USED CARS WEBSCRAPING & PREDICTION

Submitted By :

ASHUTOSH CHAUDHARY

ACKNOWLEDGMENT

I would like to express my gratitude towards my internship mentor Ms. Srishti Maan for helping me in completion of the project.

BUSINESS PROBLEM

From my understanding the problem is about scraping data from websites of used cars to make dataset & then making prediction on used cars prices.

OBJECTIVE FOR PROBLEM UNDERTAKEN

We have to study every feature’s behaviour present in the dataset with regards to target feature (Price of used cars) & make observation from its behaviour that how will it affect the target feature to build a model which would perform quite good on dataset to make prediction of prices of used cars.

ANALYTICAL PROBLEM FRAMING

- **Origin of dataset & data types of every features**

We have to create dataset ourselves by webscraping the data from various websites & then convert it into dataset. After doing this we convert the dataset to csv file & then we need to import it using various libraries. Also features present in the dataset are of both continuous & categorical data types.

- **Mathematical/Analytical modelling of the problem**

For visualization we only use three plots most of the times that were countplot, distplot & scatterplot & for model building we use Linear Regression, Decision Tree Regression, Random Forest Regression & Bagging Regression models to opt best out of them to work on dataset.

- **Assumptions related to problem statement**

No assumptions were made while working on the dataset.

- **Libraries & Tools used**

We used numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle & warnings libraries for this task.

STEPS TAKEN FOR THE TASK

1. **Importing Libraries for the task**

   Numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickles & warnings were imported for task to get completed.

2. **Importing Dataset using libraries**

   Imported the dataset using pandas library in jupyter notebook.

3. **Checking Dimension of dataset**

   By checking dimensions of dataset we get to know that it contains 5047 rows & 8 columns.

4. **Checking Description of dataset**

   From description we find the mean, min value, max value, etc of every column which contains continuous data in them.

5. **Checking for presence of null values in dataset**

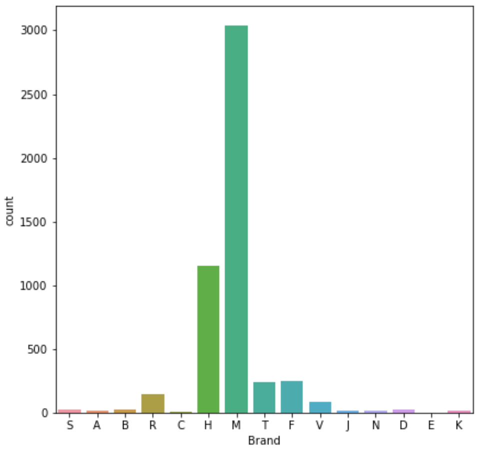
   We checked for the presence of null values in dataset & we get to know that null values are present in every column.

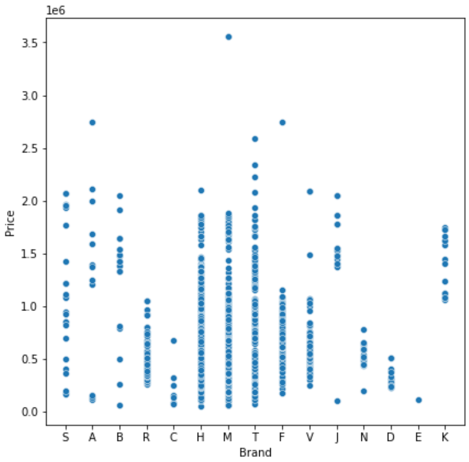
6. Identifying Target variable

We have created the dataset ourselves ,so, we know that target variable is named Price.

7. Performing EDA on whole dataset

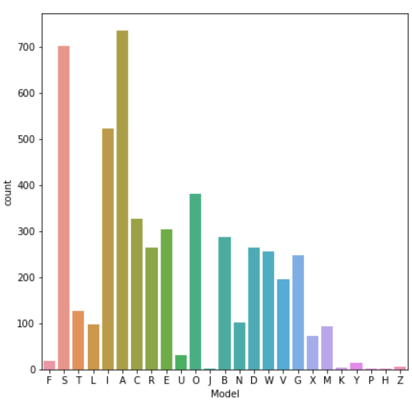
Brand

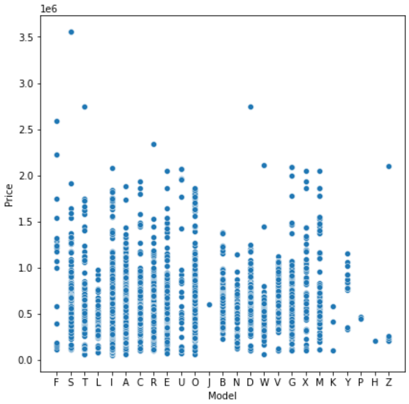




From visualization we concluded that most cars available for sale is of Brand name starting with M & least sale of used car is from Brand name starting with E.

