**J. Jedediah Smith**

**BIFX 503-1**

**Homework Set #1**

**9/9/2021**

*Instructions:*

*Use R to complete this assignment.*

*Assignment is to be submitted via Blackboard.*

Use the R dataset **airquality** to answer all questions.

1. Get familiar with the dataset by using ?, str(), and head().

Text

Description automatically generated

1. Plot a histogram for ozone, temperature, wind speed, and solar radiation. Describe each distribution – is it symmetric or skewed? Do there seem to be outliers?

Ozone is skewed right and has clear outliers. Temperature is almost symmetric and has possible outliers. Wind is more symmetric than skewed, but has a funny dip in the middle, as well as clear outliers. Solar radiation is skewed left, no outliers.

Chart, histogram

Description automatically generated

1. Generate summary statistics for each variable using summary(). Which variable has the most missing values?

Ozone has the most missing data with 37 NAs. Solar radiation is next highest with 7 NAs. Wind and temperature do not appear to have any NAs.

1. Generate side-by-side box plots showing the distribution of each of these variables separately by month. How does each appear to vary by month?

Ozone median stays around 25 during months 5, 6, and 9, but skyrocket up to around 50 during months 7 and 8. Temperature median is usually between 75 and 85, but plummets to around 67 during month 5. Wind is more stable, with the median staying between 9 and 12, with low points during months 7 and 8. Solar radiation median is usually about 200, but jumps to 250 during month 7.

Diagram, schematic

Description automatically generated

1. Calculate the mean and standard deviation of each of these variables separately by month.

Mean by Month Standard Deviation by Month

A picture containing text

Description automatically generatedA picture containing text

Description automatically generated

1. Create a subset of the data that contains only the variables Ozone, Solar.R, Wind, and Temp. Calculate the correlation matrix for these four variables, using pairwise complete observations. Which variables are positively correlated, and which are negatively correlated? Which variables have a correlation that is greater (in either the positive or negative direction) than 0.4?

The positively correlated pairs are Ozone and Solar Radiation, Temperature and Ozone, Temperature and Solar Radiation. The negatively correlated pairs are Wind and Ozone, Wind and Solar Radiation, Wind and Temperature. The pairs with greater than 0.4 in either direction are Wind and Ozone, Wind and Temperature, Temperature and Ozone.

Text

Description automatically generated

1. Create a subset of the data that contains only measurements made in May and September. Perform independent samples t tests to determine if these months differ in terms of ozone, solar radiation, wind, or temperature. What are the results?

The p-values for ozone, wind, and solar radiation were 0.2158, 0.1125, and 0.6026 respectively. All of these are greater than 0.05, so we would be unable to reject the null hypothesis for them. As a result, we conclude that there is not a statistically significant difference between the May and September data for these categories. However, temperature had a p-value of 3.309e-07. Since this is smaller than 0.05, we reject the null hypothesis and accept the alternative one. As a result, we conclude that there is a statistically significant difference between the May and September temperature data.

1. Now, perform the non-parametric version of the independent samples t test, comparing May with September for the four variables. Do these results differ from those of the parametric t tests you did for Question 7? Based on the histograms you generated for Question 2, would you expect a non-parametric test to be a better choice for any of these variables?

The p-values differ slightly from those in Question 7. However, the conclusions remain unchanged. Ozone, wind, and solar radiation are still greater than 0.05 while only temperature is smaller. Nonparametric does seem like a good idea because the distributions of each data category will follow different trends.