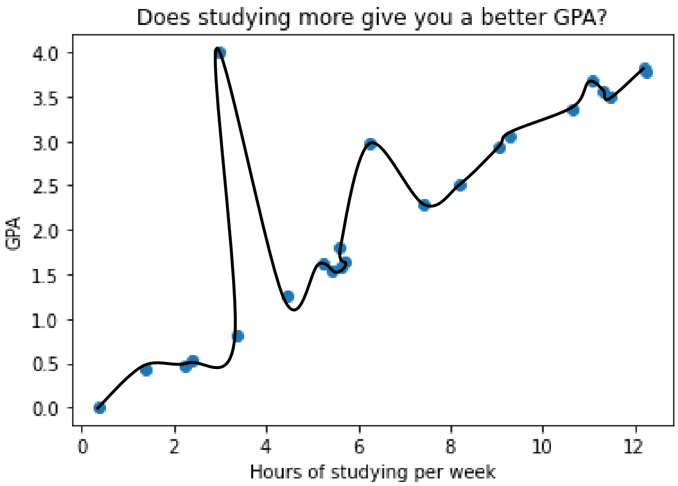


Training Data

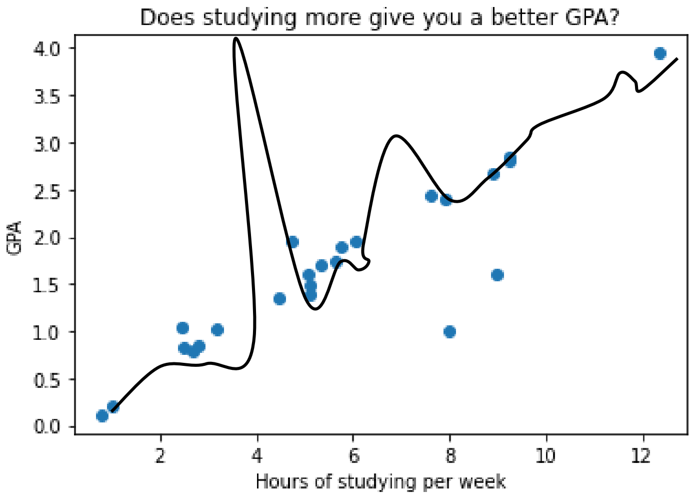


Testing Data

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).

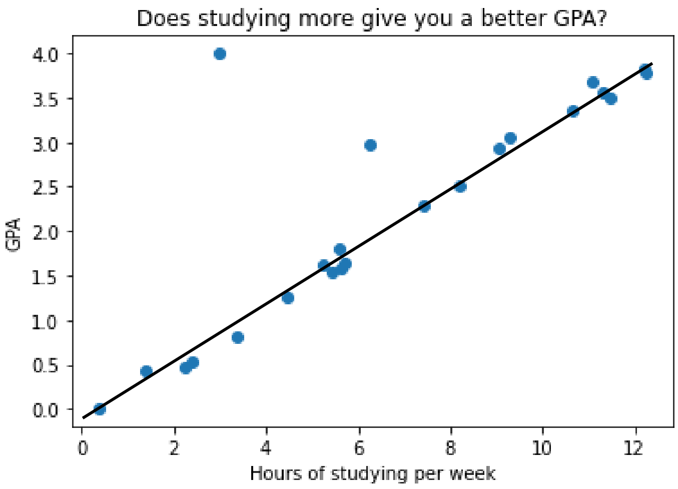


Training Data

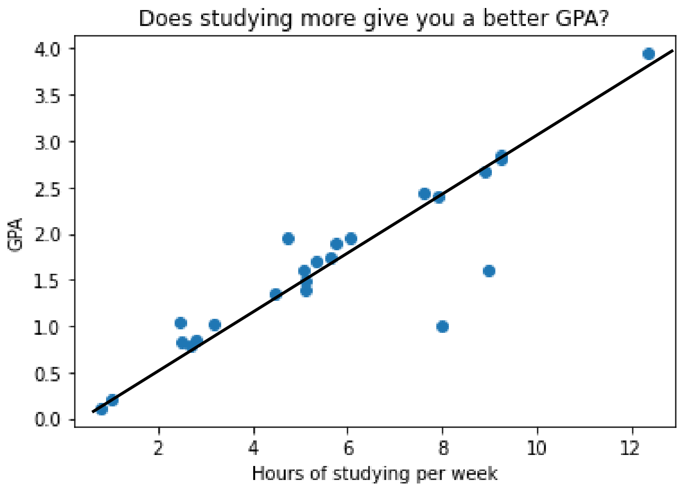


Testing Data

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

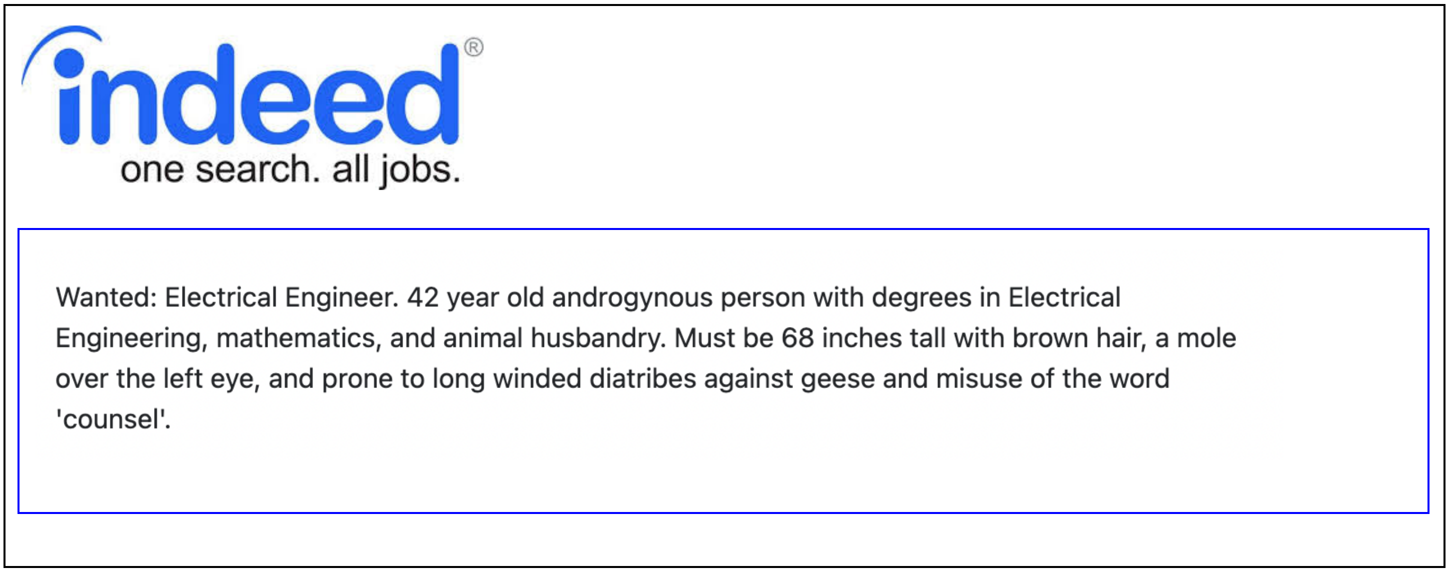
It is important to not overfit or underfit your data so that your predictions in the future are more accurate. Overfitting can also introduce bias from outliers. For example, the study may have been conducted in a school with a large amount of socioeconomic diversity. Over or underfitting can hide underlying 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.*

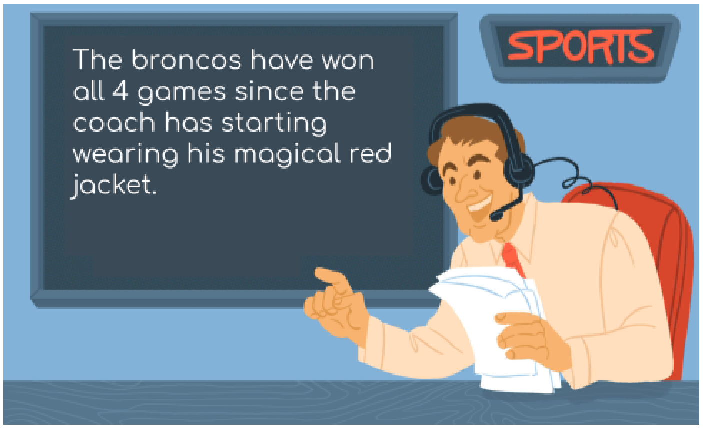
Job Posting Resource

"Agh! Pat is leaving the company. How are we ever going to find a replacement?"

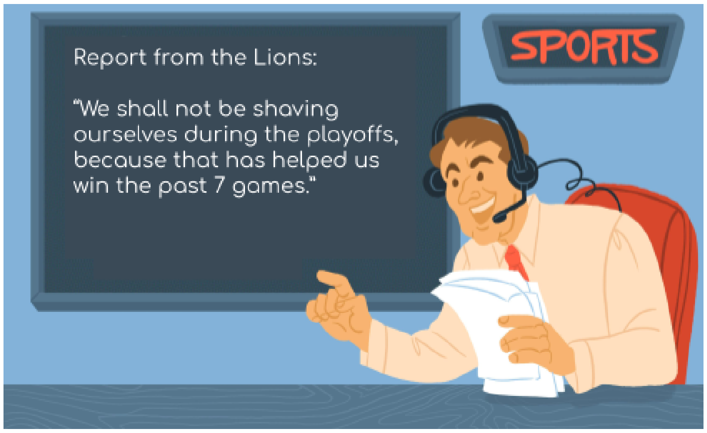


Wanted: Electrical Engineer. 42 year old androgynous person with degrees in Electrical Engineering, mathematics, and animal husbandry. Must be 68 inches tall with brown hair, a mole over the left eye, and prone to long winded diatribes against geese and misuse of the word 'counsel'.

Sports Resource

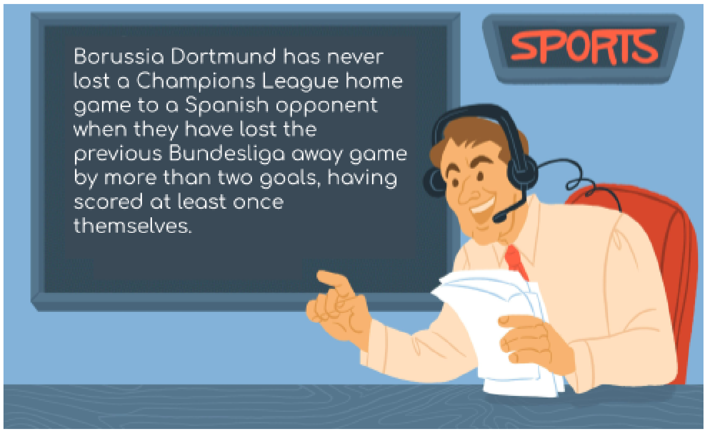


The broncos have won all 4 games since the coach has starting wearing his magical red jacket.

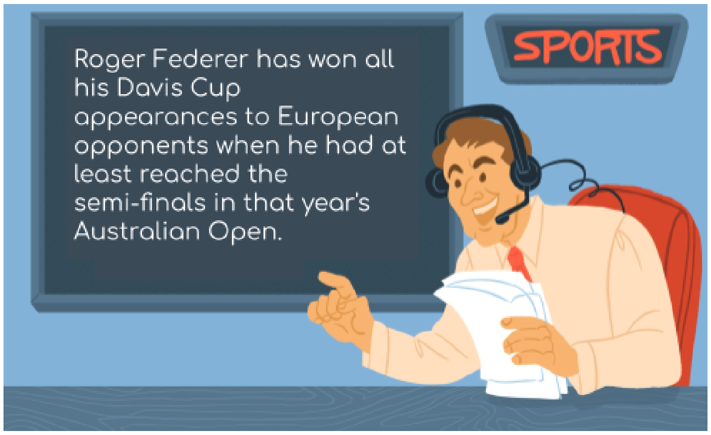


Report from the Lions:

"We shall not be shaving ourselves during the playoffs, because that has helped us win the past 7 games."



Borussia Dortmund has never lost a Champions League home game to a Spanish opponent when they have lost the previous Bundesliga away game by more than two goals, having scored at least once themselves.

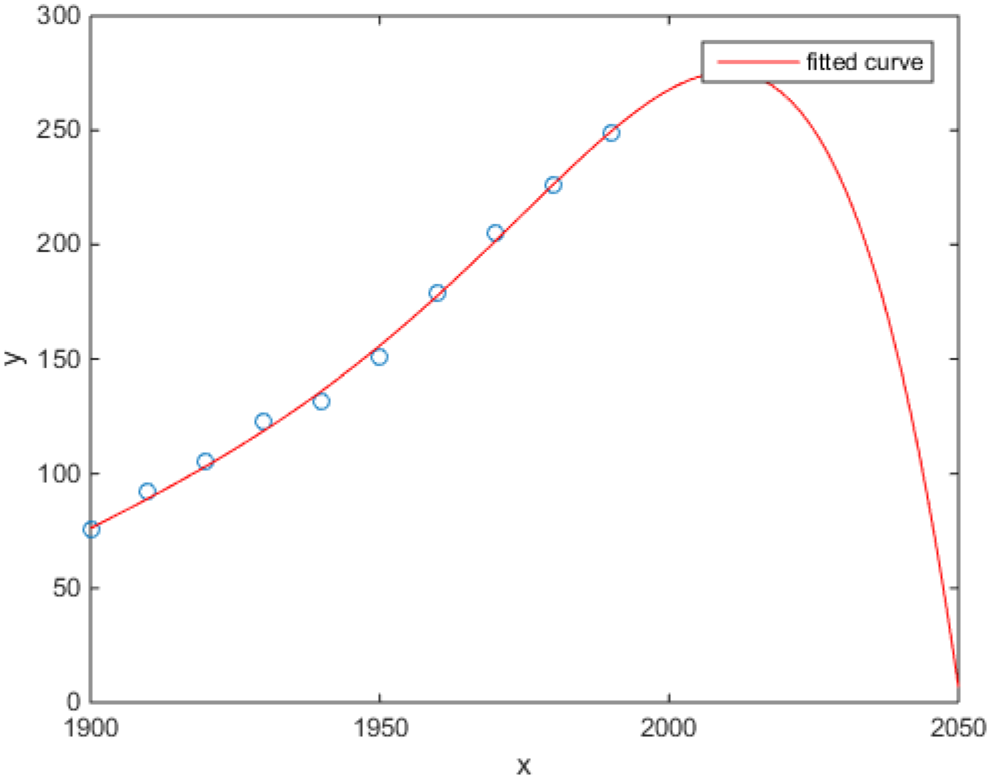


Roger Federer has won all his Davis Cup appearances to European opponents when he had at least reached the semi-finals in that year's Australian Open.

