*Applied Data Analytics*

*By*

*Onome Okobiebi*

*Group 1*

Table of Contents

[Research Question & Hypothesis 8](#_Toc144216220)

[Task two: Dataset and Data Preparation (10%) 8](#_Toc144216221)

[Data importation 9](#_Toc144216222)

[9](#_Toc144216223)

[Data Cleaning 9](#_Toc144216224)

[Incorrect Data and Missing Data 9](#_Toc144216225)

[Duplicates 11](#_Toc144216226)

[Outliers’ detection 12](#_Toc144216227)

[Outlier City FE 13](#_Toc144216228)

[Outlier highway FE 16](#_Toc144216229)

[Outlier combined FE. 17](#_Toc144216230)

[Outlier CityC02 19](#_Toc144216231)

[Highway Outlier 21](#_Toc144216232)

[Combined CO2 23](#_Toc144216233)

[Task 3 25](#_Toc144216234)

[Data Analytical Methods 25](#_Toc144216235)

[Research questions 25](#_Toc144216236)

[Variable type 25](#_Toc144216237)

[The Measure of Central Tendency 25](#_Toc144216238)

[Definition of terms 25](#_Toc144216239)

[What type of car provides the most fuel economy? 30](#_Toc144216240)

[Research Questions 1 33](#_Toc144216241)

[Research Questions 2 35](#_Toc144216242)

[Task four: Evaluation and Conclusion 36](#_Toc144216243)

[Evaluation 36](#_Toc144216244)

[RQ1 36](#_Toc144216245)

[RQ2 36](#_Toc144216246)

[Conclusions 36](#_Toc144216247)

[Section 2: Data Modelling 38](#_Toc144216248)

[Tables 38](#_Toc144216249)

[Entity Relationship Diagram (ERD) 38](#_Toc144216250)

[Design Procedure. 38](#_Toc144216251)

[Database and Table Creation. 39](#_Toc144216252)

[The Customer Table 40](#_Toc144216253)

[Data type and constraints 40](#_Toc144216254)

[The item Entity 42](#_Toc144216255)

[Data type 43](#_Toc144216256)

[Constraints 43](#_Toc144216257)

[The Drivers Entity. 43](#_Toc144216258)

[Data type 44](#_Toc144216259)

[Constraints 44](#_Toc144216260)

[45](#_Toc144216261)

[Task 3 46](#_Toc144216262)

[The Importance of Database Design to Kangaroo delivery services. 46](#_Toc144216263)

[Supporting Kangaroo's Business Scenario: 46](#_Toc144216264)

[References 48](#_Toc144216265)

[Appendixes 49](#_Toc144216266)

**List of Figures**

[Figure 1: CO2 Emissions from Car exhaust 7](#_Toc144216325)

[Figure 2: Shows the SPSS Data page 8](#_Toc144216326)

[Figure 3: importation 9](#_Toc144216327)

[Figure 4: Missing data in for the Mfrn name columns 10](#_Toc144216328)

[Figure 5: Summary table to check data missing data 10](#_Toc144216329)

[Figure 6:(a) showing dealing with incorrect data. 10](#_Toc144216330)

[Figure 7: Showing the process of filtering the data set. 11](#_Toc144216331)

[Figure 8: Duplicates identification and removed. 11](#_Toc144216332)

[Figure 9: the boxplot for City FE Outliers based other factors. 13](#_Toc144216333)

[Figure 10 14](#_Toc144216334)

[Figure 11: Showing the Selection of cases to 6 Cylinders. 14](#_Toc144216335)

[Figure 12: filtering outliers based on conditions and finding means to replace them. 15](#_Toc144216336)

[Figure 13: boxplot for highway FE Outliers based other factors. 16](#_Toc144216337)

[Figure 14: No outlier detected in highway FE. 16](#_Toc144216338)

[Figure 15: Boxplot for Outlier detection for Combined FE 17](#_Toc144216339)

[Figure 16: Showing the histogram and box plot 18](#_Toc144216340)

[Figure 17: filtering data transmission and cylinders and replacing with mean of 27. 18](#_Toc144216341)

[Figure 18: Boxplot for Outlier detection for City CO2 based on factors. 19](#_Toc144216342)

[Figure 19: Shows the histogram and box plot 20](#_Toc144216343)

[Figure 20: Boxplot for Outlier detection for High CO2 based on factors. 21](#_Toc144216344)

[Figure 21: shows the histogram and boxplot 21](#_Toc144216345)

[Figure 22: Outlier Filter on cylinder = 8 and highway CO2 all General Motors 22](#_Toc144216346)

[Figure 23: Boxplot for Outlier detection for Combined CO2 based on factors. 23](#_Toc144216347)

[Figure 24: Shows the histogram distribution 24](#_Toc144216348)

[Figure 25: Outlier Filter on cylinder = 8 and combined CO2 all General Motors 24](#_Toc144216349)

[Figure 26: Summary measures of central tendencies and dispersion for the data set 26](#_Toc144216350)

[Figure 27: data shape to show the tailed direction of the data 26](#_Toc144216351)

[Figure 28: showing the distribution number of Gears. 27](#_Toc144216352)

[Figure 29: showing the distribution of cylinders in the data. 27](#_Toc144216353)

[Figure 30: the distribution of transmission. 28](#_Toc144216354)

[Figure 31: Distribution of Aspiration Method 28](#_Toc144216355)

[Figure 32: Manufacturer distribution 29](#_Toc144216356)

[Figure 33: Drive Distribution 29](#_Toc144216357)

[Figure 34: CO2 emissions of cars released in 2014 30](#_Toc144216358)

[Figure 35: showing the distribution of fuel economy in Automatic and Manual cars. 30](#_Toc144216359)

[Figure 36: CO2 emission for Automatic and manual cars 31](#_Toc144216360)

[Figure 37: Air Aspiration on CO2 Emission 31](#_Toc144216361)

[Figure 38: Air Aspiration on Combined FE 32](#_Toc144216362)

[Figure 39: shows the Manufacturer name vs the CO2 emissions. 32](#_Toc144216363)

[Figure 40: figure showing the Manufacturer name vs Combined FE 33](#_Toc144216364)

[Figure 41: Summary description 34](#_Toc144216365)

[Figure 42: showing the corelation procedure. 34](#_Toc144216366)

[Figure 43: Shows the Pearsons Correlation table 35](#_Toc144216367)

[Figure 44 35](#_Toc144216368)

[Figure 45: Shows the summary of the One-way ANOVA table 36](#_Toc144216369)

[Figure 46: **Kangaroo Database ERD.** 39](#_Toc144216370)

[Figure 47:Sqlonline.com 40](#_Toc144216371)

[Figure 48: showing the SQL script for creating the customer table. 40](#_Toc144216372)

[Figure 49: Showing the filling of the Customer table. 41](#_Toc144216373)

[Figure 50: Output from filling the Customer table. 42](#_Toc144216374)

[Figure 51: Syntax for creating the Item entity. 42](#_Toc144216375)

[Figure 52: syntax for filling item entity 43](#_Toc144216376)

[Figure 53: this is the syntax for the creation of the Drivers table. 44](#_Toc144216377)

[Figure 54: Syntax for adding column to the Drivers table. 45](#_Toc144216378)

[Figure 55: Syntax for filling the Driver table. 45](#_Toc144216379)

**List of tables**

[Table 1: showing the research questions. 8](#_Toc144216380)

**Section 1: Statistical Data Analysis**

**Task one: Domain Understanding and Research Questions**

The dataset contains information about car types, transmission types, Engine Displacement, Cylinders, fuel economy, etc. The application domain of the dataset is fuel economy and the environmental impact of automobile fuel usage and its impact on air quality.

The significance of increasing vehicle fuel energy efficiency, or fuel economy, has never been stronger due to the quick depletion of natural resources like fossil fuels and natural gas. To lessen escalating air pollution and climate change, it's also crucial to reduce fuel usage and emissions from fossil fuel-powered automobiles. (Lee and Choi, 2017).

One of the first studies to determine fuel economy was done in the US which was an event that took place every year from 1936 by Mobil (Paul et al., 2003). Since then, manufacturers have explored different technologies based on the Volpe model to enhance vehicle fuel efficiency. The DOT/NHTSA classifies these into engine, transmission, hybrid/electrification, and "vehicle" tech (e.g., weight reduction, aerodynamics). To boost fuel efficiency, manufacturers have recently added more gears and innovations like dual-clutch transmission (DCT) (Moawad and Rousseau, 2012).



Figure 1: CO2 Emissions from Car exhaust

Studies by Moawad and Rousseau, 2012) showed that in the case of a mid-size car, the automatic transmission enhances fuel efficiency by 6.9%, while the DCT path boosts it by 10.3%. For a compact car, the transmission leads to an 11% fuel consumption improvement. A small SUV benefits from an 8.5% fuel consumption enhancement due to the transmission. Meanwhile, a mid-size SUV experiences a 5.5% fuel consumption improvement, and a pickup truck sees up to 6.3% enhanced fuel efficiency through the transmission.

Dias (2011) showed that the efficiency of a CVT is 10% to -20% lower than an automatic transmission by planetary. Using INMETRO's homologation data covering the last decade in the Brazilian market, it's evident that Continuously Variable Transmission (CVT) transmissions have demonstrated fuel consumption benefits. However, recently introduced planetary transmissions have exhibited favourable consumption figures specifically in highway conditions (Junior, Muraria, and Guarieiro, 2021).

## Research Question & Hypothesis

According to the literature, the transmission types affect fuel efficiency in towns and highways. With other factors like the number of gears, cylinders, and brands in the table. The goal of this analysis is to find the factors or the variables that are responsible for the fuel economy and C02 emissions.

Table 1: showing the research questions.

A screenshot of a document

Description automatically generated

# Task two: Dataset and Data Preparation (10%)

This cartype dataset is publicly available provided by the university, there are also different variations of this dataset. It is a CSV file named “caretype.xls” of 213kB file size. This dataset is designed to know the Co2 emission rate and fuel economy for various car brands etc.

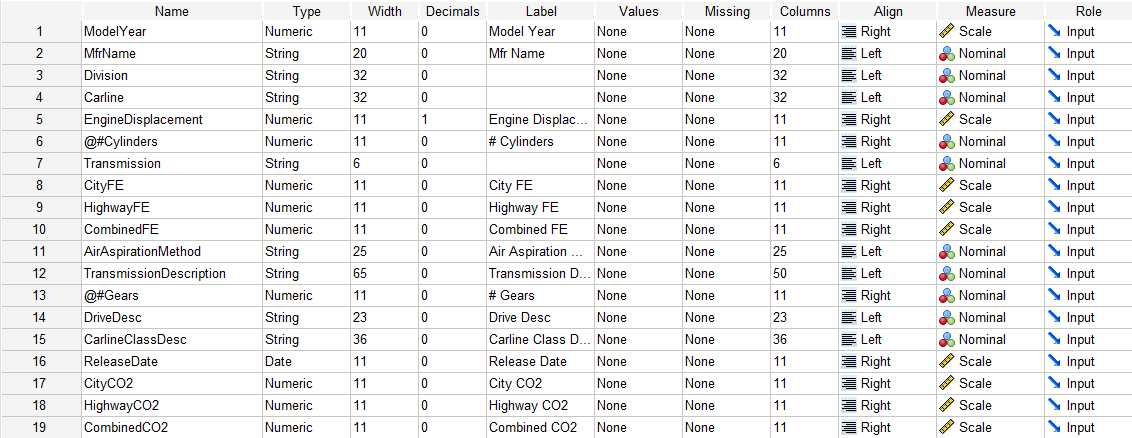


Figure 2: Shows the SPSS Data page

Data for the analysis was done through SPSS.

