Exploring the Relationship Between Car Characteristics and Environmental Factors

by conducting a Statistical Analysis

Author: **Koutsourelis Ioannis** *

Supervisor: Tsagris M.

Paper on course: Introduction to programming using R

University of Crete, Department of Economic Sciences, Rethymnon

Winter Semester, 2024-2025

* Post-Graduate Student
General Register Number: 6152

Contents

Abstract

This study analyzes the relationship between CO2 emissions and key vehicle attributes using a dataset of car specifications. The analysis focuses on understanding how factors such as engine size, fuel consumption, horsepower, and fuel type influence CO2 emissions. The study employs descriptive statistics to summarize the central tendencies and variability of these variables, followed by correlation analysis to identify significant relationships. Outliers and extreme values are systematically addressed to ensure the reliability of results. Visualizations, including histograms and scatter plots, provide an intuitive understanding of data distributions and relationships. Regression analysis further explores the predictive power of independent variables on CO2 emissions, offering insights into the environmental impact of vehicle design and fuel efficiency. Results reveal that engine size and fuel consumption are the most significant predictors of CO2 emissions, with hybrid and plug-in hybrid vehicles demonstrating substantially lower emission levels compared to traditional fuel types like gasoline and diesel. The final regression model achieved an Adjusted R-squared value of 0.9768, indicating robust explanatory power.The findings contribute to discussions on sustainable automotive practices and policies aimed at reducing emissions in the transportation sector.Policy implications include the promotion of hybrid technologies and fuel-efficient vehicles as effective strategies for reducing overall emissions.

1 Introduction

1.1 Background

In recent years, concerns about environmental sustainability have heightened the focus on vehicle emissions, particularly **carbon dioxide (CO2) emissions**, due to their significant contribution to climate change. The automotive industry plays a pivotal role in shaping global emissions, making it essential to analyze the factors influencing CO2 outputs.

According to the European Union, road transport accounts for approximately **21% of total CO2 emissions** in the region, leading to the implementation of stringent CO2 emission performance standards. Regulation (EU) 2019/631 outlines progressively stricter targets for average CO2 emissions from new passenger cars, aiming to achieve **net-zero emissions by 2035** (European Commission). These measures have spurred technological advancements and market shifts toward low-emission and zero-emission vehicles.

Attributes such as engine size, fuel consumption, horsepower, and fuel type are critical in determining vehicle emissions. Larger engines and higher fuel consumption rates typically result in increased CO2 emissions (EPA). Additionally, the type of fuel significantly influences emissions, with diesel vehicles emitting about 13% more CO2 per liter of fuel burned compared to gasoline vehicles (The ICCT).

By exploring the relationships between these vehicle attributes and CO2 emissions, this analysis provides actionable insights. These insights are relevant not only for consumers making informed choices but also for manufacturers strategizing around vehicle design and policymakers implementing effective environmental regulations. The study's findings align with global sustainability goals, emphasizing the importance of fuel-efficient and hybrid technologies in reducing emissions.

1.2 Objectives

The primary objective of this analysis is to **investigate the relationships** between CO2 emissions and various vehicle attributes to understand the influence of key factors such as engine size, fuel consumption, and fuel type on emissions. It aims to **identify patterns and trends** within the dataset

that may indicate opportunities for improving vehicle efficiency and reducing environmental impacts.

1.3 Overview of Methods

This study employs a systematic approach to analyze the dataset. **Descriptive statistics** are used to summarize central tendencies, variability, and distributions of key variables, providing an initial understanding of the data. **Correlation analysis** identifies significant relationships between CO2 emissions and independent variables, including both continuous and categorical attributes. **Regression analysis** develops a predictive model for CO2 emissions, highlighting the impact of each variable and their combined explanatory power. **Visualizations** are integrated throughout the analysis to enhance the interpretability of results and support data-driven conclusions. A lot of importance was given to the styling of the plots, by integrating unique color themes and aesthetic backgrounds, aiming to enhance readability and add visual appeal.

2 Data and Methodology

2.1 Dataset Description

The data that we are analyzing in the current paper were extracted from the Greek site **Autotriti.com** and it is about new cars in the market of Greece on 2020. The starting dataset entries a total of 1505 cars and the data spans over a number of 14 variables of car characteristics and metrics.

Continuous Variables	
Variable	**Justification**
Price	Vehicle price
Engine	Engine size in cubic centimeters
Horsepower	Vehicle power output in hp
Power_Nm	Torque, measured in Newton-meters
Acceleration	Acceleration time (0-100 km/h) in seconds
Consumption	Fuel consumption in liters per 100 km
CO2	CO2 emissions measured in grams per km
Autonomy_klm	Estimated range on a full tank or battery in kilometers
Taxation	Tax cost per year in euros

Table 1: Continuous Variables

Table 1 represents the continuous variables

Categorical Variables		
Variable	**Subtype**	**Justification**
Fuel	Nominal	Fuel type (Gasoline/Diesel/ Mild-Hybird/ Diesel Mild-Hybird/ Hybrid-Plug-in/ LPG)
Air_condition	Nominal	Indicates if air conditioning is present ("Yes/No")
Clima	Nominal	Indicates if climate control is present ("Yes/No")
Back_electric_windows	Nominal	Indicates if rear electric windows are present ("Yes/No")
Heated_mirrors	Nominal	Indicates if heated mirrors are present ("Yes/No")

Table 2: Categorical Variables

The categorical variables of the dataset are presented in **Table 2**

2.2 Preprocessing (Cleaning)

Pre – cleaning of the data was needed before we proceeded with the analysis as missing values(NA), in addition to other issues,were occurred. The cleaning procedure that took place is presented in the this section.

After a first glance at the collected data several problems were identified. Firstly, several variables showed **missing values** (NAs) and also empty values, which were immediately removed. Secondly, a number of categorical variables (such as) which had a value different than that of YES/NO were also removed from the sample. After all the excessive cleaning of the data we were left with a total of 1.191 car observations. Note that the "Electric" class from the 'Fuel' variable was entirely removed as a result of severe data loss. Also, the rows representing the top 5% in the Price variable were removed, as a part of the **outliers removal process**, in order to enhance the reliability and interpretability of the analysis. Extreme values(outliers) can distort statistical measures such as the mean, standard deviation, and regression coefficients. In addition, the top 5% of prices often represent luxury cars that do not reflect general consumer trends. Excluding these values allows the analysis to concentrate on the typical price range, yielding insights that are more applicable to the average market participant.

By removing the top 5%, we are aiming to focus on the **"core"** data, reducing the influence of outliers and keeping the "true" relationships between the variables, thus ensuring more accurate results. This percentile-based trimming is widely recognized in data analysis as an effective way to reduce the influence of outliers while preserving the dataset's essential characteristics. Finally, 11 more outlying observations were detected that were far away from the general trend and thus, had to be removed. Some of them were very close but lower than the top 5% and others towards the bottom 5%, so for both instances we had to specifically select them. After all the cleansing process that we conducted, the **final dataset** consisted of **1.120 observations** and **14 variables**.

2.3 Statistical Methods

In this analysis, **Descriptive Statistics** were used to summarize and understand the data. **T-tests** were applied to compare the means of different groups, such as fuel types, to see if they differ significantly in fuel consumption or emissions. **F-tests** were used to compare different regression models, so as to check if adding more variables improves the model. In addition, **Correlation Analysis** examined the relationships between continuous variables, like Engine and CO emissions, to see how they are linked. Finally, **Regression Analysis** was applied to model how vehicle characteristics (like engine size and fuel type) affect CO emissions, helping to predict future outcomes.

These methods were chosen to understand the data, test relationships between variables, and model the effects of different factors on CO emissions.

3 Descriptive Statistics

Descriptive statistics are the first step for any statistical analysis. That happens because of two reasons: 1) to provide basic information about variables in a dataset and 2) to highlight potential relationships between variables. Thus, they play a vital role in understanding the nature of our data and it's variables .

In the subsections following we will go through some basic descriptive statistics, some frequency tables for our variables, and comment the findings along.

