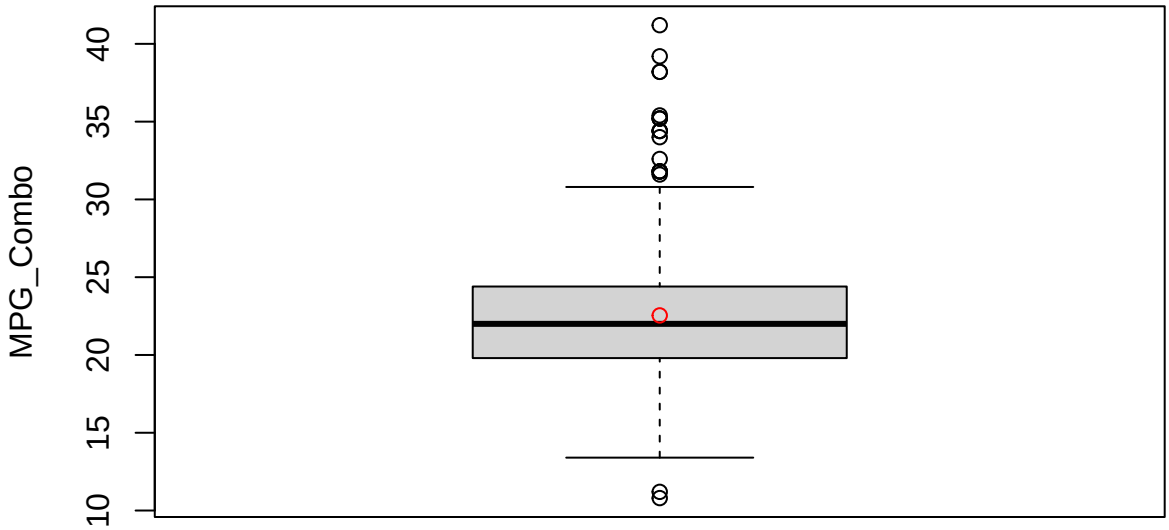
Algorithms 1 - HW 1

Erik Bratz / Diego Aldo Pettorossi / Nicholass Anderson

9/10/2021

NOTE: new version with minor changes allowed by the instruction. This is not a late submission

```
cars = read.csv("CARS.csv", header = TRUE)
MPG_Combo <- 0.6*cars$MPG_City+0.4*cars$MPG_Highway
cars=data.frame(cars, MPG_Combo)
boxplot(cars$MPG_Combo, ylab="MPG_Combo"); points(mean(cars$MPG_Combo), col="red")
```



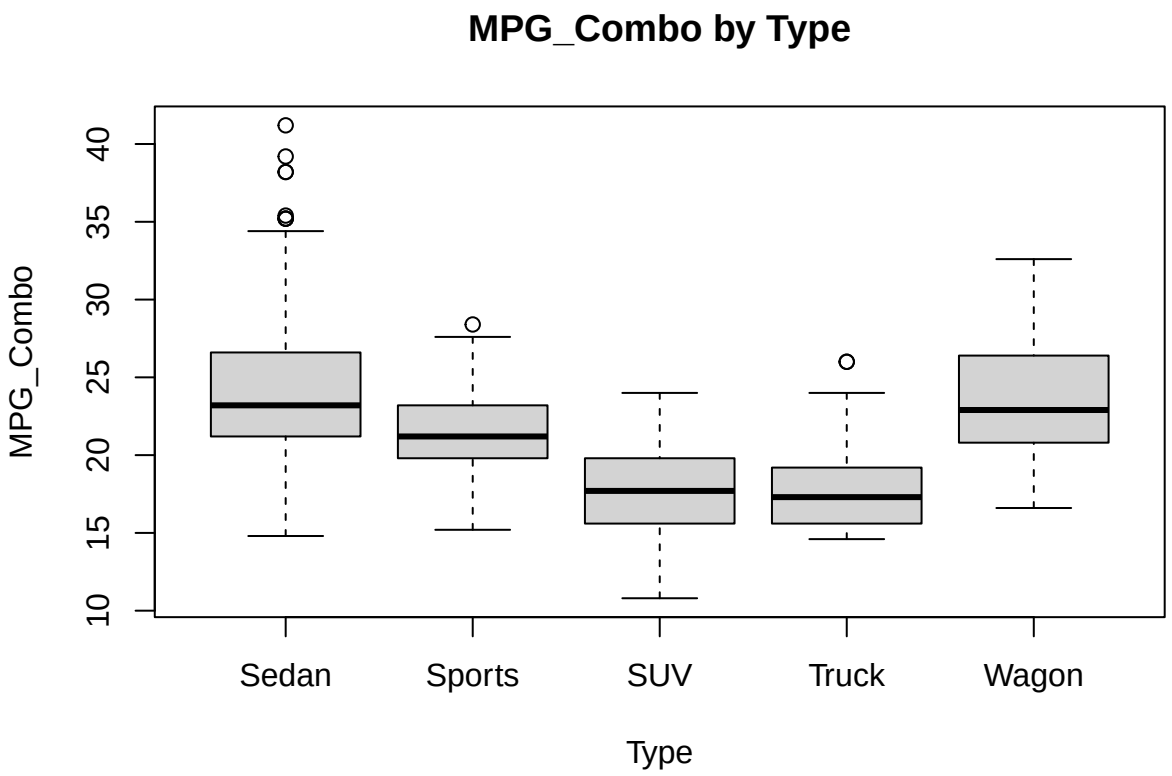

```
summary(MPG_Combo)
```

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

Beside mean and median are quite close, the bloxplot indicates we're dealing with a non-normal distribution as evident amount of outliers at both ends of the plot

(B)

```
boxplot(MPG_Combo ~ Type, data=cars, main="MPG_Combo by Type",
        xlab="Type", ylab="MPG_Combo");
```



