1º report: Predicting Car CO_2 Emissions through Machine Learning

Phillipp Würfel
Python Backend Engineer | Software Engineer
Karl Richard Georg Kutz
Mechanical Engineer
Leonel Leite Barros
Master in Economic Theory
Thomas Igor Lisowsky
Junior Web Developer

Abstract
This report presents a preliminary analysis within the *"Predicting Car CO_2 Emissions through Machine Learning"* project. Its primary objective is to provide an initial overview of the dataset exploration and the progress made so far. The report contextualizes the project, outlines its purpose and objectives, and highlights its framework and significance. Additionally, it introduces initial data visualizations, including tables and graphs that illustrate correlations between key variables. Finally, it summarizes the next steps in the project's development.

1. Introduction to the Project

Growing environmental obligations necessitate advanced, data-driven strategies for estimating vehicle CO_2 emissions. This study leverages large-scale data from the European Environment Agency (EEA) for the years 2021–2023 to develop machine learning models for both combustion and electric vehicles[1]. These models aim to improve the prediction of CO_2 emissions from newly introduced cars in the market.

Accurate emission prediction informs manufacturers' strategic decisions on engine design, material selection, and marketing compliance. For governments, transparent modeling supports setting feasible targets while identifying which powertrain types merit further incentives or regulations.

Context

[1] At first was thought the project would handle with two complimentary datasets, the above named European data and a dataset that handle with French market. Although the analysis of the second dataset showed that the approach would be to discard this data. There are a special section where the decision is make clear with comments about the data.

Context of the project's integration into business

Predicting CO_2 emissions to help car manufacturers and stakeholders design vehicles in alignment with official and governmental environmental standards, supporting decisions that can enhance brand image and meet consumer demand for eco-friendly solutions.

From a Technical Point of View

We use the Python programming language and the scikit-learn library to train machine learning models. Our data comprises a variety of officially available CSV files, which we split into training and testing sets.

From an Economic Point of View

Predicting CO_2 emissions can facilitate the creation of more fuel-efficient or emission-reduced cars, meeting official regulations and offering an early impression of emissions during the concept phase. It also helps strengthen the image of manufacturers among consumers who prioritize environmental concerns.

From a Scientific Point of View

This project contributes to understanding and mitigating the environmental impacts of automotive emissions and fuels similar scientific research on a broader scale.

2. Objectives

<u>Main Objectives:</u>

- Use machine learning to predict the CO_2 emissions of cars by focusing on two specialized models: one for combustion engines (fuel consumption) and one for electric vehicles (energy consumption).
- Investigate formulas to calculate CO_2 emissions from the predicted consumption values.

<u>Expertise of Team Members:</u>

- *Philipp Würfel e*xperienced Python backend developer and engineering manager with background in SaaS solutions related to sustainable supply chains. In his previous jobs he used to work in startups and enterprises as founder, team & tech lead with focus on managing people in Backend, Data Engineering, QA, Ops, and Data Science. Responsible for team hiring and culture building and holding strategic and technical responsibility for SDM (Sustainability Data Mining) topics.
- *Karl Richard Georg Kutz* has his background in mechanical component testing in the automobile industry. Basic programming knowledge exists in Visual Basic and Matlab, which pretty much resembles python. One relevant project was the programming and maintenance of an evaluation routine for endurance-test measurements from complete cars or combustion engines (for BMW) in matlab.
- *Leonel Leite Barros* has an academic background with expertise in Macroeconomics, where statistical analysis and data interpretation are fundamental tools. His experience extends to working with theoretical models, as well as handling data exploration and econometric modeling. During his university studies, he conducted various data-driven projects, applying econometric techniques and working with E-Views 8 to develop and test economic models. His skill set includes structuring and interpreting data, which serves as a foundational aspect of his macroeconomic research.
- *Thomas Igor Lisowsky* has no experience with automotive or emissions topics, but some web development background that helped him with understanding Python faster. He has no Science-based academic background, but recently started teaching himself web-based programming.

Business Experts

No direct consultations were made to refine the models; the work relies on publicly accessible data sources and official documentation.

Similar Projects

No known similar in-house projects have influenced or supported this one.

3. Methods

Data Description

- EEA Data (2021–2023): More than 30 million records, including vehicle mass, engine power, fuel consumption, and electric energy consumption.
- French Data (2014): Roughly 55,000 rows but excluded due to ~30% missingness in critical fields like mass_vehicle.

Preprocessing & Feature Engineering

- Missing Data Threshold: Variables with >70% missing values were removed ; minor gaps were imputed.
- Feature Selection & Encoding: Key variables were retained; categorical fields (fuel_type, category_of_vehicle) were one-hot encoded; numerical features scaled.

- Train/Test Split: 80/20 train-test split with 5-fold cross-validation for robustness.

3.3 Modeling

- Combustion Vehicles: Random Forest and Gradient Boosting on fuel_consumption (l/100 km). CO_2 derived from standard factors (~2.32 kg/liter petrol, 2.65 kg/liter diesel).
- Electric Vehicles: Gradient Boosting on electric_energy_consumption (Wh/km), multiplied by region-specific gCO_2/kWh to estimate CO_2e.
- Evaluation: R^2, RMSE, plus sensitivity analyses comparing various regions' emission factors.

4. Research on Calculating CO_2/CO_2e

CO_2e Based on Energy Consumption

External resources include the World Bank's World Development Indicators, the European Environment Agency (EEA) emission data, CaDI (The Carbon Database), and the European Commission's GHG Emission Factors for Electricity Consumption per country.
For electric vehicles, we rely on emission factors that vary by regional energy production (renewables vs. fossil fuels), making a universal conversion factor challenging.
CO_2 Based on Fuel Consumption

We plan to look for scientific formulas that map liters of fuel consumed to grams of CO_2.
Example factors:
1 liter of petrol produces roughly 2,320 g CO_2
1 liter of diesel produces roughly 2,650 g CO_2

5. Data Visualization

Fuel consumption and electric energy consumption will be our targets to estimate CO2 emissions.
We analyzed specific CO2 emissions and its distribution based on fuel type. For electric and hydrogen cars we do not have values and would need a different metric to calculate CO2 emissions.

Descriptive tables
The following tables describe the variables and some found correlations between them after the preprocessing and cleaning of the raw data.

Data Cleaning and Processing

Variable	Percentage Present	Observations
MMS	100.00%	
Enedc (g/km)	100.00%	
W (mm)	100.00%	
At1 (mm)	100.00%	
At2 (mm)	100.00%	
Ernedc (g/km)	100.00%	
De	100.00%	
Vf	100.00%	
RLFI	71.34%	
Electric range (km)	77.38%	
Electric energy consumption (Wh/km)	Differential treatment (Not removed)	Threshold >70% but retained

- *Electric energy consumption* was retained despite missing values above 70%.

Selected Variables for Modeling

Variable	Data Type	Comments
category_of_vehicle	Categorical	Encoded
mass_vehicle	Numerical	Feature
fuel_type	Categorical	Encoded
engine_capacity	Numerical	Feature
engine_power	Numerical	Feature
year	Numerical	Feature
electric_energy_consumption	Numerical	Target for EVs
fuel_consumption	Numerical	Target for combustion engines
specific_co2_emissions	Numerical	Feature

- Only the most relevant variables were kept to reduce noise.
- *Fuel type*, *category of vehicle*, and *fuel mode* were encoded.

Correlation Table for Fuel Consumption

Variable	Correlation
fuel_type_petrol	0.3326
engine_power	0.2893
fuel_type_lpg	0.1967
engine_capacity	0.1678
category_of_vehicle_M1G	0.0549
fuel_type_diesel	0.0498
fuel_type_e85	0.0116
fuel_type_ng	0.0086
category_of_vehicle_N1G	0.0007
category_of_vehicle_M1	-0.0549
mass_vehicle	-0.0643
fuel_type_diesel/electric	-0.1426
fuel_type_petrol/electric	-0.7231
year	(missing)

- *Fuel type (petrol)* and *engine power* show the strongest positive correlations.

- *Fuel type (petrol/electric)* has the highest negative correlation (-0.72), indicating lower fuel consumption in hybrid models.

Correlation Table for Electric Energy Consumption

Variable	Correlation
mass_vehicle	0.6371
engine_power	0.3633
category_of_vehicle_M1G	0.0444
category_of_vehicle_M1	-0.0444
year	(missing)

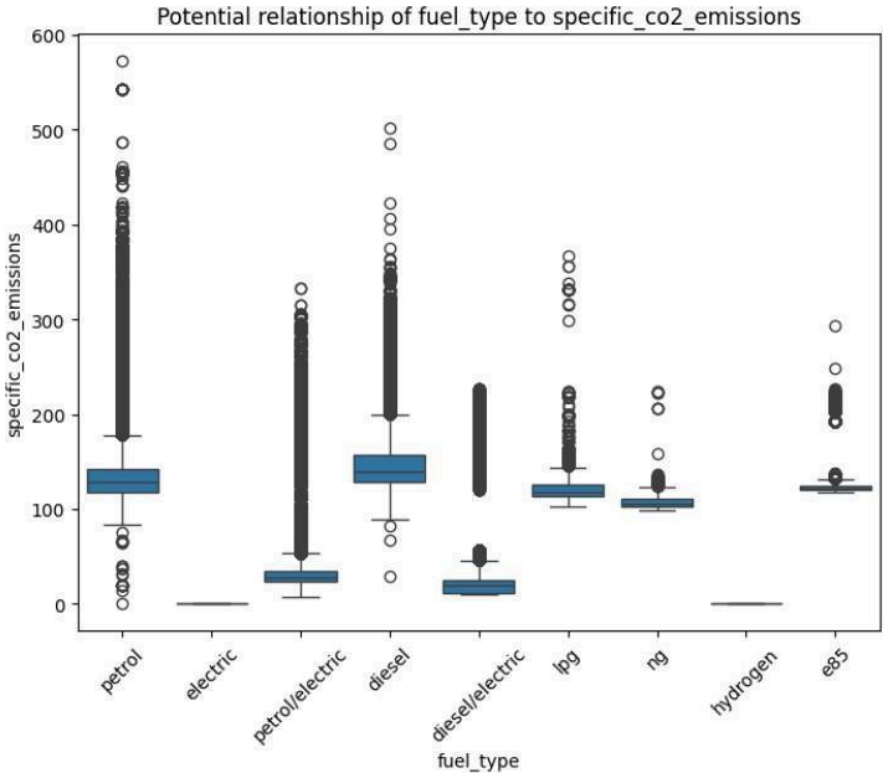
- *Mass vehicle* (0.64) has the strongest correlation with *electric energy consumption*, suggesting heavier vehicles consume more energy.
- *Year* appears to have minimal impact.

Fuel Consumption and CO2 Emissions

Fuel Type	Conversion Factor	CO_2 Emitted per Liter (g)
Petrol	23.2	2320
Diesel	26.5	2650

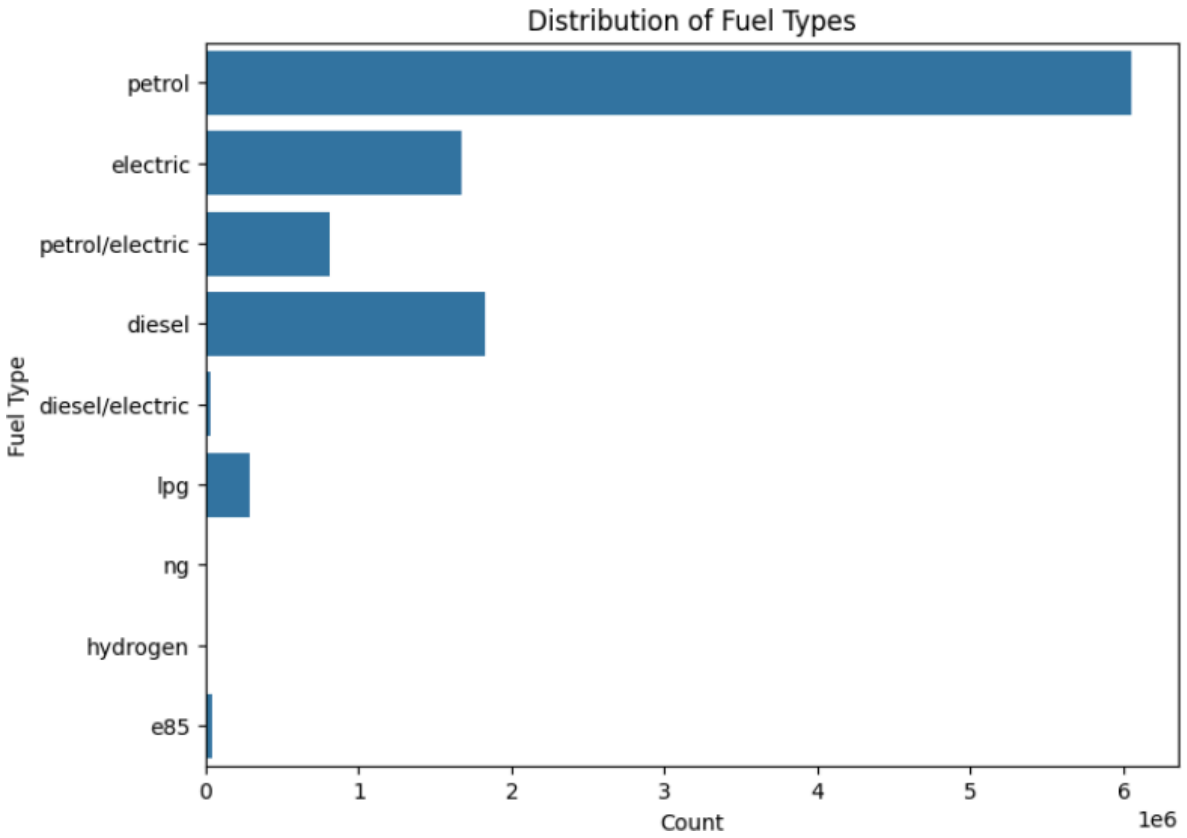
Graphs

Potential relationship of "fuel_type" to "specific_co2_emissions"



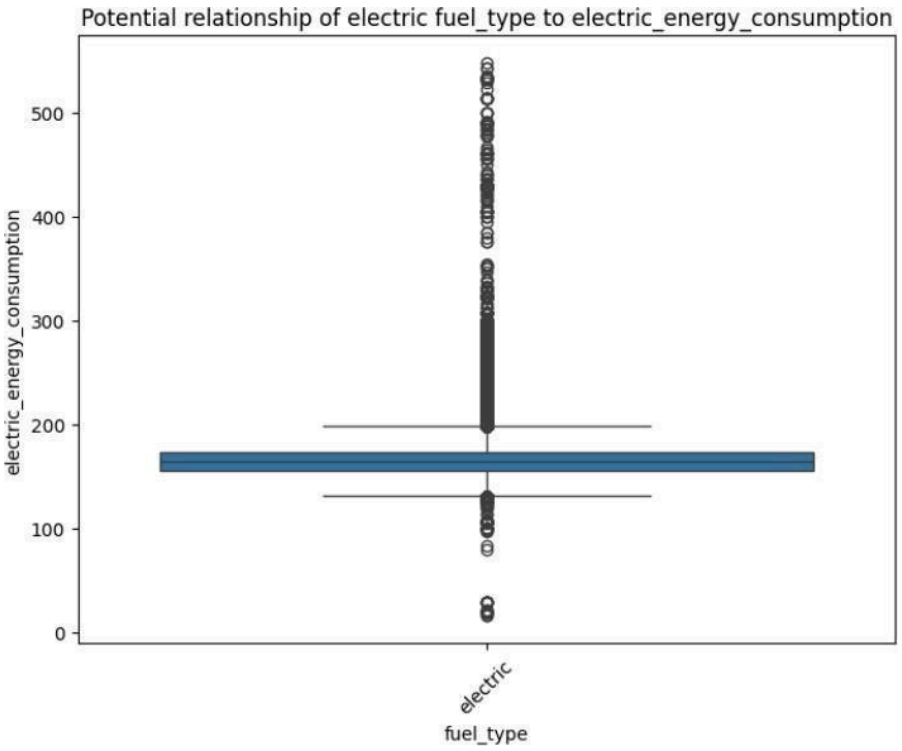
The graph above shows the relationship between fuel type and specific CO_2 emissions. Electric and hydrogen vehicles have near-zero emissions, while petrol and diesel cars exhibit the highest median CO_2 outputs, with significant variability and outliers. Hybrid (petrol/electric, diesel/electric) and alternative fuel vehicles (LPG, NG, E85) show moderate emissions, depending on energy mix and driving conditions. This highlights the environmental advantage of electrification and the impact of fuel choice on emissions reduction.

Distribution of Fuel Types



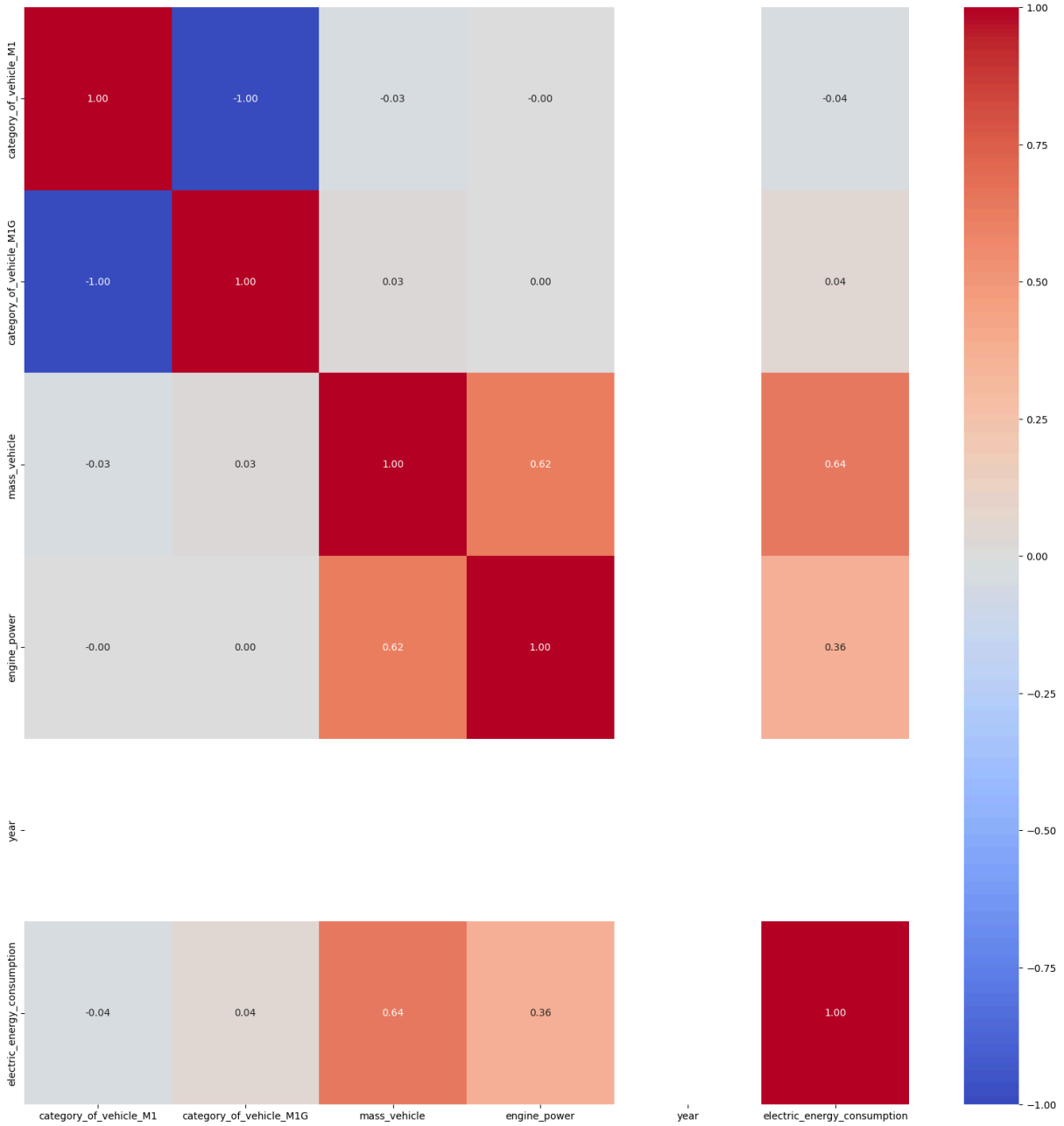
The graph above shows a horizontal bar chart shows with the distribution of vehicles by fuel type. The majority of vehicles use petrol (gasoline), followed by diesel and then electric vehicles. Other fuel types, such as petrol/electric and diesel/electric, appear in smaller numbers. Alternative fuels like LPG, natural gas (NG), hydrogen, and E85 have very low representation in the dataset.

Potential relationship of electric "fuel_type" to "electric_energy_comsuption"



The graph above shows the distribution of electric energy consumption (Wh/km) for electric vehicles. Most values are concentrated around a narrow range, indicating a consistent energy efficiency among electric cars. However, there are many outliers, especially on the higher end, suggesting that some vehicles consume significantly more energy per kilometer, likely due to larger battery capacities, higher performance, or inefficiencies in certain models. The few lower outliers may represent exceptionally efficient vehicles. This highlights the variation in energy efficiency across different electric models.

Heatmap from the "corr_matrix_electric"



Key Insights:

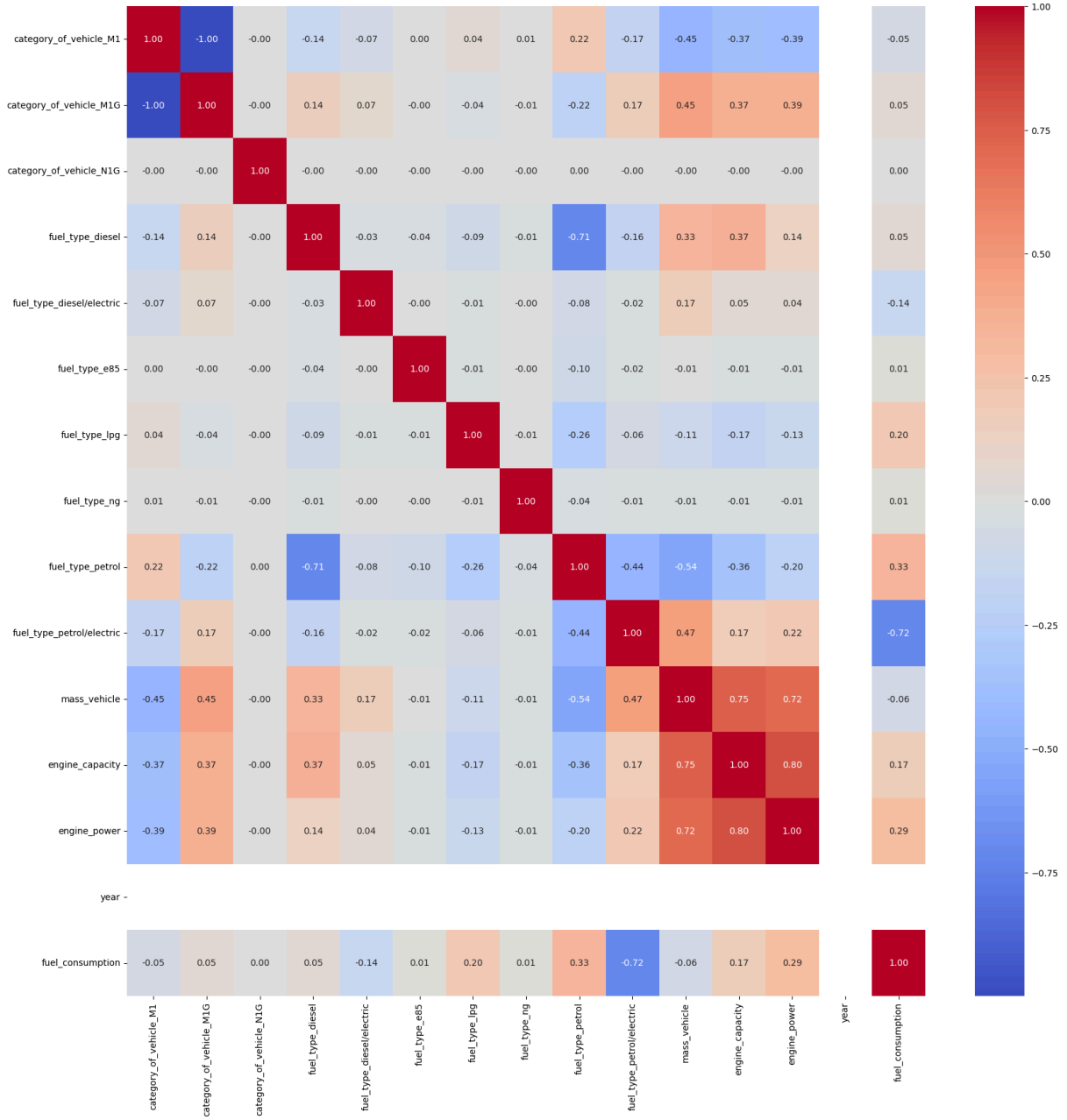
- ***Mass Vehicle & Energy Consumption:***
  - *Mass vehicle has a **strong positive correlation (0.64) with electric energy consumption**, indicating heavier electric vehicles consume more energy.*
- ***Engine Power Impact:***
  - *Engine power shows a **moderate positive correlation (0.36) with electric energy consumption**, suggesting that more powerful electric vehicles require more energy.*

- ***Vehicle Category:***
    - *Category_of_vehicle_M1 and M1G are **perfectly negatively correlated (-1.00)**, which is expected since they are mutually exclusive classifications.*
    - *These categories show **negligible correlation with electric energy consumption**, meaning classification alone is not a strong predictor.*
- ***Year Influence:***
    - *Year shows **little to no correlation** with energy consumption, implying that technological advancements or efficiency gains are not strongly reflected in the dataset.*

Overall Interpretation:

- ***Vehicle mass** is the **strongest predictor** of electric energy consumption.*
- ***Engine power also contributes**, but to a lesser extent.*
- ***Vehicle category and year have minimal influence**, suggesting other factors may play a bigger role in EV efficiency.*

Heatmap from the "corr_matrix_combustion"



Key Observations:

- Fuel Type Impact:
  - *Fuel type (petrol/electric)* shows a strong negative correlation (-0.72) with fuel consumption, indicating hybrids consume significantly less fuel.
  - *Fuel type (petrol)* has a moderate positive correlation (0.33) with fuel consumption, meaning petrol cars tend to consume more fuel.

- Vehicle Characteristics:
    - *Mass vehicle* correlates positively with engine capacity (0.75) and engine power (0.72), showing that heavier vehicles tend to have larger and more powerful engines.
    - *Engine power* has a moderate positive correlation (0.29) with fuel consumption, suggesting higher power leads to increased fuel usage.
- Category of Vehicle:
    - Some vehicle categories (e.g., M1 and M1G) show negative correlations among themselves, indicating distinct classifications.
    - *Category_of_vehicle_M1* has a moderate negative correlation (-0.39) with fuel consumption, suggesting these vehicles might be more fuel-efficient.

Overall Interpretation:

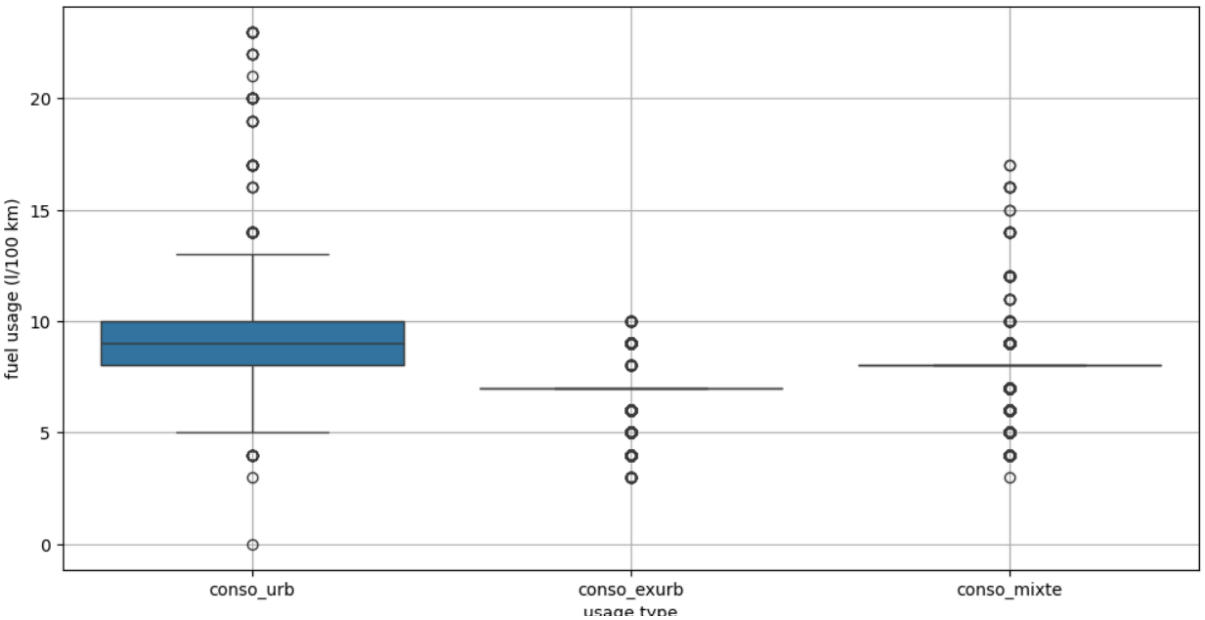
- Hybrid vehicles (petrol/electric) are associated with lower fuel consumption.
- Mass and engine power influence fuel consumption positively, meaning heavier and more powerful vehicles tend to use more fuel.
- There are strong intra-variable relationships among vehicle mass, engine power, and engine capacity.

6. French Data specifics

Most of the aspects discussed earlier are also present in the French dataset, but this section highlights additional insights.

A key focus was fuel usage, which in the French data is categorized into urban, ex-urban, and mixed conditions. The analysis aimed to determine whether mixed usage (conso_mixte) serves as a reliable average of the other two. The graph reveals that urban consumption has the highest variability and many outliers, while ex-urban consumption is more stable. The distribution of mixed usage appears well-balanced, both in spread and outliers, supporting its validity as a representative measure of fuel consumption.

Fuel Consumption Distribution Across Driving Conditions

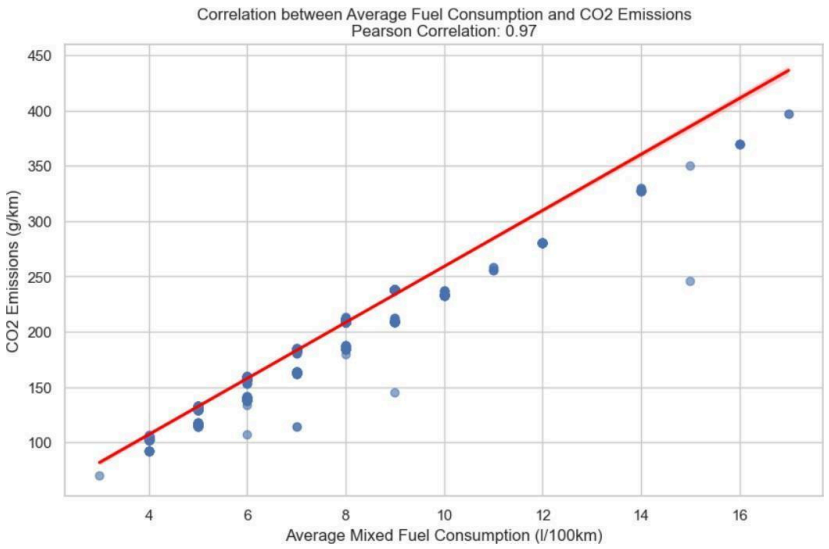


The graph above analyzes fuel usage in the French dataset, divided into urban, ex-urban, and mixed conditions, aiming to verify whether mixed usage (conso_mixte) serves as a reliable mean value of the other two.

- Urban consumption (conso_urb) shows the highest variability, with numerous outliers.
- Exurban consumption (conso_exurb) is more stable, with fewer outliers.
- Mixed consumption (conso_mixte) appears well-balanced, both in distribution and outliers.
- Conclusion: *Conso_mixte* is a reliable representation of overall fuel consumption.

We now analyze the relationship between average fuel consumption and CO_2 emissions to confirm a strong correlation.
As the following graph shows, fuel consumption is indeed highly correlated with CO_2 emissions.

Linear Correlation Between Fuel Consumption and CO_2 Emissions



The graph above illustrates the strong linear relationship between average mixed fuel consumption (l/100 km) and CO_2 emissions (g/km), with a Pearson correlation of 0.97.

