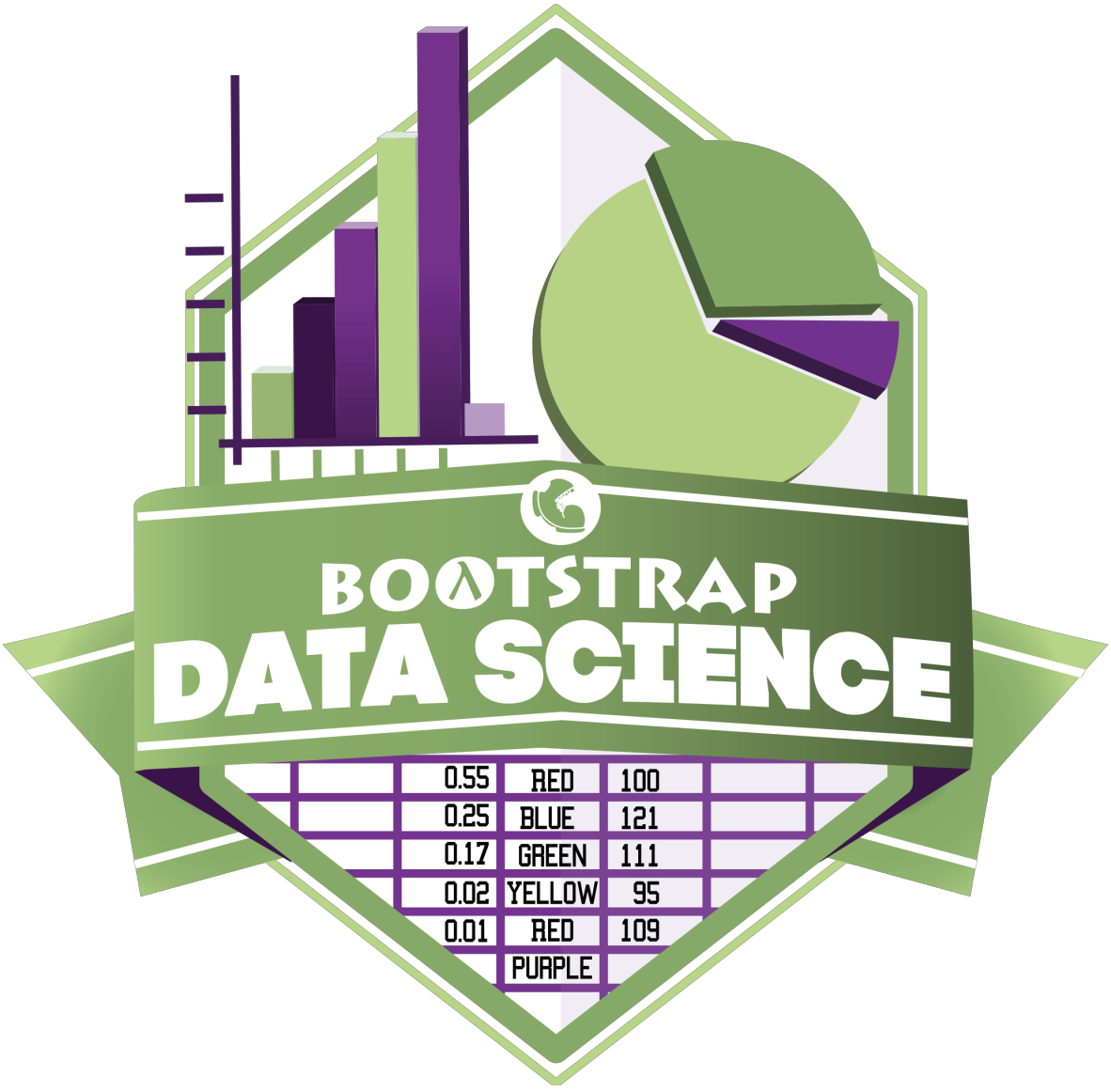


BOOTSTRAP
www.bootstrapworld.org

Workbook v1.4

Brought to you by the Bootstrap team:

- Emmanuel Schanzer
- Kathi Fisler
- Shriram Krishnamurthi
- Ed Campos
- Emma Youndtsmith
- Sam Dooman

Bootstrap is licensed under a Creative Commons 3.0 Unported License. Based on a work from www.BootstrapWorld.org. Permissions beyond the scope of this license may be available at schanzer@BootstrapWorld.org.

Unit 1

Many important questions ("what's the best restaurant in town?", "is this law good for citizens?", etc.) are answered with data. Data Scientists try and 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 data - usually numeric - that measures *quantity*, such as a person's height, a score on test, a measure of distance, etc. A list of quantitative data can be ordered from smallest to largest.
- **Categorical data** is data that specifies *categories*, such as 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.

Programming languages involves different *datatypes*, such as Numbers, Strings, Booleans and Images.

- **Operators** (like +, -, *, <, etc.) are written between values. For example: `4 + 2`
- We can use **functions** (like triangle, star, string-repeat, etc.) by writing the function name first, followed by a list of **arguments** in parentheses. For example: `star(50, "solid", "red")`
- **Methods** are special functions that are attached to pieces of data. We use them to manipulate Tables. They are different from functions in several ways:
  - Their names can't be used alone: they can only be used as part of data, separated by a dot. (For example, `shapes.row-n(2)`)
  - Their contracts are different: they include the type of the data as part of their names. (eg, `<table>.row-n :: (index :: Number) → Row`)
  - They have a "secret" argument, which is the data they are attached to
- In this course, we will use three **Table Methods** to manipulate our datasets:
  - `<Table>.order-by` – order the rows of a table based on a column
  - `<Table>.filter` – create a **subset** of the data, with only certain rows
  - `<Table>.build-column` – use the columns of a table to make a new one

Numbers and Strings

Make sure you've loaded the Unit 1 Starter File, and clicked "Run".

1. Try typing `42` into the Interactions Area and hitting "Enter". What happens?

2. Try typing in other Numbers. What happens if you try a decimal like `0.5`? A fraction like `1/3`? Try really big Numbers, and really small ones.

3. String values are always in quotes. Try typing your name (in quotes!). What happens when you hit "Enter"?

4. Try typing your name with the opening quote, but *without* the closing quote. What happens? Now try typing it without *any* quotes.

5. Is `42` the same as `"42"`? Why or why not? Write your answer below:

Operators

6. Just like in math, Pyret has *operators* like `+` and `-`. Try typing in `4 + 2`, and then `4+2` (without the spaces). What can you conclude from this? Write your answer below:

---

7. Typing in the following expressions, one at a time: `4 + 2 + 6`, `4 + 2 * 6`, and `4 + (2 * 6)`. What do you notice? Write your answer below:

---

8. Try typing in `4 + "cat"`, and then `"dog" + "cat"`. What can you conclude from this? Write your answer below:

Booleans

Boolean expressions are yes-or-no questions, and will always evaluate to either `true` ("yes") or `false` ("no"). What will each of the expressions below evaluate to? Write down the result in the blanks provided, and type them into Pyret if you're not sure.

