FLIP ROBO

CAR PRICE PREDICTION

Submitted by:

ANJAY KRISHNA T U

ACKNOWLEDGMENT

The websites Cars24 and Car Dhekho helped in giving information about the used cars and all the details need to complete this project. The data collected from this sites are used for building a machine learning model which predict the price of a vehicle.

INTRODUCTION

- Business Problem Framing

  Due to the impact of covid 19 there is a lot of changes in the car market. Some of the cars are high in demand and some others are not. The price of the cars those are high demands increased in this time. This project is to collect data of used cars from different sites and predict the price of the car using the data collected from the different sites.

- Motivation for the Problem Undertaken

  The change in the car market due to covid 19 is a problem for clients who sells used cars. By collecting data from different sites and building a model for predicting the price will help them in selling cars.

Analytical Problem Framing

• Data Sources and their formats

The data are collected from cars24 and car dhekho sites. The collected data contains the details about cars like brand, model, variant, manufacturing year, driven kilometres, fuel type, location and price. The collected dataset has 5575 rows 9 columns and the data types are integer and object data types.

• Data Preprocessing Done

The dataset has no null values, in the columns model and brand same name were showing differently all these corrected by replacing with correct name. In the location also Delhi and Bengaluru replaced as New Delhi and Bangalore. Driven kilometres and Price include ‘,’ & ’kms’ they are removed and also 'Unnamed: 0'. Outliers are present in Manufacturing year, Driven kilometres and price, by using z-score method the outliers are removed.

• Data Inputs- Logic- Output Relationships

The input variables in the dataset are Brand, Model, Variant, Manufacturing year, Driven kilometers, Fuel and Location. The output variable is the Price. There is a positive correlation between price and manufacturing year. All other inputs are negative correlated to the output variable price.

• Hardware and Software Requirements and Tools Used

Selenium Webdriver – For scraping the data

Anaconda jupyter Notebook – For Programming

Pandas, Seaborn, Matplotlib – For Visualization

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The analysis shows that there is a positive linear relationship between the price and manufacturing year. So, the manufacturing year is playing an important role in predicting the price of the car.

- Testing of Identified Approaches (Algorithms)

  For training and testing five algorithms are used here.

  1.Lasso Regression

  2.Ridge Regression

  3.Decision Tree Regressor

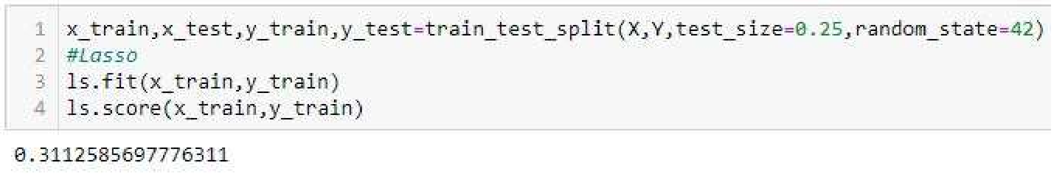
  4.Random Forest Regressor

  5.Gradient Boosting Regressor

- Run and Evaluate selected models

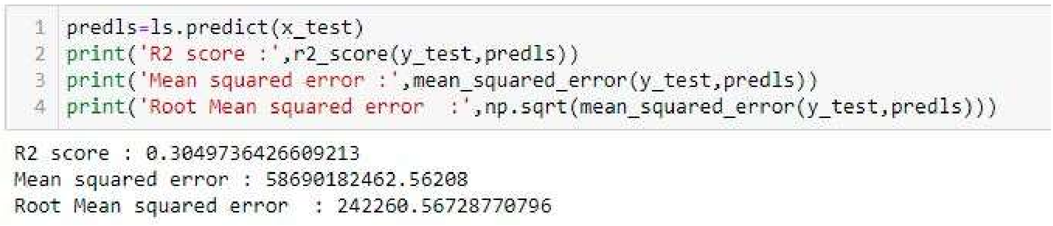
  1.Lasso Regression

  Lasso Regression is a supervised machine learning algorithm used for selecting the best features and regularizing the model for better performance. The R2 score of the model is 30.49%.



```
x_train,x_test,y_train,y_test=train_test_split(X,Y,test_size=0.25,random_state=42)
#Lasso
ls.fit(x_train,y_train)
ls.score(x_train,y_train)
```

```
0.3112585697776311
```

