CAR PRICE PREDICATION

Submitted by:

AMEY PRABHUNE

ACKNOWLEDGMENT

I take this an opportunity to thank all those who have directly inspired and helped me towards successful completion of this project report. We express our sincere thanks to SME MR. Shubham Yadav for his guidance.

INTRODUCTION

- ## Business Problem Framing

  We are looking for new machine learning models from new data. We have to make car price valuation model.

- ## Conceptual Background of the Domain Problem

  Describe the domain related concepts that you think will be useful for better understanding of the project.

- ## Motivation for the Problem Undertaken

  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

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  a. For the visualization I have used numpy, sklearn(sikit learn), pandas, matplotlib, zscore.
  b. For measuring our model accuracy I have used accuracy score, confusion matrix and classification report.
  c. For pre-processing I have used min-max scaler, power transform.
  d. For model selection I have used train_test_split, and cross validation.

- Data Sources and their formats

You have to scrape 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.

```
In [65]: df.head()
```

Out[65]:

	Model	Price	Kilometers Driven	Year	Owner	Fuel Type	Transmission	Location
0	MarutiWagonR1.0LXI	312165	82238	2014	First Owner	Petrol + CNG	MANUAL	Delhi
1	ToyotaEtiosLiva	313799	30558	2013	First Owner	Petrol	MANUAL	Coimbatore
2	MarutiAlto800	295999	22164	2018	First Owner	Petrol	MANUAL	Mumbai
3	MarutiSwift	435199	30535	2013	First Owner	Diesel	MANUAL	Hyderabad
4	MarutiWagonR1.0	289099	15738	2013	First Owner	Petrol	MANUAL	Mumbai

- Data Pre-processing Done

In our data set there are no null values. There are 2 features which are numerical. We will normalize the skewed data with transformation. And there is a positive correlation between the features and the Price. There are few outliers and we will try to remove it. For categorical columns we will use one hot encoding and then we will remove columns which are not useful to predict there price.

- Data Inputs- Logic- Output Relationships

  Given data is in tabular form of rows and column. Data-type of dataset is object means it is mixture of character and numeric type. Our output is dependent on feature columns. With the help of feature columns we can predict our target variable.

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

  We might want to remove columns. There are 4 features which are object. We try to normalize the data. And there is a positive correlation between the features and the ratings. There are few outliers and we will try to remove it.

- Hardware and Software Requirements and Tools Used

  a. Software Requirement :
     i. Excel
     ii. OS – windows , Linux
     iii. Jupyter Notebook
     iv. Internet browser
  b. Hardware Requirement:
     i. RAM: 4 GB or more than.
     ii. ROM: 50 GM or more than.
     iii. Internet connection.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  In target variable we have continuous data so it is a regression problem. For model building we will use regression model like. Random Forest, XG Boost etc. in dataset ratings column is our target variable and others are feature columns means independent and dependent variable. First I

drop column which are not useful for the model building. Then there are few columns which are in object type, I converted it into numeric type.

- Testing of Identified Approaches (Algorithms)
    a. Linear Regression
    b. Random Forest
    c. Decision Tree

- Key Metrics for success in solving problem under consideration

  In dataset, we can see the correlation with the help of plot means there are some big values in our dataset. There are null values in dataset.

- Visualizations

```
: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4363 entries, 0 to 4362
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Model              4363 non-null   object
 1   Price              4363 non-null   int64 
 2   Kilometers Driven  4363 non-null   int64 
 3   Year               4363 non-null   int64 
 4   Owner              4363 non-null   object
 5   Fuel Type          4363 non-null   object
 6   Transmission       4363 non-null   object
 7   Location           4363 non-null   object
dtypes: int64(3), object(5)
memory usage: 272.8+ KB
```

```
In [68]: df.columns
Out[68]: Index(['Model', 'Price', 'Kilometers Driven', 'Year', 'Owner', 'Fuel Type',
                'Transmission', 'Location'],
               dtype='object')
```

