HW1

Collaborators: Collin Real (yhi267), Seth Harris (dmp903)

Import libraries.

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(e1071)
library(fBasics)
```

```
Attaching package: 'fBasics'

The following objects are masked from 'package:e1071':

    kurtosis, skewness
```

```
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

Set working directory path.

```
setwd("/Users/c2cypher/codebase/msda/sta-6443-902-data_analytics_algorithms/HW1")
```

Read in HW1 dataset.

```
cars=read.csv("Cars.csv", header = TRUE)
```

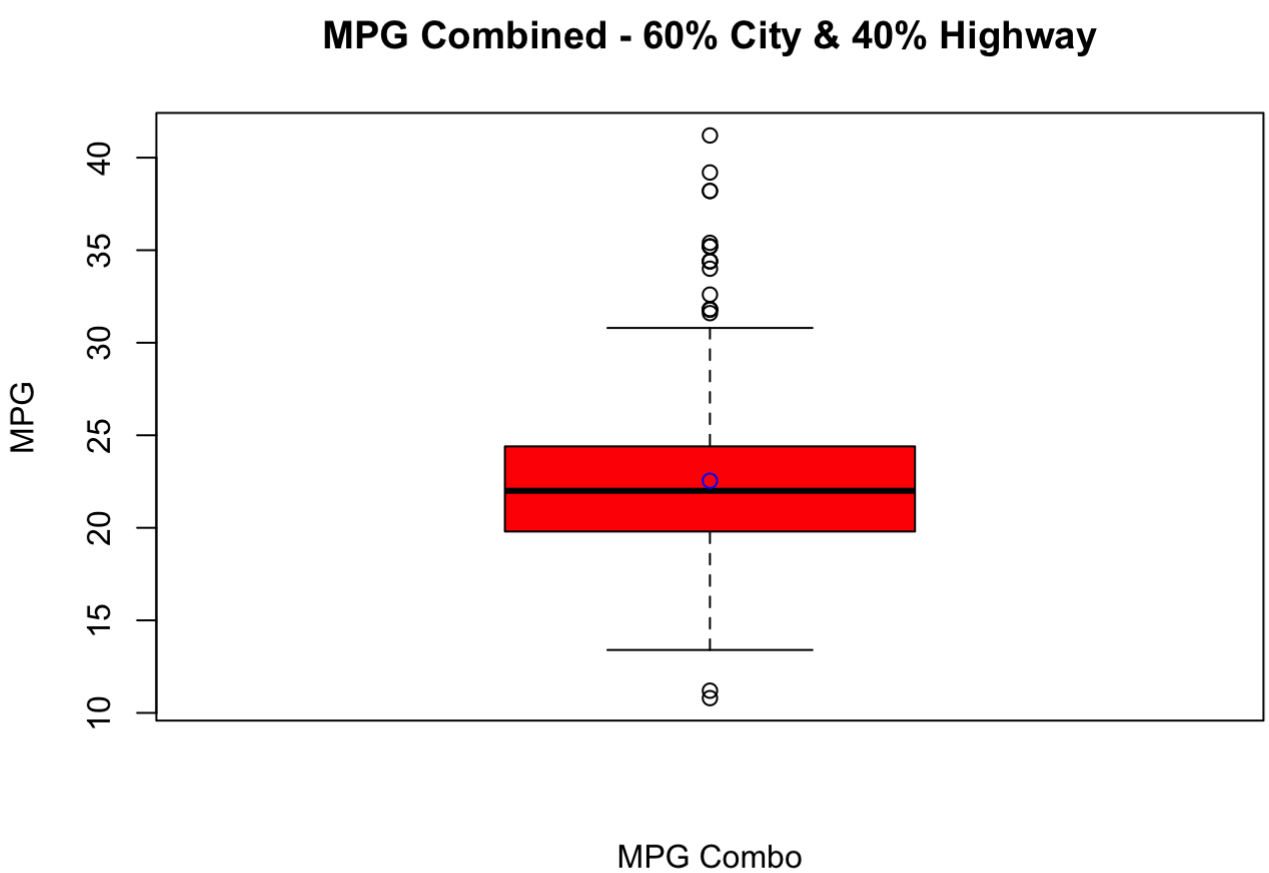
(1a) Create the combined MPG variable and create a new dataframe with the new MPG_Combo column.

```r
MPG_Combo <- 0.6*cars$MPG_City+0.4*cars$MPG_Highway
cars=data.frame(cars, MPG_Combo)
summary(cars$MPG_Combo)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.80   19.80   22.00   22.54   24.40   41.20 
```

Create box plot for the new MPG_Combo variable.

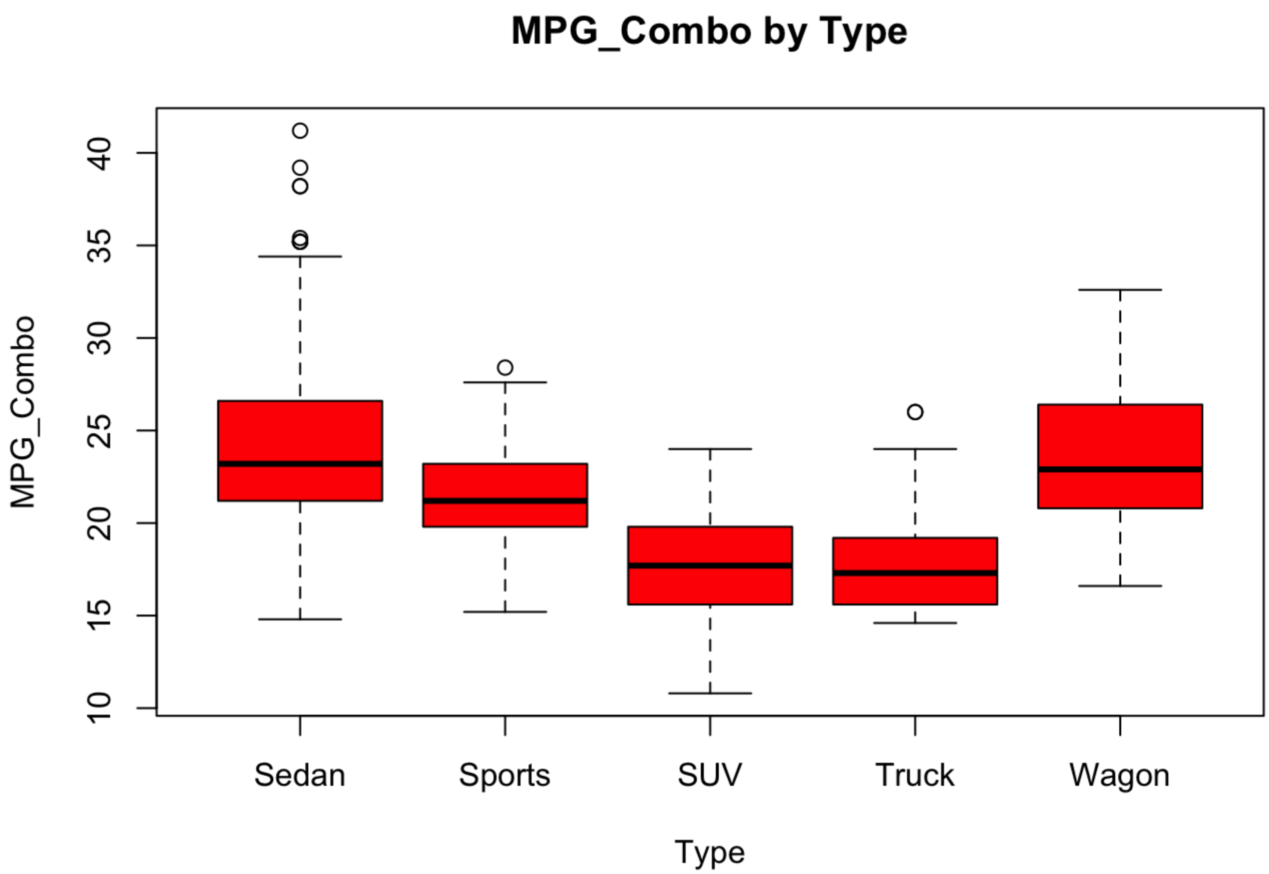
```r
boxplot(cars$MPG_Combo,
        main="MPG Combined - 60% City & 40% Highway",
        ylab="MPG",
        xlab="MPG Combo",
        col="red",
        border="black");
points(mean(cars$MPG_Combo, na.rm = TRUE), col="blue")
```



Comments: The box plot illustrates several outliers above the maximum and below the minumum. The mean and median values are almost equal indicating the possibility of normal distribution. The plot also illustrates that 50% of the cars have a 20-25 MPG range. Finally, the minimum value for MPG is ~14, and the maximum value is ~31.

