Data Visualization Lab

September 7, 2024

We will work through a series of questions in this lab, which will cover some of the topics we went over in the Introduction and Data Visualization slides. Fill in answers to the questions presented throughout this document. You may work together with others, but each student will need to submit their own version of this file to Canvas as a .pdf with all answers filled out.

Part 1 - Conceptual Questions

Question 1: I said statistics is the study of variability. Please explain that in your own words.

Question 2: What is a census? Why are censuses generally considered good? Give two reasons why we do not always want to conduct a census.

Question 3: What does it mean to say two variables are *associated* with each other?

Part 2 - Describing Data & Including Context

We have seen the terms *population, parameter, sample, statistic*, and *observation* in the Introduction. These terms are important for helping us describe data and understand what the purpose of a study is. Being able to read a summary of a study and label these individual parts is going to be an important skill we will use all semester. Our goal is to be able to communicate with each other.

When we are describing the *population*, the *sample*, and the *observations* in a study, we want to provide adequate context to explain the study and data.

To begin with, the topic that is going to be discussed and then the question(s) of interest should be stated or made clear at the beginning. Giving the audience an understanding of why the data exists will improve their understanding of the later description of the data. I want to stress this point. Context is critical in statistics and everything that is reported needs to be understood within that context.

For the techincal side, a generally accepted way is to use the 5 W's and HOW! This may be familiar to those of you who have taken media/journalism classes.

5 W's and H of Data

- **Who** – Who collected the data, who is the data collected on? This requires us to report how many "who's" we have (sample size).
- **What** – What variables was the data recorded on? That is, what pieces of information did we collect from our sample? Eg if we are interested in the voting population of the state of Iowa's views of gay marriage we would probably record if they support it, party affiliation, gender, age bracket, etc...

- **Where** – Where was the data collected? A poll done in Massachusetts offers limited information on the state of Iowa's views on gay marraige.
- **When** – When was the data collected? The voting population of the state of Iowa's views of gay marriage in 1995 offers limited information for today's views.
- **Why** – Why was the data collected? What research question(s) were the investigators trying to answer? This should be answered in the note on context above.
- **How** – How was the experiment ran? How was the data physically collected? How was a random sample randomly sampled?

We may not always use all of these in our own descriptions, but they are useful to add context to our data, and potentially see if there are any issues with the study. It is incorrect to believe that one needs to just respond to each question I presented above. Those are examples of common things that researchers convey to their audience that you need to cognizant about.

Describing Studies

Question 1: (Healthcare Opinions) In 2009, the PEW research group wanted to learn more about public opinion on the idea of the public option for health coverage. One thing that they wanted to know was the percentage of adult U.S. residents who favored a public option for health coverage in October 2009. In a poll of 1500 randomly selected Adult residents in the United States, they found that 55% of adult residents favored a government health insurance plan to compete with private plans. Source

- What is the point of the study? Why did they do this poll?
- Describe the population in this study:
- Describe the sample in this study:
- Describe an observation in this study:
- What is the variable of interest in this study? Is it categorical or quantitative?
- Do you think this data is useful for learning about healthcare opinions in 2024?

Question 2: (National household size) The American Community Survey (ACS) conducts yearly surveys. One thing that is of interest is the average household size. In April 2022, the ACS had surveyed 1,980,550 U.S. households and found the average household size to be 2.50. Source

- Describe briefly why the American Community Survey is ran. Hint: you may google the survey to learn more about it and it's goals.
- Describe the population in this study:
- Describe the sample in this study:
- Describe an observation in this study:
- What is the variable of interest in this study? Is it categorical or quantitative?

Question 3: (Consulting Problem) Making batteries last longer is an economically profitable area and many experiments occur with them. At Iowa State 64 laptop-style batteries were produced using the same standardized process. The cells were then either "slowly discharged" where the battery is slowly drained of power or "quickly discharged" where the battery is rapidly drained of power. All batteries would then be recharged and the process repeated. This continued for each battery until it fell below an acceptable operating condition (ie wouldn't hold a strong enough charge). The total number of charges/discharges for each battery was recorded.

