

Name: _____



Data Literacy

Fall 2024 Student Workbook - Pyret Edition



BOOTSTRAP
Equity • Scale • Rigor

Workbook v3.1

Brought to you by the Bootstrap team:

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

Bootstrap is licensed under a Creative Commons 4.0 Unported License. Based on a work from www.BootstrapWorld.org.
Permissions beyond the scope of this license may be available at contact@BootstrapWorld.org.

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?

Reflection: Problem Solving Advantages of Diverse Teams

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

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

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

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

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

5) Can you 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?

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?"

Introduction to Computational Data Science

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, 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 decide to sort the animals in *ascending order* (smallest-to-largest) by age. Then we 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 <i>Kujo took a long time to be adopted</i>	<i>Is it because he was so big?</i>	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
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. _____, which contains _____ data. Some example values are:

_____.

b. _____, which contains _____ data. Some example values are:

_____.

What Questions Can You Answer with the Given Data?

The following is a dataset of a bicycle rider's training rides.

date	miles	time (w/stops)	weather	average speed	max speed
04/10/2018	10	44	"cloudy"	13	30
05/30/2018	15	66	"sunny"	13.5	22
06/12/2018	12	61	"rainy"	11.2	25
07/04/2018	24	103	"sunny"	14	26
07/12/2018	24	120	"windy"	12.5	26

1) Decide whether each questions below *can* or *cannot* be answered with the given data and circle your selection.

Question	Answered by this dataset?	
	Yes	No
How many miles did the cyclist ride June 12th?		
What tire pressure produces the highest average speed?		
What is the average time it takes this cyclist to ride 1 mi?		
Does this cyclist ride slower when it is rainy?		
Does this cyclist ride faster when they are late to an appointment?		
How many miles has the cyclist ridden in total as part of their training?		

2) In the space provided below each question, explain *how* you could answer the question using the data or *why you cannot* answer the question.

★ Are there any questions that you could find the answers to more than one way?

Opening Questions

Sports

- Who is the best quarterback of all time?
- Are baseball pitchers throwing harder than ever?
- How much more do male soccer players earn than females?
- How common is it for former Olympic athletes to become coaches?
- How much does an extra inch of height help a basketball player?

Pop Culture

- What percentage of people have seen the movie that won last year's Best Picture Award?
- Who tends to be more popular: bands or solo singers?
- Are younger actors paid more than older actors?
- Are movies with female leads as profitable as movies with male leads?
- Does winning a Grammy increase sales?

Politics

- Is "Stop and Frisk" a racist policy?
- Do Republican politicians tend to come from different states than Democratic ones?
- Do people in countries that have universal healthcare live longer than people in countries that don't?
- Was press coverage slanted for or against a particular candidate?

Education

- Do small schools perform better than large ones?
- Which has a stronger correlation with student achievement: race or wealth?
- Do bilingual classes result in better outcomes for ESL/ELL students?
- How does quality of education differ in various regions of the United States?

Introduction to Programming

The **Editor** is a software program we use to write Code. Our Editor allows us to experiment with Code on the right-hand side, in the **Interactions Area**. For Code that we want to *keep*, we can put it on the left-hand side in the **Definitions Area**. Clicking the "Run" button causes the computer to re-read everything in the Definitions Area and erase anything that was typed into the Interactions Area.

Data Types

Programming languages involve different **data types**, such as Numbers, Strings, Booleans, and even Images.

- 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.
 - In Pyret, any decimal *must* start with a 0. For example, `0.22` is valid, but `.22` is not.
- Strings are values like `"Emma"`, `"Rosanna"`, `"Jen and Ed"`, or even `"08/28/1980"`.
 - All strings *must* be surrounded by quotation marks.
- Booleans are either `true` or `false`.

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

Operators

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

- Operators are written between values, for example: `4 + 2`.
- In Pyret, operators must always have spaces around them. `4 + 2` is valid, but `4+2` is not.
- If an expression has different operators, parentheses must be used to show order of operations. `4 + 2 + 6` and `4 + (2 * 6)` are valid, but `4 + 2 * 6` is not.

Applying Functions

Applying functions works much the way it does in math. Every function has a name, takes some inputs, and produces some output. The function name is written first, followed by a list of **arguments** in parentheses.

- In math this could look like $f(5)$ or $g(10, 4)$.
- In Pyret, these examples would be written as `f(5)` and `g(10, 4)`.
- Applying a function to make images would look like `star(50, "solid", "red")`.
- There are many other functions, for example `num-sqr`, `num-sqrt`, `triangle`, `square`, `string-repeat`, etc.

Functions have **contracts**, which help explain how a function should be used. Every Contract has three parts:

- The *Name* of the function - literally, what it's called.
- The *Domain* of the function - what *type(s) of value(s)* the function consumes, and in what order.
- The *Range* of the function - what *type of value* the function produces.

Strings and Numbers

Make sure you've loaded [code.pyret.org \(CPO\)](http://code.pyret.org), clicked "Run", and are working in the **Interactions Area** on the right. Hit Enter/return to evaluate expressions you test out.

Strings

String values are always in quotes.

- Try typing your name (in quotes!).
- Try typing a sentence like "I'm excited to learn to code!" (in quotes!).
- Try typing your name with the opening quote, but *without the closing quote*. Read the error message!
- Now try typing your name *without any quotes*. Read the error message!

1) Explain what you understand about how strings work in this programming language. _____

Numbers

2) Try typing `42` into the Interactions Area and hitting "Enter". Is `42` the same as `"42"`? Why or why not?

3) What is the largest number the editor can handle?

4) Try typing `0.5`. Then try typing `.5`. Then try clicking on the answer. Experiment with other decimals.

Explain what you understand about how decimals work in this programming language. _____

5) What happens if you try a fraction like `1/3`? _____

6) Try writing **negative** integers, fractions and decimals. What do you learn? _____

Operators

7) Just like math, Pyret has **operators** like `+`, `-`, `*` and `/`.

Try typing in `4 + 2` and then `4+2` (without the spaces). What can you conclude from this?

8) Type in the following expressions, **one at a time**: `4 + 2 * 6` `(4 + 2) * 6` `4 + (2 * 6)` What do you notice?

9) Try typing in `4 + "cat"`, and then `"dog" + "cat"`. What can you conclude from this?

Booleans

Boolean-producing expressions are yes-or-no questions, and will always evaluate to either **true** ("yes") or **false** ("no").

What will the expressions below evaluate to? Write down your prediction, then type the code into the Interactions Area to see what it returns.

	Prediction	Result		Prediction	Result
1) <code>3 <= 4</code>	_____	_____	2) <code>"a" > "b"</code>	_____	_____
3) <code>3 == 2</code>	_____	_____	4) <code>"a" < "b"</code>	_____	_____
5) <code>2 < 4</code>	_____	_____	6) <code>"a" == "b"</code>	_____	_____
7) <code>5 >= 5</code>	_____	_____	8) <code>"a" <> "a"</code>	_____	_____
9) <code>4 >= 6</code>	_____	_____	10) <code>"a" >= "a"</code>	_____	_____
11) <code>3 <> 3</code>	_____	_____	12) <code>"a" <> "b"</code>	_____	_____
13) <code>4 <> 3</code>	_____	_____	14) <code>"a" >= "b"</code>	_____	_____

15) In your own words, describe what `<` does. _____

16) In your own words, describe what `>=` does. _____

17) In your own words, describe what `<>` does. _____

	Prediction:	Result:
--	-------------	---------

18) `string-contains("catnap", "cat")` _____

19) `string-contains("cat", "catnap")` _____

20) In your own words, describe what `string-contains` does. Can you generate another expression using `string-contains` that returns true?

★ There are infinite string values ("a", "aa", "aaa" ...) and infinite number values out there (...-2,-1,0,-1,2...). But how many different *Boolean* values are there? _____

Functions for Tables

Open the [Animals Starter File](#) and click "Run".

In the Interactions Window on the right, type `animals-table` and hit "Enter" to see the default view of the table.

sort

Suppose we wanted to see the names of the animals in alphabetical order...

The `sort` function takes in three pieces of information:

1. A table
2. A column we want to sort the table by (declared using a String)
3. The order in which we want the column sorted (declared using a Boolean)

Test out these two expressions in the Interactions Area and record what you learn about ordering below:

- `sort(animals-table, "species", true)`
- `sort(animals-table, "species", false)`

1) `true` sorts the table... _____

2) `false` sorts the table... _____

Suppose we wanted to sort the `animals-table` by the `weeks` column to determine which animals were adopted quickest...

3) Would you use `true` or `false`? Explain. _____

4) Test it out, and write your thinking about *quantitative* columns at the end of your explanations of `true` and `false` above.

5) Which animal(s) were adopted the quickest? _____

6) Some functions produce Numbers, some produce Strings, some produce Booleans. What did the `sort` function produce? _____

There are many other functions available to us in Pyret. We can describe them using contracts. The Contract for `sort` is:

```
# sort :: Table, String, Boolean -> Table
```

- Each Contract begins with the function name: in this case `sort`
- Lists the data types required to satisfy its Domain: in this case `Table, String, Boolean`
- And then declares the data type of the Range it will return. in this case `Table`
- Contracts can also be written with more detail, by adding *variable names* in the Domain:

```
# sort :: ( Table, String, Boolean ) -> Table
           table-name column-name order
```

Suppose we wanted to sort the `animals-table` by the `legs` column to determine which animals had the most legs...

7) Fill in the blanks below with the code you'd use (We've put pieces of the Contract below each line to help you!):

_____ (_____, _____, _____)
function-name table-name :: Table column-name :: String order :: Boolean

8) Which animal(s) had the most legs? _____

9) Think of another question you might answer quickly by sorting the table.

10) What code would you write to answer your question?

_____ (_____, _____, _____)
function-name table-name :: Table column-name :: String order :: Boolean

Functions for Tables (continued)

count

count :: Table, String -> Table

1) What is the Domain of count ? _____

2) What is the Range of count ? _____

3) What do you suspect the String in the Domain will describe? _____

Suppose we wanted to know how many animals had 4 legs...

Type count(animals-table, "legs") into the Interactions Area and click "Enter"

4) What did the expression produce? _____

5) How many animals had 4 legs? _____

6) Think of another question you might be able to answer with the count function.

7) Fill in the blanks with the code you'd write.

_____ (_____ table-name :: Table , _____ column-name :: String)

8) Tables that summarize data with a count are commonly used in the real world. Give two examples of where you've seen them before:

- Example 1: _____
- Example 2: _____

9) Newscasters and journalists often incorporate data into their reporting. How else might they display this information, besides using a table?

first-n-rows

10) Type first-n-rows(animals-table, 5). What happens? _____

11) If we wanted a table of the first 3 rows of the animals-table, what code would you write? _____

12) What is the Contract for first-n-rows ? _____

★ What happens when you type first-n-rows(sort(animals-table, "pounds", true), 5)?

Note: In this case, the output of sort(animals-table, "pounds", true) is the Table first-n-rows is taking in!

★★ See if you can figure out how to compose the code that would generate a table of the 10 oldest animals!

_____ (_____ Table _____ , _____ Number _____)

Circles of Evaluation: Count, Sort, First-n-rows

For each scenario below, draw the Circle of Evaluation and then use it to write the code.

When you're done, test your code out in the [Animals Starter File](#) and make sure it does what you'd expect it to.

count :: Table, String -> Table

first-n-rows :: Table, Number -> Table

sort :: Table, String, Boolean -> Table

1) We want to see the 10 animals who were adopted the quickest.

Circle of Evaluation:

code: _____

2) We want to see the heaviest animal.

Circle of Evaluation:

code: _____

3) We want to take the first 8 animals from the table and put them in alphabetical order (by name).

Circle of Evaluation:

code: _____

4) You notice that the lightest 16 animals weigh under 10 pounds and you want to know the count (*by species*) of those animals.

Circle of Evaluation:

code: _____

Catching Bugs when Sorting Tables

Learning about a Function through Error Messages

- 1) Type `sort` into the Interactions Area of the [Animals Starter File](#) and hit "Enter". What do you learn? _____
- 2) We know that all functions need an open parenthesis and at least one input! Type `sort(animals-table)` in the Interactions Area and hit Enter/return. Read the error message. What hint does it give us about how to use this function?

What Kind of Error is it?

syntax errors - when the computer cannot make sense of the code because of unclosed strings, missing commas or parentheses, etc.
contract errors - when the function isn't given what it needs (the wrong type or number of arguments are used)

- 3) In your own words, the difference between **syntax errors** and **contract errors** is: _____

Finding Mistakes with Error Messages

The code below is **BUGGY!** Read the code and the error messages, and see if you can catch the mistake **WITHOUT** typing the code into Pyret.

- 4) `sort(animals-table, name , true)`

The name **name** is unbound:
`sort(animals-table, name , true)`
It is **used** but not previously defined.

This is a _____ error. The problem is that _____
contract / syntax

- 5) `sort(animals-table, "name" , "true")`

The **Boolean annotation**:
`fun sort(t :: Table, col :: String, asc :: Boolean)`
was not satisfied by the value
"true"

This is a _____ error. The problem is that _____
contract / syntax

- 6) `sort(animals-table "name" true)`

Pyret didn't understand your program around:
`sort(animals-table "name" true)`
You may need to add or remove some text to fix your program. Look carefully before **the highlighted text**. Is there a missing colon (:), comma (,), string marker ("), or keyword? Is there something there that shouldn't be?

This is a _____ error. The problem is that _____
contract / syntax

- 7) `sort(animals-table, "name", true`

Pyret didn't expect your program to **end** as soon as it did:
`sort(animals-table, "name", true`
You may be missing an "end", or closing punctuation like ")" or "]" somewhere in your program.

This is a _____ error. The problem is that _____
contract / syntax

- 8) `sort (animals-table, "name", true)`

Pyret thinks this code is probably a function call:
`sort (animals-table, "name", true)`
Function calls must not have space between the **function expression** and the **arguments**.

This is a _____ error. The problem is that _____
contract / syntax

Contracts for Image-Producing Functions

Log into code.pyret.org (CPO) and click "Run". Experiment with each of the functions listed below, trying to find an expression that will build. Record the contract and example code for each function you are able to successfully build!

Name	Domain	Range
# triangle	:: Number, String, String	-> Image
<code>triangle(80, "solid", "darkgreen")</code>		
# star	::	->
# circle	::	->
# rectangle	::	->
# text	::	->
# square	::	->
# ellipse	::	->
# regular-polygon	::	->

Challenge: Composing with Circles of Evaluation

What if we wanted to see your name written on a diagonal?

- We know that we can use the `text` function to make an Image of your name.
- Pyret also has a function called `rotate` that will rotate any Image a specified number of degrees.

`# rotate :: Number, Image -> Image`

But how could the `rotate` and `text` functions work together? Draw a Circle of Evaluation, translate it to code and test it out in the Editor!

Exploring Displays

Use the contracts provided below to make each type of display in the [Animals Starter File](#). Then answer the questions about each display.

Bar Charts # `bar-chart :: Table, String -> Image`

`function-name` (`table-name :: Table`, `column-name :: String`)

Sketch a bar chart below.

Bar charts summarize 1 column of `_____` data.
_____ categorical/quantitative

This kind of display tells us...

Pie Charts # `pie-chart :: Table, String -> Image`

`function-name` (`table-name :: Table`, `column-name :: String`)

Sketch a pie chart below.

Pie charts summarize 1 column of `_____` data.
_____ categorical/quantitative

This kind of display tells us...

Box Plots # `box-plot :: Table, String -> Image`

`function-name` (`table-name :: Table`, `column-name :: String`)

Sketch a box plot below.

Box plots summarize 1 column of `_____` data.
_____ categorical/quantitative

This kind of display tells us...

Histograms # `histogram :: Table, String, String, Number -> Image`

`function-name` (`table-name :: Table`, `labels :: String`, `values :: String`, `bin-width :: Number`)

Sketch a histogram below.

Histograms summarize 1 column of `_____` data.
_____ categorical/quantitative

This kind of display tells us...

Circles of Evaluation: Composing Functions to Make Displays

Using the Contracts below as a reference, draw the Circle of Evaluation for each prompt.

pie-chart :: Table, String -> Image

box-plot :: Table, String -> Image

bar-chart :: Table, String -> Image

first-n-rows :: Table, Number -> Table

histogram :: Table, String, String, Number -> Image

sort :: Table, String, Boolean -> Table

1) Make a bar-chart of the lightest 16 animals by sex.

★ What other bar chart might you want to compare this to? _____

2) Take the heaviest 20 animals and make a histogram of weeks to adoption (use "species" for your labels).

★ What other histogram might you want to compare this to? _____

3) Make a box-plot of age for the 11 animals who spent the most weeks in the shelter.

★ What other box plot might you want to compare this to? _____

4) Make a pie-chart of species for the 18 animals who spent the fewest weeks in the shelter.

★ What other pie chart might you want to compare this to? _____

Exploring Displays (2)

Use the contracts provided below to make each type of display in the [Animals Starter File](#). Then answer the questions about each display.

Line Graphs # `line-graph :: Table, String, String, String -> Image`

```

function-name (
  table-name :: Table,
  column-name :: String,
  column-name :: String,
  column-name :: String
)
  
```

Sketch a line graph below.

Line Graphs summarize 2 columns of data.

This kind of display tells us...

Scatter Plots # `scatter-plot :: Table, String, String, String -> Image`

```

function-name (
  table-name :: Table,
  column-name :: String,
  column-name :: String,
  column-name :: String
)
  
```

Sketch a scatter plot below.

Scatter Plots summarize 2 columns of data.

This kind of display tells us...

LR Plots # `lr-plot :: Table, String, String, String -> Image`

```

function-name (
  table-name :: Table,
  column-name :: String,
  column-name :: String,
  column-name :: String
)
  
```

Sketch an Linear Regression Plot below.

LR Plots summarize 2 columns of data.

This kind of display tells us...

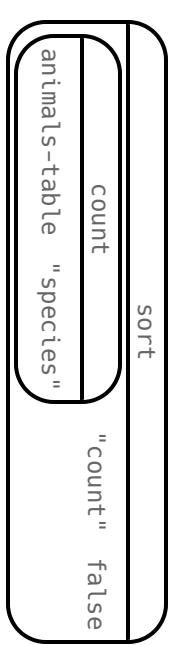
Composing Functions: Match Display Descriptions to Circles of Evaluation

Match each prompt on the left to the Circle of Evaluation used to answer it.

Make a pie-chart, showing the species of the 4 oldest animals.

1

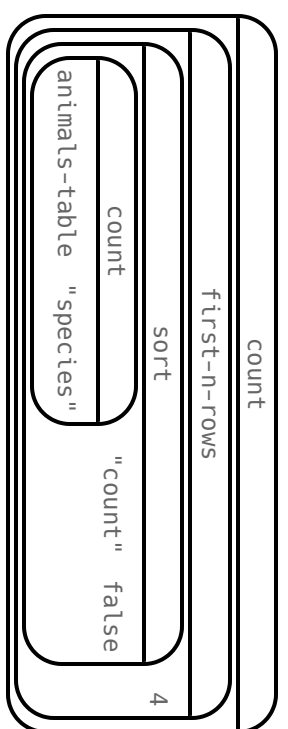
A



Take the 4 heaviest animals and make a box plot of their weight.

2

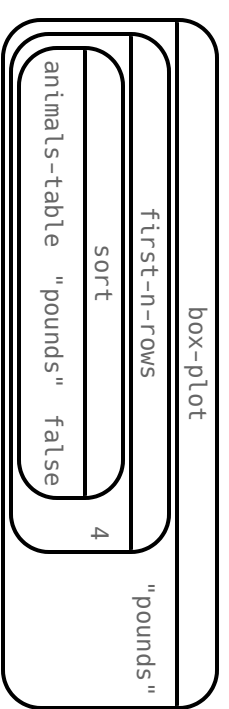
B



Make a table showing the count of the species in this dataset, sorted from most to least.

3

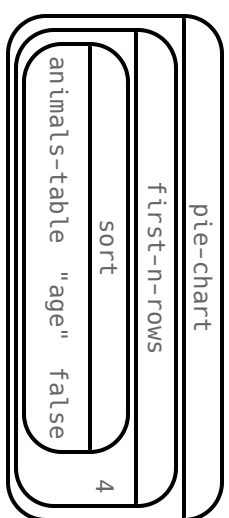
C



Make a table showing the count of the 4 species with the most animals

4

D



Circles of Evaluation: Composing Functions to Make Displays (2)

Using the Contracts below as a reference, draw the Circle of Evaluation for each prompt.

pie-chart :: Table, String -> Image

box-plot :: Table, String -> Image

bar-chart :: Table, String -> Image

first-n-rows :: Table, Number -> Table

histogram :: Table, String, String, Number -> Image

sort :: Table, String, Boolean -> Table

1) Take the youngest 12 animals and make a box-plot of pounds.