3.1 Frequencies

Variable	Air_con..	
Value	Yes	No
Frequencies	293	827
Relative Frequencies (%)	26.16	73.84
Variable	Clima	
Value	Yes	No
Frequencies	827	293
Relative Frequencies (%)	73.84	26.16
Variable	Back_ele.._win..	
Value	Yes	No
Frequencies	963	157
Relative Frequencies (%)	14.02	85.98
Variable	Heated_mir..	
Value	Yes	No
Frequencies	1050	70
Relative Frequencies (%)	93.75	6.25

Table 3: Frequency table for Binary Categorical variables

Regarding **Table 3**, the variable with the biggest difference between 'Yes' and 'No' is Heated_mirrors_ as the frequency percentage for each option are 93.75 and 6.25, respectively. Also, it's important to note the substitude nature of Air_condition and Clima, as we can see exactly opposite frequency distributions, meaning that a car that has the Air condition can't have Climatism and vise versa

Variable	**Fuel**						
Class	**Gasoline**	**Diesel**	**LPG**	**Hybrid**	**Mild-Hybrid**	**Hybrid Plug-in**	**Diesel-Mild-Hybrid**
Frequencies	**648**	**369**	**3**	**36**	**40**	**20**	**4**
Relative Frequencies (%)	**57.86**	**32.95**	**0.27**	**3.21**	**3.57**	**1.79**	**0.36**

Table 4: Frequency distribution of Fuel

As for **Table 4**, the two most dominant fuel types are Gasoline and Diesel, as expected, taking up 90.81% of the fuel categories. This can be better represented visually in **Figure 5**, which we will go through later in **Section 4** .

3.2 Basic Statistics

	Price	Engine	Horsepower	Power Nm	Acceleration
Mean	30752	1566	144.5	254.4	9.942
Median	26695	1499	130	250	10
Maximum	91550	2998	360	700	17.1
Minimum	9100	898	60	88	4.5
1st Quarter	19297	1199	112	200	8.575
3rd Quarter	37005	1969	165	306.2	11.1
Standard Error	473.51	12.55	1.63	2.86	0.06
Median Absolute Deviation	13091.36	444.78	38.55	77.84	1.78
Range	82450	2100	300	612	12.6
Standard Deviation	15846.69	420.08	54.66	95.69	2.08
Kurtosis	1.29	0.85	2	0.53	0.24
Skewness	1.24	0.72	1.36	0.67	0.14

Table 5: Descriptive statistics for Continuous variables(Part 1)

The main findings from **Table 5** consist of the following. Regarding central Tendency: the Price mean (€30.752) is higher than the median (€26.695), suggesting a positive skew which is possibly influenced by expensive vehicles. Engine (Mean = 1.566 cc, Median = 1.499 cc) shows a slight positive skew, indicating more vehicles with smaller engines. In addition Horsepower (Mean = 144.5 HP, Median = 130 HP) indicates a central range of moderate horsepower vehicles. As for Spread, Price's high range (€82450) and large standard deviation (€15846.69) indicate significant price variability. Also Horsepower's range of 300 HP with a standard deviation of 54.66 HP suggests the existence of some extreme values. For Skewness and Kurtosis, Horsepower indicates a Skewness of 1.36, showing a stronger positive skew. Kurtosis = 2, suggests sharper peaks than normal. On the

other hand, Acceleration (Skewness = 0.14 and Kurtosis = 0.24) shows the most symmetric and flat distribution among variables.

	Consumption	CO2	Autonomy klm	Taxation
Mean	5.141	122.1	1056	162.3
Median	5	119	1000	117
Maximum	10.2	237	3571	723
Minimum	1.3	31	541	0
1st Quarter	4.3	105	854	103
3rd Quarter	5.8	136	1163	163
Standard Error	0.04	0.82	9.69	3.57
Median Absolute Deviation	1.04	20.76	225.36	45.96
Range	8.9	206	3030	723
Standard Deviation	1.30	28.61	324.18	126.55
Kurtosis	1.58	2.67	21	4.92
Skewness	0.60	0.65	3.57	2.03

Table 6: Descriptive statistics for Continuous variables(Part 2)

From **Table 6** we found some notable findings. Some positive observations were that, Consumption (Mean = 5.141 L/100 km, Median = 5 L/100 km) indicates that most vehicles are fuel-efficient. In addition CO2 (Mean = 122.1 g/km, Median = 119 g/km) shows that most vehicles are aligned with moderate emissions. Consumption's low range (8.9 L/100 km) and small standard deviation (1.3 L/100 km) suggest consistent fuel efficiency across the dataset. As for the skewness and kurtosis, Autonomy (Skewness = 3.57 and kurtosis = 21) indicate extreme positive skew and a highly peaked distribution, with some outliers. Similarly, Taxation (Skewness = 2.03 and kurtosis = 4.92) indicate high positive skewness and a sharper peak due to high-tax outliers.

3.3 Overall Observations

Variables like Acceleration, Power, and Consumption have distributions closer to symmetry, making them less affected by extreme values. Price and Autonomy have high variability, reflecting a diverse range of vehicles in terms of cost and range. The skewness in CO2 and Taxation reflects the impact of regulations or market choices on vehicle sustainability and affordability.

4 Data Visualization

This section presents visual explorations of the dataset to understand patterns, relationships, and distributions. To understand the distribution of CO2 emissions and other key variables, histograms and box-plots were used. These visualizations help identify patterns, skewness, and potential outliers in the data. These visualizations provide a foundation for identifying outliers, understanding data spread, and preparing for more advanced analyses, such as correlations or regression.

4.1 Uni-variate Analysis

In this section we will analyze data based on only one variable, with purpose to describe, summarize and find possible patterns.

4.1.1 Histograms of Key Continuous Variables

This section explores the distribution of key continuous variables in the dataset. These plots will show whether the variables are normally distributed or skewed, which is important for selecting the right models.

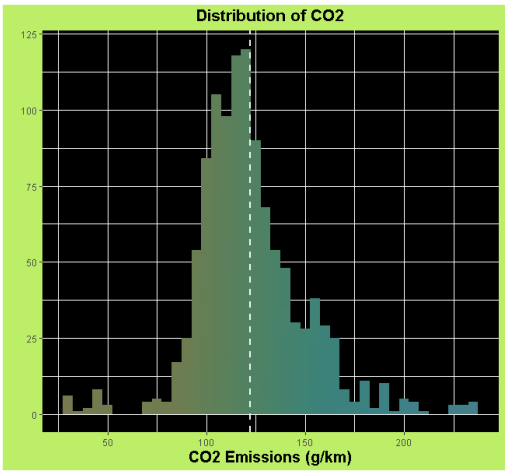

Figure 1: Distribution of CO2

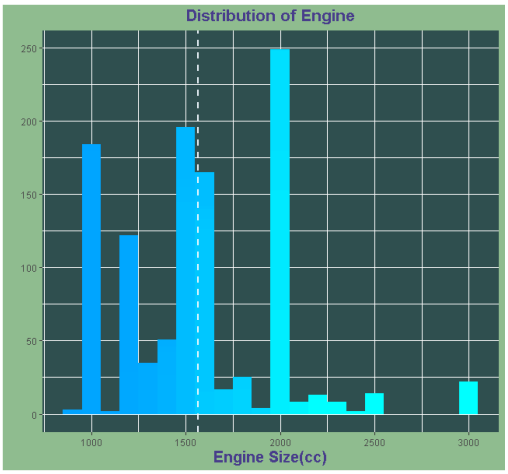

Figure 2: Distribution of Engine

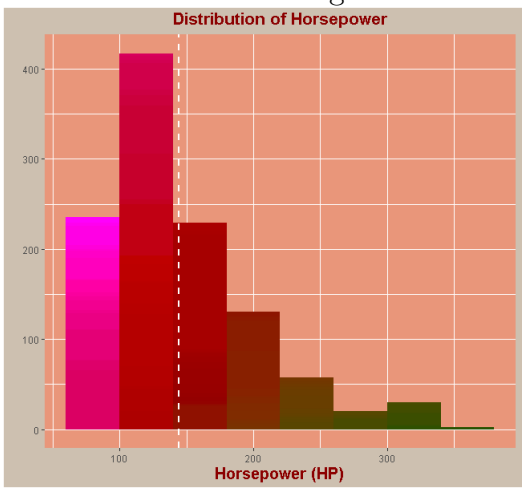

Figure 3: Distribution of Horsepower

The histograms above represent the distribution of the main variables CO2, Engine and Horsepower. The white vertical dashed line indicates the mean of each variable.

For **Figure 1**, the data appears to be slightly right-skewed, indicating that most vehicles have lower emissions, with fewer vehicles emitting higher levels of CO2. This visualization highlights the variability in CO2 emissions and may suggest the need for targeted policies or technologies to reduce emissions among high-emission vehicles. Regarding **Figure 2**, there seems

to be a concentration of vehicles with engine sizes around 1,000–1,600 cc, with fewer vehicles having larger engines. This figure suggests that most vehicles in the dataset fall within the compact to midsize engine range, possibly reflecting consumer preferences or market trends. Finally, the Histogram of Horsepower (**Figure 3**) indicates the spread of vehicle power ratings. The distribution shows that the biggest percentage of vehicles fall within a moderate horsepower range, with a small number of vehicles exhibiting very high horsepower.

