

CO2 Emission by vehicles

Group Members: Alexander Perez , Andreas Lathan, Victor Weissenbach, Tillmann Stralka

Identifying the vehicles that emit the most CO2 is important to identify the technical characteristics that play a role in pollution. Predicting this pollution in advance makes it possible to prevent the appearance of new types of vehicles (new series of cars for example).

Report 1: exploration, data visualization and data pre-processing report

Introduction to the project

Context

Context of the project's integration into your business.

Starting in 2035, the EU mandates a fleet-wide CO2 emission target of 0 g CO2/km for both passenger cars and vans, representing a 100% reduction. To meet this ambitious goal, machine learning can play a pivotal role for automotive manufacturers. By analyzing extensive datasets on vehicle design, performance, and emissions, machine learning algorithms can uncover critical factors that influence CO2 emissions.

This insight enables companies to design and develop vehicles that minimize emissions, ensuring compliance with stringent regulations and avoiding costly penalties while enhancing their brand's reputation for environmental responsibility. Additionally, machine learning can optimize production strategies through early design adjustments, providing a competitive edge in a market that increasingly prioritizes eco-friendly solutions¹.

The EU's push for zero CO2 emissions and the electrification of the transport sector poses significant challenges in developing new heaters for caravans and motorhomes that utilize alternative energy sources like hydrogen or electricity. This transition necessitates a reevaluation of the company's business strategy, emphasizing careful resource allocation, investment priorities, and timelines to ensure the development of future appliances that align with zero-emission and pollutant reduction objectives.

¹ This is also a professional insight of one of our team members, who worked at a manufacturer of LPG (liquefied petroleum gas) appliances for the caravanning industry.

From a technical point of view.

- Apply advanced data analytics and ML to predict CO2 emission test results.
- Key steps: data preprocessing, feature selection and engineering, model selection and training and model evaluation.
- Develop real-time prediction tools for new designs.
- Innovative Technologies (IT: column in EU-Dataset) have been developed and used for CO2 emission reductions. How they roughly work would be interesting in the sense of a possible “cross- application” in other combustions-processes/industries.

From an economic point of view.

- Avoid penalties and fines.
- Reduce redesign costs.
- Enable competitive pricing for greener vehicles.
- Access government incentives for low-emission vehicles.
- Get a free marketing campaign and promotion by gaining a green-label, best in class, CO2 friendliest car/Manufacturer.

From a scientific point of view.

- Contribute to climate change mitigation research
- Identify patterns in vehicle data affecting emissions
- Support development of sustainable vehicle technologies
- It is important though to keep in mind, that the scope of this analysis with respect to CO2 Emissions of cars and its possible impact on climate change mitigation is very limited: In the life cycle of a product, there are CO2 (and other greenhouse gas) emissions during resource extraction, manufacturing, trade, retail and transport to the customer, operation, and disposal/recycling. Here, we are focusing only on the operation phase. Also, we are not looking at the actual emissions during the customer’s usage, but rather at a proxy result, which is derived from a standardized test (WLTP). Although this test is intended to simulate this usage, evaluating the quality or effectiveness of this simulation is largely beyond the scope of this project.

Objectives

What are the main objectives to be achieved?

- Finding correlations of technical characteristics / reported features of cars and CO2 emission test results.
- Predict CO2 test results from technical characteristics / reported features.

For each member of the group, specify the level of expertise around the problem addressed?

- Tillmann: No background, general technical understanding of internal combustion engine, but not an engineer or experienced in car commerce.
- Andreas: Physics and economics background. Worked in environmental NGOs, specifically on the climate issue. No specific knowledge about cars.
- Alexander: Environmental and resource economics background, however, no specific knowledge in this area.
- Victor: Mechanical Engineering background, experience as a Test Engineer for LPG appliances, including measurements of CO₂, CO, NO_x, C_xH_y emissions.

Have you contacted business experts to refine the problem and the underlying models?

No.

Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?

No.

Understanding and manipulation of data

Framework

Which set(s) of data(s) did you use to achieve the objectives of your project?

We explored three different datasets.

1. A EU data set of newly registered cars and respective emissions named “CO₂ emissions from new passenger cars” (2010-2023) provided by the European Environment Agency (EEA): [EEA_Europe](#)

2. An older french Dataset (2001-2015): [French Gov \(2001-2015\)](#)

3. A newer, and in certain aspects more detailed french Dataset: [French_Gov_ADEME](#)

After extensive exploration, we concluded that the older French dataset is likely not worth including in our analysis. There are several reasons for this: the recorded and registered data only extends from 2001 to 2012, making it impossible to analyze modern vehicle models and their technical characteristics or to derive properties that favor a reduction in CO₂ emissions. Additionally, the

CO2 emissions data is less reliable, especially considering the Volkswagen diesel emissions scandal from 2015.

Comparing CO2 emissions from before 2018 to after 2018 is also challenging due to changes in EU legislation regarding emissions measurements. Furthermore, the different annual datasets show inconsistencies in data completeness and information content, reducing their trustworthiness. Nevertheless, they were merged by mapping similar columns and concatenating the years (see appendix).

We are still unsure if and how much of the both remaining datasets we should use for our further analysis, but for the time being **we settled on using the EU (EEA) dataset and possibly integrating the newer french (ADEME) dataset into our analysis (as it is very clean and has a lot of features).**

Are these data freely available? If not, who owns the data?

ADEME -Labeling- dataset is freely available:
Open Licence version 2.0

EU(EEA):
They are free and provided by the European Environment Agency.



Describe the volume of your dataset

ADEME -Labeling- dataset (last update: 27.09.2024)
Dataset as .csv is ~1MB big: 3604 rows, 52 columns
("after duplicates removal: 2309 rows")

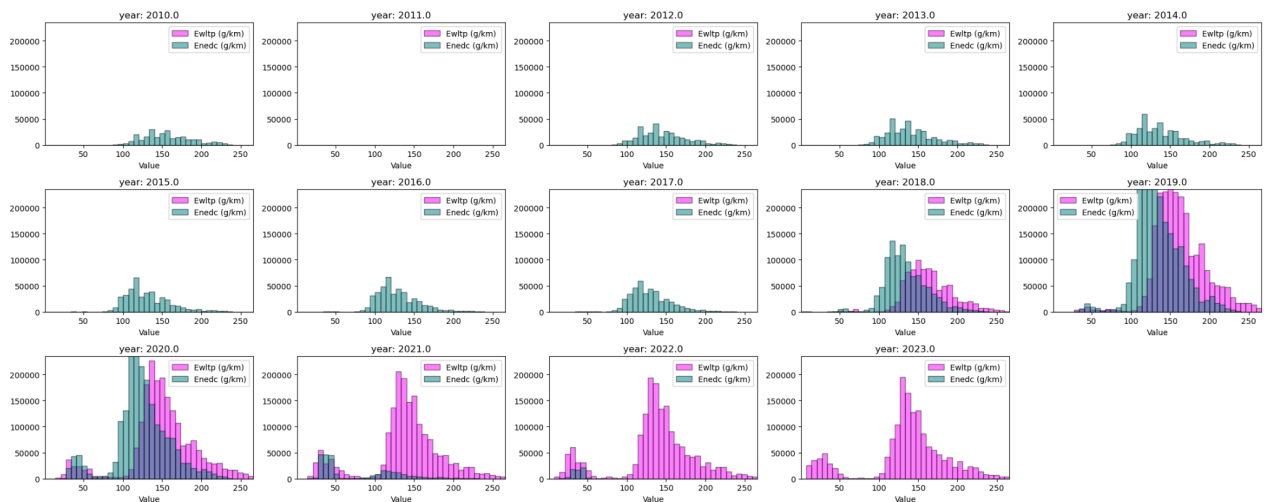
EU(EEA):

The size of the downloaded files for all years was over 16 GB. Through transformation of data types and grouping/collapsing of similar rows it was reduced to ~ 300 mb (see appendix). The shape of the resulting data set is 14.442.792 rows, 41 columns. This refers to the whole Dataset of 30 countries and the timespan of 2010 to 2023. Later on in our analysis we might focus on a subset of the Data.

