STA 6443: Algorithms I

Homework 1

Ashley Cortez

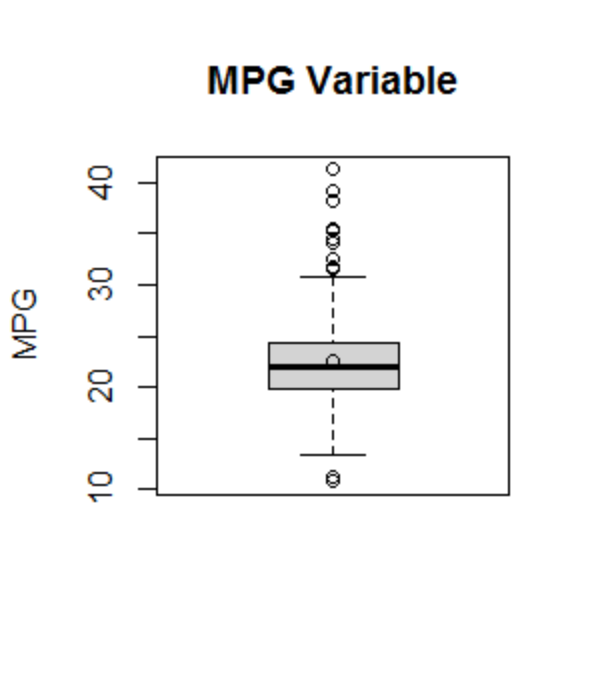
Jacob Kendall

September 9, 2021

**Exercise 1: Descriptive Statistics**

**a)** Create a combined mpg variable called **MPG Combo** which combines 60% of the **MPG\_City** and 40% of the **MPG Highway**. Obtain a box plot for **MPG\_Combo** and comment on what the plot tells us about fuel efficiencies.

**Miles per Gallon Combo Variability**

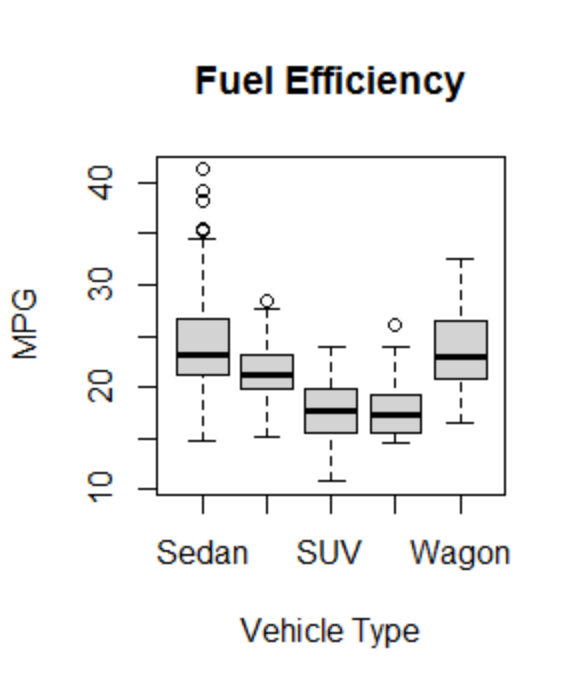


**Table 1. Boxplot of MPG Combo**

This chart indicates that the median miles per gallon among vehicles is around 22 mpg and the average miles per gallon is similar at 22.54 (Table 1). The boxplot illustrates a right skewed data set. The plot shows that there are several outliers identified in fuel efficiency among vehicles. This may indicated that there are several outliers with miles per gallon significantly above and below the median and the data is not normally distributed.

**b)** Obtain box plots for **MPG\_Combo** by **Type** and comment on any differences you notice between the different vehicle types combined fuel efficiency.

