ABDULLAH GUL UNIVERSITY

IE222

STATISTICS

TERM PROJECT

INSTRUCTOR

Uğur SATIÇ

GROUP MEMBERS

Fatma Nur BURGU

Sunay İlayda ÇAMBEL

Halef ÇOBAN

Hüseyin ALTINSOY

June 31,2022

TABLE OF CONTENT

LIST OF FIGURES

ABSTRACT

In this project, various factors such as engine size, cylinders, fuel type, fuel consumption, CO2 emissions, CO2 rating, and smog rate, which we think may affect the fuel consumption amounts of vehicles in the 2022 model year, are a data set containing 946 data for each column of the R program. analyzed with the help of the distribution of the data set was determined by Histogram plots. The direction and intensity of the effects of factors on fuel consumption were determined using t-test, core-test, and regression models. As a result, it turns out that it affects the engine size and the amount of fuel in the cylinder. In addition, fuel consumption affects the CO2 rate and the smog rate.

INTRODUCTION

Automotive production has been advancing rapidly for a very long time. Cars are systems that work without human power. Thanks to the fuel they use, cars provide comfort and time to people, as well as a great convenience. However, fuel consumption in vehicles now harms the environment and forces people financially. This project is a discussion and study on the fuel consumption of vehicles. The main topic of this project will be to address the fuel consumption of a vehicle by looking at the characteristics of vehicles such as engine size, cylinders, fuel type, fuel consumption, CO2 emissions, CO2 rating, and smog rating.

The aim of this project is to address the factors affecting fuel consumption in vehicles. Thanks to this study, it can be seen how variables (engine size, cylinders, transmissions, fuel type, etc.) are related to fuel consumption and how they affect fuel consumption. In this study, a method with a multiple linear regression model for fuel consumption in vehicles will be used based on statistical data.

Our aim to determine the possible effects of given variables on the fuel consumption. In order to answer the following questions, the hypotheses included in the proposal were asked according to the determining factors.

1-How does the number of cylinders in the vehicles affect the vehicle's fuel consumption?

2-How does engine size affect the fuel consumption of vehicles?

3-Is the fuel consumption of vehicles with a smog rate of less than 5 less than that of vehicles with a smog rate of more than 5?

4-How can the CO2 emissions of vehicles with an equal number of cylinders affect the fuel consumption of the vehicles?

Hypothesis:

H0: All factors are effective on the fuel consumption

H1: All factors are **not** effective on the fuel consumption

OBJECTIVE OF THE PROJECT

In this project, a decision mechanism will be created by using statistical tools for the fuel consumption amounts of the vehicles, thus minimizing the fuel consumption amounts of the vehicles. Our motivation is for the project to guide those who want to buy a vehicle, to ensure that vehicle owners minimize fuel consumption and minimize the damage of fuels to the environment. Otherwise, people hesitate to buy a vehicle due to high fuel consumption, and vehicle owners have financial difficulties due to high fuel consumption. In addition, our world is negatively affected by excessive fuel consumption. With this project, it is desired to make a statistical data interpretation that can be used by vehicle owners.

For this reason, it is aimed to estimate fuel consumption by examining different vehicle types including various information such as engine size, cylinders, fuel type, fuel consumption, CO2 emissions, CO2 rating, and smog rate.

COLLECTING THE DATA

A data set containing 947 rows of data with different vehicle characteristics such as engine size, cylinders, fuel type, fuel consumption, CO2 emissions, CO2 rating, and smog rating was found. These data include the fuel consumption amounts in L per 100 km of different brands of cars with the model year 2022.

FILTERING THE DATA

For use in the study, the data set was separated from the vehicle characteristics to include the fuel consumption per 100 km on highways of different brands of vehicles with the same model year, and engine size, cylinders, fuel consumption, CO2 emissions, CO2 rating, and smoke rate. The data set has been improved to make it more suitable for the study.

IDENTIFYING THE TOOLS

As a result of some research on the most efficient tools used in statistical analysis, it was thought that R would be very useful for analyzing data because of the various functions and packages it includes. The analysis was carried out by installing the readxl package in R, which will enable it to read the data set in the Excel file.

LITERATURE REVIEW

Although there is a lot of literature on multiple regression analysis,there is no literature that directly examines car models , features and fuel consumption.However, articles with different regressions on fuel consumption were examined. For example, Jereb et al. (2018) investigated the effect of traffic flow on fuel consumption in a real-world example of a road section in Celje. The aim of this study was to determine the causes of road pollution and the relationship between fuel consumption and road pollution.This study, which examines the regression of fuel consumption and pollution, was used as a guide as it may be similar to our regression model.

APPLIYING DESCRIPTIVE STATISTICS

To answer the questions we asked some descriptive statistics methods were needed to use.

To summary the data, analyze the data, and creating plots and histograms as offered in the course statistical analysis program R used.

The summary of the data is below. As it can be seen there is 7 variables included in our data. To find t-scores, the correlations between variables and for linear regression used R codes and results are given below.

```
> summary(HgwDATA)
  Engine_Size      Cylinders      CO2_Emissions    Smog_Rating     CO2_Rating       Fuel_Type
 Min.   :1.200   Min.   : 3.000   Min.   : 94.0   Min.   :1.00   Min.   : 1.000   Min.   : 4.00
 1st Qu.:2.000   1st Qu.: 4.000   1st Qu.:213.2   1st Qu.:3.00   1st Qu.: 3.000   1st Qu.:24.00
 Median :3.000   Median : 6.000   Median :257.0   Median :5.00   Median : 5.000   Median :26.00
 Mean   :3.199   Mean   : 5.668   Mean   :259.2   Mean   :4.95   Mean   : 4.539   Mean   :24.15
 3rd Qu.:3.800   3rd Qu.: 6.000   3rd Qu.:300.8   3rd Qu.:6.00   3rd Qu.: 5.000   3rd Qu.:26.00
 Max.   :8.000   Max.   :16.000   Max.   :608.0   Max.   :7.00   Max.   :10.000   Max.   :26.00
 Fuel_ConsumptionHGW
 Min.   : 3.900
 1st Qu.: 7.700
 Median : 9.200
 Mean   : 9.363
 3rd Qu.:10.700
 Max.   :20.900
>
```

Figure 1 Summary of the data

First, the main step is that the dataset must be analysed to introduce the variables and elements of descriptive statistics in this investigation. Minimum and maximum values of each column in the table are demonstrated, the values of quartiles such as 1st, median, 3rd and 4th quartiles, mean. According to the summary table above, the minimum engine size is 1.2, the maximum engine size is 8, and the mean engine size is 3.199. In addition to these, the median value for engine size is 3. The second category is the cylinders of cars. According to the summary, it is seen that the maximum cylinder is 16. It shows that there are cars which have certainly powerful engines. Moreover, when it is concerned that CO2 emissions have an important effect on the environment because the maximum CO2 emission is 608 in some cars.

The fourth category is this rating that reflects vehicle exhaust emissions that contribute to local and regional air pollution and create problems such as smoke, haze, and health problems. Vehicles with 10 points are the cleanest vehicles. According to the summary, Fuel

types are numbered because of the regression. X or 24 = Regular gasoline; Z or 26 = Premium gasoline; D or 4 = Diesel; E or 5 =Electricity. In conclusion, Fuel_ConsumptionHGW means the consumption on the highway of vehicles. Moreover, data is a term of a litre per 100 kilometres. The maximum is 20 litres in 100 km and the minimum litre is 3.9 in one kilometre.

ANALYZING THE DATA

The behavior patterns of the data we obtained and the relationship between them will be examined in this section. Histograms were used to learn data distributions and distribution types to know how to proceed in the future and to identify tests that should be used.

