CAR PRICE PREDICTION

Submitted by:

AKASHDEEP SINGH MANRAL

Batch no. :1834

ACKNOWLEDGEMENT

The dataset was scraped using Instant data Scraper, from the website- Cardekho.

I want to thank my intern mentor miss- Swati Mahaseth for providing assistance in solving my queries, with her help and guidance I was able to complete my project successfully.

INTRODUCTION

Data sources and Formats-

We have scraped data of 5000 used cars from Cardekho website. The data set contains the data regarding cars such as- company , variant , model year , gear type , fuel type , no.of photos available in the site, distance covered , place , prices etc.

A-PROBLEM STATEMENT

In this project we tried to acces the features that leads to the price of the car and tried to make machine learning prediction model based on those features.

Data Info-

- Gsc_col-xs12 src – image source
- assured- website
- photoNumber src – No.of photos in the website
- ImageTransition – image source
- views – no. of views on the car
- imageTransition src 2 –
- model_year – year at which the car was first purchased
- company – company of the car
- model – model name of the company

- variant- variant of the model
- emitextCard – stands for EMI
- emitextCard 2- for @
- emitextaCard 3- EMI of the car
- prices price of the car.

Analytical Problem Framing

A- DATA CLEANING-

Data cleaning is an important step before proceeding for data visualization. In Data Cleaning the unwanted or the columns that are not providing any valuable information is eliminated.
After combining the scraped dataset of different places , a final dataframe(d) was formed.And from Dataset we eliminated certain unwanted columns-

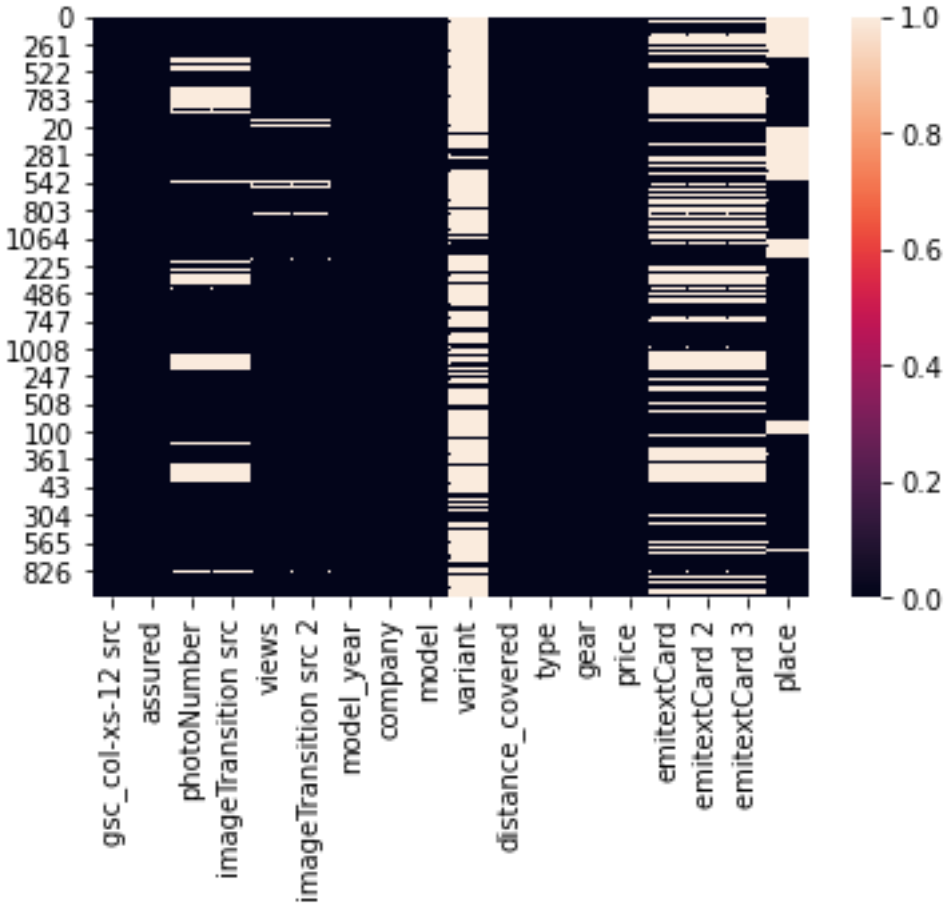
a- Gsc_col-xs12 src
b- Assured
c- ImageTransition
d- imageTransition src2
e- emitextCard
f- emitextCard2

B- EXPLORATORY DATA ANALYSIS(EDA)-

EDA is the step for craving out the valuable information from the given dataset using methods of inferential statistics.
Inferential Statistics involve univariate, bivariate and multivariate analysis using

countplots, histogram, piechart, scatterplots etc.
Descriptive Statistics involve description of data such as, mean,median,mode,std,of the data.

Results of EDA-
1-Heatmap was used to find out the null values in the dataset

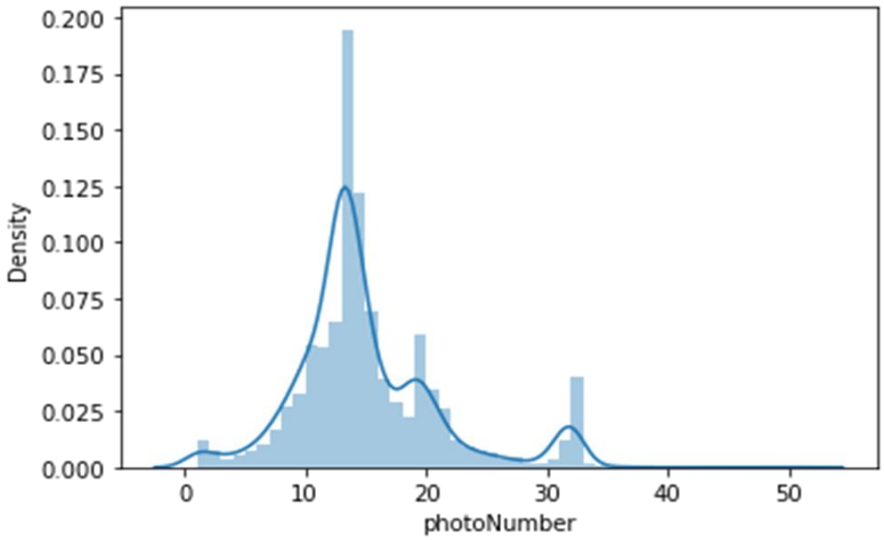


Observation-
a-We observed ther were lots of null values in our data set and we need to fix them.

