29/08/2024

Project CO2 emission by vehicles

Report 1

DATA SCIENTEST

Table of contents

1 Introduction to the project

1.1 Context

The project involves analyzing large datasets of vehicle characteristics and their corresponding CO2 emissions. This will require data preprocessing, exploratory data analysis, and the application of machine learning techniques to build predictive models.
Understanding and predicting vehicle CO2 emissions can have significant economic implications. It can inform policy decisions, influence car manufacturing processes, and potentially impact consumer choices, all of which have economic consequences.
Using common techniques in machine learning and analyzing the results can contribute to the discussion of the strengths and weaknesses of those techniques. Following a particular paper especially contributes to the repeatability of that paper and may emphasize or question particular subareas.

1.2 Objectives

The objectives of the project are: (i) analyze the relationship between vehicle characteristics and CO2 emissions (ii) develop a predictive model for CO2 emissions based on vehicle specifications (iii) identify the key factors that contribute most to vehicle CO2 emissions (iv) provide insights that could guide the development of lower-emission vehicles

Different members of the team have different experiences regarding the project:

- Koffi KOUMI: I have a moderate level of expertise. I am a mechanical engineer by training and I know how a combustion engine works.
- Arthur ROHR PASCHOAL CORREA CARDOSO: I am a Water Engineer with experience in renewable energy. My expertise in sustainable technologies and environmental impact analysis positions me well to contribute to this project, particularly in understanding the environmental implications of vehicle CO2 emissions.
- Naor Gabriel SIBONY:
  I am a university graduate, BSc Computer Science. I have 8 years of experience with both client and server side of Web Development, although most fresh experience mostly revolves around the Frontend.
- Claudia WISNIEWSKI: I have experience in Web Development with a Focus on Clean Code. My university Background in Digital Humanities allows me to understand and visualize Big Datasets.

Although we have not discussed the problem and the underlying models with any experts besides our mentor, we leverage several online materials dealing with CO2 emissions. For example, Zubair et al. studied in "Impact of Features on CO2 Emission from Fueling Vehicles" the key parameters that impact CO2 emission. They highlighted these parameters:

- Engine size. There is a direct relation of CO2 with the engine size (R2 = 0.72)
- Cylinder number. There is a direct relation of CO2 with cylinder number (R2=0.69)
- Brand. For example, Bugatti produced the highest CO2 (~ 500 g/km) while Smart brand found with less CO2 (~ 50 g/km)
- Vehicle type. For example, VAN-Passenger among the fueling vehicle class produced the highest CO2 (~ 400 g/km) while the least CO2 (~ 200 g/km) was observed in Station Wagon –small

- Transmission type. Automate manual with 5 gears (AM5) has low CO2 (150 g/km) while automatic with 7 and 10 gears (A7/A10) produced the highest CO2 (300 g/km)
- Fuel types. Natural gas (N) produced the least CO2 while ethanol (E85) and premium gasoline (Z) fuel caused the highest CO2 (> 250 g/km)

Al-Nefaie et al., in *Predicting CO2 Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model,* identified key features that are relevant for CO2 emissions forecast. Those parameters are:

- Vehicle class
- Engine size
- Cylinder transmission
- Fuel type
- Fuel consumption city (L/100 km)
- Fuel consumption hwy (L/100 km)
- Fuel consumption comb (L/100 km)
- Fuel consumption comb (mpg)

Since we are the only group in our cohort working on the CO2 emission project we didn’t look for help in our cohort either.

2 Understanding and manipulation of data

2.1 Framework

We used public data available from the European Environment Agency. The agency monitors the CO2 emissions of all the cars that come out each year. We use the most recent final data (2022). The data is in csv format with a total size of 2.4 GB. The uncleaned data has 10.7 million rows and 40 columns.

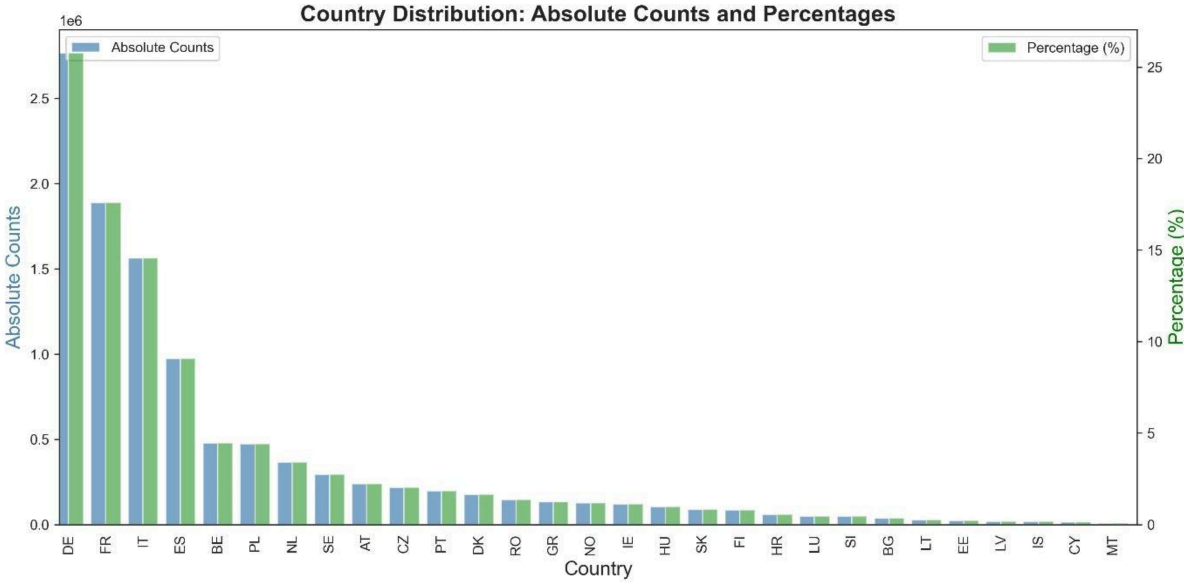


Figure 1: Distribution of the data across all the European countries

The data encompasses the 27 members of the European Union, as well as Norway and Island. Germany (DE), France (FR), and Italy (IT) represent around 60% of the overall data (Fig.1). This finding may be correlated to the countries' populations, but this won't be the focus of this work due to the goal of favoring the size of the relevant dataset to improve the results of the analysis. For the rest of the study, we will work with the overall data but also with the data of specific countries.