* Data importation
* Data pre-processing/ cleaning
* Data analysis (Descriptive and Inferential Statistical Analysis)

## Data importation

Data is imported into SPSS using the import tab and using the Excel option because it's cartype.xls.

## 

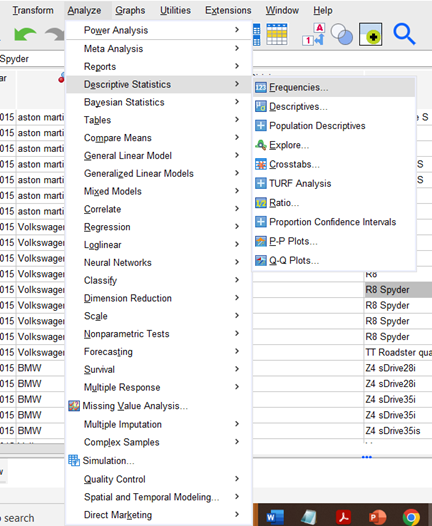
Figure 3: importation

## Data Cleaning

The data cleaning involves locating the incorrect data input by filtering the data sets and filling it based on an established dataset from the table. The missing data and incorrect analysis were done on the dataset in all columns. The first step of data cleaning is to check for incorrect data, missing data, duplicates, and outliers.

### Incorrect Data and Missing Data

The incorrect data analysis, missing data and preprocessing are done simultaneously by plotting the table of all the data input in each category and their occurrence and percentage occurrence on a table. Then duplicates were checked. The result of the analysis showed no missing data on all columns, but we noticed an incorrect data entry ‘d’ in the drive description column.

 A screenshot of a computer

Description automatically generated

Figure 4: Missing data in for the Mfrn name columns

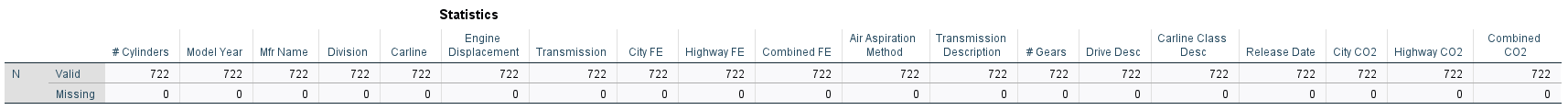


Figure 5: Summary table to check data missing data

A screenshot of a data

Description automatically generated A screenshot of a data sheet

Description automatically generated

Figure 6:(a) showing dealing with incorrect data.

* The ‘d ‘was filled using the input method to 2 Wheel Drive Rear based on data analysed from the number of carline (2 seats), and transmission (manual).

A screenshot of a computer

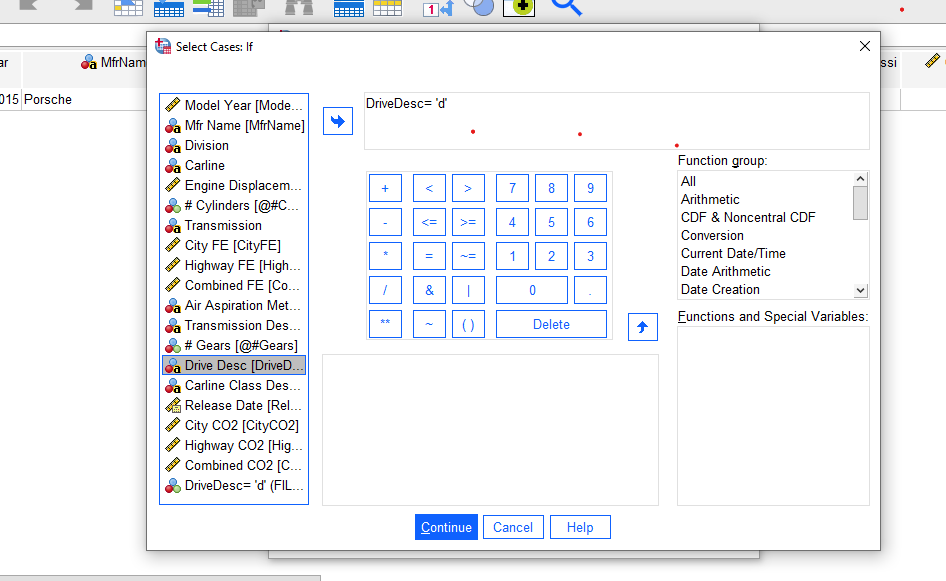
Description automatically generated 

Figure 7: Showing the process of filtering the data set.

### Duplicates

The duplicates were checked visually from the description table on all columns and input data and 7 duplicates were found.

A screenshot of a statistics

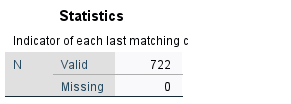
Description automatically generated

Figure 8: Duplicates identification and removed.

### Outliers’ detection

Outliers are extreme values that lie far from the other values in your data set. It is calculated based on an upper and lower boundary. Any value that is 1.5 x IQR greater than the third quartile is designated as an outlier and any value that is 1.5 x IQR less than the first quartile is also designated as an outlier. Outliers were checked for numerical data.

Note: The outlier’s detection was first done on a wholistic level using box plot as a variable then it was broken down further to narrow their occurrence based on the label of the records.

### Outlier City FE

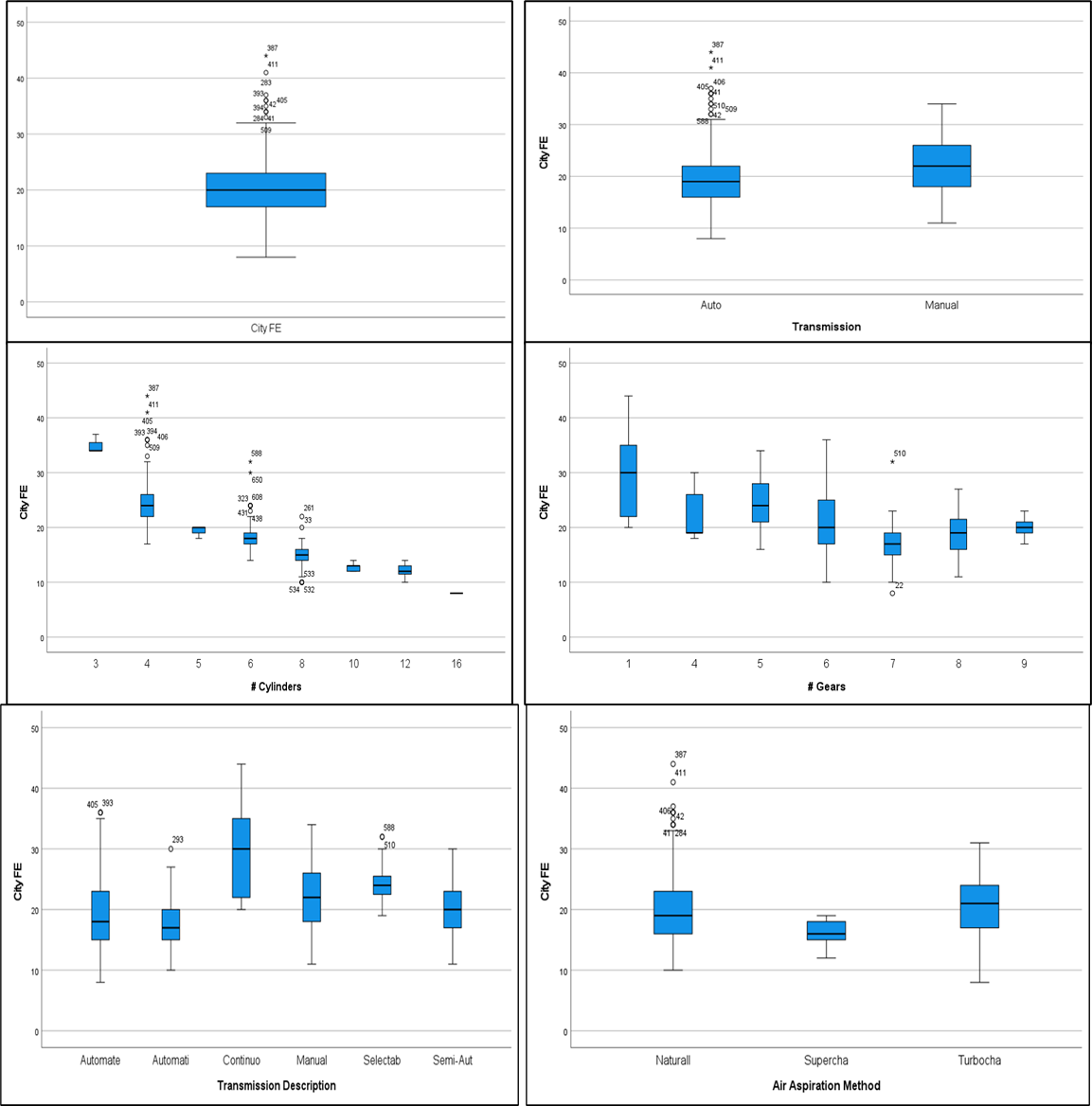


Figure 9: the boxplot for City FE Outliers based other factors.

A graph with a blue rectangle and black text

Description automatically generatedA screenshot of a graph

Description automatically generated

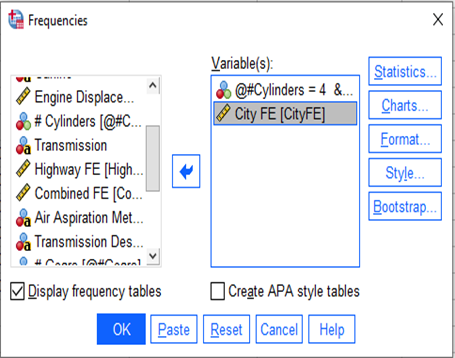
Figure 10

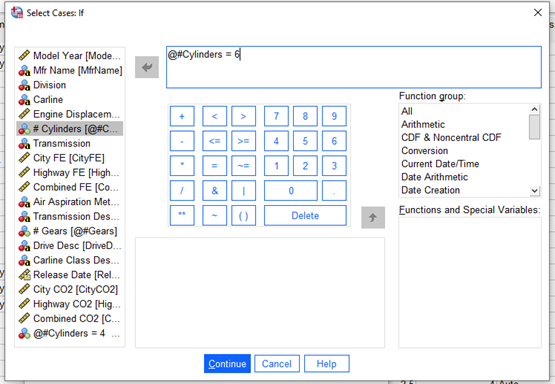
A screenshot of a computer

Description automatically generated

Figure 11: Showing the Selection of cases to 6 Cylinders.

A screenshot of a computer

Description automatically generated 

    A screenshot of a computer

Description automatically generated

Figure 12: filtering outliers based on conditions and finding means to replace them.

City FE outliers were identified at rows (411, 387) common in cars with 4 Cylinders and automatic and (588, 650) in cars with 6 cylinders. These were replaced with the means 21 and 17 based on the means of the filtered data.

### Outlier highway FE



Figure 13: boxplot for highway FE Outliers based other factors.

