b
UNIVERSITÄT
BERN

Car Fuel Consumption and Emissions

Conceptual Design Report

Martinez Alberto, Scandella Davide

CAS Applied Data Science
University of Bern
Switzerland
19^{th} October, 2020

Contents

List of Figures

1 Project Objectives

This data-science project aims at examining the relationship between car fuel consumption and emission depending on car characteristics. The answer to the following questions is given:

1. How do car characteristics such as fuel type, gearbox, year of production, area of use, influence the consumption?
2. How does the CO_2 production scale depending on the type of fuel used?
3. How more convenient is a diesel car compared to a gasoline car? Are the same conclusions valid also for the new AdBlue technology?
4. Is is possible to predict fuel consumption of a car?

Thus the analysis objective is to uncover possible correlations between car details - fuel type, gearbox, year of production, type of use, power - and emission and consumption values as well as to focus on possible correlations in order to discover other factors that might influence other variables.

2 Methods

2.1 Hardware

The hardware used for this project is:

- MacBookPro 9.2 running on an Intel Core i5 processor with 2.50 GHz for each of the two cores and 4GB of RAM. The running OS is MacOS Mojave v10.14.1.
- Lenovo L590 running on an Intel Core i5 with 1.6 GHz for each of the four cores and 8GB of RAM. The running OS is Linux Mint v20.

The former laptop is used for all phases of the project, including web scraping. The latter is used for data analysis and post-processing.

2.2 Software

The main software used for the project is python 3.8.2. All used libraries and modules are installed on a virtual environment managed by Anaconda v1.9.12.

The following Python libraries were used in this project: beautifulsoup, pandas, numpy, scipy, sklearn, matplotlib, seaborn. For a detailed list of the required modules, please see the *requirements.txt* file on the GitHub project page.

Jupyter Notebook v6.1.1 is used in order to visualise and interactively assess the data.

2.3 Analysis Methods

Analysis methods adopted for the current project are expected to be descriptive statistics and statistical inference, i.e. through linear regression. The evaluation of more complex methods will be also re-assessed throughout the duration of the project, once an overview of available methods and results will be provided.

3 Data

The data used for the project is extracted from the web by means of scraping techniques. The source is `http://www.spritmonitor.de`. The extraction consists of two main steps:

1. Extraction of all the relevant cars from the search results of spritmonitor: `https://www.spritmonitor.de/en/overview/0-All_manufactures/0-All_models.html?fueltype=<FUEL_TYPE>&vehicletype=<VEHICLE_TYPE>&minkm=<MIN_KM>&page=<PAGE_NO>&powerunit=3`. Among all pieces of information, the link to the car details is extracted for each car.

2. Parsing of all the cars information by accessing to the specific vehicle page obtained from the previous step: `https://www.spritmonitor.de/en/<CAR_URL>.html`

For a detailed explanation of the parameters and relevant data, please refer to Section 5.

Step 1 is performed for all the cars, whereas Step 2 is performed on a 10% subset of the former. This is mainly driven by time constraints, but it was also considered as sufficient in order to have statistically-significant results.

The python code written and used for web scraping is available on the GitHub project page (commit `015d05ebaba62b60159a04a11b0071ca97638fda`).

Table 1 and Figure 1 provide an overview of the data field and a preliminary analysis of the descriptive statistics characterising the dataset by means of a pair plot.

	Brand	Model	Year	Power	Fuel_Type	Consumption	No_Fuelings	Gearbox	CO2
9	Dodge	Ram	2017	396	Gasoline	3.12	164	automatic	74.0
29	Ford	Fiesta	2017	80	Gasoline	3.30	17	manual	77.0
39	Toyota	Yaris	2011	107	Gasoline	3.36	5	automatic	78.0
49	Toyota	Prius	2016	121	Gasoline	3.43	81	continuously variable	80.0
59	Rex	Rexy 50	1997	1	Gasoline	3.47	21	continuously variable	81.0
79	Hyundai	IONIQ	2017	139	Gasoline	3.52	23	direct shift gearbox	82.0
89	Opel	Corsa	2001	58	Gasoline	3.56	5	manual	83.0
99	Opel	Corsa	1994	44	Gasoline	3.59	55	manual	84.0
109	Nissan	Pixo	2013	67	Gasoline	3.62	155	manual	84.0
119	Toyota	Prius	2019	121	Gasoline	3.64	13	continuously variable	85.0

	total_km	min_cons	max_cons	motorway	city	countryroads
9	70229.0	0.65	16.74	0.339	0.331	0.331
29	8206.0	NaN	NaN	0.408	0.408	0.183
39	2800.0	NaN	NaN	0.000	0.500	0.500
49	63900.0	2.72	4.35	0.472	0.222	0.306
59	1470.0	2.79	4.48	NaN	NaN	NaN
79	17692.0	2.99	4.48	0.215	0.393	0.393
89	2470.0	NaN	NaN	0.000	0.272	0.728
99	92280.0	3.00	4.15	1.000	0.000	0.000
109	191005.0	3.21	4.04	0.500	0.000	0.500
119	9956.0	3.31	4.04	0.099	0.666	0.235

