FLIP ROBO

PROJECT REPORT

Submitted By

Aswini Arun Patil

NAME OF THE PROJECT

Car Price Prediction

Acknowledgement:

- Primarily I would like to thank God to being able to complete this project with success. Then I would like to express my special thanks of gratitude to my SME,

- And I am thankful I am part of flib rob technology of employee, who given me the golden opportunity to do this wonderful project on the given topic which is also help me in doing a lot of research and I came to know about so many new things, I am really thankful to flip robo.

DATE:11/07/2022

ASWINI A. PATIL
Data Science course
Institute: Data trained education
Internship: Flip Robo technology
@Bangalore

Abstract:

- A car price prediction has been a high interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction.

- To build a model for predicting the price of used cars the applied three machine learning techniques and artificial neural network and linear regression.

- Respective performances of different algorithms were then compared to find one that best suits the available data set. The final prediction model was integrated into java application. Furthermore, the model was evaluated using test data and the accuracy of 82% was obtained

Introduction:

- From a long time since being a continuous paradigm of transaction of commodities has been into existence. Earlier these transactions were in the from of barter system which later was translated into a monetary system. And with consideration into these, all changes that were brought about the pattern of re-selling items was affected as well. There are two ways in which the re-selling of the item is carried out. One is offline and the other being online. In offline transaction, there is mediator present in between who is very vulnerable to being corrupt and make overly profitable transaction. The second option is online wherein there is a certain platform which lets the user find the price the might get if he goes for selling.

- Vehicle price prediction especially when the vehicle is used and not coming direct from the factory , in both a critical and important task. With increase in demand for used cars more and more vehicle buyer are finding alternatives of buying new cars.

- There is need of accurate price prediction mechanism for the used cars. Prediction techniques of machine learning can be helpful in this regard.

- It is common to lease a car in many countries rather then buying a new car.

Motivation:

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.

Objective:

The main aim of this project is to predict the price of used cars using the various Machine Learning (ML) models. This can enable the customers to make decisions based on different inputs or factors namely.
To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes.

The system that is being built must be features based i.e. feature wise prediction must be possible.
Providing graphical comparisons to provide a better view.
Predict the price of a car, bike, electric vehicle and hybrid vehicle. This app can predict the price of any vehicle because of the smartly optimized.

- Brand or Type of the car one prefers like Ford, Hyundai
- Model of the car namely Ford Figo, Hyundai Creta
- Location like Delhi, Chennai, Mumbai
- Year of manufacturing like 2020, 2021
- Type of fuel namely Petrol, Diesel
- Price range or Budget
- Type of transmission which the customer prefers like Automatic or Manual
- Mileage

Case Study Description:

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.

In the dataset, I have scrapped 6000 different kinds used cars data.

The given dataset contains various Brands, Models, Kilometers driven, Manufacturing Year, Number of Owners, Fuel Type of the particular car, and finally the price of the car. These cars are selling in various locations in India. The given dataset includes all types of cars for example- SUV, Sedans, Coupe, etc.

Features:

There will be majorly two features provided in the project note that this will be not

- Re-sale platform: A centralized platform for car re-sale that will predict prices.
- Feature selection: Feature-based search and prediction.

Literature Review:

We discuss various application and methods which inspired us to build our project. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of information like the technological stack, algorithms, and shortcomings of our project which led us to build a better project.

CarWale:

CarWale app is one of the top-rated ca apps in India for new and used car research. It provides accurate on-road prices of cars, genuine user and expert reviews. It can also compare different cars with the car comparison tool. This app also helps you to connect with your nearest car dealers for the best offers available.

CarTrade:

carTrade is web and Android platform where user can research new cars in India by exploring car prices, car specs, images, mileage, reviews, and car comparisons. On this app one can sell used car to genuine buyers with ease. One can list their used car for sale along with details like image, model, and year of purchase and kilometers so that it is displayed to lakhs of interested car buyers in their city. User can read user reviews and expert car reviews with images that help in finalizing a new car buying decision.