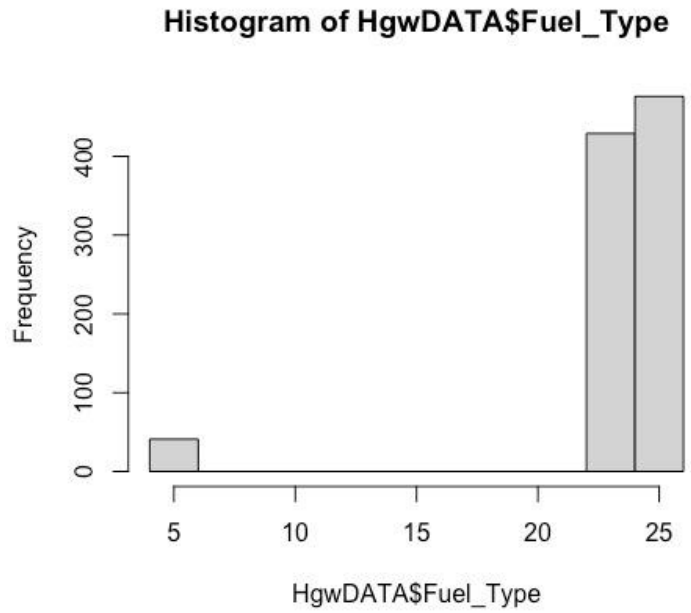


Figure 2 Histogram of Fuel Type

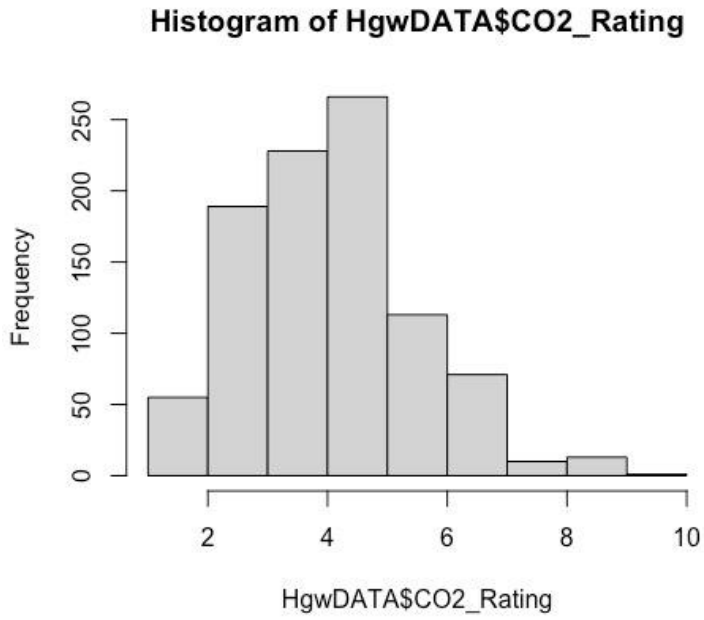


Figure 3 Histogram of CO2 Rating

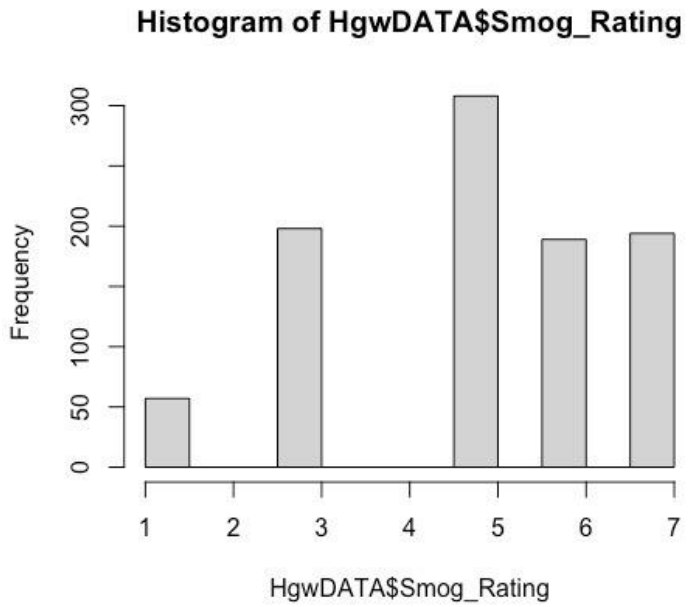


Figure 4 Histogram of Smog Rating

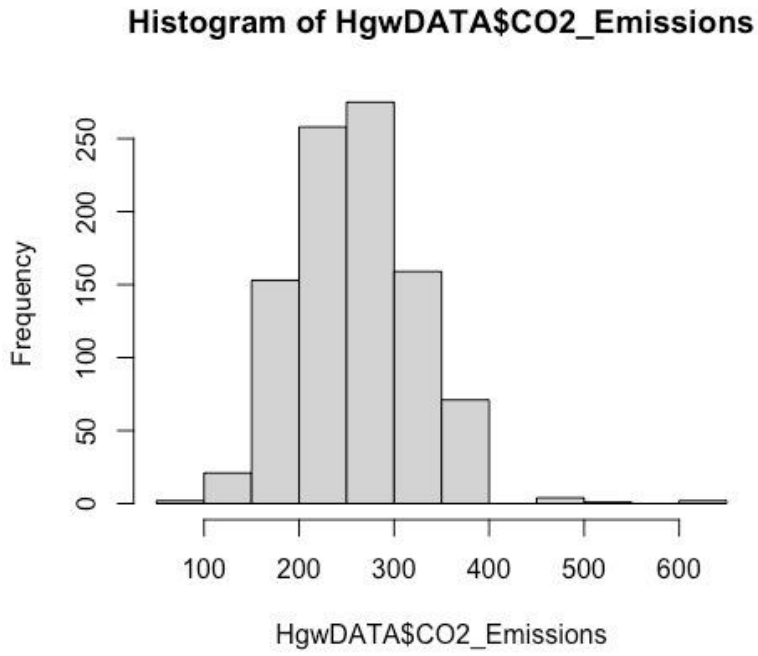


Figure 5 Histogram of CO2 Emissions

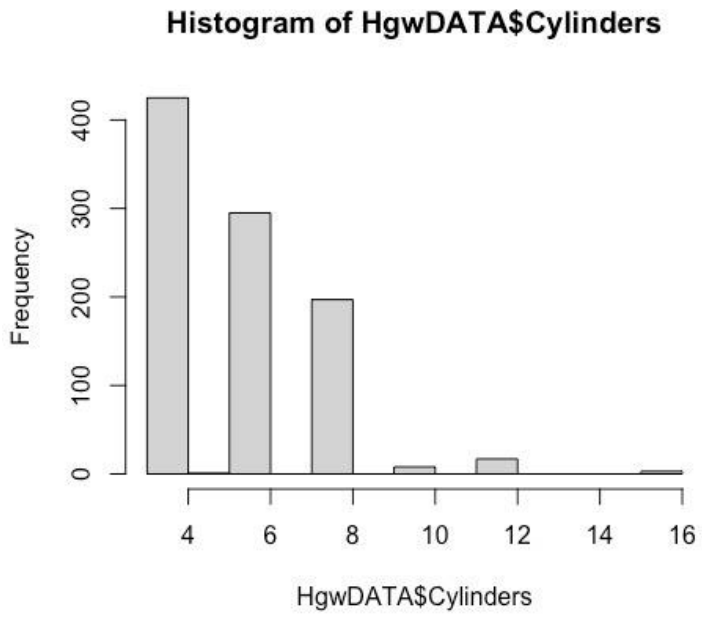


Figure 6 Histogram of Cylinders

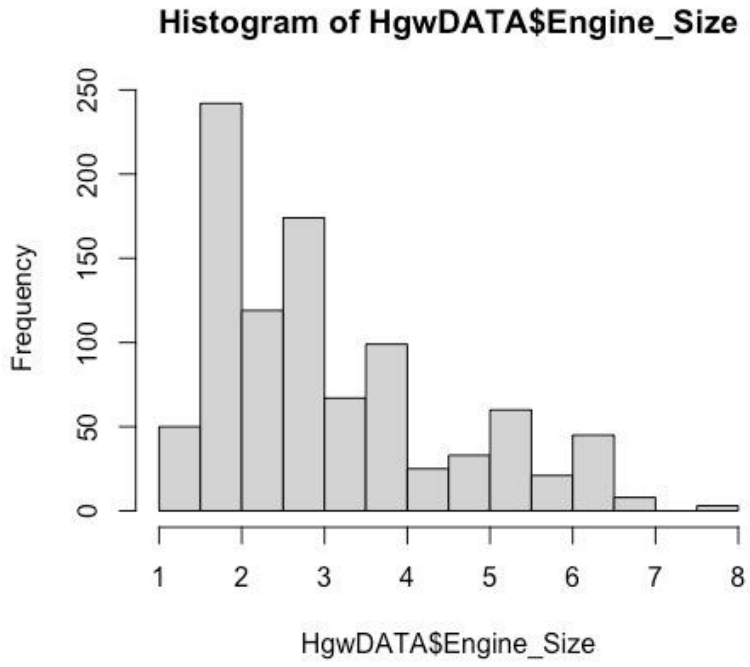


Figure 7 Histogram of Engine Size

According to histogram data, the data types are different from the normal distribution. The appearance of the histograms is the beta distribution. However, according to the Central Limit Theorem, as the sample size increases, the normal distribution-like behavior of the data will increase. Therefore, having 946 data in each column of our data will positively affect this situation.