Table 1: Subset of dataset containing relevant columns

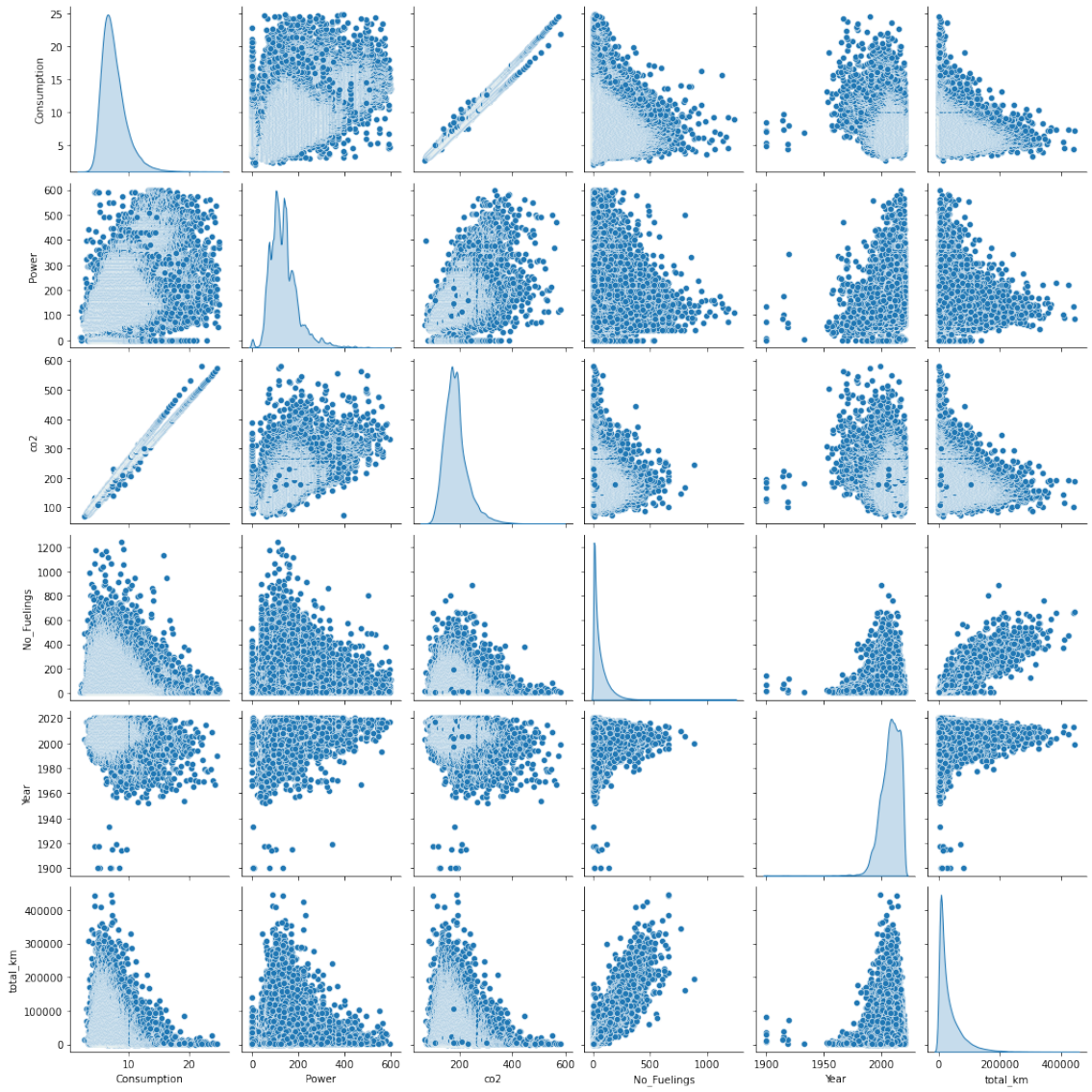

Figure 1: Pair-plot of the most relevant fields of the dataset. Frequency plots are shown on the diagonal.

4 Metadata

This section deals with the description of the metadata. Firstly, information on the web-scraping procedure is given. At a second step, the extracted variables are explained and characterised.

The information related to the web scraping procedure is:

Source URL	`http://www.spritmonitor.de`
Date of Extraction	21^{st} September, 2020
Type of Fuel	Only cars assigned to *Gasoline*, *Diesel* or *Diesel with AdBlue* fuels are considered.
Filtering	Some pre-filtering was carried out at the source. In detail, car with a consumption lower than $2L/100km$ were not considered, as well as cars with less than 5 tank re-fuelings. This in order to avoid outliers and have statistically-relevant information regarding the cars.

As stated above, the full car information was extracted for only 10% of the total extracted cars. The data available only for 10% of the total set is marked with $^{(*)}$ in the list below.

4.1 Variables Description

The variables extracted from the website are listed in this section. Some variables are redundant and used only for subsequent data extraction and validation. Nevertheless, they are reported for the sake of completeness. They are highlighted in italic and will not be used throughout the data analysis phase.

The data contains 432272 entries, grouped in 20 different columns.
Memory usage: 72.6+ MB.

ID	*Units:* [−]. Identification number used during data extraction;
Brand	*String.* Car maker;
Description	*String.* Web description provided by the users;
Vehicle_URL	*String.* URL to the specific car page;
Power	*Units:* [*hp*]. Declared car power;
Fuel_Type	*String.* Type of fuel of the car;
No_Fuelings	*Units:* [−]. Number of refuelings of the car registered;

User	*String.* Username of the car owner;
User_URL	*String.* URL to the specific user page;
Brand1 $^{(*)}$	*String.* Car maker. Duplicate of *Brand*, used as cross-check during second-step data scraping;
Model$^{(*)}$	*String.* Car model;
Year$^{(*)}$	*Units:* $[YYYY]$. Year of production.
Gearbox$^{(*)}$	*String.* Type of gearbox;
co2$^{(*)}$	*Units:* $[g/km]$. Car CO_2 production;
total_km$^{(*)}$	*Units:* $[km]$. Total amount of km of the car;
min_cons$^{(*)}$	*Units:* $[L/100km]$. Minimum consumption registered for a fueling.
max_cons$^{(*)}$	*Units:* $[L/100km]$. Maximum consumption registered for a fueling.
motorway$^{(*)}$	*Units:* $[-]$. Percentage of mileage on the motorway;
city$^{(*)}$	*Units:* $[-]$. Percentage of mileage in the city;
countryroads$^{(*)}$	*Units:* $[-]$. Percentage of mileage on the country roads;
Consumption	*Units:* $[L/100km]$. Real car consumption (average over fuelings).

5 Data Quality

Data quality is necessary to fulfil the needs of the current project in terms of targets, planning, and decision-making.

The metrics used for assessing data quality are the followings:

Accuracy In order to provide a statement on the accuracy and resolution of the data in possession, some basic assessment is done on the most important variable of the whole dataset, i.e. consumption. Consumption is measured in $[L/100km]$ and is obtained by a division between re-fueled quantity and kilometers travelled.

Uncertainty propagation [2] is computed as

$$\frac{\delta c}{c} = \sqrt{\left(\frac{\delta L}{L}\right)^2 + \left(\frac{\delta km}{km}\right)^2} \tag{1}$$

where c is the consumption, L the fuel quantity in liters, km is the distance travelled and δ denotes the error on the measurement. Assuming an average tank of 40 L and an average consumption of 7.49 $L/100km$, a car travels

534 km. The reading accuracy from the odometer is $\delta km = \pm\ 0.05km$ and from the fuel station is $\delta L = \pm\ 0.005L$. The resulting $\frac{\delta c}{c}$ is 0.02 %, which is lower than the resolution of the consumption records.

Completeness The dataset is composed of more than 400 thousand records. The least amount of records is obtained for the fields `min_cons` and `max_cons`. However, such amount of information is considered enough to be able to draw conclusions on the dataset.

Integrity Data integrity is ensured through data cleaning, as depicted in Figure 2. The aim of data cleaning is to make sure that there are no unintended data errors, i.e. integrity is guaranteed.

Timeliness Timeliness is one of the biggest challenges of a web-scraping project. The required data has to be available within the project start date. Due to time constraints, additional car details are extracted only from a subset of 10% of the total dataset. Anyway, such timeliness constraints do not affect data integrity, nor completeness.

The statements above support the conclusion that data quality is ensured for the scope defined in the current project.

6 Data Flow

The data flow for the current project is shown in Figure 2 on page 11. As explained previously, data is obtained through web scraping, performed in two steps. The first step involves the acquisition of cars list and general characteristics. More detailed information is retrieved by scraping the cars details page, whose address is extracted during the first step. All data is then collected into a `.csv` file, considered as the starting point for subsequent data analysis. Before being analysed, data is checked and cleaned, if needed. Analysis methods are defined in Section 2.3. Expected results of the analysis are plots and tables, essential in order to draw conclusions useful for the project.

