

# **CAR PRICE PREDICTION**

Submitted by:

**MOHAMMED MINHAI**

## **ACKNOWLEDGMENT**

Wikipedia, Youtube.com, google.com. The data were collected from the used car website cardheko.com. Referred in study materials of machine learning which I am studying.

# INTRODUCTION

- **Business Problem Framing**

Pandemic situation affects in the world. It affects in the different markets. In this project we are looking in the car market.as this pandemic affects, people only focuses budget cars for their daily uses. As petrol and diesel price is increasing daily basis, people of middle class focuses on mileage rather than other features. And also they focuses on the brands of the cars which will have low maintenance. So people focuses on this type of cars to buy in the market.

- **Conceptual Background of the Domain Problem**

The data were collected from the used car website cardheko.com. The data contains the brand of the cars, model, variant, no of owners, year of manufacture.

- **Review of Literature**

For this project I have researched all popular used cars websites and find the problem. We can assume that among the features we collected that for the prediction, people looking for budget friendly cars for their daily basis. They look the features like the mileage, low maintenance and brand, manufacture year of the car. They don't look for the safety the power of engine like features before buying. Here I am talking about the ordinary people not the professionals of the car. Pandemic situation affects the people so badly. They buy car for daily basis which is needed not for entertain purpose.

- **Motivation for the Problem Undertaken**

The main objective is to help the car selling people to overcome their situation and how to attract people to the car market. Each and every person in the world affects so badly the pandemic

situation. It affect in car market and trying to help them to overcome this situation.

## Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem
- Predictions are done by analysing the given data. The given data will be the past data. With this past data we want to predict future. So for analysing the past data we need to use some techniques for that. We want to relate each features, how they are related, how they related to the output. In this project the features fuel, brand, manufacture year. These features are related to the label price and some of them were will give a high impact in output.

- Data Sources and their formats

In this project data are collected from the used cars website. I collect the data from the website cardheko. Collected data from different location in the website cardheko. I research al other used cars website but the data from cardheko is more relevant easy to understand the data. In other websites some data are missing. So I scrape the data from this website

The screenshot shows a Jupyter Notebook interface with the following content:

```
In [2]: 1 df=pd.read_csv('car_price_data.csv',index_col=0)
```

```
In [3]: 1 df.head()
```

Out[3]:

	Brand	Model	Manufacture Year	variant	Fuel	Driven Kilometers	Transmission type	Location	Price	Number of owners
0	Renault	KWID	2016	1.0 RXT Optional AT 2016-2019	Petrol	26,769 kms	automatic	Bhopal	3,42,000	-
1	Renault	KWID	2019	1.0 RXT Optional AT 2016-2019	Petrol	13,342 kms	automatic	Mumbai	4,48,000	-
2	Hyundai	EON	2013	Magna Plus	Petrol	33,824 kms	manual	Mumbai	2,66,000	-
3	Maruti	Alto	2016	CNG LXI	CNG	18,679 kms	manual	Mumbai	3,46,500	-
4	Hyundai	Verna	2018	VTVT 1.6 SX	Petrol	4,694 kms	manual	Mumbai	9,54,000	-

```
In [4]: 1 df.shape
```

Out[4]: (5109, 10)

- **Data Pre-processing Done**

As we want to predict the price of the car we want to build a regression model. When we collected the data and the data type of the price feature is in object data type. We want to convert it into integer. First of all we want to replace the commas (,) and convert to numeric using to numeric function in pandas. Same step done to the feature driven kilometres because it is also in object data type.

Dataset have only one missing value we can just drop it. The feature 'no of owners' have the feature with hyphen (-), we can also replace it with the mode of the feature. Mode is 'first owner' so we can replace it by first owner. The data of first owner have given in 1<sup>st</sup> owner also let's replace and append it by the first owner.

The features like brand, model, variant, location are encoded using label encoder.

While the features like no of owners, transmission, fuel were used one hot encoder. These are the pre-processing done in data.

- **Data Inputs- Logic- Output Relationships**

As in supervised learning there should be label with respect to features the machine will predict the output. Here the features like 'transmission', 'fuel', 'brand' are related to the output price. So here the type of fuel which is diesel cars and automatic transmission type are also impact the price which is our output.

- **State the set of assumptions (if any) related to the problem under consideration**

The cars which have fuel type diesel and transmission automatic have high demands in cars market.

Demands to the diesel cars state that the cars have high mileage will have high demands in market.

- **Hardware and Software Requirements and Tools Used**

Here the dataset is small so the memory usage also small, which have only 450 KB. For data analysing and visualizing here we use pandas packages, matplotlib and seaborn for visualization technique. The packages for four algorithms tree package for decision tree, linear model packages for linear regression, neighbors packages for knearest neighbors algorithm, ensemble packages for random forest. For performing an algorithm first of all we need to split it into train and test model for that here we can use the package model selection. In same package we can use for hyper parameter tuning and to find cross validation score. And metrics packages used for finding mean squared error, mean absolute error and R2 score. For scaling and encoding imported from the package preprocessing.

## **Model/s Development and Evaluation**

- **Identification of possible problem-solving approaches (methods)**

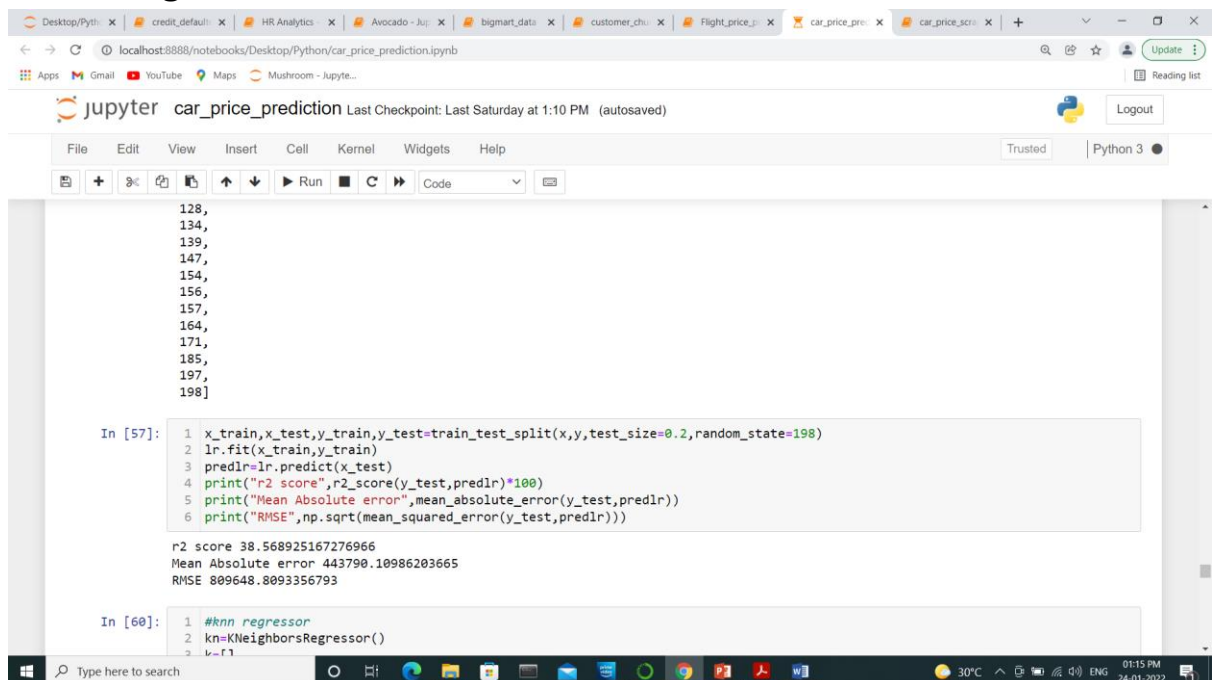
Our objective in this project is to predict the car price. For the prediction we need some features related to the label. Here for the price of the car is related to some features of the car. Electrical vehicles are releasing in our country but it is not common. May be in future it demands in car market. Currently the price of petrol increasing so the people select for the cars which have high mileage and it demands high in market.

