Name: _____

Data Literacy

Fall 2025 Student Workbook - Pyret Edition

BOOTSTRAP

Equity • Scale • Rigor

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

9 781716 654817

Table of Contents

Pioneers in Computing and Mathematics

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

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

Notice and Wonder

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

What do you Notice?	What do you Wonder?

Windows and Mirrors

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

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

Reflection: Try Thinking About Ketchup

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

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

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

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

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

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

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

Perspective: A solution to tech's lingering diversity problem? Try thinking about ketchup

By Dexter Thomas • Published March 16, 2016 6:24 PM PT in the Los Angeles Times

Diversity is a hot, and controversial, topic in Silicon Valley. But why do so many people care about it?

At first glance, the answer may seem simple: Improving minorities' access to tech jobs is the right thing to do.

But when I moderated a panel Monday at SXSW on diversity in the tech industry, I was surprised none of the panelists talked much about what was "right."

Instead, they talked about what was right for business.

Sarah Wagener, vice president of talent acquisition and diversity at Pandora, agreed during the panel that pushing to hire more diverse candidates is the "right thing" to do.

"But," she said, "it's been the 'right thing to do' for a long time, and we're still having this conversation."
If you're trying to make the case at your company for diversifying your workforce, she said, your argument needs to be focused on "real business outcomes."

In other words, recruiting people from underrepresented backgrounds should be understood not as an obligation that could lower the bar and weigh your company down, but as an opportunity that could raise the bar, and lift your company above the competition.

Instantly, Wagener's statements reminded me of ketchup.

If you haven't heard it yet, the "ketchup question" is a thought experiment that's become something of a meme in some corners of the tech community thanks to a popular episode of the Reply All podcast. It starts as an innocent question:

Where do you keep your ketchup?

If you're like most people in the United States, odds are that you keep your ketchup in the refrigerator. But depending on where you grew up, you might keep it in the cupboard.

Imagine that you reach for the ketchup bottle and find it empty. You need a substitute sauce, and grab whatever is nearby. If that bottle is in the refrigerator, you may opt for mayo. But if it's in the cupboard, the seasoning closest at hand might be malt vinegar, or Tabasco, or salt and pepper.

Start-up culture is often centered around new ways of solving "problems" — ride-sharing apps such as Lyft and Uber solve the problem of getting around town without a car, for example. The "ketchup question" shows how a slight difference in perspective can lead a coworker toward a completely different solution that might never occur to you. That extra perspective could lead to a fresh new idea that could take your company to the top.

But without a diverse team? It's gonna be mayo every time.

What do we do about it?

Most people aren't chief executives of a major company, and may feel like they have no sway in the hiring process.
So I asked two of the panelists to give some suggestions that could be useful for employees of all levels, regardless of the industry in which they work.

Karla Monterroso, vice president of programs at Code 2040, an organization that works to place black and Latino students in engineering internships at tech companies, said that job listings could be an unexpected barrier to attracting diverse talent.
Using seemingly innocent words like "hacker" or "rockstar" in job listings could unintentionally give the impression to some women that the company would not be a hospitable place to work, said Monterroso. She recommended reading articles on the topic of bias and having informal conversations with coworkers.

More directly, she said, using these articles as "evidence" to suggest small changes in recruitment practices could be an easy first step in attracting new talent.

James Talbot, a software engineer at San Francisco web publishing startup Medium, was concerned with what happens after a new recruit is hired. He suggested using social media to follow people who have different perspectives than you, for 30 days. The key, he said, is to listen to what they have to say, simply exposing yourself to their conversations — not commenting or arguing with them.
This is important, he said, because even after a recruiter hires a person from an underrepresented community, adapting to the workplace environment can be another challenge. If people get into a job but have to deal with racist or sexist comments and insensitive treatment, they may simply leave – and take their unique perspectives and talent elsewhere.
People often say that the cause of the lack of diversity in many tech companies is the lack of an easy way to find available candidates.

"People always give excuses, saying the problem is the 'pipeline,'" Talbot said.

"But who wants to be on a pipeline into a sewer?"

Dexter Thomas is from San Bernardino and is a PhD candidate in East Asian studies at Cornell University. He has taught media studies and Japanese and is writing a book about Japanese hip-hop. Thomas began working in new media as a student director of programming at KUCR-FM (88.3), independently producing podcasts as well as music and news programs. He has written for several outlets internationally on topics as diverse as Internet and youth culture, social justice and video games. He left The Times in 2016.

(handout)

Case Study: Ethics, Privacy, and Bias

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

My Case Study is _____

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

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

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

Data Science used for this purpose is bad because…

Categorical and Quantitative Data in a Nutshell

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

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

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

Categorical or Quantitative?

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

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

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

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

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

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

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

_____

_____

_____

Questions and Column Descriptions

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

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

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

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

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

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

   _____.

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

   _____.

Opening Questions

Sports

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

Pop Culture

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

Politics

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

Education

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

What Questions Can You Answer with the Given Data?

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

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

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

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

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

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

(optional)

Introduction to Programming in a Nutshell

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, decimals *must* start with a zero. 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

Functions work much the way they do 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 in Pyret, for example `sqr`, `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? _____

Applying Functions

Open code.pyret.org (CPO) and click "Run". We will be working in the Interactions Area on the right.

Test out these two expressions and record what you learn below:

- `regular-polygon(40, 6, "solid", "green")`
- `regular-polygon(80, 5, "outline", "dark-green")`

1) You've seen data types like Numbers, Strings, and Booleans. What data type did the `regular-polygon` function produce? _____

2) How would you describe what a regular polygon is? _____

_____

3) The `regular-polygon` function takes in four pieces of information (called arguments). Record what you know about them below.

	Data Type	Information it Contains
Argument 1		
Argument 2		
Argument 3		
Argument 4		

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

```
# regular-polygon :: Number, Number, String, String -> Image
```

- Each Contract begins with the function name: *in this case* `regular-polygon`
- Lists the data types required to satisfy its Domain: *in this case* Number, Number, String, String
- And then declares the data type of the Range it will return: *in this case* Image

Contracts can also be written with more detail, by annotating the Domain with *variable names*:

```
# regular-polygon :: ( Number , Number , String , String ) -> Image
                     (  size  ,  number-of-sides , fill-style , color )
```

4) We know that a square is a regular polygon because _____

_____

5) What code would you write to make a big, blue square using the `regular-polygon` function?

_____ (_____ , _____ , _____ , _____)
function-name — size :: Number — number-of-sides :: Number — fill-style :: String — color :: String

6) Pyret also has a `square` function whose contract is:

```
# square :: ( Number , String , String ) -> Image
            (  size  , fill-style , color )
```

What code would you write to make a big blue square using the `square` function?

_____ (_____ , _____ , _____)
function-name — size :: Number — fill-style :: String — color :: String

7) Why does `square` need fewer arguments to make a square than `regular-polygon`? _____

_____

★ Where else have you heard the word ***contract*** used before?

Practicing Contracts: Domain & Range

Note: The contracts on this page are not defined in Pyret and cannot be tested in the editor.

is-beach-weather

Consider the following Contract:

```
# is-beach-weather :: Number, String -> Boolean
```

1) What is the **Name** of this function? _____

2) How many arguments are in this function's **Domain**? _____

3) What is the **Type** of this function's **first argument**? _____

4) What is the **Type** of this function's **second argument**? _____

5) What is the **Range** of this function? _____

6) Circle the expression below that shows the correct application of this function, based on its Contract.

A. `is-beach-weather(70, 90)`
B. `is-beach-weather(80, 100, "cloudy")`
C. `is-beach-weather("sunny", 90)`
D. `is-beach-weather(90, "stormy weather")`

cylinder

Consider the following Contract:

```
# cylinder :: Number, Number, String -> Image
```

7) What is the **Name** of this function? _____

8) How many arguments are in this function's **Domain**? _____

9) What is the **Type** of this function's **first argument**? _____

10) What is the **Type** of this function's **second argument**? _____

11) What is the **Type** of this function's **third argument**? _____

12) What is the **Range** of this function? _____

13) Circle the expression below that shows the correct application of this function, based on its Contract.

A. `cylinder("red", 10, 60)`
B. `cylinder(30, "green")`
C. `cylinder(10, 25, "blue")`
D. `cylinder(14, "orange", 25)`

Matching Expressions and Contracts

Match the Contract (left) with the expression that uses it correctly (right).
Note: The contracts on this page are not defined in Pyret and cannot be tested in the editor.

Contract		Expression
`# make-id :: String, Number -> Image`	1	A `make-id("Savannah", "Lopez", 32)`
`# make-id :: String, Number, String -> Image`	2	B `make-id("Pilar", 17)`
`# make-id :: String -> Image`	3	C `make-id("Akemi", 39, "red")`
`# make-id :: String, String -> Image`	4	D `make-id("Raïssa", "McCracken")`
`# make-id :: String, String, Number -> Image`	5	E `make-id("von Einsiedel")`

Contract		Expression
`# is-capital :: String, String -> Boolean`	6	A `show-pop("Juneau", "AK", 31848)`
`# is-capital :: String, String, String -> Boolean`	7	B `show-pop("San Juan", 395426)`
`# show-pop :: String, Number -> Image`	8	C `is-capital("Accra", "Ghana")`
`# show-pop :: String, String, Number -> Image`	9	D `show-pop(3751351, "Oklahoma")`
`# show-pop :: Number, String -> Number`	10	E `is-capital("Albany", "NY", "USA")`

Contracts for Image-Producing Functions

Log into [code.pyret.org (CPO)] and click "Run". Experiment with each of the functions listed below in the interactions area. Try to find an expression that produces an image. Record the contract and example code for each function you are able to use!

Name		Domain		Range
`# triangle`	`::`	`Number, String, String`	`->`	`Image`
`triangle(80, "solid", "darkgreen")`				
`# star`	`::`		`->`	
`# circle`	`::`		`->`	
`# rectangle`	`::`		`->`	
`# text`	`::`		`->`	
`# square`	`::`		`->`	
`# rhombus`	`::`		`->`	
`# ellipse`	`::`		`->`	
`# regular-polygon`	`::`		`->`	
`# right-triangle`	`::`		`->`	
`# isosceles-triangle`	`::`		`->`	
`# radial-star`	`::`		`->`	
`# star-polygon`	`::`		`->`	
`# triangle-sas`	`::`		`->`	
`# triangle-asa`	`::`		`->`	

Catching Bugs when Making Triangles

Learning about a Function through Error Messages

1) Type `triangle` into the Interactions Area of code.pyret.org (CPO) and hit "Enter". What do you learn? _____

2) We know that all functions will need an open parenthesis and at least one input! Type `triangle(80)` in the Interactions Area and hit Enter/return. Read the error message. What hint does it give us about how to use this function?

_____

3) Using the hint from the error message, experiment until you can make a triangle. What is the contract for `triangle`?

_____

4) Read the explanation below. Then explain the difference in your own words.

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)

The difference between **syntax errors** and **contract errors** is: _____

_____

Finding Mistakes with Error Messages

The following lines of code are all BUGGY! Read the code and the error messages below. See if you can find the mistake WITHOUT typing it into Pyret.

5) `triangle(20, "solid" "red")`

```
Pyret didn't understand your program around
triangle(20, "solid" "red")
```

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

6) `triangle(20, "solid")`

```
This application expression errored:
triangle( 20 , "solid" )
2 arguments were passed to the operator. The operator evaluated to a function accepting 3
parameters. An application expression expects the number of parameters and arguments to be the
same.
```

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

7) `triangle(20, 10, "solid", "red")`

```
This application expression errored:
triangle(20, 10, "solid", "red" )
4 arguments were passed to the operator. The operator evaluated to a function accepting 3
parameters. An application expression expects the number of parameters and arguments to be the
same.
```

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

8) `triangle (20, "solid", "red")`

```
Pyret thinks this code is probably a function call:
triangle (20, "solid", "red")
Function calls must not have space between the function expression and the arguments.
```

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

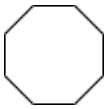
Using Contracts

For questions 1,2,4,5,8 & 9, use the contracts provided to find expressions that will generate images similar to the ones pictured.
Test your code in [code.pyret.org (CPO)] before recording it.

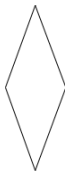
```
# ellipse :: ( Number , Number , String , String ) -> Image
             (width, height, fill-style, color)
```

1)		
2)		
3)	Write an expression using `ellipse` to produce a circle.	

```
# regular-polygon :: ( Number , Number , String , String ) -> Image
                     (side-length, number-of-sides, fill-style, color)
```

4)		
5)		
6)	Use `regular-polygon` to write an expression for a square!	
7)	How would you describe a **regular polygon** to a friend?	

```
# rhombus :: ( Number , Number , String , String ) -> Image
             (size, top-angle, fill-style, color)
```

8)		
9)		
10)	Write an expression to generate a `rhombus` that is a square!	

Triangle Contracts

Respond to the questions. Go to code.pyret.org (CPO) to test your code.

1) What kind of triangle does the `triangle` function produce? _____
There are lots of other kinds of triangles! And Pyret has lots of other functions that make triangles!

```
# triangle :: ( Number , String , String ) -> Image
#             (  size  , fill-style , color )
# right-triangle :: ( Number , Number , String , String ) -> Image
#                   (  base  , height , fill-style , color )
# isosceles-triangle :: ( Number , Number , String , String ) -> Image
#                        (  leg  , angle , fill-style , color )
```

2) Why do you think `triangle` only needs one number, while `right-triangle` and `isosceles-triangle` need two numbers?

_____

_____

3) Write `right-triangle` expressions for the images below using `100` as one argument for each.

_____

_____

4) Write `isosceles-triangle` expressions for the images below using `100` as one argument for each.

_____

_____

5) Write 2 expressions that would build **right-isosceles** triangles. Use `right-triangle` for one expression and `isosceles-triangle` for the other expression.

_____

_____

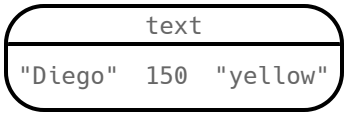
6) Which do you like better? Why? _____

Composing with Circles of Evaluation

Notice and Wonder

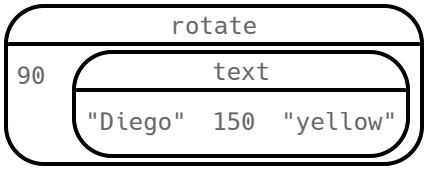
Suppose we want to see the `text` *"Diego" written vertically in yellow letters of size 150. Let's use Circles of Evaluation to look at the structure:*

We can start by generating the Diego image.



→

And then use the `rotate` function to rotate it 90 degrees.




```
text("Diego", 150, "yellow")
```

```
rotate(90, text("Diego", 150, "yellow"))
```

1) What do you Notice? _____

_____

2) What do you Wonder? _____

_____

Let's Rotate an Image of Your Name!

Suppose you wanted the computer to show your name in your favorite color and rotate it so that it's diagonal…

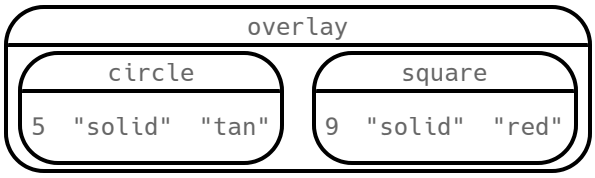
Write your name (any size), in your favorite color	**`rotate` the image so that it's diagonal**
3) Draw the circle of evaluation:	4) Draw the circle of evaluation:
5) Convert the Circle of Evaluation to code:	6) Convert the Circle of Evaluation to code:

Circle of Evaluation to Code (Scaffolded)

Complete the Code by Filling in the Blanks!

Finish the Code by filling in the blanks.

1)

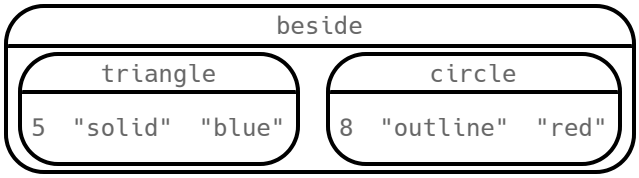



```
overlay(circle(____, "solid", _____________________), _____________________(9, _____________________, "red"))
```

Complete the Code by adding Parentheses

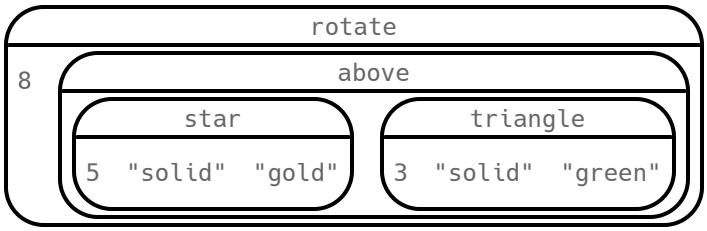
For each Circle of Evaluation, finish the Code by adding parentheses and commas.

2)



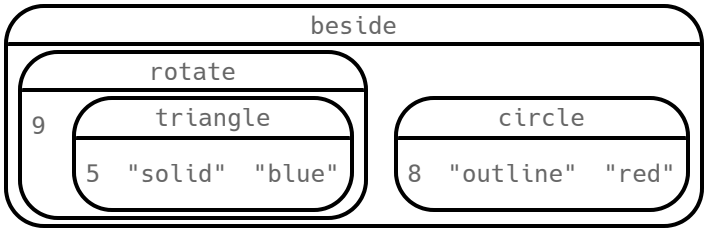

```
beside    triangle    5    "solid"    "blue"    circle    8    "outline" "red"
```

3)




```
rotate    8    above    star    5    "solid"    "gold"    triangle    3    "solid"    "green"
```

4)




```
beside    rotate    9    triangle    5    "solid"    "blue"    circle    8    "outline"    "red"
```

Frayer Model: Domain and Range

My Definition

Facts and Characteristics

Domain

Examples

Non-Examples

My Definition

Facts and Characteristics

Range

Examples

Non-Examples

(optional)

Frayer Model: Function and Variable

My Definition

Facts and Characteristics

Function

Examples

Non-Examples

My Definition

Facts and Characteristics

Variable

Examples

Non-Examples

(optional)

Radial Star

```
# radial-star :: ( Number , Number , Number , String , String ) -> Image
                   points   outer-radius  inner-radius  fill-style  color
```

Using the Contract above, match the images on the left to the expressions on the right. You can test the code at code.pyret.org (CPO).

1

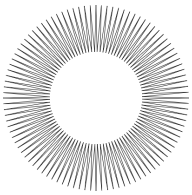
A

```
radial-star(5, 200, 50, "solid", "black")
```

2

B

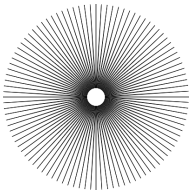
```
radial-star(7, 200, 100, "solid", "black")
```



3

C

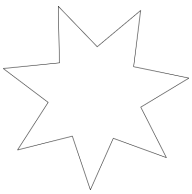
```
radial-star(7, 200, 100, "outline", "black")
```



4

D

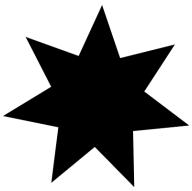
```
radial-star(10, 200, 150, "solid", "black")
```



5

E

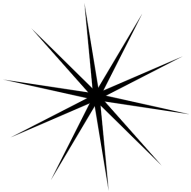
```
radial-star(10, 200, 20, "solid", "black")
```



6

F

```
radial-star(100, 200, 20, "outline", "black")
```



7

G

```
radial-star(100, 200, 100, "outline", "black")
```

(optional)

Triangle Contracts (SAS & ASA)

Type each expression (left) below into the code.pyret.org (CPO) and match it to the image it creates (right).

Expression			Image
`triangle-sas(120, 45, 70, "solid", "black")`	1	A	
`triangle-sas(120, 90, 70, "solid", "black")`	2	B	
`triangle-sas(120, 135, 70, "solid", "black")`	3	C	
`triangle-sas(70, 135, 120, "solid", "black")`	4	D	

Contracts

Think about how you would describe each `triangle-sas` argument to someone who'd never used the function before.

