Name: _____

Student Workbook

Fall, 2024 - Pyret Edition

Workbook v3.1

Brought to you by the Bootstrap team:

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

Pioneers in Computing and Mathematics

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

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

Notice and Wonder

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

What do you Notice?	What do you Wonder?

Windows and Mirrors

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

_____

_____

_____

_____

_____

_____

_____

_____

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

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Reflection: 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 *Kujo took a long time to be adopted*	*Is it because he was so big?*	Yes No
I notice that		Yes No
I notice that		Yes No
I notice that		Yes No
I notice that		Yes No
I notice that		Yes No
I notice that		Yes No

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

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

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

        _____ .

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

        _____ .

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)](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) `3 <= 4`			2) `"a" > "b"`		
3) `3 == 2`			4) `"a" < "b"`		
5) `2 < 4`			6) `"a" == "b"`		
7) `5 >= 5`			8) `"a" <> "a"`		
9) `4 >= 6`			10) `"a" >= "a"`		
11) `3 <> 3`			12) `"a" <> "b"`		
13) `4 <> 3`			14) `"a" >= "b"`		

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.

_____(_____ , _____)
　　function-name　　　　　　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!

_____(_____ , _____)
　　function-name　　　　　　　　　　　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 <u>Animals Starter File</u> 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)](code.pyret.org) 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
triangle(80, "solid", "darkgreen")				
# 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`

_____ (_____ , _____)
<small>function-name</small>   <small>table-name :: Table</small>   <small>column-name :: String</small>

Sketch a bar chart below.

Bar charts summarize 1 column of _____ data.
<small>categorical/quantitative</small>

This kind of display tells us…

_____

_____

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

_____ (_____ , _____)
<small>function-name</small>   <small>table-name :: Table</small>   <small>column-name :: String</small>

Sketch a pie chart below.

Pie charts summarize 1 column of _____ data.
<small>categorical/quantitative</small>

This kind of display tells us…

_____

_____

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

_____ (_____ , _____)
<small>function-name</small>   <small>table-name :: Table</small>   <small>column-name :: String</small>

Sketch a box plot below.

Box plots summarize 1 column of _____ data.
<small>categorical/quantitative</small>

This kind of display tells us…

_____

_____

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

_____ (_____ , _____ , _____ , _____)
<small>function-name</small>   <small>table-name :: Table</small>   <small>labels :: String</small>   <small>values :: String</small>   <small>bin-width :: Number</small>

Sketch a histogram below.

Histograms summarize 1 column of _____ data.
<small>categorical/quantitative</small>

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? _____

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 the 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?
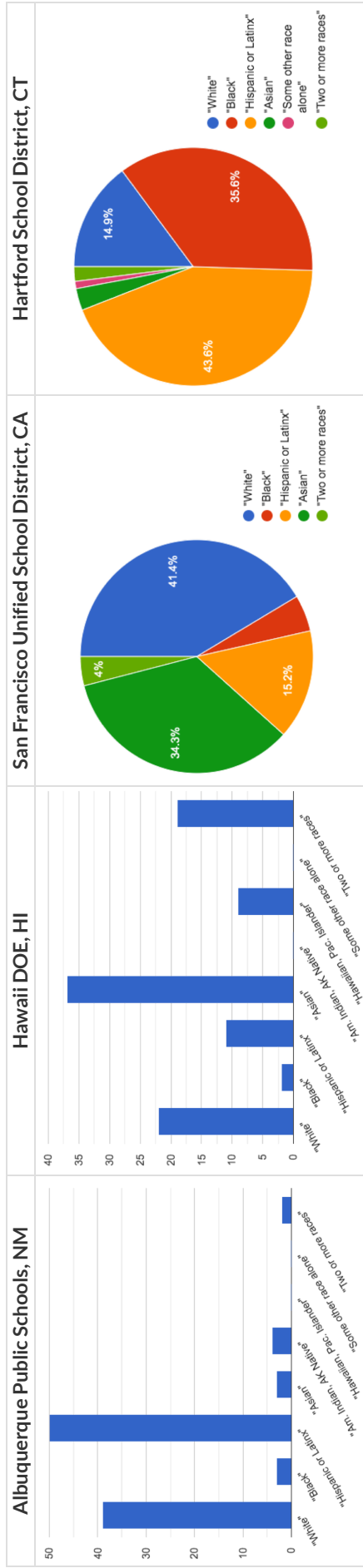
Albuquerque Public Schools, NM	Hawaii DOE, HI	San Francisco Unified School District, CA	Hartford School District, CT
What do you Notice?		**What do you Wonder?**	

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
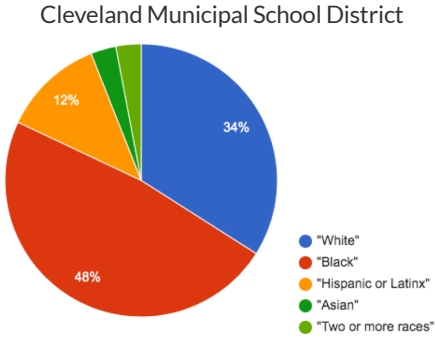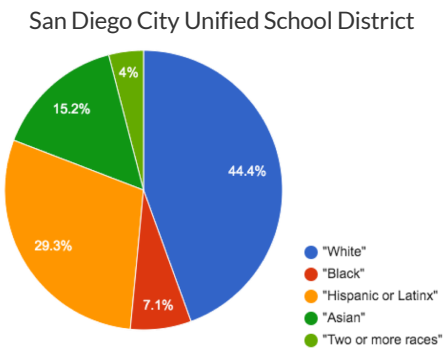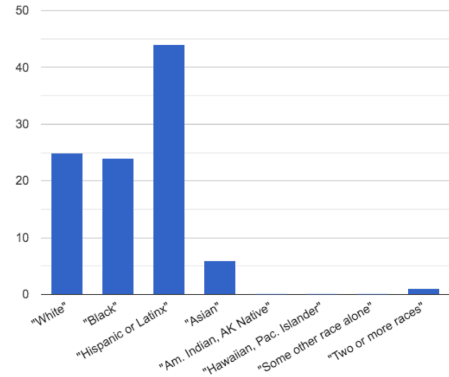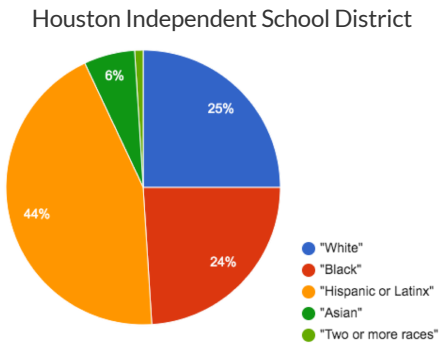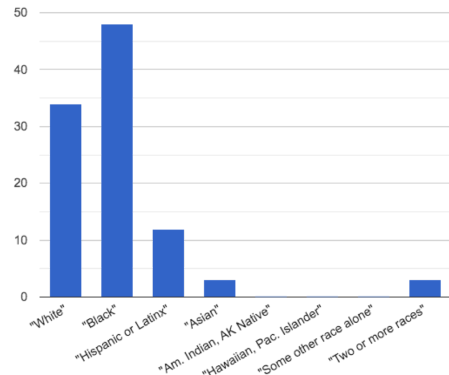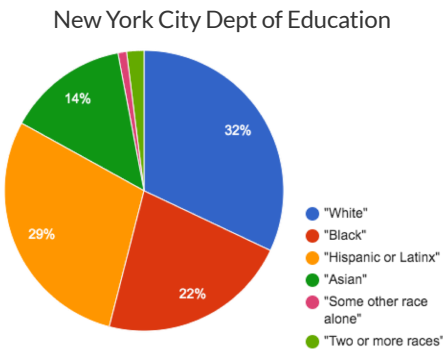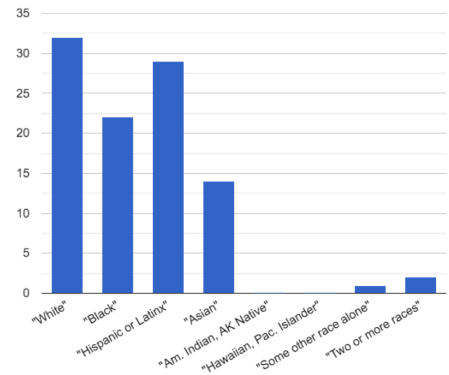
4