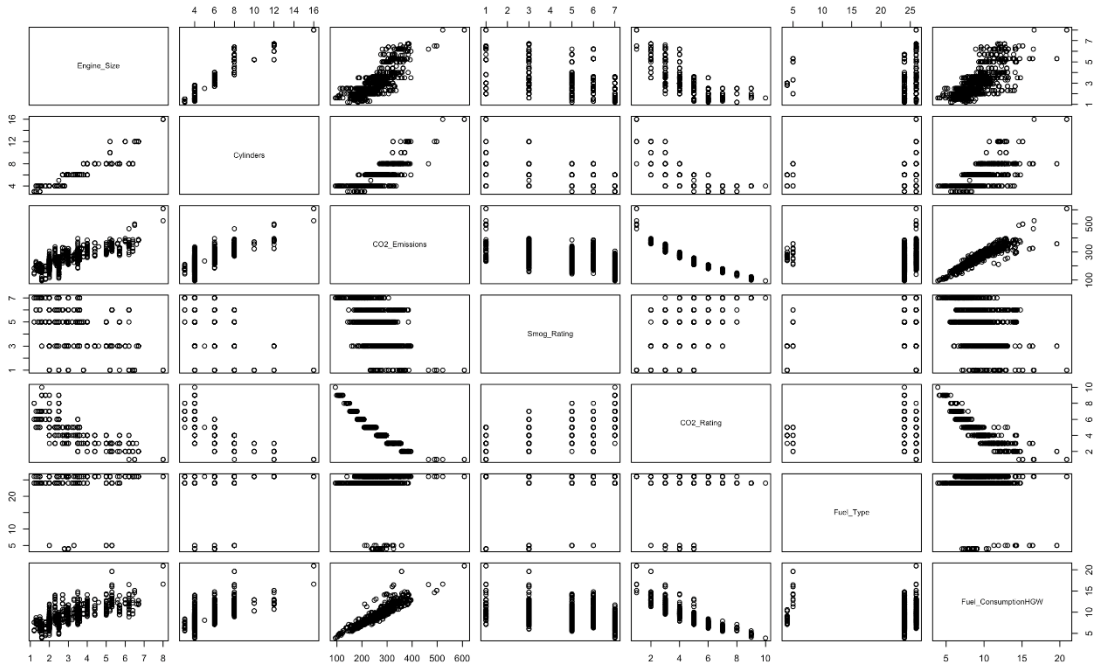


Figure 8 A Matrix of scatterplots

In Figure – 9, a plot matrix consisting of scatter plots for each variable combination of a data frame can be shown. The diagonal shows the names of the seven numeric variables of our sample data. The other cells of the plot matrix show a scatterplot of each variable combination of our data frame.

STATISTICAL ANALYSIS

CORRELATION ANALYSIS

One of the basic analysis to be done in statistical analysis is correlation analysis. This analysis is done to see if there is a correlation between two variables the number will be closer to –1 or 1.

Engine Size and Fuel Consumption

```
> cor.test(HgwDATA$Engine_Size,HgwDATA$Fuel_ConsumptionHGW,conf.level = 0.95)

        Pearson's product-moment correlation

data:  HgwDATA$Engine_Size and HgwDATA$Fuel_ConsumptionHGW
t = 34.772, df = 944, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7200265 0.7760453
sample estimates:
      cor
0.7493738

>
```

Figure 9 Correlation between Engine Size and Fuel Consumption

As it can be seen above the correlation between engine size and the fuel consumption at the high way is very high. 0.749 means there is a big correlation between engine size and fuel consumption at highway is very related to each other. This result is proving us another good point for our hypothesis. When engine sizes are getting bigger the fuel consumption is also getting bigger. To provide a better data we should continue to check the t-tests and others.

Correlation between Cylinder Count and Fuel Consumption

```
> cor.test(HgwDATA$Cylinders,HgwDATA$Fuel_ConsumptionHGW,conf.level = 0.95)

        Pearson's product-moment correlation

data:  HgwDATA$Cylinders and HgwDATA$Fuel_ConsumptionHGW
t = 33.567, df = 944, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7071616 0.7654034
sample estimates:
      cor
0.7376517

>
```

Figure 10 Correlation between Cylinder Count and Fuel Consumption

The cylinder count of a car generally related to engine size. But we didn't want to provide another correlation for it because it may not that much necessary. As like the correlation between engine size and fuel consumption, the correlation between cylinder count and fuel consumption is also high because of the relation between cylinder count and engine size. This means higher cyclinder count means more consumption which probably leads more emission.

Correlation between CO2 Emissions and Fuel Consumptions

```
> cor.test(HgwDATA$CO2_Emissions,HgwDATA$Fuel_ConsumptionHGW,conf.level = 0.95)

        Pearson's product-moment correlation

data:  HgwDATA$CO2_Emissions and HgwDATA$Fuel_ConsumptionHGW
t = 80.315, df = 944, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9253387 0.9416705
sample estimates:
      cor
0.9339908

>
```

Figure 11 Correlation between CO2 Emissions and Fuel Consumptions

CO2 emission of an engine related the combustion type and combustion amount of fuel. To prove this, the correlation between CO2 emission and Fuel Consumption must be

calculated. As it expected the correlation, which is 0.9334, is very much near to 1 which approves our relation between combustion amount and CO2 emission.

Correlation between CO2 Rating and Fuel Consumption

```
> cor.test(HgwDATA$CO2_Rating,HgwDATA$Fuel_ConsumptionHGW,conf.level = 0.95)

        Pearson's product-moment correlation

data:  HgwDATA$CO2_Rating and HgwDATA$Fuel_ConsumptionHGW
t = -61.532, df = 944, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9067018 -0.8811781
sample estimates:
       cor
-0.8946677

>
```

Figure 12 Correlation between CO2 Rating and Fuel Consumption

Most people may think that there is a positive correlation between CO2 Rating and Fuel Consumption but the correlation test shows there is not very big correlation between those two. It's because of the type of fuel doesn't affect the CO2 rating as much as most people expect. The correlation between these 2 variable is nearly 0.9 which approves that even there is a correlation, that correlation is nearly absolute negative.

Correlation between Fuel Type and Fuel Consumption

```
> cor.test(HgwDATA$Fuel_Type,HgwDATA$Fuel_ConsumptionHGW,conf.level = 0.95)

        Pearson's product-moment correlation

data:  HgwDATA$Fuel_Type and HgwDATA$Fuel_ConsumptionHGW
t = -2.4935, df = 944, p-value = 0.01282
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14388817 -0.01724158
sample estimates:
        cor
-0.08089137

>
```

Figure 13 Correlation between Fuel Type and Fuel Consumption

The correlation between fuel type and consumption is very low because fuel type is not one of the reasons which affects the fuel consumption. It doesn't mean that it doesn't

affect but as it can seen in the values absolute cor value for these two is less than 0.1 which means still there is a relation but its too uneffective.

Correlation between Smog Rating and Fuel Consumption

```
> cor.test(HgwDATA$Smog_Rating,HgwDATA$Fuel_ConsumptionHGW,conf.level = 0.95)

        Pearson's product-moment correlation

data:  HgwDATA$Smog_Rating and HgwDATA$Fuel_ConsumptionHGW
t = -13.493, df = 944, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4541972 -0.3472606
sample estimates:
       cor
-0.4020993

>
```

Figure 14 Correlation between Smog Rating and Fuel Consumption

The correlation value between these two variable is not very high but there is still a conenction with these two. Smog rating is increasing while fuel consumption is decreasing. It is because of the technology that used in high fuel consumer vehicles is a bit better than the less fuel consumer ones.

T-TESTS

The t-tests are used to determine the differences between two different sample groups or variables. Higher values of the t-value, also called t-score, indicate that a large difference exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets. A large t-score indicates that the groups are different. A small t-score indicates that the groups are similar.

The hypothesis H_0 is that true difference between two variables are not equal zero.

$H_0 \neq 0$;

Also the confidence interval is considered as 95%. 95% is general interval for common and many cases.

T-test for Engine Size and Fuel Consumption

```
> t.test(HgwDATA$`Engine_Size`,HgwDATA$Fuel_ConsumptionHGW,paired = T)

        Paired t-test

data:  HgwDATA$Engine_Size and HgwDATA$Fuel_ConsumptionHGW
t = -122.3, df = 945, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.263505 -6.065670
sample estimates:
mean of the differences
              -6.164588

>
```

Figure 15 T-test for Engine Size and Fuel Consumption

The t-score which is –6.1646 is very low. It shows these two data sets are not much different than each other. Also the confidence interval as shown is between –6.2635 and –6.0656 and this interval provides many things to us.

T-test for Cylinder Number and Fuel Consumption

```
> t.test(HgwDATA$Cylinders,HgwDATA$Fuel_ConsumptionHGW,paired = T)

        Paired t-test

data:  HgwDATA$Cylinders and HgwDATA$Fuel_ConsumptionHGW
t = -72.738, df = 945, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.794941 -3.595545
sample estimates:
mean of the differences
              -3.695243

>
```

Figure 16 T-test for Cylinder Number and Fuel Consumption

The t-score of cylinder number and fuel consumption is again very low. Having –3.6952 as a t-score is a good thing for an interval search. This t-score shows us the difference between cylinder count and fuel consumption is not much and these two variables are close to each other.

T-test for CO2 Emissions and Fuel Consumption

```
> t.test(HgwDATA$CO2_Emissions,HgwDATA$Fuel_ConsumptionHGW,paired = T)

        Paired t-test

data:  HgwDATA$CO2_Emissions and HgwDATA$Fuel_ConsumptionHGW
t = 123.3, df = 945, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 245.833 253.785
sample estimates:
mean of the differences
                249.809

>
```

Figure 17 T-test for CO2 Emissions and Fuel Consumption

T-score for CO2 emission and fuel consumption is very high. While having a t-value as 123.3, the mean of differences is nearly 250 which is far greater than our t-value at the

beginning. This test shows us the difference between the distributions and mean of the differences of these two values are not very related and not close.

