# **Sprint 1 (Data Mining + DataViz’)**

The dataset provides a comprehensive overview of the technical characteristics and environmental impact of vehicles marketed in France in 2013. Analyzing this data can help identify trends, patterns,

and insights related to vehicle emissions, fuel efficiency, and other factors influencing air quality and environmental sustainability.

## **Initial Data Exploration**

This section provides an overview of the dataset.

* **Size**: The dataset consists of 44,850 entries and 26 columns.
* **Columns:** The dataset includes various attributes, including
  + **Marque:** The brand or manufacturer of the vehicle.
  + **Modèle dossier:** Model name assigned by UTAC (Union Technique de l'Automobile, du Motocycle et du Cycle), which is a French organization responsible for vehicle approval.
  + **Modèle UTAC:** Model name assigned by UTAC.
  + **Désignation commerciale:** Commercial designation of the vehicle.
  + **CNIT:** Identification number assigned by the French government for each vehicle model.
  + **Type Variante Version (TVV):** Type, variant, and version of the vehicle.
  + **Carburant:** Type of fuel used by the vehicle (e.g., gasoline, diesel, hybrid, etc.).
  + **Hybride:** Indicates whether the vehicle is hybrid (yes/no).
  + **Puissance administrative:** Administrative horsepower of the vehicle.
  + **Puissance maximale (kW):** Maximum power of the vehicle in kilowatts.
  + **Boîte de vitesse:** Type of transmission (e.g., manual, automatic).
  + **Consommation urbaine (l/100km):** Urban fuel consumption in liters per 100 kilometers.
  + **Consommation extra-urbaine (l/100km):** Extra-urban fuel consumption in liters per 100 kilometers.
  + **Consommation mixte (l/100km):** Mixed fuel consumption in liters per 100 kilometers.
  + **CO2 (g/km):** CO2 emissions of the vehicle in grams per kilometer.
  + **CO type I (g/km):** CO emissions of type I (e.g., carbon monoxide) in grams per kilometer.
  + **HC (g/km):** Hydrocarbon emissions in grams per kilometer.
  + **NOX (g/km):** Nitrogen oxide emissions in grams per kilometer.
  + **HC+NOX (g/km):** Combined hydrocarbon and nitrogen oxide emissions in grams per kilometer.
  + **Particules (g/km):** Particulate emissions in grams per kilometer.
  + **masse vide euro min (kg):** Minimum empty weight of the vehicle in kilograms according to Euro standards.
  + **masse vide euro max (kg):** Maximum empty weight of the vehicle in kilograms according to Euro standards.
  + **Champ V9:** Additional field.
  + **Date de mise à jour:** Date of the last update for the vehicle.
  + **Carrosserie:** Body type of the vehicle (e.g., sedan, hatchback, SUV).
  + **gamme:** Range or series of the vehicle.
* **Data Types**: The dataset contains a mix of data types, including object, int64, and float64. Figure 1 shows the type of each column.

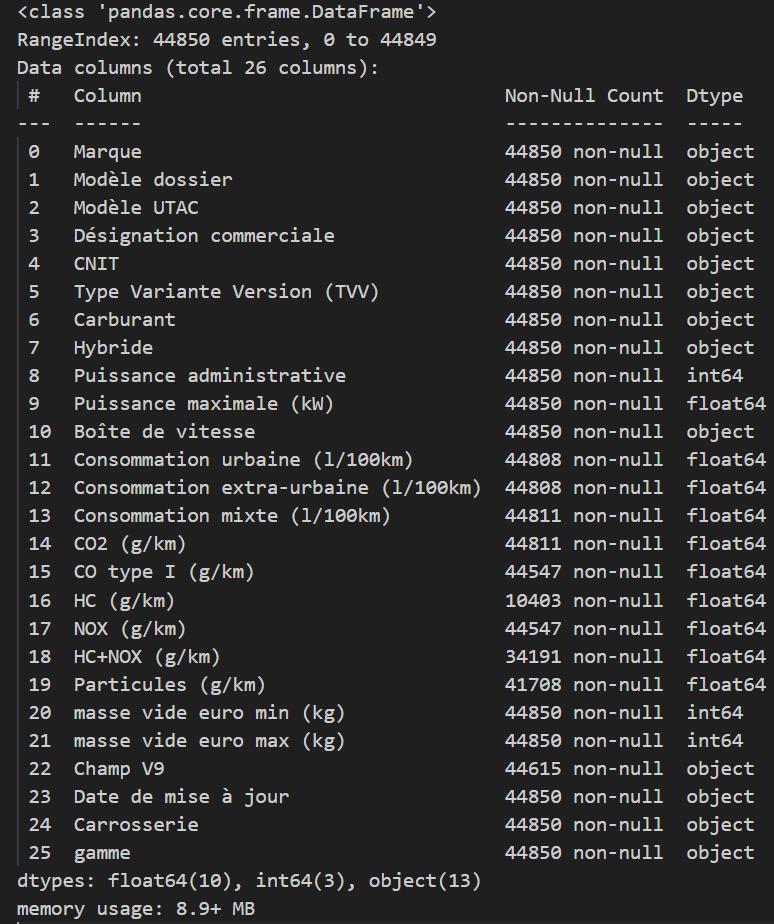


Figure 1. An overview of data types in the dataset

* **Missing values**: Some columns have missing values, such as Consommation urbaine, Consommation extra-urbaine, CO2 emissions, and others. Figure 2 shows the amount of missing values for each column.

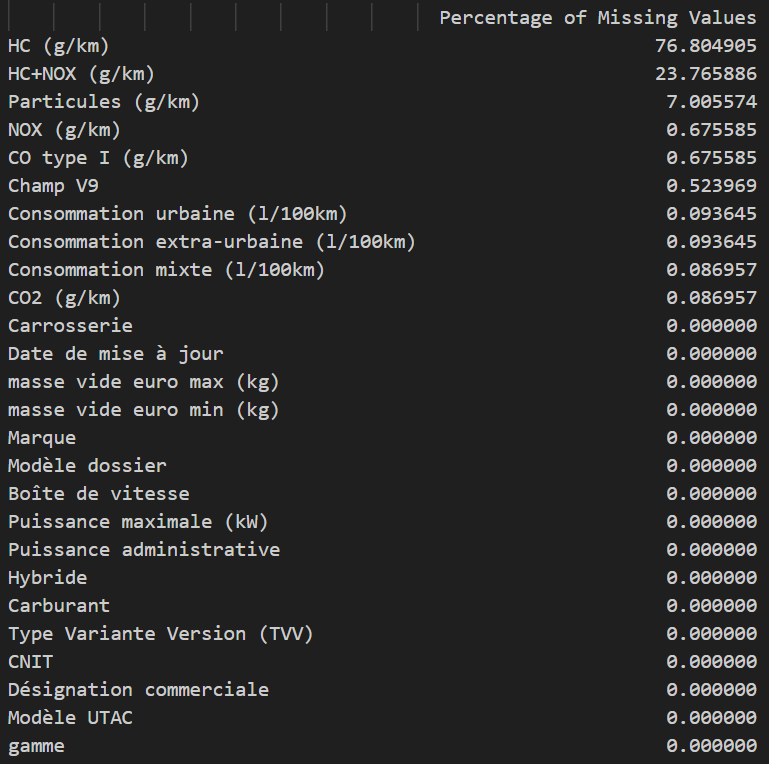


Figure 2. Percentage of missing values in the dataset

* **Hybrid Vehicles:** The dataset indicates whether a vehicle is hybrid (Hybride), with most vehicles being non-hybrid.
* **Last Updated:** The dataset includes a column indicating the date of the last update for each vehicle.
* **Target Variable:** It seems that one of the potential target variables for analysis could be CO2 emissions (CO2 (g/km)), which is essential for assessing the environmental impact of vehicles.

## **Potential Challanges**

In this section, we want to show the distribution of values for each attribute and potential outliers.

### **Numerical Variables**

In this section, potential challenges with regard to the numerical attributes are presented.

1. **Puissance administrative:** Statistics of this attribute are:

count 44850.000000

mean 11.018997

std 5.554475

min 1.000000

25% 9.000000

50% 10.000000

75% 11.000000

max 81.000000

Name: Puissance administrative, dtype: float64

Figure 3 shows the distribution of the data for this attribute.

