FLIP ROBO

CAR PRICE PREDICTION

Submitted by:

Rahul Adhwaria

ACKNOWLEDGMENT

The internship opportunity I had with Flip Robo Technology was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this internship period.

I express my deepest thanks to Shubham Yadav, SME of internship for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make work easier. I choose this moment to acknowledge his contribution gratefully.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

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. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model. This project contains two phase-

1. Data Collection Phase –
   You have to scrape at least 5000 used cars data. You can scrape more data as well, it's up to you. More the data better the model
   In this section You need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.)
   You need web scraping for this. You have to fetch data for different locations. The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

2. Model Building Phase-
   After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

- **Conceptual Background of the Domain Problem**
  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. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

- **Motivation for the Problem Undertaken**
  As this project belongs to real world problem, and working on such project will boost my knowledge in data science and machine learning field. And being on a starting phase of becoming data scientist I thought working on this project will help me a lot to make hands on data analysis, model building and data prediction. Also handling of such a big data will definitely enhanced my knowledge in data science field as per my expectation.

Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

  In our project we have numeric outputs, therefore here we used regression. In this project I used DecisionTree Regressor, RandomForest Regressor and AdaBoost Regressor. As a result I found RandomForest Regressor as the best algorithm with accuracy score of 97%.

- **Data Sources and their formats**

  The data is scraped from cartrader website to analyse the data and to work on its information, clearing of unwanted data and building model to improve the selection of used cars, the client wants some predictions that could help them in further investment and improvement in selection of used cars. In this we at the starting try to understand the features of dataset. In this we find the dimensions, type of data and null value if present or not.

- **Data Preprocessing Done**

  As we import our dataset, we first analyse the data and see all the features of our dataset. After this we find the null values in present dataset. After this we plot histogram graph, scatter plot and heat map graph (for correlation) to see the relationship between each other column.

- **Data Inputs- Logic- Output Relationships**
  We split our dataset into independent and target variable, as we have to find the best model, so then we divide our dataset into train and test data by 7:3 ratio respectively. We select 3 models DecisionTree Regressor, RandomForest Regressor and AdaBoost Regression. We train all these 3 model so that we can use these model in future to predicte the car price. The model which gives result with best accuracy is selected for future prediction.

- **State the set of assumptions (if any) related to the problem under consideration**
  As we are very much familiar to various algorithms, but maybe there is a very high chance that the best algorithm for the give dataset is something which is not used by me in this project. So in this project I assumed that the algorithm I have selected will give the best model but somehow there can be anyone who has used any other algorithm and getting accuracy more than my selected model (RandomForest Regressor).

Model/s Development and Evaluation

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

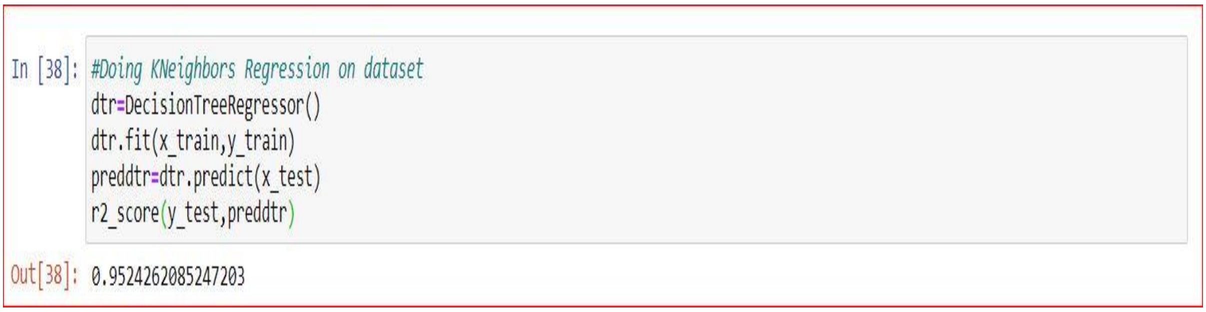
  In the starting as we import the dataset, firstly we did the basic look up on the features of our train dataset. For this we did EDA process, we find the correlation among column. We split column into target variable (Price) and independent variable (all other column). And we find the skewness of independent variable. Then we observed our output dataset, we found values are numerical. Then we train our dataset by using the 3 method and we also noted down the accuracy of all 3 model. Then we did the cross validation to find the over fitting or under fitting of any model. And by overall working we find the RandomForest Regressor gives the best result. Then we save our model for future use.