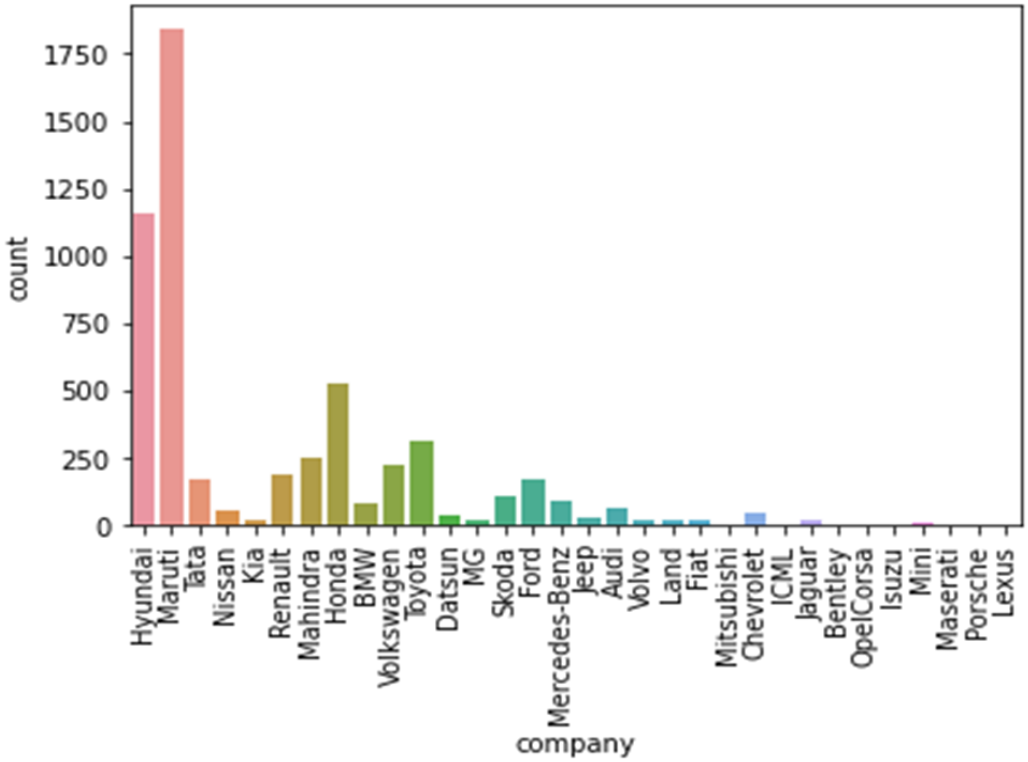
2-Distplot for column[Photonumber]

Observation-

a-Most of the cars have 10-15 photos available on the site.



3-Countplot- or column[company]



Observation-

a-Most cars are of company Maruti followed by Hyundai,Honda,Toyota,Mahindra.

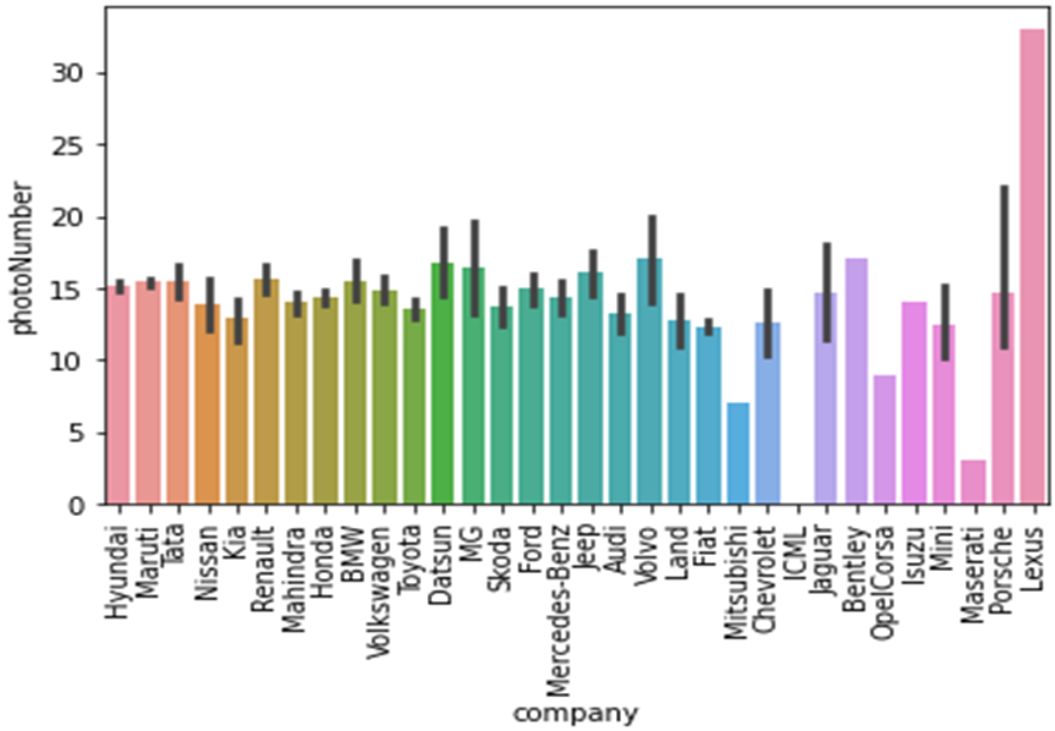
4-Barplot- Relationship between columns [company and photoNumber]

