GPM (Gallons Per Mile) Prediction from relevant variables

1. Descriptive Analysis

We start our analysis by loading the data into R by running the following commands

```
setwd("C:/Users/Anup/Downloads")
working.data = read.csv("FuelEfficiency.csv")
attach(working.data)
```

We have loaded our data file named FuelEfficiency.csv and stored it in the variable called working.data. We have also attached the data using the attach function so that we can use the variable names directly while analysing the data.

We can see the structure of our data using the str function.

```
'data.frame':  38 obs. of  8 variables:
 $ MPG: num  16.9 15.5 19.2 18.5 30 27.5 27.2 30.9 20.3 17 ...
 $ GPM: num  5.92 6.45 5.21 5.41 3.33 ...
 $ WT : num  4.36 4.05 3.6 3.94 2.15 ...
 $ DIS: int  350 351 267 360 98 134 119 105 131 163 ...
 $ NC : int  8 8 8 8 4 4 4 4 5 6 ...
 $ HP : int  155 142 125 150 68 95 97 75 103 125 ...
 $ ACC: num  14.9 14.3 15 13 16.5 14.2 14.7 14.5 15.9 13.6 ...
 $ ET : int  1 1 1 1 0 0 0 0 0 0 ...
```

We can see that our data file has a dataframe with 38 observations divided into 8 variables viz *MPG* (Miles per gallon), *GPM* (Gallon per mile), *WT* (Weight of Engine), *DIS* (Displacement), *NC* (No. of Cylinders), *HP* (Horsepower), *ACC* (Acceleration) and *ET* (Engine Type).

Let us perform the descriptive analysis on our data to find the min, 1st quartile, mean, median, 3rd quartile and the maximum value of each variable. For these i have used *for* loop so that these don't have to be calculated by running the command each time. The summary function is used to calculate the above values.

```
#Descriptive analysis of each variable
for (i in 1:7){
  cat(sprintf("Currently calculating summary for: %s",
              colnames(working.data[i])), sprintf("\n"))
  stats = summary(working.data[i])
  cat(sprintf("%s",stats),sprintf("\n"),sprintf("\n"))
```

```
}
```

The above code gives us the following output.

```
Currently calculating summary for: MPG
Min.   :15.50   1st Qu.:18.52   Median :24.25   Mean   :24.76
3rd Qu.:30.38   Max.   :37.30

Currently calculating summary for: GPM
Min.   :2.681   1st Qu.:3.292   Median :4.160   Mean   :4.331
3rd Qu.:5.398   Max.   :6.452

Currently calculating summary for: WT
Min.   :1.915   1st Qu.:2.208   Median :2.685   Mean   :2.863
3rd Qu.:3.410   Max.   :4.360

Currently calculating summary for: DIS
Min.   : 85.0   1st Qu.:105.0   Median :148.5   Mean   :177.3
3rd Qu.:229.5   Max.   :360.0

Currently calculating summary for: NC
Min.   :4.000   1st Qu.:4.000   Median :4.500   Mean   :5.395
3rd Qu.:6.000   Max.   :8.000

Currently calculating summary for: HP
Min.   : 65.0   1st Qu.: 78.5   Median :100.0   Mean   :101.7
3rd Qu.:123.8   Max.   :155.0

Currently calculating summary for: ACC
Min.   :11.30   1st Qu.:14.03   Median :14.80   Mean   :14.86
3rd Qu.:15.78   Max.   :19.20
```

The above output from the code is self-explanatory about the Minimum, 1st quartile, Median, Mean, 3rd quartile and Maximum values repectively. We can also see the standard deviation and variance of the variables using `var(working.data[,columnNumber])` and `sd(working.data[,columnNumber])`.

We use table function for *ET* (Engine Type) variable since it has two levels either V type engine is there or not.

```
ET
 0  1
27 11
```

This means that 27 vehicles don’t have V type engine and the rest have V type engine.

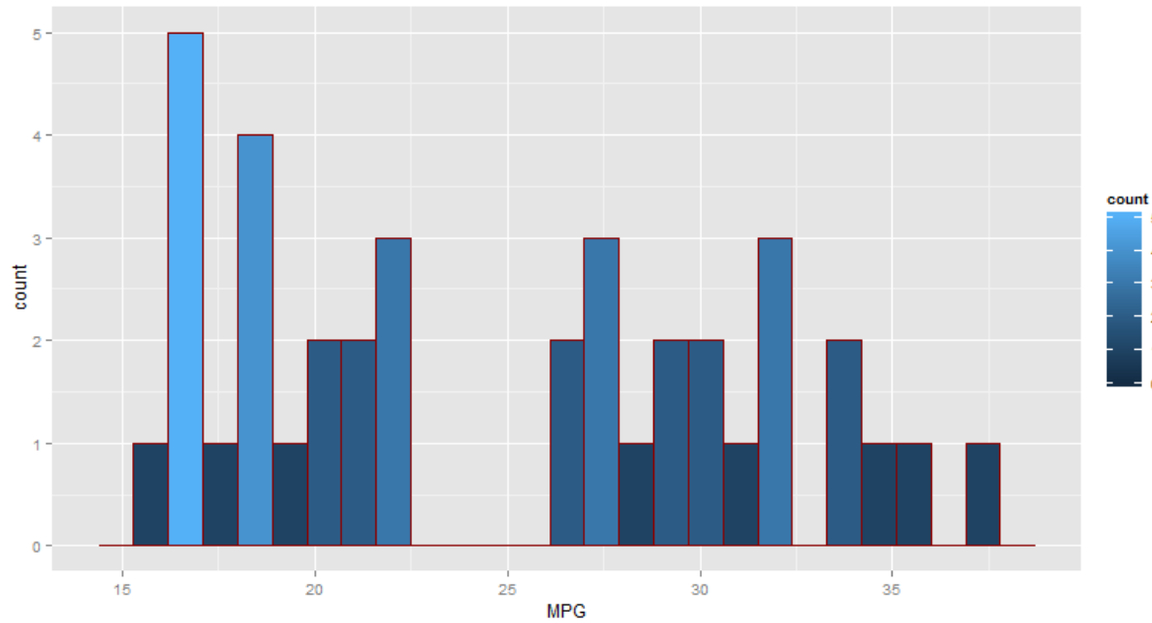
2. Visual Representation of Data

To visualize the data, I have used histograms. For that, ggplot2 library is used. Let's look at the histograms for the individual variable data.