**Fuel Efficiency by Type**



**Table 1. Boxplot of Fuel Efficiency per Vehicle Type**

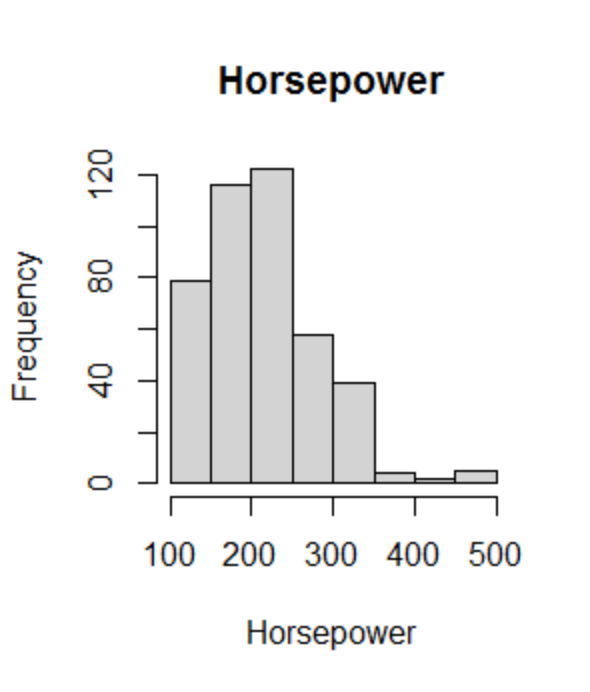
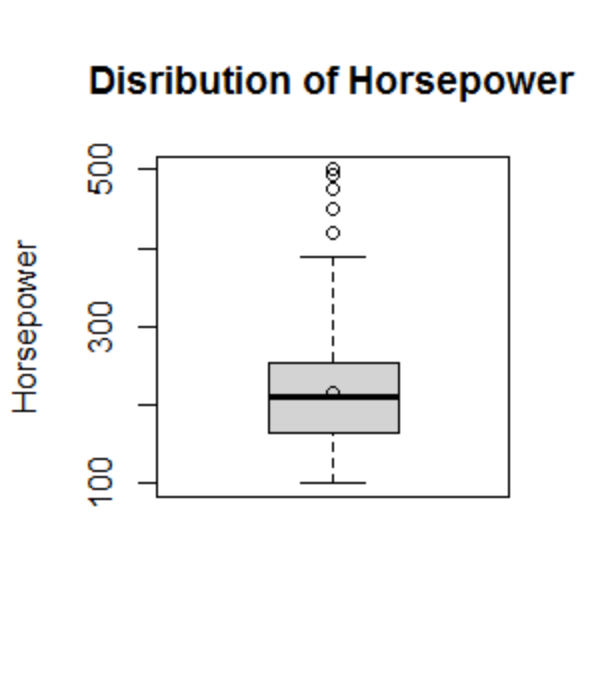
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SEDAN** | **SPORTS** | **SUV** | **TRUCK** | **WAGON** |
| **MIN** | 14.80 | 15.20 | 10.80 | 14.60 | 16.60 |
| **1st QTR** | 21.20 | 19.80 | 15.60 | 15.60 | 20.95 |
| **MEDIAN** | 23.20 | 21.20 | 17.70 | 17.30 | 22.90 |
| **MEAN** | 24.10 | 21.24 | 17.86 | 18.30 | 23.82 |
| **3rd QTR** | 26.55 | 23.30 | 19.70 | 18.80 | 26.30 |
| **MAX** | 41.20 | 28.40 | 24.00 | 26.00 | 32.60 |

**Table 2. Fuel Efficiency by Vehicle Type Statistics**

A review of the data above suggest that fuel efficiency may differ by vehicle type. The data suggest that vehicles such as “Sedans” and “Wagons” enjoy better fuel efficiency with an average of approximately 24 miles per gallon (Table 2). While “Sports”, “SUVs” and “Trucks” average approximately 21 and 18 miles per gallon respectively (Table 2). The data also states that the optimum fuel efficient vehicle type is a Sedan (40 mpg) and the least fuel efficient is an SUV (11 mpg).

