ST 314 Data Analysis 03

Topics:
- Obtaining Data, Experiments and Bias
- One Variable Visual Displays and Summary Statistics for Categorical and Quantitative Variables.
- Lessons Covered: 14 - 19
- Textbook Chapter (Optional): 1

Grading:
- Points are listed next to each question and should total 25 points overall.
- Grading will be based on the content of the data analysis as well as the overall appearance of the document.
- Late assignments will not be graded.

Deadlines:
- Final Submission: **Thursday, October 22nd**. All submissions must be PDF files.

Instructions:
- Clearly label and **type answers** to the questions on the proceeding pages in Word, Google Docs, or other word processing software.
- Insert **diagrams or plots as a picture** in an appropriate location.
- Math Formulas need to be typed with Math Type, LaTeX, or clearly using key board symbols such as +, -. *, /, sqrt() and ^
- Submit assignment to the Gradescope link as a PDF. Indicate the pages to the individual questions and also verify the correct document has been uploaded. Failing to follow this direction may result in point deductions.

Allowances:
- You may use any resources listed or posted on the Canvas page for the course.
- You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
- Outside resources need to be cited. Websites such as Chegg, CourseHero, Koofers, etc. are discouraged, but if used need to be cited and used within the boundaries of academic honesty.

ST 314 Data Analysis 03

Part 1. (7 Points)
Suppose a company that manufactures paint claims in their advertisement:

"*In a randomized comparative experiment on drying time, our paint beat out the competition! We have the fastest drying paint!*"

Suppose the following is the actual data from the experiment:

Dry time in minutes for Manufacturer's Paint	33.3	29.1	35.6
Dry time in minutes for Competition's Paint	33.4	35.7	29.3

a. (1 point) What are the average dry times for each company's paint?

Manufacturer: 32.67 minutes
Competition: 32.8 minutes

b. (3 points) Is the manufacturers claim truthful? Either way, is the advertisement misleading? Why or why not?

Claim can be considered true, as on average, their paint dries faster than the competitions, by 0.13 minutes. However, it is not totally truthful, as in some instances, the manufacturer's paint dries slower than the competition's, so stating that the paint has the fastest drying time is misleading, as it implies that the manufacturer's paint always beats out the competition, which it does not.

c. (3 points) Suppose the company advertising the faster drying paint performed the experiment. Why could this be a potential problem?

There would be a definite conflict of interest there, as they want their paint to dry faster, as they advertised. Beyond conflict of interest, the manufacturer who made the advertising could set up the test to benefit their paint, on purpose or inadvertently, which could make their results questionable (i.e. test in a pure oxygen environment where the advertiser's solvent evaporates faster, ...)

ST 314 Data Analysis 03

Part 2. (18 Points)

