FLIP ROBO

Car Price Prediction

Submitted by:

VARSHA PATANKAR

ACKNOWLEDGMENT

Various references were used for the implementation of this project. Different research papers by experts were referred for the best method and for the best result. Different sites also referred for this project which are useful for data science projects like github, kaggle, researchgate and so on.

INTRODUCTION

- Business Problem Framing

  With the covid 19 impact in the market, we have seen 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. Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities.

- Conceptual Background of the Domain Problem

  For this project, we are using the dataset on used car sales from all over the cities. Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market.

- Review of Literature

  In this the reviewed studies were critically focused towards applying data mining techniques for the prediction and classification of car price predictions. Different data mining techniques were applied for the proposed model and their performances were evaluated on various parameters. Based on these parameters the best methods was selected, explained and suggested because of its characteristics regarding the prediction of the prices.

- Motivation for the Problem Undertaken

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

The general problem one faces in the modelling is that to find best suitable algorithm and mathematical and statistical approach for our model. In real world the datasets are very large, noisy and unstructured. So to find the best strategy for our model we need such approach that combines the statistics and machine learning together.

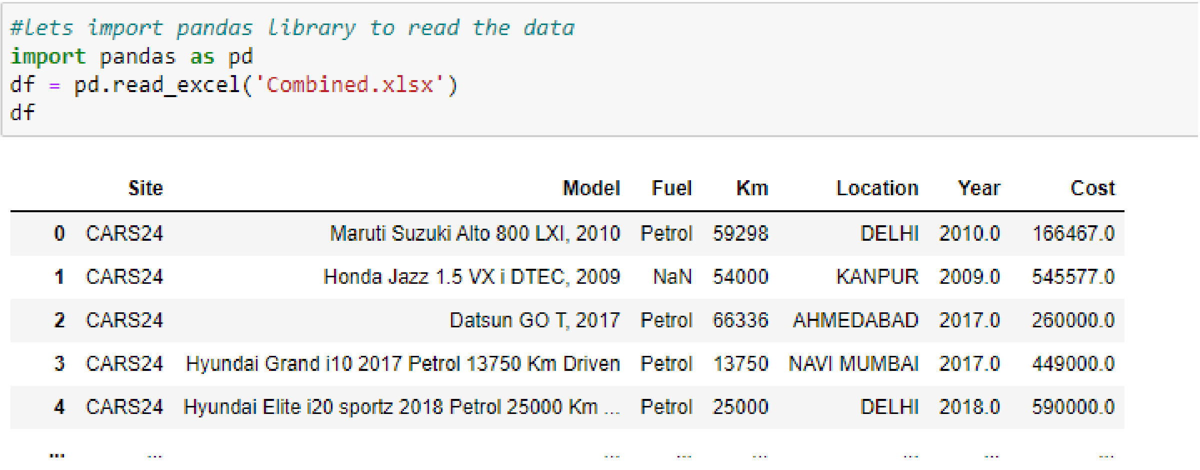
Statisticians are using so many approaches to find the best accuracy and we are also trying some of the machine learning algorithms along with the statistics. So while using those methods we will came to know that which algorithm has limitations and which algorithm is best fitted for our model.

- Data Sources and their formats

There are so many sources are available on the internet from which can get the data on which we have to work on. So for this project you can get the data from github site, kaggle site, researchgate etc.

The data you found can be in any form i.e. csv file, excel file etc.

Here I have got the data in the form of excel file which we have to import that file using pandas library. Below is the snapshot of data and how I have load the data for further process.