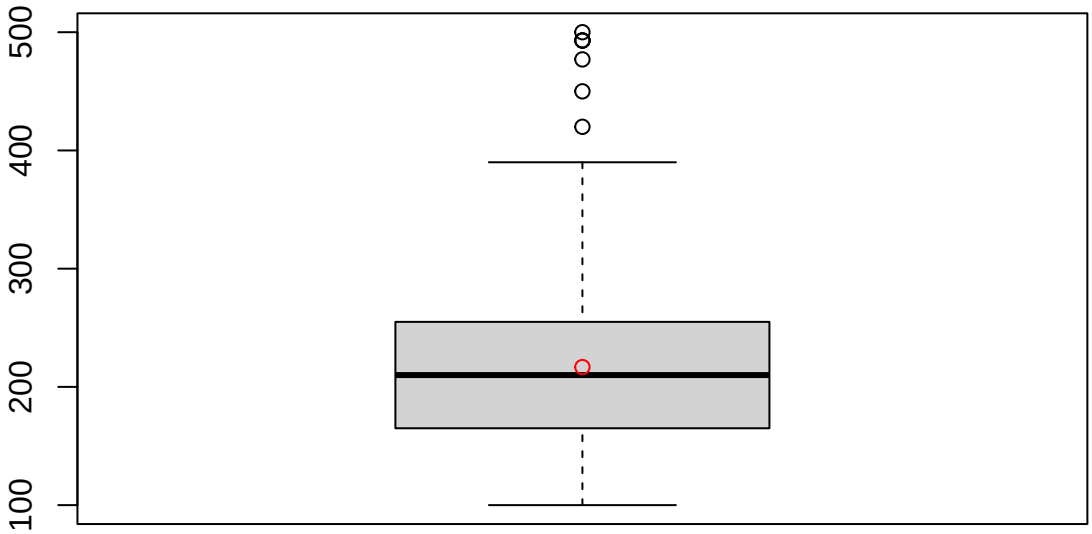
The Sedan overwhelmingly has outliers in the higher extremes. Sports, SUV, and Wagon all seem to have relatively evenly distributed data. Trucks data appears to extremely right skewed as a majority of the data falls near the bottom of the tail.

(C)