Data Processing:

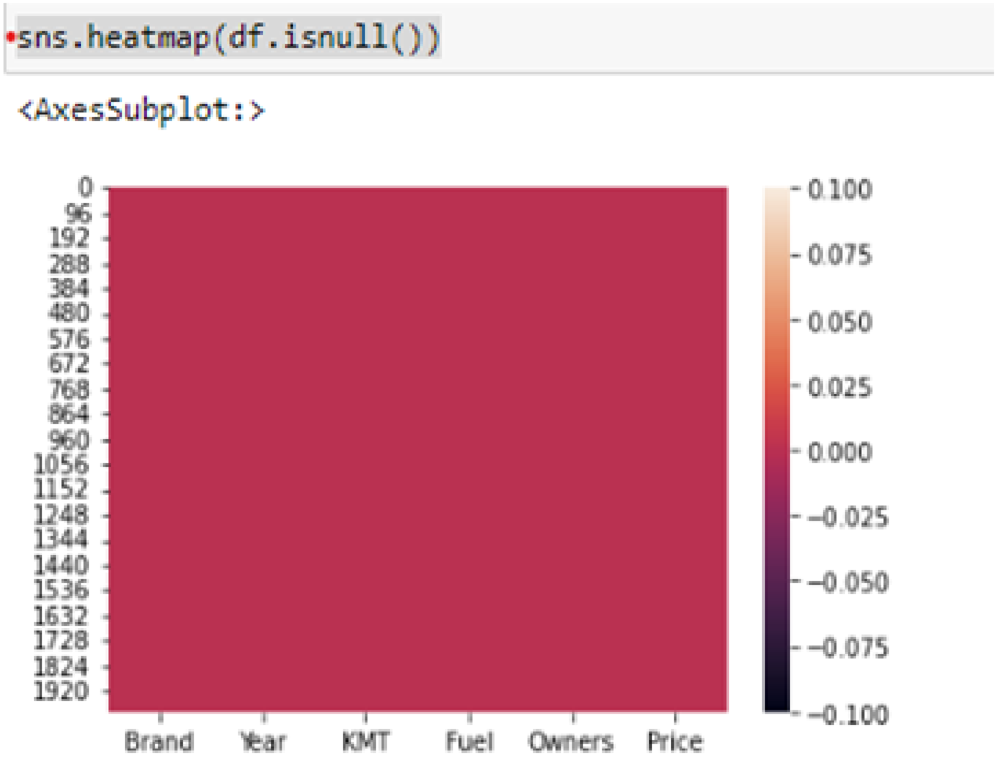
- Null Values: The dataset has no null values present in it originally.
- Unique values and Value count of each column.
- The “Kilometers Driven” column has commas in the data so I removed the commas.
- The “No of Owners” column has repetitive entries under 1st owner and First owner as well as 2nd and Second owner. So, I merged these data to 1st and 2nd owner respectively

Data Analysis We will try to Find out the below stuff:

1.Missing Values in the dataset.2. All the Numerical variables and Distribution of the numerical variables.
3. Categorical Variables
4. Outliers
5. Relationship between an independent and dependent feature (*selling price*)

