Understanding Statistical Techniques

Temitope Oduwole

Importing all Libraries

```
library(ggplot2)#To explore the dataset
library(ggpubr)#To explore and visualize dataset
library(corrplot)#for visualization of correlation
library(dplyr)#for data manipulation
library(tidyverse)
library(rmarkdown)#to knit to word or pdf
library(e1071)
```

#** Importing the first Dataset**

```
#To read dataset
Exams_data <- read.csv('exams.csv')
#To view first 6 rows
head(Exams_data)

##   gender race.ethnicity parental.level.of.education        lunch
## 1   male        group A                 high school     standard
## 2 female        group D            some high school free/reduced
## 3   male        group E                some college free/reduced
## 4   male        group B                 high school     standard
## 5   male        group E          associate's degree     standard
## 6 female        group D                 high school     standard
##   test.preparation.course math.score reading.score writing.score
## 1               completed         67            67            63
## 2                    none         40            59            55
## 3                    none         59            60            50
## 4                    none         77            78            68
## 5               completed         78            73            68
## 6                    none         63            77            76

#To print dataset size
sprintf("Dataset size: [%s]",toString(dim(Exams_data)))

## [1] "Dataset size: [1000, 8]"
```

The Exams dataset contains 8 columns and 1000 rows.

#**Dataset Description**

This is a Students Performance in Exams dataset downloaded from kaggle. It contains the below fields:

Gender - This is the student's gender, either male or female

race.ethnicity- This is the student's race/ethnicity, and it is grouped into A - D.

parent.level.of.education- This indicates how educated the student's parents are. It is divided into "high school", "college','degree' etc

lunch- This represents the type of daily lunch the student has subscribed for in school. It is divided into "standard', 'free/reduced"

math.score- This is an integer indicating the score of each student in Maths Assessment out of 100

reading.score-This is an integer indicating the score of each student in Reading Assessment out of 100

writing.score-This is an integer indicating the score of each student in Writing Assessment out of 100

**All variables are independent.

To check for null variables

```
is.null(Exams_data)

## [1] FALSE
```

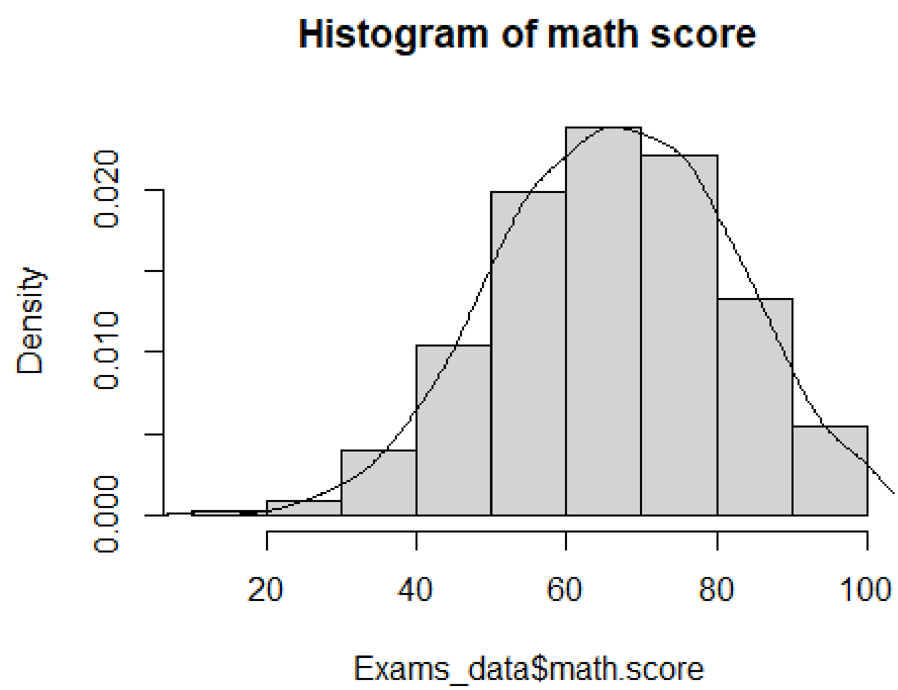
The dataset contains no missing values or null variables.