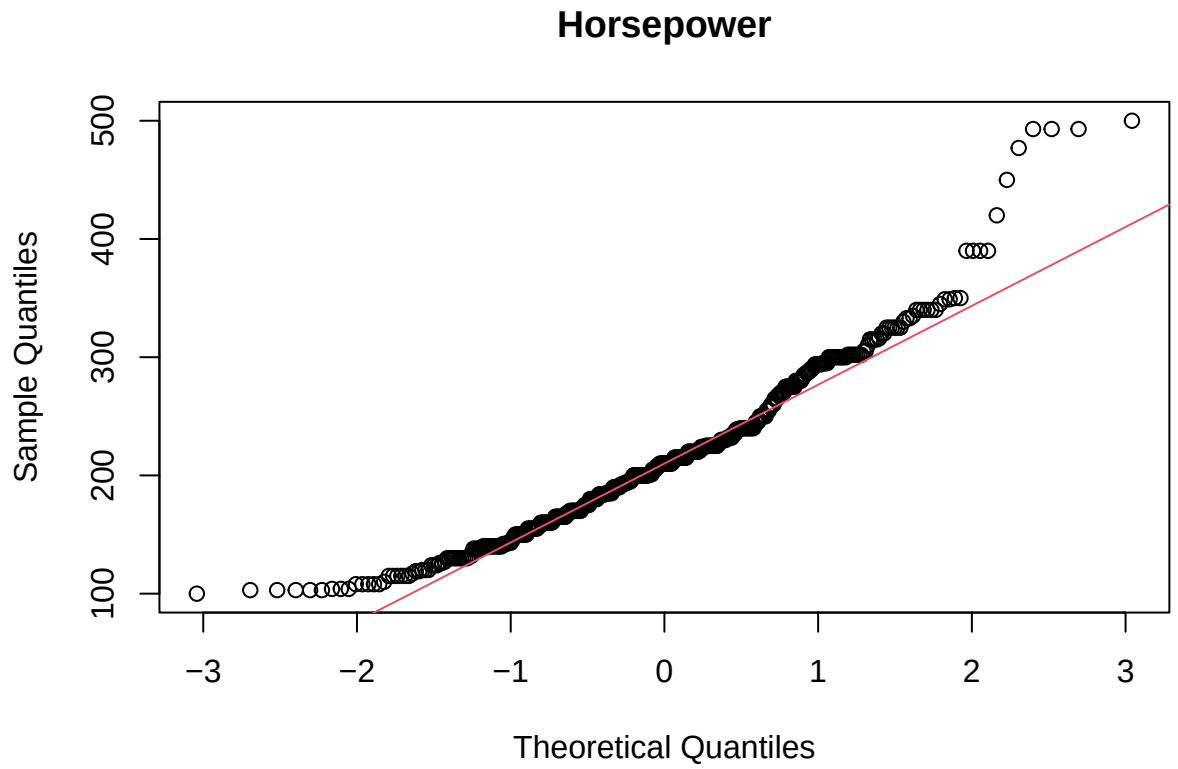
```
summary(cars$Horsepower)
```

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

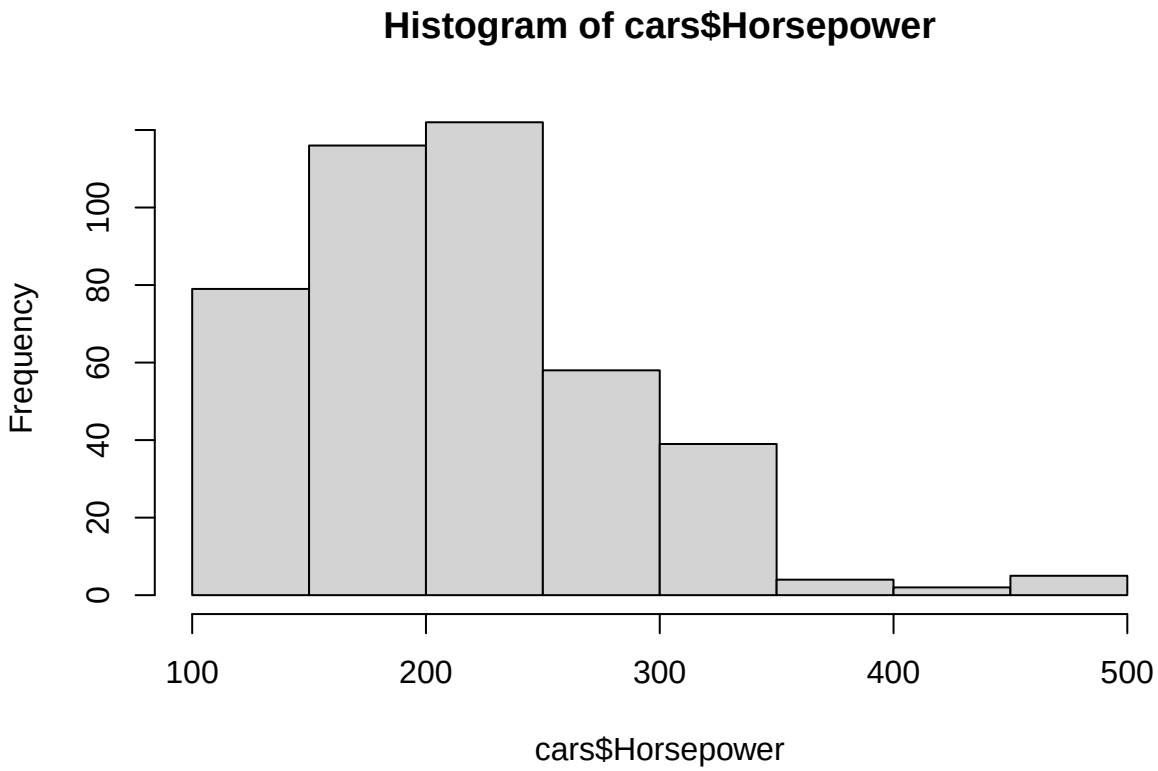
```
boxplot(cars$Horsepower); points(mean(cars$Horsepower), col="red")
```



```r
qqnorm(cars$Horsepower, main = ("Horsepower"))
qqline(cars$Horsepower, col = 2)
```




```r
hist(cars$Horsepower)
```



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

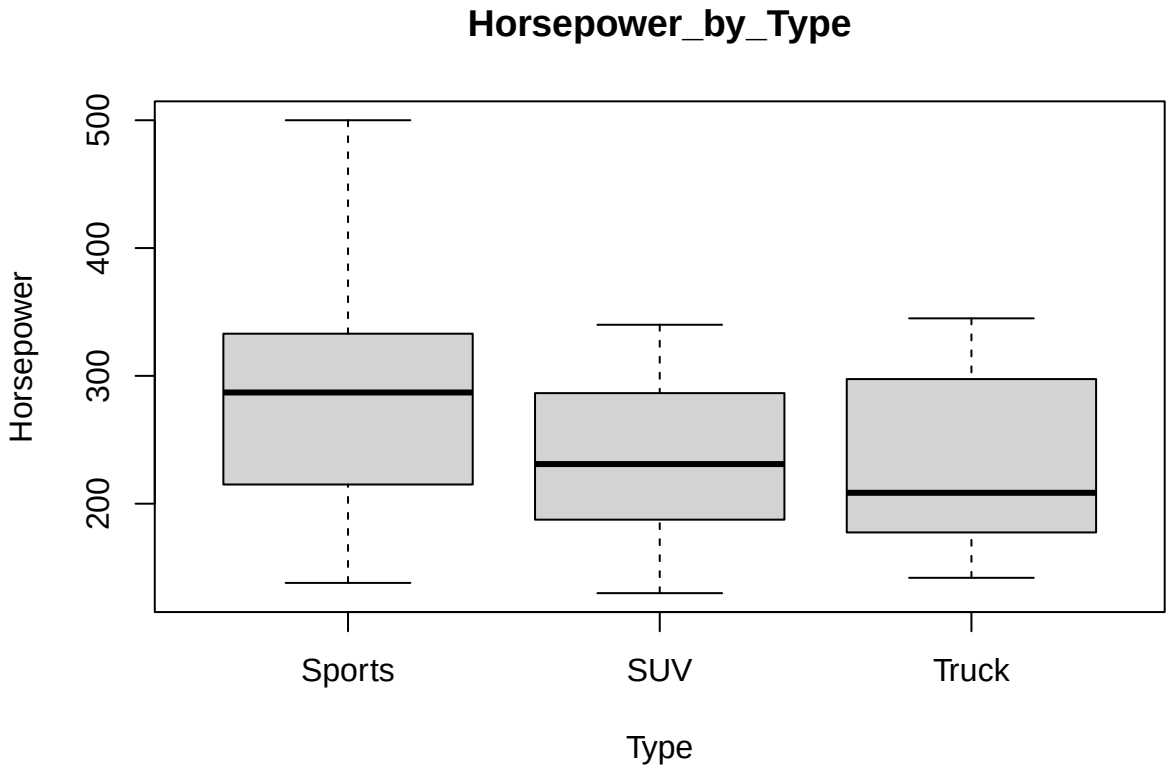
```
##
##  Shapiro-Wilk normality test
##
## data:  cars$Horsepower
## W = 0.94573, p-value = 2.32e-11
```

The presence of outliers in our boxplot, the presence of right skewed tail on the histogram and deviation from the QQline all indicates signs of non-normal distribution.

