CO2 Emission by Vehicles: An Empirical Analysis

Abstract - Environmental sustainability is a pressing concern in the modern era, with the automotive sector being a significant contributor to global CO2 emissions. This study centers around a comprehensive dataset sourced from the Canadian Government's open data portal, detailing vehicular attributes and their corresponding CO2 emissions over a span of seven years. Through rigorous exploratory data analysis and inferential statistics, we identified critical vehicular features influencing emissions, such as engine size, number of cylinders, and fuel consumption metrics. The analysis underscored the positive correlation between engine size and emissions, while also highlighting the inverse relationship between fuel efficiency (in mpg) and CO2 emissions. This research provides valuable insights for policymakers, manufacturers, and consumers, emphasizing the importance of fuel-efficient vehicles in reducing the global carbon footprint.

Keywords— CO2 Emissions, Automotive Industry, Fuel Efficiency, Engine Size, Cylinders, Environmental Sustainability, Data Analysis, Canadian Vehicle Data, Fuel Consumption, Predictive Modeling

1. OVERVIEW OF THE PROJECT

1.1 Background:

As global awareness about environmental sustainability and climate change grows, the automotive industry finds itself at the crossroads of innovation and responsibility. Carbon dioxide (CO2) emissions from vehicles account for a significant portion of total greenhouse gas emissions, having a direct impact on global warming. Given the potential consequences, there's a pressing need to understand the factors influencing these emissions.

1.2 Dataset Overview:

Our study revolves around a dataset sourced from the official open data portal of the Canadian Government [1,2]. This dataset sheds light on the nuances of how CO2 emissions by vehicles can differ based on various attributes. Incorporating data from a span of seven years, this comprehensive collection encompasses 7,385 data points across 12 columns. It meticulously details attributes ranging from the model, transmission, fuel type, fuel consumption, to the CO2 emissions.

1.3 Goals:

- **Understand Influencing Factors:** Delve into the dataset to discern the influence of various vehicular attributes on CO2 emissions.
- **Identify Key Features:** Highlight the most influential features affecting emissions, guiding future research and vehicular design.
- **Analyze Fuel Consumption:** Investigate the disparities in CO2 emissions between city and highway fuel consumption and their combined metric.

1.4 Expected Outcomes:

By the end of this project, we aim to provide clear insights into the features most influencing vehicular CO2 emissions. This analysis will not only benefit manufacturers in designing eco-friendlier vehicles but also guide policymakers in crafting informed regulations.

2. CRITICS

- Given that the dataset is compiled over a period of seven years, it's essential to understand the context in which the data was collected. Over the years, advancements in vehicle technology, changes in fuel standards, and shifts in consumer preferences could have influenced emission patterns.

- Previous studies leveraging similar datasets [2] have primarily focused on singular attributes like fuel type or transmission. While informative, these studies potentially overlook the intricate interplay of multiple vehicles features and their collective impact on emissions.

- There exists a knowledge gap, especially when considering the combined effect of features like fuel consumption (city vs. highway) and other vehicular attributes. Our analysis aims to bridge this gap, offering a holistic view of the influencing factors.

3. Specification

Tools: For the successful completion of this project, we'll employ a suite of tools optimized for data analysis and visualization:

- **Python**: As the primary programming language, Python offers robust libraries and frameworks tailored for data analysis.

- **pandas**: A versatile library for data manipulation and analysis.
- **seaborn and matplotlib:** For data visualization, these libraries offer a plethora of options to visualize complex datasets.
- **scikit-learn:** A machine learning library to help in predictive modeling and statistical analysis.

Implementation: The project is structured in phases:

1. **Data Preprocessing:** Cleaning the dataset, handling missing values, and ensuring data consistency.
2. **Exploratory Data Analysis:** Visualizing the dataset to understand patterns, correlations, and distributions.
3. **Inferential Statistics:** Delving deeper into the dataset to understand relationships and draw inferences.
4. **Predictive Modeling:** Building models to predict CO2 emissions based on vehicle attributes.

4. Milestones:

- **Week 1-2:** Data Preprocessing and Initial Exploration.
- **Week 3**: In-depth Exploratory Data Analysis.
- **Week 4:** Inferential Statistics and Hypothesis Testing.
- **Week 5:** Predictive Model Design and Training.
- **Week 6:** Model Testing, Evaluation, and Refinement.
- **Week 7:** Documentation, Report Writing, and Visualization Finalization.
- **Week 8:** Review, Final Presentation, and Submission.

