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.

2. 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.

	year	selling_price	km_driven
count	8128.00	8128.00	8128.00
mean	2013.80	638271.81	69819.51
std	4.04	806253.40	56550.55
min	1983.00	29999.00	1.00
25%	2011.00	254999.00	35000.00
50%	2015.00	450000.00	60000.00
75%	2017.00	675000.00	98000.00
max	2020.00	10000000.00	2360457.00

Figure 1. Main statistics for numerical variables

	name	fuel	seller_type	transmission	owner
count	8128	8128	8128	8128	8128
unique	2058	4	3	2	5
top	Maruti Swift Dzire VDI	Diesel	Individual	Manual	First Owner
freq	129	4402	6766	7078	5289

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.

3. 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 combined with testing the assumptions for Linear Reegression

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.

```
: cars.loc[cars.duplicated(keep = False),:].sort_values(['name'])
```

:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
1977	Audi Q3 2.0 TDI Quattro Premium Plus	2017	2825000	22000	Diesel	Dealer	Automatic	First Owner
7324	Audi Q3 2.0 TDI Quattro Premium Plus	2017	2825000	22000	Diesel	Dealer	Automatic	First Owner
7775	Audi Q5 3.0 TDI Quattro	2014	1850000	76131	Diesel	Individual	Automatic	First Owner
2129	Audi Q5 3.0 TDI Quattro	2014	1850000	76131	Diesel	Individual	Automatic	First Owner
1857	Audi Q5 35TDI Premium Plus	2018	3975000	31800	Diesel	Dealer	Automatic	First Owner
...	...	...	...	...	...	...	...	...
1119	Volvo XC40 D4 Inscription BSIV	2019	3800000	20000	Diesel	Individual	Automatic	First Owner
2667	Volvo XC40 D4 Inscription BSIV	2019	3800000	20000	Diesel	Individual	Automatic	First Owner
1100	Volvo XC40 D4 Inscription BSIV	2019	3800000	20000	Diesel	Individual	Automatic	First Owner
145	Volvo XC40 D4 R-Design	2018	3400000	22000	Diesel	Dealer	Automatic	First Owner
3251	Volvo XC40 D4 R-Design	2018	3400000	22000	Diesel	Dealer	Automatic	First Owner

1827 rows × 8 columns

Figure 3. Duplicated entries

Second step was data transformation. In this step I am removing a word ‘*Owner’* from owner variable. 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.

```
cars['owner'] = cars.owner.str.replace(" Owner", "")
cars.owner.value_counts()
```

```
First             4241
Second            1974
Third              536
Fourth & Above     169
Test Drive Car       5
Name: owner, dtype: int64
```

Figure 4. Owner parameter after transformation

```
cars['name'] = cars['name'].str.split(" ",1).str[0]
cars['name'].value_counts()
```

```
Maruti           2164
Hyundai          1267
Mahindra          723
Tata              647
Honda             362
Ford              361
Toyota            357
Chevrolet         216
Renault           206
Volkswagen        174
Nissan             73
Skoda              70
Datsun             57
BMW                47
Mercedes-Benz      46
Fiat               44
Audi               33
Jeep               22
Mitsubishi         11
Volvo               9
Jaguar              8
Isuzu               4
Force               4
Ambassador          4
Land                3
Kia                 3
Daewoo              3
MG                  3
Opel                1
Peugeot             1
Ashok               1
Lexus               1
Name: name, dtype: int64
```

Figure 5. Name parameter after transformation

There is a big scatter of the brands. As it may not make sense to include such a variable in the model I will remove it from analysis – there is a lot of categories which are not significantly represented in the data set which may lead to some problems in later stage of analysis (after implementation of one hot encoding this may create a big number of variables which will not bring any additional information).

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

```
cars.isnull().sum()
```

```
index            0
name             0
year             0
selling_price    0
km_driven        0
fuel             0
seller_type      0
transmission     0
owner            0
dtype: int64
```

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.