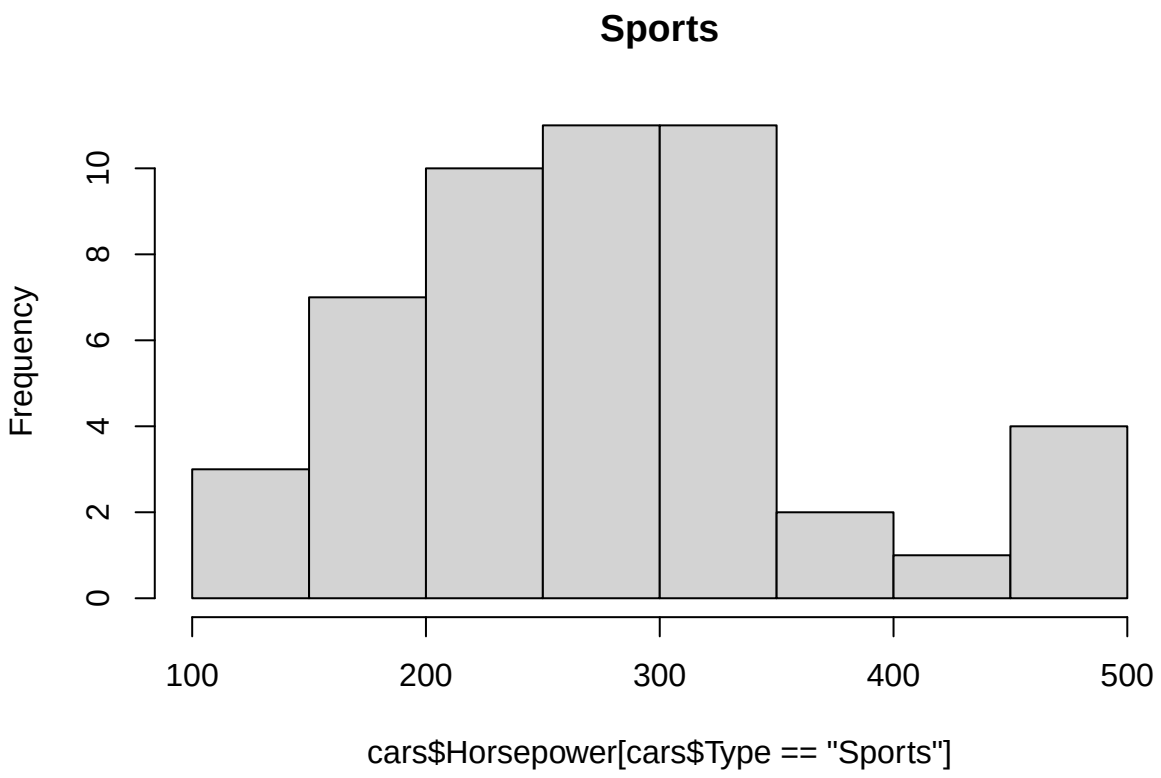
The p-value obtained from the Shapiro test valides our observations (p-value < 0.05).

(D)

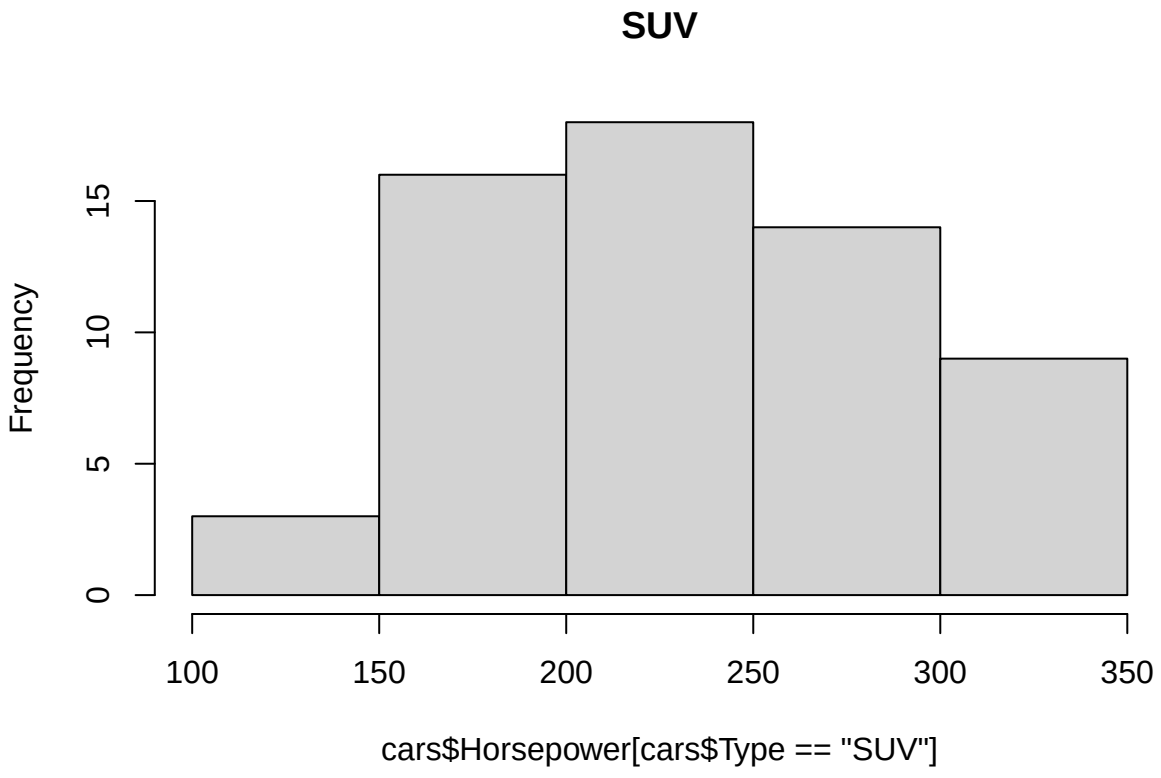
```
cars_subset1 <- subset(cars,Type == "SUV" | Type =="Truck" | Type == "Sports")
boxplot(Horsepower ~ Type, data=cars_subset1, main="Horsepower_by_Type",
        xlab="Type", ylab="Horsepower")
```