Relevance

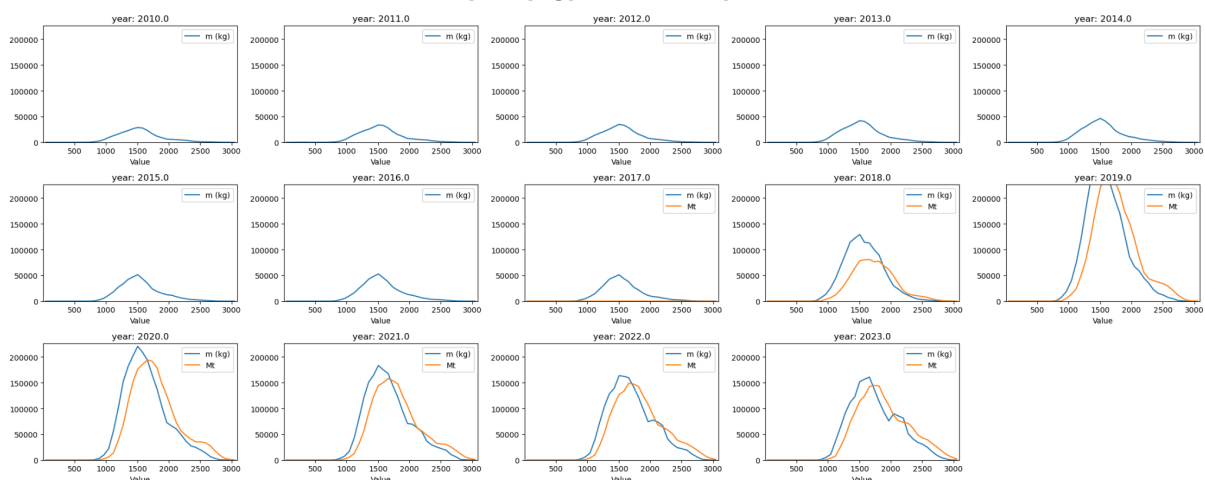
Which variables seem most relevant to you with regard to your objectives?

- Target variable(s)
 - EU (EEA)
 - Eneadc (g/km): old measurement standard until 2019/20
 - Ewltp (g/km): new measurement standard from 2018/19 onwards

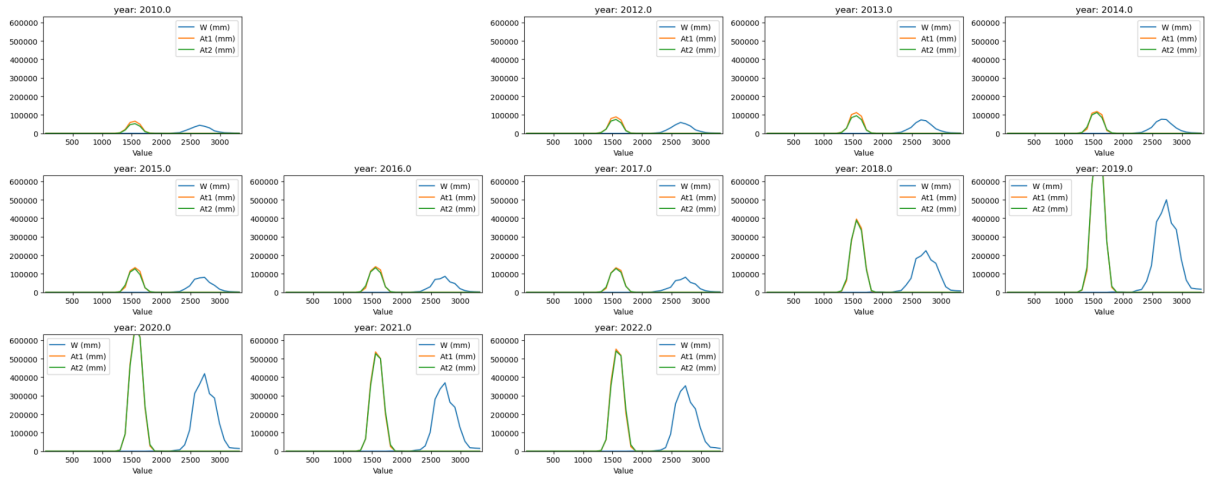


One can see how the older measure (NEDC) is gradually replaced by the newer (WLTP). In later years one can also observe the emergence of a second peak in the lower ranges due to the arrival of electric cars. For further analysis these would probably better be analyzed separately.

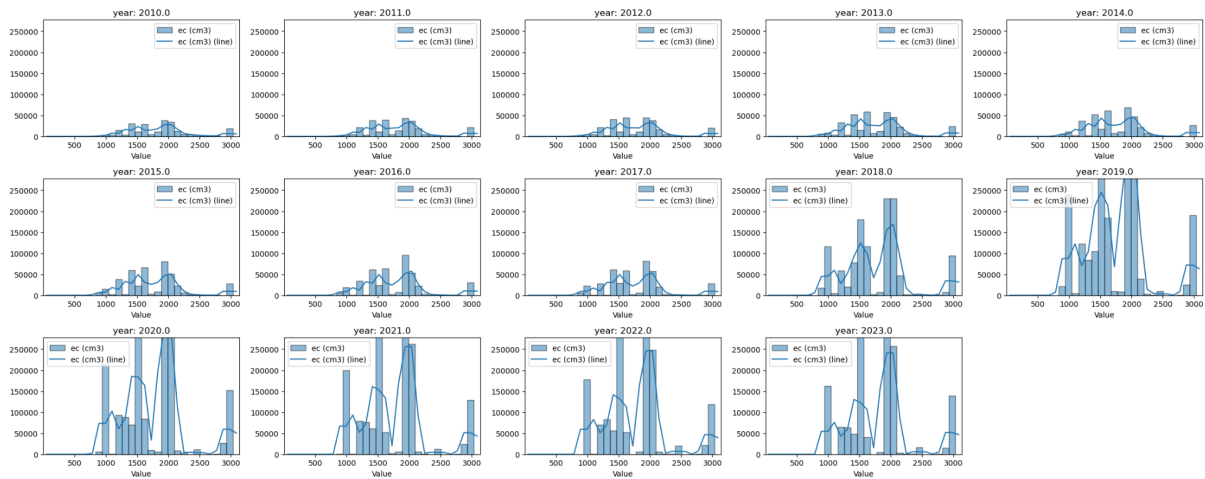
- Features/Explanatory variables:
 - EU(EEA):
 - Promising **numerical** explanatory variables:
 - mass (“m (kg)” and “Mt”)



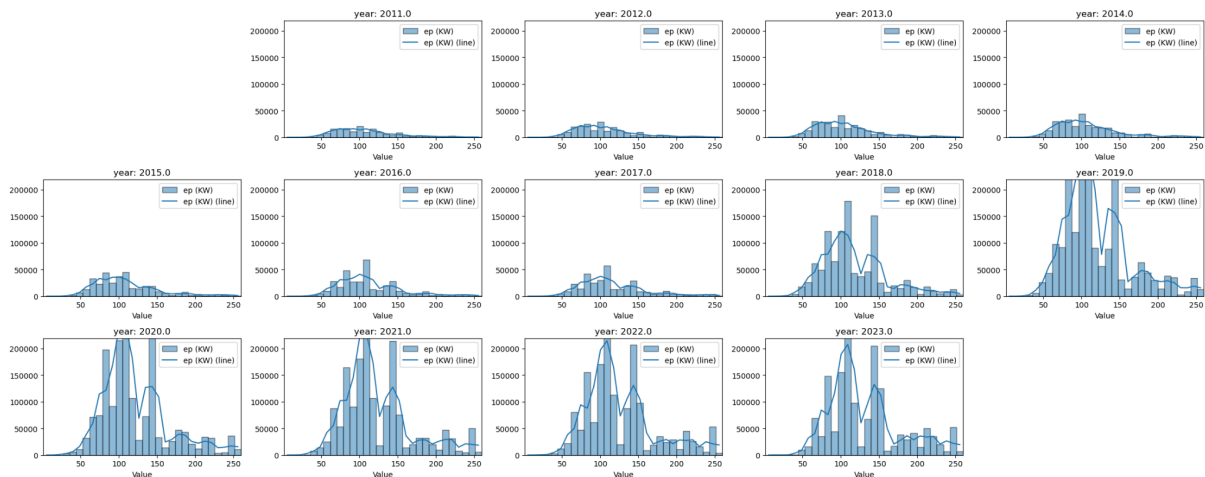
- wheelbase and axle widths (“W (mm)”, “At1 (mm)”, “At2 (mm)”)



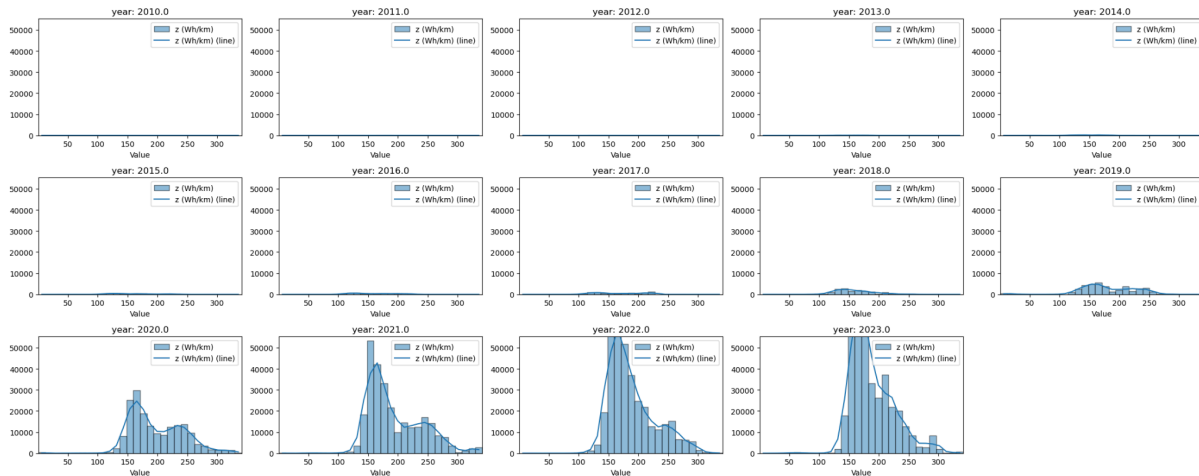
- engine capacity (“ec (cm3)”)



- engine power (“ep (KW)”)



- possibly **electric energy consumption (“z (Wh/km)”)**



We will still have to see how we handle electric and hybrid engines, possibly we will analyze them separately from combustion engines.

■ Promising **categorical** explanatory variables:

- **fuel type and mode (“Ft”, “Fm”)**
- **use of innovative technologies (“IT”, possibly in combination with numerical vars that quantify the emissions reduction through those technologies: “Ernedc (g/km)”, “Erwltp (g/km)”)**
- **possibly EU vehicle category (“Ct”, “Cr”) and manufacturer/brand (“Mh”, “Man”, “MMS”, “Mk”)** although these are not technical characteristics but could still serve as proxies for unobserved technical characteristics.

■ Promising **variables for grouping** of individual entries(cars) to types/ classes:

- **Type Approval number (“Tan”), and Type/Variant/Version (“T”, “Va”, “Ve”)**
- Possibly **“r”** should also be mentioned here, which before 2018 seems to represent the total number of cars of a specific type registered in the respective year. **After 2018 the dataset apparently switches from showing types/classes of vehicles to individual cars.** Possibly we should attempt similar groupings for the years after 2018, probably by Tan and/or T/Va/Ve.

- **ADEME:** Engine power: nominal, max., engine size: ccm, vehicle weight, further pollutants: NoX..
- **Are you limited by some of your data?**

ADEME: In the sense of data quality: time and spatial coverage are not specified.

EU (EEA): Different measures of CO2 emissions over the years. Last data is from 2023 and marked as preliminary.

Pre-processing and feature engineering

- **Did you have to clean and process the data? If yes, describe your treatment process.**
 - In the appendix (see appendix below) and GitHub repository².
 - **Major Steps:** Changing data types, consolidating rows by grouping and dropping duplicates, renaming/mapping different spellings, deleting empty columns, decreasing storage size through parquet format (see Appendix for details).
- **Did you have to carry out normalization/standardization type transformations of your data? If yes, why?**
 - Not yet, only for some visualizations.
- **Are you considering dimension reduction techniques in the modeling part? If yes, why?**
 - We are thinking about consolidating “T” “Va” and “Ve” into a single variable. Also possibly manufacturer/brand into one column - if we use it at all, because it is not a technical characteristic, albeit a potential proxy.

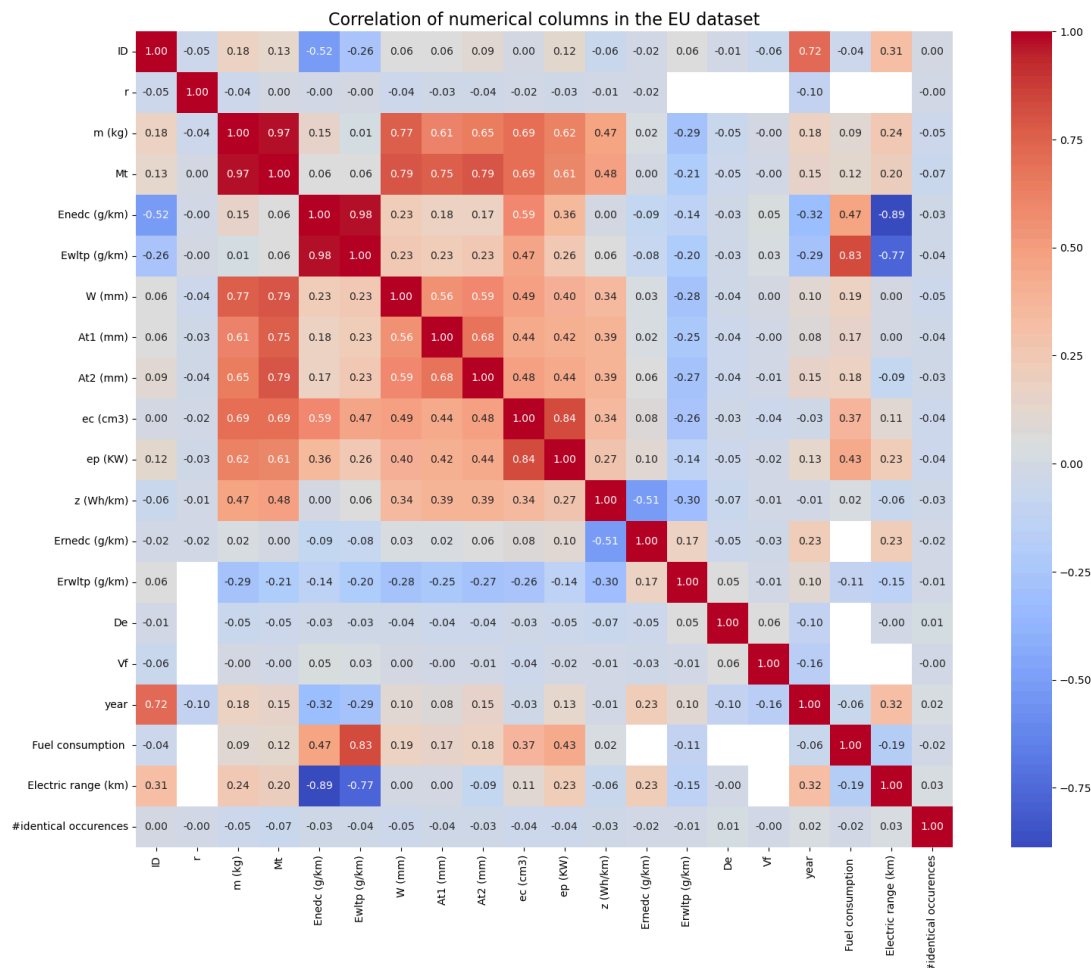
Visualizations and Statistics

Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?

The following plot shows a quick correlation heatmap of all numerical columns from the EU dataset. It provides a clear, quick analysis of the data and

² In the uploaded jupyter notebooks , the steps of each data processing for the three different sources are described.

unsurprisingly highlights that engine power ('ec' & 'ep') as well as 'Fuel consumption' do correlate with the CO2 emissions (Ewltp) of the vehicle.



Additionally, the physical dimensions ('W', 'At1', 'At2') of the car also seem to have an impact on the target variable. There is also a relationship between the explanatory variables, particularly between fuel consumption and engine power (ec, ep). Once again, the physical dimensions of the vehicle are relevant. It is reasonable to assume that the larger the car, the more fuel it consumes.

Describe the distribution of these data, distribution, outliers (pre/post processing if necessary)

See histograms above. Potential outliers have already been identified. These have been disregarded for purposes of clearer visualization but have not been dropped from the dataframe yet.

Present the statistical analyzes used to confirm the information present on the graphs.

See heat map above.

Draw conclusions from the elements noted above allowing them to project themselves into the modeling part.

The EU dataset suggests that we will be dealing with supervised machine learning models. Given the mixture of data types (categorical and numerical), we already know that categorical variables will need encoding (one-hot or label encoding), while numerical variables will need (further) scaling and outlier handling.

Regarding modeling, multiple linear regression is the first model candidate given the nature of the problem at hand, however, other models can be tested and compared, such as those that do not make assumptions about the relationship between features and target, normally distributed residuals, and absent multicollinearity. Examples of such models are decision trees, random forests and gradient boosting. The comparison in performance will include hyperparameter tuning via cross validation.


----- APPENDIX - PRE-WORK ON DATASETS -----

Work on French ADEME -Labeling- Data Victor

French ADEME-labeling- data comes from source: [Link_French_ADEME](#)


Relation with the “French Gov Data” as provided link to datasets in the ‘CO2 emissions by vehicles’ project’s definition: -> Referenced by possible discussion keeper ([Link](#))

Data update for 2016 to todayCopy discussion permalink



Ghislain Trabichet
July 7, 2023

Hello, We would like to analyze the evolution of CO2 emissions from used cars in France through a statistical study. Is it possible to have the same information from 2016 to today? Thank you in advance for your help, Have a nice day!



Laurent MORICE
July 10, 2023

Hello, The updated datasets are available here: <https://www.data.gouv.fr/fr/datasets/ademe-car-labelling/> Kind regards,

As stated by the source page, the dataset:

“provides source data, reworked from various sources, to enable the emergence of new applications that will encourage the acquisition of vehicles with the least impact or additional analyses to inform decision-makers and purchasers.”

The dataset contains information of “actual” car models with its technical characteristics, CO2 emissions, fuel/energy consumption and even the price. The last update of the dataset was on 14.09.2024 but it assumed that it also contains registered cars for the year 2023 and even 2021 since it was created on 19.04.2021. No date information for each entry is provided in the data set.

Workflow:

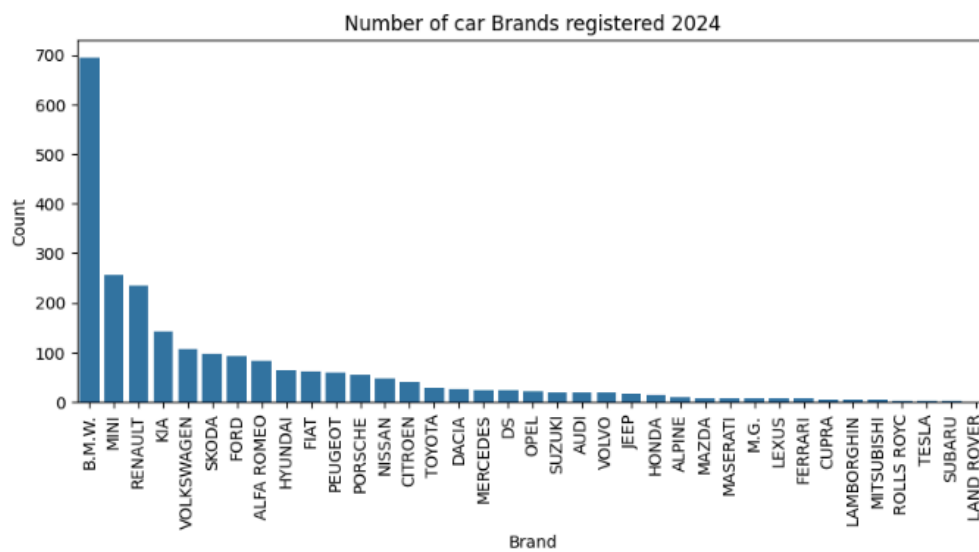
Loading Data:

- Download the csv-file and upload it to google-colab.
- Download the Field Description file.
- Create a notebook.
- Recognize the data encoding.
- Load data with the found data encoding-type and set decimal to “,” into a dataframe.
- Find and remove duplicates (1295).
- Create dictionaries for the column names and aliases (abbreviations) FR -> EN.

- Create dictionaries for the classes in the categorical variables/columns FR -> EN.
- Exploring unique values of “Max:Power” to future round values and changing var type from float to int.
- Exploring data by visualizations.
- Create a new dataframe with only numerical variables -> heatmap.

First exploratory diagrams and results:

Distribution of car-brands:

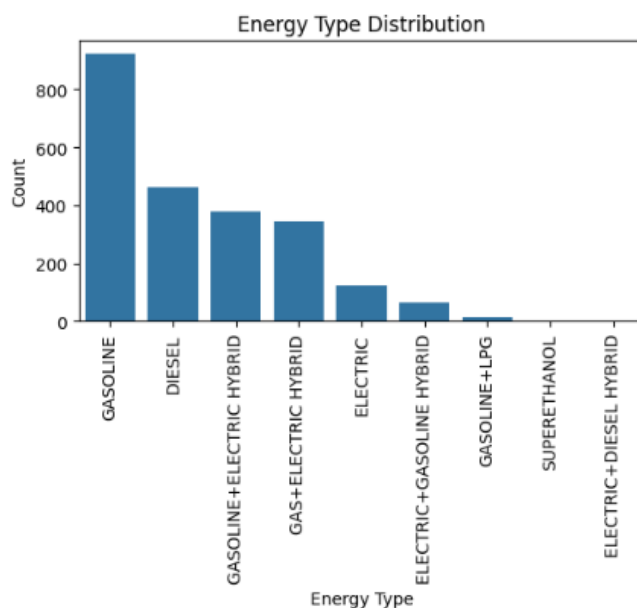


TOP 3: BMW,
MINI,
RENAULT

Distribution
of Energy

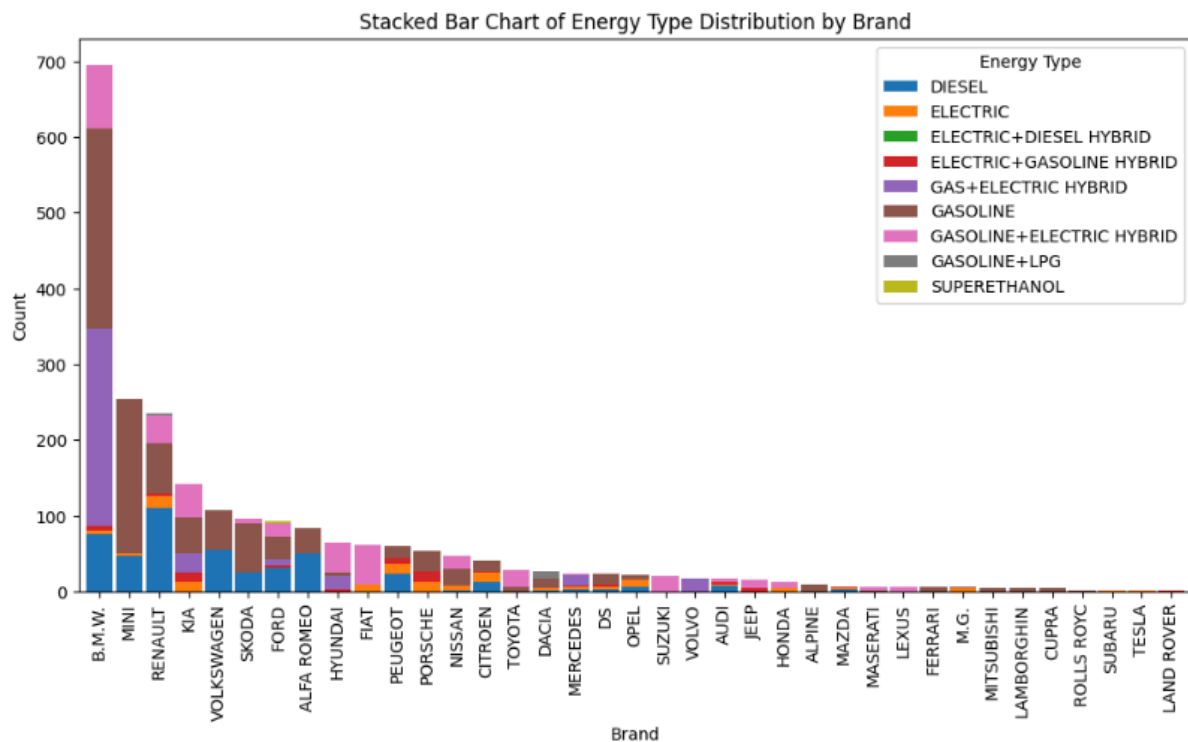
type::

TOP 3 counts: BMW, MINI, RENAULT



Distribution of Brand and Energy type:

TOP 3 counts: BMW, MINI, RENAULT with mostly Gasoline and Diesel



Heatmap of all numerical variables:

General: too many vars on a single plot.