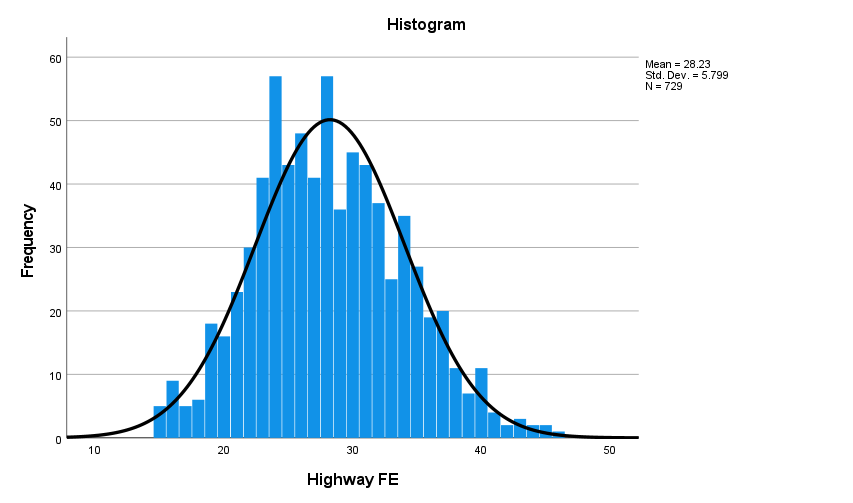
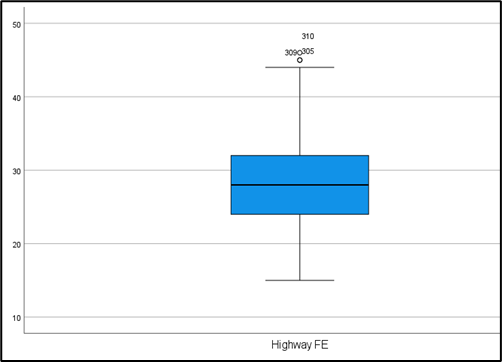
 

Figure 14: No outlier detected in highway FE.

* Highway FE no outliers detected. The presumed outliers belong to Volkswagen cars which have generally poor FE and the lower of outlier is of Bugatti which has high FE.

### Outlier combined FE.

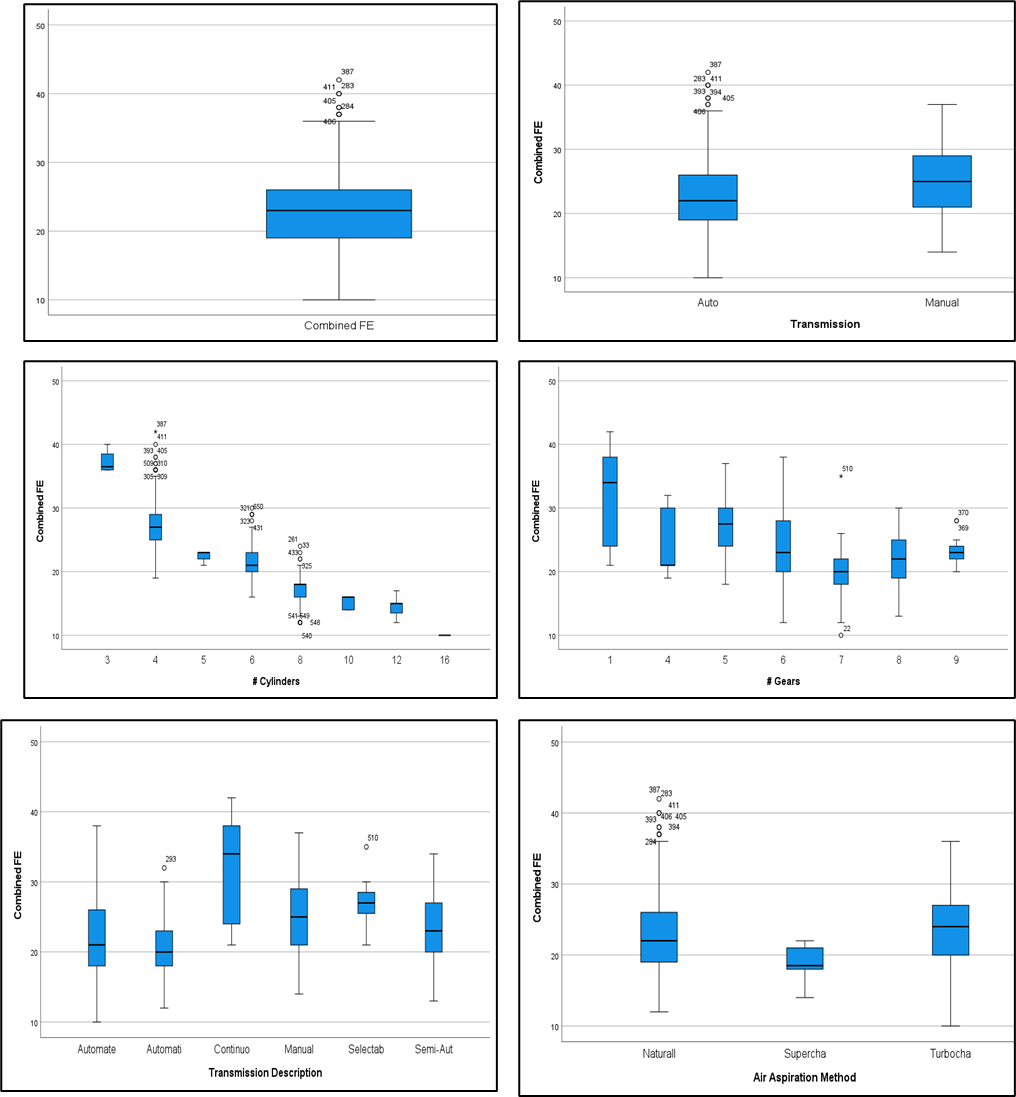


Figure 15: Boxplot for Outlier detection for Combined FE

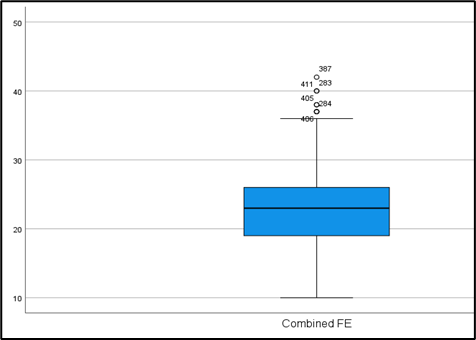
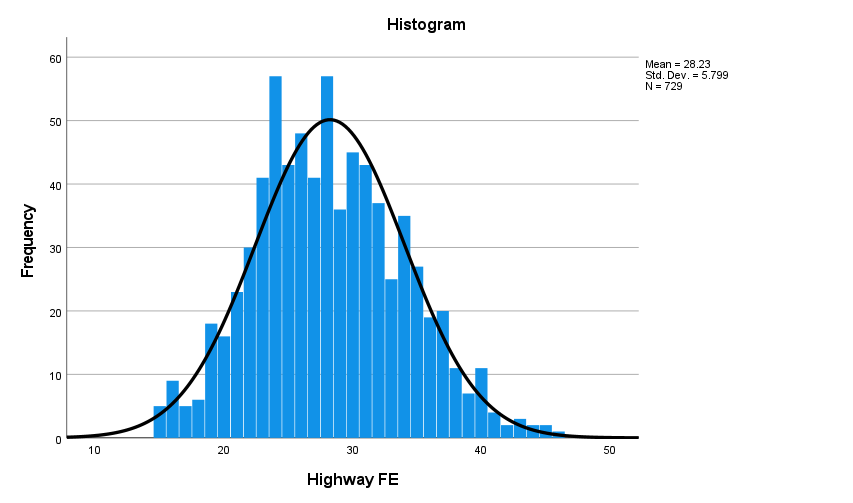


Figure 16: Showing the histogram and box plot

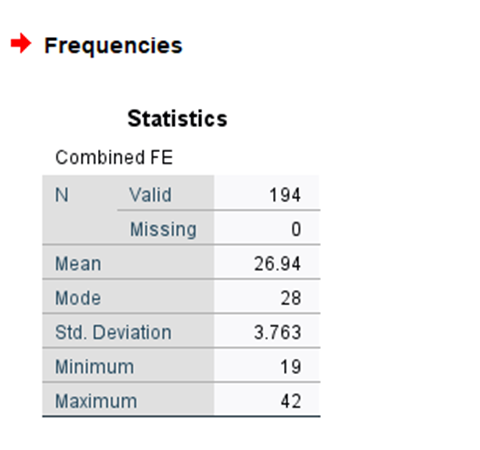
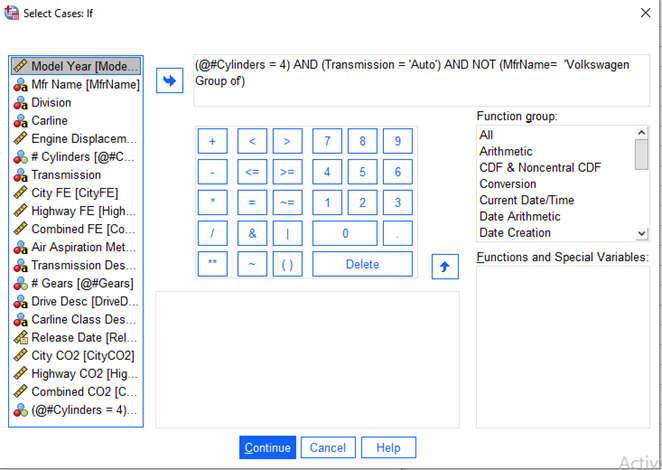


Figure 17: filtering data transmission and cylinders and replacing with mean of 27.

* Combined FE Outlier was replaced using filter (excluding Volkswagen because high FE is real data).

### Outlier CityC02

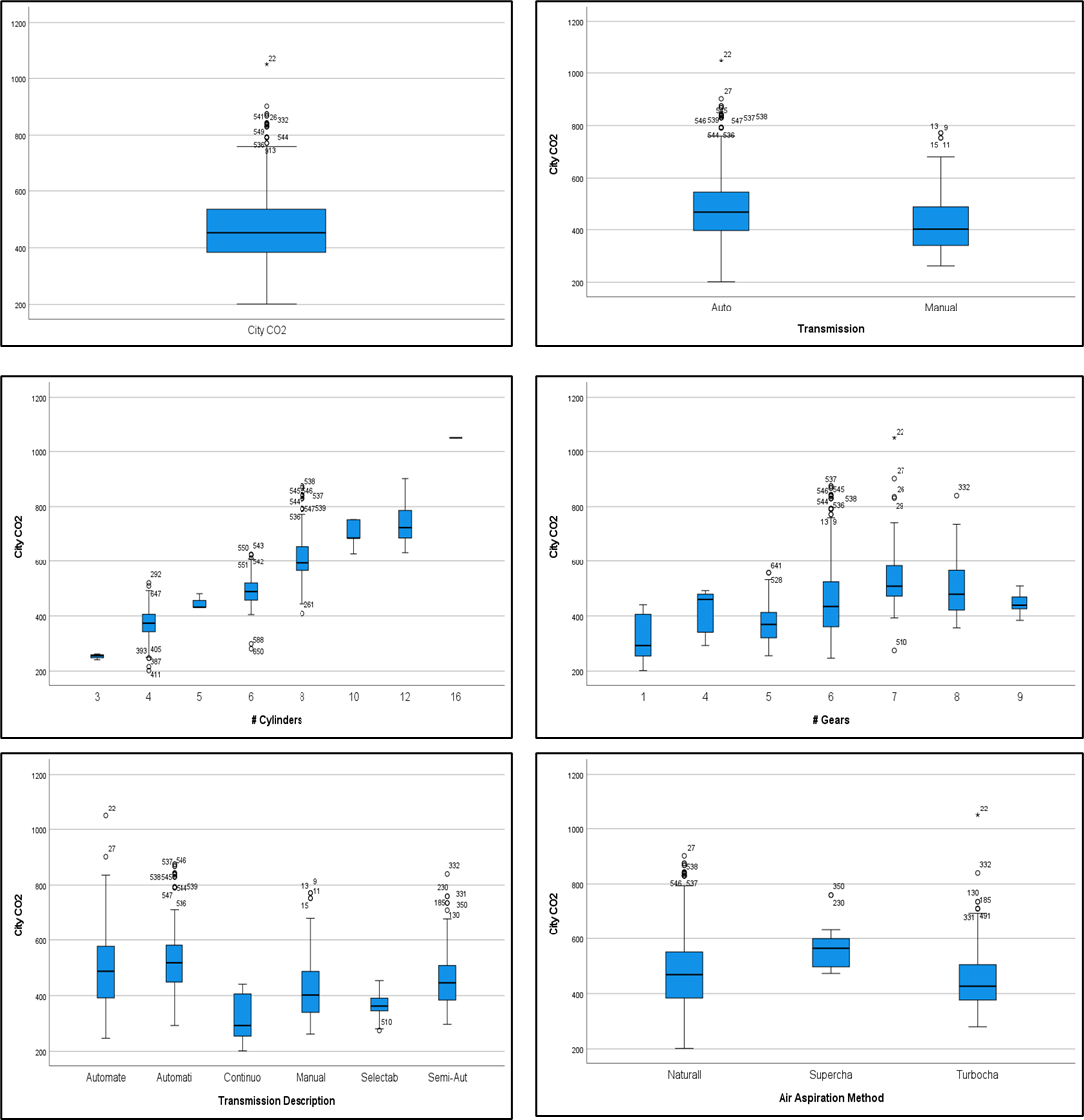


Figure 18: Boxplot for Outlier detection for City CO2 based on factors.