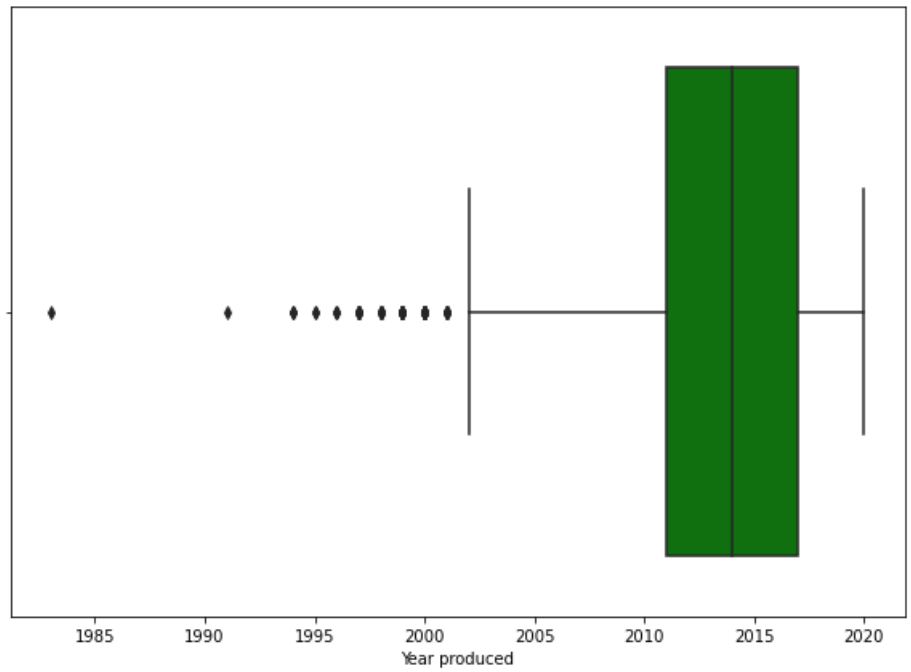

Figure 7. Boxplot for Year produced variable

There are 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. Mean selling price of cars produced before year 2000 is 6 times smaller than the mean selling price of cars produced after year 2000. This numbers seem to deny my assumption that the oldest cars might be collector's vehicles. To summarize I will leave all observations in the data set.

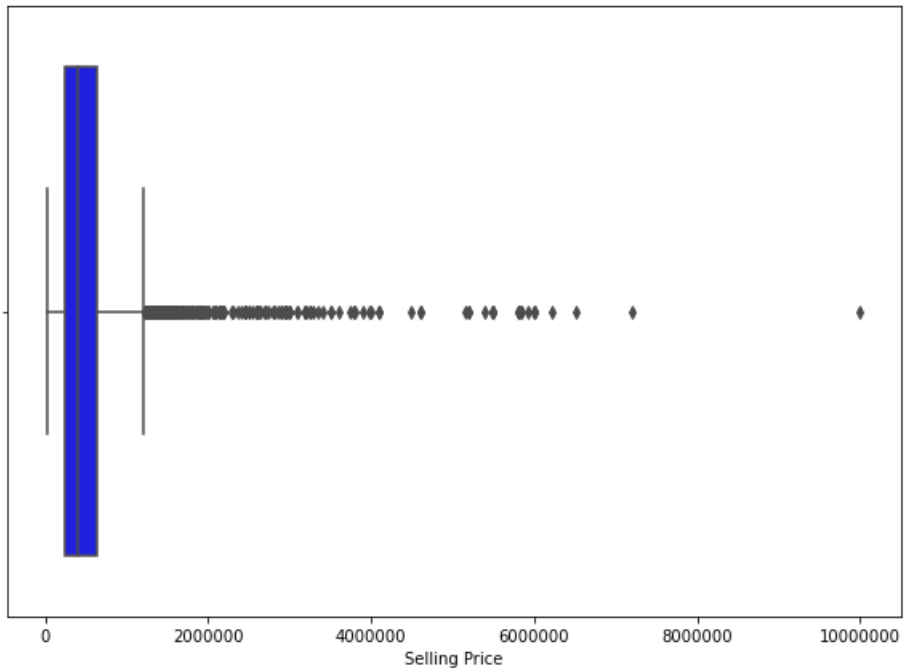

Figure 8. Boxplot for Selling Price variable

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. Two biggest observations will be removed from the analysis as they might skew the following analysis. The other part seems normal (cars with huge mileage are also being sold on auctions) and I will leave them for further analysis.

