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 numeric, and measures an amount, such as a person’s height, a score on test, distance, etc. A list of quantitative data can be ordered from smallest to largest.
- Categorical data is data that specifies qualities, such as gender, 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.

Answering questions with 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 of 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 average weight of animals from the shelter?”)
- **Relate Questions** require looking for trends across multiple columns. (E.g., “Do cats tend to be adopted sooner than dogs?”)
### The Animals Dataset

<table>
<thead>
<tr>
<th>name</th>
<th>Species</th>
<th>Gender</th>
<th>Age (years)</th>
<th>Fixed</th>
<th>Legs</th>
<th>Weight (lbs)</th>
<th>Time to Adoption (weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sasha</td>
<td>cat</td>
<td>female</td>
<td>1</td>
<td>FALSE</td>
<td>4</td>
<td>6.5</td>
<td>3</td>
</tr>
<tr>
<td>Mittens</td>
<td>cat</td>
<td>female</td>
<td>2</td>
<td>TRUE</td>
<td>4</td>
<td>7.4</td>
<td>1</td>
</tr>
<tr>
<td>Sunflower</td>
<td>cat</td>
<td>female</td>
<td>5</td>
<td>TRUE</td>
<td>4</td>
<td>8.1</td>
<td>6</td>
</tr>
<tr>
<td>Sheba</td>
<td>cat</td>
<td>female</td>
<td>7</td>
<td>TRUE</td>
<td>4</td>
<td>8.4</td>
<td>6</td>
</tr>
<tr>
<td>Felix</td>
<td>cat</td>
<td>male</td>
<td>16</td>
<td>TRUE</td>
<td>4</td>
<td>9.2</td>
<td>5</td>
</tr>
<tr>
<td>Snowcone</td>
<td>cat</td>
<td>female</td>
<td>2</td>
<td>TRUE</td>
<td>4</td>
<td>6.5</td>
<td>5</td>
</tr>
<tr>
<td>Wade</td>
<td>cat</td>
<td>male</td>
<td>1</td>
<td>FALSE</td>
<td>4</td>
<td>3.2</td>
<td>1</td>
</tr>
<tr>
<td>Hercules</td>
<td>cat</td>
<td>male</td>
<td>3</td>
<td>FALSE</td>
<td>4</td>
<td>13.4</td>
<td>2</td>
</tr>
<tr>
<td>Toggle</td>
<td>dog</td>
<td>female</td>
<td>3</td>
<td>TRUE</td>
<td>4</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>Boo-boo</td>
<td>dog</td>
<td>male</td>
<td>11</td>
<td>TRUE</td>
<td>4</td>
<td>123</td>
<td>24</td>
</tr>
<tr>
<td>Fritz</td>
<td>dog</td>
<td>male</td>
<td>4</td>
<td>TRUE</td>
<td>4</td>
<td>92</td>
<td>3</td>
</tr>
<tr>
<td>Maple</td>
<td>dog</td>
<td>female</td>
<td>3</td>
<td>TRUE</td>
<td>4</td>
<td>51.6</td>
<td>4</td>
</tr>
<tr>
<td>Bo</td>
<td>dog</td>
<td>male</td>
<td>8</td>
<td>TRUE</td>
<td>4</td>
<td>76.1</td>
<td>10</td>
</tr>
<tr>
<td>Midnight</td>
<td>dog</td>
<td>female</td>
<td>5</td>
<td>FALSE</td>
<td>4</td>
<td>112</td>
<td>4</td>
</tr>
<tr>
<td>Rex</td>
<td>dog</td>
<td>male</td>
<td>1</td>
<td>FALSE</td>
<td>4</td>
<td>28.9</td>
<td>9</td>
</tr>
<tr>
<td>Gir</td>
<td>dog</td>
<td>male</td>
<td>8</td>
<td>FALSE</td>
<td>4</td>
<td>88</td>
<td>5</td>
</tr>
<tr>
<td>Max</td>
<td>dog</td>
<td>male</td>
<td>3</td>
<td>FALSE</td>
<td>4</td>
<td>52.8</td>
<td>8</td>
</tr>
<tr>
<td>Nori</td>
<td>dog</td>
<td>female</td>
<td>3</td>
<td>TRUE</td>
<td>4</td>
<td>35.3</td>
<td>1</td>
</tr>
<tr>
<td>Mr. Peanutbutter</td>
<td>dog</td>
<td>male</td>
<td>10</td>
<td>FALSE</td>
<td>4</td>
<td>161</td>
<td>6</td>
</tr>
<tr>
<td>Lucky</td>
<td>dog</td>
<td>male</td>
<td>3</td>
<td>TRUE</td>
<td>3</td>
<td>45.4</td>
<td>9</td>
</tr>
<tr>
<td>Kujo</td>
<td>dog</td>
<td>male</td>
<td>8</td>
<td>FALSE</td>
<td>4</td>
<td>172</td>
<td>30</td>
</tr>
<tr>
<td>Buddy</td>
<td>lizard</td>
<td>male</td>
<td>2</td>
<td>FALSE</td>
<td>4</td>
<td>0.3</td>
<td>3</td>
</tr>
<tr>
<td>Gila</td>
<td>lizard</td>
<td>female</td>
<td>3</td>
<td>TRUE</td>
<td>4</td>
<td>1.2</td>
<td>4</td>
</tr>
<tr>
<td>Snuffles</td>
<td>rabbit</td>
<td>female</td>
<td>3</td>
<td>TRUE</td>
<td>4</td>
<td>3.5</td>
<td>8</td>
</tr>
<tr>
<td>Niblet</td>
<td>rabbit</td>
<td>male</td>
<td>6</td>
<td>FALSE</td>
<td>4</td>
<td>4.3</td>
<td>2</td>
</tr>
<tr>
<td>Snuggles</td>
<td>tarantula</td>
<td>female</td>
<td>2</td>
<td>FALSE</td>
<td>8</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>Daisy</td>
<td>dog</td>
<td>female</td>
<td>5</td>
<td>TRUE</td>
<td>4</td>
<td>68</td>
<td>8</td>
</tr>
<tr>
<td>Ada</td>
<td>dog</td>
<td>female</td>
<td>2</td>
<td>TRUE</td>
<td>4</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>Miaulis</td>
<td>cat</td>
<td>male</td>
<td>7</td>
<td>FALSE</td>
<td>4</td>
<td>8.8</td>
<td>4</td>
</tr>
<tr>
<td>Heathcliff</td>
<td>cat</td>
<td>male</td>
<td>1</td>
<td>TRUE</td>
<td>4</td>
<td>2.1</td>
<td>2</td>
</tr>
<tr>
<td>Tinkles</td>
<td>cat</td>
<td>female</td>
<td>1</td>
<td>TRUE</td>
<td>4</td>
<td>1.7</td>
<td>3</td>
</tr>
</tbody>
</table>
### Categorical or Quantitative?

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Categorical</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hair color</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>3</td>
<td>ZIP Code</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>4</td>
<td>Year</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>5</td>
<td>Height</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>6</td>
<td>Gender</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>7</td>
<td>Street Name</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Categorical</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>We’d like to find out the average price of cars in a lot.</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>9</td>
<td>We’d like to find out the most popular color for cars.</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>10</td>
<td>We’d like to find out which puppy is the youngest.</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>11</td>
<td>We’d like to find out which cats have been fixed.</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>12</td>
<td>We want to know which people have a ZIP code of 02907.</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
<tr>
<td>13</td>
<td>We’d like to sort a list of phone numbers by area code.</td>
<td>categorical</td>
<td>quantitative</td>
</tr>
</tbody>
</table>
# The Animals Dataset

<table>
<thead>
<tr>
<th>What do you NOTICE about this dataset?</th>
<th>What do you WONDER about this dataset?</th>
<th>Question Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lookup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compute</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can’t answer</td>
</tr>
</tbody>
</table>

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

2. Some of the columns are:

   a. __________ species __________, which contains __________ categorical __________ data. Some example values are:

   __________________________________________________________________________.

   "cat", "dog", and "rabbit"

   b. ________________________________________________________________________

   __________________________________________________________________________.

   which contains ______________ data. Some example values are:

   __________________________________________________________________________.
What’s on your mind?
Programming languages involve different *datatypes*, such as Numbers, Strings, and Booleans. Numbers are usually used for quantitative data, and other values are typically used as categorical data.