D
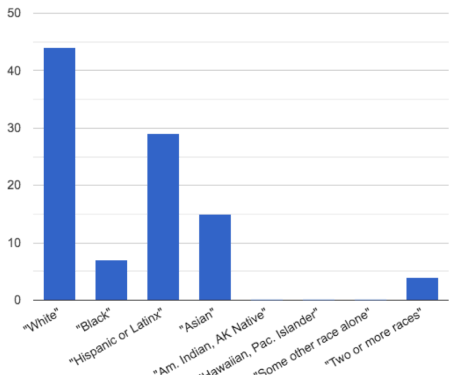
Introducing Displays for Subgroups

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 barcharts**.

Political Structure of Countries in the Americas

Distribution of sexes across Cats, Dogs and Rabbits
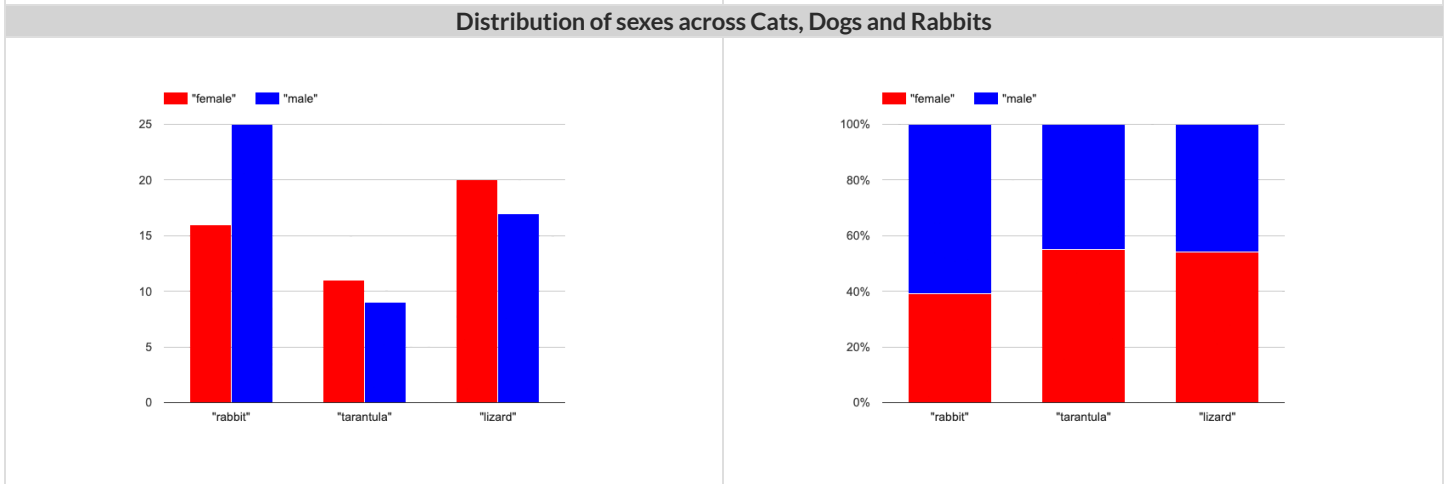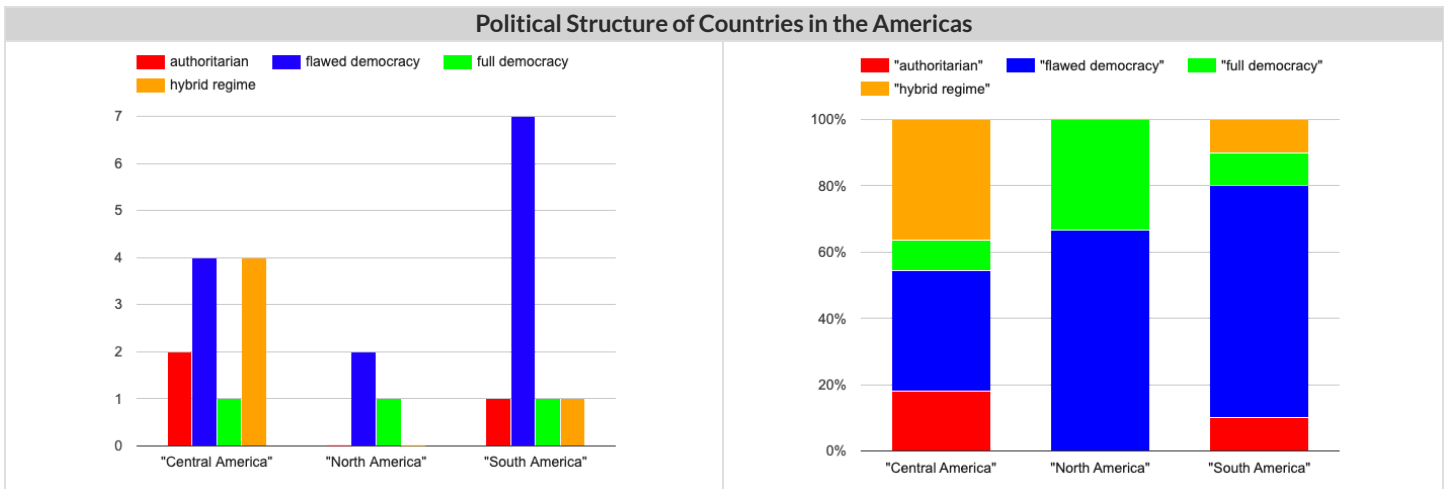
What do you Notice?	What do you Wonder?

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

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

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

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!

The Data Cycle

Ask Questions

Interpret Data · Research Topic · Consider Data

Analyze Data
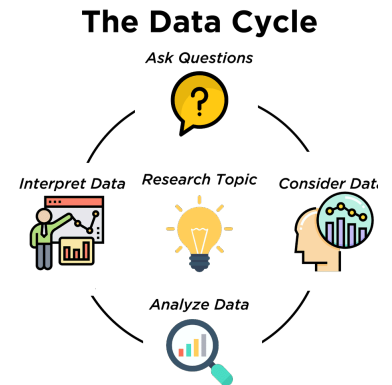
- We start by **Asking Questions** after reviewing and closely observing the data. These questions can come from initial wonderings, or as a result of previous data cycle. Most questions can be broken down into one of four categories:
  - **Lookup questions** - Answered by only reading the table, no 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 often involve multiple steps to answer, and the answer isn'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.

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

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

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

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

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

Data Cycle: 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.

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

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

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

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**.