5) Annotate the Contract below using descriptive variable names.

```
triangle-sas :: ( ___Number___ , ___Number___ , ___Number___ , ___String___ , ___String___ ) -> Image
```

If you have a printed workbook, add examples of each of the triangle functions we've explored to your contracts pages.

★ If you have time, experiment with the `triangle-asa` function.

```
# triangle-asa :: ( ___Number___ , ___Number___ , ___Number___ , ___String___ , ___String___ ) -> Image
                    top-left-angle   left-side     bottom-angle    fill-style      color
```

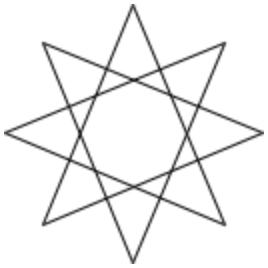
★ Why did these two functions need to take in one more Number than `right-triangle` did?

(optional)

Star Polygon

```
# star-polygon :: (  Number  ,  Number  ,  Number  ,  String  ,  String  ) -> Image
                   side-length  points-on-polygon  points-to-skip-for-star  fill-style  color
```

1. Using the Contract above, write expressions to create images like those pictured below.
2. Go to code.pyret.org (CPO) to test your code.
3. Then write expressions to generate two more star polygons of your choosing.
   Sketch them and record your working code.

1		
2		
3		
4		

(optional)

Sorting and Summarizing Tables

Open the Animals Starter File and click "Run". In the Interactions Area (right), type `animals-table`*. Hit "Enter" to see the default view of the table.*

Ordering a Table with `sort`

1) Mabel Lee wants to sort this table by `age` (youngest-to-oldest). Juan Carlos wants to sort the table by `pounds` (heaviest-to-lightest). What are some other ways we could sort the table?

a. _____

b. _____

Pyret has a function called `sort` that will produce sorted Tables!

2) Test out `sort(animals-table, "age", true)` in the Interactions Area. Try using `false` instead of `true`.
Complete the sentences below by **circling** the behavior you observed for each Boolean.

(a) `true` sorts the table…	in **ascending order** *(from least to greatest)*	in **descending order** *(from largest to smallest)*
(b) `false` sorts the table…	in **ascending order** *(from least to greatest)*	in **descending order** *(from largest to smallest)*

3) The Domain of `sort` has three inputs. One of them is the table itself. Can you identify the data types of the other two?

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

4) What code will sort the animals by alphabetical order of their *names*? _____

5) Did you use `true` or `false`? Explain why. _____

Summarizing a Column with `count`

Let's explore another table function, beginning with its contract:

```
# count :: Table, String -> Table
```

6) What do you expect the code `count(animals-table, "legs")` to produce? _____

_____

Type the code into the Interactions Area and click "Enter" to test it out.

7) How many animals had 4 legs? _____

8) Think of another question you might be able to answer by making a different table using the `count` function.

_____

9) Fill in the blanks with the code to make the table: _____ (_____, _____)
function-name — table-name :: Table — column-name :: String

10) Try using the `count` function to summarize the `pounds` column. Is the resulting summary useful? Why or why not? _____

_____

11) Tables that summarize data with a count are commonly used in the real world. Write an example of where you've seen them before:

_____

12) Newspapers often incorporate data into their reporting. How else might they display this information, besides using a table?

_____

Functions for Tables (continued)

Grabbing a Single Row

In addition to Numbers, Strings, Booleans, Images and Tables, Pyret has a *data type* for an individual `Row`.

Open the Animals Starter File and click "Run". In the Interactions Area (right), type `animals-table`. Hit "Enter" to see the default view of the table. Then type `row-n(animals-table, 2)` and compare the result to the table.

1) Write the code that generates the first row of the table. _____

2) Explain what the second input to `row-n` means, in as much detail as possible. _____

_____

Grabbing Multiple Rows

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

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

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

Defining Values

Pyret lets us *define* values that we want to use later. We can define any kind of values we like!

6) If we tell Pyret that `x = 4 * 2`, what would you expect to get back when you type `x + 1`? _____
Test it out by typing `x = 4 * 2` into the Interactions Area, hitting "Enter" and then typing `x + 1`.

7) Try typing `gt = triangle(50, "solid", "green")` and hitting "Enter".

What happens? _____

Now type `gt`. What do you get back? _____

8) Explain what is happening on Line 14 of the Animals Starter File. _____

_____

9) On line 16 of the Definitions Area, add a new definition called `my-pet`, which is defined to be your favorite animal.

code: _____

10) Add a new line at the bottom of the Definitions Area, define `first-3` to be a subset of the first 3 rows of the animals-table.

code: _____

★ 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!

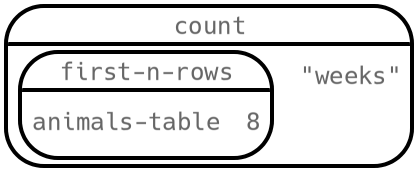
★★ See if you can figure out how to compose the code that would generate a table of the 10 oldest animals!

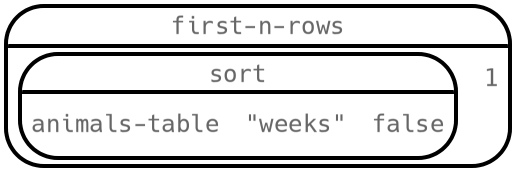
_____ (_____ , _____)
function-name Table Number

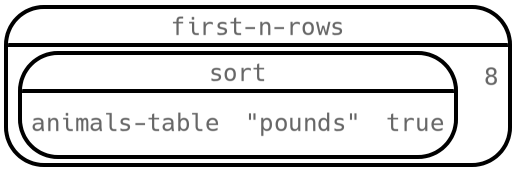
Matching Descriptions to Circles of Evaluation: sort, count, first-n-rows

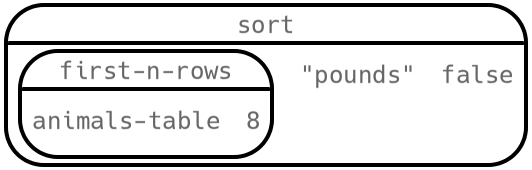
Match each prompt on the left to the Circle of Evaluation used to answer it.

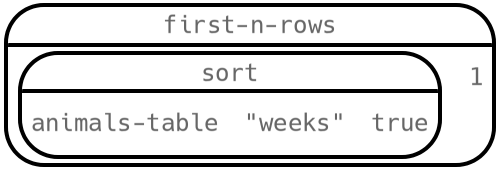
We want to see the 8 lightest animals.	1	A
We want to see the animal who was adopted the quickest.	2	B
We want to take the first 8 animals from the table and order them from heaviest to lightest.	3	C
We want to know the count by weeks to adoption of the first 8 animals.	4	D
		E











★ Translate each Circle of Evaluation into code and test it out in the Animals Starter File to confirm it does what you'd expect it to.

Hint: The Code for A is `count(first-n-rows(animals-table, 8), "weeks")`

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. Read the error message. What hint does it give us about how to use this function?

_____

3) Read the explanations below. Then explain the difference in your own words.

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)