4.1.2 Charts of Key Categorical Variables

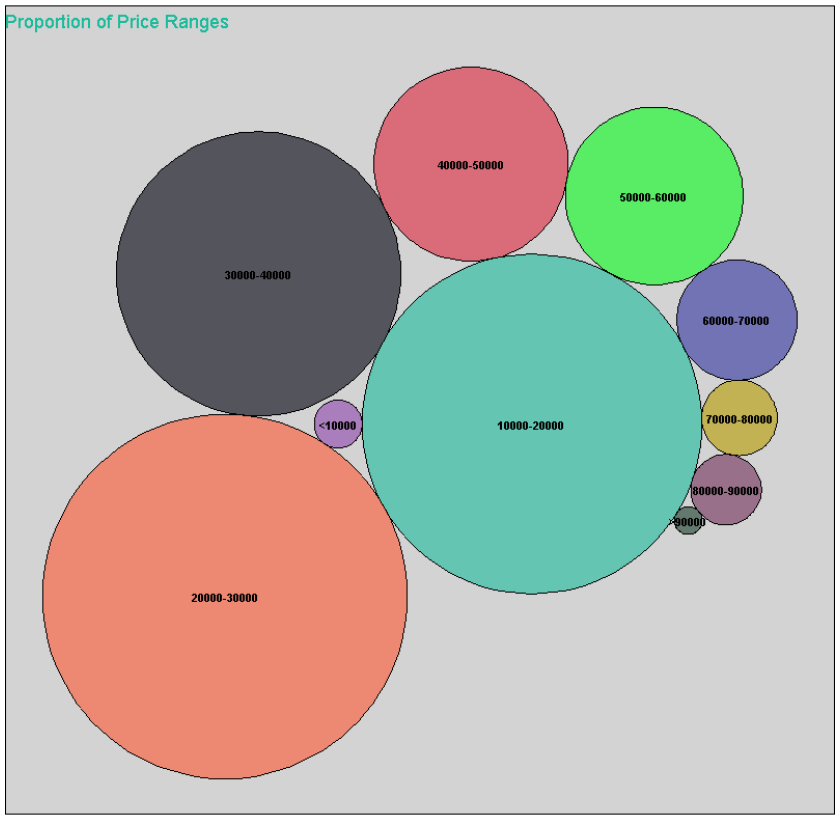

Figure 4: Circle Packing plot for ranges of Price

Figure 4 visualizes the distribution of vehicles across different price ranges using circle packing. Each circle represents a price range, with its

size proportional to the count of vehicles within that range. The main observation here is that the majority of vehicles are concentrated in mid-level price ranges (€10000 - €40000), suggesting a strong market preference for moderately priced vehicles.

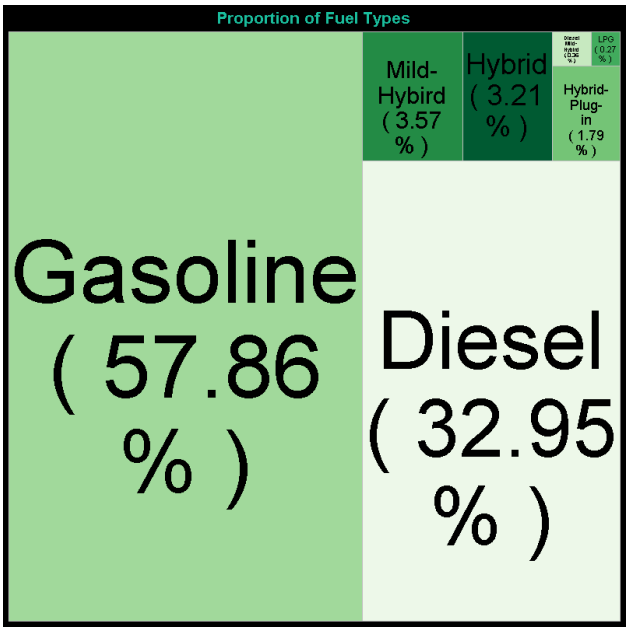

Figure 5: Tree-map of Fuel

The tree-map in **Figure 5** represents the proportions of vehicles in the dataset by their fuel types. Each rectangle corresponds to a fuel type, with its area proportional to the count of vehicles using that type. Traditional fuel types (e.g. Gasoline, Diesel) continue to dominate the market, reflecting consumer familiarity and existing infrastructure. Alternative fuels (e.g., Electric, Hybrid) are growing but remain less prevalent, signaling a gradual shift toward sustainability. These findings can be confirmed from **Table 4**, which was used for the creation of the tree-map

4.2 Bi-variate Analysis

In the following section we will go through the plots used for the quantitive statistical analysis, in order to explore relationships between variables

4.2.1 Categorical vs. Categorical

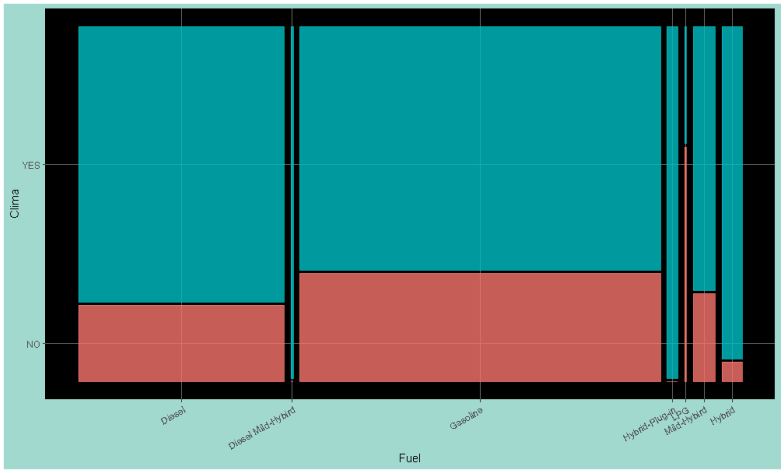

Figure 6: Mosaic Plot for Clima by Fuel categories

This Mosaic Plot in **Figure 6** helps us understand the relationship between Clima and Fuel. Almost across all fuel types, Climatism in a car is a common characteristic. Clima was used indicatively for all the other categorical variables, considering that the same mechanics that exist in this plot, expand across the other variables .

4.2.2 Continuous vs. Continuous

Both **Figure 7** and **9** below, have color coded data points by Fuel type, providing additional insights into how fuel categories impact this relationship.

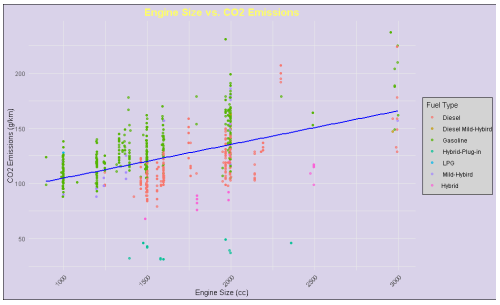

Figure 7: Scatter Plot of Engine and CO2 by Fuel types

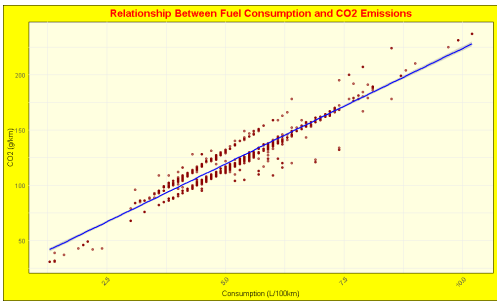

Figure 8: Scatter Plot of Consumption and CO2

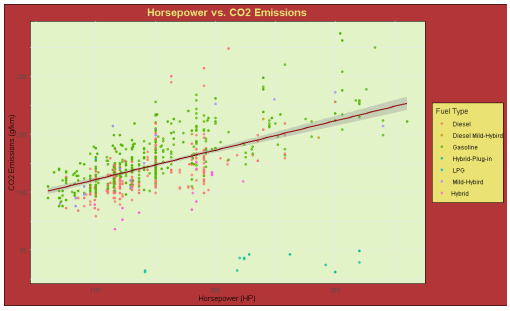

Figure 9: Scatter Plot of Horsepower and CO2 by Fuel types

Main observation from **Figure 7**, is the positive correlation between CO2 and Engine size. Larger engines are generally associated with higher CO2 emissions, a trend visible across most fuel types. Nevertheless, the distinction between Gasoline or Diesel and Electric or Hybrid vehicles is clear, with the first dominating the higher CO2 emissions range, especially for larger engines, and the latter clustering at lower CO2 levels, reflecting their environmental benefits.

