MACHINE LEARNING APPLIED: USED CAR PRICE PREDICTION.

Submitted by:

Gaurav Borole

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Khushboo Garg as well as Flip Robo Technologies who gave me the opportunity to do this project on Used Car Price Prediction, which also helped me in doing lots of research wherein I came to know about so many new things especially the data collection part.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1) https://www.google.com/

2) https://www.youtube.com/

3) https://scikit-learn.org/stable/user_guide.html

4) https://github.com/

5) https://www.kaggle.com/

6) https://medium.com/

7) https://towardsdatascience.com/

8) https://www.analyticsvidhya.com/

INTRODUCTION

Business Problem Framing.

Impact of COVID-19 on Indian automotive sector

The Indian automotive sector was already struggling in FY20. Before the Covid-19 crisis. It saw an overall DE growth of nearly 18 per cent. This situation was worsened by the onset of the Covid-19 pandemic and the ongoing lockdowns across India and the rest of the world. These two years (FY20 and FY21) are challenging times for the Indian automotive sector on account of slow economic growth, negative consumer sentiment, BS-VI transition,changes to the axle load norms, liquidity crunch, low-capacity utilisation and potential bankruptcies.

The return of daily life and manufacturing activity to near normalcy in China and South Korea, along with extended lockdown in India, gives hope for a U-shaped economic recovery. Our analysis indicates that the Indian automotive sector will start to see recovery in the third quarter of FY21. We expect the industry demand to be down 15-25 per cent in FY21. With such DE growth, OEMs, dealers and suppliers with strong cash reserves and better access to capital will be better positioned to sail through.

Auto sector has been under pressure due to a mix of demand and supply factors. However, there are also some positive outcomes, which we shall look at.

- With India's GDP growth rate for FY21 being downgraded from 5% to 0% and later to (-5%), the auto sector will take a hit. Auto demand is highly sensitive to job creation and income levels and both have been impacted. CII has estimated the revenue impact at $2 billion on a monthly basis across the auto industry in India.
- Supply chain could be the worst affected. Even as China recovers, supply chain disruptions are likely to last for some more time. The problems on

the Indo-China border at Ladakh are not helping matters. Domestic suppliers are chipping in but they will face an inventory surplus as demand remains tepid.

- The Unlock 1.0 will coincide with the implementation of the BS-VI norms and that would mean heavier discounts to dealers and also to customers. Even as auto companies are managing costs, the impact of discounts on profitability is going to be fairly steep.
- The real pain could be on the dealer end with most of them struggling with excess inventory and lack of funding options in the post COVID-19 scenario. The BS-VI price increases are also likely to hit auto demand.

There are two positive developments emanating from COVID-19. The China supply chain shock is forcing major investments in the "Make in India" initiative. The COVID-19 crisis has exposed chinks in the automobile business model and it could catalyse a big move towards electric vehicles (EVs). That could be the big positive for auto sector.

Conceptual Background of the Domain Problem

Understanding the above business problem, there are certain factors that will influence the automotive industries in the future. Some of them include digital technologies, changing customer preferences, electrical vehicles, intelligent ability, and technical advancements. Technologies such as artificial intelligence, machine learning, cloud computing, and internet of things will also play an important role in developing new business models. Apart from that, they enable customers to ensure a better mobility experience. In other words, technologies may impact automotive industry units significantly that will change the markets. The introduction of electrical cars and hybrid vehicles may transform the automobile industries in coming years.

Review of Literature.

As per the requirement of our client, I have scrubbed data from different used cars selling merchants websites, and so based on the data collected I have tried

analysing based on what factors the used car price is decided? What is the relationship between cost of the used cars and other factors like Fuel type, Brand and Model, year the car is purchased and No. Of owners before selling? And so based on all the above consideration I have developed a model that will predict the price of the used cars.

Motivation for the Problem Undertaken

I have taken this problem based on the requirement of the client and also, with a curiosity to know how the used cars markets are at the time of pandemic.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

On importing the data that is collected I have done some initial play around to understand the data and to cleanse the data.

Data Cleansing.

On data cleansing I have detecting duplicate records in the data collected and the null values in the data.

```
In [10]: df.duplicated().sum()
Out[10]: 6667
```

As like we can see I had 6667 duplicate records out of 14975, I decided to drop all the duplicate records because it will not help us in creating a perfect model.

Removing the Duplicate Records.

```
In [12]: df = df.drop_duplicates() #Because we have too many duplicate records we will detete them to have a good model
```

After dropping the duplicate records, I checked for null values in the data.

```
In [13]: df.isnull().sum()
```

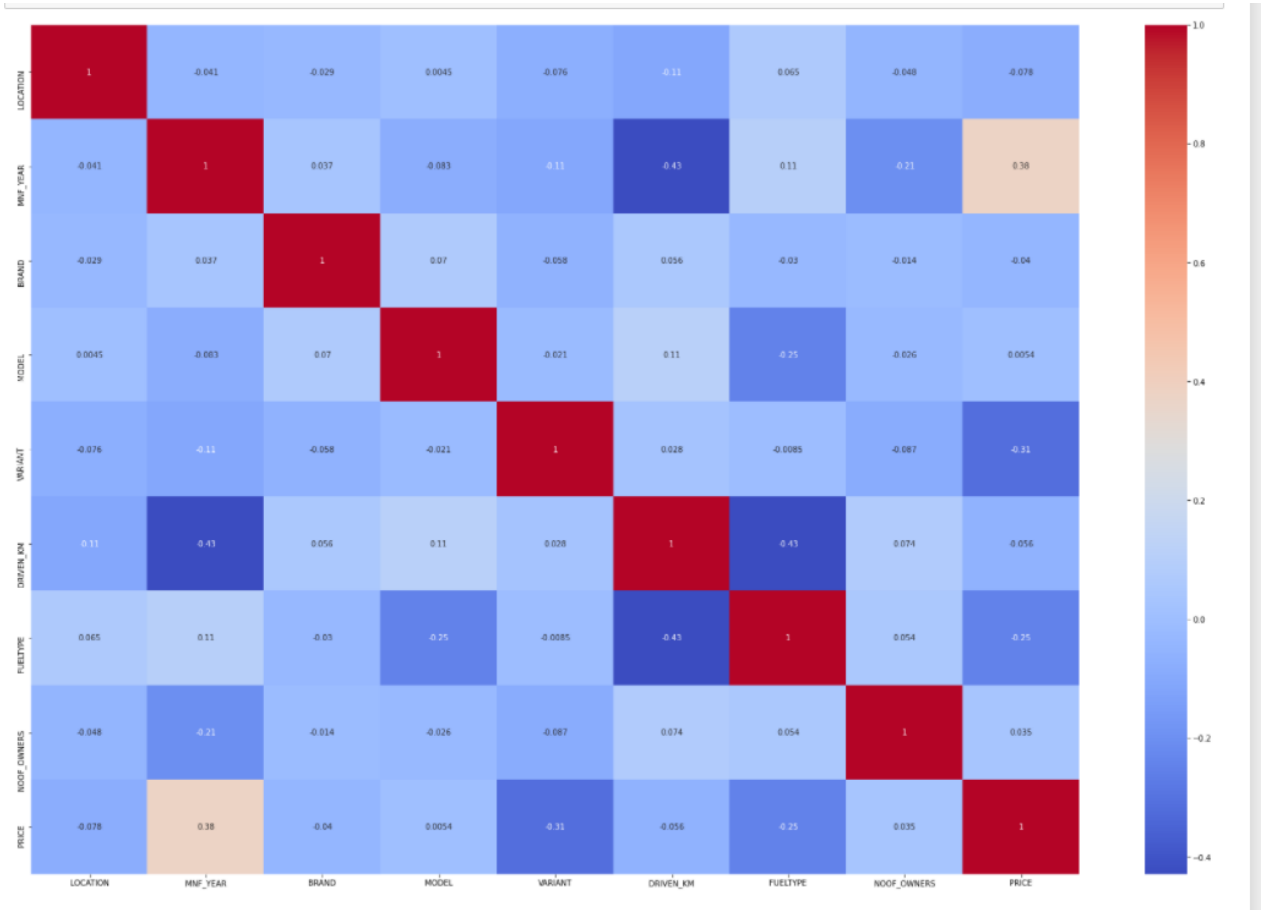
```
Out[13]: LOCATION        0
         MNF_YEAR        0
         BRAND           0
         MODEL           0
         VARIANT        10
         DRIVEN_KM       0
         FUELTYPE        0
         NOOF_OWNERS     0
         PRICE           0
         dtype: int64
```