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`

```
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 _____ (contract / syntax) error. The problem is that _____

5) `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 _____ (contract / syntax) error. The problem is that _____

6) `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 _____ (contract / syntax) error. The problem is that _____

7) `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 _____ (contract / syntax) error. The problem is that _____

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 _____ (contract / syntax) error. The problem is that _____

Exploring Data Visualizations

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…

Dot Plots: `# dot-plot :: Table, String, String -> Image`

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

Sketch a dot plot below.

Dot 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…

Composing Functions: Match Descriptions to Circles of Evaluation

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

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

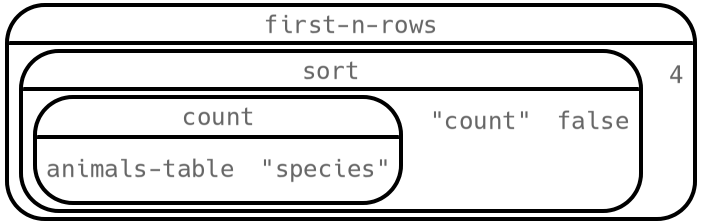
1

A

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

2

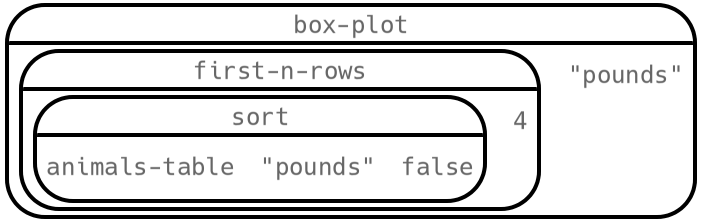
B



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

3

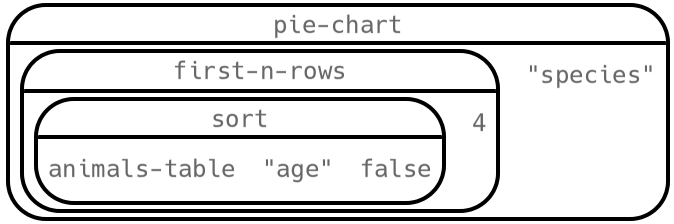
C



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

4

D



Circles of Evaluation: Composing Functions to Make Visualizations

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
# dot-plot :: Table, String -> 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 dot plot of `weeks` to adoption.

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

_____

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

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

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

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

Exploring Data Visualizations (2)

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

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

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

Sketch a line graph below.

Line Graphs summarize 2 columns of _____ data.
categorical/quantitative

This kind of display tells us…

_____

_____

_____

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

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

Sketch a scatter plot below.

Scatter Plots summarize 2 columns of _____ data.
categorical/quantitative

This kind of display tells us…

_____

_____

_____

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

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

Sketch an Linear Regression Plot below.

LR Plots summarize 2 columns of _____ data.
categorical/quantitative

This kind of display tells us…

_____

_____

_____

(optional)

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

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

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

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

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

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

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

(optional)

Displaying Categorical Data in a Nutshell

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

When it comes to displaying **Categorical Data**, there are two visualizations 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).

Frequency Tables, Bar Charts and Pie Charts

Open the Expanded Animals Starter File and click "Run".

Part 1 - Visualizations for Categorical Data

Test the following expressions in the Interactions Area:

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

1) How are they similar? _____

2) Which do you like better: the bar chart or the frequency 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 visualizations. 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!

Part 2 - 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?

_____

_____

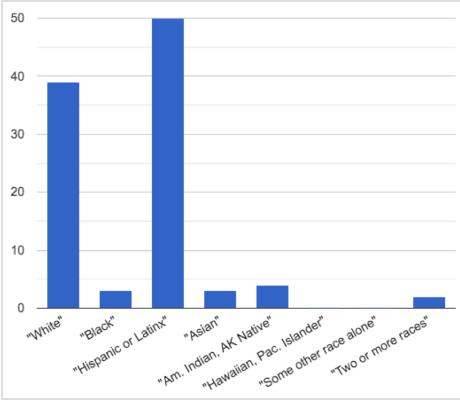
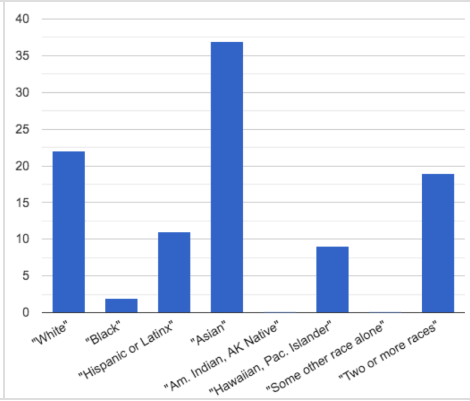
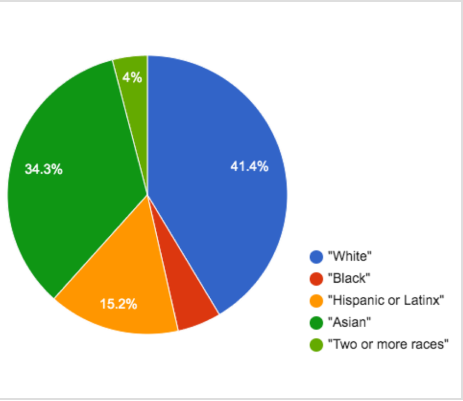
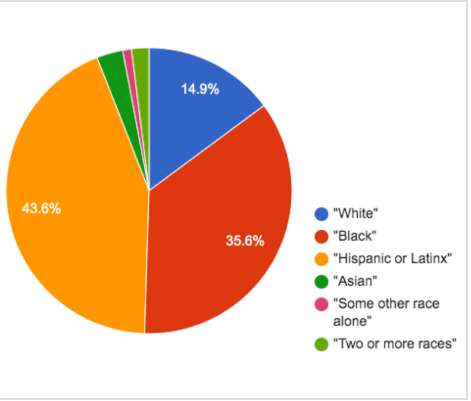
10) Look at the list of columns in the Definitions Area. For which columns do you expect pie charts to be most useful?

_____

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

Bar & Pie Chart - Notice and Wonder

What do you Notice and Wonder about the visualizations below?

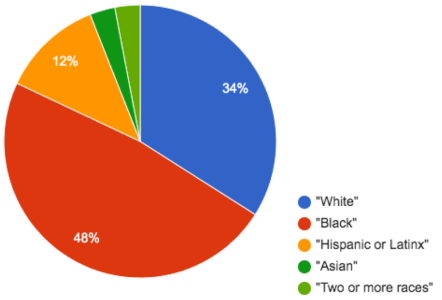
Albuquerque Public Schools, NM	Hawaii DOE, HI	San Francisco Unified School District, CA	Hartford School District, CT
			

What do you Notice?	What do you Wonder?

Matching Bar and Pie Charts

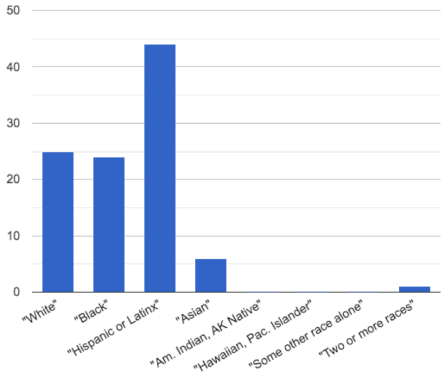
Match each bar chart below to the pie chart that visualizes the racial demographic data from the same school district.

Cleveland Municipal School District

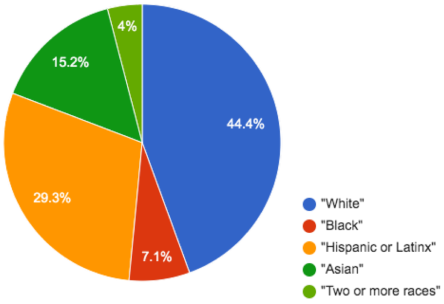


1

A

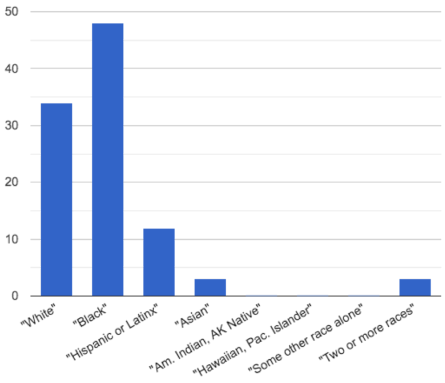


San Diego City Unified School District

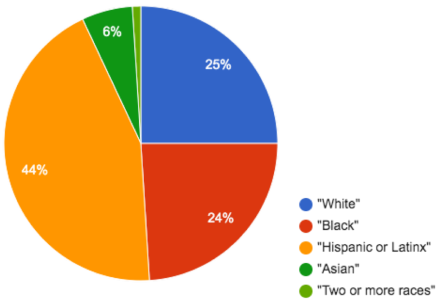


2

B

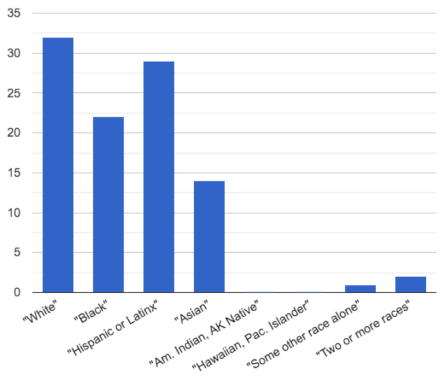


Houston Independent School District

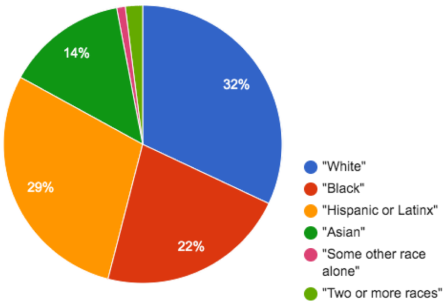


3

C

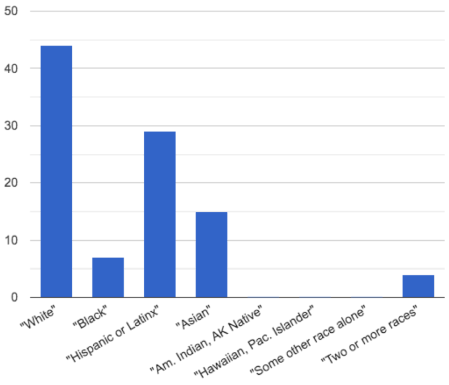


New York City Dept of Education



4

D



Introducing Visualizations 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. _____

_____

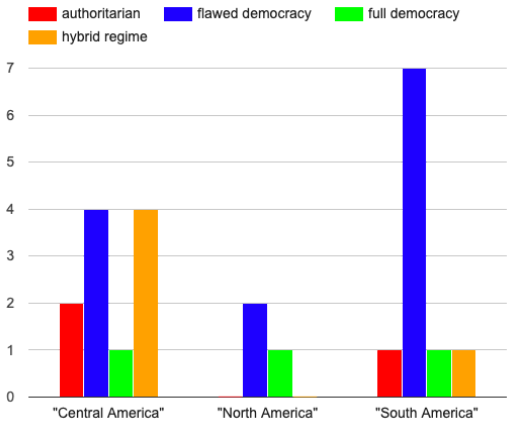
What did you learn from making this display? _____

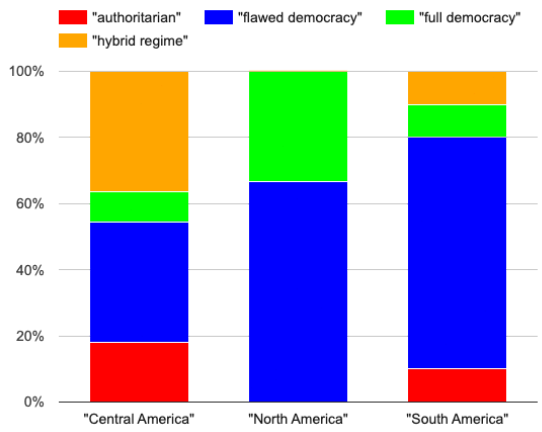
Multi Bar & Stacked Bar Charts - Notice and Wonder

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

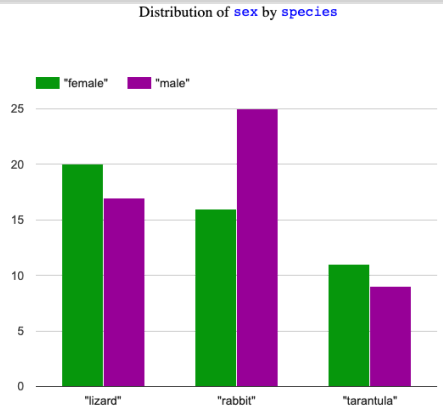
The visualizations on the right are called **stacked bar charts**.

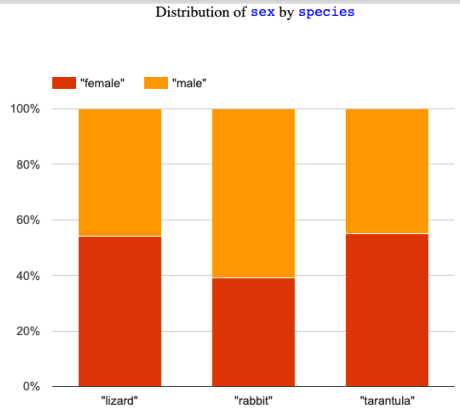
Political Structure of Countries in the Americas





Distribution of sexes across Cats, Dogs and Rabbits





What do you Notice?	What do you Wonder?

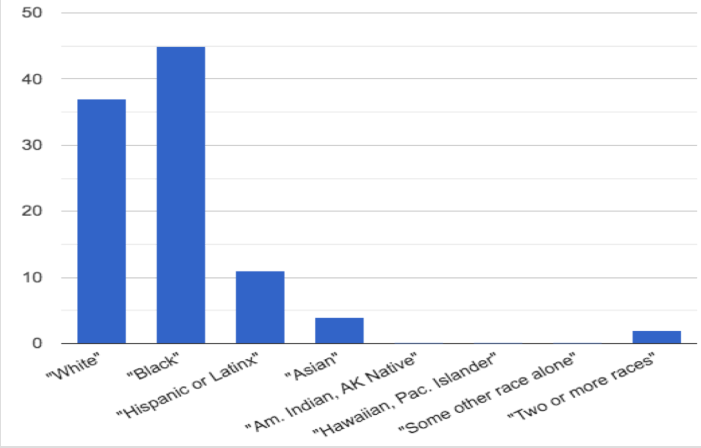
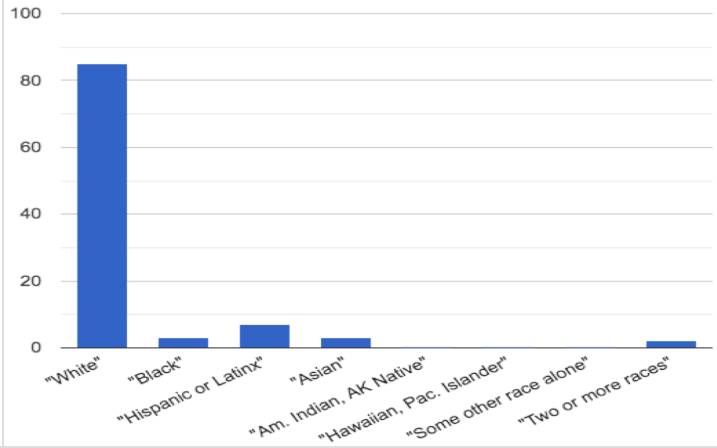
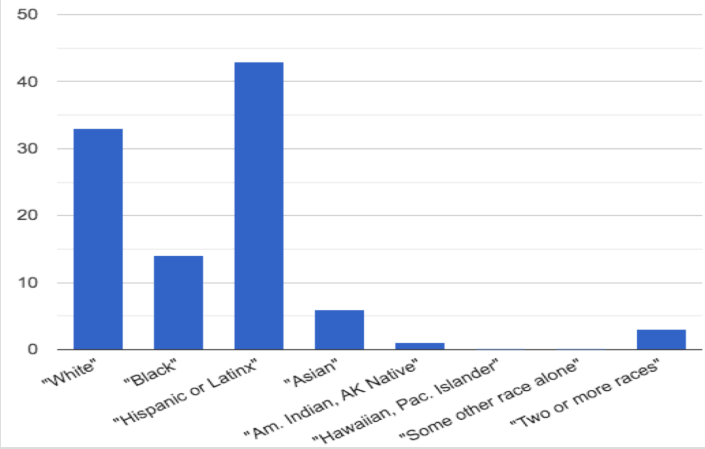
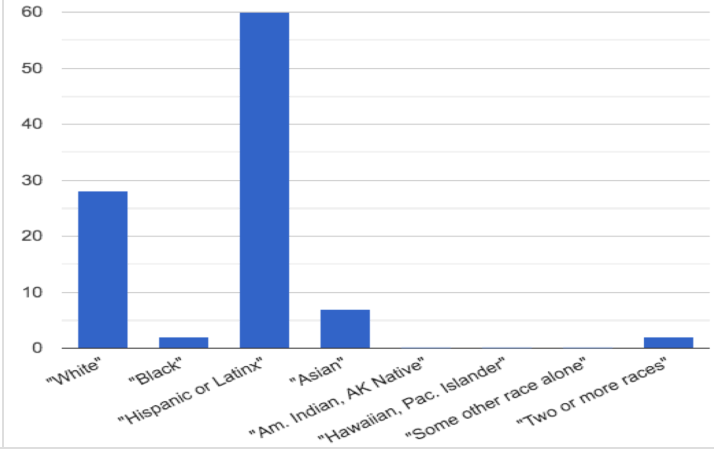
1) Is it possible that the same data was used for the multi bar charts as for the stacked bar charts? How do you know?

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

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

Bar Chart - Notice and Wonder

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

Washington, D.C. Public Schools	West Warwick Schools, RI
	
Providence Schools, RI	**Oxnard Union HS District, CA**
	

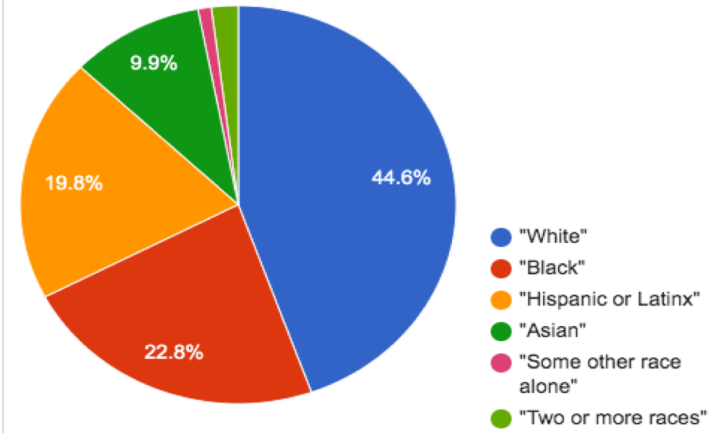
What do you Notice?	What do you Wonder?

(optional)

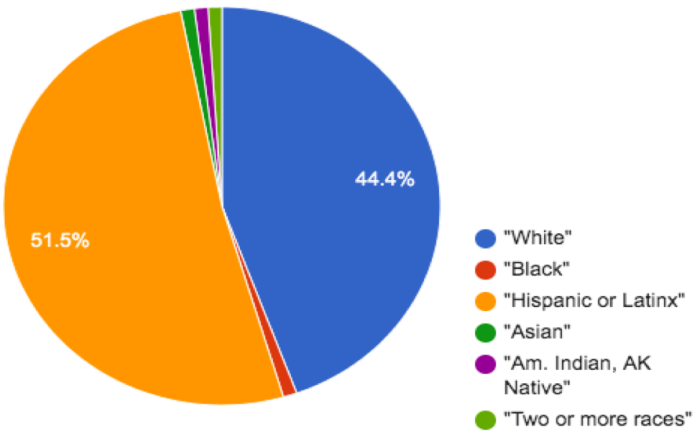
Pie Chart - Notice and Wonder

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

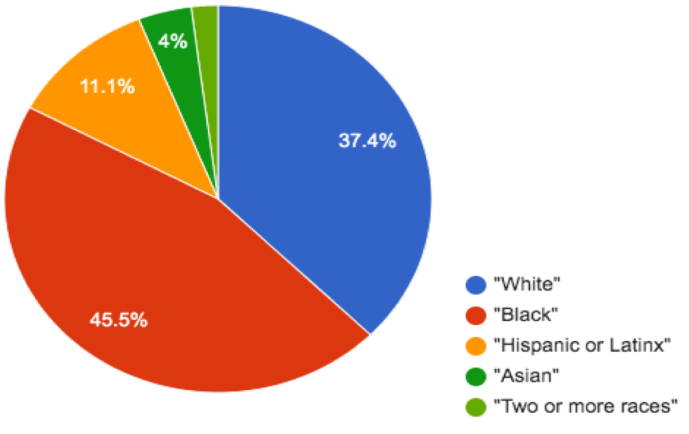
Boston School District, MA



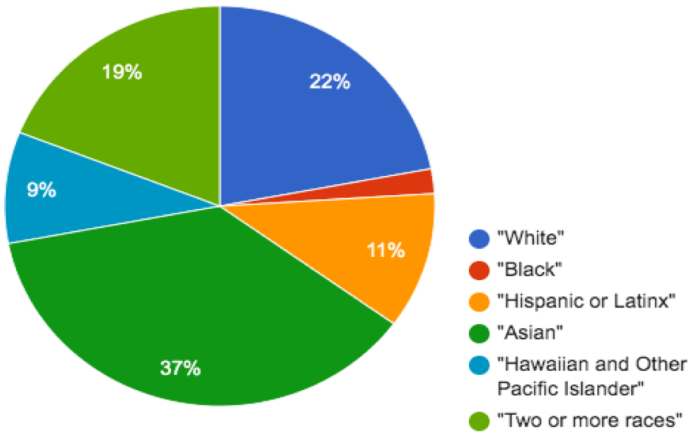
Santa Fe Public Schools, NM



Washington, D.C. Public Schools



Hawaii DOE, HI

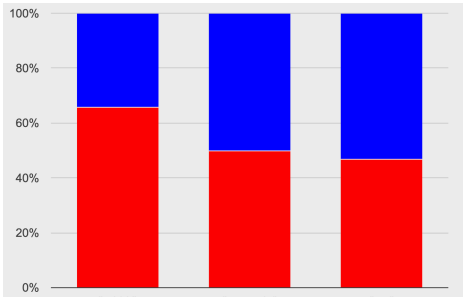


What do you Notice?	What do you Wonder?

(optional)

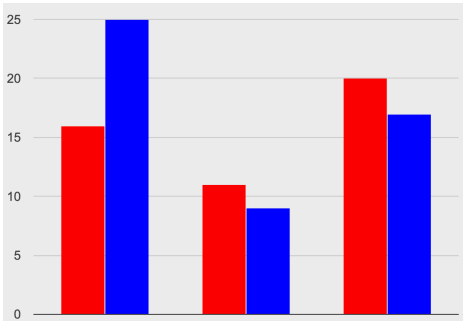
Matching Stacked and Multi Bar Charts

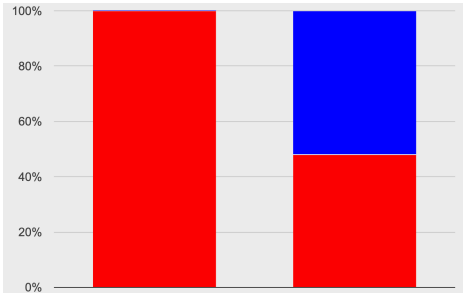
Match each stacked bar chart below to the multi bar chart that visualizes the same information.



1

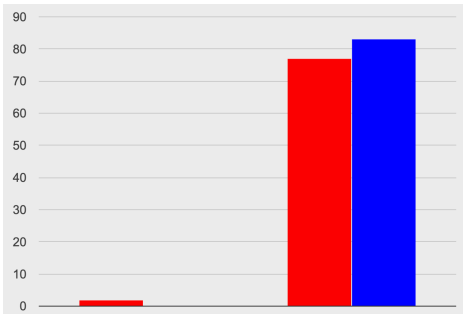
A

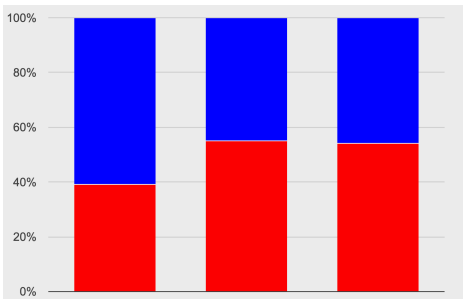




2

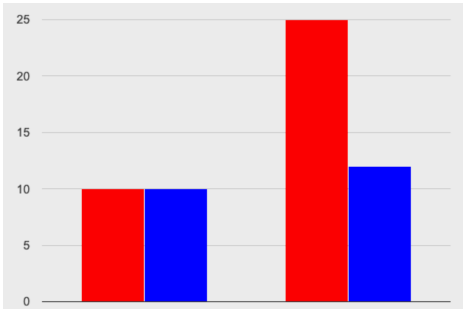
B

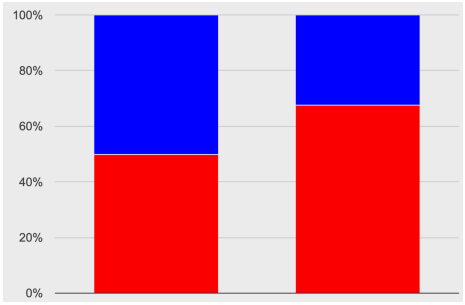




3

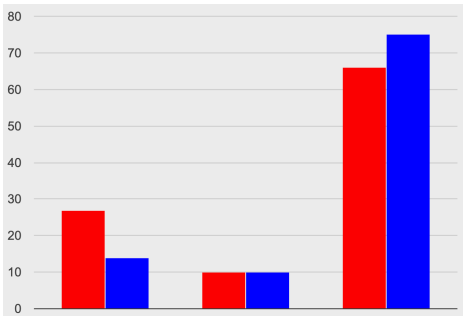
C





4

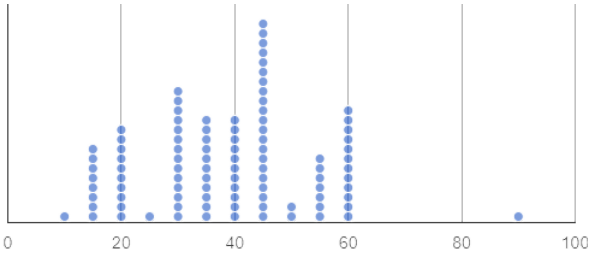
D



(optional)

Dot Plots: Distribution, Typicality, Variability in a Nutshell

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



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

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

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

Typicality of Data

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

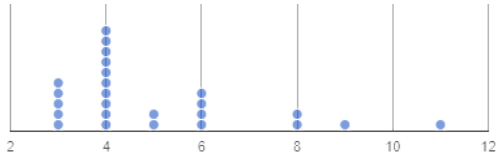
Variability of Data

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

Interpreting Dot Plots

Reading a Dot Plot (Group A)

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



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

2) What is the most common name length? _____

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

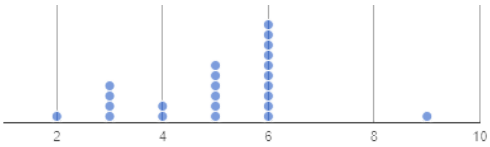
Interpreting Peaks, Clusters, Gaps, and Outliers

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

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

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

Reading a Dot Plot (Group B)



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

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

_____

_____

Typicality of Name Length Data

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

_____

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

_____

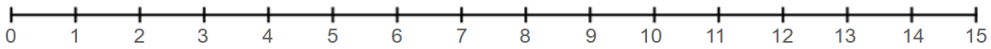
10) Would 6 letters be a good description of the typical number of letters in students' names for Group B? _____ Explain. _____

_____

Our Class' Name Length Data

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

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



Reading a Dot Plot

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

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

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

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

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

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

Typicality of Name Length Data

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

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

Compare

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

Two Ways of Thinking about Variability

Variability of Categorical Data

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

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

_____

_____

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

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

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

_____

_____

Variability of Quantitative Data

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

Do you agree or disagree? Explain your reasoning. _____

_____

_____

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

_____

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

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

_____

_____

Designing Datasets with High and Low Variability

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

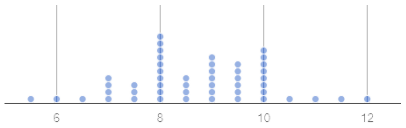
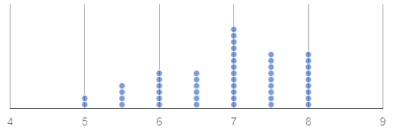
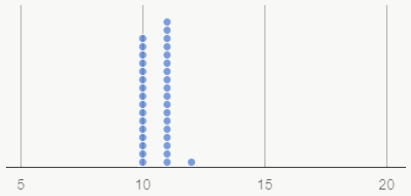
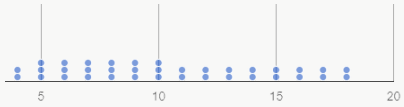
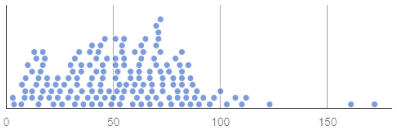
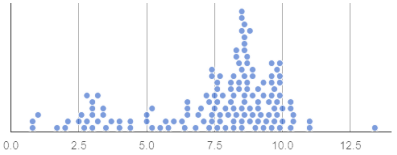
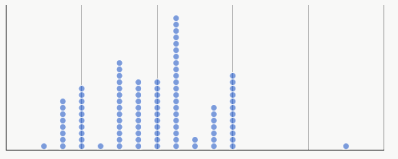
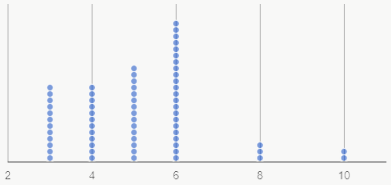
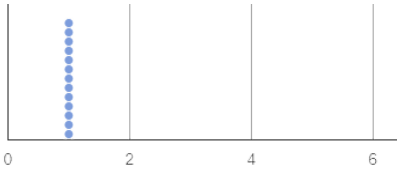
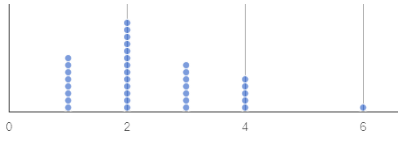
Dataset with Low Variability	Dataset with High Variability

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

Dataset with Low Variability	Dataset with High Variability

Variability of Dot Plots

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

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

Variability of Animals' Weights

Make Your Predictions

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

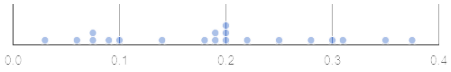
1) Imagine a *typical* animal from each of these four species. Rank the animals from lightest (1) to heaviest (4).

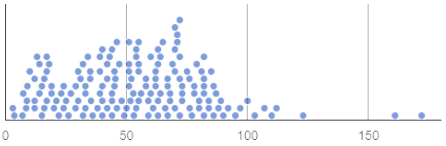
dog: _____ rabbit: _____ cat: _____ tarantula: _____

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