What other box plot might you want to compare this to? _____

2) Make a pie-chart of legs for the 10 oldest animals.

What other pie chart might you want to compare this to? _____

★ Take the 20 lightest animals, then take the 10 youngest of *those* animals and make a bar-chart of species

What other pie chart might you want to compare this to? _____

Displaying Categorical Data

Data Scientists use **displays** to visualize data. You've probably seen some of these charts, graphs and plots yourselves!

When it comes to displaying **Categorical Data**, there are two displays that are especially useful:

1. **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 shown 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).
2. **Pie charts** show the *percentage* of rows in each category.
 - Pie charts provide a visual representation of the relative frequency of values in a categorical column.
 - Pie charts have a slice for every category in a column.
 - The more rows in a category, the larger the slice.
 - Slices in a pie chart can be shown in *any order*, without changing the meaning of the chart. However, slices are usually shown in some sensible order (e.g. slices might be shown in alphabetical order or from the smallest to largest slice).

Count, Bar Charts and Pie Charts

Open the [Expanded Animals Starter File](#) and click "Run".

A - Displays for Categorical Data

Test the following expressions in the Interactions Area:

- `count(more-animals, "species")`
- `bar-chart(more-animals, "species")`

1) How are they similar?

2) Which do you like better: the bar chart or the table? Why?

Now test out the expression `pie-chart(more-animals, "species")`

3) How does the pie chart connect to the bar chart you just made?

Note: When you first build a bar chart or pie chart in Pyret, they are interactive displays. That means that you can mouse over them for more information. Hit the up arrow in the interactions area to reload your last expression and test it out!

B - Comparing Bar and Pie Charts

Best completed after [Bar & Pie Chart - Notice and Wonder](#) and [Matching Bar and Pie Charts](#)

4) How are pie charts similar to bar charts?

5) How are pie charts and bar charts different?

6) What information is provided in bar charts that is hidden in pie charts?

7) Why might this sometimes be problematic?

8) When would you want to use one chart instead of another?

C - Bar and Pie Charts for Quantitative Data?

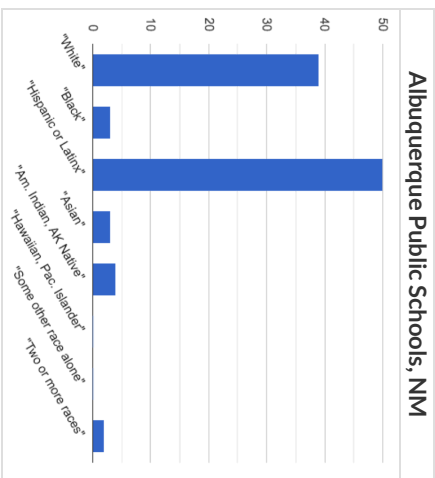
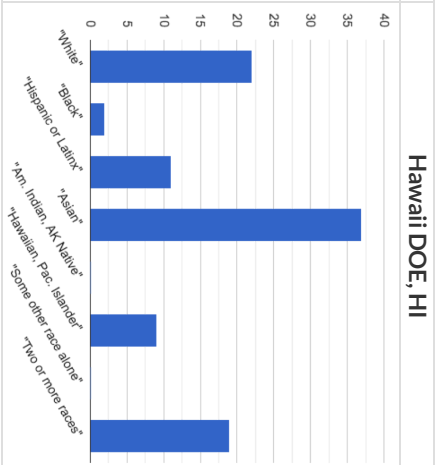
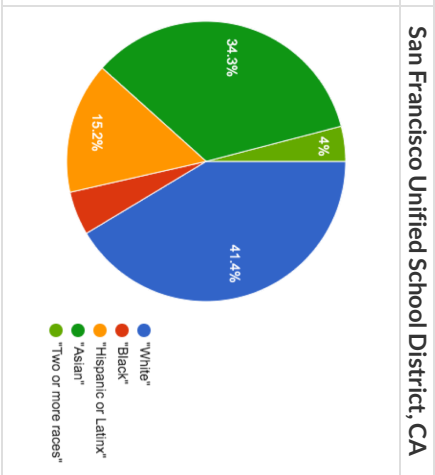
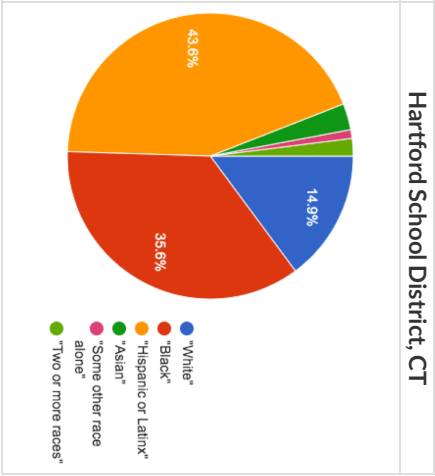
9) Make a `pie-chart` and `bar-chart` for the `pounds` column. Why isn't grouping the `pounds` column very useful?

10) Look at the list of columns in the Definitions Area. For which columns do you expect pie charts to be most useful?

★ What questions about the dataset are you curious to investigate using these displays?

Bar & Pie Chart - Notice and Wonder

What do you Notice and Wonder about the displays below?

Albuquerque Public Schools, NM	Hawaii DOE, HI	San Francisco Unified School District, CA	Hartford School District, CT																																																														
 <table border="1"> <caption>Albuquerque Public Schools, NM</caption> <thead> <tr> <th>Race/Ethnicity</th> <th>Percentage</th> </tr> </thead> <tbody> <tr><td>"White"</td><td>~45%</td></tr> <tr><td>"Black"</td><td>~5%</td></tr> <tr><td>"Hispanic or Latinx"</td><td>~48%</td></tr> <tr><td>"Asian"</td><td>~5%</td></tr> <tr><td>"Am. Indian, AK Native"</td><td>~5%</td></tr> <tr><td>"Hawaiian, Pac. Islander"</td><td>~5%</td></tr> <tr><td>"Some other race alone"</td><td>~5%</td></tr> <tr><td>"Two or more races"</td><td>~5%</td></tr> </tbody> </table>	Race/Ethnicity	Percentage	"White"	~45%	"Black"	~5%	"Hispanic or Latinx"	~48%	"Asian"	~5%	"Am. Indian, AK Native"	~5%	"Hawaiian, Pac. Islander"	~5%	"Some other race alone"	~5%	"Two or more races"	~5%	 <table border="1"> <caption>Hawaii DOE, HI</caption> <thead> <tr> <th>Race/Ethnicity</th> <th>Percentage</th> </tr> </thead> <tbody> <tr><td>"White"</td><td>~35%</td></tr> <tr><td>"Black"</td><td>~5%</td></tr> <tr><td>"Hispanic or Latinx"</td><td>~15%</td></tr> <tr><td>"Asian"</td><td>~38%</td></tr> <tr><td>"Am. Indian, AK Native"</td><td>~5%</td></tr> <tr><td>"Hawaiian, Pac. Islander"</td><td>~10%</td></tr> <tr><td>"Some other race alone"</td><td>~5%</td></tr> <tr><td>"Two or more races"</td><td>~15%</td></tr> </tbody> </table>	Race/Ethnicity	Percentage	"White"	~35%	"Black"	~5%	"Hispanic or Latinx"	~15%	"Asian"	~38%	"Am. Indian, AK Native"	~5%	"Hawaiian, Pac. Islander"	~10%	"Some other race alone"	~5%	"Two or more races"	~15%	 <table border="1"> <caption>San Francisco Unified School District, CA</caption> <thead> <tr> <th>Race/Ethnicity</th> <th>Percentage</th> </tr> </thead> <tbody> <tr><td>"White"</td><td>41.4%</td></tr> <tr><td>"Black"</td><td>4%</td></tr> <tr><td>"Hispanic or Latinx"</td><td>34.3%</td></tr> <tr><td>"Asian"</td><td>15.2%</td></tr> <tr><td>"Two or more races"</td><td>15.2%</td></tr> </tbody> </table>	Race/Ethnicity	Percentage	"White"	41.4%	"Black"	4%	"Hispanic or Latinx"	34.3%	"Asian"	15.2%	"Two or more races"	15.2%	 <table border="1"> <caption>Hartford School District, CT</caption> <thead> <tr> <th>Race/Ethnicity</th> <th>Percentage</th> </tr> </thead> <tbody> <tr><td>"White"</td><td>14.9%</td></tr> <tr><td>"Black"</td><td>35.6%</td></tr> <tr><td>"Hispanic or Latinx"</td><td>43.6%</td></tr> <tr><td>"Asian"</td><td>~1%</td></tr> <tr><td>"Some other race alone"</td><td>~1%</td></tr> <tr><td>"Two or more races"</td><td>~1%</td></tr> </tbody> </table>	Race/Ethnicity	Percentage	"White"	14.9%	"Black"	35.6%	"Hispanic or Latinx"	43.6%	"Asian"	~1%	"Some other race alone"	~1%	"Two or more races"	~1%
Race/Ethnicity	Percentage																																																																
"White"	~45%																																																																
"Black"	~5%																																																																
"Hispanic or Latinx"	~48%																																																																
"Asian"	~5%																																																																
"Am. Indian, AK Native"	~5%																																																																
"Hawaiian, Pac. Islander"	~5%																																																																
"Some other race alone"	~5%																																																																
"Two or more races"	~5%																																																																
Race/Ethnicity	Percentage																																																																
"White"	~35%																																																																
"Black"	~5%																																																																
"Hispanic or Latinx"	~15%																																																																
"Asian"	~38%																																																																
"Am. Indian, AK Native"	~5%																																																																
"Hawaiian, Pac. Islander"	~10%																																																																
"Some other race alone"	~5%																																																																
"Two or more races"	~15%																																																																
Race/Ethnicity	Percentage																																																																
"White"	41.4%																																																																
"Black"	4%																																																																
"Hispanic or Latinx"	34.3%																																																																
"Asian"	15.2%																																																																
"Two or more races"	15.2%																																																																
Race/Ethnicity	Percentage																																																																
"White"	14.9%																																																																
"Black"	35.6%																																																																
"Hispanic or Latinx"	43.6%																																																																
"Asian"	~1%																																																																
"Some other race alone"	~1%																																																																
"Two or more races"	~1%																																																																

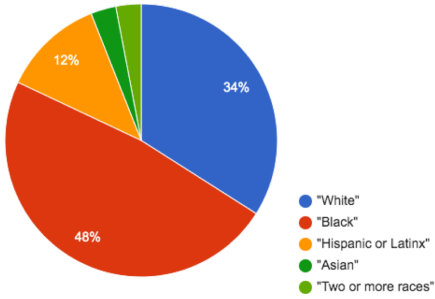
What do you Notice?

What do you Wonder?

Matching Bar and Pie Charts

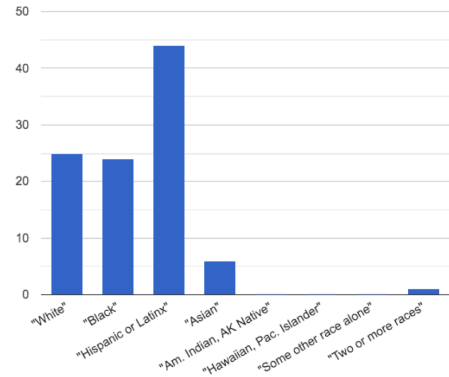
Match each bar chart below to the pie chart that displays the racial demographic data from the same school district.

Cleveland Municipal School District

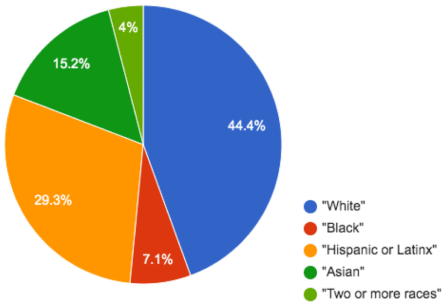


1

A

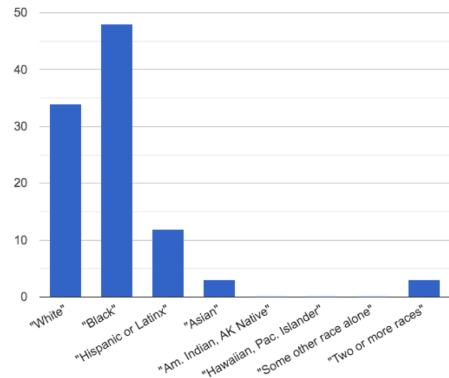


San Diego City Unified School District

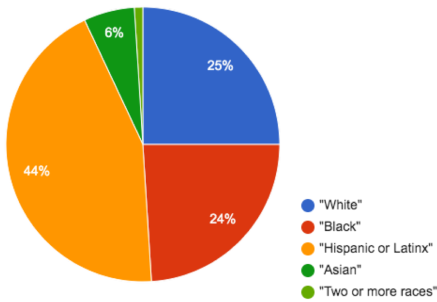


2

B

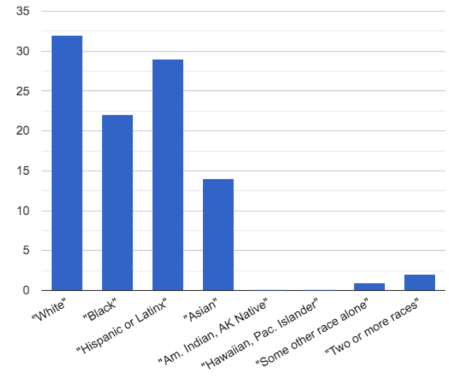


Houston Independent School District

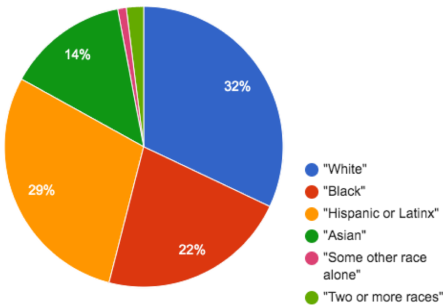


3

C

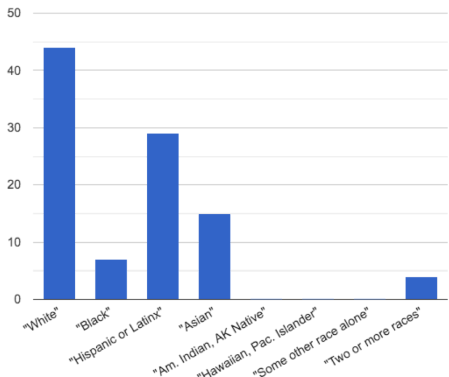


New York City Dept of Education



4

D



Introducing Displays for Subgroups

This page is designed to be used with the [Expanded Animals Starter File](#).

Part A

1) How many tarantulas are male? _____

Hint: Sort the table by species!

2) How many tarantulas are female? _____

3) 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, Pyret has two functions that let us specify both a group and a subgroup:

```
# stacked-bar-chart :: ( Table , String , String ) -> Image
                        table-name group subgroup
# multi-bar-chart :: ( Table , String , String ) -> Image
                      table-name group subgroup
```

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

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

6) What do you notice? _____

7) What do you wonder? _____

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

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

10) 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.

What did you learn? _____

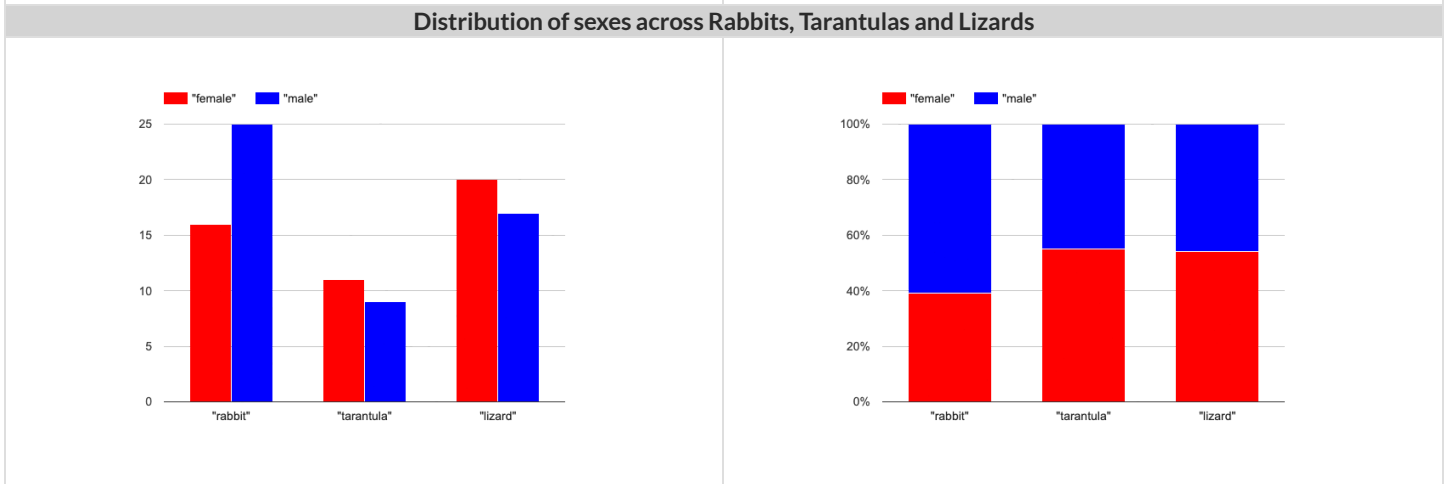
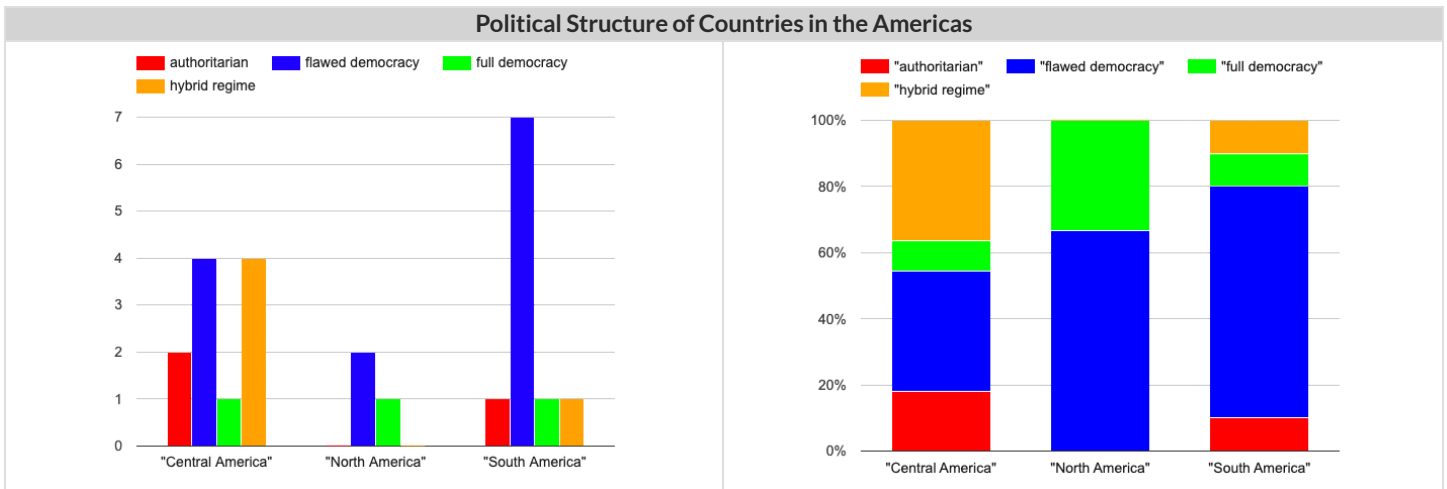
11) Write a different question that would be more efficient to answer with the other kind of display. _____

What did you learn from making this display? _____

Multi Bar & Stacked Bar Charts - Notice and Wonder

The displays on the left are called multi bar charts.

The displays on the right are called stacked bar charts.



What do you Notice?	What do you Wonder?

