

Car Price prediction

Submitted by:

Ashish Anand

**ACKNOWLEDGMENT**

This includes mentioning all the references, research papers, data sources, professionals, and other resources that helped you and guided you in the completion of the project.

**INTRODUCTION**

* Business Problem Framing

With the covid 19 impact in the market, we have seen a lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in the market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make a car price valuation model.

* Conceptual Background of the Domain Problem

The price of the car mainly varies with location, company,engine capacity, etc.

* Review of Literature

Many papers and articles have been written on this topic and Kaggle is one of the biggest open-source sites to state the research.

* Motivation for the Problem Undertaken

car price prediction is one of the most researched topics in the data science world and doing a project from scraping phage is a real-time experience so as a beginner it is always better to have good knowledge and tools in hand to work.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical, and analytics modeling done during this project along with the proper justification.

Mathematically the dataset consists of many categorical variables that need to be converter but before that, many null values should be taken care of so let’s start mathematical modeling by checking null values (df.IsNull().sum())

*city 1 0.00019739439399921041*

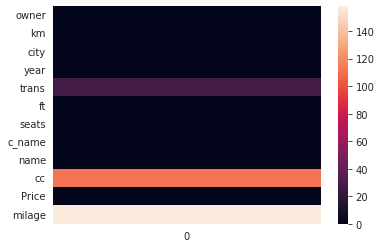
*trans 29 0.005724437425977102*

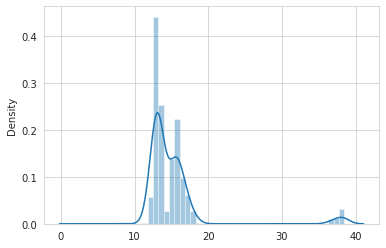
*seats 1 0.00019739439399921041*

*cc 112 0.022108172127911566*

*milage 158 0.031188314251875248*

Column name, number of null value, and percentage of the null present in the columns are shown above and all the null value is less then 5% so null rows can be dropped but engine capacity(cc) can be refiled according to car name.





distribution of Price is right-skewed so log transformation is needed. the above graph is the distribution plot of log-transformed Price. it can be seen there are some cars having very high price so lets analyse it by checking unique car name

