Name: _____

Data Literacy

Fall 2025 Student Workbook - CODAP Edition

Workbook v0.9-beta

Brought to you by the Bootstrap team:

- Emmanuel Schanzer
- Kathi Fisler
- Shriram Krishnamurthi
- Dorai Sitaram
- Joe Politz
- Ben Lerner
- Nancy Pfenning
- Flannery Denny
- Rachel Tabak

Table of Contents

Pioneers in Computing and Mathematics

The pioneers pictured below are featured in our Computing Needs All Voices lesson. To learn more about them and their contributions, visit https://bit.ly/bootstrap-pioneers.

We are in the process of expanding our collection of pioneers. If there's someone else whose work inspires you, please let us know at https://bit.ly/pioneer-suggestion.

Notice and Wonder

Write down what you Notice and Wonder from the What Most Schools Don't Teach video.
"Notices" should be statements, not questions. What stood out to you? What do you remember? "Wonders" are questions.

What do you Notice?	What do you Wonder?

Windows and Mirrors

1) Think about the stories you've just encountered. Identify something(s) from the film and/or posters that served as a mirror for you, connecting you with your own identity and experience of the world. Write about who or what you connected with and why.

2) Identify something(s) from the film or the posters that served as a window for you, giving you insight into other people's experiences or expanding your thinking in some way.

Reflection: Try Thinking About Ketchup

This reflection is designed to follow reading LA Times Perspective: A solution to tech's lingering diversity problem? Try thinking about ketchup

1) Think of a time when someone else had a strategy or idea that you would never have thought of, but was interesting to you and/or pushed your thinking to a new level.

2) Think of a time when you had an idea that felt "out of the box". Did you share your idea? Why or why not?

3) The author argues that tech companies with diverse teams have an advantage. Why?

4) What suggestions did the article offer for tech companies looking to diversify their teams?

5) What is one thing of interest to you in the author's bio?

6) Based on your experience of exceptions to mainstream assumptions, propose another pair of questions that could be used in place of "Where do you keep your ketchup?" and "What would you reach for instead?"

Categorical and Quantitative Data in a Nutshell

Many important questions ("What's the best restaurant in town?", "Is this law good for citizens?", etc.) are answered with *data*.
Data Scientists try to answer these questions by writing *programs that ask questions about data*.

Data of all types can be organized into **Tables**.

- Every Table has a **header row** and some number of **data rows**.
- **Quantitative data** is numeric and measures *an amount*, such as a person's height, a score on a test, distance, etc. A list of quantitative data can be ordered from smallest to largest.
- **Categorical data** is data that specifies *qualities*, such as sex, eye color, country of origin, etc. Categorical data is not subject to the laws of arithmetic — for example, we cannot take the "average" of a list of colors.

Categorical or Quantitative?

- **Quantitative data** measures an *amount* and can be ordered from smallest to largest.
- **Categorical data** specifies *qualities* and is not subject to the laws of arithmetic — for example, we cannot take the "average" of a list of colors. *Note: Numbers can sometimes be categorical rather than quantitative!*

For each piece of data below, circle whether it is **Categorical** or **Quantitative**.

1)	Hair color	categorical	quantitative
2)	Age	categorical	quantitative
3)	ZIP Code	categorical	quantitative
4)	Date	categorical	quantitative
5)	Height	categorical	quantitative
6)	Sex	categorical	quantitative
7)	Street Name	categorical	quantitative

For each question below, circle whether it will be answered by **Categorical** or **Quantitative** data.

8)	We'd like to find out the average price of cars in a lot.	categorical	quantitative
9)	We'd like to find out the most popular color for cars.	categorical	quantitative
10)	We'd like to find out which puppy is the youngest.	categorical	quantitative
11)	We'd like to find out which cats have been fixed.	categorical	quantitative
12)	We want to know which people have a ZIP code of 02907.	categorical	quantitative

★ We can sort the animals in *ascending order* (smallest-to-largest) by `age` and then sort the table in *alphabetical order* (A-to-Z) by `name`.

Does that mean `name` is a quantitative column? Why or why not? _____

_____

_____

_____

Questions and Column Descriptions

1) Take some time to look through the Animals Dataset. What stands out to you? Which animals are interesting? What patterns do you notice? Put your observations in the **Notice** column below.

2) Do any of these observations make you wonder? If so, write your question next to the observation in the **Wonder** column. If not, think of another question to write down.

Notice	Wonder	Answered by this dataset?
I notice that *Kujo took a long time to be adopted*	*Is it because he was so big?*	*Yes No*
I notice that		*Yes No*
I notice that		*Yes No*
I notice that		*Yes No*
I notice that		*Yes No*
I notice that		*Yes No*
I notice that		*Yes No*

Describe the table, and two of the columns, by filling in the blanks below.

1. This dataset is about _____; it contains _____ data rows.
2. Some of the columns are:

   a. _____ (column name), which contains _____ (*categorical or quantitative*) data. Some example values are:

   _____.

   b. _____ (column name), which contains _____ (*categorical or quantitative*) data. Some example values are:

   _____.

CODAP in a Nutshell

CODAP is a web-based data science tool that runs in your web browser.

Data Types

CODAP utilizes different ***data types***, including Numbers, Strings, and Booleans.

- Numbers are values like `1`, `0.4`, `1/3`, and `-8261.003`.
  - Numbers are *usually* used for quantitative data and other values are *usually* used as categorical data.
- Strings are values like `"Emma"`, `"Rosanna"`, `"Jen and Ed"`, or even `"08/28/1980"`.
  - All strings *must* be surrounded in quotation marks.

All values evaluate to themselves. The program `42` will evaluate to `42`, and the String `"Hello"` will evaluate to `"Hello"`.

Operators

Operators (like `+`, `-`, `<`, etc.) work the same way that they do in math.

- Operators are written between values, for example: `4 + 2`.

Expressions

Expressions work the same way that they do in math. Numeric expressions can be *evaluated*.

- The following are all examples of expressions: `sqrt(16)`, `sqrt(Weight)`, `m+5`, and `9+17`.

CODAP Exploration

This page will help you familiarize yourself with some of CODAP's features. Check off each item once you have completed it. Feel free to experiment and try things out! Make sure you're logged into the Animals Starter File in CODAP before beginning.

Tables in CODAP

1) A table of data is shown here, with the title at the top of the table. What is the title? _____

2) Move the table to a different location on the screen, then minimize the table (hints: Hovering your mouse over the title. What appears?).

3) Re-expand the minimized table, then add a row - also called a *case* - to the table. (Hint: Click any of the Index numbers in the left-most column. Look at the menu that appears.) Can you delete that same row? (Note: CODAP will not let you delete an *empty* case.)

4) Move the `Age` column so that it is between `Fixed` and `Legs`. Click on `Age` and choose "Sort Ascending (A→Z, 0→9)" from the drop-down menu that appears.

5) Now try "Sort Descending". How many animals have names that begin with S? _____

6) Delete a column of the table. (Columns are sometimes called *attributes*.)

7) Use the "Undo" button in the upper right to get your column back. Do the keyboard shortcuts for Redo (`Ctrl-Y` on PC, `Cmd-Opt-Z` on Mac) and Undo (`Ctrl-Z` on PC, `Cmd-Z` on Mac) work in CODAP? _____

8) Close the table. Get it back either by opening the drop-down menu that appears when you click on "Tables" in the upper left.

9) Create a new attribute. Is the column populated (filled in) or empty? _____

10) Name your new attribute. What name did you choose? _____

Graphs in CODAP

11) Click on the "Graph" icon in the upper left-hand corner of the screen. *Note: When you first make a graph, the points are randomly positioned!*

12) How many dots appeared on the graph? (Hint: How many rows - or *cases* - are on the table?) _____

13) Click on a dot. What happens? _____

14) Can you figure out a way to make *different* information appear when you click on a dot? (Hint: You may need to move a column!)

15) Drag an attribute (like `Weight`, `Name`, or `Sex`) to one of the graph's axes - or use the drop-down menu that appears when you click on an axis. Can you make the graph show *two* attributes?

16) Double click on the background of your graph. What happens? _____

17) Click on the "Rescale" icon (it looks like four arrows pointing in four different directions) to zoom back out and display all data again.

18) Once a graph shows two attributes, can you change it back to a graph with one attribute?

19) Click and drag any attribute name from the top of any column in the dataset to the center of the graph. When the graph region turns yellow, release the mouse. What happened? _____

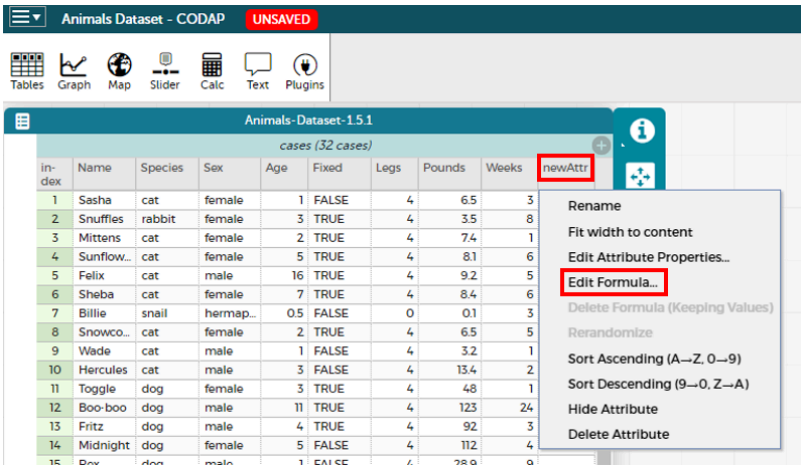
Matching

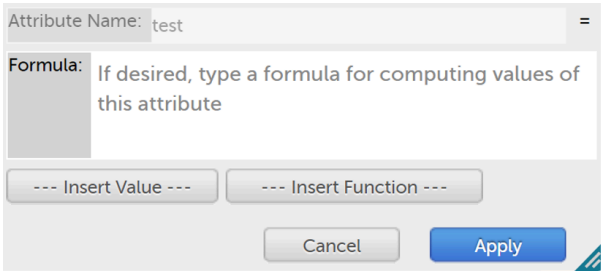
Complete the matching activity below to review what you discovered about graphs and tables in CODAP.

In order to...		I need to...	
delete a table column	**20**	**A**	click on an orange point
move a table	**21**	**B**	mouse over the title bar until a – button appears in the upper right-hand corner
minimize a table	**22**	**C**	select the attribute; from the drop-down menu that appears, select "Delete Attribute"
create a new table column	**23**	**D**	click the "Graph" icon in the upper left-hand corner of the screen
create a graph of randomly configured points	**24**	**E**	mouse over the title bar until the cursor turns into a hand
identify information about a specific point	**25**	**F**	make sure the table is selected, then click the grey plus sign

Strings

- For each of three sections below, refer Animals Starter File.
- In order to follow the directives, you must first create a new column that appears after you select the table.
- Next, click on the attribute name (`newAttr`) and select `Edit Formula`.
- **In order to follow the directives below, you must type text into the "Edit Formula" box.**





Task 1: "Hello, my name is"

The shelter wants to put a name tag in front of each animal's cage so visitors can learn their names. One shelter employee suggests populating all the rows of an entire column with "Hello, my name is" to create enough tags for all of the animals. After printing the tags, shelter employees will write in each animal's name.

1) Click on `newAttr`. Select `Edit Formula`. Type `Hello, my name is` into the formula box that appears, then select `Apply`. What error message appears in all the rows of this column? _____

2) Click `new Attr` again, then select `Edit Formula`. This time, type `"Hello, my name is"` (*with quotation marks!*) into the formula box. What happens? _____

3) Try typing `Hello, my name is` with the opening quote, but without the closing quote, and select `Apply`. What do you think a "syntax error" is? _____

4) A string is any value that is entered within _____.

Task 2: "Hello, my name is Sasha" ... "Hello, my name is Snuffles" ...

The employee who proposed this solution is happy with it... but you wonder: Wouldn't it be cool if CODAP could input each animal's unique name after "Hello, my name is"? Then, you wouldn't need to handwrite in all those animals' names.

5) Access the formula box again. Try typing in `"Hello, my name is Name"`. Did you get the result you want? _____

6) This time, try typing the `"Hello, my name is " + Name`, being sure to leave `+ Name` out of the string. What happens? _____

_____

7) Do you get the same result if you use `"Hello, my name is " + name`? Does CODAP care about capitalization of attribute names?

_____

8) Now you're feeling like you can create all kinds of nametags! Edit the formula box to create tags for all of the animals resembling this one: "Hello, my name is Felix. I am a 16 year old cat who weighs 9.2 pounds."

Numbers

Task 3: Playing with Pounds

As an employee of the shelter, you want each of these animals to be adopted! You wonder if visitors to the shelter might prefer to receive each animal's weight in kilograms, or maybe rounded to the nearest whole number.

1) But first… let's make sure we understand how numbers work in CODAP. Create a new column, then enter the specified information into the formula box. (You can delete what's in the formula box once you've observed the output.)

- Type `42` (no quotes). *Click Apply.*
- Type a fraction. *Click Apply.*
- Type a decimal. *Click Apply.*
- Type an integer. *Click Apply.*
- Enter some expressions that include operators, such as `5 * (8 + 2)`. *Click Apply.*

Does anything surprise you about how numbers behave in CODAP? Does CODAP know the order of operations? _____

_____

2) Create a new column. Name it `Kilograms`. Note that to convert pounds to kilograms, we divide by 2.205. What will you enter in the formula box to populate this column with each animal's weight in kilograms? _____

_____

3) Create another new column. Name it `Rounded Kilograms`. Here, you will use the function `round`, which returns the value of its input, rounded. Enter `round(Kilograms)` in the formula box. What place value did `round` round to? _____

4) Enter `round(Kilograms, 1)`, then change it to `round(Kilograms, 2)`. What does the round function do with that second - optional - argument?"

_____

_____

5) Click on `Kilograms`. From the drop-down menu that appears, select `Edit Attribute Properties`. Try changing the precision. How is changing precision different from rounding? _____

_____

Task 4: You're the official CODAP expert at the shelter!

You've been so successful answering people's CODAP questions that now *everyone* is ocming to you for help! You decide to spend some time playing around with more of the available functions, so you can help anyone who asks.

6) Enter `sqrt(16)` into the `Edit Formula` box. How many arguments does `sqrt` expect? _____

7) What type of argument does the function `sqrt` expect? _____
Number? String?

8) What type of output does `sqrt` produce? _____
Number? String?

9) Put a check-mark next to expression below that will successfully populate a column. If you're not sure, try them out in Animals Starter File.

`sqrt(Weight)`	`sqrt(Legs)`	`sqrt(Name)`

10) Why will some of these expressions work and some generate errors? _____

_____

Dot Plots and Bar Charts in a Nutshell

Displaying Categorical Variables

- With a table open in CODAP, select the "graph" icon to produce a scatter plot of *randomly distributed* data points.
- Drag attributes/columns to the axes (or select from a drop-down menu of attributes/columns by clicking the axes) to organize the data so that it is no longer randomly distributed.
- Once the data is organized, manipulate it further by selecting the graph menu icons:
  - **the ruler icon** provides options for calculating statistics such as mean, median, and standard deviation
  - for datasets with two variables, clicking the ruler icon will provide *additional* statistical computations (such as a least squares line or regression line)
  - the **bar graph icon** allows new configurations of the data. Select this option to group data points into bins or create a bar for each point. If the data is numeric, clicking on the bar graph icon a second time (for instance, after data is grouped into bins) allows the creation of a histogram (by fusing the dots into bars).

Exploring other Data Visualizations

Data Scientists use **data visualizations** to visualize information. You've probably seen some of these charts, graphs and plots yourselves! When it comes to visualizing ***Categorical Data***, we often rely on **dot plots** and ***bar charts***. (Pie charts display categorical data, too, but CODAP doesn't offer them largely because many find them challenging to read.)

When we want to create a data visualization in CODAP, it is important to consider the following: Which attributes on which axes? What type of data? What configuration?

Bar charts show the *count or percentage* of rows in each category.

- Bar charts provide a visual representation of the frequency of values in a categorical column.
- Bar charts have a bar for every category in a column.
- The more rows in a category, the taller the bar.
- Bars in a bar chart can be show in *any order* , without changing the meaning of the chart. However, bars are usually shown in some sensible order (bars for the number of orders for different t-shirt sizes might be presented in order of smallest to largest shirt).

Dot Plots and Bar Charts in CODAP

Open the Animals Starter File. First, create a graph of randomly generated points by selecting the `Graph` icon, and then respond to the following prompts.

Create Data Visualizations

1) Select the y-axis on your graph (where it says "Click here"). On the drop-down menu that appears, select `Fixed`. (*If you prefer, you may also drag the attribute name from the table to the y-axis.*) What do you notice?

2) Now select the x-axis on your graph and select `Fixed`. How does the graph change?

3) Select the `configuration` icon (which looks like a bar graph) to the right of the data visualization. Select `Fuse Dots into Bars`

4) Click the `ruler` icon to test `count` and `percentage`. What happens?

5) Now, make a bar chart showing how many animals there are of each species by changing the variable on the x-axis to `species`. How can reconfigure the bar chart as a dot plot?

Numeric vs. Categorical Data Visualizations

6) Create a graph with `Weeks` on the x-axis. What intervals do you see on the x-axis?

7) Now, click on `Weeks` so that a drop-down menu appears. From this drop-down menu, choose `Treat as Categorical`. How did the numbers on the x-axis change? (*Look closely!*)

8) Why do you think CODAP produced a graph with intervals on the x-axis that are *not* evenly spaced?

9) As you've discovered, CODAP can view `Age` as numeric or categorical. In which mode can we `Fuse dots into bars`? Which kind of data is used for bar graphs?

Introducing Data Visualizations for Subgroups

This page is designed to be used with the Expanded Animals Starter File.

Part A

1) How many tarantulas are male? _____ How many female? _____

Hint: Sort the table by `species`*!*

2) Would you imagine that the distribution of male and female animals will be similar for every species at the shelter? Why or why not?

_____

_____

Part B

Sometimes we want to compare *sub-groups across groups* . In this example, we want to compare the distribution of sexes across each species. Fortunately, CODAP allows us to build a variety of visualizations where we specify both a group and a subgroup.

To create a **stacked bar chart** …	To make a **multi bar chart** …
• create a graph of randomly distributed points • drag the *group* to an axis • drag the *sub-group* to the center of the display • from the Configuration menu, select "Fuse Dots into Bars" • from the Configuration menu, select "Percent" as the scale.	• create a graph of randomly distributed points • drag the *sub-group* to an axis • drag the *group* to the + in the upper left-hand corner of the graph • from the Configuration menu, select "Fuse Dots into Bars" • to the right of the graph, locate and click the "Rescale Display" button (it looks like four arrows pointing in different directions) until you can see all of the data.

3) Make a stacked bar chart showing the distribution of sexes across species in our shelter.

4) Make a multi bar chart showing the distribution of sexes across species in our shelter.

5) What do you notice? _____

_____

_____

6) What do you wonder? _____

_____

7) Which display would be most efficient for answering the question: "What percentage of cats are female?" Why?

_____

8) Which display would be most efficient for answering the question: "Are there more cats or dogs?" Why?

_____

9) Write a question of your own that involves comparing subgroups across groups. _____

_____

Which display would be most efficient for answering your question? _____ Make the display.

10) Write a different question that would be more efficient to answer with the other kind of display. _____

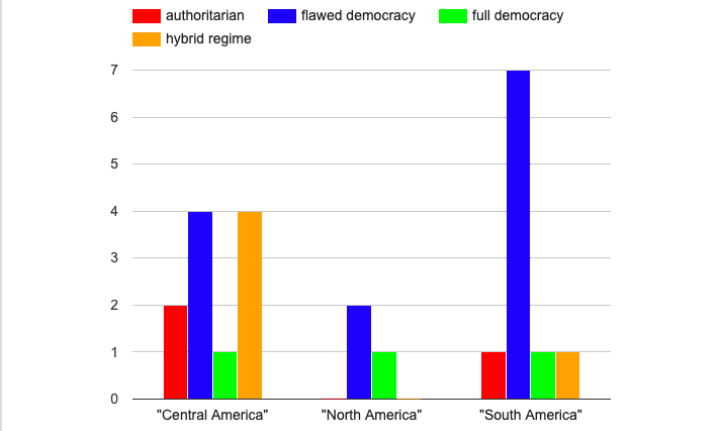
_____

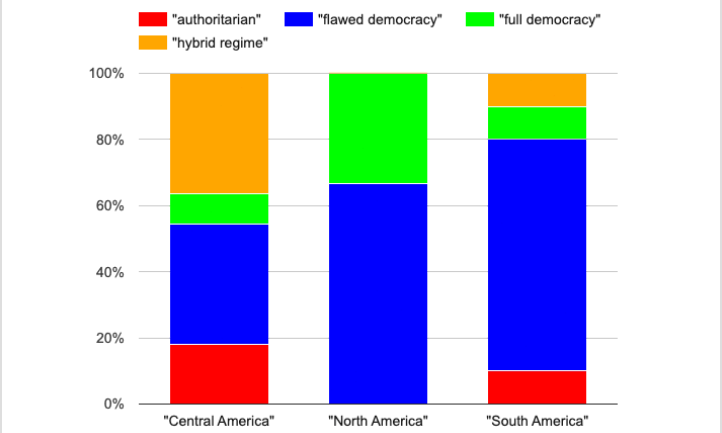
Multi Bar & Stacked Bar Charts - Notice and Wonder

The visualizations on the left are called **multi bar charts**.

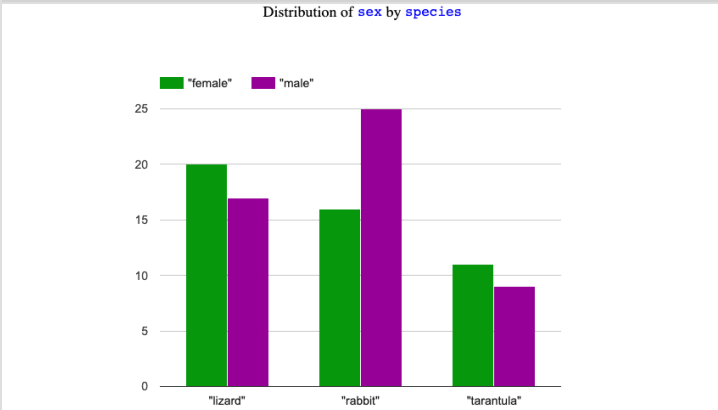
The visualizations on the right are called **stacked bar charts**.

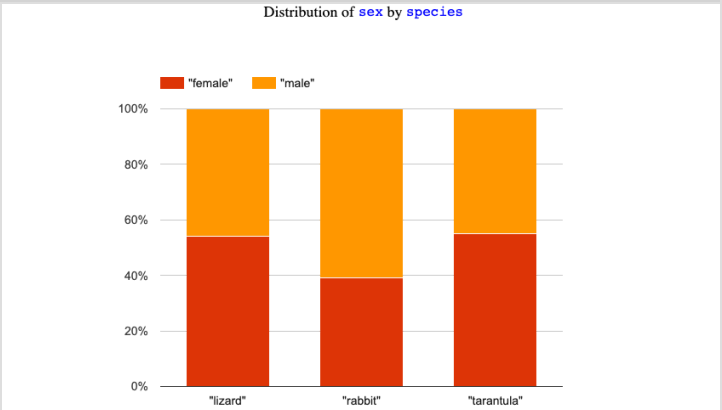
Political Structure of Countries in the Americas





Distribution of sexes across Cats, Dogs and Rabbits





What do you Notice?	What do you Wonder?

