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 Gradescope as a .pdf with all answers filled out.

Part 1 - Conceptual Questions

Question 1: In the Introduction slides, I stated that 'Statistics is about *variation*.' In your own words, explain what I meant by this statement.

Data is around us, and it is always changing. In statistics we wish to explain how data varies. Hence we apply statistical methods to data and learn about the variations.

Question 2: What is a census? Give two reasons why we do not always want to conduct a census.

Census is a full collection of data for a population. It help us to find out a parameter of the population.

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

Two variables are associated with each other if the value of one variable is related to the value of the other variable.

Part 2 - Describing Data & Including Context

We have seen the terms *population, parameter, sample, statistic*, and *observation* in the Introduction slides. 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.

When we are describing the *population*, *sample*, and *observations* in a study, we want to provide adequate context to explain the study and data. The following are some things to consider when reading a description of a study.

5 W's and H of Data

- **Who** – Who collected the data, who is the data collected on? How many observations are there?
- **What** – What variables were data recorded on?
- **When** – When was the data collected? Populations can change over time and old data does not always reflect how things are now
- **Where** – Where was the data collected? Different geographical areas can have vastly different populations
- **Why** – Why was the data collected? What research question(s) were the investigators trying to answer?
- **How** – How was the data physically collected?

We may not always use all of these terms 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.

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

- Describe the population in this study:

All adult U.S. residents in October 2009

- Describe the sample in this study:

1500 randomly selected adult U.S. residents

- Describe an observation in this study:

One adult U.S. resident's response indicating whether they favored or did not favor a government health insurance plan

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

The variable is whether a person favors a government health insurance plan to compete with private plans. This is a categorical variable since responses are either favor or not favor.

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

No, because the data is collected in 2009.

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 the population in this study:

All U.S. households in April 2022

- Describe the sample in this study:

1,980,550 U.S. households

- Describe an observation in this study:

One household surveyed in the ACS is an observation in this experiment.

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

average household size, it is a quantitative variable.

Question 3: (Real Life Engineering Example) Forty prismatic lithium-ion pouch cells were built at the University of Michigan Battery Laboratory. Cells were formed using two different formation protocols: "fast formation" and "baseline formation". After formation, the cells were put under cycle life testing at room temperature and 45degC. Cells were cycled until the discharge capacities dropped below 50% of the initial capacities and the number of cycles was recorded.

- Describe an observation in this study:

The number of cycles a cell can withstand before hitting 50%% of inital capacity.

- Describe the sample in this study:

40 prismatic lithium-ion pouch cells built at the University of Michigan Battery Laboratory and tested under different conditions.

- Describe the population in this study:

All prismatic lithium-ion pouch cells.

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

Does the "fast formation" protocol improve the cycle life of lithium-ion pouch cells compared to the "baseline formation" protocol?