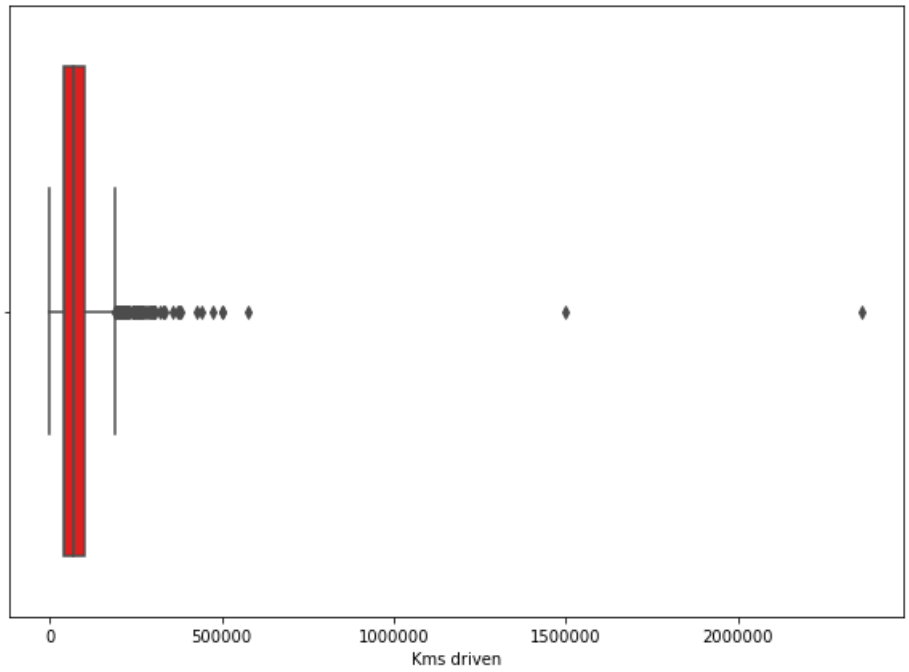

Figure 9. Boxplot for kilometres driven

In fifth step I will test the assumptions for linear regression and (if required) transform the variables so that they will better meet criteria. If the assumption will not be met and transformation will not be possible or it will not help the analysed feature might be dropped from the analysis.

First assumption I will check is the normality – normally it is good to have dependent variable which is normally distributed as it leads to better results. This can be checked through visual analysis of Q-Q plot and histogram or through running Kolmogorov-Smirnov test.

Below I present the histogram and Q-Q plot for target variable – *selling price*. Based on this graphical presentation we can see that target variable is not normally distributed. Additionally the data seems to be right-skewed.