1) Is it possible that the same data was used for the multi bar charts as for the stacked bar charts? How do you know?

2) Write a question that it would be easiest to answer by looking at one of the multi bar charts.

3) Write a question that it would be easiest to answer by looking at one of the stacked bar charts.

Practice Plotting

Use the Animals Starter File to create the following visualizations in CODAP. First, fill in the blanks and check all boxes that apply. Next, predict and sketch what the display will look like. Then, create the visualization in CODAP. We've started the first one for you!

1) A histogram of the number of `pounds` that animals weigh.

Column / Attribute	Type of Data	Configuration
pounds [column used as x-axis] n/a [column used as y-axis]	✓ Numeric ❏ Categorical	❏ Points ✓ Fuse dots into bars ❏ Bar for each point ✓ Group into bins ❏ No need to make a selection

Sketch the chart below:	What do you think the data visualization tells us?

2) A dot plot showing the `sex` of animals from the shelter.

Column / Attribute	Type of Data	Configuration
[column used as x-axis] [column used as y-axis]	❏ Numeric ❏ Categorical	❏ Points ❏ Fuse dots into bars ❏ Bar for each point ❏ Group into bins ❏ No need to make a selection

Sketch the chart below:	What do you think the visualization tells us?

Practice Plotting (2)

Use the Animals Starter File to create the following visualizations in CODAP. First, fill in the blanks and check all boxes that apply. Next, predict and sketch what the display will look like. Then, create the visualization in CODAP.

1) A bar chart showing the `species` of animals from the shelter.

Column / Attribute	Type of Data	Configuration
_____ [column used as x-axis] _____ [column used as y-axis]	❏ Numeric ❏ Categorical	❏ Points ❏ Fuse dots into bars ❏ Bar for each point ❏ Group into bins ❏ No need to make a selection

Sketch the chart below:	What do you think the visualization tells us?

2) A scatter-plot, using the animals `name` as the labels, `age` as the x-axis, and `pounds` as the y-axis, for all the animals from the shelter. *Note: The* `Measure` *menu has lots of options! On this page, we've included the two options that* ***create new visualizations***.

Column / Attribute	Type of Data	Measure
_____ [column used as x-axis] _____ [column used as y-axis] _____ [(optional) column used for labels]	❏ Numeric ❏ Categorical	❏ Box plot ❏ Least squares line ❏ No need to make a selection

Sketch the chart below:	What do you think the data visualization tells us?

Practice Plotting (3)

Use the Animals Starter File to create the following visualizations in CODAP. First, fill in the blanks and check all boxes that apply. Then, predict and draw what you think the visualization will look like. Finally, create the visualization in CODAP.

1) A box plot, using `Pounds` as the x-axis, for all the animals from the shelter. *Note: The* `Measure` *menu has lots of options! On this page, we've included the two options that* ***create new visualizations*** .

Column / Attribute	Type of Data	Measure
_____ [column used as x-axis] _____ [column used as y-axis]	❏ Numeric ❏ Categorical	❏ Box plot ❏ Least squares line ❏ No need to make a selection

Sketch the chart below:	What do you think the data visualization tells us?

2) (Challenge) A least squares line (also sometimes called a regression line), using the animals `species` as the labels, `pounds` as the x-axis, and `weeks` as the y-axis, for all the animals from the shelter.

Column / Attribute	Type of Data	Measure
_____ [column used as x-axis] _____ [column used as y-axis] _____ [(optional) column used for labels]	❏ Numeric ❏ Categorical	❏ Box plot ❏ Least squares line ❏ No need to make a selection

Sketch the chart below:	What do you think the data visualization tells us?

Data Visualizations Organizer

Put a check mark to indicate whether each chart listed below displays data from *1 variable or 2 variables*, and whether it displays data that is *categorical or numeric*. In the notes column, add any relevant reminders to yourself about when to use each kind of data visualization. You will want to revisit and add additional notes to this page as you learn more.

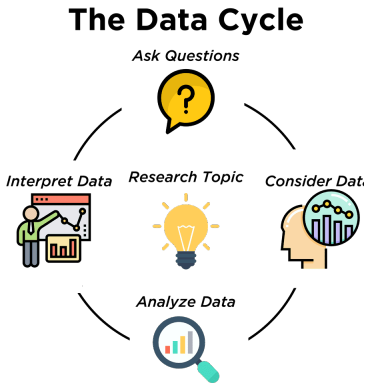
Display	How many variables? What type?	Notes ***(How do I create the visualization? What does it tell me?)***
dot plot	How many variables? ❑ 1 ❑ 2 What type? ❑ Numeric ❑ Categorical	
bar chart	How many variables? ❑ 1 ❑ 2 What type? ❑ Numeric ❑ Categorical	
histogram	How many variables? ❑ 1 ❑ 2 What type? ❑ Numeric ❑ Categorical	
scatter plot	How many variables? ❑ 1 ❑ 2 What type? ❑ Numeric ❑ Categorical	
box plot	How many variables? ❑ 1 ❑ 2 What type? ❑ Numeric ❑ Categorical	
least squares line	How many variables? ❑ 1 ❑ 2 What type? ❑ Numeric ❑ Categorical	

The Data Cycle in a Nutshell

Data Science is all about *asking questions of data*.

- Sometimes the answer is easy to compute.
- Sometimes the answer to a question is *already in the dataset* - no computation needed.
- Sometimes the answer just sparks more questions!

Each question a Data Scientist asks adds a chapter to the story of their research. Even if a question is a "dead-end", it's valuable to share what the question was and what work you did to answer it!



1) We start by **Asking Questions** after reviewing and closely observing the data. These questions can come from initial wonderings, or as a result of previous data cycle. Most questions can be broken down into one of four categories:

- **Lookup questions** - Answered by only reading the table, no further calculations are necessary! Once you find the value, you're done! Examples of lookup questions might be "How many legs does Felix have?" or "What species is Sheba?"
- **Arithmetic questions** - Answered by doing calculations (comparing, averaging, totaling, etc.) with values from one single column. Examples of arithmetic questions might be "How much does the heaviest animal weigh?" or "What is the average age of animals from the shelter?"
- **Statistical questions** - These are questions that both *expect some variability in the data* related to the question and *account for it in the answers*. Statistical questions often involve multiple steps to answer, and the answers aren't black and white. When we compare two statistics we are actually comparing two datasets. If we ask "are dogs heavier than cats?", we know that not every dog is heavier than every cat! We just want to know if it is *generally* true or *generally* false!
- **Questions we can't answer** - We might wonder where the animal shelter is located, or what time of year the data was gathered! But the data in the table won't help us answer that question, so as Data Scientists we might need to do some research beyond the data. And if nothing turns up, we simply recognize that there are limits to what we can analyze.

2) Next, we **Consider Data**, by determining which parts of the dataset we need to answer our question. Sometimes we don't have the data we need, so we conduct a survey, observe and record data, or find another existing dataset. Since our data is contained in a table, it's useful to start by asking two questions:

- What rows do we care about? - Is it all the animals? Just the lizards?
- What columns do we need? - Are we examining the ages of the animals? Their weights?

3) Then, we **Analyze the Data**, by completing calculations, creating data visualizations, creating new tables, or filtering existing tables. The results of this step are calculations, patterns, and relationships.

- Are we making a pie chart? A bar chart? Something else?

4) Finally, we **Interpret the Data**, by answering our original question and summarizing the process we took and the results we found.

Sometimes the data cycle ends once we've interpreted the data… but often our interpretations lead to new questions…
and the cycle begins again!

Which Question Type?

name	type1	hitpoint	attack	defense	speed
Bulbasaur	Grass	45	49	49	45
Ivysaur	Grass	60	62	63	60
Venusaur	Grass	80	82	83	80
Mega Venusaur	Grass	80	100	123	80
Charmander	Fire	39	52	43	65
Charmeleon	Fire	58	64	58	80
Charizard	Fire	78	84	78	100
Mega Charizard X	Fire	78	130	111	100
Mega Charizard Y	Fire	78	104	78	100
Squirtle	Water	44	48	65	43
Wartortle	Water	59	63	80	58

Start by filling out **ONLY the "Question Type"** column of the table below.

Based on the Pokemon data above, decide whether each question is best described as:

- **Lookup** - Answered by only reading the table, no further calculations are necessary!
- **Arithmetic** - Answered by doing calculations (comparing, averaging, totalling, etc.) with values from one single column.
- **Statistical** - Best asked with "in general" attached, because the answer isn't black and white. If we ask "are dogs heavier than cats?", we know that not every dog is heavier than every cat! We just want to know if it is *generally true* or *generally false*!

	Question	Question Type	Which Rows?	Which Column(s)?
1	What type is Charizard?			
2	Which Pokemon is the fastest?			
3	What is Wartortle's attack score?			
4	What is the mean defense score?			
5	What is a typical defense score?			
6	Is Ivysaur faster than Venusaur?			
7	Is speed related to attack score?			
8	What is the most common type?			
9	Does one type tend to be faster than others?			
10	Are hitpoints (hp) similar for all Pokemon in the table?			
11	How many Fire-type Pokemon have a speed of 78?			

Data Cycle: Consider Data

Part 1: For each question below, identify the type of question and fill in the Rows and Columns needed to answer the question.

Ask Questions 	*How old is Boo-boo?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	

Ask Questions 	*Are there more cats than dogs in the shelter?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	

Part 2: Think of 2 questions of your own and follow the same process for them.

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	

Data Cycle: Categorical Distributions (Animals)

Using the Expanded Animals Starter File, let's make a **bar chart** to see what we can learn about the distribution of fixed animals and what new questions it may lead us to.

Ask Questions

Are more animals fixed or unfixed?
What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

All the rows
Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

fixed
What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

The chart shows that there are _____ (more / less / about the same number of) fixed animals _____ (as / than) unfixed animals.

Some new questions this raises include:

Let's make a **stacked-bar-chart** to see if the ratio of fixed to unfixed animals differs by species.

Ask Questions

How does the ratio of fixed to unfixed animals differ by species?
What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

The stacked bar chart shows that _____ (most / some / a few) species have _____ (more / the same number of / fewer) fixed animals _____ (as / than) unfixed animals. I also notice _____

Some new questions this raises include:

Data Cycle: Categorical Distributions 2 (Animals)

Open the Expanded Animals Starter File. Explore the distribution of a categorical column using **pie-chart** or **bar-chart**.

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	❏ The chart shows that there is an even distribution of _____ (variable). ❏ The chart shows that the most common _____ (variable) is/are _____ I notice that _____ I wonder _____ • How does the distribution of _____ (variable) differ by _____ (variable)? • _____ Another question I have is…	

Explore the distribution of two categorical columns using **stacked-bar-chart** or **multi-bar-chart**.

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	When we break the distribution of _____ (variable) down by _____ (variable): • I notice that _____ • I wonder _____ Another question I have is…	

Probability, Inference, and Sample Size in a Nutshell

How can you tell if a coin is fair, or designed to cheat you? Statisticians know that a fair coin should turn up "heads" about as often as "tails", so they begin with the **null hypothesis:** they assume the coin is fair, and start flipping it over and over to record the results.

A coin that comes up "heads" three times in a row could still be fair! The odds are 1-in-8, so it's totally possible that the null hypothesis is still true. But what if it comes up "heads" five times in a row? Ten times in a row?

Eventually, the chances of the coin being fair get smaller and smaller, and a Data Scientist can say "this coin is a cheat! The chances of it being fair are one in a million!"

By sampling the flips of a coin, we can *infer* whether the coin itself is fair or not.

Using information from a sample to draw conclusions about the larger population from which the sample was taken is called **Inference** and it plays a major role in Data Science and Statistics! For example:

- If we survey pet owners about whether they prefer cats or dogs, the **null hypothesis** is that the odds of someone preferring dogs are about the same as them preferring cats. And if the first three people we ask vote for dogs (a 1-in-8 chance), the null hypothesis could still be true! But after five people? Ten?

- If we're looking for gender bias in hiring, we might start with the null hypothesis that no such bias exists. If the first three people hired are all men, that doesn't necessarily mean there's a bias! But if 30 out of 35 hires are male, this is evidence that undermines the null hypothesis and suggests a real problem.

- If we poll voters for the next election, the **null hypothesis** is that the odds of voting for one candidate are the same as voting for the other. But if 80 out of 100 people say they'll vote for the same candidate, we might reject the null hypothesis and infer that the population as a whole is biased towards that candidate!

Sample size matters! The more bias there is, the smaller the sample we need to detect it. Major biases might need only a small sample, but subtle ones might need a huge sample to be found. However, choosing a **good sample** can be tricky!

Random Samples are a subset of a population in which each member of the subset has an equal chance of being chosen. A random sample is intended to be a representative subset of the population. The larger the random sample, the more closely it will represent the population and the better our inferences about the population will tend to be.

Grouped Samples are a subset of a population in which each member of the subset was chosen for a specific reason. For example, we might want to look at the difference in trends between two groups ("Is the age of a dog a bigger factor in adoption time v. the age of a cat?"). This would require making grouped samples of *just the dogs* and *just the cats.*

Finding the Trick Coin

Open the Fair Coins Starter File, which defines `coin1`, `coin2`, *and* `coin3`. *Click "Run".*
You can flip each coin by evaluating `flip(coin1)` *in the Interactions Area (repeat for coins 2 and 3).*
*One of these coins is fair, one will land on "heads" 75% of the time, and one will land on "heads" 90% of the time. *Which one is which?**

1) Complete the table below by recording the results for five flips of each coin and *totalling* the number of "heads" you saw.
Convert the ratio of heads to flips into a *percentage*. Finally, decide whether or not you think each coin is *fair* based on your sample.

Sample	coin1		coin2		coin3	
1	H	T	H	T	H	T
2	H	T	H	T	H	T
3	H	T	H	T	H	T
4	H	T	H	T	H	T
5	H	T	H	T	H	T
#heads	/5		/5		/5	
% heads	%		%		%	
fair?	Y	N	Y	N	Y	N

2) Record 15 more flips of each coin in the table below and *total* the number of "heads" you saw *in all 20 flips of each coin*.
Convert the ratio of total heads to total flips into a *percentage*. Finally, decide whether you think each coin is fair based on this larger sample.

Sample	coin1		coin2		coin3	
6	H	T	H	T	H	T
7	H	T	H	T	H	T
8	H	T	H	T	H	T
9	H	T	H	T	H	T
10	H	T	H	T	H	T
11	H	T	H	T	H	T
12	H	T	H	T	H	T
13	H	T	H	T	H	T
14	H	T	H	T	H	T
15	H	T	H	T	H	T
16	H	T	H	T	H	T
17	H	T	H	T	H	T
18	H	T	H	T	H	T
19	H	T	H	T	H	T
20	H	T	H	T	H	T
#heads	/20		/20		/20	
% heads	%		%		%	
fair?	Y	N	Y	N	Y	N

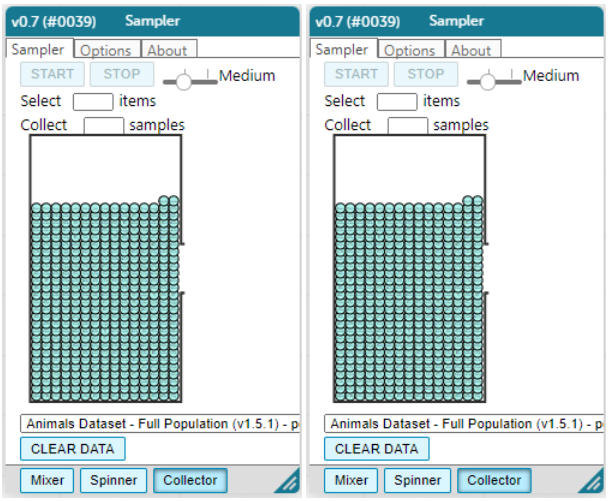
3) Which coin was the easiest to identify? fair? 75%? 90%?

4) Why was that coin the easiest to identify? _____

_____

Sampling and Inference

1) In the screenshots of the "Sampler" (below), show how you would create a small random sample of 10 animals and a large random sample of 40 animals. *To create two separate tables (rather than a single hierarchical table), re-select and re-open "Sampler" from the Plugins menu before each sampling simulation.*



2) In the options tab, did you select "with replacement" or "without replacement"? Why?

3) Make a bar chart for the animals in each sample, showing percentages of fixed and unfixed.

- The percentage of fixed animals in the entire population is: **47.7%**
- The percentage of fixed animals in the small sample is: _____
- The percentage of fixed animals in the large sample is: _____

4) Make a bar chart for the animals in each sample, showing percentages for each species.

- The percentage of tarantulas in the entire population is: **roughly 5%**
- The percentage of tarantulas in the small sample is: _____
- The percentage of tarantulas in the large sample is: _____

5) Direct the sampler to generate a different set of random samples of these sizes. Make a new bar chart for each sample, showing percentages for each species.

- The percentage of tarantulas in the entire population is: **roughly 5%**
- The percentage of tarantulas in the small sample is: _____
- The percentage of tarantulas in the large sample is: _____

6) Which repeated sample gave us a more accurate inference about the whole population? Why?

Choosing Your Dataset in a Nutshell

When selecting a dataset to explore, *pick something that matters to you!* You'll be working with this data for a while, so you don't want to pick something at random just to get it done.

When choosing a dataset, it's a good idea to consider a few factors:

1. Is it **interesting**?
   Pick a dataset you're genuinely interested in, so that you can explore questions that fascinate you!

2. Is it **relevant**?
   Pick a dataset that deals with something personally relevant to you and your community!
   Does this data impact you in any way?
   Are there questions you have about the dataset that mean something to you or someone you know?

3. Is it **familiar**?
   Pick a dataset you know about, so you can use your expertise to deepen your analysis! You wouldn't be able to make samples of the Animals Dataset properly if you didn't know that some animals are much bigger or longer-lived than others.

Consider and Analyze

Fill in the tables below by considering the rows and columns you need. If time allows, type your code into CODAP to see your display!

1) A dot plot showing the `species` of animals from the shelter.

Which Rows?	Which Column(s)?	What will you Create?
All the animals		

2) A `bar-chart` showing the `sex` of animals from the shelter.

Which Rows?	Which Column(s)?	What will you Create?
All the animals		

3) A `histogram` of the number of `pounds` that animals weigh.

Which Rows?	Which Column(s)?	What will you Create?
All the animals		

4) A `box-plot` of the number of `pounds` that animals weigh.

Which Rows?	Which Column(s)?	What will you Create?
All the animals		

5) A `scatter-plot`, using the animals' `species` as the labels, `age` as the x-axis, and `pounds` as the y-axis.

Which Rows?	Which Column(s)?	What will you Create?
All the animals		

6) A `scatter-plot`, using the animals' `name` as the labels, `pounds` as the x-axis, and `weeks` as the y-axis.

Which Rows?	Which Column(s)?	What will you Create?
All the animals		

My Dataset

The _____ dataset contains _____ data rows.

1) I'm interested in this data because _____

_____

2) My friends, family or neighbors would be interested because _____

_____

3) Someone else should care about this data because _____

_____

4) In the table below, write down what you Notice and Wonder about this dataset.

What do you Notice?	What do you Wonder?	Question
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*

5) Consider each Wonder you wrote above and Circle what type of question it is.

Choose two columns to describe below.

6) _____ (column name), which contains _____ (categorical/quantitative) data. Example values from this column include:

_____

7) _____ (column name), which contains _____ (categorical/quantitative) data. Example values from this column include:

_____

Data Cycle: Categorical Data

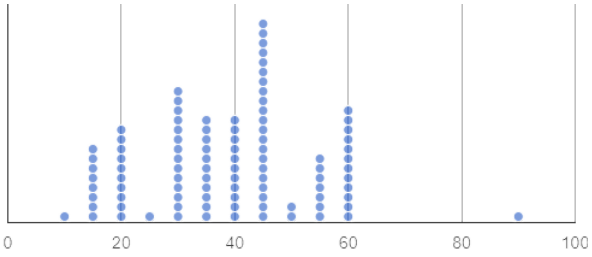
Use the Data Cycle to explore the distribution of one or more categorical columns using **pie-charts and bar-charts** , and record your findings.

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Dot Plots: Distribution, Typicality, Variability in a Nutshell

A **dot plot** (below) is a data visualization consisting of data points plotted along a number line.



On the dot plot (above), each data point represents one student in a sample.
The position of the data point indicates how many minutes it takes for that student to get ready for school. We see, for example, that there is only one students who gets ready in 10 minutes and there are 8 students who take 15 minutes to get ready.

Distribution of Data. To describe the distribution of data—the way that it is spread out on a number line—it is helpful to locate any outliers, clusters, peaks, and gaps.

- A **cluster** is a group of data points that are close together. *Most of the data in the dot plot above is clustered from 10-60, meaning that most students spend between 10 minutes and an hour getting ready for school in the morning.*
- A **gap** is an interval where there are no data points. *On the dot plot above, there is a gap from 60 to 90. In this sample, no one takes between 60 and 90 minutes to get ready.*
- An **outlier** occurs when one data point is much larger or smaller than the other data points. *There is an outlier on the above dot plot at 90. One student requires much more time to get ready in the morning.*
- A **peak** is the value(s) with the most data. *In this sample, 45 minutes is the most common amount of time spent getting ready for school.*

Typicality of Data

- Typicality in a dataset is what we expect from a dataset. We can estimate typicality by looking for peaks and clusters in a dataset.
- In looking at the dot plot above, we might estimate that students typically spend 40 or 45 minutes getting ready for school.

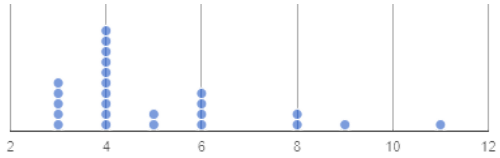
Variability of Data

- **Variability** is how different or alike the data points are. In a quantitative dataset we can measure and describe the variability using range, interquartile range, and standard deviation.
- **Statistical questions** are questions that anticipate variability.
- "In general, how tall are the students in your class?" does anticipate variability.
- "How many inches are in a foot?" does not anticipate variability. The answer is always 12.

Interpreting Dot Plots

Reading a Dot Plot (Group A)

The dot plot below is a name length data visualization created by a group of 25 students (**Group A**).



1) What is the difference (in letters) between the longest and shortest name? _____

2) What is the most common name length? _____

3) What fraction of students have first names that are 5 letters long? _____

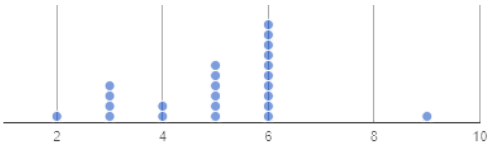
Interpreting Peaks, Clusters, Gaps, and Outliers

4) The distribution of the data is the way that it is spread out on the number line. One way to describe distribution is by identifying peaks, clusters, gaps, and outliers. As a class, label any peaks, clusters, gaps, or outliers on the dot plot **above** .

