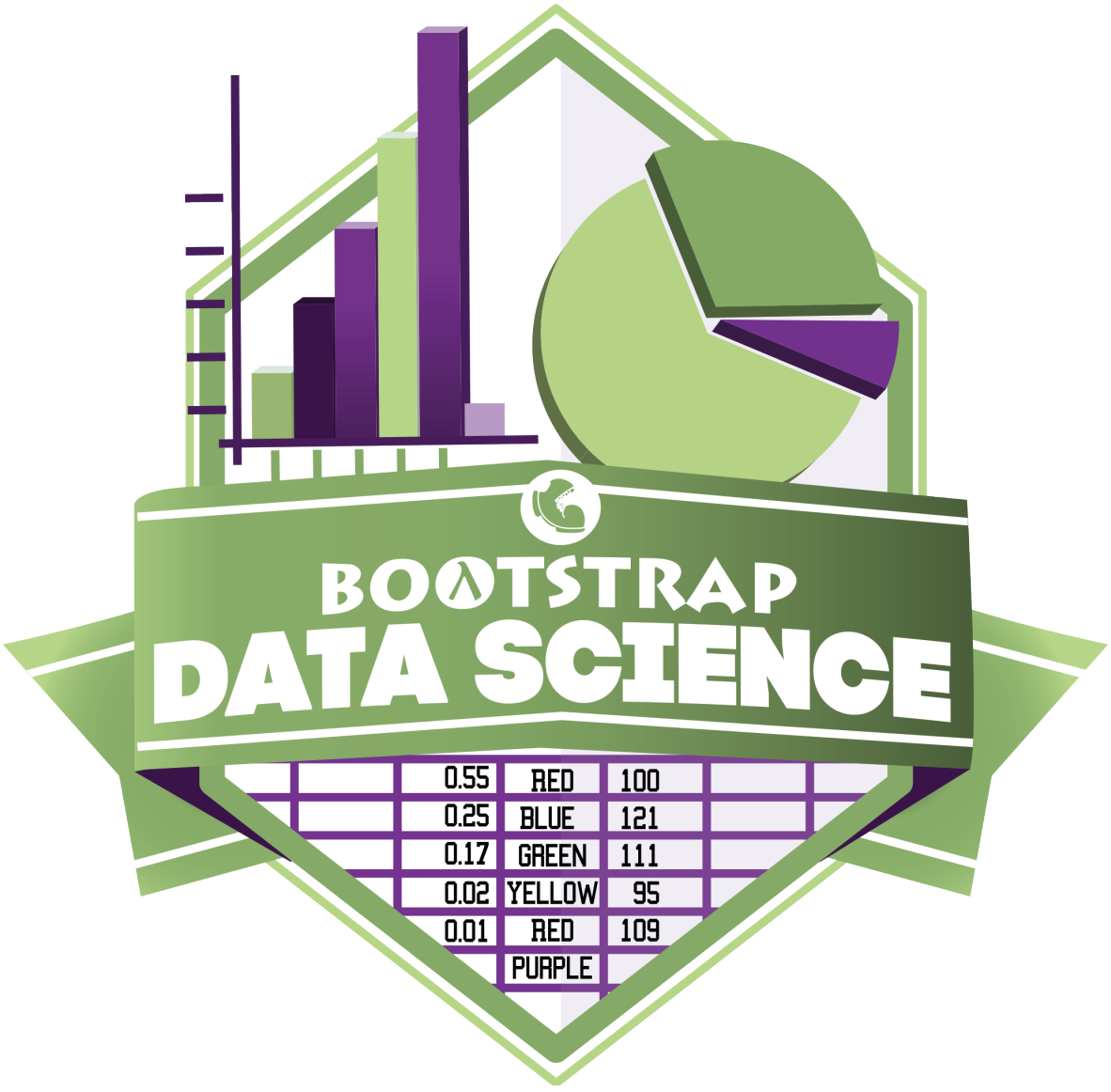


BOOTSTRAP
www.bootstrapworld.org

Workbook v1.1

Brought to you by the Bootstrap team:

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

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 of 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. A list of categorical data has no notion of "smallest" or "largest", and cannot be ordered.

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

- **Operators** (like +, -, *, <, etc.) are written between values. For example: `4 + 2`

- **Functions** (like triangle, star, string-repeat, etc.) are written first, followed by a list of **arguments** in parentheses. For example: `star(50, "solid", "red")`

- **Examples** help programmers reason about their code. Every example contains two expressions, and the example "passes" if both expressions evaluate to the same thing. For example: `4 + 2 is 6`, or `"cat" == "dog" is false`

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:

They are different data types: 42 (without quotes) is a Number, and "42" (with quotes) is a string.

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:

Operators (like +) need whitespace separating them from their operands.

7. Try typing in `4 + 2 + 6`, `4 + 2 * 6`, and `4 + (2 * 6)`. What can you conclude from this? Write your answer below:

You can use the same operator multiple times without parentheses, but you need parentheses to group order of operations if using different operators (like + and *) together.

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

The + operator can only be used with Numbers, not Strings.

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`	True	`"a" > "b"`	False
`3 == 2`	False	`"a" <> "b"`	True
`2 <> 4`	True	`"a" == "b"`	False
`3 <> 3`	True	`"a" <> "a"`	False

