EXPLORING THE RELATIONSHIP BETWEEN CAR CHARACTERISTICS AND ENVIRONMENTAL FACTORS

by conducting a Statistical Analysis

Author: Koutsourelis Ioannis - 6152
Supervisor: Tsagris M

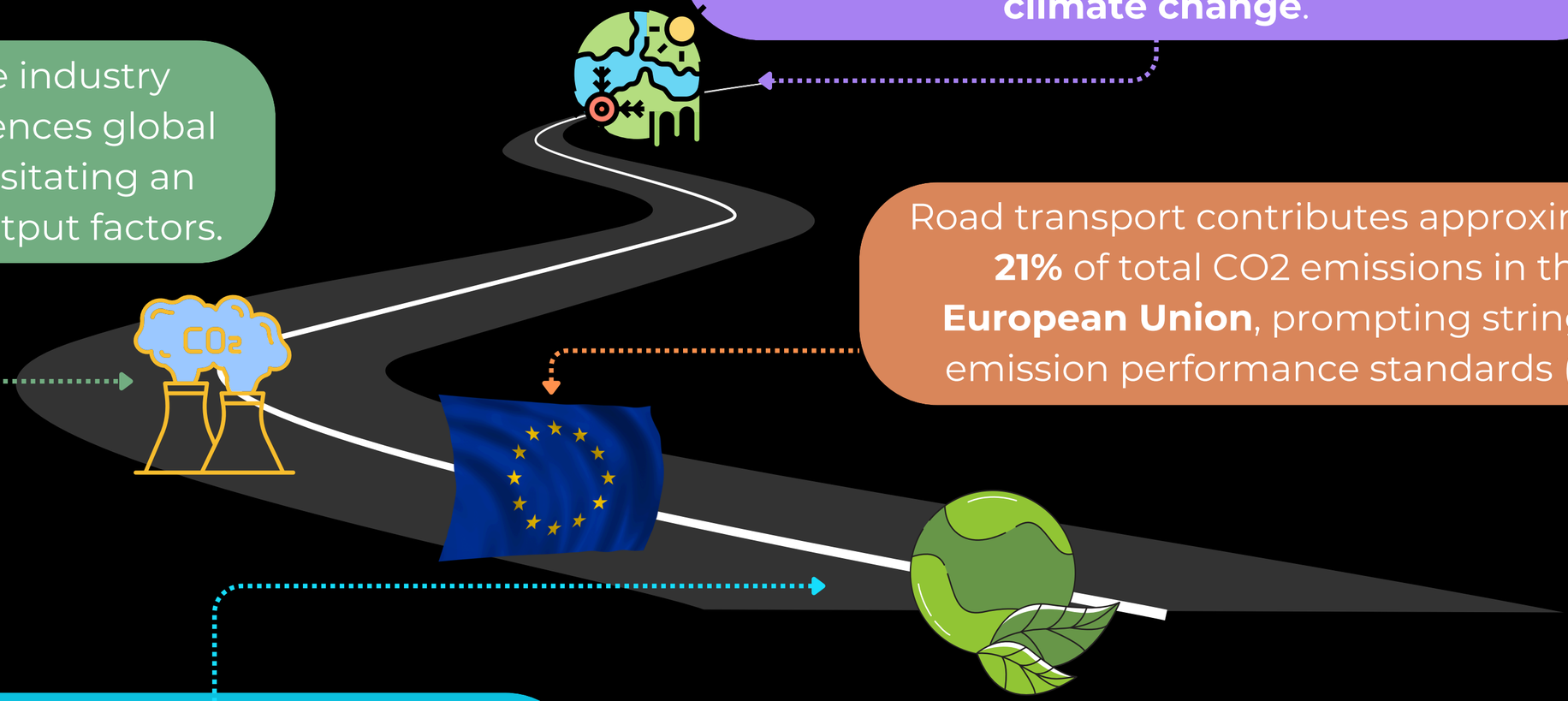
Course: Introduction to programming using R

Winter Semester | 2024-2025
University of Crete | Department of Economic Sciences | Rethymnon

INTRODUCTION

In recent years Environmental Sustainability concerns have increased the focus on vehicle emissions, particularly CO2, due to their impact on **climate change**.

The automotive industry significantly influences global emissions, necessitating an analysis of **CO2** output factors.



Road transport contributes approximately **21%** of total CO2 emissions in the **European Union**, prompting stringent emission performance standards (EU).

Regulation (EU) 2019/631 sets progressively stricter CO2 emission targets for new passenger cars, aiming for **net-zero emissions by 2035**.

INTRODUCTION

KEY FACTORS

Engine Size
Larger **engines** and higher **fuel consumption** rates generally lead to increased CO2 emissions (EPA).

Horsepower
Increased horsepower in vehicles generally leads to higher CO_2 emissions and lower fuel economy (EPA, 2021).

Fuel Type
Diesel vehicles emit about **13% more** CO2 per liter of fuel burned compared to gasoline vehicles (The ICCT).

Fuel Consumption
Real-world CO_2 emissions and fuel consumption from diesel and petrol vehicles are about **20% higher than official values** (European Commission, 2024).

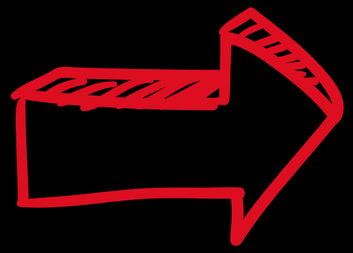
PURPOSE OF THE STUDY

The aim of this study is to:

- **Investigate** the relationships between CO_2 emissions and vehicle attributes.
- **Analyze key factors** such as engine size, fuel consumption, and fuel type.
- **Identify patterns** and trends in the dataset.
- **Highlight opportunities** to improve vehicle efficiency.
- **Focus on reducing environmental impacts** through actionable insights.

THE ORIGINAL DATASET

New cars in the Greek market from the year 2020 (Autotriti.com).



Includes 1,505 car entries and 14 variables

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

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”)

DATA CLEANING

1. Missing values (NAs), false input and other **issues were addressed**.

2. The "Electric" class from the 'Fuel' was removed due to **severe data loss**.

3. **Outliers removal process**. Top 5% of Price was trimmed to focus on general consumer trends.

4. 11 **additional outliers** were removed that were on the bottom 5% or very close, but not over the top 5%.

