Northeastern University
College of Professional Studies

Module 6 Report

ALY6010 - Probability Theory and Introductory Statistics

Instructor: Prof. Amin Karimpour

Date: 17 December 2021

By,

Jeseeka Shah

Module 6

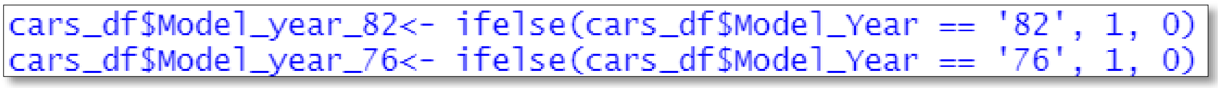
Introduction

The dataset that I am using here is about small cars . The dataset is taken from Kaggle. It has 100 rows and 11 variables. Here, I am going to examine relationships between variables. Making dummy variables to subset the dataset and re-running the regression analysis. Understanding the impact of categorical variables on the regression. After which making separate regression line for each subset and carrying out multiple regression analysis on this subset.

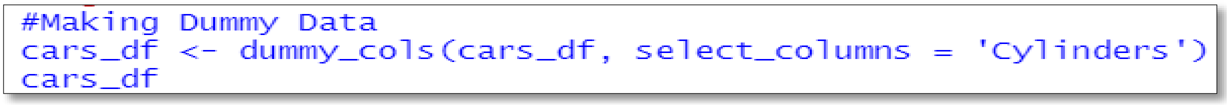
Question -create dummy variables to subset your dataset ?

Creating a subset

I have considered on making four subsets here. Initially , the variable that I have looked up at is cylinders and model year of the cars. There are three model year shown here 70,76 and 82 and three type of cars with cylinders 4,5, and 6. In total I have four dummy variables considering the formula N-1(where N is the total number of categorical values present in that columns).



```
cars_df$Model_year_82<- ifelse(cars_df$Model_Year == '82', 1, 0)
cars_df$Model_year_76<- ifelse(cars_df$Model_Year == '76', 1, 0)
```

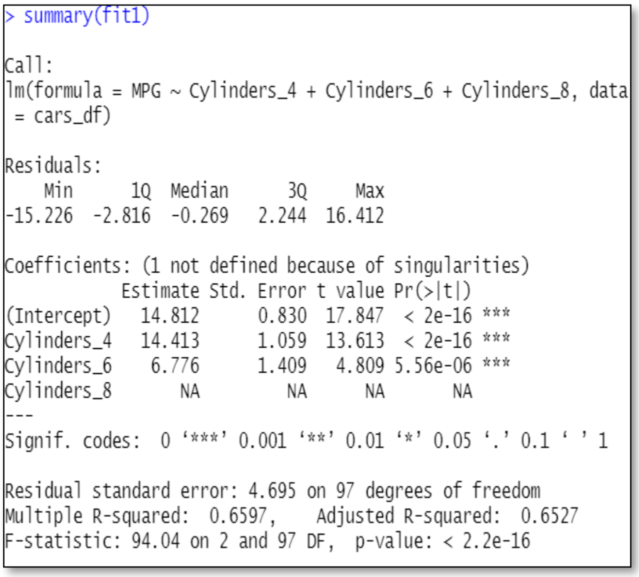


```
#Making Dummy Data
cars_df <- dummy_cols(cars_df, select_columns = 'Cylinders')
cars_df
```

Question- How does this impact your understanding of the impact of the categorical variable on the regression?