Test for Normality on the math.score variable

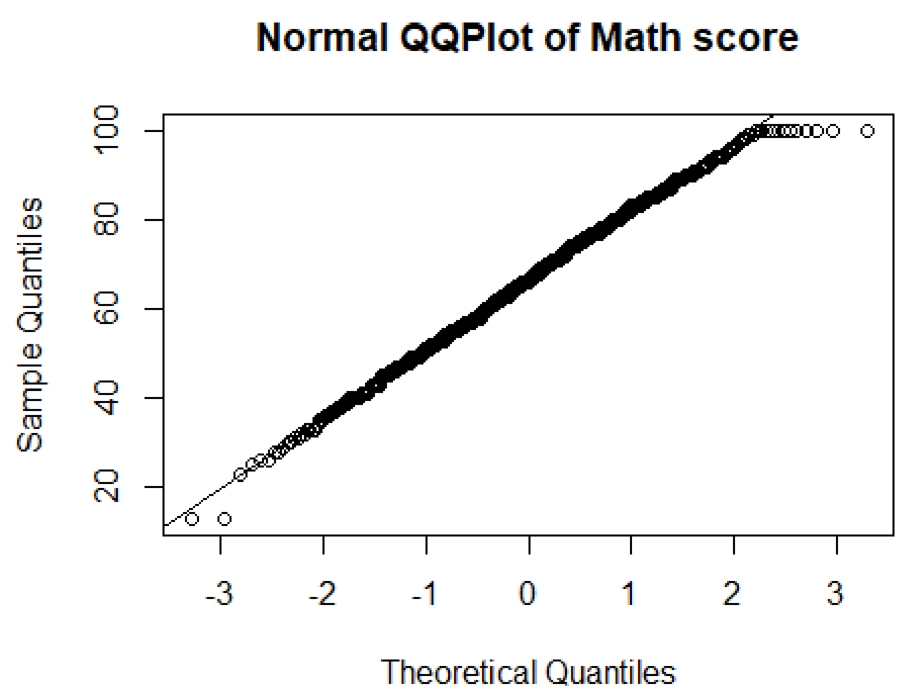
The aim here is to check whether the math score of the students is normally distributed or not. This will determine whether the statistical techniques to apply on the dataset i.e whether parametric or non-parametric.

Histogram and Lineplot of math.score variable

```
hist(Exams_data$math.score,main ="Histogram of math score", prob=TRUE)
lines(density(Exams_data$math.score))
```



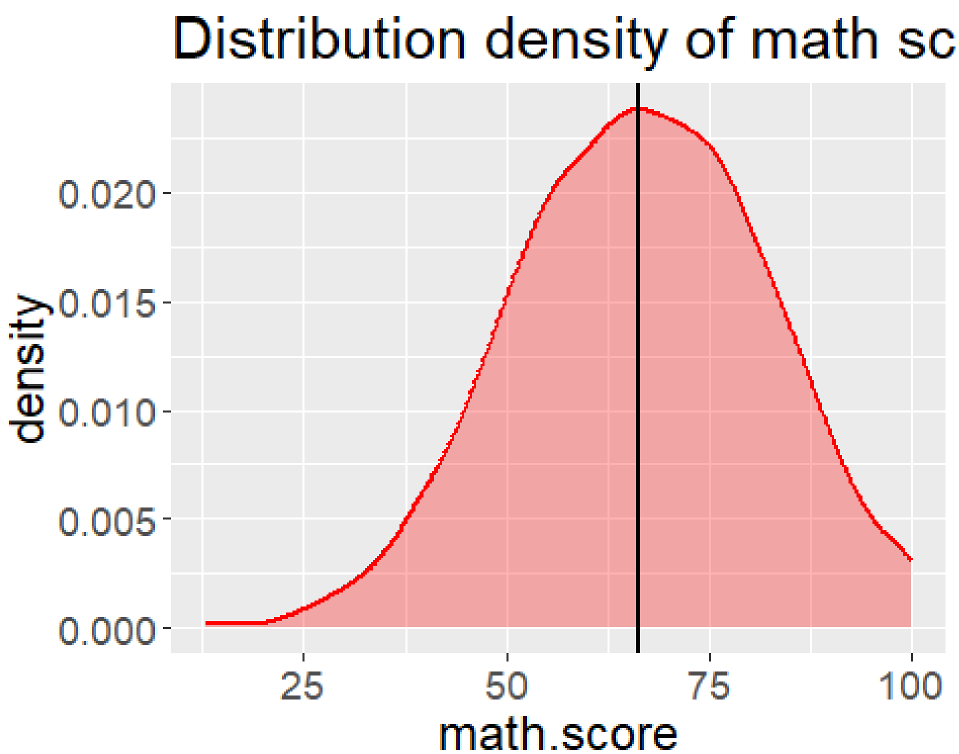

```
qqnorm(Exams_data$math.score, main="Normal QQPlot of Math score")
qqline(Exams_data$math.score)
```