2.2 Relevance

Based on the analysis of the two articles analyzed we identified the following features:

Features	**Description**	**Type**
Country	The country where the vehicle was registered or tested	Object
Pool	A designation indicating a group of vehicles or manufacturers for data aggregation or regulatory compliance	Object
Type	The type or category of vehicle (e.g., sedan, SUV), which can affect emission standards and regulations	Object
Commercial name	The name used for marketing and selling the vehicle	Object
Mass in running order (kg)	The weight of the vehicle when it is fully equipped and ready for use, which affects fuel consumption and emissions	Float64
WLTP test mass	The weight of the vehicle used for testing under the Worldwide Harmonized Light Vehicles Test Procedure (WLTP)	Float64
Fuel type	The type of fuel used by the vehicle (e.g., petrol, diesel, electric), impacts emissions and operational costs	Object
Fuel mode	The mode of fuel consumption, such as conventional or hybrid, which affects overall efficiency and emissions	Object
Engine capacity in cm3	The size of the engine, which influences power output and emissions	Float64
Emissions reduction through IT in g/km (WLTP)	The reduction in CO2 emissions achieved through innovative technologies, indicating environmental performance improvements	Float64

Fuel consumption	The amount of fuel consumed per distance traveled, is important for assessing efficiency and operational costs	Float64

The target feature is Specific CO2 Emissions in g/km (WLTP) which represents the amount of CO2 emitted per kilometer under the World Harmonized Light Vehicles Test Procedure.

2.3 Pre-processing and feature engineering

The pre-processing took the following steps :

- First, all the columns with more than 50% of empty rows were removed from the dataset.
- Second, the remaining columns were analyzed based on the state-of-the-art analysis completed using the articles of Zubair et al. and Al-Nefaie et al.
- From the selected features and target, we dropped all the columns with empty rows except the column on the "emissions reduction through IT in g/km(WLTP)." For this column specifically, we replaced all the NaN with 0. The rationale here is that cars without emissions reduction technology will have their cells empty.

The data will be normalized using the min-max normalization. For example, the fuel consumption ranges between 0.1 and 29, while the engine capacity in cm3 ranges between 658 and 7997.

Since the number of features is limited (lower than 15) we are not planning at this stage to do any dimension reduction techniques.

We are investigating what influence innovative technologies have on the CO2 Emission, as it might be a strong indicator for the prediction. A space separates the innovative technologies. For example - "e13 33 37" represents the 3 innovative technologies 13, 33, and 37.

According to the List of eco-innovations approved under NEDC, All NEDC decisions (cars and vans) have been repealed starting in 2021.

Also, according to the List of eco-innovations approved under WLTP, the only eco-innovations approved by WLTP are 28, 37, 29, 32, 33, 35, 38 (28 and 37 are identical).

From exploring the database, we can see that:

- 35 and 38 barely exist
- 29, 32, and 33 don't co-exist together
- 37 (LED lights) can exist in combinations with all the 3

Therefore, we can pre-process the "Innovative Technology" variable into a "has_LED" boolean variable, and an "additional_IT" categorical variable which can only receive the values [32, 33, 37, NONE]. Figures 2-4 show the combination of these technologies and the effect on the total emissions, per Fuel type.

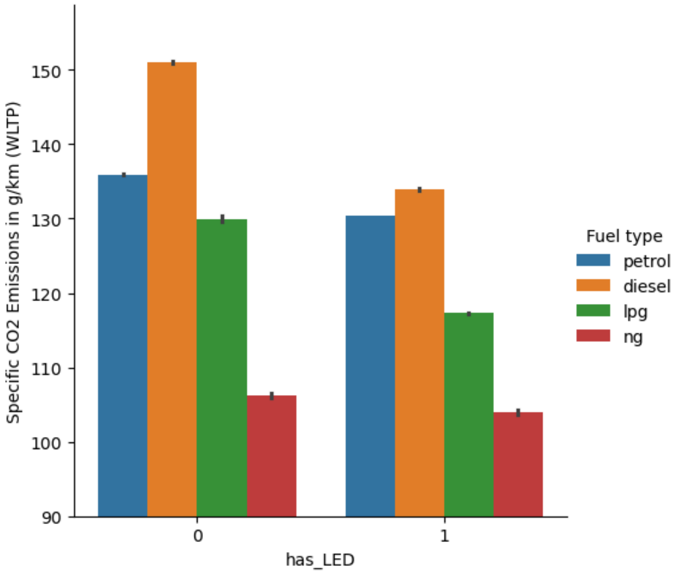


Figure 2: Influence on CO2 Emissions by Fuel type divided by has_LED.

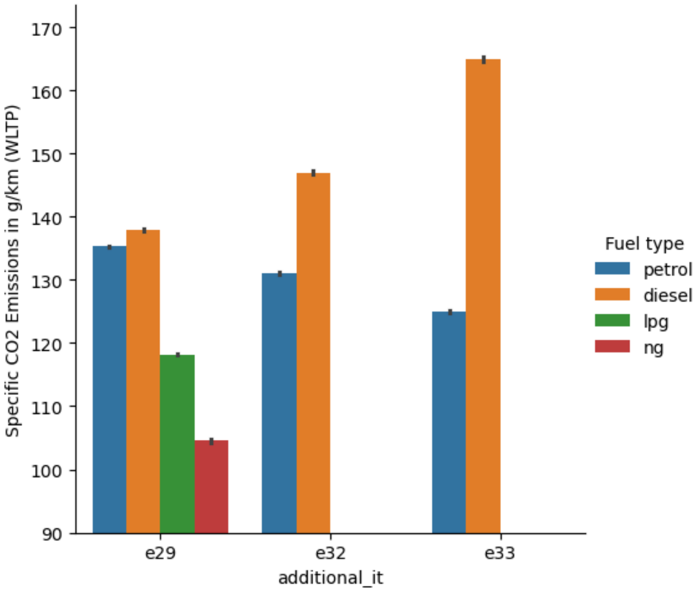


Figure 3: Influence on CO2 Emissions by Fuel type divided by technologies.