Firstly we plotted histogram for MPG

```
#Plotting the data variables as histograms
#For MPG
ggplot(aes(x=MPG), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=0.9,
                 color = 'dark red')
```

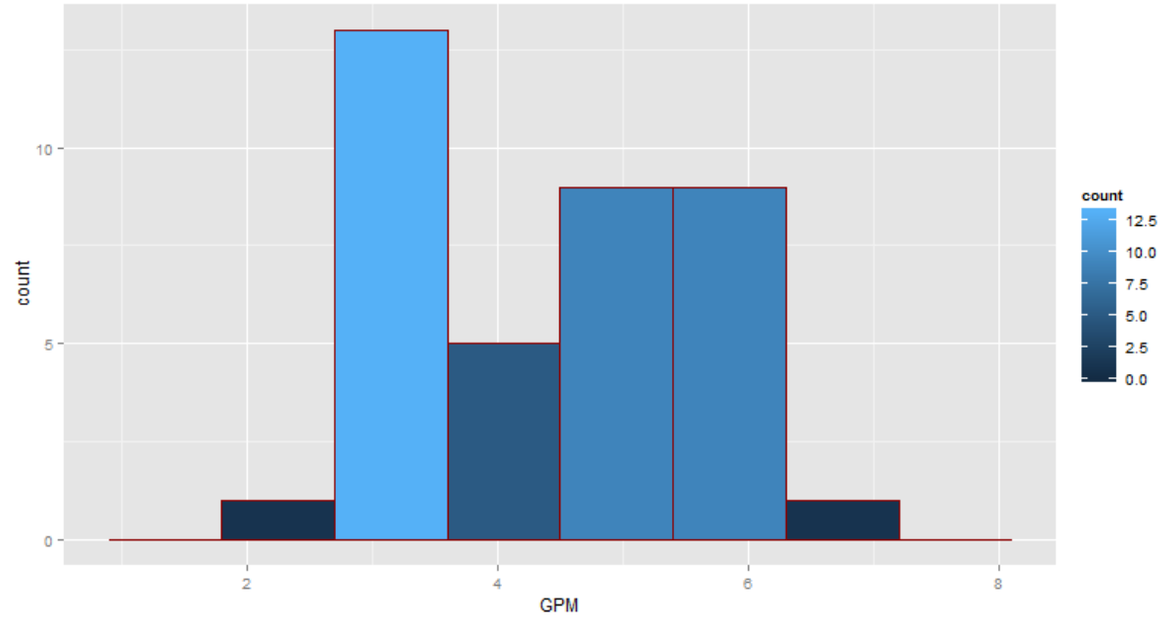
The output graph is as shown below.



We can see that none of the cars have the MPG between 22.5 to 25.5.

The next histogram is for GPM.

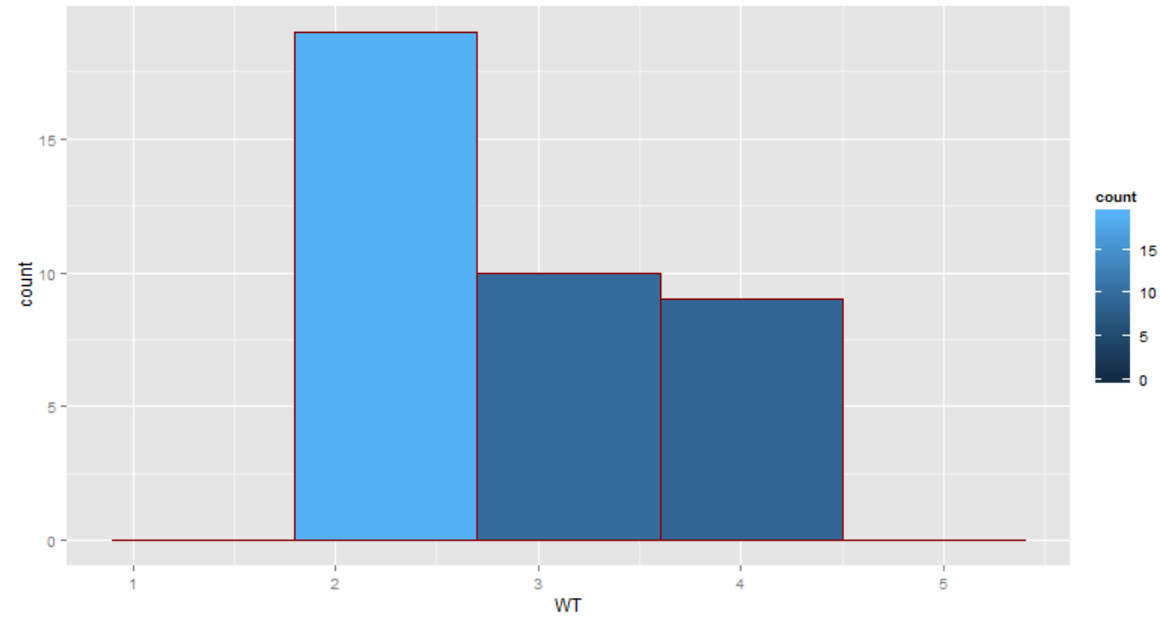
```
#For GPM
ggplot(aes(x=GPM), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=0.9,
                 color = 'dark red')
```



Here, we can see that GPM is more than 7 in most of the cars.