Population Resource

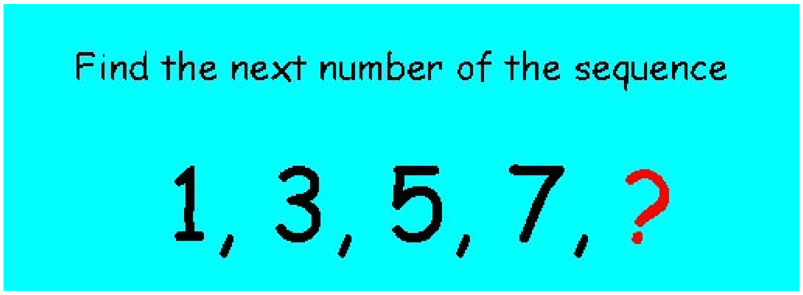
The Apocalypse is Coming!

According to recent research, we can predict that there will be no living people remaining in the United States by the year 2050. This decline will be beginning in 2010, although the cause is still unknown. This prediction fits previous data very well, so we know that our model is strong.



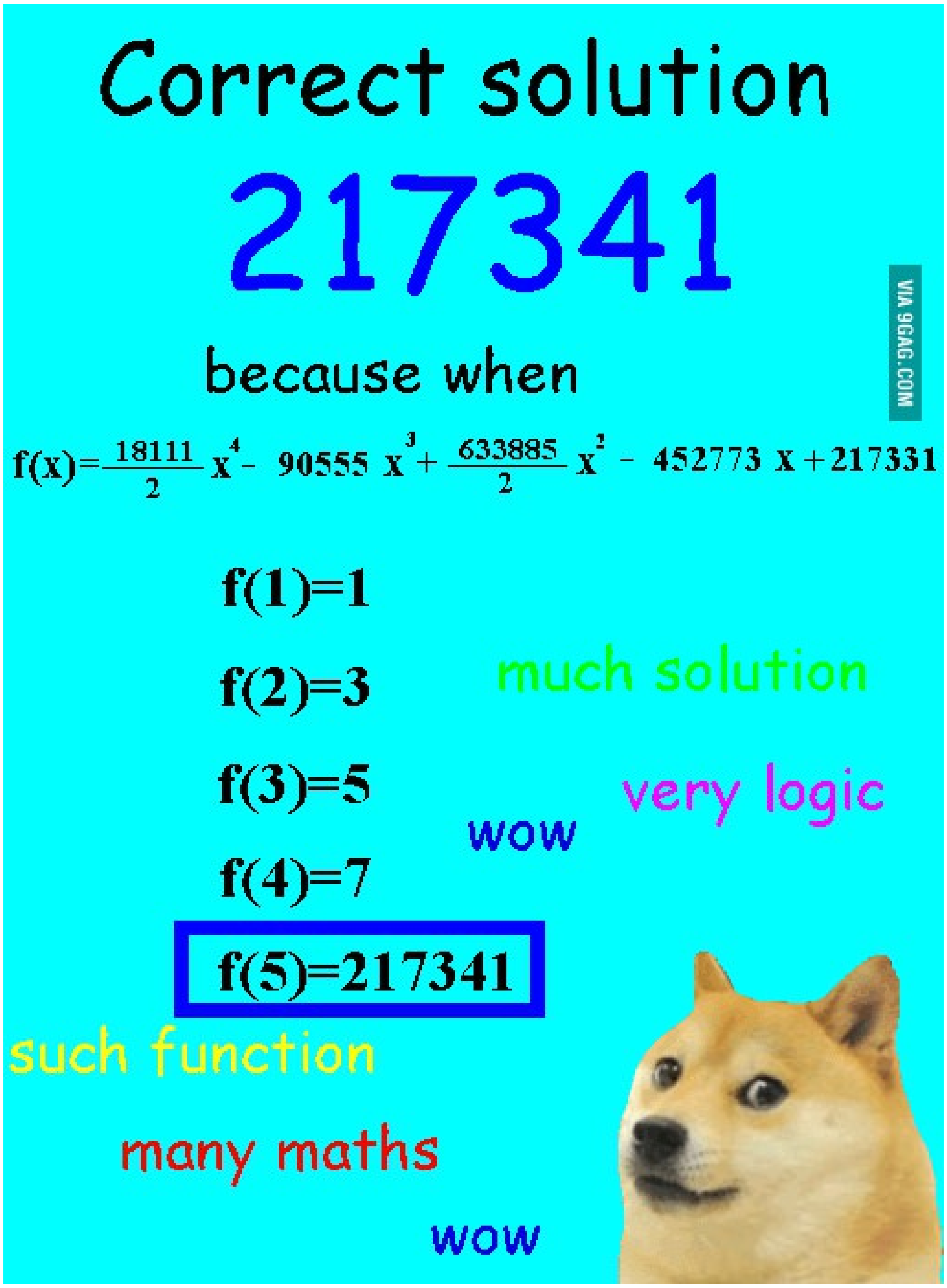
We suggest moving out of the country as soon as possible. Researchers are still working on modeling populations in other countries. It is possible that this event will not only occur in the United States.

Pattern Resource



Find the next number of the sequence

1, 3, 5, 7, ?



