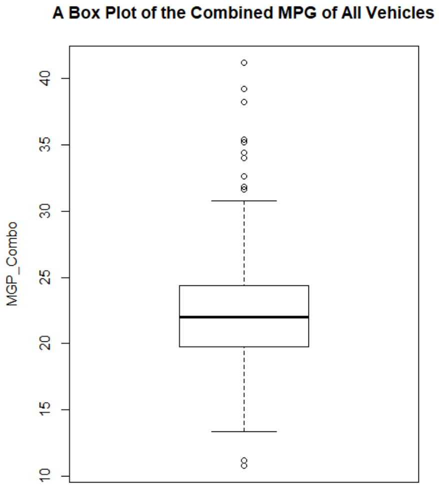
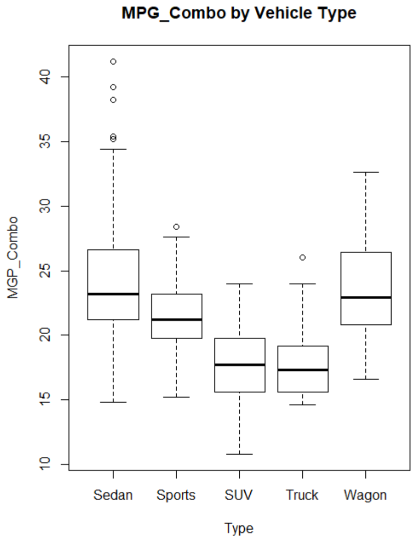
STA 6443 – Algorithms 1 HW1 solution

Exercise 1

(a) The box plot of the combined MPG has several outliers above and below the quartiles. Ignoring those outliers, the mean and median are quite close together (roughly 22 combined miles per gallon but slightly larger values of mean) and the spread of the distribution is not wide. It is slightly right skewed and most vehicles get between 20 and 25 miles per gallon (combined)

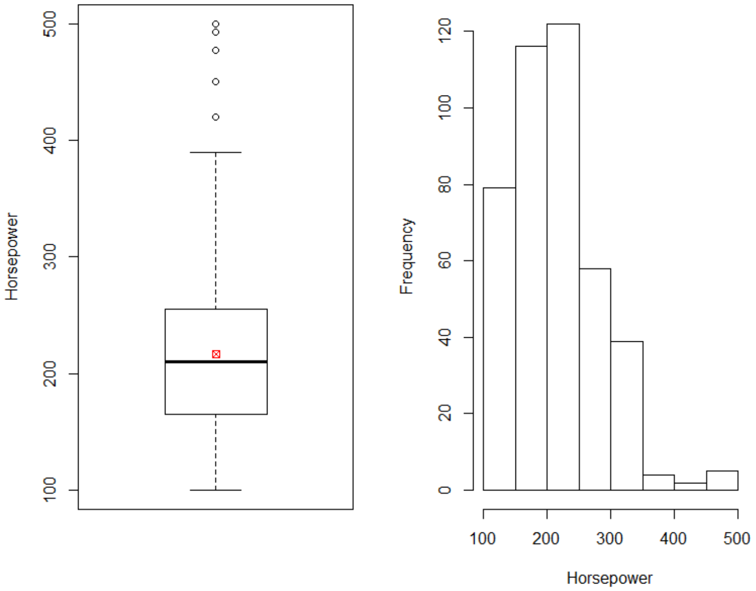


(b) When viewing the distribution of combined MPG separated by vehicle type, we see some interesting things. The combined MPG of SUVs and sports cars appear to be most like normal distributions - seemingly symmetric, with not wide spreads. Sedans' combined MPG have lots of variability and the most outliers. Combined MPG for wagons also has a widespread but may not be symmetric. Trucks appear to be least fuel efficient among the vehicle types. The distribution of combined MPG of trucks is right skewed and has at least one outlier.