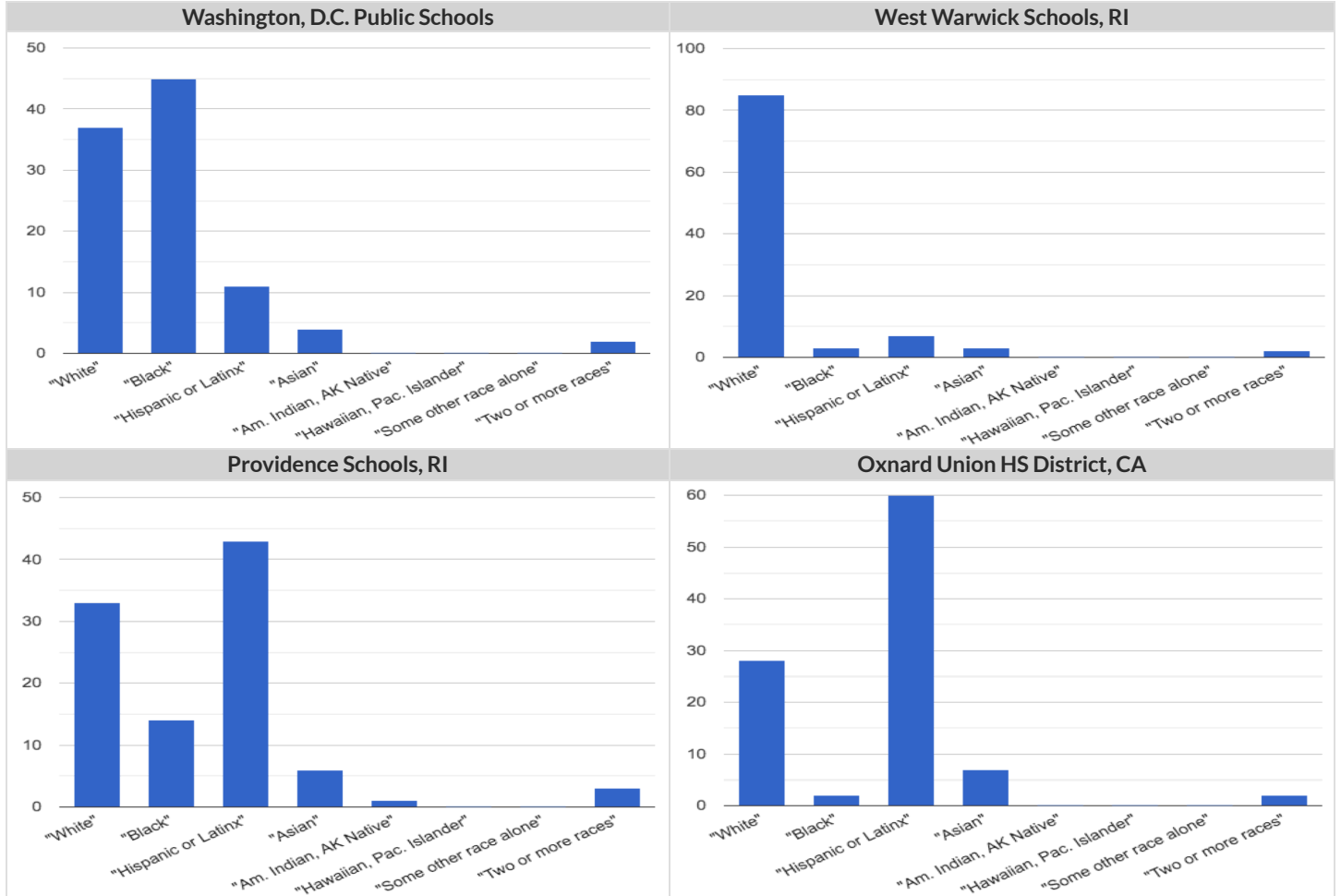
1) Is it possible that the same data was used for the multi bar charts as for the stacked bar charts? How do you know?

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.

Bar Chart - Notice and Wonder

What do you Notice and Wonder about the pie charts below?

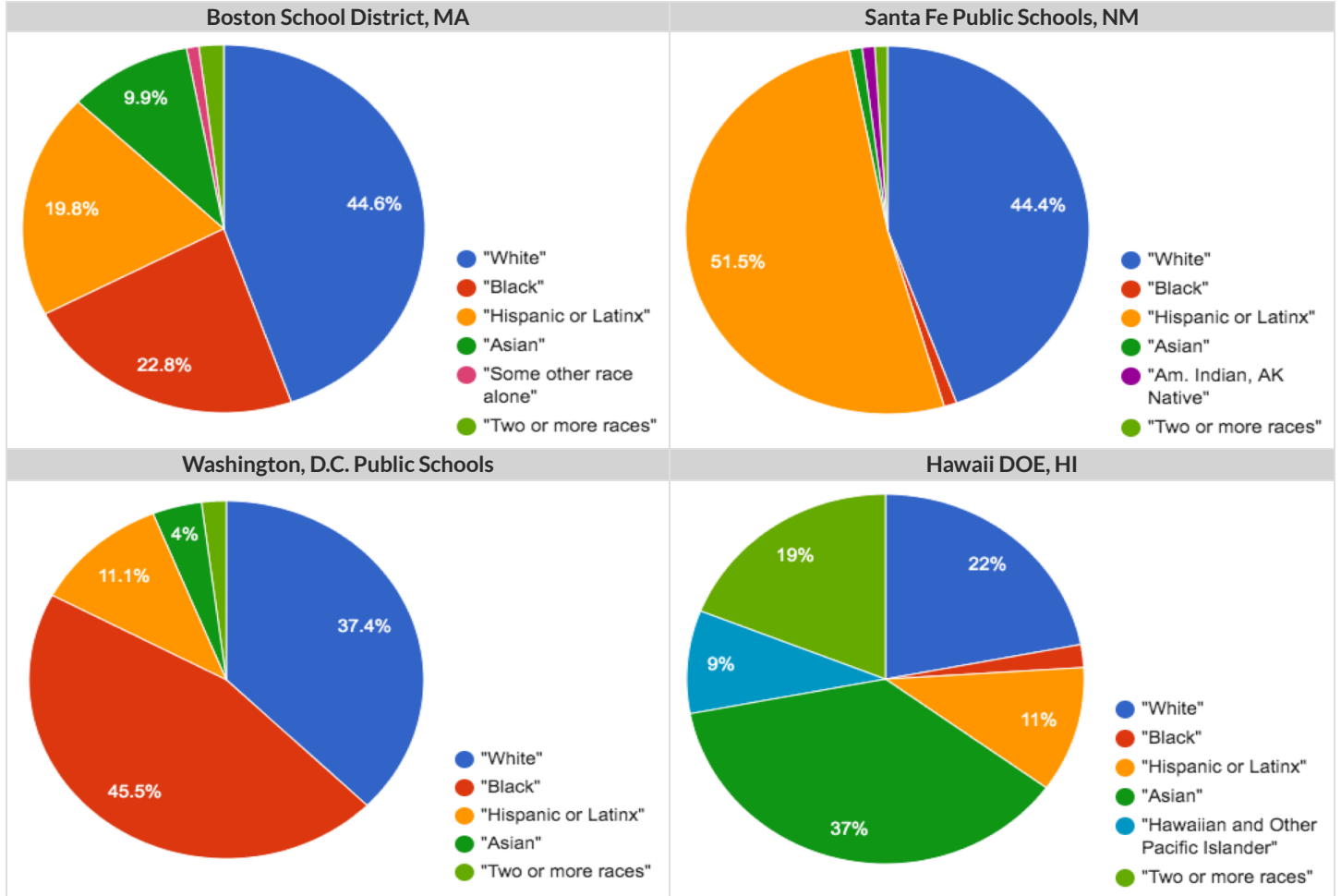


What do you Notice?

What do you Wonder?

Pie Chart - Notice and Wonder

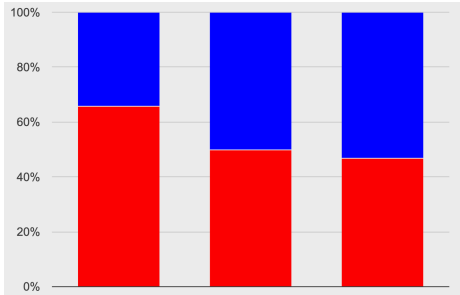
What do you Notice and Wonder about the pie charts below?



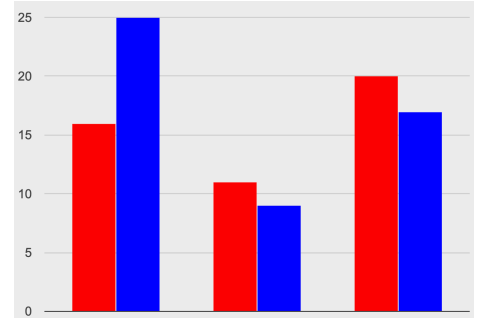
What do you Notice?	What do you Wonder?

Matching Stacked and Multi Bar Charts

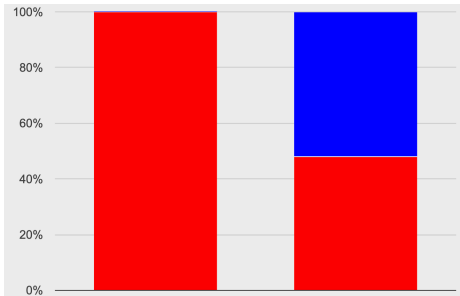
Match each stacked bar chart below to the multi bar chart that displays the same information.



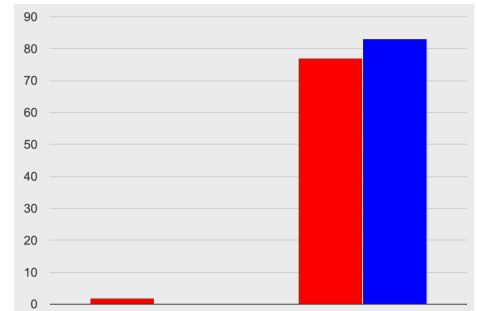
1



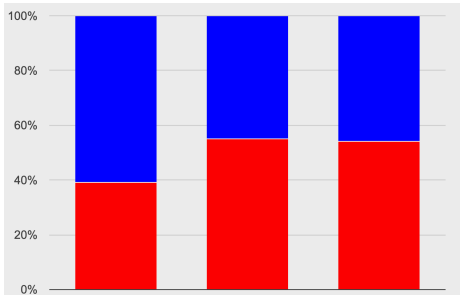
A



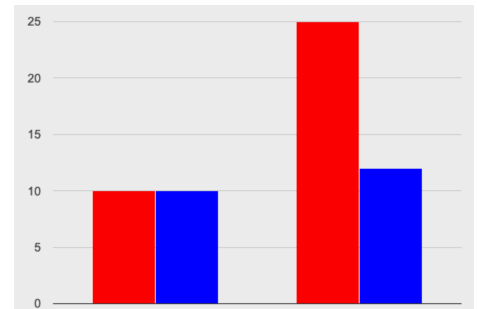
2



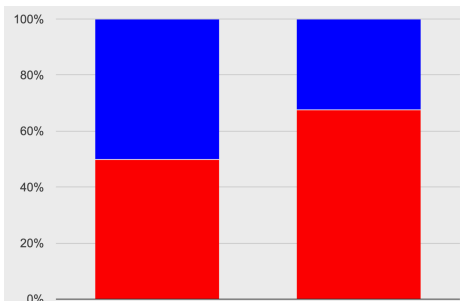
B



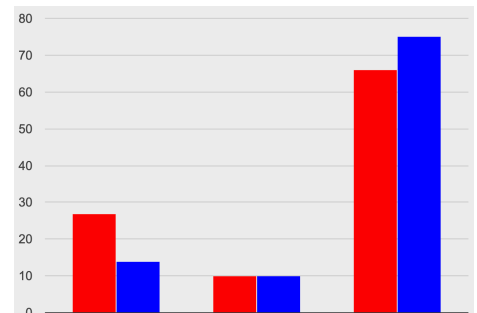
3



C



4



D

Making Infographics Rubric

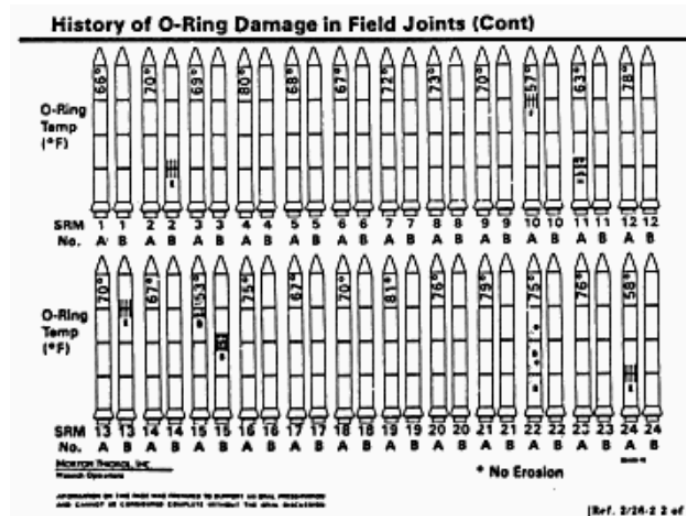
	Wow!	Getting There	Needs Improvement
Preparatory Work	The display or ratio statement formed a strong foundation for the rest of my infographic project.	The display or ratio statement needed revision in order to inspire a meaningful infographic (e.g., it was unclear or it was not interesting).	I did not create a display or ratio statement or what I produced was not conducive to creating a meaningful infographic.
Ratio statement: Impact	My ratio statement will really give those who read it something fascinating to contemplate!	My ratio statement is interesting but probably won't spark any deep conversations.	My ratio statement is dull and uninspired.
Images chosen: Accessibility	The imagery that I used when creating my infographic is inclusive. My images avoid stereotyping and help the viewer relate to and understand the topic.	The imagery that I used mostly avoids stereotyping. More inclusive imagery might help viewers connect with my topic better.	The imagery that I included reinforces stereotypes and might leave some viewers feeling disconnected from my message.
Infographic: Accuracy	The infographic is correctly drawn to scale (every element is in the same proportion).	There were some minor errors made in drawing the infographic to scale.	The infographic is not accurately scaled.
Infographic: Impact	The strategy that I chose (repeated images / bars on a grid / area model) makes sense for my ratio statement and has a strong impact.	The strategy that I chose makes sense but is not terribly impactful; another strategy might have been more effective at conveying my ratio statement.	The strategy that I chose did not make sense in this context nor did it have an impact.

Case Study: NASA Infographic

A day before the 1986 launch of the Challenger, a team of engineers urged NASA to postpone, arguing that launching in cold weather would be extremely dangerous. Parts called "O-rings", they said, were likely to crack in cold weather. A cracked O-ring could lead to a catastrophic explosion – and the death of every astronaut onboard.

Mission control asked the engineers to explain this risk with *data*.

To make their case, the engineers created an infographic that displayed outlines of 48 rockets, each representing a previous launch. Each rocket was labeled with the temperature at launch, with marks showing O-ring damage. These marks were explained in a legend, to help mission control understand what the damage was.



An infographic conveying O-ring damage in 48 rockets

Unfortunately, their infographic was very hard to read:

- Instead of sorting the rockets by *temperature* or *amount-of-damage* (the two variables the engineers claimed were related!), they were sorted by...the date they launched.
- The temperature at launch, which was the most important thing the engineers wanted mission control to see, was written *sideways*, in a tiny font that was difficult to read.
- The marks showing O-ring damage were hard to understand, and the legend that explained them *was on a separate page!*

The engineers created an infographic that failed to clearly explain the risk, and mission control made the decision to go ahead with the launch.

73 seconds into the flight, the rocket exploded over the coast of Florida, killing everyone onboard. The tragedy crippled NASA, which did not launch another rocket for nearly three years.

...The Challenger's explosion was, in the end, attributed to O-ring failure.

Which Silhouette Might Work?

Below are screenshot of the top google search results for 1) pilot transparent silhouette 2) pilot silhouette female 3) pilot silhouette African American.



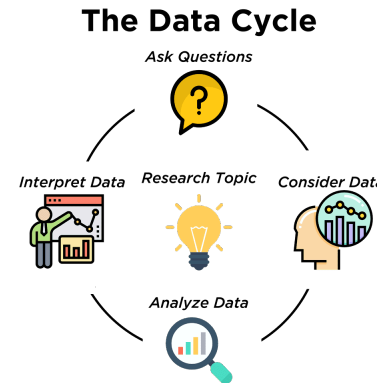
- 1) Put an x on images that read as male only.
- 2) Put a diagonal line on images that read as female only.
- 3) Put a horizontal line (--) through the images that read as a white pilot.
- 4) Circle one silhouette from the remaining images that you think could possibly work as a generalized image of a pilot.
- 5) What do you Notice? What do you Wonder? _____

The Data Cycle

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!



- 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 data sets. 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.
- Next, we **Consider Data**, by determining which parts of the data set 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?
- Then, we **Analyze the Data**, by completing calculations, creating data displays, 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?
- 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 here, but often these 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.

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

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





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

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





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

Data Cycle: Distribution of Fixed Animals

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





<p>Ask Questions</p> 	<p><i>Are more animals fixed or unfixed?</i></p> <p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p><i>All the rows</i></p> <p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p><i>fixed</i></p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>The chart shows that there are _____ fixed animals _____ unfixed animals. <small>more / less / about the same number of</small> <small>as / than</small></p> <p>Some new questions this raises include:</p> <p>_____</p> <p>_____</p> <p>_____</p>	

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





<p>Ask Questions</p> 	<p><i>How does the ratio of fixed to unfixed animals differ by species?</i></p> <p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>The stacked bar chart shows that _____ species have _____ fixed animals _____ unfixed animals. <small>all / most / some / a few / no</small> <small>more / the same number of / fewer</small> <small>as / than</small></p> <p>I also notice _____</p> <p>Some new questions this raises include:</p> <p>_____</p> <p>_____</p>	

Data Cycle: Distribution of Categorical Columns

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

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p><input type="checkbox"/> The chart shows that there is an even distribution of _____ variable _____.</p> <p><input type="checkbox"/> The chart shows that the most common _____ variable _____ is/are _____.</p> <p>I notice that _____</p> <p>I wonder _____</p> <ul style="list-style-type: none"> • How does the distribution of _____ variable _____ differ by _____ variable _____? • _____ <p>Another question I have is...</p> <p>_____</p>	

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

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>When we break the distribution of _____ variable _____ down by _____ variable _____:</p> <ul style="list-style-type: none"> • I notice that _____ • I wonder _____ <p>Another question I have is...</p> <p>_____</p>	

Question Types: Animals

A subset of the whole Animals Dataset is shown in the table below.

name	species	sex	age	fixed	legs	pounds	weeks
Sasha	cat	female	1	false	4	6.5	3
Sunflower	cat	female	5	true	4	8.1	6
Felix	cat	male	16	true	4	9.2	5
Sheba	cat	female	7	true	4	8.4	6
Billie	snail	hermaphrodite	0.5	false	0	0.1	3
Snowcone	cat	female	2	true	4	6.5	5
Wade	cat	male	1	false	4	3.2	1
Hercules	cat	male	3	false	4	13.4	2
Toggle	dog	female	3	true	4	48	1




Using this table - or the full dataset - write three questions of each type below.




- **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* !

	Type	Question
1	Lookup	
2	Lookup	
3	Lookup	
4	Arithmetic	
5	Arithmetic	
6	Arithmetic	
7	Statistical	
8	Statistical	
9	Statistical	




Data Cycle: Analyzing with Count

For each question below, complete the first three steps of the Data Cycle.
Once you know what code to write, type it into Pyret and try it out!

<p>Ask Questions</p> 	<p><i>How many of each species are at the shelter?</i></p> <p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	

<p>Ask Questions</p> 	<p><i>How many of each sex are at the shelter?</i></p> <p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	

For the final Data Cycle, develop your own question and complete the remaining steps.

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	

Snack Habits Rubric

	Wow!	Getting There	Needs Improvement
Data collection	I filled in the Google form each and every time I had a snack. If I was unable to complete the form at the time of the snack, I made a point of completing it as soon as possible. When I responded to the prompts, I gave accurate information (acquired by looking at the nutritional label). This snacking log perfectly accurately represents my snacking.	I filled in the Google form almost every time I had a snack. When I responded to the prompts, I tried my best to give accurate information, but sometimes I made guesses about the number of servings, calories per serving, etc. Overall, the accuracy of the data collected is decent, however.	I often forgot to fill in the Google form when I snacked. I had to go back and dig through my memories to make educated guesses about my snacking habits. The information recorded during the data collection phase is most likely not an accurate depiction of my snacking habits.
Part 1: Our Snacking Habits	I've reflected on the process of tracking my snacking habits, providing interesting details about what I learned. I have offered meaningful noticings and wonderings about our class' snacking habits. I shared a display that I found interesting.	My reflections on the process of tracking my snacking habits are brief and would benefit from additional detail. The observations I shared about our class' snacking habits were shallow. I shared a display, but it was not necessarily interesting.	My reflections on snack tracking and our class dataset are brief, confusing, or missing entirely.
Part 2: US Snacking Habits	I've included an interesting graph and/or statistic from a credible source to represent America's snacking habits. At the end of the slide deck, I've credited my sources. I have explained why the graph caught my attention and what it made me wonder.	I've included a graph and/or statistic to represent America's snacking habits, but the source is not entirely credible. My explanation of why I have chosen this graph is not compelling.	I have either forgotten to include a graph/statistic to represent America's snacking habits, or the graph/statistic that I chose is not appropriate for this project.
Part 3: My statistical question and its answer	I developed a compelling and interesting statistical question based on the data I collected. I clearly answered that question by presenting plots, tables, photos and thoughtful written analysis.	The statistical question I chose is not fully answered by the data presented. I have put in some effort to answer the question with plots, tables, photos and written analysis, but more detail is needed.	Either my statistical question is simple and straightforward, and answering it did not require much critical analysis by me, or my statistical question was not adequately answered by my graphics and written analysis.
Part 4: Conclusion & Sources	I truthfully and honestly answered all questions about the challenges of this project. I addressed in detail how the project's challenges might have affected the quality of my data. I've provided accurate source information.	My discussion of the challenges of this project was brief and lacking in detail. I only partially addressed how this project's challenges might have affected the quality of my data. I've provided some source information.	I did not offer enough thoughtful discussion on the challenges of collecting data. It is not clear to the reader that I understand how challenges I encountered could affect the quality of the data. My source information is missing or inaccurate.