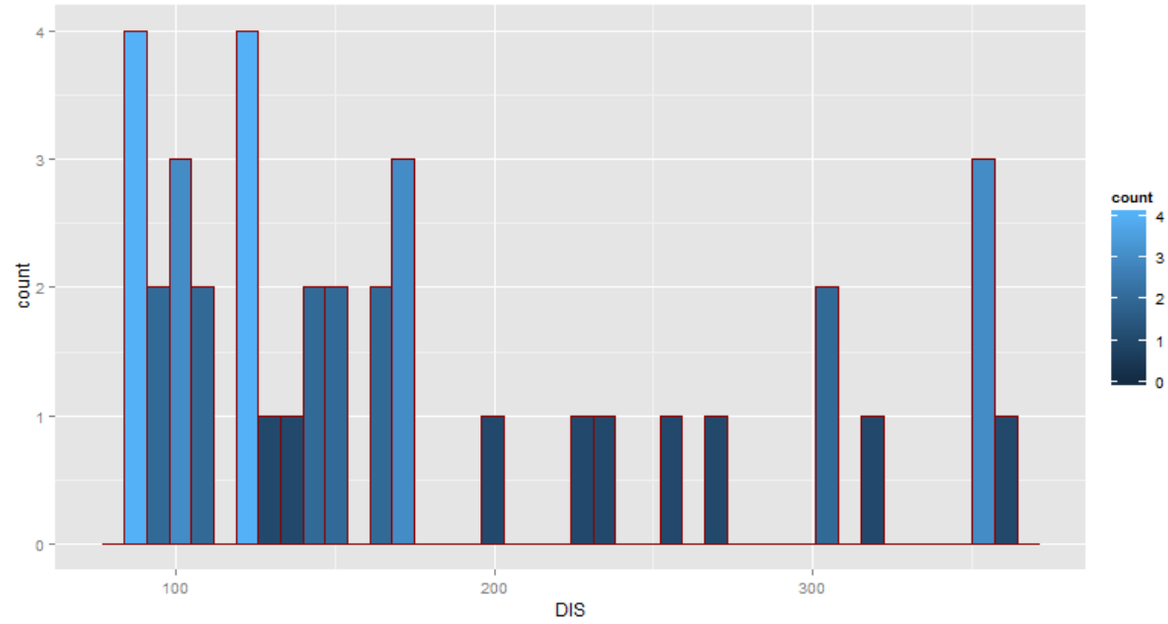
The next histogram is of WT

```
#For WT
ggplot(aes(x=WT), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=0.9,
                 color = 'dark red')
```



The next histogram is of DIS

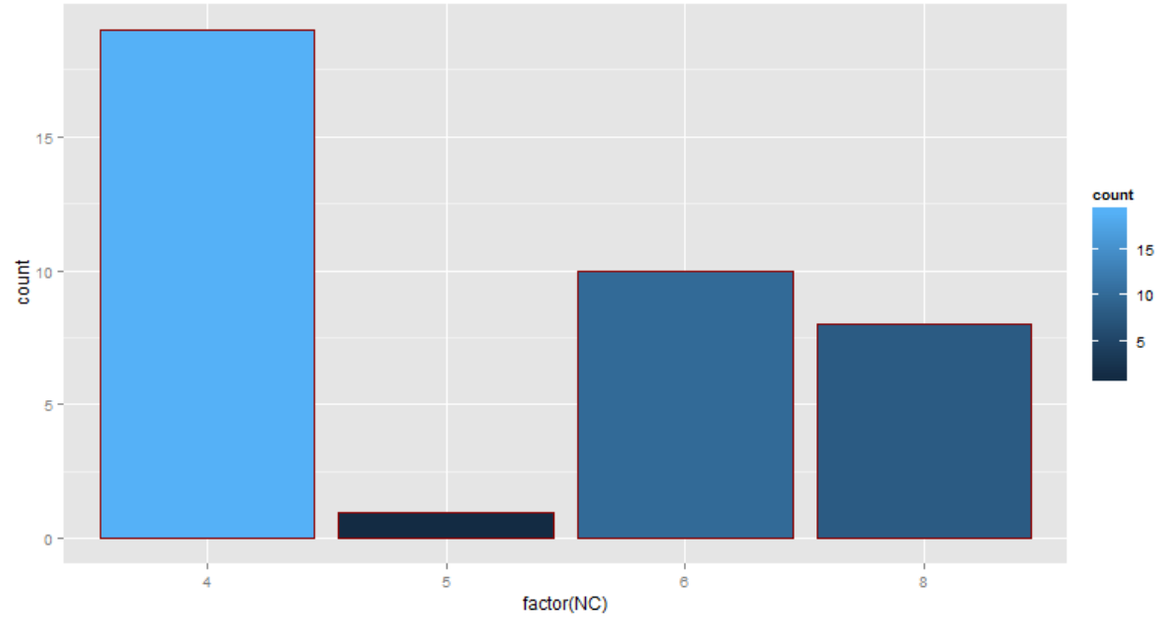
```
#For DIS
ggplot(aes(x=DIS), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=7,
                 color = 'dark red')
```



The above plot shows that the majority of the cars have displacement 100 to 180 cc.

The next histogram is of NC.

