Gas Stations Study

Alireza Molla Ali Hosseini, Francesca Zen

May 2022

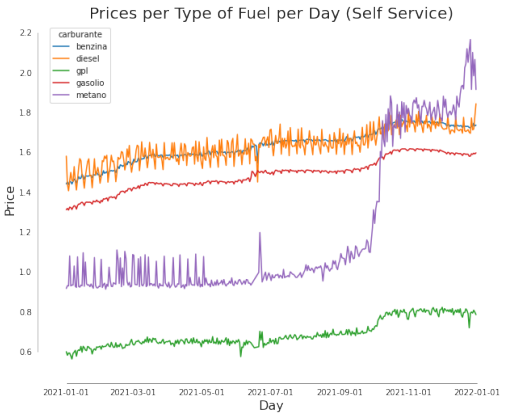
Abstract

What are the main results obtained from this work, specifying some particularities.

1 Introduction

Description of the context in which we worked, the problem Deloitte proposed and which aspects resulted to be more interesting. Specify the steps in which the analysis proceeds and what was the general idea that has been a guide for the analysis.

The goal of this project is to structure a business strategy that could be helpful to an businessman who wants to open his own gas station in Italy. Our work focuses on those factors that may contribute at the definition of the price of the fuels, looking for relationships with geographical locations, type of roads in which a gas station is placed, Brent price, and other macroeconomic elements in 2021.
From the day-to-day experience we know that the prices of the fuels are not constant and, what's more, are increasing during the years. As a matter of fact, in 2021, the prices of the fuels increment of 0.24 cents on average during the year. In the beside figure we show the trend of the prices in the mentioned period. Our objective is thus to find out a way for a good prediction on the prices, and to be competitive in the market.



We began our study from a basic analysis on the fuels' prices, distinguishing between

- self-service and served modelities,
- geographic position of the gas stations,
- type of road,
- type of fuel and the main differences among them, and
- relation with the price of the Brent.

As expected, we noticed that self-service fuels' prices are lower than the served ones, that "Autostradale" roads are more costly and that there are different distributions of the fuels depending on the previously points. In particular, this last element will help to understand which types of fuel to sell in the new gas station, as we will see later on.
Add somewhere the summed up info (general idea) given from the features importance work.

After this overview of the data, we proceeded more in detail into our case study with the aim of locating the new gas station in a strategic place, i.e. a territory with high density of the population and low concentration of petrol stations. We chose a territory with these characteristics because we though that the higher the density of the population in a specific territory, the higher the probability that they will use a gas station in that place, and thus the more the profit of the owner of the gas station. Also we have to take into account the possible competitors already present, and the "type" of cities we are considering. For this reason we studied the density of the population w.r.t. the amount of flags already present in Italy, and excluded the metropolitan cities as possible candidates, because of the high density of public transports present.
Finally we will sum up the results obtained, highlighting the trend of the fuels' prices and correlated business strategy for starting the new activity.

2 Dataset

The initial information that we used to start our analysis comes from the site of the Ministry of Economic Development [1], that provides a dataset with national prices of some type of fuels (unleaded petrol, diesel fuel, gpl). The data carried information like the location of the gas stations, the corresponding princes of the fuels sold, the type of flag and the owner of the installation, and so on, and they are summed up in Table 1 and Table 2.

For what concern the dataset "Anagrafica Impianti Attivi" we found that the already existing

Information	**Description**
idImpianto	progressive numerical code assigned by the system for the identification of the installation
Gestore	the company name of the company that manages the store
Bandiera	the distributor's sign (may be a brand identified or indicated in a general way as "white flags")
Tipo Impianto	countersignifies the type of road on which the distributor is located, and it can be of three types: autostradale, strada statale, altro (that comprehend all the other types of road)
Nome Impianto	name indicated by the operator to identify his plant
Indirizzo	address, house number and postcode
Comune	name of the Municipality where the plant is located; it is the second sorting criterion of the data contained in the file
Provincia	name of the Province where the plant is located; is the first ordering criterion for the data contained in the file
Latitudine	coordinate corresponding to the latitude expressed in decimal degrees
Longitudine	coordinate corresponding to the longitude expressed in decimal degrees

Table 1: Dataset "Anagrafica Impianti Attivi".

petrol stations in Italy are 22339 divided in 262 different flags and 13205 owners. Also, by looking at the `Comune` variable corresponding to the missing values of the `Province` column we discovered that they all belong to the Province of Naples, so we added the missing information.