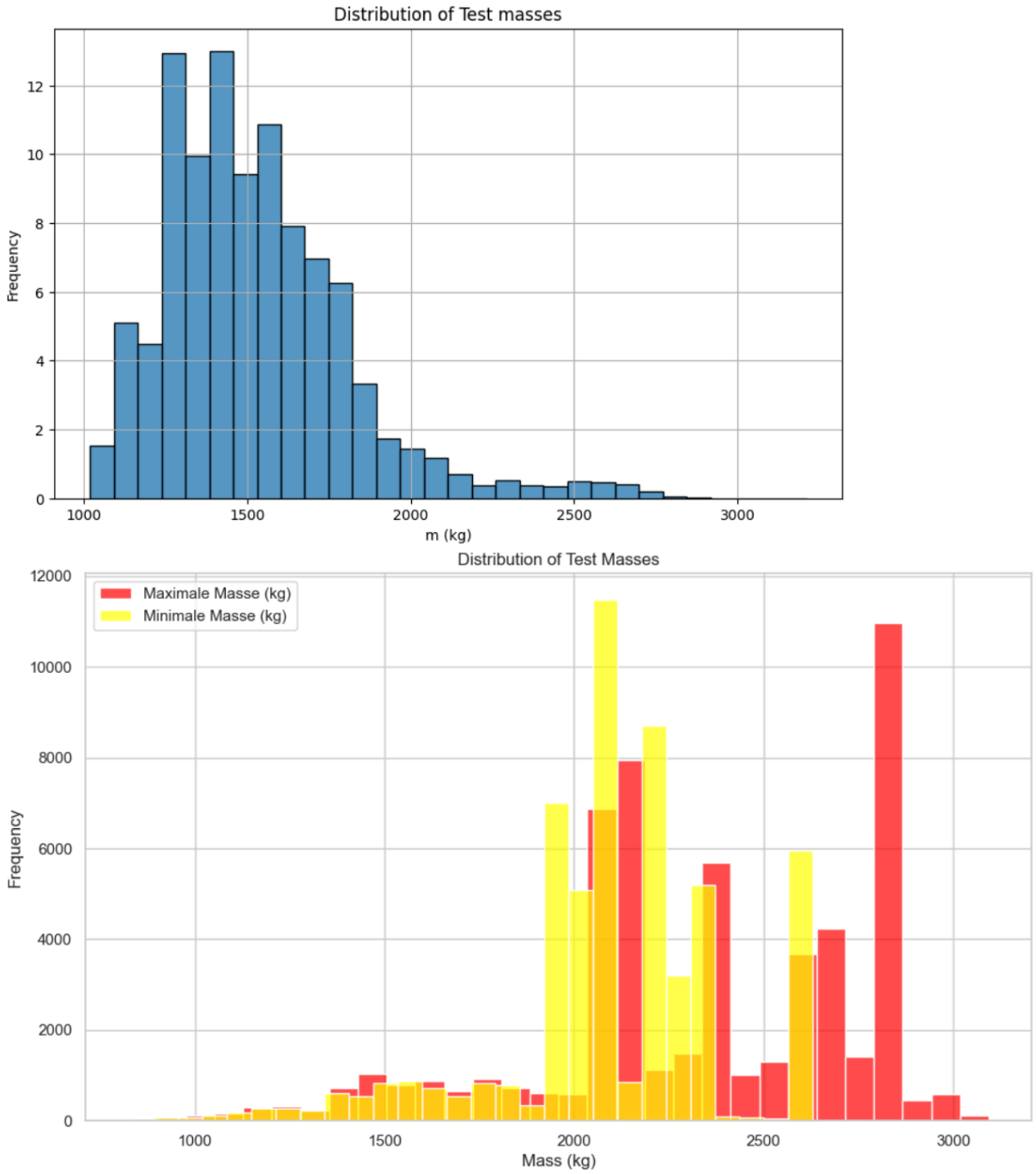
Key Insights:

- Higher fuel consumption directly increases CO_2 emissions, confirming their strong dependency.
- The linear trend (red regression line) suggests a nearly proportional relationship.
- Minimal dispersion around the trend line indicates high predictive reliability between the two variables.
- Fuel consumption is a highly effective predictor of CO_2 emissions, reinforcing the expected link between combustion and carbon output.

To reinforce the statistical analysis, the Pearson correlation between fuel consumption and CO_2 emissions is 0.97, indicating that both variables exhibit almost identical behavior.

Next, we present another graphical comparison, analyzing the test masses from the current EU dataset and the latest French dataset (2014) to assess their compatibility.

Comparison of Test Mass Distributions in EU and French Datasets



From the graph above we can observe that:

- The top histogram shows the overall distribution of test masses, with a peak around 1,400–1,600 kg and a right-skewed tail extending beyond 2,500 kg.
- The bottom histogram compares maximum (red) and minimum (yellow) test masses in the datasets:
    - Distinct peaks suggest variability in mass classification, with higher maximum values exceeding 3,000 kg.
- Minimum masses are more concentrated, mostly between 1,500–2,200 kg.
- The test mass distributions indicate compatibility between datasets, but differences in weight classification suggest variations in vehicle types or regulatory standards.