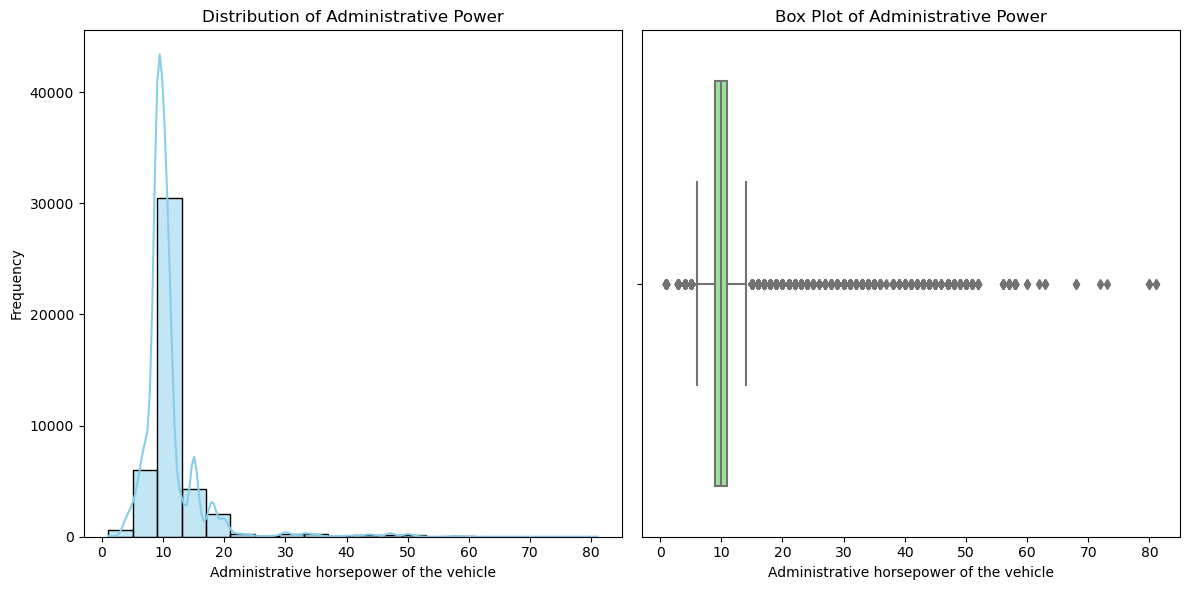


Figure 3. Distribution of Administrative horsepower of the vehicle

There is no missing value in this column. **As we can see in the boxplot, there are several outliers that we need to handle.** The calculation of outliers shows that:

Number of outliers: 7813

Lower whisker: 6.0

Upper whisker: 14.0

1. **Puissance maximale:** Statistics of this attribute are:

count 44850.000000

mean 124.780834

std 49.158804

min 10.000000

25% 100.000000

50% 120.000000

75% 125.000000

max 559.300000

Name: Puissance maximale (kW), dtype: float64

Figure 4 shows the distribution of the data for this attribute.

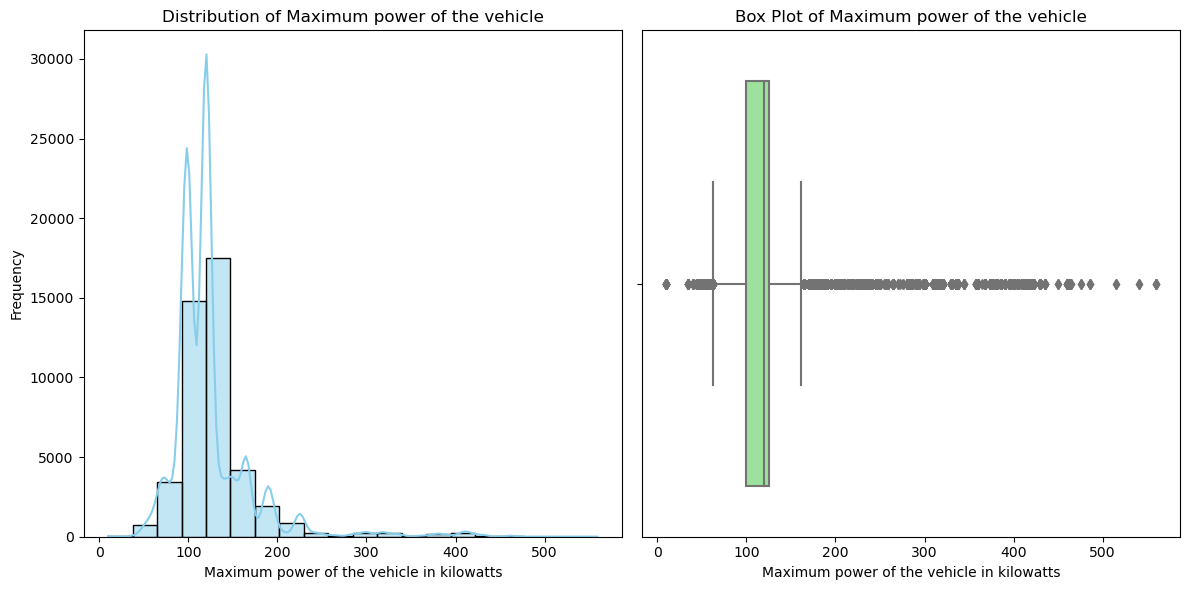


Figure 4. Distribution of maximum power

To identify outliers, we used the boxplot diagram (Figure 5). It shows that there are 7259 data points that are smaller than the Lower whisker (62.5) or greater than the Upper whisker: 162.5. **We need to select an appropriate method to handle these outliers.** There is no missing value for this attribute.

1. **Consommation urbaine:** Statistics of this attribute are:

count 44808.000000

mean 9.706744

std 2.366181

min 0.000000

25% 8.800000

50% 9.800000

75% 10.700000

max 41.100000

Name: Consommation urbaine (l/100km), dtype: float64

There are 44,808 non-null values out of a total of 44,850 entries in the dataset, **suggesting some missing values**. Furthermore, the min and max values show that there may be some erroneous or missing data entries. Figure 5 shows the distribution of data in this column. We identified 4210 data points that are smaller than the Lower whisker (5.950000000000003) or greater than the upper whisker (13.549999999999997). **We need to handle these outliers.**

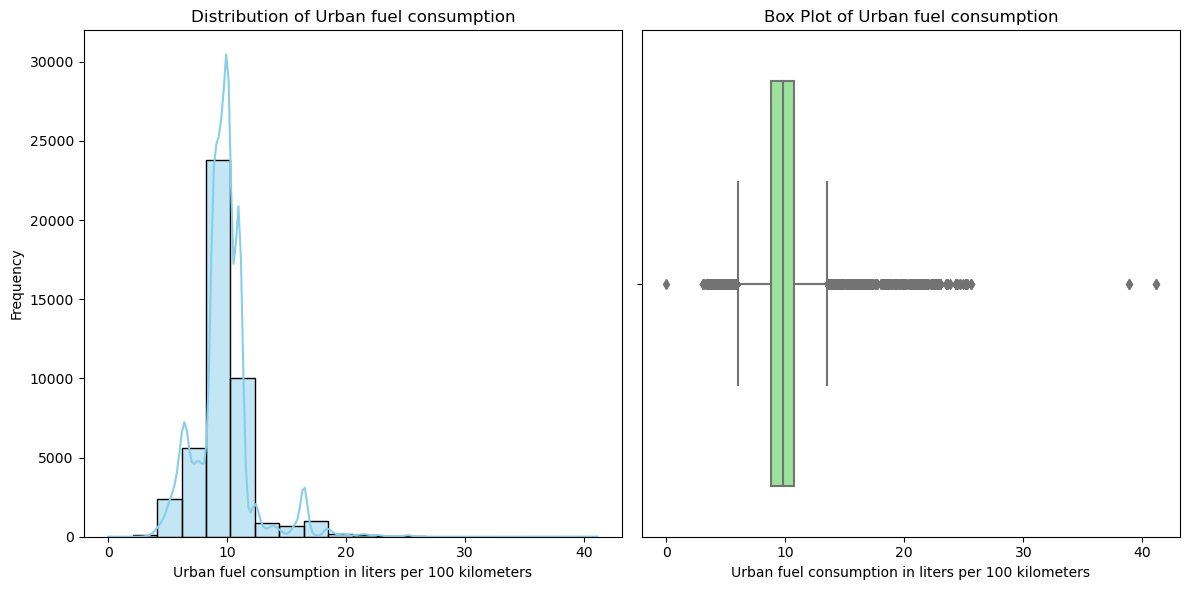


Figure 5. Distribution of Urban fuel consumption in liters per 100 kilometers

1. **Consommation extra-urbaine:** Statistics for this attribute are:

count 44808.000000

mean 6.567634

std 1.196234

min 2.800000

25% 6.300000

50% 6.700000

75% 7.100000

max 14.900000

Name: Consommation extra-urbaine (l/100km), dtype: float64

There are 44,808 non-null values out of a total of 44,850 entries in the dataset, **suggesting some missing values**. The minimum value of 2.8 l/100km suggests that there may be some vehicles with very low fuel consumption. Figure 6 shows some representations of this attribute. Analysis of the dataset shows that there are 8546 outlier or extreme values. **We need to find a suitable approach to manage them.**

