DATA UPSKILLING PROGRAM
REPORT STEP 1 - DATA MINING + DATAVIZ

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.

1.1. Selection of dataset

To begin our metadata analysis, we have opted to start with the initial dataset sourced from data.gouv.fr. Our selection of the French dataset over the European dataset is influenced by two factors. Firstly, upon preliminary examination, we observed that the French dataset offers a wider array of explanatory variables pertinent to the project's scope. Notably, it provides a detailed breakdown of fuel consumption across urban, extra-urban, and mixed driving conditions, alongside comprehensive data concerning other emissions such as NOx, CO, HC, and particulates. Additionally, the French dataset includes information on the car's body type and range enriching the depth of our analysis.

However, it is worth noting that the European Environment Agency (EEA) dataset does contain supplementary technical characteristics of vehicles, such as wheelbase, track width, and other dimensions, which may be of interest for future stages of our preprocessing efforts.

Within the French dataset, our focus centered on the most recent four years, spanning from 2012 to 2015, for our preliminary analysis. This selection was motivated by the emergence of hybrid vehicles, which began to appear prominently from 2011 onwards. To ensure that these vehicles were included in our analysis, we deemed it necessary to limit our dataset to the years 2011 and beyond. However, we made an exception for the year 2011 due to the limited availability of significant explanatory variables, including data on NOx, CO, and HC emissions, particulate emissions, mileage, body type of the car, and vehicle mass.

Another important step of our thinking was the discover of two norms to measure the CO_2 emissions: the New European Driving cycle (NEDC) and the Worldwide Harmonized Light Vehicle (WLTP). The first one is an older way to standardized the way to measure the CO_2 emissions of a car between all the different passenger vehicles. As of 1st September 2017, a new standard has been launched to provide more realistic measurements. It means that to have comparable values, it is not recommended to merge older dataset dating from before 2017 with most recent dataset. For this reason, we have decided to focus especially on dataset between 2012 and 2015 where the CO_2 emissions is measured with the NEDC norm. We have also decided to not have a too large number of years available as we expect that the technology and the legislation norms evolve each years and may impact the prediction.

2. OBJECTIVES

The purpose of this report is to have a first overview of the data we are going to use for the project.

The main objectives are:

- Interrogate the data we are working with and understand what each variable means.
- Homogenise and merge data from different years.
- Perform an initial clean up and check for missing values and duplicate values.
- Identify relationships between target variable and features using DataViz.
- Start looking for correlations.
- Identify potential features to be implemented in the feature engineering phase.
- Identify incoherences to be corrected.

3. DATA MINING

3.1. Data Structure

The Table 1 below shows the variable names and the corresponding descriptions provided by the French government website. We can see that the variables are not named the same across the available years (2015 and before). We choose to align all the names with the ones provided in the description of the website (column 'legend' here-under).

Table 1. Variable names and corresponding columns by dataset

#	nom-colonne	legend	unit	2015	2014	2013	2012	2011
1	lib_mrq_utac	The brand	-	lib_mrq_doss	lib_mrq	Marque	lib_mrq	lib_mrq
2	lib_mod_doss	The file model	-	lib_mod_doss	lib_mod_doss	Modèle dossier	lib_mod_doss	lib_mod_doss
3	lib_mod	The business model	-	mod_utac	lib_mod	Modèle UTAC	lib_mod	lib_mod
4	dscom	Commercial designation	-	dscom	dscom	Désignation commerciale	dscom	dscom
5	cnit	The National Type Identification Code (CNIT)	-	cnit	cnit	CNIT	cnit	cnit
6	tvv	The Variant-Variant (TVV) or the Mines type	-	tvv	tvv	Type Variante Version (TVV)	tvv	tvv
7	cod_cbr	The type of fuel	-	energ	cod_cbr	Carburant	typ_cbr	typ_cbr
8	hybride	Information to identify hybrid vehicles (O/N)	-	hybride	hybride	Hybride	hybride	NA
9	puiss_admin_98	Administrative power	-	puiss_admin	puiss_admin_98	Puissance administrative	puiss_admin_98	puiss_admin_98
10	puiss_max	Maximum power (in KW)	kW	puiss_max	puiss_max	Puissance maximale (kW)	puiss_max	puiss_max
11	typ_boite_nb_rapp	The type of gearbox and the number of reports,	-	typ_boite_nb_rapp	typ_boite_nb_rapp	Boîte de vitesse	typ_boite_nb_rapp	typ_boite_nb_rapp
12	conso_urb	Urban fuel consumption (in L/100km),	liter for 100 km	conso_urb_93	conso_urb	Consommation urbaine (l/100km)	conso_urb	conso_urb
13	conso_exurb	Mixed fuel consumption (in L/100km),	liter for 100 km	conso_exurb	conso_exurb	Consommation extra-urbaine (l/100km)	conso_exurb	conso_exurb
14	conso_mixte	Extra urban fuel consumption (in L/100km),	liter for 100 km	conso_mixte	conso_mixte	Consommation mixte (l/100km)	conso_mixte	conso_mixte
15	co2	CO2 emission (in G/km),	gram per km	co2_mixte	co2	CO2 (g/km)	co2	co2
16	co_typ_1	CO Type I test result	gram per km	co_typ_1	co_typ_1	CO type I (g/km)	co_typ_1	NA
17	hc	Results of test HC	gram per km	hc	hc	HC (g/km)	hc	NA
18	nox	Nox trial results	gram per km	nox	nox	NOX (g/km)	nox	NA
19	hcnox	HC+Nox trial results	gram per km	hcnox	hcnox	HC+NOX (g/km)	hcnox	NA
20	ptcl	particle test result	gram per km	ptcl	ptcl	Particules (g/km)	ptcl	NA
21	masse_ordma_min	The mass in mini walking order	kg	masse_ordma_min	masse_ordma_min	masse vide euro min (kg)	masse_ordma_min	NA
22	masse_ordma_max	the mass in maximum walking order	kg	masse_ordma_max	masse_ordma_max	masse vide euro max (kg)	masse_ordma_max	NA
23	champ_v9	Field V9 of the registration certificate which contains the Euro standard	-	champ_v9	champ_v9	Champ V9	champ_v9	champ_v9
24	date_maj	The date of the last update.	-	date_maj	date_maj	Date de mise à jour	date_maj	date_maj
25	Carrosserie	Body	-	-	Carrosserie	Carrosserie	Carrosserie	NA
26	gamme	Range	-	-	gamme	gamme	gamme	NA

We note that many of these variables are missing from the datasets of years before 2012. For this project we will then choose to work with the data from 2012 to 2015.

3.2. Initial cleanup

After aligning the column names of the data sets from 2012 to 1015 and concatenating them we end up with the following data frame:

```
<class 'pandas.core.frame.DataFrame'>
Index: 160826 entries, 0 to 40051
Data columns (total 32 columns):
 #   Column              Non-Null Count   Dtype     % Missing values
---  ------              --------------   -----                -----
 0   lib_mrq_utac        160826 non-null  object          0.000000
 1   lib_mod_doss        160826 non-null  object          0.000000
 2   mrq_utac            20880 non-null   object         87.017025
 3   lib_mod             160826 non-null  object          0.000000
 4   dscom               160826 non-null  object          0.000000
 5   cnit                160826 non-null  object          0.000000
 6   tvv                 160826 non-null  object          0.000000
 7   cod_cbr             160826 non-null  object          0.000000
 8   hybride             160826 non-null  object          0.000000
 9   puiss_admin_98      160826 non-null  int64           0.000000
 10  puiss_max           160770 non-null  object          0.034820
 11  puiss_heure         895 non-null     float64        99.443498
 12  typ_boite_nb_rapp   160826 non-null  object          0.000000
 13  conso_urb           160588 non-null  object          0.147986
 14  conso_exurb         160588 non-null  object          0.147986
 15  conso_mixte         160667 non-null  object          0.098865
 16  co2                 160667 non-null  float64         0.098865
 17  co_typ_1            159943 non-null  object          0.549041
 18  hc                  37430 non-null   object         76.726400
 19  nox                 159943 non-null  object          0.549041
 20  hcnox               122688 non-null  object         23.713827
 21  ptcl                150599 non-null  object          6.359046
 22  masse_ordma_min     160826 non-null  int64           0.000000
 23  masse_ordma_max     160826 non-null  int64           0.000000
 24  champ_v9            160448 non-null  object          0.235037
 25  date_maj            68977 non-null   object         57.110791
 26  Carrosserie         139946 non-null  object         12.982975
 27  gamme               139946 non-null  object         12.982975
 28  Unnamed: 26         0 non-null       float64       100.000000
 29  Unnamed: 27         0 non-null       float64       100.000000
 30  Unnamed: 28         0 non-null       float64       100.000000
 31  Unnamed: 29         0 non-null       float64       100.000000
dtypes: float64(6), int64(3), object(23)
memory usage: 40.5+ MB
```

For all the cleaning process refer to 'DATA MINING+DATAVIZ.ipynb'

We can see that there are many missing values. We will drop the columns that are not in the initial description (which also happens to be the ones that are empty or with a lot of missing values) These columns are the following:

- `'mrq_utac'`, `'puiss_heure'`, `'Unnamed: 26'`, `'Unnamed: 27'`, `'Unnamed: 28'` and `'Unnamed: 29'`

We will now only have the columns presented in the Variable names section.

There are also variables which are duplicated (1 046 in total) that are dropped from the data frame by keeping the last one.

Finally, some variables are still in represented as objects, but should be numbers. We will look in detail each variable in the following sections, performing an initial cleanup, and identifying the clean-up / merges that could be required for the next phase.

To do so, we will separate them in two groups:

- Categorical Variables
- Quantitative Variables

3.3. Categorical Variables

A lot of the qualitative variables have many different values because of a trailing space. These spaces have been removed for this preliminary analysis.

3.3.1. Variable names and preliminary visualization

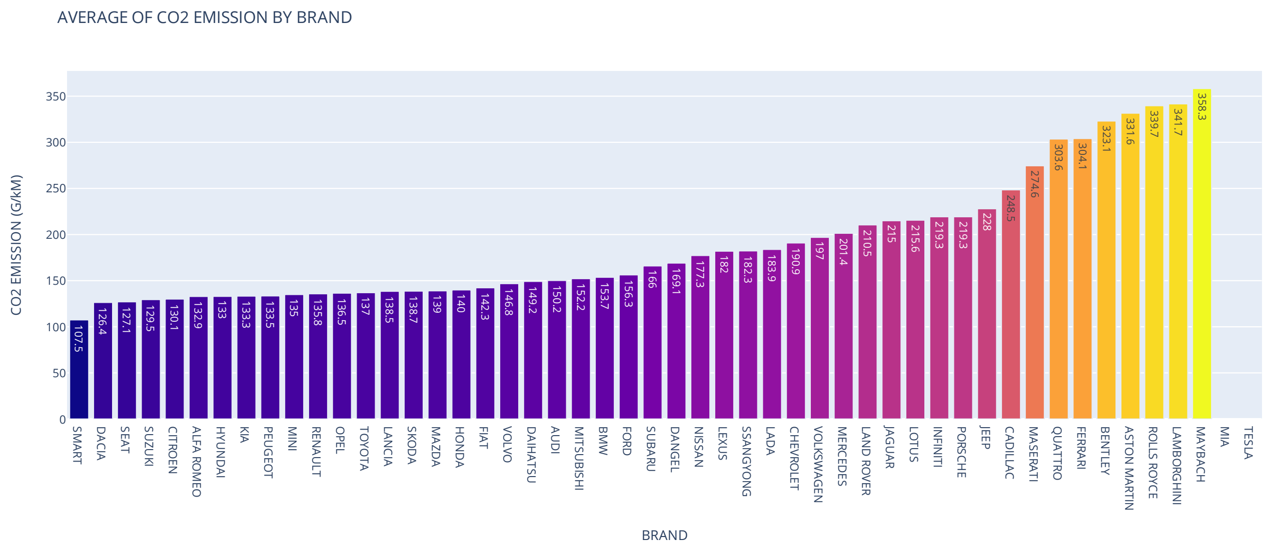
- **`lib_mrq_utac`**

The brand of the car manufacturer (the real brand of the car is indicated and not the group of manufacturers i.e Peugeot is indicated and not the PSA group which regroup different brands like Peugeot, Citroën, etc…)