5. A total of **385 observations** and 1 class was removed.

6. The final dataset consisted of **1,120** observations and 14 variables.

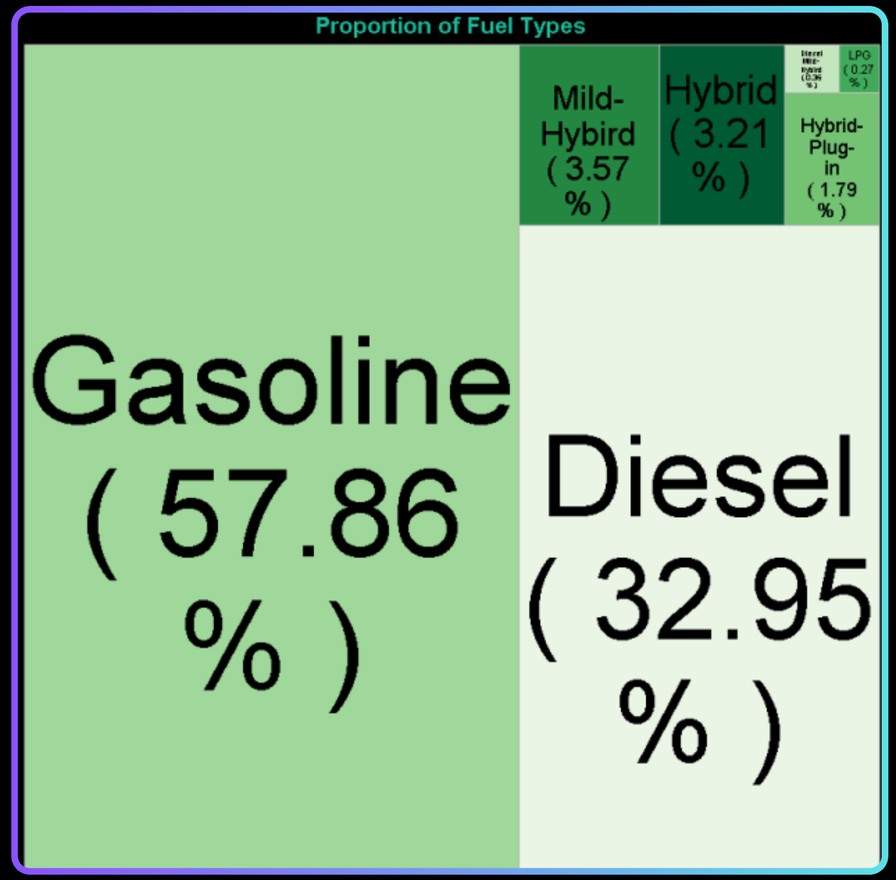
DESCRIPTIVE STATISTICS

Frequencies

Variable	Air_con..		Back_ele.._win..	
Value	Yes	No	Yes	No
Frequencies	293	827	963	157
Relative Frequencies (%)	26.16	73.84	14.02	85.98

Variable	Clima		Heated_mir..	
Value	Yes	No	Yes	No
Frequencies	827	293	1050	70
Relative Frequencies (%)	73.84	26.16	93.75	6.25

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



- Air condition and Clima are **substitutes**, with opposite frequency distributions
- Heated mirrors show the largest difference between 'Yes' (93.75%) and 'No' (6.25%).
- Gasoline and Diesel are the two **most dominant fuel types**, comprising 90.81% of the fuel categories.

DESCRIPTIVE STATISTICS

Basic statistics

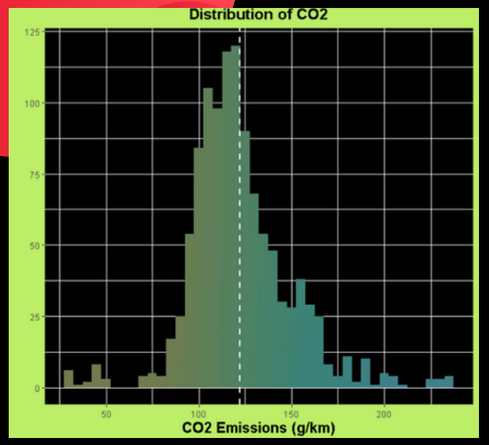
	Price	Engine	Horsepower	Power Nm	Acceleration
Mean	30752	1566	144.5	254.4	9.942
Median	26695	1499	130	250	10
Maximum	91550	2998	360	700	17.1
Minimum	9100	898	60	88	4.5
1st Quarter	19297	1199	112	200	8.575
3rd Quarter	37005	1969	165	306.2	11.1
Standard Error	473.51	12.55	1.63	2.86	0.06
Median Absolute 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

- **Price:** Mean (€30,752) > Median (€26,695), indicating a positive skew influenced by expensive vehicles.
- **Price:** High range (€82,450) and large standard deviation (€15,846.69) indicate significant variability.
- **Acceleration:** Skewness (0.14) and Kurtosis (0.24) indicate a more symmetric and flat distribution.

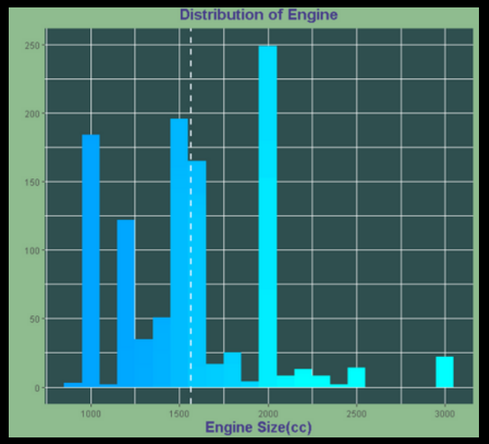
	Consumption	CO2	Autonomy klm	Taxation
Mean	5.141	122.1	1056	162.3
Median	5	119	1000	117
Maximum	10.2	237	3571	723
Minimum	1.3	31	541	0
1st Quarter	4.3	105	854	103
3rd Quarter	5.8	136	1163	163
Standard Error	0.04	0.82	9.69	3.57
Median Absolute Deviation	1.04	20.76	225.36	45.96
Range	8.9	206	3030	723
Standard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

- **Consumption:** Mean (5.141 L/100 km) ≈ Median (5 L/100 km), indicating most vehicles are fuel-efficient.
- **Autonomy:** Skewness (3.57) and Kurtosis (21) reflect extreme positive skew and a highly peaked distribution, indicating outliers.
- **Taxation:** Skewness (2.03) and Kurtosis (4.92) show high positive skewness and a sharper peak due to high-tax outliers.