From the dataset of the prices, instead, we selected the information related to the year 2021 and standardize the name of the fuels, in order to avoid duplicates of the same type. Then, we grouped together the more specific types present in the dataset into the macro categories of fuel -benzina, diesel, metano, gpl, and gasolio.
We ended up with 37 typology of fuels for the self-service mode and 38 for the served one.

[1] https://www.mise.gov.it/index.php/it/open-data/elenco-dataset/2032336-carburanti-prezzi-praticati-e-anagrafica-degli-impianti

Information	Description
idImpianto	progressive numerical code assigned by the system for the identification of the installation
descCarburante	type of fuel the price refers to. In addition to standard fuels - gasoline, diesel, GPL, methane - special fuels are named according to the specifications of the flag or identified with the name of the type of fuel followed by the indication "special"
prezzo	price - expressed in euros and three decimal places referring to the corresponding unit of measurement (liter in all cases except for methane which is kg) - in effect at the plant at 8 am on the day indicated in "Extraction date"
isSelf	binary variable 0, 1 indicating the service mode (1 in the case of self-service; 0 in the case of served) of the fuel entered in the record
dtComu	indicates the date (in dd/mm/yyyy format) and time (in hh:mm format) in which the extracted price was communicated by the operator

Table 2: Dataset "prezzi alle 8".

In the end we merged the two datasets, based on the `idImpianto` column.

To provide a complete overview of the analysis, we integrate the data about the daily fuel prices with the information related with the density of the population per province [2], the amount of cars present in Italy with the relative type of fuel[3] and the trend of the price of Brent in 2021[4].
The former of these new datasets carried the amount of people that populate each province; the second collects the information about the amount of cars and corresponding type of fuel per province (and it is summed up in Table 3); and the last one propose the price of Brent, expressed in dollars per gallon, over the year 2021.

Information	Description
Anno	the year of which the information is related
Provincia	name of the Province in which the record has been made
AL	type of fuel: other
BE	type of fuel: gasoline - benzina
BG	gasoline and liquid gas - benzina e gas liquido
BM	gasoline and methane - benzina e metano
EL	electric
GA	gasolio
GG	gasolio and gas
IB	hybrid and gasoline - ibrido e benzina
IG	hybrid and gasolio
ME	methane
ND	not declared

Table 3: Dataset about the amount of cars per type of fuel and province.

In order to have a visual representation of the future analysis we build a geo data frame, using geopandas[5], at both provinces and regional levels. We started by constructing the provinces shape file: we imported all the provinces shape files per region, and then we merged them for ending up with the provinces of Italy all together. Unfortunately, there are some missing information about

[2] http://dati.istat.it/Index.aspx?QueryId=18460
[3] https://opv.aci.it/WEBDMCircolante/
[4] https://mercati.ilsole24ore.com/materie-prime/commodities/petrolio/BRNST.IPE
[5] https://geopandas.org/en/stable/

the provinces, and for this reason there are some lack of colors in the maps.
The information about the provinces [6] and regions[7] shape files come from github.

3 Method

The programming language adopted for the development of the project is python, in particular we used the pandas library[8] for working fluently with the datasets, and pycaret[9] and seaborn[10] for visualizing, respectively, the results about the feature importance and the statistics obtained. We worked on google Colaboratory[11] -also named Colab-, an online platform in which it is possible to write code and run notebooks. Since our initial data were uploaded on google Drive, we used the os library[12] for reading those data from the corresponding folder. This methodology has a double advantage: the operations of unzipping the file and extract them are faster and we didn't have to upload our data each time we want to perform the analysis.
The datasets related to the density of the population, to the cars and their fuels, and to the price of Brent over 2021, were added manually for the local computer.
One point that we would like to mention is that our analysis had been made without the use of accelerators like GPUs, in particular our results came from a pure statistical evaluation, without the introduction of models.
The project is divided into five different notebooks, because of the heavy material produced by the pycaret library. They are divided in the following sections:

- Pycaret Analysis, in which there are analysis made distinguishing between self-service and served mode,
- Visualization, in which are plotted the results again distinguishing between self-service and served mode,
- Statistical analysis, in which we studied in detail the intrinsic characteristics of the price of the fuels and their relationships with other macroeconomic factors.

