**INFO8066: Data Analytics**

**Fuel Consumption Analysis**

**Title: Fuel Consumption Analysis and Predictive Modeling**

**Team Members:**

**Amanpreet Kaur (8940730), Akaur0730****@conestogac.on.ca**

**Nirosha Raghunath (8901892), Nraghunath1892@conestogac.on.ca**

**Mohammed zameel(8900952),**

**Mmadayantevalap0952@conestogac.on.ca,**

**Group Number: 4**

**Section 2  
Course Instructor: Kirti Dhir**

Table of Contents

[Introduction 3](#_Toc174486498)

[Data Preprocessing 6](#_Toc174486499)

[Exploratory Data Analysis 13](#_Toc174486500)

[Data Visualization 26](#_Toc174486501)

[Predictive Model 42](#_Toc174486502)

# Introduction

In this project, we are analyzing a dataset which contain information regarding multifarious vehicles, including their engine size, vehicle class, cylinder, fuel consumption, emissions, and other characteristics. The automotive industry plays a significant role in global emissions, making it crucial to understand the factors that contribute to higher emissions. Moreover, this dataset can be valuable for automotive manufacturers, policymakers, environmental agencies and consumers interested understanding and improving vehicle fuel efficiency and reducing environmental impact.

Our goal is to identify key vehicle characteristics that influence emissions and provide insights to reduce the environmental impact of vehicles.

**TASK 1**

**DATA SELECTION**

DATASET: https://www.kaggle.com/datasets/ahmettyilmazz/fuel-consumption?resource=download

**Source: Kaggle**

**Reasons for choosing dataset**

We are choosing this dataset due to multifarious reasons:

**Personal Interest:** we are really into cars, especially how they perform, how efficient they are, and how to design them to be more eco-friendly.

**Environmental Impact**: This data helps understand if cars meet emissions rules and how to make them better for the environment.

**Comprehensive Analysis:** The dataset is full of useful information that can help car makers, government officials, and people who buy cars.

**Educational Value:** This project is a great chance for us to practice and improve our skills in data analysis, creating visuals, and using machine learning.

**Real-World Implications:** The results of this project could lead to better car designs and help protect the environment.

**TASK 2:**

**BUISNESS PROBLEM STATEMENT**

How can automotive manufacturers optimize vehicle designs to improve fuel efficiency and reduce CO2 emissions?

Specific Questions to Address the Business Problem

1. What are the trends in fuel consumption and emissions across different vehicle classes?

2. How does engine size impact fuel consumption and emissions?

3. What is the effect of transmission type on vehicle fuel efficiency?

4. Which fuel types are associated with lower emissions and better fuel efficiency?

5. What are the characteristics of the most fuel-efficient and least polluting vehicles?

**Description of DATA**

A description of the columns is provided below:

YEAR: Year of the vehicle model from 2000-2022.

MAKE: Manufacturer of the vehicle

VEHICLE CLASS: Category of the vehicle (e.g., Compact, Mid-size, SUV)

ENGINE SIZE: Size of the engine in liters

CYLINDERS: Number of cylinders in the engine

FUEL: Type of fuel used (e.g., X for regular gasoline, Z for premium gasoline)

HWY (L/100 km): Fuel consumption on highways (L/100 km)

COMB (L/100 km): Combined fuel consumption (city and highway) in L/100 km

COMB (mpg): Combined fuel consumption (city and highway) in miles per gallon

EMISSIONS: CO2 emissions in grams per kilometer

MODEL:

• 4WD/4X4: Four-wheel drive

• AWD: All-wheel drive

• CNG: Compressed natural gas

• FFV: Flexible-fuel vehicle

• NGV: Natural gas vehicle

• #: High output engine that provides more power than the standard engine of the same size

TRANSMISSION

• A: Automatic

• AM: Automated manual

• AS: Automatic with select shift

• AV: Continuously variable

• M: Manual

• 3 - 10: Number of gears

Fuel Type

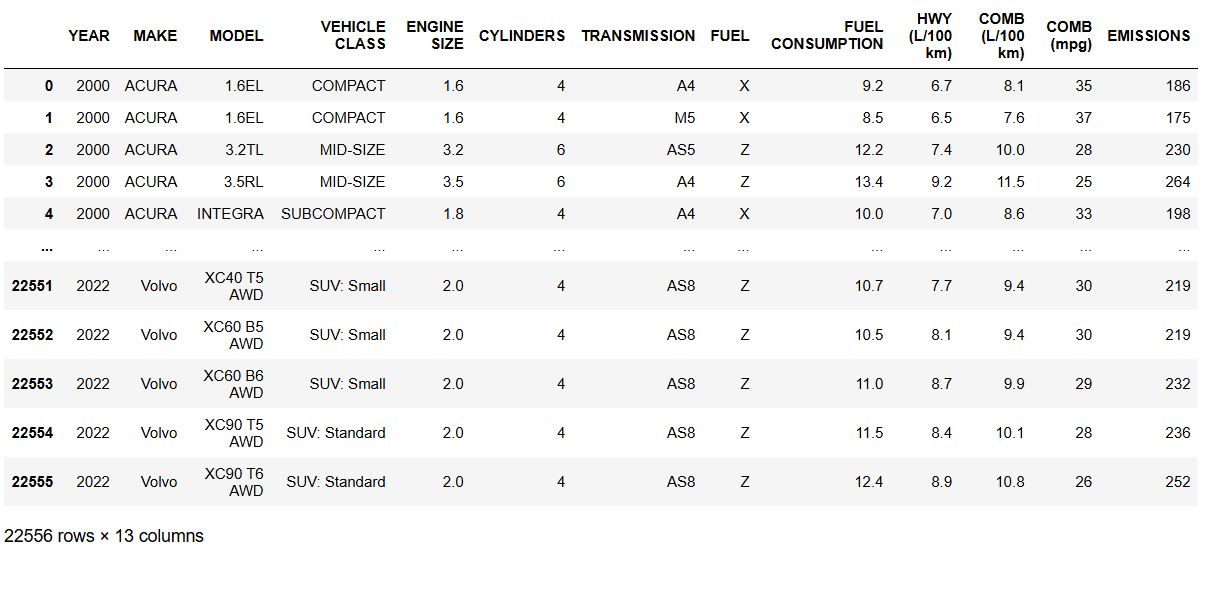
• X: Regular gasoline

• Z: Premium gasoline

• D: Diesel

• E: Ethanol (E85)

• N: Natural Gas



**TASK 3**

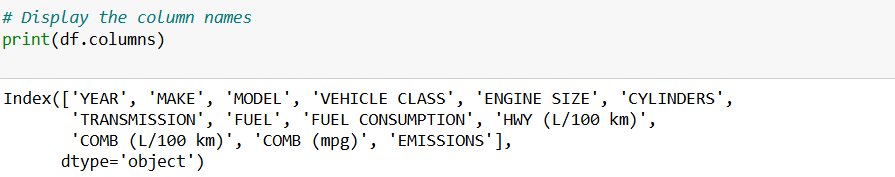
# Data Preprocessing

1. **Checking number of rows and columns**

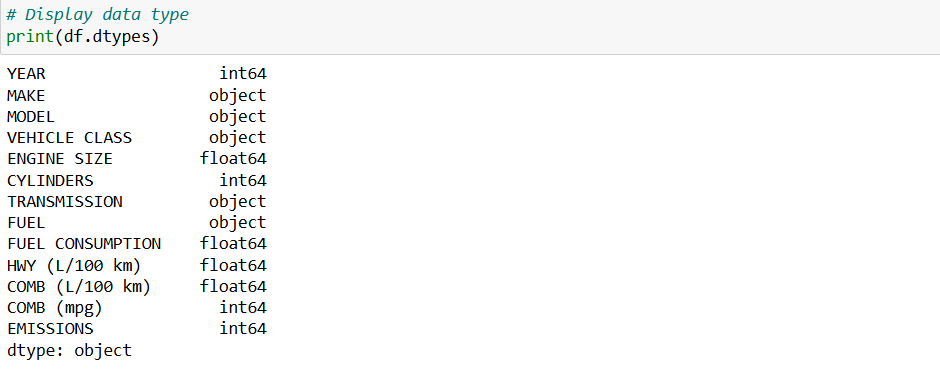


Output: There are 22556 rows and 13 columns.

1. **Display columns names**

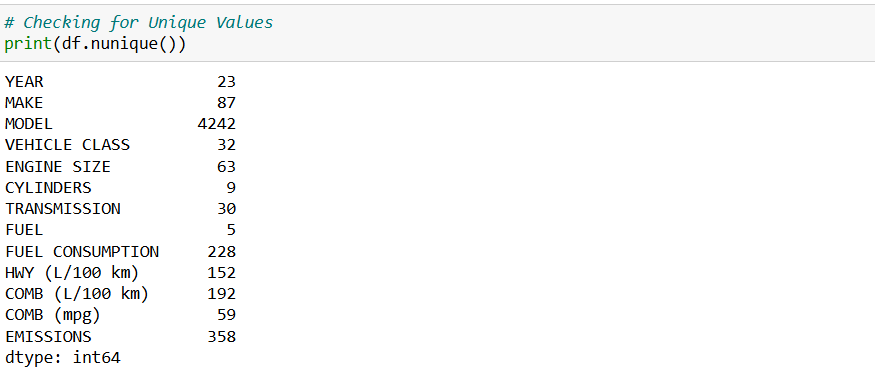


1. **Display data type**



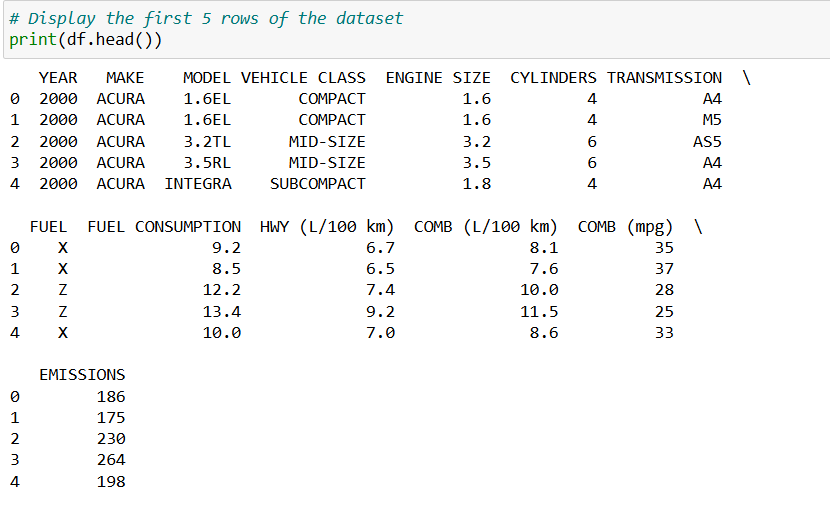
Output: Year is integer only and engine size, fuel consumption (city, highway and combined) are float and others are object.

1. **Calculating unique values**



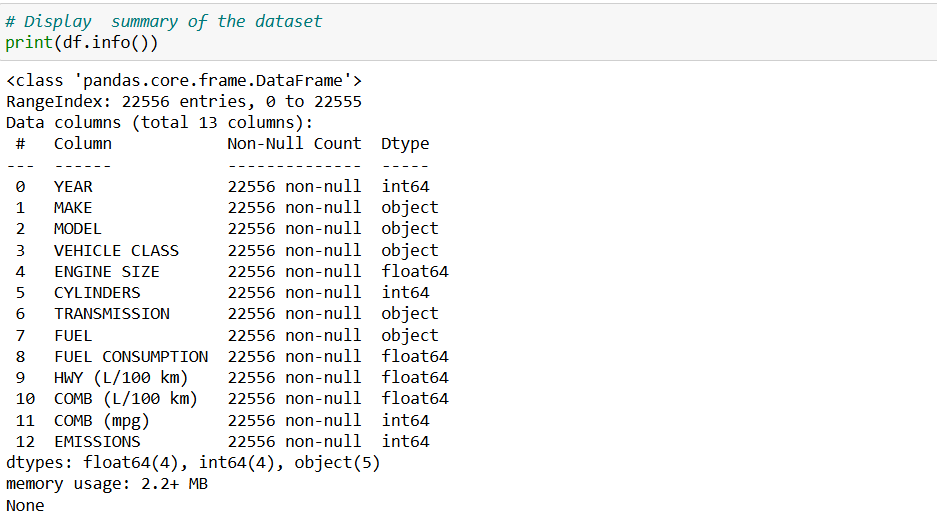
Output: vehicle class has highest number unique values that is 4242 whereas fuel has least count of that compared to other.