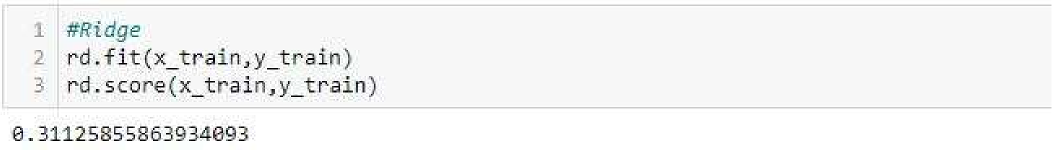


```
predls=ls.predict(x_test)
print('R2 score :',r2_score(y_test,predls))
print('Mean squared error :',mean_squared_error(y_test,predls))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predls)))
```

```
R2 score : 0.3049736426609213
Mean squared error : 58690182462.56208
Root Mean squared error  : 242260.56728770796
```

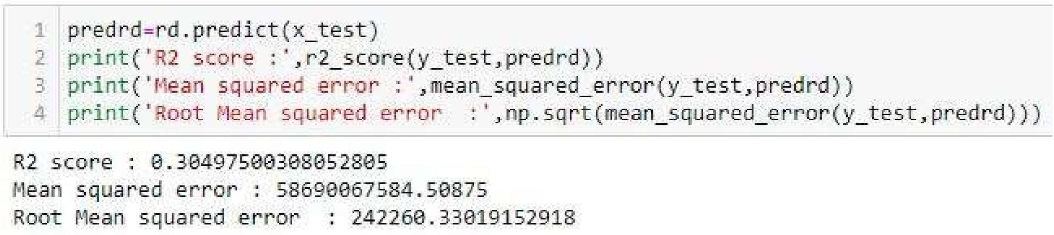
2.Ridge Regression

Ridge regression is a supervised machine learning algorithm.



```
#Ridge
rd.fit(x_train,y_train)
rd.score(x_train,y_train)
```

```
0.31125855863934093
```



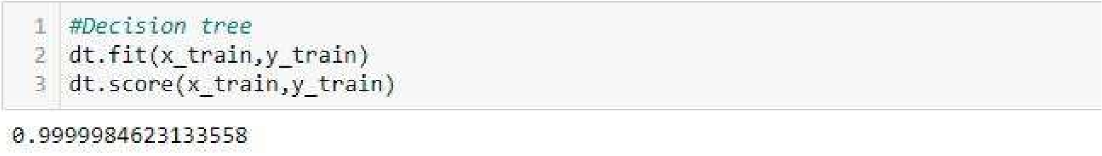
```
predrd=rd.predict(x_test)
print('R2 score :',r2_score(y_test,predrd))
print('Mean squared error :',mean_squared_error(y_test,predrd))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predrd)))
```

```
R2 score : 0.30497500308052805
Mean squared error : 58690067584.50875
Root Mean squared error  : 242260.33019152918
```

3.Decision Tree Regressor

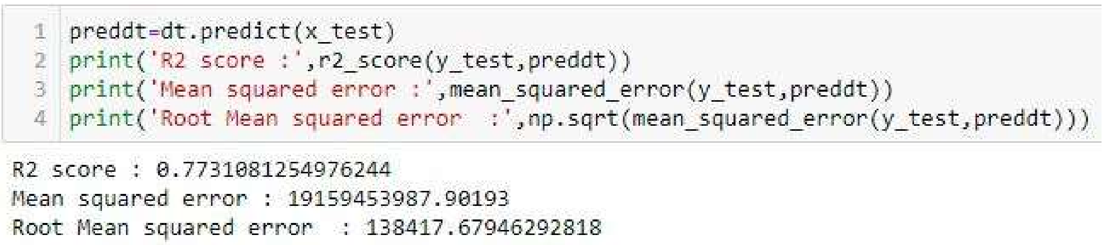
Decision tree regressor is supervised machine learning algorithm used for both classification and regression problem.