- Numbers are values like 1, 0.4, 1/3, and -0.261.003.
- Any decimal *must* start with a 0.022 is valid, but .22 is not.
- Strings are values like "Emma", "Rosanna", "Jen and Ed", or even "08/28/1980".
- All strings *must* be surrounded in quotation marks.
- Booleans are either *true* or *false*.

**Operators** (like +, -, *, <, etc.) work the same way in Pyret that they do in math. Operators are written between values (for example, 4 + 2).

- Operators must always have a space around them. 4 + 2 is valid, but 4+2 is not.
- If an expression has different operators, parentheses must be used to show order of operations. 4 + 2 + 6 and 4 + (2 * 6) are valid, but 4 + 2 * 6 is not.

**Applying Functions** (like f(5), f(g(10, 4))) also works the way it does in math. Applying a Pyret function (like num-sqr, num-sqrt triangle, star, string-repeat, etc.) involves putting the function name first, followed by a list of arguments in parentheses. For example: star(50, "solid", "red").

Functions have *contracts*, which help explain how a function should be used. Every contract has three parts:

1. The *Name* of the function - literally, what it's called.
2. The *Domain* of the function - what *types of values* the function consumes, and in what order.
3. The *Range* of the function - what *type of value* the function produces.

**Value Definitions** (like x = 4, or y = 9 + 6) also work the way they do in math. Every value definition starts with *aname*, followed by an equals sign, and then an expression. Once a value is defined, it can be referred to by name.
Numbers and Strings

Make sure you’ve loaded the Animals 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 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. Type in the following expressions, one at a time: 4 + 2 + 6, 4 + 2 * 6, 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.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>3 &lt;= 4</code></td>
<td></td>
</tr>
<tr>
<td><code>3 == 2</code></td>
<td></td>
</tr>
<tr>
<td><code>2 &lt;&gt; 4</code></td>
<td></td>
</tr>
<tr>
<td><code>3 &lt;&gt; 3</code></td>
<td></td>
</tr>
<tr>
<td><code>&quot;a&quot; &gt; &quot;b&quot;</code></td>
<td></td>
</tr>
<tr>
<td><code>&quot;a&quot; &lt;&gt; &quot;b&quot;</code></td>
<td></td>
</tr>
<tr>
<td><code>&quot;a&quot; == &quot;b&quot;</code></td>
<td></td>
</tr>
<tr>
<td><code>&quot;a&quot; &lt;&gt; &quot;a&quot;</code></td>
<td></td>
</tr>
</tbody>
</table>

9) How many different Number values are there in Pyret?

10) How many different String values are there in Pyret?

11) How many different Boolean values are there in Pyret?
### Applying Functions

Type this line of code into the interactions area and hit "Enter":
```py
triangle(50, "solid", "red")
```

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is the name of this function?</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>What did the expression evaluate to?</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>How many arguments does <code>triangle</code> expect?</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>What does the <code>triangle</code> function produce? (Numbers? Strings? Booleans?)</td>
<td></td>
</tr>
</tbody>
</table>

### Catching Bugs

The following lines of code are all BUGGY! Can you spot the mistake? In the blanks below, write down what the problem is, then type in the code and see what error message Pyret produces.

<table>
<thead>
<tr>
<th></th>
<th>Code</th>
<th>Error Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><code>triangle(20, &quot;solid&quot; &quot;red&quot;)</code></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><code>triangle(20, &quot;solid&quot;)</code></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><code>triangle(20, 10, &quot;solid&quot;, &quot;red&quot;)</code></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td><code>triangle (20, &quot;solid&quot;, &quot;red&quot;)</code></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td><code>triangle 20, &quot;solid&quot;, &quot;red&quot;)</code></td>
<td></td>
</tr>
</tbody>
</table>
Contracts

Consider the following contract:

rotate :: (degree :: Number, img :: Image) -> Image

What is the Name of this function? ____________________________
How many things are in this function’s Domain? ____________________________
What is the type of this function’s first argument? ____________________________
What is the name of this function’s second argument? ____________________________
What is the Range of this function? ____________________________.

Below, circle which expression is the correct application of this function, based on its contract.

1. rotate(45, 90)
2. rotate(circle(99, "solid", "green"))
3. rotate(25, rectangle(7, 10, "outline", "black"))
4. rotate(rectangle(7, 10, "outline", "black"), 25)
Matching Expressions and Contracts

Read the following 5 contracts carefully:

A. `make-id :: (name :: String, age :: Number) -> Image`
B. `phone-bill :: (minutes :: Number, texts :: Number) -> Number`
C. `phone-bill :: (minutes :: Number) -> Number`
D. `make-id :: (first :: String, last :: String) -> Image`
E. `make-id :: (first :: String, last :: String, age :: Number) -> Image`

For each of the expressions below, indicate which contract (A–E) describes the function being used.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Contract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  <code>make-id(&quot;Hannah&quot;, &quot;Smith&quot;)</code></td>
<td></td>
</tr>
<tr>
<td>2  <code>make-id(&quot;George&quot;, 17)</code></td>
<td></td>
</tr>
<tr>
<td>3  <code>phone-bill(31, 287)</code></td>
<td></td>
</tr>
<tr>
<td>4  <code>make-id(&quot;Jessica&quot;, &quot;Jones&quot;, 32)</code></td>
<td></td>
</tr>
<tr>
<td>5  <code>phone-bill(55)</code></td>
<td></td>
</tr>
</tbody>
</table>
Plotting and Displaying Data

Data Scientists use displays to visualize data. You’ve probably seen some of these charts, graphs and plots yourselves! When it comes to displaying Categorical Data, there are two displays that are especially useful.

1. **Bar charts** show the count or percentage of rows in each category.
   - Bar charts provide a visual representation of the frequency of values in a categorical column.
   - Bar charts have a bar for every category in a column
   - The more rows in a category, the taller the bar.
   - Bars in a bar chart can be shown in any order, without changing the meaning of the chart. However, bars are usually shown in some sensible order (bars for the number of orders for different t-Shirt sizes might be presented in order of smallest to largest shirt).

2. **Pie charts** show the percentage of rows in each category.
   - Pie charts provide a visual representation of the relative frequency of values in a categorical column.
   - Pie charts have a slice for every category in a column
   - The more rows in a category, the larger the slice.
   - Slices in a pie chart can be shown in any order, without changing the meaning of the chart. However, slices are usually shown in some sensible order (slices might be shown in alphabetical order, or from the smallest to largest slice).
### Exploring Displays

For each type of display, fill in the information below.

<table>
<thead>
<tr>
<th>Pie Charts</th>
<th>Bar Charts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketch a pie chart here</td>
<td>Sketch a bar chart here</td>
</tr>
<tr>
<td>Pie charts are constructed from ______ column(s), and show <strong>categorical</strong> data. A pie chart tells us</td>
<td>Bar charts are constructed from ______ column(s), and show ______________ data. A bar chart tells us</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Box Plots</th>
<th>Histograms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketch a box plot here</td>
<td>Sketch a histogram here</td>
</tr>
<tr>
<td>Box plots are constructed from ______ column(s), and show ______________ data. A pie chart tells us</td>
<td>Histograms are constructed from ______ column(s), and show ______________ data. Histograms tell us</td>
</tr>
</tbody>
</table>
For each type of display, fill in the information below.

<table>
<thead>
<tr>
<th>Scatter Plot</th>
<th>Linear Regression Plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketch a scatter plot here</td>
<td>Sketch a scatter plot here</td>
</tr>
</tbody>
</table>

Scatter Plots are constructed from ________ column(s), and show _______________ data. Scatter plots tell us _______________.

LR Plots are constructed from ________ column(s), and show _______________ data. An LR Plot tells us _______________.

---

15
Data Displays and Lookups

Data scientists use data visualizations to gain better insights into their data, and to communicate their findings with others. Making a display requires answering three questions:

1. **What data** is being displayed? This could be "a random sample of 2000 people", "every animal from the shelter", or "students' aged 14-17".