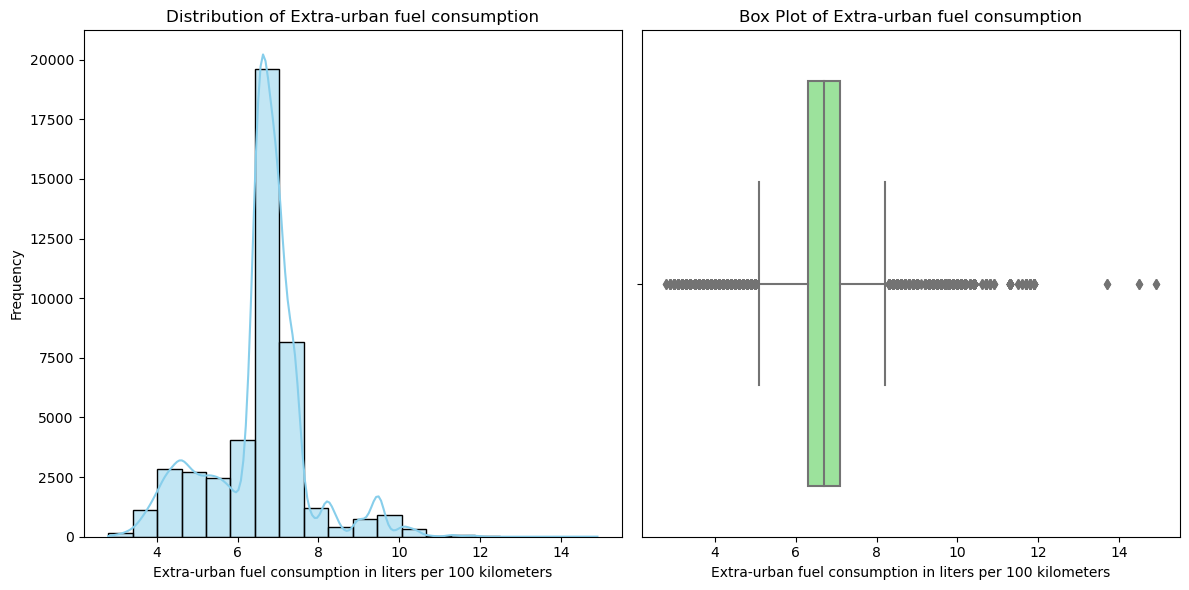


Figure 6. Distribution of Extra-Urban fuel consumption

1. **Consommation mixte:** Statistics of this attribute are:

count 44811.000000

mean 7.716254

std 1.597110

min 1.200000

25% 7.200000

50% 7.700000

75% 8.400000

max 24.500000

Name: Consommation mixte (l/100km), dtype: float64

There are 44,811 non-null values out of a total of 44,850 entries in the dataset, **suggesting some missing values.** Figure 7 shows distribution of the data for this attribute and also the boxplot to analyze the outliers or extreme values. Analysis of data shows that there are 6254 outlier or extreme values greater than Upper whisker (10.200000000000001) or lower than Lower whisker (5.4). **We need to find a suitable approach to manage them.**

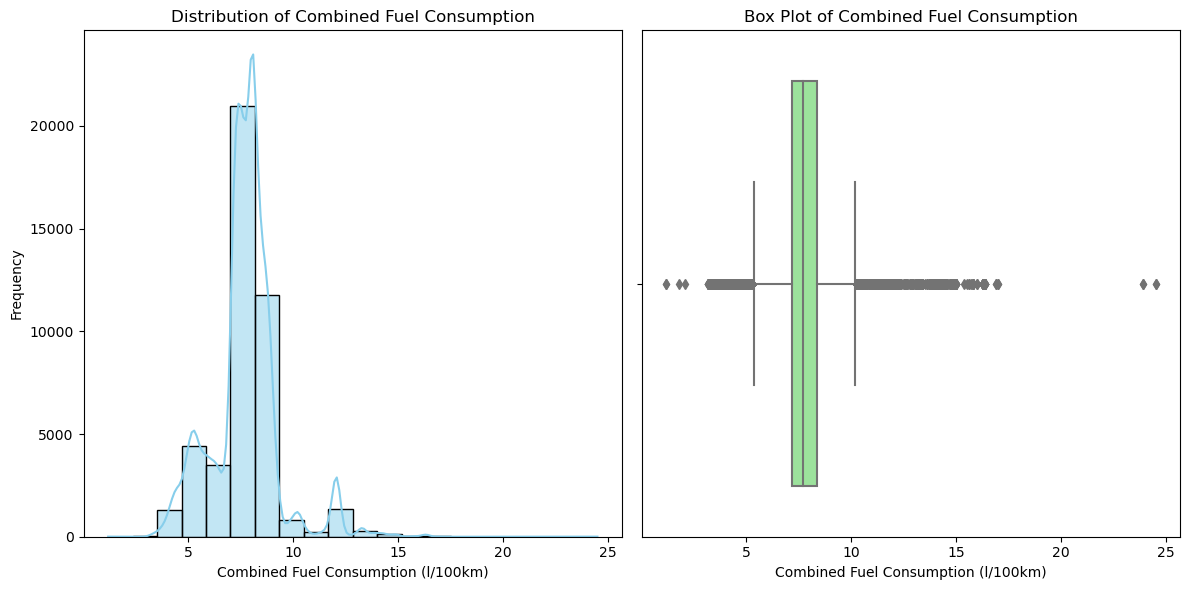


Figure 7. Representation of Combined Fuel Consumption

1. **Co2:** Statistics of this attribute are:

count 44811.000000

mean 198.910892

std 39.014678

min 27.000000

25% 187.000000

50% 203.000000

75% 221.000000

max 572.000000

Name: CO2 (g/km), dtype: float64

There are 44,811 non-null values out of a total of 44,850 entries in the dataset, **suggesting some missing values.** The minimum value of 27 g/km suggests that there may be some vehicles with very low CO2 emissions. Figure 8 shows two representations of data. Analysis of data shows that there are 5458 outlier or extreme values greater than Upper whisker (272.0) or lower than Lower whisker (136.0). **We need to find a suitable approach to manage them**

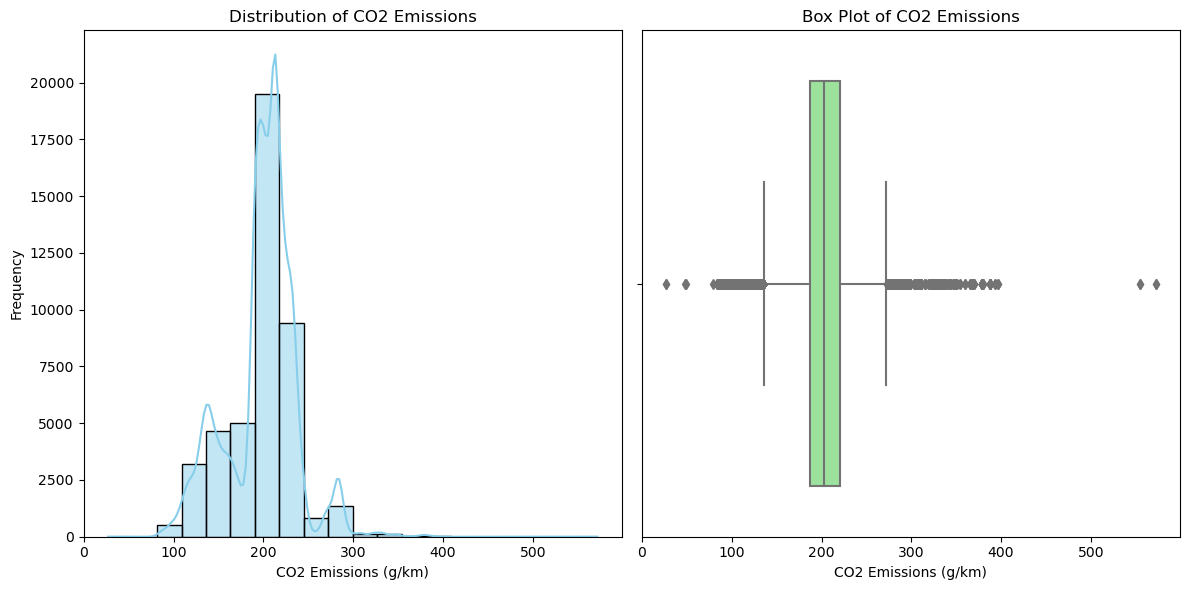


Figure 8. Representations of CO2 emissions

1. **CO type I (g/km):** Statistics of this attribute are:

count 44547.000000

mean 0.153461

std 0.138984

min 0.005000

25% 0.046000

50% 0.093000

75% 0.222000

max 0.968000

Name: CO type I (g/km), dtype: float64

There are 44,547 non-null values out of a total of 44,850 entries in the dataset, **suggesting some missing values**. Figure 9 shows two representations of this attribute. The boxplot shows that most of the data are concentrated on the right side after the median line inside of the box. It suggests that the distribution of the data is positively skewed. In other words, there are more data points with higher values than with lower values. Furthermore, outliers are concentrated on the right side of the boxplot. We have these statistics for outliers:

Number of outliers: 1009

Lower whisker: -0.21800000000000003

Upper whisker: 0.486

There are 303 NaN values in this column and therefore it affects the calculation of outliers. **We need to handle these missing values.**

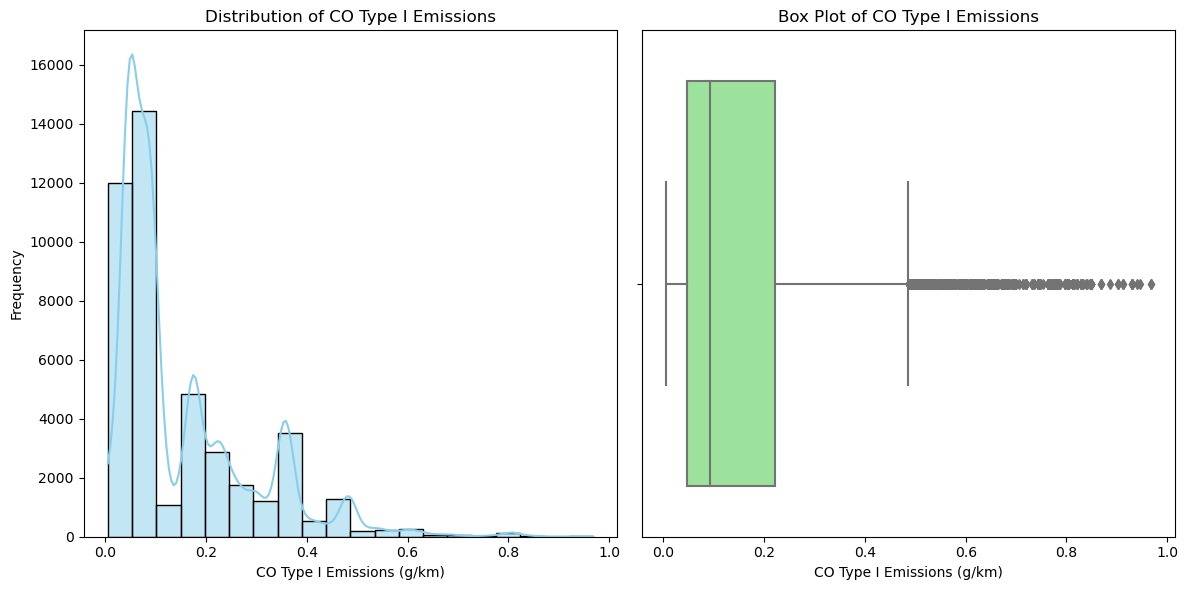


Figure 9. Distribution of CO type 1 emissions

1. **HC (g/km):** Statistics of this attribute are:

count 10403.000000

mean 0.030499

std 0.018408

min 0.008000

25% 0.008000

50% 0.031000

75% 0.044000

max 0.143000

Name: HC (g/km), dtype: float64

As we can see, there are too many missing values in this column (34447). **We need to decide how can we manage this issue.** Figure 10 shows the distribution of the data. As we have a bunch of missing values, the results for the calculation of outliers and lower and upper whiskers are:  
Number of outliers: 4

