DATA UPSKILLING PROGRAM
REPORT STEP 2 – PRE-PROCESSING AND FEATURE ENGINEERING

PROJECT MENTOR		
Antoine TARDIVON	Data Scientist	DataScientest
COHORT		
Clement ARNAUD	Process Engineer	CFT TEN - PARIS
Diego GOMEZ-OCHOA	Process Engineer	REFINING TEN - PARIS
Presheet DESHPANDE	Technical Safety & Risk Engineer	GENESIS - LONDON
Reginaldo MARINHO	Process Engineer	CFT TEN - PARIS
Simran MASOOD	Process Engineer	CFT TEN - PARIS
NAME	***POSITION***	***DEPARTMENT – CENTER***

Contents

1. CONTEXT

Global transportation sector is a major contributor to greenhouse gas emissions, with passenger cars and vans responsible for around 10% of global energy-related CO_2 emissions in 2022 according to International Energy Agency (IEA). This substantial emission rate significantly affects air quality and contributes to climate change. Therefore, identifying the vehicles emitting the most CO_2 and other pollutants is crucial for devising effective strategies to mitigate environmental impact. As automotive technology evolves, understanding the role of technical characteristics with respect to emissions is vital for promoting the development and adoption of cleaner and more efficient vehicles ultimately contributing to the realization of the Net Zero Emissions goals by 2050.

This project explores two datasets (given below) encompassing a wide array of technical specifications of vehicles, alongside their fuel consumption, CO_2 emissions, and pollutant emissions, marketed both in France and Europe. Through the application of Data Science and Machine Learning techniques, our objective is to explore the relationship between vehicle specifications and emissions. By doing so, we aim to provide valuable insights that can inform decision-making processes in environmental policy and drive advancements in automotive industry practices towards sustainable transportation solutions.

The following datasets are provided for reference:

- data.gouv.fr
- European Environment Agency

This project employs a combination of data analysis, statistical modeling, and machine learning techniques to extract actionable insights from the dataset. Exploratory data analysis (EDA) will uncover patterns and relationships within the data, providing a foundational understanding of the variables at play. Feature engineering will involve transforming or selecting relevant variables to enhance model performance. Lastly, statistical modeling techniques, such as linear regression, will help quantify the impact of technical characteristics of vehicles on CO_2 emissions. Additionally, machine learning algorithms, such as decision trees or random forests, or ensemble learning algorithms such as Bagging and Boosting may be utilized for better predictive performance of the model.

2. OBJECTIVES

The main objectives are:

- Perform the pre-processing and cleaning of the data set.
- Add new features to the data set where possible (Feature Engineering).
- End up with a final data set ready for modelling.

An initial pre-processing was performed in the data first report of this project (refer to *'REPORT STEP 1-DATA MINING + DATAVIZ'*). In the previous report we also defined an initial list of pre-processing (labelled as action list) and feature engineering we intended to do in the following phase. This is what we will develop in this report.

The lists made in the previous report were from an initial analysis of the data set. Based on the findings on this phase of the project we may choose to implement them or not, as will be shown in the following sections. Also, when trying to implement these actions / features we found new actions to implement. Finally, the actions and features listed in the previous report are fully explored but not necessarily in the same order presented in the precious report.