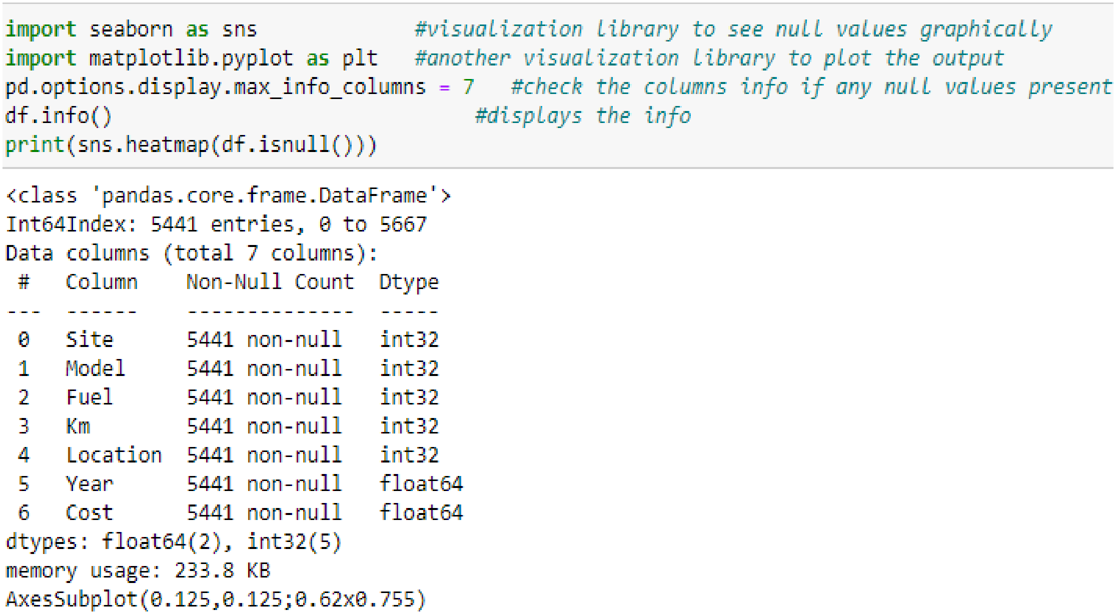
```
#lets import pandas library to read the data
import pandas as pd
df = pd.read_excel('Combined.xlsx')
df
```

	Site	Model	Fuel	Km	Location	Year	Cost
0	CARS24	Maruti Suzuki Alto 800 LXI, 2010	Petrol	59298	DELHI	2010.0	166467.0
1	CARS24	Honda Jazz 1.5 VX i DTEC, 2009	NaN	54000	KANPUR	2009.0	545577.0
2	CARS24	Datsun GO T, 2017	Petrol	66336	AHMEDABAD	2017.0	260000.0
3	CARS24	Hyundai Grand i10 2017 Petrol 13750 Km Driven	Petrol	13750	NAVI MUMBAI	2017.0	449000.0
4	CARS24	Hyundai Elite i20 sportz 2018 Petrol 25000 Km ...	Petrol	25000	DELHI	2018.0	590000.0
...	...	...	...	...	...	...	...

- Data Pre-processing Done

There is some process that we have to do while pre-processing. Following are the process that we have used-

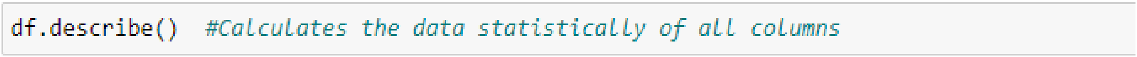
1. **Data Wrangling:** Our dataset is not always in proper format. Sometimes there maybe some missing values, sometimes there must be out of the box values. Our data is always a raw data so for that we have to do some wrangling process like.
   a. Check if there are any null values are present in the dataset or not.
   b. Check the data type of each column.