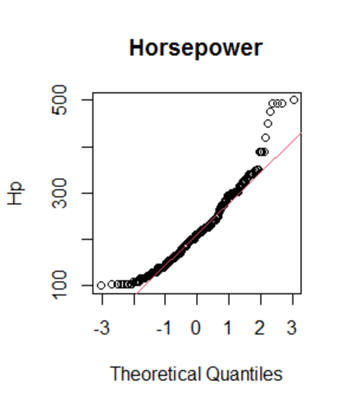
**c)** Obtain basic descriptive statistics for **Horsepower** for all vehicles. Comment on any general features and statistics of the data. Use visual and quantitative methods to comment on whether an assumption of Normality would be reasonable for **Horsepower** variable.

**Horsepower of All Vehicles**

**Table 2. Boxplot of Horsepower**

**Table 1. Histogram of Horsepower**

****

**Table 3. Normality Plot**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MIN** | **1st QTR.** | **MEDIAN** | **MEAN** | **3rd QTR** | **MAX** |
| **100** | **165** | **210** | **216.8** | **255** | **500** |

**Table 4. Horsepower Quartile Statistics**

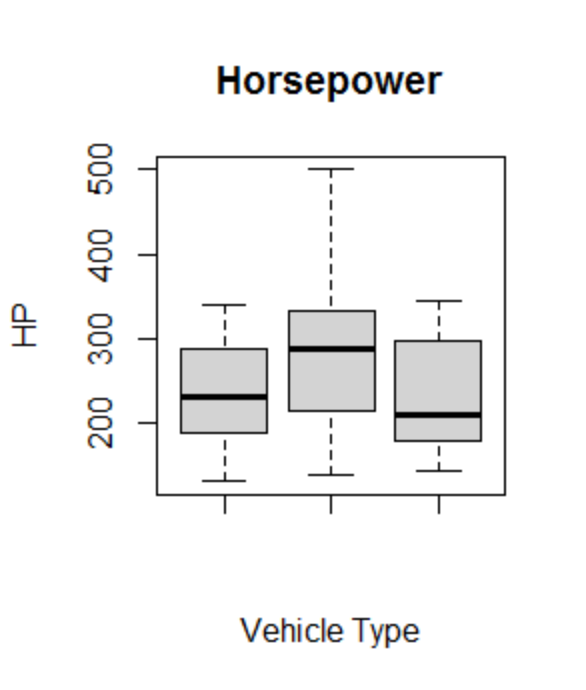
|  |  |  |  |
| --- | --- | --- | --- |
| **RANGE** | **VARIANCE** | **SKEWNESS** | **SHAPIRO WILK** |
| 100-500 | 5085.95 | .95 | 2.32e-11 |