From above we can see that the only in VARIANT we have values missing so we will change VARIANT nan values as NOT MENTIONED

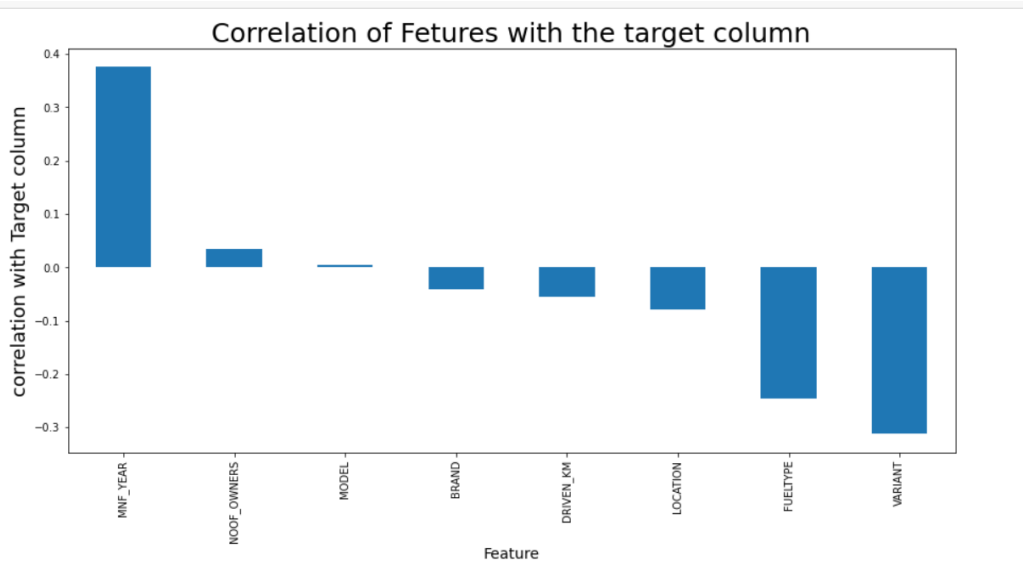
We have 10 null values in Variant column, I have used those null values in the model building but I have changed null values as not mentioned, this will also help the client to predict the values on the used cars without the Variant values.

Data correlations.

To better understanding the mathematical concept of the problem we have to see the correlation of the data.

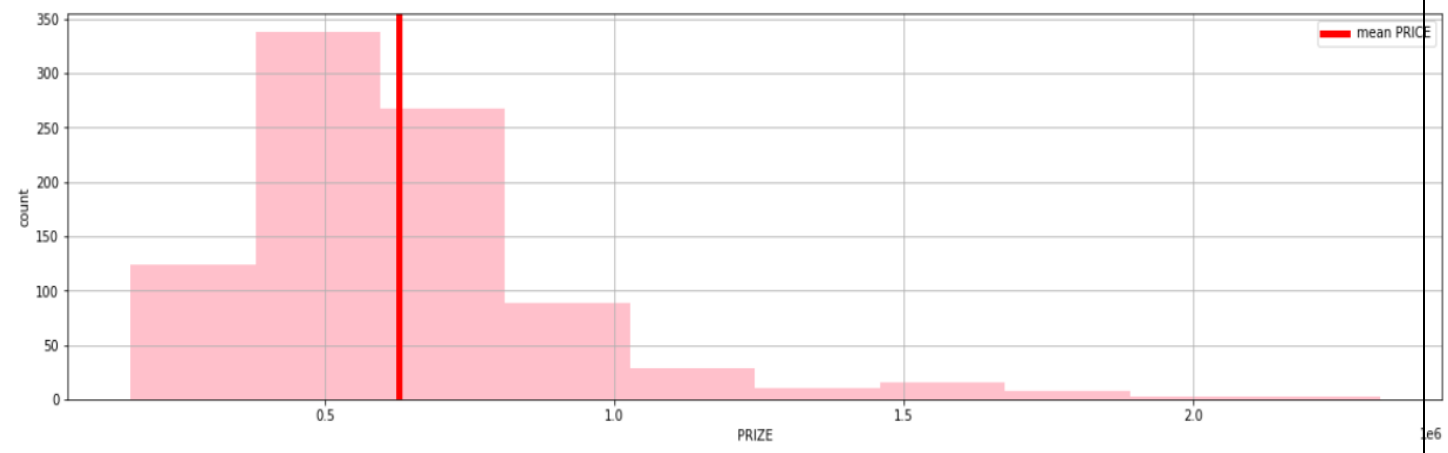


Key Observations:



- From above we can clearly see that MNF_YEAR is positively correlated to PRICE and FUEL_TYPE and VARIENT is negative correlated to PRICE.

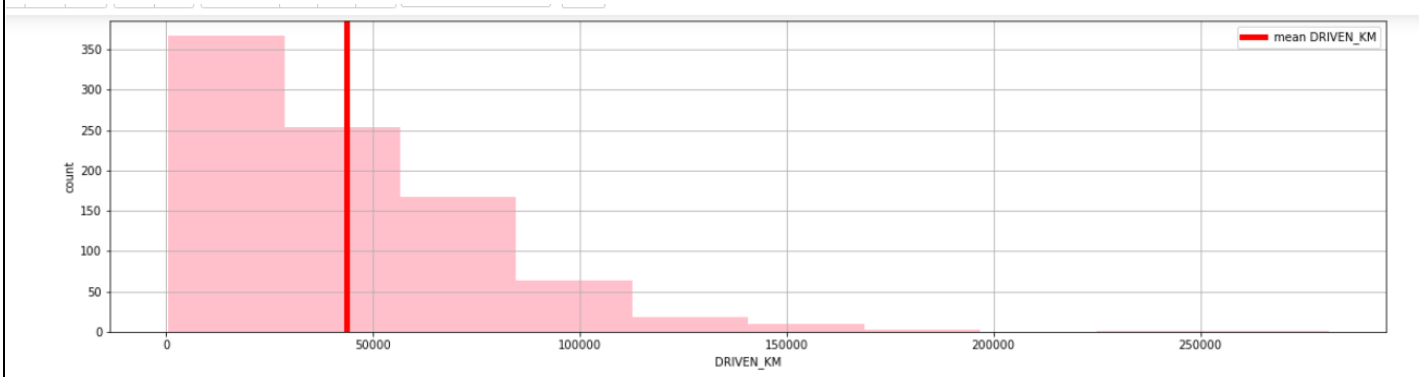
Univariate Analysis.




```
MATHEMATICAL SUMMARY OF PRIZE :
 count    8.830000e+02
mean     6.267737e+05
std      2.953709e+05
min      1.632990e+05
25%      4.264990e+05
50%      5.699990e+05
75%      7.309490e+05
max      2.322099e+06
Name: PRICE, dtype: float64
```

Key observations:

- Mean of the prize is Rs: 5,33,047, the prize is distributed between Rs: 99,162 to Rs: 41,00,000.
- Above we can understand that most of the Car price is lesser than the Mean i.e., Rs: 5,33,047




```
MATHEMATICAL SUMMARY OF PRIZE :
 count       883.000000
mean      43526.952435
std       33833.245517
min         470.000000
25%       18909.500000
50%       34119.000000
75%       62205.000000
max      280921.000000
Name: DRIVEN_KM, dtype: float64
```

Key observations:

- Mean of the DRIVEN_KM is 64489.278888kms and the maximum KMS driven is 312882kms.
- Above we can understand that most of the Car comes to selling around low kilometres driven.

