DATA 115: Introduction to Data Analytics
Week #5 Assignment
Fall 2022

Name: ____________________

Submit your work in a single .pdf compiled with knitr. The Homework Data folder has a starting .Rmd template with spaces for you to fill in your answers to the questions.

**1.** Create an account on GitHub (https://github.com) and create a repository for your personal dataset project. Submit the corresponding URL as for this problem.

**2.** In your own words, write brief definitions of

(a) Mean
(b) Median
(c) IQR
(d) Variance
(e) Skewness

**3.** Load the COL.csv dataset into R.

(a) Decide which rows are outliers in this data and describe and justify how you determined their outlier status.
(b) For each row you identified, if you were performing EDA on this dataset, would you include its values in your analysis and plots?
(c) Why or why not?

**4.** Load the Height_Weight_Age_Sex.csv data into R.

(a) Create boxplots for the height and weight columns separately. Comment on the symmetry and skewness, if any, for their distributions using these plots.
(b) Create histograms for the height and weight columns separately. Comment on the symmetry and skewness, if any, for their distributions using these plots. Are your conclusions based on the boxplots consistent with those based on densities?
(c) Create separate boxplots for the weight data separated by the Male variable. What do you observe about the two distributions?
(d) Add a BMI column to the data frame:

```
HWAS <- read.csv("./Height_Weight_Age_Sex.csv")

HWAS$BMI <- HWAS$weight/((HWAS$height/100)**2)

HWAS$underweight <- HWAS$BMI <18.5
```

(e) Create separate histograms for the BMI column separated by the Male variable. What do you observe about the two distributions?
(f) Make a scatterplot of height vs. weight for the full dataset that distinguishes both the Male variable and the under variable. What do you observe?

**5.** Read the following examples about Simpson's Paradox: How to lie with statistics? (you may also find the wikipedia page on the topic or the additional readings uploaded to Canvas to be helpful resources). Fill in the following table with ratios of hits to attempts so that player A has a higher batting average in both season 1 and season 2 but player B has a higher overall batting average for the two seasons combined.

| hits/attempts | Season 1 | Season 2 |
|---|---|---|
| Player A | 25/100 | ?/? |
| Player B | ?/? | ?/? |