Dhirubhai Ambani
Institute of Information and Communication Technology

Subject: Statistical Methods (using R)

Lab R

Name:

Tanya Jagyasi 202118039

Bhumi Bosamia 202118040

importing libraries:

```
library(readxl)
library(modeest)
library(corrplot)
install.packages("dbplyr")
library(ggplot2)
```

reading the csv:

```
> car_data <- read_excel("E:/DA/clg/STATS/project/Cars93.xlsx",
+     sheet = "Cars93")
```

 # Dropping unwanted columns:

> car_data = subset(car_data,select = -(Origin))

> car_data = subset(car_data,select = -(Make))

 # Summary of the dataset

```
> summary(Cars93)
 Manufacturer          Model               Type             Min.Price         Price
 Length:93          Length:93          Length:93          Min.   : 6.70   Min.   : 7.40
 Class :character   Class :character   Class :character   1st Qu.:10.80   1st Qu.:12.20
 Mode  :character   Mode  :character   Mode  :character   Median :14.70   Median :17.70
                                                          Mean   :17.13   Mean   :19.51
                                                          3rd Qu.:20.30   3rd Qu.:23.30
                                                          Max.   :45.40   Max.   :61.90
   Max.Price       MPG.city      MPG.highway      Cylinders       EngineSize      Horsepower
 Min.   : 7.9   Min.   :15.00   Min.   :20.00   Min.   :0.000   Min.   :1.000   Min.   : 55.0
 1st Qu.:14.7   1st Qu.:18.00   1st Qu.:26.00   1st Qu.:4.000   1st Qu.:1.800   1st Qu.:103.0
 Median :19.6   Median :21.00   Median :28.00   Median :4.000   Median :2.400   Median :140.0
 Mean   :21.9   Mean   :22.37   Mean   :29.09   Mean   :4.914   Mean   :2.668   Mean   :143.8
 3rd Qu.:25.3   3rd Qu.:25.00   3rd Qu.:31.00   3rd Qu.:6.000   3rd Qu.:3.300   3rd Qu.:170.0
 Max.   :80.0   Max.   :46.00   Max.   :50.00   Max.   :8.000   Max.   :5.700   Max.   :300.0
      RPM        Rev.per.mile  Fuel.tank.capacity   Passengers        Length        Wheelbase
 Min.   :3800   Min.   :1320   Min.   : 9.20      Min.   :2.000   Min.   :141.0   Min.   : 90.0
 1st Qu.:4800   1st Qu.:1985   1st Qu.:14.50      1st Qu.:4.000   1st Qu.:174.0   1st Qu.: 98.0
 Median :5200   Median :2340   Median :16.40      Median :5.000   Median :183.0   Median :103.0
 Mean   :5281   Mean   :2332   Mean   :16.66      Mean   :5.086   Mean   :183.2   Mean   :103.9
 3rd Qu.:5750   3rd Qu.:2565   3rd Qu.:18.80      3rd Qu.:6.000   3rd Qu.:192.0   3rd Qu.:110.0
 Max.   :6500   Max.   :3755   Max.   :27.00      Max.   :8.000   Max.   :219.0   Max.   :119.0
     Width         Turn.circle    Rear.seat.room      Weight      Luggage.room
 Min.   :60.00   Min.   :32.00   Min.   : 0.00   Min.   :1695   Min.   : 0.00
 1st Qu.:67.00   1st Qu.:37.00   1st Qu.:26.00   1st Qu.:2620   1st Qu.:11.00
 Median :69.00   Median :39.00   Median :27.50   Median :3040   Median :14.00
 Mean   :69.38   Mean   :38.96   Mean   :27.23   Mean   :3073   Mean   :12.25
 3rd Qu.:72.00   3rd Qu.:41.00   3rd Qu.:30.00   3rd Qu.:3525   3rd Qu.:15.00
 Max.   :78.00   Max.   :45.00   Max.   :36.00   Max.   :4105   Max.   :22.00
  DriveTrain          AirBags
 Length:93          Length:93
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Here we have shown the summary of our dataset, which displays the Mean, Median, Quartiles, Minimum, and Maximum values of each column.

 # Linear relation between Horsepower and Price of the car :

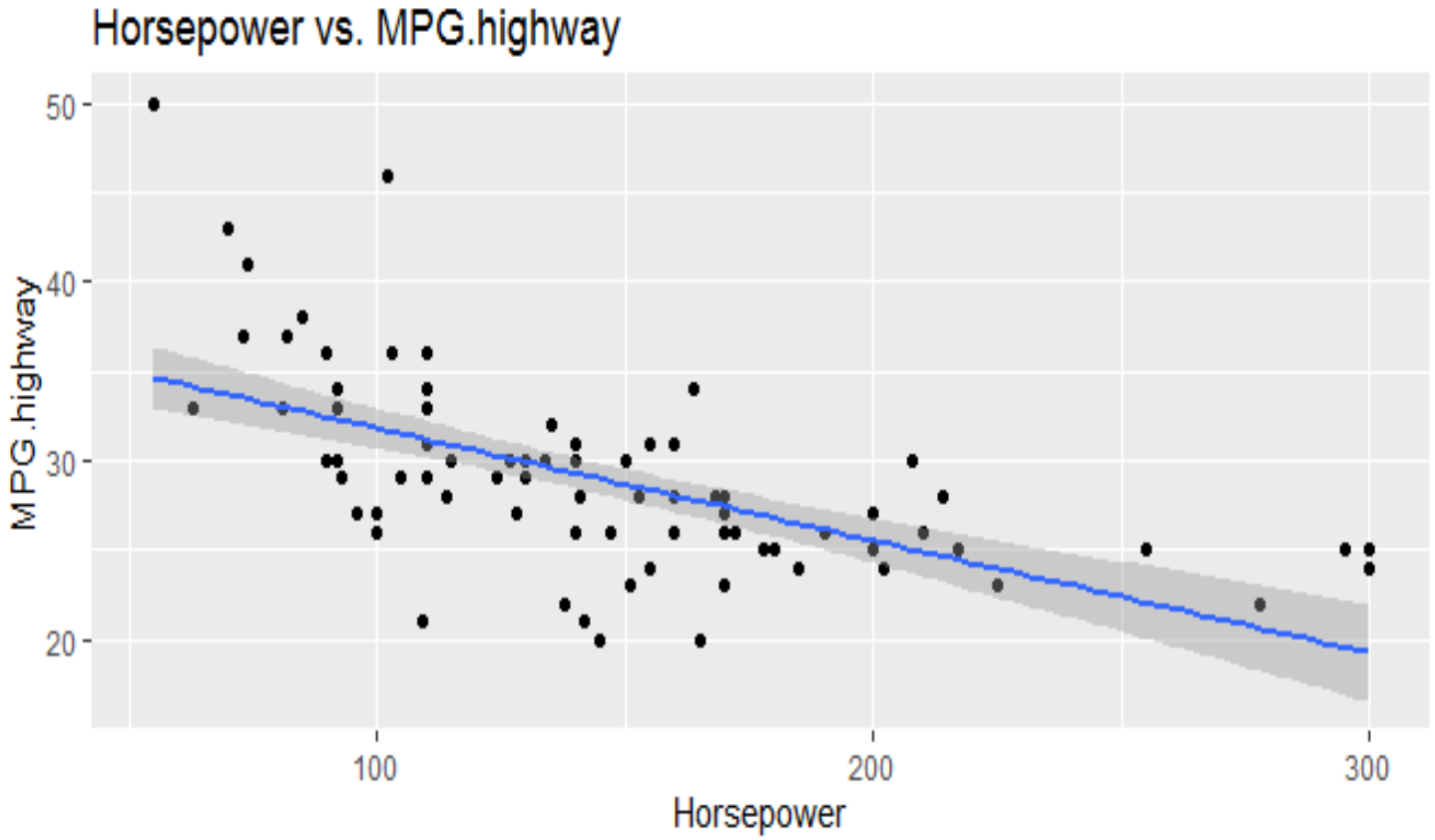
Hypothesis H0: There is no significant relationship between the variables

H1: there is a significant relationship between the variables.