5. Responsibilities of Each Group Member:

Given that each member is entrusted with analyzing one variable, responsibilities are allocated as follows:

- **K.Frances:** Analysis of the 'Engine Size(L)' variable, including visualization, statistical analysis, and relationship with CO2 emissions.
  Delving into the 'Cylinders' variable, understanding its distribution, and discerning its impact on emissions.
- **P.Nikethan** Exploring the 'Fuel Consumption Comb (L/100 km)' variable, its distribution, and correlation with emissions.
- **Mahammad khasim:** Analyzing the 'Fuel Consumption Comb (mpg)' variable, visualizing its patterns, and determining its influence on CO2 emissions.

6. Exploratory Data Analysis:

6.1 Data Cleaning and Preprocessing:

Missing Values: Our comprehensive review of the dataset revealed that there are no missing values across all columns. This is a positive indication of the dataset's completeness, reducing the need for imputations or data augmentations.

Duplicate Values: Our analysis detected 1,103 duplicate rows. The presence of these duplicates warrants further investigation. Depending on the nature of the dataset and the reason for duplication (e.g., data entry errors, data merging issues), we might need to consider removing these duplicates to prevent any potential biases in our subsequent analyses.

```
print(missing_values)
print (" number of duplicate rows are", data.duplicated().sum())
```

```
Make                              0
Model                             0
Vehicle Class                     0
Engine Size(L)                    0
Cylinders                         0
Transmission                      0
Fuel Type                         0
Fuel Consumption City (L/100 km)  0
Fuel Consumption Hwy (L/100 km)   0
Fuel Consumption Comb (L/100 km)  0
Fuel Consumption Comb (mpg)       0
CO2 Emissions(g/km)               0
dtype: int64
 number of duplicate rows are 1103
```

Fig 1: No Missing Values

Data Types Inspection:

Ensuring appropriate data types is crucial for accurate data analysis. Below is the summary of the data types associated with each column:

```
: data.dtypes
```

```
: Make                              object
  Model                             object
  Vehicle Class                     object
  Engine Size(L)                   float64
  Cylinders                          int64
  Transmission                      object
  Fuel Type                         object
  Fuel Consumption City (L/100 km)  float64
  Fuel Consumption Hwy (L/100 km)   float64
  Fuel Consumption Comb (L/100 km)  float64
  Fuel Consumption Comb (mpg)         int64
  CO2 Emissions(g/km)                 int64
  dtype: object
```

6.2 Dependent Variables:

In our exploration of the CO2 Emission by Vehicles dataset, understanding the behavior and characteristics of our primary dependent variable, CO2 Emissions(g/km), is paramount. This section of the report details our findings related to this variable, offering insights into its distribution, central tendency, and spread.

Histogram Analysis:

The histogram provides a visual representation of the frequency distribution of CO2 emissions across vehicles. Our observations indicate:

- A slightly right-skewed distribution, suggesting that while most vehicles emit CO2 within a specific range, a subset of vehicles exhibit notably higher emissions.
- This skewness potentially points towards the existence of high-emission vehicles in the dataset, possibly larger vehicles or those equipped with more powerful engines.

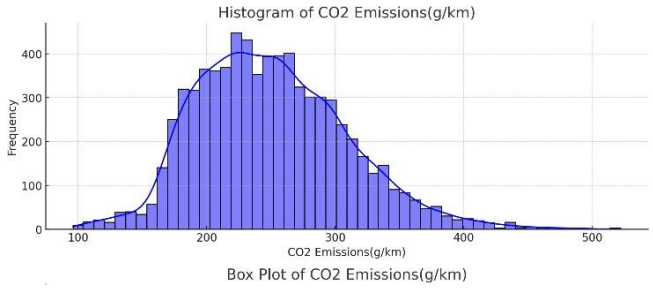

Fig2: Histogram

Box Plot Analysis (Car Price Distribution):
The box plot offers a concise summary of the distribution's central tendency, variability, and potential outliers. Our key takeaways are:

- The median CO2 emission, represented by the line within the box, is around the 246g/km mark.
- Outliers, particularly on the higher end, are discernible. These outliers corroborate our histogram's observation of vehicles with exceptionally high emissions.