By observing **Figure 8**, the strong positive correlation is visible. As fuel consumption increases, CO2 emissions also rise, which aligns with expected physical principles. Most vehicles cluster at moderate levels of fuel consumption and CO2 emissions, while vehicles with exceptionally high fuel

consumption and emissions are likely high-performance or heavy-duty models. On the other hand the low-consumption outliers represent electric or hybrid vehicles, which minimize fuel usage and emissions. This relationship underscores the need for fuel-efficient technologies to reduce emissions.

Finally **Figure 9** shows us that, in general, higher horsepower is linked to increased CO2 emissions, as more powerful engines typically consume more fuel. Gasoline and diesel vehicles show a wide spread, with higher horsepower vehicles emitting significantly more CO2. Contrariwise, electric or hybrid vehicles are clustered at lower horsepower and emissions levels, reflecting a focus on efficiency over performance. The same inference as before can be made for the outliers, as they correspond to a few high-performance vehicles with exceptionally high horsepower and emissions, likely representing luxury or sports cars. This plot highlights the inevitable trade-off between vehicle performance and environmental impact.

In conclusion , these figures highlight the environmental impact of engine size, horsepower and fuel choice, which is crucial for regulatory frameworks and consumer decision-making. These results are relevant for automotive makers, engineers and policymakers, aiming to balance consumer demand for powerful high-performing cars, with sustainability goals. In addition, with the need for environmental sustainability and growth in mind, policymakers and automakers can use this data to design strategies that incentivize the adoption of low-emission, Eco-friendly vehicles.

4.2.3 Categorical vs. Continuous

We created these Box Plots, to visualize the distribution of CO2 emissions across different fuel types and price ranges, enabling comparisons among categories

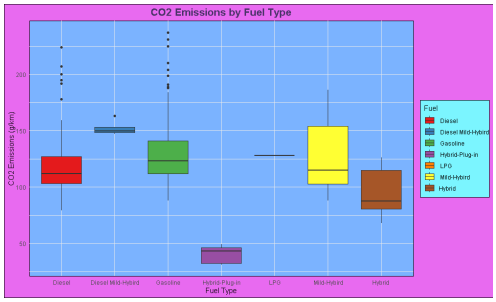

Figure 10: Box Plot of CO2 by Fuel type

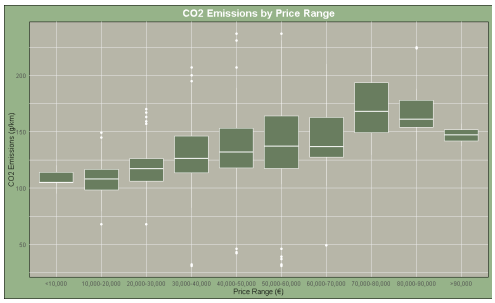

Figure 11: Box Plot of CO2 by ranges of Price

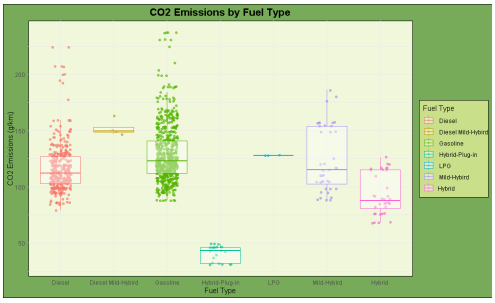

Figure 12: Box Plot with Jittered data points of CO2 by Fuel types

Gasoline and diesel vehicles in **Figure 10**, exhibit higher median CO2 emissions compared to alternative fuel types like electric and hybrid, who have notably lower medians, aligning with their environmentally friendly design. As for the data spread, gasoline and diesel vehicles show wider interquartile ranges, indicating greater variability in emissions. This contrasts electric vehicles who display minimal variation, consistent with their emission-free nature. Regarding the outliers, same conclusions can be drawn as **Subsection 4.2.2**, with several extreme values appearing for gasoline and diesel categories, suggesting high-performance models or inefficient engines.

For **Figure 11**, higher price categories generally correspond to higher CO2 emissions, likely driven by the inclusion of high-performance vehicles

in these segments. In the other hand, the lower price ranges show a relatively lower and narrower spread in CO2 emissions, dominated by economy and hybrid models. This leads to the conclusion that Price, in general, is associated with CO2 emissions. Obvious is the existence of outliers in almost every Price class, especially in the middle ranges.

While **Figure 12** consists of similar characteristics to **Figure 10**, it provides a more enhanced visualization by overlaying jittered individual data points on a box plot, allowing us to provide a more detailed view. Thus, it improves interpretability, making it easier to identify trends and anomalies. Such patterns in the spread of data consist of the distinct emission bands within each fuel category , revealing nuances and variability internally for each Fuel type, which was impossible with a simple Box Plot. For example, we can observe the existence of a pattern for the variables Hybrid-Plug-in, Mild-Hybrid and Hybrid. This pattern takes place with the existence of two general groups of observations in each variable. One that is at the higher end, and one at the lower, separated by the median line.

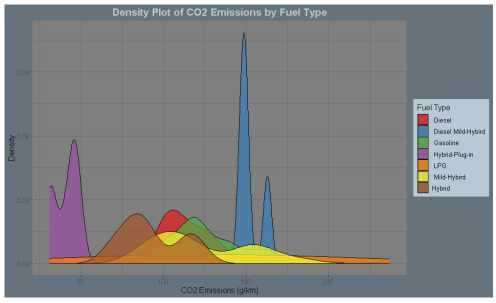

Figure 13: Density Plot of CO2 by Fuel types

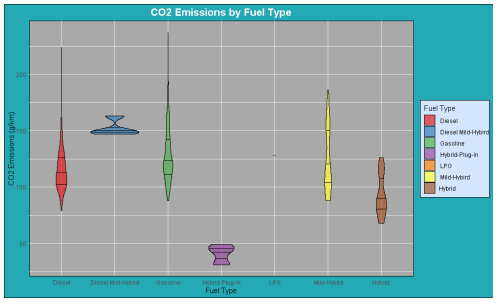

Figure 14: Violin Plot of CO2 by Fuel type

The Density Plot (**Figure 13**) represents the probability distribution of CO2 emissions across different fuel types. This is useful because we can distinct different distribution shapes across variables. For example, we can observe the existence of a Bimodal Distribution for all types except gasoline and diesel. This suggests the existence of two groups of vehicles across these variables, which could be emphasizing the need for further segmentation within those categories. Hybrids (except Diesel Mild-Hybrid), peak closer to zero, confirming their low-emission status. Specifically, Mild-Hybrid and Hybrid , show narrower distributions, clustered around lower CO2 levels,

highlighting their consistency in emission efficiency. Additionally, there is some overlap between diesel and gasoline vehicles, particularly in the mid-range of CO2 emissions ($\tilde{8}$0-150 g/km). This could indicate similar emissions for vehicles with comparable engine sizes.

Coming to the end of this section, we will analyze **Figure 14**. This plot combines box plot features with a density estimate, offering a detailed view of data distribution, variability, and central tendencies. The conclusions can be similar **Figure 13**, with Gasoline and Diesel showing wide distributions with values concentrating in the mid-range ($\tilde{8}$0-150 g/km), and hybrids and electrics inclined towards the low-range. The existence of outliers can once again be visualized, as the tails of the violins for gasoline and diesel extend significantly higher.

While the density plot (Figure 13) focuses on the overall distribution, the violin plot (Figure 14) combines this information with traditional box plot statistics, offering richer insights into variability and extremes.

5 Correlation Analysis

Understanding the relationships between variables is a crucial step in data analysis, as it provides insights into potential patterns, associations, or dependencies within the dataset. This section focuses on analyzing correlations between variables to uncover meaningful interactions. Depending on the nature of the variables (continuous or categorical) and their distribution, different statistical methods were applied to ensure robust and accurate findings.

Firstly, normality issues had to be addressed, in order to select the best representing test for each variable.

For continuous variables, correlation coefficients such as **Pearson** and **Spearman** were employed to assess the strength and direction of linear and monotonic relationships, respectively. To explore associations between categorical variables, **Chi-Square tests** were used to determine statistical significance. Relationships between continuous and categorical variables were analyzed using **t-tests** or **ANOVA**, depending on the categorical variable's levels. Note that a confidence level of a = 5% was used for the tests.