T-test for Smog Rating and Fuel Consumption

```
> t.test(HgwDATA$Smog_Rating,HgwDATA$Fuel_ConsumptionHGW,paired = T)

        Paired t-test

data:  HgwDATA$Smog_Rating and HgwDATA$Fuel_ConsumptionHGW
t = -40.684, df = 945, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.625875 -4.200129
sample estimates:
mean of the differences
              -4.413002

>
```

Figure 18 T-test for Smog Rating and Fuel Consumption

The t-score for smog rating and fuel consumption is very low which led us to check the mean differences. Mean of differences is – 4.4130 which is very low and the meaning of this is known. These two distribution is very similar to each other.

T-test for CO2 Rating and Fuel Consumption

```
> t.test(HgwDATA$CO2_Rating,HgwDATA$Fuel_ConsumptionHGW,paired = T)

        Paired t-test

data:  HgwDATA$CO2_Rating and HgwDATA$Fuel_ConsumptionHGW
t = -40.525, df = 945, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.057826 -4.590588
sample estimates:
mean of the differences
              -4.824207

>
```

Figure 19 T-test for CO2 Rating and Fuel Consumption

As it can seen from above the mean of the differences between CO2 rating and fuel consumption is very low. The mean of the differences is –4.8242 which proves the distribuition and the means of these two variables are very close to each other.

T-test for Fuel Type and Fuel Consumption

```
> t.test(HgwDATA$Fuel_Type,HgwDATA$Fuel_ConsumptionHGW,paired = T)

        Paired t-test

data:  HgwDATA$Fuel_Type and HgwDATA$Fuel_ConsumptionHGW
t = 89.929, df = 945, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 14.46824 15.11379
sample estimates:
mean of the differences
               14.79101

>
```

Figure 20 T-test for Fuel Type and Fuel Consumption

The difference between the distributions of fuel type and fuel consumption is a bit higher than expected. Since the fuel type of an engine doesn't affect the fuel consumption of a car very much this value was expecting. The mean of the differences is 14.7910 in a 945 degree of freedoom and a t value at 89.929. These numbers led us to assume that there is not much common things in these two different samples and their distributions.

MODEL 1

```
> LMMODEL <- lm(HgwDATA$Fuel_ConsumptionHGW~HgwDATA$Engine_Size+HgwDATA$Cylinders)
> summary(LMMODEL)

Call:
lm(formula = HgwDATA$Fuel_ConsumptionHGW ~ HgwDATA$Engine_Size +
    HgwDATA$Cylinders)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6266 -0.9388 -0.1271  0.7660  7.7630

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.81327    0.15704  30.650  < 2e-16 ***
HgwDATA$Engine_Size  0.76627    0.09025   8.490  < 2e-16 ***
HgwDATA$Cylinders    0.37032    0.06420   5.768 1.09e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.489 on 943 degrees of freedom
Multiple R-squared:  0.5765,    Adjusted R-squared:  0.5756
F-statistic: 641.8 on 2 and 943 DF,  p-value: < 2.2e-16
```

Figure 21 Model of fuel consumption between engine size and cylinders

First model is trained to find the regression of the fuel consumption between engine size and cylinders.

The model is trained in the first line above. Let’s check the result.

Multiple R-squared: 0.5765

Adjusted R-squared:0.5756

Since the compared variable is very low and they are very related to the fuel consumption this shows a great correlation between fuel consumption and engine size and cylinder count.

```
> anova(LMMODEL)
Analysis of Variance Table

Response: HgwDATA$Fuel_ConsumptionHGW
                    Df  Sum Sq Mean Sq  F value    Pr(>F)
HgwDATA$Engine_Size   1 2771.08 2771.08 1250.425 < 2.2e-16 ***
HgwDATA$Cylinders     1   73.73   73.73   33.269 1.087e-08 ***
Residuals           943 2089.79    2.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
> summary(regfordata)
Subset selection object
Call: regsubsets.formula(Fuel_ConsumptionHGW ~ Engine_Size + Cylinders +
    CO2_Emissions + Smog_Rating + CO2_Rating + Fuel_Type, data = HgwDATA,
    method = "exhaustive", nvmax = 2)
6 Variables  (and intercept)
              Forced in Forced out
Engine_Size       FALSE      FALSE
Cylinders         FALSE      FALSE
CO2_Emissions     FALSE      FALSE
Smog_Rating       FALSE      FALSE
CO2_Rating        FALSE      FALSE
Fuel_Type         FALSE      FALSE
1 subsets of each size up to 2
Selection Algorithm: exhaustive
         Engine_Size Cylinders CO2_Emissions Smog_Rating CO2_Rating Fuel_Type
1  ( 1 ) " "         " "       "*"           " "          " "         " "
2  ( 1 ) " "         " "       "*"           "*"          " "         " "
> summary(regfordata)$adjr2
[1] 0.8722035 0.8817620
>
```

```
> LMMODEL3 <- lm(HgwDATA$Fuel_ConsumptionHGW~Cylinders,data = HgwDATA)
> summary(LMMODEL)

Call:
lm(formula = HgwDATA$Fuel_ConsumptionHGW ~ HgwDATA$Engine_Size +
    HgwDATA$Cylinders)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6266 -0.9388 -0.1271  0.7660  7.7630

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.81327    0.15704  30.650  < 2e-16 ***
HgwDATA$Engine_Size 0.76627    0.09025   8.490  < 2e-16 ***
HgwDATA$Cylinders   0.37032    0.06420   5.768 1.09e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.489 on 943 degrees of freedom
Multiple R-squared:  0.5765,    Adjusted R-squared:  0.5756
F-statistic: 641.8 on 2 and 943 DF,  p-value: < 2.2e-16

> anova(LMMODEL)
Analysis of Variance Table

Response: HgwDATA$Fuel_ConsumptionHGW
                   Df  Sum Sq Mean Sq  F value    Pr(>F)
HgwDATA$Engine_Size   1 2771.08 2771.08 1250.425 < 2.2e-16 ***
HgwDATA$Cylinders     1   73.73   73.73   33.269 1.087e-08 ***
Residuals           943 2089.79    2.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 22 Model of fuel consumption between engine size and cylinders and CO2 emission and smog rating

This model is created to understand the regression of engine size and cylinder count and co2 emission and co2 rating and fuel type and smog rating.

```
Call:
lm(formula = HgwDATA$Fuel_ConsumptionHGW ~ HgwDATA$Engine_Size +
    HgwDATA$Cylinders + HgwDATA$CO2_Emissions + HgwDATA$CO2_Rating +
    HgwDATA$Smog_Rating + HgwDATA$Fuel_Type)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8381 -0.3349 -0.0472  0.3140  6.0202

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.393947   0.602843   0.653    0.514
HgwDATA$Engine_Size  0.070335   0.046928   1.499    0.134
HgwDATA$Cylinders   -0.159053   0.035064  -4.536 6.47e-06 ***
HgwDATA$CO2_Emissions  0.038559   0.001510  25.536  < 2e-16 ***
HgwDATA$CO2_Rating   0.018701   0.056311   0.332    0.740
HgwDATA$Smog_Rating  0.168764   0.017398   9.700  < 2e-16 ***
HgwDATA$Fuel_Type   -0.052492   0.005665  -9.265  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.741 on 939 degrees of freedom
Multiple R-squared:  0.8955,    Adjusted R-squared:  0.8948
F-statistic:  1341 on 6 and 939 DF,  p-value: < 2.2e-16

> anova(alldata)
Analysis of Variance Table

Response: HgwDATA$Fuel_ConsumptionHGW
                     Df  Sum Sq Mean Sq   F value    Pr(>F)
HgwDATA$Engine_Size   1 2771.08 2771.08 5046.2765 < 2.2e-16 ***
HgwDATA$Cylinders     1   73.73   73.73  134.2633 < 2.2e-16 ***
HgwDATA$CO2_Emissions   1 1490.38 1490.38 2714.0510 < 2.2e-16 ***
HgwDATA$CO2_Rating    1    0.17    0.17    0.3071    0.5796
HgwDATA$Smog_Rating   1   36.47   36.47   66.4126 1.162e-15 ***
HgwDATA$Fuel_Type     1   47.14   47.14   85.8477 < 2.2e-16 ***
Residuals           939  515.64    0.55
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 23 Model of fuel type and CO2 emissions