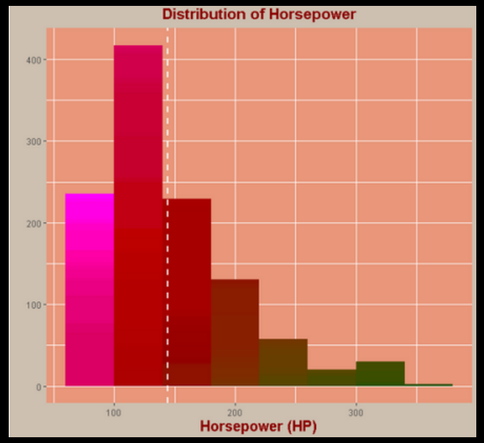
DATA VISUALIZATION | Univariate



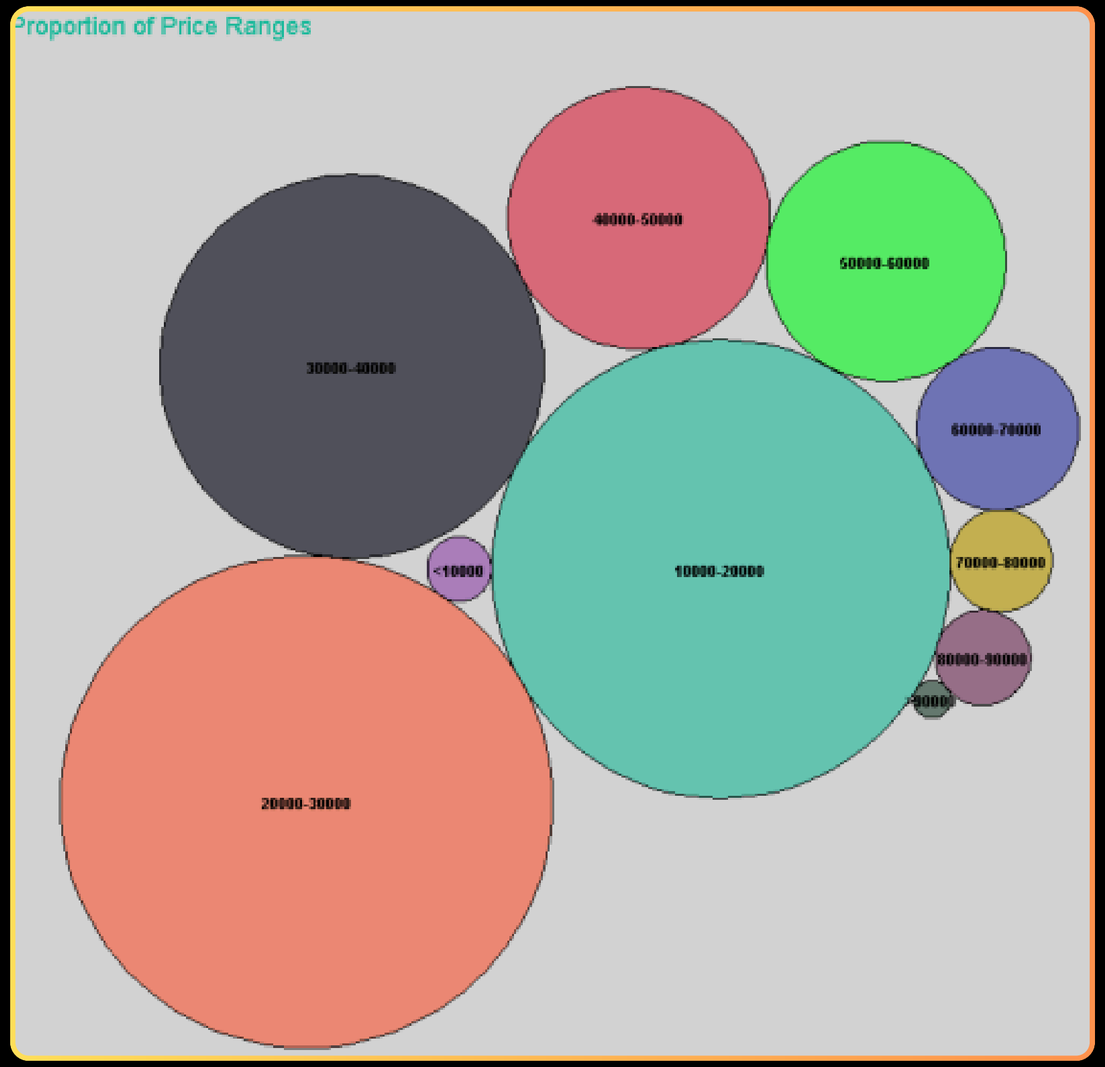
Slightly right-skewed, indicating most vehicles have lower emissions.



Concentrated around 1,000–1,600 cc.