Lower whisker: -0.04599999999999999

Upper whisker: 0.09799999999999999

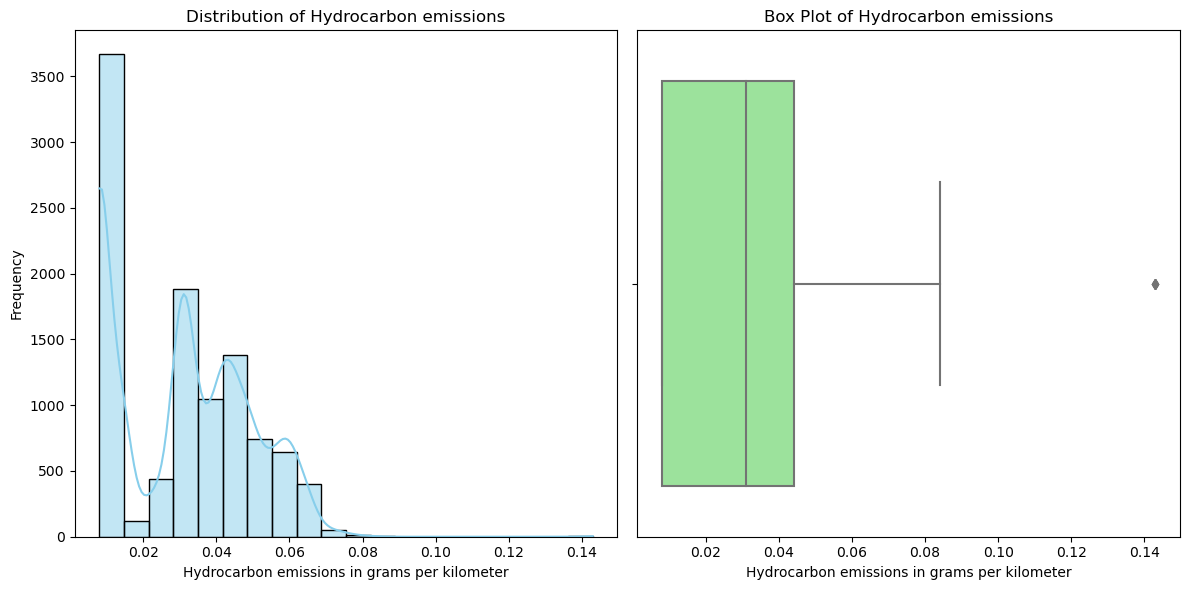


Figure 10. Distribution of Hydrocarbon emissions

1. **NOX (g/km):** Statistics of this attribute are:

count 44547.000000

mean 0.311837

std 0.463112

min 0.001000

25% 0.158000

50% 0.197000

75% 0.228000

max 1.846000

Name: NOX (g/km), dtype: float64

**This column has 303 missing values.** Figure 11 shows the distribution of the data for this attribute. We calculated the number of outliers for this attribute:

Number of outliers: 10426

Lower whisker: 0.05299999999999999

Upper whisker: 0.333

**Therefore, we need to handle these outliers.**

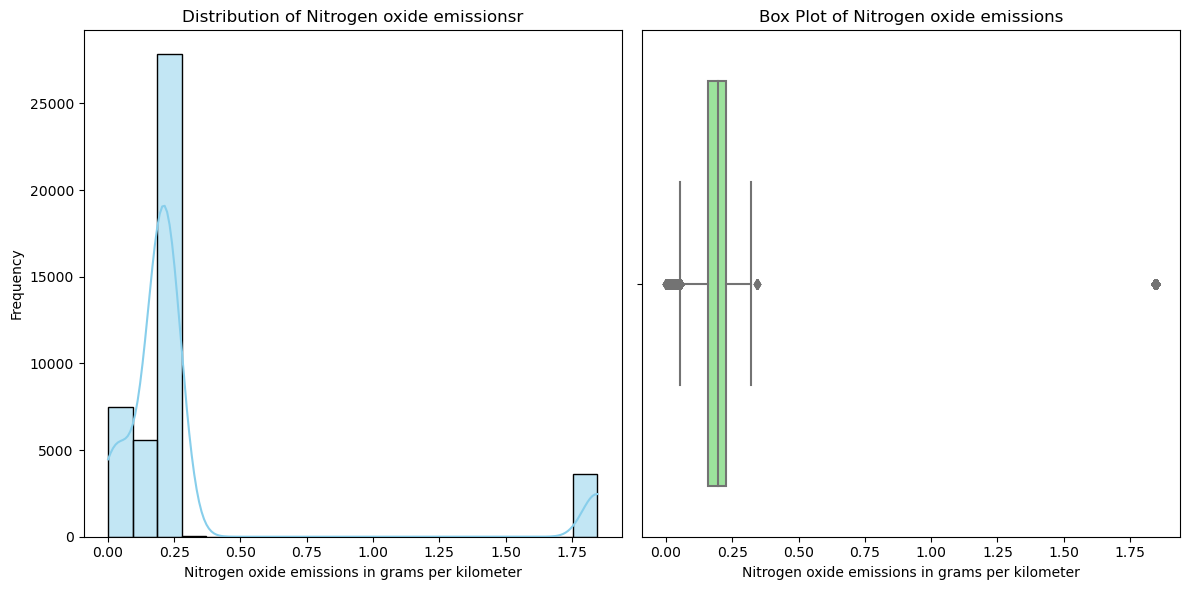


Figure 11. Distribution of data for Nitrogen oxid emissions

1. HC+NOX (g/km): Statistics of this attribute:

count 34191.000000

mean 0.224788

std 0.041681

min 0.038000

25% 0.201000

50% 0.220000

75% 0.248000

max 0.306000

Name: HC+NOX (g/km), dtype: float64

There are 10659 missing values in this column and **we need to decide about the approach to handle them.** Figure 12 shows the distribution of the data in this column. The title of this column says that this attribute is the combination of two previous columns. We need to examine whether they are directly concluded from the two previous columns or they are new data. We calculated the number of outliers:

Number of outliers: 883

Lower whisker: 0.13050000000000003

Upper whisker: 0.3185

**Therefore, we need to decide on which approach is suitable to handle these outliers.**

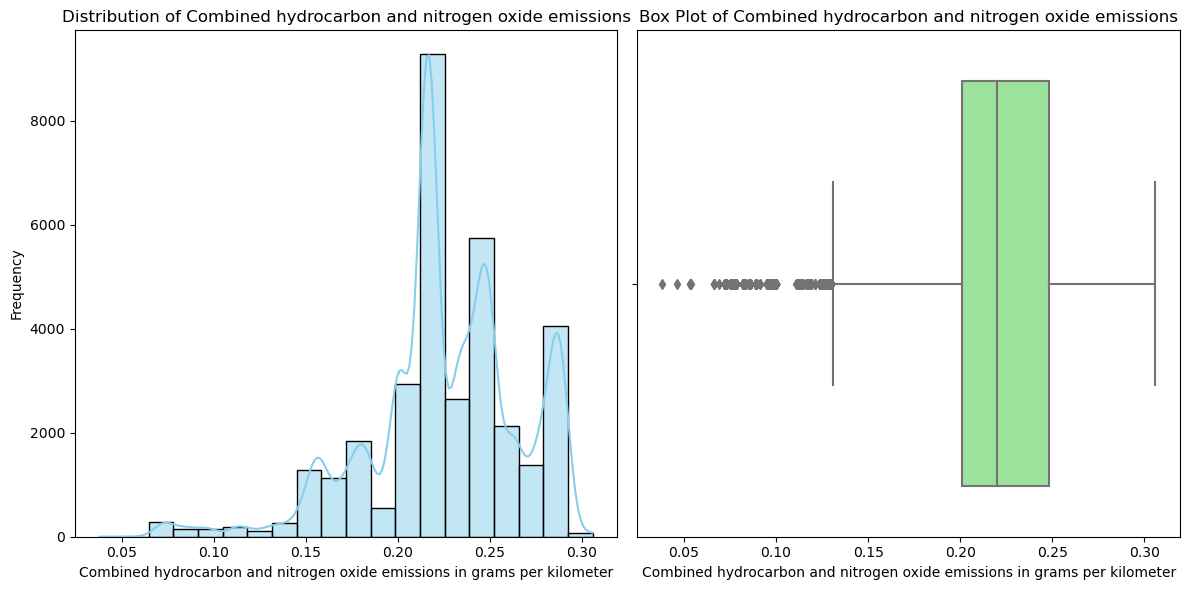


Figure 12. Distribution of Combined hydrocarbon and nitrogen oxide emissions

1. **Particules (g/km):** Statistics of this attribute are:

count 41708.000000

mean 0.000961

std 0.006469

min 0.000000

25% 0.000000

50% 0.001000

75% 0.001000

max 0.610000

Name: Particules (g/km), dtype: float64

**There are 3142 missing values in this column**. Figure 13 shows distribution of data in this column. **This figure shows that there are some anomalies in the data.** This is the results of calculation for outliers:

Number of outliers: 3696

Lower whisker: -0.0015

Upper whisker: 0.0025

We need to decide how can we handle these anomalies.