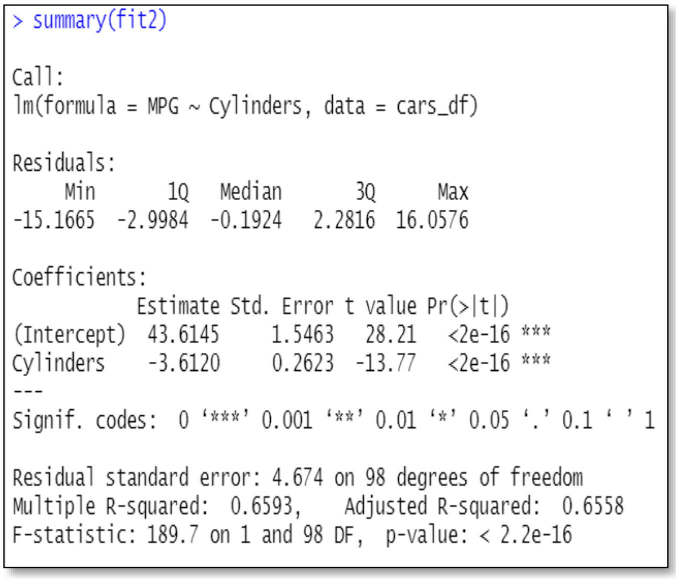
Regression Analysis

After creating the dummy variables, I have checked the regression of them individual and all together. The result shows that there is not much difference in the result. But the first method is preferred more as the information is clearer in fig 1.a than combined information shown in fig 1.b.



```
> summary(fit1)

Call:
lm(formula = MPG ~ Cylinders_4 + Cylinders_6 + Cylinders_8, data
 = cars_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.226  -2.816  -0.269   2.244  16.412 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   14.812      0.830  17.847  < 2e-16 ***
Cylinders_4   14.413      1.059  13.613  < 2e-16 ***
Cylinders_6    6.776      1.409   4.809 5.56e-06 ***
Cylinders_8       NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.695 on 97 degrees of freedom
Multiple R-squared:  0.6597,	Adjusted R-squared:  0.6527 
F-statistic: 94.04 on 2 and 97 DF,  p-value: < 2.2e-16
```

Fig 1.a : Adjusted R-sq. with individual cylinders



```
> summary(fit2)

Call:
lm(formula = MPG ~ Cylinders, data = cars_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.1665  -2.9984  -0.1924   2.2816  16.0576 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  43.6145     1.5463   28.21   <2e-16 ***
Cylinders    -3.6120     0.2623  -13.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.674 on 98 degrees of freedom
Multiple R-squared:  0.6593,	Adjusted R-squared:  0.6558 
F-statistic: 189.7 on 1 and 98 DF,  p-value: < 2.2e-16
```

Fig 1.b Adjusted R-sq. with total cylinders

Question - Create separate regression lines for each subset. How do these regression lines differ from the regression lines ? How does this method of looking at the data impact your understanding of the data?

Module 6

Plot 1: Miles per gallon Vs. Horsepower With respect to cylinders

The below figure shows that negative linear relationship between the two variables MPG and Horsepower in accordance with cylinders. The regression plot (a) does not clearly show the bifurcation of the impact of different cylinder on the relationship of mile per gas and the horsepower of the car. Whereas, the figure (b), indicates the impact of each cylinder on the MPG and horsepower. It shows that as the horsepower increases and number of cylinders also increases then the relationship is steeper in negative direction.