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

Interpret Data

❏ The chart shows that there is an even distribution of _____ .
　　　　　　　　　　　　　　　　　　　　　　　　　　　　 variable

❏ The chart shows that the most common _____ is/are _____
　　　　　　　　　　　　　　　　　　　　　　 variable

I notice that _____

I wonder _____

- How does the distribution of _____ differ by _____ ?
　　　　　　　　　　　　　　　 variable　　　　　　　　　 variable

- _____
Another question I have is…

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

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

Interpret Data

When we break the distribution of _____ down by _____ :
　　　　　　　　　　　　　　　　 variable　　　　　　　　　 variable

- I notice that _____

- I wonder _____

Another question I have is…

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?

_____

_____

_____

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?
All the animals		

code: _____

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

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

code: _____

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

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

code: _____

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

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

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?
All the animals		

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?
All the animals		

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
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*
		Lookup *Arithmetic* *Statistical* *Can't Answer*

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

Choose two columns to describe below.

6) _____, 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.

Ask Questions

What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

What did you find out? What can you infer?

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

Ask Questions

What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

What did you find out? What can you infer?

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

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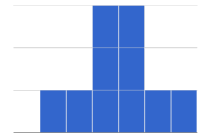
- Some extreme values may be far greater or far lower than the other values in a dataset. These extreme values are called **outliers**.

- 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.
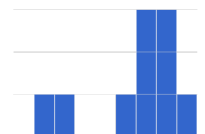
- A dataset that is **skewed right** has a few values that are unusually high. The histogram for a skewed right dataset has a few data points that are stretched out to the right (higher) end of the x-axis.
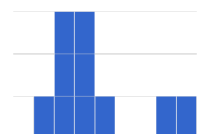
- One way to visualize the difference between a histogram of data that is **skewed left** or **skewed right** is to think about the lengths of our toes on our left and right feet.

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.
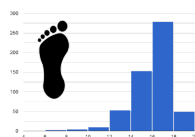
Our right feet have the big toe on the left and smaller toes on the right, more closely resembling the shape of a histogram of "skewed right" data.
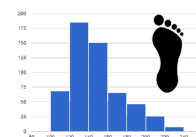
Summarizing Columns with Bar Charts & Histograms

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?
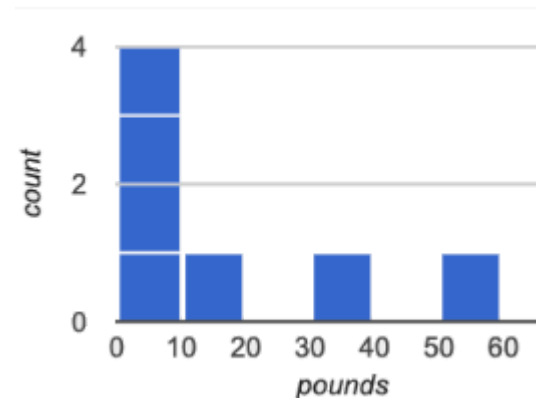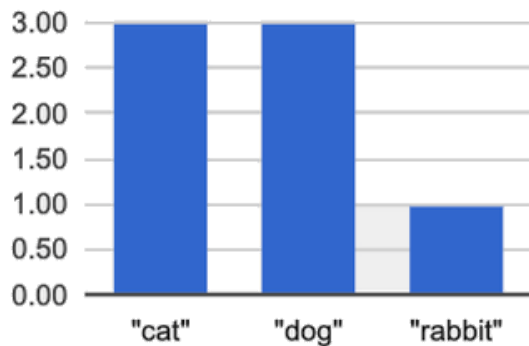
Similarities	Differences

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.

Number of Teeth
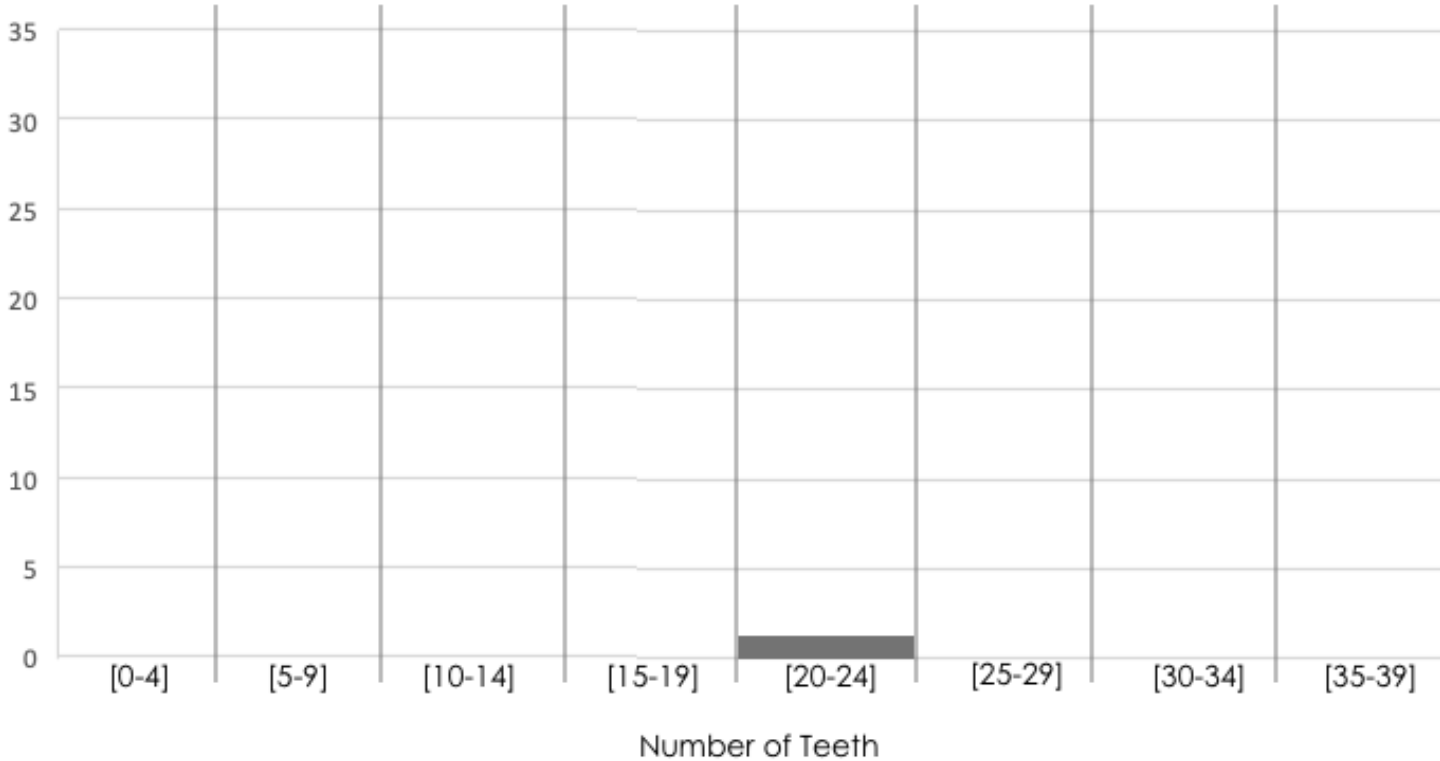
Reading Histograms

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

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

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

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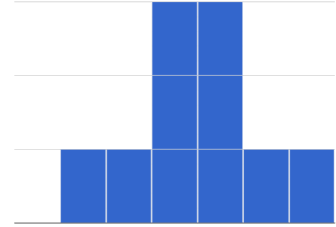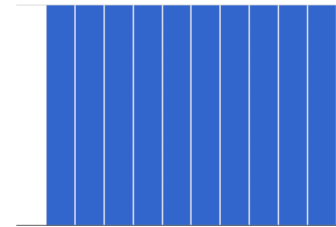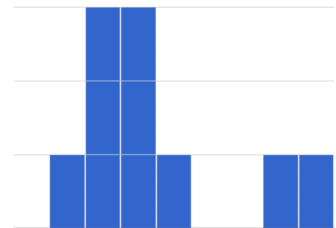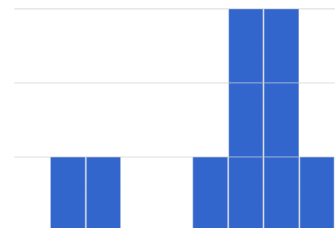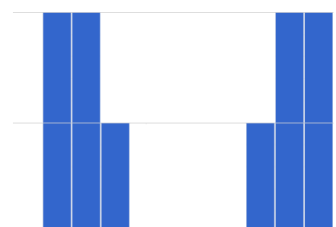
Choosing the Right Bin Size

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**.