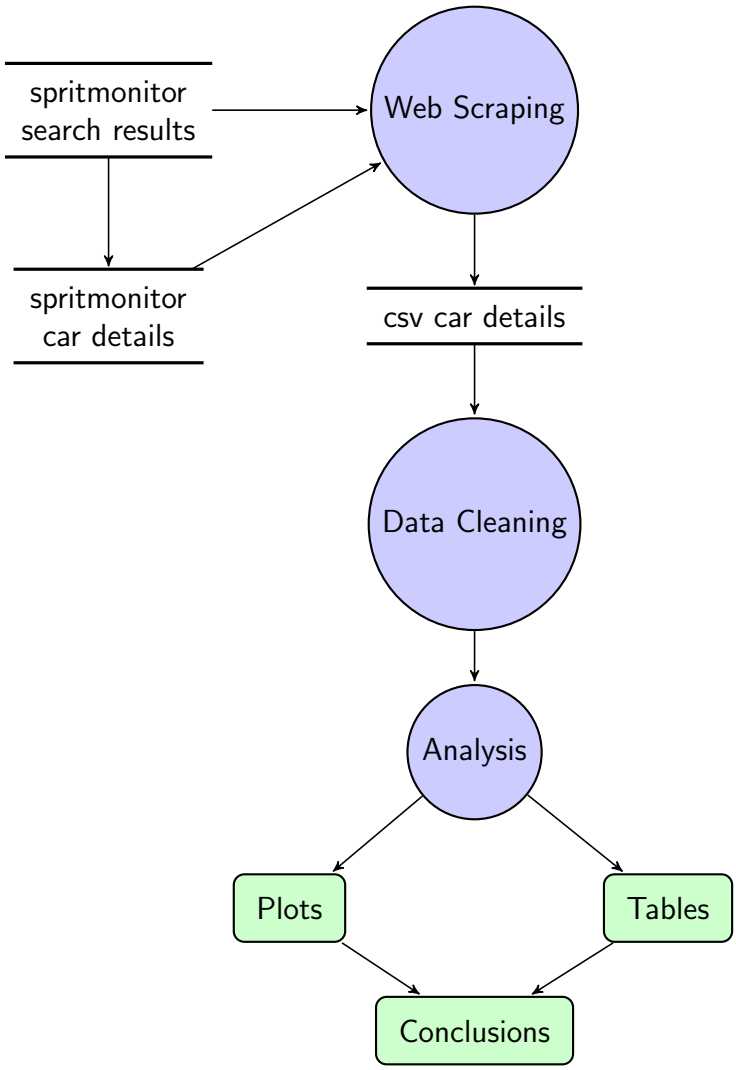

Figure 2: Data flow diagram for the current project

7 Data Model

7.1 Physical Data Model

The physical infrastructure needs are defined in this section.
The storage needed for the extracted data is estimated beforehand according to [3]. The expected number of entries is retrieved by the search page of spritmonitor, by selecting the fuel types of interest. The expected entries are 450'000.

In Table 2 below, details about the estimation procedure are given:

Field	Records Expected	Type	Unitary Size	Total Size
ID	450′000	Integer	4 B	1.80 MB
Brand	450′000	String (8 chars)	5 B	3.60 MB
Description	450′000	String (20 char)	10 B	9.00 MB
Vehicle_URL	450′000	String (20 char)	10 B	9.00 MB
Power	450′000	Integer	4 B	1.80 MB
Fuel_Type	450′000	String (7 char)	7 B	3.15 MB
No_Fuelings	450′000	Integer	4 B	1.80 MB
User_URL	450′000	String (25 char)	25 B	11.3 MB
User	450′000	String (10 char)	6 B	4.50 MB
Brand1	40′000	String (8 char)	5 B	200 kB
Model	40′000	String (8 char)	5 B	320 kB
Year	40′000	Integer	4 B	160 kB
Gearbox	40′000	String (12 char)	12 B	480 kB
co2	40′000	Integer	4 B	160 kB
total_km	40′000	Integer	4 B	160 kB
min_cons	40′000	Float	4 B	160 kB
max_cons	40′000	Float	4 B	160 kB
motorway	40′000	Float	4 B	160 kB
city	40′000	Float	4 B	160 kB
countryroads	40′000	Float	4 B	160 kB
Consumption	450′000	Float	4 B	1.8 MB
commas	9′000′000	Char	1 B	9.00 MB
Total				≈ 60 MB

Table 2: Estimation of storage required for the dataset

The estimation here reported does not differ too much from the final dataset

storage of 72.6 MB mentioned in Section 4.1 on page 8. The deviation could be explained by the presence of special characters or variation in records size.
However, the dataset dimensions do not represent an issue for the hardware in use. See Section 2.1 for more details.

7.2 Conceptual Data Model

The conceptual data model is shown in Figure 3, where the structure of the dataset is sketched.
The *Car Stats* table contains all the statistics related to a specific car. The relation between *Car Stats* and *Car* is a many-to-one relationship, i.e. there are many specific cars associated to the same car model. Similar reasoning is done for the relationship *Car Stats* and *User*, as the same user could own many cars. More detail regarding the entities is provided in the following section.

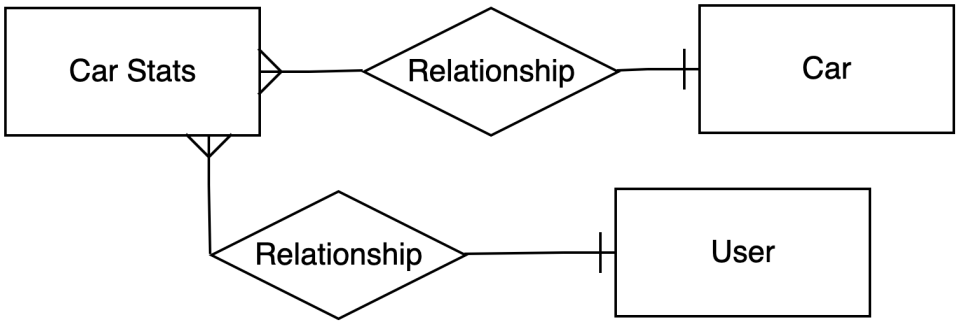

Figure 3: Conceptual Data Model for the dataset

7.3 Logical Data Model

The logical data model is shown in Figure 3 aims at showing the specific attributes contained in each entity.
The *Car Stats* table contains all the statistics related to a specific car. It is associated to a car model, whose emission, year of production, gearbox type and power are known. Each car unit is owned by a single user, who is identified by the *User_ID* attribute.

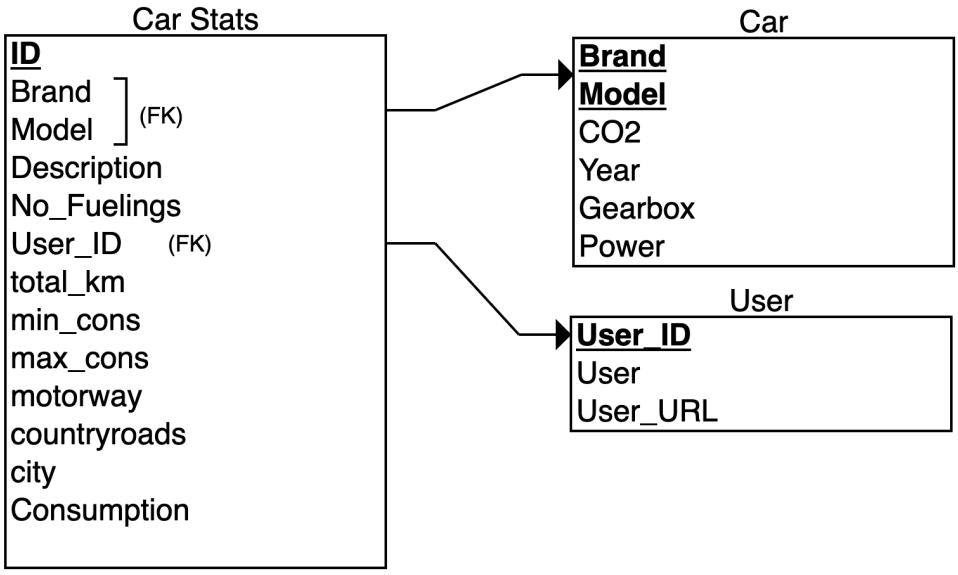

Figure 4: Logical Data Model for the dataset

8 Risks

Three main risks have been identified during the development of this data science project.