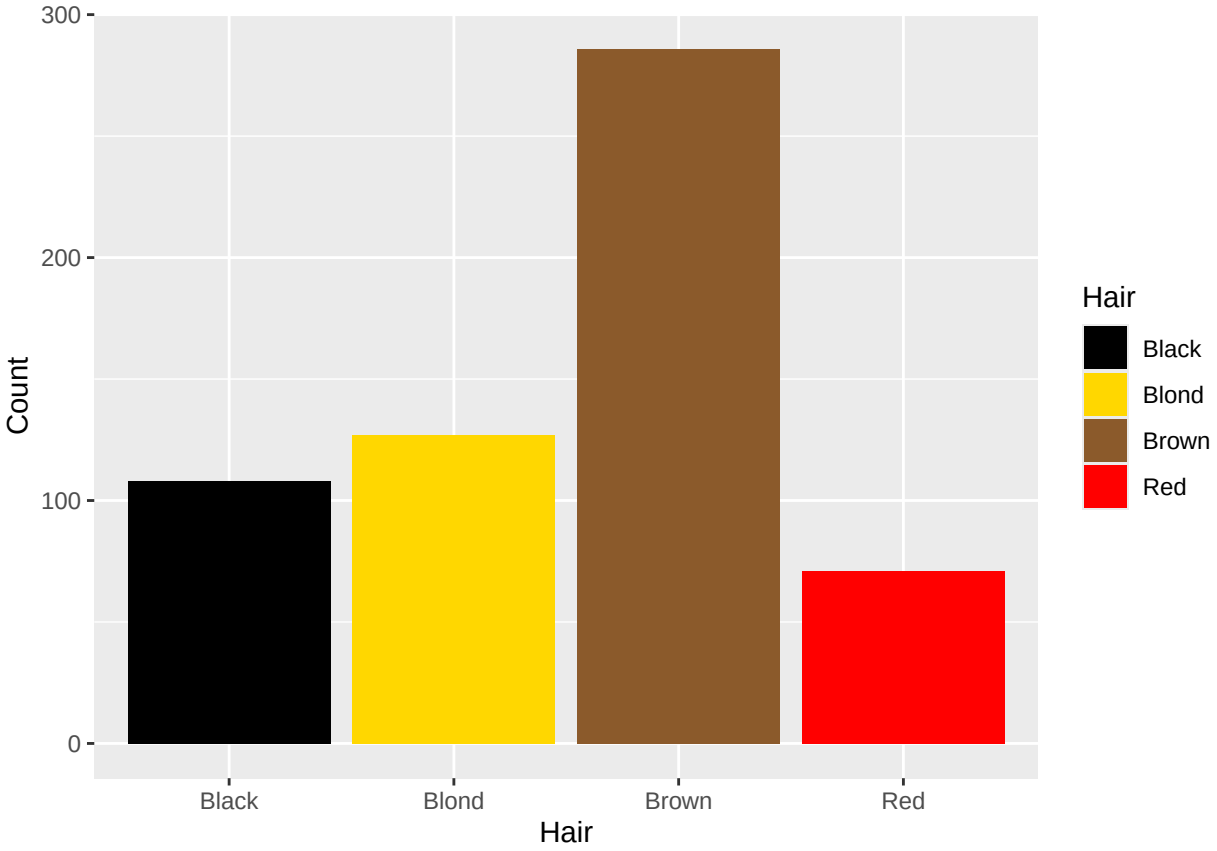
Part 3 - Distributions

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 (color may be exaggerated in the chart). Describe the distribution of the haircolor variable.

The distribution of hair color in the statistics class shows that Brown hair is the most common, with approximately 300 students. Black and Blond hair have similar frequencies, each between 100 and 150 students. Red hair is the least common, with fewer than 100 students. The data shows a notable preference for Brown hair in the class.

(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.



Note: We will come back to distributions of quantitative variables on Friday, after we have learned a bit more about describing histograms and box plots.

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).

```
nrow(colleges)
```

```
## [1] 1095
```

```
ncol(colleges)
```

```
## [1] 23
```