Correct solution

217341

because when

$$f(x)=\frac{18111}{2}x^4-90555x^3+\frac{633885}{2}x^2-452773x+217331$$

$f(1)=1$

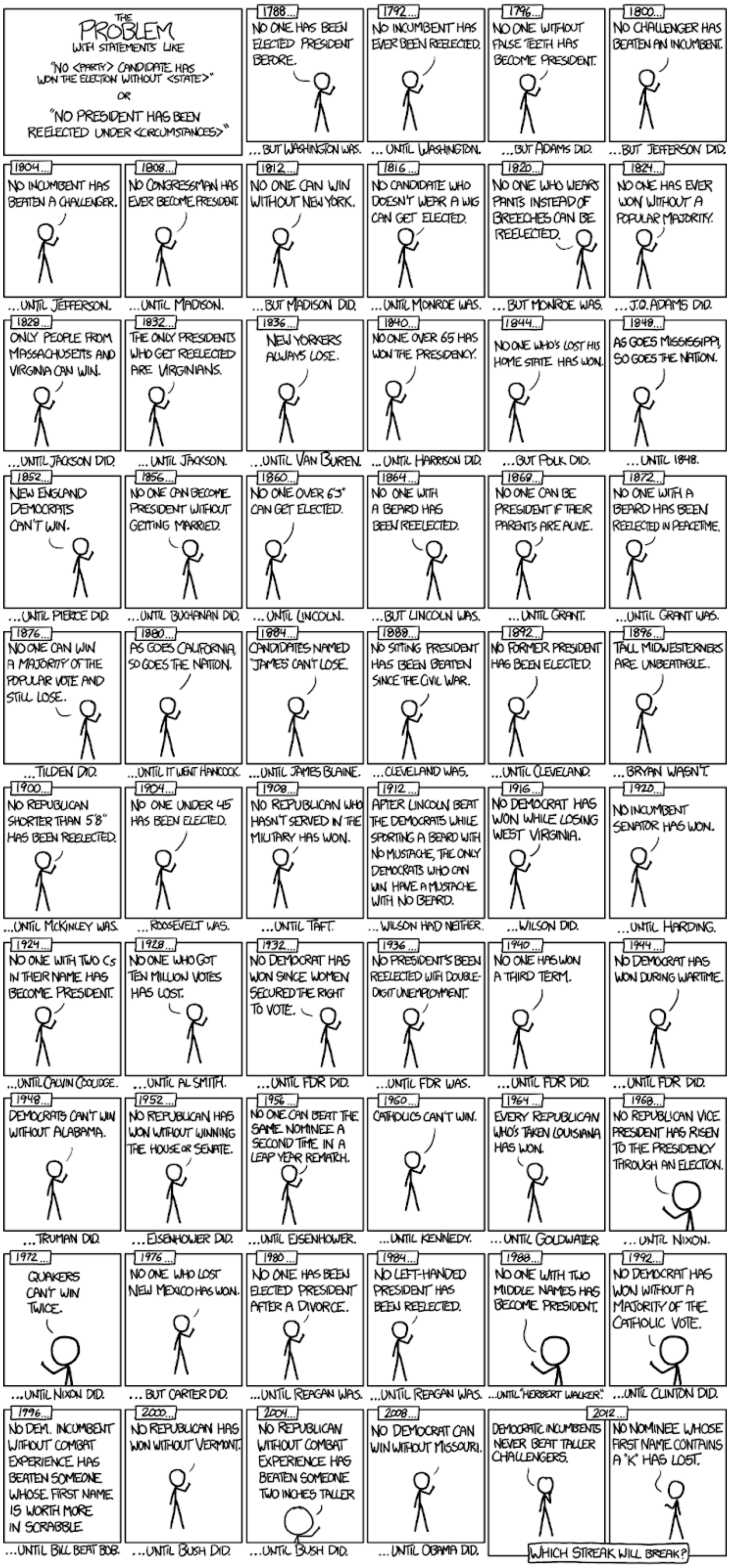
$f(2)=3$

$f(3)=5$

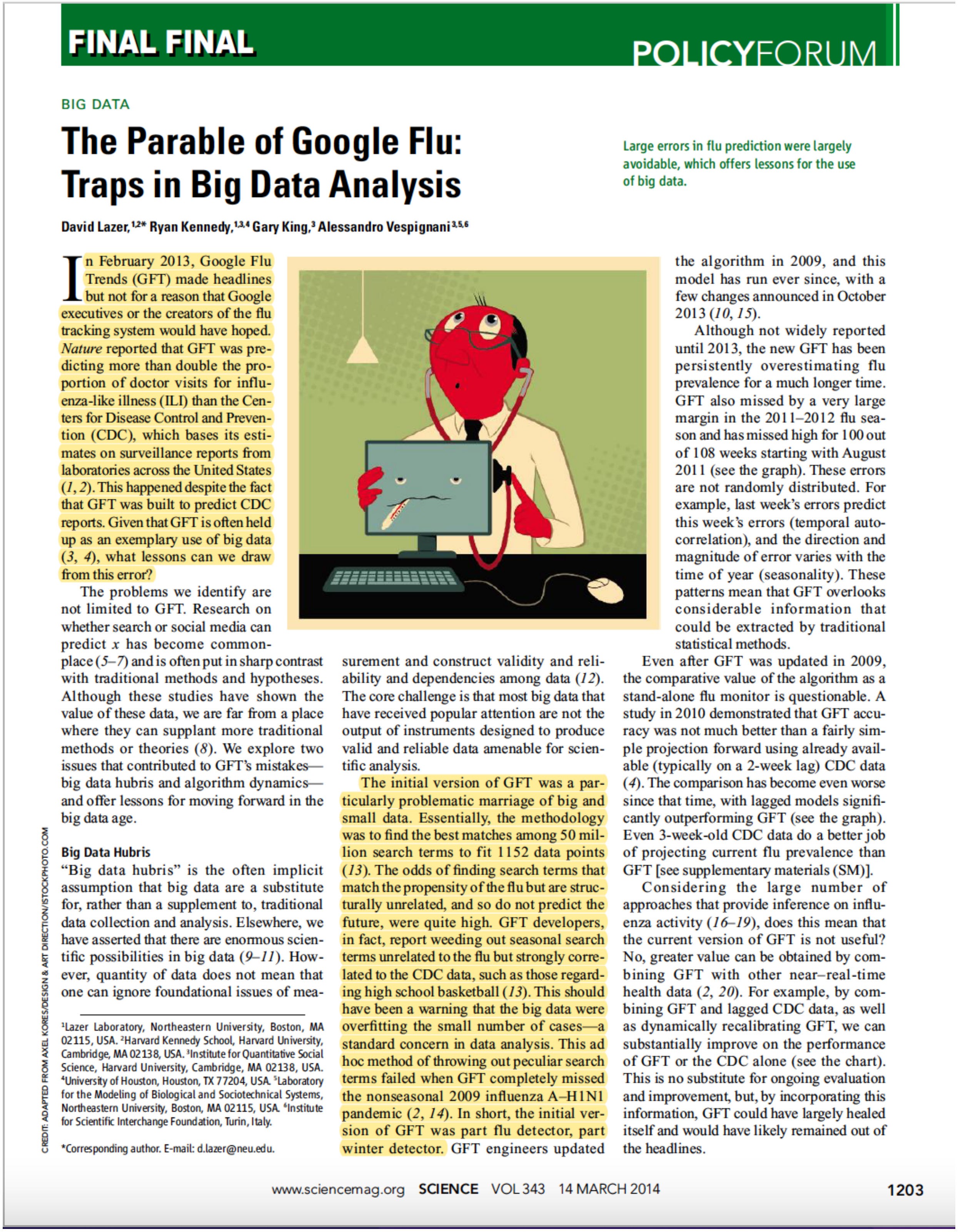
$f(4)=7$

$\boxed{f(5)=217341}$

XKCD President Comic Resource


THE PROBLEM WITH STATEMENTS LIKE "NO <PARTY> CANDIDATE HAS WON THE ELECTION WITHOUT <STATE>" OR "NO PRESIDENT HAS BEEN REELECTED UNDER <CIRCUMSTANCES>"
1788... NO ONE HAS BEEN ELECTED PRESIDENT BEFORE.
...BUT WASHINGTON WAS.
1792... NO INCUMBENT HAS EVER BEEN REELECTED.
... UNTIL WASHINGTON.
1796... NO ONE WITHOUT FALSE TEETH HAS BECOME PRESIDENT.
...BUT ADAMS DID.
1800... NO CHALLENGER HAS BEATEN AN INCUMBENT.
...BUT JEFFERSON DID.
1804... NO INCUMBENT HAS BEATEN A CHALLENGER.
...UNTIL JEFFERSON.
1808... NO CONGRESSMAN HAS EVER BECOME PRESIDENT.
...UNTIL MADISON.
1812... NO ONE CAN WIN WITHOUT NEW YORK.
...BUT MADISON DID.
1816... NO CANDIDATE WHO DOESN'T WEAR A WIG CAN GET ELECTED.
...UNTIL MONROE WAS.
1820... NO ONE WHO WEARS PANTS INSTEAD OF BREECHES CAN BE REELECTED.
...BUT MONROE WAS.
1824... NO ONE HAS EVER WON WITHOUT A POPULAR MAJORITY.
...J.Q. ADAMS DID.
1828... ONLY PEOPLE FROM MASSACHUSETTS AND VIRGINIA CAN WIN.
...UNTIL JACKSON DID.
1832... THE ONLY PRESIDENTS WHO GET REELECTED ARE VIRGINIANS.
... UNTIL JACKSON.
1836... NEW YORKERS ALWAYS LOSE.
...UNTIL VAN BUREN.
1840... NO ONE OVER 65 HAS WON THE PRESIDENCY.
...UNTIL HARRISON DID.
1844... NO ONE WHO'S LOST HIS HOME STATE HAS WON.
...BUT POLK DID.
1848... AS GOES MISSISSIPPI, SO GOES THE NATION.
...UNTIL 1848.
1852... NEW ENGLAND DEMOCRATS CAN'T WIN.
...UNTIL PIERCE DID.
1856... NO ONE CAN BECOME PRESIDENT WITHOUT GETTING MARRIED.
...UNTIL BUCHANAN DID.
1860... NO ONE OVER 6'3" CAN GET ELECTED.
...UNTIL LINCOLN.
1864... NO ONE WITH A BEARD HAS BEEN REELECTED.
...BUT LINCOLN WAS.
1868... NO ONE CAN BE PRESIDENT IF THEIR PARENTS ARE ALIVE.
...UNTIL GRANT.
1872... NO ONE WITH A BEARD HAS BEEN REELECTED IN PEACETIME.
...UNTIL GRANT WAS.
1876... NO ONE CAN WIN A MAJORITY OF THE POPULAR VOTE AND STILL LOSE.
...TILDEN DID.
1880... AS GOES CALIFORNIA, SO GOES THE NATION.
...UNTIL IT WENT HANCOCK.
1884... CANDIDATES NAMED "JAMES" CAN'T LOSE.
...UNTIL JAMES BLAINE.
1888... NO SITTING PRESIDENT HAS BEEN BEATEN SINCE THE CIVIL WAR.
...CLEVELAND WAS.
1892... NO FORMER PRESIDENT HAS BEEN ELECTED.
...UNTIL CLEVELAND.
1896... TALL MIDWESTERNERS ARE UNBEATABLE.
...BRYAN WASN'T.
1900... NO REPUBLICAN SHORTER THAN 5'8" HAS BEEN REELECTED.
...UNTIL MCKINLEY WAS.
1904... NO ONE UNDER 45 HAS BEEN ELECTED.
...ROOSEVELT WAS.
1908... NO REPUBLICAN WHO HASN'T SERVED IN THE MILITARY HAS WON.
...UNTIL TAFT.
1912... AFTER LINCOLN BEAT THE DEMOCRATS WHILE SPORTING A BEARD WITH NO MUSTACHE, THE ONLY DEMOCRATS WHO CAN WIN HAVE A MUSTACHE WITH NO BEARD.
...WILSON HAD NEITHER.
1916... NO DEMOCRAT HAS WON WHILE LOSING WEST VIRGINIA.
...WILSON DID.
1920... NO INCUMBENT SENATOR HAS WON.
...UNTIL HARDING.
1924... NO ONE WITH TWO Cs IN THEIR NAME HAS BECOME PRESIDENT.
...UNTIL CALVIN COOLIDGE.
1928... NO ONE WHO GOT TEN MILLION VOTES HAS LOST.
...UNTIL AL SMITH.
1932... NO DEMOCRAT HAS WON SINCE WOMEN SECURED THE RIGHT TO VOTE.
...UNTIL FDR DID.
1936... NO PRESIDENT'S BEEN REELECTED WITH DOUBLE-DIGIT UNEMPLOYMENT.
...UNTIL FDR WAS.
1940... NO ONE HAS WON A THIRD TERM.
...UNTIL FDR DID.
1944... NO DEMOCRAT HAS WON DURING WARTIME.
...UNTIL FDR DID.
1948... DEMOCRATS CAN'T WIN WITHOUT ALABAMA.
...TRUMAN DID.
1952... NO REPUBLICAN HAS WON WITHOUT WINNING THE HOUSE OR SENATE.
...EISENHOWER DID.
1956... NO ONE CAN BEAT THE SAME NOMINEE A SECOND TIME IN A LEAP YEAR REMATCH.
...UNTIL EISENHOWER.
1960... CATHOLICS CAN'T WIN.
...UNTIL KENNEDY.
1964... EVERY REPUBLICAN WHO'S TAKEN LOUISIANA HAS WON.
...UNTIL GOLDWATER.
1968... NO REPUBLICAN VICE PRESIDENT HAS RISEN TO THE PRESIDENCY THROUGH AN ELECTION.
...UNTIL NIXON.
1972... QUAKERS CAN'T WIN TWICE.
...UNTIL NIXON DID.
1976... NO ONE WHO LOST NEW MEXICO HAS WON.
... BUT CARTER DID.
1980... NO ONE HAS BEEN ELECTED PRESIDENT AFTER A DIVORCE.
...UNTIL REAGAN WAS.
1984... NO LEFT-HANDED PRESIDENT HAS BEEN REELECTED.
...UNTIL REAGAN WAS.
1988... NO ONE WITH TWO MIDDLE NAMES HAS BECOME PRESIDENT.
...UNTIL "HERBERT WALKER."
1992... NO DEMOCRAT HAS WON WITHOUT A MAJORITY OF THE CATHOLIC VOTE.
...UNTIL CLINTON DID.
1996... NO DEM. INCUMBENT WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE WHOSE FIRST NAME IS WORTH MORE IN SCRABBLE
...UNTIL BILL BEAT BOB.
2000... NO REPUBLICAN HAS WON WITHOUT VERMONT.
...UNTIL BUSH DID.
2004... NO REPUBLICAN WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE TWO INCHES TALLER
...UNTIL BUSH DID.
2008... NO DEMOCRAT CAN WIN WITHOUT MISSOURI.
...UNTIL OBAMA DID.
2012...
DEMOCRATIC INCUMBENTS NEVER BEAT TALLER CHALLENGERS.
NO NOMINEE WHOSE FIRST NAME CONTAINS A "K" HAS LOST.
WHICH STREAK WILL BREAK?

Google Flu Resource: *Read just the highlighted paragraphs of this article.*



FINAL FINAL
POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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.

CREDIT: ADAPTED FROM AXEL KORES/DESIGN & ART DIRECTION/ISTOCKPHOTO.COM

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)

Creating Decision Tree Models

A guide to creating and interpreting decision tree models by Sara Fergus

Summary

In this lesson, students will read and create decision trees. In doing so, students will have to identify what sort of datasets are effectively modeled using decision trees, and will have to clean datasets to prepare for numeric decision trees using sklearn. At the end of the lesson, students will practice translating between mathematical models and representations (decision trees) and real-world conclusions.

Objectives

The students will be able to . . .

- Clean data sets to prepare for analysis
- Carefully consider the ethical and statistical implications of data cleaning decisions
- Identify appropriate datasets to be analyzed and communicated as decision trees
- Draw conclusions from decision trees

Standards Alignment

- **DS.6:** The student will justify the design, use and effectiveness of different forms of data visualizations.
- **DS.8:** The student will be able to acquire and prepare big data sets for modeling and analysis.
- **DS.9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.10:** The student will be able to summarize and interpret data represented in both conventional and emerging visualizations.
- **DS.12:** The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.
- **DS.13:** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

Materials

- Flow charts for warm-up (see below)
- Slideshow for discussion (Google Slides or make a copy)
- Teacher guide for Python decision trees (.ipynb, heart_2020_cleaned.csv)
- Student worksheet (.ipynb, & key) & datasets (titanic.csv, Social_Network_Ads.csv)
- Decision Tree Checklist (see below, Google Doc, PDF, or make a copy)

Vocabulary

Term	Definition
Decision Tree	A decision tree is a flowchart used to classify cases in machine learning algorithms. Using the Python library sklearn, decision trees are read from top to bottom. At each node, the reader should follow the left arrow if the condition is true for a given case and the right arrow if the condition is false.
Node	A node is one rectangle in the decision tree. Most nodes give a statement about a case which can be true or false (for example, have you had a stroke?). The truth value of the node for the given case indicates which direction to follow the tree. A leaf node, however, does not give a true or false statement. Instead, it gives the most likely categorization for a case given the attributes previously considered. For example, a leaf node may say "likely to develop heart disease in the next 10 years" or "unlikely to develop heart disease in the next 10 years".
Sample	A sample is one case in the machine learning algorithm that creates a decision tree. In the sklearn decision tree, each node has a number that represents the number of samples in the training set that fall into that particular node.

Day 1 Outline

Formative Assessment Notes

1. **Warm Up:** Have students choose one of these flowcharts. Then, instruct them to come up with a fake person/situation. In their journals, write the person/situation, the path, and the result.

   For example: *My Flowchart: Should I Dump this Person By Text? Situation: I hung out with this guy four times, I won't see them at work, they are not seeing anyone new because they are in a coma. In fact, they have been in a coma the whole time. According to the flowchart, it is okay for me to dump them by text.*

2. **Interpreting the Decision Tree:** Display this slide show.

   - Show slide 2. For one minute, have students write in their journal what they notice, what they wonder, and any conclusions they can draw from the tree. Tell them to not worry about what they don't understand, but to interpret what they do understand.
   - Work through the rest of the slides, facilitating discussion about how to understand the decision tree.

As students share notice and wonders, write their observations on the board. As you go through the slide show, refer to observations that relate to each slide.

3. **Decision Trees Tutorial**: Walk students through this decision tree using heart disease data.

   Live-code each block, while having students copy down onto their own notebook. Stop with students to discuss as you go, where indicated.

Teacher notes are included within the decision tree teacher guide notebook.

4. **Decision Trees Practice:** Have students create a decision tree using the Titanic data with this worksheet (worksheet includes link to data and the example you did as a class)

   As an extension or for additional practice, students may also create a tree for the Social Network Ads data

The answer key to the Titanic decision tree is available here.

5. **Assessment:** Instruct students to find their own data set to create a decision tree. Have them fill out this worksheet as they go, where they will be justifying data choices, sketching the tree, and interpreting.

Encourage students to check in with you after each step, to ensure that they are on the right track.

Assessment Strategies

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

Decision Tree Checklist

The *Decision Tree Checklist* below is designed to provide an opportunity for you to assess student understanding of decision trees. Use the rubric below to evaluate students' work on the checklist:

	The student is able to . . .	***Yes***	***No***	***Notes***
Sourcing & Preparing Data	Find data suitable for decision tree predictive modeling, and explain their decision-making			
Creating the Decision Tree	Write code to produce a decision tree model using Python			
Tree Interpretation	Explain the following in their own words: 1. What their tree is predicting 2. What attributes are most significant for their prediction 3. Whether or not their model is a good predictive tool, and why			

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 who may need additional support in live coding could be given the student worksheet, which includes the finished code from the tutorial, from the start of the lesson.

Extensions

As an extension, students could create another decision tree using either the social network ads data, or data of their choice.

Students may also experiment more with the heart disease data set. Prompt students to consider why the decision tree is so different when all of the cases are included. You may even encourage them to experiment with the pandas sample function to get random subsets of the data and observe the changes in the decision tree.

Decision Tree Checklist

View the Google Doc, PDF, or make a copy

Objectives: In this activity, you will be showing that you can:

1. Prepare data for interpretation
2. Create a decision tree
3. Interpret the decision tree

Directions: To start, find a data set that interests you, and could be analyzed using a decision tree. You can use Kaggle, the Google dataset search, or any other website.

Explore the data

Justify: Why does a decision tree make sense for this data? Be sure to mention which attribute you plan to predict, and which could be used as predictors.

Clean the Data

Now, clean your data. Keep these questions in mind as you work:

- Which cases, if any, will need to be dropped entirely?
- Which attributes, if any, need to be changed to dummies using `get_dummies`?
- Which attributes, if any, need to be changed to numeric using `replace`?
- How will I translate between the values in the dataset and numeric values? What data cleaning considerations should I make? Could any of my decisions lead to bias?

