Winter 2019 – 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: **Monday, January 28th**. All submissions must be PDF files.

Instructions:
- Clearly label and **type answers** to the questions on the proceeding pages, ***without*** question prompts, 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 Canvas link as a PDF. Verify the correct document has been uploaded. If not, resubmit. You can submit up to three times.

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.

Winter 2019 – 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 Manufacturers Paint	33.3	29.1	35.6
Dry time in minutes for Competitions Paint	33.4	35.7	29.3

a. (1 point) What are the average dry times for each company's paint?
b. (3 points) Is the manufacturers claim truthful? Either way, is the advertisement misleading? Why or why not?
c. (3 points) Suppose the company advertising the faster drying paint performed the experiment. Why could this be a potential problem?

Part 2. (18 Points)

In this section, use the R script, `One_Variable_Display_and_Summary_Stats.R` and the ST314 student survey dataset, `ST314SISW19.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.*
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.

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

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.
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.