`3 <= 4`	_____	`"a" > "b"`	_____
`3 == 2`	_____	`"a" <> "b"`	_____
`2 <> 4`	_____	`"a" == "b"`	_____
`3 <> 3`	_____	`"a" <> "a"`	_____

Boolean Operators

Pyret also has operators that work on *Booleans*. For each expression below, *write down your guess* about what it will evaluate to. Then type them in and see if you were right!

```
(3 <= 4) and (3 == 2)
```
_____

```
("a" == "b") and (3 <> 4)
```
_____

```
(3 <= 4) or (3 == 2)
```
_____

```
("a" == "b") or (3 <> 4)
```
_____

1. How many different Number values are there in Pyret? _____
2. How many different String values are there in Pyret? _____
3. How many different Boolean values are there in Pyret? _____

Lookups

The table below represents four shapes in a table:

shapes

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

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

`shapes.row-n(3)`	Evaluates to `4`
`shapes.row-n(0)`	Evaluates to the last row in the table
`shapes.row-n(1)["corners"]`	Evaluates to `"square"`
`shapes.row-n(2)["is-round"]`	Evaluates to `true`
`shapes.row-n(1)["name"]`	Evaluates to `12`
`shapes.row-n(0)["corners"]`	Evaluates to `false`
`shapes.row-n(3)["name"]=="circle"`	Evaluates to `3`
`shapes.row-n(2)["corners"] * 3`	Evaluates to the first row in the table

2. Fill in the blanks (left) with the Pyret lookup code that will produce the value (right).

a. *shapes.row-n(2)["name"]*	`"rectangle"`
b.	`"triangle"`
c.	`4`
d.	`0`
e.	`true`

Unit 2

Answering Questions from Data can take many forms. Here are a few types of questions, each requiring a different kind of analysis:

- **Lookup Questions** can be answered just by finding the right row and column a table. (e.g. – "How old is Toggle?")

- **Compute Questions** can be answered by computing over a single row or column. (e.g. – "What is the heaviest animal at the shelter?")

- **Analyze Questions** require looking for trends across multiple rows or columns. (e.g. – "Do cats tend to be adopted sooner than dogs?")

We can **define our own functions**, using a technique called the **Design Recipe**.

- We use the Design Recipe to help us define functions **without making mistakes**.

- The first step is to write a **Contract** and **Purpose Statement** for the function, which specify the Name, Domain and Range of the function and give a summary of what it does.

- The second step is to **write at least two examples**, which show how the function should work for specific inputs. These examples help us see patterns, and we express those patterns by **circling and labeling** what changes.

- The final step is to **define the function**, which generalizes our examples.

The Animals Dataset

1. This dataset is Animals from an animal shelter, which contains 31 data rows.

2. Some of the columns are:

   i. name, which contains categorical data, and is of type String. Some example values from this column are: "Toggle", "Fritz", and "Nori".

   ii. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

   iii. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

   iv. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

3. Some questions I have about this dataset:

My question is…	Lookup, Compute or Analyze?

Practicing Lookups

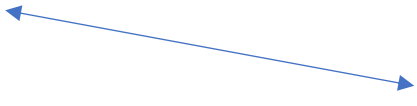
The table below represents four pets at an animal shelter, and four value definitions for rows in that table:

animals-table

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

```
animalA = animals-table.row-n(0)
animalB = animals-table.row-n(1)
animalC = animals-table.row-n(2)
animalD = animals-table.row-n(3)
```

v. Match each Pyret expression (left) to the description of what it looks up(right).



`animalD`	Evaluates to `51.6`
`animalA`	Evaluates to the last row in the table
`animalB["gender"]`	Evaluates to `"male"`
`animalC["age"]`	Evaluates to `"Toggle"`
`animalB["weight"]`	Evaluates to `92`
`animalA["name"]`	Evaluates to `6`
`animalD["weight"]`	Evaluates to "`female`"
`animalC["gender"]`	Evaluates to the first row in the table

vi. Fill in the blanks (left) with the Pyret lookup code that will produce the value (right).

animalD["name"]	`"Maple"`
	`"male"`
	`4`
	`48`
	`"Nori"`

The Design Recipe

For the word problems below, assume you have `animalA` and `animalB` defined in your code.

Define a function called `is-fixed`, which looks up whether or not an animal is fixed

```
# ___is-fixed___ :: ___(animal :: Row)___ → ___Boolean___
      name                domain                range
# Consumes an animal, and looks up the value in the fixed column
examples:

    ____________(_________) is ____________________________

    ____________(_________) is ____________________________
end

fun ____________ (________) : ____________________________

end
```

Define a function called `gender`, which consumes a Row of the animals table and looks up the gender of that animal

```
# ________________::________________________ → ________________
      name                domain                range
# ____________________________________________________________
examples:

    ____________(_________) is ____________________________

    ____________(_________) is ____________________________
end

fun ____________ (________) : ____________________________

end
```

The Design Recipe

For the word problems below, assume you have `animalA` and `animalB` defined in your code.

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

```
# is-cat :: (animal :: Row) → Boolean
  name      domain             range
# Consumes an animal, look up the species column, and computer if species = "cat"
examples:

    is-cat(animalA) is _____

    _____(_____) is _____
End

fun _____ (_____) : _____

end
```

Define a function called `is-young`, which consumes a Row of the animals table and computes whether it is less than four years old.

```
#  _____::_____ → _____
     name   domain   range
#  _____
examples:

    _____(_____) is _____

    _____(_____) is _____
end

fun _____ (_____) : _____

end
```

Unit 3

Functions can contain value definitions

We use **Table Plans** to help us use table methods correctly, without making mistakes:

- Like functions, we start with a Contract and Purpose Statement
- But instead of writing *programmed examples*, we sketch out **Sample Tables** and **Results**, based on the Contract and Purpose.
- Then we define the function based on our Sample Table and Result. Every function includes both the table definition (using methods) and a table expression.

Design Recipe

For the word problems below, assume you have `animalA` and `animalB` defined in your code.

