DEPARTMENT OF STATISTICS, ASUTOSH COLLEGE

Affiliated to

UNIVERSITY OF CALCUTTA

NAME OF PROJECT: ANALYSIS AND PREDICTION OF CO_2 EMISSION FROM MID-SIZED CARS

NAME: TATHAGATA CHATTOPADHYAY

SEMESTER: VI

PAPER: DSE-B2

ROLL NUMBER: 193012-21-0389

REGISTRATION NUMBER: 012-1111-0695-19

Under supervision of

Dr. Dhiman Dutta

CONTENTS

ABSTRACT

Environmental pollution is not a new phenomenon, yet it remains the world's greatest problem facing humanity and the leading environmental cause of morbidity and mortality. Man's activities through urbanization, industrialization, mining, and exploration are at the forefront of global environmental pollution. In this project, we take the initiative to predict the CO_2 emission caused by mid-sized vehicles and also analyze the given data to understand what measures can be taken by both the consumer and the producer to reduce the CO_2 emissions in the environment. The analytical and predictive study has been conducted using the Government of Canada dataset which gives us a comparative view of different brands and vehicle models by their fuel consumption and carbon dioxide emissions. We then use various statistical methods and inferences to understand and visualize the data, so that we can list recommendations that the consumers and producers may follow. We also use multiple linear regression to predict the CO_2 emissions from various car brands.

INTRODUCTION

The transportation sector accounts for a large proportion of global greenhouse gas and toxic pollutant emissions. Even though alternative fuel vehicles such as all-electric vehicles will be the best solution in the future, mitigating emissions by existing gasoline vehicles is an alternative countermeasure in the near term. This project aims to predict the vehicle CO_2 emission per kilometer of mid-sized vehicles and determine an eco-friendly path that results in minimum CO_2 emissions.

Estimating and visualizing fuel consumption and exhaust emissions are critical for reducing the energy cost and air pollution caused by transportation, as well as detailing emission control strategies. As, in the past decade, there has been a pressing concern about climate change, estimation models of CO_2 emissions and fuel consumption from vehicles are of increasing significance. Therefore, this has invoked a global interest in statistical research for sustainability among global researchers and analysts. As an outcome of this project, we shall be able to make recommendations and necessary changes for an efficient and sustainable future of minimal emission vehicles.

METHODOLOGY

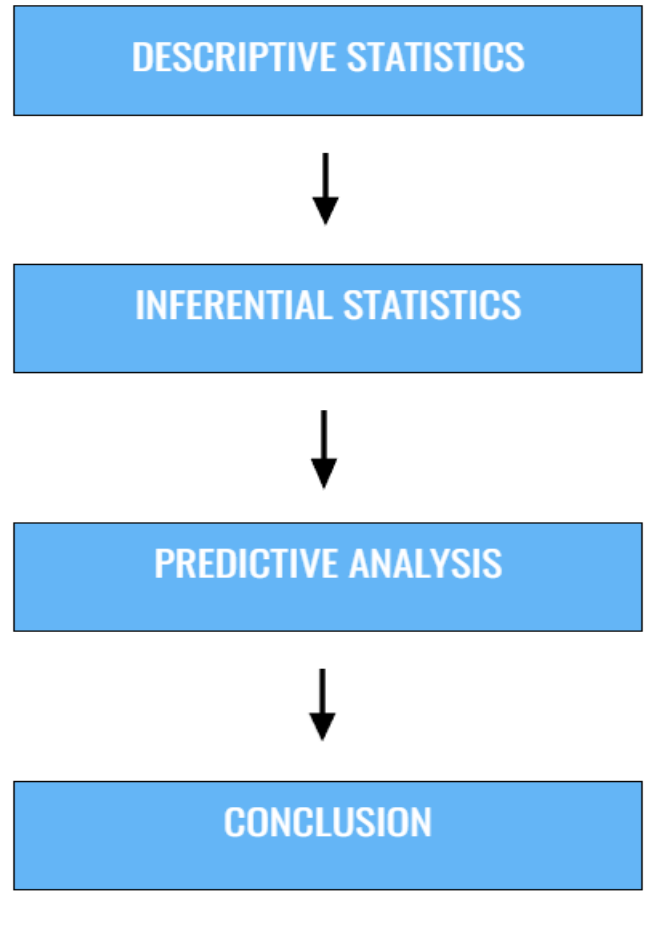
To successfully understand and interpret the dataset, we must follow a step-by-step statistical analysis. We follow these methods one after the other, to analyze the data and give important recommendations and conclusions:

- **Descriptive Statistics**: In this section, we compute various descriptive measures such as central tendencies (mean, median, etc.). In addition to that, we visualize the data for easy comparison among different features and factors. This section is extremely important as it helps us give recommendations on which feature/factor to use to achieve minimum CO_2 emission from the vehicles. We also understand the frequency distribution for various features in this section.

- **Inferential Statistics**: In this section, we run various parametric or non-parametric tests, whichever is suitable, to understand the difference in location, spread, skewness, kurtosis, or any other relevant area of interest between the various features in the dataset. In this project, Wilcoxon rank-sum test has been used to compare the location shift (median) between fuel consumption in City and Highway. Kruskal-Wallis test has also been used to check whether there is a significant difference between the distributions of various fuel types in predicting CO_2 emission. We have also calculated the principal component among all the features. Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. So, if the dataset we are working with needs more columns in the future, PCA may help us to reduce the complexity of the dataset.

- **Predictive Analysis**: In this section, we build our main prediction model, i.e., the multiple linear regression model. We first clean the dataset and remove unnecessary columns from the same. We also try to remove outliers from features of interest. In this project, two regression models

have been constructed and compared to analyze and understand which of them is the better model. In the first model, the dataset has been left uncleaned with 249 observations. The second model has been constructed after data cleaning and has 232 observations. When we look at the results, we can easily conclude that the second model gives us the best fit. We have also evaluated the model using mean squared error, root mean squared error, and adjusted r-squared values. This section mainly aims to provide a suitable regression model which the consumers and producers may use to predict how much CO_2 a vehicle might emit, before manufacturing or purchasing the vehicle.



Software used:

1) R studio
2) Google Collab (Python IDE)
3) Microsoft Excel

DATASET

Before we start our analysis and get deeper into the project, we shall first take a quick look at the dataset, understand the different columns and what they mean and also understand how they are relevant to the project.

	Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
1	ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	7.7	10.0	28	230
2	ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	9.0	11.1	25	255
3	ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	9.5	11.6	24	267
4	AUDI	A6 QUATTRO TDI (modified)	MID-SIZE	3.0	6	AS8	D	9.8	6.2	8.1	35	217
5	AUDI	A7 QUATTRO	MID-SIZE	3.0	6	AS8	Z	13.3	8.5	11.2	25	262
6	AUDI	A8	MID-SIZE	4.0	8	AS8	Z	13.7	8.3	11.3	25	265
7	AUDI	RS 7	MID-SIZE	4.0	8	AS8	Z	15.5	8.8	12.5	23	293
8	AUDI	S6	MID-SIZE	4.0	8	A7	Z	14.2	9.7	12.2	23	281
9	AUDI	S7	MID-SIZE	4.0	8	A7	Z	14.2	9.7	12.2	23	281
10	AUDI	S8	MID-SIZE	4.0	8	AS8	Z	15.8	9.2	12.8	22	300

Dataset showing 10 out of 249 rows

Data columns:

1) Make: This column shows us all the mid-sized vehicle brands we have considered while making the regression model.

2) Model: This column shows us different models under a car brand. For example, Acura has 2 models, namely, RLX and TL AWD.

3) Vehicle class: We are concerned with the CO_2 emission analysis of mid-sized vehicles only. Hence all the values under this column are named 'MID-SIZED'.

4) Engine Size (L): This column shows us various engine sizes of each of the car models we have, in litres, we shall see later in the project that these values are extremely important for our prediction model.

5) Cylinders: This column shows us the total number of cylinders in each car.

6) Transmission: This column shows us various transmission types of each of the car models we have, even these values are extremely important for our prediction model.

7) Fuel Type: This column shows us four different fuel types namely X (Regular gasoline), Z (Premium gasoline), E (Ethanol) and D (Diesel). We shall later analyze

the fuel types and see the average CO_2 emission caused by each fuel.

8) Fuel consumption City (L/100 Km): This column shows city fuel consumption ratings in litres per 100 kilometers.

9) Fuel consumption Hwy (L/100 Km): This column shows highway fuel consumption ratings in litres per 100 kilometers.

10) Fuel consumption Comb (L/100 Km): This column shows the combined fuel consumption rating (55% city, 45% hwy) in litres per 100 kilometers.

10) Fuel consumption Comb (mpg): This column shows the combined fuel consumption rating (55% city, 45% hwy) in miles per imperial gallon (mpg).

11) CO_2 emission (g/Km): This column shows us the CO_2 emission of each vehicle. This is the most important column of the dataset as we are interested in predicting the same. We shall use all the other related columns to help us create a regression model and predict the CO_2 emission caused by a certain car with given specifications. We shall discuss more about the methods as we go deeper into the project.

Type of data:

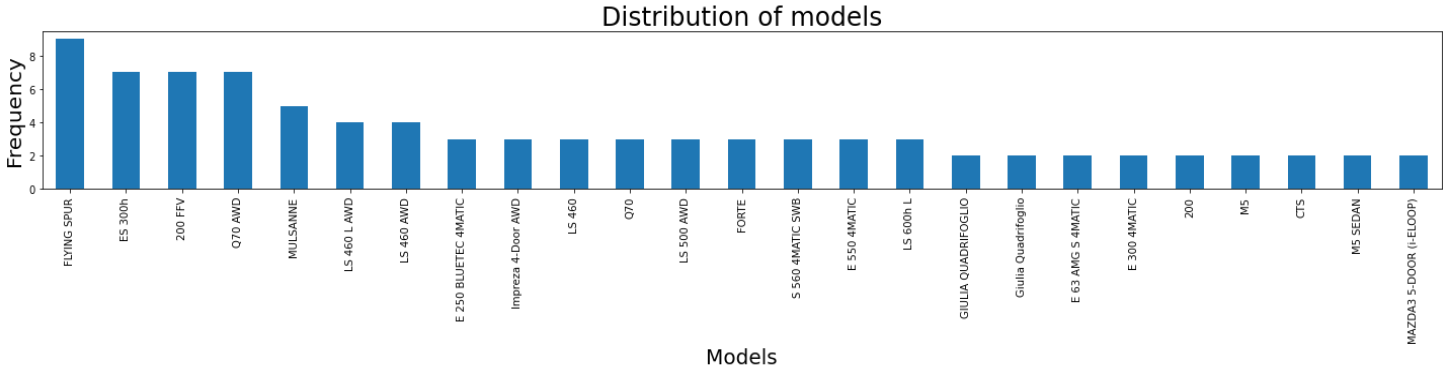
Make	**Character**
Model	Character
Vehicle class	Character
Engine Size (L)	Numeric
Cylinder	Numeric
Transmission	Numeric
Fuel Type	Character
Fuel consumption City (L/100 Km)	Numeric
Fuel consumption Hwy (L/100 Km)	Numeric
Fuel consumption Comb (L/100 Km)	Numeric
Fuel consumption Comb (Mpg)	Numeric