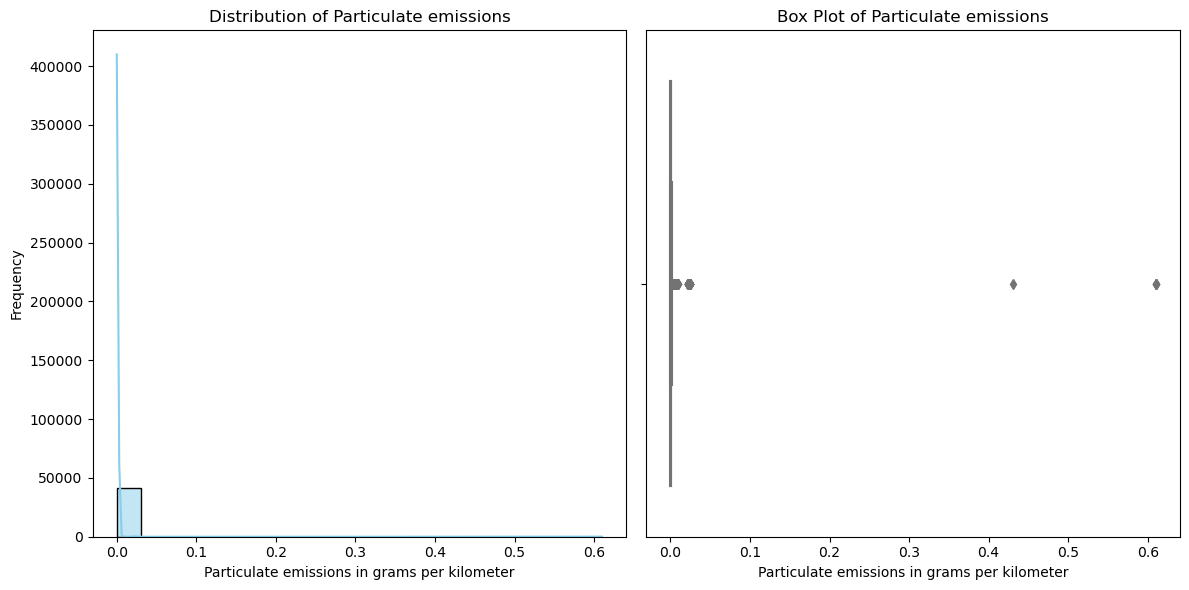


Figure 13. Distribution of Particular emissions

1. **masse vide euro min (kg):** Statistics of this attribute are:

count 44850.000000

mean 2070.961650

std 342.872975

min 825.000000

25% 1976.000000

50% 2076.000000

75% 2256.000000

max 3115.000000

Name: masse vide euro min (kg), dtype: float64

There is no missing value in this column. Figure 14 shows the distribution of data. This is the results of calculations for outliers; **we need to manage them.**

Number of outliers: 4397

Lower whisker: 1556.0

Upper whisker: 2676.0

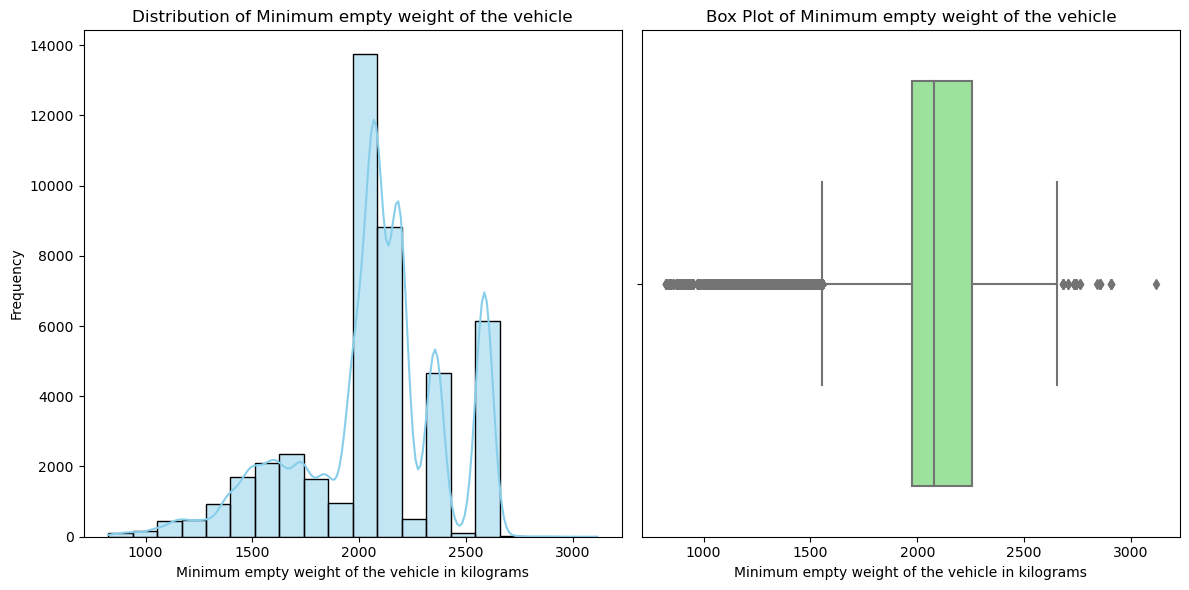


Figure 14. Distribution of Minimum empty weight of the vehicle

1. **masse vide euro max (kg):** Statistics of this attribute are:

count 44850.000000

mean 2169.545284

std 410.600541

min 825.000000

25% 2043.500000

50% 2185.000000

75% 2355.000000

max 3115.000000

Name: masse vide euro max (kg), dtype: float64

There is no missing value in this column. Figure 15 shows the distribution of data.

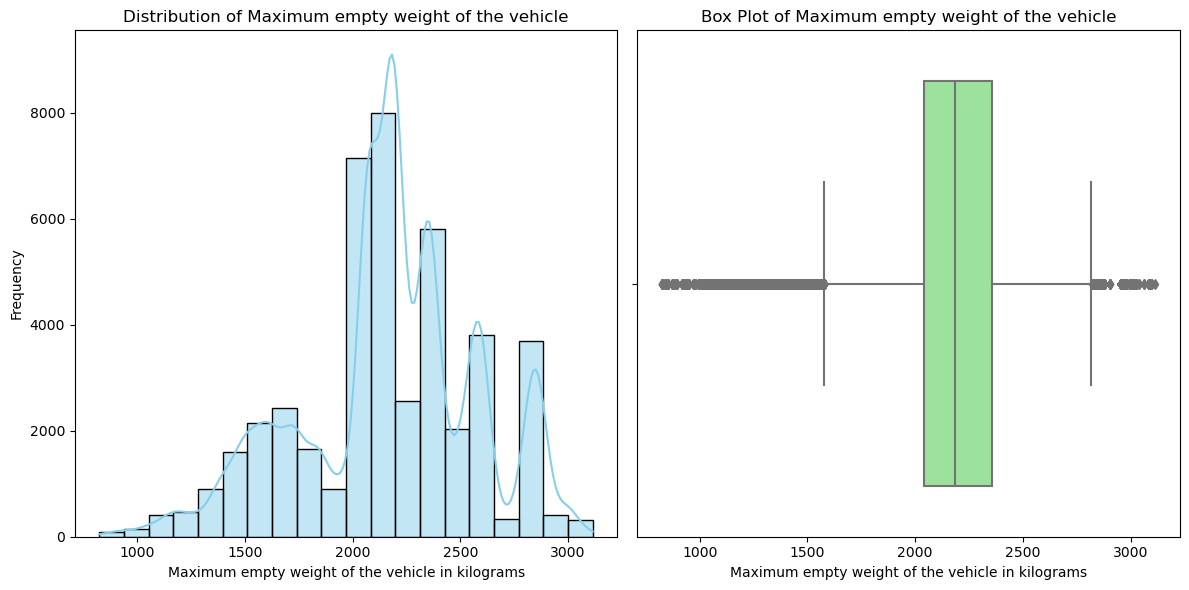


Figure 15. Distribution of Maximum empty weight of the vehicle

As we can see in the boxplot, **there are many outliers that we need to handle:**

Number of outliers: 8240

Lower whisker: 1576.25

Upper whisker: 2822.25

### **Categorical Variables**

There are 13 categorical attributes in the dataset. The number of categories for each variable are:

* 'Marque': 51
* 'Modèle dossier': 458
* 'Modèle UTAC': 419
* 'Désignation commerciale': 3582
* 'CNIT': 44191
* 'Type Variante Version (TVV)': 28781
* 'Carburant': 13
* 'Hybride': 2
* 'Boîte de vitesse': 16
* 'Champ V9': 13
* 'Date de mise à jour': 3
* 'Carrosserie': 10
* 'gamme': 7

As we can wee, there are variables with so many categories. When preparing data for machine learning models, categorical variables need to be transformed into numerical representations. There are several approaches to handle categorical variables with many categories, including One-Hot Encoding, Frequency Encoding, and Target Encoding. **We need to decide which approach is suitable for encoding the mentioned categorical variables.**

## **Data Visualization**

In this section, five representations of the data are presented.