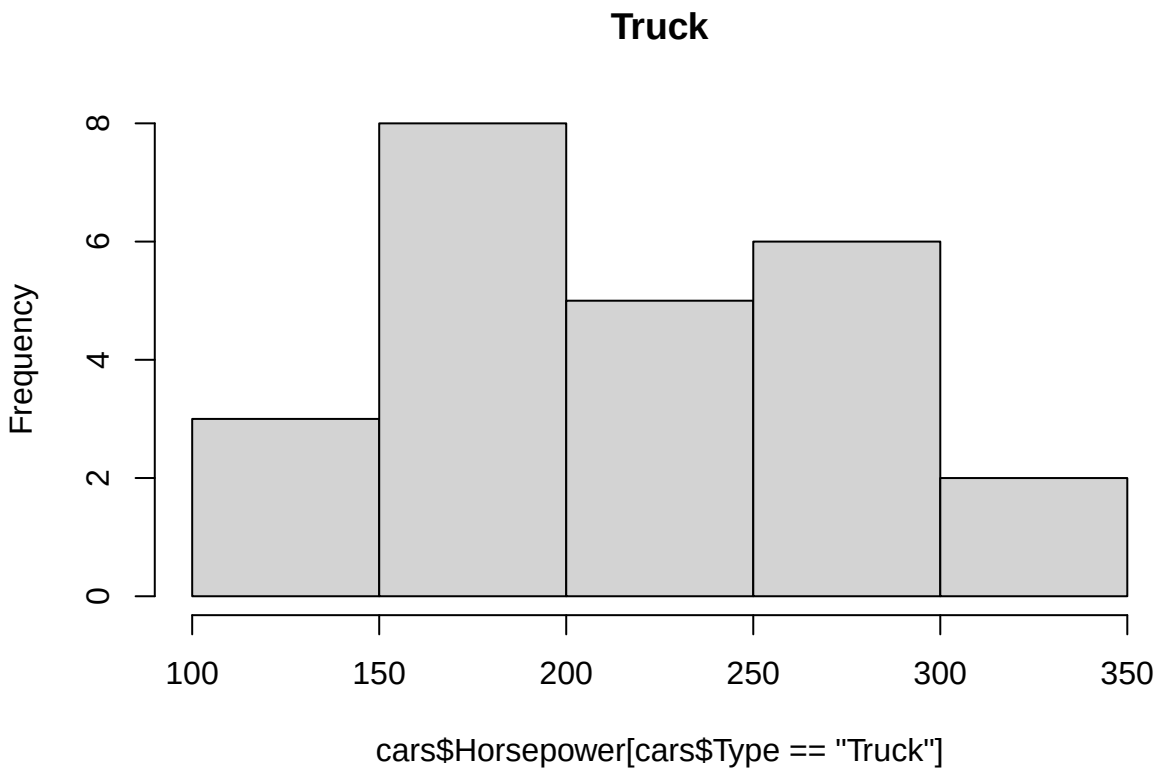

```
hist(cars$Horsepower[cars$Type=="Sports"], main ="Sports")
```



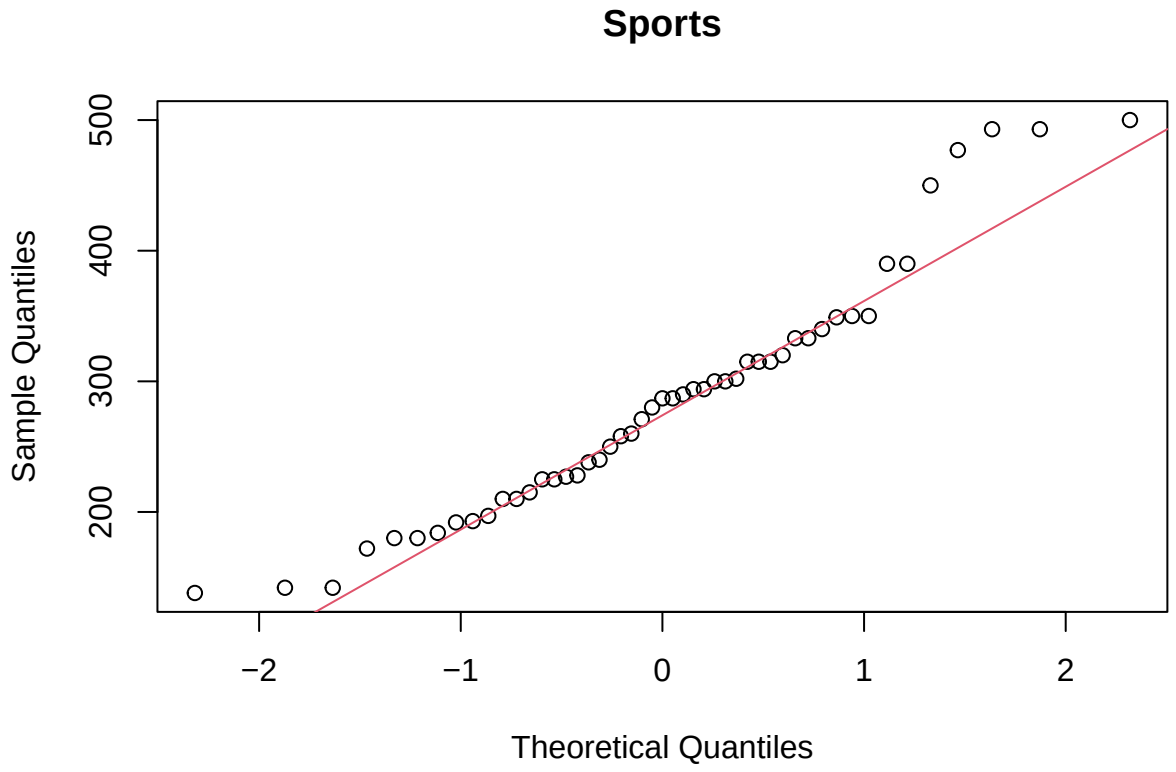

```
hist(cars$Horsepower[cars$Type=="SUV"], main ="SUV")
```