Ask Questions

What is the shape of the age column of the Animals dataset?
What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

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

The bin size I chose is _____, which resulted in a histogram with _____ bins. I chose this bin size because
                                      bin size   how many?

I would describe the shape of this histogram as _____

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

I wonder _____

Ask Questions

What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

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

The bin size I chose is _____, which resulted in a histogram with _____ bins. I chose this bin size because
                                      bin size   how many?

I would describe the shape of this histogram as _____

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

I wonder

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.

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

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

Identifying Shape - Histograms

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

1

2

3

4

5
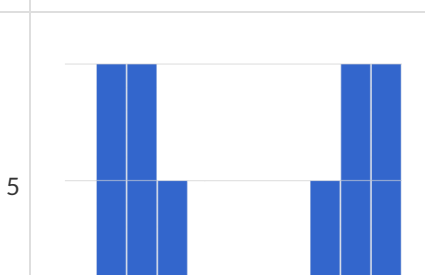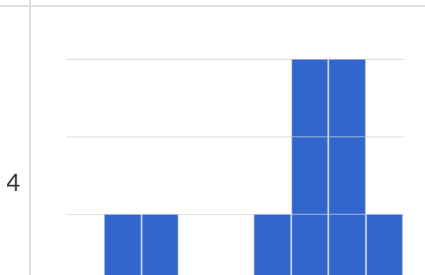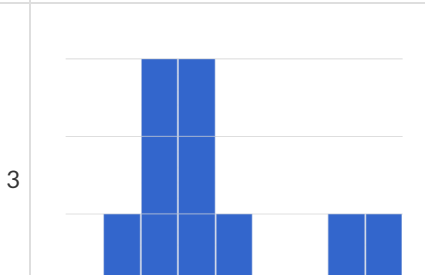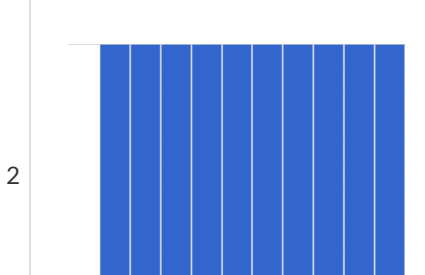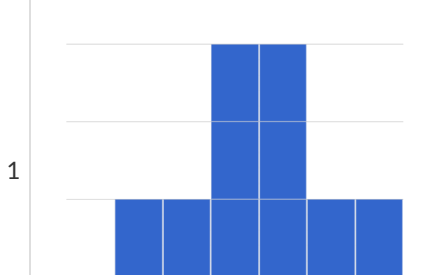
Data Cycle: Shape of the Animals Dataset

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!

Ask Questions

What question do you have?

What is the distribution of weight among all animals at the shelter?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

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

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

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

I wonder _____

Ask Questions

What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

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

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

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

I wonder _____

Outliers: Should they Stay or Should they Go?

Tahli and Fernando are looking at a scatter plot showing the relationship between poverty and test scores at schools in Michigan. They find a trend, with low-poverty schools generally having higher test scores than high-poverty schools. However, one school is an extreme outlier: the highest poverty school in the state also has higher test scores than most of the other schools!
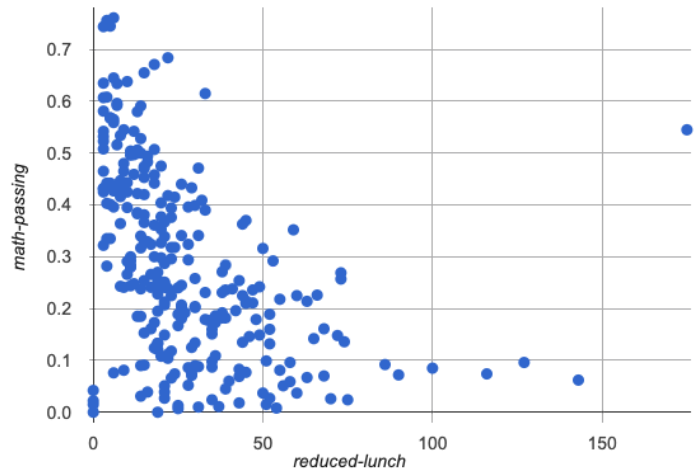
Tahli thinks the outlier should be removed before they start analyzing, and Fernando thinks it should stay. Here are their reasons:

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

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

_____

_____

_____

_____

_____

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 decriptive measure of center than the median.
  - 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:
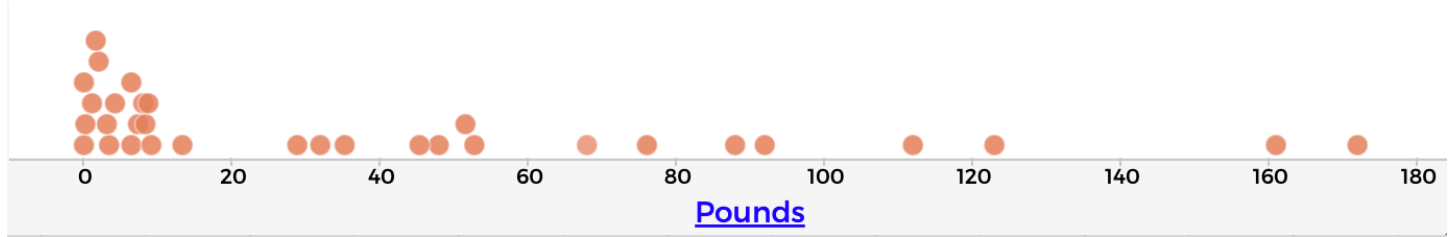
1) What do you Notice?

_____

_____

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.

Ask Questions

What is the mean age for cats at the shelter?
What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)

If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)

What code will make the table or display you want?

Interpret Data

What did you find out? What can you infer?

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

Ask Questions

What is the median time it takes for an animal to be adopted?
What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)

If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)

What code will make the table or display you want?

Interpret Data

What did you find out? What can you infer?

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

Data Cycle Practice

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

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

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

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**.
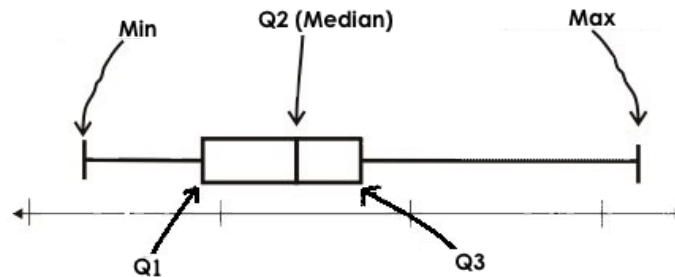
- Each of the four sections of the box plot contains 25% of the data.
    - If the values are distributed evenly across the range, the four sections of the box plot will be equal in width.
    - Uneven distributions will show up as differently-sized sections of a box plot.