As observed, the density curve on the histogram is bell shaped which is characteristic of a normal distribution, and the with the QQplot, the points mostly fall on the line. Even though the distribution looks normal, it would be further tested for skewness, and Shapiro Wilk's Test.

#Test for skewness

```
#set_plot_dimensions(10,8)
ggplot(Exams_data, aes(x=math.score)) +
    geom_density(alpha=.3, fill="red", color="red", size=1)+
    geom_vline(aes(xintercept=mean(math.score)), size=1, color ="black")+
    ggtitle("Distribution density of math score") +
    theme(text = element_text(size = 18))

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```




```
sprintf("Skewness of math score: [%s]",
toString(skewness(Exams_data$math.score)))

## [1] "Skewness of math score: [-0.150694341991838]"
```

The skewness of the math.score is (-0.15). This means the variable is slightly skewed to the left. A skewness of 0 indicates a perfectly symmetrical data A shapiro Wilk's test will be done to conclude on Normality of the variable.

Shapiro Wilk's Test for Normality

H0: The math.score is normally distributed
H1: The math.score is not normally distributed

If the p-value is lower than 0.05, the null hypothesis will be rejected, but if more than 0.05, I will fail to reject the null hypothesis

```
shapiro.test(Exams_data$math.score)

##
##  Shapiro-Wilk normality test
##
## data:  Exams_data$math.score
## W = 0.99508, p-value = 0.002512
```

A p-value of 0.000122,0.000119 which are both lesser than 0.05 is observed. This is sufficient evidence to reject the null hypothesis that math.score and reading.score variables is normally distributed

#**Conclusion on Test for Normality**

The math.score does not follow a normal distribution, hence, non-parametric tests would be used for further hypothesis testing

HYPOTHESIS TESTING

Mann-Whitney U test:

The non-parametric alternative to the independent t-test, the Mann-Whitney U test, is used in Comparing the difference in observation of the medians of males and females math.score
H0: The average math score of the two groups are equal
H1: The average math score of the two groups are not equal

This test will be carried out at *95% confidence interval*. This means that if the p-value is below 0.05, the null hypothesis will be rejected and if it is above 0.05, then we have evidence not to reject the null hypothesis (this means we will fail to reject it).