3. INITIAL DATA SET

The initial data set information is shown below:

```
<class 'pandas.core.frame.DataFrame'>
Index: 159780 entries, 0 to 40051
Data columns (total 26 columns):
 #   Column             Non-Null Count   Dtype     % Missing values
---  ------             --------------   -----                -----
 0   lib_mrq_utac       159780 non-null  object            0.000000
 1   lib_mod_doss       159780 non-null  object            0.000000
 2   lib_mod            159780 non-null  object            0.000000
 3   dscom              159780 non-null  object            0.000000
 4   cnit               159780 non-null  object            0.000000
 5   tvv                159780 non-null  object            0.000000
 6   cod_cbr            159780 non-null  object            0.000000
 7   hybride            159780 non-null  object            0.000000
 8   puiss_admin_98     159780 non-null  float64           0.000000
 9   puiss_max          159724 non-null  float64           0.035048
 10  typ_boite_nb_rapp  159780 non-null  object            0.000000
 11  conso_urb          159543 non-null  float64           0.148329
 12  conso_exurb        159543 non-null  float64           0.148329
 13  conso_mixte        159622 non-null  float64           0.098886
 14  co2                159622 non-null  float64           0.098886
 15  co_typ_1           159090 non-null  float64           0.431844
 16  hc                 36813 non-null   float64          76.960195
 17  nox                159090 non-null  float64           0.431844
 18  hcnox              122452 non-null  float64          23.362123
 19  ptcl               150181 non-null  float64           6.007635
 20  masse_ordma_min    159780 non-null  float64           0.000000
 21  masse_ordma_max    159780 non-null  float64           0.000000
 22  champ_v9           159595 non-null  object            0.115784
 23  date_maj           68352 non-null   object           57.221179
 24  Carrosserie        138900 non-null  object           13.067968
 25  gamme              138900 non-null  object           13.067968
dtypes: float64(13), object(13)
memory usage: 32.9+ MB
```

4. PRE-PROCESSING

4.1. Fuel Type correction

Some of the hybrid cars are listed as Gasoline only (ES) or Diesel only (GO), which is not possible. They are, in reality, Non-Plug-In Hybrid (EH) and Non-Plug-In Electric Diesel (GH) cars respectively. This information is found by looking for the same vehicles in previous years (using the tvv).

4.1.1. 'GO' hybrid cars

After looking for the cars with same tvv in the FRENCH database itself, we realized that the GO hybrid cars are GH cars. We apply his change to our dataset, considering that it was wrongly entered.

4.1.2. 'ES' hybrid cars

After looking for the cars with same tvv in the FRENCH database itself we found out that most of the cars have EH fuel type records for the same tvv. When comparing the only 'ES' cars (14 out of the 105 ES hybrids) with the EEA cars with same tvv, we found that many of them are entered also with petrol fuel type. We consider that these cars where wrongly entered, and we correct 'ES' to 'EH' for these cars.

4.1.3. Number of cars using each Fuel Type

Before moving forward, we have an action from Section 5.1 – Actions List of the DATA MINING + DATAVIZ Report:

"Identify the number of car brands using each fuel type and compare the proportions to get a bigger picture for why CO2 emissions differ between each brand."

The count of the values for each fuel type is show below:

```
GO        134646
ES         22566
EH          1148
GH           548
GN/ES        232
EL           158
ES/GP        152
GN           152
EE           104
FE            71
GL             3
```

We can see that most of the cars are GO and ES. This can create an imbalanced problem when trying to estimate the co2 emissions of minorities like 'FE' and 'GL' as these cars that also have very lower average co2 emissions (refer to previous report).

A catplot of co2 emissions vs brand and differentiated by fuel type is show below in section 7.2. It confirms most of the cars in most of the brands are 'GO' and 'ES', and the lower consuming cars are 'EE' and 'GL' cars. Please note that the brand is not a explanatory variable, as explained in section 6.

4.2. hcnox correction

In addition to Hydrocarbon (hc) and NOx (nox) emissions, a third feature is reported, and it corresponds to the sum of them called hcnox.

Those three features have NAN values, and they are different in number as follow:

```
hc        122967
nox          690
hcnox      37328
```

It is proposed for each missing value in hcnox to perform hc + nox if both are available (no missing values).

Once the preprocessing is performed, the NAN distribution is as follows:

```
hc        122967
nox          690
hcnox        692
```

For 690 cars only one parameter is specified either hc or nox and for two of them only the nox is specified. Calculating back the hc for missing values, the resulting 'hc' missing values is 692. The remaining records are missing the 3 variables (hc, nox and hcnox). We therefore drop them. At least 158 of them are electric only cars, that should be drop anyways, as they always present 'co2' emissions equal to zero (so there is no modelling required for co2 emissions for electric cars).

The variable 'hc' is then back calculated as hcnox – nox for all hc values that are missing. A problem found in this process was getting negative 'hc' values, as some of the 'hcnox' where lower than nox or even zero. For these cases, hcnox was set to nox value and hc was set to zero. Finaly, records with all values equal to zero (8 cases) where dropped. This solved the negative 'hc' values problem. A box plot for this variable is shown in section 7.1.

4.3. Gamme and Carrosserie

There are too many values missing that cannot be recovered form EEA dataset. We choose to drop these two variables instead of missing 20k records.

Variable drop:

> The ***Gamme*** *and* ***Carrosserie*** have many missing values that cannot be recovered.

4.4. Mass in min/max Walking Order: masse_ordma_min and masse_ordma_max

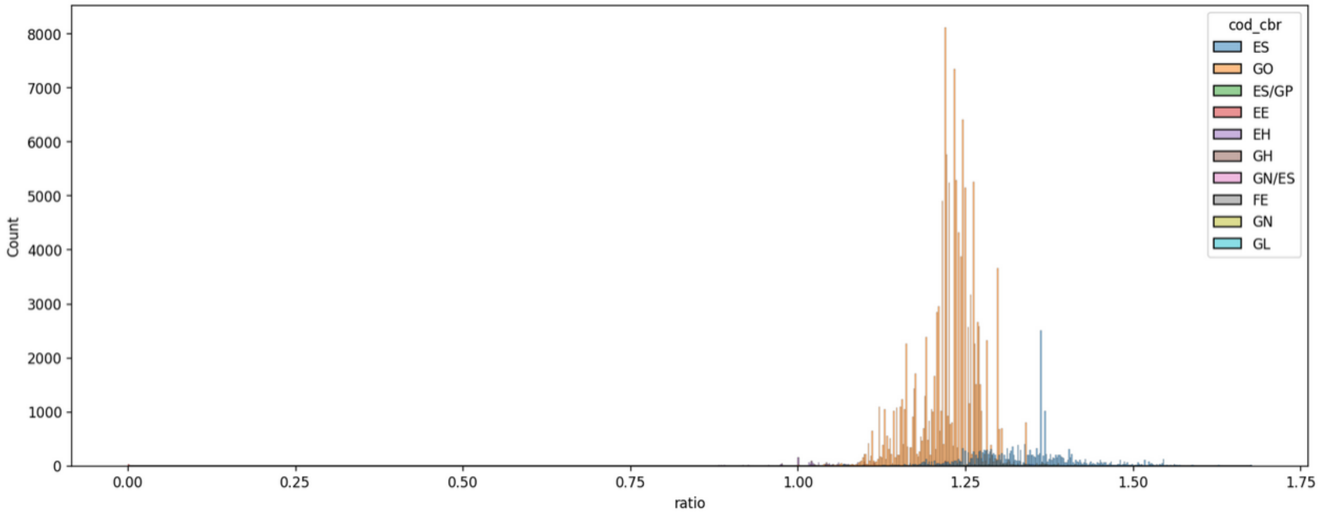
There are no missing values for these variables. However, they are very close in values. We most of them being less then 2% apart from each other. We consider that they are the same information, and we will keep only the 'masse_ordma_max' variable.

Variable drop:

> The ***masse_ordma_min*** has the same information as ***masse_ordma_max***, we drop the first one.

4.5. Fuel Consumption: conso_urb, conso_exurb, conso_mixte

To find out if there is a relationship (ratio) between conso_urb and conso_mixte, we can plot a graph of the ratios per fuel type:



There is no apparent direct correlation. So, we cannot recover the 79 records missing for 'conso_rub', same applies for conso_exurb. We will keep only, 'conso_mixte', as there is no value missing.

4.6. Particle Trial Results Measurement: *ptcl*

The 'ptcl' variable cannot be recovered from the EEA dataset and was found to have low correlation with CO2 emissions in previous report. Additionally, almost half of the values of this variable are 0.

We will choose to drop this column instead of losing 9 000 rows of data.

Variable drop:

The ***ptcl*** has low correlation with the target variable (refer to previous report). And many missing values.

4.7. Update date_maj

We will drop this column as the information we needed was the year and was already added from the dataset name.

Variable drop:

The information we want from ***date_maj*** is already in the year column.

4.8. Duplicate check

Duplicate are removed from the data set. Note that 'date_maj' and 'champ_v9' are not considered as same car can have different values for these two features, so if they are taken into account, some duplicates could remain in the dataset. However, as proposed in § 4.7 and 5.4 these two features are removed.

5. FEATURE ENGINEERING

5.1. Extract engine power and cylinder volume from dscom

It is possible to retrieve the engine power and the cylinder volume from the dscom column of the vehicle, as highlighted below:

```
0                            159 1750 Tbi (200ch)
1                        159 2.0 JTDm (170ch) ECO
2                            159 2.0 JTDm (136ch)
3                            159 2.0 JTDm (136ch)
4                            159 2.0 JTDm (170ch)
                            ...
159764                         500 1.4 16V (100ch)
159765                            500C 1.4 (100ch)
159766                  500 1.4 16V Dualogic Euro 5
159767                          500 1.4 16V Euro 5
159779    Delta 1.9 MultiJet Twinturbo (190ch) DPF
Name: dscom, Length: 103256, dtype: object
```

- Engine power

When we extract the engine power, however, and compare with the variable puiss_max we realize it is the same information. For example, for the first record 200ch = 147 kW. This feature thus already exists in the data set.

- Cylinder volume

We can see that not all the records have a cylinder volume. Indeed, the first 5 and last 5 records of the data set is not representative of the entire data. After extraction we end up with 71 828 missing values, which is more than 50%. Finaly, this value cannot be directly retrieved from EEA data set (this is further discussed in section 5.3). This feature, therefore, cannot be used for our model.

Variable drop:

The ***dscom*** variable has many unique values and cannot easily be used in the modelling, we will therefore drop it.

5.2. typ_boite_nb_rapp

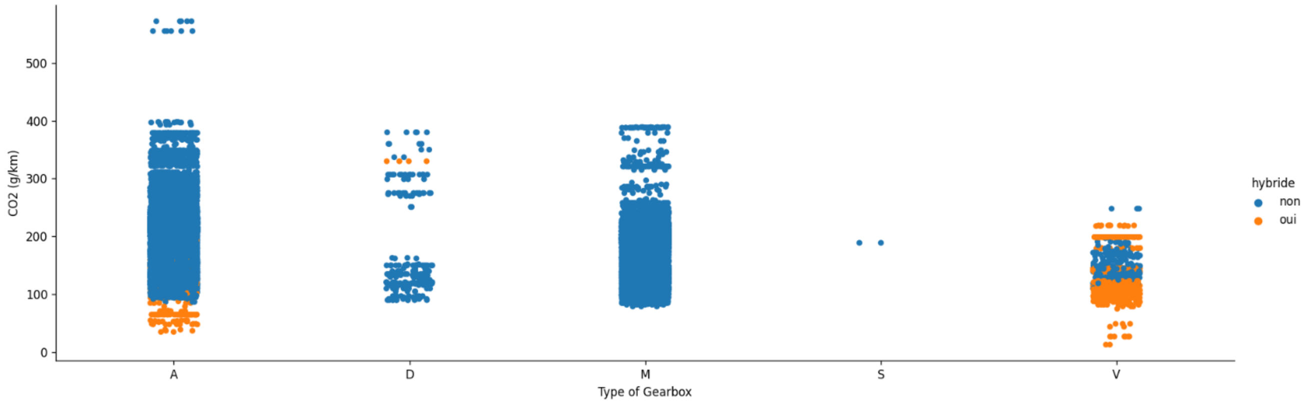
The variable typ_boite_nb_rapp is composed of two parts:

- A letter indicating the type of gearbox
- A number indicating the number of reports

```
0         M 6
1         M 6
2         M 6
3         M 6
4         M 6
         ...
159764    M 6
159765    M 6
159766    M 5
159767    M 6
159779    M 6
Name: typ_boite_nb_rapp, Length: 103256, dtype: object
```

- Type of gearbox

If we separate the letter and plot the CO2 emissions for each type of gearbox, we have the plot below:



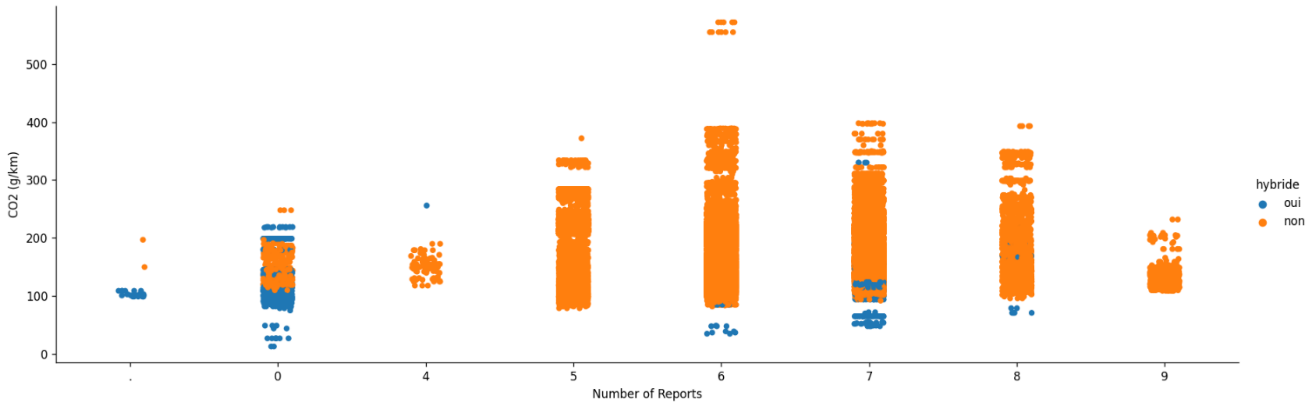
We can see that they can impact the value and the distribution of the CO2 emissions. We will then keep this feature.

Feature:

The ***type_of_gearbox*** feature is added to the data set.

- Number of reports

If we separate the number of reports and plot the CO2 emissions for each value, we have the plot below:



We can see that they can impact the value and the distribution of the CO2 emissions. We will then keep this feature.

Feature:

The ***nbr_reports*** feature is added to the data set.

We no longer need the variable ***typ_boite_nb_rapp*** as the new two added features describe better the data set. We will then drop it:

Variable drop:

The ***typ_boite_nb_rapp*** is no longer required, we drop it.

5.3. Add variables from EEA

The EEA data set has some additional information about the vehicles as axel length, wheelbase length etc. This data can be added by using the tvv as an unique identifier.

When the merging is completed, however, there are 74 604 missing values. These features, therefore, cannot be added to the data set.

5.4. champ_v9

The ***champ_v9*** corresponds to the certification of the vehicle. We could then separate them into certified and uncertified classes.

However, after pre-processing (refer to section 1), we end up with 0 missing values for ***cham_v9*** variable. We cannot therefore implement this feature.

Variable Drop

The variable ***cham_v9*** has no identified use for our modelling and presents too many unique values, we will drop it.

6. REMAINING VARIABLES TO DROP

As discussed in previous report, we also drop the following variables:

- ***'mrq_utac***': high quantity of missing values, as not present in all years.
- '***puiss_admin_98***': can be used to back calculate co2 emissions, no sense to make a ML model if we have this value.

Additionally, 'lib_mod_doss' and 'lib_mod' have many values and cannot be easily associated with the co2 emissions, we will drop these columns.

'lib_mrq_utac' will be kept only for exploratory visualization of the results, as it cannot be used neither in the model due to its high number of unique values.

