Name: _____

Data Science

Fall 2024 Student Workbook - 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

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

Pioneers in Computing and Mathematics

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

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

Notice and Wonder

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

What do you Notice?	What do you Wonder?

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

   _____.

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

   _____.

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)](), 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 Animals Starter File and hit "Enter". What do you learn? _____

2) We know that all functions need an open parenthesis and at least one input! Type `sort(animals-table)` in the Interactions Area and hit Enter/return. Read the error message. What hint does it give us about how to use this function?

_____

What Kind of Error is it?

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

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

_____

Finding Mistakes with Error Messages

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contracts for Image-Producing Functions

Log into [code.pyret.org (CPO)](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`

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

Sketch a bar chart below.

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

This kind of display tells us...

_____

_____

_____

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

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

Sketch a pie chart below.

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

This kind of display tells us...

_____

_____

_____

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

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

Sketch a box plot below.

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

This kind of display tells us...

_____

_____

_____

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

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

Sketch a histogram below.

Histograms summarize 1 column of _____ data.
categorical/quantitative

This kind of display tells us...

_____

_____

_____

Circles of Evaluation: Composing Functions to Make Displays

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

```
# pie-chart :: Table, String -> Image
# bar-chart :: Table, String -> Image
# histogram :: Table, String, String, Number -> Image
# box-plot :: Table, String -> Image
# first-n-rows :: Table, Number -> Table
# 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?

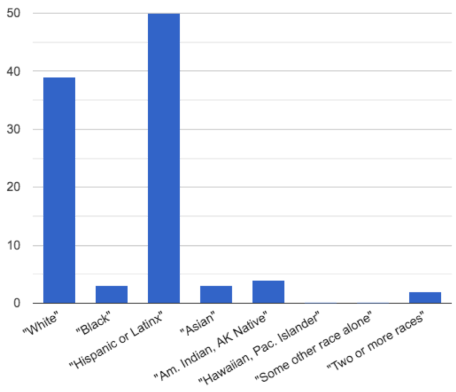
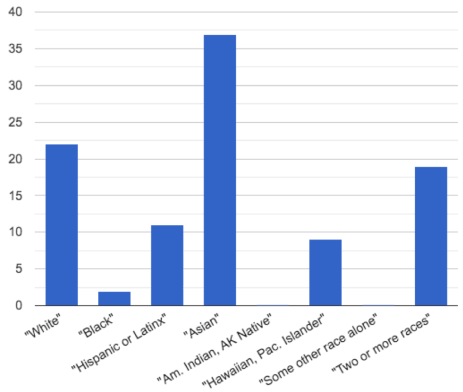
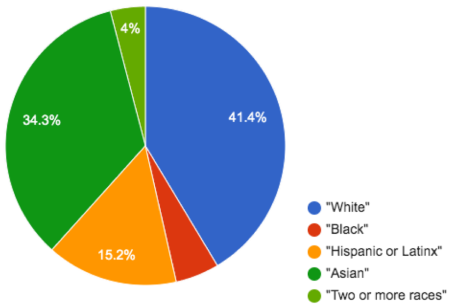
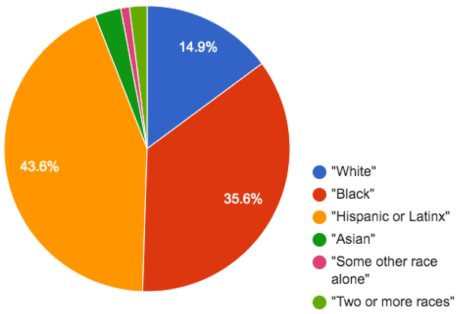
9) Make a `pie-chart` and `bar-chart` for the `pounds` column. Why isn't grouping the `pounds` column very useful?

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?

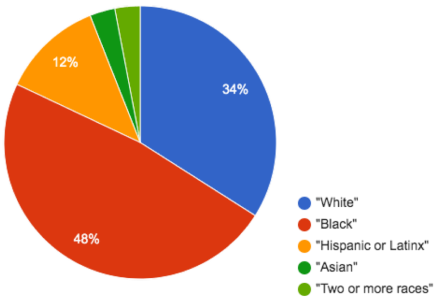
Albuquerque Public Schools, NM	Hawaii DOE, HI	San Francisco Unified School District, CA	Hartford School District, CT
			

What do you Notice?	What do you Wonder?

Matching Bar and Pie Charts

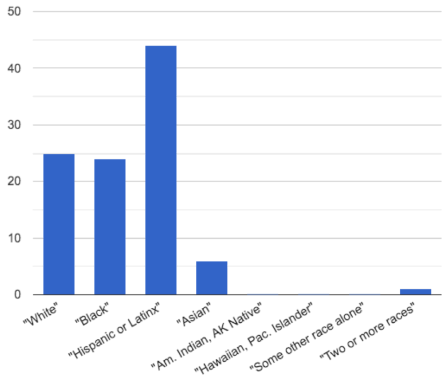
Match each bar chart below to the pie chart that displays the racial demographic data from the same school district.

Cleveland Municipal School District

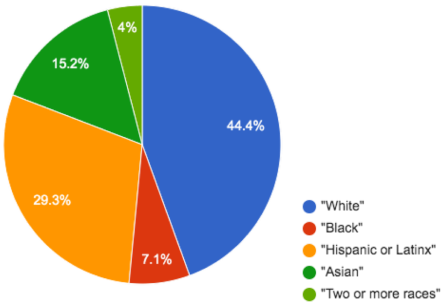


1

A

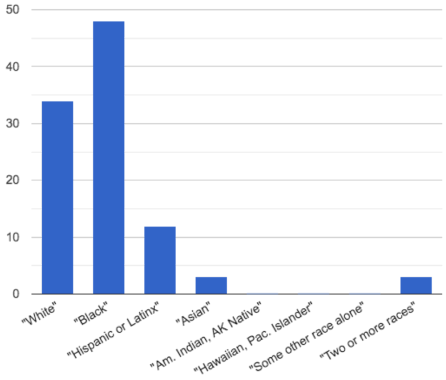


San Diego City Unified School District

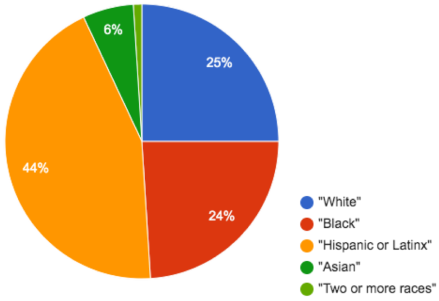


2

B

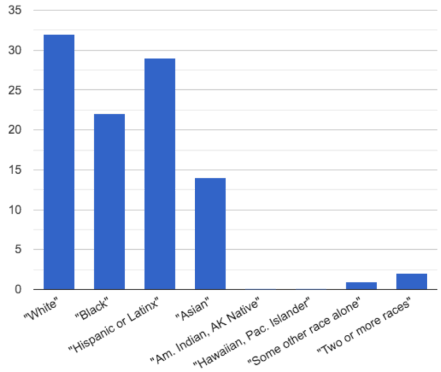


Houston Independent School District

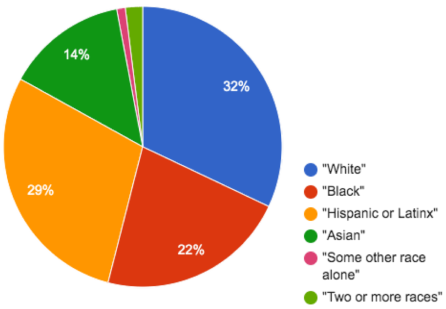


3

C

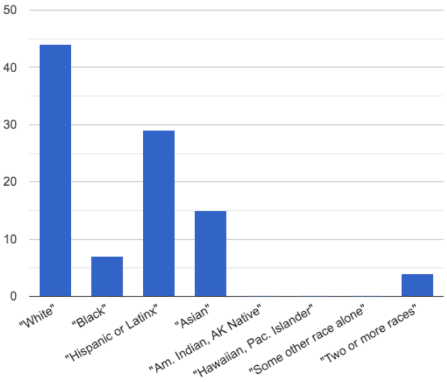


New York City Dept of Education



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. _____

_____

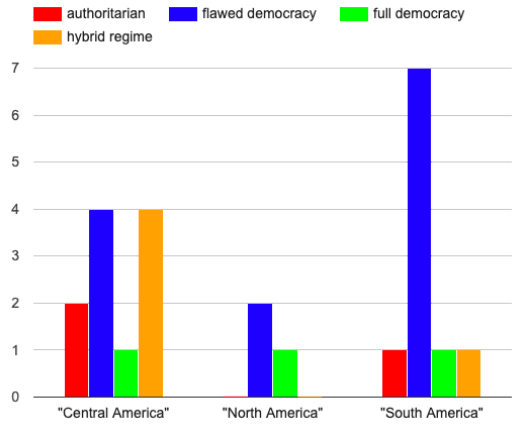
What did you learn from making this display? _____

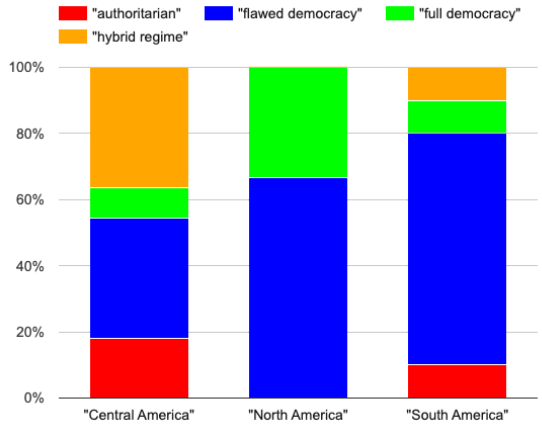
Multi Bar & Stacked Bar Charts - Notice and Wonder

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

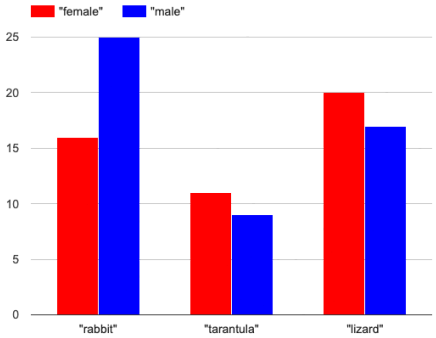
The displays on the right are called **stacked barcharts**.

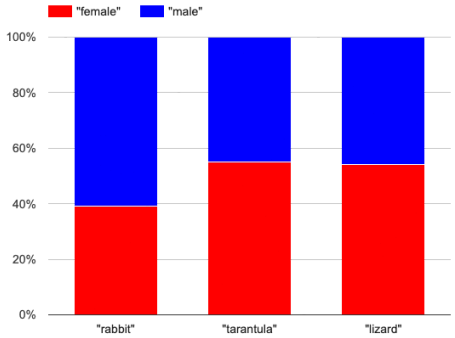
Political Structure of Countries in the Americas





Distribution of sexes across Rabbits, Tarantulas and Lizards





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.

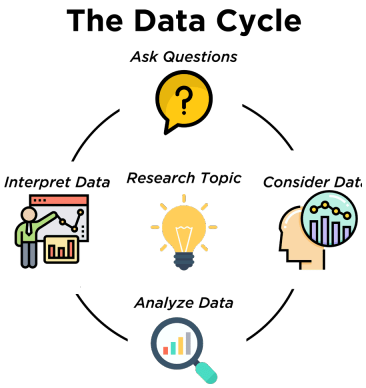
3) Write a question that it would be easiest to answer by looking at one of the stacked bar charts.

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!



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

- Next, we **Consider Data**, by determining which parts of the data set we need to answer our question. Sometimes we don't have the data we need, so we conduct a survey, observe and record data, or find another existing dataset. Since our data is contained in a table, it's useful to start by asking two questions:
  - What rows do we care about? - Is it all the animals? Just the lizards?
  - What columns do we need? - Are we examining the ages of the animals? Their weights?

- Then, we **Analyze the Data**, by completing calculations, creating data displays, creating new tables, or filtering existing tables. The results of this step are calculations, patterns, and relationships.
  - Are we making a pie chart? A bar chart? Something else?

- Finally, we **Interpret the Data**, by answering our original question and summarizing the process we took and the results we found. Sometimes the data cycle ends here, but often these interpretations lead to new questions… and the cycle begins again.

Which Question Type?

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

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

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

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

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

Data Cycle: Consider Data

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

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 _____ (more / less / about the same number of) fixed animals _____ (as / than) unfixed animals. Some new questions this raises include: _____ _____ _____	

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

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

Data Cycle: 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 _____ (variable) is/are _____ I notice that _____ I wonder _____ • How does the distribution of _____ (variable) differ by _____ (variable)? • _____ Another question I have is…	

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

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

Probability, Inference, and Sample Size

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

_____

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

_____

Data Cycle: Categorical Data

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

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

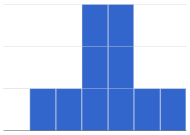
- Bar charts provide a visual representation of the frequency of values in a **categorical** column.
- There's no strict numerical way to order these bars.
  - The count of red, yellow and blue balloons would make sense no matter what order they get presented in.
  - But **sometimes there's an order 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!

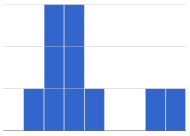
The **shape** of a dataset tells us which values are more or less common.

- 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.



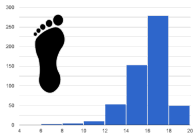
- Some extreme values may be far greater or far lower than the other values in a dataset. These extreme values are called **outliers**.
- A dataset that is **skewed left** has a few values that are unusually low. The histogram for a skewed left dataset has a few data points that are stretched out to 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.

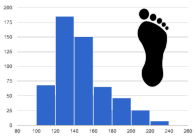


- One way to visualize the difference between a histogram of data that is **skewed left** or **skewed right** is to think about the lengths of our toes on our left and right feet.

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.

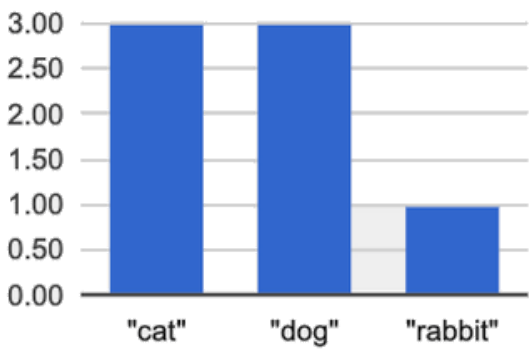


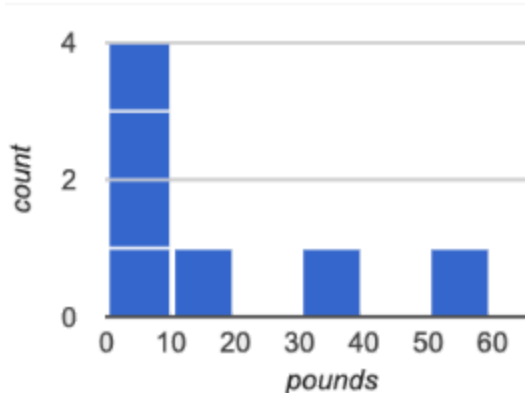
Summarizing Columns with Bar Charts & Histograms

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?





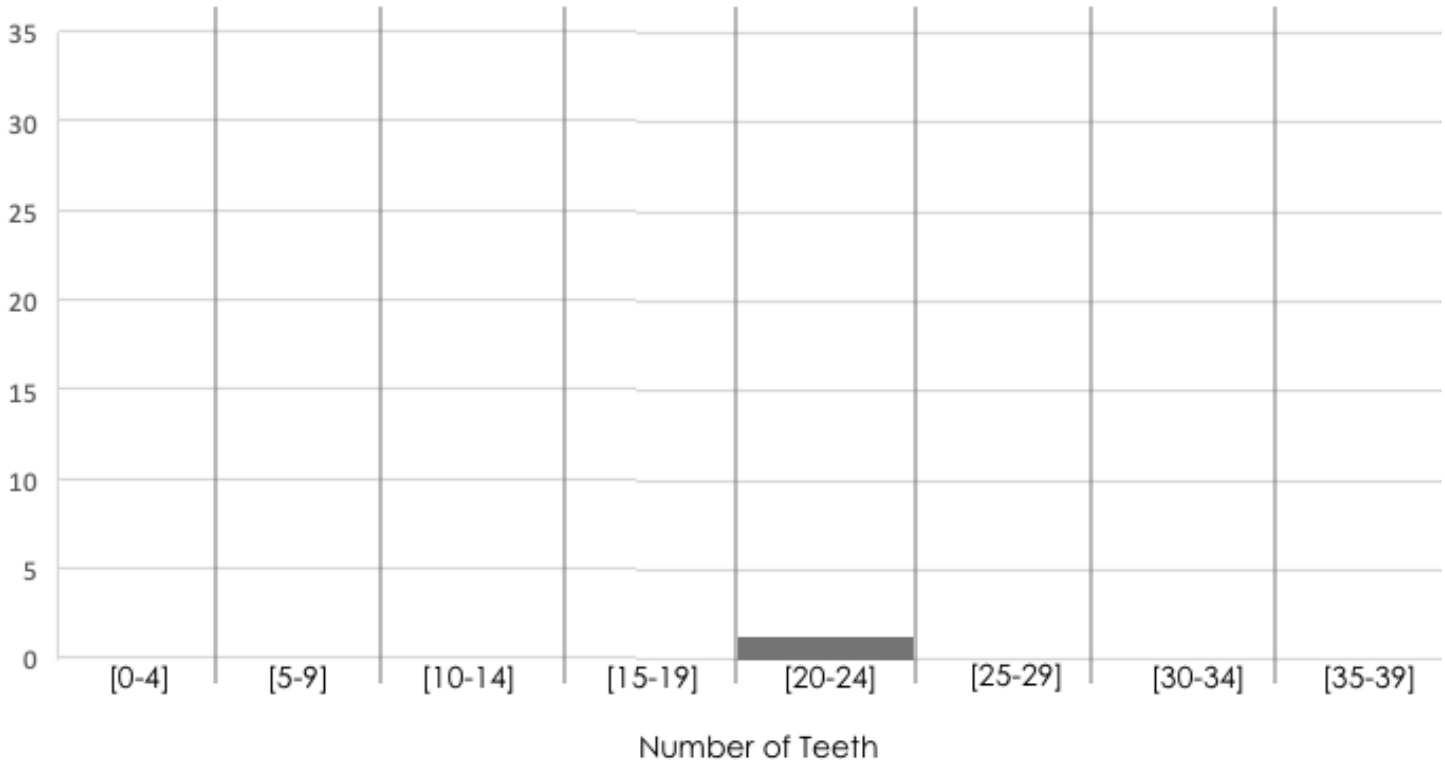
Similarities	Differences

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.



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).

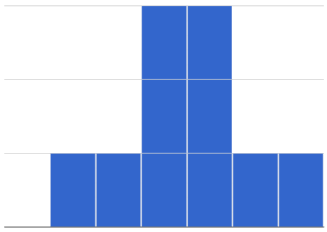
Match the summary description (left) with the *shape* of the histogram of student ratings (right).

- 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

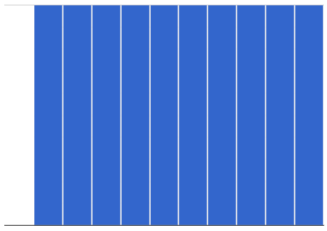
A



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

2

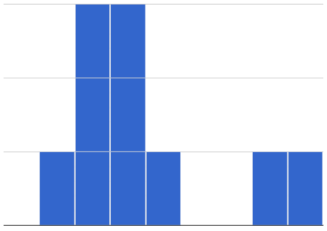
B



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

3

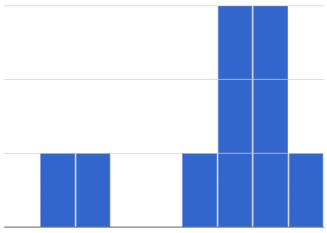
C



Students either really liked or really disliked the video.

4

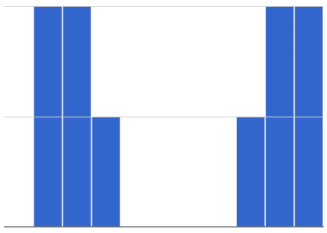
D



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 _____ (x-variable in context) from _____ (dataset or subset).

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

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 _____ (x-variable in context) from _____ (dataset or subset).

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

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

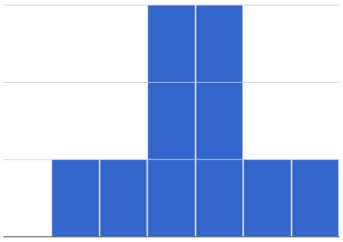
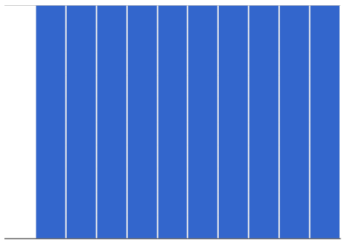
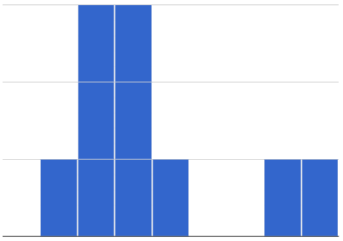
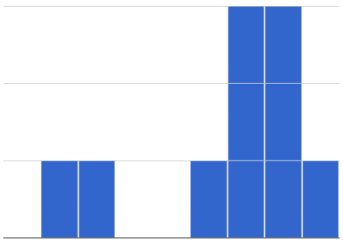
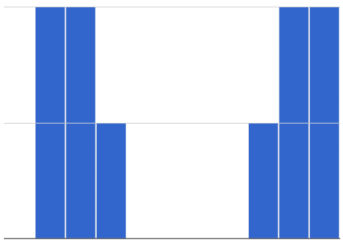
Use the Data Cycle to explore the distribution of one or more quantitative columns from your chosen dataset using **histograms**, and write down your findings.

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

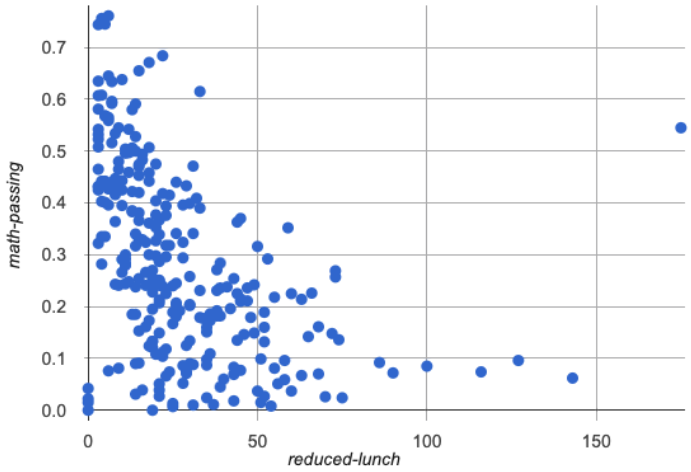
Describe two **histograms** made from columns of the animals dataset.
The first question is provided. You'll need to come up with the second question on your own!

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

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

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.

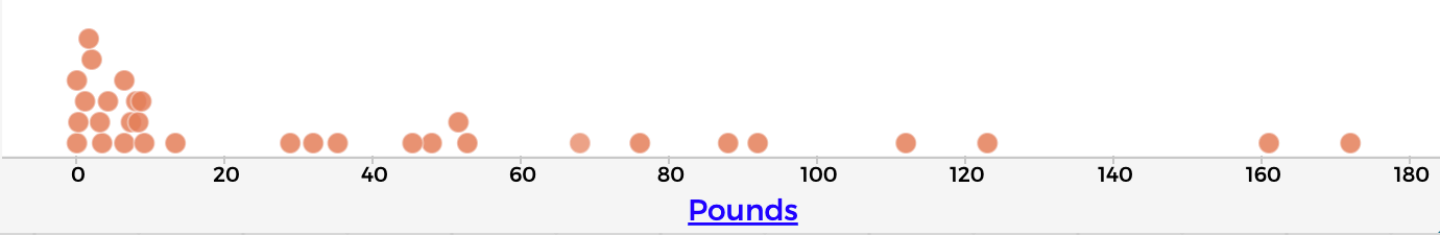
- Each of these measures of center summarizes a whole column of quantitative data using just one number:
  - The **mean** of a dataset is the average of all the numbers.
  - The **median** of a dataset is a value that is smaller than half the dataset, and larger than the other half. In an ordered list 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 animals at the shelter?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	If you only need some rows, 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.

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?	

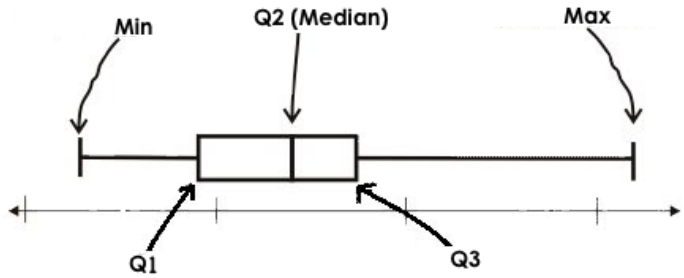
Ask Questions	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	If you only need some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

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: _____.

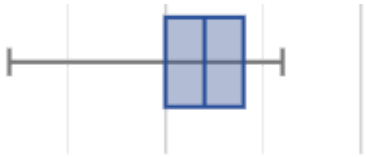
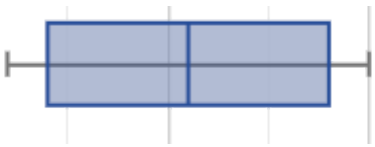
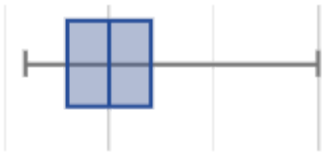
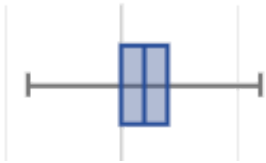
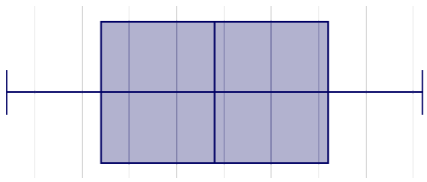
8) From this summary and box plot, I conclude that:

_____

_____

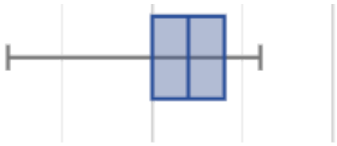
Identifying Shape - Box Plots

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		

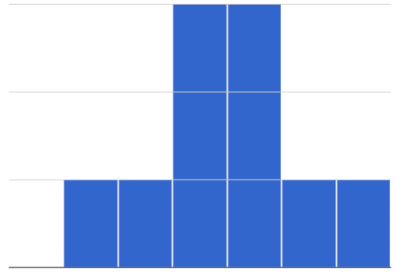
Matching Box Plots to Histograms

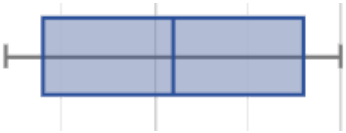
Students watched 5 videos, and rated them on a scale of 1 to 10. For each video, their ratings were used to generate box plots and histograms. Match each box plot to the histogram that displays the same data.



1

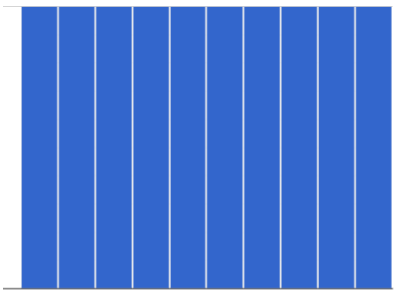
A

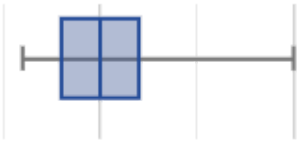




2

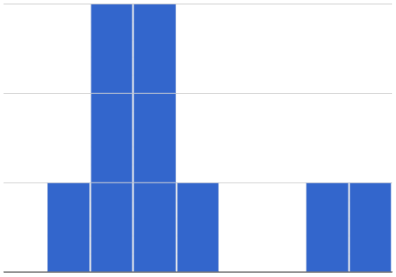
B

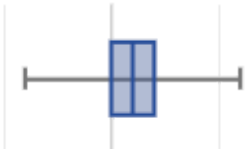




3

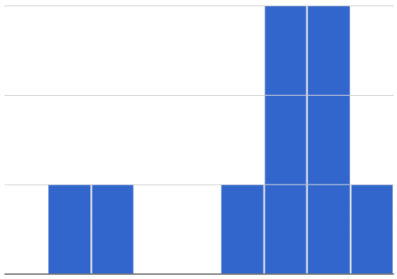
C

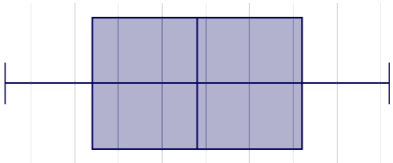




4

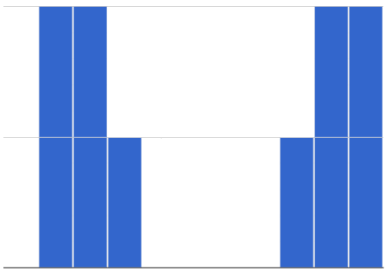
D





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?	

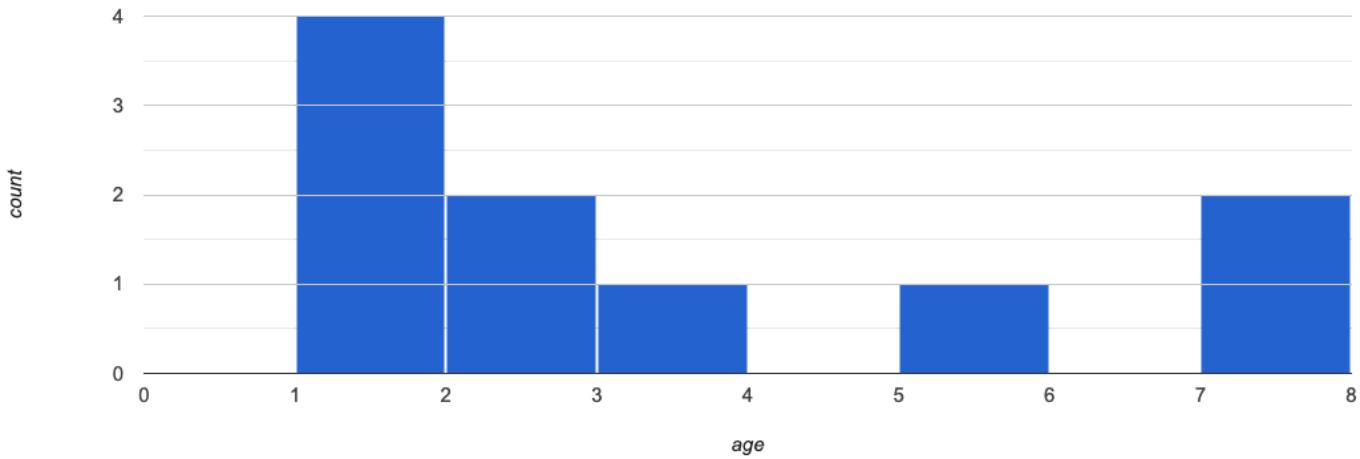
Ask Questions	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	If you only need some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

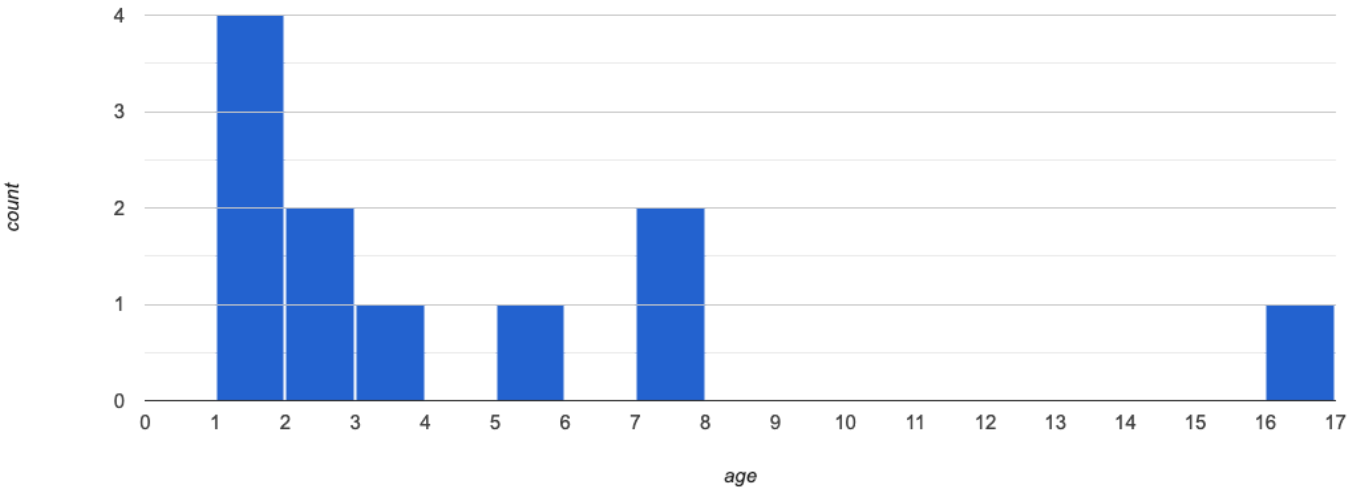
8) Add all the squared distances. What is their sum? _____

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 $N\text{-}1$. 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.

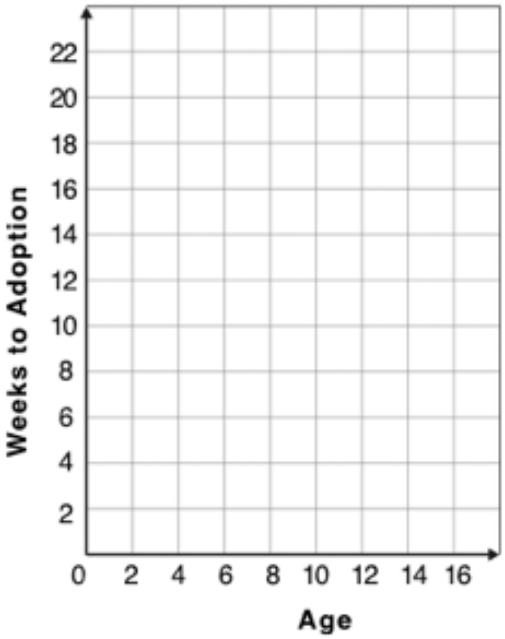
    Note: These cutoffs are not an exact science! In some contexts an r-value of ±0.50 might be considered impressively strong!

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? _____

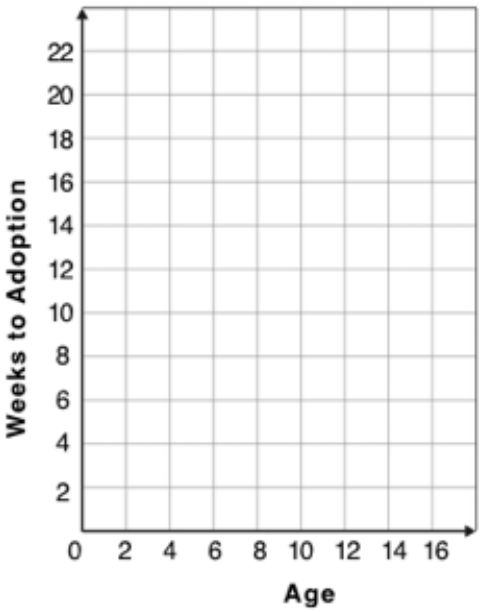
4) Use a straight edge to draw a line on the graph that best represents the pattern you see, then circle the cloud of points around that line.

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.

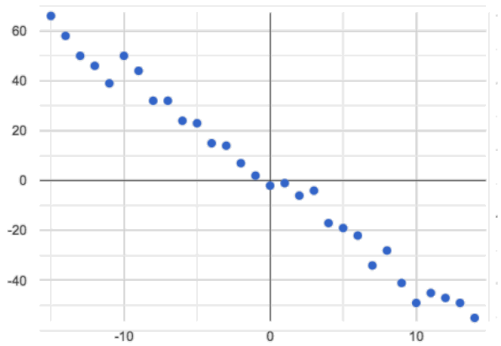
Ask Questions	What question do you have? _____	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) _____ What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.) _____	
Analyze Data	What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

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

Identifying Form, Direction and Strength

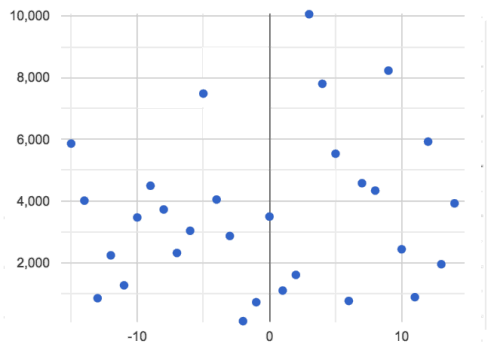
What do your eyes tell you about the Form, Direction, & Strength of these displays?
Note: If the form is nonlinear, we shouldn't report direction - a curve may rise and then fall.

A



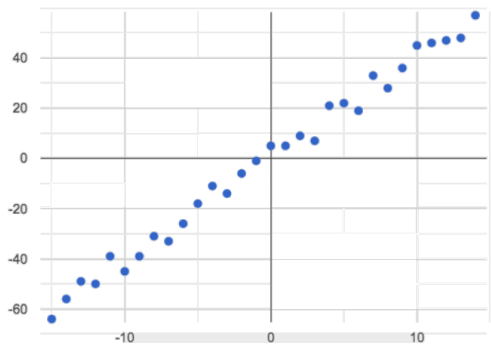
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

B



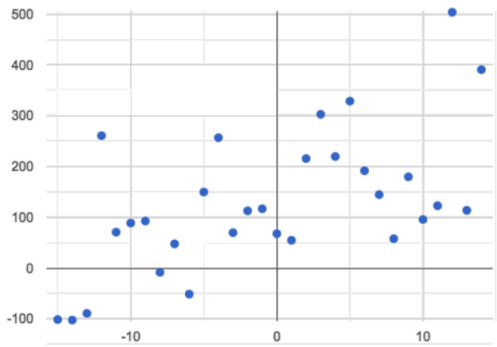
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

C



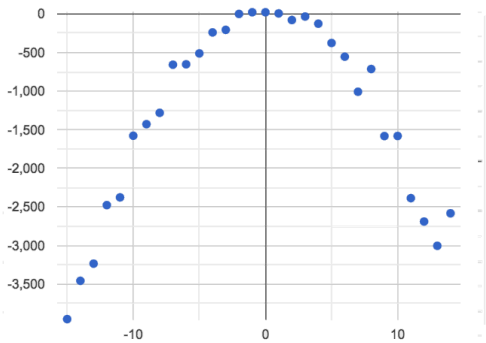
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

D



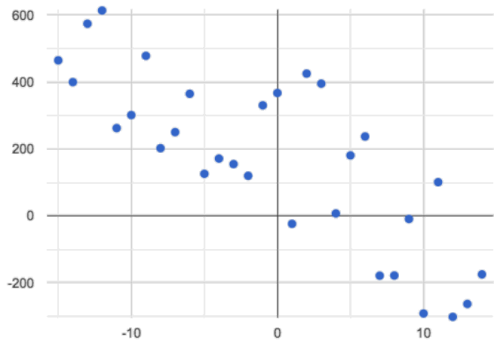
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

E



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

F



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

Reflection on Form, Direction and Strength

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

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

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

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

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

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

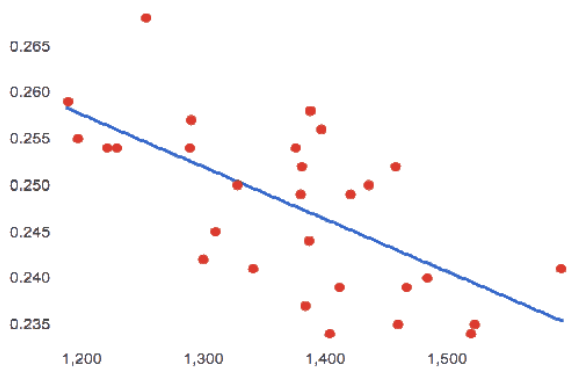
Identifying Form and r-Values

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

Reminder:

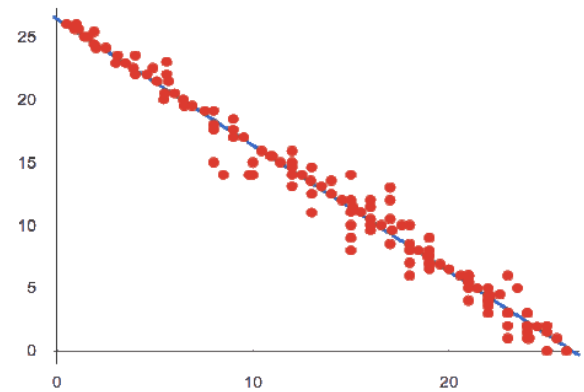
- −1 is the strongest possible *negative* correlation, and +1 is the strongest possible *positive* correlation
- 0 means no correlation
- ±0.65 or ±0.70 or more is typically considered a "strong correlation"
- ±0.35 to ±0.65 is typically considered "moderately correlated"
- Anything less than about ±0.25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

A



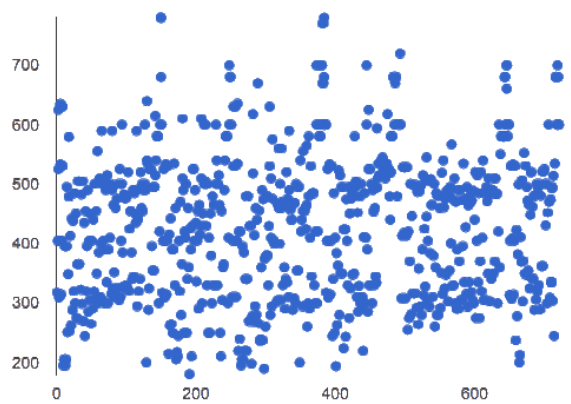
Form:
r close to:

B



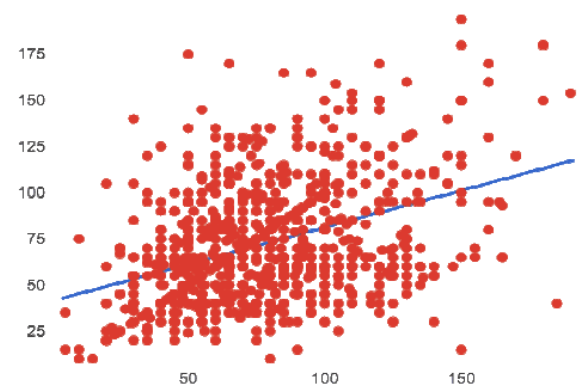
Form:
r close to:

C



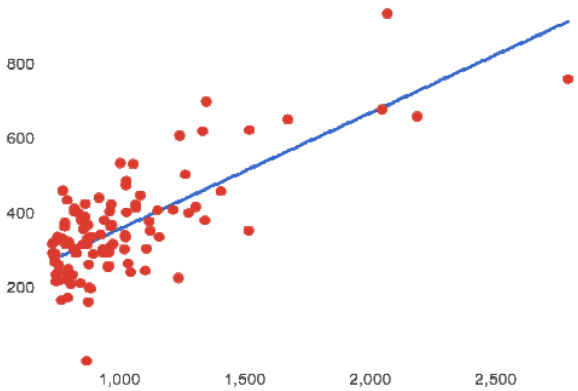
Form:
r close to:

D



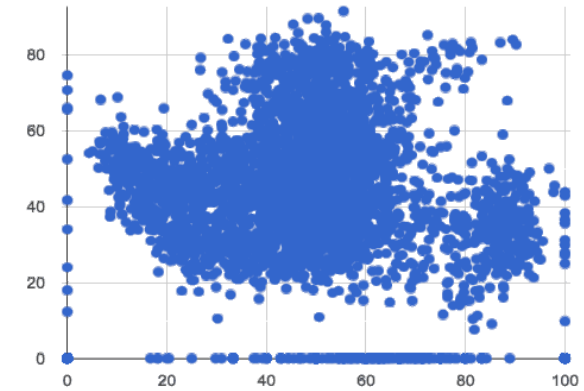
Form:
r close to:

E



Form:
r close to:

F



Form:
r close to:

Correlation Does Not Imply Causation!

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

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

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

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

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

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

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

Correlations in the Animals Dataset

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

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

_____

_____

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

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

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

4) What does that *mean*? _____

_____

_____

_____

Correlations in My Dataset

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

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

because _____

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

---

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

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

because _____

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

---

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

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

because _____

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

---

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

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

because _____

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

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? _____

_____

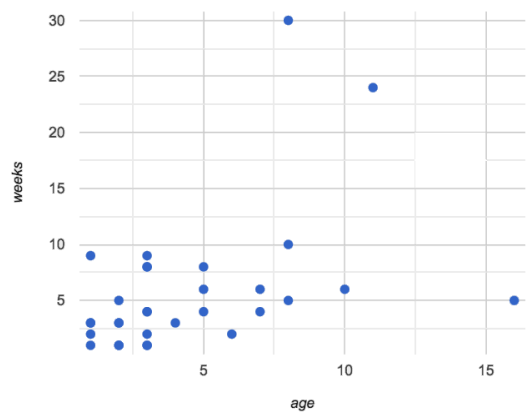
5) What is the highest r-value you can get? _____ Where did you place P? (_____, _____)

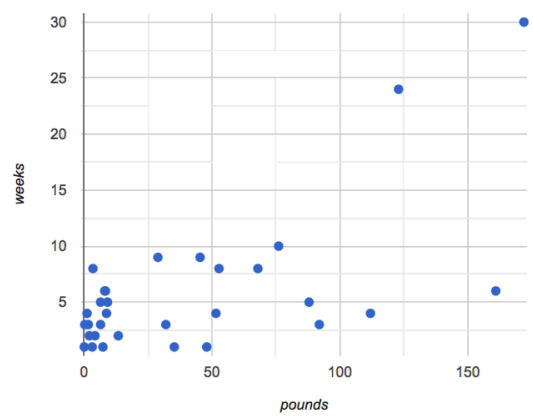
6) What function describes the regression line with this value of P? $y =$ _____ $x +$ _____

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?

_____

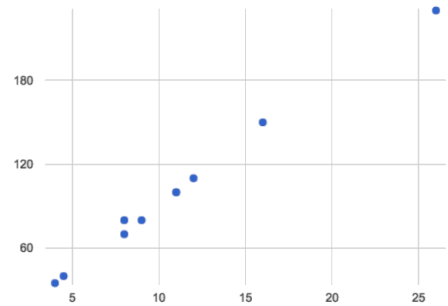
Drawing Predictors

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

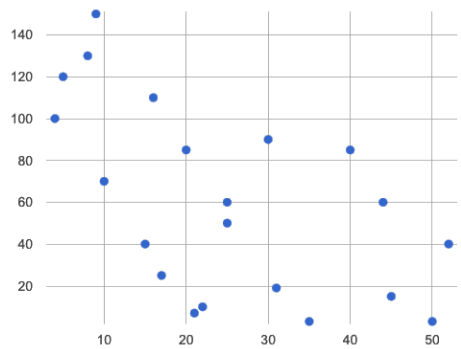


Direction: Positive Negative None

Strength: Stronger Weaker

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

B

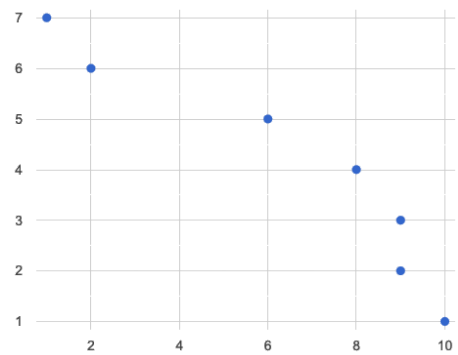


Direction: Positive Negative None

Strength: Stronger Weaker

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

C

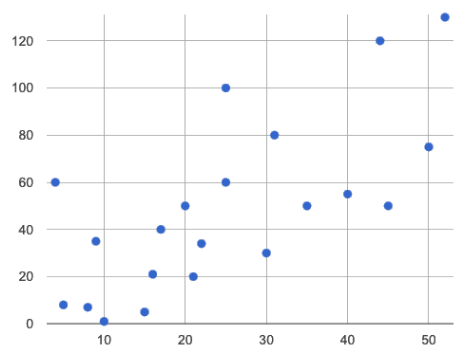


Direction: Positive Negative None

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* [*Animals Starter File*]*.*

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*? _____

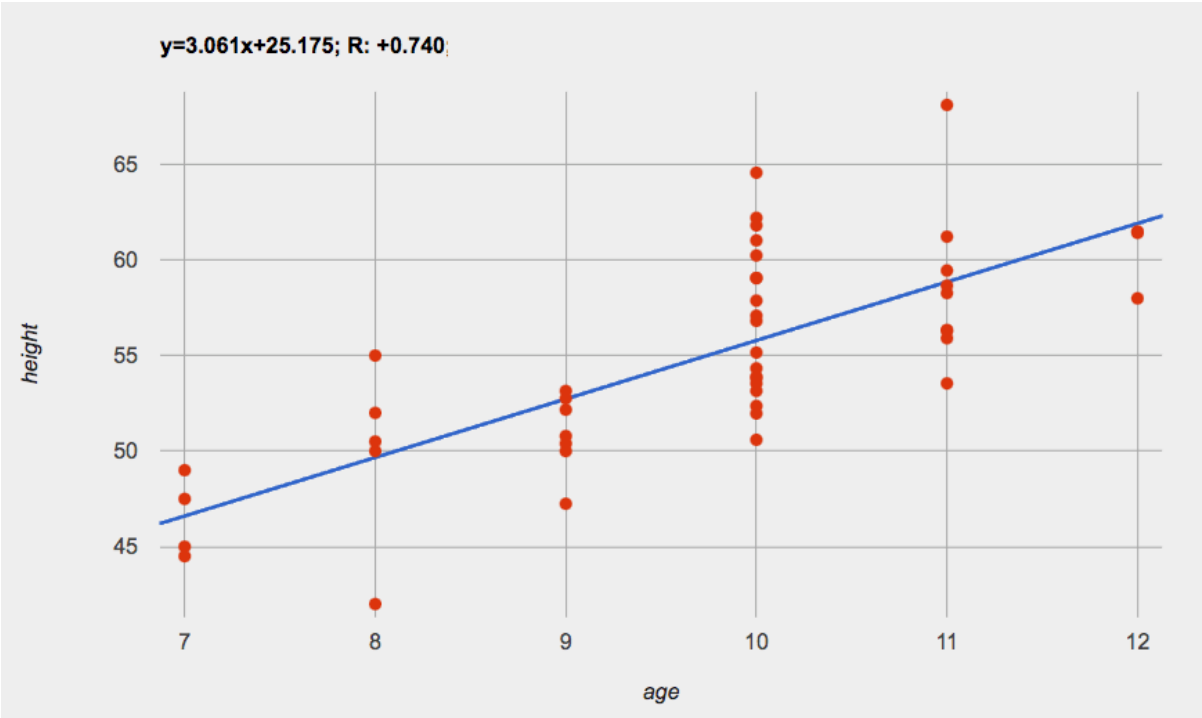
_____

_____

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



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

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___ [dataset or subset] and found a ___moderate (R=.448), positive___ [weak / strong / moderate (R=...), positive / negative] correlation between ___age___ [x-axis] and ___time to adoption___ [y-axis]. I would predict that a 1 ___year___ [x-axis units] increase in ___age___ [x-axis] is associated with a ___.789 week___ [slope, y-units] ___increase___ [increase / decrease] in ___time to adoption___ [y-axis].

Ask Questions

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What code will make the table or display you want?

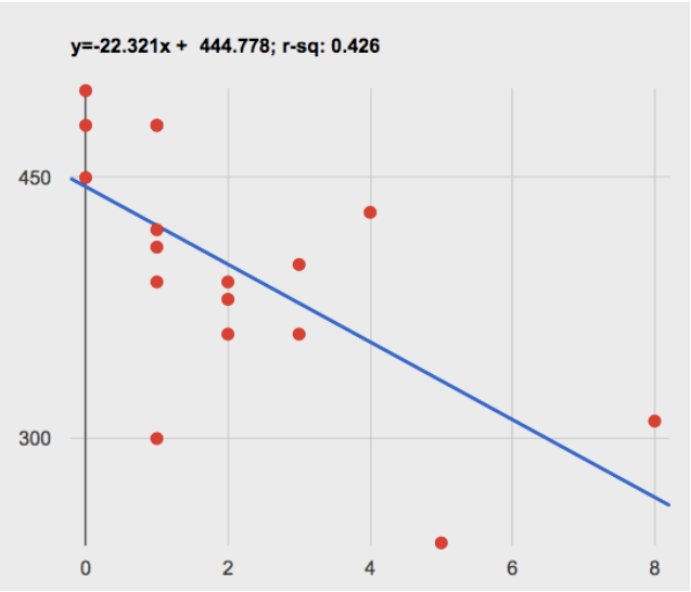
Interpret Data

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].

Describing Relationships

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

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



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

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

Data Cycle: Regression Analysis

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

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		

Looking up Rows and Columns

We can define names for values in Pyret, the same way we do in math:

```
name = "Shanti"
age = 16
logo = star(50, "solid", "red")
```

When **looking up a data Row** from a Table, programmers use the `row-n` function. This function takes a Table and a Number as its inputs. The numbers tell the computer which Row we want from the Table. *Note: Rows are numbered starting at zero!*
For example:

```
sasha   = row-n(animals-table, 0)  # define Sasha to be the first row
mittens = row-n(animals-table, 2)  # define Mittens to be the third row
```

When we define these rows, it's more useful to name them based on their *properties* , rather than their identifiers:

```
cat-row = row-n(animals-table,  0) # Sasha is a cat
dog-row = row-n(animals-table, 10) # Toggle is a dog
```

When **looking up a column** from a Row, programmers use square brackets and the name of the column they want.
For example:

```
# these two lines do the same thing! We can use the defined name to simplify our code
row-n(animals-table, 0)["age"] # look up Sasha's age (in row 0)
cat-row["species"]             # look up Sasha's age (using the defined name)
dog-row["age"]                 # look up Toggle's age (using the defined name)
```

Lookup Questions

The table below represents four pets at an animal shelter:

`pets-table`

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

1) *Match* each Lookup Question (left) to the code that will give the answer (right).

"How much does Maple weigh?"	**1**	**A**	`row-n(pets-table, 3)`
"Which is the last row in the table?	**2**	**B**	`row-n(pets-table, 2)["name"]`
"What is Fritz's sex?"	**3**	**C**	`row-n(pets-table, 1)["sex"]`
"What's the third animal's name?"	**4**	**D**	`row-n(pets-table, 3)["age"]`
"How much does Nori weigh?"	**5**	**E**	`row-n(pets-table, 3)["pounds"]`
"How old is Maple?"	**6**	**F**	`row-n(pets-table, 0)`
"What is Toggle's sex?"	**7**	**G**	`row-n(pets-table, 2)["pounds"]`
"What is the first row in the table?"	**8**	**H**	`row-n(pets-table, 0)["sex"]`

2) For each value on the left, write the Pyret expression that will produce that value on the right. The first one has been completed for you.

a.	"Maple"	row-n(pets-table, 3)["name"]
b.	"male"	
c.	4	
d.	48	
e.	"Nori"	

More Practice with Lookups

Consider `shapes-table` below, and the four value definitions that follow.

name	corners	is-round
`"triangle"`	`3`	`false`
`"square"`	`4`	`false`
`"rectangle"`	`4`	`false`
`"circle"`	`0`	`true`