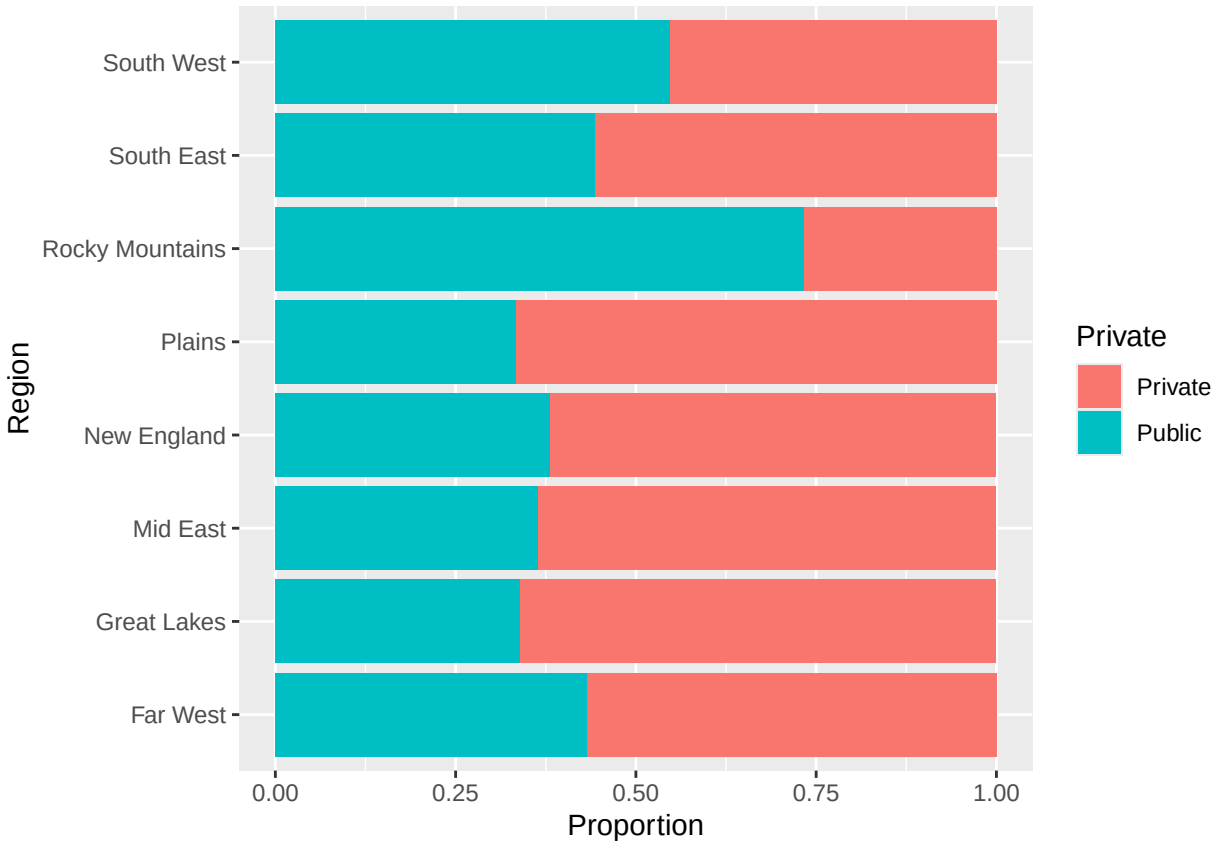
```
str(colleges)
```

```
## 'data.frame':    1095 obs. of  23 variables:
##  $ X                   : int  1 3 4 5 8 9 10 11 12 14 ...
##  $ Name                : chr  "Abilene Christian University" "Adelphi University" "Adrian College" "A
##  $ City                : chr  "Abilene" "Garden City" "Adrian" "Orlando" ...
##  $ State               : chr  "TX" "NY" "MI" "FL" ...
##  $ Enrollment          : int  3524 5307 1781 1166 4990 3903 857 1113 1512 3106 ...
##  $ Private             : chr  "Private" "Private" "Private" "Private" ...
##  $ Region              : chr  "South West" "Mid East" "Great Lakes" "South East" ...
##  $ Adm_Rate            : num  0.57 0.742 0.648 0.869 0.899 ...
##  $ ACT_median          : int  24 25 23 20 18 18 26 18 23 20 ...
##  $ ACT_Q1              : int  21 22 19 18 16 16 24 15 20 17 ...
##  $ ACT_Q3              : int  21 22 19 18 16 16 24 15 20 17 ...
##  $ Cost                : int  48046 49008 51626 24338 22489 21476 47221 47496 56722 23966 ...
##  $ Net_Tuition         : int  16177 24971 14136 15360 7413 10160 24852 12493 12849 2668 ...
##  $ Avg_Fac_Salary      : int  69804 111339 72873 69759 63909 69786 84078 65700 67968 61164 ...
##  $ PercentFemale       : num  0.612 0.721 0.422 0.825 0.564 ...
##  $ PercentWhite        : num  0.795 0.667 0.886 0.762 0.468 ...
##  $ PercentBlack        : num  0.0814 0.1785 0.0692 0.1395 0.4798 ...
##  $ PercentHispanic     : num  0.1635 0.1292 0.0318 0.1338 0.0379 ...
##  $ PercentAsian        : num  0.0287 0.0673 0.0121 0.0259 0.0148 ...
##  $ FourYearComp_Males  : num  0.412 0.611 0.232 0.476 0.147 ...
##  $ FourYearComp_Females: num  0.528 0.7 0.332 0.413 0.231 ...
##  $ Debt_median         : int  16000 19500 18468 16646 15000 18950 25000 23944 16721 20037 ...
##  $ Salary10yr_median   : int  43000 58500 38600 56000 31000 27700 124700 51200 47400 31100 ...
```