7. FINAL DATA SET

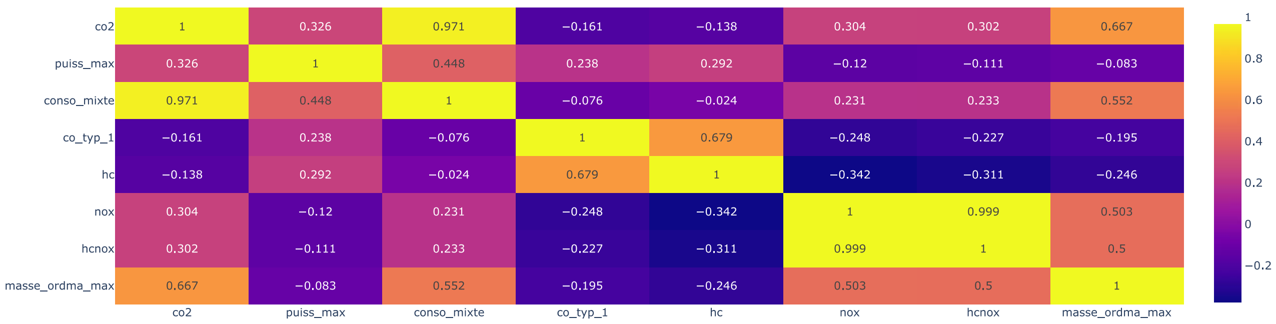
The final data set information is shown below and is saved for modelling under the name: '***data_phase_2.csv***'.

```
<class 'pandas.core.frame.DataFrame'>
Index: 103248 entries, 0 to 159779
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   lib_mrq_utac    103248 non-null  object
 1   cod_cbr         103248 non-null  object
 2   hybride         103248 non-null  object
 3   puiss_max       103248 non-null  float64
 4   conso_mixte     103248 non-null  float64
 5   co2             103248 non-null  float64
 6   co_typ_1        103248 non-null  float64
 7   hc              103248 non-null  float64
 8   nox             103248 non-null  float64
 9   hcnox           103248 non-null  float64
 10  masse_ordma_max 103248 non-null  float64
 11  year            103248 non-null  int64
 12  type_of_gearbox 103248 non-null  object
 13  nbr_reports     103248 non-null  object
dtypes: float64(8), int64(1), object(5)
memory usage: 11.8+ MB
```

7.1. Quantitative Variables

Let's have a quick look in the correlation between the variables after finishing the cleaning and the feature engineering.

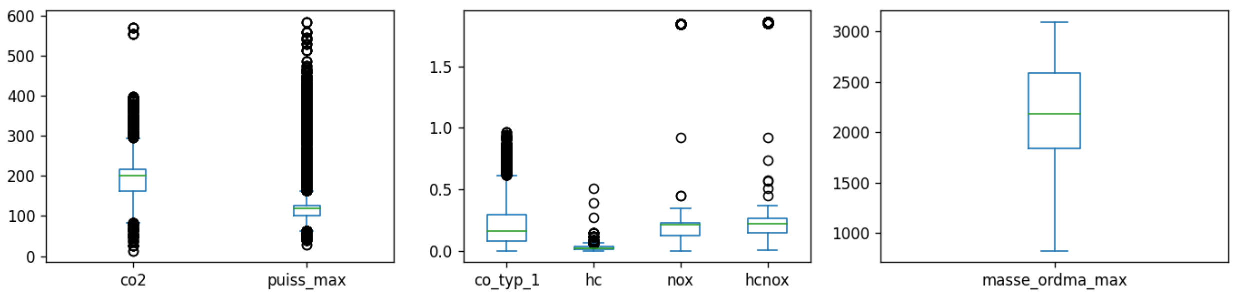
CORRELATION MATRIX



We can see that all the features but ***co_type_1*** are correlated with the target variable. We will nevertheless keep all variables for our modelling.