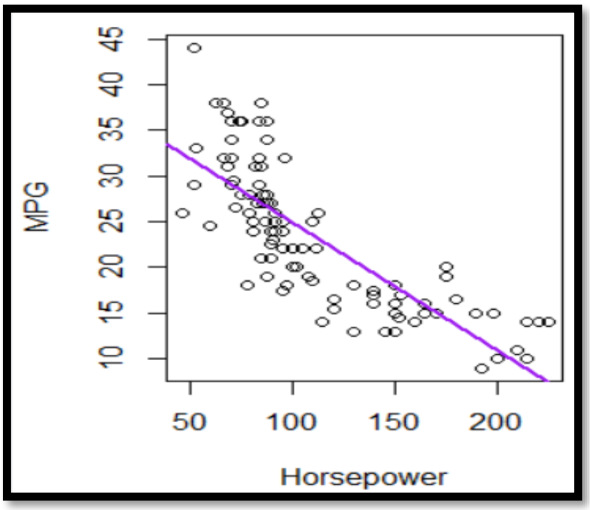


Fig 2.a : Horsepower Vs. MPG w.r.t Cylinders

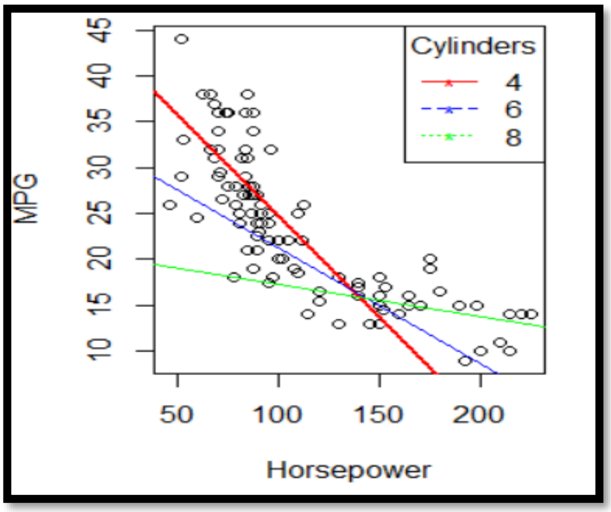


Fig 2.b : Horsepower Vs. MPG w.r.t different Cylinders

Plot 2: Displacement against Horsepower with Cylinders

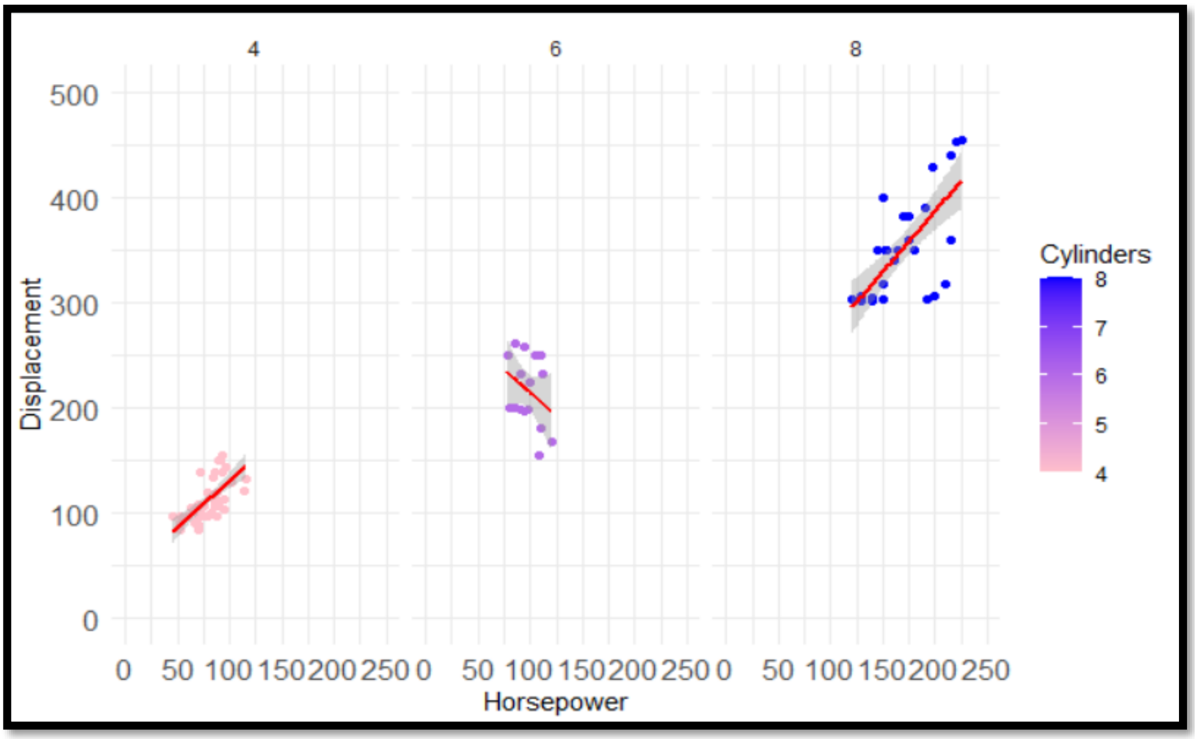
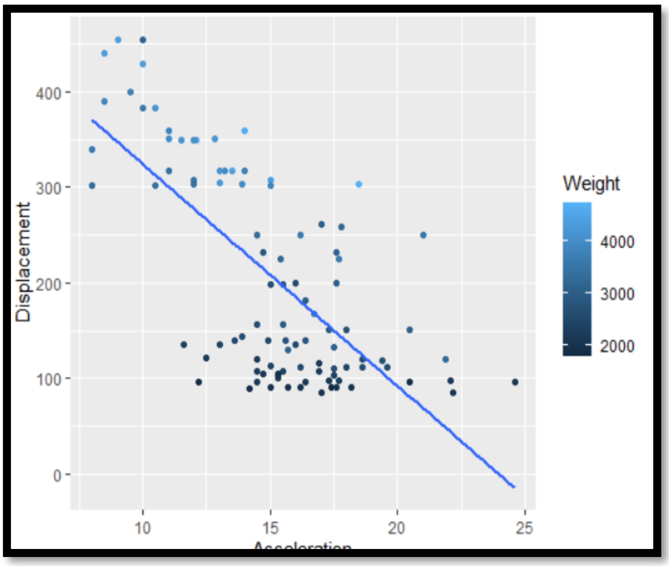