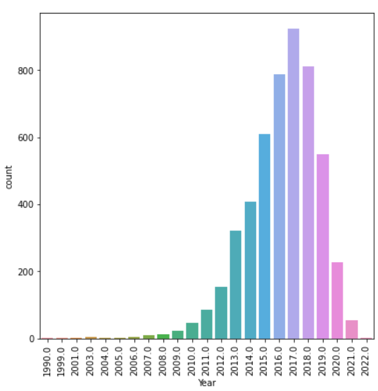
Model

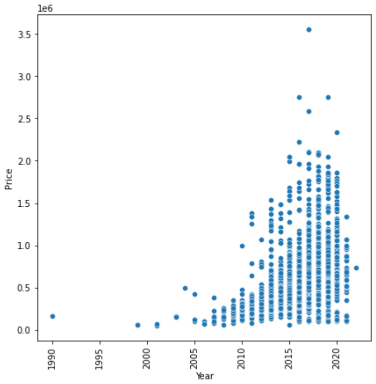




From visualization we concluded that most cars available for sale is of Model name starting with A & least cars available for sale is of Model name starting with H. Also the highest price for a car is of model name starting with S.

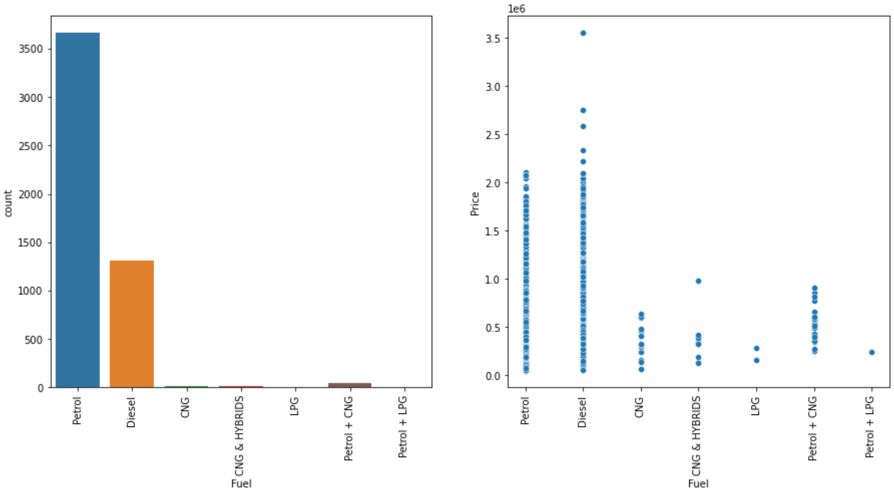
Year





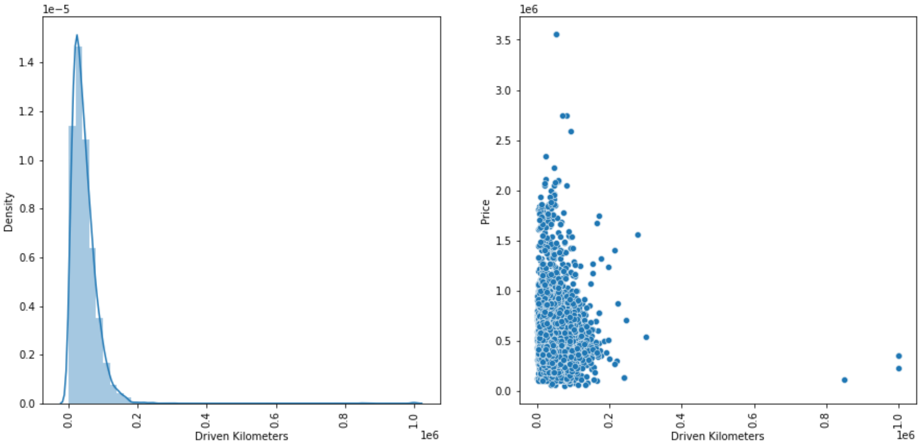
From visualization we concluded that most cars available for sale are manufactured in year 2017 & cars available for sale that were manufactured before year 2000 are very few & highest price for the car for sale is from year 2017.