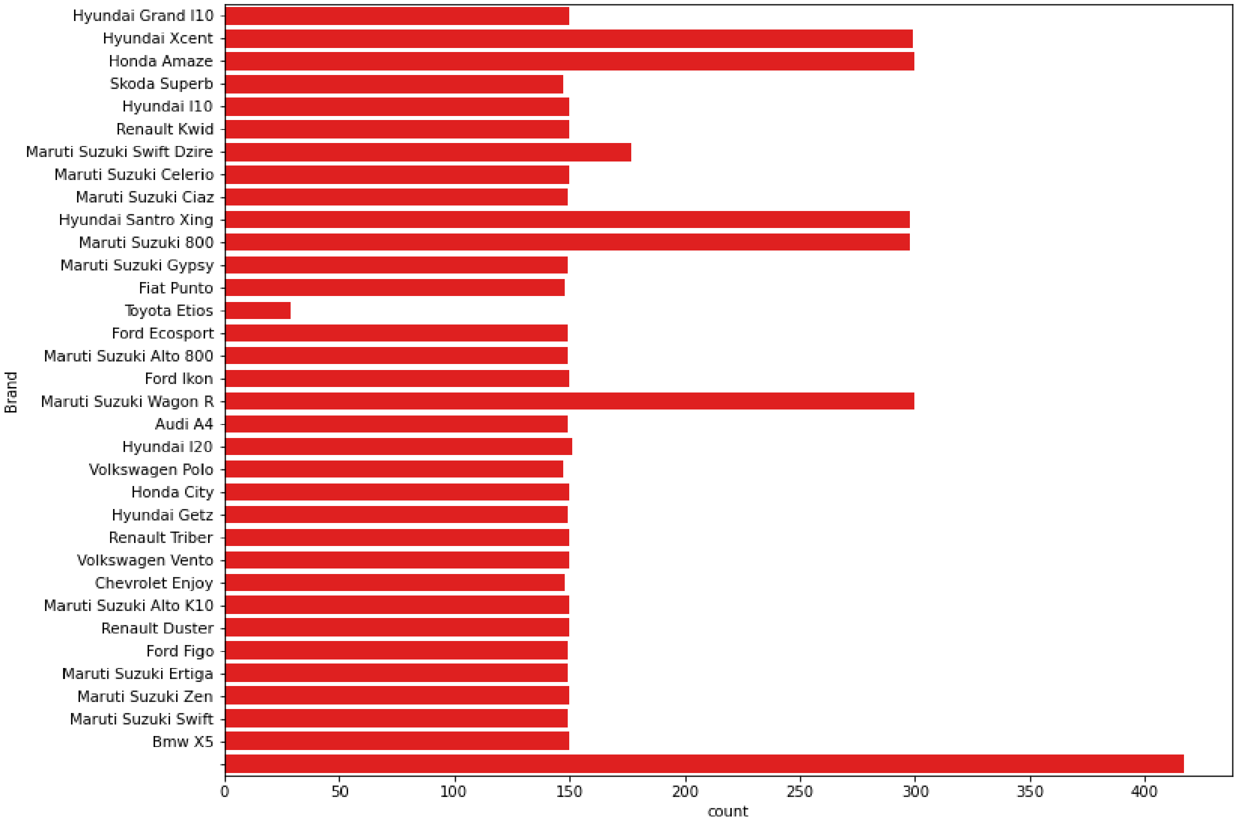
Visualizing variables and relationships:

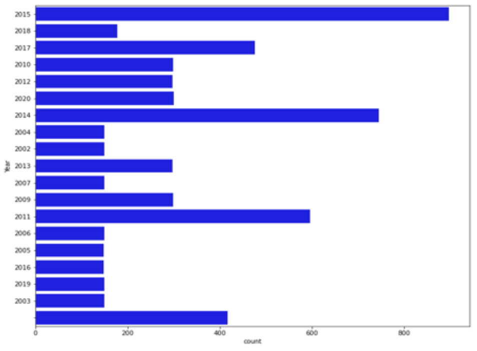
After cleaning the data, we can visualize data and better understand the relationships between different variables. There are many more visualizations that you can do to learn more about your dataset, like scatterplots, histograms, boxplots, etc Using **sns.heatmap()**, we can see that the '*Present_Price'* is positively correlated with '*Selling_Price'* and '*Fuel_Type_Petrol'* and '*Fuel_Type_Diesel'* is negatively correlated with '*Selling_Price'.*