Additionally, visualizations, including **scatter-plots**, **mosaic-plots**, and **box-plots**, were utilized to complement the statistical analysis and provide a more intuitive understanding of the relationships between variables.

This comprehensive approach aims to provide a detailed examination and understanding of variable interactions.

This section provides an in-depth examination of the relationships between continuous variables in the dataset using a correlation matrix.

5.1 Normality Test

Before the correlation analysis, we wanted to proceed with a **Normality Test** so as to check if our main continuous variables follow normal distribution. After running the **Shapiro - Wilk** test we got values very small values back (ex. 2.2e-16) for all our variables. We then proceeded with some possible solutions for this including, transforming our variables using logarithm or square root but the problem persisted. After following a visual inspection so as to check the normality, using the histograms (already from before) and a Q-Q Plot, we figured out that the **CO2 variable** indeed **follows a Normal Distribution**. All the points fall approximately along the reference line, so we can assume normality. We concluded the same from examining the CO2 Histogram.

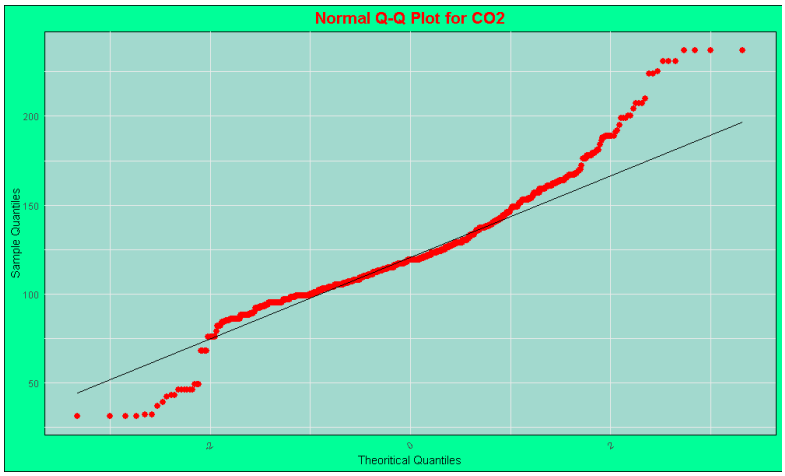

Figure 15: Q-Q Plot for CO2

After some research, the possible reason that we ended up to for this occurrence is the large sample size (1120 obs.). According to **Central Limit Theorem**, with large sample sizes, the sampling distribution of the mean

tends to be normal, even if the population distribution is not. That happens because even tiny deviations from normality will result in very small p-values. Essentially, the test is detecting that the data is not perfectly normal. While this applies to sampling distributions, it emphasizes that 1120 data points provide substantial statistical power. Considering that we are working with data mined from the site Autotriti.com, we can justify this finding for the reason that this is a common occurrence in real-world data, because most data do not follow a perfect normal distribution. In addition the **Shapiro-Wilk** test is sensitive to even small deviations from normality when the sample size is large. For smaller sample sizes (¡50), the test might not detect subtle deviations. In contrast, with a sample size of 1120, even minor deviations from a perfect normal distribution will likely result in rejecting the null hypothesis of normality. Considering the above we executed the **Kolmogorov-Smirnov test** but the problem continued, still showing that the CO2 does not follow a normal distribution. Even with non parametric tests the we had the same result. For the reason that we had to proceed with the analysis, we used the finding from the Histograms and the Q-Q Plot and assumed normal distribution for the CO2 variable.

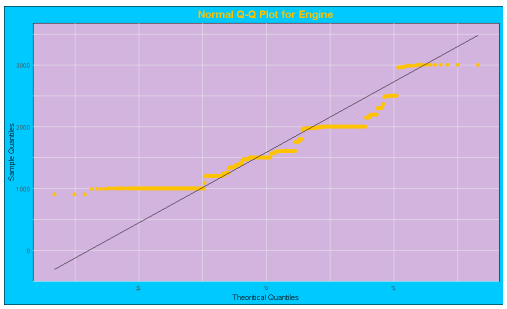

Figure 16: Q-Q Plot of Engine

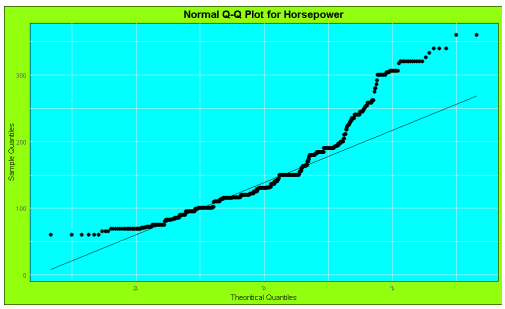

Figure 17: Q-Q Plot of Horsepower

By also examining the Q-Q Plots and Histograms of Engine and Horsepower, we came to the conclusion that both are not normally distributed as the data points significantly differ from the reference line, with this difference being much greater for Engine.

5.2 Continuous vs. Categorical

5.2.1 T-tests for Binary categorical variables

For Binary categorical variables **Welch's Two Sample T-tests** were conducted

	CO2	
	t-stat	**p-value**
Clima	-8.7655	2.2e-16
Air_cond..	8.9895	2.2e-16
Back_el.._win..	-2.1844	0.03008
Heated_mi..	-3.9647	0.000158

Table 7: Welch's Two Sample T-tests results

The results from **Table 7** indicate p-value < 0.05 for all the binary categorical variables, indicating statistical significance, allowing us to reject the null hypothesis (that the difference in group means is 0) and suggesting a meaningful difference between group means. That suggests that the mean of CO2 was significantly different between vehicles with, or without the above characteristics.

5.2.2 ANOVA for Non-Binary categorical variables

For Fuel (Categorical Variable with More Than Two Levels) **ANOVA** was used to test differences between the 3 main continuous variables.

	Fuel
	p-value
Engine	2e-16
CO2	2e-16
Horsepower	2e-16

Table 8: ANOVA results

The results indicate that again the null hypothesis is rejected , thus indicating that Engine size, CO2 emissions and Horsepower varied significantly across fuel types.

5.3 Continuous vs. Continuous

After addressing the normality issues, we started the correlation analysis by calculating correlations between continuous variables. The graphs below were created to provide an intuitive understanding of these relationships, in addition to scatter plots from the previous sections with regression lines.

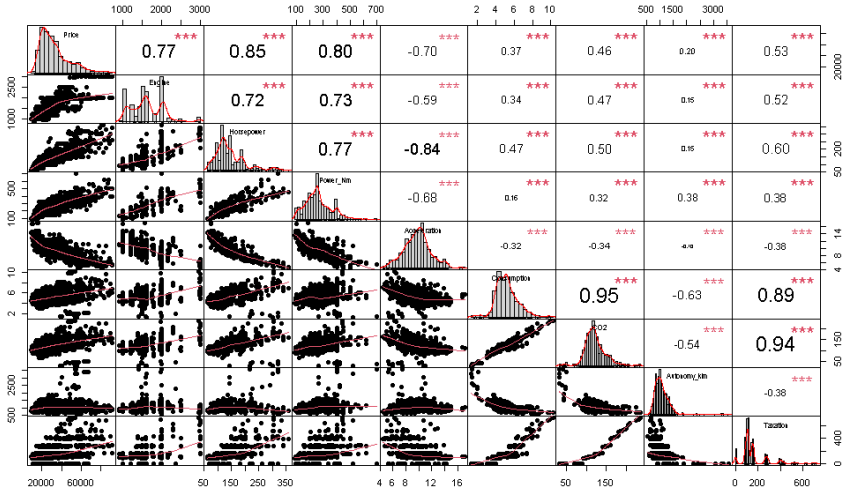

Figure 18: Correlation chart of continuous variables

This correlation chart is consisted of :the distribution graphs of each variable in the diagonal, the Pearson's correlation coefficient between every variable in the upper part, and scatter plots with reference line (see **Subsection 4.2.2**) between the variables in the lower part.

Firstly, CO2 is greatly positively correlated with Consumption and Taxation, representing a coefficient of r = 0.95 and r = 0.94, respectively. Same conclusion can be extracted from **Figure 8** in **Subsection 4.2.2**. In addition, strong positive correlations occurred for the variables: Consumption and Taxation (r = 0.89), Price and Horsepower (r = 0.85), Price and Power_Nm(r = 0.80), Price and Engine(r = 0.77), Horsepower and Power_Nm(r = 0.77), Engine and Power_Nm(r = 0.72). Lastly a strong negative correlation was noticed between the variables Horsepower and Acceleration(r = -0.84).