- Testing of Identified Approaches (Algorithms)

- Linear Regression
- Decision Tree Regressor.
- K-nearest Neighbors Regressor
- Random Forest Regressor

- Run and Evaluate selected models

### Linear Regression



```
128,  
134,  
139,  
147,  
154,  
156,  
157,  
164,  
171,  
185,  
197,  
198]  
  
In [57]: 1 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=198)  
2 lr.fit(x_train,y_train)  
3 predlr=lr.predict(x_test)  
4 print("r2 score",r2_score(y_test,predlr)*100)  
5 print("Mean Absolute error",mean_absolute_error(y_test,predlr))  
6 print("RMSE",np.sqrt(mean_squared_error(y_test,predlr)))  
  
r2 score 38.568925167276966  
Mean Absolute error 443790.10986203665  
RMSE 809648.8093356793  
  
In [60]: 1 #knn regressor  
2 kn=KNeighborsRegressor()  
3 lr=lr
```

Here observing the above picture shows that the r2 score for linear regression is very low it showing only 38%. While MSE and RMSE are higher. First of all split the model into test and train model. 20% of data were used for testing and remaining 80% of data were for training. And we check for best random states and selecting one of the best random state 198 and get the r2 score 38%.

### K-Neighbors Regressor

```
At random state 199,the training r2_score is:- 0.7161352188704648
At random state 199,the testing r2_score is:- 0.6114599857093749

In [63]: 1 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=64)
2 kn.fit(x_train,y_train)
3 predkn=kn.predict(x_test)
4 print("r2 score",r2_score(y_test,predkn)*100)
5 print("Mean Absolute error",mean_absolute_error(y_test,predkn))
6 print("RMSE",np.sqrt(mean_squared_error(y_test,predkn)))

r2 score 67.88834927269761
Mean Absolute error 268875.5457925636
RMSE 536378.0882105023

In [64]: 1 #dt regression
2 dt=DecisionTreeRegressor()
3 dt.fit
```

When we go through k neighbors as compared linear regression it showing better r2 score, MSE, RMSE are better. Knn giving around 67% of r2 score at random state 64. 20% of data were used for testing and remaining 80% for training.

## Decision Tree Regressor

```
At random state 186,the training r2_score is:- 0.9999856839438148
At random state 186,the testing r2_score is:- 0.7463286411951675

At random state 187,the training r2_score is:- 0.9999860866573191
At random state 187,the testing r2_score is:- 0.7478197659358756

At random state 188,the training r2_score is:- 0.9999842600878576
At random state 188,the testing r2_score is:- 0.8181212599847519

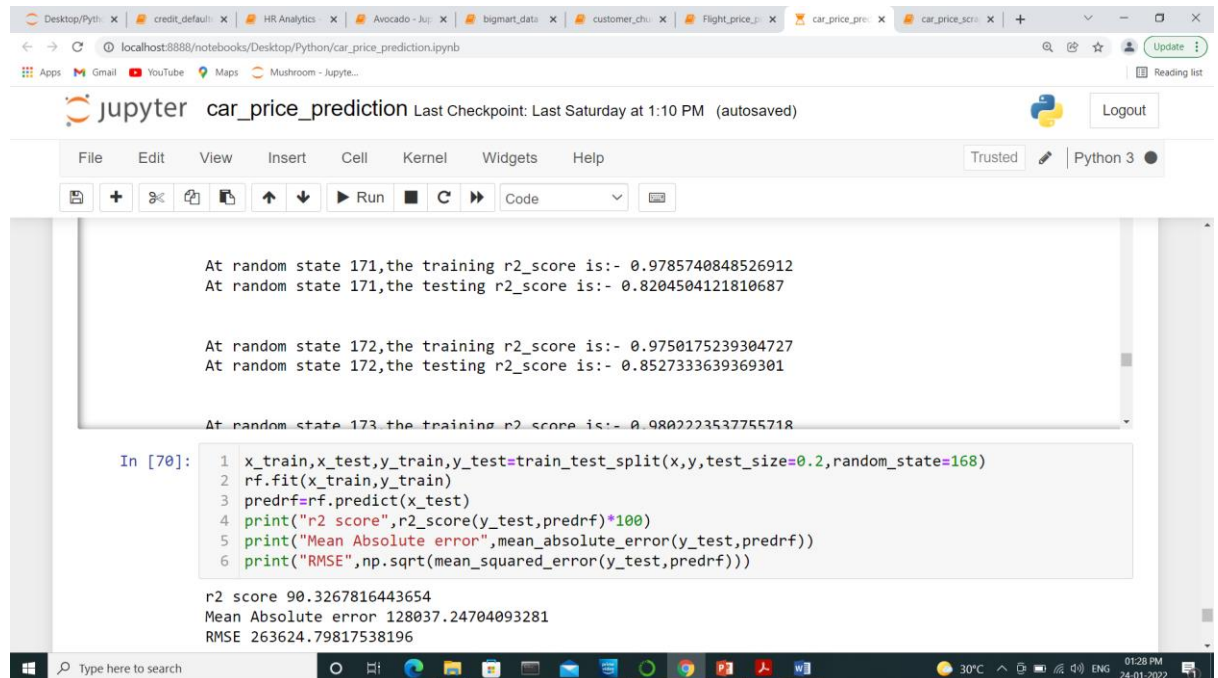
In [66]: 1 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=143)
2 dt.fit(x_train,y_train)
3 preddt=dt.predict(x_test)
4 print("r2 score",r2_score(y_test,preddt)*100)
5 print("Mean Absolute error",mean_absolute_error(y_test,preddt))
6 print("RMSE",np.sqrt(mean_squared_error(y_test,preddt)))

r2 score 84.0544415297478
Mean Absolute error 154078.5376712329
RMSE 377019.11837365205
```



As compared to both of the algorithms used, decision tree gives better r2 score and MSE and RMSE as compared to them. Decision tree is giving 84% for r2 score at random state 143. 20% of data were split for testing and remaining for training.