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

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





species: _____ species: _____ species: _____

5) Explain how you made your decisions. _____

_____

_____

_____

Test Your Predictions Using Pyret

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

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

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

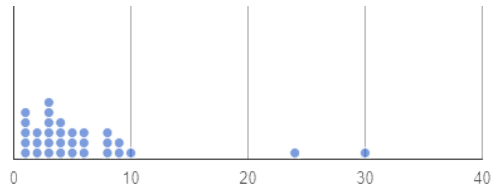
_____

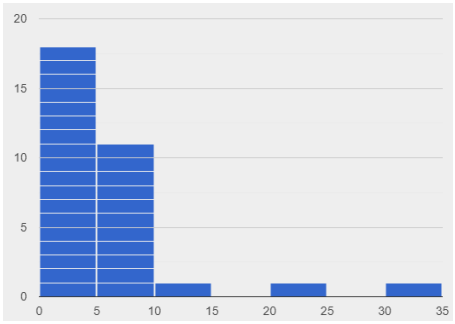
_____

Comparing Dot Plots and Histograms

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

Notice and Wonder





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

_____

_____

_____

Dot Plots versus Histograms

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

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

Reflect

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

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

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

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

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

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

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

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

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

_____

_____

_____

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

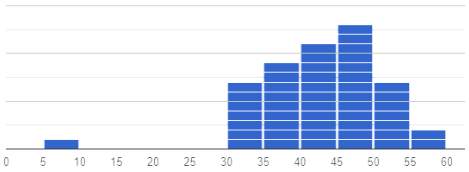
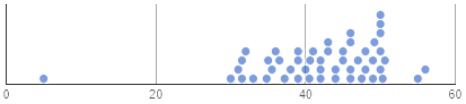
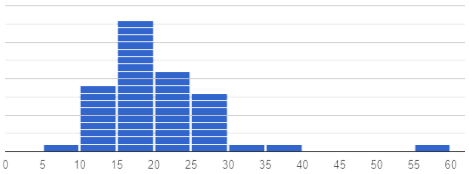
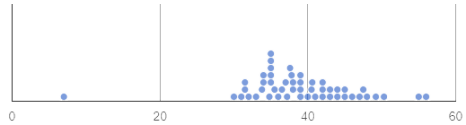
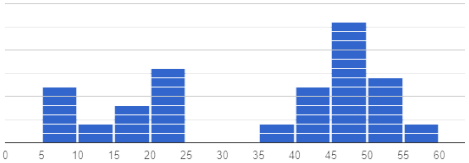
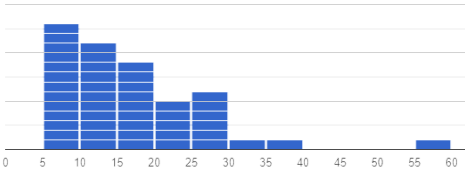
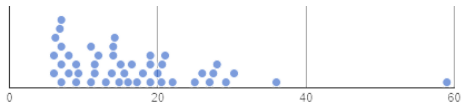
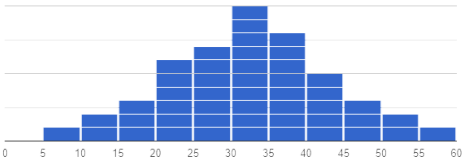
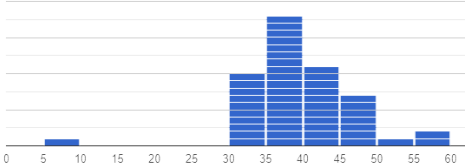
_____

_____

_____

Matching Dot Plots and Histograms

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

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

Making Histograms

By Hand

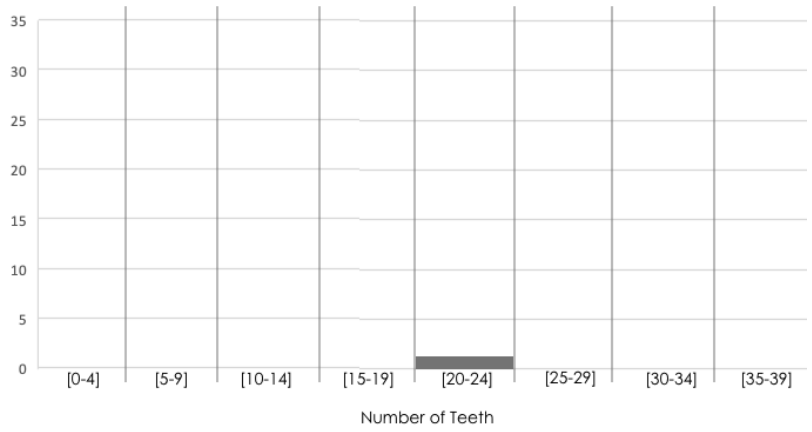
Suppose we have a dataset for a group of 50 adults, showing the number of teeth each person has…

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

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

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

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



In Pyret

Open the Tooth Data Starter File. Make a copy, and click "Run".

3) Type `tooth-table` in the Interactions window. Press enter. What do you see? _____

4) Type `count(tooth-table, "num-teeth")` in the Interactions window and press enter. How is the frequency table created in Pyret different from the one that you created, above? _____

_____

5) What bin sized was used for the Tooth Data frequency table and the histogram above? _____

6) Build `tooth-table`. Does this data appear to be the same or different from the tooth data that appeared in the first section? _____

7) Use the contract below to build a histogram in Pyret of the distribution of teeth.

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

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

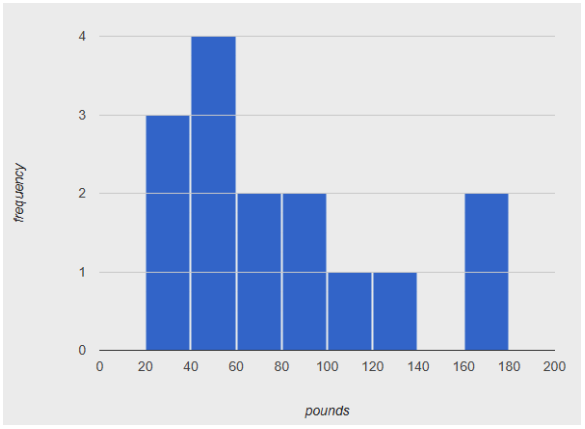
_____

_____

Reading Histograms

Small Local Animal Shelter

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



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

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

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

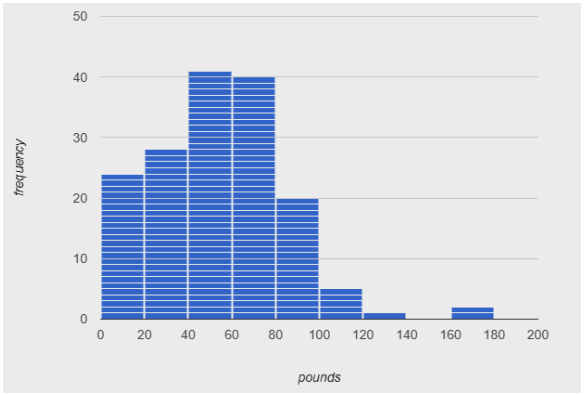
4) True or False: The dogs weigh between 20 and 180 pounds.

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

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

Larger Animal Shelter

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



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

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

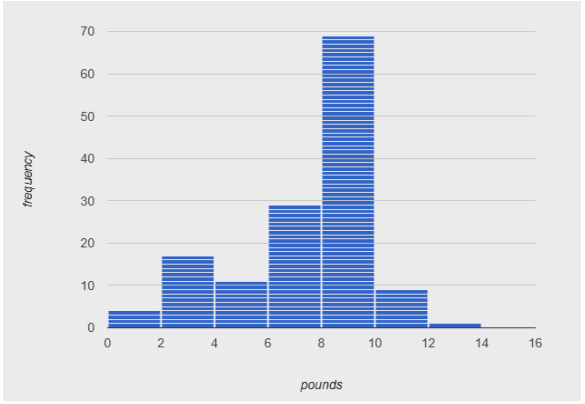
9) True or False: The lightest dog weighs zero pounds.

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

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

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

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



13) _____

_____

_____

14) _____

_____

_____

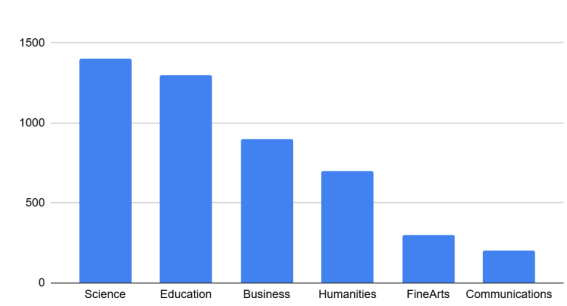
15) _____

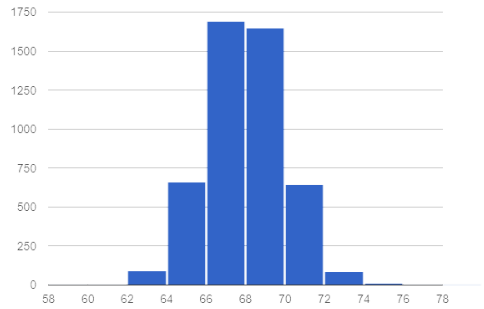
_____

_____

Bar Charts Versus Histograms

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





Differences and Similarities

Respond to the prompts to complete the table below.

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

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

_____

_____

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

_____

_____

Distribution of College Choice

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

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

_____

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

_____

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

_____

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

_____

Choosing the Right Bin Size

Open your saved Animals Starter File, or make a new copy, 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 a picture of the distribution with between 5 and 10 bins. _____

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

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

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

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

_____

_____

_____

_____

_____

_____

_____

_____

_____

9) What was a typical time to adoption?

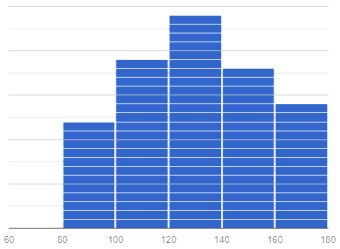
_____

_____

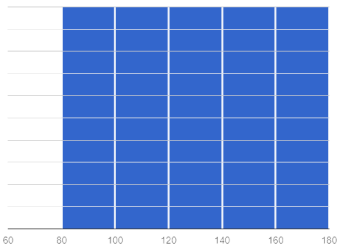
Histograms Card Sort

Cut out one set of cards for every pair of students. Instructions for this sorting activity can be found on Sorting Histograms.

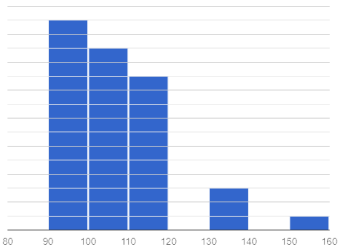
A



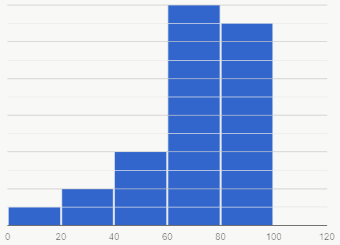
B



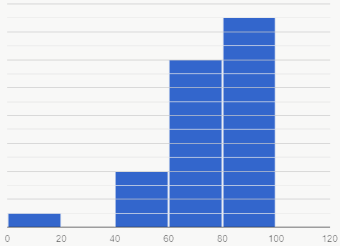
C



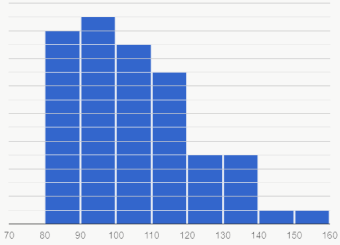
D



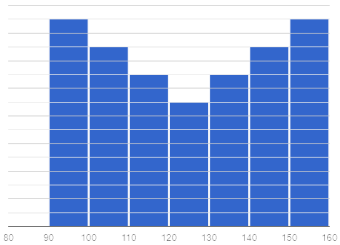
E



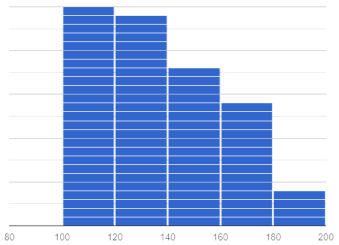
F



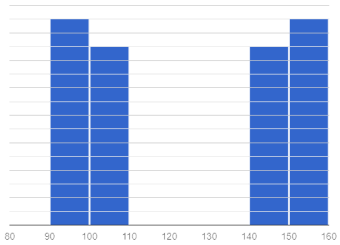
G



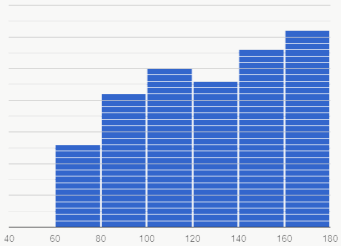
H



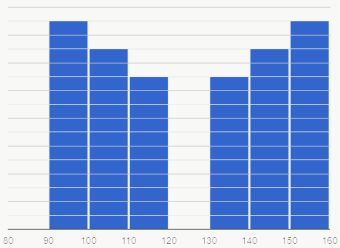
I



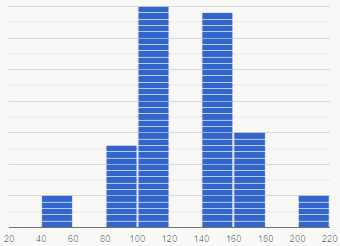
J



K



L



Sorting Histograms

With your partner, sort the histograms into two piles: approximately symmetrical and definitely not symmetrical. Then, follow the prompts and respond to the questions below.

Symmetrical Histograms

Put your asymmetrical cards aside (or back in their envelope).

1) List out the letters of the histograms that were symmetrical: _____

2) Sort the symmetrical cards into two or three logical groups. *Hint: It may be useful to think about peaks, gaps, clusters, center, and spread!* What do the cards in your first group have in common? _____

3) What do the cards in your second group have in common? (Describe your third group as well, if you have a third group.) _____

4) Can you think of a different way of grouping these histograms? Explain. _____

5) Describe how you can determine what's typical of a symmetrical histogram. _____

Asymmetrical Histograms

Put the symmetrical histograms away, and take out the asymmetrical histograms.

6) List out the letters of the histograms that were asymmetrical: _____

7) Sort the asymmetrical histograms into two or three logical groups. What do the cards in your first group have in common? _____

8) What do the cards in your second group have in common? (Describe your third group as well, if you have a third group.) _____

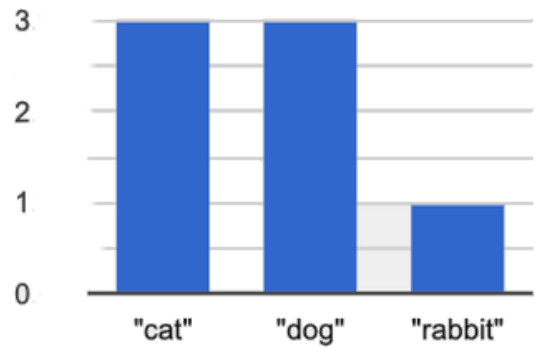
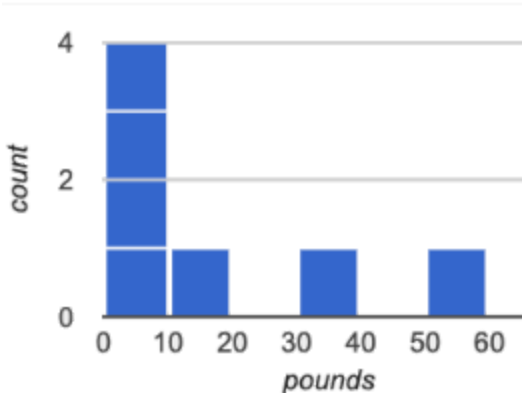
9) Can you think of a different way of grouping these histograms? _____

10) Describe how you can determine where the outliers are on an asymmetrical histogram. _____

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

The two data visualizations below both summarize columns of this table. In some ways they are similar. In other ways they are quite different!

	Bar Chart	Histogram
		
1) What type of data is displayed?		
2) Could we rearrange the order of the bars?		
3) Does having a sense that the bars are taller on the left than the right tell us anything about the dataset as a whole?		

What else do you Notice?	What else do you Wonder?

(optional)

Data Collection in a Nutshell

With Great Power Comes Great Responsibility

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

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

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

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

Data Cleaning

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

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

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

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

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

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

Survey Validation

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

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

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

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

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

Analyzing Survey Results When Data is Dirty

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

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

Make the `pie-chart` in Pyret to see what he's looking at. Why is this display misleading? How is the data "dirty"?

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

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

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

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

Dirty Data!

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

What do you Notice?	What do you Wonder?

There are many different ways that data can be dirty!

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

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

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

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

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

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

_____

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

_____

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

_____

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

_____

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

_____

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

_____

_____

Bad Questions Make Dirty Data

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

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

Probability, Inference, and Sample Size in a Nutshell

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

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

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

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

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

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

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

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

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

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

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

Finding the Trick Coin

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

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

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

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

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