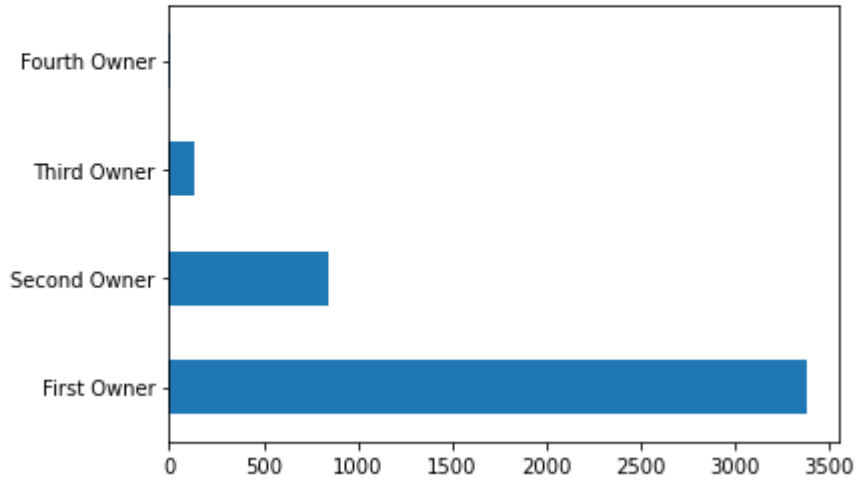
In [69]:
```
df.isnull().sum()
```

Out[69]:
```
Model                0
Price                0
Kilometers Driven    0
Year                 0
Owner                0
Fuel Type            0
Transmission         0
Location             0
dtype: int64
```

In [70]:
```
sb.heatmap(df.corr(), cmap="YlGnBu", annot = True)
plt.show()
```

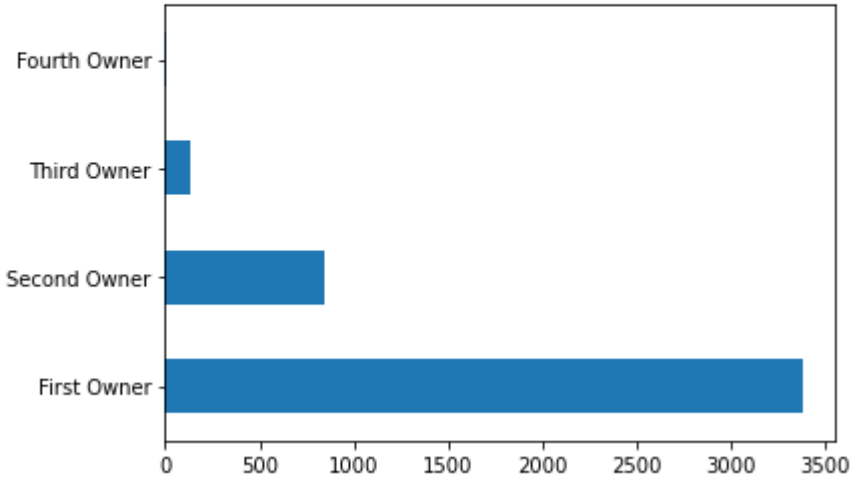
In [73]:
```
df['Owner'].value_counts().plot(kind='barh')
```

Out[73]: <AxesSubplot:>



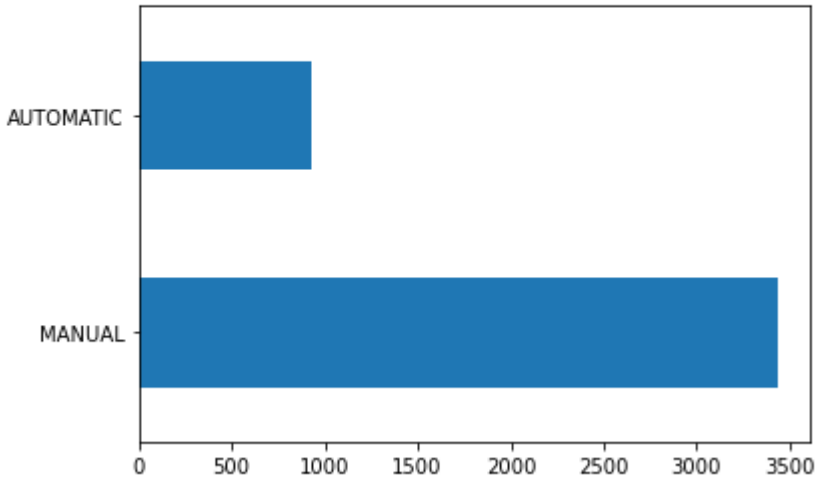
In [75]:
```
df['Owner'].value_counts().plot(kind='barh')
```