(1b) Create box plots for MPG_Combo by Type.

```r
boxplot(MPG_Combo~Type,
        data = cars,
        main="MPG_Combo by Type",
        xlab="Type",
        ylab="MPG_Combo",
        col="red",
        border="black")
```



Comments: The box plot illustrates that Sedan has several outliers above the maximum value. Truck and SUV have the lowest fuel efficiencies, whereas Wagon and Sedan have the highest. Sports vehicles fall between Truck/SUV and Wagon/Sedan in terms of fuel efficiency.

(1c) Basic descriptive statistics for Horsepower for all vehicles.

```r
summary(cars$Horsepower)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  100.0   165.0   210.0   216.8   255.0   500.0 
```

```r
mean(cars$Horsepower)
```

```
[1] 216.76
```

```r
median(cars$Horsepower)
```

```
[1] 210
```

```r
var(cars$Horsepower)
```

```
[1] 5085.952
```

```r
range(cars$Horsepower)
```

```
[1] 100 500
```

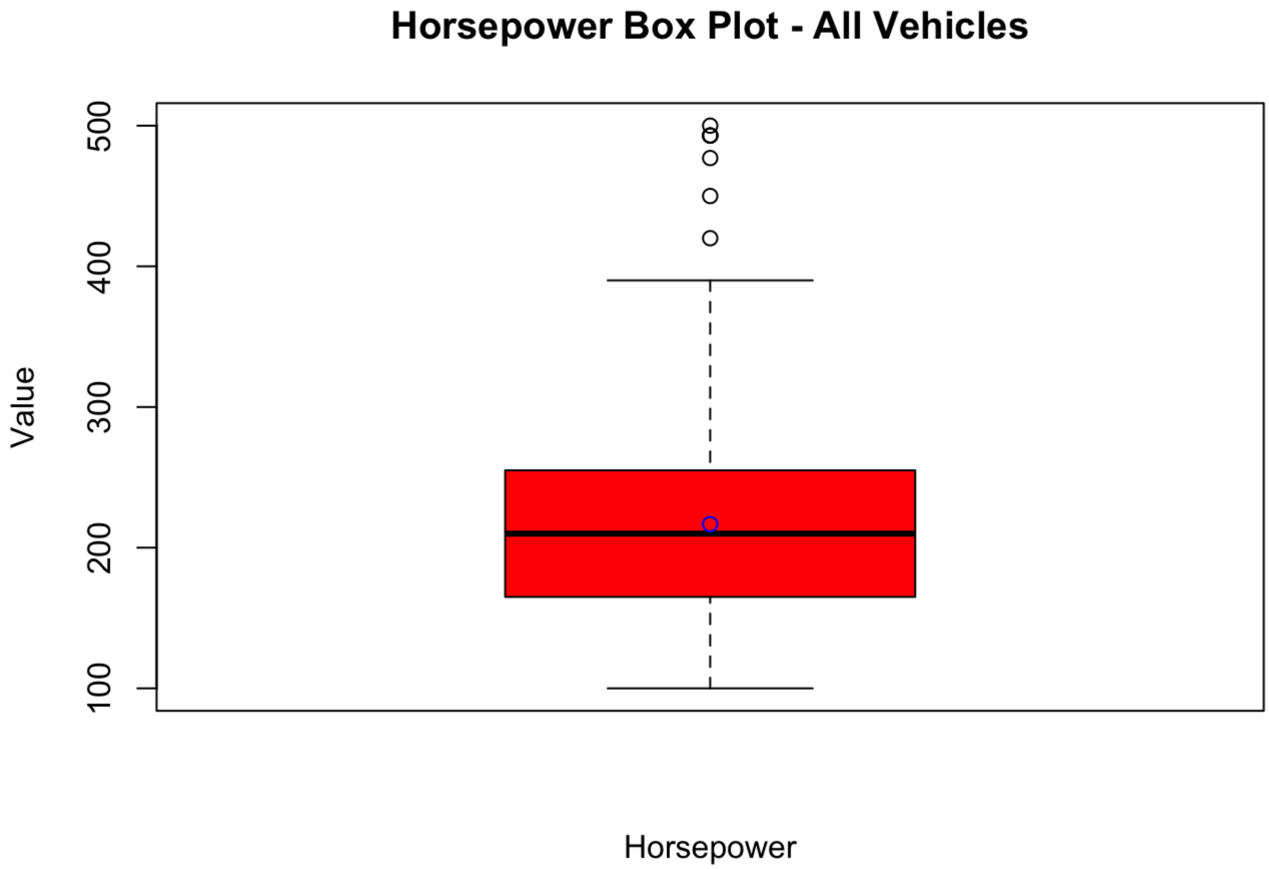
```r
skewness(cars$Horsepower)
```

```
[1] 0.9528091
attr(,"method")
[1] "moment"
```

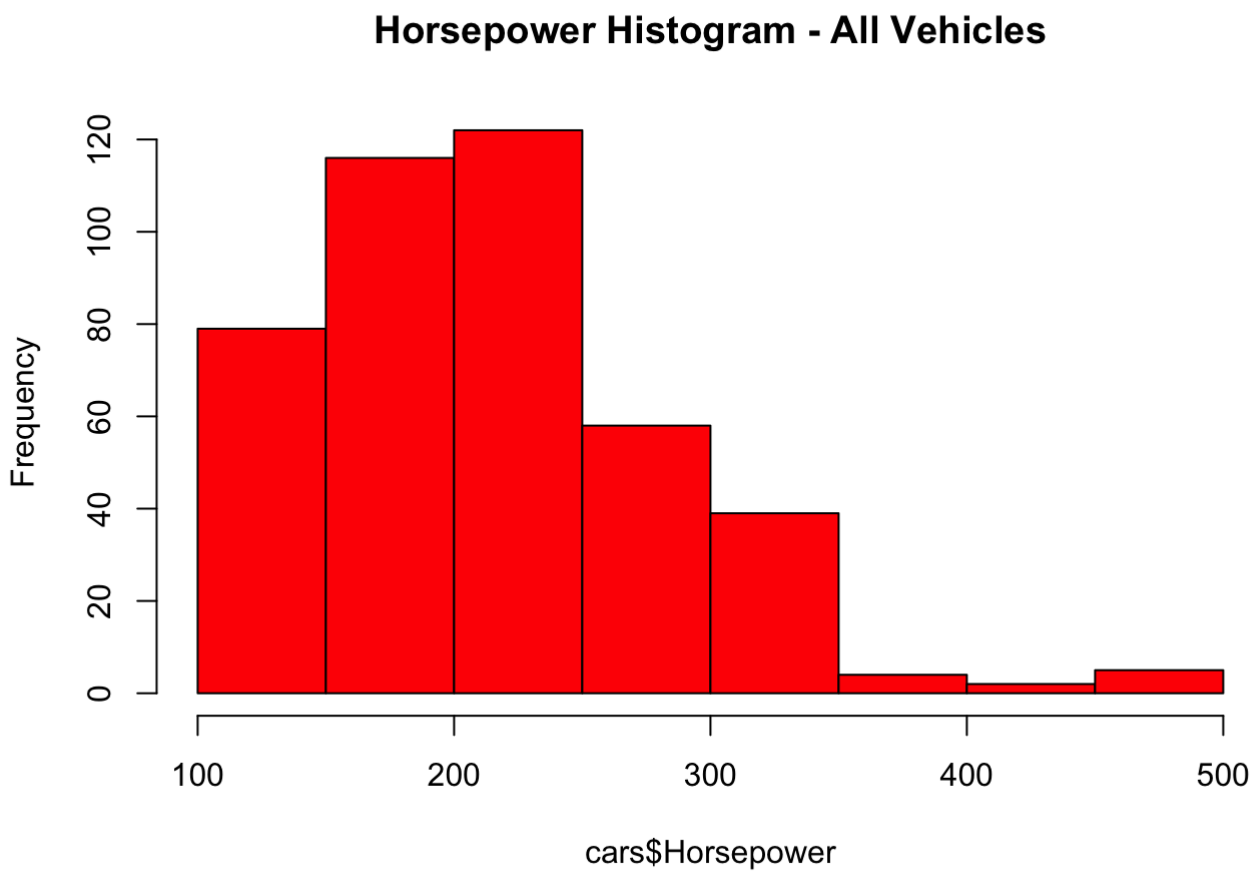
Comments: The mean (216.8) is greater than the median (210) indicating that the data is skewed to the right. Our skewness assumption is confirmed since our calculated skewness (0.9528091) is greater than zero. The values range from 100 to 500.