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

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

_____

Sampling and Inference

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 **47.2%**
- 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 4.9%**
- The percentage of tarantulas in `small-sample` is _____
- The percentage of tarantulas in `large-sample` is _____

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

_____

_____

_____

Predictions from Samples

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

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

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

_____

- Click "Run" for a *new* random `tiny-sample`, and make *another* pie-chart for species. What animals are in this sample?

_____

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

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

_____

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

_____

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

_____

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

_____

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

_____

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

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

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

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

_____

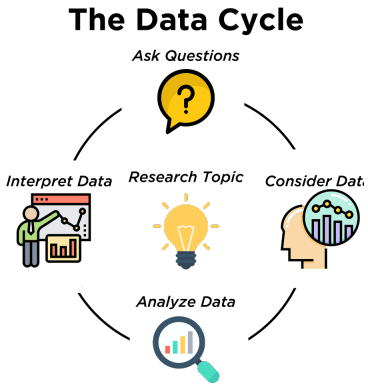
(optional)

The Data Cycle in a Nutshell

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

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

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



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

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

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

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

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

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

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

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

Which Question Type?

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

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

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

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

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

Data Cycle: Consider Data

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

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

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

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

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

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

Data Cycle: Categorical Distributions (Animals)

Using the Expanded Animals Starter File, let's make a **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 _____ (most / some / a few) species have _____ (more / the same number of / fewer) fixed animals _____ (as / than) unfixed animals. I also notice _____

Some new questions this raises include:

Data Cycle: Categorical Distributions 2 (Animals)

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

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

Question Types: Animals

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

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

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

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

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

(optional)

Data Cycle: Analyzing with Count

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

Ask Questions 	*How many of each species are 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?	

Ask Questions 	*How many of each sex are 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?	

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

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?	

(optional)

Rubric: Exploration Project (1)

About this Dataset

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

Criteria for Visualizations

☐ Wow!	☐ Getting There	☐ Needs Improvement
I either included multiple visualizations of this type or wrote about why I my data didn't allow for multiple. I indicated which column(s) I used and added the relevant code. I made a strong attempt to interpret the interesting visualizations and report about the visualizations that weren't useful. I added the questions that emerged to the "My Questions" section.	I included one display of this type. I provided the column name and relevant code. My interpretation lacked detail. I added the questions that emerged to the "My Questions" section.	I included one or no visualizations of this type. My slides may be missing a correct column name or code. My data interpretation may be missing or inaccurate. I may not have added to the "My Questions" section.

Visualizations	Rating	Teacher Feedback
Bar Chart	❏ Wow ❏ Getting There ❏ Needs Improvement	
Pie Chart	❏ Wow ❏ Getting There ❏ Needs Improvement	
Dot Plot	❏ Wow ❏ Getting There ❏ Needs Improvement	
Histogram	❏ Wow ❏ Getting There ❏ Needs Improvement	
Box Plot	❏ Wow ❏ Getting There ❏ Needs Improvement	

(handout)

Rubric: Exploration Project (2)

Measures of Center

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

Correlation and Linear Regression

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

(handout)

My Questions

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

Additional Teacher Feedback

Choosing Your Dataset in a Nutshell

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

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

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

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

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

Consider and Analyze

Fill in the tables below by considering the rows and columns you need. 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?	

Datasets and Starter Files

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

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

The Environment & Health

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

Politics

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

Sports

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

Entertainment

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

Education

College Majors	[Dataset Starter File]

U.S. Colleges 2019-2020	[Dataset Starter File]
★R.I. Schools	[Dataset Starter File]
Evolution of College Admissions in California	[Dataset Starter File]

Nutrition

Soda, Coffee & Other Drinks	[Dataset Starter File]
Fast Food Nutrition	[Dataset Starter File]

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

Relationships Between Quantitative Columns

Scatter Plots

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

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

Line of Best Fit

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

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

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

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

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

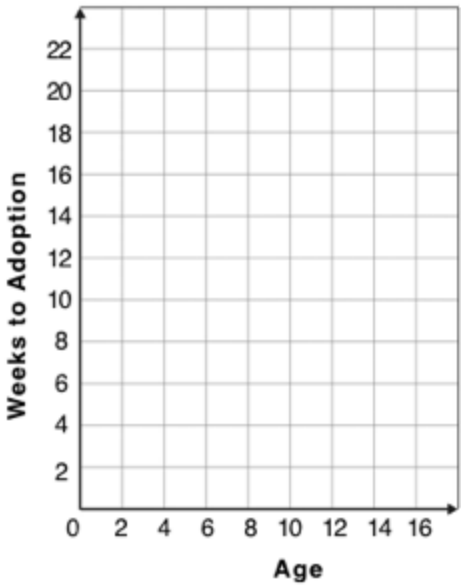
Creating a Scatter Plot

New Animals

1) The table below has some new animals!

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

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



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

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

Generalizing the pattern

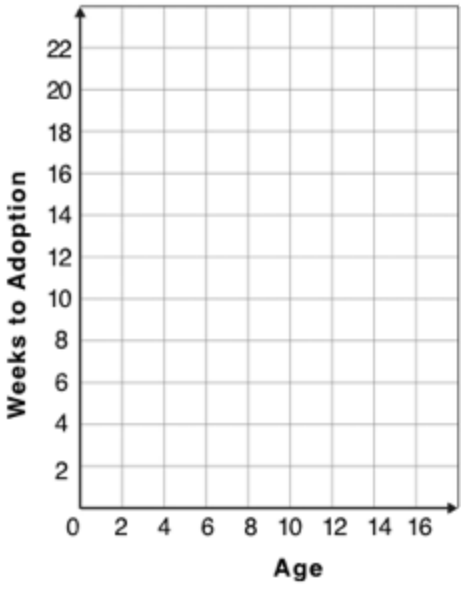
4) Use a straight edge to draw a line on the graph that best represents the pattern you see, then circle the cloud of points around that line.

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

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

_____

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



Exploring Relationships Between Columns

This page is designed to be used with the Animals Starter File. Log into 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` (explanatory variable) an animal weighs related to its `age` (response variable)?

- What would you expect? _____
- What did you learn from your scatter plot? _____

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

- What would you expect? _____
- What did you learn from your scatter plot? _____

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

- What would you expect? _____
- What did you learn from your scatter plot? _____

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

5) Are there any unusual observations?

Data Cycle: Looking for Relationships (Animals)

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

Ask Questions 	*Is there a relationship between weight and adoption time?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	What 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: Looking for Relationships (My Dataset)

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

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

Measures of Center in a Nutshell

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

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

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

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

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

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

Mean, Median, Mode(s) Practice

Mean

1) Find the mean of each dataset.

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

Median

2) Find the median of each dataset.

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

Mode(s)

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

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

Choosing the Best Measure of Center

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

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

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

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

_____

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

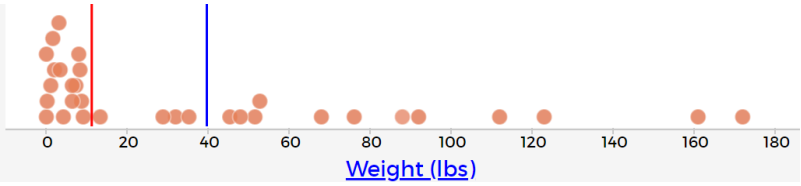
- Outliers on the right pull the mean toward the right, causing the mean to be _____ (greater than / less than) the median.

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

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

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

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



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

_____

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

_____

6) In the Interactions area of the Animals Starter File, type `modes(animals-table, "species")`. What does Pyret return? _____

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

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

_____

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

_____

_____

Critiquing Written Findings

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

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

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

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

Mean (Average)	Median	Mode(s)

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

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

Data Cycle: Measures of Center (Animals)

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

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

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

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

Using Shape to Interpret Data

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

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