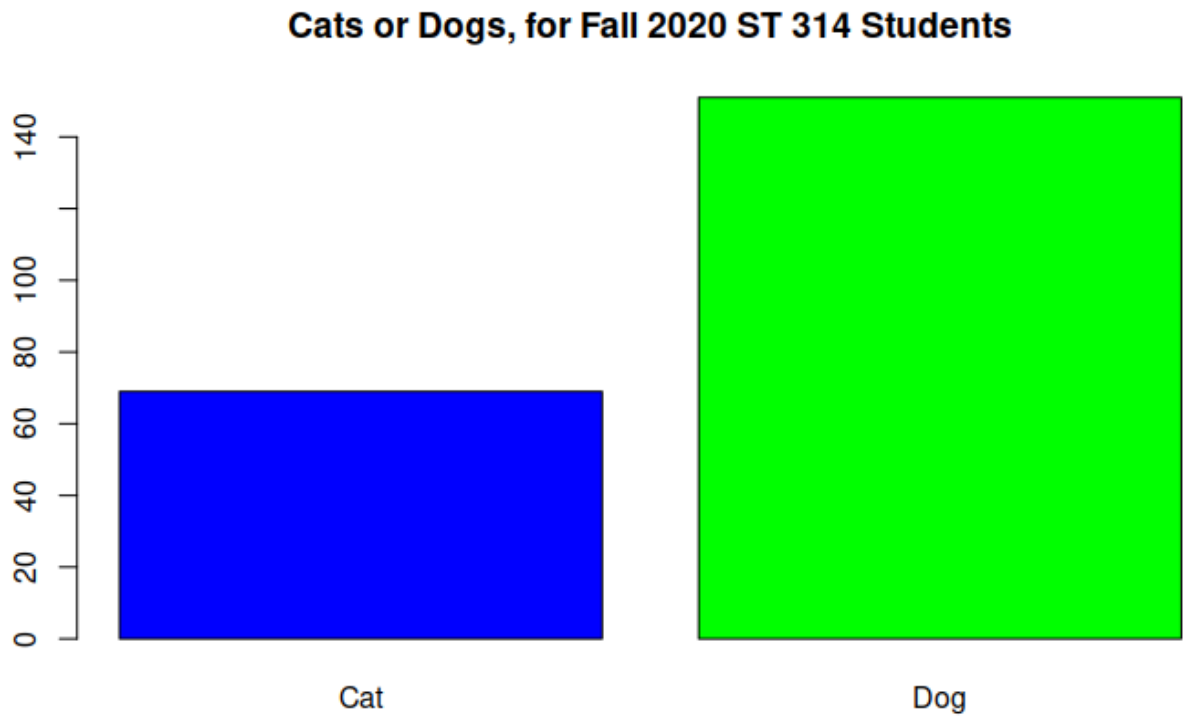
In this section, use the R script, One_Variable_Display_and_Summary_Stats.R and the ST314 student survey dataset, f20-st314-student-survey.csv to explore one categorical and one quantitative variable of your choice. Download the R script and the dataset, open the R script and follow the command instructions. Check out the dataset legend to see what variables represent. Then answer the following questions:

Categorical Variable

a. **(1 point) Choose a categorical variable to explore. Which variable did you choose?** ***Note: "SubjectPreferred" is off limits given this was my example. Choose something else.***

I chose variable CatsOrDogs

b. **(2 point) Paste the table of counts and bar chart for the categorical variable of your choosing. Include color and an appropriate title on your plot.**



c. **(2 point) Briefly, describe the distribution in context. Recall, categorical variables are summarized by counts and/or percents.**

Cats are less popular than dogs among the ST314 class of Fall 2020. Cats are the favorite animal when choosing between dogs and cats for about 31.37% of respondents, While dogs are the favorite when asked to choose between cats and dogs for 68.63% of respondents.