Visual Methods for Normality Assumption

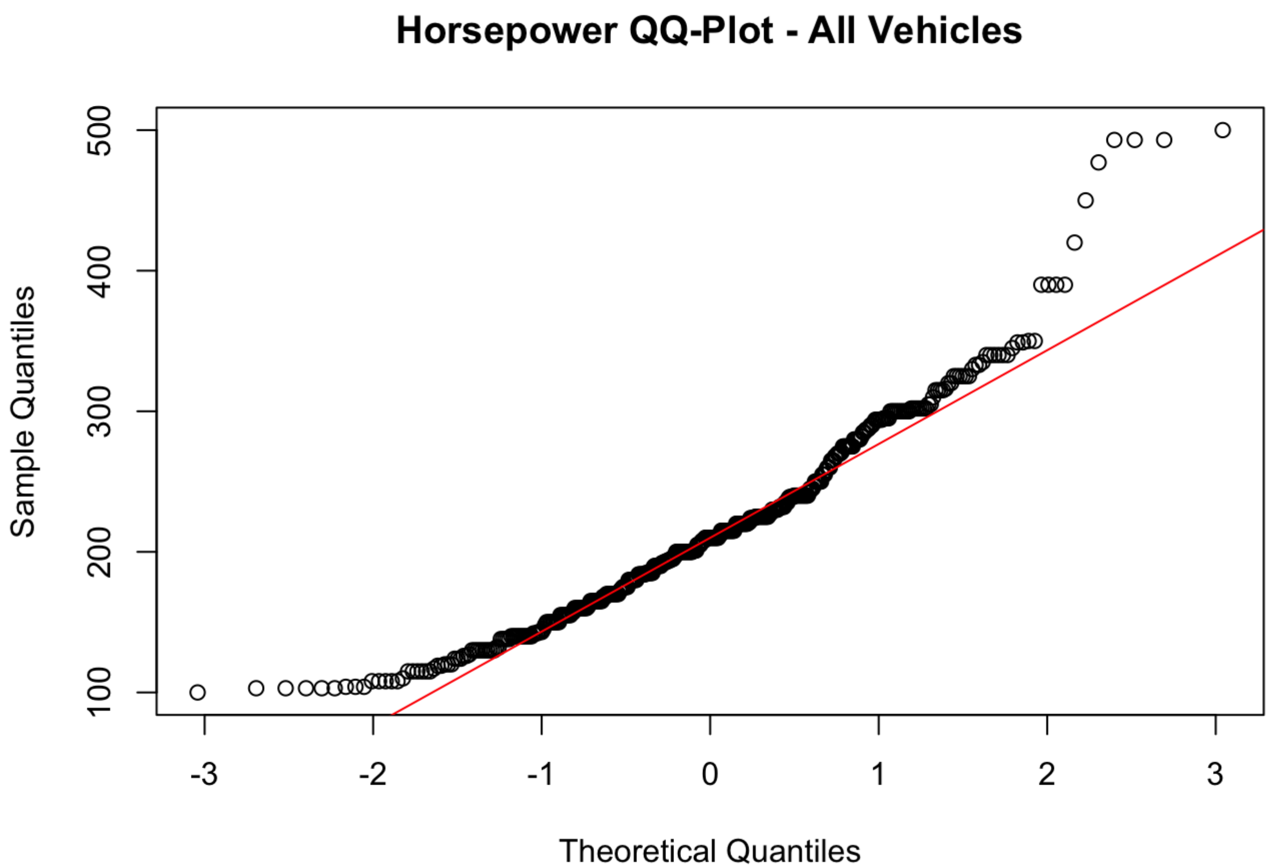
```r
boxplot(cars$Horsepower,
        main = "Horsepower Box Plot - All Vehicles",
        xlab = "Horsepower",
        ylab = "Value",
        col = "red",
        border = "black");
points(mean(cars$Horsepower, na.rm = TRUE), col="blue")
```




```
hist(cars$Horsepower,
     main = "Horsepower Histogram - All Vehicles",
     col = "red",
     border = "black")
```




```r
qqnorm(cars$Horsepower,
       main = "Horsepower QQ-Plot - All Vehicles");
qqline(cars$Horsepower, col = "red")
```



Comments: The histogram and box plot further illustrate the right skewness of the Horsepower data. The QQ-plot demonstrates a majority of the values near the red line, but there are various outliers.

Quantitative: Shapiro-Wilks Test is the quantitative method used to comment on the assumption of Normality for the Horsepower variable.

```
shapiro.test(cars$Horsepower)
```