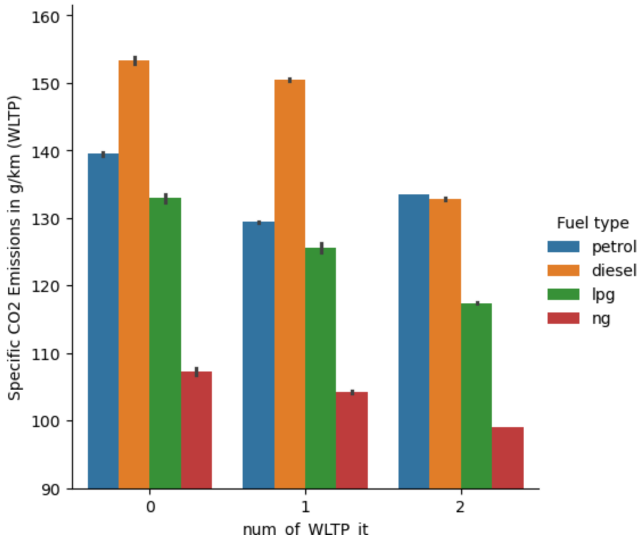


Figure 4: Influence on CO2 Emissions by Fuel type divided by WLTP it.

2.4 Visualizations and Statistics

We did some basic statistical analysis on the target variable to decide if we would work on all data or only on the data of particular countries. Figure 5 and Table 1 present the figure and the parameters of the box plot analysis. Based on the analysis we can drop Italy for the rest of the analysis because it's not adding insight:

- The Mean of Italy is close to that of the overall data while the mean of Germany is higher and that of France is lower
- In the case of Italy, the data is less dispersed, so it does not bring any additional specificity to the analysis that is worth exploring.

Type of data	Size of the data set	Mean	Std	Min	25%	50%	75%	max
Overall	7063310.0	126.761808	39.731316	3.0	116.0	128.0	142.0	461.0
Germany	2140118.0	138.999046	45.440096	10.0	124.0	136.0	155.0	456.0
France	1543750.0	116.502216	32.389164	10.0	111.0	122.0	133.0	456.0
Italy	1360690.0	125.401174	25.366009	10.0	112.0	124.0	135.0	456.0

Table 1: Box plot parameters

Figure 6 shows that the Correlation between Fuel consumption and CO2 Emissions is 95%. As the Correlation is very high it is recommended to not further investigate the Fuel consumption to reduce bias. As the Correlation between WLTP test mass and Mass in running order (kg) is 99%, it is recommended to remove one of the parameters to reduce redundancy in the Dataset.