Human Resources The main risk when it comes to human resources is losing one member of the team. Since the current project is being developed by two people, the field of knowledge is spread among two people. The absence of a person would imply the risk of losing knowledge. This risk is mitigated by ensuring exhaustive documentation throughout the development process.

Data Protection As mentioned before, the data is obtained by means of web-scraping. The impossibility to access the webpage has been identified as one of the main risks of the project since it would mean that the data is not anymore at the source. In order to make the data accessible for the project, the scraped webpage has been stored in a .csv file. Despite the amount of cars stored in the file, the size of 70 MB is considered manageable. Thus, the data is stored locally. Moreover, backups on cloud and personal NAS are performed on a daily basis.

Version Control and Concurrent Development In order to ease the concurrent development of the project and provide a comfortable environment to the project owners, a repository on GitHub has been created. This reduces the risk of conflicts and loss of data during the development phase.

9 Preliminary Studies

The preliminary studies and answers to the objectives defined in Section 1 are dealt with in this chapter.

The full dataset is shown in Figure 5 with the aim of checking whether the data makes sense or not with a general overview. The biggest amount of data available belongs to cars from the 1990's till today. On one hand, older cars have low number of refuelings likely because they are owned as vintage cars and were registered on the website at a later stage. On the other hand, more recent cars did not travel a significant amount of kilometres yet. As expected, the cars with highest mileage have also the highest number of refuelings. The data is acceptable from the consistency point of view.

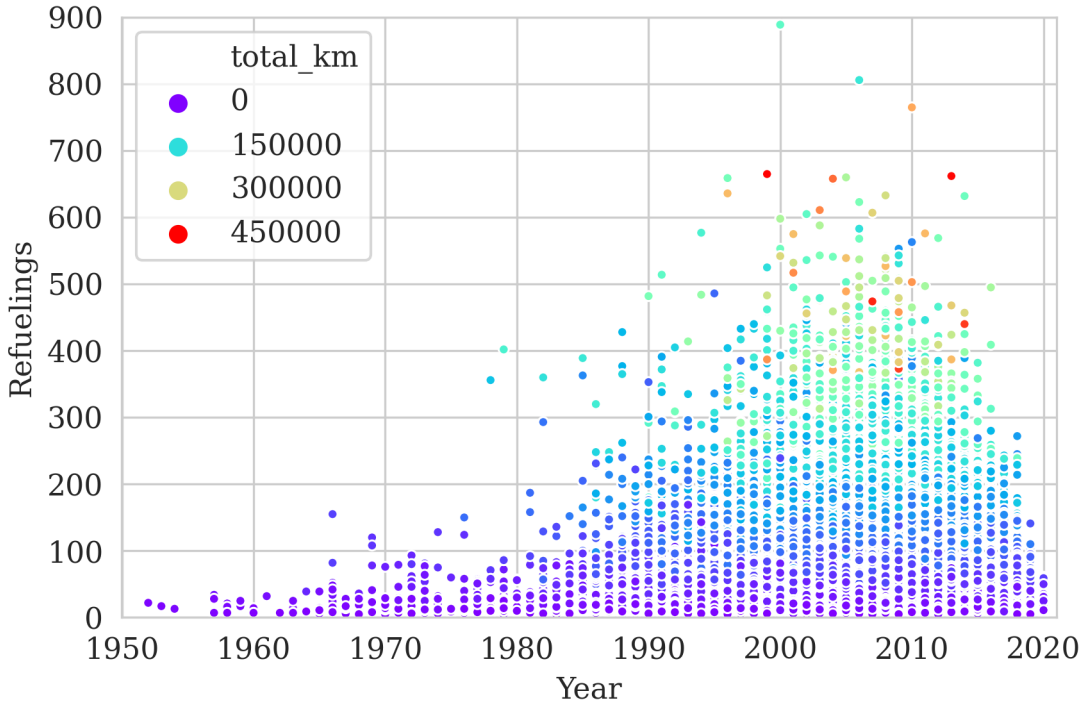

Figure 5: Pointcloud with all the available data.

It is also interesting to check the histogram of the fuel consumption shown in Figure 6, where an average of around $7.5L/100km$ is obtained. This is verified with the statistical description of the more interesting variables in Table 3. It is

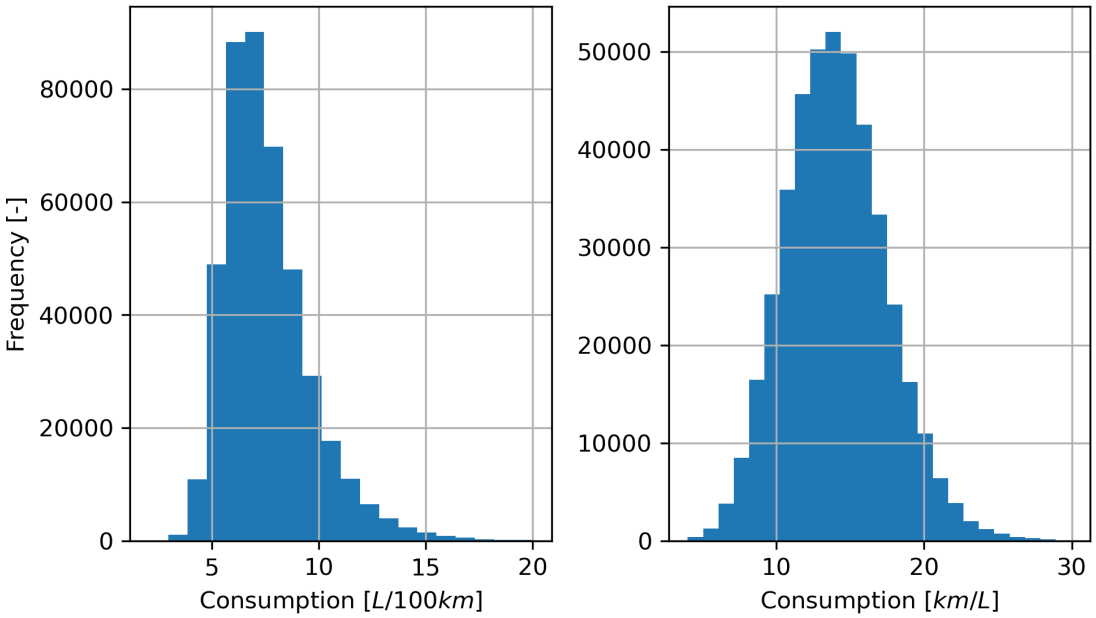

Figure 6: Fuel consumption histogram for all the available data.

also interesting to notice how the histogram presents a skeweness to the right. If the consumption units are instead converted to litres per kilometre $[L/km]$, the histogram plot resembles a normal distribution. This statement will be verified later in the chapter.