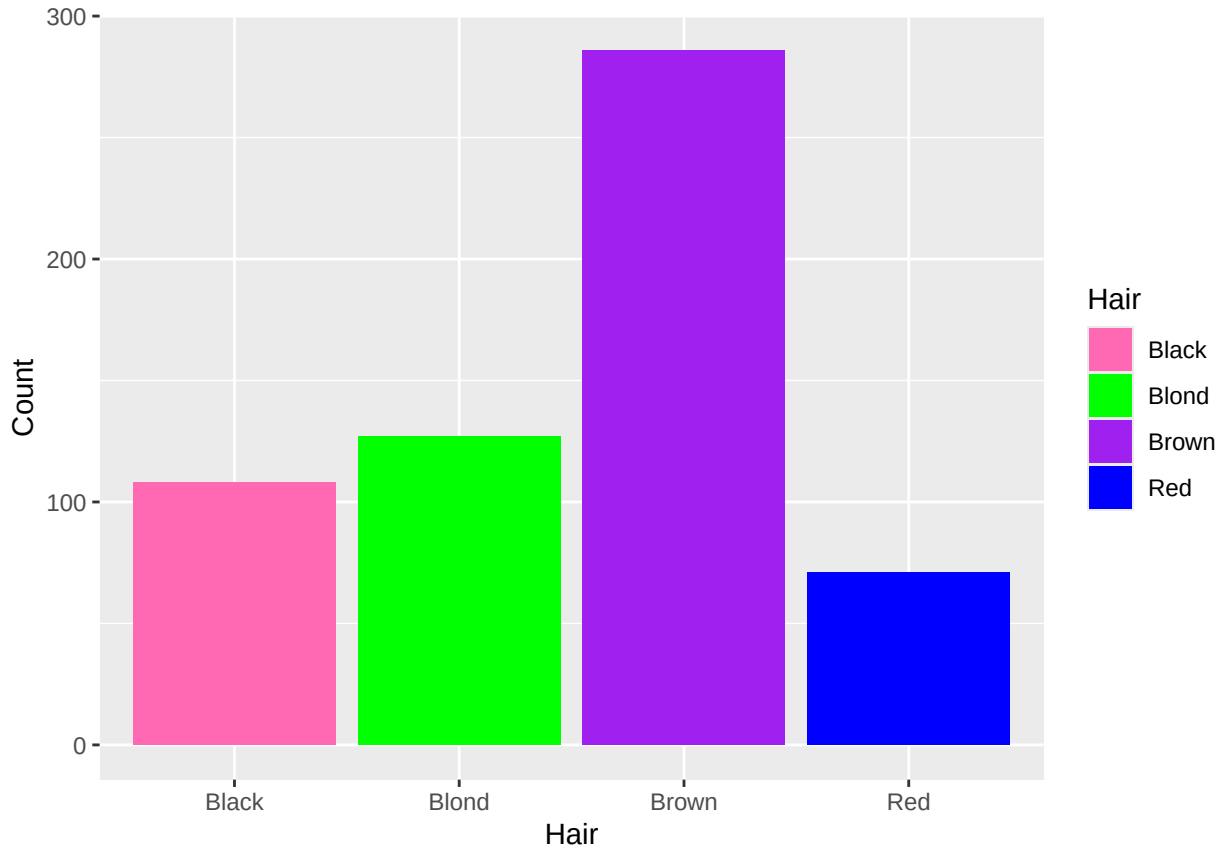
- Describe an observation in this study:
- Describe the sample in this study:
- Describe the population in this study:
- What question do you think the researchers were trying to answer?
- What are the two variables that are being collected?

Part 3 - Distributions

The *distribution* of a variable is a description of how frequently values of that variable show up. We saw that the way in which we describe the distribution of a variable is different depending on if the variable is categorical or quantitative.

Question 1: Below is a bar chart representing the hair color of students in a statistics class. Describe the distribution of the haircolor variable.

(In the graph code chunks I have put the term 'echo=FALSE' in the brackets. This stops RStudio from showing the code in the pdf to save a little bit of space). We will talk more about how to make these graphs on Wednesday (possibly Friday)



Question 2: Comment on the appropriateness of the color choices in the graph. Then, try to edit the code to make the colors seem more natural.