- **Testing of Identified Approaches (Algorithms)**
  1. DecisionTreeRegressor
  2. RandomForestRegressor
  3. AdaBoost Regressor

- **Run and Evaluate selected models**

  As in our target variable we have numerical values, and we know that this type of dataset falls under regression. So we used 3 algorithms to predict the test data i.e DecisionTree Regressor, RandomForest Regressor and AdaBoost Regressor

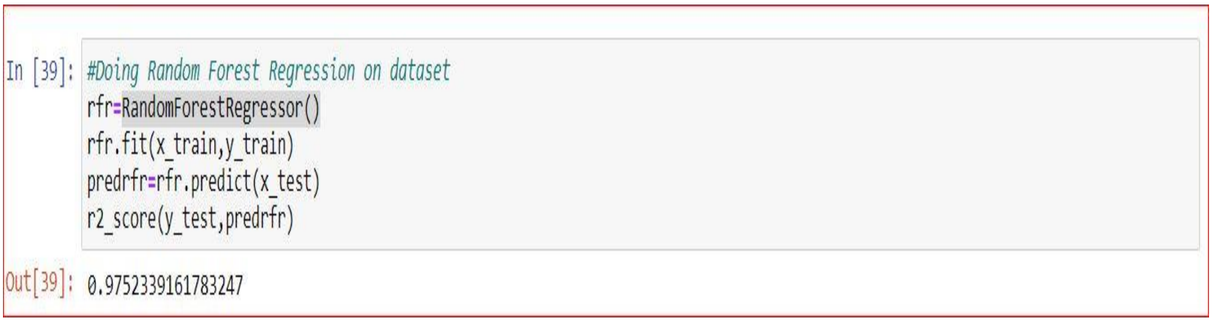
1. DecisionTreeRegressor : In this algorithm we are getting accuracy of 95%.



```
In [38]: #Doing KNeighbors Regression on dataset
         dtr=DecisionTreeRegressor()
         dtr.fit(x_train,y_train)
         preddtr=dtr.predict(x_test)
         r2_score(y_test,preddtr)

Out[38]: 0.9524262085247203
```

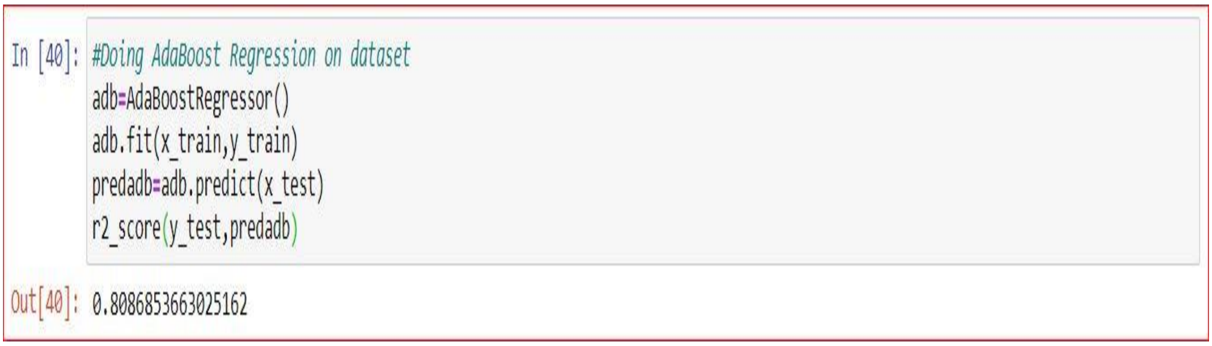
2. RandomForestRegressor : In this algorithm we are getting accuracy of 97%.



```
In [39]: #Doing Random Forest Regression on dataset
         rfr=RandomForestRegressor()
         rfr.fit(x_train,y_train)
         predrfr=rfr.predict(x_test)
         r2_score(y_test,predrfr)

Out[39]: 0.9752339161783247
```