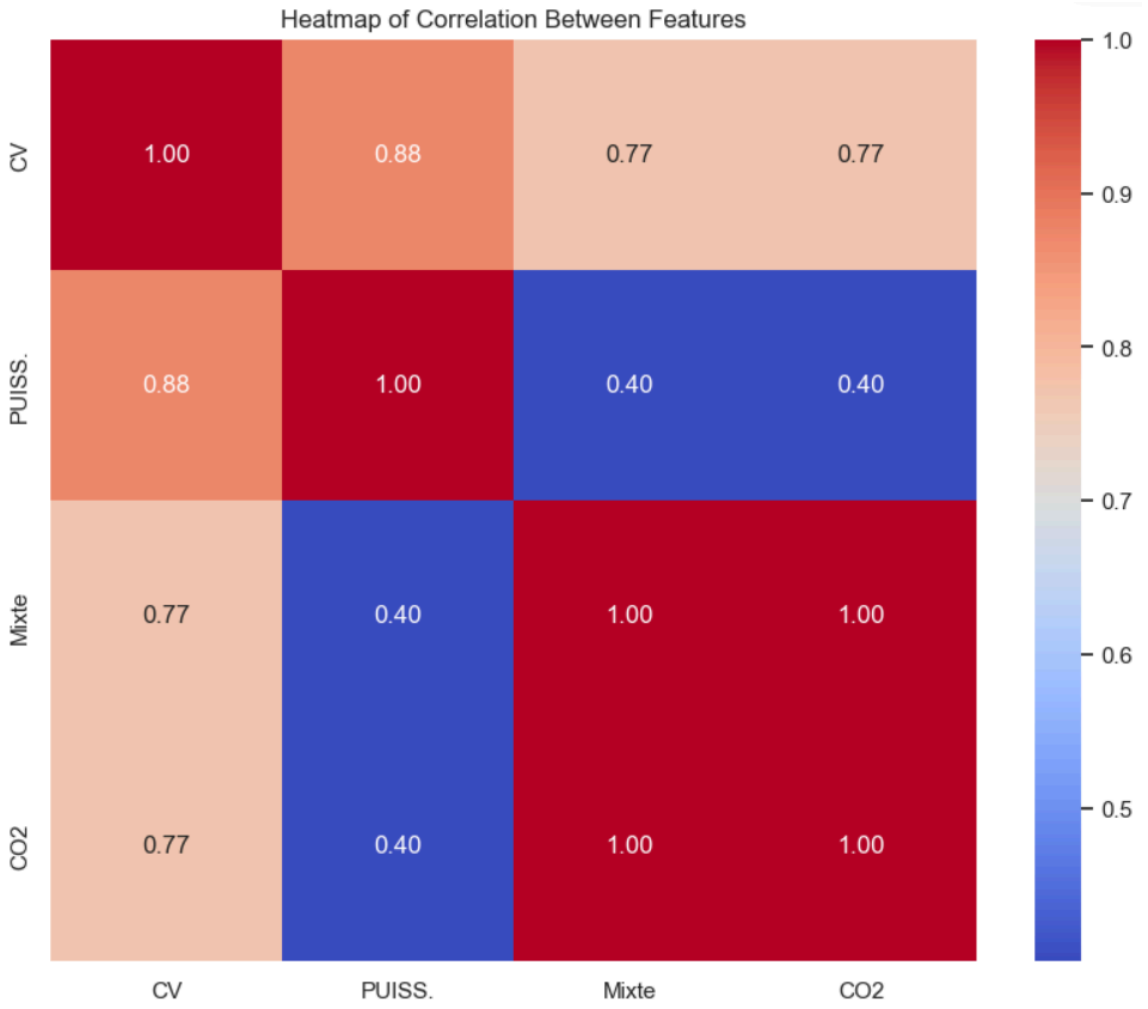
Despite the graphs showing that the french data does indeed look generally valid and comparable to EU data, we ultimately decided not to use the french data.
The biggest argument to use it was that it provided data from before 2010, where the EU data did not provide any. In this specific time period though, the French data is far more limited, e.g. doesn't include vehicle mass at all, one of our most important explanatory variables.
We concluded that the data from before 2010 thus was less usable to us, and the data from after 2010 would just produce data noise in our dataset.

As an additional visualisation for the difference/incompatibility between the recent EU data and the past French data, here is a simple head() output of pre-2010 french datasets:

	MARQUES	MODELE	TYP. MINES	CNIT	CARB	CV	PUISS.	BV	Urbain	Ex.Urb	Mixte	CO2
0	VOLKSWAGEN	LUPO 3L TDI	MVW70C1R4385	6ESCANYX01AGFD5850021N0H	GO	3	45.0	A 5	3.6	2.7	3.0	81
1	VOLKSWAGEN	LUPO 3L TDI	MVW70C1RX572	6ESCAYZX01AGFD5850021N0H	GO	3	45.0	A 5	3.6	2.7	3.0	81
2	VOLKSWAGEN	LUPO 3L TDI	MVW70C1R5386	6ESCANYX01AGFD5850021N0I	GO	3	45.0	A 5	3.8	2.8	3.2	86
3	VOLKSWAGEN	LUPO 3L TDI	MVW70C1R6387	6ESCANYX01AGFD5850021N1H	GO	3	45.0	A 5	4.0	2.7	3.2	86
4	VOLKSWAGEN	LUPO 3L TDI	MVW70C1RY573	6ESCAYZX01AGFD5850021N0I	GO	3	45.0	A 5	3.8	2.8	3.2	86

We see the sparsity of the data columns in general compared to recent EU data, which featured 30 columns, here we are down to 11. Even though we have also reduced the EU dataset for the modeling part a bit, this reduces our possibilities drastically if we want to try re-introducing some of the EU data columns later in during the modeling.

Additionally here is a heatmap of the numeric variables:



("Urbain" / "Ex.Urb" were dropped, as "Mixte" gives us their relevant mean value already)

Even in this reduced dataset, we of course still see the important high correlation of fuel consumption and co2 emission, but we also see the limitedness of the french data before 2010.

This should serve as an additional indicator on why we should drop the data.

7. Conclusion/Next Steps

Until now in this project, we:

Explored two major datasets (EU and French), eventually removing the French data because it lacked vital columns in earlier years and risked duplicating or introducing noise.

Focused on the most relevant features for predicting fuel or energy consumption (mass_vehicle, fuel_type, engine_power, etc.), aiming to derive CO_2 or CO_2e.
Used recognized formulas (e.g., KBA factors) to convert predicted consumption into emissions.
Dealt with outliers and correlation analyses to validate the strong link between consumption and CO_2.

Key Takeaways:

- Fuel consumption strongly correlates with CO_2 for combustion vehicles, offering a reliable path for emission estimation.
- Electric vehicles require factoring in the local electricity supply's emission intensity for accurate CO_2e calculations.
- In future work, incorporating region-specific emission factors for electricity and exploring additional (or updated) data sources could refine these predictions.

Next Steps:

With the data exploration completed, the next stage of the project will focus on modeling. This will involve defining the type of machine learning problem, selecting appropriate algorithms, and optimizing model performance. Key aspects will include evaluating different approaches, refining hyperparameters, and assessing interpretability techniques.
As the project progresses, a detailed analysis of errors and model performance will guide improvements, ensuring reliability and relevance. The final phase will focus on assessing challenges encountered, refining the methodology, and identifying avenues for further optimization.

References & Resources

European Environment Agency (EEA)
https://www.eea.europa.eu/en/datahub/datahubitem-view/fa8b1229-3db6-495d-b18e-9c9b3267c02b
French Government Data
https://www.data.gouv.fr/fr/datasets/emissions-de-co2-et-de-polluants-des-vehicules-commercialises-en-france/
World Bank - World Development Indicators
European Commission: GHG Emission Factors for Electricity Consumption per country
CaDI - The Carbon Database
KBA Reference: CELEX:31993L0116