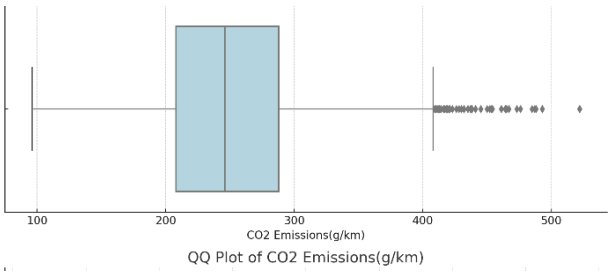

Fig3: Box Plot

QQ Plot Analysis (Normality Check for Car Price):

The Quantile-Quantile (QQ) plot is instrumental in assessing the normality of our data distribution. Key insights from the QQ plot include:

- The deviation from the straight line, especially at the tails, signals that our emission data doesn't adhere perfectly to a normal distribution.
- While the central data points align closely with the theoretical quantiles of a normal distribution, the tails deviate, emphasizing the skewness detected in the histogram.

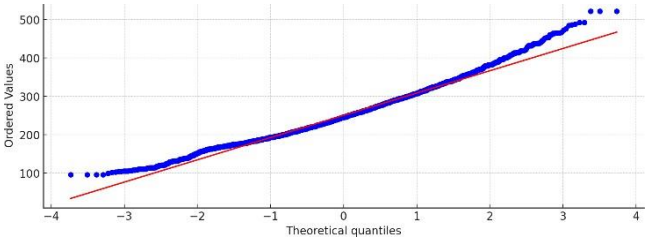

Fig4: QQ Plot

Descriptive Statistics
A deeper quantitative understanding of the CO2 emissions is attained through the computed descriptive statistics:

The insights derived from these statistics are:

- Central Tendency: The mean emission, at approximately 250.58 g/km, provides an average measure across the dataset. The median, at 246.00 g/km, offers a more robust measure, indicating that half of the vehicles in the dataset have emissions below this value and half above.
- Spread: A standard deviation of 58.51 g/km underscores the variability in the CO2 emissions across the dataset. This moderate spread indicates the diversity of vehicles in terms of their emission characteristics.
  - Range: The emissions span a range from a low of 96.00 g/km to a high of 522.00 g/km, highlighting the breadth of vehicle types and performances encapsulated in the dataset.

	Value
count	7385.000000
mean	250.584699
std	58.512679
min	96.000000
25%	208.000000
50%	246.000000
75%	288.000000
max	522.000000

4.2 Independent Variables:

6.2.1 Engine Size:

Engine size, often measured in liters, plays a pivotal role in determining a vehicle's performance, fuel consumption, and, consequently, its CO2 emissions. In this section, we delve into the analysis of the Engine Size(L) variable, shedding light on its distribution, relationship with CO2 emissions, and general characteristics.

Scatter Plot Analysis:

The scatter plot accentuates the positive correlation between engine size and CO2 emissions. Larger engines, while offering enhanced performance, tend to consume more fuel, leading to elevated emissions. The upward trend in the plot substantiates this observation.

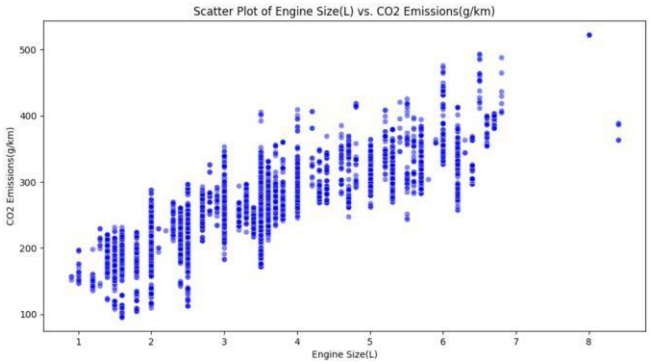

Fig 5: Scatter Plot

Histogram Analysis:

The histogram of engine sizes unveils a concentration between 1.5 and 5 liters, indicating the prevalence of mid-sized engines in the dataset. However, the right-skewed distribution also highlights the existence of vehicles with notably larger engines, potentially luxury or performance vehicles.