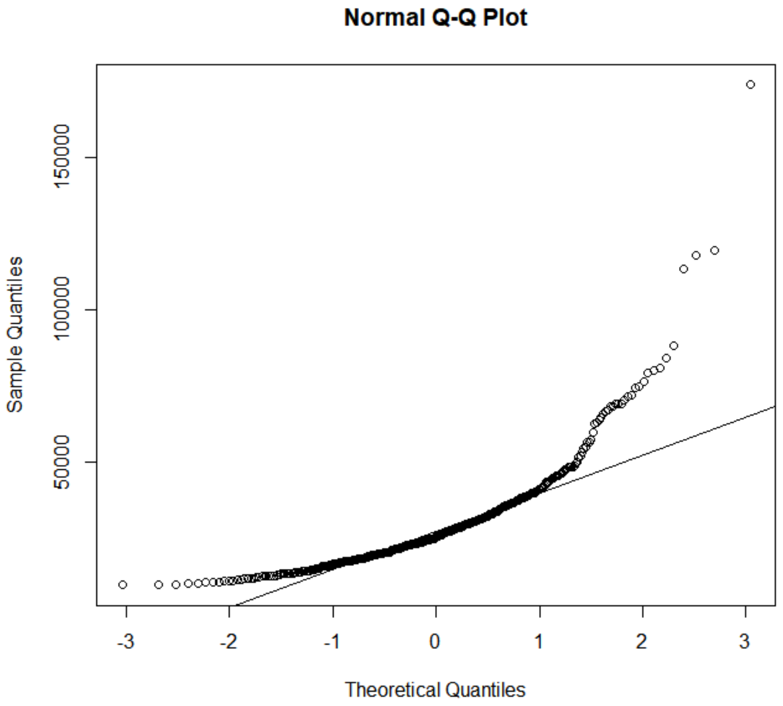


(c) The distribution of Horsepower is not normal as visible from the qqplot and strongly skewed as visible from the histogram. The mean and median are quite different from each other and it clarifies the asymmetry of the Horsepower variable's distribution. The qqplot does not follow the straight line and shapiro-Wilk test shows very small p-value, thus Horsepower does not follow the Normal distribution.

```
> summary(cars$Horsepower)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  100.0   165.0   210.0   216.8   255.0   500.0 
> mean(cars$Horsepower); var(cars$Horsepower)
[1] 216.76
[1] 5085.952
> range(cars$Horsepower) # [min,max]
[1] 100 500
> skewness(cars$Horsepower) # skewness
[1] 0.9528091
```

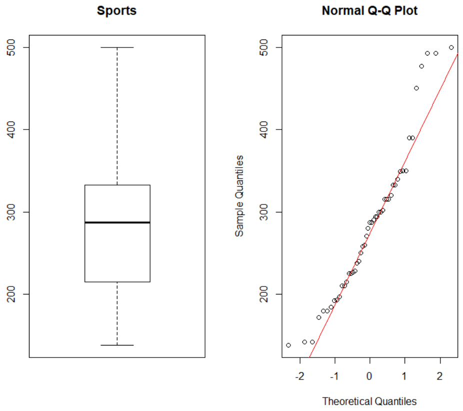


```
> shapiro.test(cars$Horsepower)

        Shapiro-Wilk normality test

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



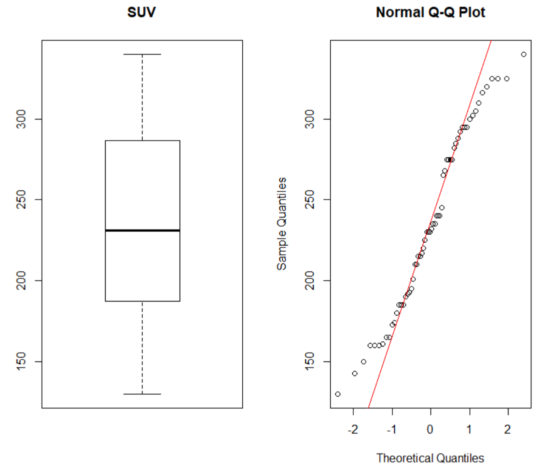
(d) There are 49 Sports cars, 60 SUV cars, and 24 Trucks in the data set. None of the distributions of Horsepower variable by Type are normal according to the tests of normality (all p-values are small, less than 0.05), histograms, and qqplots. The distribution of Horsepower is skewed and asymmetric for each of the 3 types. The respective mean and median for each of the 3 distributions are quite different from each other.

#Sports



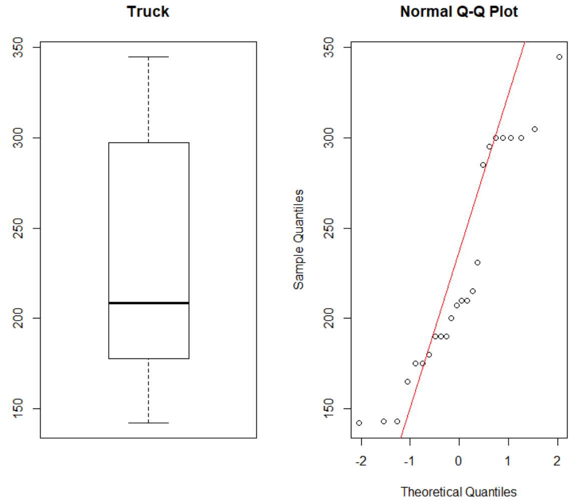
```
Shapiro-Wilk normality test