2. **What variables** are being explored? Are we looking at the species column? The number of kilograms that an animal weighs? Searching for a relationship between a person's income and their height?

3. **What display** is being used, given the variables being explored? If it's a quantitative variable, we might use a histogram or box plot. If it's categorical, we could use a pie or bar chart. If it's two quantitative variables, we probably want a scatter plot.

When looking up a data Row from a Table, programmers use the `row-n` method. This method takes a single number as its input, which tells the computer which Row we want. **Note: Rows are numbered starting at zero!**

For example:

```
animals-table.row-n(0)    # access the 1st data row
animals-table.row-n(16)   # access the 17th data row
```

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

For example:

```
animals-table.row-n(0)["age"]  # look up the age of the animal in the 1st data row
animals-table.row-n(16)["species"]  # look up the species of the animal in the 17th data row
```
<table>
<thead>
<tr>
<th>Display</th>
<th>Column Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pie Charts</td>
<td>A 1 column of Quantitative Data</td>
</tr>
<tr>
<td>Bar Charts</td>
<td>B 2 columns of Quantitative Data</td>
</tr>
<tr>
<td>Histograms</td>
<td></td>
</tr>
<tr>
<td>Box Plots</td>
<td></td>
</tr>
<tr>
<td>Scatter Plots</td>
<td>C 1 column of Categorical Data</td>
</tr>
</tbody>
</table>
Displaying Data

Fill in the tables below, then use Pyret to make the following displays. The first column has been filled in for you.

1) A pie-chart showing the species of animals from the shelter.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3) A pie-chart showing the how many animals are fixed or not.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4) A histogram of the number of pounds that animals weigh.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6) A histogram of the number of legs that animals have.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7) A scatter-plot, using the animals name as the labels, age as the x-axis, and pounds as the y-axis.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8) A scatter-plot, using the animals name as the labels, age as the x-axis, and weeks as the y-axis.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9) A scatter-plot, using the animals name as the labels, pounds as the x-axis, and weeks as the y-axis.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Animals Dataset</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Lookup Questions

The table below represents four pets at an animal shelter:

<table>
<thead>
<tr>
<th>name</th>
<th>gender</th>
<th>age</th>
<th>pounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Toggle&quot;</td>
<td>&quot;female&quot;</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>&quot;Fritz&quot;</td>
<td>&quot;male&quot;</td>
<td>4</td>
<td>92</td>
</tr>
<tr>
<td>&quot;Nori&quot;</td>
<td>&quot;female&quot;</td>
<td>6</td>
<td>35.3</td>
</tr>
<tr>
<td>&quot;Maple&quot;</td>
<td>&quot;female&quot;</td>
<td>3</td>
<td>51.6</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Question</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;How much does Maple weigh?&quot;</td>
<td>A animals-table.row-n(3)</td>
</tr>
<tr>
<td>&quot;Which is the last row in the table?&quot;</td>
<td>B animals-table.row-n(2)[&quot;name&quot;]</td>
</tr>
<tr>
<td>&quot;What is Fritz's gender?&quot;</td>
<td>C animals-table.row-n(1)[&quot;gender&quot;]</td>
</tr>
<tr>
<td>&quot;What's the third animal's name?&quot;</td>
<td>D animals-table.row-n(3)[&quot;age&quot;]</td>
</tr>
<tr>
<td>&quot;How much does Nori weigh?&quot;</td>
<td>E animals-table.row-n(3)[&quot;pounds&quot;]</td>
</tr>
<tr>
<td>&quot;How old is Maple?&quot;</td>
<td>F animals-table.row-n(0)</td>
</tr>
<tr>
<td>&quot;What is Toggle's gender?&quot;</td>
<td>G animals-table.row-n(2)[&quot;pounds&quot;]</td>
</tr>
<tr>
<td>&quot;What is the first row in the table?&quot;</td>
<td>H animals-table.row-n(0)[&quot;gender&quot;]</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Blank</th>
<th>Code</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>animals-table.row-n(3)[&quot;name&quot;]</td>
<td>&quot;Maple&quot;</td>
</tr>
<tr>
<td>b.</td>
<td></td>
<td>&quot;male&quot;</td>
</tr>
<tr>
<td>c.</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>d.</td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>e.</td>
<td></td>
<td>&quot;Nori&quot;</td>
</tr>
</tbody>
</table>
What’s on your mind?
Defining Functions

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

- We use the Design Recipe to help us define functions and think through problems clearly.
- 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

Directions: Define a function called \( \text{gt} \), which makes solid green triangles of whatever size we want.

Contract and Purpose Statement

Every contract has three parts…

\[
\begin{align*}
\underline{\text{function name}} & : \quad \underline{(\text{size :: Number})} & \quad \rightarrow & \quad \underline{\text{Image}} \\
\text{Consumes a size, and produces a solid green triangle of that size.}
\end{align*}
\]

Examples

Write some examples, then circle and label what changes…

\[
\begin{align*}
\text{examples:} & \\
\underline{\text{function name}} & \quad \underline{(\text{input(s)})} & \quad \underline{\text{is}} & \quad \underline{\text{what the function produces}} \\
\underline{\text{function name}} & \quad \underline{(\text{input(s)})} & \quad \underline{\text{is}} & \quad \underline{\text{what the function produces}}
\end{align*}
\]

Definition

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

\[
\begin{align*}
\text{fun} & \quad \text{gt} \quad \underline{(\text{size })}: \\
\text{triangle(} & \text{size, "solid", "green"} & \underline{\text{what the function does with those variable(s)}}
\end{align*}
\]

Directions: Define a function called \( \text{bc} \), which makes solid blue circles of whatever radius we want.

Contract and Purpose Statement

Every contract has three parts…

\[
\begin{align*}
\underline{\text{function name}} & : \quad \underline{\text{domain}} & \quad \rightarrow & \quad \underline{\text{range}} \\
\text{Consumes a radius, and produces a solid blue circle with that radius.}
\end{align*}
\]

Examples

Write some examples, then circle and label what changes…

\[
\begin{align*}
\text{examples:} & \\
\underline{\text{function name}} & \quad \underline{(\text{input(s)})} & \quad \underline{\text{is}} & \quad \underline{\text{what the function produces}} \\
\underline{\text{function name}} & \quad \underline{(\text{input(s)})} & \quad \underline{\text{is}} & \quad \underline{\text{what the function produces}}
\end{align*}
\]

Definition

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

\[
\begin{align*}
\text{fun} & \quad \text{bc} \quad \underline{\text{variable(s)}}: \\
\underline{\text{what the function does with those variable(s)}}
\end{align*}
\]
The Design Recipe

Directions: Define a function called `sticker`, which draws 50px stars in whatever color is input.

Contract and Purpose Statement
Every contract has three parts…

```
# function name :: domain -> range
# what does the function do?
```

Examples
Write some examples, then circle and label what changes…

```
examples:

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

function name (input(s)) is what the function produces
```

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

```
fun (variable(s)):

function name variable(s)

what the function does with those variable(s)
```

Directions: Define a function called `nametag`, which consumes a `Row` of the animals table and draws their name in purple, 10px letters. (Assume you have rows `animalA` and `animalB` defined.)

Contract and Purpose Statement
Every contract has three parts…

```
# function name :: domain -> range
# what does the function do?
```

Examples
Write some examples, then circle and label what changes…

```
examples:

nametag (animalA) is what the function produces

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

function name (input(s)) is what the function produces
```

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

```
fun nametag (r)

text(r["name"], 10, "purple")

what the function does with those variable(s)
```

end
Using Table Methods

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. (E.g., `shapes.row-n(2)`.)
- Their contracts are different: they include the type of the data as part of their names. (E.g., `<table>.row-n :: (index :: Number) -> Row`.)
- They have a "secret" argument, which is the data they are attached to.
- In this course, the methods we’ll be using are `row-n`, `order-by`, `filter`, and `build-column`. 

Make sure you have the "Table Methods Starter File" open on your computer, and click "Run".