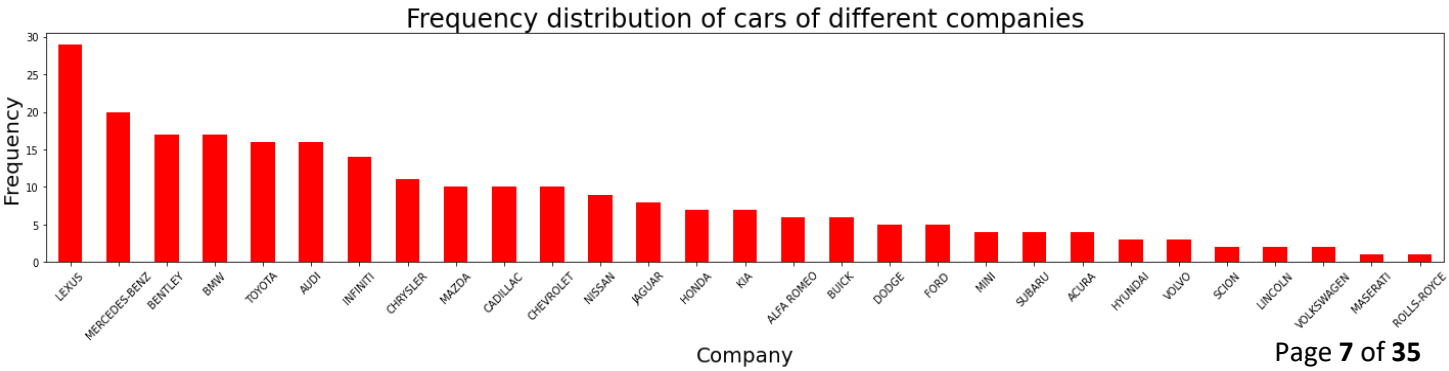
DESCRIPTIVE STATISTICS

The purpose of descriptive statistics is to observe the different columns of our data set like fuel consumption and carbon dioxide emissions from different brands, vehicle models, vehicle class, cylinders, engine size, transmission, etc., and provide a statistical understanding of the dataset quality. Descriptive statistics for all numerical columns in the dataset have been conducted to evaluate the data distribution. The purpose of descriptive statistics is to provide a statistical understanding of the dataset quality.

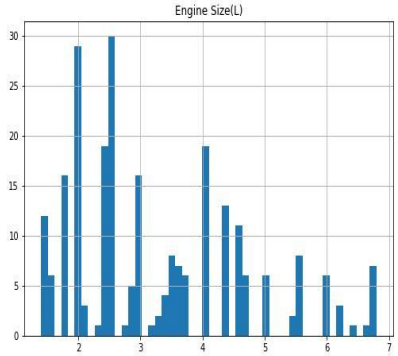
FEATURES	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Engine Size(L)	1.400	2.000	3.000	3.294	4.400	6.800
Cylinders	3.000	4.000	6.000	5.831	8.000	12.000
Fuel cons. City	4.400	9.100	11.900	11.960	14.400	24.500
Fuel cons. Hwy	4.600	6.500	8.100	8.293	9.600	14.900
Fuel cons. Comb.	4.400	8.000	10.200	10.310	12.300	20.000
Co2 Emission	105.000	186.000	230.000	238.800	283.000	465.000

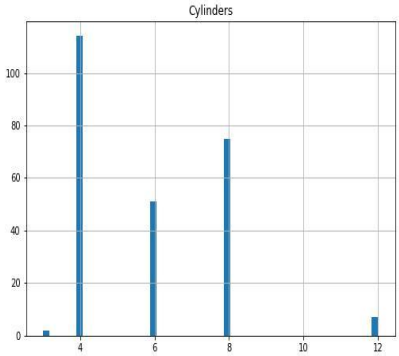
The following charts show us the frequency distribution of all the vehicle brands and also of the various models of these brands.

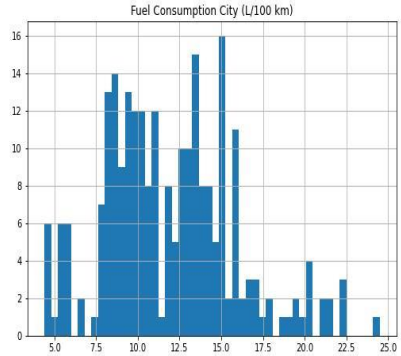


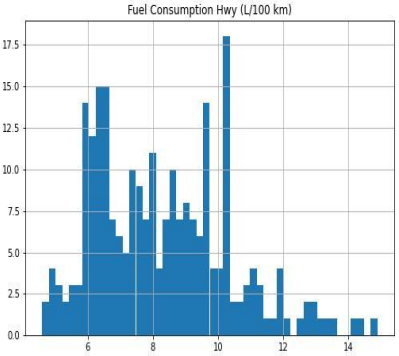


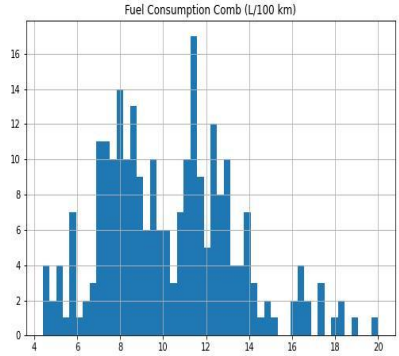
Distribution of numerical features:

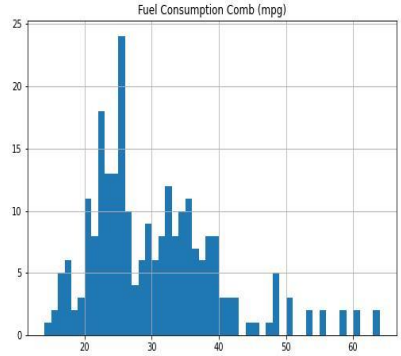


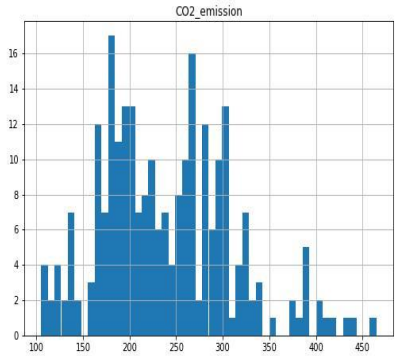






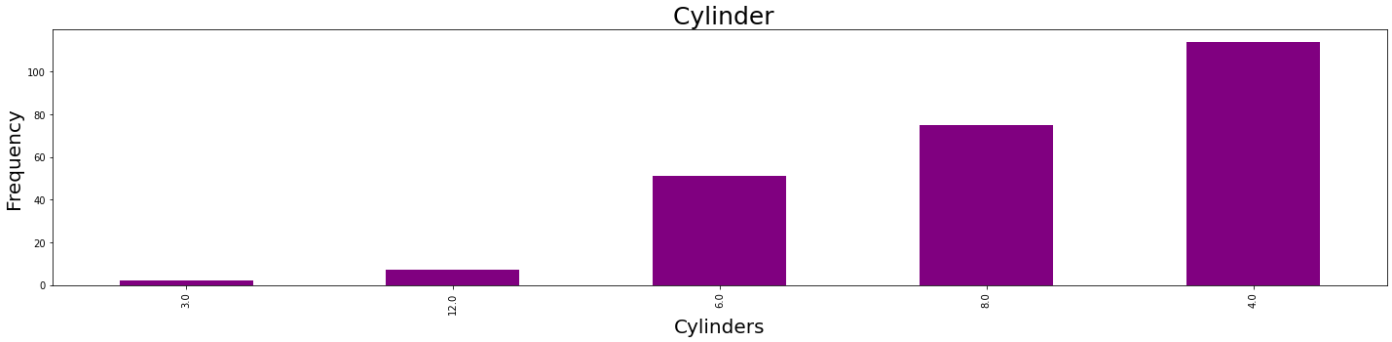


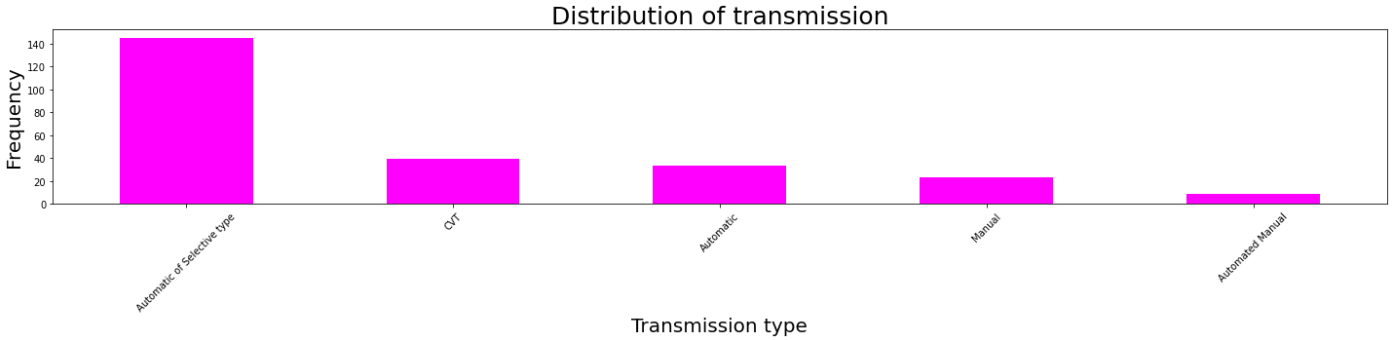




In this figure, we notice that some of the distributions are positively skewed (Fuel consumption, Co2 emission) and some are discrete distributions like Engine size and cylinders.