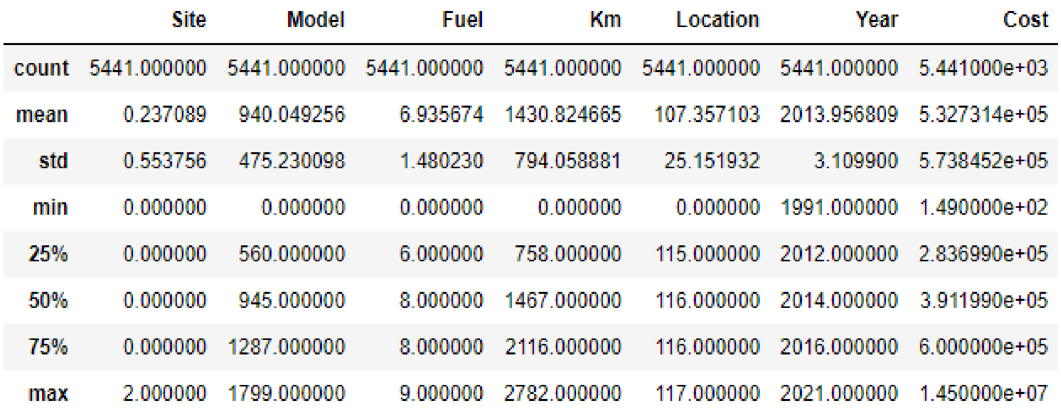
```
import seaborn as sns           #visualization library to see null values graphically
import matplotlib.pyplot as plt #another visualization library to plot the output
pd.options.display.max_info_columns = 7   #check the columns info if any null values present
df.info()                                 #displays the info
print(sns.heatmap(df.isnull()))
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5441 entries, 0 to 5667
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Site      5441 non-null   int32  
 1   Model     5441 non-null   int32  
 2   Fuel      5441 non-null   int32  
 3   Km        5441 non-null   int32  
 4   Location  5441 non-null   int32  
 5   Year      5441 non-null   float64
 6   Cost      5441 non-null   float64
dtypes: float64(2), int32(5)
memory usage: 233.8 KB
AxesSubplot(0.125,0.125;0.62x0.755)
```

c. Check the summary of our dataset using



```
df.describe()  #Calculates the data statistically of all columns
```

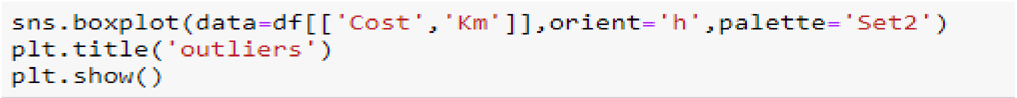


	Site	Model	Fuel	Km	Location	Year	Cost
count	5441.000000	5441.000000	5441.000000	5441.000000	5441.000000	5441.000000	5.441000e+03
mean	0.237089	940.049256	6.935674	1430.824665	107.357103	2013.956809	5.327314e+05
std	0.553756	475.230098	1.480230	794.058881	25.151932	3.109900	5.738452e+05
min	0.000000	0.000000	0.000000	0.000000	0.000000	1991.000000	1.490000e+02
25%	0.000000	560.000000	6.000000	758.000000	115.000000	2012.000000	2.836990e+05
50%	0.000000	945.000000	8.000000	1467.000000	116.000000	2014.000000	3.911990e+05
75%	0.000000	1287.000000	8.000000	2116.000000	116.000000	2016.000000	6.000000e+05
max	2.000000	1799.000000	9.000000	2782.000000	117.000000	2021.000000	1.450000e+07

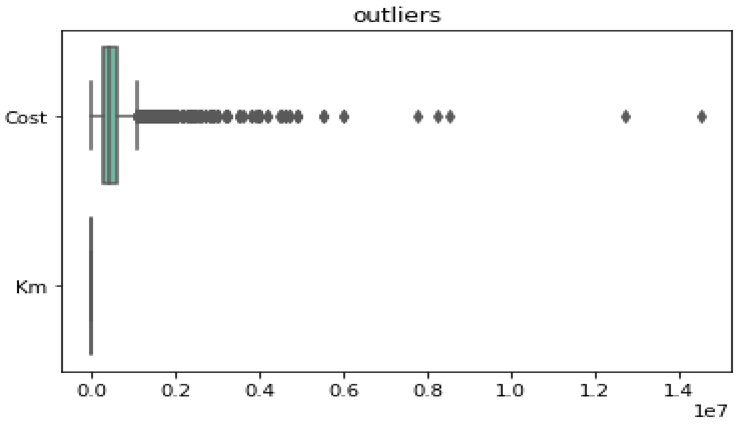
From above we can say that all values for each column is different.

