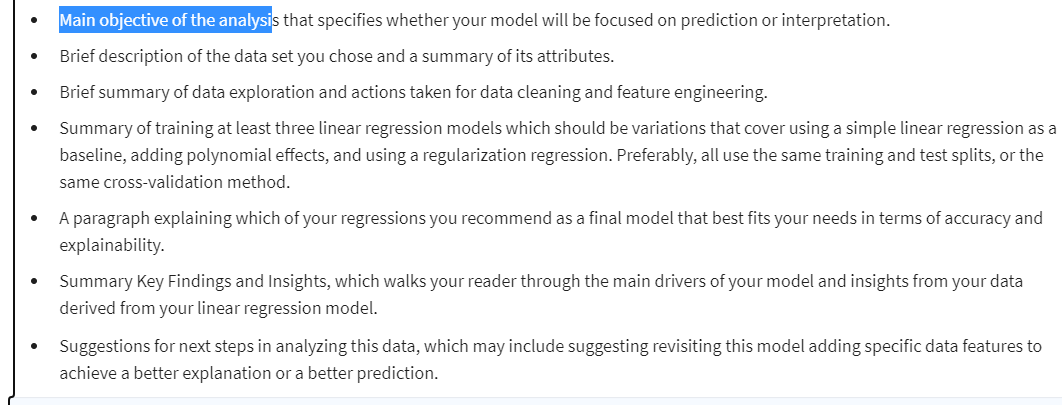
COURSERA IBM MACHINE LEARNING –REGRESSION – COURSE PROJECT

Author: Wojciech Gołębiowski



1. Main objective of the analysis

A main objective of this analysis is to predict selling price of the car given some attributes (criterias). This work will focus mostly on predictive aspect of modelling though I will also check which coefficients are relatively the largest. This will give additional information to the analysis – it will reveal the most influential attributes.

1. Brief description of the data set

Dataset used in analysis presents used vehicles data set. In general the datasets consists of 8128 records and presents variables such as:

* Car’s name
* Year produced
* Selling price (target variable – in India Rupias)
* Km’s driven
* Fuel type
* Seller type
* Transmission
* Owner (First/Second etc)

There are 3 numeric variables and 5 categorical variables. This dataset is available under public access rights on Kaggle.com - <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho>.

In the below tables I present a short summary of main statistics for each of the analysed variables.

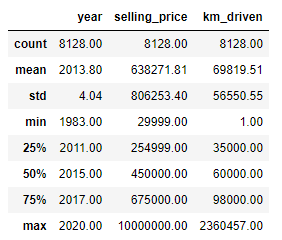


Figure 1. Main statistics for numerical variables

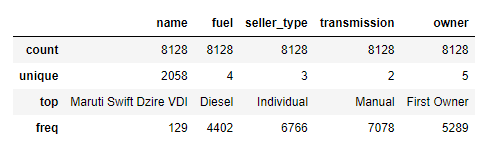


Figure 2. Main statistics for categorical variables

As presented in Figure 1 there is a wide range of years when the analysed cars were produced (the oldest one is from 1973 and the newest from 2020). Selling price has an enormous standard deviation which might point to the existence of outliers. In later analysis we might take a look at this observations.

As presented in Figure 2 all categorical variables seem to be filled in. Some of the car names seem to appear in the data set repeatedly. This might indicate existence of duplicated records.

1. Brief summary of data exploration and actions taken for data cleaning and feature engineering

The plan of data exploration will contain the following steps:

* Searching for duplicates
* Variable transformations
* Examining missing data and potential data imputation process (as present in the Figure 1 there is one variable (price) which seems to be missing many records of the data
* Examining outliers (as per Figure 1 the variable ‘price’ seems to contain outliers as there is a huge difference between mean and median. Additionally standard deviation for this variable is relatively big)
* Feature engineering – adding new columns and transforming existing ones.