## Random Forest Regressor



```
At random state 171,the training r2_score is:- 0.9785740848526912
At random state 171,the testing r2_score is:- 0.8204504121810687

At random state 172,the training r2_score is:- 0.9750175239304727
At random state 172,the testing r2_score is:- 0.8527333639369301

At random state 173,the training r2_score is:- 0.9802223537755718

In [70]: 1 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=168)
         2 rf.fit(x_train,y_train)
         3 predrf=rf.predict(x_test)
         4 print("r2 score",r2_score(y_test,predrf)*100)
         5 print("Mean Absolute error",mean_absolute_error(y_test,predrf))
         6 print("RMSE",np.sqrt(mean_squared_error(y_test,predrf)))

r2 score 90.3267816443654
Mean Absolute error 128037.24704093281
RMSE 263624.79817538196
```

As we can say that, random forest is best model among them. It is giving r2 score of 90% and better MSE and RMSE. At random state 168 it gives the best r2 score. Random forest is selected as our final model for our project because it have least difference in cv score and r2 score.

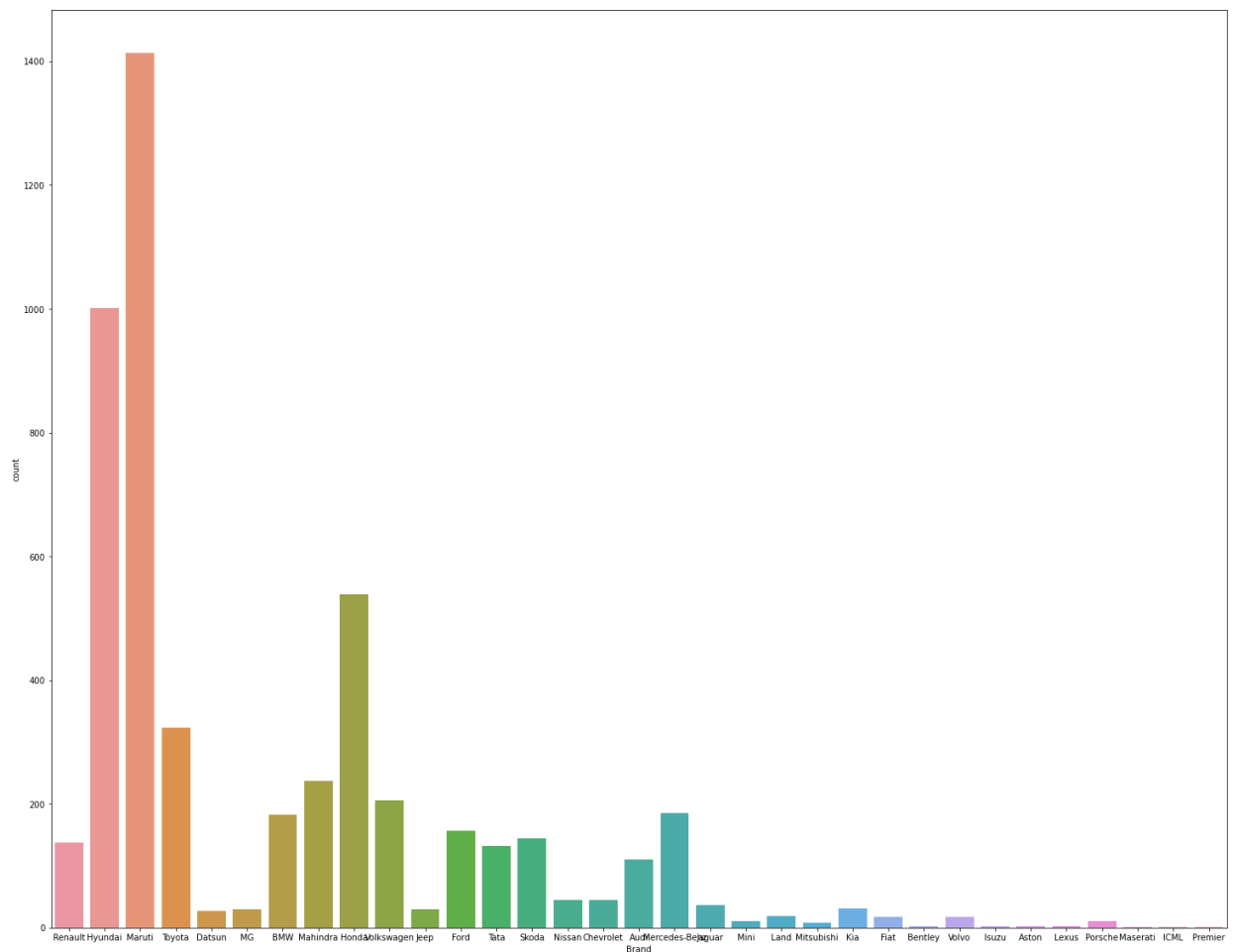
- Key Metrics for success in solving problem under consideration

R2 score, Mean absolute error (MSE) and Root Mean Squared error (RMSE) were used in this project. Here the R2 score of the project using random forest regression is 90%. As we know the Mean squared error told us that how close the regression line with set of points. Least difference makes a good regression line. While RMSE

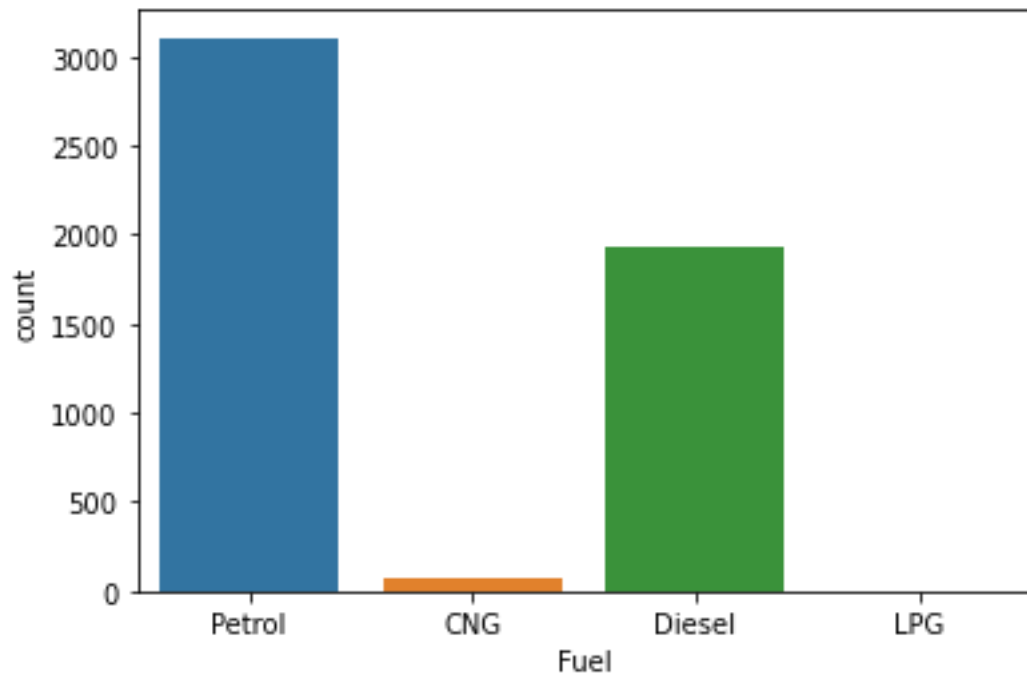
is the standard deviation of the residuals or errors. These are the key metrics used in this project.