5) Let's think about what those peaks, clusters, gaps and outliers **tell** us about the dataset. In the dot plot above:

- the peak indicates that _____ letters is the most common name length
- the cluster indicates that many students' names are _____ letters
- the gaps tell us that, in this sample, no students have names that are _____ letters or _____ letters
- the outlier is _____ letters, telling us that longer names are uncommon in this sample.

Reading a Dot Plot (Group B)



6) Label the peaks, clusters, gaps, and outliers of this new dot plot representing the name lengths of a different group of 25 students (Group B).

7) What do the peaks, clusters, gaps, and outliers tell you about the dataset?

_____

_____

Typicality of Name Length Data

8) What do you think is a typical value in Group A? _____(There is more than one correct response.) Explain your reasoning. _____

_____

9) Identify another value someone else might claim is typical of Group A. _____ Why would they choose that value? _____

_____

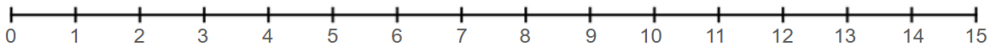
10) Would 6 letters be a good description of the typical number of letters in students' names for Group B? _____ Explain. _____

_____

Our Class' Name Length Data

Create a Dot Plot: Length of First Names in My Class

1) Your class just created a communal dot plot. Copy all of its dots onto the number line below.



Reading a Dot Plot

2) What is the difference (in letters) between the longest name and the shortest name? _____

3) What is/are the most common name length(s)? _____

4) What fraction of students have first names that are 5 letters long? _____

Peaks, Clusters, Gaps, and Outliers in Name Length Data

5) Label any peaks, clusters, gaps, and outliers on the class dot plot (above).

6) Describe what you can conclude about students' name lengths in your class, based on those peaks, clusters, gaps, and outliers. _____

Typicality of Name Length Data

7) What is one possible typical value for class name length? _____ Explain. _____

8) Give another possible typical value: _____. Explain why it is appropriate. _____

Compare

9) Compare and contrast your class dataset with either Group A or Group B from Interpreting Dot Plots. Give at least one way that the distributions are alike, and at least one way that they are different. _____

Two Ways of Thinking about Variability

Variability of Categorical Data

Sana's Groceries	Juliette's Groceries
12 apples and 1 banana	4 peaches, 4 kiwis, 4 oranges, and 1 lime

1) Which dataset has greater variability - Sana's groceries or Juliette's groceries? Explain. _____

_____

_____

2) You ask a group of sixth grade students to respond to two different statements with either "true" or "false."

- Statement A: *I am in sixth grade.*
- Statement B: *I am wearing blue today.*

Which statement do you predict will produce greater variability? Explain. _____

_____

_____

Variability of Quantitative Data

3) Someone looks at your class roster and says, *"In general, students in our class have the same number of letters in their first names."*

Do you agree or disagree? Explain your reasoning. _____

_____

_____

4) Which dataset do you predict will have greater variability for a group of ninth graders who attend the same school - wake-up times on Wednesday or Saturday? Explain. _____

_____

5) Below are the students' responses for their wake-up times on Wednesday versus Saturday. Was your prediction correct? Explain.

- Wednesday: 6:30, 6:15, 6, 6:45, 6:30, 5:45, 6:45, 6:30, 6:30, 6:15
- Saturday: 7:00, 8:00, 8:30, 6:30, 9:45, 10:30, 6:00, 5:45, 10:15, 9:30

_____

_____

Designing Datasets with High and Low Variability

6) Make up two **categorical** datasets with 5 items, each.

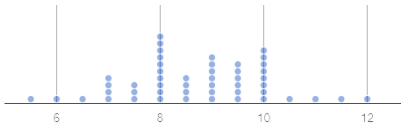
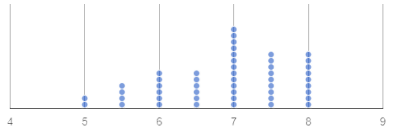
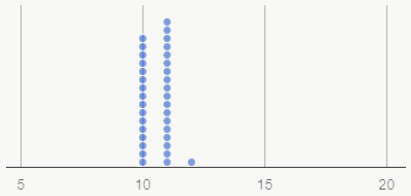
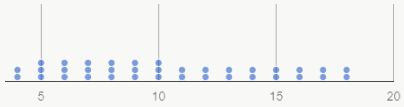
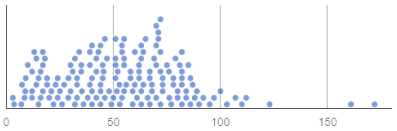
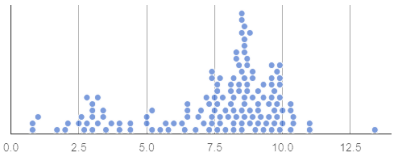
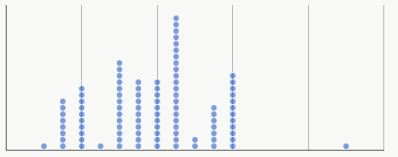
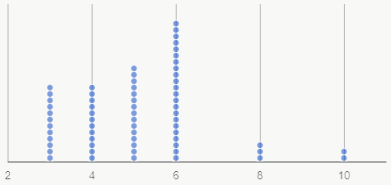
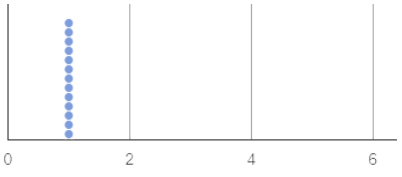
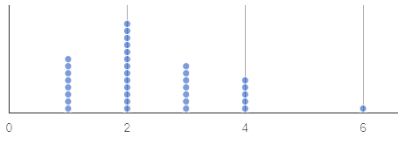
Dataset with Low Variability	Dataset with High Variability

7) Make up two **quantitative** datasets with ten quantities, each.

Dataset with Low Variability	Dataset with High Variability

Variability of Dot Plots

The person who created the dot plots below forgot to label them. For each row, decide which description matches which dot plot. Then explain your choice.

	Which dot plot corresponds, A or B?	Dot Plot A	Dot Plot B	Explain your choice
1)	Students' hours of sleep: • on Monday night: _____ • on Saturday night: _____			
2)	Ages: • of all sixth graders at a K-12 school: _____ • of all students at a K-12 school: _____			
3)	Weights: • of cats in the shelter: _____ • of dogs in a shelter: _____			
4)	Number of minutes: • spent brushing teeth in a day: _____ • spent getting ready for school: _____			
5)	Number of televisions: • per household: _____ • per bedroom: _____			

Variability of Animals' Weights

Make Your Predictions

The staff at the shelter know there is a relationship between how much an animal weighs and how much it eats. They're about to order food for the month, and need some help analyzing the distribution of animals' weights!

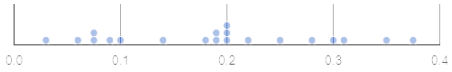
1) Imagine a *typical* animal from each of these four species. Rank the animals from lightest (1) to heaviest (4).

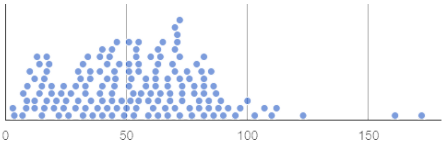
dog: _____ rabbit: _____ cat: _____ tarantula: _____

2) Circle the species you expect to have the *greatest* variability in weight: dog rabbit cat tarantula

3) Circle the species you expect to have the *least* variability in weight: dog rabbit cat tarantula

4) The dot plots below display the weight distributions of dogs, rabbits, and tarantulas. Identify the species of each plot.





species: _____ species: _____ species: _____

5) Explain how you made your decisions. _____

_____

_____

_____

Test Your Predictions Using Pyret

6) Using the Dogs, Rabbits, Cats & Tarantulas Starter File, build a dot plot for each species. In your code, use the tables defined on lines 22-25. Use information from your dot plots to fill in the cells. You can hover your mouse over specific points on the dot plot for additional information on an individual animal. Some cells have been completed for you.

	dogs	cats	rabbits	tarantula
Range/variability	3-172 lbs			
Gaps	123-161 lbs		*No significant gaps*	*No significant gaps*
Outliers	Kujo (172 lbs) Mr. PB (161 lbs)		*No significant outliers*	*No significant outliers*
Peak(s)	72 pounds			

7) Purchasing dog food would be easier if every dog ate roughly the *same amount of food!* But is that true for dogs? What about rabbits, or *any* of the four species in the Dogs, Rabbits, Cats & Tarantulas Starter File? Can you make any recommendations about quantity of food to purchase? _____

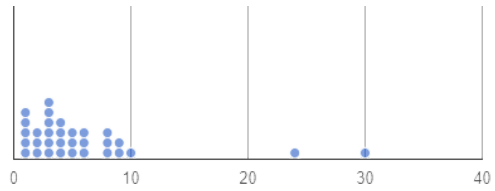
_____

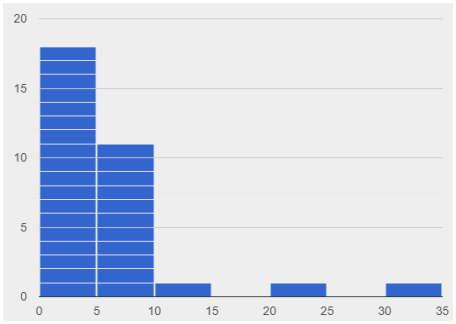
_____

Comparing Dot Plots and Histograms

The displays below both show the distribution of weeks that animals spend at the shelter.

Notice and Wonder





1) What do you Notice about the dot plot (left) and the histogram (right)? What do you Wonder? _____

_____

_____

_____

Dot Plots versus Histograms

Answer the questions below using only the dot plot, and then only the histogram. If you cannot answer a question precisely, write "X".

Question	Dot Plot	Histogram
2) How many animals were in the shelter for fewer than 10 weeks?		
3) How many animals were in the shelter for exactly 30 weeks?		
4) What is the longest amount of time that an animal stayed in the shelter?		
5) How many animals were in the shelter for at least 5 weeks but not more than 25?		
6) Are there any gaps in the data?		
7) Are there any peaks in the data?		

Reflect

8) When you answered the questions using the dot **plot**:

i. Which questions were **easy** to answer? _____

ii. Which questions were **hard** to answer? _____

iii. Which questions were **impossible** to answer? _____

9) When you answered the questions using the **histogram**:

i. Which questions were **easy** to answer? _____

ii. Which questions were **hard** to answer? _____

iii. Which questions were **impossible** to answer? _____

10) When might a histogram be more useful than a dot plot?

_____

_____

_____

11) When might a dot plot be more useful than a histogram?

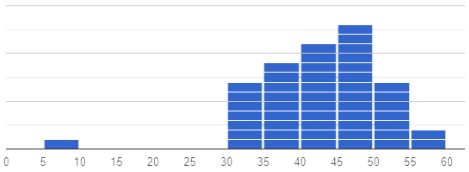
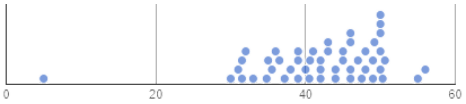
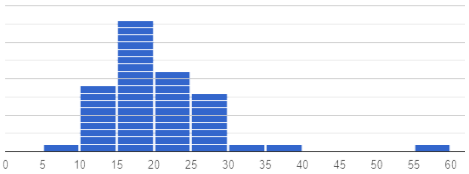
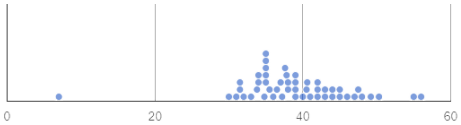
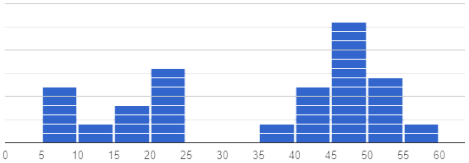
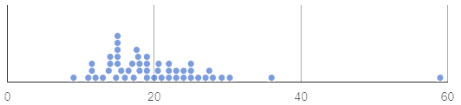
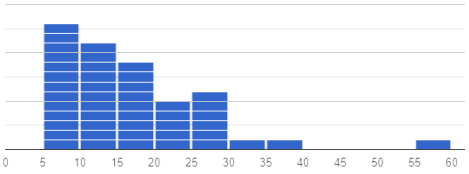
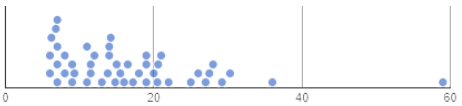
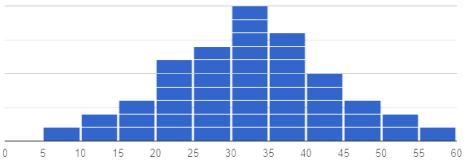
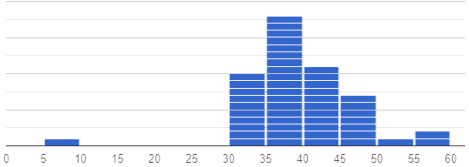
_____

_____

_____

Matching Dot Plots and Histograms

Draw a line from each dot plot on the left to the corresponding histogram on the right.

Dot Plot		Histogram
	1	A 
	2	B 
	3	C 
	4	D 
	5	E 
	6	F 

Making Histograms

By Hand

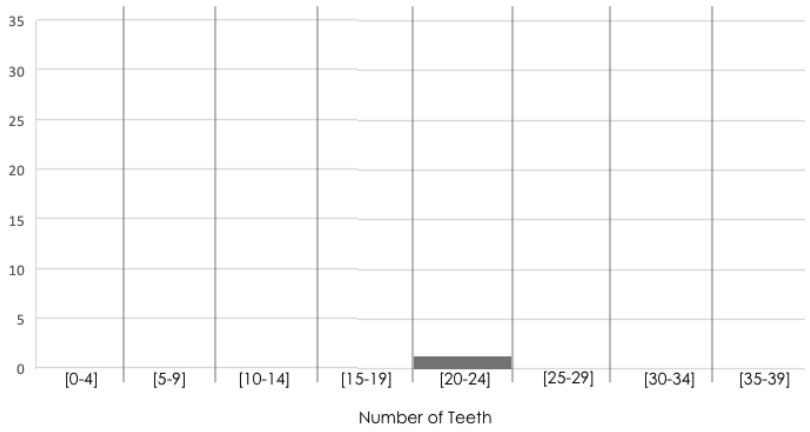
Suppose we have a dataset for a group of 50 adults, showing the number of teeth each person has…

0, 0, 0, 0, 0, 22, 26, 27, 28, 28, 28, 28, 29, 29, 29, 30, 30, 30, 30, 30, 31, 31, 31, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32, 32

1) Use the data to complete the frequency table below. (The last cell has been completed for you.)

number of teeth	0-4	5-9	10-14	15-19	20-24	25-29	30-34
frequency							35

2) Use the frequency table to draw a histogram below, filling in each interval so that its height is equal to the frequency.



In CODAP

Open the Tooth Data Starter File. Make a copy.

3) Look at the table that appears in the starter file. What do you see? _____

4) What bin sized was used for the histogram above? _____

5) Review the tooth table. Does this data appear to be the same or different from the tooth data that appeared in the first section? _____

6) Drag `Num Teeth` to the x-axis and select Group into Bins from the Configuration menu. Fuse dots into bars, then enter the desired bin width of 5.

7) How does the histogram you created in CODAP look *similar* to the one that you drew? Are there any ways in which the histogram you created in CODAP is *different* than the one you created by hand? _____

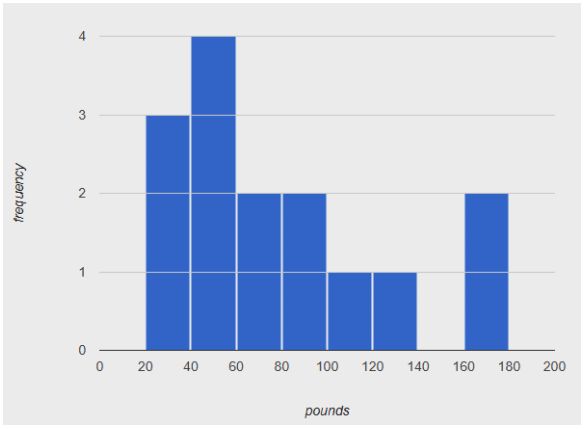
_____

_____

Reading Histograms

Small Local Animal Shelter

Using the histogram below, respond to the questions about the distribution of dogs' weights at a small local animal shelter.



1) How many dogs are represented on the histogram? _____

2) How many dogs weigh less than 100 pounds? _____

3) True or False: The majority of dogs weigh between 40 and 60 pounds.

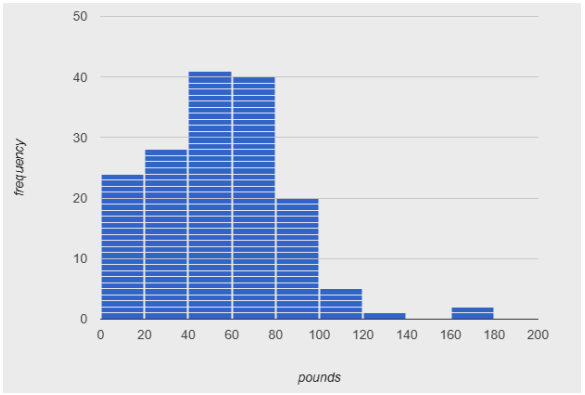
4) True or False: The dogs weigh between 20 and 180 pounds.

5) True or False: The heaviest dog weighs between 40 and 60 pounds.

6) True or False: The histogram shows us that one dog weighs exactly 140 pounds.

Larger Animal Shelter

Using the histogram below, respond to the questions about **dogs** ' weights at a different (much larger) animal shelter.



7) True or False: There are two dogs that weigh at least 160 pounds.

8) True or False: The majority of the dogs weigh between 40 and 60 pounds.

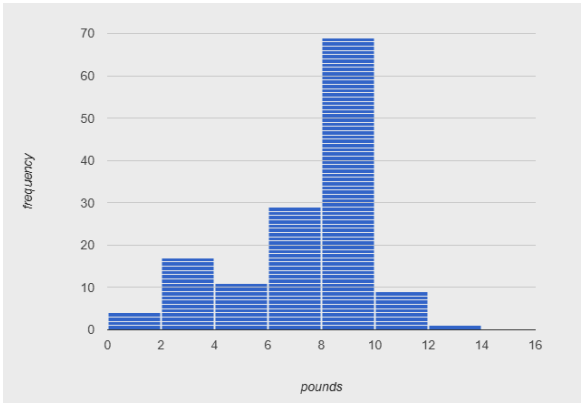
9) True or False: The lightest dog weighs zero pounds.

10) True or False: Most commonly, dogs at this shelter weigh 40-60 pounds.

11) True or False: There are 180 dogs at this animal shelter.

12) True or False: There are more than 150 dogs at this animal shelter.

Using the histogram below, write three statements about the **cats** ' weights and their distribution at the large animal shelter.



13) _____

_____

_____

14) _____

_____

_____

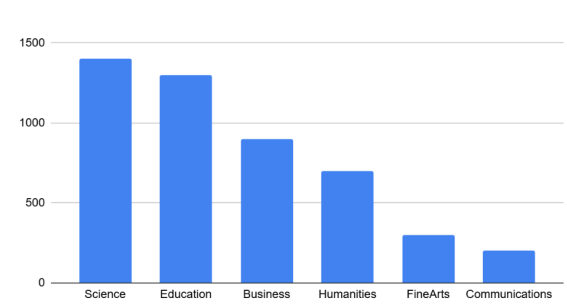
15) _____

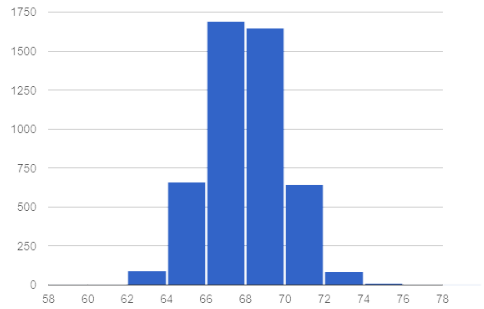
_____

_____

Bar Charts Versus Histograms

A university consists of six colleges. Each student in the university has chosen to enroll in one of these colleges. The **bar chart** below shows the distribution of college choice. The **histogram** below shows the distribution of students by height in inches.





Differences and Similarities

Respond to the prompts to complete the table below.

	Bar Chart	Histogram
Displays frequency: yes or no?		
Type of data: categorical or quantitative?		
Bars touch: yes or no?		
Bars can be reordered: yes or no?		
The shape of the data matters: yes or no?		

1) What are some of the ways that bar charts and histograms are **alike**? *Summarize your conclusions from the table.* _____

_____

_____

2) What are some of the ways that bar charts and histograms are **different**? *Summarize your conclusions from the table.* _____

_____

_____

Distribution of College Choice

Four different students share their conclusions about the **bar graph** displayed above. Only **one** of those conclusions is correct. Respond whether you agree or not, and then explain your stance.

Student A: "The distribution is skewed to the left." _____

_____

Student B: "The distribution is skewed to the right." _____

_____

Student C: "The majority of students are enrolled in the college of science." _____

_____

Student D: "After science and education, there is a large drop in enrollments for the other colleges." _____

_____

Choosing the Right Bin Size

Open your saved Animals Starter File, or make a new copy. After dragging an attribute to an axes, select Group into Bins from the Configuration menu. Fuse dots into bars, then enter the desired bin width.

Make a histogram for the "weeks" column in the animals-table, using a bin size of 10 and the "name" column for your labels.

1) How many animals took between 0 and 10 weeks to be adopted? _____

2) How many animals took between 10 and 20 weeks to be adopted? _____

Try some other bin sizes (be sure to experiment with bigger and smaller bins!)

3) What shape emerges? _____

4) What bin size gives you a picture of the distribution with between 5 and 10 bins. _____

5) Are there any outliers? If so, are they high or low? _____

6) How many animals took between 0 and 5 weeks to be adopted? _____

7) How many animals took between 5 and 10 weeks to be adopted? _____

8) What else do you Notice? What do you Wonder?

_____

_____

_____

_____

_____

_____

_____

_____

9) What was a typical time to adoption?

_____

_____

Using Shape to Interpret Data

Read each scenario. Draw a **rough** histogram sketch (you do not need to label the axes), then decide if the histogram is skew left, skew right, or symmetric. Explain your interpretation.

1) In the United States, there are a few billionaires that have far greater incomes than the average (about $28,000).

Rough histogram sketch:	**Circle one:** *skew left* *skew right* *symmetric* Explain your choice: _____

2) A school cafeteria mostly buys canned goods in huge sizes (48-64 ounces), but also purchases a few ingredients in smaller sizes.

Rough histogram sketch:	**Circle one:** *skew left* *skew right* *symmetric* Explain your choice: _____

3) It's just as likely for a newborn baby to be a certain number of ounces below the average weight (approximately 7.5 pounds) as it is to be that number of ounces above the average weight.