        data:  Sports$Horsepower
        W = 0.94276, p-value = 0.01898
```

#SUV



```
        Shapiro-Wilk normality test

data:  SUV$Horsepower
W = 0.95945, p-value = 0.04423
```

# Truck



```
Shapiro-Wilk normality test

data:  Truck$Horsepower
W = 0.8951, p-value = 0.01697
```

Exercise 2

(a) The distributions of Horsepower for SUV and Truck are not Normal from the normality tests above. Thus we should perform nonparametric **Wilcoxon rank-sum test**.

(b) H0: Distributions of Horsepower for SUV and Truck cars are from the same distribution
H1: One of the groups tends to be more efficient (either SUV or Truck) .

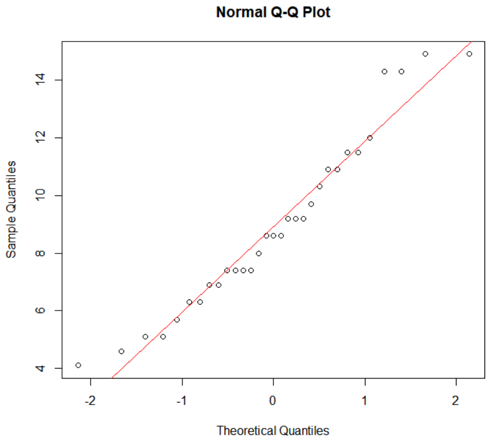
(c) We see the Wilcoxon rank-sum test do not reject the null hypothesis, with the larger p-value (greater than 0.05). Thus we conclude that the distributions of Horsepower for SUV and Truck cars are from the same distribution.

```
Wilcoxon rank sum test with continuity correction

data:  Truck$Horsepower and SUV$Horsepower
W = 633.5, p-value = 0.3942
alternative hypothesis: true location shift is not equal to 0
```

Exercise 3

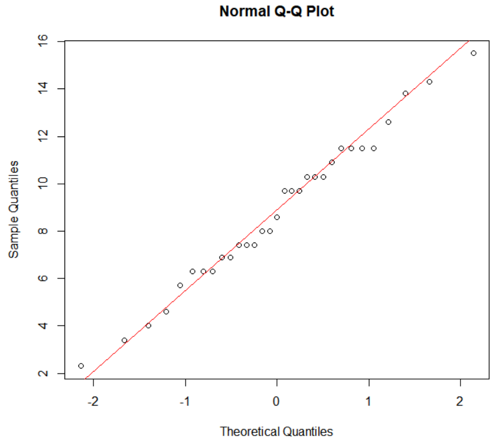
(a) The distributions of wind from July and August both follow Normal as we see the almost straight line in qqplots and large p-values (greater than significance level 0.05) on Shapiro-Wilk test. Thus we perform two-sample t-test. Then we check equal variance of two groups through equal variance test and find that two groups have the same variance with large p-values. So we can perform **pooled two-sample t-test**.

# July



```
            Shapiro-Wilk normality test

data:  July$Wind
W = 0.95003, p-value = 0.1564
```

August



```
        Shapiro-Wilk normality test

data:  Aug$Wind
W = 0.98533, p-value = 0.937
```

```
        F test to compare two variances

data:  July$Wind and Aug$Wind
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
```

(b) H0: mean(Wind of July) = mean(Wind of Aug)
H1: mean(Wind of July) != mean(Wind of Aug)

(c) We see large p-values on pooled two sample t-test and do not have enough evidence to reject the null hypothesis. The mean of Wind from July is equal to the mean of wind from August.

```
> t.test(July$Wind, Aug$Wind, var.equal = TRUE)

        Two Sample t-test

data:  July$Wind and Aug$Wind
t = 0.1865, df = 60, p-value = 0.8527
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.443108  1.739883
sample estimates:
mean of x mean of y
 8.941935  8.793548
```