Define a function called `birth-year`, which consumes a Row of the animals table and produces the year that animal was born.

```
# ________________::________________________ → ________________
      name                  domain                    range
# ____________________________________________________________
examples:

    ____________(________) is ____________________________

    ____________(________) is ____________________________
end

fun ____________ (________) : ____________________________

end
```

Define a function called `nametag`, prints out each animal's name in big red letters.

```
# ___nametag_____::___(animal :: Row)___ → ___Image___
      name                  domain                  range
# Consumes an animal, and produces an image of their name in big, red letters
examples:

    ___nametag___(___animalB___) is ____________________________

    ____________(________) is ____________________________
end

fun ____________ (________) : ____________________________

end
```

Playing with Methods

You have the following functions defined below (read them *carefully*!):

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

The table **t** below represents four animals at the shelter:

name	gender	age	fixed	weight
“Toggle”	“female”	3	true	48
“Fritz”	“male”	4	true	92
“Nori”	“female”	6	true	35.3
“Maple”	“female”	3	true	51.6

Match each Pyret expression (left) to the description of what it does (right).

```
t.order-by(“age”, true)
```

Produces a table containing *only* Toggle and Maple

```
t.filter(is-fixed)
```

Produces a table, sorted oldest-to-youngest.

```
t.build-column(“sticker”, nametag)
```

Produces a table, sorted youngest-to-oldest

```
t.filter(is-young)
```

Produces a table with an extra column, named “sticker”

```
t.order-by(“age”, false)
```

Produces a table containing Maple and Toggle, in that order.

```
t.filter(is-young)
 .order-by(“weight”, false)
```

Produces a table containing the same four animals.

```
t.order-by(“age”, true)
 .build-column(“label”, nametag)
```

Produces a table with an extra “label” column, sorted youngest-to-oldest

Table Plan

The shelter wants to print up bar charts showing young animal's ages, in alphabetical order. Sometimes they want to do this for *every* animal, but sometimes they just need it for the cats, or for animals that are fixed.
Define a function `sorted-age-bar`, which takes in a table of animals and computes a bar-chart showing their ages (in alphabetical order), *for only the young animals*.

Contract and Purpose

sorted-age-bar :: *(animals :: Table)* → *Image*

Consume a table of animals, and compute a bar chart showing their ages, in alphabetical order

Where I start, what I type, and what I get back

An example table to start with:

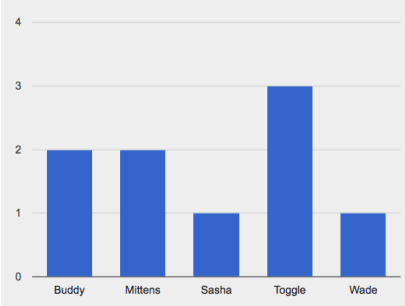
`example-table`

name	...	age
Sasha		1
Toggle		3
Buddy		2
Wade		1
Mittens		2

→

To use the function, I would type:

`sorted-age-bar(example-table)`



Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

```
fun sorted-age-bar ( animals ) :
  t = animals                                  Define the table
  .build-column(                              )  Are there more columns?
  .filter(                                    )  Are there fewer rows?
  .order-by(                                  )  Are the rows ordered?
                                                 Produce the result
end
```

Table Plan

The shelter wants to see if there's a relationship between how old an animal is, and how long it takes them to be adopted. Sometimes they want to do this for *every* animal, but sometimes they just need it for the cats, or for animals that are young. Define a function `age-adopted-scatter`, which takes in a table of animals and computes a scatter-plot showing only the fixed animals, with their ages on the x-axis and weeks to be adopted on the y-axis.

Contract and Purpose

age-adopted-scatter **::** *(animals :: Table)* → *Image*

Consume a table of animals, and compute a scatterplot for only the fixed animals, showing their ages on the x-axis,

Where I start, what I type, and what I get back

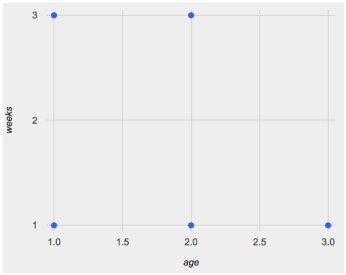
A sample table to start with:

name	...	age	weeks	fixed
Sasha		1	3	true
Toggle		3	1	true
Buddy		2	3	false
Wade		1	1	true
Mittens		2	1	true

→

To use the function, I would type:

`age-adopted-scatter(sample)`



Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

```
fun age-adopted-scatter ( animals ) :
    t = animals                                    Define the table
    .build-column(                                 Are there more columns?
    .filter(                                       Are there fewer rows?
    .order-by(                                     Are the rows ordered?
                                                   Produce the result
end
```

Unit 4

Bar charts - In bar charts, each bar has a height corresponding to the count or proportion of data values in a given category. Visually, we consider how heights of the bars compare to one another.

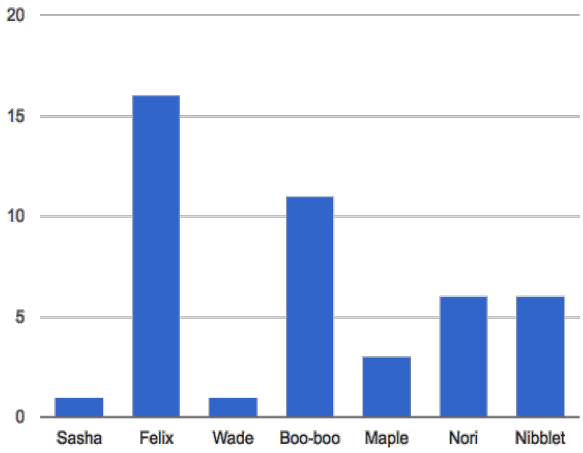
Pie charts - Pie charts show the relative proportion (or %) of a column's data values that fall into each category. The greater the proportion, the larger the pie slice. Visually, we consider how areas of the slices compare to one another, and to the whole area of 100%.

Choosing a Sample Table is important when coming up with small examples for Table Plans. A good sample table has:

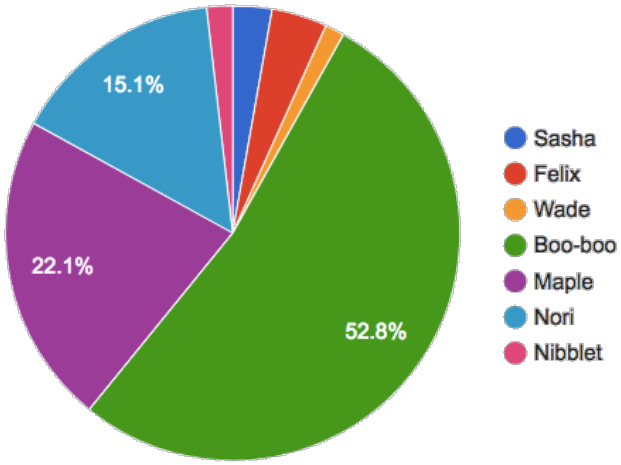
- At least all the relevant columns
- Enough rows to accurately represent the dataset
- Rows that aren't obviously presented in order

Quantity Charts in the `Animals` Dataset

Below are two **quantity charts** made from subsets of the animals table



Animals Ages (yrs)



Animals Weights (lbs)

What do you NOTICE about these charts?	What do you WONDER about these charts?

Why are some questions easier to answer with one kind of chart or another?

Bad Sample Tables!

For each word problem, a Sample Table must have (1) all the columns that matter, (2) a representative sample of the rows, and be in (3) random order. For each problem below, check the boxes if the Sample Table meets those criteria.

