****

**USED CAR PRICE PREDICTION**

**Submitted by:**

**ABHISHEK KUMAR**

**ACKNOWLEDGMENT**

I would like to thanks to Flip Robo Technologies to give me a wonderful opportunity. This project is given by my SME Ms Sapna Verma. I have referred below resources that helped and guided me in completion of this project as below:-

<https://www.cars24.com/buy-used-cars/?utm_source=bing&utm_medium=CPC&utm_campaign=C2C_Search&utm_source=bing&utm_medium=CPC&utm_keyword=buy%20cars%20used&utm_matchtype=e&utm_device=c&adgroup_id=1231453256395046&campaign_id=412354997&msclkid=d7f61619d15b1aff5e357768e94ed891&utm_campaign=Search%20%7C%20Buy%20%7C%20NCR_BUYER&utm_term=buy%20cars%20used&utm_content=Buy%20second%20hand%20Car>

<https://www.cardekho.com/used-cars+in+faridabad>

<https://in.search.yahoo.com/search?fr=mcafee&type=E210IN1316G0&p=used+cars+for+sale>

<https://www.olx.in/olxautos/?utm_source=bing&utm_medium=cpc&utm_campaign_id=413114177&utm_adgroup_id=1363395291294270&utm_campaign=IN|DNCR|GS|Pros|Web|CPA|Leads-Conversions|Booking|PriceReference&utm_term=used%20cars%20for%20sale&utm_content=&utm_source=bing&utm_medium=cpc&utm_campaign_id=413114177&utm_adgroup_id=1363395291294270&utm_campaign=IN|DNCR|GS|Pros|Web|CPA|Leads-Conversions|Booking|PriceReference&utm_term=used%20cars%20for%20sale&utm_content=&msclkid=8000e09e9ca6140c0abd81efa4b311c7>

**INTRODUCTION**

* **Business Problem Framing**

New cars are a tempting purchase for anyone looking to replace their old vehicle or add to their garage. Financing is often easier with new cars, and they typically come with all the latest technology and safety features. However, no matter how you look at it, buying a new car is generally not a financially sound decision for several reasons.

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.

**Conceptual Background of the Domain Problem**

Take a look at the reasons why buying used is a smarter choice in the long run.

**New Cars Depreciate Immediately**

**Get More For Your Money**

**Certified Pre-Owned Cars Provide Peace of Mind**

**Used Car Variety is Hard to Beat**

**Used Cars Have Data**

**Cut Your Insurance Costs**

**Cut Your Registration Fees**

**Cars Last Longer**

**Vehicle History Reports**

**The Aftermarket Community is Thriving**

* **Review of Literature**

We are required to model the price of used car with the available independent variables.

Technical Requirements:

• Data contains 935 entries each having 20 variables.

• Data contains very few Null values. So no need to treat them.

• Extensive EDA has to be performed to gain relationships of important variable and price.

• Data contains numerical as well as categorical variable. We need to handle them accordingly.

• We have to build Machine Learning models, apply regularization and determine the optimal values.

• We need to find important features which affect the price positively or negatively.

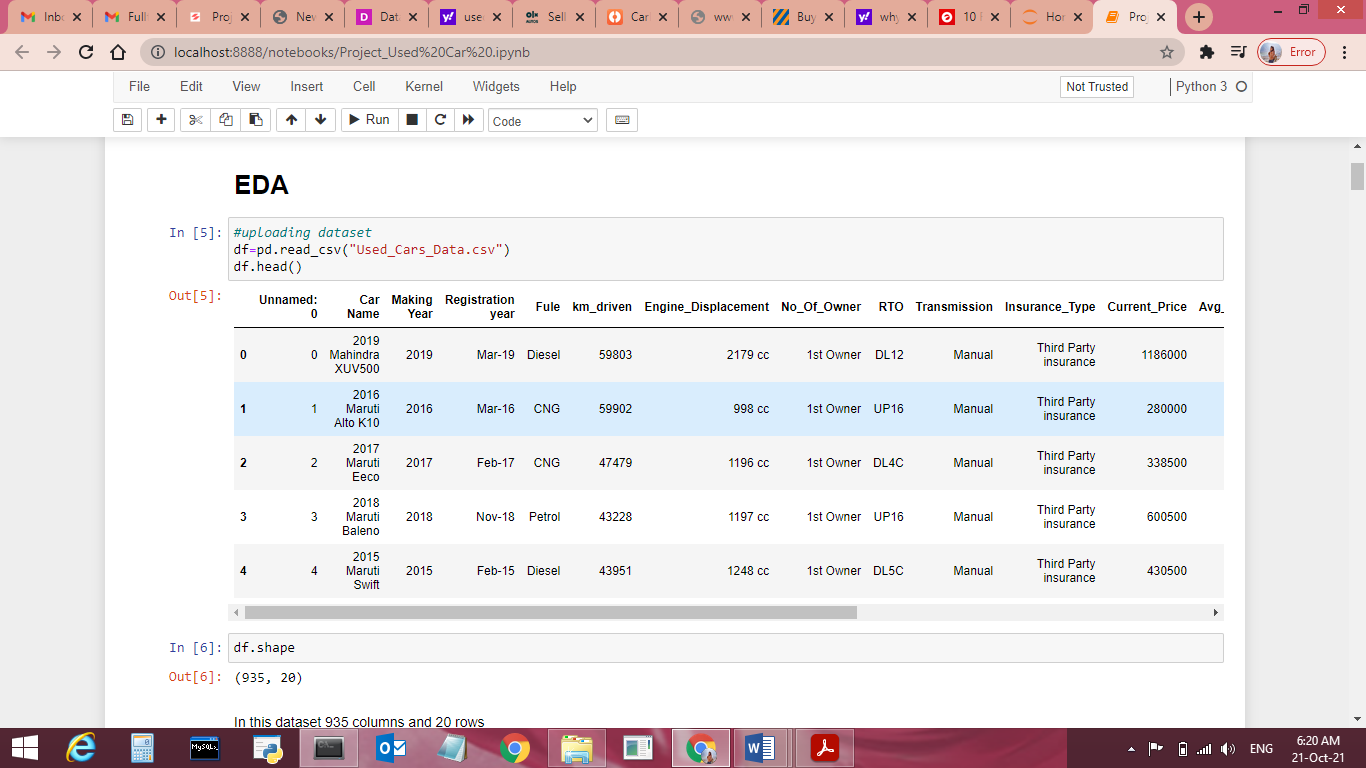
• Two datasets are being used in (.csv).

The “Data file.csv” and “Data description.txt” are enclosed with this file.

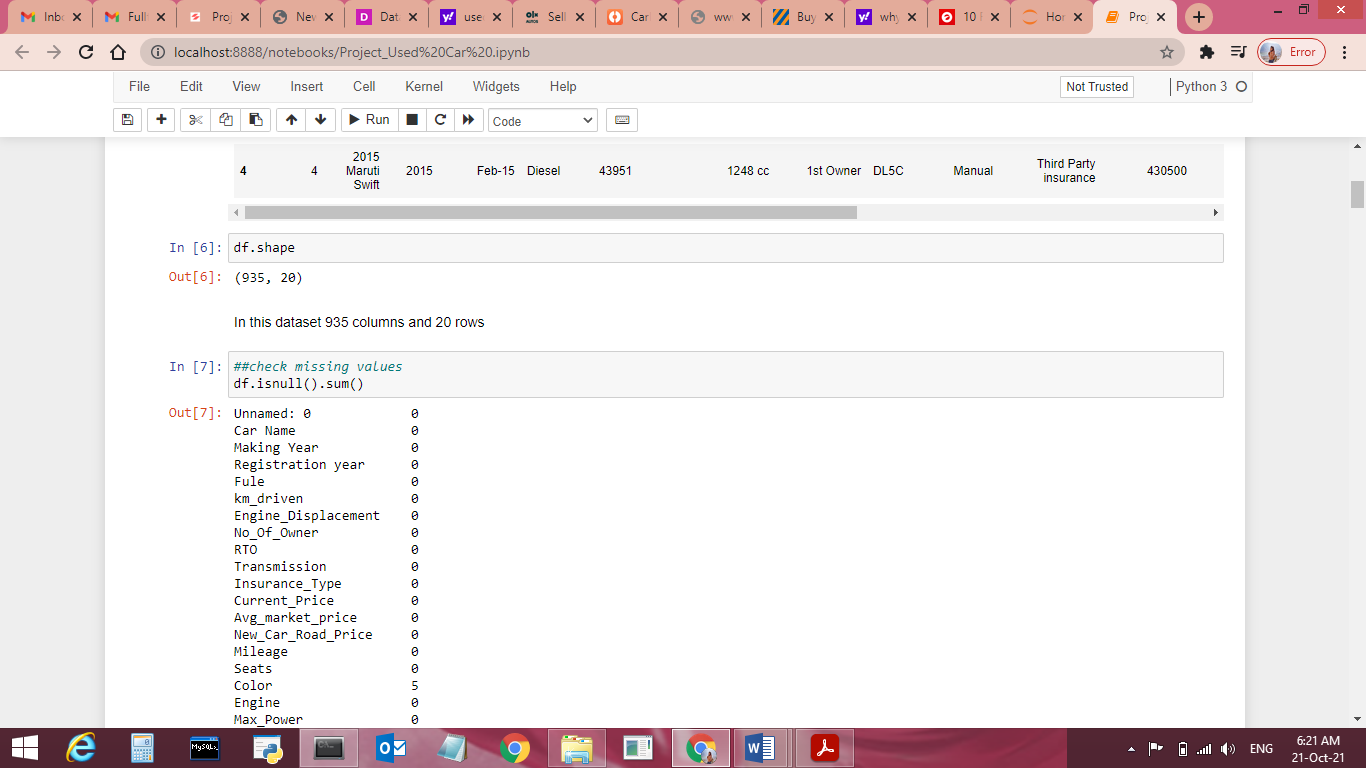
* **Motivation for the Problem Undertaken**
* **1. New Cars Depreciate Immediately**
* If you're trying to decide whether to buy new or used cars, remember that the moment you take your new car off the lot, it goes down in value. Some makes of cars do hold their value better than others, and new cars do last much longer today than they used to. Nevertheless, new cars can still lose as much as 30 percent of their value within the first year.
* Also consider that you've most likely borrowed money to buy that new car, and you're already losing as much as 30 percent of its actual value while making payments on it, including interest. No matter how you view it, buying a new car cannot qualify as a good investment when it loses so much money so quickly.
* **2. Get More For Your Money**
* The depreciation on a new car can actually be a boon when you're a used car shopper. Your budget for a new car might only get you the base trim of your favored model, but if you shop that model a few years back, you might be able to afford a mid-level or top trim. For example, if you buy a vehicle coming out of a two- or three-year lease, you might be able to save as much as 50 percent off the original sticker price.
* **3. Certified Pre-Owned Cars Provide Peace of Mind**
* If you're concerned about not getting a new-vehicle warranty on something used, then it's a good idea to investigate the options available with certified pre-owned cars. Nearly every automaker offers some form of this program to make buying a used car less anxiety-inducing. The CPO system is different with each manufacturer, and it's important to remember that dealer-certified and manufacturer-certified vehicles are not the same.
* Manufacturer certified pre-owned vehicles typically offer a much higher level of protection. The program usually includes some type of warranty based on time or mileage, plus extras like roadside assistance or a free rental when your vehicle needs to spend time in the shop.
* However, don't buy a CPO vehicle just because you think it's the only way to get a warranty on a used vehicle. Many dealerships and individual companies allow you to purchase a warranty on a used vehicle as a way to buy an extra layer of peace of mind.
* **4. Used Car Variety is Hard to Beat**
* Somewhere between 300 and 400 new car models go up for sale every year in the U.S., but the used car market holds even more variety. Carmakers discontinue trims and models on a regular basis, and your perfect car may be lurking a few years back in the trim made for only one year. For example, maybe you're looking for a small pickup. That type of vehicle is harder to find in the current market with crossovers dominating, but if you look used, then Ford Rangers, older Toyota Tacomas and Chevy Colorados and S-10s are everywhere.
* **Cars Last Longer**
* The basic truth is that cars are lasting longer than ever. 100,000 miles is no longer the end of a car's life without Toyota or Honda in the name. Many automakers have models that can go well past that and several even offer warranties that go up to that mileage. It's always good to have your prospective used car purchase inspected by a mechanic, but 100,000 miles or more on the odometer should no longer be a major deterrent.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**
* This problem is a Linear Regression problem. The dataset is in CSV format and It contains 935  training data points and 20 features that might help us predict the selling price of a used car.
* Build a model of housing prices to predict median house values in California using the provided dataset.
* Train the model to learn from the data to predict the median housing price in any district, given all the other metrics.
* Predict used car prices based on data and information we collected.
* **Data Sources and their formats**
* The dataset we collectedfrom different diffrenrt sources and websire like- Car24,Olx, etc..andused it in CSV format. In this dataset, there are 935 rows and 20 columns.



Data Shape:



Dataset null Values:

Unnamed: 0 0

Car Name 0

Making Year 0

Registration year 0

Fule 0

km\_driven 0

Engine\_Displacement 0

No\_Of\_Owner 0

RTO 0

Transmission 0

Insurance\_Type 0

Current\_Price 0

Avg\_market\_price 0

New\_Car\_Road\_Price 0

Mileage 0

Seats 0

Color 5

Engine 0

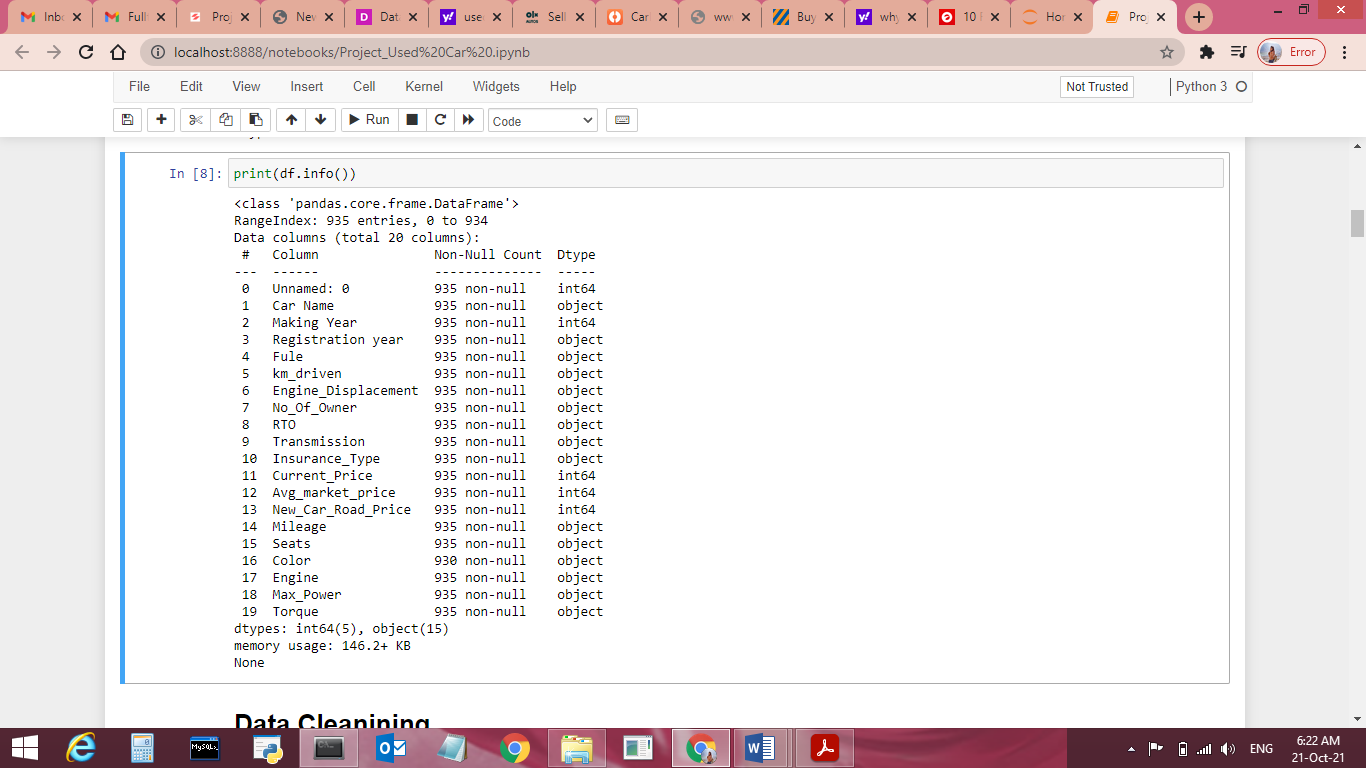
Max\_Power 0

Torque 0

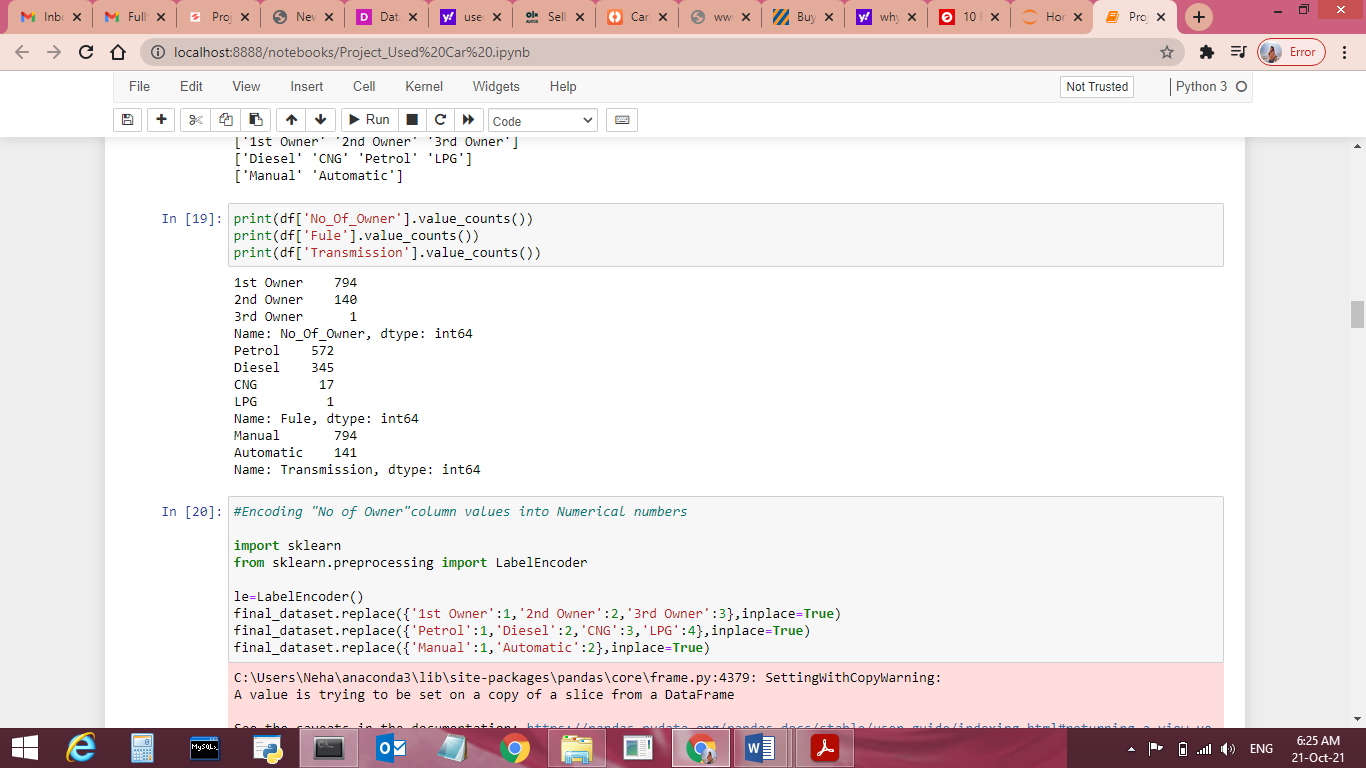
dtype: int64

Dataset Informations:

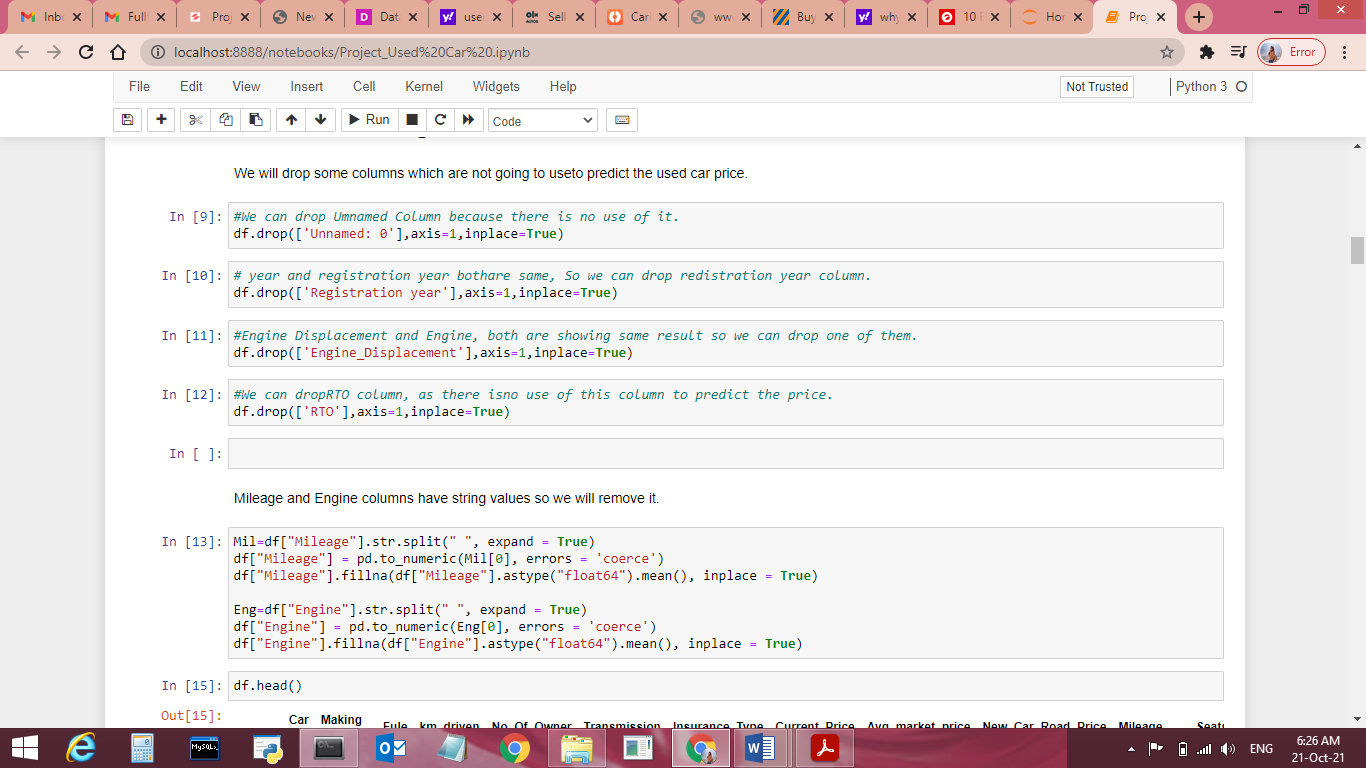
[{"metadata":{"trusted":false},"cell\_type":"code","source":"print(df.info())","execution\_count":8,"outputs":[{"name":"stdout","output\_type":"stream","text":"\nRangeIndex: 935 entries, 0 to 934\nData columns (total 20 columns):\n # Column Non-Null Count Dtype \n--- ------ -------------- ----- \n 0 Unnamed: 0 935 non-null int64 \n 1 Car Name 935 non-null object\n 2 Making Year 935 non-null int64 \n 3 Registration year 935 non-null object\n 4 Fule 935 non-null object\n 5 km\_driven 935 non-null object\n 6 Engine\_Displacement 935 non-null object\n 7 No\_Of\_Owner 935 non-null object\n 8 RTO 935 non-null object\n 9 Transmission 935 non-null object\n 10 Insurance\_Type 935 non-null object\n 11 Current\_Price 935 non-null int64 \n 12 Avg\_market\_price 935 non-null int64 \n 13 New\_Car\_Road\_Price 935 non-null int64 \n 14 Mileage 935 non-null object\n 15 Seats 935 non-null object\n 16 Color 930 non-null object\n 17 Engine 935 non-null object\n 18 Max\_Power 935 non-null object\n 19 Torque 935 non-null object\ndtypes: int64(5), object(15)\nmemory usage: 146.2+ KB\nNone\n"}]}]



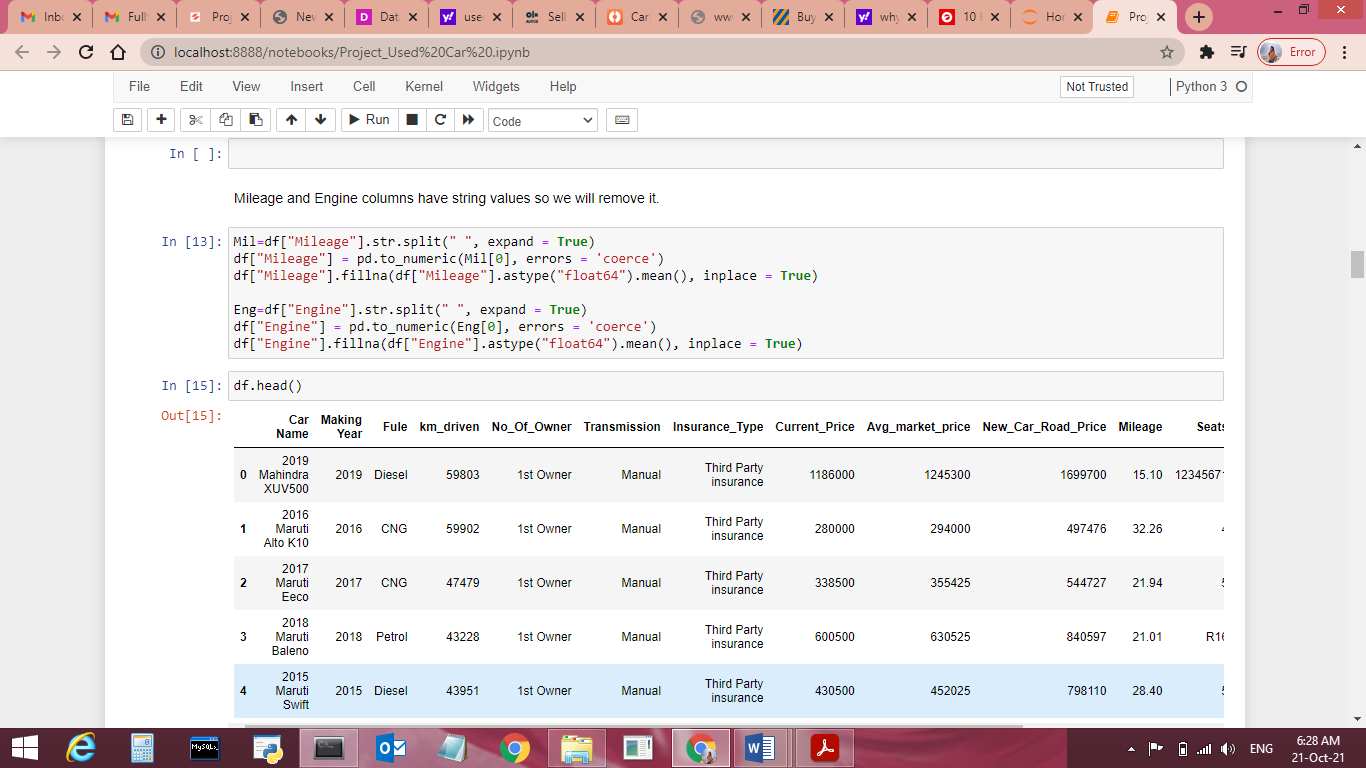
* **Data Preprocessing Done**
  1. I checked the information, data types, null values, correlation of the independent and dependent features and f**rom the correlation table.**
  2. Some columns can’t have any negative value, so those columns were treated accordingly.
  3. Dataset does not have null values so no need to treat them.
  4. Applied LabelEncoer.
  5. Applied various machine learning model and compared it.

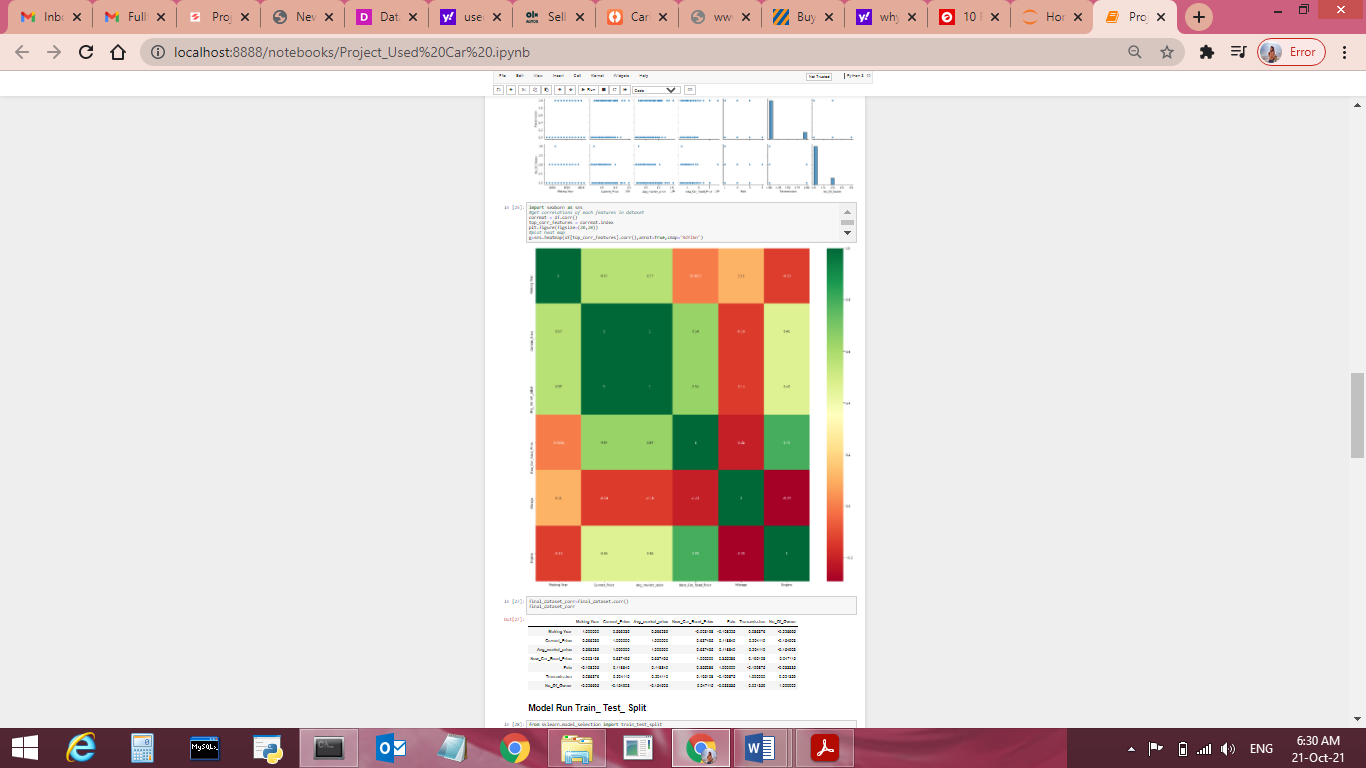


**Dropped some columns which were not used for used car price**

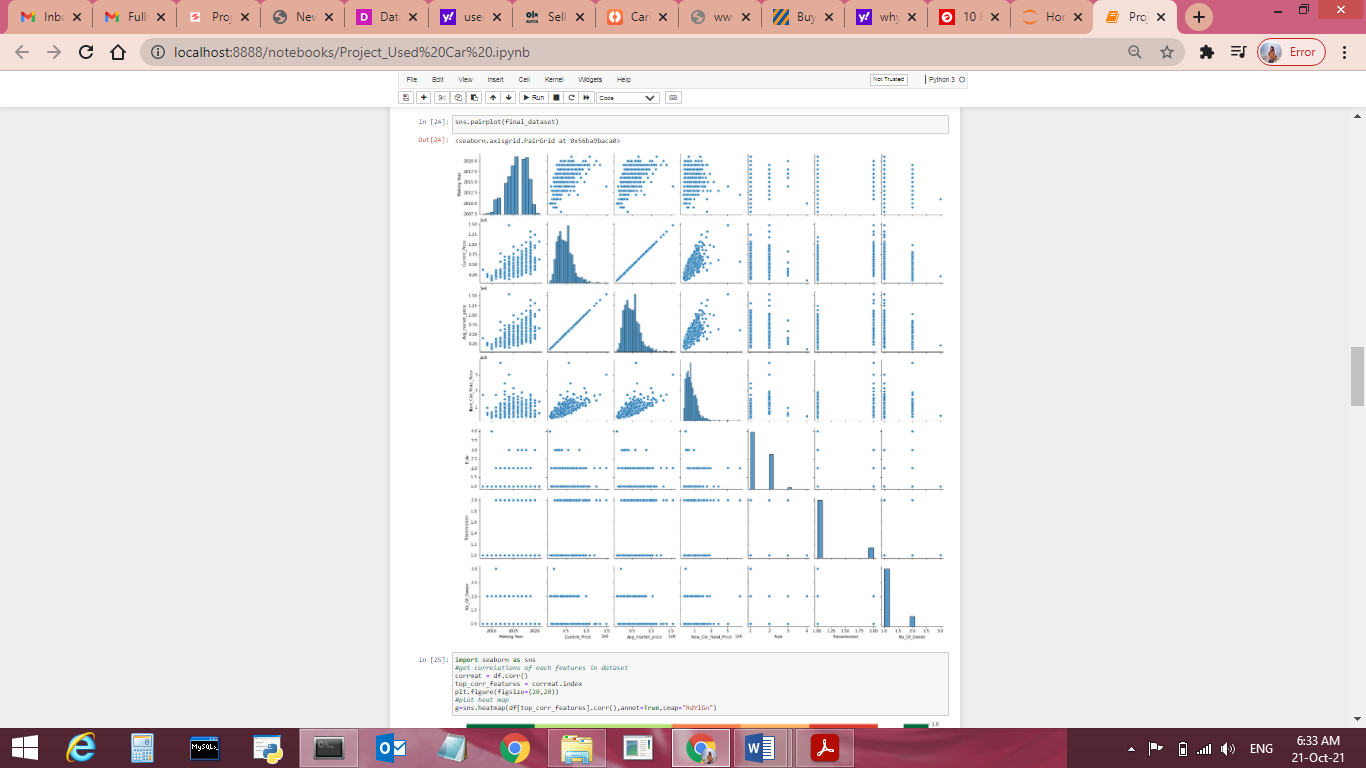


**Convert string values into Numirical values**



* **Data Inputs- Logic- Output Relationships**
* Lets check the correlation with target variable “UsedCarsprice”.
* 

**Pair plotting**



* **State the set of assumptions (if any) related to the problem under consideration**
* I have not consider any pre-assumption , project performance from beginning to end is based on data facts only.
* **Hardware and Software Requirements and Tools Used**

Windows Edition-Windows 8.1 Pro

Processor-Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz 2.00GHz

Installed memory R AM- 4 GB

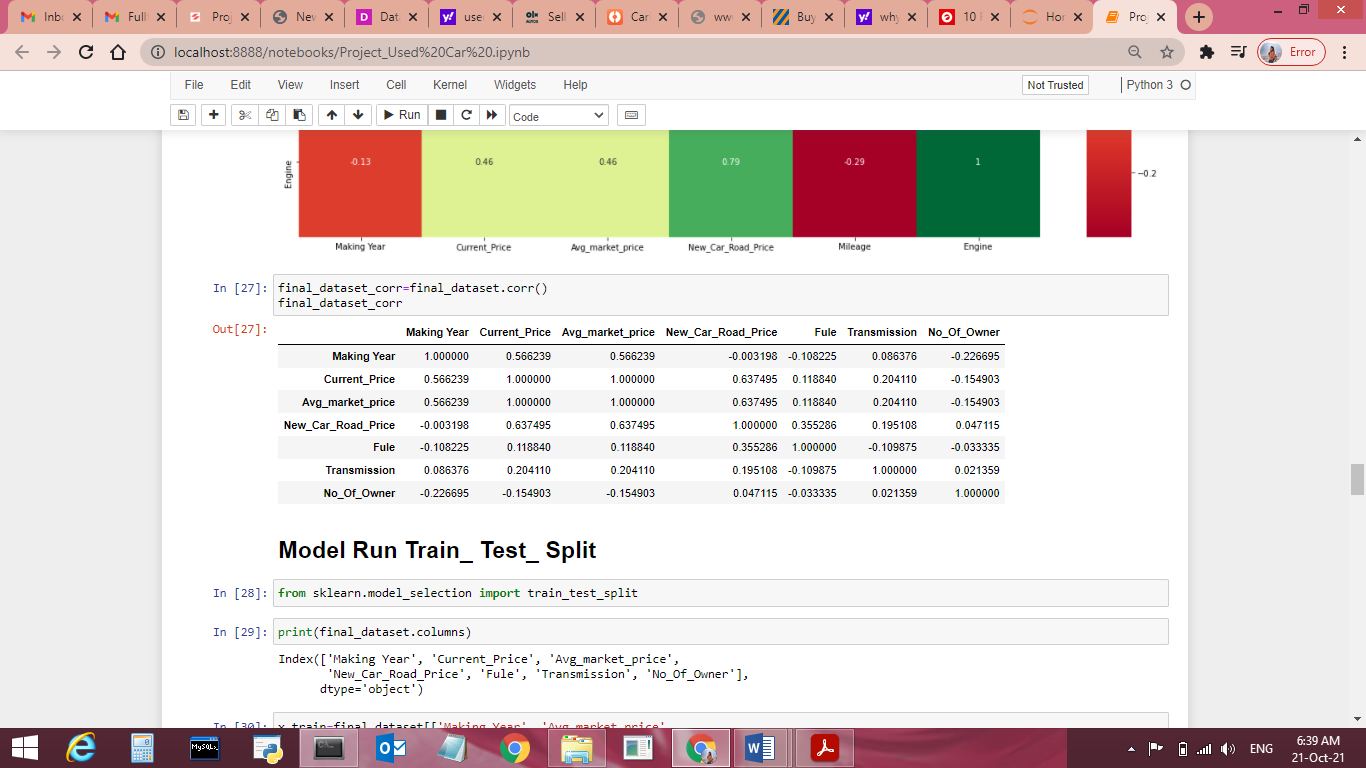
System Type-64 bit OS, x64 based processor

**Software Requirement-** Anaconda 4.9.2 , Python 3.8.5, Jupiter Notebook.

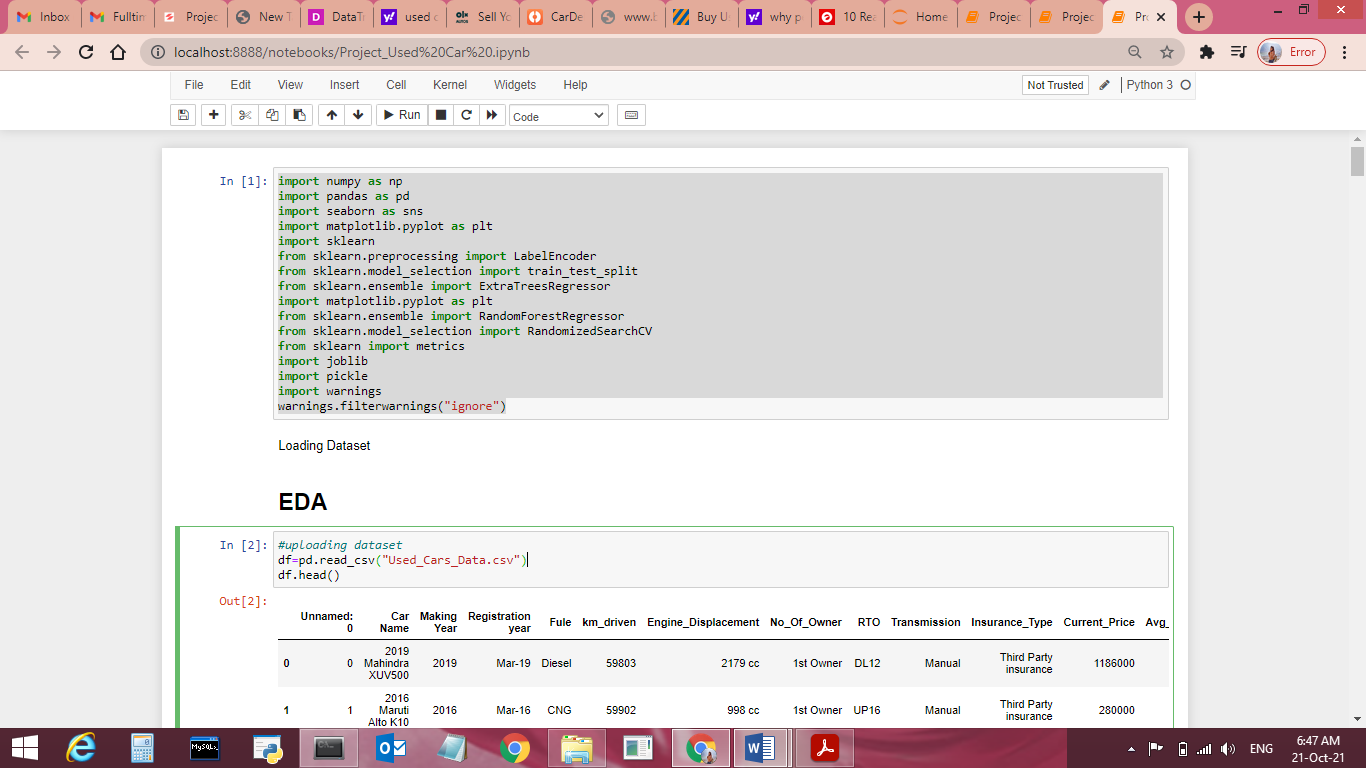
**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**
* **Analytical Approach –**Based on type of data by performing EDA I have decided which model to be used for this data.
* **Statistical Approach –** Data should be in scaled manner, it should not be distorted, for that all values using mean method due to continuous data numbers.

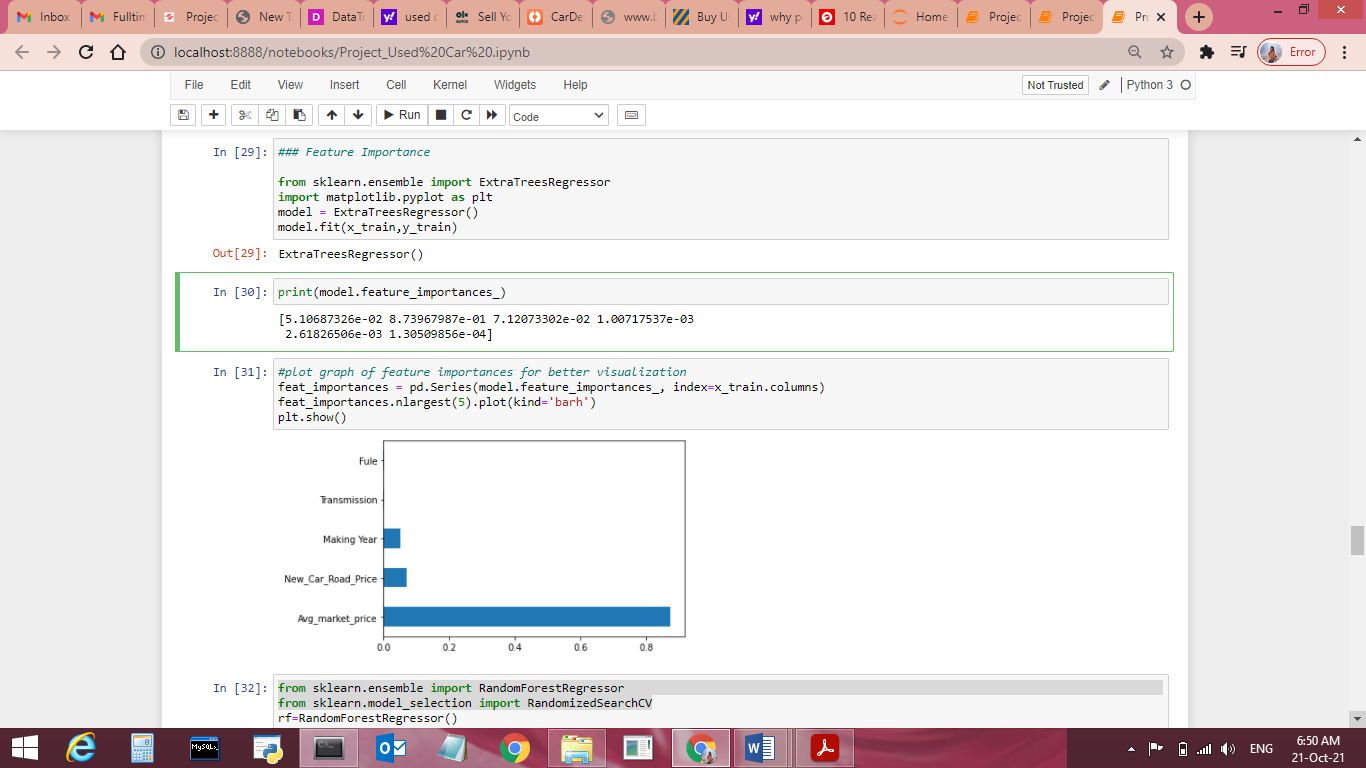
**Statistical Apporch**

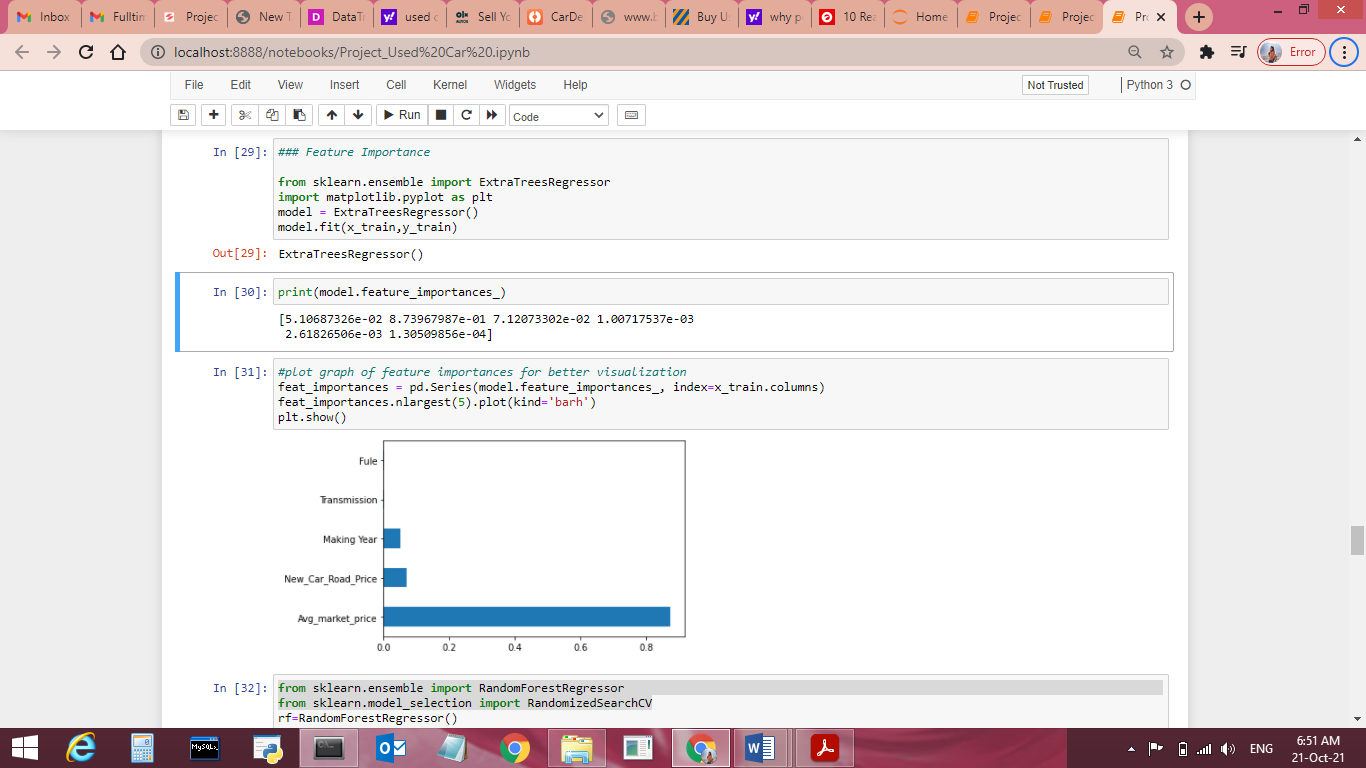


* **Testing of Identified Approaches (Algorithms)**
* All Algorithms list
* import numpy as np
* import pandas as pd
* import seaborn as sns
* import matplotlib.pyplot as plt
* import sklearn
* from sklearn.preprocessing import LabelEncoder
* from sklearn.model\_selection import train\_test\_split
* from sklearn.ensemble import ExtraTreesRegressor
* import matplotlib.pyplot as plt
* from sklearn.ensemble import RandomForestRegressor
* from sklearn.model\_selection import RandomizedSearchCV
* from sklearn import metrics
* import joblib
* import pickle
* import warnings
* warnings.filterwarnings("ignore")

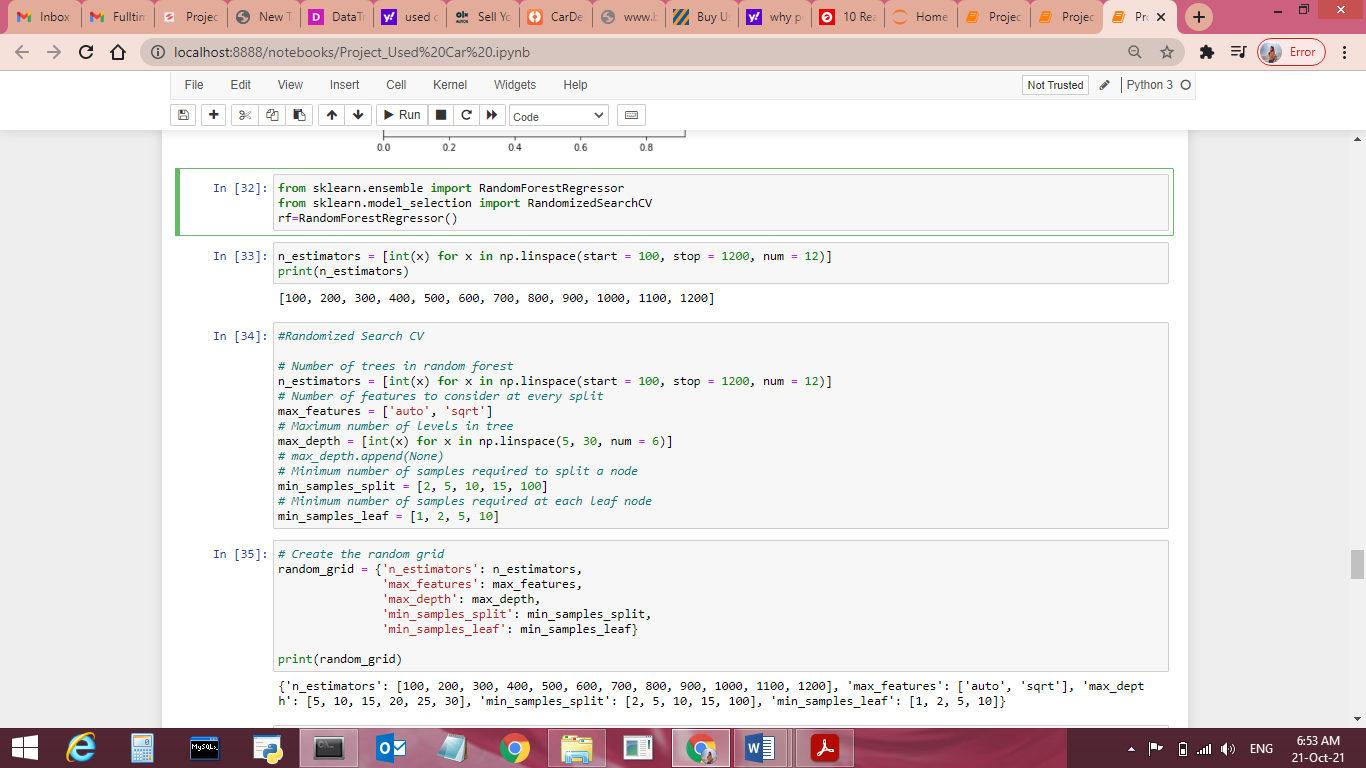


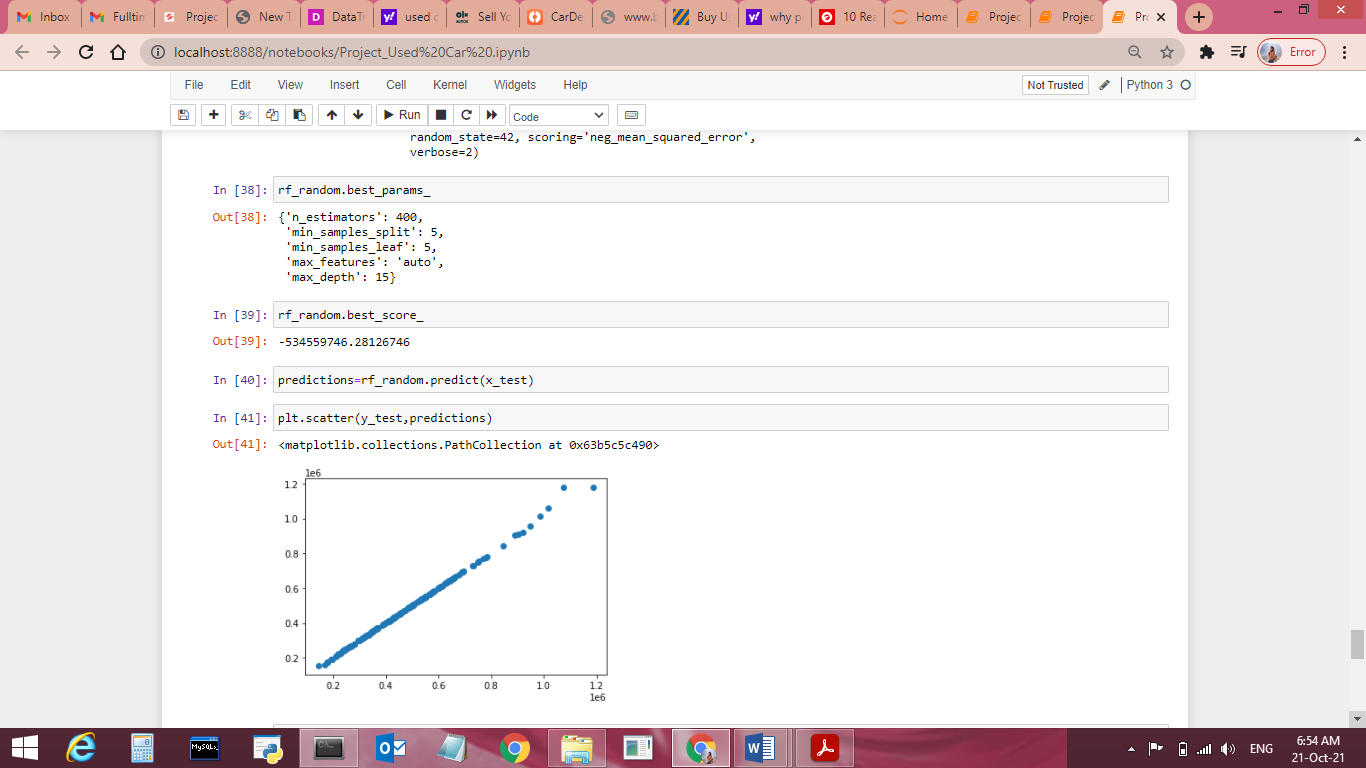
* **Below are Linear Regression algorithms used for the training and testing this dataset**.

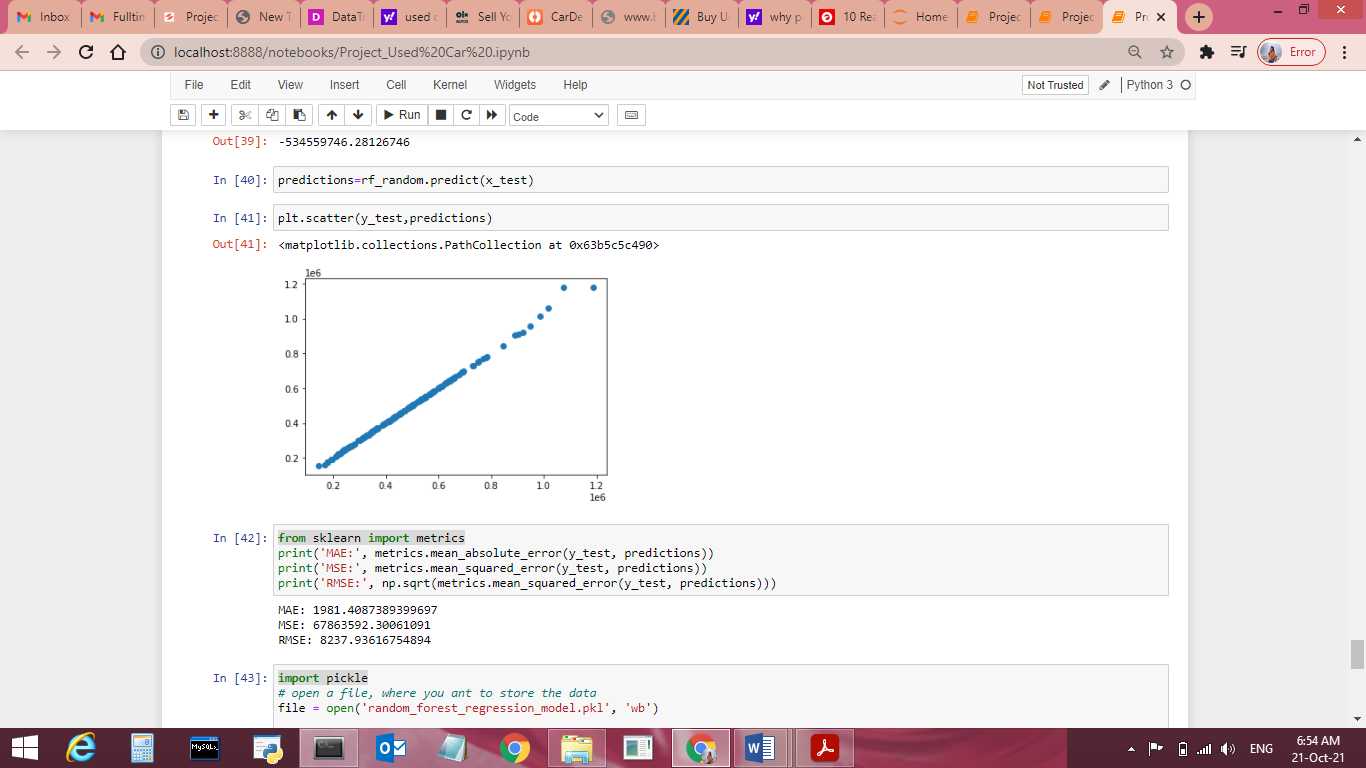




* **Run and Evaluate selected models**

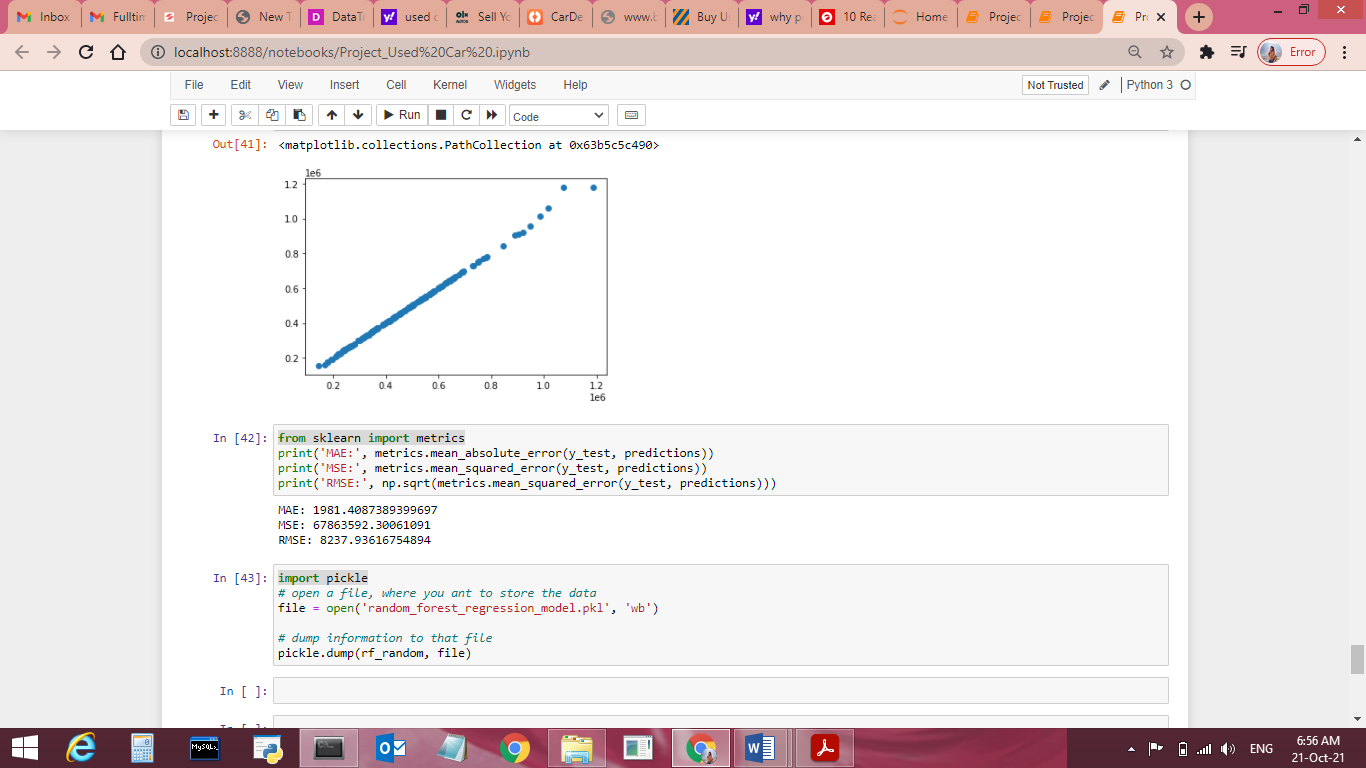






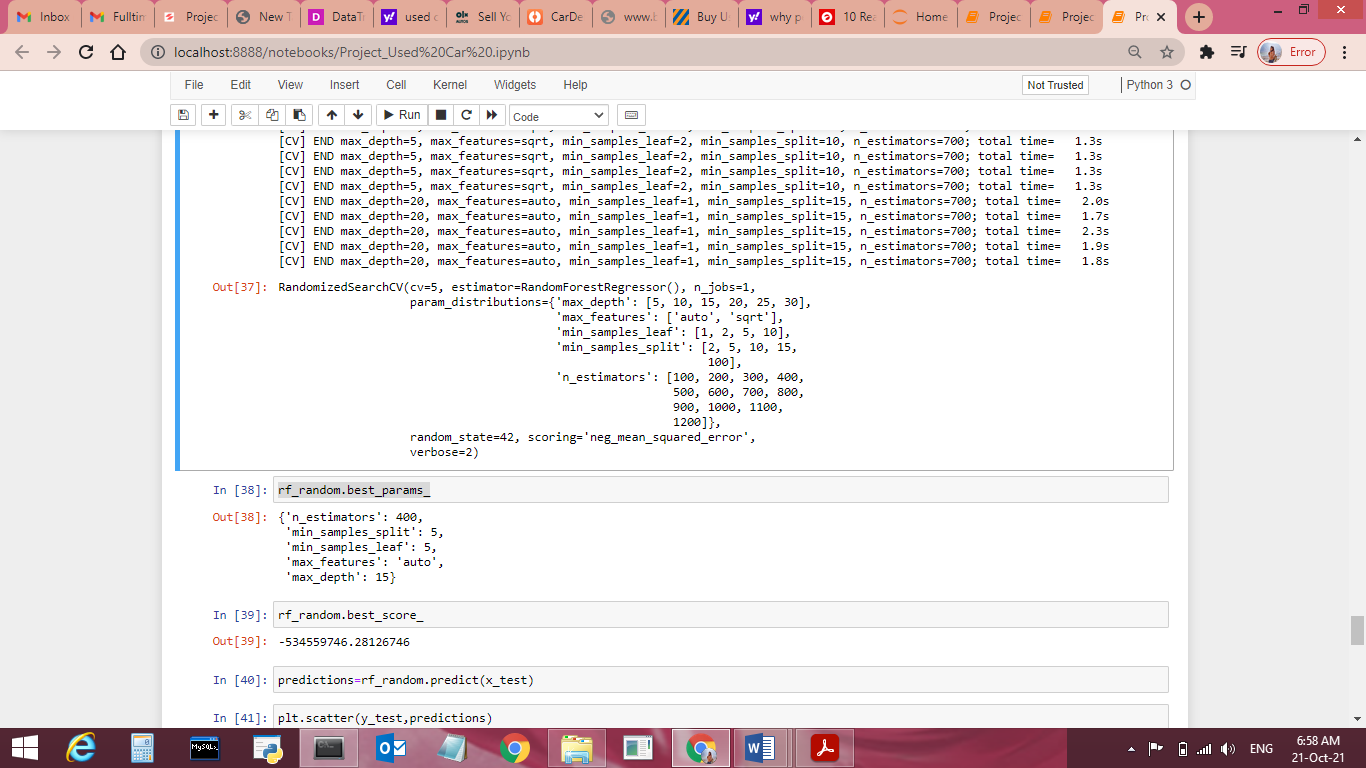
**Key Metrics for success in solving problem under consideration**

* Key Metrices used were the r2 Score and GridsearcCV score as this was Linear Regression problem and we focus more on R2score metrics to observe Mean absolute error, Mean squared error and Root Mean Squared Error.
* **Visualizations**

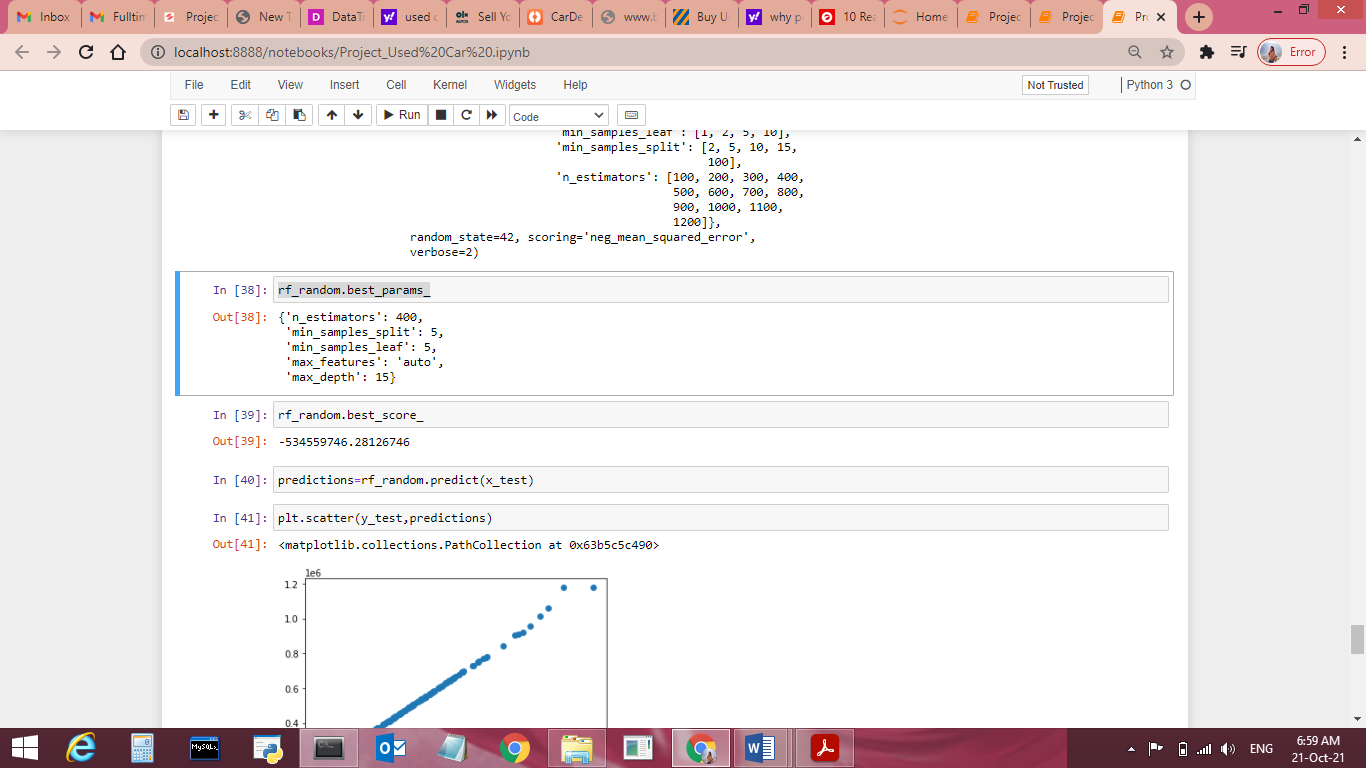


* **Interpretation of the Results**
* Data Pre-processing done by performing EDA (Exploratory Data Analysis), checking for best r2 score.
* We will save our Model by Random forest Regression.

Parameters are as below:



**R2Score and Peramaeter Score:**



**CONCLUSION**

* **Key Findings and Conclusions of the Study**
* Linear regression models assume that the relationship between a dependent continuous variable Y and one or more explanatory (independent) variables X is linear (that is, a straight line). It’s used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).
* **Learning Outcomes of the Study in respect of Data Science**
* This dataset is Linear Regression in nature, we can verify data by using read method & get stats related information for each column using describe method.
* Visualizations, Pre-processing and Data Cleaning part was very crucial as without all these all method we were not able to judge the data effectively and won’t be able to remove the outliers, handling null values and adding into the errors.
* Data contains numerical as well as categorical variable. So we handled them accordingly
* Check the r2 score using Mean absolute error, Mean squared error & get root mean squared error score.
* Train data using Linear Regression models to get the best score & finalise best score giver model for this dataset.
* Get the test score for same model.
* Save file using pickle library.
* **Limitations of this work and Scope for Future Work**
* Visualizations and EDA process, Pre Processing helped a lot in finding out dataset values and helped in finding out the features having direct relation between the feature and the label.
* Its always good to to have complete data while performing model but 7-8 % of data can be excluded based on performance impact.