# CASE: AUTOMOBILE DATA

#### Dataset

* The dataset we chose is the Automobile dataset from Kaggle.com.
* This is a set of data used to quantify the miles per gallon of different models of cars taking in account their cylinders, displacement, horsepower, weight, acceleration year and origin.
* This data has 398 observations and 9 variables
  + Response variable: MilesPerGallon, it is the miles per gallon of each of the automobile model given in the dataset.
* Regressors:
  + Cylinders: It is the number of cylinders an automobile model has.
  + Displacement: It is the displacement produced by each model of automobile.
  + Horse Power: It is the horsepower produced by each model of automobile
  + Weight: It is the weight of each model of automobile
  + Acceleration: It is the acceleration produced by each model of automobile.
  + Year: It is the year of manufacture of the automobiles
  + Origin: It is the country where the automobile was manufactured.
  + Name: It is the name of the model of the automobile.

Since HorsePower has 6 null values and is very less compared to the dataset observations, we are going to omit the rows which have null horsepower values.

Now our dataset has 392 rows

**Hypothesis:**

1. There exists a relationship between the dependent variable i.e. mpg and the independent variables i.e. cylinders, displacement, horsepower, weight and acceleration
2. There exist correlations between the independent variables.

**Data Summary:**

summary(auto)

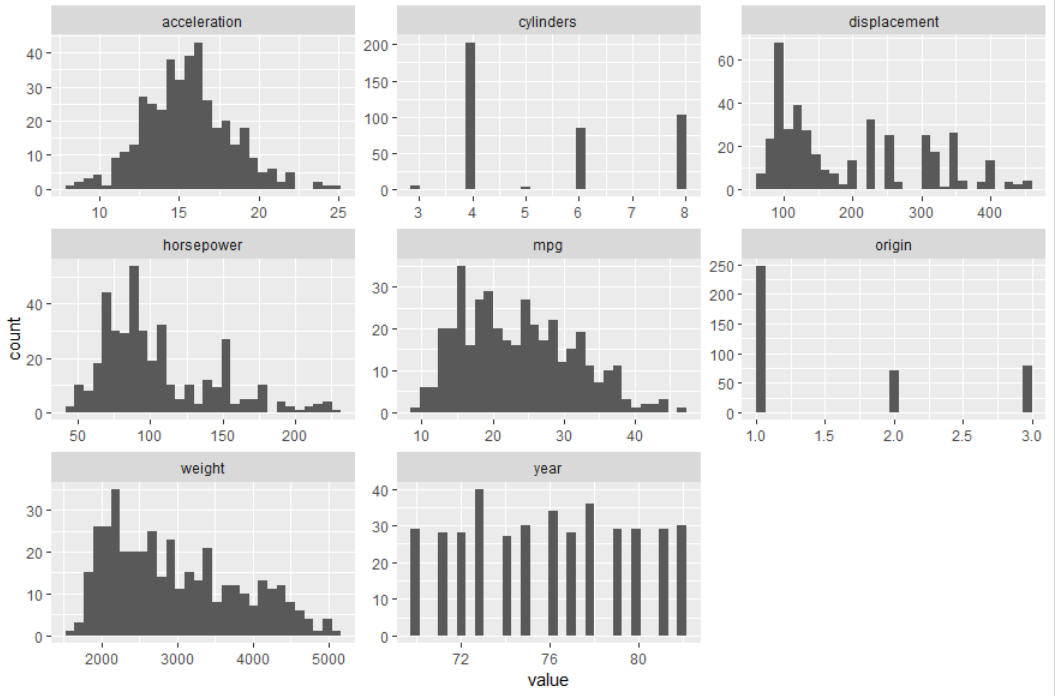
## mpg cylinders displacement horsepower   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Min. : 46.0   
## 1st Qu.:17.50 1st Qu.:4.000 1st Qu.:104.2 1st Qu.: 75.0   
## Median :23.00 Median :4.000 Median :148.5 Median : 93.5   
## Mean :23.51 Mean :5.455 Mean :193.4 Mean :104.5   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:262.0 3rd Qu.:126.0   
## Max. :46.60 Max. :8.000 Max. :455.0 Max. :230.0   
## NA's :6   
## weight acceleration year origin   
## Min. :1613 Min. : 8.00 Min. :70.00 Min. :1.000   
## 1st Qu.:2224 1st Qu.:13.82 1st Qu.:73.00 1st Qu.:1.000   
## Median :2804 Median :15.50 Median :76.00 Median :1.000   
## Mean :2970 Mean :15.57 Mean :76.01 Mean :1.573   
## 3rd Qu.:3608 3rd Qu.:17.18 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :5140 Max. :24.80 Max. :82.00 Max. :3.000   
##   
## name   
## ford pinto : 6   
## amc matador : 5   
## ford maverick : 5   
## toyota corolla: 5   
## amc gremlin : 4   
## amc hornet : 4   
## (Other) :369