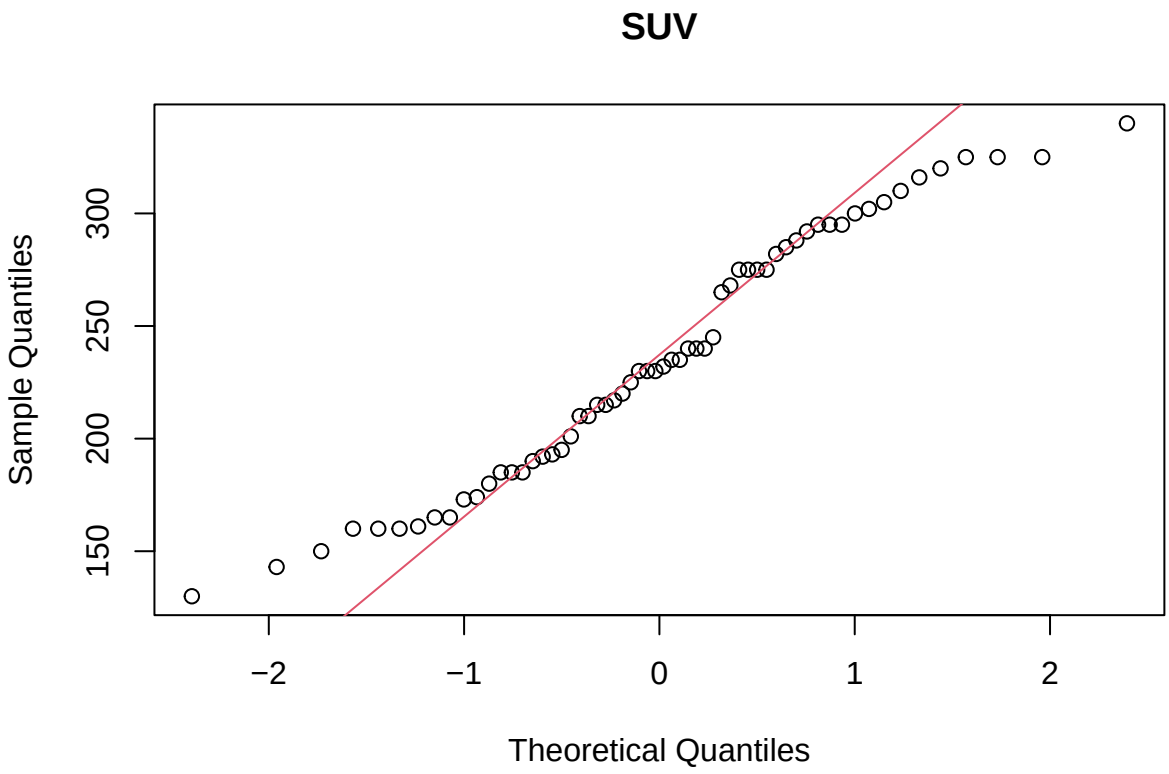

```
hist(cars$Horsepower[cars$Type=="Truck"], main ="Truck")
```




```
qqnorm(cars$Horsepower[cars$Type=="Sports"], main ="Sports")
qqline(cars$Horsepower[cars$Type=="Sports"], col = 2)
```




```
qqnorm(cars$Horsepower[cars$Type=="SUV"], main ="SUV")
qqline(cars$Horsepower[cars$Type=="SUV"], col = 2)
```




```
qqnorm(cars$Horsepower[cars$Type=="Truck"], main ="Truck")
qqline(cars$Horsepower[cars$Type=="Truck"], col = 2)
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  cars$Horsepower[cars$Type == "Sports"]
## W = 0.94276, p-value = 0.01898
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  cars$Horsepower[cars$Type == "SUV"]
## W = 0.95945, p-value = 0.04423
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  cars$Horsepower[cars$Type == "Truck"]
## W = 0.8951, p-value = 0.01697
```

While sports and trucks have a right skewed tail distribution, SUV distribution appears to be symmetrical based on the boxplot. All vehicle types present a right skewed tail distribution based on the histograms. On the QQPlot we can see how the central SUV observations look normally distributed excluding the tail the tail end, which appear to be deviating from the normal. The shapiro test results indicates that SUVs, sports and trucks are not normally distributed.

Exercise 2

```
cars_subset <- subset(cars,Type == "SUV" | Type =="Truck")
wilcox.test(Horsepower ~ Type, data=cars_subset, 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
```

(a)

We performed the Wilcoxon rank-sum test since the distributions of SUVs and trucks are not normal. In fact, the p-values obtained in the Shapiro tests are greater than the significance level.

(b)

H0: The horsepower of SUVs and Trucks are not statistically different

H1: The horsepower of SUVs and Trucks are statistically different

(c)

Since the p-value obtained is greater than the significance level we can't reject the null hypotesis. Therefore, the horsepower of SUVs and Trucks are not statistically different (p-value= 0.3942)

Exercise 3

(a)