Observation-

a- Lexus has highest number of photos

b- ICML has the least number of photos i.e 0

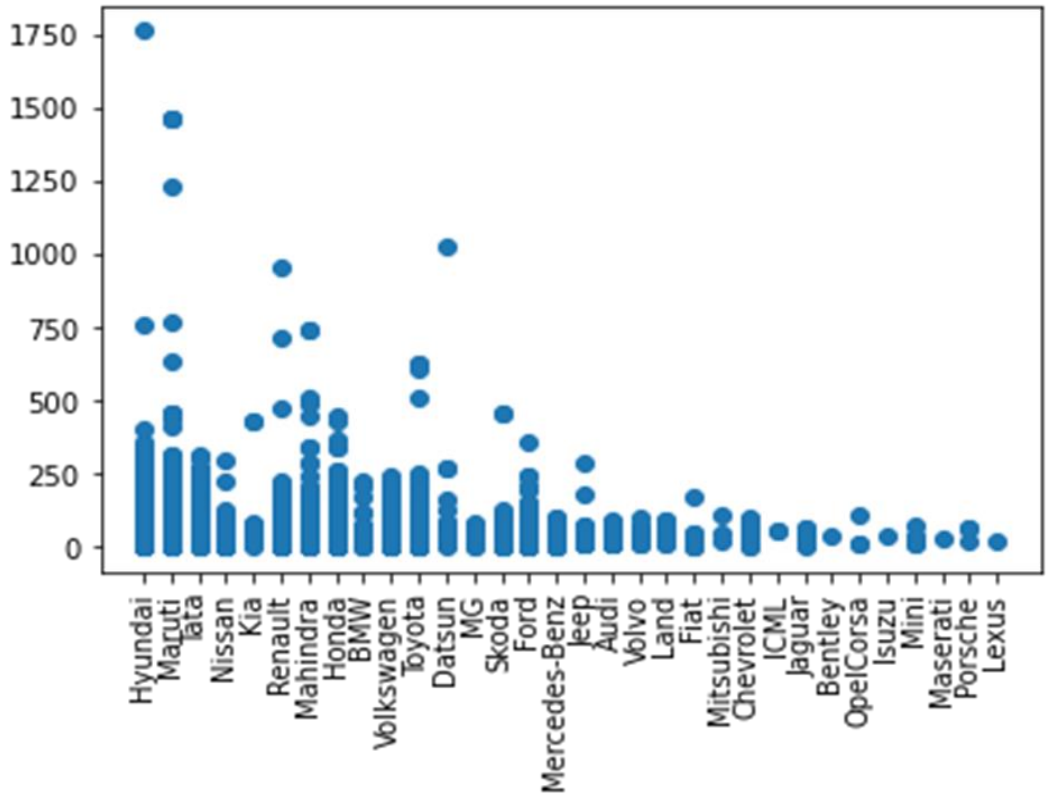
c- most of the cars have around 13-15 photos at least.



5-Scatterplot- Relationship between Company and Views column

Observation-

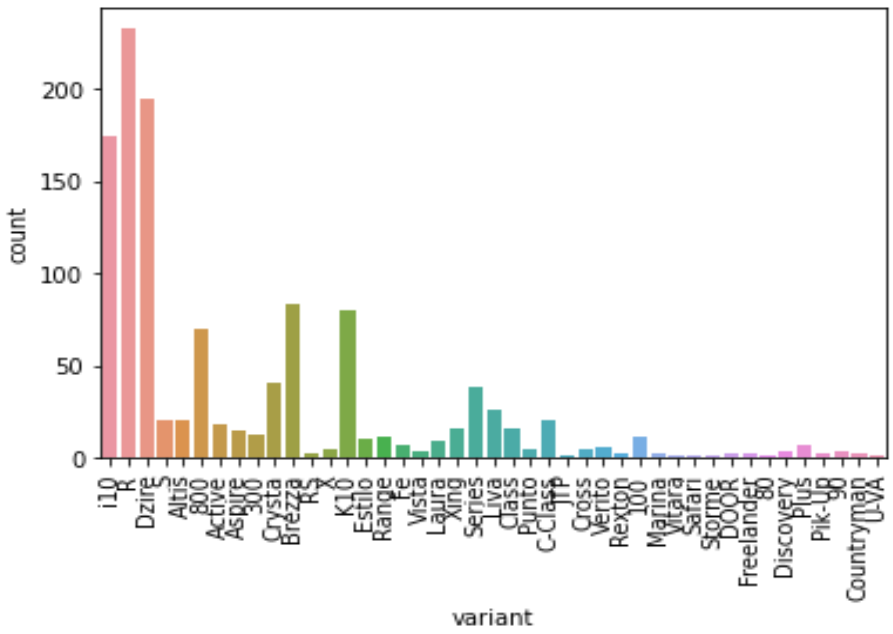
a- Hyundai has highest number of views
followed by Maruti