Snack Habits Data

For our purposes, a snack is any food or beverage other than water that you consume between meals.

1) Below is a table of the prompts you will see in the google form you will be completing for each snack you consume over the next 5 days. What do you Notice? What do you Wonder?

2) Complete the table by defining each variable's data type (Number,String, Boolean, Image...).

Prompt	Variable Name	Data Type
Time you ate the Snack <i>Format: The nearest hour on the 24-hour clock (e.g. 4am = 4, 4pm = 16)</i>	time	
Date you ate the Snack <i>Format: 09/23/24</i>	date	
True or False: You ate this snack on a day you went to school?	is-school-day	
What's the name of the snack?	name	
Is your snack salty? sweet? Or neither?	salty-sweet	
How many servings did you eat?	servings	
How many calories per serving?	calories	
How many grams of total fat per serving?	fat	
How many milligrams of sodium per serving?	sodium	
How many grams of sugar per serving?	sugar	
How healthy do you think the snack is? <i>(1- very unhealthy; 5- very healthy)</i>	health-level	
In one word, describe why you are eating the snack.	why	
How much does this snack cost?	cost	
How many ingredients are in this snack?	ingredients	
Take a photo of your snack or beverage. <i>(Your teacher may or may not have included this in the actual google form, but having some images of your snacks will probably be useful for your final project.)</i>	snack-image	

Note: Most snacks come in packages with nutritional value labels that will help you to answer many of these questions. When eating a snack whose package does not include the nutritional value, a simple google search will return an image that looks just like those labels, e.g. "Nutritional Value of an Apple". Similarly, if you get a snack from the cupboard rather than the store, you can google for the price.

Snack Habits Check-In

Name: _____

1) How well have you done collecting data for this project? Circle one of the choices below and explain why you ranked it at that level.
(5) Excellent (4) Very Well (3) Average (2) Below Average (1) Not as well as I wanted (0) Collected no data

2) If you are struggling with data collection, what changes are you going to make so that you can do a better job moving forward?

3) Have you faced any obstacles when it comes to data entry? What were they and how did you overcome them?





4) Do you have any tips for someone who is struggling to stay on top of data entry?

5) Has the process of collecting your own snack data influenced or altered your snacking habits at all? Explain.





6) Do you think it will affect the quality of data? What types of snacks might people not be entering?

Data Cycle: Distribution of Categorical Columns

Explore the distribution of categorical columns in your class' snacking data using **pie-chart**, **bar-chart**, **stacked-bar-chart** or **multi-bar-chart**.

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p><input type="checkbox"/> The chart shows that there is an even distribution of _____ variable _____.</p> <p><input type="checkbox"/> The chart shows that the most common _____ variable _____ is/are _____.</p> <p>I notice that _____</p> <p>I wonder _____</p> <ul style="list-style-type: none"> • How does the distribution of _____ variable _____ differ by _____ variable _____? • _____ <p>Another question I have is...</p> <p>_____</p>	

Explore the distribution of categorical columns in your class' snacking data using **pie-chart**, **bar-chart**, **stacked-bar-chart** or **multi-bar-chart**.

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>When we break the distribution of _____ variable _____ down by _____ variable _____:</p> <ul style="list-style-type: none"> • I notice that _____ • I wonder _____ <p>Another question I have is...</p> <p>_____</p>	

U.S. Snack Habits

1) Choose one statistic (or the title of a display) you found about US snacking habits: _____

2) What is the source? _____

3) What information is available re: the data collection process? (For example: year of data collection, sample size, how the sample was selected, reason for data collection, etc.) You may need to do some digging. _____

4) Based on the above information, what makes you think this data is credible? _____

5) What are some **similarities** between our class snacking habits and the US snacking habits data you found?

6) What are some **differences** between our class snacking habits and the US snacking habits data you found?

7) Do you have any guesses about why the data are similar / different in the ways that you have identified? _____

8) Does anything you turned up in your research surprise you? _____

Probability, Inference, and Sample Size

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

Open the [Expanded Animals Starter File](#), and save a copy.

1) Evaluate the `more-animals` table in the Interactions Area. This is the *complete* population of animals from the shelter!

Here is a true statement about that population: *The population is 47.7% fixed and 52.3% unfixed.*

Type each of the following lines into the Interactions Area and hit "Enter".

```
random-rows(more-animals, 10)
```

```
random-rows(more-animals, 40)
```

2) What do you get? _____

3) What is the Contract for `random-rows`? _____

4) What does the `random-rows` function do? _____

5) In the Definitions Area,

- define `small-sample` to be `random-rows(more-animals, 10)`
- define `large-sample` to be `random-rows(more-animals, 40)`

6) Make a `pie-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 `small-sample` is _____
- The percentage of fixed animals in `large-sample` is _____

7) Make a `pie-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 `small-sample` is _____
- The percentage of tarantulas in `large-sample` is _____

8) Click "Run" to direct the computer to generate a different set of random samples of these sizes. Make a new `pie-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 `small-sample` is _____
- The percentage of tarantulas in `large-sample` is _____

9) Which sample size gave us a more accurate inference about the whole population? Why?

Predictions from Samples

1) In the Definitions Area of the [Expanded Animals Starter File](#), define the following samples:

```
tiny-sample = random-rows(more-animals, 10)
small-sample = random-rows(more-animals, 20)
medium-sample = random-rows(more-animals, 40)
large-sample = random-rows(more-animals, 80)
```

2) Click "Run" and make a pie-chart of the species in the tiny-sample. What animals are in the sample? _____

- Click "Run" for a *new* random tiny-sample, and make *another* pie-chart for species. What animals are in this sample? _____
- Click "Run" for a *new* random sample, and make *yet another* pie-chart for species. Based on these 3 samples, how many species do you think are at the shelter? _____
- Which is the *most common* species at the shelter? _____

3) What did you learn from taking multiple samples that you wouldn't have known if you'd only taken one?

4) Repeat the steps above, but for small-sample. What animals are in the sample?

5) Now that you've seen small-sample, how has your sense of the distribution of the species changed?

6) Now use medium-sample to make a pie-chart of the species. If there are about 400 animals at the shelter, how many of each species would you predict there to be?

7) Now use large-sample to make a pie-chart of the species. If there's anything you'd like to change about your prediction now that you've seen large-sample, record it here.

8) Let's see how accurate your prediction is... *feel free to click "Run" and build a few more pie charts from your samples if you want to collect more information first!* When you're ready, make a pie-chart of more-animals.

- Which predictions were closest? _____
- Which predictions were off? _____
- Were there any surprises? _____

9) In the real world, we usually don't have access to a whole dataset to check predictions against! How could we test...

- *Every giraffe on the planet?*
- *Everyone who has ever come in contact with a covid-positive person?*
- *Every person who identifies as queer?*
- **What strategies can we use to make sure that predictions from samples are as close to accurate as possible?**

Choosing Your Dataset

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. Look up the [Contract](#) for the display and record the Pyret code you'd need to make it. If time allows, type your code into code.pyret.org (CPO) to see your display!

1) A pie-chart showing the species of animals from the shelter.

Which Rows?	Which Column(s)?	What will you Create?
<i>All the animals</i>		

code: _____

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

Which Rows?	Which Column(s)?	What will you Create?
<i>All the animals</i>		

code: _____

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

Which Rows?	Which Column(s)?	What will you Create?
<i>All the animals</i>		

code: _____

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

Which Rows?	Which Column(s)?	What will you Create?
<i>All the animals</i>		

code: _____

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?
<i>All the animals</i>		

code: _____

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?
<i>All the animals</i>		

code: _____

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
		<i>Lookup</i> <i>Arithmetic</i> <i>Statistical</i> <i>Can't Answer</i>
		<i>Lookup</i> <i>Arithmetic</i> <i>Statistical</i> <i>Can't Answer</i>
		<i>Lookup</i> <i>Arithmetic</i> <i>Statistical</i> <i>Can't Answer</i>
		<i>Lookup</i> <i>Arithmetic</i> <i>Statistical</i> <i>Can't Answer</i>
		<i>Lookup</i> <i>Arithmetic</i> <i>Statistical</i> <i>Can't Answer</i>
		<i>Lookup</i> <i>Arithmetic</i> <i>Statistical</i> <i>Can't Answer</i>

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

Choose two columns to describe below.

6) _____, which contains _____ data. Example values from this column include:

column name

categorical/quantitative





7) _____, which contains _____ data. Example values from this column include:





column name

categorical/quantitative

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.

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Datasets and Starter Files

Click through the datasets below. (Your teacher might also ask you to work with Global Food Supply [[Dataset](#)] [[Starter File](#)].) When you find one you'd like to use in Pyret, (1) click the "Starter File" link to open it in a new tab and (2) select "Save a copy" from the "File" menu.

★ Looking for a shorter list? We've starred a few good beginner datasets.

The Environment & Health

Global Waste by Country 2019	[Dataset Starter File]
World Cities' Proximity to the Ocean	[Dataset Starter File]
Earthquakes	[Dataset Starter File]
Air Quality, Pollution Sources & Health in the U.S.	[Dataset Starter File]
Health by U.S. County	[Dataset Starter File]
COVID in the U.S. by County	[Dataset Starter File]
Arctic Sea Ice	[Dataset Starter File]

Politics

Countries of the World	[Dataset Starter File]
Gerrymandering	[Dataset Starter File]
Marijuana Laws & Arrests by State 2018	[Dataset Starter File]
LAPD Arrests 2010-2019	[Dataset Starter File]
NYPD Stop, Search & Frisk 2019	[Dataset Starter File]
Refugees 2018	[Dataset Starter File]
State Demographics	[Dataset Starter File]
U.S. Income	[Dataset Starter File]
U.S. Jobs	[Dataset Starter File]
U.S. Voter Turnout 2016	[Dataset Starter File]

Sports

Esports Earnings	[Dataset Starter File]
MLB Hitting Stats	[Dataset Starter File]
NBA Players	[Dataset Starter File]
NFL Passing	[Dataset Starter File]
NFL Rushing	[Dataset Starter File]

Entertainment

★Movies	[Dataset Starter File]
IGN video game Reviews	[Dataset Starter File]
International Exhibition of Modern Art	[Dataset Starter File]
North American Pipe Organs	[Dataset Starter File]
Pokemon	[Dataset Starter File]
Music	[Dataset Starter File]

Education

College Majors	[Dataset Starter File]
----------------	--

U.S. Colleges 2019-2020 [[Dataset Starter File](#)]

★R.I. Schools [[Dataset Starter File](#)]

Evolution of College Admissions in California [[Dataset Starter File](#)]

Nutrition

Soda, Coffee & Other Drinks [[Dataset Starter File](#)]

Fast Food Nutrition [[Dataset Starter File](#)]

[Would you like to contribute a dataset of your own, or is there something you'd like to change about one of ours?](#)

Rubric: Exploration Project (1)

About this Dataset

Wow!	<input type="checkbox"/> Getting There	<input type="checkbox"/> Needs Improvement	<input type="checkbox"/>
I explained why this dataset is interesting to me, others like me, and why others should care about it. I considered why the dataset was collected, and what purpose it might serve. I correctly identified all rows, columns, and types in my dataset.	I explained why this dataset was interesting to me and at least one other person/group, and shared <i>something</i> about where it came from. I correctly identified most of the rows, columns, and types in my dataset.	I explained why this dataset was interesting to me, and shared <i>something</i> about where it came from. I correctly identified some rows, columns, and types in my dataset.	

Criteria for Displays

Wow!	Getting There	Needs Improvement
I either included multiple displays of this type or wrote about why I my data didn't allow for multiple. I indicated which column(s) I used and added the relevant code. I made a strong attempt to interpret the interesting displays and report about the displays that weren't useful. I added the questions that emerged to the "My Questions" section.	I included one display of this type. I provided the column name and relevant code. My interpretation lacked detail. I added the questions that emerged to the "My Questions" section.	I included one or no displays of this type. My slides may be missing a correct column name or code. My data interpretation may be missing or inaccurate. I may not have added to the "My Questions" section.
Displays	Rating	Teacher Feedback
Bar Chart	<input type="checkbox"/> Wow <input type="checkbox"/> Getting There <input type="checkbox"/> Needs Improvement	
Pie Chart	<input type="checkbox"/> Wow <input type="checkbox"/> Getting There <input type="checkbox"/> Needs Improvement	
Histogram	<input type="checkbox"/> Wow <input type="checkbox"/> Getting There <input type="checkbox"/> Needs Improvement	
Box Plot	<input type="checkbox"/> Wow <input type="checkbox"/> Getting There <input type="checkbox"/> Needs Improvement	

Rubric: Exploration Project (2)

Measures of Center

Wow!	Getting There	Needs Improvement
<input type="checkbox"/> I selected at least two columns in my dataset, and correctly filled out the entire summary table for each one (or wrote about why my data didn't allow for this). Based on these measures, I decided which measure of center was best for each column, and I provided a detailed interpretation of what these measures tell me about the dataset.	<input type="checkbox"/> I selected at least two columns in my dataset (or wrote about why my data didn't allow for this), and correctly filled out the entire summary table for each one. I tried to interpret what these measures tell me about the dataset, but my interpretation lacked detail.	<input type="checkbox"/> I filled out most of the table but didn't demonstrate understanding of what these measures tell about the dataset.

Correlation and Linear Regression

Wow!	Getting There	Needs Improvement
<input type="checkbox"/> I either included multiple scatterplots or wrote about why my data didn't allow for multiple. I described my observations, including identifying outliers and patterns that could point to possible correlations. If the scatter plot didn't reveal any patterns or outliers, I wrote about that. When the corresponding linear regression plot(s) showed a correlation, I included an additional slide and a thoughtful interpretation.	<input type="checkbox"/> I included at least one scatter plot with cursory descriptions and observations. I included a slide of a linear regression plot showing a correlation or described why I didn't include any linear regression plots.	<input type="checkbox"/> I added at least one slide about a scatter plot. The description and/or display may be lacking. I may have left out the linear regression, included one that didn't reveal a correlation, or offered an incorrect interpretation of it.

My Questions

Wow!	Getting There	Needs Improvement
<input type="checkbox"/> I had lots of questions by the end of the exploration, and I chose at least two that I thought were most interesting. I explained why I thought they were interesting, and wrote about grouped samples that might be good to explore when answering those questions.	<input type="checkbox"/> I had a few questions by the end of the exploration, and I chose at least one that was interesting. I wrote about grouped samples that might be good to explore.	<input type="checkbox"/> I picked a question, and wrote about grouped samples.

Additional Teacher Feedback

Histograms

To best understand histograms, it's helpful to contrast them first with bar charts.

Bar charts show the number of rows belonging to a given category. The more rows in each category, the taller the bar.

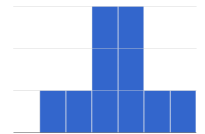
- Bar charts provide a visual representation of the frequency of values in a **categorical** column.
- There's no strict numerical way to order these bars.
 - The count of red, yellow and blue balloons would make sense no matter what order they get presented in.
 - But **sometimes there's an order that makes sense**. For example, it would be logical to show the count of t-shirt sizes in order of smallest to largest shirt.

Histograms show the number of rows that fall within certain intervals, or "bins", on a horizontal axis. The more rows that fall within a particular "bin", the taller the bar.

- *Histograms provide a visual representation of the frequencies (or relative frequencies) of values in a **quantitative** column.*
- Quantitative data **can always be ordered**, so the bars of a histogram always progress from smallest (on the left) to largest (on the right).
- When dealing with histograms, it's important to select a good **bin size**. If the bins are too small or too large, it is difficult to see the shape of the dataset. Choosing a good bin size can take some trial and error!

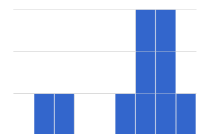
The **shape** of a dataset tells us which values are more or less common.

- In a **symmetric** dataset, values are just as likely to occur a certain distance above the mean as below the mean. Each side of a symmetric distribution looks almost like a mirror-image of the other.

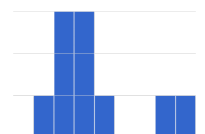


- Some extreme values may be far greater or far lower than the other values in a dataset. These extreme values are called **outliers**.

- A dataset that is **skewed left** has a few values that are unusually low. The histogram for a skewed left dataset has a few data points that are stretched out to the left (lower) end of the x-axis.

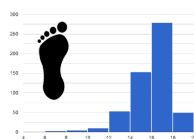


- A dataset that is **skewed right** has a few values that are unusually high. The histogram for a skewed right dataset has a few data points that are stretched out to the right (higher) end of the x-axis.

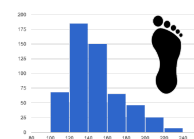


- One way to visualize the difference between a histogram of data that is **skewed left** or **skewed right** is to think about the lengths of our toes on our left and right feet.

Much like the bar lengths of a histogram that is "skewed left", our left feet have smaller toes on the left and a bigger toe on the right.



Our right feet have the big toe on the left and smaller toes on the right, more closely resembling the shape of a histogram of "skewed right" data.

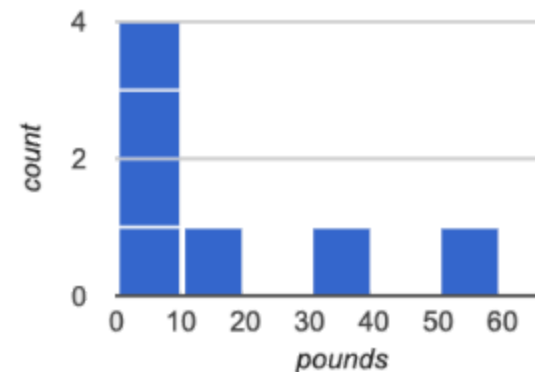
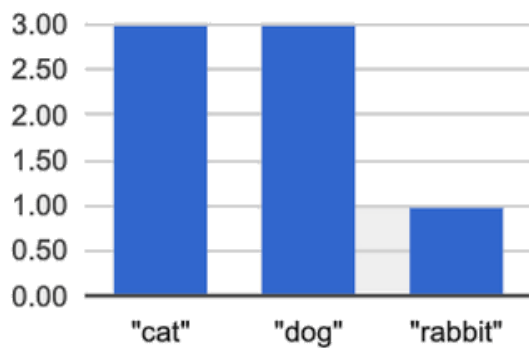


Summarizing Columns with Bar Charts & Histograms

name	species	age	pounds
"Sasha"	"cat"	1	6.5
"Boo-boo"	"dog"	11	12.3
"Felix"	"cat"	16	9.2
"Nori"	"dog"	6	35.3
"Wade"	"cat"	1	3.2
"Nibblet"	"rabbit"	6	4.3
"Maple"	"dog"	3	51.6

1	How many cats are there in the table above?	
2	How many dogs are there?	
3	How many animals weigh between 0 and 20 pounds?	
4	How many animals weigh between 20 and 40 pounds?	
5	Are there more animals weighing 40-60 pounds than 60-140 pounds?	

The two displays below both summarize this table. The display on the left is a **Bar Chart**, while the one on the right is a **Histogram**. What is similar about them? What is different?