The R2 score of the model is 77.31%.



```
#Decision tree
dt.fit(x_train,y_train)
dt.score(x_train,y_train)
```

```
0.9999984623133558
```

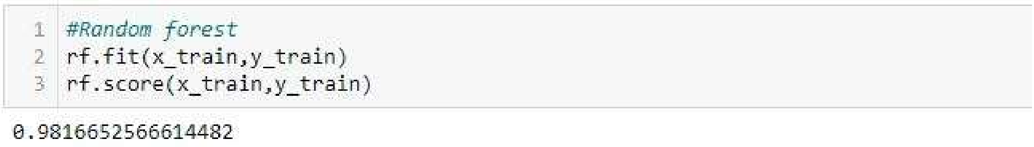


```
preddt=dt.predict(x_test)
print('R2 score :',r2_score(y_test,preddt))
print('Mean squared error :',mean_squared_error(y_test,preddt))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,preddt)))
```

```
R2 score : 0.7731081254976244
Mean squared error : 19159453987.90193
Root Mean squared error  : 138417.67946292818
```

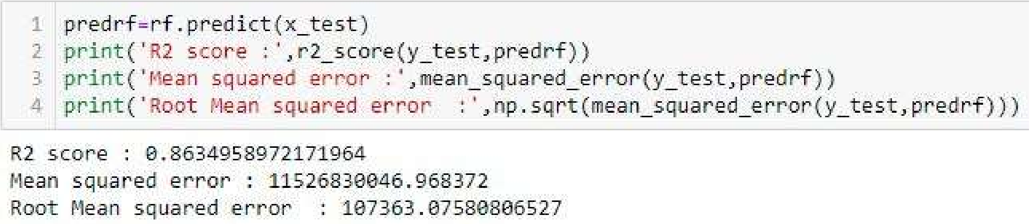
4.Random Forest Regressor

Random forest regressor is supervised machine learning algorithm uses an ensemble learning method for regression. The R2 score of the model is 86.34%.



```
#Random forest
rf.fit(x_train,y_train)
rf.score(x_train,y_train)
```

```
0.9816652566614482
```

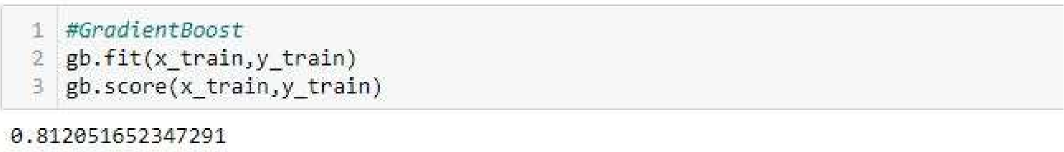


```
predrf=rf.predict(x_test)
print('R2 score :',r2_score(y_test,predrf))
print('Mean squared error :',mean_squared_error(y_test,predrf))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predrf)))
```

```
R2 score : 0.8634958972171964
Mean squared error : 11526830046.968372
Root Mean squared error  : 107363.07580806527
```

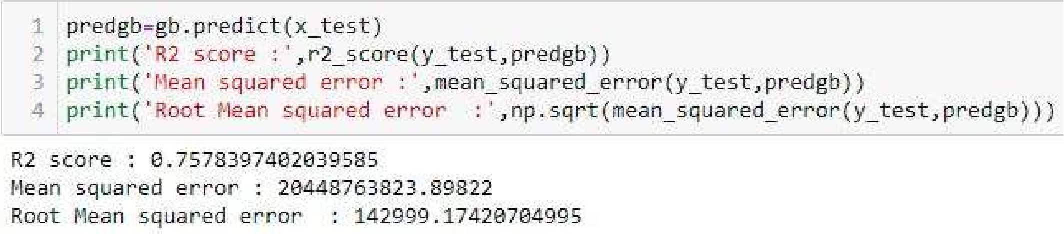
5.Gradient Boosting Regressor

Gradient boosting is an ensemble method used for both classification and regression problem. Weak models are combined to give better result in this algorithm. The R2 score of the model is 75.78%.



```
#GradientBoost
gb.fit(x_train,y_train)
gb.score(x_train,y_train)
```

```
0.812051652347291
```

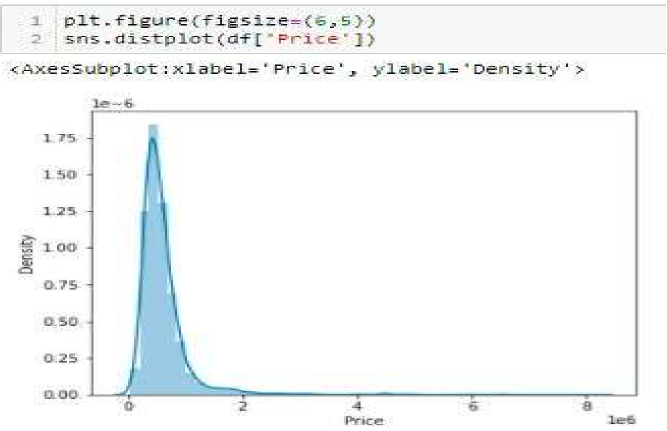


```
predgb=gb.predict(x_test)
print('R2 score :',r2_score(y_test,predgb))
print('Mean squared error :',mean_squared_error(y_test,predgb))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predgb)))
```

```
R2 score : 0.7578397402039585
Mean squared error : 20448763823.89822
Root Mean squared error  : 142999.17420704995
```

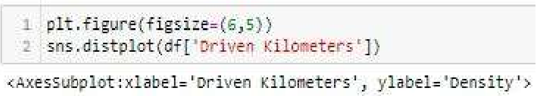
- Visualizations

The Distribution plot shows that the data distribution is good in Driven kilometers, Manufacturing year and Price.



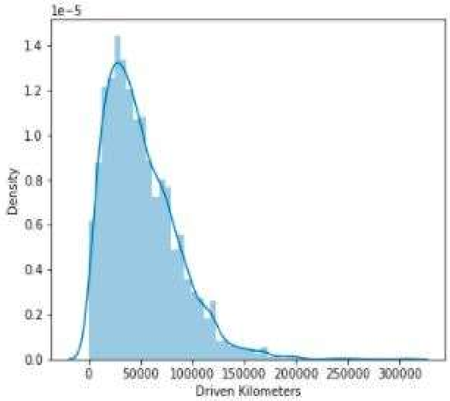
```
plt.figure(figsize=(6,5))
sns.distplot(df['Price'])
```

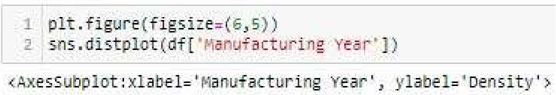
```
<AxesSubplot:xlabel='Price', ylabel='Density'>
```



```
plt.figure(figsize=(6,5))
sns.distplot(df['Driven Kilometers'])
```

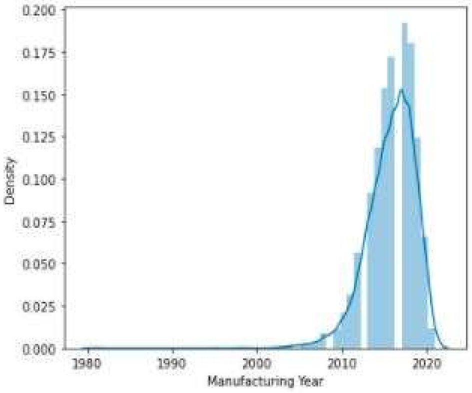
```
<AxesSubplot:xlabel='Driven Kilometers', ylabel='Density'>
```



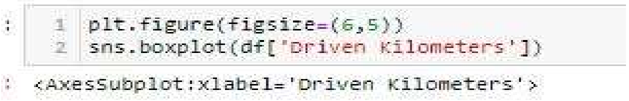


```
plt.figure(figsize=(6,5))
sns.distplot(df['Manufacturing Year'])
```

```
<AxesSubplot:xlabel='Manufacturing Year', ylabel='Density'>
```

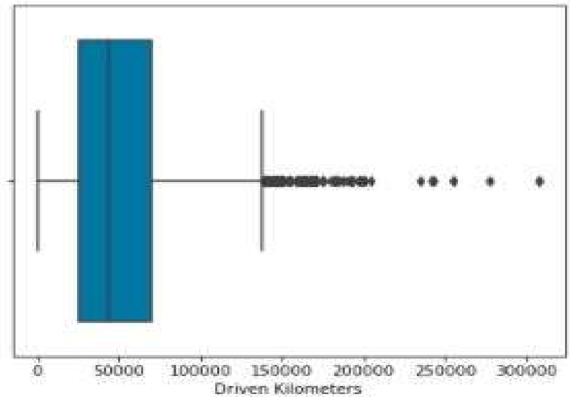


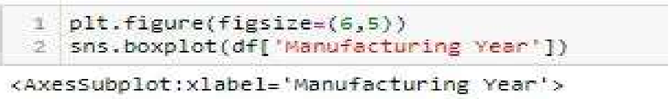
The boxplot shows that outliers are present in Driven kilometers, Manufacturing year and Price.



```
plt.figure(figsize=(6,5))
sns.boxplot(df['Driven Kilometers'])
```

