From this assignment, our aim is to understand the supervised learning technique “Linear Regression”. As part of the assignment we worked on the following steps:

Research goal (“identify your story”)

    Used car market is a big market around the world. Increase in demand for used cars is not only in the developing countries but also in the developed countries. This increase in demand for used cars ( either from the direct seller or the cars whose lease have ended), there has to be a correct way to determine the residual value of the car that is the “price” by evaluating features of the car. This is important problem to solve, else people not knowing the residual value of the car could fall trap in the hands of sellers/dealers with unrealistic price. Also, rise in prices of new cars has lead to people choosing used cars. So we decided to use “price” as our target variable.

we took the dataset of used cars from Europe from kaggle : <https://www.kaggle.com/mirosval/personal-cars-classifieds>

This dataset has nearly 3.5 million records. Also, this dataset has good number of feature columns ( maker, model, mileage, manufacture year and so on) with high range of values including bad data which is a perfect dataset for machine learning exercise to prepare/clean the data.

Columns:

Maker: Maker of the car ( this dataset has all normalized names in lowercase)

model: Model of the car.

mileage: mileage of the car in kms

manufacture\_year

engine displacement in cc

engine power

body type, color\_slug, stk\_year: Values of these columns are almost not available. So eventually dropped these columns from analysis

transmission: transmission type if its automatic or manual

door

seat count

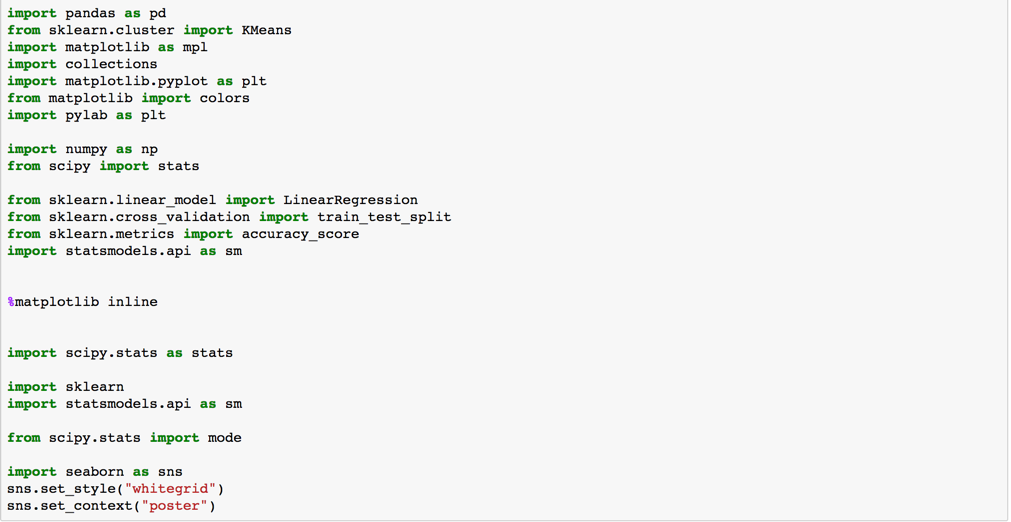
fuel type of the car

price ( this dataset has the price in euro’s)

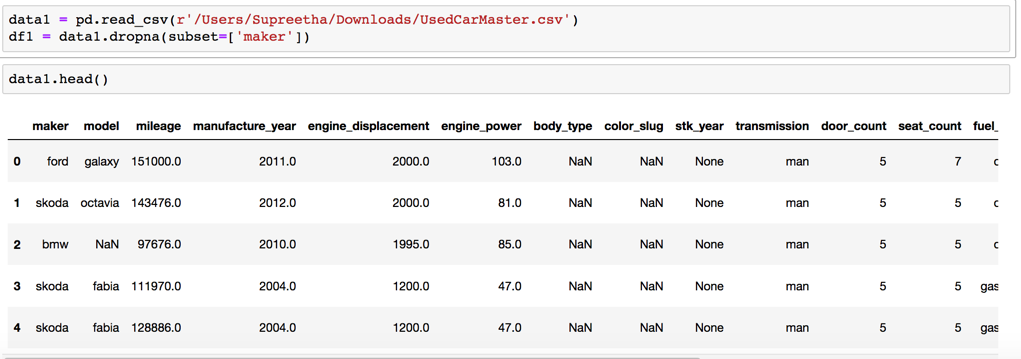
Data Cleaning and Processing (“data preparation”)

    In this step, we wanted to understand the dataset in depth so that we can get idea on the values, anomalies if any exist and how we could enrich the data based on the outcome from our prediction.