First step was to look for duplicates. It turns out there are a lot of duplicated entries in our dataset. The approach to these records is a bit tricky. Hypothetically there might be a case when identical cars (with the very same parameter) were sold for the same selling price. Nonetheless in this research I presume these are duplicates as the records were gathered in web scrapping process from websites with car offers. This mean that on the different (or even on the same) website there might be duplicated car offerings. Going forward all duplicated records were removed – we are left with 6925 records.

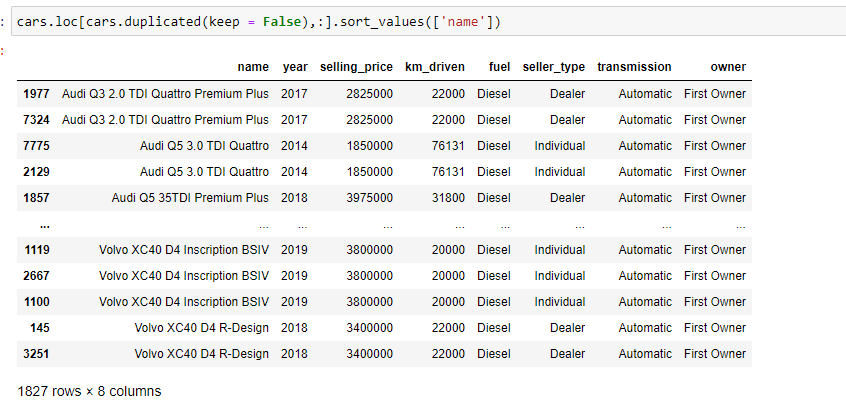


Figure 3. Duplicated entries

Second step was data transformation. In this step I am removing a word ‘*Owner’* from ownervariable. Also I am changing name variable. I will leave just the company name as it might bring additional information whereas the particular car model is too granular level to bring interesting insights. Results of these actions are presented in Figure 4 and 5.

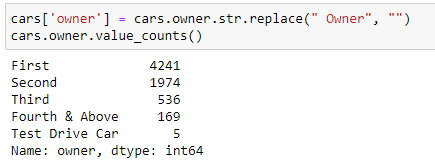


Figure 4. Owner parameter after transformation

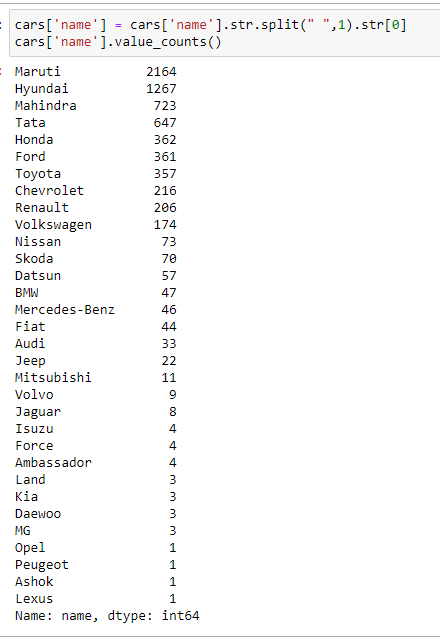


Figure 5. Name parameter after transformation

As we can see some of the categories are not very well represented in the dataset. This may lead us to transform them further to exclude not well-represented categories or to remove these variables from analysis entirely.

Third step was to examine missing data. In figure 3 I present the number of missing records for each of the analysed variables.

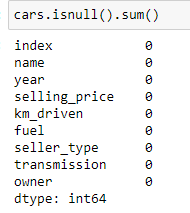


Figure 6. Missing values in variables

None of the variables contain missing entries.

Fourth step is recognizing outliers. To do that boxplots for each of analysed numerical variables were created. Below I present the boxplots generated using python’s seaborn package.