1. For Location column max value is 117 and 75% value is 116. So we can say that in such case data is ok.
2. But for the Model column max value is 1799 and 75% value is 1287 so in this case we can say that some outliers are present in this column.

Checking Outliers: Our data is a raw data so we have to check whether any value is out of the box or not. If any outliers are present in any column, so in that case we have remove that particular value from our dataset. Otherwise it will give us wrong predictions.

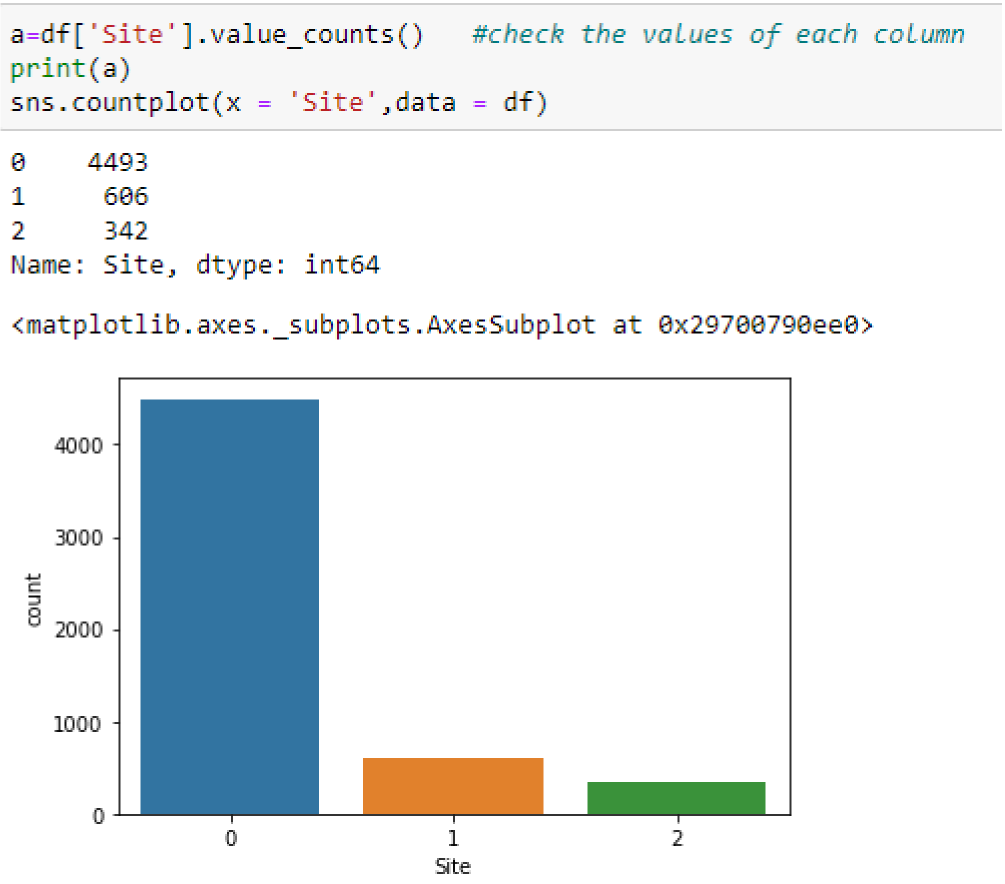


```
sns.boxplot(data=df[['Cost','Km']],orient='h',palette='Set2')
plt.title('outliers')
plt.show()
```



- Data Inputs- Logic- Output Relationships

  In this step we have to do the analysis of each column using univariate method. And analysis of each column with target variable using bivariate method. So for this I have used Count method to check the number of unique values present in that particular column and also some visualization libraries to understand our data more clearly.