*'Hyundai Santro', 'Maruti Suzuki Swift', 'Ford Ecosport',*

*'Honda Amaze', 'Renault Duster', 'Honda Civic',*

*'Toyota Etios Liva', 'Maruti Suzuki Swift DZire', 'Honda City',*

*'Maruti Suzuki Celerio', 'Hyundai Creta', 'Mahindra XUV500',*

*'Hyundai i20 Active', 'Maruti Suzuki DZire', 'BMW 5-Series',*

*'MG Hector', 'Maruti Suzuki Ciaz', 'Hyundai Xcent',*

*'Toyota Innova', 'Renault Kwid', 'Ford Figo', 'Volkswagen Vento',*

*'Honda Brio', 'Honda Jazz', 'Hyundai Venue', 'Mahindra KUV100',*

*'Audi A4', 'Audi Q5', 'Hyundai i20', 'Maruti Suzuki Eeco',*

*'Maruti Suzuki Alto', 'Tata Nexon', 'Hyundai i10', 'Hyundai Verna',*

*'Skoda Laura', 'Maruti Suzuki Vitara Brezza', 'Nissan Sunny',*

*'Maruti Suzuki Wagon R', 'Audi Q3', 'Volkswagen Polo',*

*'Toyota Corolla Altis', 'Nissan Terrano', 'Honda Mobilio',*

*'Skoda Octavia', 'Datsun Redigo', 'Maruti Suzuki Baleno',*

*'Ford Fiesta', 'Ford Fiesta/Classic', 'Nissan Micra',*

*'Maruti Suzuki S-Presso', 'Hyundai Santa Fe', 'Renault Lodgy',*

*'Maruti Suzuki Alto 800', 'Toyota Fortuner', 'BMW 7-Series',*

*'Ford Figo Diesel', 'Maruti Swift', 'Maruti Eeco', 'Tata Tiago',*

*'Nissan Evalia', 'Maruti Alto LXi', 'Hyundai i20 Sportz',*

*'Toyota Innova Crysta', 'Hyundai Grand i10 Sportz', 'Tata Harrier',*

*'Maruti Wagon', 'Maruti Swift Dzire', 'Honda City i', 'Audi', '',*

*'Hyundai Grand i10 Magna', 'Hyundai Grand i10 Asta',*

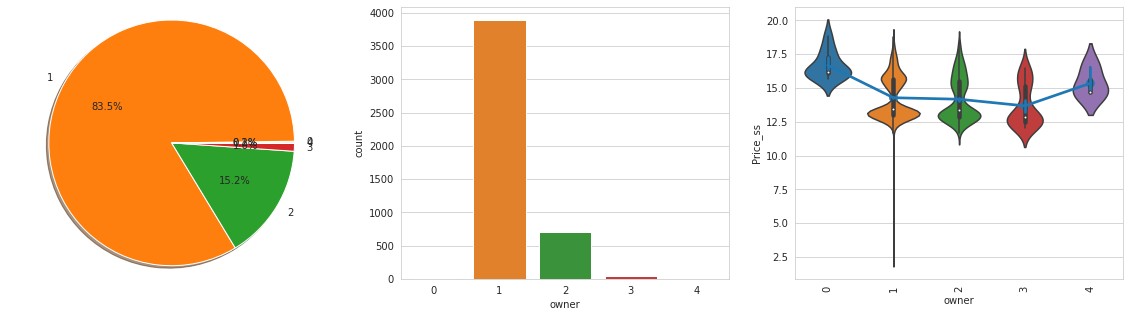
*'Maruti Zen Estilo', 'Mahindra', 'Mahindra Marazzo',*

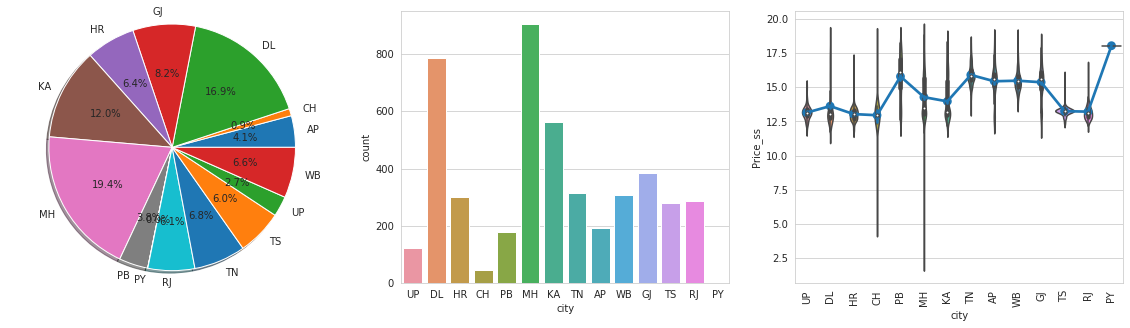
*'Maruti Celerio', 'Ford EcoSport', 'Hyundai i20 Magna', 'Hyundai',*

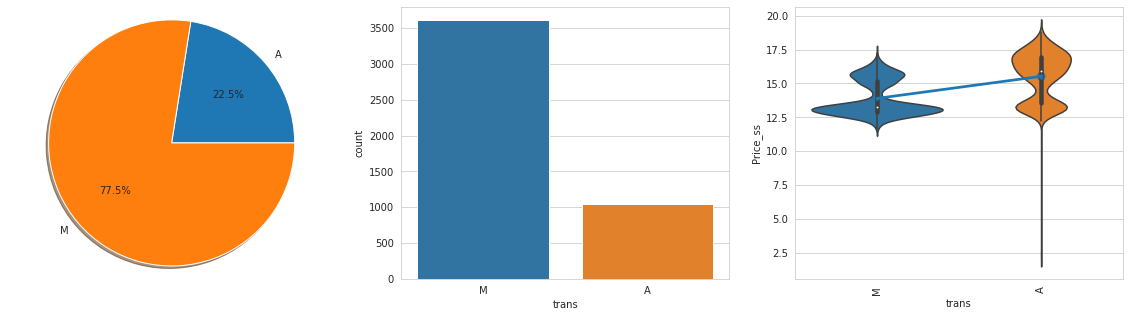
*'Hyundai Elantra', 'Volkswagen Vento Diesel Comfortline',*

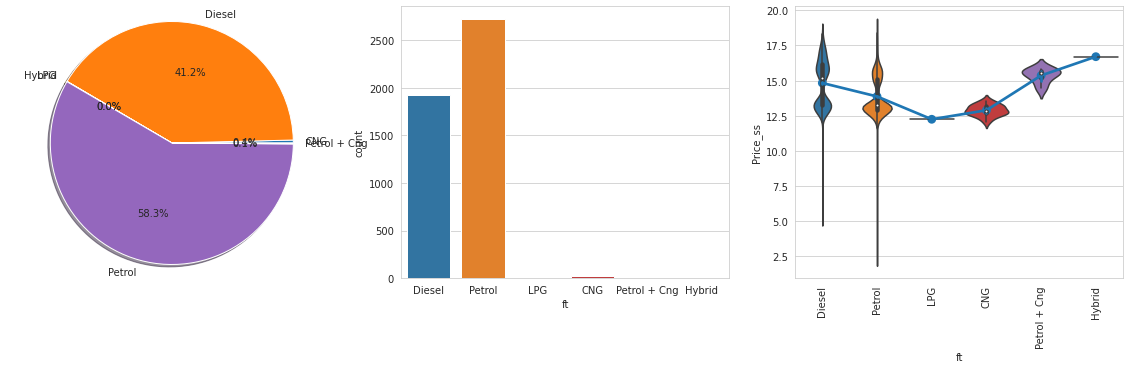
*'Maruti Alto', 'Maruti Swift VXi', 'Hyundai i10 Era'*

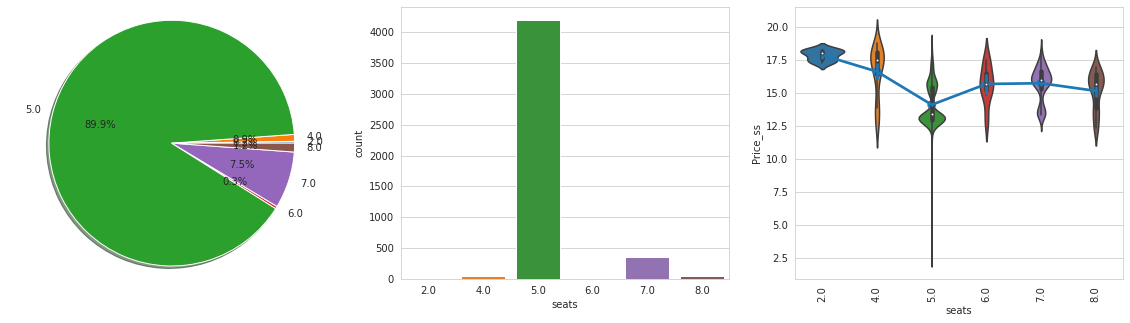
so after analusing it can be concluded that most car should not have high price so we will drop the rows as it will mislead the data.