As it can be seen results given above, adding smog rating and CO2 emission affects the regression. These results show us that Fuel type is not very effective on the CO2 emission as much as other variables.

```
> regfordata <- regsubsets(Fuel_ConsumptionHGW ~  Engine_Size  + Cylinders + CO2_Emissions + Smog_Rating
 +CO2_Rating + Fuel_Type, data = HgwDATA, method="exhaustive", nvmax=2)
> summary(regfordata)
Subset selection object
Call: regsubsets.formula(Fuel_ConsumptionHGW ~ Engine_Size + Cylinders +
    CO2_Emissions + Smog_Rating + CO2_Rating + Fuel_Type, data = HgwDATA,
    method = "exhaustive", nvmax = 2)
6 Variables  (and intercept)
             Forced in Forced out
Engine_Size       FALSE      FALSE
Cylinders         FALSE      FALSE
CO2_Emissions     FALSE      FALSE
Smog_Rating       FALSE      FALSE
CO2_Rating        FALSE      FALSE
Fuel_Type         FALSE      FALSE
1 subsets of each size up to 2
Selection Algorithm: exhaustive
         Engine_Size Cylinders CO2_Emissions Smog_Rating CO2_Rating Fuel_Type
1  ( 1 ) " "         " "       "*"           " "          " "         " "
2  ( 1 ) " "         " "       "*"           "*"          " "         " "
> summary(regfordata)$adjr2
[1] 0.8722035 0.8817620
>
```

```
> summary(regfordata)$adjr2
[1] 0.8722035 0.8817620
> #higher is better
> summary(regfordata)$cp
[1] 205.1843 120.2566
> #lower better
> summary(regfordata)$bic
[1] -1933.518 -2001.210
> #lower better
>
```

```
> summary(alldata)$cov.unscaled
                     (Intercept) HgwDATA$Cylinders HgwDATA$CO2_Emissions HgwDATA$CO2_Rating
(Intercept)         0.6606460408      5.077820e-03        -1.495500e-03      -5.795561e-02
HgwDATA$Cylinders   0.0050778197      9.909876e-04        -3.391224e-05      -4.918366e-04
HgwDATA$CO2_Emissions -0.0014955000     -3.391224e-05         4.119022e-06       1.367198e-04
HgwDATA$CO2_Rating  -0.0579556109     -4.918366e-04         1.367198e-04       5.759679e-03
HgwDATA$Fuel_Type   -0.0009699983     -1.024146e-05        -5.606096e-07      -1.727736e-05
HgwDATA$Smog_Rating -0.0028849785      1.159879e-04         2.653416e-06      -8.426207e-05
                    HgwDATA$Fuel_Type HgwDATA$Smog_Rating
(Intercept)             -9.699983e-04       -2.884978e-03
HgwDATA$Cylinders       -1.024146e-05        1.159879e-04
HgwDATA$CO2_Emissions   -5.606096e-07        2.653416e-06
HgwDATA$CO2_Rating      -1.727736e-05       -8.426207e-05
HgwDATA$Fuel_Type        5.837988e-05       -3.199008e-05
HgwDATA$Smog_Rating     -3.199008e-05        5.444161e-04
>
```

Figure 24 The Summary of Full Model

DISCUSSION

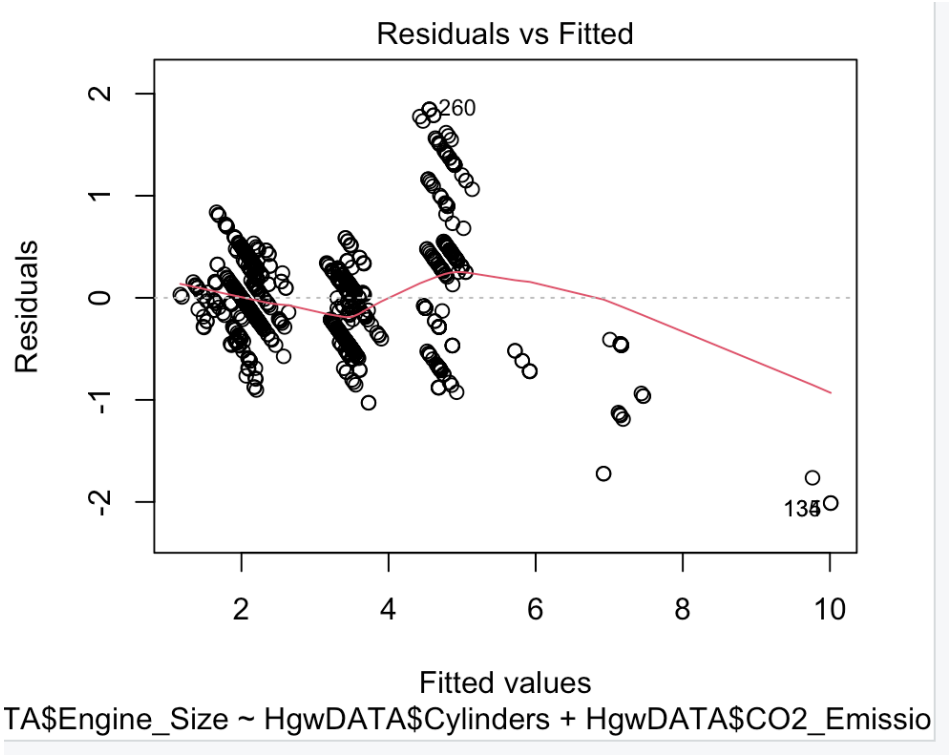


Figure 25 The Plot of Residuals and Fitted Values

Looking at this graph, it is seen that the values are distributed separately from each other. If the inserted values increase, the spread will be less.

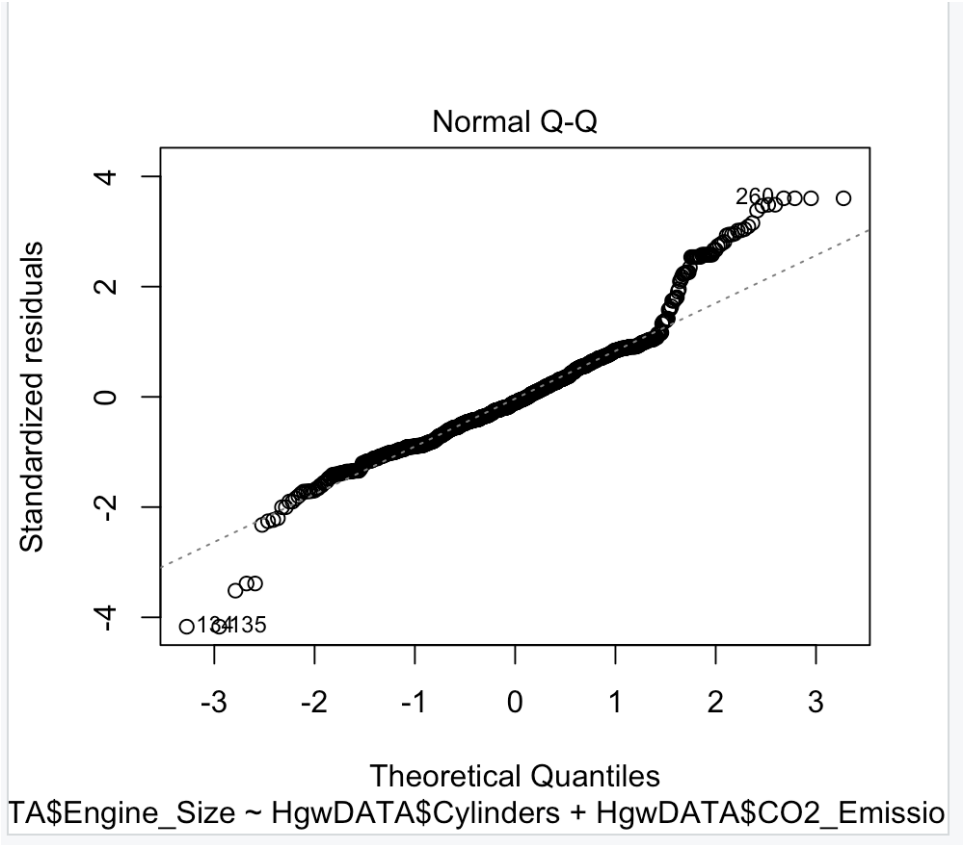


Figure 26 Normal Q-Q Plot of Full Model

Normal Q-Q plots are helpful in understanding whether the residuals are approximately normally distributed or not. If the residual slopes are larger than would be expected from the normal distribution, the p-values and confidence intervals will be more optimistic than normal.This situation tells us that the data can be variable and there may be confusion when these are taken into account. In the Normal Q-Q graphs obtained in the study, the data do not continue normally. This explains that the data are not normally distributed.

In summary, in the study, it was observed that the factors tested with statistical analyzes affected factors such as fuel consumption and CO2 emissions to a varying extent.

CONCLUSION

Research has been carried out on the fuel consumption of vehicles by considering various factors. In this research, we tried to determine how the determined factors affect fuel consumption in vehicles. In addition, data from 946 vehicles from Kaggle were used in this study. Some data and summaries are provided for all data after filtering and for the instruments included in the study. As a result, the results of this research, engine size and cylinder number are the factors affecting fuel consumption. Also, fuel consumption affects the CO2 rate and the smog rate. Finally, to support future studies, we will try to obtain more reliable data that can strengthen the purpose of the research by using different techniques that will help make our findings more effective.

