**Table of Contents**

Contents

[INTRODUCTION 3](#_Toc93766990)

[DATA IMPORT/DATA CLEANING/DATA PRE-PROCESSING AND TRANSFORMATION 4](#_Toc93766991)

[2.1 Data Import 4](#_Toc93766992)

[2.2 Data Cleaning, Pre-processing & Transformation 4](#_Toc93766993)

[ASSUMPTION ANALYSIS 6](#_Toc93766994)

[CONCLUSION 25](#_Toc93766995)

[4 References 26](#_Toc93766996)

# INTRODUCTION

The following documentation details an examination of a dataset's fuel consumption, carbon dioxide emissions, and tax bands using the computer language R. The  R is a widely used programming language for data analysis and development by a large number of statisticians and data miners. Additionally, it is a free software environment, which implies that anybody may use it to study and analyze data in order to make judgments.

The Vehicle Certification Agency (VCA) donated the dataset, which spans over 45,000 rows and includes vehicles manufactured between 2000 and 2013. For each vehicle entry, 28 additional parameters relevant to CO2 emissions, fuel consumption, and tax bands are stated. This documentation begins with an analysis of the data in general, followed by the creation of four major questions that define the study objectives. Each question will be researched, adjusted, and visualized, along with an explanation of the strategies used.

**OBJECTIVES/ASSUMPTIONS:-**

* Analysis on group for cheaper vehicles:- This analysis deals with analyzing different features form the dataset to evaluate the price factor.
* Analysis on group for lesser noise pollution:- This analysis deals with analyzing the noise feature form the dataset with other features to evaluate the effect of noise on groups.
* Analysis on group for Environmental impacts.:- This analysis deals with analyzing the co2, co and thc emission features from the dataset with other features to evaluate the effect of the selected features on environment based on different groups.
* Analysis on suitability to users.:- This analysis deals with analyzing several features such as engine, transmission, euro standard such as to provide the user with a selection of best data based on the requirement.

# DATA IMPORT/DATA CLEANING/DATA PRE-PROCESSING AND TRANSFORMATION

## 2.1 Data Import

The function for importing data from an external file into a variable in RStudio is shown in Figure 3.2. The read.csv() method require two parameters: one for the file name, and other one for checking the columns and including it in the environment.



Figure 1: Data read

## 2.2 Data Cleaning, Pre-processing & Transformation

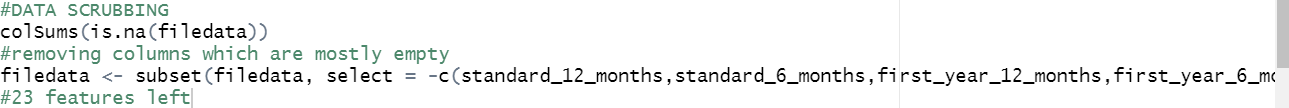


Figure : Data scrubbing

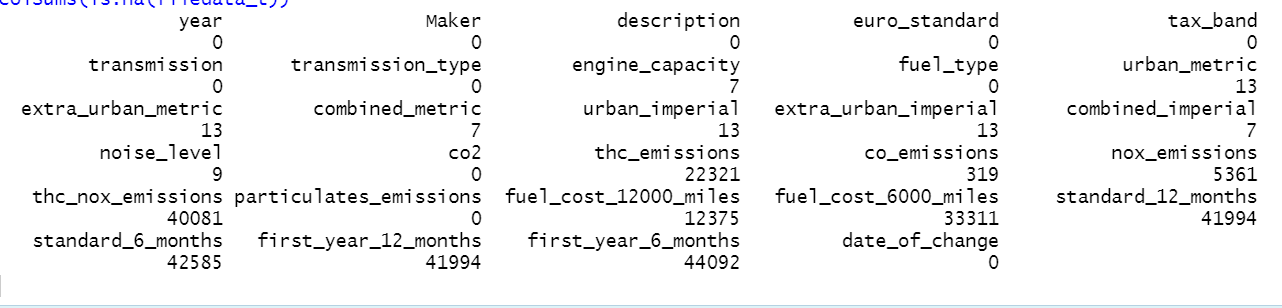


Figure : Output

Figure 2 shows the data scrubbing techniques was used to clean the data or in other words removing unnecessary features which will affect the analysis. colSums function was used in which is.na function with parameters name of the file is included. This returns number of Na data available in the dataset as per column. After analyzing

the output as shown in figure 3 it was evident that last few columns has almost all Na data hence it will be of no use. Subset command is used to with -c function to remove all the unnecessary columns from the dataset and file data variables stored the new data.