**Table 5. Horsepower Descriptive Statistics**

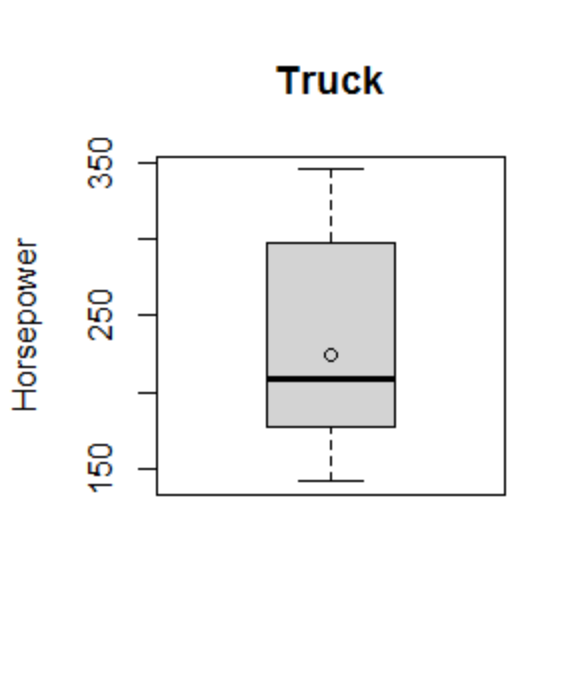
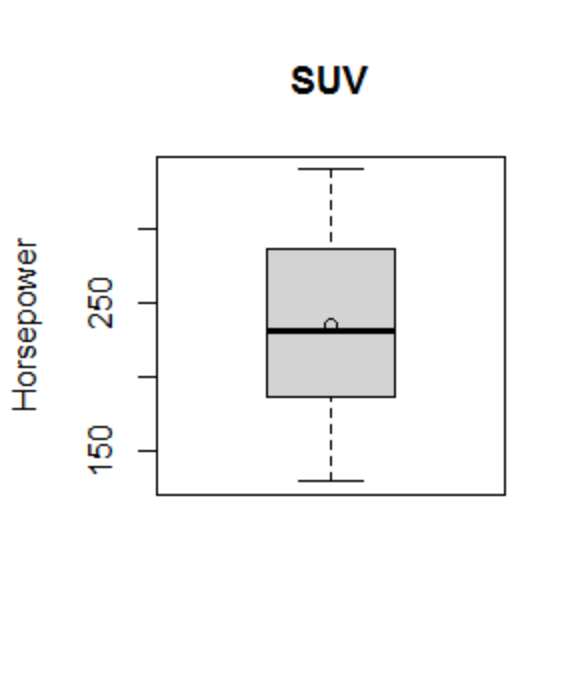
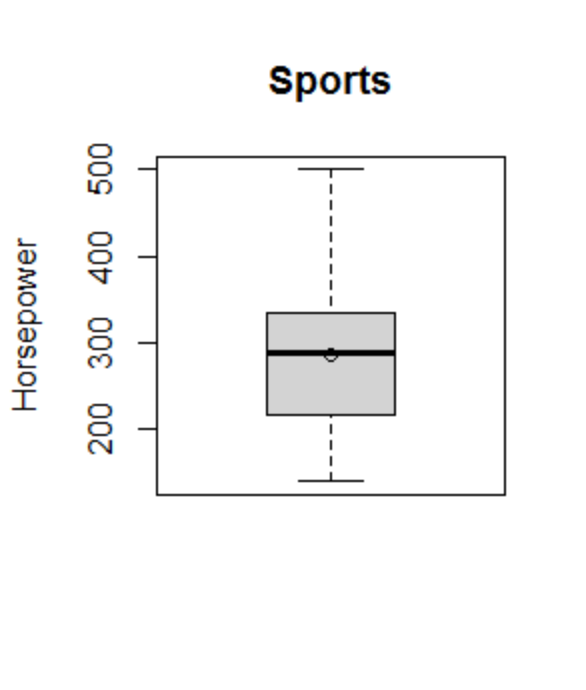
Table 1, illustrates the Horsepower for all vehicles has a slight right skew with a concentration of the data being between 100 and 250 horsepower. Further, Table 2 indicates that there are several outliers in this data set with a max value of 500. A quick visual representation from the normality plot suggest that there is not a normal distribution, specifically among cars with a horsepower above 400 (Table 3). The skewness is captured in Table 4, which identifies that the mean is greater than the median. These visuals suggest that the data is not normally distributed. To be certain, we reviewed the Shapiro Wilk test and found the p value to be less than 5% (Table 5), as a result we have an enough evidence to support that horsepower does not follow a normal distribution.

**d)** Obtain basic descriptive statistics for **Horsepower** for all vehicles. Comment on any general features and statistics of the data. Use visual and quantitative methods to comment on whether an assumption of Normality would be reasonable for **Horsepower** variable.