REFERENCES

1-Jereb, B., Kumperščak, S., & Bratina, T. (2018). The impact of traffic flow on fuel consumption increase in the urban environment. FME Transaction, 46(3), 278–284. **https://doi.org/10.5937/fmet1802278j**

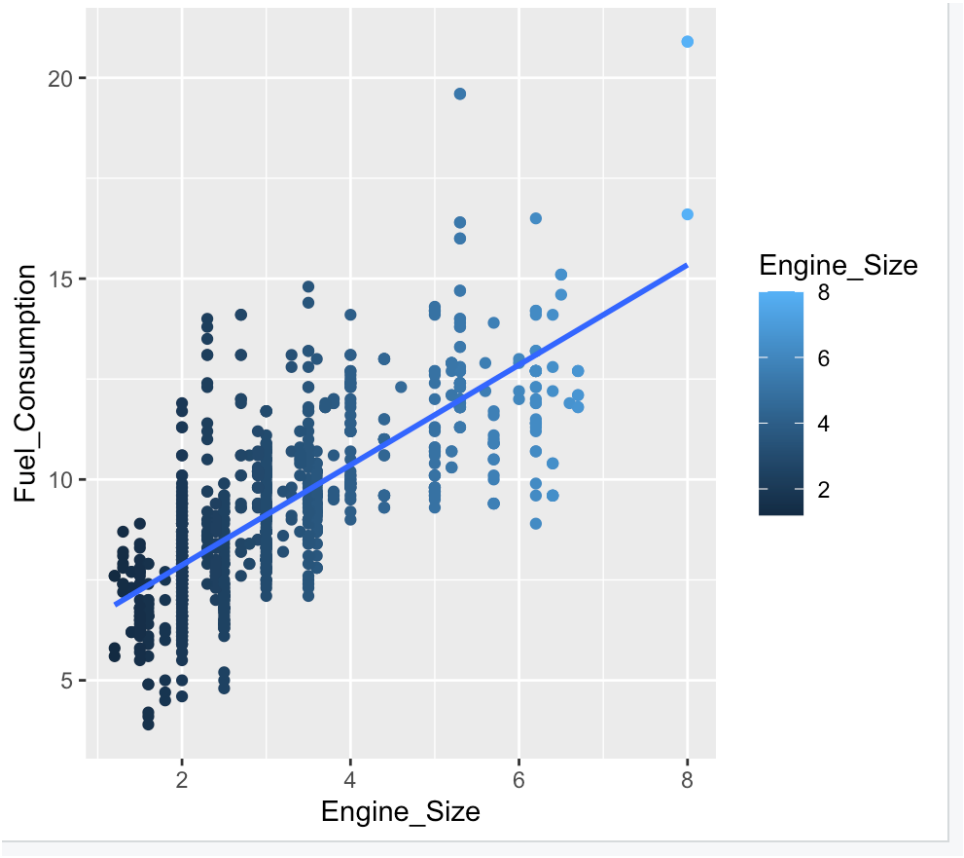
2-https://www.rdocumentation.org/packages/inlabru/versions/2.5.2/topics/gg.prediction

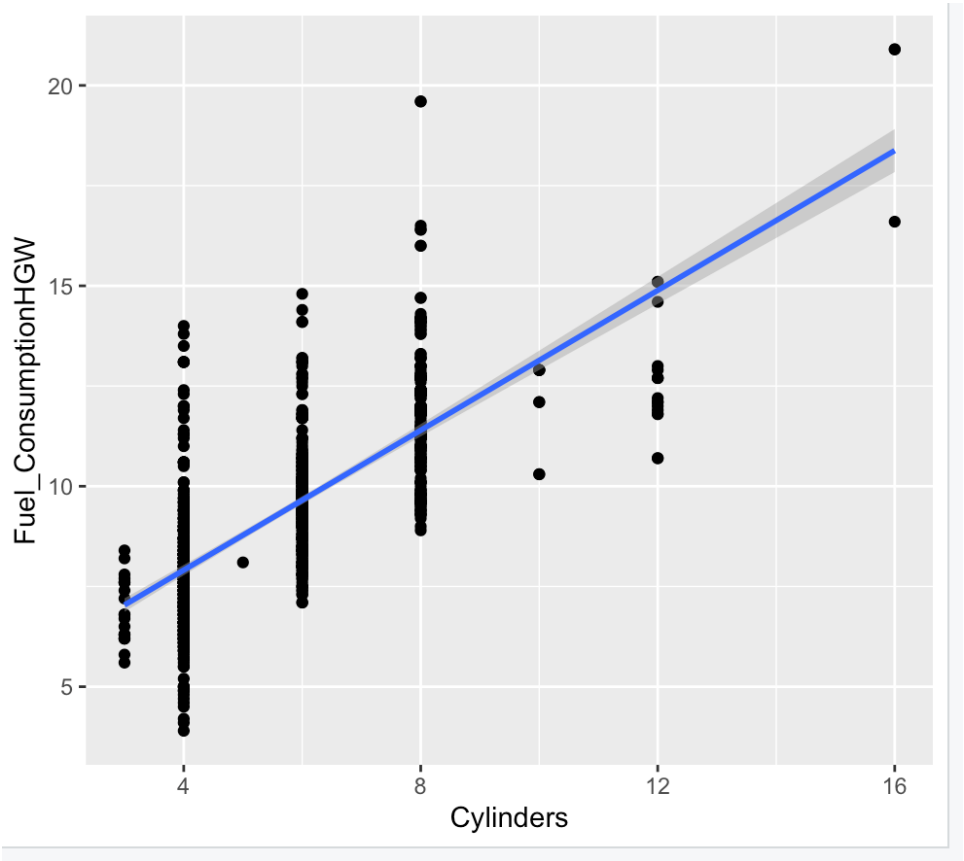
3-https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html

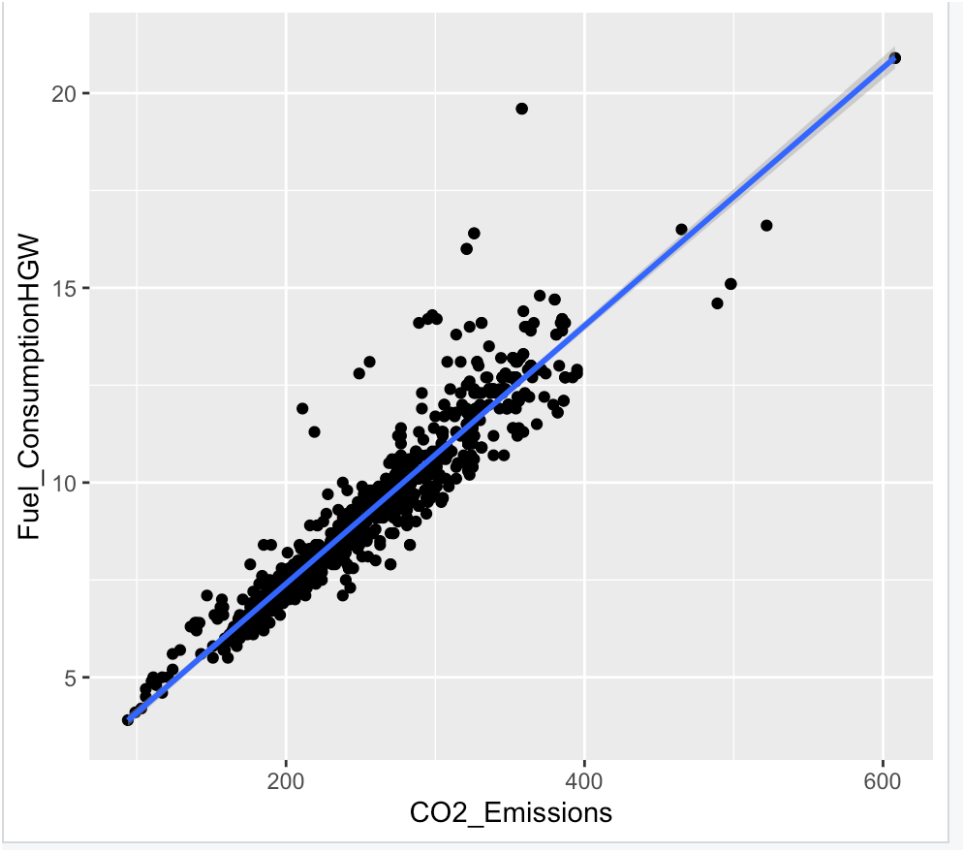
4-https://libguides.library.kent.edu/spss/independentttest