Fig 3. Displacement Vs. Horsepower w.r.t cylinders 4,5 & 6

The Fig 3 shows that displacement verses acceleration in an increase order for the cylinder 4 and 8. More the cylinders the car will be able to catch up on more distance as the horsepower will be more. This is clear seen for graph with eight cylinders.

Plot 3 : Displacement Vs. Acceleration with respect to weight

The relationship between displacement and acceleration with respect to weight. As the weight increases the displacement decrease even if the acceleration is more. The figure 4 a shows only negative linearity rather but figure 4.b gives us more information based on the different range of weights of the cars.



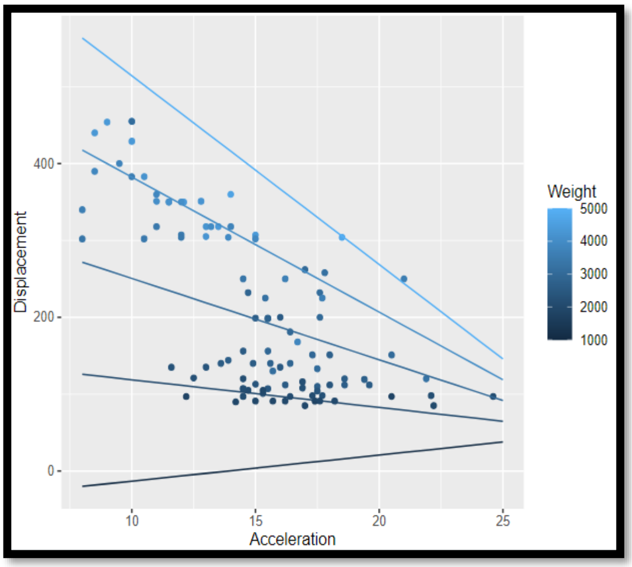


Fig 4.a Displacement Vs. Acceleration w.r.t weight   Fig 4.b Displacement vs. Acceleration w.r.t different weight

Plot 4: Model year Vs. Mile per gallon

As the model evolves per year the miles covered per gallon increase. Therefore, increasing the efficience every year of the car. Below the graph shows the same. The make of 82 has the most coverage mile per gas compared to others.