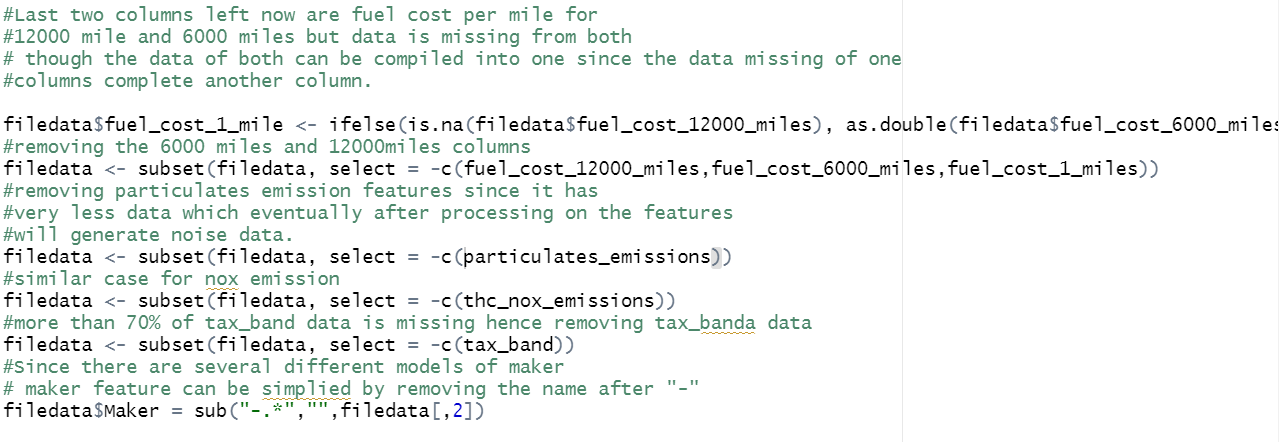


Figure :- Data scrubbing.

In the two columns – fuel\_cost\_6000\_miles and fuel\_cost\_12000 miles there was a pattern that was analyzed, one contains the data which other did not. Basically, if both columns will be merged it will form a complete data. As an alternative to combining the two columns into one, a new column fuel\_cost\_1\_mile was created, each data from 6000 miles was divided by 6000 while the 12000-mile columns data was divided by 12000. The resulted data was stored in fuel\_cost\_1\_mile and populate it with the matching data before rounding it off and removing the two columns. NA values are excluded from the following mathematical calculations. Following that most of the data in the columns of thc\_nox emission, particulates\_emission and tax\_band emission was missing. Hence these three columns were removed separately. In the maker column the model number was included with the maker name with “-“ symbol. To remove the model number from the maker column sub() function was used, the first parameter was the symbol after which the words will be removed, and second parameter is the filedata frame.

# ASSUMPTION ANALYSIS

Question 1:- Which brand and which features are related to cheaper value?

This analysis is executed to get the details for brands and features which relates with the lower price of the vehicles. Such as which brands consumes less fuel, which transmission type consumes low fuel and which fuel type is related for consumption of less fuel. This analysis will help user in selecting the brand and the feature types.

**ANALYSIS 1:- Analysis based on brands.**

The first   analysis looks at which makers have the lowest fuel costs. In this study, the fuel cost 1 mile vector will be utilized as a measure of cost.

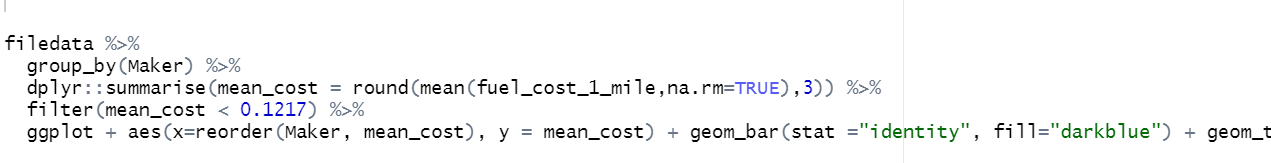


Figure :- Q1A1 Code.

The code for the above-mentioned analysis is seen in Figure 5. Since there are multiple rows of same group name in Maker column, group\_by() function is used to perform the group operation. Dyplyr library is used for orgranizing the data in particular set, such as in this case it is the Maker column with fuel\_cost\_1\_mile column. mean\_cost() function is provided as the parameter to the summarize group, this is to know the mean value of the fuel\_cost\_1\_mile column. If the row contains Na value, it will not be included. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is sorted by increasing mean fuel cost using the reorder, and the mean fuel cost is shown by the y axis.geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and the data colour dark blue. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 6 shows the graph plotted for analysis 1 which shows that maker micro compact car, smart and Daihatsu are the three makers which has the lowest fuel cost per mile.