```
Male_score <- Exams_data$math.score[Exams_data$gender == 'male']
Female_score <- Exams_data$math.score[Exams_data$gender == 'female']


wilcox.test(Male_score,Female_score,alternative = "greater", conf.int =TRUE)

##
##  Wilcoxon rank sum test with continuity correction
##
```

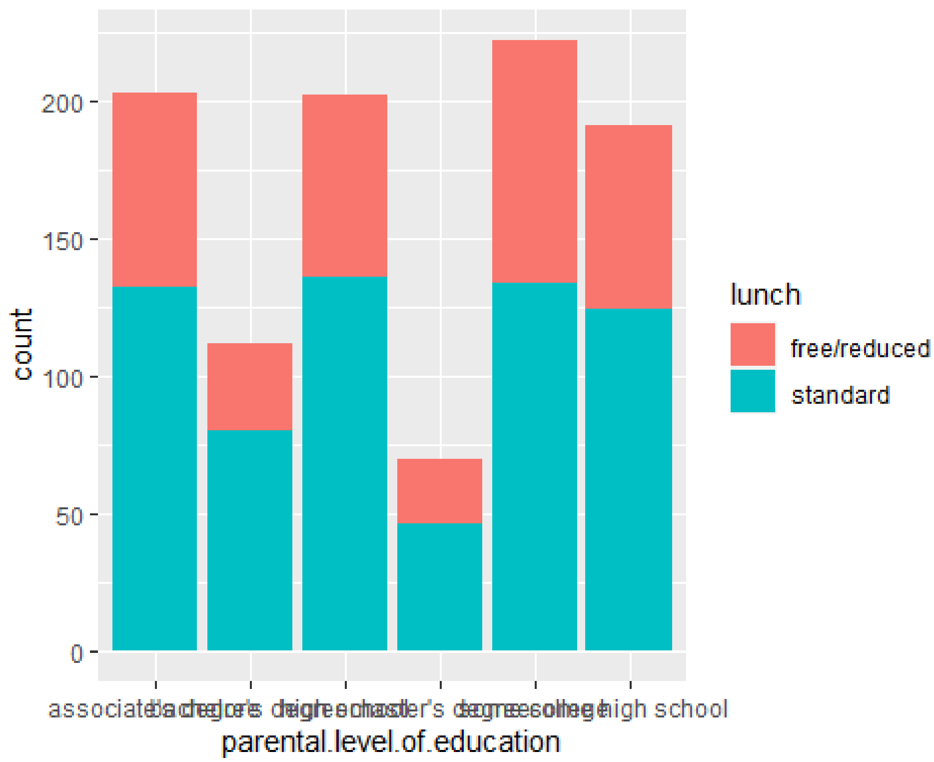
```
## data:  Male_score and Female_score
## W = 152698, p-value = 5.253e-10
## alternative hypothesis: true location shift is greater than 0
## 95 percent confidence interval:
##  4.999984      Inf
## sample estimates:
## difference in location
##               6.000021
```

With a p-value of 0.000000005.25 which is lesser than the significant level of 0.05, there is enough evidence to reject the null hypothesis which is that the average scores of Males and Females are similar

Pearson's Chi-Squared Test

H0: There is no relationship between how educated the parent is and the type of lunch the child gets **H1:** There is a relationship between how educated the parent is and the type of lunch the child gets

```
#Create a contingency of the ParentalLevelOfEducation and lunch

ParentalEducation <- table(Exams_data$parental.level.of.education,
Exams_data$lunch)
ParentalEducation

##
##                         free/reduced standard
##   associate's degree             71      132
##   bachelor's degree              32       80
##   high school                    66      136
##   master's degree                24       46
##   some college                   88      134
##   some high school               67      124

#Chi-squared test
chisq.test(Exams_data$lunch, Exams_data$parental.level.of.education, correct
=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  Exams_data$lunch and Exams_data$parental.level.of.education
## X-squared = 4.6268, df = 5, p-value = 0.4631

#Barplot of ParentalLevelOfEducation and lunch

ggplot(Exams_data) +
  aes(x=parental.level.of.education, fill=lunch) +
  geom_bar()
```



A p-value of 0.46 which is greater than the significant level of 0.05 means we fail to reject the null hypothesis. This means there is no correlation between ParentalLevelOfEducation and lunch.

#** Importing the second Dataset**

```
#To read the file, view first few rows and get summary of Dataset
Cars_data = read.csv("cars.csv")
head(Cars_data)
```

