HW 1

Ashley Cortez and Jacob Kendall

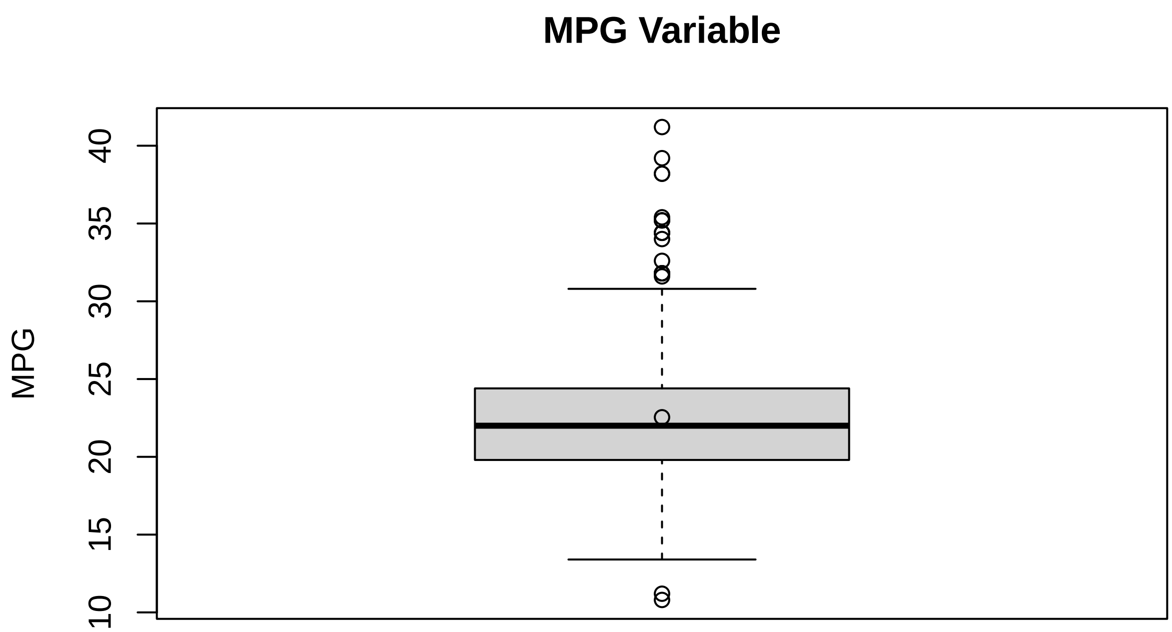
9/5/2021

HW1

setwd() # set your own path

Exercise 1: Descriptive Statistics

```
#1 (a)

cars=read.csv("Cars.csv", header = TRUE)  # read data set
MPG_Combo <- 0.6*cars$MPG_City+0.4*cars$MPG_Highway  # combined mpg variable
cars=data.frame(cars, MPG_Combo)   # data frame with MPG_Combo
boxplot(MPG_Combo, main = "MPG Variable", ylab = "MPG") #Boxplot of MPG_Combo
points(mean(MPG_Combo, na.rm = TRUE)) #Mean point on boxplot
```

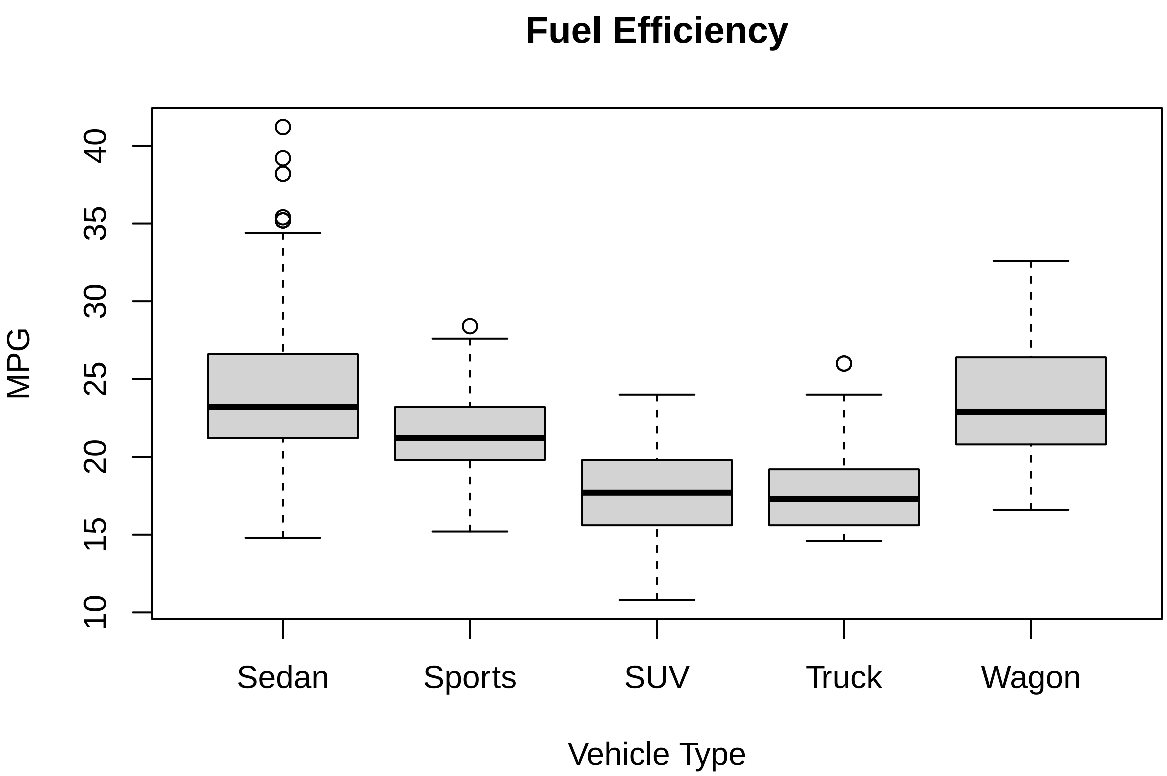



```
summary(MPG_Combo)
```

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

```
#The box plot indicates that the median miles per gallon among vehicles is 22 mpg. The mean of 22.54 is
```

```
# 1 (b) Obtain box plots for MPG_Combo by Type and comment on any differences you notice between the di

plot(factor(cars$Type), cars$MPG_Combo, main ="Fuel Efficiency", ylab = "MPG", xlab = "Vehicle Type") #
```




```
SUV = subset(cars, Type == "SUV")  #Subset SUV as dataframe
summary(SUV$MPG_Combo) #Summarize SUV
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.80   15.60   17.70   17.86   19.70   24.00
```

```
Sports = subset(cars, Type == "Sports") #Subset Sports as dataframe
summary(Sports$MPG_Combo) #Summarize Sports
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.20   19.80   21.20   21.24   23.20   28.40
```

```
Truck = subset(cars, Type == "Truck") #Subset Truck as dataframe
summary(Truck$MPG_Combo) #Summarize Sports
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.6    15.6    17.3    18.3    18.8    26.0
```

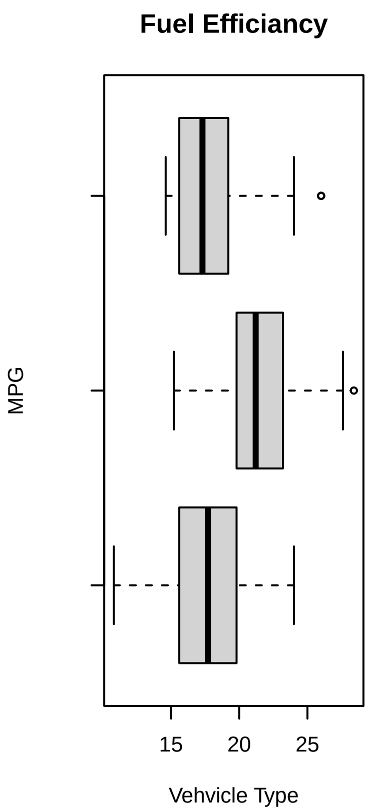
```r
Sedan =subset(cars, Type == "Sedan")  #Subset Sedan as dataframe
summary(Sedan$MPG_Combo) #Summarize Sedan
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.80   21.20   23.20   24.10   26.55   41.20
```

```r
Wagon = subset(cars, Type == "Wagon")  #Subset Wagon as dataframe
summary(Wagon$MPG_Combo) #Summarize Wagon
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.60   20.95   22.90   23.82   26.30   32.60
```

```r
####Combine Boxplots####
par(mfrow = c(1,3))
boxplot(SUV$MPG_Combo, Sports$MPG_Combo, Truck$MPG_Combo, main= "Fuel Efficiancy", ylab = "MPG", xlab =
```




```r
# 1 (c) Obtain basic descriptive statistics for Horsepower for all vehicles. Comment on any general fea

library(moments)

summary(cars$Horsepower, na.rm = TRUE) #Descriptive Stats Snapshot
```

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

```r
#####Qualitative#####
#Min 100, 1st Qtr 165, Median 210, Mean 216.8, 3rd Qtr 255, Max 500
#Note that the mean > median, slightly right skewed
var(cars$Horsepower, na.rm = TRUE) #Variance
```

```
## [1] 5085.952
```