Importing all the required libraries that we used:



From our initial analysis, we found out that there were lot of missing values. Mainly the maker column itself had missing values. It is purely bad data. Since maker of the car has to be present ( else we never know which car is that) we decided to drop those records, code below



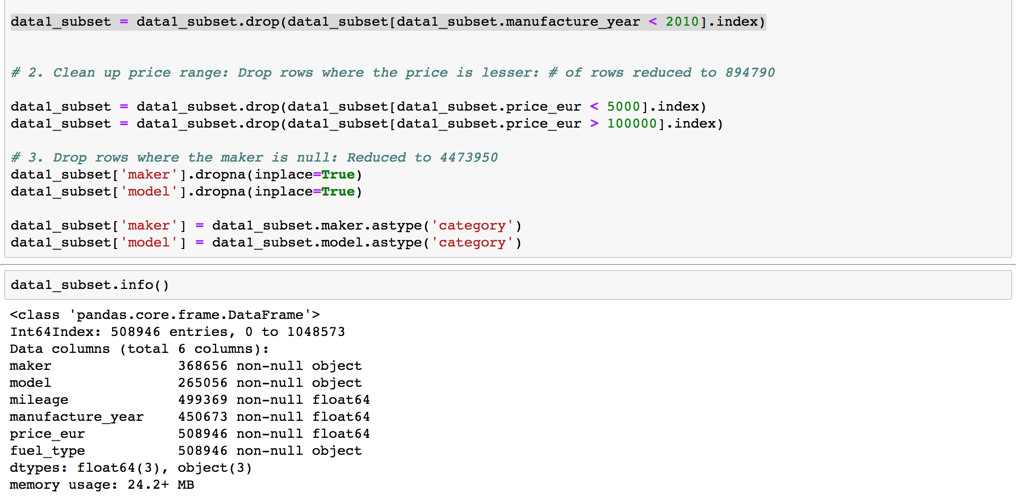
As mentioned above, since this dataset is huge and had too many anomalies, we came to a conclusion, that we will look at the data, only for following features :

maker, model, mileage, manufacture year, price, fuel type.

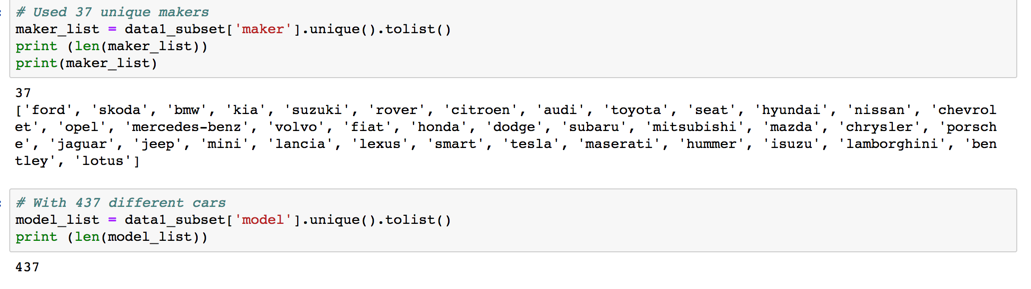
data1\_subset = data1[['maker', 'model', 'mileage', 'manufacture\_year', 'price\_eur', 'fuel\_type’]]

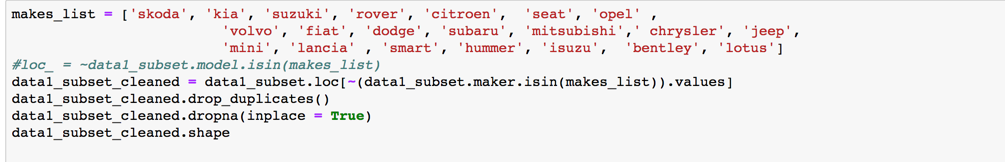
We did a quick analysis on the number of vehicles spread across years and saw that there was good enough data  for cars that were manufactured after 2010.

We also looked at the distribution of data for price, we found out that majority of the dataset was between the range 5000 to 100000. So we applied all the filters to eventually end up at well distributed dataset



When we did distinct values for make, we saw there were 37 unique values. this was very huge. So we decided to limit the dataset to a few makers so that for those smaller set of makers, we wanted to predict the price using linear regression and determine the accuracy.