- Visualizations

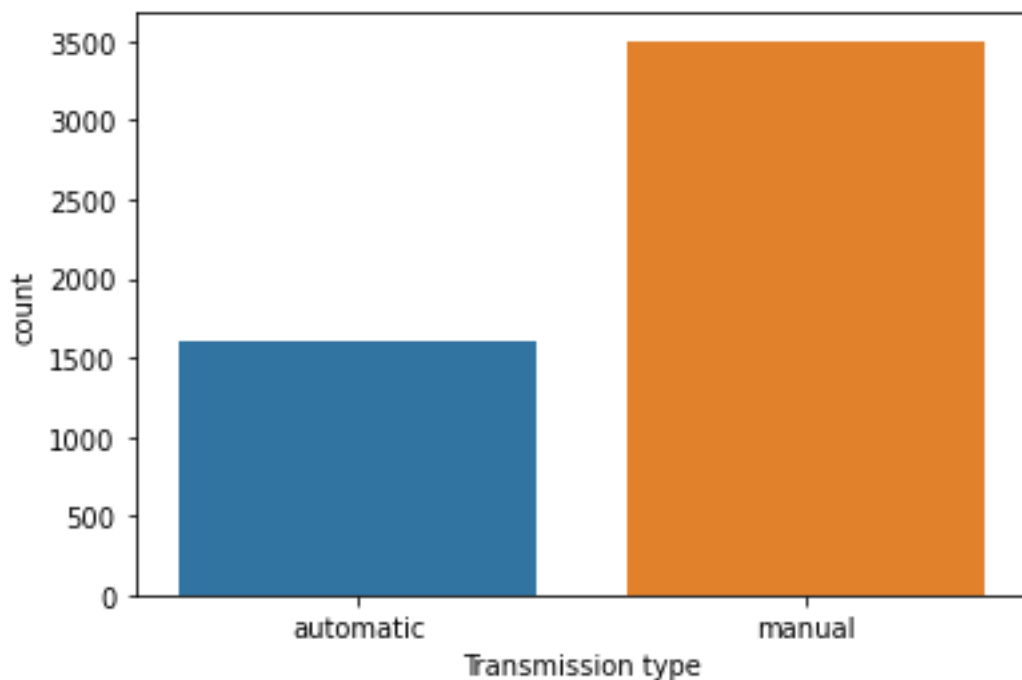
Visualizations are one of the analytics step used while analysing the data. Let's look the visualizations of this project.



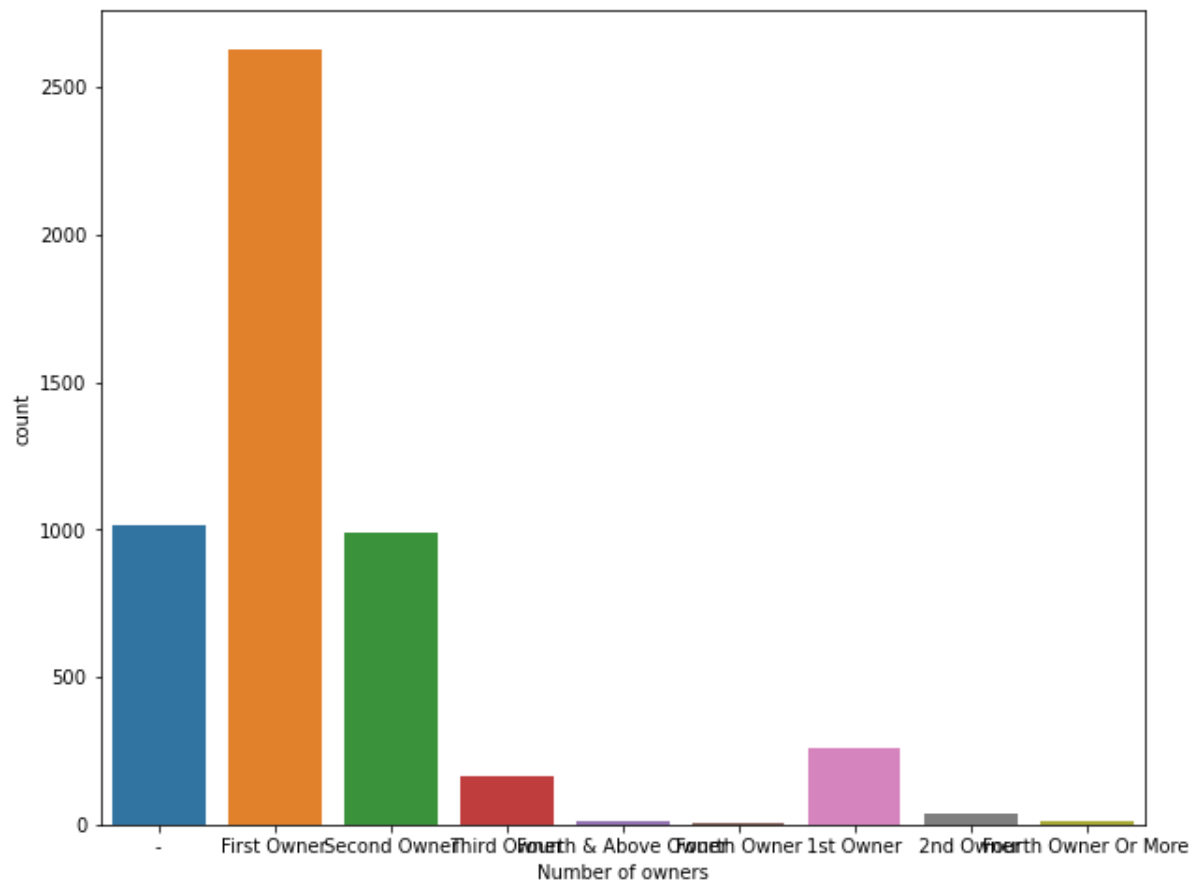
The above visualization is the count plot of the feature 'Brand' here we can observe that in the given data 'Maruti Suzuki' brands are more compared to others. The second most cars were of 'Hyundai' cars.