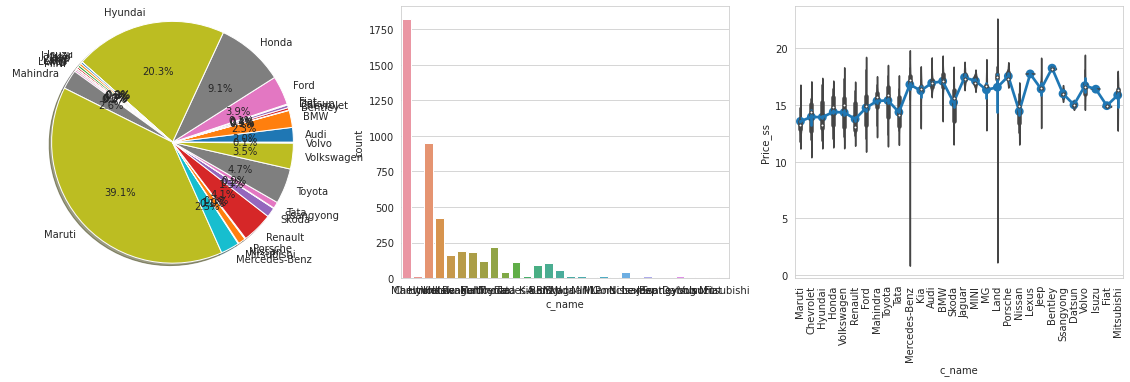






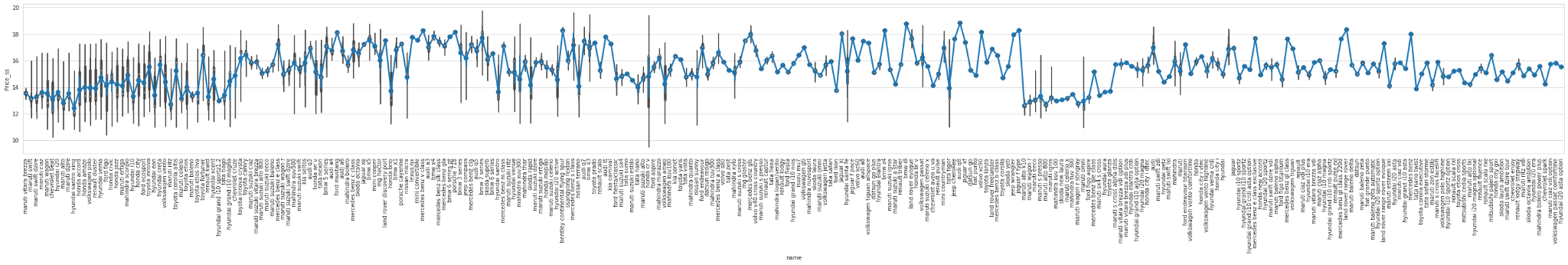






Observations:-

1. for the first owners, the price is higher than the rest.
2. The city has very good variance but some of the city has approximately same mean so they can be clustered together for better performance of the model.
3. The automatic transmission has a higher price than a manual transmission.
4. Fuel type has a very good variance in its distribution.
5. The number of seats has a little bit of variance but it can be classified into three e categories i.e Less than 5 equal to 5 and greater than 5.
6. The company name has a very good variance along with its distribution.