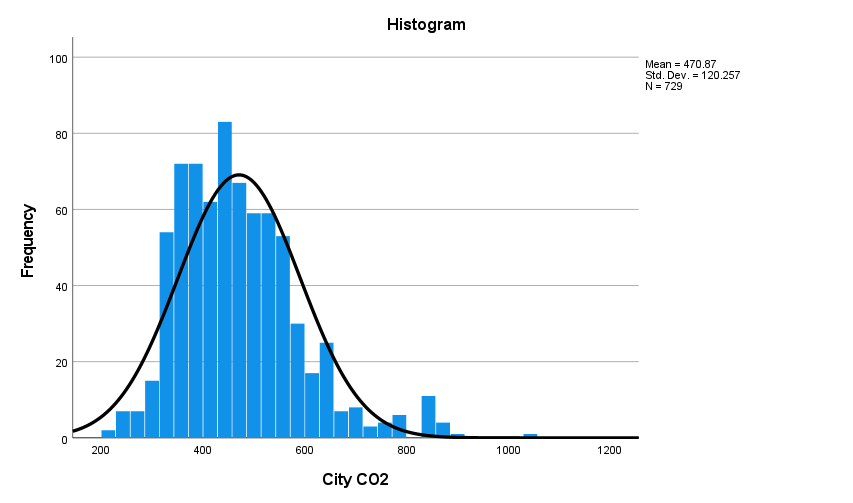
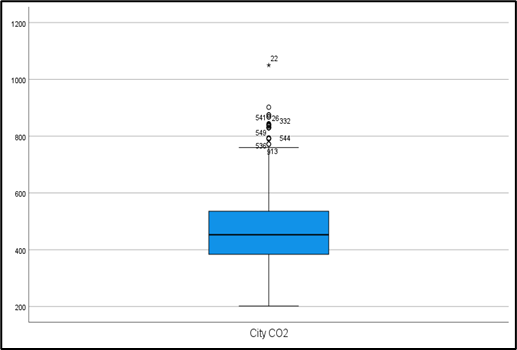
 

Figure 19: Shows the histogram and box plot

* City CO2 Outliers analysed and were all in General Motors and record (22 is replaced by /2)

### Highway Outlier

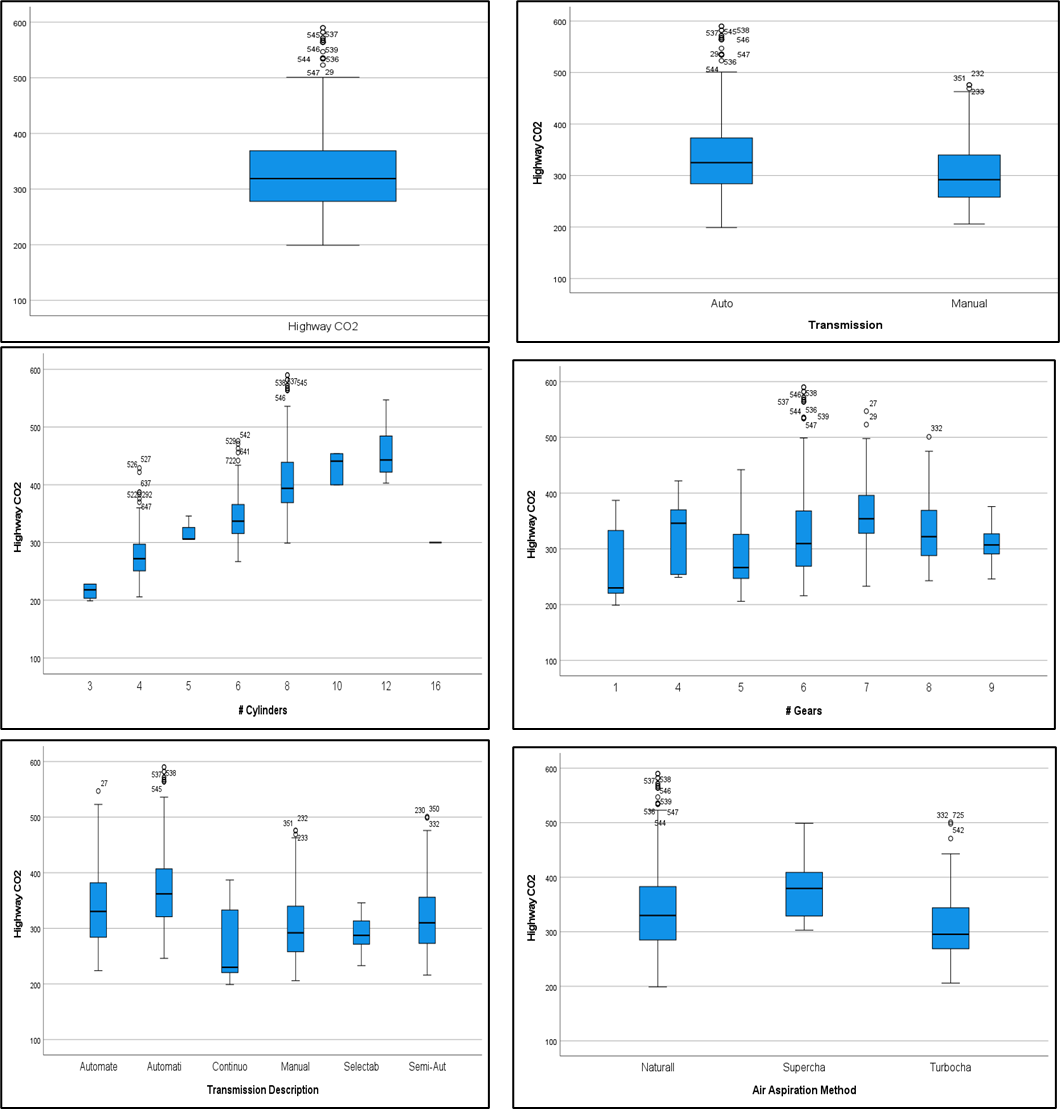


Figure 20: Boxplot for Outlier detection for High CO2 based on factors.

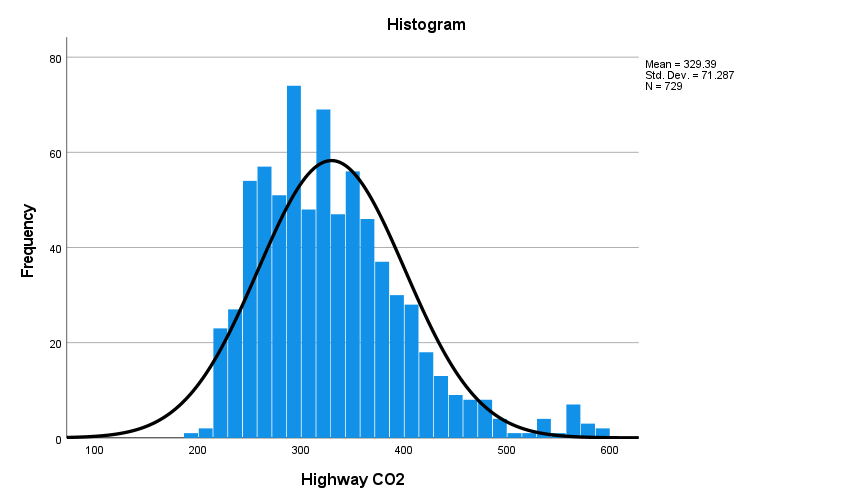
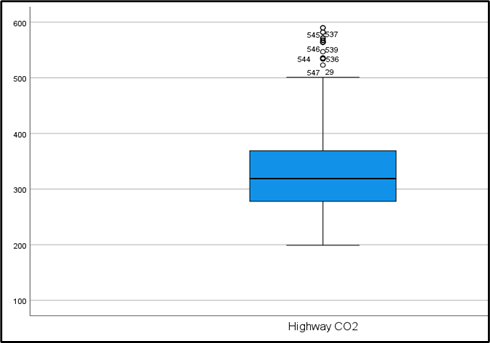
 

Figure 21: shows the histogram and boxplot

A screenshot of a computer

Description automatically generated

Figure 22: Outlier Filter on cylinder = 8 and highway CO2 all General Motors

* Highway CO2 Outliers analysed and were all in General Motors and record (22 is replaced by /2), other outliers were related to certain General Motors in Mr Name or carline class (Vans, Pick-up).

### Combined CO2

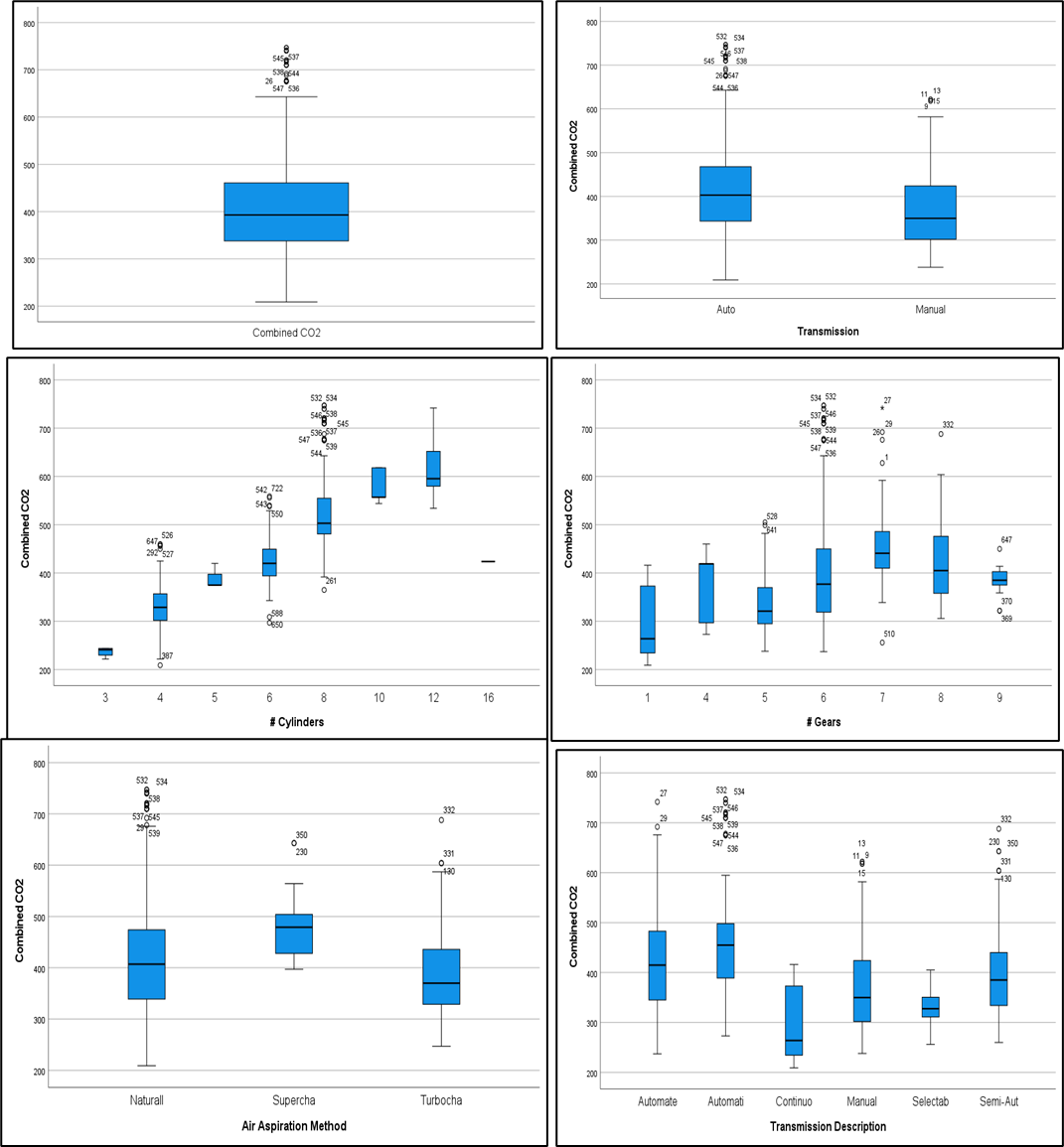


Figure 23: Boxplot for Outlier detection for Combined CO2 based on factors.

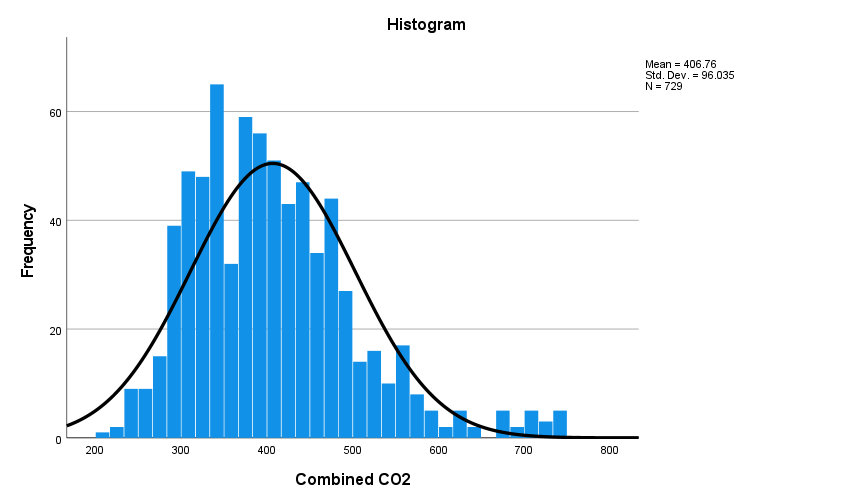


Figure 24: Shows the histogram distribution

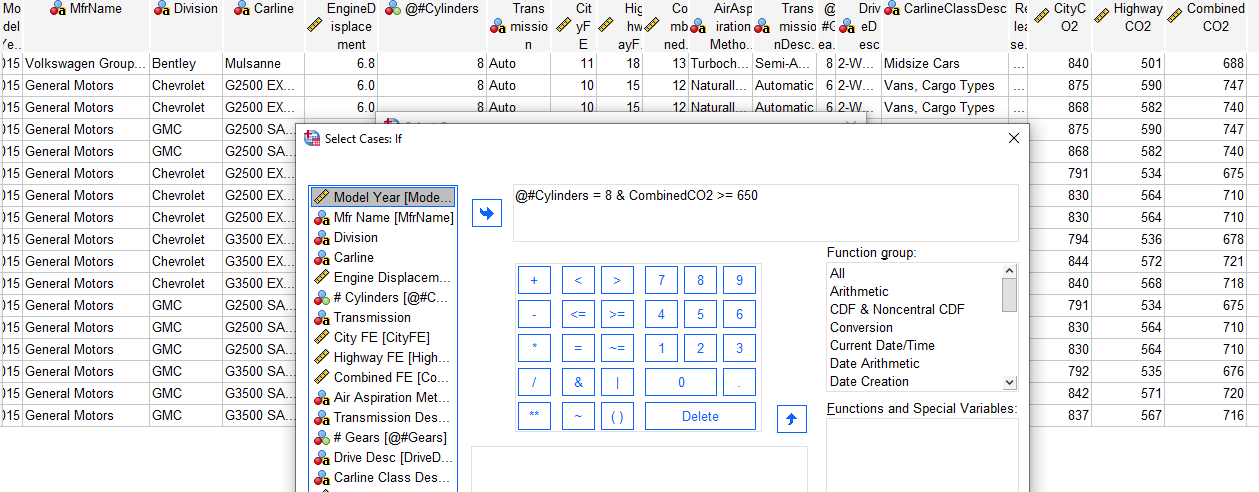


Figure 25: Outlier Filter on cylinder = 8 and combined CO2 all General Motors

Presumed outliers are real data because the high CO2 values higher than 650 are common to cars with from General Motors.

# Task 3

## Data Analytical Methods

Descriptive statistics are a critical part of initial data analysis that involves collecting, organizing, summarizing, and presenting data in a meaningful way and provide the foundation for comparing variables with inferential statistical tests. It includes different types of variables (nominal, ordinal, interval, and ratio) as well as measures of frequency ((mean, median, mode), central tendency (range, variance, standard deviation), and shape (skewness, kurtosis) (Kaur, Stoltzfus, Yellapu, 2018).

## Research questions

## Variable type

Before analysing the dataset, you have to be familiar with different types of variables. Categorical variables may be further classified as nominal, ordinal, or dichotomous.

**Nominal variables**: they are the simplest in nature, include two or more categories that lack intrinsic order (e.g., male or female).

**Ordinal variables:** they have two or more categories that can be ranked or ordered, but there is no objective value to the rankings (e.g., strongly disagree,” “agree,” and “strongly agree”).

**Continuous variables** (also known as quantitative or numerical) are further categorized as either interval or ratio.

The Measure of Central Tendency

This is used to determine the position of the dataset, and they are the mean, mode and median. The mean and median are mostly used in ratio or scaled variables, while the mode is used for the qualitative and categorical variables. It is imperative to clean the data as the mean is influenced by the presence of outliers in the data and this might skew our decisions.

**Frequency**: This is used to measure the distribution of value or range of across the dataset. It is mostly applying to categorical variables using the simple bar chart or the histogram.

**Dispersion**: these helps to quantify the spread or variability of the data. The common measures of dispersion are the Range, Variance and Standard Deviation.

**Measures of Shape**: These measures describe the distribution's characteristics. The skewness indicates the asymmetry of the data distribution and Kurtosis reflects the "tailedness" or concentration of data around the mean.

### Definition of terms

FE= fuel economy in MPG miles per gallon (higher miles covered per gallons means better fuel efficiency)

A screenshot of a computer

Description automatically generated

Figure 26: Summary measures of central tendencies and dispersion for the data set

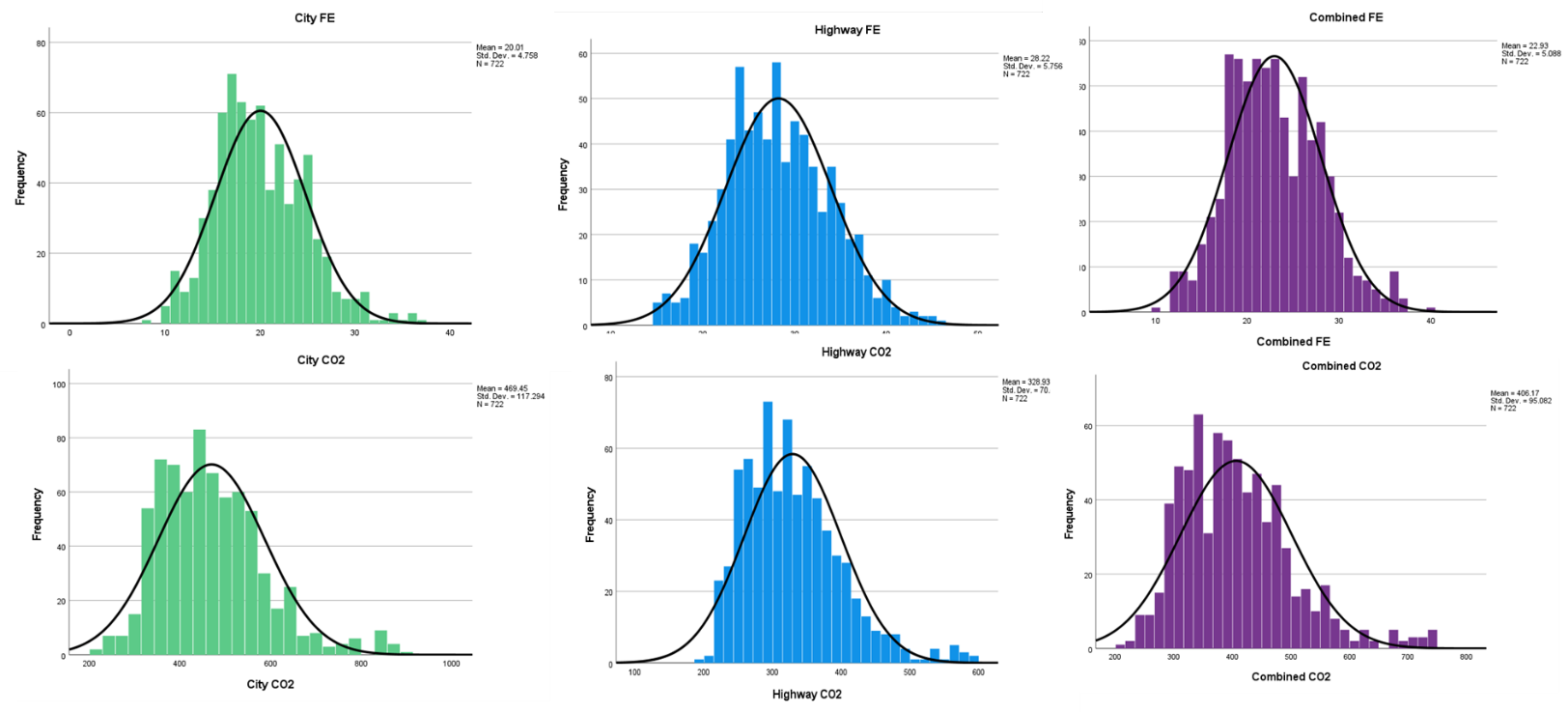


Figure 27: data shape to show the tailed direction of the data

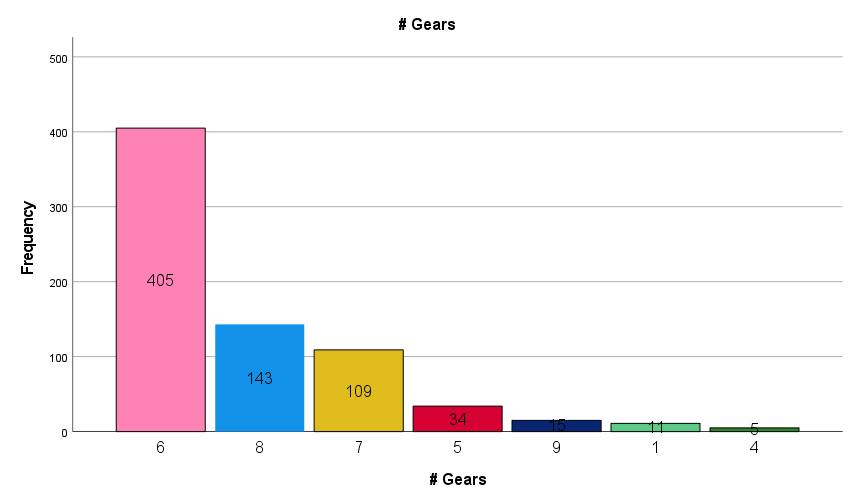


Figure 28: showing the distribution number of Gears.

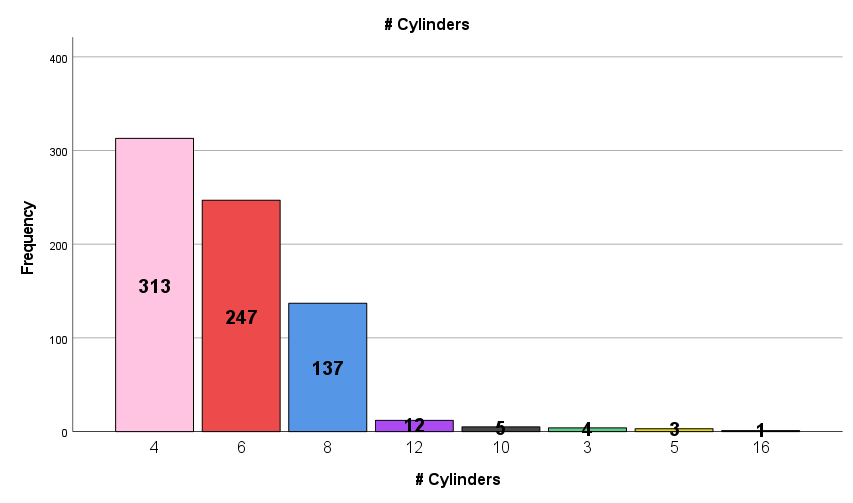


Figure 29: showing the distribution of cylinders in the data.

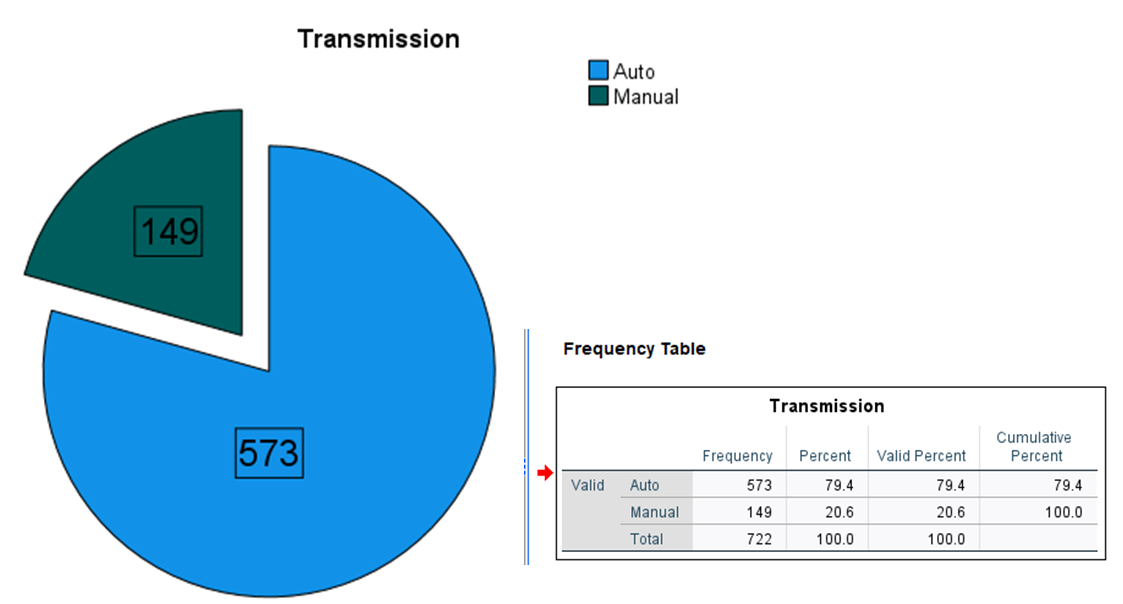


Figure 30: the distribution of transmission.

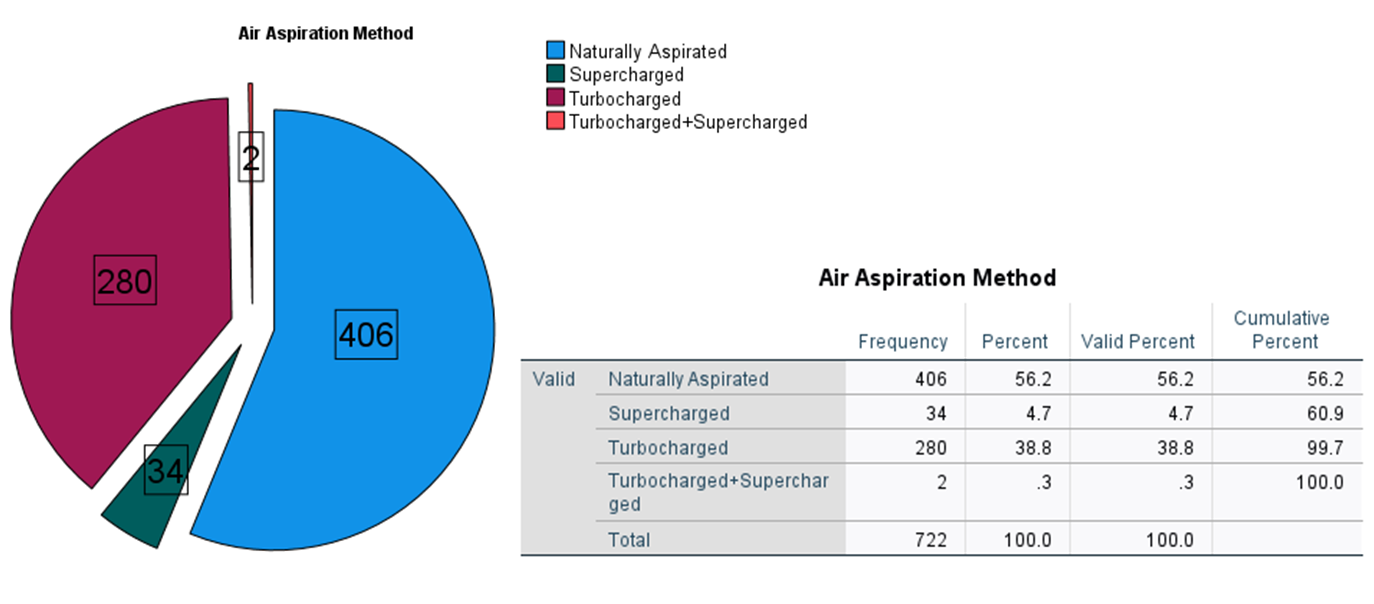


Figure 31: Distribution of Aspiration Method

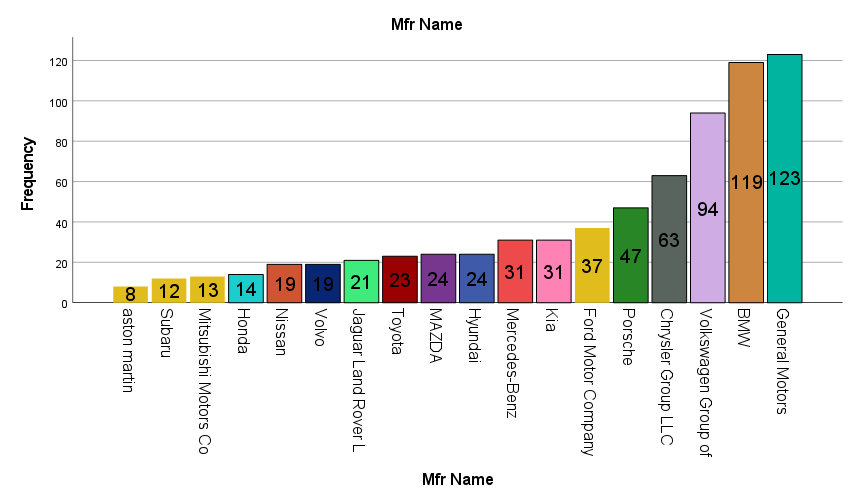


Figure 32: Manufacturer distribution

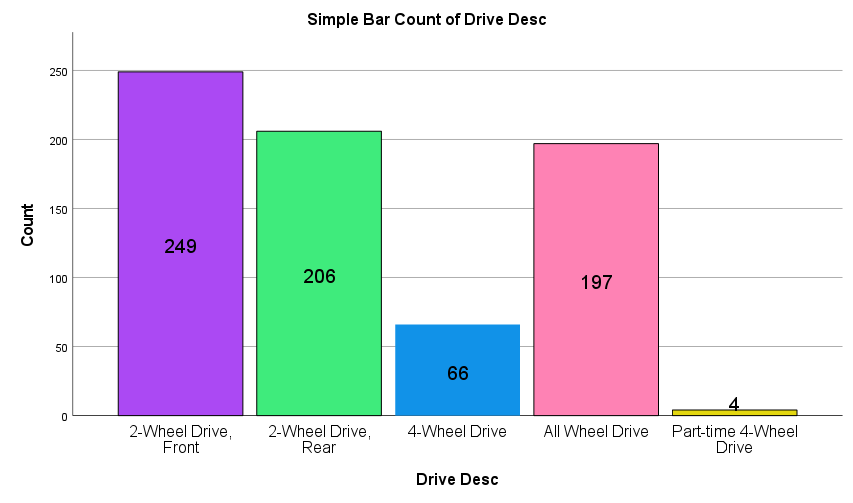


Figure 33: Drive Distribution

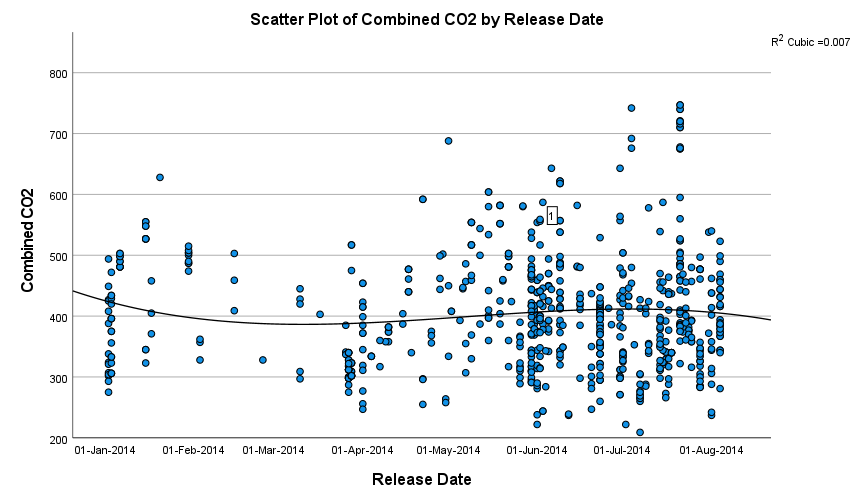


Figure 34: CO2 emissions of cars released in 2014

## What type of car provides the most fuel economy?

We are testing to check the transmission type that has the most fuel economy.

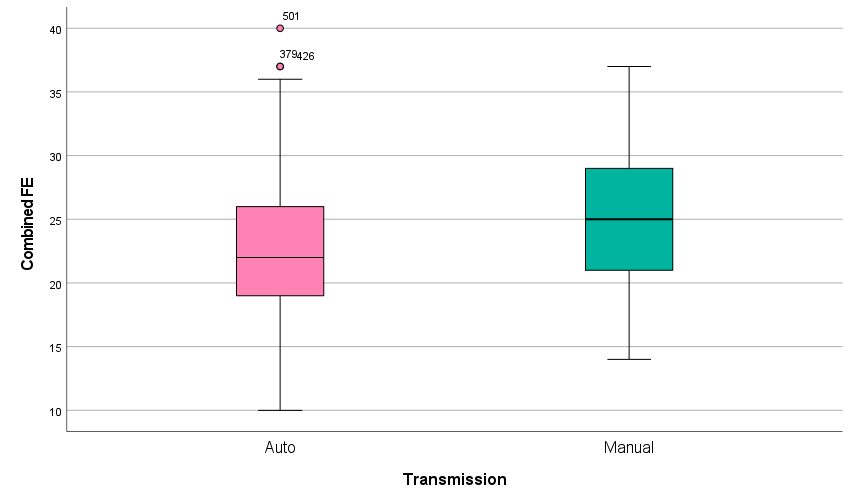
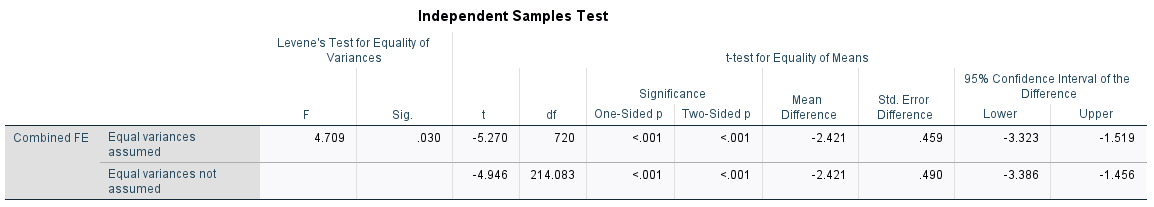


Figure 35: showing the distribution of fuel economy in Automatic and Manual cars.



In both cases, the p-values are less than 0.001, indicating strong evidence of a significant difference in means between the two groups. The negative mean differences suggest that the auto cars to have lower combined FE than the Manual cars**. This suggest that manual cars have a better combined fuel efficiency.**

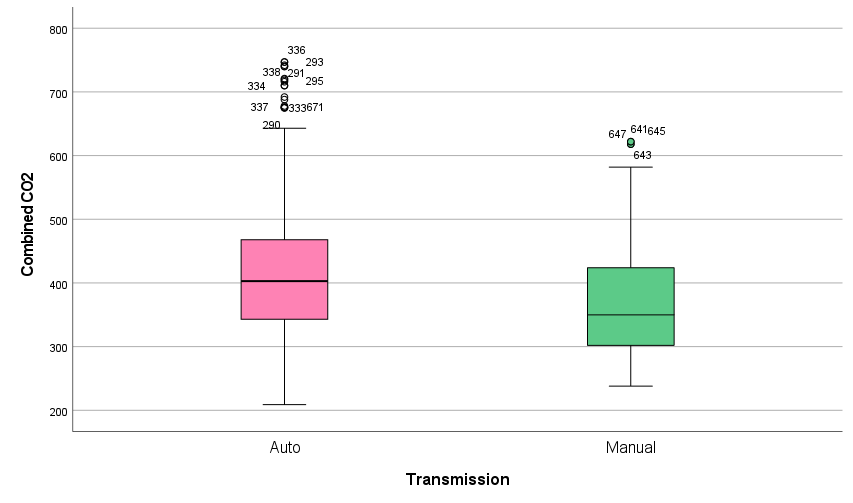
****

Figure 36: CO2 emission for Automatic and manual cars

This shows that Manual cars emits less CO2 and are more environmentally friendly.

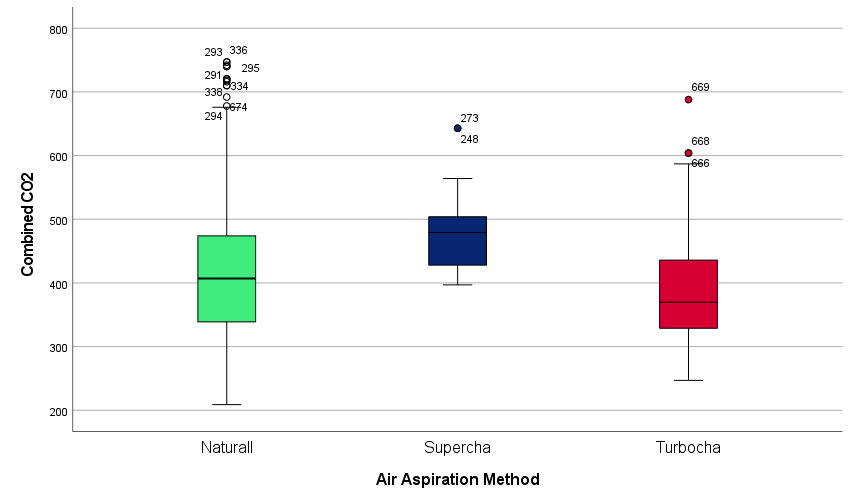
****

Figure 37: Air Aspiration on CO2 Emission

This shows that Turbocharged air aspiration method type of cars emits the least CO2 are more environmentally friendly.

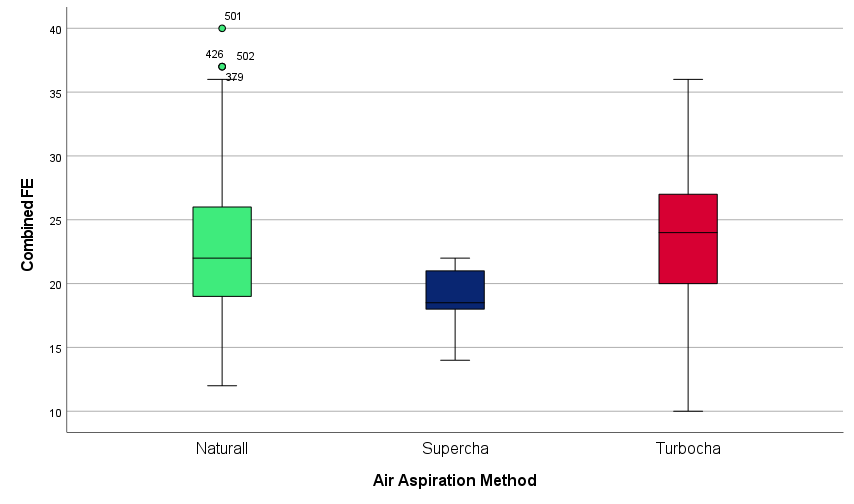
****

Figure 38: Air Aspiration on Combined FE

This shows that Turbocharged cars have the highest Combined FE and hence the most fuel efficient.

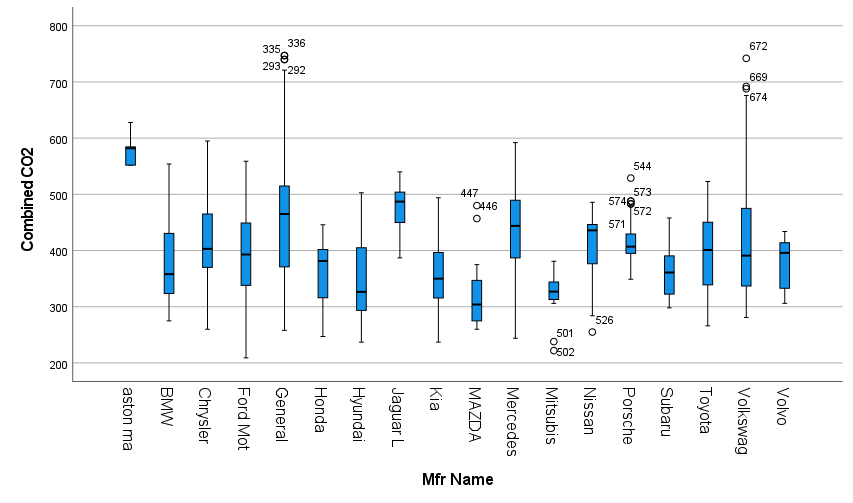
****

Figure 39: shows the Manufacturer name vs the CO2 emissions.

This shows that the Aston Martin has the highest CO2 emissions, and the Mazda has the least CO2 Emission hence the Mazda is the most environmentally friendly brand.

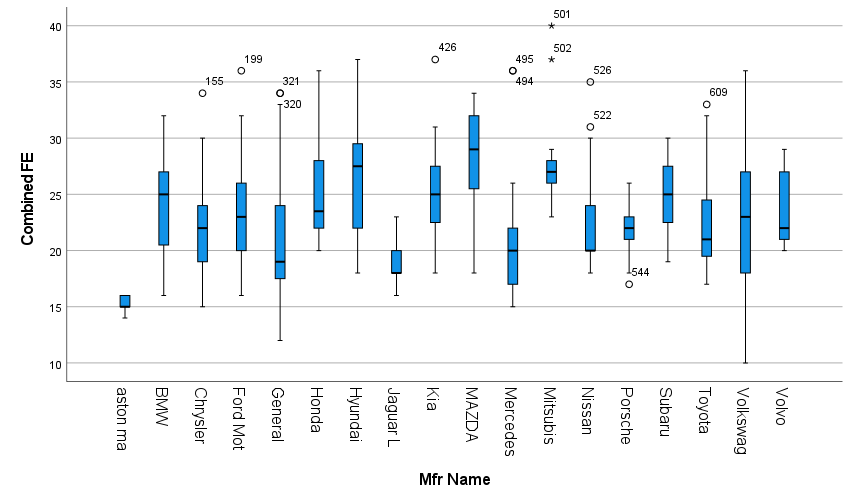
****

Figure 40: figure showing the Manufacturer name vs Combined FE

This showsAston martin has the least range of Combined FE and the Mazda has the highest range hence the Mazda is a more Fuel-efficient car.

### Research Questions 1

Is the variable number of cylinders affecting the combined City CO2 emissions of the vehicle?

H0: The number of cylinders does not significantly affect its City CO2 emissions.

H1: The number of cylinders car has a significant impact on its City CO2 emissions.

A screenshot of a computer

Description automatically generated

Figure 41: Summary description

• Analyze > Correlate > Bivariate

• Then move the two variables of interest into the variable dialogue box.

A screenshot of a computer

Description automatically generated

Figure 42: showing the corelation procedure.

A screenshot of a computer

Description automatically generated

Figure 43: Shows the Pearsons Correlation table

Used the Pearson correlation because we are analysing two numerical variables.

### Research Questions 2

Does the number of gears affect the combined City Combined FE of the vehicle?

H0: The number of gears does not significantly affect its Combined FE.

H1: The number of gears has a significant impact on its Combined FE.

A screenshot of a computer

Description automatically generated

Figure 44

A screenshot of a computer

Description automatically generated

Figure 45: Shows the summary of the One-way ANOVA table

I used the one- way ANOVA for the test because I am testing a categorical data with two or more groups against the numerical data (Ostertagová and Ostertag, 2013), as number of gears is a categorical data in this dataset.