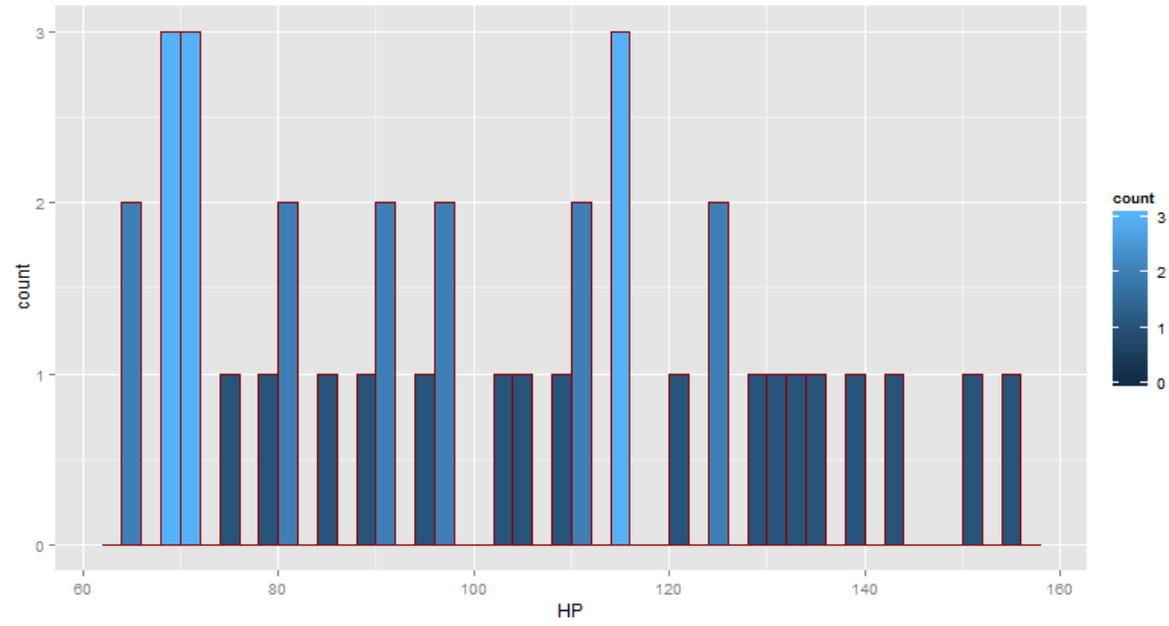
```
#For NC
ggplot(aes(x=factor(NC)), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=0.9,
                 color = 'dark red')
```



Here, we can see that most of the vehicles have 4 cylinders and only few vehicles have one cylinder installed in them.

The next histogram is of HP.

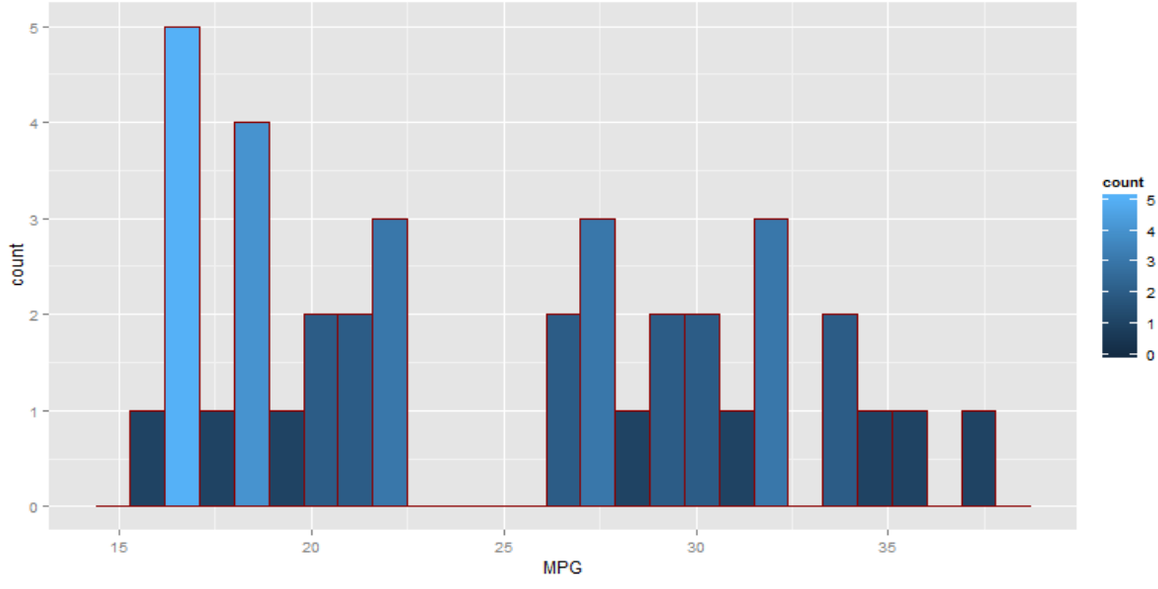
```
#For HP
ggplot(aes(x=HP), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=2,
                 color = 'dark red')
```



This plot shows that there is atleast one car with Horse powe in the range 70 to 160 with irregular gaps with no cars having few specific horse powers.

The next histogram is of ACC.

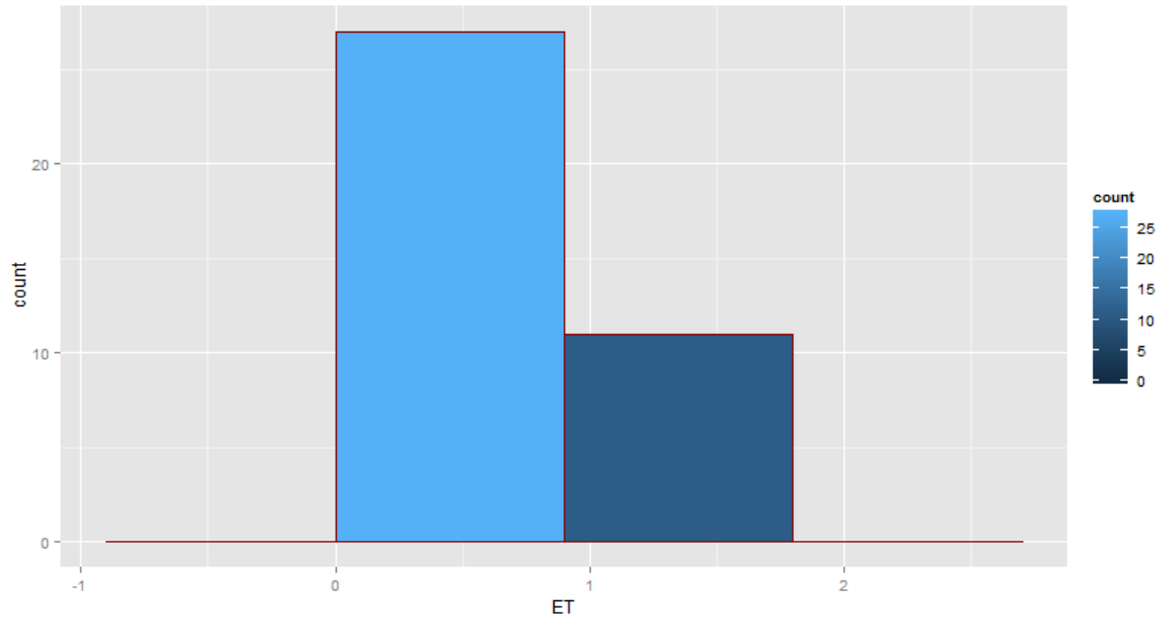
```
#For ACC
ggplot(aes(x=MPG), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=0.9,
                 color = 'dark red')
```