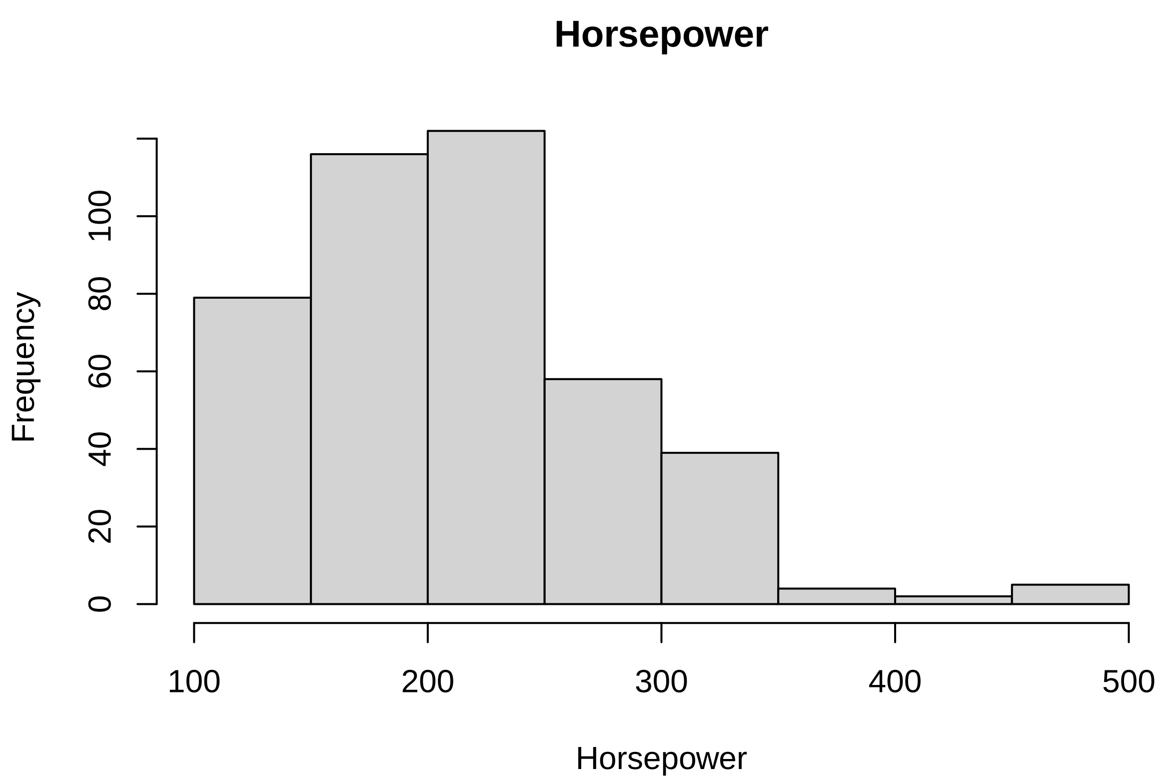
```r
skewness(cars$Horsepower, na.rm = TRUE) #Skewness
```

```
## [1] 0.9561818
```

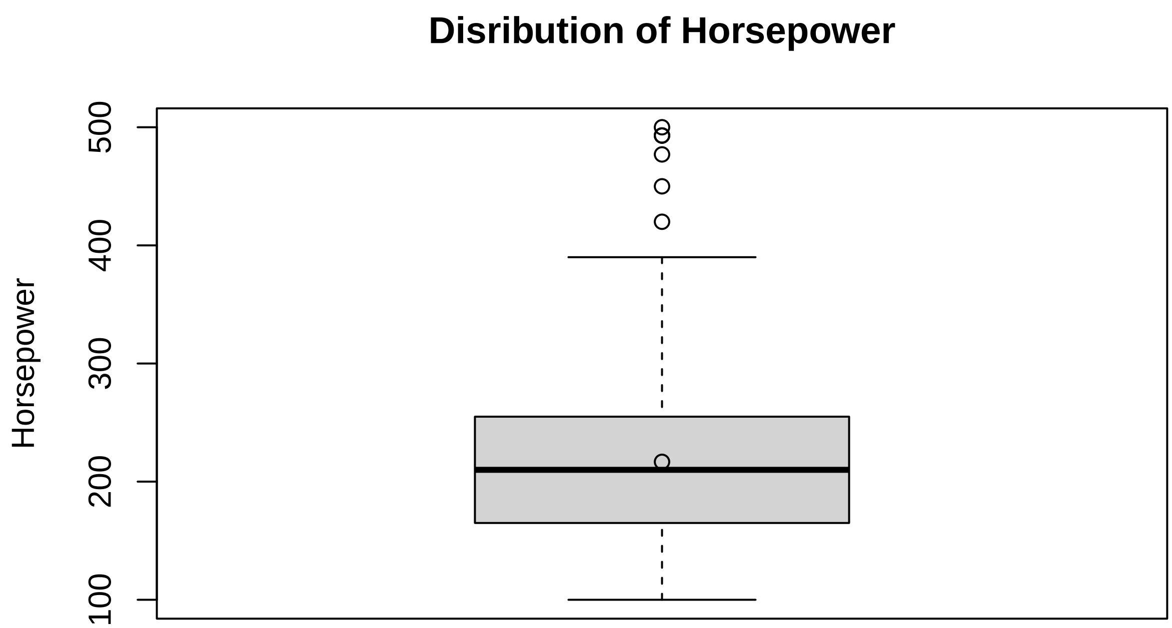
```r
range(cars$Horsepower, na.rm = TRUE)  #Range
```

```
## [1] 100 500
```

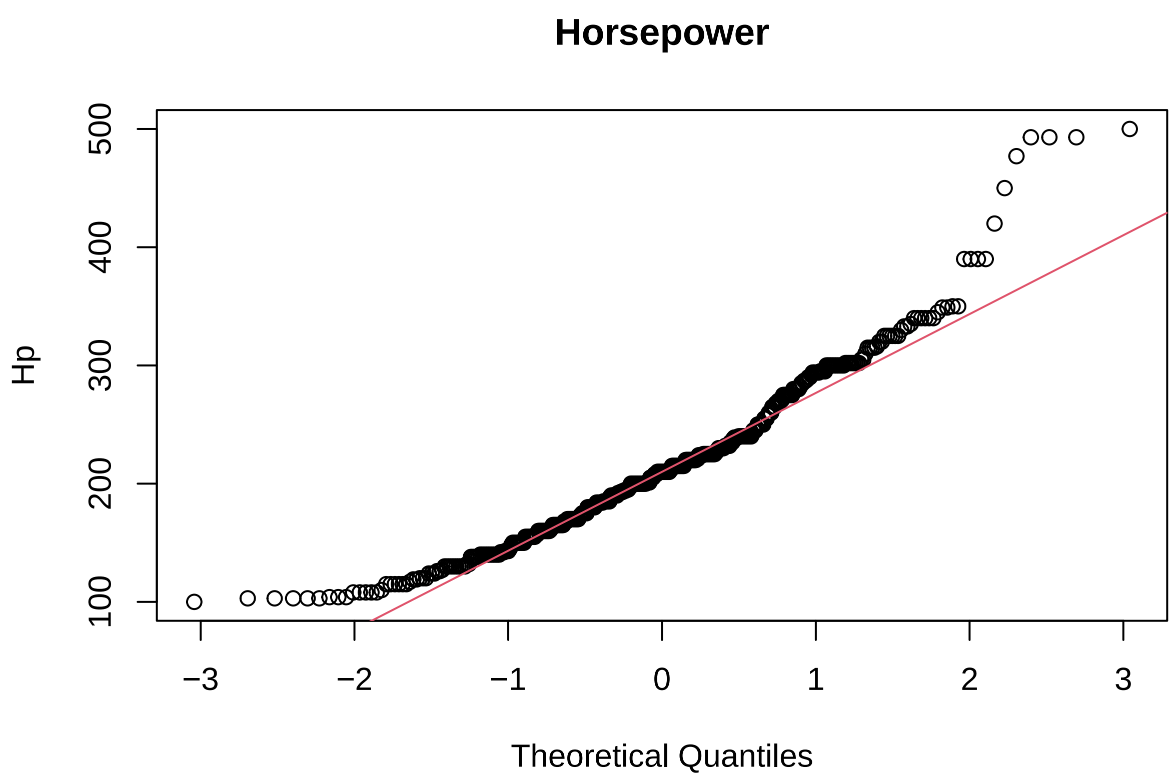
```r
hist(cars$Horsepower, main = "Horsepower", xlab = "Horsepower") #Histogram of Horsepower
```




```r
boxplot(cars$Horsepower, main = "Disribution of Horsepower", ylab = "Horsepower") #Boxplot of Horsepowr
points(mean(cars$Horsepower, na.rm = TRUE)) #Mean point on boxplot
```




```
qqnorm(cars$Horsepower, main = "Horsepower", ylab = "Hp"); qqline(cars$Horsepower, col = 2) #Normality .
```

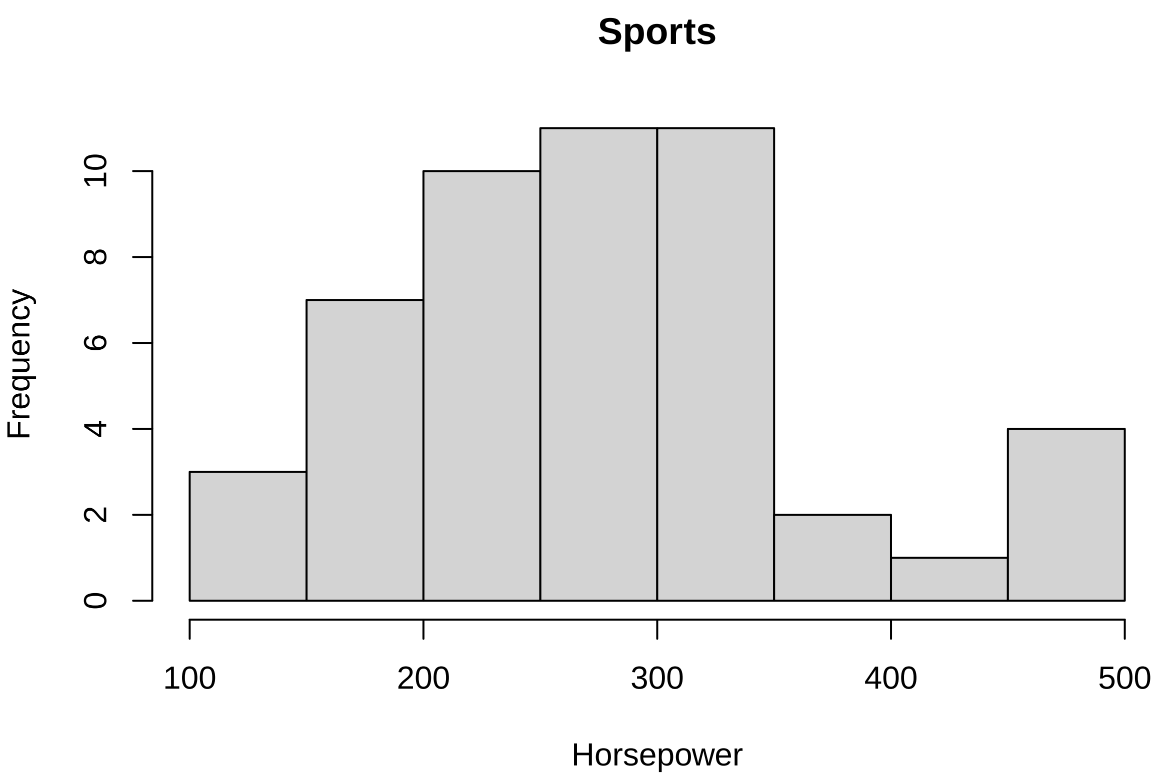



```
#####Quantitative#####
shapiro.test(cars$Horsepower)
```