* Figure 16 shows the relationship between consomation Mixte and Co2 emissions.

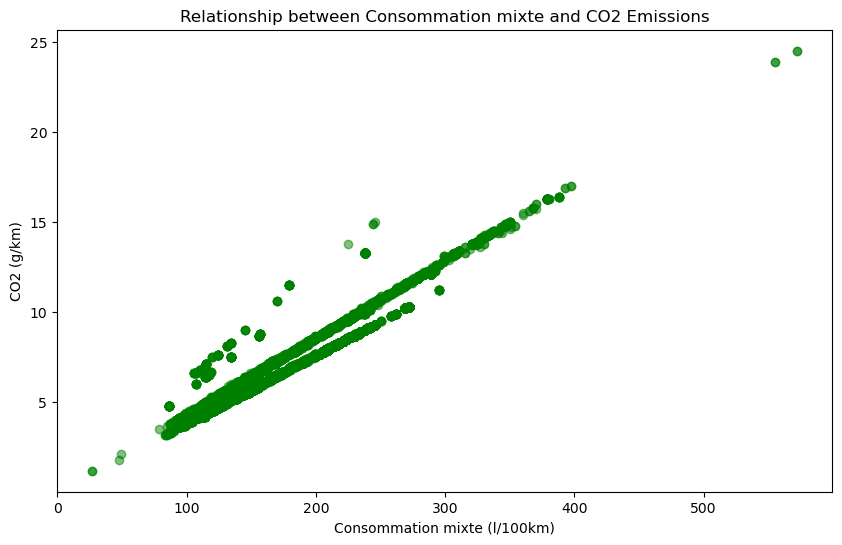


Figure 16. Relationship between Consomation Mixte and Co2 Emissions

The scatterplot illustrates a clear positive correlation between fuel consumption ('Consommation mixte') and CO2 emissions. The main trendline slopes from the bottom left to the top right, indicating that vehicles with higher fuel consumption tend to have higher CO2 emissions.

There are two additional lines observed at approximately 70 l/100 km. One line runs below the main trendline up to around 300 l/100 km, while the other runs at the same height but above the main trendline. These additional lines suggest potential subgroups or distinct characteristics within the data:

* Below the main line up to around 300 l/100 km: Vehicles in this range may possess specific attributes leading to lower CO2 emissions despite higher fuel consumption.
* Above the main line up to around 300 l/100 km: Vehicles in this range may have characteristics resulting in higher CO2 emissions, even with relatively lower fuel consumption.

Further exploration and subgroup analysis could provide insights into the unique features influencing CO2 emissions within these fuel consumption ranges.

* Figure 17 shows the influence of weight and hybrid on Co2 emmissions.

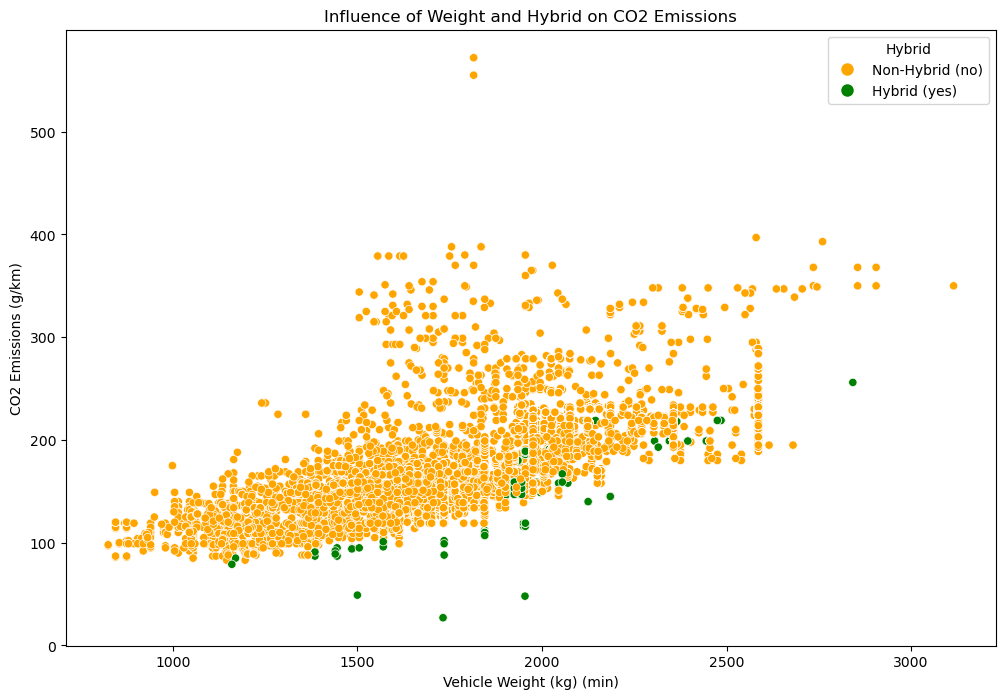


Figure 17. Influence of weight and hybrid on Co2 emissions

Weight and Hybrid Influence on CO2 Emissions:

* For vehicles with a weight below approximately 1800 kg (minimum weight), the green points (Hybrid vehicles) seem to have generally lower CO2 emissions compared to the orange points (Non-Hybrid vehicles). This could suggest that Hybrid vehicles may be more efficient in terms of CO2 emissions at lighter weights.
* From a weight of about 1800 kg onwards, there appears to be a mixing of green and orange points, indicating that the influence of vehicle weight on CO2 emissions becomes more similar between Hybrid and Non-Hybrid vehicles.

It would be interesting to conduct additional analyses, such as dividing the data into weight ranges and examining average CO2 emissions in these ranges. This could help further explore the relationship between weight, hybrid properties, and CO2 emissions.

* Figure 18 presents the distribution of CO2 emissions by car fuel type.

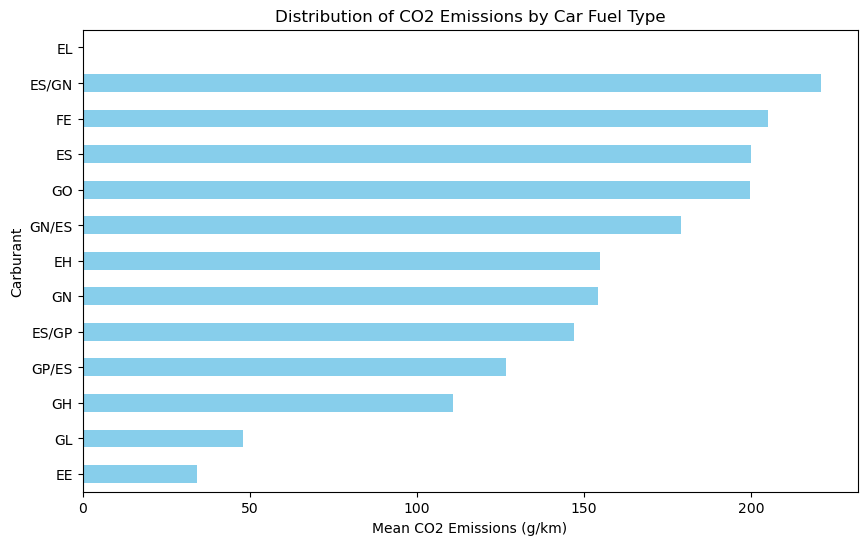


Figure 17. Distribution of CO2 emissions by car fuel type

The horizontal bar plot we created shows the mean CO2 emissions for different car fuel types.

The order of the bars from top to bottom represents the ascending order of mean CO2 emissions. Here's an interpretation based on our observation:

* ES/GN (Ethanol/Gasoline): This category has the highest mean CO2 emissions among the displayed fuel types. It indicates that vehicles using a combination of Ethanol and Gasoline tend to emit more CO2 on average.
* EE (Electric/Electric): This category represents vehicles that are fully electric (Electric) and use an electric powertrain exclusively.

The "EE" category has the lowest mean CO2 emissions among the displayed fuel types. The low CO2 emissions in this category indicate a reduced carbon footprint compared to vehicles with other fuel types.

* EL (Electric/LPG): The absence of a bar for this category in the plot suggests that there might be very few or no vehicles with the combination of Electric and LPG in the dataset. As a result, the dataset does not provide sufficient information to calculate a meaningful mean CO2 emissions value for this particular fuel type.

This is information about other fuel types:

* + FE (Electric)
  + ES (Electric/Gasoline) and Go (Gasoline)
  + GN/ES (Natural Gas/Electric)
  + EH (Electric/Hybrid)
  + GN (Natural Gas)
  + ES/GP (Electric/LPG)
  + GP/ES (LPG/Electric)
  + GH (Gasoline/Hybrid)
  + GL (LPG/Hybrid)

We need to consider that these interpretations are based on the mean values, and individual vehicles within each category may vary in their CO2 emissions.