Frequency distributions of Cylinder, Transmission, and Fuel type:





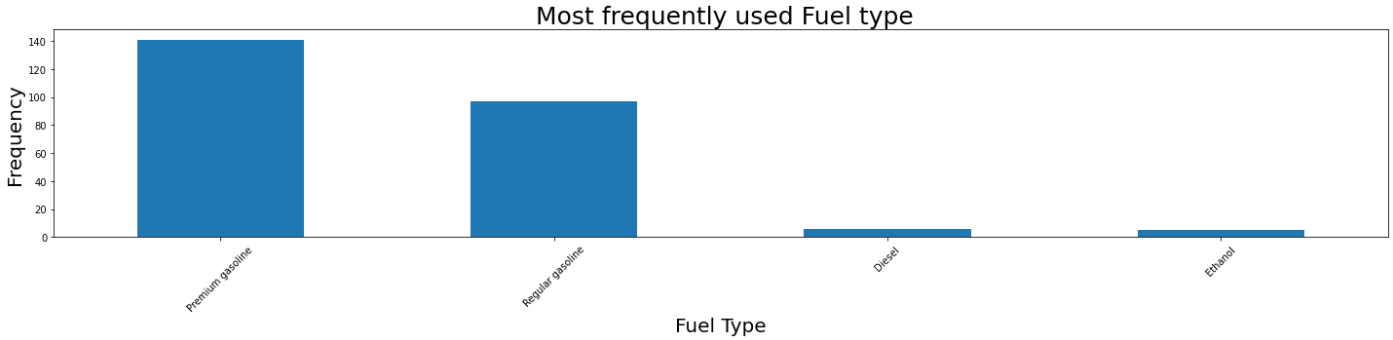
To remove complexity from the project, we have interpreted various transmission types as follows:

A4, A5, A6, A7, A8, A9, A10 – Automatic

AS4, AS5, AS6, AS7, AS8, AS9, AS10 - Automatic Selective type

AV, AV6, AV7, AV8, AV10 – CVT

M5, M6, M7 - Manual



We have done the same with Fuel types:

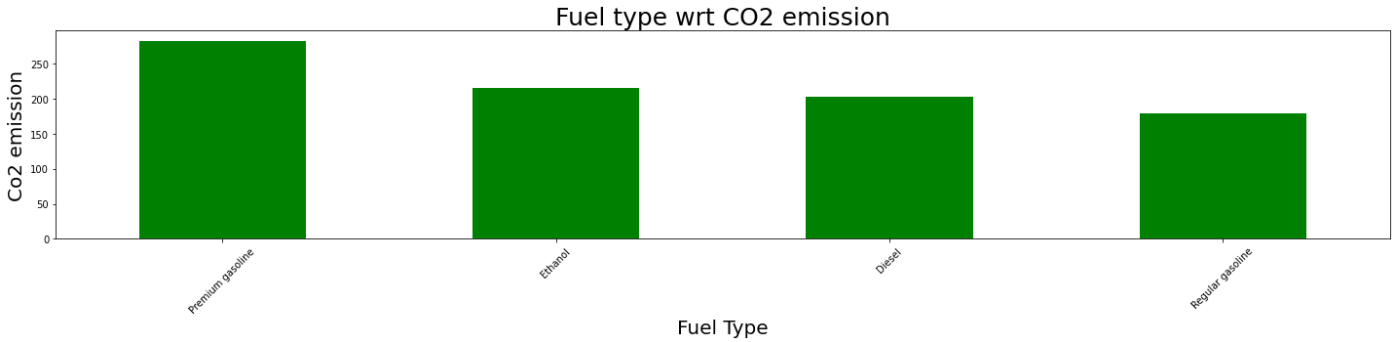
X - Regular gasoline

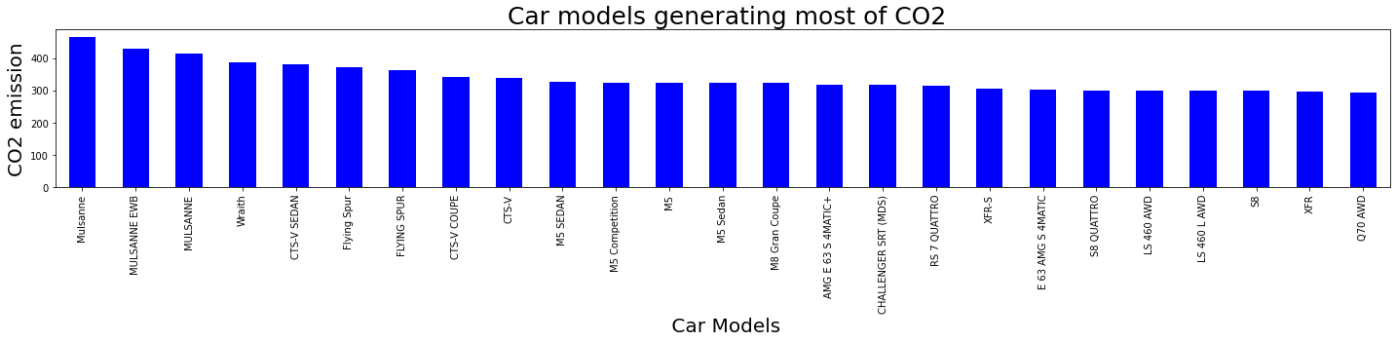
Z - Premium gasoline

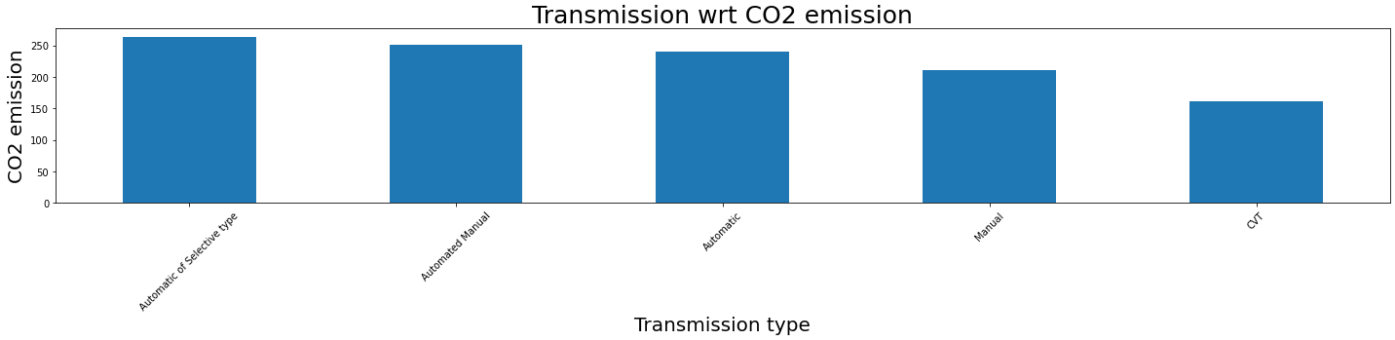
E – Ethanol

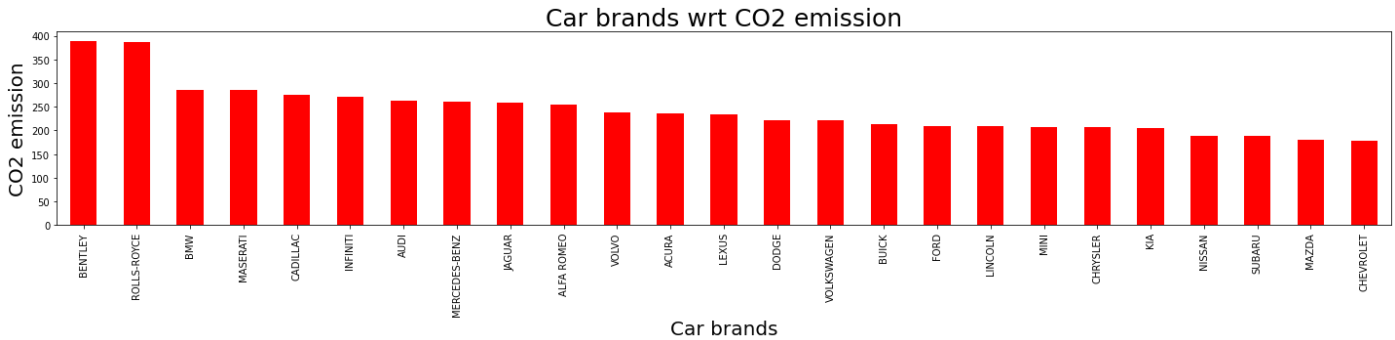
D - Diesel

Feature distributions concerning CO2 emission:

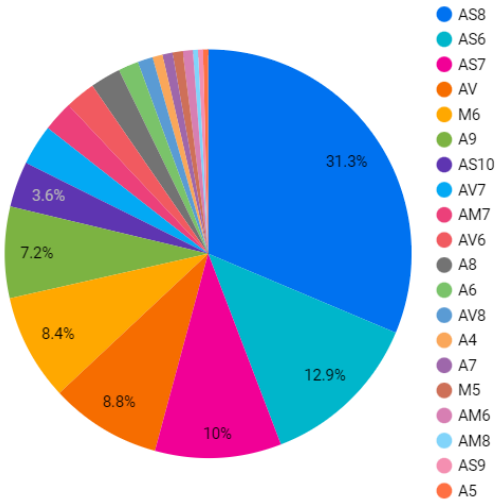






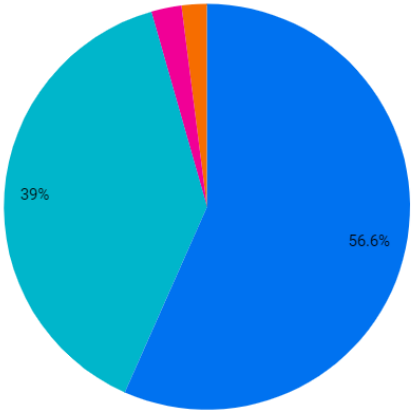


Various **Transmission types** distributed in the dataset:



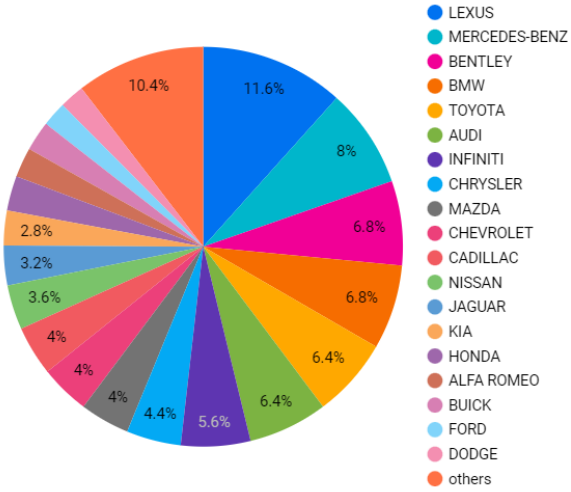
From the pie chart we see that transmission type AS8 is the most commonly used transmission type in case of mid-sized vehicles.

Various **Fuel types** distributed in the dataset:



From the pie chart we see that fuel type Z and X i.e., premium and regular gasoline are the most commonly used fuels in case of mid-sized vehicles.

Various **Car** Brands distributed in the dataset:



This pie chart shows us how many different car brands are present in the dataset. We can see that there are numerous brands with more or less same frequency, which indicates that our analysis will not be biased towards one brand.

INFERENTIAL STATISTICS

With an idea of what the data set looks like, we now may proceed to conduct various analytical tests to infer important conclusions from the data.

Wilcoxon Rank-Sum Test (Mann Whitney U):

In statistics, the T-test is one of the most common tests which is used to determine whether the mean of the two groups is equal to each other. However, the assumption for the test is that both groups are sampled from a normal distribution with equal fluctuation.

Assumptions before performing the Wilcoxon Rank-Sum test: The two samples are mutually independent.

Now here, if we conduct a Shapiro-test of normality we see that,

- **Data: Fuel Consumption City (L/100 km)**

  W = 0.9755, p-value = 0.0002694

- **Data: Fuel Consumption Hwy (L/100 km)**

  W = 0.9632, p-value = 5.237e-06

In both cases, the p-value is significantly smaller than 0.05 which indicates that the groups are **not normally distributed**.

Therefore, we have to use **Wilcoxon rank-sum test (Mann-Whitney U test)** which is a general test to compare two distributions in independent samples. It is a commonly used alternative to the two-sample t-test when the assumptions are not met.