3.1 Pycaret

Pycaret is an open library in python that can help to work on the statistical analysis of the data. It does have many different features but the most important feature that it has for our analysis is that we can pass the data then it made different models that are represented famously in statistical field such as linear regression, gradient boosting, and random forest. Then it calculate the loss based on various metrics such as mean absolute error, mean squared error, and R2. Therefore in this way we were able to compare different models and find the best one for our analysis which was light gradient boosting(the objective metric for our comparison was MSE). The advantage of this library is that we can also get the most important features that are affecting the target(Prices in our analysis).

3.2 Light Gradient Boosting Model

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.

[6] https://github.com/sramazzina/italian-maps-shapefiles/tree/master/regions-with-provinces
[7] https://github.com/sramazzina/italian-maps-shapefiles/tree/master/italy-with-regions
[8] https://pandas.pydata.org/
[9] https://pycaret.org/
[10] https://seaborn.pydata.org/
[11] https://colab.research.google.com
[12] https://docs.python.org/3/library/os.html

- Lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.

4 Analysis of the Intrinsic Characteristics of Fuel Prices

Describe the preprocessing techniques adopted, motivating your choices for the work; introduce more in detail the path that guided the analysis, including a description of the models used to get to the final solution and state what were the regions why you prefer it to others. Describe the choice and the metric based on the price.
After the merging of the two main datasets presented in Section 2, we went more in detail analysing the types of fuel present and we discovered that there were duplicates of the same type with slightly different names. For this reason we first put the same format for every element of the `descCarburante` column and then we group them into five macro categories of fuel namely Diesel, Benzina, Metano, GPL, and Gasolio, summarized in Table 4.

Category of Fuel	**Subcategory of Fuel**
Diesel	blu diesel alpino, blue diesel, diesel e+10, diesel shell v power, dieselmax, e-diesel, excellium diesel, gp diesel, hi-q diesel, hiq perform+, s-diesel, supreme diesel, v-power, v-power diesel
Benzina	benzina, benzina 100 ottani, benzina energy 98 ottani, benzina plus 98, benzina shell v power, benzina speciale, benzina wr 100, blue super, f101, r100
Metano	gnl, l-gnl, metano
Gpl	gpl
Ssp98	ssp98
Gasolio	gasolio, gasolio alpino, gasolio artico, gasolio ecoplus, gasolio energy d, gasolio gelo, gasolio oro diesel, gasolio premium, gasolio speciale

Table 4: Grouping of the specific typologies of fuels into the corresponding macro category.

Doing so made the results of the analysis more readable and easier to work with.
After this we began our study on the intrinsic characteristics of the price of fuels. We started by looking more in detail on the differences between self-service and served modalities, searching for features that characterize the prices of the fuels in each of the two cases and that could explain the lack of similarity of the two modalities. Then we studied the different type of road in which the gas stations are placed and finally we analyzed the relationship of the fuels with the Brent.

4.1 Self-Service and Served Mode analysis

For this part of the analysis we divided the data into two major categories which are self-service and served modalities. As expected, the prices of the served modality are higher that those of the self-service mode, since in the former the price comprehend also the payment of the person that has to fill the fuel into the means of transport. From the self-service prices there is a rise of about 15 cent, and we can see in Figure 1 a visual representation of the bull.

Our main goal was to find the most important factors that influence the price of the fuels, in each of the two modalities mentioned previously.
From a first overview of the data, which comprehend all the columns, we discovered that the most important factors were `Latitude`, `Longitude`, `Flags` and `Provinces`, but since latitude and longitude are correlated with the provinces, we focused our analysis only on this last and on the