<table>
<thead>
<tr>
<th></th>
<th>Reading Function Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How many functions are defined here?</td>
</tr>
<tr>
<td>2</td>
<td>What are their names?</td>
</tr>
<tr>
<td>3</td>
<td>What is the domain of is-dog?</td>
</tr>
<tr>
<td>4</td>
<td>What is the range of is-old?</td>
</tr>
<tr>
<td>5</td>
<td>What is the range of get-name?</td>
</tr>
<tr>
<td>6</td>
<td>What does is-fixed(animalA) evaluate to?</td>
</tr>
<tr>
<td>7</td>
<td>What does get-name(animalB) evaluate to?</td>
</tr>
<tr>
<td>8</td>
<td>What does is-old(animalA) evaluate to?</td>
</tr>
<tr>
<td>9</td>
<td>What does is-dog(animalA) evaluate to?</td>
</tr>
<tr>
<td>10</td>
<td>What does is-fixed do?</td>
</tr>
<tr>
<td>11</td>
<td>What does get-name do?</td>
</tr>
<tr>
<td>12</td>
<td>What does is-old do?</td>
</tr>
</tbody>
</table>
The Design Recipe

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

**Directions:** Define a function called `get-fixed`, which looks up whether or not an animal is fixed.

### Contract and Purpose Statement

Every contract has three parts...

```
# get-fixed :: (r :: Row) -> Boolean
# Consumes an animal, and looks up the value in the fixed column.
```

### Examples

Write some examples, then circle and label what changes...

```
examples:

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

end
```

### Definition

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

```
fun get-fixed(r):
  r["fixed"]
end
```

**Directions:** Define a function called `get-gender`, which consumes a Row of the animals table and looks up the gender of that animal.

### Contract and Purpose Statement

Every contract has three parts...

```
# :: ->
# what does the function do?
```

### Examples

Write some examples, then circle and label what changes...

```
examples:

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

end
```

### Definition

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

```
fun get-gender:
  what the function does with those variable(s)
end
```
The Design Recipe

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

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

### Contract and Purpose Statement
Every contract has three parts…

```
# is-cat :: (r :: Row) -> Boolean
# Consumes an animal, and computes whether the species == "cat"
```

### Examples
Write some examples, then circle and label what changes…

```
examples:

is-cat (animalA) is

end
```

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

```
fun is-cat (r):
  r["species"] == "cat"
end
```

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

### Contract and Purpose Statement
Every contract has three parts…

```
# :: ->
# domain range
```

### Examples
Write some examples, then circle and label what changes…

```
examples:

is-young (r) is

end
```

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

```
definition (r):
  r["age"] < 4
end
```
**Method Chaining**

Method chaining allows us to apply multiple method with less code: For example, instead of using multiple definitions, like this:

```java
with-labels = table.build-column("labels", nametag)
cats = with-labels.filter(is-cat)
cats.order-by("age", true)
```

We can use method-chaining to write it all on one line, like this:

```java
table.build-column("labels", nametag).filter(is-cat).order-by("age", true)
```

**Order Matters!** The methods are applied in the order they appear. For example, trying to order a table by a column that hasn’t been built will result in an error.
The Design Recipe

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

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

**Contract and Purpose Statement**

Every contract has three parts...

```
#  is-dog  ::  ( r  ::  Row )  ->  Boolean

# Consumes an animal, and computes whether the species == "dog"
what does the function do?
```

**Examples**

Write some examples, then circle and label what changes...

**Examples:**

```
is-dog  (   animalA   )  is

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

end
```

**Definition**

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

```
fun  is-dog  (   r   ):

   r["species"] == "dog"

what the function does with those variable(s)

end
```

**Directions:** Define a function called `is-old`, which consumes a Row of the animals table and computes whether it is more than 12 years old.

**Contract and Purpose Statement**

Every contract has three parts...

```
#  ::  ->  

# what does the function do?
```

**Examples**

Write some examples, then circle and label what changes...

**Examples:**

```
function name  (input(s))  is
what the function produces

end
```

**Definition**

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

```
fun  is-old  (   ):

what the function does with those variable(s)

end
```
The Design Recipe

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

**Directions:** Define a function called `is-female`, which consumes a Row of the animals table and returns true if the animal is female.

**Contract and Purpose Statement**

Every contract has three parts...

\[
\begin{array}{c}
\text{function name} \\
\text{domain} \\
\text{range} \\
\text{what does the function do?}
\end{array}
\]

**Examples**

Write some examples, then circle and label what changes...

**examples:**

\[
\begin{array}{c}
\text{function name} \\
\text{input(s)} \\
\text{what the function produces}
\end{array}
\]

**Definition**

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

\[
\begin{array}{c}
\text{fun} \\
\text{variable(s)} \\
\text{what the function does with those variable(s)}
\end{array}
\]

**Directions:** Define a function called `name-has-s`, which returns true if an animal’s name begins with the letter “s”

**Contract and Purpose Statement**

Every contract has three parts...

\[
\begin{array}{c}
\text{function name} \\
\text{domain} \\
\text{range} \\
\text{what does the function do?}
\end{array}
\]

**Examples**

Write some examples, then circle and label what changes...

**examples:**

\[
\begin{array}{c}
\text{function name} \\
\text{input(s)} \\
\text{what the function produces}
\end{array}
\]

**Definition**

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

\[
\begin{array}{c}
\text{fun} \\
\text{variable(s)} \\
\text{what the function does with those variable(s)}
\end{array}
\]
Chaining Methods

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

```py
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 from the shelter:

<table>
<thead>
<tr>
<th>name</th>
<th>gender</th>
<th>age</th>
<th>fixed</th>
<th>pounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Toggle&quot;</td>
<td>&quot;female&quot;</td>
<td>3</td>
<td>true</td>
<td>48</td>
</tr>
<tr>
<td>&quot;Fritz&quot;</td>
<td>&quot;male&quot;</td>
<td>4</td>
<td>true</td>
<td>92</td>
</tr>
<tr>
<td>&quot;Nori&quot;</td>
<td>&quot;female&quot;</td>
<td>6</td>
<td>true</td>
<td>35.3</td>
</tr>
<tr>
<td>&quot;Maple&quot;</td>
<td>&quot;female&quot;</td>
<td>3</td>
<td>true</td>
<td>51.6</td>
</tr>
</tbody>
</table>

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

1. `t.order-by("age", true)`
   - A Produces a table containing only Toggle and Maple

2. `t.filter(is-fixed)`
   - B Produces a table of only young, fixed animals

3. `t.build-column("sticker", nametag)`
   - C Produces a table, sorted youngest-to-oldest

4. `t.filter(is-young)`
   - D Produces a table with an extra column, named "sticker"

5. `t.filter(is-young).filter(is-fixed)`
   - E Produces a table containing Maple and Toggle, in that order

6. `t.filter(is-young).order-by("pounds", false)`
   - F Produces a table containing the same four animals

7. `t.build-column("label", nametag).order-by("age", true)`
   - G Won’t run: will produce an error

8. `t.order-by("gender", false)`
   - H Produces a table with an extra "label" column, sorted youngest-to-oldest
Chaining Methods 2: Order Matters!

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