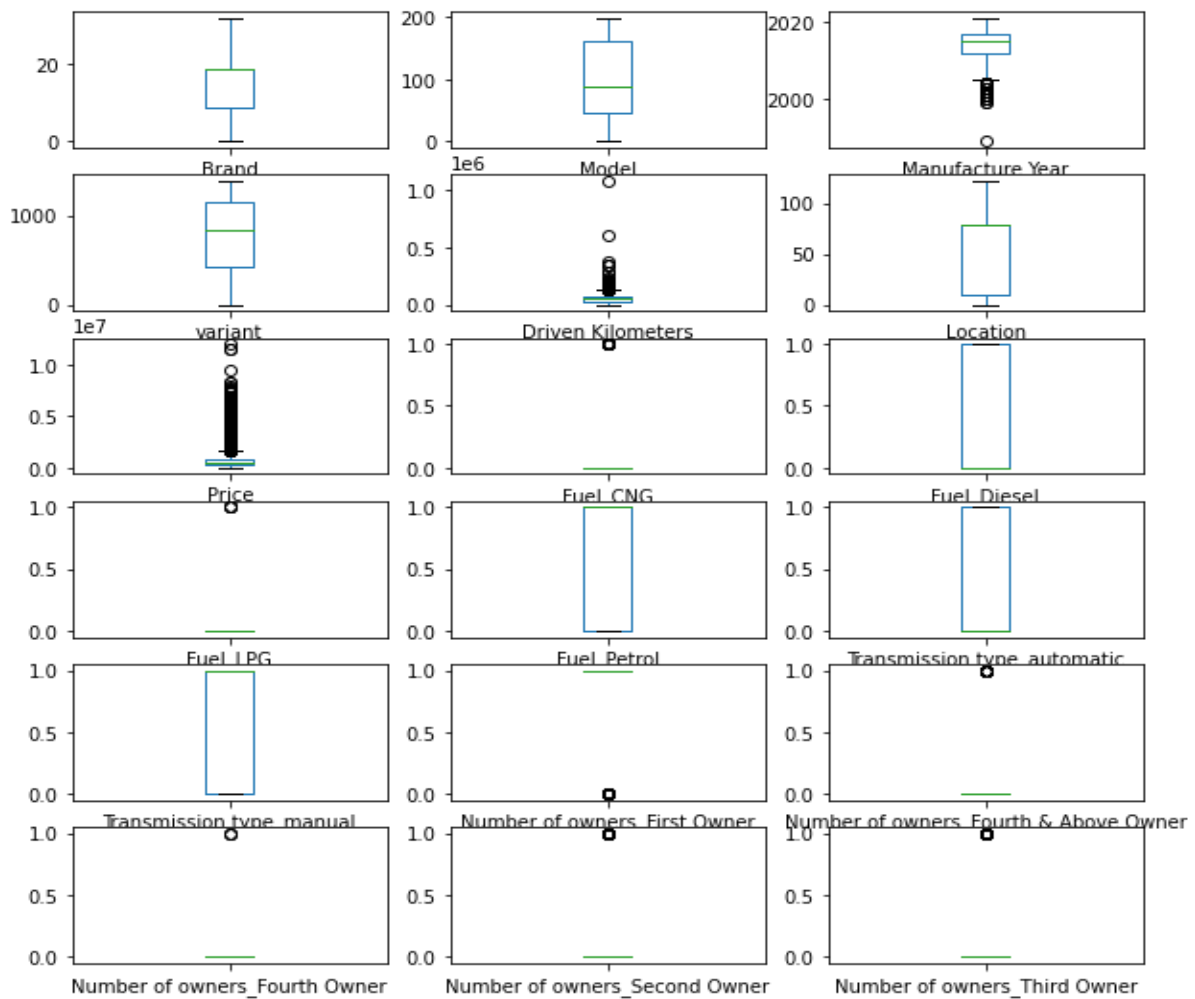
In this figure I used to check the count of the fuels in the given data. Here in this data the given cars fuel were petrol. And other fuels are diesel, CNG, LPG.



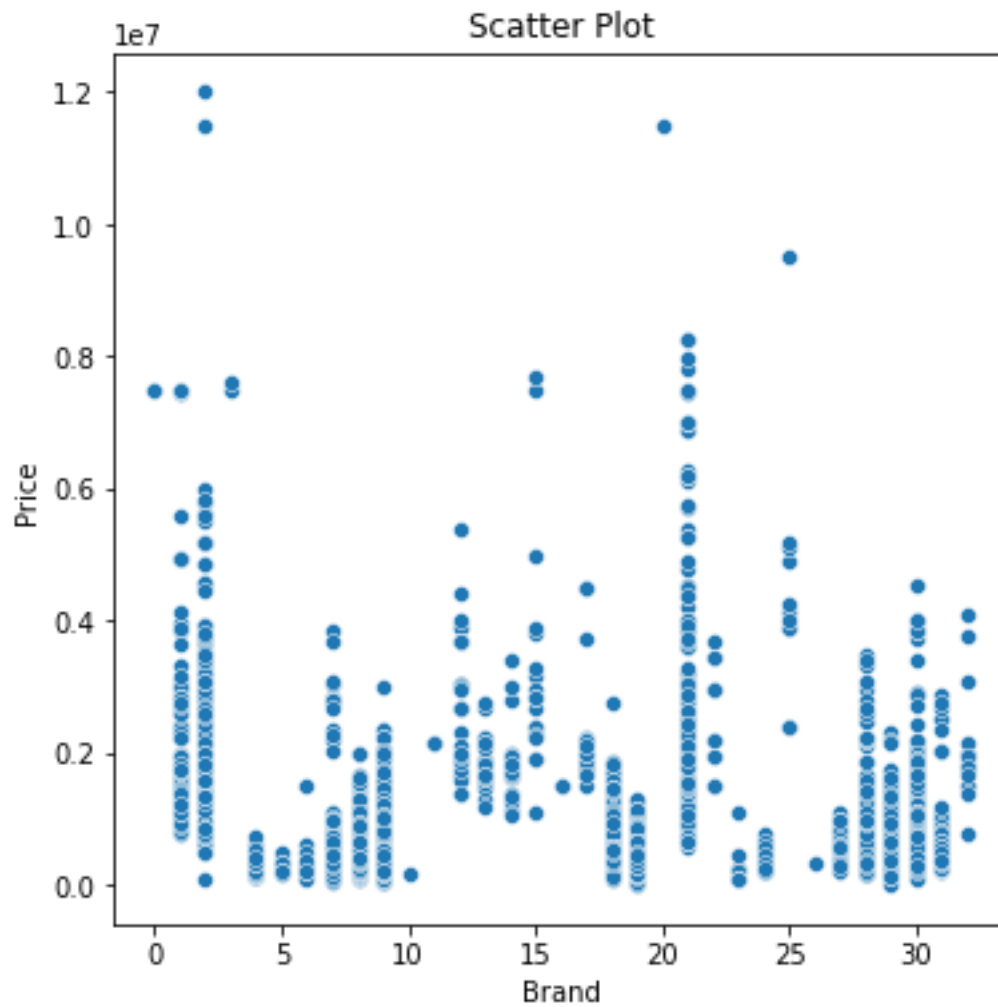
In the above figure we can observe that the count of transmission type in the given data. Here we can observe that manual transmission type of data were given mostly.



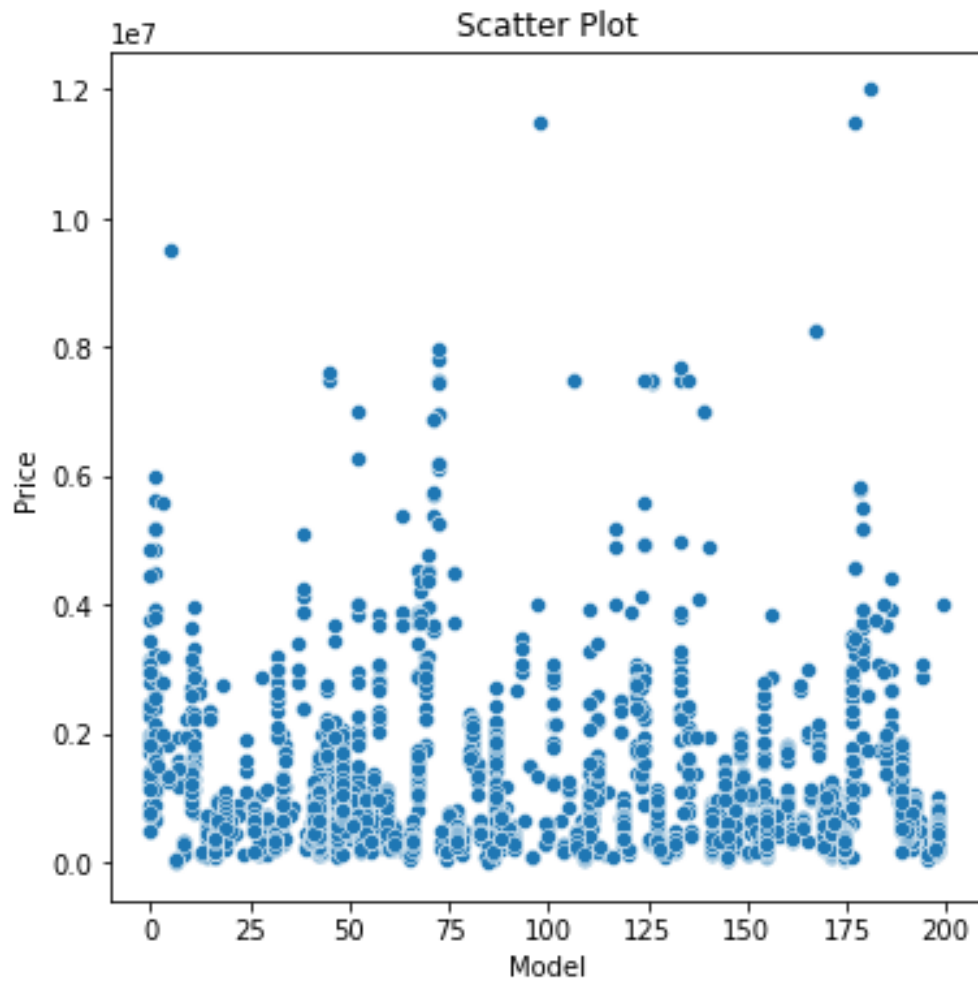
Here we can observe the number of owners used the car. Here most of them were first owners. We replaced the hyphen (-) given in the features with the most occurred value that is First owner. And the most of the car were used by first owner.



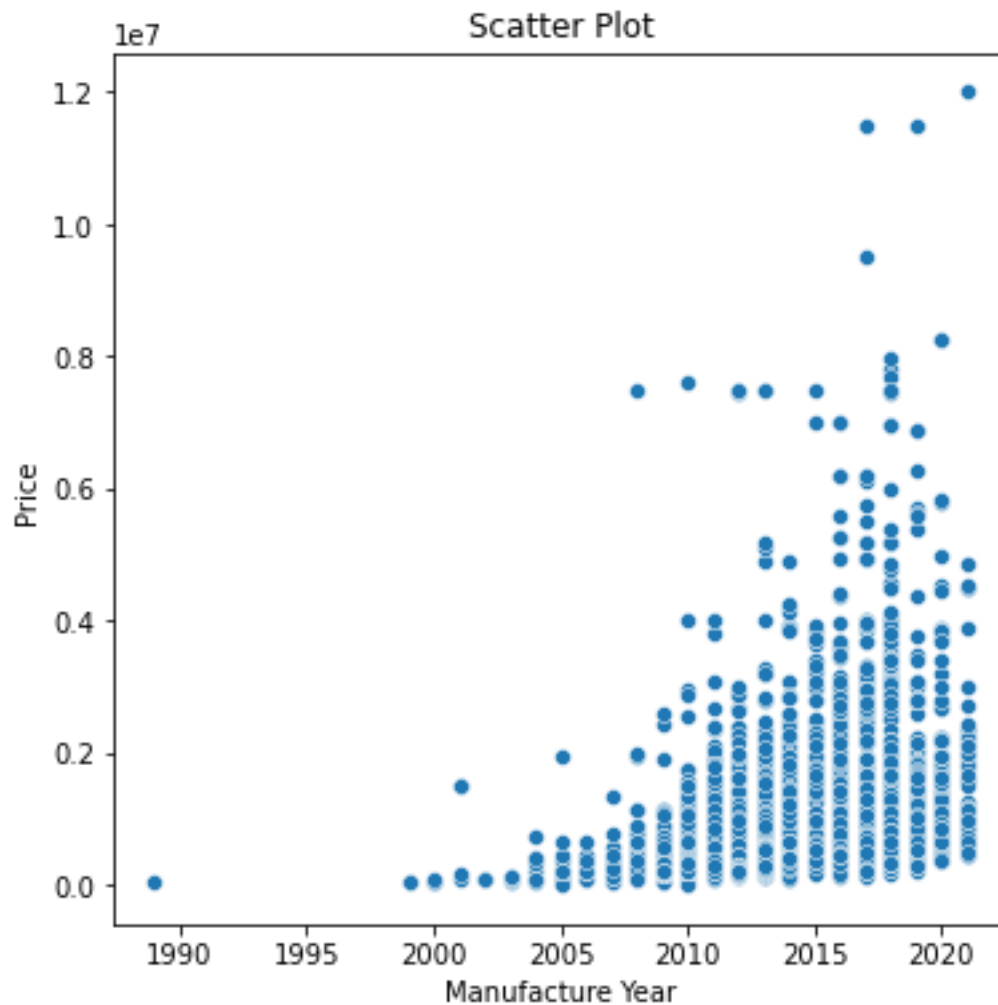
Here I used boxplot for checking the outliers, here we can observe some outliers present. While removing outliers the around 8% of data were loses. So I proceed with outliers.



When we compare the brand with price, most of the brand have a medium price and brand were correlated to price.

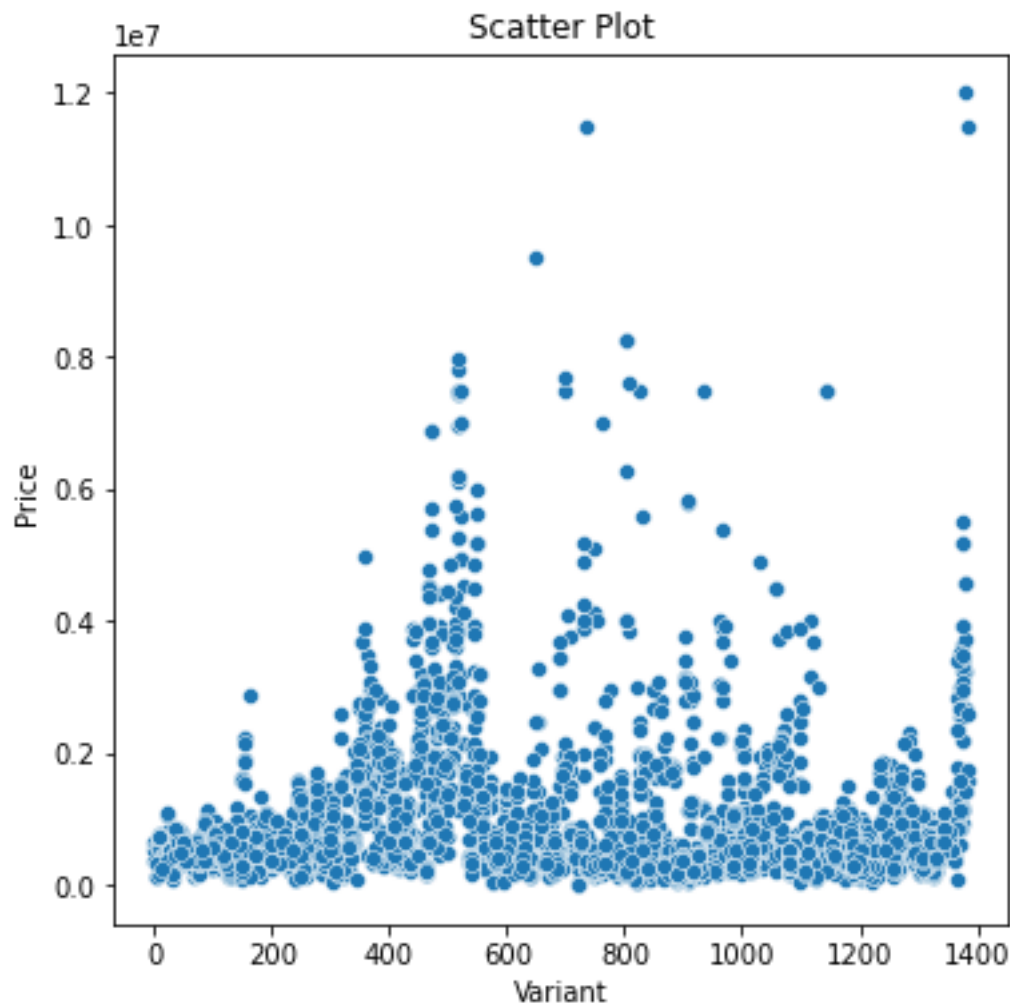


When we compare the model and price of the car, the model is up to brand which brand's model have a good condition and the customer will expect mainly from brand.