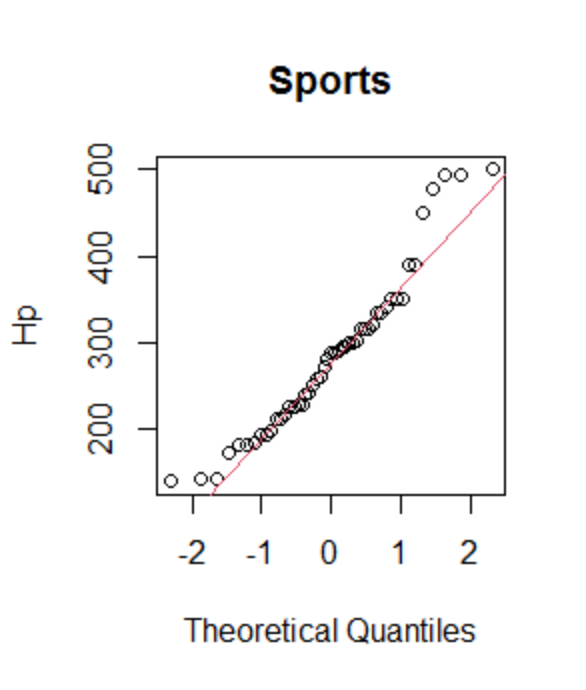
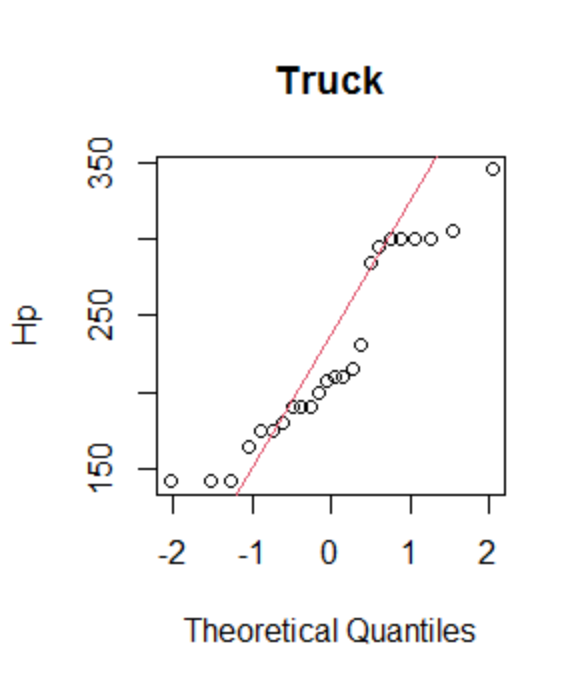
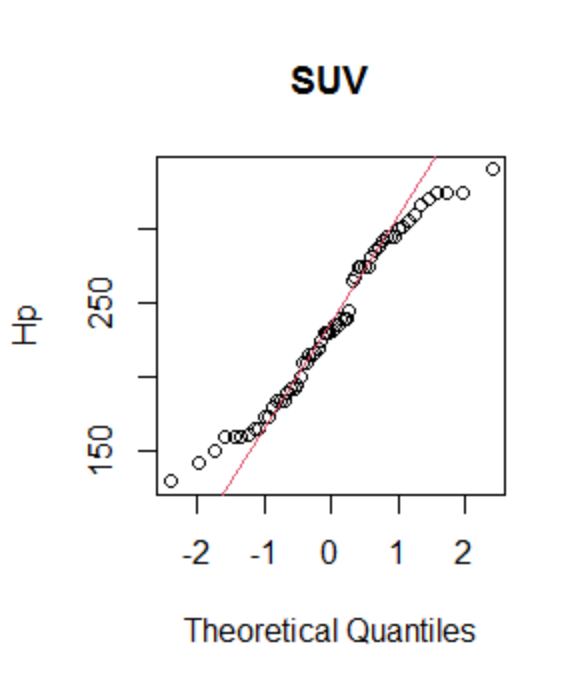
**Horsepower of Vehicle by Type (Sports, SUV, Truck)**



