**Name:**

**Classwork 4**

**Difficulties from Chapter 2**

* Putting data into DataCamp
* factor()
* head(select())
* Printing out just one of the variables from a data frame ($)
* filter(), !=, ==
* Making a new variable by combining already existing variables

**Exploring Variation in Our Data (Lecture)**

Today we’re going to look at our Head Sketch Survey data in the DataCamp Sandbox: <http://bit.ly/datacampsandbox>

1. First, we need to put this google spreadsheet into the DataCamp Sandbox. Then we can use our fancy R coding skills on this data. Make some notes here so that you will be able to do this yourself in lab.

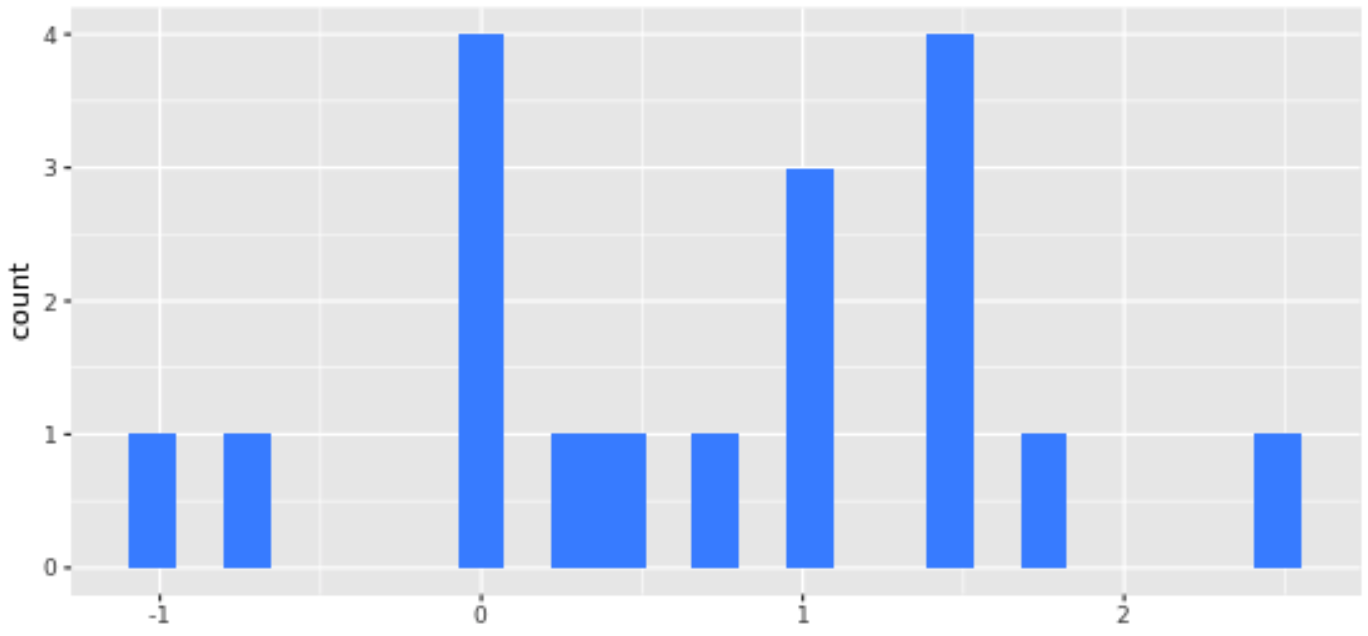
General code for importing csv file (more instructions on Page 2.6):

DataFrameName <- read.csv("https://url.com", header=TRUE)

1. We’ve got some choices here. What variable are you interested in exploring? What research question(s) do you have about that variable?
2. As a class, let’s explore variation in how much in the distance from eye to midline (EyesHigherMid).
3. Predict what that histogram might look like. Roughly sketch it below.

(b) Now let’s actually make the histogram in R. Write the code and roughly sketch it below. Make some observations (think: shape, center, spread, weird things).

1. The weird things that we see – what happened here? Should we filter the data in some way? How should we fix this?
2. Here is new histogram for EyesHigherMid here. Let’s also get the 5-number summary for this variable and label your histogram sketch below. Why aren’t the quartiles equally sized?



1. If a different class of college students measured their how much higher their eyes are above the midline, would histogram of their data be similar? Would it be more similar to 3b or 5?
   1. How would it be similar or different from 3b?
   2. How would it be similar or different from 5?

**Explaining Variation**

We start with variation in general. Part of our job as statisticians will be to “explain” the variation.

1. We talked about this a bit in the previous class, what do we mean to “explain variation”? (This is going to be our working definition but we’ll keep refining our understanding of this as we go.)
2. What are some variables that you might want to collect data on in the future if your goal was to “explain variation” in this variable, EyesHigherMid?

We won’t be able to explain all of the variation, so some of it will be unexplained variation, sometimes called “random variation,” or “error.”

1. The **random variation** will include a few different kinds of variation such as the types listed below.
   1. Not-yet-explained Variation:
   2. Measurement error:
   3. Sampling variation (think back to the many samples we drew for the Martin v. Westvaco simulation...)

**Different Types of Variables (Lab)**

1. Take a moment to import the csv file for the Head Sketch data into DataCamp Sandbox.
2. Which of these variables are categorical and which are quantitative? Which variables can we make a histogram for? Which ones should we make a bar chart for?
3. Should we tell R that some of these variables are categorical? Which ones? How would we do that?
4. Create a new variable for whether someone was correct or incorrect about where their eyes would fall relative to the midline. What should we name it? Write the code here.
5. How would you print out just that variable (not the whole data set but just that variable)?
6. How many people were correct about where their eyes would be? What percentage of people in our data were correct?
7. Do people know their face well? If we knew where a person *thought* his/her eyes would be (EyeThink), do you think we could make a better prediction about *how much higher* their eyes were relative to midline? Why or why not?
8. How would we explore our data to see if what they thought (EyeThink) helps us make a better prediction about their head size? Which measure of EyeHigherMid should we use (self or other)?