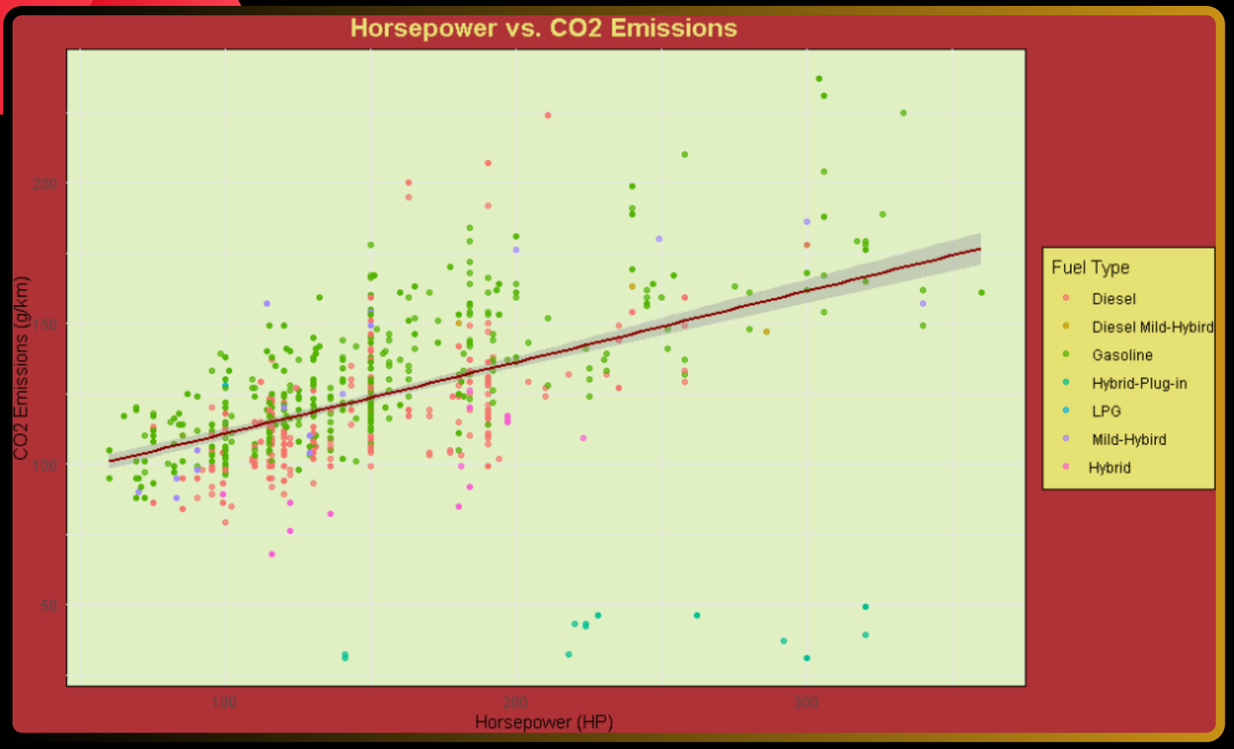


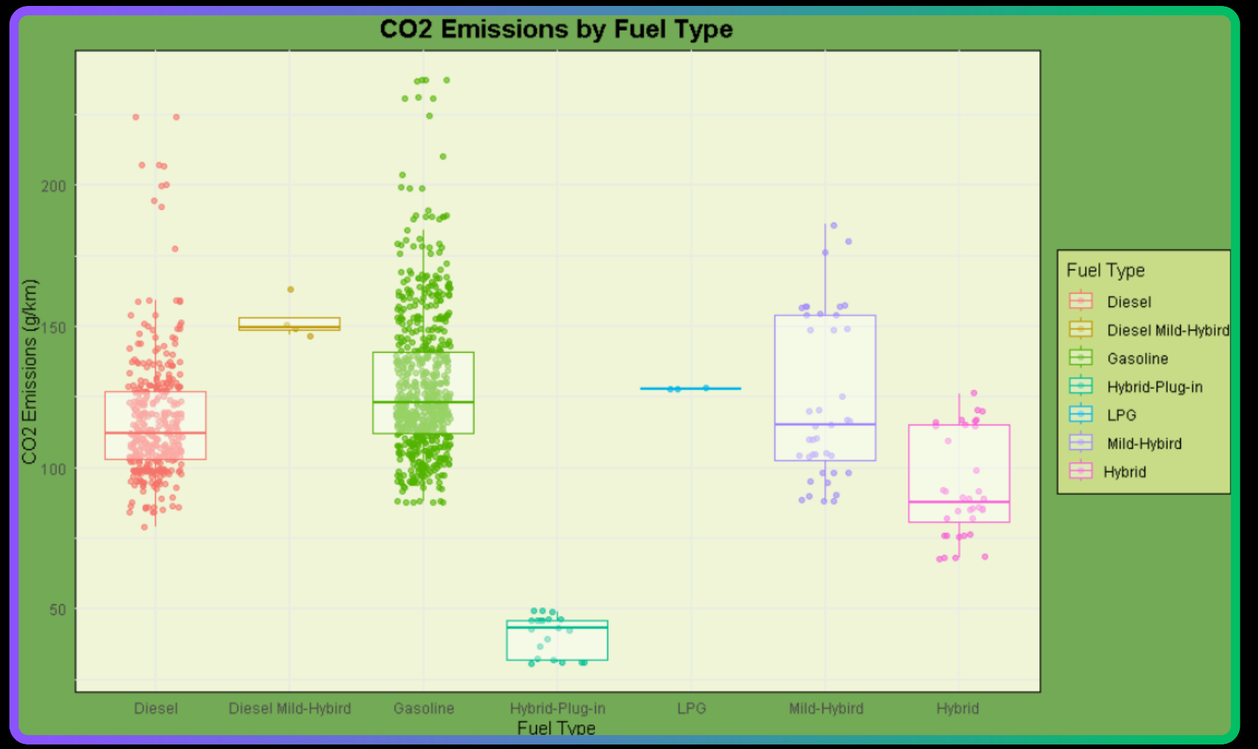
Most vehicles have moderate power, with a smaller number having very high horsepower.



- Majority of vehicles are concentrated in mid-level price ranges (€10,000 - €40,000).
- Indicating a strong market preference for moderately priced vehicles.