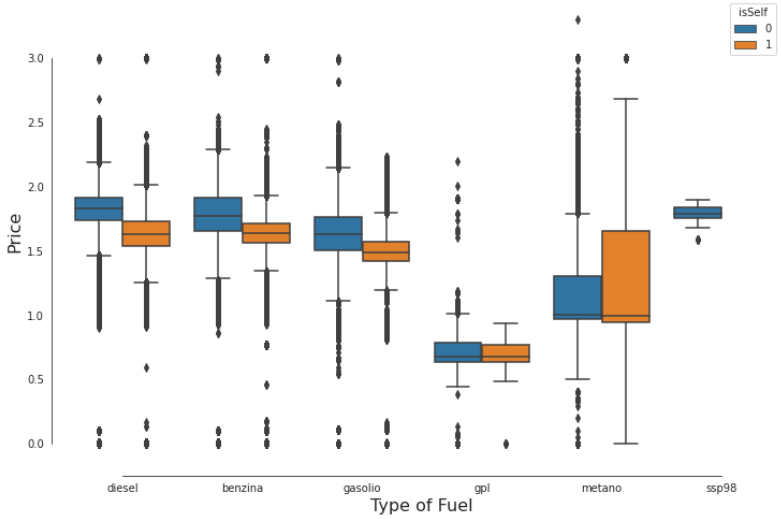


Figure 1: Plot of the distribution of the price of the fuels based on the different modalities of fuel delivery. 0 stand for served mode while 1 stands for self-service mode.

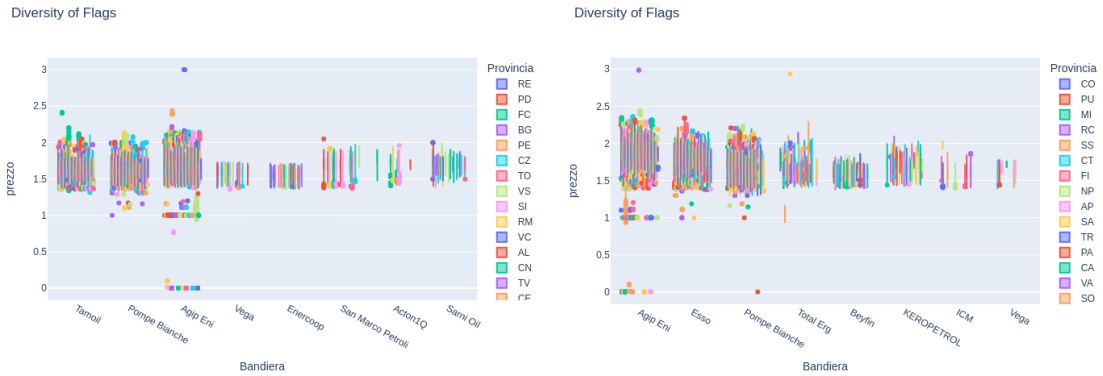


Figure 2: Diversity of Flags in Benzina(left picture is related to self data and right one is related to not Self data

flags.
In next step of the work we found the most important Provinces and Flags per category of fuel. The mutual characteristic of the important flags in each category was the diversity of them which means that those flags that are more branches in different provinces have more influence on the price. Similarly, those provinces that have more different kinds of flag on it have major impact on the price.

4.2 Metrics: Important factors

In order to check how important are those Flags and Provinces, we took all data related to them and tried to build a model for the prices with them and then tune whole data to the model and extract the related Error (MSE) . The results are provided in Table 5 and 6.

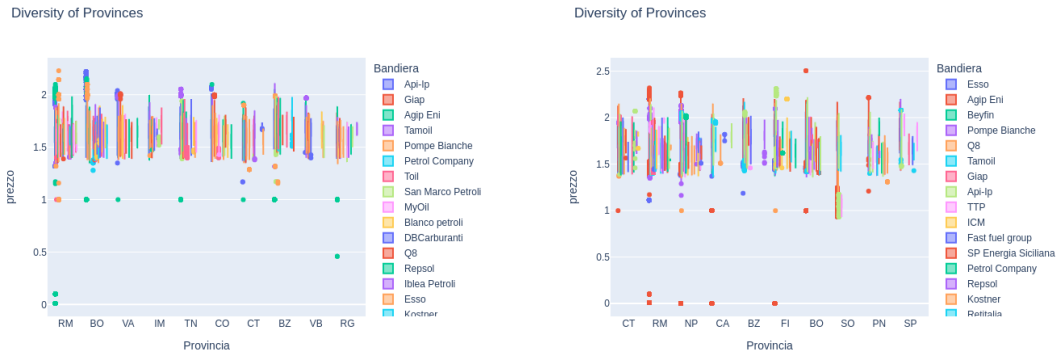


Figure 3: Diversity of Provinces in Benzina(left picture is related to self data and right one is related to not Self data