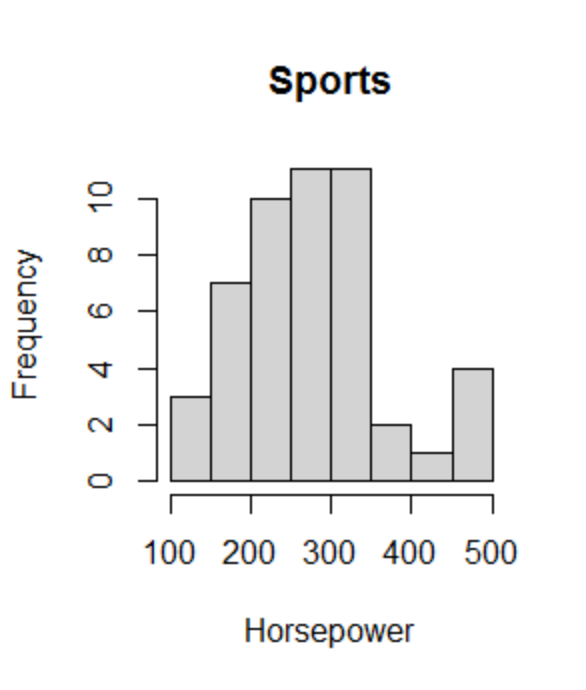
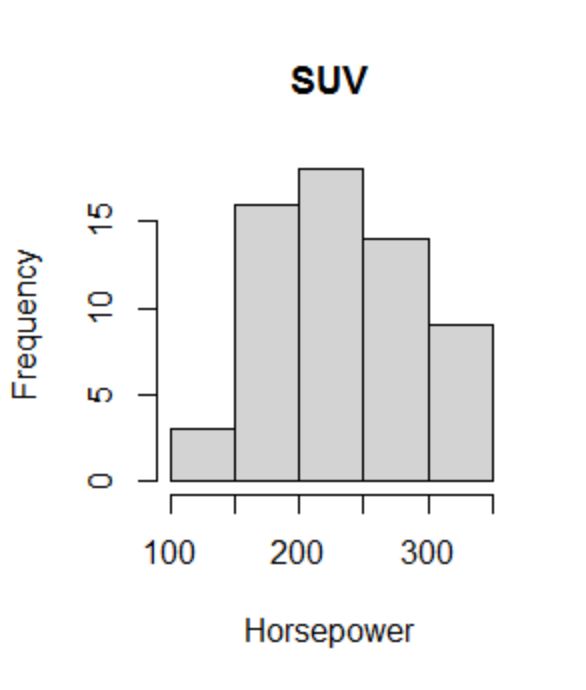
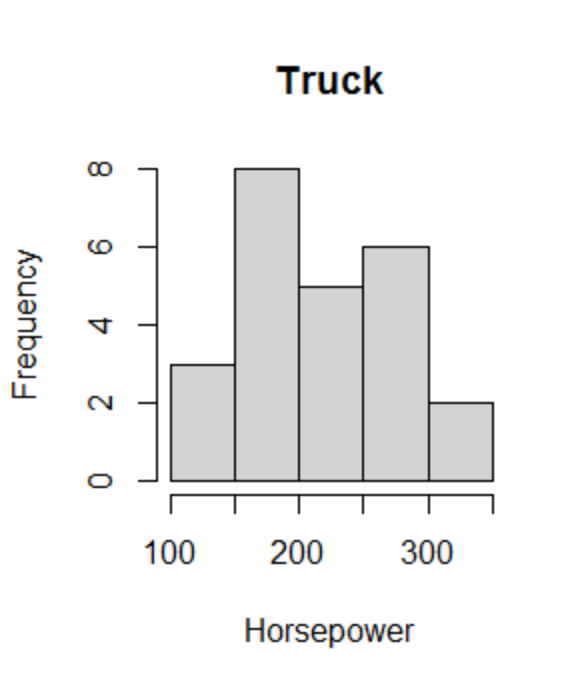
**Table 1. Horsepower by Vehicle Type**



**Table 2. Sports Horsepower Table 3. SUV Horsepower Table 4. Truck Horsepower**

**Table 5. Sports Normality Plot Table 6. SUV Normality Plot Table 7. Truck Normality Plot**

**Table 8. Sports HP Histogram Table 9. SUV HP Histogram Table 10. Truck HP Histogram**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **SPORTS** | **SUV** | **TRUCK** |
| **RANGE** | 138-500 | 130-340 | 142-345 |
| **VAR** | 8609.93 | 3162.25 | 3824.84 |
| **SKEWNESS** | 0.66 | 0.06 | 0.38 |
| **SHAPIRO WILK** | 0.02 | 0.04 | 0.02 |

**Table 11. Horsepower by Type Descriptive statistics**

To determine the normality of horsepower by vehicle type we conducted qualitative analysis specific to Sports, SUVs and Trucks. We reviewed histograms, boxplots and normality plots to determine if horsepower by vehicle type was normally distributed. Visually, each vehicle type did not seem to follow a normal distribution. To confirm this, we conducted a Shapiro Wilk test for each vehicle type (Sports, SUV and Truck) and identified that each vehicle type had a p value of less than 5%. This supports our hypothesis that horsepower by vehicle type (for Sports, SUV and Truck) does not follow a normal distribution. Tables 1-11 support our claim.