Blank spots: Electric consumptions, ranges and urban ranges

-> suggest a separation of energy type: fossil fuels + electric

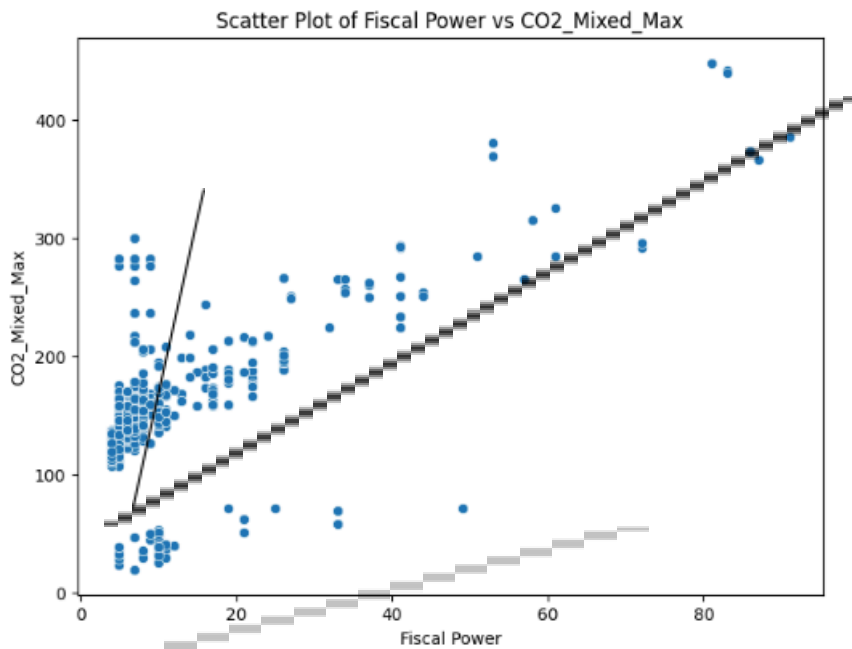
-> suggest dropping electric powered vehicles (no CO2 emissions)

Engine capacity correlates with engine capacity.

CO2 emissions correlate with fuel consumptions

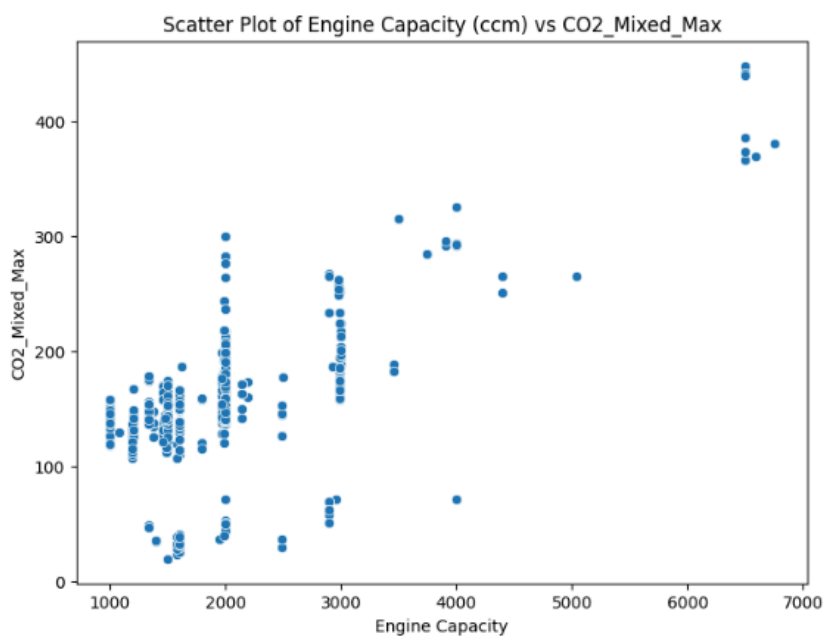
Scatterplot of Fiscal Power against CO2_Mixed_Max:

Suggest a linear relationship, three groups -> further separation needed



Scatterplot of Engine Capacity against CO2_Mixed_Max:

Linear relationship, two groups -> suggest further separation

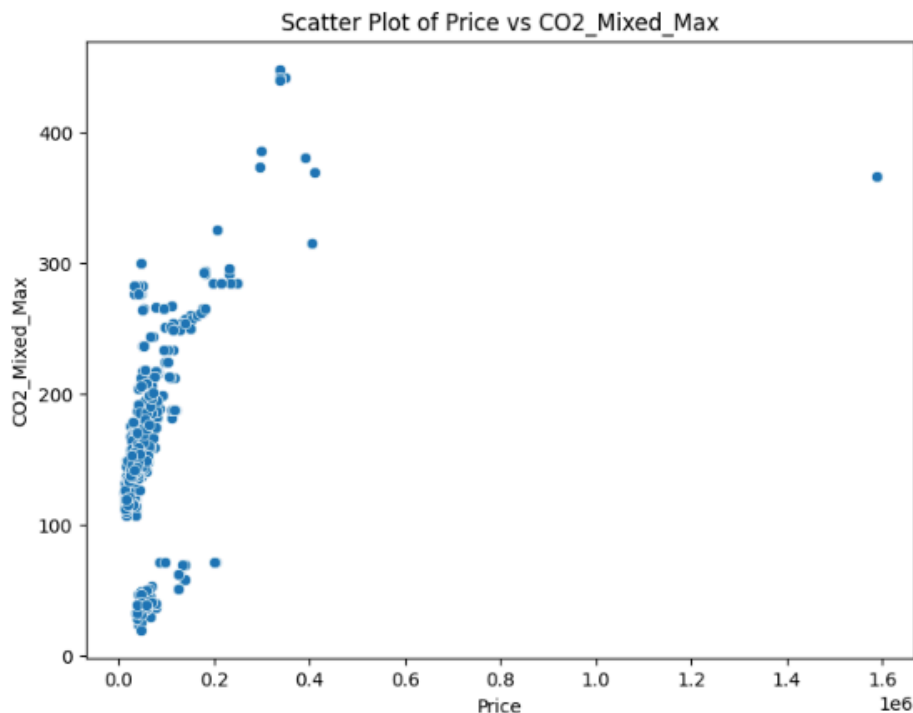


Scatterplot of Price (€) against CO2_Mixed_Max:

Linear relationship, two groups -> suggest further separation

Potential price outlier "Ferrari?"

Subgroup of expensive cars with high CO2 emissions



Result:

Dataset must be further analyzed and preprocessed []

Starting j-notebook to be pushed on GitHub [x]

Other work also done during exploring the datasets for the project:

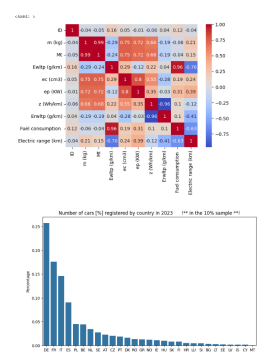
CO2 EU-Dataset: - "BIG-Data"

Downloaded Year 2023, all countries (4,3 GB as csv file)

Unable to open it, even google colab with 12,7GB of RAM did reset...