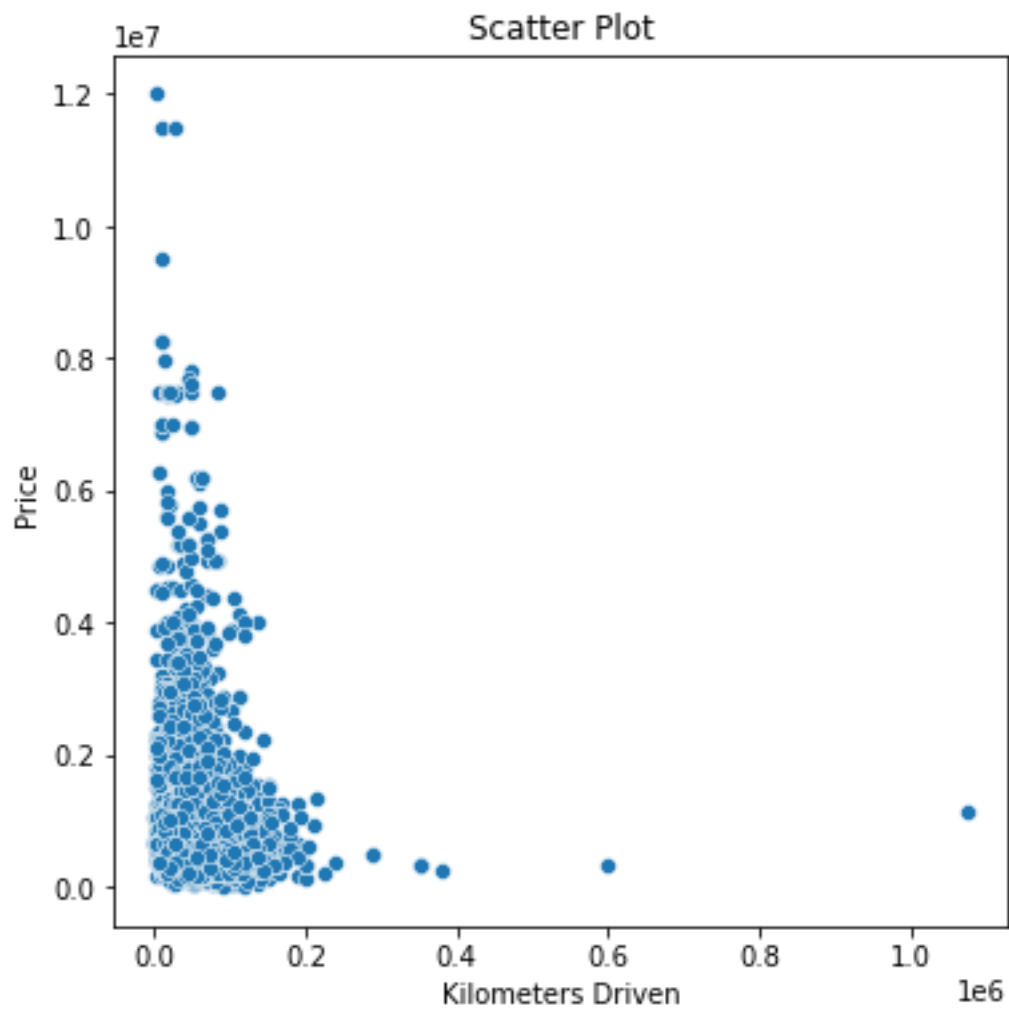


The year of the manufacture one of the main feature of car price. All brands are releasing new model of car and also they provide some variant and give update to features in existing model. So as usual the price of the old variant or model price will reduce. The customer will first check the manufacture year of the car. As the age of the car increasing the price of the car decreases.

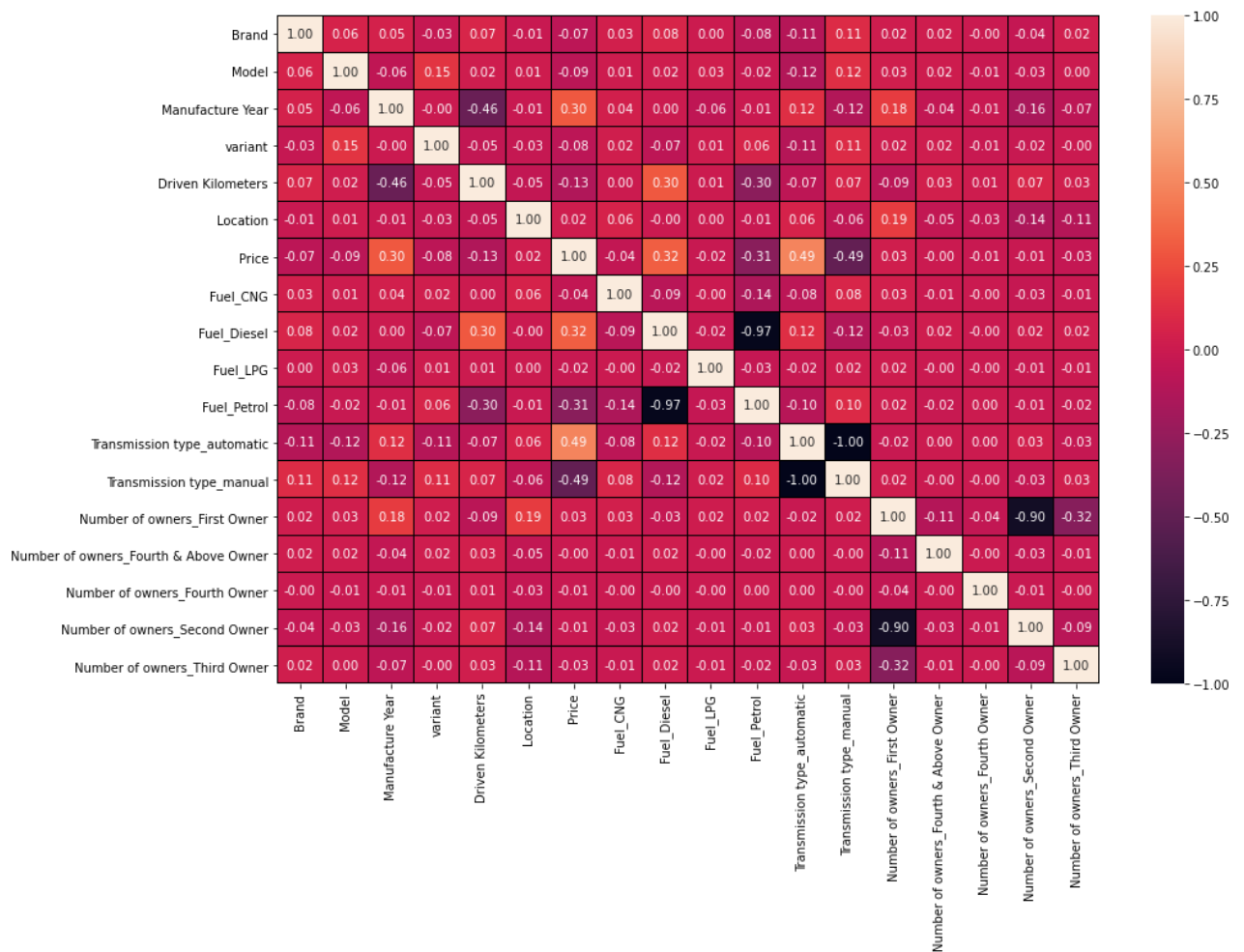




The variant of the car also related to car price. Cars have so many features like power window or not having power window, airbags, stereos, alloy wheels these are some basic features given to the car. The difference in variant varies in price. So the variant also have something to do with car price.



Driven kilometres are the one of the major feature affect to the price.  
More kilometres driven ay affect the price.



Here we can observe the correlation of the features and features with target variable price. The manufacture year, transmission type, fuel type of the car are correlated to the price. It means these features were impact the output price.

- **Interpretation of the Results**

So after all the steps, we can say that fuel type plays a major role in price. That is according to the given data of this project the car with fuel diesel have high price. That is in India price of petrol is increasing in daily basis so the customer attract to the cars which have high mileage. The diesel cars have good mileage compared to petrol cars so it have high price in market. And year of manufacture of the car also have impact in features. Brand, model and it's variant also plays a major role in price.

## **CONCLUSION**

- **Key Findings and Conclusions of the Study**

Pandemic situation affects the world. Every region of business were trying to raise and overcome the situation. According to our project the some features may affect the price of the car. Every people look for budget car for daily basis. So having car with high mileage and low maintenance have high demands in the market.

- **Learning Outcomes of the Study in respect of Data Science**

Visualization gives a good understanding of data how the features are related to the label and how features are related to each other. So such basic and detailed understanding of data is given by the visualization tools. In this projects I used four different algorithms. Linear regression perform very badly. It is because we can assume that the data are not linearly scattered. While this problem solved by bagging technique random forest. It gives high performance on the project. After tuning with best parameters it gives 67% of accuracy. So it is common for a data scientist occurrence of over fitting and under fitting of data so it is overcome by many techniques, which one is choosed by the data scientist.

- Limitations of this work and Scope for Future Work

Now a days the electric vehicles and CNG vehicles are provided by some major companies. It is good for environment, pollution control, cost control are benefits to people and environment and people. So people will attracts to those vehicles and demands of the diesel and petrol vehicles will get low. So the market of this type of cars gives a good result.