Table 3: Statistical description of the main variables

	Power	Consumption	Year	No_Fuelings
count	42,986.00	42,986.00	42,986.00	42,986.00
mean	137.48	7.69	2,007.56	62.71
std	63.60	2.03	8.15	68.91
min	0.00	2.66	1,900.00	5.00
25%	99.00	6.29	2,003.00	16.00
50%	129.00	7.42	2,009.00	38.00
75%	168.00	8.50	2,014.00	85.00
max	597.00	24.59	2,020.00	889.00

9.1 Sensitivity of Fuel Consumption

This section deals with the first target of the project, that is the study of the influence of car parameters on fuel consumption.

9.1.1 Fuel consumption over year and power range

During the process of buying a car, one parameter that is often taken into consideration is the fuel consumption, since it is one of the biggest variable costs that the owner will face during the lifetime of the vehicle.

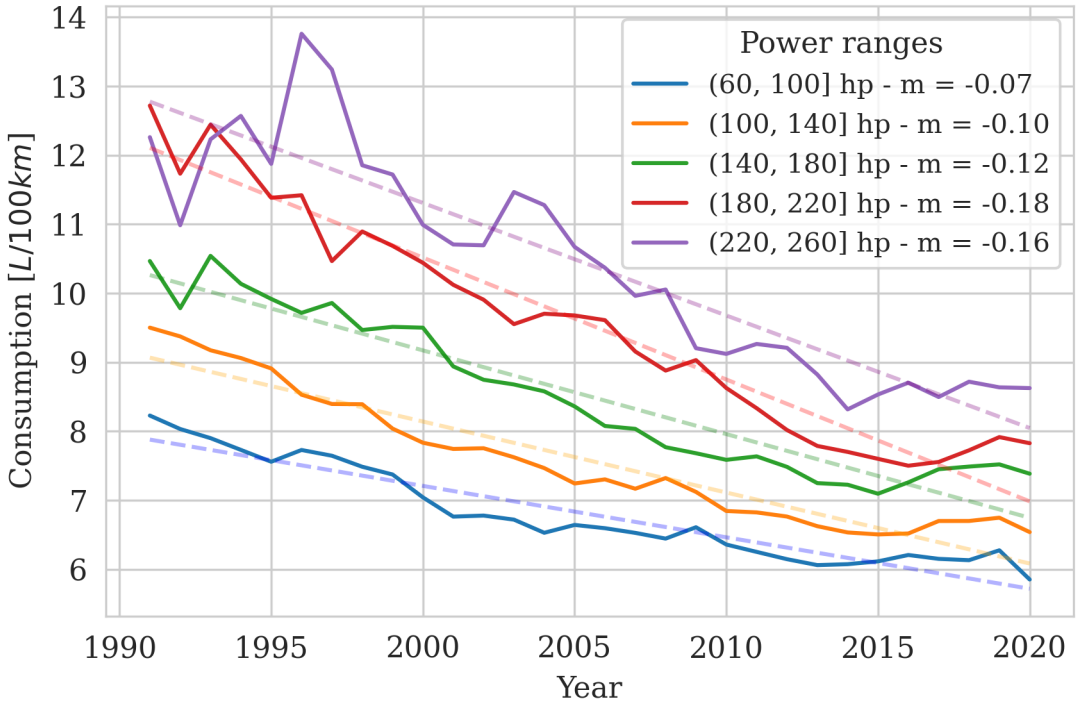

Figure 7: Fuel consumption and linear fit over the years grouped by power range.

Attending to Figure 7 one can clearly note that for all power ranges the fuel consumption has decreased over the years, from 1990 to today. The rates of reduction of fuel consumption are bigger for higher power ranges, most probably due to new technologies, that is used first in the higher end cars. It is also worth to note how from 2015 the fuel consumption reduction has stabilised. This could be due to several reasons:

- The internal combustion engine is approaching a technological limit.
- The more and more stringent emission requirements limit engine combustion efficiency.
- According to the European Environment Agency, *about 38% of new car registrations in Europe for 2018 were SUVs. Compared to other cars in the same segment, SUVs are typically heavier and have more powerful engines and larger frontal areas - all features that increase fuel consumption. The average mass of new cars increased by 30 kg from 2018 to 2019. The mass increase was observed for all vehicle segments (small, medium, large regular cars, and SUVs) and for both petrol and diesel cars* [4].

9.1.2 Fuel consumption variation with city usage

Fuel consumption is also influenced by the driving conditions. The most interesting check is to verify how much driving in the city affects the final fuel consumption.

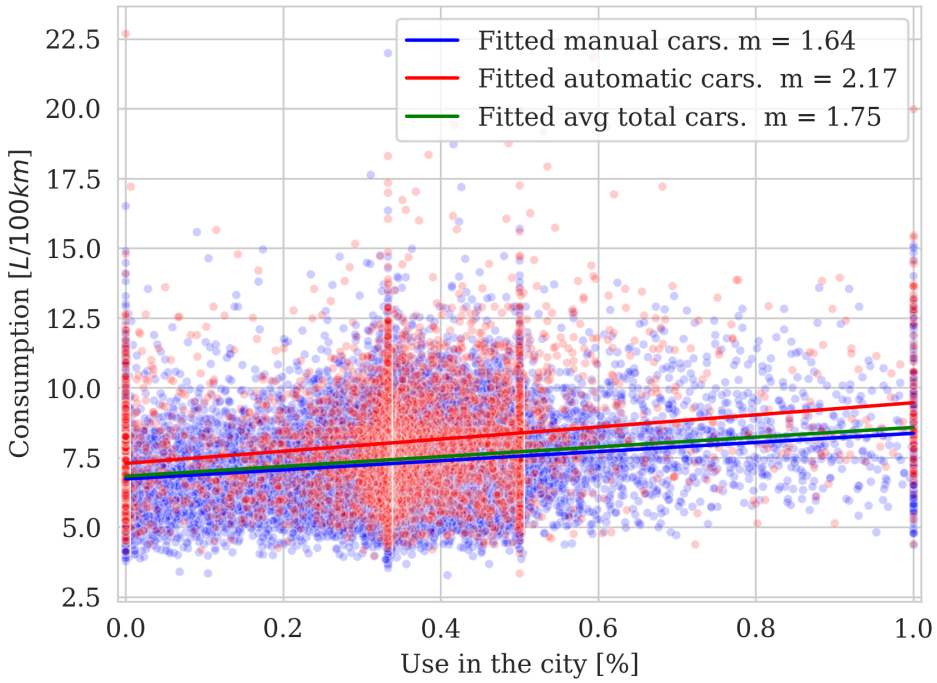

Figure 8: Fuel consumption depending on city mileage percentage.

Table 4: Statistical description of consumption by brand

	Consumption by brand
mean	7.63
std	1.91

In order to have more reliable data regarding current cities we have only computed those entries between years 1994 and 2020 and a power range $\mu_{power} \pm 1 \cdot \sigma_{power}$. A differentiation between manual and automatic gearbox was also taken into account in order to identify potentially-different behaviours of these two systems.

In Figure 8 the scatter plot and the related linear fit are shown for each type of gearbox. As expected, the fuel consumption increases with the mileage done in the city. Automatic cars seem to behave worse the more they are driven in the city. In average there is an increase of fuel consumption of 0.175 litres per each 10% increase on the mileage travelled in the city.