3. AdaBoost Regressor: In this algorithm we are getting accuracy of 80%.



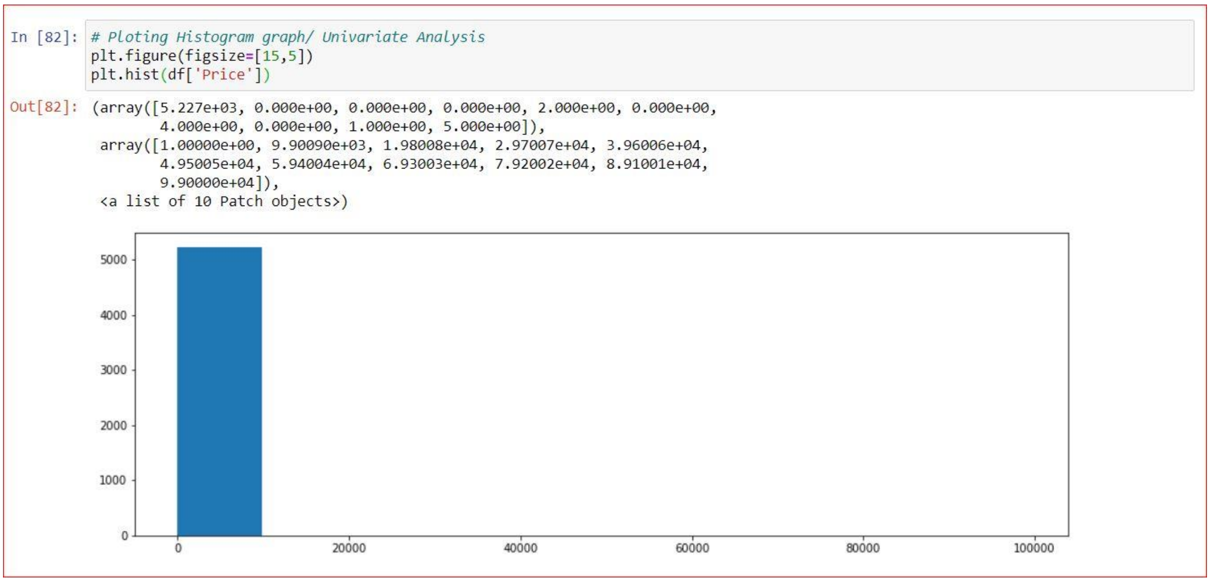
```
In [40]: #Doing AdaBoost Regression on dataset
         adb=AdaBoostRegressor()
         adb.fit(x_train,y_train)
         predadb=adb.predict(x_test)
         r2_score(y_test,predadb)

Out[40]: 0.8086853663025162
```