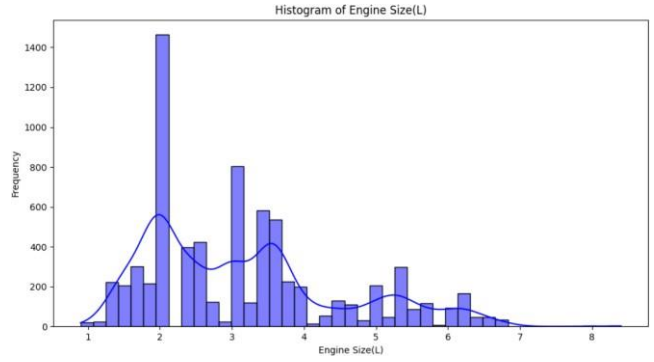

Fig6: Histogram Plots

Box Plot Analysis:

Through the box plot, we observe that the median engine size hovers around the 3-liter mark. Outliers, especially on the larger end, signify vehicles equipped with exceptionally large engines, providing a testament to the diversity of vehicles in the dataset.

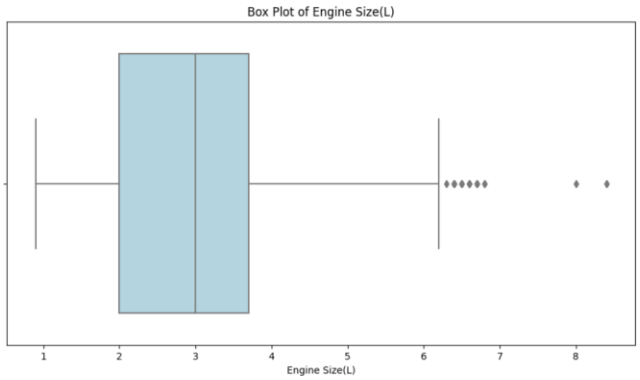

Fig 7: Box Plots

QQ Plot Analysis:

The QQ plot's deviation from a straight line, particularly at the tails, underscores the non-normal distribution of engine sizes. This deviation is consistent with our earlier observations from the histogram.

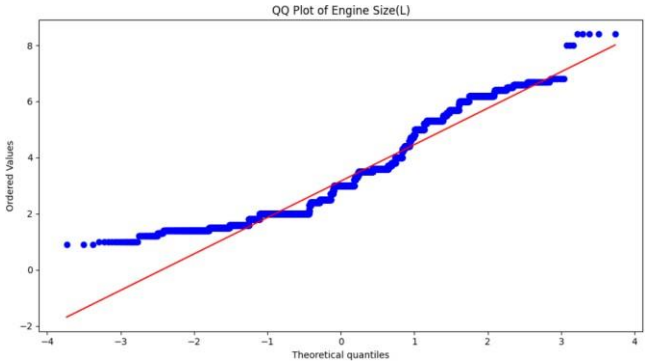

Fig 8: QQ Plots

Parameter Estimation Analysis:

Proportion:

- Small Engine: 13.45% of vehicles have engines ≤ 2 liters.
- Medium Engine: 63.63% of vehicles have engines greater than 2 liters but ≤4 liters.
- Large Engine: 22.92% of vehicles have engines > 4 liters.

Mean: The average engine size across the dataset is approximately 3.16 liters**.**

Standard Deviation: The standard deviation of 1.35 liters indicates a moderate variability in engine sizes around the mean.

```
Engine Size(L)
Small Engine     0.134462
Medium Engine    0.636290
Large Engine     0.229248
Name: proportion, dtype: float64
the mean of the engine size is, 3.160067704807041
the standard deviation of the engine size is, 1.3541704555622656
```

6.2.2 Cylinders:

The number of cylinders in an engine is a pivotal factor influencing its power, capacity, and fuel consumption. Consequentially, it plays a significant role in determining the CO2 emissions of a vehicle. In this segment, we thoroughly analyze the Cylinders variable, exploring its nuances, distribution, and its correlation with CO2 emissions.

Scatter Plot Analysis:

The scatter plot reveals a pronounced upward trend between the number of cylinders and CO2 emissions. At the lower end, vehicles with fewer cylinders (typically compact or economy cars) have lower emissions. As we move to vehicles with more cylinders, which are often associated with performance, luxury, or heavy-duty applications, the emissions noticeably increase. This trend underscores the direct relationship between engine power (often associated with more cylinders) and fuel consumption.