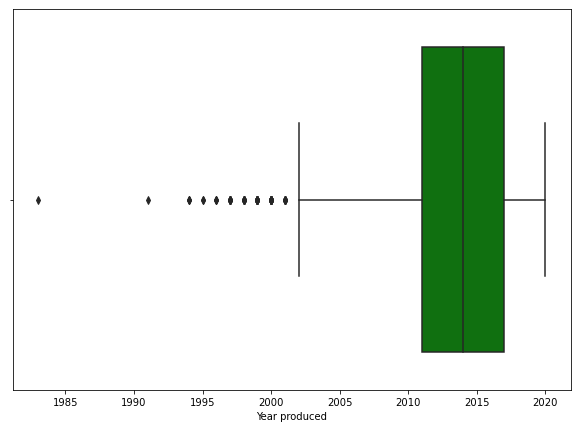


Figure 7. Boxplot for Year produced variable

There are at least 10 outliers in Year produced variable. Nonetheless these records are outliers only in theoretical manner. It seems normal that there are cars from 1980s’ or 1990s’ which are offered on the auctions. These might be collector’s vehicle which can have a higher price than the other cars which potentially might deteriorate the future models. We can investigate that by comparing the mean selling price of cars produced before and after year 2000. High difference might point out that the oldest examples are indeed collector’s vehicles and we might be pushed to eliminate them from the data set.

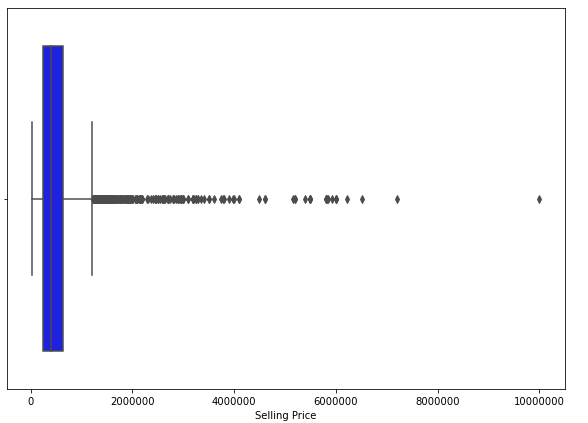


Figure 8. Boxplot for Selling Price variable

There seems to be a lot of outliers. One observation is particularly interesting – one of the cars was sold for 10m INR. This one might be mistakenly entered into the database. This record is a Volvo vehicle produced in year 2017 with 30 000 kilometres driven. This seems like a record which was not entered correctly into the database. Therefore it will be removed.

There were some outliers spotted for kilometres driven variable. Nonetheless they seems normal (cars with huge mileage are also being sold on auctions) and I will leave them for further analysis.

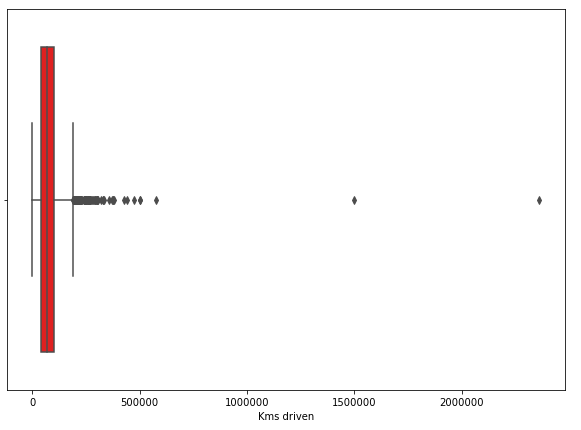


Figure 9. Boxplot for kilometres driven

The last – fifth step – is feature engineering. In the current stage the only action which I will take is creating a two new variables showing whether the flat was already built and price per squared meter which will be calculated by dividing flat price by its area in squared meters. Statistics for newly created price per square meter variable you can find in Figure 10.