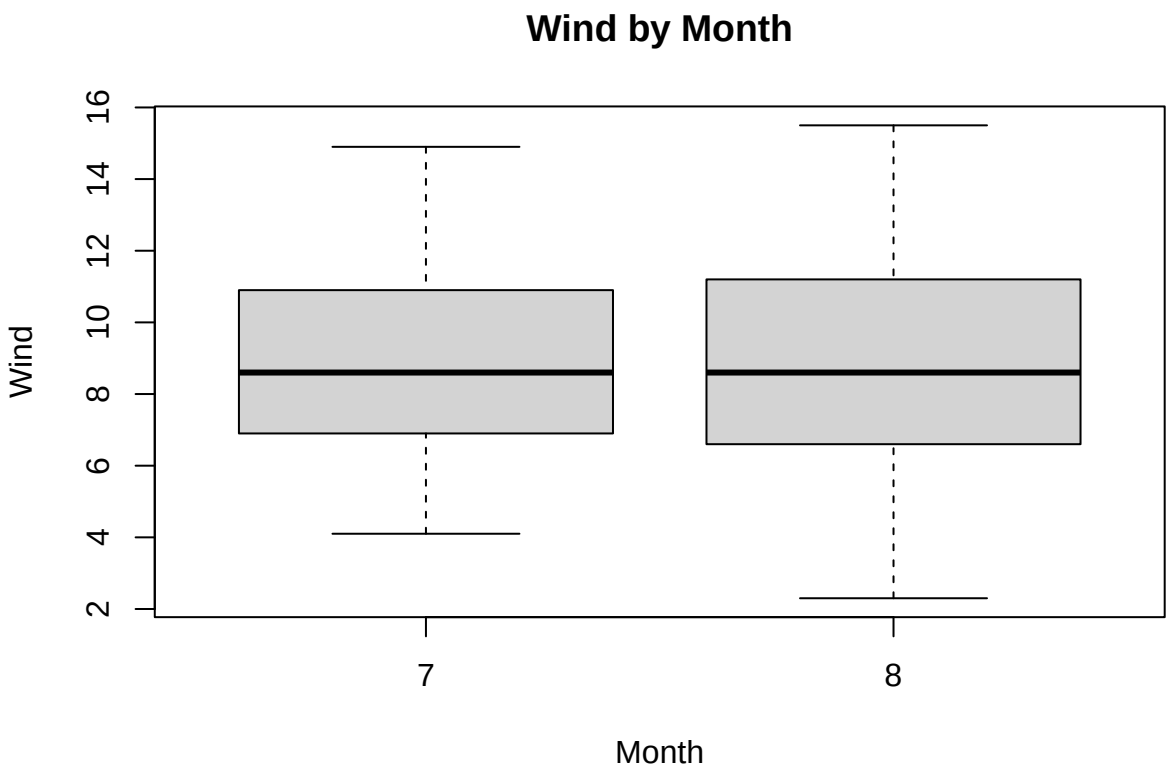
```
aq <- airquality
july <- subset(aq, Month == 7)
aug <- subset(aq, Month == 8)
julaug <- subset(aq, Month == "7" | Month == "8")

summary(july)
```

```
##      Ozone           Solar.R          Wind             Temp          Month
##  Min.   :  7.00   Min.   :  7.0   Min.   : 4.100   Min.   :73.0   Min.   :7
##  1st Qu.: 36.25   1st Qu.:175.0   1st Qu.: 6.900   1st Qu.:81.5   1st Qu.:7
##  Median : 60.00   Median :253.0   Median : 8.600   Median :84.0   Median :7
##  Mean   : 59.12   Mean   :216.5   Mean   : 8.942   Mean   :83.9   Mean   :7
##  3rd Qu.: 79.75   3rd Qu.:273.0   3rd Qu.:10.900   3rd Qu.:86.0   3rd Qu.:7
##  Max.   :135.00   Max.   :314.0   Max.   :14.900   Max.   :92.0   Max.   :7
##  NA's   :5
##       Day
##  Min.   : 1.0
##  1st Qu.: 8.5
##  Median :16.0
##  Mean   :16.0
##  3rd Qu.:23.5
##  Max.   :31.0
##
```

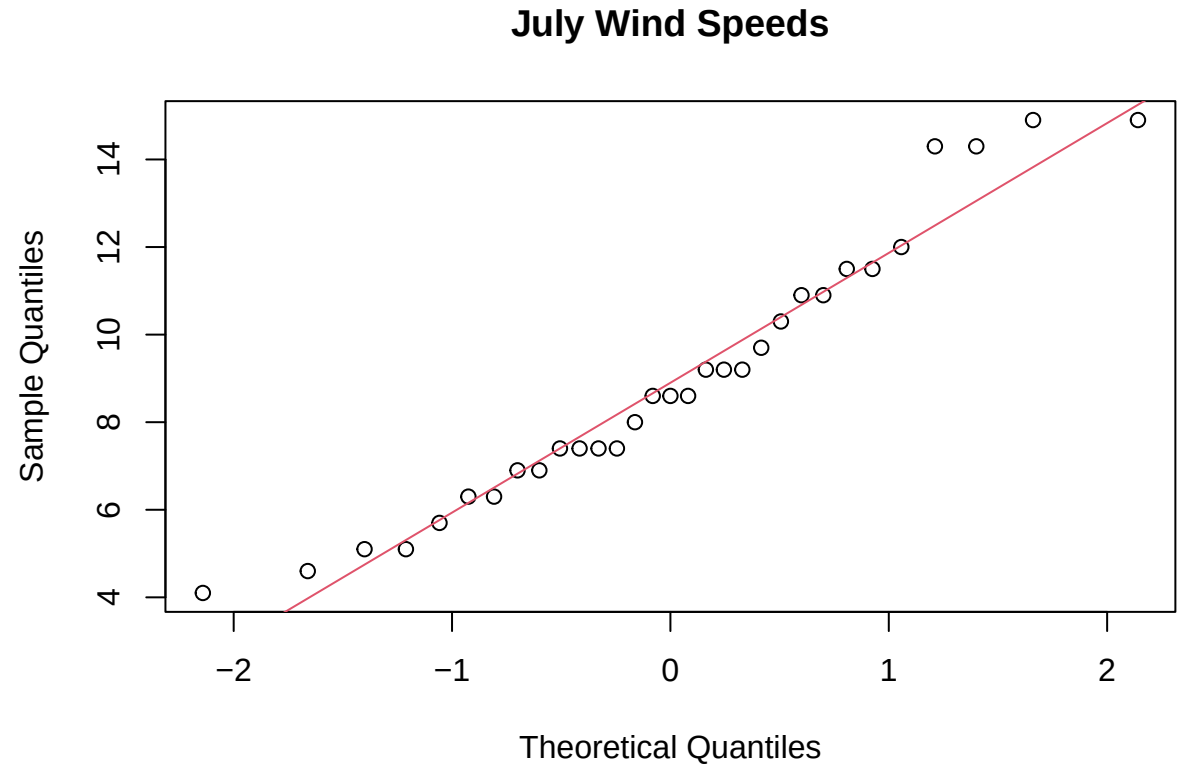
```
summary(aug)
```

```
##      Ozone           Solar.R          Wind             Temp          Month
##  Min.   :  9.00   Min.   : 24.0   Min.   : 2.300   Min.   :72.00   Min.   :8
##  1st Qu.: 28.75   1st Qu.:107.0   1st Qu.: 6.600   1st Qu.:79.00   1st Qu.:8
##  Median : 52.00   Median :197.5   Median : 8.600   Median :82.00   Median :8
##  Mean   : 59.96   Mean   :171.9   Mean   : 8.794   Mean   :83.97   Mean   :8
##  3rd Qu.: 82.50   3rd Qu.:231.0   3rd Qu.:11.200   3rd Qu.:88.50   3rd Qu.:8
##  Max.   :168.00   Max.   :273.0   Max.   :15.500   Max.   :97.00   Max.   :8
##  NA's   :5        NA's   :3
##       Day
##  Min.   : 1.0
##  1st Qu.: 8.5
##  Median :16.0
##  Mean   :16.0
##  3rd Qu.:23.5
##  Max.   :31.0
##
```