1. The shelter wants to a scatter plot showing the age of the cats v. their weight

name	species	age	fixed	legs	pounds	weeks
Sasha	cat	1	FALSE	4	6.5	3
Mittens	cat	2	TRUE	4	7.4	5
Sunflower	cat	5	TRUE	4	8.1	10

- ☐ Relevant columns
- ☐ Representative sample of rows
- ☐ Random order

2. The shelter wants a pie chart showing all the dogs' weight

name	species	age
Fritz	dog	4
Wade	cat	2
Nibblet	rabbit	6
Daisy	dog	5

- ☐ Relevant columns
- ☐ Representative sample of rows
- ☐ Random order

3. Sort all the animals alphabetically by name

name	species	age	fixed	legs	pounds	weeks
Ada	dog	2	TRUE	4	32	3
Bo	dog	4	TRUE	4	76.1	10
Boo-boo	dog	11	TRUE	4	123	10

- ☐ Relevant columns
- ☐ Representative sample of rows
- ☐ Random order

4. Make a bar chart for all the fixed animals

name	species	age	fixed	legs	pounds	weeks
Sasha	cat	1	FALSE	4	6.5	3

- ☐ Relevant columns
- ☐ Representative sample of rows
- ☐ Random order

Table Plan

Define a function `pie-pounds-young`, which takes in a Table of animals and creates a pie chart of the animals' weight, but only for animals that are young.

Contract and Purpose

 # pie-pounds-young **::** _____ → _____

 # Consumes a table of animals, filters to show only young animals, and produces a pie chart of their weight

Where I start, what I type, and what I get back

A sample table to start with: *To use the function, I would type:*

sample-table → *pie-pounds-young(sample-table)*

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

```
fun pie-pounds-young ( animals ) :
    t = animals
    .filter(is-young)
    pie-chart(t, "name", "pounds")
end
```

Define the table

Are there more columns?

Are there fewer rows?

Are the rows ordered?

Produce the result

My Dataset

1. This dataset is _____, which contains _____ data rows.

2. Some of the columns are:

   i. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

   ii. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

   iii. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

   iv. _____, which contains _____ data, and is of type _____. Some example values from this column are: _____.

3. Some questions I have about this dataset:

My question is…	**Lookup, Compute or Analyze?**

My Dataset

What are two ways you might want to *order* this dataset?

1) _____

2) _____

What are two subsets into which you might *filter* this dataset?

1) _____

2) _____

What are two new columns you might want to *build* from this dataset?

1) _____

2) _____

Design Recipes – Filtering Rows

What are two criteria you might want to *filter* by? Write your own word problems below, and solve them using the Design Recipe.

Define a function called _____ , which consumes a Row of the _____ table and _____

_____

```
# ______________::_______________________ → ______________
      name                domain                  range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (________) :  ______________________________

end
```

```
# ______________::_______________________ → ______________
      name                domain                  range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (________) :  ______________________________

end
```

Design Recipes – Building Columns

What are two columns you might want to *build* for your dataset? Write your own word problems below, and solve them using the Design Recipe.

```
# ________________::________________________ → ________________
      name                  domain                    range
# ____________________________________________________________
examples:

    ____________(________) is ____________________________

    ____________(________) is ____________________________
end

fun ____________ (_______) :  ____________________________

end
```

```
# ________________::________________________ → ________________
      name                  domain                    range
# ____________________________________________________________
examples:

    ____________(________) is ____________________________

    ____________(________) is ____________________________
end

fun ____________ (_______) :  ____________________________

end
```

Quantity Charts in My Dataset

Describe two of the pie or bar charts you made from your dataset.

1) I made a _____ chart, showing the _____ for
   (pie / bar) (column in your dataset)

   _____.
   (your subset (for example, "fixed dogs at the shelter"))

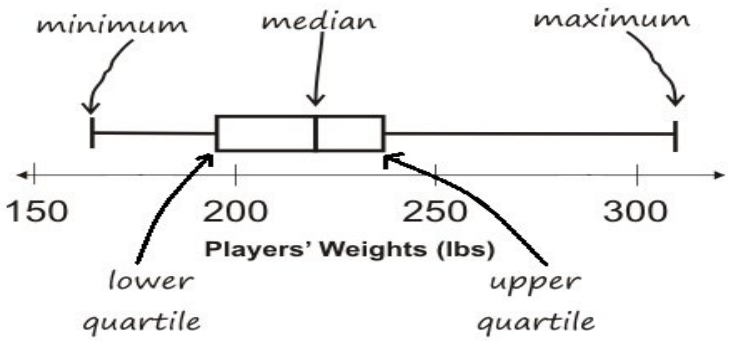
2) I made a _____ chart, showing the _____ for

   _____.

What do you NOTICE about these charts?	What do you WONDER about these charts?

Unit 5

- There are three ways to measure the "center" of a dataset, to summarize a whole column of data using just one number:
    - The **mean** of a dataset is the average of all the numbers
    - The **median** of a dataset is a value that is smaller than half the dataset, and larger than the other half
    - The **modes** of a dataset are the numbers that appear the most often.
- The **shape** of a dataset gives us an idea of which values are more or less common. In a *symmetric* data set, values are just as likely to occur a certain distance below the mean as above it. Outliers or **skew** can shift result in a mean that is higher than the mean (high outliers or right skew) or lower than the mean (low outliers or left skew).
- Data Scientists can also measure the "variation" or "spread" of a dataset using a **five number summary:**
    - The **minimum** – the smallest value in the dataset
    - The **first, or "lower" quartile (Q1)** – the middle of the smaller half of values, that separates the smallest quarter from the next smallest quarter
    - The **second quartile (Q2)** – the median value which separates the entire dataset into "top" and "bottom" halves.
    - The **third, or "upper" quartile (Q3)** – the middle of the larger half of values, that separates the second largest quarter from the largest quarter
    - The **maximum** – the largest value in the dataset
- The **five number summary** can be used to draw a **box-and-whisker plot.**



Summarizing Columns in `Animals`

2) The column I choose to measure is _____weeks_____

Measures of Center

The three measures for this column are:

Mean (Average)	Median	Mode(s)

3) Since the mean is _____ than the median, this suggests that there may
[higher/lower]

be outliers or skewness due to values that are unusually _____.
[high / low]

Measures of Variation

My five-number summary is:

Minimum	Q1	Q2 (Median)	Q3	Maximum

A box plot can be drawn from this summary on the number line below:

From this summary and box-plot, I conclude:

Interpreting Variation

Consider the following dataset, representing the annual income of ten people:

`$65k, $12k, $14k, $280k, $15k, $22k, $45k, $34k, $45k, $175k`

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

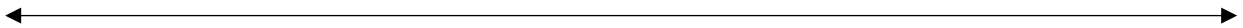
2. In the table below, compute the **measures of center** for this dataset.

Mean (Average)	Median	Mode(s)

3. In the table below, compute the **five number summary** of this dataset.

Minimum	Q1	Q2 (Median)	Q3	Maximum

4. On the number line below, draw a **box plot** for this dataset.



5. The following statements are *correct*…but misleading. Write down the reason why.

Statement	Why it's misleading
"They're rich! The average person makes more than $70k dollars!"	
"It's a middle-income list: the most common salary is $45k/yr!"	
"This group is really diverse, with people making as little as 12k and as much as $280k!"	