```
##    mpg cylinders cubicinches  hp weightlbs time.to.60 year    brand
## 1 14.0         8         350 165      4209         12 1972      US.
## 2 31.9         4          89  71      1925         14 1980  Europe.
## 3 17.0         8         302 140      3449         11 1971      US.
## 4 15.0         8         400 150      3761         10 1971      US.
## 5 30.5         4          98  63      2051         17 1978      US.
## 6 23.0         8         350 125      3900         17 1980      US.
```

```
sprintf("Dataset size: [%s]",toString(dim(Cars_data)))
```

```
## [1] "Dataset size: [256, 8]"
```

```
summary(Cars_data)
```

```
##       mpg          cylinders     cubicinches          hp
weightlbs
##  Min.   :10.00   Min.   :3.00   Min.   : 70.0   Min.   : 46.0   Min.
:1613
```

```
##  1st Qu.:16.80   1st Qu.:4.00   1st Qu.:100.2   1st Qu.: 75.0   1st
Qu.:2246
##  Median :22.00   Median :5.00   Median :156.0   Median : 95.0   Median
:2832
##  Mean   :23.19   Mean   :5.59   Mean   :201.4   Mean   :106.8   Mean
:3006
##  3rd Qu.:28.85   3rd Qu.:8.00   3rd Qu.:304.0   3rd Qu.:139.0   3rd
Qu.:3666
##  Max.   :46.60   Max.   :8.00   Max.   :455.0   Max.   :230.0   Max.
:4997
##    time.to.60        year         brand
##  Min.   : 8.0   Min.   :1971   Length:256
##  1st Qu.:14.0   1st Qu.:1974   Class :character
##  Median :16.0   Median :1977   Mode  :character
##  Mean   :15.5   Mean   :1977
##  3rd Qu.:17.0   3rd Qu.:1980
##  Max.   :25.0   Max.   :1983
```

The Cars Data dataset contains 8 columns, each with 256 rows.

#**Dataset Description**

The Cars Data has information about 3 brands or make of cars. Namely; US, Japan, Europe kaggle. It contains the below fields:

mpg - This represents miles per galon which is a measure of fuel economy in cars.

cylinders- This is the number of cylinders in the car engine. It ranges from 4-8

cubicinches- This is a measure of engine displacement in cubic inches

hp- This is the car's horsepower. This is used to measure the power produced by the engine

weightlbs- This the weight of the car including a full tank and all standard equipment

time.to.60-This is the time it takes a car to accelerate from 0-60mph

year-This is the year of production of the car

brand- This is the country of production of the car

**All variables are independent.

To check for null variables

```
#Checking for Null values
is.null(Cars_data)