- **Visualizations**

  1. Univariate Analysis

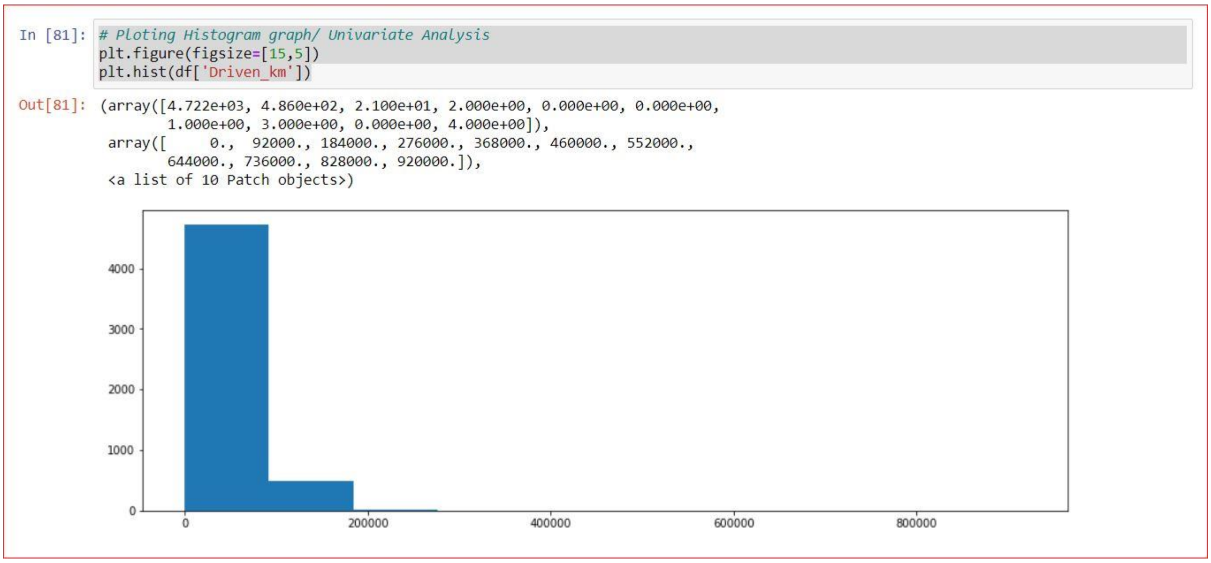
➢ From this garph we can observe that Price of cars in our dataset. The value of price is in lakhs.



```
In [82]: # Ploting Histogram graph/ Univariate Analysis
         plt.figure(figsize=[15,5])
         plt.hist(df['Price'])
```

```
Out[82]: (array([5.227e+03, 0.000e+00, 0.000e+00, 0.000e+00, 2.000e+00, 0.000e+00,
                 4.000e+00, 0.000e+00, 1.000e+00, 5.000e+00]),
          array([1.00000e+00, 9.90090e+03, 1.98008e+04, 2.97007e+04, 3.96006e+04,
                 4.95005e+04, 5.94004e+04, 6.93003e+04, 7.92002e+04, 8.91001e+04,
                 9.90000e+04]),
          <a list of 10 Patch objects>)
```

➢ We can observe that in this graph driven_km is maximum from 0 to 100000 km in our data and above 200000 is minimum.

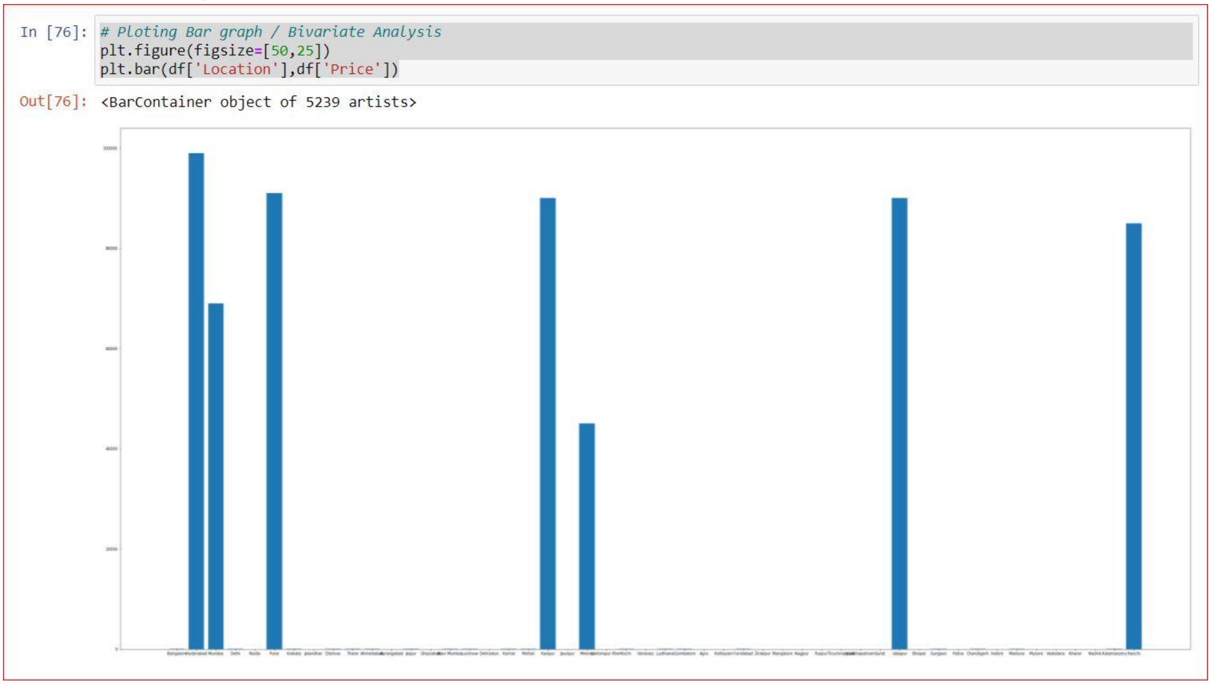


```
In [81]: # Ploting Histogram graph/ Univariate Analysis
         plt.figure(figsize=[15,5])
         plt.hist(df['Driven_km'])
```

```
Out[81]: (array([4.722e+03, 4.860e+02, 2.100e+01, 2.000e+00, 0.000e+00, 0.000e+00,
                 1.000e+00, 3.000e+00, 0.000e+00, 4.000e+00]),
          array([     0.,  92000., 184000., 276000., 368000., 460000., 552000.,
                 644000., 736000., 828000., 920000.]),
          <a list of 10 Patch objects>)
```