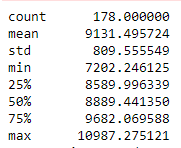


Figure 13. Statistics for Price per square meter

Maximum price per square meter in the neighbourhood equals almost 11 000 PLN whereas the minimal value equals 7 200 PLN. As all of the flats are located in the same neighbourhood this difference seems to be big.

The second variable refers to whether flat was already built or not. Statistics for newly created categories you can find in Figure 11.

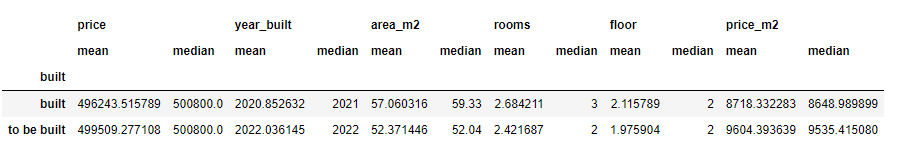


Figure 14. Statistics for built/to be build categories

There are two interesting things you can see there:

* Flats not yet build tends to be smaller (5 square meters difference in terms of average)
* Flats not yet built tends to be more expensive (almost 1000 PLN more expensive when it comes to price per square meter)

1. Summary of training linear regression models

Key findings from Exploratory Data Analysis:

* Data regarding number of rooms, floor number and area is quite flat.
* For many of the offers the developer doesn’t present the price. These records were replaced by the median value of price of flats.
* There was one outlier showing price of more than 4000000 PLN which seems to be wrong. The hunch is that too many 0 were entered into the offer. This record was excluded.
* Some of the flats differs from the mean price by maximum of 100000 PLN. This might be caused by some additional facilities present/missing in the flat or by number of square meters.
* The price per square meter shows a big difference in pricing. All of the flats are located in the same neighbourhood, though for some of them the price per square meters reaches 11 000 PLN where for the others it is around 7 500 PLN.
* Flats not yet build tends to be smaller but more expensive (they are around 5 square meters smaller but 1000 PLN more expensive in regards to price per square meter)

1. Recommendation for a final model

Based of the key findings presented in point 4 we can enrich the conclusions by formulating the below hypothesis:

* 1st hypothesis: There is a difference in price per square meter between flats already built and ones to be built
* 2nd hypothesis: There is a difference in number of square meters in flats already built and ones to be built
* The mean price of flats located on 1st, 2nd, 3rd floor are the same

1. Summary Key Findings and Insights

I will conduct a hypothesis test for the first hypothesis presented in point 5. First I will define hypothesis:

H0: There is no difference in price per square meter between flats already built and ones to be built

H1: There is a difference in price per square meter between flats already built and ones to be built

Next, using pythons ttest\_ind function, we will conduct a t-test at a alpha = 0.05 significance level. Results of the test are presented in Figure 12.

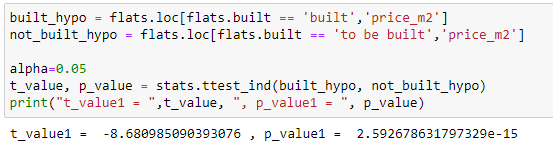


Figure 15. Results of t-test conducted for formualted hypothesis

As the p-value is less than alpha (0.05) we reject the null hypothesis that there is no difference in price per square meter between flats already built and flats to be built. This confirms our assumption that developers are asking for higher bids for flats which are currently under construction.

1. Suggestions for next steps in analyzing this data

In our analysis we formulated very interesting points and discovered powerful insights regarding local housing market in Wrocław. To enhance our knowledge regarding the reasons staying behind higher prices for some of the flats in the next steps of the analysis we might need to gather additional data regarding existence/non-existence of additional facilities in flats e.g. pools, solar systems, balcony, terrace etc. This binary variables might show why some of the flats are priced higher than the others. Having this binary variables as well as already possessed ones we would be able to predict the price of a new flat by features such as floor number, number of rooms, year build and existence of facilities. That might help people considering buying a new flat in the future in preparing sufficient money contribution and evaluating bank mortgage offers.