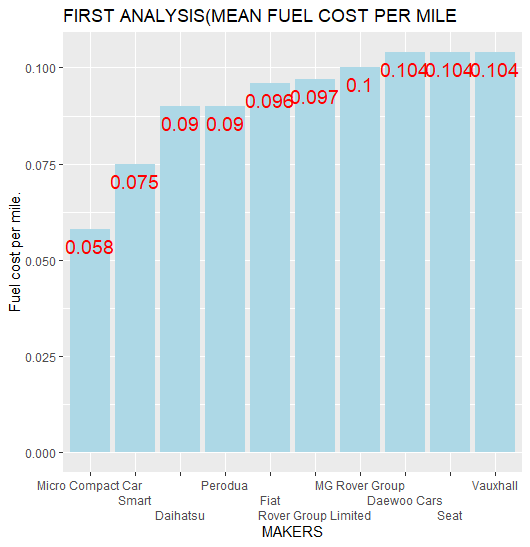


Figure : Q1A1 chart.

**ANALYSIS 2:- Analysis based on Transmission type.**

The second analysis looks at which transmission type consumes less as per urban metric standard. In this study, the urban\_metric will be utilized as a measure of cost.

A picture containing Word

Description automatically generated

Figure 7:- Q1A2 Code.

The code for the above-mentioned analysis is seen in Figure 7. Since there are multiple rows of same transmission name in transmission\_type column, group\_by() function is used to perform the group operation. Summarise\_at() function is used to get the information about the selected two columns, the information include the number of row, mean and median value. If the row contains Na value, it will not be included. Following this, the ggplot() function is called, the transmission type is represented by the x axis, which is sorted by increasing mean urban condition, and the mean urban conditon is shown by the y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and the width is set as 1. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed which was implemented by label function.

Figure 8 shows the graph plotted for analysis 2 which shows the transmission type is shown by the y-axis, while the fuel consumption is represented by the x-axis. From this chart, we can observe that manual transmissions utilize less gasoline than automated transmissions from year to year.

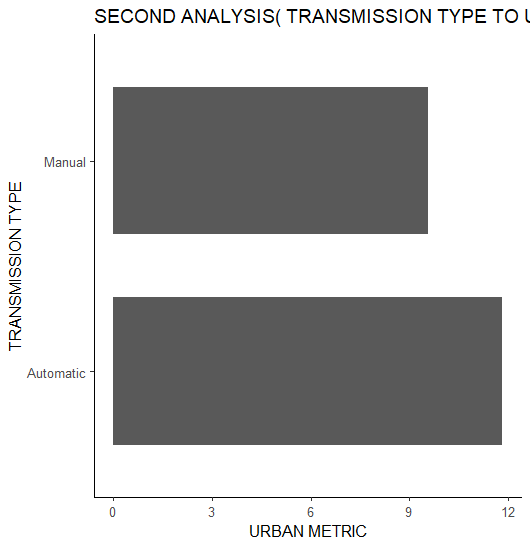


Figure 8: Q1A2 chart.

**ANALYSIS 3:- Analysis based by fuel type and fuel consumption.**

The third analysis looks at which fuel type consumes less as per fuel consumption. In this study, the fuel\_cost\_1\_mile will be utilized as a measure of cost.

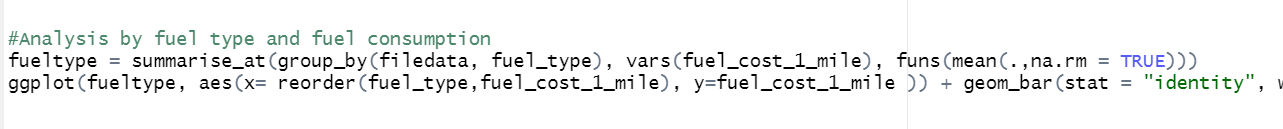


Figure 9:- Q1A3 Code.

The code for the above-mentioned analysis is seen in Figure 9. Since there are multiple rows of same fuel type in fuel\_tyepe column, group\_by() function is used to perform the group operation. Summarise\_at() function is used to get the information about the selected two columns, the information include the number of row, mean and median value. If the row contains Na value, it will not be included. Following this, the ggplot() function is called, the fuel\_type is represented by the x axis, which is sorted by increasing mean fuel\_cost, and the fuel\_cost is shown by the y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parameter is set to identity value and the width is set as 0.3. To adjust the x label font size theme() function was used. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed which was implemented by label function.

Figure 9 shows the graph plotted for analysis 3 which shows the fuel type is shown by the x-axis, while the fuel consumption is represented by the y-axis. From this chart, we can observe that manual transmissions utilize less gasoline than automated transmissions from year to year.

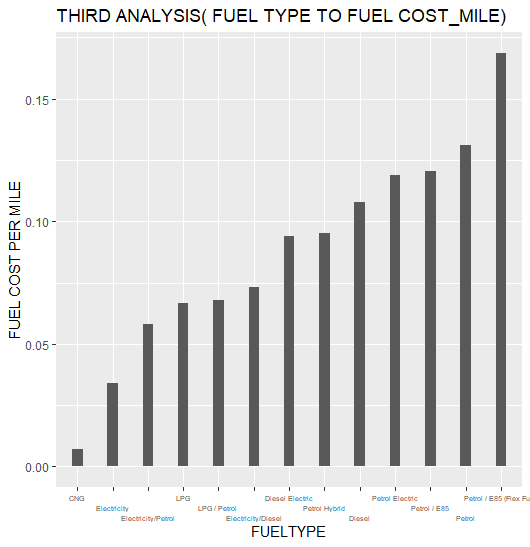


Figure 10: Q1A3 chart.