## [1] FALSE
```

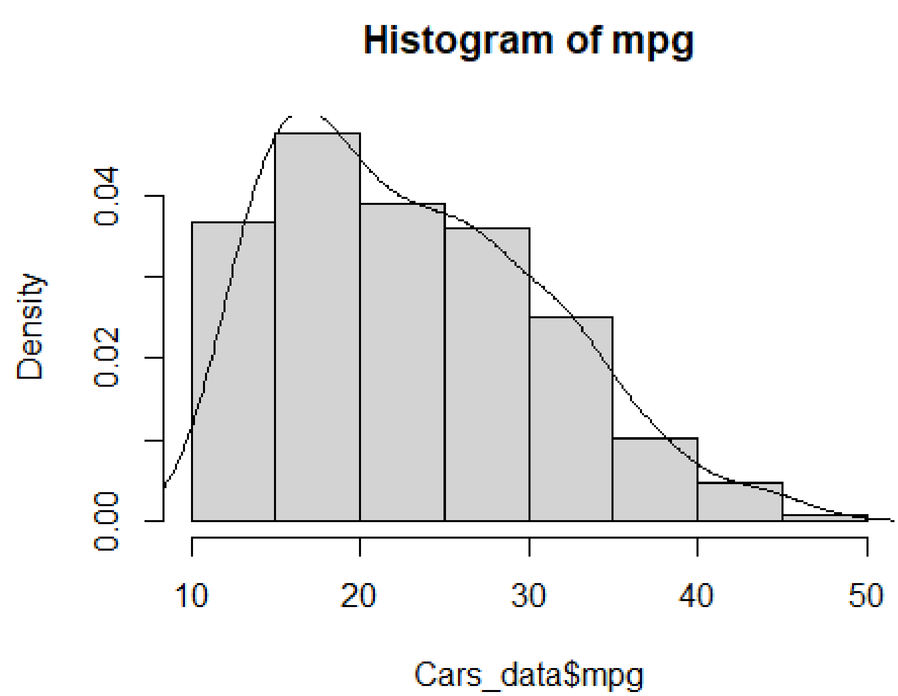
```
sum(is.na(Cars_data))

## [1] 0
```

Testing for Normality on the mpg variable

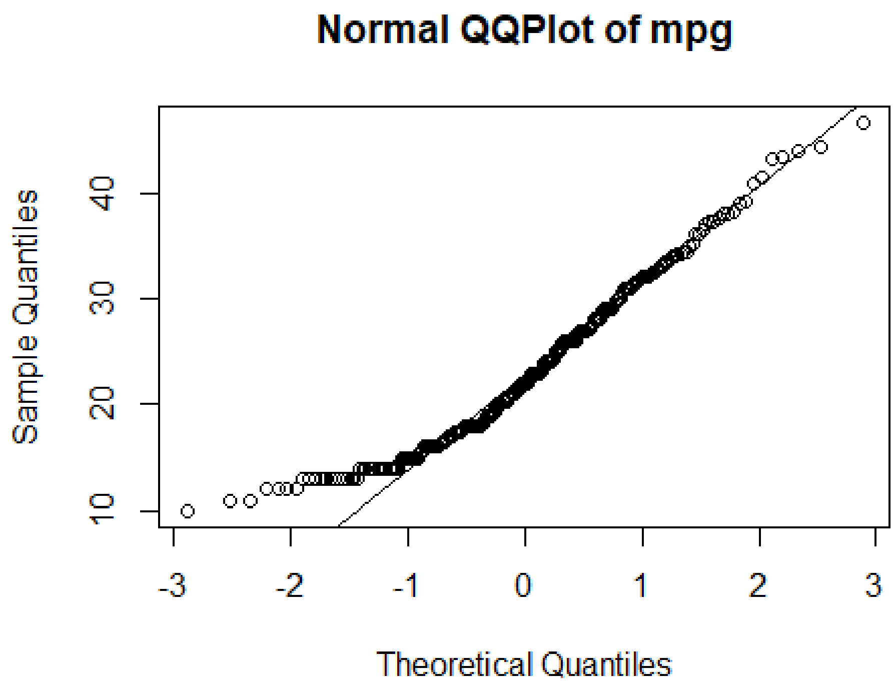
H0: The mpg is normally distributed
H1: The mpg is not normally distributed

If the p-value from the Shapiro-Wilk's test is lower than 0.05, the null hypothesis will be rejected, but if more than 0.05, I will fail to reject the null hypothesis

```
#Histogram of mpg
hist(Cars_data$mpg,main ="Histogram of mpg", prob=TRUE)
lines(density(Cars_data$mpg))
```




```
#Normal QQplot for mpg
qqnorm(Cars_data$mpg, main="Normal QQPlot of mpg")
qqline(Cars_data$mpg)
```




```
#Shapiro-wilk's test for mpg
shapiro.test(Cars_data$mpg)

##
##  Shapiro-Wilk normality test
##
## data:  Cars_data$mpg
## W = 0.9552, p-value = 4.15e-07
```

The histogram is skewed to the left, and the line plot is far from the origin with many points outside the line. This is characteristic of distributions that are not normal. The observed p-value is far lesser than 0.05 . This is sufficient evidence to reject the null hypothesis that mpg variable is normally distributed.

Kruskal-Wallis Test (ANOVA)

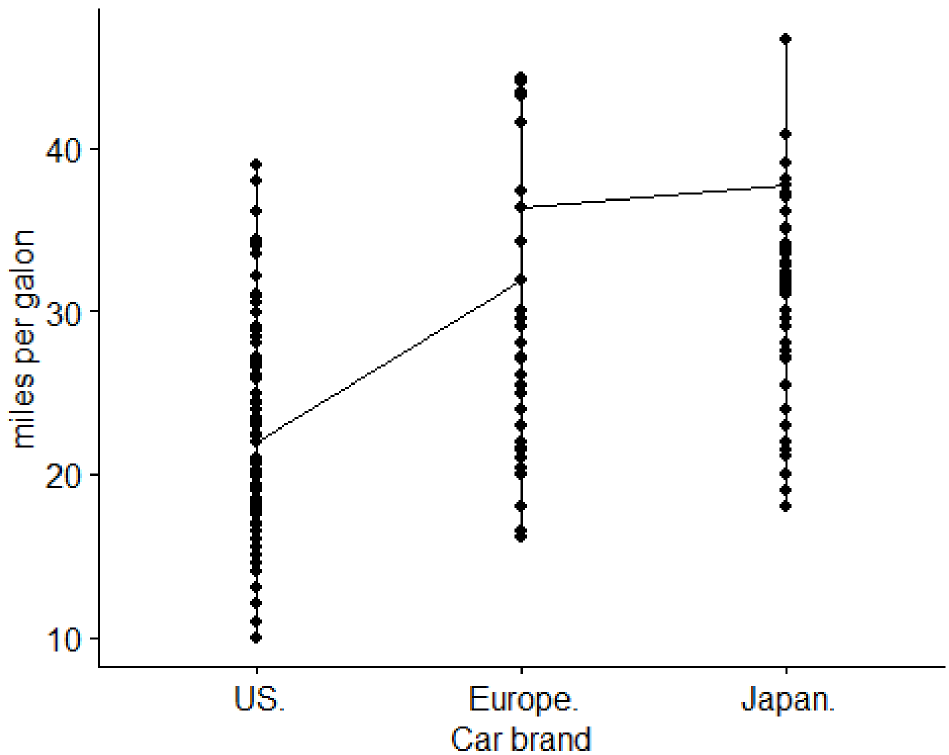
The Kruskal-Wallis test is the non-parametric alternative to compare analysis of variance. This test is to compare average rank mpg based on the brand of cars.

H0: The 3 brands of cars have similar mpg

H1: Atleast one of the brands has a different mpg

This test will be carried out at *95% confidence interval*. This means that if the p-value is below 0.05, the null hypothesis will be rejected and if it is above 0.05, then we have evidence not to reject the null hypothesis (this means we will fail to reject it).

```
kruskal.test(mpg ~ brand, data = Cars_data)

##
##  Kruskal-Wallis rank sum test
##
## data:  mpg by brand
## Kruskal-Wallis chi-squared = 90.706, df = 2, p-value < 2.2e-16

#Plot to compare average miles per gallon by car brand
ggline(Cars_data, x = "brand", y = "mpg",ylab= "miles per galon", xlab="Car
brand")
```



There is no similarity in the central tendency in mpg of the cars based on the different brands, as the p-value is less than the significance level of 0.05. Thereby the null hypothesis is rejected. This is also further confirmed by the ggplot

Multivariate Correlation test

Checking for correlation between mpg and other variable factors (Cylinders, cubicinches, hp, weightlbs, year)