There is an irregular trend in the accelarations of cars with many cars with accelarations between 17 – 23 mph and then again between 26 to 32 mph.

Finally we have the histogram of ET.

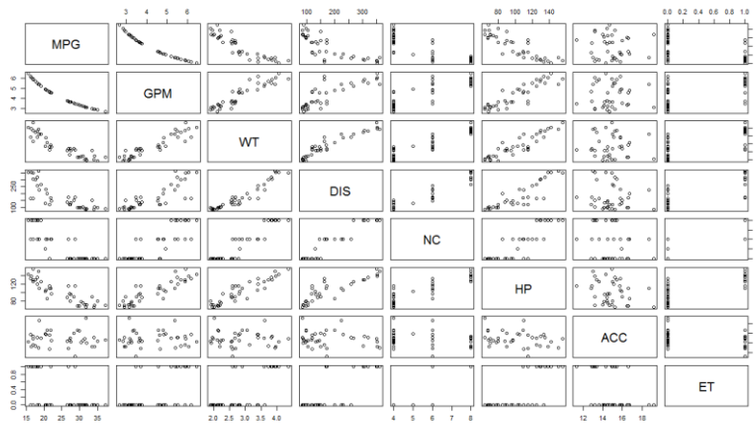
```
#For ET
ggplot(aes(x=ET), data=working.data) +
  geom_histogram(aes(fill = ..count..), binwidth=0.9,
                 color = 'dark red')
```



The histogram depicts that most of the cars don’t have the V-type engine in them.

Let us see the relationship of the variables among one another with the help of a scatter plot using pairs function.

```
#Showing relationship between all the variables using a
#scatter plot
pairs(working.data)
```

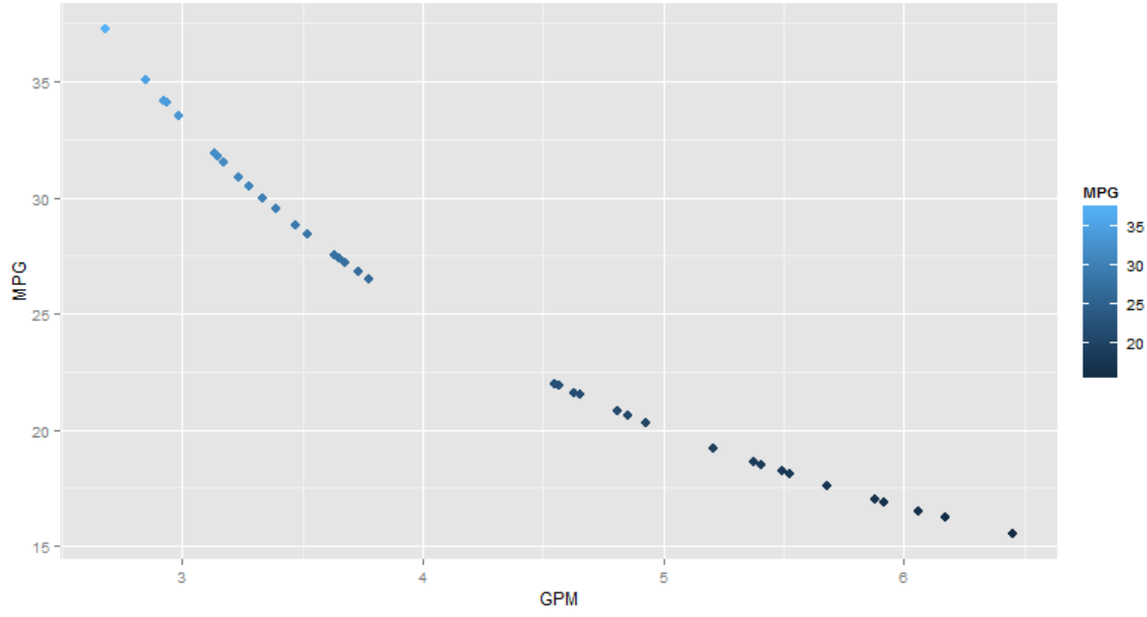


See attached scatter plot with the file for full view.

Now, let us visualize the dependence of GPM on different variable with the help of point or scatter plots. I have used ggplot2 library for the visualizations.

The first plot is of GPM with respect to MPG.