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)` False

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

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

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

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

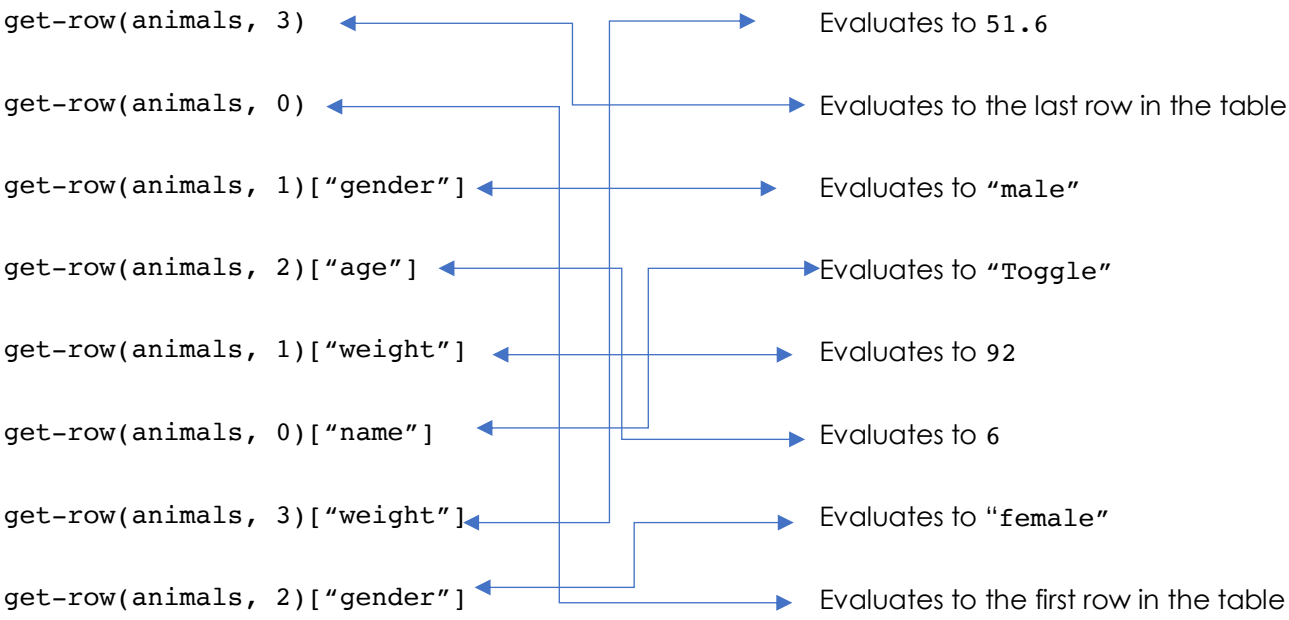
Lookups

The table below represents four animals at the shelter:

animals

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

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



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

a. *get-row(animals, 3)["name"]* — "Maple"

b. get-row(animals, 1)["gender"] — "male"

c. get-row(animals, 1)["age"] — 4

d. get-row(animals, 0)["weight"] — 48

e. get-row(animals, 2)["name"] — "Nori"

Writing Examples

1. In the examples block below, **put an "X" next to the examples that will *fail*.** Remember: examples only pass if the left- and right-hand expressions evaluate to the same thing!

```
examples:
  1 + 2 + 9                      is 19      ✖
  num-sqrt(16)                   is 2 + 2
  3 > 99                         is true    ✖
  square(10, "solid", "red")     is rectangle(10, 10, "solid", "red")
end
```

2. In the examples block below, **fill in the blank on the right-hand side so the example will *pass*.**

```
examples:
  string-repeat("yeah! ", 3)      is "yeah! yeah! yeah! "
  string-contains("Maya", "May")  is true
  "apples" <> "oranges"           is true
end
```

3. The examples block below refers to the `shapes` table on the right, using row-accessors and the `get-row` function. For each example, **fill in the blank so the example will *pass*.**

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

```
examples:
  get-row(shapes, 3)["name"]      is "square"
  get-row(shapes, 0)["corners"]   is 3
  get-row(shapes, 2)["is-round"]  is true
end
```

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?")

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** – if they person surveying the animals has a piece of bacon in their pocket, they will incorrectly find that all dogs are friendly!

The Animals Dataset

1. This dataset is Animals from an animal shelter

2. Some of the columns are….
   (For each column in this dataset, fill out the datatype, and whether it contains Qualitative or Categorical data in the table below)

Headers	Name	species	age	fixed	pounds
Sample 1	"Sasha"	"cat"	1	false	6.5
Sample 2	"Nori"	"dog"	6	true	35.3
Sample 3	"Snowcone"	"cat"	2	true	6.1
Datatype	String	String	Number	Boolean	Number
Quantitative or Categorical?	C	C	Q	C	Q

3. For the questions below, check the box to the left of the questions you CAN answer given this dataset:

	Question?
✓	*How old is Nori?*
	What color is Snowcone's fur?
✓	*What is the average age of the animals in the table?*
✓	*Are there more fixed or unfixed animals in the table?*
	Are families with children more likely to adopt kittens?

4. Some questions I have about this dataset:

Do younger animals get adopted faster?

What is the average weight of all the animals in the shelter?

Are male or female animals more likely to get adopted?

What Questions Can You Answer?

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

date	miles	time	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
06/22/2018	15	61	"cloudy"	13	28
07/04/2018	24	103	"sunny"	14	26
07/12/2018	24	120	"windy"	12.5	26

What <u>can</u> you answer? For each of the following questions, check the box to the left of questions you <u>can</u> answer. For each *checked* question, write whether the question is a **lookup**, **compute**, or **analyze** question.

	Question?	**Category?**
✓	*What is the cyclist's average speed across all rides?*	Compute
✓	*How many miles did they ride in June?*	Compute
	What is the tallest hill this cyclist climbed?	
✓	*Does this cyclist ride slower when it is rainy?*	Analyze
	Does this cyclist ride faster when they are late to an appointment?	

What <u>*can't*</u> you answer? For each of the following questions, check the box to the left of questions you <u>cannot</u> answer. For each *un-checked* question, write whether the question is a **lookup**, **compute**, or **analyze** question.

	Question?	**Category?**
✓	*What tire pressure produces the highest avg speed?*	
	What is the avg time it takes this cyclist to ride 1mi?	Compute
	Does this cyclist ride more in April or July?	Compute
✓	*What is the average temperature while this cyclist is riding?*	
✓	*How many flat tires did this cyclist fix in June?*	

Threats to Validity

Some volunteers from the animal shelter surveyed a group of pet owners at a local 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?

Not many people are likely to walk their cats at the park, so if the volunteers only surveyed pet owners at the park, dogs are likely to be more highly represented in their sampling.

The animal shelter noticed a large increase in pet adoptions between Thanksgiving and New Year's Eve. They conclude that at this current rate, there will be a huge demand for pets between January and April.

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

Lots of people may be adopting animals during the holiday season, so these past patterns are unlikely to predict future patterns in adoption rates.

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 random sample may not be representative of the whole group of pets. In this case, there are many more dogs and cats than spiders and rabbits at the shelter, so using this random sample to draw conclusions about the whole group is wrong!