**Exercise 2: Hypothesis Testing**

Perform a hypothesis test of whether SUV has different horsepower than Truck, and state your conclusions

**a)** Which test should we perform, and why? Justify your answer based on findings on Exercise 1 (d).

**Normality:** Referring to tables 6, 7, 9, 10 and 11 in exercise 1d we were able to identify the normality of the horsepower for both SUVs and Trucks. The histograms for both SUV and Trucks does not indicate a large variation in data, however a look at the qq plots indicate that the data may not be normally distributed. To determine this we ran a shapiro wilk test for both SUVs and Trucks (Table 11) and determined that both P Values were less than 5%. This indicates that we can reject the Null hypothesis and confirm that horsepower for both SUVs and Trucks is not normally distributed.

**b)** Specify null and alternative hypotheses.

**Hypothesis:**

Null (Ho): Median of Horsepower for SUVs is equal to the median of Horsepower for Trucks

Ho: MSUV = MTruck

Alternative (Ha): Median of Horsepower for SUVs is not equal to the median of Horsepower for Trucks

Ha: MSUV != MTruck

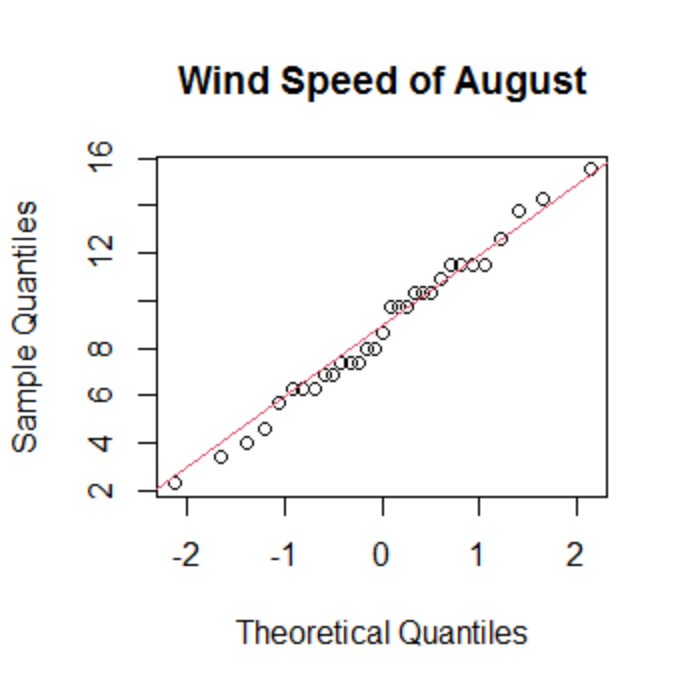
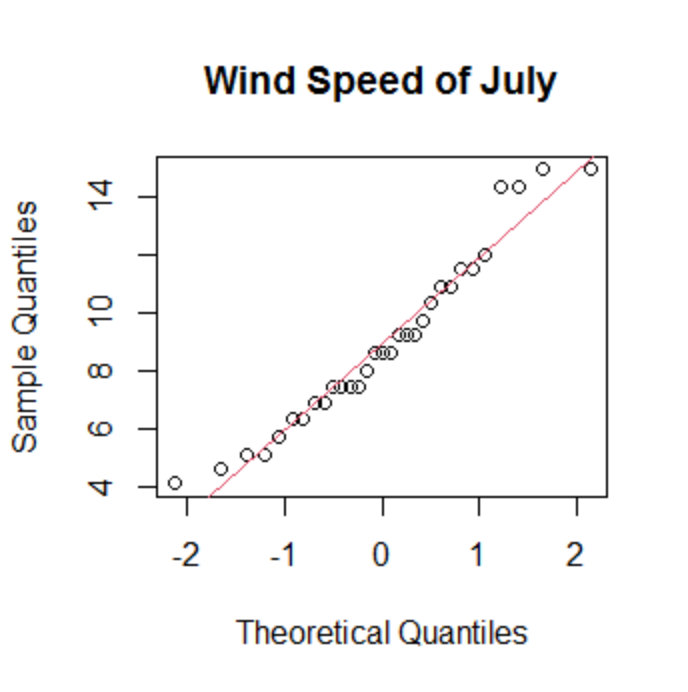
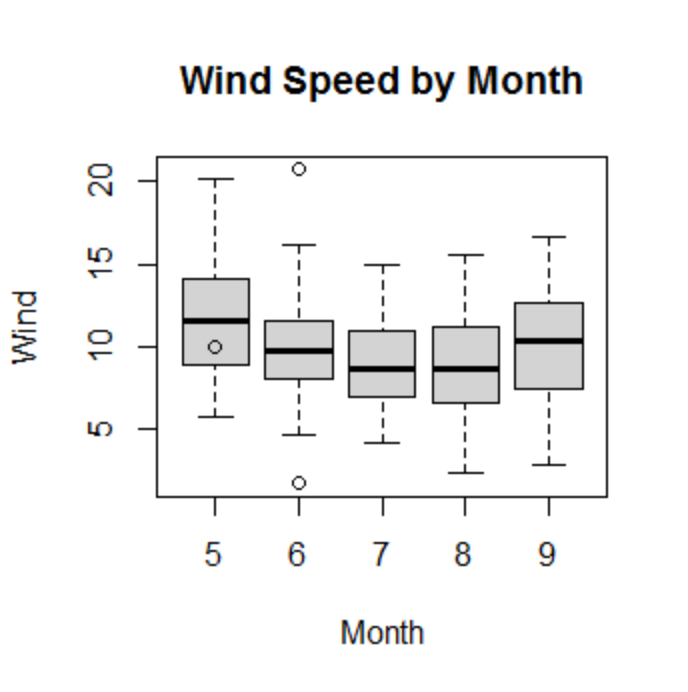
**c)** State the conclusion based on the test result.