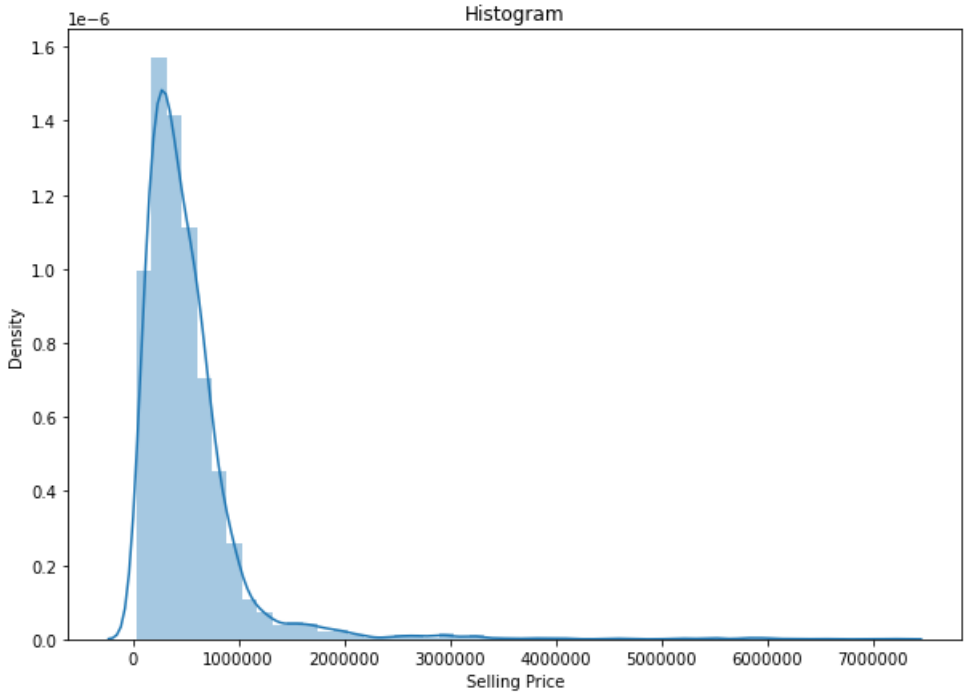

Figure 10. Histogram for selling price variable

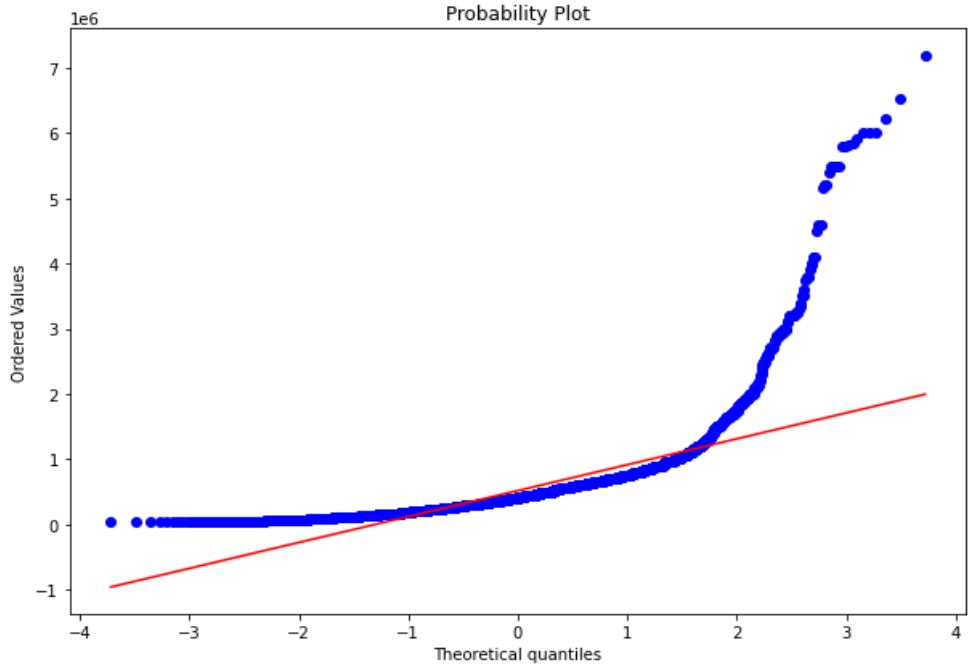

Figure 11. Q-Q Plot for selling price variable

Additionally we can perform normality test. P-value for selling price variable equals 0 which confirms that target variable is not normally distributed. To overcome this issue we can transform our target using one of the most popular methods. First let's see the results after log-transforming our target. Results of this action can be seen on subsequent figures.

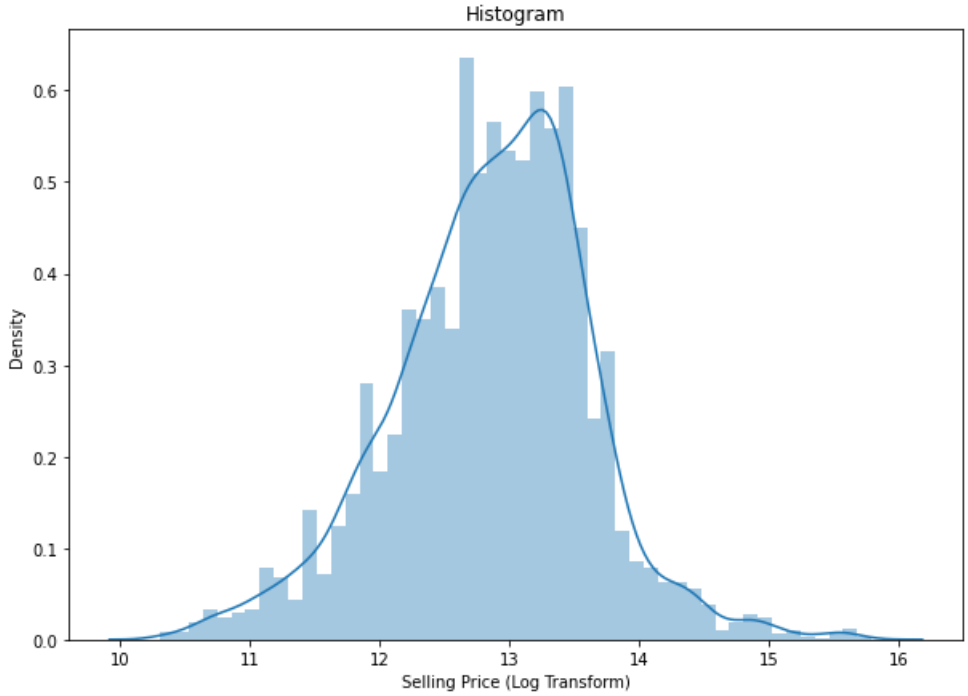

Figure 12. Histogram for selling price variable after performing log transformation

