Flight Price Prediction

**Problem identification:**

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 flight bookers saying that flight ticket prices are so unpredictable.

To avoid that we need to automate the Flight price Prediction process, from the given data provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities are analysis and predict by EDA and Machine Learning Models.

**Import libraries:**

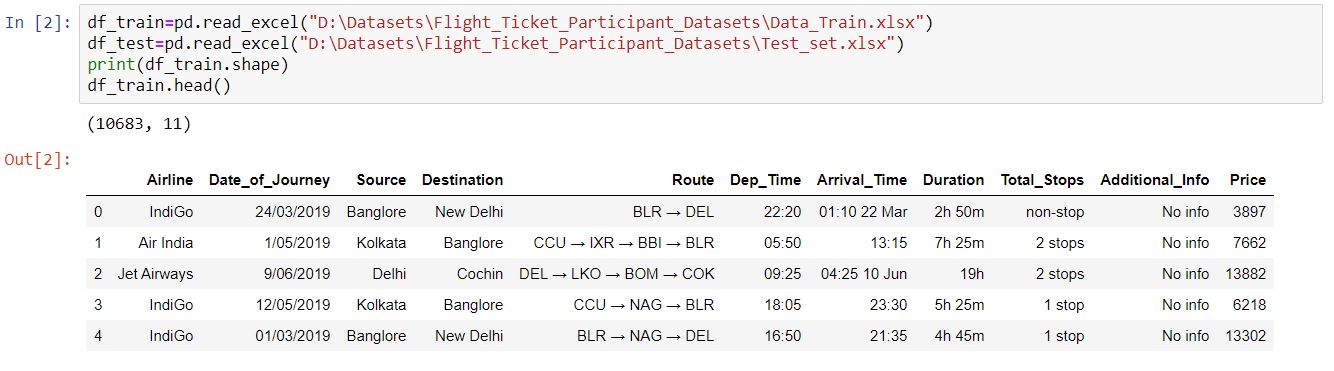
Have to import the basic required libraries for Exploratory Data Analysis process



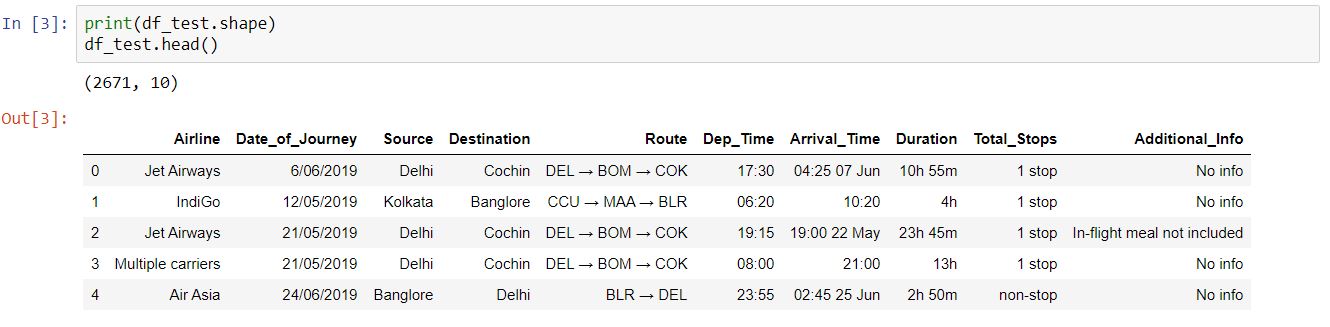
Here Pandas and numpy are libraries used for data processing, matplotlib and seaborn are visualization library used to visual the given data graphically in a easy understandable way.

**Reading Data:**

To load the data by using pandas



Here data is divided into two as train and test dataset we have to read it separately and then combine it for analysis process. Above it shows that shape of whole train dataset and it’s top 5 rows.

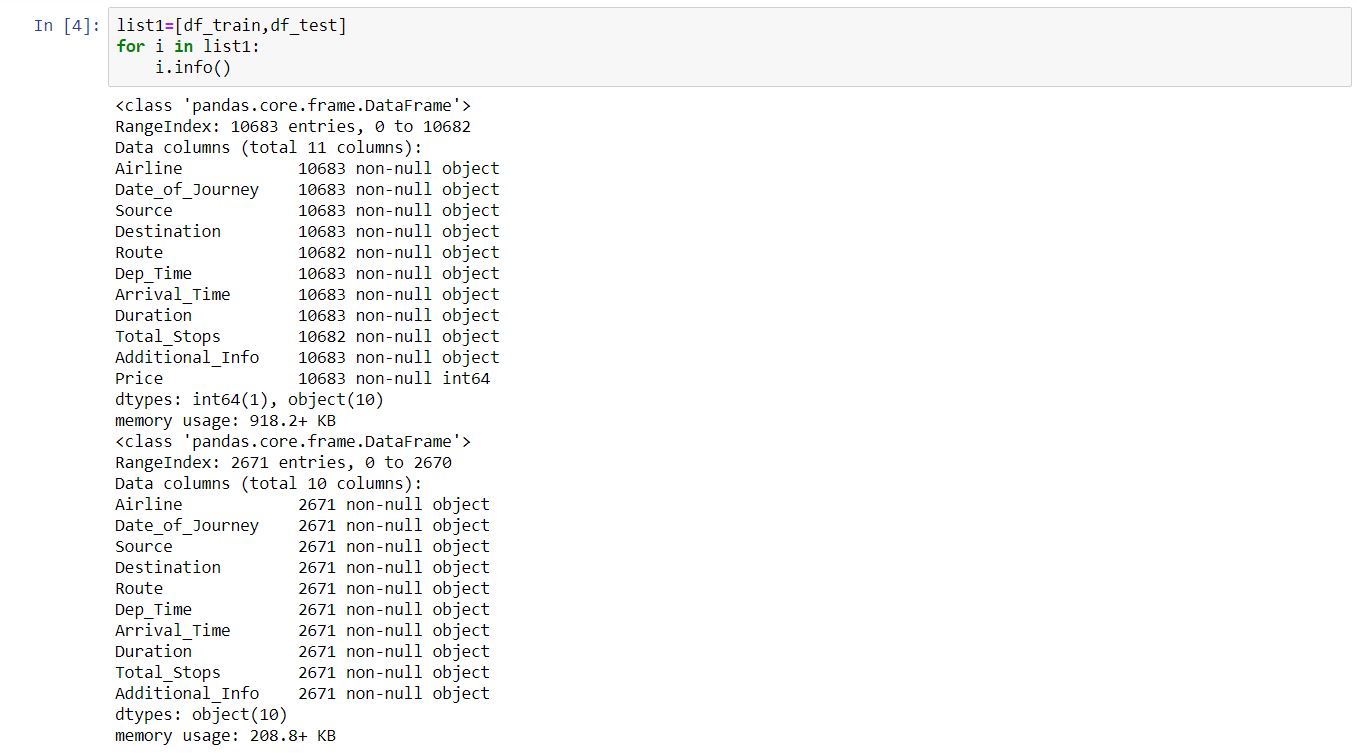


Above it shows that shape of whole test dataset and it’s top 5 rows.

It shows that test dataset have only input data there is no presence of output data price.