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?

There may be other reasons the dogs are happier at mid-day than morning and evening- for instance, mid-day is when they eat lunch, which is likely to make the dogs very excited!

My Dataset

1. My dataset is [specific to each student]

2. Some of my columns are….
   (Copy six columns from your dataset, and for each column write its datatype, and whether it contains Qualitative or Categorical data in the table below)

Headers				
Sample 1				
Sample 2				
Sample 3				
Datatype				
Quantitative or Categorical?				

3. Some questions I have about this dataset:

4. What are some possible threats to validity you may encounter in your analysis?

Unit 3

- Programming languages let us **define our own function**.

- 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 Design Recipe

Define a function called `is-fixed`, which tells us whether or not an animal is fixed

```
# is-fixed :: (animal :: Row) → Boolean
  name       domain             range
# Consumes an animal, and produces the value in the fixed column
```

examples:

```
is-fixed ( sasha ) is sasha["fixed"]
is-fixed ( felix ) is felix["fixed"]
end

fun is-fixed ( animal ) : animal["fixed"]
end
```

Define a function called `gender`, which consumes a Row of the animals table tells us the gender of that animal

```
# gender :: (animal :: Row) → String
  name     domain             range
# Consumes an animal, and produces the value in the gender column
```

examples:

```
gender ( snowcone ) is snowcone["gender"]
gender ( toggle ) is toggle["gender"]
end

fun gender ( animal ) : animal["gender"]
end
```

Define a function called `is-cat`, which consumes a Row of the animals table and produces `true` if it's a cat.

```
# is-cat :: (animal :: Row) → Boolean
  name      domain             range
# Consumes an animal, and return true if the species is "cat"
```

examples:

```
is-cat ( sasha ) is sasha["species"] == "cat"
is-cat ( snuggles ) is snuggles["species"] == "cat"
end

fun is-cat ( animal ) : animal["species"] == "cat"

end
```

Define a function called `is-young`, which consumes a Row of the animals table and produces `true` if it's an animal that is less than two years old.

```
# is-young :: (animal :: Row) → Boolean
  name       domain             range
# Consumes an animal, returns true if the animal is less than 2 years old
```

examples:

```
is-young ( wade ) is wade["age"] < 2
is-young ( sheba ) is sheba["age"] < 2
end

fun is-young ( animal ) : animal["age"] < 2

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 ( sasha ) is text(sasha["name"], 50, "red")
nametag ( felix ) is text(felix["name"], 50, "red")
end

fun nametag ( animal ) : text(animal["name"], 50, "red")
end
```

Define a function called `is-kitten`, which consumes a Row of the animals table and produces true if it's a cat younger than two years old.

```
# is-kitten :: (animal :: Row) → Boolean
#   name        domain          range
# Consumes an animal, returns true if it's a cat less than two years old
```

examples:

```
is-kitten ( snuggles ) is (snuggles["species"] == "cat") and (snuggles["age"] < 2)
is-kitten ( wade ) is (wade["species"] == "cat") and (wade["age"] < 2)
end

fun is-kitten ( animal ) : (animal["species"] == "cat") and (animal["age"] < 2)
end
```

Unit 4

- **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, `animals.order-by`)
  - Their contracts are different: they include the type of the data as part of their names. (eg, `<table>.order-by :: (column :: String) → Table`)
  - They have a "secret" argument, which is the data they are attached to

- 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

Reviewing Functions

1. **One of the examples for the last function is broken!** Fix this example in the Definitions Area.
2. How many *values* are defined in this file? 4
3. How many *functions* are defined in this file? 7
4. What is the *name* of the last function? is-old-dog
5. What is the *Domain* of the last function? Row
6. What is the *Range* of the last function? Boolean
7. What is the variable name that the last function uses? animal
8. Which function will tell us if an animal is a kitten? is-kitten
9. Which function will print out "<name> the <species>"? sentence
10. Which function will tell us if an animal is a dog older than 10? is-old-dog
11. Which function will tell us if an animal has been fixed? is-fixed
12. Which function will draw a nametag for an animal? nametag

Plans for the `Animals` Dataset

What are two ways you might want to *order* the `animals` dataset?

1) Order by weight

2) Order by age

What are two subsets into which you might *filter the `animals` dataset*?

1) Filter animals heavier than 20 pounds

2) Filter animals that have been fixed

What are two new columns you might want to *build* from the `animals` dataset?

1) Add a column for time in the shelter by months

2) Add a column for whether or not each animal is a kitten

Methods

You've seen the `get-row` function before. Given a table and an index, it produces the data row at that index. Here is an example that uses the `get-row` function:

```
get-row(animals-table, 2) # produces the 3rd data row in the animals-table
```

Here is an example that uses ***method,*** which will do the exact same thing:

```
animals-table.row-n(2) # produces the 3rd data row in the animals-table
```

Even though they both do the same thing, both lines of code *look* different!

What do you NOTICE about these lines of code?	**What do you NOTICE about these lines of code?**

Methods are a lot like functions, but they differ in three important ways:

- They can only be called as **part of a value**, using the **dot-accessor**. For example: **`animals.`**`row-n(2)`
- Their Contracts are different, because they contain a **Type** as part of their name. For example: **`<Table>.`**`row-n :: (index :: Number) -> Row`
- They have a "secret argument", which is the value they are attached to. In the examples above, the `row-n` method consumes only a `Number` as part of its Domain, but it *also* consumes the `Table` to which it is attached.