```
> relation = lm(y$Horsepower~y$Price)
> summary(relation)

Call:
lm(formula = y$Horsepower ~ y$Price)

Residuals:
    Min      1Q  Median      3Q     Max
-107.996  -18.401   -5.281   17.973  129.288

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.4476     7.6068   7.947 4.94e-12 ***
y$Price       4.2738     0.3498  12.218  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.41 on 91 degrees of freedom
Multiple R-squared:  0.6213,      Adjusted R-squared:  0.6171
F-statistic: 149.3 on 1 and 91 DF,  p-value: < 2.2e-16
```

 # Scatter plot of Horsepower vs. MPG.highway

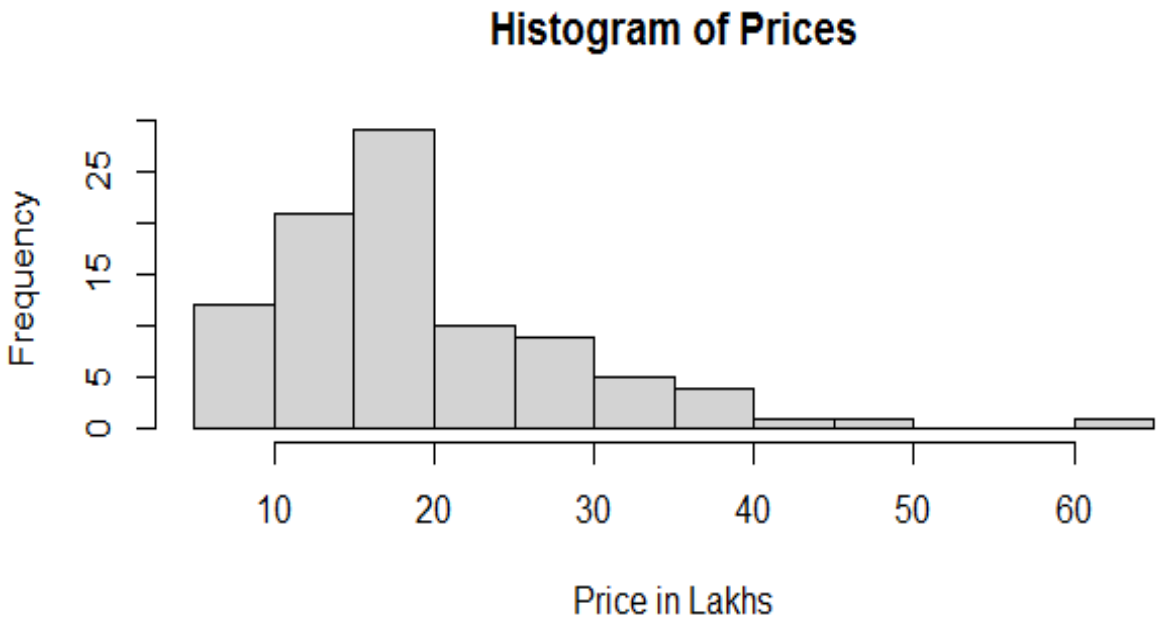
```
>ggplot(data = x, aes(x = Horsepower, y = MPG.highway)) +
   geom_point() +
   geom_smooth(method='lm') +
   xlab('Horsepower') +
   ylab('MPG.highway') +
   ggtitle('Horsepower vs. MPG.highway')
```



Here the plot clearly shows that the Horsepower and MPG are inversely related. We can see that as the Horsepower increases, the MPG highway decreases.

 # Visualizing Histogram:

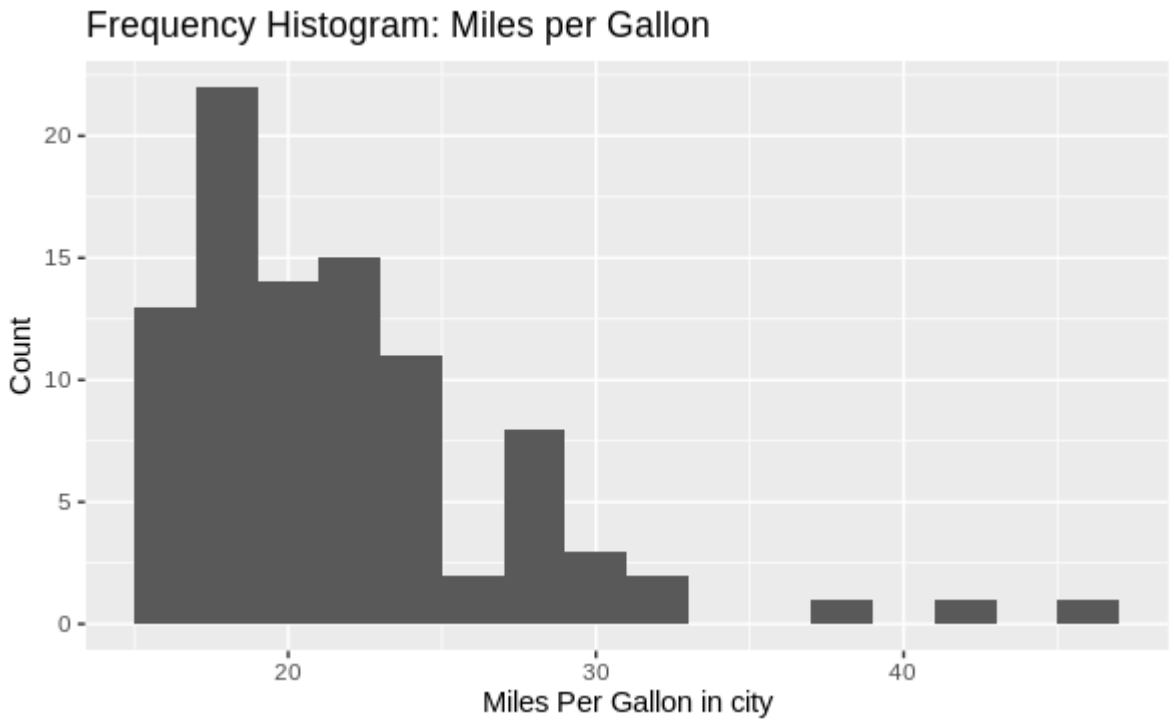
hist(y$Price,main="Histogram of Prices" , xlab='Price' , ylab='Frequency')



Here it shows that the maximum value of the car lies between 10 lakh to 20 lakh.

 # Histogram of cars with respect to its mileage mpg cities

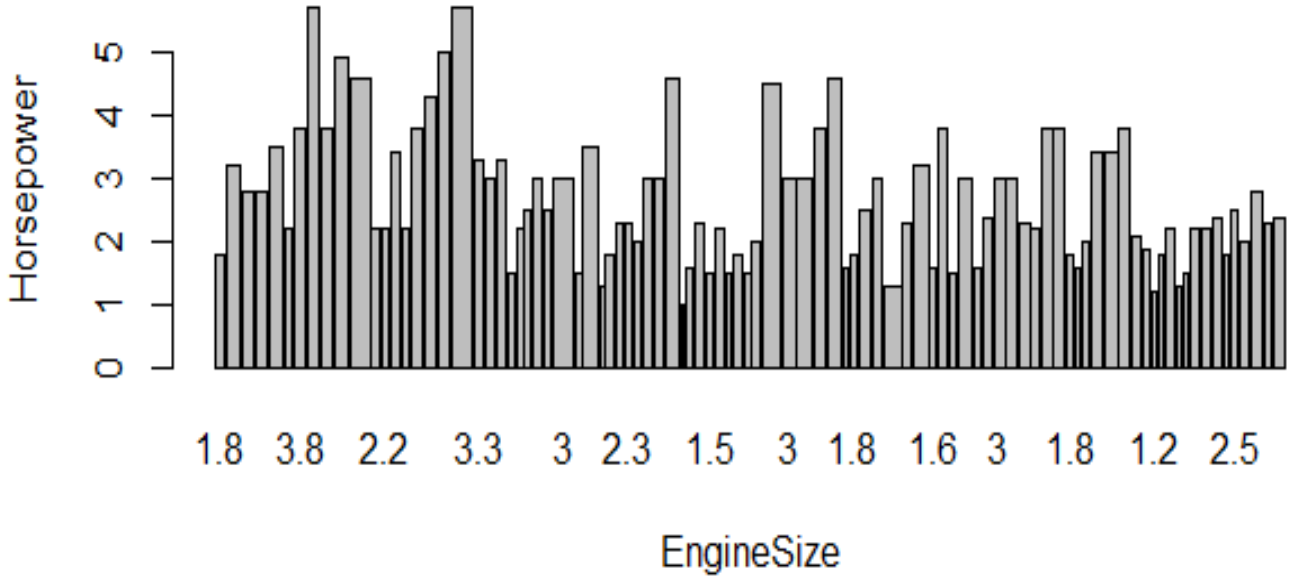
>qplot(Cars93$MPG.city, xlab = 'Miles Per Gallon in city', ylab = 'Count', binwidth = 2, main = 'Frequency Histogram: Miles per Gallon')



It shows that the highest value of MPG lies near 10.

 # Barplot between Engine size and Horsepower:

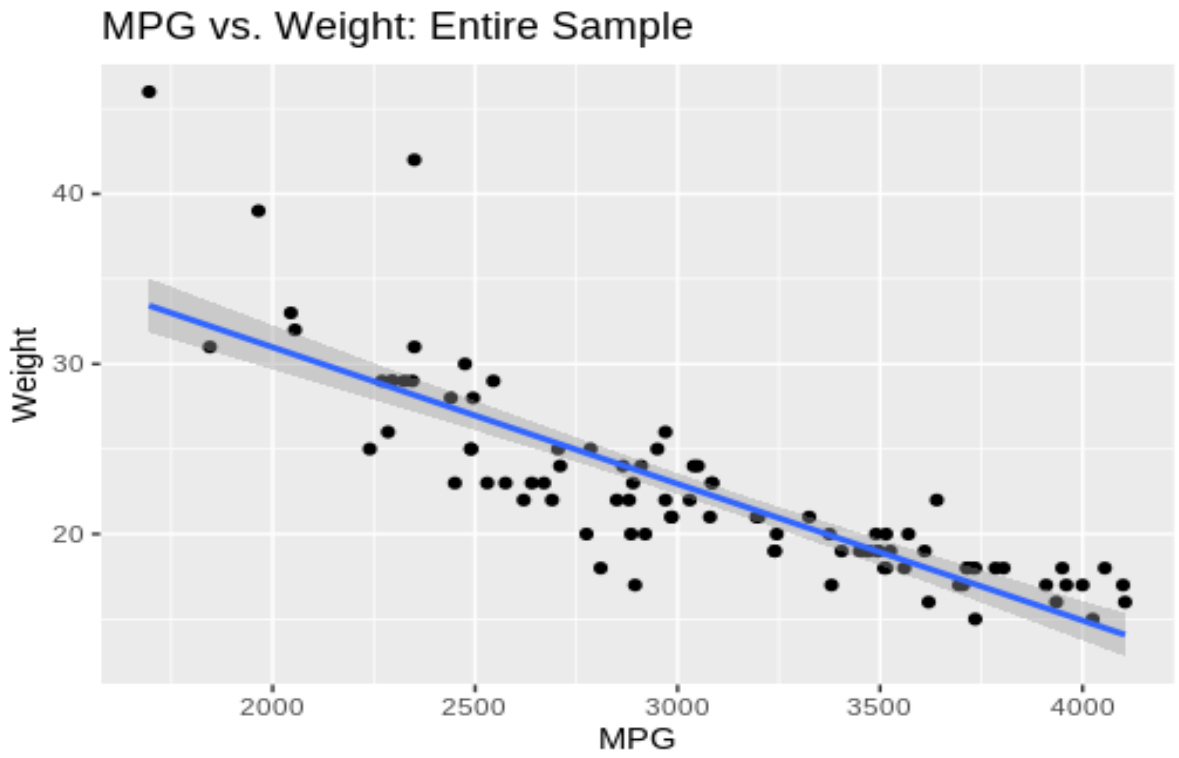
> barplot(y$EngineSize,y$Horsepower,xlab = "EngineSize",ylab = "Horsepower")



It shows the different engine size values for different Horsepower values.

 # Scatter plot of MPG vs. Weight: Entire Sample

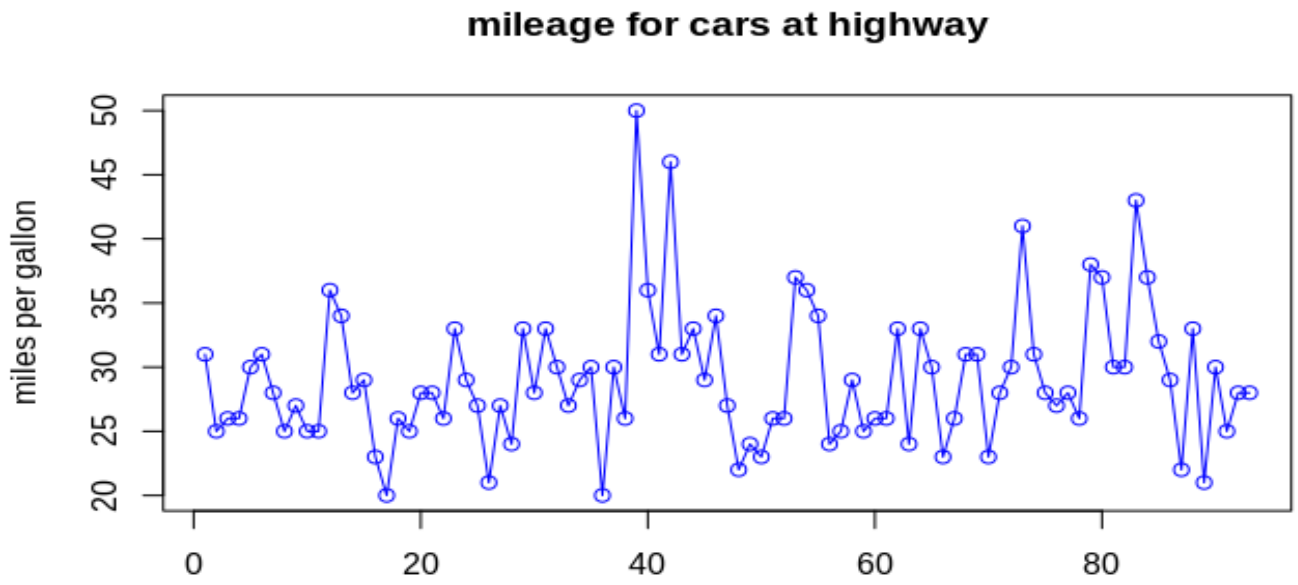
```
> ggplot(data = x, aes(x = Weight, y = MPG.city)) +
+     geom_point() +
+     geom_smooth(method='lm') +
+     xlab('MPG') +
+     ylab('Weight') +
+     ggtitle('MPG vs. Weight: Entire Sample')
```



Here the scatter plot shows MPG and Weight are negatively correlated.

 # line graph of mileage for cars at highway

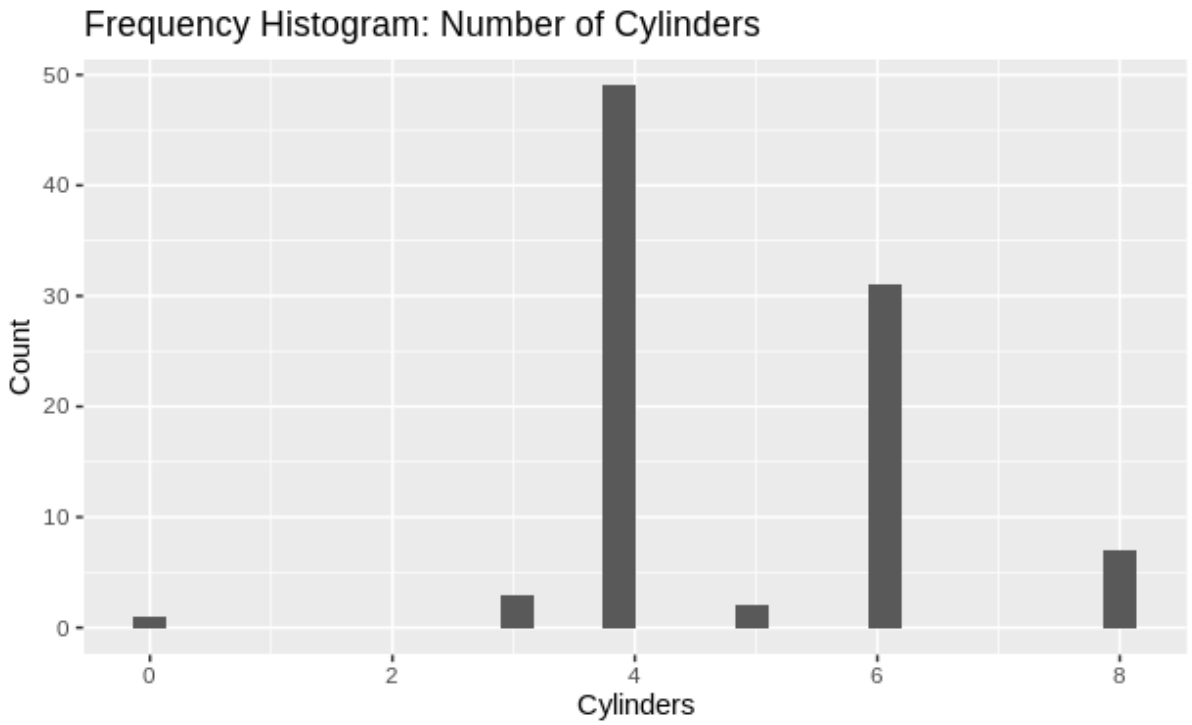
> plot(Cars93$MPG.highway, type = "o", col="blue",xlab = "Model", ylab = "miles per gallon", main = "mileage for cars at highway")



Here we can see at the speed rate of 40 miles per hour, it gives the highest mileage of around 50 miles per gallon.

 # Histogram showing no.of cylinders.

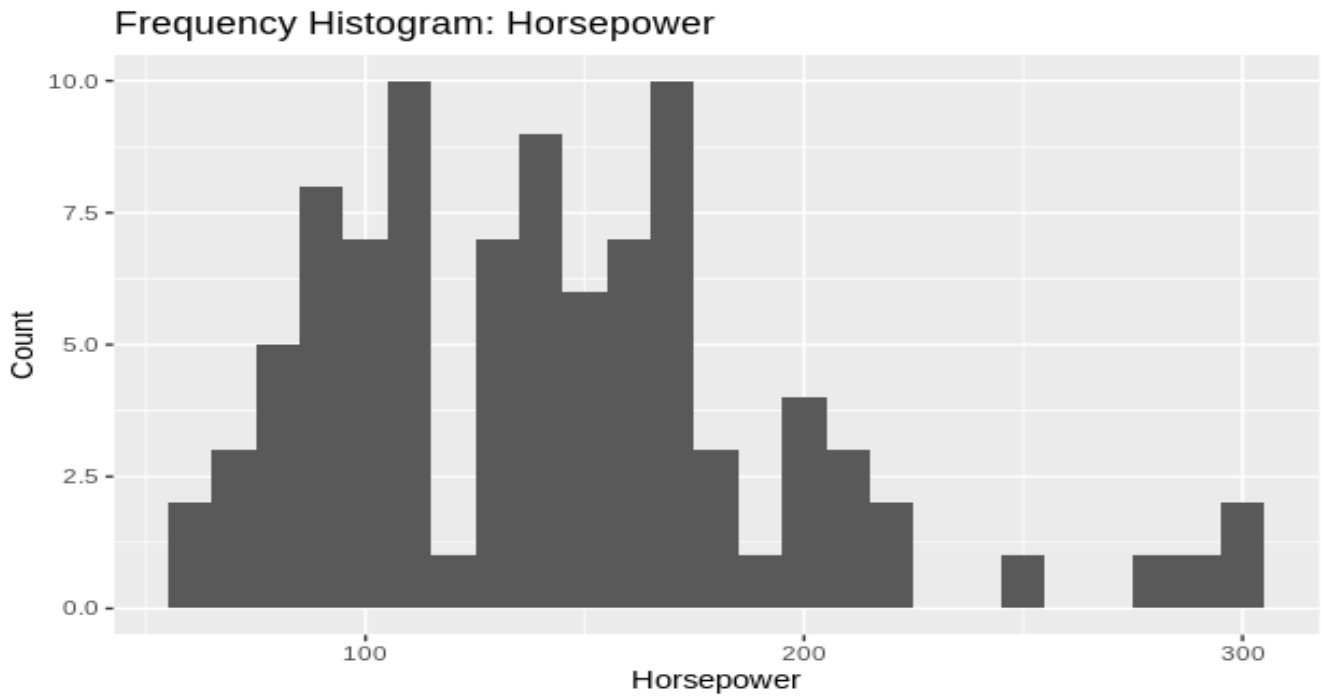
```
> qplot(Cars93$Cylinders, xlab = 'Cylinders', ylab = 'Count',
    main='Frequency Histogram: Number of Cylinders')
> table(Cars93$Cylinders)

 0  3  4  5  6  8
 1  3 49  2 31  7
```



In this data, we can see that maximum cars have four cylinders.

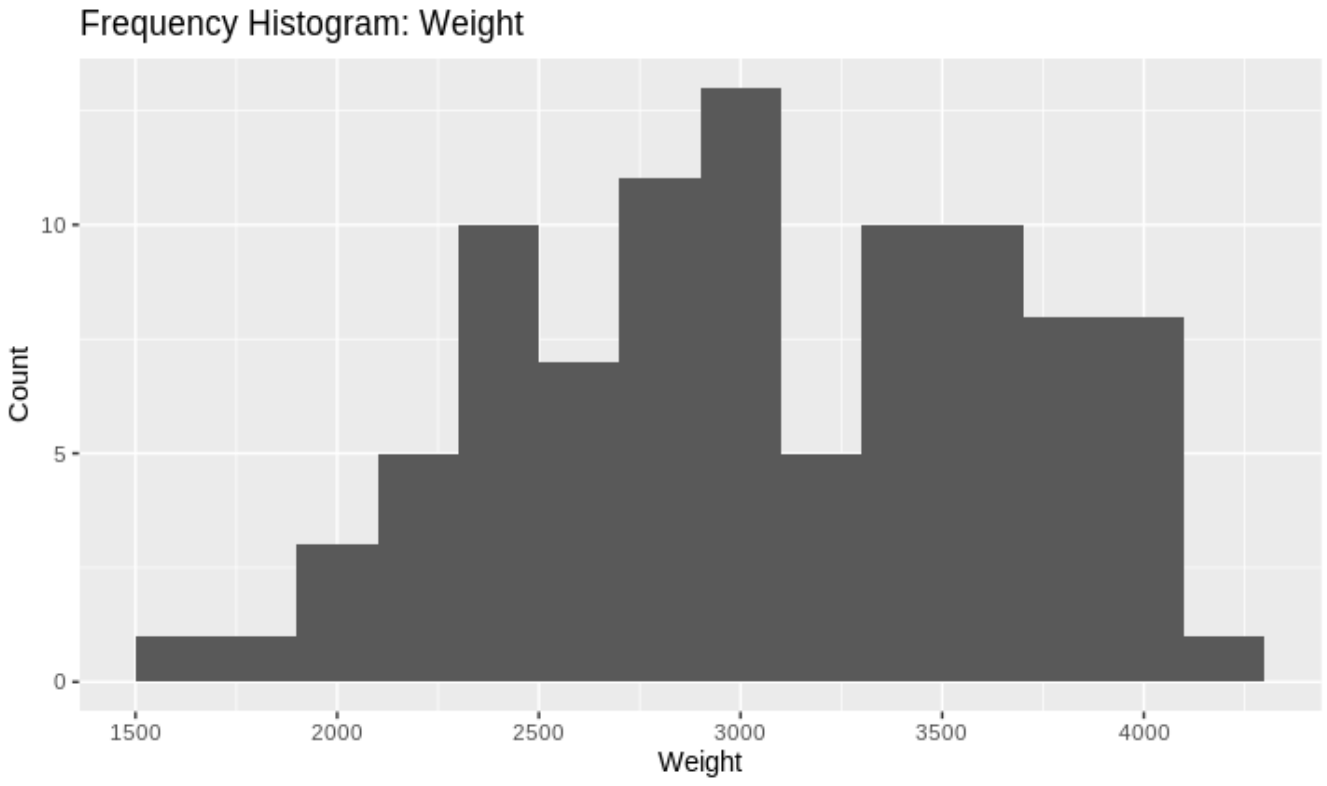
 # Frequency of horsepower

> qplot(Cars93$Horsepower, xlab = 'Horsepower', ylab = 'Count', binwidth = 10, main='Frequency Histogram: Horsepower')



 # Frequency of Weight

```
qplot(Cars93$Weight, xlab = 'Weight', ylab = 'Count', binwidth = 200,
    main='Frequency Histogram: Weight')
```



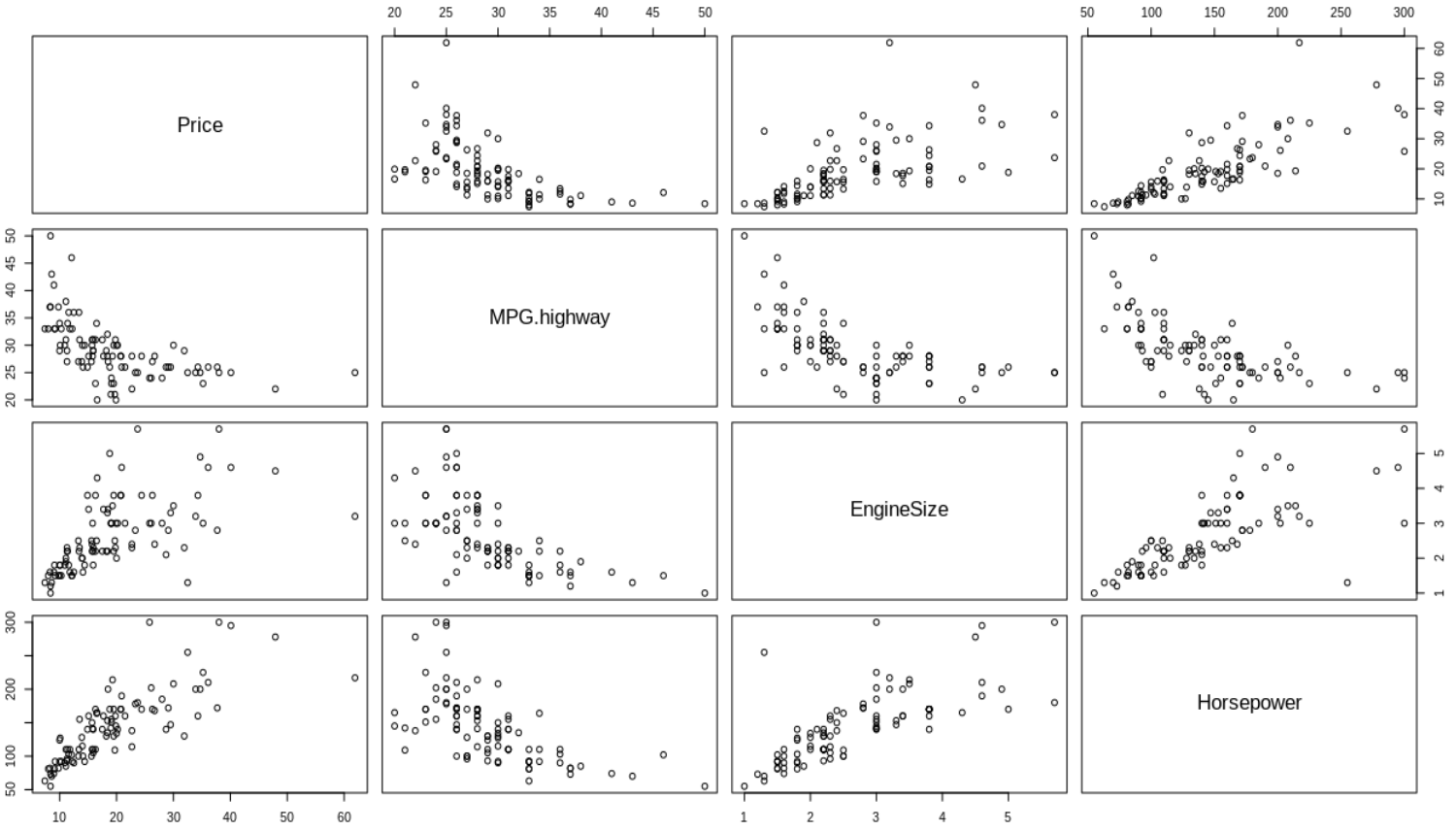
 # correlation between Weight and Horsepower

```
> cor(Cars93[ , c('Weight', 'Horsepower')], use='complete')
            Weight Horsepower
Weight     1.0000000  0.7387975
Horsepower 0.7387975  1.0000000
```

The correlation value is 0.738 so we can say that the it is positively correlated.

 # Pairplot of different columns:

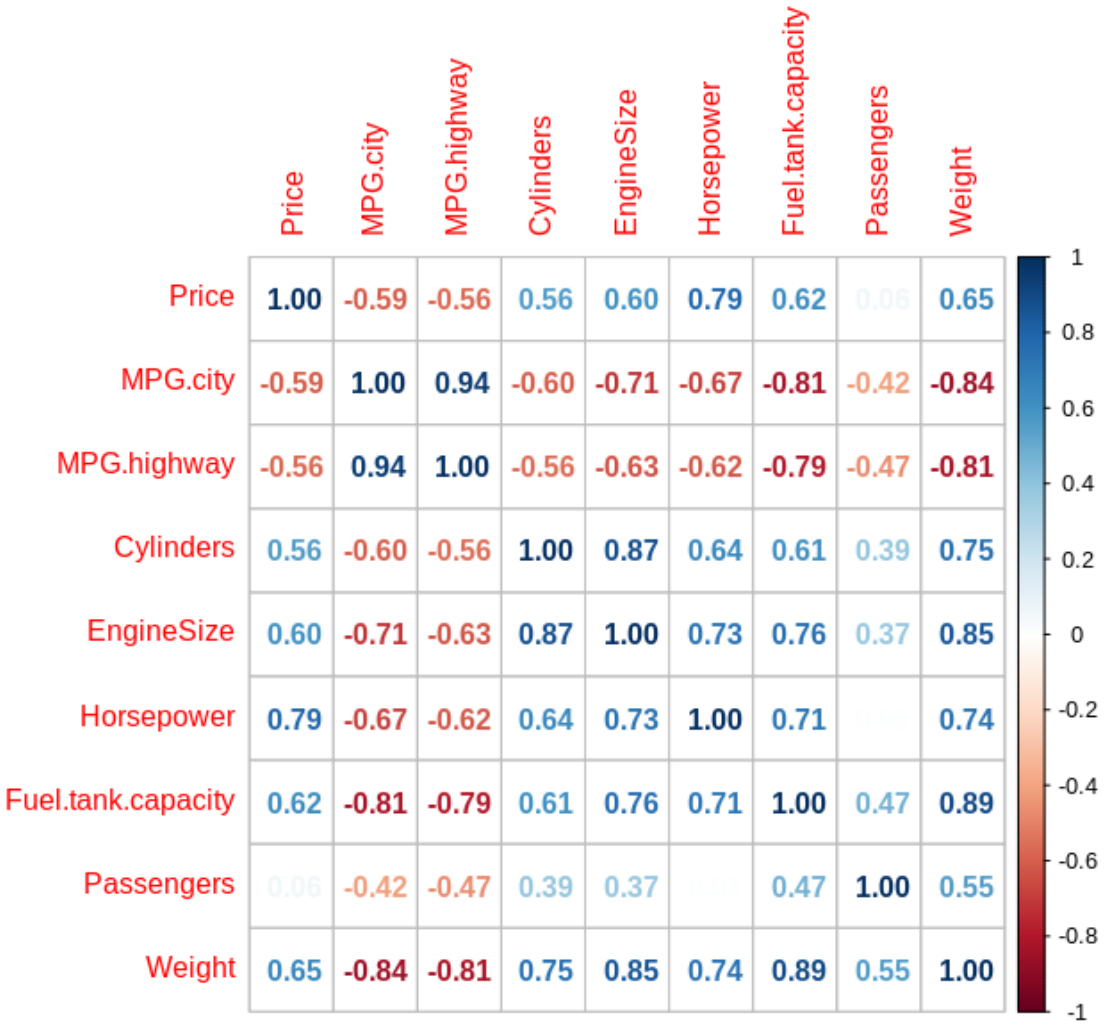
```
> pairs(Cars93[c(5,8,10,11)])
```



Here for the Price and Horsepower, we can see that the values almost lie in a straight line, so they are positively correlated. The following example we can take for Price and MPG Highway the values are going down, so they are negatively correlated.

 # corrplot showing the coorelation between each columns

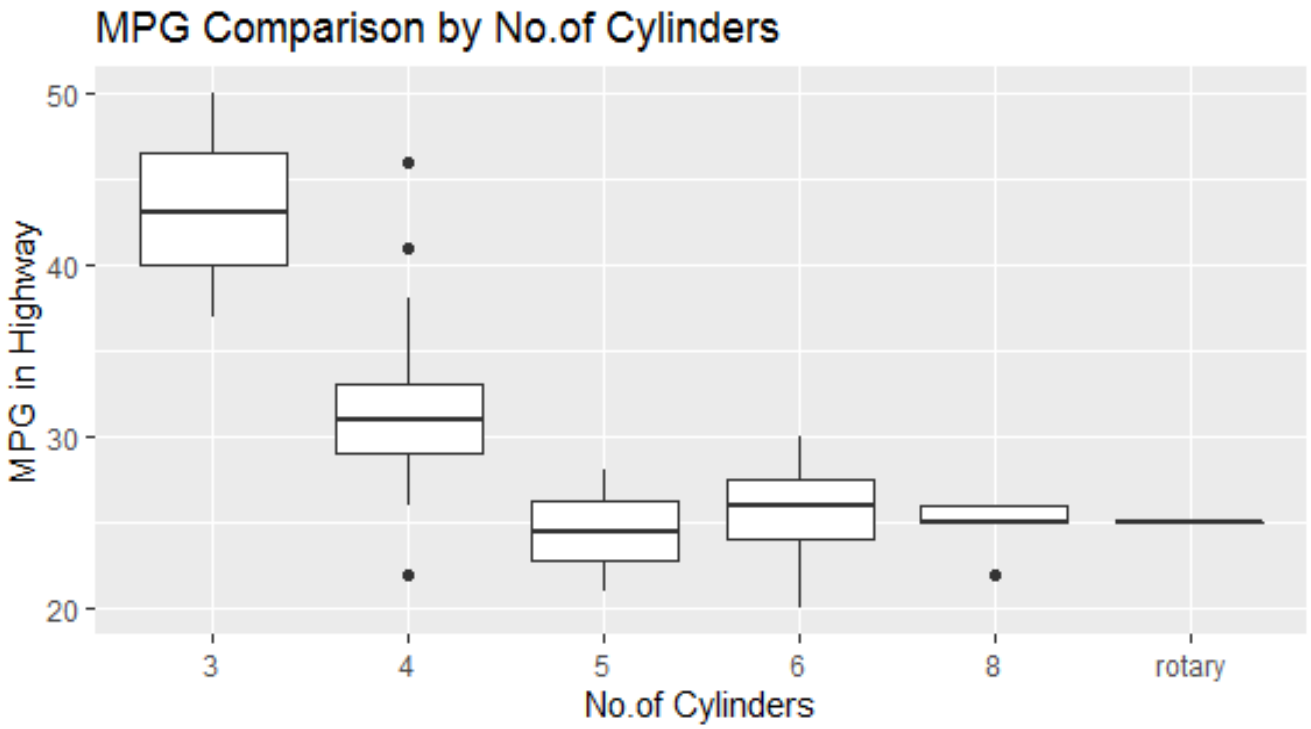
> corrplot(cor(Cars93[c(5,7,8,9,10,11,14,15,21)]),method="number")



Here we can see that the correlation value between weight and horsepower is 0.74, so we can say that they are positively correlated. Similarly, we can see the value for weight and MPG.city is -0.84, so they are negatively correlated.

 # box plot relating no. of cylinders and its mpg

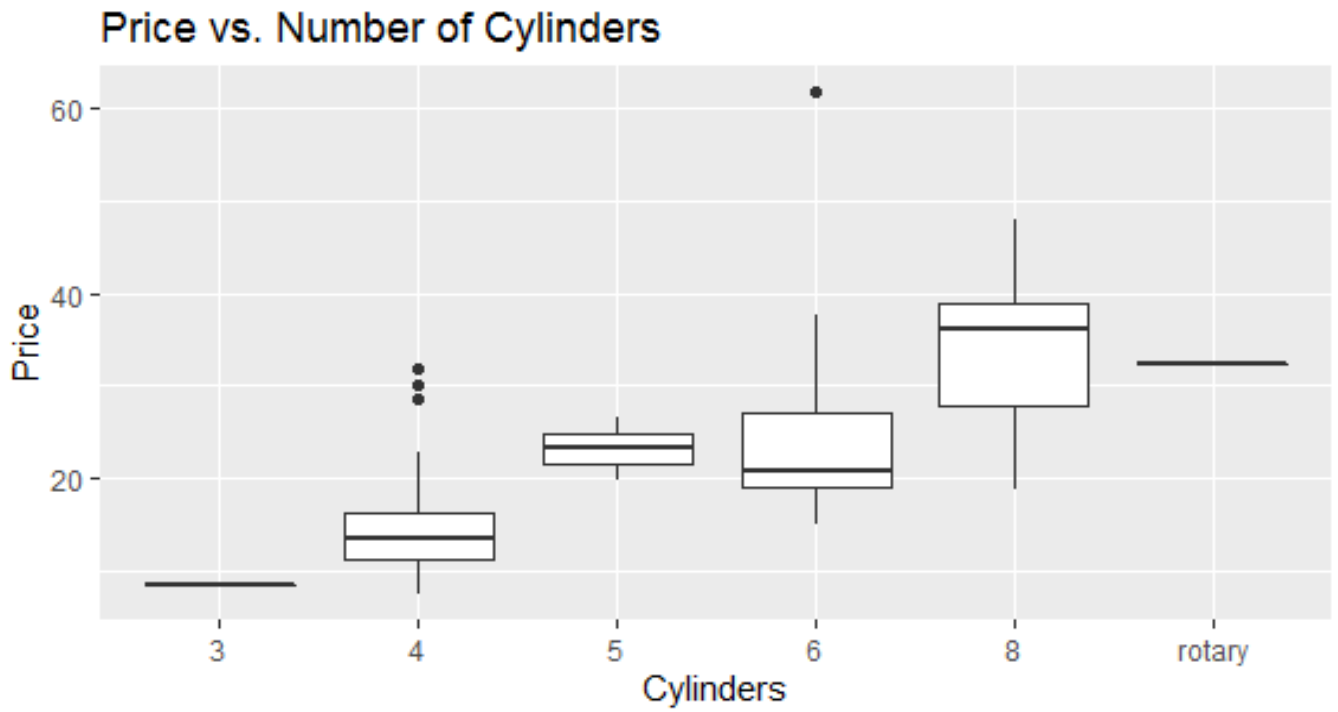
```
> ggplot(data = x, aes(x = Cylinders, y = MPG.highway)) +
+     geom_boxplot() +
+     xlab('No.of Cylinders') +
+     ylab('MPG in Highway') +
+     ggtitle('MPG Comparison by No.of Cylinders')
```



The boxplot shows that according to the number of cylinders and MPG in highway, the maximum values for cylinders lying in the box, and the dots showing the outliers. For the cars having four cylinders, the minimum value is near to 26, the maximum value lies near 39.

 # box plot relating Price vs. Number of Cylinders

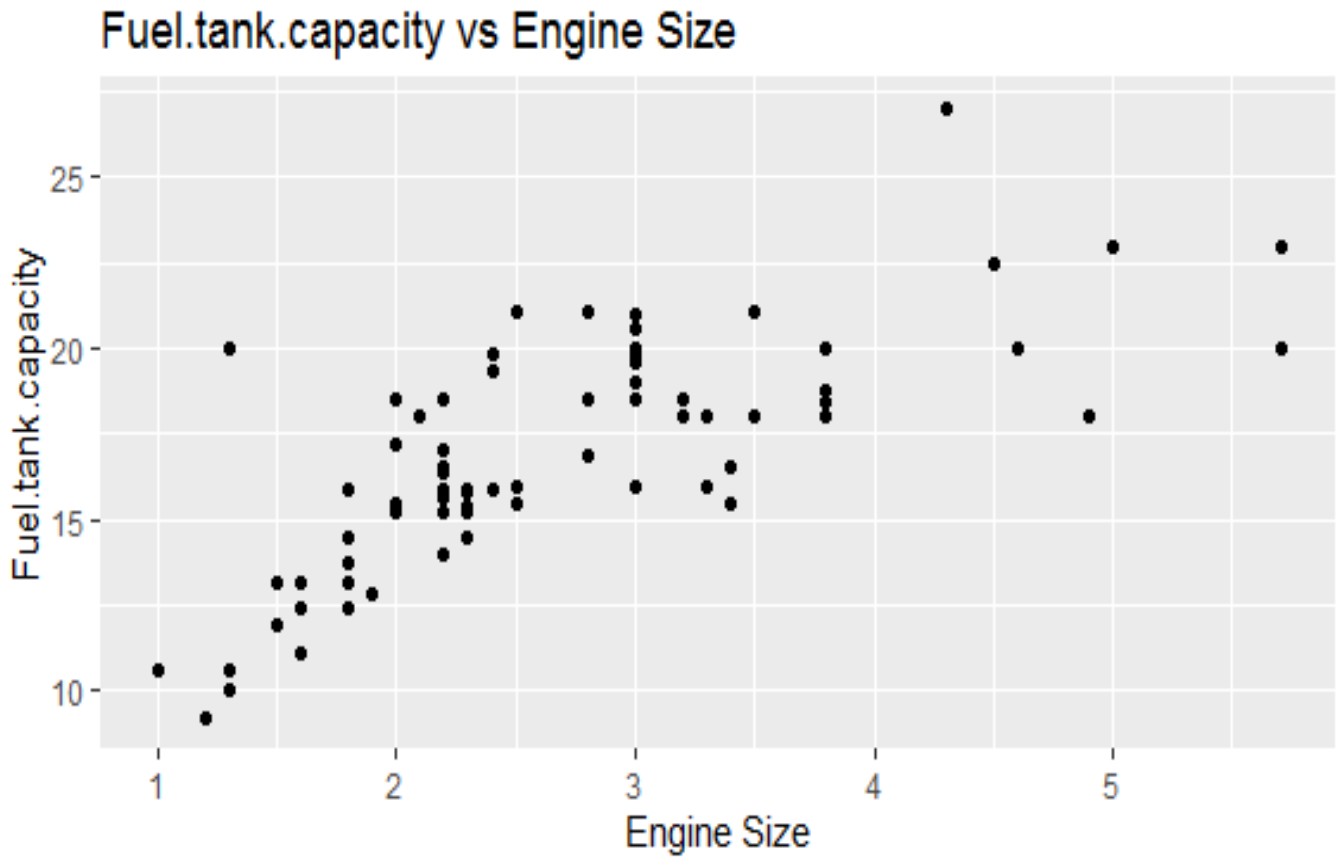
```
> ggplot(data = x, aes(x = Cylinders, y = Price)) +
+     geom_boxplot() +
+     xlab('Cylinders') +
+     ylab('Price') +
+     ggtitle('Price vs. Number of Cylinders')
```



Here for the car having 8 cylinders have the box lying between the value near to 29 to 40 lakhs so we can essay that the price for the car having 8 cylinders lies in this range. The dots shows the outliers. We can say that we have values lying near that point also.

#scatter plot of Fuel.tank.capacity vs Engine Size

```
> ggplot(data = x, aes(x = EngineSize, y = Fuel.tank.capacity)) +
+     geom_point() +
+     xlab('Engine Size') +
+     ylab('Fuel.tank.capacity') +
+     ggtitle('Fuel.tank.capacity vs Engine Size')
```



Here the graph shows that the Fuel.tank.capacity and Engine Size are positively correlated but not highly correlated as the values are scattering more.

 # Conclusion:

Data visualization in R and RStudio makes it easy to plot basic functions or do statistical analysis or apply more advanced functions through packages.

As we have noticed throughout this project, the undeniable added value of R/RStudio compared to the more classical resources such as Excel is the ability to produce publication-ready graphic.We used default functions and options, which already produce a highly controlled output quality, pre-definedfine advanced options and use them in variables.

Moreover, we ensured reproducibility of our output by writing our codes in script. This would allow us to apply changes in our data.Through this analysis, we got to know that how different

factors are connected in cars. Using histograms and bar plots we know the peak values of our dataset. Dplyr used for Data Manipulation. Corrplot to see the relationship between the attributes we have. Using different data in graphical format helped us to visualize the data more efficiently. Overall, it was a great learning experience