MOTOR'S PRICE PREDICTION

HARSH SHRIVASTAVA

PGDDS BATCH 8

CONTENTS

- Problem Identification and Definition
- Implementation of the solution
- Evaluation of the solution
- Conclusion

PROBLEM IDENTIFICATION AND DEFINITION

The analysis required to build the prediction of motor car prices based on various attributes associated with the car and it will be used for decision making to understand how exactly prices vary with respect to independent variables is one of the most important criteria accordingly we can manipulate the business strategy to increase the profit of the company.

The model will be a good way for the company management to understand the pricing dynamics of a new market.

By considering past results, we need to train a model to accurately predict future outcomes.

KNOWING THE DATASET:

THIS DATASET CONSIST OF (13)STRING VARIABLES AND (6) NUMERICAL VARIABLES.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 47243 entries, 0 to 50000
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   dateCrawled          47243 non-null  object
 1   mtor_name            47243 non-null  object
 2   vendor               47243 non-null  object
 3   offerType            47243 non-null  object
 4   price                47243 non-null  int64
 5   abtest               47243 non-null  object
 6   vehicleType          47243 non-null  object
 7   yearOfRegistration   47243 non-null  int64
 8   gearbox              47243 non-null  object
 9   powerPS              47243 non-null  int64
 10  model                47243 non-null  object
 11  kilometer            47243 non-null  int64
 12  reg_month            47243 non-null  int64
 13  fuelType             47243 non-null  object
 14  brand                47243 non-null  object
 15  notRepairedDamage    47243 non-null  object
 16  dateCreated          47243 non-null  object
 17  postalCode           47243 non-null  int64
 18  lastSeen             47243 non-null  object
dtypes: int64(6), object(13)
memory usage: 7.2+ MB
```

Five-step process of critical thinking

1. Identify the Problem.
2. Gather Information.
3. Evaluate the Evidence.
4. Consider Solutions.
5. Choose and Implement.

We will carry our further analysis by these 5 steps of critical thinking which we would be helpful to predict an efficient model

EDA :

Data Mining Implementation Process

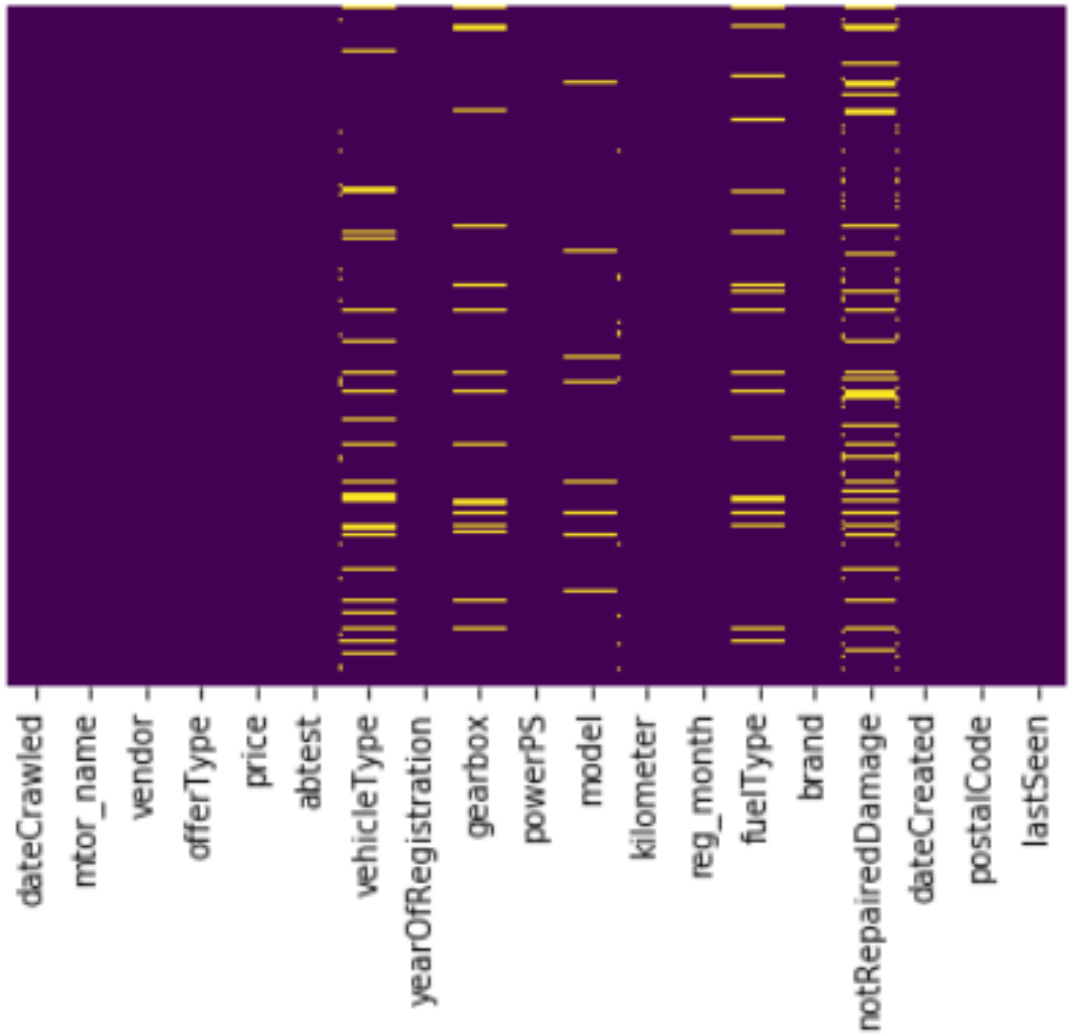
Steps in Data Preprocessing

Here are the steps I have followed;

1. Import libraries
2. Read data
3. Checking for missing values
4. Checking for categorical data
5. Standardize the data
6. PCA transformation
7. Data splitting

DATACLEANING:

IN REAL WORLD THERE ARE SOME INSTANCES WHERE PARTICULAR ELEMENT IS ABSENT BECAUSE OF VARIOUS REASONS SUCH AS CORRUPT DATA,FAILURE TO LOAD THE INFORMATION,OR INCOMPLETE EXTRACTION.HANDLING THE MISSING VALUES IS ONE OF THE GREATEST CHALLENGE BECAUSE MAKING THE RIGHT DECISSION ON HOW TO HANDLE IT GENERATES ROBUST DATA MODELS.



YELLOW COLOUR WHICH IS HIGHLIGHTED GIVES US REPRESENTATION OF THE COLUMNS HAVING THE NULL VALUES.(VEHICLE TYPE,GEARBOX,MODEL,FUEL TYPE,NOTREPAIRED DAMAGE))

REMOVED THE NULL VALUES:

1:DEALT WITH THE MISSING VALUES BY DROPPING FEW ROWS AND COLUMNS THAT CONTAIN THEM AND ALSO DONE MODE IMPUTATION FOR FEW FILLING FEW MISSING VALUES AS NULL VALUES WOULD RESULT IN BIAS RESULTING FROM DIFFERENCES BETWEEN MISSING AND COMPLETE DATA.

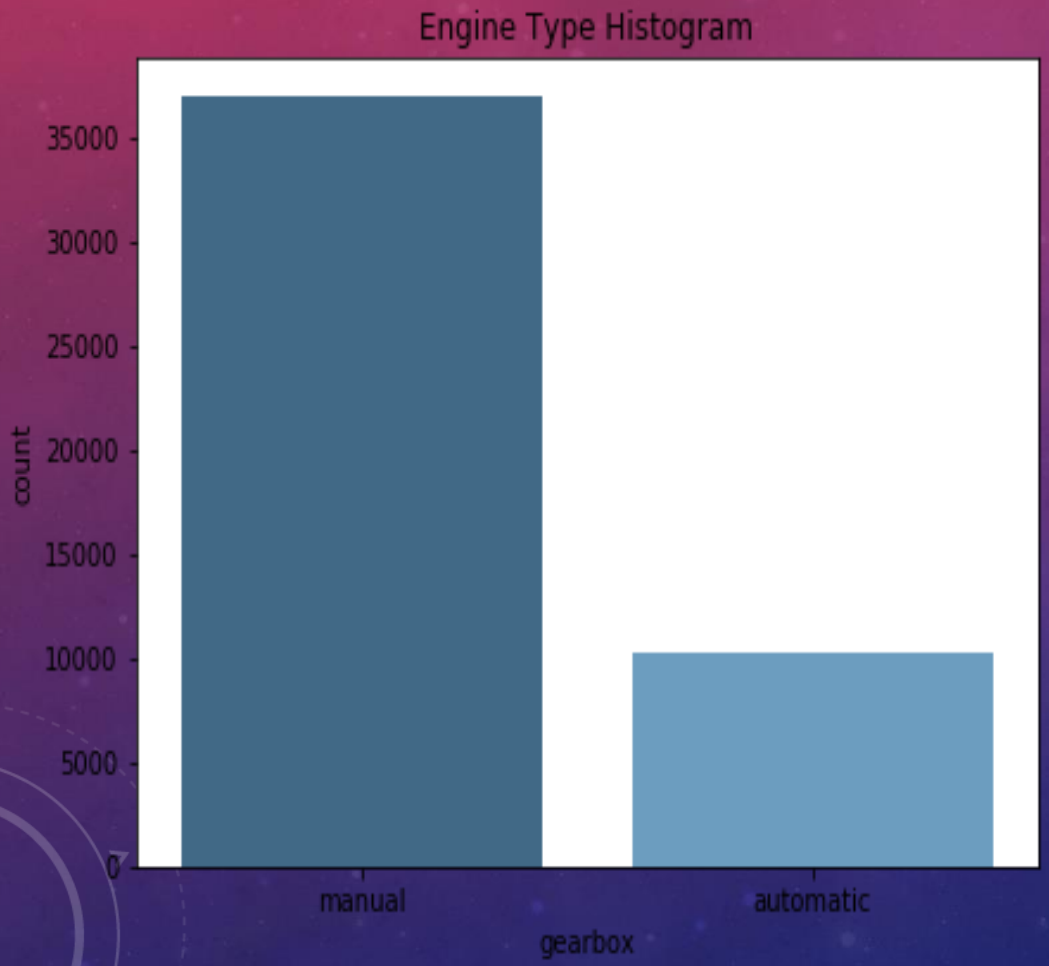
2:THE DATASET ALSO CONTAIN OUTLIERS AND THESE ARE THE VALUES OR OBSERVATIONS THAT ARE DISTANT FROM OTHER OBSERVATION IN THE COLUMN SO OUTLIERS WERE INDENTIFIED FROM RELEVANT COLUMNS AND REMOVED.

Now updated null value in data set is:

```
data.isnull().sum()
```

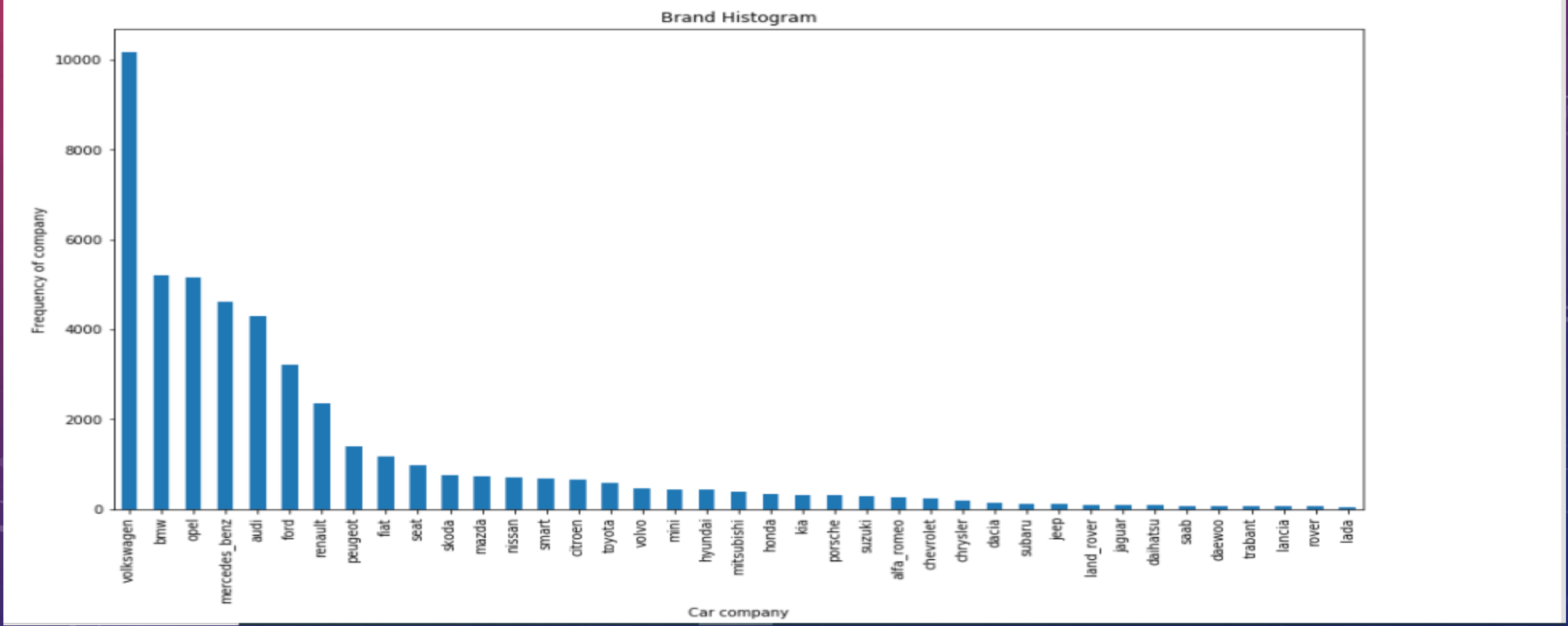
```
dateCrawled            0
mtor_name              0
vendor                 0
offerType              0
price                  0
abtest                 0
vehicleType            0
yearOfRegistration     0
gearbox                0
powerPS                0
model                  0
kilometer              0
reg_month              0
fuelType               0
brand                  0
notRepairedDamage      0
dateCreated            0
postalCode             0
lastSeen               0
dtype: int64
```

EDA ANALYSIS:

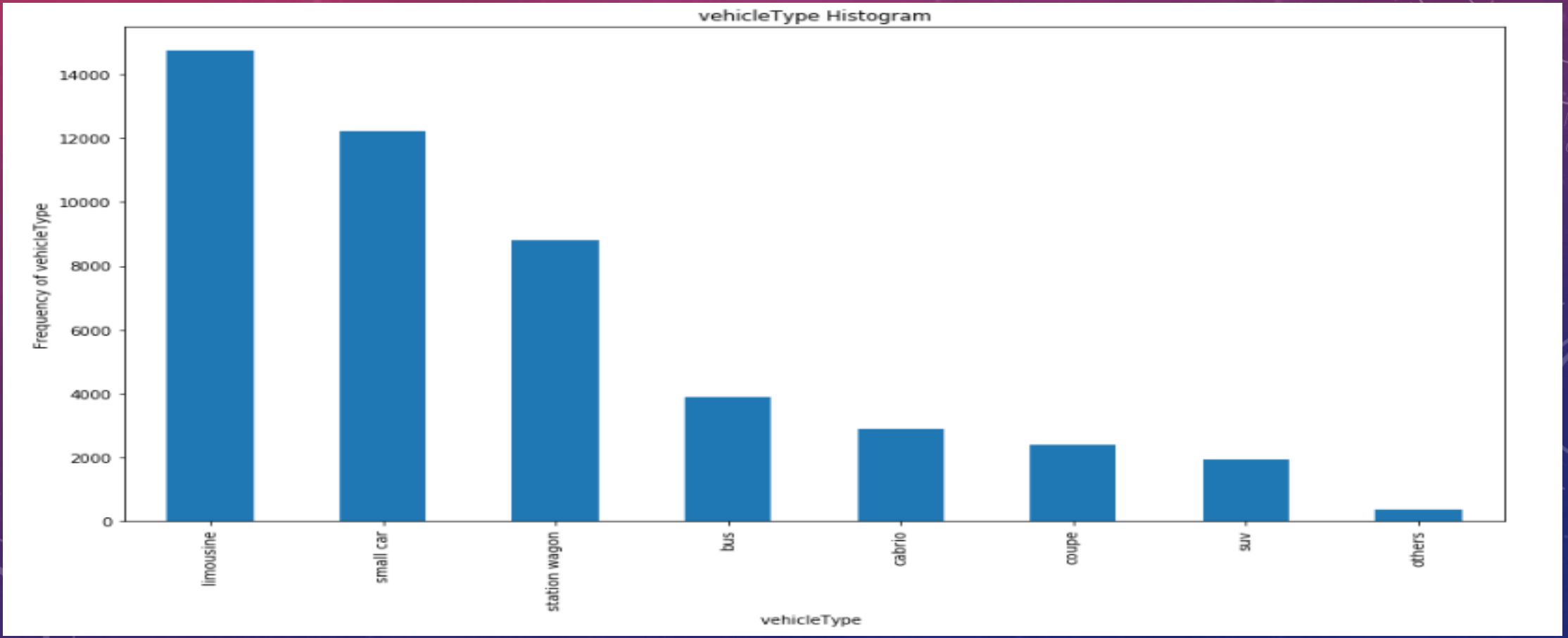


- INFERENCE: The representation shows that resale of manual gearbox cars are more as compared to automatic gearbox cars.

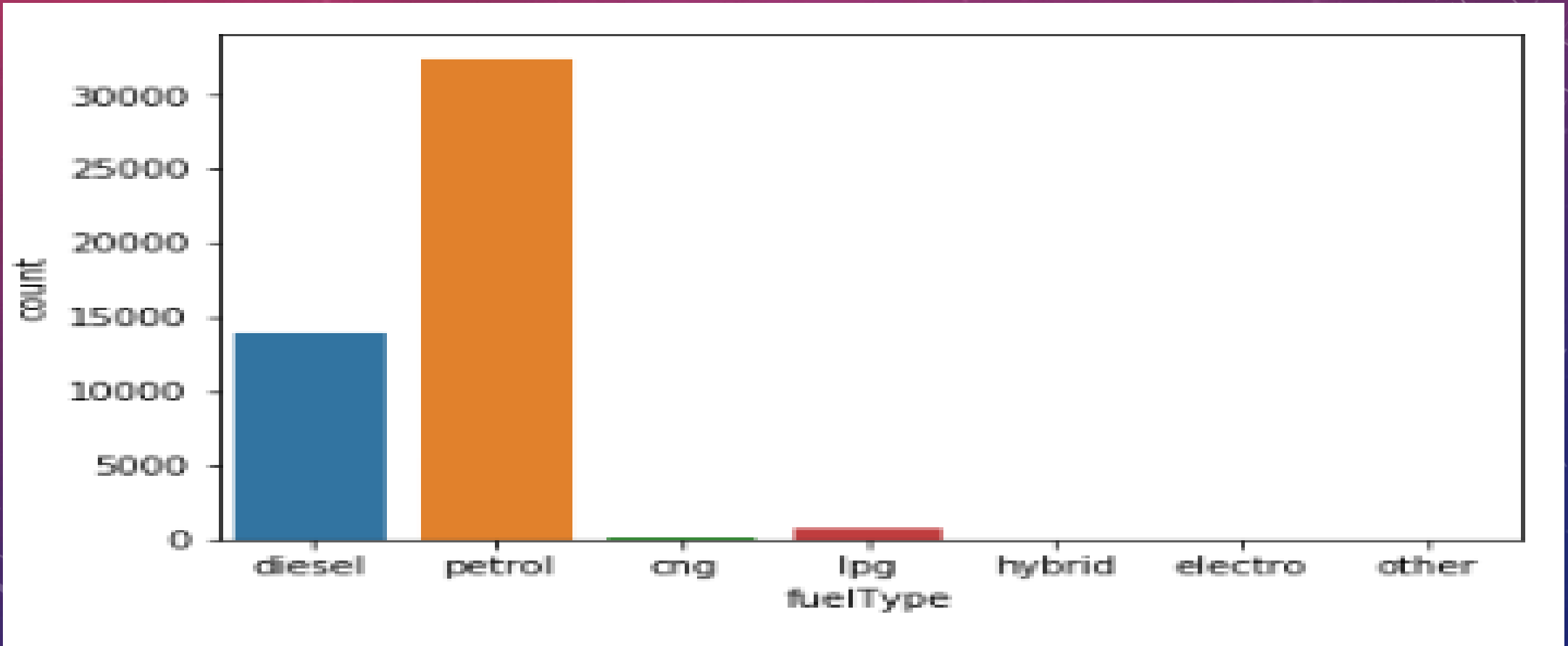
INFERENCE:The graph shows that VOLKSWAGEN seemed to be favored car company or brand for the resale .



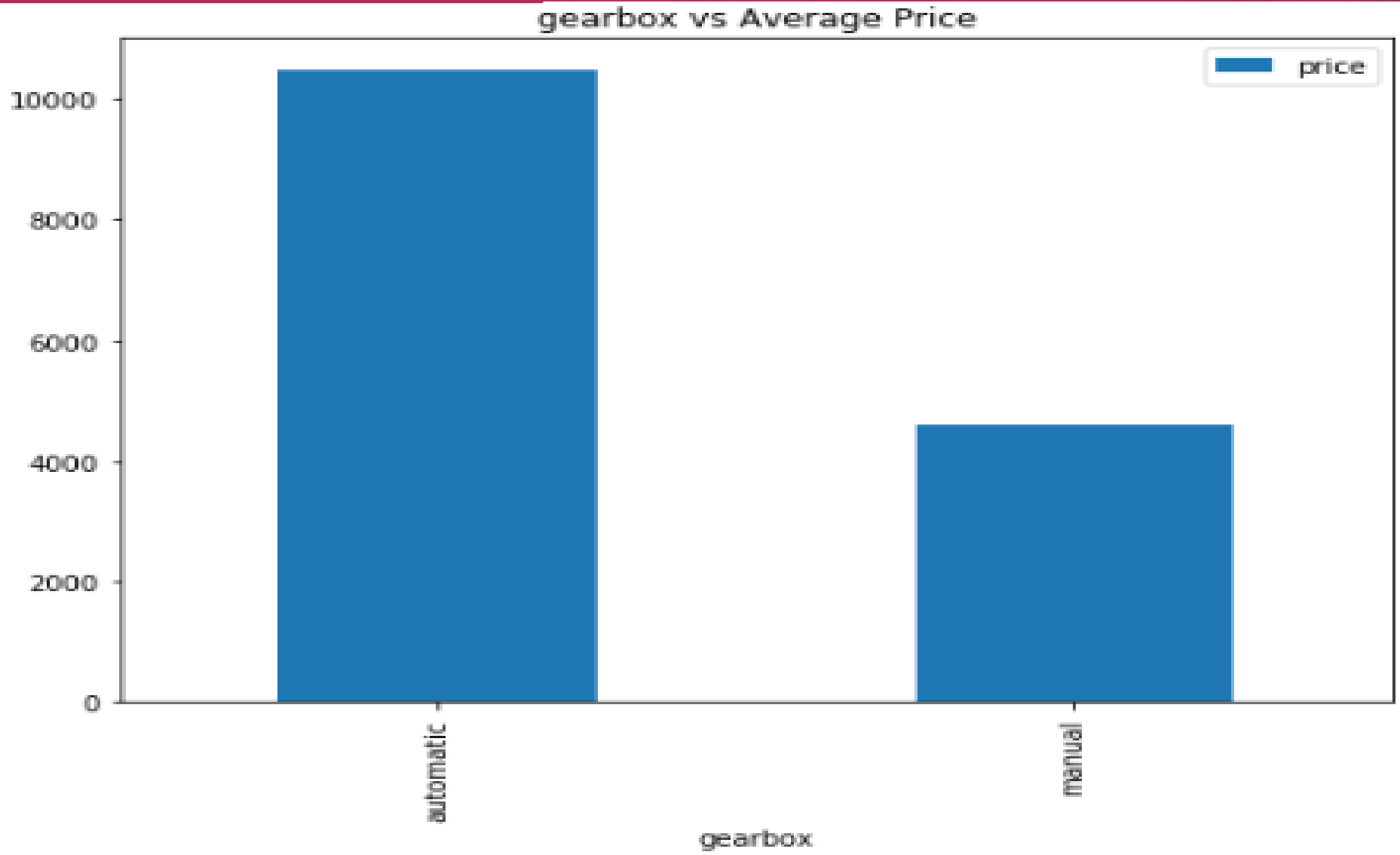
INFERENCE: The above representation shows that resale of vehicle Type limousine are much higher then others.



The representation shows that petrol type fueled cars are mostly being preferred by our customers.

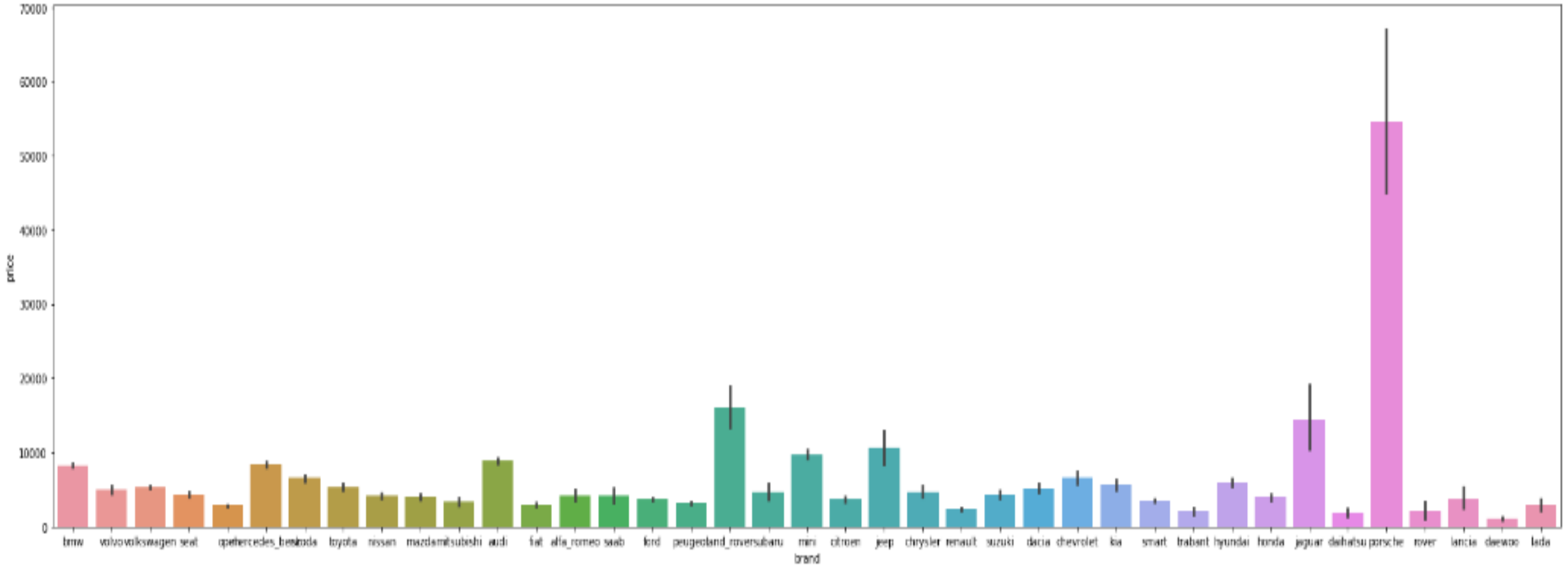


EDA ANALYSIS:

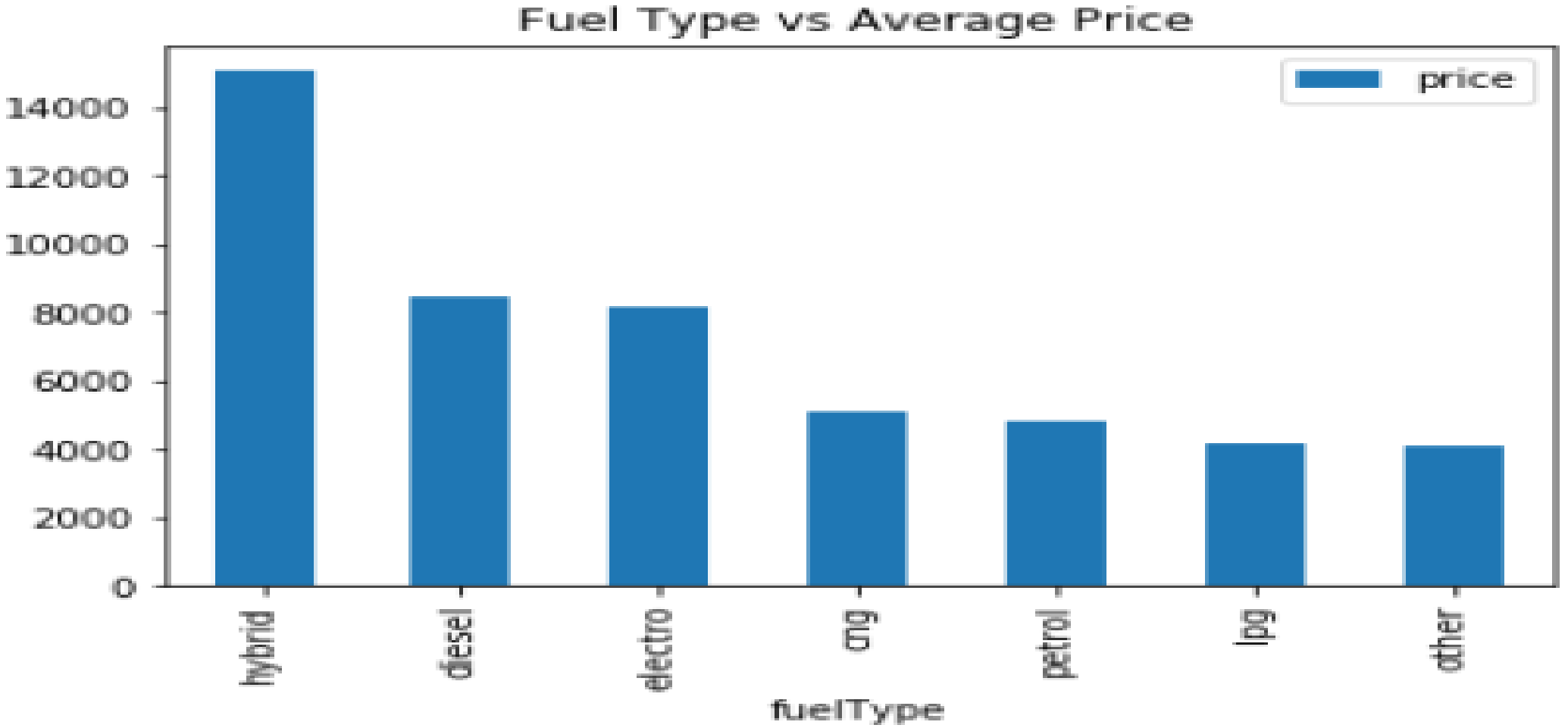


- Average price of automatic gearbox vehicles are more than manual gearbox vehicles.

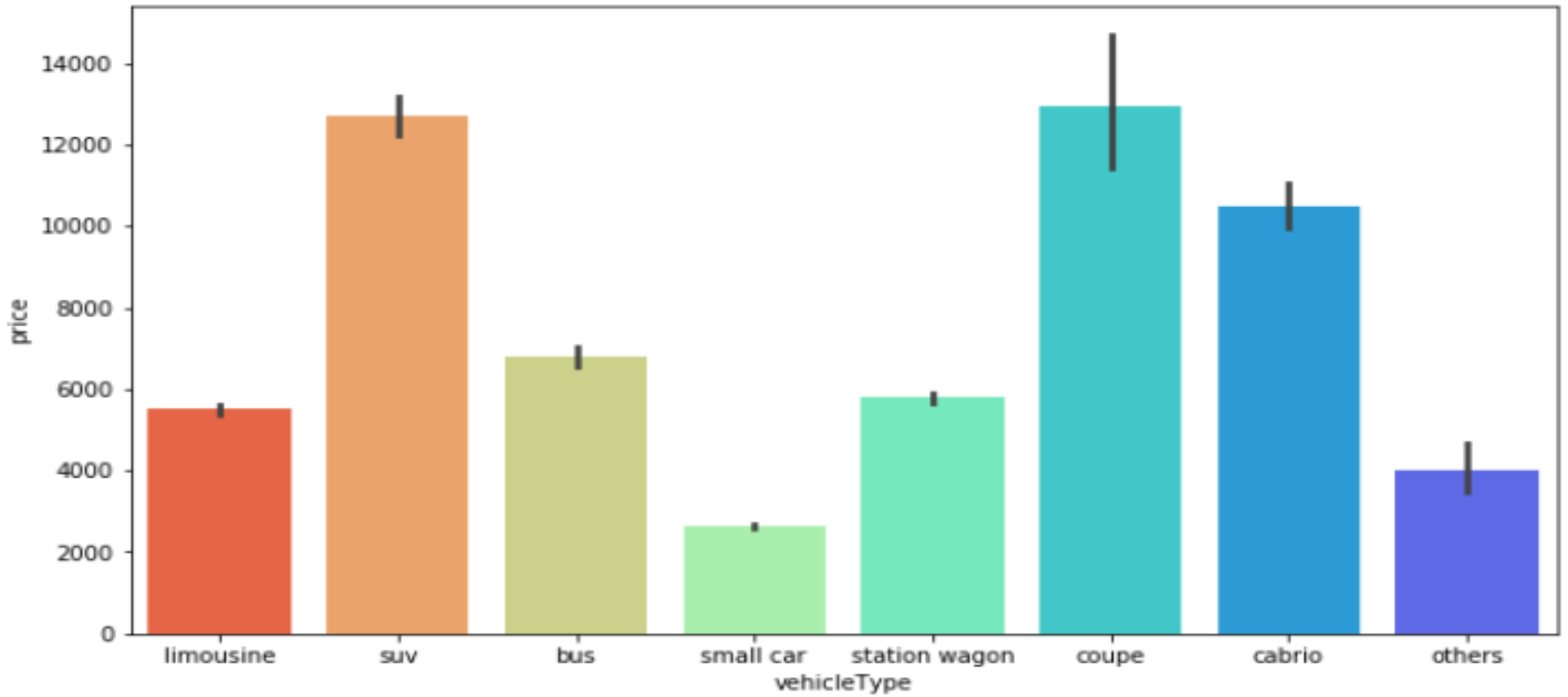
INFERENCE: From presentation we can infer that price of porsche brand is higher than other vehicles.



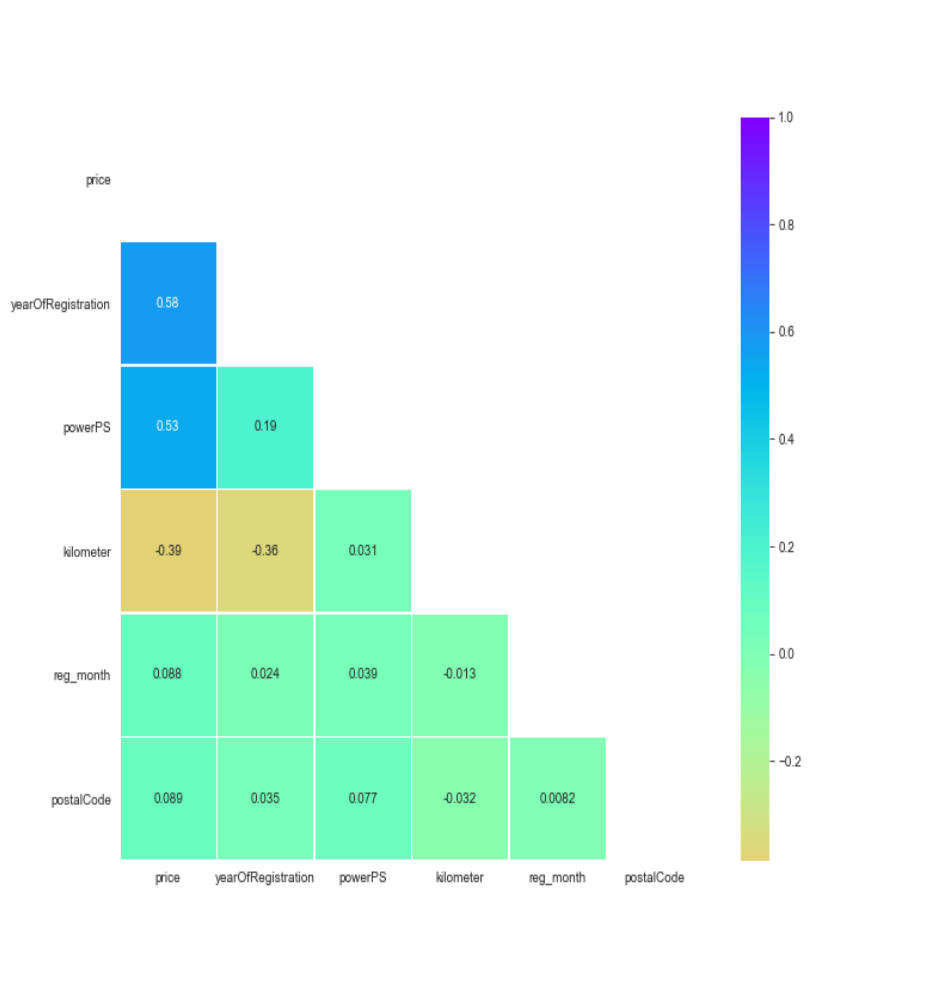
INFERENCE: From presentation we can infer that Hybrid fueltype has higher average price than others.

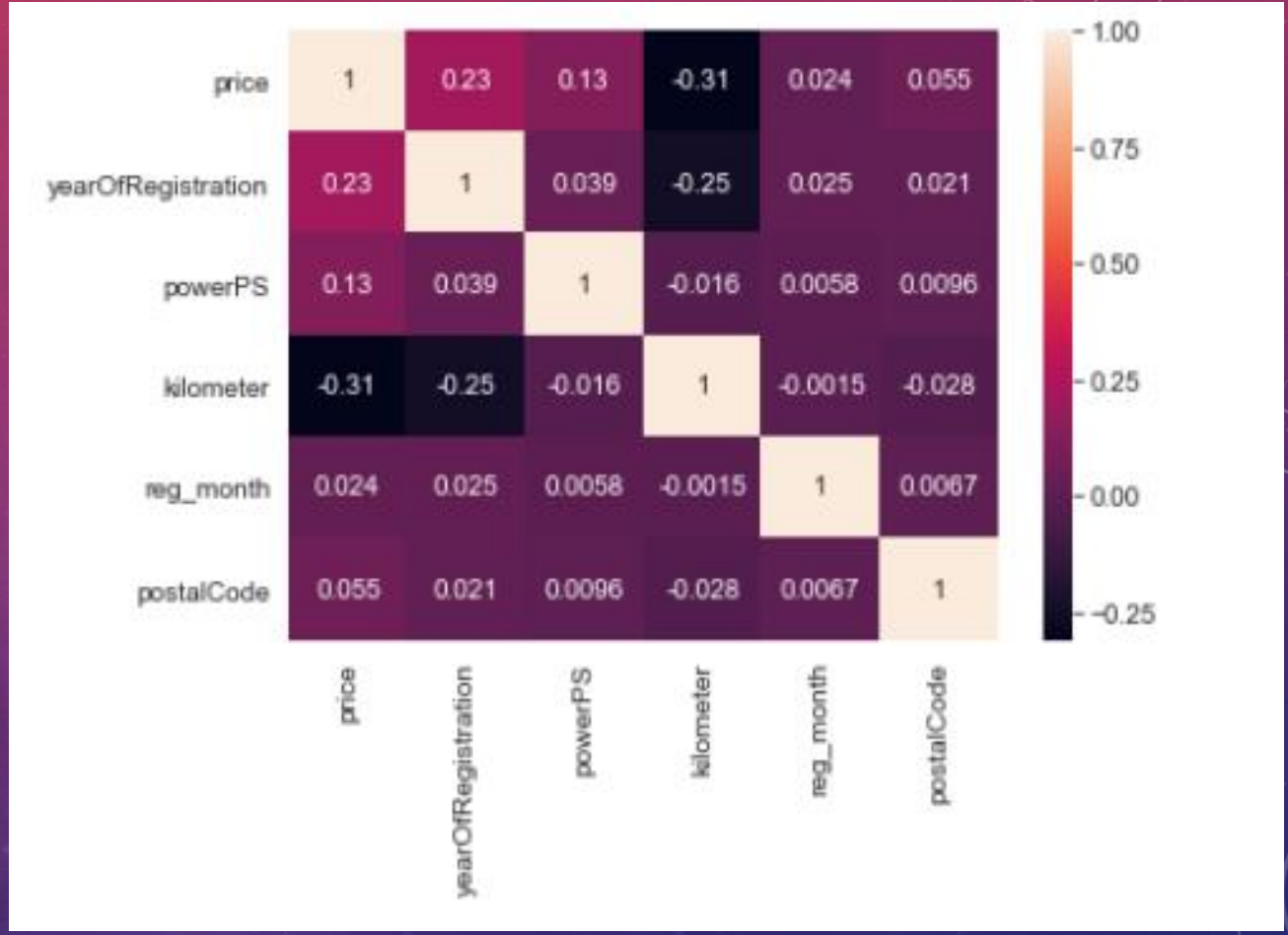


Coupe vehicle type value is impacting more in price .

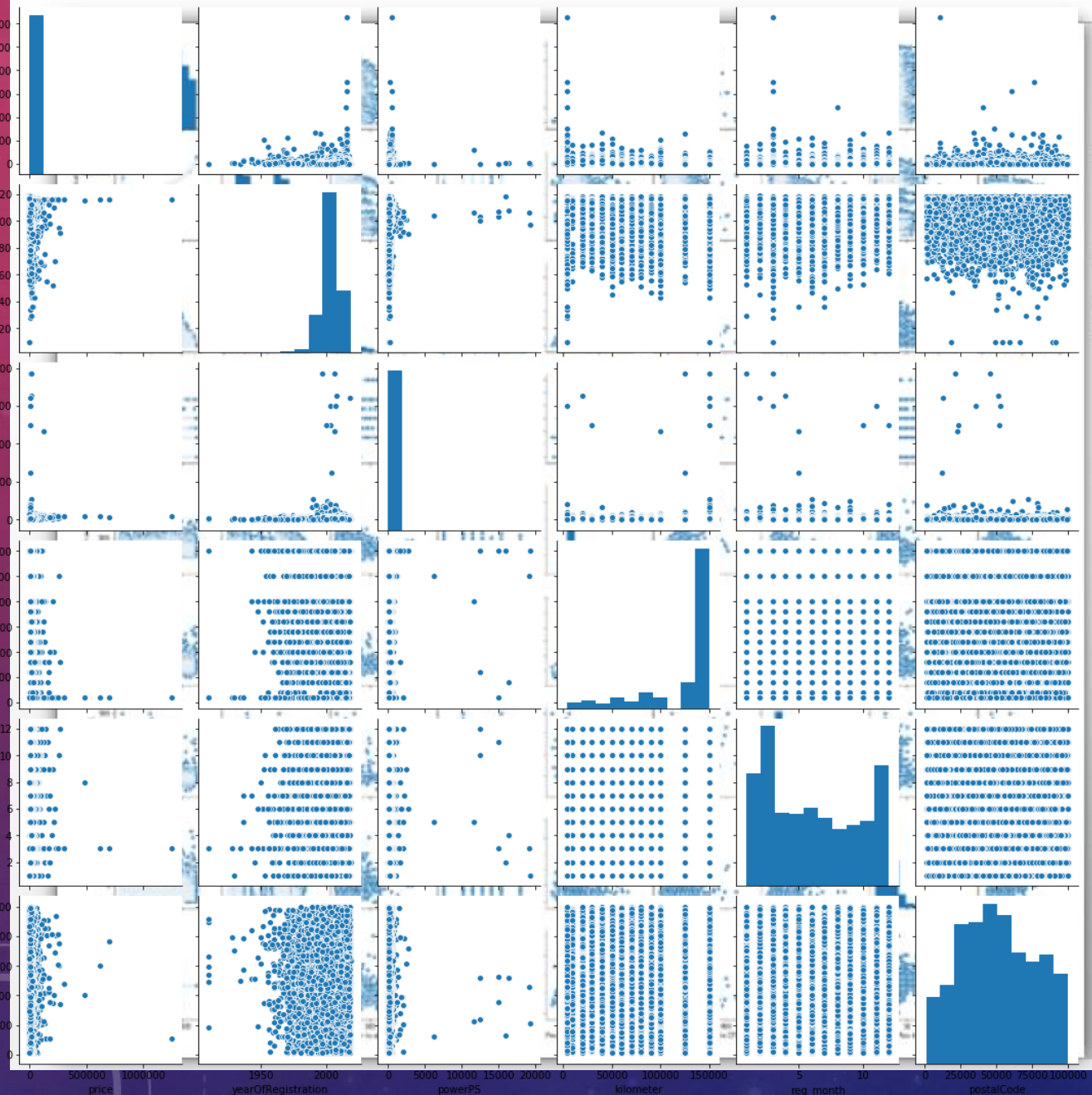


- **INFERENCE:with respect to price coupe vehicleType is on high demand higher than others.**





We can see the correlation between the different attributes.