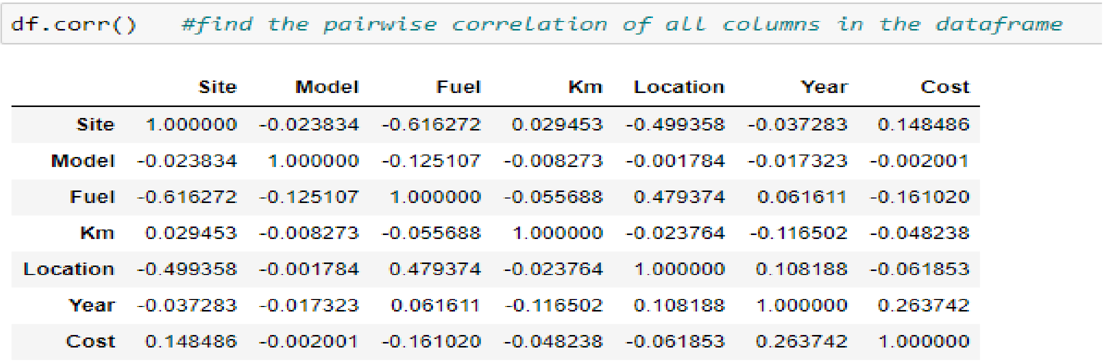


```
a=df['Site'].value_counts()   #check the values of each column
print(a)
sns.countplot(x = 'Site',data = df)
```

```
0    4493
1     606
2     342
Name: Site, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x29700790ee0>
```

Now we have to find the correlation between all variables.

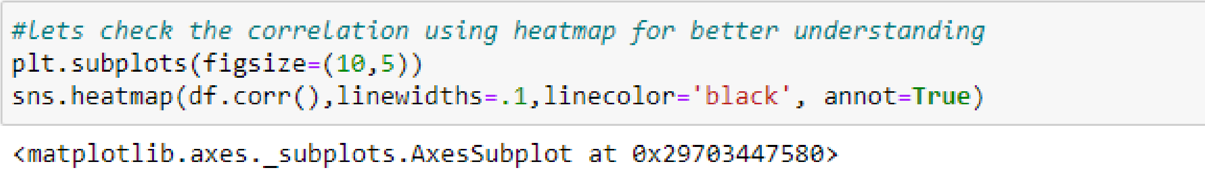


```
df.corr()    #find the pairwise correlation of all columns in the dataframe
```

	Site	Model	Fuel	Km	Location	Year	Cost
Site	1.000000	-0.023834	-0.616272	0.029453	-0.499358	-0.037283	0.148486
Model	-0.023834	1.000000	-0.125107	-0.008273	-0.001784	-0.017323	-0.002001
Fuel	-0.616272	-0.125107	1.000000	-0.055688	0.479374	0.061611	-0.161020
Km	0.029453	-0.008273	-0.055688	1.000000	-0.023764	-0.116502	-0.048238
Location	-0.499358	-0.001784	0.479374	-0.023764	1.000000	0.108188	-0.061853
Year	-0.037283	-0.017323	0.061611	-0.116502	0.108188	1.000000	0.263742
Cost	0.148486	-0.002001	-0.161020	-0.048238	-0.061853	0.263742	1.000000

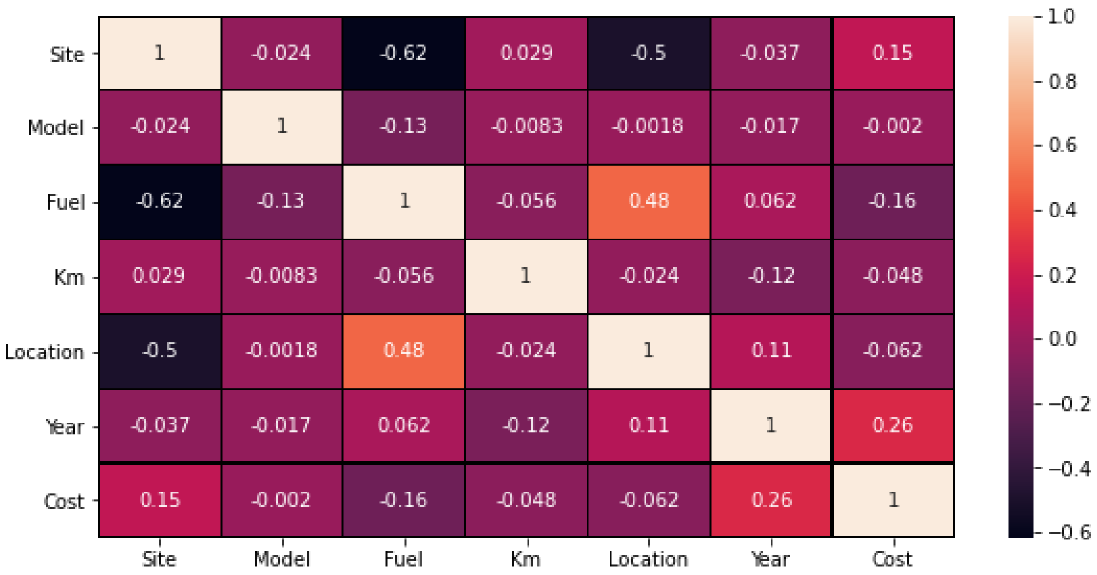
- State the set of assumptions (if any) related to the problem under consideration

  This above method is used to check the correlation between all columns. Heatmap method gives the graphical visualization from which we can easily understand the correlation. So from the diagram below we can say that-



```
#lets check the correlation using heatmap for better understanding
plt.subplots(figsize=(10,5))
sns.heatmap(df.corr(),linewidths=.1,linecolor='black', annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x29703447580>
```



	Site	Model	Fuel	Km	Location	Year	Cost
Site	1	-0.024	-0.62	0.029	-0.5	-0.037	0.15
Model	-0.024	1	-0.13	-0.0083	-0.0018	-0.017	-0.002
Fuel	-0.62	-0.13	1	-0.056	0.48	0.062	-0.16
Km	0.029	-0.0083	-0.056	1	-0.024	-0.12	-0.048
Location	-0.5	-0.0018	0.48	-0.024	1	0.11	-0.062
Year	-0.037	-0.017	0.062	-0.12	0.11	1	0.26
Cost	0.15	-0.002	-0.16	-0.048	-0.062	0.26	1

Some columns are making good positive correlation with our target variable and some has negative correlation. So the variables that doesn’t make any good correlation with any variable so we to drop that column from the dataset.

- Hardware and Software Requirements and Tools Used

  No external hardware required for this project. Only you have to do in laptop.

  Jupyter notebook with any latest version is required as a software.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Machine learning helps to solve the all above problems. It will not only used for the recent data but also helpful to predict the future outcomes with the help of past outcomes.

Machine learning algorithms help to create a new model using existing and historical data that we can use for the training and testing our model to predict the future outcome.

In this project we have used various machine learning algorithms to predict the proper outcome which gives as much as accuracy for our model.

- Testing of Identified Approaches (Algorithms)

Following algorithms we have used for our model which gives better performance for the dataset.

1. Linear regression – Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

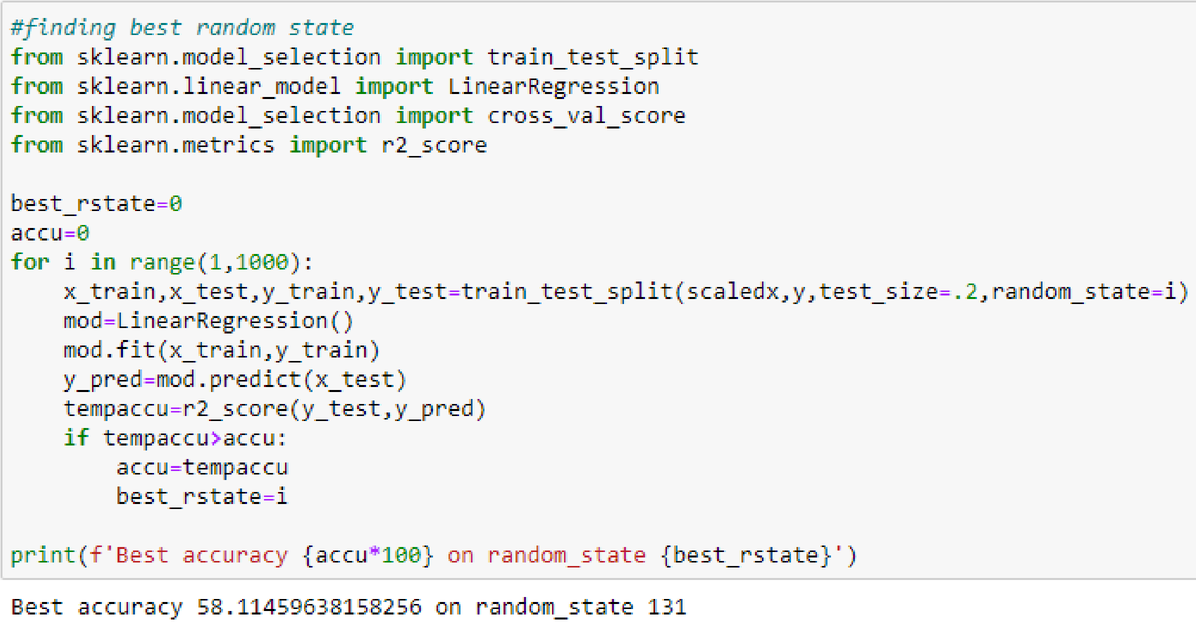
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable

2. RandomForestRegressor – A random forest is a meta estimator that fits a number of classifying decision trees on various sub-

samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

- Run and Evaluate selected models

Following are the methods used for our model-



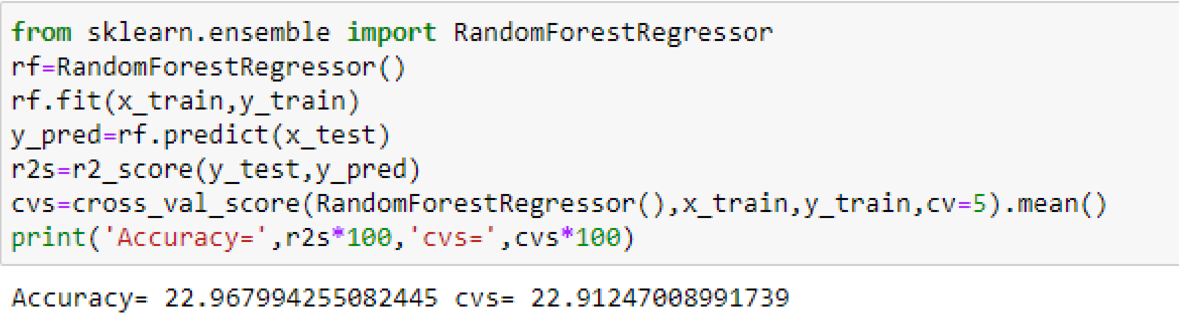
```
#finding best random state
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
from sklearn.metrics import r2_score

best_rstate=0
accu=0
for i in range(1,1000):
    x_train,x_test,y_train,y_test=train_test_split(scaledx,y,test_size=.2,random_state=i)
    mod=LinearRegression()
    mod.fit(x_train,y_train)
    y_pred=mod.predict(x_test)
    tempaccu=r2_score(y_test,y_pred)
    if tempaccu>accu:
        accu=tempaccu
        best_rstate=i

print(f'Best accuracy {accu*100} on random_state {best_rstate}')
```

```
Best accuracy 58.11459638158256 on random_state 131
```

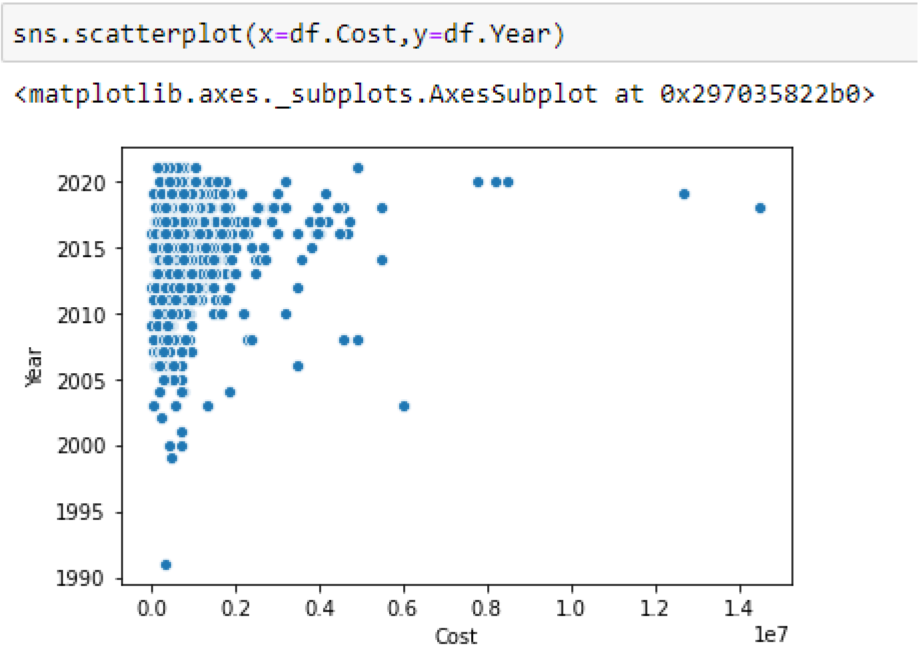
From above diagram we can say that linear regression gives the accuracy above 58%. We have check the accuracy for best random state that best suited to our model.



```
from sklearn.ensemble import RandomForestRegressor
rf=RandomForestRegressor()
rf.fit(x_train,y_train)
y_pred=rf.predict(x_test)
r2s=r2_score(y_test,y_pred)
cvs=cross_val_score(RandomForestRegressor(),x_train,y_train,cv=5).mean()
print('Accuracy=',r2s*100,'cvs=',cvs*100)
```

```
Accuracy= 22.967994255082445 cvs= 22.91247008991739
```

- Visualizations



```
sns.scatterplot(x=df.Cost,y=df.Year)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x297035822b0>
```

Seaborn and Matplotlib are two visualization libraries that I used to visualize graphically my data. This is the bivariate method because I have visualized two variable target variable with other columns. The same method I have used for remaining variables.

- Interpretation of the Results

From all above process we can say that whatever algorithms or method we used for prediction of our model all process with visualization is very beneficial. Using any visualization library it is easy for us to understand the all data.

CONCLUSION

- Key Findings and Conclusions of the Study

This study has proposed a comprehensive research and model development for the prediction of the used car prices.

For this project, we are using the dataset on used car sales from all over the cities. Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market.

1. In the experimentation phase two different machine learning algorithms were employed for the proposed model development and their performances were evaluated on various parameters.

2. From all above process we can say that whatever algorithms or method we used for prediction of our model all process with visualization is very beneficial. Using any visualization library it is easy for us to understand the all data.

3. For this data Linear regression proves the best and others we have used also has good result but while using Linear regression with different random state, it also gives the more than 58% accuracy.

Learning Outcomes of the Study in respect of Data Science

To cope up with this problem a comprehensive amount of literature was reviewed to study the significant factors that lead to such problems. Moreover, these reviewed studies were critically focused towards the employed techniques and methods of data mining for the prediction and classification of the used car prices. Based on the all parameters that we have used for the different algorithm the best method was selected.

- ## Limitations of this work and Scope for Future Work

Financial services hold a great amount of significance for any individual, business or enterprise. There must be always advancement in technology.

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for over fitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.