

Data Visualization Lab

Daisy Cutter

2025-09-15

Part 1 - Conceptual Questions

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

“Statistics is the study of variability” means that we are studying the trends and patterns within the data we are collecting or are given, using the tools statisticians and other contributors to the field have designed. We are looking at visualizing, interpreting, and summarizing data to draw conclusions on what it could mean to answer questions we have on the dataset.

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.

A census is a large, complete study of an entire population sample. They are considered good because they are very thorough, and very extensive. They are not always done because they are 1: expensive, and 2: time consuming. This is why the US census only happens every 10 years.

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

Two variables being associated with each other means that there is a trend that can be noticed once the data is visualized or parsed in some way, either through a table or through a graph.

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?

This was done before the Affordable Care Act as a poll on healthcare options, given the debate on healthcare in the US at the time.

- Describe the population in this study:

The population is adult US residents.

- Describe the sample in this study:

The sample is 1500 randomly selected adult US residents.

- Describe an observation in this study:

An observation is whether they support one particular issue or not, for example, whether one individual wants to prevent healthcare reform or supports it strongly, or moderately. ☐

- What is the variable of interest in this study? Is it categorical or quantitative?

It is categorical data, not quantitative. ☐

- Do you think this data is useful for learning about healthcare opinions in 2024?

It is somewhat useful, but given that opinions change over time, it should be re-studied to see what the modern-day landscape looks like in healthcare, especially after the pandemic.

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:

The population is American households across US states, DC, and Puerto Rico. It is done more frequently than the US census.

- Describe the sample in this study:

The sample is around 1.98 million households.

- Describe an observation in this study:

One observation is the sex ratio, which is 97.9, which means that there are either slightly more women in the population or women were slightly more likely to respond. ☐

- What is the variable of interest in this study? Is it categorical or quantitative?

It is numerical data that is collected in categorical variables. ☐

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:

The observation would be the number of total charges/discharges for each battery. ☐

- Describe the sample in this study:

The sample is 64 laptop-style batteries.

- Describe the population in this study:

The population is battery cells, but there isn't necessarily a study population to draw from because it's not about people.

- What question do you think the researchers were trying to answer?

How quickly charging or discharging a battery will degrade it.

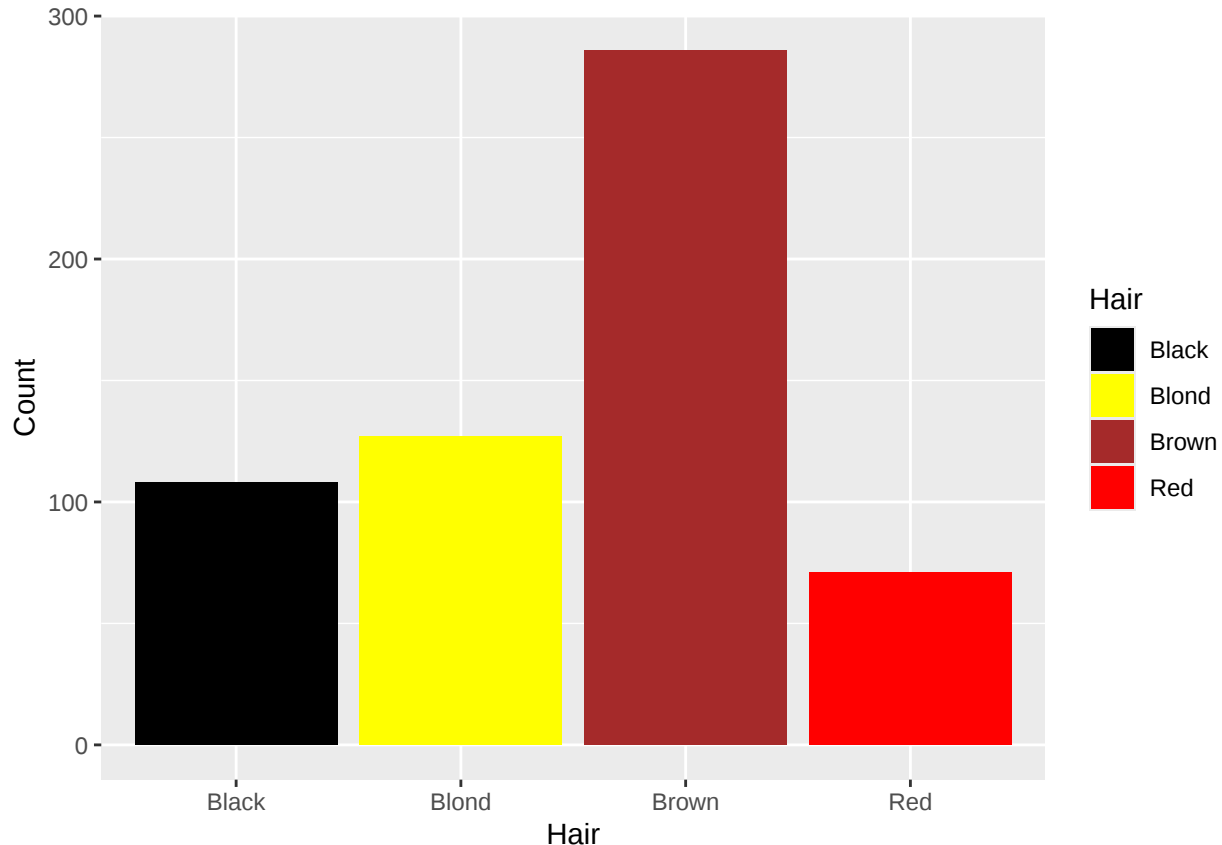
- What are the two variables that are being collected?

Whether it was discharged quickly or slowly and how long the battery lasts in terms of total number of charge cycles.

Part 3 - Distributions

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

Brown hair is the most common, followed by blond and black, then red.



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.

The colors are garish and do not match the subject matter. I have matched them to the colors in the data set.

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 its own help page). Comment on any piece of information presented about the data that would be relevant to any possible discrepancies.

I don't think it's representative of the college in general, maybe the class but definitely not the college. I feel that black hair is a bit more common and red hair is slightly more uncommon than that. It is an old piece of data, from 1974. College is much more diverse now, so this data set likely includes mostly white students.

Part 4 - Relationships between Variables

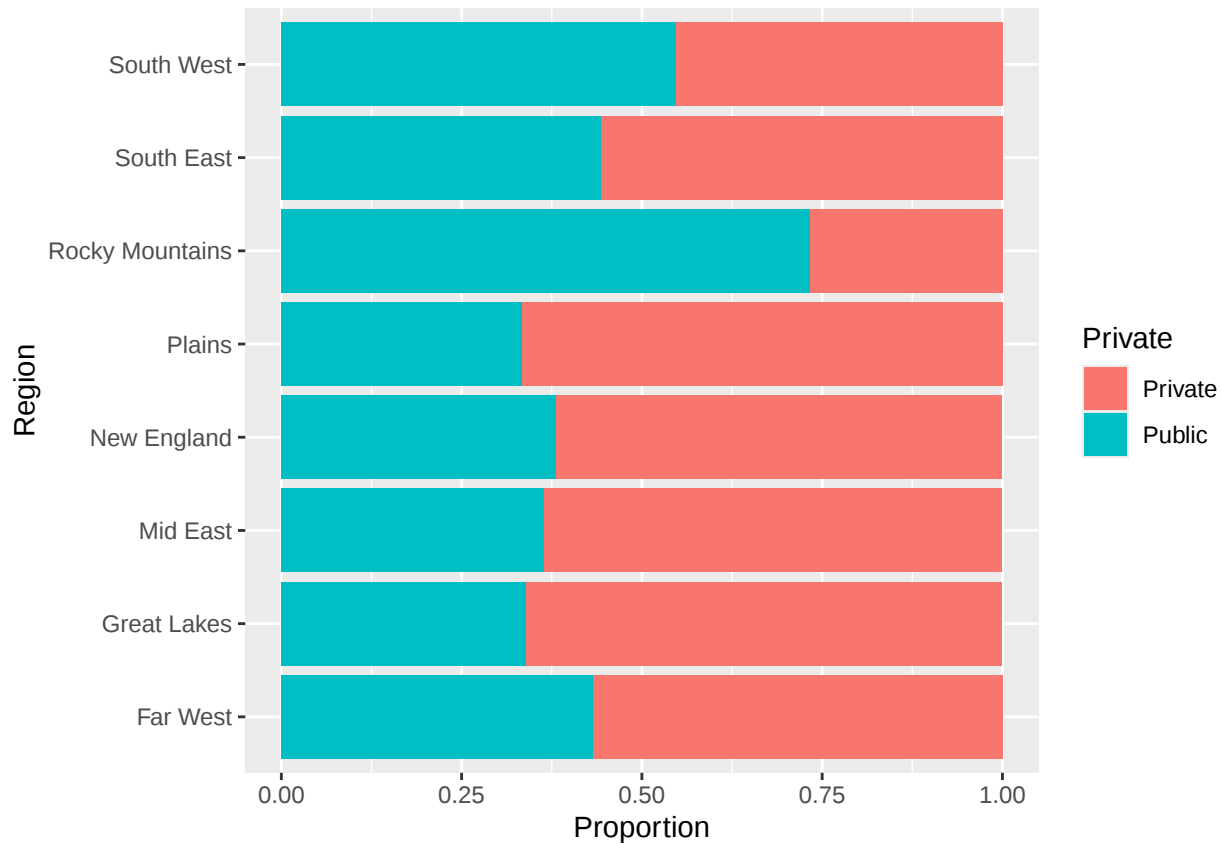
For this set of questions we are going to use the College data set presented in the last few sets of slides. Read in the dataset for the College data using the following code.

```
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).

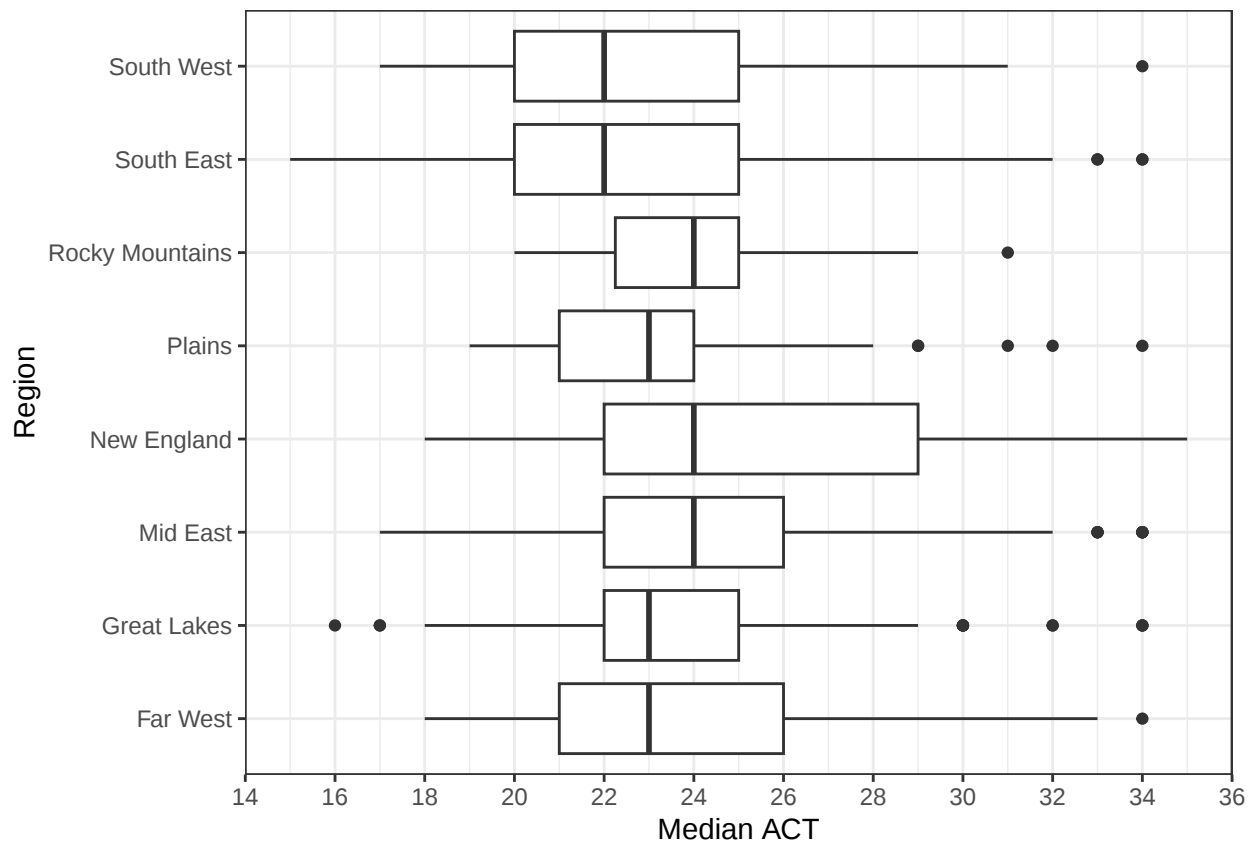
There are 1095 observations and 23 variables. This can be seen in the environment tab in R Studio.

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.



Yes, the Southwest, Southeast, and Rocky Mountains especially have more public schools than private.

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’?

The South East is skewed right. The Mid East is more central, but also skewed right. It has a smaller overall range and a smaller IQR.

- Which region’s boxplot has the largest median and what is the value of the median?

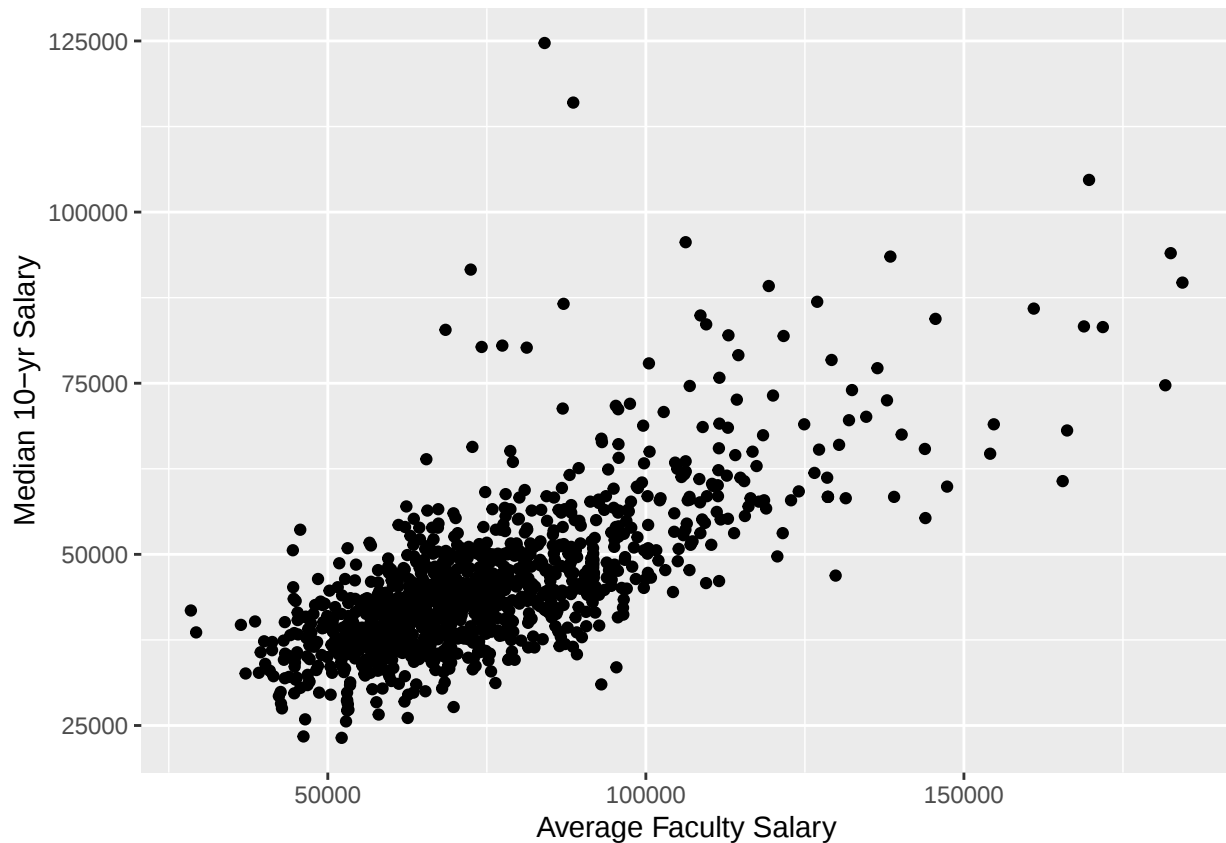
The highest median is tied between New England, the Mid East, and the Rocky Mountains, all at 24.

- Which region has the largest IQR? Give an approximate value of the IQR for this region and show your calculation

The largest IQR is New England, at $29 - 22 = 7$.

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.



The scatterplot has a positive direction, with outliers in the top right corner, showing that the average salary is between 50k to 100k, and a lower median 10-year salary, meaning the faculty are paid less and the students earn less. There is a medium association between median 10 year salary and average faculty salary, with some outliers, where faculty earn less than 100k but median 10-yr salary is very high.

Question 5: Below is another scatterplot similar to the one in Question 4, but it now includes 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).

We can see that private colleges are more likely to have higher 10 year median salaries and higher faculty salaries, and public schools have lower salaries for faculty overall. The outliers tend to be private schools.