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

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

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

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

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

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

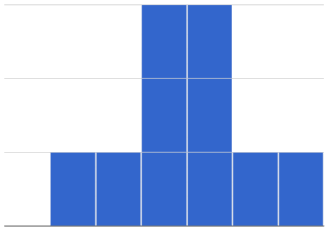
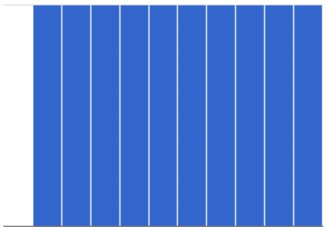
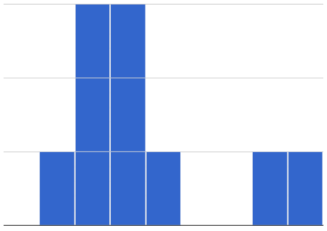
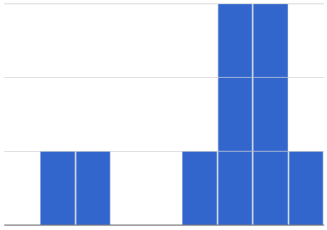
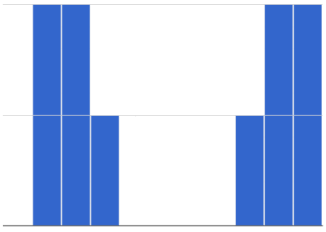
Rough histogram sketch:	**Circle one:** *skew left* *skew right* *symmetric* Explain your choice: _____

Reading Histograms

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

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

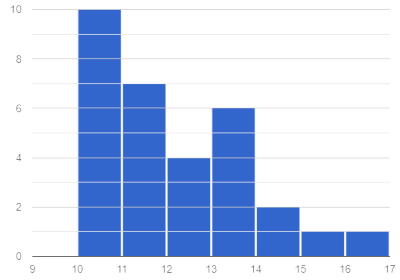
- The x-axis shows the score, and the y-axis shows the number of students who gave it that score.
- These axes are intentionally unlabeled - the **shapes** of the ratings distributions were very different! And that's the focus here.

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

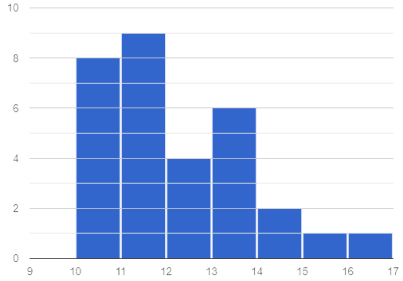
Histograms and Measures of Center

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

Distribution of Travel Times (Sixth Grade)



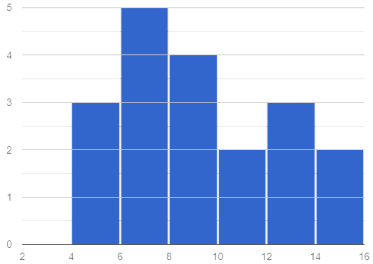
Distribution of Travel Times (Eighth Grade)



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

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

Distribution of Rock Climbers' Ages



The median age for these 19 climbers is:

□ about 6 or 7

□ about 8 or 9

□ about 10 or 11

□ about 12 or 13

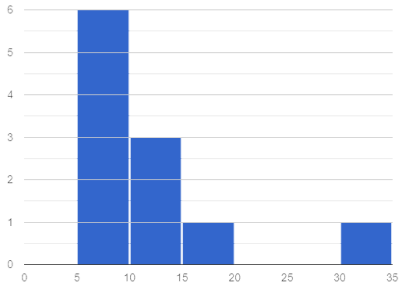
Explain how you determined the median value: _____

_____

_____

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

Distribution of Time Spent Solving a Logic Puzzle



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

□ The mean time is less than the median time.

□ The mean time is equal to the median time.

□ The mean time is greater than the median time.

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

Explain how you arrived at your choice: _____

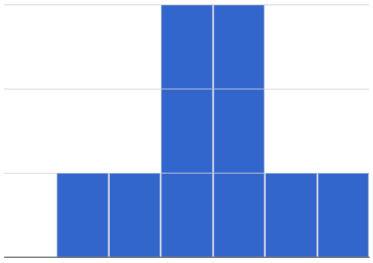
_____

_____

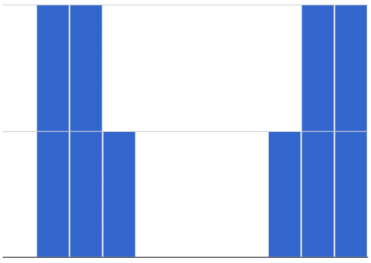
Histograms and Variability

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

Movie A



Movie B



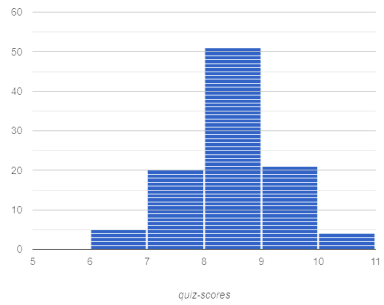
Comparing the two graphs, we know that:

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

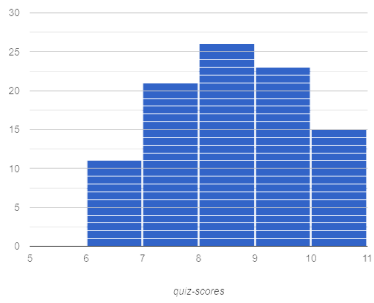
Explain how you arrived at your answer:

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

Class 1



Class 2



Comparing the two graphs, we know that:

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

Explain how you arrived at your answer:

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

Data Cycle: Quantitative Distributions (Animals) - Histograms

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

Ask Questions

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

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What 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 _____

Data Cycle: Quantitative Distributions (My Dataset) - Histograms

Open your chosen dataset. Use the Data Cycle to explore the distribution of one or more quantitative columns 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 	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?	

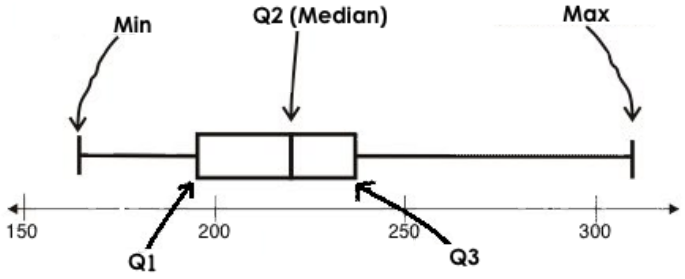
(optional)

Measures of Spread in a Nutshell

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

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

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



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

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

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

Distribution of a Dataset

Family Gatherings by the Numbers

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

1) Order the Ages from Least to Greatest: _____

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

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

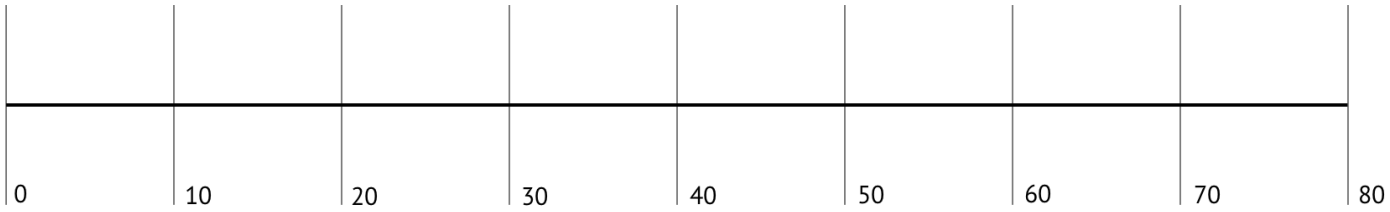
2) Order the Ages from Least to Greatest: _____

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

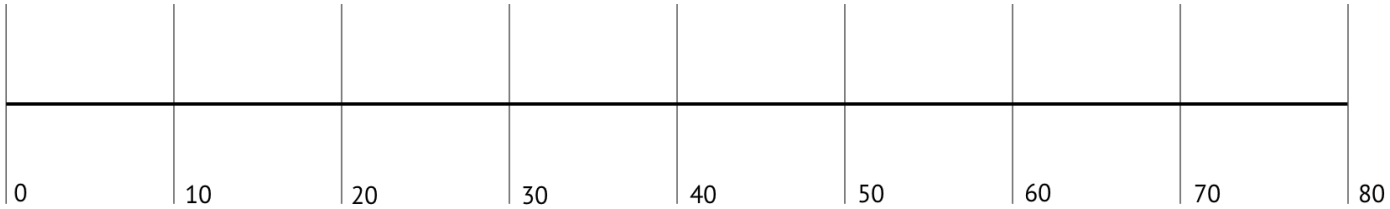
Box Plots - Visualizing Shape

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

3) Ledet:



4) Watson:



Compare and Contrast

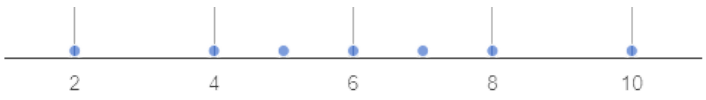
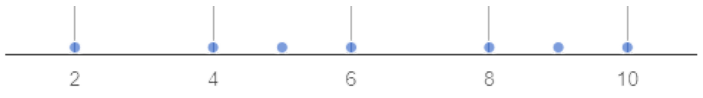
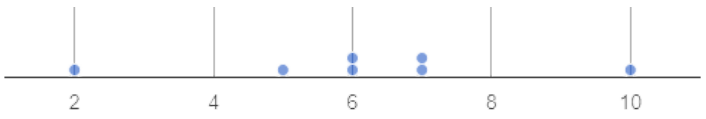
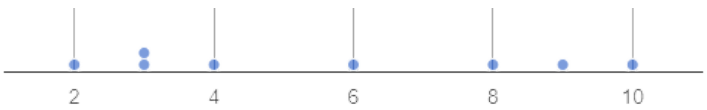
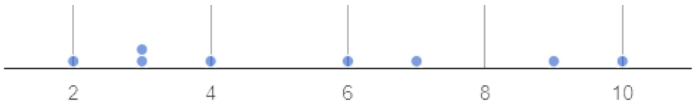
5) For which family gathering was the average age more typical? How do you know? _____

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

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

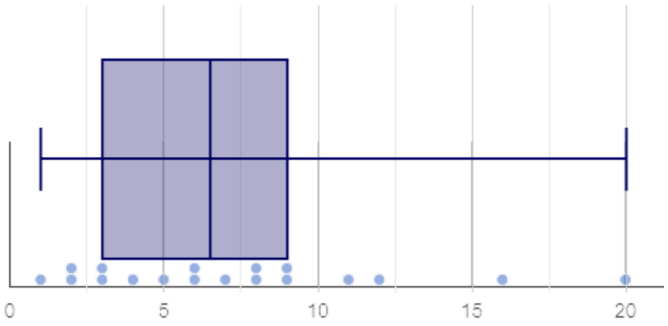
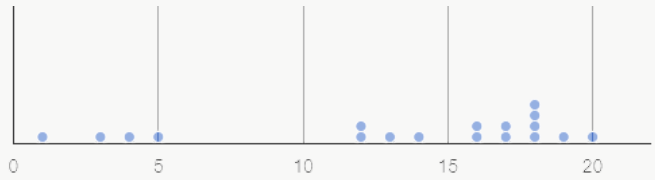
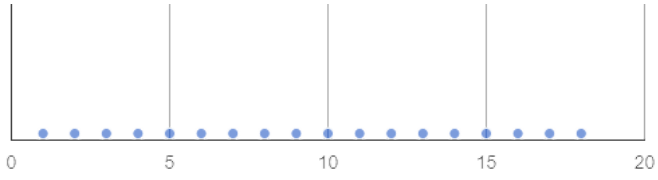
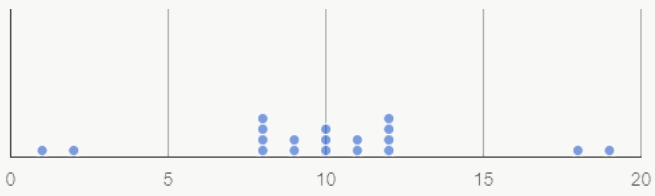
Matching Dot Plots and Five-Number Summaries

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

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

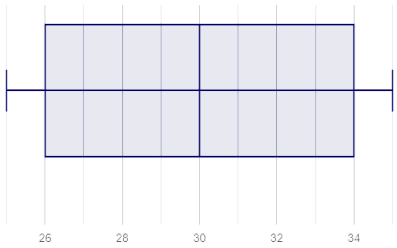
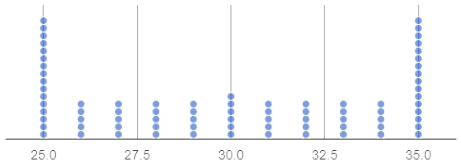
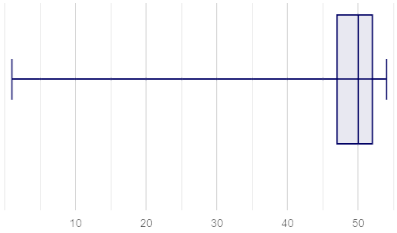
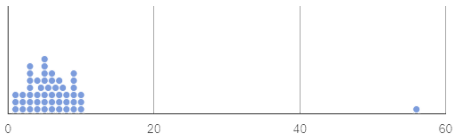
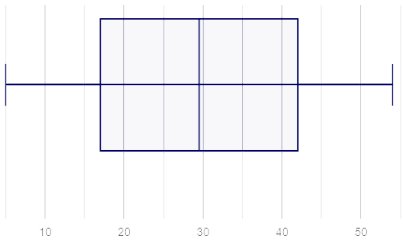
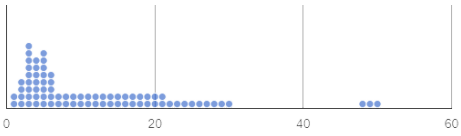
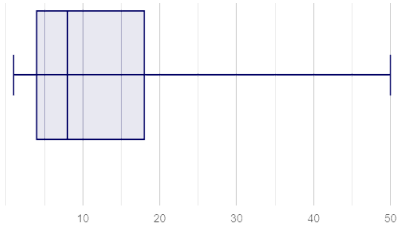
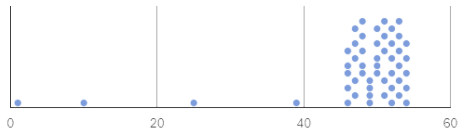
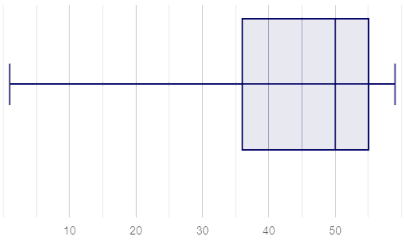
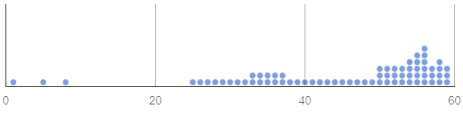
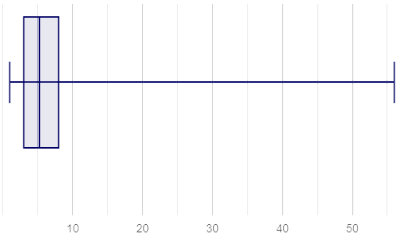
Create Box Plots from Dot Plots

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

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

Matching Dot Plots and Box Plots

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

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

Summarizing Columns with Measures of Spread

Summarizing the Pounds Column

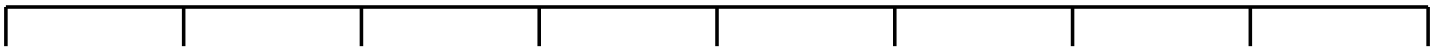
Get the values to summarize the spread of the _____pounds_____ column of the Animals Starter File by typing

`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:

_____

_____

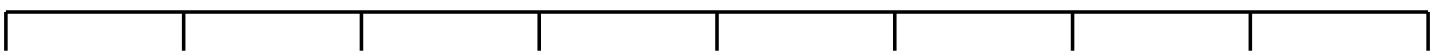
Summarizing the _____ Column

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: _____.

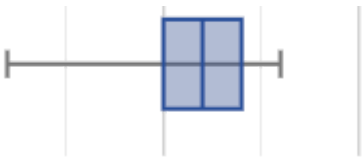
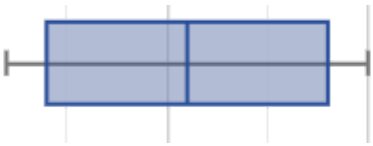
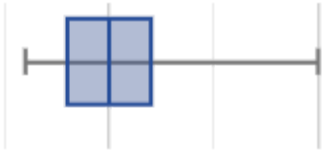
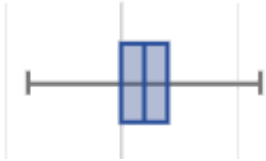
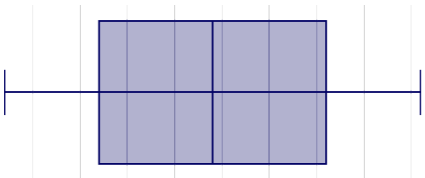
8) From this summary and box plot, I conclude that:

_____

_____

Identifying Shape - Box Plots

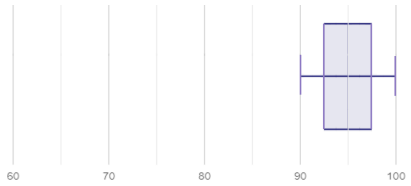
Describe the shape of the box plots on the left. *Do your best to incorporate the vocabulary you've been introduced to.*

1		
2		
3		
4		
5		

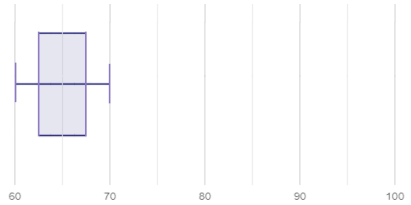
(optional)

Reading Box Plots

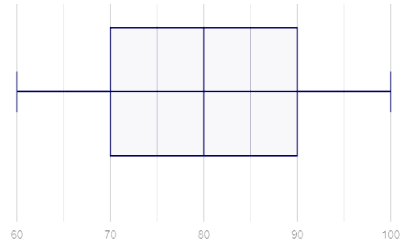
There are six different retirement homes in Retirement City. Each box plot (left) shows the spread of ages at one of the retirement homes. Match each box plot with the appropriate description (right) of residents' ages.



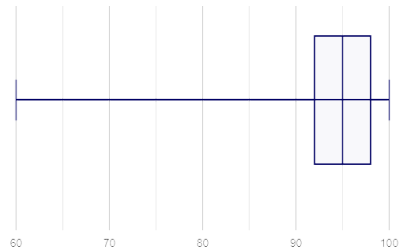
1



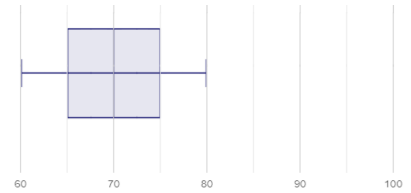
2



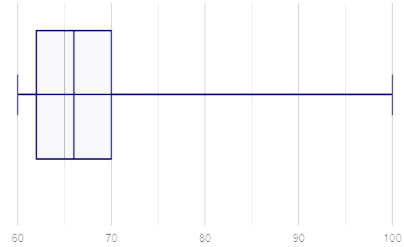
3



4



5



6

A At Apple Tree Independent Living, the majority of residents are at least 90 years old, but there are few younger residents as well.

B The residents of Cherry Hill Retirement Home are all at least 90 years old.

C The majority of seniors at the Edgewood Home are younger than 70, but there are some exceptions.

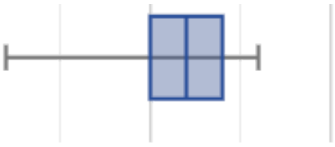
D The residents of Juneau Independent Living are mostly in their sixties.

E At Horizon Retirement Center, residents' ages are evenly distributed from 60 to 100.

F The median age at Live Oak Assisted Living is 70.

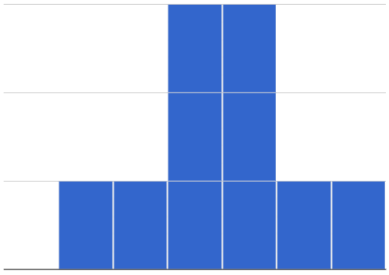
Matching Box Plots to Histograms

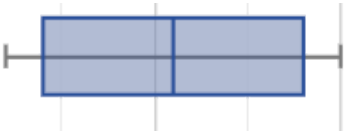
Students watched 5 videos, and rated them on a scale of 1 to 10. For each video, their ratings were used to generate box plots and histograms. Match each box plot to the histogram that displays the same data.



1

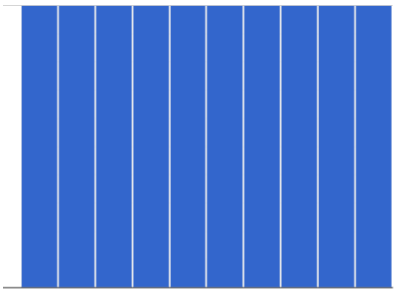
A

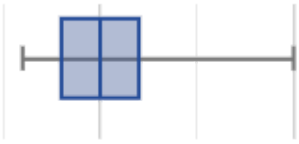




2

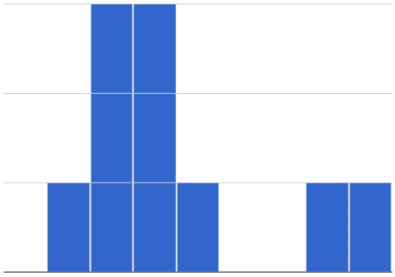
B

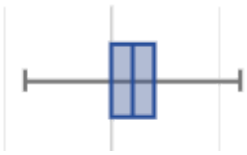




3

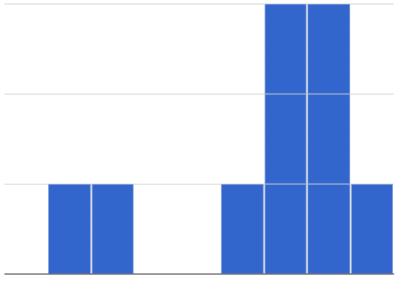
C

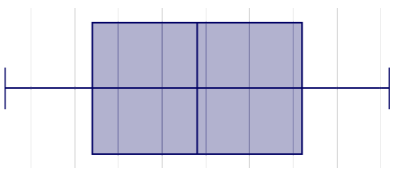




4

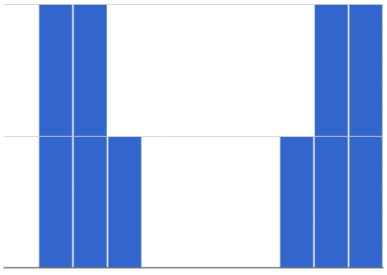
D





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: Quantitative Distributions - Box Plots (Animals)

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 _____ (x-variable in context) is _____ (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 _____ (x-variable in context) is _____ (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: Quantitative Distributions - Box Plots (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.

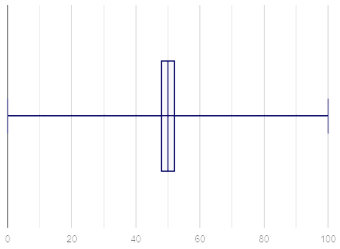
Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some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

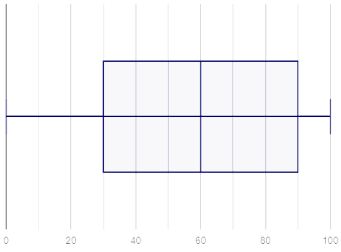
Box Plots Card Sort

Cut out one set of cards for every pair of students. Instructions for this sorting activity can be found on Sorting Box Plots.

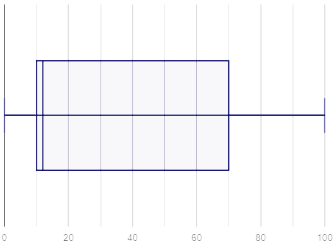
A



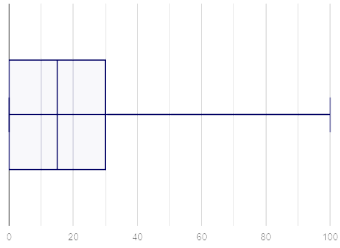
B



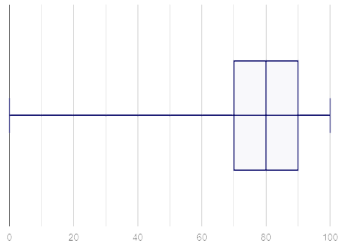
C



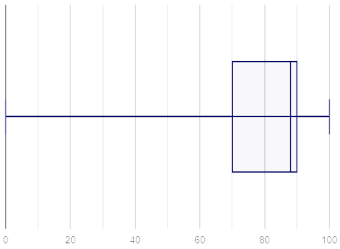
D



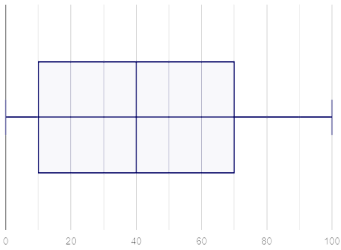
E



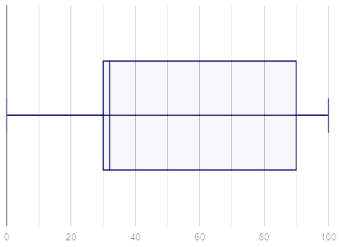
F



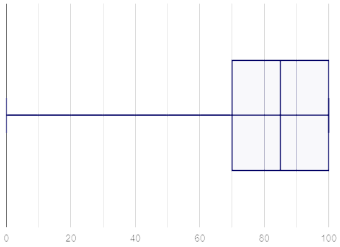
G



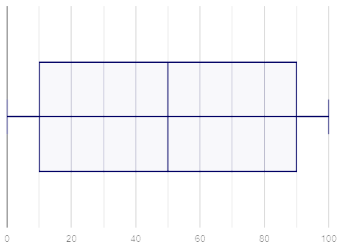
H



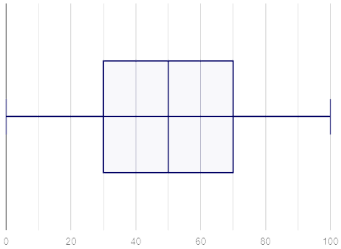
I



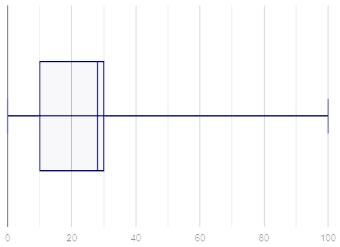
J



K



L



Sorting Box Plots

Notice & Wonder

1) What do you Notice about the box plots? _____

_____

2) What do you Wonder about the box plots? _____

_____

Grouping Option #1

Identify two or three logical and related groupings that you could use to sort each of the box plot cards.

3) Group 1 Description: _____ Cards that belong: _____

4) Group 2 Description: _____ Cards that belong: _____

5) *Optional:* Group 3 Description: _____ Cards that belong: _____

6) Why are these groupings useful? What can you learn about a box plot based on the feature that you've chosen to focus on? _____

_____

_____

_____

Grouping Option #2

Identify two or three logical and related groupings that you could use to sort each of the box plot cards.

7) Group 1 Description: _____ Cards that belong: _____

8) Group 2 Description: _____ Cards that belong: _____

9) *Optional:* Group 3 Description: _____ Cards that belong: _____

10) Why are these groupings useful? What can you learn about a box plot based on the feature that you've chosen to focus on? _____

_____

_____

_____

Grouping Option #3

Identify two or three logical and related groupings that you could use to sort each of the box plot cards.

11) Group 1 Description: _____ Cards that belong: A, D, E, F, I, K, L

12) Group 2 Description: _____ Cards that belong: _____

13) *Optional:* Group 3 Description: _____ Cards that belong: _____

14) Why are these groupings useful? What can you learn about a box plot based on the feature that you've chosen to focus on? _____

_____

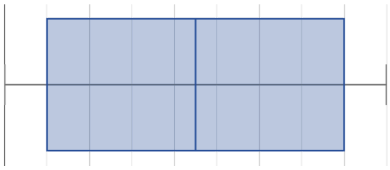
_____

_____

(optional)

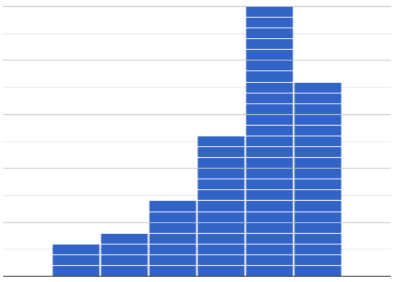
Matching Box Plots to Histograms 2

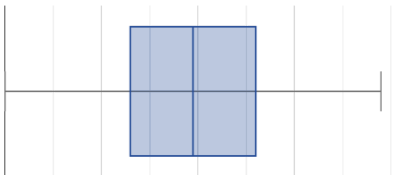
Match each box-plot to the histogram that displays the same data.



1

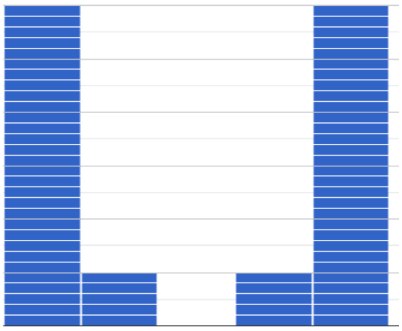
A

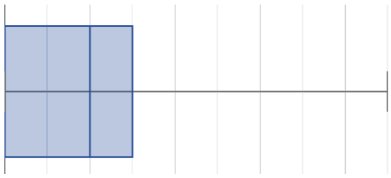




2

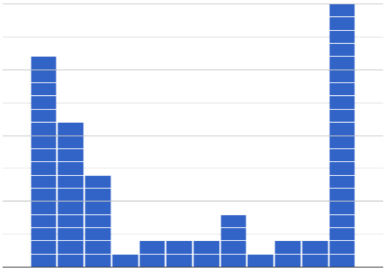