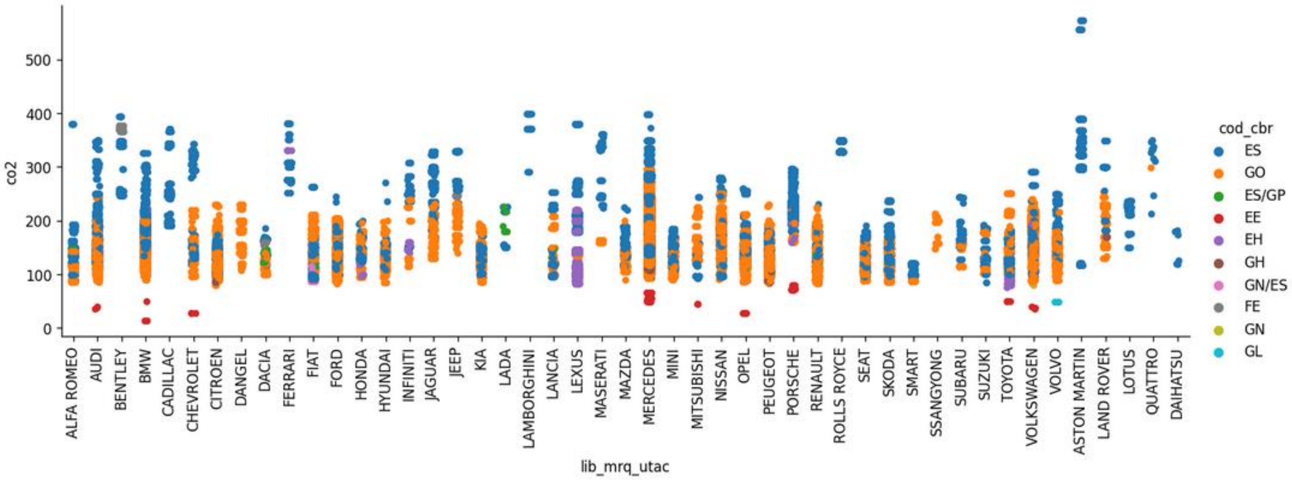
The year is an added numerical variable but is intended to be used only to separate the train and test sets. Training being the older years and test the latest available year.

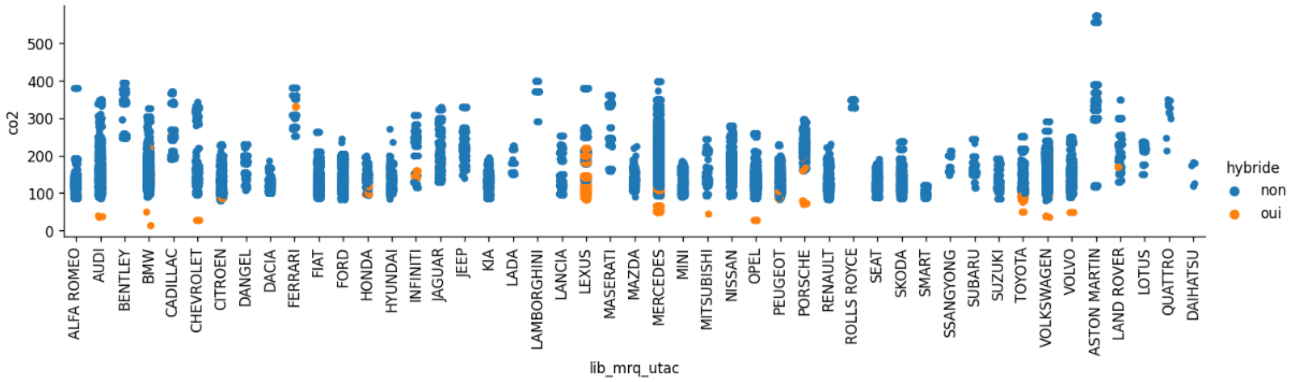
The image below shows the distribution of each quantitative variable. For more details refer to the previous report.