DATA VISUALIZATION | Bivariate

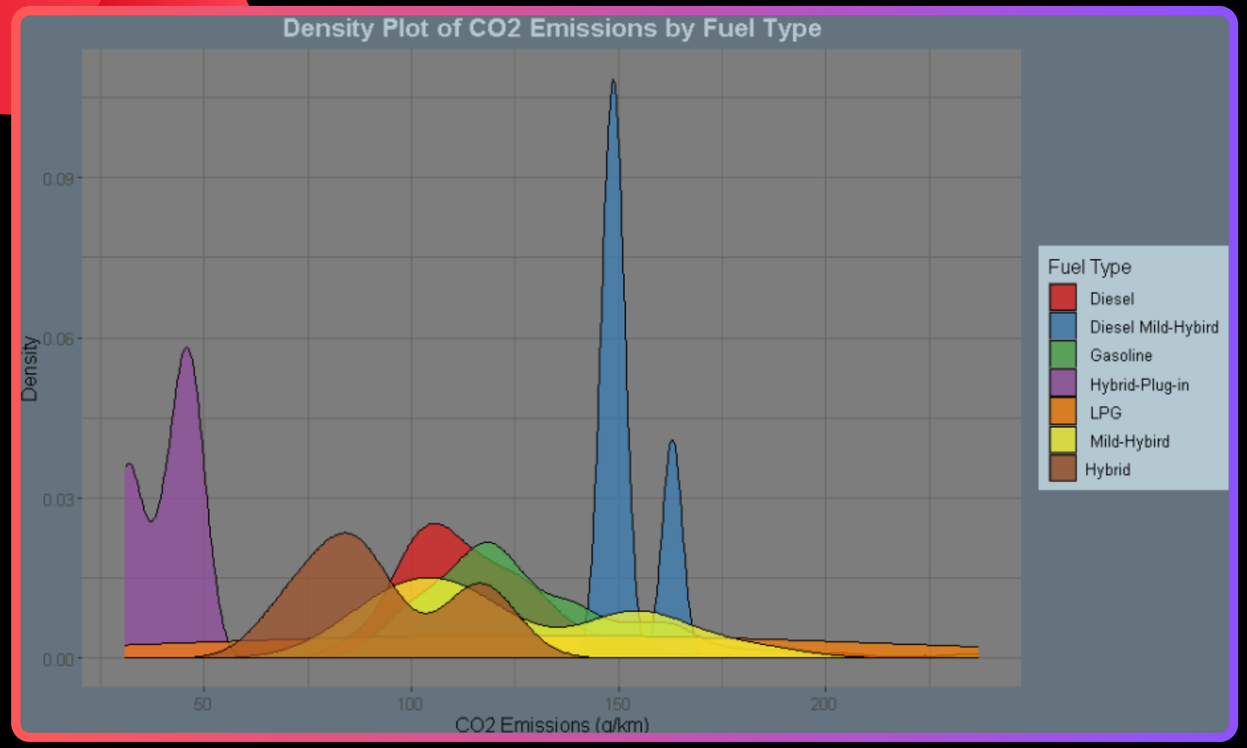


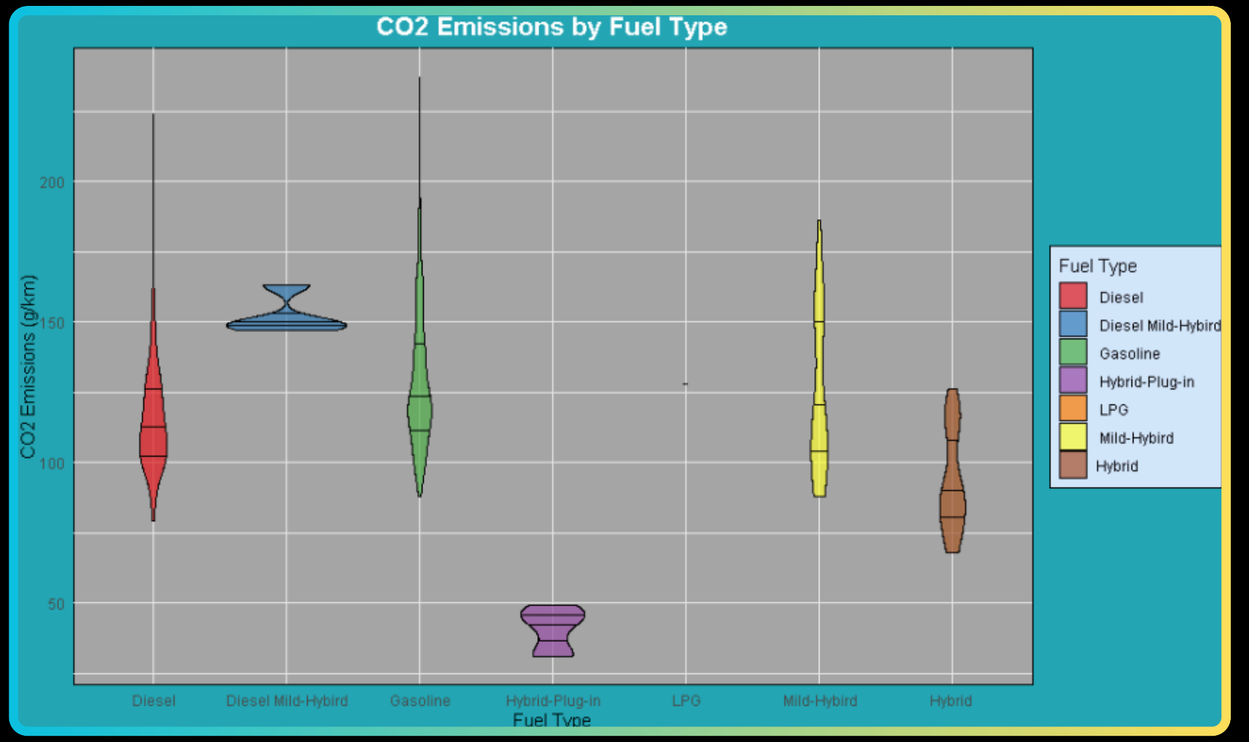


- **Higher Horsepower:** Linked to increased CO2 emissions, as **more powerful engines consume more fuel**.
- **Electric/Hybrid Vehicles:** Clustered at lower horsepower and emissions, emphasizing **efficiency over performance**.

- **Gasoline and Diesel Vehicles:** Higher Medians aligning with their **less environmentally friendly** design.
- **Electric and Hybrid Vehicles:** Display notably lower median CO2 emissions and Minimal Variation showing **consistency**.