Question 2:- Which brand and which features are related to produce less noise pollution?

This analysis is executed to get the details for brands and features which relates with the noise pollution of the vehicles. Such as which brands cars have lesser noise pollution , which transmission type consumes low fuel and which fuel type is related for consumption of less noise pollution. This analysis will help user in selecting the brand and the feature types.

**ANALYSIS 1:- Analysis based on brands.**

The first   analysis looks at which makers have the lowest fuel costs. In this study, the fuel cost 1 mile vector will be utilized as a measure of cost.

Text, letter

Description automatically generated

Figure 11:- Q2A1 Code.

The code for the above-mentioned analysis is seen in Figure 11. Since there are multiple rows of same group name in Maker column, group\_by() function is used to perform the group operation. To know the mean value of noise\_level column mean function is used inside the summarise\_at() function. If the row contains Na value, it will not be included. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is sorted by increasing noise\_level using the reorder() function, and the noise level is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and width is set to 0.2. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 12 shows the graph plotted for analysis 1 which shows that maker Perodua, Daihatsu and Mazda are the three makers which has the lowest noise\_level.

Table

Description automatically generated with low confidence

Figure 12: Q2A1 chart.

**ANALYSIS 2:- Analysis based on Transmission type.**

The second analysis looks at which transmission type produces less noise . In this study, the noise\_level will be utilized as a measure of feature.

Text

Description automatically generated

Figure 13:- Q2A2 Code.

The code for the above-mentioned analysis is seen in Figure 13. Since there are multiple rows of same transmission name in transmission\_type column, group\_by() function is used to perform the group operation. Summarise\_at() function is used to get the information about the selected two columns, the information include the number of row, mean and median value. If the row contains Na value, it will not be included. Following this, the ggplot() function is called, the transmission type is represented by the x axis, which is sorted by increasing mean noise\_level, and the noise\_level is shown by the y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and the width is set as 1. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed which was implemented by label function.

Figure 14 shows the graph plotted for analysis 2 which shows the transmission type is shown by the y-axis, while the noise level is represented by the x-axis. From this chart, we can observe that automatic transmission generates lesser noise than automated transmissions.

Chart

Description automatically generated with medium confidence

Figure 14: Q2A2 chart.

**ANALYSIS 3:- Analysis based by fuel type and fuel consumption.**

The third analysis looks at which fuel type generates less noise as per noise\_level. In this study, the no8ise\_level will be utilized as a measure of parameter.

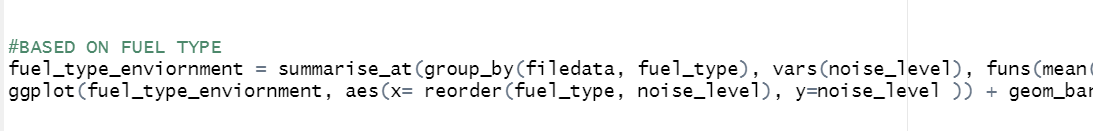


Figure 15:- Q1A3 Code.

The code for the above-mentioned analysis is seen in Figure 15. Since there are multiple rows of same fuel type in fuel\_type column, group\_by() function is used to perform the group operation. Summarise\_at() function is used to get the information about the selected two columns, the information include the number of row, mean and median value. If the row contains Na value, it will not be included. Following this, the ggplot() function is called, the fuel\_type is represented by the x axis, which is sorted by increasing mean noise level and the noise\_level is shown by the y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parameter is set to identity value and the width is set as 0.2. To adjust the x label font size theme() function was used. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed which was implemented by label function.

Figure 16 shows the graph plotted for analysis 3 which shows the fuel type is shown by the x-axis, while the noise level is represented by the y-axis. From this chart, we can observe that Electricity type generates the least noise.