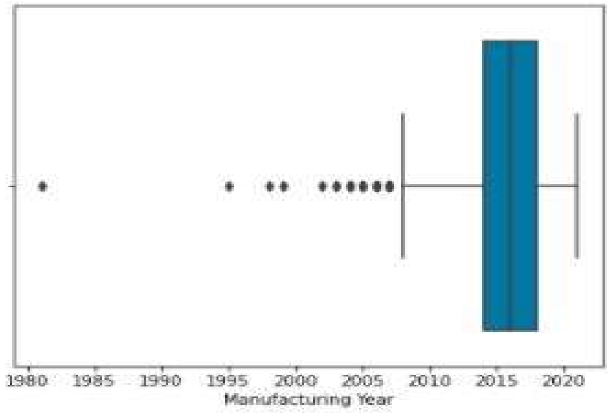
```
<AxesSubplot:xlabel='Driven Kilometers'>
```

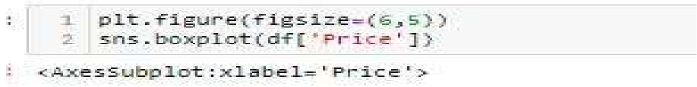




```
plt.figure(figsize=(6,5))
sns.boxplot(df['Manufacturing Year'])
```

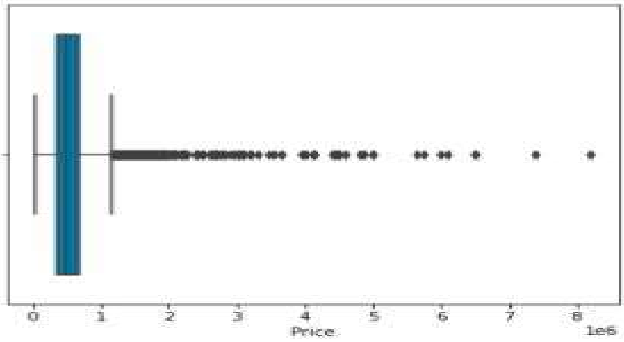
```
<AxesSubplot:xlabel='Manufacturing Year'>
```



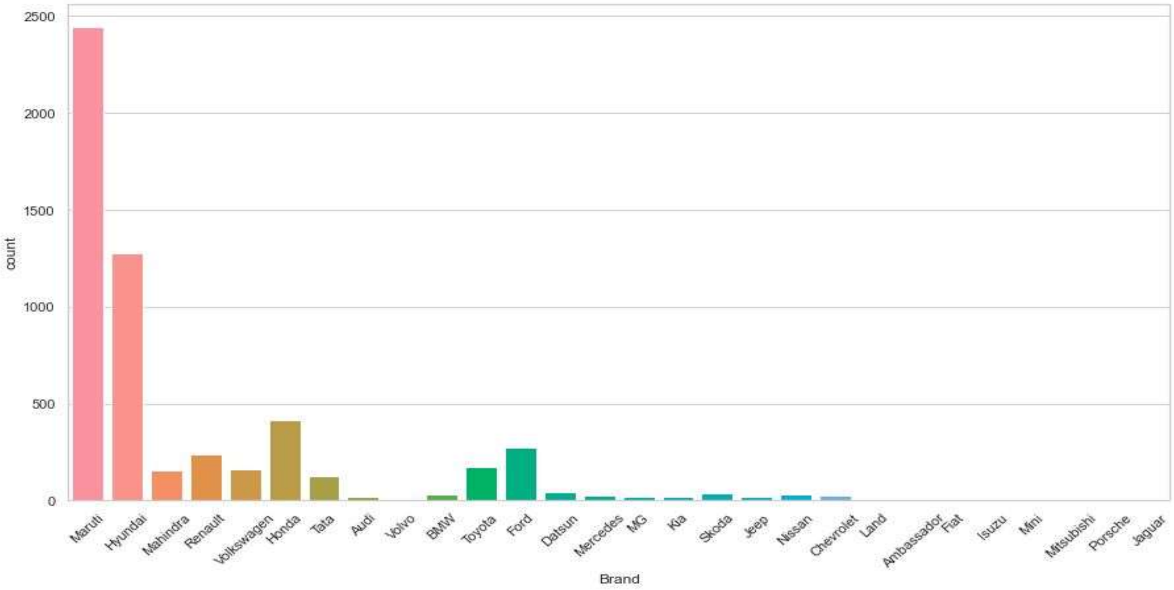


```
plt.figure(figsize=(6,5))
sns.boxplot(df['Price'])
```

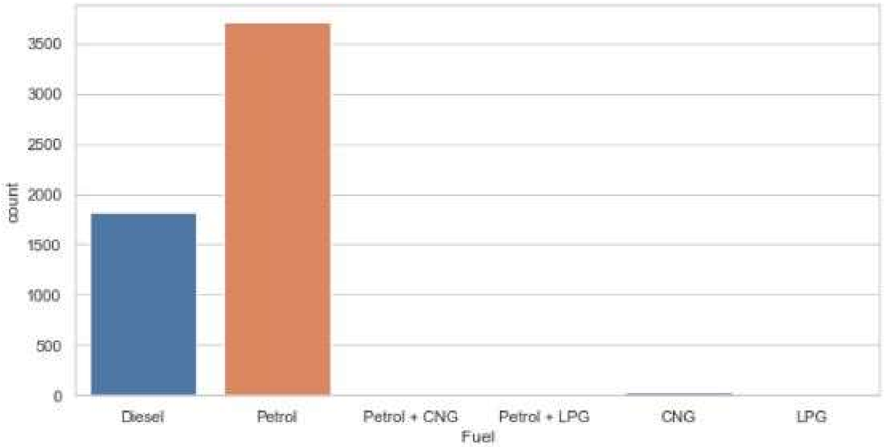
```
<AxesSubplot:xlabel='Price'>
```



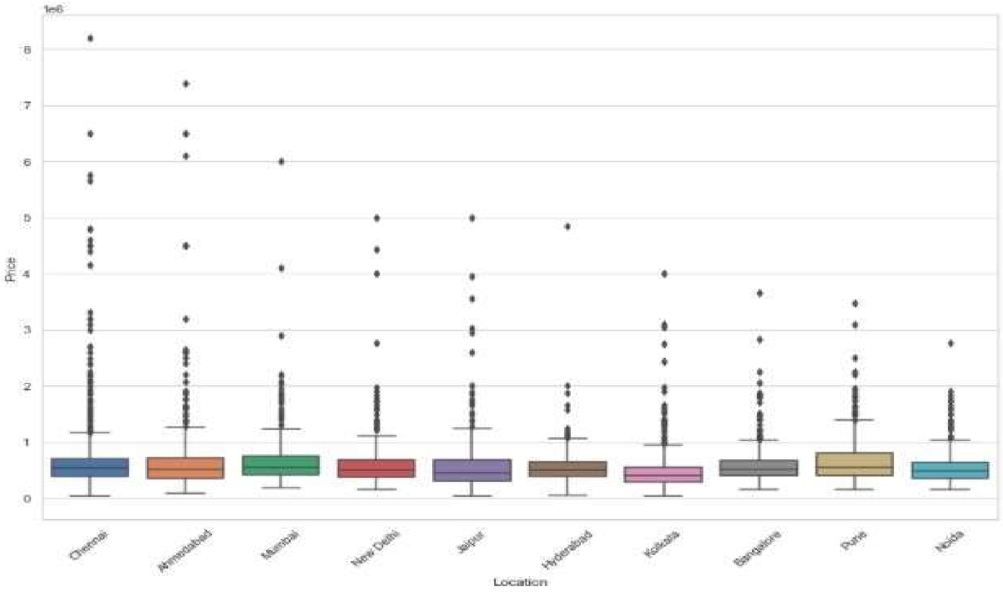
The Count plot shows the count of vehicles in the dataset. Large number vehicles from the brands Maruti, Hyundai, Honda, Ford are available and a smaller number of vehicles from Isuzu, Jaguar, Porsche and Mini.