Playing with Methods

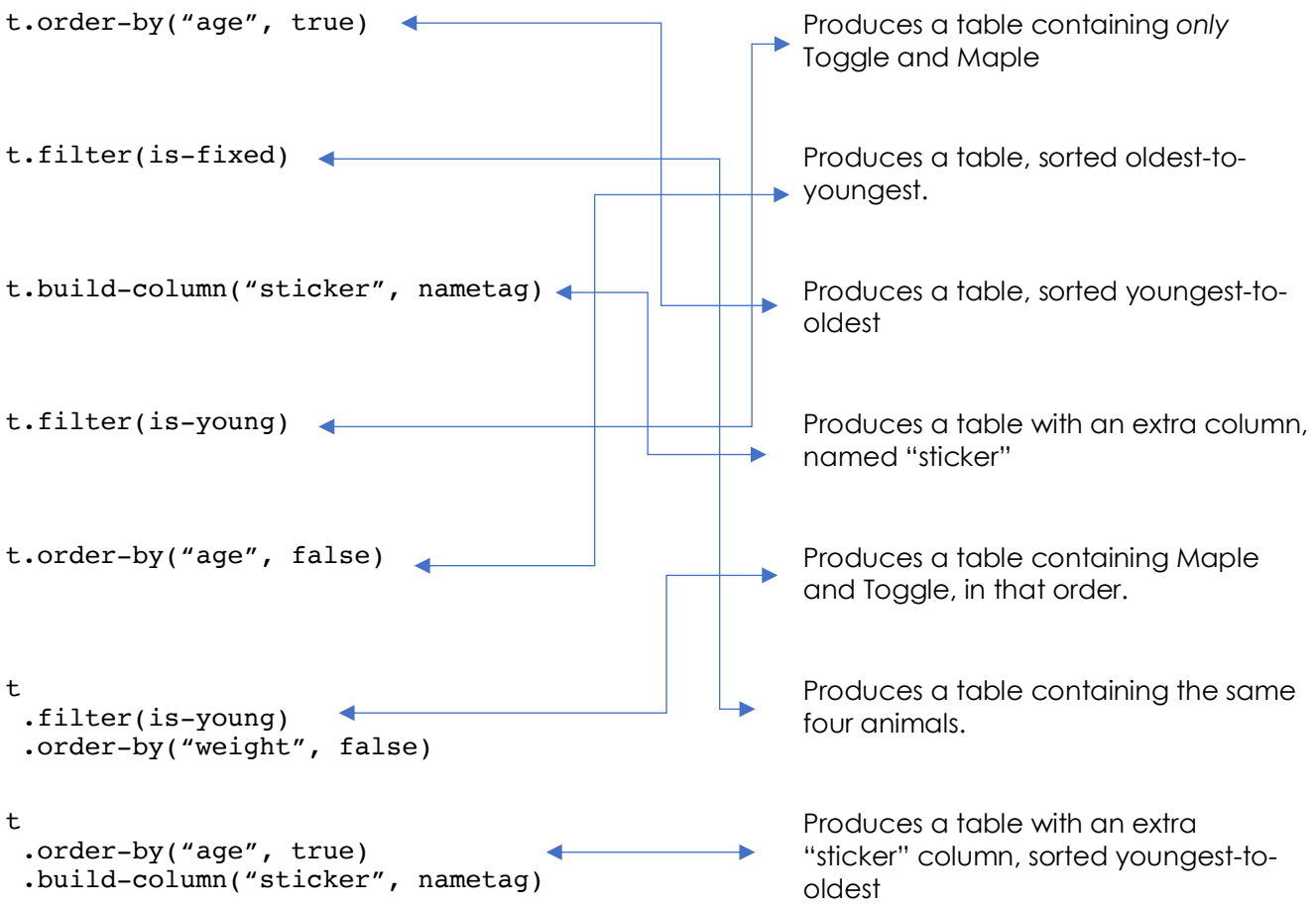
You have the following functions defined below (read them carefully!):

```
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).



Unit 5

- 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 **Start** and **End Tables**, based on the Contract and Purpose.
  - Then we define the function based on our Start and End Tables. Every function includes both the table definition (using methods) and a table expression.

Review

1. In the Interactions Area, use table methods to sort your table by one column. **Try sorting your table in both ascending and descending order.**

2. If a researcher is looking at a dataset of students, they might want to divide the data into separate populations of boys and girls. A veterinarian might want to look at only the cats at a shelter. **Come up with one criteria you could use for animals at the shelter, and describe it below.**

Filter animals heavier than 20 pounds

3. In the space below, **use the Design Recipe to write a function that checks if a row in your dataset fits that criteria.** Whatever criteria you choose, it should be true for some rows and false for others. **Type this function into the Definitions Area.**

```
# is-large :: (animal :: Row) → Boolean
    name          domain          range
# Consumes an animal and produces true if its weight is greater than 20 lbs
examples:
    is-large (sample1) is sample1["weight"] > 20
    is-large (sample2) is sample2["weight"] > 20
end

fun is-large ( animal ) : animal["weight"] > 20
end
```

4. **Use the function to filter your dataset.**

5. Instead of using the function you wrote to *filter* your dataset, **use another table method to build a new column** that shows whether or not each row meets the criteria.

Table Plan

On Kitten Day, the shelter prints up a list of all the cats in their database that are less than 2 years old, and makes nametags for them. They need a function that will help them out! Define a function called `get-kittens-tags`, which takes in the dataset and produces the correct table.

Contract and Purpose

 # get-kittens-tags :: (animals :: Table) → Table

 # Consume a table of animals, and produce a table containing kittens with nametags, sorted by name

Example Tables

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

`animals-table` → `get-kittens-tags(animals-table)`

name	species	age	fixed	legs	weight	adopt
Sasha	cat	1	FALSE	4	6.5	4
Toggle	dog	3	TRUE	4	48	3
Buddy	lizard	2	FALSE	4	0.3	12
Wade	cat	1	FALSE	4	3.2	4
Mittens	cat	2	TRUE	4	7.4	5

name	species	age	fixed	legs	weight	adopt	tag
Sasha	cat	1	FALSE	4	6.5	4	Sasha
Wade	cat	1	FALSE	4	3.2	4	Wade

Define the function

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

```
fun get-kittens-tags ( animals ) :
  t = animals                                          Define the table
  .build-column( nametag )                             Are there more columns?
  .filter( is-kitten )                                 Are there fewer rows?
  .order-by( "name", true )                            Are the rows ordered?
  t                                                    Produce the result
end
```

Table Plan

The first weekend of every month, the shelter holds a "meet the dogs" picnic, to encourage families to adopt their dogs. Write a function called `get-dogs-by-age`, that takes their database and produces a table of all the dogs in the shelter, sorted from youngest to oldest.

Contract and Purpose

 # get-dogs-by-age :: (animals :: Table) → Table

 # Consume a table of animals, and produce a table containing only the dogs, sorted by age

Examples

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

`animals-table` → `get-dog-by-age(animals-table)`

name	species	age	fixed	legs	weight	adopt
Snowcone	cat	2	TRUE	4	6.1	5
Wade	cat	1	FALSE	4	3.2	4
Hercules	cat	3	FALSE	4	13.4	7
Toggle	dog	3	TRUE	4	48	3
Fritz	dog	4	TRUE	4	92	6

name	species	age	fixed	legs	weight	adopt
Toggle	dog	3	TRUE	4	48	3
Fritz	dog	4	TRUE	4	92	6

Define the function

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