Rough histogram sketch:	**Circle one:** *skew left* *skew right* *symmetric* Explain your choice: _____

4) At many restaurants, the busiest dinner time is around 7pm, but there are always a few people who want to eat earlier or later.

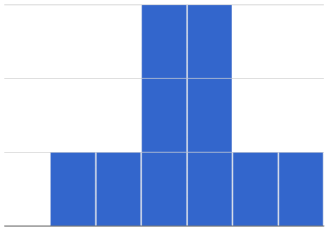
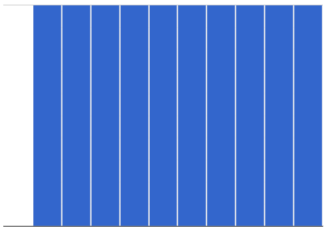
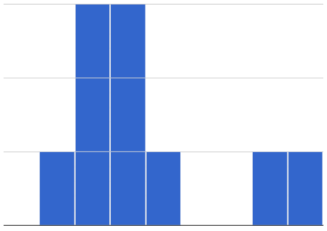
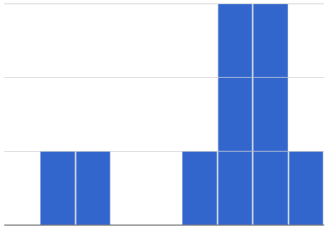
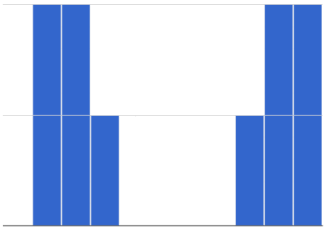
Rough histogram sketch:	**Circle one:** *skew left* *skew right* *symmetric* Explain your choice: _____

Reading Histograms

Students watched 5 videos, and rated them on a scale of 1 to 10. The average score for every video is the same (5.5).

Match the summary description (left) with the *shape* of the histogram of student ratings (right).

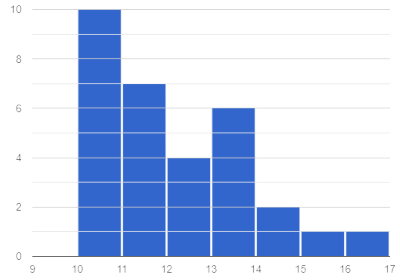
- The x-axis shows the score, and the y-axis shows the number of students who gave it that score.
- These axes are intentionally unlabeled - the **shapes** of the ratings distributions were very different! And that's the focus here.

Most of the students were fine with the video, but a couple of them gave it an unusually low rating.	**1**	**A** 
Most of the students were okay with the video, but a couple students gave it an unusually high rating.	**2**	**B** 
Students tended to give the video an average rating, and they weren't likely to stray far from the average.	**3**	**C** 
Students either really liked or really disliked the video.	**4**	**D** 
Reactions to the video were all over the place: high ratings and low ratings and inbetween ratings were all equally likely.	**5**	**E** 

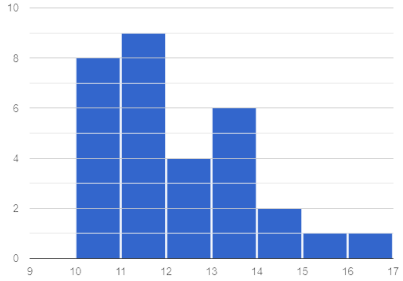
Histograms and Measures of Center

1) The two histograms below show the number of minutes students spent traveling to school: one represents a sample of sixth grade students and the other represents a sample of eighth grade students. All travel times in the dataset are whole numbers.

Distribution of Travel Times (Sixth Grade)



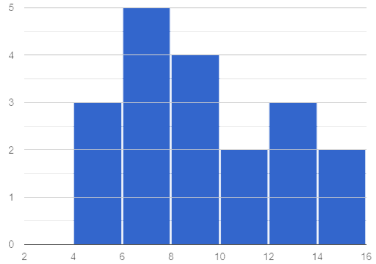
Distribution of Travel Times (Eighth Grade)



2) Which group has the larger mode(s). sixth graders eighth graders the modes are roughly the same

3) The histogram below shows the ages of the 19 children who signed up for rock climbing camp.

Distribution of Rock Climbers' Ages



The median age for these 19 climbers is:

□ about 6 or 7

□ about 8 or 9

□ about 10 or 11

□ about 12 or 13

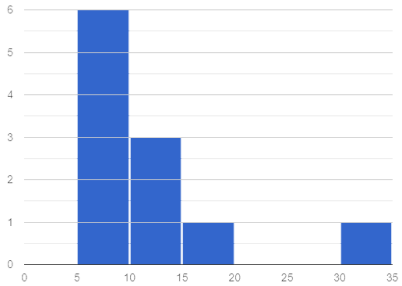
Explain how you determined the median value: _____

_____

_____

4) Eleven students were asked to solve a logic puzzle. The minimum time was 5 minutes, and the maximum time was 35 minutes. The distribution of their times is shown on the histogram below.

Distribution of Time Spent Solving a Logic Puzzle



Which of the following statements is likely a correct comparison of the mean and median number of minutes spent working?

□ The mean time is less than the median time.

□ The mean time is equal to the median time.

□ The mean time is greater than the median time.

□ There isn't enough information provided to determine which is greater.

Explain how you arrived at your choice: _____

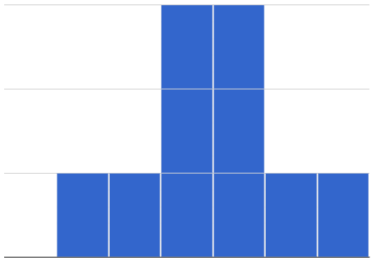
_____

_____

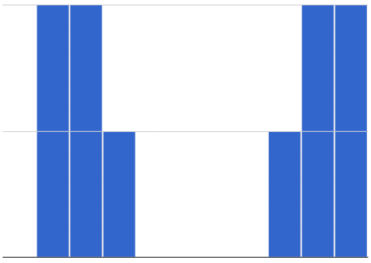
Histograms and Variability

1) Students watched 2 videos, and rated them on a scale of 1 to 10. The average score for every video is the same (5.5).

Movie A



Movie B



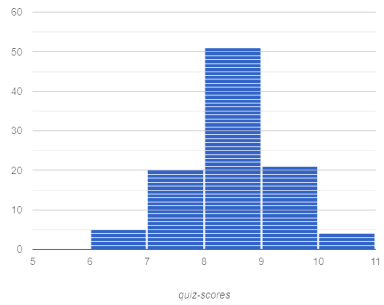
Comparing the two graphs, we know that:

- ❏ The scores for Movie A have greater variability.
- ❏ The scores for Movie B have greater variability.
- ❏ The scores for Movie A and Movie B have equal variability.
- ❏ It is impossible to tell from the given information.

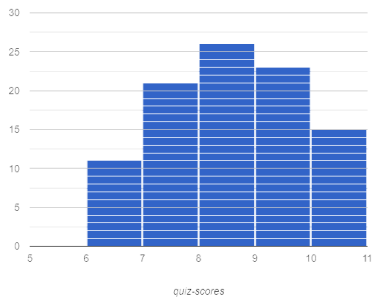
Explain how you arrived at your answer:

2) The following graphs show the distribution of quiz scores for two classes.

Class 1



Class 2



Comparing the two graphs, we know that:

- ❏ The quiz scores of Class 1 have greater variability.
- ❏ The quiz scores of Class 2 have greater variability.
- ❏ The quiz scores of Class 1 and Class 2 have equal variability.
- ❏ It is impossible to tell from the given information.

Explain how you arrived at your answer:

3) Caro says, "Flatter histograms always show less variability." Is she correct? Explain why you agree or disagree with Caro.

Data Cycle: Quantitative Distributions (Animals) - Histograms

Describe two **histograms** made from columns of the animals dataset.
The first question is provided. You'll need to come up with the second question on your own!

Ask Questions

What is the distribution of weight among all animals at the shelter?
What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

The histogram I created is for _____ (x-variable in context) from _____ (dataset or subset).

The shape of this histogram is _____ (skewed left, skewed right, symmetric). There are peaks at _____ and gaps at _____.

I notice that _____
Consider statements like: Most of the histogram's area is… / A small amount of the histograms area trails out… / etc

I wonder _____

Ask Questions

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

The histogram I created is for _____ (x-variable in context) from _____ (dataset or subset).

The shape of this histogram is _____ (skewed left, skewed right, symmetric). There are peaks at _____ and gaps at _____.

I notice that _____
Consider statements like: Most of the histogram's area is… / A small amount of the histograms area trails out… / etc

I wonder _____

Measures of Center in a Nutshell

There are three values used to report the *center* of a dataset .

- Each of these measures of center summarizes a whole column of quantitative data using just one number:

- **Mean** is the average of all the numbers in a dataset .
- **Median** : Half of the dataset will always be greater than or equal to the median. Half of the dataset will always be less than or equal to the median. In an ordered list, the median will either be the middle number or the average of the two middle numbers.
- **Mode(s)** of a dataset is the value (or values) occurring most often. When all of the values occur equally often, a dataset has no mode.

Which Measure of Center is most typical, depends on the shape of the data and the number of values.

- *When a dataset is symmetric* , values are just as likely to occur a certain distance above the mean as below the mean, and the median and mean are usually close together.
- *When a dataset is asymmetric* , the median is a more descriptive measure of center than the mean.

- **Left skew** datasets have a few values that are unusually low, which pull the mean *below* the median.
- **Right skew** datasets have a few values that are unusually high, which pull the mean *above* the median.
  - When a dataset contains a small number of values, the mode(s) may be the most descriptive measure of center. (Note that a small number of *values* is not the same as a small number of *data points* !)

Mean, Median, Mode(s) Practice

Mean

1) Find the mean of each dataset.

17, 23, 25, 23, 22	11, 3, 7, 4, 5	11, 3, 7, 4	5, 7, 11, 11, 7, 7	2, 3, 5, 4, 3, 7, 4

Median

2) Find the median of each dataset.

17, 23, 25, 23, 22	5, 11, 3, 7, 4	11, 3, 7, 4	5, 7, 11, 11, 7, 7	2, 3, 5, 4, 3, 7, 4

Mode(s)

3) Find the mode(s) of each dataset.

17, 23, 25, 23, 22	5, 11, 3, 7, 4	11, 3, 7, 4	5, 7, 11, 11, 7, 7	2, 3, 5, 4, 3, 7,4

Choosing the Best Measure of Center

Find the measures of center to summarize the `pounds` *column of the Animals Starter File, then respond to the prompts.*

1) The three measures of center for this column are:

Mean (Average)	Median	Mode(s)

2) If we scan the dataset, we can quickly see that **most** of the animals weigh less than the mean weight. Why is the average so high? _____

_____

3) Referring to the pounds column of the Animals dataset, fill in the blanks:

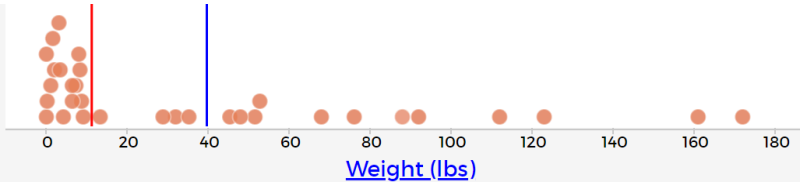
- Outliers on the right pull the mean toward the right, causing the mean to be _____ (greater than / less than) the median.

  When the mean is greater than the median, the shape of the data is _____ (skewed right / skewed left).

- Outliers on the left pull the mean toward the left, causing the mean to be _____ (greater than / less than) the median.

  When the mean is less than the median, the shape of the data is _____ (skewed right / skewed left).

4) In the dot plot below, identify which line is the median and which is the mean. Then label the lines.
Hint: You can refer to the table at the top of the page.



- Which has more data clustered quite close to it, the median or the mean? _____
- Which do you think better represents the data, the median or the mean? Why? _____

_____

5) What did you learn from calculating the mode(s)? _____

_____

6) Are there any measures of center that we can use for categorical data? _____

7) For which quantitative column(s) in the animals table do you think the modes might be a good measure of center? Why? _____

_____

8) To take the average of a column, we add all the numbers in that column and divide by the number of rows. Will that work for every column?

_____

_____

Critiquing Written Findings

Consider the following dataset, representing the heaviest bench press (in lbs) for ten powerlifters:

`135, 95, 230, 135, 203, 55, 1075, 135, 110, 185`

1) In the space below, rewrite this dataset in sorted order.

2) In the table below, compute the measures of center for this dataset.

Mean (Average)	Median	Mode(s)

3) The following statements are correct … but misleading. Write down the reason why.

Statement	Why it's misleading
"More personal records are set at 135 lbs than any other weight!"	
"The average powerlifter can bench press about 236 lbs."	
"With a median of 135, that means that half the people in this group can't even lift 135 lbs."	

Data Cycle: Measures of Center (Animals)

Open the Animals Starter File. Complete both of the Data Cycles shown here, which have questions defined to get you started.

Ask Questions 	*What is the mean age for animals at the shelter?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Ask Questions 	*What is the median time it takes for an animal to be adopted?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Data Cycle: Measures of Center (My Dataset)

Open your chosen dataset. Complete both of the Data Cycles shown here.

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

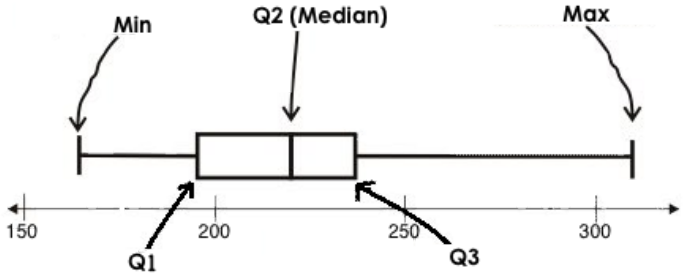
Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Measures of Spread in a Nutshell

Data Scientists measure the *spread* of a dataset using a *five-number summary*:

- ***Minimum***: the smallest value in a dataset - it starts the first quarter
- ***Q1 (lower quartile)***: the number that separates the first quarter of the data from the second quarter of the data
- ***Q2 (Median)*** : the middle value (median) in a dataset
- ***Q3 (upper quartile)***: the value that separates the third quarter of the data from the last
- ***Maximum***: the largest value in a dataset - it ends the fourth quarter of the data

The **five-number summary** can be used to draw a **box plot**.



- Each of the four sections of the box plot contains 25% of the data.
  - If the values are distributed evenly across the range, the four sections of the box plot will be equal in width.
  - Uneven distributions will show up as differently-sized sections of a box plot.
- The left **whisker** extends from the minimum to Q1.
- The **box**, or **interquartile range**, extends from Q1 to Q3. It is divided into 2 parts by the **median**. Each of those parts contains 25% of the data, so the whole box contains the central 50% of the data.
- The right **whisker** extends from Q3 to the maximum.

The box plot above, for example, tells us that:

- The minimum weight is about 165 pounds. The median weight is about 220 pounds. The maximum weight is about 310 pounds.
- The data is not evenly distributed across the range:
  - 1/4 of the players weigh roughly between 165 and 195 pounds
  - 1/4 of the players weigh roughly between 195 and 220 pounds
  - 1/4 of the players weigh roughly between 220 and 235 pounds
  - 1/4 of the players weigh roughly between 235 and 310 pounds
  - 50% of the players weigh roughly between 165 and 220 pounds
  - 50% of the players weigh roughly between 195 and 235 pounds
  - 50% of the players weigh roughly between 220 and 310 pounds
- The densest concentration of players' weights is between 220 and 235 pounds.
- Because the widest section of the box plot is between 235 and 310 pounds, we understand that the weights of the heaviest 25% fall across a wider span than the others.
  - 310 may be an outlier
  - the weights of the players weighing between 235 pounds 310 pounds could be evenly distributed across the range
  - or all of the players weighing over 235 pounds may weigh around 310 pounds.

Distribution of a Dataset

Family Gatherings by the Numbers

Ledet Family Ages : 1, 44, 3, 42, 46, 74, 75, 21, 74, 70, 40, 41, 45 **Average:** 44.3 years old

1) Order the Ages from Least to Greatest: _____

Then compute: _____ Minimum _____ Q1 _____ Median _____ Q3 _____ Maximum _____ Range _____ Interquartile Range (IQR)

Watson Family Ages: 70, 68, 69, 72, 65, 75, 65, 78, 70, 72, 71, 70 **Average:** 70.4 years old

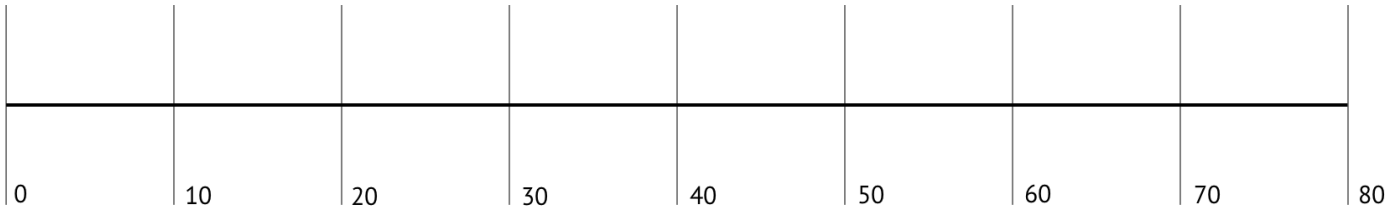
2) Order the Ages from Least to Greatest: _____

Then compute: _____ Minimum _____ Q1 _____ Median _____ Q3 _____ Maximum _____ Range _____ Interquartile Range (IQR)

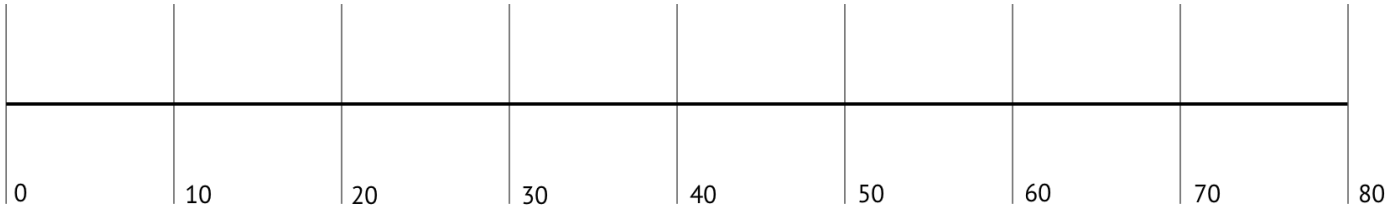
Box Plots - Visualizing Shape

Make box plots for each family's age distribution on the number lines below. *Hint: Plot the 5-Number Summaries, draw a box around the IQR (from Q1 to Q3), let the median split the box into 2 parts, and add whiskers from the box to the minimum and maximum values.*

3) Ledet:



4) Watson:



Compare and Contrast

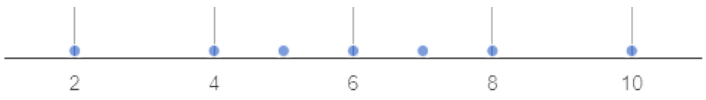
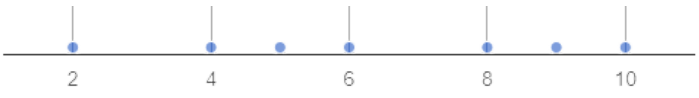
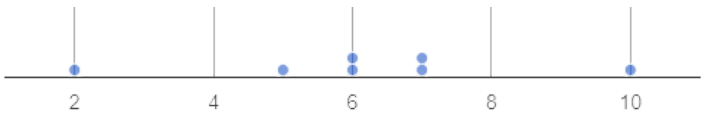
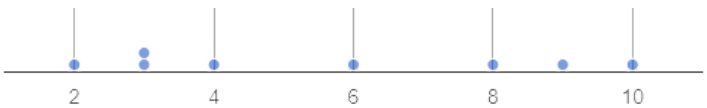
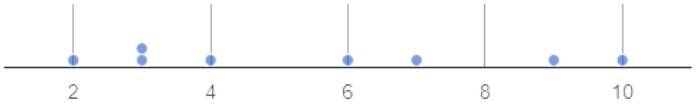
5) For which family gathering was the average age more typical? How do you know? _____

6) What else do you Notice and Wonder about the data from these two family gatherings?

7) We plotted both of these box plots on number lines with the same scale. What are the pros and cons of that choice?

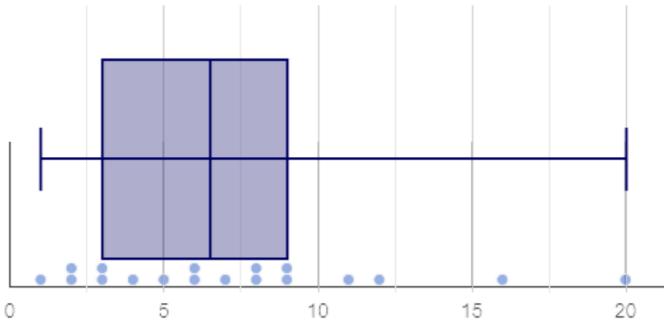
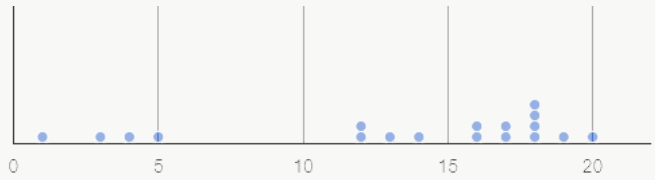
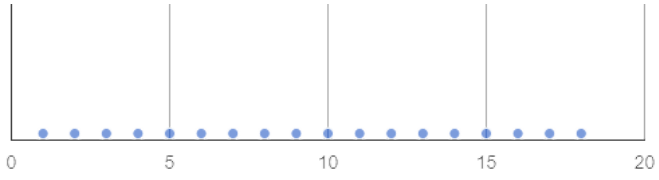
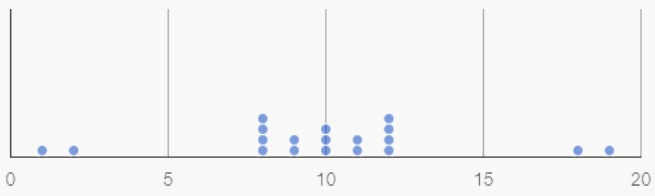
Matching Dot Plots and Five-Number Summaries

Draw a line from each dot plot on the left to the corresponding five-number summary on the right. You might find it useful to label the five-number summaries before you begin matching (see question 1 for an example).