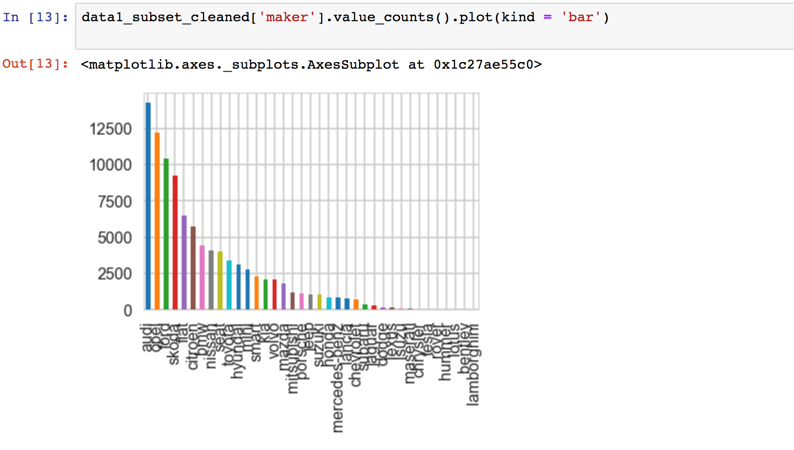
/var/folders/mq/xk0scgln5bg7x423qyt2gwkm0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/p18

After reducing the dataset based on the above analysis, we wanted to check the number of cars by year to see if there was any inference we could conclude. We found out that our dataset has highest manufactured car count for the year 2015. this analysis is in-line with the actual reality. As per the article "<https://www.washingtonpost.com/business/economy/us-car-sales-hit-record-high-in-2015/2016/01/05/363aaf30-af14-11e5-b820-eea4d64be2a1_story.html?noredirect=on&utm_term=.958cb298154d>”, 2015 did indeed have the highest number of cars manufactured.

Exploring data with Visualization

    In this step, we wanted to understand the patterns, relations within the variables in the dataset so that we could identify anomalies if any.

    as shown below, we wanted to analyze by maker, how many findings were available in the dataset?

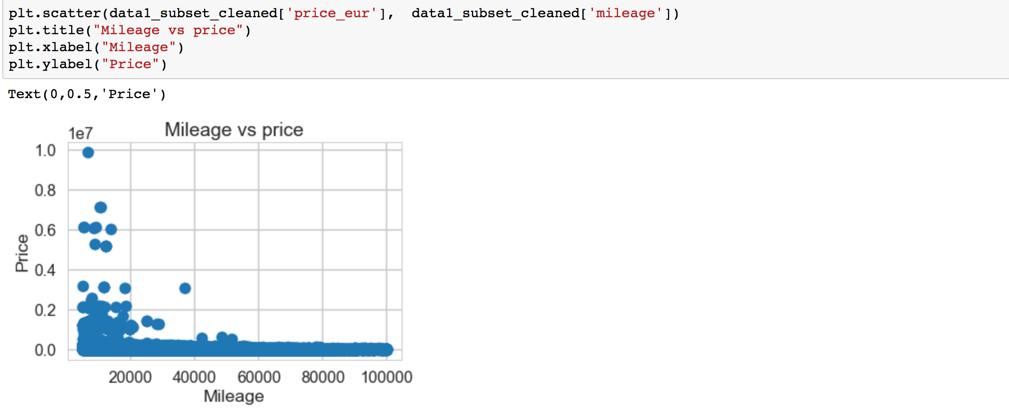


What we could infer from the above screenshot is that cars from makers such as audi, opel, ford etc had highest number of cars available in the used car market for re-sale.