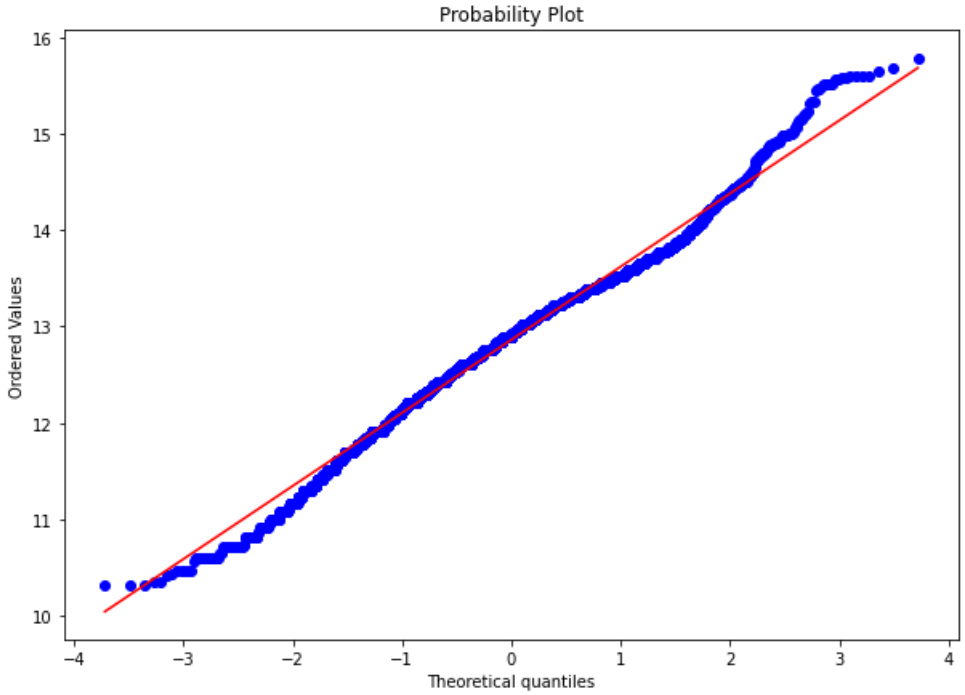

Figure 13. Q-Q Plot for selling price variable after log transformation

This looks a way better. Unfortunately it is still not normally distributed which is confirmed by results of normality test in which P-value equals 4.56368736193202e-24.

Second approach I will investigate is box-cox transformation. Subsequent figures present the results of implementation of box-cox transformation on our target.

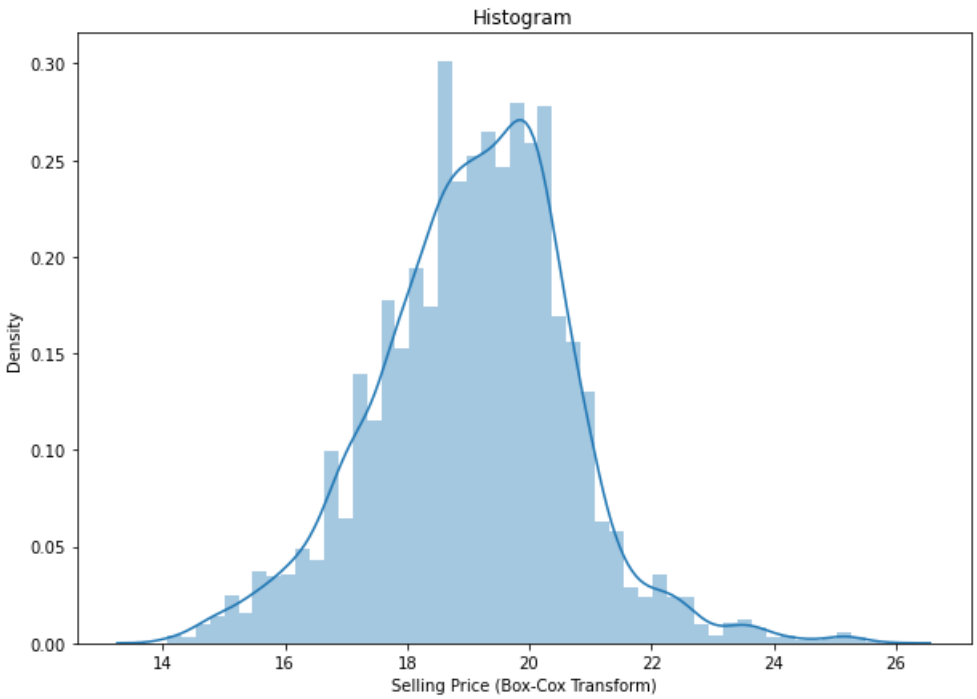

Figure 14. Histogram for selling price variable after performing box-cox transformation