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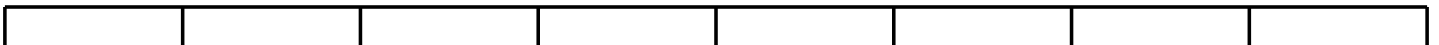
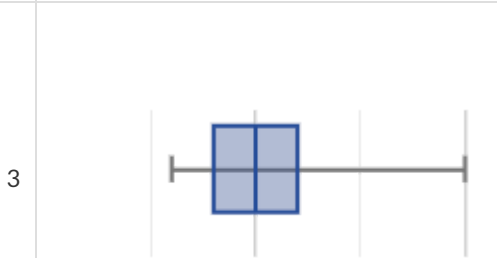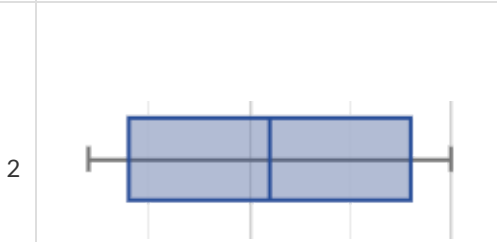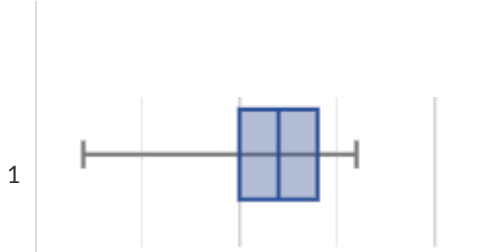7) The **Range** is: _____ and the **Interquartile Range(IQR)** is: _____ .

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.*
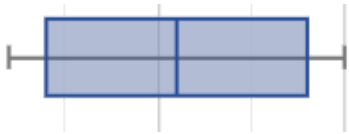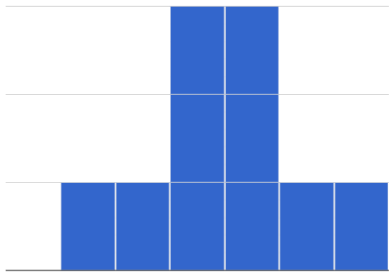
1

2

3

4

5
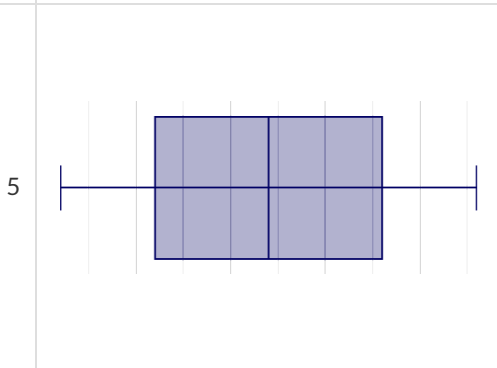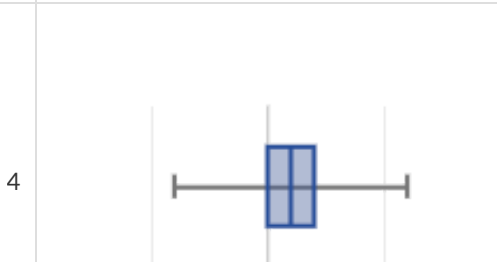
Matching Box Plots to Histograms

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.
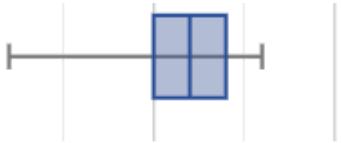
1   A

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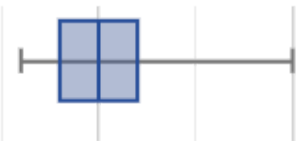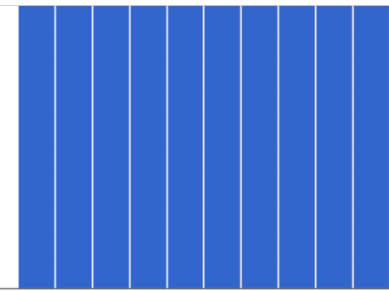
3   C
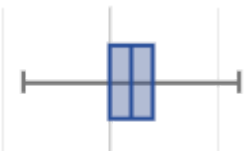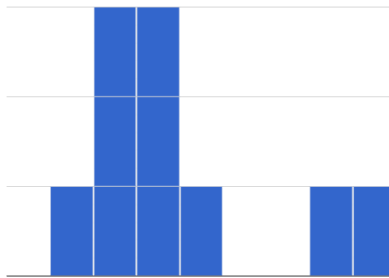
4   D
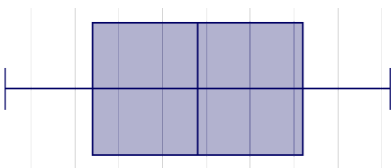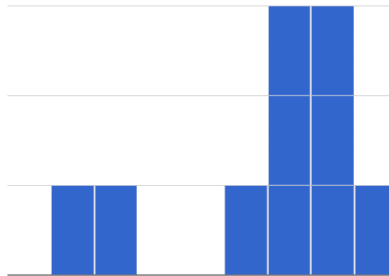
5   E
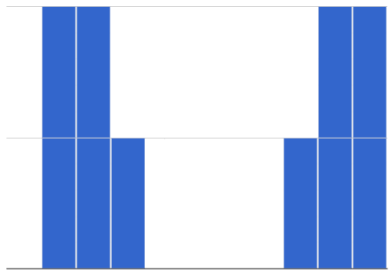
Directions: Connect each item on this page to at least one other item by drawing an arrow and writing an explanation of how they are connected along the arrow. (Arrows may curve.)

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**.

|		
Ask Questions	*What is the distribution of the weeks column from the animals dataset?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	What code will make the table or display you want?	

Interpret Data

The box plot for _____ is _____ .
  x-variable in context   skewed left / skewed right / symmetric / etc.

The 5-number summary is: min = _____ Q1 = _____ median = _____ Q3 = _____ max = _____

The middle 50% of the data lies between _____ and _____ so the Interquartile Range is _____

I notice that _____
  Consider statements like: 75% of the data fall below … / The top 25% of the data fall between … / etc

_____

I wonder _____

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

Interpret Data

The box plot for _____ is _____ .
  x-variable in context   skewed left / skewed right / symmetric / etc.

The 5-number summary is: min = _____ Q1 = _____ median = _____ Q3 = _____ max = _____

The middle 50% of the data lies between _____ and _____ so the Interquartile Range is _____

I notice that _____
  Consider statements like: 75% of the data fall below … / The top 25% of the data fall between … / etc

I wonder _____

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.