6-Countplot- Variant of cars

Observation-

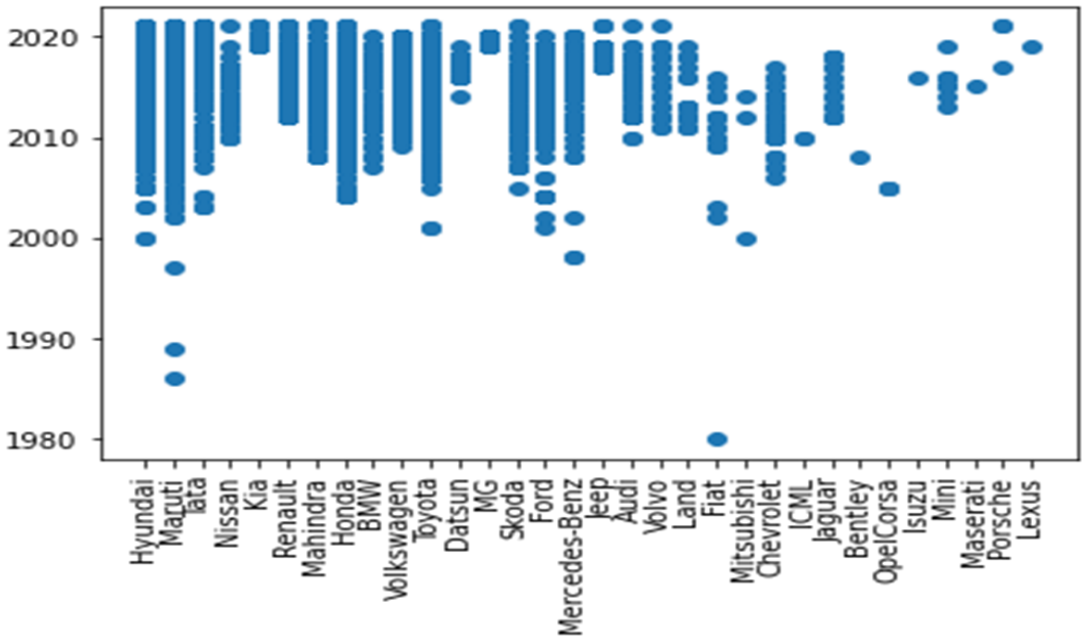
a- most number of cars - Wagon R, followed by- Swift Dzire and Hyundai 10



7-Scatterplot- Relationship between Company and model_year

Observation-

a-most of the car models are of the year 2012-2018



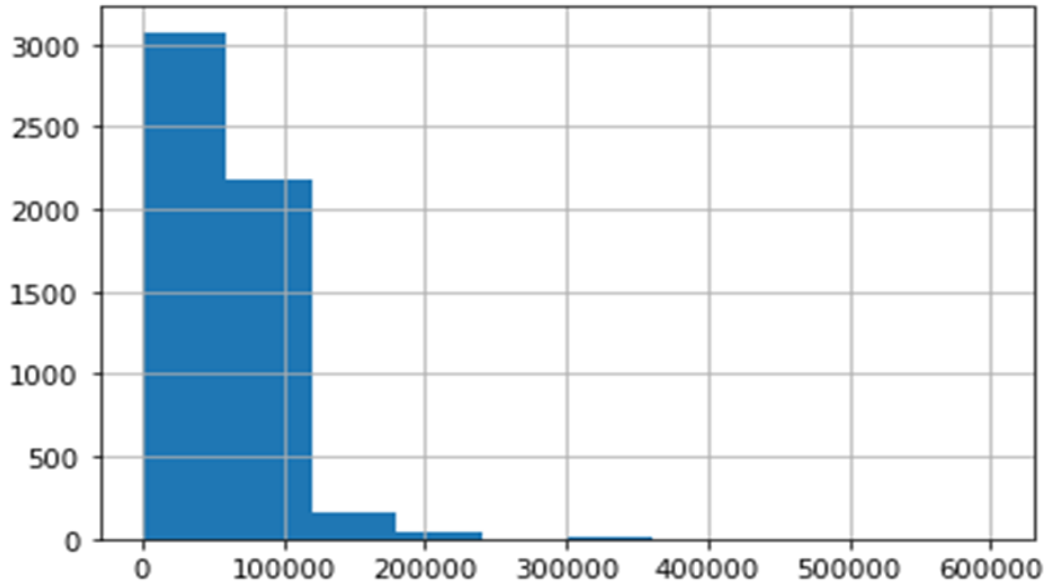
8-Histogram- DistanceCovered

Observation-

a-1-majority of cars have covered the distance upto : 0-60000kms

b-followed by cars who have covered the distance upto : 60000-120000kms

c-some cars have even covered the distance upto :350000kms



9-countplot –column[type]