```
shapeA = row-n(shapes-table, 0)
shapeB = row-n(shapes-table, 1)
shapeC = row-n(shapes-table, 2)
shapeD = row-n(shapes-table, 3)
```

1) *Match* each Pyret expression (left) to the description of what it evaluates to (right).

`shapeD`	**1**	**A**	Evaluates to 4
`shapeA`	**2**	**B**	Evaluates to the last row in the table
`shapeB["corners"]`	**3**	**C**	Evaluates to `"square"`
`shapeC["is-round"]`	**4**	**D**	Evaluates to `true`
`shapeB["name"]`	**5**	**E**	Evaluates to `false`
`shapeA["corners"]`	**6**	**F**	Evaluates to 3
`shapeD["name"] == "circle"`	**7**	**G**	Evaluates to the first row in the table

2) For each value on the left, write the Pyret expression that will produce that value on the right. The first one has been completed for you.

a.	`"rectangle"`	`shapeC["name"]`
b.	`"square"`	
c.	`4`	
d.	`0`	
e.	`true`	

Defining Rows

Remember: rows start at index zero!
We've already given you two row definitions, `cat-row` and `dog-row`:

```
cat-row = row-n(animals-table,  0)  # Sasha is a cat
dog-row = row-n(animals-table, 10)  # Toggle is a dog
```

1) Use the Animals Table to identify the index of a row containing...

a lizard _____

a rabbit _____

a fixed animal _____

a male animal _____

a female animal _____

a hermaphroditic animal _____

an unfixed animal _____

a young animal (<2 years) _____

an old animal (>10 years) _____

2) What code would you write to define `lizard-row`?

_____

3) What code would you write to define `rabbit-row`?

_____

4) What code would you write to define `fixed-row`?

_____

5) What code would you write to define `male-row`?

_____

6) What code would you write to define `female-row`?

_____

7) What code would you write to define `hermaphrodite-row`?

_____

8) What code would you write to define `young-row`?

_____

9) What code would you write to define `old-row`?

_____

Add this code to your Animals Starter File! You'll want these rows for later!

Defining Functions

Functions can be viewed in *multiple representations*. You already know one of them: ***Contracts***, which specify the Name, Domain, and Range of a function. Contracts are a way of thinking of functions as a *mapping* between one set of data and another. For example, a mapping from Numbers to Strings:

```
# f :: Number -> String
```

Another way to view functions is with ***Examples***. Examples are essentially input-output tables, showing what the function would do for a specific input:

How f is used	What f does
f(1)	$1 + 2$
f(2)	$2 + 2$
f(3)	$3 + 2$
f(4)	$4 + 2$

In our programming language, we focus on the last two columns and write them as code:

```
examples:
  f(1) is 1 + 2
  f(2) is 2 + 2
  f(3) is 3 + 2
  f(4) is 4 + 2
end
```

Finally, we write a formal **function definition** ourselves. The pattern in the Examples becomes *abstract* (or "general"), replacing the inputs with ***variables***. In the example below, the same definition is written in both math and code:

$$f(x) = x + 2$$

```
fun f(x): x + 2 end
```

Look for connections between these three representations!

- The function name is always the same, whether looking at the Contract, Examples, or Definition.
- The number of inputs in the Examples is always the same as the number of types in the Domain, which is always the same as the number of variables in the Definition.
- The "what the function does" pattern in the Examples is almost the same in the Definition, but with specific inputs replaced by variables.

The Great gt domain debate!

Kermit: The domain of `gt` is `Number, String, String`.

Oscar: The domain of `gt` is `Number`.

Ernie: I'm not sure who's right!

In order to make a triangle, we need a size, a color and a fill style…
but all we had to tell our actor was `gt(20)` …and they returned `triangle(20, "solid", "green")`.

Please help us!

1) What is the correct domain for gt?

2) What could you tell Ernie to help him understand how you know?

Let's Define Some New Functions!

1) **Let's define a function `rs` to generate solid red squares of whatever size we give them!**
If I say `rs(5)`, what would our actor need to say?

_____

Let's write a few more examples:

`rs(____)` → _____

`rs(____)` → _____

`rs(____)` → _____

What changes in these examples? Name your variable(s): _____
Let's define our function using the variable:

`fun rs(_____):` _____ `end`

2) **Let's define a function `bigc` to generate big solid circles of size 100 in whatever color we give them!**
If I say `bigc("orange")`, what would our actor need to say?

_____

Let's write a few more examples:

`bigc(_____)` → _____

`bigc(_____)` → _____

`bigc(_____)` → _____

What changes in these examples? Name your variable(s): _____
Let's define our function using the variable:

`fun bigc(_____):` _____ `end`

3) **Let's define a function `ps` to build a pink star of size 50, with the input determining whether it's solid or outline!**
If I say `ps("outline")`, what would our actor need to say?

_____

Write examples for all other possible inputs:

`ps(_____)` → _____

`ps(_____)` → _____

What changes in these examples? Name your variable(s): _____
Let's define our function using the variable:

`fun ps(_____):` _____ `end`

4) Add these new function definitions to your gt Starter File and test them out!

Let's Define Some More New Functions!

1) **Let's define a function `sun` to write SUNSHINE in whatever color and size we give it!**

If I say `sun(5, "blue")`, what would our actor need to say?

_____

Let's write a few more examples:

`sun(____, _____)` → _____

`sun(____, _____)` → _____

`sun(____, _____)` → _____

What changes in these examples? Name your variable(s): _____

Let's define our function using the variable(s):

```
fun sun(_____, _____):
_____
end
```

2) **Let's define a function `me` to generate your name in whatever size and color we give it!**

If I say `me(18, "gold")`, what would our actor need to say?

_____

Let's write a few more examples:

`me(____, _____)` → _____

`me(____, _____)` → _____

`me(____, _____)` → _____

What changes in these examples? Name your variable(s): _____

Let's define our function using the variable(s):

```
fun me(_____, _____):
_____
end
```

3) **Let's define a function `gr` to build a solid, green rectangle of whatever height and width we give it!**

If I say `gr(10, 80)`, what would our actor need to say?

_____

Let's write a few more examples:

`gr(____, ____)` → `rectangle(____, ____, "solid", "green")`

`gr(____, ____)` → `rectangle(____, ____, "solid", "green")`

`gr(____, ____)` → `rectangle(____, ____, "solid", "green")`

What changes in these examples? Name your variable(s): _____

Let's define our function using the variable(s):

```
fun gr(_____, _____):
_____
end
```

4) Add these new function definitions to your gt Starter File and test them out!

Describe and Define Your Own Functions!

1) **Let's define a function** _____ **to generate...**

_____

If I say _____, what would our actor need to say? _____

Let's write a few more examples:

_____(_____) → _____(_____)

_____(_____) → _____(_____)

_____(_____) → _____(_____)

What changes in these examples? Name your variable(s): _____

Let's define our function using the variable.

```
fun _____(_____): _____ end
```

2) **Let's define a function** _____ **to generate...**

_____

If I say _____, what would our actor need to say? _____

Let's write a few more examples:

_____(_____) → _____(_____)

_____(_____) → _____(_____)

_____(_____) → _____(_____)

What changes in these examples? Name your variable(s): _____

Let's define our function using the variable.

```
fun _____(_____): _____ end
```

3) **Let's define a function** _____ **to generate...**

_____

If I say _____, what would our actor need to say? _____

Let's write a few more examples:

_____(_____) → _____(_____)

_____(_____) → _____(_____)

_____(_____) → _____(_____)

What changes in these examples? Name your variable(s): _____

Let's define our function using the variable.

```
fun _____(_____): _____ end
```

4) Add your new function definitions to your gt Starter File and test them out!

Matching Examples and Contracts

Match each set of examples (left) with the Contract that best describes it (right).

Examples			Contract

`examples:`
`  f(5) is 5 / 2`
`  f(9) is 9 / 2`
`  f(24) is 24 / 2`
`end`
	1	A	`# f :: Number -> Number`

`examples:`
`  f(1) is rectangle(1, 1, "outline", "red")`
`  f(6) is rectangle(6, 6, "outline", "red")`
`end`
	2	B	`# f :: String -> Image`

`examples:`
`  f("pink", 5) is star(5, "solid", "pink")`
`  f("blue", 8) is star(8, "solid", "blue")`
`end`
	3	C	`# f :: Number -> Image`

`examples:`
`  f("Hi!") is text("Hi!", 50, "red")`
`  f("Ciao!") is text("Ciao!", 50, "red")`
`end`
	4	D	`# f :: Number, String -> Image`

`examples:`
`  f(5, "outline") is star(5, "outline", "yellow")`
`  f(5, "solid") is star(5, "solid", "yellow")`
`end`
	5	E	`# f :: String, Number -> Image`

Matching Examples and Function Definitions

(1) Find the variables in `gt` and label them with the word "size".

```
examples:
    gt(20) is triangle(20, "solid", "green")
    gt(50) is triangle(50, "solid", "green")
end
fun gt(size): triangle(size, "solid", "green") end
```

(2) Highlight and label the variables in the example lists below.

(3) Then, using `gt` as a model, match the examples to their corresponding function definitions.

Examples			Definition

`examples:`
`  f("solid") is circle(8, "solid", "red")`
`  f("outline") is circle(8, "outline", "red")`
`end`
	1	A	`fun f(s): star(s, "outline", "red") end`

`examples:`
`  f(2) is 2 + 2`
`  f(4) is 4 + 4`
`  f(5) is 5 + 5`
`end`
	2	B	`fun f(num): num + num end`

`examples:`
`  f("red") is circle(7, "solid", "red")`
`  f("teal") is circle(7, "solid", "teal")`
`end`
	3	C	`fun f(c): star(9, "solid", c) end`

`examples:`
`  f("red") is star(9, "solid", "red")`
`  f("grey") is star(9, "solid", "grey")`
`  f("pink") is star(9, "solid", "pink")`
`end`
	4	D	`fun f(s): circle(8, s, "red") end`

`examples:`
`  f(3) is star(3, "outline", "red")`
`  f(8) is star(8, "outline", "red")`
`end`
	5	E	`fun f(c): circle(7, "solid", c) end`

Creating Contracts From Examples

Write the contracts used to create each of the following collections of examples. The first one has been done for you.

1) `# big-triangle :: Number, String -> Image`

```
examples:
  big-triangle(100, "red") is triangle(100, "solid", "red")
  big-triangle(200, "orange") is triangle(200, "solid", "orange")
end
```

2) _____

```
examples:
  purple-square(15) is rectangle(15, 15, "outline", "purple")
  purple-square(6) is rectangle(6, 6, "outline", "purple")
end
```

3) _____

```
examples:
  sum(5, 8) is 5 + 8
  sum(9, 6) is 9 + 6
  sum(120, 11) is 120 + 11
end
```

4) _____

```
examples:
  banner("Game Today!") is text("Game Today!", 50, "red")
  banner("Go Team!") is text("Go Team!", 50, "red")
  banner("Exit") is text("Exit", 50, "red")
end
```

5) _____

```
examples:
  twinkle("outline", "red") is star(5, "outline", "red")
  twinkle("solid", "pink") is star(5, "solid", "pink")
  twinkle("outline", "grey") is star(5, "outline", "grey")
end
```

6) _____

```
examples:
  half(5) is 5 / 2
  half(8) is 8 / 2
  half(900) is 900 / 2
end
```

7) _____

```
examples:
  Spanish(5) is "cinco"
  Spanish(30) is "treinta"
  Spanish(12) is "doce"
end
```

Contracts, Examples & Definitions - bc

We've already found the Contract for gt, generated Examples and described the pattern with a Function Definition. Let's review our process, beginning with the Word Problem.

Directions: Define a function called `gt`, which makes solid green triangles of whatever size we want.

Contract and Purpose Statement

Every contract has three parts…

`# gt :: Number -> Image`

function name | Domain | Range

Examples

Write some examples, then circle and label what changes…

examples:

`gt(10)` **is** `triangle(10, "solid", "green")`

function name | input(s) | what the function produces

`gt(20)` **is** `triangle(20, "solid", "green")`

function name | input(s) | what the function produces

end

Definition

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

fun `gt(size)`:

function name | variable(s)

`triangle(size, "solid", "green")`

what the function does with those variable(s)

end

Now, let's apply the same steps to think through a new problem!

Directions: Define a function called `bc`, which makes solid blue circles of whatever radius we want.

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____

function name | Domain | Range

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

Contracts, Examples & Definitions - Stars

Directions: Define a function called `sticker`, which consumes a color and draws a solid 50px star of the given color.

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____
function name — Domain — Range

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: Define a function called `gold-star`, which takes in a radius and draws a solid gold star of that given size.

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____
function name — Domain — Range

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

Contracts, Examples & Definitions - Name

Directions: Define a function called `name-color`, which makes an image of your name at size 50 in whatever color is given.

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____
function name — Domain — Range

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: Define a function called `name-size`, which makes an image of your name in your favorite color (be sure to specify your name and favorite color!) in whatever size is given.

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____
function name — Domain — Range

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

Solving Word Problems

Being able to see functions as Contracts, Examples or Definitions is like having three powerful tools. These representations can be used together to solve word problems! We call this **The Design Recipe**.

1) When reading a word problem, the first step is to figure out the **Contract** for the function you want to build. Remember, a Contract must include the Name, Domain and Range for the function!

2) Then we write a **Purpose Statement**, which is a short note that tells us what the function *should do*. Professional programmers work hard to write good purpose statements, so that other people can understand the code they wrote! Programmers work on teams; the programs they write must outlast the moment that they are written.

3) Next, we write at least two **Examples**. These are lines of code that show what the function should do for a *specific* input. Once we see examples of at least two inputs, we can *find a pattern* and see which parts are changing and which parts aren't.

4) To finish the Examples, we circle the parts that are changing, and label them with a short **variable name** that explains what they do.

5) Finally, we **define the function** itself! This is pretty easy after you have some examples to work from: we copy everything that didn't change, and replace the changeable stuff with the variable name!

Matching Word Problems and Purpose Statements

Match each word problem below to its corresponding purpose statement.

Annie got a new dog, Xavier, that eats about 5 times as much as her little dog, Rex, who is 10 years old. She hasn't gotten used to buying enough dogfood for the household yet. Write a function that generates an estimate for how many pounds of food Xavier will eat, given the amount of food that Rex usually consumes in the same amount of time.	**1**	**A**	Consume the pounds of food Rex eats and add 5.
Adrienne's raccoon, Rex, eats 5 more pounds of food each week than her pet squirrel, Lili, who is 7 years older. Write a function to determine how much Lili eats in a week, given how much Rex eats.	**2**	**B**	Consume the pounds of food Rex eats and subtract 5.
Alejandro's rabbit, Rex, poops about 1/5 of what it eats. His rabbit hutch is 10 cubic feet. Write a function to figure out how much rabbit poop Alejandro will have to clean up depending on how much Rex has eaten.	**3**	**C**	Consume the pounds of food Rex eats and multiply by 5.
Max's turtle, Rex, eats 5 pounds less per week than his turtle, Harry, who is 2 inches taller. Write a function to calculate how much food Harry eats, given the weight of Rex's food.	**4**	**D**	Consume the pounds of food Rex eats and divide by 5.

Writing Examples from Purpose Statements

We've provided contracts and purpose statements to describe two different functions. Write examples for each of those functions.

Contract and Purpose Statement

Every contract has three parts…

`# triple::` `Number` `->` `Number`

function name — Domain — Range

`# Consumes a Number and triples it.`

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

Contract and Purpose Statement

Every contract has three parts…

`# upside-down::` `Image` `->` `Image`

function name — Domain — Range

`# Consumes an image, and turns it upside down by rotating it 180 degrees.`

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

Fixing Purpose Statements

Beneath each of the word problems below is a purpose statement (generated by ChatGPT!) that is either missing information or includes unnecessary information.

- Write an improved version of each purpose statement beneath the original.
- Then, explain what was wrong with the ChatGPT-generated Purpose Statement.

1) **Word Problem:** *The New York City ferry costs $2.75 per ride. The Earth School requires two chaperones for any field trip. Write a function* `fare` *that takes in the number of students in the class and returns the total fare for the students and chaperones.*

ChatGPT's Purpose Statement: Take in the number of students and add 2.

Improved Purpose Statement: _____

_____

Problem with ChatGPT's Purpose Statement: _____

_____

2) **Word Problem:** *It is tradition for the Green Machines to go to Humpy Dumpty's for ice cream with their families after their soccer games. Write a function* `cones` *to take in the number of kids and calculate the total bill for the team, assuming that each kid brings two family members and cones cost $1.25.*

ChatGPT's Purpose Statement: Take in the number of kids on the team and multiply it by 1.25.

Improved Purpose Statement: _____

_____

Problem with ChatGPT's Purpose Statement: _____

_____

3) **Word Problem:** *The cost of renting an ebike is $3 plus an additional $0.12 per minute. Write a function* `ebike` *that will calculate the cost of a ride, given the number of minutes ridden.*

ChatGPT's Purpose Statement: Take in the number of minutes and multiply it by 3.12.

Improved Purpose Statement: _____

_____

Problem with ChatGPT's Purpose Statement: _____

_____

4) **Word Problem:** *Suleika is a skilled house painter at only age 21. She has painted hundreds of rooms and can paint about 175 square feet an hour. Write a function* `paint` *that takes in the number of square feet of the job and calculates how many hours it will take her.*

ChatGPT's Purpose Statement: Take in the number of square feet of walls in a house and divide them by 175 then add 21 years.

Improved Purpose Statement: _____

_____

Problem with ChatGPT's Purpose Statement: _____

_____

Word Problem: rocket-height

Directions: A rocket blasts off, and is now traveling at a constant velocity of 7 meters per second. Use the Design Recipe to write a function `rocket-height`, which takes in a number of seconds and calculates the height.

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

Defining Table Functions

The steps of the Design Recipe don't change just because we're working with Rows, but we can make some adjustments when using Row-consuming functions to filter tables and build columns!

Let's try a concrete example: *Write a function* `is-lizard`*, which tells us whether an animal is a lizard.*

Contract and Purpose

- We still want to pick good names. Since we're writing a function to check if an animal is a lizard, call it `is-lizard`!
- The Domain is a lot easier — it's *always* a `Row`!
- The Range is easier, too. If we're writing a function to filter a Table, we know the Range *has to be a Boolean* . (What would it be if we were building a column of Numbers? Images? Strings?)

Examples

The goal of the Examples step is to *find the pattern* that represents what the function does. When working with Rows, sometimes we have to start by just focusing on what the answer should be.
Suppose we have two rows defined:

`lizard-row` (which happens to be a lizard) `cat-row` (which happens to be a cat)

We can imagine the answers for an `is-lizard` function to be…

```
examples:
    is-lizard(lizard-row) is true
    is-lizard(cat-row)    is false
end
```

But why do we think these expressions will evaluate to `true` and `false`?
We KNOW `lizard-row` *is a lizard, and we KNOW* `cat-row` *is a cat and not a lizard…*
If we replace our answers with the Boolean expressions that compare their species, someone else would be able to follow our logic.

```
examples:
    is-lizard(lizard-row) is "lizard" == "lizard"  # will produce true
    is-lizard(cat-row)    is "cat"    == "lizard"  # will produce false
end
```

And what work would the computer need to do to know that `lizard-row` is a lizard and `cat-row` is a cat? Look in the `species` column!

```
examples:
    is-lizard(lizard-row) is lizard-row["species"] == "lizard" # will produce true
    is-lizard(cat-row)    is cat-row["species"]    == "lizard" # will produce false
end
```

Sometimes we can get straight to this final form in one step, but sometimes it helps to break our thinking down into pieces.
Once we see the pattern, we can *circle and label what changes* .
In this case, only the Row representing the animal changes! So we might use `r` as our label, to represent the Row.

Definition

The final step in the Design Recipe is to take the pattern from our examples and *generalize it* to work with any input.
It's no different when working with Rows.
Our previous step is a huge help. We can **copy everything that stays the same**, and replace the part that changes with the label we used.
Combining the Contract, Purpose, Examples and Definitions, we end up with:

```
# is-lizard :: Row -> Boolean
# Consumes a Row, and checks to see if the species column is "lizard"
examples:
    is-lizard(lizard-row) is lizard-row["species"] == "lizard" # will produce true
    is-lizard(cat-row)    is cat-row["species"]    == "lizard" # will produce false
end
fun is-lizard(r): r["species"] == "lizard" end
```

Making Connections

*Open the [Row Functions Starter File]() on your computer, save a copy, and **Click "Run"**!*

1) Write the code to lookup the value of the `weeks` column for each of the rows listed (the first one has been completed for you).

row	code to lookup the value of the `weeks` column
`cat-row`	`cat-row["weeks"]`
`young-row`	
`old-row`	

2) Write the code that uses the `circle` function to draw a solid, green circle whose radius is the *number of weeks* it took to get adopted (the first one has been completed for you).

row	code to draw a circle using the "weeks" of the row as the radius
`cat-row`	`circle(cat-row["weeks"], "solid", "green")`
`young-row`	
`old-row`	

3) Check with your partner or another student to confirm that your code matches.

4) What is the name of the animal defined in `old-row`? _____ How many weeks did it take for them to be adopted? _____

weeks-dot

Scroll down in the [Row Functions Starter File]() until you find the Contract, Purpose, Examples and Definition for `weeks-dot`.

5) What is the Domain of this function? _____ The Range? _____ How many examples does this function have? _____

6) Does the Purpose Statement make it clear what this function should do, when given a Row? _____

7) Look at the first two examples. How do they satisfy the Contract and Purpose Statement?

_____

_____

These examples show us exactly what should be produced for `cat-row` and `young-row` - the two Rows representing "Sasha" and "Wade", based on their `weeks` to adoption (1 and 3). But they don't show us where the computer should get the number of `weeks` from!

8) The last two examples do the same thing as the first two examples, but the numbers 3 and 1 have been replaced!
Where do they get the number of weeks from?

_____

9) How is the definition for the `weeks-dot` function connected to our examples?

_____

_____

10) Add an example for `old-row` to match first pair of examples (using the actual number of weeks). Then add an example for the second pair (using a lookup).

★ Choose one more row that's defined at the top of the file, and add examples for that as well.

Advanced Displays

Functions as Data

You've learned that functions are *machines that consume and produce data* .

In the real world, we see machines consume things to produce things all the time:

- Bulbs consume electricity and produce light.
- Toasters consume bread and produce toast.

Sometimes, machines consume other machines:

- A school bus is a machine. It comes with a stereo, which could be swapped out for a new one with more features. A stereo is a machine. And the bus needs one of them in order to play music over the speakers.
- A blender might have different attachments. Each attachment is a machine of its own and the blender needs one of them to work!

This is true of function machines in math and programming, as well! By now you've learned plenty of data types (e.g. - Numbers, Strings, Images, Booleans, Rows and Tables). **... And Functions can be their own kind of data type!**

- Imagine a function `species-dot`, that consumes a Row from the Animals Dataset, and produces a different-colored square depending on the species.
- What if we used `species-dot` to customize the dots on our scatterplot, instead of using the same blue dot for each animal?
  In this example, we'd be using the `species-dot` *function as an input to our* `scatter-plot` *function!*

Here are the Contracts for some special display **functions that consume functions**, including the scatterplot we just described:
Look carefully at the last argument in each Domain. In each case, ***the function consumes a Row and produces an Image***.

```
# image-scatter-plot :: Table, String, String, (Row -> Image) -> Image
# image-histogram :: Table, String, Number, (Row -> Image) -> Image
# image-bar-chart :: Table, String, (Row -> Image) -> Image
# image-pie-chart :: Table, String, (Row -> Image) -> Image
```

Piecewise Functions

Functions always apply a particular rule to their input.

- In an earlier lesson, you saw how `gt` always draws a solid, green triangle using the input as the size.
- In the `species-dot` example above, there's no single rule that will generate a different color for each species.

We need a way for functions to change rules, depending on their input.

Piecewise Functions are functions that can behave one way for part of their Domain, and another way for a different part.

- Piecewise functions are divided into "pieces".
- Each piece has two parts: the "if" and the "then".
- This tells the computer *when* to apply each rule, and *what* the rule is.

In our `species-dot` example, our function might draw black squares when the input is a dog, but orange squares when the input is a cat. The function definition would look like this:

```
# species-dot :: (Row) -> Image
fun species-dot(r):
  if      (r["species"] == "dog"):       square(5, "solid", "black")
  else if (r["species"] == "cat"):       square(5, "solid", "orange")
  else if (r["species"] == "rabbit"):    square(5, "solid", "pink")
  else if (r["species"] == "tarantula"): square(5, "solid", "red")
  else if (r["species"] == "lizard"):    square(5, "solid", "green")
  end
end
```

age-dot

1) Write the code to lookup the value of the `age` column for each of the rows listed (the first one has been completed for you).

row	code to lookup the value of the `age` column
`dog-row`	`dog-row["age"]`
`old-row`	
`young-row`	

2) Write the code that uses the `circle` function to draw a solid, blue circle whose radius is the *age of the animal* for each of the rows listed (the first one has been completed for you).

row	code to draw a circle using the "age" of the row as the radius
`dog-row`	`circle(dog-row["age"], "solid", "blue")`
`old-row`	
`young-row`	

3) Check with your partner or another student to confirm that your code matches.

Instead of writing repetitive code like this over and over for each animal, let's define a function to do it for us!

Defining the Function

Directions: Define a function called `age-dot`, which takes in a row from the Animals Table and draws a solid, blue circle whose radius is the age of the animal. *HINT: Use the rows from above in your examples!*

Contract and Purpose Statement

Every contract has three parts…

age-dot:: *Row* -> *Image*

function name — Domain — Range

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 *age-dot*(_____):

function name — variable(s)

_____

what the function does with those variable(s)

end

species-tag

To help you with this page, we've re-printed the Contract for the `text` *function, and an example of how to use it.*
(Remember, you can always refer to the Contracts Pages. *If you're working with a printed workbook, they are included in the back.*)

```
# text :: (String, Number, String) -> Image
          message  size   color
text("hello", 24, "green")
```

1) On the three lines below, write the code to lookup the value of the `species` column from `dog-row`, `old-row`, and `young-row`.

_____

_____

_____

2) On the three lines below, write the code that uses the `text` function to show the species of those same three rows in *red, 15px letters*.

_____

_____

_____

3) Check with your partner or another student. Do you have the same code? Why or why not?

Instead of writing this out over and over for each animal, let's define a function to do it for us!

Defining the Function

Directions: Define a function called `species-tag`, which takes in a row from the Animals Table and draws its name in red, 15px letters.
HINT: Use of the rows from above in your examples!

Contract and Purpose Statement

Every contract has three parts…

 # species-tag::_____*Row*_____->_____*Image*_____
function name — Domain — Range

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 *species-tag*(_____):
function name — variable(s)

_____
what the function does with those variable(s)

end

Exploring Image Scatter Plots

Look at the code in the [Custom Scatter Plot Starter File](#).

1) Compare the definitions of `age-dot` and `species-tag` to what you wrote. Are they the same? If not, what is different?

_____

_____

Answer the following questions about the last line of code in the file, which is commented out.

2) What is the **name** of the function being used here? _____ How many things are in its **Domain**? _____

3) What is the 1st argument? _____ What is its **data type?** _____

4) What is the 2nd argument? _____ What is its **data type?** _____

5) What is the 3rd argument? _____ What is its **data type?** _____

6) What is the 4th argument? _____

7) What is the **data type** of the fourth argument in the Domain? If you're not sure, write down your thinking. What can you rule out? What do you think it *might* be? _____

_____

8) **Uncomment the last line at the bottom of the file, and click "Run".** What does `image-scatter-plot` do with its 4th argument?

_____

_____

9) Try changing your `age-dot` function to use different colors, or even different shapes! Can you make the size of the shape be *one half* the age of the animal?

10) **On a new line in the Definitions Area, try making an `image-scatter-plot` using the `species-tag` function.**

Click run, and describe what you see. _____

_____

Understanding Custom Displays

11) **Look at the image scatter plot that has dots of different sizes.**

Can you draw any conclusions about animals that are both *young* and *lightweight*? _____

_____

12) Looking at that same scatter plot, the director of the shelter says: "Animals that are older *and* that weigh more than 50 pounds generally take at least 5 weeks to be adopted." Do you agree with this statement? Explain. _____

_____

13) **Look at your image scatter plot with species nametags in red.**

What does this chart reveal that we couldn't see on the original (pounds v. weeks) scatter plot? _____

_____

Exploring Conditional / Piecewise Functions

Here's an example of a piecewise function with 3 "pieces" (or "conditions"):

```
# species-dot :: (Row) -> Image
fun species-dot(r):
  if      (r["species"] == "dog"):       square(5, "solid", "black")
  else if (r["species"] == "cat"):       square(5, "solid", "orange")
  else if (r["species"] == "lizard"):    square(5, "solid", "green")
  end
end
```

What do you Notice about this code?	What do you Wonder?

1) What will this function produce for a dog? _____

2) What will this function produce for a cat? _____

Open the Piecewise Displays Starter File, and click "Run".

3) Compare the regular scatter plot with the image scatter plot. What can you see now that you couldn't see before?

_____

_____

4) Compare the regular histogram with the image histogram. What can you see now that you couldn't see before?

_____

_____

5) What do you think will happen if we run the function on a species that it has no condition for? _____

6) **On line 45, add a comment (#) to "turn off" the condition for snails.** Click Run and test your prediction. In your own words, describe how piecewise / conditional functions work.

_____

_____

_____

★ *Optional:* Make a **new function** (don't delete `species-dot`!), which uses piecewise functions to draw something different! For example, have it draw different shapes depending on whether an animal is younger than 3 years old or not.

Advanced Table Manipulation

You've seen that Pyret has special functions that we can use to manipulate Tables:

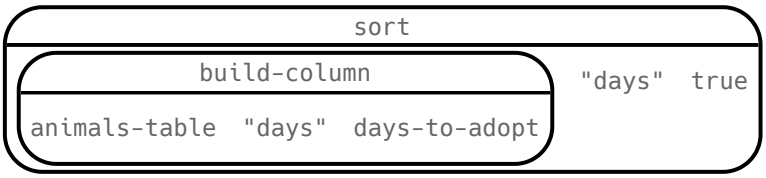
- `sort` - consumes the name of a column and a Boolean value to determine if that table should be sorted by that column in ascending order
- `row-n` - consumes an index (starting with zero!) and produces a row from a table
- `filter` - consumes a *Boolean-producing function* , and produces a table containing only rows for which the function returns `true`
- `build-column` - consumes the name of a new column, and a function that produces the values in that column for each Row

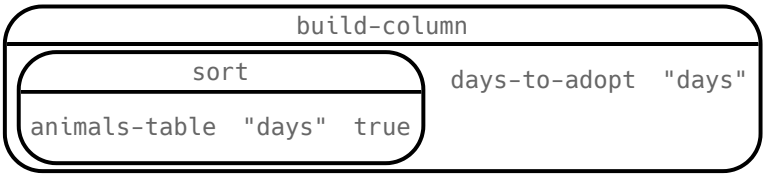
You've also seen the "Order of Operations" for things like addition, subtraction, multiplication and division. **Is there an Order of Operations for manipulating tables?**

Suppose you have the following function defined:

```
fun days-to-adopt(r): r["weeks"] * 7 end
```

One of the Circles of Evaluation below will sort the table by the number of days it took for each animal to be adopted. **One of them will produce an error!** Can you figure out which one?





We can't `sort` *by a column that doesn't exist yet!* In fact, we can't `filter` by that column either. When filtering and building tables, it's important to keep the order of operations in mind!

Reading Row and Function Definitions

Open the [Table Functions Starter File]() on your computer, save a copy, and click "Run".

1) What is the name of the Table defined on line 5? _____

2) How many columns does it have? _____

3) What is the name of the Row defined on line 17? _____

4) Is `red-circle` a Number, String, Image, Boolean, Table, or Row? _____

5) Type `red-circle` into the Interactions Area. What do you get? _____

6) In the space provided on lines 19 and 20, add new definitions for two more Rows from this table.

7) A Contract for a function is written on line 22. What is its name? _____

8) What is its Domain? _____

9) What is its Range? _____

10) What other functions are defined in the starter file?

_____

11) Use the function `is-red`, passing in a Row. For example, type `is-red(blue-triangle)`. What do you get? _____

12) What does the `is-red` function do? _____

For the remaining functions, read the code and try to guess what it does *before* testing it out by passing in a Row.

13) What does `is-polygon` do? _____

14) What does `lookup-name` do? _____

15) What does `is-triangle` do? _____

16) Define two new functions: `is-green` and `is-blue`.

★ There is a hidden function called `draw-shape`. What is its Domain and Range? What does it do?

_____

_____

★ Is there another way to define `is-polygon`, so that it doesn't use the `"polygon"` column at all?

_____

_____

Exploring Table Functions

Open your copy of the [Table Functions Starter File]() and click "Run".

Filtering Rows

1) What does `filter(shapes-table, is-red)` evaluate to? Describe the value you get back below.

_____

2) What does `filter(shapes-table, is-polygon)` evaluate to? Describe the value you get back below.

_____

3) Write the code to generate a table of only triangles. _____

4) At the *bottom* of the Definitions Area, define `triangles` by writing `triangles = filter(shapes-table, is-triangle)`. Click "Run" and evaluate `triangles` in the Interactions Area.

5) Define `reds` to be a table of only red shapes. _____

6) What do the contracts for `is-red`, `is-polygon`, and `is-triangle` have in common?

_____

7) **Find the Contract for `filter` on the Contracts Page.** *If you're working with a printed workbook, the contracts pages are included in the back.* Explain how `filter` uses the two inputs specified in the Domain.

_____

_____

8) What happens if you evaluate `filter(shapes-table, lookup-name)`? Why?

_____

Building Columns

9) What does `build-column(shapes-table, "red", is-red)` evaluate to?

_____

10) What does `build-column(shapes-table, "img", draw-shape)` evaluate to?

_____

11) Find the Contract for `build-column` on the Contracts Page. *If you're working with a printed workbook, the contracts pages are included in the back.* Explain how `build-column` uses the three inputs specified in the Domain.

_____

_____

_____

Define your own table!

★ In the Definitions Area, define a table of your own using `filter` or `build-column`. Add a comment to describe what's in it!

What Table Do We Get?

Consider the table below, and the four function definitions that follow:

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

name	sex	age	fixed	species	pounds
"Toggle"	"female"	12	true	"dog"	48
"Fritz"	"male"	4	false	"dog"	92
"Nori"	"female"	6	true	"dog"	35.3
"Sunflower"	"female"	2	false	"cat"	51.6

```
fun lookup-fixed(animal): animal["fixed"]                    end
fun is-dog(animal):       animal["species"] == "dog"         end
fun is-old(animal):       animal["age"] > 10                 end
fun label(animal):        text(animal["name"], 20, "red")    end
```

Below is a list of expressions, eaching using one table function. *Match* each expression to the description of the table it will produce.

`sort(t, "age", true)`	1	A	Produces a table with Toggle, Fritz, and Nori - but not Sunflower.
`sort(t, "pounds", false)`	2	B	Produces a table of all four animals, sorted youngest-to-oldest
`build-column(t, "sticker", label)`	3	C	Produces a table, with only Toggle.
`filter(t, is-old)`	4	D	Produces an identical table with an extra column called "dog", whose values are true, true, true, false
`filter(t, lookup-fixed)`	5	E	Produces a table containing only Nori and Toggle.
`filter(t, is-dog)`	6	F	Produces a table with all four animals, sorted from heaviest to lightest.
`build-column(t, "dog", is-dog)`	7	G	Won't run: will produce an error. (Why?)
`filter(t, label)`	8	H	Produces an identical table with an extra column called "sticker", whose values are images

Putting it All Together

Open the Putting it All Together Starter File and **take a look at the helper functions** in the Definitions Area.

Write the names of those functions here: _____

Filter and Building with our Helper Functions

Example: Make a table of only dogs (define it as `dogs`*)*

```
dogs = filter(animals-table, is-dog)
```

1) Make a table of only fixed animals (define it as `fixed`)

```
fixed =
```

2) Make a table with a column called "sticker", containing a label for every animal

```
stickers =
```

3) Make a table of only fixed dogs (define it as `fixed-dogs`)

```
fixed-dogs =
```

★ Make a table of old, fixed dogs… with a "sticker" column! (define it as `sticker-table`)

```
sticker-table =
```

Define Additional Helper Functions

4) Define a function called `is-lizard`, which consumes a `Row` of the animals table and *computes* whether the animal is a lizard.

5) Define a function called `months`, which consumes a `Row` of the animals table and divides the weeks by `4.435` to get the approximate equivalent number of months to adoption.

★ Make a table with a "months" column (define it as `months-table`)

Make Displays Using Your New Tables

6) Make a pie chart showing the sex of only fixed dogs.

7) Make a box plot showing the distribution of months to adoption.

★ Make a scatter plot of old, fixed dogs, comparing age to pounds using the "sticker" as the label!

The Design Recipe: is-dog / is-female

Directions: Define a function called `is-dog`, which consumes a `Row` of the animals table and *computes* whether the animal is a dog. HINT: use predefined rows like `dog-row` to make your examples easier!

Contract and Purpose Statement

Every contract has three parts…

```
# is-dog:: Row -> Boolean
```

function name — Domain — Range

```
# Consumes an animal, and checks whether the species == "dog"
```

what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

```
is-dog(dog-row) is dog-row["species"] == "dog"
```

function name — input(s) — what the function produces

```
is-dog(                    ) 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: Define a function called `is-female`, which consumes a Row of the animals table and returns true if the animal is female. HINT: use predefined rows like `female-row` to make your examples easier!

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 Design Recipe: is-old / name-has-s

Directions: Define a function called `is-old`, which consumes a Row of the animals table and *computes* whether it is more than 12 years old. HINT: use predefined rows like `old-row` to make your examples easier!

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: Define a function called `name-has-s`, which returns true if an animal's name contains the letter "s". HINT: The name of the animal in `cat-row` is "Sasha" and the name of the animal in `dog-row` is "Toggle".

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:

name-has-s (*cat-row*) **is** _____
function name — input(s) — what the function produces

name-has-s (*dog-row*) **is** _____
function name — input(s) — what the function produces

end

Definition

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

fun *name-has-s* (*r*):
function name — variable(s)

_____
what the function does with those variable(s)

end

Composing Table Operations

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

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

```
fun is-fixed(r): r["fixed"]                end
fun is-young(r): r["age"] < 4              end
fun nametag(r):  text(r["name"], 20, "red") end
```

Match each table description on the left, to the Circle of Evaluation that will produce it.

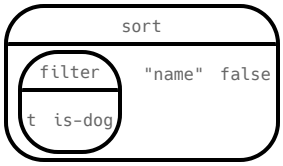
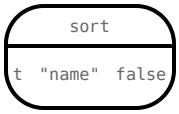
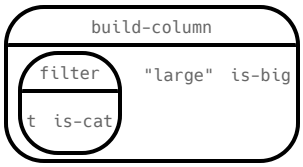
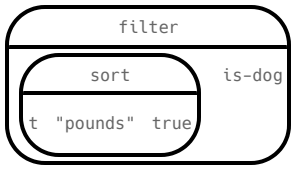
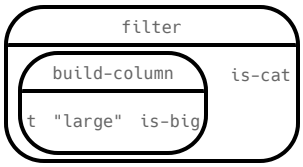
A table containing only Toggle and Sasha	1	A	`sort(t, "age", true)`
Produces a table of only young, fixed animals	2	B	`filter(t, is-fixed)`
Produces a table, sorted youngest-to-oldest	3	C	`build-column(t, "sticker", nametag)`
Produces a table with an extra column, named "sticker"	4	D	`filter(t, is-young)`
Produces a table containing Toggle and Sasha, in that order	5	E	`filter(filter(t, is-young), is-fixed)`
Produces a table containing Toggle, Fritz, and Nori	6	F	`sort(filter(t, is-young), "pounds", false)`
Won't run: will produce an error (why?)	7	G	`sort(build-column(t, "label", nametag), "age", true)`
Produces a table with an extra "label" column, sorted youngest-to-oldest	8	H	`sort(t, "cuteness", false)`

From Circles to Code

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

name	sex	age	fixed	species	pounds
"Toggle"	"female"	12	true	"dog"	48
"Fritz"	"male"	4	false	"dog"	92
"Nori"	"female"	6	true	"dog"	35.3
"Sunflower"	"female"	2	false	"cat"	51.6

Convert each Circle of Evaluation below into Pyret code. What do you think the resulting table will be?
The first one has been done for you.

	Circle of Evaluation	Pyret code
1		`sort(filter(t, is-dog), "name", false)`
2		
3		
4		
5		

Define the functions specified below by filling in the blanks. The first one has been done for you.

6	A function `is-cat`, which returns true if the animal is a cat.	`fun is-cat(r): r["species"] == "cat" end`
7	A function `is-dog`, which returns true if the animal is a dog.	`fun is-dog(r):` _____ `end`
8	A function `is-big`, which returns true if an animal weighs more than 50 pounds.	`fun is-big(r):` _____ `end`

Planning Table Operations

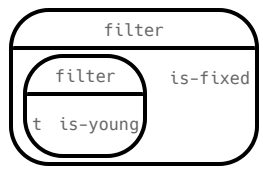
Consider the table below, and the function definitions that follow:

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

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

```
fun is-female(r): r["sex"] == "female" end
fun is-young(r):  r["age"] < 4         end
fun is-fixed(r):  r["fixed"]           end
```

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

	Prompt	Circle of Evaluation
1	Produce a Table containing all young, fixed animals	
2	Produce a Table showing all fixed female animals, sorted by age	
3	Produce a box-plot for all fixed female animals, showing the distribution of `age`	
4	Produce a pie-chart for all young, fixed animals, showing the distribution of `sex`	

Grouped Samples from the Animals Dataset

Use function composition to define the **grouped samples** below. We've given you the solution for the first sample, to get you started.
Assume that the following helper functions are defined exactly the way they are in the [Grouped Samples Starter File](#): `is-old`, `is-young`, `is-cat`, `is-dog`, `is-female`, `is-fixed`, and `name-has-s`.

	Subset	The code to define that subset
1	Kittens	`kittens = filter(filter(animals-table, is-cat), is-young)`
2	Puppies	
3	Fixed Cats	
4	Cats with "s" in their name	
5	Old Dogs	
6	Fixed Animals	
7	Old Female Cats	
8	Fixed Kittens	
9	Fixed Female Dogs	
10	Old Fixed Female Cats	

Displaying Data

Fill in the tables below, then use Pyret to make the following displays. Record the code you used in the line below.
The first table has been filled in for you.

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

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

code: `bar-chart(filter(filter(animals-table, is-dog), is-young), "fixed")`

2) A `pie-chart` showing how many heavy dogs are fixed or not.

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

code: _____

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

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

code: _____

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

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

code: _____

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

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

code: _____

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

_____

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

code: _____

Data Cycle: Analyzing Categorical Data

Use the [Animals Starter File]() to analyze categorical data with the data cycle.

Ask Questions	*How many of each species are fixed at the shelter?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	If you only need some rows, 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	*Are there more female cats than male cats at the shelter?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data	If you only need some rows, 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?	

Samples from My Dataset

Think back to when we defined grouped samples from the Animals Table, like "puppies", "old cats", etc. What grouped samples would be useful for *your* dataset? List a few of these in the first column.
Then, for each one, define a function that will identify if a row `r` is in the subset. *Hint: you can always use a blank design recipe page.*

Grouped Sample	A function that returns true if a row `r` is in the subset

	`fun _____(r):`
	
	`end`

	`fun _____(r):`
	
	`end`

	`fun _____(r):`
	
	`end`

	`fun _____(r):`
	
	`end`

	`fun _____(r):`
	
	`end`

The Design Recipe

Write helper functions for **your** dataset, which you can use to define grouped samples. Since all helper functions will consume Rows, their Domains have already been filled in for you.

Directions: Define a function called _____, which consumes a Row of the _____

table and _____.

Contract and Purpose Statement

Every contract has three parts…

_____ :: `(r :: Row)` -> `Boolean`

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: Define a function called _____, which consumes a Row of the _____

table and _____.

Contract and Purpose Statement

Every contract has three parts…

_____ :: `(r :: Row)` -> `Boolean`

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

"Trust, but verify ..."

This page requires that you also open the Trust but Verify Starter File.

A "helpful" Data Scientist gives you access to the following function:

```
# fixed-cats :: Table -> Table
# consumes a table of animals, and produces a table containing only cats that have been fixed, sorted
from youngest-to-oldest
```

You can use the function, *but you can't see the code for it!* **How do you know if you can trust their code?**

- You could make a *verification subset* that contains one of every species, and make sure that the function filters out everything but cats.
- You could make sure this subset has multiple cats not already ordered youngest-to-oldest, and make sure the function puts them in the right order.

1) What other qualities would this subset need to have?

2) Create your verification subset! In the space below, list the name of each animal in your subset.

Name

"Trust, but verify..." (2)

This page requires that you also open the Trust but Verify Starter File.

A "helpful" Data Scientist gives you access to the following function:

```
# old-dogs-nametags :: Table -> Table
# consumes a table of animals, and produces a table containing only dogs 5 years or older, with an extra
column showing their name in red
```

You can use the function, *but you can't see the code for it!* **How do you know if you can trust their code?**

1) What qualities would a verification subset need to have?

2) Create your verification subset! In the space below, list the name of each animal in your subset.

Name

We listened to more than 3 hours of US Congress testimony on facial recognition so you didn't have to go through it

Long story short: Models are ineffective, racist, dumb...

Katyanna Quach Wed 22 May 2019 // 23:50 UTC

ANALYSIS AI experts, lawyers, and law enforcement urged US Congress to regulate the use of facial recognition technology during a hearing held by the House Committee on Oversight and Reform on Wednesday, May 22, 2019.

The technical issues and social impacts of using AI software to analyse images or videos are well known. There have been repeated reports of how inaccuracies lead to people being misidentified in research and in real life. San Francisco just passed an ordinance banning the local government using facial recognition technology.

In some cases, like the experiment conducted by the American Civil Liberties Union's (ACLU), a nonprofit based in New York, that showed Amazon Rekognition incorrectly matched members of the US Congress to criminal mugshots, the effects have been minimal. It's simply absurd for elected politicians to be wanted criminals. But what happens when the technology is turned on civilians who have less power?

At a hearing of the House Committee on Oversight and Reform on Wednesday, Joy Buolamwini, founder of Algorithmic Justice League, an activist collective focused on highlighting the shortcomings of facial recognition, found that commercial computer models struggled most when it came to recognizing women with darker skin. IBM's system was incorrect for 34.7 per cent of the time when it came to identifying black women, she said…

The problem boiled down to biased training datasets, Buolamwini told the House committee. AI systems perform worse on data that they haven't seen before. So, if most datasets mainly represent white men then it's not surprising that they find it difficult when faced with an image of women of colour.

When it comes to databases of mugshots, however, the reverse is true. Black people are overrepresented in mugshot databases, explained Clare Garvie, Senior Associate at Georgetown University Law Center's Center on Privacy & Technology. If law enforcement are using these flawed models to target the group of people that it struggles to identify most then it will undoubtedly lead to police stopping and searching the wrong people. "It's a violation of the first and fourth amendment," Garvie said during the hearing.

Law enforcement and lack of transparency

Cedric Alexander, the former president of the National Organization of Black Law Enforcement Executives who was also a witness at the hearing, estimated that at least a quarter of law enforcement agencies across the US use facial recognition to some degree.

Police from Washington County and Orlando are an example of some bureaus that are using Rekognition. Michael Punke, Amazon's VP of Global Public Policy, said at the time it has "not received a single report of misuse by law enforcement." It's difficult to verify that claim, however, considering that the police haven't been transparent about how it's used.

It's all done in secrecy, according to testimony. Elijah Cummings, the chair of the Oversight Committee, said that 18 states had shared data like passport photos or driver licenses with the FBI without explicit consent. When the witnesses were pressed with questions on what kind of information law agencies share with one another, nobody knew.

Neema Guliani, senior legislative counsel for the ACLU, took a tough stance and called for a moratorium on the technology. She urged the committee to "take steps to halt the use of face recognition for law enforcement and immigration enforcement purposes until Congress passes a law dictating what, if any, uses are permissible and ensures that individuals' rights can be protected." Unregulated use of the technology could also potentially lead to an "Orwellian surveillance state," where citizens are constantly tracked Guliani said. In the opening statement, Cummings said there are about 50 million surveillance cameras in the US, and that half of all American adults are probably part of facial recognition databases and they don't even know it.

Andrew Ferguson, professor of law at the University of the District of Columbia, agreed that the Congress needed to act now to prohibit facial recognition until Congress establishes clear rules. "Unregulated facial recognition should not be allowed to continue unregulated. It is too chilling, too powerful. The fourth amendment won't save us. The Supreme Court is trying to make amendments but it's not fast enough. Only legislation can react in real time to real time threats," he warned.

Alexander was more cautious about a blanket ban on the technology, however. He believed that there were still ways that law enforcement could positively use facial recognition. "There is a place for the technology, but the police need to be trained properly. They can't just be passed the technology by software companies." Effective policing is about building relationships in the local community, and it can't afford the effects of misidentifying people. How can we utilise the technology, whilst developing some standards?, he asked.

Benchmark tests simply aren't good enough

The National Institute of Standards and Technology (NIST), a laboratory part of the US Department of Commerce, is currently conducting official benchmark tests for commercial facial recognition systems. But they need to be better, Buolamwini said. She brought up the issue of what she called "pale male datasets". "The gold standard benchmark dataset is biased and can lead to a false understanding of progress," she said.

Even if there was a facial recognition system with near-perfect accuracy in the testing phase, it doesn't solve the problem that most data used by law enforcement is often grainy and low resolution. A recent report by Georgetown University found that in some cases police were even trying to match people by composite artist sketches.

"Faces maybe the final frontier of privacy," Buolamwini said.

The hearing took place at the same time as Amazon shareholders tried to stop Rekognition being sold to law enforcement. The proposal was defeated, but the vote tallies were not immediately disclosed. © **The Register.**

Can Software be Biased?

This page is designed as a reflection on either this article, summarizing US Congress Testimony on Artificial Intelligence or this video The Coded Gaze: Bias in Artificial Intelligence.

1) Describe three concerns experts and activists have raised about Artifical Intelligence.

2) What are some solutions that would address these concerns?

3) How would you test whether or not a facial recognition system was equally accurate for everyone?

Threats to Validity

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

Some examples of threats are:

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

Identifying Threats to Validity

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

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

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

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

Identifying Threats to Validity (2)

The animal shelter wanted to find out what kind of food to buy for their animals. They took a random sample of two animals and the food they eat, and they found that spider and rabbit food was by far the most popular cuisine!

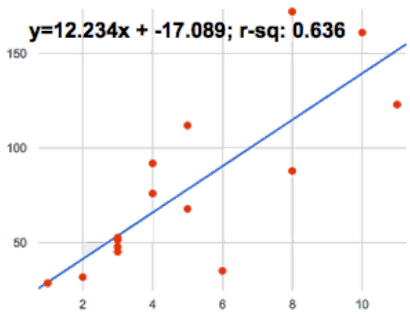
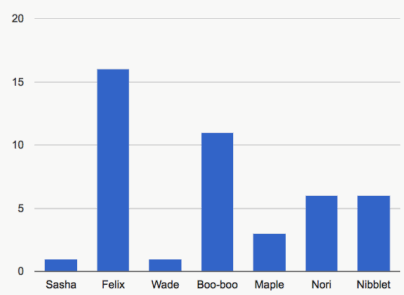
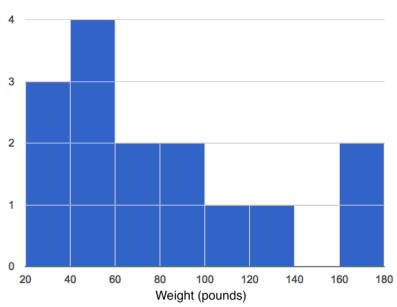
Explain why sampling just two animals can result in unreliable conclusions about what kind of food is needed.

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

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

Fake News

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

	Data	Claim	What's Wrong
1	The average player on a basketball team is 6′1″.	*"Most of the players are taller than 6′."*	
2	Linear regression found a positive correlation (r=0.42) between people's height and salary.	*"Taller people are more qualified for their jobs."*	
3		*"According to the predictor function indicated here, the value on the x-axis will predict the value on the y-axis 63.6% of the time."*	
4		*"According to this bar chart, Felix makes up a little more than 15% of the total ages of all the animals in the dataset."*	
5		*"According to this histogram, most animals weigh between 40 and 60 pounds."*	
6	Linear regression found a negative correlation (r= -0.91) between the number of hairs on a person's head and their likelihood of owning a wig.	*"Owning wigs causes people to go bald."*	

Lies, Darned Lies, and Statistics

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

Data	Claim	Why it's wrong

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

Exploring the States Dataset

Open the [Preview: State Demographics Starter File](#).
Then, click "Run" and type `states-table` into the Interactions Area on the right to see the dataset.

What do you Notice about this dataset?	What do you Wonder about this dataset?

1) What code will produce a table showing the number of states in each region? _____

2) Which states do you **think** have the most people? _____

3) What code will produce a table containing the five states with the largest population in 2020?

_____

4) Which states do you **think** have the most poverty? _____

5) What code will produce a table containing the ten states with the highest % of people in poverty?

_____

6) What code will produce a table containing the states with the lowest **median** income?

_____

7) What code will produce a table containing the states with the lowest **per-capita** ("average" or "mean") income?

_____

★ What does it mean if a state has a higher `per-capita income` than `median-income`? _____

_____

The two lines of code under # `Define some rows` extract rows 0 and 1 from the table, and define them as `alabama` and `alaska`.

8) Type `alabama` into the Interactions Area. What do you get back? _____

9) Underneath the definition of those rows, **add a new definition** for `california` and click "Run", so that Pyret reads your new definition.

10) Add a definition for your own state, then **click "Run"** and test it out in the Interactions Area!

11) Add any additional Notices or Wonderings you have about this dataset to the table at the top.

Looking for Patterns

Open the [Preview: State Demographics Starter File](#).

Part 1

1) What columns do you think might be related to one another? (e.g. - is the number of veterans related to the amount of land-area? Is the population in 2010 related to the population in 2020?) List three possible relationships below.

a. I think that _____ may be related to _____

b. I think that _____ may be related to _____

c. I think that _____ may be related to _____

```
# scatter-plot :: (Table, String, String, String) -> Image
```

labels explanatory response

2) Use the Contract above to make a scatter-plot for the **first relationship** you wrote above. (Use `"state"` as the label, so that clicking on a point will show you which state you're looking at.)

a. If there's a pattern in this scatter-plot, what does that mean? If there isn't, what does *that* mean? _____

_____

b. In your own words, describe the pattern you see in the scatter plot so someone else could sketch it. _____

_____

_____

3) Make a scatter-plot for the **second relationship** you wrote.

a. If there's a pattern in this scatter-plot, what does that mean? If there isn't, what does *that* mean? _____

_____

b. In your own words, describe the pattern you see in the scatter plot so someone else could sketch it. _____

_____

_____

4) Make a scatter-plot for the **third relationship** you wrote.

a. If there's a pattern in this scatter-plot, what does that mean? If there isn't, what does *that* mean? _____

_____

b. In your own words, describe the pattern you see in the scatter plot so someone else could sketch it. _____

_____

_____

Part 2

Wait to complete this until after diving deeper into statistical relationships!

Revisit the three scatter plots you made and add the following labels to the descriptions you wrote in Question 1:

- Place an "L" by any relationships that you think might be linear.
- Place a "P" by any relationships that appear to be positive.
- Place an "N" by any relationships that appear to be negative.
- Place an "S" by the strongest-looking relationship.
- Place a "W" by the weakest-looking relationship.

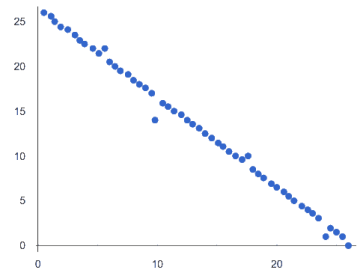
Identifying Form, Direction and Strength (Matching)

Match the description (left) with the scatter plot (right).

Note: The computer won't tell us if the relationship we see is linear, so we must train our eyes to decide this ourselves. For linear relationships, we should train our eyes to assess their direction and get a feel for their strength, rather than relying completely on what numbers the computer reports.

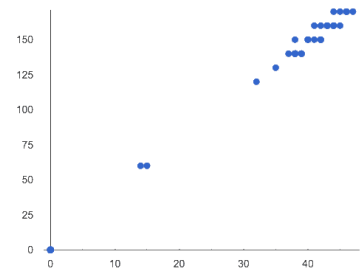
The relationship appears to be linear, negative, and of moderate strength. **1**

A



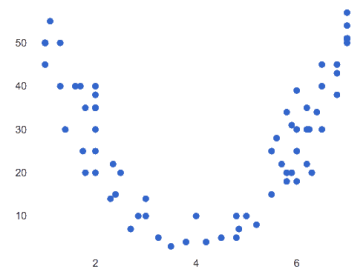
This relationship is nonlinear. **2**

B



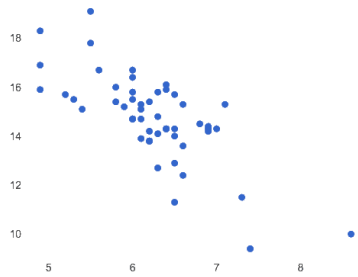
The x and y variables in this dataset do not appear to be related. **3**

C



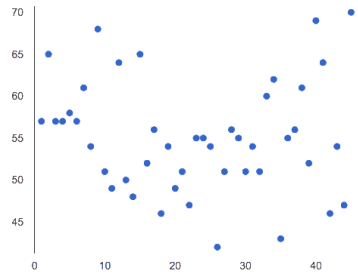
The relationship appears to be linear, positive, and strong. **4**

D



The relationship appears to be linear, negative, and strong. **5**

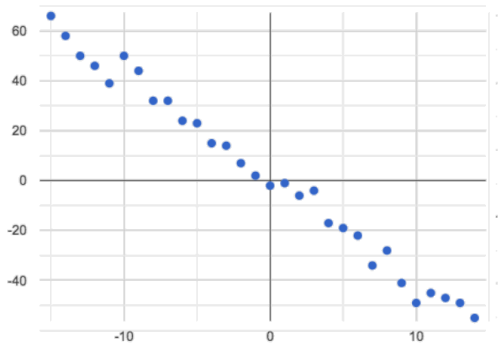
E



Identifying Form, Direction and Strength

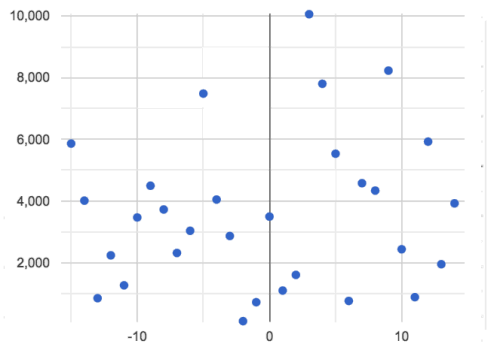
What do your eyes tell you about the Form, Direction, & Strength of these displays?
Note: If the form is nonlinear, we shouldn't report direction - a curve may rise and then fall.

A



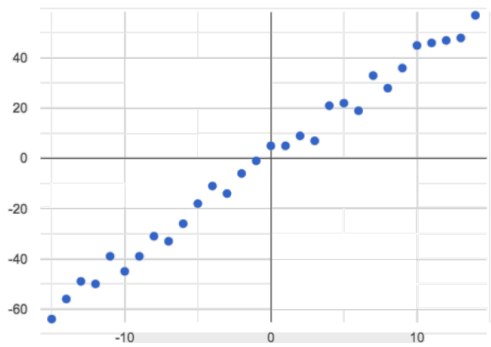
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

B



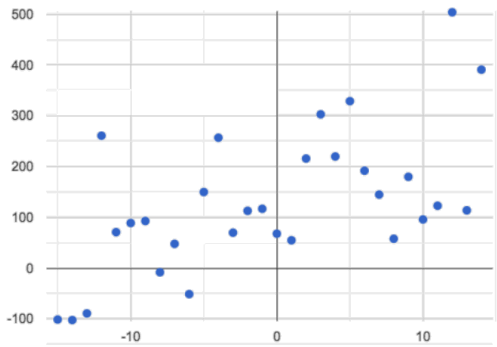
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

C



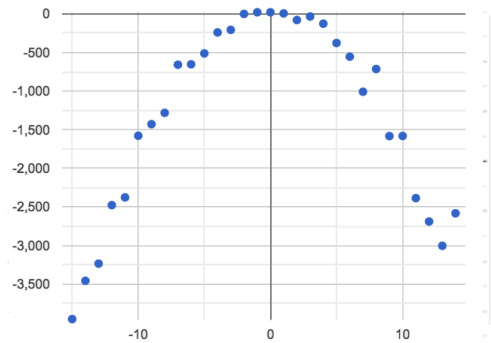
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

D



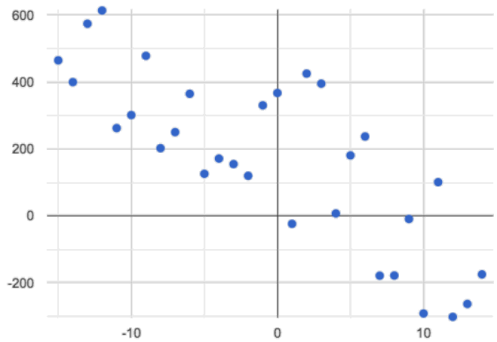
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

E



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

F



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

Build a Model from Samples: College Degrees v. Income

Open the [Preview: State Demographics Starter File](#).

1) Record the `pct-college-or-higher` and `median-income` values for the `alabama` and `alaska` rows, as (x,y) pairs below:

(_____, _____) (_____, _____)

AL pct-college-or-higher, AL median-income; AK pct-college-or-higher, AK median-income

2) Using the space below, compute the equation of the line passing between these two points. **This line will be your linear model** (also known as the "predictor function", or "line of best fit"), which predicts `median-income` as a function of `pct-college-or-higher`.

3) Write the complete model below (in both Function and Pyret notation):

al-ak(x) = _____ x + _____

slope(m); y-intercept / vertical shift

```
fun al-ak(x): (_____ * x) + _____ end
```

Return to your copy of the starter file and add the code you just wrote to the Definitions Area. Then Click "Run".
(If there are any errors or warnings, fix them and click "Run" again.)

4) In the Interactions Area, try plugging in the `pct-college-or-higher` value for Alabama by typing `al-ak(22.6)`.

- How well does it predict the correct median income for Alabama? _____
- What expression would predict median income for Alaska? _____.
- How well does it predict the correct median income for Alaska? _____
  Consider: If it doesn't predict it perfectly, why might that be?

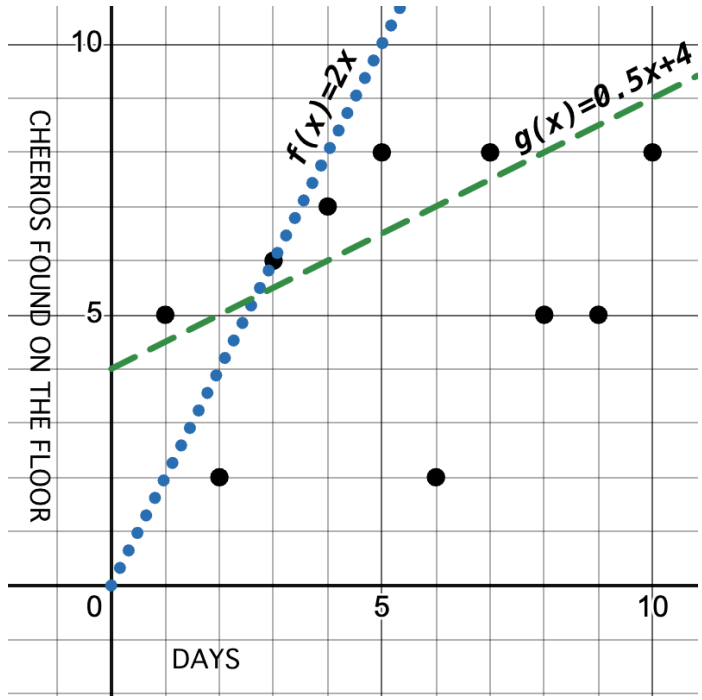
Try different `pct-college-or-higher` values from *other* states, to see how well our Alabama-Alaska model fits the rest of the country.

5) Identify a state for which this model works well: _____

6) Identify a state for which this model works poorly: _____

7) What median income does this model expect a state without ANY college graduates (0%) to earn? _____

How could we Measure Whether a Model is a Good Fit?



id	Days	Cheerios found on the floor
a	1	5
b	2	2
c	3	6
d	4	7
e	5	8
f	6	2
g	7	8
h	8	5
i	9	5
j	10	8

1) Do you think `f(x)` or `g(x)` is a better model for this data? _____

2) What makes you think that? _____

3) What could we measure, to calculate *how much better of a model* it is? _____

4) *Neither of these models is the best possible model!* What would have to be true of a third model, for us to know that it was a better fit than these two? _____

Fit a Model: College Degrees v. Income

*Open the [Fitting a Model: State Demographics Starter File]() and **Save a Copy** of the file that's just for you.*

The `al-ak` Model

Type `fit-model(states-table, "state", "pct-college-or-higher", "median-income", al-ak)` in the Interactions Area, then find the points for AL and AK along the predictor line. *Hint: You know their coordinates and they will help you know where to look!*

1) What do you Notice?

_____

_____

2) What do you Wonder?

_____

_____

3) Find S in the upper left corner. What is the S value (the number after S)? _____

Other Models

In the definitions area, find the section titled *Define some other models by modifying al-ak* .

- For now, all three definitions in this section are exactly the same as `al-ak` .
- You will be changing them according to the directions below.

4) If you wanted the model to be *less steep* , what slope could you use? _____

- Change the definition for `less-steep` to use the slope you wrote above.
  - Click "Run" to load your new definition. In the Interactions Area type:
    `fit-model(states-table, "state", "pct-college-or-higher", "median-income", less-steep)`
  - What is the S value of `less-steep` ? _____
- Identify a y-intercept that would make the model fit the data better: _____
  - Adjust the definition to use the new y-intercept and click "Run".
  - Hit the up arrow in the Interactions Area and click return/Enter to fit the model again.
  - What is the S value of `less-steep` now? _____

5) Change the definition of `negative` so that it models the data with a slope that is *negative* .

- Click "Run" and type the code to fit this model to the data.
- What slope did you use? _____ What is the S value now? _____

6) Change the definition of `horizontal` so that it draws a horizontal model. Click "Run" and fit this model. What is the S value? _____

7) Change the y-intercept so that the horizontal line passes through more of the points. Click "Run" and fit this model.

- What y-intercept did you use? _____ What is the S value now? _____

8) What do you think S tells us? _____

_____

_____

_____

What does *S* tell us about the fit of these models?

For each model below, decide whether the fit is "poor", "ok", or "good". *Then* rank the models from 1 (best fit) to 8 (worst fit).

How good is the model?	Ranking
1 A data scientist is working with data from animals at a shelter. • The range of days to adoption in this dataset are from 0 to 400. • An S value of 300 means predicted adoption times could be off by 300 days. This is a(n) _____ model for the dataset. poor, ok, good	
2 A student is exploring a dataset on climate change. • The range of Arctic Sea Ice is from 3,920,000 to 7,670,000 square kilometers • An S value of 300 means predicted Arctic Sea Ice coverage could be off by 300 square kilometers. This is a(n) _____ model for the dataset. poor, ok, good	
3 A data scientist is working with data from US public schools. • The range of graduates per school per year is 2 to 2003. • An S value of 300 means predicted graduate values could be off by 300 students. This is a(n) _____ model for the dataset. poor, ok, good	
4 A student is exploring a dataset on earthquakes. • The range of earthquake depths in this dataset are from 4200m to 664000m. • An S value of 300 means predicted earthquake depths could be off by 300 meters. This is a(n) _____ model for the dataset. poor, ok, good	
5 A student is exploring a dataset on arrests in Los Angeles. • The age range in this dataset is from 0 to 92. • An S value of 1 means predicted ages could be off by 1 year. This is a(n) _____ model for the dataset. poor, ok, good	
6 A data scientist is working with data about snowflakes. • The range of snowflake weights is from 0.001 grams to 0.02 grams. • An S value of 1 means predicted values could be off by 1 gram. This is a(n) _____ model for the dataset. poor, ok, good	
7 A data scientist is working with data from animals at a shelter. • The range of ages is from 0.5 years to 16 years. • An S value of 1 means predicted ages could be off by 1 year. This is a(n) _____ model for the dataset. poor, ok, good	
8 A student is working with a dataset of adult blue whales. • The range of weights is 200,000 to 330,000 pounds. • An S value of 1 means predicted weights could be off by 1 pound. This is a(n) _____ model for the dataset. poor, ok, good	

Better Modeling: College Degrees v. Income

Open your copy of the [Fitting a Model: State Demographics Starter File](#).

Build a Model through Trial & Error

Find `# Define some rows` in the Definitions Area.
Add two new definitions for `MA` (row 21) and `NV` (row 28), using the definitions for `alaska` and `alabama` as a model.

1) Record the `college-or-higher` and `median-income` values for `MA` and `NV`, as (x,y) pairs below:

(_____, _____) (_____, _____)
MA college-or-higher, MA median-income; NV college-or-higher, NV median-income

2) Derive the `MA-NV model` (using the same steps you followed to derive the `AL-AK model` on [Fit a Model: College Degrees v. Income](#)) and write it below (in both Function and Pyret notation), then fit the model and record the S-value:

$ma\text{ - }nv(x) =$ _____ x + _____ `fun ma-nv(x): (` _____ `* x) +` _____ `end` S: _____
slope (m); y-intercept / vertical shift

3) Identify two other states that you think would make a better model: _____ and _____.

Add two new definitions for these states to your [Fitting a Model: State Demographics Starter File](#).

4) Record the `college-or-higher` and `median-income` values for these states, as (x,y) pairs below:

(_____, _____) (_____, _____)
college-or-higher, median-income; college-or-higher, median-income

5) Derive your model and write it below (in both Function and Pyret notation), then fit the model and record the S-value:

$other(x) =$ _____ x + _____ `fun other(x): (` _____ `* x) +` _____ `end` S: _____
slope (m); y-intercept / vertical shift

6) Adjust the slope and y-intercept of your model to get the **smallest S possible.** Write the best model you find (and corresponding S) below:

$best(x) =$ _____ x + _____ `fun best(x): (` _____ `* x) +` _____ `end` S: _____
slope (m); y-intercept / vertical shift

Optimizing and Interpreting Linear Models

Open your copy of the [Fitting a Model: State Demographics Starter File](#).

Build a Model Computationally

`lr-plot` computes the *optimal linear model* using all of the data points.

1) Evaluate `lr-plot(states-table, "state", "pct-college-or-higher", "median-income")`. What is S? _____

2) On the line below, write the optimal linear model that was computed through linear regression:

$optimal(x) =$ _____ x + _____
slope (m) — y-intercept / vertical shift

`fun optimal(x): (` _____ `* x) +` _____ `end`

Interpret the Model

We started with a model based on Alabama and Alaska `fun al-ak(x): (5613.67 * x) + -83616.02 end` `S: ~36164.68`
which we can interpret as follows:

The ___Alabama-Alaska___ (sensible name) *model predicts that a 1* ___percent___ (x-axis units) *increase in* ___percent college degrees___ (x-axis) *is associated with a* ___5613 dollar___ (slope, y-units) ___increase___ (increase / decrease) *in* ___median household income___ (y-axis). *With an S - value of* ___~36,164.68___ (S-value) ___dollars___ (y-units) *and* ___median household income___ (y-axis) *ranging from* ___$39,031___ (lowest y-value) *to* ___$73,538___ (highest y-value), *this model fits* ___really, really poorly___ (really well, decently, poorly, etc.).

3) Describe the optimal model YOU created via linear regression:

The _____ (sensible name) model predicts that a 1 _____ (x-axis units) increase in _____ (x-axis) is associated with a _____ (slope, y-units) _____ (increase / decrease) in _____ (y-axis). With an S-value of _____ (S-value) _____ (y-units) and _____ (y-axis) ranging from _____ (lowest y-value) to _____ (highest y-value), this model fits _____ (really well, decently, poorly, etc.).

4) What does the **slope (m)** of this linear model tell us? _____

_____

5) What does the **y-intercept / vertical shift** of this linear model tell us? _____

_____

6) Suppose a state goes from 10% to 11% college graduation. According to this model,

- What kind of change would we expect to see in the median household income? _____
- What if it goes from 50% to 51%? _____
- What if it goes from 90% to 91%? _____

7) Does this model predict the same increase in income for *every* additional 1% `college-or-higher`? Why or why not? _____

_____

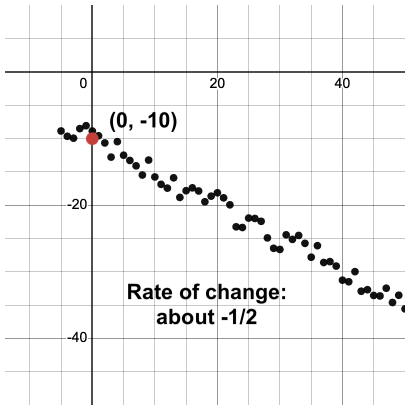
_____

Which Form is Best?

For each set of data provided below,

- Decide which form of the line would be the easiest to build from the available information.
- Write a definition of the linear model in that form.
- Translate the definition into Pyret notation.

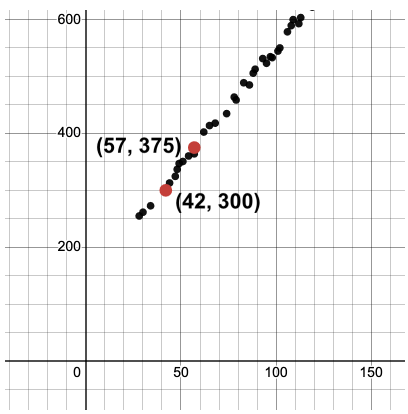
1



Linear Model: _____
Your model slope-intercept, point-slope, or standard form - which ever is easiest!

`fun f(x) :` _____ `end`

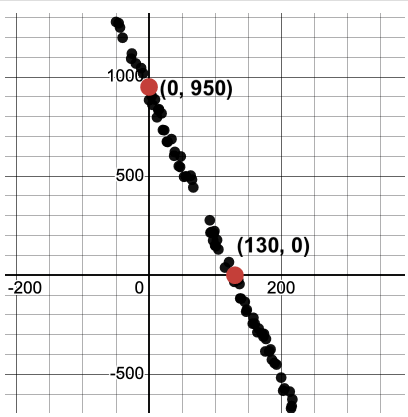
2



Linear Model: _____
Your model slope-intercept, point-slope, or standard form - which ever is easiest!

`fun f(x) :` _____ `end`

3



Linear Model: _____
Your model slope-intercept, point-slope, or standard form - which ever is easiest!

`fun f(x) :` _____ `end`

Exploring the Fuel Efficiency Dataset

*For this page, you'll need to open the Fuel Efficiency Starter File on your computer. If you haven't already, select **Save a Copy** from the "File" menu to make a copy of the file that's just for you. **Read the comments at the top of the file**, which describe what each column in the dataset means.*

Fitting Linear Models

1) Evaluate `A15`, `A45` and `A75` in the Interactions Area. What **model** of car is used in all three rows? _____

2) At what three **speeds** is this model being tested in these rows? _____

3) Does there appear to be a relationship between speed and miles-per-gallon? _____.

4) *Looking at the numbers in the* `mpg-table`, describe its **form** (e.g. - linear, non-linear, or none) and **strength** (strong, moderate, or weak). If it appears to be linear, what is the **direction**? If it does *not* appear to be linear, describe its shape.

_____

_____

5) Use `lr-plot(mpg-table, "model", "speed", "mpg")` to find the optimal **linear** model. What is S for this model? _____
Write the model below, in both math and Pyret notation.

$y =$ _____ x + _____
slope — y-intercept / vertical shift

```
fun y(x): (                        * x) +                        end
```

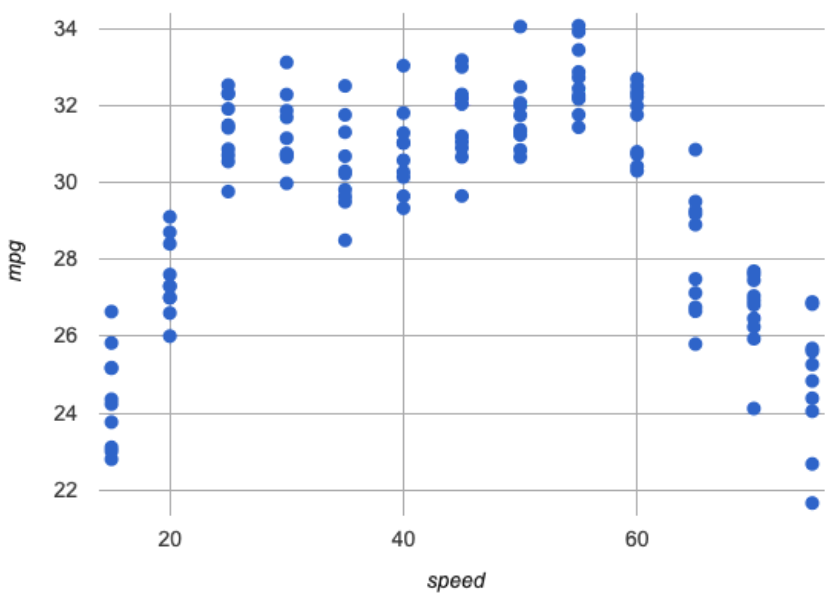
6) Is the best-possible linear model a good fit? _____. Why or why not? _____

_____

_____

Fitting Curves

7) Sketch your lr-plot in the space below, showing the relationship between `speed` and `mpg`. Be sure to label your axes, and draw the linear model!



8) What do you **Notice?** _____

_____

_____

_____

9) What do you **Wonder?** _____

_____

_____

_____

10) Do you think a **curve** would fit better?

_____

11) Draw a **curve** on your scatter-plot, which shows the overall shape in the data. At what speed is the "peak"? _____

12) Based on your best-guess curve, what do you predict `mpg` would be for a new test run at 25mph _____? 60mph _____? 90mph _____?

What Kind of Model? (Descriptions)

Decide whether each situation sounds like it would be better modeled by a linear or quadratic function, and circle your answer.

1) A car is 50 miles away, traveling at 65mph. How far away is the car after each hour?

Linear Quadratic

2) A ball is dropped from the top of the Empire State Building, and it keeps dropping faster and faster. **How far has the ball dropped** after x seconds?

Linear Quadratic

3) The data plan for a cell phone bill costs $5/gb, plus $15/mo. How much is the bill for a given month, after x number of gigabytes?

Linear Quadratic

4) A ball is dropped from the top of the Empire State Building, and it keeps dropping faster and faster. **How fast is the ball moving** after x seconds?

Linear Quadratic

5) A cannonball is fired from the deck of the S.S. Parabola, and arcs through the sky before hitting its target, 17 miles away.

Linear Quadratic

6) The area of a circle, as its radius increases.

Linear Quadratic

7) The circumference of a circle, as its radius increases.

Linear Quadratic

What Kind of Model? (Tables)

Decide whether each representation is best described by a linear model, a quadratic model or neither! Show any work that you feel is useful.

For Class Discussion:

1

x	0	1	2	3	4	5	6
y	5	6	9	14	21	30	41

Linear
Quadratic
Neither

2

x	0	1	2	3	4	5	6
y	0	3	6	9	12	15	18

Linear
Quadratic
Neither

For Independent Practice:

3

x	1	2	3	4	5	6	7
y	1	3	5	7	9	11	13

Linear
Quadratic
Neither

4

x	-3	-2	-1	0	1	2	3
y	-23	-38	-47	-50	-47	-38	-23

Linear
Quadratic
Neither

5

x	-3	-2	-1	0	1	2	3
y	1	2	1	2	1	1	1

Linear
Quadratic
Neither

6

x	1	2	3	4	5	6	7
y	2	5	10	17	26	37	50

Linear
Quadratic
Neither

7

x	-3	-2	-1	0	1	2	3
y	12	7	2	-3	-8	-13	-18

Linear
Quadratic
Neither

8

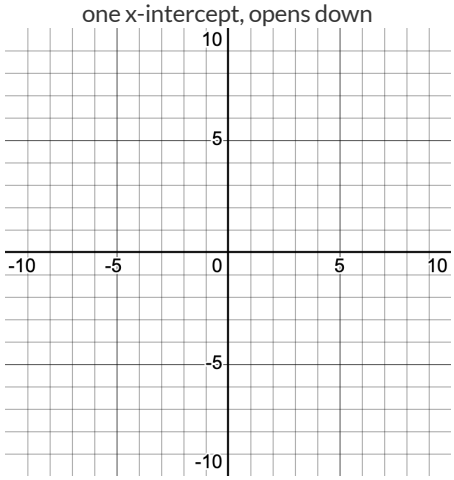
x	1	2	3	4	5	6	7
y	100	102	105	109	114	120	127

Linear
Quadratic
Neither

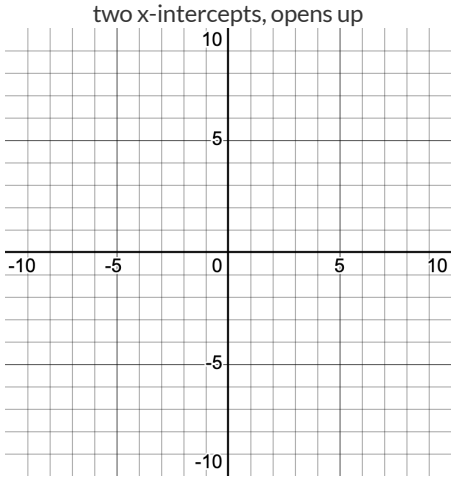
Parabolas

1) Sketch a ***parabola*** on each of the grids below that matches the description.