```py
fun is-female(animal): animal["gender"] == "female" end
fun kilograms(animal): animal["pounds"] / 2.2 end
fun is-heavy(animal): animal["kilos"] > 25 end
```

The table below represents four animals from the shelter:

<table>
<thead>
<tr>
<th>name</th>
<th>gender</th>
<th>age</th>
<th>fixed</th>
<th>pounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Toggle&quot;</td>
<td>&quot;female&quot;</td>
<td>3</td>
<td>true</td>
<td>48</td>
</tr>
<tr>
<td>&quot;Fritz&quot;</td>
<td>&quot;male&quot;</td>
<td>4</td>
<td>true</td>
<td>92</td>
</tr>
<tr>
<td>&quot;Nori&quot;</td>
<td>&quot;female&quot;</td>
<td>6</td>
<td>true</td>
<td>35.3</td>
</tr>
<tr>
<td>&quot;Maple&quot;</td>
<td>&quot;female&quot;</td>
<td>3</td>
<td>true</td>
<td>51.6</td>
</tr>
</tbody>
</table>

Match each Pyret expression (left) to the description of what it does (right). **Note: one description might match multiple expressions!**

1. `t.order-by("kilos", true)`
   - **A** Produces a table containing Toggle, Nori and Maple, with an extra column showing their weight in kilograms

2. `t.filter(is-female)
  .build-column("kilos", kilograms)`
   - **B** Produces a table containing Maple, Nori and Toggle (in that order)

3. `t.build-column("kilos", kilograms)
  .filter(is-heavy)`
   - **C** Produces a table containing only Fritz.

4. `t.filter(is-heavy)
  .build-column("kilos", kilograms)`

5. `t.build-column("kilos", kilograms)
  .filter(is-heavy)
  .order-by("gender", true)`
   - **D** Won't run: will produce an error
   - **E** Produces a table containing only Fritz, with two extra columns.

6. `t.build-column("female", is-female)
  .build-column("kilos", kilograms)
  .filter(is-heavy)`
   - **F** Produces a table containing Maple and Fritz
Mood Generator

1) Open the Mood Generator starter file, and read through the code you find there. This code contains new programming that you haven’t seen yet! Take a moment to list everything you Notice, and then everything you Wonder…

<table>
<thead>
<tr>
<th>Notice</th>
<th>Wonder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2) Add another line of code to the definition, so that `mood("mad")` produces the same emoji as `mood("angry")`.

3) Add another example to the examples: section for "laughing", using the appropriate emoji.

4) Come up with some new moods, and add them to the code. Make sure you include examples:!

5) In your own words, how do `if`-expressions work in Pyret? Write your answer below.

6) Write down at least 2 ways you could use `if`-expressions when analyzing the Animals Dataset.
**Word Problem: species-color**

**Directions:** We want to generate a custom dot for our image-scatter-plot, such that every species gets a unique color. Write a function called `species-color`, which takes in a Row from the animals table and returns a solid, 5px circle using a color you’ve chosen.

**Contract and Purpose Statement**

Every contract has three parts…

# _function name_ :: _domain_ -> _range_

# what does the function do?

**Examples**

Write some examples, then circle and label what changes…

examples:

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

end

**Definition**

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

fun _function name_ (variable(s)):

what the function does with those variable(s)

end
Randomness and Sample Size

Computer Scientists may take **samples** that are subsets of a data set. If their sample is well chosen, they can use it to test if their code does what it’s supposed to do. However, choosing a good sample can be tricky!

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

**Grouped Samples** are a subset of a population in which each member of the subset was chosen for a specific reason.
1) Evaluate the `big-animals-table` in the Interactions Area. This is the complete population of animals from the shelter!

   The population is roughly 50% male and 50% female.

2) How close to these percentages do we get with random samples? Type each of the following lines into the Interactions Area and hit “Enter”.

   ```
   random-rows(big-animals-table, 10)
   random-rows(big-animals-table, 40)
   ```

3) What do you get? ____________________________________________

4) What is the contract for `random-rows`? ____________________________

5) What does the `random-rows` function do?

   ____________________________________________

6) In the Definitions Area, define `small-sample` and `large-sample` to be these two random samples.

7) Make a pie-chart for the animals in each sample, showing percentages of males and females. Note that the percentage of males in the entire population is ~50%.
   - The percentage of males in `tiny-sample` is ________________
   - The percentage of males in `small-sample` is ________________

   The population is 48.6% fixed and 51.4% unfixed.

8) How close to these percentages do we get with random samples? Make a pie-chart for the animals in each sample, showing the percentages of fixed and unfixed.
   - The percentage of fixed animals in `tiny-sample` is ________________
   - The percentage of fixed animals in `small-sample` is ________________

9) Which repeated sample gave us a more accurate inference about the whole population? Why?
## Grouped Samples from the Animals Dataset

Use method chaining to define the grouped samples below, using the helper functions that you've already defined: `is-old`, `is-young`, `is-cat`, `is-dog`, `is-female`, `get-fixed`, and `has-s-name`. We've given you the solution for the first sample, to get you started.

<table>
<thead>
<tr>
<th>Subset</th>
<th>The code to define that subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kittens</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>kittens = animals-table.filter(is-cat).filter(is-young)</td>
</tr>
<tr>
<td>Puppies</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>young-dogs =</td>
</tr>
<tr>
<td></td>
<td>animals-table.</td>
</tr>
<tr>
<td>Fixed Cats</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>fixed-cats =</td>
</tr>
<tr>
<td></td>
<td>animals-table.</td>
</tr>
<tr>
<td>Cats with “s” in their name</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>s-cats =</td>
</tr>
<tr>
<td></td>
<td>animals-table.</td>
</tr>
<tr>
<td>Old Dogs</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>old =</td>
</tr>
<tr>
<td></td>
<td>animals-table.</td>
</tr>
<tr>
<td>Fixed Animals</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>fixed =</td>
</tr>
<tr>
<td></td>
<td>animals-table.</td>
</tr>
<tr>
<td>Old Female Cats</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>old-cats =</td>
</tr>
<tr>
<td></td>
<td>animals-table.</td>
</tr>
<tr>
<td>Fixed Kittens</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>young-fixed-cats=animals-table.</td>
</tr>
<tr>
<td>Fixed Female Dogs</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>fixed-female-dogs=animals-table.</td>
</tr>
<tr>
<td>Old Fixed Female Cats</td>
<td>```</td>
</tr>
<tr>
<td></td>
<td>old-fixed-female-cats =</td>
</tr>
</tbody>
</table>

---

42
### Displaying Data

Fill in the tables below, then use Pyret to make the following displays. The first answer has been filled in for you.

1) A pie-chart showing the gender of kittens from the shelter.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kittens</td>
<td>gender</td>
<td></td>
</tr>
</tbody>
</table>

2) A bar-chart showing how many puppies are fixed or not

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

3) A pie-chart showing the how many heavy dogs are fixed or not.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

4) A histogram of the number of pounds that fixed cats weigh.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

7) A scatter-plot of puppies, using name as the labels, pounds as the x-axis, and age as the y-axis.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>

9) A scatter-plot of fixed cats, using name as the labels, pounds as the x-axis, and weeks as the y-axis.

<table>
<thead>
<tr>
<th>What Data?</th>
<th>Which Variable(s)?</th>
<th>What Display?</th>
</tr>
</thead>
</table>
Choosing Your Dataset

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

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

1. Is it **interesting**? This should be data you are curious about, that answers questions you’d want to ask. Pick a dataset you’re genuinely interested in, so that you can explore questions that matter to you!

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

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

<table>
<thead>
<tr>
<th>What do you NOTICE?</th>
<th>What do you WONDER?</th>
<th>Question Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lookup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compute</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can’t answer</td>
</tr>
</tbody>
</table>

1. This dataset is ______________________________, which contains _______ data rows.

2. Some of the columns are:

   a. ________________________, which contains __________________ data. Some example values from this column are: ________________________________.

   b. ________________________, which contains __________________ data. Some example values from this column are: ________________________________.
How can we define subsets? For a given row \( r \), what function will identify if that row is in the subset?

<table>
<thead>
<tr>
<th>Subset</th>
<th>A function that returns true if a row ( r ) is in the subset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
<tr>
<td></td>
<td>fun ____________________(r); end</td>
</tr>
</tbody>
</table>
The Design Recipe

Write helper functions for your dataset, which you can use to define subsets.

Directions: Define a function called ____________, which consumes a Row of the ____________ table and ____________.

Contract and Purpose Statement

Every contract has three parts…

# ____________ :: ____________ -> ____________

# function name domain range

what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

( ____________ ) is ____________

function name input(s) what the function produces

( ____________ ) is ____________

function name input(s) what the function produces

end

Definition

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

fun ____________ ( ____________ ):

function name variable(s)

what the function does with those variable(s)

end

Directions: Define a function called ____________, which consumes a Row of the ____________ table and ____________.

Contract and Purpose Statement

Every contract has three parts…

# ____________ :: ____________ -> ____________

# function name domain range

what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

( ____________ ) is ____________

function name input(s) what the function produces

( ____________ ) is ____________

function name input(s) what the function produces

end

Definition

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

fun ____________ ( ____________ ):

function name variable(s)

what the function does with those variable(s)

end
The Design Recipe

Write your own word problems below, and solve them using the Design Recipe.

Directions: Define a function called ______________________, which consumes a Row of the table and ______________________.

Contract and Purpose Statement

Every contract has three parts...

# __________:: __________ -> __________

# function name domain range

what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

_________ (_________) is __________

_________ (_________) is __________

end

Definition

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

fun __________ (_________):

_________ __________ __________

_________ __________ __________

what the function does with those variable(s)

end

Directions: Define a function called ______________________, which consumes a Row of the table and ______________________.

Contract and Purpose Statement

Every contract has three parts...

# __________:: __________ -> __________

# function name domain range

what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

_________ (_________) is __________

_________ (_________) is __________

end

Definition

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

fun __________ (_________):

_________ __________ __________

_________ __________ __________

what the function does with those variable(s)

end
Histograms

To best understand histograms, it’s helpful to contrast them first with bar charts.

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

- *Bar charts provide a visual representation of the frequency of values in a categorical column.*
- There’s no strict numerical way to order these bars, but *sometimes there’s an order* that makes sense. For example, bars for the sales of different t-Shirt sizes might be presented in order of smallest to largest shirt.

**Histograms** show the number of rows that fall within certain intervals, or “bins” on a horizontal axis. The more rows that fall within a particular “bin”, the taller the bar.

- *Histograms provide a visual representation of the frequencies (or relative frequencies) of values in a quantitative column.*
- Quantitative data **can always be ordered**, so the bars of a histogram always progress from smallest (on the left) to largest (on the right).
- When dealing with histograms, it’s important to select a good **bin size**. If the bins are too small or too large, it is difficult to see the shape of the dataset.

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

- In a **symmetric** data set, values are just as likely to occur a certain distance above the mean as below the mean.
- A data set that is **skewed left** and/or has low outliers has a few values that are unusually low. The histogram for a skewed left dataset has a few data points that are stretched out to the left (lower) end of the x-axis.
- A data set that is **skewed right** and/or high outliers means there are a few values that are unusually high. The histogram for a skewed right dataset has a few data points that are stretched out to the right (higher) end of the x-axis.
The Design Recipe

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

**Directions:** Define a function called `kilos`, which consumes a Row of the animals table and divides the pounds column by 2.2 to compute the animal’s weight in kilograms.

### Contract and Purpose Statement

Every contract has three parts…