```
fun get-dogs-by-age ( animals ) :
  t = animals                                          Define the table
  .build-column(                                )      Are there more columns?
  .filter(                is-dog                )      Are there fewer rows?
  .order-by(              "age", true           )      Are the rows ordered?
                                                       Produce the result
end
```

Table Plan

The first weekend of every month, the shelter holds a "meet the dogs" picnic, to encourage families to adopt their dogs. Write a function called `get-dogs-by-age`, that takes their database and produces a table of all the dogs in the shelter, sorted from youngest to oldest.

Contract and Purpose

get-dogs-by-age :: (animals :: Table) → Table

Consume a table of animals, and produce a table containing only the dogs, sorted by age

Examples

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

`animals-table` → `get-dogs-by-age(animals-table)`

name	species	age	fixed	legs	weight	adopt
Snowcone	cat	2	TRUE	4	6.1	5
Lucky	dog	3	TRUE	3	45.4	9
Mr. PB	dog	10	FALSE	4	161	6
Boo-boo	dog	11	TRUE	4	123	24
Snuggles	tarantula	2	FALSE	8	0.1	1

name	species	age	fixed	legs	weight	adopt
Toggle	dog	3	TRUE	4	48	3
Fritz	dog	4	TRUE	4	92	6

Define the function

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

```
fun get-dogs-by-age ( animals ) :
  t = animals                                          Define the table
  .build-column(                                )      Are there more columns?
  .filter(                  is-dog              )      Are there fewer rows?
  .order-by(                "age", true)               Are the rows ordered?
                                                       Produce the result
end
```

Table Plan

The shelter is tracking birth-years for all the animals who've been fixed. They need a function that takes in their database and returns a table that contains the birth-year for each one. Define `get-fixed-birth` that will do this for them.

Contract and Purpose

get-fixed-birth : : (animals :: Table) → Table

Consumes a table of animals, produces a new table of animals who have been fixed, with a new column for birth year

Examples

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

`animals-table` → `get-fixed-by-legs(animals-table)`

name	species	age	fixed	legs	weight	adopt
Snowcone	cat	2	TRUE	4	6.1	5
Lucky	dog	3	TRUE	3	45.4	9
Hercules	cat	3	FALSE	4	13.4	7
Toggle	dog	3	TRUE	4	48	3
Snuggles	tarantula	2	FALSE	8	0.1	1

name	species	age	fixed	legs	weight	adopt	year
Snowcone	cat	2	TRUE	4	6.1	5	2015
Lucky	dog	3	TRUE	3	45.4	9	2014
Toggle	dog	3	TRUE	4	48	3	2014

Define the function

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

```
fun get-fixed-birth ( animals ) :
  t = animals                                          Define the table
  .build-column( birth-year )                          Are there more columns?
  .filter( is-fixed )                                  Are there fewer rows?
  .order-by( )                                         Are the rows ordered?
  _____                                                Produce the result
end
```

My Dataset

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

1) [specific to each student]

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)

Unit 6

- **Bar charts** show the *absolute* quantity of each row in a dataset. The larger the quantity, the longer the bar. Bar charts provide a visual representation of values in a dataset.

- **Pie charts** show the *relative* quantity of each row in a dataset. The greater the percentage, the larger the pie slice. Pie charts provide a visual representation of proportions in a dataset.

- **Choosing a Sample Table** is important when coming up with small examples for Table Plans. A good sample table has:
  1. At least all the relevant columns
  2. Enough rows to accurately represent the dataset
  3. Rows that are randomly-ordered

Statements about Columns

Use the Table below to help you answer the questions.

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

1. Which animal(s) is/are the heaviest? Boo-boo

2. Which animal(s) is/are the youngest? Sasha, Wade

3. How much of the *total weight* comes from Maple? 22%

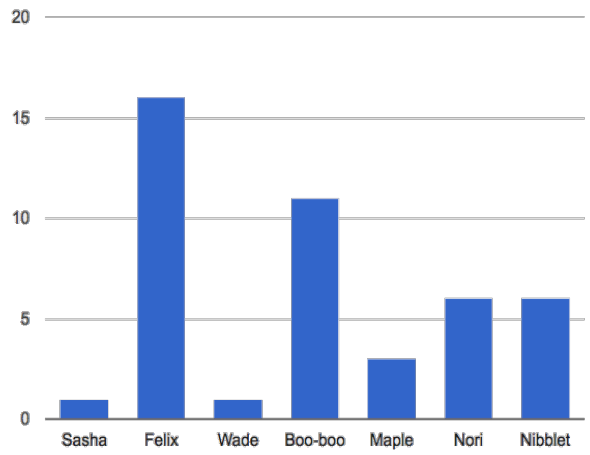
4. How much of the *combined age* comes from Nori? 13%

5. Would these questions be harder to answer if the table had 100 rows? If so, why?

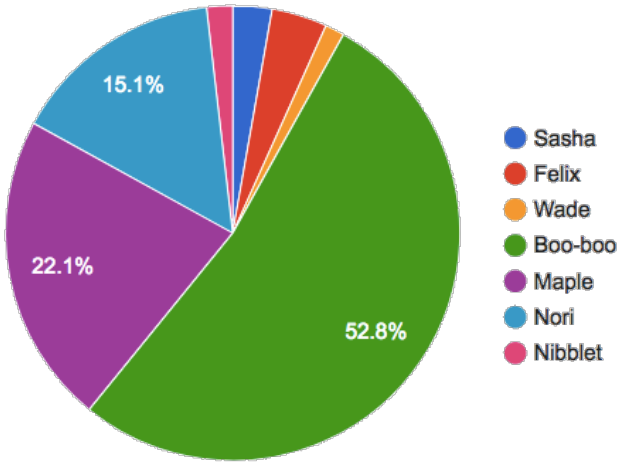
Much harder if you were estimating, because it is harder to calculate a large number of entries without using software.

Visualizing Quantity

In the table below, there are two observations drawn from the following charts. Add two more.



Animals Ages (yrs)



Animals Weights (lbs)

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

Table Plan

Dogs are generally a lot bigger heavier than cats, so the shelter wants to look at a chart of *only* the dogs to determine who needs more exercise time. Define a function `pie-dog-weight`, which will make a pie chart showing the relative weights of all the dogs in the shelter.

Contract and Purpose

_____ : : _____ → _____

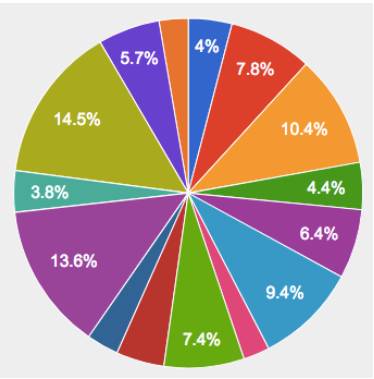
_____

Examples

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

`animals-table` → `pie-dog-weight(animals-table)`

name	...	pounds
Snowcone	...	6.1
Lucky	...	45.4
Hercules	...	13.4
Toggle	...	48
Snuggles	...	0.1



Define the function

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