Question 3: Comment on whether you believe that distribution is representative of the class and the college in general. Then, go skim the help page for HairEyeColor (it is a data set that comes with R, but, despite not being a function, it has it's own help page). Comment on any piece of information presented about the data that would be relevant to any possible discrepancies.

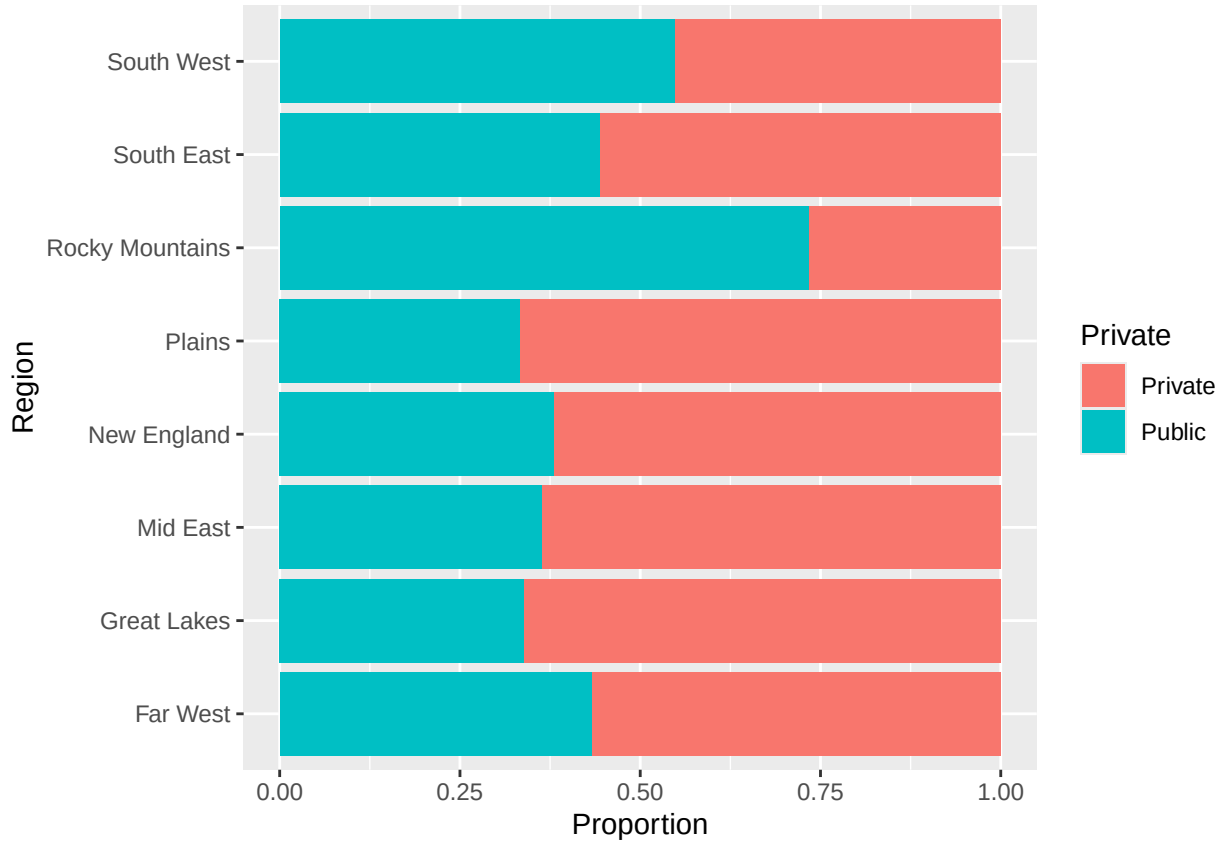
Part 4 - Relationships between Variables

For this set of questions we are going to use the College data set presented in the last few sets of slides. Read in the dataset for the College data using the following code.