1. The name has a very large amount of giving data and it has a good variance to It is wise to use mean encoding.

Continuous variables are skewed

owner 2.190681

km 3.721166

year 0.801432

seats 2.413290

cc 2.647826

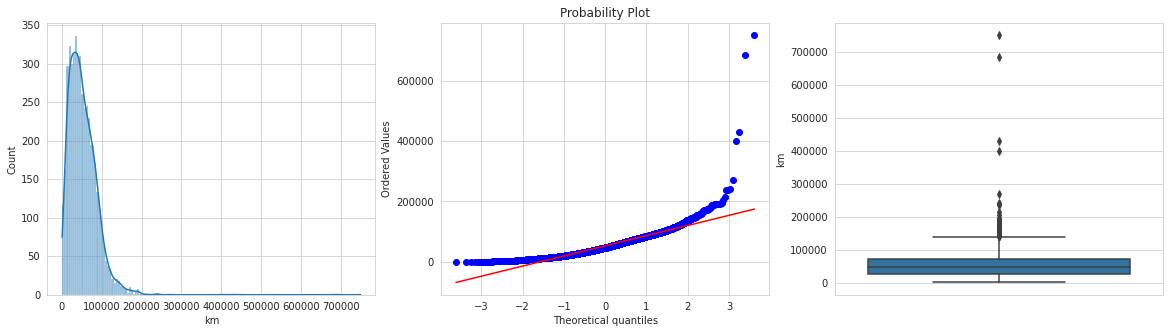
Price 4.480155

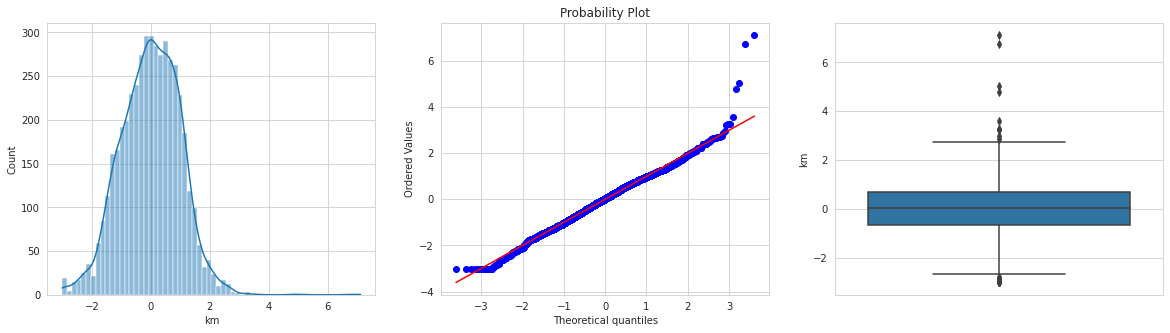
milage -0.245654

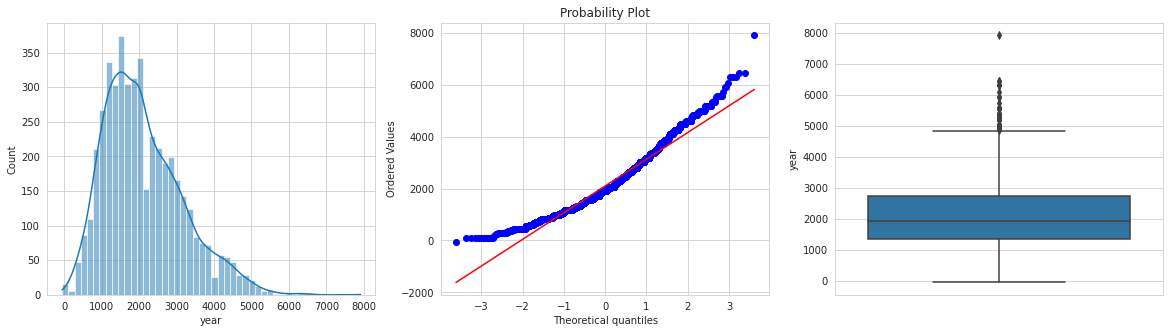
Price\_ss 0.585904

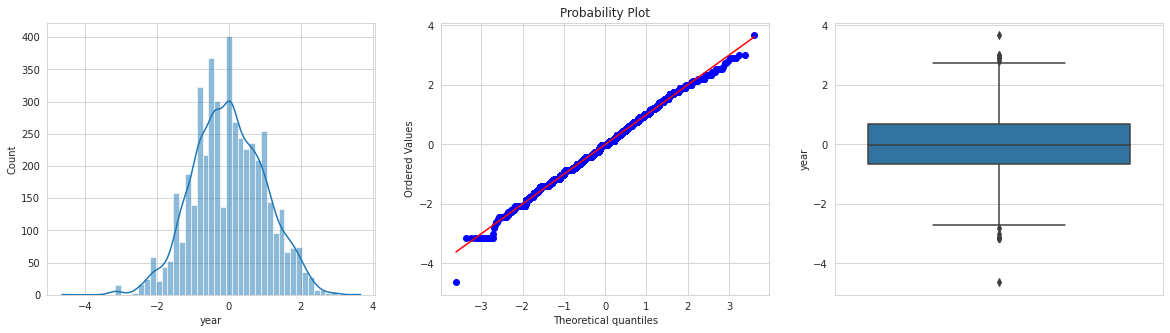
Owner and sweets will be considered as categorical variables. Price has already been log transform as Price\_ss

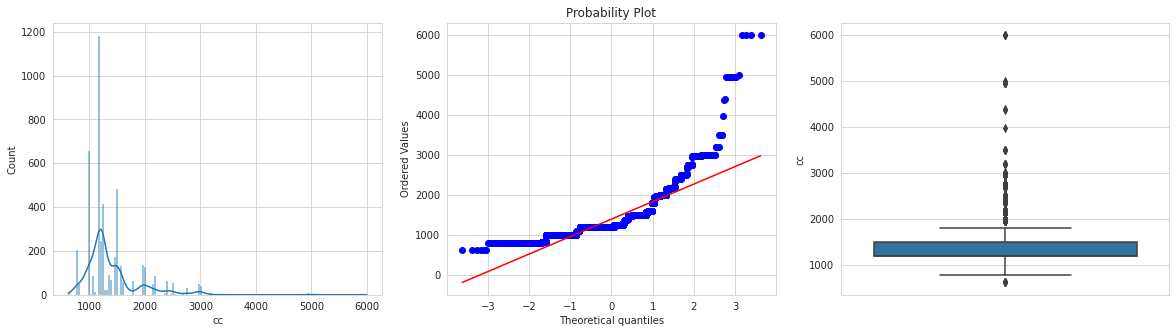
By using power transformation let's check skewness standardization and outliers.