Opted for sampling 10%, still 1 Million rows, 40 columns, which at least google colab could open/load as a dataframe. Table of definition did not match, 3 vars are not included and changed by other vars/column-names. Dropped all empty columns in the 10% sample dataframe. Made a heatmap. Made a count-plot normalized of Number of cars [%] registered by country in 2023.

Thanks to Andreas who shrunk the dataset it is now manageable.



French Gov: - Messy - not updated -> "Go for Data Quality"

Downloaded Year 2015 only, Dataset is ~4.2MB big 20880 rows, 26 columns.

Downloaded dictionary/table of definition having guessed only a third of matching with the col. names.

Created and translated dicts: FR -> EN

No further exploration.

Work on French Gov Data

Tillmann

French Government data comes from source: [Link_French_Gov](#)

They contain the administration of cars in the years 2001 to 2012. The Excel files are downloaded and converted to CSV format, with `_year.csv` appended to the end of each file name. Process steps of the **jupyter notebook** (see **GitRepository**) are as followed:

Loading Data:

- Remove empty columns and rows
- Delete index columns
- Collect column names in a list (`column_names`)
- Append the year as a global variable to each DataFrame

Working with Specific Data:

- Replace commas with dots for float values
- Remove duplicate columns containing redundant information

Combine DataFrames into a List:

- Merge all DataFrames into a list to work simultaneously with multiple DataFrames

Overview:

- Get an overview of the shape and column names of each DataFrame

Mapping:

- Relabel column names into groups with common entries
- Assign clear, descriptive English names

Merging DataFrames:

- Combine all DataFrames into a single DataFrame
- Rename any remaining duplicate columns
- Save the combined DataFrame as `data_all`

Data Type Conversion:

- Convert columns with object data types to floats, where possible

Handling Missing Data:

- Fill missing values with NaNs

Saving the Data:

- Save the final DataFrame as a CSV file named `'data_all_french.csv'`

Result:

The combined dataset for the years 2001 to 2012, with 180,000 entries, is relatively small compared to the EU dataset. The key question is how many technically relevant data points it contains for building a predictive model for CO2 emissions. In particular, for the **earlier years**, there is very **limited technical data**, and its **reliability is questionable** since it was recorded over 20 years ago.

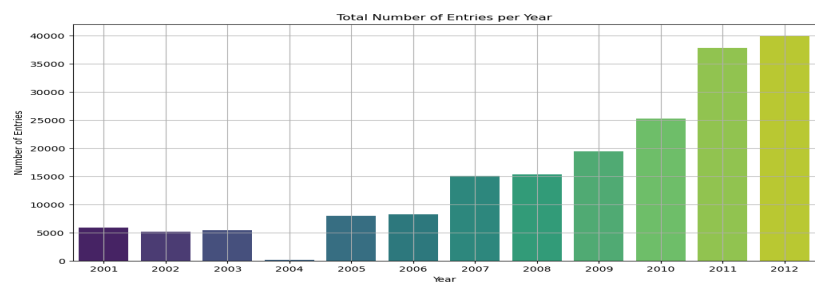
Additional variables appear in the years 2010, 2011, and 2012, these are **customer-oriented**, such as bonus malus systems, rather than technically significant. The technical variables are **less than 20%** of total row numbers (many missing entries).

```
[186030, 36]
<class 'pandas.core.frame.DataFrame'>
Int64Index: 186030 entries, 0 to 40051
Data columns (total 36 columns):
#   Column              Non-Null Count  Dtype
---  --
0   Brand               186030 non-null  object
1   Model              186030 non-null  object
2   Type Mines         110002 non-null  object
3   CMT                139945 non-null  object
4   Fuel Type          167469 non-null  object
5   CV                 14178 non-null   object
6   Engine Power        185838 non-null  object
7   Transmission        186030 non-null  object
8   Urban Consumption   185920 non-null  object
9   Extra Urban Consumption 185920 non-null  object
10  Combined Consumption 185923 non-null  object
11  CO2 Emissions       185923 non-null  object
12  YEAR                186030 non-null  int64
13  Unnamed             13701 non-null   object
14  Energy Label        93846 non-null   object
15  Fiscal Power        93982 non-null   object
16  Registration        43206 non-null   object
17  Puissance max       138 non-null     float64
18  Unnamed.2           3 non-null       object
19  Unnamed.3           4 non-null       object
20  CMT                8259 non-null   object
21  Bonus/Malus        60171 non-null   object
22  puiss_admin_98      77868 non-null   float64
23  V9 Field           77764 non-null   object
24  Model.2            40052 non-null   object
25  dscom              40052 non-null   object
26  hybride            40052 non-null   object
27  co_typ_1           39905 non-null   object
28  hc                 8357 non-null   object
29  nox                39905 non-null   object
30  hcnox             31686 non-null   object
31  ptcl              37355 non-null   object
32  masse_ordma_min     40052 non-null   float64
33  masse_ordma_max     40052 non-null   float64
34  Body Type          40052 non-null   object
35  gamme             40052 non-null   object
dtypes: float64(4), int64(1), object(31)
```

Car Registrations in France

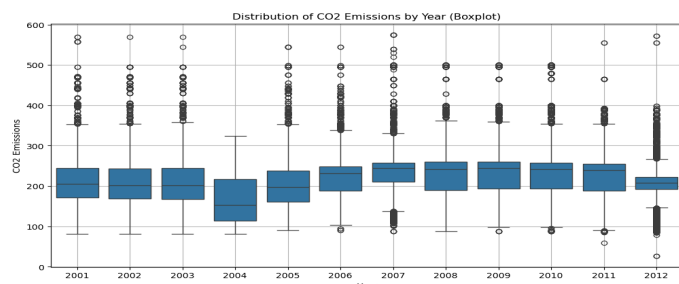
The data shows car registrations in France over the past 12 years. It's notable that registrations increase each year, which appears unrealistic, especially to such a

degree, as the total number of car registrations per year should remain roughly within the same range. Additionally, it's clear that the data for 2004 is highly corrupted, with significantly fewer data points available.



Outdated Data Concerns):

It's crucial to question whether working with such outdated data is appropriate, given that technical specifications, combustion engine types, and technologies have evolved significantly over the past 20 years. As a result, the data being used may no longer be relevant. This is also reflected in the noticeable decline in average CO2 emissions between 2001 and 2012.



ETL work on EU Data

Andreas

Initial ETL Process to create a single dataframe:

I **focused mainly on ETL work and on the EU data** set of newly registered cars and respective emissions named “CO2 emissions from new passenger cars”

[EEA Europe](#)

The goal was to get as much Data as possible into a workable format.

The EU site has Data from 2010 - 2023. Data for each given year in csv format was between 2 and 3 GB for recent years, and less for older, **16 GB altogether**. At first my machine wasn't even reading the file before I closed all other apps.

Therefore I looked for **ways to reduce the size** and came up with the following routine which I then successively applied to all years (see [notebook "EU Data read and reduce_year by year.ipynb"](#)):