Unique Values

```
000: ALFA ROMEO     001: ALFA-ROMEO                 002: ASTON MARTIN    003: AUDI          004: BENTLEY
005: BMW            006: BMW I                      007: CADILLAC        008: CHEVROLET     009: CITROEN
010: DACIA          011: DAIHATSU                   012: DANGEL          013: FERRARI       014: FIAT
015: FORD           016: FORD-CNG-TECHNIK           017: HONDA           018: HYUNDAI       019: INFINITI
020: JAGUAR         021: JAGUAR LAND ROVER LIMITED  022: JEEP            023: KIA           024: LADA
025: LAMBORGHINI    026: LANCIA                     027: LAND ROVER      028: LEXUS         029: LOTUS
030: MASERATI       031: MAYBACH                    032: MAZDA           033: MERCEDES      034: MERCEDES AMG
035: MERCEDES BENZ  036: MERCEDES-BENZ              037: MIA             038: MINI          039: MITSUBISHI
040: NISSAN         041: OPEL                       042: PEUGEOT         043: PORSCHE       044: QUATTRO
045: RENAULT        046: RENAULT TECH               047: ROLLS ROYCE     048: ROLLS-ROYCE   049: SEAT
050: SKODA          051: SMART                      052: SSANGYONG       053: SUBARU        054: SUZUKI
055: TESLA          056: TOYOTA                     057: VOLKSWAGEN      058: VOLVO
```

We can see that some brands highlighted in blue are repeated with similar names. They have been grouped before this analysis (refer to notebook).



The bar chart above shows the average release of CO2 emissions (g/km) from 50 brands of cars between the year 2012 and 2015. The average release of CO2 emissions ranges from 107.5 g/km to 358.3 g/km. We can see that certain brands produce cars with lower CO2 emissions. The five cars that produce the least CO2 emissions are namely 'Smart', 'Dacia', 'Seat', 'Suzuki' and 'Citroën'. On the other hand, the five cars that produce the most CO2 emissions are 'Maybach', 'Lamborghini', 'Rolls Royce', 'Aston Martin', and 'Bentley'. Of course, MIA and Tesla do not show values as they electric cars and do not give off CO2 emissions while on the road.

It is worth noting that the bottom five cars use 'GO', 'ES', 'GH' and 'FE' fuel types, which are diesel, gasoline, non-plug-in electric diesel and E85 super-ethanol respectively. In contrast, the top five cars all use gasoline ('ES') fuel types.

However, further investigation will be required to explain the stark differences in CO2 emissions between different car brands. An initial hypothesis can be put forward that the following variables will influence the average release of CO2 emissions the most:

- masse_ordma_max & masse_ordma_min – Mass in max & min walking order (kg).
- cod_cbr – Fuel Type of the car.
- dscom – Commercial designation, which includes the engine volume, engine power and gearbox parameters.
- hybride – Information to identify hybrid vehicles.
- conso_urb/exurb/mixte – Various fuel consumptions.

This hypothesis will be tested as we further explore and investigate all the variables in the dataset during the next phase of the project.

- **`lib_mod_doss:`**

The file name of the car

Unique Values:

```
000: 107          001: 108              002: 159          003: 2008          004: 206+          005: 207
006: 208          007: 2171             008: 2172         009: 3008          010: 308           011: 370Z
...
582: XKR COUPE    583: XKR-S CONVERTIBLE  584: XKR-S COUPE   585: YARIS         586: YARIS HYBRID  587: YETI
588: YPSILON      589: ZAFIRA           590: ZAFIRA TOURER  591: ZOE
```

Many unique values. To be further studied later as some info are common other columns.

- **`lib_mod`**

The business name of the car: The manufacturer chooses the business name which will be written on the vehicle registration document. In the dataset, the file name of the car can be identical to the business name but can also be slightly different (i.e. for the same car: lib_mod_doss name = AR8C SPIDER and lib_mod = 8C SPIDER)

Unique Values:

```
000: 107      001: 108      002: 114     003: 116        004: 118    005: 120
006: 123      007: 125      008: 130     009: 135        010: 159    011: 2
...
516: XV       517: YARIS    518: YETI    519: YPSILON    520: Z4     521: ZAFIRA
522: ZAGATO   523: ZOE
```

Many unique values. To be further studied later as some info are common other columns.

- **`dscom`**

Commercial Designation: It regroups different information about the car model i.e. : 3008 1.6 THP (156ch) BVM6 "3008" is the business name of the car "1.6" is the volume of all engine cylinders, here it's 1.6 liters or 1600 cm3 "THP" is a name of a motor brand (Turbo High Pressure) "(156ch)" is the power of an engine "BVM6" means 6 speed manual gearbox.

```
0                             159 1750 Tbi (200ch)
1                         159 2.0 JTDm (170ch) ECO
2                             159 2.0 JTDm (136ch)
                                ...
40050    Delta 1.6 MultiJet (120ch) DPF Selectronic
40051      Delta 1.9 MultiJet Twinturbo (190ch) DPF
```

Feature engineering:

Extract engine power and cylinder volume from dscom.

- **`cnit`**

The National Type Identification Code (CNIT) which is a number attributed to all the cars. This number is mandatory to register the vehicle and is written on the vehicle registration document. It is a sequence of 15 characters (i.e. "M10ALFVP0000324").

```
0         M10ALFVP000G340
1         M10ALFVP000U221
              ...
40051       MLC5802B7581
```

Many unique values. We do not find, in this first analysis an interest on these values, it will therefore be dropped.

- **`tvv`**

The Variant-Variant (TVV) or the Mines type. It corresponds to an alphanumeric sequence (i.e. KW01B5B) which is specific to each manufacturer and allow to identify the specific finishing of a car. The manufacturer provides a unique identifier for each type, version, and variant of a car. It means that all identical models have the same Variant-Variant numbers. The TVV is divided in 3 main information:

- The type which regroups all the identical information on some technical points.
- The variant if the car has different model.
- The version which gives the different finishing of a car.

This variable can be used during the feature engineering phase to merge the French Government dataset and European Union dataset. Indeed, a preliminary test has been performed between both datasets on a common tvv variable to check if a similar car appears. The ‘TVV’ variable has been used on the French government Dataset and compared to the 3 variables ‘T’, ‘Va’ and ‘Ve’ variables of the European Union datasets which is the TVV cuts in the 3 main information detailes above. The results are the following:

French Government Dataset		**European Union Dataset**	
Name of the variable	Value	Name of the variable	Value
lib_mrq_utac	ASTON MARTIN	id	192476
lib_mod_doss	DB9	MS	FR
lib_mod	DB9	MP	na
dscom	DB9	Mh	ASTON MARTIN
cnit	M10SCFVP000J200	MAN	ASTON MARTIN LAGONDA LTD
tvv	VH1A103L4MAAE	MMS	ASTON MARTIN LAGONDA
cod_cbr	ES	TAN	e11*2001/116*0229*18
hybride	non	T	VH1
puiss_admin_98	44	Va	A103
puiss_max	381	Ve	L4MAAE
typ_boite_nb_rapp	A 6	Mk	ASTON MARTIN
conso_urb	21.60	Cn	DB9
conso_exurb	10	Ct	M1
conso_mixte	14.30	r	5
co2	333	e (g/km)	333
co_typ_1	0.19	m (kg)	1860
hc	0.04	w (mm)	2740
nox	0.03	at1 (mm)	1590
hcnox	NaN	at2 (mm)	1580
ptcl	NaN	Ft	Petrol
masse_ordma_min	1860	Fm	M
masse_ordma_max	1860	ec (cm3)	5935
champ_v9	715/2007*630/2012EURO5	ep (KW)	381
date_maj	mars-13	z (Wh/km)	NaN
Carrosserie	COUPE	IT	NaN
gamme	LUXE	Er (g/km)	NaN

By using the same identifier (in this case ‘VH1A103L4MAAE’), the same car has been found between both datasets with some common variables and results (name of the manufacturer, name of the car, CO2 emissions, weight, …). It also gives access to new variables that has been describe in the audit report.

Feature engineering:

> The 'tvv', as an unique identification number can be used to complete the French government data with new variables from EEA dataset. Ex.: wheelbase, track width…

Unique values:

```
0          939AXN1B52C
1          939AXP1B54C
2           939AXR1B64
                ...
40049      198AXN1B12D
40050      844AXC1105C
40051      844AXE1A04C
```

- **`cod_cbr`**

The type of fuel of the car

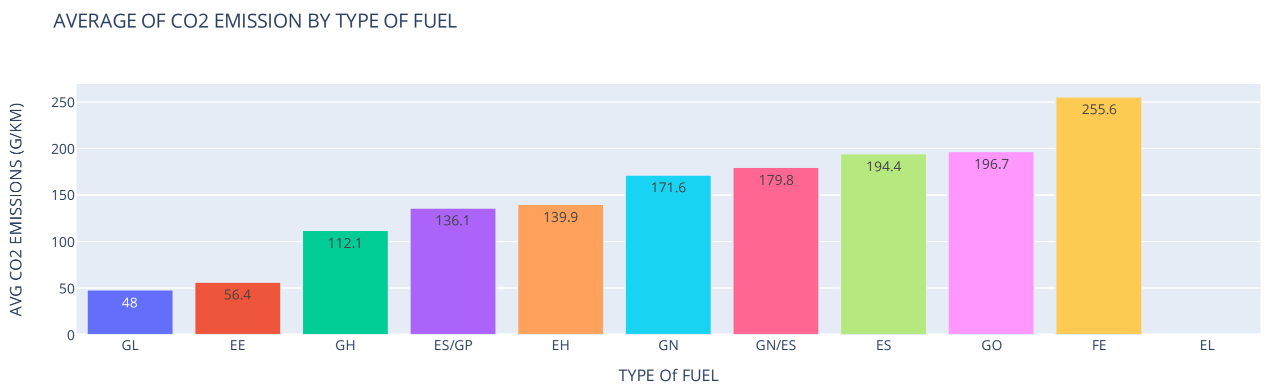
Unique Values:

```
000: EE      001: EH   002: EL   003: ES   004: ES/GN   005: ES/GP
006: FE      007: GH   008: GL   009: GN   010: GN/ES   011: GO
012: GP/ES
```

GO: Diesel; **ES**: Gasoline; **EH**: Non-plug-in hybrid vehicle; **GN/ES**: Natural Gas/Gasoline; **GH**: Non-plug-in electric diesel; **ES/GP:** Gasoline/liquefied petroleum gas; **EL**: Electric; **GN**: Gas Natural; **EE**: Gasoline electricity plug-in hybrid; **FE**: E85 super-ethanol; **GL**: Diesel plug-in electricity.

The following values are merged before this analysis: 'ES/GN' → 'GN/ES' and 'GP/ES' → 'ES/GP'.

We expect that the type of fuel will have a big impact in the CO2 emissions. Let's try to visualize that in a plot:



The bar chart above depicts how the type of fuel influences the average release of CO2 emissions from the 10 fuel types studied. The emissions range from 48 g/km to 255.6 g/km. The three fuel types with lowest carbon footprint are namely 'GL' (diesel plug-in electric), 'EE' (gasoline plug-in electric hybrid) and 'GH' (diesel non-plug-in electric). By contrast, the three fuel types with the highest carbon footprint are 'ES' (gasoline), 'GO' (diesel) and 'FE' (E85 super-ethanol).