Describe: What data cleaning decisions did you make, and why did you make them? Is it possible that your data cleaning decisions introduced bias?

Create the Tree

Create your decision tree in Python. Keep these questions in mind as you work:

- What will I choose for my `max_leaf_nodes`? Why?
- What are my class names?

Sketch: Sketch your decision tree. Do not copy exactly. Instead, sketch a version that includes the most important information and would make sense to someone with no Data Science background.

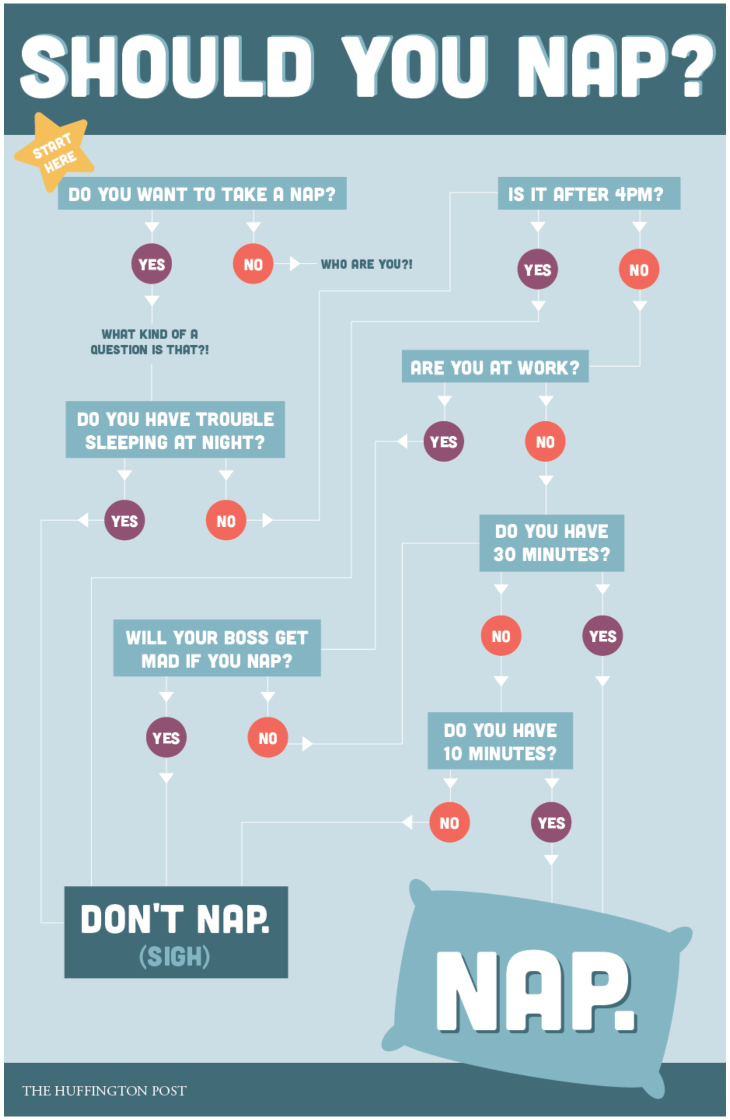
Interpret the Tree

Now, interpret your tree. Keep these questions in mind:

- Which attributes did the decision tree end up using?
- What key values for those attributes make the difference in final prediction (ex, Sleep Time being greater or less than 8.5 hours)?
- Where do most samples fall?
- Does the tree include all possible outcomes?

Summarize: Write a summary of your tree. Your interpretation should make sense to someone with no Data Science background.

Warm-Up Flow Charts



SHOULD YOU NAP?

START HERE

DO YOU WANT TO TAKE A NAP?

YES

NO

WHO ARE YOU?!

WHAT KIND OF A QUESTION IS THAT?!

IS IT AFTER 4PM?

YES

NO

ARE YOU AT WORK?

YES

NO

DO YOU HAVE TROUBLE SLEEPING AT NIGHT?

YES

NO

DO YOU HAVE 30 MINUTES?

NO

YES

WILL YOUR BOSS GET MAD IF YOU NAP?

YES

NO

DO YOU HAVE 10 MINUTES?

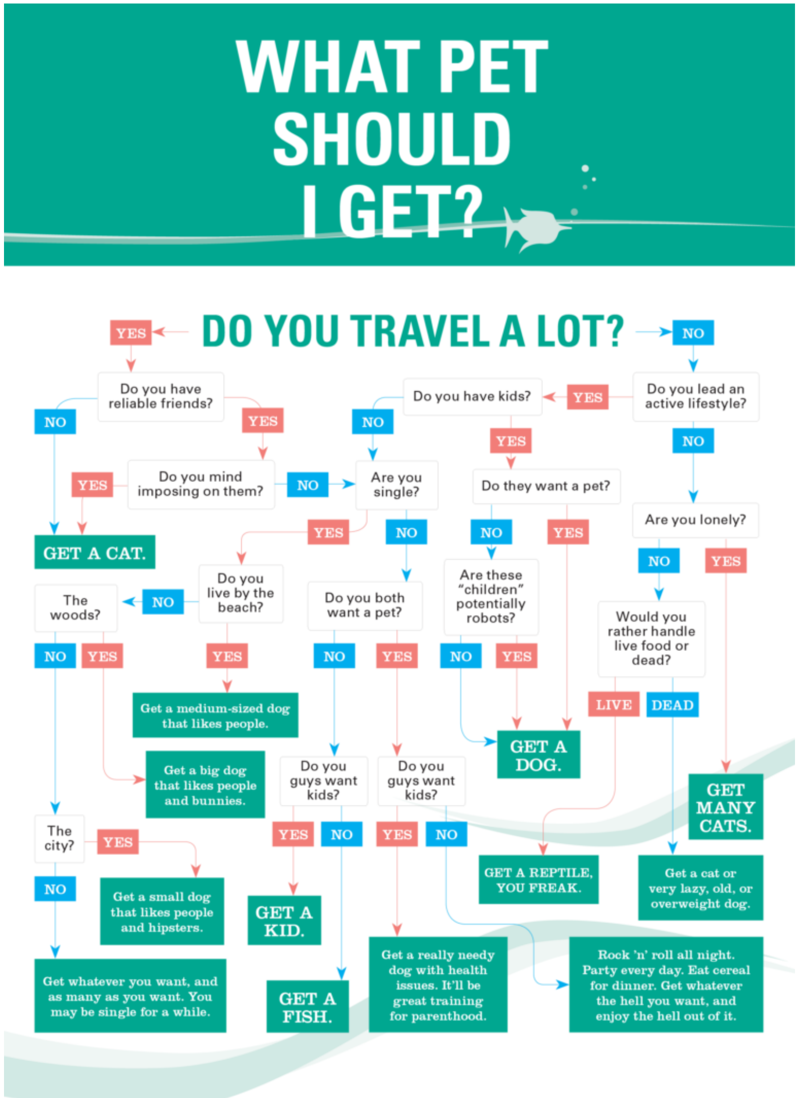
NO

YES

DON'T NAP. (SIGH)

NAP.

THE HUFFINGTON POST



WHAT PET SHOULD I GET?

DO YOU TRAVEL A LOT?

YES

NO

Do you have reliable friends?

Do you have kids?

Do you lead an active lifestyle?

NO

YES

NO

YES

YES

NO

Do you mind imposing on them?

NO

Are you single?

Do they want a pet?

Are you lonely?

YES

YES

NO

NO

YES

NO

YES

GET A CAT.

The woods?

NO

Do you live by the beach?

Do you both want a pet?

Are these "children" potentially robots?

Would you rather handle live food or dead?

NO

YES

YES

NO

YES

NO

YES

LIVE

DEAD

Get a medium-sized dog that likes people.

Get a big dog that likes people and bunnies.

Do you guys want kids?

Do you guys want kids?

GET A DOG.

GET MANY CATS.

The city?

YES

YES

NO

YES

NO

GET A REPTILE, YOU FREAK.

Get a cat or very lazy, old, or overweight dog.

NO

Get a small dog that likes people and hipsters.

GET A KID.

Get whatever you want, and as many as you want. You may be single for a while.

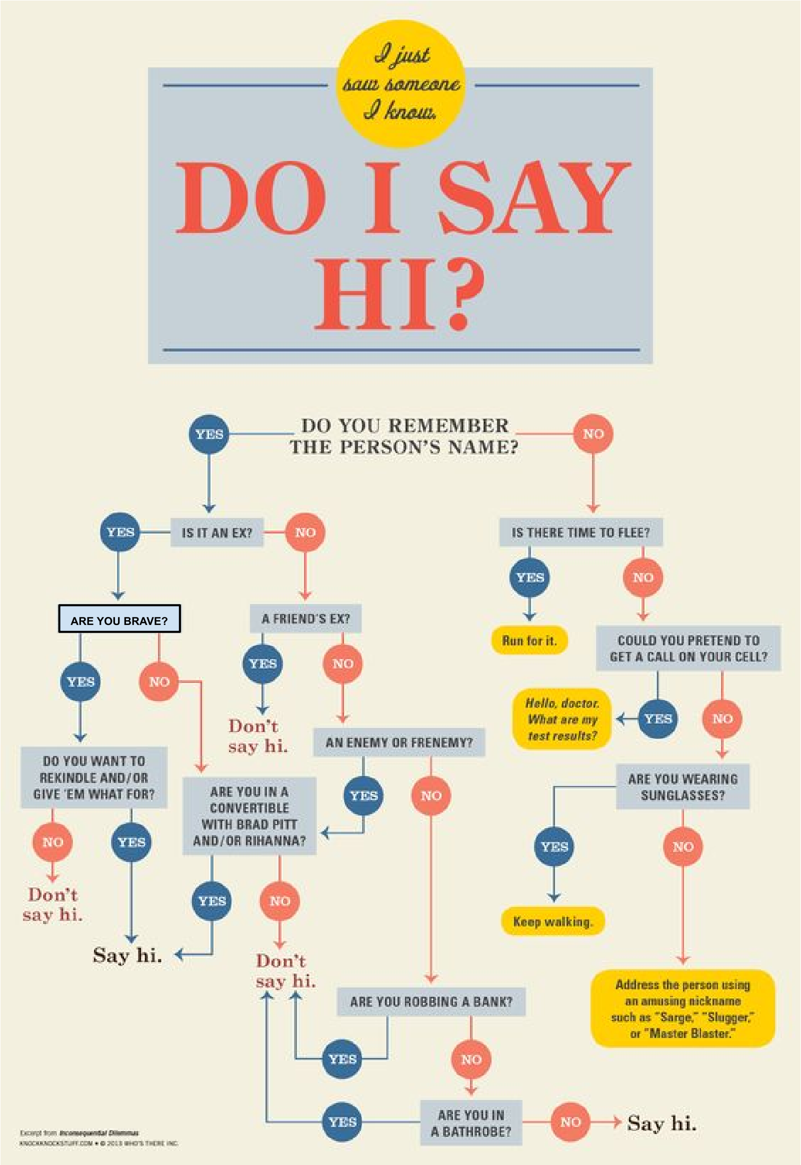
GET A FISH.

Get a really needy dog with health issues. It'll be great training for parenthood.

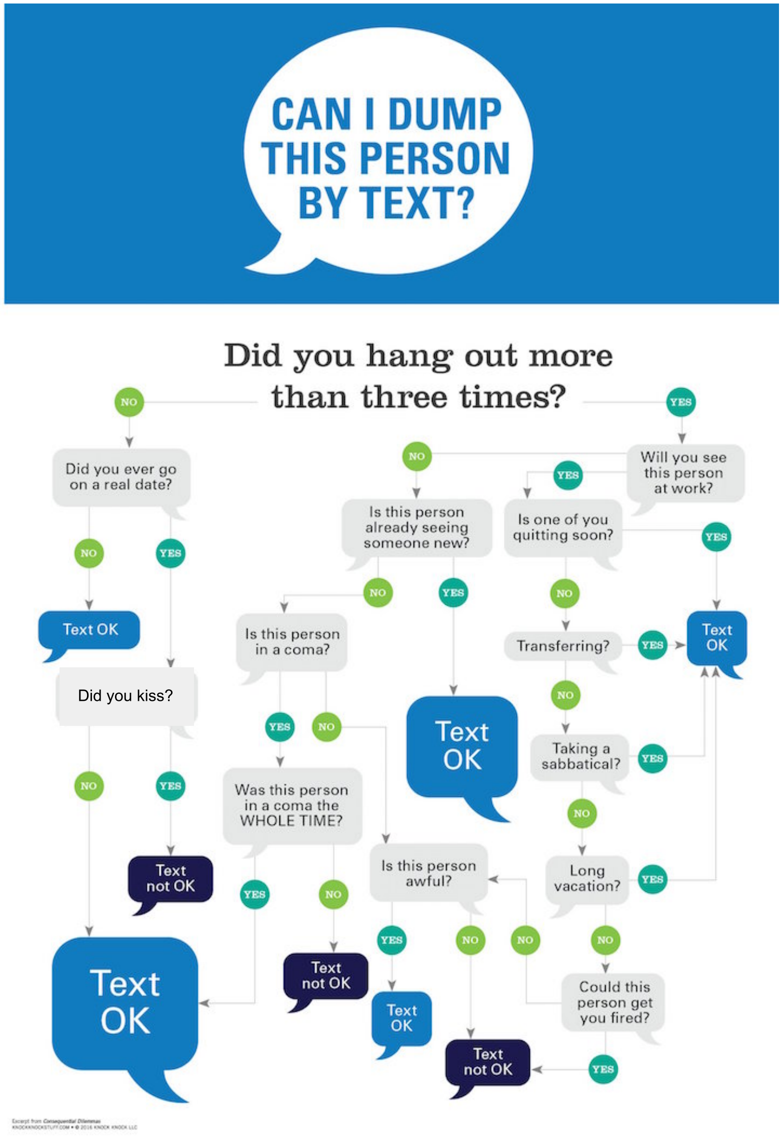
Rock 'n' roll all night. Party every day. Eat cereal for dinner. Get whatever the hell you want, and enjoy the hell out of it.