observation-

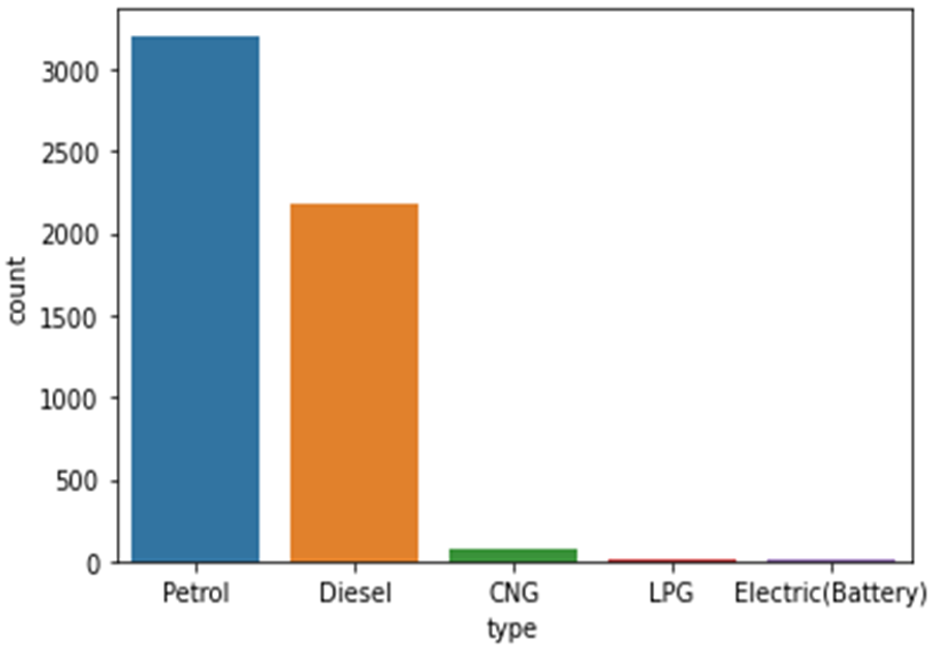
a-petrol-3207 cars

b-diesel-2177cars

c-CNG-75cars

d-LPG-10cars

e-Battery-6cars

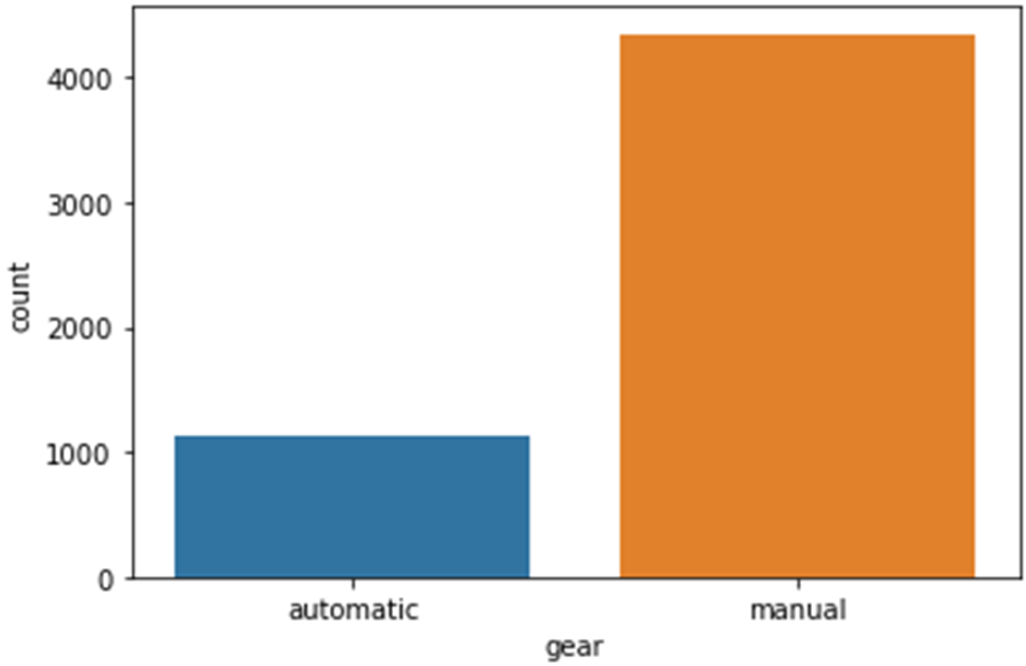


10- Countplot- on column[gear]

Observation-

a-majority of cars have manual gears- 4352

b-automatic gear cars- 1123



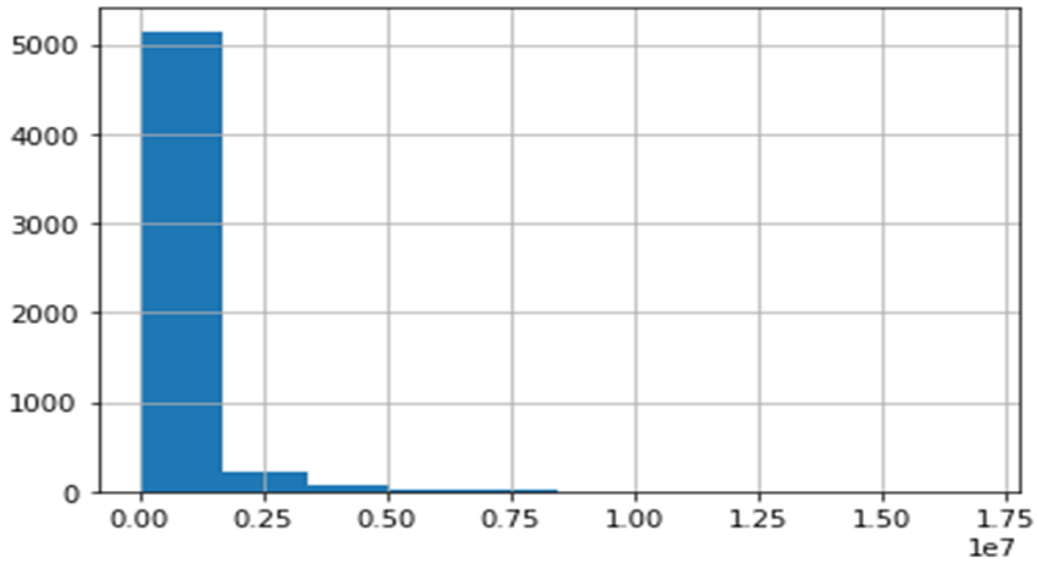
11- Histogram – on column[prices]

Observation-

a-Most car prices falls in between: (0-0.17le) (0.17le= 0.17^10000000=Rs1700000)

b-some cars also have price ranging between:(0.50le-0.80le) or (Rs5000000-RS8000000)