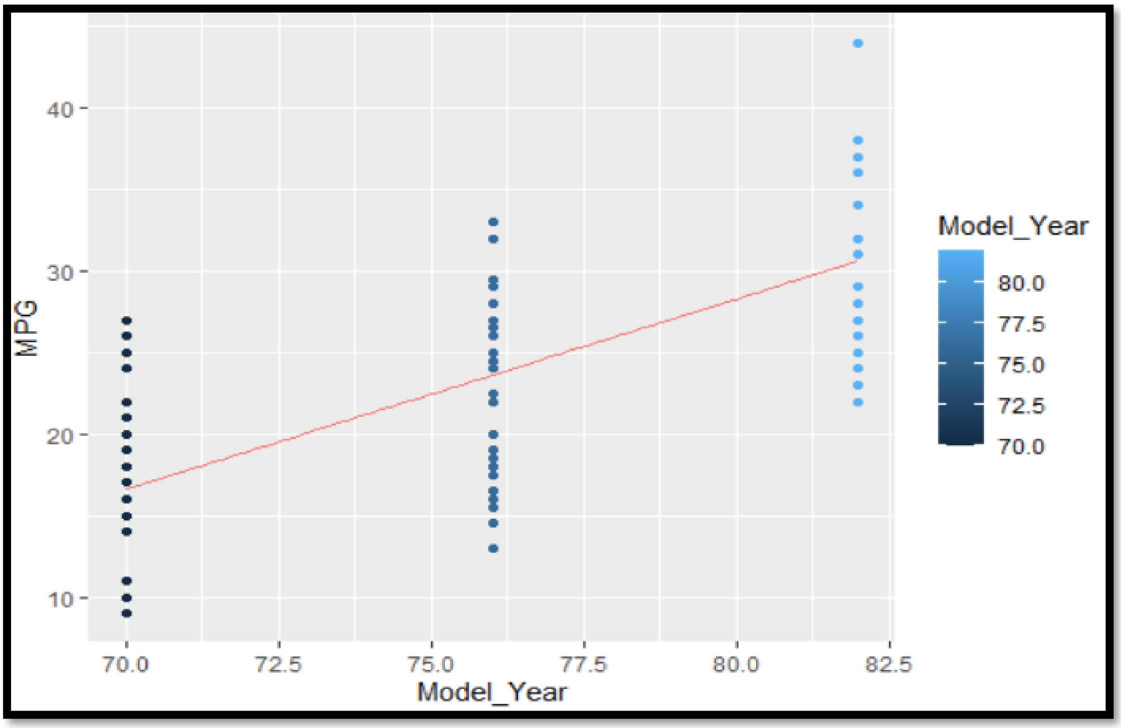
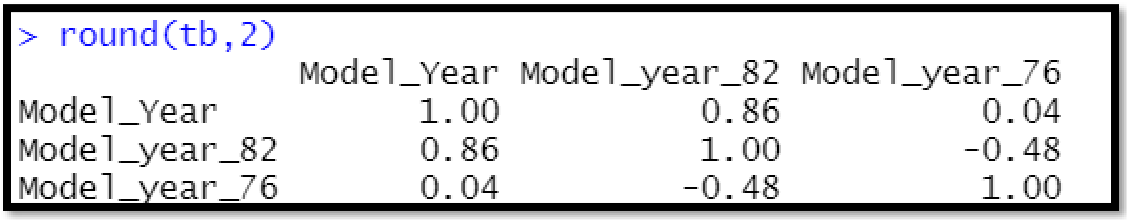


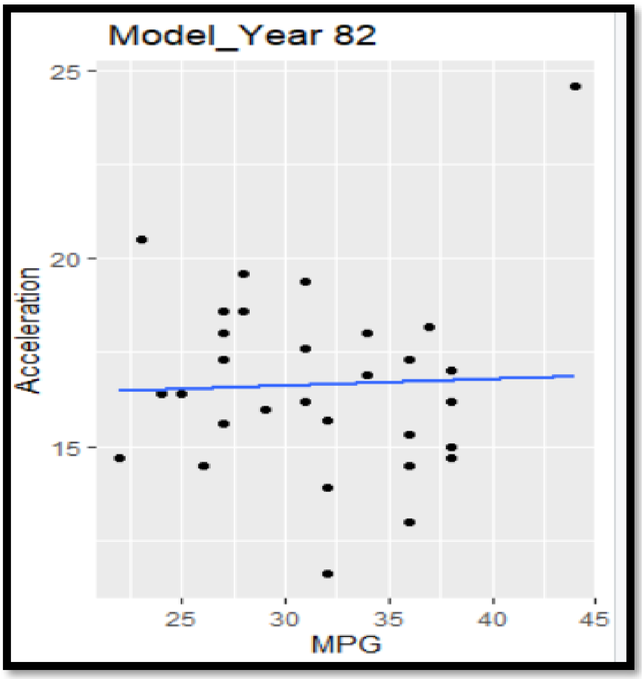
Fig 5 .the year of make with miles per gas



```
> round(tb,2)
              Model_Year Model_year_82 Model_year_76
Model_Year          1.00          0.86          0.04
Model_year_82       0.86          1.00         -0.48
Model_year_76       0.04         -0.48          1.00
```

Fig 6: S the correlation between the different the Model_year

The correlation shown in the figure 6 shows the relation between the subset made with the original variable. The most significant relation that impacts the original variable model_year is of model made in the year 82 which has 86 percent significance. Therefore, I have considered the model year 82 for further analysis . The figure 7 a. shows the relationship based on MPG and acceleration with respect to cyclinders but then figure 7.b gives better understanding with repect to mpg and acceleration.



