**Article**

**On**

**Flight Price Prediction**

**Author – Sidharth Das**

1. Problem Definition :

This article is on ‘prediction on flight rate’. Flight ticket expenses can be something hard to wager, Now a days we would see a rate, test out the charge of the same flight tomorrow, it will likely be another story. We'd have regularly heard tourists announcing that flight ticket charges are so unpredictable.

Let’s takes you through each step-in element a way to predict the flight rate and enables to understand the entire ml model building system.

**Dataset Information :**

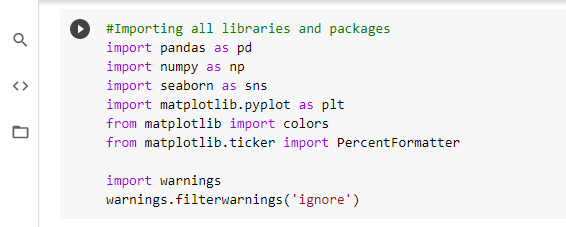
* **Airline: The name of the airline.**
* **Date\_of\_Journey : The date of the journey**
* **Source: The source from which the service begins.**
* **Destination: The destination where the service ends.**
* **Route: The route taken by the flight to reach the destination.**
* **Dep\_Time: The time when 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**

1. Data Analysis :

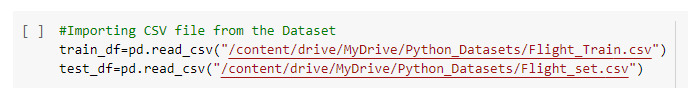
**Size of training set: 10683 records**

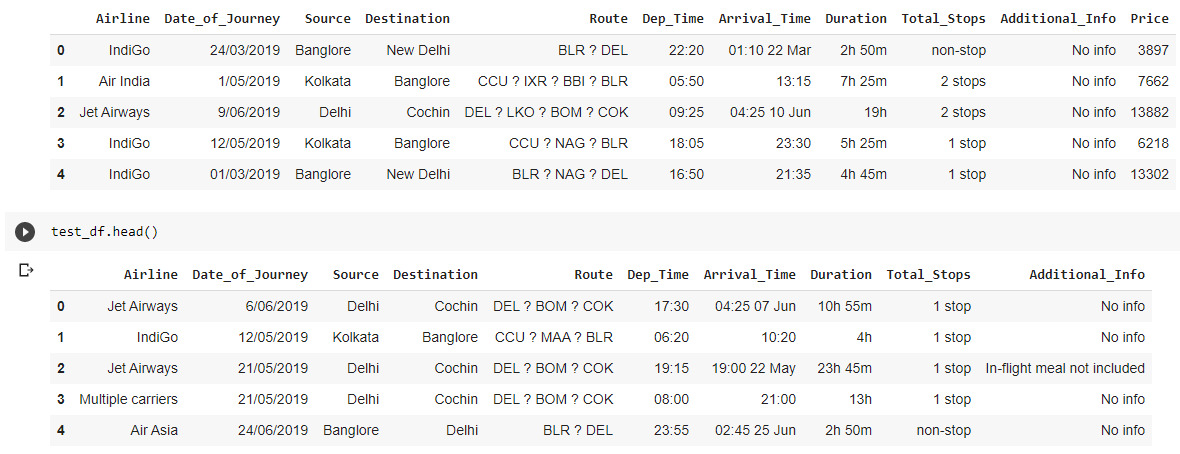
**Size of test set: 2671 records**

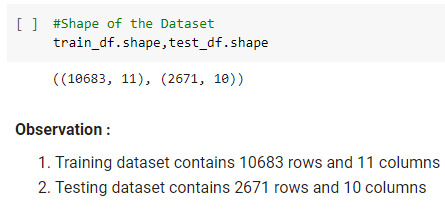
**Libraries :**

****

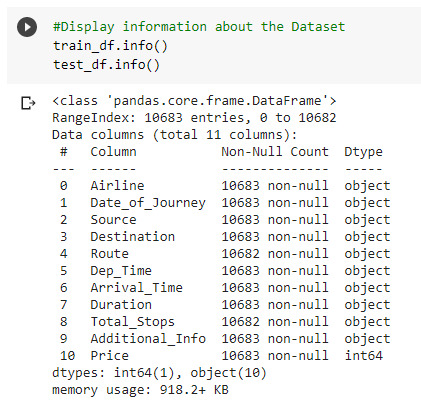
**Importing the Dataset :**

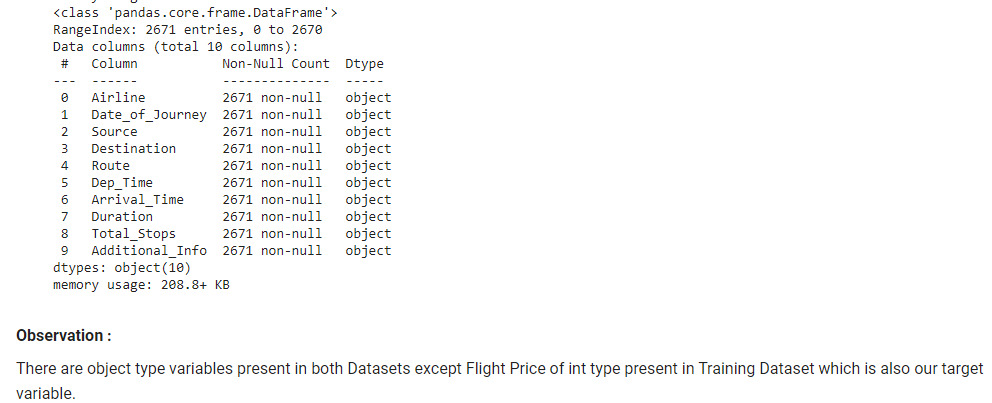
****

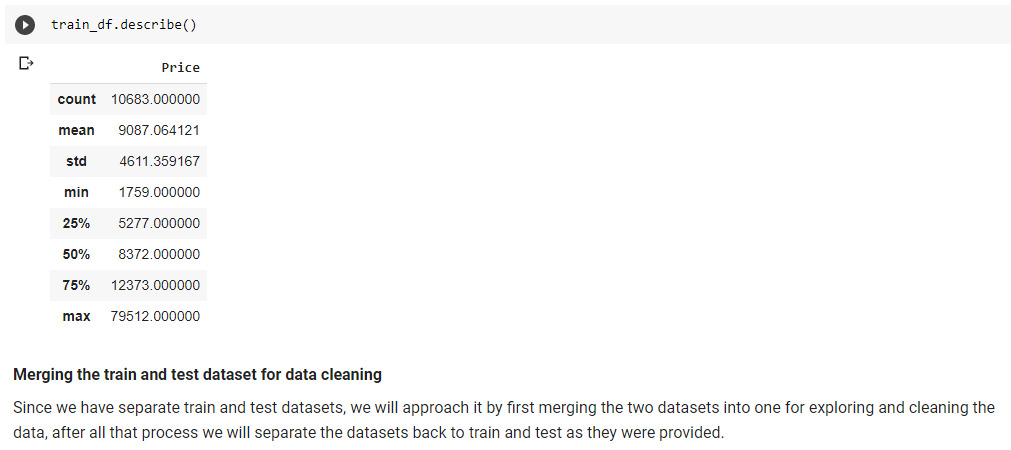
****

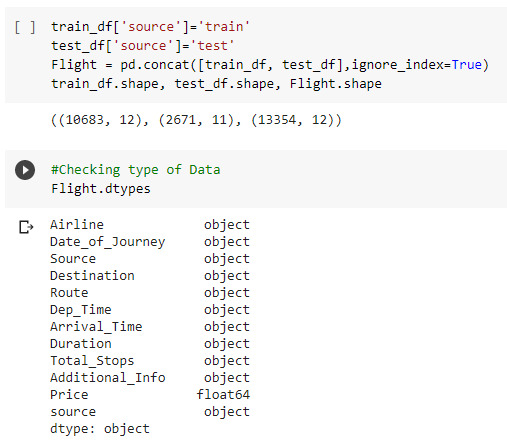
****

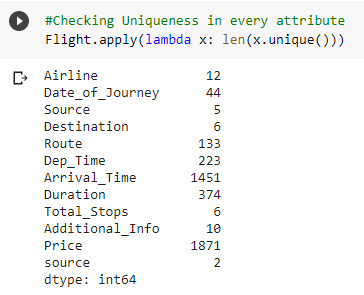
**Data Wrangling :**

****

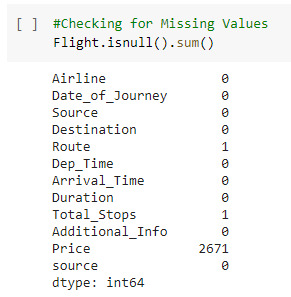
****

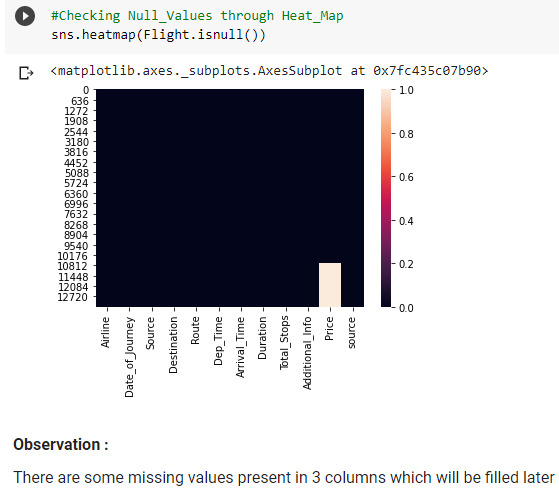






Checking for Missing Values :

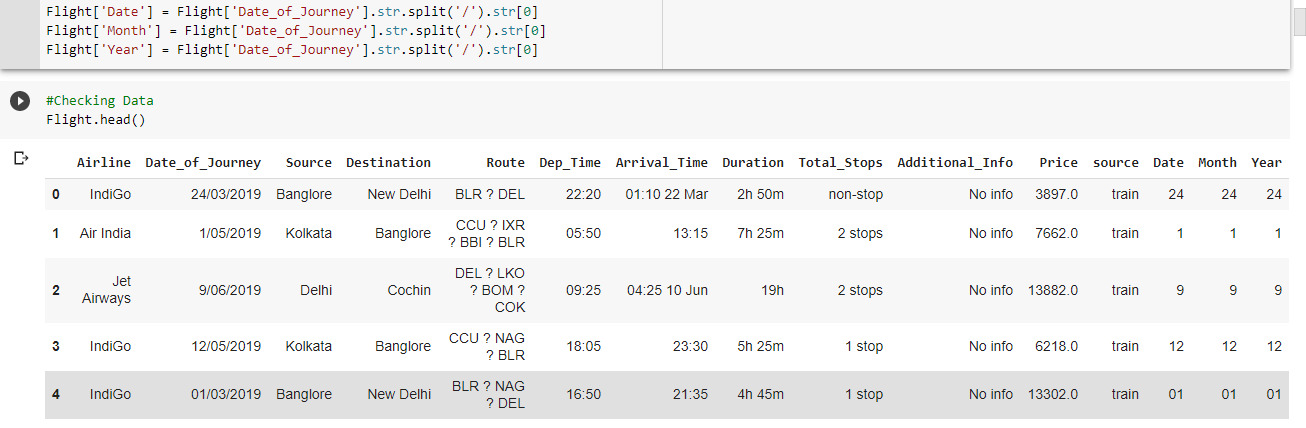




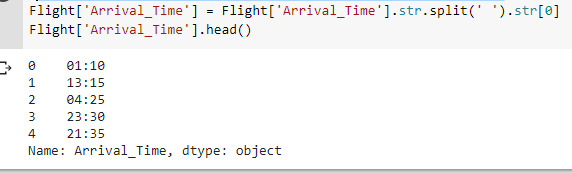
Feature Engineering of Data :

**On this step we will se on the facts set and do a little transformation like creating extraordinary packing containers of precise column’s, easy the messy statistics in order that it may be utilized in our Machine Learning model. This step is very important because for high prediction score, we want to continuously make modifications in it.**

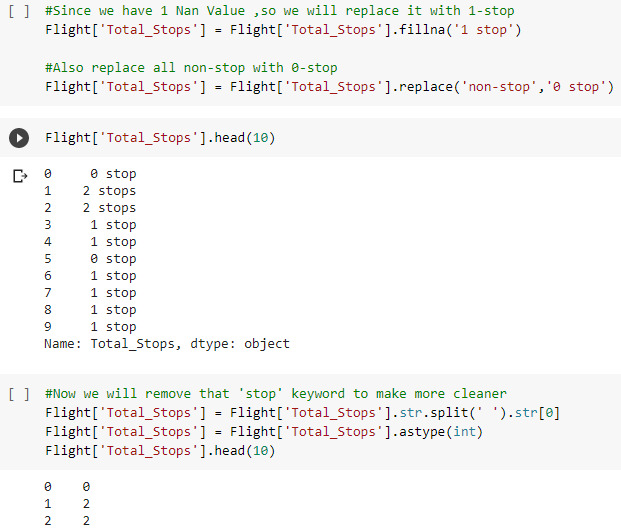
**Split Date\_of\_Journey column values with independent years , months and date .**



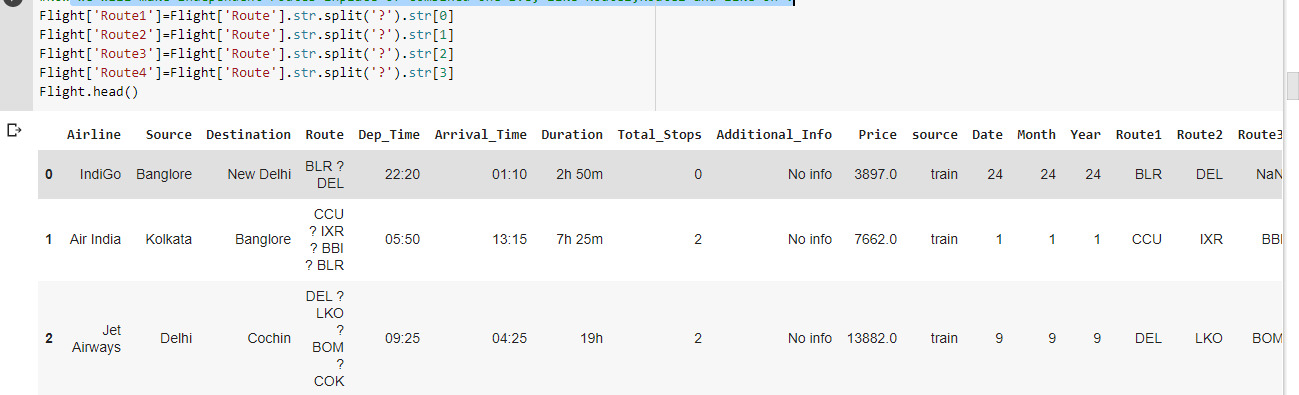
**Next we will remove some dates written in Arrival\_Time column**

****

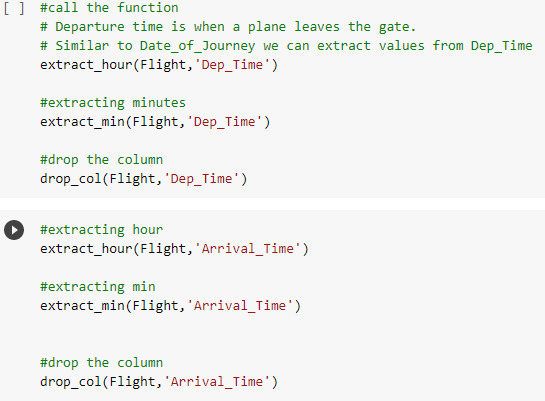
**Feature Engineering on Total Stops :**

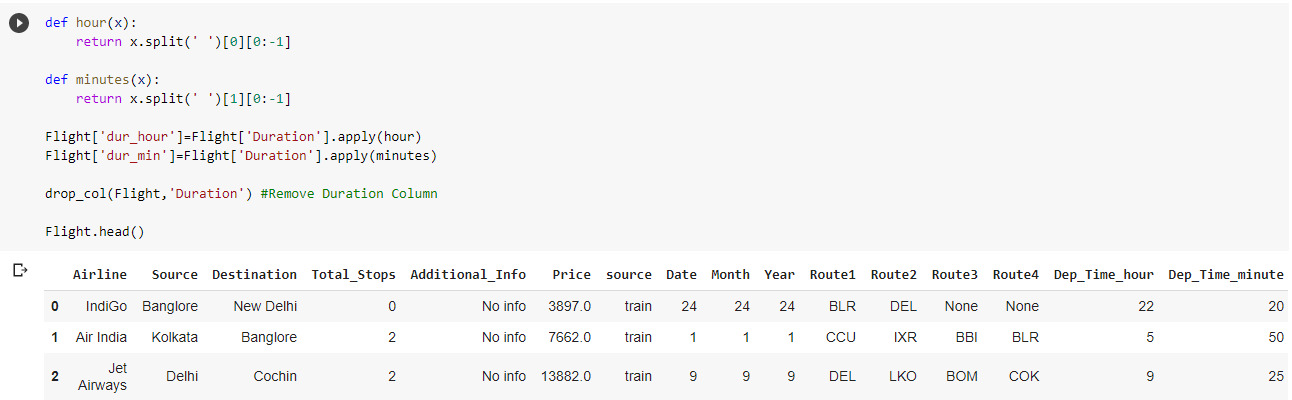
****

**We will make independent routes inplace of combined one i.e, like Route1,Route2 and like on .**



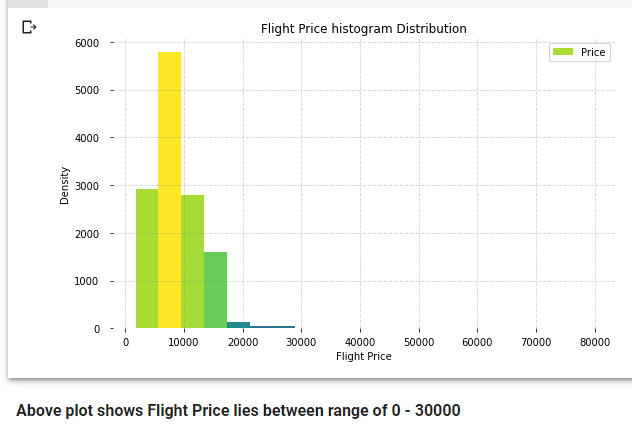
**Extracting time, hour for departure\_time and arrival\_time .**



****

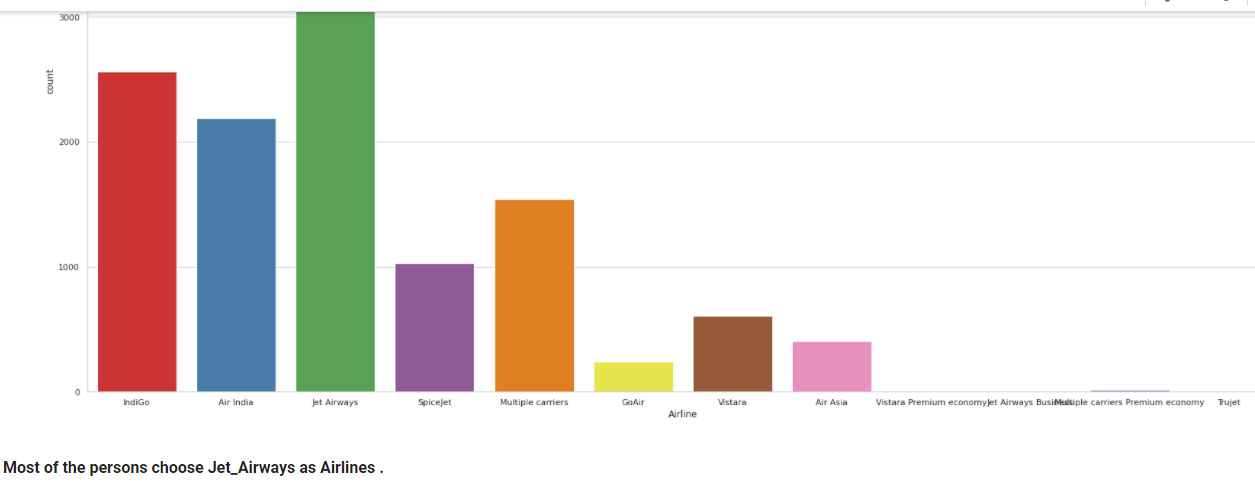
1. EDA concluding Remarks :

**'Flight Price histogram Distribution'**

****

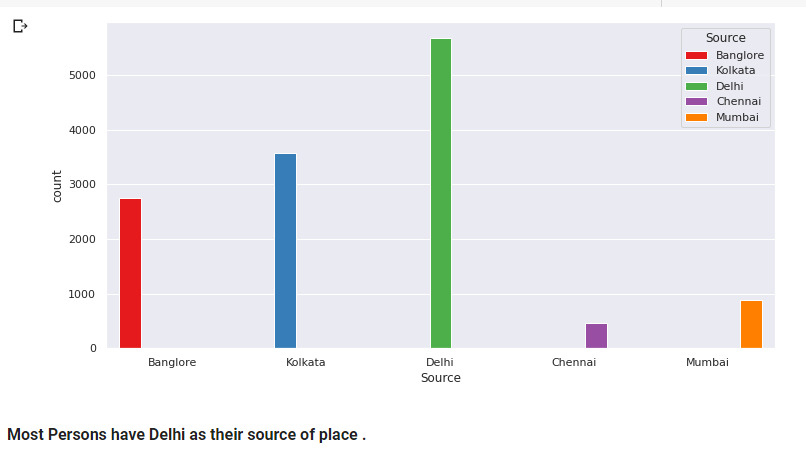
**Remarks : Above fig shows that flight price is up to 30k. And maximum fare is within 10k .**

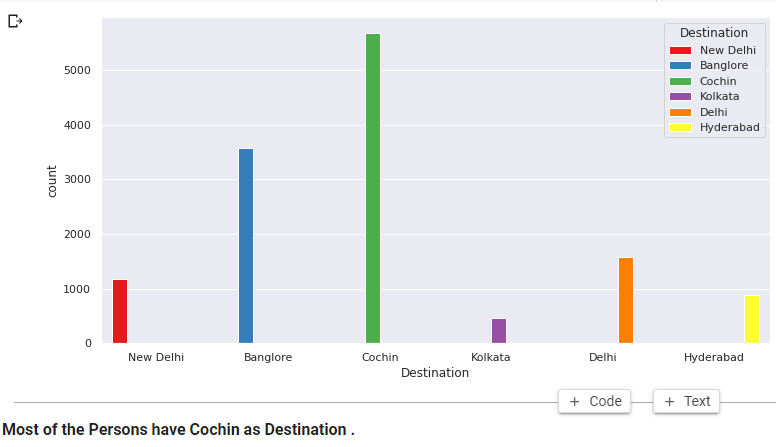
**EDA on Airline Distribution :**

****

**Remarks : Most of the customers choose jet airways as their flight with Air India and Indigo have also good number customers flying with them .**

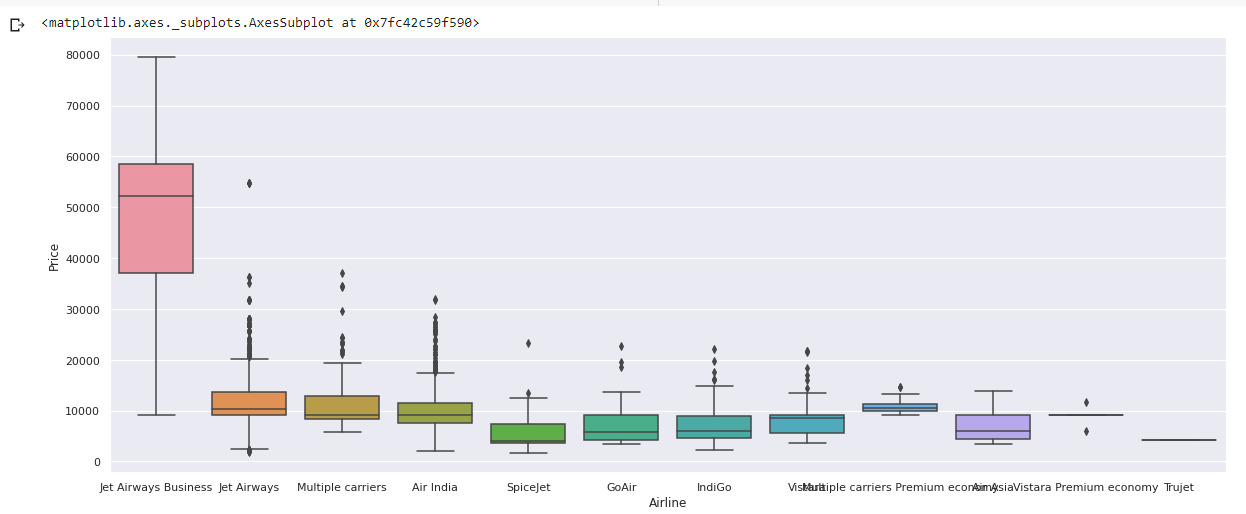
**Let’s see for Source and Destination Distribution .**

****

****

**Remarks : From the above fig’s it looks most of the travellers have their source as Delhi and also most of the travellers have their destination as Cochin .**

**Let’s look for relation between Airline vs Price :**

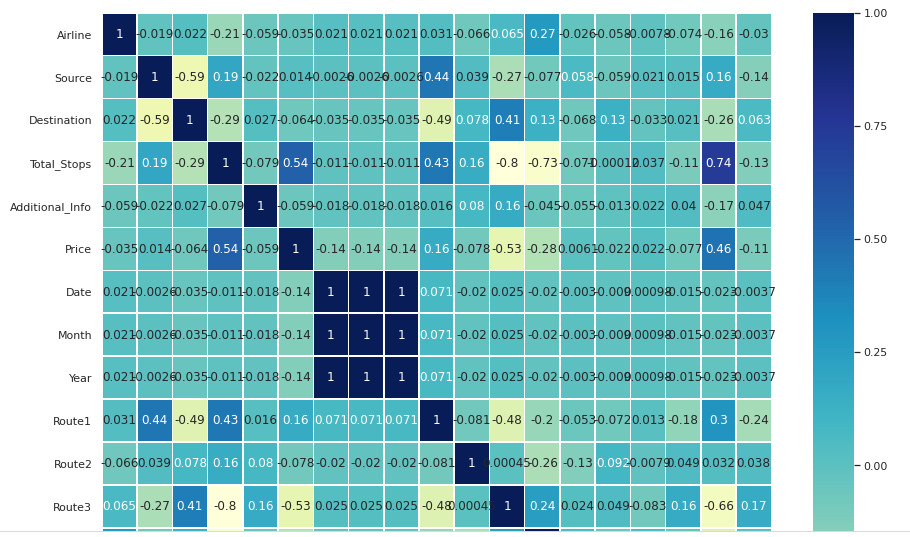
****

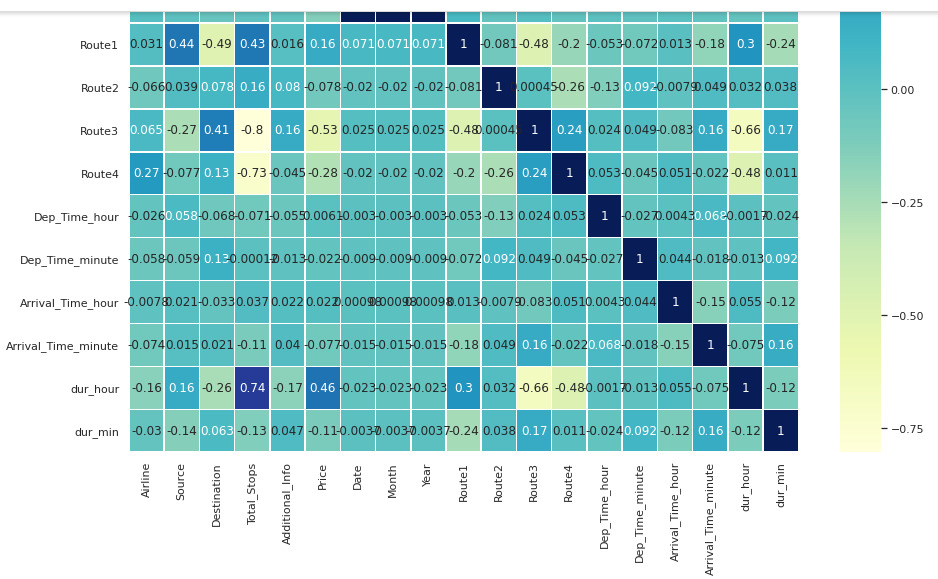
**Remarks :**

**Above bar plot shows the flight price according to the Airline.**

**We can clearly see that Jet Airways Business has the highest flight price among all the Airlines**.

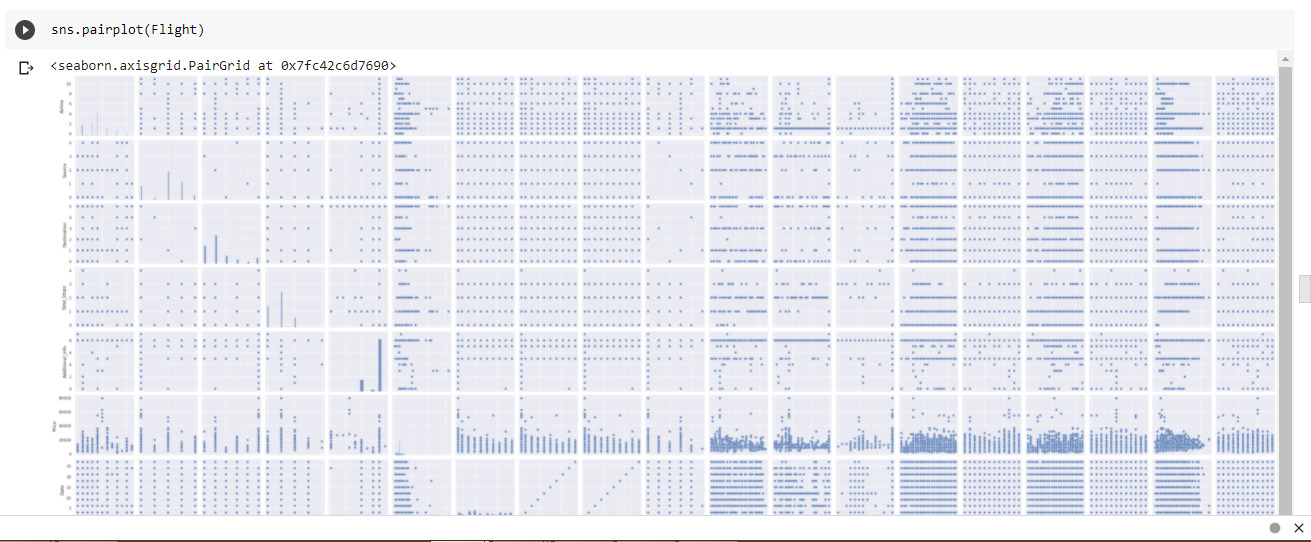
**After all process is done, we plot heatmap of correlation matrix of the train data and see the correlaion of the train data**.

****

****

**Remarks : From the above fig , it shows that columns namely Airline , Destination , Additional\_Info , Date , Month , Year , Route2 , Route3, etc have –ve correlation between Target price column . And Total\_stops , Route, Dur\_hour have +ve correlation between target column .**

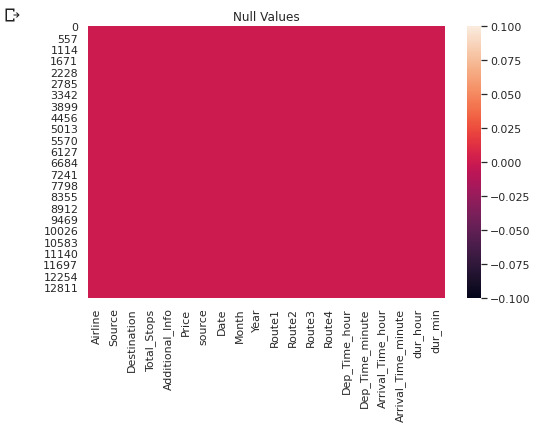
**Pairplot :**

****

**Remarks :**

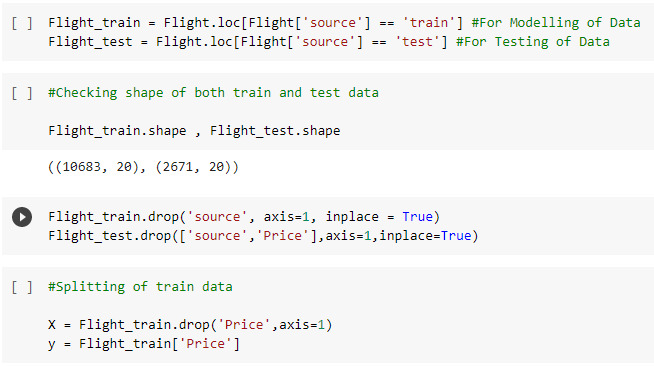
**In pairplot of the records we can see there may be lot of plots are shown, in which scatter plot, density plot, skewness of the information and plenty of other statistics is proven.**

**Plotting Null Values :**

****

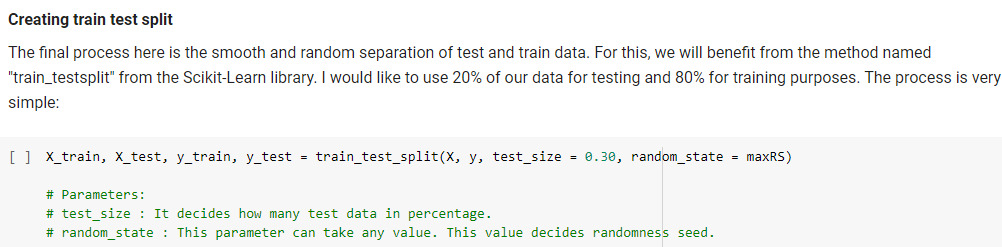
1. Pre-processing Pipeline :

**Splitting Data :**

****

**Remarks : Now after splitting the data into x and y variable, now we go for Train Test Split.Here we assign how much of the our data go for training and testing, for doing so we have to import first train\_ test\_split and some other codes also which we needed later after train\_test\_split.**

**Train-Test Split Data :**

****

**Remarks :**

**So above is the python coding for the train\_test\_split. Here we doing the training of the 70% data and testing will be done on the 30% data with random\_state of 42 i.e. taken randomly. And we can see the shape of the x\_train and y\_train of the data and also the shape of the x\_test and y\_test of the data.**

**Now our splitting and the training and testing of the data part is done, now our data is fully ready for the model building process. So lets build the model and predict the price of the flight.**

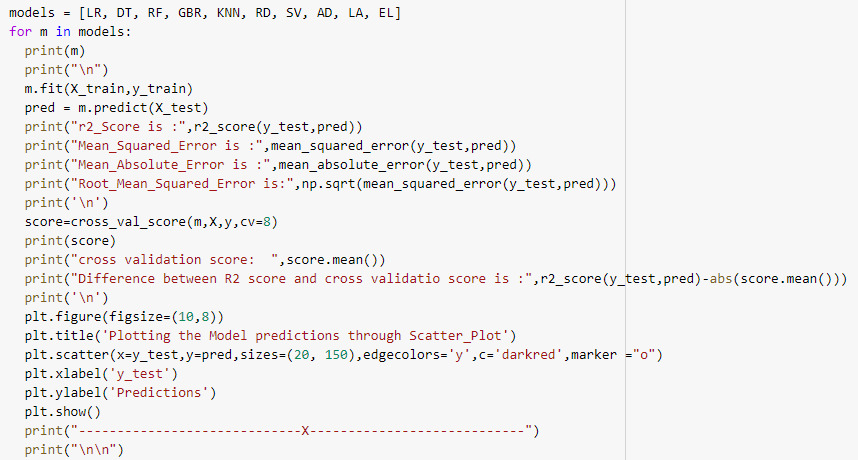
1. Building Machine Learning Models :

The goal on this step is to increase a benchmark model that serves us a baseline, upon which we are able to degree the overall performance of a better and more tunned algorithim. We're the usage of specific regressin method and comparing them to look which algorithim is giving higher overall performance then other and we can additionally go for a move validation check for due to the fact that the model isn't in overfitting/underfitting .

I'm the usage of linear regression from linear model, svr from svm, decidion tree regressor from the tree and also the use of a few ensemble techinques like random forest regressor, gradient boosting regressor and ada improve regressor.

Due to the fact our version is regression type version so we use r2\_score, mean\_squared\_error, and mean\_absolute\_error.

****

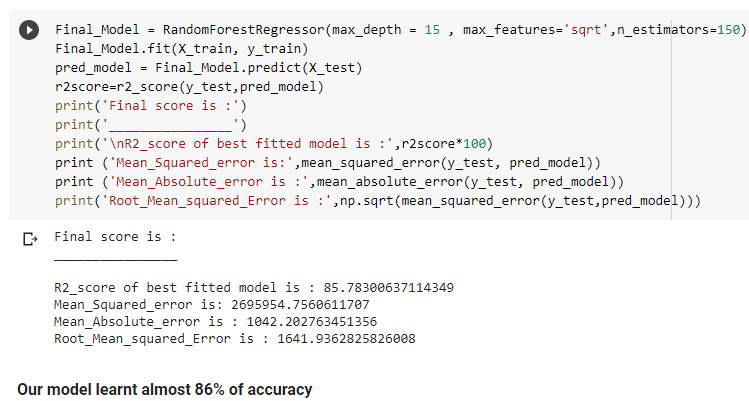
****

**Remarks :**

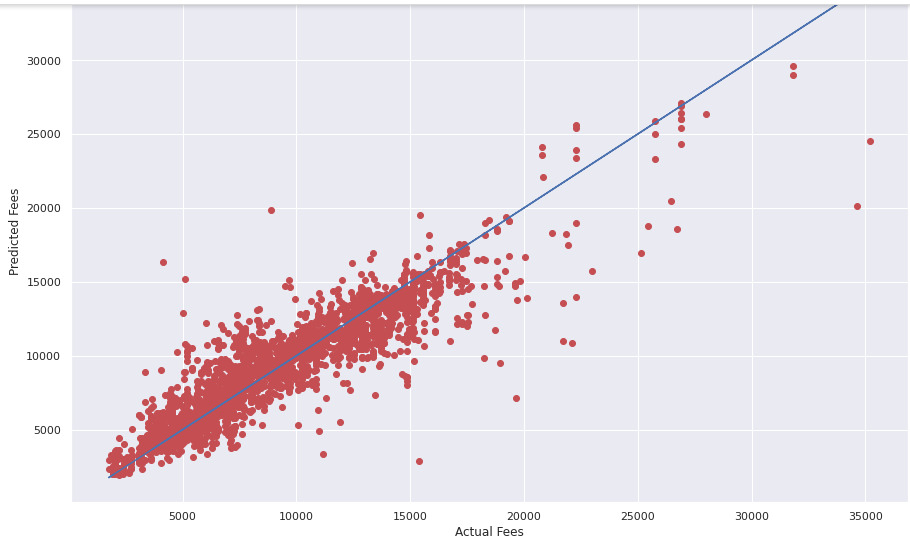
1. **We see Regression model - RandomForestRegressor has the highest r2\_score of 0.85 while SVR has the lowest r2\_score of 0.04 .**
2. **But in RandomForestRegressor we see that difference between r2\_score and cross validation score is least among all i.e, 0.002 . Hence RandomForestRegressor is our best fitted Model .**

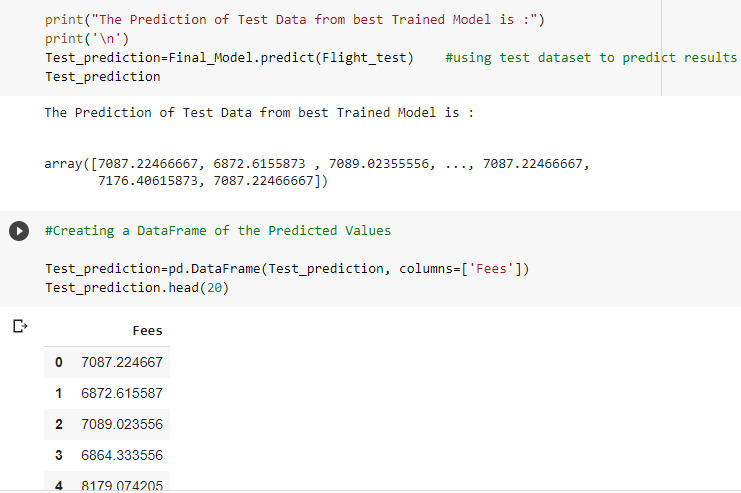
**Perform Hyper-Parameter Tuning**

****

****

**Plotting the Model :**

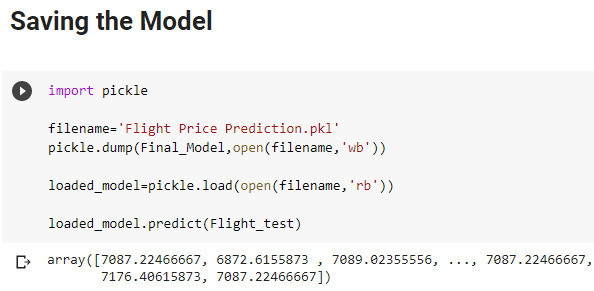
**Predicting Flight Price :**

****

**We can see above the final prediction for the test data.**

**And loading the final prediction for the test data to the DataFrame and seeing the Price columns for the Test data which is predicted Price for the flight.**

**Now saving the model by importing joblib using .pkl method.**



1. Concluding Remarks :

**From the statistics retrieved and through exploratory information analysis, we are able to decide the subsequents below :**

* 1. **The trend of flight prices range over diverse months and throughout the holiday**
  2. **There are two groups of airlines: the least expensive group and the high priced organization. Spicejet, Air-Asia, indigo, cross air are in the budget friendly elegance, whereas jet airways and air india inside the other. Vistara has a more spread out trend.**
  3. **The Flight Price varies depending at the time of departure, making timeslot used in analysis an critical parameter.**
  4. **The Flight Price will increase during a vacation season. We haven’t considered vacation season as a parameter now, considering that we are searching at information for a few months.**
  5. **Ai**rfare **varies according to the day of the week of travel. It is higher for weekends and Monday and barely lower for the opposite days.**
  6. **There are some instances when an offer is run through an airline due to which the expenses drop all of sudden. These are difficult to comprise in our mathematical models, and for this reason lead to error.**
  7. **Alongside the Mumbai-Delhi path, we find that the charge of flights will increase or stays steady because the days to departure decreases. That is because of the excessive frequency of the flights, excessive demand and additionally may be due to heavy opposition.**
  8. **Best about 8-10% of the instances, a person ought to wait according to the records accumulated across the Mumbai-Delhi route, compared to 30-forty% in Delhi-Guwahati course**