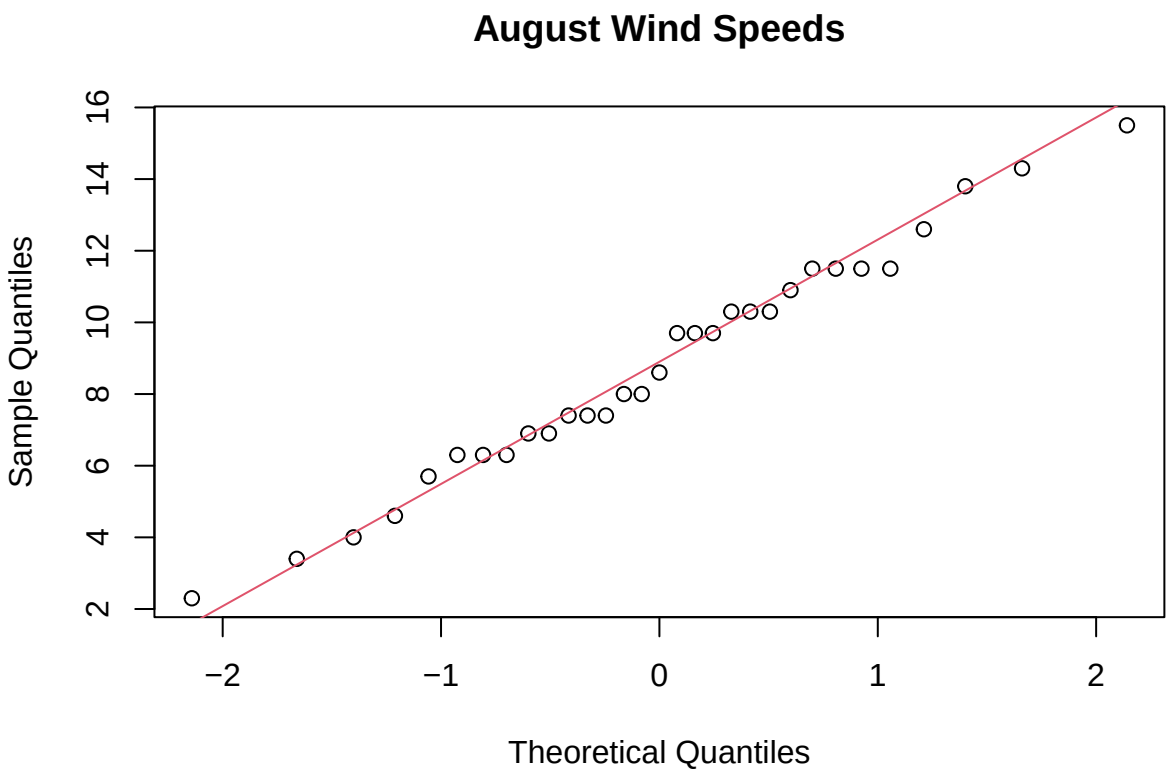
```
boxplot(Wind ~ Month, data = julaug, main = "Wind by Month",
        xlab = "Month", ylab = "Wind")
```




```
qqnorm(july$Wind, main = "July Wind Speeds")
qqline(july$Wind, col = 2)
```




```
qqnorm(aug$Wind, main = "August Wind Speeds")
qqline(aug$Wind, col = 2)
```




```
shapiro.test(july$Wind)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  july$Wind
## W = 0.95003, p-value = 0.1564
```

```
shapiro.test(aug$Wind)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aug$Wind
## W = 0.98533, p-value = 0.937
```

```
var.test(Wind ~ Month, julaug, 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
```

```
bartlett.test(Wind ~ Month, julaug)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Wind by Month
## Bartlett's K-squared = 0.10861, df = 1, p-value = 0.7417
```

Both the p values obtained from the Shapiro Test for July and August's wind speeds are greater than alpha. Thus indicating that both distributions are normal. This is corroborated by the visual representation provided by the qqplot generated, as both months showed few outliers and deviations from the qqline.

After performing a Bartlett test, the p value was greater than alpha, our level of significance. Providing evidence that the variance between the two variables were the same.

Since the two variances were the same, this justifies using a pooled t-test. The hypothesis for this is below in section (b).

(b)

H0: There is no statistical difference between the mean Wind speed in July and August

H1: There is a statistical difference between the mean Wind speed in July and August

```
t.test(Wind ~ Month, julaug, 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
```

(c)

Since the p-value obtained is greater than the significance level we can't reject the null hypothesis. Therefore, the mean Wind Speed in July and August are not statistically different (p-value= 0.8527)