Fuel	**MAE**	**MSE**	**RMSE**	**R2**	**RMSLE**
Diesel	0.0911	0.01151	0.123	0.4814	0.0512
Benzina	0.0857	0.0127	0.1127	0.062	0.0457
Metano	0.3583	0.2007	0.448	0.018	0.1891
GPL	0.2174	0.0838	0.2895	-5.7158	0.2018
Gasolio	0.1064	0.0203	0.1424	-0.3545	0.061

Table 5: Results of the Models of Important factors(self data)

4.3 Latitude and Longitude Analysis

As previously mention, for latitude and longitude which are highly important factors in prices, we connect them to Provinces to have a better understanding of the locations. Therefore we needed to analyze the data only based on Provinces with the same categories as before. After the analysis we took those important provinces and plot data with their corresponding latitude and longitude. We noticed that the most important provinces from geographical point of view are those in north west, middle, south, and south east.(4)

4.4 Type of Road

One topic in which we focus our attention on was the difference that we may encounter when pay for fuel in gas stations situated in different types of road. For this reason we first look at the different addresses of the gas stations and discovered that there are three main type of roads: autostrade, strade statali and altro (in which are summed up all the remaining types).
In addition, in correspondence of the `altro` road we found out that there were some addresses with the "s.s." format, which normally stands for "strada statale", and thus we labeled those samples with the correct `Tipo Impianto` variable name.

Fuel	**MAE**	**MSE**	**RMSE**	**R2**	**RMSLE**
Diesel	0.0999	0.0201	0.1419	0.3953	0.0612
Benzina	0.1147	0.0222	0.1492	0.2621	0.0572
Metano	0.304	0.1451	0.381	0.0214	0.1592
GPL	0.0724	0.0075	0.0867	0.096	0.0504
Gasolio	0.1242	0.0249	0.1577	0.1832	0.0623

Table 6: Results of the Models of Important factors(not self data)

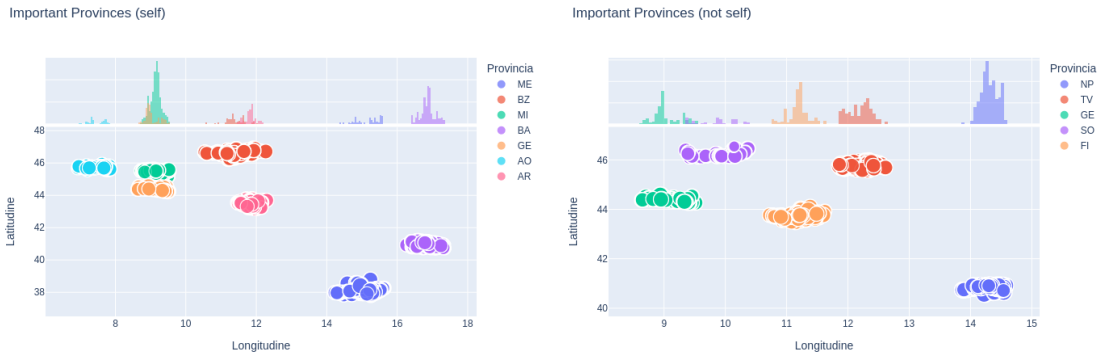


Figure 4: Location of Important Provinces in Gasolio

In Table 7 we reported the number of installation per type of road with corresponding statistics about the price of the fuels.

Tipo Impianto	**Count**	**Mean(€)**	**Std(€)**	**Min(€)**	**Max(€)**
Altro	17,926	1.57	0.143	0.001	2.999
Autostradale	441	1.691	0.171	0.001	2.324
Strada Statale	2,962	1.57	0.149	0.001	2.999

Table 7: Summary statistics about the different type of roads in which are present gas stations, considering the self-service modality.

From Figure 5, instead, we can see that for the self-service modality, on average, the prices in the autostrade are higher than the ones in the other two types and that the methane is not sell in those roads. This last one, in particular, has a wider range of prices, and this means more freedom on the choice of the price by the owner of the gas station and/or the corresponding flag.