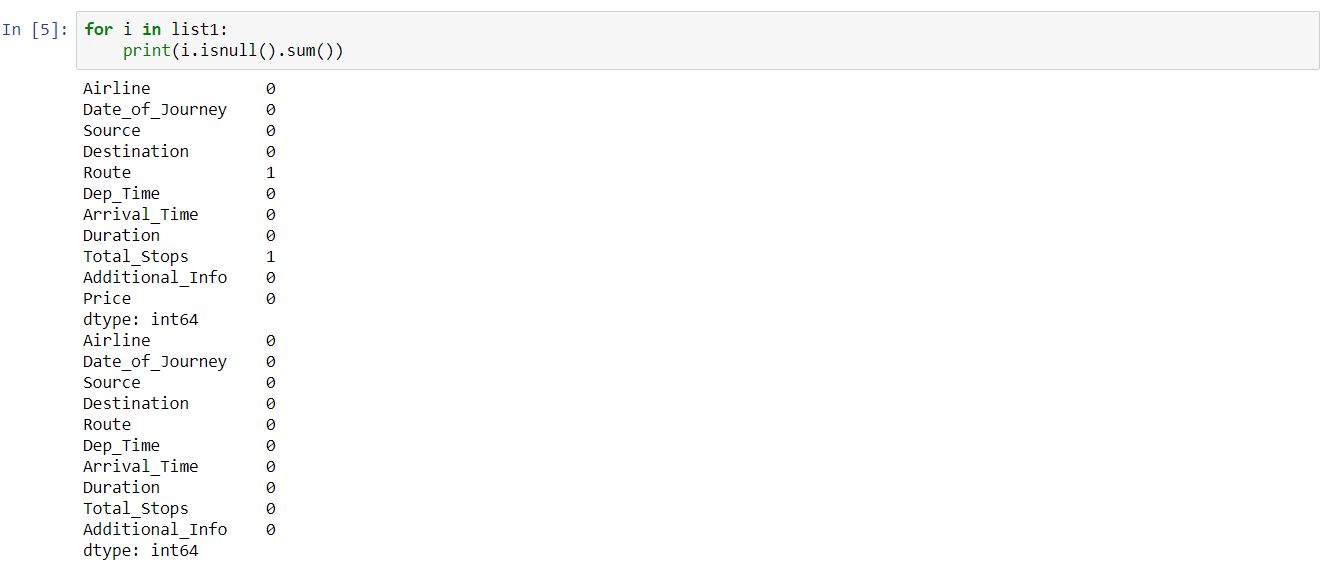
**Exploratory Data Analysis:**

To know the data type of both the dataset and check the presence of null value can done by



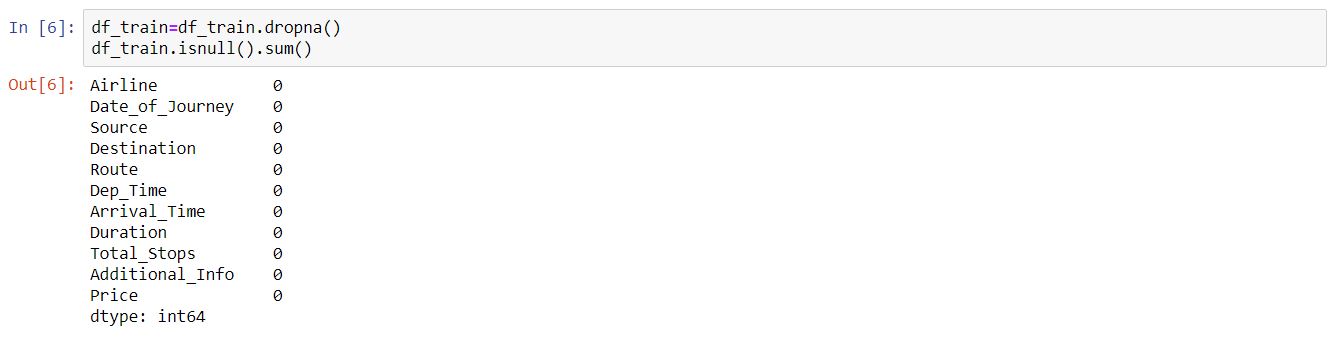
In these dataset all input columns are object data type and price outcome is in integer data type.

It show the count of null values in respective columns

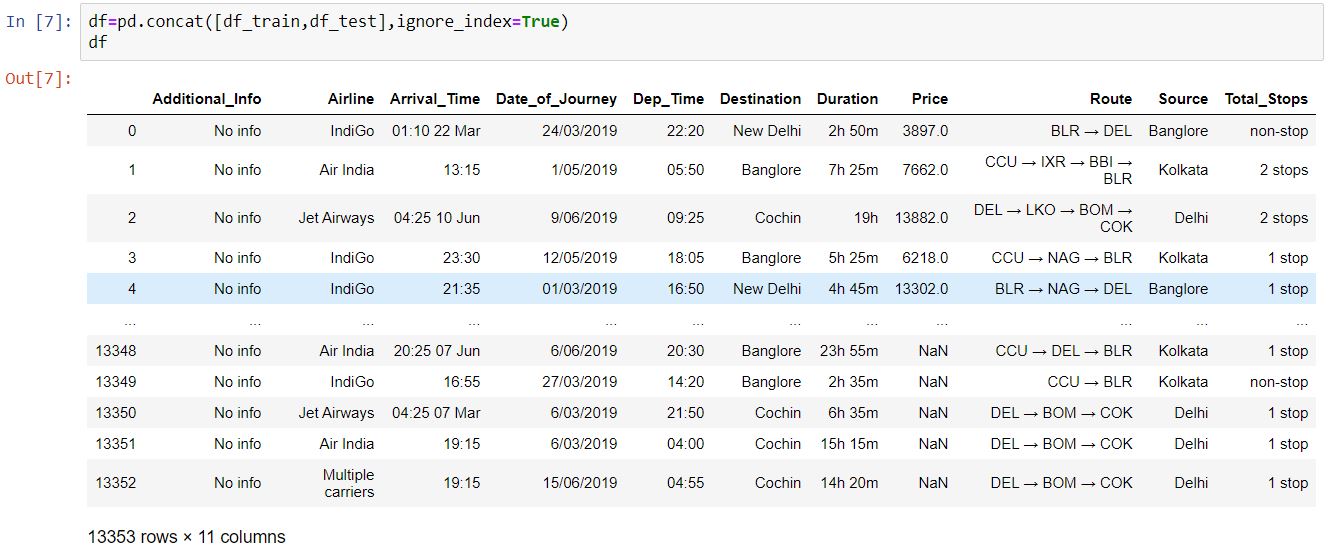


**Treating of null values:**

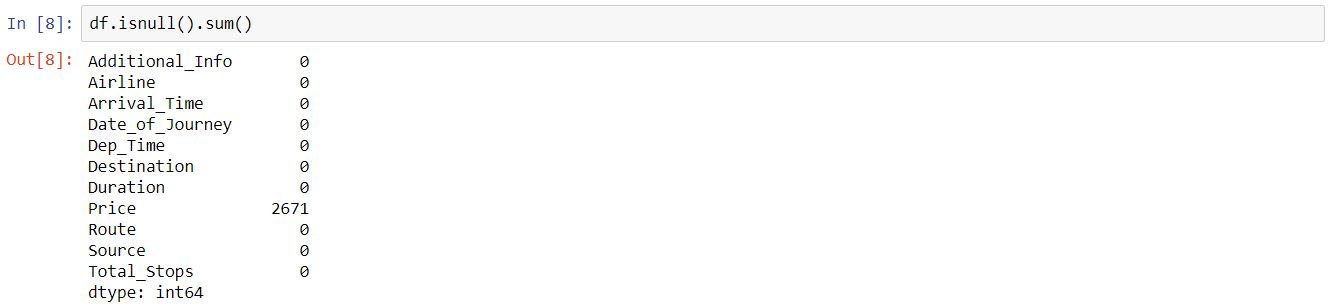
Replace the null values with mean or median (but median is better than mean because of presence of outliers) for unique numerical data and for categorical data it can be fill by mode operation or also can use drop option (but not recommended).



Combine the train and test dataset into a single data for analysis purpose by joining of dataset we can use concat, merge, append operations.

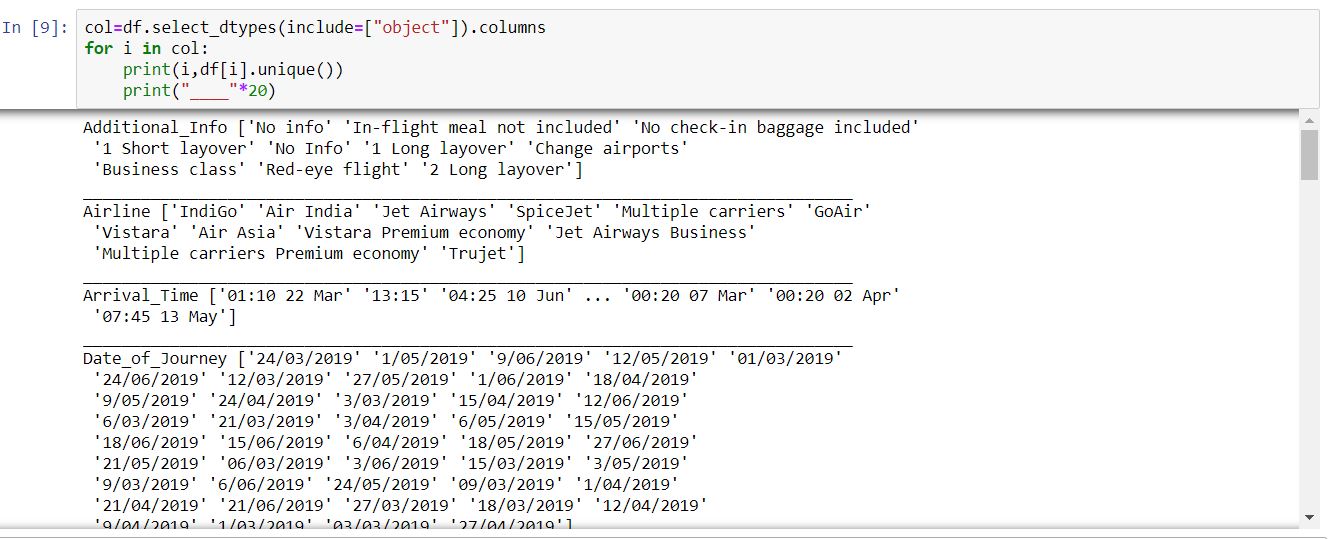


Again to check the presence of null values in the combined dataset



In price column there is occur of null values it is due to shape of train and test dataset. That is due to missing of price data in test dataset.

To check the unique value of each column by

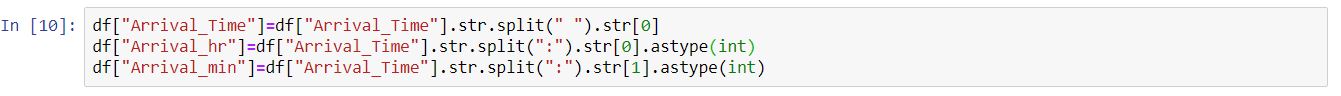


Total Stops, Source, Destination, Airline, Additional info have minimum unique values and we have to treat the remaining columns with maximum unique value.

To convert some of object data into required format and as a integer type for model to understand

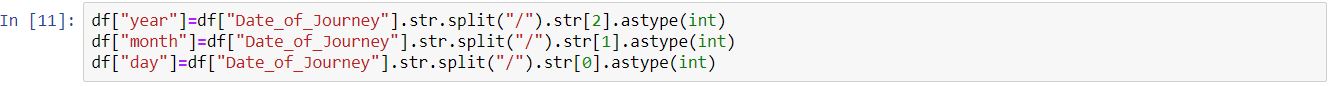
Arrival Time is in the format of hour: minutes date month and it is a object data.

We have to separate Arrival hour and arrival minute from that format, so



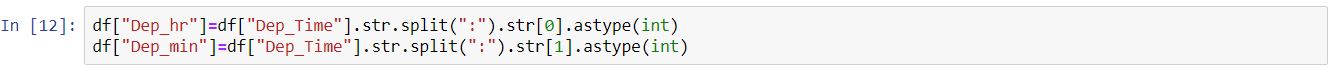
By use of split function we can able to get the required data from the combined data.

Date of journey is in the form of dd/mm/yyyy and it is a object data.



By use of split function we can able to get the required data from the combined data.

Dep Time is in the format of hour: minute, we have to separate it into hours and minutes columns and also as a integer type



Duration is in the format of hours minute, we have to convert it into only minutes columns and also as a integer type



For Route columns

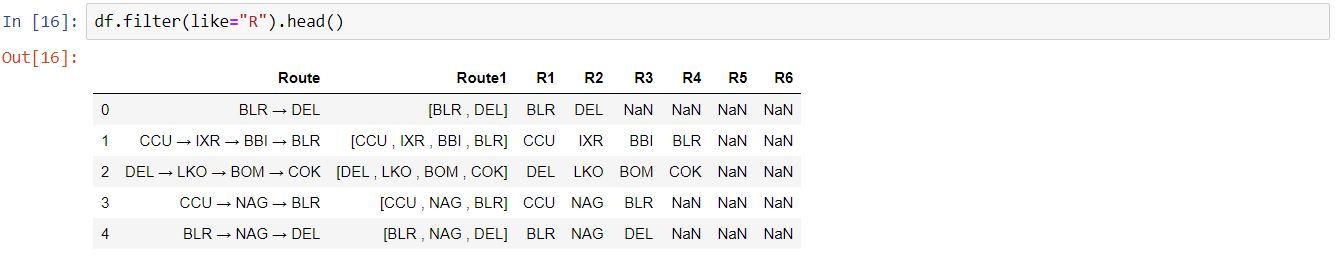
It shows the route of flight fling, in a format city → city → city →….. and so on .



The maximum number of city in a route of flight fly is 6.then,



With the help of split function every Route of flight is got split and stored in separate column but each column data type will be object type.



To drop the unwanted columns from that we converted into required format we need

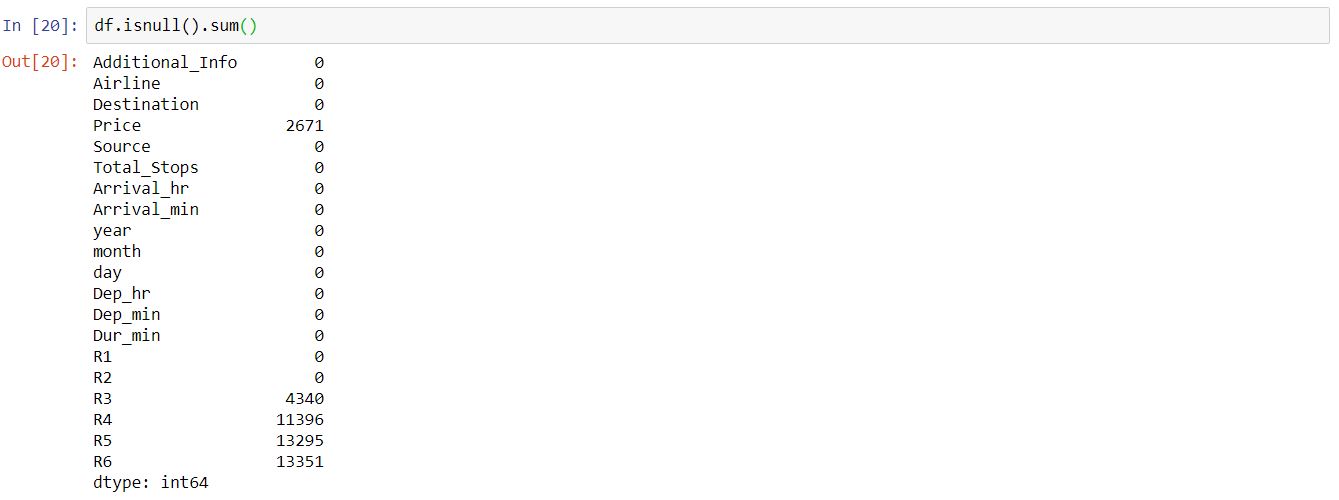
17.JPG

In additional info there is presence typing error

No Info and No info both are same so we have to replace it ,if we did not replace it may affect the accuracy a little

18.JPG

To show the count of null values in respective columns that we edit for our requirement



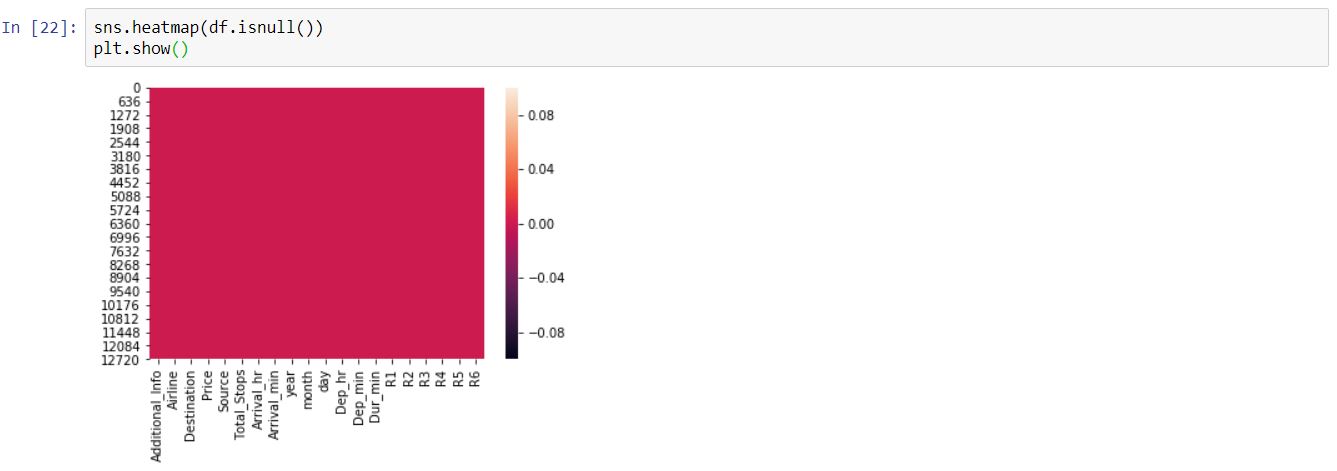
There is presence of null values in R3, R4, R5, R6 because of irregular route of every flight.

To replace it,

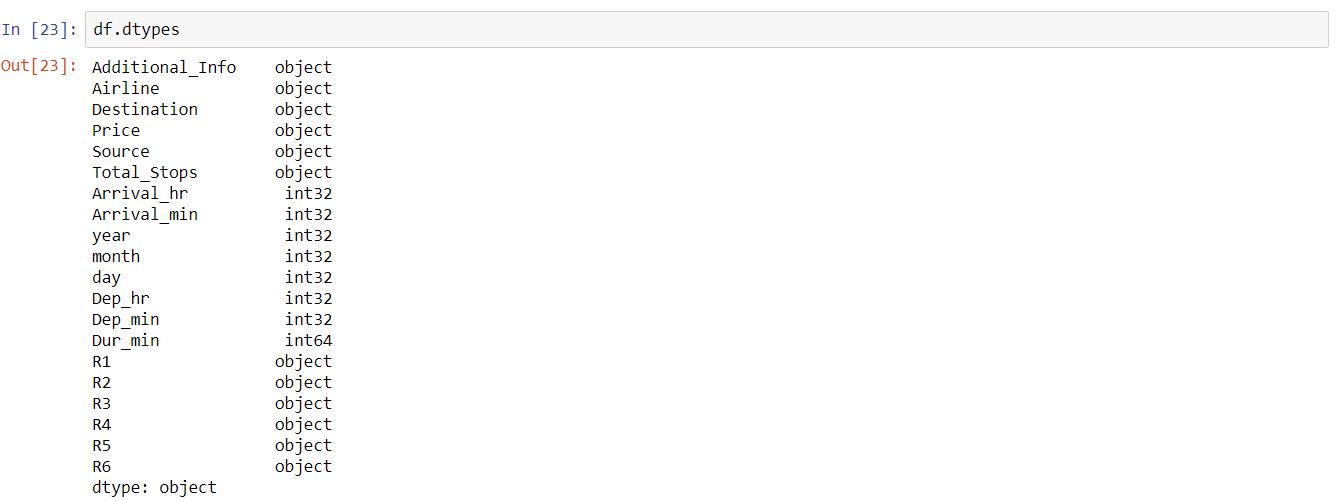
21.JPG

Null value into None because we can’t use mean or median because it is a object data type mode filler is also not able to use because it will change the route of flight

To check whether null value presence is there are not can again done by above format (df.isnull().sum()) or by Graphical method

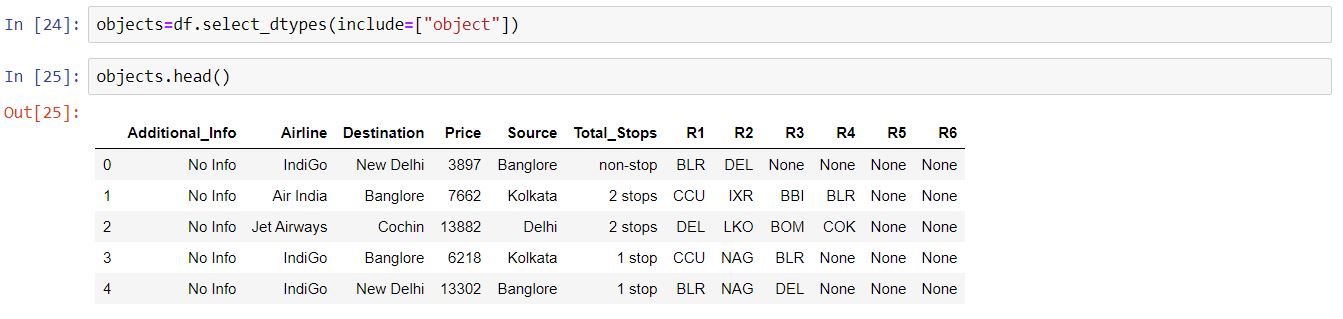


**To check the columns which are in what type :**



There some data are in integer type and some are in object type.

To separate only object data from whole dataset:



**Visualization:**

Visualize the given data Graphically is to make a quick decision about the data

**Categorical data:**

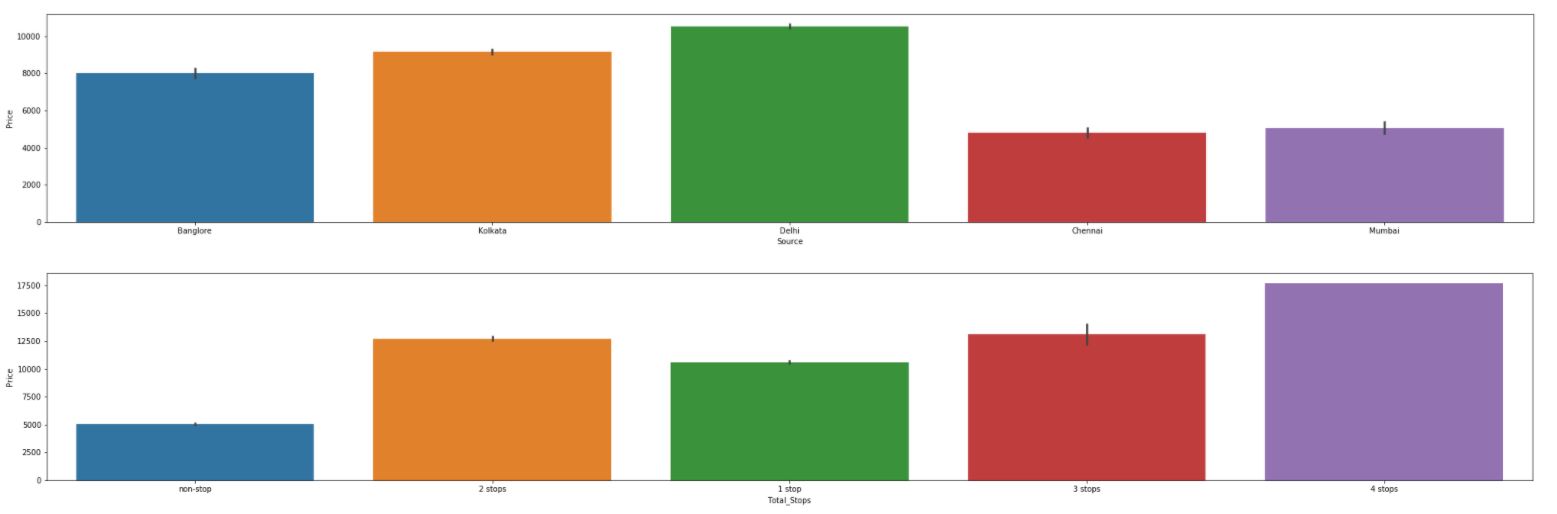
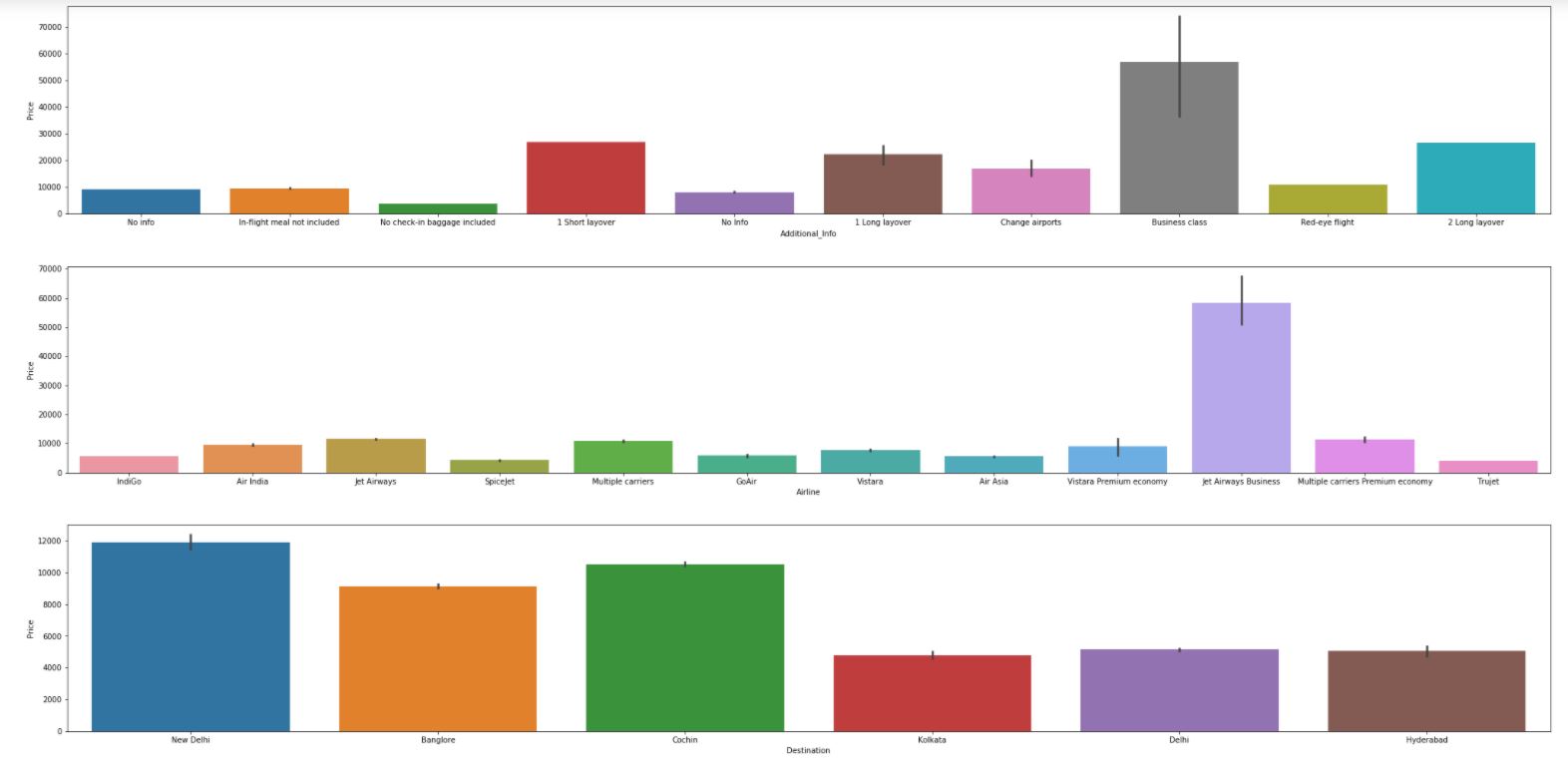
**Bi- variant analysis:**

Compare the object data type with Flight Price

26.JPG

Drop these columns is for to visualize the object data with less unique values





The above bar graphs shows that

Business Class shows high price in Additional\_Info

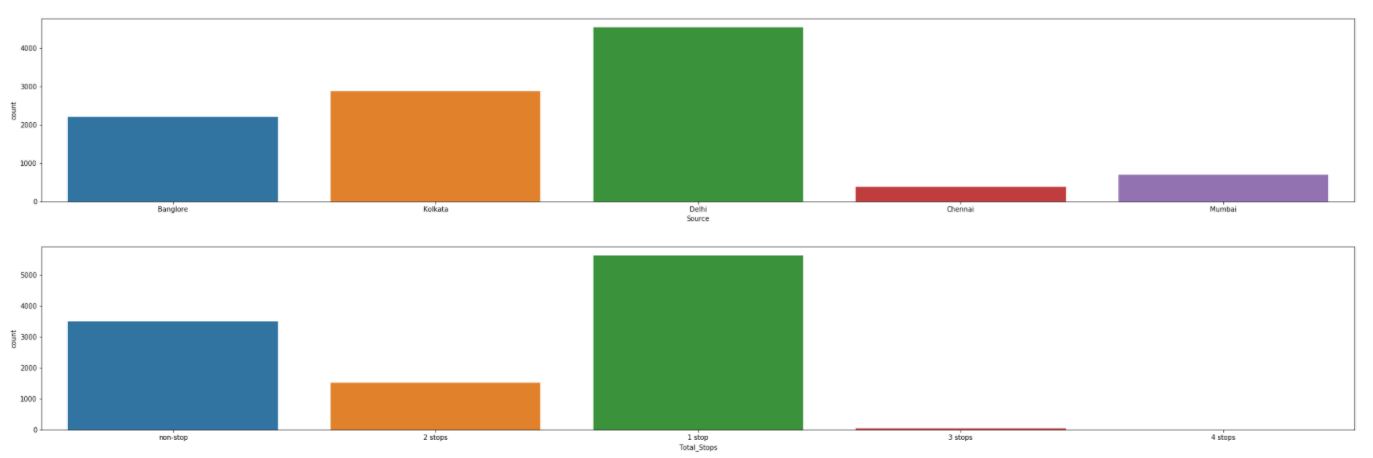
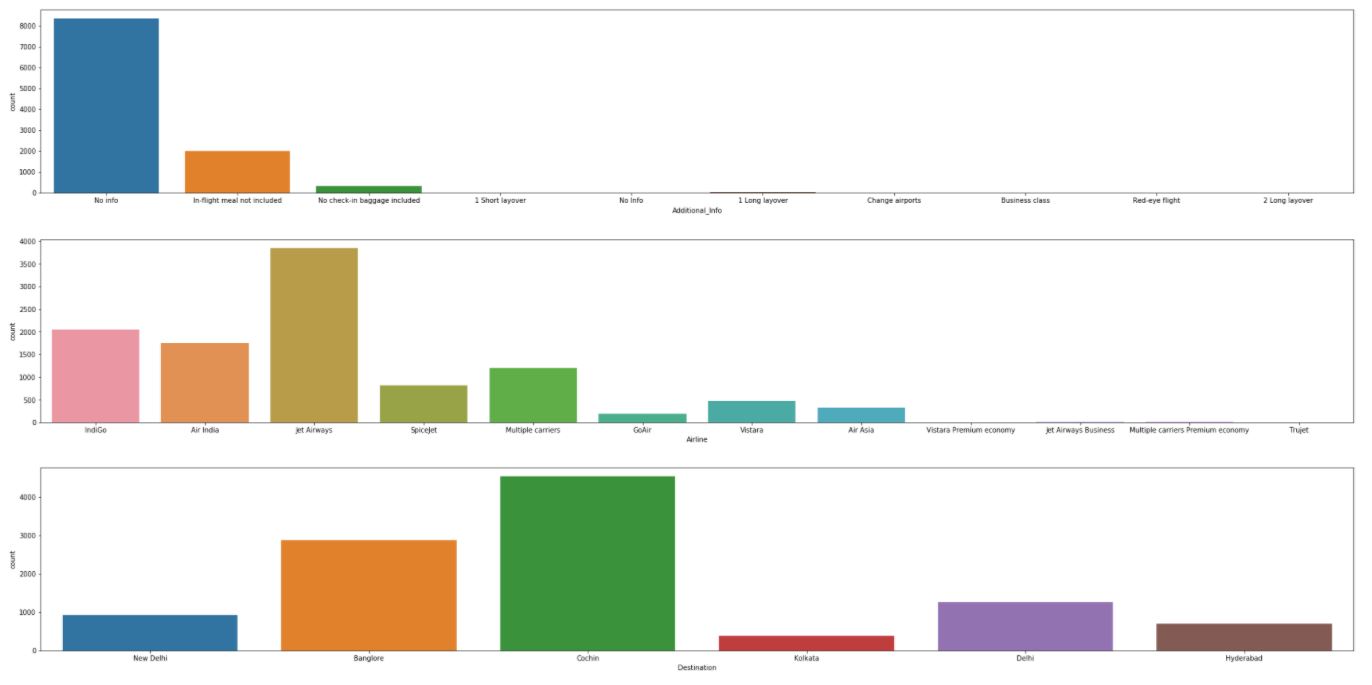
Jet Airways Business Air line is at high price when compare with others

High price place to reach are New Delhi , Banglore, Cochin than others

Increase in no of stops leads to high in price

**Uni- variant analysis:**





High amount of destination booked is Cochin and the lowest preference is Kolkata.

Flights start from Delhi the most when compare with others and low number of flights are from Chennai.

Lot of Flights are operated by Jet Airways.

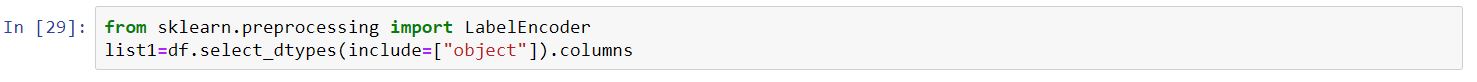
Lot of people selects 1 stop flight. The preference of 3 stops and 4stops are very low.

**Encode the data:**

To convert object data into integer data

Before apply the Machine Learning model we have to check required data is in integer, if not we have to convert it as integer data, because the model will only consider integer data as their input and output data to train and test the data.

Encoding technique can help to convert it as a required type.



Encode can be done by either label encoder or one hot encoder these are available in scikit learn library.

One Hot Encoder is for to encode the categorical data column with only two possible answers (yes or no) or (True or False).

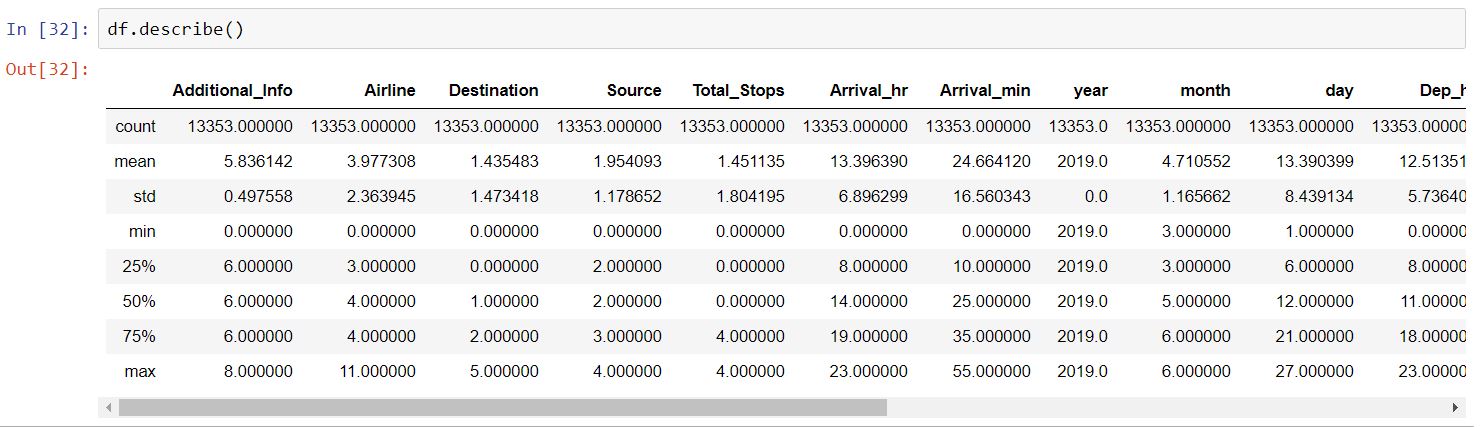
Label Encoder is applicable for the object data with two and more variety of data in a column.

Here I used Label Encoder because all of the columns have more than two types of data.

There One Hot Encoder is can’t applicable.



**To describe the columns:**



The above figure shows that there is no presence of outliers in columns because it have no huge difference between variation in (min,25%,50%,75%) and max(maximum number)

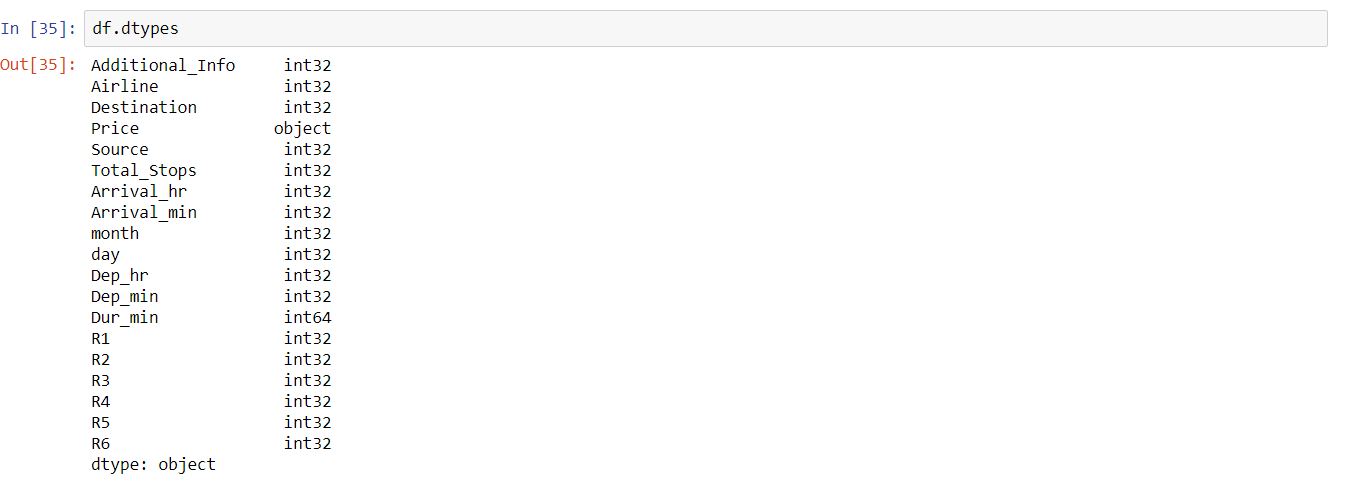
If the standard deviation is 0 then that column will not use for prediction because it won’t affect or make impact in model

It also show presence of data is skewed

**Drop the unwanted columns:**

33.JPG

To check all the columns are converted to integer data:



Here all the data type of columns are showing either integer type except Price Columns ,because of presence of null string in price of test data it will be omitted while training phase of model, So it will be a good data for to apply in model.

**Finding of correlation between data present in dataset:**

Correlation is can either positive correlation or negative correlation

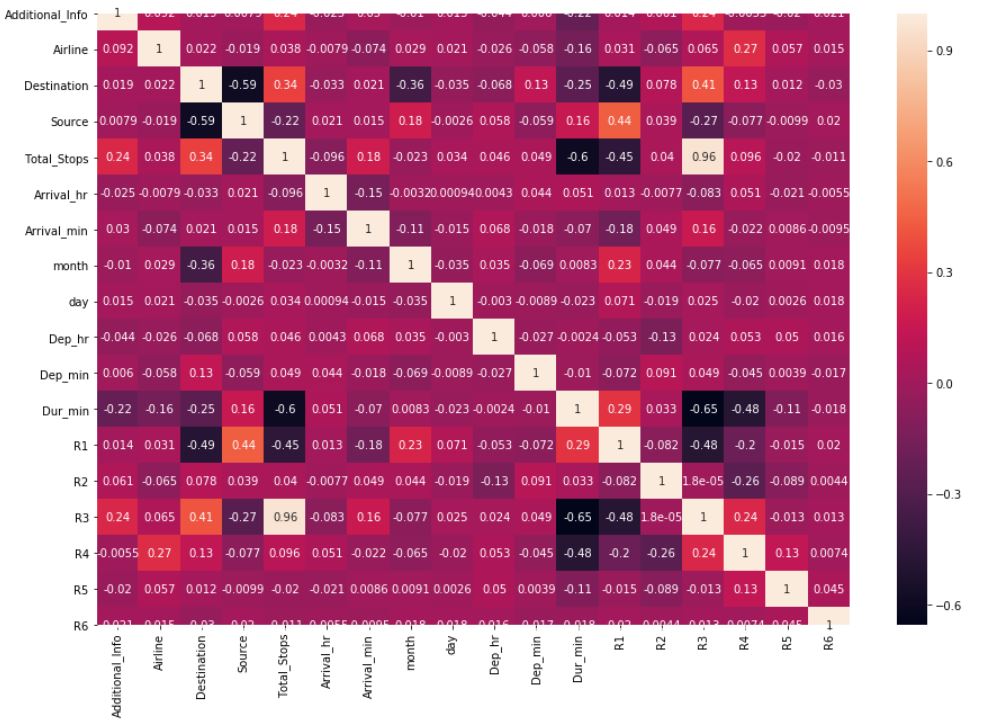
Positive Correlation - (0 to 1)

Negative Correlation – (0 to -1)

If the correlation value is nearer to 1 or -1 this data will give great impact in values to the corresponding data.

Or if the correlation value is nearer to 0 this data will not impact in values to the corresponding data.





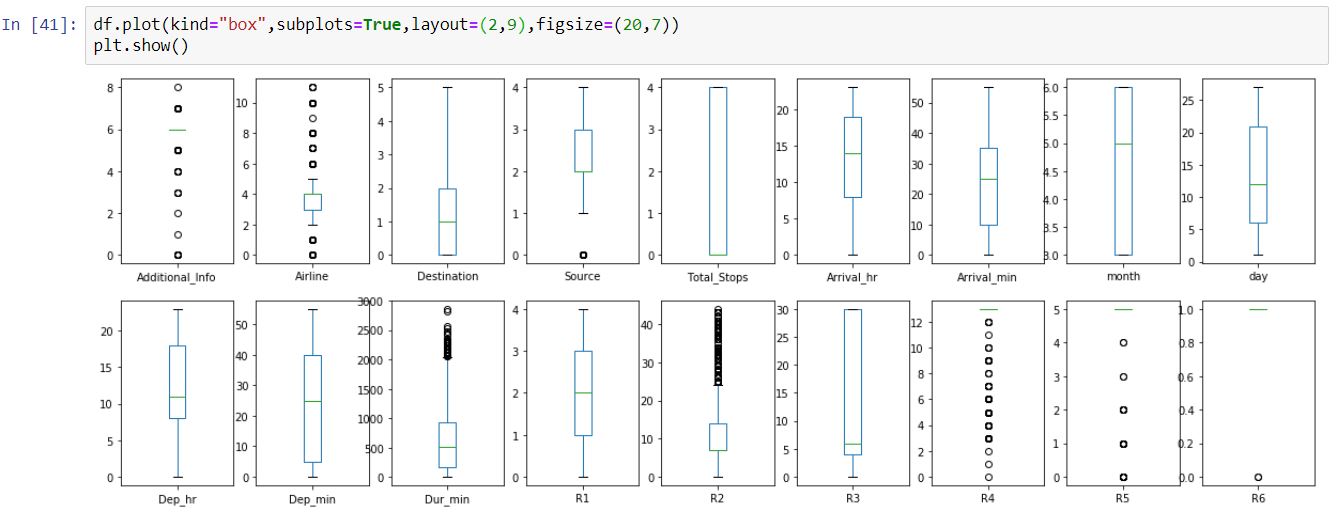
**Outliers Mining:**

Outliers are presence of unwanted data present in dataset, it also may be a fake or by typing mistake.

To remove the presence of outliers we have apply Z-scores on the data, before that we have to find the presence of outliers in the given data.

Outliers Identification

It can be able to done by box plot



Circles present on top of maximum value and bottom of minimum value in given data will be consider as outlier.

Here it showed that there is present of outliers. But these are categorical data, So no need to remove outliers, if you have to remove it can be by

Outliers Removal:

It is going to done by Z-Score.

Z-score measures how far the data points are present away from mean of the data, and also there will be a presence of threshold value(-3 to +3)

If the distance between data points and mean of data is above that fixed value it will be consider as a outliers.

**Skewness Identifier:**

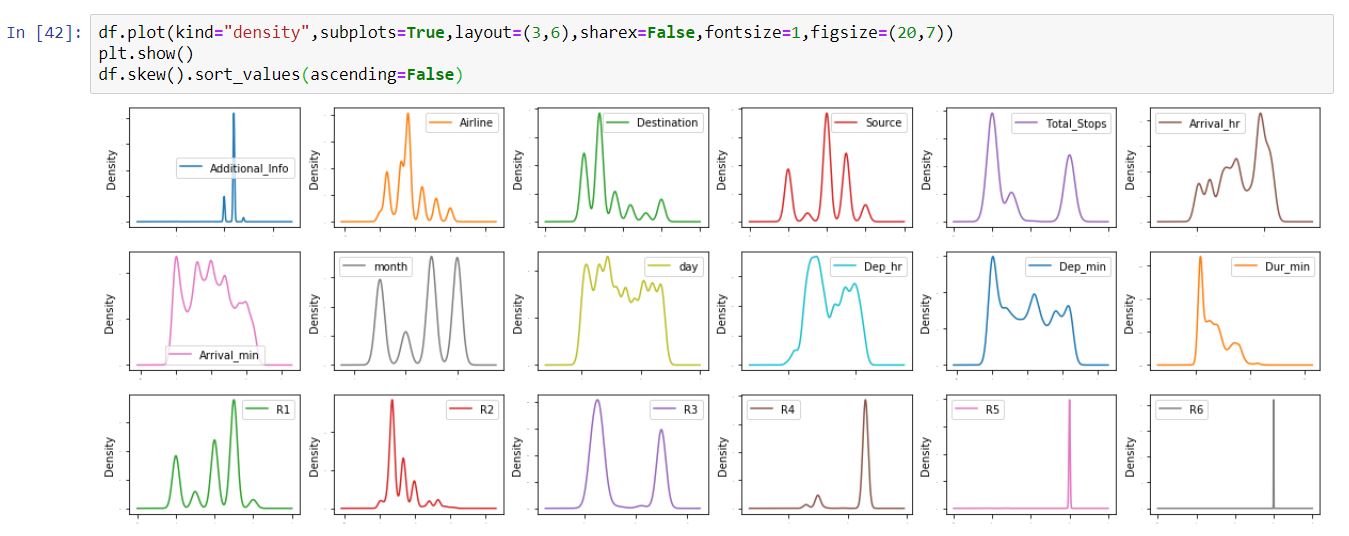
Skewed data refers as it is not like a Normal Distribution (Standard Deviation = 1), there curve will be asymmetric that is curve may be left or right skewed

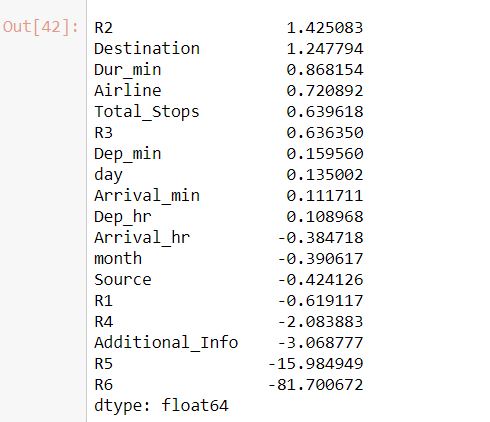
Left Skewed = Negative

Right Skewed = Positive

Normal Distribution = Zero Skewed value

If the data is right skewed or left skewed it will affect the model accuracy, else if it is Normal Distribution data it will good for model to predict





For to get a good normal curve the skew value should between (-0.55 to +0.55)

**EDA Concluding Remark:**

**From the above visualizations it concludes,**

From the above two it is clearly states that Jet Airways are lot in number so their price amounts are high.

4 Stops flight are less in number, So its price is also high.

In Delhi there are lot of flights are get operated when compare with other so price income from Delhi Airport is high.

There occur of removal of null values and shrinking and expand of required columns are occur.

Some unwanted columns are get removed.

All columns data are converted as integer type.

**Split data process:**

Before building Models separate train (input data) and test data (output data)





Test data which concat with train data have to remove

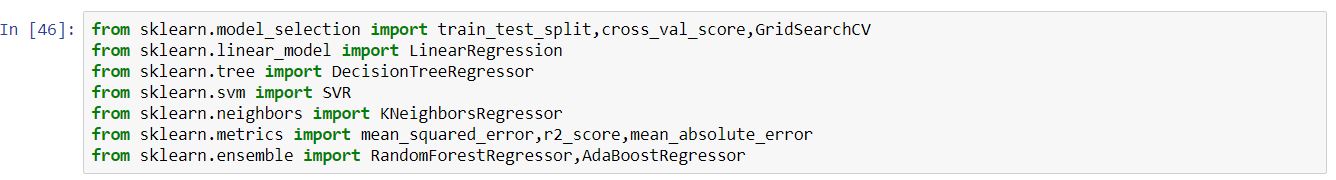
Except Price column every data columns will be consider as a input data

Price column will consider as output data

**Building Machine Learning Models:**

**Model Import:**

Import the required model to apply in prediction process.



These models are not inbuilt in python so we have to import it from sklearn (Scikit Learn).

Models are stored in particular repository and these repositories are have to import from sklearn.

Here the given output data is not have only two outcome values 1 or 0, So it will be consider as Regression data model,

Model comes under Regression are

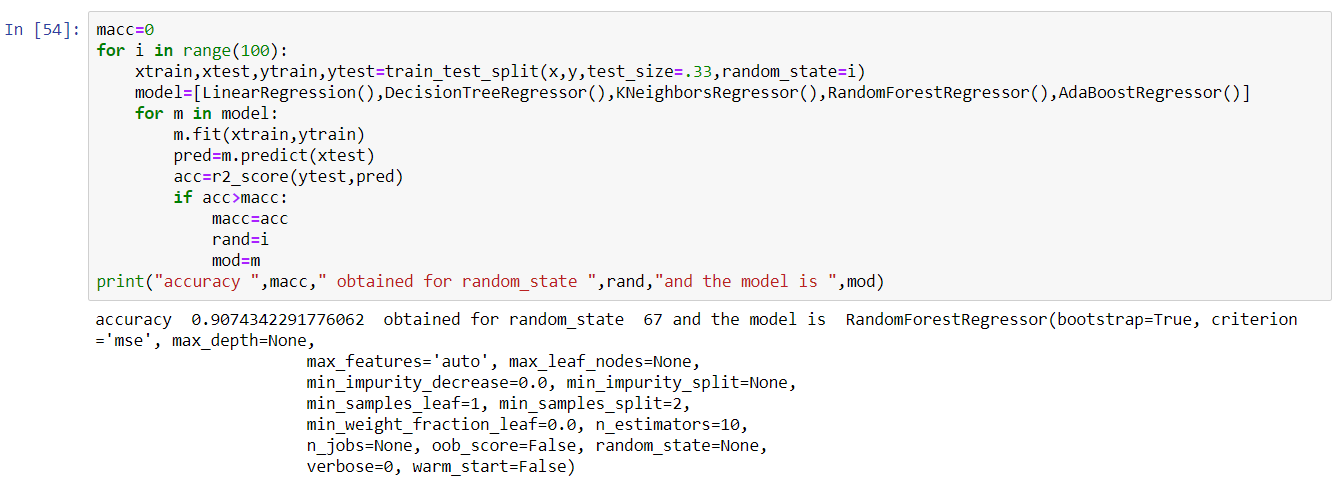
1. LogisticRegression
2. DecisionTreeRegressor
3. SVR
4. KNeighborsRegressor
5. RandomForestRegressor
6. AdaBoostRegressor

In metrics repository

1. r2\_score is for to check the accuracy score of the prediction value.
2. Mean squared error
3. Mean absolute error

**To find best random state and model:**

For select best model and random state which give high accuracy score can able to find by iterate different random state and classification model through the model train and test process



Train test split is helps to split the train and test data with some tuning parameter.

Test size = It decides ratio of train and test data if 0.30 is test size then test data percentage is 30% and train data percentage will 70%.

Random state = It is a type of tuning where it randomly desired which data should go for train process and which data is for test process.

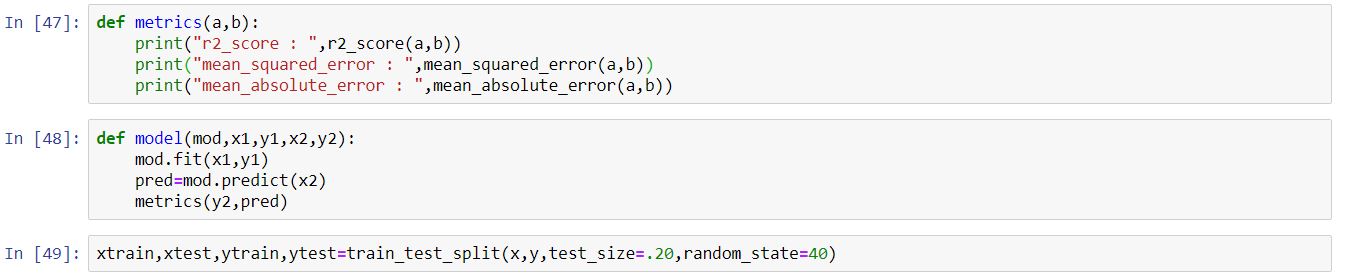
In training phase we have to fit both the input and target train data.

Prediction is has to done for input test data to get target test data.

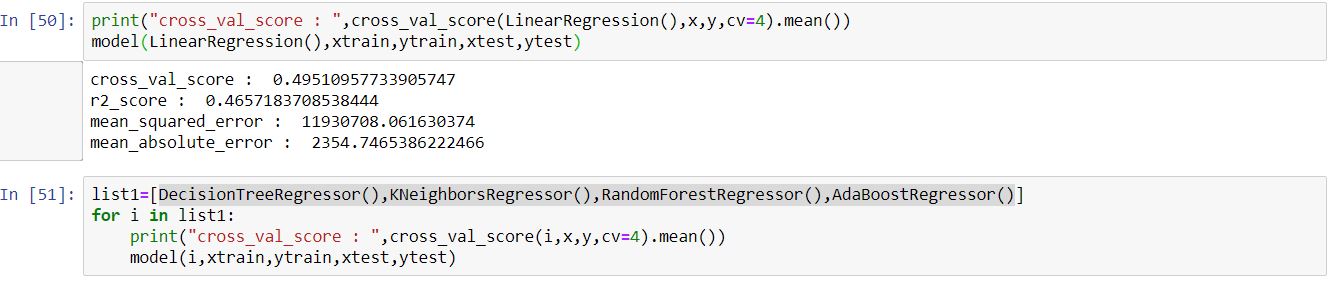
For to find the r2 score we have to compare the actual target test data with predicted target test data.

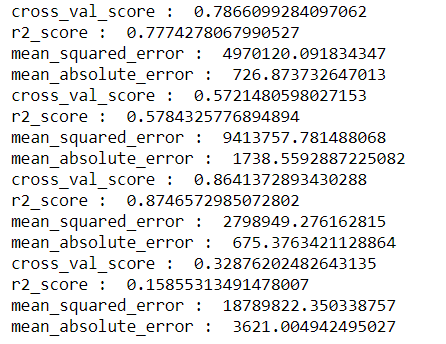
**Cross validation:**

There may be over fitting can occur to avoid that we have to do cross validation by cross\_val\_score model in sklearn library.



Here we implemented model train and test structure in user defined function named as model and r2\_score, Mean squared error and Mean absolute error in matrix user defined function.





From the above model Random Forest Regressor had a good r2 score because the difference of r2 score and cross validation score is minimum for Random Forest Regressor when compare with other model score.

**Hyper parameter tuning:**

Hyper parameter tuning can be done by Grid Search CV for Random Forest Regressor



For Random Forest Regressor there are certain parameters to tune for getting good accuracy score

After applying Grid Search CV for Random Forest Regressor model there may increase in accuracy score by the process of tuning.

**Conclusion:**

Dump the model prediction in a file for the developer use



Joblib or pickle will help you to dump the predicted model in object file.

It can be saved for later prediction process for the same size of input data.

With the help of model saved file by input data, we can able to predict the price of flight as outcome .