A useful exercise to carry out in the next phase, would be to identify the number of cars using each fuel type by brand. In doing so, we can ascertain the proportion of cars using each fuel type. As even if certain fuel types such as 'FE' and 'GO' are more polluting, this may not necessarily indicate how many cars actually use these more polluting fuel types.

Action:

> 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. This analysis can further inform the findings from the graph above.

If we do a value count of the fuel types for hybrid cars, we retrieve the following result:

```
cod_cbr
EH     497
GH     135
ES     105
GO      27
EE      14
GL       2
Name: count, dtype: int64
```

We can see that 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). This must be corrected during next phase before starting the modelling.

Action:

> Replace fuel type by the correct values. To do this, it's possible to check the fuel type of a car of another year of the French Government dataset by using the variable 'tvv' (if the 'tvv' variable is similar, it means that this is exactly the same car with the same characteristics).

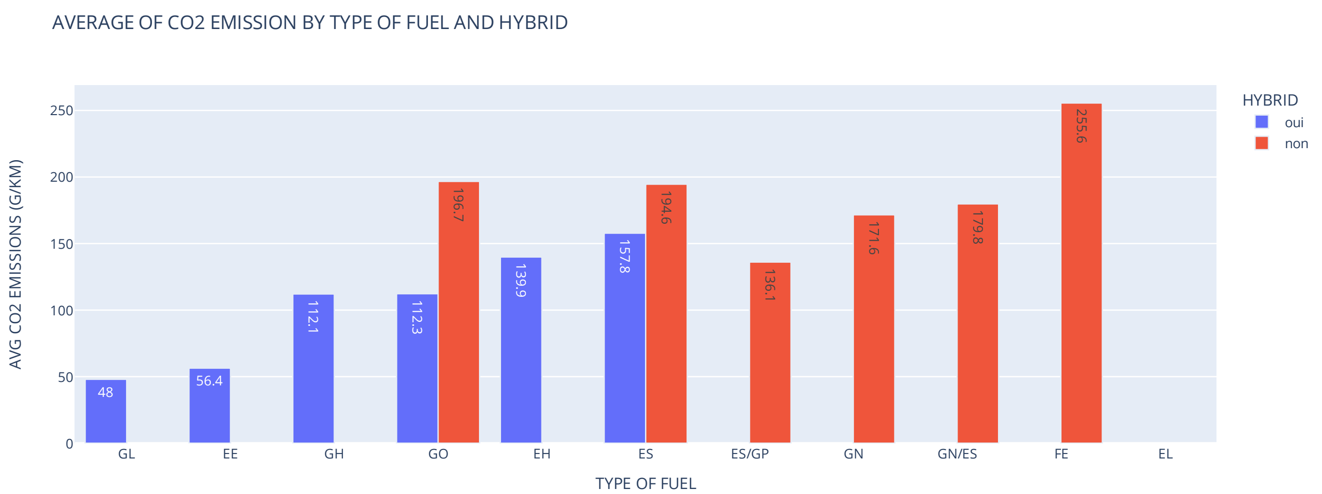
- **`hybride`**

Information to identify hybrid vehicles.

Unique Values:

```
000: oui
001: non
```

We expect that a hybrid car will produce less CO2 than a non-hybrid. Let's try to visualize that in a plot:



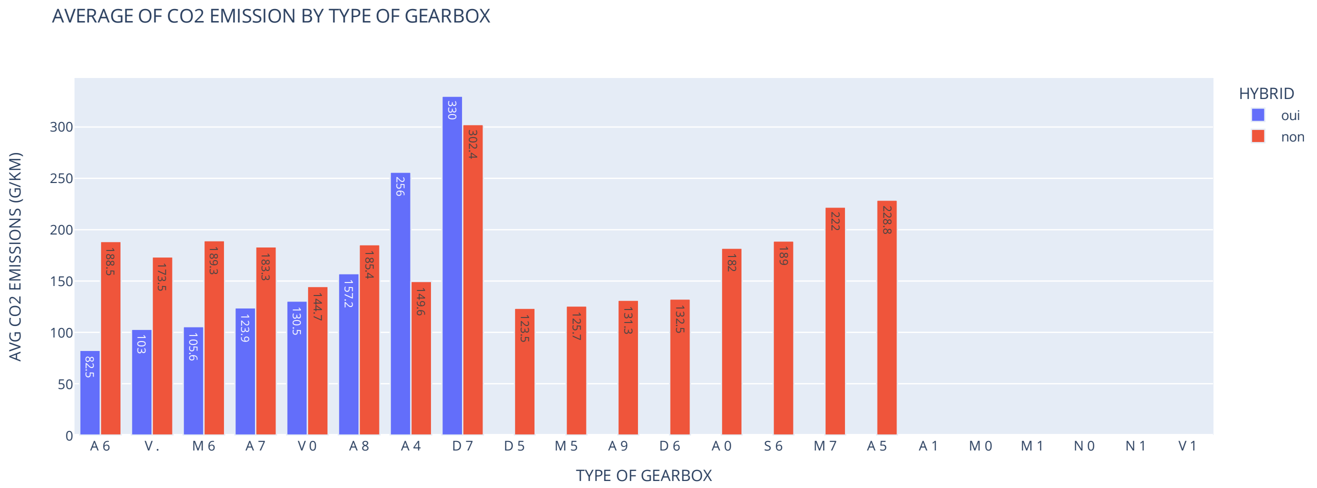
That bar chart above shows, as expected, that a hybrid vehicle is on average less CO2 intensive, compared to a non-hybrid vehicle. This relationship will be further investigated in the next phase of the project.

- **`typ_boite_nb_rapp`**

The type of gearbox (first letter) and the number of reports (the following number) i.e. 'M 6' means Manual Gearbox and 6 reports.

Unique Values:

```
000: A 0   001: A 1   002: A 4   003: A 5   004: A 6   005: A 7   006: A 8
007: A 9   008: D 5   009: D 6   010: D 7   011: M 0   012: M 1   013: M 5
014: M 6   015: M 7   016: N 0   017: N 1   018: S 6   019: V .   020: V 0
021: V 1
```



As can be seen in the bar chart above, the type of gearbox has a visible impact on the CO2 emissions. To investigate this impact, it has been proposed that the letter and number of the 'typ_boite_nb_rapp' variable can be separated to explore the correlation of each category with the average CO_2 emissions.

Feature engineering:

> Separate first letter and number in different features and check the correlation of each category with CO_2 emissions.

- **`champ_v9`**

Field V9 of the registration certificate which contains the Euro standard.

Unique Values:

```
0        715/2007*692/2008EURO5
1        715/2007*692/2008EURO5
2        715/2007*692/2008EURO5
                  ...
40049    715/2007*692/2008EURO5
40050    715/2007*692/2008EURO5
40051    715/2007*692/2008EURO5
```

Many unique values. To be further studied as described below.

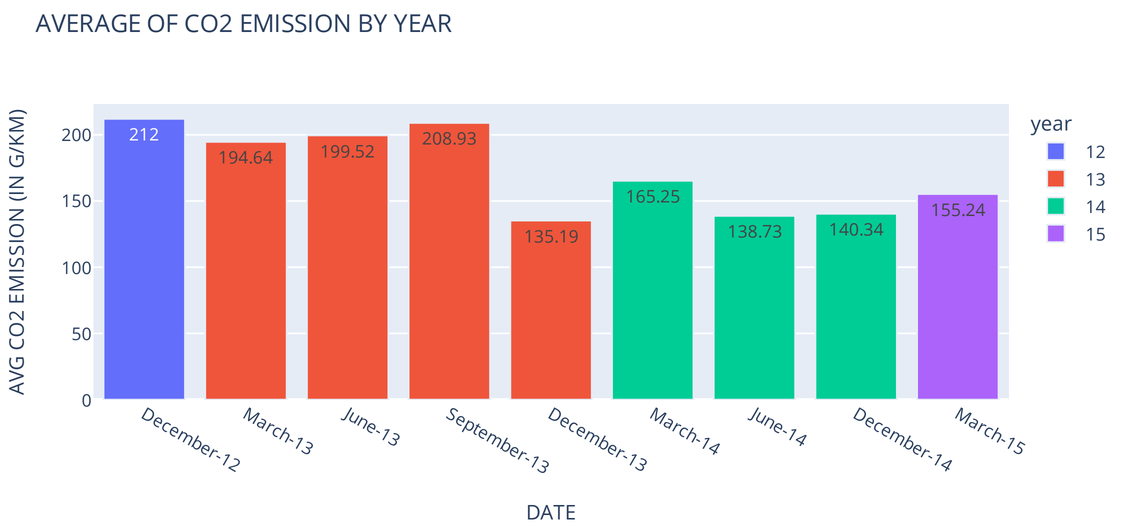
Feature engineering:

> Investigate if certified or no certified car has an impact on the CO_2 emissions. A binary variable can be created.

- **`date_maj`**

The date of the last update.

We expect the date of the last update to not to have a correlation with the CO2 emissions. But let's investigate further, as it contains the year and this could be interesting to extract later to see the trend of CO2 emissions by year.



The bar char above shows that from December 2012 to March 2015, the average release of CO2 emissions gradually declines each year. However, we must be careful as electrical cars were not removed from the dataset and the production of these cars grew significantly in the later years. This could impact the average if included in the calculation. Therefore, it is possible that the increase in electric cars gave rise to the reduction in CO2 emissions. Overall, a 27% decrease in CO2 emissions is seen from December 2012 to March 2015. The largest decrease in CO2 emissions is seen between September 2013 and December 2013, which was 73.7 g/km (35%).

Limitations: It is worth nothing that in section 3.2, we saw that around 57% of the variable data is missing. Furthermore, the step change between each period is not consistent. For example, the year 2014 does not contain any data for September like the year 2013 does. Nevertheless, if what we are interested in is the year, this data is available in the file name.

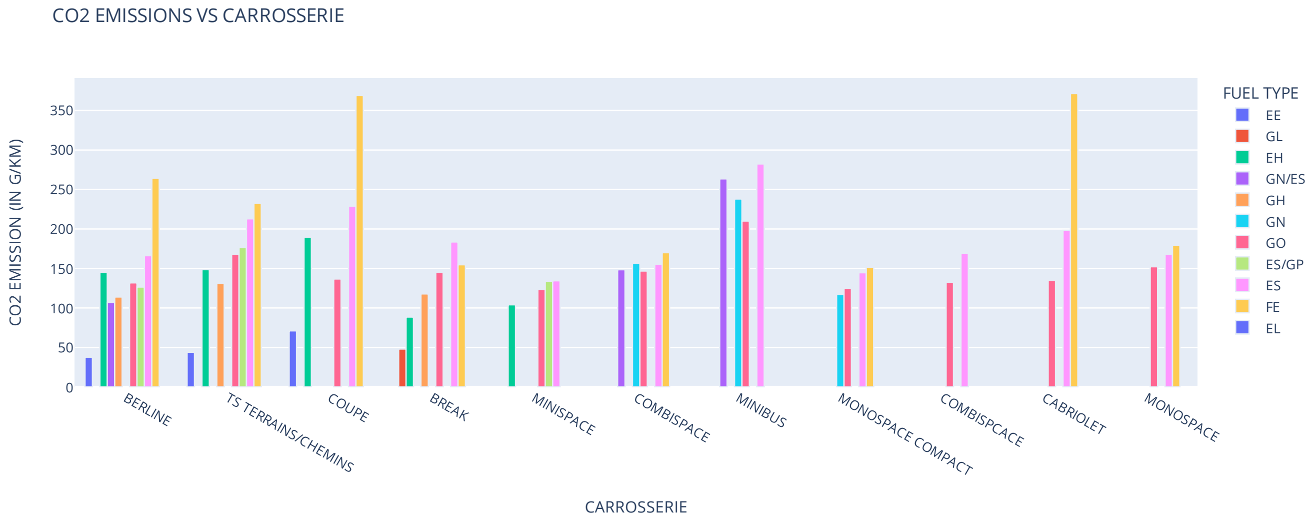
Create a column year from file name. This can be used to split the test / train data by year. We can train the dataset in older years and test it in the most recent data.