* Figure 18 shows the relationship between power and CO2 emissions

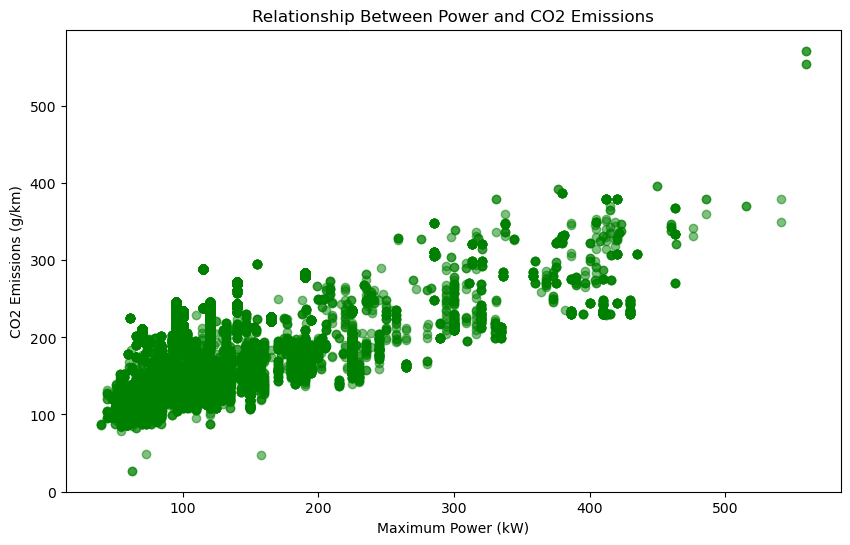


Figure 18. Relationship between power and CO2 emissions

We can identify two observations based on this diagram:

* Vehicles with more power tend to have higher CO2 emissions and vice versa. This observation aligns with general expectations, as vehicles with higher power often have larger engines and may consume more fuel.
* A large number of data points are concentrated within a specific range of power values, such as between approximately 50 and 250. But, we have a large number of data points concentrated within a specific range of power values, such as between approximately 50 and 250. The concentration of data points in a specific power range may indicate that many vehicles in the dataset share similar power characteristics. But, we need to explore the dataset further to see popular models or manufacturers within this power range. Vehicles with extremely high-power values (e.g., greater than 400 kW) are often high-performance or specialty vehicles. These could include sports cars, luxury vehicles, or other niche segments. The limited number of data points in this range might be reflective of the relatively lower production volume of such vehicles compared to mainstream models.
* Figure 19 shows the distribution of CO2 emissions by Car Make (Marque).

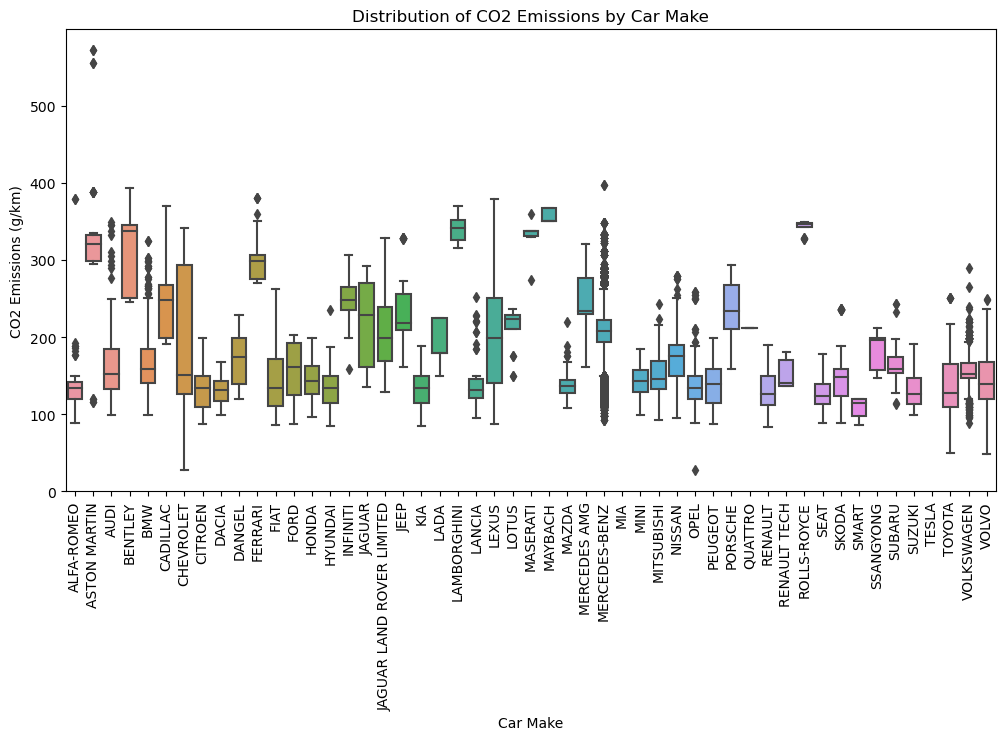


Figure 19. Distribution of CO2 emissions by Car make

Based on this figure, we can make some observations about which car makes are performing relatively well or poorly in terms of CO2 emissions:

1. **Low CO2 emissions**: Car makes with lower mean and median CO2 emissions compared to the overall average of approximately 191.6 g/km can be considered as performing relatively well in terms of emissions efficiency. Examples of such car makes include:
   * ALFA-ROMEO: Mean = 134.78 g/km, Median = 134.0 g/km
   * CITROEN: Mean = 133.01 g/km, Median = 133.5 g/km
   * DACIA: Mean = 130.93 g/km, Median = 131.5 g/km
   * KIA: Mean = 133.21 g/km, Median = 134.5 g/km
   * PEUGEOT: Mean = 137.76 g/km, Median = 139.0 g/km
2. **High CO2 Emissions**: Car makes with higher mean and median CO2 emissions compared to the overall average indicate relatively poorer emissions performance. Examples of such car makes include:
   * BENTLEY: Mean = 313.55 g/km, Median = 338.0 g/km
   * LAMBORGHINI: Mean = 339.75 g/km, Median = 341.0 g/km
   * MAYBACH: Mean = 358.31 g/km, Median = 350.0 g/km
   * MERCEDES-BENZ: Mean = 205.04 g/km, Median = 208.0 g/km
   * ROLLS-ROYCE: Mean = 342.75 g/km, Median = 347.0 g/km

It should be noted that there are various factors that we need to be considered while categorizing car makes as “doing good” or “poorly”, such as the vehicle type, market segment, and the size and performance of vehicles.

# **2. Sprint 2 (Pre-processing and Feature Engineering)**

In data processing, "preprocessing" refers to a series of steps performed before the actual analysis or processing of data. The goal of preprocessing is to prepare the data in a way that is optimally suitable for specific applications or algorithms. Here are some common tasks in preprocessing:

* **Data Cleaning:** Removing faulty or incomplete data to improve data quality.
* **Data Normalization/Standardization:** Adjusting data to ensure it is on a consistent scale. This facilitates comparison and application of algorithms.
* **Feature Engineering:** Creating new features from existing data to enhance model performance.
* **Outlier Removal**: Identifying and removing extreme values that could affect analysis or modeling.
* **Handling Missing Values**: Deciding how to deal with missing data, whether through imputation, removal, or other methods.
* **Data Reshaping**: Adjusting the data structure to make it more compatible with the requirements of algorithms or models.
* **Data Splitting**: Dividing the data into training and test sets to evaluate model performance.

Preprocessing is crucial to ensure that the data is of the highest quality possible and that models can be trained on it correctly and effectively.

## **2.1. Handling Missing Values**

As explained in the previous section, some of the columns have missing values.

* **HC (g/km):** Given that this attribute has a high percentage of missing values (~76.8%), we decided to drop this column from the dataset.
* **HC+NOX(g/km):** With nearly 24% missing values in this column, using mean or median imputation might not be the best approach, especially if the missingness is not completely at random. Therefore, we used **K-Nearest Neighbors (KNN) Imputation** to handle missing values in this column.
* **Particules (g/km):** The percentage of missing values in this column is approximately 7%. We can use mean or median imputation to handle these missing values. We used median approach for this purpose.
* **NOX (g/km), CO type I (g/km), Consommation urbaine (l/100km), Consommation extra-urbaine (l/100km), Consommation mixte (l/100km), CO2 (g/km):** Since the percentage of missing values is relatively low for these columns, we used mean or median imputation to fill in the missing values. We used mean imputation to handle missing values in these columns.
* **Champ V9:** This is a categorical column and has approximately 0.52% missing values. We used the mode imputation approach to replace missing values with the most frequent category (mode) in the column.
* **Other Numeric Columns (e.g., Puissance maximale (kW), Puissance administrative, etc.):** For these columns with no missing values, no action is needed.
* **Other Categorical Columns (e.g., Carrosserie, Carburant, etc.):** Since there are no missing values in these columns, no action is needed.

## **2.2. Handling Duplicates**

In the next step, we removed the duplicate values.

## **2.3. Handling Categorical Variables**

The frequency of categorical variables in the dataset is not the same. The number of categories for each categorical variable is as follows:

* Marque: 51 categories
* Modèle dossier: 458 categories
* Modèle UTAC: 419 categories
* Désignation commerciale: 3582 categories
* CNIT: 44191 categories
* Type Variante Version (TVV): 28781 categories
* Carburant: 13 categories
* Hybride: 2 categories
* Boîte de vitesse: 16 categories
* Champ V9: 13 categories: whether the information provided in this column is meaningful?
* Date de mise à jour: 3 categories
* Carrosserie: 10 categories
* gamme: 7 categories

As it can be seen, the number of categories for some of the variables are very high. Therefore, we need to decide how to handle each categorical variable based on the number of categories and our specific requirements for the analysis.