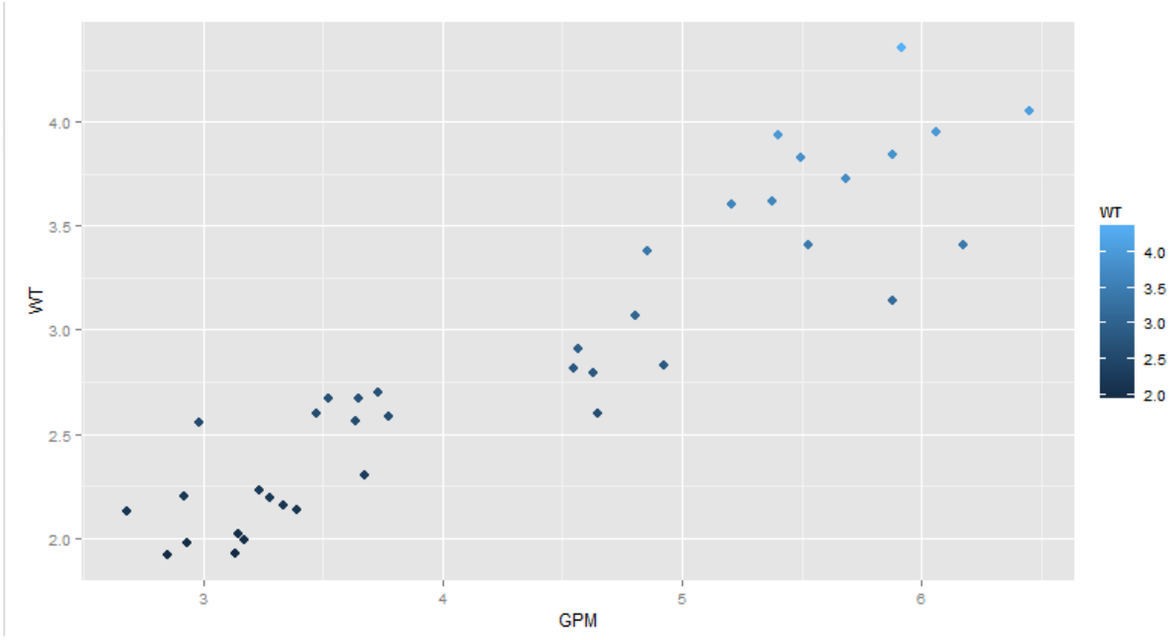
```
#Plotting the relationship of GPM with other variables as
scatterplots
#For MPG
ggplot(aes(x=GPM, MPG), data=working.data) +
  geom_point(aes(color = MPG), size=3)
```



This plot shows the GPM and MPG seem to have a negative exponential relationship between them. This means that as MPG increases GPM decreases exponentially.

Second plot is of GPM with respect to WT.

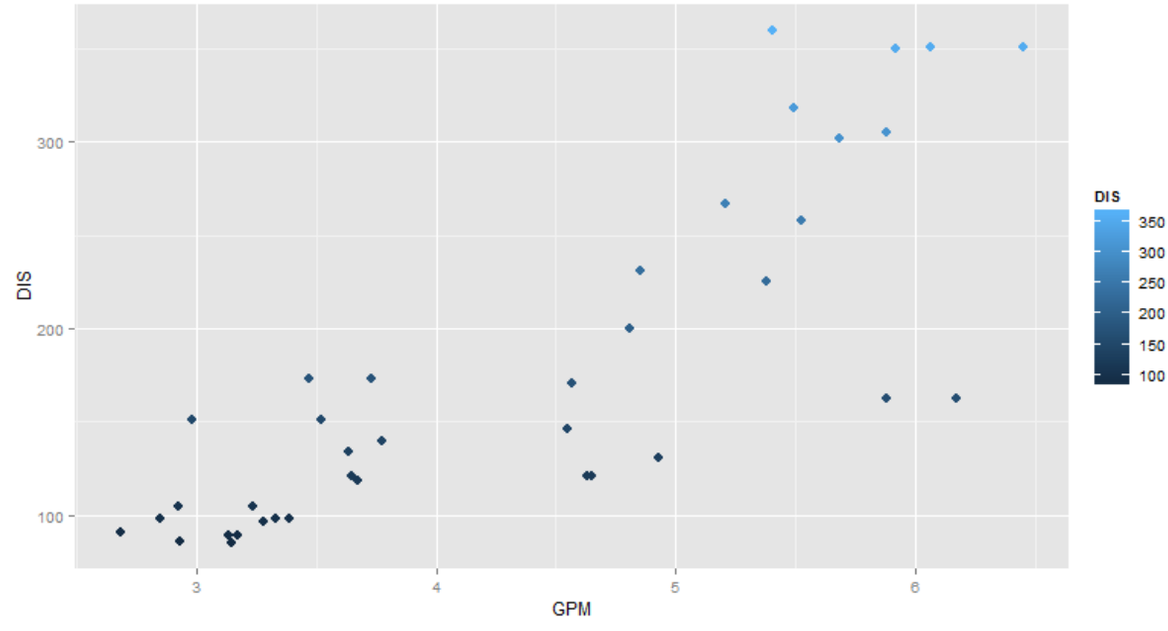
```
#For WT
ggplot(aes(x=GPM,WT), data=working.data) +
  geom_point(aes(color = WT), size=3)
```



This plot shows the GPM and WT are positively corelated and with increase in weight of the car, the gallons per mile increase.

Third plot is of GPM with respect to DIS.

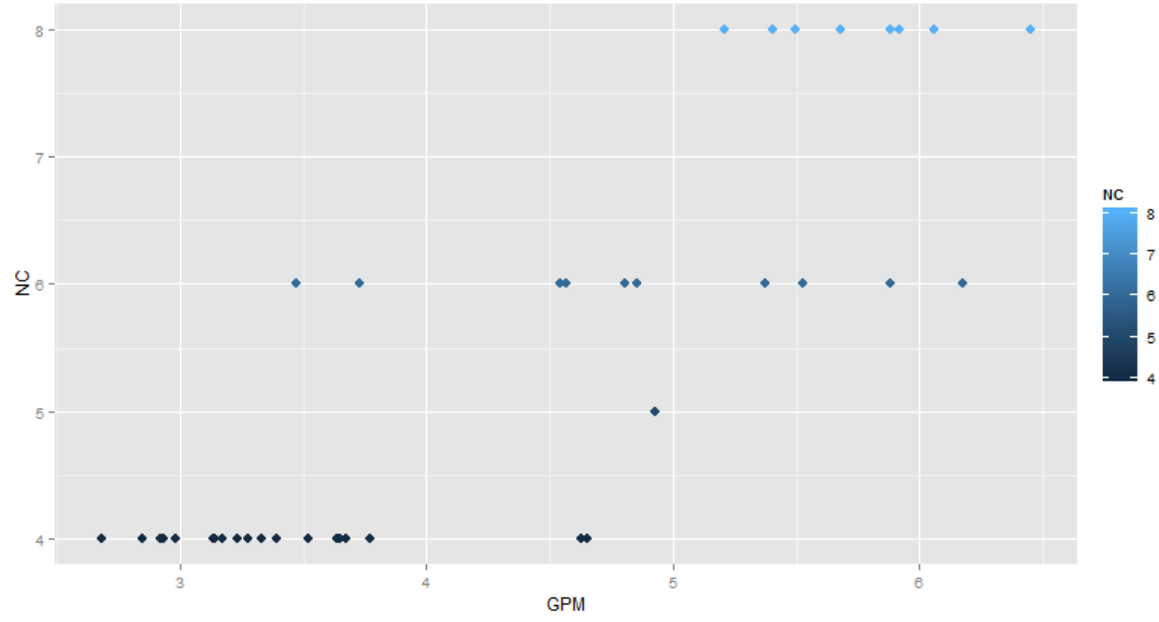
```
#For DIS
ggplot(aes(x=GPM, DIS), data=working.data) +
  geom_point(aes(color = DIS), size=3)
```