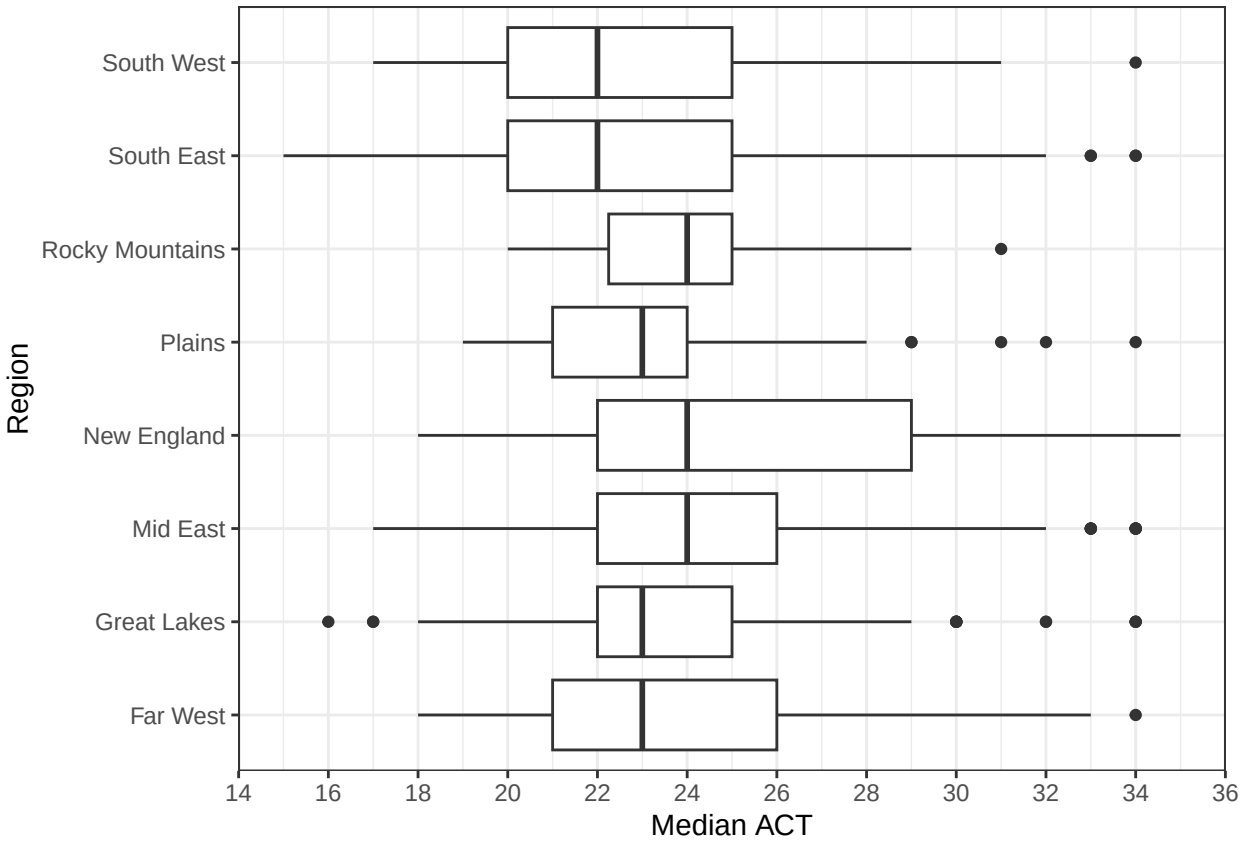
```
colleges <- read.csv("https://remiller1450.github.io/data/Colleges2019_Complete.csv")
```

Question 1: How many observations and variables are there in the dataset? Explain how you found this answer and show any code (if you used any).

Question 2: Look at the conditional bar chart below. Is there an *association* between the region and the type of college (public vs private) in our sample? Justify your answer using 1 or 2 sentences.