Data Sources and their formats

The Data is scrubbed on multiple ecommerce websites that sells used cars in India, Website cars 24. These data are scrubbed and stored in a CSV format. Data contains following columns.

1. 'LOCATION' – It will tell which location the car is sold.
2. 'MNF_YEAR' – At what year the car is manufactured
3. 'BRAND' – Brand is manufacturer or which company made
4. 'MODEL' – It is basically the model of the car.
5. 'VARIANT' – Gear shift variant is (Automatic, Manual, Semi-Automatic)
6. 'DRIVEN_KM' – no of Kms driven before selling
7. 'FUELTYPE' – Petrol, diesel, CNG, LPG, Electric
8. 'NOOF_OWNERS' – 1end, 2end or 3end car
9. 'PRICE' – our target variable that tells what is the price of the used car.

Data Preprocessing Done

On pre-processing the data, I have tried in finding out the skewness of the data and the outliers, have changed the data into numbers with label encoders.

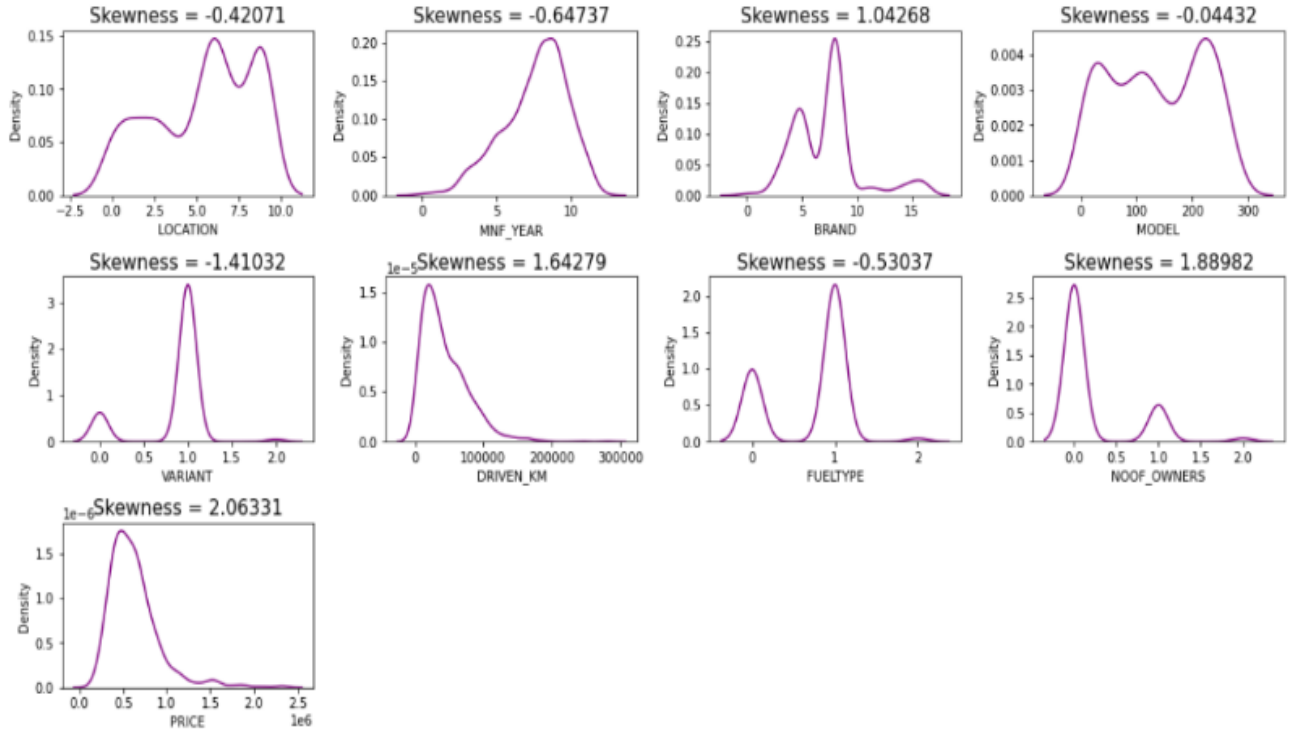
In [16]:
```
from sklearn.preprocessing import LabelEncoder
LE= LabelEncoder()
catagorical_data = ['LOCATION' , 'MNF_YEAR', 'BRAND', 'MODEL', 'VARIANT', 'FUELTYPE' , 'NOOF_OWNERS' ]

for i in catagorical_data:
    DF[i]=DF[i].astype('str')
    DF[i]=LE.fit_transform(DF[i])
```

In [17]:
```
DF['PRICE'] = DF['PRICE'].str.replace(r'\D', '').astype(int)
DF['DRIVEN_KM'] = DF['DRIVEN_KM'].str.replace(r'\D', '').astype(int)
```

After changing all the data with label encoders, I have tried in identifying skewness and outliers as follows.

In [33]:
```
plt.figure(figsize=(15,15))
for i in range (0, len(DF.columns)):
    plt.subplot(6,4,i+1)
    sns.kdeplot(DF[DF.columns[i]], color = "purple")
    plt.title(f"Skewness = {round(DF[DF.columns[i]].skew(),5)}",fontsize=15)
    plt.tight_layout()
```



We actually can see there are more skewness in the data let also see about the outliers in the data.

```
In [34]: from scipy.stats import zscore

         z= np.abs(zscore(DF))
         threshold= 3
         df_new = DF[(z < 3).all(axis=1)]
```

```
In [35]: print(f"Orginal Data {DF.shape}\nAfter Removing outliers {df_new.shape}\nThe percentage of data loss {((8308-7765)/8308)*100}%")

         Orginal Data (883, 9)
         After Removing outliers (823, 9)
         The percentage of data loss 6.535869041887338%
```

We have many outliers and we also have skewness in the data. Because it's more, correcting them will have loss in the data or data will be deformed. I am deciding to build the model with the skewness and outliers present in the data and this will also help the client to get an accurate prediction on the car prices with the data he gets.