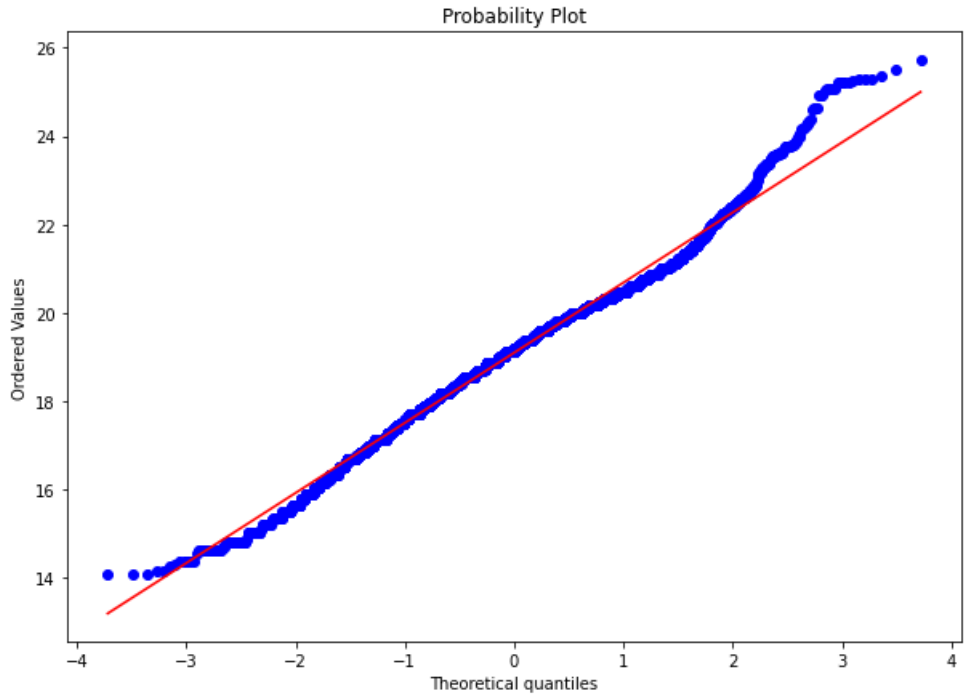

Figure 15. Q-Q Plot for selling price variable after box-cox transformation

Again it looks a bit better but unfortunately it is still not normally distributed which is again confirmed by results of normality test in which P-value equals 6.0955846054342926e-21.

Nonetheless this is a way closer to be normally distributed than in previous cases. Therefore going forward I will treat selling price after box-cox transformation as my target variable.

Second assumption that will be checked is homoscedasticity (i.e. if the residuals fluctuate randomly around the line). I will check that assumption for a pair of variables including the target and others continuous variables (In my case only Kilometres driven). It is a crucial assumption and can be checked using residual plot which is presented in figure 16.

In figure 16 error variance across the true line seems to be dispersed uniformly. There are some values which are pulling out from the rest in the middle which might be worrying. Nonetheless I decided to leave this variable in for further analysis.

