**FLIGHT FARE PREDICTION**

**Problem Statement:**

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 travelers 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.

Size of training set: 10683 records

Size of test set: 2671 records

This Train dataset includes variables like:

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.

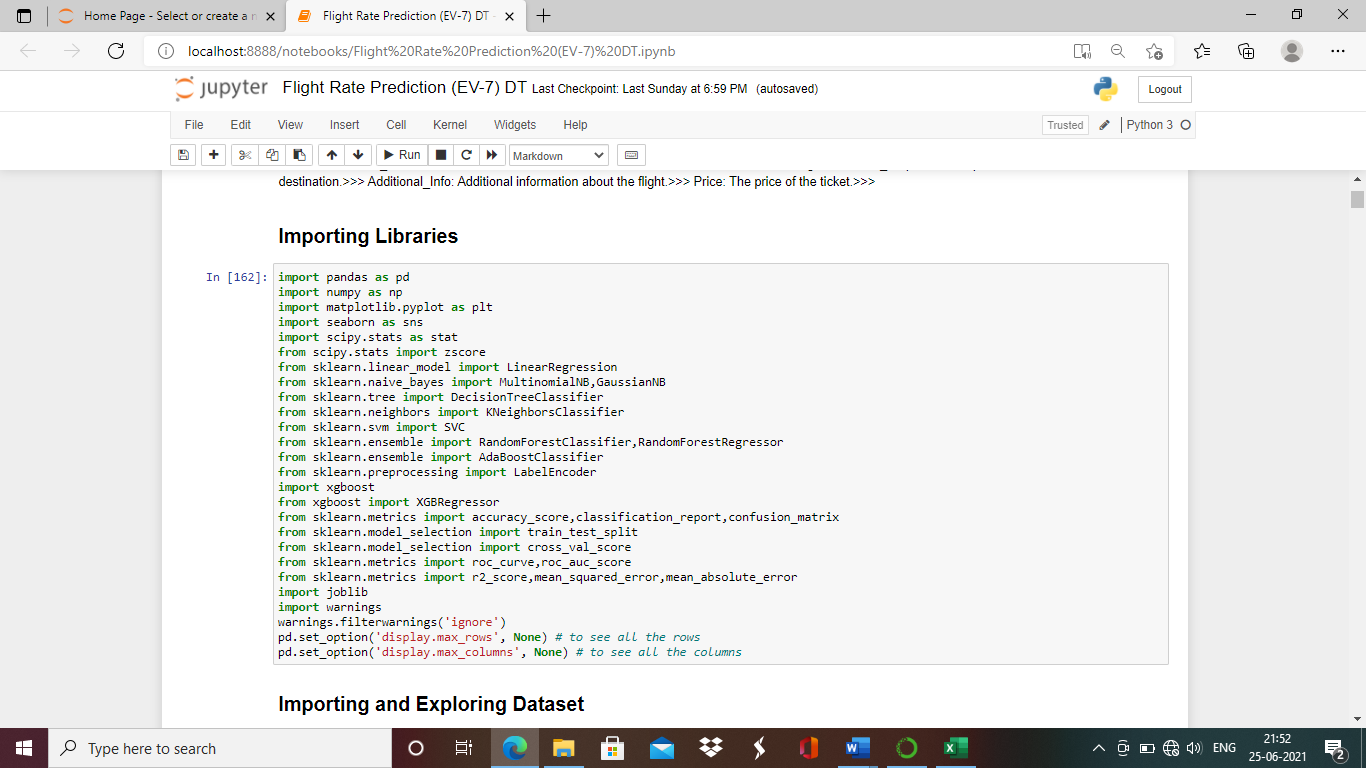
Total\_Stops: Total stops between the source and destination.

Additional\_Info: Additional information about the flight

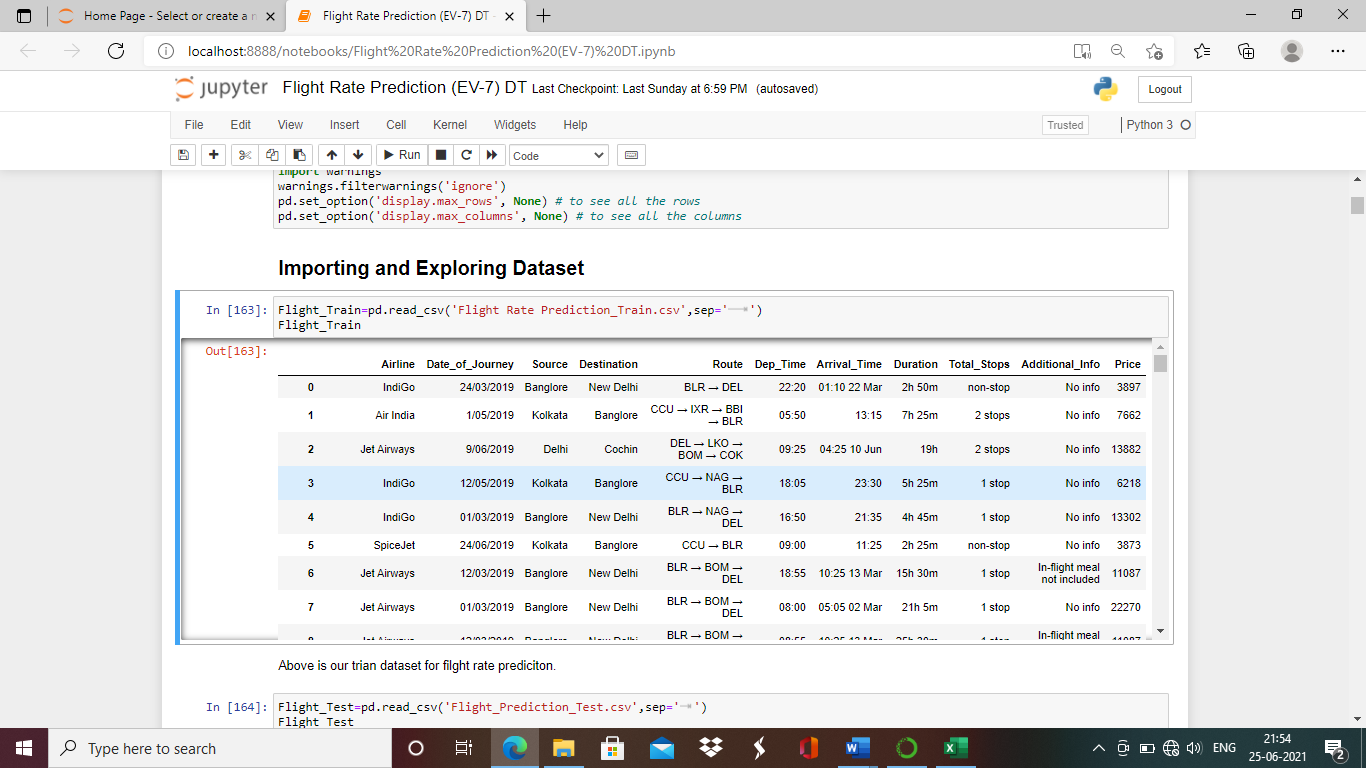
Price: The price of the ticket.

**Data Analyzing:**

We will import all the required libraries to call the dataset and execute the models.

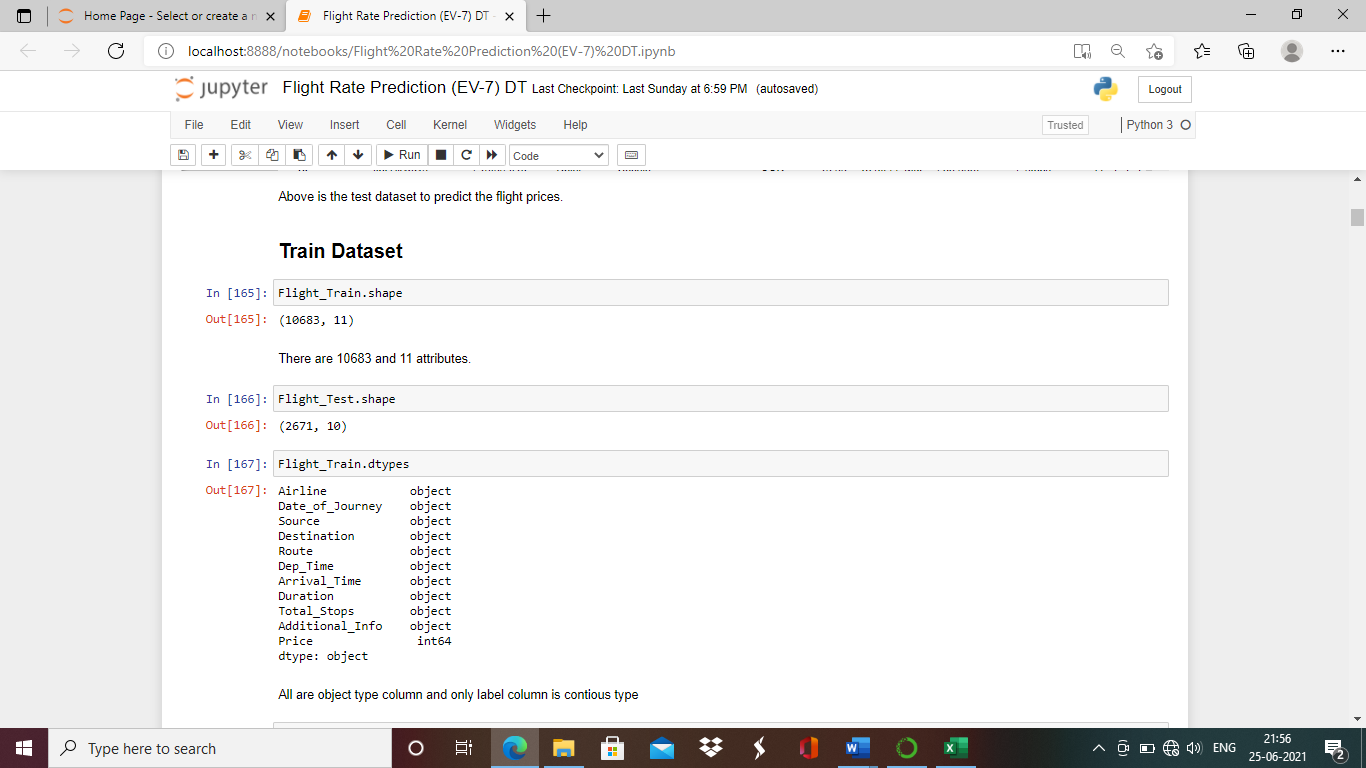


Importing the dataset for analyzing .



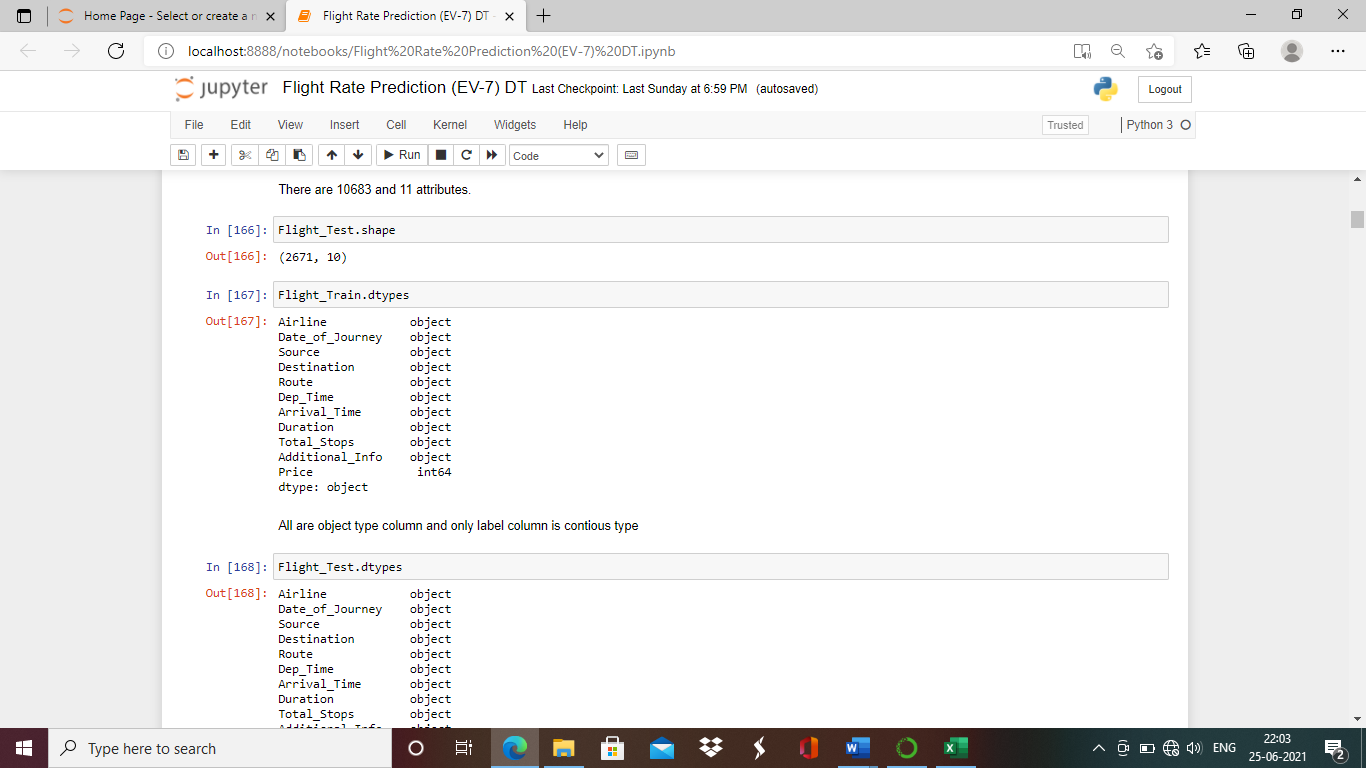
Dataset is called as Flight\_Train further for model execution.Now lets check number of rows and columns in dataset.

Check number of rows and columns in dataset.



There are 10683 rows and 11 attributes in this Flight\_Train dataset.

Check the data types of all the columns.



All the columns in this dataset are object type and only target column is integer type.

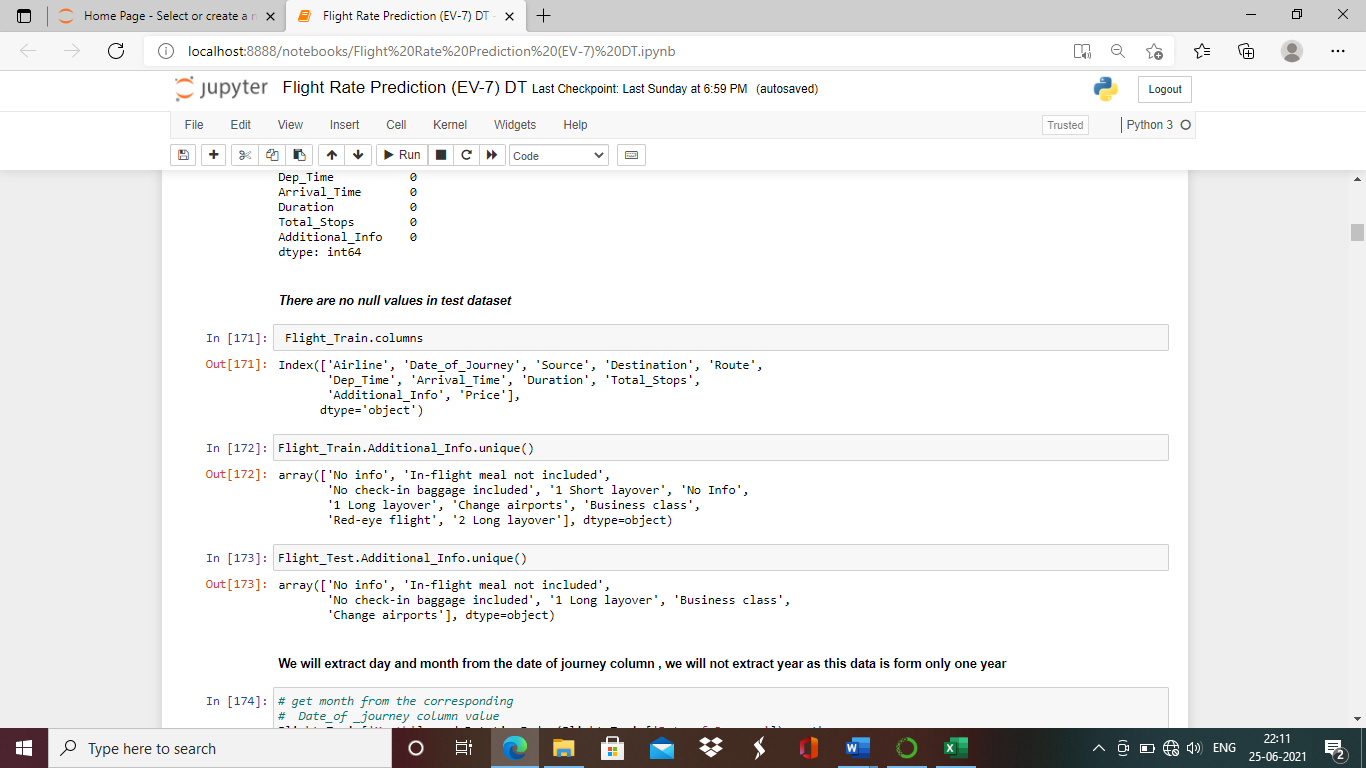
Let’s ,check if any null values are present in dataset.



We can conclude from above code that only route and total stops column have null values .

Each of those column have 1 null vlaue.

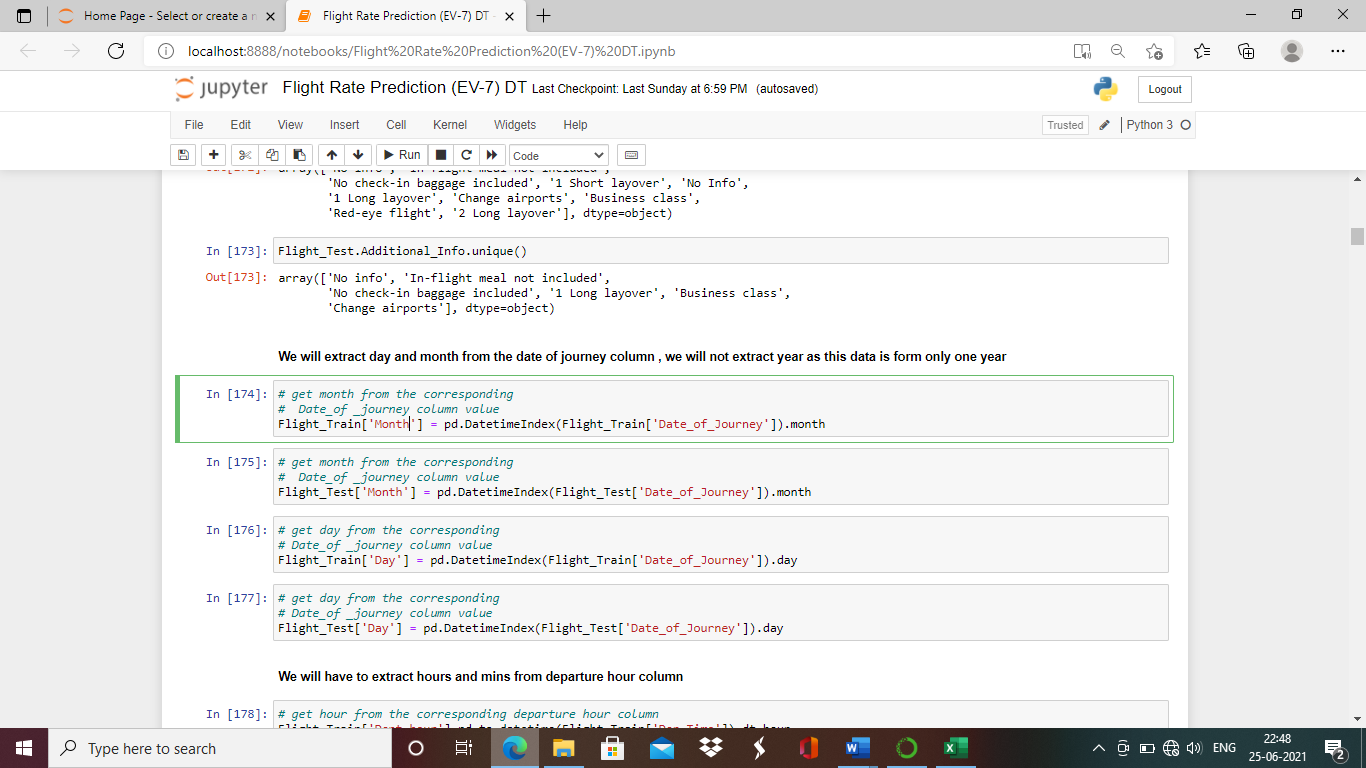
Check the column names:



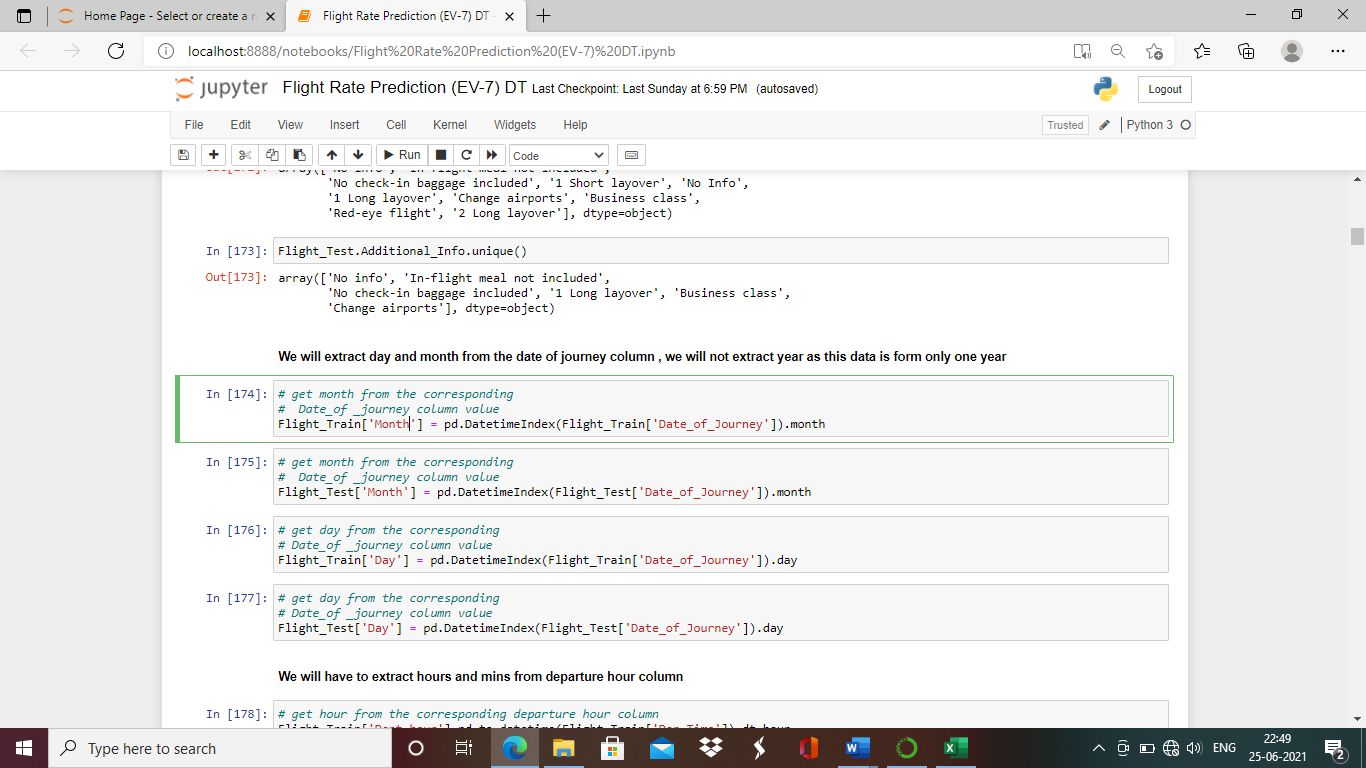
After importing and analyzing the dataset we could understand Airline column has categorical data,Date of journey has data in date format which we will split in month and year for better analyzing,Source & destination is categorical column,Route is 2 or 3 places with special character which also we will have to treat for better analyzing,Dep\_Time,Arrival\_Time has data in time format which is not suitable to be sused without treating hence we will split data in mins and hours column,Total\_stops and Additional\_info has categorical data.

**Date Of Journey**

We will extract month from the date of journey column.



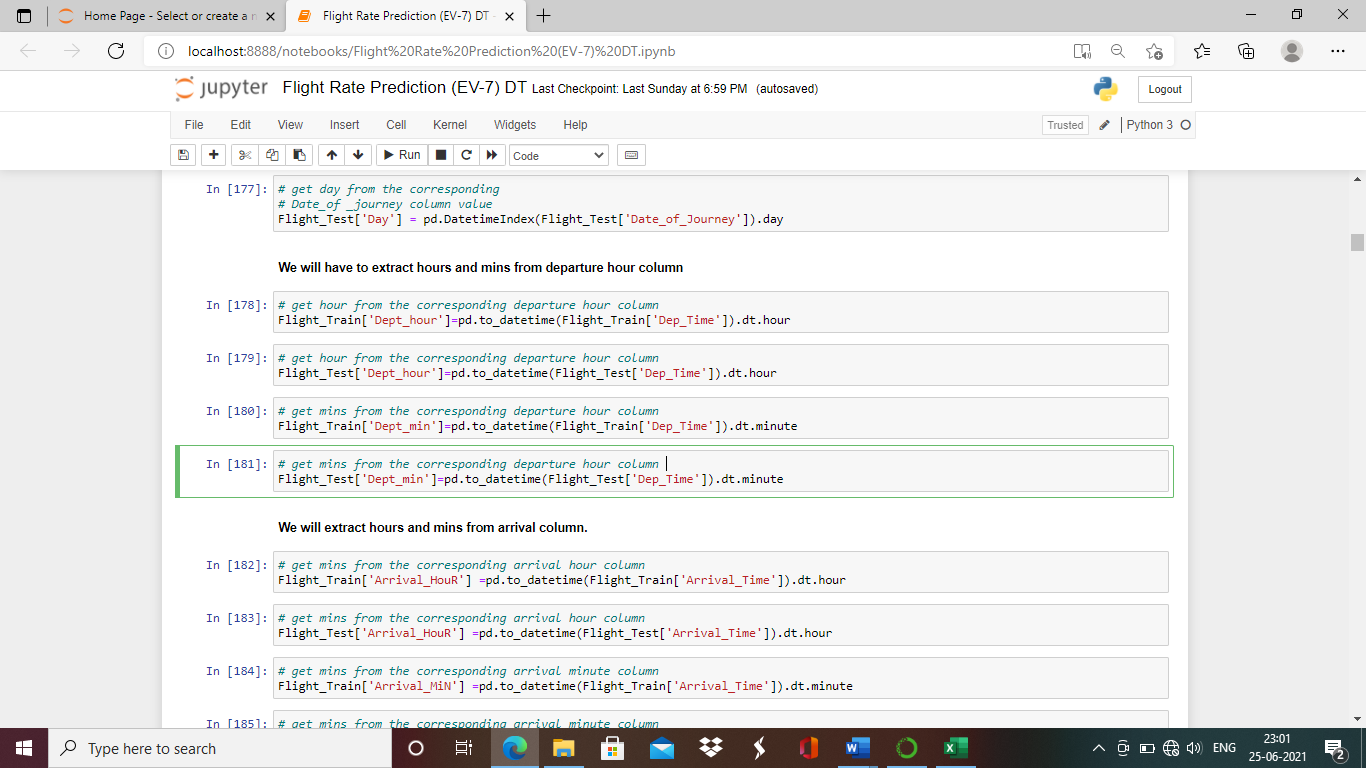
We will extract day from the date of journey column.



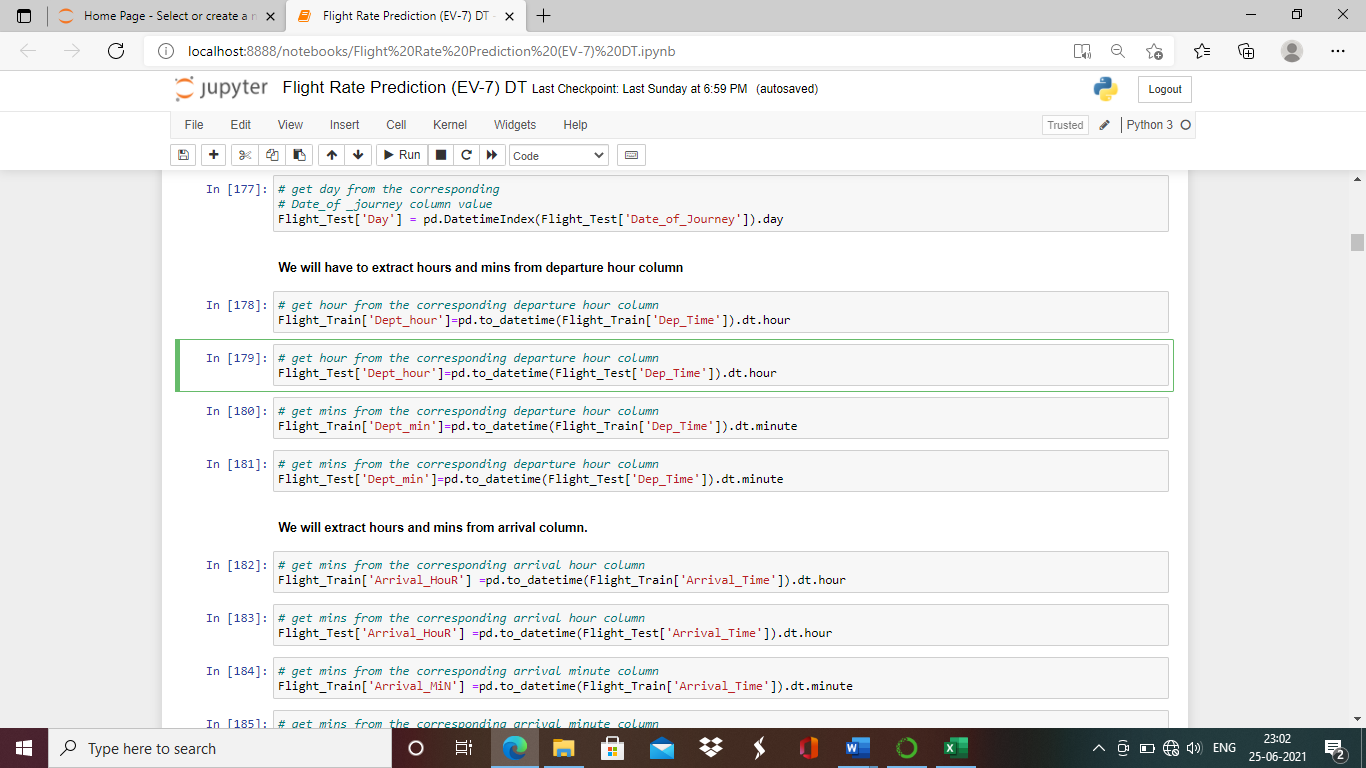
We will not extract year from the date of journey column as this data is only for one year.

**Dept\_Time**

We will extract hours from departure time column.

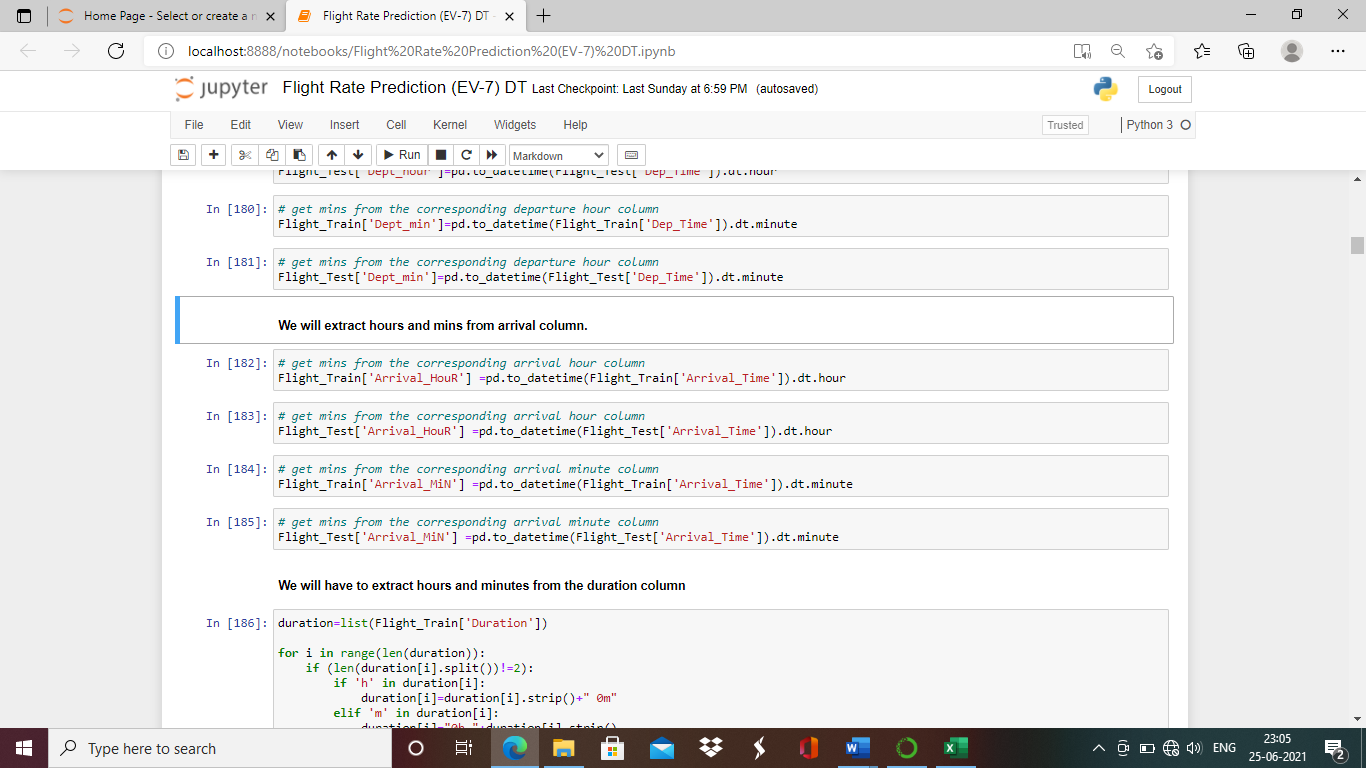


We will extract hours from departure time column.

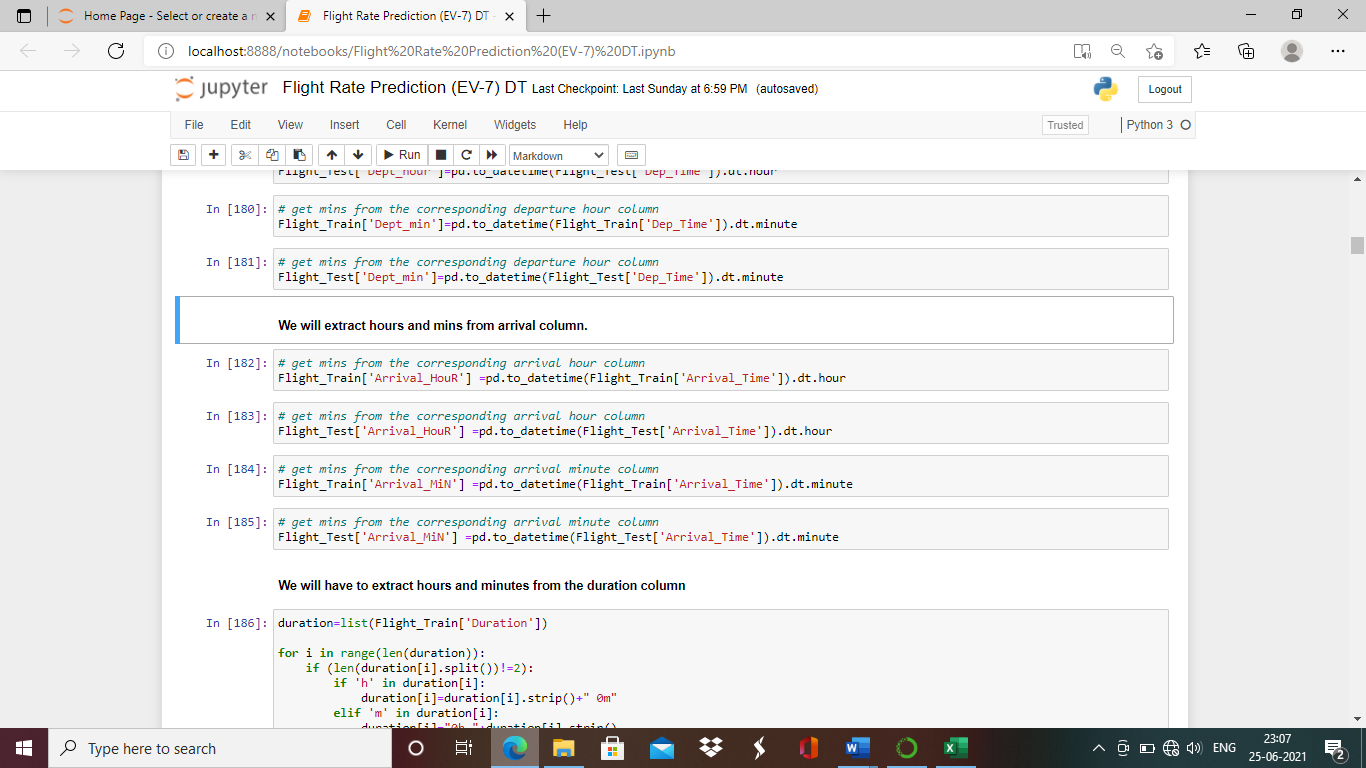


**Arrival\_Time:**

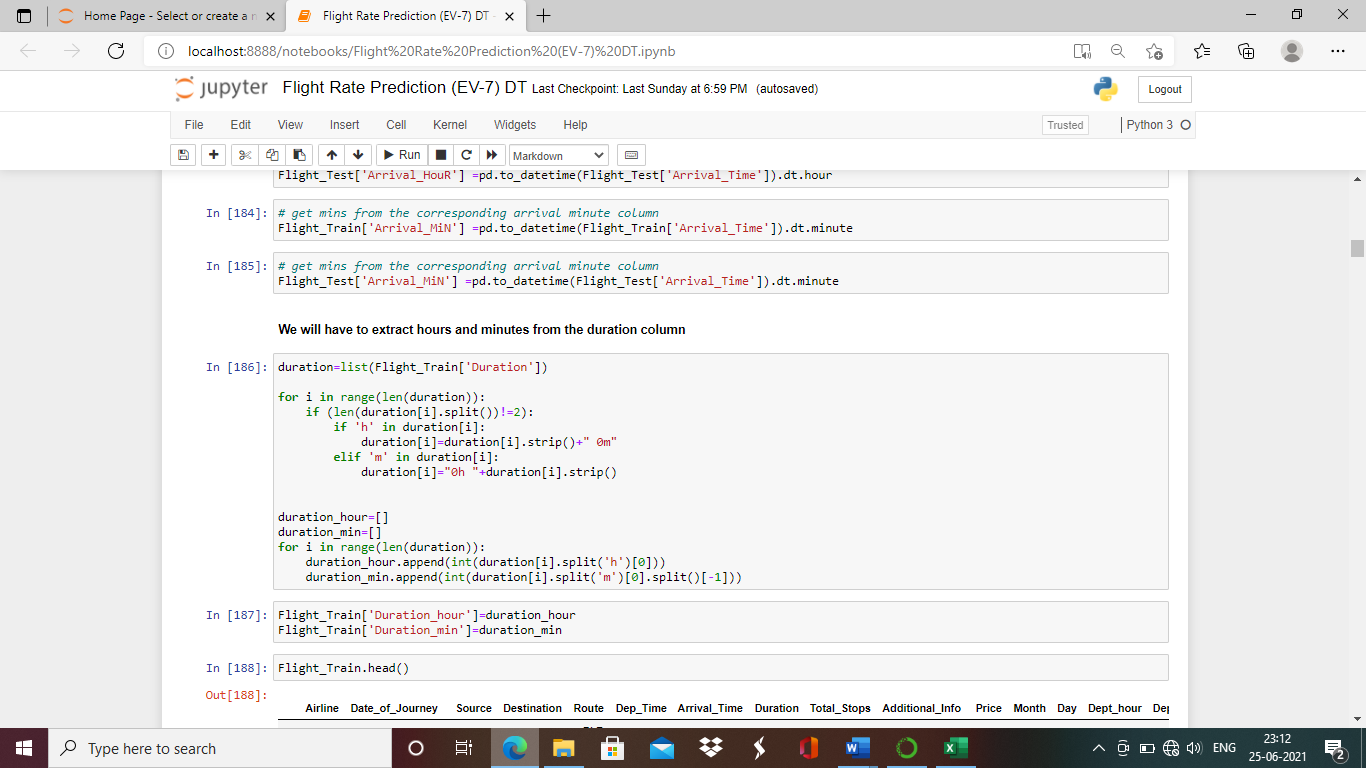
We will extract hours from arrival column.



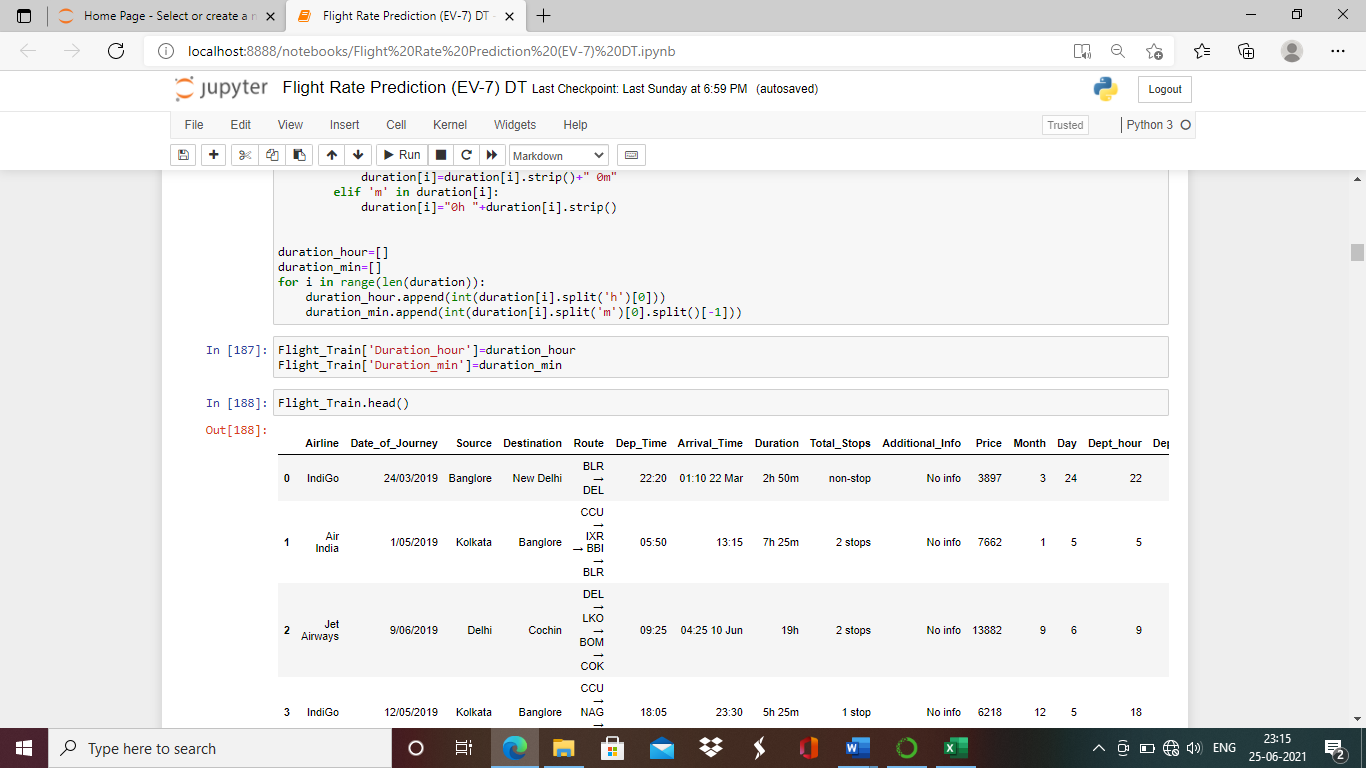
We will extract minutes from arrival column.



We will now extract hours and mins from duration column, but in this column time isn’t in same format as of arrival and departure time column.Hence we will make for loop .

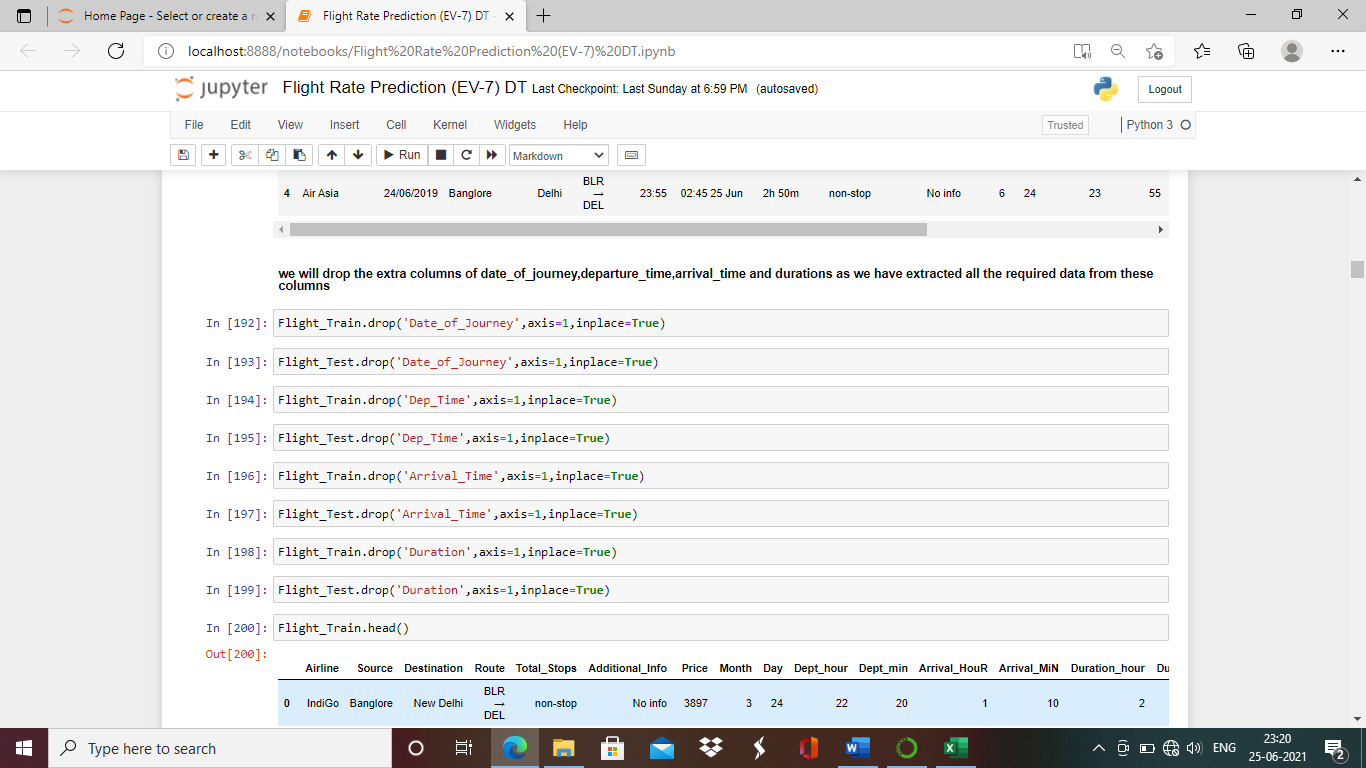


We will save hour and minutes from duration column in duration hour and duration min arrays.



After splitting and extracting all the required data from departure time, arrival time, date of journey and duration columns, we will drop all the extra columns

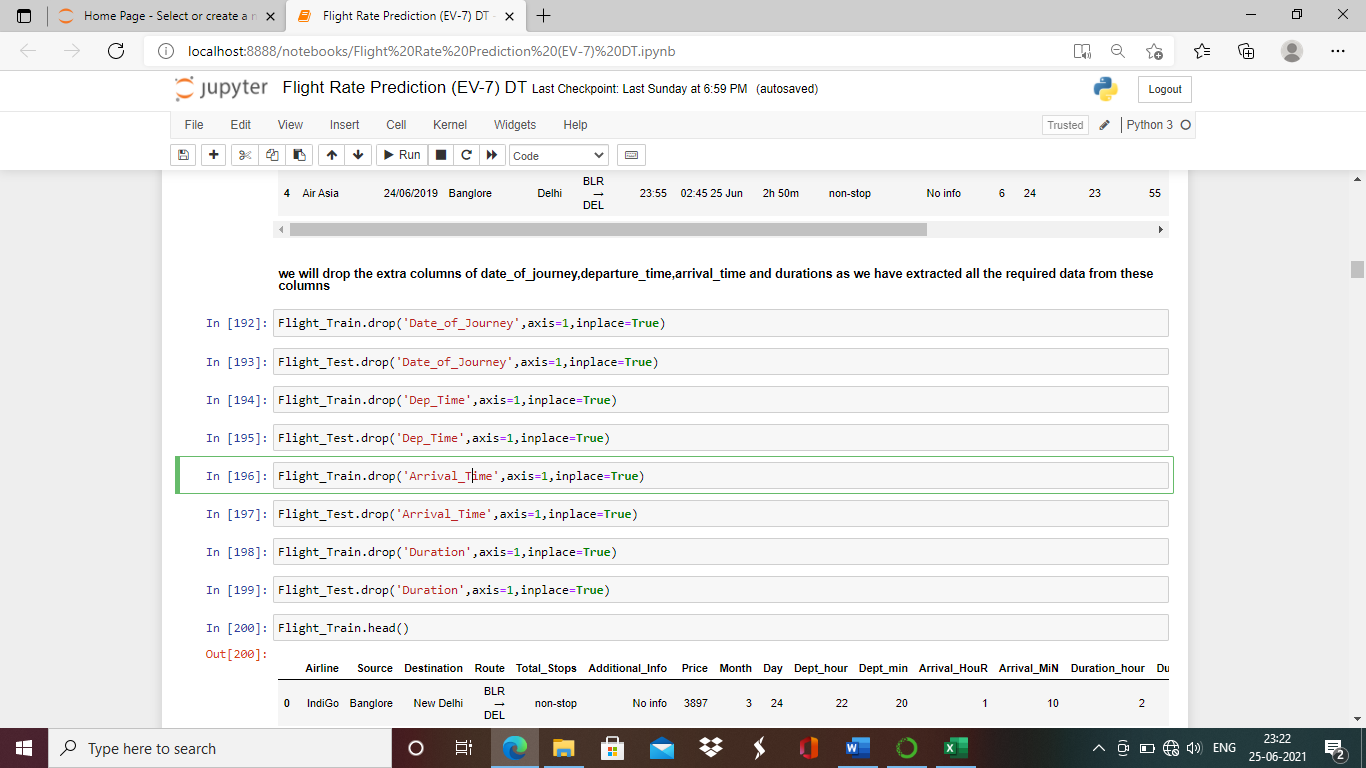
Drop Date\_of\_journey.



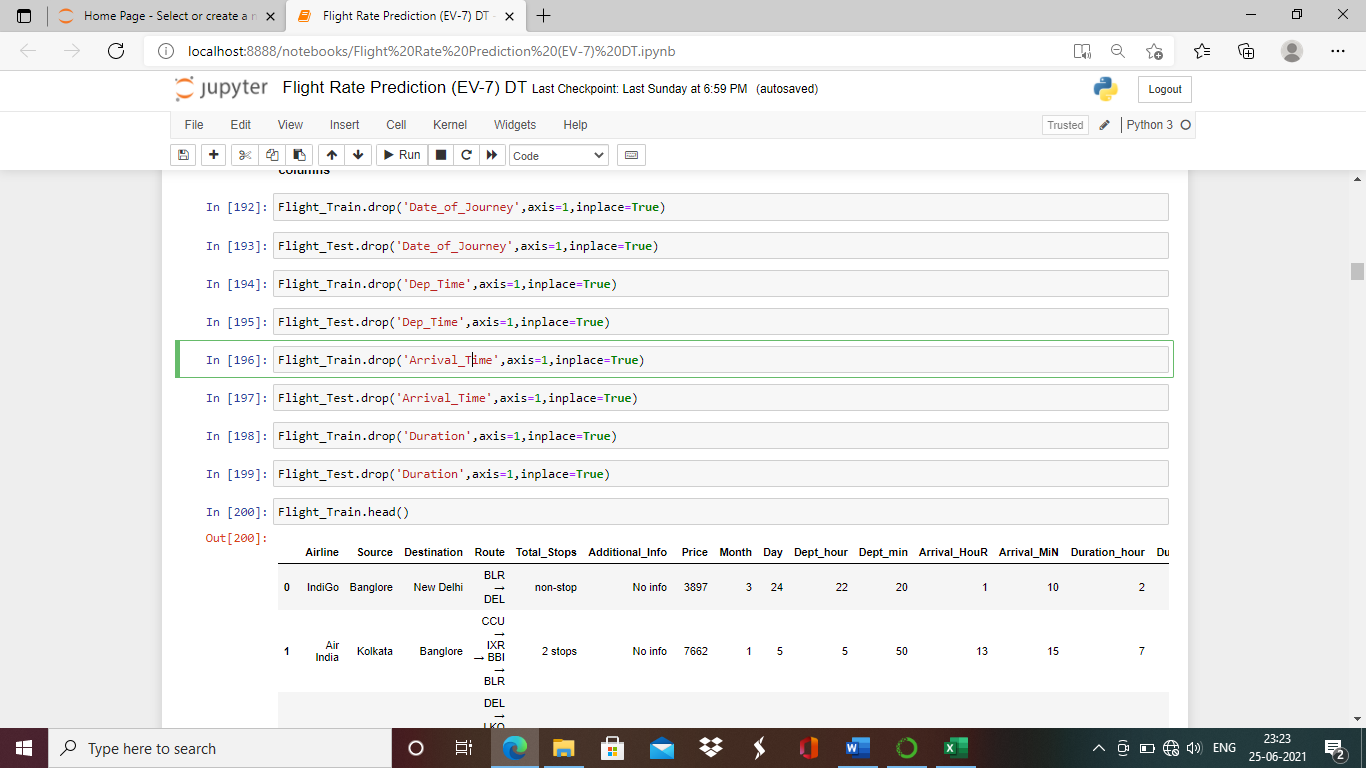
Drop Arrival\_Time



Drop Departure\_Time



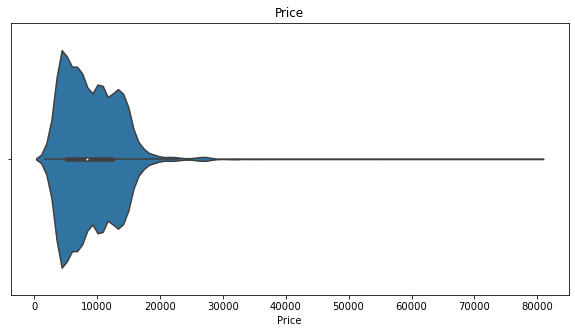
Drop Duration



**EDA Concluding Remarks**

**Univariate Analysis**

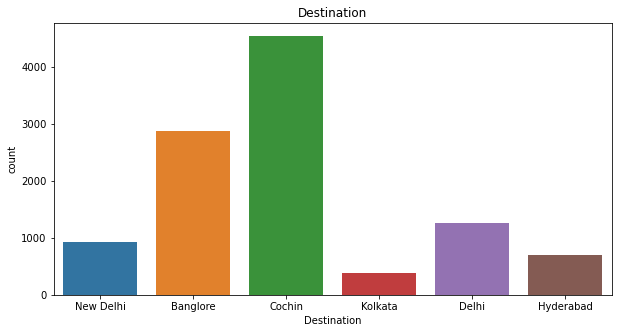
**Violin Plot for Price :**



Observation:

Flight prices range highly from Rs:0 to 20000 and are peak at around Rs:6000.

**Count Plot for Destination:**

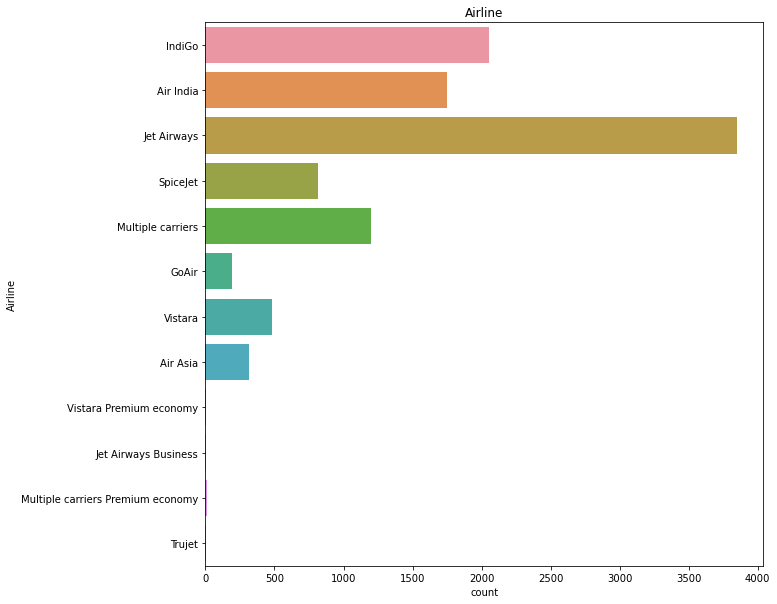


Observation:

1)Most of the flights are having final destination Cochin.

2)Lest flights in dataset have destination Kolkata.

3)Other destinations where flights are flying up to are New Delhi, Banglore, Delhi and Hyderabad.



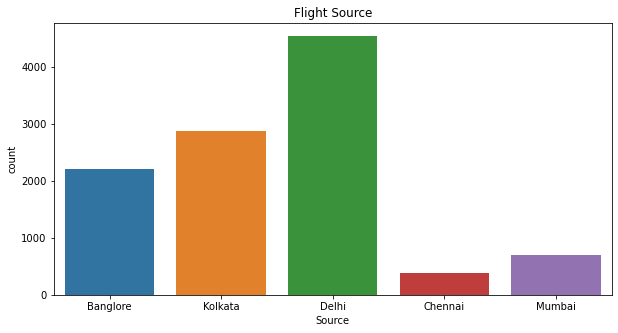
Observation:

1)Jet airways flights are the most in the dataset.

2)Vistara Premium Economy, Jet Airways Business Multiple carriers’ premium economy and Trujet are least flights in dataset.

3)Other flying flights are Indigo,Air India,Spicejet,Multiple carriers,GoAir,Vistara and Air asia .

**Count plot for Flight Source:**

****

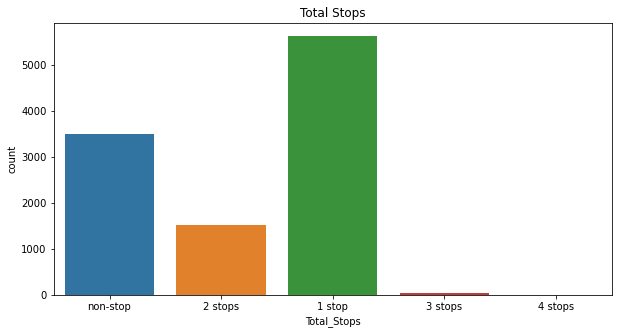
Observation:

1)Delhi is the source place for most of the flights in dataset.

2)Chennai is the source place for very flights.

3)Flights are also from Banglore ,Kolkata and Mumbai.

**Count plot for Total Stops :**

****

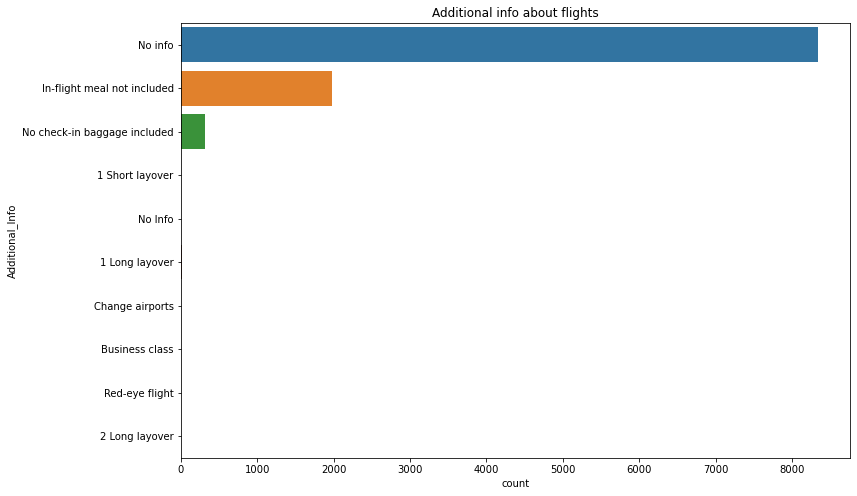
Observation:

1)Flights with 1 stop in the journey are the highest in dataset.

2)Non stop flights are the second highest in the dataset.

3)Very few flights are with 3 and 4 stops.

**Count Plot for Additional Info:**



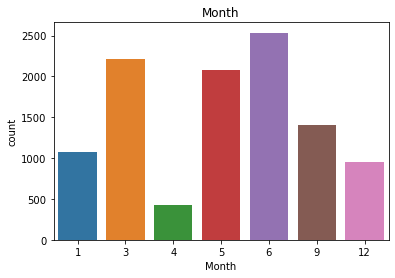
Observation:

1)Most of the columns are not giving us any info.

2)For few fights we get to know that meals is not included and also for few check in baggage isn't included.

3)Other info's what we have got can be considered as not essential.

**Count plot for Flight in Month:**



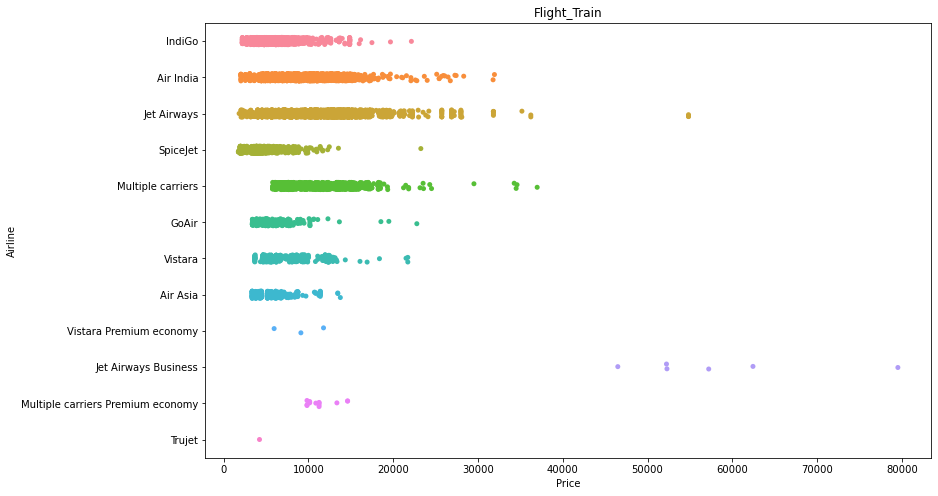
Observation:

1)We have data for most of the flights on june.

2)We have data of very few flights from month of april.

**Bivariate Analysis:**

**Strip Plot for Airlines and Price:**



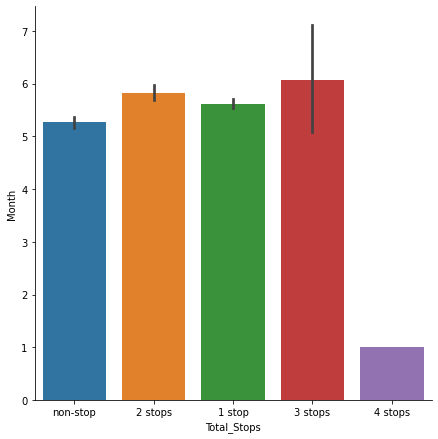
Observation:

1)Jet Airways Business ailine price are very high compared to other airlines.

2)IndiGo,Air India,Jet Airways,Spice jet have prices in similar range.

3)GoAir,Vistara and Air Asia have price range in almost similar range.

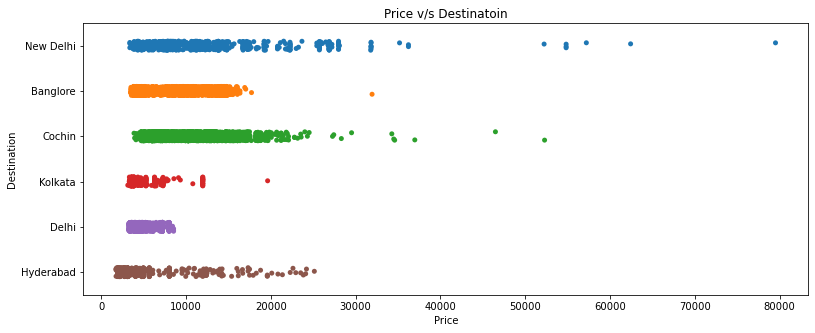
**Cat Plot for Total stops V/s Month:**

****

Observation:

1)Flights in month of jan are having large number of stops.

**Strip Plot for Destination v/s Price**

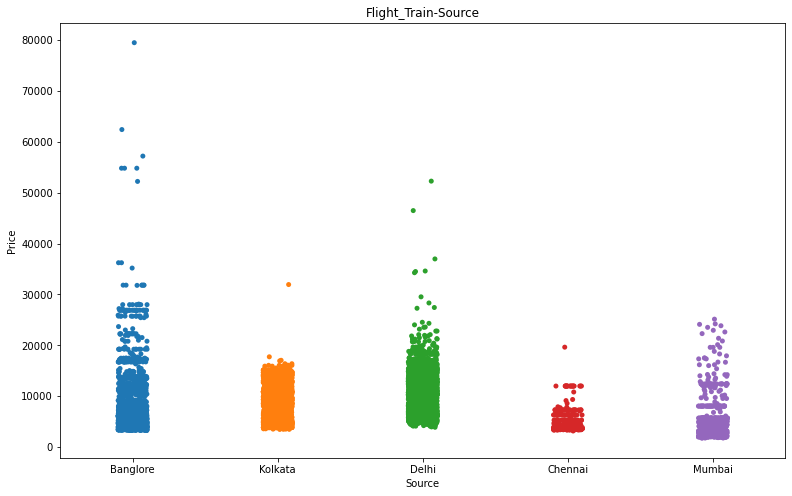
****

Observation:

1)The price of flights flying to Delhi is high.

2)Flights with destination kolkata have least fare.

**Strip Plot for Source v/s Price:**

****

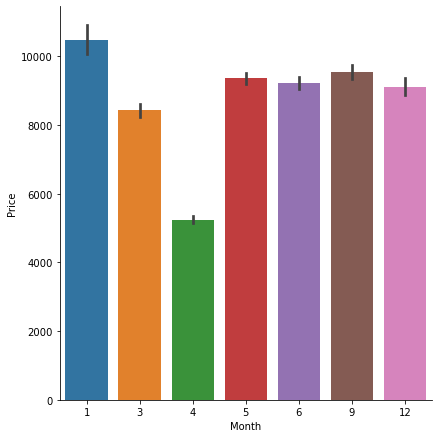
Observation:

1)Charges are at higher side person is flying from Bamglore.

2)Flights taking off from delhi are at second number after banglore in terms of flight charges.

3)Flight flying from Mumbai have the cheap prices

**Cat Plot for Month v/s Price**

****

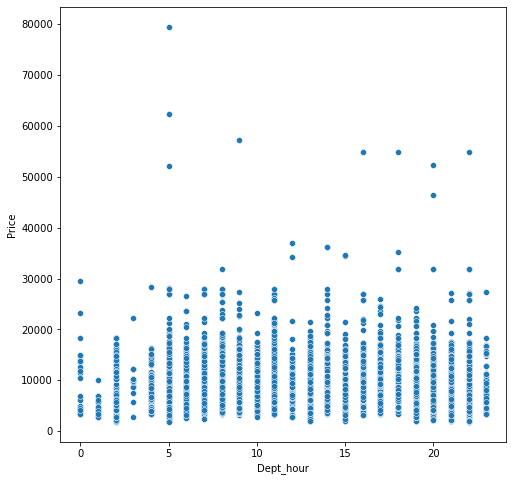
Observation:

1)Flight prices are high in month of jan.

2)Flight prices are least in month of april.

3)Flight prices are similar in month of june,sept and december.

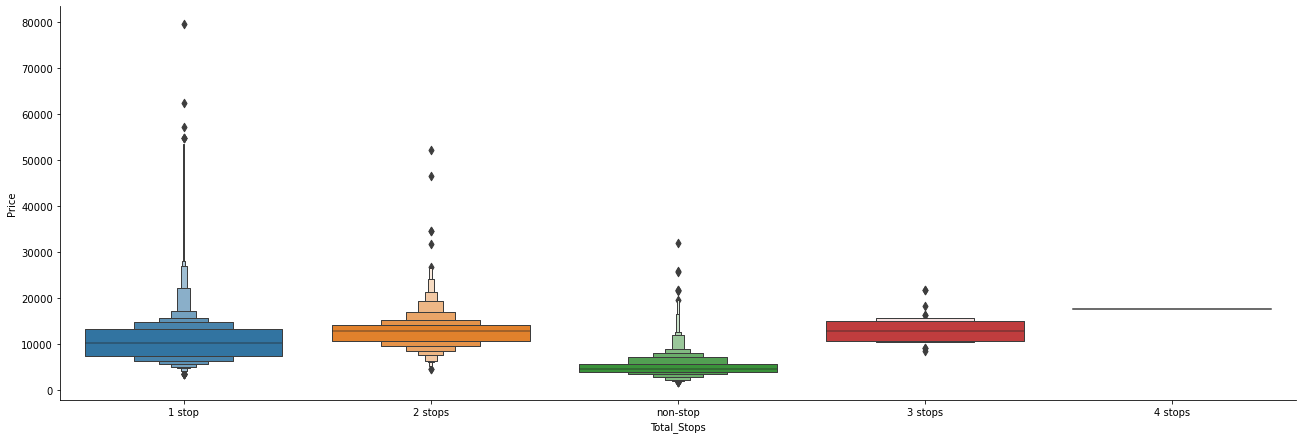
**Scatter Plot for Dept\_hour v/s Price**

****

Observation:

1)Flight prices are less during late nights and early mornings.

**Cat plot Total stops v/s price using box pattern:**

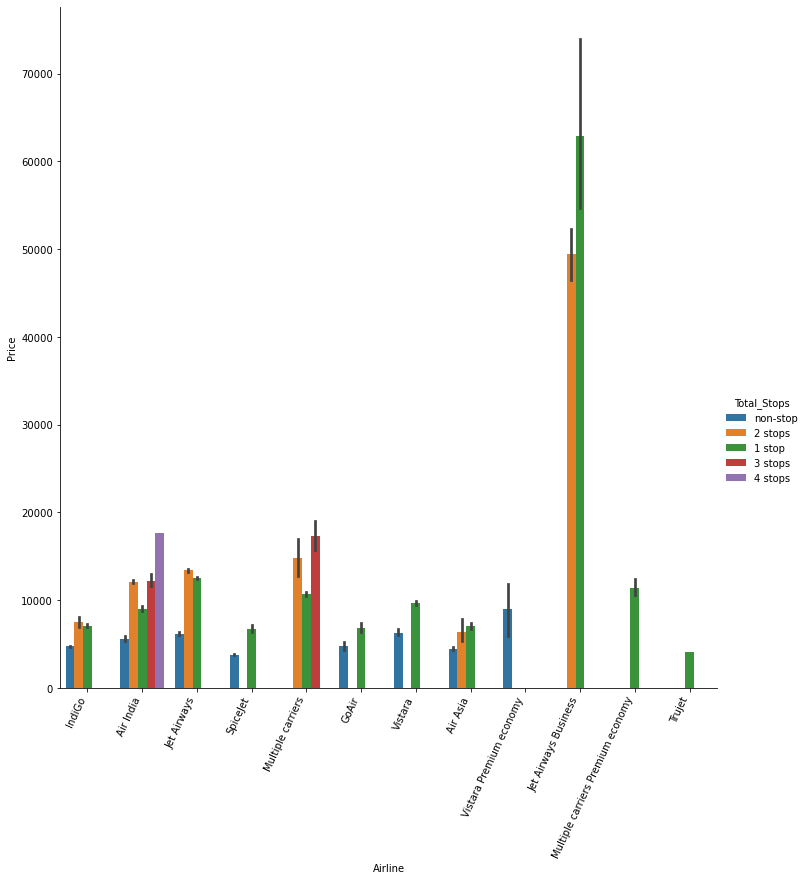


Observation:

1)Prices are high with 1stop.

2)Flights with no stop tend to have best price.

**Cat Plot for Airline v/s Price:**

****

Observation:

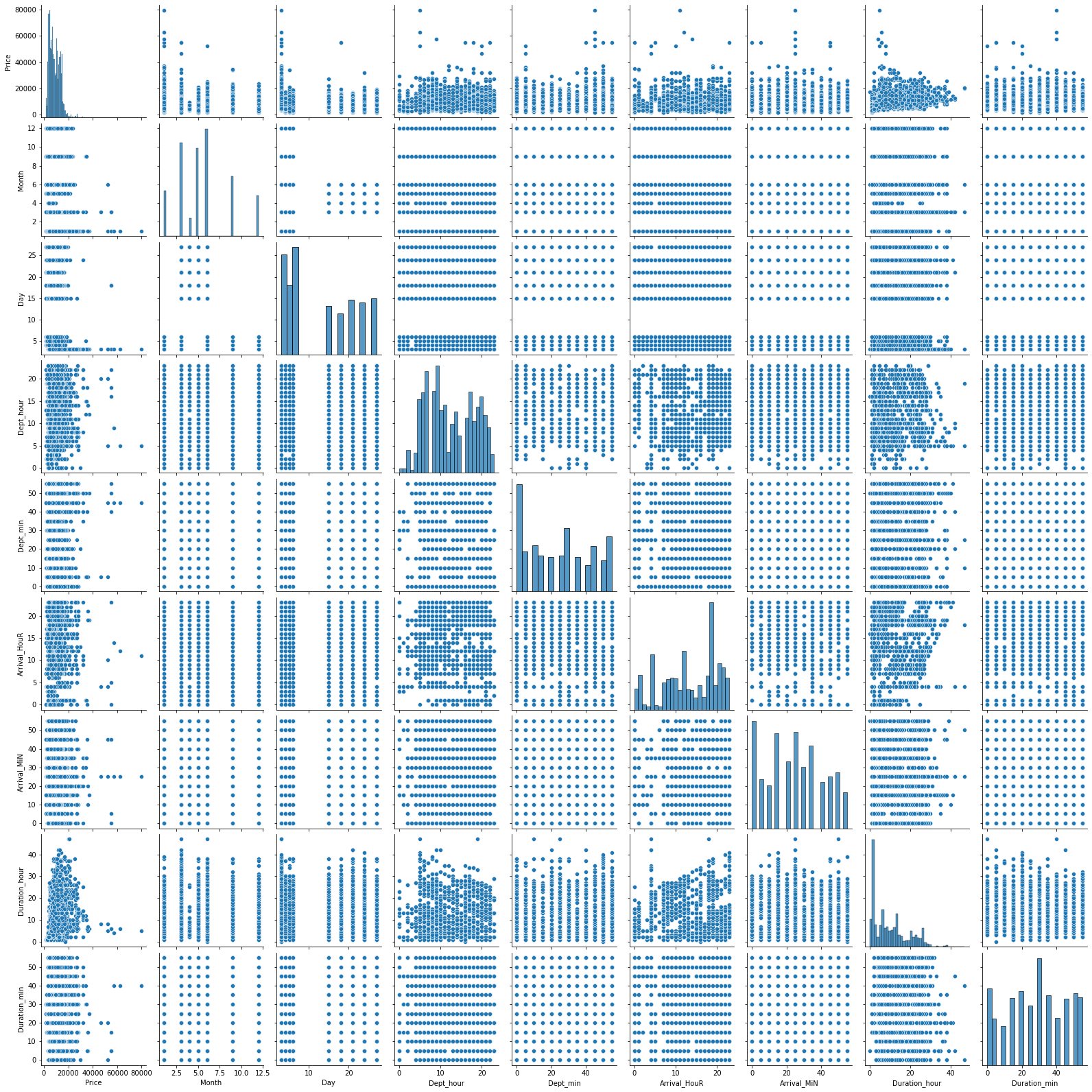
1)Air India and Multiple carriers are the only airlines with 3 stops.

2)Trujet Multiple carriers premium economy airlines have flights with 1 stop and Vistara airlines have flights with no stops.

3)Only Air India has got flights with 4 stops.

4)Spice jet, Go Air, Vistara have flights with no stop or 1 stop.

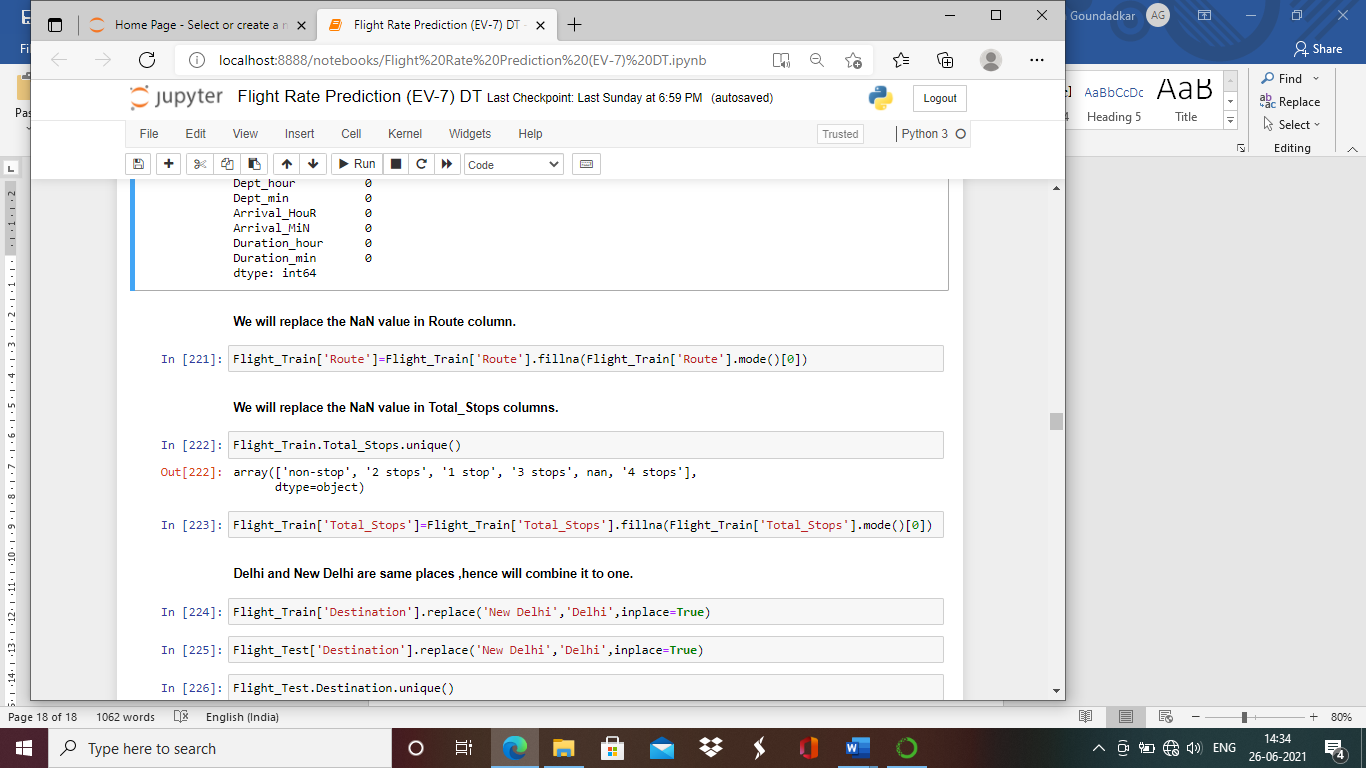
**Multivariate Analysis**

****

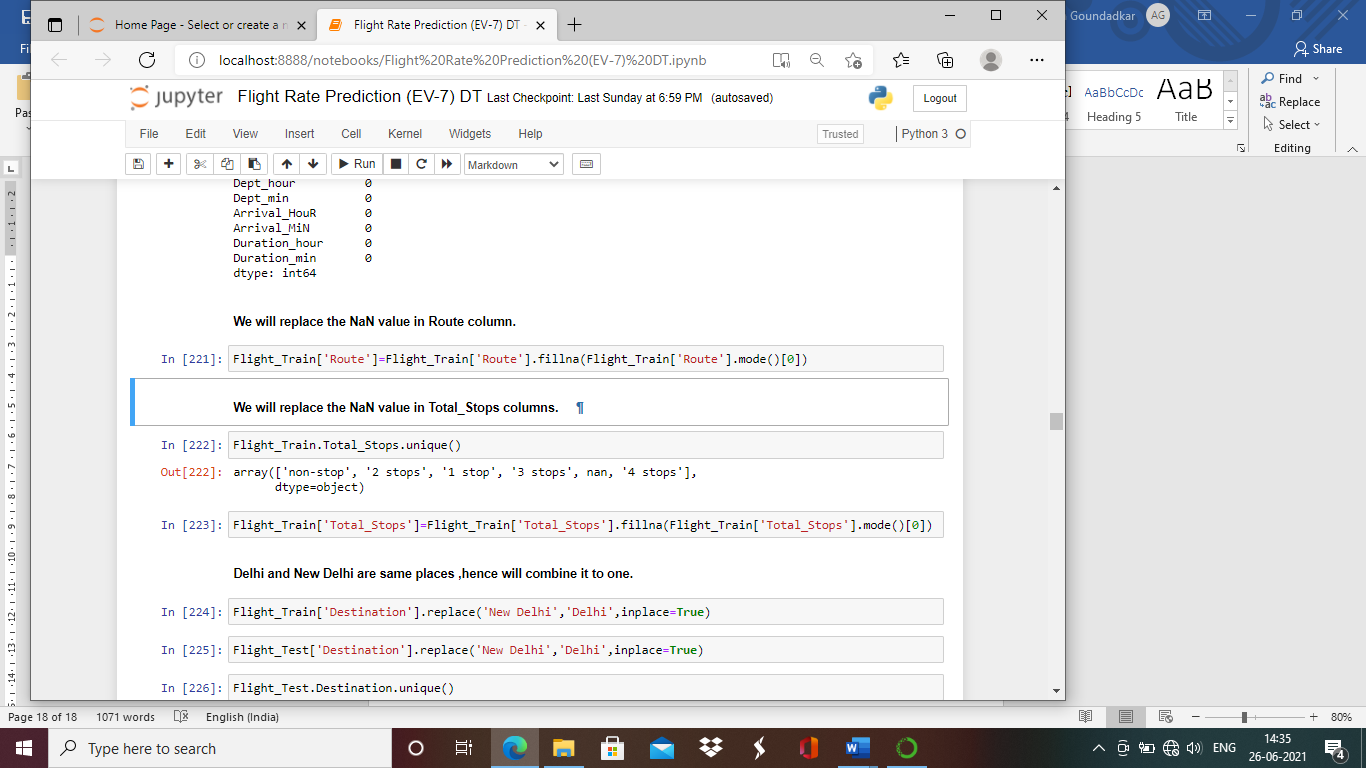
**Pre-processing Pipeline**

Treat the missing values.

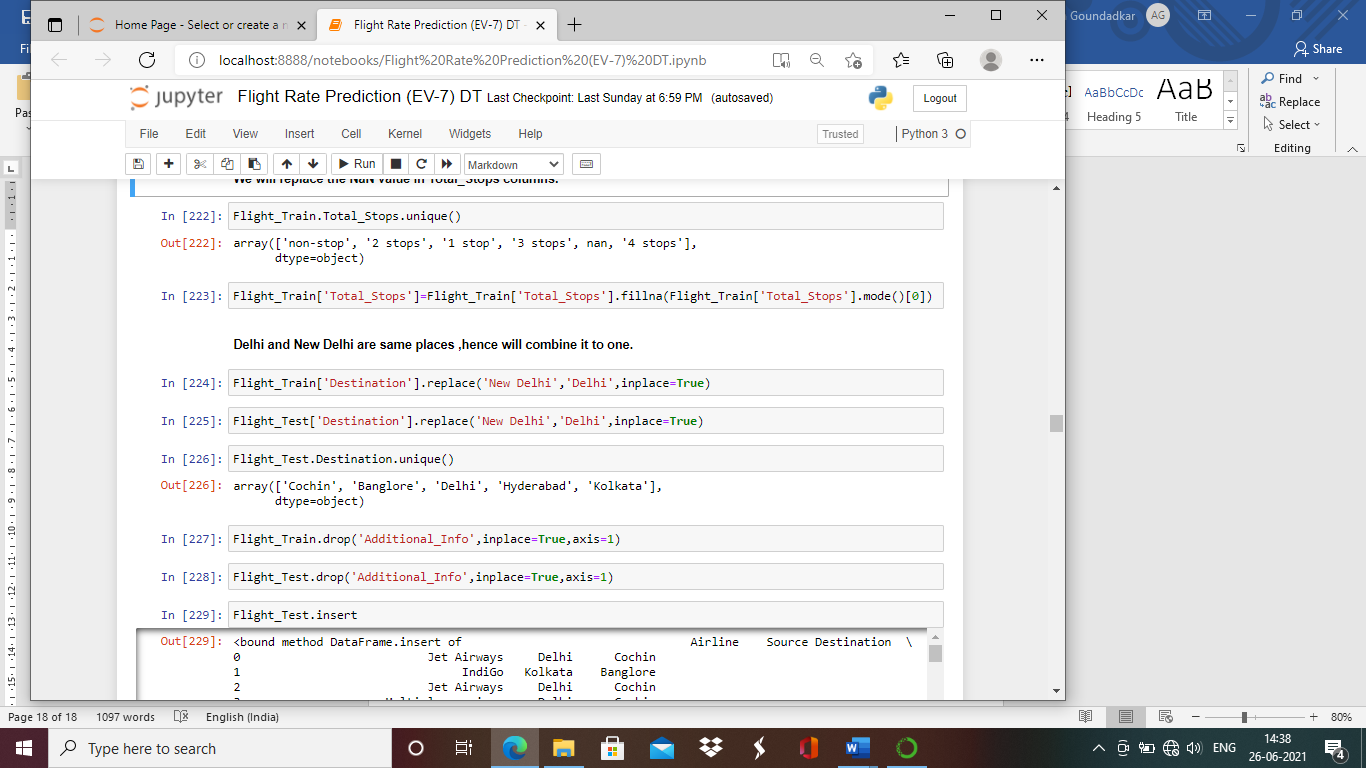
#### We will replace the NaN value in Route column.



#### We will replace the NaN value in Total\_Stops columns

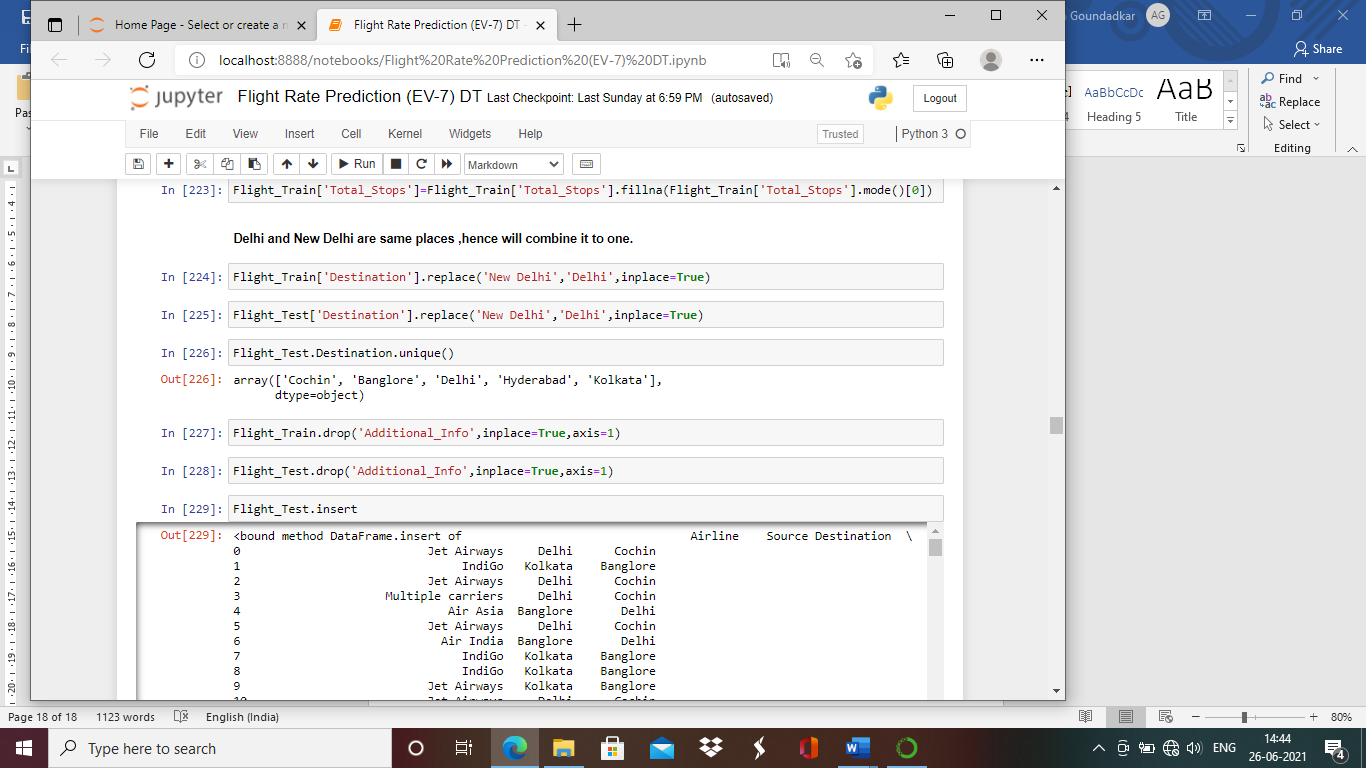


While performing EDA we had noticed there are two variables New and Delhi, which are actually same place. Hence, we will replace new Delhi with Delhi.

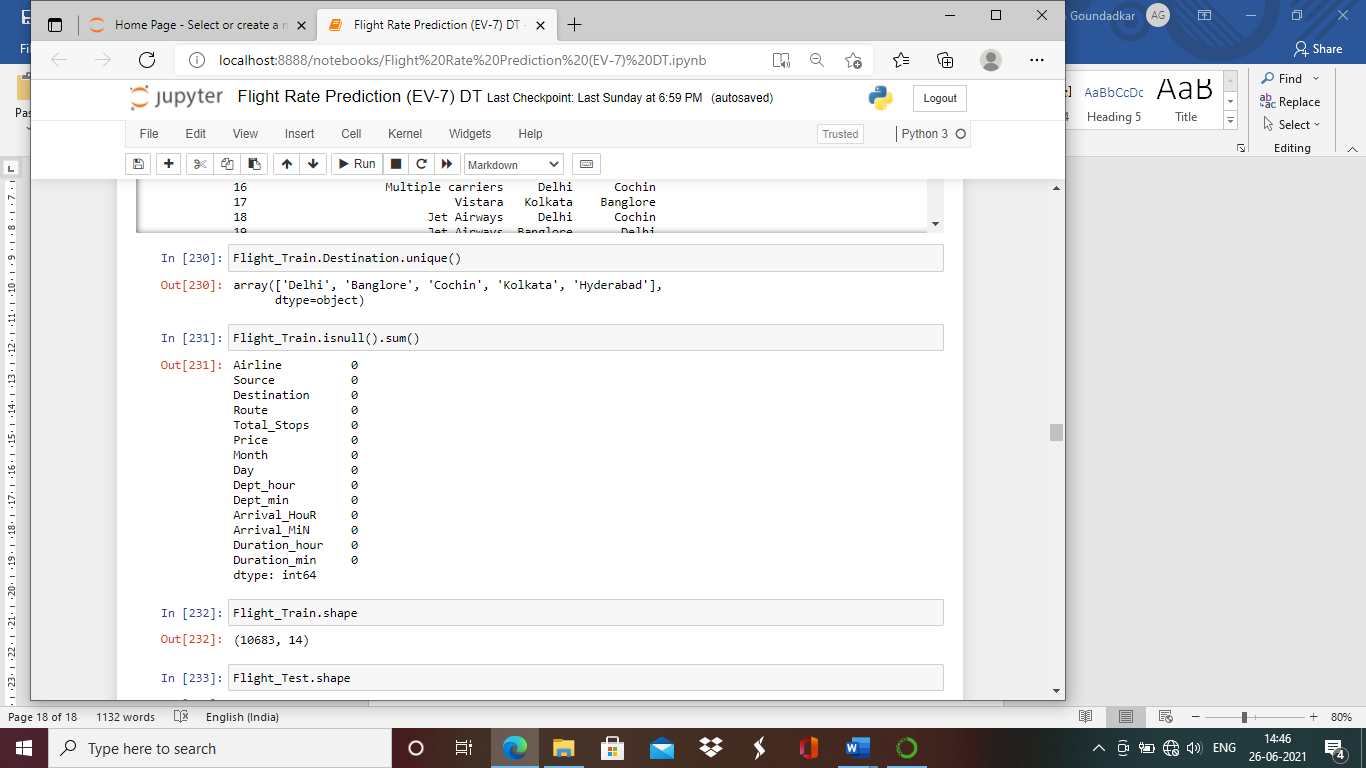


Also after doing EDA we can coclude the Additionla\_Info column has 80% No\_info data which is not helpful for model prediciton.

We will drop Additional\_Info column.



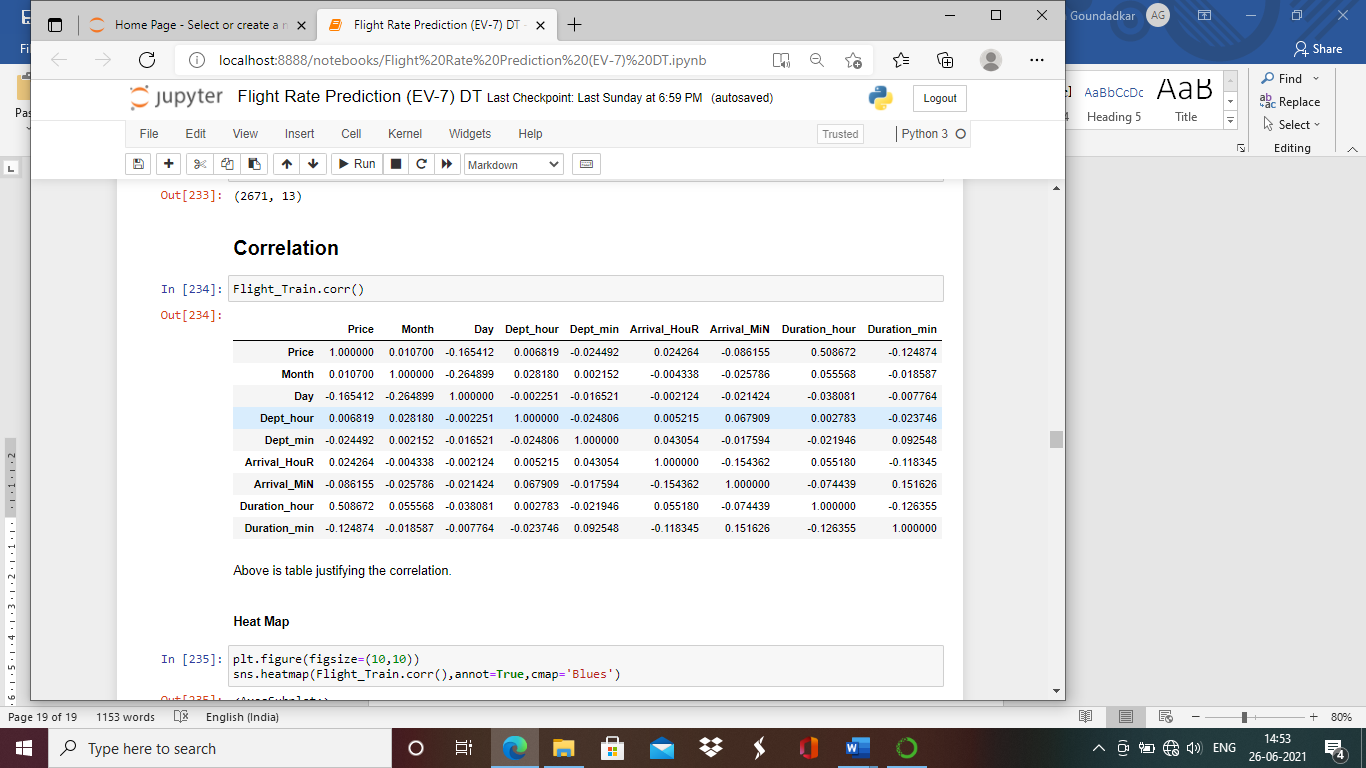
Checking if all null values are cleared from dataset.



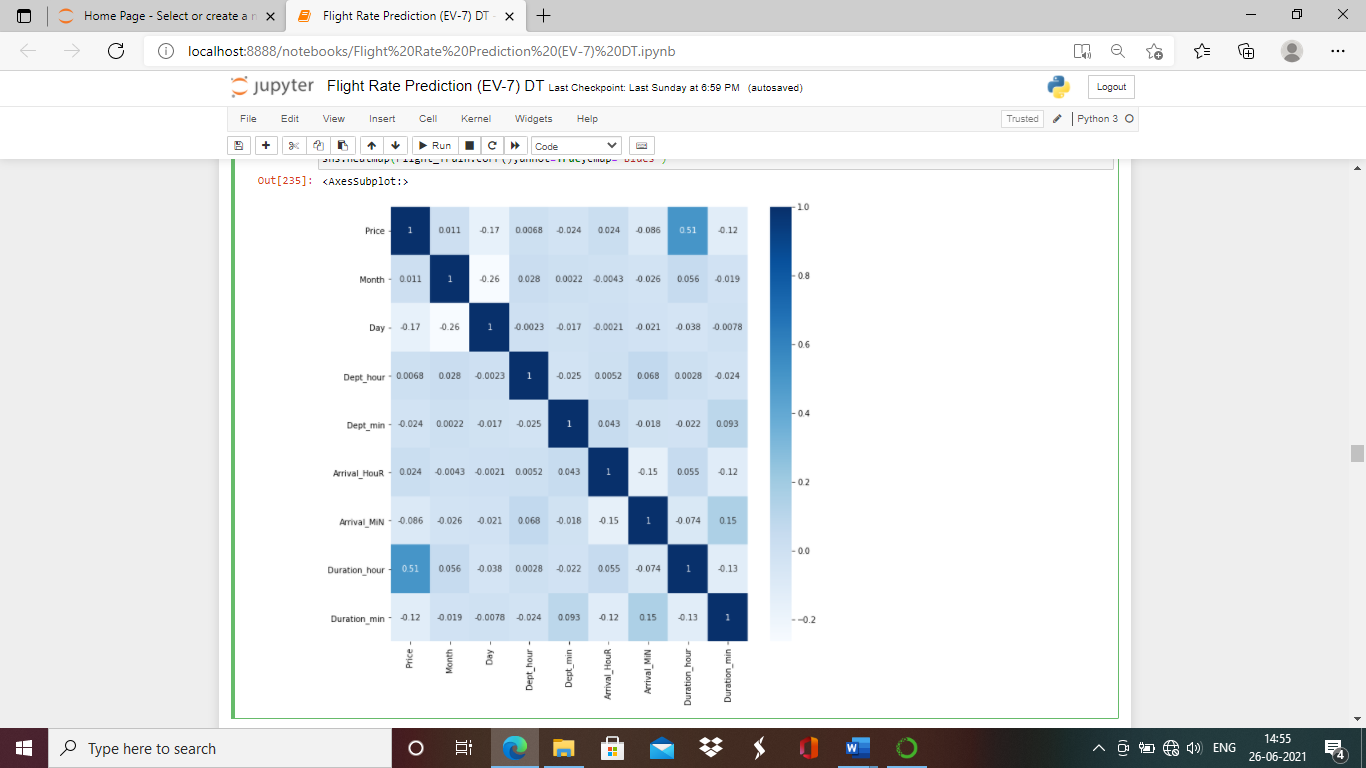
No null values in dataset.

**We will check correlation of the columns**

Below is table justifiying the correlation of the columns.



For better understanding of the correlation between columns we will plot the correlation with help of heatmap.



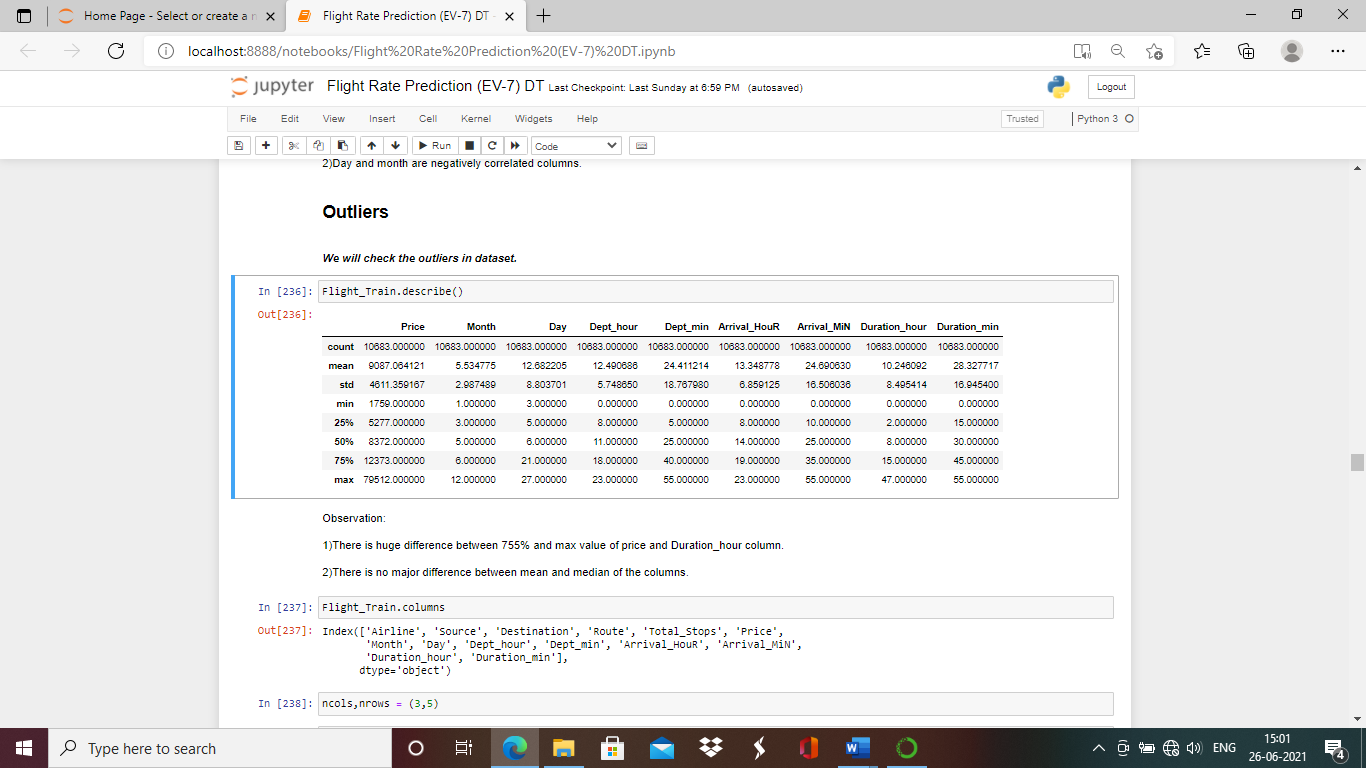
Observation:

1)Duration\_hour and price are most correlated columns in the dataset.

2)Day and month are negatively correlated columns.

**Outliers**

Let’s treat the outliers in Flight\_Train dataset. Outliers are to be treated only from numeric and continuous type of columns.

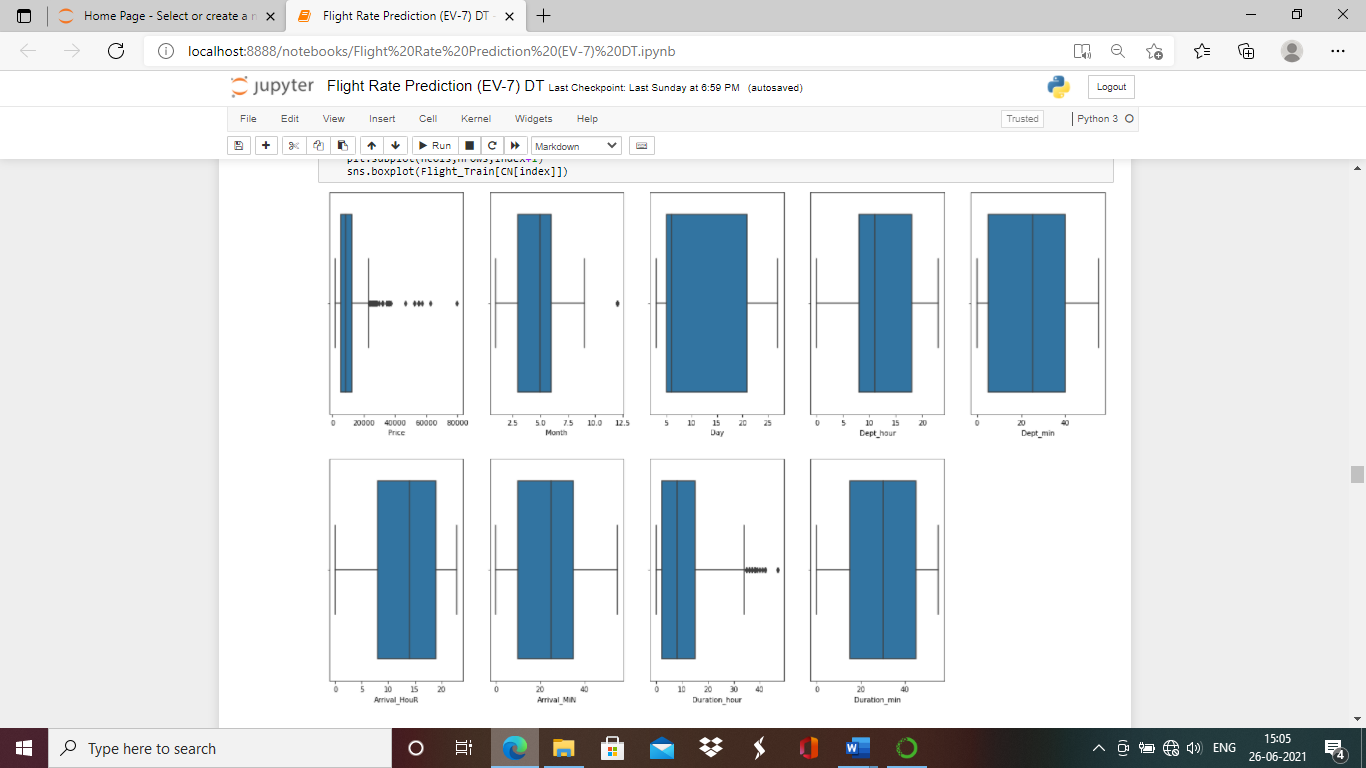


Observation:

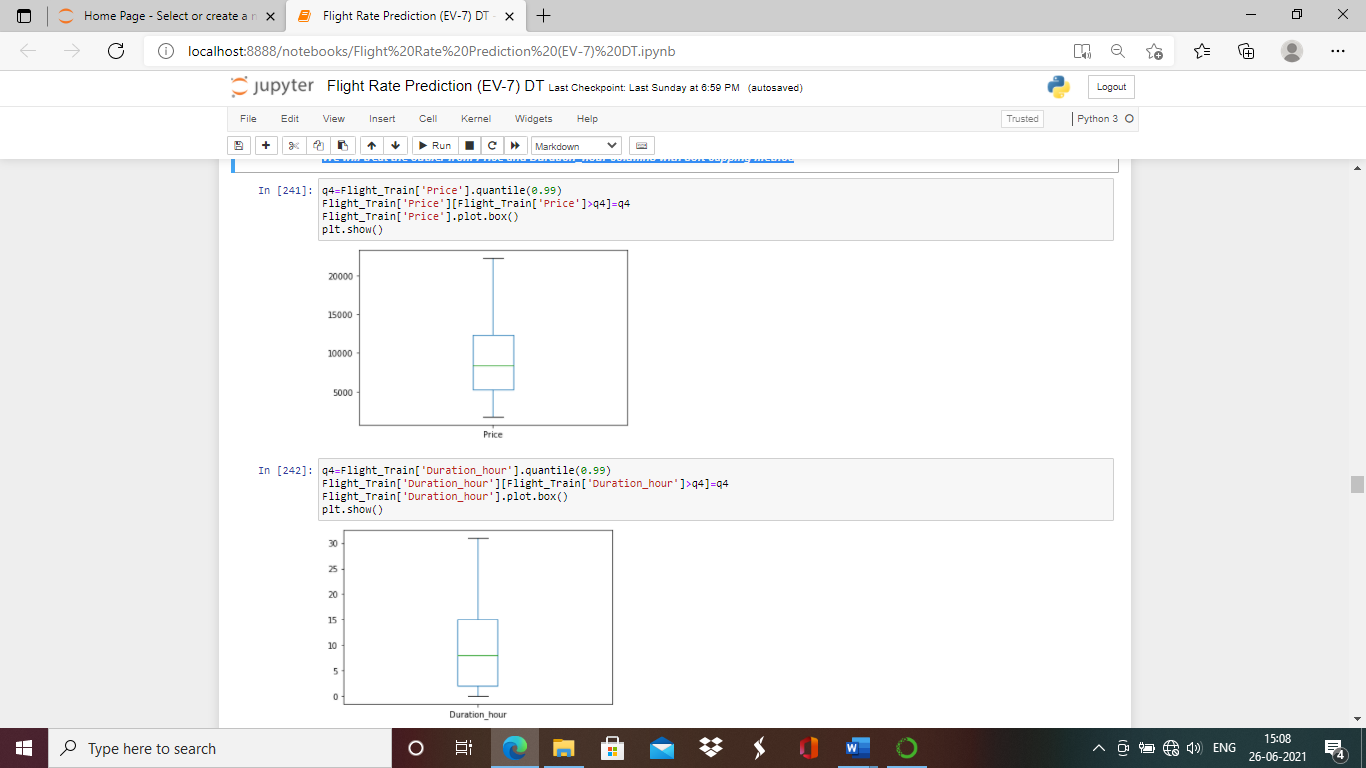
1)There is huge difference between 75% and max value of price and Duration\_hour column.

2)There is no major difference between mean and median of the columns.

**Let’s plot box plot for better understanding of outliers.**



We will treat the outlier from Price and Duration\_hour columns with soft capping method.



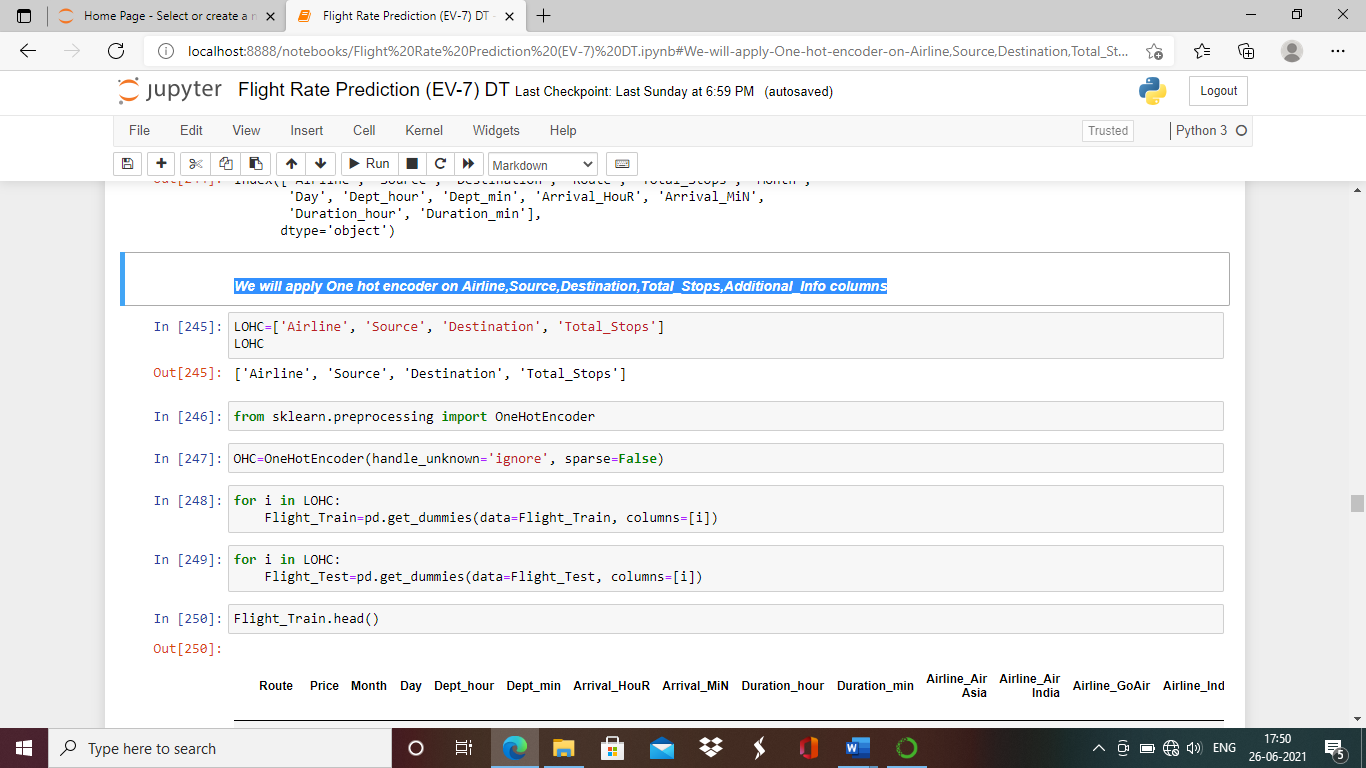
Outlier from both the columns are removed.

**Encoding**

All object data type of columns to be converted from alphabetical variable to numerical numbers.

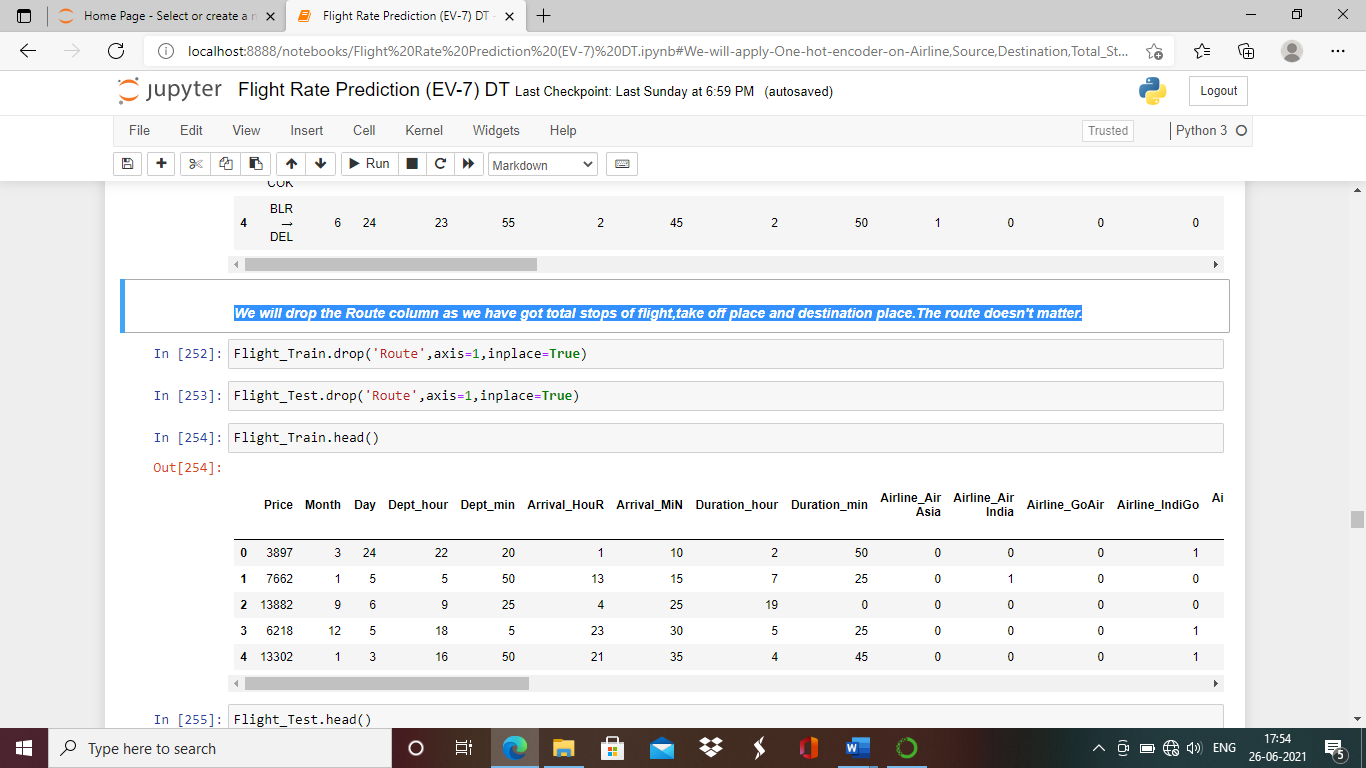
We will perform that with encoding method.

We will apply One hot encoder on Airline,Source,Destination,Total\_Stops,Additional\_Info columns.



We will drop the Route column as we have got total stops of flight.

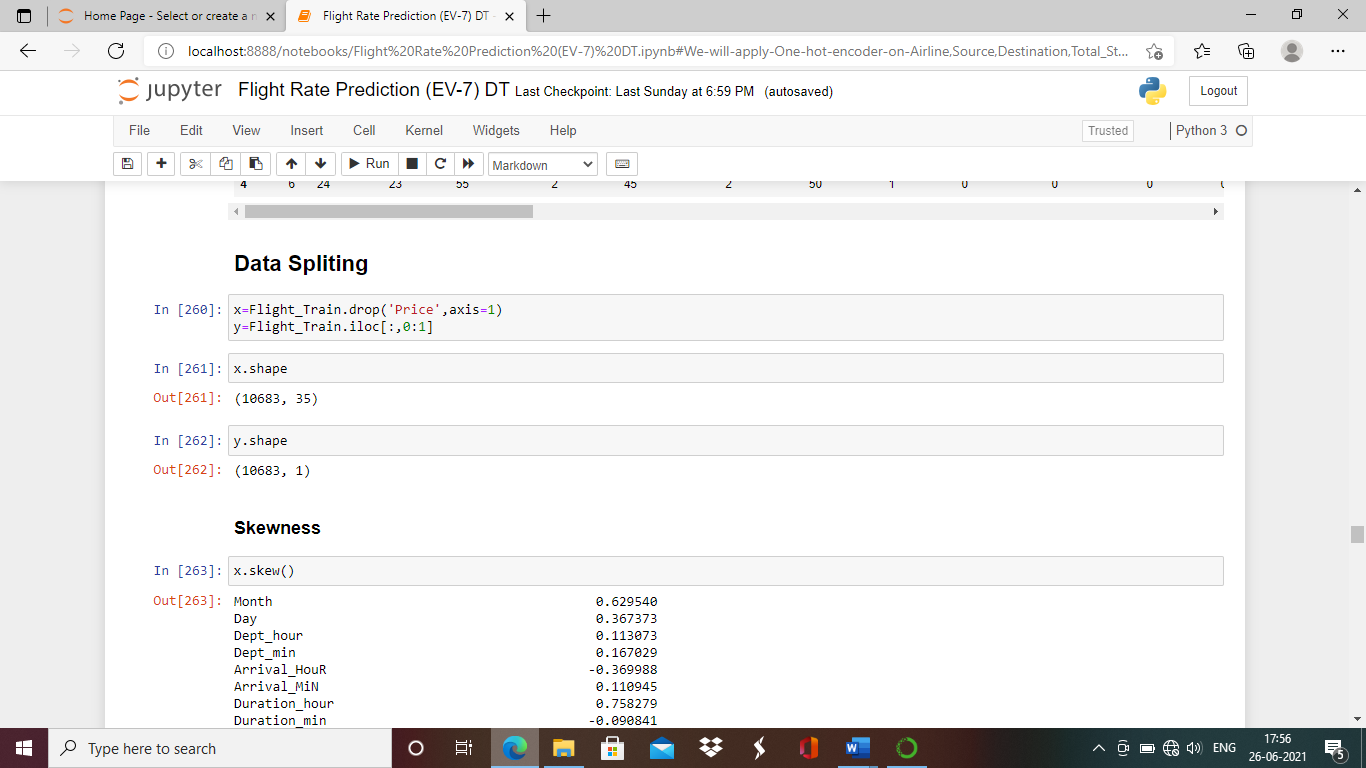
Take off place and Destination place makes us understand from where passenger is going to take off and where passenger is going to land. The route o cities in the journey where flight is going doesn't matter.



**Data Splitting.**

We will split the data in X and Y.

X set contains all the independent columns and Y set is dependent column.

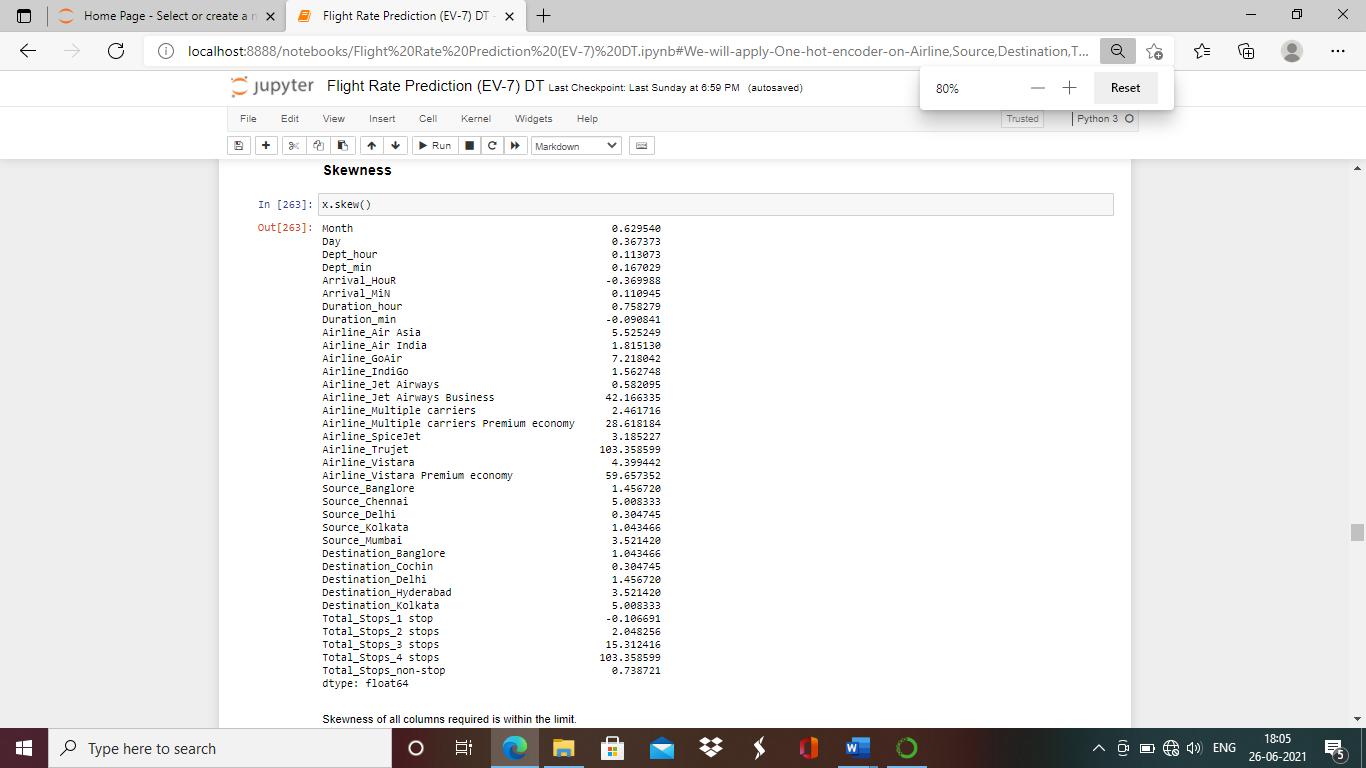


**Skewness**

Skewness measures the deviation of a random variable’s given distribution from the normal distribution, which is symmetrical on both sides. Models don’t perform better when data is skewed and the reason behind this is that the tapering ends or the tail region of the skewed data distributions are basically the outliers in the data and it is known to us that outliers can severely damage the performance of a statistical model the best example of this being regression-models which show very bad results when trained over skewed data. Hence, before running X set the we should check the skewness of each numerical column and not categorized column.

Skew value of column should always be between -0.5 to 0.5.

We will check the skewness of each column now.



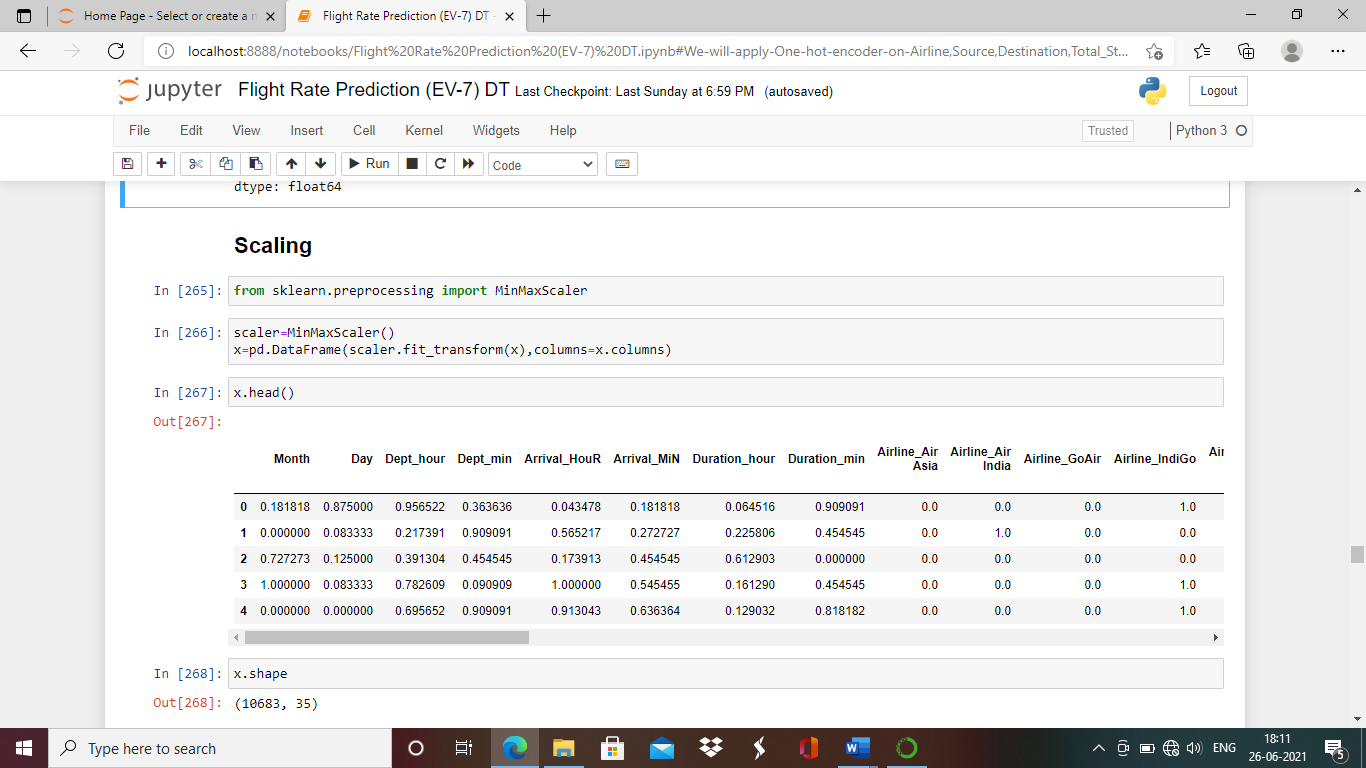
Skewness of all columns required is within the limit.

**Scaling**

Scaling is the final treatment we need to do before executing the data in model for prediciotn.

Scaling is performed when there is vast difference in range of data within column.There are various methods we can perform feature scaling. (i.e Minmax scaler ,standard scaler,Robust scaler etc.

We are using Min max scaler for data set as standard deviation is small.



Data is ready for model execution after treating the missing values, encoding, removing skewness and scaling.

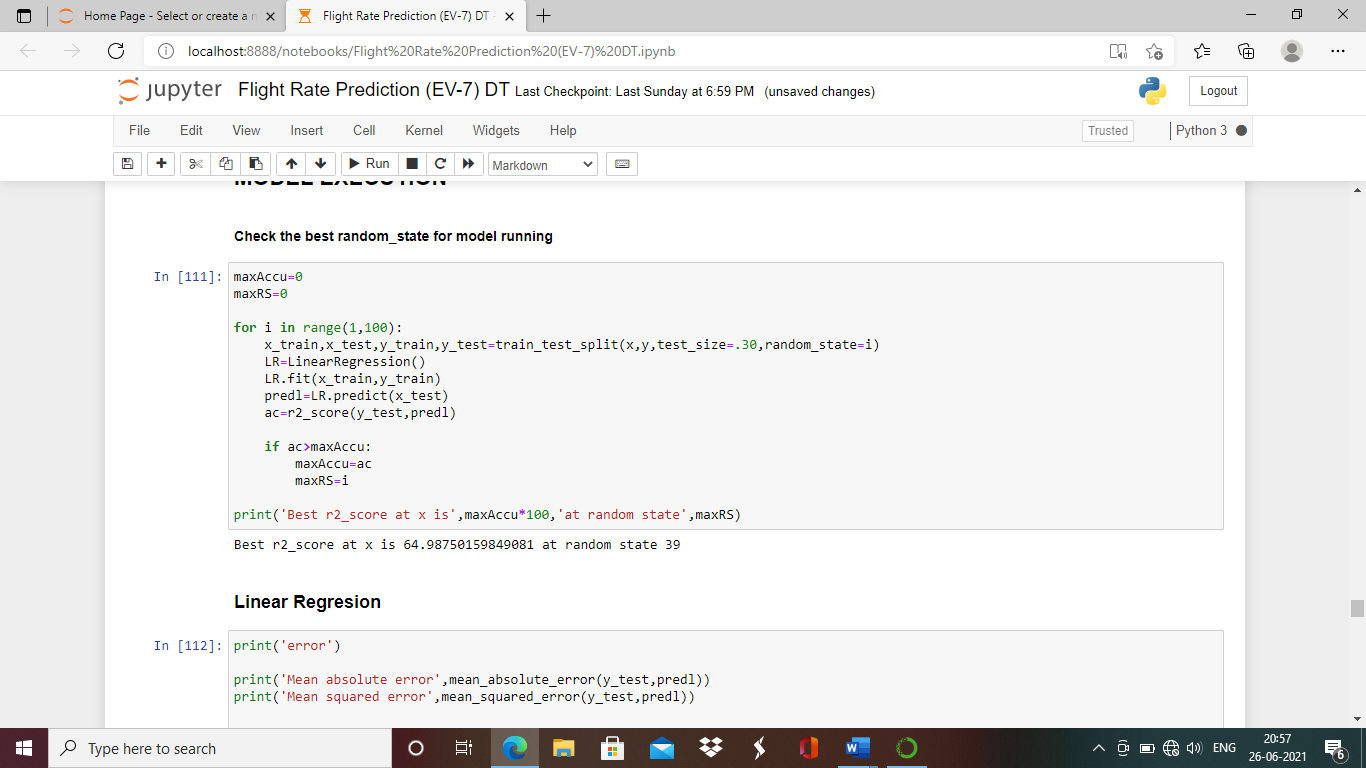
**Building Machine Learning Models**

For model execution we have to split the data in train and test data which plays a very critical role in model performance, as imbalanced fitting of test and train data can impact model performance. We should always try keeping training data close to 70% for better performance.

This being a regression problem we can use various models like linear regression, Random Forest regressor, XgBoost regressor, regularization like (Lasso, Ridge and Elastic) etc.

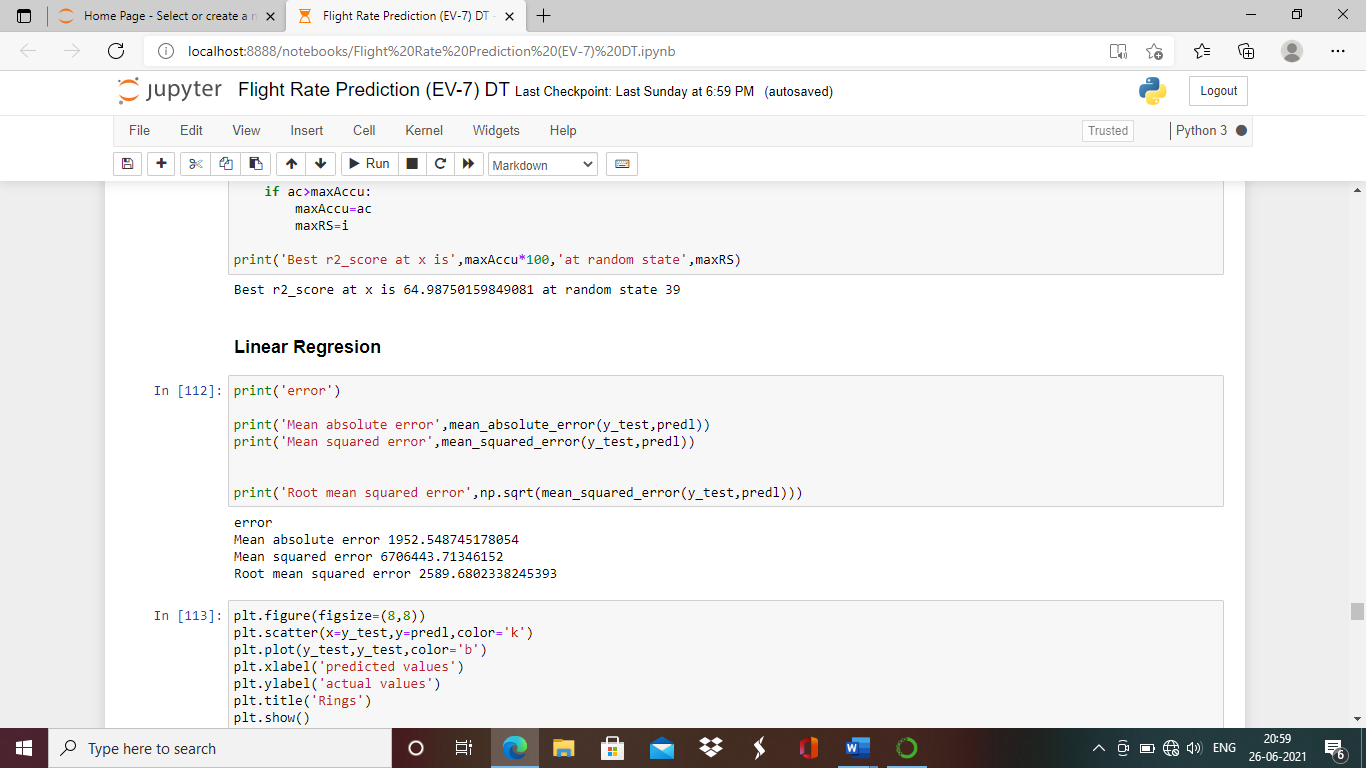
I have tried running all the above-mentioned model and also have also applied hyper parameter tunning to improve the performance of each model. After running all models I have to come to conclusion at random forest regressor model is best performing model ,we will check the performance and details further.

Initially we will check the best random state for model performance using logistic regression.

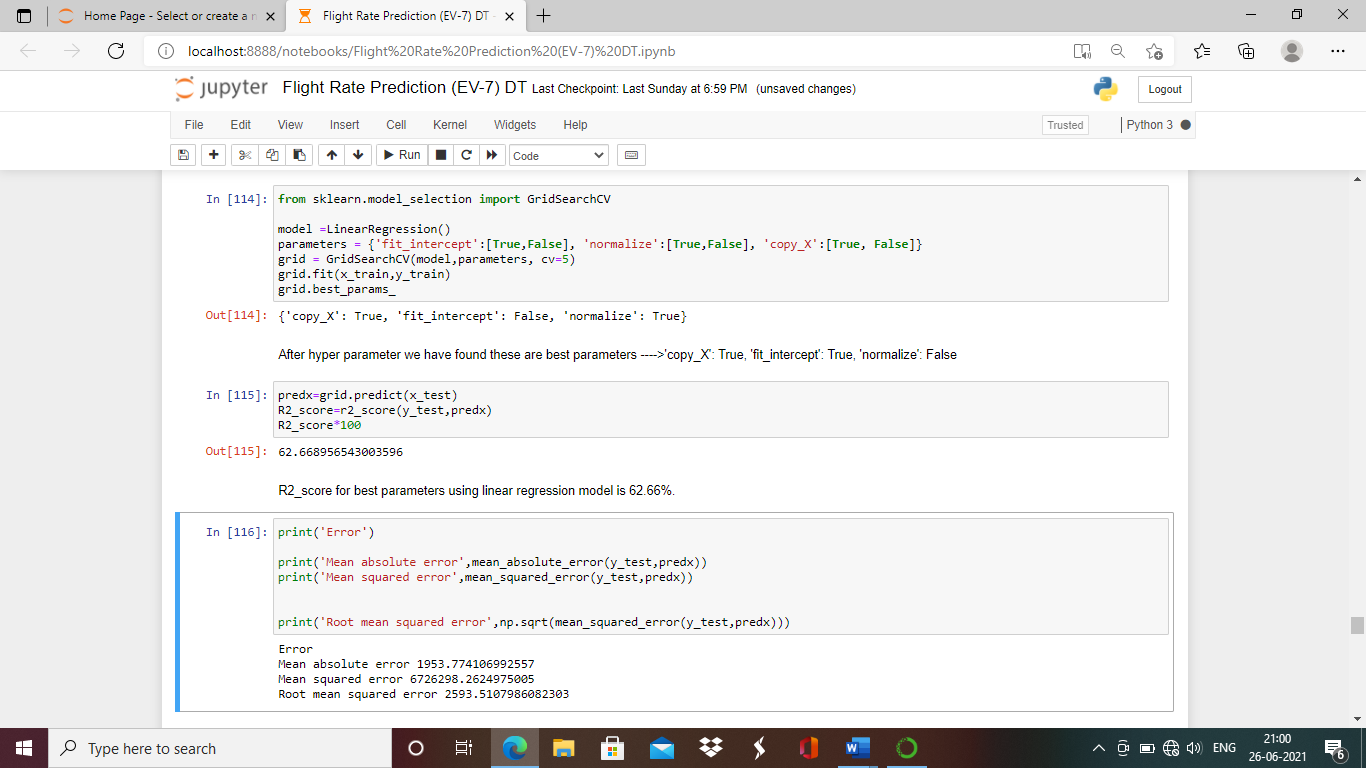


The best r2\_Score is 64.98 at random state 39.

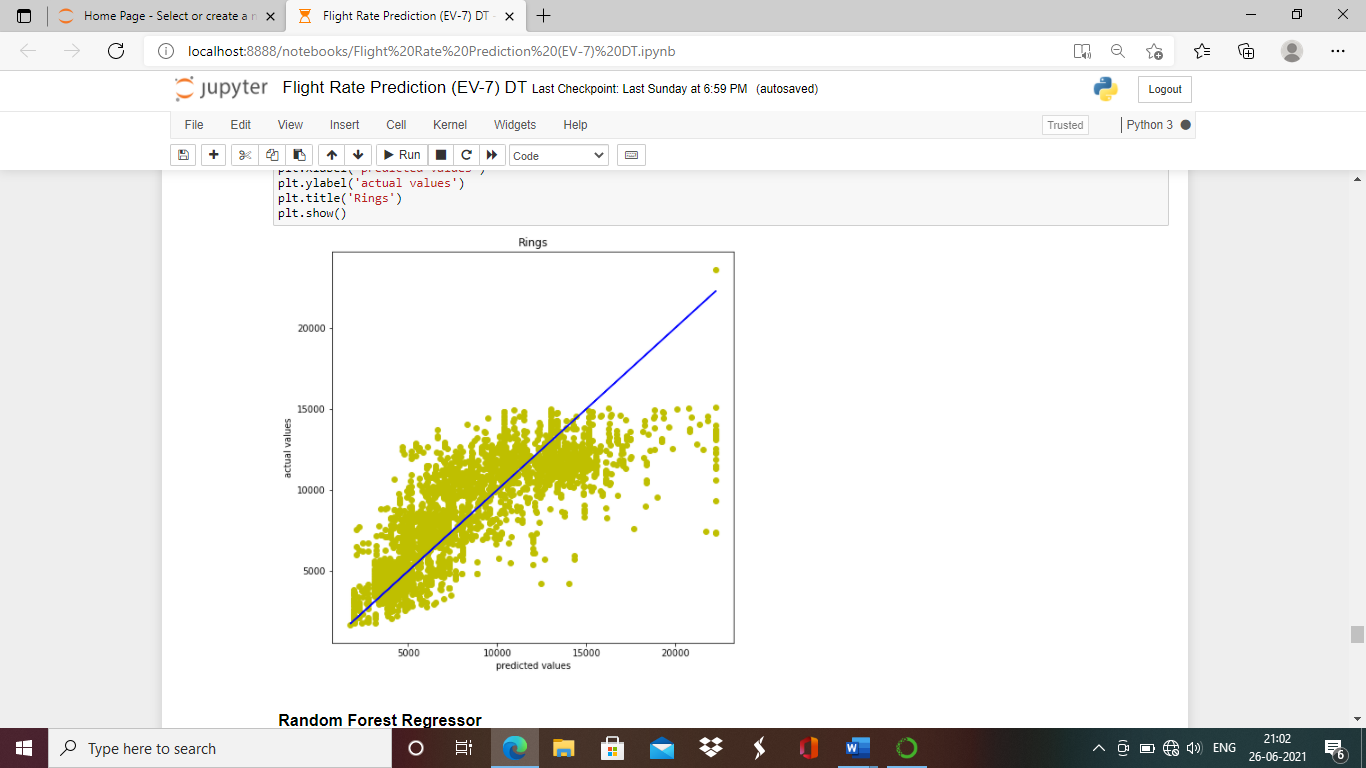
Linaer Regression



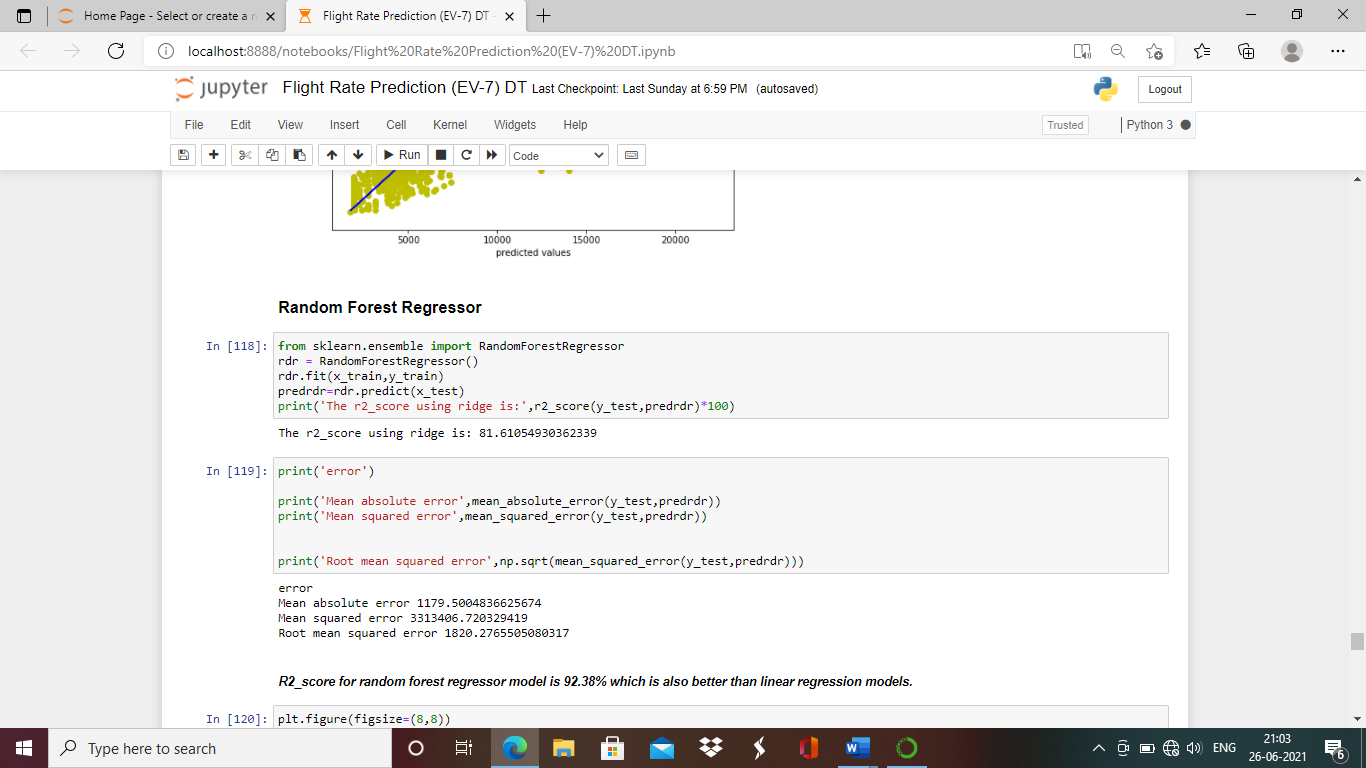
### Hyper Parameter Tunning for Linear Regression Model



We will check scatter plot of predicted values v/s acctual values



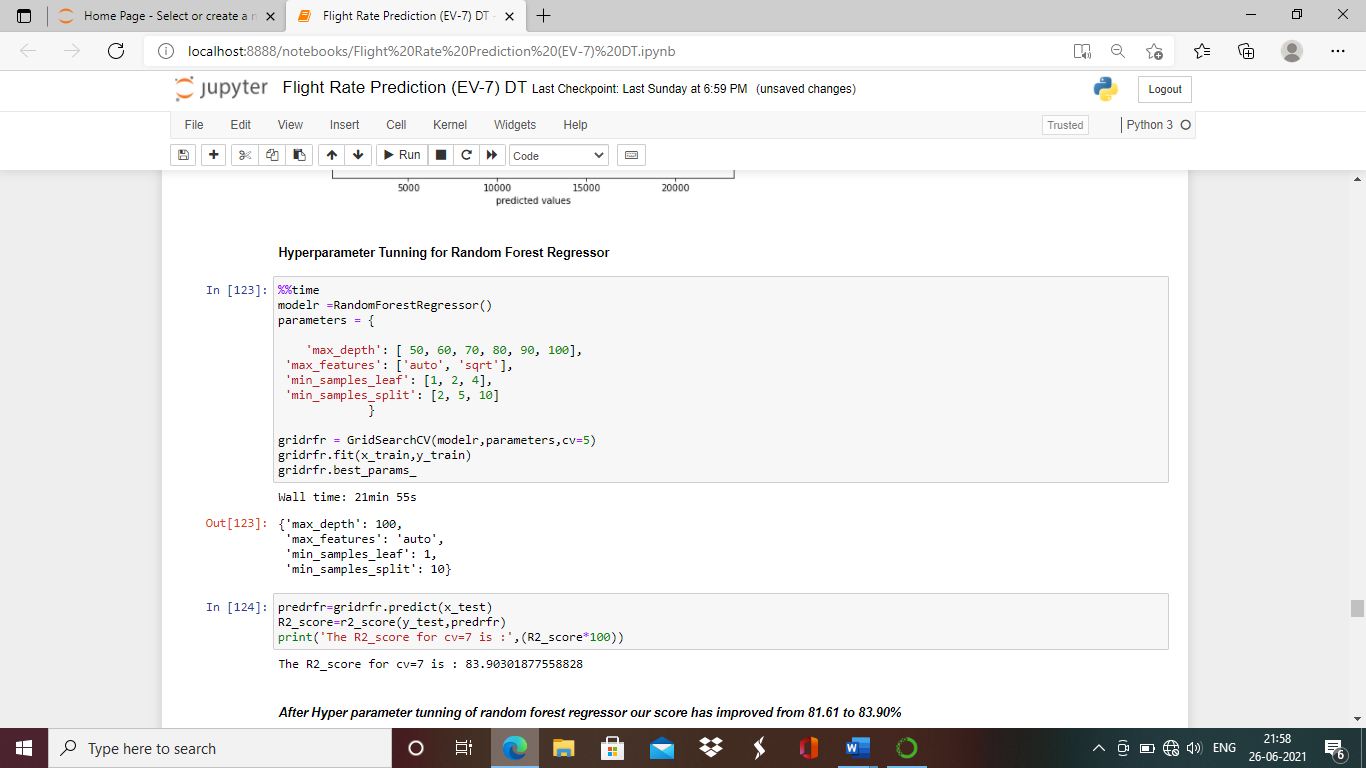
**Random Forest Regressor**



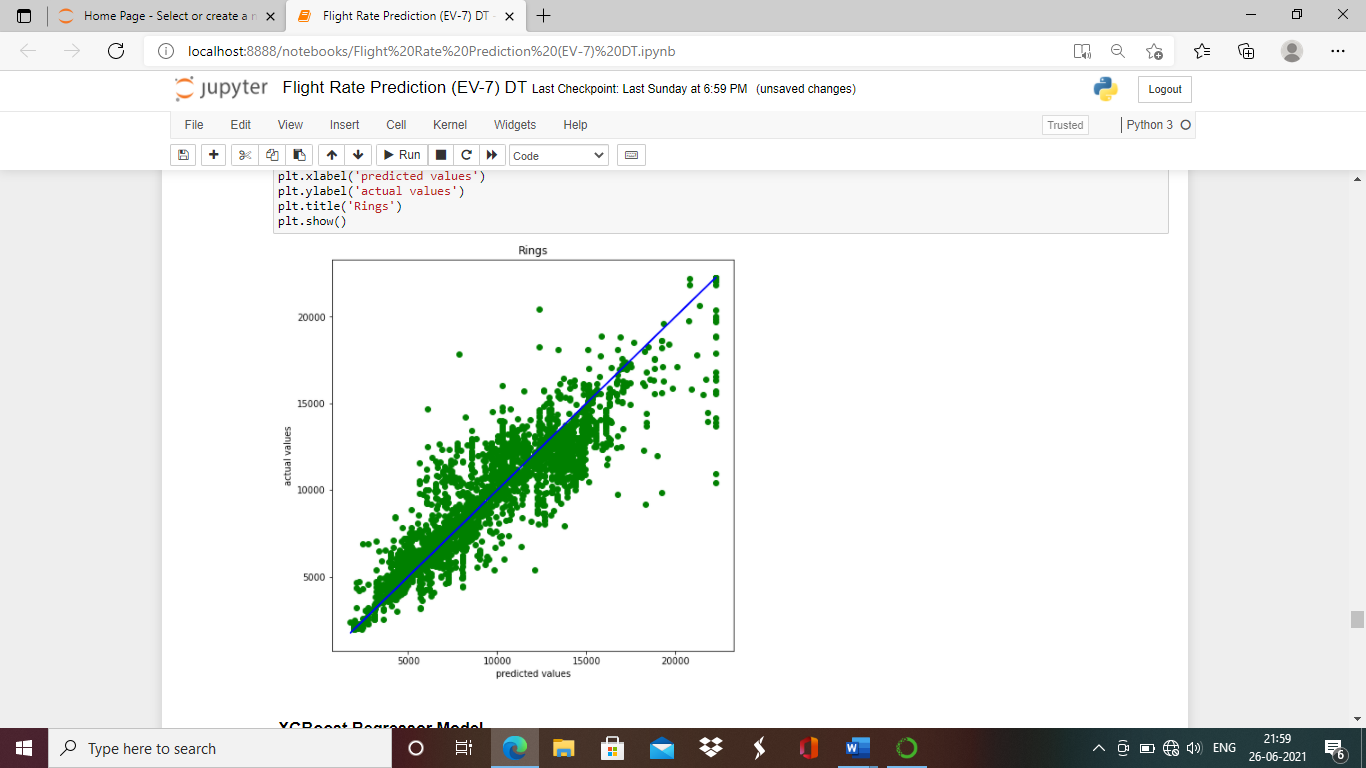
R2\_score without hyper tunning is 81.61%.

Now let’s apply hyper parameter tunning

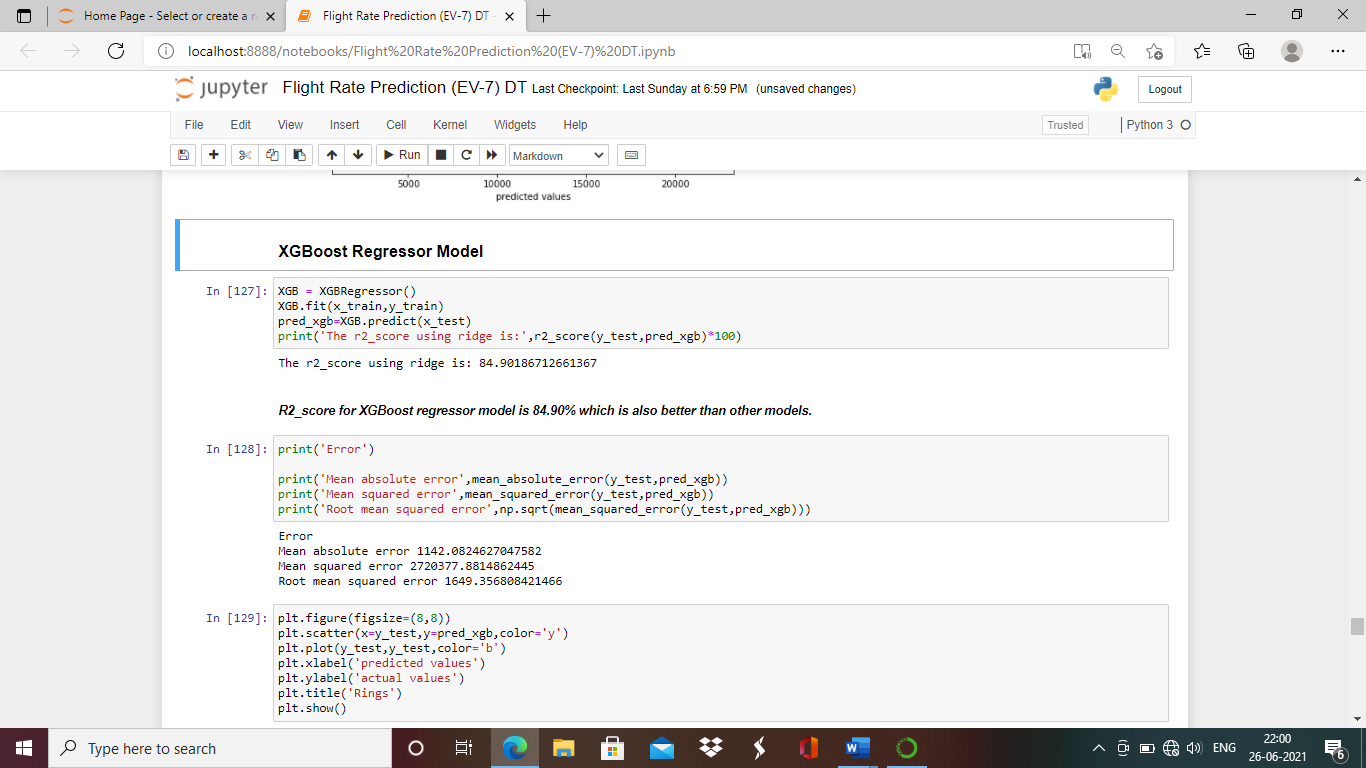
#### Hyperparameter Tunning for Random Forest Regressor



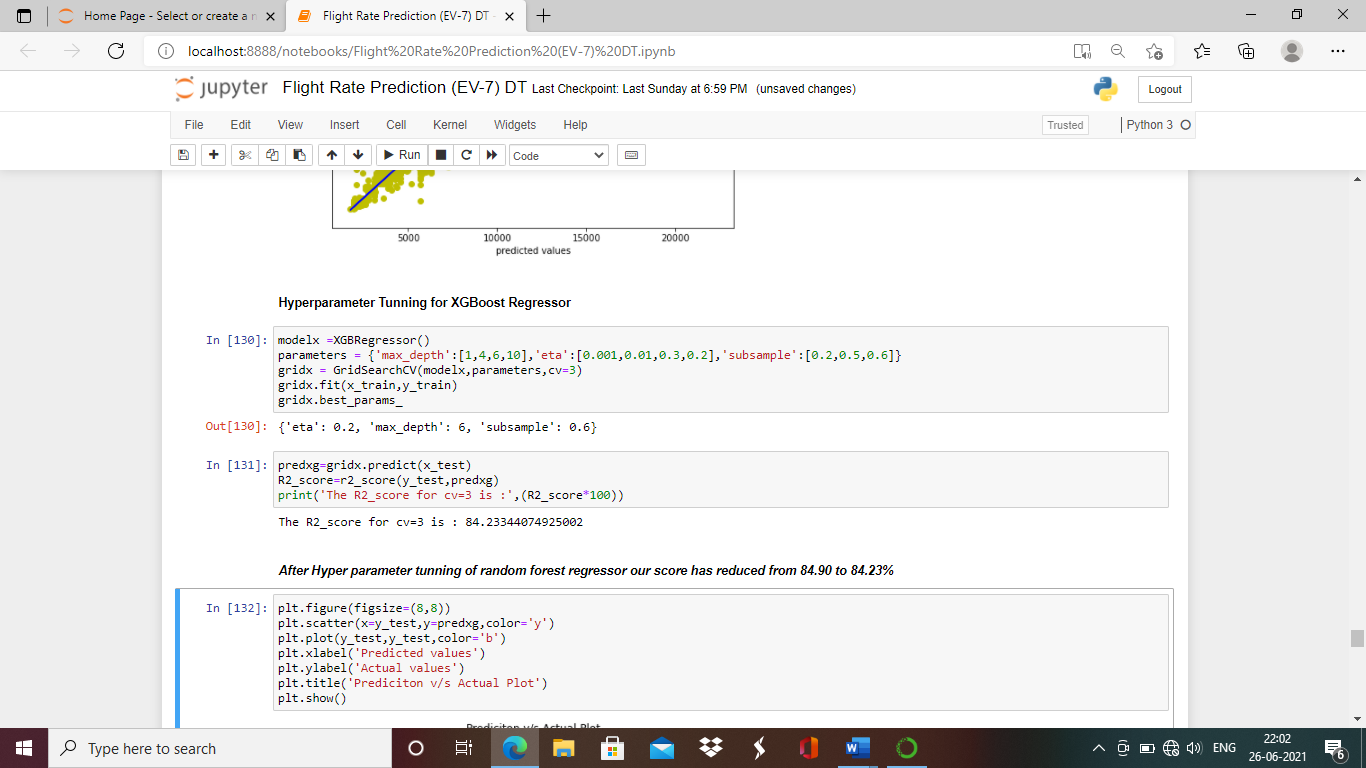
We will check scatter plot of predicted values v/s acctual values

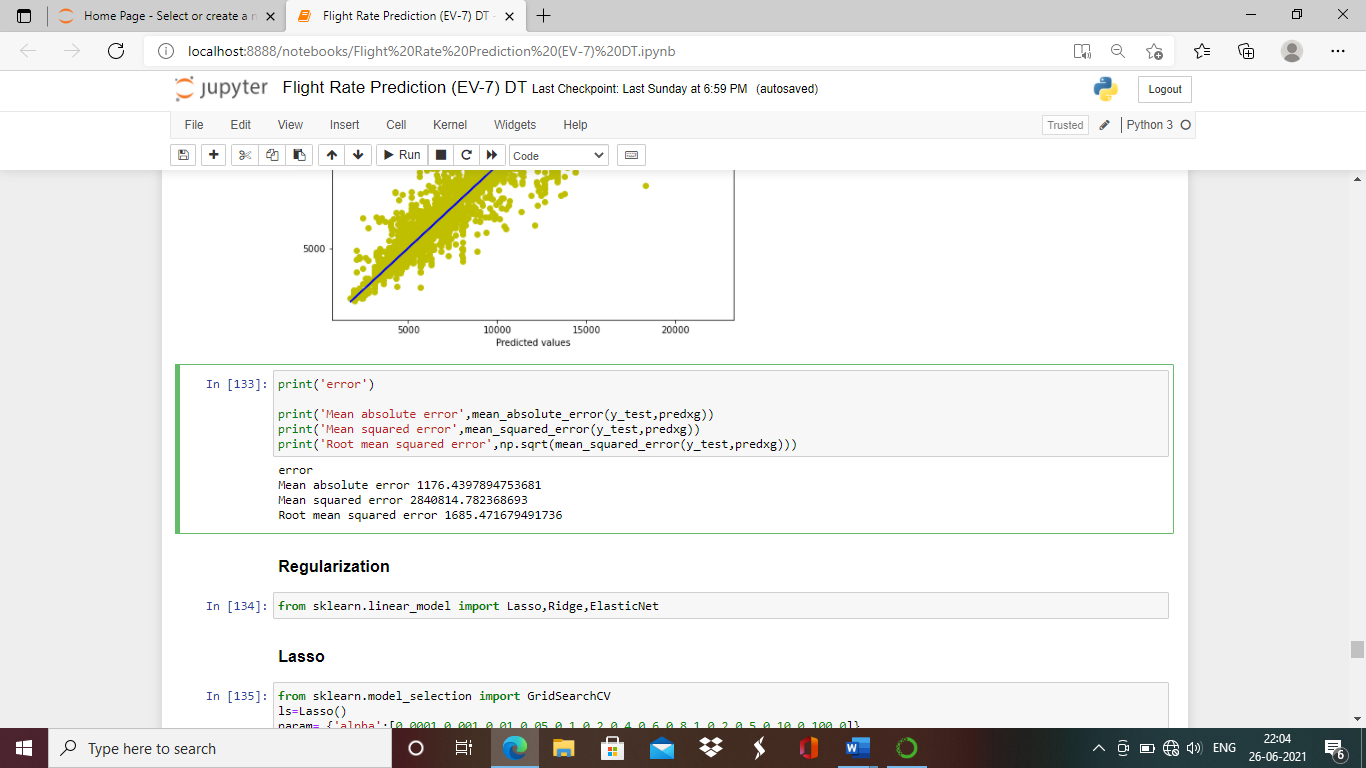


### **XgBoost Regressor Model**



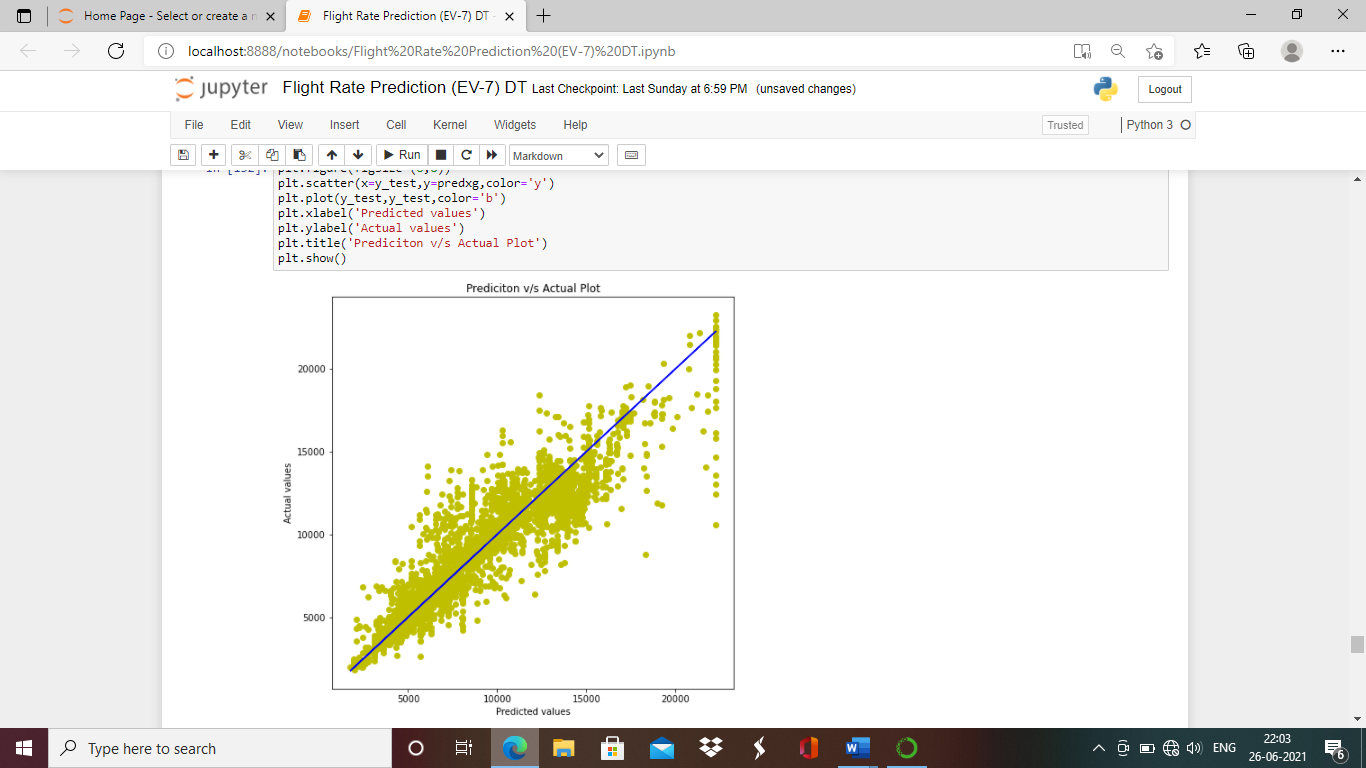
#### Hyperparameter Tunning for XgBoost Regressor





***After Hyper parameter tunning of XgBoost regressor our score has reduced from 84.90 to 84.23%***

We will check scatter plot of predicted values v/s acctual values



#### After running the different models and applying hyper tunning we can understand that XgBoost Regressor model is the best model supporting our dataset, with R2\_score of 84.23% which is best from all the models.

Even we get the least root mean squared error and mean squared error for this model

(MAE: 1176.43 and RMSE: 1685.4)

**Concluding Remarks**

In Flight fare prediction dataset, we have introduced the whole pipeline of an end-to-end machine learning. We have described the dataset, and explored all the variables in the dataset. Checked if there are any missing values. Exploratory data analysis is performed using univariate, bivariate and multivariate columns for better understanding of data.

In this dataset problem featuring play’s critical role. We had to handle many categorical columns. Data featuring is performed by treating the missing the values, checking correlation between columns, encoding all the alphabetical columns to numerical. Outliers are cleared in dataset with quantile method and made ready to run the model. We have run various models to find the best model and found flight fare prediction has used XgBoost regressor model as best model .MAE and RMSE are very important values to be checked for verifying the performance of model in this dataset and we have found Low MAE and RMSE for XgBoost model.

Thank You!!!