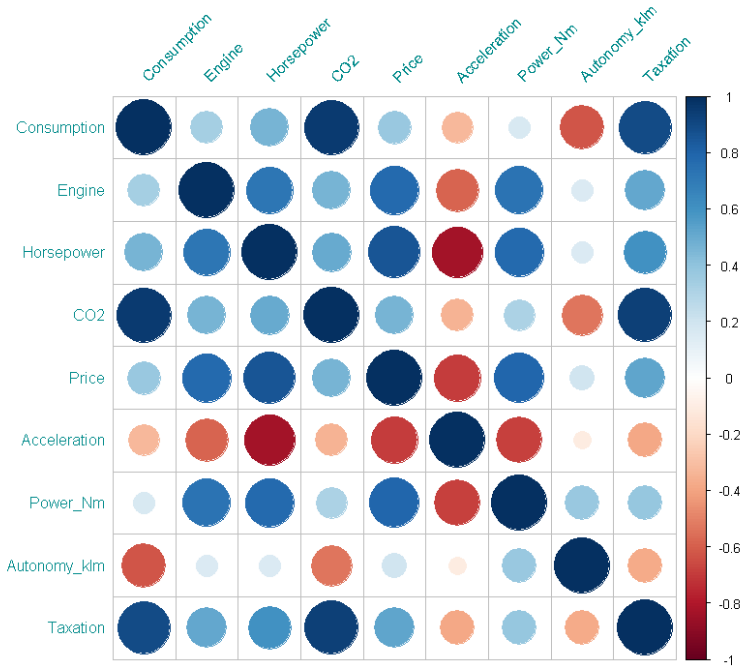

Figure 19: Correlation Heat-map of continuous variables

These relationships can be more obvious in the above heat-map, where the tensity of the color and the size of the circle represent the distance of r from 1 or -1, with deep blue and deep red representing r = 1 and r = - 1 , respectively. Note that the diagonal has a value of r = 1 throughout the plot, as it shows the correlation of a variable with itself.

5.4 Categorical vs. Categorical

	Clima		Air_cond..		Back_electric_wi..		Heated_mi..	
Fuel	**NO**	**YES**	**NO**	**YES**	**NO**	**YES**	**NO**	**YES**
Diesel	80	289	288	81	40	329	13	356
Diesel Mild-Hybird	0	4	4	0	0	4	0	4
Gasoline	199	449	451	197	109	539	54	594
Hybrid-Plug-in	0	20	20	0	0	20	0	20
LPG	2	1	1	2	1	2	1	2
Mild-Hybird	10	30	29	11	6	34	0	40
Hybrid	2	34	34	2	1	35	2	34

Table 9: Contingency table

The contingency table above was used to perform the **Pearson's Chi-squared tests**.

	Fuel	
	Chi-Squared	**p-value**
Clima	29.769	4.35e-05
Air_cond..	28.418	7.837e-05
Back_elec.._win..	15.963	0.01395
Heated_mi..	17.535	0.007506

Table 10: Pearson's Chi- squared test results

In the above table we can see the results of the Pearson's Chi-squared test between Fuel and all the other categorical variables. The result from all the test was the same: p-value < 0.05, with differences in the Chi Squared test statistics, so we rejected the null hypothesis (Ho) that states that there is no association between the two variables (the alternative hypothesis H1 is that there is an association of any kind). So we concluded that there should, maybe, be a significant association between the Fuel and the other categorical variables, indicating a dependence between them.

To save the reader's time, a Mosaic plot was used just for the Clima variable by Fuel categories, which is presented in **Subsection 4.2.1** (**Figure 6**)

,knowing that similar properties would apply to all the categorical variables, illustrating the proportionate relationships between them.

6 Regression Analysis

6.1 Introduction

Regression analysis is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. It helps in identifying patterns, making predictions, and testing hypotheses. In this project, the primary goal of the regression analysis is to identify the key predictors of CO2 emissions in vehicles and assess their impact. The CO2 emissions variable will be treated as the dependent variable, while various vehicle features such as engine size, fuel consumption, and fuel type will be the independent predictors.

Multiple Linear Regression was chosen for this analysis due to its simplicity and interpretability. It assumes a linear relationship between the dependent and independent variables, which is appropriate for examining continuous variables like CO2 emissions and engine size.

6.2 Full Model Specification

The initial regression model assumes all the variables of the dataset as potential predictors that could influence CO2 emissions, thus including them in the formula.

The starting model is specified as:

$$CO2 = \beta_0 + \beta_1 Engine + \beta_2 Consumption + \beta_3 Horsepower +$$

$$\beta_4 Price + \beta_5 Taxation + \beta_6 Power_N m + \beta_7 Acceleration + \beta_8 Autonomy\_klm +$$

$$\beta_9 Diesel + \beta_{10} Diesel\_Mild\_Hybird + \beta_{11} Hybrid\_Plug\_in + \beta_{12} LPG +$$

$$\beta_{13} Mild\_Hybrid + \beta_{14} Hybrid + \beta_{15} Air\_condition + \beta_{16} Clima +$$

$$\beta_{17} Back\_electric\_windows + \beta_{18} Heated\_mirrors + \epsilon$$

Explanation:

- CO2: Dependent variable, representing CO2 emissions.
- β_0: Intercept of the model.
- $\beta_1, \beta_2, \ldots, \beta_{18}$: Coefficients for each predictor variable, representing the change in CO2 emissions for a one-unit change in the predictor while keeping other variables constant.
- Variables like Engine, Consumption, Horsepower, etc., are continuous predictors.
- Variables like Diesel, Hybrid, Mild-Hybrid, etc., are dummy variables that represent Fuel type, where each coefficient compares the given category to the reference level.
- Variables like Clima, Air_condition, etc., are dummy variables that represent the average change in CO2 emissions when the variable is 1 (Yes), compared to when it is 0 (No), while keeping all other variables constant.
- ϵ : accounting for unexplained variability in CO2 emissions

Gasoline is omitted as the reference group and thus, it has not been added in the model as a dummy variable. The coefficients of the remaining dummies though, will be interpreted relative to Gasoline. When the vehicle type is Gasoline, all the dummy variables will take value equal to 0 and Gasoline will be interpreted using the intercept. This happened, to avoid multicollinearity and the **Phenomenon of Singularity (or Dummy Trap)**.

Provide a summary table or list of coefficients, Adjusted R^2, and statistical significance of predictors.

Comment on the overall performance of the full model (goodness-of-fit).

Predictor	Coefficient	Standard Error	t-Statistic	p-Value
Intercept	2.722e+01	3.435e+00	7.924	5.62e-15
Engine	2.708e-03	6.284e-04	4.310	1.78e-05
Consumption	1.389e+01	3.780e-01	36.758	¡ 2e-16
Horsepower	-4.901e-02	8.564e-03	-5.723	1.35e-08
Price	5.018e-05	1.808e-05	2.776	0.00560
Taxation	8.021e-02	3.200e-03	25.066	¡ 2e-16
Power_Nm	3.059e-02	4.218e-03	7.251	7.77e-13
Acceleration	1.794e-01	1.347e-01	1.331	0.18337
Autonomy_klm	1.206e-03	8.666e-04	1.392	0.16430
Diesel	2.980e+00	6.197e-01	4.808	1.73e-06
Diesel Mild Hybird	3.366e+00	2.266e+00	1.485	0.13777
Hybrid Plug in	-2.421e+01	1.799e+00	-13.460	¡ 2e-16
LPG	1.846e+01	2.452e+00	7.527	1.07e-13
Mild Hybrid	-2.479e+00	7.111e-01	-3.486	0.00051
Hybrid	-3.019e+00	9.463e-01	-3.190	0.00146
Air_conditionYES	2.514e-01	2.123e+00	0.118	0.90577
ClimaYES	7.588e-01	2.125e+00	0.357	0.72108
Back_electric_windowsYES	9.226e-01	4.407e-01	2.093	0.03655
Heated_mirrorsYES	-9.578e-01	6.405e-01	-1.495	0.13510

Table 11: Full Model Regression Results *

6.3 Backward Selection

Backward Selection (Elimination) is a stepwise regression method used to remove non-significant predictors from the model. The process begins with the full model and iteratively removes the least significant predictors based on their p-values or AIC (Akaike Information Criterion).

Iteratively remove non-significant predictors based on p-values or AIC.