B





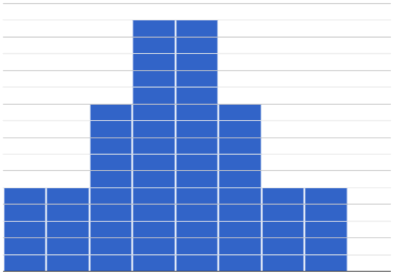
3

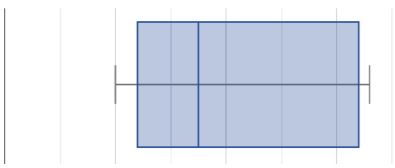
C



4

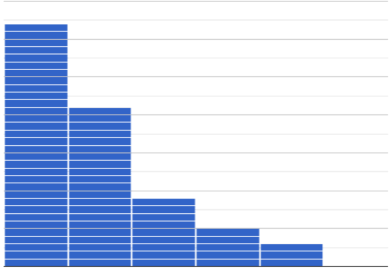
D





5

E



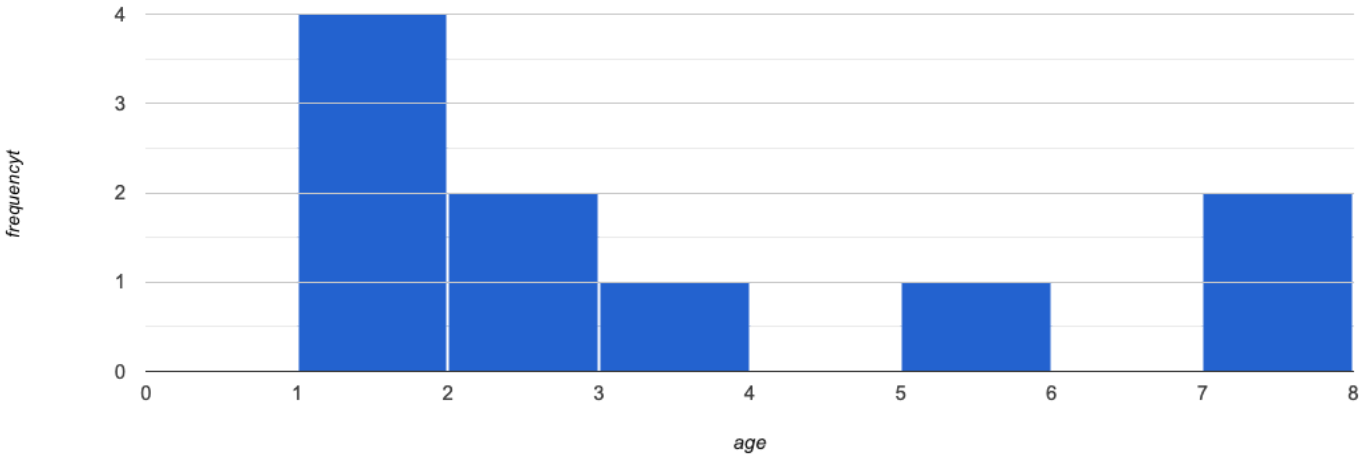
(optional)

Computing Standard Deviation

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

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

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



2) Describe the shape of this histogram. _____

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

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

age	1	2	3	4	5	6	7
frequency	4						

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

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

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

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

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

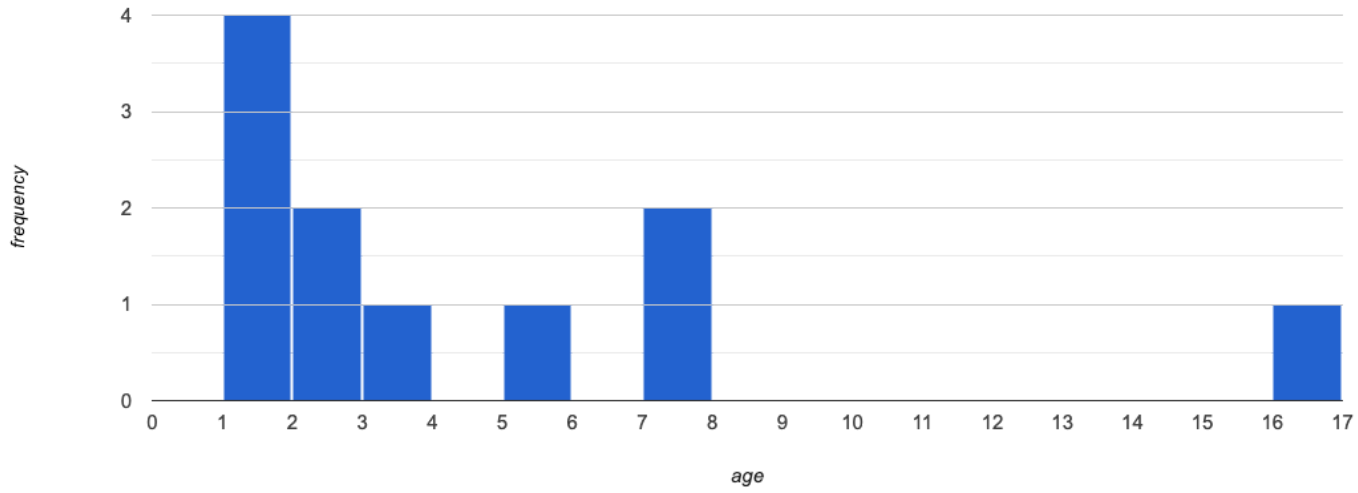
8) Add all the squared distances. What is their sum? _____

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

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

The Effect of an Outlier

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



1) Describe the shape of this histogram. _____

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

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

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

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

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

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

age of cat											
distance from mean											
squared distance											

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

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

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

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

_____

Data Cycle: Measure of Spread (Animals)

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

Ask Questions 	*Do the animals all get adopted in around the same length of time?* What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, 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: Measure of Spread (My Dataset)

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

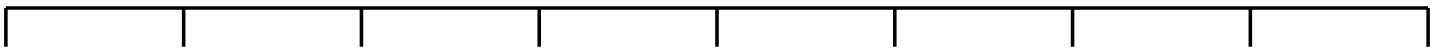
Ask Questions 	What question do you have?	**Question Type** (circle one): Lookup Arithmetic Statistical
Consider Data 	Which Rows should we investigate? (All the rows, just the cats, fixed dogs, etc.) What Column(s) do we need? (age, weight-in-kilograms, weeks, etc.)	
Analyze Data 	If you only need some rows, 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 (2)

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

(optional)

Matching Mean & Standard Deviation to Data

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

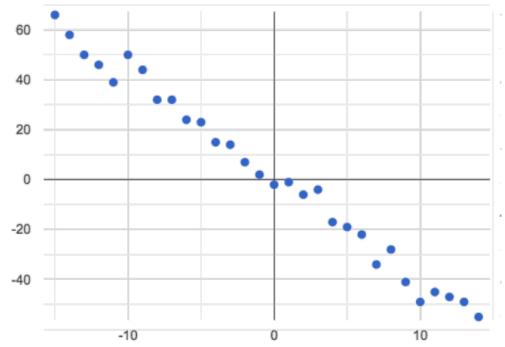
Mean: 4 *StDev:* 0	**1**	**A**	`-1,-2,-3,-4,-5,-6,-7`
Mean: -5 *StDev:* ~5.66	**2**	**B**	`1, 2, 3, 4, 5, 6, 7`
Mean: 4 *StDev:* ~2.16	**3**	**C**	`-1, -9`
Mean: 4 *StDev:* ~2.65	**4**	**D**	`0, 2, 3, 4, 5, 6, 8`
Mean: -4 *StDev:* ~2.16	**5**	**E**	`4, 4, 4, 4, 4`

(optional)

Identifying Form, Direction and Strength

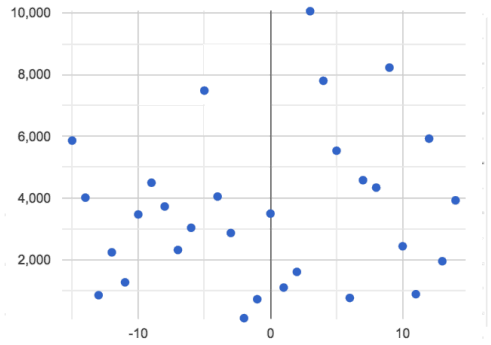
What do your eyes tell you about the Form, Direction, & Strength of these visualizations?
Note: If the form is nonlinear, we shouldn't report direction - a curve may rise and then fall.

A



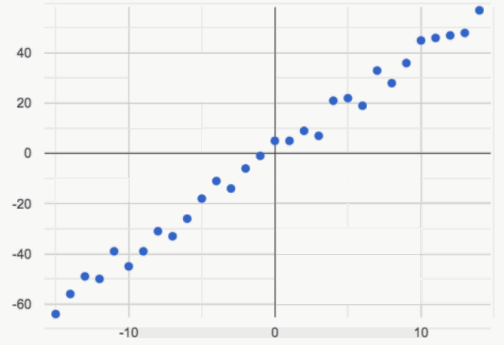
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

B



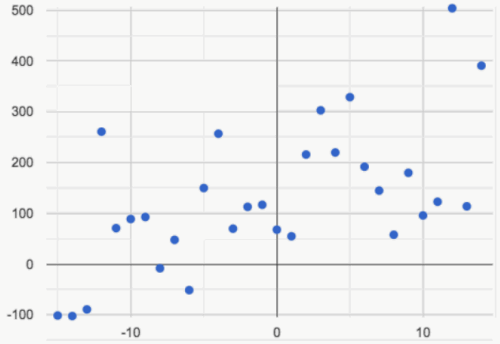
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

C



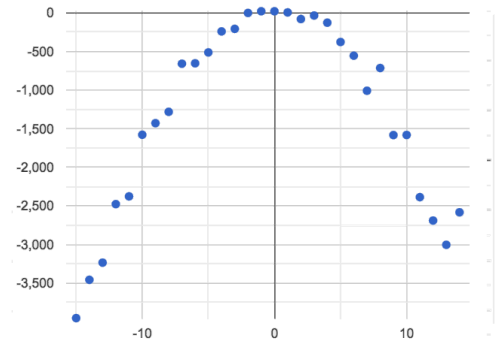
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

D



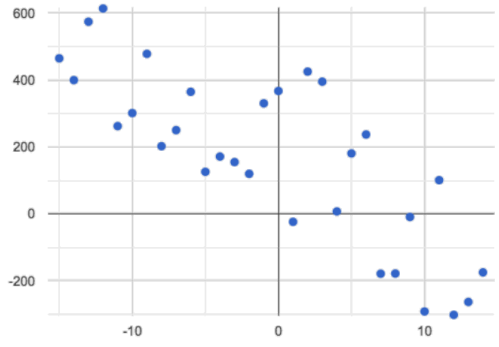
Form:	Linear	Nonlinear	None
Direction:	Positive	Negative	N/A
Strength:	Strong	Weak	

E



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

F



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

Reflection on Form, Direction and Strength

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

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

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

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

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

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

Summarizing Correlations with r-values

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

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

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

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

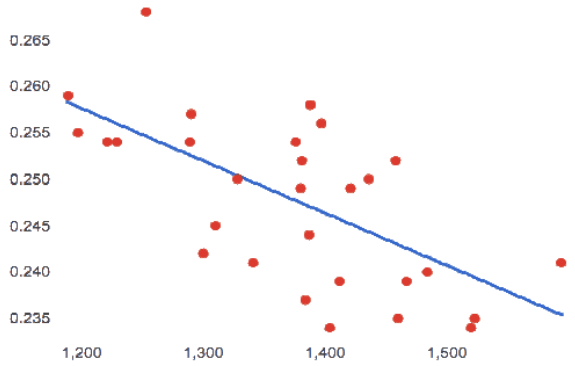
Identifying Form and r-Values

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

Reminder:

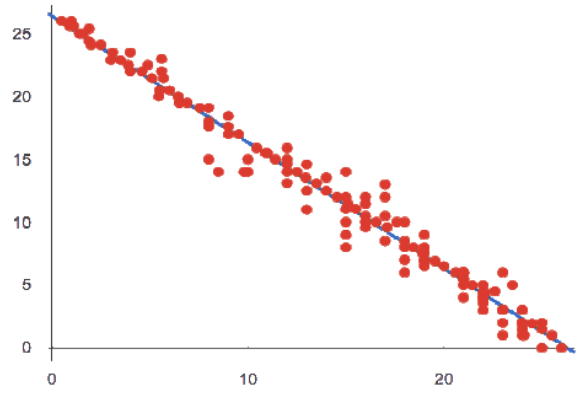
- −1 is the strongest possible *negative* correlation, and +1 is the strongest possible *positive* correlation
- 0 means no correlation
- ±0.65 or ±0.70 or more is typically considered a "strong correlation"
- ±0.35 to ±0.65 is typically considered "moderately correlated"
- Anything less than about ±0.25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

A



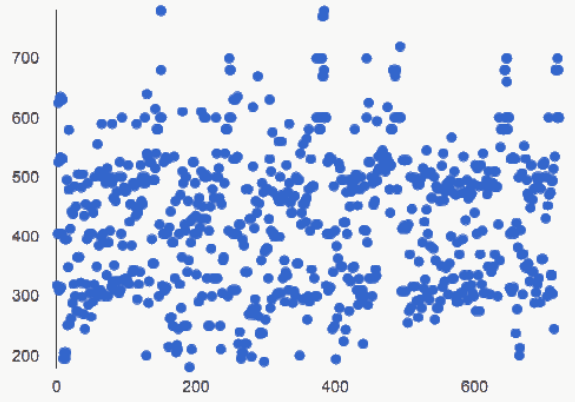
Form:
r close to:

B



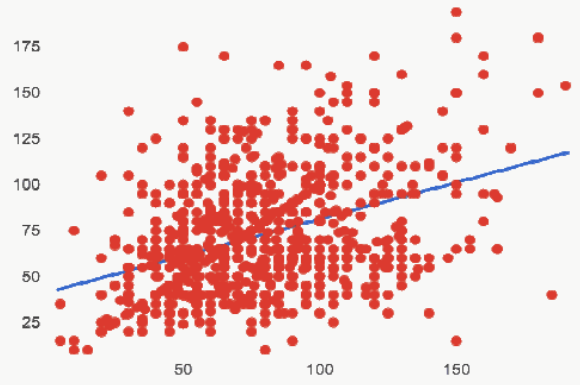
Form:
r close to:

C



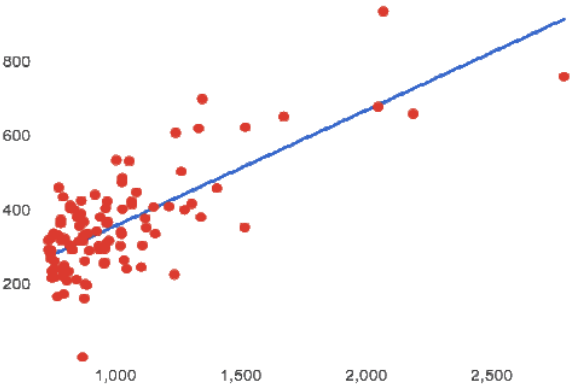
Form:
r close to:

D



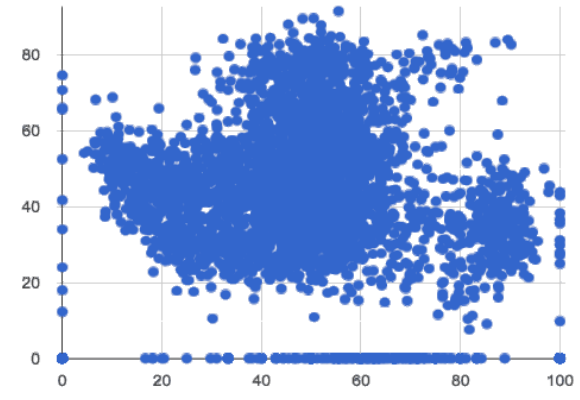
Form:
r close to:

E



Form:
r close to:

F



Form:
r close to:

Correlation Does Not Imply Causation!

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

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

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

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

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

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

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

Correlations in the Animals Dataset

1) 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 _____ (column) and _____ (column).

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

because _____

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

---

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

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

because _____

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

---

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

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

because _____

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

---

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

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

because _____

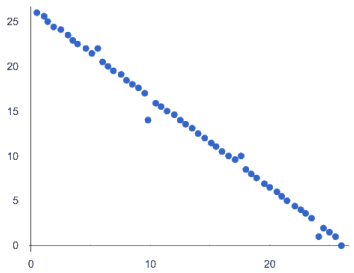
It might be stronger if I looked at _____ (a sample or extension of my data)

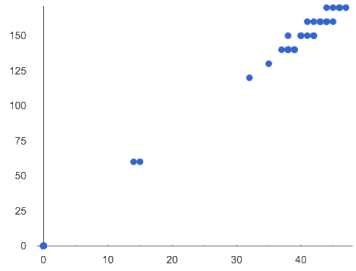
Identifying Form, Direction and Strength (Matching)

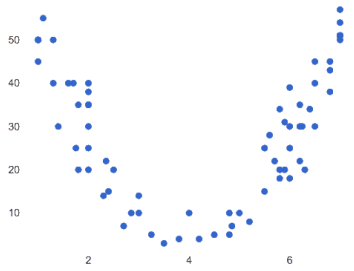
Match the description (left) with the scatter plot (right).

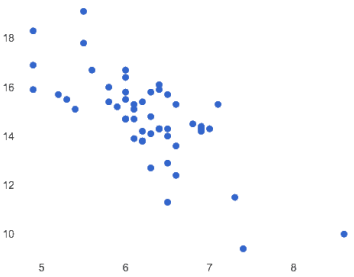
Note: The computer won't tell us if the relationship we see in a scatter plot is linear, so it's important to train our eyes to decide this ourselves. For linear relationships, we should train our eyes to assess their direction and get a feel for their strength, so that we have a sense of whether the computed results make sense.

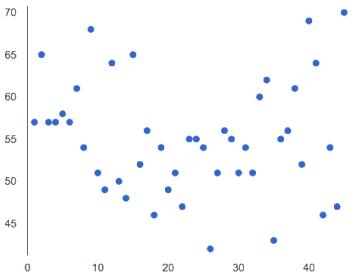
The relationship appears to be linear, negative, and of moderate strength.	1	A
This relationship is nonlinear.	2	B
The x and y variables in this dataset do not appear to be related.	3	C
The relationship appears to be linear, positive, and strong.	4	D
The relationship appears to be linear, negative, and strong.	5	E











(optional)

Introduction to Linear Regression

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

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

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

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

_____

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

_____

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

_____

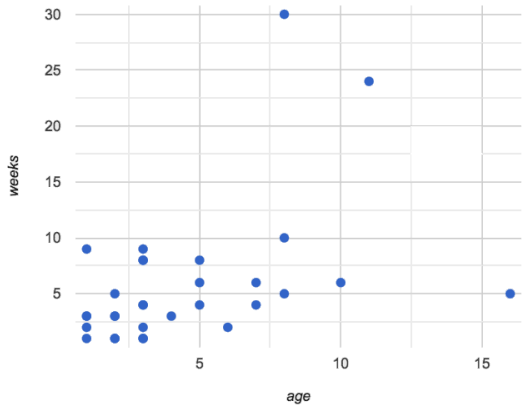
5) What is the highest r -value you can get? _____ Where did you place P? (_____, _____)

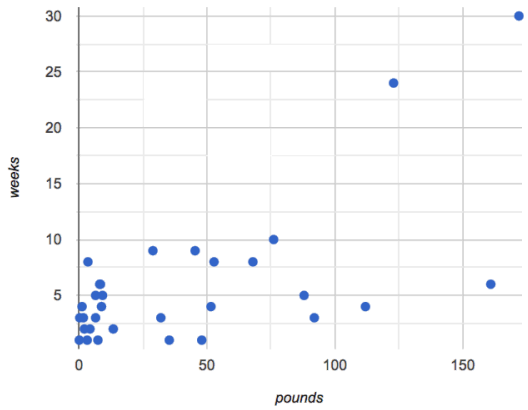
6) What function describes the regression line with this value of P? $y =$ _____ $x +$ _____

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

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

Predictions from Scatter Plots





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

_____

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

_____

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

_____

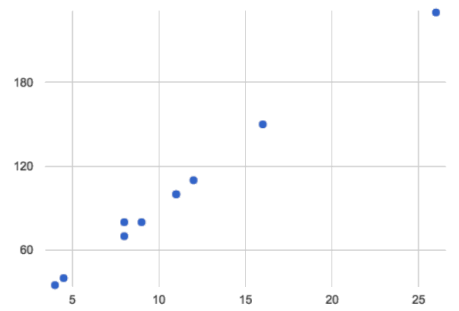
Drawing Predictors

Remember what we learned about r-values…

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

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

A



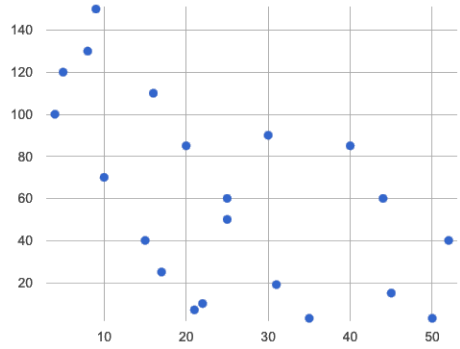
Direction: Positive Negative None

Strength: Stronger Weaker

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

-1 -0.5 0 0.5 1

B



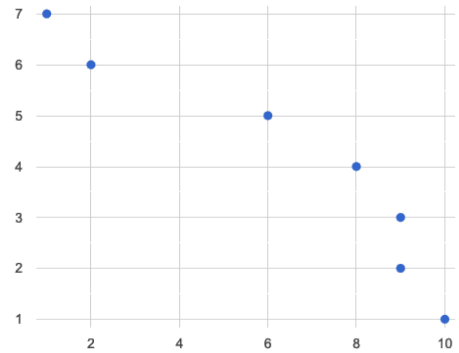
Direction: Positive Negative None

Strength: Stronger Weaker

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

-1 -0.5 0 0.5 1

C



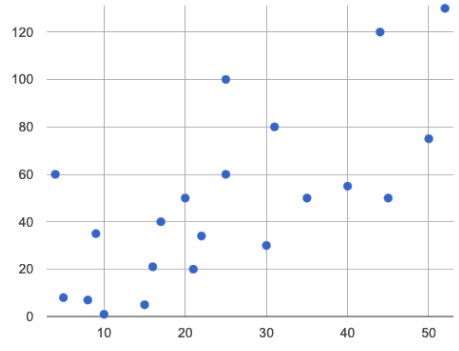
Direction: Positive Negative None

Strength: Stronger Weaker

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

-1 -0.5 0 0.5 1

D



Direction: Positive Negative None

Strength: Stronger Weaker

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

-1 -0.5 0 0.5 1

Exploring lr-plot

age

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

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

2) What is the slope? _____

3) What is the y-intercept? _____

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

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

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

_____

weight

Make another 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? $y =$ _____ $x +$ _____ $r =$ _____

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*? _____

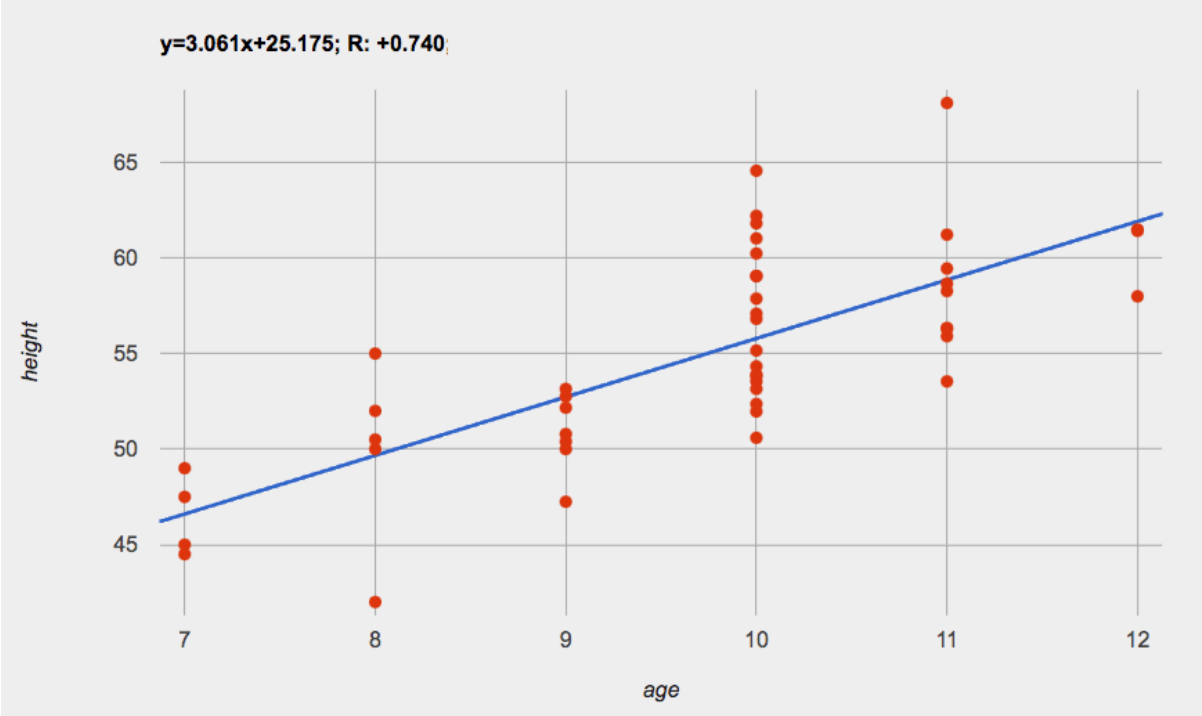
_____

_____

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

Data Cycle: Regression Analysis (Animals)

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

Ask Questions

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

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

`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?

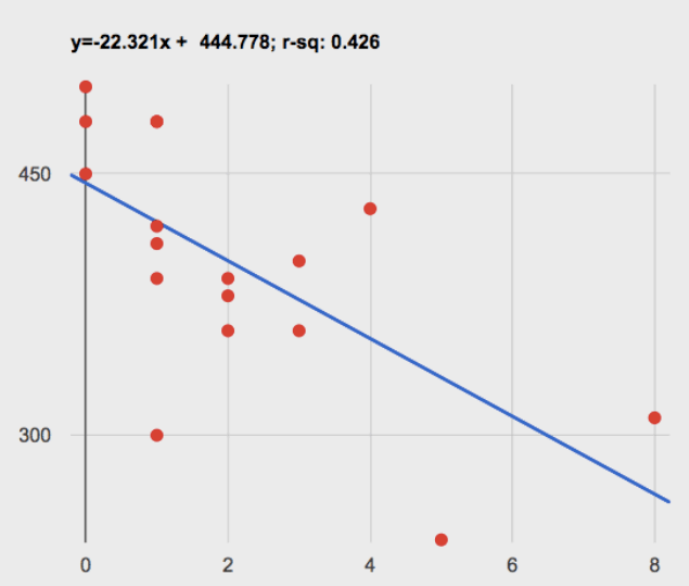
Interpret Data

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 (My Dataset)

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

Ask Questions

What question do you have?

Question Type (circle one):
Lookup
Arithmetic
Statistical

Consider Data

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

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

Analyze Data

What 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].

Age vs. Height Explore

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

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

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

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

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

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

4) What makes it difficult to compare these visualizations?

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

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

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

5) How do these visualizations compare now that their windows match?

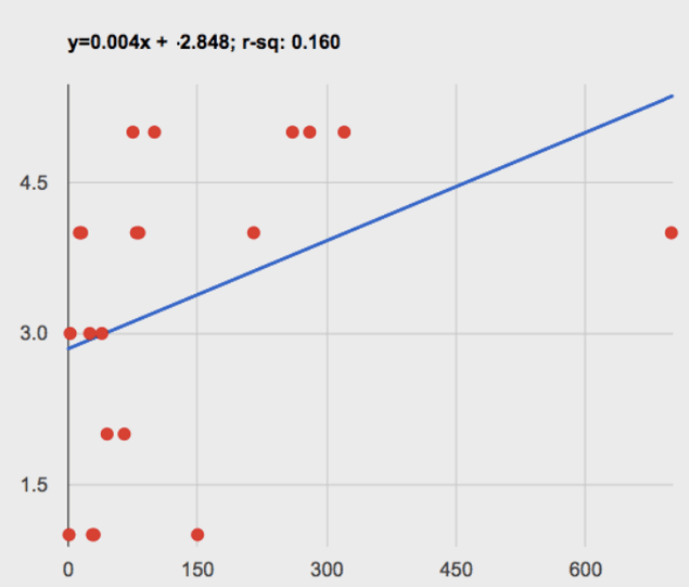