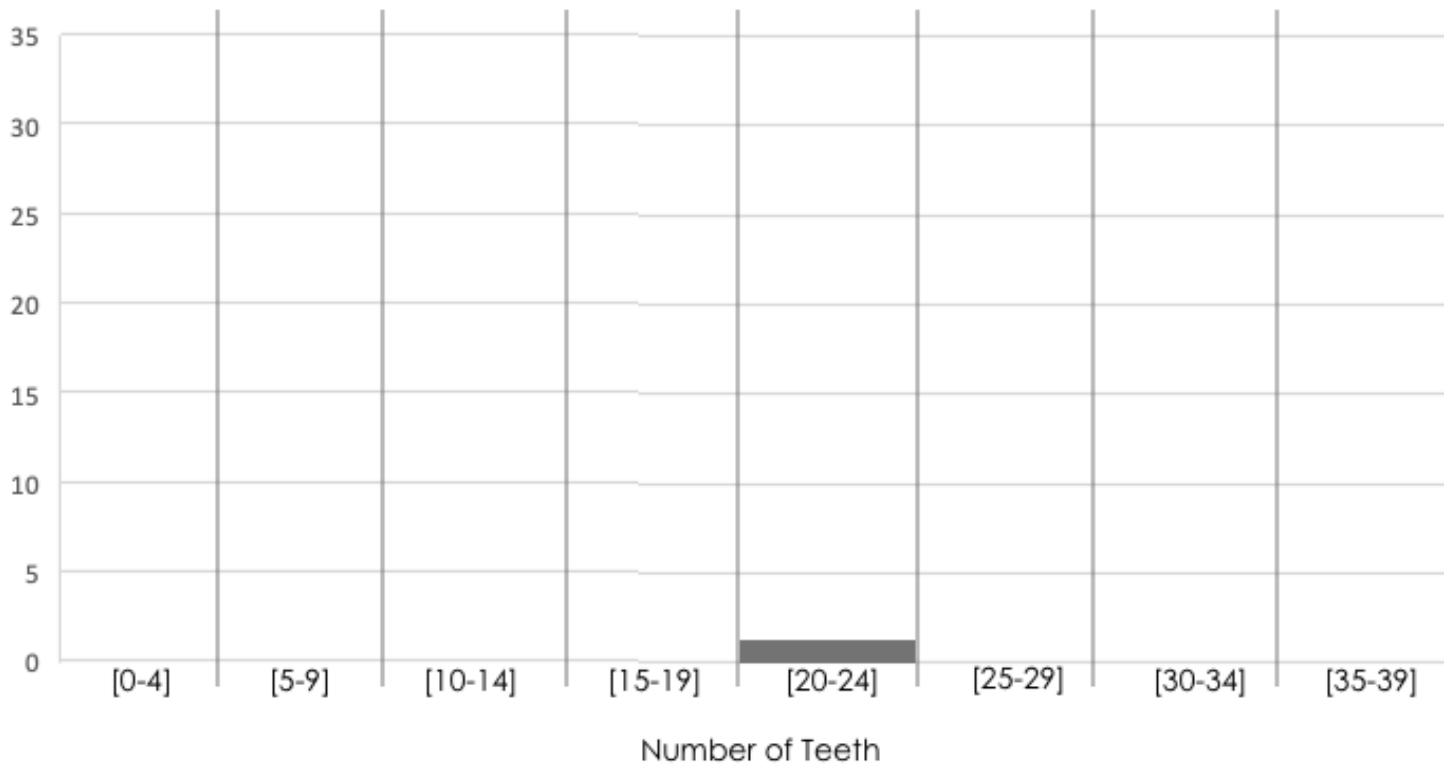
Similarities	Differences

Making Histograms

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

Number of teeth	Count
0	5
22	1
26	1
27	1
28	4
29	3
30	5
31	3
32	27

Draw a histogram for the table in the space below. For each row, find which interval (or “bin”) on the x-axis represents the right number of teeth. Then fill in the box so that its height is equal to the *sum of the counts* that fit into that interval. One of the intervals has been completed for you.



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).

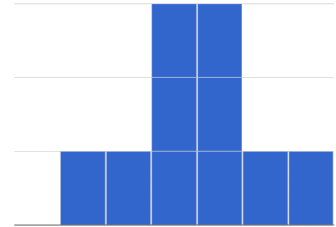
Match the summary description (left) with the *shape* of the histogram of student ratings (right).

- 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.

1 Most of the students were fine with the video, but a couple of them gave it an unusually low rating.

1

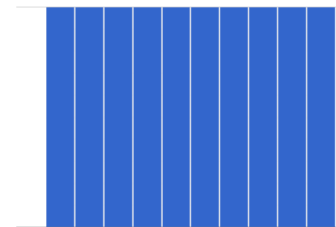
A



2 Most of the students were okay with the video, but a couple students gave it an unusually high rating.

2

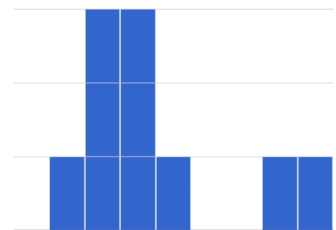
B



3 Students tended to give the video an average rating, and they weren't likely to stray far from the average.

3

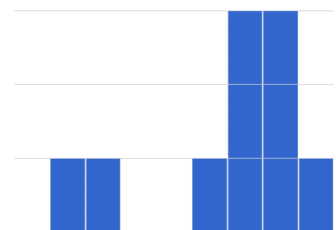
C



4 Students either really liked or really disliked the video.

4

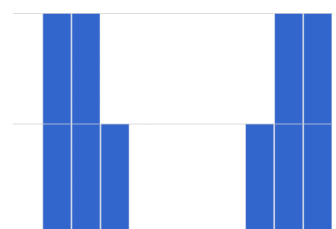
D



5 Reactions to the video were all over the place: high ratings and low ratings and inbetween ratings were all equally likely.

5

E



Choosing the Right Bin Size

Open your saved [Animals Starter File](#), or make a new copy, and click "Run".

```
# histogram :: ( Table , String , String , Number ) -> Image
               table-name labels column-name bin-size
```

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 the best picture of the distribution? (Note: *ideally your histogram should have between 5 and 10 bars*) _____
- 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? _____
-
-

Data Cycle: Shape of the Animals Dataset



Use the Data Cycle to explore the distribution of one or more quantitative columns in [Animals Starter File](#) using histograms.





<p>Ask Questions</p> 	<p>What is the shape of the age column of the Animals dataset? What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>The histogram I created is for _____ x-variable in context _____ from _____ dataset or subset _____.</p> <p>The bin size I chose is _____ bin size _____, which resulted in a histogram with _____ bins. I chose this bin size because _____</p> <p>_____</p> <p>I would describe the shape of this histogram as _____</p> <p>I notice that _____ Consider statements like: Most of the histogram's area is... / A small amount of the histograms area trails out... / etc</p> <p>I wonder _____</p>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>The histogram I created is for _____ x-variable in context _____ from _____ dataset or subset _____.</p> <p>The bin size I chose is _____ bin size _____, which resulted in a histogram with _____ bins. I chose this bin size because _____</p> <p>_____</p> <p>I would describe the shape of this histogram as _____</p> <p>I notice that _____ Consider statements like: Most of the histogram's area is... / A small amount of the histograms area trails out... / etc</p> <p>I wonder _____</p>	

Data Cycle: Shape of My Dataset

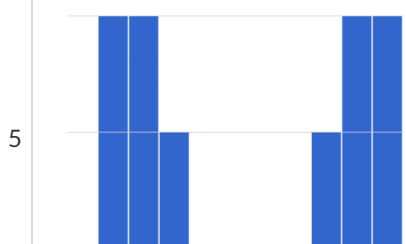
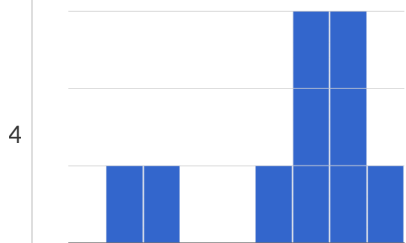
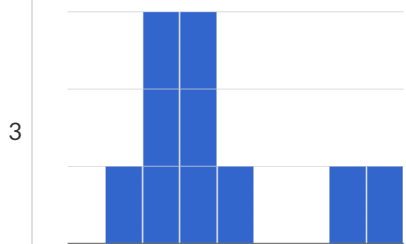
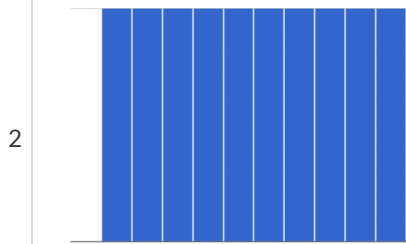
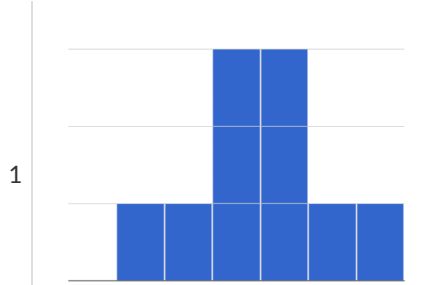
Use the Data Cycle to explore the distribution of one or more quantitative columns from [your chosen dataset](#) using **histograms**, and write down your findings.

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Identifying Shape - Histograms





Describe the shape of the histograms on the left. Do your best to incorporate the vocabulary you've been introduced to.







Data Cycle: Shape of the Animals Dataset

Describe two **histograms** made from columns of the animals dataset.

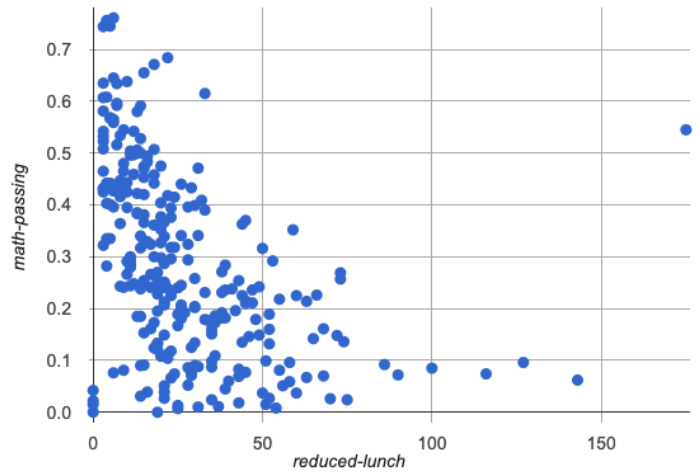
The first question is provided. You'll need to come up with the second question on your own!

<p>Ask Questions</p> 	<p><i>What is the distribution of weight among all animals at the shelter?</i> What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one): Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>The histogram I created is for _____ x-variable in context _____ from _____ dataset or subset _____.</p> <p>The shape of this histogram is _____. There are peaks at _____ and gaps at _____. <small>skewed left, skewed right, symmetric</small></p> <p>I notice that _____ <small>Consider statements like: Most of the histogram's area is... / A small amount of the histograms area trails out... / etc</small></p> <hr/> <p>I wonder _____</p> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one): Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>The histogram I created is for _____ x-variable in context _____ from _____ dataset or subset _____.</p> <p>The shape of this histogram is _____. There are peaks at _____ and gaps at _____. <small>skewed left, skewed right, symmetric</small></p> <p>I notice that _____ <small>Consider statements like: Most of the histogram's area is... / A small amount of the histograms area trails out... / etc</small></p> <hr/> <p>I wonder _____</p> <hr/>	

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 <i>has</i> 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?

Measures of Center

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

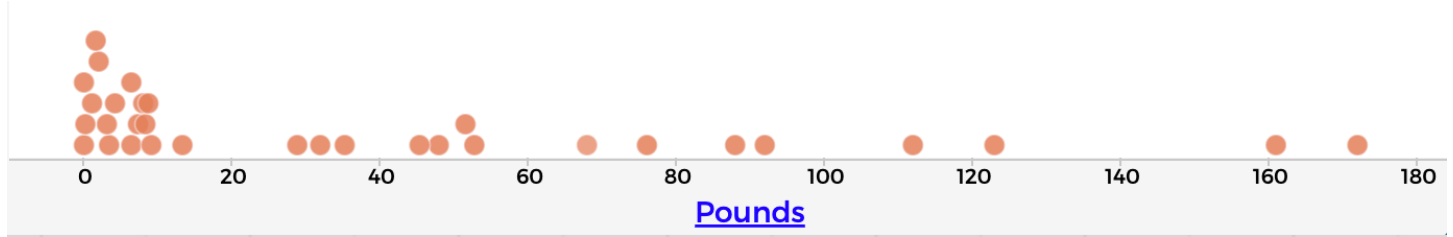
- Each of these measures of center summarizes a whole column of quantitative data using just one number:
 - The **mean** of a dataset is the average of all the numbers.
 - The **median** of a dataset is a value that is smaller than half the dataset, and larger than the other half. In an ordered list the median will either be the middle number or the average of the two middle numbers.
 - The **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.
 - A dataset with **left skew** has a few values that are unusually low, which pull the mean *below* the median.
 - A dataset with **right skew** has 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 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*!)

What Value is Typical?

If we plotted all 32 animals' weights as points on a number line, it would look something like this:



1) What do you Notice?

2) What do you Wonder?

3) What do you think is a typical value in this sample? Why?

4) Identify another value someone might claim is typical in this sample. Why would they choose that value?

5) Do you think there is a midpoint of this sample? Why or why not?

6) Do you think there is a value that's repeated more than any other value? Why or why not?

Summarizing Columns with Measures of Center

Summarizing the Pounds Column

Find the measures of center to summarize the _____ pounds _____ column of the [Animals Starter File](#).

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

Mean (Average)	Median	Mode(s)
mean(animals-table, "pounds")	median(animals-table, "pounds")	modes(animals-table, "pounds")

2) 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?

3) The mean is _____ the median, which suggests the shape is _____.

higher than/lower than/about equal to
skewed right (high outliers) / skewed left (low outliers) / symmetric

4) Which do you think is the most useful measure for this column of data? Why? _____

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

Summarizing the _____ Column

Find the measures of center to summarize the _____ column of the [Animals Starter File](#).

a column of your choosing!

The three measures of center for this column are:

Mean (Average)	Median	Mode(s)

The mean is _____ the median, which suggests the shape is _____.

higher than/lower than/about equal to
skewed right (high outliers) / skewed left (low outliers) / symmetric

★ Four animals weighing 5, 5, 10, and 100 pounds will have an average mean of 30 pounds.

(because $5 + 5 + 10 + 100 = 120$ and $120 \div 4 = 30$)

Can you think of another set of four animals that would have the same average? How many sets can you come up with?

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 235 lbs."	
"With a median of 135, that means that half the people in this group can't even lift 135 lbs."	

Data Cycle Practice

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


<p>Ask Questions</p> 	<p><i>What is the mean age for cats at the shelter?</i></p> <p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p><i>What is the median time it takes for an animal to be adopted?</i></p> <p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Data Cycle Practice

Open [your chosen dataset](#). Complete both of the Data Cycles shown here.

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Mean, Median, Mode(s) Practice

Mean

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

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)

Find the mode(s) of each dataset:

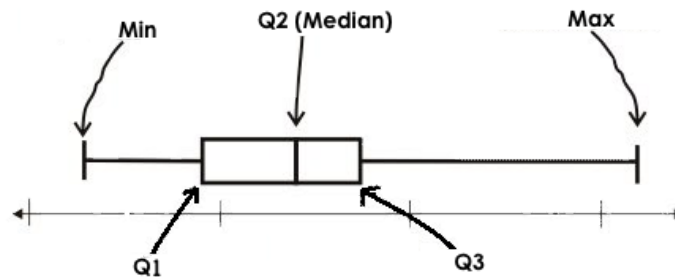
17, 23, 25, 23, 22	5, 11, 3, 7, 4	11, 3, 7, 4	5, 7, 11, 11, 7, 7	2, 3, 5, 4, 3, 7, 4

Measures of Spread

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.

Summarizing Columns with Measures of Spread

Summarizing the Pounds Column

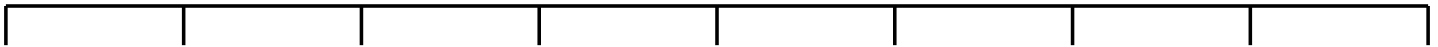
Get the values to summarize the spread of the _____ pounds _____ column of the [Animals Starter File](#) by typing

`box-plot(animals-table, "pounds")` into the Interactions Area.

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.*

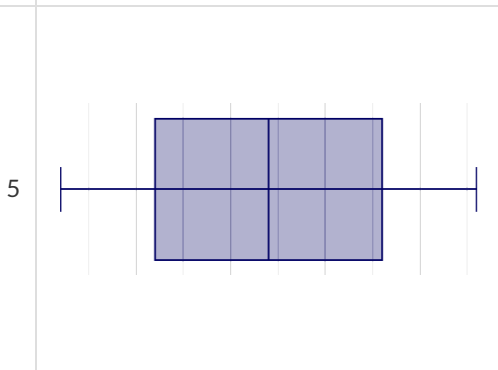
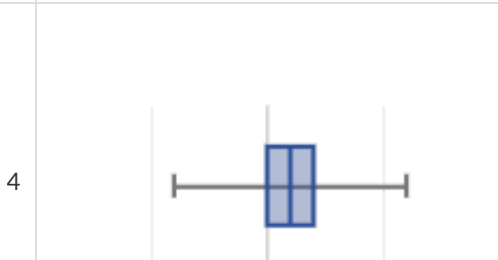
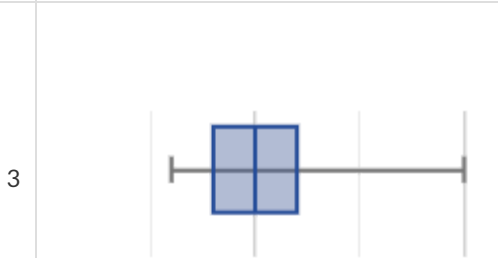
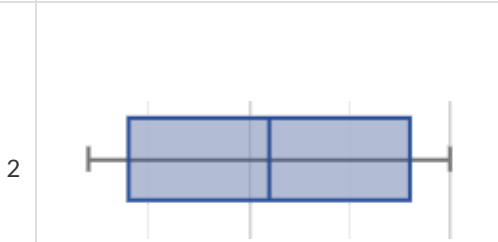
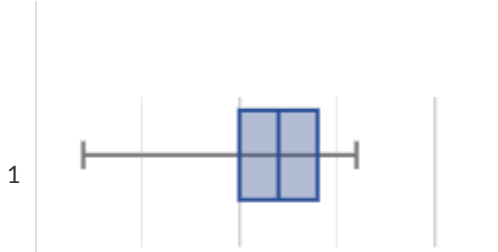


7) The **Range** is: _____ and the **Interquartile Range(IQR)** is: _____.

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

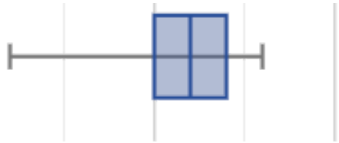
Identifying Shape - Box Plots

Describe the shape of the box plots on the left. Do your best to incorporate the vocabulary you've been introduced to.



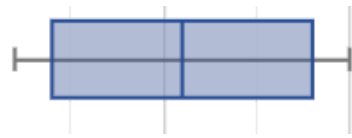
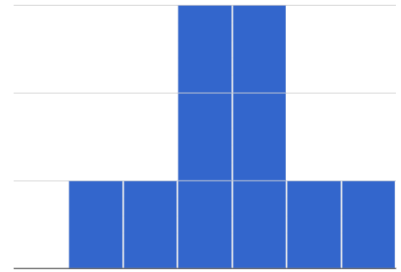
Matching Box Plots to Histograms

Students watched 5 videos, and rated them on a scale of 1 to 10. For each video, their ratings were used to generate box plots and histograms. Match each box plot to the histogram that displays the same data.



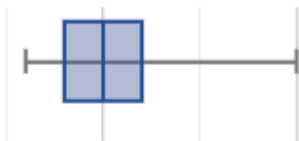
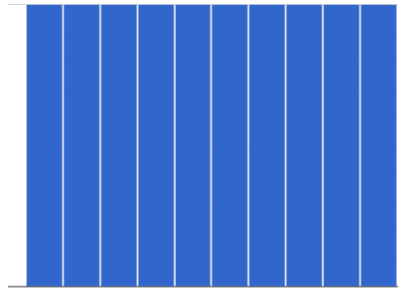
1

A



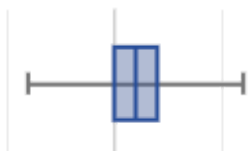
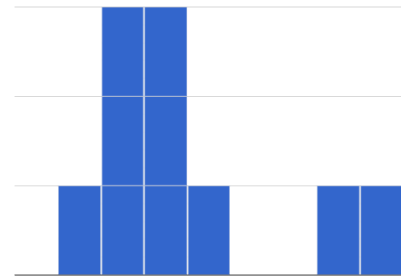
2

B



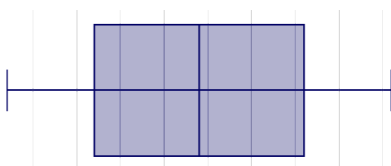
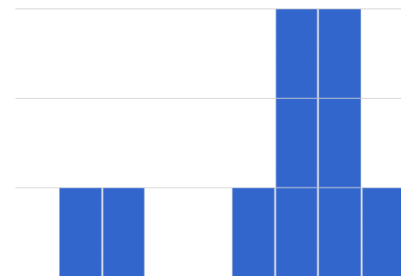
3

C



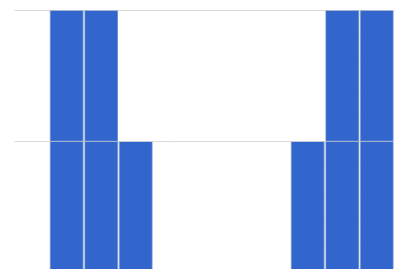
4

D



5

E



Directions: Connect each item on this page to at least one other item by drawing an arrow and writing an explanation of how they are connected along the arrow. (Arrows may curve.)