```
Carscorr <-Cars_data %>% select(mpg, cylinders,cubicinches,hp,weightlbs,year)
library(corrplot)
cor_data <- cor(Carscorr)
corrplot(cor_data, method = "number")
```

The closer to 1 or -1, the stronger the correlation, but the closer to zero, the weaker the correlation. The mpg shows strong negative correlation with cylinders, cubicinches, hp, weightlbs. This means that as number of cylinders, horsepower, wieight, and engine size increases, the gas mileage goes down. However, there is a lesser positive correlation with between mpg and year which means more fuel efficient cars have been manufactured over the years.

Linear Modelling

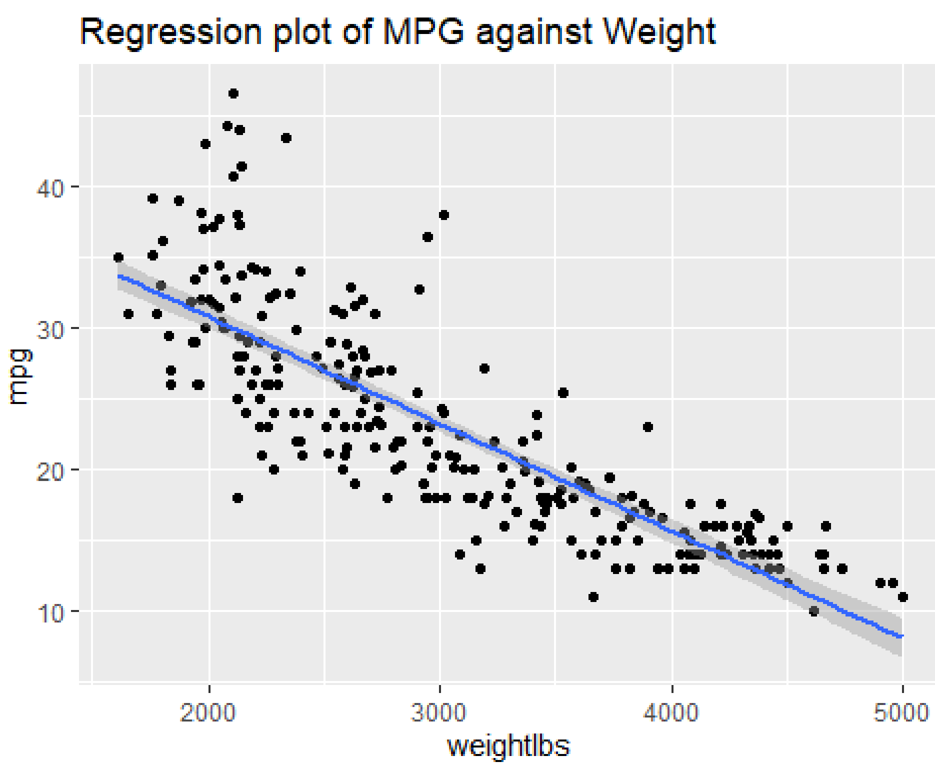
Weightlbs has the highest correlation with mpg. A linear model can be constructed for the two variables

```
#Simple Linear Regression
lm1 <-lm(mpg ~ weightlbs,data = Cars_data)
summary(lm1)

##
## Call:
## lm(formula = mpg ~ weightlbs, data = Cars_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8838  -2.8311  -0.3101   2.2717  16.6099
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.0024967  1.0196355   45.12   <2e-16 ***
```

```
## weightlbs   -0.0075888  0.0003262  -23.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.457 on 254 degrees of freedom
## Multiple R-squared:  0.6805, Adjusted R-squared:  0.6793
## F-statistic: 541.1 on 1 and 254 DF,  p-value: < 2.2e-16
```

```
#Plot of mpg against weighlbs with trendline to graphically Visualize the Regression
ggplot(Cars_data,aes(weightlbs,mpg)) +
  geom_point() + geom_smooth(method=lm)+
  ggtitle("Regression plot of MPG against Weight")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Mathematically,Linear Regression is written as y = B0 + B1*x + e.

B0 is intercept and B1 is the slope of the regression line(predicted) e is the residual error.

From the summary, the intercept B0 is 46.0 and coefficient B1 is (-0.0075), the estimated regression line equation for this model can then be written as follow: mpg = 46.00 + (-0.0075)*weightlbs The correlation coefficient between the observed values and the predicted values is represented by R-squared, the value of which is 0.68. This means the model is a good fit and able to explain the variability. Also the RSE, which is the standard deviation of residual errors is 4.45. This is low and indicates that the model prediction is not far from the observed values