```plaintext
# function name :: domain -> range
# what does the function do?
```

### Examples

Write some examples, then circle and label what changes…

**examples:**

```plaintext
function name (input(s)) is what the function produces
```

### Definition

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

```plaintext
fun (variable(s)):
```

### Directions

Define a function called `smart-dot`, which consumes a Row of the animals table and computes the image of a solid red circle using the weight as the radius.

### Contract and Purpose Statement

Every contract has three parts…

```plaintext
# function name :: domain -> Image
# Consumes an animal, and computes a solid red circle using the weight as the radius
# what does the function do?
```

### Examples

Write some examples, then circle and label what changes…

**examples:**

```plaintext
smart-dot (animalA) is what the function produces
```

### Definition

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

```plaintext
fun (variable(s)):
```

52
## Summarizing Columns

<table>
<thead>
<tr>
<th>name</th>
<th>species</th>
<th>age</th>
<th>pounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Sasha&quot;</td>
<td>&quot;cat&quot;</td>
<td>1</td>
<td>6.5</td>
</tr>
<tr>
<td>&quot;Boo-boo&quot;</td>
<td>&quot;dog&quot;</td>
<td>11</td>
<td>12.3</td>
</tr>
<tr>
<td>&quot;Felix&quot;</td>
<td>&quot;cat&quot;</td>
<td>16</td>
<td>9.2</td>
</tr>
<tr>
<td>&quot;Nori&quot;</td>
<td>&quot;dog&quot;</td>
<td>6</td>
<td>35.3</td>
</tr>
<tr>
<td>&quot;Wade&quot;</td>
<td>&quot;cat&quot;</td>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>&quot;Nibblet&quot;</td>
<td>&quot;rabbit&quot;</td>
<td>6</td>
<td>4.3</td>
</tr>
<tr>
<td>&quot;Maple&quot;</td>
<td>&quot;dog&quot;</td>
<td>3</td>
<td>51.6</td>
</tr>
</tbody>
</table>

1. How many cats are there in the table above?
2. How many dogs are there?
3. How many animals weigh between 0-20 pounds?
4. How many animals weigh between 20-40 pounds?
5. Are there more animals weighing 40-60 than 60-140 pounds?

The charts below are both based on this table. What is similar about them? What is different?

<table>
<thead>
<tr>
<th>Similarities</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Making Histograms

Suppose we have a data set for number of teeth in a group of 50 adults:

<table>
<thead>
<tr>
<th>Number of teeth</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>33</td>
</tr>
</tbody>
</table>

*Draw a histogram for the table in the space below.* For each row, find which interval (or “bin”) on the x-axis represents the right number of teeth. Then fill in the box so that the height of the box is equal to the sum of the counts that fit into that interval. One of the intervals has been completed for you.
Reading Histograms

Students watched 5 videos, and rated them on a scale of 1 to 10. While the average score for every video is the same (5.5), the shapes of the ratings distributions were very different! Match the summary description (left) with the histogram of student ratings (right). For each histogram, the x-axis is the score, and the y-axis is the number of students who gave it that score.

Most of the students were fine with the video, but a couple of them gave it an unusually low rating.

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

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

Students either really liked or really disliked the video.

Reactions to the video were all over the place: high ratings and low ratings and inbetween ratings were all equally likely.
Identifying Shape

Describe the shape of histograms on the left in complete sentences, using vocabulary like "Skewed Left", "Symmetric", or "Skewed Right".

1

2

3

4

5
The Shape of the Animals Dataset

Describe two of the histograms you made from your dataset.

1. I made a histogram, showing the distribution of __________________________ pounds __________________________ for column in your dataset
   animals from the shelter __________________________. your subset, e.g., “fixed dogs from the shelter”

2. I made a histogram, showing the distribution of __________________________ pounds __________________________ for column in your dataset
   animals from the shelter __________________________. your subset, e.g., “fixed dogs from the shelter”

3. What do you Notice and Wonder about these histograms? What shape do they have?

<table>
<thead>
<tr>
<th>What do you NOTICE?</th>
<th>What do you WONDER?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Shape of the My Dataset

Describe two of the histograms you made from your dataset.

1. I made a histogram, showing the distribution of [column name] for [your subset, e.g., "fixed dogs from the shelter"]

2. I made a histogram, showing the distribution of [column name] for [your subset, e.g., "fixed dogs from the shelter"]

3. In the table below, describe the histograms. Are they symmetric? Do they show left skewness and/or low outliers? Right skewness and/or high outliers?

<table>
<thead>
<tr>
<th>What do you NOTICE about these displays?</th>
<th>What do you WONDER about these displays?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measures of Center and Spread

There are three ways to measure the “center” of a dataset, to summarize a whole column of quantitative data using just one number:

- **The mean** of a dataset is the average of all the numbers.
- **The median** of a dataset is a value that is smaller than half the dataset, and larger than the other half.
- **The mode(s)** of a dataset is the value (or values) occurring most often.
  - In a symmetric data set, values are just as likely to occur a certain distance above the mean as below the mean, and the median and mean are usually close together.
  - A data set with left skew, and/or low outliers, has a few values that are unusually low, pulling the mean below the median.
  - A data set with right skew, and/or high outliers, means there are a few values that are unusually high, pulling the mean above the median.

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

- **The minimum** – the lowest value in the dataset
- **The first, or “lower” quartile (Q1)** – the middle of the lower half of values which separates the lowest 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 higher half of values which separates the second highest quarter from the highest quarter.
  - **The maximum** – the largest value in the dataset.

- **The five-number summary** can be used to draw a **box plot**.
# Summarizing Columns in the Animals Dataset

1) The column I chose to summarize is **pounds**.

## Measures of Center

The three measures of center for this column are:

<table>
<thead>
<tr>
<th></th>
<th>Mean (Average)</th>
<th>Median</th>
<th>Mode(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2) Since the mean is **compared to the median**, this suggests the shape is

[higher/lower/about equal]

[skewed right (or high outliers) / skewed left (or low outliers) / symmetric]

## Measures of Spread

My five-number summary is:

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Displaying Center and Spread with a Box Plot

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

From this summary and box-plot, I conclude:

---

---

---

---

---
Interpreting Spread

Consider the following dataset, representing the annual income of ten people (note: $65k means $65,000, etc.):

$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.

<table>
<thead>
<tr>
<th>Mean (Average)</th>
<th>Median</th>
<th>Mode(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Q1</th>
<th>Q2 (Median)</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Why it’s misleading</th>
</tr>
</thead>
<tbody>
<tr>
<td>“They’re rich! The average person makes more than $70k dollars!”</td>
<td></td>
</tr>
<tr>
<td>“It’s a middle-income list: the most common salary is $45k/yr!”</td>
<td></td>
</tr>
<tr>
<td>“This group is really diverse, with people making as little as 12k and as much as $280k!”</td>
<td></td>
</tr>
</tbody>
</table>
Identifying Shape

Describe the shape of box plots on the left in complete sentences, using vocabulary like "Skewed Left", "Symmetric", or "Skewed Right".

1

2

3

4

5
Shape of My Dataset

1) The column I chose to summarize is ____________________________.

Measures of Center

The three measures of center for this column are:

<table>
<thead>
<tr>
<th>Mean (Average)</th>
<th>Median</th>
<th>Mode(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2) Since the mean is _______ compared to the median, this suggests the shape is _______.

   [higher/lower/about equal]

   [skewed right (or high outliers) / skewed left (or low outliers) / symmetric]

Measures of Spread

My five-number summary is:

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Q1</th>
<th>Q2 (Median)</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Displaying Center and Spread with a Box Plot

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

From this summary and box plot, I conclude:

__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
Data Scientists have to know whether or not they can trust their tools. Fortunately, then can use Data Science to verify that their tools do what they’re supposed to!
“Trust, but verify …”

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

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

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

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

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

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

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
A “helpful” Data Scientist gives you access to the following functions:

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

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

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

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

<table>
<thead>
<tr>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Visualizing Relationships

- **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. If the points are tightly clustered around a line, it is a **strong** correlation. If they are loosely scattered, it is a **weak** correlation.

- Points that are far above or below the cloud of points in a scatter plot are called **outliers**.

- We graphically summarize 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.
“Younger animals 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?
