Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

**Problem Statement: -**

***Flight Price Prediction***



**Submitted By: -** *Ojasav Sahu*

**Batch Number: -** *1836*

Nowadays, Travelling has become a very important part of each and everyone’s Life, Almost all the people in the entire world are travelling from one place to another, either due to some important office work, or for a trip with their family/friends or may be at times solo. But most of the people around the word are travelling. So, the major problem that comes in each and every one’s mind is the Price of the Fare, Price of the Flight. As, we all know that the prices of the flight are so unpredictable.

But, now I must say our this problem is solved to a great level, as with the help of Machine Learning(ML), we can build a model and with the help of our model, we can predict the future prices of the Flights based on the past data.

* <https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects>

Dataset link:-

Here, in the link we have 2 dataset present; one is ‘Training Dataset’ and second is ‘Testing Dataset’. So, in simple word’s if I say then, I’ll be making our model learn in training data and then I’ll be predicting the data in our test data.

**Lists of Available Features: -**

* **Airline**: The name of the airline.
* **Date\_of\_Journey**: The date of the journey
* **Source**: The source from which the service begins.
* **Destination**: The destination where the service ends.
* **Route**: The route taken by the flight to reach the destination.
* **Dep\_Time**: The time when the journey starts from the source.
* **Arrival\_Time**: Time of arrival at the destination.
* **Duration**: Total duration of the flight.
* **Total\_Stops**: Total stops between the source and destination.
* **Additional\_Info**: Additional information about the flight
* **Price**: The price of the ticket

**Index of the Article: -**

**Training Data**

1. Importing the Libraries
2. Data Collecting(fetching the data in our jupyter notebook)
3. Data Handling
4. Exploratory Data Analysis
5. Visualization
6. Outlier Detection
7. Treating the outliers
8. Skewness

**Test Data**

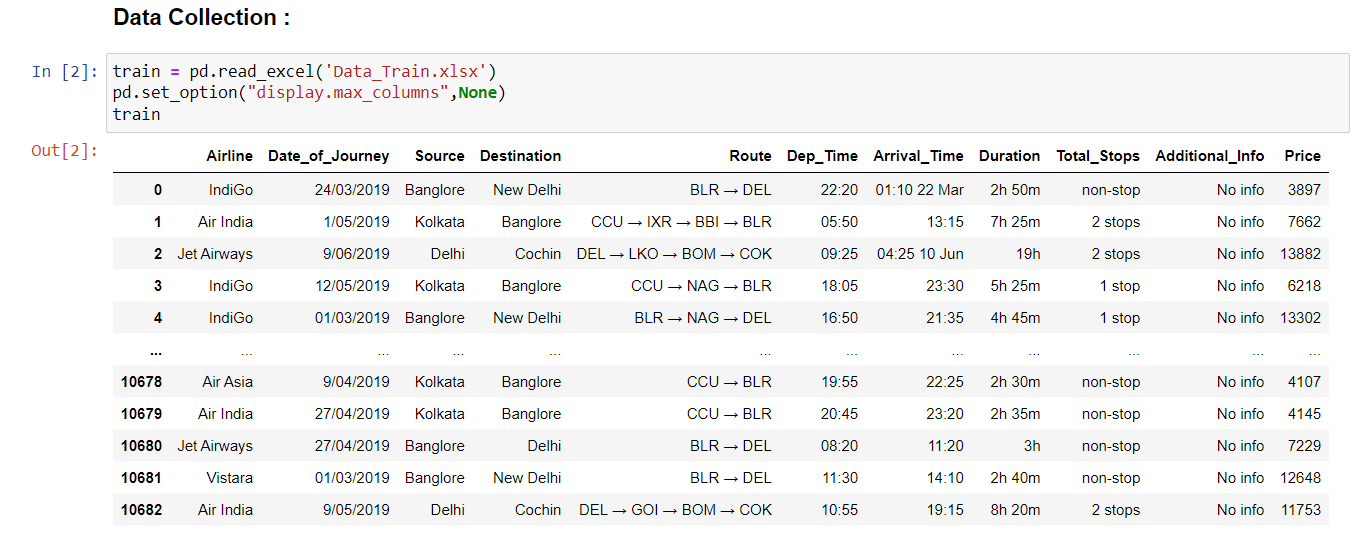
1. Data Collecting(fetching the data in our jupyter notebook)
2. Pre-Processing of the data
3. Data Handling
4. Checking the Outliers.
5. Model building

* Importing the Lib. Needed for the model building
* Building the model
* Cross validation
* Regularization (Hyper Parameter Tuning)
* Saving the Model
* Graph for Predicted and Actual Value

1. Conclusion
2. Concluding Remarks



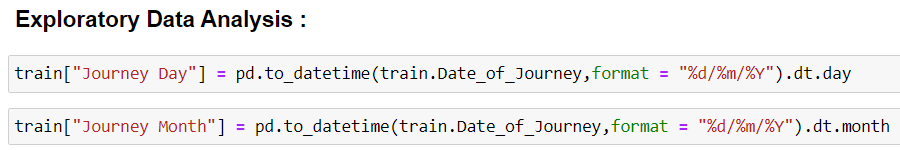
**Training Data: -**



Here, at the very first I have Imported the Libraries and then, Importing our training dataset.

After, that I have checked for the null values, as there were only 2 null values present, So I have simply dropped them.



i. The column, "Date\_of\_Journey" was "object" type data so I have converted this into numerical data type so that I can use it for the model prediction.

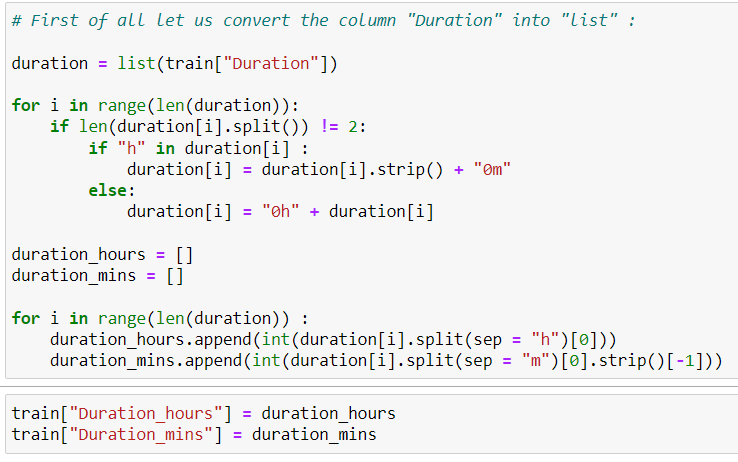
ii. For, the conversion, I need “pandas.to\_datetime” to convert this object into “datetime” data type.

iii. Here I'll use ".dt.day" method for extracting day of that date and also I'll use ".dt.month" method for extracting month of that date.

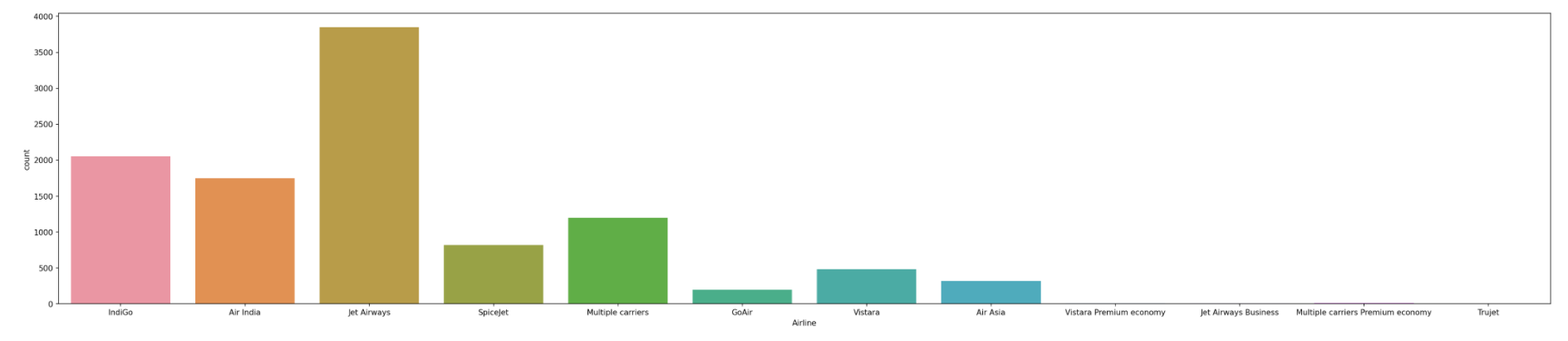
iv. After that, I have made a new column for data and month and dropped the column, “Date\_of\_Journey”.

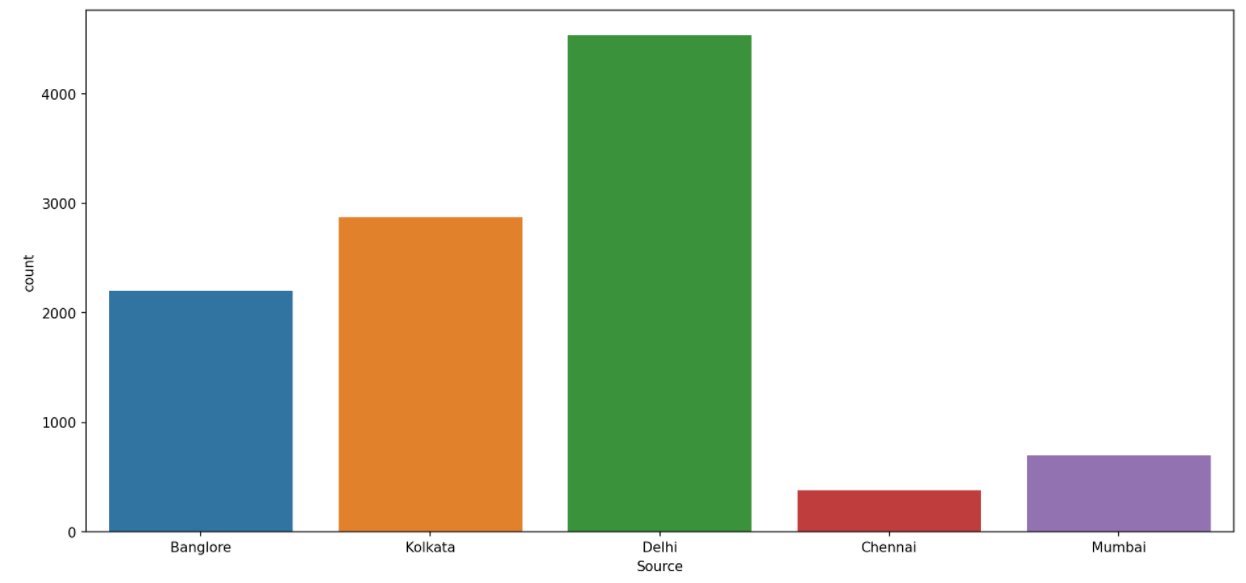
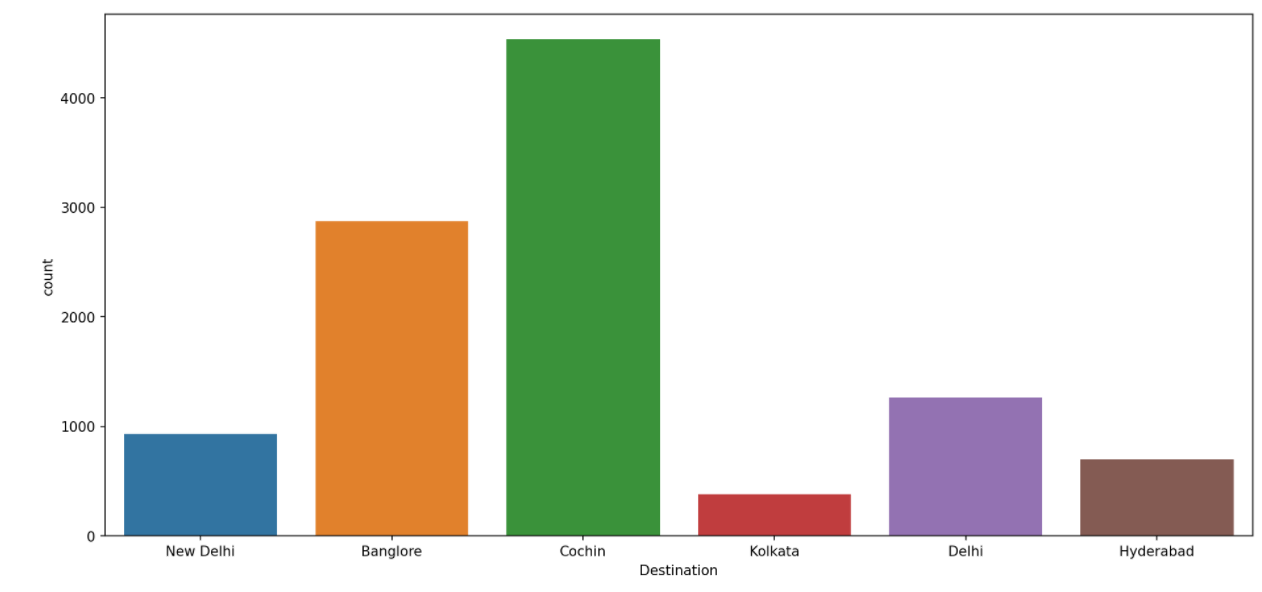
v. I have done the same thing in two other columns: - “Dep\_Time”,” Arrival\_time”

vi. For the “Duration” column, I have first converted the data of duration in a list, and then I have separated Hour’s and minutes.

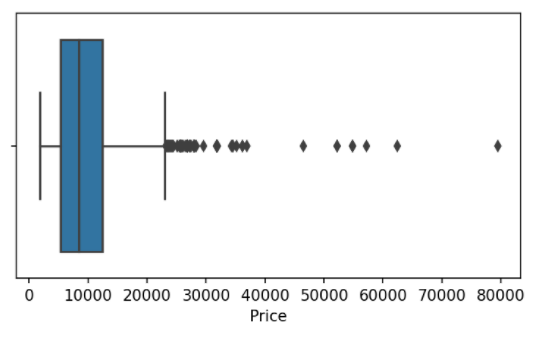
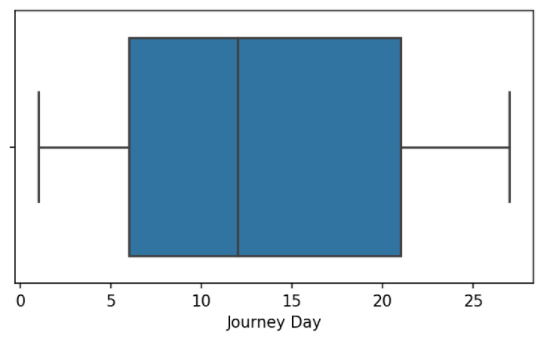




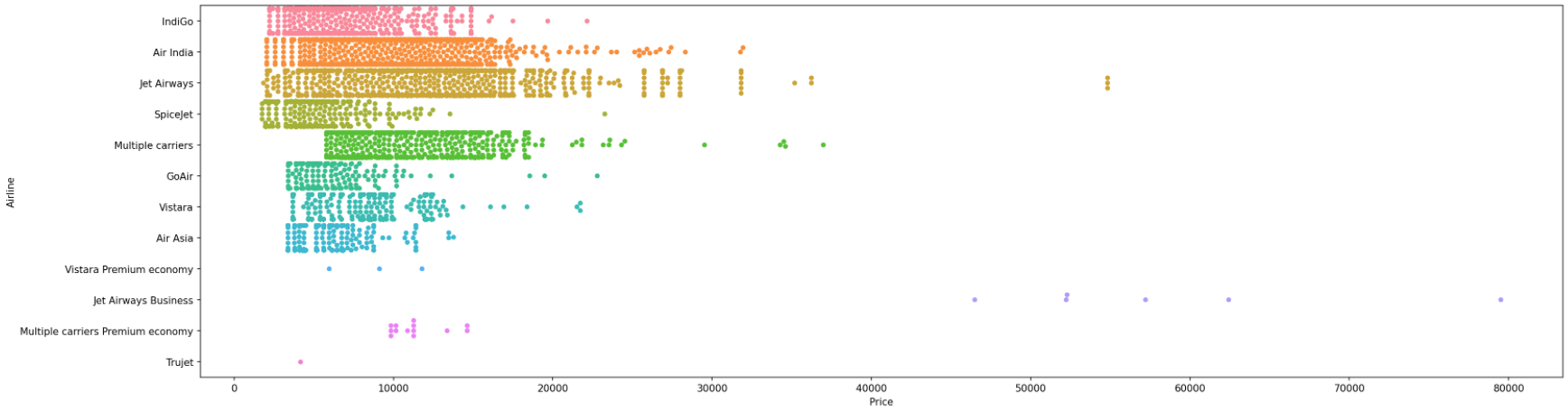


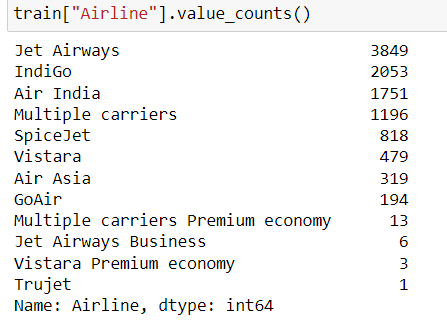
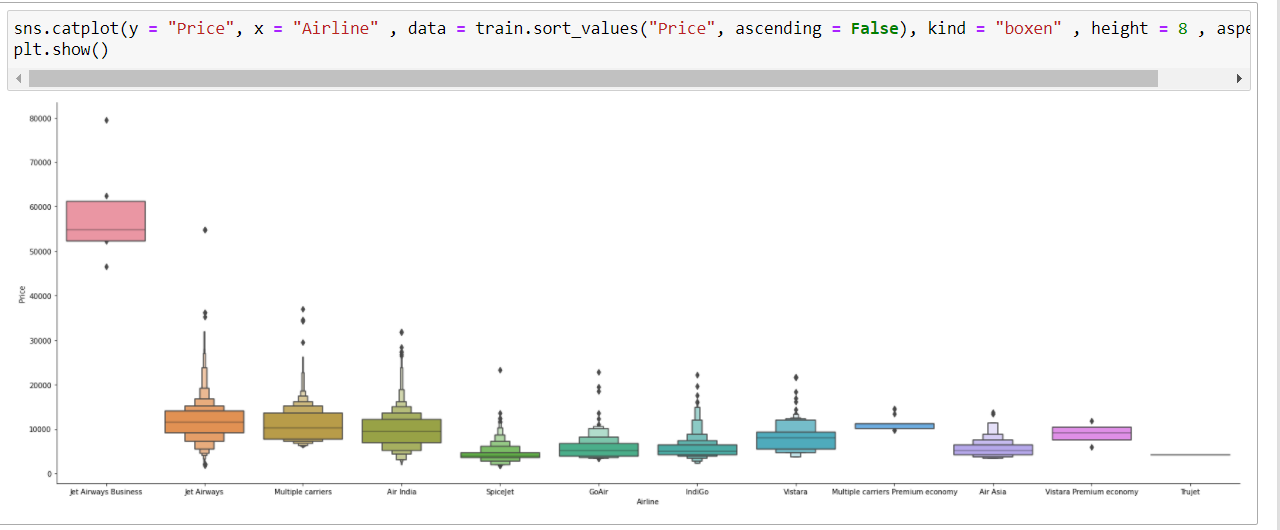
Here, I can see that a lot of people prefer to travel through, “Jet Airways”, and the source is “Delhi” i.e most of the people took flight from delhi to “Cochin”.

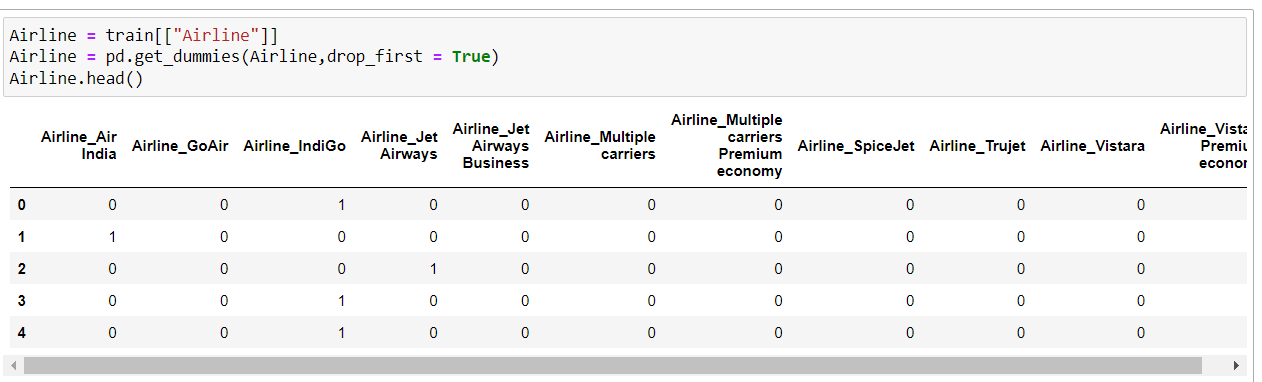
Here, I can see that a lot of outlier’s are present in the “price” column and in “Journey Day” there are no outlier’s present, The same thing I have done for the other columns also.



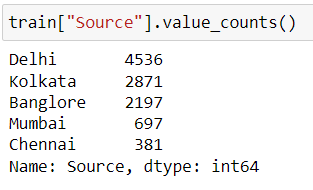
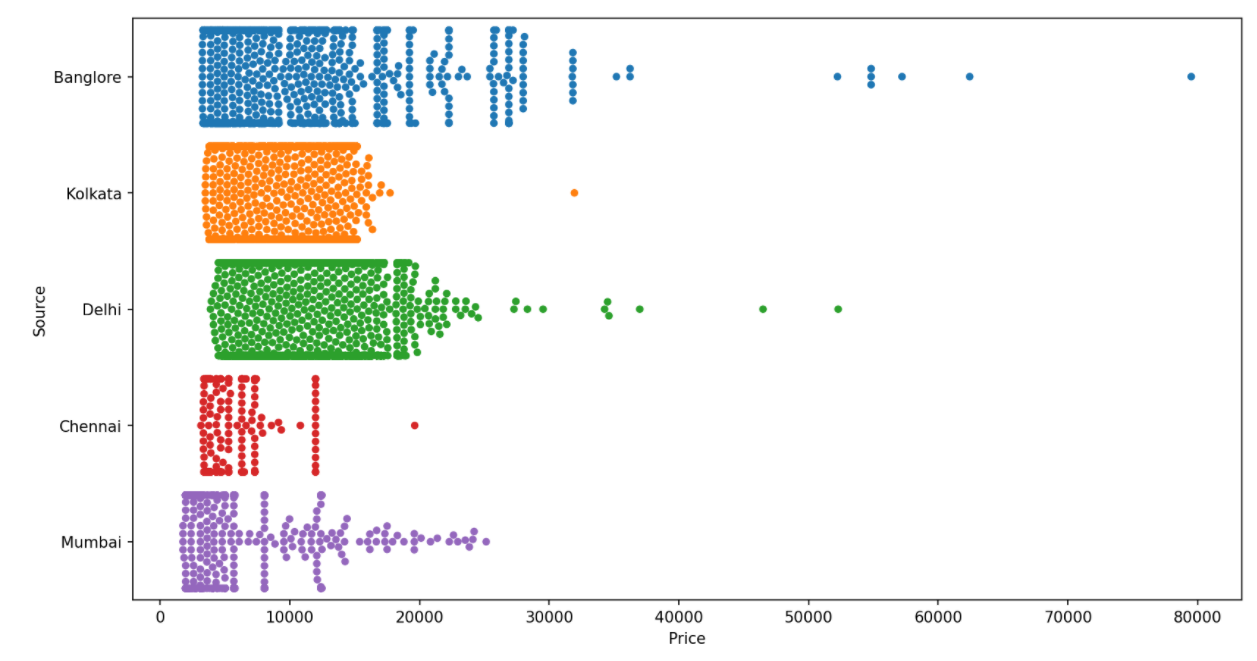
Here, I can see that “Jet Airways” has the highest “Price” among all other flight’s. I can also see that the minimum lowest price is of “Spicejet” and the minimum highest price is of “Multiple Carriers”.

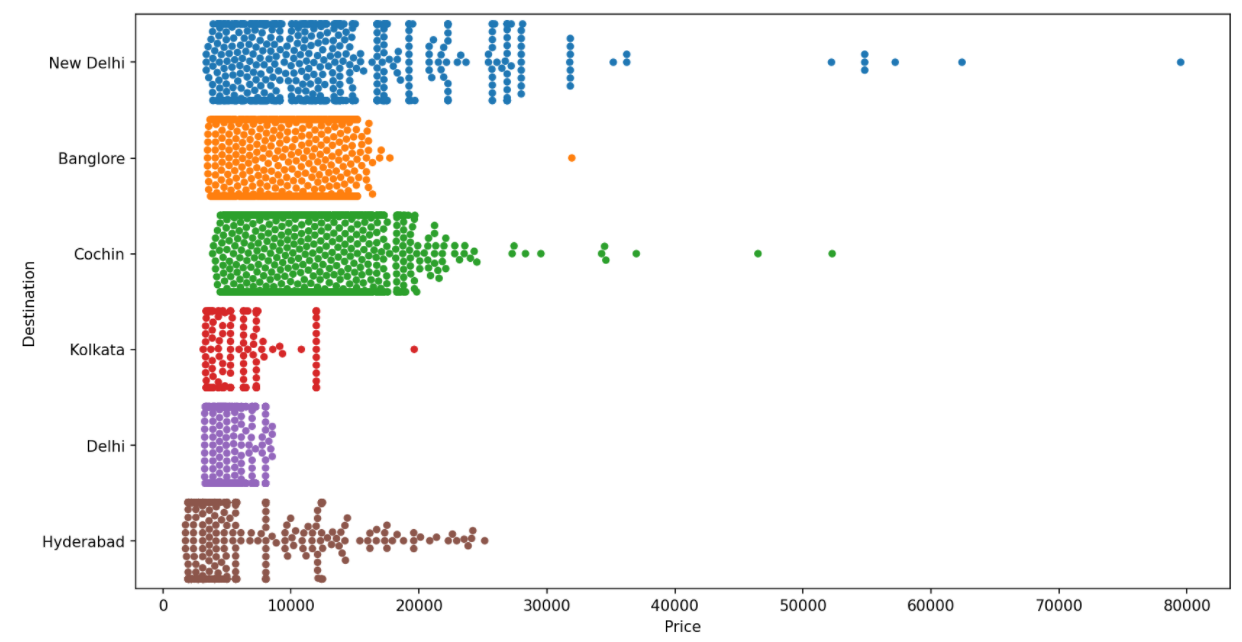
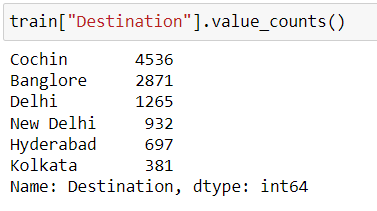
Here, in the 1st pic I can see the value count for each flight, and from the 2nd pic I come to know “price” variation between the flights, I can see, most costly flight is, “Jet Airways Buisness” and after that somewhat or the other there is not much difference in other flights.



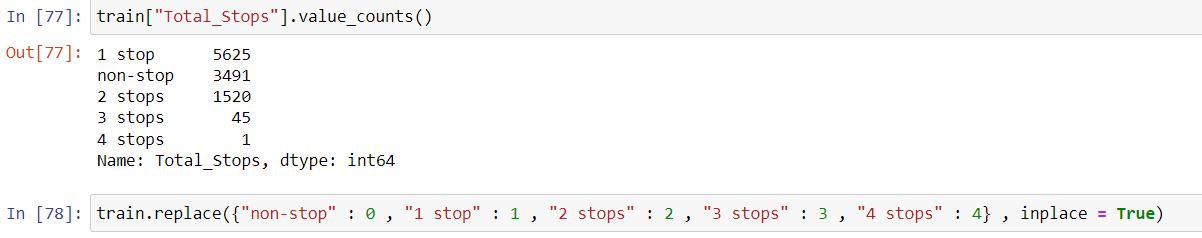
Here, I have used, “One Hot Encoding”, for the “Airline” column and created dummies for our easy processing of the model.

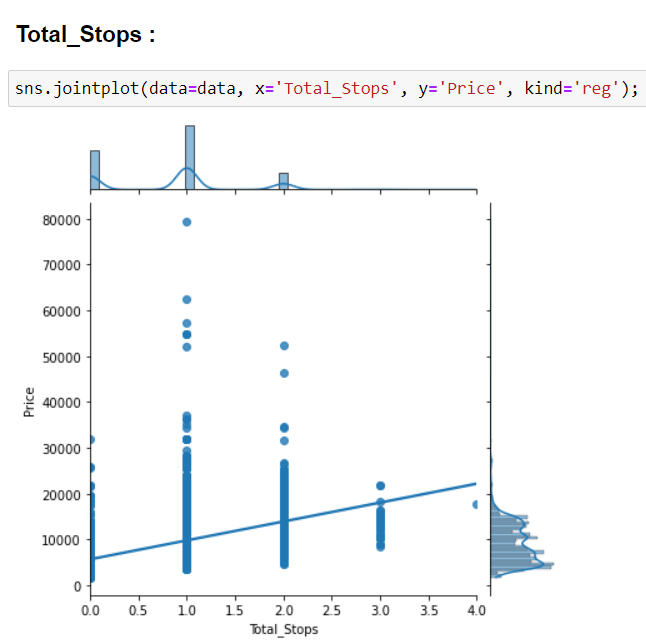
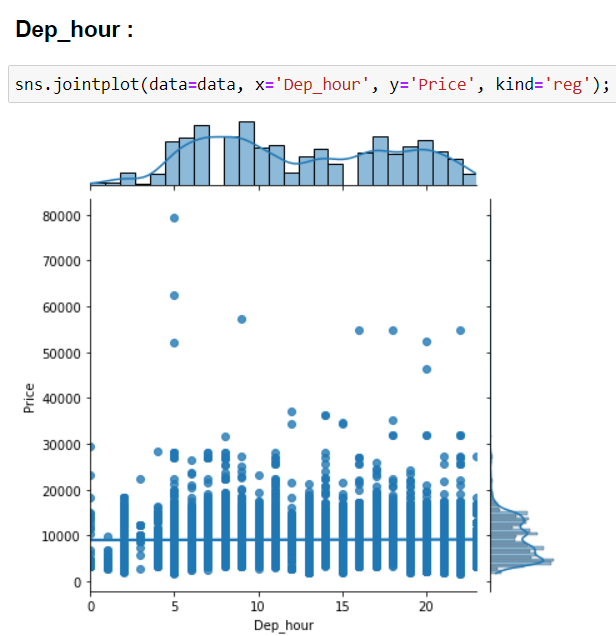
Here, in the 1st pic I can see the value count for each source, and from the 2nd pic I come to know “price” variation depending on the ”Source” location. I can see, most costly is for, “Banglore”.

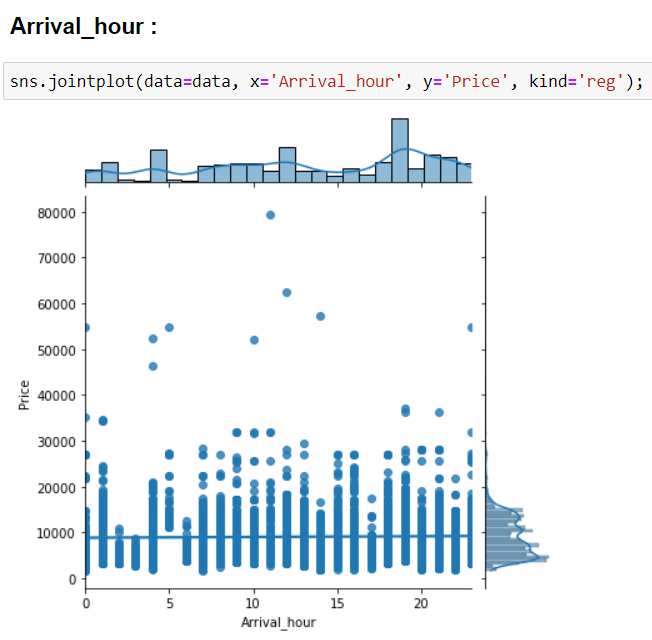
Here, I can see the “price” variation depending on the ”Destination” location. I can see, that most costly is for, “New Delhi” location and In the second pic I can see the value count for the “Destination” column.

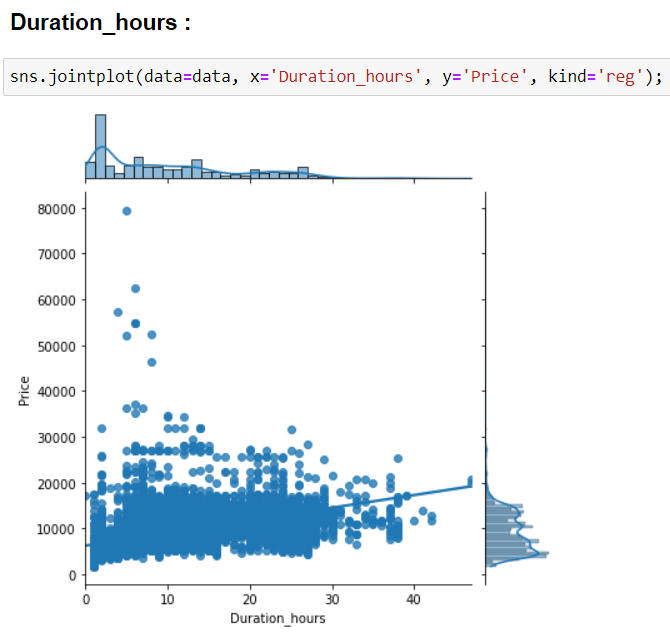


Here, First I have find out the value count of the Column, ”Total\_Stops” and then I have replaced the data, “non-stop” with “0”, “1 stop” with “1”, “2 stops” with “2”, “3 stops” with “3” and “4 stops ” with “4.”

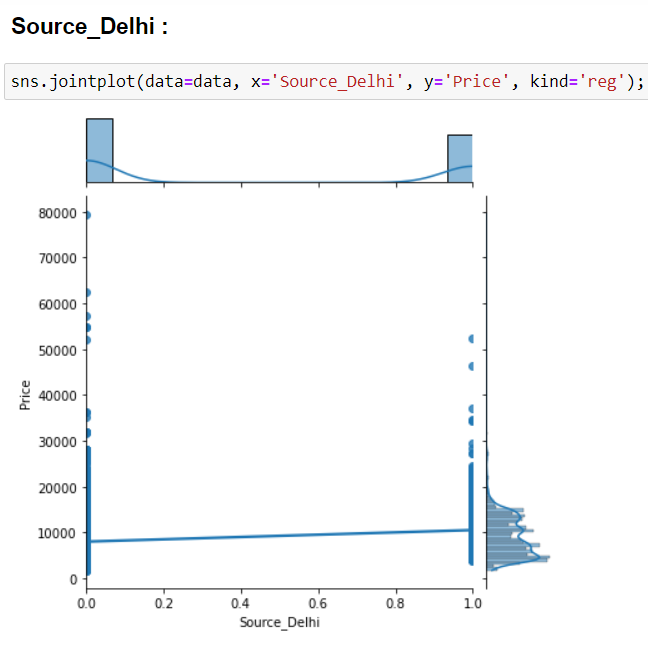
 

Here, In the First Pic I can see that with the increasing number of stops, Price is also increasing of the flight and in the second pic I can see that as the departure hour is going on increasing there is more, less or no change in the price. So, probably at the 5TH Departure hour the highest price get’s reached.

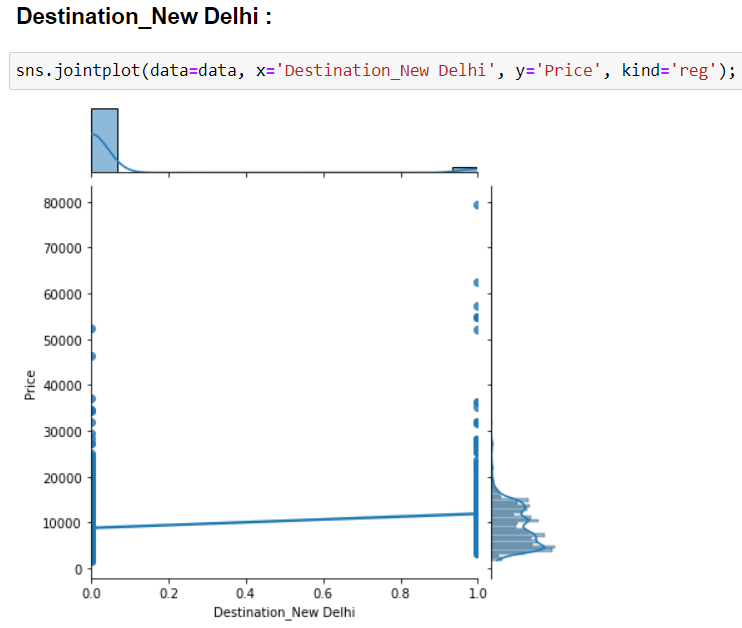


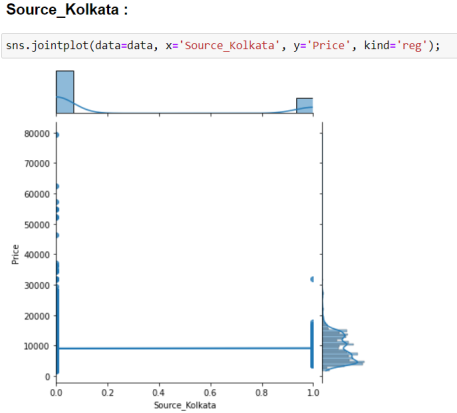
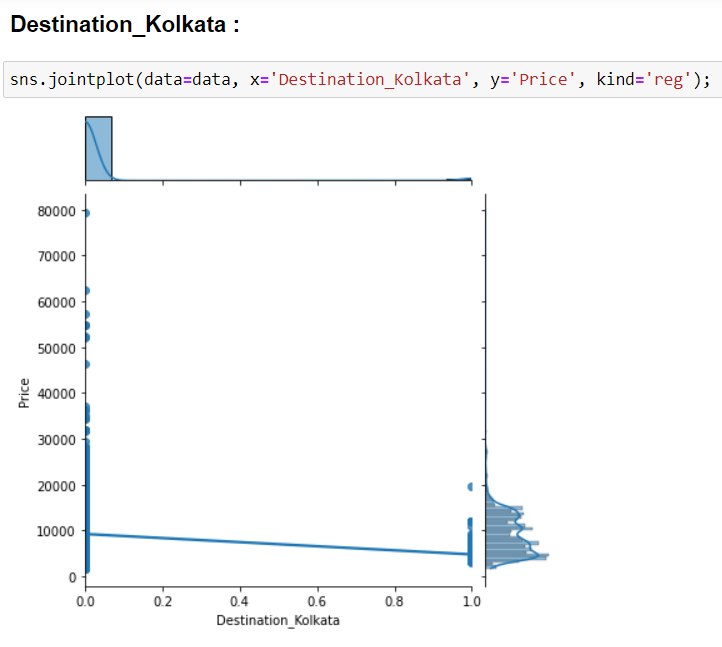


Here, In the 1st Pic I can see that as the arrival hour is increasing then there is more, less or no change in the price but the maximum price reached in the arrival hour reached at the range between 10-15 and in the 2nd pic I can see that as there is increase in the Duration hours the high distribution of price is within the range between 0-30hours and as reached to maximum hours the price is low.

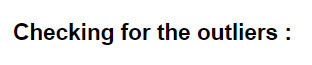


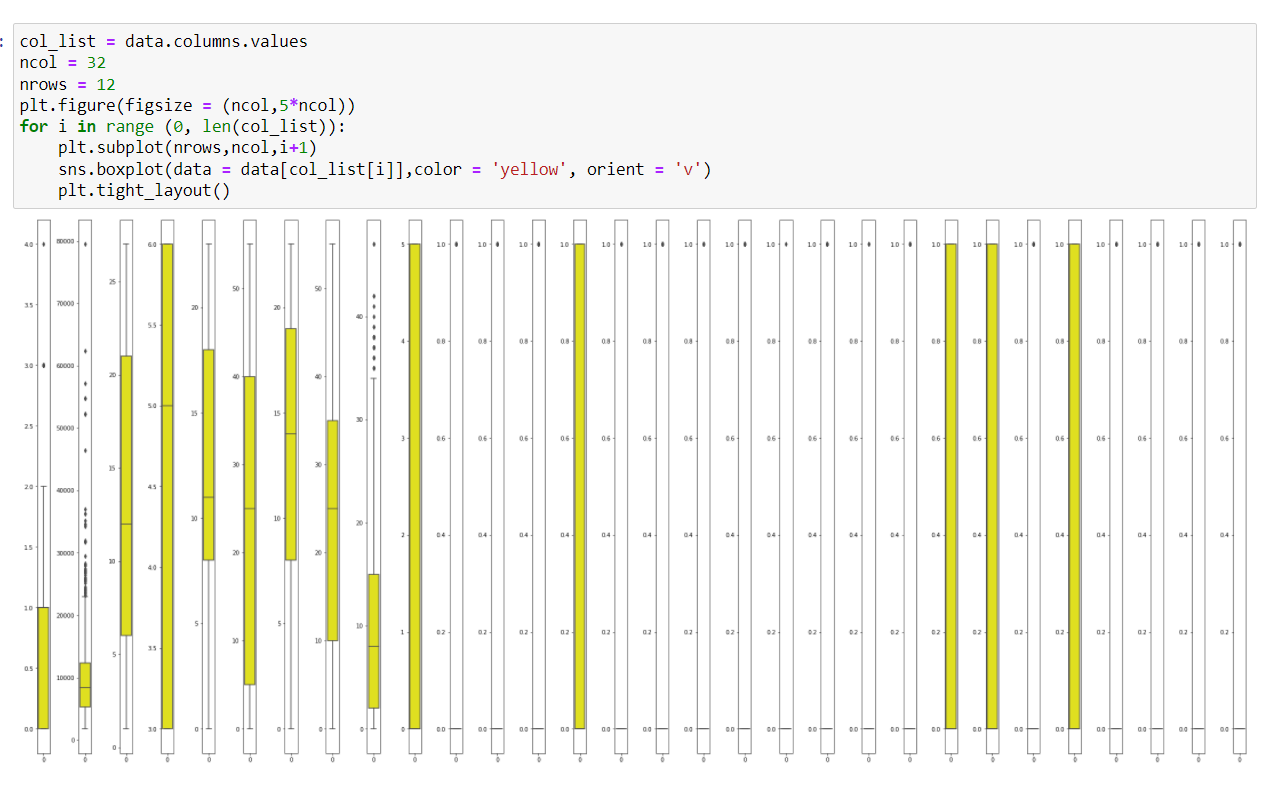
Here, in the 1st pic I can see that as the value of the attribute are increasing there is more, less or no change in the price but the distribution is high for the category 0.0 and followed by 1.0 and in the 2nd pic I can see that as the value of the attributes is increasing there is increase in the price and the distribution is also high for the value 1.0 of the attribute and followed by 0.0



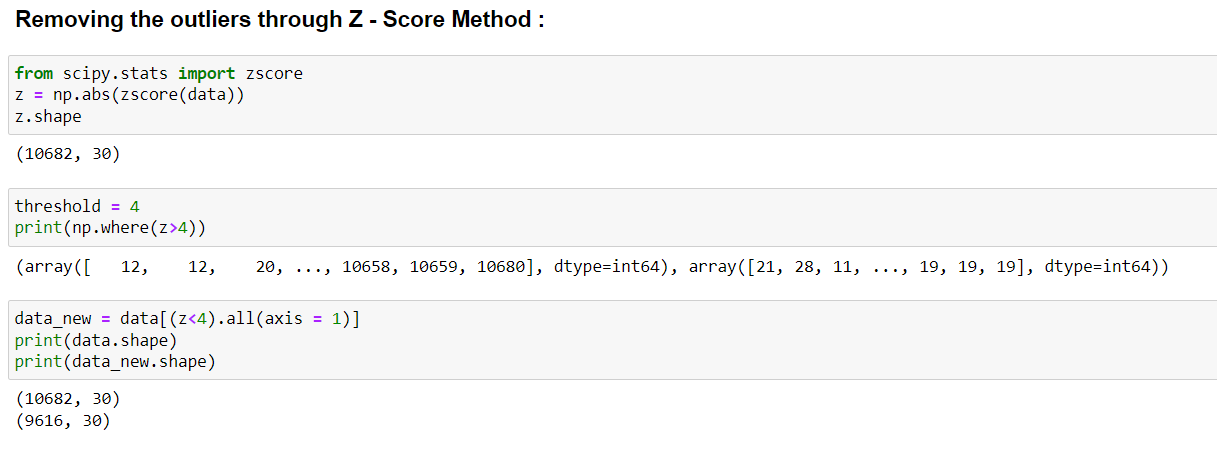
 

Here, In the 1st pic, I can see that as the value of the attribute are increasing there is more, less or no change in the price but the distribution is high for the category 0.0 and then followed by 1.0 and in the 2nd pic, I can see that the highest distribution is for the category 0.0 of the attribute and as the values of the attributes are increasing there is slight decrease in the price and least distribution is for the maximum value of the attribute.

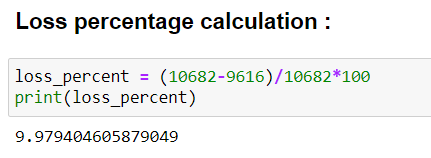




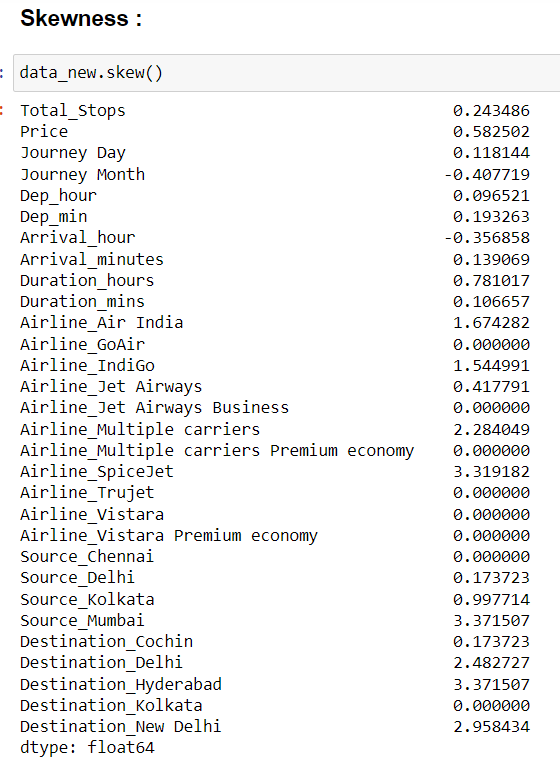
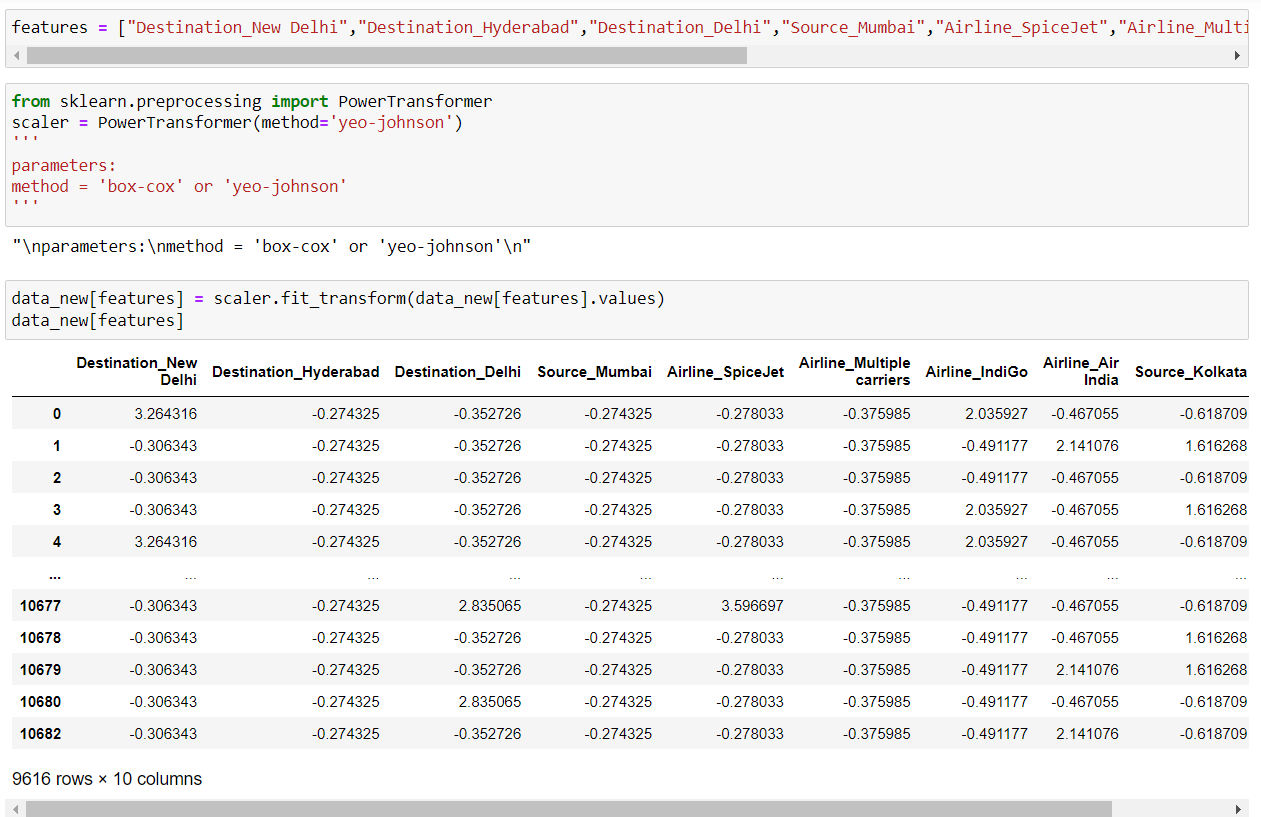
Here, I can see a few columns, have some outlier’s present, but for the better accuracy, I have to treat them.



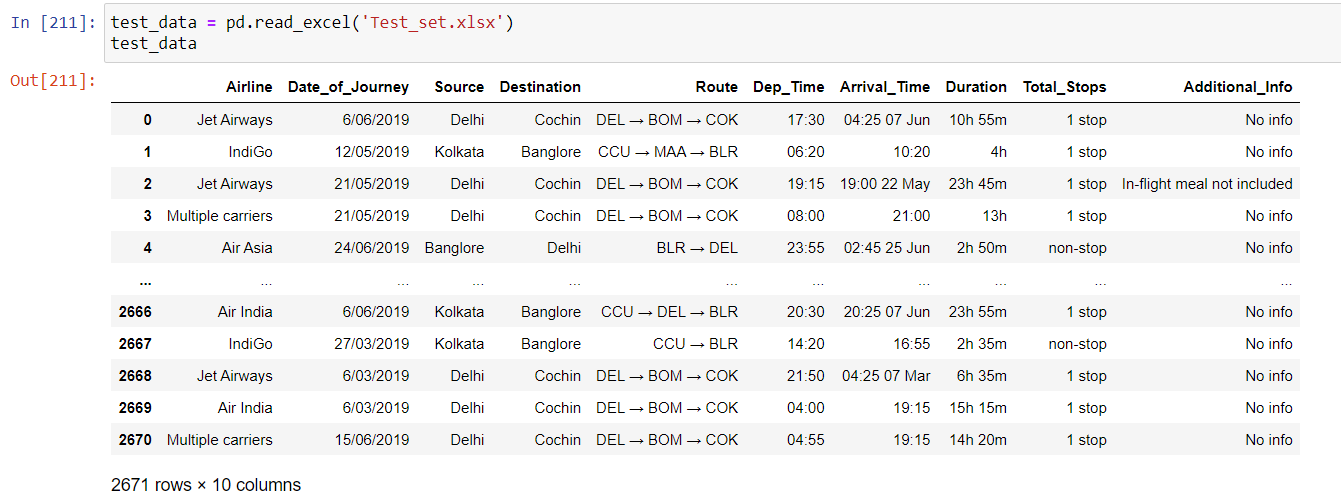
As, I have seen a few outlier’s were present, So by using the “Z-Score Method”, I have treated the outlier’s and at the last I can also see that some data has been dropped, So I have also mentioned the shape of the old data and the new data that is left with us.



Here, I can see that there is around 10% of the data loss. In removing the outlier’s.

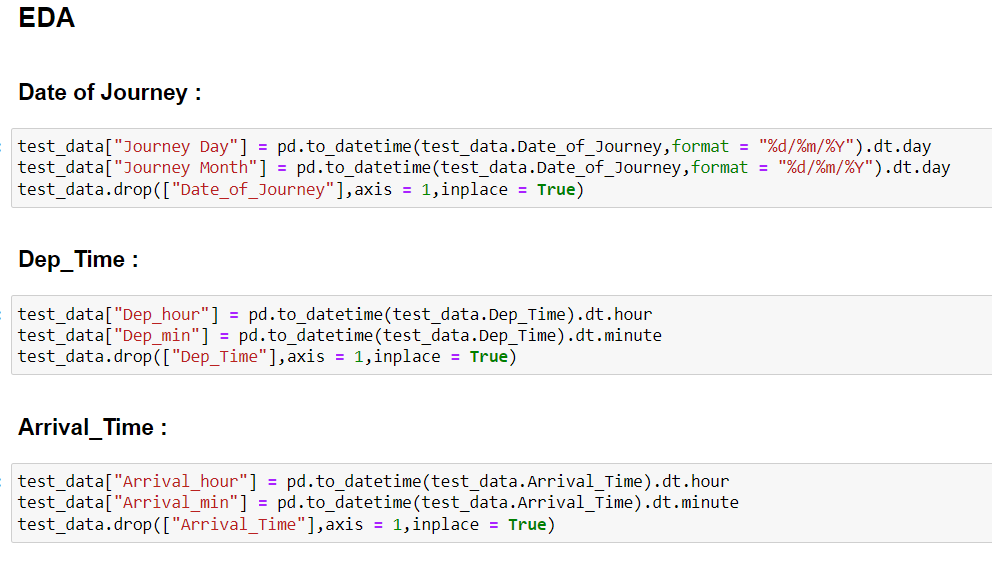
 

Here, I have checked the skewness of the data and then, all the columns, which had skewness in them, I have mentioned them in new list named, “features”. After that I have tried to remove the skewness of the data by using “Power Transform” method.



Now, I’m extracting the test data In my Jupyter Notebook.

**Test Data: -**



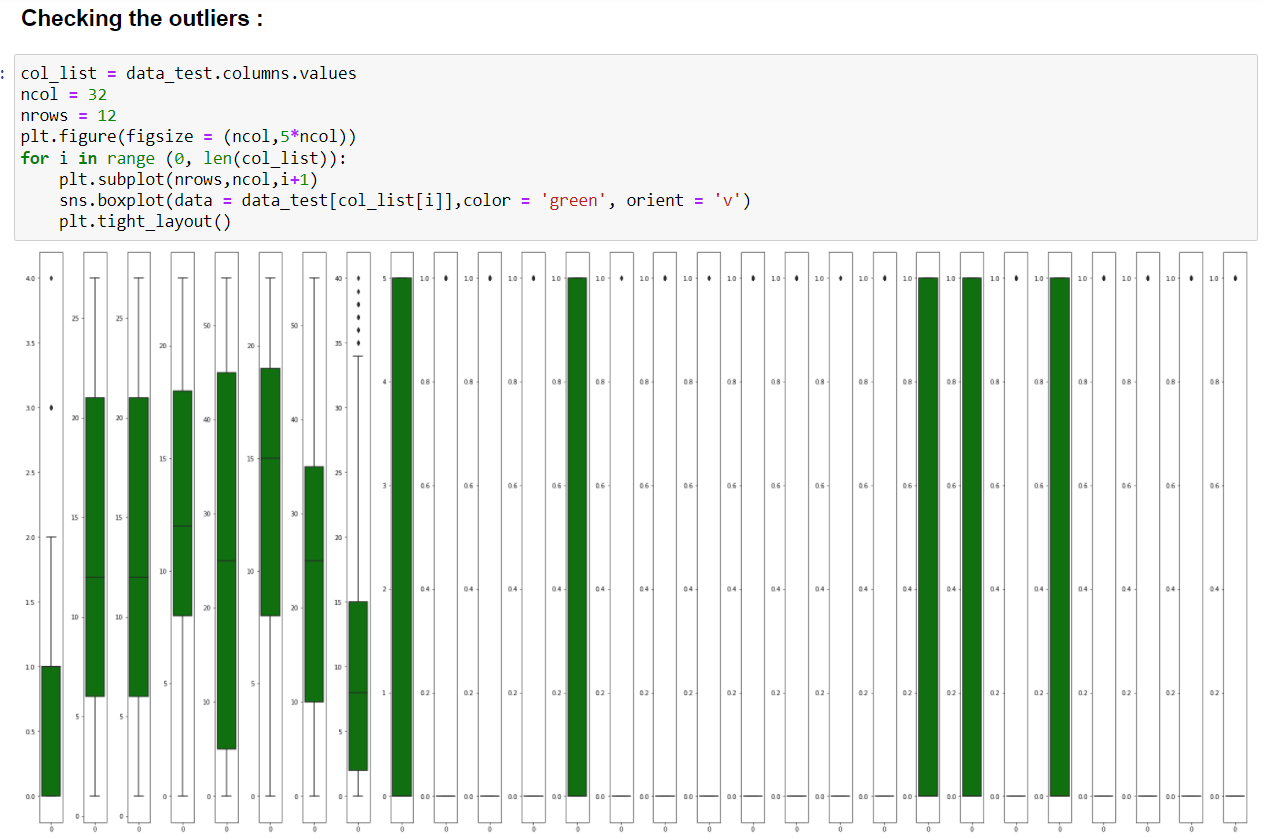
i. The column, "Date\_of\_Journey" was "object" type data so I have converted this into numerical data type so that I can use it for the model prediction.

ii. For, the conversion, I need “pandas.to\_datetime” to convert this object into “datetime” data type.

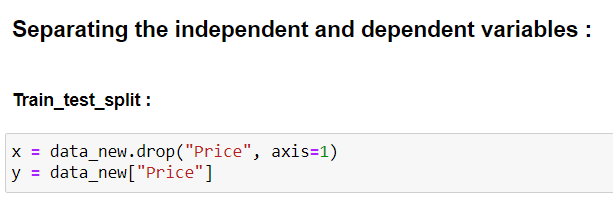
iii. Here I'll use ".dt.day" method for extracting day of that date and also I'll use ".dt.month" method for extracting month of that date.

iv. After that, I have made a new column for data and month and dropped the column, “Date\_of\_Journey”.

v. I have done the same thing in two other columns: - “Dep\_Time”,” Arrival\_time”



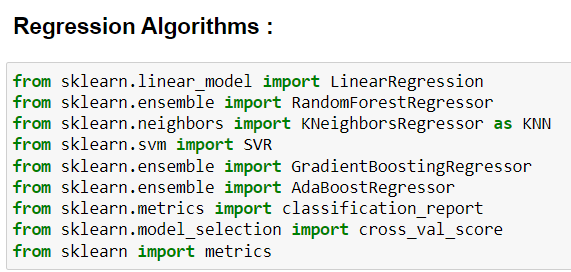
Here, for the Test data also, I have checked the outlier’s and same as training data, Here also I have used, “Z-Score method” for treating the outlier’s.



Here, I have splitted our “label” column, with the features, I have assigned our data to “x” and I have dropped “Price” from it and assigned “price” to “y”.

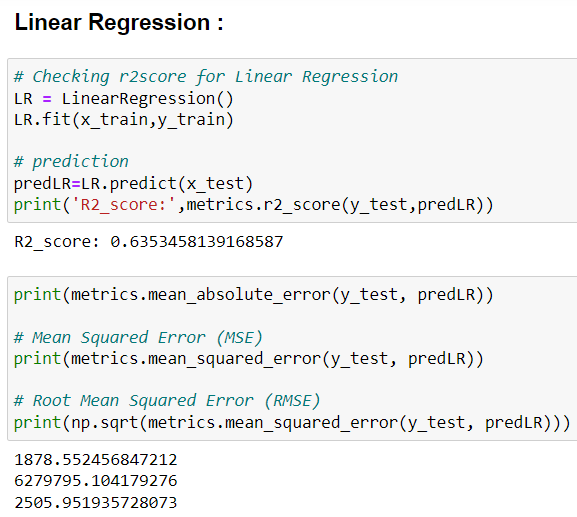
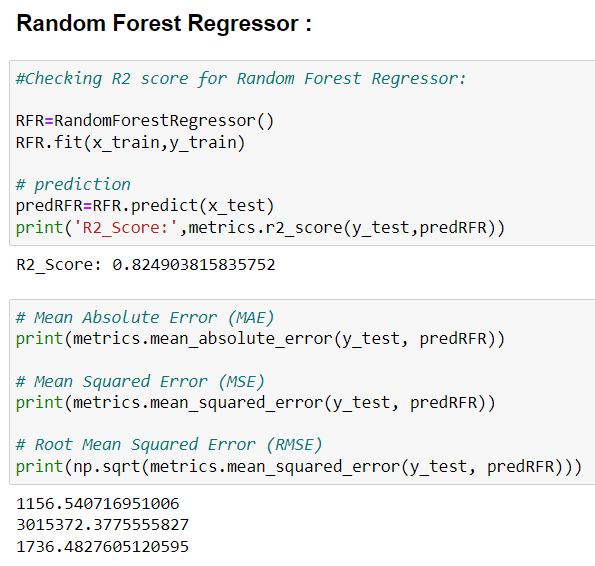
**Model Building: -**

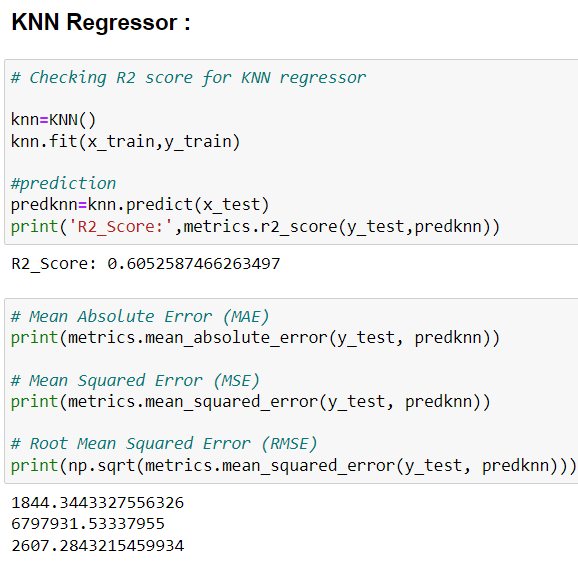
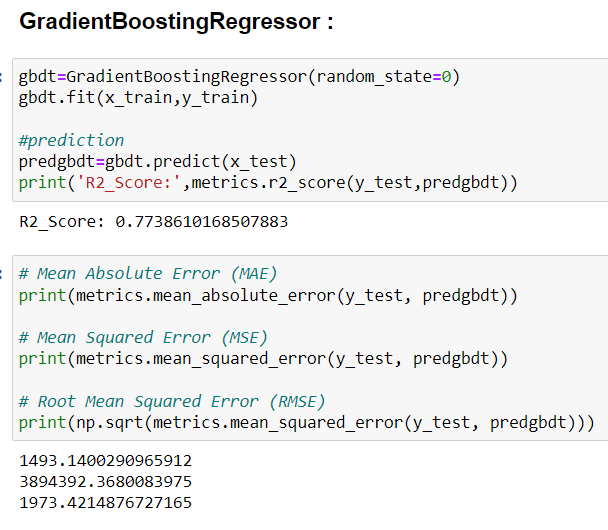
Importing the Required Libraries: -

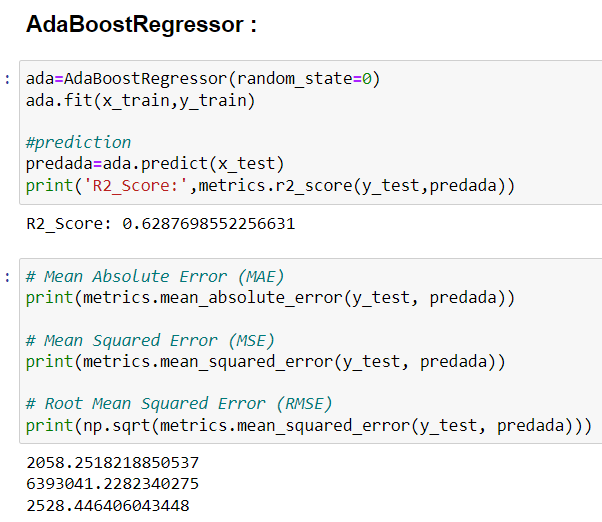


Here, I have imported the required libraries for our model building,

Now, I can proceed with the “Model Building”.





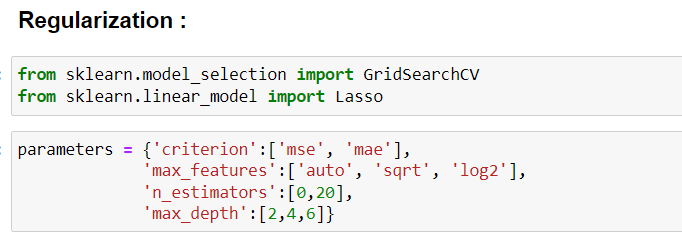
Here, I have made total 5 models, In that As I can see “Random Forest” has the highest “R2 Score”.

Now, I’ll further move for the Checking the CV(Cross Validation Score) of the predicted models. So, that I can get to know that my model is not “Overfit”.



Here, I have checked the Cross Validation Score of all the predicted models, and I come to know that “Random Forest Regression” has the highest “Cross Validation Score”.

**Hyper Parameter Tuning**

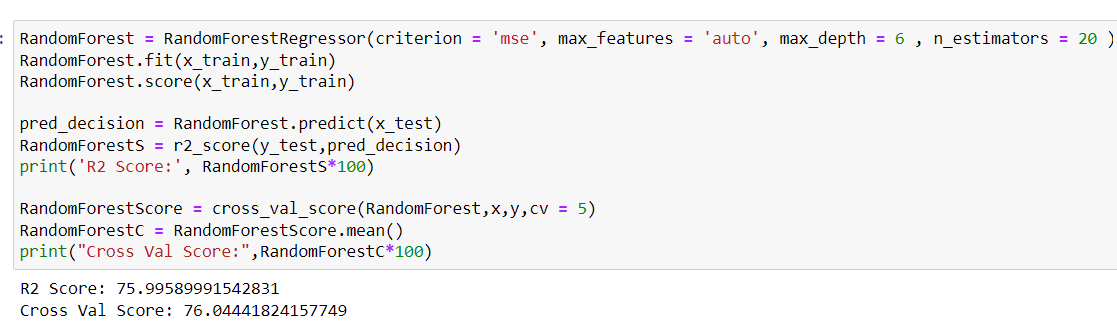


Here, I am doing Hyper Parameter Tuning of my predicted model and I have imported the required libraries for that.



Here, I used “Grid Search CV” and I have given some parameter’s, so that I can get to know which all parameter, Best Suit’s my model, and which all parameter will be helping me to gain the accuracy of my model and I have received the best parameter’s.

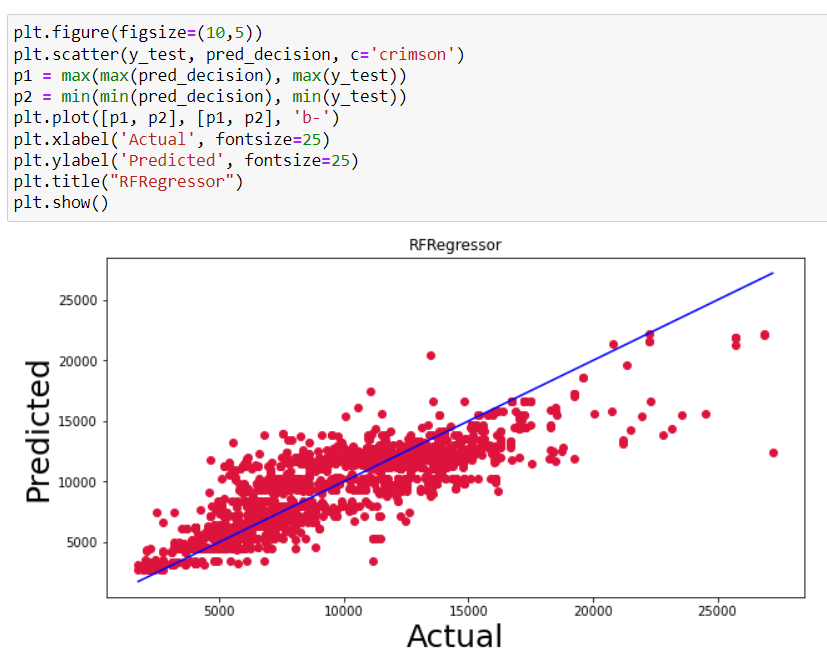
Next, I’ll be applying these parameter in my model.



Here, I have applied, all the above best parameter’s in, “Random Forest Regressor”.

Here, I have received 75.99% in “R2 Score” and 76.04% in “Cross Val Score”.

**Graph for predicted and Actual Values: -**



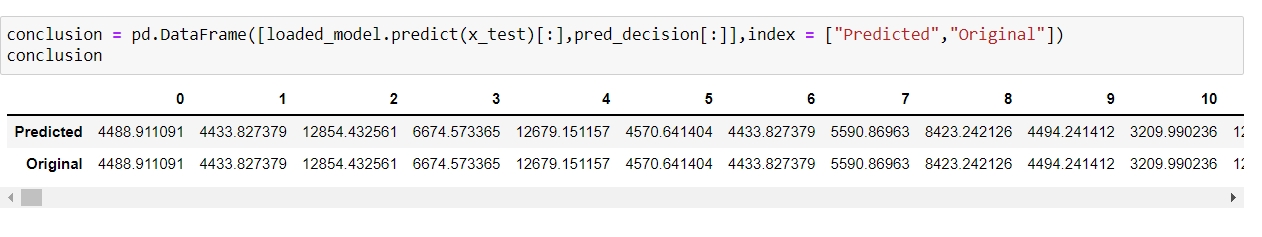
Here, I have Plotted a Graph between, “Actual” and “Predicted Value”. By seeing I can say that the graph is looking Linear.



Here, At the last I have saved the model and now, I have predicted the test score, and assigned it to a variable(“loaded\_model”).

**Conclusion: -**

**Predicted and Original Values: -**



Here, I have presented the predicted and the actual values of my dataset.

**Concluding Remarks: -**

* In This Model, I have a lot of “Feature Engineering” .
* Here, I had to extract multiple data from the same column.
* Here, I can also see that, People from, “New Delhi” travels the most.
* I have also seen that, people prefer “Jet Airways” the most for their travelling.



**Thank You**