The assignment done by the website users to city mileage is significantly biased towards fractional values, that is 0, 1/3, 1/2, 1.

9.1.3 Brand Influence on fuel consumption

In Figure 9 fuel consumption boxplots for the 15 most common brands are shown. Table 4 lists the mean and the standard deviation of the consumption grouped by brand.

If a buyer would only focus on pure consumption numbers, this plot would be useful to make a decision. In that case, *Toyota* would be the preferred choice of consumption-aware drivers. However, as almost every car owner bases his choice to other aspects than pure consumption - i.e. increased power - a better picture of car efficiency is drawn by Figure 10, where the consumptions values have been divided by the power of each car.

With the power-normalised values of consumption, the previous brand selection of *Toyota* is now replaced by premium brands like *BMW*, *Audi* and *Volvo*, that seem to be the most efficient in relation to their engines power. Car weight is not taken into account at this stage, as not available from the dataset. Nevertheless, it could have a significant impact on the conclusions drawn here.

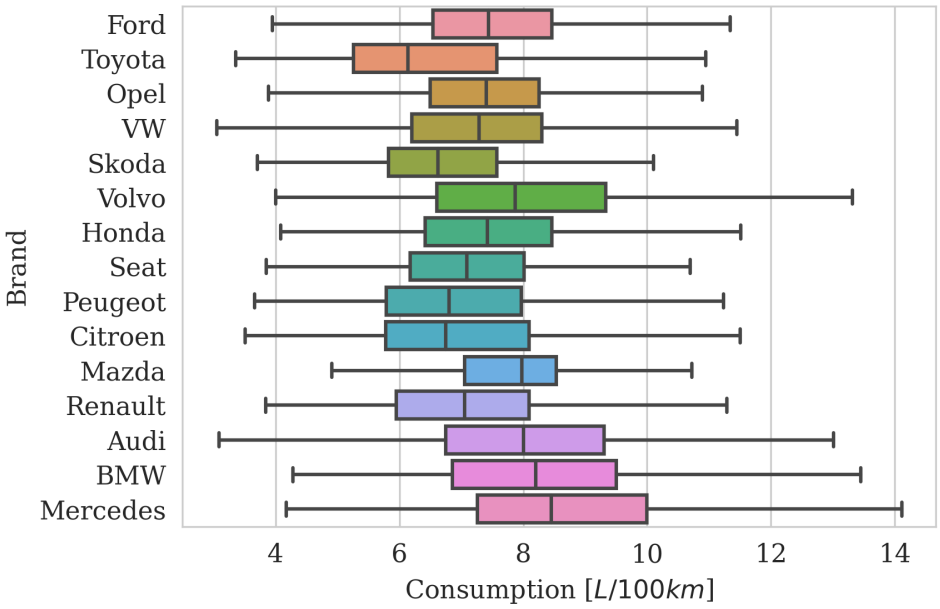

Figure 9: Fuel consumption boxplots by brand

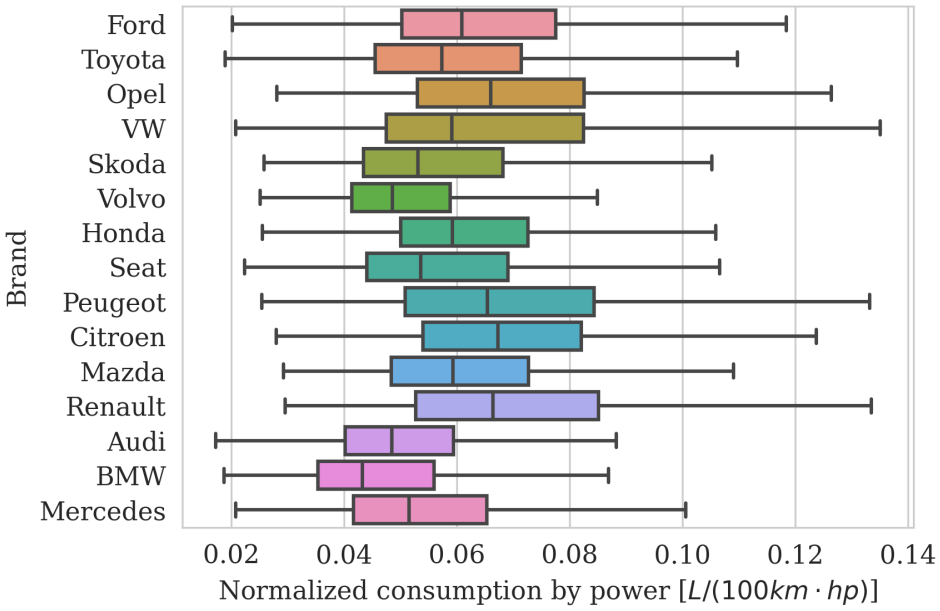

Figure 10: Fuel consumption normalised with power boxplots by brand

9.1.4 Fuel consumption by gearbox type

Another parameter that affects the fuel consumption on cars is the gearbox type. A histogram is shown in Figure 11 and the main statistical description is given in Table 5. The gearbox with lowest fuel consumption is the continuously variable one. This result is expected since the continuously variable gearbox is design to operate in the most efficient operating point of the engine. Direct shift and manual have a really similar behaviour since the operation concept is basically the same. Automatic gearbox has the worst performance, however, the mean consumption value of the automatic is probably affected by the low-efficiency gearboxes that were used at the beginning of their development.