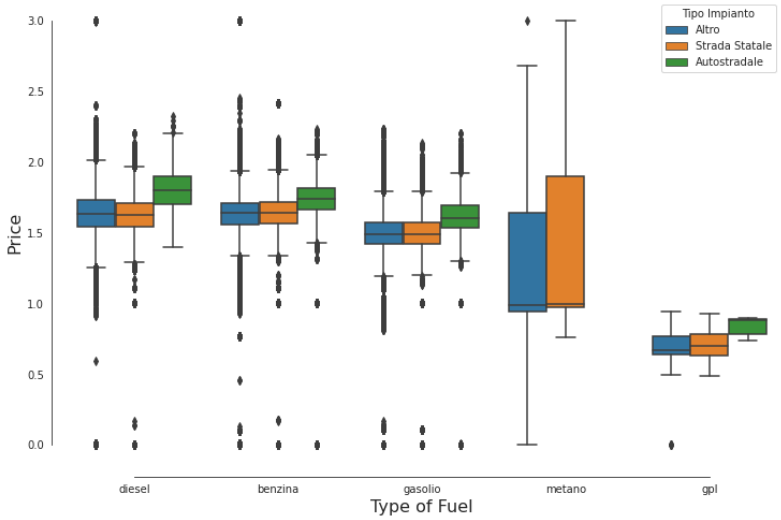


Figure 5: Plot of the distribution of the price of the fuels based on the different type of roads. Zoom on the self-service modality.

Even thought the prices on autostrade are higher that those on the other two types, we can see

that in these last there are more gas stations that impose prices that are very different from the corresponding mean, deviating from the average and overcoming also the prices on the autostrade. The major examples here are the prices of benzina and diesel in the smaller roads, in which the flag Agip Eni impose a price of €2.999 against the mean of €1.633 for benzina and €1.642 for diesel. This can due to a more freedom on the decision of the prices or other particularities of the territory in which the installation is placed (autonomy of the Region and so on).
We perform the same analysis also in the case of served modality, noticing that there were not many differences: the only changes came on the methane that is served in autostrade, and on the prices that present a small increment.

4.5 Geographical Distribution of the Prices

In this section we show the distribution of the prices in the regions of Italy, taking into consideration the fuels diesel, benzina and gasolio. In order to do a fair comparison between them, we consider the self-service modality, where the increments in price due to the hand-work of filling the fuel into the means of transport are not present, and the can evaluate better the differences among the regions. In Figure 6 it can be seen the results of our work: the colors are given taking into

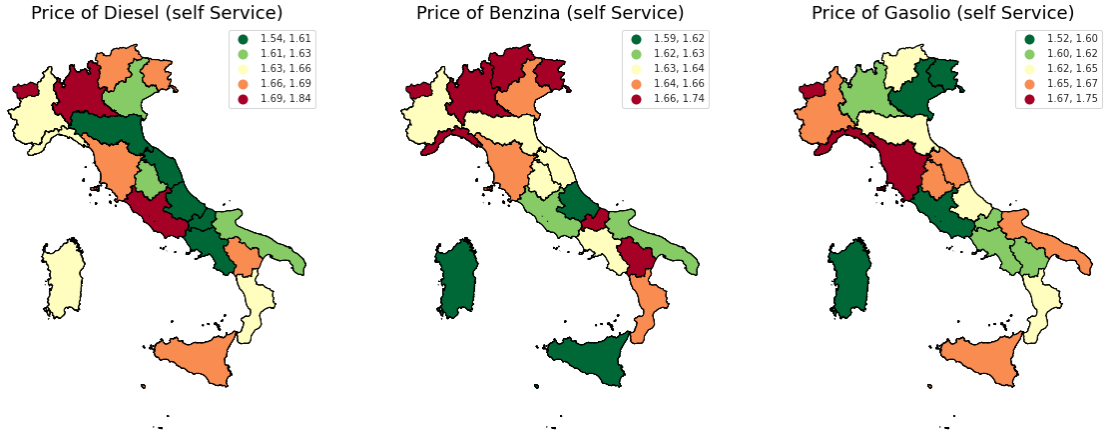


Figure 6: Geographical representation of the distribution of the prices of three type of fuel: diesel (left), benzina (center) and gasolio (left) for the self-service modality.

consideration the division with quantiles and from the legend we can see that diesel comprehend a wider range of prices with respect to the others.
From the image it can be seen that there is not a clear separation of prices bands through the North, Center and South of Italy, thought we made an analysis taking into consideration latitude and longitude, using pycaret library. We discover that ...

4.6 Relation with the Brent Price

The price of the fuel is composed by three factors:

- Platts, the value of fuels internationally,
- the gross margin of the oil industry,
- taxation (consisting of excise duties and VAT).