2. Bivariate analysis

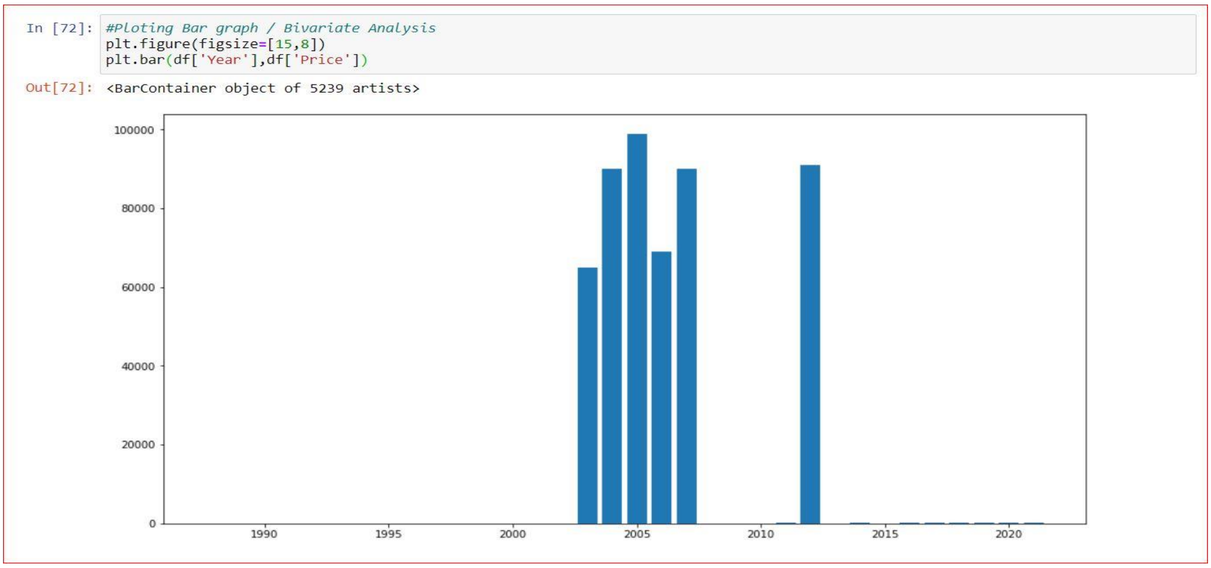
From this graph we can observe that only few city have maximum saling of cars, excluding those we have very less cities in which cars are solding. And counting of such cities are maximum in which cars are solding.



```
In [76]: # Ploting Bar graph / Bivariate Analysis
         plt.figure(figsize=[50,25])
         plt.bar(df['Location'],df['Price'])

Out[76]: <BarContainer object of 5239 artists>
```