Dot Plot		5-Number Summary
	**1**	**A** Min: 2 Q1: 4 Median: 7 Q3: 9 Max: 10
	**2**	**B** Min: 2 Q1: 4 Median: 6 Q3: 8 Max: 10
	**3**	**C** Min: 2 Q1: 3 Median: 7 Q3: 9 Max: 10
	**4**	**D** Min: 2 Q1: 4 Median: 6 Q3: 9 Max: 10
	**5**	**E** Min: 2 Q1: 5 Median: 6 Q3: 7 Max: 10
	**6**	**F** Min: 2 Q1: 4 Median: 7 Q3: 8 Max: 10
	**7**	**G** Min: 2 Q1: 3 Median: 5 Q3: 8.5 Max: 10
	**8**	**H** Min: 2 Q1: 3 Median: 5 Q3: 8 Max: 10

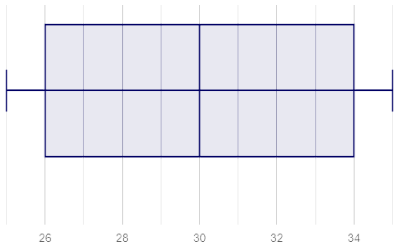
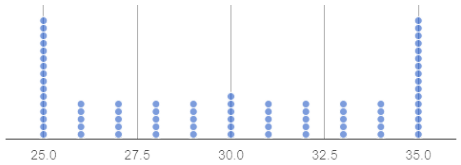
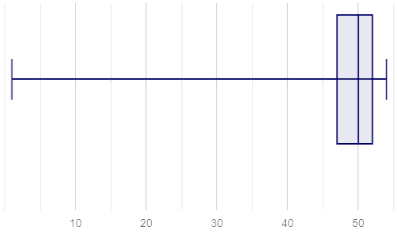
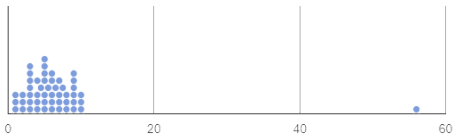
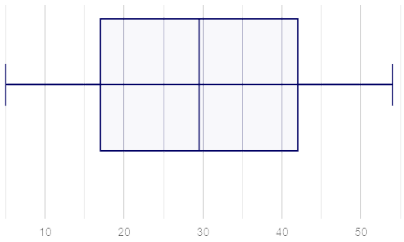
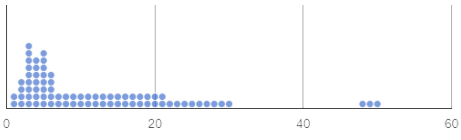
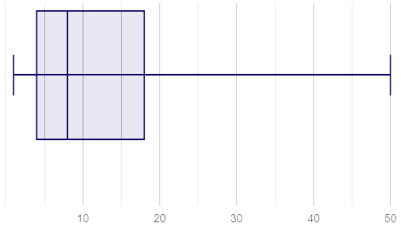
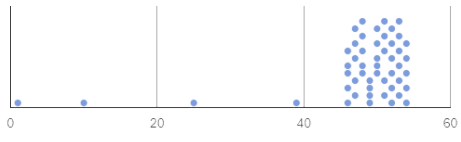
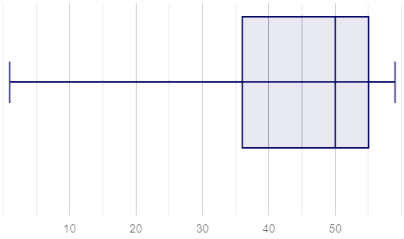
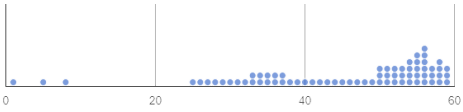
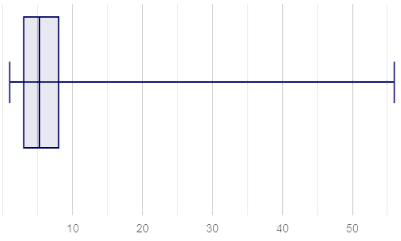
Create Box Plots from Dot Plots

Use the five-number summary to draw a box plot above the corresponding dot plot. When you're finished, identify which quarter(s) of the data are packed the densest, and which quarter(s) of the data are the most dispersed. The first row has been completed as a sample.

	Five-Number Summary	Dot Plot	Densest Packed Quarter(s)	Most Dispersed Quarter(s)
1	Min: 1 Q1: 3 Median: 6.5 Q3: 9 Max: 20		first	fourth
2	Min: 1 Q1: 12 Median: 16 Q3: 18 Max: 20			
3	Min: 1 Q1: 5 Median: 9.5 Q3: 14 Max: 18			
4	Min: 1 Q1: 8 Median: 10 Q3: 12 Max: 19			

Matching Dot Plots and Box Plots

Draw a line from each dot plot on the left to the corresponding box plot on the right.

Dot Plot		Box Plot
	1	A 
	2	B 
	3	C 
	4	D 
	5	E 
	6	F 

Summarizing Columns with Measures of Spread

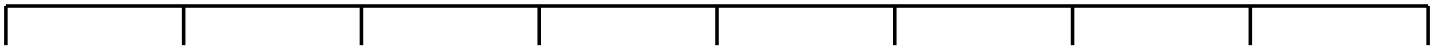
Summarizing the Pounds Column

Get the values to summarize the spread of the _____pounds_____ column of the Animals Starter File by creating a Box Plot and hovering over the minimum, Q1, median, Q3, and maximum.

1) My five-number summary is:

Minimum	Q1	Median	Q3	Maximum

2) Draw a box plot from this summary on the number line below. *Be sure to label the number line with consistent intervals.*



3) The **Range** is: _____ and the **Interquartile Range(IQR)** is: _____.

4) From this summary and box plot, I conclude that:

_____

_____

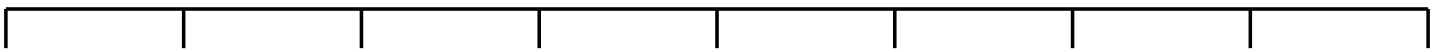
Summarizing the _____ Column

Choose another column to investigate by making a `box-plot`

5) My five-number summary is:

Minimum	Q1	Median	Q3	Maximum

6) Draw a box plot from this summary on the number line below. *Be sure to label the number line with consistent intervals.*



7) The **Range** is: _____ and the **Interquartile Range(IQR)** is: _____.

8) From this summary and box plot, I conclude that:

_____

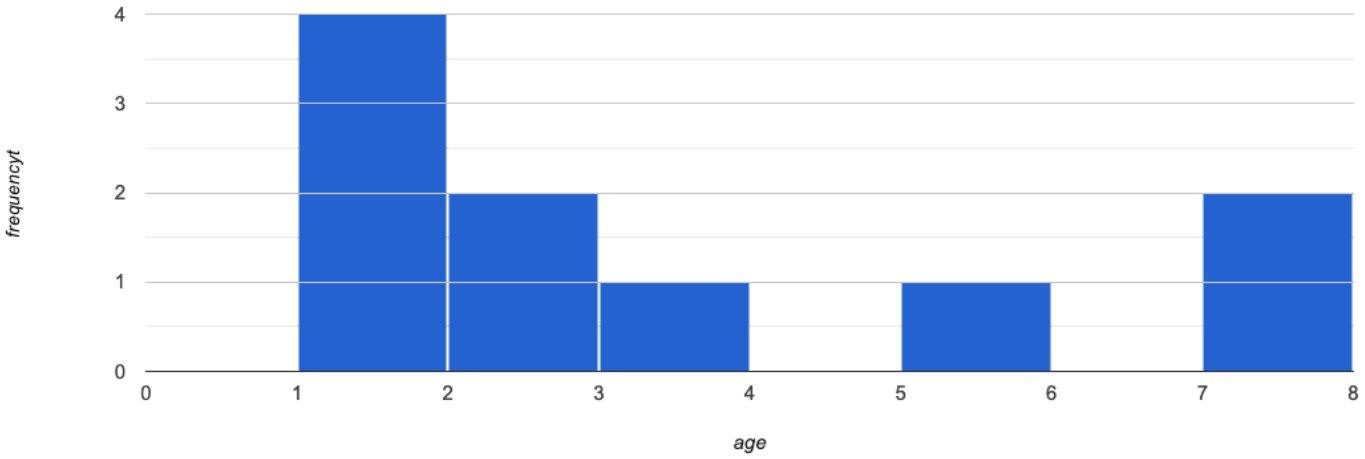
_____

Computing Standard Deviation

Here are the ages of different cats at the shelter: *1, 7, 1, 1, 2, 2, 3, 1, 5, 7*

1) How many cats are represented in this sample? _____

The **distribution** of these ages is shown in the ***histogram*** below:



2) Describe the shape of this histogram. _____

3) What is the mean age of the cats in this dataset? _____

4) How many cats are 1 year old? 2 years old? Fill in the table below. The first column has been done for you.

age	1	2	3	4	5	6	7
frequency	4						

5) **Draw a star to locate the mean on the x-axis of the histogram above** .

6) For each cat in the histogram above, **draw a horizontal arrow** under the axis from your star to the cat's interval, and **label the arrow with its distance from the mean** . (For example, if the mean is 3 and a cat is in the 1yr interval, your arrow would stretch from 1 to 3, and be labeled with the distance "2")

To compute the standard deviation we square each distance and take the average, then take the square root of the average.

7) We've recorded the ages (N=10) shown in the histogram above in the table below, and listed the distance-from-mean for the four 1-year-old cats for you. As you can see, 1 year-olds are 2 years away from the mean, so their squared distance is 4. Complete the table.

age of cat	1	1	1	1	2	2	3	5	7	7
distance from mean	2	2	2	2						
squared distance	4	4	4	4						

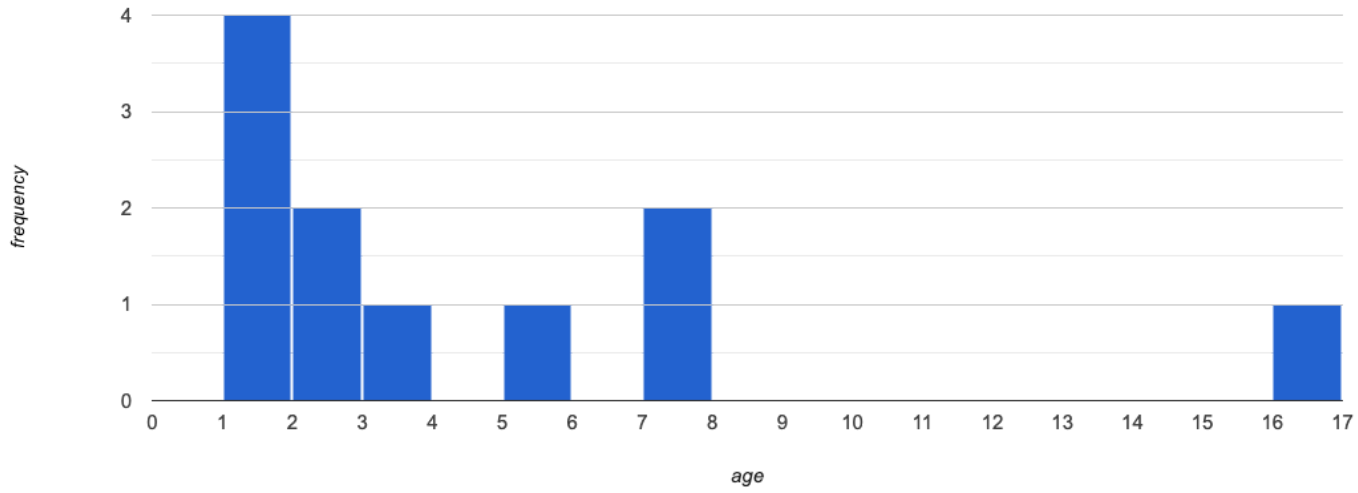
8) Add all the squared distances. What is their sum? _____

9) There are N=10 distances. What is N-1? _____ Divide the sum by *N-1* . What do you get? _____

10) Take the square root to find the **standard deviation** ! _____

The Effect of an Outlier

The histogram below shows the ages of eleven cats at the shelter:



1) Describe the shape of this histogram. _____

2) How many cats are 1 year old? 2 years old? Fill in the table below by reading the histogram. The first column has been done for you.

age	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
frequency	4															

3) What is the mean age of the cats in this histogram? _____

4) **Draw a star to identify the mean on the histogram above** .

5) For each cat in the histogram above, **draw a horizontal arrow** from the mean to the cat's interval, and **label the arrow with its distance from the mean** . (If the mean is 2 and a cat is 5 years old, your arrow would stretch from 2 to 5, and be labeled with the distance "3")
To compute the standard deviation we square each distance and take the average, then take the square root of the average.

6) Recorded the 11 ages shown in the histogram in the first row of the table below. For each age, compute the distance from the mean and the squared distance.

age of cat											
distance from mean											
squared distance											

7) Add all the squared distances. What is their sum? _____

8) Divide the sum by *N-1* . What do you get? _____

9) Take the square root to find the **standard deviation** ! _____

10) How did the outlier impact the standard deviation? _____

_____

Data Cycle: Measure of Spread (Animals)

Open the Animals Starter File. The mean time-to-adoption is 5.75 weeks. Does that mean most animals generally get adopted in 4-6 weeks? Use the Data Cycle to find out. Write your findings on the lines below, in response to the question.

Ask Questions 	*Do the animals all get adopted in around the same length of time?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Turn the Data Cycle above into a Data Story, which answers the question *"If the average adoption time is 5.75 weeks, do all the animals get adopted in roughly 4-6 weeks?"*

Data Cycle: Measure of Spread (My Dataset)

Open your chosen dataset. Use the Data Cycle to find the standard deviation in two distributions, and write down your thinking and findings.

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, write an expression for your Filter Transformer here. If you need to Transform or Build an attribute, write the expression for your Transformer here. What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Relationships Between Quantitative Columns

Scatter Plots

Scatter plots can be used to look for relationships between columns. Each row in the dataset is represented by a point, with one column providing the x-value (***explanatory variable***) and the other providing the y-value (***response variable***). The resulting "point cloud" makes it possible to look for a relationship between those two columns.

- *Form*
  - If the points in a scatter plot appear to follow a straight line, it suggests that a ***linear relationship*** exists between those two columns.
  - Relationships may take other forms (u-shaped for example). If they aren't linear, it won't make sense to look for a correlation.
  - Sometimes there will be no relationship at all between two variables.
- *Direction*
  - The correlation is **positive** if the point cloud slopes up as it goes farther to the right. This means larger y-values tend to go with larger x-values.
  - The correlation is **negative** if the point cloud slopes down as it goes farther to the right.
- *Strength*
  - It is a **strong** correlation if the points are tightly clustered around a line. In this case, knowing the x-value gives us a pretty good idea of the y-value.
  - It is a **weak** correlation if the points are loosely scattered and the y-value doesn't depend much on the x-value.

Line of Best Fit

Linear Relationships can be graphically summarized by drawing a straight line through the data cloud. This summary line is called a "model", as it attempts to provide a simple summary for trends in the dataset. For most datasets, there is no line that will touch every dot, so *all possible models will have some error!* But if the line is close enough to enough of the dots, the model can still help us reason and make predictions about y-values from x-values

$$\text{Data} = \text{Model} + \text{Error}$$

The line that is *closest* to all the other points is known as the ***line of best fit***, meaning it is the *best possible summary* of the relationship and therefore the *best possible model*.

Linear Regression is a way of computing the **line of best fit**. It considers every single data point to generate the optimal linear model, with the smallest possible vertical distance between the line and all the points taken together. *(More specifically, the computer minimizes the sum of the squares of the vertical distances from all of the points to the line. There's a reason we use computers to do this!)*

Points that do not fit the trend line in a scatter plot are called **unusual observations**.

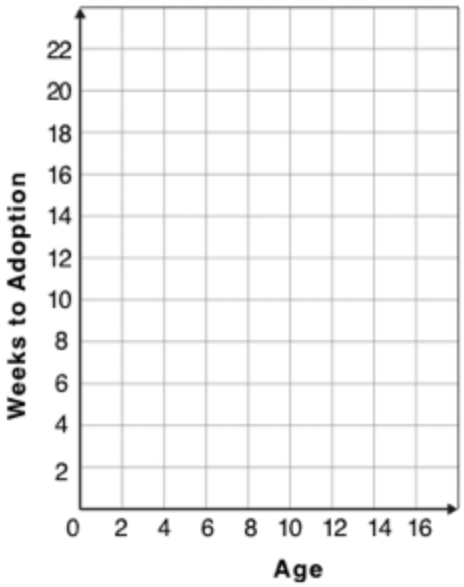
Creating a Scatter Plot

New Animals

1) The table below has some new animals!

- Choose an animal and plot a dot for it on the scatter plot on the right using its age and weeks values. *(Pay careful attention to how the axes are labelled.)*
- Then write the animal's name next to the dot you made.

name	species	age	weeks
"Alice"	"cat"	1	2
"Bob"	"dog"	17	2
"Callie"	"cat"	14	16
"Diver"	"lizard"	1	20
"Eddie"	"dog"	6	9
"Fuzzy"	"cat"	8	5
"Gary"	"rabbit"	4	2
"Hazel"	"dog"	3	3
"Chelsea"	"cat"	12	14
"Josie"	"dog"	9	12
"Cheetah"	"dog"	10	8



2) Plot the rest of the animals - one at a time - labeling each point as you go. After each animal, ask yourself whether or not you see a pattern in the data.

3) After how many animals did you begin to see a pattern? _____

Generalizing the pattern

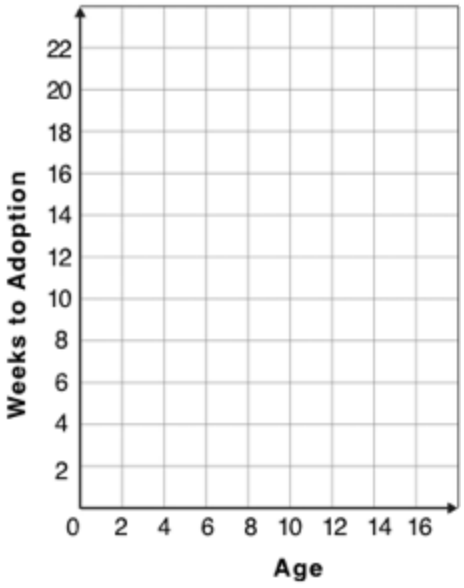
4) Use a straight edge to draw a line on the graph that best represents the pattern you see, then circle the cloud of points around that line.

5) Are the points tightly clustered around the line or loosely scattered? _____

6) Does this display support the claim that younger animals get adopted faster? Why or why not?

_____

7) Now place 10 points on the graph to make a scatter plot that appears to have NO relationship.



Exploring Relationships Between Columns

This page is designed to be used with the Animals Starter File. Log into CODAP to open your saved copy.

As you consider each of the following relationships, first think about what you *expect* , then make the scatter plot to see if it supports your hunch.

1) How are the `pounds` (explanatory variable) an animal weighs related to its `age` (response variable)?

- What would you expect? _____

- What did you learn from your scatter plot? _____

2) How are the number of `weeks` (explanatory variable) it takes for an animal to be adopted related to its number of `legs` (response variable)?

- What would you expect? _____

- What did you learn from your scatter plot? _____

3) How are the number of `legs` (explanatory variable) an animal has related to its `age` (response variable)?

- What would you expect? _____

- What did you learn from your scatter plot? _____

4) Do any of these relationships appear to be linear (straight-line)?

5) Are there any unusual observations?

Data Cycle: Looking for Relationships (Animals)

Open the Animals Starter File. Use the Data Cycle to search for relationships between columns. *The first cycle has a question to get you started. What question will you ask for the second?*

Ask Questions 	*Is there a relationship between weight and adoption time?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?	
Interpret Data 	What did you find out? What can you infer? What - if any - new question(s) does this raise?	

Data Cycle: Looking for Relationships (My Dataset)

Open your chosen dataset. Use the Data Cycle to search for relationships between columns.

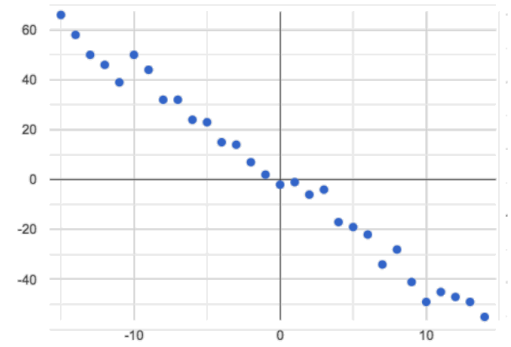
Ask Questions 	What question do you have? _____ _____	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) _____ What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.) _____	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)? _____	
Interpret Data 	❏ There appears to be no relationship between _____ (x-variable) and _____ (y-variable). ❏ There appears to be a _____ (strong / weak / moderate), _____ (positive / negative), _____ (linear / non-linear) relationship between _____ (x-variable) and _____ (y-variable). Some possible outliers might be _____	

Ask Questions 	What question do you have? _____ _____	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) _____ What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.) _____	
Analyze Data 	What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)? _____	
Interpret Data 	❏ There appears to be no relationship between _____ (x-variable) and _____ (y-variable). ❏ There appears to be a _____ (strong / weak / moderate), _____ (positive / negative), _____ (linear / non-linear) relationship between _____ (x-variable) and _____ (y-variable). Some possible outliers might be _____	

Identifying Form, Direction and Strength

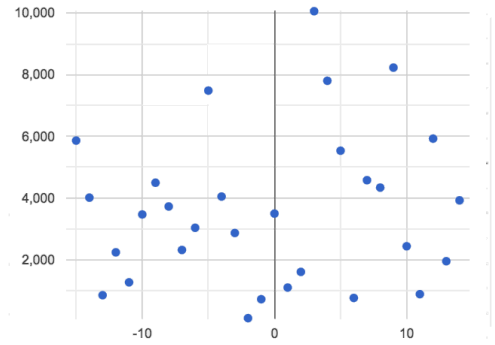
What do your eyes tell you about the Form, Direction, & Strength of these visualizations?
Note: If the form is nonlinear, we shouldn't report direction - a curve may rise and then fall.

A



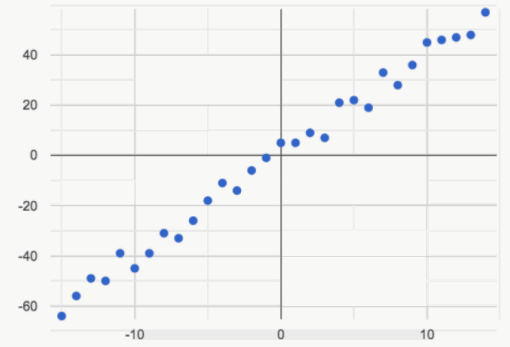
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

B



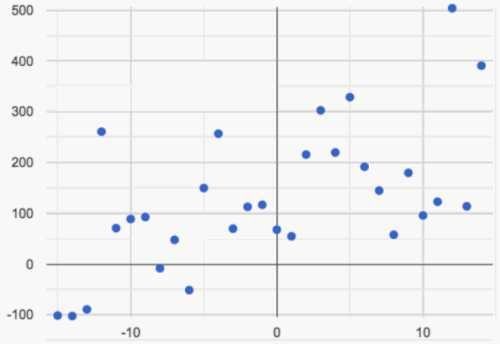
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

C



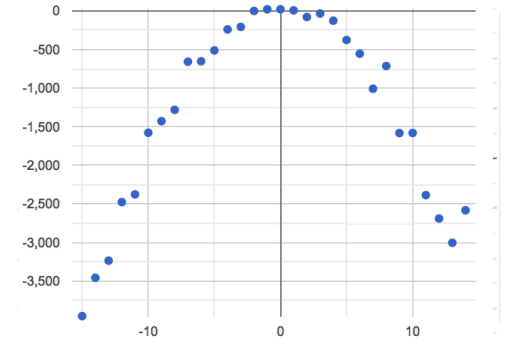
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

D



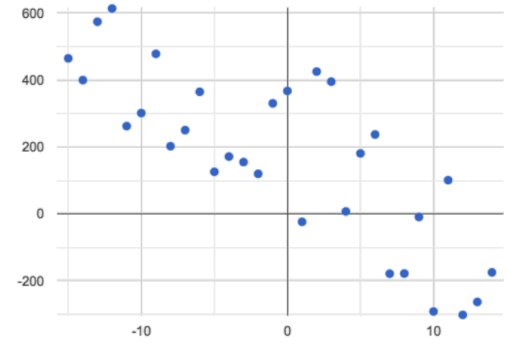
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

E



Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

F



Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

Reflection on Form, Direction and Strength

1) What has to be true about the *shape* of a relationship in order to start talking about the correlation's *direction* being positive or negative?