The pair plot shows how the data is spread with respect to individual column

```
data.groupby(["brand","vehicleType"])["fuelType"].count().reset_index().sort_values(by="fuelType",ascending=False).head(10)
```

	brand	vehicleType	fuelType
251	volkswagen	limousine	3788
17	bmw	limousine	2822
253	volkswagen	small car	2415
166	opel	small car	2183
132	mercedes_benz	limousine	2141
9	audi	limousine	1892
254	volkswagen	station wagon	1651
12	audi	station wagon	1578
248	volkswagen	bus	1367
76	ford	small car	1305

VOLKSWAGEN,BMW ,AUDI ARE THE MOST TOP BRANDS WHICH ARE MOSTLY OF LIMOUSINE VEHICLE TYPE BEEN PRUCHASED BY OUR CUSTOMERS.

The Machine Learning Process

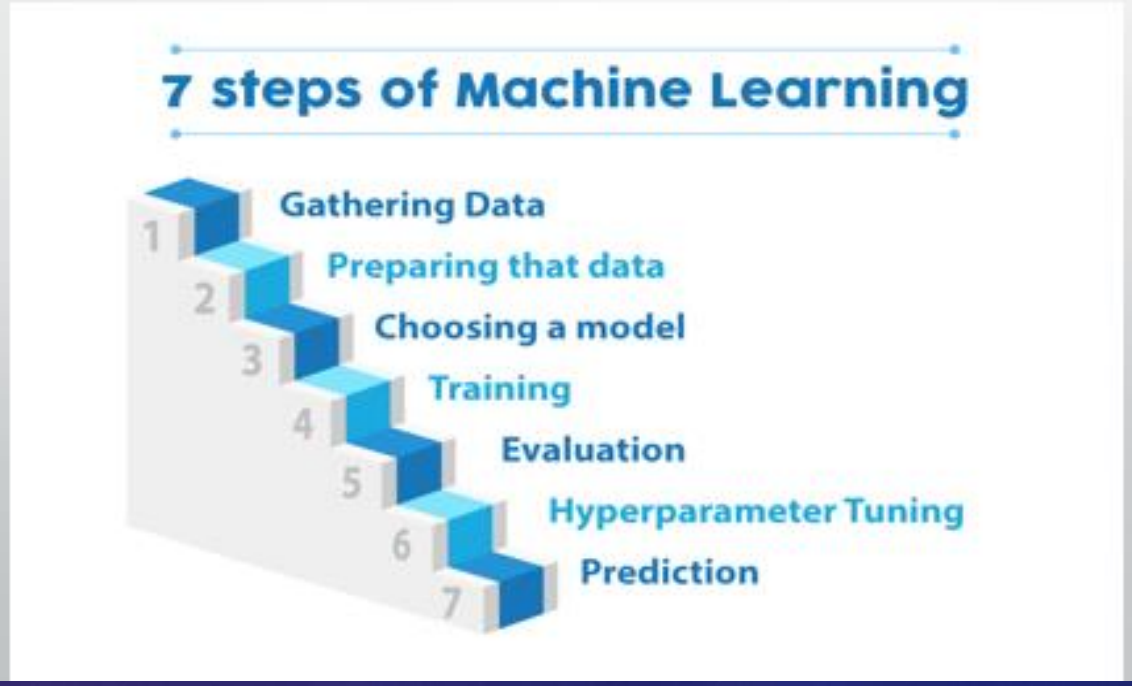
Step 1
Gathering data from various sources

Step 2
Cleaning data to have homogeneity

Step 3
Model Building- Selecting the right ML algorithm

Step 4
Gaining insights from the model's results

Step 5
Data Visualization- Transforming results into visuals graphs



SPLITTING THE DATA IN TO TEST AND TRAIN SETS:

1: The dataset was split in to training data and test data.

2: The training data is the data on which we train and fit our model basically to fit the parameters where as test data is used only to assess performance of model.

3:Training data’s output is available to model where as testing data is the unseen data for which predictions have to be made.

TRAINING MODELS:

1:Predicting the price of a used car is a regression problem.

2:Different types of regression models can be implemented using python Scikit-Learn.

3:The Scikit-learn is a python machine learning library.

4:The training dataset would be trained or fitted using LINEAR

ACCURACY PREDICTIONS:

1: Get the predictions by providing the test data to the linear Regression models.

2:This would give us the prediction accuracy score.

3:The prediction accuracy score from each model would be used as one the basis to determine the best model.

LINEAR REGRESSION MODEL:

1:In linear regression model is constructed that enables us to predict the value of new data considering the training data used to train the model. Regression models a target prediction value based on independent variables.it is mostly used for finding out the relationship between variables and forecasting.

2: Prediction Accuracy Score of this model is: 56.50 %

3:Root Mean Squared Error Metric: 0.5650