DATA VISUALIZATION | Bivariate

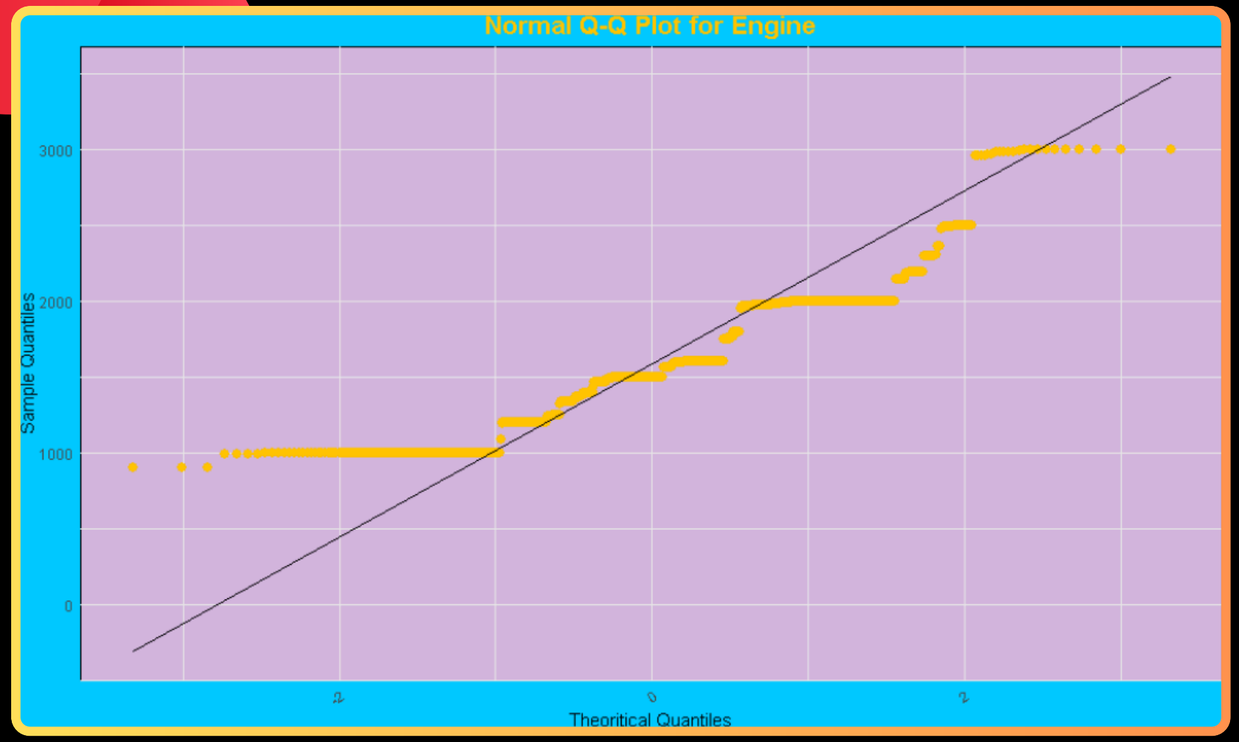


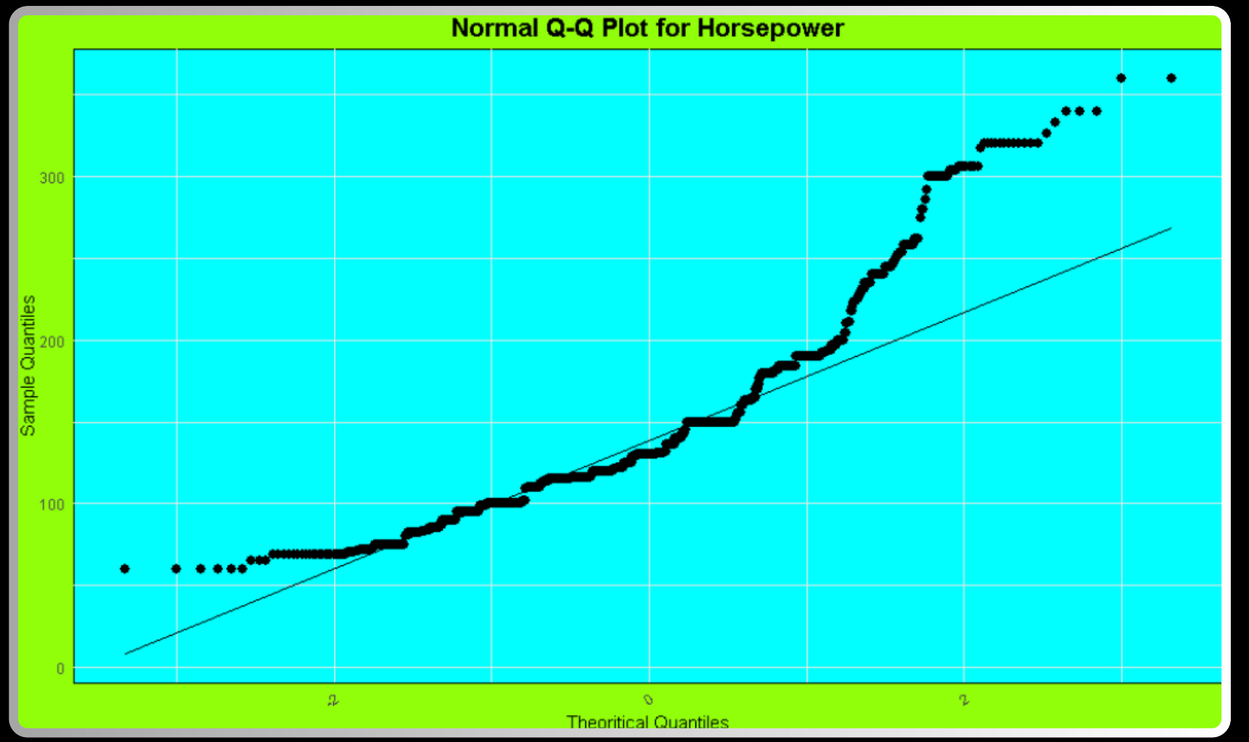


- **Bimodal Distribution:** Observed in all fuel types except gasoline and diesel, suggesting **two distinct vehicle groups** within these categories.
- **Hybrid Vehicles:** (Except Diesel Mild-Hybrid) peak closer to zero, confirming their **low-emission status**.

- **Gasoline and Diesel:** Display wide distributions, with values concentrated in the mid-range
- **Outliers:** Visualized as extended tails in the violin plots for gasoline and diesel, suggesting the **presence of high-emission vehicles** within these categories.