Minimum

Maximum

Quartile

Median

50%

Interquartile Range





Upper Quartile





Lower Quartile

25%

Data Cycle: Shape of the Animals Dataset

Open the [Animals Starter File](#). Use the Data Cycle to explore the distribution of one or more quantitative columns using **box plots**.



<p>Ask Questions</p> 	<p>What is the distribution of the weeks column from the animals dataset? What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one): Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>The box plot for _____ x-variable in context _____ is _____ skewed left / skewed right / symmetric / etc.</p> <p>The 5-number summary is: min = _____ Q1 = _____ median = _____ Q3 = _____ max = _____</p> <p>The middle 50% of the data lies between _____ and _____ so the Interquartile Range is _____</p> <p>I notice that _____ Consider statements like: 75% of the data fall below ... / The top 25% of the data fall between ... / etc</p> <p>_____</p> <p>I wonder _____</p>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <p>_____</p> <p>_____</p>	<p>Question Type (circle one): Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <p>_____</p> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <p>_____</p>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <p>_____</p>	
<p>Interpret Data</p> 	<p>The box plot for _____ x-variable in context _____ is _____ skewed left / skewed right / symmetric / etc.</p> <p>The 5-number summary is: min = _____ Q1 = _____ median = _____ Q3 = _____ max = _____</p> <p>The middle 50% of the data lies between _____ and _____ so the Interquartile Range is _____</p> <p>I notice that _____ Consider statements like: 75% of the data fall below ... / The top 25% of the data fall between ... / etc</p> <p>_____</p> <p>I wonder _____</p>	

Data Cycle: Shape of My Dataset

Open [your chosen dataset](#). Use the Data Cycle to explore the distribution of one or more quantitative columns using **box plots**, and write down your findings.

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

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

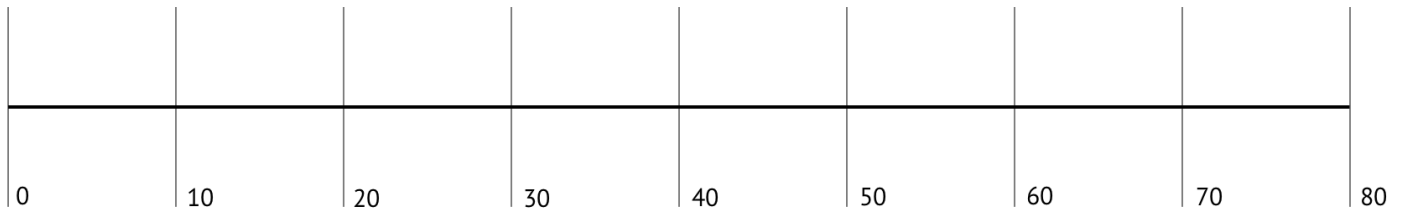
2) Order the Ages from Least to Greatest: _____

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

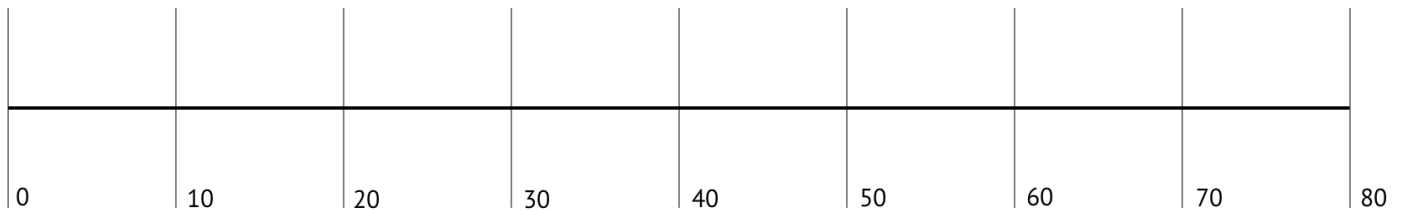
Box Plots - Visualizing Shape

Make box plots to 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? _____

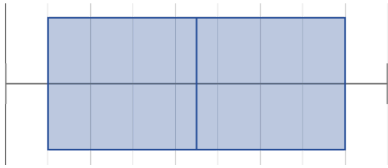
6) What else do you Notice and Wonder about the data from these two family gatherings?

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

Reading Box Plots

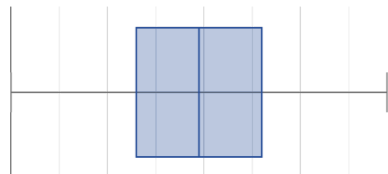
A class of students took five different exams this year, and each distribution of their scores has been plotted in one of the five box plots below.

Match the summary description (left) with the *shape* of the box plot of student scores (right).



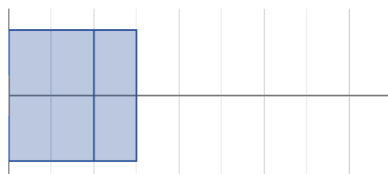
1

A Most students did pretty well on this exam, but there were some mediocre scores and a handful of very low scores.



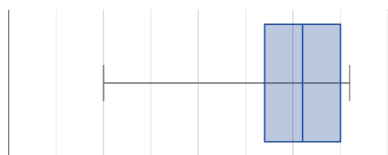
2

B This exam featured one question worth a lot of points that many of the students got completely right, while many others got it completely wrong. Nobody actually got the "average" score.



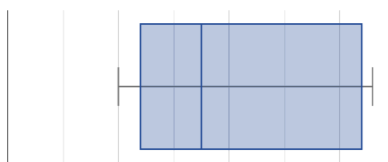
3

C A lot of students did poorly on this exam. Relatively few did just OK. Still, a bunch of students who really knew what they were doing completely aced it.



4

D Performance on this exam resulted in a classic "bell curve" shape: most students performed close to the average and scores far from the average in either direction were increasingly unlikely.

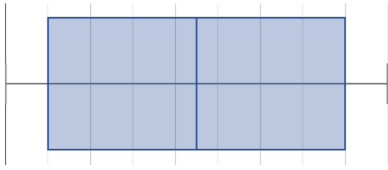


5

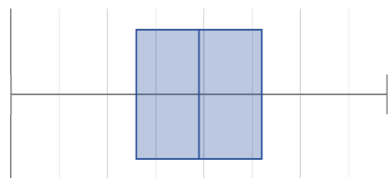
E This was a hard exam. Most students did poorly, with scores tapering to the point where hardly anyone got an A.

Matching Box Plots to Histograms 2

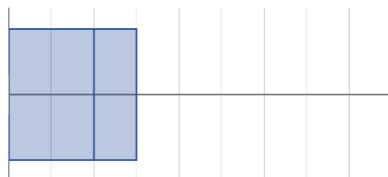
Match each box-plot to the histogram that displays the same data.



1



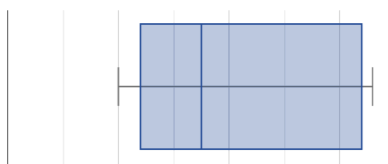
2



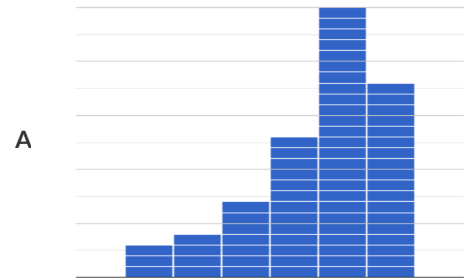
3



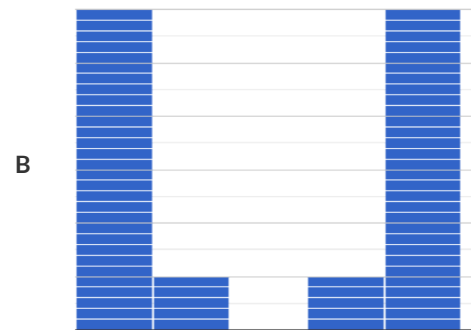
4



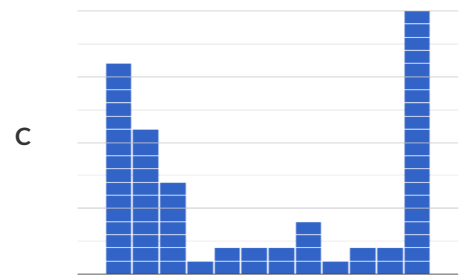
5



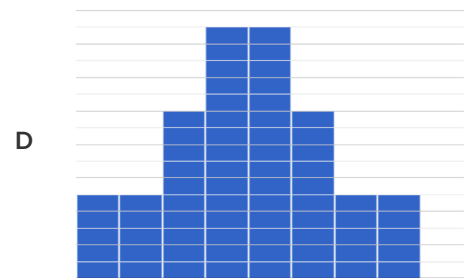
A



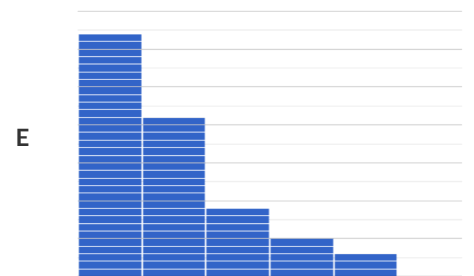
B



C



D



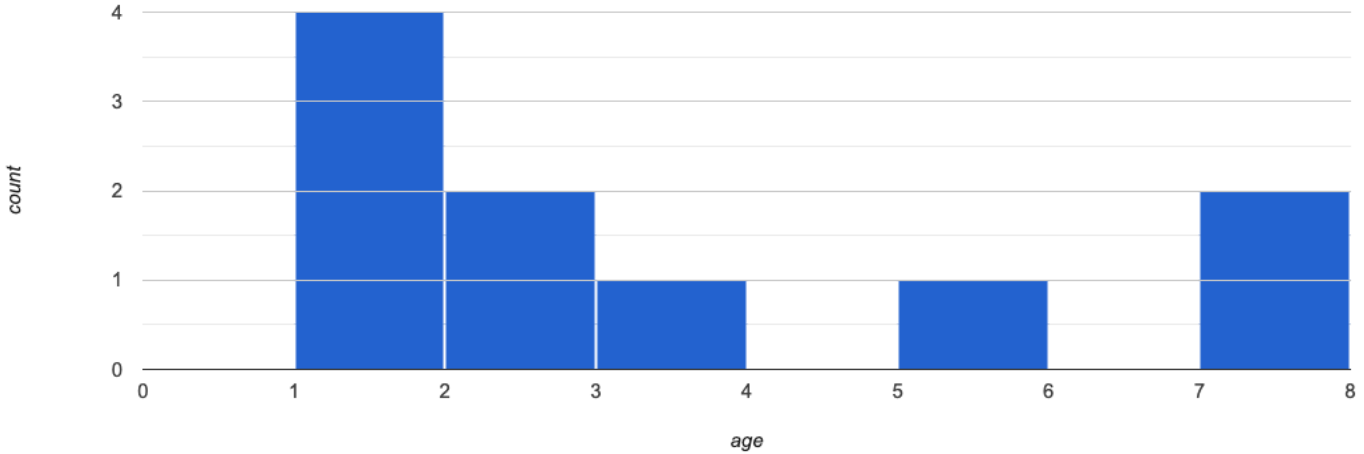
E

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
count	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						

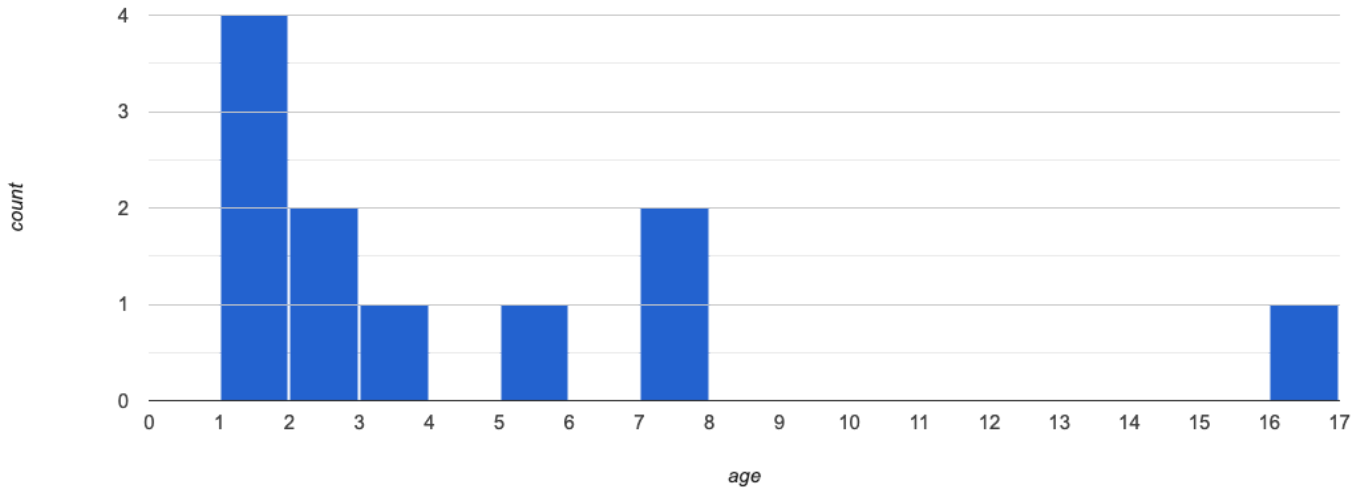
8) Add all the squared distances. What is their sum? _____

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
count	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. (For example, 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: Standard Deviation in the Animals Dataset

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.



<p>Ask Questions</p> 	<p><i>Do the animals all get adopted in around the same length of time?</i></p> <p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/> <hr/>	

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: Standard Deviation in 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.

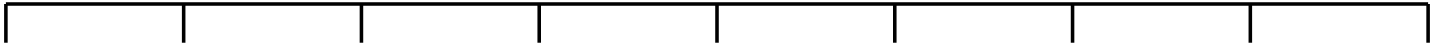
<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Computing Standard Deviation (2)

Here are ten different family incomes: \$43k, \$62k, \$39k, \$141k, \$58k, \$82k, \$41k, \$73k, \$68k, \$73k

1) Draw the **distribution** of these incomes by placing a dot on the number line below. If two families have the same income, put one dot on top of the other. Finally, draw a **box plot** on the number line, making sure to label the axis and show each quartile.



2) Describe the shape of this box-plot. _____

3) What is the mean income of the families in this dataset? _____

4) How many families earn \$39k? \$43k? Fill in the table below. The first column has been done for you.

income	\$39k	\$41k	\$43k	\$58k	\$62k	\$68k	\$73k
\$82k	\$141k	count	1				

5) Draw a star to locate the mean on the number line above.

6) For each family on the number line you drew,

- Draw a **horizontal arrow** under the axis from the star you drew in #5 to the dot for that family's income
- **Label the arrow with its distance from the mean.**
e.g. if the mean is \$50k and a family's income is \$82k, your arrow would stretch from \$50k to \$82k, and be labeled with the distance "\$32k"

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

7) For each of the 10 incomes in the table below, list the distance-from-mean for each income, using the mean you computed above. Then fill in the squared distance in the next row to complete the table.

income (in 10s of thousands)	39	41	43	58	62	68	73	73	82	141
distance from mean										
squared distance										

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!** _____

Matching Mean & Standard Deviation to Data

In the table below, match the mean and standard deviation to the list of data it describes.

Mean: 4 StDev: 0	1	A	-1, -2, -3, -4, -5, -6, -7
Mean: -5 StDev: ~5.66	2	B	1, 2, 3, 4, 5, 6, 7
Mean: 4 StDev: ~2.16	3	C	-1, -9
Mean: 4 StDev: ~2.65	4	D	0, 2, 3, 4, 5, 6, 8
Mean: -4 StDev: ~2.16	5	E	4, 4, 4, 4, 4

Correlations in Scatter Plots

Scatter Plots can be used to show a relationship between two quantitative columns.

Each row in the dataset is represented by a point, with one column providing the x-value and the other providing the y-value. 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.

Line of Best Fit

We graphically summarize a relationship by drawing a straight line through the data cloud, so that the vertical distance between the line and all the points taken together is as small as possible. This allows us to predict y-values (the **response variable**) based on x-values (the **explanatory variable**).

- *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.

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

r-value

We can summarize the **correlation** between two quantitative columns in a single number.

- The *r*-value will always fall between -1 and $+1$.
- 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.

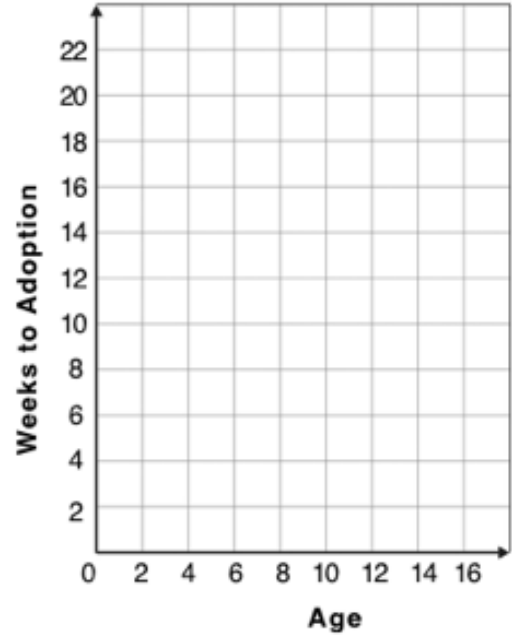
Note: These cutoffs are not an exact science! In some contexts an *r*-value of ± 0.50 might be considered impressively strong!

Correlation is not causation! Correlation only suggests that two column variables are related, but 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!

Creating a Scatter Plot

1) The table below has some new animals!
 Choose one and (*paying careful attention to how the axes are labelled*)
 plot their age/weeks values by adding a dot to the scatter plot on the right.
 Then write the animal's name next to the dot you made.

name	species	age	weeks
"Alice"	"cat"	1	3
"Bob"	"dog"	11	5
"Callie"	"cat"	16	4
"Diver"	"lizard"	2	24
"Eddie"	"dog"	6	9
"Fuzzy"	"cat"	1	2
"Gary"	"rabbit"	6	12
"Hazel"	"dog"	3	2



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.

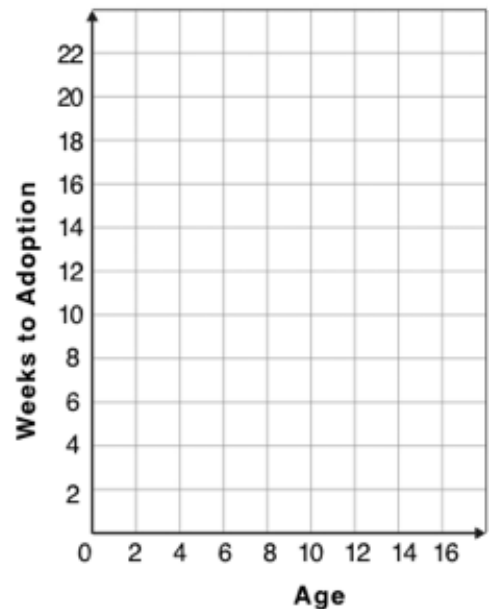
3) After how many animals did you begin to see a pattern? _____

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) Place points on the graph to create a scatter plot with NO relationship.



Exploring Relationships Between Columns

This page is designed to be used with the [Animals Starter File](#). Log into [code.pyret.org\(CPO\)](http://code.pyret.org(CPO)) 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 an animal weighs related to its age?

- What would you expect? _____

- What did you learn from your scatter plot? _____

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

- What would you expect? _____

- What did you learn from your scatter plot? _____

3) How are the number of legs an animal has related to its age?

- 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: Relationships in the Animals Dataset

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?*

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

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Linear Regression

- **We compute linear relationships to predict the future!** Well...sort of. Given a dataset, like ages of animals v. how long before they're adopted, we try to compute the relationship between age and weeks so that we can *predict* how long a new animal might stay, based on their age.
- When we compute linear relationships, we're talking about **straight-line patterns** that appear on a scatter plot.
- A scatter plot has an x-axis and a y-axis. When looking for relationships, the y-axis is called the **response variable**, and the x-axis is called the **explanatory variable**. In our example, we are trying to figure out how much of the weeks variable is *explained by* the age variable.
- **Linear Regression** is a way of computing the **line of best fit**, which tries to draw a line as close as possible to all the points. (Want details? It minimizes the *sum of the squares* of the vertical distances from the points to the line. There's a reason we use computers to do this!)
- **Slope** is how much we predict the **response variable** will increase or decrease for each unit that the **explanatory variable** increases. In our example, a slope of 0.5 would mean "we predict that each additional year of age means an extra half-week in the shelter". (What would a slope of 3 mean?)
- **Sample size matters!** The number of data values is also relevant. 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.

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? _____

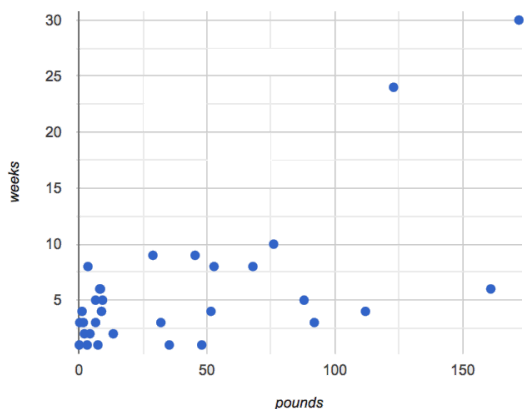
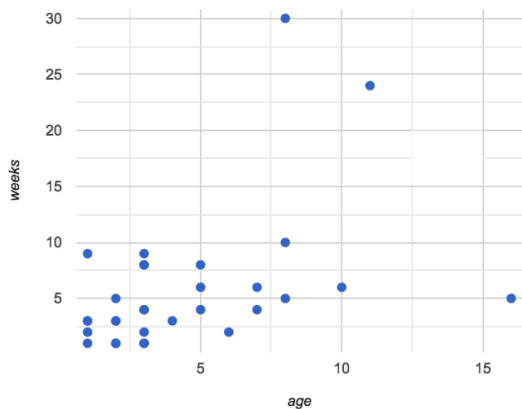
- 5) What is the highest r -value you can get? _____ Where did you place P? (_____, _____)

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

- 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? $y = \text{_____} x + \text{_____}$

Predictions from Scatter Plots



- 9) Draw the line of best fit for age-v-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) Draw the line of best fit for pounds-v-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?

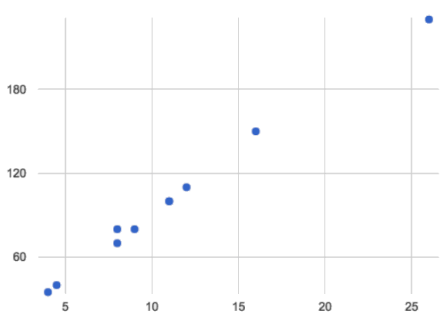
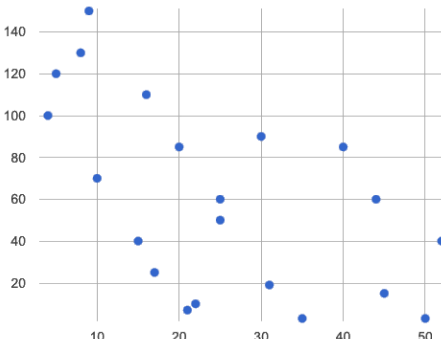
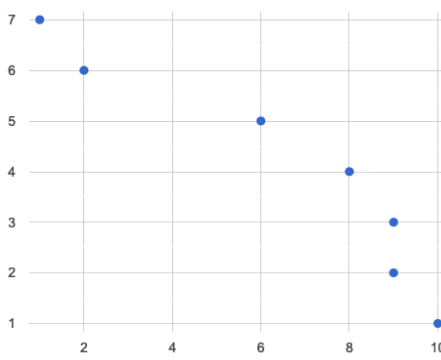
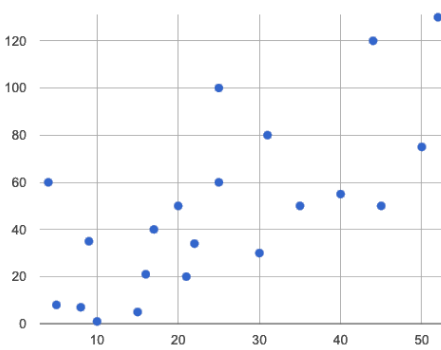
- 11) Do either or both of the relationships appear to be linear?

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		<p>Direction: Positive Negative None</p> <p>Strength: Stronger Weaker</p> <p>estimated r: -1 -0.5 0 0.5 1</p>
B		<p>Direction: Positive Negative None</p> <p>Strength: Stronger Weaker</p> <p>estimated r: -1 -0.5 0 0.5 1</p>
C		<p>Direction: Positive Negative None</p> <p>Strength: Stronger Weaker</p> <p>estimated r: -1 -0.5 0 0.5 1</p>
D		<p>Direction: Positive Negative None</p> <p>Strength: Stronger Weaker</p> <p>estimated r: -1 -0.5 0 0.5 1</p>

Exploring lr-plot

age

You should already have plotted `lr-plot(animals-table, "name", "age", "weeks")` in the [Animals Starter File](#).

- 1) What is the predictor function? $y = \underline{\hspace{2cm}} x + \underline{\hspace{2cm}}$
- 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 lr-plot, 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

Make another lr-plot, comparing the age v. weeks columns for **only the cats** using the following code:

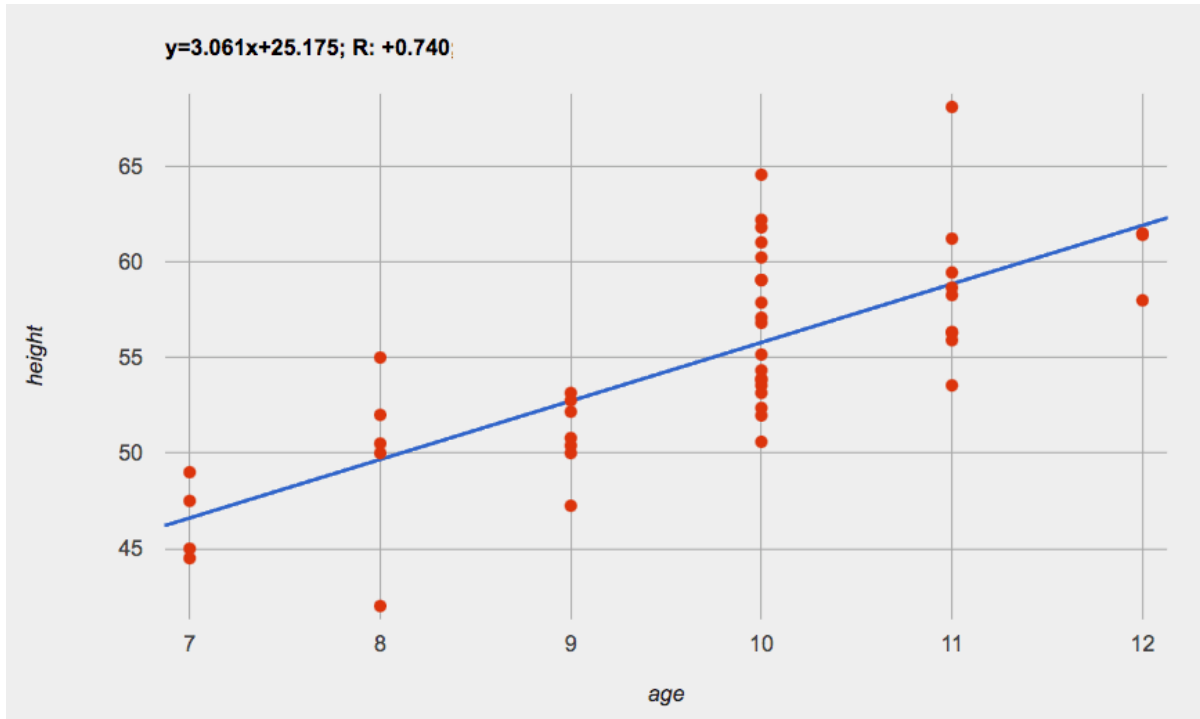
```
fun is-cat(r): r["species"] == "cat" end
lr-plot(filter(animals-table, is-cat), "name", "age", "weeks")
```

- 9) What is the predictor function? $y = \underline{\hspace{2cm}} x + \underline{\hspace{2cm}}$
- 10) What is the slope? _____
- 11) What is the y-intercept? _____
- 12) How does this line of best fit for *cats* compare to the line of best fit for *all animals*? _____

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

★ Make another lr-plot, 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.

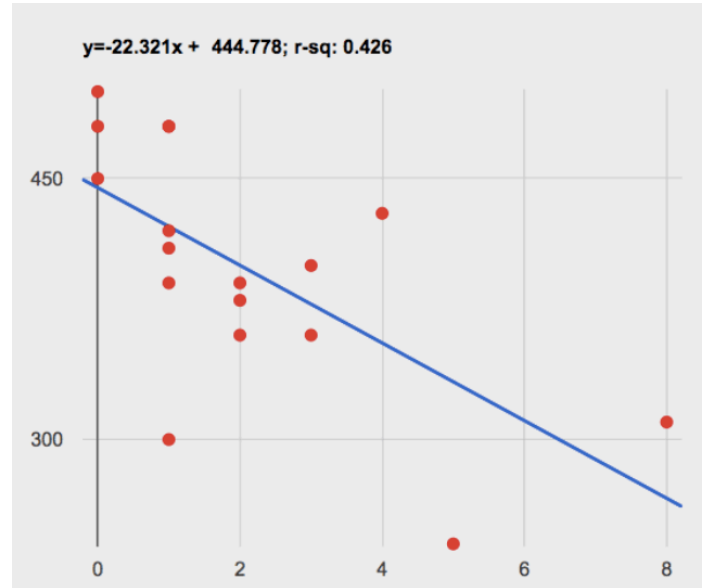
1	$\text{sugar}(m) = -3.19m + 12$ $r = -0.05$	<p>For every additional Marvel Universe movie released each year, the average person is predicted to consume _____ pounds of sugar! This correlation is _____.</p>
2	$\text{height}(s) = 1.65s + 52$ $r = 0.89$	<p>Shoe size and height are _____, _____ correlated. If person A is one size bigger than person B, we predict that they will be roughly _____ inches taller than person B as well.</p>
3	$\text{babies}(u) = 0.012u + 7.8$ $r = 0.01$	<p>There is _____ relationship found between the number of Uber drivers in a city and the number of babies born each year.</p>
4	$\text{score}(w) = -15.3w + 1150$ $r = -0.65$	<p>The correlation between weeks-of-school-missed and SAT score is _____ and _____. For every week a student misses, we predict a _____ point _____ in their SAT score.</p>
5	$\text{weight}(n) = 1.6n + 160$ $r = 0.12$	<p>There is a _____, _____ correlation between the number of streaming video services someone has, and how much they weigh. For each service, we expect them to be roughly _____ pounds heavier.</p>

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

Open [your chosen dataset](#). Ask a question about your data to tell your Data Story.

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>I performed a linear regression on a sample of _____ [dataset or subset] _____ and found a _____ 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] _____.</p>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>I performed a linear regression on a sample of _____ [dataset or subset] _____ and found a _____ 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] _____.</p>	

Age vs. Height Explore

Open the [Age vs. Height Starter File](#) and click "Run" to interact with data from another sample of students.

1) Take a look at the code in the Definitions Area. What do you notice? What do you wonder?

2) Build `image-scatter-plot(h-table, "age", "height", dot)`. Try to visualize the line of best fit for just the blue dots. Then try to visualize the line of best fit for just the red stars. How do you think they would compare? Which line do you think would be steeper?

3) Make three linear regression plots comparing `age` and `height`, and record the results for each in the table below:

- The whole population: `lr-plot(h-table, "gender-id", "age", "height")`
- Females only: `lr-plot(filter(h-table, is-f), "gender-id", "age", "height")`
- Males only: `lr-plot(filter(h-table, is-m), "gender-id", "age", "height")`

Sample	rate of change	y-intercept	R value
All			
Females			
Males			

4) What makes it difficult to compare these plots visually?

Rebuild `lr-plot(filter(h-table, is-f), "gender-id", "age", "height")`, adjust the window of the interactive plot using the numbers in the table below, and click `Redraw`.

x-min:	x-max:	y-min:	y-max:
6.5	12.5	45	70

Then, do the same for `lr-plot(filter(h-table, is-m), "gender-id", "age", "height")`.

5) How do the plots compare now that their windows match?

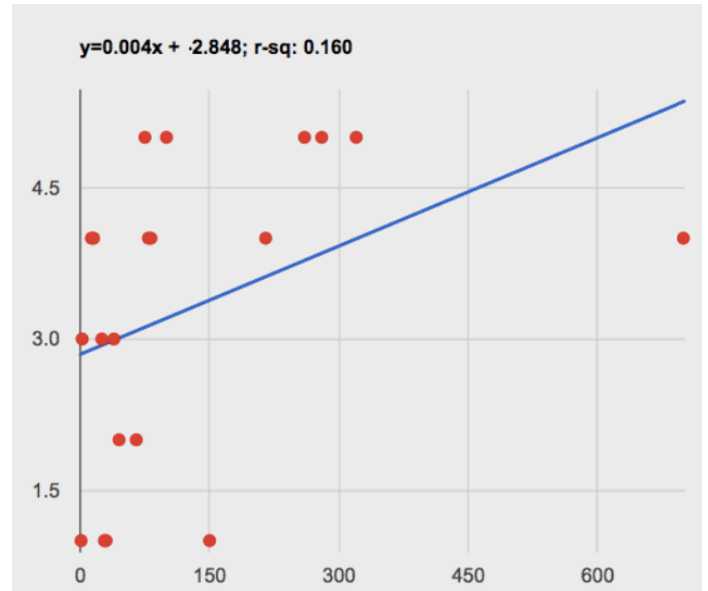
6) What happens if you compare the students' height in inches to their height in centimeters by plotting `lr-plot(h-table, "gender-id", "height-cm", "height")`?

Describing Relationships (2)

A small sample of people were surveyed about their satisfaction with their most recent purchase using a scale from 1 (very unsatisfied) to 5 (extremely satisfied).

NOTE: this data is made up for instructional purposes!

Dollars	Satisfaction
15.5	4
280	5
0.99	1
2.3	3
39	3
82	4
215	4
700	4
25	3
79	4
99.99	5
30	1
75	5
13	4
320	5
260	5
150	1
28	1
45	2
65	2



Describe the relationship between dollars spent and satisfaction shown in the data above.

Data Cycle: Regression Analysis 2

Open [your chosen dataset](#). Ask a question about your data to tell your Data Story.

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Write your Data Story below:

I performed a linear regression on a sample of _____ dataset or subset and found _____ correlation between _____ a weak/strong/moderate (R=...), positive/negative _____ and _____ [x-axis] _____ [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 lesson](#).

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...

Collecting Data

"In a survey of three hundred thousand people, the average height was less than four feet tall"

Politicians pass laws, shoppers choose brands, and countries go to war based on studies that sounds reliable. But is everything that *seems* reliable actually reliable? **Can we really trust these studies?**

There are many ways for a study to be flawed. Some flaws sneak in by accident, and data scientists have an obligation to look for these flaws and minimize them.

- A survey of people's favorite restaurants will be flawed, if it's only given to vegetarians.
- Some people might not fill out a survey that requires them to share their religion. This might change the results of the survey!
- A survey that lets people write whatever they want for "sex" might get some answers that are left blank, misspelled, or answers that aren't really about sex. Removing these responses from the dataset might change the results of the survey - especially if a certain group is more likely to leave it blank.

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

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 pie-chart of the dessert column and was surprised to discover that **Fruit** was the most popular dessert among 8th graders! Make the pie-chart. 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) Shani made a bar-chart of the gender-id column. In her analysis she stated that the most common gender identity among eighth graders in her class is male. Make the bar-chart. Do you agree? Why or Why Not?

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

5) 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?

Dirty Data!

Open the [New Animals Dataset](#) 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!

- Missing Data** - A column containing some cells with data, but some cells left blank.
- 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".
- 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.
- 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 and make some recommendations for how to improve the survey!

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

Design a Survey Rubric

	Wow!	Getting There	Needs Improvement
Brainstorming Phase and Survey Creation	We developed at least eight questions, and correctly identified which would be answered by categorical or quantitative data. We correctly determined which data type each question will produce, and created a digital version of our survey.	We developed eight questions, but weren't always sure which would be answered by categorical vs. quantitative data. We couldn't always determine which data type each question would produce, but we created a google form with our questions.	Our questions were often incorrectly categorized as categorical vs. quantitative, and we had a lot of confusion about which data type each question would produce. We did not finish making the digital survey.
Required Questions	We correctly indicated all questions that are required.	We sometimes indicated required questions.	We forgot to indicate required questions.
Question Format	We strategically used multiple choice answers, checkboxes, and dropdown menus when possible to prevent dirty data.	We missed one or more opportunities to use multiple choice answers, checkboxes, or dropdown menus to prevent dirty data.	We did not consider question format as a tool to prevent dirty data.
Description	Each question has appropriate and helpful instructions that help collect maximally clean data.	Most questions have helpful instructions and / or the instructions could be clearer.	We often forgot to include instructions with questions and / or our instructions were confusing.
Validation	When relevant, we specified answer data types and / or parameters to prevent dirty data.	We sometimes forgot to specify data types and / or parameters or we did not correctly specify data types.	We did not specify data types and / or parameters in order to guard against dirty data.
Survey Hacking	We outlined several examples of realistic, dirty data that we entered on another group's survey. We offered compelling and practical suggestions to guard against dirty data, and shared insights that could help us improve our own survey.	We outlined a few examples of dirty data that we entered on another group's survey, but the examples were not always realistic. Our suggestions to guard against dirty data needed to be more specific. We shared one insight to help us improve our own survey.	Our examples of dirty data were not realistic. Our suggestions to guard against dirty data were not useful or helpful to the other group. We did not demonstrate that we learned how to improve our own survey.
Address Bad Data Entered	We have modified our survey so that it would no longer accept any of the bad data entered during the hacking process.	We have modified our survey to account for most of the bad data entered during the hacking process.	We didn't address most of the concerns revealed through the hacking process.

Survey Brainstorming

Team Members: _____

1) What is your group's topic?

2) What data do you plan to gather? Be sure to include a mix of categorical and quantitative!

Question	Categorical or Quantitative?	Expected Data Type of Response

3) What displays would you be interested in seeing as part of your analysis?

4) What grouped samples might you want to explore separately?
(*Just the teenagers, just the 8th graders, just the students with siblings, etc.*)

5) Are there any other questions you would need to ask as part of your survey in order to be able to identify the subgroups you want to study?

Question	Categorical or Quantitative?	Expected Data Type of Response

Threats to Validity

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

Some examples of threats are:

- **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.

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.

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.

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 spider and rabbit food was by far the most popular cuisine!

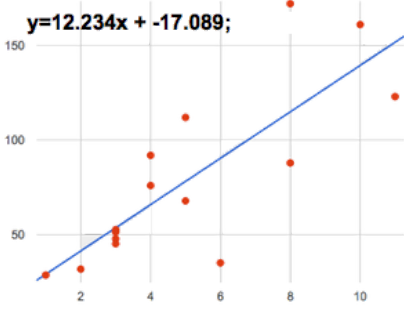
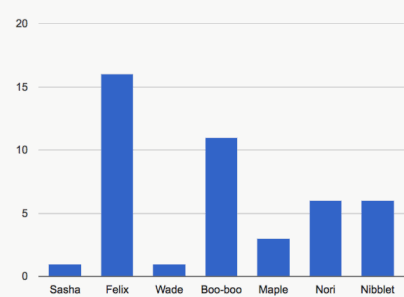
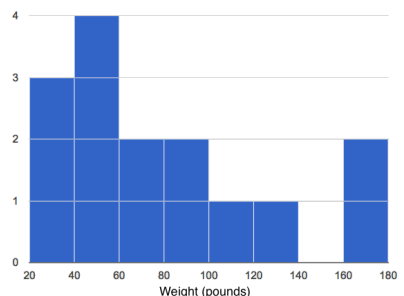
Explain why sampling just two animals can result in unreliable conclusions about what kind of food is needed.

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.

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

Fake News

There are six separate, *unrelated* claims below, and ALL OF THEM ARE WRONG! Your job is to figure out why by looking at the data.