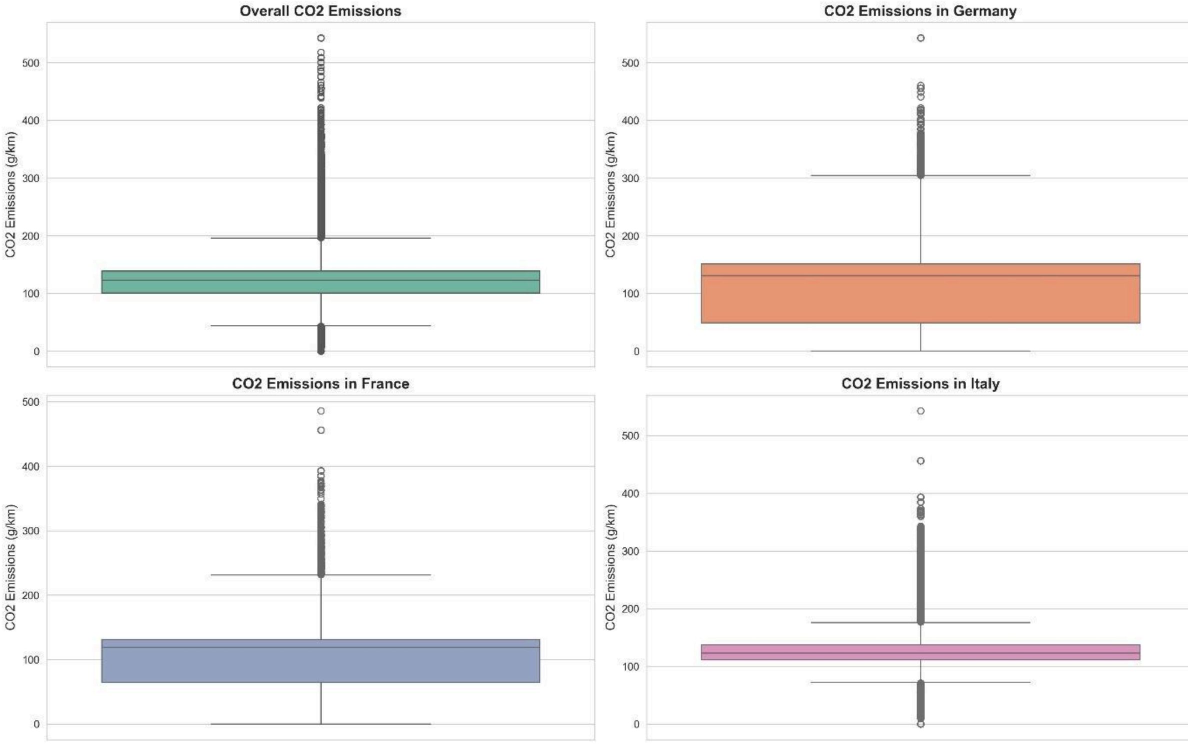
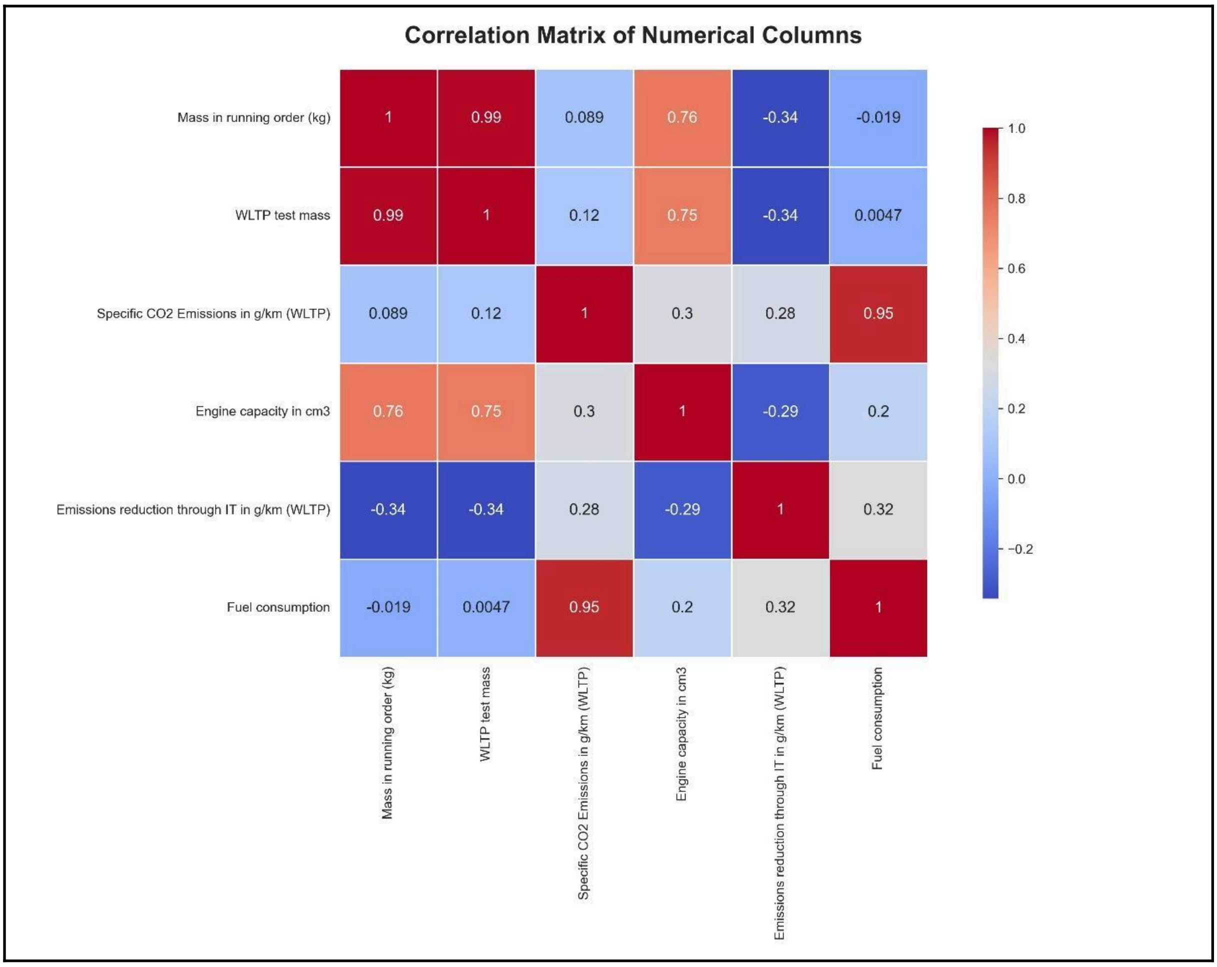


Figure 5: CO2 emissions distributions: All, Germany, France, Italy



	Mass in running order (kg)	WLTP test mass	Specific CO2 Emissions in g/km (WLTP)	Engine capacity in cm3	Emissions reduction through IT in g/km (WLTP)	Fuel consumption
Mass in running order (kg)	1	0.99	0.089	0.76	-0.34	-0.019
WLTP test mass	0.99	1	0.12	0.75	-0.34	0.0047
Specific CO2 Emissions in g/km (WLTP)	0.089	0.12	1	0.3	0.28	0.95
Engine capacity in cm3	0.76	0.75	0.3	1	-0.29	0.2
Emissions reduction through IT in g/km (WLTP)	-0.34	-0.34	0.28	-0.29	1	0.32
Fuel consumption	-0.019	0.0047	0.95	0.2	0.32	1

Figure 6: Correlation between the numerical variables

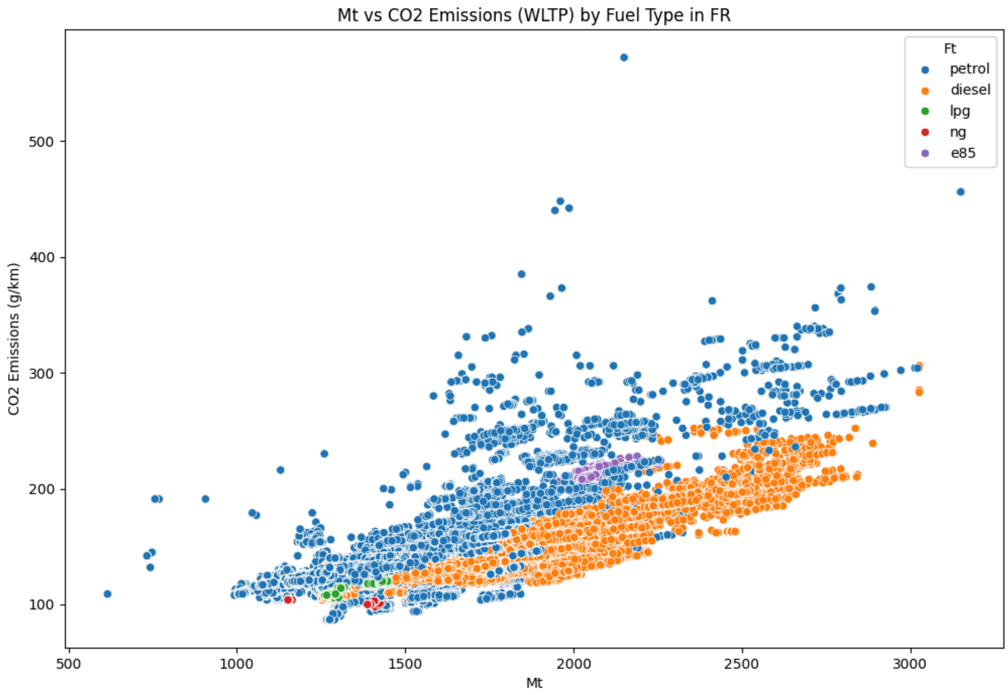


Figure 7: Mt (WLTP test mass) vs CO2 Emissions (WLTP) by Fuel Type in France.