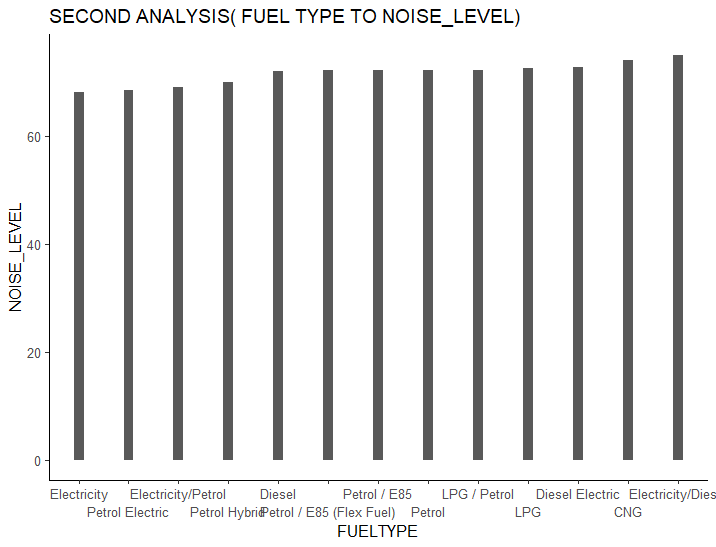


Figure 16: Q2A3 chart.

**Question 3:- Which brand and which features are related to produce less environment pollution?**

This analysis is executed to get the details for brands and features which relates with the environment pollution generates from the vehicles. Such as which brands cars generates lesser co, co2 and thc , which transmission type generates less environment pollution and which fuel type is related for generating less environment pollution. This analysis will help user in selecting the brand and the feature types.

**ANALYSIS 1:- Analysis based on brands co emission and Maker.**

The first   analysis looks at which makers have the lowest fuel costs. In this study, the fuel cost 1 mile vector will be utilized as a measure of cost.

Graphical user interface, text, application, email

Description automatically generated

Figure 17:- Q3A1 Code.

The code for the above-mentioned analysis is seen in Figure 17. Since there are multiple rows of same group name in Maker column, group\_by() function is used to perform the group operation. To know the mean value of co emission level column mean function is used inside the summarise\_at() function. If the row contains Na value, it will not be included. To sort the emission level order() function is used and stored in the sorted\_co\_level variable. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is sorted by increasing co emission level using the fct\_inorder() function, and the noise level is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and width is set to 0.1. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 18 shows the graph plotted for analysis 1 which shows that maker infiniti ,LTI and Mercedes are the three makers which has the lowest co emission level. In which Infiniti produces the least co emission level.

A picture containing chart

Description automatically generated

Figure 18: Q3A1 chart.

**ANALYSIS 2:- Analysis based on Maker type and Co2 emission.**

The second analysis looks at which Maker type produces less Co2 emission. In this study, the Co2 emission will be utilized as a measure of feature.

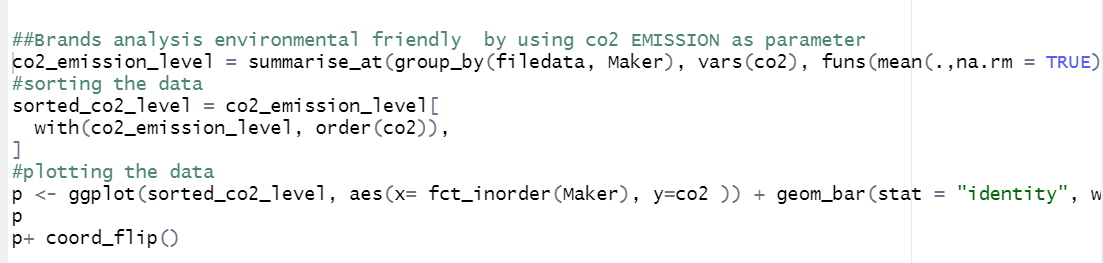


Figure 19:- Q3A2 Code.

The code for the above-mentioned analysis is seen in Figure 19. Since there are multiple rows of same group name in Maker column, group\_by() function is used to perform the group operation. To know the mean value of co2 emission level column mean function is used inside the summarise\_at() function. If the row contains Na value, it will not be included. To sort the co2 emission level order() function is used and stored in the sorted\_co2\_level variable. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is sorted by increasing co2 emission level using the fct\_inorder() function, and the noise level is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and width is set to 0.1. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 20 shows the graph plotted for analysis 2 which shows that maker smart ,micro and dacia are the three makers which has the lowest co emission level. In which Smart produces the least co2 emission level.

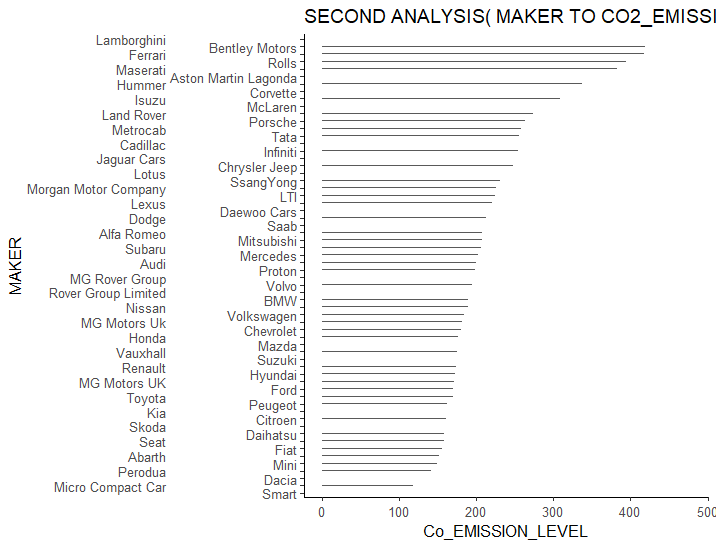


Figure 21: Q3A2 chart.