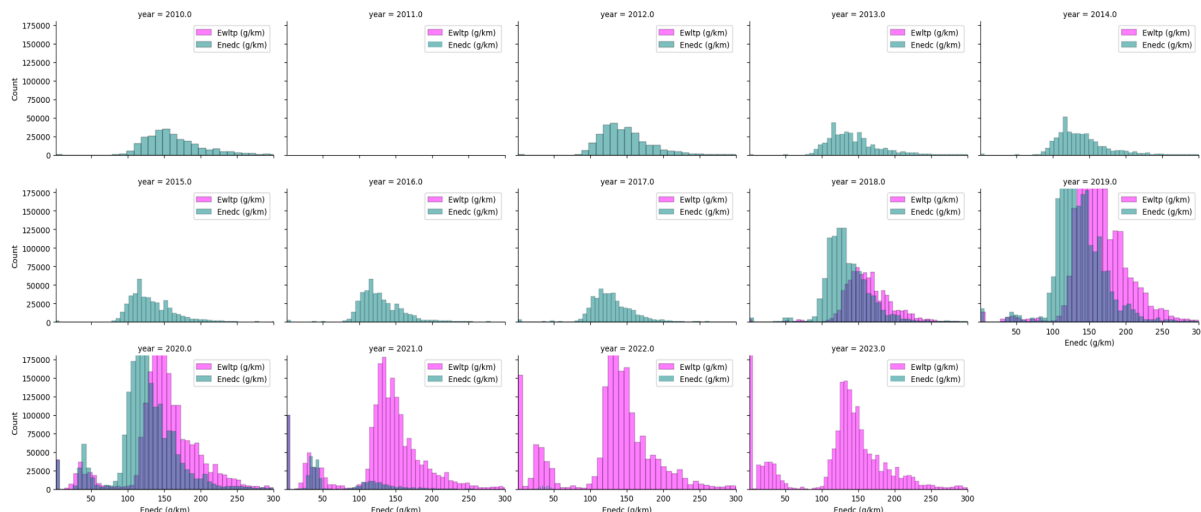
1. **Read Dataframe from csv with specified data types:** read all object type directly as category
2. **Further optimize data types for memory usage.** Downcast numerical types as much as possible
3. **As the dataset contains data of individual cars** I figured there must be a lot of cars that can be regarded as identical for our purposes and can therefore be removed w/o losing any information that could be used to improve our predictions. This was especially true for those cars/rows that are identical except for “ID” and “date of registration”.
Still, the information about how many identicals there were, could be valuable, so I calculated and stored this info in a new column “#identical occurrences” before dropping the “duplicate” rows. This step was not as easy as I expected, as the calculations repeatedly used too much memory and also the identification of “duplicate” rows failed because of NaN != NaN and it took a lot of experimenting until it finally worked.
So in effect the dataset was aggregated by collapsing records of identical car configurations, consolidating the data to focus on vehicle types rather than individual registration events.
4. **Stored the files in parquet format.**

These measures combined reduced the volume of the bigger, most recent files from > 2GB to under 50 MB each.

After that I **merged all years (2010-2023) into one big dataframe of around 14.4 mio. rows and about 300 MB**, which then became the basis of our preliminary exploration of the dataset. (see [notebook "EU Data_merge all years.ipynb"](#))

Exploration of variables:

Here e.g. is the **evolution of the distribution of the two possible target variables** over the years. One can see how the older measure (NEDC) is gradually replaced by the newer (WLTP). In later years one can also observe the emergence of a second peak in the lower ranges due to the arrival of electric cars. For further analysis these would probably better be analyzed separately. (generation of plots: [notebook "EU Data explore and further clean up.ipynb"](#), section "Plots")



Further ETL measures:

I also explored multiple further ways of cleaning, reducing and refining the data in meaningful ways. I was especially looking for **ways to remove redundant information** by further collapsing records of identical car configurations and consolidating the data to focus on vehicle types rather than individual registration events.(see [notebook "EU Data explore and further clean up.ipynb"](#), sections "Cleaning Columns" and following)

- **Consolidating Categories: fixed different spelling** of several categorical (string) variables: "Tan", "Ct", "Cr", "Fm", "Ft", "Country", "Mp".
- **Grouping/Removing/collapsing rows with identical features** in all but specific rows, e.g. "Country" and other non-technical categorical attributes which might not matter for our specific questions and modelisation.
- **Grouping by "Tan"** (Type Approval Number) and collapsing the data into a single row per group, **while retaining information about the features of the collapsed rows** by calculating and storing the means, variance, and other possible descriptors or measures of their respective distributions within each 'TAN' group or bin.
- Similarly **grouping by "tv"** a combination of T (type), "Va" (Variant) and "Ve"(Version), also **to possibly merge with the (older) french dataset** on this attribute ("tv").

Work on EU Data (all countries, time span: 2010 - 2023)

Alexander

These were the steps:

1-Check the meaning of each variable (building some domain knowledge). I decided to tackle first numeric variables:

2-Checking the ones that can be dropped at the onset - crude removal:

Removing columns with >80 % of missing values.

Rationale: Improving data quality (high levels of missing data can introduce bias and reduce the reliability of future analysis), enhance model performance (many machine learning algorithms struggle with missing data), and reduce computational cost. It is also unlikely that that much missing data can be replaced with imputation techniques without hampering the credibility of the model.

This was the output:

Columns dropped: Index(['z (Wh/km)', 'Ernedc (g/km)', 'Erwltp (g/km)', 'De', 'Vf', 'ech', 'RLFI', 'Electric range (km)']

Important is to note that our target is named 'Erwltp (g/km)', but the dataset shows two instances of the target. The one removed means: “emissions reduction through innovative technologies”, whereas the target variable meaning is “specific CO2 Emissions (WLTP)”, which was of course not removed.

Removing redundant variables

That is, different names of variables but containing the same information: Mp, Mh, Man, and Mk variables state the name of manufacturer according to different regulatory guidelines. The one with the fewer NaNs was retained, namely Mh.

```
Missing percentage for Mp: 10.65207475119769
Missing percentage for Mh: 0.0
Missing percentage for Man: 2.1517169256470634
Missing percentage for MMS: 41.877526173609645
Missing percentage for Mk: 0.29708937163950017
```

This is the state:

	Column Type	Missing Percentage
ID	float32	0.000000
Country	category	0.000000
Mh	category	0.000000
Tan	category	3.396455
T	category	0.429536
Va	category	0.695738
Ve	category	1.559415
Cn	category	0.341831
Ct	category	0.170943
Cr	category	19.463328
r	float32	0.000028
m (kg)	float32	0.097453
Ewltp (g/km)	float32	24.687734
W (mm)	float32	15.593183
At1 (mm)	float32	17.470348
At2 (mm)	float32	18.960551
Ft	category	0.046314
Fm	category	0.558604
ec (cm3)	float32	4.250529
ep (KW)	float32	9.703242
Status	category	0.000000
year	float32	0.000000
#identical occurrences	float32	0.000000

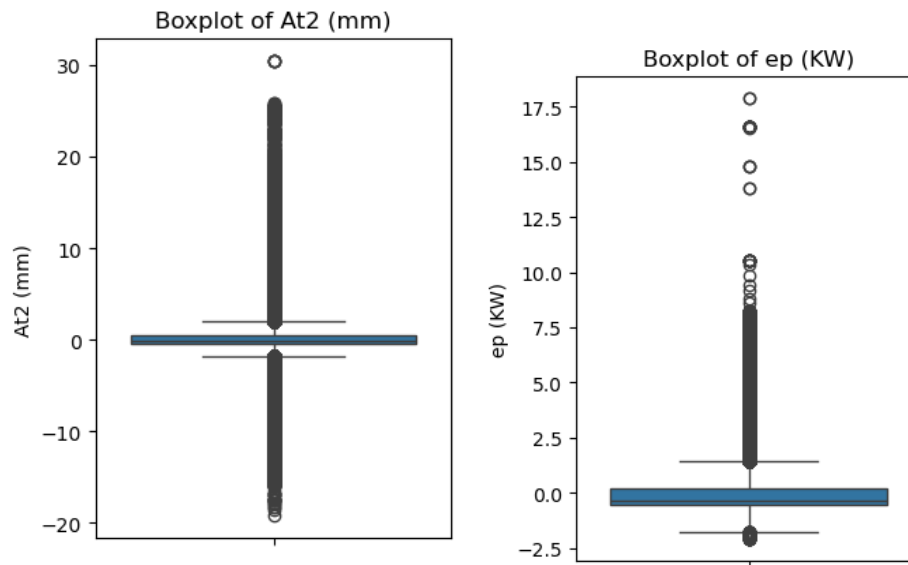
Removing rows of columns with low percentage of missing variables (<10%)

The dataset is large, hence removing rows with missing variables should not impair the data set. For instance, 'Tan', 'T', 'Ve', etc'.

Replacing missing data via imputation by variables with a high percentage of missing values (10%-80%):

After running some visualizations of the distribution of each of these variables, missing data points in variables with an apparent normal distribution were replaced by the mean (W, AT1 and AT2). Missing data points in variables with an apparent skewed distribution were replaced by the median (ec (cm3) and ep (KW)).

As an example, see the (normal) distribution of AT2 (mm) and (skewed) ep (KW):



Note: I did run a log-transformation, which in theory can help to reduce the skewness and to better visualize the data. However, the results were not superior.

The result of all of these procedures was an almost³ data set clean up of nans, large enough (10199165 rows), but with severely reduced variables (23 in total).