Steps:

1. **Start with the full model** of linear multi-regression
2. **Iteratively remove non-significant predictors based on p-values or AIC.** Remove the predictor with the highest p-value greater than

0.05. In this case, Air_condition could be removed first due to its high p-value.

3. **Reassess the model** and remove any other non-significant predictors. Similarly, these variables were removed (in chronological order) : Clima, Acceleration.

4. **Stop when all remaining predictors are statistically significant**.

Final Model After Backward Elimination:

$$\widehat{CO2} = 29.61 + 0.0027 * Engine + 13.99 * Consumption - 0.0546 * Horsepower +$$

$$0.00005 * Price + 0.0805 * Taxation + 0.0292 * Power\_Nm + 0.0014 * Autonomy\_klm +$$

$$3.188 * Diesel + 3.684 * Diesel\_Mild\_Hybird - 23.74 * Hybrid\_Plug\_in + 18.59 * LPG$$

$$-2.316 * Mild\_Hybrid - 2.860 * Hybrid +$$

$$0.9966 * Back\_electric\_windows - 0.9083 Heated\_mirrors + \epsilon$$

Where all predictors remain as statistically significant.

Predictor	Coefficient	Standard Error	t-Statistic	p-Value
Intercept	2.961e+01	1.943e+00	15.241	¡ 2e-16
Engine	2.710e-03	6.224e-04	4.354	1.46e-05
Consumption	1.399e+01	3.742e-01	37.384	¡ 2e-16
Horsepower	-5.469e-02	7.344e-03	-7.448	1.91e-13
Price	5.602e-05	1.778e-05	3.151	0.001671
Taxation	8.051e-02	3.135e-03	25.683	¡ 2e-16
Power_Nm	2.927e-02	4.122e-03	7.102	2.20e-12
Autonomy_klm	1.435e-03	8.566e-04	1.675	0.094233
Diesel	3.188e+00	5.990e-01	5.322	1.25e-07
Diesel Mild Hybird	3.684e+00	2.251e+00	1.637	0.101938
Hybrid Plug in	-2.374e+01	1.757e+00	-13.517	¡ 2e-16
LPG	1.859e+01	2.450e+00	7.588	6.87e-14
Mild Hybrid	-2.316e+00	6.998e-01	-3.310	0.000963
Hybrid	-2.860e+00	9.400e-01	-3.042	0.002403
Back_electric_windowsYES	9.966e-01	4.310e-01	2.312	0.020954
Heated_mirrorsYES	-9.083e-01	6.165e-01	-1.473	0.140931

Table 12: Final Model Regression Results *

	AIC	Adjusted R-squared
Full Model	6411.093	0.9768
Final Model	6408.514	0.9768

Table 13: Goodness of fit comparison

Table 13 compares the goodness-of-fit metrics for the full model and the final (reduced) model.

- **AIC (Akaike Information Criterion):** A lower AIC value indicates a better fit of the model while penalizing for complexity. The reduction in AIC suggests that the final model has a slightly better balance between goodness-of-fit and simplicity compared to the full model.
- **Adjusted R-squared:** The Adjusted R-squared for both models is **0.9768**, indicating that approximately **97.68%** of the variance in the dependent variable (CO2 emissions) is explained by the predictors in

each model. Since Adjusted R-squared does not change, it suggests that removing the variables in the final model did not significantly reduce the explanatory power.

To conclude, the final model achieves a similar level of explanatory power (Adjusted R-squared remains the same) while improving slightly in terms of model simplicity (lower AIC). This makes the final model more explanatory without compromising its predictive ability. This is desirable in regression analysis as it avoids overfitting and enhances interpretability.

6.4 Marginal Effect Interpretation (or Ceteris Paribus Analysis

Marginal Effect Interpretation helps us economically interpret the regression coefficients in a comparative format.

1. **Intercept** : The baseline CO emissions for a gasoline-powered vehicle (as it is the reference group), when all predictors are set to zero, is expected to be **29.61 g/km**, while all other variables remain constant (ceteris paribus).

2. **Engine**: An increase of 1 cubic centimeter (cc) in engine size is expected to increase CO emissions by **0.00271 g/km**, while all other variables remain constant (ceteris paribus). This suggests that larger engines tend to emit more CO, aligning with expectations in vehicle design.

3. **Horsepower**: An increase of 1 horsepower is expected to decrease CO emissions by **0.05469 g/km**, while all other variables remain constant (ceteris paribus). This counterintuitive result may indicate that higher horsepower engines are more efficient due to advanced technologies.

4. **Consumption**: An increase of 1 liter/100 km in fuel consumption is expected to increase CO emissions by **13.99 g/km**, while all other variables remain constant (ceteris paribus). This is highly significant and emphasizes the direct relationship between fuel usage and emissions

5. **Price**: An increase of 1 euro in the price of a vehicle is expected to increase CO emissions by **0.00005602 g/km**, while all other variables

remain constant (ceteris paribus). This could imply that luxury vehicles often prioritize performance over efficiency.

6. **Taxation:** An increase of 1 euro in annual taxation is expected to increase CO emissions by **0.08051 g/km**, while all other variables remain constant (ceteris paribus). This might reflect the tendency of governments to tax higher-emission vehicles more.

7. **Power_Nm :** An increase of 1 Newton-meter (Nm) in power is expected to increase CO emissions by **0.02927 g/km**, while all other variables remain constant (ceteris paribus). This suggests that vehicles designed for high power output emit more CO.

8. **Autonomy:** An increase of 1 kilometer in autonomy is expected to increase CO emissions by **0.001435 g/km**, while all other variables remain constant (ceteris paribus), but this effect is not statistically significant ($p > 0.05$).

Dummy Variables

9. **Diesel:** The mean difference in CO emissions between diesel-powered and gasoline-powered vehicles is expected to be **3.188 g/km**, while all other variables remain constant (ceteris paribus). This confirmis their higher emissions compared to other fuels.

10. **Diesel Mild-Hybrid:** The mean difference in CO emissions between diesel mild-hybrid vehicles and gasoline-powered vehicles is expected to be **3.684 g/km**, but this effect is not statistically significant ($p<0.05$), while all other variables remain constant (ceteris paribus).

11. **Hybrid Plug-in:** The mean difference in CO emissions between hybrid plug-in vehicles and gasoline-powered vehicles is expected to be **-23.74 g/km**, while all other variables remain constant (ceteris paribus). Hybrid plug-ins show a significant reduction in emissions, indicating their environmental benefits.

12. **LPG:** The mean difference in CO emissions between LPG vehicles and gasoline-powered vehicles is expected to be **18.59 g/km**, while all other variables remain constant (ceteris paribus). LPG vehicles

increase emissions by 18.59 g/km reflecting their inefficiency compared to other alternatives.

13. **Mild-Hybrid:** The mean difference in CO emissions between mild-hybrid vehicles and gasoline-powered vehicles is expected to be **-2.316 g/km**, while all other variables remain constant (ceteris paribus). Mild hybrids reduce emissions by 2.316 g/km , emphasizing their role in lowering environmental impact.

14. **Hybrid:** The mean difference in CO emissions between hybrid vehicles and gasoline-powered vehicles is expected to be **-2.860 g/km**, while all other variables remain constant (ceteris paribus). Hybrid vehicles also reduce emissions significantly by 2.86 g/km .

15. **Back_Electric_Windows (Yes):** The mean difference in CO emissions between vehicles with rear electric windows and those without is expected to be **0.9966 g/km**, while all other variables remain constant (ceteris paribus). Vehicles with rear electric windows emit 0.9966 g/km more, a surprising result that might reflect correlation rather than causation.

16. **Heated Mirrors (Yes) :** The mean difference in CO emissions between vehicles with heated mirrors and those without is expected to be **-0.9083 g/km**, but this effect is not statistically significant ($p>0.05$), while all other variables remain constant (ceteris paribus).

6.5 Model Diagnostics

Interpreting diagnostic plots is crucial in assessing the quality and validity of our regression model.