In the above summary we can see that mpg has a minimum value of 9 and maximum of 46 with a mean of 23.51. Displacement has a minimum value of 68 and maximum of 455 with a mean of 193.4. Horse power has a minimum value of 46 and maximum value of 230 with a mean value of 104.5. Weight has minimum value of 1613 and maximum value of 5140 with a mean of 2970. Acceleration has a minimum value of 8 and maximum of 24.8 with a mean of 15.57. It seems like year, origin and name are categorical variables which can be verified using their histogram plots.

#### Data Visualization

**Univariate plots:**

Histograms:



The distributions for MPG, displacement, horsepower, and weight are all right skewed–a longer tail toward the higher end of the scale. And there are many more four-cylinder cars than six- or eight-cylinder cars. This supports our hypothesis that there is a strong correlation between these variables.

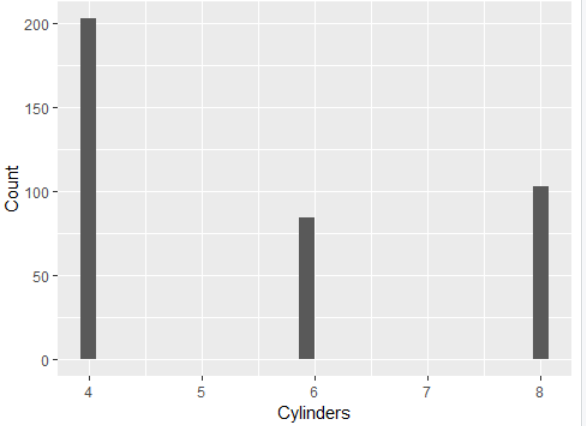
Acceleration has a normal distribution, so this variable has less correlation with other variables. The count for each year is almost same across the sample.

table(auto$cylinders)



Based on the relatively tiny counts of 3 and 5 cylinders, these were removed in the later plots. There are many automobiles with 4 cylinders in the sample as compared to other number of cylinders.

Plot after dropping 3 and 5 cylinders:



table(auto$origin)

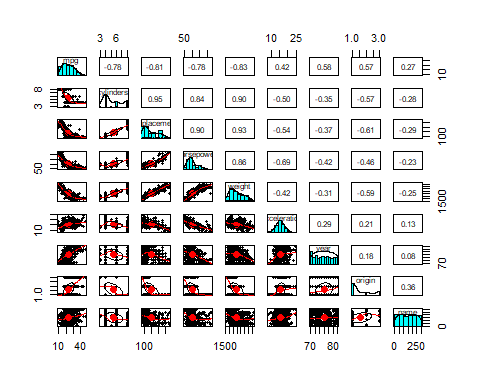


There are more than 3 times as many automobiles from country 1 as compared to 2 and 3.

**Correlation plots:**

We can visualize the relationship plots of all the variable through the following plot.

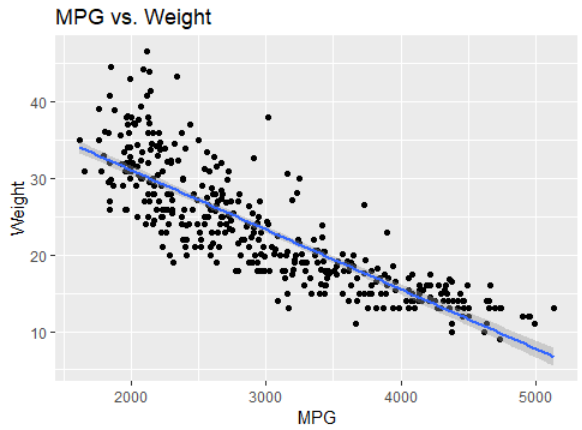
pairs.panels(auto)

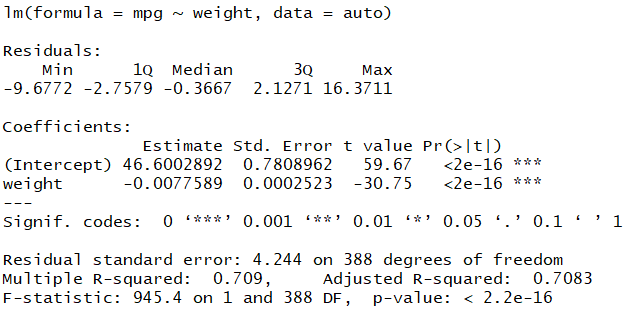


We can observe in the above visualization that horse-weight, horsepower-displacement, weight-displacement have a strong correlation. Thus, this data supports our hypothesis that these variables have correlation. Mpg has highest correlation with weight followed by horsepower and displacement. Acceleration has less correlation with weight, displacement and horsepower.

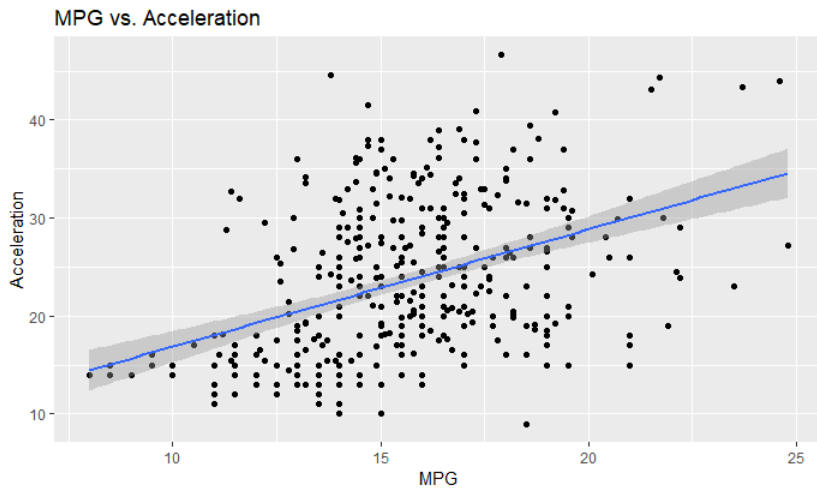
#### Analysis: Bivariate and Multivariate

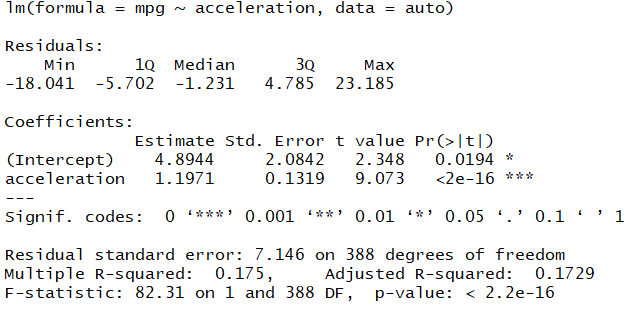
* **Bivariate plots:**



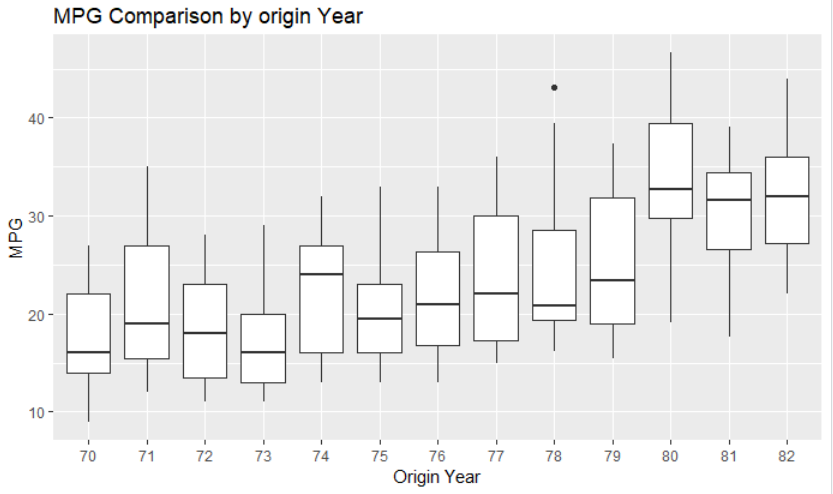


The above scatter plot fitted with a linear line shows that mpg and weight are inversely related i.e. as weight of the automobile increases the mpg decreases. The adjusted R2 is around 70%   
Lets check the relationship between mpg and acceleration now:

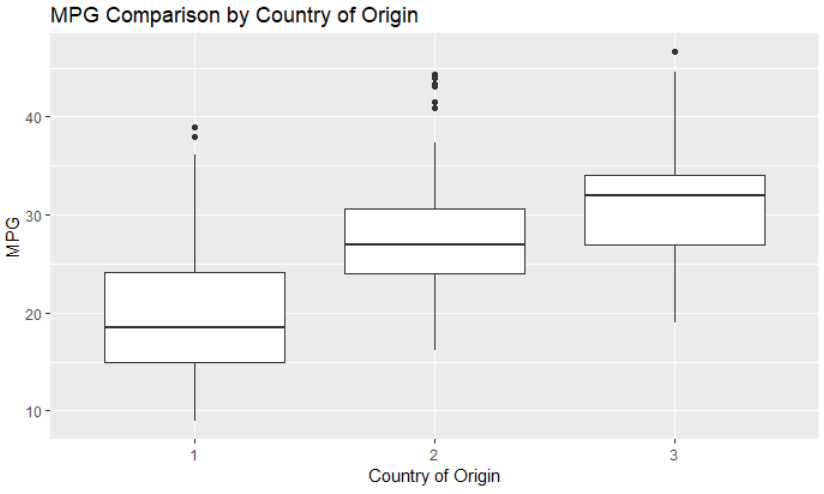




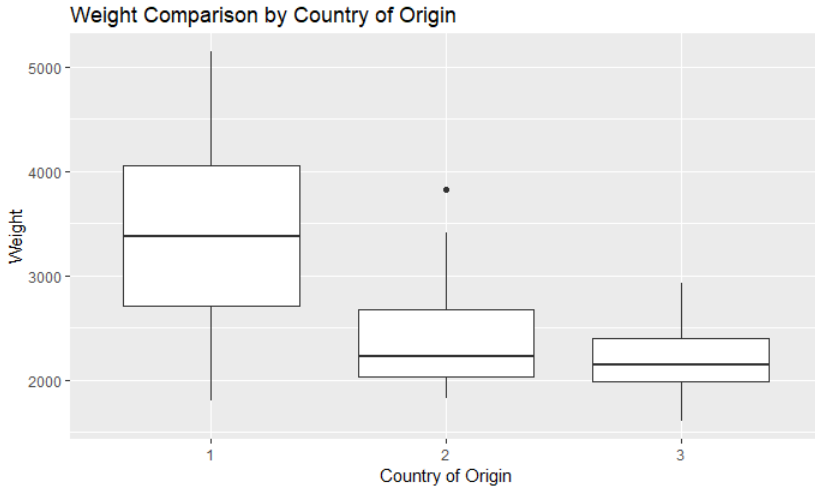
This shows that mpg and acceleration have a positive linear relationship. The adjusted R2 was found to be around 17% which is not very high.



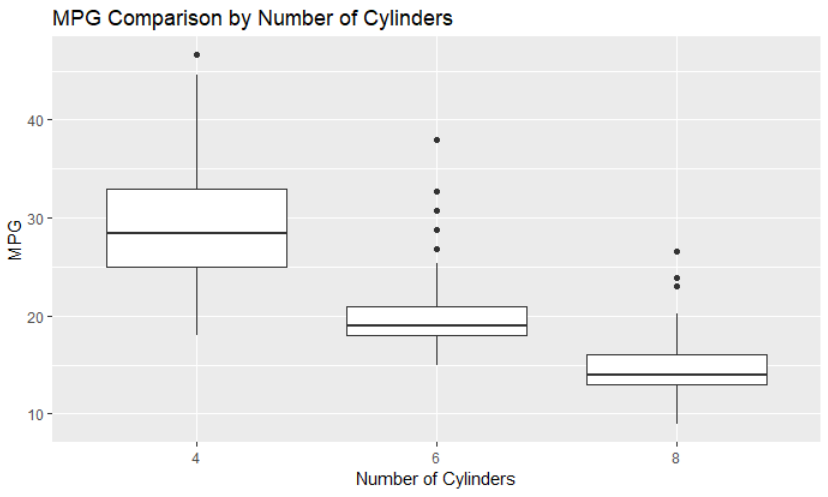
The plot above shows that there has been a significant increase in mpg from the early 70s to early 80s.



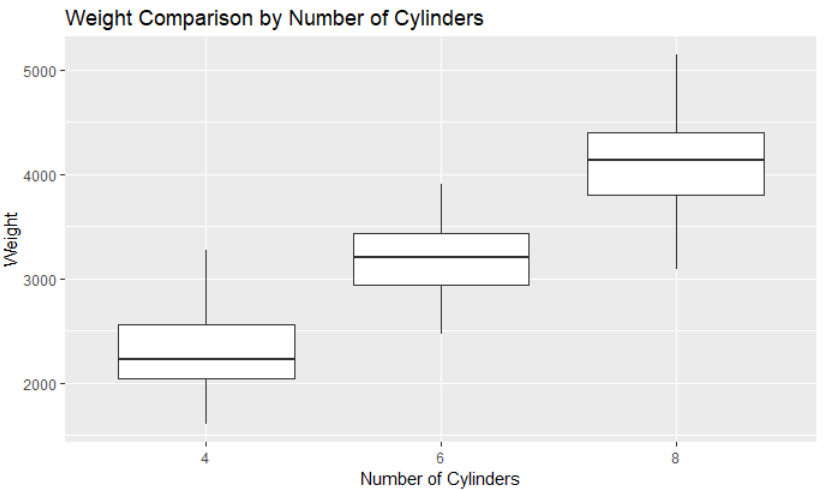
The above plot shows that the mpg of autos from country 1 is lesser compared to country 2 and 3. This is could be because of mpg’s inverse relation with weight and country 1 autos weigh more. To verify this, we can plot another box plot for weight and origin.



Here we can see that weight of autos in country 1 are indeed much higher than country 2 and 3.



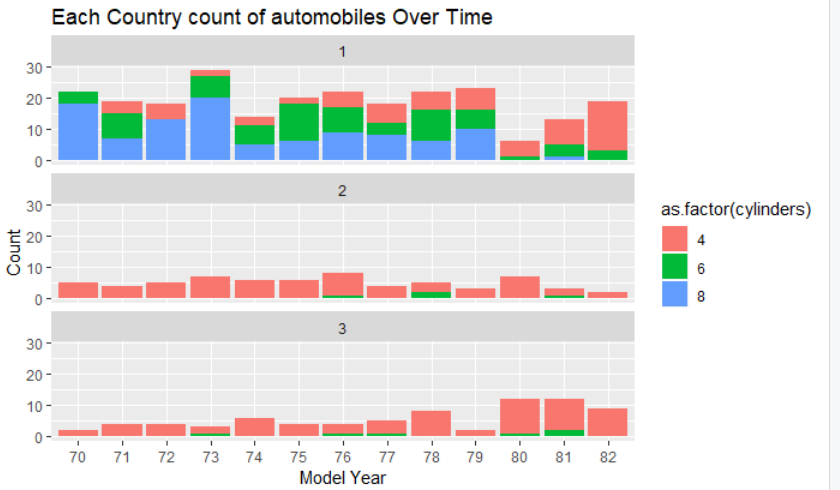
More number of cylinders causes the mpg to decrease. Thus, they are inversely related. This may be because a greater number of cylinders causes the weight to increase and hence mpg will decrease. This can also be verified using another boxplot.



The above plot verifies that autos with 4 cylinders weight less thus having higher mpg.

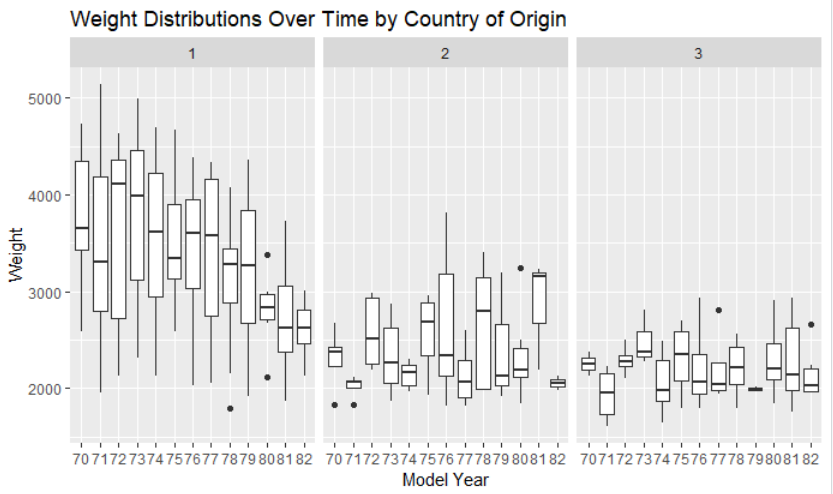
Adding all the above information, it seems that country 1 is manufacturing cars with more number of cylinders, which increases their weight and thus decreases their mpg. While country 2 and 3 autos have higher mpg.

Multivariate Plots:



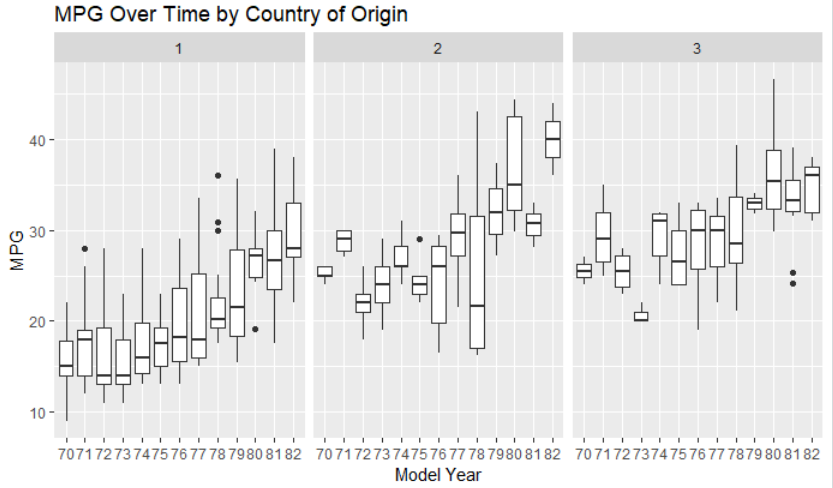
In the above plot, we can see that for country 1, the number of automobiles with 4 cylinders increase over time and 4 and 6 cylinders automobiles decrease over time. While for country 2 and 3, there are mostly automobiles with 4 number of cylinders and very few 6 cylinders automobiles manufactured in few years.

Similarly, we can check the weight distribution:



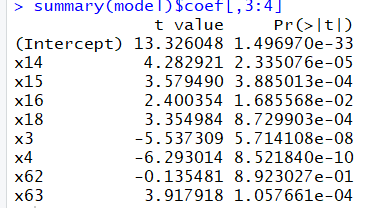
In the above plot we can see that, for country 1, the weight of the automobiles has significantly decreased from 70s to 80s while country 2 and 3 have manufactured less weighted automobiles since earlier years.

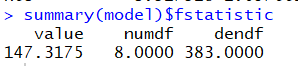
Next, we plot mpg over time:



In the above plot we can see that mpg of automobiles in country 1 has drastically increased but for country 2 and 3 it has remained somewhat same until 78 and then increased from there.

The above relationships can also be verified by t-test and f-test:





The automobiles with 4, 5, 6 and 8 cylinders, weight and acceleration are significant based on their p-values and f-statistic value.

**Conclusion:**

Weight of automobiles are the strongest determinant of their mpg. The automobiles having 4 cylinders have least weight and with 8 cylinders have highest weight. The increase in weight causes the mpg to decrease. The mpg of the automobiles has increased from 70s to 80s. Country 2 and 3 have been manufacturing 4 cylinders automobiles since the early 70s whereas country 1 started to mimic this feature later after 80s. With these observations we saw that cylinders, weight and mpg have a strong relationship i.e this supports our first hypothesis that these covariates have strong relationship with the response variable. We also saw that weight and cylinders have a strong relationship which supports our second hypothesis that the covariates are also related among one another.