2) What is the difference between a *weak* relationship and a *negative* relationship?

3) What is the difference between a *strong* relationship and a *positive* relationship?

4) If we find a strong relationship in a sample from a larger population, will that relationship *always hold* for the whole population? Why or why not?

5) If two correlations are both positive, is the stronger one *more positive* (steeper slope) than the other?

6) A news report claims that after surveying *10 million people*, a positive correlation was found between how much chocolate a person eats and how happy they are. Does this mean eating chocolate almost certainly makes you happier? Why or why not?

Summarizing Correlations with r-values

The ***correlation*** between two quantitative columns can be summarized in a single number, the r -value.

- The sign tells us whether the correlation is positive or negative.
- Distance from 0 tells us the strength of the correlation.
- Here is how we might interpret some specific r-values:
  - −1 is the strongest possible negative correlation.
  - +1 is the strongest possible positive correlation.
  - 0 means no correlation.
  - ±0.65 or ±0.70 or more is typically considered a "strong correlation".
  - ±0.35 to ±0.65 is typically considered “moderately correlated”.
  - Anything less than about ±0.25 or ±0.35 may be considered weak.

Note: These cutoffs are not an exact science! In some contexts an r -value of ±0.50 might be considered impressively strong! And sample size matters! We'd be more convinced of a positive relationship in general between cat age and time to adoption if a correlation of +0.57 were based on 50 cats instead of 5.

Correlation is not causation! Correlation only suggests that two variables are related. It does not tell us if one causes the other. For example, hot days are correlated with people running their air conditioners, but air conditioners do not cause hot days!

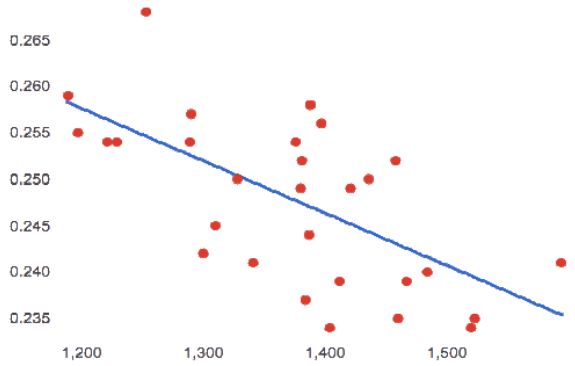
Identifying Form and r-Values

What do your eyes tell you about the Form and Direction of the data? If the form is linear, approximate the r -value.

Reminder:

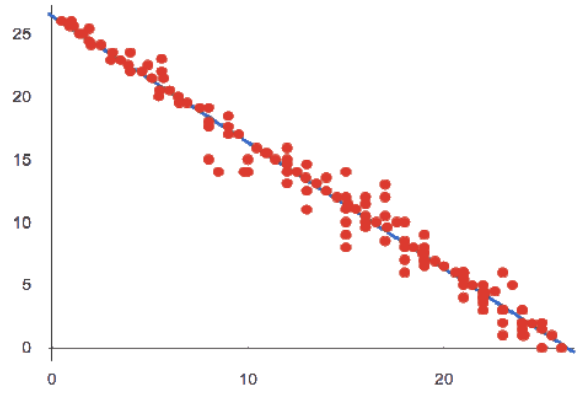
- −1 is the strongest possible *negative* correlation, and +1 is the strongest possible *positive* correlation
- 0 means no correlation
- ±0.65 or ±0.70 or more is typically considered a "strong correlation"
- ±0.35 to ±0.65 is typically considered "moderately correlated"
- Anything less than about ±0.25 or ±0.35 may be considered weak

A



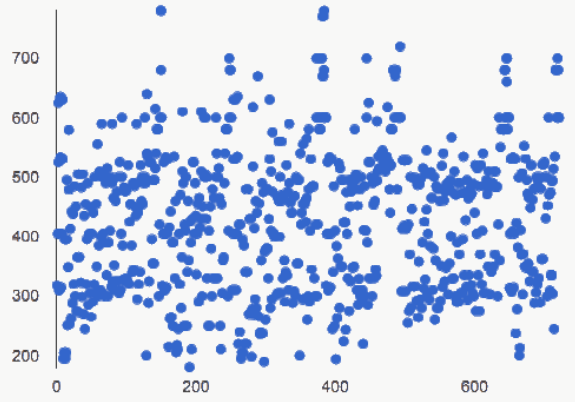
Form :
r close to :

B



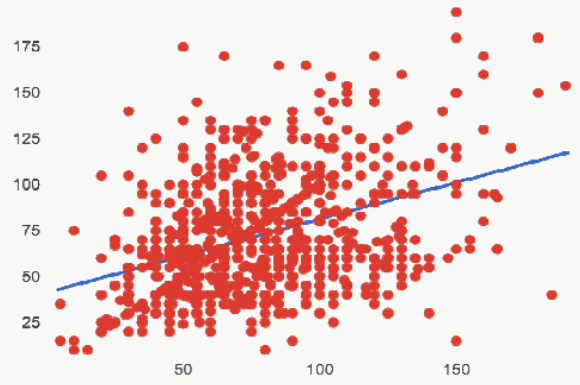
Form :
r close to :

C



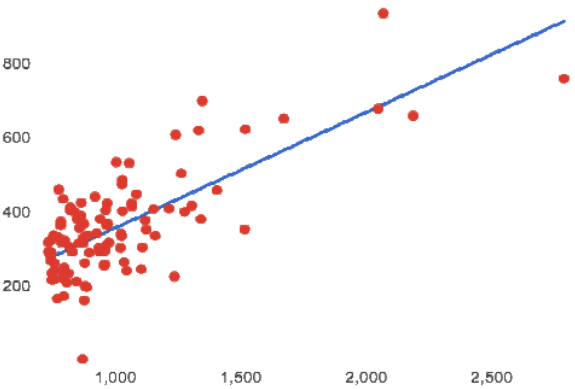
Form :
r close to :

D



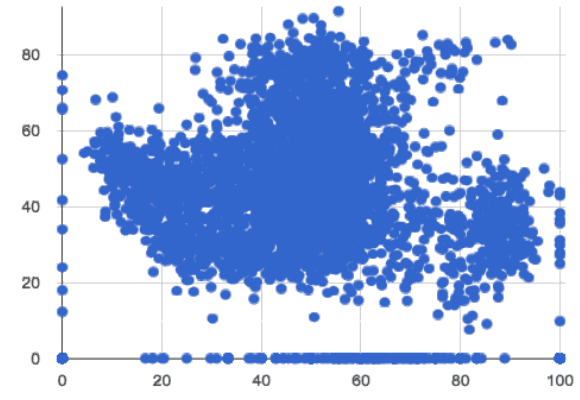
Form :
r close to :

E



Form :
r close to :

F



Form :
r close to :

Correlation Does Not Imply Causation!

Here are some possible correlations and the nonsense headlines a confused journalist might report as a result. In reality, the correlations have absolutely no causal relationship; they come about because both of them are related to another variable that's lurking in the background.

Can you think of another variable for each situation that might be the actual cause of the correlation and explain why the headlines the paper ran based on the correlations are nonsense?

1) **Correlation:** For a certain psychology test, the amount of time a student studied was negatively correlated with their score!
Headline: "Students who study less do better!"

2) **Correlation:** Weekly data gathered at a popular beach throughout the year showed a positive correlation between sunburns and shark attacks.
Headline: "Sunburns Attract Shark Attacks!"

3) **Correlation:** A negative correlation was found between rain and ski accidents.
Headline: "Be Safe - Ski in the Rain!"

4) **Correlation:** Medical records show a positive correlation between Tylenol use and Death Rates.
Headline: "Tylenol use increases likelihood of dying!"

5) **Correlation:** A positive correlation was found between hot cocoa sales and snow ball fights.
Headline: "Beware: Hot Cocoa Drinking encourages Snow Throwing!"

Correlations in the Animals Dataset

1) Create a scatter plot for the Animals Starter File, using `"pounds"` as the xs and `"weeks"` as the ys.

- **Form:** Does the point cloud appear linear or nonlinear? _____
- **Direction:** If it's linear, does it appear to go up or down as you move from left to right? _____
- **Strength:** Is the point cloud tightly packed, or loosely dispersed? _____
- Would you predict that the r-value is positive or negative? _____
- Will it be closer to zero, closer to ±1, or in between? _____
- What r-value, does CODAP compute when you type `r-value(animals-table, "pounds", "weeks")`? _____
- Does this match your predictions? _____

_____

_____

2) Create a scatter plot for the Animals Dataset, using `"age"` as the xs and `"weeks"` as the ys.

- **Form:** Does the point cloud appear linear or nonlinear? _____
- **Direction:** If it's linear, does it appear to go up or down as you move from left to right? _____
- **Strength:** Is the point cloud tightly packed, or loosely dispersed? _____
- Would you predict that the r-value is positive or negative? _____
- Will it be closer to zero, closer to ±1, or in between? _____
- What r-value does CODAP compute? _____
- Does this match your prediction? _____

3) Is this correlation **stronger** or **weaker** than the correlation for `"pounds"`? _____

4) What does that *mean*? _____

_____

_____

_____

Correlations in My Dataset

1) There may be a correlation between _____ (column) and _____ (column).

I think it is a _____ (strong/weak), _____ (positive/negative) correlation,

because _____

It might be stronger if I looked at _____ (a sample or extension of my data)

---

2) There may be a correlation between _____ (column) and _____ (column).

I think it is a _____ (strong/weak), _____ (positive/negative) correlation,

because _____

It might be stronger if I looked at _____ (a sample or extension of my data)

---

3) There may be a correlation between _____ (column) and _____ (column).

I think it is a _____ (strong/weak), _____ (positive/negative) correlation,

because _____

It might be stronger if I looked at _____ (a sample or extension of my data)

---

4) There may be a correlation between _____ (column) and _____ (column).

I think it is a _____ (strong/weak), _____ (positive/negative) correlation,

because _____

It might be stronger if I looked at _____ (a sample or extension of my data)

Introduction to Linear Regression

How much can one point move the line of best fit?

Open the Interactive Regression Line (Geogebra). Move the blue point "P", and see what effect it has on the red line.

1) Move P so that it is **centered amongst** the other points. Now move it all the way to top and bottom of the screen.

2) Move P so that it is **far to the left or right** of the other points. Now move it all the way to top and bottom of the screen. How - if at all - does the x-position of P impact on the line of best fit? _____

_____

3) Could the ***regression line*** ever be above or below *all* the points *(including the blue one you're dragging)*? Why or why not? _____

_____

4) Would it be possible to have a line with more points on one side than the other? Why or why not? _____

_____

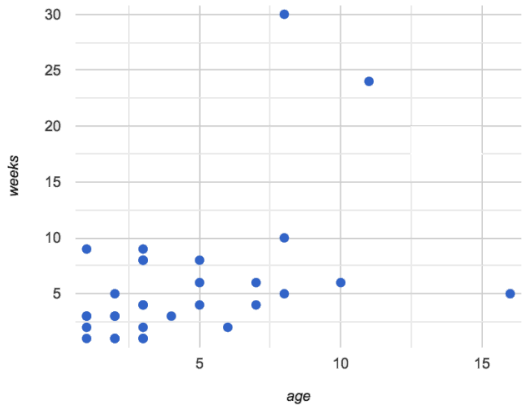
5) What is the highest r -value you can get? _____ Where did you place P? (_____, _____)

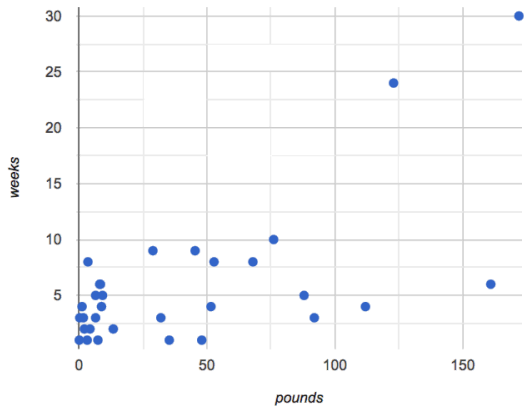
6) What function describes the regression line with this value of P? $y =$ _____ $x +$ _____

7) What is the lowest r -value you can get? _____ Where did you place P? (_____, _____)

8) What function describes the regression line with this value of P? $y =$ _____ $x +$ _____

Predictions from Scatter Plots





9) Use a straight edge to draw what you think would be the line of best fit for **age vs. weeks** (on the left).
Is this a strong correlation that will allow us to make a good prediction of an animal's adoption time just by knowing how old it is?

_____

10) Use a straight edge to draw what you think would be the line of best fit for **pounds vs. weeks** (on the right).
Is this a strong correlation that will allow us to make a good prediction of an animal's adoption time just by knowing how heavy it is?

_____

11) Do either or both of the relationships appear to be linear?

_____

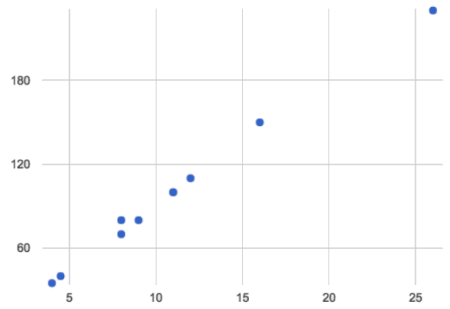
Drawing Predictors

Remember what we learned about r-values…

r = -1	r = -0.5	r = 0	r = 0.5	r = 1
perfect negative correlation	moderate negative association	no correlation	moderate positive association	perfect positive correlation

For each of the scatter plots below, draw a **predictor line** that seems like the best fit. Describe the correlation in terms of Direction and whether you think it is **generally** *stronger* or *weaker*, then estimate the r -value as being close to -1, -0.5, 0, +0.5, or +1.

A



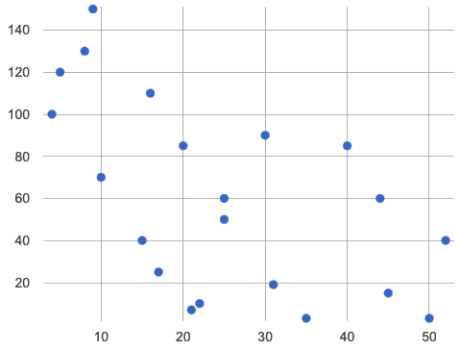
Direction: Positive Negative None

Strength: Stronger Weaker

I would guess that **r** is closest to…

-1 -0.5 0 0.5 1

B



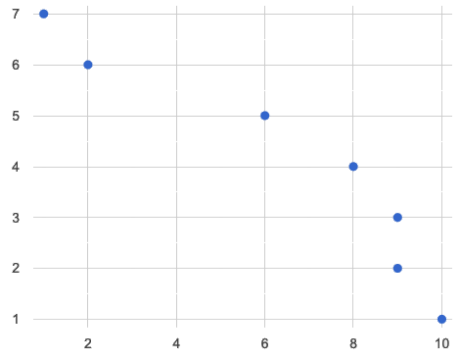
Direction: Positive Negative None

Strength: Stronger Weaker

I would guess that **r** is closest to…

-1 -0.5 0 0.5 1

C



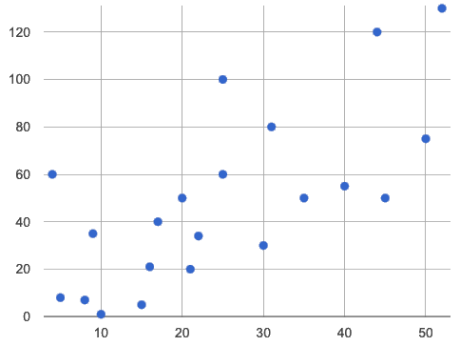
Direction: Positive Negative None

Strength: Stronger Weaker

I would guess that **r** is closest to…

-1 -0.5 0 0.5 1

D



Direction: Positive Negative None

Strength: Stronger Weaker

I would guess that **r** is closest to…

-1 -0.5 0 0.5 1

Exploring lr-plot

age

You should already have created a Least Squares line with `Weeks` *on the x-axis and* `Age` *on the y-axis in the Animals Starter File.*

1) What is the predictor function? $y =$ _____ $x +$ _____ $r =$ _____

2) What is the slope? _____

3) What is the y-intercept? _____

4) How long would our line of best fit predict it would take for a 5 year-old animal to be adopted? _____

5) What if they were a newborn, or just 0 years old? _____

6) Does it make sense to find the adoption time for a newborn using this predictor function? Why or why not?

_____

weight

Make another Least Squares Line, but this time use the animals' weight as our explanatory variable instead of their age.

7) How long would our line of best fit predict it would take for an animal weighing 21 pounds to be adopted? _____

8) What if they weighed 0.1 pounds? _____

cats

9) Using the Animals Dataset - Cats Only, make another Least Squares Line, comparing the `age` v. `weeks` columns for *only the cats*.

10) What is the predictor function? $y =$ _____ $x +$ _____ $r =$ _____

11) What is the slope? _____

12) What is the y-intercept? _____

13) How does this line of best fit for *cats* compare to the line of best fit for *all animals*? _____

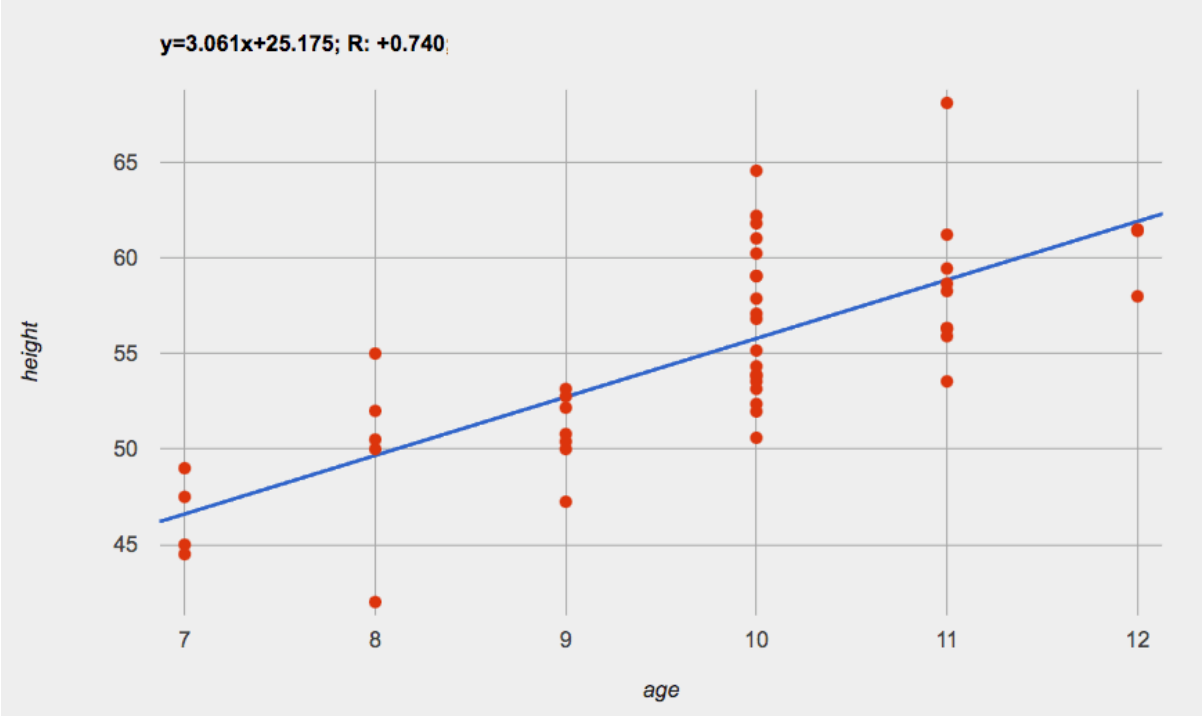
_____

_____

14) How long would our line of best fit predict it would take for a 5 year-old cat to be adopted? _____

★ Using Animals Dataset - Dogs Only, make another Least Squares line, comparing the `age` v. `weeks` columns for *only the dogs*.

Making Predictions



1) About how many inches are kids in this dataset expected to grow per year? _____

2) At that rate, if a child were 45" tall at age eight, how tall would you expect them to be at age twelve? _____

3) At that rate, if a ten-year-old were 55" tall, how tall would you expect them to have been at age 9? _____

4) Using the equation, how tall would you expect a seven-year-old child to be? _____

5) How many of the seven-year-olds in this sample are actually that height? _____

6) Using the equation, determine the expected height of someone who is…

7.5 years old	13 years old	6 years old	newborn	90 years old

7) For which ages is this predictor function likely to be the **most** accurate? Why? _____

_____

_____

8) For which ages is this predictor function likely to be the **least** accurate? Why? _____

_____

_____

Interpreting Regression Lines & r-Values

Use the predictor function and r-value from each linear regression finding on the left to fill in the blanks of the corresponding description on the right.

1	`sugar(m) = −3.19m + 12` `r = −0.05`	For every additional Marvel Universe movie released each year, the average person is predicted to consume _____ [amount] _____ [more / fewer] pounds of sugar! This correlation is _____ [strong, moderate, weak, practically non-existent].
2	`height(s) = 1.65s + 52` `r = 0.89`	Shoe size and height are _____ [strongly, moderately, weakly, not], _____ [positively / negatively] correlated. If person A is one size bigger than person B, we predict that they will be roughly _____ [amount] inches taller than person B as well.
3	`babies(u) = 0.012u + 7.8` `r = 0.01`	There is _____ [a strong, a moderate, almost no] relationship found between the number of Uber drivers in a city and the number of babies born each year.
4	`score(w) = −15.3w + 1150` `r = −0.65`	The correlation between weeks-of-school-missed and SAT score is _____ [strong, moderate, weak, practically non-existent] and _____ [positive / negative]. For every week a student misses, we predict a _____ [amount] point _____ [gain / drop] in their SAT score.
5	`weight(n) = 1.6n + 160` `r = 0.12`	There is a _____ [strong, moderate, weak, practically non-existent], _____ [positive / negative] correlation between the number of streaming video services someone has, and how much they weigh. For each service, we expect them to be roughly _____ [amount] pounds heavier.

Data Cycle: Regression Analysis (Animals)

Open the Animals Starter File. Before completing a data cycle on your own, read the provided example.