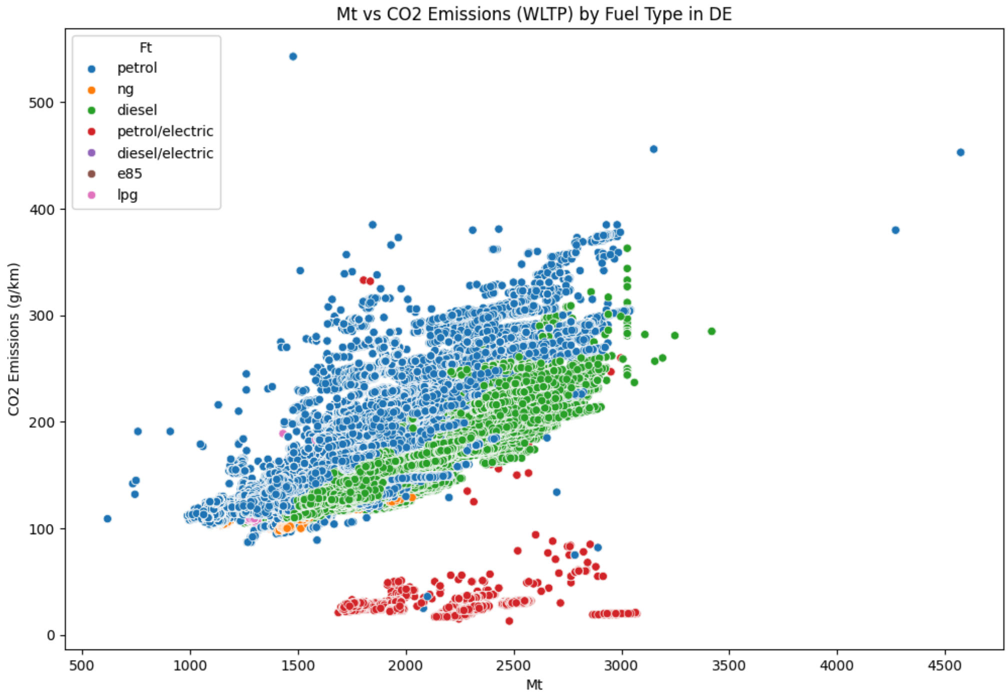


Figure 8: Mt (WLTP test mass) vs CO2 Emissions (WLTP) by Fuel Type in Germany.

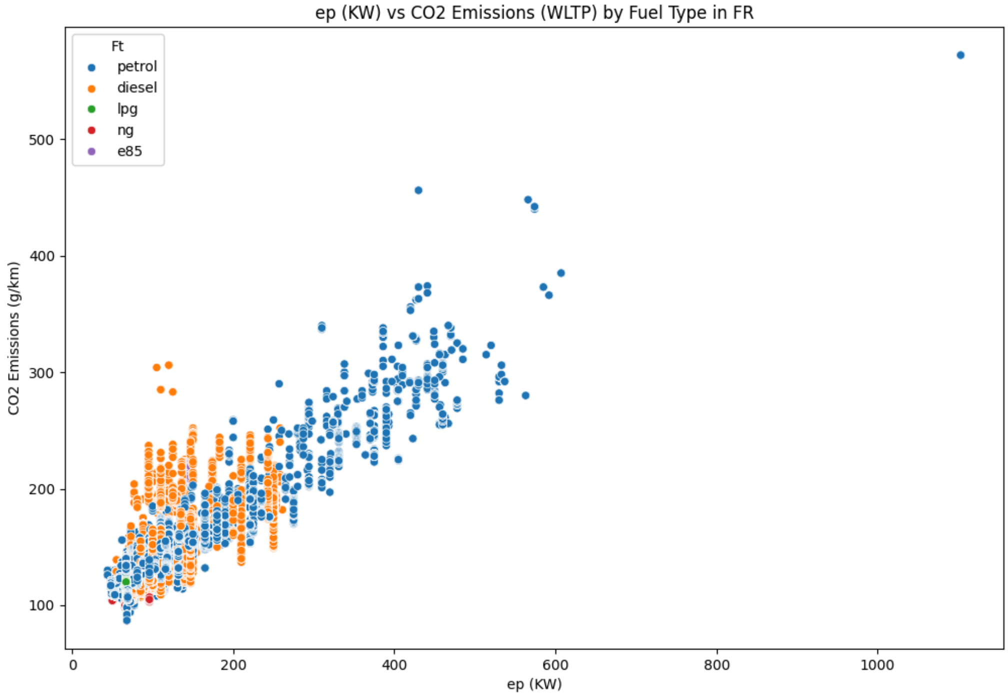


Figure 9: ep (KW) vs CO2 Emissions (WLTP) by Fuel Type in France.