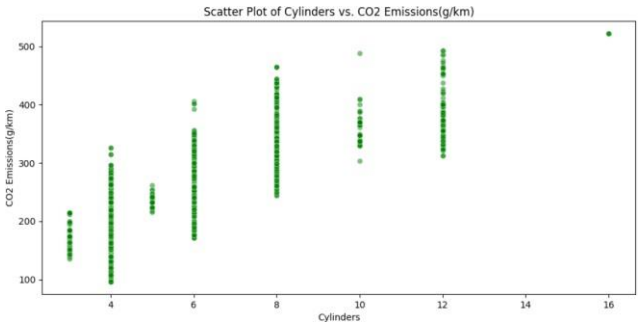

Fig 9: Scatter Plot

Histogram Analysis:

Through the histogram, we observe that the majority of vehicles in our dataset possess 4, 6, or 8 cylinders. These configurations are commonly found in most passenger cars and SUVs. However, the presence of vehicles with configurations like 3, 5, 10, 12, and even 16 cylinders indicates the dataset's diversity, encapsulating everything from fuel-efficient compact cars to high-performance and luxury vehicles.

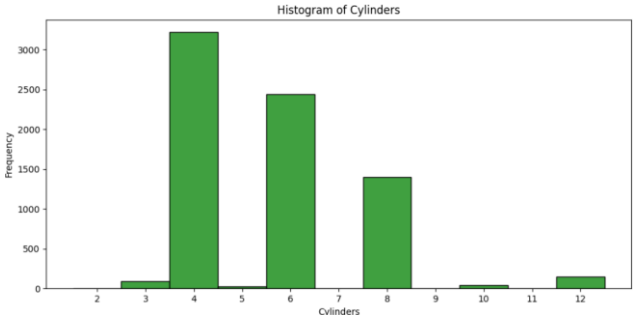

Fig10: Histogram Plots

Box Plot Analysis:

The box plot offers insights into the central tendency and spread of the number of cylinders in vehicles. The median, represented by the line inside the box, stands at 6 cylinders. This suggests that half of the vehicles in our dataset have 6 or fewer cylinders, while the other half have more than 6. Additionally, the presence of dots beyond the "whiskers" of the plot highlights outliers. Specifically, vehicles with 3, 5, 10, 12, and 16 cylinders can be considered outliers, as they deviate significantly from the typical 4-, 6-, or 8-cylinder configurations. These outliers are a testament to the diverse range of vehicles in the dataset.

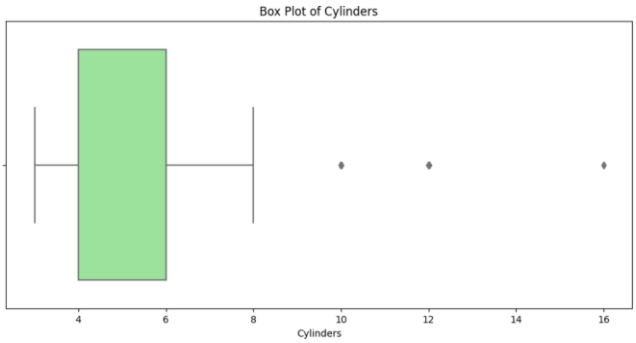

Fig 11: Box Plots

QQ Plot Analysis:

The QQ plot is designed to assess the normality of a data distribution. Given the discrete nature of the Cylinders variable and its limited range, a perfect normal distribution isn't expected. The observed deviations, validate this expectation.

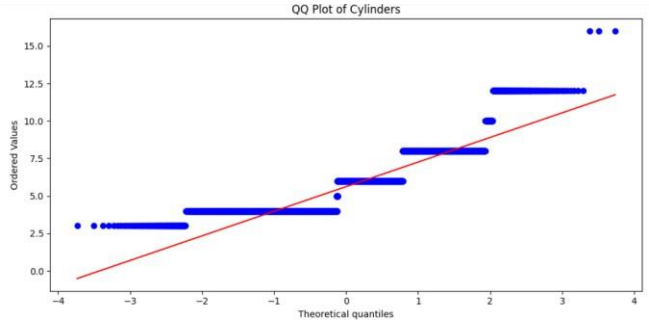

Fig 12: QQ Plots

Parameter Estimation Analysis:

- **Mean**: The average number of cylinders across all vehicles in the dataset is approximately 5.62
- **Standard Deviation:** The standard deviation is 1.83, suggesting a moderate variability in the number of cylinders among vehicles.
- **Proportion**: Among the vehicles in our dataset, approximately 1.29% are equipped with a small engine configuration of 4 cylinders or fewer. Vehicles with a medium engine configuration, having between 5 and 6 cylinders, constitute a significant portion, accounting for roughly 43.95% of the dataset. Finally, vehicles with a large engine configuration of more than 6 cylinders form the majority, comprising about 54.76% of the data.

```
Cylinders
3     0.012864
4     0.436019
5     0.003521
6     0.331212
8     0.189844
10    0.005687
12    0.020447
16    0.000406
Name: proportion, dtype: float64
The mean of the number of cylinders is, 5.615030467163169
The standard deviation of the number of cylinders is, 1.8283065156997242
```

6.2.3 Fuel Consumption Comb (L/100 km):

Fuel consumption, particularly combined fuel consumption encompassing both city and highway driving, is a critical determinant of a vehicle's environmental impact. The more fuel a vehicle consumes per 100 kilometers, the higher its CO2 emissions are likely to be. This section delves into the analysis of the Fuel Consumption Comb (L/100 km) variable, highlighting its distribution, relationship with CO2 emissions, and inherent characteristics.