And final step in pre-processing I have split the data and converted it into an array with Pre-processing Standard Scaler.

```
In [36]: x_1=DF.drop(["PRICE"], axis = 1)
         y_1=DF.PRICE
```

```
In [37]: x_1
```

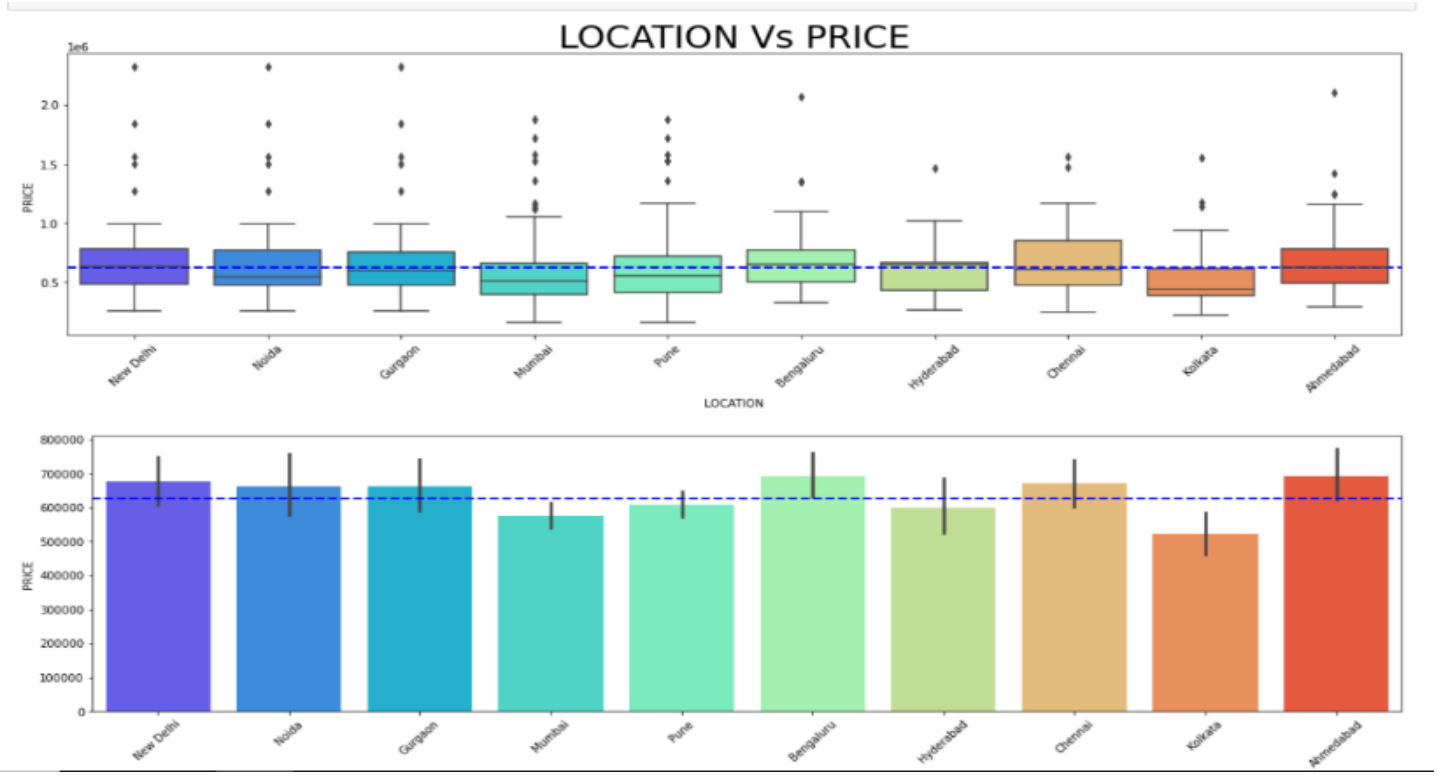
Out[37]:

	LOCATION	MNF_YEAR	BRAND	MODEL	VARIANT	DRIVEN_KM	FUELTYPE	NOOF_OWNERS
0	7	9	3	107	1	21169	0	0
1	7	10	5	228	0	7676	1	0
2	7	9	8	241	1	40458	0	0
3	7	9	8	241	1	82601	0	0
4	7	9	5	78	1	39294	1	1
...	...	...	...	...	...	...	...	...
1180	0	11	8	10	1	11061	1	0
1181	0	9	0	1	0	39254	0	1
1182	0	9	6	70	1	57010	0	0
1183	0	3	15	138	1	184782	0	2
1184	0	10	6	69	1	60802	0	0

883 rows × 8 columns

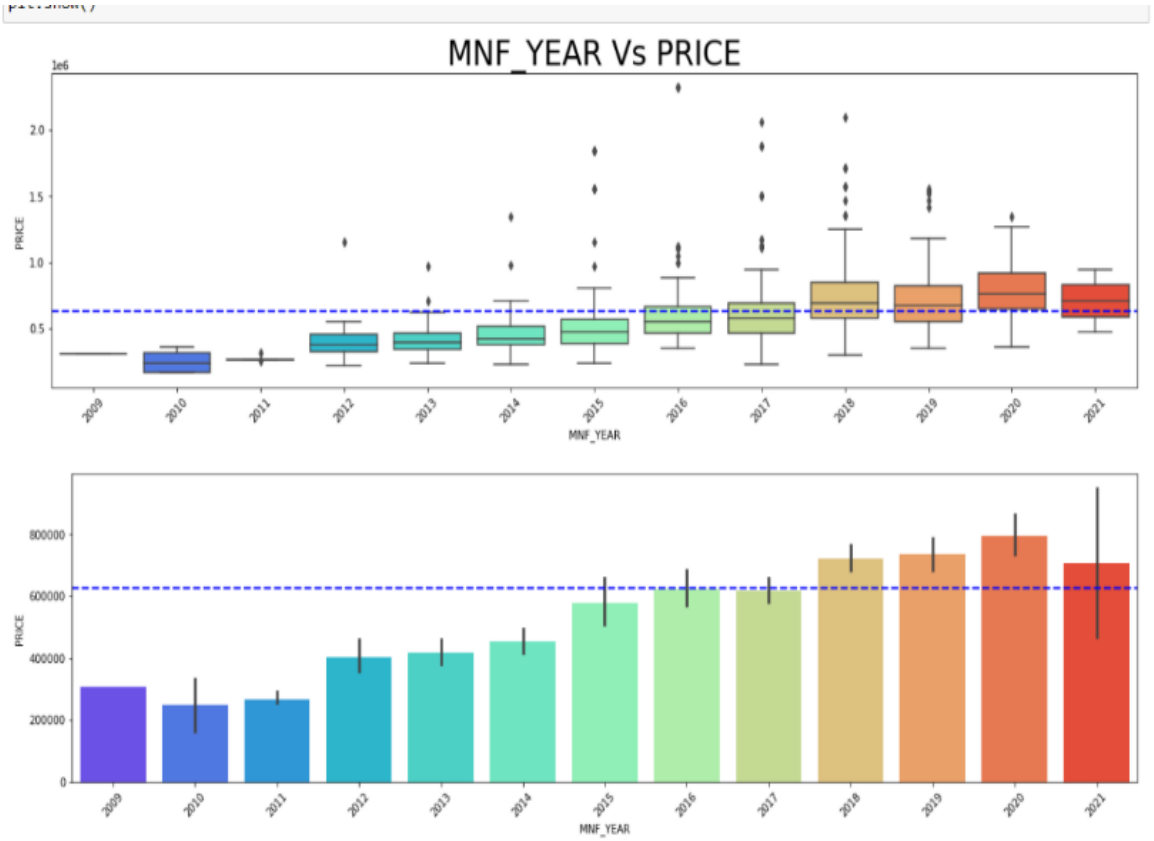
Data Inputs- Logic- Output Relationships

I have done some visualizations to understand the input output logic of the data collected.



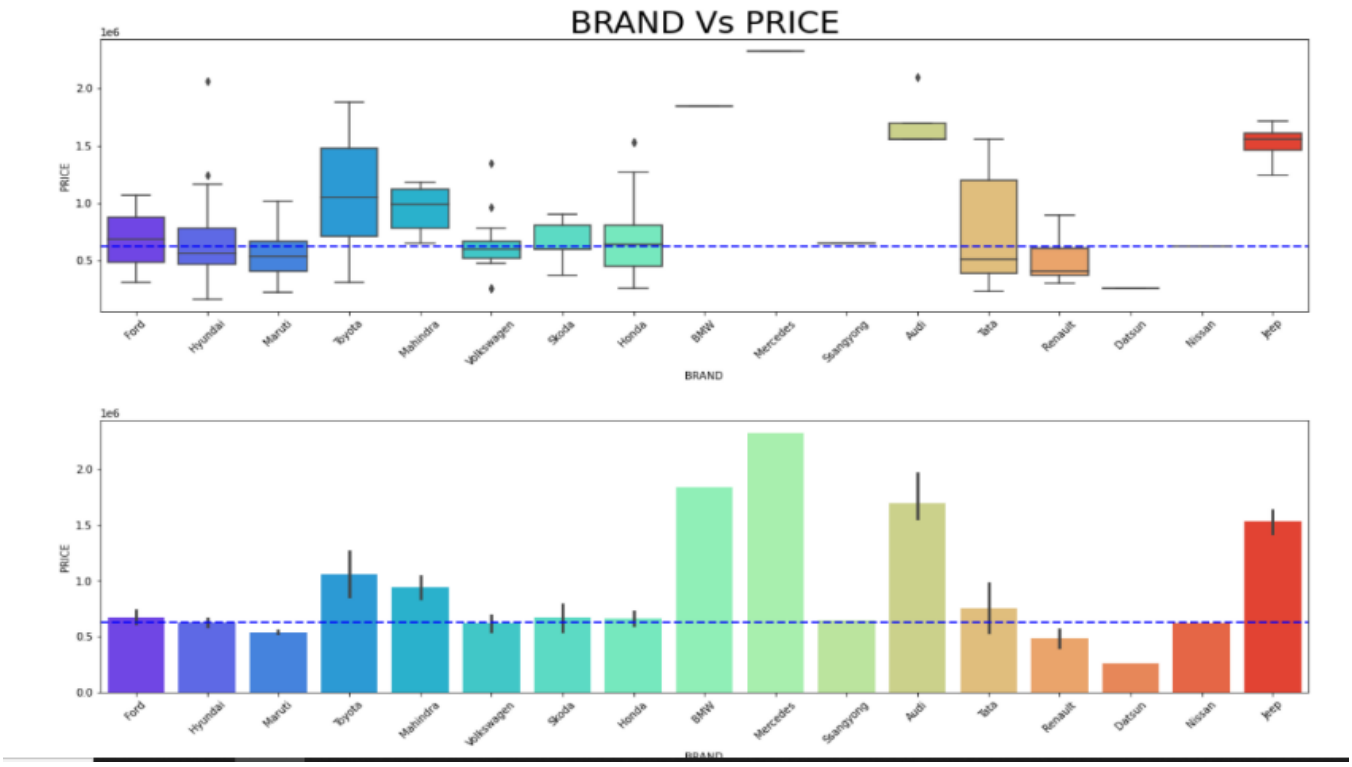
Key observations:

1. New Delhi, Noida, Gurgaon have the costliest cars and Mumbai, Pune, Ahmedabad have most cars being sold.
2. We have Bengaluru, Hyderabad, Chennai, Kolkata has least cars being sold and also comparatively cheaper.



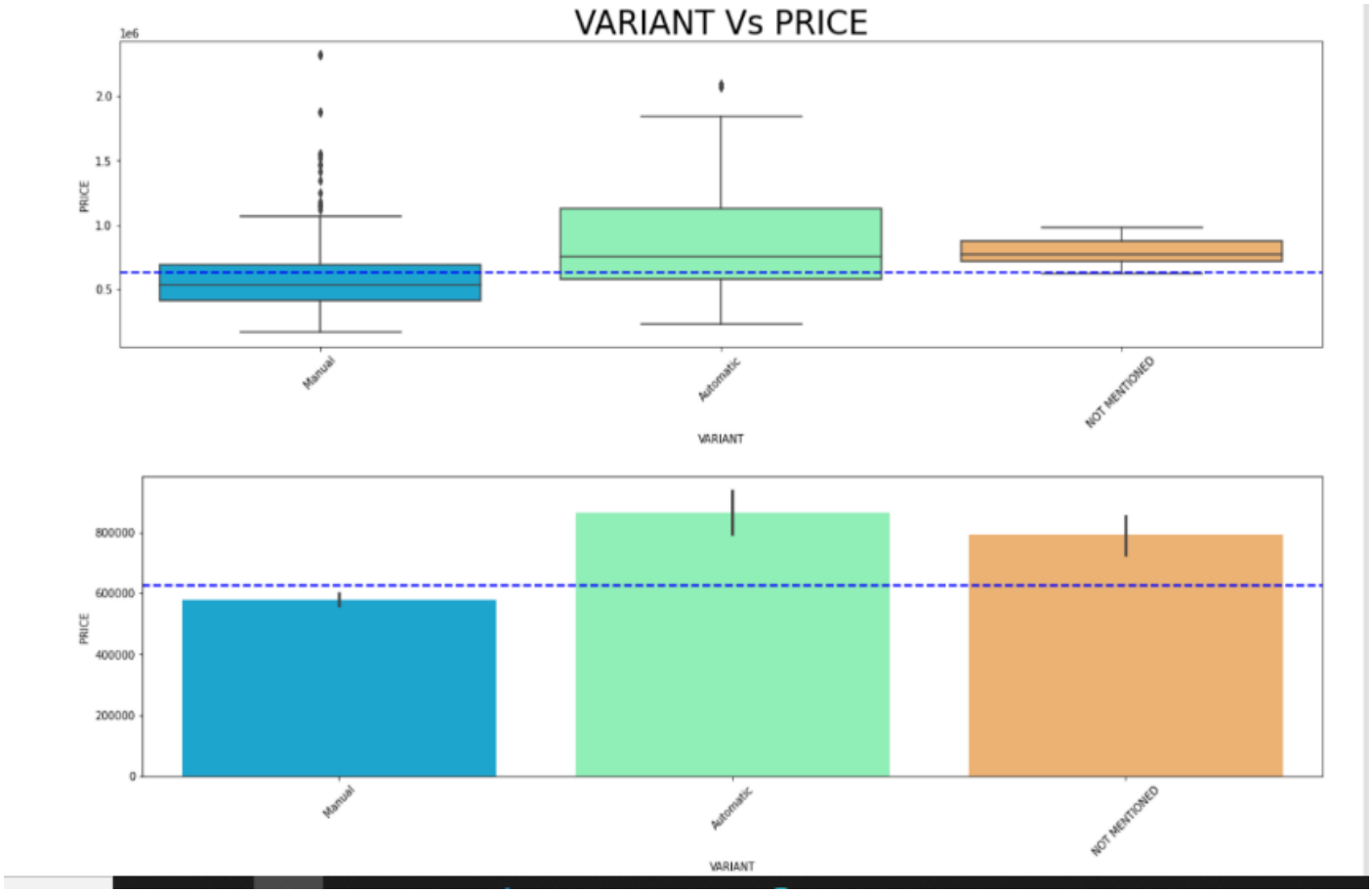
Key observations:

1. 2019, 2020, 2021 model are being sold higher in PRICE and also above average PRICE.
2. Above we can understand that cars sold in lesser kms driven and also in lesser years used are sold in high price.



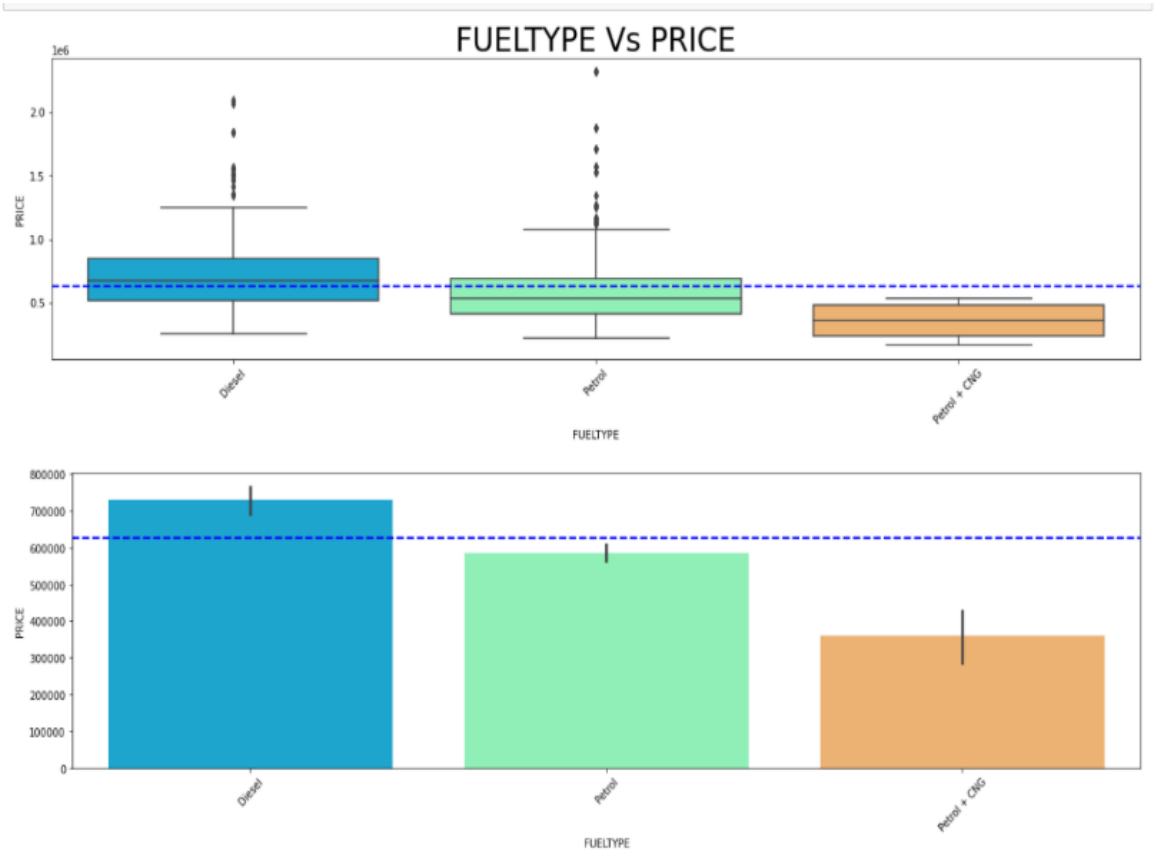
Key observations:

1. When comes to the Brand Land rover are being the costliest in country followed by Jaguar
2. And most of the other brands including most of the foreign brands are below the PRICE mean line



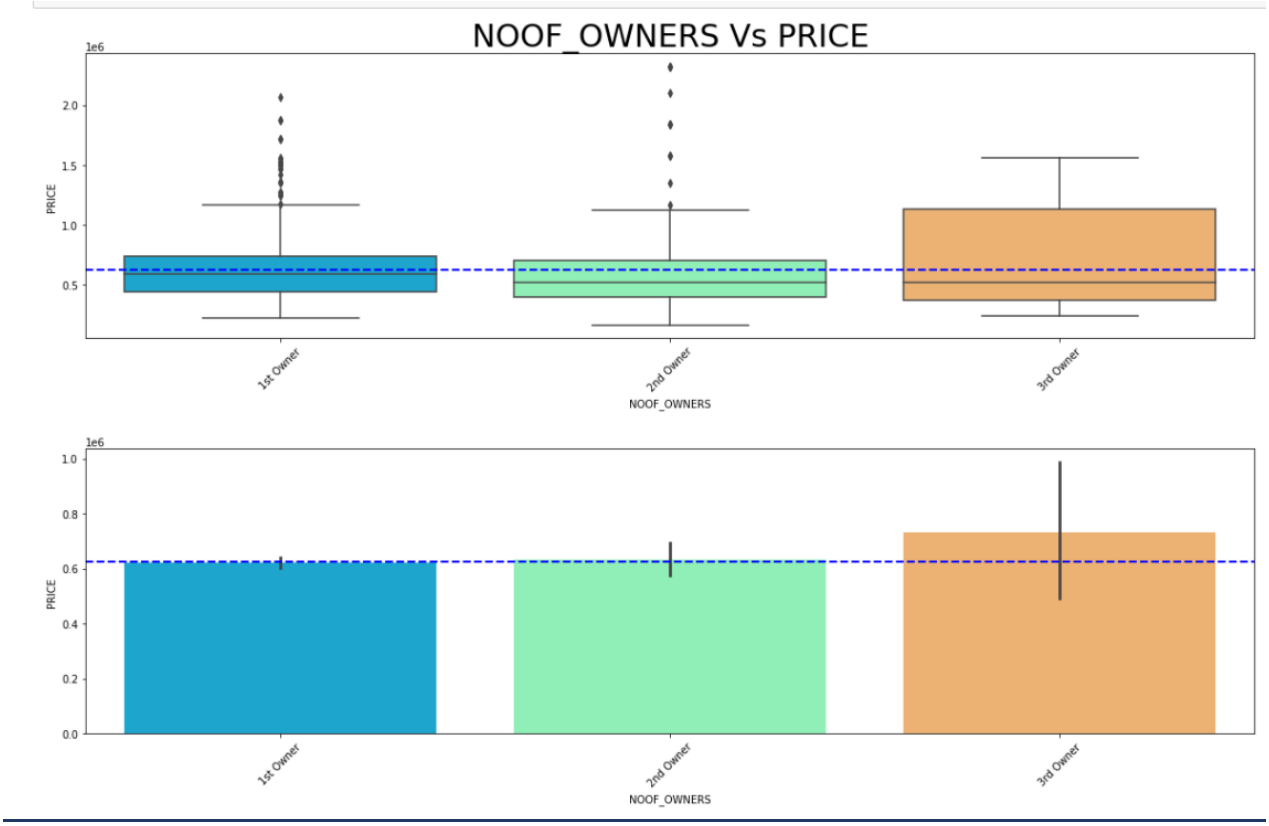
Key observations:

1. We can see that the automatic engines are costliest in the market. And also, most costlier cars come in Automatic shift.



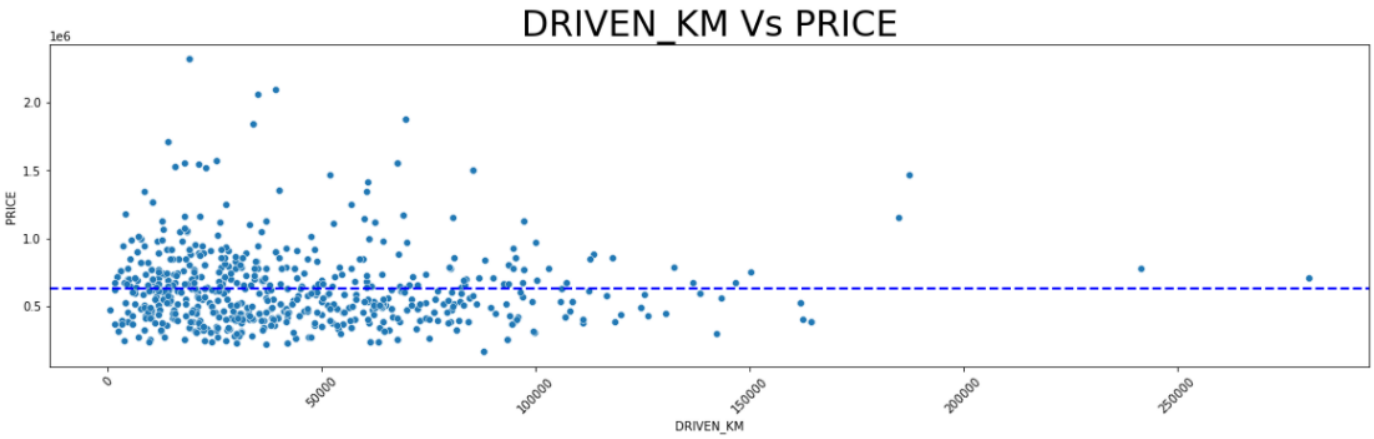
Key observations:

1. Diesel being the cheapest and most efficient fuel type, and so the Diesel engines are being the costliest fuel types.
2. Petrol bend second preferred followed by CNG and LPG fuel types.



Key observation:

1st owner cars are costliest followed by second and third.



Key Observations:

The lesser kms driven are evidently sold costlier.

Hardware and Software Requirements and Tools Used

1. Python 3.8.
2. NumPy.
3. Pandas.
4. Matplotlib.
5. Seaborn. 6. Data science.
6. SciPy
7. Sklearn.
8. Anaconda Environment, Jupyter Notebook.

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods).

Considering the business requirement of the client, I have collected the precise data to predict the used car price but there where multiple data of the car that are available. Data like colour of the car, sun roof attached, music system brand, electronics in the car, tyre brands, seat colours and much more. But after analysing all these data I have selected the data that have more correlation with the price of the car. Data like manufacturing year, number of owners used before, mode, fuel variant, gear shift variant, Brand of the car. I experimented and visualized how these variables contributed more towards the deciding factor of the car price. Based on such visualization I have built the model.

Testing of Identified Approaches (Algorithms)

After the pre-processing of the data that is collected, I have split the data as x_1 and y_1 and I have imported the required libraries to train my model. ##

In [36]:
```
x_1=DF.drop(["PRICE"], axis = 1)
y_1=DF.PRICE
```

Selecting parameters for training

In [39]:
```
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score, cross_val_predict, cross_validate
from sklearn.metrics import r2_score, mean_absolute_error,mean_squared_error

accu = 0
for i in range(0,1000):
    x_train_1, x_test_1, y_train_1, y_test_1 = train_test_split(x_1,y_1,test_size = .25, random_state = i)
    mod = LinearRegression()
    mod.fit(x_train_1,y_train_1)
    y_pred_1 = mod.predict(x_test_1)
    tempacc = r2_score(y_test_1,y_pred_1)
    if tempacc> accu:
        accu= tempacc
        best_rstate=i

print(f"Best Accuracy {accu*100} found on randomstate {best_rstate}")
```

```
Best Accuracy 46.22071850505682 found on randomstate 451
```

In [40]:
```
x_train, x_test, y_train, y_test = train_test_split(x_1,y_1,test_size = .25, random_state = best_rstate)
```

In [41]:
```
from sklearn.linear_model import LinearRegression, Lasso, Ridge, ElasticNet
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor
```

On selecting the best random state parameter, I have used nine regression algorithms to train my model. Based on the best model score and best CV score I have selected Random Forest Regressor as the final model.

Run and evaluate selected models

I have used nine different regression algorithms to shortlist the best model.

Shortlisting the best model

```
In [42]: models = [LinearRegression(), Lasso(), Ridge(alpha=1, random_state=42), ElasticNet(), SVR(), KNeighborsRegressor(), DecisionTreeF

         model_names = ["LinearRegression", "Lasso", "Ridge", "ElasticNet", "SVR", "KNeighborsRegressor", "DecisionTreeRegressor", "AdaBoo
```

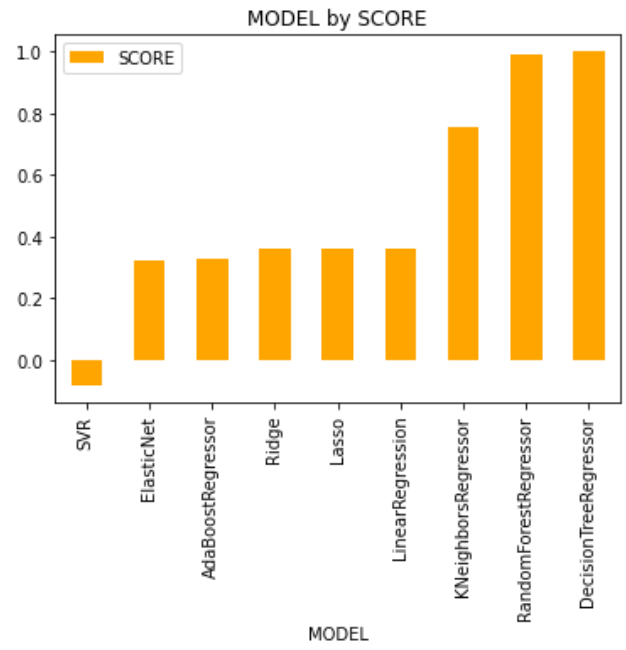
Cross validation mean score and the model score. ##

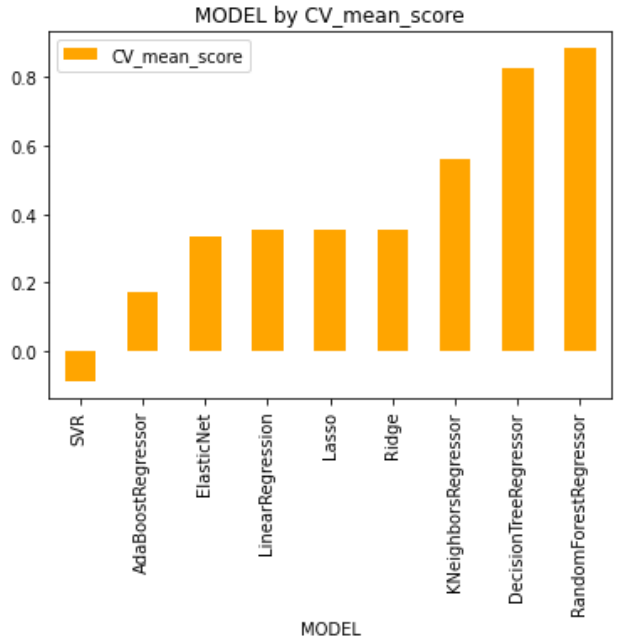
Out[45]:

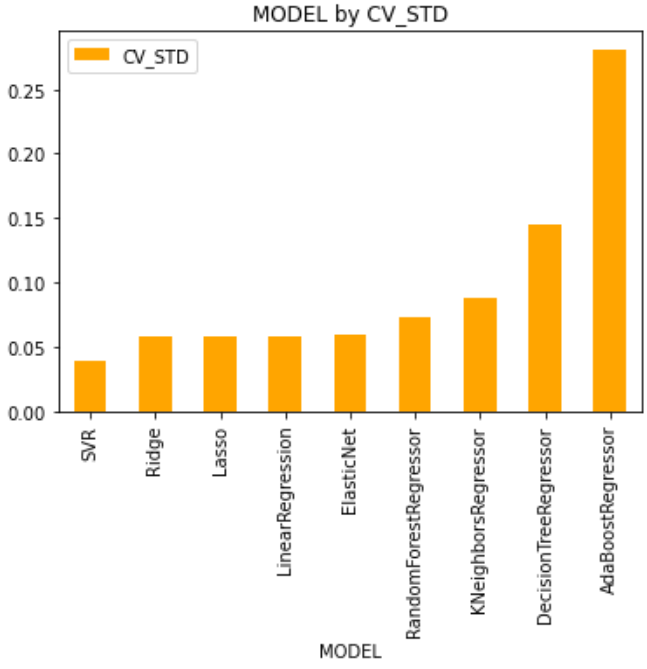
	MODEL	SCORE	CV_mean_score	CV_STD	MBE	MSE	RMSE	R2
8	RandomForestRegressor	0.975757	0.759224	0.147056	65272.110317	1.215593e+10	110253.938888	0.845412
6	DecisionTreeRegressor	1.000000	0.678654	0.292954	62931.212670	2.005406e+10	141612.342280	0.744970
7	AdaBoostRegressor	0.607379	0.492784	0.122012	152785.663087	3.790838e+10	194700.750053	0.517915
5	KNeighborsRegressor	0.601392	0.369354	0.131639	144375.382805	4.443459e+10	210795.129951	0.434921
2	Ridge	0.276441	0.270238	0.073170	153270.273518	4.229475e+10	205656.877072	0.462133
1	Lasso	0.276442	0.270139	0.073228	153317.120356	4.228892e+10	205642.690206	0.462207
0	LinearRegression	0.276442	0.270137	0.073229	153317.621135	4.228893e+10	205642.732718	0.462207
3	ElasticNet	0.239197	0.246453	0.059093	156908.329520	4.807144e+10	219251.995990	0.388670
4	SVR	-0.038704	-0.062426	0.067144	203659.116682	8.153680e+10	285546.495497	-0.036912

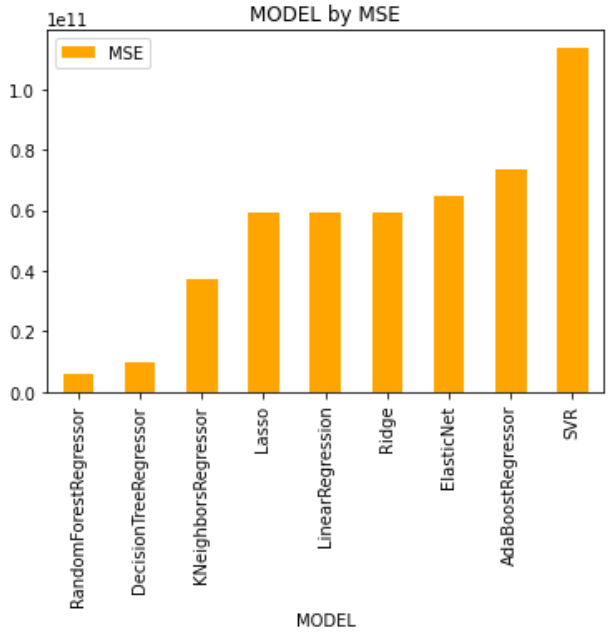
As live saw above Random forest Regression model stands at the top with the model score of 98.76 with the CV score of 88.54 further I am going to hyperparameter tune the model to reduce over fitting and to increase the performance of the model.

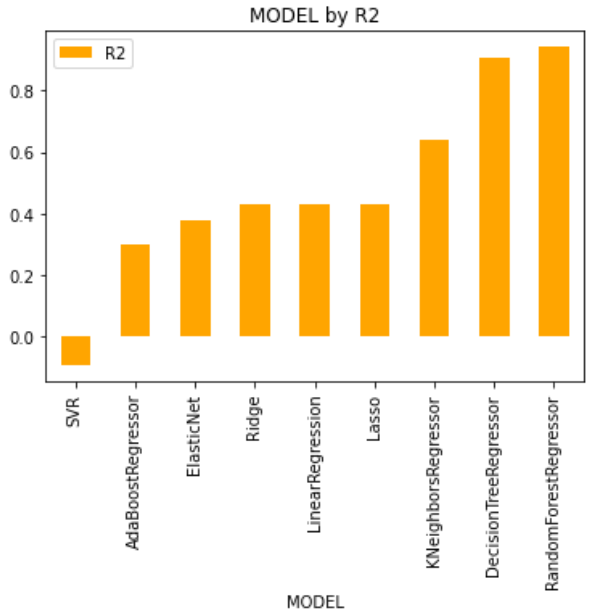
Model selecting Visualization:

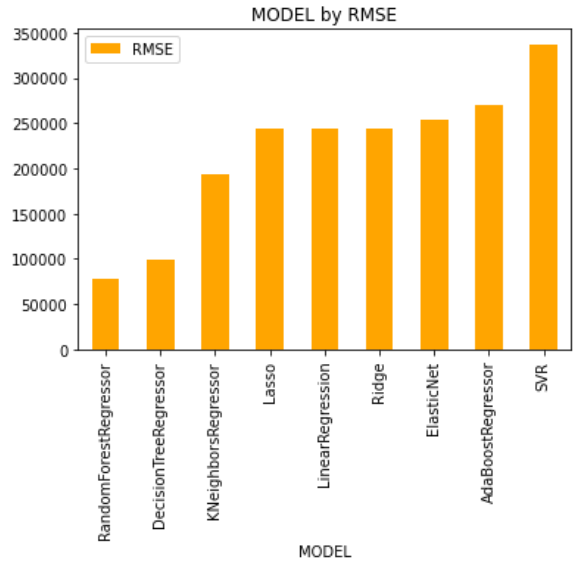












From above observation, we can come to a conclusion that Random Forest is the best model with Score of 98.78 let's try in Hyper tuning the same for improved performance and also to reduce the over fitting the Data.

Interpretation of the Results

From the visualization above we can clearly understand that the used car price factors are decided by the factors such as brand, location, model, year made, number of owners used the car before, fuel type of the car.

From that we can clearly say that the used car price depending on the Brand that is the manufacturer and model it varies. The manufacturer like Land Rover, Benz, BMW cars are costliest used car in the market comparatively to other cars, the low kilometres driven and also if the manufacturing year is lesser on these brands those card sells in much higher rates or closest to the buying new car rates. The Diesel variant and Automatic shift variants are also costliest user car variants in the used car market

CONCLUSION

Key Findings and Conclusions of the Study

The manufacturer like Land Rover, Benz, BMW cars are costliest used car in the market comparatively to other cars, the low kilometres driven and also if the manufacturing year is lesser on these brands those card sells in much higher rates or closest to the buying new car rates. The Diesel variant and Automatic shift variants are also costliest user car variants in the used car market.

Learning Outcomes of the Study in respect of Data Science

The above research will help our client to study about the latest used car market and with the help of the model built he can easily predict the price ranges of the cars, and also will helps him to understand based on what factors the Car Price is decided.

Limitations of this work and Scope for Future Work

The limitation of the study is that in the volatile changing market we have taken the data, to be more precise we have taken the data at the time of pandemic, so when the pandemic ends the market correction might happen slowly. So based on that again the deciding factors of the used car prize might change and we have shortlisted and taken these data from the important cities across India, if the seller is from the different city our model might fail to predict the accurate prize of that used car.