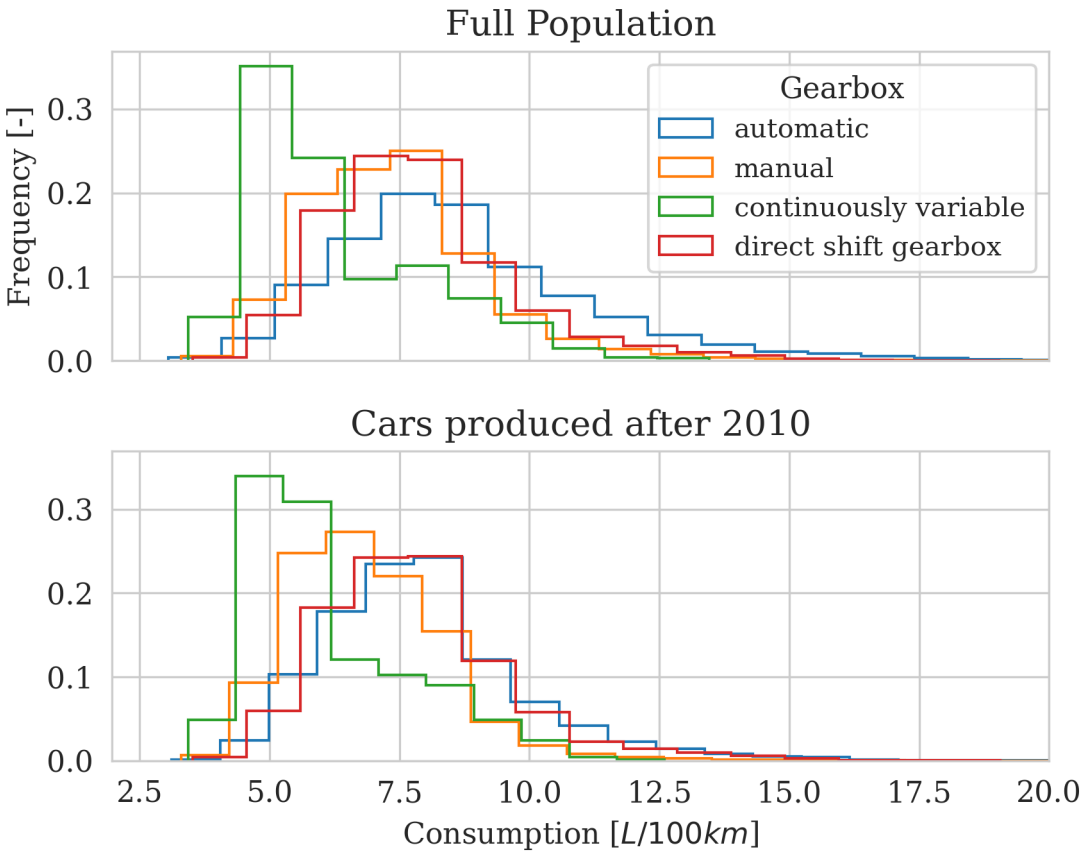

Figure 11: Fuel consumption by gearbox type histogram

Table 5: Statistical description for each type of gearbox in relation to fuel consumption

Gearbox	mean	std	min	25%	50%	75%	max
automatic	8.68	2.53	3.05	7.03	8.26	9.88	24.59
continuously variable	6.31	1.71	3.43	5.04	5.69	7.52	13.46
direct shift gearbox	7.90	1.82	3.52	6.63	7.67	8.71	19.06
manual	7.42	1.75	3.30	6.19	7.28	8.27	22.39

9.2 CO2 generation by fuel type

This section deals with the first target of the project, that is how the CO_2 production scales depending on the type of fuel used.

As shown in Figure 12, CO_2 emission over fuel consumption can be well represented by a linear trend, different for each fuel type. The carbon dioxide emissions are the result of a the chemical reaction of fuel combustion. As reported by the Minister of the Natural Resources of Canada [5], 2.29 kg and 2.66 kg of CO_2 are produced from the combustion of $1L$ of gasoline and diesel respectively. Such linear trends fit well with available data, as shown in Figure 12.

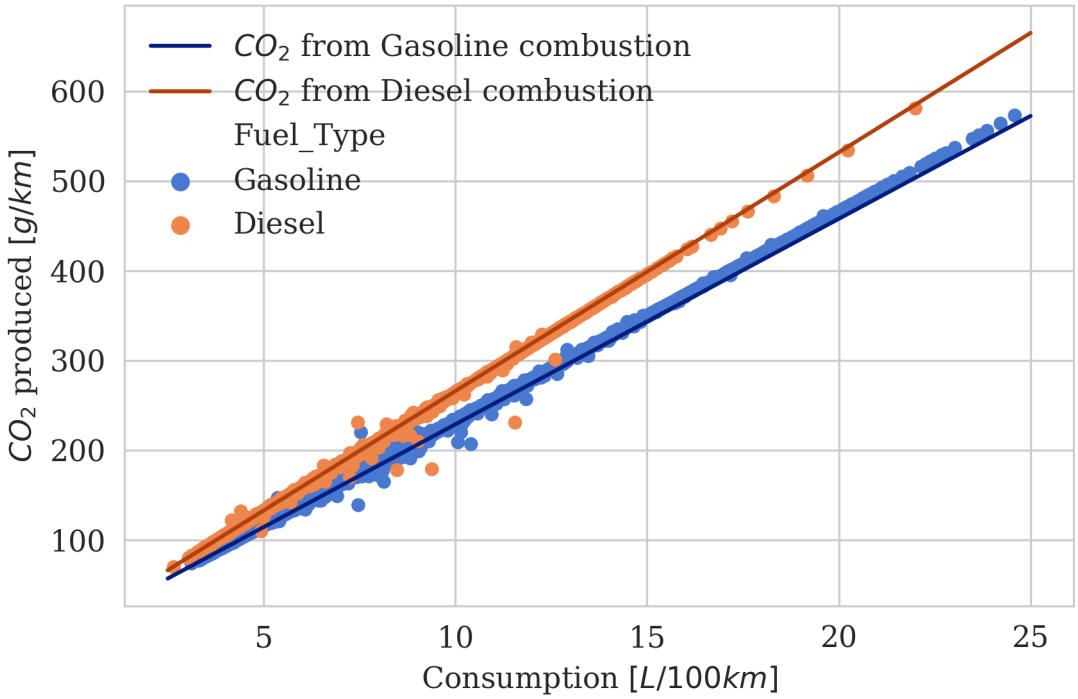

Figure 12: CO2 generation data.

9.3 Fuel Analysis

The effect of fuel type on consumption is assessed in this section. In particular, the frequency plots of gasoline, diesel, and diesel with AdBlue fuels are compared to the normal distribution by means of q-q plots. At a second step, hypothesis testing is carried out on the consumption data in order to draw statistical conclusions on the consumption based on fuel types.

Figure 13 and Figure 14 shows the different histograms for the different fuels and the related q-q plots respectively. From a visual assessment, the assumption that the samples belong to a normal distribution is confirmed.

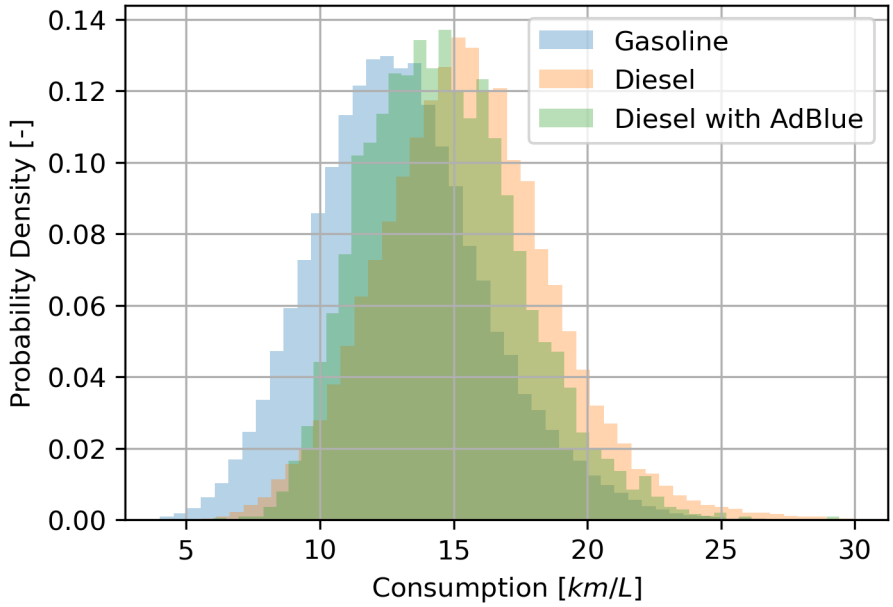

Figure 13: Histogram of consumption grouped by type of fuel.