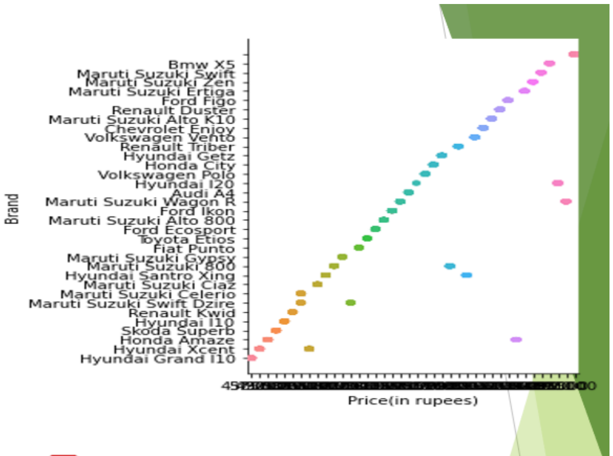


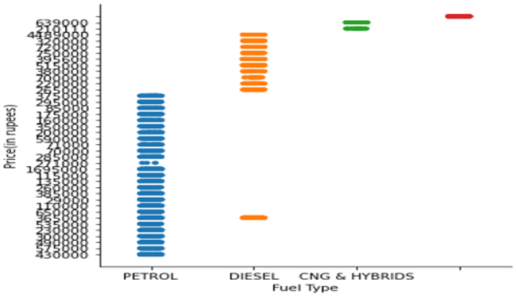
```
sns.heatmap(df.isnull())
```

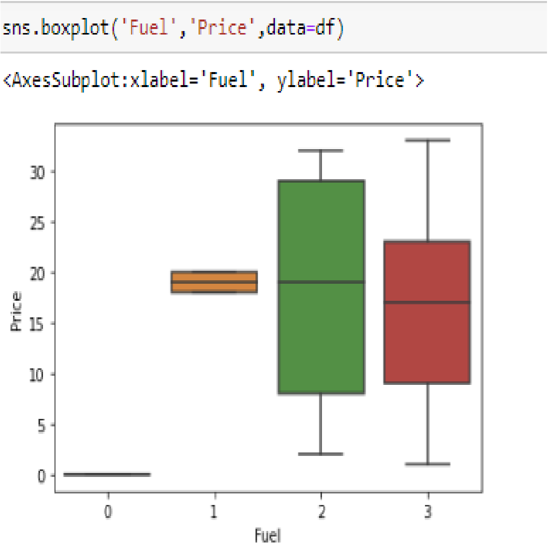
```
<AxesSubplot:>
```


Hyundai Grand I10
Hyundai Xcent
Honda Amaze
Skoda Superb
Hyundai I10
Renault Kwid
Maruti Suzuki Swift Dzire
Maruti Suzuki Celerio
Maruti Suzuki Ciaz
Hyundai Santro Xing
Maruti Suzuki 800
Maruti Suzuki Gypsy
Fiat Punto
Toyota Etios
Ford Ecosport
Maruti Suzuki Alto 800
Ford Ikon
Maruti Suzuki Wagon R
Audi A4
Hyundai I20
Volkswagen Polo
Honda City
Hyundai Getz
Renault Triber
Volkswagen Vento
Chevrolet Enjoy
Maruti Suzuki Alto K10
Renault Duster
Ford Figo
Maruti Suzuki Ertiga
Maruti Suzuki Zen
Maruti Suzuki Swift
Bmw X5
Brand
0
50
100
150
200
250
300
350
400
count


2015
2018
2017
2010
2012
2020
2014
2004
2002
2013
2007
2009
2011
2006
2005
2016
2019
2003
Year
0
200
400
600
800
count