We are interested in testing the following hypothesis,

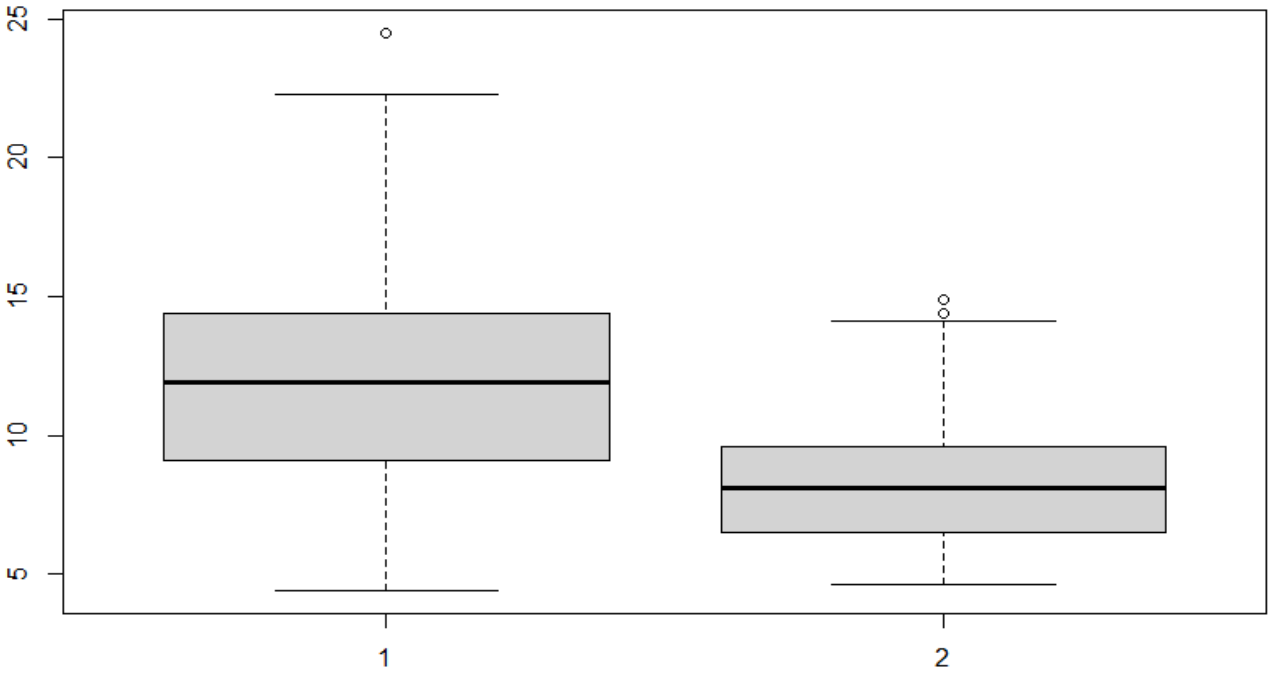
Null Hypothesis (H_0) = Median of fuel consumption in the city = Median of fuel consumption on a highway

Alternative Hypothesis (H_1): Median of fuel consumption in a city ≠ Median of fuel consumption on the highway

On conducting a **Wilcoxon rank-sum test** at the level of significance α = 0.05, we see that,

W = 49061, p-value < 2.2e-16, which concludes that alternative hypothesis is true, i.e., location shift is not equal to 0. Hence the median fuel consumption in a city and on a highway for the same individual has a significant difference.

BOX-PLOT FOR INTUITIVE INTERPRETATION:



"1" indicates mean fuel consumption in the city (L/100 Km) and "2" indicates mean fuel consumption on the highway (L/100 Km).

Kruskal-Wallis Test:

Kruskal-Wallis test by rank is a **non-parametric alternative** to the one-way **ANOVA test**, which extends the two-samples Wilcoxon test in the situation where there are more than two groups. It's recommended when the assumptions of the one-way ANOVA test are not met.

Assumptions before performing the Kruskal Wallis test: The two samples are mutually independent.

If we conduct the Shapiro test for normality, we see that,

- **Data: Diesel**

W = 0.95557, p-value = 0.785

- **Data: Ethanol**

W = 0.82408, p-value = 0.1255

- **Data: Regular gasoline**

W = 0.96926, p-value = 0.02239

- **Data: Premium gasoline**

W = 0.94802, p-value = 3.949e-05

We see that the Premium gasoline or 'Z' does not follow the normality assumption required for ANOVA, hence we have to conduct a Kruskal Wallis test to check the following hypothesis,

Null Hypothesis (H_0) = Mean ranks of the fuel groups are the same

Alternative Hypothesis (H_1): Mean ranks of the fuel groups are not the same

We take the level of significance at $\alpha = 0.05$.

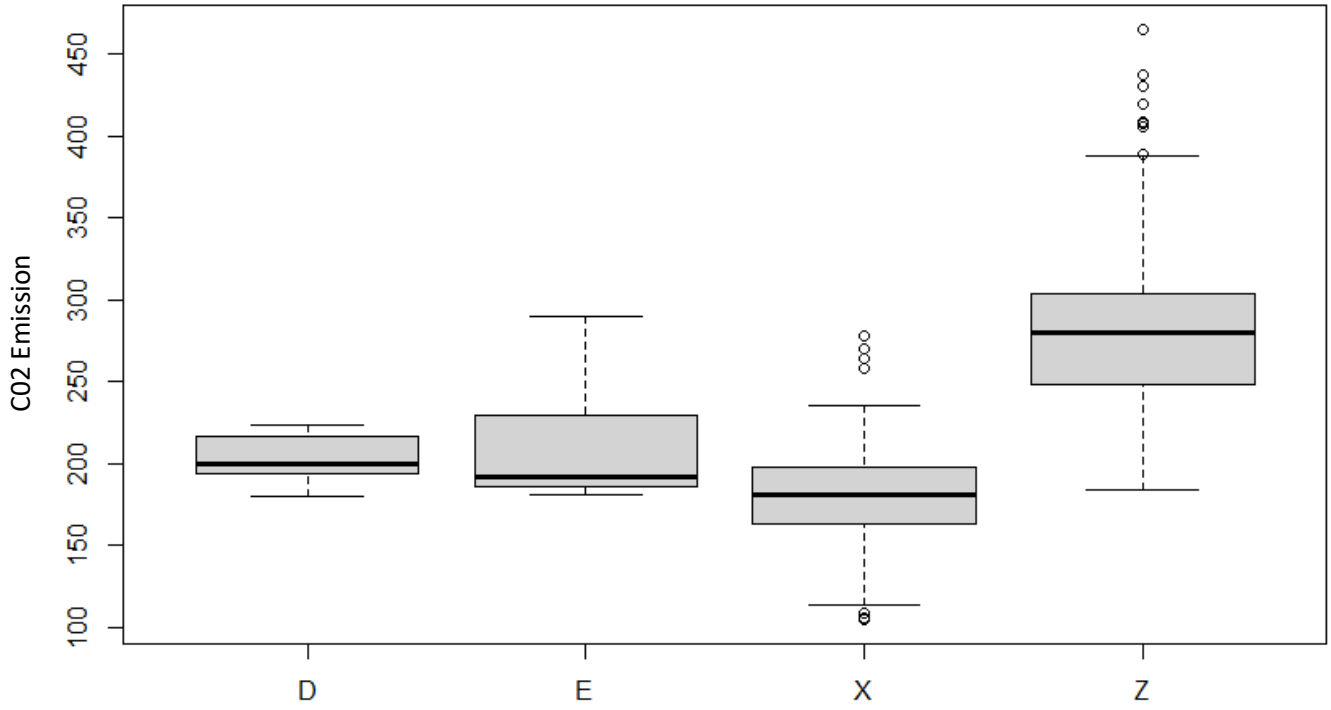
On conducting the test, we find that,

Kruskal-Wallis rank-sum test

Data: Fuel groups (X, Z, D, and E)

Kruskal-Wallis chi-squared = 152.91, df = 3, p-value < 2.2e-16. We can see that the p-value is much less than the level of significance, hence we reject the null hypothesis and accept the alternate hypothesis that the **mean ranks of the fuel groups are not the same.**

<u>BOX-PLOT FOR INTUITIVE INTERPRETATION:</u>



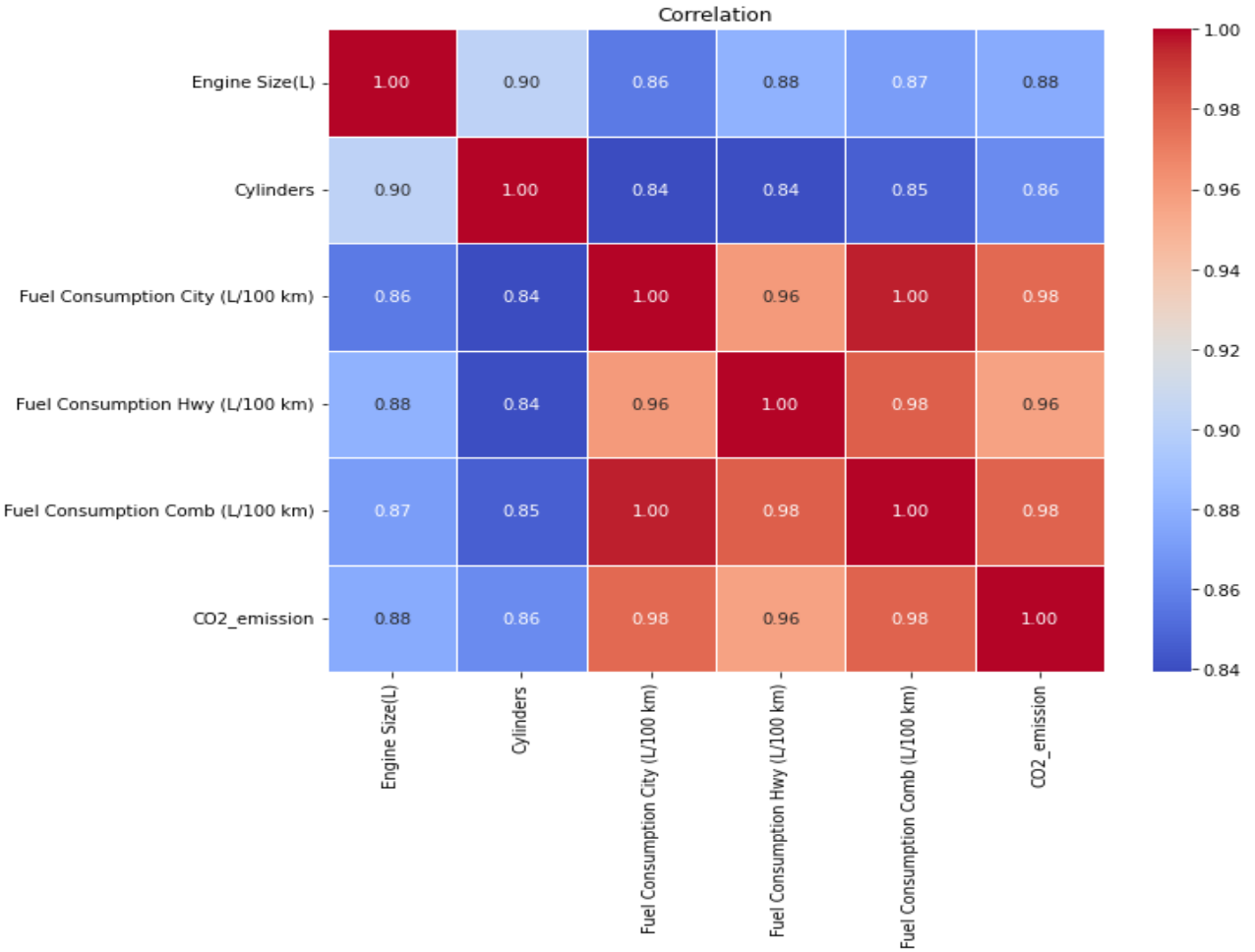
Even visually we can notice the difference in the mean for the various fuel types, our interpretation is further solidified by the Kruskal-Wallis test done above.