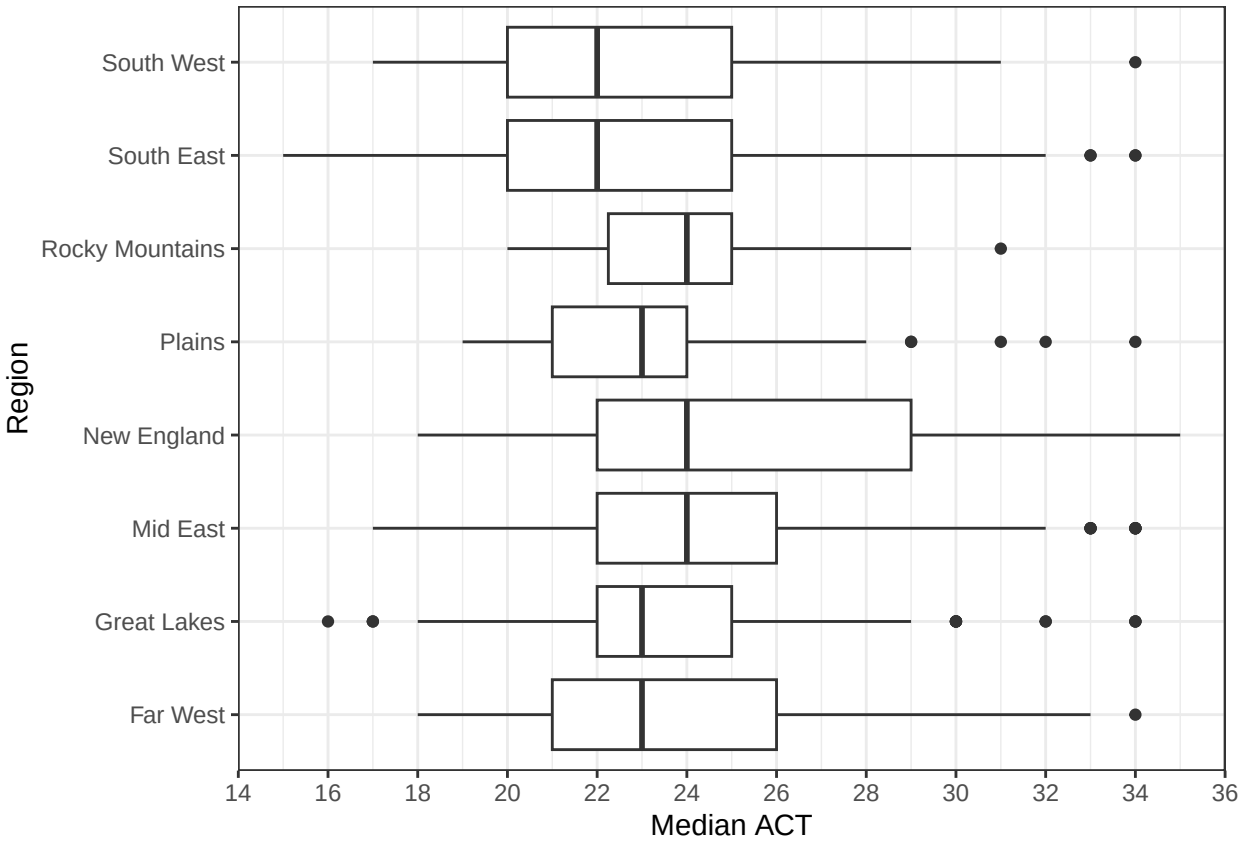
There are 1095 observations and 23 variables in the dataset. I find it by applying nrow() and ncol() to the dataset. And use str() to check the structure of the dataset.

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, there appears to be an association between region and the type of college in the sample. They vary significantly by region. Regions like the Rocky Mountains and Plains have a much higher proportion of public colleges, while regions such as New England have a higher proportion of private colleges.



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 shape of South East's box plot is slightly skewed to the right because the upper box is longer than the lower box. The shape of Mid East's box plot is roughly symmetric.

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

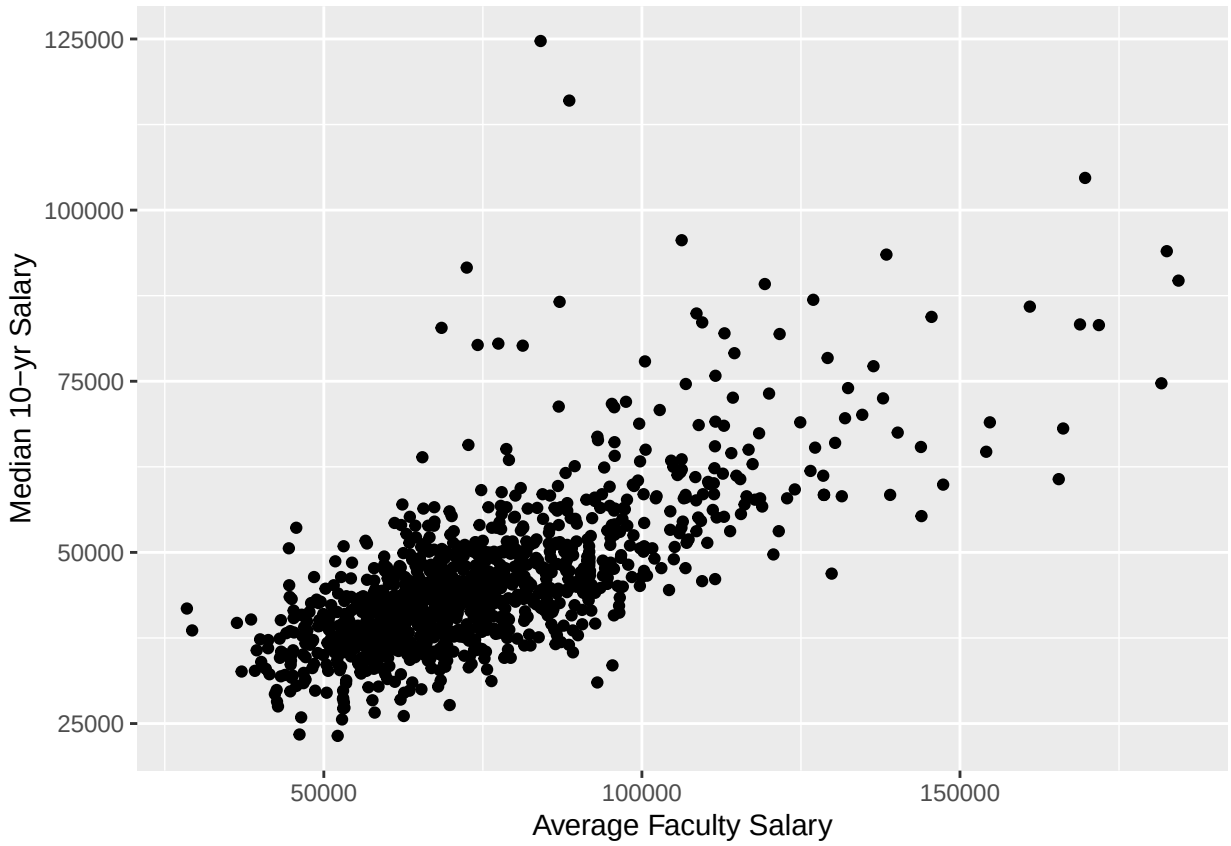
Rocky Mountains, New england, and mid east have the largest median, which is 24.

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

New England has the largest IQR, which is 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 relationship between Average Faculty Salary and Median 10-year Salary is a linear, moderate to strong relationship. As it increases, the median 10-year salary also increases hence it is a positive relationship. There is few outliers at upper of the plot.

Question 5: Below is another scatterplot similar to the one in Question 3, but I have added 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).

Yes it is different. The relationship between Average Faculty Salary and Median 10-year Salary have higher slope plot for private colleges, but slightly lower for public colleges. Also, we can see that the private colleges have weaker relationship compare to public colleges plots. It has more outliers and the points are more spread out.