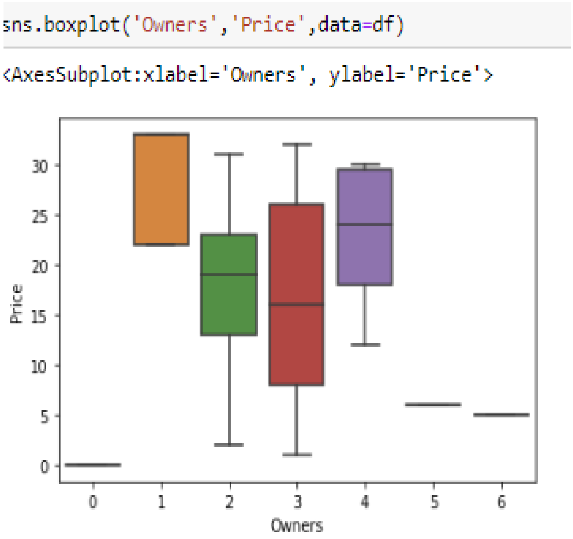

Bmw X5
Maruti Suzuki Swift
Maruti Suzuki Zen
Maruti Suzuki Ertiga
Ford Figo
Renault Duster
Maruti Suzuki Alto K10
Chevrolet Enjoy
Volkswagen Vento
Renault Triber
Hyundai Getz
Honda City
Volkswagen Polo
Hyundai I20
Audi A4
Maruti Suzuki Wagon R
Ford Ikon
Maruti Suzuki Alto 800
Ford Ecosport
Toyota Etios
Fiat Punto
Maruti Suzuki Gypsy
Maruti Suzuki 800
Hyundai Santro Xing
Maruti Suzuki Ciaz
Maruti Suzuki Celerio
Maruti Suzuki Swift Dzire
Renault Kwid
Hyundai I10
Skoda Superb
Honda Amaze
Hyundai Xcent
Hyundai Grand I10
Brand
Price(in rupees)


639000
210111
4489000
320000
720000
750000
395600
515000
430000
Price(in rupees)
PETROL
DIESEL
CNG & HYBRIDS
Fuel Type



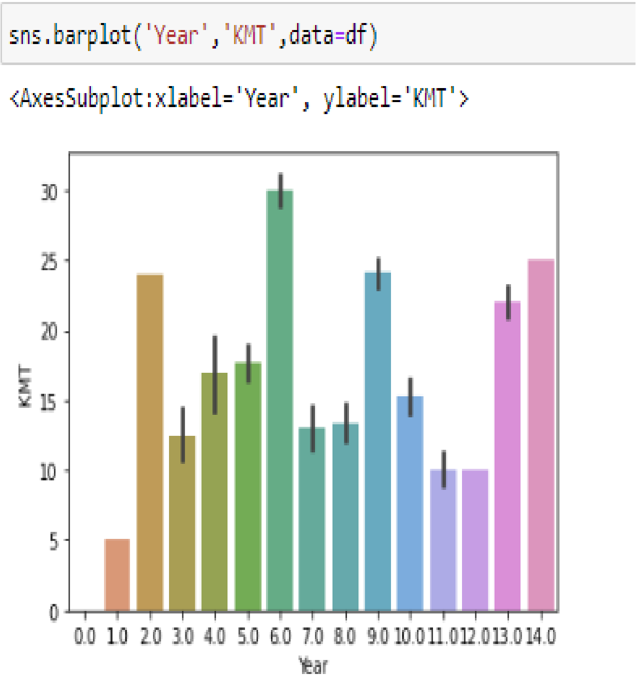
```
sns.boxplot('Fuel','Price',data=df)
```

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



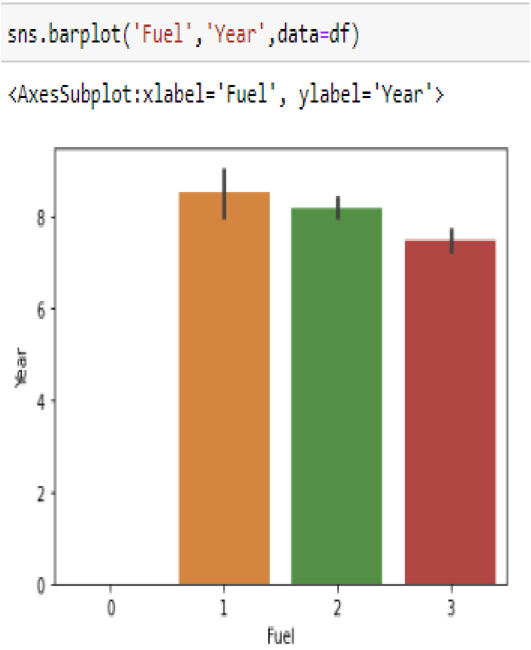
```
sns.boxplot('Owners','Price',data=df)
```

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



```
sns.barplot('Year','KMT',data=df)
```

```
<AxesSubplot:xlabel='Year', ylabel='KMT'>
```



```
sns.barplot('Fuel','Year',data=df)
```

```
<AxesSubplot:xlabel='Fuel', ylabel='Year'>
```

Model Training:

Train-test data splits were conducted. In this situation, we split the data into training and test sets, then fit candidate models on the training set, evaluate and select them on the test set.