Scatter Plot Analysis:

The scatter plot vividly demonstrates the strong positive correlation between combined fuel consumption and CO2 emissions. As the fuel consumption rate increases, so do the CO2 emissions. This trend is in line with our understanding that vehicles consuming more fuel tend to emit more CO2. The plot underscores the environmental advantages of fuel-efficient vehicles, as they emit comparatively lesser CO2 for every 100 kilometers driven.

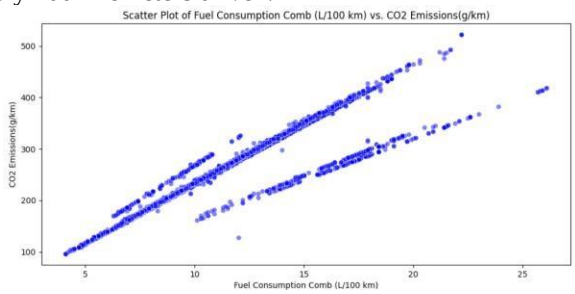

Fig 13: Scatter Plot

Histogram Analysis:

The histogram showcases the distribution of combined fuel consumption rates. The bulk of the vehicles in our dataset exhibit a consumption rate between 5 and 15 L/100 km. The peak concentration around 8 to 9 L/100 km indicates the popularity of vehicles with this fuel consumption rate, likely representing a balance between performance and efficiency.

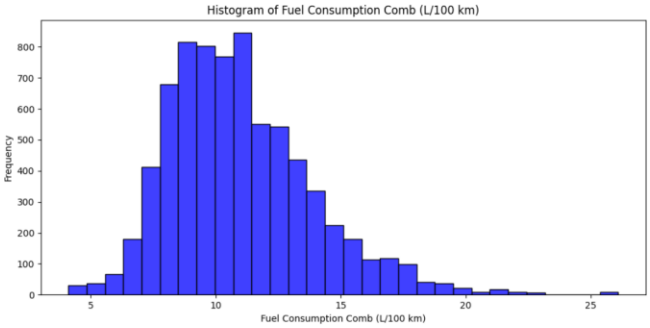

Fig14: Histogram Plots

Box Plot Analysis:

The box plot provides a succinct view of the Fuel Consumption Comb (L/100 km) distribution. The median, indicated by the line within the box, is slightly above 10 L/100 km, revealing that half the vehicles in our dataset have a consumption rate of 10 L/100 km or lower, while the other half consume more. The plot also highlights outliers—vehicles that have notably high fuel consumption rates, diverging significantly from the typical range. These outliers may represent high-performance or luxury vehicles.

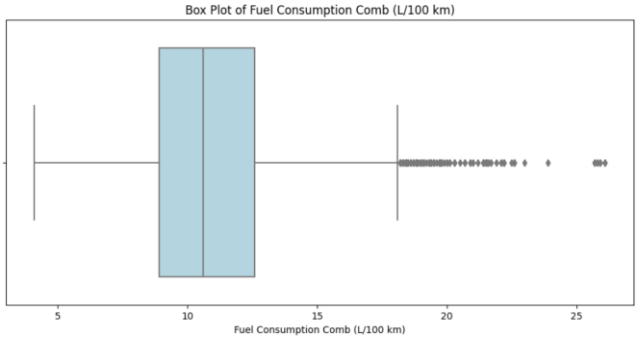

Fig 15: Box Plots

QQ Plot Analysis:

The QQ plot, designed to assess data normality, suggests that the fuel consumption distribution isn't perfectly normal.