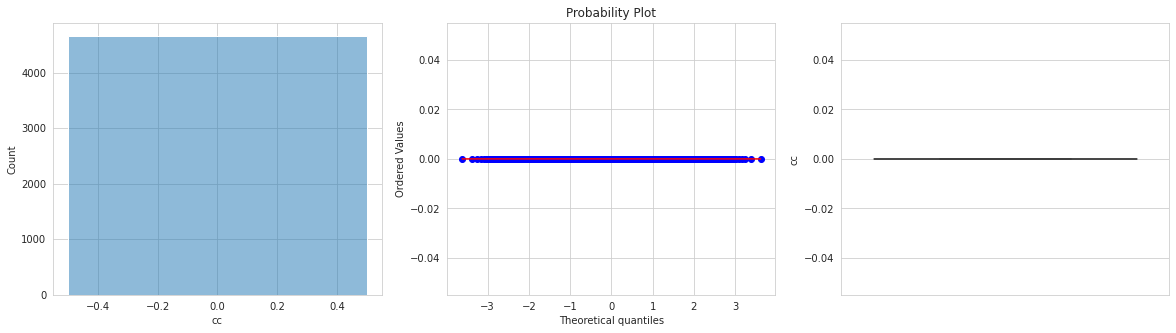


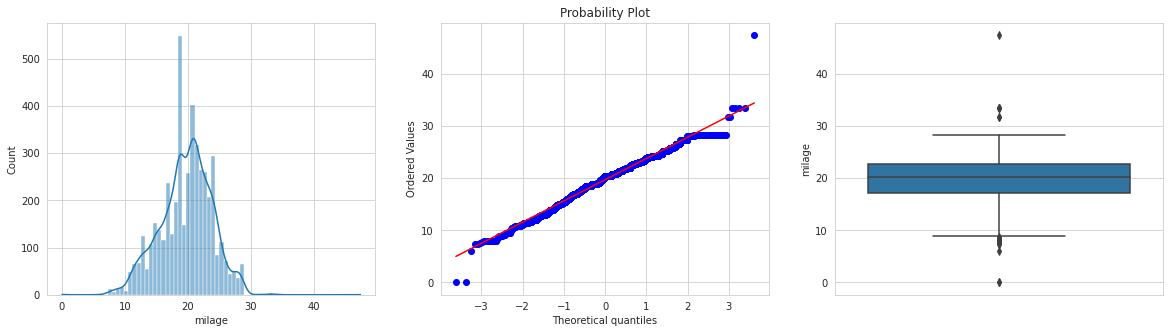


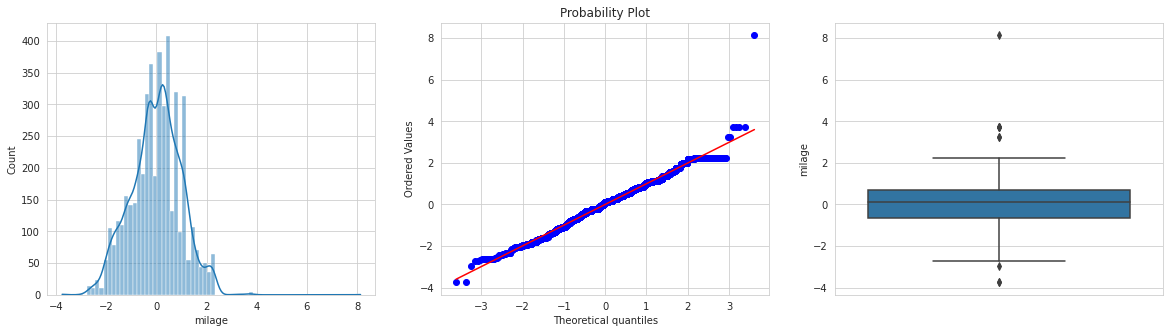




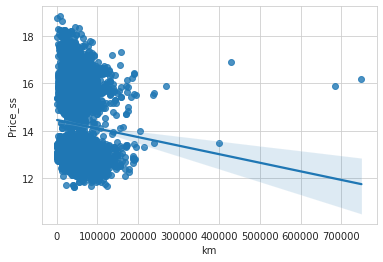


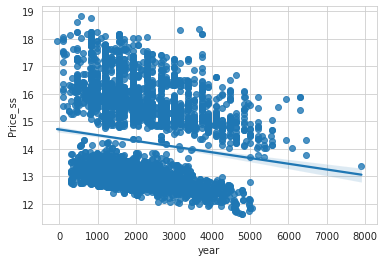


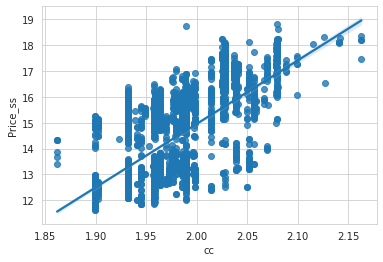


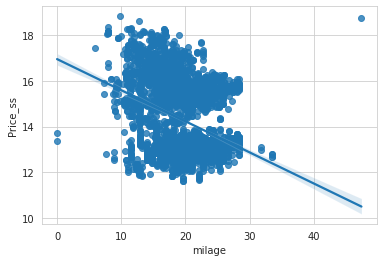


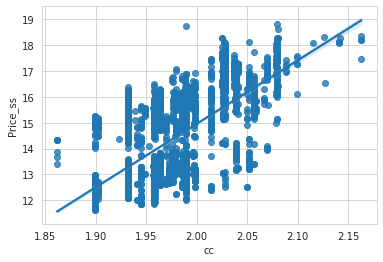
Maximum continuous variable outliers and tune has been taken care of using power transformation but engine capacity Has been flattened out to 0 so we will use log transformation to remove outliers in the engine capacity.



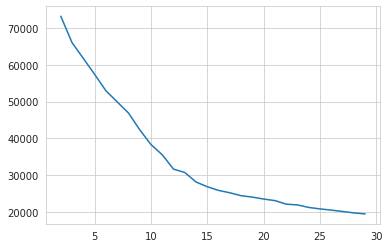




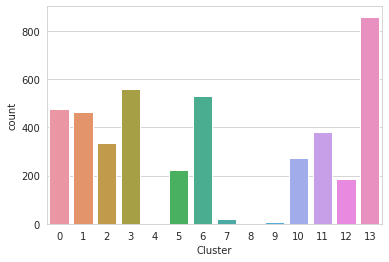




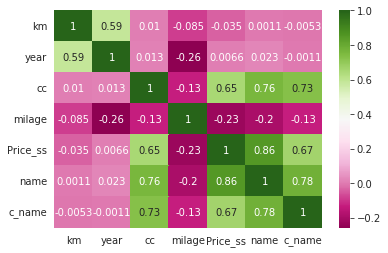
1. Price and kilometer are negatively related.
2. Price and here are negatively related and interestingly to clusters of form by plotting this graph.
3. Engine capacity is positively related to price
4. Price and mileage are negatively related
5. By analyzing these continuous variables this can be concluded that clustering these variables can generate a very good feature.



mean clustering for the entire data set



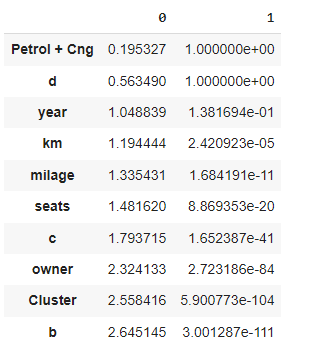
count plot col clustering



correlation plot of numerical data.

Name Engine capacity and company name has the highest effect on the price

ANOVA analysis:-



The top three columns are not related to output.

* Data Sources and their formats

The data source is different car sites such as car24, carwale, cardekho etc.