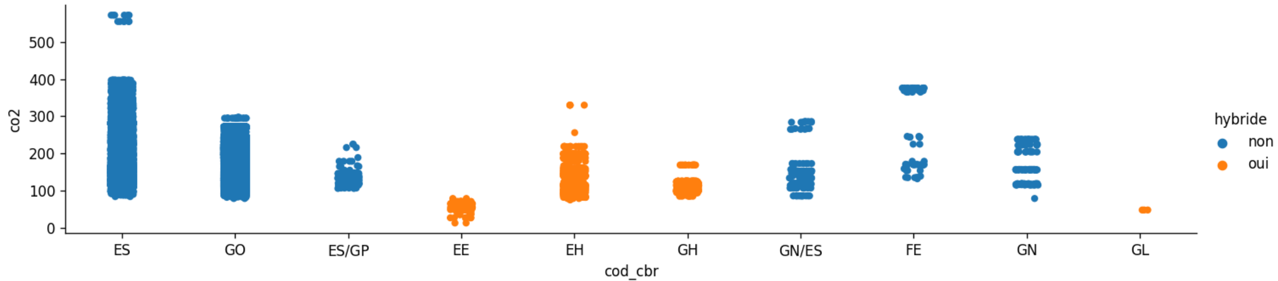


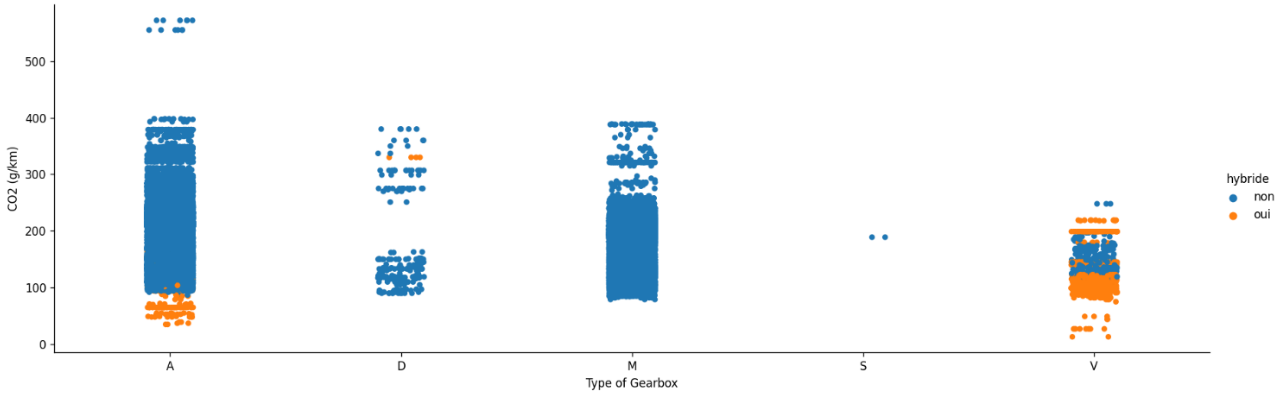
7.2. Categorical Variables

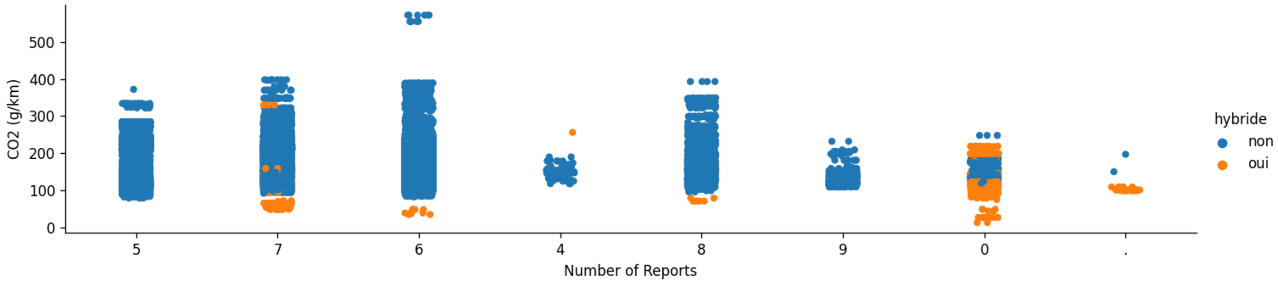
Here under the distribution of CO2 emissions for each categorical variable:











8. CONCLUSION AND WAY FORWARD

The actions and the feature engineering identified in previous phase where all implemented. More deep analysis for each variable was made to try to keep a maximum of data. Some extra pre-processing may be done to the dataset if any problem in found during modelling, but for now we have a dataset ready for the modelling phase.