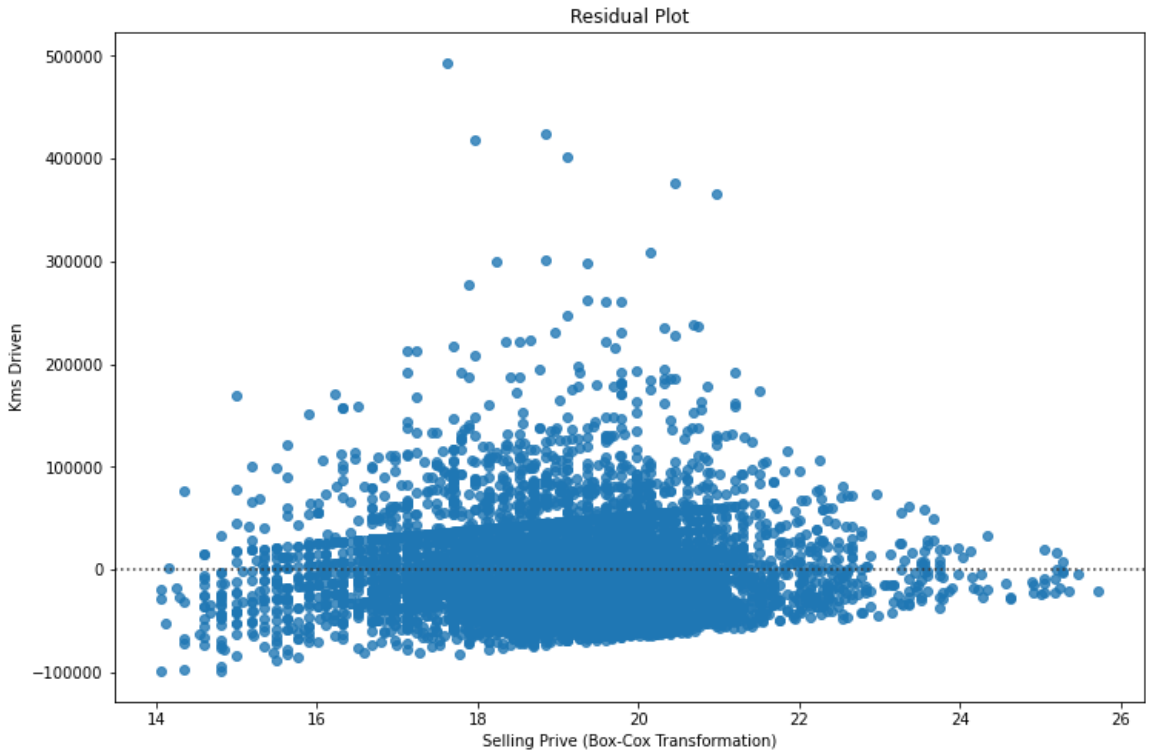

Figure 1616. Residual Plot for selling price (Box-Cox transformation) and Kilometres driven

The third assumption that will be checked is linearity assumption that can be evaluated using scatter plots. Those will be created for pairs of variables including the target and two numerical variables: kilometres driven and year. Result of this analysis can be found on Figure 17. Analysis of the graph leads to conclusion that there seems to be linear relationship between variables year and selling price and there is no linear relationship between variables kilometres driven and selling price. Therefore kilometres driven will not be taken into consideration when feeding the data into the linear regression model.

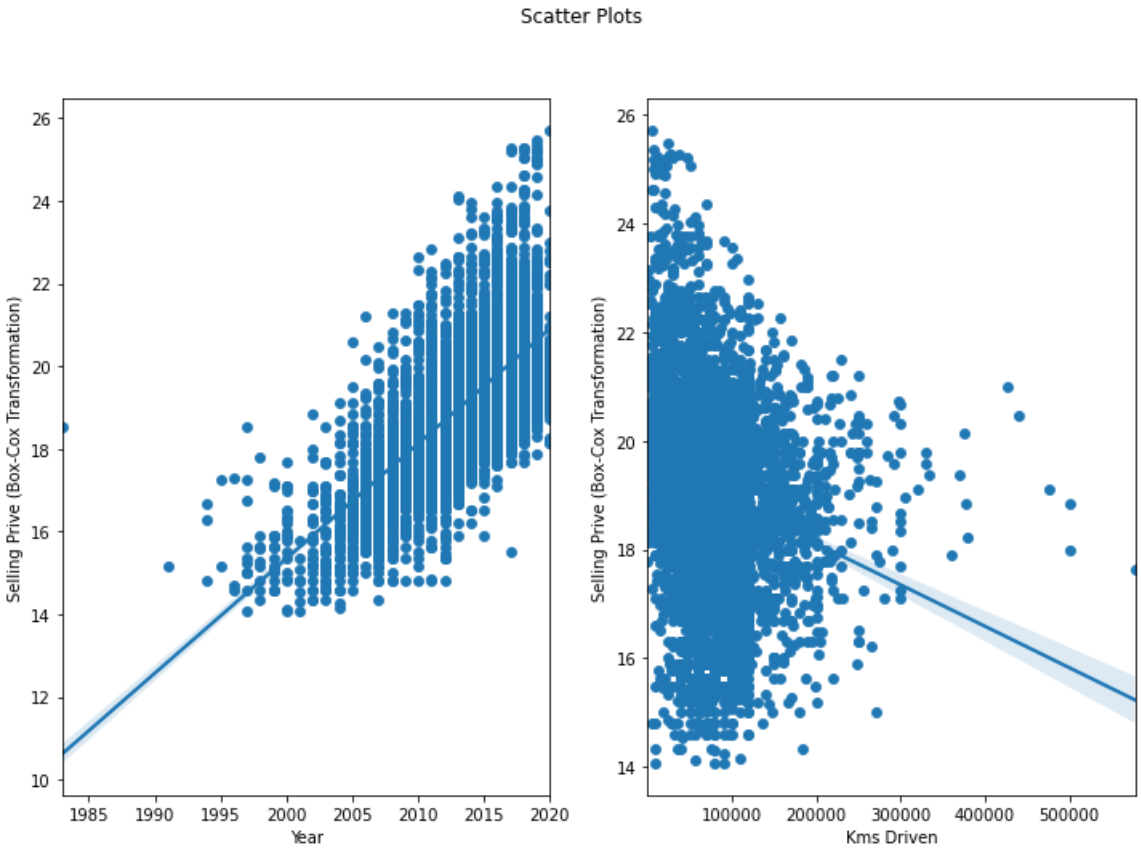

Figure 17. Linearity between Selling Price (Box-Cox transformation) and Year and Kilometres Driven variables

Fourth assumptions that will be tested will be multicollinearity between the independent variables. To allow this analysis also for categorical variables I will first calculate V-Cramer's association measure which will present the intercorrelation between categorical variables.