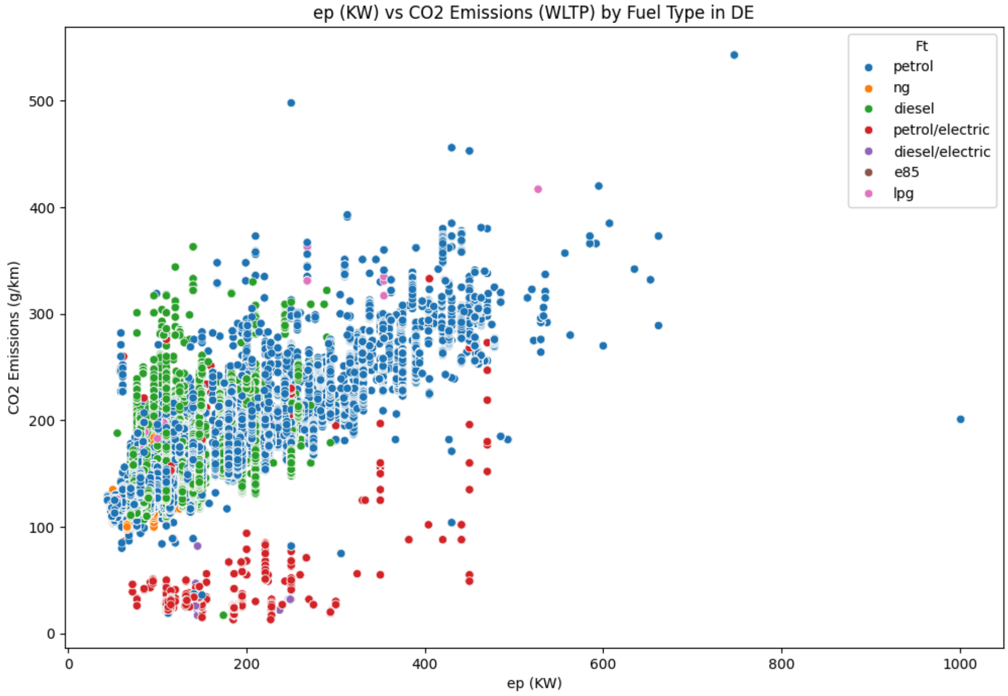


Figure 10: ep (KW) vs CO2 Emissions (WLTP) by Fuel Type in Germany.

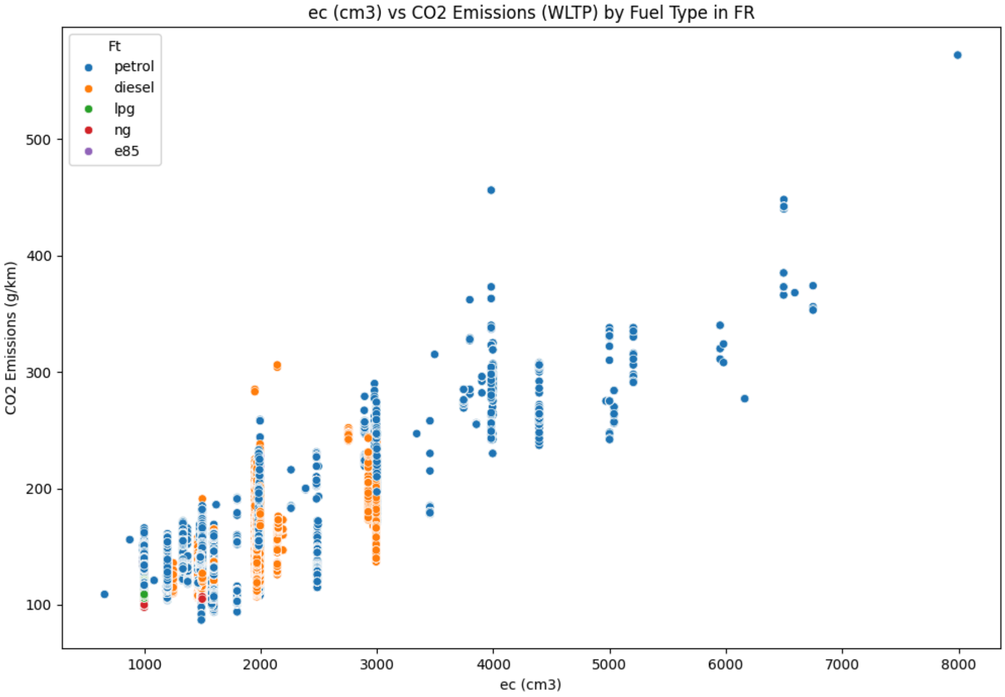


Figure 11: ec (cm3) vs CO2 Emissions (WLTP) by Fuel Type in France.

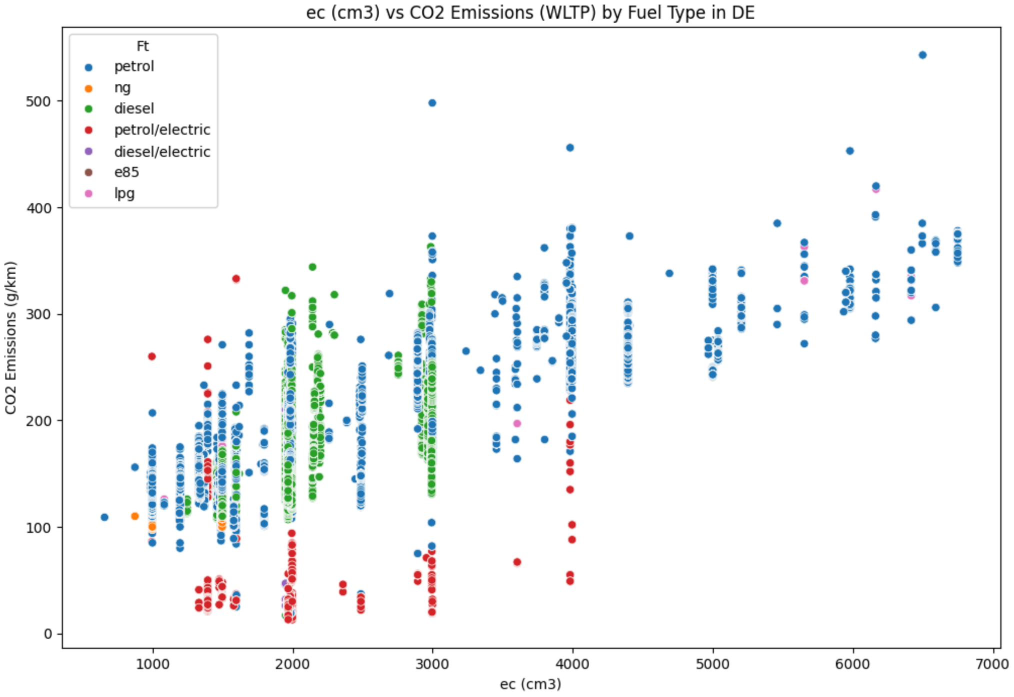


Figure 12: ec (cm3) vs CO2 Emissions (WLTP) by Fuel Type in Germany.