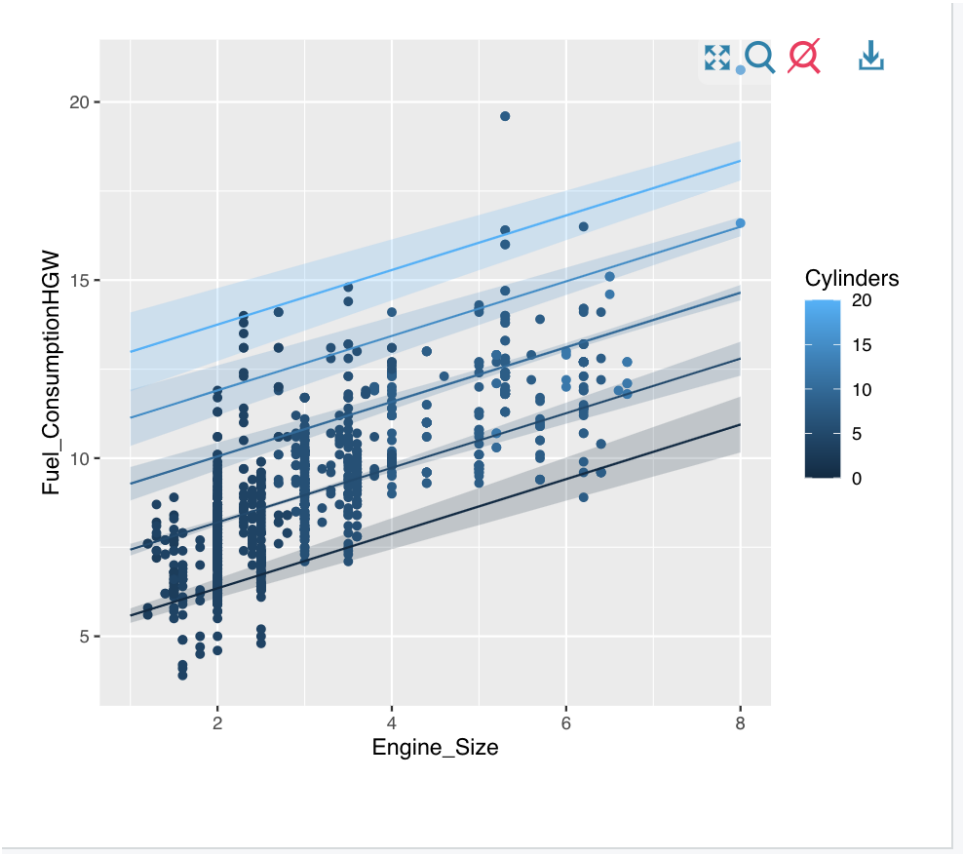
5-https://rdrr.io/cran/ggiraphExtra/man/ggPredict.html