12-



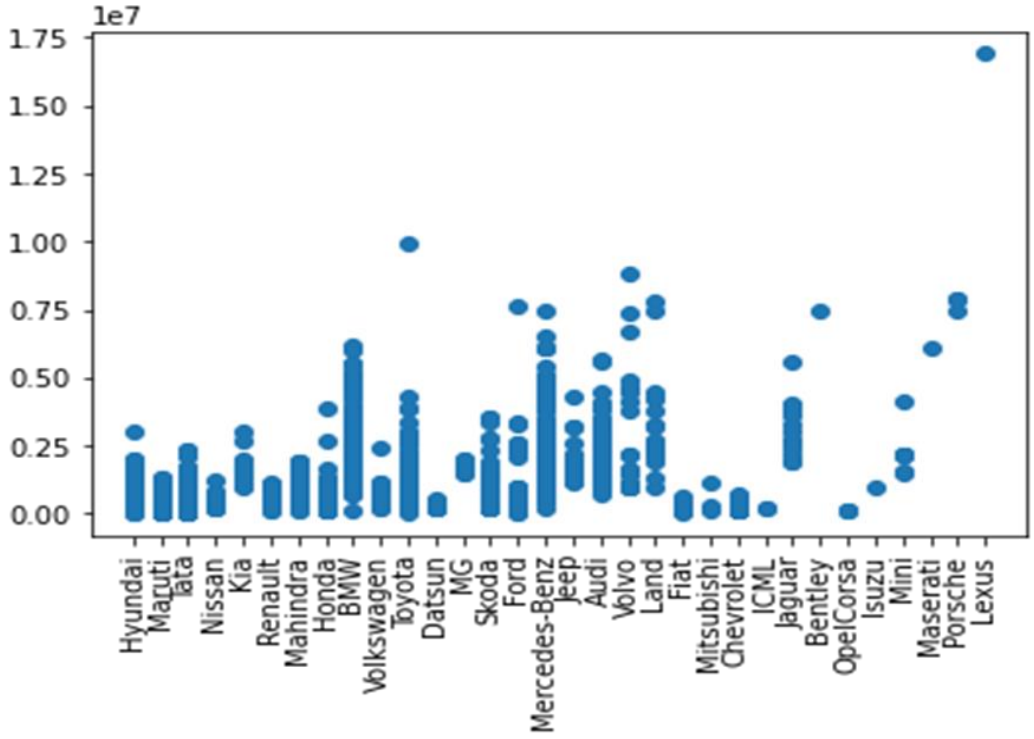
12- Scatter- Relationship between Company and Price.

Observation-

a-Lexus car have the highest price

b-Hyundai,Maruti,Tata, Mahindra,Honda cars have moderate prices

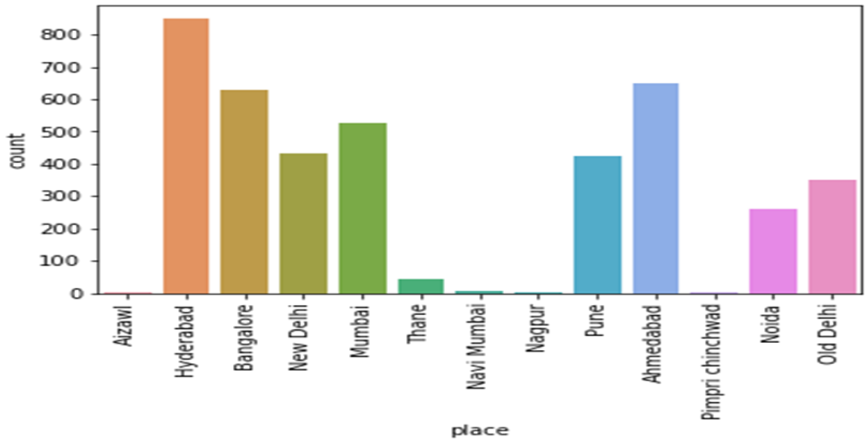
c-Toyota have the 2nd highest car price



13- Countplot- Place

Observation-

a- the dataset contain cars mostly from Hyderabad region followed by Ahemdabad, Bangalore, Mumbai, NewDelhi,OldDelhi, Noida respectively.



C- DATA PRE-PROCESSING

Data pre-processing of Feature Scaling involves-

a-Filling NaN values- columns having Nan values are filled.

In categorical column NaN values are filled with mode value of the column

In numerical column NaN values re filled with either mean or median values.

b-Encoding the DataFrame- Encoding is done to convert categorical values to numerical format,so that machine could learn it better.

Types of Encoders-

1-Ordinal Encoder- used in feature variables

2-One Hot Encoder- used in Target variables when the value are-yes or no , true or false type, havin only two outcomes.

3-Label Encoder- used when the target column have more than two outcomes.

c-Outliers removal- Outliers are the values that are too far from the mean value,either in positive or negative direction.

1-Outlier Detection- it is done through zscore method

Zscore- Z score is an important concept in statistics. Z score is also called standard

score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.

Z score = (x -mean) / std. deviation

If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.
Such data points are removed from the dataset , so that it does not affect the algorithm of Machine Learning models.

d-Selecting x and y variable, x variable includes all the input feature , while y variable includes only target column.

e-Scaling- Scaling is done to equalise the ranges of values of different type of data in the data set.
Types of Scaling techniques:
a-Standardization- In this we use StandardScaler() for scaling the data. This scaler makes the mean value to zero and std as 1.

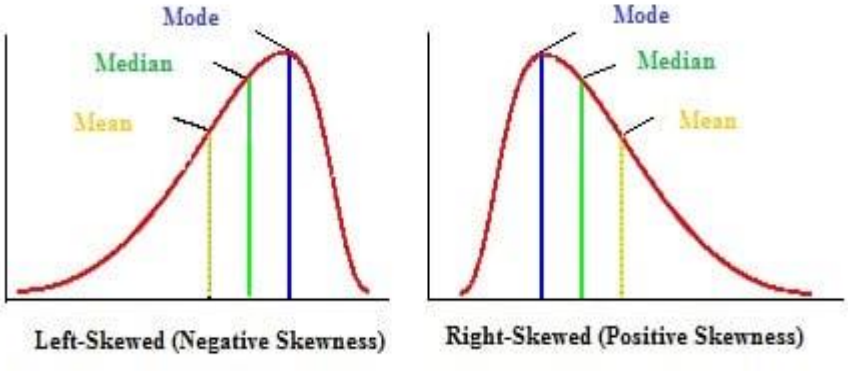
b-Normalization- In this we uses MinMaxScaler() for scaling.This technique ranges the data to (-1 - +1) .
Both the techniques are efficient. We do Scaling only on Input variables.
In our Dataset we made use of StandardScaler.

f- Skewness removal- When a data is normally aligned,i.e- it shows a bell shaped curve,thus no skewness is present and mean=median=mode.