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

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

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:
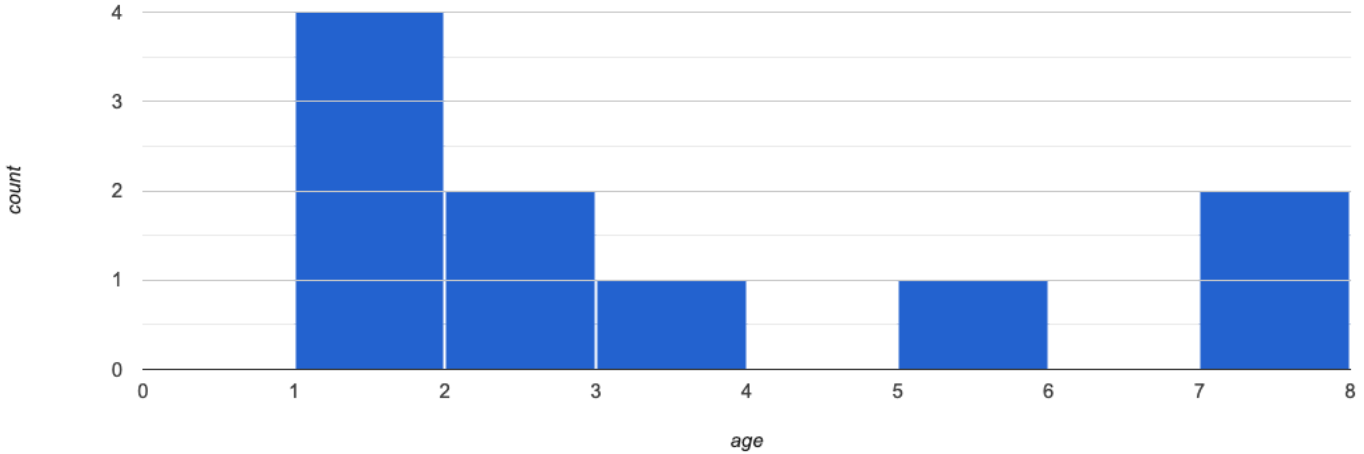
2) Describe the shape of this histogram. _____

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:
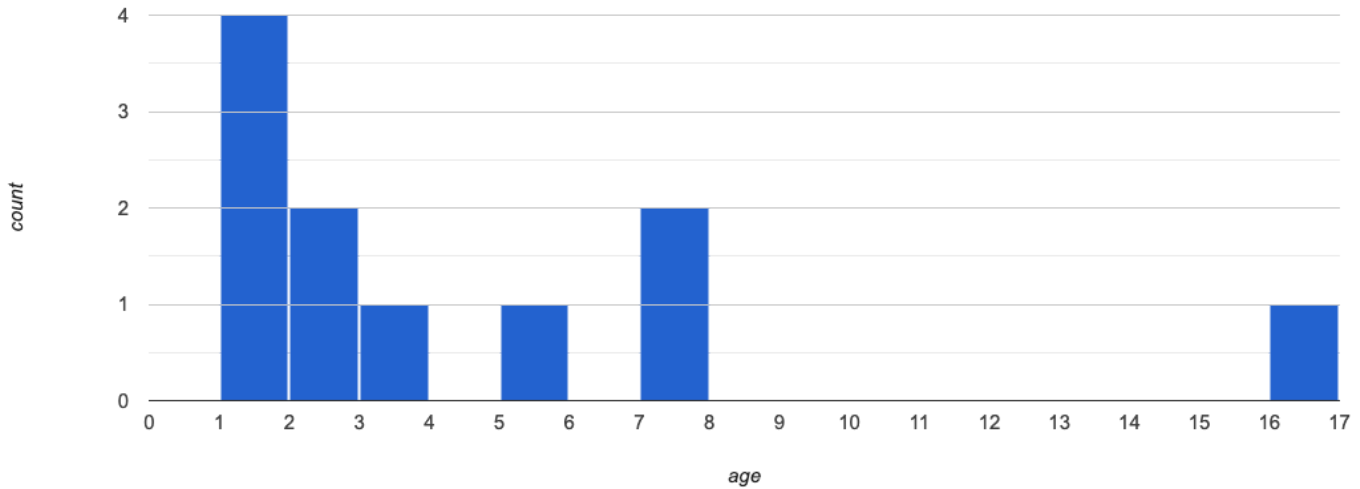
1) Describe the shape of this histogram. _____

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.

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

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

Data Cycle: 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.

Ask Questions	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical

Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)
	What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data	If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)
	If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)
	What code will make the table or display you want?

Interpret Data	What did you find out? What can you infer?
	What - if any - new question(s) does this raise?

Ask Questions	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical

Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.)
	What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)

Analyze Data	If you only need some rows, define your filter function here (Need help? Use the Design Recipe!)
	If you need to make a new column, define your builder function here (Need help? Use the Design Recipe!)
	What code will make the table or display you want?

Interpret Data	What did you find out? What can you infer?
	What - if any - new question(s) does this raise?

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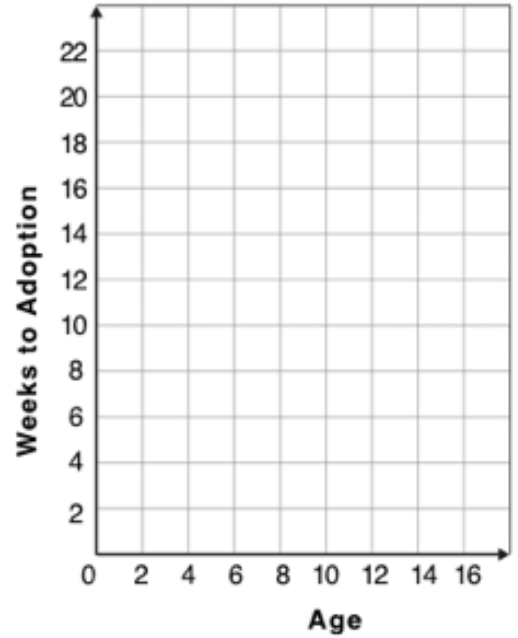
2) Plot the rest of the animals - one at a time - labeling each point as you go. After each animal, ask yourself whether or not you see a pattern in the data.

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?

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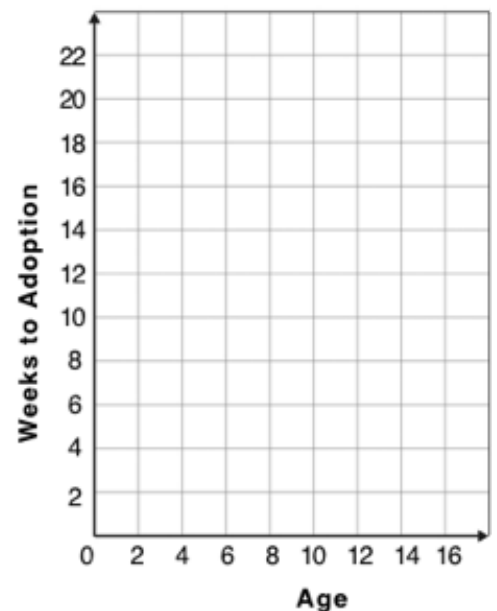
7) Place points on the graph to create a scatter plot with NO relationship.

Exploring Relationships Between Columns

This page is designed to be used with the [Animals Starter File](#). Log into [code.pyret.org (CPO)](#) 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?

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

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

Data Cycle: Relationships in Your Dataset

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

Ask Questions

What question do you have?

_____

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

❏ There appears to be no relationship between _____ and _____.
x-variable   y-variable

❏ There appears to be a _____, _____, _____ relationship
strong / weak / moderate   positive / negative   linear / non-linear

between _____ and _____.
x-variable   y-variable

Some possible outliers might be _____

Ask Questions

What question do you have?