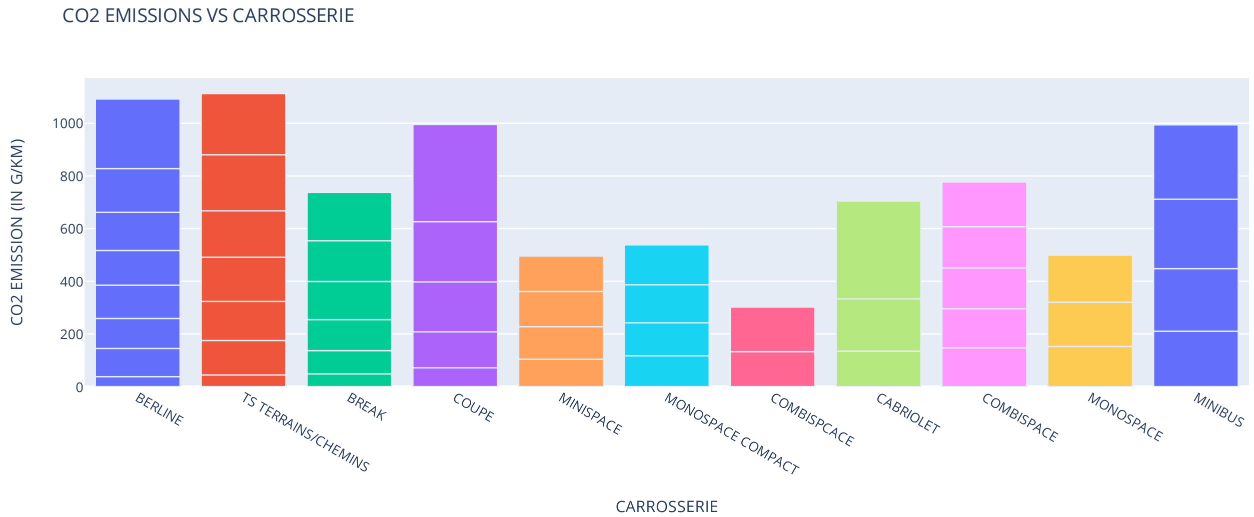
- **Carrosserie**

Car’s body type

Unique Values:

```
000 : BERLINE            001 : BREAK              002 : CABRIOLET
003 : COMBISPACE         004: COMBISPCACE         005: COUPE
006: MINIBUS             007: MINISPACE           008: MONOSPACE
009: MONOSPACE COMPACT   010: TS TERRAINS/CHEMINS
```





The bar chart above illustrates how the vehicle body type influences the average release of CO2 emissions. The four body types with the highest carbon footprint seem to be ‘Berline’, ‘TS Terrains/Chemins’, ‘Minibus’ and ‘Coupe’. Body Type ‘Coupe’ and fuel type ‘FE’ gives the highest absolute CO2 footprint. On the other hand, body type ‘Break’ and fuel type ‘GL’ gives the lowest absolute CO2 footprint. Again the fuel type data, as mentioned in section 3.3, will need to be validated further before exploring the ‘Carrosserie’ variable.

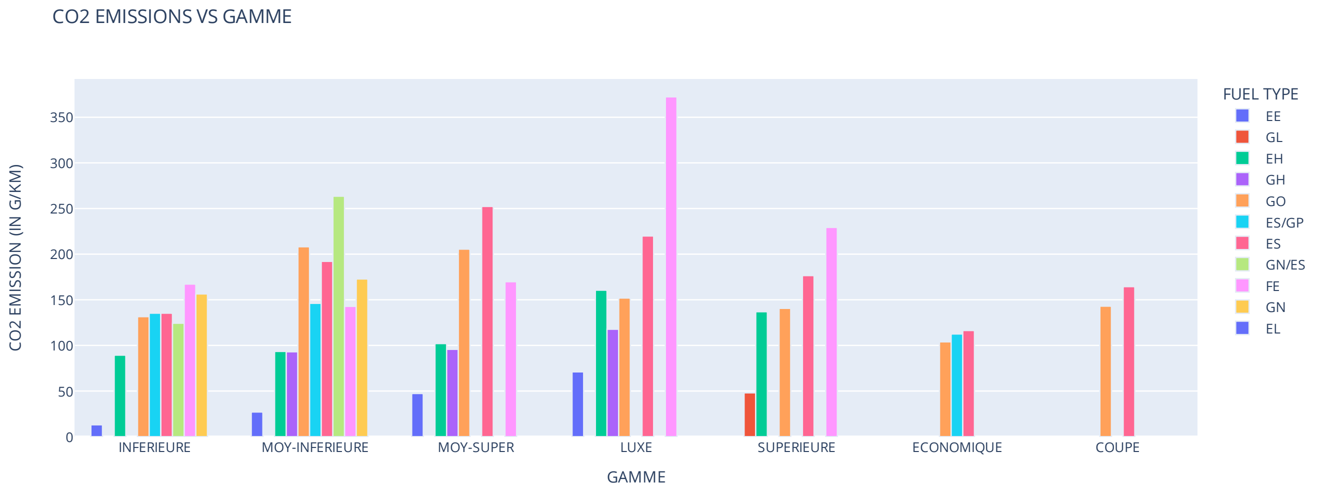
- **`gamme`**

The range of the car in term of quality (luxury car…).

Unique Values:

```
000: COUPE            001: ECONOMIQUE    002: INFERIEURE
003: LUXE             004: MOY-INF       005: MOY-INFER
006: MOY-INFERIEURE   007: MOY-SUPER     008: SUPERIEURE
```

We can see that some types highlighted in blue are repeated with similar names. They have been grouped before this analysis (refer to notebook).



From the bar chart above, we can see that the highest CO2 emissions are from type 'LUXE' with Fuel type 'FE' and the lowest for type 'INFERIEURE' fuel type 'EL'. This last one comes from the incorrect available data for some hybrid car as discussed in section 3.3 'cod_cbr'. What is much more logical, is that the cars from type 'ECONOMIQUE' seems to have average lower emissions for all fuel types. This will be further investigated in the next phase of the project.

3.4. Quantitative Variables

3.4.1. Variable names

The quantitative variables are the following:

- **`co2 [TARGET VARIABLE]`**

CO2 emission (in g/km) of the car. For the French dataset (2012-2015), the measure is according the NEDC norm.

This is our target variable. In following steps of this project, we will define if we treat this as a regression problem or a classification problem (or both). When treating the problem as a classification this variable will be split in bins which represent a certain category of emissions (ex.: low, high, average, high and very high).

> Split co2 emission in classes (ex.: low, high, average, high and very high) to use it in classification models. Ranges to be defined later.

- **`puiss_admin_98`**

Administrative power is expressed in 'CV' (tax horsepower) and is used to estimate the tax amount on the car during registration of renewal of the vehicle registration document.

In France, it exists two formulas to convert the motor power (kW) to administrative power (CV):

Approval from January 1, 2020:

$$Admin\ Power\ (CV) = 1.34 + \left(1.8 \times \frac{Motor\ Power\ (kW)}{100}\right)^2 + \left(3.87 \times \frac{Motor\ Power\ (kW)}{100}\right)$$

Approval before December 31, 2019:

$$Admin\ Power\ (CV) = \left(\frac{CO2\ Emission\ \left(\frac{g}{km}\right)}{45}\right) + \left(\frac{Motor\ Power\ (kW)}{40}\right) \times 1.6$$

There is a correlation between the target variable CO_2 Emissions and this variable. To be discussed during the next phase if this variable should be removed or not from the dataset.

- **`puiss_max`**

Maximum power of the motor expressed in KW.

- **`conso_urb`**

Urban fuel consumption (in L/100km). This consumption corresponds to drive in an urban area with an acceleration up to 15 km/h, 30 km/h and 50 km/h. Including the most frequent stop, the urban fuel consumption is typically the higher consumption.

- **`conso_exurb`**

Extra urban fuel consumption (in L/100km). This consumption corresponds to drive in an extra urban area with a drive on several speed levels up to 120 km/h/ It allows to optimize the driving and the fuel consumption of the car. Therefore, this consumption is generally the lower consumption.

- **`conso_mixte`**

Mixed fuel consumption (in L/100km). This consumption includes the drive in urban and extra urban area. Therefore, the fuel consumption is typically between the urban fuel consumption and the extra urban fuel consumption.

- **`co_typ_1`**

Carbon monoxide (CO) type I trial results measurement (g/km).

- **`hc`**

Unburned Hydrocarbons (HC) trial results measurement (g/km).

- **`nox`**

NOx trial results measurement (g/km).

- **`hcnox`**

HC+NOx trial results measurement (g/km).

It seems that we don't have a value of the hc, nox and hcnox variables in the same time (for example we can only have the variable nox and hcnox but a NaN value for hc). Using the assumption that hc + nox = hcnox, it is possible to fill the NaN values if we have 2 filled value out of 3 (which is mainly the case).

> If we have 2 filled value out of 3 for the variables hc, nox and hcnox, replace the NaN value using the assumption that hc + nox = hcnox.

- **`ptcl`**

Particle trial results measurement (g/km)

- **`masse_ordma_min`**

The mass in minimum walking order (kg). It corresponds to the empty weight of the car with a gas bottle, 90% of the fluid necessary for the car to work and one driver (75 kg).

- **`masse_ordma_max`**

The mass in maximum walking order (kg). It corresponds to the weight that the vehicle must not exceed (include passengers and bags).Action:

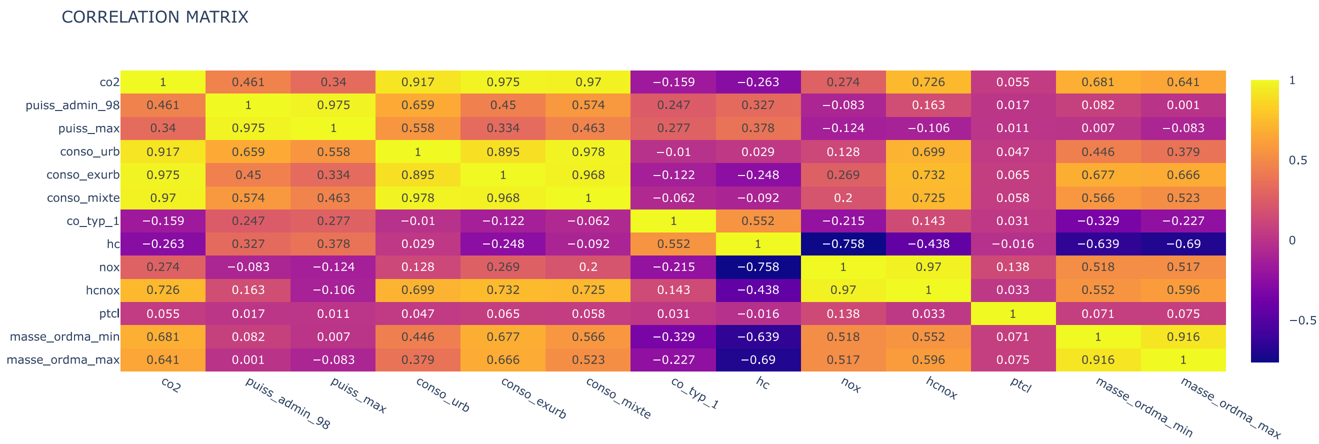
> Some cars have the same min and max mass order (around 47 000 cars) while it is different for the others. If the variable max mass is kept during the next phase, this problematic must be investigated.