Results of V-Cramer calculations are presented in the Figure 18. Values close to 1 suggest perfect association between the categorical variables. Nonetheless in this dataset the strongest value is 0,21 (between transmission and seller type). This indicate that there is no autocorrelation between categorical variables.

	features	resuls
0	fuel / seller_type	0.050539
1	fuel / transmission	0.042713
2	fuel / owner	0.032458
3	seller_type / fuel	0.050539
4	seller_type / transmission	0.211931
5	seller_type / owner	0.136735
6	transmission / fuel	0.042713
7	transmission / seller_type	0.211931
8	transmission / owner	0.112560
9	owner / fuel	0.032458
10	owner / seller_type	0.136735
11	owner / transmission	0.112560

Figure 18. V-Cramer's association measure between categorical variables

To summarize the variables which are selected as predictor's and will be considered in modelling phase are:

- Year produced
- Fuel type
- Seller type
- Transmission
- Owner (First/Second etc)

Cars name and kilometres driven were rejected. First variable because of big category dispersion (many not numerous classes). Second one because of not meeting linearity assumption.

4. Summary of training linear regression models

First step before moving into the modelling phase was to convert categorical variables into binary classes. This changes the dimensions of our dataset. We moved from 5 to 11 variables (approach used was to leave k-1 levels of categorical variables).

I will fit 3 linear regression models with analyzed data:

- Basic Linear Regression model
- Linear Regression model with polynomial features

- Ridge Regression model

30% of the data will be taken into training set (around 2077 observations) and around 4845 observations will constitute training set.

Results of key evaluation metrics can be found in a Table 1.

Table 1. Key evaluation metrics' results

Model	R2 score on training data	R2 score on test set	MSE
Simple Linear Regression	68%	69%	0.841
LR with polynomial features	68%	69%	0.78
Ridge Regression	70%	70%	0.75

Interestingly Linear Regression model with polynomial shows identical results. Its because after implementing Grid Search CV it turned out that the most optimal approach would be to keep polynomial at a first degree. Difference in MSE score between simple Linear Regression and one using Polynomials might be caused by standardization which was implemented in the 2nd approach.

For Ridge Regression the most optimal approach was to first use 2nd degree polynomial and then during modelling regularization parameter equal to 4.

5. Recommendation for a final model

To decide which model is the best I will compare the values of goodness of fit statistic (coefficient of determination) and mean square error (MSE). In the first instance the closer to 1 the better the model and in terms of the second measure the lower the value the better.

Based on the comparison visible in Table 1 I recommend Ridge regression model as the final model. It is featured by the highest values of coefficient determination (both for training and testing set) and the smallest value of MSE comparing to other models. R2 score equal to 0.7 means that around 70% of variability in the data is explained by the model. This is a strong result. Nonetheless it will not constitute a perfect model.

6. Summary Key Findings and Insights

To summarize – for the sold cars dataset firstly I've made exploratory data analysis which was then followed by modelling exercise. During EDA I found out that there is a huge distortion in terms of brands analysed in the data set. Additionally it turned out that the data contains single records describing cars produced before year 2000 which mean selling price was about 6 times smaller than average selling price of cars produced after year 2000. As well there was one outlier in terms of selling price itself which can be assumed as an entry error.

In the next step I tested the linear regression assumptions. Firstly it turned out that target variable doesn't meet the normality assumption so it was transformed using box-cox transformation method. It didn't bring the ideal state but it moved the target closer to the normality. Homoscedasticity was met for pair of variables selling price and kilometres driven. Nonetheless mileage variable eventually didn't met linearity assumption so it was removed from modelling phase. Last check performed was multicollinearity assumption which was negated for both numerical variables and categorical variables (using V-Cramer statistic).

Eventually the data had been put into the 3 models from which the best one turned out to be the Ridge Regression model. It achieved 70% coefficient of determination level which can be perceived as good but not perfect.

7. Suggestions for next steps in analysing this data

In pursuit of finding a better explanation of selling price I would strongly suggest to gather more records. The final analysis was run on around 6000 observations which doesn't constitute a huge sample. Additional bits of data might change the currently observed statistics and change the results of the linear regression assumptions. Especially the brand variable can benefit from higher number of rows analysed. Currently this category is dispersed significantly – there are brands which have thousands of records but most of them contains hundreds or even dozens of records. In my opinion brand could be a game changer in terms of modelling as it strongly divide the cars into premium/non-premium segment. Brands like Mercedes or Audi are synonyms of high quality and they are recognized with higher prices. Eventually this might improve the accuracy of the model.