```
fun _____ ( animals ) :
  ____(t, "name", "pounds")
  _____
  _____
  _____
  _____
  _____
end
```

Define the table

Produce the result

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 to determine if the Sample Table meets those criteria.

1. The shelter wants to know the median age of all the cats

name	species	age	fixed	legs	pounds	weeks
Sasha	cat	1	FALSE	4	6.5	3
Mittens	cat	2	TRUE	4	7.4	5
Sunfower	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 `bar-kitten-adoption`, which takes in a Table of animals and creates a bar chart showing how many weeks it took for each kitten to be adopted

Contract and Purpose

#_____::_____ → _____

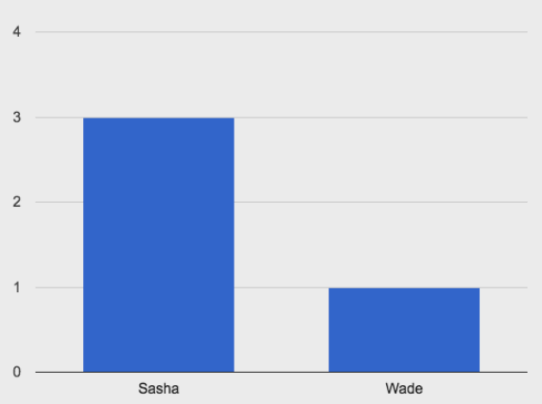
_____

Examples

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

_____ → _____

name	species	age	fixed	pounds	weeks
Sasha	cat	1	FALSE	6.5	3
Wade	cat	1	FALSE	7.4	1
Sunfower	cat	5	TRUE	8.1	10
Boo-boo	dog	11	TRUE	123	10



Define the function

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

```
fun _____ (_____) :
    t = animals                                          Define the table
    .build-column(                                  )    Are there more columns?
    .filter(          is-kitten                     )    Are there fewer rows?
    .order-by(                                      )    Are the rows ordered?
        bar-chart(t, "name", "weeks")                    Produce the result
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 = _____                                          Define the table
   _____                                              Are there more columns?
   _____                                              Are there fewer rows?
   _____                                              Are the rows ordered?
   _____                                              Produce the result
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

Visualizing My Dataset

What quantity charts did you make, and what do you notice? Fill in the table below.

Based on a _____ chart of _____	**I notice that _____**

Unit 7

- There are three ways to measure the "center" of a dataset, to talk about a whole column of data using just one number:

    1. The **mean** of a dataset is the average of all the numbers

    2. The **median** of a dataset is a value that is smaller than half the dataset, and larger than the other half

    3. The **modes** of a dataset are the numbers that appear the most often.

- Data Scientists can also measure the "variation" of a dataset using a **five number summary:**

    1. The **minimum** – the smallest value in the dataset

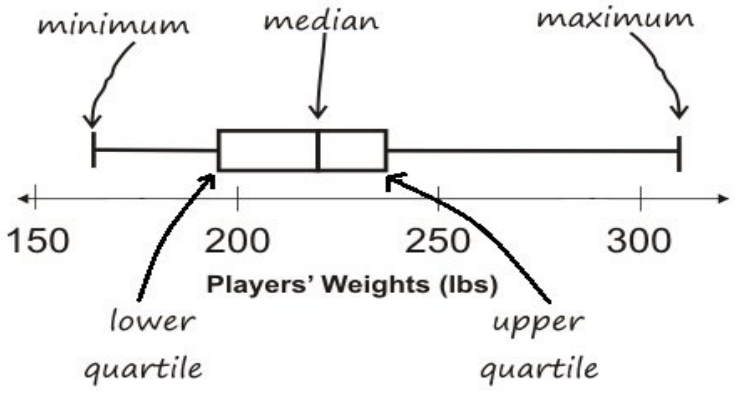
    2. The **first, or "lower" quartile (Q1)** – the median value that separates the first quarter of the values in the dataset from the second quarter

    3. The **second quartile (Q2)** – the median value which separates the entire dataset into "top" and "bottom" halves.

    4. The **third, or "upper" quartile (Q3)** – the median value that separates the third quarter of the values in the dataset from the fourth quarter

    5. 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`

The column I choose to measure is _____

Measures of Center

The three measures for this column are:

Mean (Average)	Median	Mode(s)

Based on the differences between mean and median, I conclude :

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:

Table Plan

The shelter wants a summary of the variation in ages among the dogs. Write a function called `variation-dog-age` that will take in a table of animals and produce a box-plot that shows this variation.