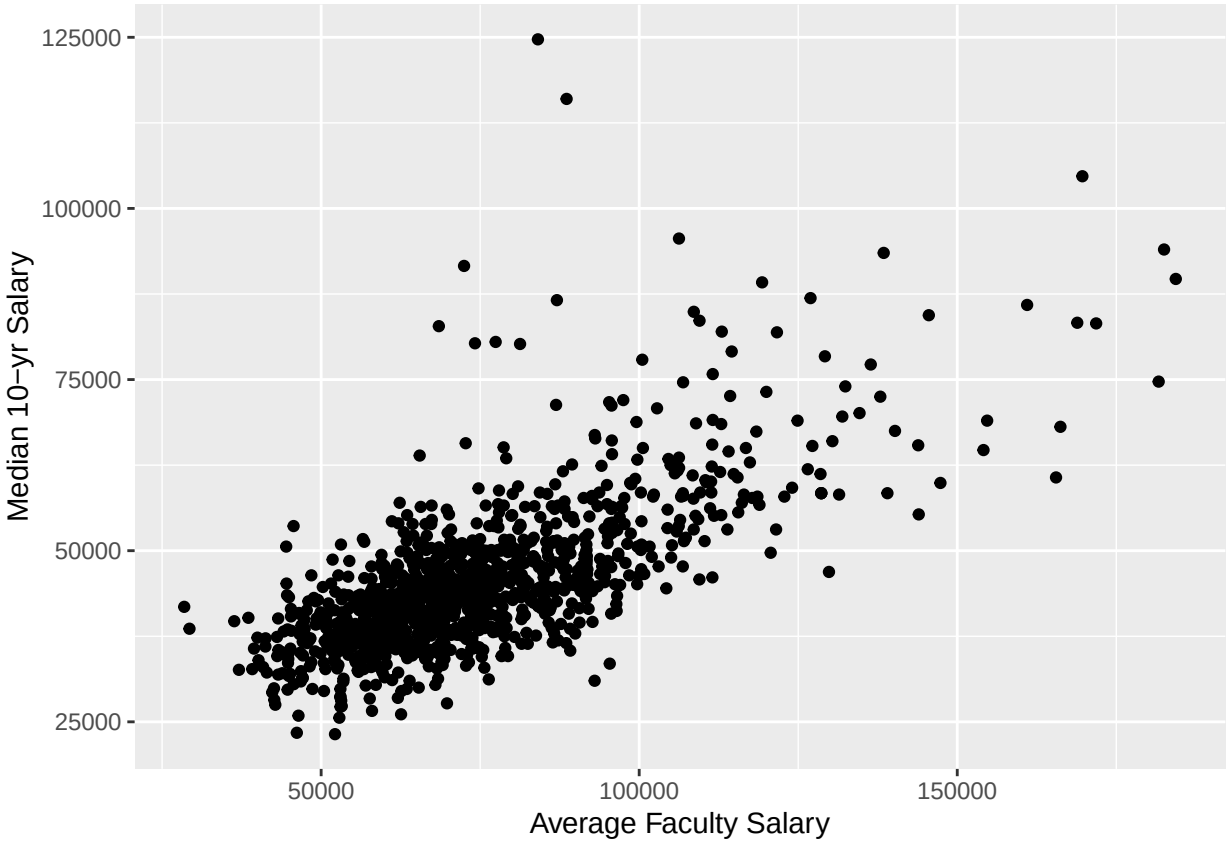
Question 3: Using the side-by-side box plots below, answer the following questions.



- What is the shape of 'South East's' box plot? What about 'Mid East'?
- Which region's boxplot has the largest median and what is the value of the median?
- Which region has the largest IQR? Give an approximate value of the IQR for this region and show your calculation

Question 4:

- When we describe scatterplots, we need to talk about **form, strength, direction,** and **outliers**. Using the scatterplot below, describe the relationship between Average Faculty Salary and Median 10-year Salary (the median salary of graduates from the college 10 years after receiving their degree) for our sample of colleges. Use full sentences and include context.



Question 5: Below is another scatterplot similar to the one in Question 3, but it now includs information on whether the colleges are public or private. Is the relationship between Average Faculty Salary and Median 10-year Salary different for public and private colleges? *Briefly* explain (1 or 2 sentences).


125000
100000
75000
50000
25000
Median 10−yr Salary
50000
100000
150000
Average Faculty Salary
Private
Private
Public