# Task four: Evaluation and Conclusion

## Evaluation

### RQ1

The From the result obtained, the Pearson Correlation Coefficient is obtained as 0.813. This is an indication that the relationship between the number of cylinders in a car and combined CO2 is a directly relationship. This means that the higher the number of cylinders, the higher the CO2 emission.

Therefore, there is a strong relationship between the cars’ cylinders and CO2 emission.

**Thus, the Null Hypothesis is Accepted**.

### RQ2

The result shows that p-value is much smaller than 0.05 (it's <0.001), so there are statistically significant differences in means between the groups and the **null hypothesis is rejected**. The number of gears affects the fuel efficiency.

### Conclusions

In conclusion, from all the analysis carried out we can conclude that;

1. Madza is the most environmentally friendly car and has the best fuel economy.
2. Automatic cars are better fuel efficient.
3. Manual cars are more environmentally friendly.
4. Volkswagen cars have high CO2 emissions.
5. Turbo charged cars are more fuel efficient and more environmentally friendly.
6. The higher cylinders affect the CO2 emission.
7. Finally, the number of gears affects fuel efficiency.

# References

Islam, T., and Mojumder, S (2020) ‘Online Store using E-Commerce and Database Design and Implementation’, International Journal of Engineering Research & Technology (IJERT), 9(10), pp. 373-380

Junior, S., L., Muraria, T., B., Lilian and Guarieiro, L., N. (2021) ‘Proposing a Method for Assessing Fuel Consumption and Pollutants Emissions with the Use of Continuously Variable Transmission in Town Cars’ Journal of Bioengineering and Technology Applied to Health, 4(4), pp.134-140

Kaur P, Stoltzfus J, Yellapu V. (2018) “Descriptive statistics”. International Journal of Academic Medicine and Pharmacy. 4, pp. 60-63

Lee, H and Choi, H (2016). ‘Analysis of vehicle fuel efficiency and survival patterns For the prediction of total energy consumption from Ground transportation in korea’ International Journal of Automotive Technology, 17(4), pp. 605−616

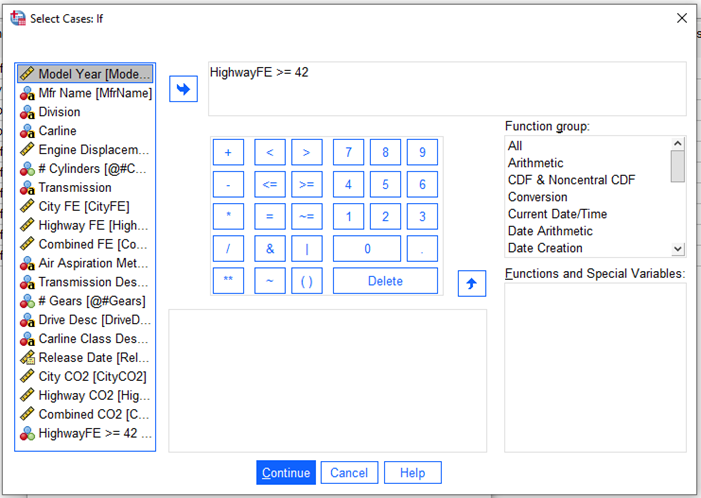
Marcelino, J (2001) Oracle Guide to Data Modeling and Relational Database Design. 1st Edition. Oracle Corporation, California

Ostertagová, E and Ostertag, O. (2013), “Methodology and Application of One-way ANOVA.” American Journal of Mechanical Engineering 1(7), pp. 256-261. doi: 10.12691/ajme-1-7-21.

Paul R. Portney; Ian W.H. Parry; Howard K. Gruenspecht; Winston Harrington (2003). "The Economics of Fuel Economy Standards" (PDF). Resources for the Future. Archived from the original (PDF) on 1 December 2007. Retrieved 4 January 2008.

# Appendixes

Indexes Support your explanation with screenshots.



A screenshot of a computer

Description automatically generated

Showing the outliers of above 42 highway FE corresponding to all Volkswagen cars

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Lower outlier in highway FE seen to be Bugatti which is a high-end vehicle with high FE

A screenshot of a computer

Description automatically generated 

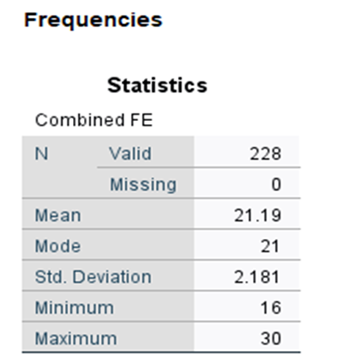
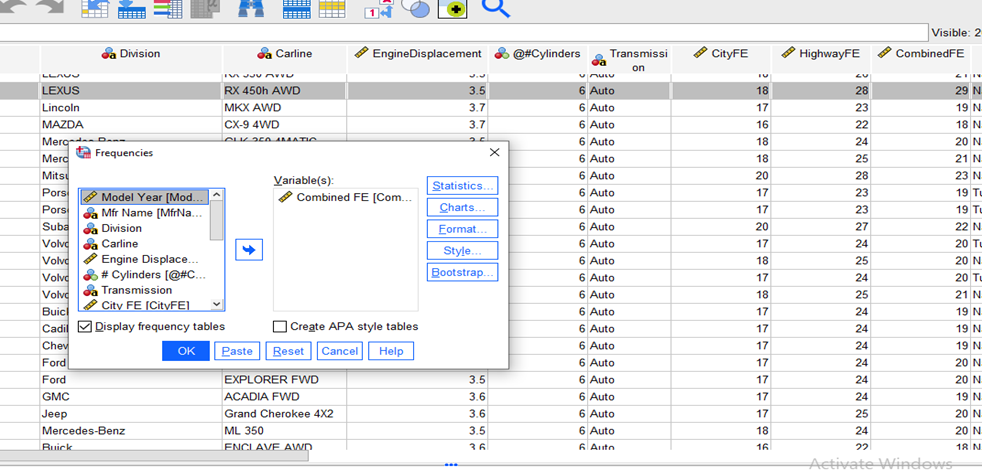
Combine FE for vehicles with Cylinder = 8 and not volkswagen

A screenshot of a computer

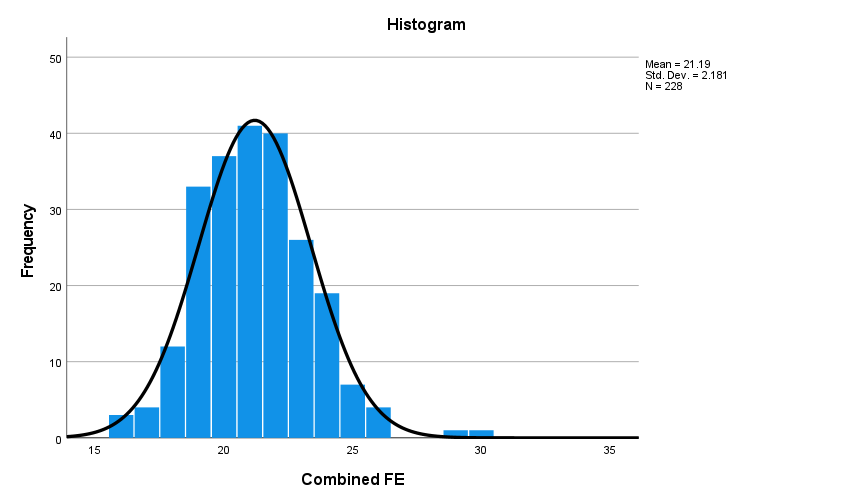
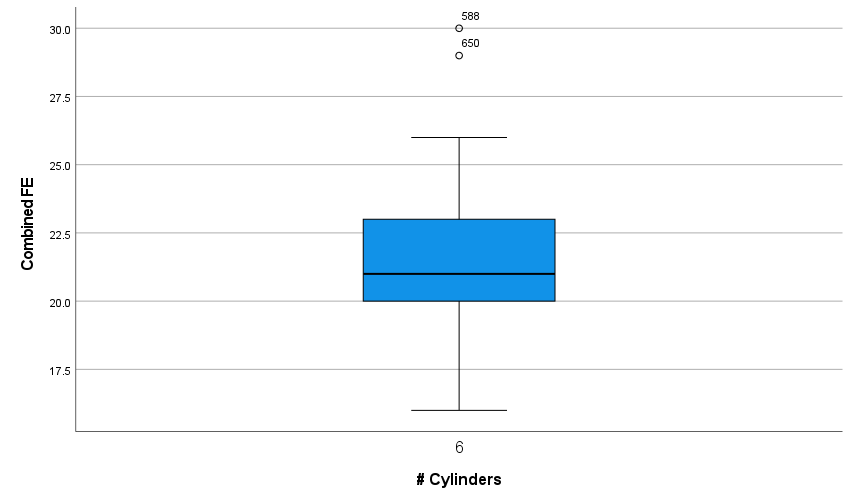
Description automatically generatedA screenshot of a computer

Description automatically generated

Means of combined FE cylinder = 8and not volkswagen



showing mean of combined FE of cylinder =6 and not Volkswagen

Combine FE for vehicles with Cylinder = 6 and not volkswagen

A screenshot of a computer

Description automatically generated

Supporting figure showing Cylinder = 6 and City Co2 levels >= 800 shows mostly GMC cars

A screenshot of a computer

Description automatically generated

Supporting slides showing aspirator = turbocharged, transmission= auto and City CO2 levels>= 700 mostly Volkswagens.

A screenshot of a computer

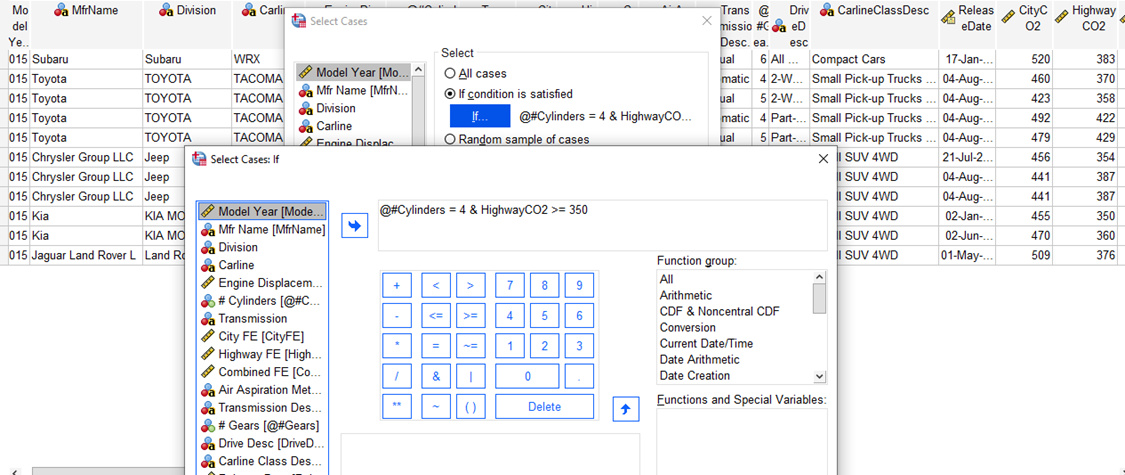
Description automatically generated

Showing the replaced C02 records divided by 2.

A screenshot of a computer

Description automatically generated

Supporting figure showing Cylinder = 6 and highway Co2 levels >= 420 shows mostly Heavy-duty SUVs and Vans.



Supporting figure showing Cylinder = 4 and highway Co2 levels >= 350 shows mostly Heavy-duty SUVs and Vans.

A screenshot of a computer

Description automatically generated

Supporting figure showing Cylinder = 6 and highway Co2 levels >= 540 shows mostly Heavy-duty SUVs and Vans. Ford and Jaguar not outliers