We do not know beforehand which model will perform best on this problem, as it is unknowable. Therefore, we fit and evaluate a suite of different models on the problem. We used Random Forest Regression and Linear Regression Models (Feature Selection) on the train set. we also check which model gives us the best result. you can try any number of regression models and choose one among them which is best suitable.

we use the random grid to search for the best hyperparameters. A random search of parameters, using 3 fold cross-validation search across 100 different combinations.

Model selection is the process of choosing one among many candidate models for a predictive modeling problem. after analyzing both models we chose the Random Forest Regression model.

Fitting parametric distributions:

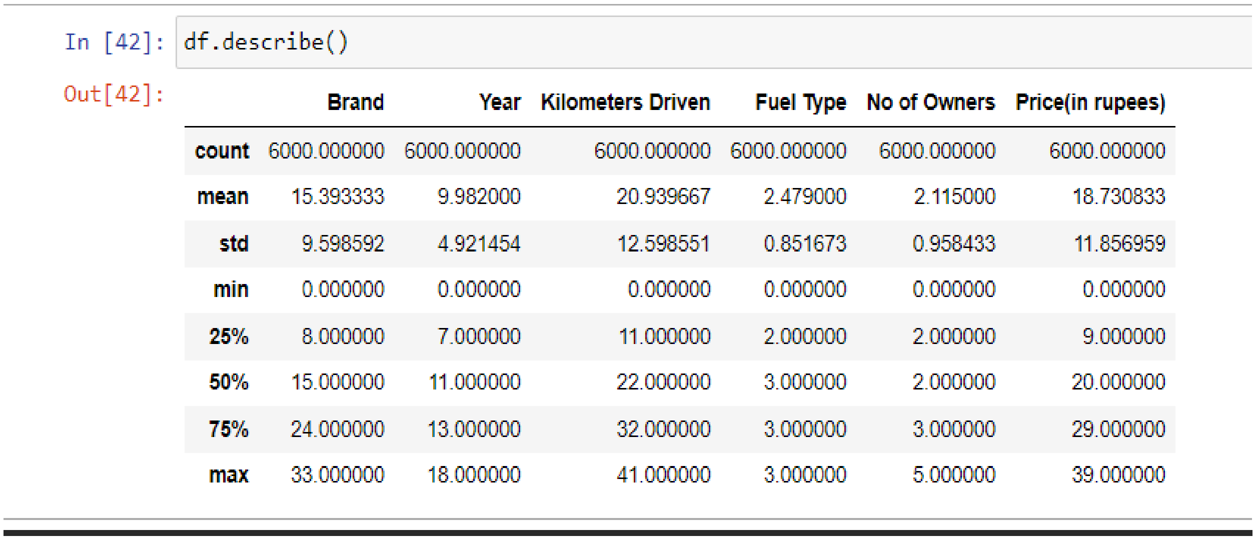
You can also use**distplot()**to fit a parametric distribution to a dataset and visually evaluate how closely it corresponds to the observed data. it should be a closed Gaussian distributed graph and the difference between ‘y_test’(real value)and ‘predictions’ should also be minimal.

Checking accuracy of the model:

Evaluating the model accuracy is an essential part of the process of creating machine learning models to describe how well the model is performing in its predictions. The MSE, MAE, and RMSE metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.