3.5. Basic statistics and correlations

Many of these quantitative variables were represented as string objects. After doing the required cleaning to be able to convert the strings to numerical variables, we end up with the following data description for the quantitative variables:

	co2	puiss_admin_98	puiss_max	conso_urb	conso_exurb	conso_mixte	co_typ_1	hc	nox	hcnox	ptcl	masse_ordma_min	masse_ordma_max
count	159 622	159 780	159 724	159 543	159 543	159 622	159 090	36 813	159 090	122 452	150 181	159 780	159 780
mean	195.58	11.02	125.12	9.46	6.5	7.58	0.17	0.03	0.29	0.22	0.0	2 059.24	2 200.12
std	40.45	6.08	52.89	2.43	1.23	1.64	0.14	0.02	0.43	0.05	0.01	337.51	437.0
min	13.0	1.0	10.0	0.0	2.6	0.6	0.0	0.0	0.0	0.0	0.0	825.0	825.0
25%	182.0	9.0	100.0	8.7	6.3	7.1	0.06	0.01	0.15	0.2	0.0	1 976.0	2 000.0
50%	203.0	10.0	120.0	9.5	6.7	7.7	0.12	0.03	0.2	0.23	0.0	2 076.0	2 185.0
75%	216.0	11.0	120.0	10.3	7.1	8.3	0.26	0.04	0.23	0.25	0.0	2 219.0	2 585.0
max	572.0	81.0	585.0	41.1	15.9	24.5	0.97	0.51	1.85	0.57	0.7	3 115.0	3 115.0

Let's look at the correlation between these variables:



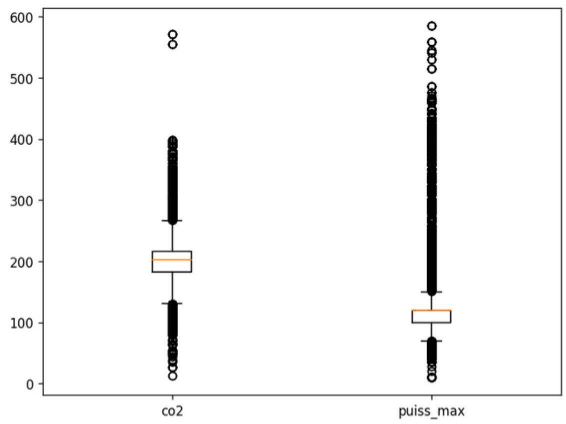
A heatmap of the correlation matrix has been used to measure the relationships between each pair of quantitative variables. As evident in the heatmap, a strong correlation can be seen between the variables 'conso_urb', 'conso_exurb' and consequently 'conso_mixte' with 'co2' (target variable). This can be justified by the fact that fuel consumption directly impacts the CO2 emissions of the car. As the car consumes more fuel, it emits more CO2 (depending upon the type of fuel).

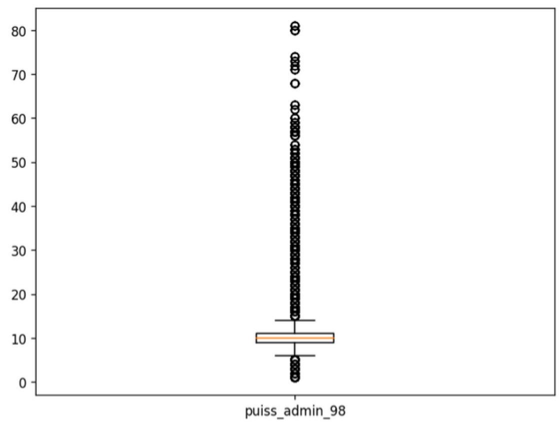
Another fairly strong correlation can be seen between 'co2' and 'masse-ordma_min' and 'masse_ordma_max'. The same reasoning can be applied to their correlation with variables 'conso_urb', 'conso_exurb' and 'conso_mixte'. This is mainly because the mass of the vehicle can have a direct impact on its fuel consumption which in turn affects the CO2 emissions. The higher the mass of the vehicle, the higher the fuel consumed by the vehicle and therefore higher the CO2 emissions.

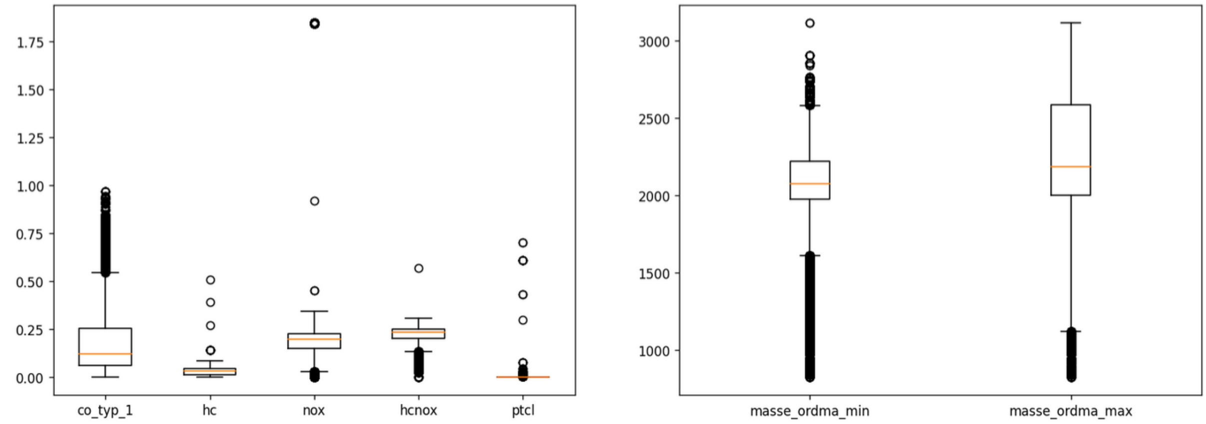
Moreover, there is another significant correlation of 'puiss_admin_98' and "puiss_max' with 'conso_urb', 'conso_exurb' and 'conso_mixte'. This is also expected as the power output of an engine is related to its fuel consumption, which in turn affects the CO2 emissions of the car. However, one must be careful while establishing a correlation between these variables since it is not always straightforward. For instance, a car with a more efficient engine may be able to produce the same power output as a less efficient engine while consuming less fuel and emitting less CO2.

Moving on the variables related to the emissions measurement, we noticed a correlation between 'hc and 'nox' as these two emissions are interdependent since 'hcnox' is the combined HC + NOx emissions for the vehicle. However, an unusual negative correlation is witnessed between the variables 'hc' and 'nox'. This is mostly linked to the high quantity of missing values of `hc`. We can see that `nox` has almost 1 as correlation with `hcnox` as expected. This may be corrected when the actions from the corresponding section is taken to complete messing values.

The only variables with weakest correlation with CO2 emissions are 'co_type_1', 'hc' and 'ptcl'. Amongst which, particle emissions 'ptcl' is the least correlated. As seen in the table above, this variable contains at least 75% of the data points lying at exactly zero which indicates a highly skewed data with some outliers going up to the maximum value.







If we look at the distribution of each of these numerical variables in the boxplot above, we notice quite a complex distribution with most of the variables (excluding ‘masse_ordma_max’ and ‘mass_ordma_min’) having a very narrow interquartile range (IQR) indicating that the data is mostly clustered around a narrow range of values.

However, most of the variables are either moderately or highly skewed on either the right (‘co_type_1’, ‘puiss_max’, ‘puiss_admin_98’) or left side (‘mass_ordma_min’, ‘mass_ordma_max’, ‘hcnox’) represented by a very high number of outliers on either side of the box.

For some of the variables (‘co2’, ‘puiss_admin_98’, ‘nox’, ‘hc’), the median falls almost in the center of the box making the distribution likely to be symmetrical but the presence of outliers complicates the interpretation of the data in these cases. And similar reasoning can be used for variables that have equidistant whiskers from the box (such as ‘mass_ordma_min’, ‘co2’, ‘puiss_max’, ‘nox’, ‘hcnox’, ‘puiss_admin_98’) suggesting that the data might be roughly symmetric around the median however the presence of outliers on either side of the distribution makes it difficult to interpret.

3.1. Final Data Set

After this initial grouping and cleanup, we end-up with the following data set:

```
<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
```

This data set is saved as 'data_2012-2015.csv' and will be the start point for this project next phase.

It is important however to keep in mind that, after the required corrections listed in this report in the action boxes, another check must be made on the duplicate values of the dataset. Indeed, we dropped the duplicated values in the beginning this first look but using all columns. The fact that the same car can have different license numbers and even the `date_maj` (update date) can lead to keeping the same information multiple times in the dataset.

A quick check shows that there are 56 290 identical values in the final dataset. This leads us with a dataset of around 103 k rows for the next phase.

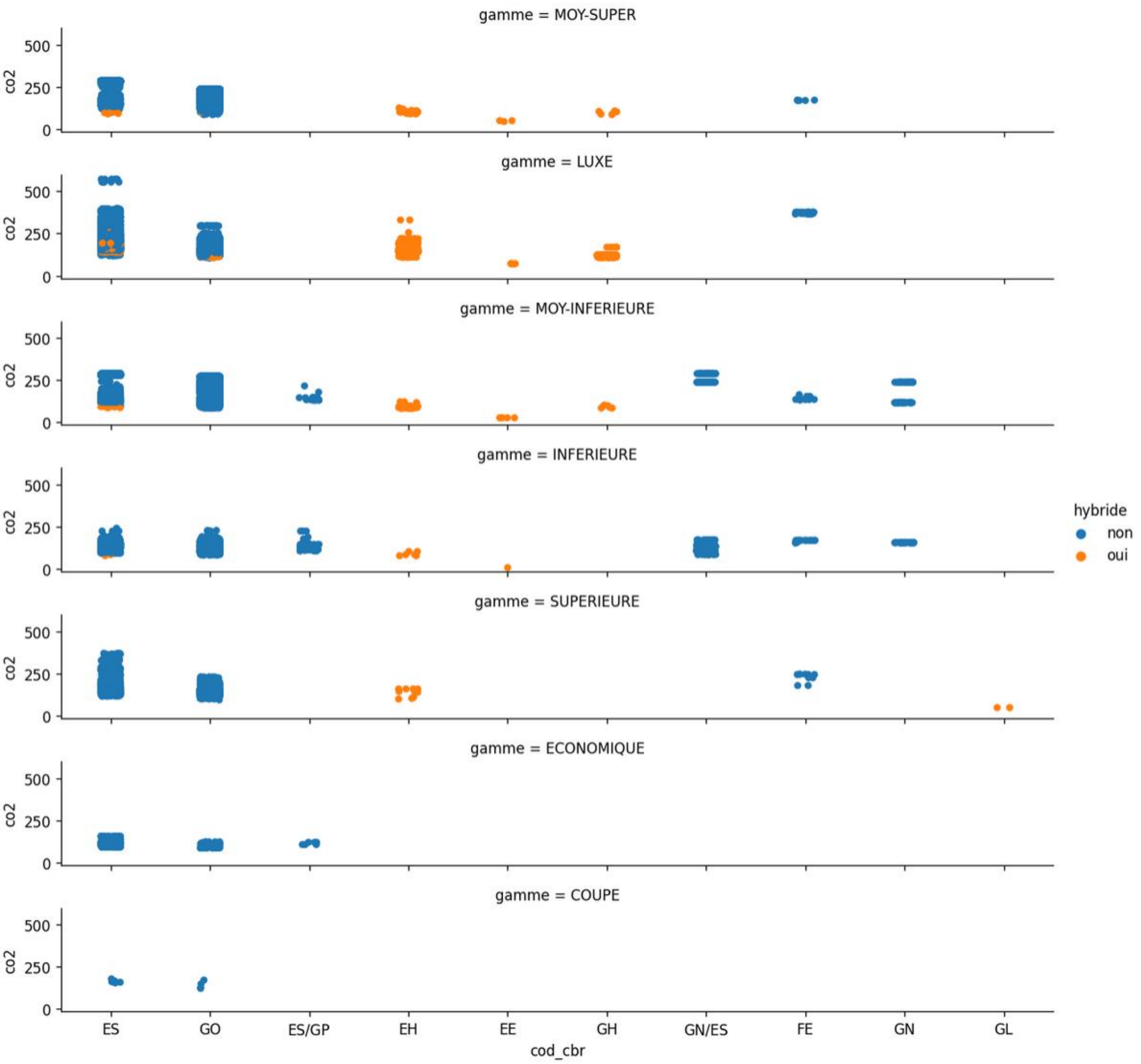
```
In [1]: columns = ['tvv', 'cod_cbr', 'hybride', 'puiss_admin_98', 'puiss_max', 'typ_boite_nb_rapp',
   ...:            'conso_urb', 'conso_exurb', 'conso_mixte', 'co2', 'co_typ_1', 'hc',
   ...:            'nox', 'hcnox', 'ptcl', 'masse_ordma_min', 'masse_ordma_max',
   ...:            'Carrosserie', 'gamme']
In [2]: df[columns].duplicated().sum()
Out[2]: 56290
```

We will choose, however, to perform the required corrections first in the next phase prior to drop these rows.