**Conclusion:** Since our data was not normally distributed and we are testing two populations, we conducted a Wilcoxon Rank-Sum test and determined that median horsepower of SUVS is equal to the median Horsepower of Trucks. This is based on our P Value of .39, which is greater than 5% so we cannot reject the null hypothesis.

**Exercise 3: Hypothesis Testing**

Perform a hypothesis test -whether Wind in July has a different speed (mph) than Wind in August.

**a)** Which test should we perform, and why? See QQ-plot and perform Shapiro-Wilk test for normality check.



**Table 1. Boxplot of Wind Speed Table 2. QQ Plot of July** **Table 3. QQ Plot of August**

**Normality:** To determine normality of the data we ran a boxplot (Table 1) and qq plot of wind speed by month (Tables 2 and 3). Visually, it appears that the data is normally distributed, this is best captured in tables 2 and 3 that show that wind speeds seem to be consistent with the regression line. To confirm we ran a shapiro wilk test for the wind in speed in the month of July and August and determined P Values of .16 and .94 respectively. Since our P Values are greater than 5% we cannot reject the null and determine that the data is normally distributed.

**Variance:**  Since we are conducting a two population test and have determined that the data is normally distributed, we will test if the variance of wind speed in July is equal to the wind speed in August. To test this we ran a variance test received a P Value of .74. This does not give us enough evidence to reject the null and we can determine that the variance is equal for both data sets.

**b)** Specify null and alternative hypotheses

**Hypothesis:**

Null (Ho): Mean speed of wind in July is equal to the mean speed of wind in August

Ho: M7 = M8

Alternative (Ha): Mean speed of wind in July is not equal to the mean speed of wind in August

Ho: M7 != M8

**c)** State the conclusion based on the test result.

**Conclusion:** Since both are data was normally distributed and their variance was equal, we ran a pooled t test. We received a P Value of .85 which is greater than 5%. We do not have enough evidence to reject the null and may conclude that there is NOT a significant difference in wind speed between July and August.