[N.B: Non-Parametric Tests such as Kruskal-Wallis and Wilcoxon Rank Sum test are insensitive to outliers, given that the outliers are not extremely large as compared to sample values. In this case, we see that the outliers are not huge and hence they may be ignored]

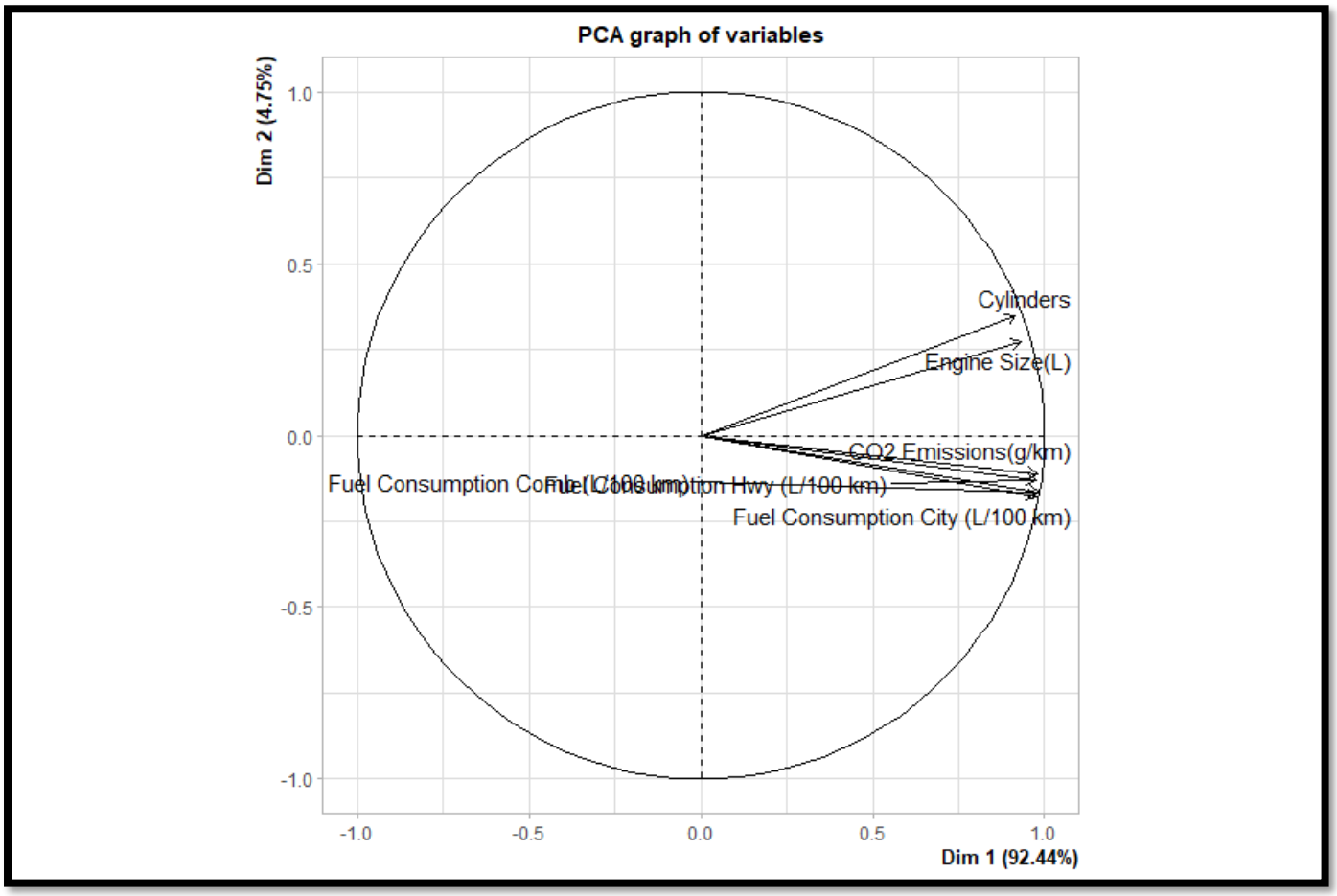
Correlation:

Correlation is a statistical measure that expresses the strength of the relationship between two variables. The two main types of correlation are positive and negative. Positive correlation occurs when two variables move in the same direction; as one increases, so does the other.

To define how the brand, model, vehicle class, cylinder, engine size, transmission type, and fuel type correlate with CO2 emissions, a correlation algorithm has been introduced to generate correlation coefficients. The most commonly used algorithm of this type in statistics is Pearson correlation, which estimates the direction and strength of a linear relationship among two variables. In this study, the objective of this statistic is to define which parameter has the strongest correlation with CO2 emission. To achieve this, Pearson's correlation coefficients have been applied and computed between all features through all vehicles and presented in a correlation heat map. From the heat map, all the correlation coefficients have been calculated, showing the correlation between corresponding parameters on the left and the corresponding parameters at the bottom. The higher the correlation coefficient, the warmer color was presented.

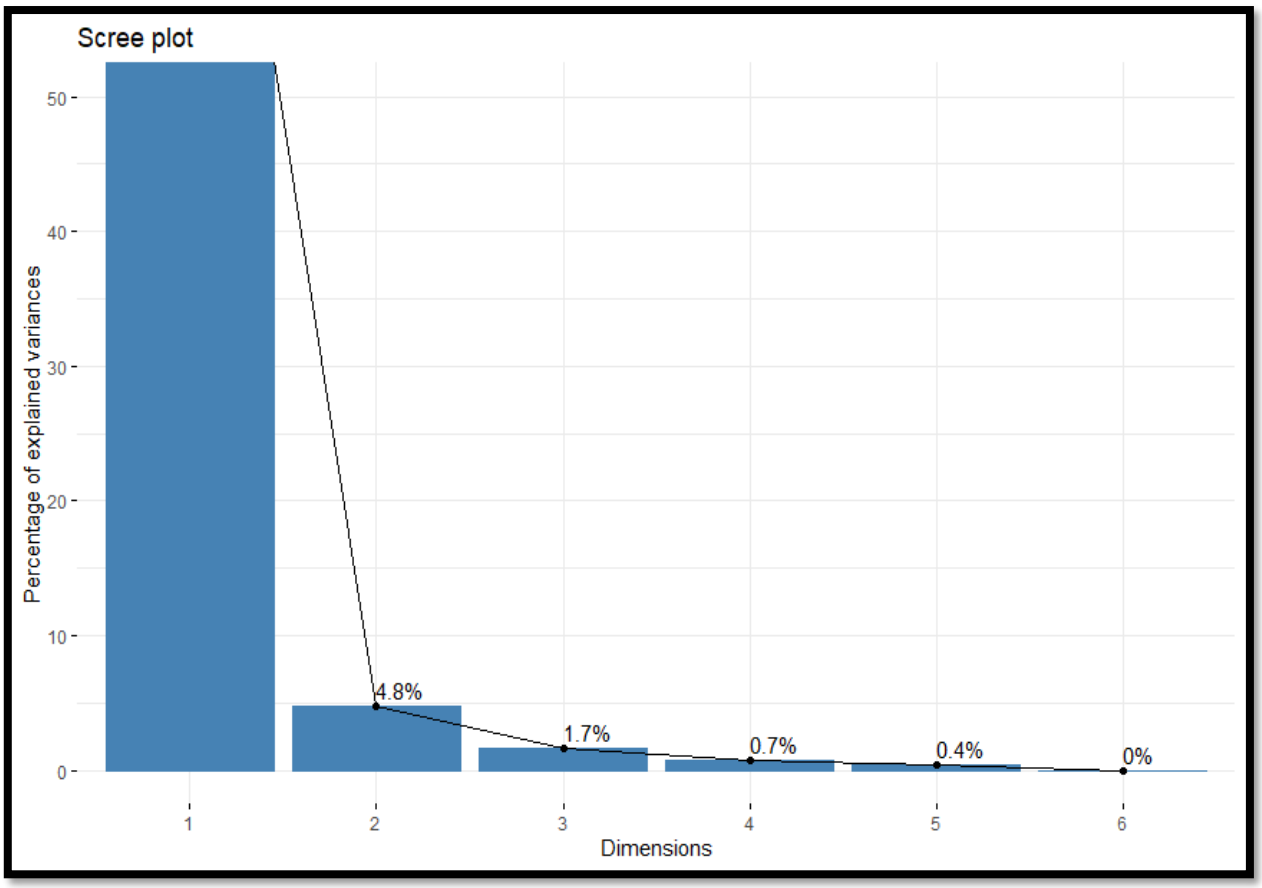


We see that some of the predictor variables are highly correlated with each other. We may use PCA (principal component analysis) to determine which component may reduce the dimension of the data without loss of information. A few graphs with their interpretation have been given as follows:



The plot above is also known as a variable correlation plot. It shows the relationships between all variables. It can be interpreted as follow:

- Positively correlated variables are grouped together.
- Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants).
- The distance between variables and the origin measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map.



An alternative method to determine the number of principal components is to look at a **Scree Plot**, which is the plot of eigenvalues ordered from largest to the smallest. The number of components is determined at the point, beyond which the remaining eigenvalues are all relatively small and of comparable size. Here we see on analysis that the first component (engine size) can explain 92.436% of the variation and the second component (cylinder) can explain 4.754% of the variation in the model. From the plot above we may stop at the second observation as 98% of the information contained in the data is retained by the first two principal components. Therefore, if we were to reduce the dimension, we could easily choose Engine size and Cylinder as the principal components for predicting CO2 emission.

PREDICTIVE ANALYSIS

(MULTIPLE LINEAR REGRESSION)

Regression analysis is a very widely used statistical tool to establish a relationship model between dependent and independent variables. One of these variables is called the response (dependent) variable, whose value is derived from the predictor variables. The other variables are called the predictor variables whose values are gathered through experiments. In Multiple linear Regression, the dependent and independent variables are related through an equation, where the exponent (power) of all the variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph. The general mathematical equation for multiple linear regression is,

Actual data = Fitted model+ Error, where the fitted model is represented as:

$Y= c+\beta 1x1+ \beta 2x2+ \beta 3x3+ \beta 4x4+\ldots\ldots\ldots\ldots\ldots..+ \beta pxp$

Here Y is the response variable, X_i's are the predictor variables. B_i's are constants which are called the coefficients. The error term is also known as residual. The "Residual" term represents the deviations of the observed values y from their means, which are normally distributed.