Key Observations:

Overall Trends:

- **Consistent Positive Correlation:** Across all three variables (engine capacity - **Figures 11 and 12**, mass - **Figures 7 and 8**, and engine power - **Figures 9 and 10**) and both countries (DE and FR), there's a clear positive correlation between these variables and CO2 emissions (Ewltp), regardless of fuel type. This indicates that larger, heavier vehicles with more powerful engines tend to emit more CO2.
- **Fuel Type Influence:** While the overall trend is positive, different fuel types exhibit distinct patterns:
    - **Electric and Hybrid:** These vehicles consistently have significantly lower CO2 emissions compared to traditional petrol and diesel options in both countries.
    - **Petrol and Diesel:** There might be slight variations in the distribution of emissions for these fuel types across countries, potentially influenced by factors like driving habits, regulations, and vehicle availability.
    - **LPG and e85:** The performance of these fuel types can vary across countries, depending on factors such as infrastructure, availability, and government incentives.

Specific Relationships:

- **Engine Capacity (ec) vs. CO2 Emissions:** As engine capacity increases, CO2 emissions generally rise, regardless of fuel type. However, the rate of increase might vary slightly between different fuel types.
- **Vehicle Mass (Mt) vs. CO2 Emissions:** Heavier vehicles tend to emit more CO2, regardless of fuel type. The relationship between mass and emissions might be more pronounced for certain fuel types.
- **Engine Power (ep) vs. CO2 Emissions:** Higher engine power is generally associated with higher CO2 emissions, but the impact might vary depending on factors like engine efficiency and vehicle design.

Clustering and Outliers:

- The data points are clustered in certain regions within each fuel type, indicating that there might be groups of vehicles with similar characteristics.
- Outliers are visible in all countries, particularly for some petrol and diesel vehicles. These might represent vehicles with less efficient engines or other factors that influence emissions.

Potential Insights:

- **Vehicle Design and Efficiency:** The consistent positive correlation between these variables and CO2 emissions highlights the importance of vehicle design and efficiency in reducing emissions. Lighter, more efficient vehicles with lower engine capacities and power outputs can significantly lower CO2 emissions.
- **Fuel Type Impact:** Electric and hybrid vehicles offer a significant advantage in terms of CO2 emissions compared to traditional petrol and diesel options. However, the performance of LPG and e85 varies, suggesting that further research is needed to understand their environmental impact.

- **Technological Advancements:** The presence of outliers, especially in the petrol and diesel categories, indicates that there might be technological advancements or specific vehicle designs that can improve fuel efficiency and reduce emissions.
- **Country-Specific Factors:** While the overall trends are similar, there might be country-specific factors influencing the distribution of CO2 emissions within each fuel type. These factors could include regulations, driving habits, and the availability of different vehicle models.

Further Analysis:

- **Multivariate Analysis:** Explore the relationships between multiple variables simultaneously using techniques like principal component analysis (PCA) or multiple regression analysis.
- **Time Series Analysis:** Having data over time, analyze trends and changes in CO2 emissions for different vehicle types and countries.
- **Machine Learning:** Develop machine learning models to predict CO2 emissions based on various vehicle characteristics and fuel types.