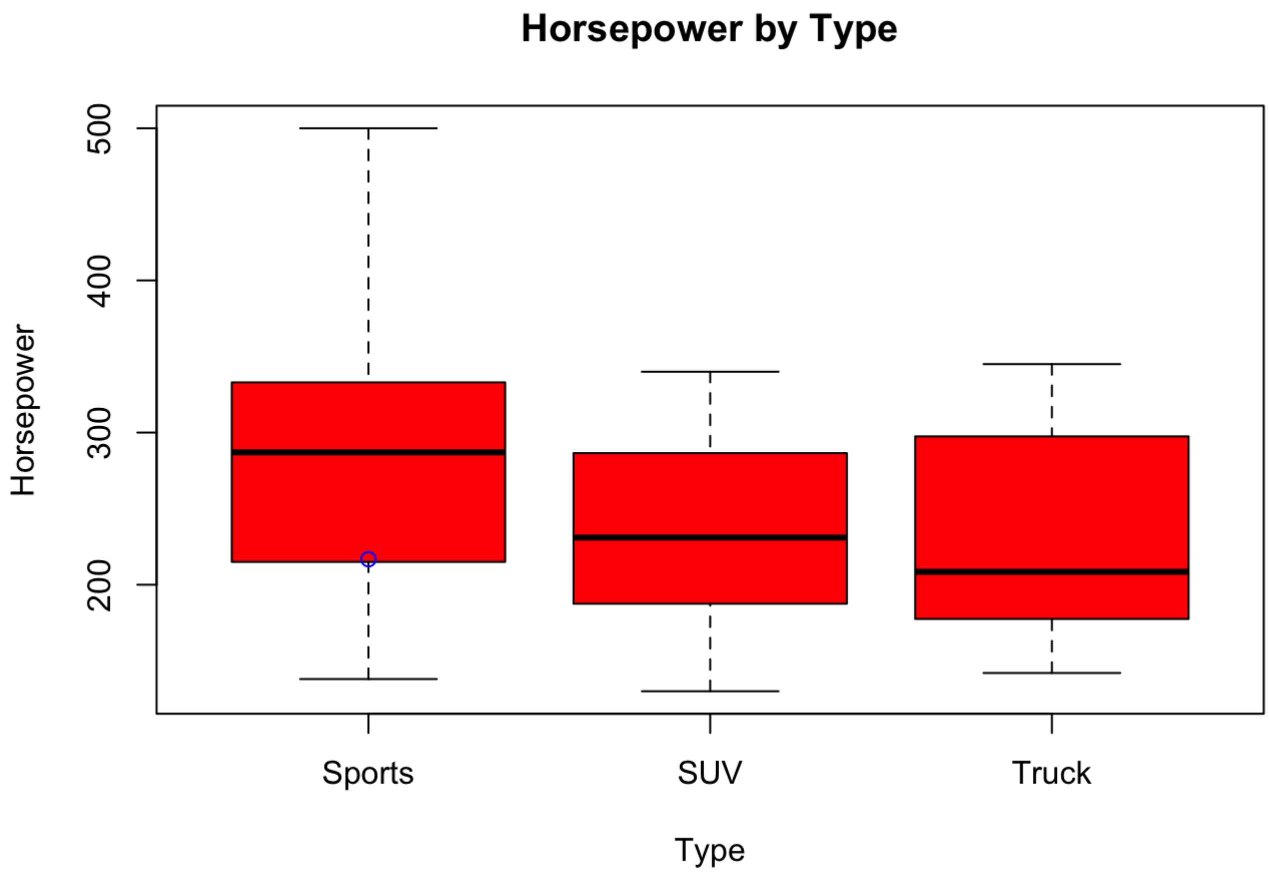
```
    Shapiro-Wilk normality test

data:  cars$Horsepower
W = 0.94573, p-value = 2.32e-11
```

Conclusion: Since the p-value (2.32e-11) of is less than the significance value (α=0.05), Horsepower **does not** follow normal distribution.

(1d) Assumption of normality for Horsepower by Type - Sports, SUV, Truck only. H0: Data follows normal distribution.; H1: Data does not follow normal distribution.

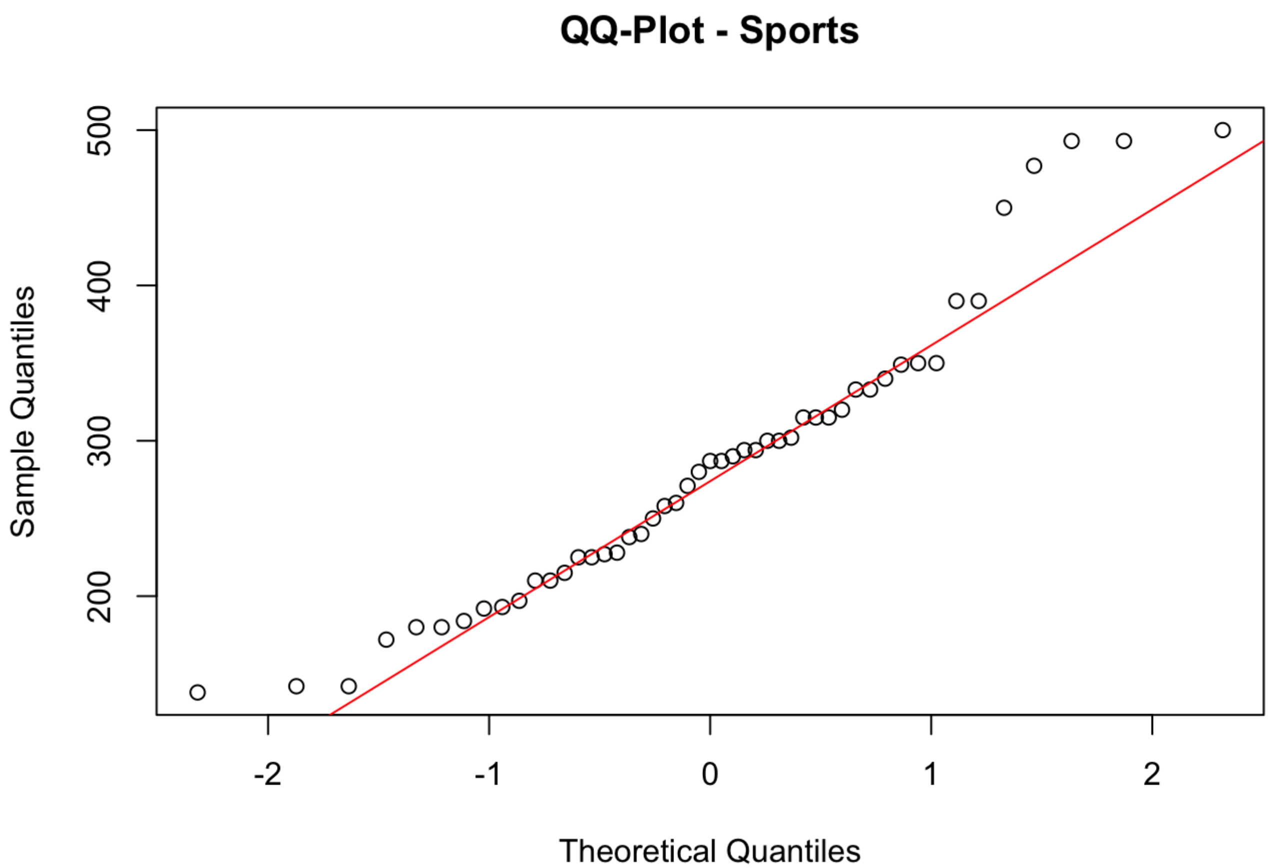
```
specified_types <- c("Sports", "SUV", "Truck")
cars2 = filter(cars, Type %in% specified_types)
boxplot(Horsepower ~ Type,
        data = cars2,
```

```
        main="Horsepower by Type",
        xlab="Type",
        ylab="Horsepower",
        col="red",
        border="black");
points(mean(cars$Horsepower), col="blue")
```