Assumptions of Multiple Linear Regression:

- There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.
- Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
- No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
- Homoscedasticity–This assumption states that the variance of error terms is similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

CLEANING THE DATASET:

Before running the multiple linear regression model, it is very essential to clean the data set. By cleaning we mean,

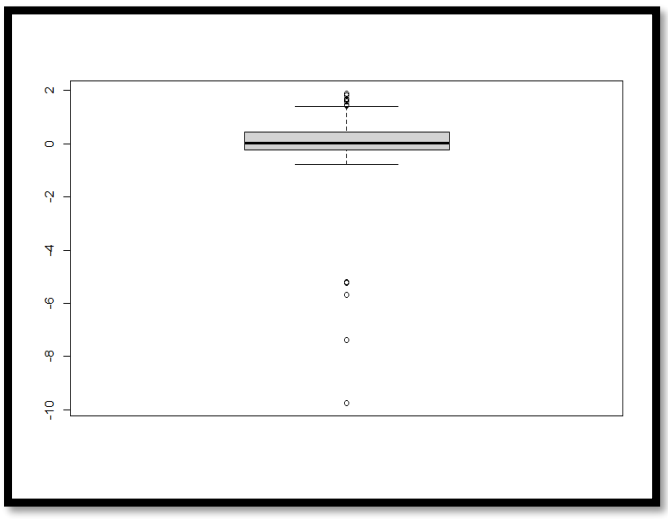
- Analysing and removing unnecessary columns from the data
- Removing null values if any
- Removing outliers from relevant sources

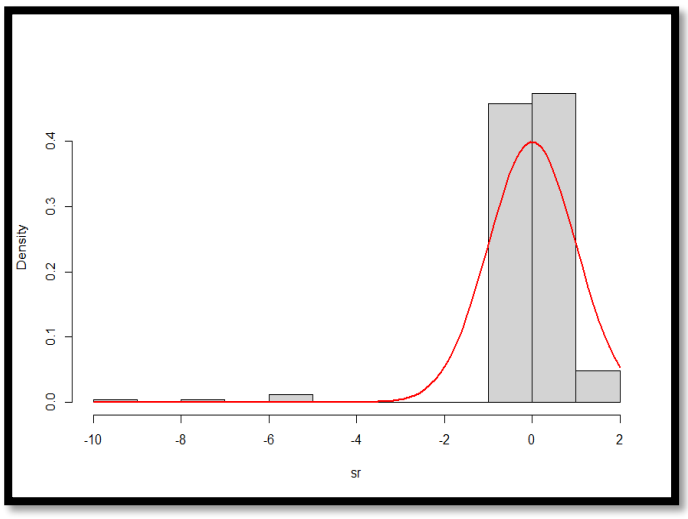
Firstly, we see that there are a few columns containing categorical variables. We shall not use them in the model. Also, we have analyzed their mean differences previously so there is no need to include them here.

Second, we get rid of the combined fuel consumption (mpg) column. This column is redundant as we have already considered the combined fuel consumption in Litres per 100 km and this is just a duplication of this column.

After this, we construct a multiple linear regression model and compare the predicted and actual values to obtain the residual. We then obtain studentized residual values. Studentized residual is computed as regression model residual divided by its adjusted standard error.

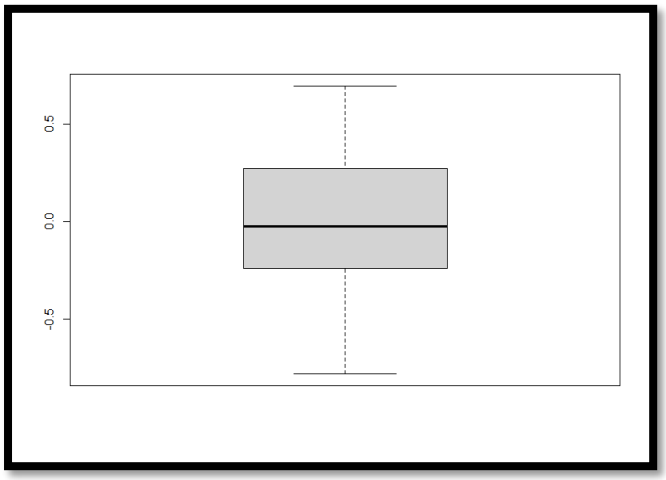
Before removal of outliers, we obtain the following box-plot and histogram for the **studentized residual values**,

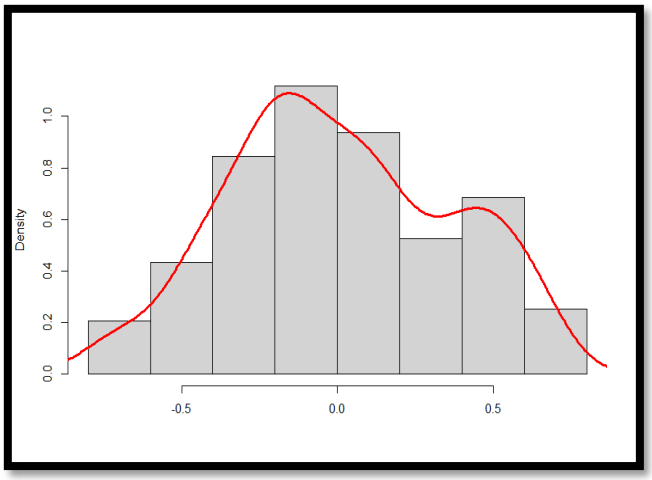




We see that there are a lot of outliers as shown by the boxplot. The histogram also shows that the studentized residuals are not normal as there lie values to the far left of the median.

To take care of this, we conduct an outlier test for studentized residuals using the Inter-quartile method where values greater than 3rd quartile + 1.5*(Inter quartile range) and values below 1st quartile - 1.5*(Inter quartile range) are removed. We get the following outcome after all the cleaning has been done.





As we can see, now, the studentized residuals almost follow a normal-like distribution.

After everything is complete, we conduct a final multiple linear regression model and compare the new model (after cleaning) with the old one (before cleaning).

Aim: To predict CO2 emission with suitable predictor variables.

No. of rows and columns in the data set:

249 rows and 10 columns before cleaning, 232 rows and 9 columns after cleaning.

We shall now compare the two regression models, one before data cleaning, and one after data cleaning.

INTERPRETATION OF REGRESSION MODEL (BEFORE DATA CLEANING)

Residuals:

The residuals are the difference between the actual values and the predicted values. We can generate these same values by taking the actual values of CO2 emission and subtracting them from the predicted values of the model. The following table shows the descriptive statistics for the residual values.

Min	1Q	Median	3Q	Max
-107.538	-3.092	0.411	5.728	25.135

Coefficients table:

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	17.550	4.886	3.592	0.000397
Engine Size(L)	2.522	1.670	1.510	0.132370
Cylinders	3.382	1.030	3.284	0.001174
Fuel Consumption City (L/100 km)	15.867	12.625	1.257	0.210039
Fuel Consumption Hwy (L/100 km)	7.921	10.645	0.744	0.457491
Fuel consumption Comb (L/100 km)	-6.041	22.992	-0.263	0.792979

Based on the above table, our linear regression model is,

Y = 17.550 + 2.522*X1 + 3.382*X2 + 15.867*X3 + 7.921*X4 + (-6.041)*X5, where Y is the response variable and X1, X2, X3, X4 and X5 are the predictor variables.

The p-value is calculated using the t-statistic from the T distribution. The p-value, in association with the t-statistic, helps us to understand how *significant* our coefficient is to the model. In practice, any p-value below 0.05 is usually deemed as *significant*.

X1 = Engine Size (**statistically insignificant** in predicting the CO2 emission)

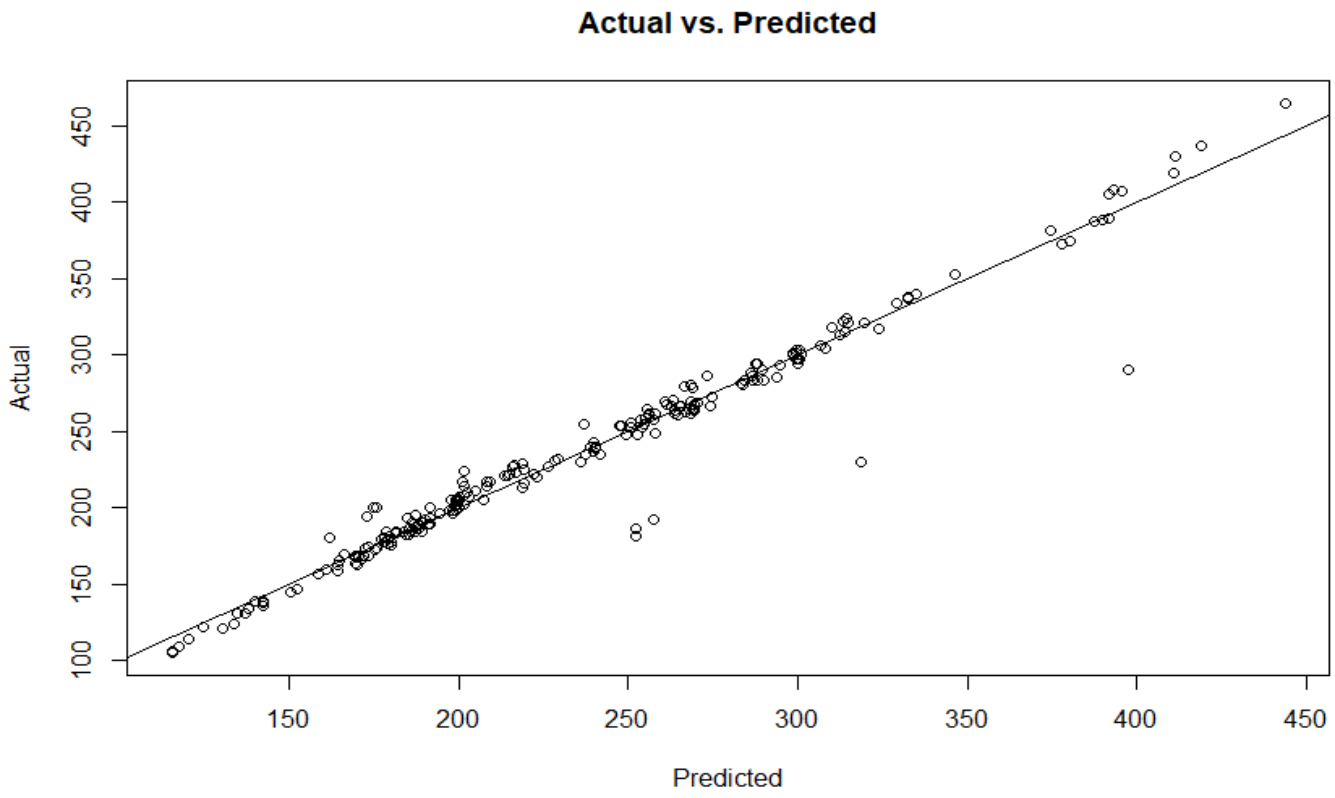
X2 = Cylinders (**statistically significant** in predicting the CO2 emission)

X3 = Fuel consumption in the city (**statistically insignificant** in predicting the CO2 emission)

X4 = Fuel consumption on highway (**statistically insignificant** in predicting the CO2 emission)

X5 = Fuel consumption combined (**statistically insignificant** in predicting the CO2 emission)

We see that in this model none of the predictor variables other than 'cylinder' are significant in predicting CO2 emission.



Residual standard error:

The residual standard error is a measure of how well the model fits the data.