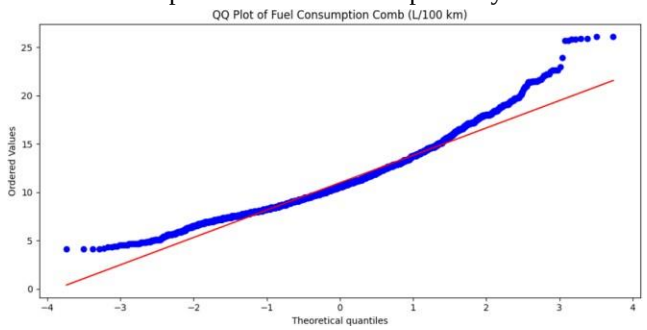

Fig 16: QQ Plots

Parameter Estimation Analysis:

Proportion:

- Low Consumption (≤8 L/100 km): A total of 12.02% of the vehicles in the dataset exhibit low fuel consumption, making them relatively more environmentally friendly.
- Medium Consumption (8-12 L/100 km): The majority of vehicles, 56.66%, have medium fuel consumption rates. This range likely represents a balance between performance, luxury, and fuel efficiency.
- High Consumption (>12 L/100 km): Vehicles with high fuel consumption rates, approximately 31.32%, are possibly performance-oriented, larger, or luxury vehicles that prioritize attributes other than fuel efficiency.

Mean: On average, vehicles in our dataset have a combined fuel consumption rate of approximately 10.98 L/100 km.

Standard Deviation: A standard deviation of 2.89 L/100 km reflects the variability in fuel consumption rates among the vehicles. This variability emphasizes the broad spectrum of vehicles present in our dataset, from highly efficient models to those designed for performance or luxury.

```
Fuel Consumption Comb (L/100 km)
Low Consumption (≤8 L/100 km)         0.120244
Medium Consumption (8-12 L/100 km)    0.566554
High Consumption (>12 L/100 km)       0.313202
Name: proportion, dtype: float64
The mean combined fuel consumption is, 10.975071090047393 L/100 km.
The standard deviation of the combined fuel consumption is, 2.8925063028984717 L/100 km.
```

6.2.4 Fuel Consumption Comb mpg:

Fuel efficiency, measured in miles per gallon (mpg), is an integral aspect of a vehicle's design and operation. A higher mpg value typically signifies better fuel efficiency, implying the vehicle can travel more miles for every gallon of fuel consumed. Consequently, fuel-efficient vehicles tend to emit less CO2. This section focuses on analyzing the Fuel Consumption Comb (mpg) variable, providing insights into its distribution, relationship with CO2 emissions, and general characteristics.

Scatter Plot Analysis:

The scatter plot illustrates the inverse correlation between combined fuel efficiency (in mpg) and CO2 emissions. Vehicles with higher fuel efficiency, signifying more miles per gallon, exhibit lower CO2 emissions. This trend reinforces the environmental benefits of fuel-efficient vehicles, emphasizing their role in reducing the overall carbon footprint.

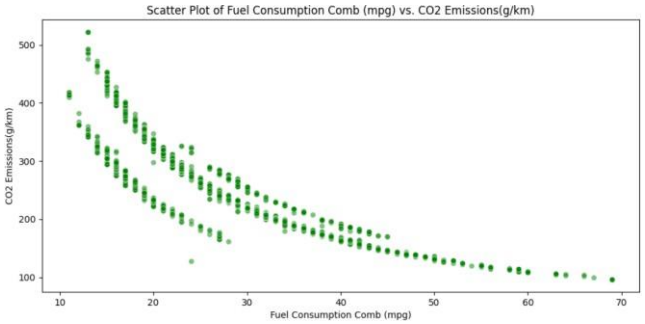

Fig 17: Scatter Plot

Histogram Analysis:

The histogram offers a snapshot of the distribution of combined fuel efficiency. A substantial segment of the vehicles in our dataset has a fuel efficiency ranging between 15 and 35 mpg, with a pronounced peak around 25 mpg. This range likely encapsulates a broad spectrum of vehicles, from compact cars to medium-sized SUVs.

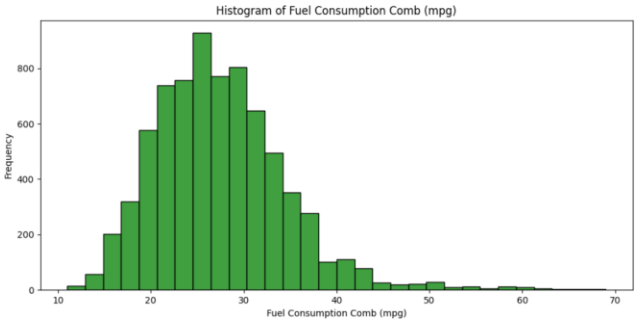

Fig18: Histogram Plots

Box Plot Analysis:

Through the box plot, we discern that the median fuel efficiency lies slightly below 25 mpg. Furthermore, outliers on the higher end of the mpg spectrum highlight vehicles with superior fuel efficiency. These outliers could encompass hybrid vehicles, smaller cars, or models specifically designed for optimized fuel consumption.