Ask Questions

How big of a factor is age in determining adoption time?
What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

all animals at the shelter
Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

name, age, and weeks
What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

Set y-axis to weeks, set x-axis to age. Select least squares line from the Measure menu.
What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

I performed a linear regression on a sample of animals at the shelter [dataset or subset] and found a moderate (R=.448), positive [weak / strong / moderate (R=...), positive / negative] correlation between age [x-axis] and time to adoption [y-axis]. I would predict that a 1 year [x-axis units] increase in age [x-axis] is associated with a .789 week [slope, y-units] increase [increase / decrease] in time to adoption [y-axis].

Ask Questions

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

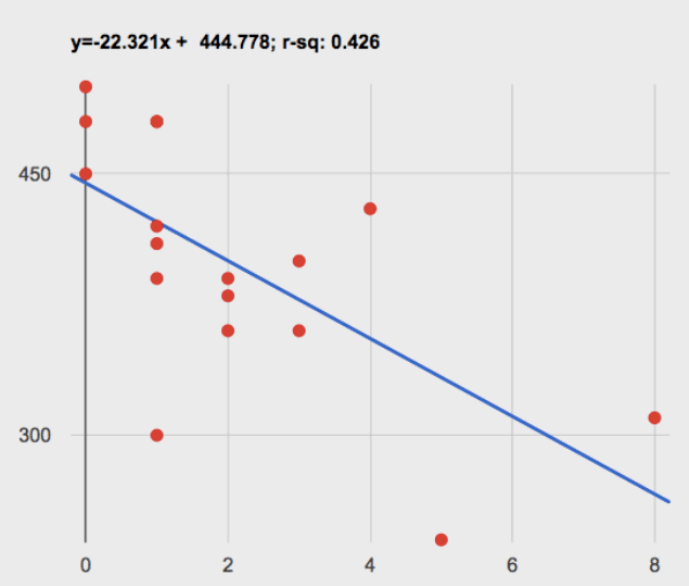
Interpret Data

I performed a linear regression on a sample of _____ [dataset or subset] and found a _____ [weak / strong / moderate (R=...), positive / negative] correlation between _____ [x-axis] and _____ [y-axis]. I would predict that a 1 _____ [x-axis units] increase in _____ [x-axis] is associated with a _____ [slope, y-units] _____ [increase / decrease] in _____ [y-axis].

Describing Relationships

A small sample of people were surveyed about their coffee drinking and sleeping habits. Does drinking coffee impact one's amount of sleep?
NOTE: this data is made up for instructional purposes!

Daily Cups of Coffee	Sleep (minutes)
3	400
0	480
8	310
1	300
1	390
2	360
1	410
0	500
2	390
1	480
3	360
4	430
0	450
5	240
1	420
2	380
1	480



1) Describe the relationship between coffee intake and minutes of sleep shown in the data above.

2) Why is the y-axis of the display above misleading?

Data Cycle: Regression Analysis (My Dataset)

Open your chosen dataset. Ask a question about your data to tell your Data Story.

Ask Questions

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

I performed a linear regression on a sample of _____ [dataset or subset] and found a _____ [weak / strong / moderate (R=...), positive / negative] correlation between _____ [x-axis] and _____ [y-axis]. I would predict that a 1 _____ [x-axis units] increase in _____ [x-axis] is associated with a _____ [slope, y-units] _____ [increase / decrease] in _____ [y-axis].

Ask Questions

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

I performed a linear regression on a sample of _____ [dataset or subset] and found a _____ [weak / strong / moderate (R=...), positive / negative] correlation between _____ [x-axis] and _____ [y-axis]. I would predict that a 1 _____ [x-axis units] increase in _____ [x-axis] is associated with a _____ [slope, y-units] _____ [increase / decrease] in _____ [y-axis].

Case Study: Ethics, Privacy, and Bias

These questions are designed to accompany one of the case studies provided in the Ethics, Privacy, and Bias

My Case Study is _____

1) Read the case study you were assigned, and write your summary here.

2) Is this a good thing or a bad thing? Why?

3) What are the arguments on *each* side?
Data Science used for this purpose is good because…

Data Science used for this purpose is bad because…

Data Collection in a Nutshell

With Great Power Comes Great Responsibility

Politicians pass laws, shoppers choose brands, and countries go to war based on studies that sound reliable. But sometimes the data those decisions are made on is unreliable and misleading!

There are many ways for a study and its analysis to be flawed, whether by accident, by incompetence or by intent.

Being an ethical data scientist means making sure that every element of your study is designed to minimize bias in the data and analysis.

It is also best practice to acknowledge any limitations of datasets we create by writing a Datasheet for the Dataset that describes how the data was collected, what efforts were made to avoid bias, and what data may have been left out, so that people who are trying to make sense of studies that use the dataset don't have to wonder about how reliable it is for the purposes they want to use it for.

Data Cleaning

In order to process data, it needs to be clean. Four ways that data can be dirty include:

1) **Missing Data** - A column containing some cells with data, but some cells left blank.

2) **Inconsistent Types** - A column where some values have one data type and some cells have another. For example, a years column where almost every cell is a Number, but one cell contains the string "5 years old".

3) **Inconsistent Units** - A column where the data types are the same, but they represent different units. For example, a weight column where some entries are in pounds but others are in kilograms.

4) **Inconsistent Naming** - Inconsistent spelling and capitalization for entries lead to them being counted as different. For example, a species column where some entries are "cat" and others are "Cat" will not give us a full picture of the cats.

Once the data is dirty, we have to make careful choices about how to clean it. It's never as simple as just deleting dirty rows! That might, for example, lead us to draw conclusions about the world in general based on a dataset the underrepresents the reality for developing countries.

Survey Validation

We can design a survey to improve the odds of getting clean data. A few design features that improve results include:

1) **Required Questions** - By making a question "required", we can eliminate missing data and blank cells.

2) **Question Format** - When you have a fixed number of categories, a drop-down can ensure that everyone selects one - and only one! - category.

3) **Descriptive Instructions** - Sometimes it's helpful to just add instructions! This can remind respondents to use inches instead of centimeters, for example, or give them extra guidance to answer accurately.

4) **Adding Validation** - Most survey tools allow you to specify whether some data should be a number or a string, which helps guard against inconsistent types. Often, you can even specify parameters for the data as well, such as "strings that are email addresses", or "numbers between 24 and 96".

Analyzing Survey Results When Data is Dirty

These questions are designed to accompany the Survey of Eighth Graders and their Favorite Desserts Starter File.

1) Paolo made a `dot plot` of the `dessert` column and was surprised to discover that **Fruit** was the most popular dessert among 8th graders!

Make the `dot plot` in CODAP to see what he's looking at. Why is this display misleading? How is the data "dirty"?

2) What ideas do you have for how the survey designer could have made sure that the data in the `dessert` column would have been cleaner?

3) Make a data visualization showing the ages of the 8th graders surveyed. What "dirty" data problems do you spot and how are they misleading?

4) What ideas do you have for how the survey designer could have made sure that the data in the `age` column would have been cleaner?

5) Experiment with making data visualizations for other columns. What other issues can you spot? What other suggestions do you have for how the survey could have been improved?

Dirty Data!

Open the New Animals Spreadsheet and take a careful look. A bunch of new animals are coming to the shelter, and that means more data!

What do you Notice?	What do you Wonder?

There are many different ways that data can be dirty!

a. **Missing Data** - A column containing some cells with data, but some cells left blank.
b. **Inconsistent Types** - A column with inconsistent data types. For example, a `years` column where almost every cell is a Number, but one cell contains the string `"5 years old"`.
c. **Inconsistent Units** - A column with consistent data types, but inconsistent units. For example, a `weight` column where some entries are in pounds but others are in kilograms.
d. **Inconsistent Naming** - Inconsistent spelling and capitalization for entries lead to them being counted as different. For example, a `species` column where some entries are `"cat"` and others are `"Cat"` will not give us a full picture of the cats.

1) Which animals' row(s) have **missing data**? _____

2) Which column(s) have **inconsistent types**? _____

3) Which column(s) have **inconsistent units**? _____

4) Which column(s) have **inconsistent naming**? _____

5) If we want to analyze this data, what should we do with the rows for Tanner, Toni, and Lizzy? _____

_____

6) If we want to analyze this data, what should we do with the rows for Chanel and Bibbles? _____

_____

7) If we want to analyze this data, what should we do with the rows for Porche and Boss? _____

_____

8) If we want to analyze this data, what should we do with the row for Niko? _____

_____

9) If we want to analyze this data, what should we do with rows for Mona, Rover, Susie Q, and Happy? _____

_____

10) Sometimes data cleaning is straightforward. Sometimes the problem is evident but the solution is less certain. For which questions were you certain of your data cleaning suggestion? For which were you less certain? Why? _____

_____

_____

Bad Questions Make Dirty Data

The **Height v Wingspan Survey** has *lots* of problems, which can lead to many kinds of dirty data: Missing Data, Inconsistent Types, Inconsistent Units and Inconsistent Language! Using the link provided by your teacher to your class' copy of the survey, try filling it out with bad data. Record the problems for each question and make some recommendations for how to improve the survey!

	What examples of bad data were you able to submit?	How could the survey be improved to avoid bad data?
A Age		
B Grade		
C Height		
D Wingspan		

Filter and Booleans

A **Boolean** is a type of data with two values: true and false.

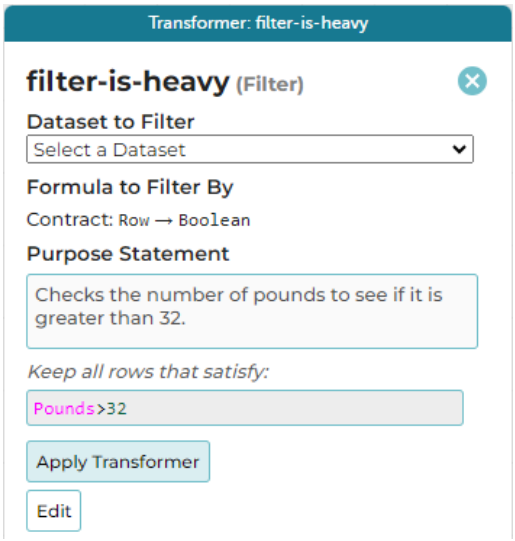
Transformers allow us to transform datasets to produce new, distinct output datasets, instead of modifying the original input dataset itself. We use them to manipulate tables and enable low-stakes "what if?" exploration.

We must provide the `Filter` Transformer with a Boolean expression, which evaluates to true or false. `Filter` then produces a copy of the input dataset that only has the cases for which the expression evaluated to true.

Every Transformer we make requires a unique expression. It's important to get the expression just right, or the Transformer will produce an error. Strings belong inside quotation marks, but Booleans do not!

Booleans and Filters (1)

Notice & Wonder



1) What do you Notice about the `Filter` Transformer on the left, which you can also view in the Boolean Starter File?

2) What do you Wonder about the `Filter` Transformer?

3) In the Boolean Starter File, open `filter-is-heavy`. (To do so, select the `-` that appears on the left when you hover.) Select "Apply Transformer". In your own words, describe what happened when you applied the Filter transformer to the Animals Dataset.

Some Booleans You Might Know

In the `filter-is-heavy` Transformer (above), we used a *Boolean* expression to tell CODAP that we wanted to keep all rows where Pounds was greater than 32. The *greater than* symbol (>) is an example of a Boolean operator that you're probably already familiar with.

4) Here are six different Booleans that we will use in CODAP.

- Put a check mark by the Booleans where you can guess what they do.
- Put a question mark by any Booleans that you're not sure about.

>	<	=	>=	>=	!=

Boolean-producing expressions are yes-or-no questions and will always evaluate to either `true` ("yes") or `false` ("no"). What will each of the expressions below evaluate to? Write down your prediction in the space provided. You'll get a chance to see if you were correct on the next page.

5) `3 <= 4`

6) `"a" > "b"`

7) `3 = 2`

8) `"a" < "b"`

9) `2 < 4`

10) `"a" = "b"`

11) `5 >= 5`

12) `"a" != "a"`

13) `4 >= 6`

14) `"a" >= "a"`

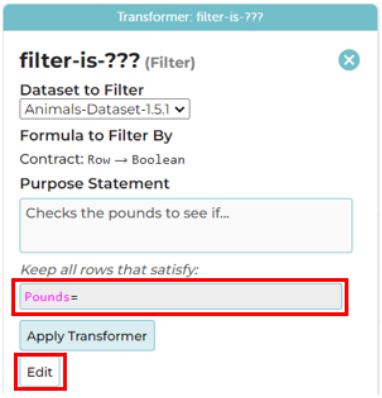
15) `3 != 3`

16) `"a" != "b"`

Booleans and Filters (2)

Booleans and Numbers

In the Boolean Starter File, open the Transformer called `filter-is-???`, pictured below. For each prompt below, you will select "Edit" in the Transformer, and then enter the specified Boolean expression. (Relevant boxes are highlighted in red in the image on the right.)



1) Click Edit. Change `Pounds=` so that it says `Pounds=32`. What happened? _____

_____

2) What would be a good **name** for a Transformer with the expression `Pounds=32`?

_____

3) What would be a good Purpose Statement for a Transformer with the expression `Pounds=32`?

_____

4) With your partner, test out each of the Booleans listed below, using `Pounds` _____ `32` as the Transformer's expression.

- What happens if you put `<` in the blank? _____
- What happens if you put `>` in the blank? _____
- What happens if you put `<=` in the blank? _____
- What happens if you put `>=` in the blank? _____
- What happens if you put `!=` in the blank? _____

Booleans and Strings

5) Click Edit. This time, type `Name>"Maple"` in the expression box. What happened? _____

_____

6) Predict what will happen if you edit the expression so that it says `Name<="Maple"` (then try it!). _____

_____

7) With your partner, test out each of the Booleans listed below, using `Name` _____ `Maple` as the expression.

- What happens if you put `<` in the blank? _____
- What happens if you put `=` in the blank? _____
- What happens if you put `>=` in the blank? _____
- What happens if you put `<=` in the blank? _____
- What happens if you put `!=` in the blank? _____

8) Edit the Transformer's expression so that it says: `beginsWith(Name, "Sn")`. What happened? _____

_____

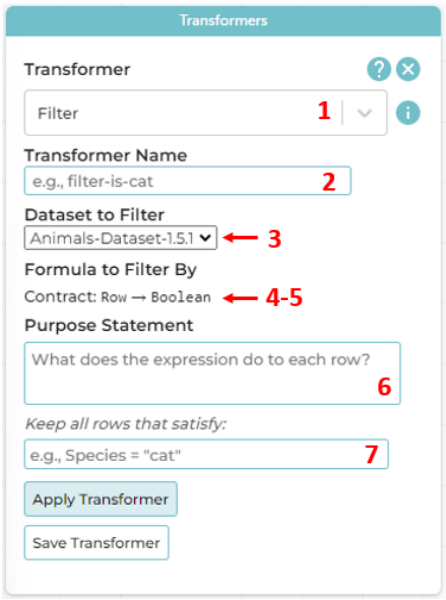
9) Now try this expression: `beginsWith(Name, "sn")`. Did you get the result you expected? _____

★ Go back to Booleans and Filters (1) and use a different color pen to correct any questions (4-15) that you got wrong.

Filter

Make sure you're logged into the Animals Starter File in CODAP. Select the Plugins icon, then choose Transformers.

Create, Apply, and Save a Filter Transformer (Step by Step)



1) Choose Filter from the drop-down menu that appears (Box 1).

2) Name the Transformer `filter-is-dog`. Type the name into Box 2 (left).

3) Click on "Dataset to Filter" to confirm that the Animals Dataset is selected.

4) The Contract's Domain is Row. Why does that makes sense?

_____

5) The Range - is Boolean. Why does that make sense?

_____

6) What Purpose Statement will you type into Box 6?

7) Enter `Species = "dog"` as the expression (Box 7). Select `Apply Transformer`. What happens? _____

_____

8) Try typing `species = "dog"` as the expression (instead of `Species = "dog"`). What happens? _____

_____

9) What are some other possible reasons you might get an error message for the expression? _____

_____

10) Select "Save Transformer." Describe what happens. Why might it be useful to save a Transformer? _____

_____

More Filtering (On Your Own)

11) Create, save, and apply a Transformer called `filter-is-old` that creates a new dataset with animals older than 5 years.

- How many rows does the resulting table have? _____
- How many datasets appeared in the drop-down menu for you to choose from? _____
- Which dataset did you choose and why? _____

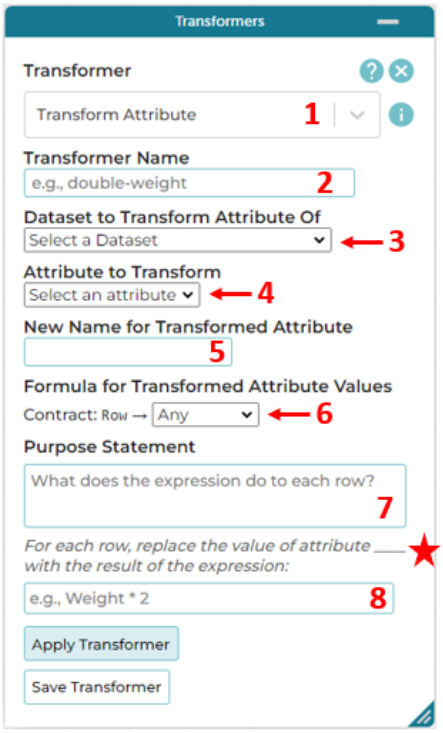
12) Create, save, and apply a Transformer called `filter-is-fixed` that creates a new dataset with only fixed animals.

- How many fixed animals are there at the shelter? _____

Transform Attribute

Make sure you're logged into the Animals Starter File in CODAP. Select the Plugins icon, then choose Transformers.

Create, Apply, and Save a Transform Attribute Transformer (Step by Step)



1) Choose Transform Attribute from the drop-down menu.

2) We want to create a Transformer that will replace all ages less than 1 with the Boolean `true`. In other words, it will *transform* our "age" column into a column that tells us if an animal is young or not. What is a good name for this Transformer?

_____

3) Select the dataset you'd like to transform.

4) What attribute will we be transforming? _____

★ Select the attribute. Notice that CODAP replaced the blank in the starred line of text (left) with the attribute name you selected!

5) What would be an appropriate name for our transformed attribute? _____

6) The Contract includes a Domain (row) only. What is the Range? _____

7) Let's write a Purpose Statement: *Checks each* _____ *to see if* _____

8) What is the expression? _____

9) Apply, the Transformer, and then Save it.

More Transforming (On Your Own)

Create a Transformer called `transform-pounds-kg`. (Note: To convert pounds to kilograms, divide pounds by 2.205.)

10) How many kilograms is the heaviest animal in the shelter? *Hint: If you want to see the animals listed in order by weight, select the attribute name and select "Sort Ascending."*

_____

Create a Transformer called `transform-pounds-round` that uses this expression: `round(Pounds)`.

11) What do you think the `round` function does? _____

_____

Create a Transformer called `transform-Name+Species` that transforms Name using this expression: `concat(Name, Species)`. Let's call the Transformed Attribute `Name+Species`.

Write a Purpose Statement that describes what this expression does to each row. _____

_____

Create a Transformer to change the number of weeks to adoption to instead show the number of days to adoption.

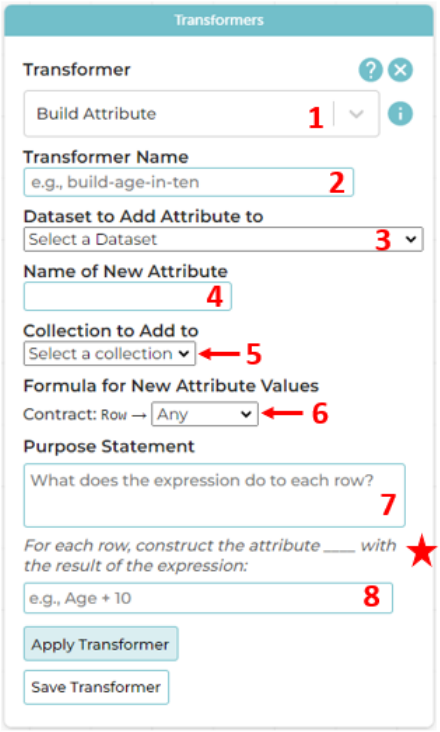
12) What is your Purpose Statement? _____

13) What expression will you use? _____

Build Attribute

Make sure you're logged into the Animals Starter File in CODAP. Select the Plugins icon, then choose Transformers.

Create, Apply, and Save a Build Attribute Transformer (Step by Step)



1) Choose Build Attribute from the drop-down menu (Box 1).

2) At the shelter, animals are considered heavy when they weigh more than 40 pounds. Enter `build-is-heavy` as the Transformer Name (Box 2). What does this name tell you about the Transformer we are creating?

_____

3) Select the dataset you'd like to transform (Box 3).

4) Let's name our new attribute `Heavy` (Box 4). What happened to the starred text (left) when you named the attribute?

_____

5) Ensure that the collection you are adding to is "cases" (Box 5).

6) A domain is provided (row), but not a range. What is the desired output for

`build-is-heavy?` _____

7) Write a purpose statement (Box 7). What do we want the expression to do?

_____

8) Enter `Pounds > 40` as the expression (Box 8).

9) Apply the Transformer. To define the Transformer for future use, select `Save`.

More Building (On Your Own)

Create a Transformer called `build-updated-age`, which will give the animals' ages one year from today.

10) How many animals are 9 years-old one year from today? _____

Create a Transformer that builds a column with the number of letters in each animal's name.

11) What did you name your Transformer and the new attribute? _____

12) How many animals have exactly 8 letters in their names? (Feeling adventurous? Try using the `Count` Transformer here!) _____

Create a Transformer to build a column that returns `true` if the number of letters in an animal's name (the column you created in Question 11!) is less than or equal to five.

Note: Does your new attribute name have a space or a hyphen? If so, CODAP will produce an error when you apply your Transformer. Either change the name of the attribute or wrap your entire attribute name inside tick marks (`) when you type in your expression. (The tick mark key is in the upper left-hand corner of your keyboard.)

13) What expression will you use? _____

14) Which dataset will you need to apply this Transformer to? Why? _____

_____

Create Transformer Cards

The table `t` below represents three animals from the shelter:

name	sex	age	fixed	pounds
`"Toggle"`	`"female"`	`3`	`true`	`48`
`"Fritz"`	`"male"`	`4`	`true`	`92`
`"Nori"`	`"female"`	`6`	`true`	`35.3`

Create a Transformer card that responds to the given prompt on the left. When you're done, give the Transformer a useful name. We've done the first one to get you started.