CORRELATION | Normality

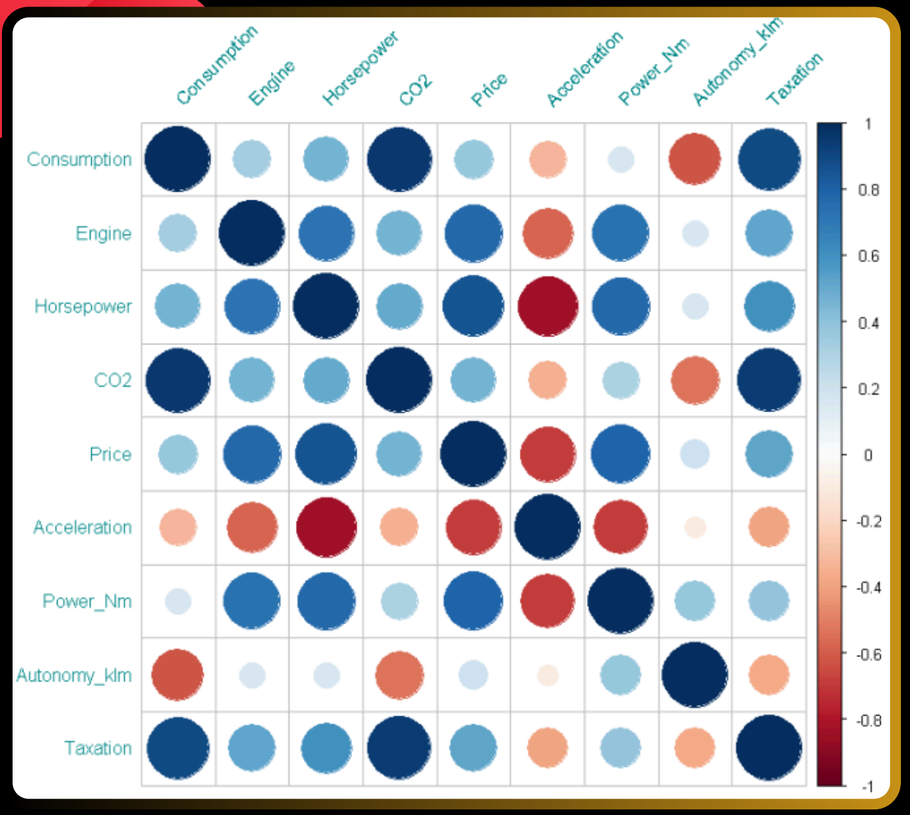


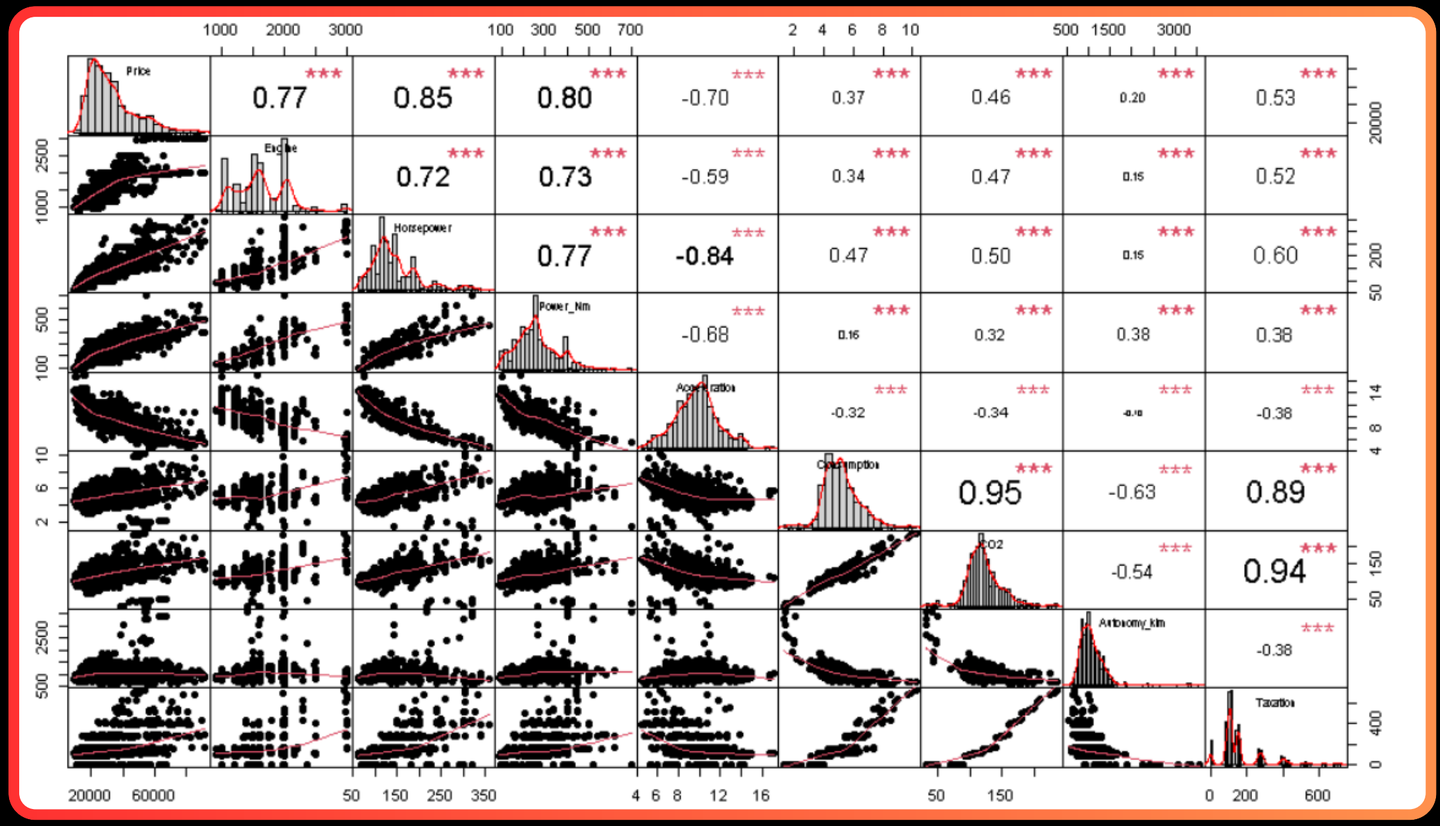


- **Q-Q Plots:** Data points significantly deviate from the reference line, **indicating non-normality** for both Engine and Horsepower.

- **Engine:** Shows a greater deviation from normality compared to Horsepower, suggesting a more pronounced skew or variability in engine sizes.

CORRELATION | Between continuous variables





Key Findings

- CO2: Strong positive correlations between Consumption (r = 0.95) and Taxation (r = 0.94).
- Consumption and Taxation (r = 0.89).
- Price and Horsepower (r = 0.85).
- Horsepower and Acceleration (r = -0.84).

CORRELATION | Tests

Welch's Two Sample T-tests

	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

- P-value < 0.05: Null Hypothesis rejected, as differences in group means are significant.
- Indicates **statistical significance** for all binary categorical variables.

ANOVA

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

- P-value < 0.05: Null Hypothesis rejected: Indicates significant differences in means across groups.
- Engine size, CO2, and Horsepower show **significant variation** across different fuel types.

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

- P-value < 0.05: Null Hypothesis rejected.
- Suggests a **significant association** between Fuel and other categorical variables, **indicating dependency**.

REGRESSION | Model

Full Model

$$CO2 = \beta_0 + \beta_1 Engine + \beta_2 Consumption + \beta_3 Horsepower + \beta_4 Price + \beta_5 Taxation + \beta_6 Power_Nm + \beta_7 Acceleration + \beta_8 Autonomy\_klm + \beta_9 Diesel + \beta_{10} Diesel\_Mild\_Hybird + \beta_{11} Hybrid\_Plug\_in + \beta_{12} LPG + \beta_{13} Mild\_Hybrid + \beta_{14} Hybrid + \beta_{15} Air\_condition + \beta_{16} Clima + \beta_{17} Back\_electric\_windows + \beta_{18} Heated\_mirrors + \epsilon$$

Backwards Selection

(Air condition, Clima and Acceleration were removed)

Final Model

$$\widehat{CO2} = 29.61 + 0.0027 * Engine + 13.99 * Consumption - 0.0546 * Horsepower + 0.00005 * Price + 0.0805 * Taxation + 0.0292 * Power\_Nm + 0.0014 * Autonomy\_klm + 3.188 * Diesel + 3.684 * Diesel\_Mild\_Hybird - 23.74 * Hybrid\_Plug\_in + 18.59 * LPG - 2.316 * Mild\_Hybrid - 2.860 * Hybrid + 0.9966 * Back\_electric\_windows - 0.9083 Heated\_mirrors + \epsilon$$