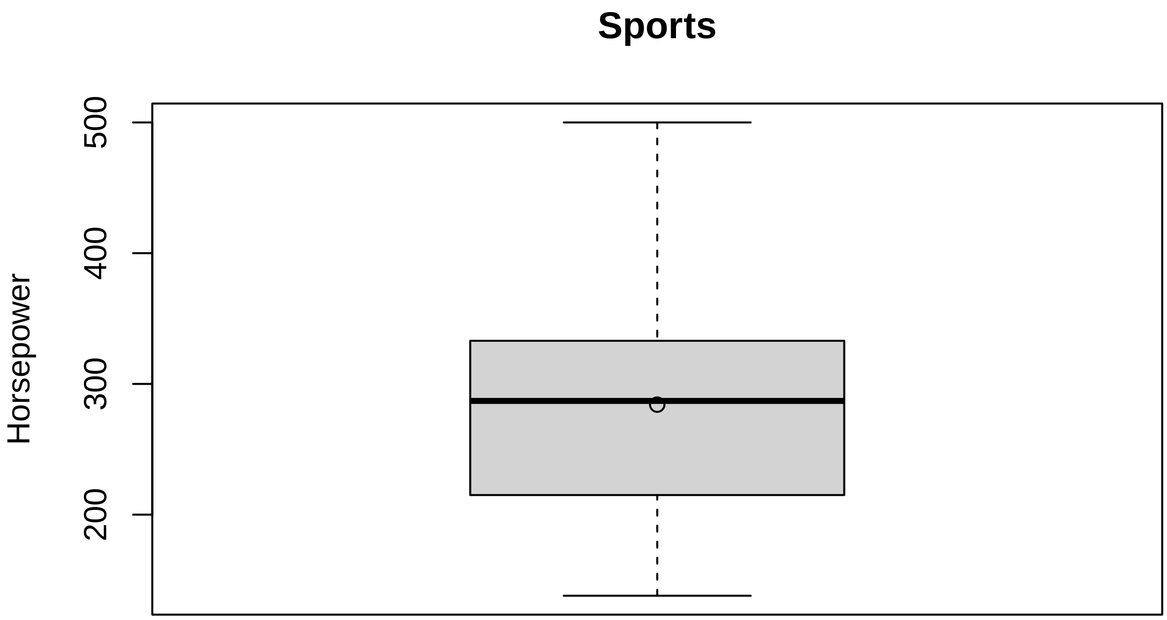
```
##
##  Shapiro-Wilk normality test
##
```

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

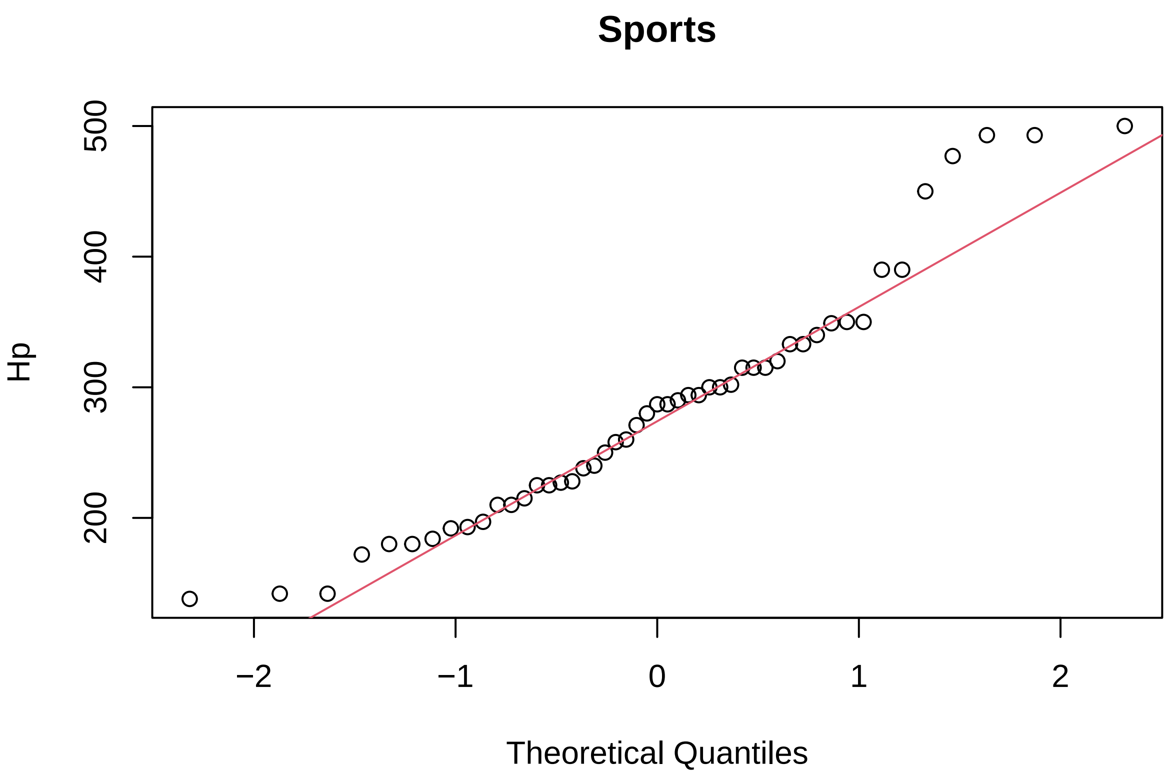
```r
#1d

library(moments)

#i) Sports

hist(Sports$Horsepower, main = "Sports", xlab = "Horsepower") #Histogram of Sports
```




```r
boxplot(Sports$Horsepower, main = "Sports", ylab = "Horsepower") # Boxplot for Sports
points(mean(Sports$Horsepower, na.rm = TRUE)) #Mean point on boxplot
```




```
qqnorm(Sports$Horsepower, main = "Sports", ylab = "Hp"); qqline(Sports$Horsepower, col = 2) #Normality .
```




```
#Normality Information
var(Sports$Horsepower, na.rm = TRUE) #Variance
```

```
## [1] 8609.931
```

```r
skewness(Sports$Horsepower, na.rm = TRUE) #Skewness
```

```
## [1] 0.6808686
```

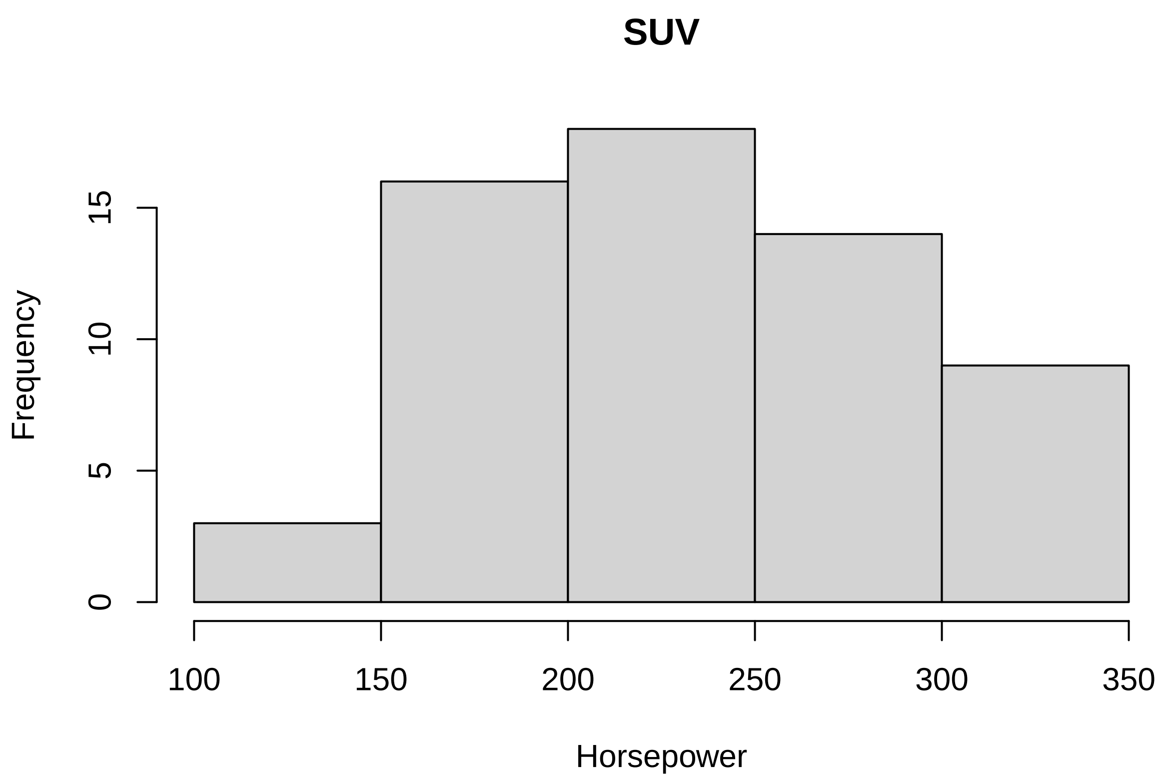
```r
range(Sports$Horsepower, na.rm = TRUE)  #Range
```

```
## [1] 138 500
```