Out[75]: <AxesSubplot:>



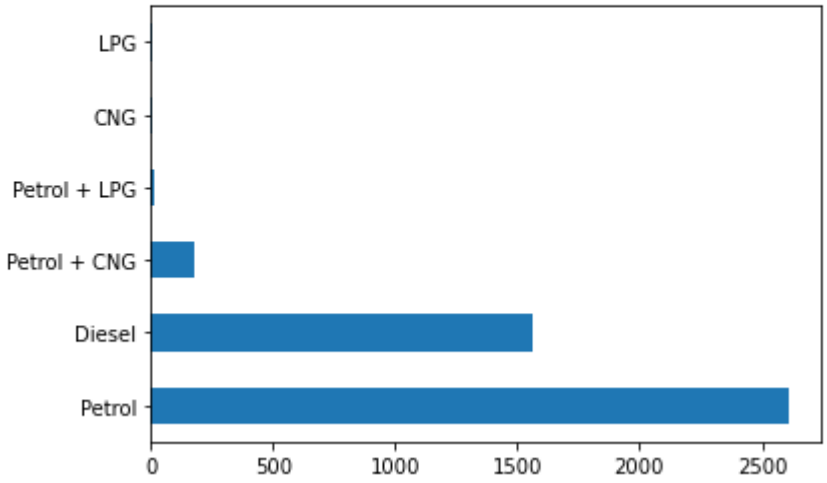
```
In [77]: df['Transmission'].value_counts().plot(kind='barh')
```

```
Out[77]: <AxesSubplot:>
```



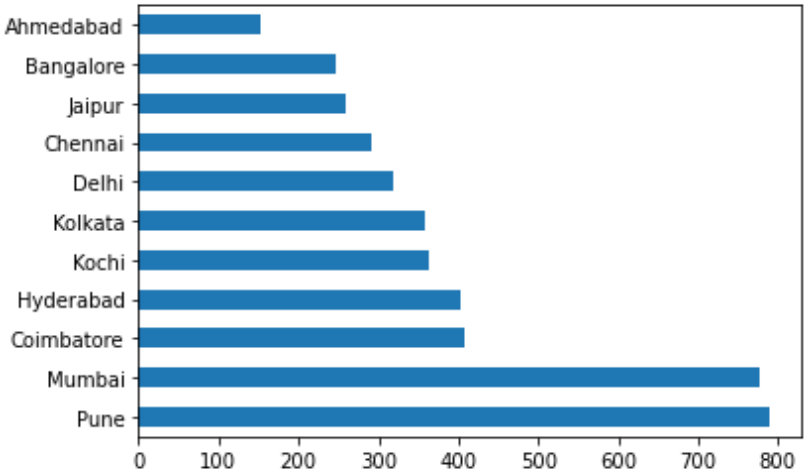

```
n [78]: df['Fuel Type'].value_counts().plot(kind='barh')
```

```
ut[78]: <AxesSubplot:>
```



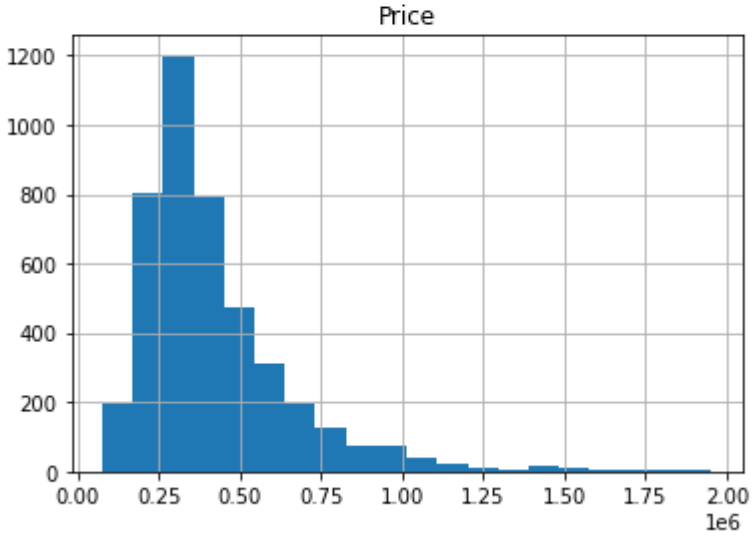
In [80]:
```
df['Location'].value_counts().plot(kind='barh')
```

Out[80]: <AxesSubplot:>

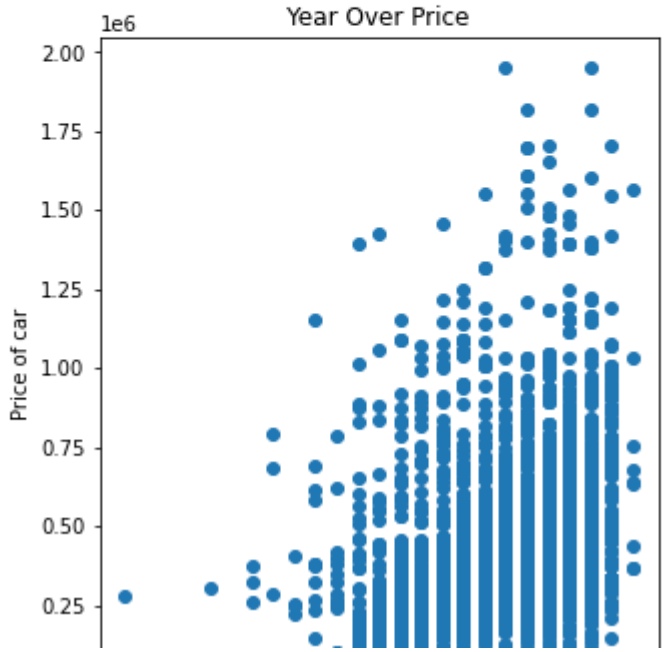


In [82]:
```
df.hist(column='Price', bins=20)
```

Out[82]: array([[<AxesSubplot:title={'center':'Price'}>]], dtype=object)