After corrections, check for duplicated values without considering columns like: `date_maj`, `champ_v9`, etc. That can be different for the same car with same characteristics.
Also, variables that has a lot of missing values like: `hcnox`, `ptcl`, `hc`, etc. must be removed from the comparison list.

4. ADDITIONAL DATA VISUALISATION

4.1. Categorical Variables



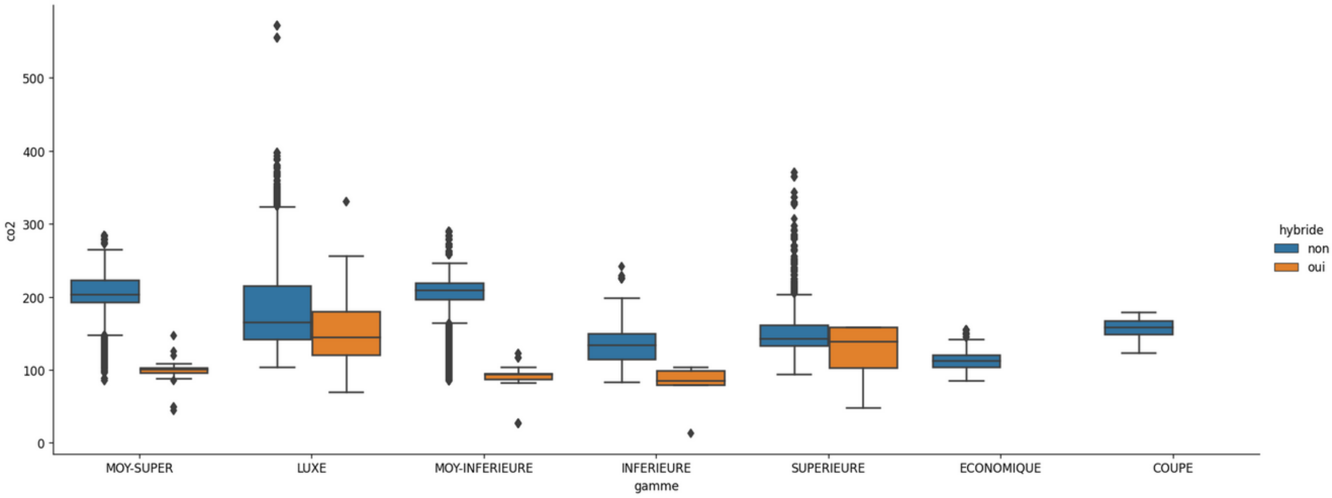
The pair plot above shows the distribution of the CO2 emissions by fuel type in the x axis and separated the variable `gamme` on each subplot. Each plot is also split between hybrid (orange dots) and non-hybrid (blue dots).

We can see that the higher CO2 emissions are found for the games ‘LUXE’ and ‘SUPERIEURE’ both for fuel type ‘ES’ (gasoline).

The lowest CO2 emissions are found in the games ‘ECONOMIQUE’, ‘MOY-INF’ and ‘COUPE’ for fuel type ‘ES’ (gasoline). These last range do not seem to have hybrid cars. We can now visually see the wrongly assigned hybrid cars to the fuel type ‘ES’ and ‘GO’ as discussed in section 3.3.1 cod_cbr on the top 3 curves.

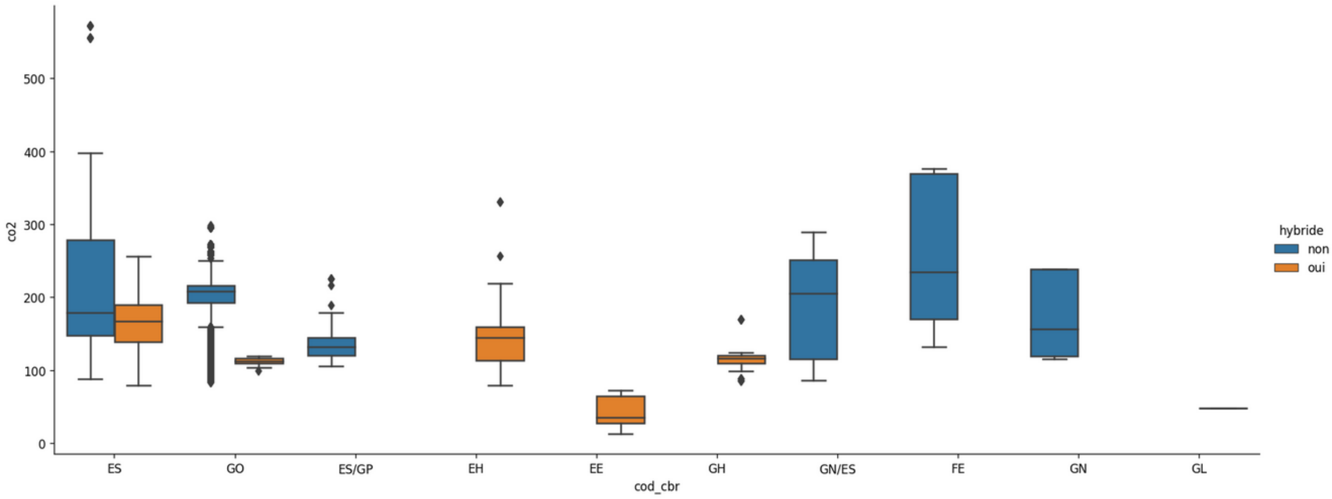
It is also possible to notice that the hybrid cars for each game are found in the lower part of the curve.

This information can be seen in another form in the plot below:

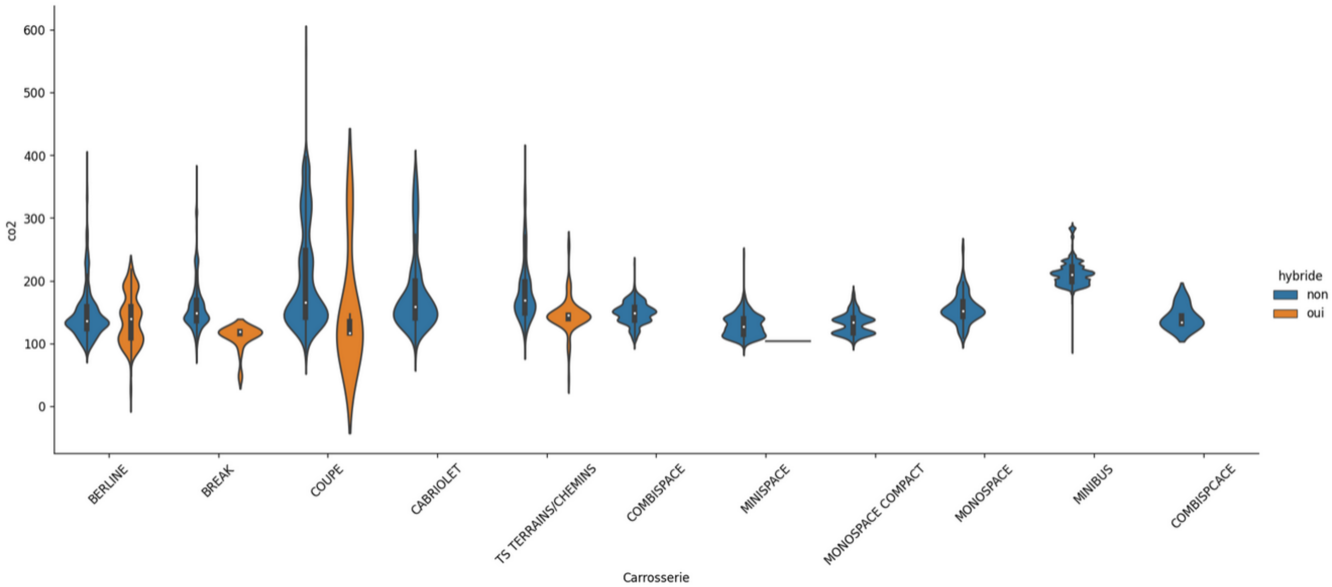


We see that the cars of game 'LUXE' non-hybrid have the highest emission values, represented by outliers. While the lowest emission values are found in hybrid cars of game 'INFERIEURE' and 'MOY-INFERIEURE'.