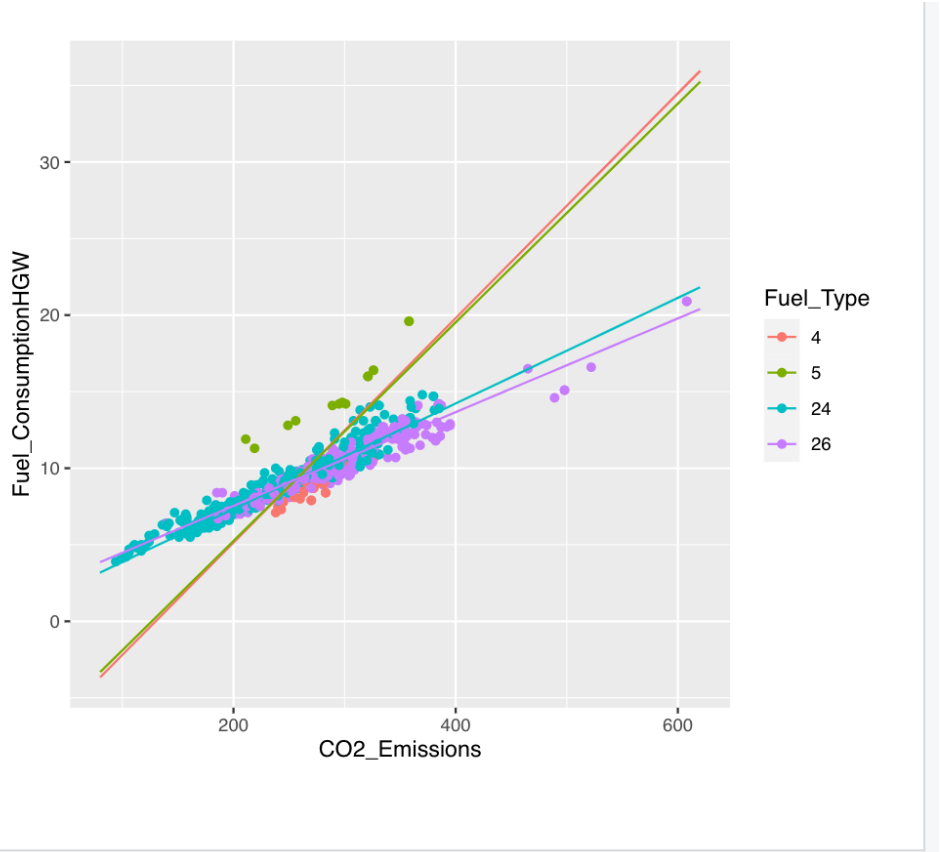
APPENDICES


Fuel_Consumption
Engine_Size
Engine_Size
8
6
4
2
20
15
10
5
2
4
6
8

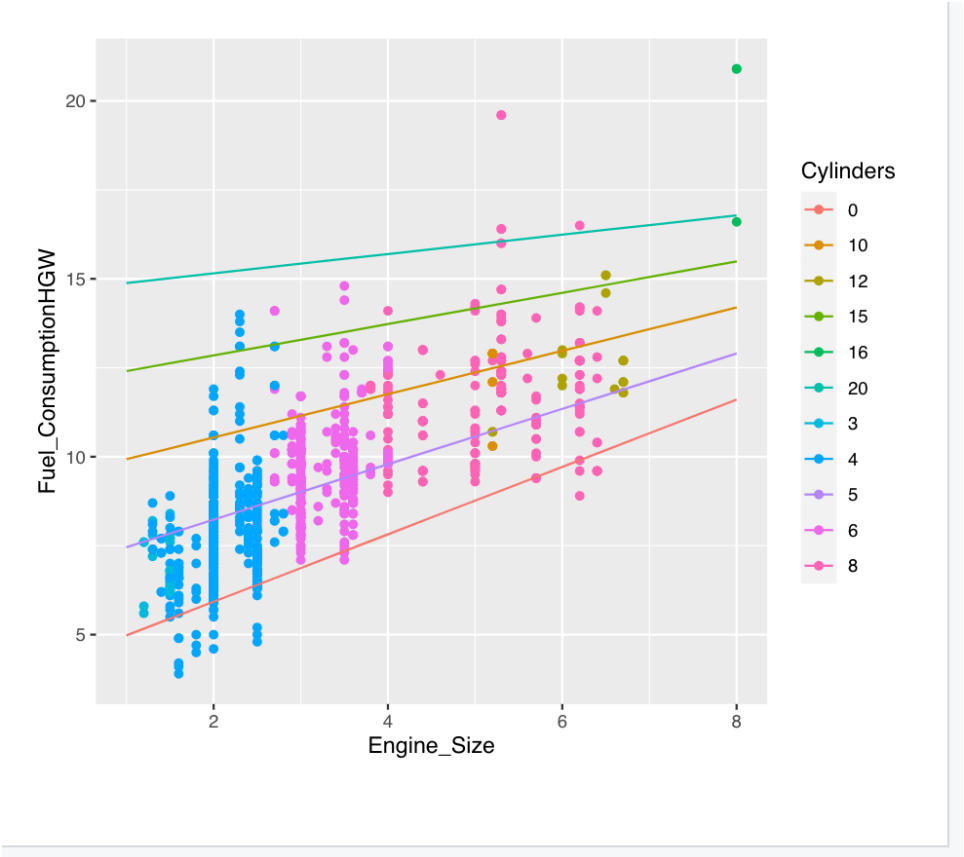

Fuel_ConsumptionHGW
Cylinders
20
15
10
5
4
8
12
16

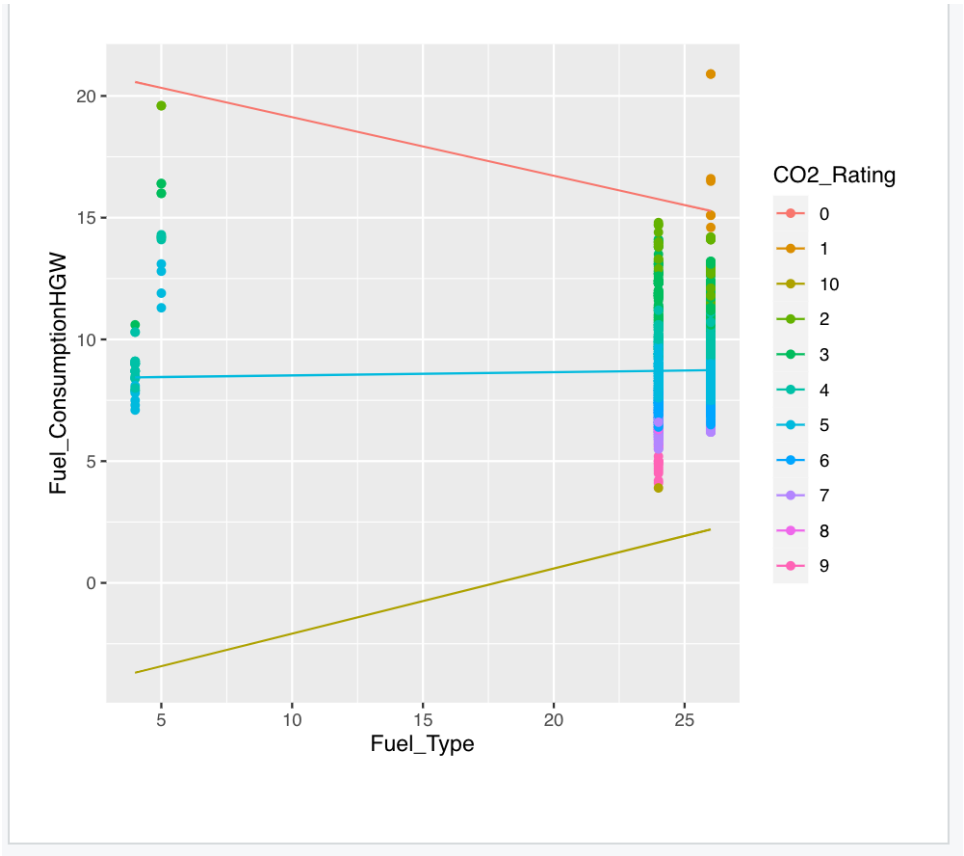

Fuel_ConsumptionHGW
20
15
10
5
200
400
600
CO2_Emissions

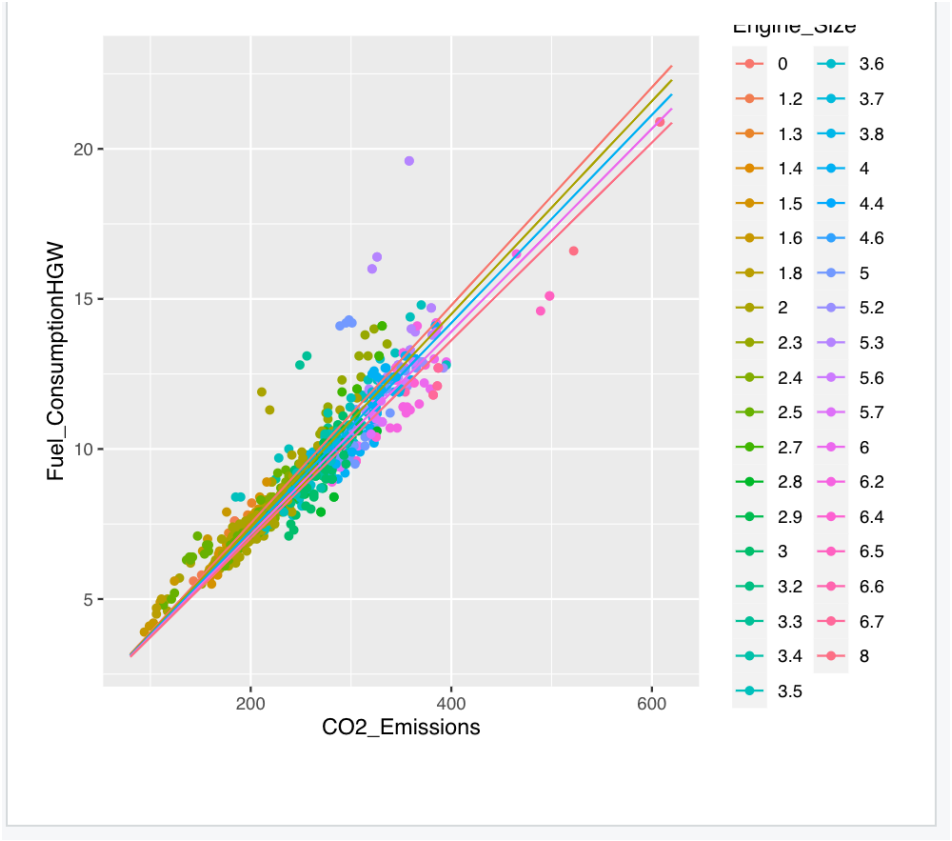

Fuel_ConsumptionHGW
20
15
10
5
2
4
6
8
Engine_Size
Cylinders
20
15
10
5
0

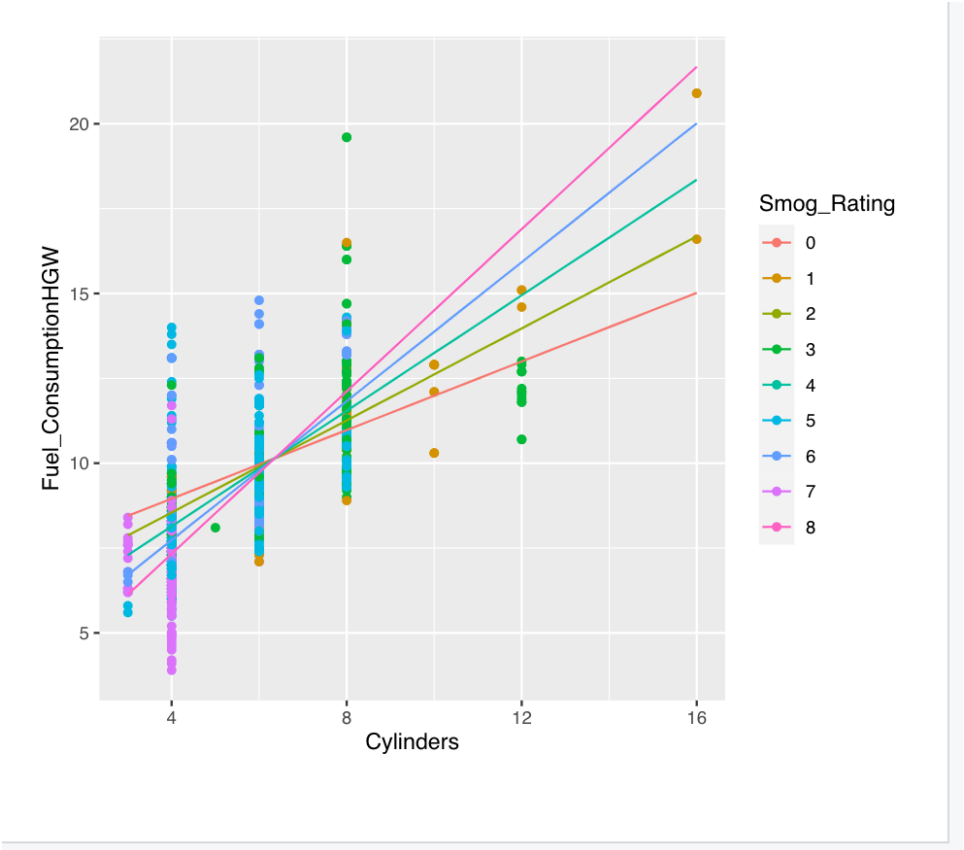



```
Engine_Size | Predicted |               95% CI
-----------------------------------------------
       1.20 |  3.37e+05 | [3.23e+05, 3.52e+05]
       1.60 |  3.37e+05 | [3.23e+05, 3.52e+05]
       2.50 |  3.37e+05 | [3.23e+05, 3.52e+05]
       3.00 |  3.37e+05 | [3.23e+05, 3.52e+05]
       3.50 |  3.37e+05 | [3.23e+05, 3.52e+05]
       4.40 |  3.37e+05 | [3.23e+05, 3.52e+05]
       5.30 |  3.37e+05 | [3.23e+05, 3.52e+05]
       8.00 |  3.37e+05 | [3.23e+05, 3.52e+05]

Adjusted for:
*     Cylinders =     5.67
* CO2_Emissions = 71318.81

$Cylinders
# Predicted values of Fuel_ConsumptionHGW

Cylinders | Predicted |               95% CI
-----------------------------------------------
        3 |  3.37e+05 | [3.23e+05, 3.52e+05]
        4 |  3.37e+05 | [3.23e+05, 3.52e+05]
        5 |  3.37e+05 | [3.23e+05, 3.52e+05]
        6 |  3.37e+05 | [3.23e+05, 3.52e+05]
        8 |  3.37e+05 | [3.23e+05, 3.52e+05]
       10 |  3.37e+05 | [3.23e+05, 3.52e+05]
       12 |  3.37e+05 | [3.23e+05, 3.52e+05]
       16 |  3.37e+05 | [3.23e+05, 3.52e+05]

Adjusted for:
*   Engine_Size =     3.20
* CO2_Emissions = 71318.81

$CO2_Emissions
# Predicted values of Fuel_ConsumptionHGW
```


Fuel_ConsumptionHGW
Engine_Size
Cylinders
0
10
12
15
16
20
3
4
5
6
8


Fuel_ConsumptionHGW
Fuel_Type
CO2_Rating
0
1
10
2
3
4
5
6
7
8
9


Engine_Size
0
1.2
1.3
1.4
1.5
1.6
1.8
2
2.3
2.4
2.5
2.7
2.8
2.9
3
3.2
3.3
3.4
3.5
3.6
3.7
3.8
4
4.4
4.6
5
5.2
5.3
5.6
5.7
6
6.2
6.4
6.5
6.6
6.7
8
Fuel_ConsumptionHGW
20
15
10
5
200
400
600
CO2_Emissions


Smog_Rating
0
1
2
3
4
5
6
7
8
Fuel_ConsumptionHGW
20
15
10
5
4
8
12
16
Cylinders

R CODES

1. **NewRCode.R (R Script)**
2. **FuelConsumptionData.xlsx**
3. **highwaydata.xlsx**