Here we have Residual standard error **= 13.48 on 243 degrees of freedom**. If we look at the least-squares regression line, we notice that the line fits through some of the points and that there is a "residual" between the point and the line. The residual standard error tells us the **average** amount that the actual values of Y (the dots) differ from the predictions (the line) in units of Y. In general, we want the

smallest residual standard error possible, because that means our model's prediction line is very close to the actual values, on average.

Multiple R-squared: 0.9636, Adjusted R-squared: 0.9628

The Multiple R-squared value is most often used for simple linear regression (one predictor). It tells us what percentage of the variation within our dependent variable the independent variable is explaining.

The **Adjusted R-squared value** can be used for **Multiple linear regression**. It shows what percentage of variation within our dependent variable that all predictors are explaining. Here it **explains 96.28% of the variation**.

F-statistic: 1285 on 5 and 243 DF, p-value: < 2.2e-16

When running a regression model, either simple or multiple, a hypothesis test is being run on the global model. The null hypothesis is that there is no relationship between the dependent variables and the independent variable and the alternative hypothesis is that there is a relationship. The F-statistic and overall p-value help us determine the result of this test.

We can see from our model, that the F-statistic is moderately large and our p-value is so small it is almost zero. **This would lead us to reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between all the predictor variables and the response variable.**

INTERPRETATION OF REGRESSION MODEL (AFTER DATA CLEANING)

Residuals:

The residuals are the difference between the actual values and the predicted values. We can generate these same values by taking the actual values of CO2 emission and subtracting them from the predicted values of the model. The following table shows the descriptive statistics for the residual values.

Min	1Q	Median	3Q	Max
-4.5182	-1.7236	0.2404	1.7103	10.2621

Coefficients table:

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.4444	0.8926	-0.498	0.619027
Engine Size(L)	-1.6096	0.2988	-5.386	1.80e-07
Cylinders	1.0788	0.2033	5.306	2.67e-07
Fuel Consumption City (L/100 km)	8.0286	2.1922	3.662	0.000311
Fuel Consumption Hwy (L/100 km)	6.4676	1.8448	3.506	0.000549
Fuel consumption Comb (L/100 km)	8.7083	3.9954	2.180	0.030322

Based on the above table, our linear regression model is,

Y = (-0.4444) + (-1.6096) *X1 + 1.0788 *X2 + 8.0286*X3 + 6.4676*X4 + 8.7083*X5, where Y is the response variable and X1, X2, X3, X4 and X5 are the predictor variables.

The p-value is calculated using the t-statistic from the T distribution. The p-value, in association with the t-statistic, helps us to understand how *significant* our

coefficient is to the model. In practice, any p-value below 0.05 is usually deemed as *significant*.

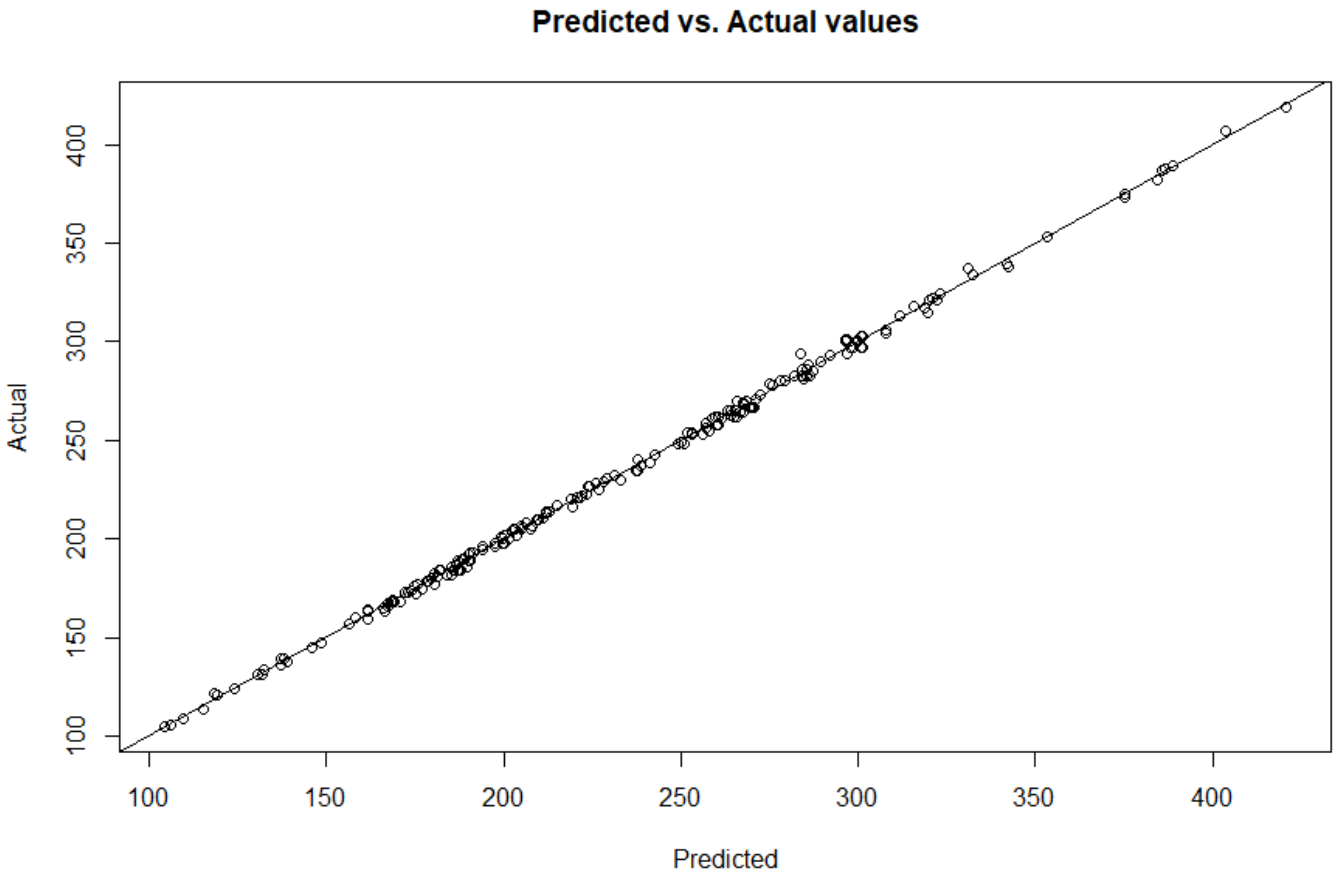
X1 = Engine Size (**statistically significant** in predicting the CO2 emission)

X2 = Cylinders (**statistically significant** in predicting the CO2 emission)

X3 = Fuel consumption in the city (**statistically significant** in predicting the CO2 emission)

X4 = Fuel consumption on highway (**statistically significant** in predicting the CO2 emission)

X5 = Fuel consumption combined (**statistically significant** in predicting the CO2 emission)



Residual standard error:

The residual standard error is a measure of how well the model fits the data.

Here we have Residual standard error **= 2.225 on 226 degrees of freedom**. If we look at the least-squares regression line, we notice that the line fits almost perfectly through each of the points and that there is a very small "residual"

between the point and the line. The residual standard error tells us the **average** amount that the actual values of Y (the dots) differ from the predictions (the line) in units of Y. In general, we want the smallest residual standard error possible, because that means our model's prediction line is very close to the actual values, on average.

Multiple R-squared: 0.9989, Adjusted R-squared: 0.9989

The Multiple R-squared value is most often used for simple linear regression (one predictor). It tells us what percentage of the variation within our dependent variable the independent variable is explaining.

The **Adjusted R-squared value** can be used for **Multiple linear regression**. It shows what percentage of variation within our dependent variable that all predictors are explaining. Here it **explains 99.89% of the variation**.

F-statistic: 4.062e+04 on 5 and 226 DF, p-value: < 2.2e-16

When running a regression model, either simple or multiple, a hypothesis test is being run on the global model. The null hypothesis is that there is no relationship between the dependent variables and the independent variable and the alternative hypothesis is that there is a relationship. The F-statistic and overall p-value help us determine the result of this test.

We can see from our model, that the F-statistic is very large and our p-value is so small it is almost zero. **This would lead us to reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between all the predictor variables and the response variable.**

EVALUATION OF OUR MODELS

	Mean squared error	*Root MSE*	*Adjusted R-squared*
MLR model (without data cleaning)	-8.065071e-17	13.31517	0.9628
MLR model (after data cleaning)	7.411258e-17	2.195705	0.9989

- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squaring the average difference over the data set.

- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.

- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

Now, let us understand the differences among these evaluation metrics,

Mean Squared Error (MSE) and Root Mean Square Error penalize the large prediction errors. However, RMSE is more widely used than MSE to evaluate the performance of the regression model with other random models as it has the same units as the dependent variable (Y-axis).

The lower value of MSE and RMSE implies higher accuracy of a regression model. However, a higher value of R square is considered desirable.

R Squared & Adjusted R Squared are used for explaining how well the independent variables in the linear regression model explains the variability in the dependent variable. R Squared value always increases with the addition of the independent

variables which might lead to the addition of the redundant variables in our model. However, the adjusted R-squared solves this problem.

Adjusted R squared takes into account the number of predictor variables, and it is used to determine the number of independent variables in our model. The value of Adjusted R squared decreases if the increase in the R square by the additional variable isn't significant enough.

For comparing the accuracy among different linear regression models, RMSE is a better choice than R Squared.

Conclusion:

From the above table, we can see that the RMSE of the new model is significantly lesser than the old model. We can also see that the adjusted R-squared for the new model explains 99.89% of the variability whereas the older model explains 96.28%. Hence, we conclude that the new model constructed after data cleaning is better than the older model.

We also conclude that this model is fairly good as we have a **high** Adjusted R-squared score and all of our **predictors are significant** in predicting the response variable, CO2 emission.

RESIDUAL ANALYSIS

An important way of checking whether a regression, simple or multiple, has achieved its goal to explain as much variation as possible in a dependent variable while respecting the underlying assumption, is to check the ***residuals*** of a regression. In other words, having a detailed look at what is left over after explaining the variation in the dependent variable using independent variable(s), i.e., the unexplained variation.

Ideally, all residuals should be small and unstructured; this then would mean that the regression analysis has been successful in explaining the essential part of the variation of the dependent variable. If, however residuals exhibit a structure or present any special aspect that does not seem random, it sheds a "bad light" on the regression. Most problems that were initially overlooked when diagnosing the variables in the model or were impossible to see, will turn up in the residuals, for instance:

- Outliers that have been overlooked, will show up ... as, often, very big residuals.
- If the relationship is not linear, some structure will appear in the residuals
- Non-constant variation of the residuals (heteroscedasticity)
- If groups of observations were overlooked, they'll show up in the residuals
- etc.

In one word, the analysis of residuals is a powerful diagnostic tool, as it will help us to assess, whether some of the underlying assumptions of regression have been violated.