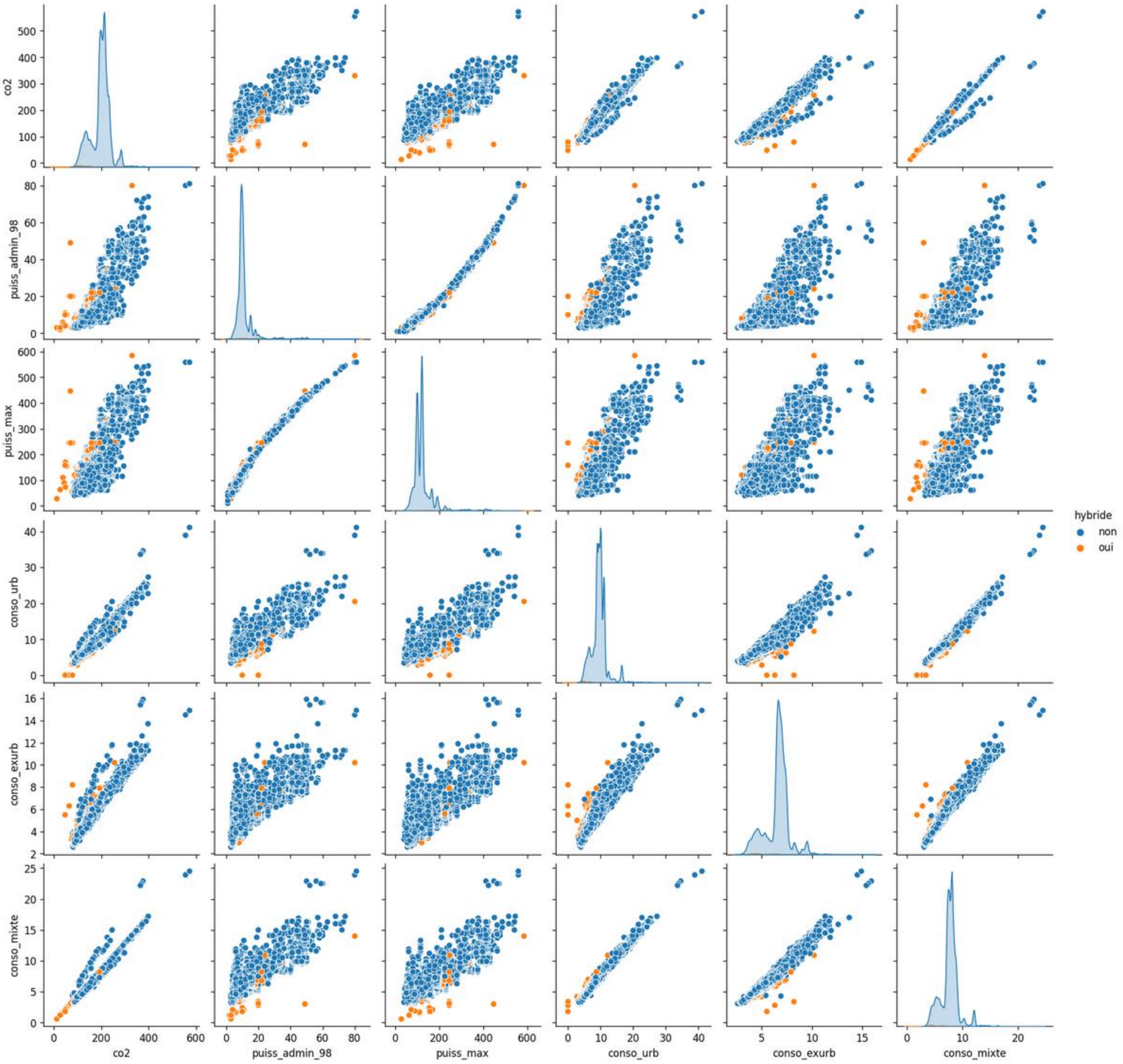
If we plot the same boxplot by fuel type we see that the highest emissions are outliers of 'ES' type, but the fuels with higher average emissions is 'FE' (Super Methanol) as already seen in section 3.3.1 cod_cbr.



If we now make a violin plot of the CO2 emissions by car body, we see that most of the values of the emissions are around 100 to 300 g/km. And the maximum emissions are found for car body 'COUPE' non-hybrid, and the minimum for this same body type but hybrid version.

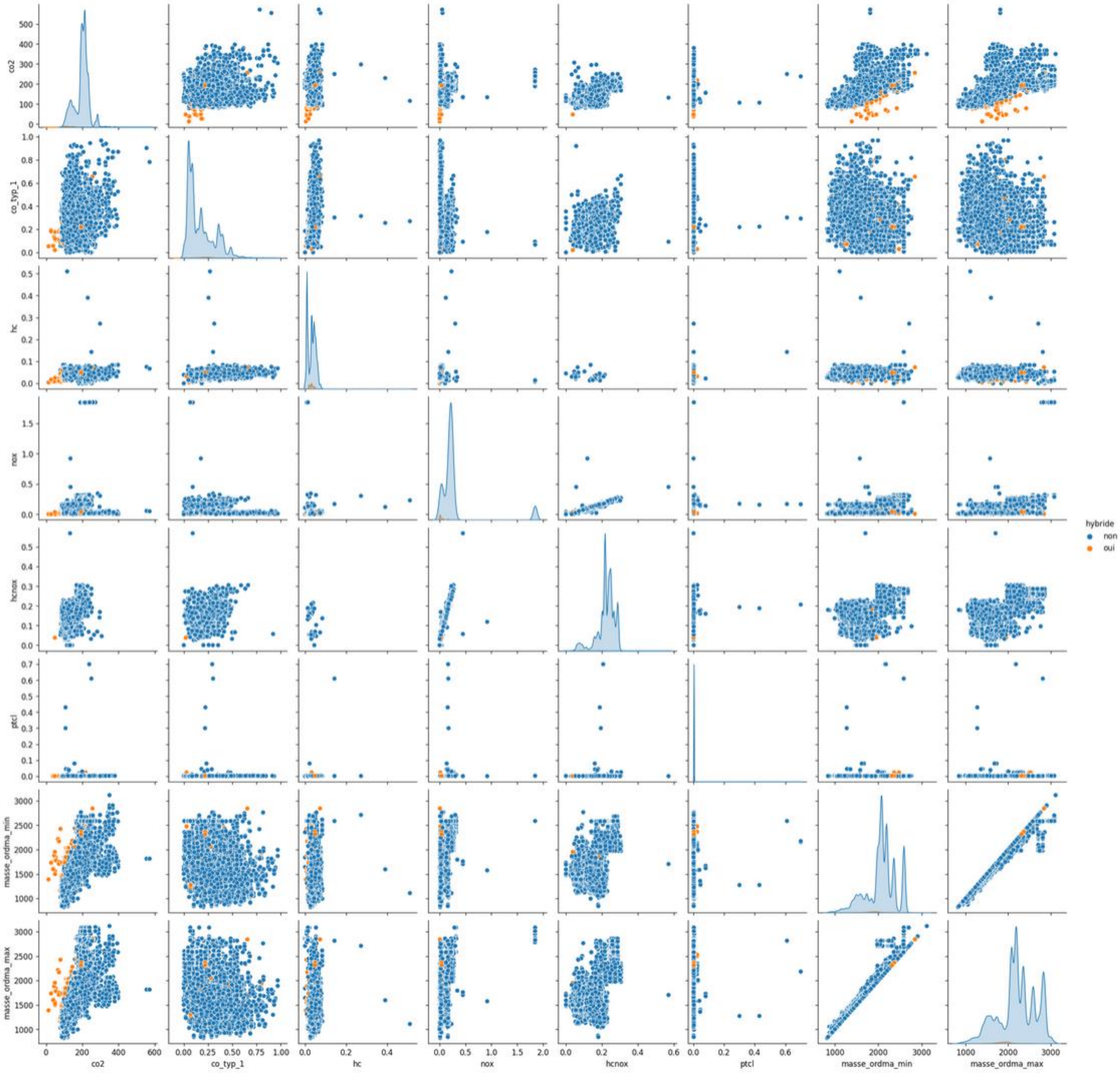


4.2. Quantitative Variables



From the pairplot shown above between the quantitative variables ‘co2’, ‘puiss_admin_98’, ‘puiss_max’, ‘conso_urb’, ‘conso_exurb’ and ‘conso_mixte’, we notice a strong linear relationship between ‘puiss_max’ and ‘puiss_admin_98’ for non-hybrid vehicles. Thus, we can say that higher the maximum power of the motor, higher its administrative power. However, a statistical test must be conducted to confirm with an appropriate statistical test.

Another linear relationship can be seen between ‘co2’ and ‘conso_urb’, ‘conso_exurb’ and ‘conso_mixte’ especially for non-hybrid cars. This can be justified higher fuel consumption of the car results in higher CO2 emissions.



From the pairplot shown above, we only a strong linear relationship between ‘mass_ordma_min’ and ‘mass_ordma_max’ for non-hybrid vehicles. We can equally see the correlation between ‘nox’ and ‘hcnox’.

It is clear again for both of previous plots that the hybrid cars are found in the lower part of the distribution of the quantitative variables, same behaviour was observed for categorical variables, which is logic for hybrid cars.

5. CONCLUSION AND WAY FORWARD

This first overview of the available dataset allowed us to have a first understanding of the data we are working with. Additionally, it was possible to identify the points described in the following sections.

5.1. Action list

A set of actions were added to this report on a case-by-case basis when analysing each variable. These actions must be done before starting the data modelling on next phase. The actions are:

- 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.
- Replace fuel type by the correct values. To do this, it's possible to check the fuel type of a car of another year of the French Government dataset by using the variable 'tvv' (if the 'tvv' variable is similar, it means that this is the same car with the same characteristics).
- Some cars have the same min and max mass order (around 47 000 cars) while it is different for the others. If the variable max mass is kept during the next phase, this problematic must be investigated.
- After corrections, check for duplicated values without considering columns like: \`date_maj\`, \`champ_v9\`, etc. That can be different for the same car with same characteristics.
  Also, variables that has a lot of missing values like: \`hcnox\`, \`ptcl\`, \`hc\`, etc. must be removed from the comparison list.

5.2. Feature Engineering list

A set of possible features to be developed was identified on a case-by-case basis when analysing each variable. These features can be implemented to the dataset if found to be interesting for the model. The identified feature engineering are:

- Extract engine power and cylinder volume from dscom.
- The 'tvv', as a unique identification number can be used to complete the French government data with new variables from EEA dataset.
- Separate First letter and number in different features and check the correlation of each category with co2 emissions.
- Investigate if certified or no certified car has an impact on the CO2 emissions. A binary variable can be created.
- Create a column year from file name. This can be used to split the test / train data by year. We can train the dataset in older years and test it in the most recent data.
- Split co2 emission in classes (ex.: low, high, average, high and very high) to use it in classification models. Ranges to be defined later.
- If we have 2 filled value out of 3 for the variables hc, nox and hcnox, replace the NaN value using the assumption that hc + nox = hcnox.

5.3. Variables to drop

The following variables do not bring any interesting information to the modelling and / or their content do not allow to differentiate the co2 emissions (refer to analysis in the corresponding sections):

- mrq_utac: high quantity of missing values, as not present in all years.
- puiss_heure: high quantity of missing values, as not present in all years.
- Unnamed: 26: empty.
- Unnamed: 27: empty.
- Unnamed: 28: empty.
- Unnamed: 29: empty.
- cnit: We do not find, in this first analysis an interest on these values, it will therefore be dropped.
- puiss_admin_98: can be used to back calculate co2 emissions, no sense to make a ML model if we have this value.

There are surely other variables to be dropped. This will, nevertheless, be done after the action list is completed in next phase, to prevent avoidable data losses (by completing missing values for example).