Excerpt from *Consequential Dilemmas* by Knock Knock

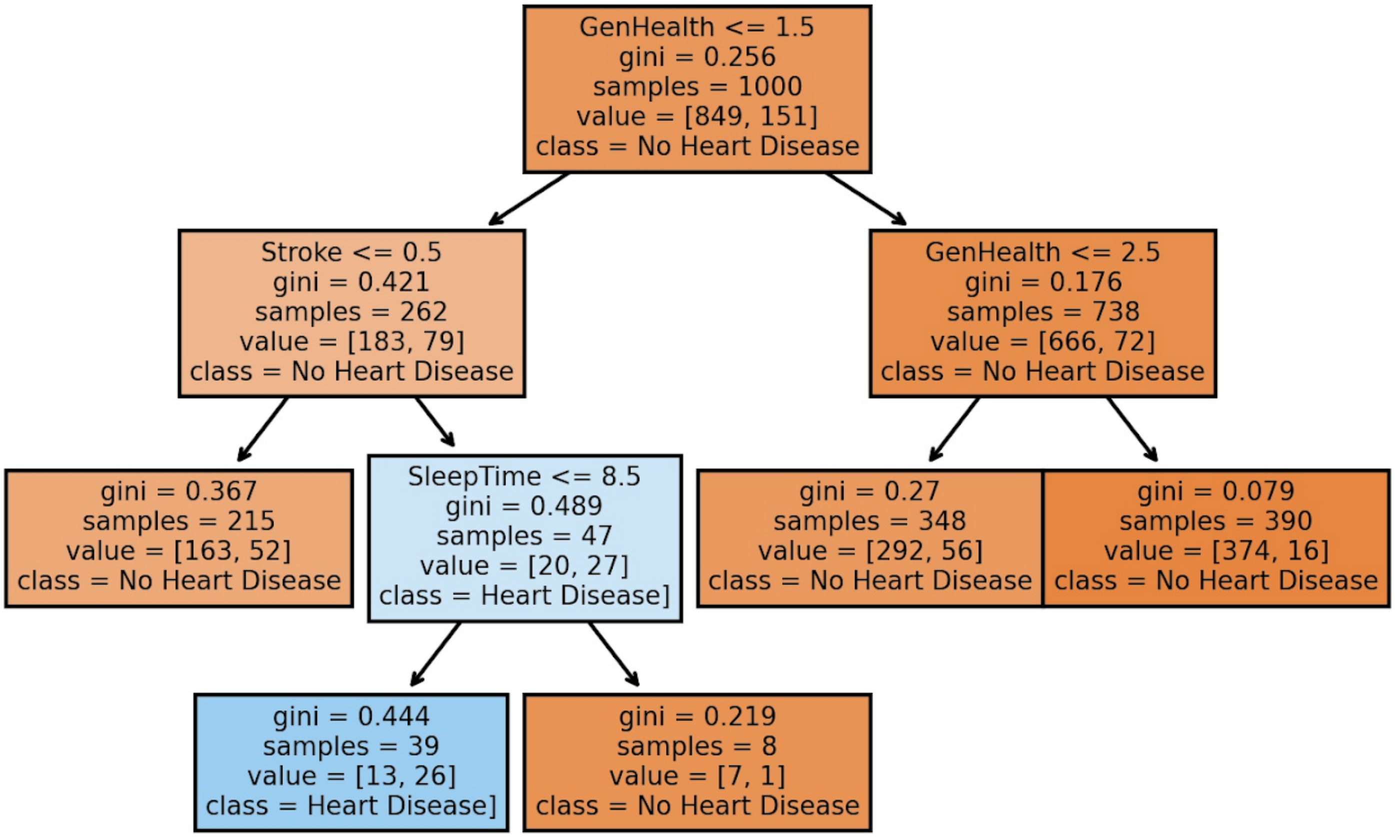
KNOCKKNOCKSTUFF.COM ▪ © 2017 KNOCK KNOCK LLC



I just saw someone I know.
DO I SAY HI?
DO YOU REMEMBER THE PERSON'S NAME?
YES
NO
IS IT AN EX?
YES
NO
ARE YOU BRAVE?
YES
NO
DO YOU WANT TO REKINDLE AND/OR GIVE 'EM WHAT FOR?
NO
Don't say hi.
YES
Say hi.
A FRIEND'S EX?
YES
Don't say hi.
NO
AN ENEMY OR FRENEMY?
YES
ARE YOU IN A CONVERTIBLE WITH BRAD PITT AND/OR RIHANNA?
YES
NO
Don't say hi.
NO
ARE YOU ROBBING A BANK?
YES
NO
ARE YOU IN A BATHROBE?
YES
NO
Say hi.
IS THERE TIME TO FLEE?
YES
Run for it.
NO
COULD YOU PRETEND TO GET A CALL ON YOUR CELL?
YES
Hello, doctor. What are my test results?
NO
ARE YOU WEARING SUNGLASSES?
YES
Keep walking.
NO
Address the person using an amusing nickname such as "Sarge," "Slugger," or "Master Blaster."
Excerpt from Inconsequential Dilemmas
KNOCKKNOCKSTUFF.COM • © 2013 WHO'S THERE INC.



CAN I DUMP THIS PERSON BY TEXT?
Did you hang out more than three times?
NO
Did you ever go on a real date?
NO
Text OK
YES
Did you kiss?
NO
Text OK
YES
Text not OK
YES
Will you see this person at work?
NO
Is this person already seeing someone new?
NO
Is this person in a coma?
YES
Was this person in a coma the WHOLE TIME?
YES
NO
Text not OK
NO
Is this person awful?
YES
Text OK
NO
Text not OK
YES
Text OK
YES
Is one of you quitting soon?
YES
Text OK
NO
Transferring?
YES
NO
Taking a sabbatical?
YES
NO
Long vacation?
YES
NO
Could this person get you fired?
YES
NO
Excerpt from Consequential Dilemmas
KNOCKKNOCKSTUFF.COM • © 2016 KNOCK KNOCK LLC


GenHealth <= 1.5
gini = 0.256
samples = 1000
value = [849, 151]
class = No Heart Disease
Stroke <= 0.5
gini = 0.421
samples = 262
value = [183, 79]
class = No Heart Disease
GenHealth <= 2.5
gini = 0.176
samples = 738
value = [666, 72]
class = No Heart Disease
gini = 0.367
samples = 215
value = [163, 52]
class = No Heart Disease
SleepTime <= 8.5
gini = 0.489
samples = 47
value = [20, 27]
class = Heart Disease]
gini = 0.27
samples = 348
value = [292, 56]
class = No Heart Disease
gini = 0.079
samples = 390
value = [374, 16]
class = No Heart Disease
gini = 0.444
samples = 39
value = [13, 26]
class = Heart Disease]
gini = 0.219
samples = 8
value = [7, 1]
class = No Heart Disease

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 Unplugged Data Science *&* Data Science with CODAP *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

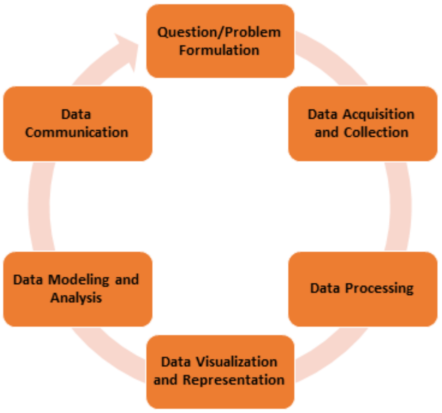
Standards Alignment

- **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.  Question/Problem Formulation Data Acquisition and Collection Data Processing Data Visualization and Representation Data Modeling and Analysis Data Communication
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

Throughout the lesson, students will complete a full iteration of the data cycle by modeling question formulation, data collection, analysis, visualization and the modeling processes. This 2 day lesson includes a guide for students to use on a class Data Science Research Project. On day 2, students will be given local data sets to conduct an Exploratory Analysis Data Science Project to reiterate the data cycle skills.

Note: Variations on this lesson appear in the CodeVA Unplugged Data Science *&* Data Science with CODAP *sequences.*

Objectives

The students will be able to . . .

- Complete at least two iterations of the data cycle
- Employ data cycle skills developed throughout the course, including:
    a. constructing a strong data question
    b. collecting/acquiring reliable and useful data,
    c. processing data effectively, controlling for bias,
    d. creating visualizations and models to understand data,
    e. presenting findings as a data story or a data science write-up
- Conduct a Research Project
- Conduct an Exploratory Analysis Project

Standards Alignment

- **DS.1:** The student will identify specific examples of real-world problems that can be effectively addressed using data science.
- **DS.2:** The student will be able to formulate a top-down plan for data collection and analysis, with quantifiable results, based on the context of a problem.
- **DS.3:** The student will recognize the importance of data literacy and develop an awareness of how the analysis of data can be used in problem solving to effect change and create innovative solutions.
- **DS.6:** The student will justify the design, use and effectiveness of different forms of data visualizations.
- **DS.9:** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.13:** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

Materials

- Data Science Project Design Menu & Design Scaffold (view Google Drawing or make a copy)
- Project Write-Up Template (see below)
- Data Science Design Flowchart (view Google Doc or make a copy) & example
- Data Science Research Question Distiller (view Google Doc or make a copy)
- 11X17 Graph Paper (view Google Doc printable or make a copy)
- Grocery Store Marketing Analytics (view Google Sheet or make a copy)
- Data Sets Websites: Data World (Virginia), Open Data Network (Virginia), Kaggle Datasets
- Data Sets: Diamond Prices (CSV), Netflix Daily Top 10 (Kaggle), U.S. International Air Traffic Data (1990-2020, Kaggle), Forbes Highest Paid Athletes (1990-2020, Kaggle), Antarctica Penguin Data (Kaggle), Climate Change: Earth Surface Temperature Data (Kaggle), Real / Fake Job Posting Prediction (Kaggle), Bigfoot Sightings (data.world), Harry Potter Data Sets (Kaggle)

Vocabulary

Term	Definition
The Data Cycle	The Data Cycle is a framework of data science. In the data cycle, a data scientist will ask a question, collect or acquire the data needed to answer that question, perform a data analysis (including pre-processing and processing, visualizations, models, and general analysis) and then communicate their findings. Once they have communicated their findings, they will notice that new questions have been brought to light. These may come from a further question, the need for an additional attribute of the data, the need for different data (different people, different place, different time, etc), or a critique of the data analysis process. Once that question is created, the data cycle is repeated.
Research Project	A research project follows the traditional data cycle: 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.

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 for a rubric.

Day 2 Outline

Formative Assessment Notes

1. **Warm Up:** Given the following Diamonds Data Set, have students use Python and exploratory analysis to support or disprove:

   "The bigger the diamond the better it is."

Students should identify attributes that are impacted by the size of a diamond including clarity, cut, and color.

2. **Exploratory Data Analysis Project:** Group students in pairs (or individually if desired). Provide the students with 2-3 data sets. Use the following sites to explore local data sets:

   - Data World (Virginia)
   - Open Data Network (Virginia)
   - Kaggle Datasets

   Or have students choose from the following:

   - Netflix Daily Top 10 (March 2020 - March 2022)
   - U.S. International Air Traffic Data (1990-2020)
   - Forbes Highest Paid Athletes (1990-2020)
   - Antarctica Penguin Data
   - Climate Change: Earth Surface Temperature Data
   - Real / Fake Job Posting Prediction
   - Bigfoot Sightings
   - Harry Potter Data Sets

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

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 for a rubric.

Some Accommodations & Extensions

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

Accommodations

You may choose to create groups strategically in order to balance student's strengths and weaknesses, or in order to create groups that you intend to spend more time supporting.

Some students may benefit from an abbreviated version of the write up template, that includes the title, research question, data analysis and findings (together), and conclusion.

Extensions

For students who finish early, you may encourage them to create a tactile model (see project examples). You could also choose to have them supplement their project with an analysis of an additional data set, or ask them to collect supplemental data to address questions that arise in their initial analysis.

Practice Project Rubric

	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) for a data set that can be used for your project
- **Plan Visualizations:** Based on the types of data you collected, what sorts of visualizations make sense? What pieces of the data relate to your research question, and how can you represent them? Write some ideas here:
- **Plan Models:** Determine whether a descriptive or predictive model can help you tell your story.

Data to Represent	Possible Visualizations and/or models

- **Create Visualizations:** Make sure they are accurate, clear and clearly labeled, and presented in a way that meets the goals you set out above.
- **Create Models:** Create your model using Python.
- **Answer your Question:** Using the Write Up Template, draw connections between your data and multiple representations of your data, and how they answer your research question. Make sure your findings are clear and directly related to the research question. Make sure your final argument is clear.
- **Reflect on the Data:** Consider your findings and how they relate to the real world. Share your reflection through a solution or call to action, an infographic, or a reflective portion of your write-up..

Part 3: Share your Findings

Share your project with your community. If you created an infographic or video, you could share on your personal social media account, or ask to share on your school's social media account.

- Create an artifact to communicate your findings. You should use visualizations created with Python, but you may choose to add to the overall infographic using sites like Canva.