Platts is a platform where international fuel supply and demand intersect, determining the value of each petroleum product. On this item ends the gain of the initial part of the supply chain, i.e. that of the oil companies. In our case, instead of the the Platts we will use the price of Brent,

which identifies the oil extracted in the North Sea, which served as reference for the majority of the mondial prices. This represent the baseline price from which are added the profits of the other two factors. The component that most contribute in the price of the fuel is the taxation which, in the case of gasoline, corresponds of about the 67% of the final price.
In Figure 7 is represented the trend of the price of Brent during 2021, in dollars/bbl. As it can be

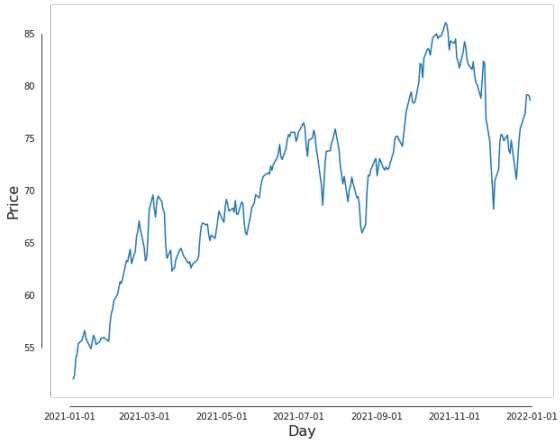


Figure 7: Plot of the distribution of the price of the Brent in 2021.

seen from the graph, the price of Brent increased over 2021, passing from 55.22 $/bbl$ in January to 74.66 $/bbl$ in December of the same year.
Unfortunately, the information about this type was not complete and we have covered only 259 days out of 365. We decided to go on with the analysis taking into consideration this fact and to work with the available data.
In order to do a fair comparison with the prices of the fuels in Italy, we transform the price per barrel (bbl) into price per liter (ℓ). A barrel of oil corresponds to approximately 42 US gallon (gal), and one gallon correspond to about $3,79\,\ell$:

$$1\,bbl \simeq 42\,US\,gal$$
$$1\,US\,gal \simeq 3,79\,\ell$$
$$1\,bbl \simeq 42\,US\,gal \simeq (42 \times 3,79)\,\ell \simeq 159,18\,\ell.$$

So we divided the price of Brent in bbl to get the corresponding price in liter.
Now, in order to see if there was a linear relationship between the Brent and the price of the fuels, we made a ratio between the two, ending up with the plot in Figure 8. From the graph we can see a decreasing trend of this new quantity, meaning that the price of Brent increase faster than that of the fuels.
We guessed that the price of fuels could undergo variations after few days of the variation of the price of Brent, so we take the same analysis that involves the ratio between the two quantities but this time taking into consideration the price of the fuels one weak after the change in the price of the Brent. Also in this new scenario the decreasing trend is maintained.

5 Business Strategy

Describe the business strategy adopted and that you recommend to use to a new player on the fuel sales, describing the more important characteristics of the major competitors, the factors that contribute the most at the definition of the price and identifying the principles at which to pay attention at the moment of the price strategy description (based on the adopted metrics).
Since now we have seen how the geographical location, the modality of fuel delivery and the type