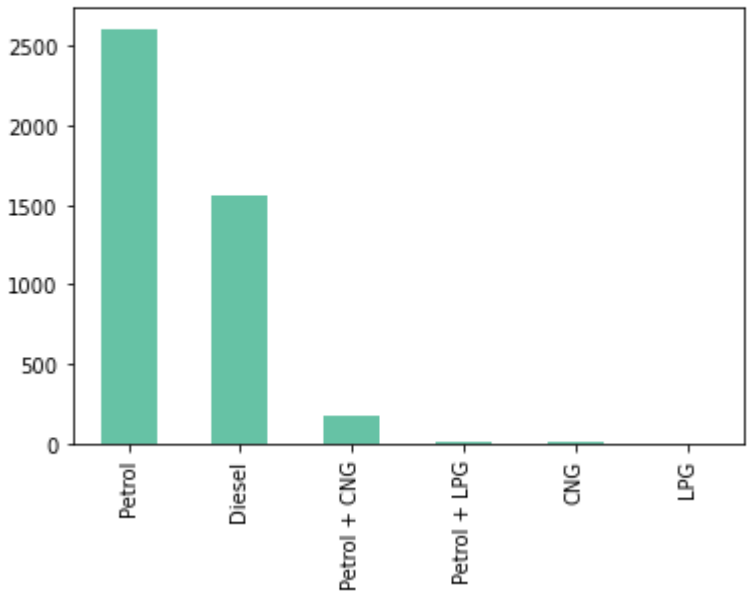


In [87]:
```
plt.figure(figsize = (5, 6))
plt.title('Year Over Price')
plt.scatter(df['Year'], df['Price'])
plt.xticks(rotation = 90)
plt.xlabel('Year')
plt.ylabel('Price of car')
plt.show()
```

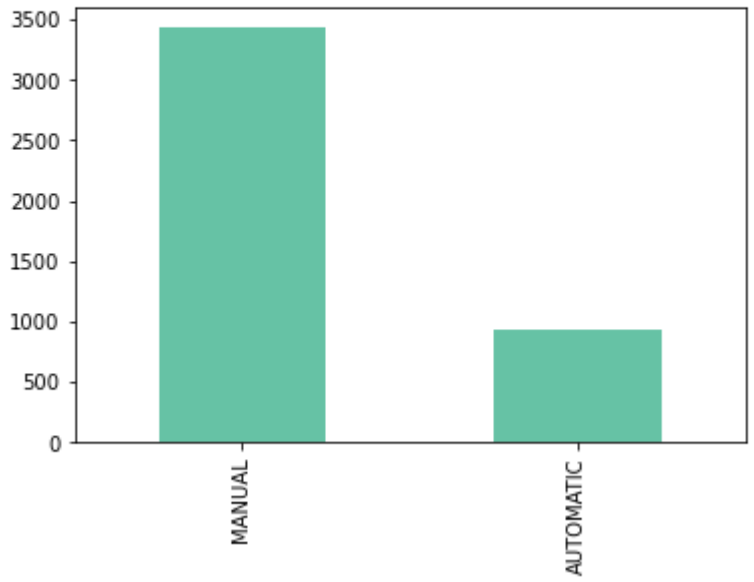


:
```
df['Fuel Type'].value_counts().plot.bar(cmap='Set2')
```

: <AxesSubplot:>



```
In [90]: df['Transmission'].value_counts().plot.bar(cmap='Set2')
Out[90]: <AxesSubplot:>
```



- Interpretation of the Results

  Most column are right and left skewed. There are only few columns which are normally distributed. Few columns are highly correlated with our target variable and few are less.

CONCLUSION

- Key Findings and Conclusions of the Study
  a. We have continuous values in our target variable.
  b. There are null values.
  c. 7 feature values

- Learning Outcomes of the Study in respect of Data Science

  With the help of visualization we can easily understand our data. Visuals and diagrams makes it easier for us to identify strongly correlated parameters. Visualization can improve speed of decision making. Clean data can give us more accurate result. If data is noisy then our model won't work as we expect.

- Limitations of this work and Scope for Future Work
  a. Data: Lack of Good Data.
  b. Time: building a machine learning model is time consuming.
  c. Performance: Performance cannot guaranteed.