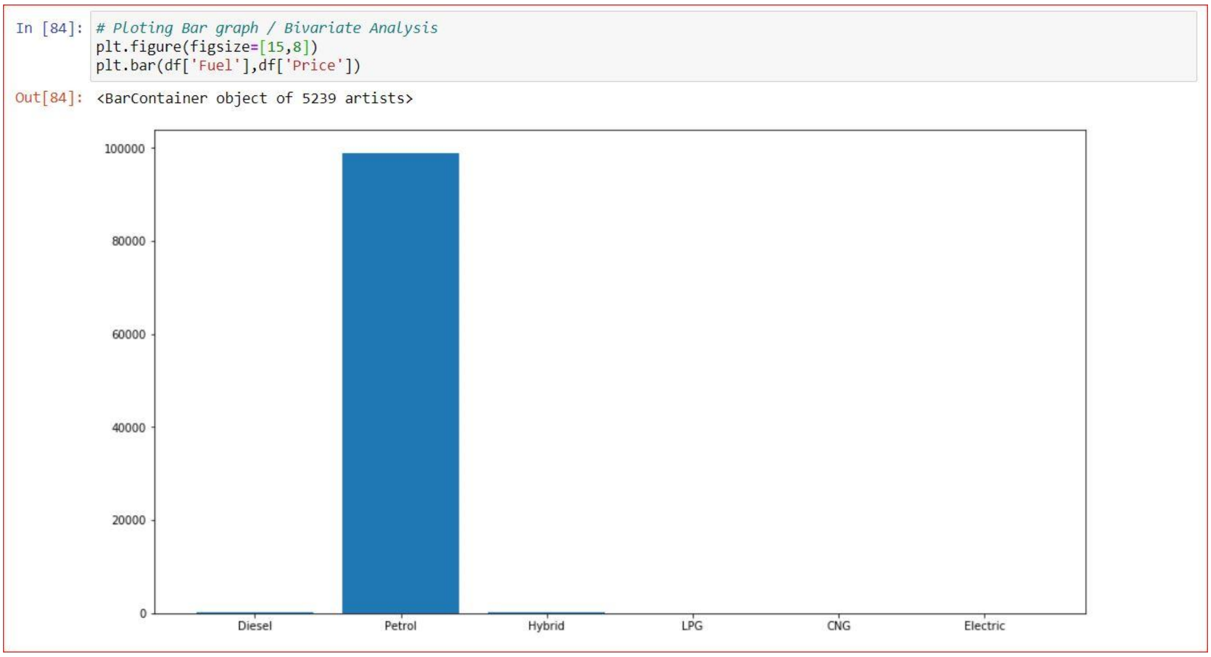
- From this graph we can conclude that car manufactured in 2003 to 2007 is more otherwise recent manufactured car in last 5-6 years are solding less.



```
In [72]: #Ploting Bar graph / Bivariate Analysis
         plt.figure(figsize=[15,8])
         plt.bar(df['Year'],df['Price'])

Out[72]: <BarContainer object of 5239 artists>
```