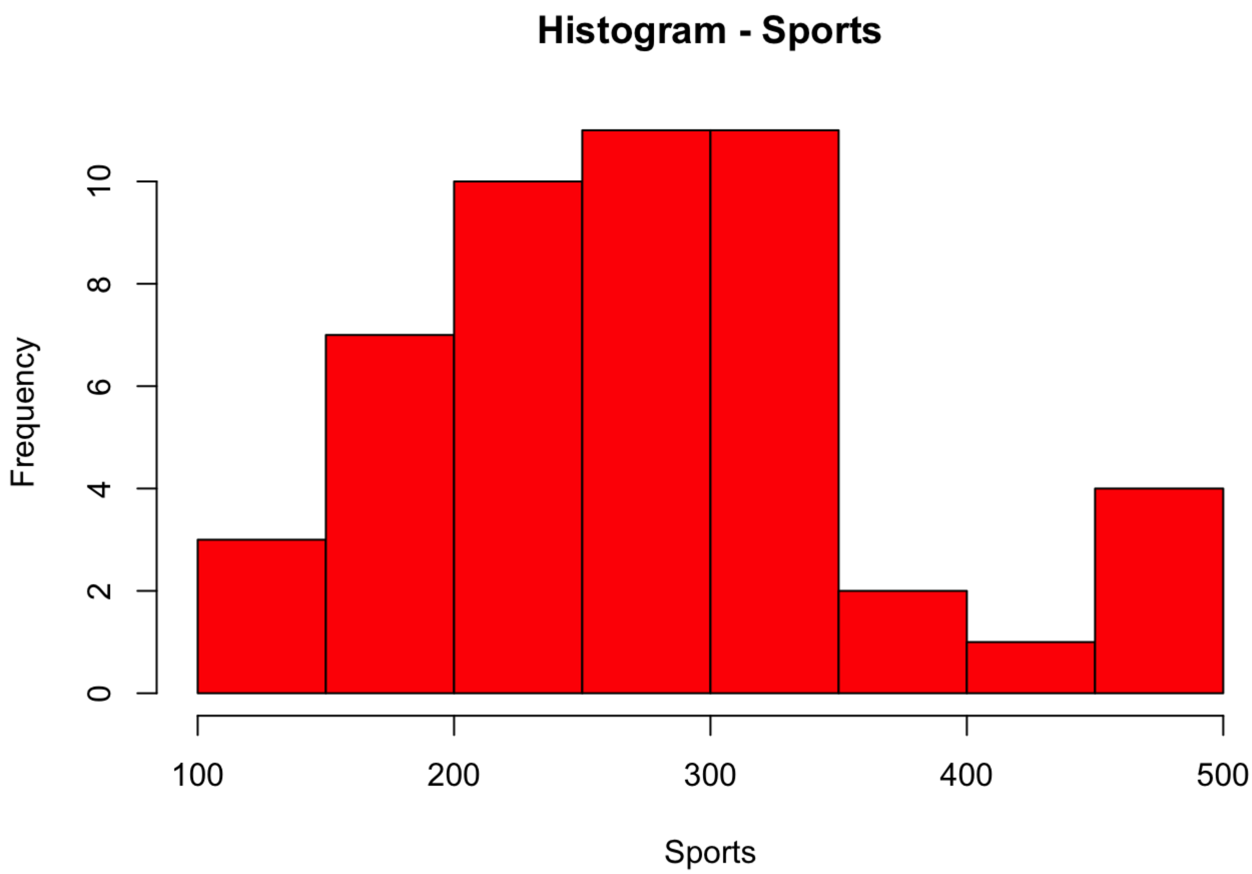


Sports - visual and quantitative methods for normality check.

```
qqnorm(cars2$Horsepower[cars2$Type=="Sports"],
       main = "QQ-Plot - Sports");
qqline(cars2$Horsepower[cars2$Type=="Sports"], col = "red")
```




```
hist(cars2$Horsepower[cars2$Type=="Sports"],
     main = "Histogram - Sports",
     xlab="Sports",
     col = "red",
     border = "black")
```



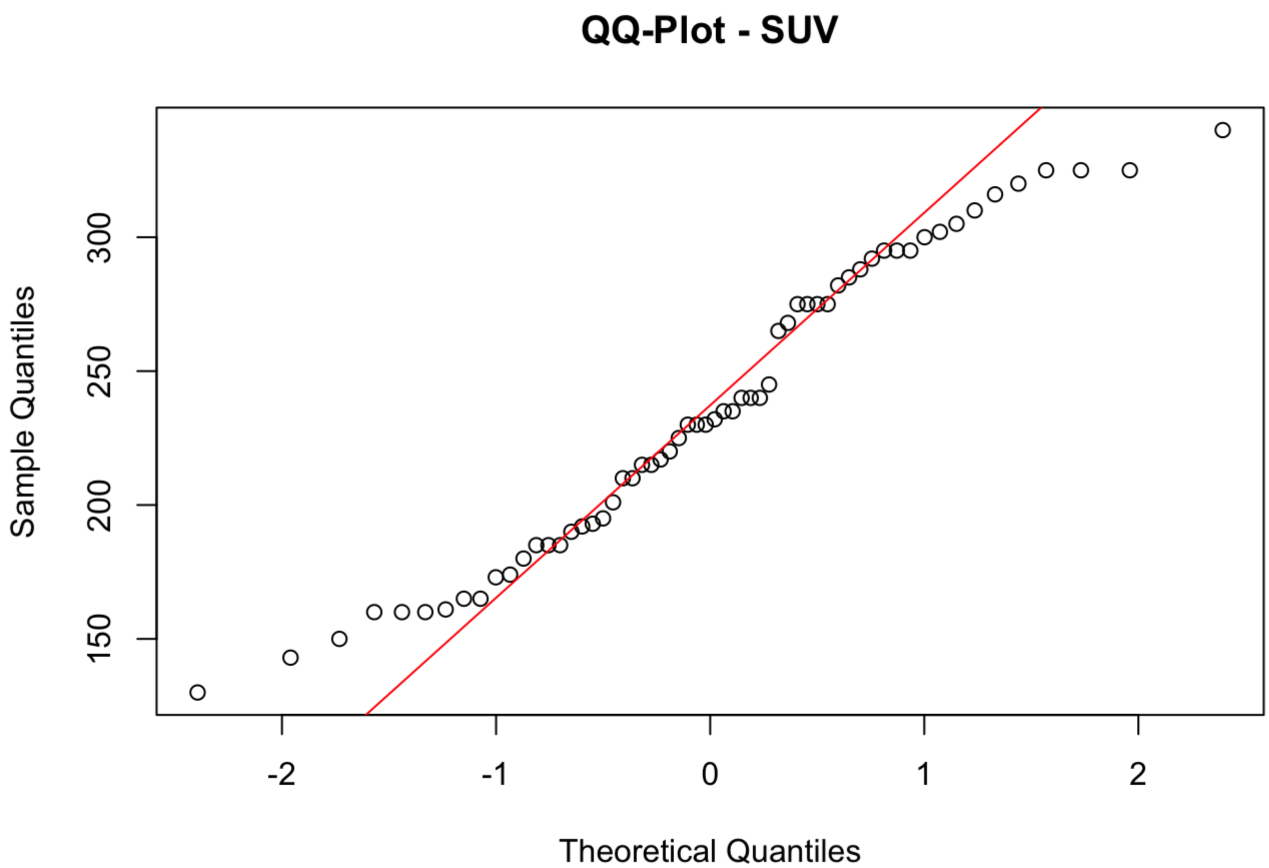

```
shapiro.test(cars2$Horsepower[cars2$Type=="Sports"])
```

```
	Shapiro-Wilk normality test

data:  cars2$Horsepower[cars2$Type == "Sports"]
W = 0.94276, p-value = 0.01898
```