1. **For each row in the Sample Table on the left, add a point to the scatter plot on the right** 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 progress 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 tightly clustered around the line? Loosely scattered?
Identifying Form, Direction and Strength

Can you identify the Form, Direction, & Strength of these displays? If the form is linear, is the r-value closer to -1, 0, or +1?

A
Form: Linear  Non-Linear  None
Direction: Positive  Negative  None
Strength: Strong  Weak  None
r close to: -1  -0.5  0  +0.5  +1

B
Form: Linear  Non-Linear  None
Direction: Positive  Negative  None
Strength: Strong  Weak  None
r close to: -1  -0.5  0  +0.5  +1

C
Form: Linear  Non-Linear  None
Direction: Positive  Negative  None
Strength: Strong  Weak  None
r close to: -1  -0.5  0  +0.5  +1

D
Form: Linear  Non-Linear  None
Direction: Positive  Negative  None
Strength: Strong  Weak  None
r close to: -1  -0.5  0  +0.5  +1

E
Form: Linear  Non-Linear  None
Direction: Positive  Negative  None
Strength: Strong  Weak  None
r close to: -1  -0.5  0  +0.5  +1

F
Form: Linear  Non-Linear  None
Direction: Positive  Negative  None
Strength: Strong  Weak  None
r close to: -1  -0.5  0  +0.5  +1
1. There may be a correlation between __________________________ and __________________________. I think it is a correlation, because __________________________. It might be stronger if I looked at __________________________.

2. There may be a correlation between __________________________ and __________________________. I think it is a correlation, because __________________________. It might be stronger if I looked at __________________________.

3. There may be a correlation between __________________________ and __________________________. I think it is a correlation, because __________________________. It might be stronger if I looked at __________________________.
Computing Relationships

- **Linear Regression** is a way of computing the **line of best fit**, which minimizes the sum of 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 (response variable) to increase or decrease, for each unit that the x-variable (explanatory variable) increases.
  - **r** is the name of the **correlation statistic**, which is also computed by linear regression. The r-value will always fall between −1 and +1. The sign tells us whether the correlation is positive or negative, and distance from 0 tells us the strength of the correlation (−1 or +1 is really strong, 0 means no correlation).

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

- **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.
### Drawing Predictors

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

<table>
<thead>
<tr>
<th></th>
<th>Direction:</th>
<th>Strength:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>Positive</td>
<td>Strong</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>Positive</td>
<td>Strong</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>Positive</td>
<td>Strong</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>Positive</td>
<td>Strong</td>
</tr>
</tbody>
</table>

**Direction:** Positive, Negative, None  
**Strength:** Strong, Weak
# Interpreting Regression Lines & r-Values

Each description on the left is written about the linear regression findings on the right. Fill in the blanks using the information in the line of best fit and the r-value.

<table>
<thead>
<tr>
<th></th>
<th>For every additional Marvel Universe movie released each year, the average person is predicted to consume ___________ pounds of sugar! This correlation is ___________.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[y = -3.19x + 12] [r = -0.05]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Shoe size and height are ___________ correlated. If person A is one size bigger than person B, we predict that they will be roughly ___________ inches taller than person B as well.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>[y = 1.65x + 52] [r = 0.89]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>There is ___________ relationship found between the number of Uber drivers in a city and the number of babies born each year.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>[y = -15.3x + 1150] [r = 0.01]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>The correlation between weeks-of-school-missed and SAT score is ___________ and ___________. For every week a student misses, we predict a more than a ___________ point ___________ in their SAT score.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>[y = -5.35x - 16] [r = -0.65]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>There is a ___________ correlation between the number of streaming video services someone has, and how much they weigh. For each service, we expect them to be roughly ___________ pounds heavier.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>[y = 1.6x + 160] [r = 0.12]</td>
</tr>
</tbody>
</table>
Regression Analysis in the Animals Dataset

1. I performed a linear regression on a sample of cats from the shelter and found a moderate \((r=0.566)\), positive correlation between age of the cats (in years) and number of weeks to adoption. I would predict that a 1 year increase in age is associated with a 0.23 week increase in adoption time.

2. I performed a linear regression on a sample of dataset or subset and found a weak/strong/moderate \((R=\ldots)\), positive/negative correlation between the x-axis and the y-axis. I would predict that a 1 increase in the x-axis is associated with a \([\text{slope, y-units}]\) increase/decrease in the y-axis.

3. I performed a linear regression on a sample of dataset or subset and found a weak/strong/moderate \((R=\ldots)\), positive/negative correlation between the x-axis and the y-axis. I would predict that a 1 increase in the x-axis is associated with a \([\text{slope, y-units}]\) increase/decrease in the y-axis.
Regression Analysis in Your Dataset

1. I performed a linear regression on ______________________________ and found ______________________________ correlation between ______________________________ and ______________________________. I would predict that a 1 ______________________________ increase in ______________________________ is associated with a ______________________________ in ______________________________. ______________________________.

2. I performed a linear regression on ______________________________ and found ______________________________ correlation between ______________________________ and ______________________________. I would predict that a 1 ______________________________ increase in ______________________________ is associated with a ______________________________ in ______________________________. ______________________________.

3. I performed a linear regression on ______________________________ and found ______________________________ correlation between ______________________________ and ______________________________. I would predict that a 1 ______________________________ increase in ______________________________ is associated with a ______________________________ in ______________________________. ______________________________.
What’s on your mind?
Case Study: Ethics, Privacy, and Bias

1. Read the case study, and write your summary here.

____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________

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

____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________

3. What are the arguments on each side?

   *Data Science used for this purpose is good because*…

____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________

   *Data Science used for this purpose is bad because*…

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

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

2. **Study bias** - If someone is supposed to assess how much cat food is eaten each day on average, but they only measure how much cat food is put in the bowls (instead of how much is actually consumed), they'll end up with an over-estimate.

3. **Poor choice of summary** - Suppose a different shelter that had 10 animals recorded adoption times (in weeks) as 1, 1, 1, 7, 7, 8, 8, 9, 9, 10. Using the mode (1) to report what's typical would make it seem like the animals were adopted much quicker than they really were, since mean is 6.1 and median is 7.5.

4. **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 from the shelter longer” means “less likely to be adopted”.


Identifying Threats to Validity

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

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

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

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

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

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________
Identifying Threats to Validity

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

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

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

What are some possible threats to the validity of this conclusion?
## Fake News!

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

<table>
<thead>
<tr>
<th>Data</th>
<th>Claim</th>
<th>What's Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 The average player on a basketball team is 6'1&quot;.</td>
<td>“Most of the players are taller than 6’.”</td>
<td></td>
</tr>
<tr>
<td>2 Linear regression found a positive correlation (r=0.18) between people's height and salary.</td>
<td>“Higher salaries can make people taller!”</td>
<td></td>
</tr>
<tr>
<td>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.”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 “According to this bar chart, Felix makes up a little more than 15% of the total ages of all the animals in the dataset.”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 “According to this histogram, most animals weigh between 40 and 60 pounds.”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Linear regression found a negative correlation (r = -0.91) between the number of hairs on a person’s head and their likelihood of owning a wig.</td>
<td>“Owning wigs causes people to go bald.”</td>
<td></td>
</tr>
</tbody>
</table>
# Lies, Darned Lies, and Statistics

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

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

<table>
<thead>
<tr>
<th>Data</th>
<th>Claim</th>
<th>Why it's wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Design Recipe

Directions:

Contract and Purpose Statement
Every contract has three parts…