1. **Displaying 5 rows of dataset**



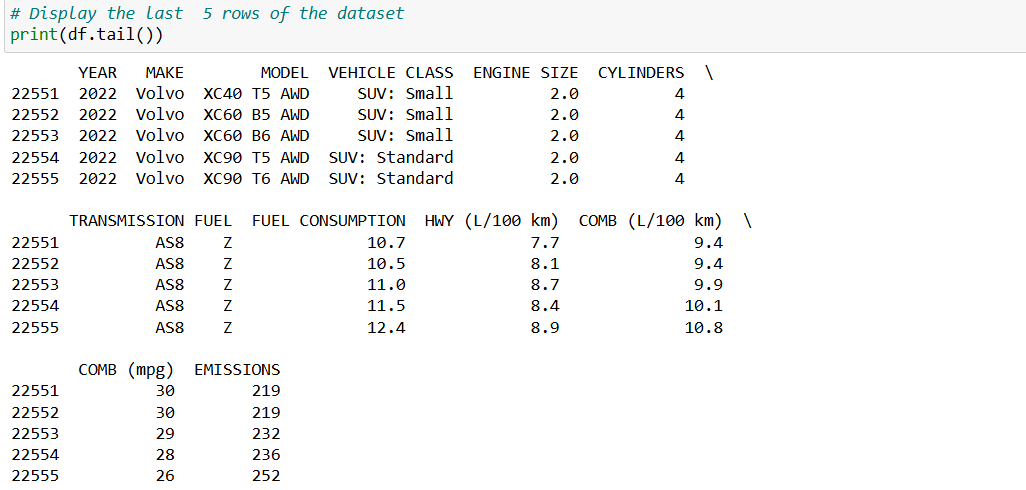
Output: This gives us quick overview of data which includes columns names, incorrect data types.

1. **Displaying summary of dataset**

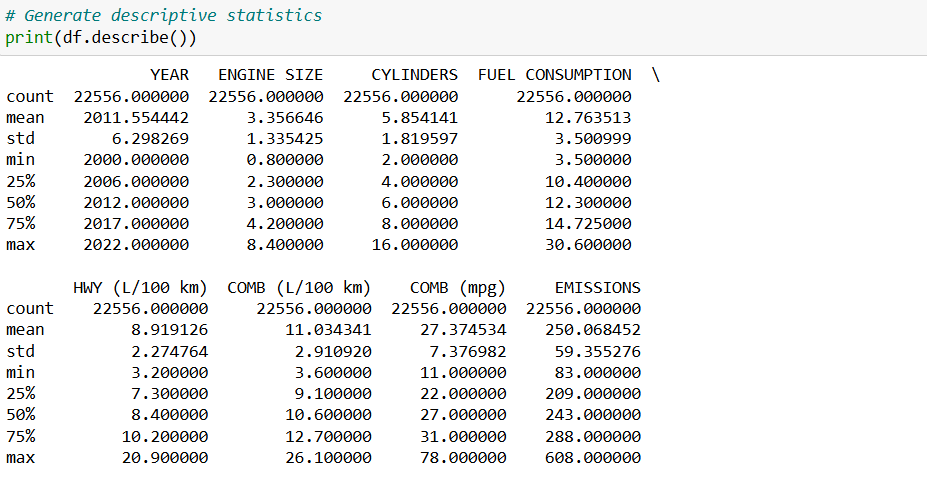


Output: This display that there are 22556 eateries and there are no null values in each column and data type, moreover the memory used by data frame is 2.2 +mb.

1. **Displaying last 5 rows of dataset**.

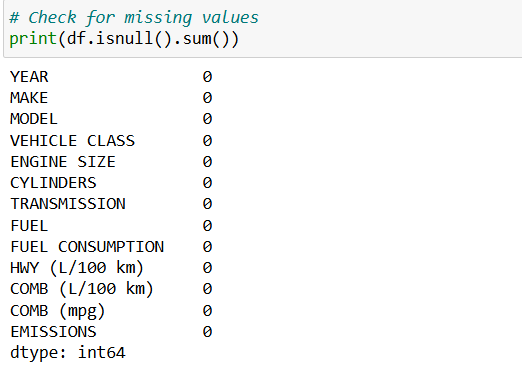


1. **Displaying descriptive statistics.**



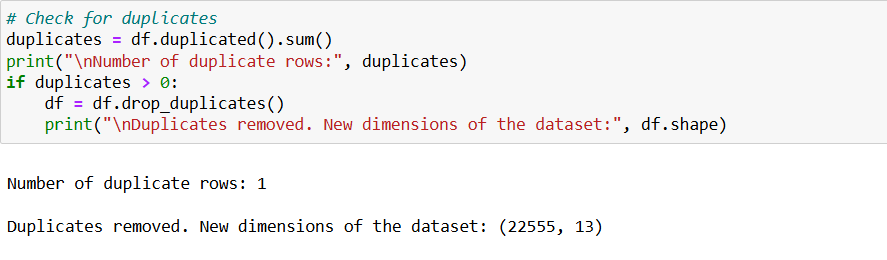
Output: The dataset provides a detailed overview of vehicles from 2000 to 2022. The average engine size is 3.36 liters, with most vehicles having around 6 cylinders. Fuel consumption averages 12.76 liters per 100 km, and highway consumption averages 8.92 liters per 100 km. Combined fuel efficiency averages 27.37 mpg. Emissions data ranges from 83 to 608 grams of CO2 per kilometer, with an average of 250.07 grams. These statistics highlight the dataset's diversity and offer a foundation for analyzing vehicle performance, fuel efficiency, and environmental impact.

1. **Checking missing values**



Output: By analyzing we found that there are no missing values in the dataset.

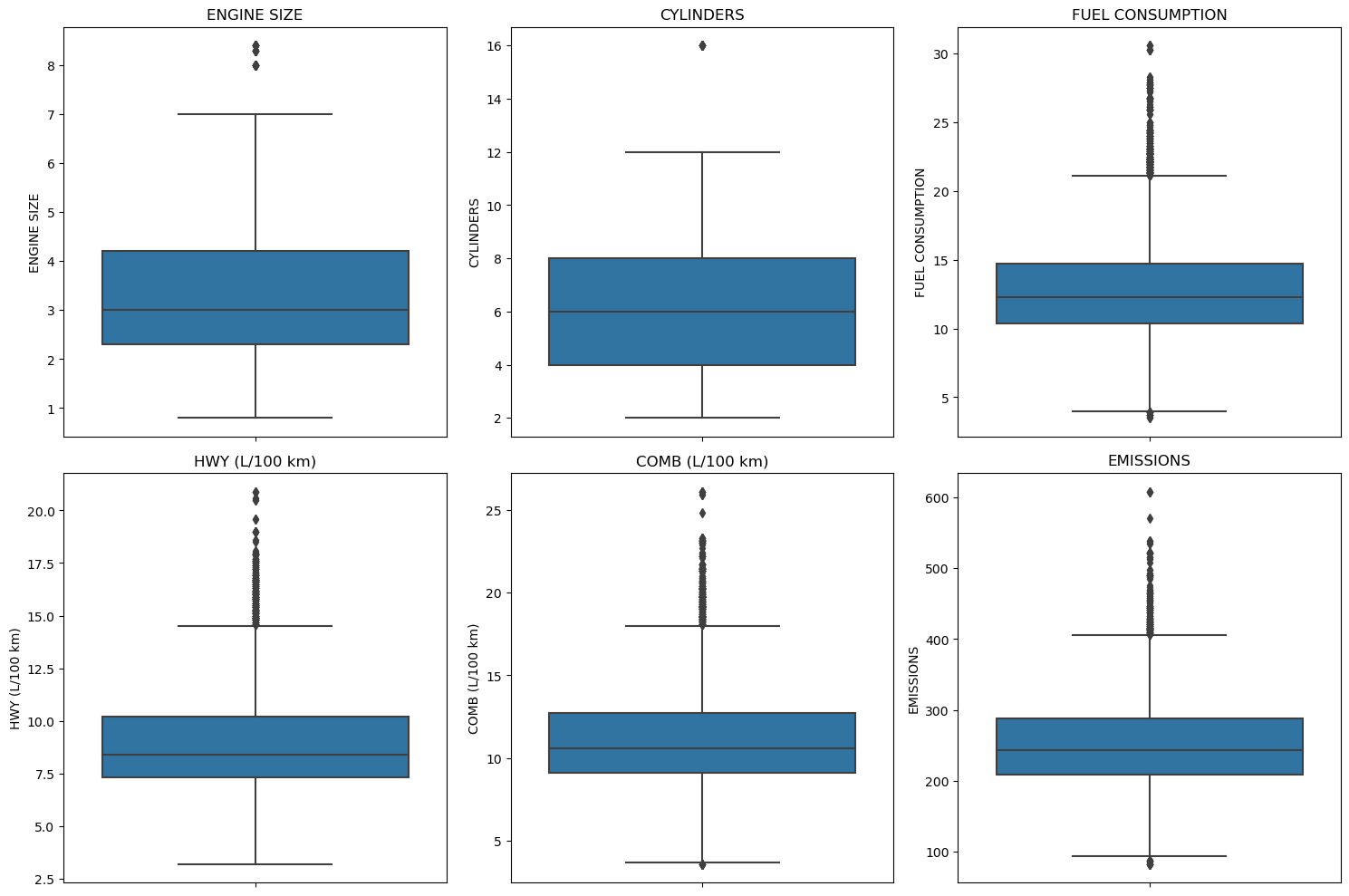
**10.) Checking for duplicates**



Output: We found a 1 duplicate in dataset so by using function drop, we drop those duplicate rows, now are dataset dimension is changes there are 22555 rows and 13 columns.

# Exploratory Data Analysis

1. **Identifying the outliers and removing**



Output: The box plots provide an overview of the key attributes in the vehicle dataset, highlighting the distribution and outliers for each attribute:

Engine Size: The median engine size is around 3 liters. Most vehicles have engines between 2 and 4 liters, with a few outliers above 7 liters.

Cylinders: The median number of cylinders is 6, with most vehicles having 4 to 8 cylinders. One outlier has 16 cylinders.

Fuel Consumption: The median fuel consumption is approximately 12 L/100 km. Most vehicles consume between 8 and 15 L/100 km, with outliers above 25 L/100 km and below 5 L/100 km.

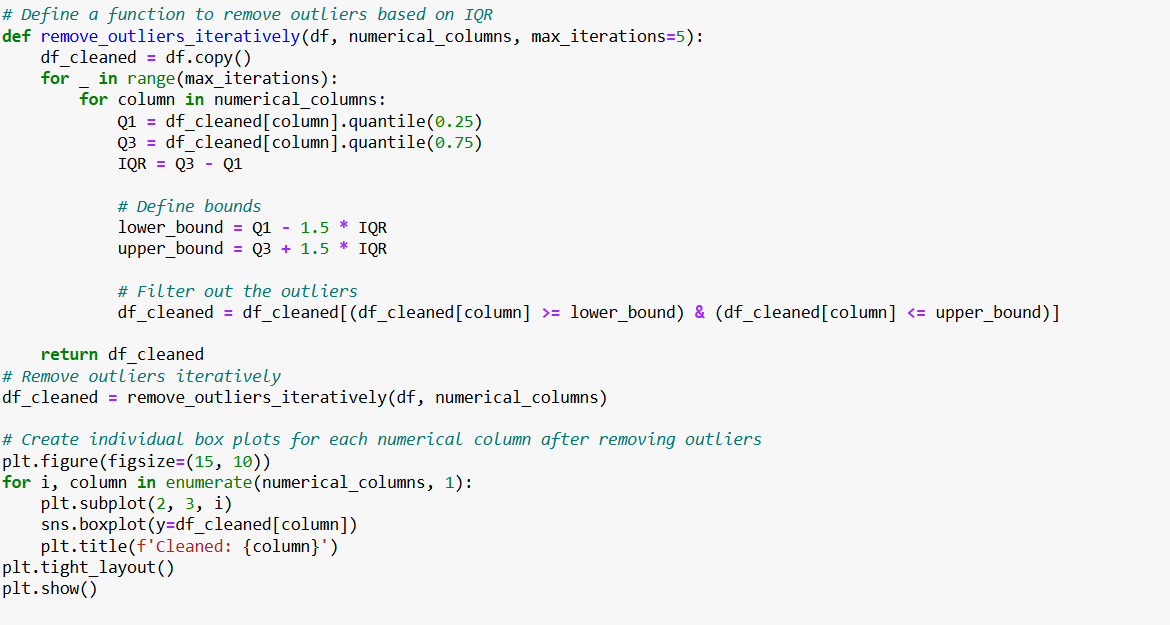
Highway Fuel Consumption (HWY L/100 km): The median is around 7.5 L/100 km, with most vehicles consuming 6 to 10 L/100 km on highways. Outliers exceed 17.5 L/100 km.