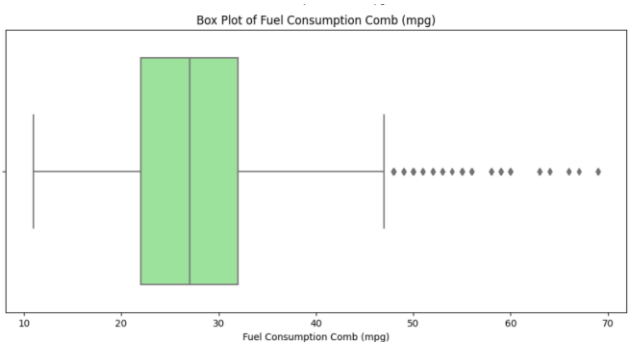

Fig 19: Box Plots

QQ Plot Analysis:

Employing the QQ plot to gauge data normality, we notice minor deviations from the straight line, especially at the top. This observation suggests that the fuel efficiency distribution, while approximating normality, isn't perfectly normal.

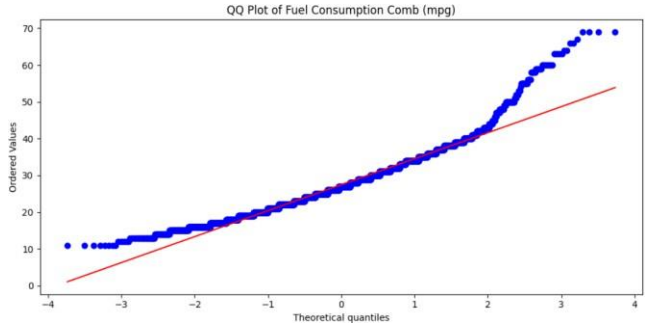

Fig 20: QQ Plots

Parameter Estimation Analysis:

Proportion:

- Low Efficiency (≤20 mpg): A segment of vehicles, approximately 11.59%, are categorized under low fuel efficiency. These could be larger vehicles, performance-oriented cars, or models that prioritize other features over fuel efficiency.

- Medium Efficiency (20-30 mpg): Forming the majority, 53.84%, this range likely represents a balance between performance, size, and fuel efficiency, encompassing a broad spectrum of vehicles from compact cars to medium-sized SUVs.

- High Efficiency (>30 mpg): Vehicles with superior fuel efficiency, accounting for about 34.57 percent , could include hybrid vehicles, smaller cars, or models specifically designed for optimized fuel consumption.

Mean:

Vehicles in our dataset, on average, exhibit a combined fuel efficiency of approximately 27.48 mpg.

Standard Deviation:

A standard deviation of 7.23 mpg conveys the variability in fuel efficiency among the vehicles. This figure underscores the diverse nature of our dataset, capturing vehicles with varying designs, purposes, and fuel efficiencies.

```
Fuel Consumption Comb (mpg)
Low Efficiency (≤20 mpg)        0.115911
Medium Efficiency (20-30 mpg)   0.538389
High Efficiency (>30 mpg)       0.345701
Name: proportion, dtype: float64
The mean combined fuel efficiency is, 27.48165199729181 mpg.
The standard deviation of the combined fuel efficiency is, 7.231879172141879 mpg.
```

Conclusion:

Our study of the Canadian vehicle dataset highlighted the relationship between car features, such as engine size and number of cylinders, and their CO2 emissions. We found that larger engines and more cylinders typically result in higher emissions, while vehicles with better fuel efficiency tend to emit less CO2. These insights emphasize the importance of choosing and promoting fuel-efficient vehicles for a sustainable future. As we continue our project, we aim to delve deeper, exploring other influencing factors and predicting emissions through advanced modeling. This knowledge can guide car manufacturers, policymakers, and consumers towards more environmentally-conscious decisions.

References:

[1] Canadian Government. (n.d.). CO2 emissions by vehicles. Open Government Portal. Retrieved from https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6.

[2] Podder, D. (n.d.). CO2 emission by vehicles. Kaggle. Retrieved from: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles.