

PROJECT ON FLIGHTS PRICE PREDICTION

Submitted by:

Sahib Hussain

**ACKNOWLEDGMENT**

This project would not have been possible without the support of many people. Many thanks to our SME Keshav Bansal, who gave suggestion and ideas on creating this project and helped make some sense of the confusion. Also thanks to my Internship Company FlipRobo Technologies Pvt. Ltd for giving me this opportunity to work on such a project which we would be very beneficial for the Tourism and Travel Industry. For the completion of this project we took help of google to get domain knowledge about flights and what factors does people look into mostly while booking flights tickets. Then in order to collect the data we did scrapping of Flights booking site which was Easemytrip as it had all the information we needed to complete the project.

**INTRODUCTION**

* Business Problem Framing

With the covid 19 impact in the market, we have seen lot of changes in the travel industry. Flights were non-operational during the pandemic which resulted in huge losses to many Airlines Companies and has impacted a lot in the price. With the change in market due to covid 19 impact, We would be making a model which can help us determine the price of flights taking various features into consideration.

* Motivation for the Problem Undertaken

This model will be used by the management to understand how exactly various factors affect the price of a flight and how much accurate result they can predict if we give all those factors as input to our model.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Dataset which we worked on had 2054 rows and 9 columns when we collected it using selenium, we needed to clean it for making the model.

By using Distplot from Seaborn Library available in python we found that Duration variable had skewness.

By using Boxplot from seaborn library we found the outliers in each of the features.

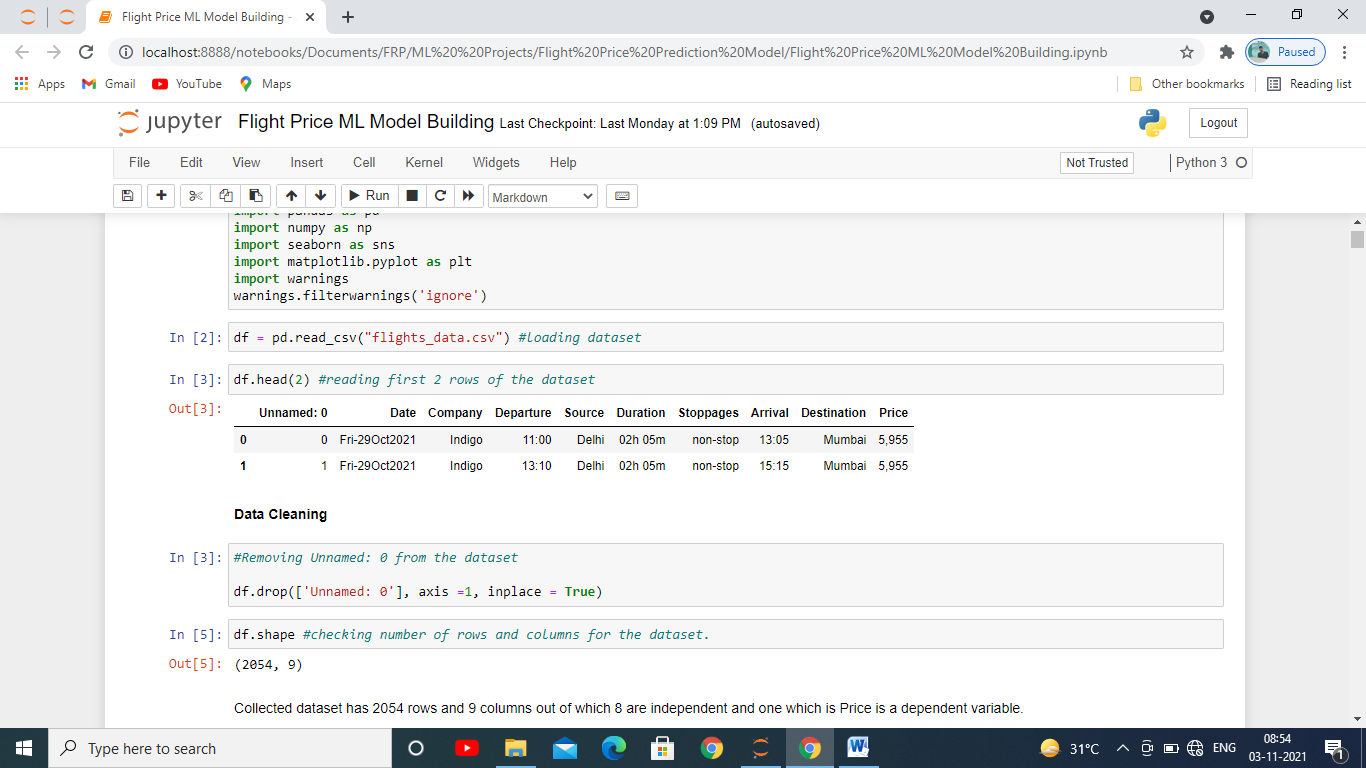
Heatmap available in Seaborn helped us to find the collinearity among various features which we will see later in this documentation.

We used countplot from seaborn to analyse the categorical features.

To make this model we used Linear Regression, Random Forest Regressor, Decision Tree Regressor, Xtreme Gradient Boost Regressor and AdaBoost Regressor.

* Data Sources and their formats

To collect data we used selenium and with the help of it we scrapped Easemytrip website as it had all the important information which we needed to analyize and create this model.



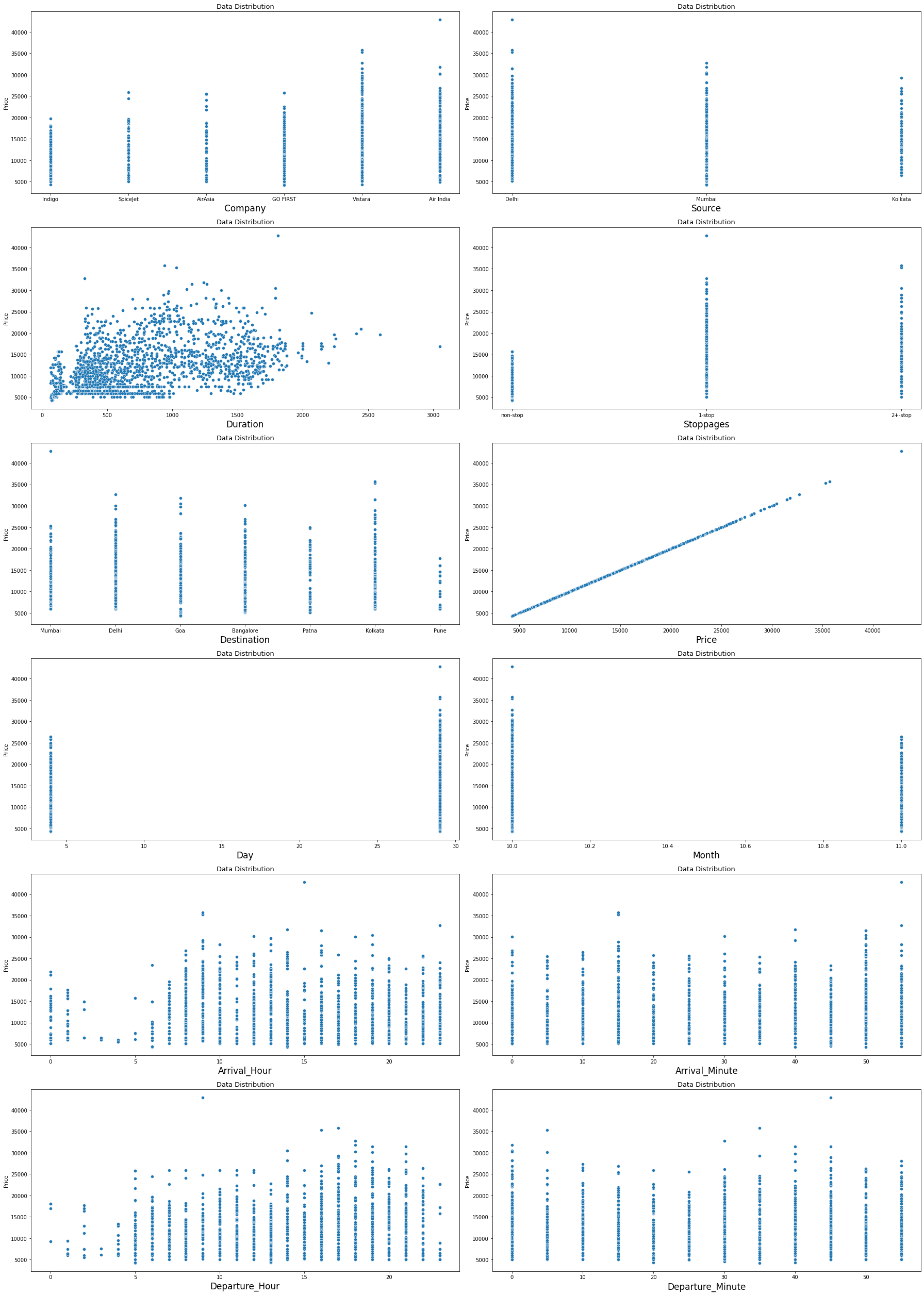
* Data Pre-processing Done

To clean the data we had to delete few features as they would not had effect in getting the predictions and made new features with existing columns to make it better for model building

Skewness were removed with the help of quantile method.

Outliers were removed from the feature Duration with the help of z-score method.

* Data Inputs- Logic- Output Relationships

****

In this figure we can clearly see how variables are related with each of the features and what effects it.

Above figure is showing some positive linear relation between Duration and Price, Stoppages and Pice, Departure\_Hour and Price, and Negative relation between Month and Price

* Hardware and Software Requirements and Tools Used

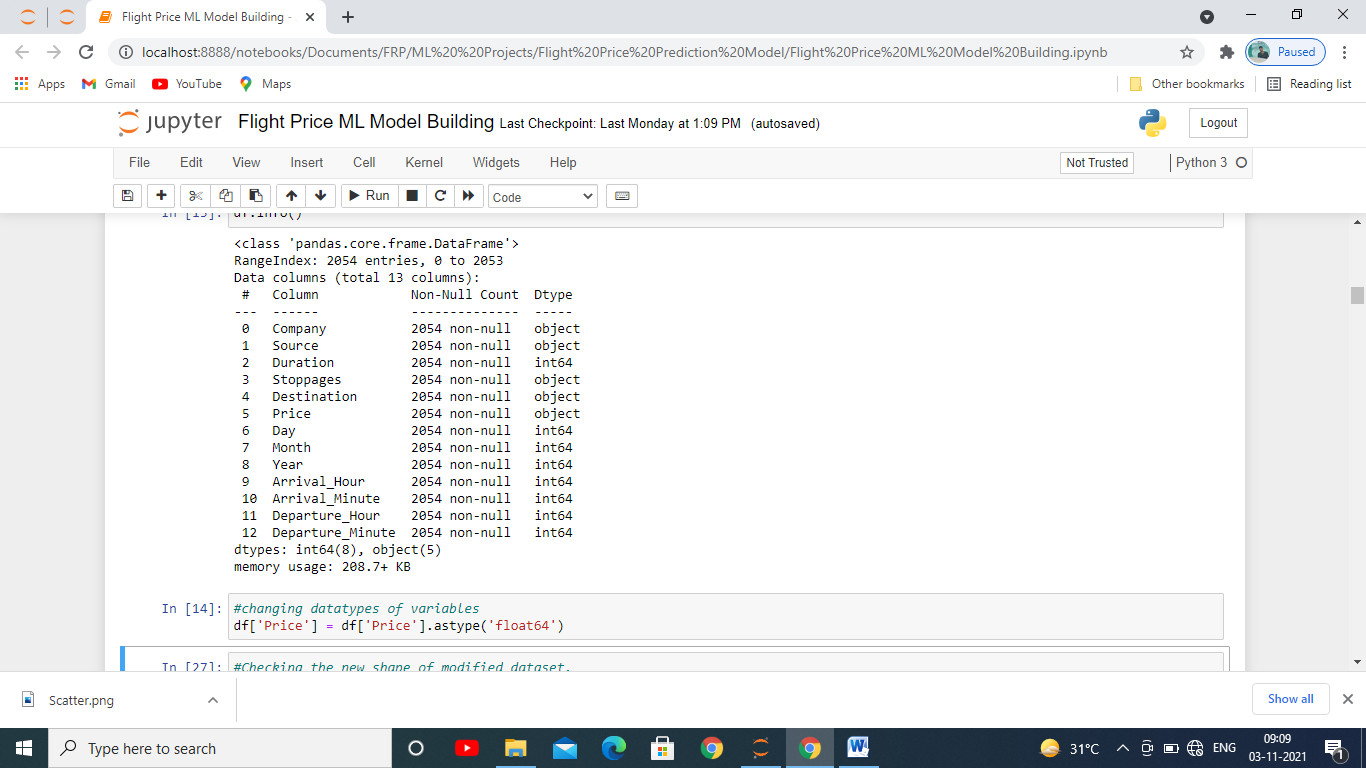
**Software used** : In this we used Jupyter Notebook from the Anaconda to build the model, Miscrosoft powerpoint to make the powerpoint presentation and Miscrosoft word for documentation.

**Libraries used**: Pandas to read the dataset, Matplotlib and seaborn to analyse the data and Scikit Learn to build our models.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

As we can see that dataset had 2054 rows and 9 columns originally, we cleaned the dataset, created new features from the existing one and finally we had 13 columns with 1 as label and 12 as features with various datatypes which we can see in the below picture.



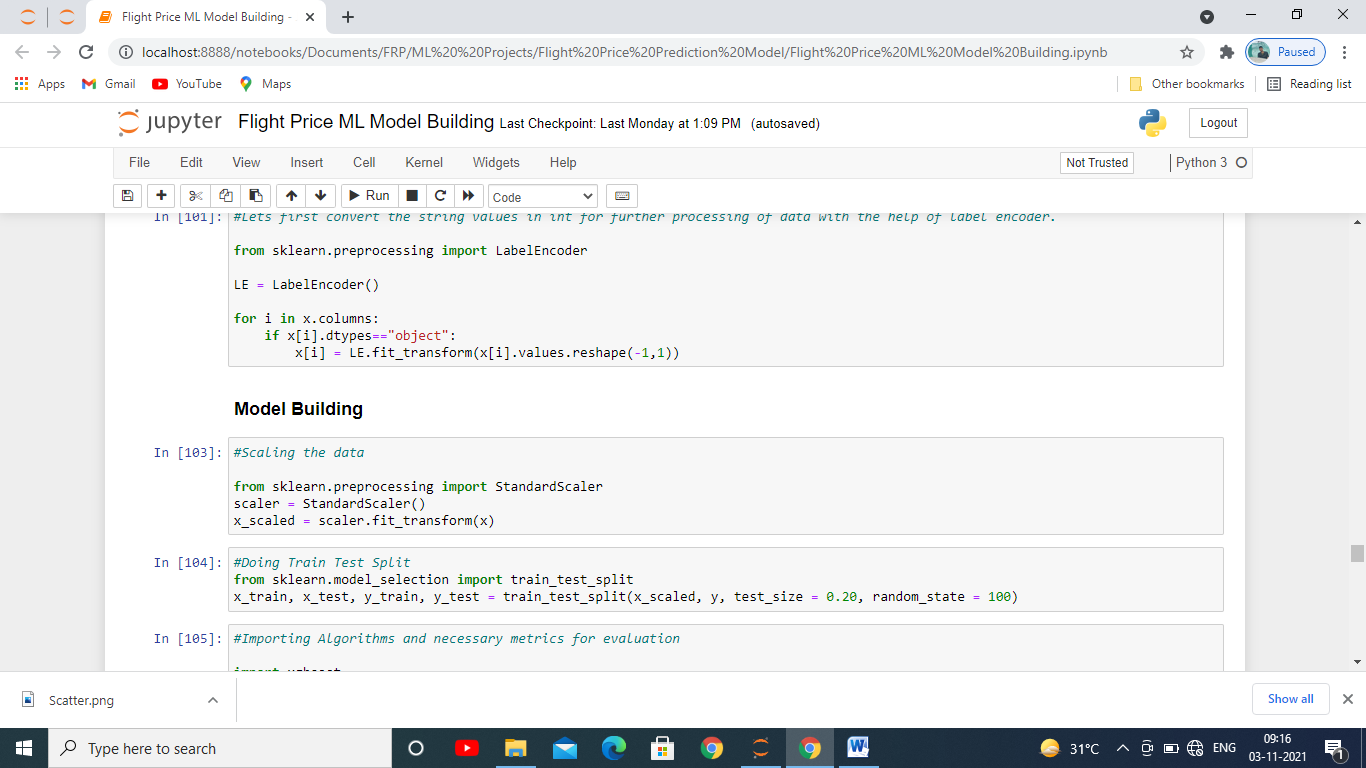
Then we checked for outliers which is abnormal data is not possible or had occurred due to some incorrect entry with the help of box plot to later on it was removed with the help of z-score. For the dataset which we made it was in the feature Duration.

With the help of distplot we checked for skewness, whether the data is skewed towards left or right or whether it was Normal distribution if it had normal distribution /or forms bell curve , it would have been considered good for building the model but since the data was skewed in the feature Duration mostly towards right these had to be cleaned with the help of quantile method so that we the model doesn’t always predict the output towards which the data is skewedmostly.

We plotted heatmap to see if there is collinearity among the features so that repetition of the data in predicting model can be avoided. Since there was very less collinearity we didn’t had to remove any features for model building.

We scaled the data for all the independent variable which is very important in building the model

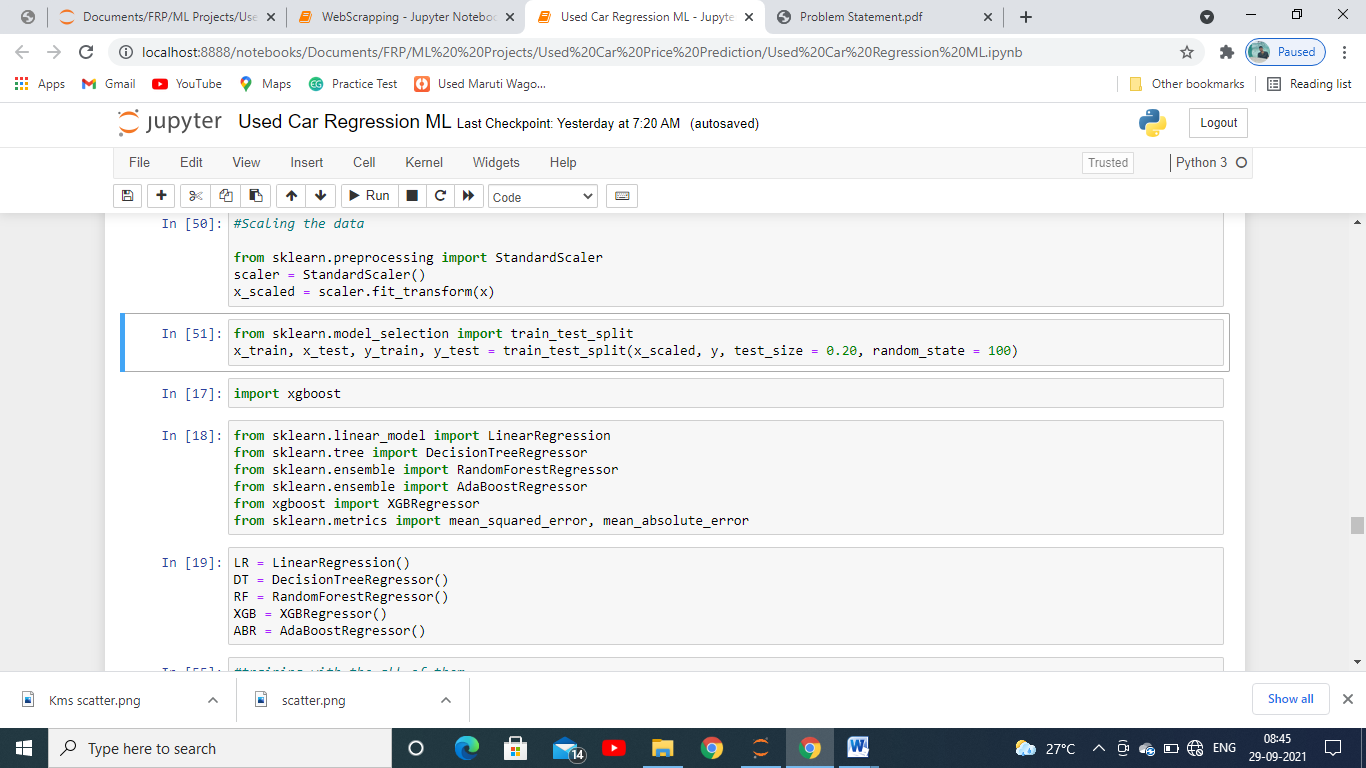
* Testing of Identified Approaches (Algorithms)



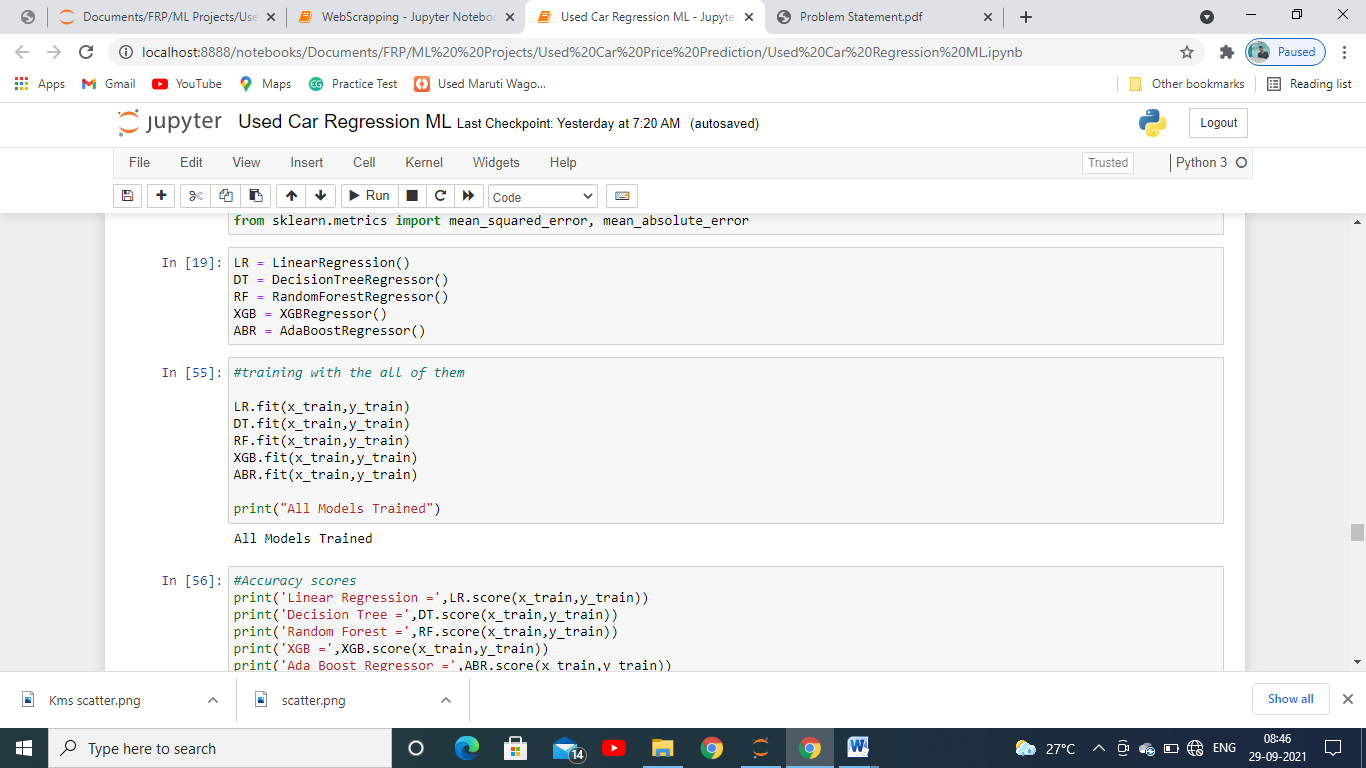
We imported the train test split as you can see in the above image.

The whole dataset after being scaled is broken down into 4 parts each for train and test, 2 parts for traning the model and 2 parts for testing, in it we kept 80% of the data for training and remaining 20% for testing the data.

* Run and Evaluate selected models

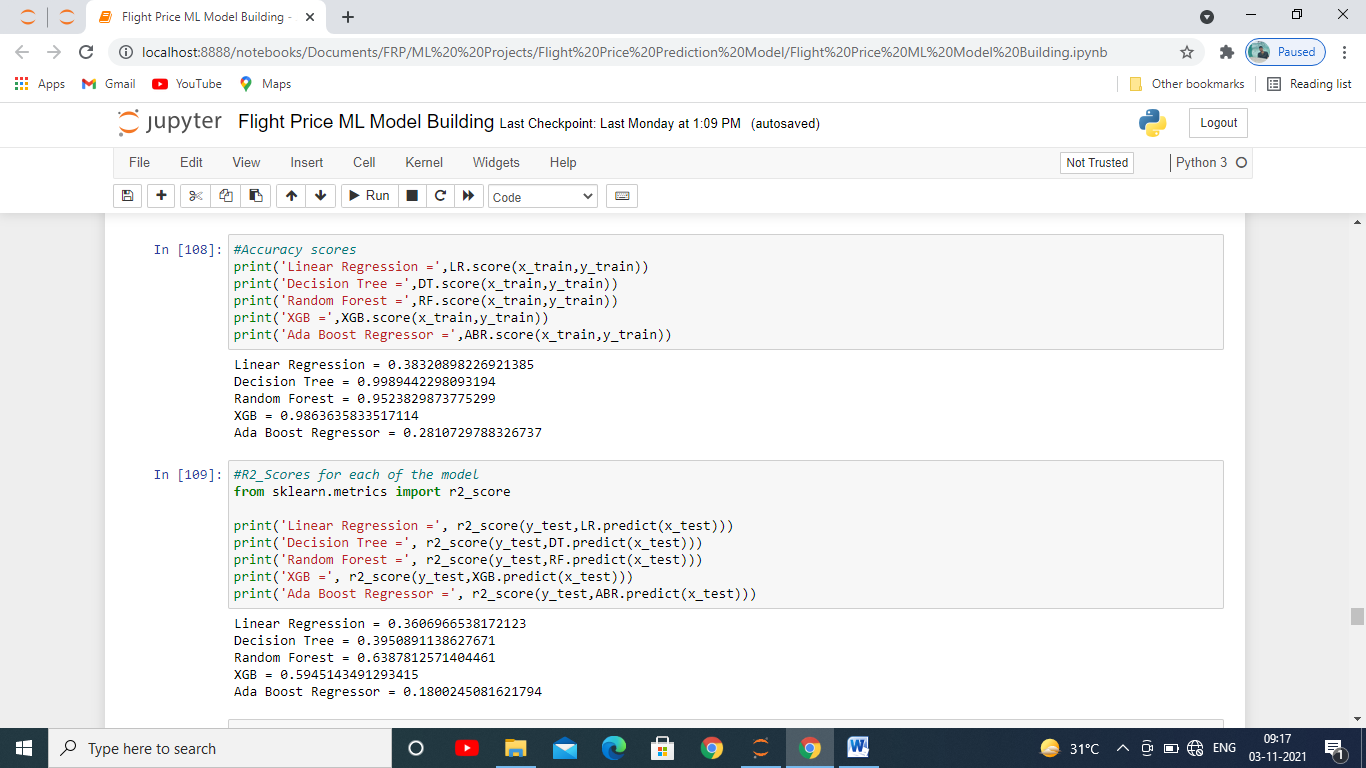


First we imported various evaluation metrics used for evaluating a regression model since the label had to be predicted is in continuous form then imported various models we are used for building classification model as you can see we imported 5 models: Linear Regression, Ada Boost Regressor, Decision Tree Regressor, RandomForest Regressor, XGBoost Regressor then we stored them in variables for each one of them which we are going to use further in building the model.



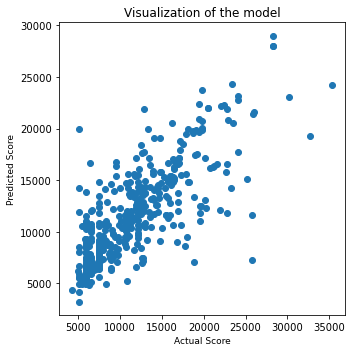
We then trained it with x\_train and y\_train which we got from the dataset after splitting it in train and test set. This train set composes of 80% of the data which is going to be used to study the data by the model to predict the test set.

Then we captured the accuracy and r2 score which helped us to determine performance each model is giving of each with our test set divided in 2: x\_test, y\_test.



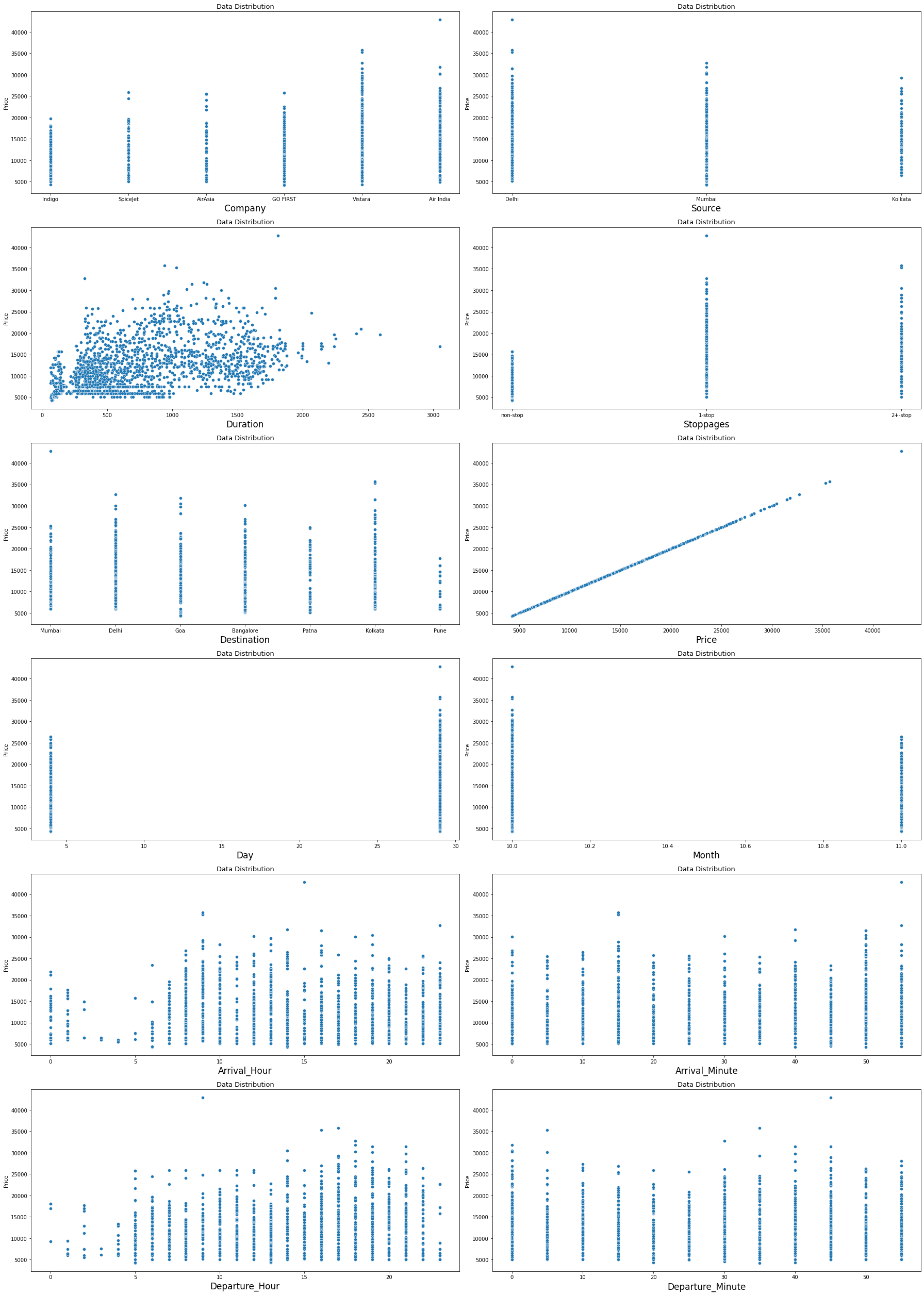
We found out score of each model which is the accuracy each model is giving for this dataset and we can see that Random Forest, Decision Tree, XGB Regressor is giving us the highest prediction with nearly about 95%, 99%, 98% approx respectively. The least was given by AdaBoost Regressor which is of 18% approx.

* Key Metrics for success in solving problem under consideration

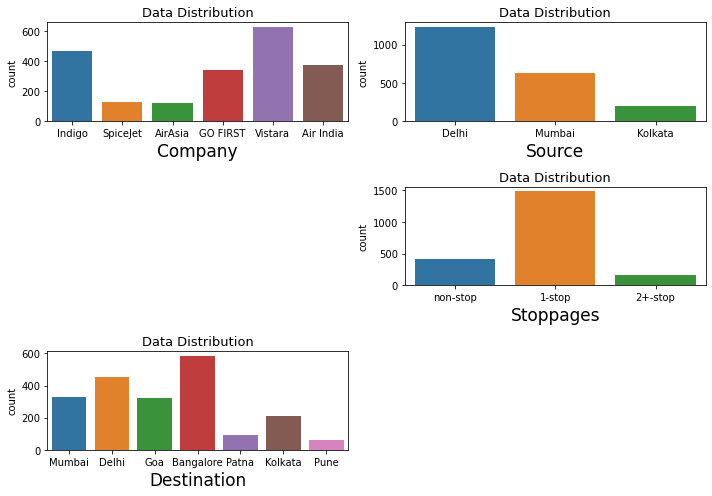


The above image is showing us the relation between our predicted values and actual values as we can see that there is not much gap for between predicted and actual datapoints for XGB Regressor which shows that the it is best fit for our dataset. Also we saw the scores of each one of them and XGB Regressor had delivered us the best scores among them when we checked all the metrics used in evaluating a model

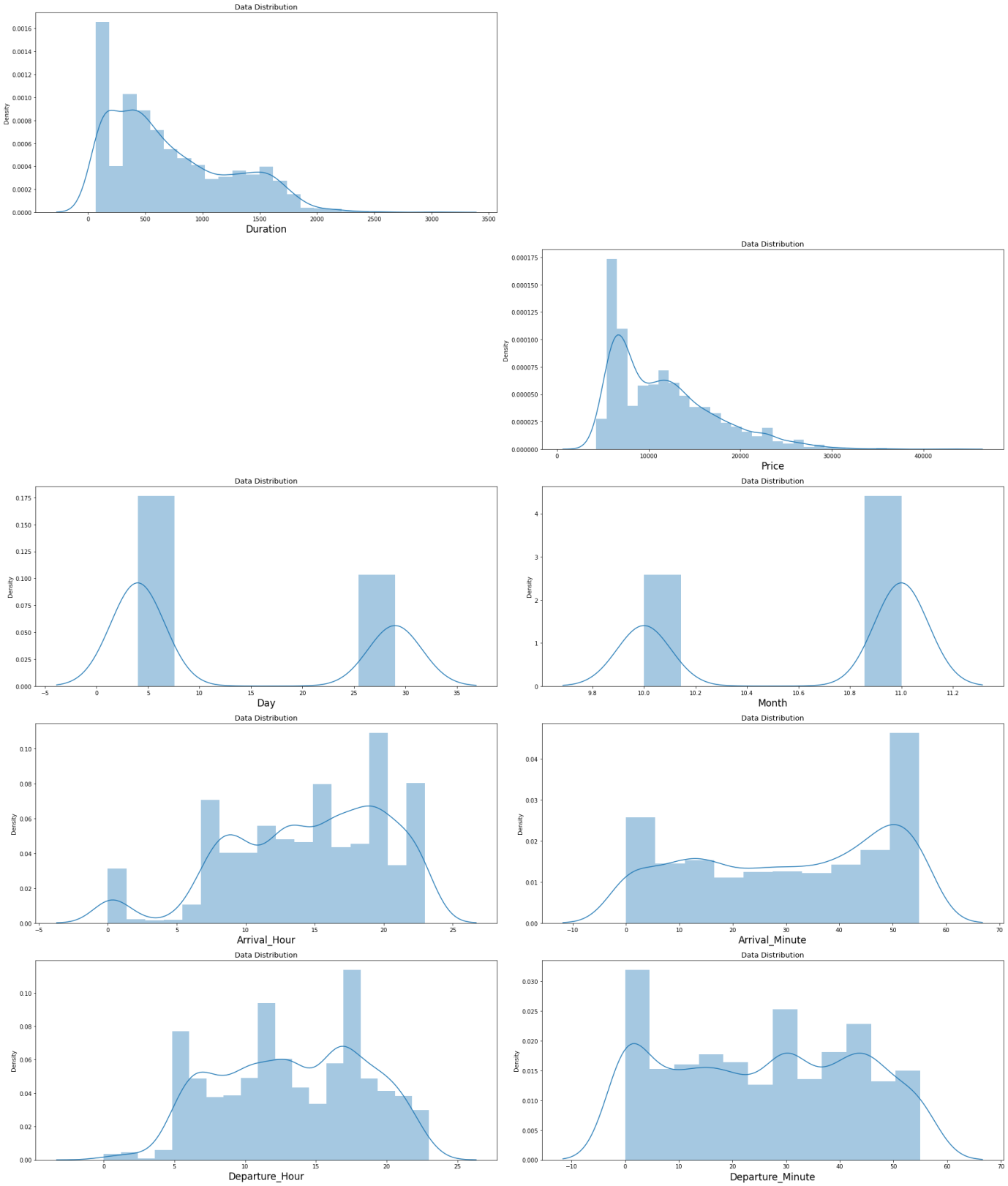
* Visualizations



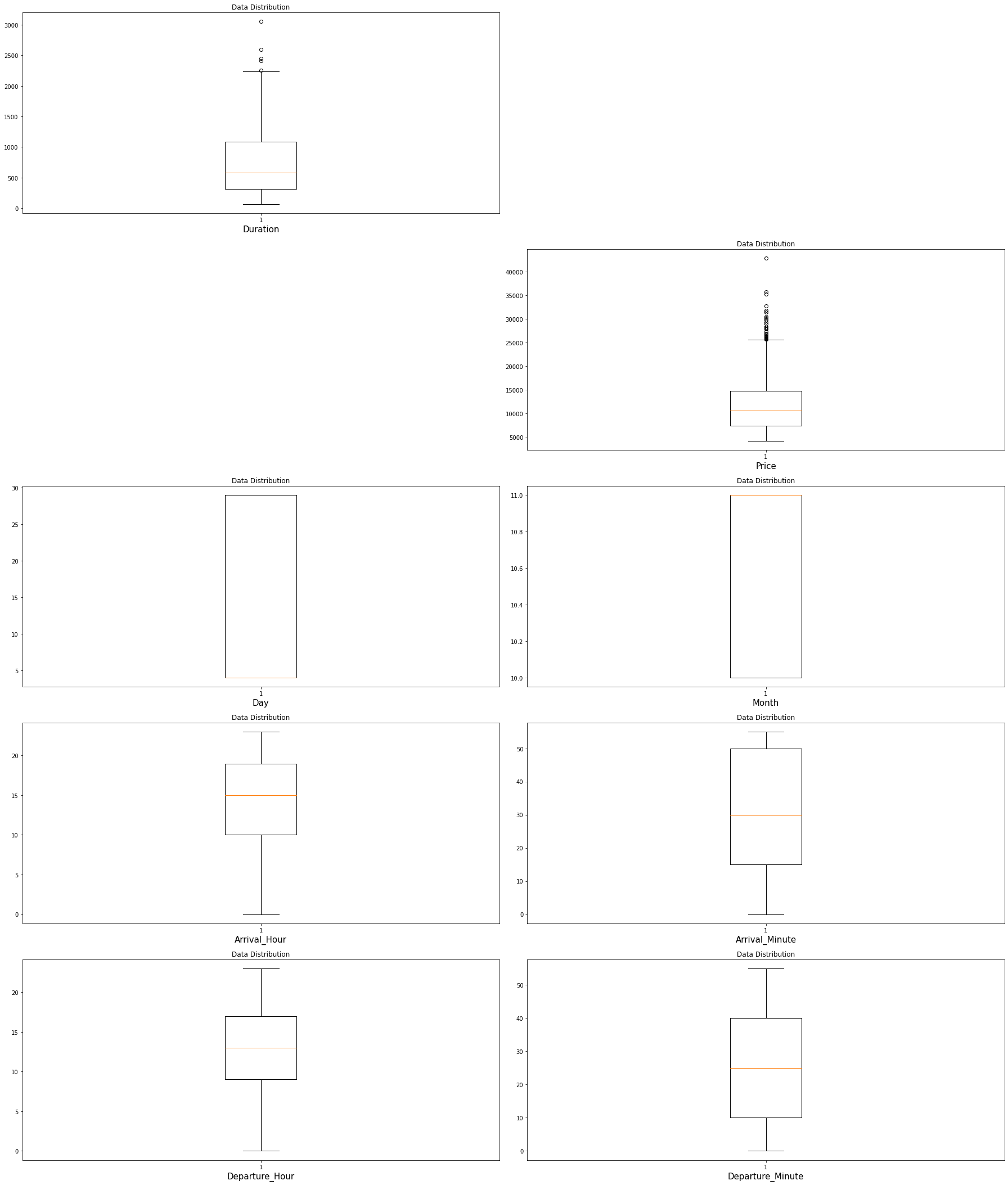
With the above image which is a scatter plot we checked how each features are affecting the label and how they are related as we saw in the beginning of this documentation. Above figure is showing some positive linear relation between Duration and Price, Stoppages and Pice, Departure\_Hour and Price, and Negative relation between Month and Price.



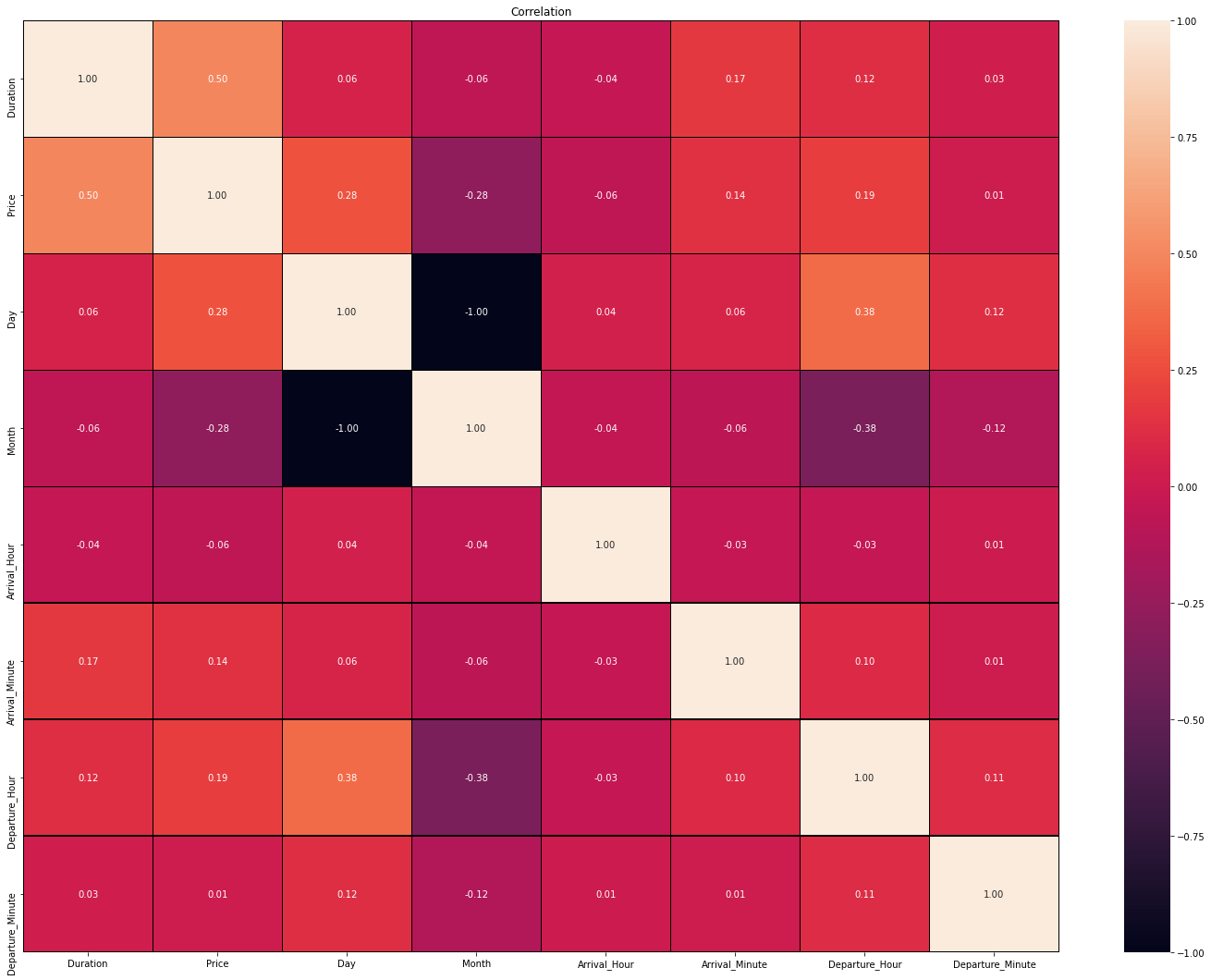
With countplot we saw the categorical columns and which category of each feature is mostly affecting the Price.



The above figure shows us the skewness for the continuous variables and we can see that only Duration was skewed towards right, a perfect curve would be the one which forms bell curve.



In this figure we can see the outliers in each continuous variable which is not normal or not possible in real life and it occurs mostly because of incorrect entry these had to be removed for creating a better model, it is present in Duration variable of our dataset.



It is a heatmap which shows us the colinearity amoung features how much they are co-related near to 1 has high colinearity among them. We don’t see any multi-collinearity among the features in our dataset.

* Interpretation of the Results

After building the model we found that XGB Regressor gave us the best result taking into consideration all the metrics scores and parameters, this model had least difference between the predicted data points and actual data points.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project we found how various factors can hugely impact the price of Flights. A little bit of change in the Timings of booking, departure hour and destination can drastically bring change to it’s price. How important the features are in determining the price of flights.

* Learning Outcomes of the Study in respect of Data Science

While doing this project we observed that data visualization plays a key role in analysis the data and further processing it to build a model which has very less error and more accuracy. Analysis helps us to find out which are the feature mostly affecting the label and has more weightage and on this basis the data needs to be cleaned so that the data which we get to build our model gives us the best results. Shape is used to find the number of rows and columns, info for getting what are the data types of features and whether there are any null values in it or not, if there are this needs to be either filled up or removed from the dataset so that it does not build an erroneous model for us, describe has also been used to check the quartiles and mean, standard deviation, min value and maximum value of each of the features.

* Limitations of this work and Scope for Future Work

As we saw while doing Exploratory Data Analysis there are various models which are common and some of the models are not present in this though we didn’t considered that in building our model however if we could have gotten the information of flights as well along with some more features the model could have been much more better and efficient in predicting better results in predicting the price of flights.