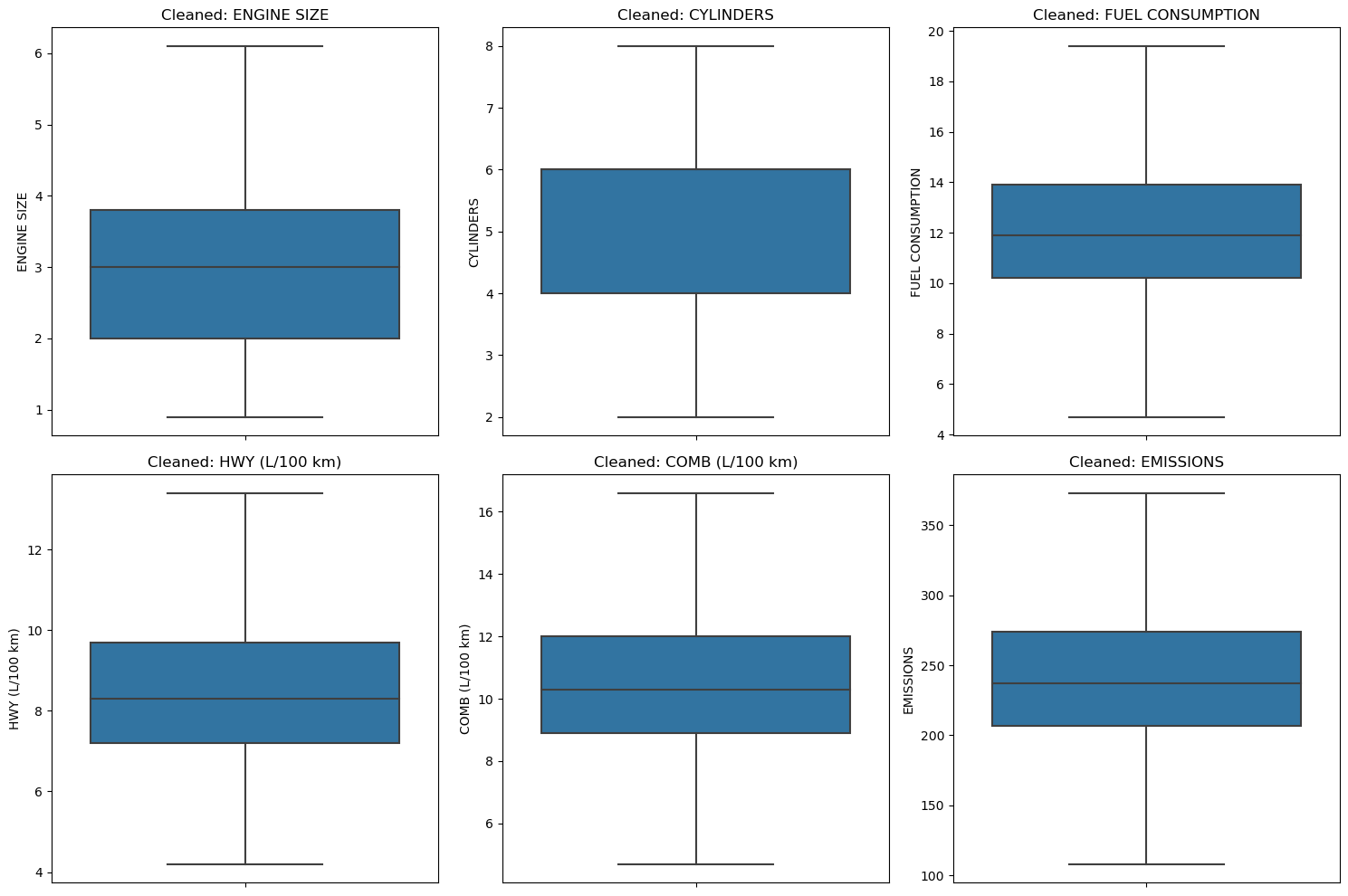
Combined Fuel Consumption (COMB L/100 km): The median is about 10 L/100 km. Most vehicles have combined consumption between 8 and 12.5 L/100 km, with outliers above 20 L/100 km.

Emissions: The median CO2 emissions are around 250 g/km, with most vehicles emitting 200 to 300 g/km. Outliers are above 500 g/km and below 100 g/km.

These box plots provide a quick summary of the central tendency, spread, and presence of outliers for each attribute. This helps in understanding the overall characteristics of the dataset and identifying any anomalies that may need further investigation.

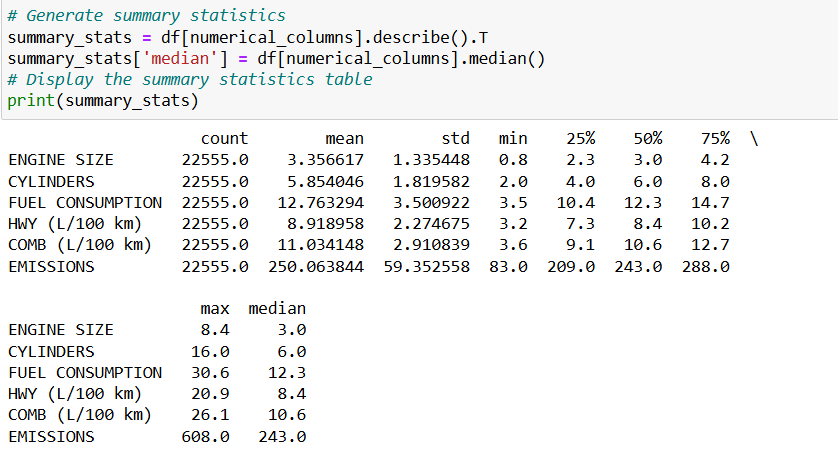
**Removing outliers**



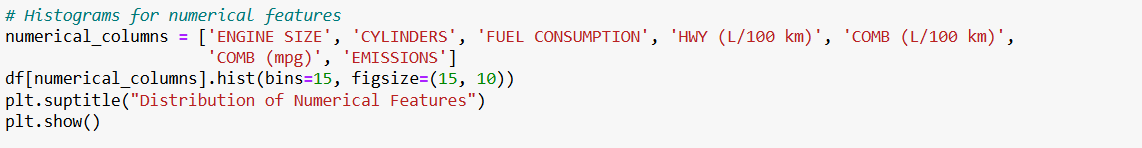


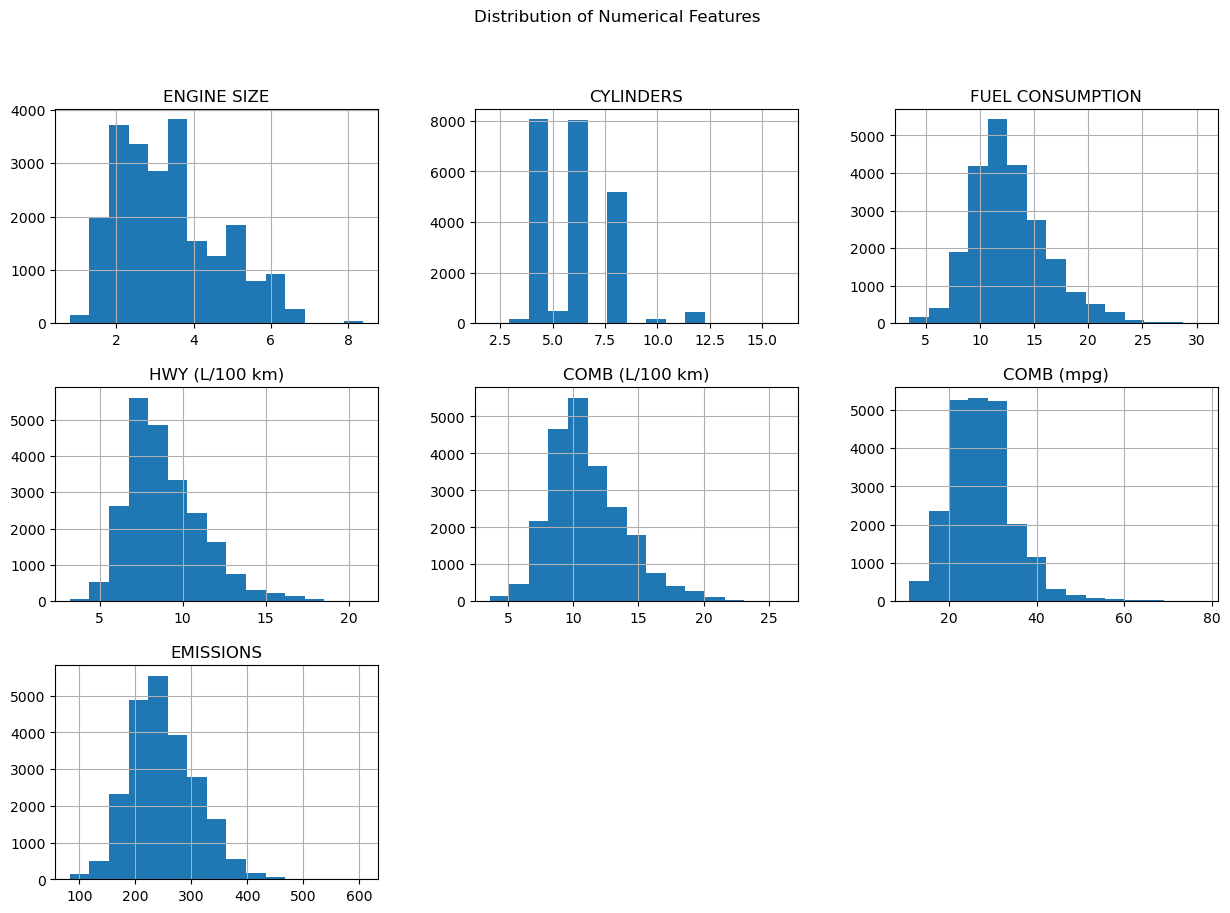
Output: To ensure the accuracy of the analysis, outliers were removed using the Interquartile Range (IQR) method. Values falling outside 1.5 times the IQR from the first and third quartiles were identified as outliers and removed iteratively (up to five times) to handle extreme cases effectively.

1. **Displaying summary statistics**



1. **Understanding the distribution**

Distribution of numerical features:



**OUTPUT:** The histogram explains various features:

Engine size: The engine sizes are mostly concentrated between 2.0 to 4.0 liters, with a peak around 3.0 liters. There is a right skew indicating fewer vehicles with larger engine sizes.

Cylinders: Most vehicles have 4, 6, or 8 cylinders. There is a significant drop in the number of vehicles with cylinders outside these values, indicating these are the most common configurations.

Fuel Consumption: The fuel consumption data is roughly normally distributed, with a peak around 10-15 L/100 km. The distribution shows that most vehicles consume between 5 and 20 L/100 km.

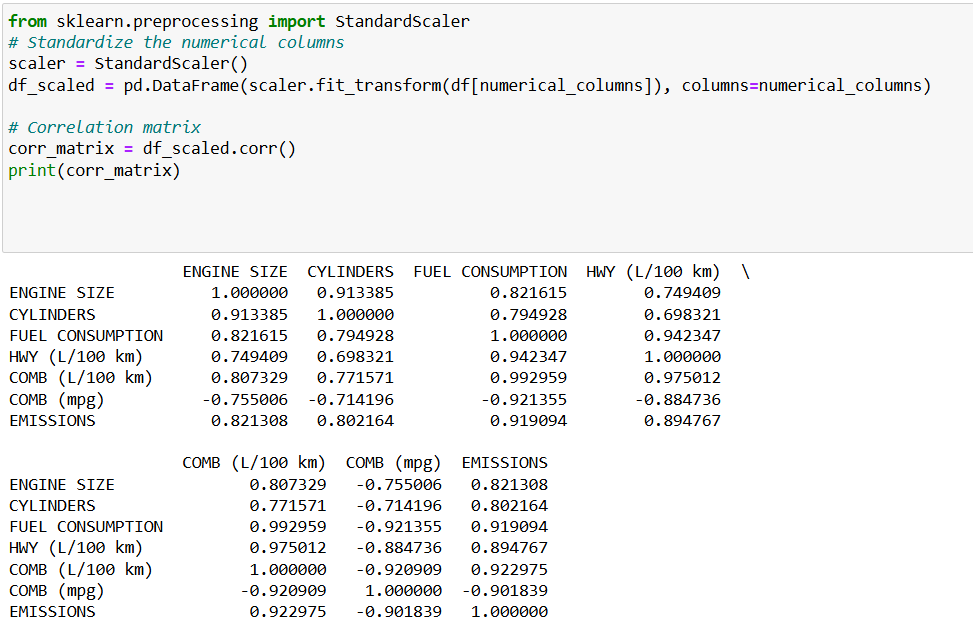
HWY (L/100 km): Highway fuel consumption follows a similar pattern to overall fuel consumption, with a peak around 5-10 L/100 km. The distribution is skewed to the right, indicating fewer vehicles with higher highway fuel consumption.

COMB (L/100 km): Combined fuel consumption also follows a normal distribution pattern, peaking around 10-15 L/100 km. This histogram is similar to the overall fuel consumption distribution.

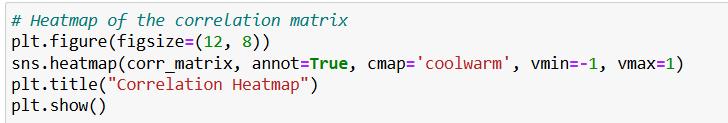
COMB (mpg): The combined miles per gallon (mpg) is inversely related to L/100 km. The distribution peaks around 20-25 mpg, indicating most vehicles fall within this range. There is a long right tail, showing fewer vehicles with higher mpg.

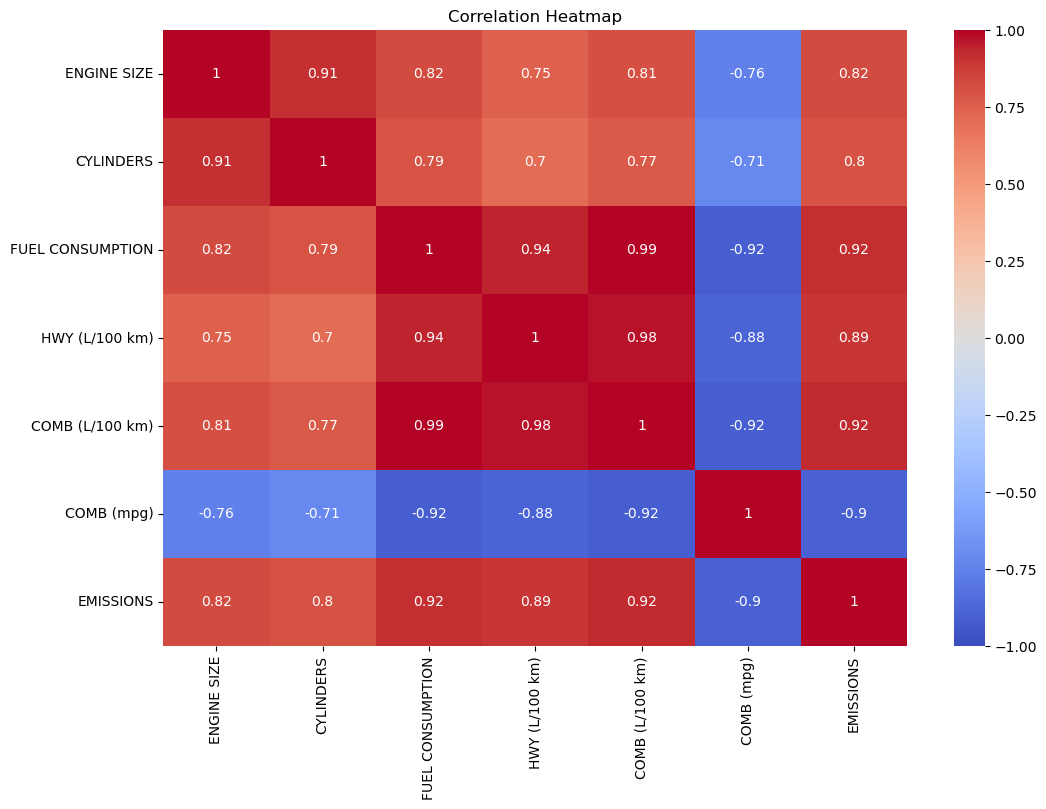
Emission: Emissions are roughly normally distributed, with a peak around 200-300 grams of CO2 per kilometer. The histogram shows a right skew, indicating fewer vehicles with higher emissions.

1. **Correlation Matrix**



**Heatmap**





**High Positive Correlations:**

**ENGINE SIZE and CYLINDERS** (0.91): Larger engines usually have more cylinders.

**FUEL CONSUMPTION and COMB (L/100 km) (0.99):** As fuel consumption goes up, combined fuel consumption (in L/100 km) almost always goes up in the same way.

**FUEL CONSUMPTION and EMISSIONS (0.92**): Vehicles that use more fuel tend to produce more emissions.

**COMB (L/100 km) and EMISSIONS (0.92):** Higher combined fuel consumption (in L/100 km) leads to higher emissions.

**High Negative Correlations:**

**COMB (mpg) and FUEL CONSUMPTION** (-0.92): Better fuel efficiency (in mpg) means lower fuel consumption.

**COMB (mpg) and EMISSIONS** (-0.90): Vehicles with better fuel efficiency produce fewer emissions.

**Moderate Positive Correlations:**

**ENGINE SIZE and EMISSIONS (0.82):** Larger engines are linked to higher emissions.

**CYLINDERS and EMISSIONS (0.80):** More cylinders usually mean more emissions.

**Moderate Negative Correlations:**

**COMB (mpg) and ENGINE SIZE (-0.76):** Bigger engines generally have lower fuel efficiency.

Implications:

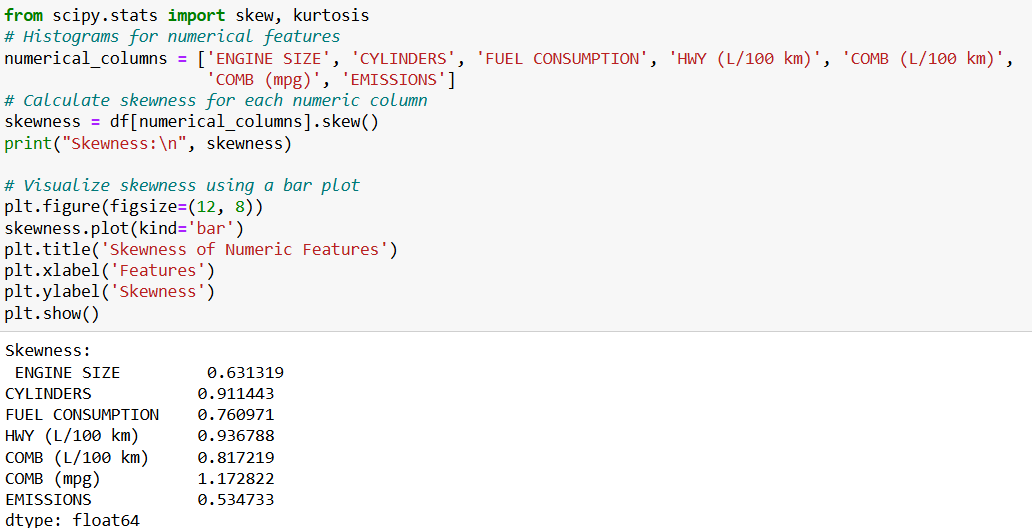
Vehicles with larger engines and more cylinders tend to use more fuel and emit more CO2.

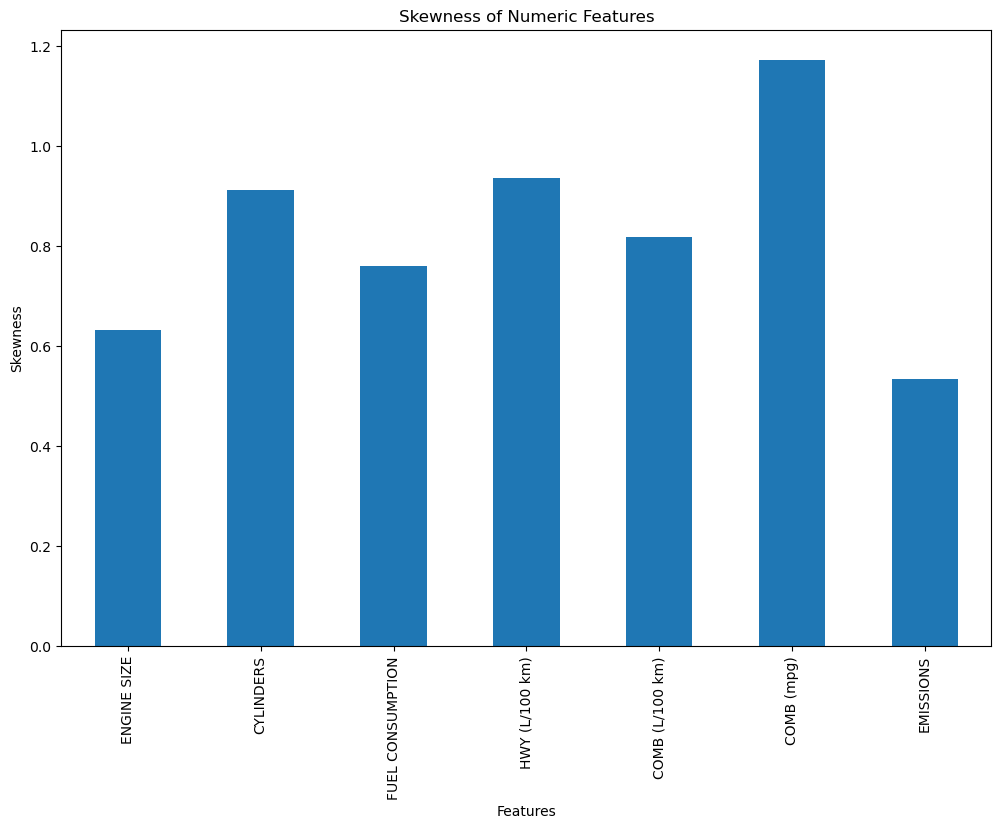
Improving fuel efficiency not only lowers fuel consumption but also reduces emissions.

This analysis highlights the importance of targeting fuel consumption in efforts to make vehicles more environmentally friendly.

1. **Skewness and kurtosis:**

**Skewness:**





Engine Size: The distribution of engine sizes is moderately skewed to the right, meaning there are more smaller engine sizes with fewer larger ones.

Cylinders: The number of cylinders in vehicles has a noticeable right skew, suggesting that most vehicles have fewer cylinders, with fewer vehicles having a larger number of cylinders.

Fuel Consumption: The fuel consumption data is moderately skewed to the right, indicating that more vehicles have lower fuel consumption, with fewer having higher consumption.

HWY (L/100 km): Highway fuel consumption also shows a moderate right skew, implying that most vehicles have relatively lower highway fuel consumption.

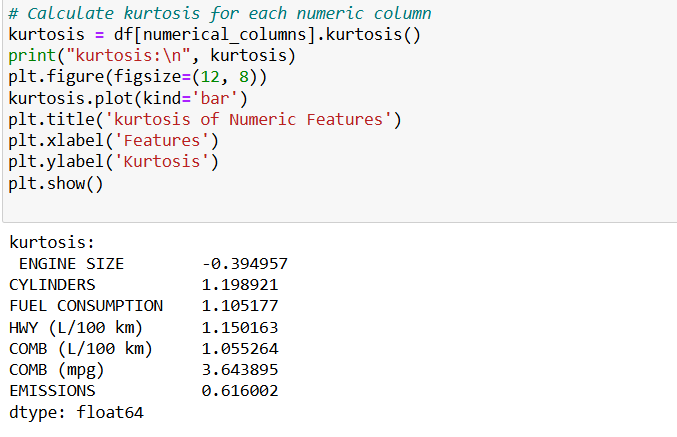
COMB (L/100 km): Combined fuel consumption is similarly skewed to the right, suggesting that most vehicles consume less fuel, with fewer consuming more.

COMB (mpg): Combined miles per gallon (mpg) is the most skewed to the right, indicating a significant number of vehicles with lower fuel efficiency (mpg) and fewer with higher efficiency.

Emission: Emissions data is slightly skewed to the right, showing that most vehicles emit lower amounts of CO2, with fewer emitting more.

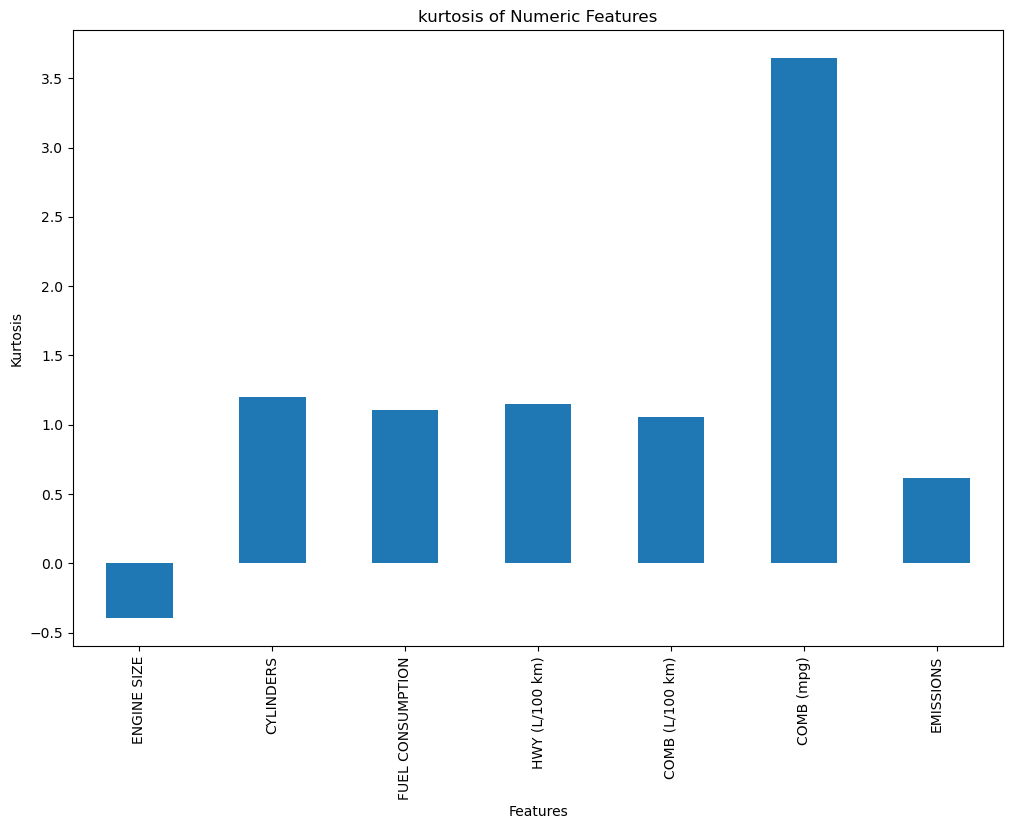
Overall, the chart indicates that most of the numerical features in our dataset are right-skewed, which means that majority of the data points are concentrated on the left side of the distribution, with fewer data points extending to the right.

**Kurtosis:**

****

**Kurtosis is a measure of the "tailedness" of a distribution, indicating how heavy or light the tails of a distribution are compared to a normal distribution.**

**Kurtosis is a measure of the "tailedness" of a distribution, indicating how heavy or light the tails of a distribution are compared to a normal distribution.**



**Output:**

Kurtosis is a measure of the "tailedness" of a distribution, indicating how heavy or light the tails of a distribution as compared to a normal distribution.

**High Kurtosis (COMB (mpg)):** The combined mpg feature stands out with a high kurtosis, indicating a lot of variability in fuel efficiency, with many vehicles having either very low or very high mpg values.

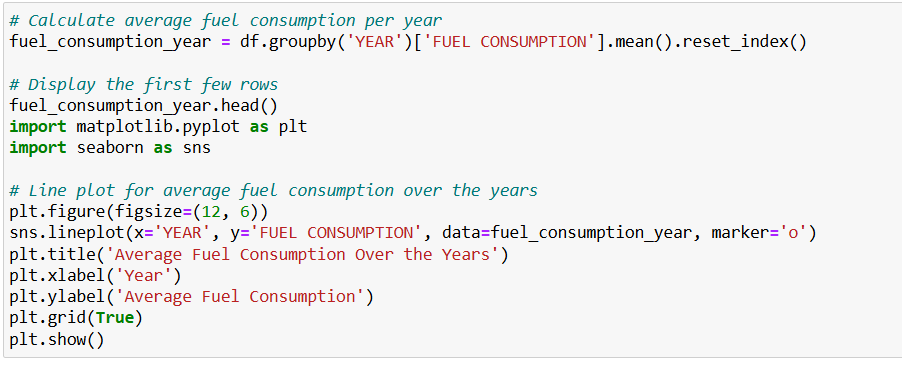
**Moderate Kurtosis** (CYLINDERS, FUEL CONSUMPTION, HWY, COMB (L/100 km)): These features have moderate kurtosis, indicating some extreme values but nothing too unusual.

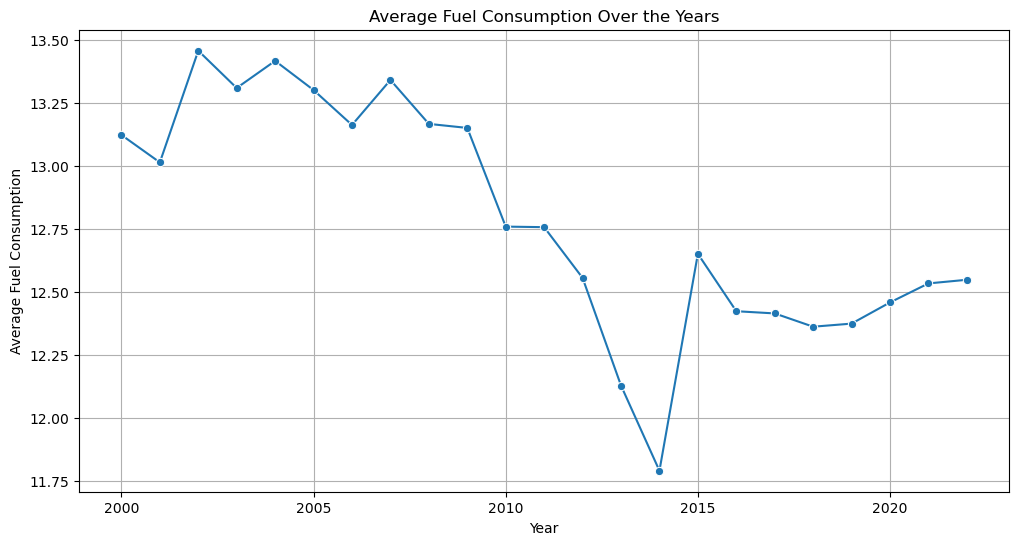
**Negative Kurtosis** (ENGINE SIZE, EMISSIONS): These features have relatively light tails, meaning fewer extreme values.

**Task 5:**

# Data Visualization

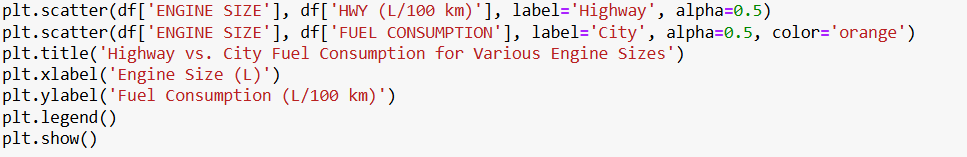
1. **Distribution of Fuel consumption versus year**

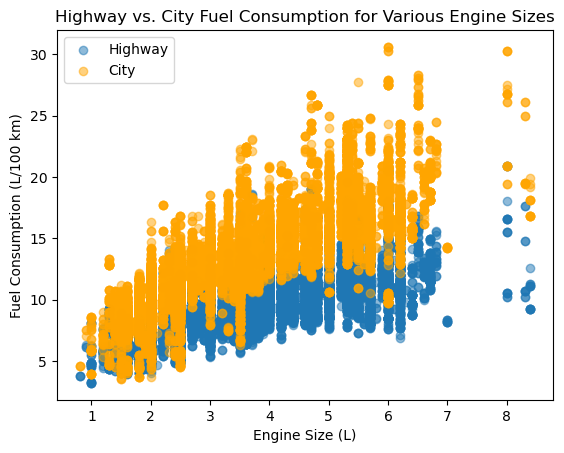
****



**Output:** From the graph, it can be seen that in 2000 majority of vehicle were consuming around 13 but between 2010 and 2015 average fuel consumption were reduced to 11, however after 2015 it again increased to approximately 12.

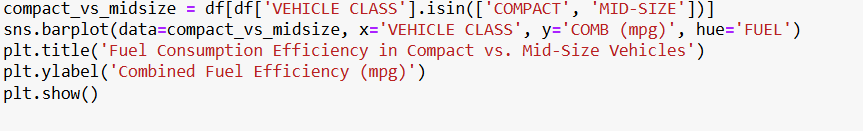
1. **Highway vs. City Fuel Consumption for Various Engine Sizes**

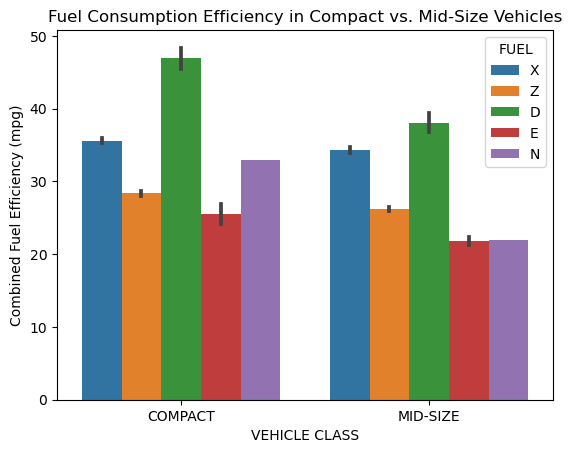
****



**Output:** From graph it can be seen that on highway fuel consumption is low as compared to city and greater the engine size more is fuel consumption.

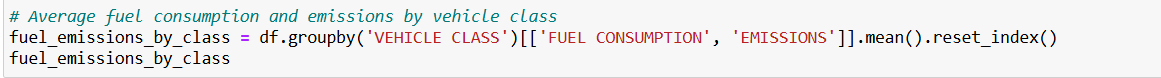
1. **Fuel Consumption Efficiency in Compact vs. Mid-Size Vehicles**

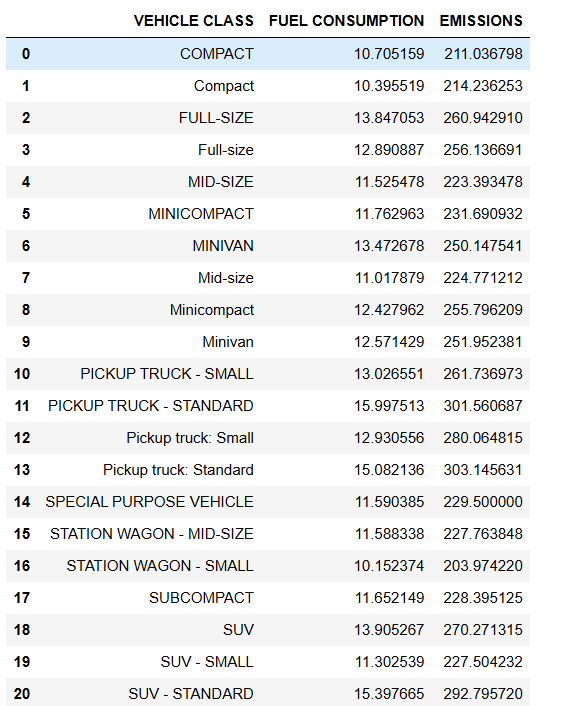
****

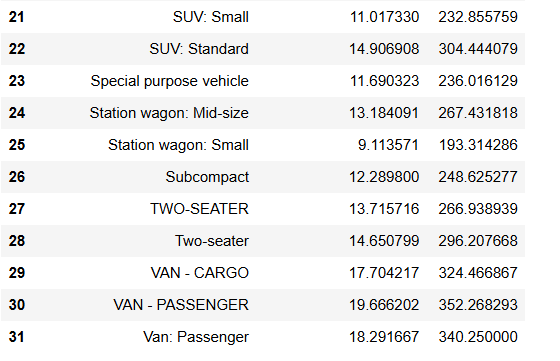


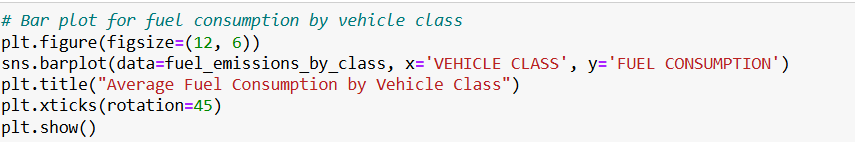
**Output:** Both vehicle class compact or mid-size which use diesel are more efficient as compared to other fuel type.

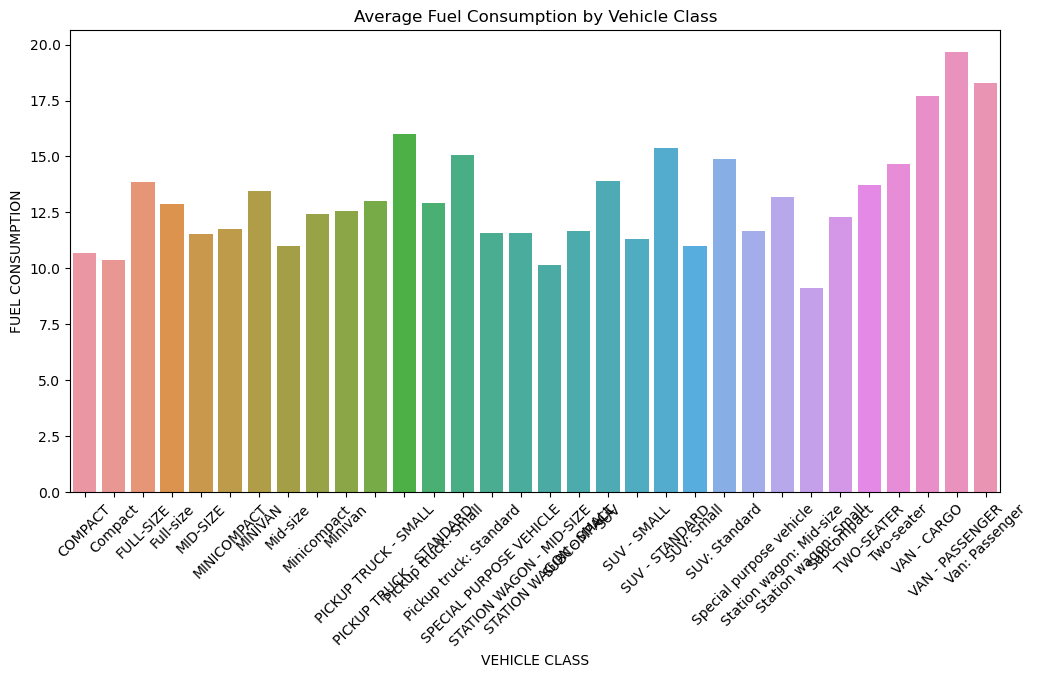
1. **Trends in Fuel Consumption and Emissions Across Different Vehicle Classes**











**Output:** This bar chart shows how much fuel different types of vehicles use on average.

**Vehicle Classes with High Fuel Consumption:**

Passenger Vans and Cargo Vans: These vehicles use the most fuel, with passenger vans using a bit more than cargo vans. Their large size and heavy weight likely make them less fuel-efficient.

Standard SUVs and Standard Pickup Trucks: These vehicles also consume a lot of fuel, which makes sense given their larger engines and heavier builds.

**Vehicle Classes with Moderate Fuel Consumption:**

Small Pickup Trucks, Small SUVs, and Special Purpose Vehicles: These vehicles fall in the middle when it comes to fuel consumption. They strike a balance between size, weight, and fuel efficiency.

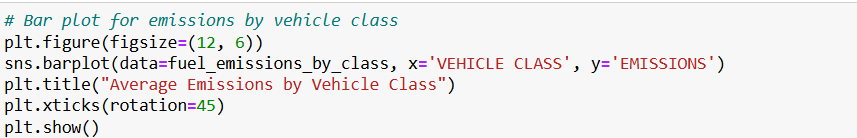
**Vehicle Classes with Low Fuel Consumption:**

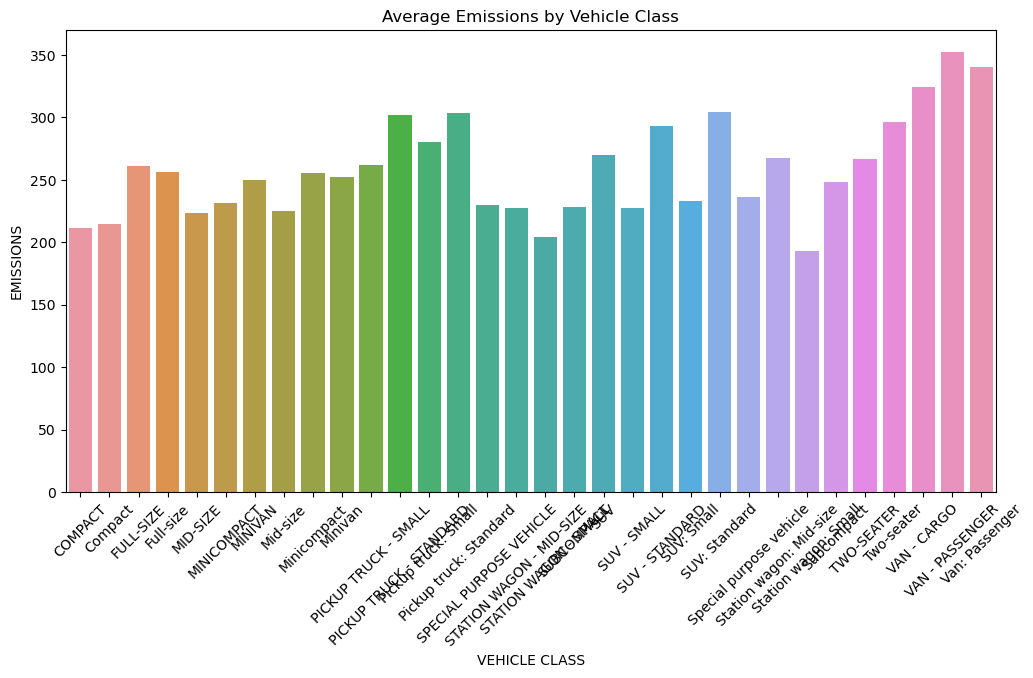
Compact and Subcompact Cars: These are the most fuel-efficient vehicles, using the least amount of fuel. Their smaller size and lighter weight contribute to better fuel economy.

Two-Seaters and Mini-Compact Cars: These also have lower fuel consumption, in line with their smaller and lighter designs.

**Interpretation:**

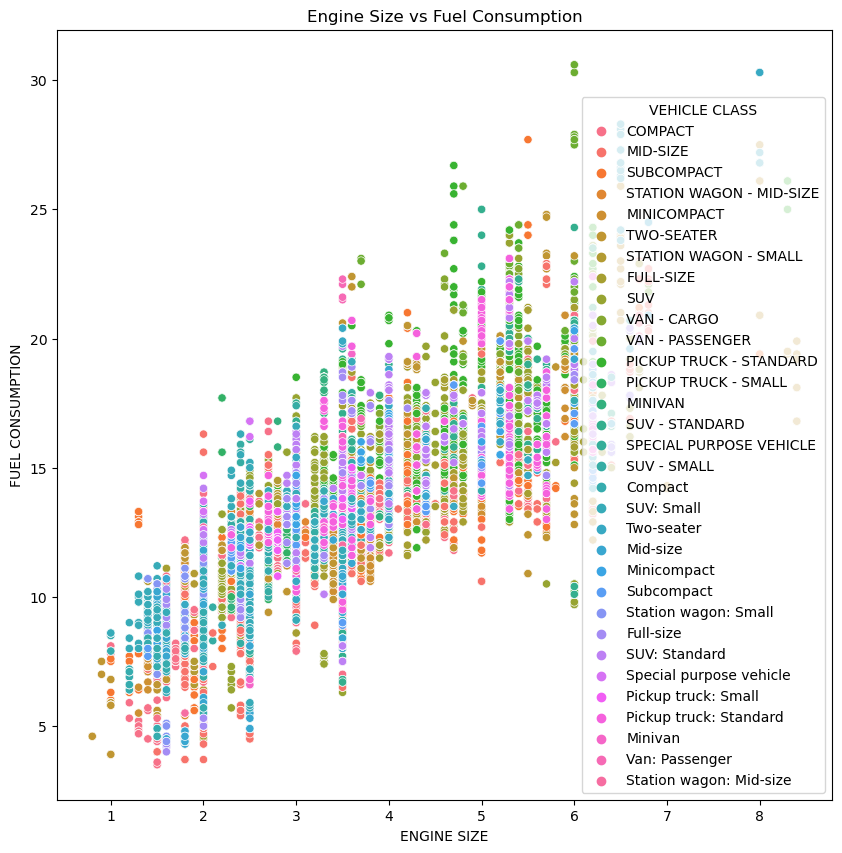
Larger and heavier vehicles tend to use more fuel, which is evident from the high consumption seen in vans, standard SUVs, and pickup trucks. Smaller vehicles, like compact and subcompact cars, are more fuel-efficient, making them better for those who want to save on fuel costs and reduce their environmental impact. This information can help consumers and policymakers make better decisions about which vehicles to buy or promote. Opting for more fuel-efficient vehicle classes could lead to lower overall fuel consumption and emissions.

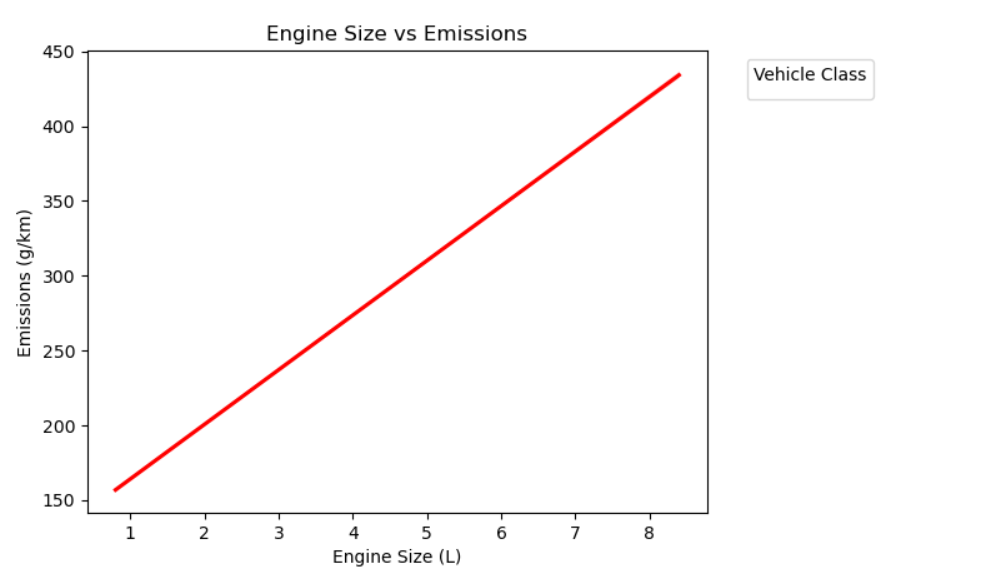




There is a clear link between how much fuel a vehicle uses and how much it pollutes. Vehicle classes that use more fuel tend to emit more pollutants. Compact and subcompact cars are better for the environment because they produce fewer emissions.

1. **Impact of Engine Size on Fuel Consumption and Emissions.**



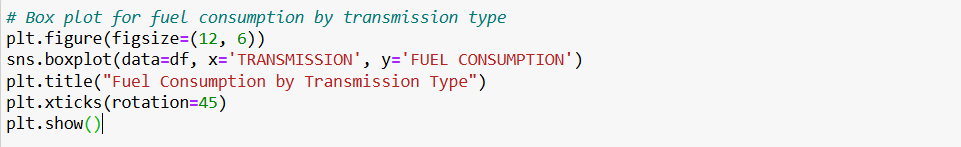
****

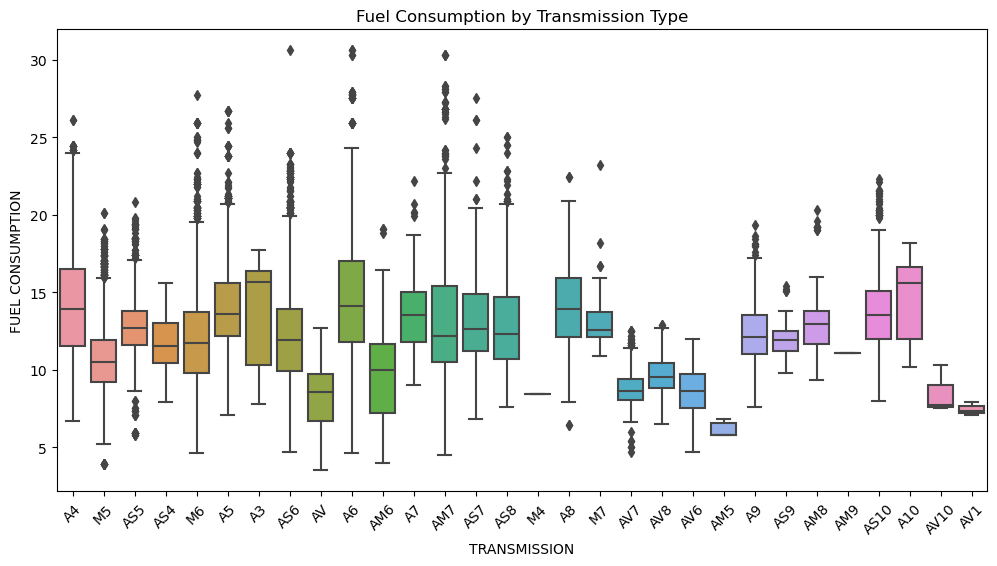
Vehicles with larger engines (4-8 liters) typically have higher fuel consumption, with few exceptions. Whereas, vehicles with smaller engines (1-3 liters) have a wide range of fuel consumption, suggesting that other factors like vehicle weight or design also impact fuel use. Moreover, the plot shows that there is positive correlation engine size and fuel consumption.

Compact, subcompact, and minicompact cars, typically represented by lighter colors in the scatter plot, are characterized by their smaller engine sizes and lower fuel consumption. On the other hand, SUVs, vans, and pickup trucks, which are shown in darker colors, generally have larger engines and consequently consume more fuel.

There are a few points on the plot where vehicles with smaller engines still use a lot of fuel. This might be due to heavy loads, poor fuel efficiency, or specific design choices.

1. **Effect of Transmission Type on Vehicle Fuel Efficiency**

****



**Output:**

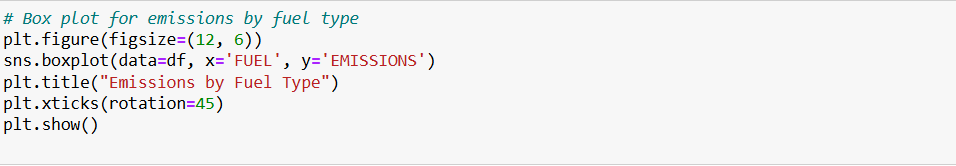
**Automatic (A):** These tend to have a wider range of fuel consumption, with more outliers, indicating that fuel efficiency can vary quite a bit depending on the model.

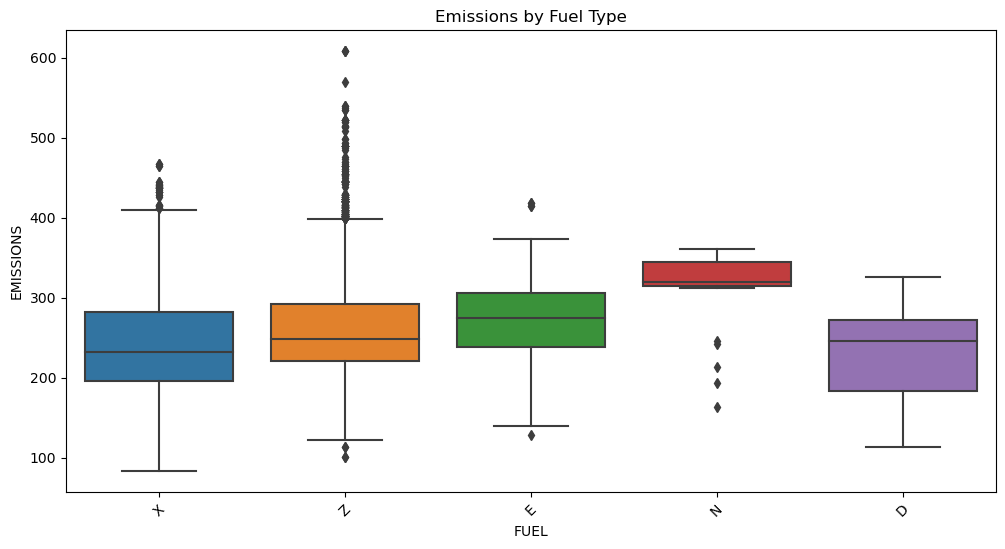
**Manual (M):** These show different levels of fuel consumption, with some types like M5 using more fuel and others like M6 using less.

**Automated Manual (AM):** These typically have higher fuel consumption and a wide range of values.

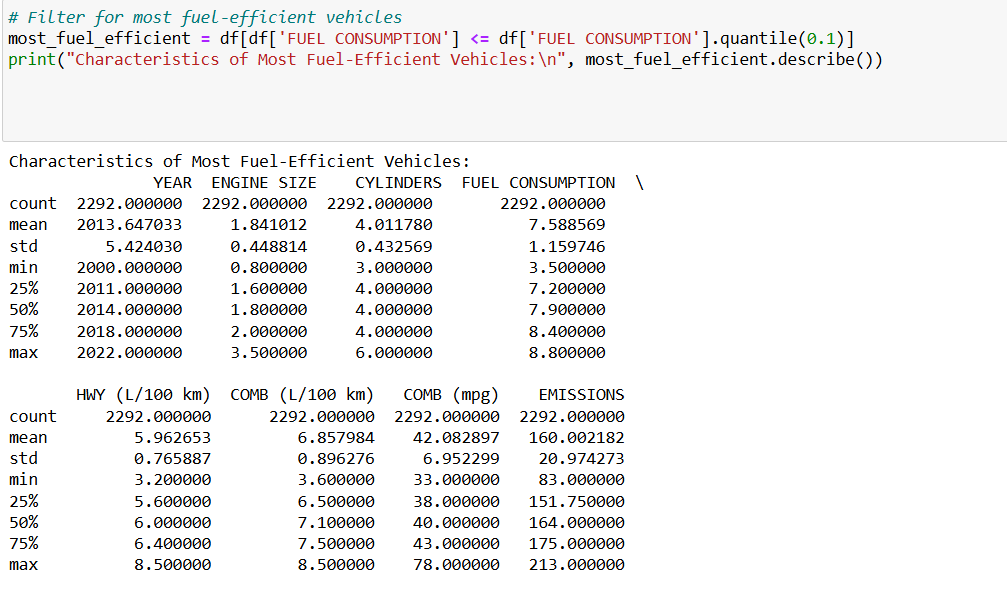
**Continuously Variable (AV):** These generally show lower and more consistent fuel consumption, particularly types like AV6 and AV10.

1. **Fuel Types Associated with Lower Emissions and Better Fuel Efficiency**

****



1. **Characteristics of the Most Fuel-Efficient and Least Polluting Vehicles**

****

**Output:**

The average model year for these vehicles is approximately 2013.6, with most cars ranging from 2011 to 2018, which suggests that newer vehicles tend to be more fuel-efficient.

The average engine size is around 1.84 liters, with most engines falling between 1.6 and 2.0 liters. These smaller engines are typical of vehicles designed for fuel efficiency.

Most fuel-efficient vehicles have around 4 cylinders, which is common for small to mid-sized engines. The data shows that 4-cylinder engines contributing to better fuel economy.

City Fuel Consumption: On average, these vehicles consume about 7.59 liters per 100 km in city driving.

Highway Fuel Consumption: They average about 5.96 liters per 100 km on highways.

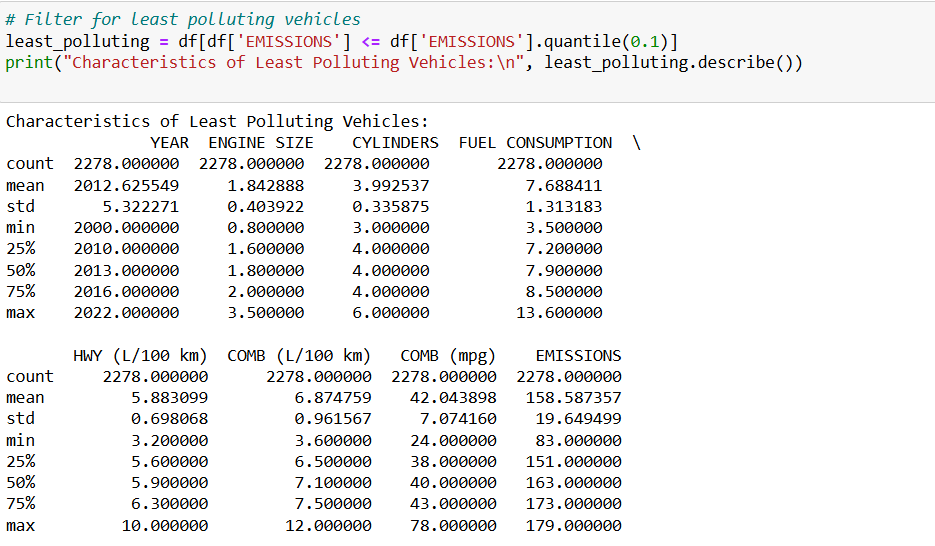
Combined Fuel Consumption: The combined average fuel consumption is 6.86 liters per 100 km.

The average fuel efficiency is about 42 miles per gallon (mpg).

The average CO2 emissions for these vehicles are around 160 grams per kilometer.

Lower emissions suggest that these vehicles are not only fuel-efficient but also environmentally friendly.

**Least polluting vehicles:**

****

**Output:**

Vehicles from around 2010 to 2016 with smaller engines and 4 cylinders are typically the least polluting. These vehicles are fuel-efficient, which directly leads to lower emissions.

The data visualizations provided clear insights into how automotive manufacturers can design vehicles to improve fuel efficiency and reduce CO2 emissions:

Vehicle Class Trends:

Larger vehicles like vans and SUVs use more fuel and emit more CO2. Smaller vehicles like compacts are more fuel-efficient and produce fewer emissions.

Engine Size Impact:

Bigger engines lead to higher fuel consumption and emissions. Manufacturers should focus on developing smaller, more efficient engines.

Transmission Types:

Some transmissions, like CVT, are more fuel-efficient. Using these in new vehicle designs can help improve fuel efficiency.

Fuel Types:

Vehicles using alternative fuels, like hybrids, have lower emissions. Manufacturers should explore these fuel options to reduce CO2.

Efficient Vehicle Characteristics:

The most efficient vehicles have smaller engines and fewer cylinders. Designing vehicles with these features can help reduce emissions.

**Overall Recommendation:**

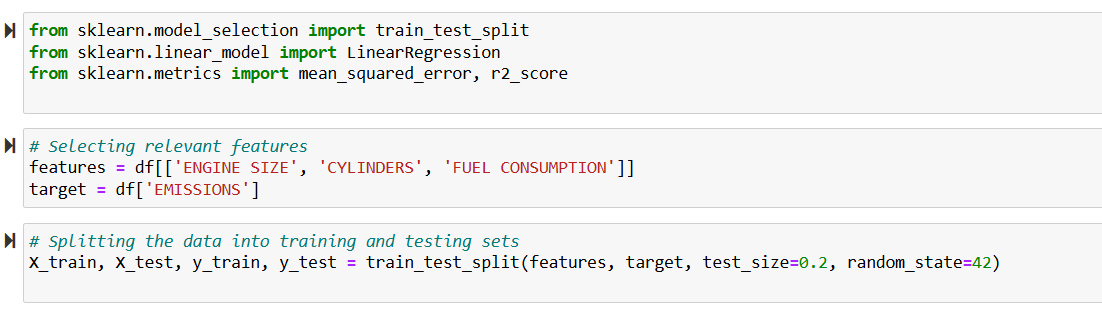
Manufacturers should focus on smaller, efficient engines, fuel-saving transmissions, and alternative fuels to create vehicles that are both fuel-efficient and environmentally friendly. This approach will help reduce CO2 emissions and meet consumer demands for greener vehicles.

**Task 6:**

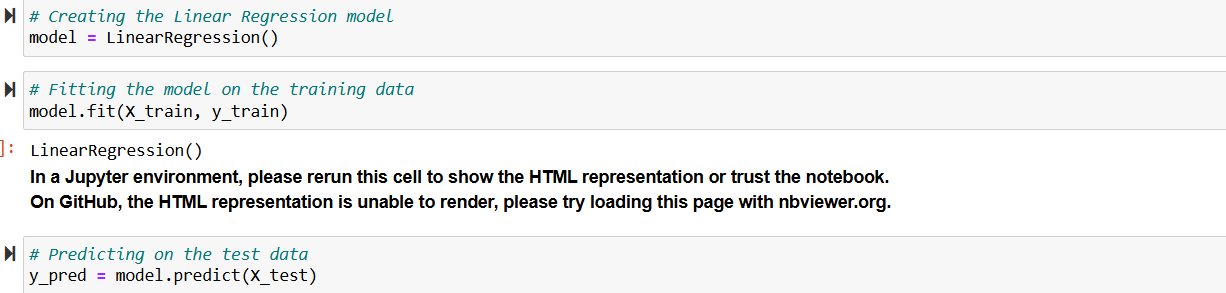
# Predictive Model

**Linear Regression Model**

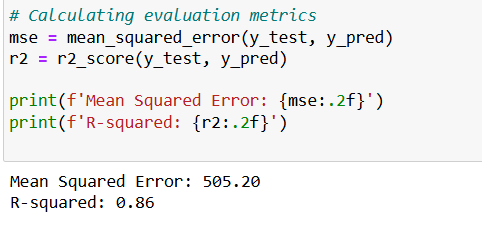
**Step 1: Data preparation**

****

**Step 2: Building the model:**

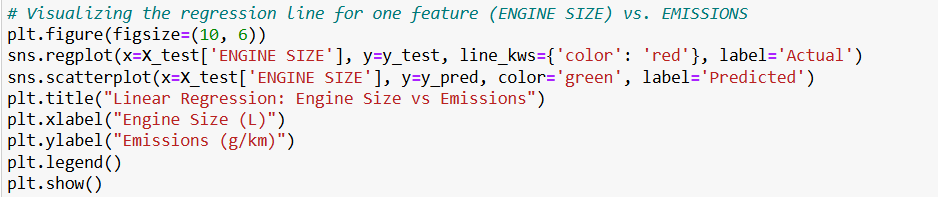
****

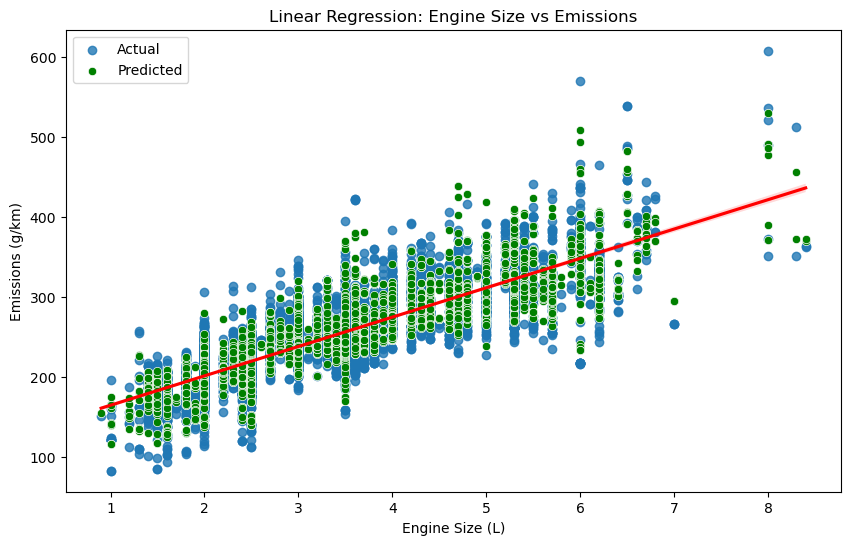
**Step 3: Evaluating the model**

****

**As R² value of 0.86 suggests that our model is performing quite well, as it indicates a strong relationship between the features and the target variable.**

**Visualization**

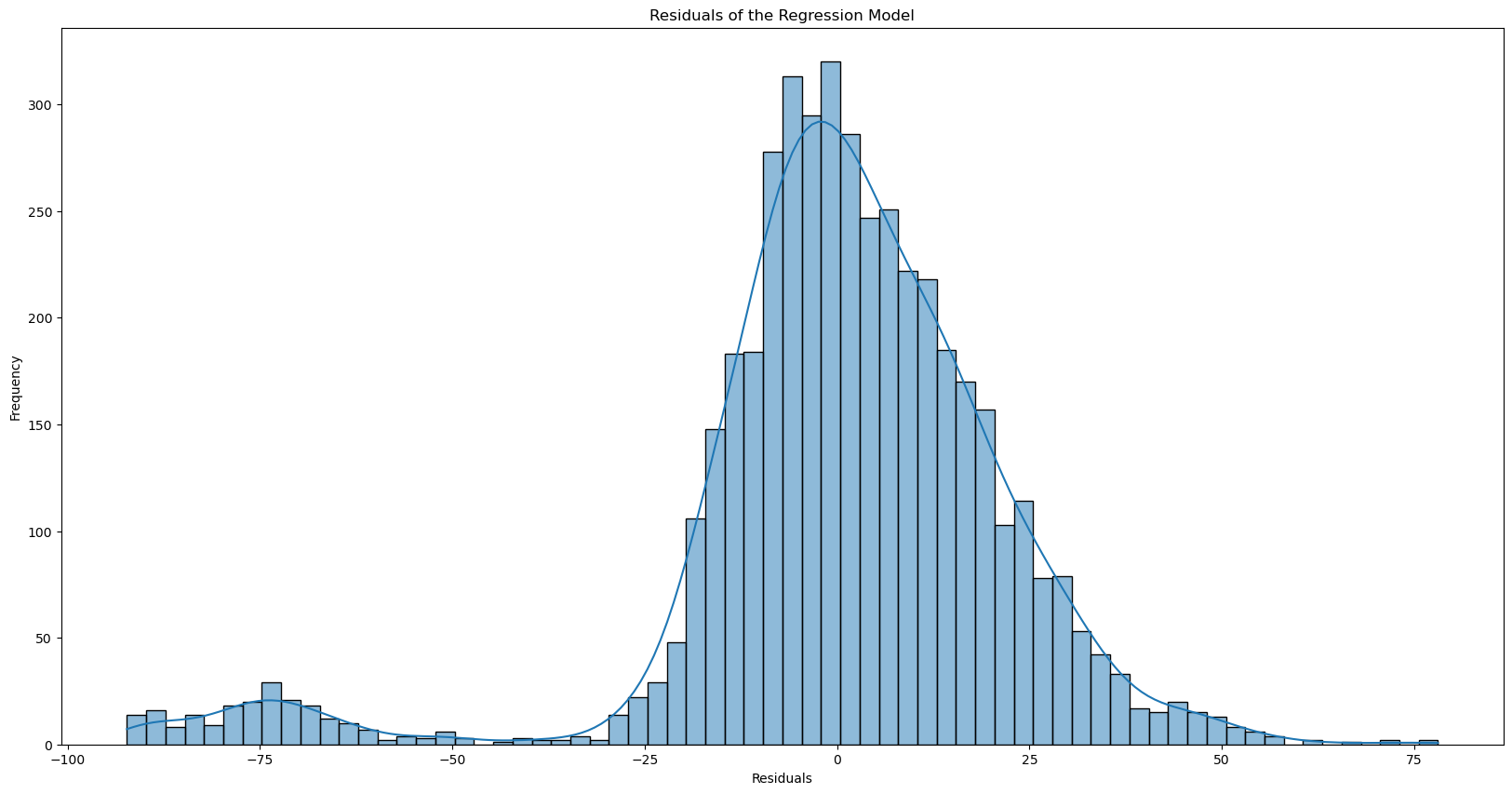
****



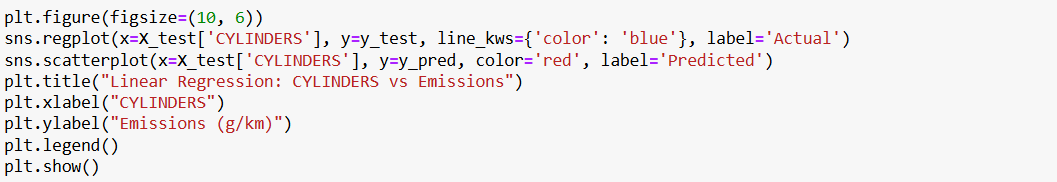
**The linear regression line suggests a positive correlation between engine size and emissions, meaning that larger engines are generally associated with higher emissions.**

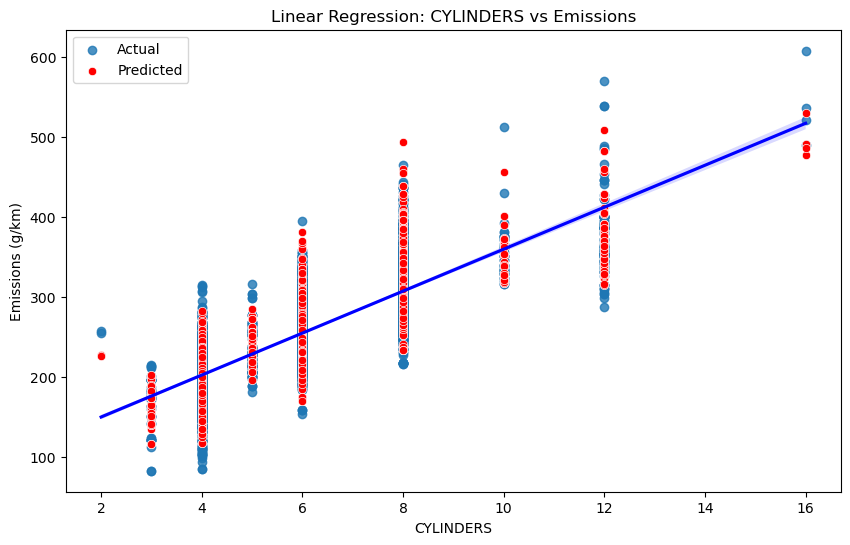
**The scatter of the data points around the regression line indicates the model's fit. If the points were closer to the line, the model would be a better fit.**

**Residual plot:**



**Output:** The histogram of residuals shows that they are centered around zero and approximately follow a normal distribution. This suggests that the regression model fits the data well, with most predictions being close to the actual values**.**

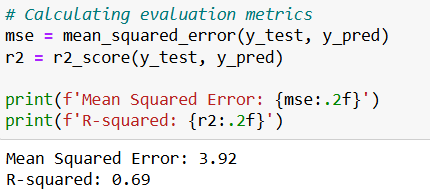
****



**The regression line shows a positive correlation between the number of cylinders and emissions. This suggests that vehicles with more cylinders generally produce higher emissions.**

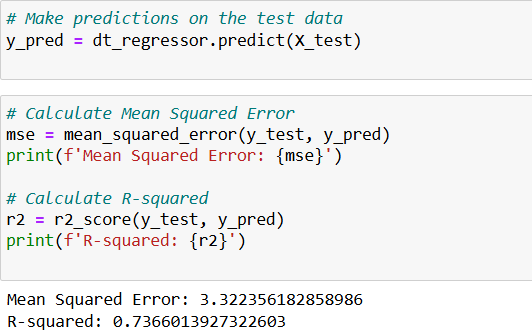
**The predicted values (red dots) closely follow the actual values (blue dots), especially for vehicles with fewer cylinders (between 2 and 8). This indicates that the model is relatively accurate in this range. However, there is more deviation in the actual values for vehicles with a higher number of cylinders (12 to 16), which suggests the model not capturing all the variability in emissions for these vehicles.**

**Changing target variable to fuel consumption:**

****

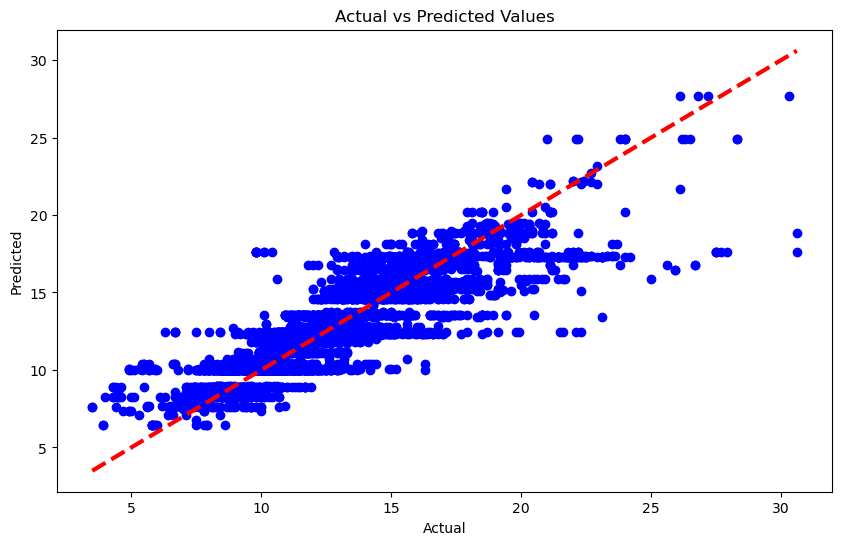
**Output: As R^2 is 0.69 which shows that it is 69 % accurate so we choose another model.**

**Decision Tree regressor**

****

**Output: By evaluating the through decision tree regressor we found 73% accuracy.**

**Visualization:**



**Output: Dots close to the red line indicate good predictions. Here, most dots are near the line, showing the model is fairly accurate.**