Types of skewness:

a-Positive skewness- When most of the data is concentrated on right hand side.
Mean>median>mode.

b- Negative Skeweness- When most of the data is concentrated on left hand side.
Mode>median>mean.



We removed skewness using log and sqrt method.Skewness is removed so that it does not affect the machine learning algorithm.

Model/s Development and Evaluation

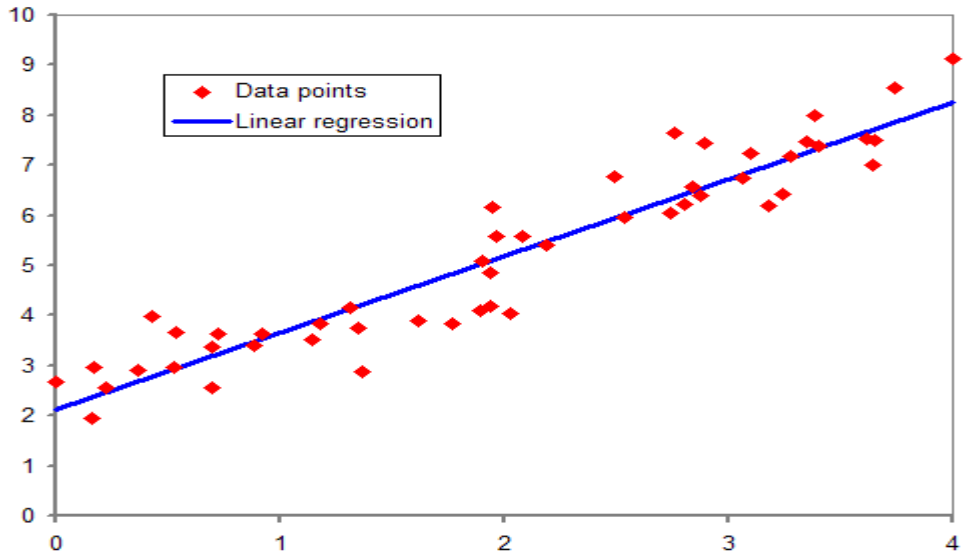
1-MODEL SELECTION-

After selecting the x and y variable,we then perform the train_test_split and and check the training and testing in various models.

train_test_split: It splits the variables and sends them ,according to random_state, for training and testing.In our model we choose the random_state as 30,it means 70% will go for trainingand 30% for testing.

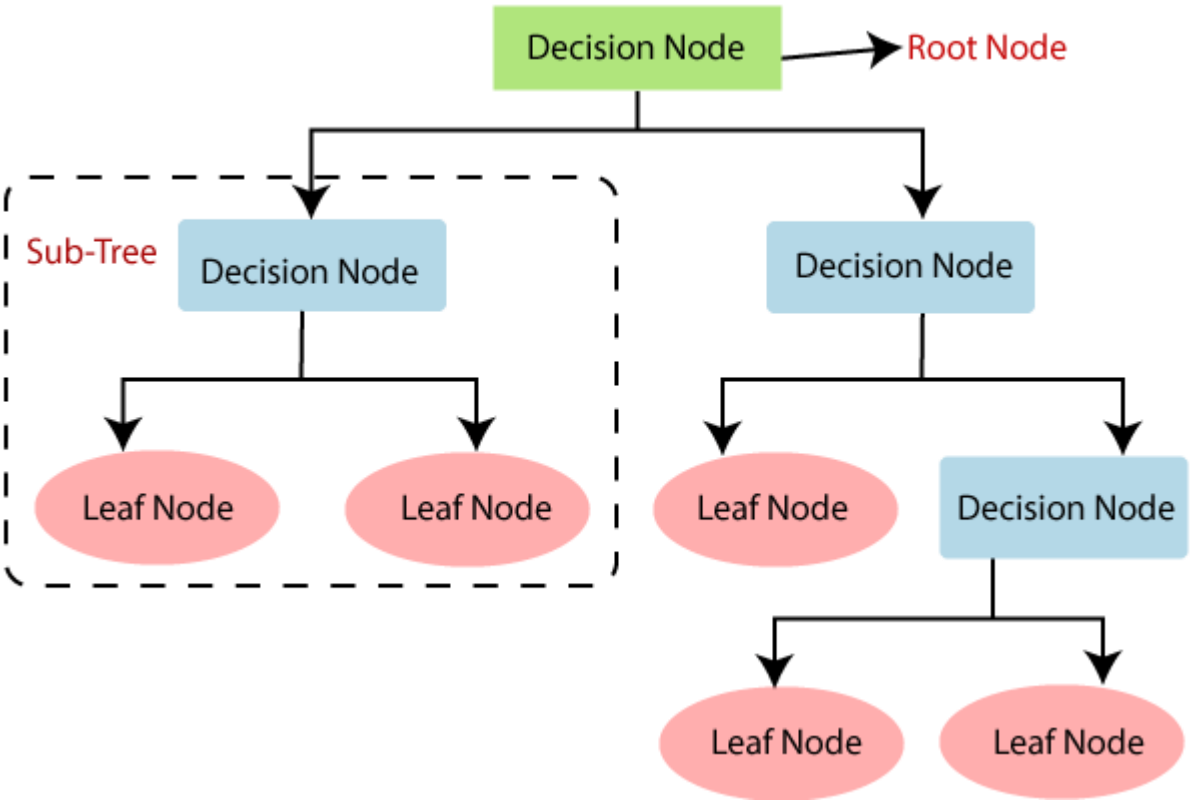
As the problem is a Regressor problem so we called various regressor models:

1-Linear Regressor- Linear regression is used for finding linear relationship between target and one or more predictors.The idea is to obtain the best fit line between the feature and target variable.



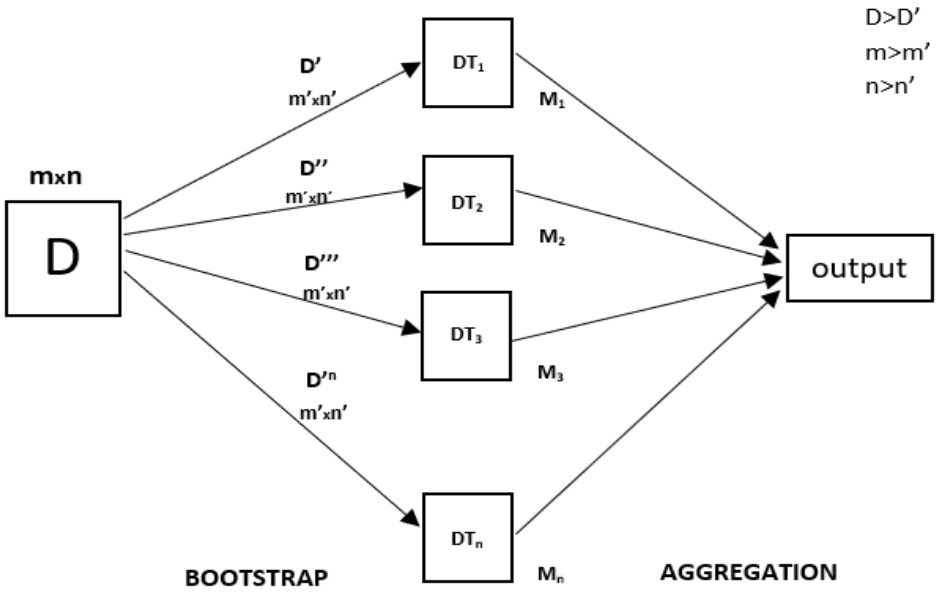
2-DecisionTree Regressor- It is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility.

The branches represent the result of nodes.Nodes either have conditions(decision) or end /leaf nodes(result).



3-KNeighborsRegressor- The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.

4-RandomForestRegressor- A RandomForest is an ensemble technique capable of performing regression with the help of multiple decision trees.



5-AdaBoostRegressor- An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.

6-GradientBoostingRegressor- GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

After this we checked the accuracy score of model in cross_val_score.
From comparing the scores of the model we found out the best suitable model was- **GradientBoostingRegressor.**

2-HYPER TUNING USING GRID SEARCH CV

After the selection of model hypertuning of the model using GridSearchCV is done to bring out te best performance of the model.
The best parameters for the model are selected using GridSearcgCV and are given in th model for the best result.In our model- GradientBoostingRegressor the best parameters selected were- {'loss': 'huber', 'max_depth': 8, 'max_features': 'auto'}.

3-METRICS

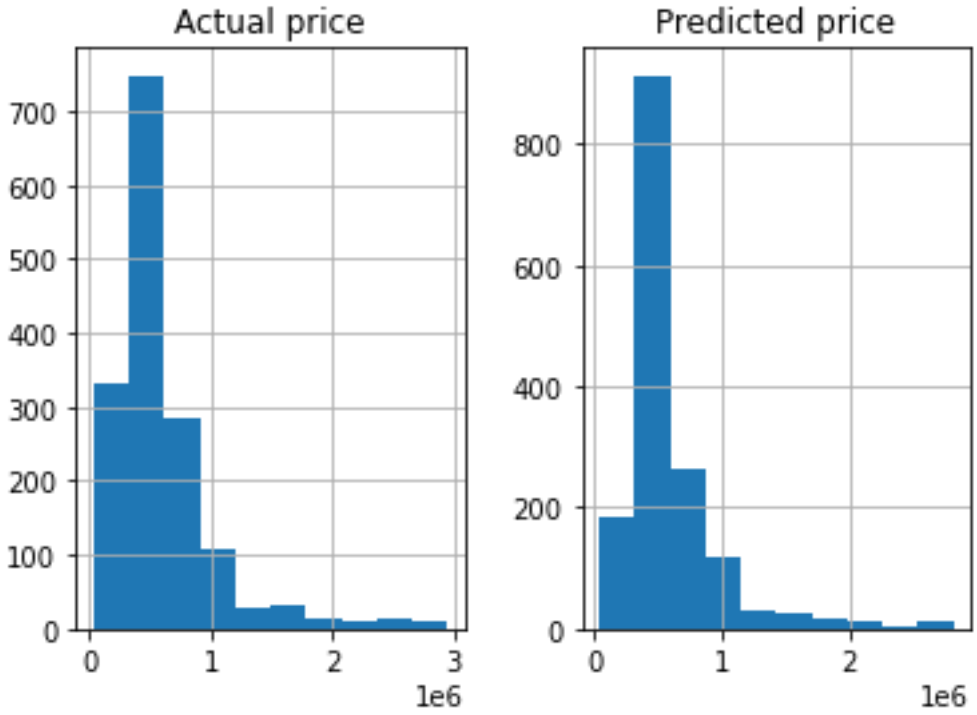
Metrics used in the regression models:
a-Mean_squared_error- It tells us how close a regression line to set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them.
The lower the value , the better the forecast.

b- Mean_abolute_error- Mean Absolute Error calculates the average difference between the calculated values and actual values. The lower the value , the better the forecast.

c- Root_mean_squared_error- Root mean squared error (RMSE) is the square root of the mean of the square of all of the error.

4-CONCLUSION

Our best model-GradientBoostingRegressor worked out with 89% accuracy after hypertuning.



5-LIMITATIONS OF THE WORK

The model works with 89% accuracy so all the predicted values could not be taken as true values. The predicted values do have a margin of failure of 10%.

THANK YOU