	Data	Claim	What's Wrong
1	The average player on a basketball team is 6'1".	"Most of the players are taller than 6'."	
2	Linear regression found a positive correlation ($r=0.42$) between people's height and salary.	"Taller people are more qualified for their jobs."	
3		"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."	
4		"According to this bar chart, Felix makes up a little more than 15% of the total ages of all the animals in the dataset."	
5		"According to this histogram, most animals weigh between 40 and 60 pounds."	
6	Linear regression found a negative correlation ($r= -0.91$) between the number of hairs on a person's head and their likelihood of owning a wig.	"Owning wigs causes people to go bald."	

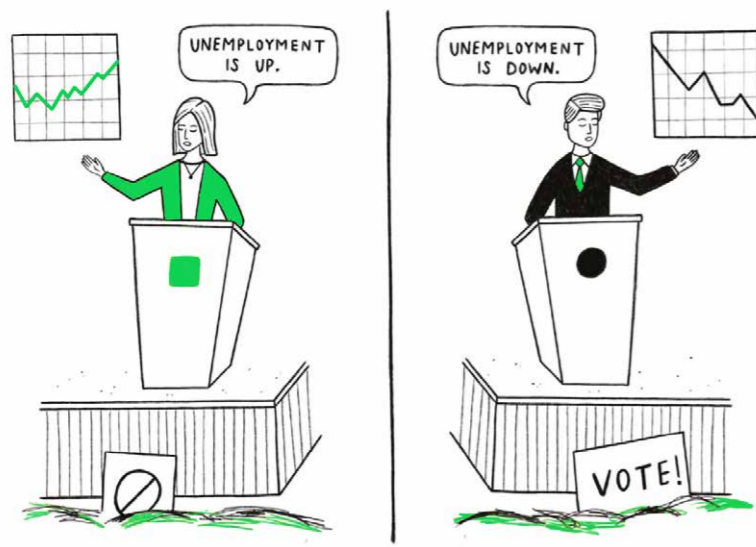
Lies, Darned Lies, and Statistics

1) Using real data and displays from your dataset, come up with a misleading claim.

Data	Claim	Why it's wrong

2) Trade papers with someone and figure out why their claims are wrong!

Data Fallacies to Avoid



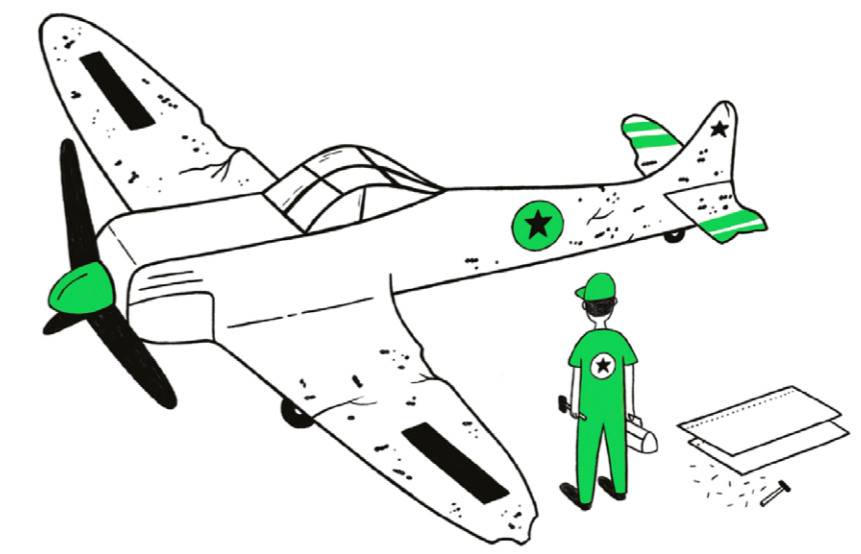
Cherry Picking

Selecting results that fit your claim and excluding those that don't.



Data Dredging

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.



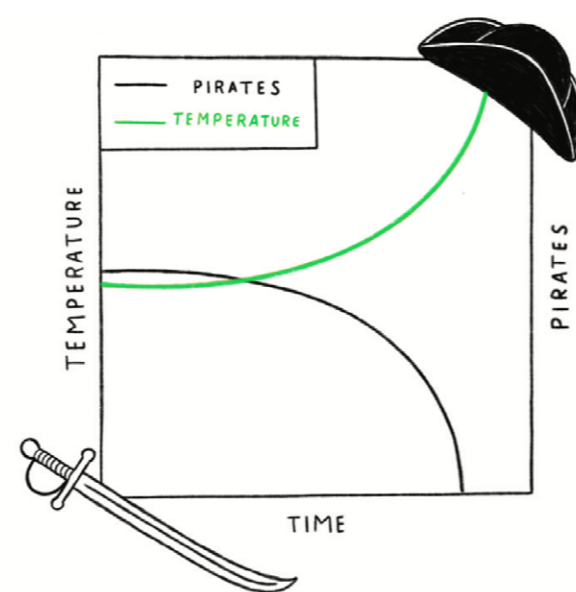
Survivorship Bias

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.



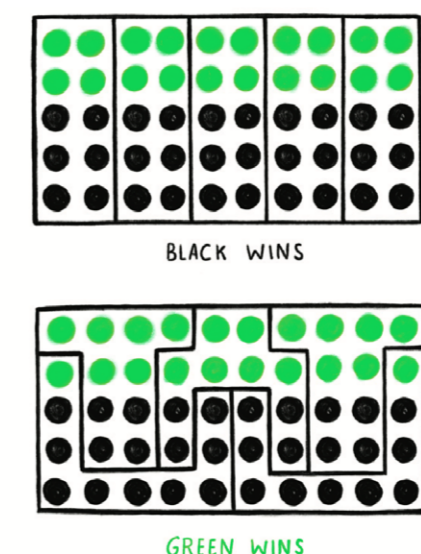
Cobra Effect

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



False Causality

Falsely assuming when two events appear related that one must have caused the other.



Gerrymandering

Manipulating the geographical boundaries used to group data in order to change the result.



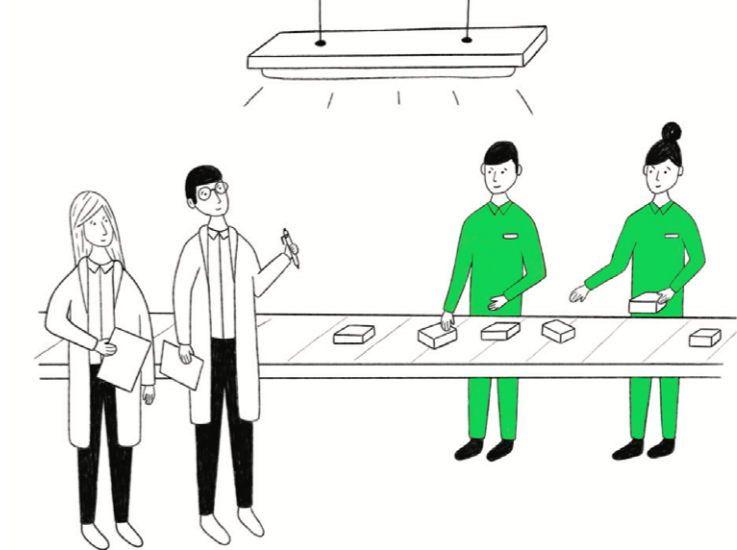
Sampling Bias

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.



Gambler's Fallacy

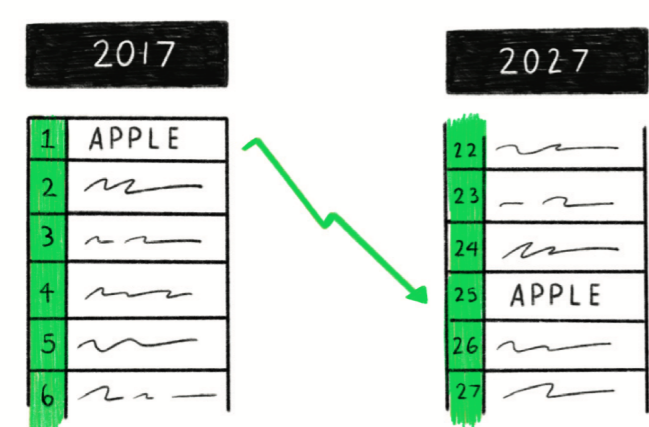
Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



Hawthorne Effect

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.

TOP COMPANIES



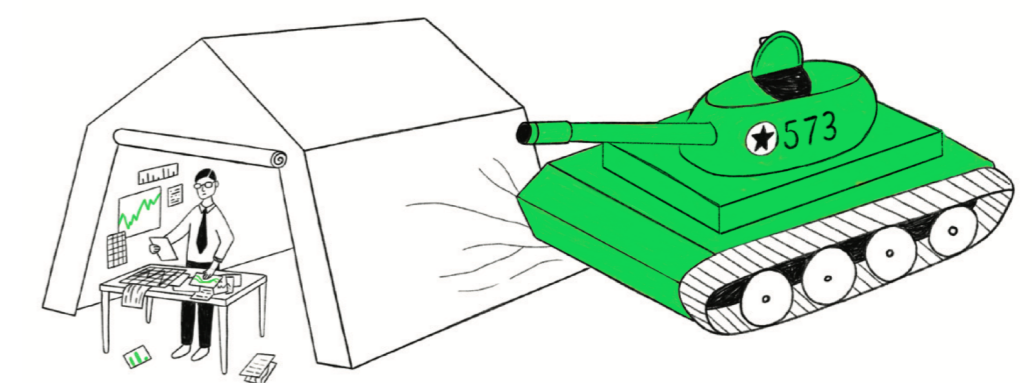
Regression Towards the Mean

When something happens that's unusually good or bad, it will revert back towards the average over time.

APPLICATION SUCCESS RATE		
	MALE	FEMALE
SUBJECT 1	14 % (168 of 1200)	15 % (270 of 1800)
SUBJECT 2	50 % (400 of 800)	51 % (102 of 200)
TOTAL	28 % (568 of 2000)	19 % (372 of 2000) ??

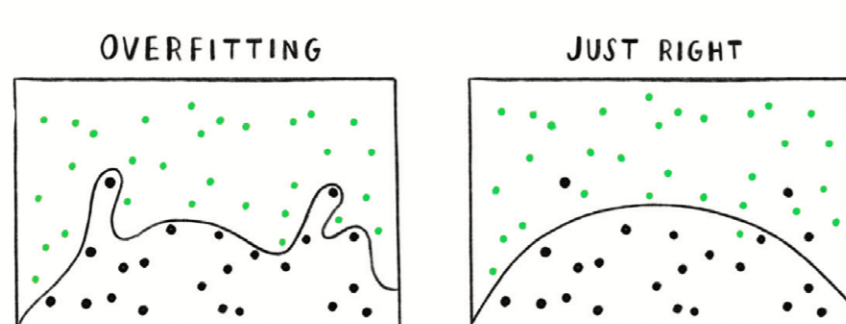
Simpson's Paradox

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



McNamara Fallacy

Relying solely on metrics in complex situations and losing sight of the bigger picture.



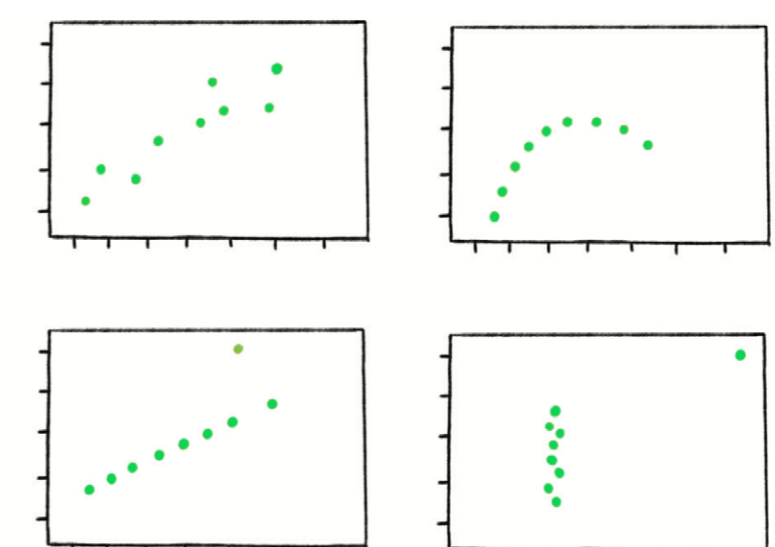
Overfitting

Creating a model that's overly tailored to the data you have and not representative of the general trend.



Publication Bias

Interesting research findings are more likely to be published, distorting our impression of reality.



Danger of Summary Metrics

Only looking at summary metrics and missing big differences in the raw data.

Data Cycle

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one): Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one): Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Data Cycle

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

<p>Ask Questions</p> 	<p>What question do you have?</p> <hr/> <hr/>	<p>Question Type (circle one):</p> <p>Lookup Arithmetic Statistical</p>
<p>Consider Data</p> 	<p>Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)</p> <hr/> <p>What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)</p> <hr/>	
<p>Analyze Data</p> 	<p>If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)</p> <hr/> <p>If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)</p> <hr/> <p>What code will make the table or display you want?</p> <hr/>	
<p>Interpret Data</p> 	<p>What did you find out? What can you infer?</p> <hr/> <hr/> <p>What - if any - new question(s) does this raise?</p> <hr/> <hr/>	

Design Recipe

Directions:

Contract and Purpose Statement

Every contract has three parts...

_____ :: _____ -> _____
function name Domain Range

what does the function do?

Examples

Write some examples, then circle and label what changes...

examples:

_____ (_____) is _____
function name input(s) what the function produces

_____ (_____) is _____
function name input(s) what the function produces

end

Definition

Write the definition, giving variable names to all your input values...

fun _____ (_____):
function name variable(s)

what the function does with those variable(s)

end

Directions:

Contract and Purpose Statement

Every contract has three parts...

_____ :: _____ -> _____
function name Domain Range

what does the function do?

Examples

Write some examples, then circle and label what changes...

examples:

_____ (_____) is _____
function name input(s) what the function produces

_____ (_____) is _____
function name input(s) what the function produces

end

Definition

Write the definition, giving variable names to all your input values...

fun _____ (_____):
function name variable(s)

what the function does with those variable(s)

end

Design Recipe

Directions:

Contract and Purpose Statement

Every contract has three parts...

_____ :: _____ -> _____
function name Domain Range

what does the function do?

Examples

Write some examples, then circle and label what changes...

examples:

_____ (_____) is _____
function name input(s) what the function produces

_____ (_____) is _____
function name input(s) what the function produces

end

Definition

Write the definition, giving variable names to all your input values...

fun _____ (_____):
function name variable(s)

what the function does with those variable(s)

end

Directions:

Contract and Purpose Statement

Every contract has three parts...

_____ :: _____ -> _____
function name Domain Range

what does the function do?

Examples

Write some examples, then circle and label what changes...

examples:

_____ (_____) is _____
function name input(s) what the function produces

_____ (_____) is _____
function name input(s) what the function produces

end

Definition

Write the definition, giving variable names to all your input values...

fun _____ (_____):
function name variable(s)

what the function does with those variable(s)

end

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

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
<code># above</code>	<code>:: (<u>Image</u>_{above} , <u>Image</u>_{below})</code>	<code>-> Image</code>
	<code>above(circle(10, "solid", "black"), square(50, "solid", "red"))</code>	
<code># bar-chart</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{column})</code>	<code>-> Image</code>
	<code>bar-chart(animals-table, "species")</code>	
<code># bar-chart-summarized</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{labels} , <u>String</u>_{values})</code>	<code>-> Image</code>
	<code>bar-chart-summarized(count(animals-table, "species"), "value", "count")</code>	
<code># box-plot</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{column})</code>	<code>-> Image</code>
	<code>box-plot(animals-table, "weeks")</code>	
<code># box-plot-scaled</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{column} , <u>Number</u>_{low} , <u>Number</u>_{high})</code>	<code>-> Image</code>
	<code>box-plot-scaled(animals-table, "weeks", 1, 40)</code>	
<code># count</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{column})</code>	<code>-> Table</code>
	<code>count(animals-table, "species")</code>	
<code># first-n-rows</code>	<code>:: (<u>Table</u>_{table-name} , <u>Number</u>_{num-rows})</code>	<code>-> Table</code>
	<code>first-n-rows(animals-table, 15)</code>	
<code># histogram</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{labels} , <u>String</u>_{values} , <u>Number</u>_{bin-size})</code>	<code>-> Image</code>
	<code>histogram(animals-table, "species", "weeks", 2)</code>	
<code># line-graph</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{labels} , <u>String</u>_{xs} , <u>String</u>_{ys})</code>	<code>-> Image</code>
	<code>line-graph(animals-table, "name", "pounds", "weeks")</code>	
<code># lr-plot</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{labels} , <u>String</u>_{xs} , <u>String</u>_{ys})</code>	<code>-> Image</code>
	<code>lr-plot(animals-table, "name", "pounds", "weeks")</code>	
<code># mean</code>	<code>:: (<u>Table</u>_{table-name} , <u>String</u>_{column})</code>	<code>-> Number</code>
	<code>mean(animals-table, "pounds")</code>	

Name	Domain	Range
# median	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> Number
<i>median(animals-table, "pounds")</i>		
# modes	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> List
<i>modes(animals-table, "pounds")</i>		
# modified-box-plot	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> Image
<i>modified-box-plot(animals-table, "pounds")</i>		
# modified-box-plot-scaled	:: (<u>Table</u> , <u>String</u> , <u>Number</u> , <u>Number</u>) <small>table-name column low high</small>	-> Image
<i>modified-box-plot-scaled(animals-table, "weeks", 1, 40)</i>		
# modified-vert-box-plot	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> Image
<i>modified-vert-box-plot(animals-table, "pounds")</i>		
# modified-vert-box-plot-scaled	:: (<u>Table</u> , <u>String</u> , <u>Number</u> , <u>Number</u>) <small>table-name column low high</small>	-> Image
<i>modified-vert-box-plot-scaled(animals-table, "weeks", 1, 40)</i>		
# multi-bar-chart	:: (<u>Table</u> , <u>String</u> , <u>String</u>) <small>table-name group subgroup</small>	-> Image
<i>multi-bar-chart(animals-table, "species", "sex")</i>		
# overlay	:: (<u>Image</u> , <u>Image</u>) <small>top bottom</small>	-> Image
<i>overlay(circle(10, "solid", "black"), square(50, "solid", "red"))</i>		
# pie-chart	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> Image
<i>pie-chart(animals-table, "species")</i>		
# pie-chart-summarized	:: (<u>Table</u> , <u>String</u> , <u>String</u>) <small>table-name labels values</small>	-> Image
<i>pie-chart-summarized(count(animals-table, "species"), "value", "count")</i>		
# put-image	:: (<u>Image</u> , <u>Number</u> , <u>Number</u> , <u>Image</u>) <small>front x-coordinate y-coordinate behind</small>	-> Image
<i>put-image(circle(10, "solid", "black"), 10, 10, square(50, "solid", "red"))</i>		
# random-rows	:: (<u>Table</u> , <u>Number</u>) <small>table-name num-rows</small>	-> Table
<i>random-rows(animals-table, 10) # select 10 random rows from the table</i>		
# rotate	:: (<u>Number</u> , <u>Image</u>) <small>degrees img</small>	-> Image
<i>rotate(45, star(50, "solid", "dark-blue"))</i>		
# scale	:: (<u>Number</u> , <u>Image</u>) <small>factor img</small>	-> Image
<i>scale(1/2, star(50, "solid", "light-blue"))</i>		
# scatter-plot	:: (<u>Table</u> , <u>String</u> , <u>String</u> , <u>String</u>) <small>table-name labels xs ys</small>	-> Image
<i>scatter-plot(animals-table, "name", "pounds", "weeks")</i>		

Name	Domain	Range
# <code>sort</code>	:: (<u>Table</u> , <u>String</u> , <u>Boolean</u>) <small>table-name column ascending</small>	-> Table
<code>sort(animals-table, "species", true)</code>		
# <code>stacked-bar-chart</code>	:: (<u>Table</u> , <u>String</u> , <u>String</u>) <small>table-name group subgroup</small>	-> Image
<code>stacked-bar-chart(animals-table, "species", "sex")</code>		
# <code>stdev</code>	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> Number
<code>stdev(animals-table, "pounds")</code>		
# <code>string-contains</code>	:: (<u>String</u> , <u>String</u>) <small>haystack needle</small>	-> Boolean
<code>string-contains("hotdog", "dog")</code>		
# <code>vert-box-plot</code>	:: (<u>Table</u> , <u>String</u>) <small>table-name column</small>	-> Image
<code>vert-box-plot(animals-table, "weeks")</code>		
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->
	::	->



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.