	Column Type	Missing Percentage
ID	float32	0.000000
Country	category	0.000000
Mh	category	0.000000
Tan	category	0.000000
T	category	0.000000
Va	category	0.000000
Ve	category	0.000000
Cn	category	0.000000
Ct	category	0.000000
Cr	category	0.000000
r	float32	0.000000
m (kg)	float32	0.001892
Ewltp (g/km)	float32	0.000000
W (mm)	float32	0.000000
At1 (mm)	float32	0.000000
At2 (mm)	float32	0.000000
Ft	category	0.000000
Fm	category	0.000000
ec (cm3)	float32	0.000000
ep (KW)	float32	0.000000
Status	category	0.000000
year	float32	0.000000
#identical occurrences	float32	0.000000
(10199165, 23)		

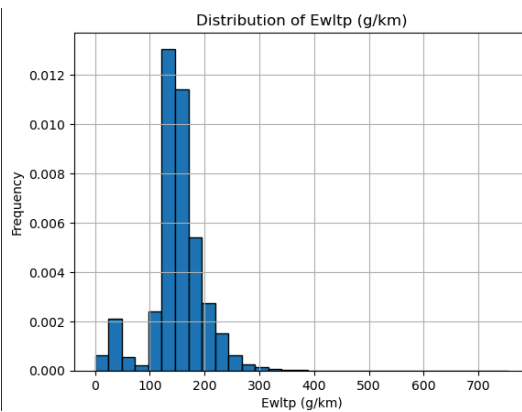
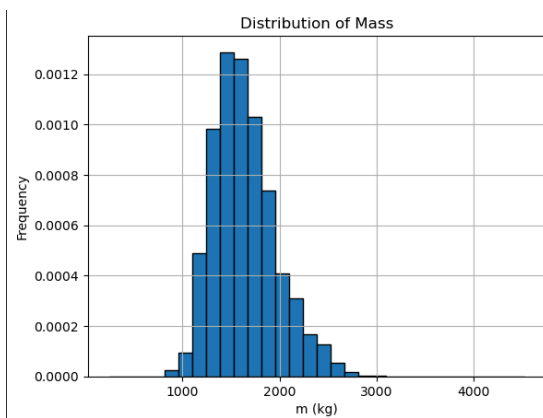
³ The variable still shows a low percentage of missing values.

3- Sample of data visualizations of numerical variables

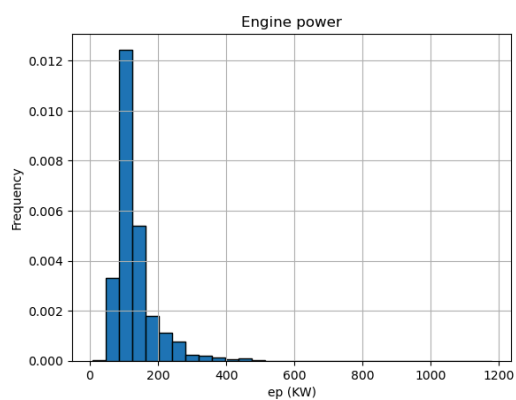
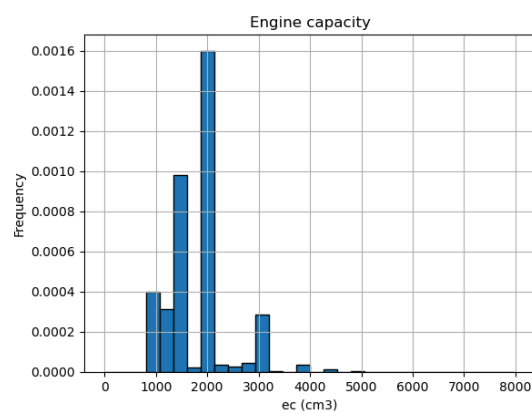
For the following graphs, binification was performed, in order to adjust the scale:

```
min_value = df['ec (cm3)'].min()
max_value = df['ec (cm3)'].max()
bin_width = (max_value - min_value) / 30
bins = np.arange(min_value, max_value + bin_width, bin_width)
```

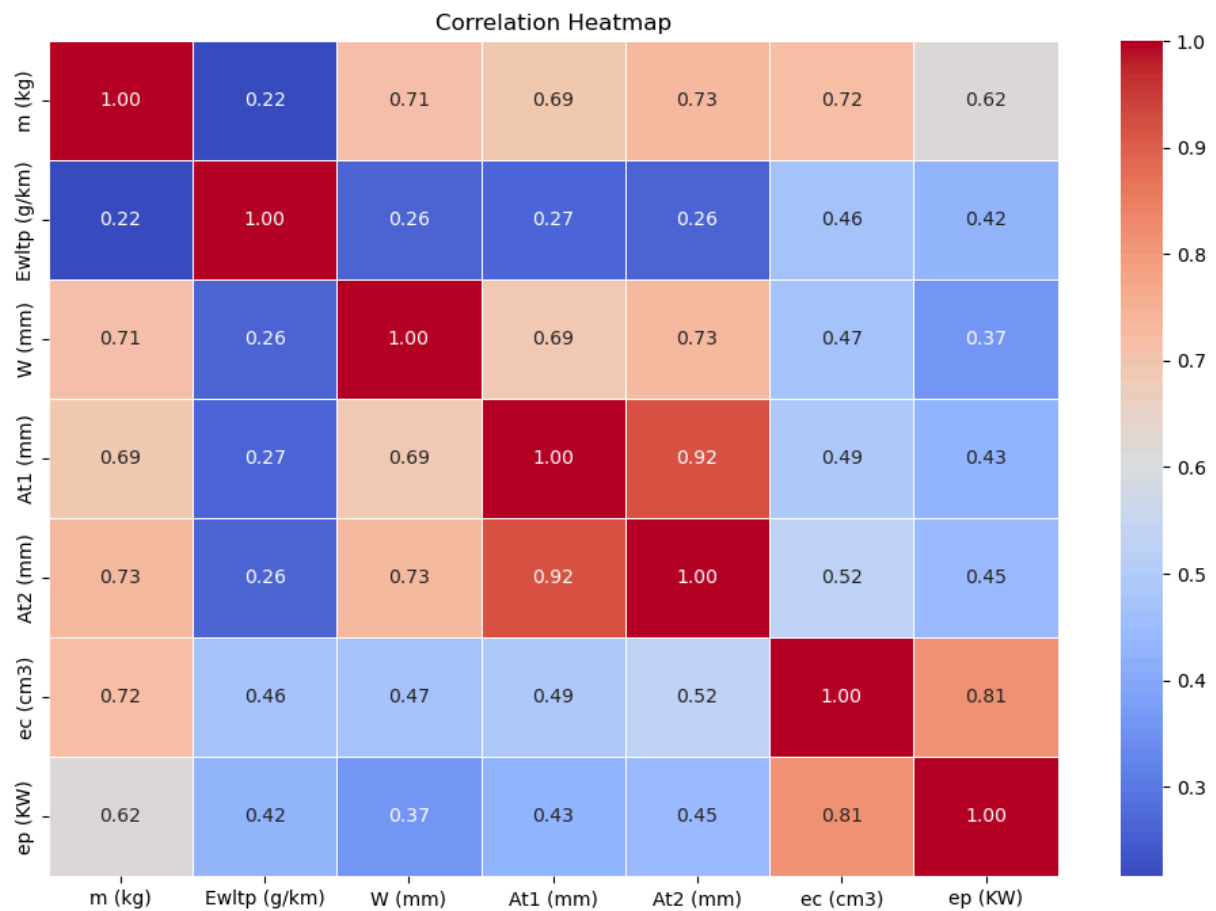
Also a graph enabling Kernel Density Estimation to get a probability density function (PDF) was attempted, but given the size of the data set, it did not work.



The target variable seems to be skewed.



Correlation heatmap



The At2 (mm), ec (cm³), and ep (kW) features should be given more attention, as they have the highest correlations with Ewltip and may offer useful insights or predictive power when modeling. Despite these correlations, none of the features are highly correlated with Ewltip, meaning that it could be challenging to model emissions using these variables alone.

Given this disappointing result of these correlations, intended statistical tests for outliers (IQR and Z-Score) and distributions (Kolmogorov-Smirnov Test) were not performed. Also moving towards categorical variables (meaning, encoding, etc) then statistical tests (ANOVA) were also stopped.