Table Plan

The Animal Shelter Bureau would like to study the distribution of weeks-until-adoption for fixed animals housed at shelters around the country. They need a function that consumes a table of animals, filters to show only the fixed animals, and produces a box-plot for the `weeks` column. Define a function called `fixed-weeks-box` below.

Contract and Purpose

_____ : : _____ → _____

_____

Where I start, what I type, and what I get back

A sample table to start with: *To use the function, I would type:*

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

fun _____ (_____) :

t = _____ *Define the table*

_____ *Are there more columns?*

_____ *Are there fewer rows?*

_____ *Are the rows ordered?*

_____ *Produce the result*

end

Summarizing a Column in My Dataset

1) The column I choose to measure is _____

Measures of Center

The three measures for this column are:

Mean (Average)	Median	Mode(s)

2) Since the mean is _____ than the median, this suggests that there may
[higher/lower]

be outliers or skewness due to values that are unusually _____.
[high / low]

Measures of Variation

My five-number summary is:

Minimum	Q1	Q2 (Median)	Q3	Maximum

A box plot can be drawn from this summary on the number line below:

From this summary and box-plot, I conclude:

Unit 6

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

- *Bar charts provide a visual representation of the frequency of values in a* ***categorical*** *column.*

- Often there is no defined way to order these bars, but sometimes there is a natural progression that makes sense. For example, bars for T-Shirt sizes might be presented in order of S, M, L, and XL.

Histograms show the number of rows that fall within certain ranges, or "bins" of a dataset. The more rows that that fall within a particular "bin", the taller the bar.

- *Histograms provide a visual representation of the frequency of values in a* ***quantitative*** *column.*

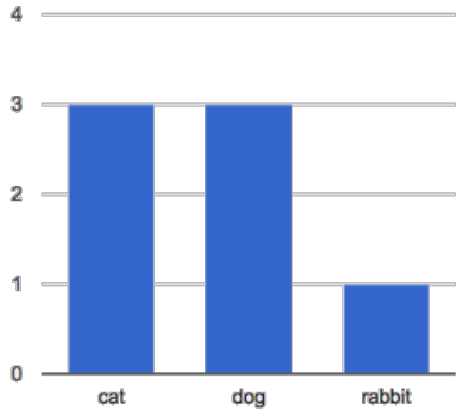
- Quantitative data can be ordered, so the bars of a histogram always progress from smallest (on the left) to largest (on the right).

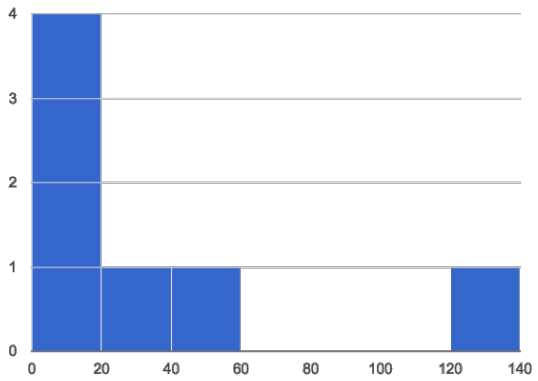
- When dealing with histograms, it's important to select a good **bin size**. If the bins are too small or too large, it is difficult to see the distribution in the dataset.

Bar Charts & Histograms in the `Animals` Dataset

name	species	age	pounds
"Sasha"	"cat"	1	6.5
"Boo-boo"	"dog"	11	123
"Felix"	"cat"	16	9.2
"Nori"	"dog"	6	35.3
"Wade"	"cat"	1	3.2
"Nibblet"	"rabbit"	6	4.3
"Maple"	"dog"	3	51.6

1. How many cats are there? _____
2. How many dogs are there? _____
3. How many animals are between 3-6 years old? _____
4. How many animals weigh between 0-5 pounds? _____
5. Are there more animals weighing 0-5 than 6-10 pounds? _____
6. The charts below are based on the Sample Table above. For each chart: is it measuring quantitative or categorical data?





_____ _____

_____ _____

_____ _____

_____ _____

Table Plan

Define a function `freq-bar-gender`, which takes in a Table of animals and creates a frequency bar chart showing how many *fixed* animals are male v. female.

Contract and Purpose

_____ : : _____ → _____

_____

Where I start, what I type, and what I get back

A sample table to start with: *To use the function, I would type:*

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

fun _____ (_____) :

t = _____

_____

_____

_____

_____

end

Define the table

Are there more columns?

Are there fewer rows?

Are the rows ordered?

Produce the result

Table Plan

Define a function `histogram-cats-adoption`, which takes in a Table of animals and creates a histogram showing how long it took for cats in the dataset to get adopted

Contract and Purpose

_____ : : _____ → _____

_____

Where I start, what I type, and what I get back

A sample table to start with: *To use the function, I would type:*

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

fun _____ (_____) :

t = _____

_____

_____

_____

_____

end

Define the table

Are there more columns?

Are there fewer rows?

Are the rows ordered?

Produce the result

Visualizing My Dataset

Describe two of the histograms or frequency bar charts you made from your dataset.

1) I made a _____, showing the _____ for
histogram / frequency bar chart — column in your dataset

_____.
your subset (for example, "fixed dogs at the shelter"

2) I made a _____, showing the _____ for

_____.

What do you NOTICE about these charts?	What do you WONDER about these charts?

Matching Charts to Questions

For each of the questions below, draw a line to the chart that will best answer it. (You may find that more than one question is best answered by the same chart!)

1. Are there more of the animals at the shelter fixed or unfixed?	
2. How many weeks did each cat wait to be adopted?	**Pie Chart**
3. How many male v. female dogs are there?	
4. How many animals have 4 legs? 8? 3?	**Bar Chart**
5. What percent of the total weight at the shelter is made up by Boo-boo?	
6. What is the distribution of weights across all the animals older than 3?	**Histogram**
7. How many animals are there of each species?	
8. Who waited the longest to be adopted?	

Unit 7

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

- If the points in a scatter plot appear to follow a straight line, it is possible that a linear relationship exists between those two columns. A number called a **correlation** can be used to summarize this relationship.

- The correlation is **positive** if the point cloud slopes up as it goes farther to the right. It is **negative** if it slopes down as it goes farther to the right. The points are tightly clustered around a line, it is a **strong** correlation. If they are loosely scattered, it is a **weak** correlation.

- If there is a pattern to the points in a scatter plot, points that are far away from the pattern are called **outliers**.

- We can graph this relationship by drawing a straight line through the data cloud, so that the vertical distance between the line and each of the points is as small as possible. This line is called the **line of best fit** and allows us to predict y-values based on x-values.

(Dis)Proving a Claim

"Younger animals are cuter, so they get adopted faster."