The format of data was different for scrapped data from the different web site so it’s a very time taking and tedious process to accumulate all the different formate of data into one format.

df.info()

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 owner 5066 non-null int64

1 km 5066 non-null float64

2 city 5065 non-null object

3 year 5066 non-null int64

4 trans 5037 non-null object

5 ft 5066 non-null object

6 seats 5065 non-null float64

7 c\_name 5066 non-null object

8 name 5066 non-null object

9 cc 4954 non-null float64

10 Price 5066 non-null float64

11 milage 4908 non-null float64

The object data type is of categorical type and float64 and int64 id of numerical type.

numerical type can be further classified into discrete(<7 unique feature), continuous, and date type(contain ‘year or yr’)

* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

The null value of engine capacity has been replaced according to the name

The remaining null value has been dropped.

All categorical variables are converted to ordered labels according to the price mean as inspected utilities have only one feature so we can delete it.

All continuous variables are skewed and they needed to be transformed so power transformation is applied to remove skewness.

standard scaler is used to scale down the variable from zero to one.

Clustering has been done to generate new features that mark the same kind of features together. Below are some features created for better prediction.

* Hardware and Software Requirements and Tools Used

Hardware- Computer

Software:- google collab

libraries:- pandas, matplotlib.pyplot, seaborn, scipy, sklearn, catboost etc.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The process involves cleaning the dataset(null value), separating variables into appropriate categories and dtype, transforming data to build an appropriate model, encoding categorical data, generating new data for better performance.

* Testing of Identified Approaches (Algorithms)

Linear Regression

Bayesian Regression

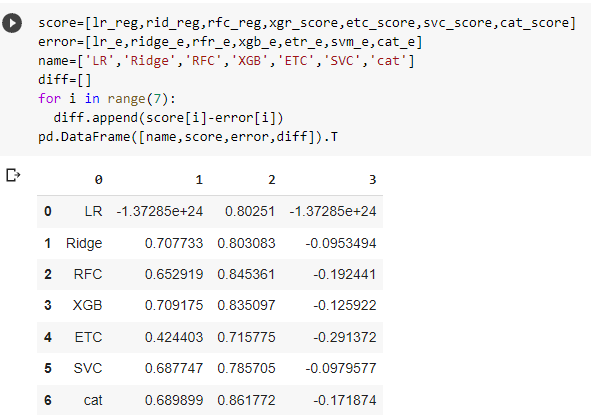
Random Forest Regression

XG Boost Regression

Extra Tree Regression

Cat boost Regression

* Run and Evaluate selected models



column1 represents the CV R2 score column 2 represents the R2 score of test data and column and column 3 represents the difference. we should look for low difference and high R2 score

* Key Metrics for success in solving problem under consideration

R2 Key metrics are used for evaluation as it denotes the fitness of prediction it one of the best scores for regression.

* Interpretation of the Results

From the extensive analysis, it can be inferred that:

1. car price has a 0.8 correlation with car price.
2. car price of owners 1 and four are higher than the rest.
3. The manual transmission has a lower price than automatic transmission.
4. LPG and CNG e v type car has lower price than the rest of the car
5. Company name has very high variance in its price range
6. Km driven is negatively correlated with the price of the car
7. Year use of ka is negatively correlated with the price of the car
8. The engine capacity of the car is very highly correlated with the price of the car
9. Mileage is negatively correlated with the price of the car

**CONCLUSION**

* car price has a 0.8 correlation with car price.
* Min price of owners 1 and four are higher than the rest.
* The manual transmission has a lower price than automatic transmission.
* LPG and CNG e v type car has lower price than the rest of the car
* Company name has very high variance in its price range
* Km driven is negatively correlated with the price of the car
* Year use of ka is negatively correlated with the price of the car
* The engine capacity of the car is very highly correlated with the price of the car
* Mileage is negatively correlated with the price of the car
* Learning Outcomes of the Study in respect of Data Science

Data collection, assembling into single formate and data cleaning almost took 80% of project time

* Limitations of this work and Scope for Future Work

Due to time constrain it was difficult to do more feature engineering. doing more feature engineering will give more insight into data. length of the training set was very small so collecting more datasets will result in much better results and correlation.