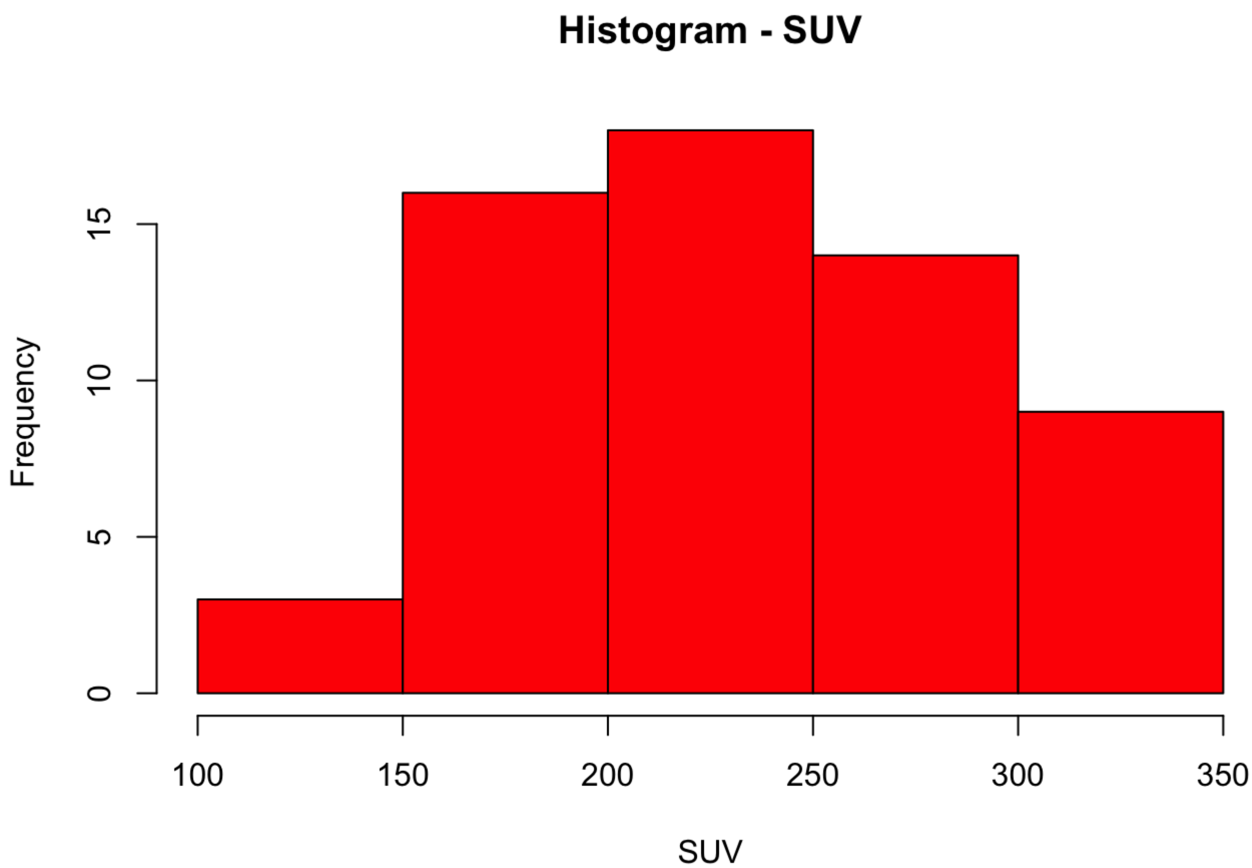
Comments: The histogram and box plots illustrate slight right skewness for Sports. The QQ-plot illustrates the majority of values close to the red line, but there are many outliers; therefore, the Shapiro-Wilk test was used to determine the normality of the data. Since the p-value (0.01898) is less than the significance level (α=0.05), we reject the null hypothesis. The data does not follow normal distribution.

SUV - visual and quantitative methods for normality check.

```
qqnorm(cars2$Horsepower[cars2$Type=="SUV"],
       main = "QQ-Plot - SUV");
qqline(cars2$Horsepower[cars2$Type=="SUV"], col = "red")
```




```r
hist(cars2$Horsepower[cars2$Type=="SUV"],
     main = "Histogram - SUV",
     xlab="SUV",
     col = "red",
     border = "black")
```



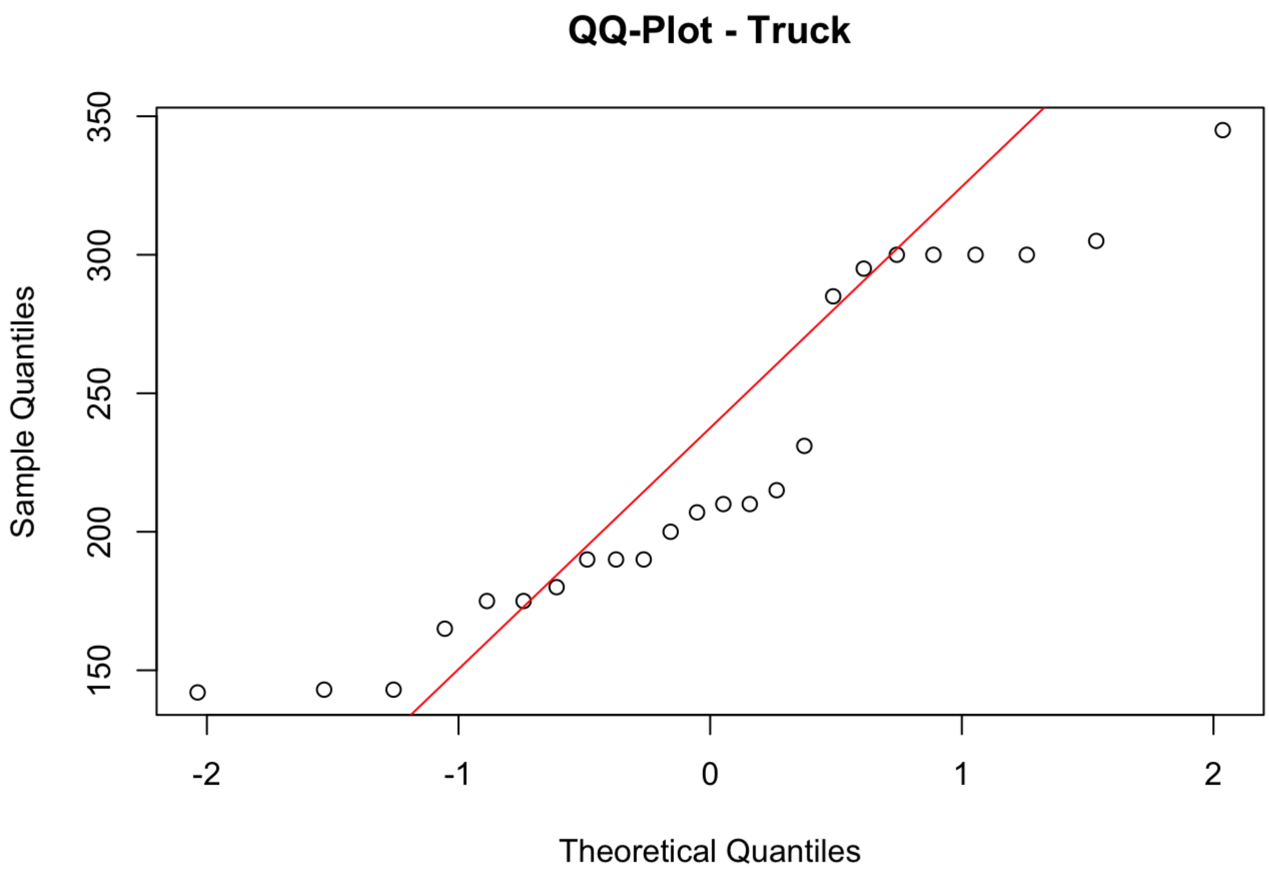

```
shapiro.test(cars2$Horsepower[cars2$Type=="SUV"])
```

```
	Shapiro-Wilk normality test

data:  cars2$Horsepower[cars2$Type == "SUV"]
W = 0.95945, p-value = 0.04423
```