**ANALYSIS 3:- Analysis based on Maker type and thc emission.**

The third  analysis looks at which Maker type produces less thc emission. In this study, the thc emission will be utilized as a measure of feature.

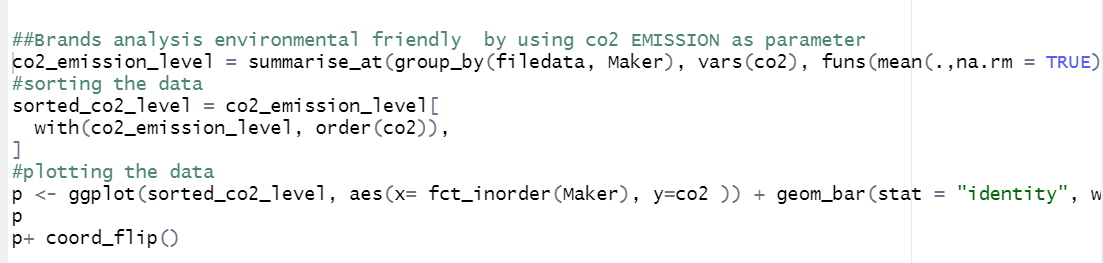


Figure 22:- Q3A2 Code.

The code for the above-mentioned analysis is seen in Figure 22. Since there are multiple rows of same group name in Maker column, group\_by() function is used to perform the group operation. To know the mean value of thc emission level column mean function is used inside the summarise\_at() function. If the row contains Na value, it will not be included. To sort the thc emission level order() function is used and stored in the sorted\_thc\_level variable. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is sorted by increasing co2 emission level using the fct\_inorder() function, and the noise level is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and width is set to 0.1. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 23 shows the graph plotted for analysis 3 which shows that maker metrocab ,infiniti and Abarth are the three makers which has the lowest thc emission level. In which metrocab produces the least thc emission level.

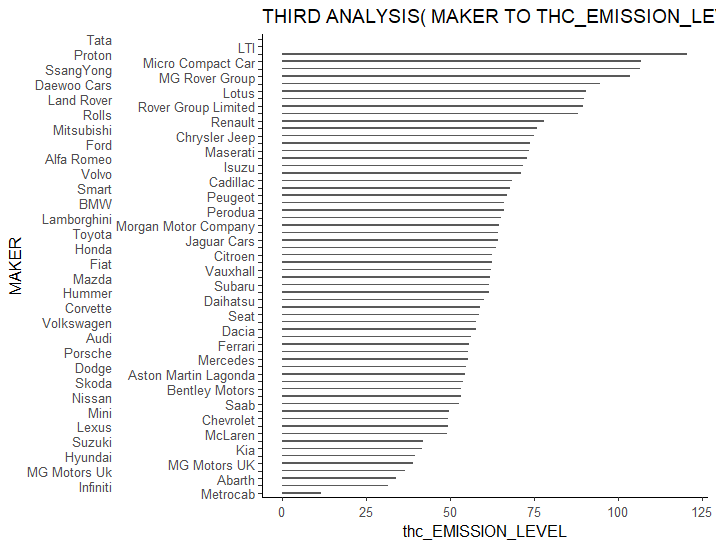


Figure 23: Q3A3 chart.

**Question 4:- Which brand are related to provide great suitability to users?**

This analysis is executed to get the details for brands which relates with the suitable as per user. Such as which brands has most powerful engine, which brands have number of transmission and which brands have higher euro standards. This analysis will help user in selecting the brand and the feature types.

**ANALYSIS 1:- Analysis based on brands and Engine capacity.**

The first   analysis looks at which makers have most powerful engine or higher capacity engine. In this study, the engine\_capacity will be utilized as a measure of cost.

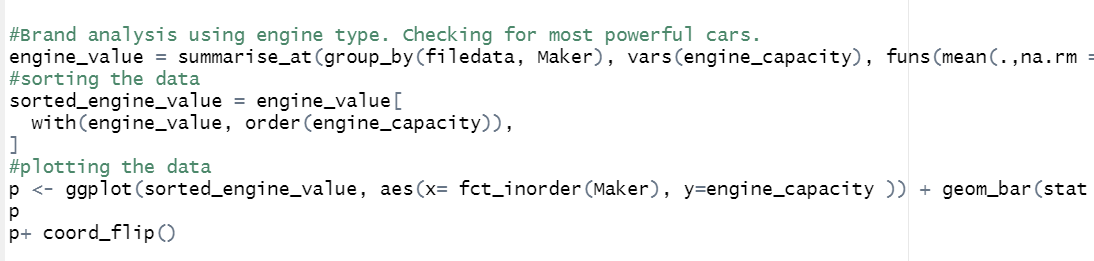


Figure 24:- Q4A1 Code.