Contract and Purpose

_____ : : _____ → _____

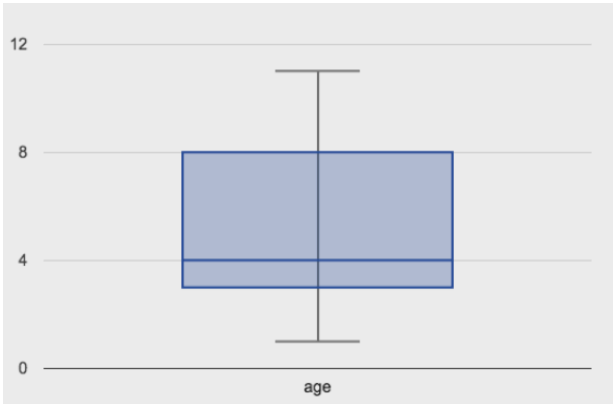
_____

Examples

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

`animals-table` → `variation-dog-age(animals-table)`

name	species	age	fixed	legs	weight	weeks
Snowcone	cat	2	TRUE	4	6.1	5
Lucky	dog	3	TRUE	3	45.4	9
Hercules	cat	3	FALSE	4	13.4	7
Toggle	dog	3	TRUE	4	48	3
Snuggles	tarantula	2	FALSE	8	0.1	1



Define the function

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

```
fun _____ (_____) :
    t = animals                                          Define the table
    .build-column(                                )      Are there more columns?
    .filter(          is-dog                      )      Are there fewer rows?
    .order-by(                                    )      Are the rows ordered?
        box-plot(t, "age")                               Produce the result
end
```

Interpreting Variation

Consider the following list 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**.

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

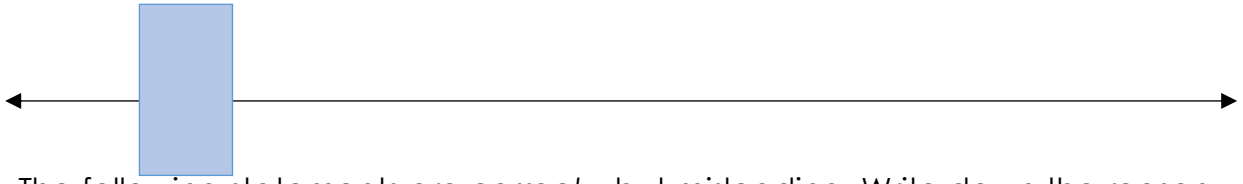
2. In the table below, compute the **measures of center** for this dataset.

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!"	While the mean is close to $70k, there are some very high earning outliers pushing the average up
"It's a middle-income list: the most common salary is $45k/yr!"	Looking at the full dataset, more than half of the entries are people making less than $45k, making the mode misleading
"This group is really diverse, with people making as little as 12k and as much as $280k!"	While the spread of incomes is large, the vast majority are still making less than $65k, with very high earning outliers.

Summarizing a Column in My Dataset

The column I choose to measure is [Specific to each student]

Measures of Center

The three measures for this column are:

Mean (Average)	Median	Mode(s)

Based on the differences between mean and median, I conclude :

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 8

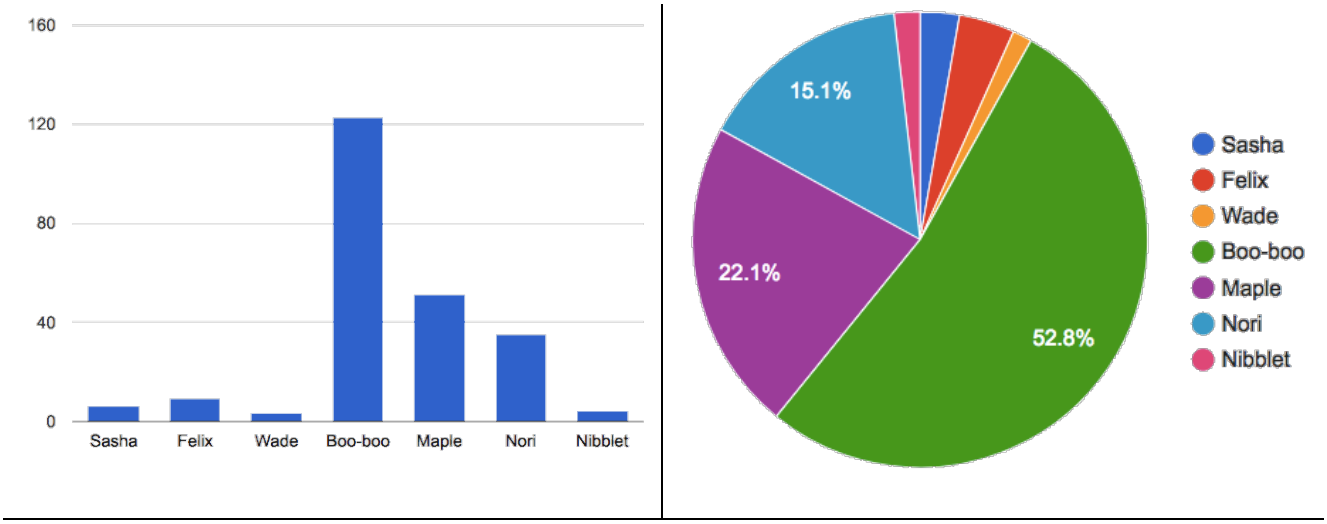
- **Frequency Bar charts** show the number of rows belonging to a given category. The more rows in each category, the longer the bar. Frequency bar charts provide a visual representation of the frequency of values in a **categorical** column. Since categorical data cannot be ordered, there is no strict ordering of bars in a frequency bar chart.

- **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 longer 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 are always sorted.

- 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.

Visualizing Quantity (Review)

Use the charts below to help you answer the questions.