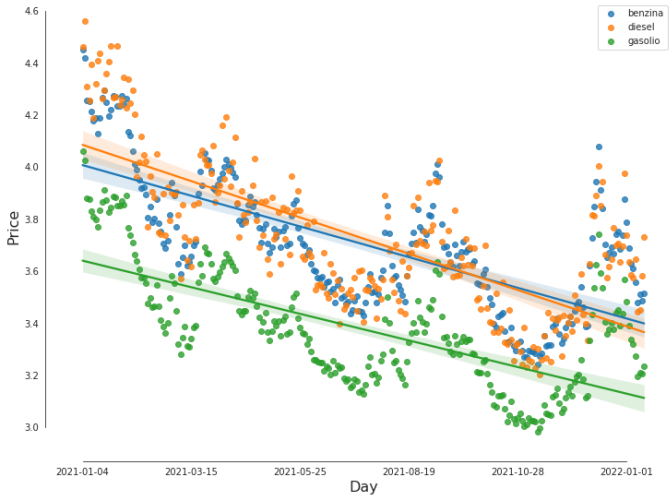


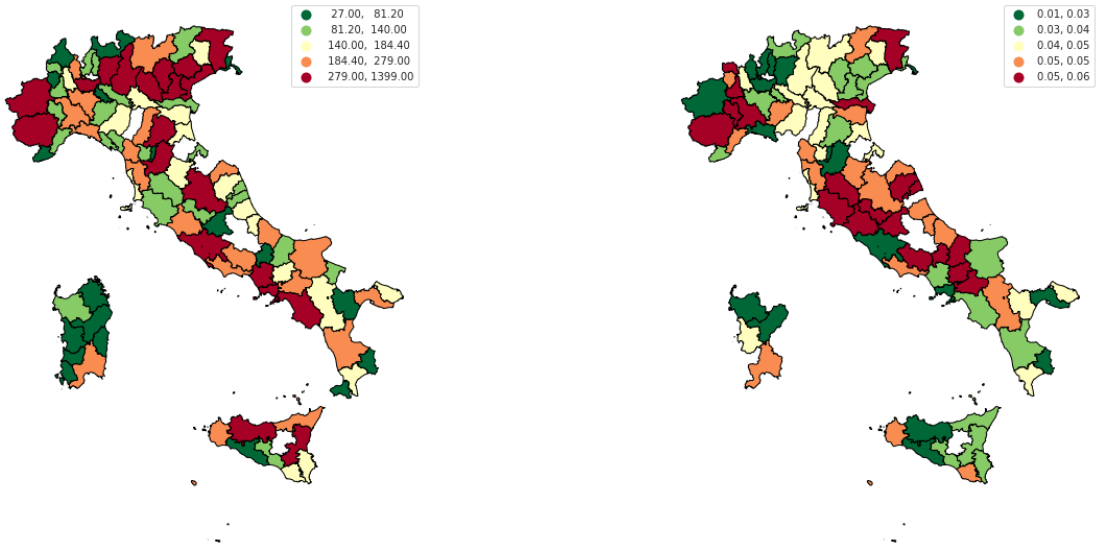
Figure 8: Plot of the ratio between the prices of fuels over that of Brent during 2021.

of road can affect the price of the fuel. In this section we will integrate other external factors that may contribute on the decision for a strategical placing of a new gas station. In particular, we wanted to find out a place with a low number of gas stations with respect to the density of the population in that territory, taking in mind that metropolitan cities have a huge concentration of public transports that not necessarily use the fuels (e.g. tram, metro, train) and that consequently are not of our interest. Once we have located the gas station we can then pass at the decision of which type of fuel to sell, taking into account the major requests.

One interest that a gas station may have is related to the amount of clients that it may have: the more the clients, the more the demand of fuel, the more the profit. Since there are no available (open) information of this kind, we tried to guess this data by working with the amount of people that populate a specific territory, taking also into consideration the number of existing gas station in that specific territory. The idea is the following: in provinces with high density of population, but with low number of gas station, we expect that there are a higher possibility of having a higher number of clients, whit less competitors, and thus a higher profit.
To reach our goal we imported from the istat site[13] the information about the density of population per province and we merge it with the amount of flags distributed across Italy, represented in Figure 9a. We ended up with Figure 9b in which is shown the percentage of gas station with respect to the amount of people who lived in each province in 2021.
The provinces with highest percentages are Grosseto (Toscana), Rovigo (Veneto) and Viterbo (Lazio) with respectively percentage of 6.15%, 6.06% and 5.99%; while the provinces with lowest percentages are Agrigento (Sicilia), Trieste (Friuli-Venezia-Giulia) and Milan (Lombardia) with respectively percentage of 1.44%, 1.47% and 2.17%. This last, in particular, is one of the metropolitan cities identifed by the Parliament in the law of the 7th April 2014, n. 56.
Since Agrigento is not a metropolitan city we could think of establishing the new gas station there.

The next step is to identify which type of fuels are more interesting to have, in order to increment the number of customers.

[13] http://dati.istat.it/Index.aspx?QueryId=18460



(a) (b)

Figure 9: (a) is the visual representation of the number of gas station in Italy in 2021; while (b) is the corresponding percentage of flags per density of population in 2021, at province level.

6 Conclusions

Describe the results obtained from a technical point of view but also from a conceptual one, like if we have to explain the work to a technical audience.