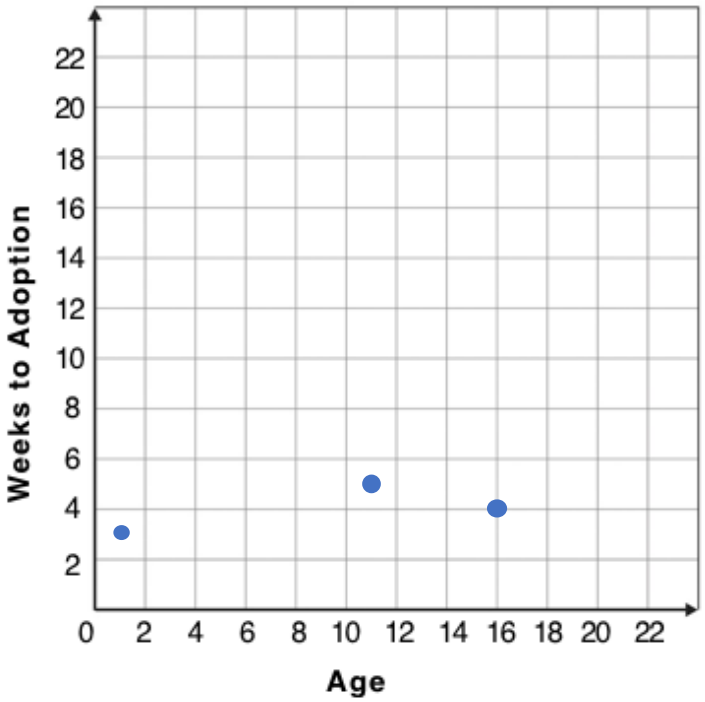
Do you agree? If so, why?

I hypothesize…

What would you look for in the dataset to see if you are right?

Creating a Scatter Plot

name	species	age	weeks
"Sasha"	"cat"	1	3
"Boo-boo"	"dog"	11	5
"Felix"	"cat"	16	4
"Buddy"	"lizard"	2	24
"Nori"	"dog"	6	9
"Wade"	"cat"	1	2
"Nibblet"	"rabbit"	6	12
"Maple"	"dog"	3	2



1. **For each row in the Sample Table on the left, add a point to the scatter plot on the right**. The first 3 rows have been completed for you. Use the values from the `age` column for the x-axis, and values from the `weeks` column for the y-axis.

2. Do you see a pattern? Do the points seem to shift up or down as age increases? **Draw a line on the scatter plot to show this pattern**.

3. Does the line slope upwards or downwards? _____

4. Are the points clustered around the line? Loosely scattered? _____

Table Plan

Define a function `cats-age-weeks`, which takes in a Table of animals and creates a scatter plot of all the cats, tracking their `age` on the x-axis and the number of `weeks` it took for them to be adopted on the y-axis.

Contract and Purpose

_____ : : _____ → _____

_____

Examples Where I start, what I type, and what I get back

A sample table to start with: *To use the function, I would type:*

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

fun _____ (_____) :

  t = _____

  _____

  _____

  _____

  _____

end

Define the table

Are there more columns?

Are there fewer rows?

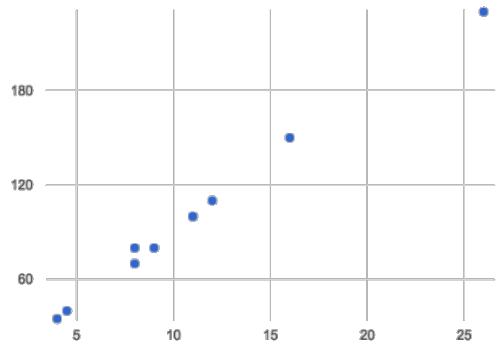
Are the rows ordered?

Produce the result

Drawing Predictors

For each of the scatter plots below, draw a **predictor line** that fits best.

A

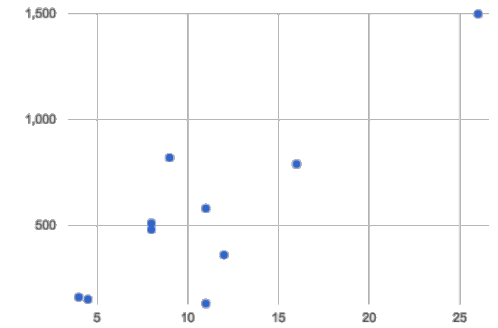


fat (g) v. calories-from-fat in common menu items

Direction: Positive Negative None

Strength: Strong Weak

B

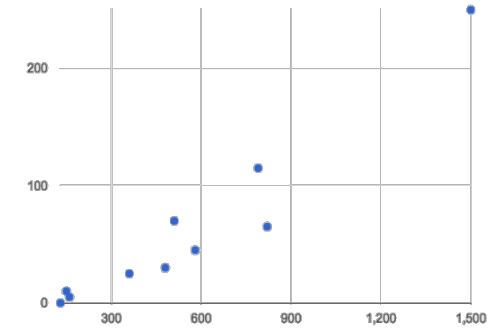


fat (g) v. sodium (g) in common menu items

Direction: Positive Negative None

Strength: Strong Weak

C

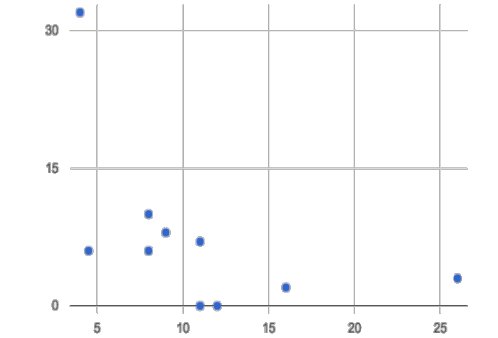


sodium (g) v. cholesterol (mg) in common menu items

Direction: Positive Negative None

Strength: Strong Weak

D



fat (g) v. sugar (g) in common menu items

Direction: Positive Negative None

Strength: Strong Weak

Correlations in My Dataset

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

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

correlation, because _____

_____. It might be stronger if I looked

at _____.
a subset or extension of my data

---

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

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

correlation, because _____

_____. It might be stronger if I looked

at _____.
a subset or extension of my data

---

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

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

correlation, because _____

_____. It might be stronger if I looked

at _____.
a subset or extension of my data

Unit 8

- **Linear Regression** is a way of computing the line of best fit, which minimizes the sum of squared vertical distances of all scatter plot points from the line. Calculating the slope and intercept of this line is a task best left to computing or statistical software. Slope provides us with the easiest summary to grasp: it's how much we predict the y-variable to increase, for each unit that the x-variable increases

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

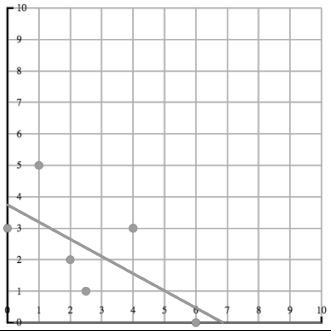
- **Sample size matters!** The number of data values is also relevant. We'd be more convinced of a positive relationship in general between cat age and time to adoption if a correlation of +0.57 were based on 50 cats instead of 5.

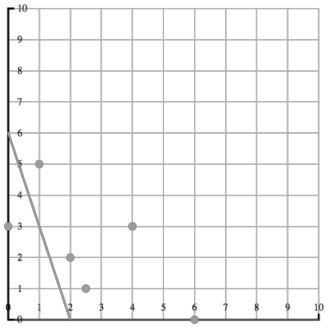
Grading Predictors

Below are the scatter plots for data sets A-D, with two different predictor lines drawn on top. For plots A-D:

1. Circle the plot with the line that fits better
2. Give the plot you circled a grade between 0 (no correlation) and 1 (perfect correlation)

A

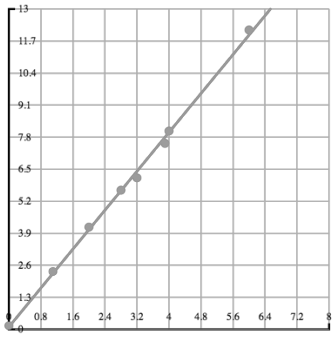


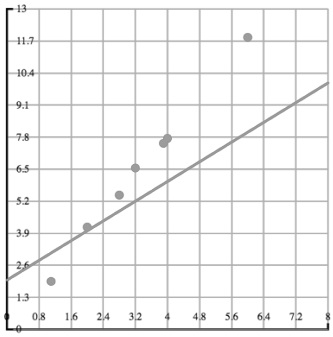


Strength of Correlation:

_____

B

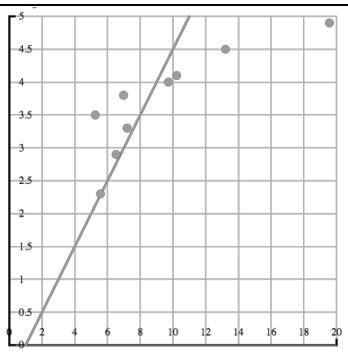


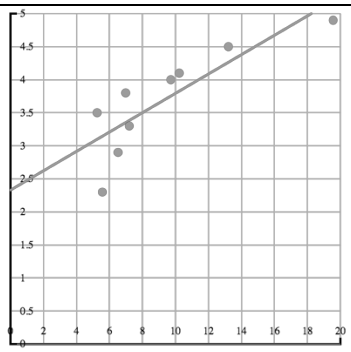


Strength of Correlation:

_____

C

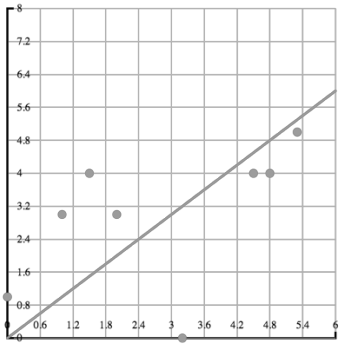


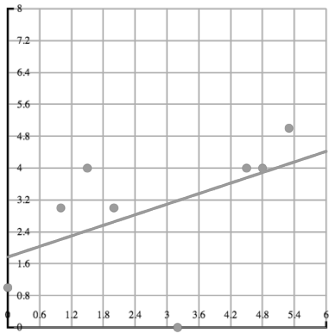


Strength of Correlation:

_____

D





Strength of Correlation:

_____

Regression Analysis in the `animals` Dataset

I performed a linear regression on _____cats at the shelter_____, and
dataset or subset

found ___a weak (r^2=0.321), positive___ correlation between
a strong/weak (r^2=__), positive/negative

__age of the cats (in weeks)__ and __number of weeks to adoption__. From this, I
[x-axis] [y-axis]

conclude that ____32.1% of the variability in adoption time is explained by the_____
r^2 % of the variation in [y-axis] is explained by [x-axis]

__age of the cat__. I would predict that a 1 ____year_____ increase in
[x-axis units]

_____age_____ is associated with a ____0.23 week____ ____increase____ in __adoption time__.
[x-axis] [slope, y-units] [increase/decrease] [y-axis]

I performed a linear regression on _____, and
dataset or subset

found _____ correlation between
a strong/weak (r^2=__), positive/negative

_____ and _____. From this, I
[x-axis] [y-axis]

conclude that _____
r^2 % of the variation in [y-axis] is explained by [x-axis]

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

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

I performed a linear regression on _____, and
dataset or subset

found _____ correlation between
a strong/weak (r^2=__), positive/negative

_____ and _____. From this, I
[x-axis] [y-axis]

conclude that _____
r^2 % of the variation in [y-axis] is explained by [x-axis]

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

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

Regression Analysis in My Dataset

I performed a linear regression on _____, and
dataset or subset

found _____ correlation between
a strong/weak (r^2=__), positive/negative

_____ and _____. From this, I
[x-axis] [y-axis]

conclude that _____
r^2 % of the variation in [y-axis] is explained by [x-axis]

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

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

I performed a linear regression on _____, and
dataset or subset

found _____ correlation between
a strong/weak (r^2=__), positive/negative

_____ and _____. From this, I
[x-axis] [y-axis]

conclude that _____
r^2 % of the variation in [y-axis] is explained by [x-axis]

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

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

I performed a linear regression on _____, and
dataset or subset

found _____ correlation between
a strong/weak (r^2=__), positive/negative

_____ and _____. From this, I
[x-axis] [y-axis]

conclude that _____
r^2 % of the variation in [y-axis] is explained by [x-axis]

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

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

Unit 9

Threats to Validity can undermine a conclusion, even if the analysis was done correctly. Some examples of threats are:

- **Selection bias** – identifying the favorite food of the rabbits won't tell us anything reliable about what all the animals eat.

- **Sample size** – averaging the age of only three animals won't tell us anything reliable about the age of animals at the shelter!

- **Sample error** – surveying dogs when they are puppies won't tell us anything reliable about overall dog behavior, since their behavior changes as they age.

- **Confounding variables** – shelter workers might steer people towards newer animals, because they've become attached to the animals that have been there for a while, making it *appear* that "staying at the shelter longer" means "less likely to be adopted".

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, and from this survey they concluded that dogs are the most popular pet in the region.