The code for the above-mentioned analysis is seen in Figure 24. Since there are multiple rows of same group name in Maker column, group\_by() function is used to perform the group operation. To know the mean value of engine capacity column mean function is used inside the summarise\_at() function. If the row contains Na value, it will not be included. To sort the engine\_capacity order() function is used and stored in the sorted\_engine\_value variable. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is sorted by increasing engine\_capacity level using the fct\_inorder() function, and the engine capacity is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and width is set to 0.1. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 25 shows the graph plotted for analysis 1 which shows that maker rolls, Bentley motors and corvette are the three makers which has the most powerful engine capacity models. While Micro, smart and Perodua has the lowest engine capacity car models.

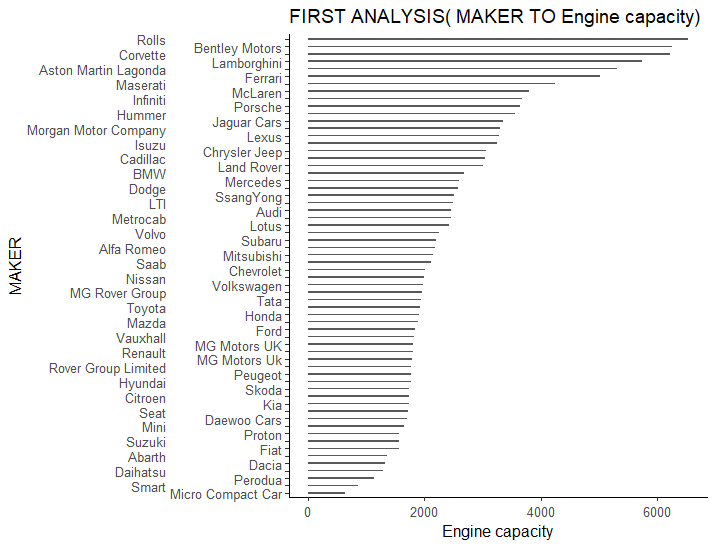


Figure 25: Q4A1 chart.

**ANALYSIS 2:- Analysis based on Maker type and number of transmission type.**

The second analysis looks at which Maker type has more number of transmission type. In this study, number of transmission type will be utilized as a measure of feature.

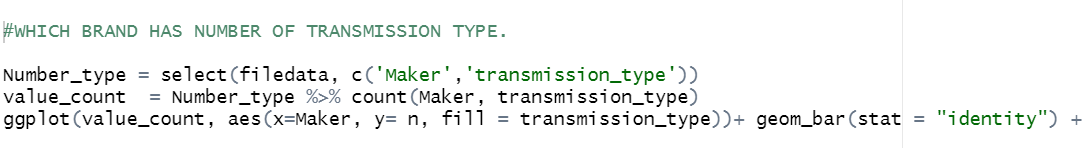


Figure 26:- Q4A2 Code.

The code for the above-mentioned analysis is seen in Figure 26. Select function was used for selection of the two columns maker and transmission type. Count() function is used to count the number of makers and their transmission type. “Following this, the ggplot() function is called, the maker is represented by the x axis, which is, and the count is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parmeter is set to identity value and width is set to 0.1. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 27 shows the graph plotted for analysis 2 which shows that Mercedes has most number of cars and more than half of the models are automatic.

Chart, histogram

Description automatically generated

Figure 27: Q4A2 chart.

**ANALYSIS 3:- Analysis based on Maker and euro standards.**

The third  analysis looks at which Maker type has highest number of euro standards car. In this study, the euro standard will be utilized as a measure of feature.

A picture containing text

Description automatically generated

Figure 28:- Q4A3 Code.

The code for the above-mentioned analysis is seen in Figure 28. Select function was used for selection of the two columns maker and euro standards. Count() function is used to count the number of makers and their euro standard. “Following this, the ggplot() function is called, the maker is represented by the x axis, and the count is shown in y axis. geom\_bar() function is used to display the data in the form of bar chart where the stat parameter is set to identity value and width is set to 0.1. Scale\_x\_discrete() function was used to stop the overlapping of x labels. It is then necessary to insert text within the bars, and finally, aesthetic qualities and labels must be developed.

Figure 29 shows the graph plotted for analysis 2 which shows that Audi has most number of 6 euro standard cars, while Mercedes has most number of 5 and 4 euro standard cars.