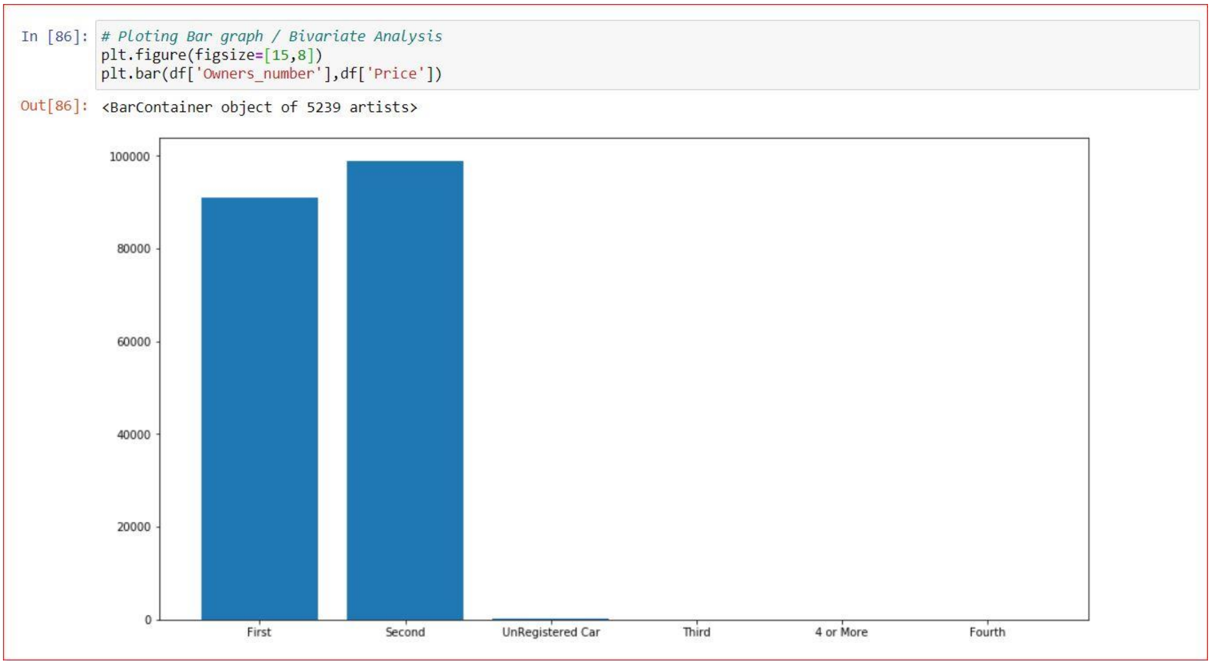
- From this graph we can conclude that car use petrol as fuel are solding more as compare to Diesel and hybrid. And no car using CNG of electric on solding in our dataset.



```
In [84]: # Ploting Bar graph / Bivariate Analysis
         plt.figure(figsize=[15,8])
         plt.bar(df['Fuel'],df['Price'])

Out[84]: <BarContainer object of 5239 artists>
```

- From this graph we can conclude that second owner car is having more price then first owner and unregistered car.

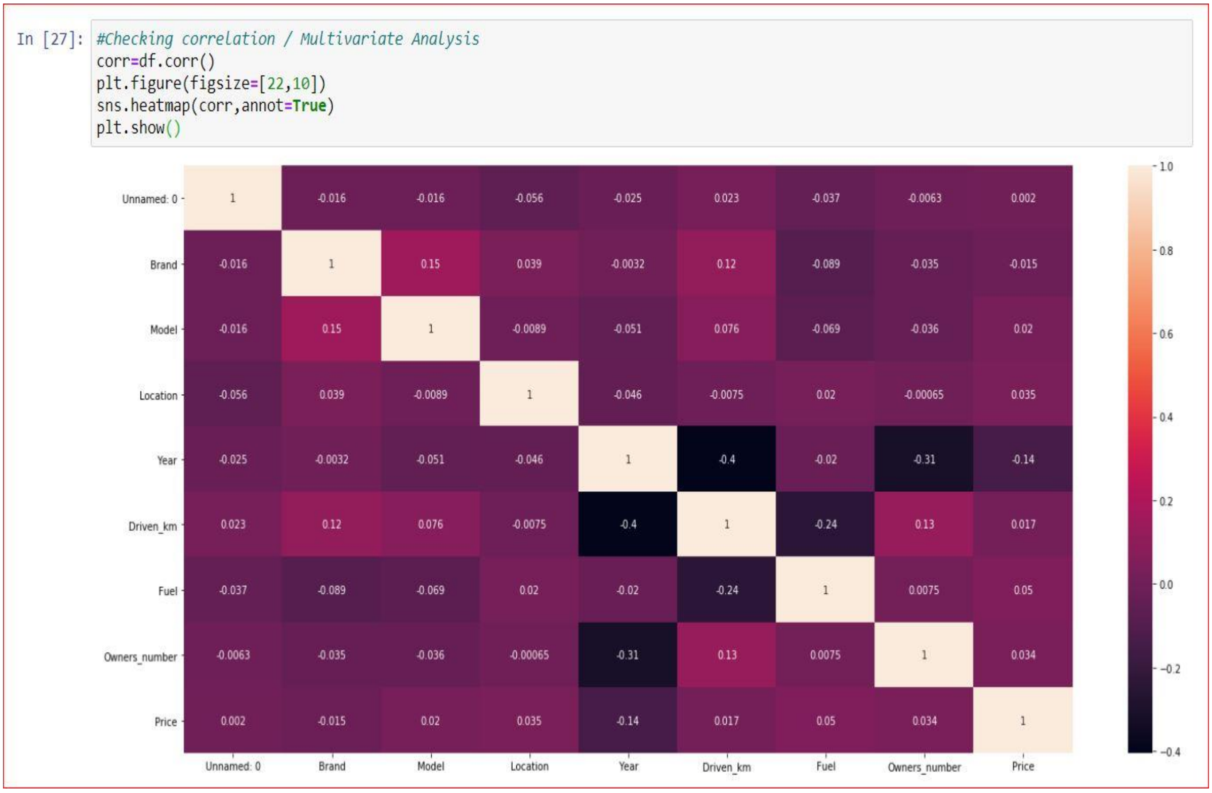


```
In [86]: # Ploting Bar graph / Bivariate Analysis
         plt.figure(figsize=[15,8])
         plt.bar(df['Owners_number'],df['Price'])

Out[86]: <BarContainer object of 5239 artists>
```