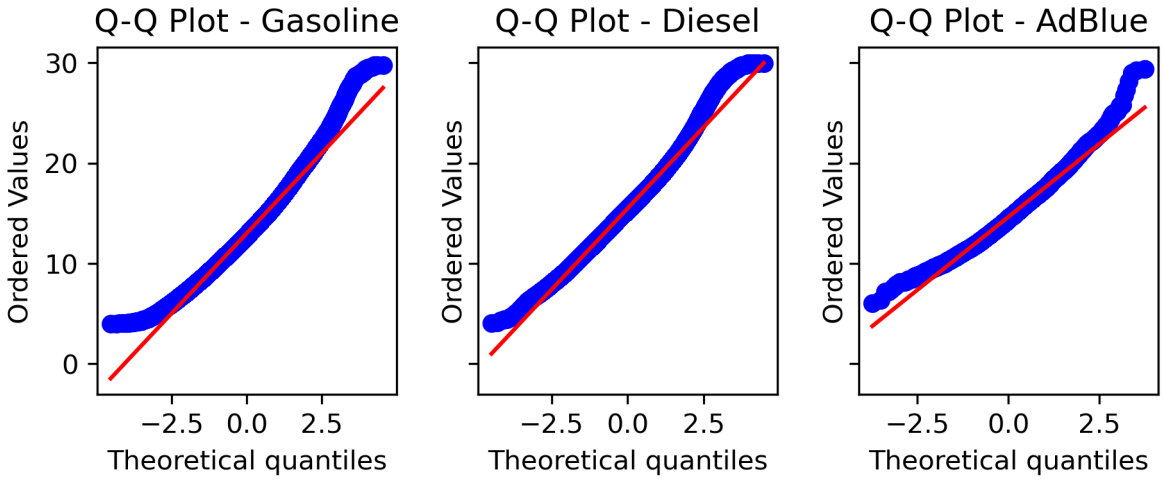

Figure 14: Q-Q plots of consumption grouped by type of fuel.

The unpaired t-test with Welsch's correction is performed in order to test the hypothesis that diesel cars consume less fuel than diesel AdBlue cars, which consume

less than gasoline vehicles. Both tests yielded zero p-values, thus the null hypotheses of equal averages are rejected.

9.4 Prediction of Fuel Consumption

This sections deals with the preliminary considerations in light of the development of a car consumption prediction algorithm.
A linear regression method was applied on the reduced dataset containing values for all the variables included in the dataset. The linear regression method was trained on 80% of the dataset and tested on the remaining 20%. The method is successful if the true consumption value and the predicted one are the same, except for some acceptable deviation.

Figure 15 contains the result of the current analysis, for which a mean squared error (mse) of 1.17 $L/100km$ is reported for the test subset. Such value is considered not sufficient for the acceptance of the linear regression method as a prediction algorithm for fuel consumption.
More complex machine learning methods will be investigated during the next steps of the project.

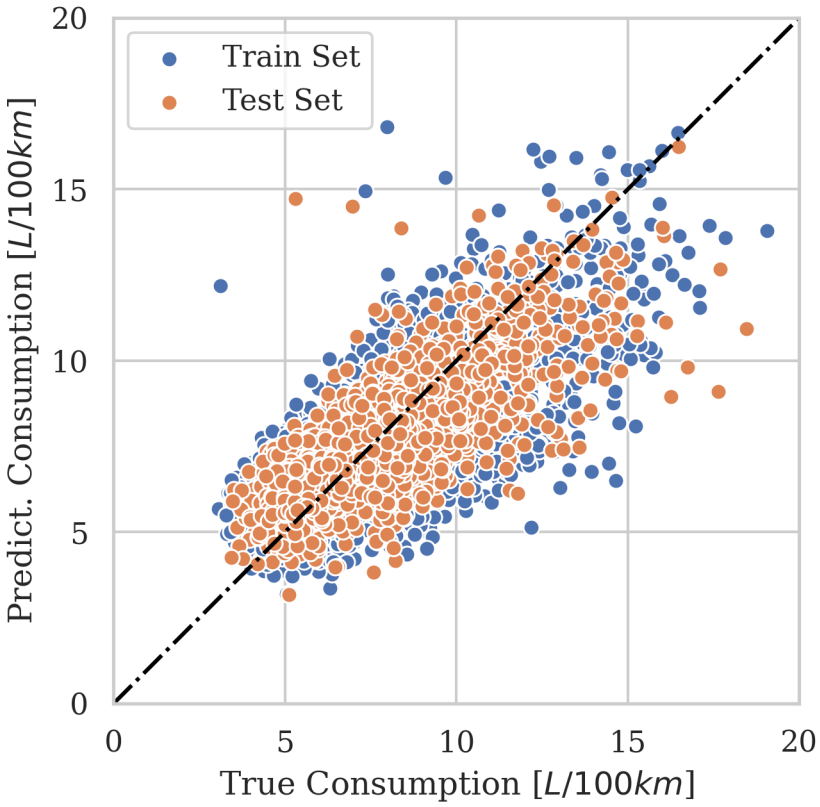

Figure 15: Fuel consumption estimation through linear regression.

10 Conclusions

The current conceptual design report defined the methods and criteria adopted for fuel car consumption data science project.
Analysis methods, data, and metadata were described in the first part. The second portion of the project was dedicated to the description of data quality, data flow, and data model, with a focus on the specific risks associated to the project.
Preliminary studies were carried out in order to reach the targets set for the analyses:

- Fuel consumption has been decreasing since the 90's, being more pronounced for the higher power ranges. However, the reduction has stabilised over time. Around 2015 the fuel consumption reduction has not been linear. The reasons are to be found on the closer technological limits of ICE engines, more stringent requirements in terms of emissions, and increasing weight of the vehicles.

- A fuel consumption increment of 0.175 litres for each 10% of mileage done in the city was highlighted. Moreover, city driving tends to affect more the automatic cars than the manual cars. There is a tendency of the users to keep the same proportion of the mileage done in the city fixed over the car refuelings.

- A study on power-specific fuel consumption shows that german brands BMW and Audi produce the most efficient engines for their cars. Weight factor shall be taken into account for future assessments. The standard deviation is around 1/3 of the mean, showing how spread the fuel consumption is on the cars population.

- Among the different gearboxes, the continuously variable one is the most efficient, having a mean fuel consumption around 1.2 litres lower with respect to the others. Automatic cars are characterised by higher consumption, due to the presence of old automatic gearbox technology in the dataset.

- The amount of CO_2 emissions generated by each car is probably computed proportionally from the consumption values by the website algorithms. Consistency of such calculation with the values obtained from the chemical balance of combustion is proven.

- Consumption grouped by fuel type is distributed normally. Hypothesis testing confirms that diesel cars consume less than gasoline cars. Diesel with AdBlue lies between the two mentioned types of fuel.

- Prediction of fuel consumption by means of linear regression methods is not satisfying. More complex predicting methods need to be adopted for such scope.

References

[1] Scikit-learn: Machine Learning in Python.

[2] Harvard University, *A Summary of Error Propagation*, Fall 2007 (`http://ipl.physics.harvard.edu/wp-uploads/2013/03/PS3_Error_Propagation_sp13.pdf`).

[3] Brown B., *Data Structures And Number Systems*, 1984-1999 (`http://www6.uniovi.es/datas/data1.html`).

[4] European Environment Agency, *Average CO2 emissions from new cars and new vans increased again in 2019*, (`https://www.eea.europa.eu/highlights/average-co2-emissions-from-new-cars-vans-2019`).

[5] Minister of Natural Resources Canada, *Learn the facts: Fuel consumption and* CO_2, (`https://www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/oee/pdf/transportation/fuel-efficient-technologies/autosmart_factsheet_6_e.pdf`).