_____

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

❏ There appears to be no relationship between _____ and _____.
x-variable   y-variable

❏ There appears to be a _____, _____, _____ relationship
strong / weak / moderate   positive / negative   linear / non-linear

between _____ and _____.
x-variable   y-variable

Some possible outliers might be _____

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? $y =$ _____ $x +$ _____

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

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

Predictions from Scatter Plots
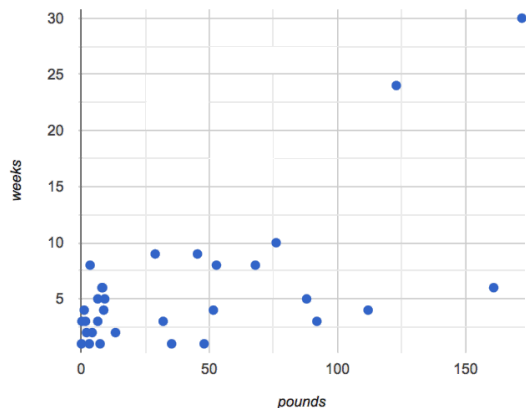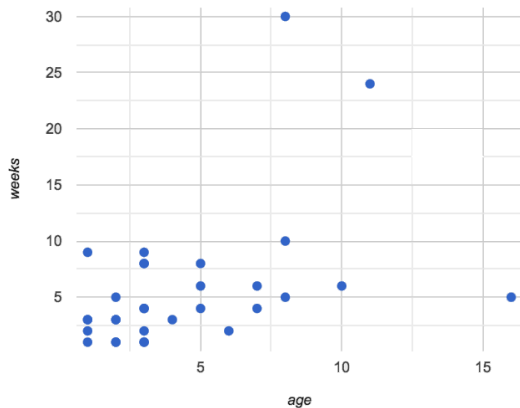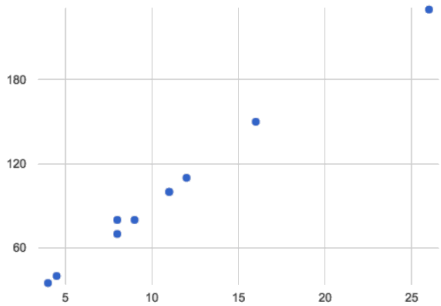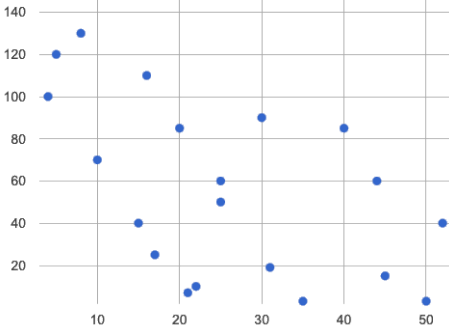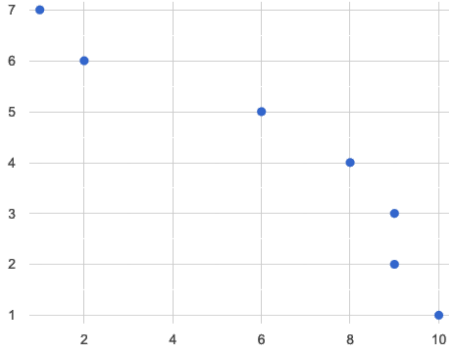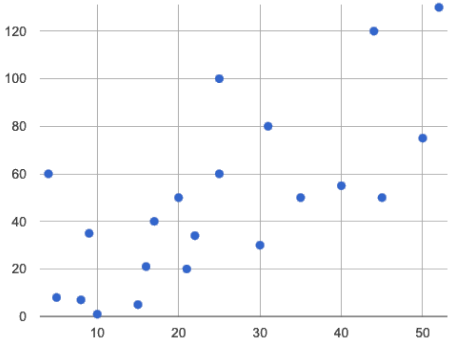
9) Draw the line of best fit for age-v-weeks (on the left). Is this a strong correlation that will allow us to make a good prediction of an animal's adoption time just by knowing how old it is?

_____

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

Direction:	Positive	Negative	None

Strength: Stronger   Weaker

estimated **r:** -1   -0.5   0   0.5   1

B

Direction:	Positive	Negative	None

Strength: Stronger   Weaker

estimated **r:** -1   -0.5   0   0.5   1

C

Direction:	Positive	Negative	None

Strength: Stronger   Weaker

estimated **r:** -1   -0.5   0   0.5   1

D

Direction:	Positive	Negative	None

Strength: Stronger   Weaker

estimated **r:** -1   -0.5   0   0.5   1

Exploring lr-plot

age

You should already have plotted `lr-plot(animals-table, "name", "age", "weeks")` *in the* <u>Animals Starter File</u>.

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

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? $y =$ _____ $x +$ _____

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

y=3.061x+25.175; R: +0.740
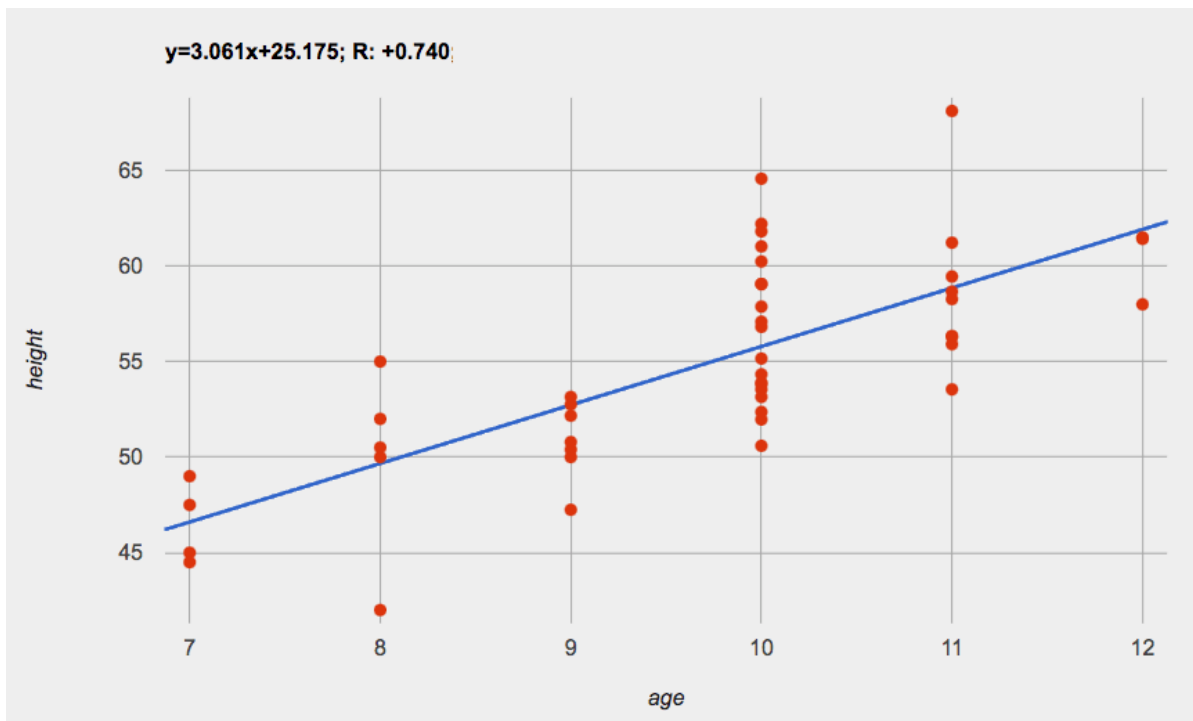
1) About how many inches are kids in this dataset expected to grow per year? _____

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

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

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

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

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

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

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

_____

_____

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

_____

_____

Interpreting Regression Lines & r-Values

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

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

Data Cycle: Animals Regression Analysis

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

Ask Questions

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

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

`lr-plot(animals-table, "name", "age", "weeks")`
What code will make the table or display you want?