Part 4: Reflection

Student Reflection
Describe what went well.
Describe what you struggled with.
Describe one way you would improve on your project.
Describe a future step for data collection or analysis.
Share your personal progress throughout the project.
Reflect on your management of this project. ● Did you meet most deadlines? ● Did you use your class time wisely?
Describe your overall experience with this project.

Template: A Write-Up

[Title of Write-Up]

[Subtitle]

The Title: The most important thing about your title is that it communicates what the paper is about. Be creative! If you have a creative title that does not fully communicate the topic of the paper, add a subtitle.

Research Question and Background

Here, clearly state your research question. Make sure you take time before you start to develop a strong research question. Briefly explain any context that is necessary for understanding your research question. Be sure to explain *why* your question is important, why should the reader care about your question?

Then, provide some background research on the question.

Tone: The tone of your paper should be relatively scientific. Avoid "talking to your reader" ("I bet you are wondering…"). However, it is not necessarily bad to describe your personal interest.

Data and Data Collection

First, describe where and how you collected / found your data. Then, describe the data itself— for example say how many entries there are, or how many questions were asked. If you had to do any data cleaning, describe the decisions you made and why you made those decisions.

Data Analysis

Now you get to share your findings! This section is where you put your well-labeled visualizations and models. In text, make connections between the visualizations and the models. Briefly discuss what each visualization or model shows. Make sure any visualizations are referenced by number and labeled. For example, in Figure 1 you see a brief description of what is being shown.

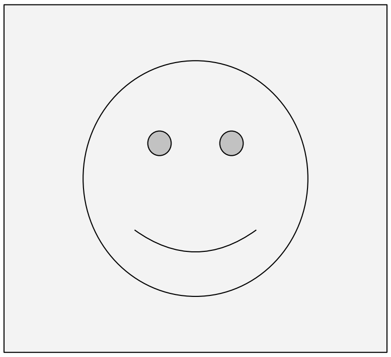


Figure 1. A smiley face

Findings

Now that you have run your data analysis, answer your research question. In answering your research question, directly reference research that you conducted and trends/patterns shown in the visualizations and models you created. Make sure your question is clearly answered.

Conclusion

Here, discuss any problems in your analysis. For example, how might different data cleaning decisions have affected your findings? Or different data collection techniques?

Then, discuss any questions that arise from your analysis. Maybe there is a part of your research question you were not able to fully answer. Maybe there is an interesting follow-up question you thought of while conducting your data analysis.

This is also where you can give recommendations on how to enact positive change based on the findings in your data analysis.

Data Science HS with Python

Python Summative Project

Using coding to tell an engaging data story by Sara Fergus

Summary

Throughout this unit, students have learned to use Python to tell a data story. In this project, students will use or collect data to explore and analyze a topic that interests them. They will run a data analysis, and then present their findings in a meaningful deliverable that can inspire deep thought or action. To tell their story, they will draw on the data questioning, visualization, and modeling skills they have learned.

Objectives

The students will be able to . . .

- Ask a relevant question that can be answered with data
- Conduct an exploratory data analysis
- Produce high quality and relevant visualizations to communicate their findings
- Produce a model to represent their data or make a prediction
- Summarize a data analysis by connecting the question to findings and visualizations to next steps and proposals
- Propose solutions to structural problems to representatives from the community

Standards Alignment

- **DS.1** The student will identify specific examples of real-world problems that can be effectively addressed using data science.
- **DS.2** The student will be able to formulate a top-down plan for data collection and analysis, with quantifiable results, based on the context of a problem.
- **DS.5** The student will use storytelling as a strategy to effectively communicate with data.
- **DS.8** The student will be able to acquire and prepare big data sets for modeling & analysis.
- **DS.9** The student will select and analyze data models to make predictions, while assessing accuracy and sources of uncertainty.
- **DS.12** The student will be able to select and utilize appropriate technological tools and functions within those tools to process and prepare data for analysis.
- **DS.13** The student will be able to select and utilize appropriate technological tools and functions within those tools to analyze and communicate data effectively.

Materials

- The student-facing *Project Frame* document (see below, view Google Doc, or make a copy)
- The *Data Science Project Design Menu* (view Google Drawing or make a copy)
- The *Project Write-Up Template* (see below, view Google Doc, or make a copy)
- A Python IDE that can support .ipynb or .py files. See *A Note on Python IDEs.*
- *Political Tweets* example project (see below)

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 *15 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 *16 Project Practice* lesson plan from this sequence to help students go through the entire project process in a less open-ended way so they can see what sort of work they should plan to do during their self-directed project.
- **Analyze Project Examples:** Have students analyze the examples linked in the *Materials* section (and in the project frame document) as a group to develop a sense of what a successful project might look like.

Using This Document

The following pages in this document are intended to be filled out by the *student* as they work through their summative project. It takes the form of a checklist, showing the different steps students should go through as they plan, execute, and share their project. You can distribute this document by printing it (leaving off the first two pages), or digitally by making a copy of this Google Doc.

You can structure students' engagement with the project in several ways. We do not provide explicit guidance because your scaffolding and guideline should be responsive to your students (who we will never know). Here are some suggestions:

- **Pacing:** If you want to provide a pacing guide, set deadlines for each step in the checklist based on how much time you think students should take to complete them.
- **Scaffolding Choice:** Sometimes, students will have a difficult time finding a data set relevant to their interests. Consider providing options for them to choose from (see *16 Project Practice*) if they are not successful in choosing their own.
- **Modifying the Design Menu:** The *Project Design Menu* is available as an editable Google Drawing; replace the options we have provided with options that you think are better suited to your students' inquiries. Consider filling out the design menu in collaboration with students so you can guide their project goal-setting process.

Data Science with Python Project Prompt

Relationships in data help us write a data story, and our data stories help us make meaning of the world around us. In this project, you will explore a dataset to highlight a meaningful relationship. You could look for a simple relationship, like correlation, or a different pattern of data relationships. Then, organize your findings into a simple visual that you can share with the community, so that they can be aware of the relationship you found. This may be an infographic, a video, a physical representation, or anything else.

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

- Python Project: Political Tweets (see below)

1. Design the Project

□ **Brainstorm Ideas**: Consider the following questions as you plan your data science project:

- Where in your community might something be underrepresented or hidden?
- How could your data analysis show something that might be underrepresented or hidden?
- How could it contribute to a cause that you are passionate about?
- How does it change with time or interpretation?
- How does it show an experience that many people in your community have?

Write some ideas here:

□ **Choose One Idea**: Base your idea on interest and the feasibility of data collection. Turn your idea into a well-constructed *data question*. Write your data question here:

□ **Plan Your Project:** Use the *Project Design Menu* (view Google Drawing, or make a copy) to create a project plan, where you define your data question, set goals for what data science skills you will use, and define who your audience will be when you present your work at the end of the project.

2. Complete the Project

☐ **Acquire the Data**: Explore the resources provided throughout the module (or another resource) for a data set that can be used for your project, or (if appropriate) collect the data you need.

☐ **Plan Visualizations:** Based on the types of data you collected, what sorts of visualizations make sense? What pieces of the data relate to your research question, and how can you represent them? Write some ideas here:

Data to Represent	Possible Visualizations

□ **Plan Models:** Determine what kinds of modeling can help you tell your data story.

Possible Predictor(s)	Possible Prediction	Appropriate Model

□ **Create Visualizations:** Make sure they are accurate, clear and clearly labeled, and presented in a way that meets the goals you've articulated 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 <u>not</u> make sense with your data question, explain why here:

- ☐ **Answer your Question:** In a write-up, draw connections between your data and multiple representations of your data, and how they answer your research question. Make sure your findings are clear and directly related to the research question. Make sure your final argument is clear.

- ☐ **Communicate your Findings:** Create an artifact to communicate your findings with the general public. You should use visualizations created with Python, 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.

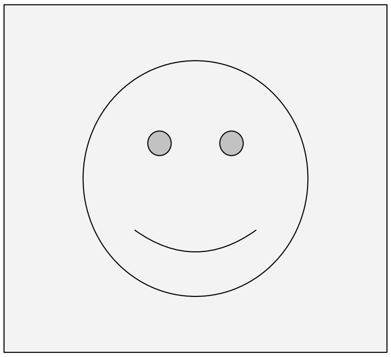


Figure 1. A smiley face

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.

Political Tweets Example Project 🐦

This project is Python-based exploratory data analysis.

Classroom Highlights

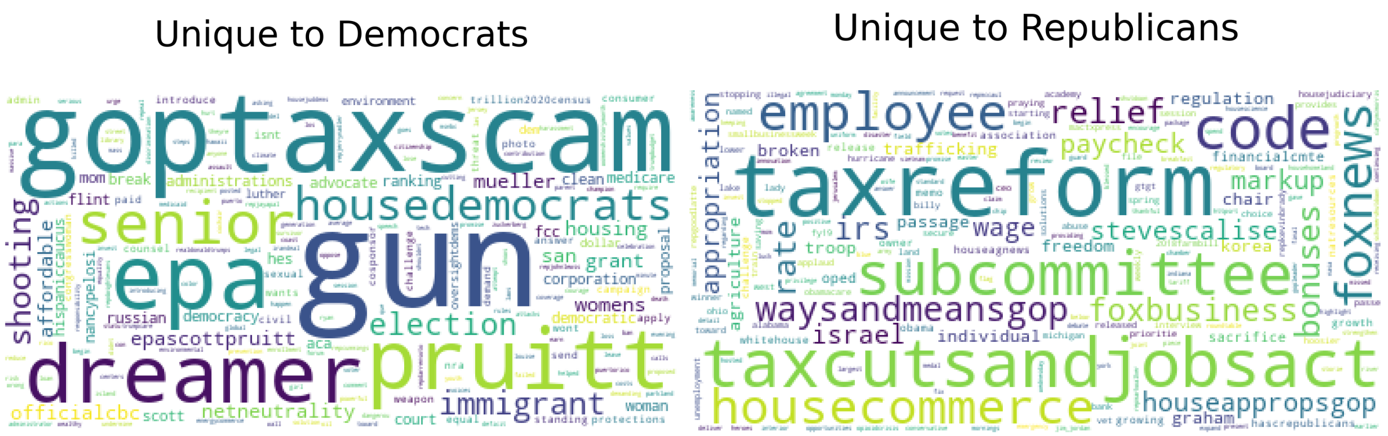
This exemplar demonstrates:

1. Basic Data Science Python
2. Python Data Science Libraries (`pandas`, `matplotlib`, `wordcloud`, `collections`)
3. Controversial / Political topics
4. Data Decisions and Ethics in Representations
5. Emerging Visualizations

The Project

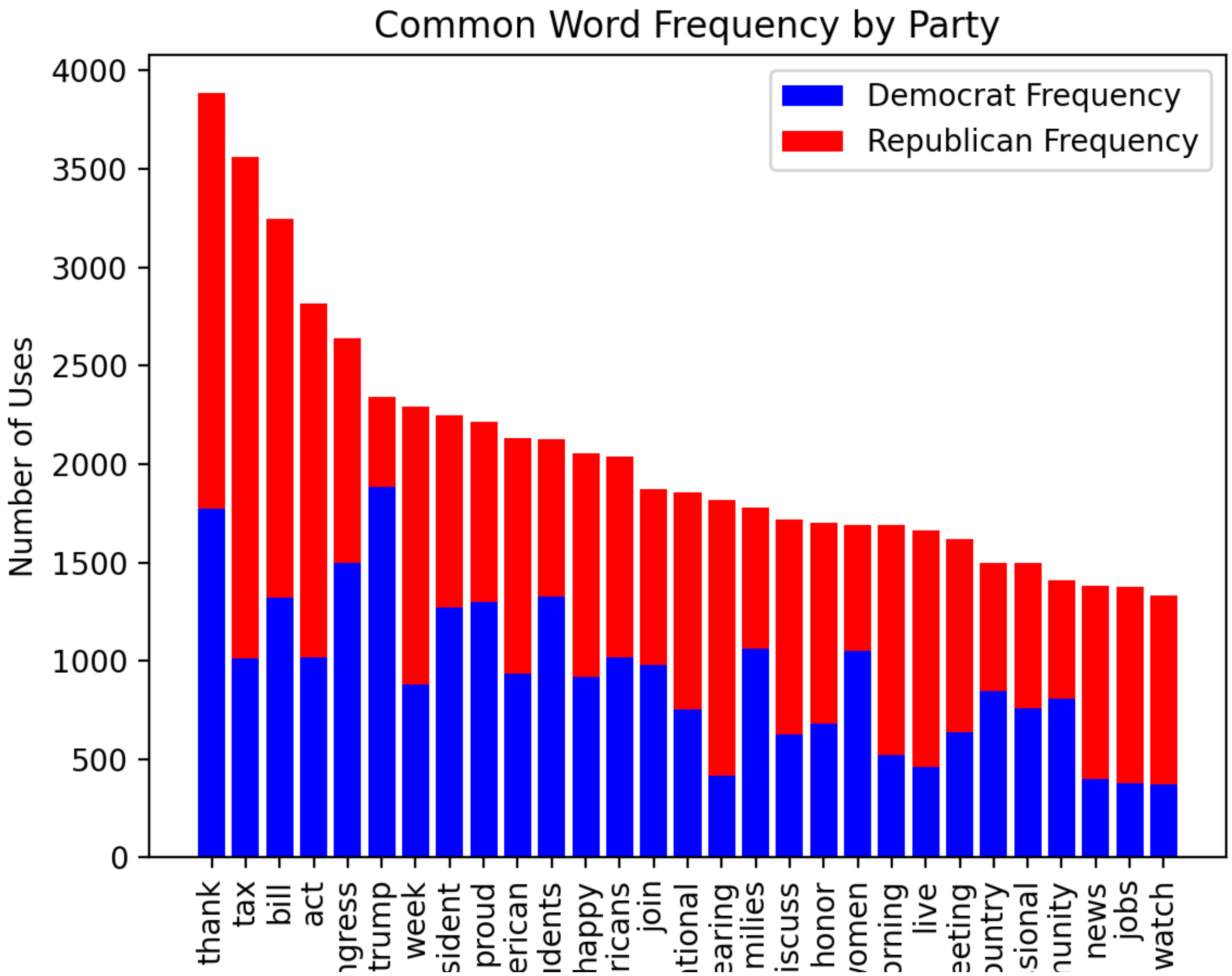
In this project, I used a dataset of tweets by various members of Congress, sorted by party (I got the tweets from Kaggle). My goal was to visualize what words politicians used in their tweets, and how those words were different based on the politician's political party.