We used the following approaches to handle categorical variables:

* **Label Encoding:** Label encoding assigns a unique numerical value to each category. We used this approach to handle the variables Hybride and Date de mise à jour. We applied this method to the CNIT column. Although the values in this column are alphanumeric codes rather than ordinal categories and label encoding might not be appropriate for such data because it would imply an order that may not exist, we used this approach in this stage of the project. In the next stages, we maybe consider using other approaches like PCA.
* **One-Hot Encoding (OHE):** This approach creates binary columns for each category, indicating the presence or absence of the category. However, it can lead to a significant increase in the dimensionality of the dataset, which might not be feasible with extremely large categories. We applied this method to the variable gamme. We have also applied this to the Marque column; although this approach will add 51 new columns to the dataset, we will use some techniques for dimensions reduction in the next sprints. We also used this method for Carburant, Boîte de vitess, and Carrosserie coulmns.

## **Feature engineering:**

Analyzing the dataset shows that "Modèle dossier" and "Modèle UTAC" are both indicating the model name, which is for example "RANGE ROVER." On the other hand, "Désignation commerciale" provides a more detailed description, including additional information such as the engine specification ("TDV6 (258ch)"), seating capacity ("4PL"), and possibly some other specifications. Therefore, we decided to drop the Modèle dossier column and keep the Modèle UTAC column. We need to extract useful features from Désignation commerciale. And also TVV

More depth to the research was added by Feature Engineering, which introduced ratios that capture efficiency and performance features of automobiles that potentially influence emissions. In our case, it is useful to create 'fuel\_efficiency', 'engine\_efficiency' and 'Power\_to\_Weight\_Ratio'. Those three columns were all added to the the dataset.

Secondly, some features needed transformations like ‘Carburant’, ‘ hybride’ and ‘Year’ etc… into suitable data type prior to modelling.

## **2.4. Normalizing Numerical Variables**

Normalizing Numerical Variables refers to the process of adjusting numerical data to be on a consistent scale or brought into a specific distribution. This step is crucial to ensure that numerical features are comparable and to avoid biases due to different scales or units.There are two common methods for normalizing numerical variables:

**Scaling to a Unit Interval (Min-Max Scaling):** This method transforms the data by scaling each value to a range between 0 and 1. The formula is:

=

Here, *X* is the original value, *X*new is the scaled value, and *X*min and *X*max are the minimum and maximum values of the variable.

**Standardization (Z-Score Normalization):** This method transforms the data to have a standard normal distribution (mean 0, standard deviation 1). The formula is:

=

In this case, each value *X* is divided by the mean of the variable and then by the standard deviation of the variable.

Normalization is particularly important in algorithms based on distance measures, such as k-nearest neighbors (k-NN) or gradient descent in neural networks. It ensures that all features are equally weighted and improves the convergence speed of optimization algorithms.

We used the Min-Max Scaling method to normalize the numerical variables. Min-Max Scaling ensures the preservation of relative distances between data points. This preservation of relationships is particularly meaningful when the connections between values play a crucial role, especially in applications like k-NN (k-nearest neighbors).

## **2.5. Outlier Detection and Treatment**

We used the Zscore method identify outliers and handled them via the winsorization technique.

The Z-score, also known as the standard score, measures how many standard deviations a data point is from the mean of a distribution. It is calculated using the formula:

*Z*=

where:

* *Z* is the Z-score,
* *X* is the individual data point,
* *μ* is the mean of the distribution, and
* *σ* is the standard deviation of the distribution.

And the winsorization is a statistical technique used to handle outliers in a dataset. Instead of removing extreme values, Winsorization involves replacing values beyond a certain threshold with the nearest value within that threshold. The idea is to limit the impact of extreme values on statistical analyses while retaining the information they provide. This method is named after the concept of "trimming" or "capping" the tails of a distribution, and it's particularly useful when a dataset has outliers that might skew the analysis.

The number of outliers in each column after treatment:

|  |  |  |  |
| --- | --- | --- | --- |
| **Colomn name** | **Outlier number** | **Colomn name** | **Outlier number** |
| Puissance administrative | 1041 | Marque\_QUATTRO | 1 |
| Puissance maximale (kW) | 1049 | Marque\_RENAULT | 155 |
| Consommation urbaine (l/100km) | 504 | Marque\_RENAULT TECH | 20 |
| Consommation extra-urbaine (l/100km) | 302 | Marque\_ROLLS-ROYCE | 16 |
| Consommation mixte (l/100km) | 394 | Marque\_SEAT | 105 |
| CO2 (g/km) | 263 | Marque\_SKODA | 364 |
| CO type I (g/km) | 573 | Marque\_SMART | 52 |
| NOX (g/km) | 3644 | Marque\_SSANGYONG | 9 |
| HC+NOX (g/km) | 617 | Marque\_SUBARU | 27 |
| Particules (g/km) | 59 | Marque\_SUZUKI | 31 |
| masse vide euro min (kg) | 298 | Marque\_TESLA | 2 |
| masse vide euro max (kg) | 122 | Marque\_TOYOTA | 111 |
| gamme\_ECONOMIQUE | 219 | Marque\_VOLKSWAGEN | 900 |
| gamme\_INFERIEURE | 1622 | Marque\_VOLVO | 106 |
| gamme\_LUXE | 0 | Carburant\_EE | 3 |
| gamme\_MOY-INF | 2 | Carburant\_EH | 199 |
| gamme\_MOY-INFER | 0 | Carburant\_EL | 39 |
| gamme\_MOY-SUPER | 0 | Carburant\_ES | 0 |
| gamme\_SUPERIEURE | 1956 | Carburant\_ES/GN | 23 |
| Marque\_ALFA-ROMEO | 103 | Carburant\_ES/GP | 17 |
| Marque\_ASTON MARTIN | 78 | Carburant\_FE | 8 |
| Marque\_AUDI | 242 | Carburant\_GH | 54 |
| Marque\_BENTLEY | 22 | Carburant\_GL | 1 |
| Marque\_BMW | 525 | Carburant\_GN | 59 |
| Marque\_CADILLAC | 44 | Carburant\_GN/ES | 23 |
| Marque\_CHEVROLET | 63 | Carburant\_GO | 0 |
| Marque\_CITROEN | 207 | Carburant\_GP/ES | 17 |
| Marque\_DACIA | 30 | Boîte de vitesse\_A 0 | 10 |
| Marque\_DANGEL | 40 | Boîte de vitesse\_A 4 | 39 |
| Marque\_FERRARI | 21 | Boîte de vitesse\_A 5 | 0 |
| Marque\_FIAT | 415 | Boîte de vitesse\_A 6 | 1031 |
| Marque\_FORD | 296 | Boîte de vitesse\_A 7 | 0 |
| Marque\_HONDA | 51 | Boîte de vitesse\_A 8 | 446 |
| Marque\_HYUNDAI | 52 | Boîte de vitesse\_D 5 | 39 |
| Marque\_INFINITI | 26 | Boîte de vitesse\_D 6 | 11 |
| Marque\_JAGUAR | 43 | Boîte de vitesse\_D 7 | 21 |
| Marque\_JAGUAR LAND ROVER LIMITED | 55 | Boîte de vitesse\_M 5 | 1191 |
| Marque\_JEEP | 74 | Boîte de vitesse\_M 6 | 0 |
| Marque\_KIA | 78 | Boîte de vitesse\_M 7 | 12 |
| Marque\_LADA | 9 | Boîte de vitesse\_N 0 | 1 |
| Marque\_LAMBORGHINI | 16 | Boîte de vitesse\_N 1 | 2 |
| Marque\_LANCIA | 37 | Boîte de vitesse\_S 6 | 1 |
| Marque\_LEXUS | 175 | Boîte de vitesse\_V 0 | 180 |
| Marque\_LOTUS | 26 | Carrosserie\_BERLINE | 0 |
| Marque\_MASERATI | 8 | Carrosserie\_BREAK | 2229 |
| Marque\_MAYBACH | 13 | Carrosserie\_CABRIOLET | 611 |
| Marque\_MAZDA | 46 | Carrosserie\_COMBISPACE | 901 |
| Marque\_MERCEDES AMG | 174 | Carrosserie\_COUPE | 1104 |
| Marque\_MERCEDES-BENZ | 0 | Carrosserie\_MINIBUS | 0 |
| Marque\_MIA | 21 | Carrosserie\_MINISPACE | 147 |
| Marque\_MINI | 79 | Carrosserie\_MONOSPACE | 115 |
| Marque\_MITSUBISHI | 39 | Carrosserie\_MONOSPACE COMPACT | 610 |
| Marque\_NISSAN | 173 | Carrosserie\_TS TERRAINS/CHEMINS | 1176 |
| Marque\_OPEL | 520 |  |  |
| Marque\_PEUGEOT | 160 |  |  |
| Marque\_PORSCHE | 89 |  |  |

## **2.7. Splitting Data**

We split the dataset into training, validation, and test sets to evaluate the performance of our machine learning models properly in the next sprints.

## **2.7. Feature Selection**

Feature selection is the process of choosing a subset of relevant features from the original set of features to improve model performance, reduce overfitting, and decrease computational cost. For now, we can use methods like Correlation Matrix or Feature importance methods (e.g. decision trees and random forests) to identify features with higher importance scores.