Goodness-of-Fit

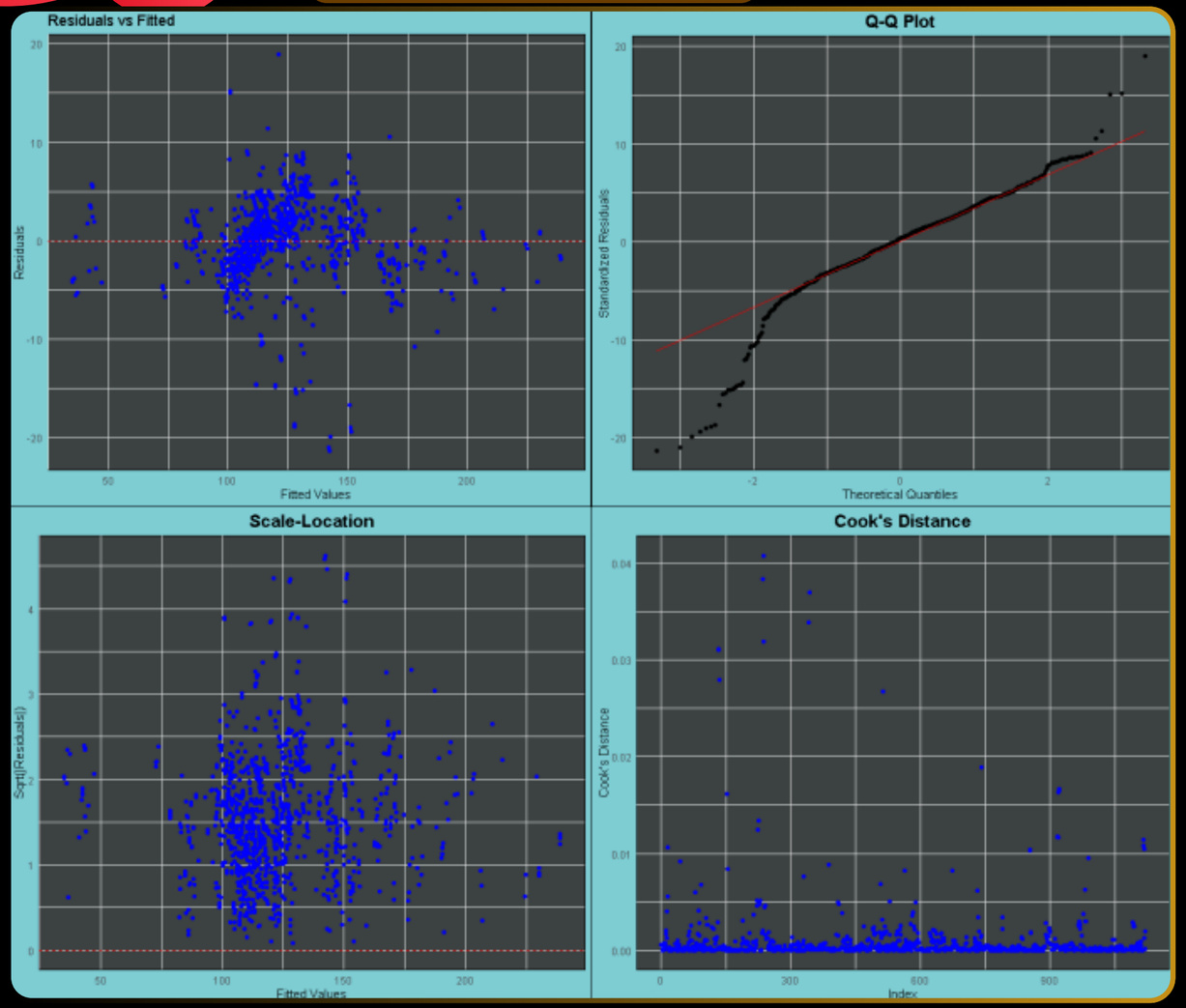
	AIC	Adjusted R-squared
Full Model	6411.093	0.9768
Final Model	6408.514	0.9768

- **AIC:** The final model shows a **reduction** in AIC, suggesting an improved balance between fit and simplicity compared to the full model.
- **Adjusted R-squared:** For both models, approximately 97.68% of the variance in CO2 emissions is explained by the predictors.

Final Model = **More Efficient**

REGRESSION | Diagnostics

Diagnostic Plots



- **Residuals vs. Fitted Plot:**

The residuals appear randomly scattered around the horizontal line at zero, indicating a generally good fit of the model with no obvious pattern.
However, the increasing or decreasing spread of residuals suggests **possible heteroscedasticity**, indicating that the variance of errors may not be constant.

- **Q-Q Plot:**

The deviations of points from the reference line, particularly at the ends, suggest that residuals might **not be normally distributed**.

- **Scale-Location Plot:**

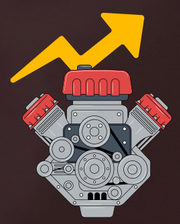
The funnel shape of points in this plot further confirms the presence of **heteroscedasticity**.

- **Cook’s Distance:**

Some points exhibit high Cook’s Distance, indicating they could be influential **outliers**. These points may disproportionately affect the regression coefficients.

CONCLUSIONS

- There is a **direct correlation** between higher fuel **consumption** and increased **CO2** emissions. Vehicles with larger engines or inefficient fuel use are significant contributors to pollution.



- Improving **fuel efficiency**, possibly through smaller engines or advanced technologies, can **reduce both fuel consumption and emissions**.

- **Gasoline** or **Diesel** vehicles demonstrate a **higher environmental impact** compared to Hybrid vehicles.

- Promoting the **adoption of hybrid and electric vehicles** can significantly contribute to **emission reduction**.

- Features like **air conditioning**, while enhancing consumer comfort, can **increase fuel consumption and emissions**.

- This highlights the need for **balancing comfort with fuel efficiency** in vehicle design.

REFERENCES

- Chang, W. (2018). R graphics cookbook: practical recipes for visualizing data. O'Reilly Media.
- Datanovia. (n.d.). Datanovia: Statistics Made Easy. Retrieved from https://www.datanovia.com/en/
- 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
- Gujarati, D. N., & Porter, D. C. (2009). Basic Econometrics (5th ed.). McGraw-Hill.
- 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
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models (4th ed.). McGraw-Hill.
- Pavía, J. M. (2020). R Graphics. Journal of Statistical Software, 92(1), 1-4.
- R-bloggers. (n.d.). R-bloggers: The R Community Blog. Retrieved from https://www.r-bloggers.com/
- R Graph Gallery. (n.d.). R Graph Gallery: Data Visualization with R. Retrieved from https://r-graph-gallery.com/index.html
- Statistics How To. (n.d.). Statistics How To: Statistics for Beginners. Retrieved from https://www.statisticshowto.com/
- Wooldridge, J. M. (2016). Introductory econometrics: A modern approach (6th ed.). Cengage Learning.

THANK YOU!

KOUTSOURELIS IOANNIS
6152

Winter Semester | 2024-2025
University of Crete | Department of Economic Sciences | Rethymnon