6) What happens if you compare the students' height in inches to their height in centimeters by plotting `lr-plot(h-table, "gender-id", "height-cm", "height")`?

(optional)

Describing Relationships (2)

A small sample of people were surveyed about their satisfaction with their most recent purchase using a scale from 1 (very unsatisfied) to 5 (extremely satisfied).
NOTE: this data is made up for instructional purposes!

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



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

(optional)

Data Cycle: Regression Analysis 2 (My Dataset)

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

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

Write your Data Story below:

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).

(optional)

Threats to Validity in a Nutshell

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

People Make Mistakes

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

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

Fake News

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

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

Outliers: Do they stay or do they go?

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

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

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

Identifying Threats to Validity

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

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

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

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

Identifying Threats to Validity (2)

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

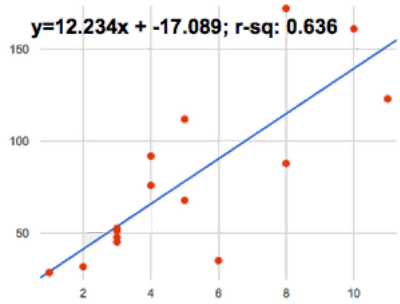
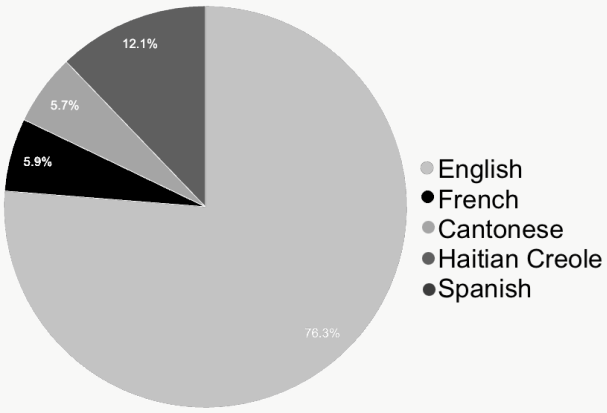
1) Explain why sampling just two animals can result in unreliable conclusions about what kind of food is needed.

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

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

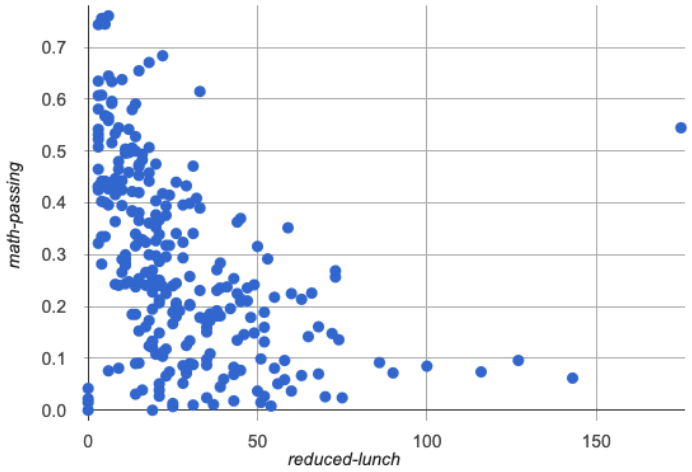
Fake News

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

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

Outliers: Should they Stay or Should they Go?

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

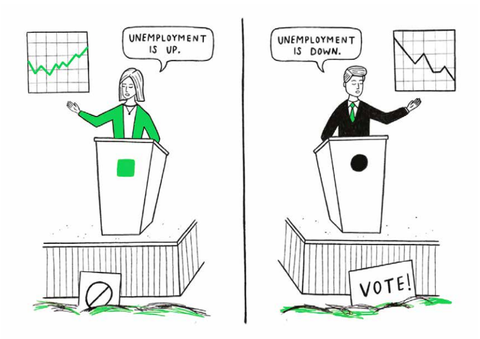


Tahli thinks the outlier should be removed before they start analyzing, and Fernando thinks it should stay. Here are their reasons:

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

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

Data Fallacies to Avoid

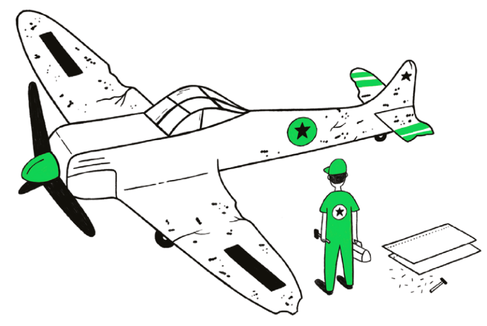


Cherry Picking

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

Data Dredging

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

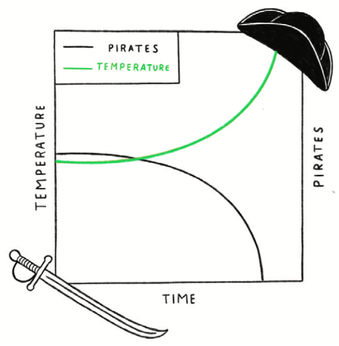


Survivorship Bias

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

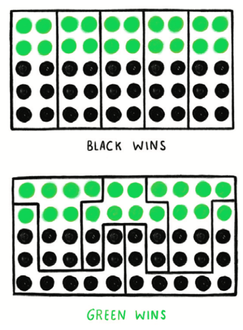
Cobra Effect

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



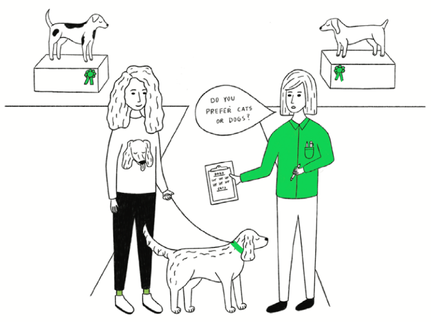
False Causality

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



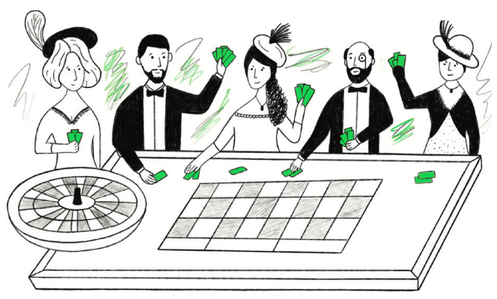
Gerrymandering

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



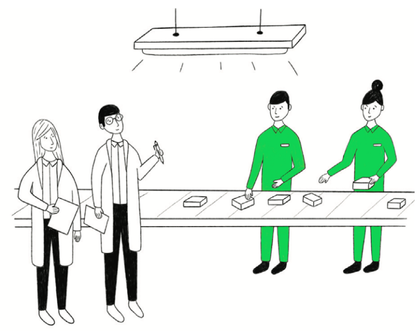
Sampling Bias

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



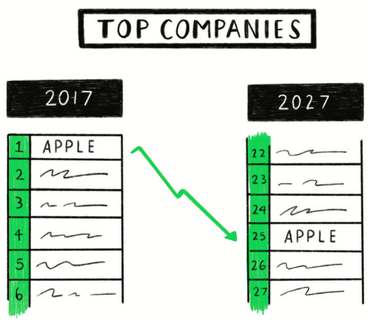
Gambler's Fallacy

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



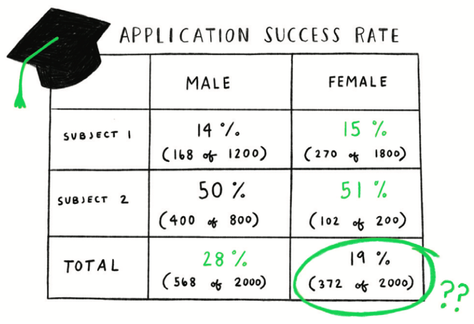
Hawthorne Effect

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



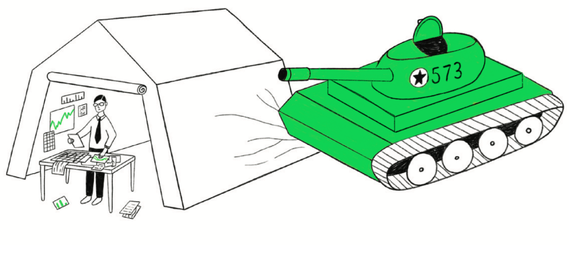
Regression Towards the Mean

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



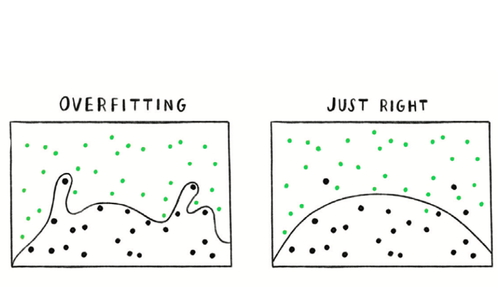
Simpson's Paradox

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



McNamara Fallacy

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

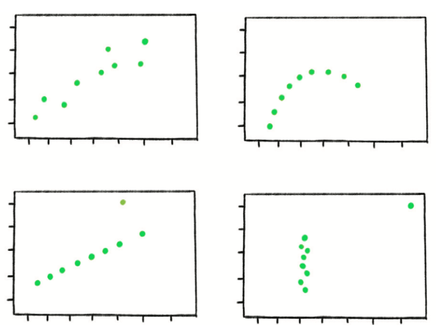


Overfitting

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

Publication Bias

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



Danger of Summary Metrics

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

geckoboard

Read more at
geckoboard.com/data-fallacies

Selection Bias or Biased Study?

The school newspaper ran an article stating that chicken was more popular than pork in the East Village.

Kendell thinks the study was biased.

Would you rather eat pork or delicious crispy fried chicken? That's such a leading question! It encouraged people to pick chicken. I bet the results would have been different if they had asked about crispy bacon!

Carson thinks the study suffered from selection bias.

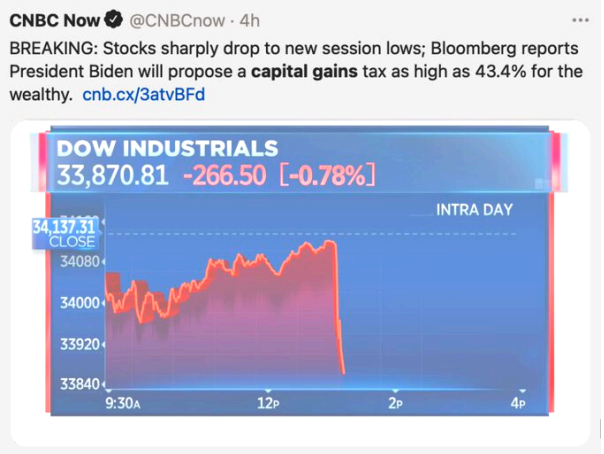
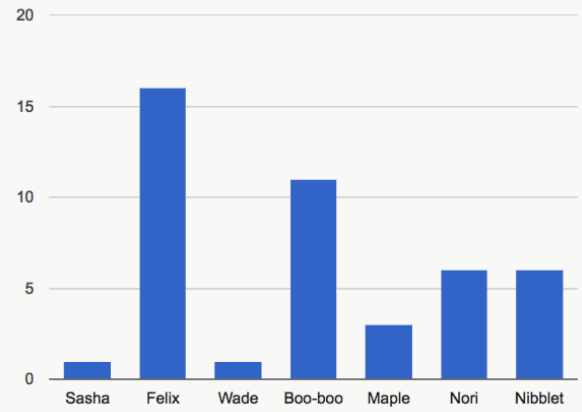
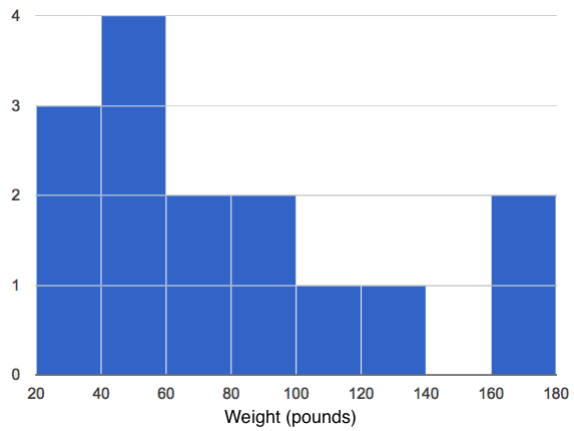
One of the survey sites was outside of a mosque?! Muslims don't even eat pork!

Who's right? How do you know?

(optional)

Fake News (2)

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

	The News Says...	Why it's Fake
1	 *"Stocks fell off a cliff today, plummeting sharply. Some fear the economic unrest will lead to rioting and looting!"*	
2	 *"According to this bar chart, Felix makes up a little more than 15% of the total ages of all the animals in the dataset."*	
3	 *"According to this histogram, most animals weigh between 40 and 60 pounds."*	

(optional)

Identifying Threats to Validity (3)

Data scientists want to know if listening to music or podcasts reduces symptoms of stress in individuals.

- They conducted **a study of 1,000 people who were brought into a laboratory office for testing**.
- While wearing a heart-rate monitor, participants were asked to listen to either music or a podcast of their choosing while completing a series of complicated puzzles.
- The data scientists discovered that on average, **participants who listened to music had a 5% lower heart rate while completing the tasks than those who listened to podcasts**.

Before publishing their findings, the data scientists have asked you to review their claim. In the space below, indicate possible **threats to validity** faced by this study.

(optional)

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 Literacy

Contracts tell us how to use a function, by telling us three important things:

1. The **Name**
2. The **Domain** of the function - what kinds of inputs do we need to give the function, and how many?
3. The **Range** of the function - what kind of output will the function give us back?

For example: The contract `triangle :: (Number, String, String) -> Image` tells us that the name of the function is `triangle`, it needs three inputs (a Number and two Strings), and it produces an Image.
With these three pieces of information, we know that typing `triangle(20, "solid", "green")` will evaluate to an Image.

Name	Domain	Range
`# 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")`		
`# 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)`		
`# circle`	`:: ( Number (radius) , String, (fill-style) , String (color) )`	`-> Image`
`circle(50, "solid", "purple")`		
`# count`	`:: ( Table (table-name) , String (column) )`	`-> Table`
`count(animals-table, "species")`		
`# dot-plot`	`:: ( Table (table-name) , String (labels) , String (values) )`	`-> Image`
`dot-plot(animals-table, "name", "pounds")`		
`# ellipse`	`:: ( Number (width) , Number (height) , String (fill-style) , String (color) )`	`-> Image`
`ellipse(100, 50, "outline", "orange")`		
`# first-n-rows`	`:: ( Table (table-name) , Number (num-rows) )`	`-> Table`
`first-n-rows(animals-table, 15)`		
`# histogram`	`:: ( Table (table-name) , String (labels) , String (values) , Number (bin-size) )`	`-> Image`
`histogram(animals-table, "species", "weeks", 2)`		
`# isosceles-triangle`	`:: ( Number (size) , Number (vertex-angle) , String (fill-style) , String (color) )`	`-> Image`
`isosceles-triangle(50, 20, "solid", "grey")`		
`# line-graph`	`:: ( Table (table-name) , String (labels) , String (xs) , String (ys) )`	`-> Image`
`line-graph(animals-table, "name", "pounds","weeks")`		
`# lr-plot`	`:: ( Table (table-name) , String (labels) , String (xs) , String (ys) )`	`-> Image`
`lr-plot(animals-table, "name", "pounds","weeks")`		
`# mean`	`:: ( Table (table-name) , String (column) )`	`-> Number`
`mean(animals-table, "pounds")`		

Name	Domain	Range
# median	:: (Table (table-name) , String (column))	-> Number
median(animals-table, "pounds")		
# modes	:: (Table (table-name) , String (column))	-> List
modes(animals-table, "pounds")		
# modified-box-plot	:: (Table (table-name) , String (column))	-> Image
modified-box-plot(animals-table, "pounds")		
# modified-box-plot-scaled	:: (Table (table-name) , String (column) , Number (low) , Number (high))	-> Image
modified-box-plot-scaled(animals-table, "weeks", 1, 40)		
# 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")		
# 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")		
# r-value	:: (Table (table-name) , String (xs) , String (ys))	-> Number
r-value(animals-table, "pounds","weeks")		
# radial-star	:: (Num (points) , Num (outer) , Num (inner) , Str (fill-style) , Str (color))	-> Image
radial-star(6, 20, 50, "solid", "red")		
# 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")		
# rhombus	:: (Number (size) , Number (top-angle) , String (fill-style) , String (color))	-> Image
rhombus(100, 45, "outline", "pink")		
# 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"))		

Name	Domain	Range
# row-n	:: (Table (table-name) , Number (index))	-> Row
row-n(animals-table, 2)		
# scatter-plot	:: (Table (table-name) , String (labels) , String (xs) , String (ys))	-> Image
scatter-plot(animals-table, "name", "pounds","weeks")		
# sort	:: (Table (table-name) , String (column) , Boolean (ascending))	-> Table
sort(animals-table, "species", true)		
# sqr	:: (Number)	-> Number
sqr(4)		
# sqrt	:: (Number)	-> Number
sqrt(4)		
# square	:: (Number (size) , String (fill-style) , String (color))	-> Image
square(50, "solid", "red")		
# stacked-bar-chart	:: (Table (table-name) , String (group) , String (subgroup))	-> Image
stacked-bar-chart(animals-table, "species", "sex")		
# star	:: (Number (radius) , String (fill-style) , String (color))	-> Image
star(50, "solid", "red")		
# star-polygon	:: (Number (size) , Number (point-count) , Number (step-count) , String (fill-style) , String (color))	-> Image
star-polygon(100, 10, 3 ,"outline", "red")		
# stdev	:: (Table (table-name) , String (column))	-> Number
stdev(animals-table, "pounds")		
# string-contains	:: (String (haystack) , String (needle))	-> Boolean
string-contains("hotdog", "dog")		
# string-length	:: (String)	-> Number
string-length("rainbow")		
# text	:: (String (message) , Number (size) , String (color))	-> Image
text("Zari", 85, "orange")		
# triangle	:: (Number (size) , String (fill-style) , String (color))	-> Image
triangle(50, "solid", "fuchsia")		
# triangle-asa	:: (Number (top-left-angle) , Number (left-side) , Number (bottom-angle) , String (fill-style) , String (color))	-> Image
triangle-asa(90, 200, 10, "solid", "purple")		
# triangle-sas	:: (Number (bottom-R-side) , Number (top-R-angle) , Number (top-side) , String (fill-style) , String (color))	-> Image
triangle-sas(50, 20, 70, "outline", "dark-green")		
# vert-box-plot	:: (Table (table-name) , String (column))	-> Image
vert-box-plot(animals-table, "weeks")		

These materials were developed partly through support of the National Science Foundation (awards 1042210, 1535276, 1648684, and 1738598) and are licensed under a Creative Commons 4.0 Unported License. Based on a work at www.BootstrapWorld.org. Permissions beyond the scope of this license may be available by contacting contact@BootstrapWorld.org.

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.