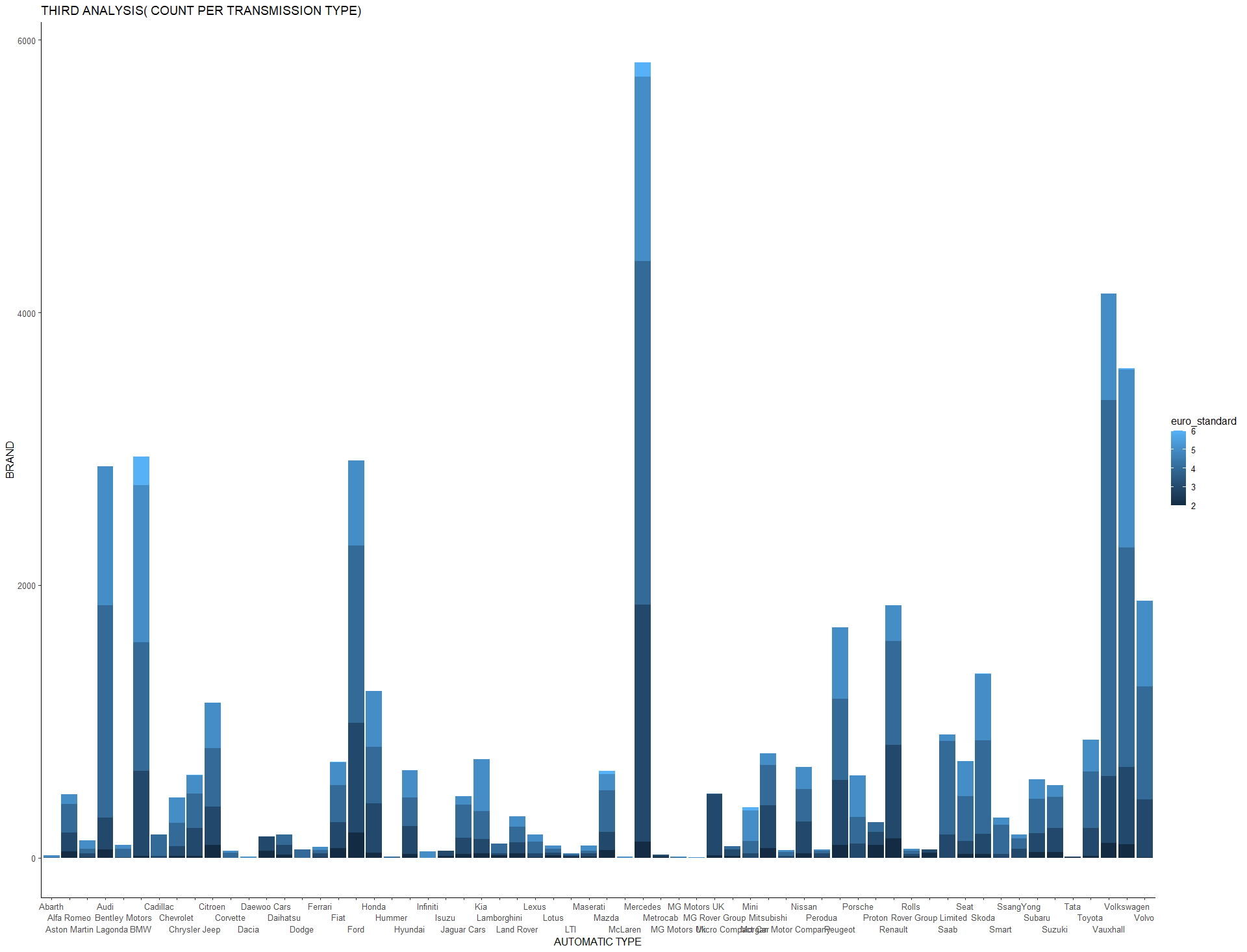


Figure 29: Q4A3 chart.

# CONCLUSION

The assignment requires the data to be imported, processed and analyzed. There were four assumptions that was identified, these assumptions were related to how a user would have choice of selection and know the impact of brands on overall experience. There were multiple columns which doesn’t have much data in it, those columns were removed during the data scrubbing process. All the four assumptions resulted into analyzation of several results. The code and analysis were explained for each assumptions. Overall, all the objectives were achieved and the assignment was made as required.

# 4 References

Atabani, A.E., Badruddin, I.A., Mekhilef, S. and Silitonga, A.S., 2011. A review on global fuel economy standards, labels and technologies in the transportation sector. *Renewable and Sustainable Energy Reviews*, *15*(9), pp.4586-4610.

Alessandrini, A., Cattivera, A., Filippi, F. and Ortenzi, F., 2012, August. Driving style influence on car CO2 emissions. In 2012 international emission inventory conference.

Orecchini, F. and Sabatini, D., 2003. Cars and the environment: a new approach to assessment through ISO 14001 certification of the car process. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *217*(1), pp.31-40.