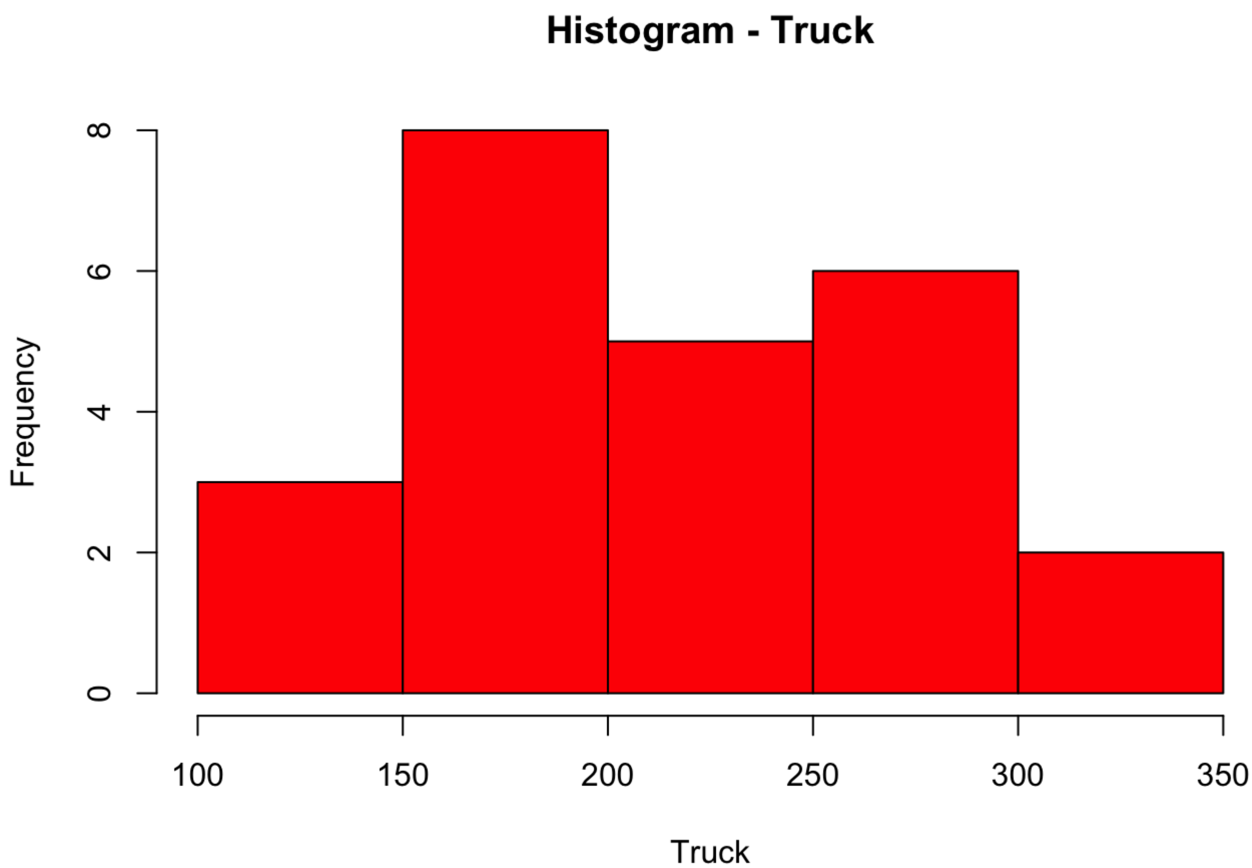
Comments: The histogram and box plots illustrate slight left skewness for SUV. The QQ-plot illustrates the majority of values close to the red line, but there are many outliers; therefore, the Shapiro-Wilk test was used to determine the normality of the data. Since the p-value (0.04423) is less than the significance level (α=0.05), we reject the null hypothesis. The data does not follow normal distribution.

Truck - visual and quantitative methods for normality check.

```
qqnorm(cars2$Horsepower[cars2$Type=="Truck"],
       main = "QQ-Plot - Truck");
qqline(cars2$Horsepower[cars2$Type=="Truck"], col = "red")
```




```
hist(cars2$Horsepower[cars2$Type=="Truck"],
     main = "Histogram - Truck",
     xlab="Truck",
     col = "red",
     border = "black")
```




```
shapiro.test(cars2$Horsepower[cars2$Type=="Truck"])
```

```
    Shapiro-Wilk normality test

data:  cars2$Horsepower[cars2$Type == "Truck"]
W = 0.8951, p-value = 0.01697
```

Comments: The histogram and box plots illustrate that the data is not normal for Truck. The QQ-plot illustrates that the majority of values are not close to the red line. The Shapiro-Wilk test was used to further demonstrate the non-normality of the data. Since the p-value (0.01697) is less than the significance level (α=0.05), we reject the null hypothesis. The data does not follow normal distribution.

Exercise 2: Hypothesis Testing

(2a) The p-values of SUV (0.04423) and Truck (0.01697) are less than the significance level (α=0.05), so we reject the null hypothesis; therefore, we should perform the **nonparametric Wilcoxon rank-sum test** because SUV and Truck are **not** normally distributed.

(2b) H0: SUV and Truck have similar horsepower because the two groups are from the same distribution with equal median values.; H1: SUV and Truck have different horsepower because one of the groups has a larger median value.

(2c)

```r
# Filter DataFrame to include only SUV & Truck
suv_truck_types <- c("SUV", "Truck")
cars3 = filter(cars2, Type %in% suv_truck_types)

# Perform the Wilcoxon rank-sum test
wilcox.test(Horsepower~Type, data = cars3, exact=FALSE)
```

```
    Wilcoxon rank sum test with continuity correction

data:  Horsepower by Type
W = 806.5, p-value = 0.3942
alternative hypothesis: true location shift is not equal to 0
```

Conclusion: Since p-value (0.3942) is larger than significance level (α=0.05), we do **not** have enough evidence to reject the null hypothesis (HO); therefore, SUV and Truck have similar horsepower because the two groups are from the same distribution with equal medians.