Statistical Summary:

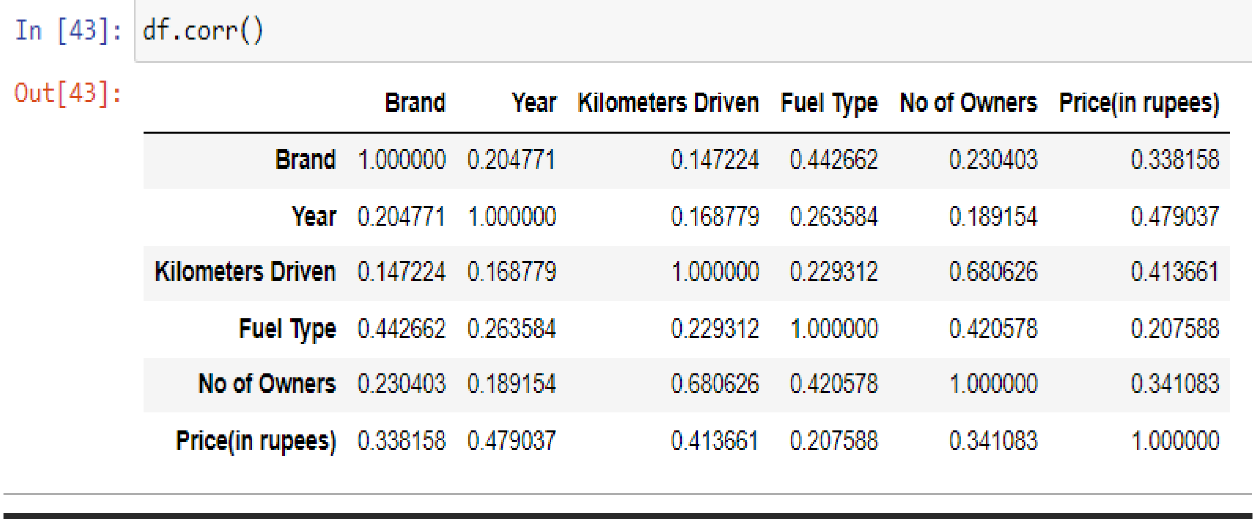


In [42]:
```
df.describe()
```

Out[42]:

	Brand	Year	Kilometers Driven	Fuel Type	No of Owners	Price(in rupees)
count	6000.000000	6000.000000	6000.000000	6000.000000	6000.000000	6000.000000
mean	15.393333	9.982000	20.939667	2.479000	2.115000	18.730833
std	9.598592	4.921454	12.598551	0.851673	0.958433	11.856959
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	8.000000	7.000000	11.000000	2.000000	2.000000	9.000000
50%	15.000000	11.000000	22.000000	3.000000	2.000000	20.000000
75%	24.000000	13.000000	32.000000	3.000000	3.000000	29.000000
max	33.000000	18.000000	41.000000	3.000000	5.000000	39.000000

Correlation:

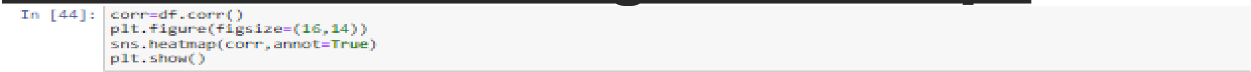


In [43]:
```
df.corr()
```

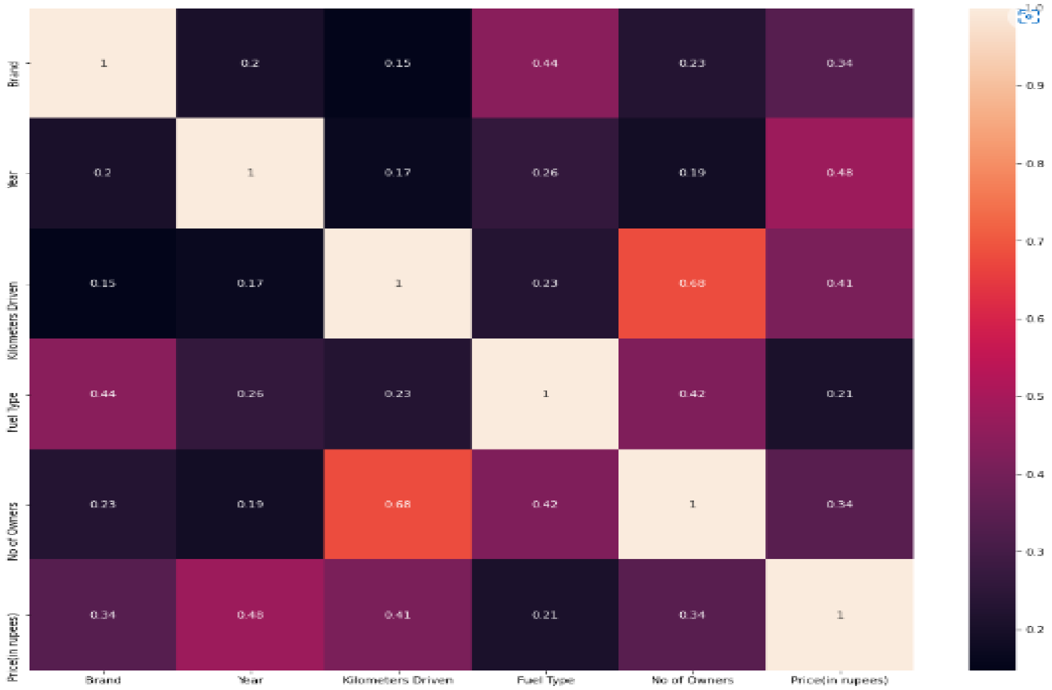
Out[43]:

	Brand	Year	Kilometers Driven	Fuel Type	No of Owners	Price(in rupees)
Brand	1.000000	0.204771	0.147224	0.442662	0.230403	0.338158
Year	0.204771	1.000000	0.168779	0.263584	0.189154	0.479037
Kilometers Driven	0.147224	0.168779	1.000000	0.229312	0.680626	0.413661
Fuel Type	0.442662	0.263584	0.229312	1.000000	0.420578	0.207588
No of Owners	0.230403	0.189154	0.680626	0.420578	1.000000	0.341083
Price(in rupees)	0.338158	0.479037	0.413661	0.207588	0.341083	1.000000

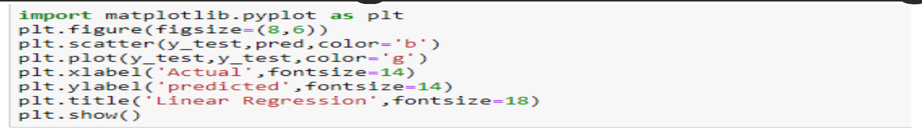
Correlation using Heatmap:



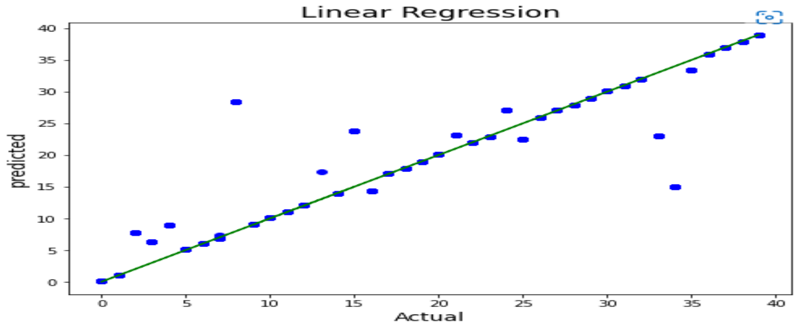
```
In [44]: corr=df.corr()
         plt.figure(figsize=(16,14))
         sns.heatmap(corr,annot=True)
         plt.show()
```



Overfitting & Underfitting:



```
import matplotlib.pyplot as plt
plt.figure(figsize=(8,6))
plt.scatter(y_test,pred,color='b')
plt.plot(y_test,y_test,color='g')
plt.xlabel('Actual',fontsize=14)
plt.ylabel('predicted',fontsize=14)
plt.title('Linear Regression',fontsize=18)
plt.show()
```



Conclusion:

So, the conclusion is the **average selling price** and average present price of the cars are not the same they are different. Regression: Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x).