Fuel



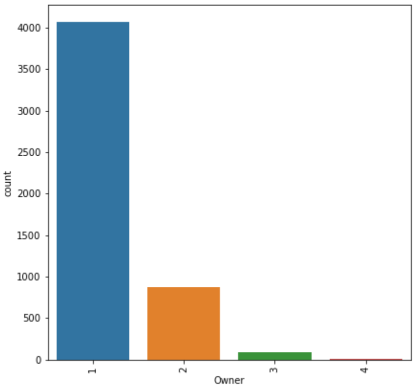
From visualization we concluded that most cars for sale use petrol as fuel & least cars available for sale use petrol + LPG as fuel. Also the highest price for a car available for sale use diesel as fuel.

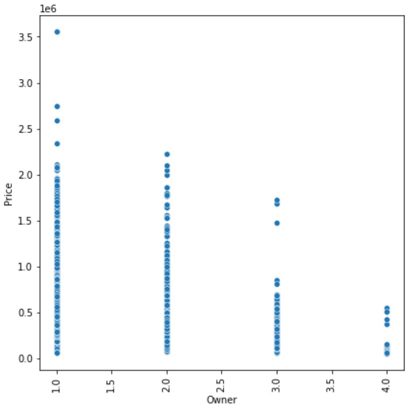
Driven Kilometers



From visualization we concluded that data is positively skewed with having highest density at around 0.05 & most cars available for sale are concentrated with in range from 0.0 to 0.2

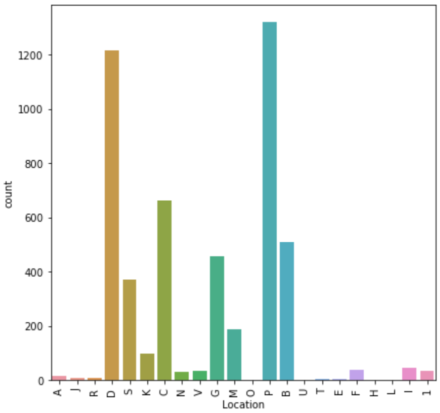
Owner

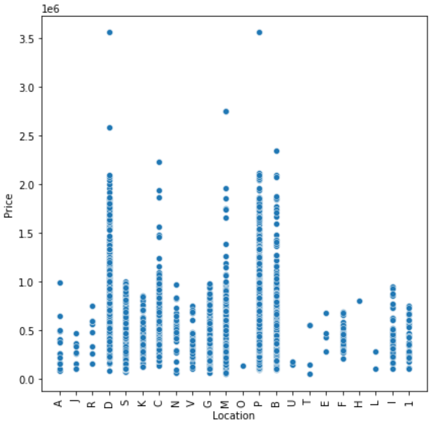




From visualization we concluded that most cars available for sale have 1st owner & least cars available for sale have 4th owner.

Location





From visualization we concluded that most cars available for sale are from location name starting with P & least cars available for sale are from location name starting with H.

8. Plotting Heatmap

We plotted heatmap to check correlation of every column with target column & we concluded that following (Model, Year & Fuel) columns have high positive correlation with target column & also we dropped column named Brand to reduce multicollinearity.

9. Splitting & Scaling data

Then we splitted data into X & y in which y contains target variable & X contains other variables. Then we further splitted X & y into X_train, X_test, y_train, y_test sets using train_test_split & after that we scaled X_train & X_test set using StandardScalar to remove outliers if present any in the columns.

10. Applying GridSearchCV

We chose 4 regression models (Linear Regression, Decision Tree Regression, Random Forest Regression, Bagging Regression) to check performance of them on the dataset to find out the best suited model for the dataset. Then one by one we applied GridSearchCV on each model to find best hyper parameter tuning while modelis working on dataset. After applying GridSearchCV on every model we concluded that Random Forest Regression model is giving best Test r2_score which was 0.3395 with hyper parameter combination (max_depth : 5, min_samples_leaf : 1, min_samples_split : 5)

11. Applying Model on dataset

Chose Random Forest Regression Model & applied it on dataset to make prediction without hyper parameter tuning just to check that whether would we get better r2_score or not. After making prediction we made a dataframe which contained actual & predicted values side by side to compare them.

12. Applying metrices

After making predictions we checked performance of model through various metrices :

- R2_score : 0.4430
- Root mean squared error : 237229.23
- Mean absolute error : 143524.34

As we had thought earlier that without hyperparameter tuning the model should give better r2_score & our assumption was correct as it was giving far better r2_score & model was performing quite good on dataset.

13. Saving Model

After checking performance of model we saved it using pickle library.

CONCLUSION

In the end we can conclude that prices of used car is getting mostly affected by Year in which it was manufactured & amount of kilometers it was driven. Also most cars in the dataset were manufactured in the year 2017 & most cars present for the sale are of model name starting with A.