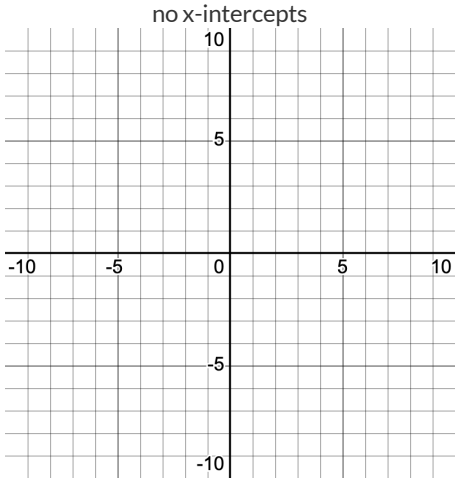
one x-intercept, opens down



two x-intercepts, opens up



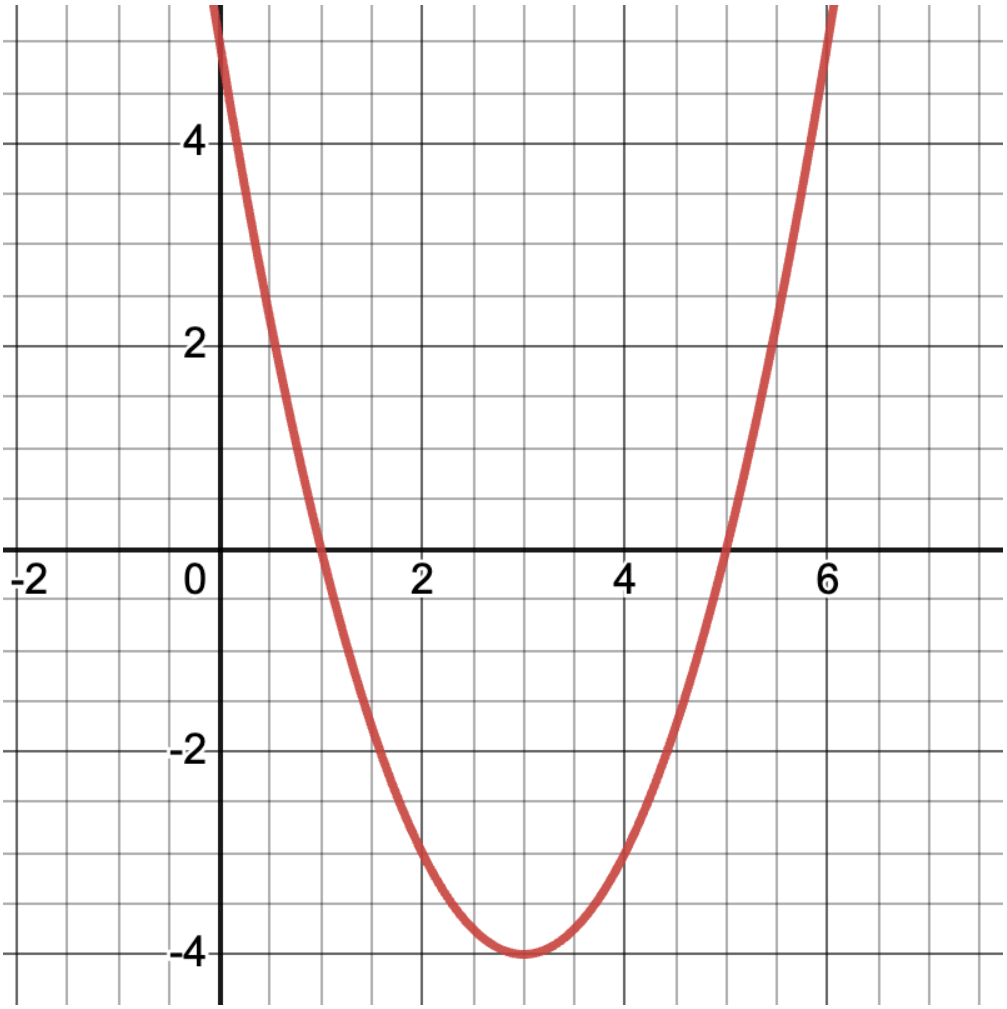
no x-intercepts



2) Label the ***vertex***, ***root(s)***, and ***y-intercept*** of the parabola below with:

A) their coordinates

B) the vocabulary word (above) that describes each



3) Draw a dotted line representing the ***axis of symmetry*** and label it with the equation that defines it.

Graphing Quadratic Models

For this page, you'll need to have ***Exploring Quadratic Functions(Desmos)*** *open on your computer.*
The parabola you'll see is the graph of $g(x) = x^2$. Another, **identical** parabola is hiding behind it.
This second parabola is written in Vertex Form: $f(x) = a(x - h)^2 + k$. Each coefficient starts at values to make $f(x)$ equivalent to $g(x)$.

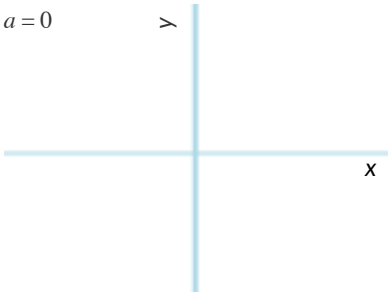
1) Using the starting values of a, h, and k you see for $f(x)$ in Desmos, rewrite $g(x) = x^2$ in Vertex Form. $g(x) =$ _____

Magnitude a

2) Try changing the value of a to -4, 0, and 2, graphing each parabola in the squares below. **Label the vertex "V" and any roots with "R"!**

$a = -4$

$a = 0$



$a = 2$

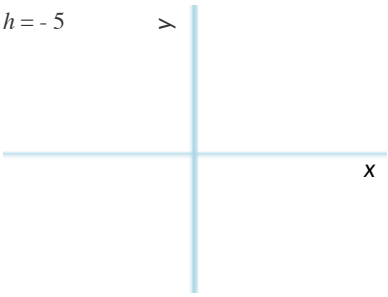
3) What does a tell us about a parabola? _____

_____

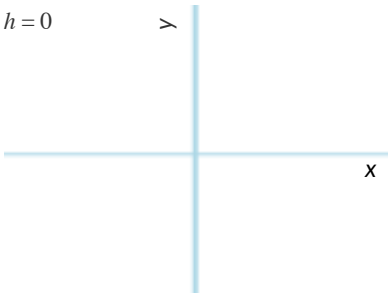
Horizontal Translation h

4) Set a back to 1. Change the value of h to -5, 0, and 5, graphing each parabola in the squares below. **Label the vertex "V" and any roots "R"!**

$h = -5$



$h = 0$



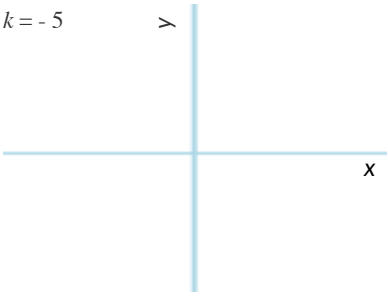
$h = 5$

5) What does h tell us about a parabola? _____

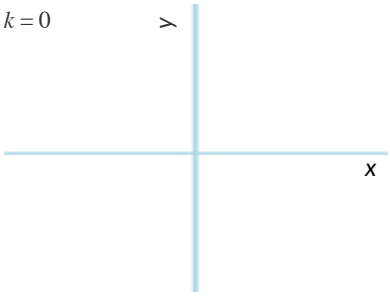
Vertical Translation k

6) Set h back to 0. Change the value of k to -5, 0, and 5, graphing each parabola in the squares below. **Label the vertex "V" and any roots "R"!**

$k = -5$



$k = 0$



$k = 5$

7) What does k tell us about a parabola? _____

Modeling Fuel Efficiency v. Speed

Open your copy of the [Fuel Efficiency Starter File]() and click "Run".

num-sqr

Before we try to model our fuel-efficiency data, we need to learn a new Pyret function!

1) Can you predict what the output of the `num-sqr` expressions below will be?

Test them out in the Interactions Area, and record the results. `num-sqr(4)` _____ `num-sqr(6 - 2)` _____

2) What is the Contract for `num-sqr`? _____

3) What does `num-sqr` do? _____

Interpreting a Quadratic Model

In the Definitions Area of your [Fuel Efficiency Starter File](), you'll find the definition of a quadratic model `quad1`.

4) In `quad1`, the value of a is _____, the value of h is _____, and the value of k is _____

5) Fit this model to your dataset, using `fit-model`. What S-value did you get? _____

Hint: If you forgot the contract for `fit-model`*, look it up in the [contracts pages]()!*

6) In your own words, describe what needs to change about this model to fit the data. _____

_____

Modeling Fuel Efficiency

Vertex Form: $f(x) = a(x-h)^2 + k$	• a: determines whether the parabola opens up or down and how steep the curve is • h: horizontal shift (also the x-coordinate of the vertex! h is often 0) • k: vertical shift (also the y-coordinate of the vertex!)

7) We've determined that peak fuel efficiency is around 45 mph. What variable in the equation should we replace with 45? _____

Update the definition of `quad1`, click "Run" and re-fit the model. What S-value did you get? _____

8) What y-coordinate of the vertex (*vertical shift*) would best match the shape of the curve? _____

Update the definition of `quad1`, click "Run" and re-fit the model. What S-value did you get? _____

9) What value of a best matches the shape of the curve? _____

Update the definition of `quad1`, click "Run" and re-fit the model. What S-value did you get? _____

10) Make any small changes you'd like, trying to get S as low as you can. Write your final definition below.

fun `f(x) :` _____ **end** S: _____

What does this model actually mean?

After experimenting, I came up with a quadratic model for this dataset showing that _____ (x-variable) is correlated to _____ (y-variable). The error in the model is described by an ***S-value*** of about _____ (S units), which is _____ (insignificant, moderate, significant, extreme) considering that _____ (y-variable) in this dataset range from _____ (lowest y-value) to _____ (highest y-value). The vertex of the parabola drawn by this model is a _____ (minima or maxima?) at about _____ (x, y) which means that _____

Before this point, as speed increases, `mpg` _____. *After* this point, as speed increases `mpg` _____

What Kind of Model? (Definitions)

Decide whether each representation describes a **linear** function, a **quadratic** function, or neither. If the function is quadratic, identify whether the **form** used is Vertex, Standard, or Factored.

$$f(x) = 3x^2 + 22$$

1) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

$$g(x) = 2(x - 11)(x - 243)$$

2) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

$$h(y) = 100 - 4y$$

3) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

$$z(x) = \frac{3}{5}x + 7$$

4) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

```
fun graph(x): 12 * x end
```

5) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

```
fun m(p): 2 * ((p - 5) * (p - 16)) end
```

6) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

$$r(s) = 42(s - 10)^2 - 3$$

7) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

```
fun f(x): (2 * num-sqr(x - 1)) + 15 end
```

8) Linear Quadratic Neither

If Quadratic, is it Vertex, Standard, or Factored?

If Quadratic, what does the form tell you?

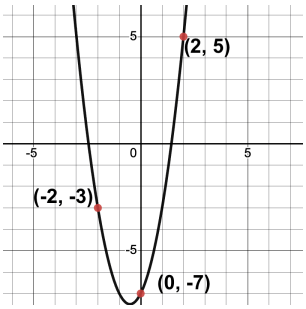
Matching Standard Form to Parabolas

Factored Form: $y = ax^2 + bx + c$

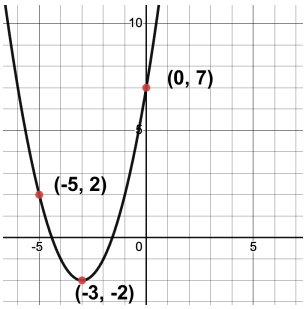
- a: determines whether the parabola opens up or down and how steep the curve is
- c: y-intercept

Match each definition below to the graph it describes.

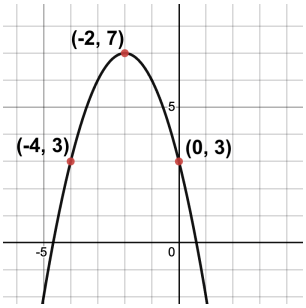
$y = -x^2 - 4x - 3$ **1** **A**



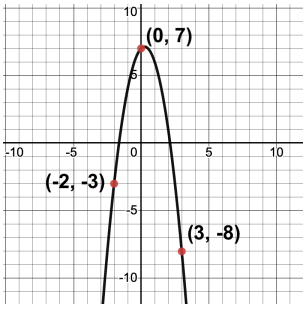
$y = 2x^2 + 2x - 7$ **2** **B**



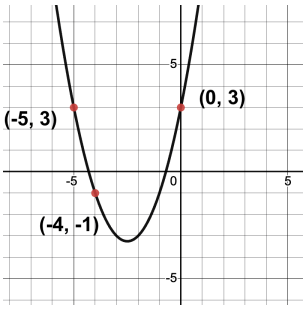
$y = x^2 + 5x + 3$ **3** **C**



$y = x^2 + 6x + 7$ **4** **D**



$y = -2x^2 + x + 7$ **5** **E**



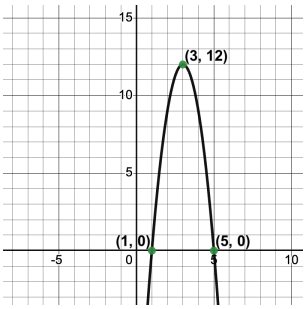
Matching Factored Form to Graphs

Factored Form: $y = a(x - r_1)(x - r_2)$

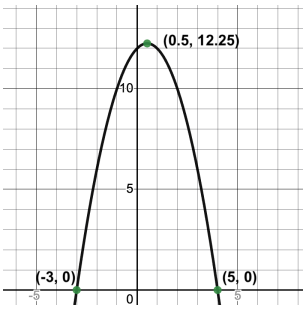
- a: determines whether the parabola opens up or down and how steep the curve is
- r_1 and r_2: roots, x-intercepts

Match each definition below to the graph it describes.

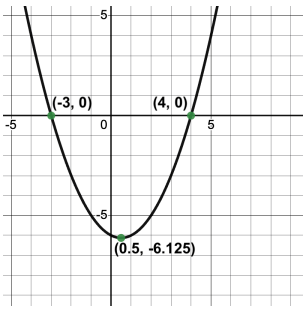
$y = 2(x - 1)(x + 5)$ **1** **A**



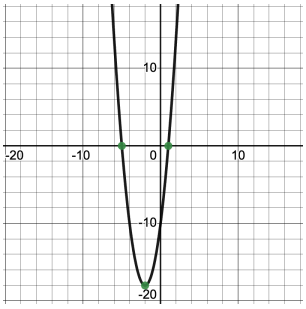
$y = (x + 3)(x + 4)$ **2** **B**



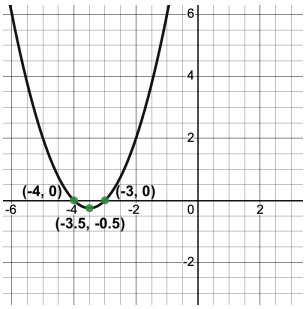
$y = - 3(x - 1)(x - 5)$ **3** **C**



$y = \frac{1}{2}(x + 3)(x - 4)$ **4** **D**



$y = - (x - 5)(x + 3)$ **5** **E**



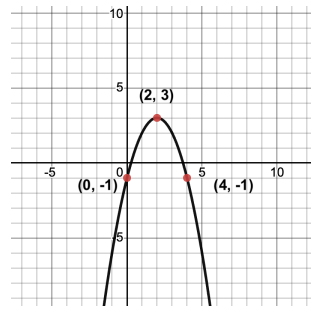
Matching Vertex Form to Graphs

Vertex Form: $y = a(x - h)^2 + k$

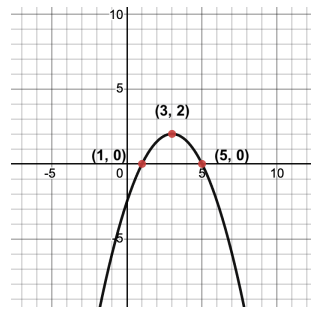
- a: determines whether the parabola opens up or down and how steep the curve is
- h: x-coordinate of the vertex
- k: y-coordinate of the vertex

Match each definition below to the graph it describes.

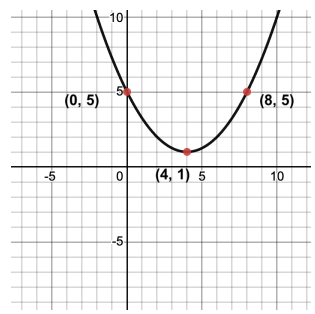
$f(x) = -0.5(x - 3)^2 + 2$ **1** **A**



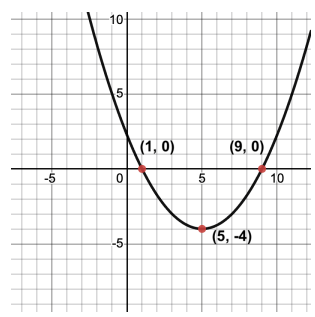
$g(x) = 2(x + 1)^2 - 4$ **2** **B**



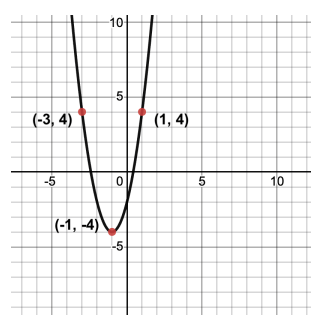
$h(x) = -(x - 2)^2 + 3$ **3** **C**



$j(x) = 0.25(x - 5)^2 - 4$ **4** **D**



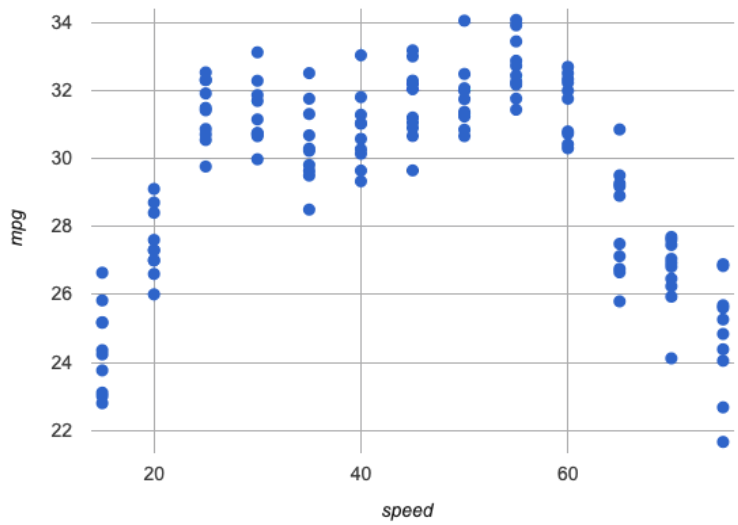
$k(x) = \frac{1}{4}(x - 4)^2 + 1)$ **5** **E**



Build a Model from Samples

*For this page, you'll need to open the [Fuel Efficiency Starter File]() on your computer. If you haven't already, select **Save a Copy** from the "File" menu to make a copy of the file that's just for you. **Read the comments at the top of the file**, which describe what each column in the dataset means.*

The **standard form** of a quadratic equation is $y = Ax^2 + Bx + C$



1) Choose a point from the **left-most column** of dots, and fill in the **standard form** equation below:

_____ = A(_____)^2 + B(_____) + C
y (mpg) x (speed) x (speed)

2) Choose a point from the **center-most column** of dots, and fill in the **standard form** equation below:

_____ = A(_____)^2 + B(_____) + C
y (mpg) x (speed) x (speed)

3) Choose a point from the **right-most column** of dots, and fill in the **standard form** equation below:

_____ = A(_____)^2 + B(_____) + C
y (mpg) x (speed) x (speed)

4) In the space below - or on another sheet of paper - solve this series of equations for A, B, and C:

5) Write your finished model in function and Pyret notation below, then define the function `f` in Pyret and try it out using `fit-model`!

Function Notation	Pyret Notation
$f(x) = A($_____$)^2 + B($_____$) + C$ (x (speed))	`fun f(x): ((_____ * num-sqr) + (_____ * x)) + _____ end`

Exploring the Covid Dataset

*For this page, you'll need to have the [Covid Spread Starter File]() open on your computer. If you haven't already, select **Save a Copy** from the "File" menu to make a copy of the file that's just for you.*

1) Take a look at the Definitions Area and find the "Notes on Columns". What is the start date for the data in this table? _____

2) Click "Run", and evaluate `covid-table` in the Interactions Area. What do you notice? _____

_____

3) Evaluate `MA1` in the Interactions Area. What does it return? _____

4) Evaluate `CT1`. What information do you learn? _____

5) Evaluate `NH1`. Why is it "unbound" and how could we make it work? _____

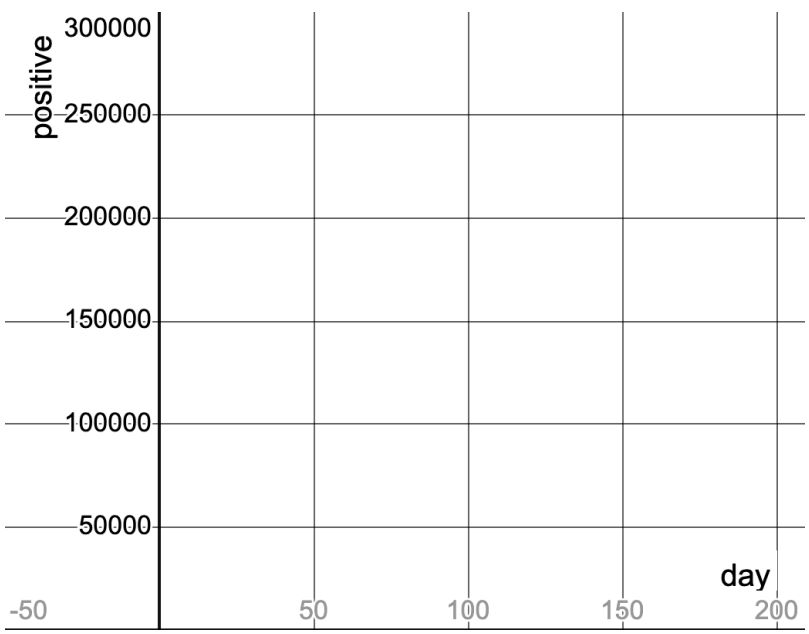
6) Define three new Rows called `NH1`, `RI1` and `VT1`, for New Hampshire, Rhode Island and Vermont. Click "Run" and test them out.

a. How many people in **Vermont** had tested positive by June 10th, 2020? _____

b. How many people in **New Hampshire** tested positive by June 10th, 2020? _____

c. How many people in **Rhode Island** tested positive by June 10th, 2020? _____

7) In Pyret, make a scatter plot showing the relationship between `day` and `positive`, using `state` as your labels, then sketch the resulting scatter plot below.



8) In which state did the number of cases grow *fastest*?

_____

9) In which state did the number of cases grow *slowest*?

_____

10) Are these strong or weak relationship(s)?

_____

11) What do you **Notice?** _____

_____

_____

12) What do you **Wonder?** _____

_____

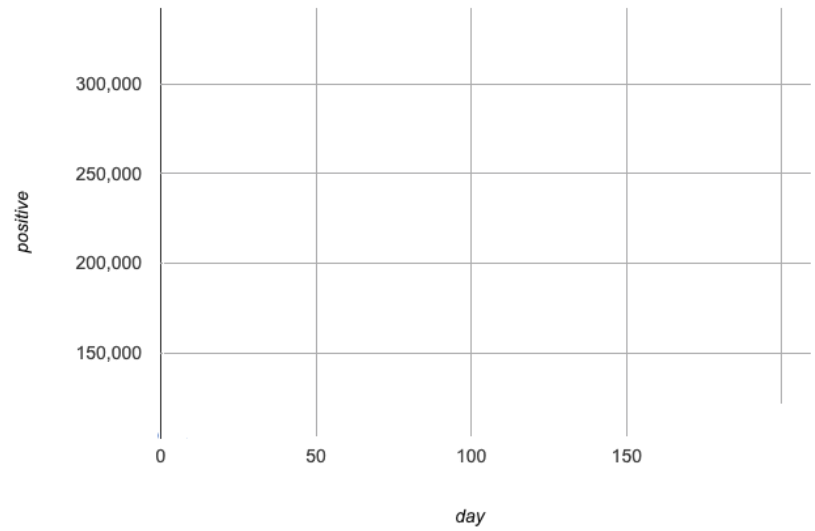
_____

Linear Models for MA-table

*For this page, you'll need to have the Covid Spread Starter File open on your computer. If you haven't already, select **Save a Copy** from the "File" menu to make a copy of the file that's just for you.*

This starter file defines a table *just for MA data*, called `MA-table`: `MA-table = filter(covid-table, is-MA)`

1) Make a scatter plot from `MA-table` showing the relationship between `day` and `positive`, using `state` as the labels. Sketch the plot on the right.



As we've seen, it's easy to fit a linear model to any dataset in Pyret, so let's start by testing how well a linear function could model this data.

2) Use `lr-plot` to obtain the best-possible **linear model** for the MA Covid dataset, and write it below:

y = _____ $S =$ _____

Note: Pyret uses `e` *for scientific notation. For example:* $2.46e3 = 2.46 \times 10^3 = 2460$

3) The optimized linear model for this dataset predicts an _____ (increase / decrease) of about _____ (slope) _____ (y-variable) per _____ (x-variable).

The error in the model is described by an ***S-value*** of about _____ (S) _____ (units), which is a _____ (poor, ok, good) fit considering that

_____ (y-variable) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

4) Change the definition of the `linear` function in the Covid Spread Starter File to match the model produced by `lr-plot` and "Save".

5) Do you think a linear function is a good model for this dataset? Why or why not? _____

_____

_____

_____

★ What do you think the code that defines `MA-table` is actually doing? _____

_____

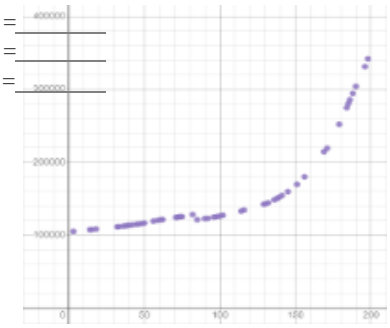
Quadratic Models for MA-table

Fitting the Model Visually $f(x) = a(x - h)^2 + k$

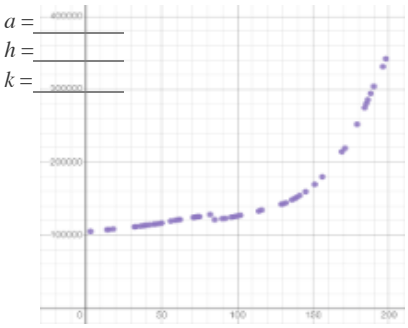
*For this section, you'll need to have **Slide 1: Quadratic Model for MA** of **Modeling Covid Spread (Desmos)** open on your computer.*

1) Try changing the values of a, h and k to find three promising quadratic models, graphing each one and labeling your values in the grids below.