/var/folders/mq/xk0scgln5bg7x423qyt2gwkm0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/p20

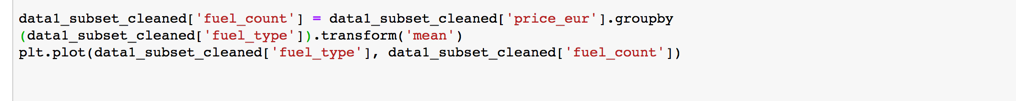
price\_mileage\_df = data1\_subset\_cleaned[["price\_eur", "mileage”]]

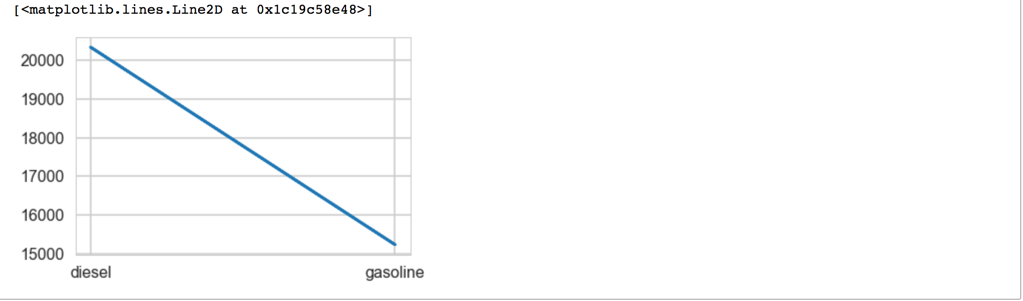
We also, wanted to understand the distribution comparison of price vs mileage ( to see the co-relation between these two variables)



From the above analysis, we can infer that lesser the mileage of the car, higher the price ( or residual value) of the car.

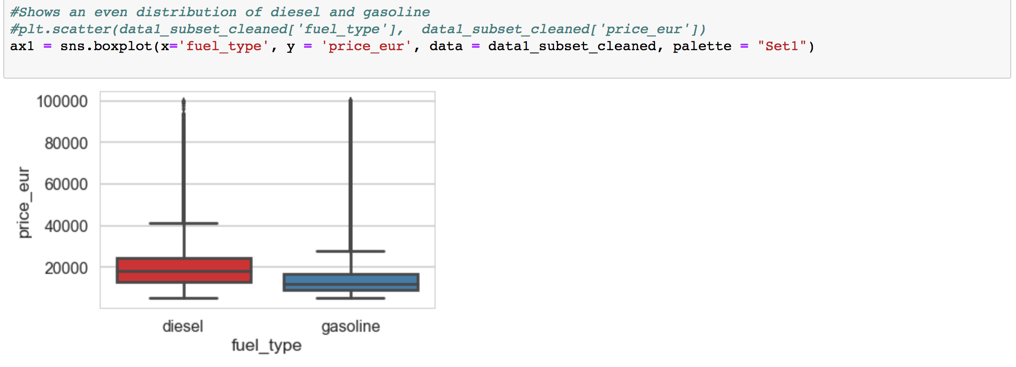
Going one more step ahead, we wanted to understand the average price of the car given the fuel type.





From the above screenshot, it is very clear that the price of a diesel car, was more expensive than gasoline car

we also ran the price vs fuel\_type box plot to understand the standard deviation of the dataset and if its within the acceptable range.



We can conclude from this analysis is that, the average sale price of used diesel vehicle is much higher than the corresponding gasoline fueled car.

Mean price for diesel car is higher even after one deviation ( both the One deviation is above and below the box plot)

Data Model

    In this step, we try to build a model to predict our target variable “price” based on other feature variables using linear regression.



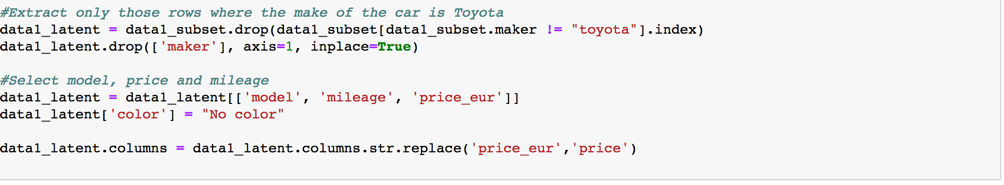
<Write Linear/Multiple Regression>

From the above we can infer that..

Data Enrichment

    In our original dataset, we had cars from various manufacturers. While searching for similar dataset, found another dataset that had the similar features, but it was only for toyota. The new dataset had additional column “color”. We wanted to see if bringing in color into our analysis would have any influence on the prediction

In the original dataset, apply filter to bring in only Toyota cars, add additional column ‘color’.



Import the second dataset, cleaning to get the subset of data in similar range as before



Combining the two data set.

/var/folders/mq/xk0scgln5bg7x423qyt2gwkm0000gn/T/com.microsoft.Word/WebArchiveCopyPasteTempFiles/p31

—