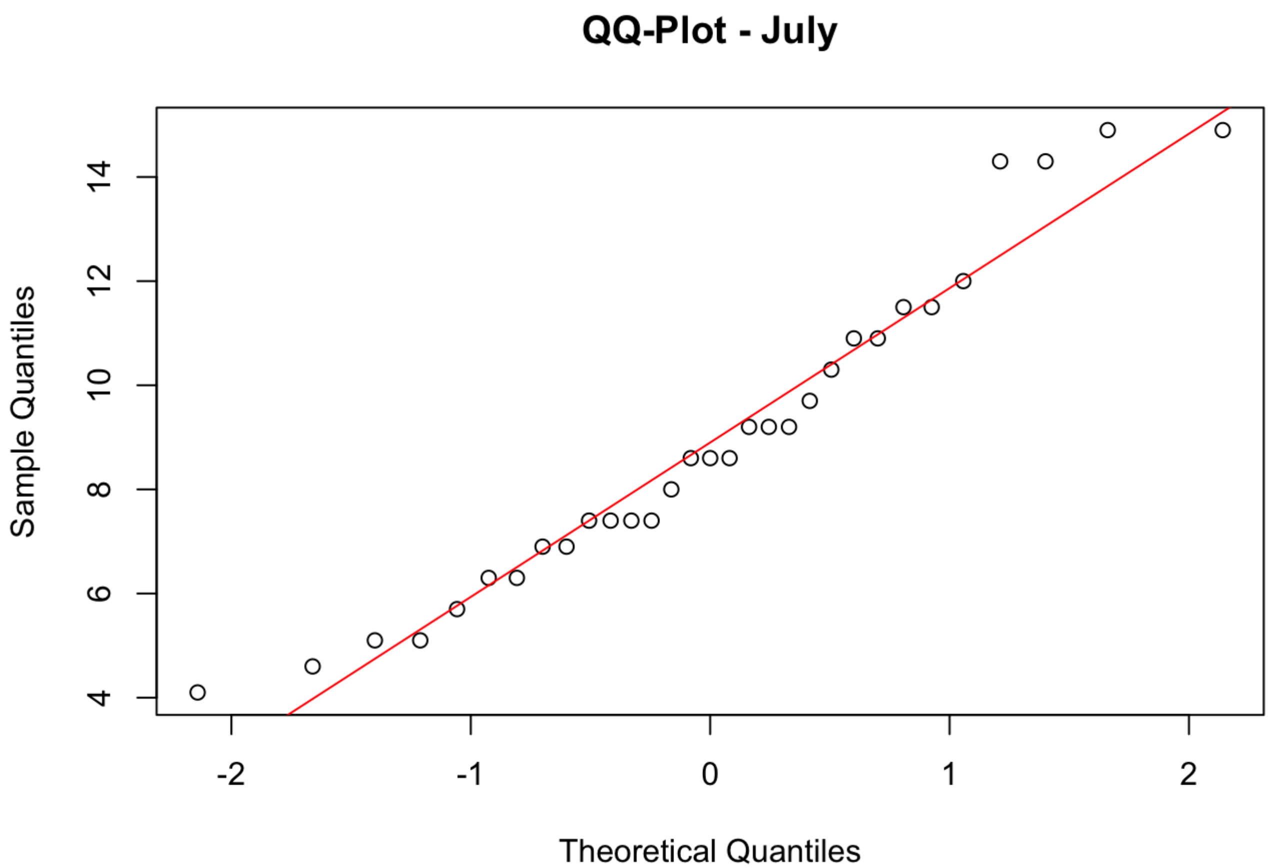
Exercise 3: Hypothesis Testing

(3a) Import dataset and filter DataFrame to include only July & August

```r
# Import dataset
df = airquality

# Filter DataFrame to include only July & August
months <- c(7, 8)
aug_july_df = filter(df, Month %in% months)
```

Create QQ-plots and perform the Shapiro-Wilk test for July and August normality check.

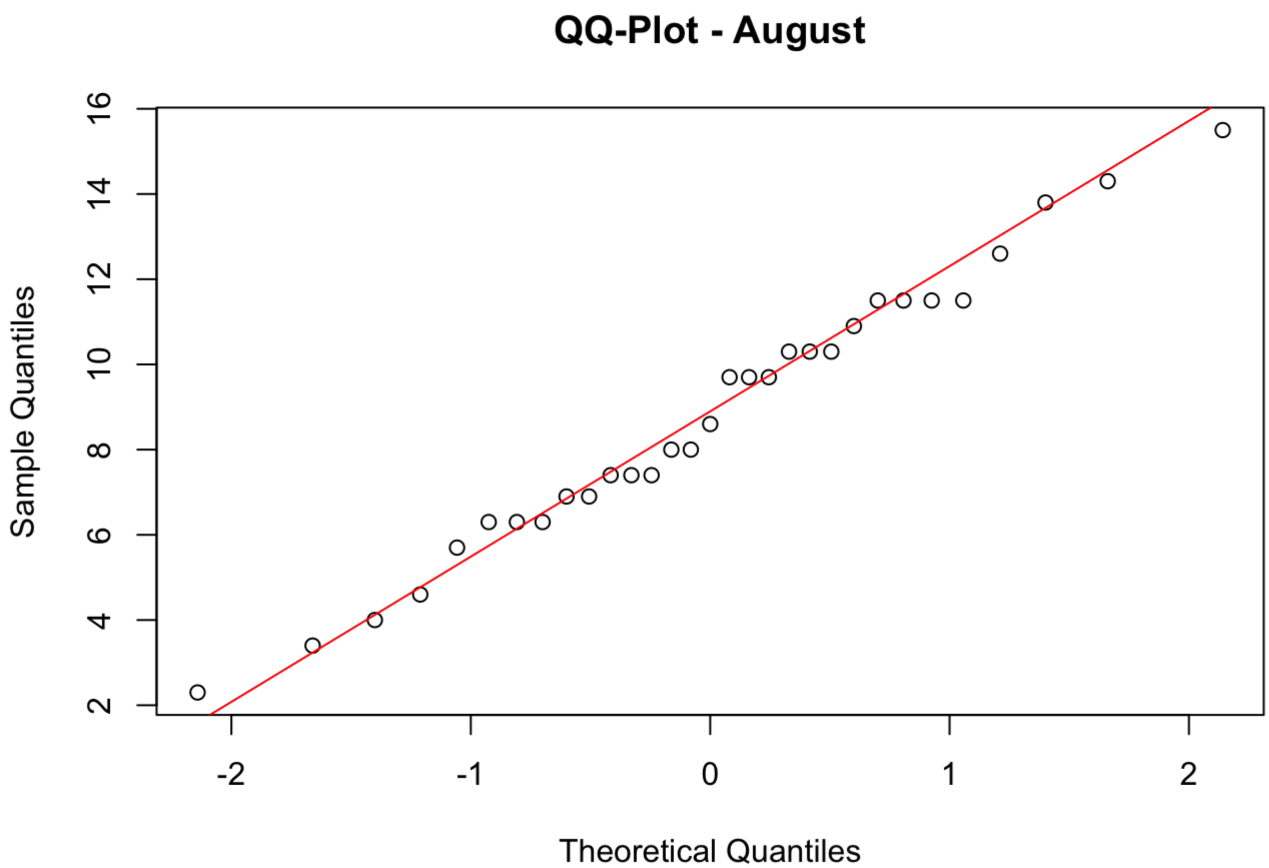
```r
# Create QQ-plot for July normality check.
qqnorm(aug_july_df$Wind[aug_july_df$Month==7],
       main = "QQ-Plot - July");
qqline(aug_july_df$Wind[aug_july_df$Month==7], col = "red")
```




```r
# Perform the Shapiro-Wilk test for July normality check.
shapiro.test(aug_july_df$Wind[aug_july_df$Month==7])
```

```
	Shapiro-Wilk normality test

data:  aug_july_df$Wind[aug_july_df$Month == 7]
W = 0.95003, p-value = 0.1564
```

```r
# Create QQ-plot for August normality check.
qqnorm(aug_july_df$Wind[aug_july_df$Month==8],
       main = "QQ-Plot - August");
qqline(aug_july_df$Wind[aug_july_df$Month==8], col = "red")
```




```
# Perform the Shapiro-Wilk test for August normality check.
shapiro.test(aug_july_df$Wind[aug_july_df$Month==8])
```

```
	Shapiro-Wilk normality test

data:  aug_july_df$Wind[aug_july_df$Month == 8]
W = 0.98533, p-value = 0.937
```

Since both p-values for July (0.1564) and August (0.937) follow the normal distribution (p-values > significance level), we perform the equal variance test to choose between the pooled t-test (equal variance case) and Satterthwaite t-test (unequal variance case).

```
var.test(Wind ~ Month, data=aug_july_df, alternative="two.sided")
```

```
	F test to compare two variances

data:  Wind by Month
F = 0.8857, num df = 30, denom df = 30, p-value = 0.7418
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4270624 1.8368992
```

```
sample estimates:
ratio of variances
         0.8857035
```

Since the p-value (0.7418) of the equal variance test is greater than the significance level (α=0.05), we do not have enough evidence to reject the null hypothesis; therefore, we will perform the **pooled t-test** because July and August have the same variance.

(3b) H0: mean of July = mean of August; H1: mean of July != mean of August

(3c)

```
t.test(Wind ~ Month, data = aug_july_df, alternative = "two.sided", var.equal=TRUE)
```

```

	Two Sample t-test

data:  Wind by Month
t = 0.1865, df = 60, p-value = 0.8527
alternative hypothesis: true difference in means between group 7 and group 8 is not equal
to 0
95 percent confidence interval:
 -1.443108  1.739883
sample estimates:
mean in group 7 mean in group 8 
       8.941935        8.793548 
```

Conclusion: Since the p-value (0.8527) of the pooled t-test is greater than the significance level (α=0.05), we do not have enough evidence to reject the null hypothesis; therefore, the data is normally distributed and the two groups have equal means.