Interpret Data

I performed a linear regression on a sample of _____ animals at the shelter _____ and found a
[dataset or subset]

_____ moderate (R=.448), positive _____ correlation between _____ age _____ and
weak / strong / moderate (R=...), positive / negative   [x-axis]

_____ time to adoption _____ . I would predict that a 1 _____ year _____ increase in _____ age _____ is
[y-axis]   [x-axis units]   [x-axis]

associated with a _____ .789 week _____ _____ increase _____ in _____ time to adoption _____ .
[slope, y-units]   increase / decrease   [y-axis]

Ask Questions

What question do you have?

Question Type
(circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

Interpret Data

I performed a linear regression on a sample of _____ and found a
[dataset or subset]

_____ correlation between _____ and
weak / strong / moderate (R=...), positive / negative   [x-axis]

_____ . I would predict that a 1 _____ increase in _____ is
[y-axis]   [x-axis units]   [x-axis]

associated with a _____ _____ in _____ .
[slope, y-units]   increase / decrease   [y-axis]

Describing Relationships

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

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

y=-22.321x + 444.778; r-sq: 0.426
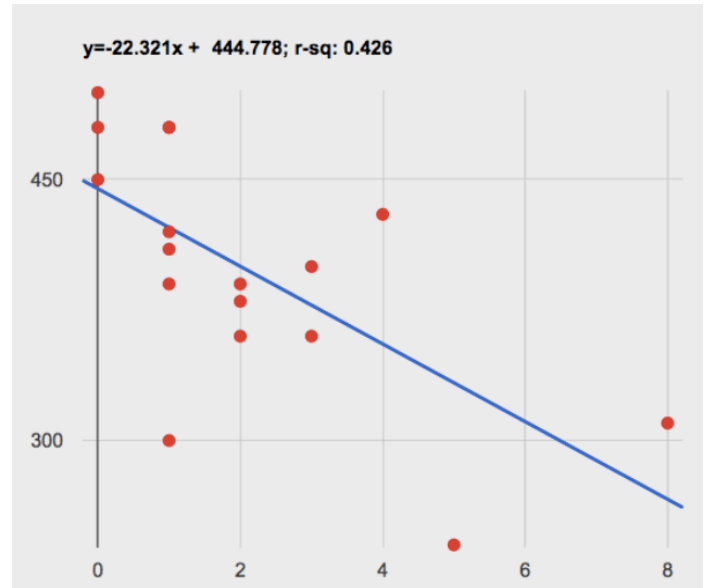
1) Describe the relationship between coffee intake and minutes of sleep shown in the data above.

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.

Ask Questions	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	What code will make the table or display you want?	
Interpret Data	I performed a linear regression on a sample of _____ [dataset or subset] and found a _____ weak / strong / moderate (R=…), positive / negative correlation between _____ [x-axis] and _____ [y-axis]. I would predict that a 1 _____ [x-axis units] increase in _____ [x-axis] is associated with a _____ [slope, y-units] _____ increase / decrease in _____ [y-axis].	

Ask Questions	What question do you have?	**Question Type** (circle one): Lookup Arithmetic ~~Statistical~~
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	What code will make the table or display you want?	
Interpret Data	I performed a linear regression on a sample of _____ [dataset or subset] and found a _____ weak / strong / moderate (R=…), positive / negative correlation between _____ [x-axis] and _____ [y-axis]. I would predict that a 1 _____ [x-axis units] increase in _____ [x-axis] is associated with a _____ [slope, y-units] _____ increase / decrease in _____ [y-axis].	

Case Study: Ethics, Privacy, and Bias

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

1. **Missing Data** - A column containing some cells with data, but some cells left blank.
2. **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"`.
3. **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.
4. **Inconsistent Naming** - Inconsistent spelling and capitalization for entries lead to them being counted as different. For example, a `species` column where some entries are `"cat"` and others are `"Cat"` will not give us a full picture of the cats.

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		

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?

_____

_____

_____

_____

_____
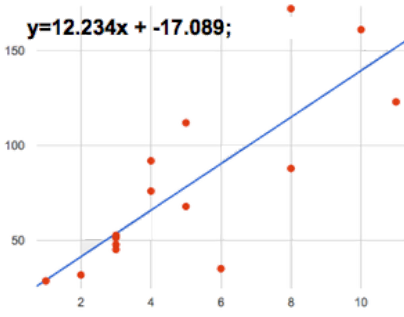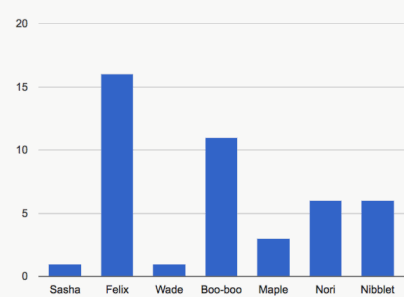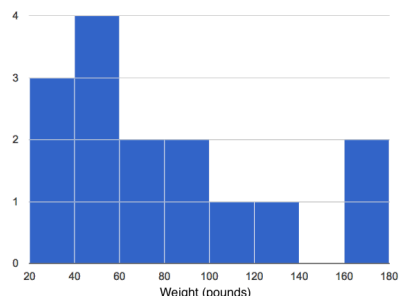
Fake News

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	$y=12.234x + -17.089;$	*"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 Cycle

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

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

Data Cycle

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

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

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

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

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

Name		Domain		Range	
# modes	::	(Table , String) table-name column	->	List	
modes(animals-table, "pounds")					
# modified-box-plot	::	(Table , String) table-name column	->	Image	
modified-box-plot(animals-table, "pounds")					
# multi-bar-chart	::	(Table , String , String) table-name group subgroup	->	Image	
multi-bar-chart(animals-table, "species", "sex")					
# overlay	::	(Image , Image) top bottom	->	Image	
overlay(circle(10, "solid", "black"), square(50, "solid", "red"))					
# pie-chart	::	(Table , String) table-name column	->	Image	
pie-chart(animals-table, "species")					
# pie-chart-summarized	::	(Table , String , String) table-name labels values	->	Image	
pie-chart-summarized(count(animals-table, "species"), "value", "count")					
# put-image	::	(Image , Number , Number , Image) front x-coordinate y-coordinate behind	->	Image	
put-image(circle(10, "solid", "black"), 10, 10, square(50, "solid", "red"))					
# random-rows	::	(Table , Number) table-name num-rows	->	Table	
random-rows(animals-table, 10) # select 10 random rows from the table					
# rotate	::	(Number , Image) degrees img	->	Image	
rotate(45, star(50, "solid", "dark-blue"))					
# scale	::	(Number , Image) factor img	->	Image	
scale(1/2, star(50, "solid", "light-blue"))					
# scatter-plot	::	(Table , String , String , String) table-name labels xs ys	->	Image	
scatter-plot(animals-table, "name", "pounds","weeks")					
# sort	::	(Table , String , Boolean) table-name column ascending	->	Table	
sort(animals-table, "species", true)					
# stacked-bar-chart	::	(Table , String , String) table-name group subgroup	->	Image	
stacked-bar-chart(animals-table, "species", "sex")					
# stdev	::	(Table , String) table-name column	->	Number	
stdev(animals-table, "pounds")					
# string-contains	::	(String , String) haystack needle	->	Boolean	
string-contains("hotdog", "dog")					