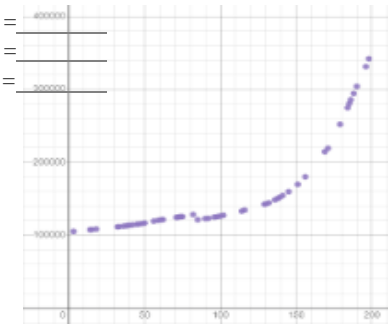
$a =$ _____
$h =$ _____
$k =$ _____



$a =$ _____
$h =$ _____
$k =$ _____



$a =$ _____
$h =$ _____
$k =$ _____



2) Do your quadratic models open up or down? _____. What does that tell us about a? _____.

3) Describe one of your models: Where is the vertex? (_____, _____) What is the horizontal shift? _____ The vertical shift? _____
x y h k

4) Which quadratic form would be the easiest to fit to this data? standard □ factored □ vertex□

Fitting the Model Programmatically $f(x) = a(x - h)^2 + k$

For this section, open your copy of the Covid Spread Starter File.

5) In the space below, define `quadratic1` to be the first model you fit in Desmos.

`fun quadratic1(x): (` _____ `* (num-sqr( x -` _____ `))) +` _____ `end`
a h k

6) Return to Covid Spread Starter File and update the definitions for `quadratic1`, `quadratic2`, and `quadratic3`.
Then click "Run" to load your updated definition.

7) Use `fit-model` to determine the S-value of each model using the `MA-table`.
Hint: If you forgot the contract for `fit-model`, look it up in the contracts pages!

S for `quadratic1`: _____ S for `quadratic2`: _____ S for `quadratic3`: _____

What does this model actually mean?

After experimenting, the best quadratic model I came up with for this dataset shows that _____ are correlated to _____.
x-variable y-variable

The vertex of the parabola drawn by this model is a _____ at about _____, which predicts that _____
minima or maxima? (x, y)

_____.

The error in the model is described by an ***S-value*** of about _____, which is a _____
S units bad, ok, good

fit considering that _____ in this dataset range from _____ to _____.
y-variable lowest y-value highest y-value

Are Quadratic Models a Good Fit for This Data?

8) Would you feel good about making predictions based on these models? Why or why not? _____

_____

_____

What Kind of Model? (Tables)

Decide whether each table is best described by a linear, quadratic, or exponential model. **If the model is exponential:** What is the growth factor? Doubling (factor of 2)? Tripling (factor of 3)? Factor of 5? 10?
HINT: Can you draw the arrows to calculate the first difference? The second? *What does it mean if neither one is constant?*

x	y
1	5
2	10
3	15
4	20
5	25
6	30
7	35

1) Linear Quadratic Exponential _____ factor

x	y
0	10
1	100
2	1000
3	10000
4	100000
5	1000000
6	10000000

2) Linear Quadratic Exponential _____ factor

x	y
70	-169
71	-126
72	-81
73	-34
74	15
75	66
76	119

3) Linear Quadratic Exponential _____ factor

x	y
-3	36
-2	16
-1	4
0	0
1	4
2	16
3	36

4) Linear Quadratic Exponential _____ factor

x	y
0	3
1	6
2	12
3	24
4	48
5	96
6	192

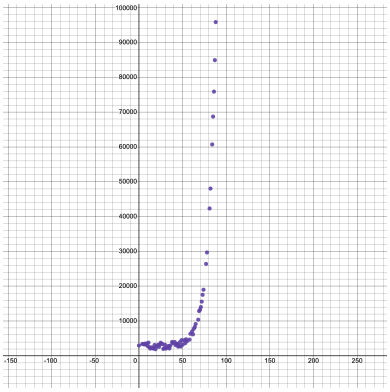
5) Linear Quadratic Exponential _____ factor

x	y
-5	466656
-4	7776
-3	1296
-2	216
-1	36
0	6
1	1

★ Linear Quadratic Exponential _____ factor

What Kind of Model? (Graphs & Plots)

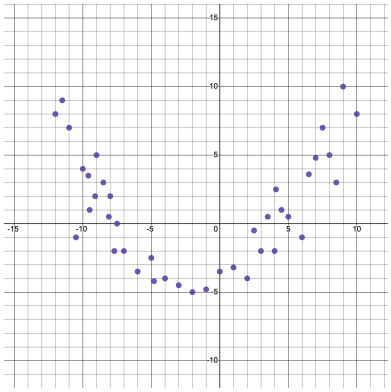
Are these scatter plots best be described by linear, quadratic, or exponential models? **If it's exponential, draw the asymptote!**



1) Linear Quadratic Exponential

How did you know? _____

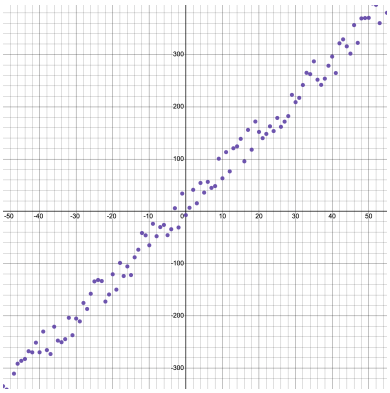
_____



2) Linear Quadratic Exponential

How did you know? _____

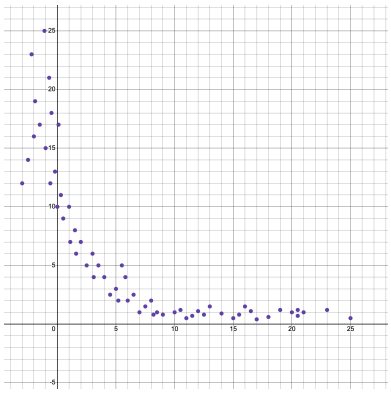
_____



3) Linear Quadratic Exponential

How did you know? _____

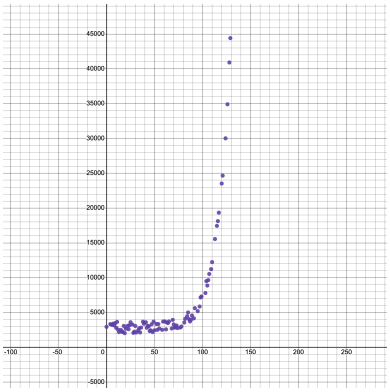
_____



4) Linear Quadratic Exponential

How did you know? _____

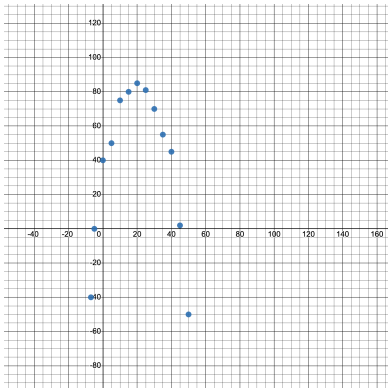
_____



5) Linear Quadratic Exponential

How did you know? _____

_____



6) Linear Quadratic Exponential

How did you know? _____

_____

Graphing Exponential Models: $f(x) = ab^x + k$

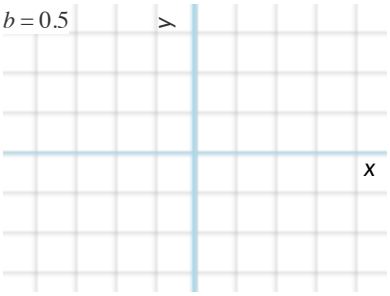
For this page, you'll need to have ***Slide 3: Exploring Exponential Models*** *of* ***Modeling Covid Spread (Desmos)*** *open on your computer.*
The curve you'll see is the graph of $h(x) = 2^x$. Another curve $f(x)$ is hiding behind it with identical coefficients: $k = 0$, $b = 2$ and $a = 1$.

Base b

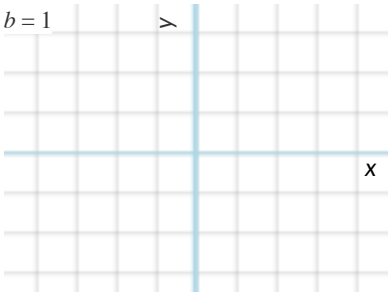
1) Make sure $k = 0$ and $a = 1$. Experiment with b. **For what values of b is the function *undefined*** , with the line disappearing? _____

2) Keeping $a = 1$ and $k = 0$, change b to 0.5, 1, and 2, graphing each curve below. For each curve, **label the coordinates at x=1, 2, and 3.**

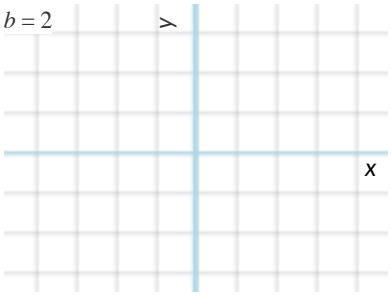
$b = 0.5$



$b = 1$



$b = 2$



3) What does b tell us about an exponential function, when $b > 1$? _____

_____

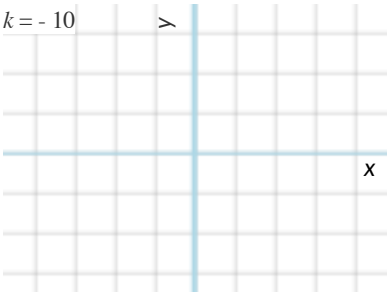
4) What does b tell us about an exponential function, when $0 < b < 1$? _____

_____

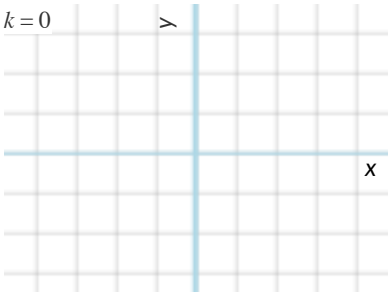
Vertical Shift…and Horizontal Asymptote k

5) Keeping $a = 1$ and $b = 2$, try changing the value of k to -10, 0, and 10, graphing each curve in the squares below. For each curve, find and label the y-value where the curve is "most horizontal", then **draw a horizontal line at that y-value**.

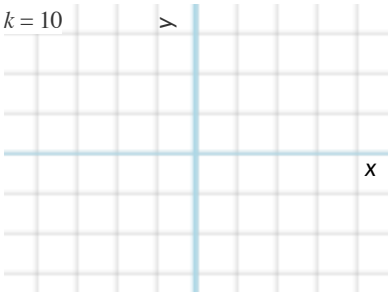
$k = -10$



$k = 0$



$k = 10$



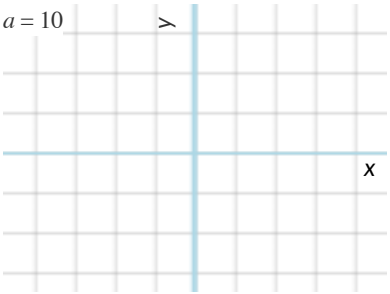
6) What does k tell us about an exponential function? _____

_____

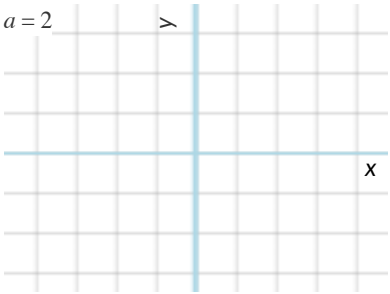
Initial Value a

7) Set $k = 0$ and $b = 2$. Change the value of a to 10, 2, and -5, graphing each curve in the squares below.
For each curve, label the y-intercept (x=0).

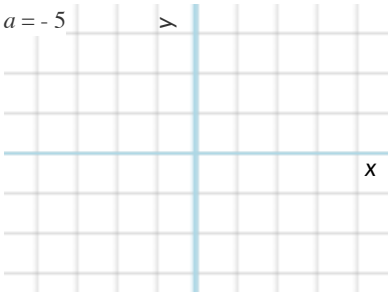
$a = 10$



$a = 2$



$a = -5$



8) What does a tell us about an exponential function? _____

What Kind of Model? (Descriptions)

Decide whether each situation is best described by a linear, quadratic, or exponential function.
If the function is exponential: What is the growth factor. Is it doubling (factor of 2)? Tripling (factor of 3)? Factor of 5? 10?

Car Values

A particular kind of car sells for $32,000, and its resale value drops by 12.5% each year.

1) Is the function increasing or decreasing? _____

2) When the car is brand-new (x=0), how much is it worth? _____

3) How much is it worth after…

(1 year) x=1	(2 years) x=2	x=3	x =4

4) What is the **form** of this function? linear □ quadratic □ exponential□

5) **If it's exponential**,

Fill in the coefficients to write a function that shows the value of the car after a given number of years:

$f(x) =$ _____ _____ ^{x} + _____

initial value a growth factor b horizontal asymptote k

Is it exponential *growth* ? □ or *decay* ? □

Lemonade Stand

Sally is selling lemonade, for $1.25 a glass in hopes of finally be able to get the power drill she's been wanting. She starts with $20 cash.

6) Is the function increasing or decreasing? _____

7) When Sally starts the day (x=0), how many dollars does she have? _____

8) How many dollars will she have after…

(first sale) x=1	(second sale) x=2	x =3	x = 4

9) What is the **form** of this function? □ linear □quadratic □ exponential

10) **If it's exponential**,

Fill in the coefficients to write a function that shows how much Sally has saved after a given number of sales:

$f(x) =$ _____ _____ ^{x} + _____

initial value a growth factor b horizontal asymptote k

Is it exponential *growth* ? □ or *decay* ? □

What Kind of Model? (Definitions)

Decide whether each representation describes a linear, quadratic, or exponential function.
If the function is exponential: Identify the growth factor and the initial value.

$$f(x) = 6x^2 - 5$$

1) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$\text{miles}(\text{hours}) = \frac{22 \times \text{hours} + 14}{12 - 9}$$

2) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$\text{cost}(w) = 5(1.2^w) + 16$$

3) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$t(g) = 42 - 2g^2$$

4) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$\text{price}(d) = d^2 + 6d$$

5) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$j(x) = \frac{1}{2}^x + 22$$

6) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$f(a) = 20000 - 4.1^a$$

7) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

$$g(x) = 8(3^{-4x})$$

8) Linear Quadratic Exponential

How did you know? _____

_____

If it's exponential, what's the _____ (growth factor) _____ (initial value)?

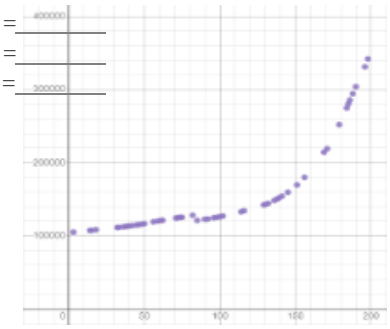
Exponential Models: $f(x) = ab^x + k$

Fitting the Model Visually - MA

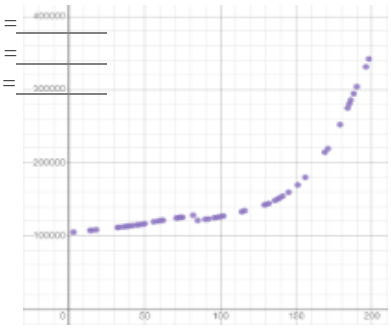
*For this section, you'll need to have **Slide 4: Exponential Model for MA** of **Modeling Covid Spread (Desmos)** open on your computer.*

1) Try changing the value of k, then a, then b to find three promising exponential models, graphing each one and labeling your values on the grids below.

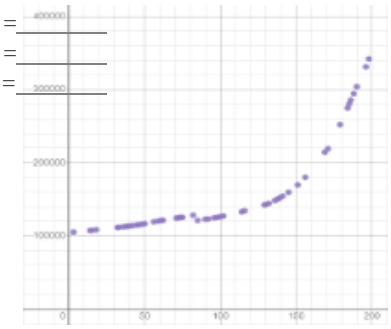
$a =$ _____
$b =$ _____
$k =$ _____



$a =$ _____
$b =$ _____
$k =$ _____



$a =$ _____
$b =$ _____
$k =$ _____



Fitting the Model Programmatically - MA

For this section, open your copy of the Covid Spread Starter File.

2) In the space below, define `exponential` for one of the models you fit in Desmos.

`fun exponential(x): (` _____ `* num-expt(` _____ `, (~1 * x))) +` _____ `end`
(a) (b) (k)

Two Notes on this function definition:

- `num-expt` *is the function that we use for exponents. It takes in 2 numbers: the base and the power, in this case b and x.*
- `(~1 * x)` *at first it may appear that x is being multiplied by negative 1, but it is actually being multiplied by* **~1** *(literally the value "roughly 1"). This tells Pyret to round off the calculation, prioritizing **speed** over **precision** to get a result that is "roughly accurate". We've added this to the function definition so that you won't have to wait for several minutes for Pyret to run* `fit-model` *to get an answer for question 4.*

3) Update the definition for `exponential` in the Definitions Area and click "Run" to reload it.
Then use `fit-model` to determine how closely `exponential` fits the `MA-table` and fill in the blanks below to interpret the model.
Hint: If you forgot the contract for `fit-model`*, look it up in the contracts pages!*

According to this exponential model, on June 9, 2020 (day zero) there were about _____ (a) + _____ (k) _____ (y-units) in MA, for a total of about _____ (a + k). This number grew exponentially, increasing by a factor of _____ (Growth Factor: b) or _____ (Growth Rate: (b - 1) × 100) % every day.

The error in the model is described by an ***S-value*** of about _____ (S) _____ (units), which is a(n) _____ (bad, ok, good) model considering that _____ (y-units) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

4) Estimate how many positive cases there will be after X days by **looking at graph with your eyes**, then use your model to find the answer.

Using your...	Eyes	Model	*Using your...*	Eyes	Model	*Using your...*	Eyes	Model
50 days	_____	_____	150 days	_____	_____	250 days	_____	_____
350 days	_____	_____	450 days	_____	_____	550 days	_____	_____

★ Rewrite the model to make Pyret do these calculations with extreme precision. (Remove the part where it multiplies by ~1 .)
WARNING: Be sure to save your work first, as there's a good chance this will lock up your browser and require force-quitting!

What changed? _____

Data scientists perform calculations to do things like send satellites to far-away planets, or analyze large populations of a billion or more. You know that the pros of using ~1 involve speed. **What are the potential downsides of using ~1 to speed up a calculation?**

_____

_____

Modeling Other States

*For this page, you'll need to have the Covid Spread Starter File open on your computer. If you haven't already, select **Save a Copy** from the "File" menu to make a copy of the file that's just for you.*

1) Find the function called `is-MA` in the Definitions Area under "Define some helper functions" and read the comments carefully!

a. What is the Domain of `is-MA` ? _____ What is its Range? _____

b. What do you *think* `is-MA(MA1)` will evalute to? _____. `is-MA(CT1)` ? _____. `is-MA(ME1)` ? _____

Try typing each of the `is-MA` *expressions into the Interactions Area on the right and confirm you were correct.*

2) Find `MA-table` in the Definitions Area under "Define some grouped and/or random samples". What is that code doing? _____

_____

_____

3) **Define a new function** `is-VT` and **create a new grouped sample** called `VT-table` .
Hint: You can use the code for `is-MA` *and* `MA-table` *as a model.*

Modeling VT

*For this section, in addition to Pyret, you will need to have **Slide 5: Exponential Model for VT** of **Modeling Covid Spread (Desmos)** open on your computer.*

4) Use `lr-plot` to obtain the best-possible linear model for the relationship between `day` and `positive` in the `VT-table` , then fill in the blanks below:

The optimized linear model for this dataset predicts an _____ (increase / decrease) of about _____ (slope) _____ (y-variable) per _____ (x-variable).

The error in the model is described by an ***S-value*** of about _____ (S) _____ (units), which is _____ (insignificant, moderate, significant, extreme)

considering that _____ (y-variable) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

5) Use **Slide 5: Exponential Model for VT** of **Modeling Covid Spread (Desmos)** to come up with the best exponential model you can for the Vermont dataset, and write it below:

_____

6) Add a definition for `exponential-VT` to the Definitions area of Covid Spread Starter File using the model you just found.

- Click "Run" to load your definition.
- Then fit the model using `VT-table`

According to this exponential model, on June 9, 2020 (day zero) there were about _____ (a) + _____ (k) _____ (y-units) in VT, for a total

of about _____ (a + k). This number grew exponentially, increasing by a factor of _____ (Growth Factor: b) or _____ (Growth Rate: (b - 1) × 100) % every

day. The error in the model is described by an ***S-value*** of about _____ (S) _____ (units), which is

_____ (insignificant, moderate, significant, extreme) considering that _____ (y-units) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

7) Are exponential models a good fit for this data? Why or why not? _____

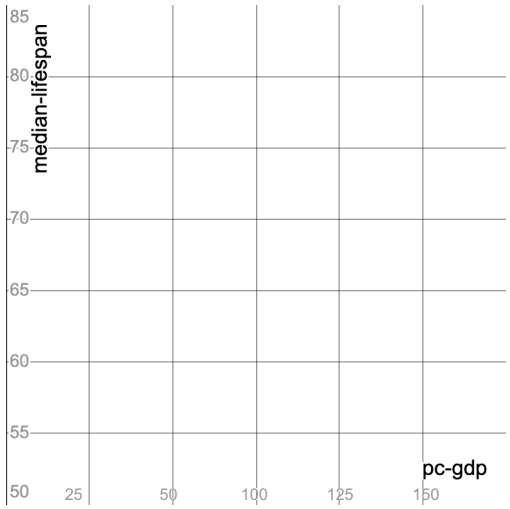
_____

Exploring the Countries Dataset

*For this section, you'll need the [Countries of the World Starter File]() open on your computer. If you haven't already, select **Save a Copy** from the "File" menu to make a copy of the file that's just for you. The columns in this dataset are described below:*

- **country** - name of the country
- **gdp** - total Gross Domestic Product of the country. GDP is often used to measure the economic health of a country.
- **population** - number of people in the country
- **pc-gdp** - the average GDP *per-person, in thousands* of $US
- **has-univ-healthcare** - indicates if the country has universal healthcare
- **median-lifespan** - the median life expectancy of people in the country

1) Make a scatter plot showing the relationship between `pc-gdp` and `median-lifespan`, and sketch its plot below.



2) What do you **Notice?** _____

_____

_____

3) What do you **Wonder?** _____

_____

_____

4) Are there any countries that stand out? Why or why not? _____

_____

_____

5) Suppose a wealthy country is suffering heavy casualties in a war. Draw a star on the plot, showing where you might expect it to be.

6) Do you think you see a relationship? If so, describe it. Is it linear or nonlinear? Strong or weak?

_____

_____

_____

Fitting Models for the Countries Dataset

*For this page you will be working with both the [Countries of the World Starter File]() and the **Desmos** file **Fitting Wealth-v-Health and Exploring Logarithmic Models**.*

Find linear, quadratic and exponential models for the relationship between `pc-gdp` and `median-lifespan`. As you find each model:

- update the corresponding definition in the [Countries of the World Starter File]()
- click "Run" to load your new definition
- use `fit-model` to calculate the ***S-value*** *Hint: If you forgot the contract for* `fit-model` *(to calculate S), look it up in the [contracts pages]()!*

1) Find the optimized **linear model** for this data using `lr-plot`.

$linear(x) =$ _____ (slope (m)) x + _____ (y-intercept / vertical shift) _____ (S-value)

The optimized linear model for this dataset predicts that a _____ (x-units) _____ (increase / decrease) in __per-capita gdp__ (x-variable) will increase

_____ (y-variable) by _____ (y-units). The error in the model is described by an S - *value* of about _____ (S) _____ (y-units),

which is _____ (insignificant / reasonable / significant / extreme) considering _____ (y-units) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

2) Find the best **quadratic model** you can, using the second slide *(Wealth-v-Health Quadratic)* in the Desmos activity.

$quadratic(x) =$ _____ (quadratic coefficient (a)) $(x -$ _____ (horizontal shift (h)) $)^2 +$ _____ (vertical shift (k)) _____ (S-value)

The vertex of the parabola drawn by my model is a _____ (minima or maxima?) at about (_____ (x, y)).

- *Before* this point, as _____ (x-variable) increases, _____ (y-variable) _____ (increases or decreases?).
- *After* this point, as _____ (x-variable) increases, _____ (y-variable) _____ (increases or decreases?).

The error in the model is described by an S - *value* of about _____ (S) _____ (y-units), which is _____ (insignificant / reasonable / significant / extreme)

considering _____ (y-units) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

3) Find the best **exponential model** you can, using the third slide *(Wealth-v-Health Exponential)* in the Desmos activity.

$exponential(x) =$ _____ (initial value (a)) (_____ (growth factor (b)) ^{x}) + _____ (vertical shift (k)) _____ (S-value)

According to this exponential model, a country with a _____ (x-variable) of zero _____ (x-unit) would have a

_____ (y-variable) of _____ (a) + _____ (k) _____ (y-units), for a total of about _____ (a + k). This number grows exponentially,

increasing by a factor of _____ (Growth Factor: b) or _____ (Growth Rate: (b - 1) × 100) % with every _____ (x-unit) increase in _____ (x-variable).

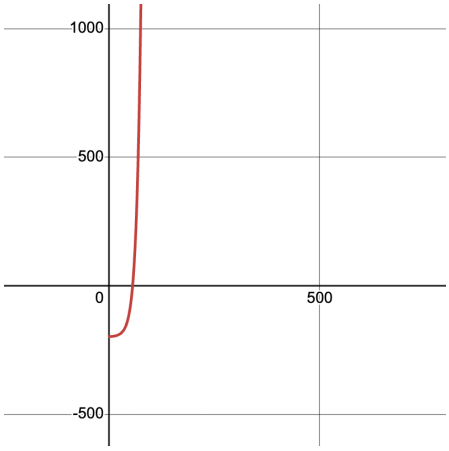
The error in the model is described by an S - *value* of about _____ (S) _____ (y-units), which is _____ (insignificant / reasonable / significant / extreme)

considering _____ (y-units) in this dataset range from _____ (lowest y-value) to _____ (highest y-value).

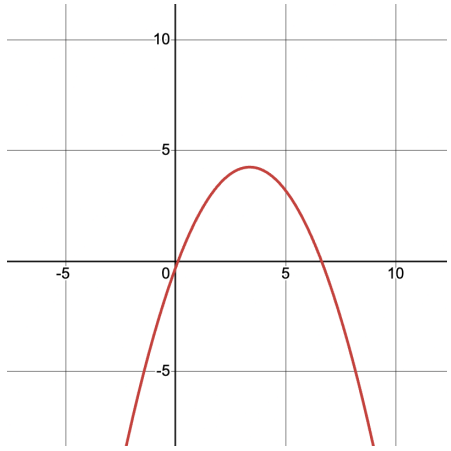
4) Are any of these models a good fit for this data? Why or why not?