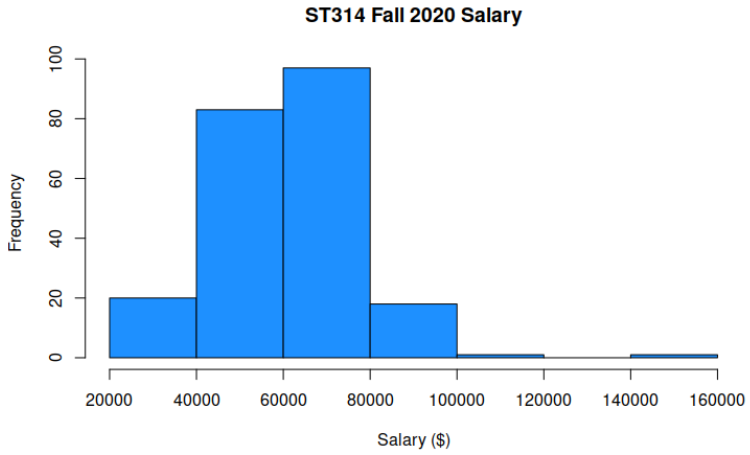
ST 314 Data Analysis 03

Quantitative Variable

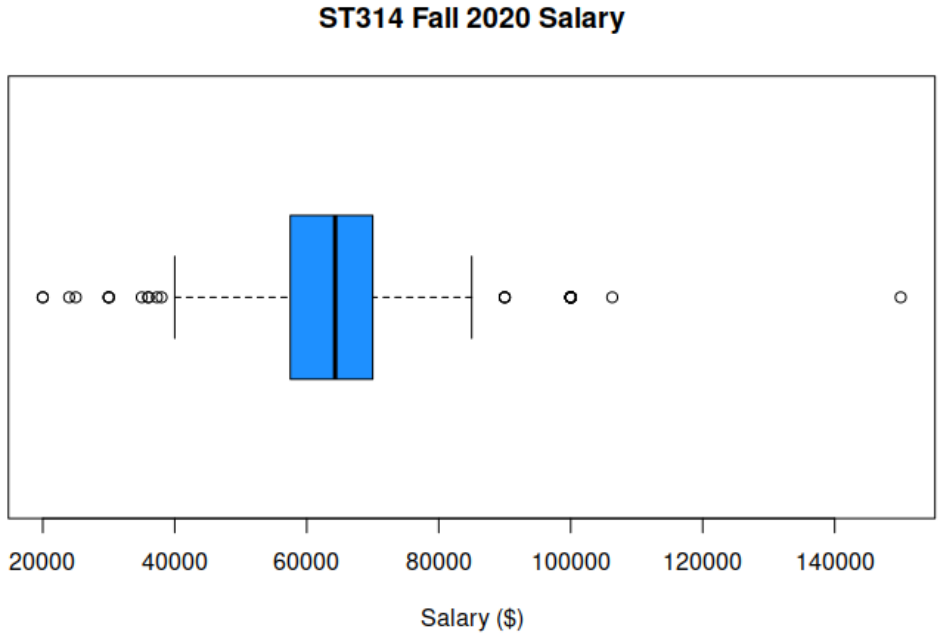
a. **(1 point) Choose a quantitative variable to explore. Which variable did you choose?** ***Note: "Email" is off limits given this was my example. Choose something else.***

I chose the variable Salary

b. **(2 point) Create a histogram of the variable. Include color and an appropriate title on your plot. Paste plot.**



c. **(2 point) Create a boxplot of the variable. Include color and an appropriate title on your plot. Paste plot.**



d. **(1 point) Which plot do you prefer to visualize the variable? Why?**

I prefer the visualization provided by the boxplot, as it shows outliers, and average. It contains more information in a given space, as opposed to the histogram which requires interpretation to get the same amount of information.

If I must choose between the horizontal and vertical box plot, I would choose horizontal, as it is intuitive to see that the further out to the right on the X axis we go, the larger the reported salary is (possibly because of habit).

e. **(2 points) Give a table that includes the mean, standard deviation, minimum, 1st quartile, median, 3rd quartile, maximum and IQR.**

Min	1st Qu.	Median	Mean	3rd Qu.	Max	IQR
20,000	57,500	64,308.4	64,308.4	70,000	150,000	57,500-70,000

f. **(3 points) Use the plots and summary statistics to describe the data in the context of the problem. Include the shape, center and spread in your description. State whether there are any outliers.**

The historgram is right skewed, as it has a long 'tail' to the right. This data represents response data for the Fall 2020 ST 314 Class. The data as presented shows that there are several outliers, notable all the values reported below $40000, and all those above about $85000 are disregarded. The distribution appears to be right shifted, when examined as a hisogram. The center is around $65000, and the spread includes about +/- $25000 from $65000.

The maximum entered value was $150000, and the minimum value entered was $20000. The mean and median of the dataset was $64308.40. The data had an IQR of $57,500-$70,000

g. **(2 points) Given the shape of the data which measure, the mean, median or either, would be a more appropriate to represent the center of the data? Explain your reasoning.**

The distribution of salaries is a right/positively skewed distribution, therefore the mean will be higher than the median, typically. In this case, the calculated mean and median are the same value, so I can choose either to be the center of the data, but were this not the case, we would use the median as the center of the data, as the outliers (in this case) to the right drag the mean up (to the right) faster than they do the median.