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

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

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

Threats to Validity

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

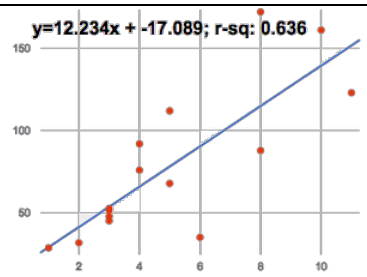
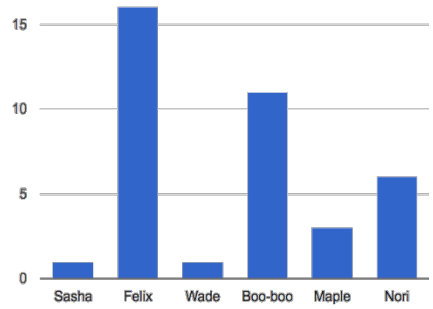
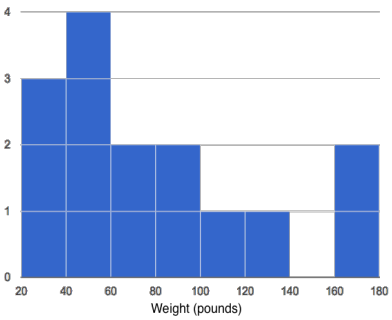
What are some possible threats to the validity of this conclusion?

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 then the others.

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

Fake News!

Every claim below is *wrong*! Your job is to figure out why, by looking at the data.

	Data	Claim	Why it's *wrong*
1	The average player on a basketball team is 6'1".	*"Most of the players on the team are taller than 6'."*	
2	After performing linear regression on census data, a positive correlation (r^2=0.18) was found between people's height and salary.	*"Taller people get paid more."*	
3		*"According to the predictor function indicated here, the value on the x-axis is will predict the value on the y-axis 63.6% of the time."*	
4	 Bar Chart of Pet Ages	*"According to this bar chart, Felix makes up a little more than 15% of the total ages of all the animals in the dataset."*	
5		*"According to this histogram, most animals weigh between 40 and 60 pounds."*	
6	After performing linear regression, a negative correlation (r^2=0.91) was found between the number of hairs on a person's head and their likelihood of owning a wig.	*"Owning wigs causes people to go bald."*	

Blank Recipes, Table Plans, and References

Design Recipes

```
# ______________::_______________________ → ______________
      name                domain                  range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (_______) :  ______________________________

end
```

```
# ______________::_______________________ → ______________
      name                domain                  range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (_______) :  ______________________________

end
```

Design Recipes

```
# ______________::_______________________ → ______________
      name                 domain                 range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (_______) :  ______________________________

end
```

```
# ______________::_______________________ → ______________
      name                 domain                 range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (_______) :  ______________________________

end
```

Design Recipes

```
# ______________::_______________________ → ______________
      name                 domain                 range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (_______) :  ______________________________

end
```

```
# ______________::_______________________ → ______________
      name                 domain                 range
# ____________________________________________________________
examples:

    ____________(________) is ______________________________

    ____________(________) is ______________________________
end

fun ____________ (_______) :  ______________________________

end
```

Table Plan

Contract and Purpose

_____ : : _____ → _____

_____

Examples

Make a Start Table and a result based on that table.

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

```
fun _____ (_____) :
   t = _____
   _____
   _____
   _____
   _____
end
```

Define the table

Are there more columns?

Are there fewer rows?

Are the rows ordered?

Produce the result

Table Plan

Contract and Purpose

_____ : : _____ → _____

_____

Examples

Make a Start Table and a result based on that table.

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

```
fun _____ (_____) :
  t = _____
      _____
      _____
      _____
      _____
end
```

Define the table

Are there more columns?

Are there fewer rows?

Are the rows ordered?

Produce the result

Table Plan

Contract and Purpose

_____ : : _____ → _____

_____

Examples

Make a Start Table and a result based on that table.

_____ → _____

Define the function

Use the relevant methods (circle your helper functions!), then produce a result with the new table.

```
fun _____ (_____) :
  t = _____
      _____
      _____
      _____
      _____
end
```

Define the table

Are there more columns?

Are there fewer rows?

Are the rows ordered?

Produce the result

Contracts

Contracts tell us how to use a function. For example: `num-sqr :: (n :: Number) → Number` tells us that the name of the function is **`num-sqr`**, that it takes one input (a **`Number`**), and that it evaluates to a number. From the contract, we know **`num-sqr(4)`** will evaluate to a **`Number`**.

Name	Domain	Range
`triangle`	`:: (side-length :: Number, style :: String, color :: String) →`	`Image`
`circle`	`:: (radius :: Number, style :: String, color :: String) →`	`Image`
`star`	`:: (radius :: Number, style :: String, color :: String) →`	`Image`
`rectangle`	`:: (width :: Num, height :: Num, style :: Str, color :: Str) →`	`Image`
`ellipse`	`:: (width :: Num, height :: Num, style :: Str, color :: Str) →`	`Image`
`square`	`:: (size-length :: Number, style :: String, color :: String) →`	`Image`
`text`	`:: (str :: String, size :: Number, color :: String) →`	`Image`
`overlay`	`:: (img1 :: Image, img2 :: Image) →`	`Image`
`rotate`	`:: (degree :: Number, img :: Image) →`	`Image`
`scale`	`:: (factor :: Number, img :: Image) →`	`Image`
`string-repeat`	`:: (text :: String, repeat :: Number) →`	`String`
`string-contains`	`:: (text :: String, search-for :: String) →`	`Boolean`
`num-sqr`	`:: (n :: Number) →`	`Number`
`num-sqrt`	`:: (n :: Number) →`	`Number`
`num-min`	`:: (a :: Number, b:: Number) →`	`Number`
`num-max`	`:: (a :: Number, b:: Number) →`	`Number`

Contracts

Contracts tell us how to use a function. For example: `<Table>.filter :: (test :: (Row→Boolean) → Row` tells us that the name of the function is **`.filter`** and that it is a **`Table`** method. The domain says it one input (a function that comsumes **`Rows`** and produces **`Booleans`**), and that the method evaluates to a Table. From the contract, we know **`animals-table.filter(is-cat)`** will evaluate to a **`Table`**.

Name	Domain		Range
<Table>.row-n	:: (n :: *Number*)	→	*Row*
<Table>.order-by	:: (col :: *String*, *increasing* :: *Boolean*)	→	*Table*
<Table>.filter	:: (test :: *(Row → Boolean)*)	→	*Table*
<Table>.build-column	:: (col :: *String*, *builder* :: *(Row → Value)*)	→	*Table*
mean	:: (t :: *Table*, col :: *String*)	→	*Number*
median	:: (t :: *Table*, col :: *String*)	→	*Number*
modes	:: (t :: *Table*, col :: *String*)	→	*List<Number>*
bar-chart	:: (t :: *Table*, labels :: *String*, values :: *String*)	→	*Image*
pie-chart	:: (t :: *Table*, labels :: *String*, values :: *String*)	→	*Image*
box-plot	:: (t :: *Table*, col:: *String*)	→	*Image*
freq-bar-chart	:: (t :: *Table*, values :: *String*)	→	*Image*
histogram	:: (t :: *Table*, values :: *String*, bin-width :: *Number*)	→	*Image*
scatter-plot	:: (t :: *Table*, labels :: *String*, xs :: *String*, ys :: *String*)	→	*Image*
lr-plot	:: (t :: *Table*, labels :: *String*, xs :: *String*, ys :: *String*)	→	*Image*