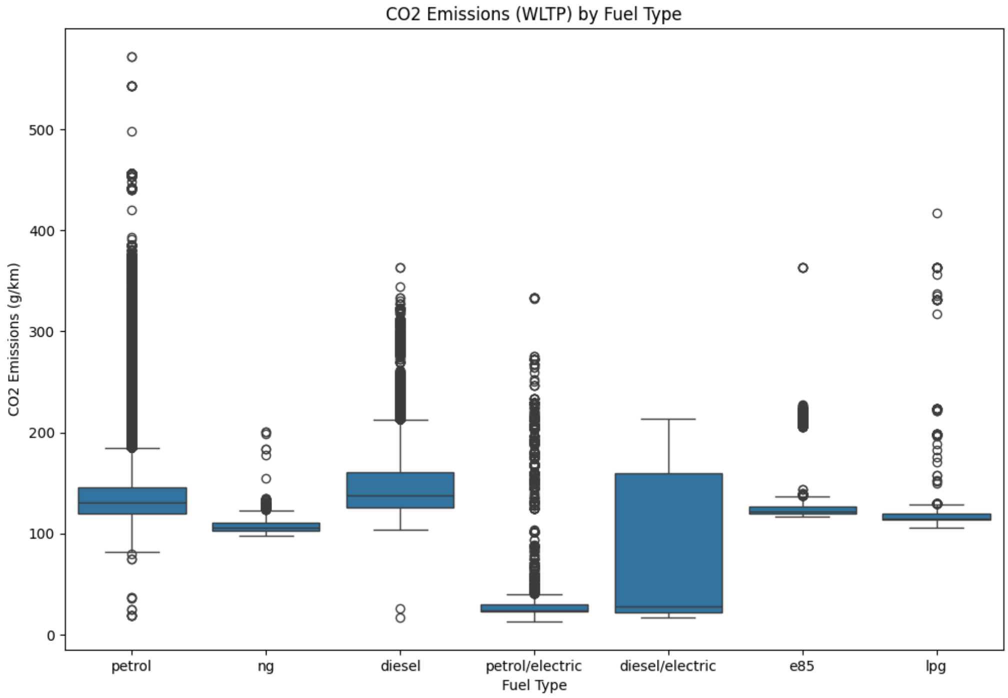
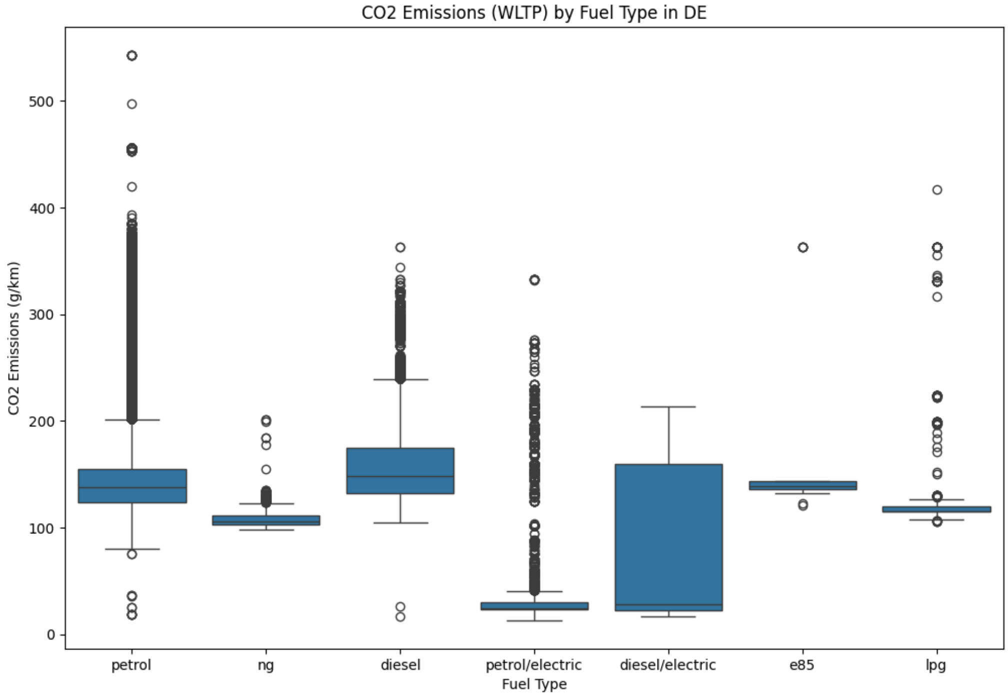
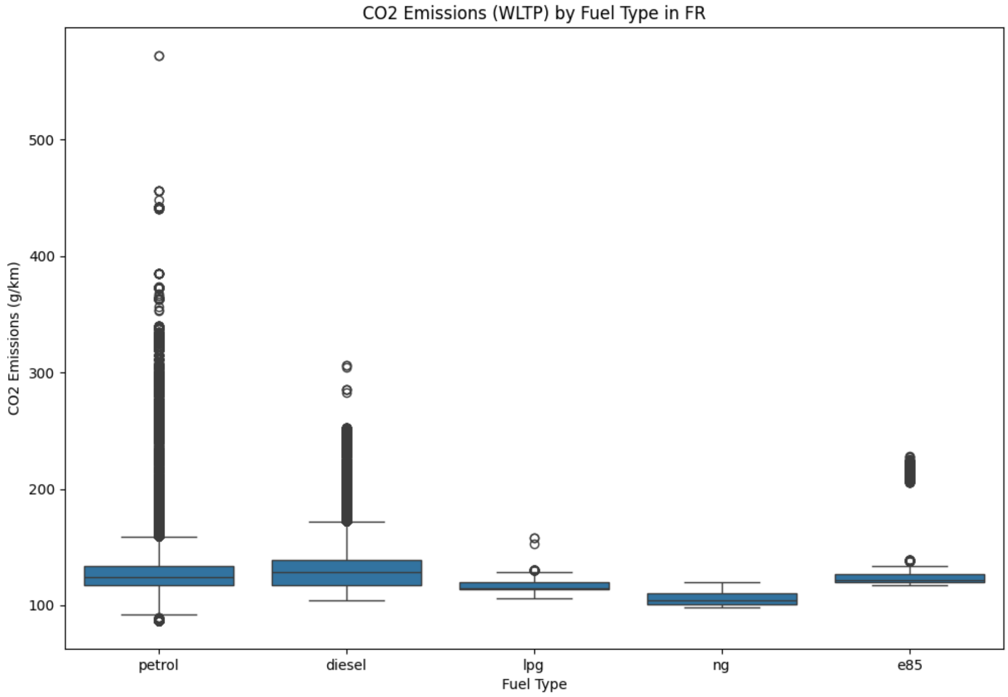


Figure 13: CO2 Emissions (WLTP) by Fuel Type in Germany and France.



<u>Figure 14</u>: CO2 Emissions (WLTP) by Fuel Type in Germany.



<u>Figure 15</u>: CO2 Emissions (WLTP) by Fuel Type in France.

Key Findings:

- **Electric vehicles consistently have lower CO2 emissions:** Across all three datasets (**Figures 13, 14, and 15**), electric vehicles (petrol/electric, diesel/electric, and e85) demonstrate significantly lower emissions compared to traditional fuel types.
- **LPG is a more eco-friendly option:** LPG (lpg) exhibits consistently lower emissions than petrol, diesel, and CNG (ng).
- **CNG has a wider range of emissions:** CNG shows a greater variability in emissions compared to other fuel types, with a higher median and a larger interquartile range.
- **Petrol and diesel have similar emissions:** These traditional fuel types exhibit comparable emission distributions.

General Trends:

- **Vehicle age and technology:** Newer vehicles, especially electric vehicles, tend to have lower emissions due to advancements in technology.
- **Driving conditions:** Factors like traffic congestion, driving style, and road conditions can influence fuel consumption and emissions.
- **Government policies:** Government incentives and regulations can impact the adoption of different fuel types and the overall emissions landscape.

Implications for Policy and Future Research:

- **Promote electric vehicles and LPG:** Governments should continue to incentivize the adoption of electric vehicles and LPG to reduce CO2 emissions.
- **Consider additional factors:** Factors like vehicle weight, engine size, and driving distance should be explored to gain a more comprehensive understanding of CO2 emissions.
- **Monitor emissions over time:** Ongoing monitoring of CO2 emissions is essential to track progress and inform future policies.

Overall, the analysis suggests that electric vehicles and LPG are promising options for reducing CO2 emissions in both Germany and France. However, it's important to address other factors that influence fuel consumption and emissions. By implementing effective policies and promoting cleaner technologies, we can make significant strides towards a more sustainable transportation sector.