This plot shows the similar trends as that of GPM and WT. These two too are positively corelated and with increase in displacement of the car, the gallons per mile increases.

Fourth plot is of GPM with respect to NC.

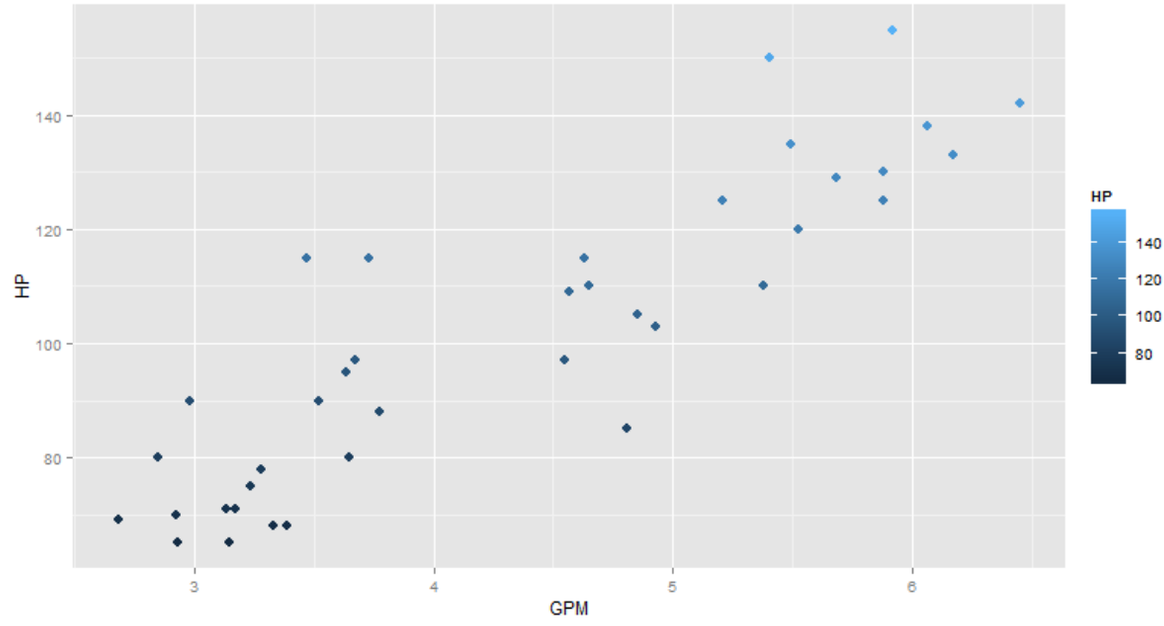
```
#For NC
ggplot(aes(x=GPM, NC), data=working.data) +
  geom_point(aes(color = NC), size=3)
```



This plot shows the clusters at four levels which can be seen in the structure of data too.

Fifth plot is of GPM with respect to HP.

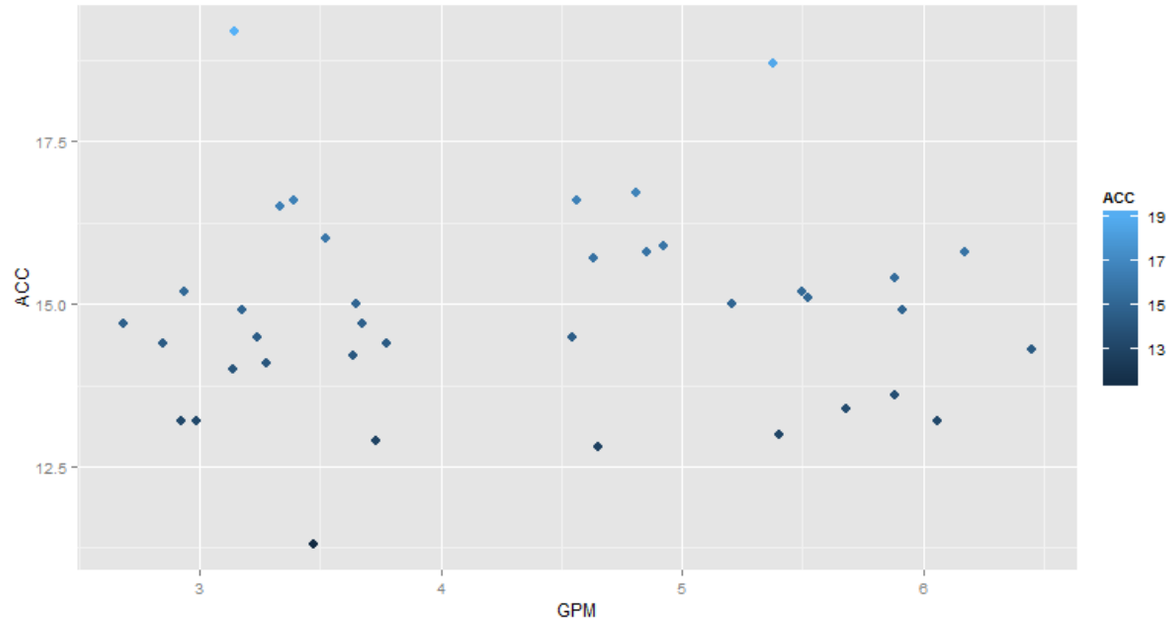
```
#For HP
ggplot(aes(x=GPM, HP), data=working.data) +
  geom_point(aes(color = HP), size=3)
```



These two too are positively corelated as well and with increase in horsepower of the car, the gallons per mile increases.