3. Multivariate analysis

This is not very much clear to see as the picture get blur. But we can understand this by its explanation. In multivariate analysis we check correlation between all columns among each other. From this we observed that best correlation of our target variable ‘Price’ is with ‘Brand’and least with ‘year’.



```
In [27]: #Checking correlation / Multivariate Analysis
         corr=df.corr()
         plt.figure(figsize=[22,10])
         sns.heatmap(corr,annot=True)
         plt.show()
```

	Unnamed: 0	Brand	Model	Location	Year	Driven_km	Fuel	Owners_number	Price
Unnamed: 0	1	-0.016	-0.016	-0.056	-0.025	0.023	-0.037	-0.0063	0.002
Brand	-0.016	1	0.15	0.039	-0.0032	0.12	-0.089	-0.035	-0.015
Model	-0.016	0.15	1	-0.0089	-0.051	0.076	-0.069	-0.036	0.02
Location	-0.056	0.039	-0.0089	1	-0.046	-0.0075	0.02	-0.00065	0.035
Year	-0.025	-0.0032	-0.051	-0.046	1	-0.4	-0.02	-0.31	-0.14
Driven_km	0.023	0.12	0.076	-0.0075	-0.4	1	-0.24	0.13	0.017
Fuel	-0.037	-0.089	-0.069	0.02	-0.02	-0.24	1	0.0075	0.05
Owners_number	-0.0063	-0.035	-0.036	-0.00065	-0.31	0.13	0.0075	1	0.034
Price	0.002	-0.015	0.02	0.035	-0.14	0.017	0.05	0.034	1

- **Interpretation of the Results**

  From the exploratory Data analysis, we could generate insight on the data. How each of the features relates to the target. Also, it can be seen from the evaluation of 3 models that RandomForestRegressor performed better than other models with accuracy score of 97%

CONCLUSION

- **Learning Outcomes of the Study in respect of Data Science**
  1. In univariate analysis we find have many times a particular car price and driven_km is present in our data.
  2. In bivariate analysis we find the relation between Location vs Price, Year vs Price, Fuel vs Price and Price vs Owners_number.
  3. In multivariate analysis we find the correlation between all the columns among each other
  4. In model building we find RandomForestRegresor is the best algorithm as compare to other. We find the accuracy score of 97% in this algorithm.

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

  This work is able to fulfil the future scope of our client as they want some predictions that could help them in further investment and improvement in selection of customers.