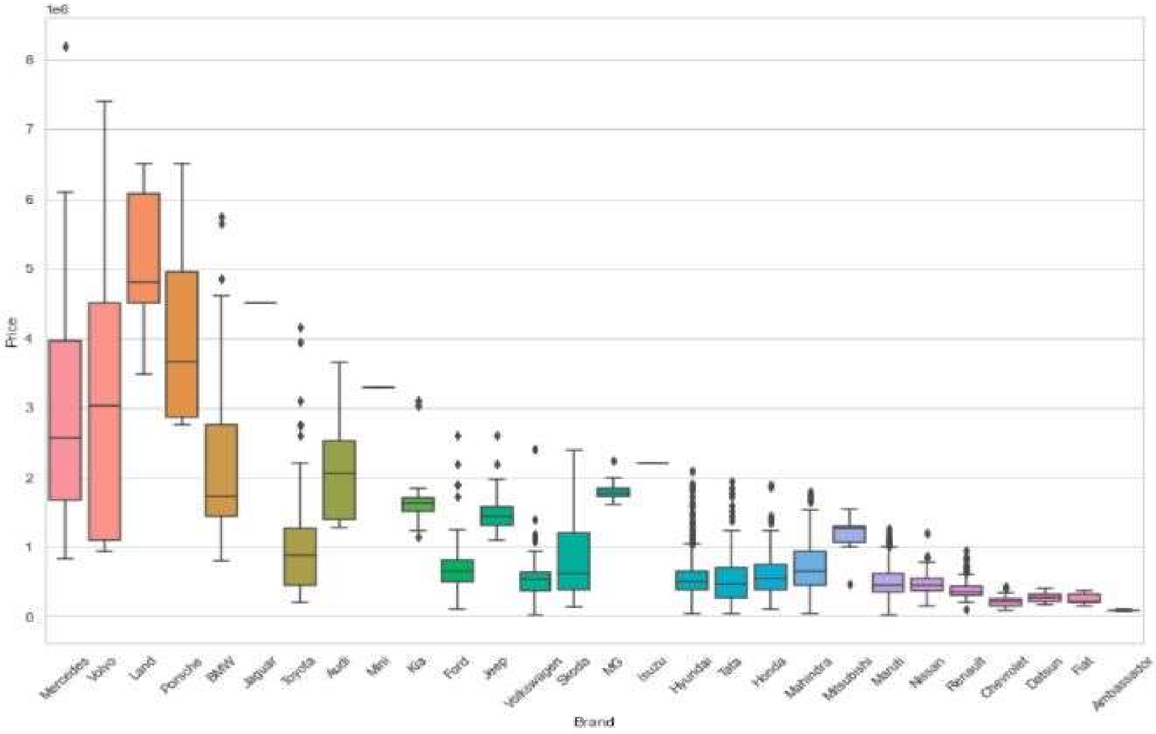
Count plot also shows mainly petrol vehicles are using compared to diesel and CNG.



The Price of the cars are almost similar in all the locations, but compared to other places price at pune is little bit higher and the lowest price is at Kolkata.



The price range of brands Mercedes, Volvo, Porsche and BMW is highest compared to other brands. The price range of brands Ford, Jeep, Skoda, Hyundai, Tata, Honda, Mahindra and Maruti is in the middle range. The price range of Nissan, Renault, Chevrolet, Datsun and Fiat is the lowest range.



- Interpretation of the Results

The only feature that positively correlated to the output variable/label is Manufacturing year, the price of the car mainly dependents on the year of manufacturing and the other input variables like Brand, model, etc has less role in predicting the price.

CONCLUSION

- Key Findings and Conclusions of the Study

From the given dataset it shows that a large number of cars from the brands like Maruti, Hyundai, Honda and Ford are available in the market and their price is correlated to the manufacturing year. The change in the manufacturing year will also change the price of the car. Majority of the vehicles using fuel type as petrol and diesel only small number of vehicles are using CNG and LPG. The price of the vehicle mainly depends on the manufacturing year.

- Learning Outcomes of the Study in respect of Data Science

Five machine learning algorithms are used here for predicting the price of the car, after hyper parameter tuning decision tree, random forest and gradient boosting models gives better result. From these three models, the gradient boosting model gives the best result with R2 score of 82.35% and cross validation score 72.39%.