Sixth plot is of GPM with respect to ACC.

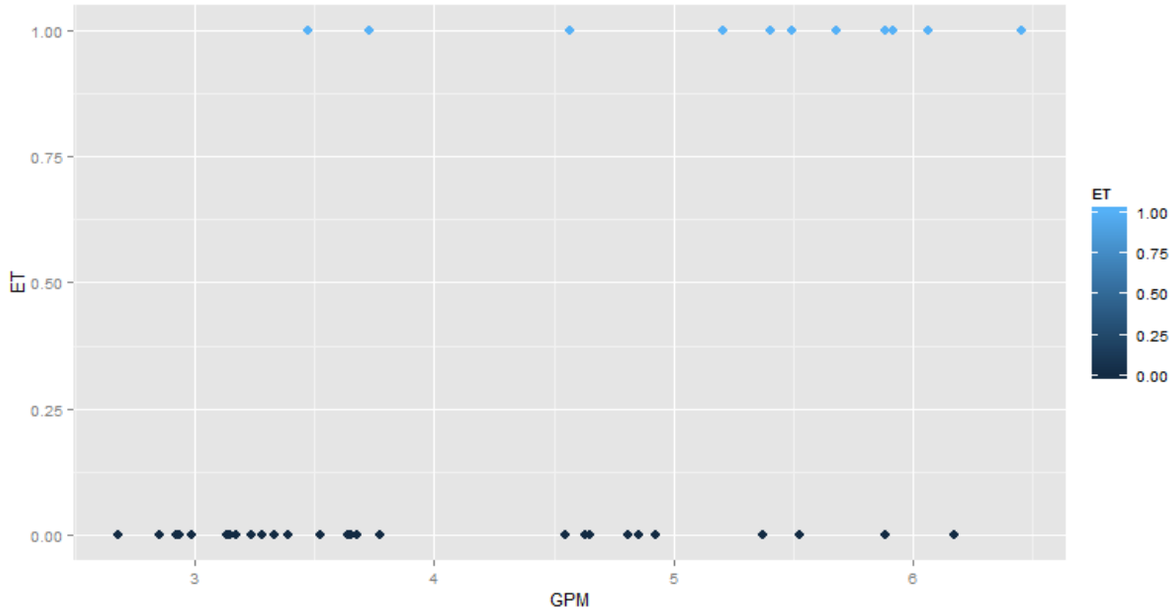
```
#For ACC
ggplot(aes(x=GPM, ACC), data=working.data) +
  geom_point(aes(color = ACC), size=3)
```



This plot shows that accelaration doesn't really have much significance on the gallon per mile since as we can see from the graph that most of the vehicles have accelaration in the range of 12.5 to 17.5 even when the GPM is large.

Our final plot is of GPM with respect to ET:

```
#For ET
ggplot(aes(x=GPM, ET), data=working.data) +
  geom_point(aes(color = ET), size=3)
```



This plot shows the clusters at two levels 0 and 1 which it itself explanatory and can be seen in the structure of data too.

3. Predictive analysis of GPM with random variables

We can firstly start with the analysis of the variation of means among the different variables so that we are able to see the significant variables and according to it we can provide the linear regression model.

```
## Predictive Analysis of GPM using ANNOVA
annova.model = aov(GPM ~ MPG + WT + DIS + NC + HP + ACC + ET)
summary(annova.model)
```

The above command gives us the following output

```
            Df Sum Sq Mean Sq   F value Pr(>F)
MPG          1  47.54   47.54 2.722e+31 <2e-16 ***
WT           1   0.44    0.44 2.521e+29 <2e-16 ***
DIS          1   0.00    0.00 2.712e+00  0.110
NC           1   0.00    0.00 3.300e-02  0.858
HP           1   0.00    0.00 4.800e-01  0.494
ACC          1   0.00    0.00 3.190e-01  0.577
ET           1   0.00    0.00 1.340e-01  0.717
Residuals   30   0.00    0.00
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From here, we can conclude that the most significant variables are MPG and WT. So, we will use these in modelling.

Now, let us model our GPM data using linear regression model.

```
## Here we can see that the most significant variable in the
## data set is MPG and the second most significant variable is WT

# Now let us do analysis using regression model with MPG and WT
# as significant variables

regression.model = lm(GPM ~ MPG + WT)
summary(regression.model)
```

The above commands gives us the following output

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.395e-15 -3.937e-16 -6.550e-17  3.430e-16  6.501e-15

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  6.722e+00  3.797e-15  1.770e+15   <2e-16 ***
MPG         -1.381e-01  7.578e-17 -1.822e+15   <2e-16 ***
WT           3.593e-01  7.019e-16  5.119e+14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.296e-15 on 35 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 1.428e+31 on 2 and 35 DF,  p-value: < 2.2e-16
```

This output shows that both are strongly significant variables.
Now, let us see the attributes of the summary of the above model to get the coefficients of MPG and WT, so that we can provide the final data model.

```
# Checking attributes of the regression model
attributes(regression.model)

# Exctracting the coefficients of the most signifact variables
for
# predictive model

regression.model$coefficients
```

The above code gives us the following output

```
(Intercept)         MPG          WT
     6.7219     -0.1381      0.3593
```

Thus, we can present our data model for MPG as

```
GPM = 0.3593*WT - 0.1381*MPG + 6.7219
```