ANNEX 1: AUDIT TABLE

# Col	Name of the Column	Dataset origins of the variable	Variable's type	Description	Is the variable available before prediction	Variable's type	Percentage of missing values	Categorical / Quantitative	Distribution	Comments
		In which dataset can we find the variable ?	*Is the variable a feature or the target ? (Only applicable for supervised learning projects)*	*What does this variable represent (from a business perspective ?)*	*Is this variable known before the prediction is made ? (Only applicable for supervised learning projects)*	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*
1	lib_mrq_utac	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The brand of the car manufacturer (the real brand of the car is indicated and not the group of manufacturer i.e Peugeot is indicated and not the PSA group which regroup different brands like Peugeot, Citroën, etc...)		object: brand of the car manufacturer so it's a string variable	0.0%	Categorical		
2	lib_mod_doss	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The file name of the car		object: file name of the car so it's a string variable	0.0%	Categorical		
3	lib_mod	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The business name of the car. The manufacturer chooses the business name which will be written on the vehicle registration document. In the dataset, the file name of the car can be identical to the business name but can also be slightly different (i.e. for the same car: lib_mod_doss name = AR8C SPIDER and lib_mod = 8C SPIDER)		object: business name of the car so it's a string variable	0.00%	Categorical		
4	dscom	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Commercial Designation. It regroups different information about the car model i.e. : 3008 1.6 THP (156ch) BVM6 "3008" is the business name of the car "1.6" is the volume of all engine cylinders, here it's 1.6 liters or 1600 cm3 "THP" is a name of a motor brand (Turbo High Pressure) "(156ch)"is the power of an engine "BVM6" means 6 speed manual gearbox		object: can regroup different type of information inside one variable like string content (car name) and some numbers (engine displacements)	0.00%	Categorical		
5	cnit	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The National Type Identification Code (CNIT) which is a number attributed to all the cars. This number is mandatory to register the vehicule and is written on the vehicule registration document. it is a sequence of 15 characters (i.e. "M10ALFVP0000324")		object: unique sequence of 15 characters	0.00%	Categorical		
6	tvv	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The Variant-Variant (TVV) or the Mines type. It corresponds to a alphanumeric sequence (i.e. KW01B5B) which is specific to each manufacturer and allow to to identify the specific finition of a car. The manufacturer provide a unique identifier for each type, version and variant of a car. It means that all identical models have the same Variant-Variant numbers. The TVV is divided in 3 main information: 1. The type which regroup all the identical information on some technical points 2. The variant if the car have different model 3. The version which give the different finition on a car		object: alphanumeric sequence	0.00%	Categorical		
7	cod_cbr	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The type of fuel of the car		object: Acronym for the fuel type	0.00%	Categorical	'GO' (Diesel), 'ES' (Gasoline), 'EH' (non-plug-in hybrid vehicle), 'GN/ES' (Natural Gas/Gasoline), 'GH' (non-plug-in electric diesel), 'ES/GP' (Gasoline/liquefied petroleum gas), 'EL' (Electric), 'GN' (Gas Natural), 'EE' (gasoline electricity plug-in hybrid), 'FE' (E85 super-ethanol), 'GL' (diesel plug-in electricity)	
8	hybride	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Information to identify hybrid vehicles		object: 'oui' or 'non'	0.00%	Categorical	'oui' for hybrid car 'non' for non-hybrid car	
9	puiss_admin_98	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Administrative power is expressed in 'CV' (tax horsepower) and is used to estimate the tax amount on the car during registration of renewal of the vehicle registration document. In France, it exists two formulas to convert the motor power (kW) to administrative power (CV) : Approval from January 1, 2020: $Admin\ Power\ (CV) = 1.34 + \left(1.8 \times \frac{Motor\ Power\ (kW)}{100}\right)^2 + \left(3.87 \times \frac{Motor\ Power\ (kW)}{100}\right)$ Approval before December 31, 2019: $Admin\ Power\ (CV) = \left(\frac{CO2\ Emission\ (\frac{g}{km}))}{45}\right) + \left(\frac{Motor\ Power\ (kW)}{40}\right) \times 1.6$		int64	0.00%	Quantitative	Refer to the section 3.4.2 of the report for the data description of the quantitative variables and the correlation matrix between the quantitative variables. Refer to the section 4.2 of the report for the pairplot graph of the quantitative variables (including the distribution).	
10	puiss_max	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Maximum power of the motor expressed in KW		int64	0.04%	Quantitative		
11	typ_boite_nb_rapp	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The type of gearbox (first letter) and the number of reports (the following number) i.e. 'M 6' means Manual Gearbox and 6 reports		object : First letter for type of geaborx (Manuel, Automatic, ...) Then a number for the number of reports of the gearbox	0.00%	Categorical		This variable could be split into 2 dedicated variables: One for the type of gearbox and one for the number of reports.
12	conso_urb	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	urban fuel consumption (in L/100km). This consumption corresponds to drive in an urban area with an acceleration up to 15 km/h, 30 km/h and 50 km/h. Including the most frequent stop, the urban fuel consumption is typically the higher consumption.		float64	0.15%	Quantitative	Refer to the section 3.4.2 of the report for the data description of the quantitative variables and the correlation matrix between the quantitative variables. Refer to the section 4.2 of the report for the pairplot graph of the quantitative variables (including the distribution).	
13	conso_exurb	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Extra urban fuel consumption (in L/100km). This consumption corresponds to drive in an extra urban area with a drive on several speed levels up to 120 km/h/ It allows to optimize the driving and the fuel consumption of the car. Therefore, this consumption is generally the lower consumption.		float64	0.15%	Quantitative		
14	conso_mixte	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Mixed fuel consumption (in L/100km). This consumption includes the drive in urban and extra urban area. Therefore, the fuel consumption is typically between the urban fuel consumption and the extra urban fuel consumption.		float64	0.10%	Quantitative		
15	co2	French Gouvernment (data.gouv.fr) - 2012 - 2015	Target variable	CO2 emission (in g/km) of the car. For the French dataset (2012-2015), the measure is according the NEDC norm.		float64	0.10%	Quantitative		
16	co_typ_1	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Carbon monoxide (CO) type I trial results measurement (g/km)		float64	0.55%	Quantitative		
17	hc	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Unburned Hydrocarbons (HC) trial results measurement (g/km)		float64	76.73%	Quantitative		
18	nox	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	NOx trial results measurement (g/km)		float64	0.55%	Quantitative		
19	hcnox	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	HC+NOx trial results measurement (g/km)		float64	23.71%	Quantitative		
20	ptcl	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Particle trial results measurement (g/km)		float64	6.36%	Quantitative		
21	masse_ordma_min	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The mass in minimum walking order (kg). It corresponds to the empty weight of the car with a gas bottle, 90% of the fluid necessary for the car to work and one driver (75 kg).		int64	0.00%	Quantitative		
22	masse_ordma_max	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The mass in maximum walking order (kg). It corresponds to the weight that the vehicle must absolutely not exceed (include passengers and bags).		int64	0.00%	Quantitative		
23	champ_v9	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	Field V9 of the registration certificate which contains the Euro standard.		object	0.24%	Categorical		
24	date_maj	French Gouvernment (data.gouv.fr) - 2012 - 2015	Feature variable	The date of the last update.		Date	57.11%	Date	juin-13, mars-13, déc-12, mars-14, dÚc-13, sept-13 , juin-14, mars-15, déc-14	
25	Carrosserie	French Gouvernment (data.gouv.fr) - 2012 - 2014	Feature variable	Car's body type.		object	12.98%	Categorical	BERLINE' , 'COUPE', 'CABRIOLET' , 'BREAK', 'TS TERRAINS/CHEMINS', 'MONOSPACE', 'COMBISPACE', 'MINISPACE', 'MINIBUS', 'MONOSPACE COMPACT', 'COMBISPCACE'	
26	gamme	French Gouvernment (data.gouv.fr) - 2012 - 2014	Feature variable	The range of the car in term of quality (luxury car,...)		object	12.98%	Categorical	'MOY-SUPER' , 'INFERIEURE' , 'LUXE', 'SUPERIEURE', 'MOY-INFER', 'ECONOMIQUE', 'COUPE' , 'MOY-INF', 'MOY-INFERIEURE'	
27	puiss_heure	French Gouvernment (data.gouv.fr) - Only 2015	Feature variable	Electric motor power (KW).		float64	99.44%	Quantitative		
28	mrq_utac	French Gouvernment (data.gouv.fr) - Only 2015	Feature variable	The brand of the car manufacturer		object	87.00%	Categorical		
Variables potentialy to be added present on the European Union Dataset										
29	id	European Union Dataset - 2012 - 2015	Feature variable	Identification of the car.		int64	0.00%	Categorical		
30	MS	European Union Dataset - 2012 - 2015	Feature variable	Country where the car has been bought.		object	0.00%	Categorical		
31	MP	European Union Dataset - 2012 - 2015	Feature variable	Group of Manufacturers (here PSA group and not the manufacturer like Peugeot, this variable is different from the variable lib_mrq_utac).		object	13.66%	Categorical		
32	Mh	European Union Dataset - 2012 - 2015	Feature variable	Manufacturer name (EU Standards) (i.e. VOLKSWAGEN)- Common with the lib_mrq_utac variable.		object	0.00%	Categorical		
33	MAN	European Union Dataset - 2012 - 2015	Feature variable	Manufacturer name (OEM declaration) (i.e. VOLKSWAGEN AG).		object	0.00%	Categorical		
34	MMS	European Union Dataset - 2012 - 2015	Feature variable	Manufacturer name (MS Registry Denomination) (i.e. VOLKSWAGEN AG).		object	10.73%	Categorical		
35	TAN	European Union Dataset - 2012 - 2015	Feature variable	Type Approval Number.		object	6.70%	Categorical		
36	T	European Union Dataset - 2012 - 2015	Feature variable	The Variant-Variant (TVV) - 1. The type which regroup all the identical information on some technical points		object	0.86%	Categorical		The combination of the 3 variables give the TVV variable of the French Government Dataset.
37	Va	European Union Dataset - 2012 - 2015	Feature variable	The Variant-Variant (TVV) - 2. The variant if the car have different model		object	2.98%	Categorical		
38	Ve	European Union Dataset - 2012 - 2015	Feature variable	The Variant-Variant (TVV) - 3. The version which give the different finition on a car		object	4.96%	Categorical		
39	Mk	European Union Dataset - 2012 - 2015	Feature variable	Manufacturer name (i.e. VOLKSWAGEN)- Common with the lib_mrq_utac variable.		object	3.97%	Categorical		
40	Cn	European Union Dataset - 2012 - 2015	Feature variable	Business name of the car.		object	0.28%	Categorical		
41	Ct	European Union Dataset - 2012 - 2015	Feature variable	Category of the Vehicule type approved: - M = Vehicule having at least four wheels and used for the carirage of passengers - N = Power-driven vehicles having at least four wheels and used for the carriage of goods		object	1.05%	Categorical	M1', 'M1G', 'N1', 'N1G', 'N1 inc'	
42	r	European Union Dataset - 2012 - 2015	Feature variable	Total new registration.		int64	0.00%	Categorical		
43	e (g/km)	European Union Dataset - 2012 - 2015	Target variable	CO2 Emission (in g/km) of the car - Common with the Co2 variables of the French dataset.		float64	0.29%	Quantitative	Distribution on this dataset has not been done.	
44	m (kg)	European Union Dataset - 2012 - 2015	Feature variable	Mass in running order (kg).		float64	0.13%	Quantitative	Distribution on this dataset has not been done.	
45	w (mm)	European Union Dataset - 2012 - 2015	Feature variable	Wheelbase in mm.		float64	4.39%	Quantitative	Distribution on this dataset has not been done.	
46	at1 (mm)	European Union Dataset - 2012 - 2015	Feature variable	Axle width steering axle in mm (track width).		float64	5.26%	Quantitative	Distribution on this dataset has not been done.	
47	at2 (mm)	European Union Dataset - 2012 - 2015	Feature variable	Axle width other Axle in mm.		float64	12.25%	Quantitative	Distribution on this dataset has not been done.	
48	Ft	European Union Dataset - 2012 - 2015	Feature variable	Fuel Type.		object	0.77%	Categorical	Diesel', 'Petrol', 'LPG', 'NG-biomethane', 'Electric', 'E85', 'Petrol-electric', 'Diesel-electric', 'Hybride petrole', 'Petrol PHEV', 'Hydrogen', 'Biodiesel', 'Petrol-Gas', 'CNG', 'other'	
49	Fm	European Union Dataset - 2012 - 2015	Feature variable	Fuel mode.		object	1.98%	Categorical	M', 'B', 'F', 'E', 'n', 'NA'	
50	ec (cm3)	European Union Dataset - 2012 - 2015	Feature variable	Engine capacity in cm3 (volume of all the cyclinders).		float64	1.88%	Quantitative	Distribution on this dataset has not been done.	
51	ep (KW)	European Union Dataset - 2012 - 2015	Feature variable	Engine Power (kW) - Common with puiss_max variable of the French Government dataset.		float64	19.49%	Quantitative	Distribution on this dataset has not been done.	
52	z (Wh/km)	European Union Dataset - 2012 - 2015	Feature variable	Electricity energy consumption in Wh/km		float64	99.69%	Quantitative	Distribution on this dataset has not been done.	
53	IT	European Union Dataset - 2012 - 2015	Feature variable	Innovative technology or group of innovative technologies.		object	98.97%	Categorical		
54	Er (g/km)	European Union Dataset - 2012 - 2015	Feature variable	Emissions reduction through innovative technologies in g/km.		float64	99.21%	Quantitative	Distribution on this dataset has not been done.	