```
# function name :: domain -> range
```

Examples
Write some examples, then circle and label what changes…

```
examples:

function name (input(s)) is what the function produces
```
end

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

```
fun __________(__________):
    function name variable(s)

what the function does with those variable(s)
```
end

Directions:

Contract and Purpose Statement
Every contract has three parts…

```
# function name :: domain -> range
```

Examples
Write some examples, then circle and label what changes…

```
examples:

function name (input(s)) is what the function produces
```
end

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

```
fun __________(__________):
    function name variable(s)

what the function does with those variable(s)
```
end
Design Recipe

Directions:

Contract and Purpose Statement
Every contract has three parts...

# function name :: domain --> range
# what does the function do?

Examples
Write some examples, then circle and label what changes...

eamples:

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

end

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

fun function name (variable(s)):

function name variable(s)

what the function does with those variable(s)

end

Directions:

Contract and Purpose Statement
Every contract has three parts...

# function name :: domain --> range
# what does the function do?

Examples
Write some examples, then circle and label what changes...

eamples:

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

end

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

fun function name (variable(s)):

function name variable(s)

what the function does with those variable(s)

end
Directions:

Contract and Purpose Statement

Every contract has three parts…
# :: ->

function name domain range
#
what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

function name ( ) is what the function produces
function name (input(s)) is what the function produces
function name (input(s)) is what the function produces
end

Definition

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

fun ( ):

function name variable(s)

what the function does with those variable(s)
end

Directions:

Contract and Purpose Statement

Every contract has three parts…
# :: ->

function name domain range
#
what does the function do?

Examples

Write some examples, then circle and label what changes…

examples:

function name ( ) is what the function produces
function name (input(s)) is what the function produces
function name (input(s)) is what the function produces
end

Definition

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

fun ( ):

function name variable(s)

what the function does with those variable(s)
end
Contracts

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`, it takes one input (a `Number`), and it evaluates to a `Number`. From the contract, we know `num-sqr(4)` will evaluate to a `Number`.

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>triangle</td>
<td><code>(side-length :: Number, style :: String, color :: String)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>circle</td>
<td><code>(radius :: Number, style :: String, color :: String)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>star</td>
<td><code>(radius :: Number, style :: String, color :: String)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>rectangle</td>
<td><code>(width :: Num, height :: Num, style :: Str, color :: Str)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>ellipse</td>
<td><code>(width :: Num, height :: Num, style :: Str, color :: Str)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>square</td>
<td><code>(size-length :: Number, style :: String, color :: String)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>text</td>
<td><code>(str :: String, size :: Number, color :: String)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>overlay</td>
<td><code>(img1 :: Image, img2 :: Image)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>beside</td>
<td><code>(img1 :: Image, img2 :: Image)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>above</td>
<td><code>(img1 :: Image, img2 :: Image)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>put-image</td>
<td><code>(img1 :: Image, x :: Number, y :: Number, img2 :: Image)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>rotate</td>
<td><code>(degree :: Number, img :: Image)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>scale</td>
<td><code>(factor :: Number, img :: Image)</code></td>
<td>-&gt; Image</td>
</tr>
<tr>
<td>string-repeat</td>
<td><code>(text :: String, repeat :: Number)</code></td>
<td>-&gt; String</td>
</tr>
<tr>
<td>string-contains</td>
<td><code>(text :: String, search-for :: String)</code></td>
<td>-&gt; Boolean</td>
</tr>
<tr>
<td>num-sqr</td>
<td><code>(n :: Number)</code></td>
<td>-&gt; Number</td>
</tr>
<tr>
<td>num-sqrt</td>
<td><code>(n :: Number)</code></td>
<td>-&gt; Number</td>
</tr>
<tr>
<td>num-min</td>
<td><code>(a :: Number, b:: Number)</code></td>
<td>-&gt; Number</td>
</tr>
<tr>
<td>num-max</td>
<td><code>(a :: Number, b:: Number)</code></td>
<td>-&gt; Number</td>
</tr>
</tbody>
</table>
Contracts also help us use methods. The domain says it has one input (a function that consumes Rows and produces Booleans), and that the method evaluates to a Table.

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>(t :: Table, col :: String)</td>
<td>Table</td>
</tr>
<tr>
<td>random-rows</td>
<td>(t :: Table, num-rows :: Number)</td>
<td>Row</td>
</tr>
<tr>
<td>row-n</td>
<td>(t :: Table, col :: String, num-rows :: Number)</td>
<td>Row</td>
</tr>
<tr>
<td>order-by</td>
<td>(t :: Table, col :: String, increasing :: Boolean)</td>
<td>Table</td>
</tr>
<tr>
<td>filter</td>
<td>(test :: (Row -&gt; Boolean))</td>
<td>Table</td>
</tr>
<tr>
<td>build-column</td>
<td>(col :: String, builder :: (Row -&gt; Boolean))</td>
<td>Table</td>
</tr>
<tr>
<td>mean</td>
<td>(t :: Table, col :: String)</td>
<td>Number</td>
</tr>
<tr>
<td>median</td>
<td>(t :: Table, col :: String)</td>
<td>Number</td>
</tr>
<tr>
<td>modes</td>
<td>(t :: Table, col :: String)</td>
<td>List&lt;Number&gt;</td>
</tr>
<tr>
<td>bar-chart</td>
<td>(t :: Table, col :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>pie-chart</td>
<td>(t :: Table, labels :: String, values :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>box-plot</td>
<td>(t :: Table, col :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>bar-chart-summed</td>
<td>(t :: Table, labels :: String, values :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>pie-chart-summed</td>
<td>(t :: Table, labels :: String, values :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>modified-box-plot</td>
<td>(t :: Table, col :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>histogram</td>
<td>(t :: Table, values :: String, bin-width :: Number)</td>
<td>Image</td>
</tr>
<tr>
<td>scatter-plot</td>
<td>(t :: Table, labels :: String, xs :: String, ys :: String)</td>
<td>Image</td>
</tr>
<tr>
<td>image-scatter-plot</td>
<td>(t :: Table, xs :: String, ys :: String, f :: (Row -&gt; Image))</td>
<td>Image</td>
</tr>
<tr>
<td>r-value</td>
<td>(t :: Table, xs :: String, ys :: String)</td>
<td>Number</td>
</tr>
<tr>
<td>lr-plot</td>
<td>(t :: Table, labels :: String, xs :: String, ys :: String)</td>
<td>Image</td>
</tr>
</tbody>
</table>