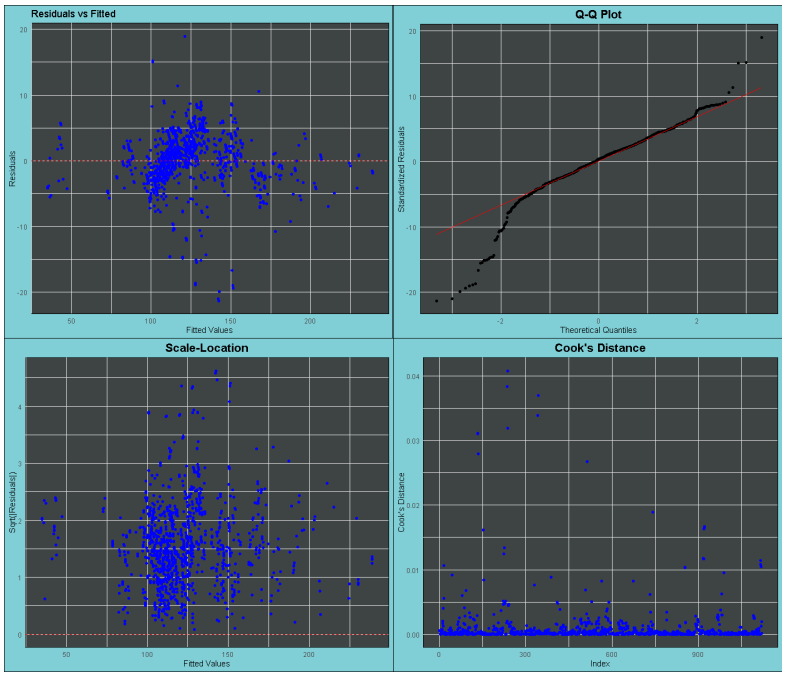

Figure 20: Regression Diagnostic Plots

The plots shown in **Figure 20** are:

- **Residuals vs Fitted Plot** This plot helps identify if there is a linear relationship between the predictors and the response variable. It also helps to check for non-linearity, heteroscedasticity (non-constant variance of errors), and outliers.
- **Q-Q Plot** This plot compares the distribution of the residuals with a normal distribution. If the residuals follow a normal distribution, the points should lie approximately along the 45-degree line.
- **Scale-Location Plot** This plot shows the square root of the absolute residuals vs the fitted values. It helps to detect heteroscedasticity, which occurs when the variance of the residuals changes across levels of the fitted values.

- **Cook's Distance Plot** Cook's distance is a measure of the influence of each data point on the regression model. High Cook's distance values indicate that the corresponding data point is an influential observation, meaning it has a large effect on the estimated coefficients.

So for Residuals vs Fitted Plot, the residuals seem to be randomly scattered around the horizontal line at zero. This indicates that the model fits the data well and there is no discernible pattern in the residuals. Nevertheless, there is a wide spread of residuals (the spread of residuals increases or decreases as fitted values change) in the middle. This could suggest heteroscedasticity, meaning that the variance of the errors is not constant. In this case, we might had to apply transformations like logarithms or use robust standard errors.

By observing the Q-Q Plot, we notice some deviations of the points from the line (especially at the ends), suggesting that the residuals might not be normally distributed.

From the Scale-Location Plot, we confirm the suggestion of heteroscedasticity from before, as there is a funnel shape for the points. So the model could be inappropriate and the transformation we referred to before, could come in handy.

Finally, there are some points with a high Cook's Distance, so they could be considered influential. These points might be outliers or have a disproportionate effect on the regression coefficients.

6.6 Conclusion

The regression analysis provides significant insights into the factors affecting CO emissions, highlighting both continuous predictors and categorical variables' impacts. The key takeaways from this analysis are summarized below:

For **Continuous Predictors**, **Fuel Consumption** emerged as the most influential factor, with higher fuel consumption strongly linked to increased CO . **Engine Size** and **Power (Nm)** also contributed positively to CO emissions, emphasizing the trade-offs between vehicle performance and environmental sustainability. Interestingly, **Horsepower** showed a negative relationship, suggesting that higher efficiency engines may achieve greater power output with lower emissions.

For **Categorical Predictors**, fuel type plays a critical role in emissions.

Gasoline was used as the reference group, and **diesel** vehicles exhibited higher emissions, while **hybrid plug-in** and **mild hybrid** vehicles showed notable reductions in emissions. This underscores the environmental benefits of hybrid technologies. **LPG-powered** vehicles, while an alternative to traditional fuels, exhibited higher emissions than gasoline vehicles, warranting further examination.

Additionally, the presence of features like **rear electric windows** was surprisingly associated with marginally higher emissions, suggesting that additional electrical components might indirectly impact energy efficiency. Features like **heated mirrors** did not show a statistically significant effect, indicating limited influence on emissions.

In conclusion, this regression analysis not only sheds light on the determinants of CO emissions but also offers actionable insights for stakeholders in the automotive industry, policymakers, and researchers aiming to combat climate change through informed decision-making.

7 Conclusions

7.1 Key Insights and Interpretations

The analysis of vehicle data provides useful insights for both environmental policies and car design.

To start with, there's a clear link between higher fuel consumption and increased CO emissions. Cars with bigger engines or less efficient fuel use tend to produce more emissions, suggesting that improving fuel efficiency could help lower pollution. Also cars with larger engines consume more fuel and emit more CO. This shows that smaller engines or more efficient technologies could reduce both fuel consumption and emissions. Additionally, hybrid cars, tend to be more environmentally friendly compared to traditional **gasoline** or **diesel** cars. Encouraging the use of these alternative fuels can help reduce emissions.

Similarly, features like **air conditioning** and **electric windows** can increase fuel consumption. While these are important for consumer comfort, they may have a negative impact on fuel efficiency and emissions.

Consumers are becoming more interested in **sustainable vehicle technologies**. Policies that promote **hybrid** and **electric vehicles** are likely

to push the market toward greener alternatives, reducing reliance on fossil fuels. These insights suggest that for both **environmental policies** and **car manufacturers**, focusing on fuel-efficient and low-emission technologies, such as **electric or hybrid vehicles**, is key to reducing environmental impact and meeting future consumer demand.

7.2 Practical Implications

The strong link between fuel consumption and CO emissions suggests the need for stricter fuel-efficiency rules and taxes on high-emission vehicles. With this in mind, the analysis shows that hybrid and plug-in hybrid vehicles produce much less CO. This could support policies that encourage their use, like tax incentives and expanding charging stations. By implementing these strategies, Greece can align its automotive sector with global sustainability targets while promoting technological innovation.

7.3 Limitations of the Analysis

First of all, the dataset we used dates back to 2020 in the market of Greece. Both of these facts could have limited our analysis, first because the market is rapidly changing and second because our results possibly could not apply worldwide for the reason that every location and country has it's own, different characteristics. Additionally, important factors, like car weight and driving habits, weren't included and could affect the results. Also, high correlation between some variables, such as Engine and Horsepower and the effect of heteroscedasticity in the regression, could distort the analysis.

7.4 Suggestions for Future Research

Future research should including factors like fuel prices and economic conditions by region. This could give a better understanding of CO emissions. Also, tracking vehicle emissions over time could help evaluate the effectiveness of current policies and innovations.

8 Bibliography

- **Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004).** "Applied Linear Regression Models" (4th ed.). McGraw-Hill.
- **Wooldridge, J. M. (2016).** *Introductory econometrics: A modern approach* (6th ed.). Cengage Learning.
- **Gujarati, D. N., & Porter, D. C. (2009).** *Basic Econometrics* (5th ed.). McGraw-Hill.
- **Chang, W. (2018).** *R graphics cookbook: practical recipes for visualizing data*. O'Reilly Media.
- **Pavía, J. M. (2020).** R Graphics. *Journal of Statistical Software*, *92*, 1-4
- **Datanovia. (n.d.).** *Datanovia: Statistics Made Easy*. Retrieved from https://www.datanovia.com/en/
- **R Graph Gallery. (n.d.).** *R Graph Gallery: Data Visualization with R*. Retrieved from https://r-graph-gallery.com/index.html
- **R-bloggers. (n.d.).** *R-bloggers: The R Community Blog*. Retrieved from https://www.r-bloggers.com/
- **Statistics How To. (n.d.).** *Statistics How To: Statistics for Beginners*. Retrieved from https://www.statisticshowto.com/
- **European Commission. (n.d.)**. CO2 emission performance standards for cars and vans. Retrieved from https://climate.ec.europa.eu/eu-action/transport/road-transport-reducing-co2-emissions-vehicles/co2-emission-performance-standards-cars-and-vans_en
- **EPA. (n.d.)**. Highlights of the automotive trends report. Retrieved from https://www.epa.gov/automotive-trends/highlights-automotive-trends-report
- **International Council on Clean Transportation (ICCT). (2019).** Gasoline vs. diesel CO2 emissions. Retrieved from https://theicct.org/wp-content/uploads/2021/06/Gas_v_Diesel_CO2_emissions_FV_20190503_1.pdf