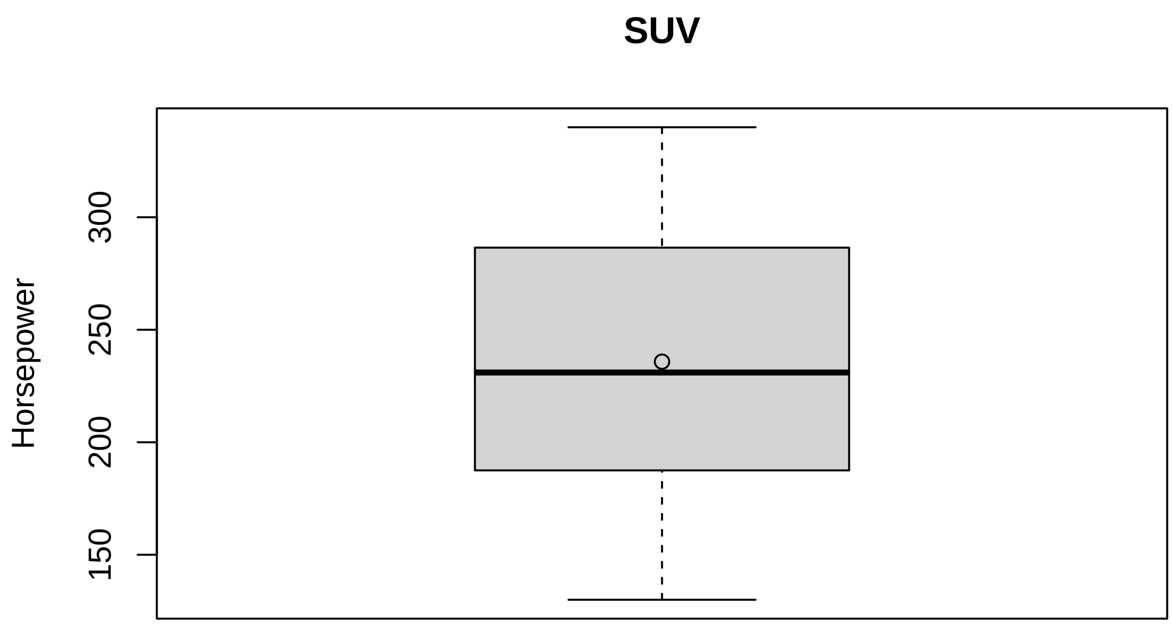
```r
#Qunatitative Test
shapiro.test(Sports$Horsepower)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Sports$Horsepower
## W = 0.94276, p-value = 0.01898
```

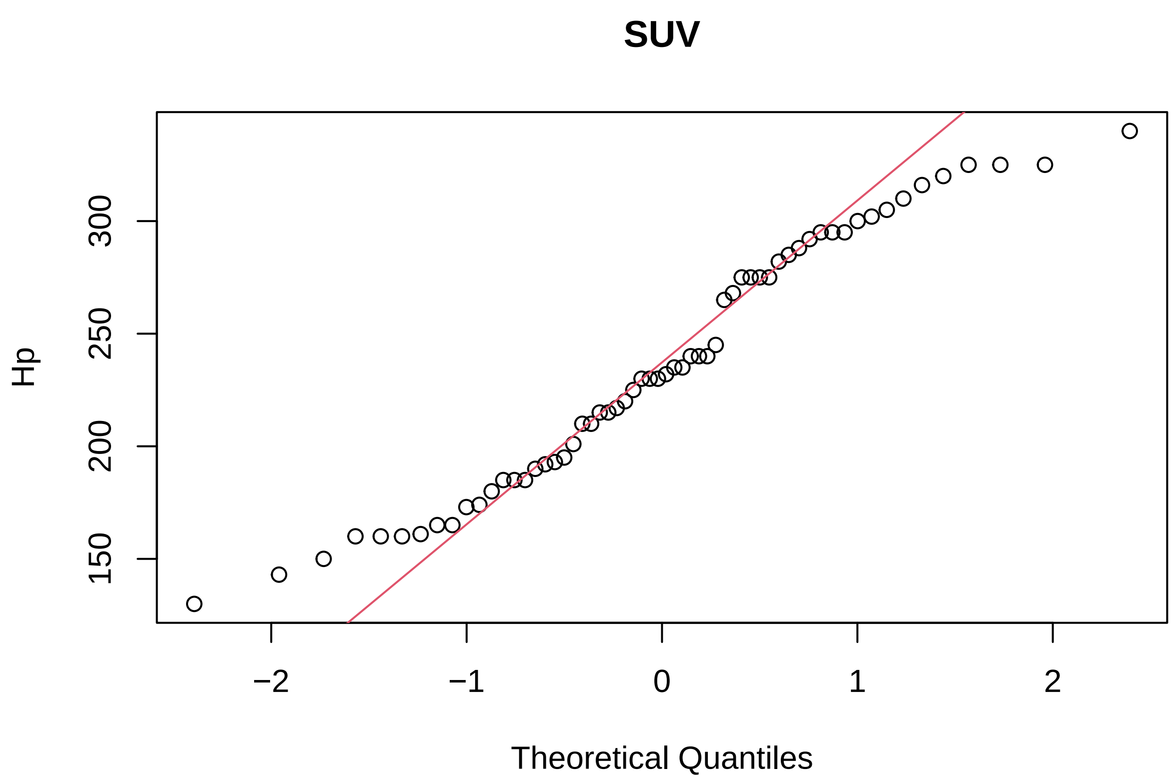
```r
#ii) SUV

hist(SUV$Horsepower, main = "SUV", xlab = "Horsepower") #Histogram of SUV
```




```r
boxplot(SUV$Horsepower, main = "SUV", ylab = "Horsepower") # Boxplot for SUV
points(mean(SUV$Horsepower, na.rm = TRUE)) #Mean point on boxplot
```




```
qqnorm(SUV$Horsepower, main = "SUV", ylab = "Hp"); qqline(SUV$Horsepower, col = 2) #Normality Plot
```




```
#Normality Information
var(SUV$Horsepower, na.rm = TRUE) #Variance
```

```
## [1] 3162.254
```

```r
skewness(SUV$Horsepower, na.rm = TRUE) #Skewness
```

```
## [1] 0.05619688
```

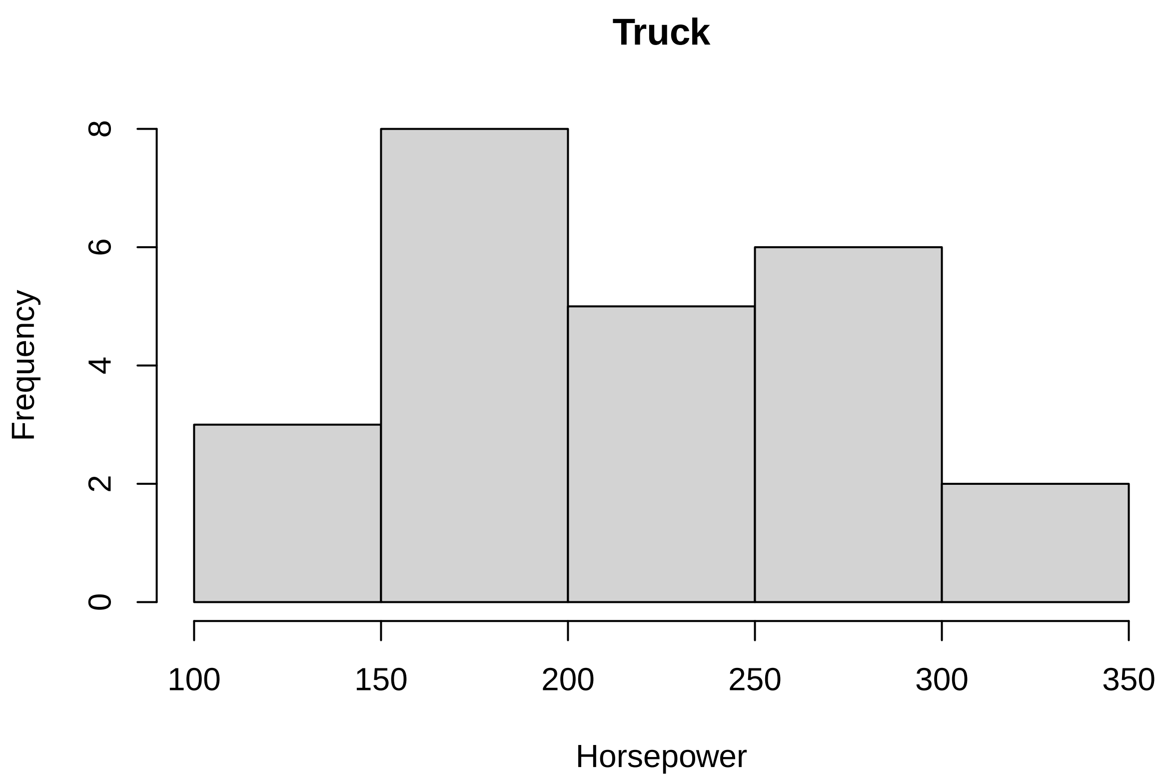
```r
range(SUV$Horsepower, na.rm = TRUE)  #Range
```

```
## [1] 130 340
```

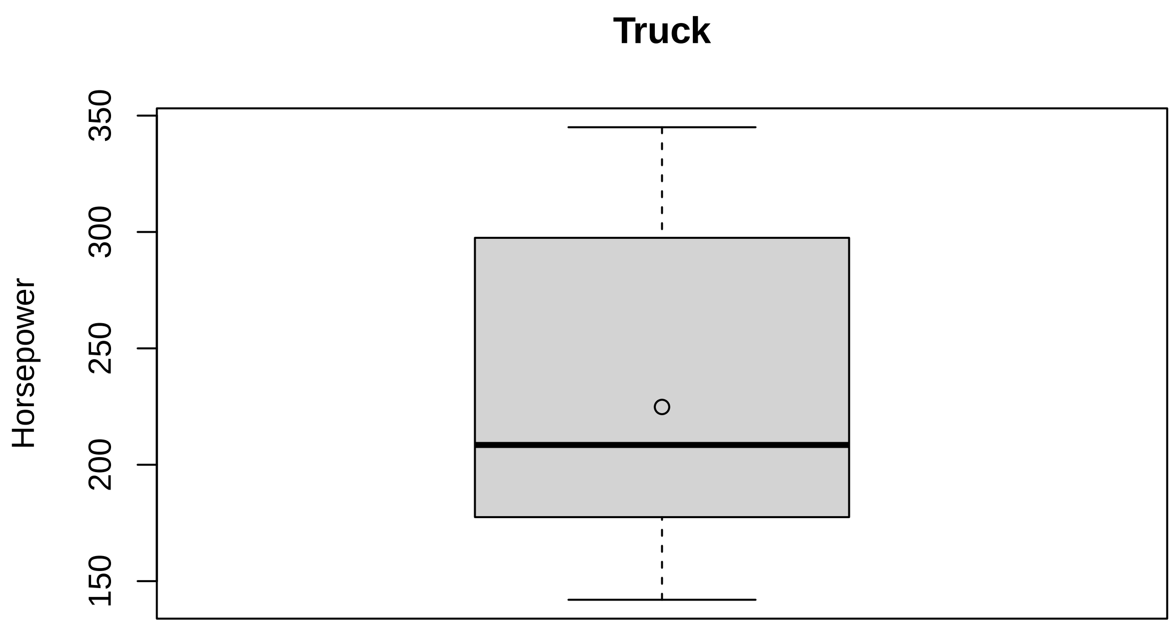
```r
#Qunatitative Test
shapiro.test(SUV$Horsepower)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SUV$Horsepower
## W = 0.95945, p-value = 0.04423
```

```r
#iii) Truck

hist(Truck$Horsepower, main = "Truck", xlab = "Horsepower") #Histogram of Truck
```




```r
boxplot(Truck$Horsepower, main = "Truck", ylab = "Horsepower") # Boxplot for Truck
points(mean(Truck$Horsepower, na.rm = TRUE))
```




```
var(Truck$Horsepower, na.rm = TRUE) #Variance
```

```
## [1] 3824.841
```

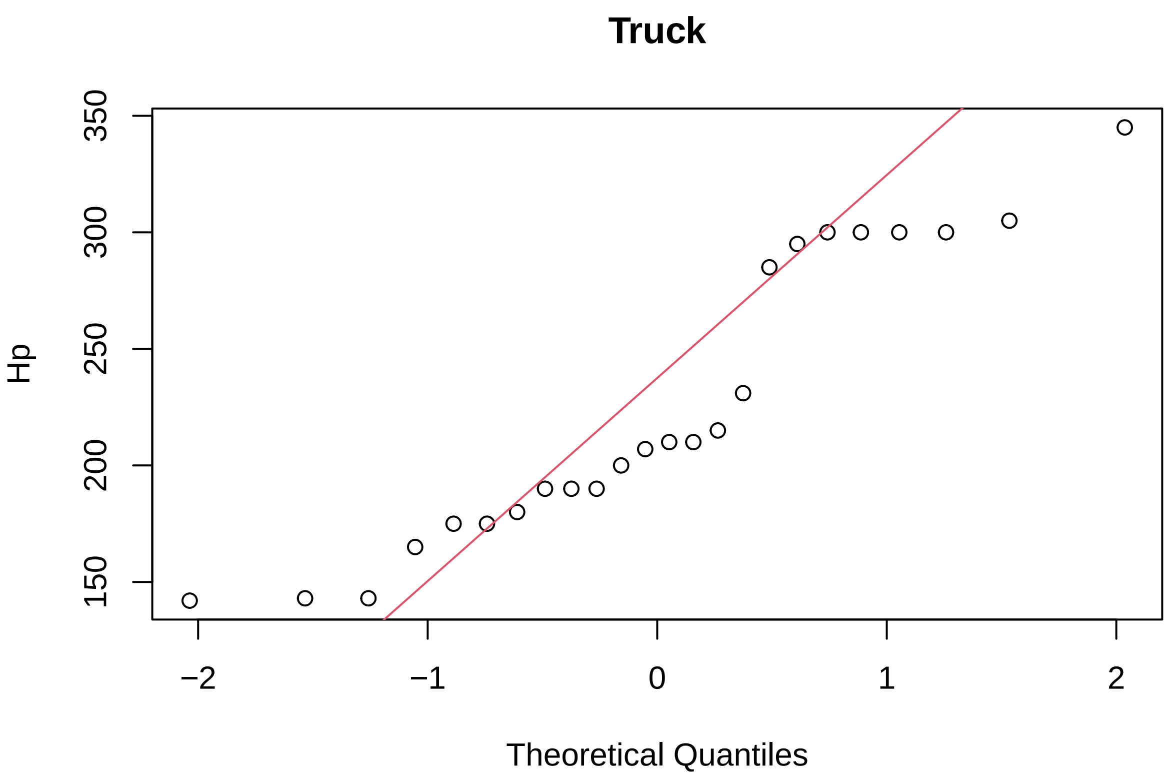
```
skewness(Truck$Horsepower, na.rm = TRUE) #Skewness
```

```
## [1] 0.4020273
```

```
range(Truck$Horsepower, na.rm = TRUE)   #Range
```

```
## [1] 142 345
```

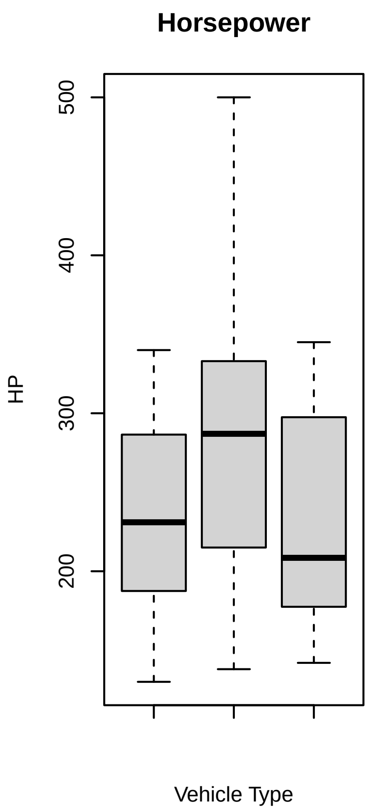
```
qqnorm(Truck$Horsepower, main = "Truck", ylab = "Hp"); qqline(Truck$Horsepower, col = 2) #Normality Plo
```




```
#Qunatitative Test
shapiro.test(Truck$Horsepower)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Truck$Horsepower
## W = 0.8951, p-value = 0.01697
```

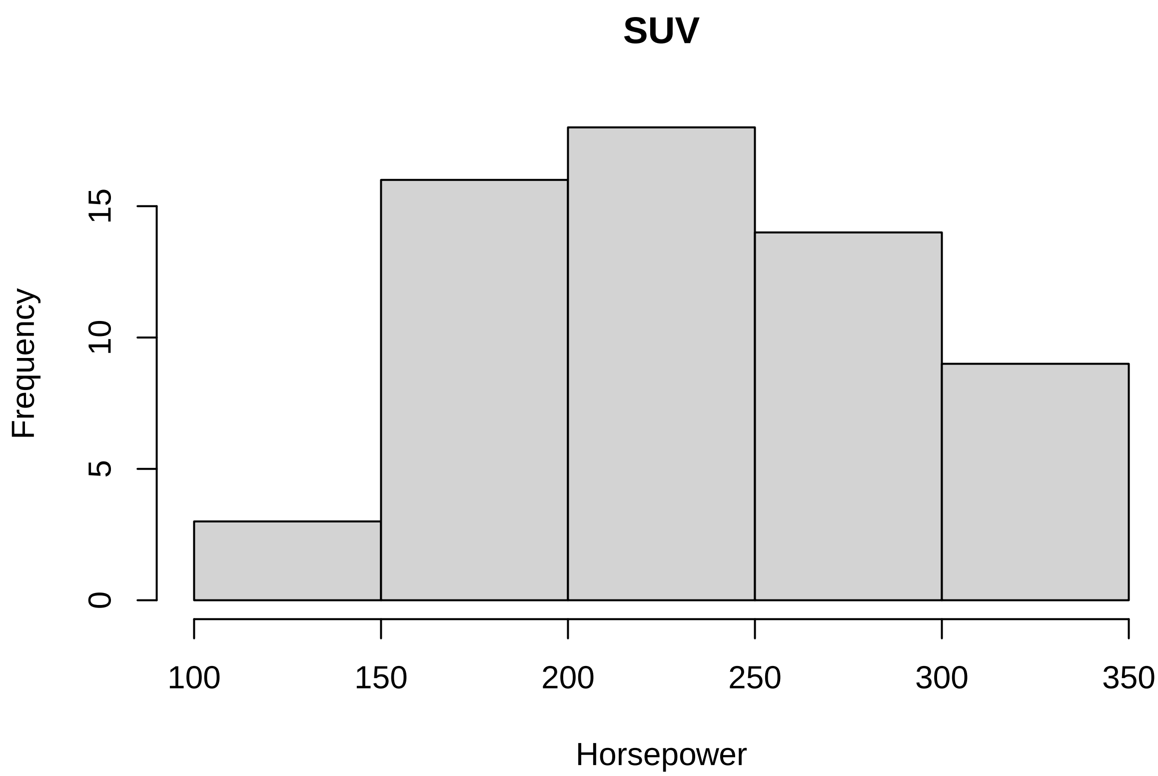
```
####Combine Boxplots####
par(mfrow = c(1,3)) #Combine boxplots
boxplot(SUV$Horsepower, Sports$Horsepower, Truck$Horsepower, main= "Horsepower", ylab = "HP", xlab = "Ve
```



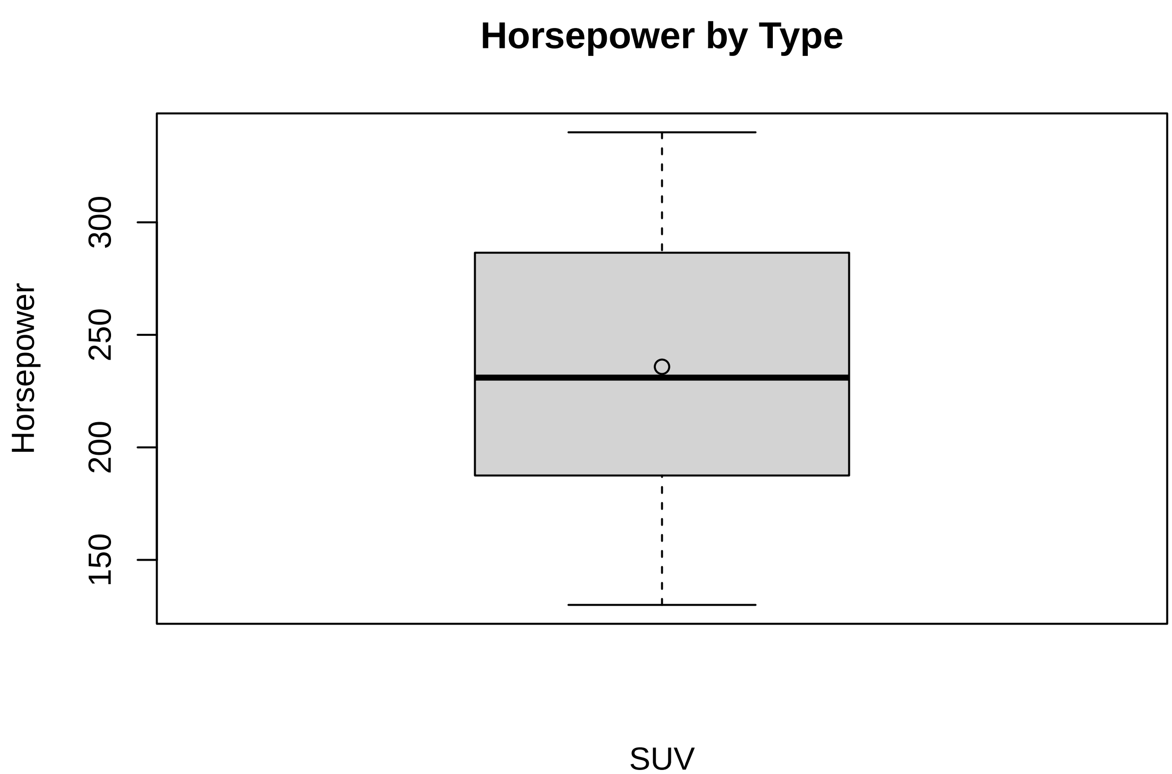
Exercise 2: Hypothesis Testing

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

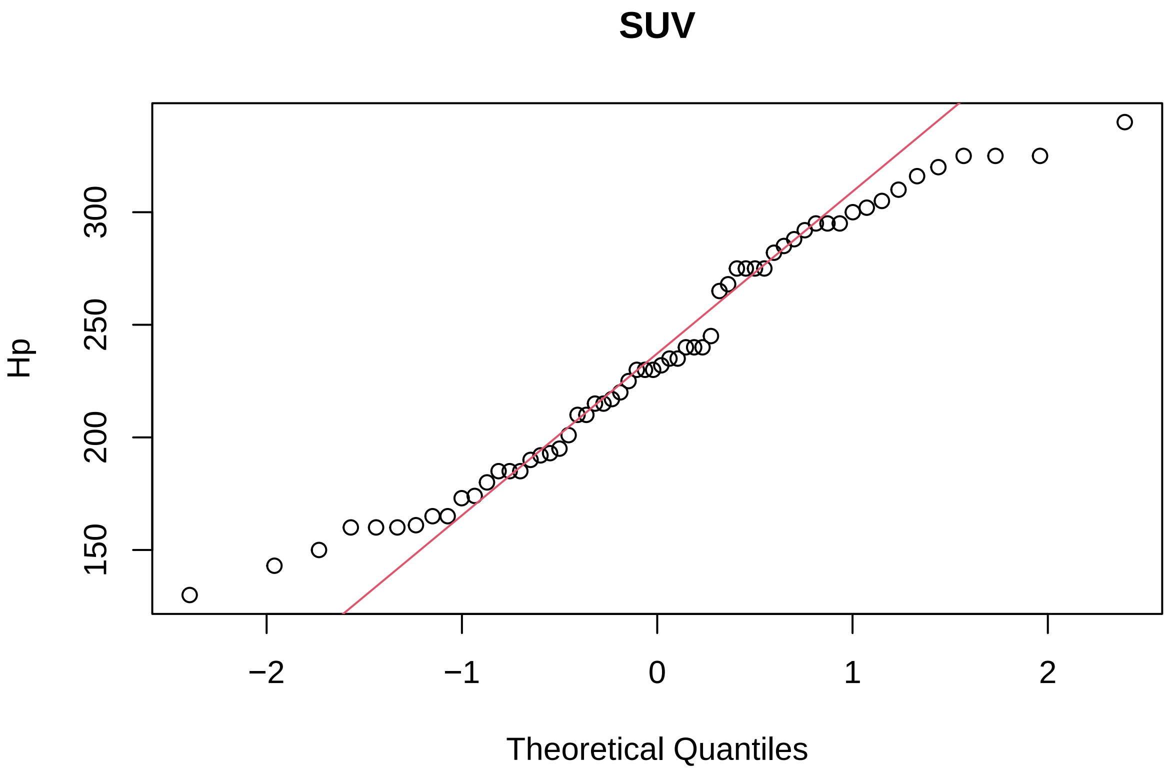
```
#2(a) Which test should we perform, and why? Justify your answer based on findings on Exercise 1 (d).

#Normality of SUV
HPSUV <- cars[cars$Type =="SUV", "Horsepower"]

hist(SUV$Horsepower, main = "SUV", xlab = "Horsepower") #Histogram of SUV
```




```
boxplot(SUV$Horsepower, main = "Horsepower by Type", ylab = "Horsepower", xlab = "SUV") # Boxplot for S
points(mean(SUV$Horsepower, na.rm = TRUE)) #Mean point on boxplot
```



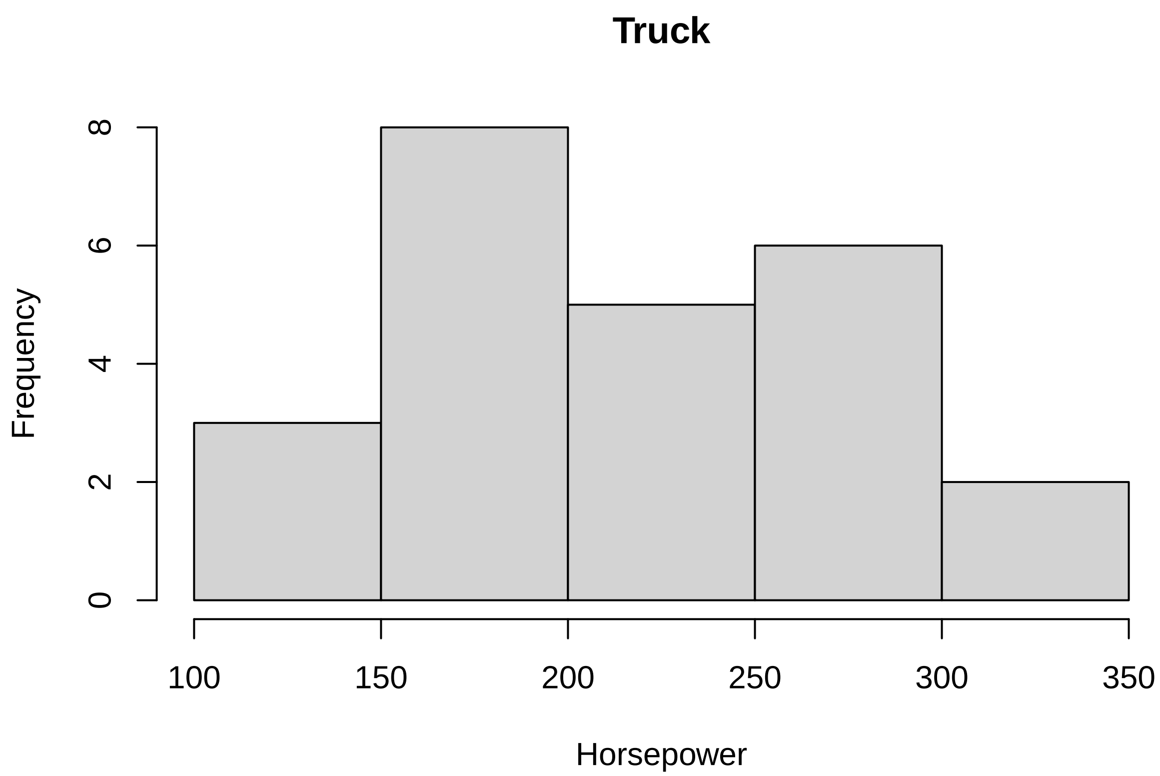
```r
qqnorm(SUV$Horsepower, main = "SUV", ylab = "Hp"); qqline(SUV$Horsepower, col = 2) #Normality Plot
```



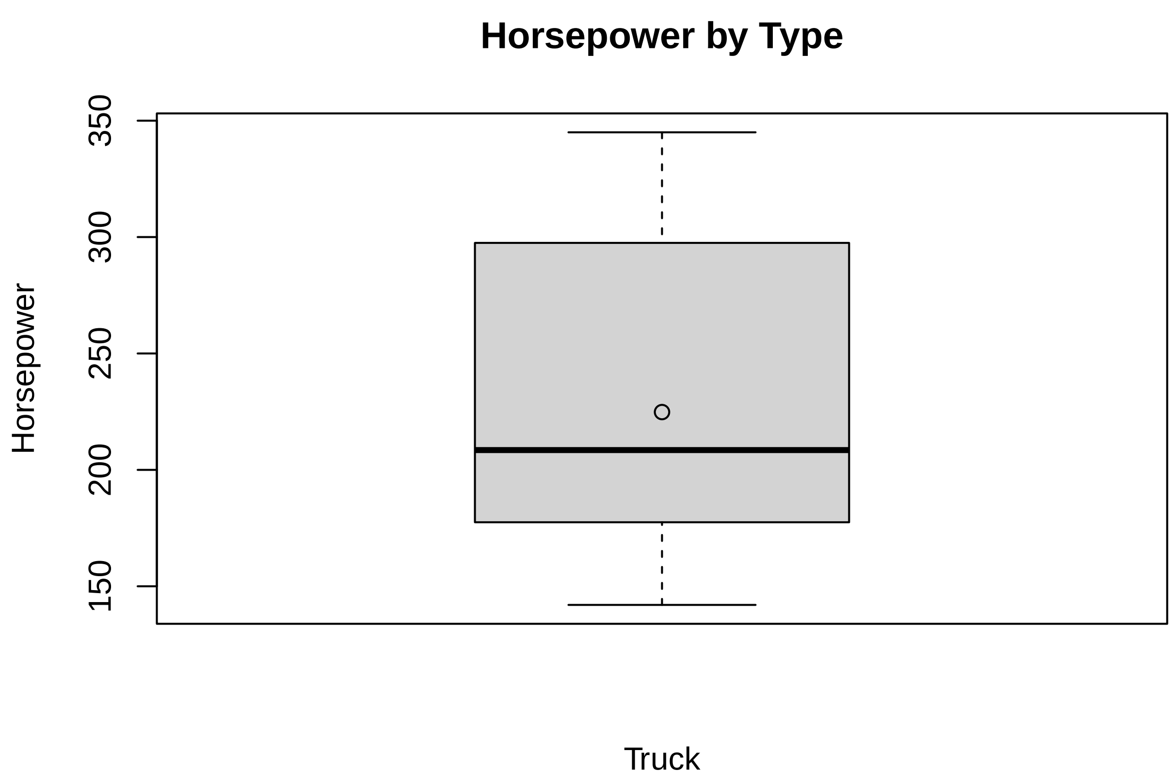
```r
#Qunatitative Test
shapiro.test(SUV$Horsepower) #PValue is less than 5%, Reject Null We do not have evidence that Data is
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SUV$Horsepower
## W = 0.95945, p-value = 0.04423
```

```r
#Normality of Truck
HPTruck <- cars[cars$Type =="Truck", "Horsepower"]

hist(Truck$Horsepower, main = "Truck", xlab = "Horsepower") #Histogram of Truck
```




```
boxplot(Truck$Horsepower, main = "Horsepower by Type", ylab = "Horsepower", xlab = "Truck") # Boxplot f
points(mean(Truck$Horsepower, na.rm = TRUE))
```



```r
var(Truck$Horsepower, na.rm = TRUE) #Variance
```

```
## [1] 3824.841
```

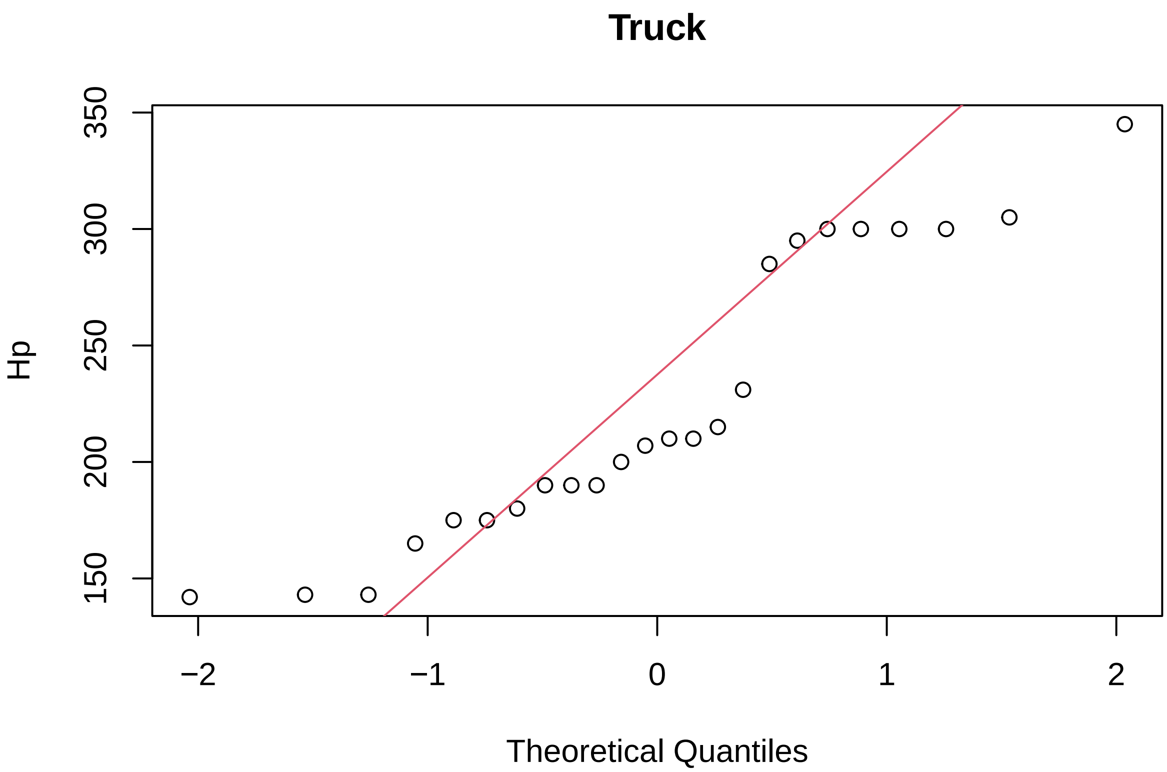
```r
skewness(Truck$Horsepower, na.rm = TRUE) #Skewness
```

```
## [1] 0.4020273
```

```r
range(Truck$Horsepower, na.rm = TRUE)  #Range
```

```
## [1] 142 345
```

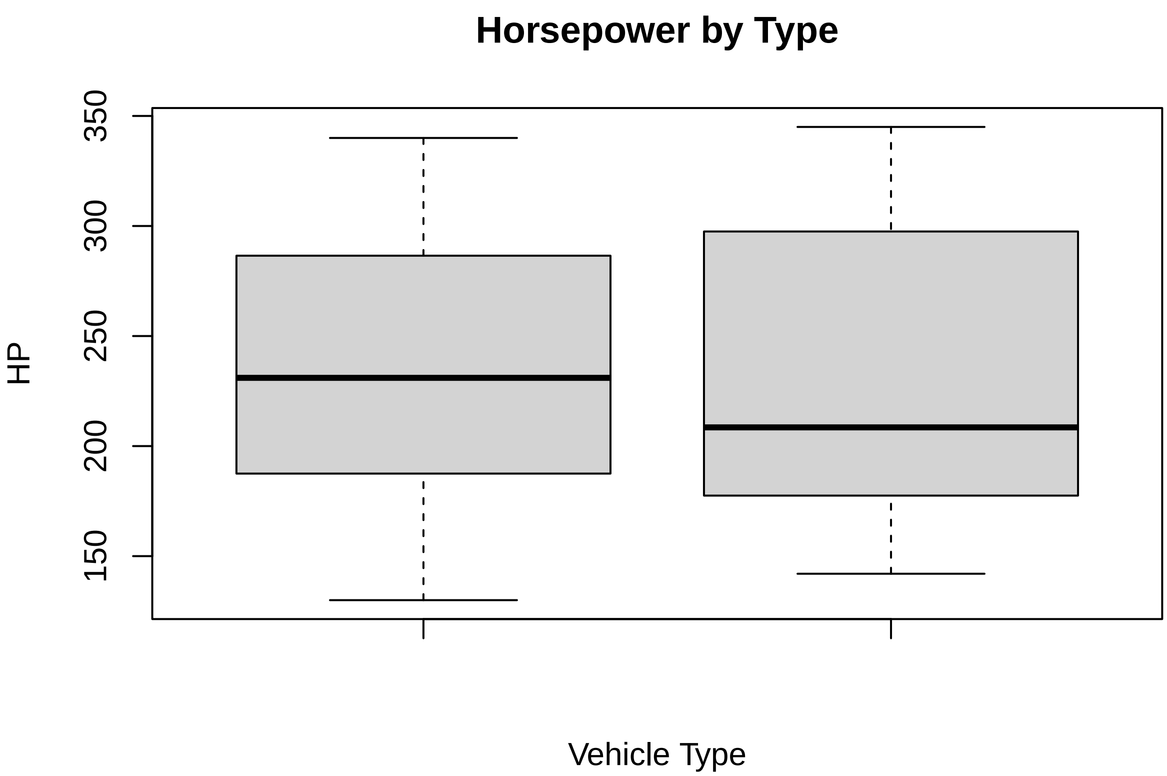
```r
qqnorm(Truck$Horsepower, main = "Truck", ylab = "Hp"); qqline(Truck$Horsepower, col = 2) #Normality Plo
```




```r
#Qunatitative Test
shapiro.test(Truck$Horsepower) #PValue is less than 5%, Reject Null We do not have evidence that Data i
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Truck$Horsepower
## W = 0.8951, p-value = 0.01697
```

```r
####Combine Boxplots####
boxplot(SUV$Horsepower, Truck$Horsepower, main= "Horsepower by Type", ylab = "HP", xlab = "Vehicle Type
```




```r
#Data for SUV AND Truck is not normaly distributed, non-parametric test

#Run Wilcoxon Rank-Sum Test
```

```r
#2 (b) Specify null and alternative hypotheses.

#Ho: Median of Horsepower for SUVs is equal to the median of Horsepower for Trucks
#Ho: MSUV = MTruck
#Ha: Median of Horsepower for SUVs is not equal to the median of Horsepower for Trucks
#Ha: MSUV != MTruck
```

```r
#2 (c) State the conclusion based on the test result.

wilcox.test(HPSUV, HPTruck, exact = FALSE)
```

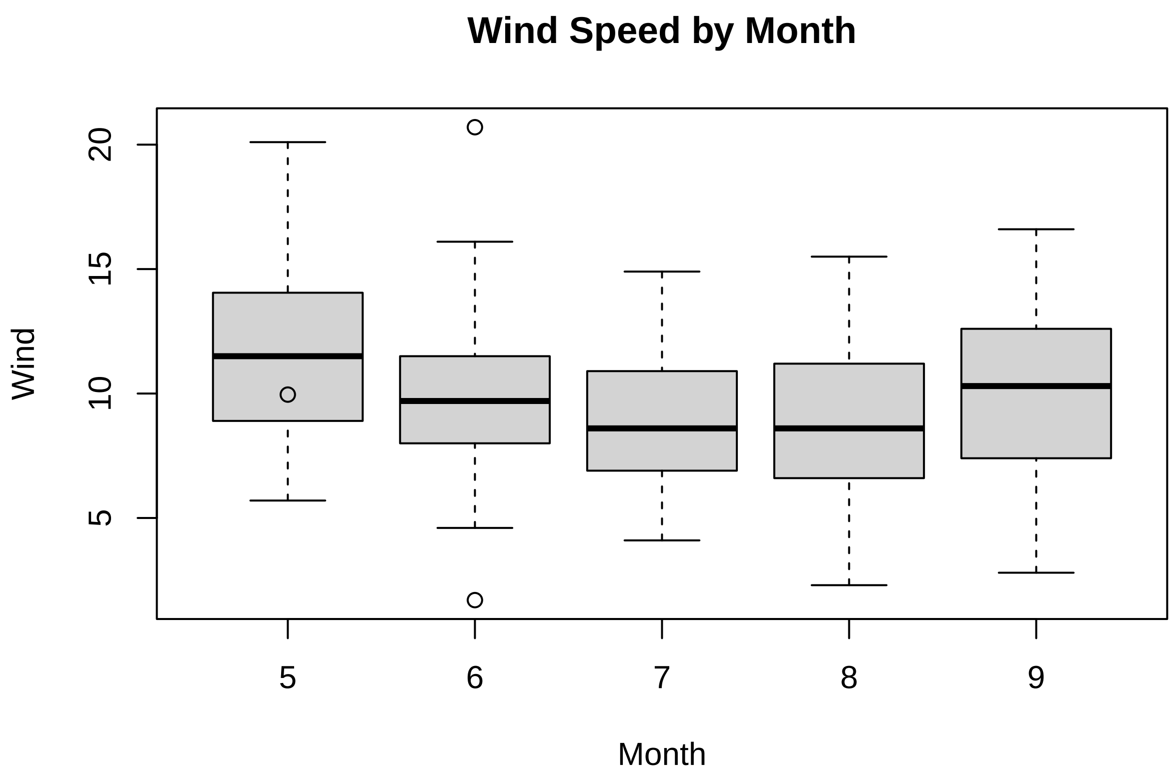
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  HPSUV and HPTruck
## W = 806.5, p-value = 0.3942
## alternative hypothesis: true location shift is not equal to 0
```

```r
#PValue is greater than 5%, we do not have enough evidence to reject the null.
#We can conclude that median horsepower for SUVs is equal to the median horsepower for Trucks
```

###Exercise 3: Hypothesis Testing

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

```
#3 (a) Which test should we perform, and why? See QQ-plot and perform Shapiro-Wilk test for normality c

boxplot(Wind ~ Month, data = airquality, main = "Wind Speed by Month") #Boxplot of Wind
points(mean(airquality$Wind, na.rm = TRUE))
```




```
#Wind Speed for July
month7.wind <- airquality[airquality$Month ==7, "Wind"]
var(month7.wind, na.rm = TRUE) #Variance
```

```
## [1] 9.217183
```

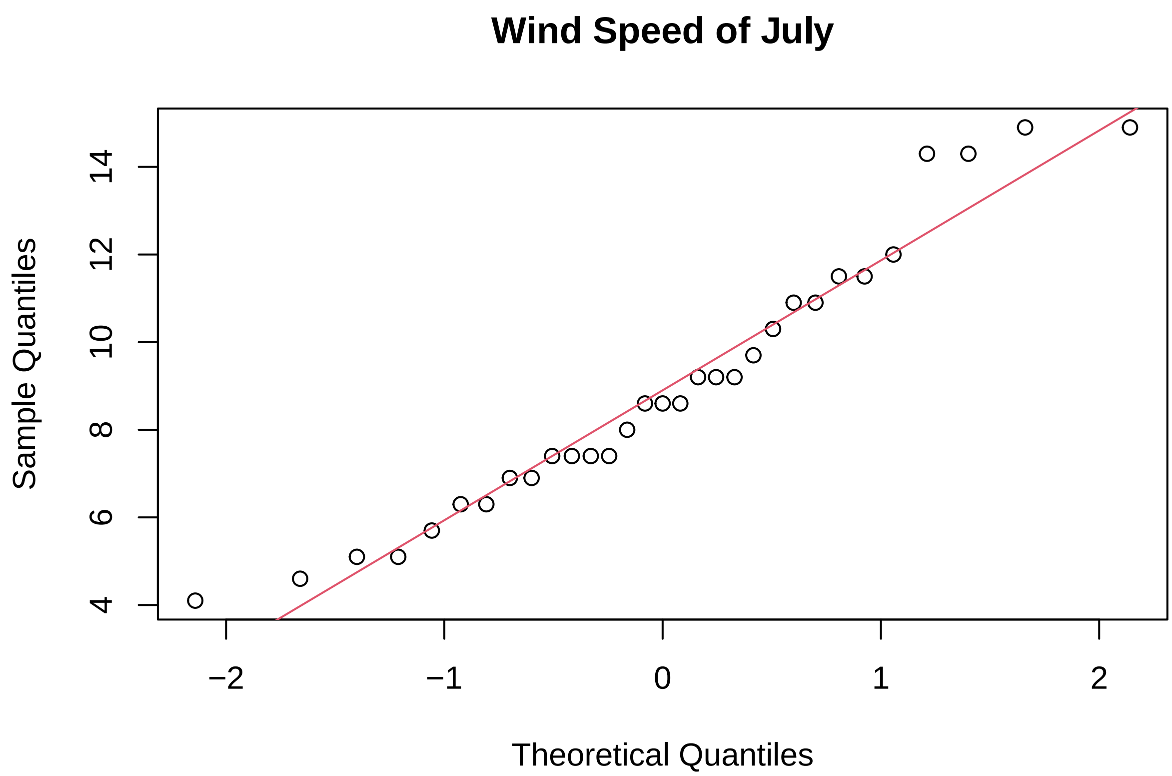
```
skewness(month7.wind, na.rm = TRUE) #Skewness
```

```
## [1] 0.482349
```

```
range(month7.wind, na.rm = TRUE)  #Range
```

```
## [1]  4.1 14.9
```

```
qqnorm(airquality$Wind[airquality$Month == "7"], main = "Wind Speed of July"); qqline(airquality$Wind[a
```




```
#Quantitative Test
shapiro.test(airquality[airquality$Month ==7, "Wind"]) #PValue is greater than 5%, We do not have evide
```

```
##
##  Shapiro-Wilk normality test
##
## data:  airquality[airquality$Month == 7, "Wind"]
## W = 0.95003, p-value = 0.1564
```

```
#Wind Speed for August
month8.wind <- airquality[airquality$Month ==8, "Wind"]
var(month8.wind, na.rm = TRUE) #Variance
```

```
## [1] 10.40662
```

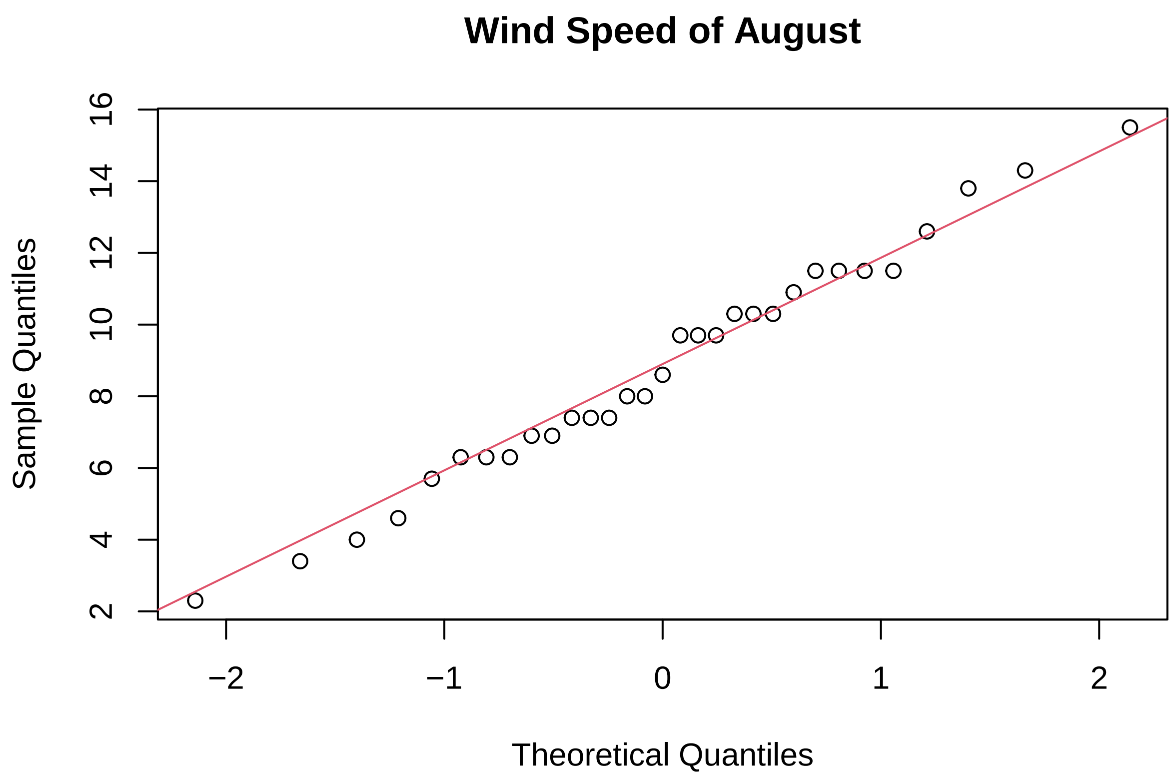
```
skewness(month8.wind, na.rm = TRUE) #Skewness
```

```
## [1] 0.03728531
```

```
range(month8.wind, na.rm = TRUE)  #Range
```

```
## [1]  2.3 15.5
```

```
qqnorm(airquality$Wind[airquality$Month == "8"], main = "Wind Speed of August"); qqline(airquality$Wind
```




```
#Quantitative Test
shapiro.test(airquality[airquality$Month ==8, "Wind"]) #PValue is greater than 5%, We do not have evide
```

```
##
##  Shapiro-Wilk normality test
##
## data:  airquality[airquality$Month == 8, "Wind"]
## W = 0.98533, p-value = 0.937
```

```
#Variance Test by Month
var.test(month7.wind, month8.wind, data = airquality)
```

```
##
##  F test to compare two variances
##
## data:  month7.wind and month8.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
```

```
#PValue is .7418 which is greater than 5% so we do not have enough evidence to reject the Null.  The tw
```

```r
#3 (b) Specify null and alternative hypotheses

#Ho: Mean speed of wind in July is equal to the mean speed of wind in August
#Ho: M7 = M8
#Ha: Mean speed of wind in July is not equal to the mean speed of wind in August
#Ha: M7 != M8
```

```r
#3 (c) State the conclusion based on the test result.


t.test(month7.wind, month8.wind, alternative = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  month7.wind and month8.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
```

```r
#PValue is .8527 which is greater than 5%.  We do not have enough evidence to reject the NUll and may c
```