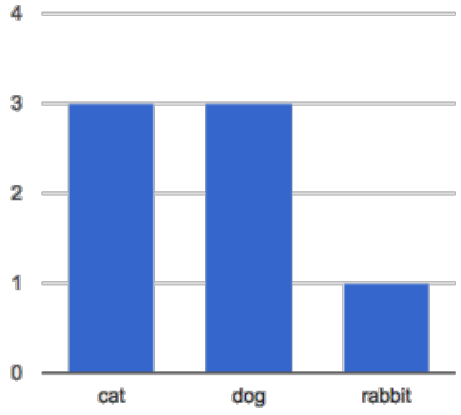
Animals Weights (lbs)

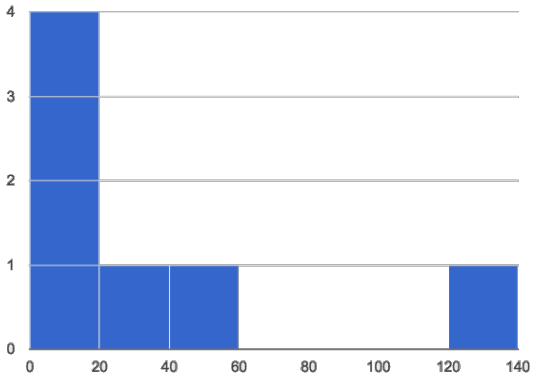
1. Which animal is the heaviest? _____
2. Which animal is the lightest? _____
3. How much of the *total weight* comes from Maple? _____
4. How much of the *total weight* comes from Nori? _____
5. Which chart did you use for questions 1 and 2? _____
6. Which chart did you use for questions 3 and 4? _____
7. Why are some questions easier to answer with one kind of chart or another?

Visualizing Frequency

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 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. What is each one measuring? Write down your guess underneath each one.





_____ _____

_____ _____

_____ _____

_____ _____

Table Plan

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

Contract and Purpose

_____ : : _____ → _____

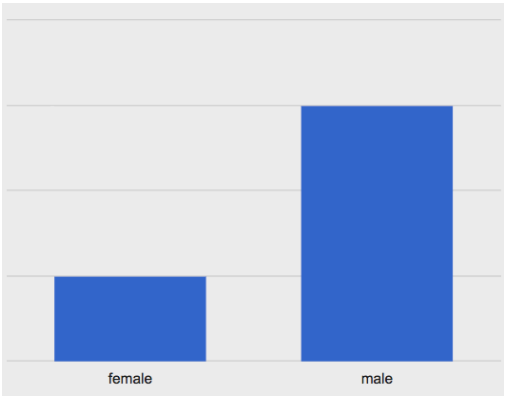
_____

Examples

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

_____ → _____

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



Define the function

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

```
fun _____ (_____) :
    t = animals
    _____
    _____
    _____
    freq-bar-chart(t, "gender")
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-adoption`, which takes in a Table of animals and creates a histogram showing how long it took for animals to get adopted

Contract and Purpose

_____ : : _____ → _____

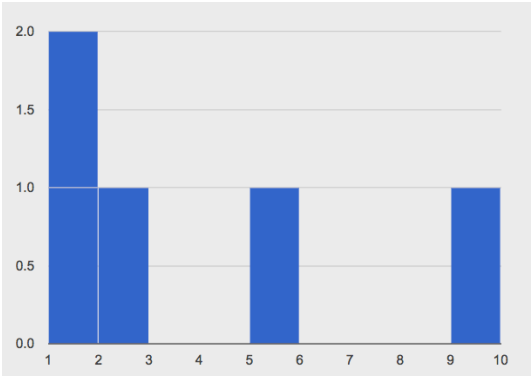
_____

Examples

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

_____ → _____

name	species	age	fixed	legs	weight	weeks
Snowcone	cat	2	TRUE	4	6.1	5
Lucky	dog	3	TRUE	3	45.4	9
Hercules	cat	3	FALSE	4	13.4	7
Toggle	dog	3	TRUE	4	48	3
Snuggles	tarantula	2	FALSE	8	0.1	1



Define the function

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

fun _____ (_____) :

Define the table

t = animals _____

Are there more columns?

_____

Are there fewer rows?

_____

Are the rows ordered?

_____

Produce the result

histogram(t, "weeks", 1) _____

end

Visualizing My Dataset

What frequency charts did you make? Sketch at least one in the space below.

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

Matching Charts to Questions

that younger animals [specific/tend] get adopted faster, possibly because they are considered cuter, but there may be other factors causing them to get adopted faster

I would look at both the ages and number of weeks until adoption for each animal to see if there was a correlation. I would also want to collect more data, such as conduct a survey of adopters.

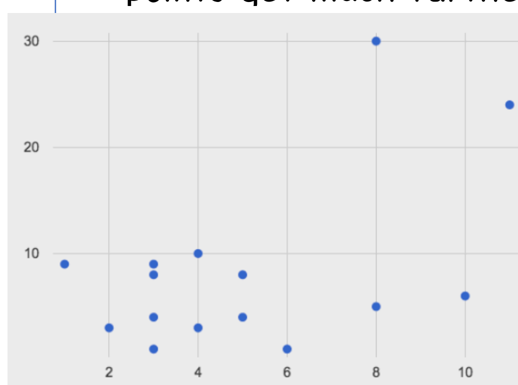
Consumes a table of animals and produces a scatter plot showing the relationship between age and weeks to adoption

w

A few points are close to the line, but as ages increase the points get much farther apart.

Slightly upwards

Image



t = animals-table

Produce the result

.filter(is-dog)

scatter-plot(t, "age", "weeks")

0.25

0.5

0.9

1	T	" a subset or extension dataset or subset	positive / negative

a weak (r² = 0.25) positive

a weak (r² = 0.025), positive

25% of the variability in adoption time is explained

r² % of the variation in [y-axis] is explained by [x-axis]

by the weight of the cat

by the age of the animal

dataset or subset

The average is based on all the players, and there may be outliers pushing the average height up-average tells you nothing about the majority of the players.

2	A	“	Only 18% of the variation in salary is based on height, which is not a large enough r-squared value to say that taller people get paid more.
3		“	The r-squared value of 0.636 does not mean how often the y-value will be predicted, rather what percent of variation in the y-value is based on the x-value.
4		“	Bar charts are not the most appropriate [illegible] pie charts should be used for that info.
5		“	More animals [illegible] for the histogram [illegible] This chart shows that Felix is a little more than 15 years old.
6	A	“	Although there is a strong correlation between hair and owning a wig, correlation does NOT equal causation.

t = *Produce the result* *Define the table*

Table Plan

t =

Produce the result
Define the table

N	D	
t<	: :	→
c<	: :	→
s<	: :	→
r<	: :	→
e_m	: :	→
s_m	: :	→
t_m	: :	→
o_b	: :	→
r_p	: :	→
s_b	: :	→
s_f	: :	→
s_h	: :	→
n_s	: :	→
n_l	: :	→
n_l	: :	→
n	: :	→
q	:	→