First, I used an "emerging" visualization (a word cloud) to represent the words used by politicians of a particular party. I did this using words which were "unique" to the party (see the "Data Decisions" part of this post for what I mean by "unique"). This is a great emerging visualization to show students, because it practices proportional thinking and is not difficult to make in Python using the `wordcloud` library.


Unique to Democrats

Unique to Republicans

I then visualized the words in a more traditional way, a stacked bar chart using Python's `matplotlib` library. This visualization brings focus to the words which are commonly used by politicians in general, while the previous visualization considered words primarily in the context of political party.



An interesting adaptation of this project may be to localize the issues by, for example, examining the tweets of more local politicians.

You can see my full report here and my Python code here.

Python Libraries

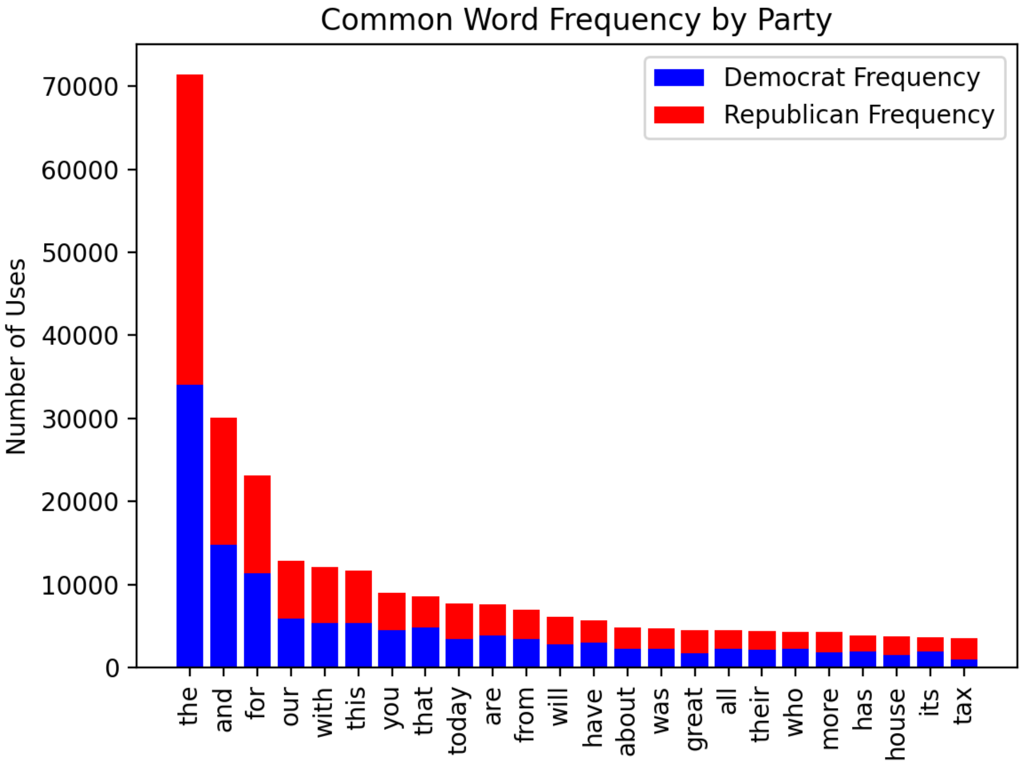
One advantage of using Python to practice data science is that there are many libraries available to use for lots of different kinds of analysis or visualization. Often, you can simplify difficult tasks (such as counting word occurrences or creating word clouds) by using a library. When I teach basic data science in Python, I always spend a lot of time training my students to look for and use these kinds of resources. I explicitly teach students what to search for online (always include the word "Python"!) and what resources are commonly good to use (docs.python, stackoverflow, w3schools, geeksforgeeks, github, and others). We start with using the `random` library at the very beginning of the year, when I answer almost every question with "I don't know, did you look at the documentation?". While students often meet this question with eye rolls, it communicates to students the importance of using resources and the idea that you don't need to know everything to be a good programmer. The best programmers are not those who can create everything on their own or know every command, but rather those who can find resources and make sense of documentation. Professional programmers almost always have a documentation tab and a StackOverflow tab open when they work on their programs.

Data Decisions

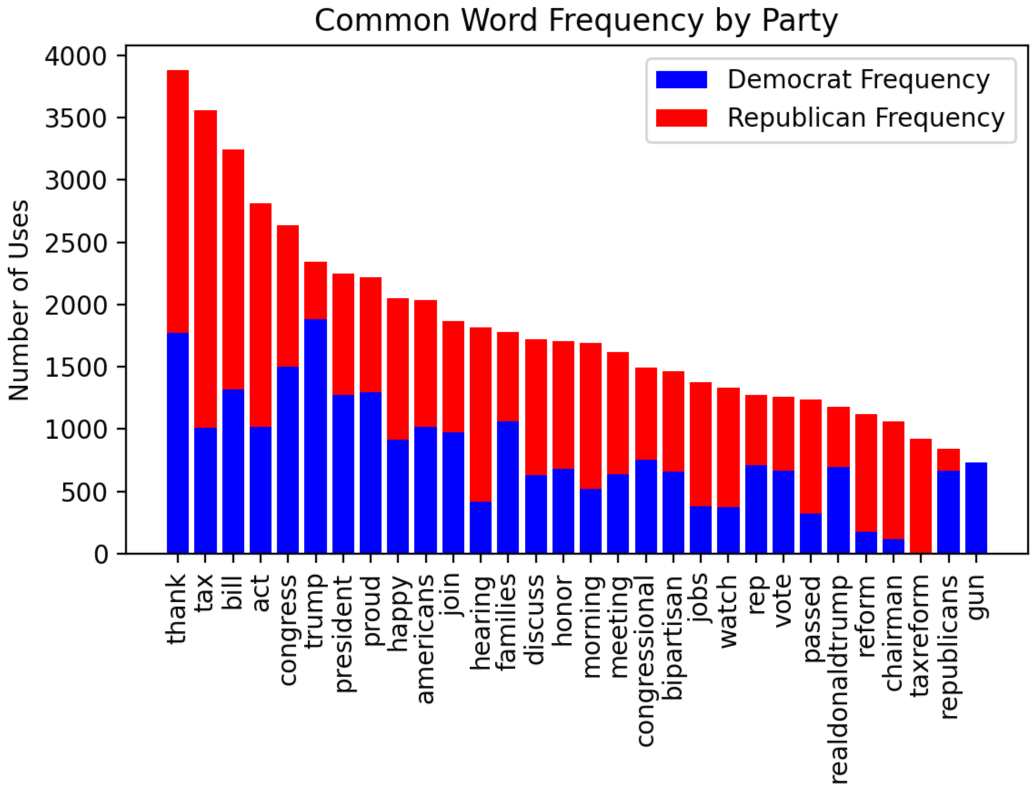
This project is a great example of the importance of basic decisions in data analysis. Once data is selected, a data analysis is largely apolitical. However, choosing and preparing data involves a lot of important decisions that have a big impact on the *results* of that mathematical analysis. Choosing what to measure, how you generate data to measure it, and what kind of analysis you subject that data to can create situations where the data scientist's implicit biases and assumptions about the world affect the results of their analysis in profound ways. In this project, for example, I made decisions about

- What tweets to include/how many
- What words to leave out (for example, `RT`, which stands for "retweet" was not included. This may have been an interesting consideration if included)
- What how many common words to leave out ("the" makes sense to exclude, but "he" or "go" or "here" may or may not be meaningful)
- How to define findings (for example, my decision to explore words unique to one party, and how I went about determining what a "unique" word is)

Some of these decisions can result in surprisingly major changes in the visualizations. For example, this bar graph doesn't exclude any common words:



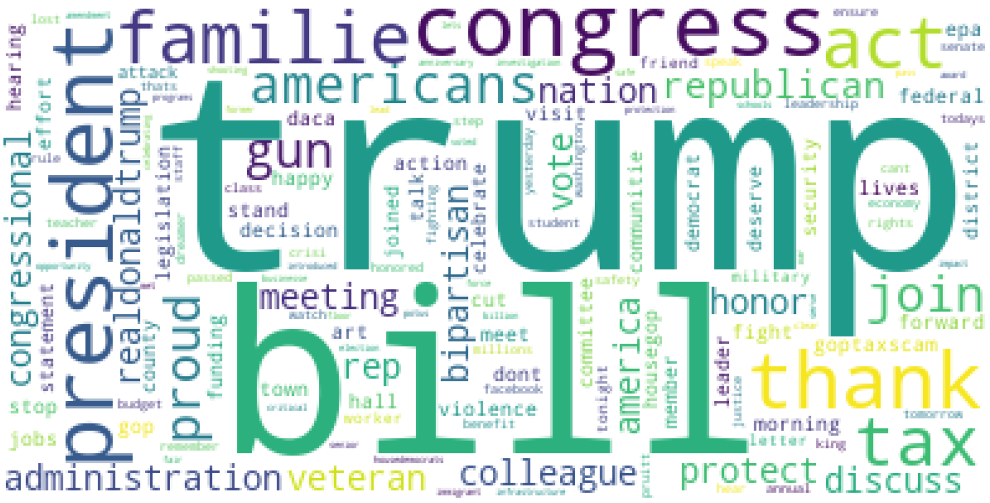
...and this is plot excludes the 500 most common words:



One may come to very different conclusions based on the data decision made here. Note, for example, that "house", which could be politically meaningful, is included in the first plot, but not the second.

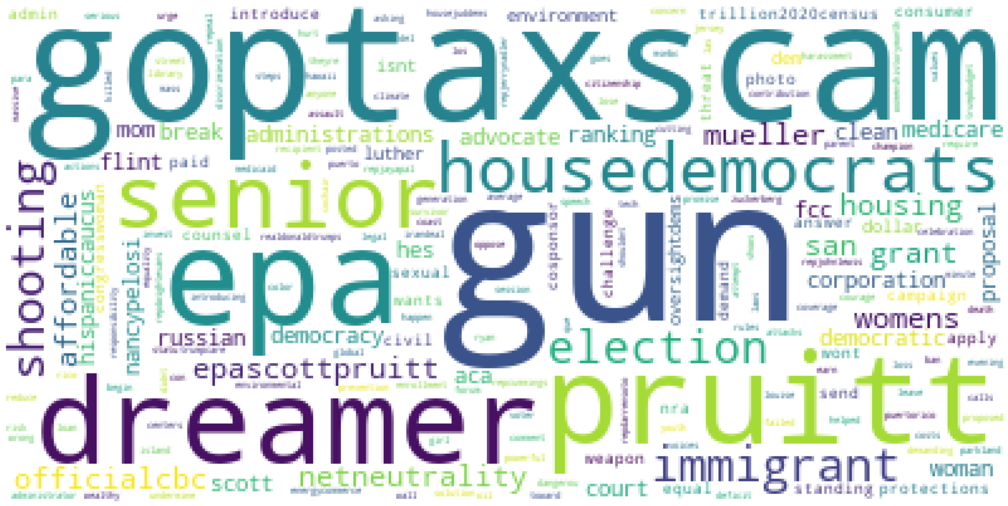
For another example, these are the most common words in tweets by Democrats (excluding the 500 most common words):

Democrats



And these are the most common "unique" to Democrats excluding the 1000 most common Republican words:

Unique to Democrats



Again, you might draw very different conclusions based on a seemingly small data decision. You see the same sorts of changes when looking at the Democrats' words. You can see, for example, that one cloud emphasizes Democrats talking about "Trump", while the other does not.

While there are ways to optimize these decisions, I tend to only have students consider them informally at the K-12 level. It's a great way to introduce the idea of bias in data science, a topic that might not be obvious or intuitive to students before engaging with the idea of "data decisions" in the context of a project.

Controversial Topics

The key to exploratory data analysis is the follow-up (this is especially true with controversial topics). I believe that it is important to (almost) never limit a student's choice in topic. To support this, I encourage students to critically consider their data decisions, and then follow-up their data analysis in a way that makes no claims besides the clear, data-driven patterns.

Political Tweets Example Write-Up

An analysis of the words used in politician's tweets.

Research Question and Background

As of 2022, we know that the political climate in the United States is dangerously polarized. This polarization exists politically and socially. Many people hold very negative opinions of Americans of the other party. We also know that many Americans get their news from social media, particularly Twitter. They get news on Twitter from traditional news outlets, but also from politicians themselves. Given these facts, understanding the trends in political tweets can help us to get an understanding of the political climate in general. In this project, I explored the words used in Tweets by politicians, sorted by party, to see if the polarization exists in these tweets, what exactly this polarization looks like, and what messages American citizens are consuming.