What Kind of Model? (Graphs & Plots)

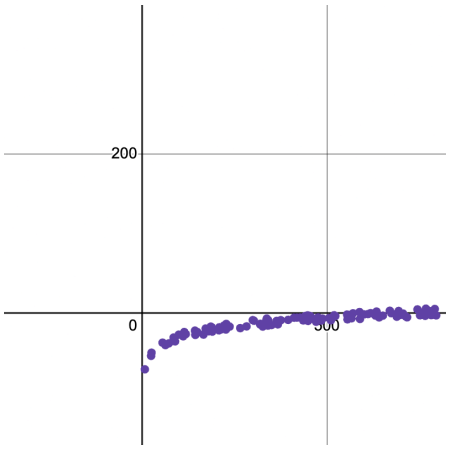
Decide whether each representation is best described by a quadratic, exponential, or logarithmic function. **If you think it's exponential OR logarithmic, draw a diagonal line for $y = x$, and then sketch the reflection of the curve.**



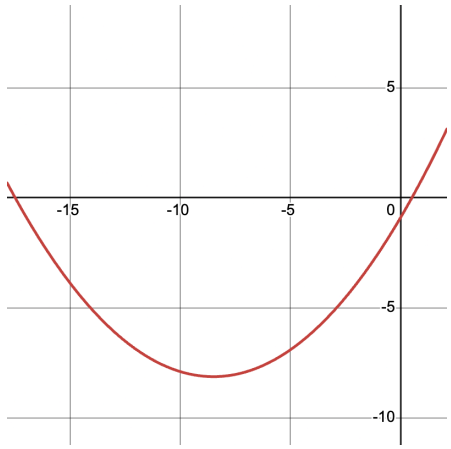
1) Quadratic Exponential Logarithmic



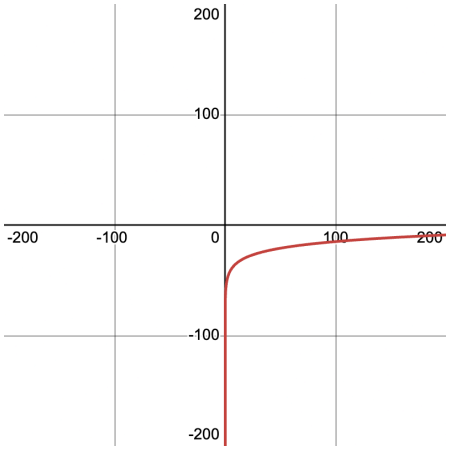
2) Quadratic Exponential Logarithmic



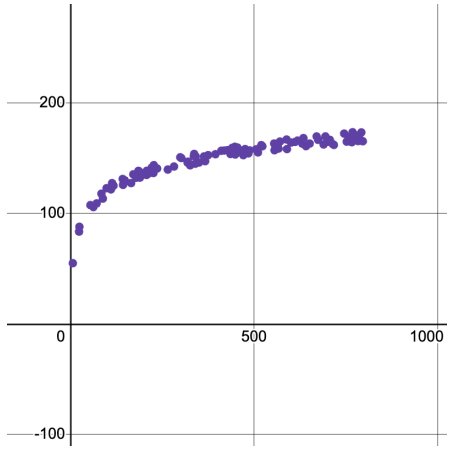
3) Quadratic Exponential Logarithmic



4) Quadratic Exponential Logarithmic



5) Quadratic Exponential Logarithmic



6) Quadratic Exponential Logarithmic

What Kind of Model? (Tables)

Decide whether each representation is best described by a quadratic, exponential, or logarithmic function.
If the function is exponential, find the *base* (also called the *growth factor*): How much does y increase (2x? 10x?) for a single increase in x?
If the function is logarithmic, find the *base*: How much does x need to increase (2x? 10x?) just to get a single increase in y?
HINT: Can you draw the arrows to calculate the first difference? The second? *What does it mean if neither one is constant?*

x	y
1	0
10	1
100	2
1000	3
10000	4
100000	5
1000000	6

1) Quadratic Exponential ____ base Logarithmic ____ base

x	y
0	1
1	10
2	100
3	1000
4	10000
5	100000
6	1000000

2) Quadratic Exponential ____ base Logarithmic ____ base

x	y
70	-169
71	-126
72	-81
73	-34
74	15
75	66
76	119

3) Quadratic Exponential ____ base Logarithmic ____ base

x	y
5	1
10	2
20	3
40	4
80	5
160	6
320	7

4) Quadratic Exponential ____ base Logarithmic ____ base

x	y
-3	36
-2	16
-1	4
0	0
1	4
2	16
3	36

5) Quadratic Exponential ____ base Logarithmic ____ base

x	y
1	0
6	1
36	2
216	3
1296	4
7776	5
466656	6

6) Quadratic Exponential ____ base Logarithmic ____ base

Evaluating Logarithmic Expressions

	Expressions	Translation	Evaluates to:
1	$\log_2(8)$	"The power you raise 2 to get 8"	3
2	$\log_2(1)$	"The power you raise 2 to get 1"	0
3	$\log_5(25)$	"The power you raise _____ to get _____"	
4	$\log_5(1)$	"The power you raise _____ to get _____"	
5	$\log_3(81)$	"The power you raise _____ to get _____"	
6	$\log_3(1)$	"The power you raise _____ to get _____"	
7	$\log_2(16)$		
8	$\log_2(32)$		
9	$\log_{10}(1000)$		
10		"The power you raise 0.1 to get 0.01"	
11		"The power you raise 4 to get 64"	
12		"The power you raise 4 to get 1"	

Graphing Logarithmic Models: $f(x) = a \log_b x + k$

Use this page with ***Slide 4: Exploring Logarithmic Functions*** *of* ***Fitting Wealth-v-Health and Exploring Logarithmic Models (Desmos)****.*

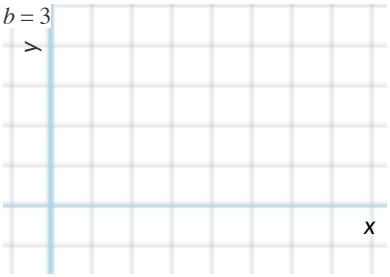
- The **blue curve** is the graph of $h(x) = 1 \log_2 x + 0$. Its constants will remain set at $a = 1$, $b = 2$, and $k = 0$.
- You can modify the **red curve** $g(x)$ (which is hiding behind $h(x)$!) by changing its coefficients: a, b, and k.

Base b

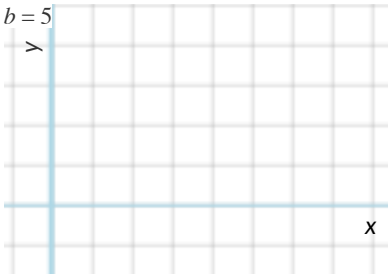
Keep ***k*** *at 0 and* ***a*** *at 1. Change the value of* ***b*** *as indicated on each grid below.*

1) Sketch each graph and **label the coordinates where** $x = 1$, $y = 1$, $y = 2$ **and** $y = 3$.

$b = 3$



$b = 5$



$b = 10$

2) How does the value of b impact the shape of a logarithmic function? _____

3) What connections can you draw between the value of b and exponents? _____

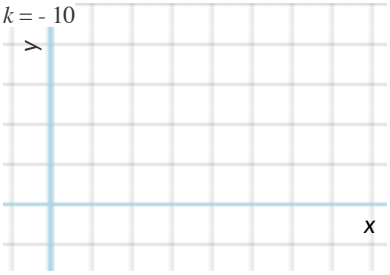
_____

Vertical Shift k

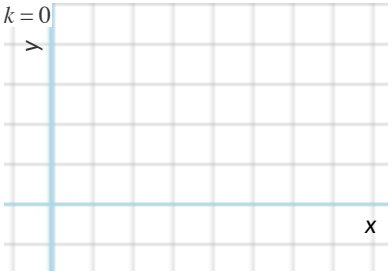
Set ***a*** *to 1 and* ***b*** *to 2. Change the value of* ***k*** *as indicated on each grid below.*

4) Sketch each graph and **label the coordinate where** $x = 1$.

$k = -10$



$k = 0$



$k = 10$

5) How does the value of k impact the shape of a logarithmic function? _____

6) Why does $y = k$ when $x = 1$? _____

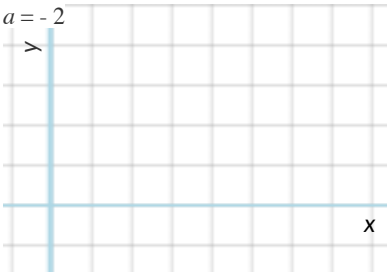
Logarithmic Coefficient a

Set ***k*** *to 0 and* ***b*** *to 10, then zoom out out so you can see as far as $x = 1{,}000$.*

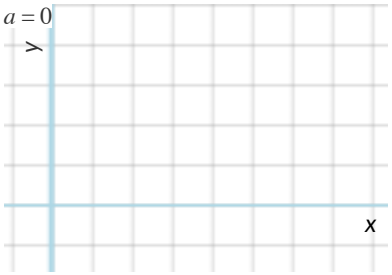
Change $h(x)$ to $h(x) = 1 \log_{10}(x) + 0$ so that the blue curve lands on top of the red curve.

7) In each graph, **label the coordinates where** $x = 10$ **and** $x = 100$ **and** $x = 1000$.

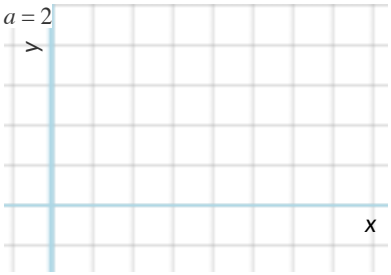
$a = -2$



$a = 0$



$a = 2$



8) What is the value of x when $1\log_2(x) = 4$? _____ What about when $2 \log_4(x) = 4$? _____ When $3 \log_8(x) = 4$? _____

★ How are a and b related? _____

What Kind of Model? (Descriptions)

Decide whether each situation describes a quadratic, exponential, or logarithmic function. **HINT:** draw a table and plug in some points!

1) Earthquakes release enormous amounts of energy, which we can compare to the energy released by blowing up pounds of dynamite. For example, $\text{richter}(12{,}000) = 4.0$, meaning that the force of blowing up 12,000 pounds of dynamite produces a 4.0 on the Richter scale! $\text{richter}(400{,}000) = 5.0$, $\text{richter}(12{,}540{,}000) = 6.0$, and $\text{richter}(398{,}000{,}000) = 7.0$.

Quadratic Exponential Logarithmic

2) A car accelerates at a constant rate of 5mph/s. After 1 second, $\text{distance}(1) = 2.5\text{miles}$.
$\text{distance}(2) = 10$, $\text{distance}(3) = 22.5$, and $\text{distance}(4) = 40$

Quadratic Exponential Logarithmic

3) Moore's law says that the number of transistors in a microprocessor will double roughly every 1.5 years. Starting with 16 transistors, how many years will it take to reach 4,294,967,296 transistors?

Quadratic Exponential Logarithmic

4) The population of a colony of bacteria can double every 20 minutes, as long as there is enough space and food. Starting with 1 bacteria, $f(20) = 2, f(40) = 4, f(60) = 8, f(80) = 16\ldots$

Quadratic Exponential Logarithmic

5) Sequan puts \$100 in a savings account, earning 4% interest. After a year, $\text{savings}(1) = \$104$.
$\text{savings}(2) = \$108.16$, $\text{savings}(2) = \$112.49\ldots$

Quadratic Exponential Logarithmic

6) If the *width and length* of a rectangle doubles, how much does the *area* change?

Quadratic Exponential Logarithmic

Changing the Scale

For this page, you'll need to have ***Slide 5: Wealth-v-Health (Logarithmic)*** *of* ***Fitting Wealth-v-Health and Exploring Logarithmic Models (Desmos)*** *and Countries of the World Starter File open on your computer.*

Fitting a Logarithmic Model $f(x) = a \log_b x + k$

Open the `Data Table` folder by clicking on the triangle (▶)

- x_1 is the per-capita income for each country in thousands of $US, and y_1 is the median lifespan.
- Next to y_1 you'll see a dark circle with spots (∷) inside. If the circle is dark, that means that those points are visible on our graph. Click the circle to "turn off" those dots, then click it again to turn them back on.
- Move the graph by clicking and dragging the background.
- Notice that a magnifying glass (⊕) appears to the bottom left of the table. (You may have to scroll down to see the bottom of the table!) Clicking on the magnifying glass resizes/rescales the graph to fit all the points in the table.

1) Write the numbers you see along the x-axis, from left to right: _____

Continue this pattern - what would the next three numbers be? _____

2) Circle the type of function that describes this pattern: Linear Quadratic Exponential

3) Move the sliders for a and c to create the best-fitting logarithmic model you can find, and write it below.
Note: The Bootstrap Pyret function `log` *always uses* $b = 10$.

$logarithmic(x) =$ _____ (log coefficient (a)) $\log_{10}(x) +$ _____ (vertical shift (k))

```
fun logarithmic(x): (_______ * log(x)) + _______ end
```

4) Modify `logarithmic(x)` in Countries of the World Starter File to define this model, and fit it using `fit-model`.

The error in the model is described by an S - *value* of about _____ (S) _____ (units), which is _____ (insignificant / reasonable / significant / extreme)

considering _____ (y-variable) in this dataset ranges from _____ (lowest y-value) to _____ (highest y-value).

Scaling the x-Axis

- Click on the wrench button (🔧) in the top-right corner of the Desmos graph to **Open the "Graph Settings" window.**
- **Expand the "More Options" section** by clicking the triangle (▶).
- **Change the x-axis scale** from `Linear` to `Logarithmic`.
- Adjust the view by zooming and dragging the graph to get all of the points in view on the screen and filling most of it.

5) What is the shape of the point cloud *now*, after changing the scale? Linear Quadratic Exponential

6) Write the numbers you see along the x-axis, from left to right: _____

Continue this pattern - what would the next three numbers be? _____

7) Circle the type of function that describes this pattern: Linear Quadratic Exponential

8) Adjust the sliders for a and c to improve the model. *Toggle back and forth between* `logarithmic` *and* `linear` *x-axis scales as you work.*
When you are satisfied with your model, record both forms of the definition below.

$logarithmic2(x) =$ _____ (log coefficient (b)) $\log_{10}(x) +$ _____ (vertical shift (k))

```
fun logarithmic2(x): (_______ * log(x)) + _______ end
```

9) Modify the definition of `logarithmic2(x)` in Pyret to match this model. Use the `fit-model` function to find its ***S-value***: _____

10) Why do you think transforming the **x-axis** makes our data look linear? _____

_____

Transforming the Data

For this page, you'll need to have ***Slide 6: Wealth-v-Health (Transformed)*** *of* ***Fitting Wealth-v-Health and Exploring Logarithmic Models (Desmos)*** *open on your computer.*

- **Find the `Wealth vs. Health` folder, which is open at the top of the expression list**
- This is the same table we've seen before, and the "points" circle (∴) shows us that these dots are "on" and visible.
- Underneath the `Wealth vs. Health` folder, you'll see a **function** $g(x)$ and a **list** y_2 defined to be the same as y_1.
- **Open the second folder, called `Log(Wealth) vs. Health`, by clicking on the triangle (▶)**

1) Compare the two tables. (Here is a side by side comparison of how they each begin.)

Wealth vs. Health **Log(Wealth) vs. Health** **Compare the 2 tables. What do you notice? What do you wonder?**

x_1	y_1
1.99051	52.1
11.76559	78.6
15.19295	77.2
6.26897	60.6
24.95776	76.9
20.5888	77.5

$g(x_1)$	y_2
0.29896436	52.1
1.0706137	78.6
1.1816421	77.2
0.79719619	60.6
1.3972056	76.9
1.313631	77.5

2) Read the comments in rows 3 to 6 of the Desmos file. Where do the x-values in the second table come from? _____

_____

3) Why is the second column of both tables the same? _____

- **Turn the points for the first table OFF, then turn the points for our new table ON.**
  Our log transformation is so drastic that it looks like all the black datapoints are smashed against the y-axis!
- **Rescale** the graph (⊕) to see the cloud.

4) What is the shape of this point cloud? linear □ quadratic □ exponential □

5) Why do you think transforming the **x-values** make our data look linear? _____

_____

_____

6) Through trial and error, move the sliders for m and b to create the best-fitting linear model you can find, and write it below.

$f(x) =$ _____ $x +$ _____
slope (m) y-intercept / vertical shift

Let's compare the coefficients from your models.

Linear (From above) _____ slope (m) _____ y-intercept / vertical shift

Logarithmic (From Changing the Scale) _____ log coefficient (a) _____ vertical shift (k)

7) How are they similar? _____

_____

Logarithmic Models

Open your copy of the [Countries of the World Starter File]() and click "Run".

Transforming: From Logarithmic *Plots* to Linear Ones

1) Find the definition of `g(r)`. What does this function do? _____

_____

2) Find the Contract for `build-column` on the Contracts Page.

What is its **Range**? _____ What is its **Domain**? _____

3) At the end of the program, you'll find this code:

```
countries-transformed = build-column(countries-table, "log(pc-gdp)", g)
```

What do you think it does? _____

_____

4) Click "Run", and evaluate `countries-transformed` in the Interactions Area on the right to test it out!

a. What is different about this Table? *Hint: Find the last column!* _____

b. Where did the column on the right come from? _____

5) Use this new table to make an `lr-plot` comparing `log(pc-gdp)` and `median-lifespan`, with `country` as the label. Record the regression line and S value below:

$y =$ _____ (slope) $x +$ _____ (vertical shift) S: _____

Inverting: From Linear *Models* to Logarithmic Ones

6) Use the coefficients of the *linear* model you just made to complete the *logarithmic* model below:

$logarithmic3(x) =$ _____ (log coefficient (a)) $log_{10}(x) +$ _____ (vertical shift (k))

```
fun logarithmic3(x): ( _____ * log(x)) + _____ end
```

7) Let's interpret this model:

A country where the _____ (x-axis) *is* _____ (base (b)) *times higher than another is also predicted to have a* _____ (y-axis) *that is* _____ (log coefficient (a)) _____ (y-axis units) *longer.*

8) Add the definition of `logarithmic3` to your starter file, and use it with `fit-model` to calculate the value of S: _____

9) Complete the table below, copying your S values from the previous models:

Linear	Quadratic	Exponential	Logarithmic

10) Compare the two smallest S values using percent change. *How much better* is the logarithmic model? _____

_____

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

Directions:

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____
function name — Domain — Range

_____
what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

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

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

end

Definition

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

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

_____
what the function does with those variable(s)

end

Directions:

Contract and Purpose Statement

Every contract has three parts…

_____ :: _____ -> _____
function name — Domain — Range

_____
what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

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

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

end

Definition

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

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

_____
what the function does with those variable(s)

end

The Animals Dataset

This is a printed version of the animals spreadsheet.
**The numbers on the left side are NOT part of the table!* They are provided to help you identify the index of each row.*

	name	species	sex	age	fixed	legs	pounds	weeks
0	Sasha	cat	female	1	false	4	6.5	3
1	Snuffles	rabbit	female	3	true	4	3.5	8
2	Mittens	cat	female	2	true	4	7.4	1
3	Sunflower	cat	female	5	true	4	8.1	6
4	Felix	cat	male	16	true	4	9.2	5
5	Sheba	cat	female	7	true	4	8.4	6
6	Billie	snail	hermaphrodite	0.5	false	0	0.1	3
7	Snowcone	cat	female	2	true	4	6.5	5
8	Wade	cat	male	1	false	4	3.2	1
9	Hercules	cat	male	3	false	4	13.4	2
10	Toggle	dog	female	3	true	4	48	1
11	Boo-boo	dog	male	11	true	4	123	24
12	Fritz	dog	male	4	true	4	92	3
13	Midnight	dog	female	5	false	4	112	4
14	Rex	dog	male	1	false	4	28.9	9
15	Gir	dog	male	8	false	4	88	5
16	Max	dog	male	3	false	4	52.8	8
17	Nori	dog	female	3	true	4	35.3	1
18	Mr. Peanutbutter	dog	male	10	false	4	161	6
19	Lucky	dog	male	3	true	3	45.4	9
20	Kujo	dog	male	8	false	4	172	30
21	Buddy	lizard	male	2	false	4	0.3	3
22	Gila	lizard	female	3	true	4	1.2	4
23	Bo	dog	male	8	true	4	76.1	10
24	Nibblet	rabbit	male	6	false	4	4.3	2
25	Snuggles	tarantula	female	2	false	8	0.1	1
26	Daisy	dog	female	5	true	4	68	8
27	Ada	dog	female	2	true	4	32	3
28	Miaulis	cat	male	7	false	4	8.8	4
29	Heathcliff	cat	male	1	true	4	2.1	2
30	Tinkles	cat	female	1	true	4	1.7	3
31	Maple	dog	female	3	true	4	51.6	4

Sentence Starters

Use these sentence starters to help describe patterns, make predictions, find comparisons, share discoveries, formulate hypotheses, and ask questions.

Patterns:

- I noticed a pattern when I looked at the data. The pattern is _____
- I see a pattern in the data collected so far. My graph shows _____

Predictions:

- Based on the patterns I see in the data collected so far, I predict that _____
- My prediction for _____ is _____

Comparisons:

- When I compared _____ and _____, I noticed that _____
- The similarities I see between _____ and _____ are _____
- The differences I see between _____ and _____ are _____

Surprises and Discoveries:

- I discovered that _____
- I was surprised by _____
- I noticed something unusual about _____

Hypotheses:

- A possible explanation for what the data showed is _____
- A factor that affected this data might have been _____
- I think this data was affected by _____

Questions:

- I wonder why _____
- I wonder how _____
- How are _____ affected by _____
- How will _____ change if _____

Contracts for Data Science

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 (above), Image (below))	-> Image
`above(circle(10, "solid", "black"), square(50, "solid", "red"))`		
`# bar-chart`	:: (Table (table-name), String (column))	-> Image
`bar-chart(animals-table, "species")`		
`# bar-chart-summarized`	:: (Table (table-name), String (labels), String (values))	-> Image
`bar-chart-summarized(count(animals-table, "species"), "value","count")`		
`# beside`	:: (Image (left), Image (right))	-> Image
`beside(circle(10, "solid", "black"), square(50, "solid", "red"))`		
`# box-plot`	:: (Table (table-name), String (column))	-> Image
`box-plot(animals-table, "weeks")`		
`# box-plot-scaled`	:: (Table (table-name), String (column), Number (low), Number (high))	-> Image
`box-plot-scaled(animals-table, "weeks", 1, 40)`		
`# build-column`	:: (Table (table-name), String (column), (Row -> Value) (builder-function))	-> Table
`build-column(animals-table, "kilos", kilograms)`		
`# circle`	:: (Number (radius), String, (fill-style), String (color))	-> Image
`circle(50, "solid", "purple")`		
`# count`	:: (Table (table-name), String (column))	-> Table
`count(animals-table, "species")`		
`# filter`	:: (Table (table-name), (Row -> Boolean) (tester-function))	-> Table
`filter(animals-table, is-dog)`		
`# first-n-rows`	:: (Table (table-name), Number (num-rows))	-> Table
`first-n-rows(animals-table, 15)`		

Name	Domain	Range
# fit-model	:: (Table , String , String , String , (Num -> Num)) (table-name, labels, xs, ys, model-function)	-> Image
fit-model(animals-table, "name", "pounds","weeks", f)		
# histogram	:: (Table , String , String , Number) (table-name, labels, values, bin-size)	-> Image
histogram(animals-table, "species", "weeks", 2)		
# image-bar-chart	:: (Table , String , (Row -> Image)) (table-name, values, draw-function)	-> Image
image-bar-chart(animals-table, "species", f)		
# image-histogram	:: (Table , String , Number ,(Row -> Image)) (table-name, values, bin-size, draw-function)	-> Image
image-histogram(animals-table, "pounds", 2, f)		
# image-pie-chart	:: (Table , String , (Row -> Image)) (table-name, values, draw-function)	-> Image
image-pie-chart(animals-table, "sex", f)		
# image-scatter-plot	:: (Table , String , String ,(Row -> Image)) (table-name, xs, ys, draw-function)	-> Image
image-scatter-plot(animals-table, "pounds","weeks", f)		
# line-graph	:: (Table , String , String , String) (table-name, labels, xs, ys)	-> Image
line-graph(animals-table, "name", "pounds","weeks")		
# log	:: (Number) (n)	-> Number
log(4)		
# log-base	:: (Number , Number) (base, n)	-> Number
log-base(2, 4)		
# lr-plot	:: (Table , String , String , String) (table-name, labels, xs, ys)	-> Image
lr-plot(animals-table, "name", "pounds","weeks")		
# mean	:: (Table , String) (table-name, column)	-> Number
mean(animals-table, "pounds")		
# median	:: (Table , String) (table-name, column)	-> Number
median(animals-table, "pounds")		
# modes	:: (Table , String) (table-name, column)	-> List
modes(animals-table, "pounds")		
# modified-box-plot	:: (Table , String) (table-name, column)	-> Image
modified-box-plot(animals-table, "pounds")		
# modified-box-plot-scaled	:: (Table , String , Number , Number) (table-name, column, low, high)	-> Image
modified-box-plot-scaled(animals-table, "weeks", 1, 40)		

Name	Domain	Range
# modified-vert-box-plot	:: (Table (table-name), String (column))	-> Image
`modified-vert-box-plot(animals-table, "pounds")`		
# modified-vert-box-plot-scaled	:: (Table (table-name), String (column), Number (low), Number (high))	-> Image
`modified-vert-box-plot-scaled(animals-table, "weeks", 1, 40)`		
# multi-bar-chart	:: (Table (table-name), String (group), String (subgroup))	-> Image
`multi-bar-chart(animals-table, "species", "sex")`		
# num-sqr	:: (Number)	-> Number
`num-sqr(4)`		
# overlay	:: (Image (top), Image (bottom))	-> Image
`overlay(circle(10, "solid", "black"), square(50, "solid", "red"))`		
# pie-chart	:: (Table (table-name), String (column))	-> Image
`pie-chart(animals-table, "species")`		
# pie-chart-summarized	:: (Table (table-name), String (labels), String (values))	-> Image
`pie-chart-summarized(count(animals-table, "species"), "value", "count")`		
# put-image	:: (Image (front), Number (x-coordinate), Number (y-coordinate), Image (behind))	-> Image
`put-image(circle(10, "solid", "black"), 10, 10, square(50, "solid", "red"))`		
# r-value	:: (Table (table-name), String (xs), String (ys))	-> Number
`r-value(animals-table, "name", "pounds","weeks")`		
# random-rows	:: (Table (table-name), Number (num-rows))	-> Table
`random-rows(animals-table, 10) # select 10 random rows from the table`		
# rectangle	:: (Number (width), Number (height), String (fill-style), String (color))	-> Image
`rectangle(100, 50, "outline", "green")`		
# regular-polygon	:: (Number (size), Number (vertices), String (fill-style), String (color))	-> Image
`regular-polygon(25,5, "solid", "purple")`		
# right-triangle	:: (Number (leg1), Number (leg2), String (fill-style), String (color))	-> Image
`right-triangle(50, 60, "outline", "blue")`		
# rotate	:: (Number (degrees), Image (img))	-> Image
`rotate(45, star(50, "solid", "dark-blue"))`		
# row-n	:: (Table (table-name), Number (index))	-> Row
`row-n(animals-table, 2)`		

Name		Domain		Range
# S	::	(Table , String , String , (Num -> Num)) table-name xs ys model-function	->	Number
`S(animals-table, "name", "pounds","weeks", f)`				
# 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)`				
# square	::	(Number , String , String) size fill-style color	->	Image
`square(50, "solid", "red")`				
# stacked-bar-chart	::	(Table , String , String) table-name group subgroup	->	Image
`stacked-bar-chart(animals-table, "species", "sex")`				
# star	::	(Number , String , String) radius fill-style color	->	Image
`star(50, "solid", "red")`				
# stdev	::	(Table , String) table-name column	->	Number
`stdev(animals-table, "pounds")`				
# string-contains	::	(String , String) haystack needle	->	Boolean
`string-contains("hotdog", "dog")`				
# sum	::	(Table , String) table-name column	->	Number
`sum(animals-table, "pounds")`				
# text	::	(String , Number , String) message size color	->	Image
`text("Zari", 85, "orange")`				
# vert-box-plot	::	(Table , String) table-name column	->	Image
`vert-box-plot(animals-table, "weeks")`				

::		->

::		->

::		->

9 781304 336422

These materials were developed partly through support of the National Science Foundation (awards 1042210, 1535276, 1648684, and 1738598) and are licensed under a Creative Commons 4.0 Unported License. Based on a work at www.BootstrapWorld.org. Permissions beyond the scope of this license may be available by contacting contact@BootstrapWorld.org.