	Prompt	Transformer Card	Name & Purpose Statement
1	Create a Transformer that produces a Table containing all animals younger than 5.	**Type:** filter [filter/build/transform] **Dataset:** `t` **Expression:** `age<5`	`filter-if-young` Checks the row to see whether age is less than 5.
2	Create a Transformer that produces a Table showing all fixed animals.	**Type:** _____ [filter/build/transform] **Dataset:** _____ **Expression:** _____	
3	Create a Transformer that produces a Table with a new column ("age next year") that adds 1 year to each age.	**Type:** _____ [filter/build/transform] **Dataset:** _____ **Name of New Attribute:** _____ **Expression:** _____	
4	Create a Transformer that produces a Table that transforms pounds to kilos (divide by 2.205) but does **not** add a new column.	**Type:** _____ [filter/build/transform] **Dataset:** _____ **Attribute to Transform:** _____ **Name of New Attribute:** _____ **Expression:** _____	
5	Create a Transformer that produces a Table that doubles pounds but does **not** add a new column.	**Type:** _____ [filter/build/transform] **Dataset:** _____ **Attribute to Transform:** _____ **Name of New Attribute:** _____ **Expression:** _____	

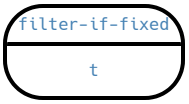
Matching Composed Transformers

The table `t` below represents four animals from the shelter:

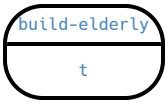
name	sex	age	fixed	pounds
`"Toggle"`	`"female"`	`3`	`true`	`48`
`"Fritz"`	`"male"`	`4`	`true`	`92`
`"Nori"`	`"female"`	`6`	`true`	`35.3`
`"Maple"`	`"female"`	`3`	`true`	`51.6`

Match each Circle of Evaluation (left) to the description of what it does (right).

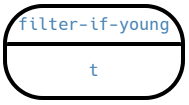
1



2

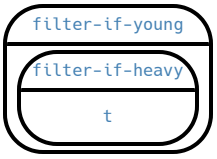


3

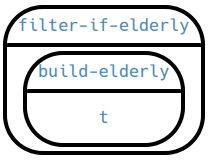


4

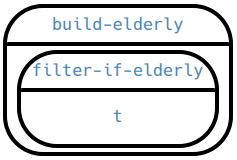
5



6



7



8

A Produces a table containing only Toggle and Maple

B Produces a table with only Maple

C Produces a table that no longer has an "age" column

D Produces a table with an extra column, named "elderly"

E Produces an empty table

F Produces a table containing the same four animals

G Won't run: will produce an error (if so, why?)

H Produces a table with only Nori

Planning Transformer Composition

The table `t` below represents four animals from the shelter:

name	sex	age	fixed	*pounds*
`"Toggle"`	`"female"`	3	`true`	48
`"Fritz"`	`"male"`	4	`true`	92
`"Nori"`	`"female"`	6	`true`	35.3
`"Sasha"`	`"female"`	1	`false`	6.5

You have several Transformers already defined:

filter-if-young filters out animals younger than 4	**filter-if-female** filters out animals that are female	**filter-if-heavy** filters out animals whose weight is greater than 20 kilos	**build-kilos** builds a new column that converts pounds to kilos	**transform-kilos** transforms kilos to grams

For each prompt on the left, draw the Circle of Evaluation that will produce the desired table or display.

	Prompt	Circle of Evaluation
	Produce a Table containing all young, fixed animals	
	Produce a Table showing all animals that weigh more than 20 kilograms	
	Produce a Table showing all female animals that weigh more than 20 kilograms	
	Produce a Table that provides all animals' weights in grams	
	Produce a Table for all female animals, which includes their weight in grams	

Grouped Samples from the Animals Dataset

You've already created and saved the following transformers: `filter-is-old`, `filter-is-young`, `filter-is-cat`, `filter-is-dog`, `filter-is-female`, `filter-is-fixed`, and `filter-has-s-name`. Provide the transformers you would use in the order you would use them. We've given you the solution for the first sample, to get you started.

	Subset	List the transformers *in order*	Use function notation
1	Kittens	`filter-is-young`, `filter-is-cat`	filter-is-young(filter-is-cat(animals-table))
2	Puppies		
3	Fixed Cats		
4	Cats with "s" in their name		
5	Old Dogs		
6	Fixed Animals		
7	Old Female Cats		
8	Fixed Kittens		
9	Fixed Female Dogs		
10	Old Fixed Female Cats		

Visualizing Data

Fill in the tables below, then use CODAP to make the following visualizations. The first table has been filled in for you.

1) A `bar-chart` showing how many puppies (young dogs) are fixed or not.

What Rows?	Which Column(s)?	What will you Create?
puppies	*fixed*	*bar-chart*

2) A `box-plot` showing how many heavy dogs are fixed or not.

What Rows?	Which Column(s)?	What will you Create?

3) A `dot-plot` of the number of `weeks` it takes for a random sample of animals to be adopted.

What Rows?	Which Column(s)?	What will you Create?

4) A `box-plot` of the number of `pounds` that kittens (young cats) weigh.

What Rows?	Which Column(s)?	What will you Create?

5) A `scatter-plot` of a 35 random animals using `species` as the labels, `age` as the x-axis, and `weeks` as the y-axis.

What Rows?	Which Column(s)?	What will you Create?

6) Describe **your own grouped sample** here, and fill in the table below.

What Rows?	Which Column(s)?	What will you Create?

Data Cycle: Analyzing Categorical Data

Use the Animals Starter File to analyze categorical data with the data cycle.

Ask Questions

How many of each species are fixed at the shelter?

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

If you only need some rows, write an expression for your Filter Transformer here.

If you need to Transform or Build an attribute, write the expression for your Transformer here.

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

What did you find out? What can you infer?

What - if any - new question(s) does this raise?

Ask Questions

Are there more female cats than male cats at the shelter?

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)

What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data

If you only need some rows, write an expression for your Filter Transformer here.

If you need to Transform or Build an attribute, write the expression for your Transformer here.

What display, measure, or table do you want to create (i.e., median, bar chart, scatterplot, etc.)?

Interpret Data

What did you find out? What can you infer?

What - if any - new question(s) does this raise?

Threats to Validity in a Nutshell

Threats to Validity can undermine a conclusion, even if the analysis was done correctly.

People Make Mistakes

Sometimes even well-meaning Data Scientists can make mistakes if they're not careful. Data Scientists need to be careful to avoid the four threats below.

- **Selection bias** - identifying the favorite food of the rabbits won't tell us anything reliable about what all the animals eat.
- **Study bias** - If someone is supposed to assess how much cat food is eaten each day on average, but they only measure how much cat food is put in the bowls (instead of how much is actually consumed), they'll end up with an over-estimate.
- **Poor choice of summary** - Suppose a different shelter that had 10 animals recorded adoption times (in weeks) as 1, 1, 1, 7, 7, 8, 8, 9, 9, 10. Using the mode (1) to report what's typical would make it seem like the animals were adopted more quickly than they really were, since 7 out of 10 animals took at least 7 weeks to be adopted.
- **Confounding variables** - Some shelter workers might prefer cats, and steer people towards cats as a result. This would make it appear that "cats are more popular with people", when the real variable dominating the sample is what *workers at the shelter* prefer.

Fake News

But sometimes, it's not an accident: **some people deliberately misuse statistics to create "Fake News" and manipulate others!** An evil Data Scientist might make the four mistakes above *on purpose!* Here are some other slimy ways to make an analysis invalid:

- **Using the Wrong Measure of Center** - With heavily-skewed data (like income in America), using the mean is deeply misleading.
- **Using a Correlation to Imply Causation** - Just because two variables are correlated doesn't mean one is *causing* the other!
- **Incorrect Interpretation of a Visualization** - Someone might point to the tallest bar in a bar chart or histogram and say "See? Most of the people surveyed said…", even if the tallest bar represents only a small percentage of the people surveyed!
- **Intentionally Using the Wrong Chart** - Surveying pet-owners at a dog park to ask about their favorite animal is obviously misleading. A Bar Chart will show empty space for the "Cat" category, which would be a huge red-flag that the survey used a biased sample. But using a Pie Chart will hide the problem, because there's no such thing as an "empty pie slice"!
- **Changing the Scale of a Chart** - A change in poverty from 10.1% to 10.3% is really small, but if the y-axis of the graph goes from 10 to 10.5 it will look like a HUGE climb! The same trick can be played with bar charts, histograms, or box-plots, to exaggerate small differences or hide large ones.

Outliers: Do they stay or do they go?

In any population, there are often one or two samples that are way outside the range of the group. These outliers can really change the results of your analysis, by altering up the average or skewing the shape of the data.

- It can be tempting to remove outliers, and *sometimes* there's a good reason to do it! You might spot an obvious typo, or an answer that you can tell was written by accident.
- But *some* outliers are completely valid, and very important! A small town that has a 30x higher rate of cancer than everywhere else might point to something really important!

As Data Scientists, outliers require us to investigate more closely. And whether we decide to keep or remove them, we should *always* explain our reasoning.

Identifying Threats to Validity

Some volunteers from the animal shelter surveyed a group of pet owners at a local dog park. They found that almost all of the owners were there with their dogs. From this survey, they concluded that dogs are the most popular pet in the state.

1) What are some possible threats to the validity of this conclusion?

The animal shelter noticed a large increase in pet adoptions between Christmas and Valentine's Day. They conclude that at the current rate, there will be a huge demand for pets this spring.

2) What are some possible threats to the validity of this conclusion?

Identifying Threats to Validity (2)

The animal shelter wanted to find out what kind of food to buy for their animals. They took a random sample of two animals and the food they eat, and they found that 100% of animals surveyed ate spider food!

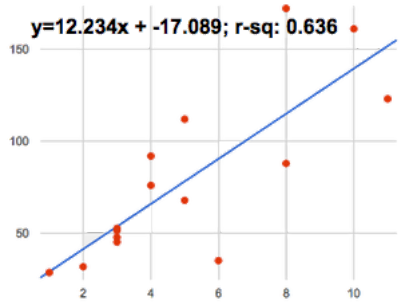
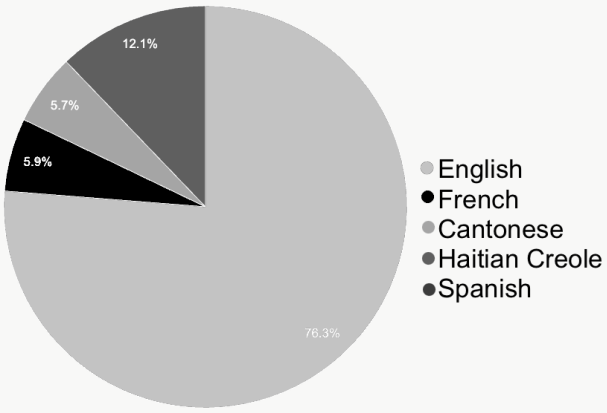
1) Explain why sampling just two animals can result in unreliable conclusions about what kind of food is needed.

A volunteer opens the shelter in the morning and walks all the dogs. At mid-day, another volunteer feeds all the dogs and walks them again. In the evening, a third volunteer walks the dogs a final time and closes the shelter. The volunteers report that the dogs are much friendlier and more active at mid-day, so the shelter staff assume the second volunteer must be better with animals than the others.

2) What are some possible threats to the validity of this conclusion?

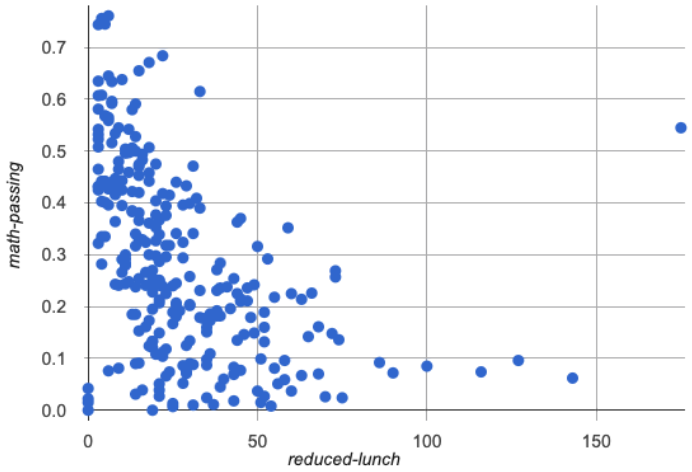
Fake News

The *unrelated* claims below are ALL WRONG! Your job is to figure out why by looking at the data.

	The News Says…	Why it's Fake
1	 *"According to the predictor function indicated here, the value on the x-axis will predict the value on the y-axis 63.6% of the time."*	
2	*The average player on the Cranston East basketball team is 6'1", so we know that most of the players are taller than 6'!*	
3	*Linear regression found a positive correlation (r=0.42) between people's height and salary, so businesses are recognizing that taller people are more qualified.*	
★	 *"Despite El Paso High School being near the Mexican border, a sample of 67 students found that Haitian Creole was the most-commonly spoken language at home (after English)."*	

Outliers: Should they Stay or Should they Go?

Tahli and Fernando are looking at a scatter plot showing the relationship between poverty and test scores at schools in Michigan. They find a trend, with low-poverty schools generally having higher test scores than high-poverty schools. However, one school is an extreme outlier: the highest poverty school in the state also has higher test scores than most of the other schools!



Tahli thinks the outlier should be removed before they start analyzing, and Fernando thinks it should stay. Here are their reasons:

Tahli's Reasons:	Fernando's Reasons:
This outlier is so far from every other school - it *has* to be a mistake. Maybe someone entered the poverty level or the test scores incorrectly! We don't want those errors to influence our analysis. Or maybe it's a magnet, exam or private school that gets all the top-performing students. It's not right to compare that to non-magnet schools.	Maybe it's not a mistake or a special school! Maybe the school has an amazing new strategy that's different from other schools! Instead of removing an inconvenient data point from the analysis, we should be focusing our analysis on what is happening there.

Do you think this outlier should stay or go? Why? What additional information might help you make your decision?

Design Recipe

Directions:

Transformer (check one) ☐ **Filter** ☐ **Transform** ☐ **Build**

Transformer name

Example Tables

What gets filtered/transformed/built? In the sample tables below, (if needed) add the relevant columns.

Original Table

Transformed Table

Contents (Contract, Purpose Statement, and Expression)

Row ->
Domain Range

Purpose: what does the formula do for each row?

i.e. Weight < 20 or Species = "rabbit". Pay careful attention to capitalization and quotation marks.

Directions:

Transformer (check one) ☐ **Filter** ☐ **Transform** ☐ **Build**

Transformer name

Example Tables

What gets filtered/transformed/built? In the sample tables below, (if needed) add the relevant columns.

Original Table

Transformed Table

Contents (Contract, Purpose Statement, and Expression)

Row ->
Domain Range

Purpose: what does the formula do for each row?

i.e. Weight < 20 or Species = "rabbit". Pay careful attention to capitalization and quotation marks.

Design Recipe

Directions:

Transformer (check one) ☐ **Filter** ☐ **Transform** ☐ **Build**

Transformer name

Example Tables

What gets filtered/transformed/built? In the sample tables below, (if needed) add the relevant columns.

Original Table

Transformed Table

Contents (Contract, Purpose Statement, and Expression)

Row ->
Domain Range

Purpose: what does the formula do for each row?

i.e. Weight < 20 or Species = "rabbit". Pay careful attention to capitalization and quotation marks.

Directions:

Transformer (check one) ☐ **Filter** ☐ **Transform** ☐ **Build**

Transformer name

Example Tables

What gets filtered/transformed/built? In the sample tables below, (if needed) add the relevant columns.

Original Table

Transformed Table

Contents (Contract, Purpose Statement, and Expression)

Row ->
Domain Range

Purpose: what does the formula do for each row?

i.e. Weight < 20 or Species = "rabbit". Pay careful attention to capitalization and quotation marks.

Design Recipe

Directions:

Transformer (check one) ☐ **Filter** ☐ **Transform** ☐ **Build**

Transformer name

Example Tables

What gets filtered/transformed/built? In the sample tables below, (if needed) add the relevant columns.

Original Table

Transformed Table

Contents (Contract, Purpose Statement, and Expression)

Row ->
Domain Range

Purpose: what does the formula do for each row?

i.e. Weight < 20 or Species = "rabbit". Pay careful attention to capitalization and quotation marks.

Directions:

Transformer (check one) ☐ **Filter** ☐ **Transform** ☐ **Build**

Transformer name

Example Tables

What gets filtered/transformed/built? In the sample tables below, (if needed) add the relevant columns.

Original Table

Transformed Table

Contents (Contract, Purpose Statement, and Expression)

Row ->
Domain Range

Purpose: what does the formula do for each row?

i.e. Weight < 20 or Species = "rabbit". Pay careful attention to capitalization and quotation marks.

The Animals Dataset

This is a printed version of the animals spreadsheet.
**The numbers on the left side are NOT part of the table!* They are provided to help you identify the index of each row.*

	name	species	sex	age	fixed	legs	pounds	weeks
0	Sasha	cat	female	1	false	4	6.5	3
1	Snuffles	rabbit	female	3	true	4	3.5	8
2	Mittens	cat	female	2	true	4	7.4	1
3	Sunflower	cat	female	5	true	4	8.1	6
4	Felix	cat	male	16	true	4	9.2	5
5	Sheba	cat	female	7	true	4	8.4	6
6	Billie	snail	hermaphrodite	0.5	false	0	0.1	3
7	Snowcone	cat	female	2	true	4	6.5	5
8	Wade	cat	male	1	false	4	3.2	1
9	Hercules	cat	male	3	false	4	13.4	2
10	Toggle	dog	female	3	true	4	48	1
11	Boo-boo	dog	male	11	true	4	123	24
12	Fritz	dog	male	4	true	4	92	3
13	Midnight	dog	female	5	false	4	112	4
14	Rex	dog	male	1	false	4	28.9	9
15	Gir	dog	male	8	false	4	88	5
16	Max	dog	male	3	false	4	52.8	8
17	Nori	dog	female	3	true	4	35.3	1
18	Mr. Peanutbutter	dog	male	10	false	4	161	6
19	Lucky	dog	male	3	true	3	45.4	9
20	Kujo	dog	male	8	false	4	172	30
21	Buddy	lizard	male	2	false	4	0.3	3
22	Gila	lizard	female	3	true	4	1.2	4
23	Bo	dog	male	8	true	4	76.1	10
24	Nibblet	rabbit	male	6	false	4	4.3	2
25	Snuggles	tarantula	female	2	false	8	0.1	1
26	Daisy	dog	female	5	true	4	68	8
27	Ada	dog	female	2	true	4	32	3
28	Miaulis	cat	male	7	false	4	8.8	4
29	Heathcliff	cat	male	1	true	4	2.1	2
30	Tinkles	cat	female	1	true	4	1.7	3
31	Maple	dog	female	3	true	4	51.6	4

Sentence Starters

Use these sentence starters to help describe patterns, make predictions, find comparisons, share discoveries, formulate hypotheses, and ask questions.

Patterns:

- I noticed a pattern when I looked at the data. The pattern is _____
- I see a pattern in the data collected so far. My graph shows _____

Predictions:

- Based on the patterns I see in the data collected so far, I predict that _____
- My prediction for _____ is _____

Comparisons:

- When I compared _____ and _____, I noticed that _____
- The similarities I see between _____ and _____ are _____
- The differences I see between _____ and _____ are _____

Surprises and Discoveries:

- I discovered that _____
- I was surprised by _____
- I noticed something unusual about _____

Hypotheses:

- A possible explanation for what the data showed is _____
- A factor that affected this data might have been _____
- I think this data was affected by _____

Questions:

- I wonder why _____
- I wonder how _____
- How are _____ affected by _____
- How will _____ change if _____

Contracts for Data Literacy Codap

Contracts tell us how to use a function, by telling us three important things:

1. The **Name**
2. The **Domain** of the function - what kinds of inputs do we need to give the function, and how many?
3. The **Range** of the function - what kind of output will the function give us back?

For example: The contract `triangle :: (Number, String, String) -> Image` tells us that the name of the function is `triangle`, it needs three inputs (a Number and two Strings), and it produces an Image.
With these three pieces of information, we know that typing `triangle(20, "solid", "green")` will evaluate to an Image.

Name	Domain	Range
`# bar-chart`	`:: ( Table , String )` (table-name, column)	`-> Image`
`bar-chart(animals-table, "species")`		
`# bar-chart-summarized`	`:: ( Table , String , String )` (table-name, labels, values)	`-> Image`
`bar-chart-summarized(count(animals-table, "species"), "value","count")`		
`# box-plot`	`:: ( Table , String )` (table-name, column)	`-> Image`
`box-plot(animals-table, "weeks")`		
`# box-plot-scaled`	`:: ( Table , String , Number , Number )` (table-name, column, low, high)	`-> Image`
`box-plot-scaled(animals-table, "weeks", 1, 40)`		
`# histogram`	`:: ( Table , String , String , Number )` (table-name, labels, values, bin-size)	`-> Image`
`histogram(animals-table, "species", "weeks", 2)`		
`# line-graph`	`:: ( Table , String , String , String )` (table-name, labels, xs, ys)	`-> Image`
`line-graph(animals-table, "name", "pounds","weeks")`		
`# lr-plot`	`:: ( Table , String , String , String )` (table-name, labels, xs, ys)	`-> Image`
`lr-plot(animals-table, "name", "pounds","weeks")`		
`# mean`	`:: ( Table , String )` (table-name, column)	`-> Number`
`mean(animals-table, "pounds")`		
`# median`	`:: ( Table , String )` (table-name, column)	`-> Number`
`median(animals-table, "pounds")`		
`# modes`	`:: ( Table , String )` (table-name, column)	`-> List`
`modes(animals-table, "pounds")`		
`# modified-box-plot`	`:: ( Table , String )` (table-name, column)	`-> Image`
`modified-box-plot(animals-table, "pounds")`		
`# modified-box-plot-scaled`	`:: ( Table , String , Number , Number )` (table-name, column, low, high)	`-> Image`
`modified-box-plot-scaled(animals-table, "weeks", 1, 40)`		

Name	Domain	Range
# modified-vert-box-plot	:: (Table (table-name) , String (column))	-> Image
modified-vert-box-plot(animals-table, "pounds")		
# modified-vert-box-plot-scaled	:: (Table (table-name) , String (column) , Number (low) , Number (high))	-> Image
modified-vert-box-plot-scaled(animals-table, "weeks", 1, 40)		
# pie-chart	:: (Table (table-name) , String (column))	-> Image
pie-chart(animals-table, "species")		
# pie-chart-summarized	:: (Table (table-name) , String (labels) , String (values))	-> Image
pie-chart-summarized(count(animals-table, "species"), "value", "count")		
# r-value	:: (Table (table-name) , String (xs) , String (ys))	-> Number
r-value(animals-table, "pounds","weeks")		
# random-rows	:: (Table (table-name) , Number (num-rows))	-> Table
random-rows(animals-table, 10) # select 10 random rows from the table		
# scatter-plot	:: (Table (table-name) , String (labels) , String (xs) , String (ys))	-> Image
scatter-plot(animals-table, "name", "pounds","weeks")		
# stdev	:: (Table (table-name) , String (column))	-> Number
stdev(animals-table, "pounds")		
# vert-box-plot	:: (Table (table-name) , String (column))	-> Image
vert-box-plot(animals-table, "weeks")		
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->

These materials were developed partly through support of the National Science Foundation (awards 1042210, 1535276, 1648684, and 1738598) and are licensed under a Creative Commons 4.0 Unported License. Based on a work at www.BootstrapWorld.org. Permissions beyond the scope of this license may be available by contacting contact@BootstrapWorld.org.