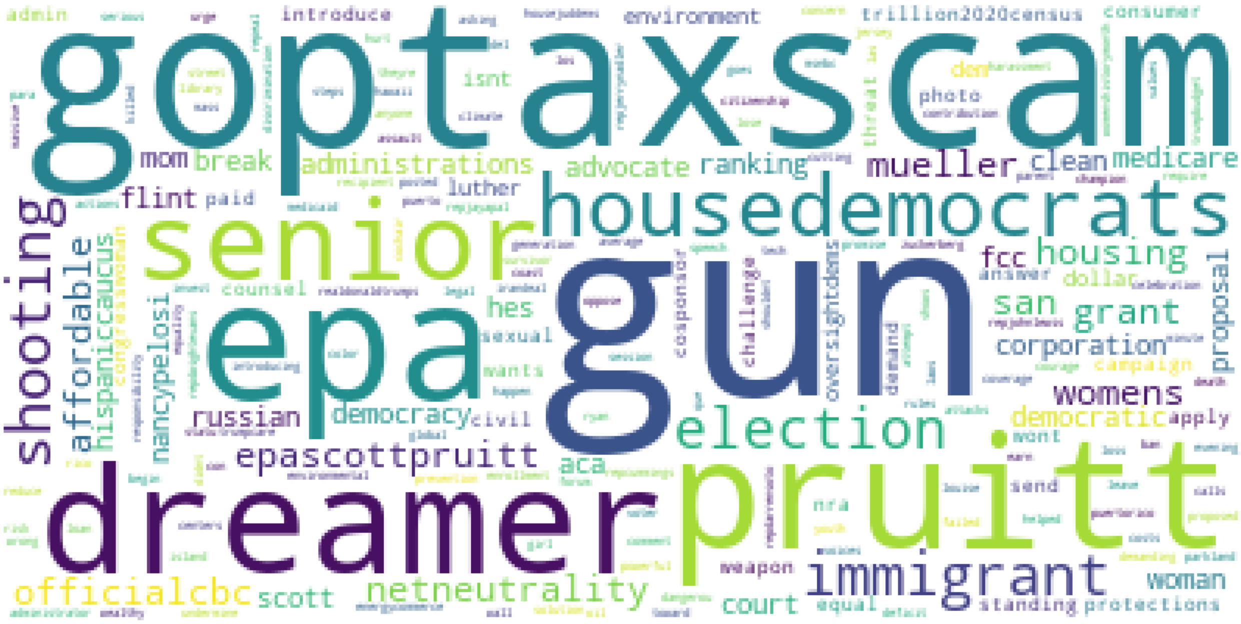
Data and Data Collection

I retrieved my data from Kaggle. The file that I used in this data set has three pieces of information per item (one tweet): the politician's political party, their twitter handle, and the tweet itself. There were a total of 433 unique politicians at the national level, 51% of whom are Republicans, 49% are Democrats. The data set included 86,460 tweets in total.

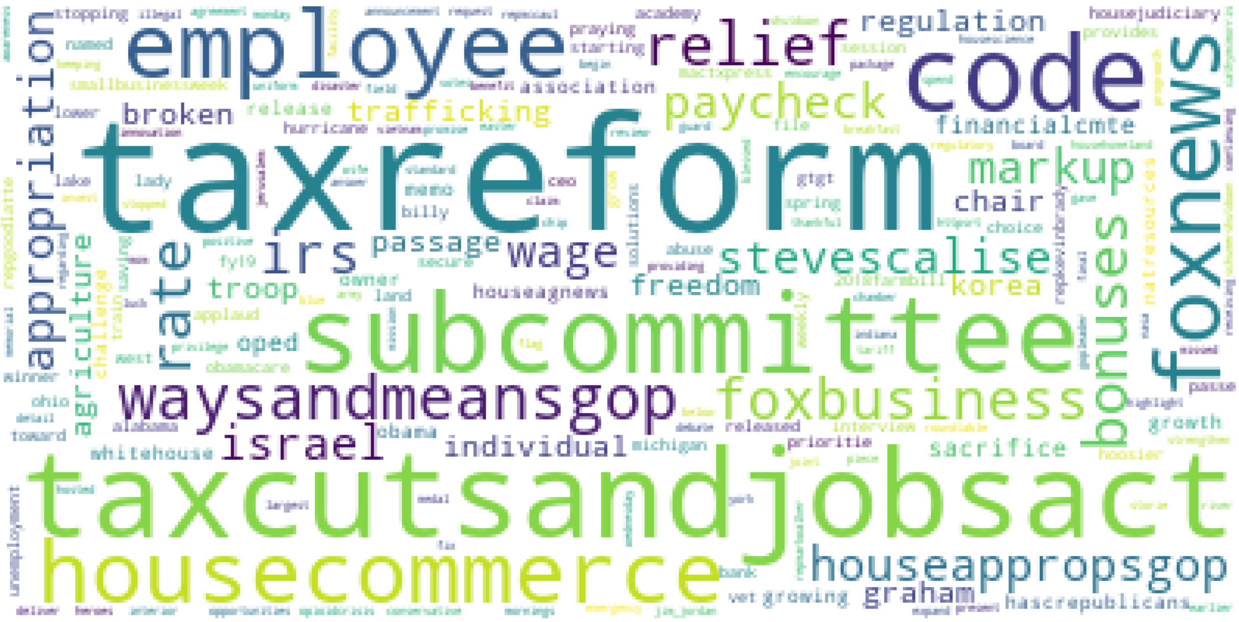
Data Analysis

First, I created word clouds to represent common words used by politicians of either party. In this analysis, I did not include the 500 most common words in the English language (the, and, for, etc), "words" included solely because of the nature of Twitter (https, &amp, RT, etc.), and the 1000 most common words used by the other party. The choice to exclude the 1000 most common words from the other party was two-fold. First, I wanted to exclude words that are often used by politicians in general. For example, politicians all use the word 'thank' often, regardless of party affiliation. Second, I wanted to explore what really made the parties different from each other. Words that were in the top 1000 used for one party but not in the top 1000 for the other seemed to be pretty unique to the former party. Figure 1 shows the word cloud I created with the most common words which are unique to Democrats. Larger words represent more usage. Figure 2 shows the results I got when doing the same analysis on the Republican party.

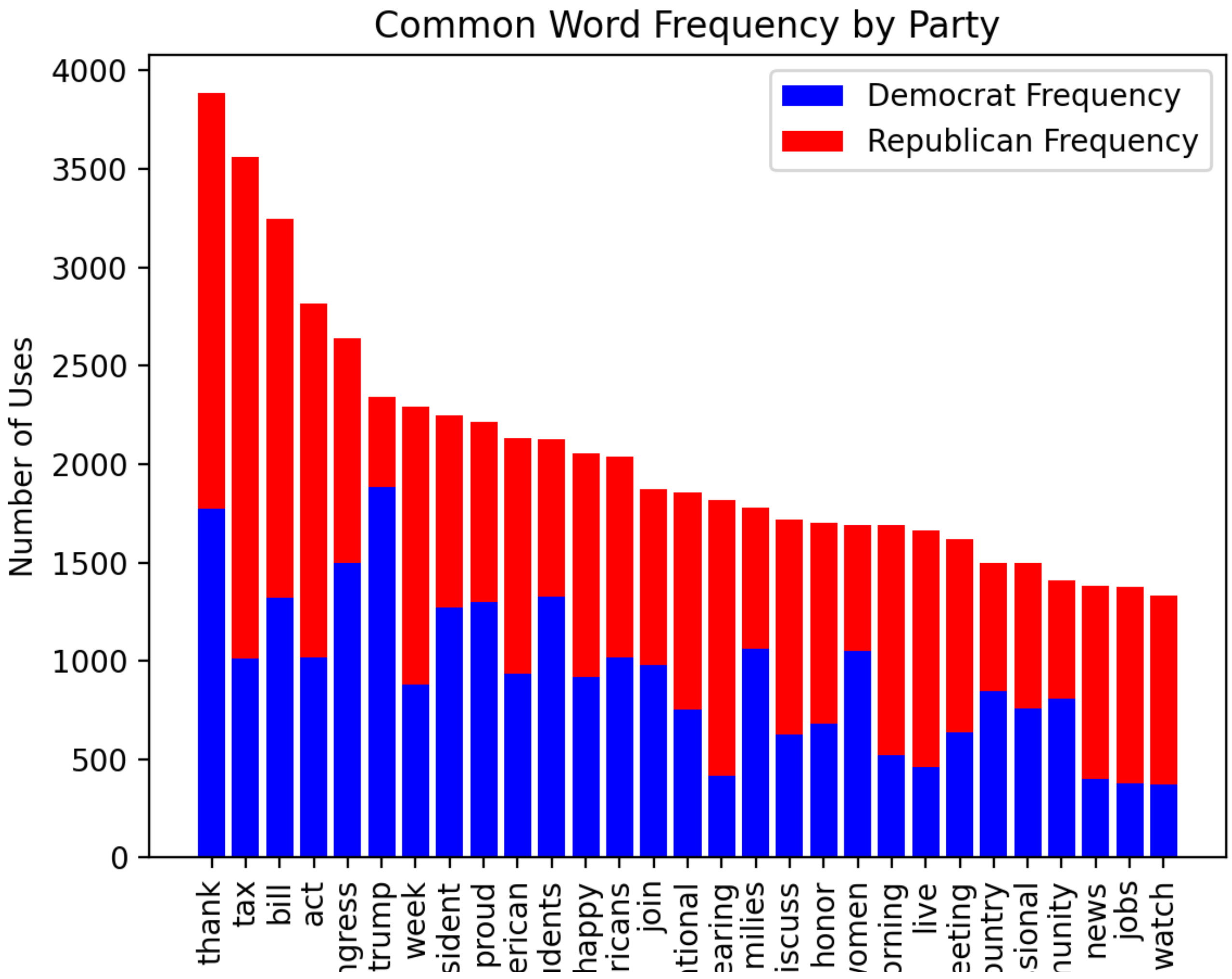
Unique to Democrats



Unique to Republicans



After looking at the unique words, I decided to take a look at generally popular words for politicians, and how those words broke down by party. The result is shown in Figure 3.



Findings

When I take a closer look at the word clouds, I notice a few interesting aspects. The first thing I notice is that Democrats tweet about a wide range of different issues. They talk about financial issues (GOP Tax Scam, affordable), gun control (gun, shooting), the environment (EPA), immigration (dreamer, immigration), the political process (pruitt, election), and other notable issues (women, housing, Russia, net neutrality, and more). On the other hand, a large portion of words used by Republicans are related to economic issues. In the word cloud, I see: tax reform, Tax Cuts and Jobs Act, house commerce, employee, paycheck, bonuses, rate, IRS, wage, Ways and Means GOP, and more. While there are words related to other issues (freedom, Israel, Fox News, and more), these words are often used less and appear to be far outnumbered by financially-related words. We see a number of different words used when we explore words which are commonly used both by Republicans and Democrats. A lot of these words are used in similar amounts by each party. A few notable words which are more used by Republicans, however, include tax, hearing, reform, chairman, jobs, watch and tax reform. This mirrors the findings in the word clouds, that Republicans tend to tweet about economic issues. Democrats use words like Trump, families, Republicans, and gun more. This mirrors findings in the word cloud that Democrats have a wide range of issues, but also introduces another trend. While both parties talk about Donald Trump and the Republican Party, Democrats talk about them substantially more.

In my analysis, I looked to answer two major questions. First, which issues are unique to the Democratic party, and which issues are unique to the Republican party? Then, for issues that are not unique to either party but instead are important to everyone, which party is talking about it more?

To answer the first question, I visualized the most common words used by Republicans that are not used by Democrats, and vice versa. The major pattern I found was that Republicans are talking about the economy and financial issues a lot, while Democrats are talking about a range of issues. This makes anecdotal sense to me, I often notice Republicans focusing political conversation around money and Democrats sharing concern about a variety of things. In fact, I often hear people describe themselves as a "fiscal Republican and social Democrat". This makes sense, since these people are likely hearing arguments, which are presumably well-constructed or widely approved of (given the amount of attention given to the arguments) about financial issues by Republicans, and about social issues by Democrats.

Then, I created a visualization to look at words which are used by all politicians in my data set. I found a few words that are common: 'thank', 'tax', 'bill', 'act', etc. First, it makes sense that these are common words. First and foremost, politicians tend to work to consistently thank their constituents (likely out of some combination of gratitude and dreams of re-election). After that, "bill" and "act" make sense, since they are words used in politics regardless of the issue the bill or act addresses. When we look at more loaded words (tax reform, guns, jobs, etc.), we see a pattern similar to what I found previously: Republicans are thinking about financial issues, Democrats are thinking about a number of different issues. Interestingly, this analysis also uncovered a different pattern. A lot of political tweets talk about former president Donald Trump or about Republicans as a whole. However, these words are far more often used by Democrats. On a brief review, it seems that most of these tweets are critical. For example, "republicans are using an obscure tactic to stifle debate and democracy in congress", "republicans voted for the trumpcare monstrosity that would have sent health costs soaring", and "republicans need to take a good long look at themselves". This suggests that Democrats tend to be more vocally critical of their opponents than Republicans.

Conclusion

The general sentiment of political tweets seems to mirror the political climate at large, although the polarization is not as apparent. When I looked at tweets from all politicians, I saw that most people are saying similar things. Talking about bills, other politicians, and thanking their constituents. There are some differences at this level, for example Democrats use the word "gun" much more than Republicans, but in general the most used words are similar.

The meaningful results of this project came from looking at the unique words. By looking at words that are only used by one party or the other, you are not getting an idea of the polarization, but you are learning what issues each party is concerned with. I found that Democrats are worried about a wide range of issues, from the environment to gun control to immigration. On the other hand, Republicans are particularly concerned with financial issues and the economy. Knowing these differences can, hopefully, help us to understand our fellow citizens in the opposite party. It does not seem that the schism is so much from pure differences in beliefs, but largely from how people prioritize issues that the country as a whole is concerned about.