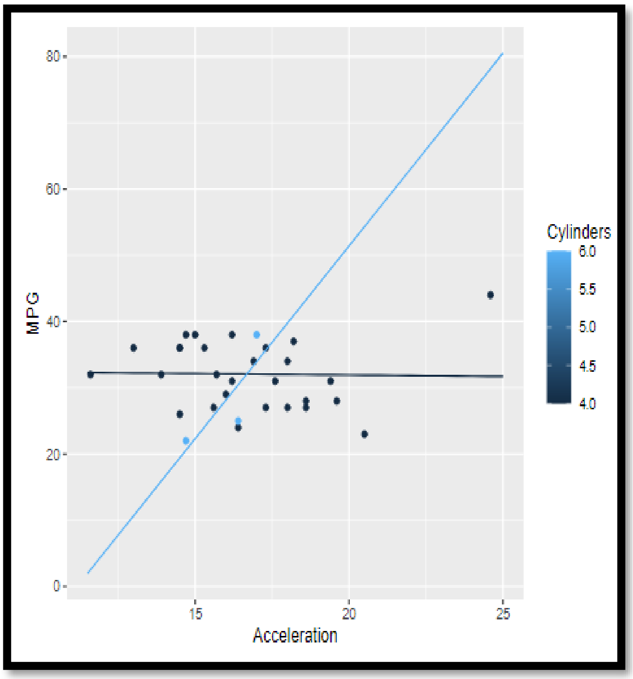


Fig 7.a : Mpg vs Acceleration with cyclinders Fig 7b, Mpg vs Acceleration with different cylinders

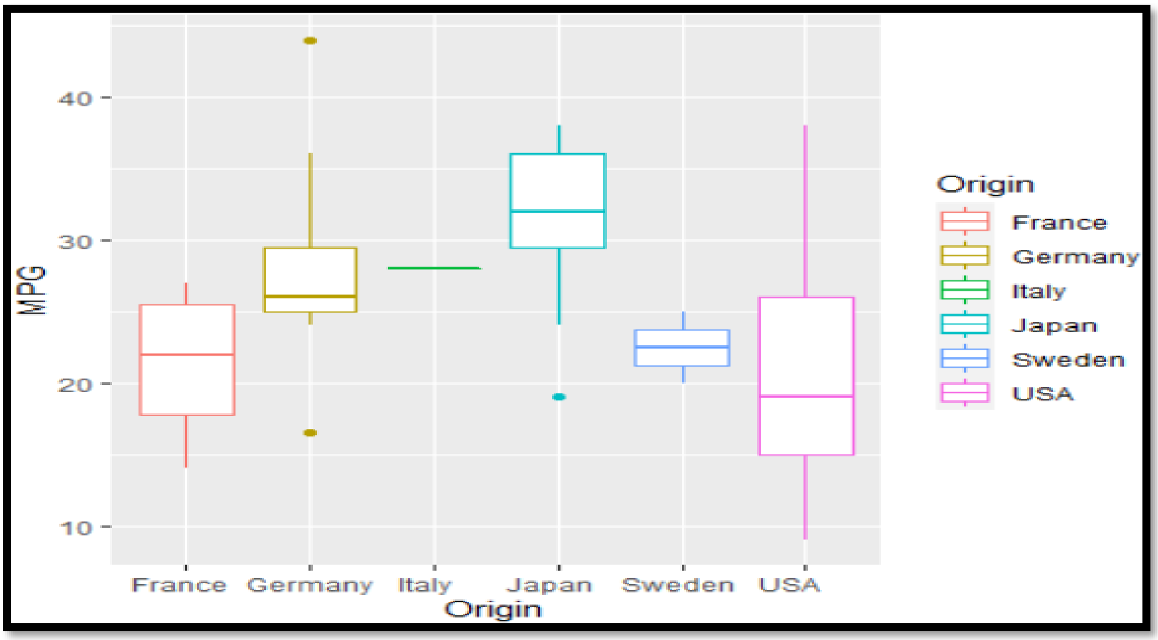


Fig 8 : Boxplot showing Origin Vs. Miles per gas

The figure 8 shows that germany has the highest miles per gas coverage but the outlier markes the highest and the lowest. Whereas the actual distribution is gives miles per gas less than 30. Therefore, Japan has the best miles per gas coverage followed by the USA. Hence, for further analysis I have considered looking at this origins.