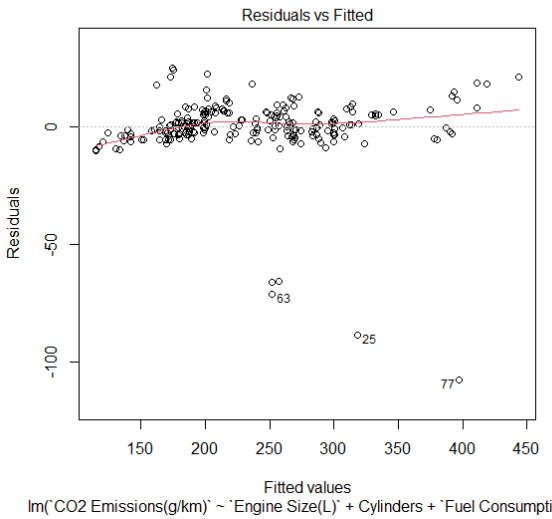
Methods for analyzing residuals:

Since we are using R, four standard plots can be accessed using the **plot()** function with the fit variable once the model is generated. In this project, we discuss three of these plots. These plots can be used to show if there are problems with the dataset and the model produced that need to be considered in looking at the validity of the model. These are:

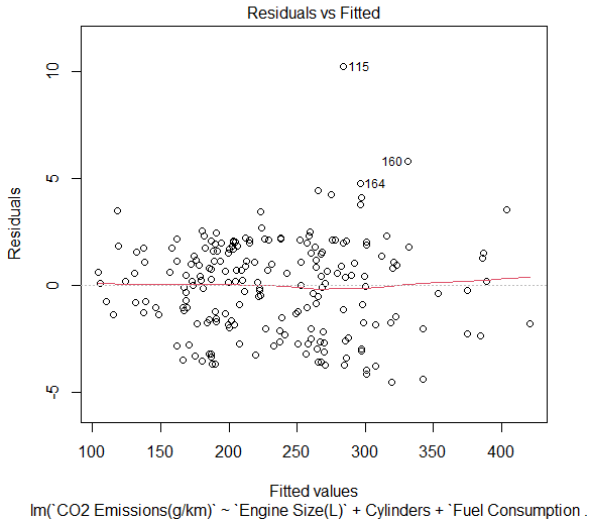
- Residuals vs Fitted Plot
- Normal Q–Q (quantile-quantile) Plot
- Scale-Location

Residuals vs Fitted Plot (Before and after cleaning)

Before



After



When conducting a residual analysis, a **residuals versus fits plot** is the most frequently created plot. It is a scatter plot of residuals on the y-axis and fitted values (estimated responses) on the x-axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

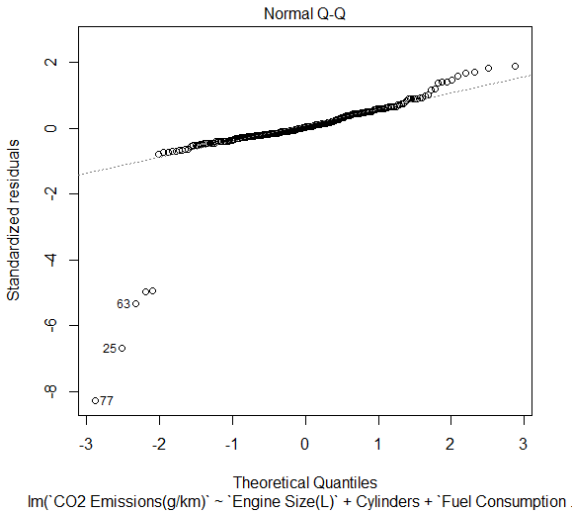
A well-behaved residual plot can be identified by the following characteristics:

- The residuals **bounce randomly** around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
- The residuals roughly form a **horizontal band** around the 0 line. This suggests that the variances of the error terms are equal.
- No one residual **stands out** from the basic random pattern of residuals. This suggests that there are no outliers.

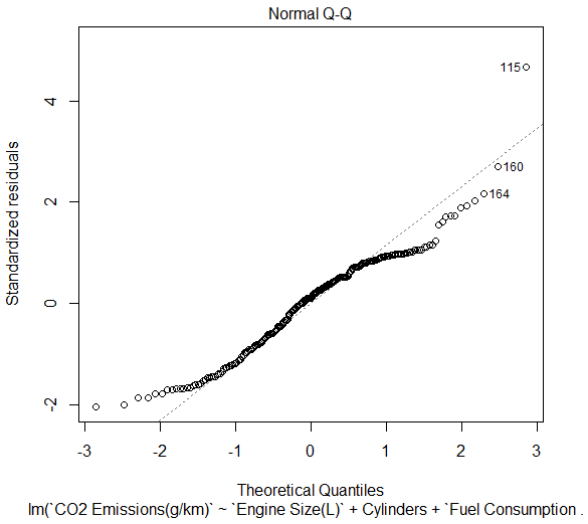
Comparing the two plots we can clearly see that the multiple linear regression model after data cleaning shows a better and well-behaved residual plot.

Quantile – Quantile plot (Before and after cleaning)

Before



After

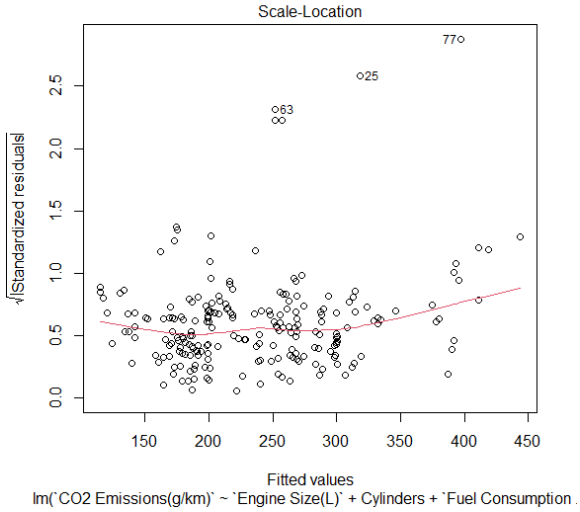


In Statistics, Q-Q(quantile-quantile) plots play a very vital role in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x. We can also interpret the skewness and kurtosis of the data using Q-Q plots.

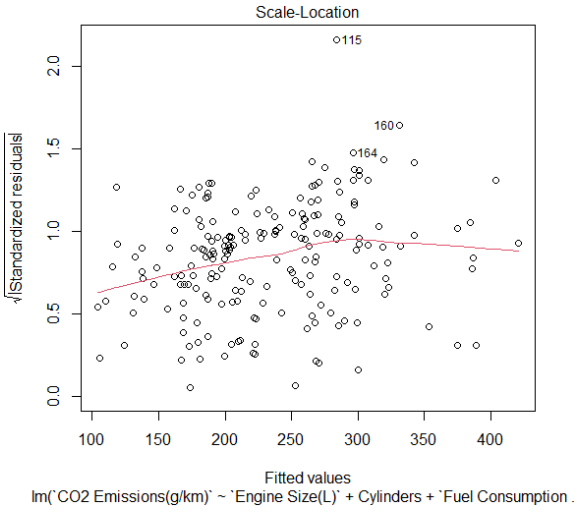
On careful analysis, we see that before data cleaning the standardized residuals were negatively skewed as shown by the first Q-Q plot. But we see that after data cleaning, we almost obtain a normal distribution for the standardized residuals, as on a Q-Q plot normally distributed data appears as roughly a straight line (although the ends of the Q-Q plot often start to deviate from the straight line).

Scale - Location plot (Before and after cleaning)

Before



After



The scale-location plot is very similar to residuals vs fitted but simplifies the analysis of the homoscedasticity assumption where homoscedasticity means constant variance in linear regression. It takes the square root of the absolute value of standardized residuals instead of plotting the residuals themselves.

We want to check two things:

1. That the **red line is approximately horizontal**. Then the average magnitude of the standardized residuals isn't changing much as a function of the fitted values.
2. That the **spread around the red line** doesn't vary with the fitted values. Then the variability of magnitudes doesn't vary much as a function of the fitted values.

Analyzing the plots, we see that the first plot shows clear signs of heteroscedasticity. We may also conduct a **Breusch-Pagan Test** to prove the same. In the case of the second plot, i.e., of the model after data cleaning, we see that it satisfies both the above-mentioned conditions and it has a p-value of 0.0569 in the Breusch-Pagan test, which shows that we may not reject the null hypothesis which says that the data is homoscedastic (with 95% confidence).

CONCLUSION AND RECOMMENDATIONS

In this project, descriptive, inferential, and predictive analysis has been performed using data from the Government of Canada, which includes 249 mid-sized vehicles, to provide a comparative view of various brands in terms of their CO2 emissions. This research analyses different vehicle brands using vehicle measurements, to predict CO2 emissions from various mid-size vehicle models. By using descriptive and inferential statistics methodologies, we observe various plots and graphs which help us visualize the data. Important information can be obtained by descriptive analysis alone. We also perform inferential analysis to understand the difference between the predictor variables in various features of interest. We finally build a multiple linear regression model to predict the CO2 emission. We also conduct residual analysis to find the efficiency of our model to make it market-ready.

Based on our analysis, we can make the following recommendations, which will reduce the overall CO2 emission for various brands in mid-sized vehicles:

- As far as engine size and cylinders are considered, the higher the engine size and the number of cylinders, the higher the CO2 emission. So, steps must be taken to make efficient engines that may be compact yet provide great performance.
- Premium gasoline (Z) and ethanol (E) are the fuel types that emit maximum CO2. Hence diesel (D) and regular gasoline (X) are the preferred fuel types.
- When we compare transmission concerning CO2 emission, we see that CVT emits the least CO2, whereas Automatic of selective type transmission and Automated manual transmission emits the highest CO2. Hence preferred transmission types are AV, AV6, AV7, AV8, and AV10.
- Fuel consumption in cities and highway can be used to predict CO2 emissions. Car manufacturers may keep this in mind while creating various car features so that the overall fuel consumption is reduced.
- On analysis of the given data, we can also see that mid-size vehicle brands like Bentley and Rolls-Royce emit the most CO2. On the other hand, Nissan, Mazda, Chevrolet, and Subaru emit the least CO2.

A larger dataset with more vehicle features should be studied for building a predictive model in vehicle design. Vehicle consumers and producers can adopt the recommendations from the findings of this study and use the multiple linear regression model to design, as well as implement appropriate action plans for reducing their environmental impacts.

Acknowledgment:

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them. I am highly indebted to Dr. Apurba Roy Vice-Principal, Asutosh College, University of Calcutta. Without his help, I couldn't have been a part of this prestigious college.

I would like to express my special gratitude and thanks towards my parents & my supervisor Dr. Dhiman Dutta (Head of the Department of Statistics) for his necessary guidance of this dissertation with his valuable observations and guidance. I owe a special thanks to other faculty members: Dr.Parthasarathi Bera, Dr. Shirsendu Mukherjee, Dr. Sankha Bhattacharya, and Ms. Oindrila Bose who helped me in building up the foundational knowledge throughout my degree. Finally, my earnest thanks and appreciations also go to my colleagues who were part of the journey in completing the degree. This project is a testimony to all the memories I shared with everyone in my college. The teachings and passion are something I would always carry with me as I move head-on toward my future.

Signature

Data source: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64

References:

1) Stack overflow - https://stackoverflow.com/
2) Geeksforgeeks - https://www.geeksforgeeks.org/
3) Statisticsglobe - https://statisticsglobe.com/
4) Fundamentals of Statistics Vol 1
5) Fundamentals of Statistics Vol 2
6) Outline of Statistics Vol 1
7) Kaggle - https://www.kaggle.com/
8) STHDA - http://www.sthda.com/