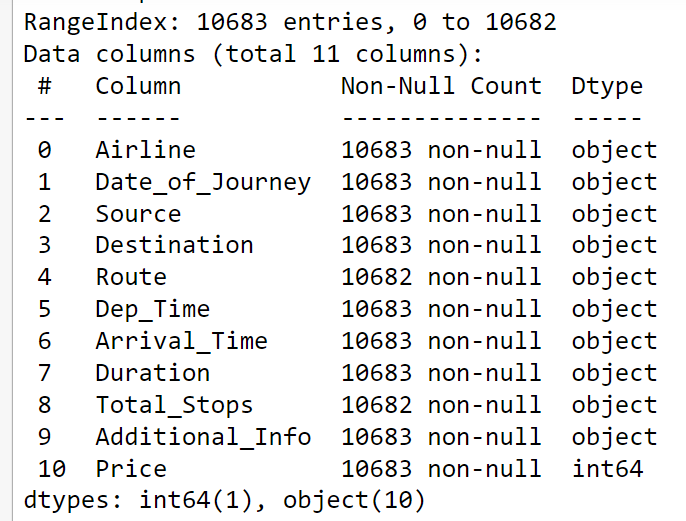
# Problem Statement:

Travelling through flights has become an integral part of today’s lifestyle as more and more people are opting for faster travelling options. The flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, destination, and duration of flights various occasions such as vacations or festive.

# Data Description

Data is in form of excel file, for reading data though this we will use pandas. Below is column wise description of dataset.



After loading it we need to check complete information of data as it can be indication many of them may have null rows, columns or values.

Check whether any null values are there or not If preent the following needs to be done:

* + Imputing data using imputation method in sklearn.
  + Filling NaN values with mean, median and mode using fillna() method.

### EDA

From description we can see that Data\_of\_Journey is a object data type, Therefore, we have to convert this datatype into timestamp so as to use this column properly for prediction.

For this we require pandas to\_datetime to convert object data type to datetime dtype.

.dt.day method will extract only day of that date.

.dt.month method will extract only month of that date.

Arrival time is when the plane pulls up to the gate. Similar to Date\_of\_Journey we can extract values from Arrival\_Time.

Time taken by plane to reach destination is called Duration. It is the difference between Departure Time and Arrival time.

### Handling Categorical Data

One can find many ways to handle categorical data. Some of them categorical data are,

1. Nominal data: data are not in order --> OneHotEncoder is used in this case.
2. Ordinal data: data in order --> LabelEncoder is used in this case.

As Airline is Nominal Categorical data we will perform OneHotEncoding.

As Source is Nominal Categorical data we will perform OneHotEncoding.

As Destination is Nominal Categorical data we will perform OneHotEncoding.

As stops case of Ordinal Categorical type we perform LabelEncoder.

### Feature Selection

Finding out the best feature which will contribute and have good relation with target variable.

Following are some of the feature selection methods,

* heatmap
* feature\_importance\_
* SelectKBest

In this case we used heat map to find correlation between Independent and dependent attributes.

### Fitting model using Random Forest

Split dataset into train and test set in order to prediction w.r.t X\_test If needed do scaling of data

Scaling is not done in Random forest

Import model

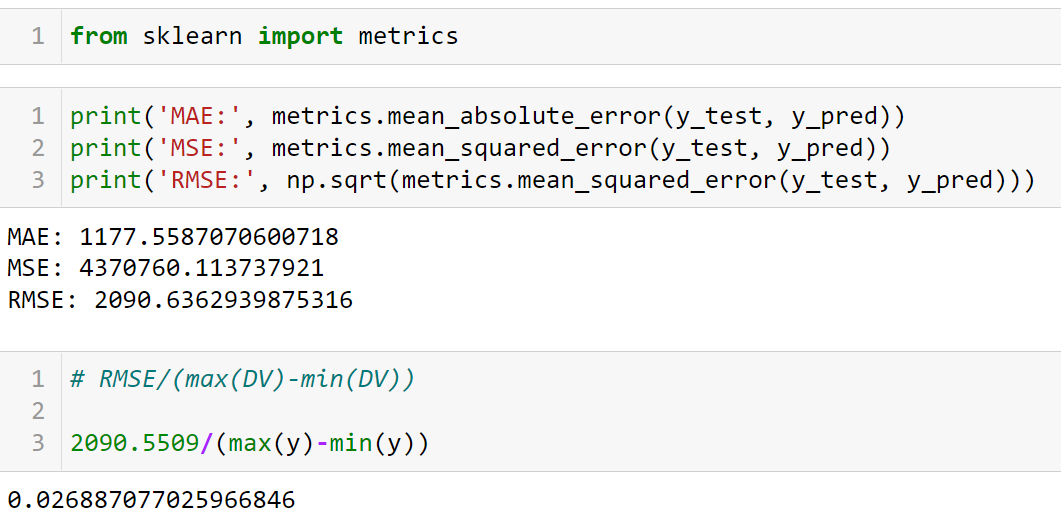
Fit the data

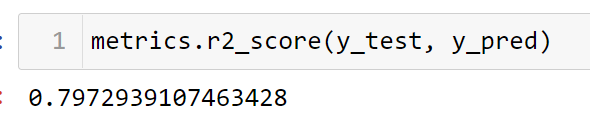
Predict w.r.t X\_test

In regression check RSME Score

Plot graph

Below are various parameters for regression problem statement of this case:





### Hyperparameter Tuning

Choose following method for hyperparameter tuning

* RandomizedSearchCV --> Fast
* GridSearchCV

Assign hyperparameters in form of dictionery

Fit the model

Check best paramters and best score.

Below is screen shot reference for the hyper parameter performed in this case.