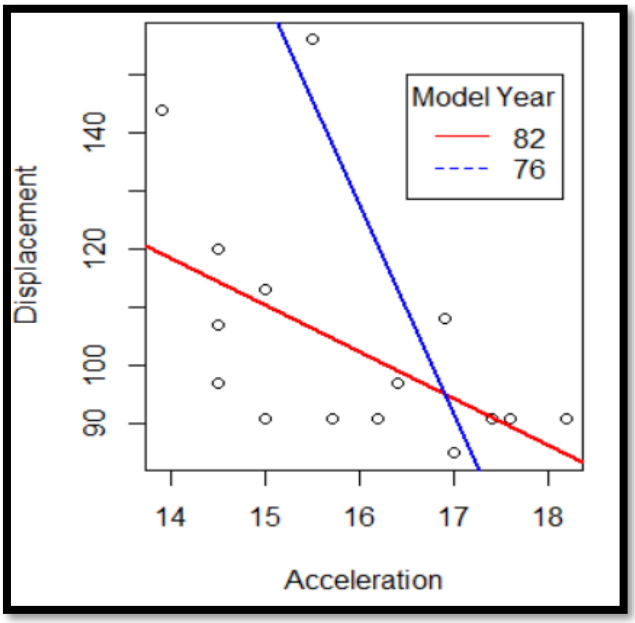


Fig 9.a Displacement Vs. Acceleration for Japan

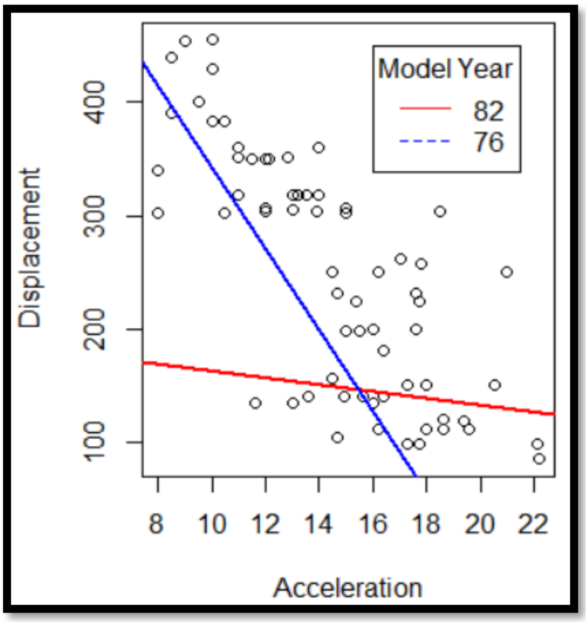


Fig 9.b Displacement Vs Acceleration for USA

The Figure 9 a shows the relation for model year 82 and 76 for the origin Japan in accordance of displacement and acceleration. Figure 9.b is shows same for the origin USA. The country USA has better steeper slope compared to Japan for the similar acceleration range.

Summary

In this last module,I have considered to make four subset on bases of model year and cyclinders. And have performed the subset analysis to derive different relationship between the rest of the variables present. The result should the cylinder 8 has the best displacement covered with respect to horsepower. The car model manufactured in the year 82 is most efficient with respect to miles per gas.Then after finding the model year then I have considered for checking the origin with the most efficient fuel usage. The result pointed at Japan & USA . Lastly, I have plotted the graph to show the relaionship of displacement and acceleration for the countries Japan and USA for different model years.

Reference

- Coder, R. (2020, December 17). *Add legend to a plot in R*. R CODER. https://r-coder.com/add-legend-r/#:%7E:text=%2C%20lwd%20%3D%202)-,Change%20legend%20size,smaller%20legends%20than%20the%20default.
- *Subsetting Data in R - Lab*. (n.d.). Subsetting. https://johnmuschelli.com/intro_to_r/Subsetting_Data_in_R/lab/Subsetting_Data_in_R_Lab_Key.html
- Long, J. (2021, July 2). *Exploring interactions with continuous predictors in regression models*. Continous Prediction. https://cran.r-project.org/web/packages/interactions/vignettes/interactions.html
- *ggplot2 box plot : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA*. (n.d.). Ggplot. http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization