**Flight Ticket Price**

**Problem Definition**

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 prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities are provided .We have to analyse and pre-process to get useful data and then predict the price of Flight Ticket by model building. Let’s start the journey.

**Dataset**

We are provided with 2 datasets train data and test data. Train data is for analysis and model building and test data for predicting the Price of Flight Ticket.

**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.

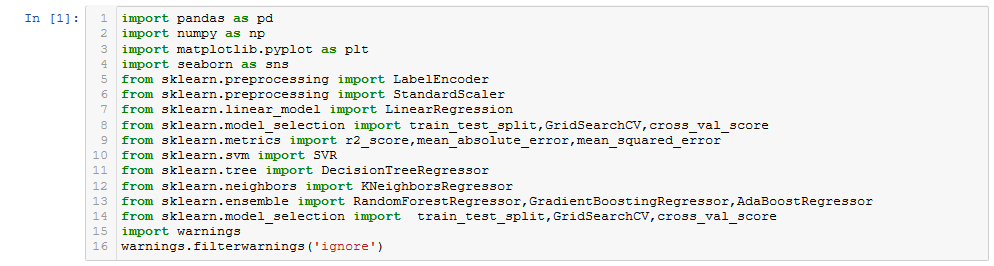
**Total\_Stops**: Total stops between the source and destination.

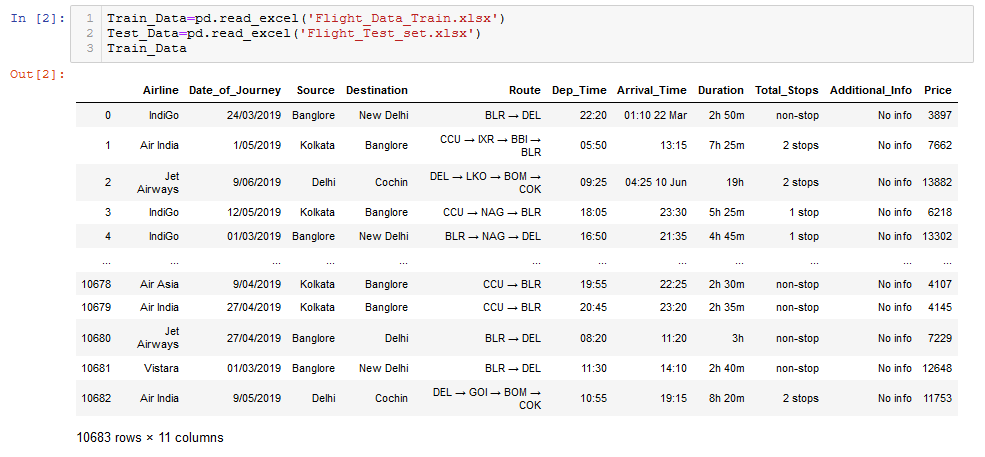
**Additional\_Info**: Additional information about the flight

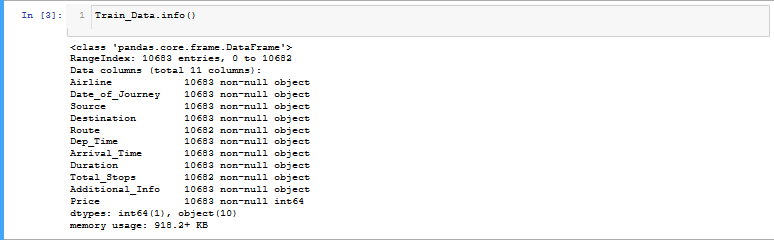
**Price**: The price of the ticket

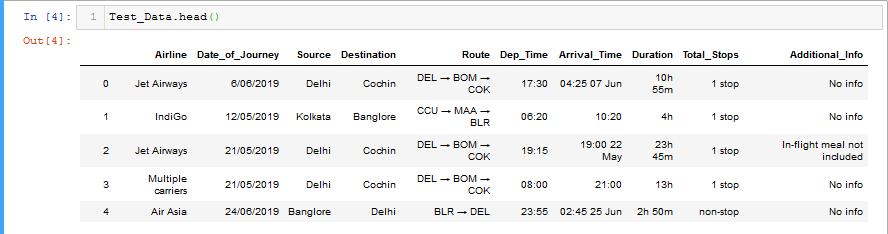
**Step-1:**

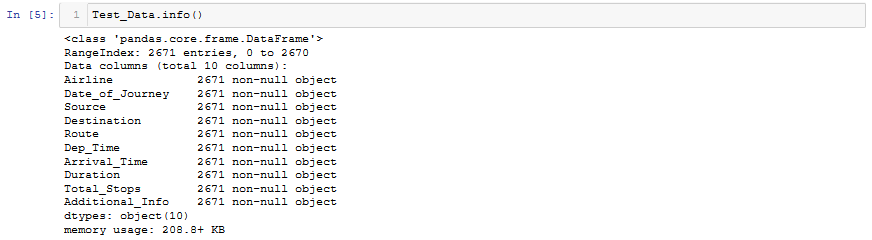
Predicting Price hints that it is a Regression Problem. So loading the libraries for Regression Problem.



**Step-2:**  Loading the datasets and checking their shape, data type and see if there any null values.



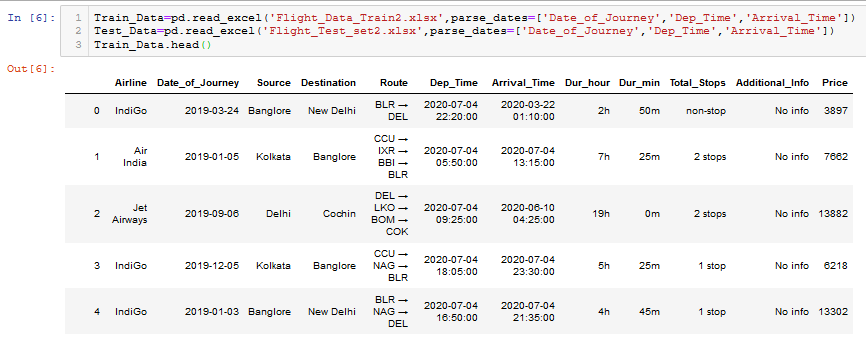


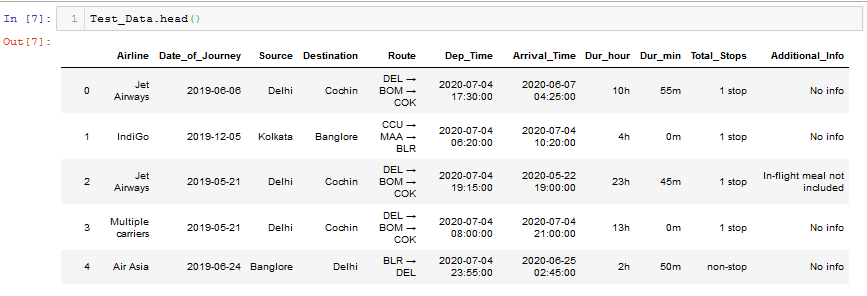


The train dataset have 10683 observations with 11 features and the test dataset have 2671 observations with 10 features (because it has no target column).

Data\_of\_Journey,Dep\_Time,Arrival\_Time are in Object datatype let’s parse their datatype to datetime while again calling the datasets to python after Delimiting the Duration column making Dur\_hour and Dur\_min by excel.

Let’s call the datasets after delimiting the column

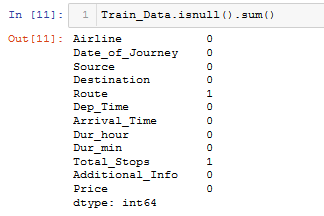


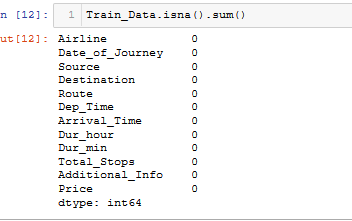


**Step-3**

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA for see what the data can tell us beyond the format modelling or hypothesis testing task. We can say that EDA is statistician’s way of storytelling where we explore data, find patterns and tells insights. EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like:-main feature of the data-variables and relationships that hold between them- identify which variable are important for our problem We shall look at various exploratory data analysis methods like :- Descriptive Statistics, which is a way of giving a brief overview of the dataset we are dealing with, including some measures and features of the sample. So Data analysis is very important to work with any dataset. Let’s begin

Before we have checked shape and datatypes of train and test datasets now we check if there any null values present in the datasets because if any null values present in the datasets we can’t work with them.



In the train dataset in column Route and in Total\_Stops single singe null value present. As the number of null values are very less compared the dataset so we drop the columns having null values by



For Better Analysis firstly we do feature Engineering then Data Analysis

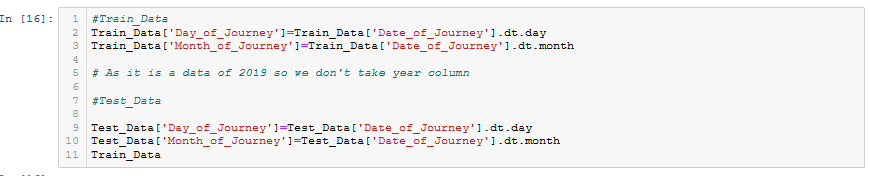
Feature Engineering-

It is very important in Machine Learning because without feature no one can’t do anything and by Feature engineering using domain knowledge to extract features from raw data via data mining techniques. Features can be used to improve the performance of machine learning algorithms so that more accurate prediction can be predicted. Also Feature engineering can be considered as applied machine learning itself.

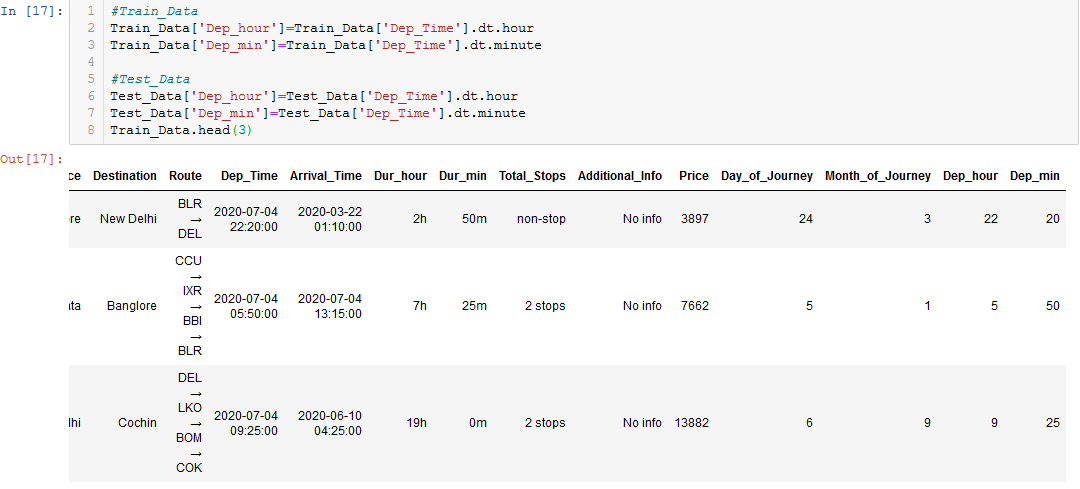
All the time in the dataset we do same FE with both the datasets so that test data is featured like train data.

Day\_of \_Journey:-

We Spllit Data\_of\_Journey column to Day\_of\_Journey and Month\_of\_Journey as-

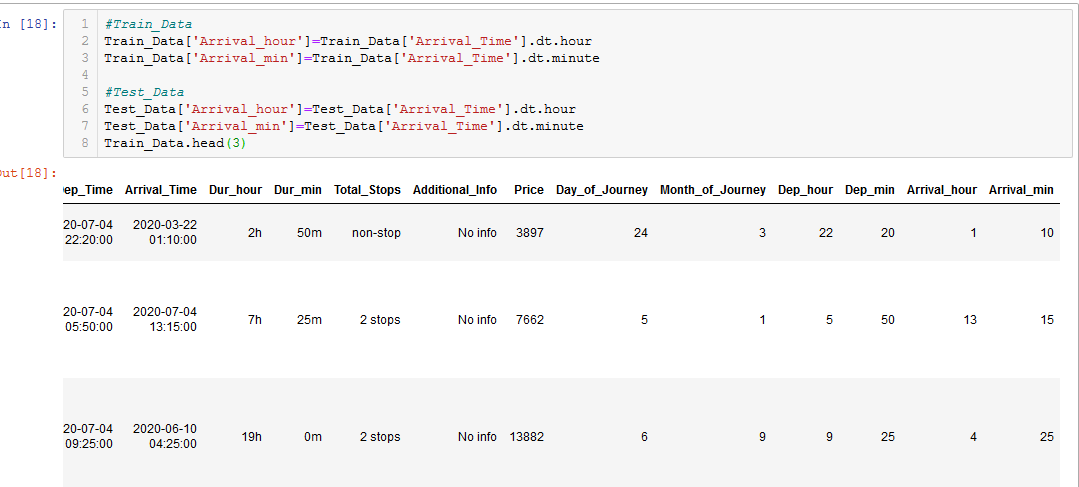
Deparature\_Time:-

Dep\_Time is in datatime dtype from it we create Dep\_hour and Dep\_min.



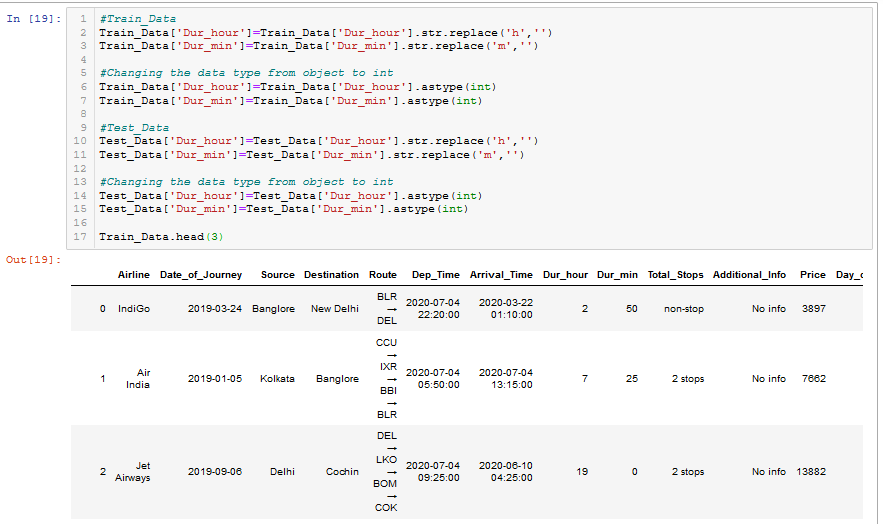
Arrival\_Time:-

We also changed Arrival\_Time dtype as datatime , from Arrival\_Time we create feature Arrival\_hour and Arrival\_min.

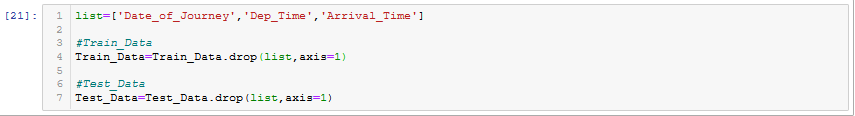


Duration\_hour & Duration\_minute:-

Duration\_hour and Duration\_min are in object dtype with h and m alphabet, so removing the alphabets first the change the dtype to int.



Now feature like 'Date\_of\_Journey', 'Dep\_Time’ and 'Arrival\_Time' are not required for Machine Learning as we derived Cleaned Data from these features.So we drop these features in the two datasets

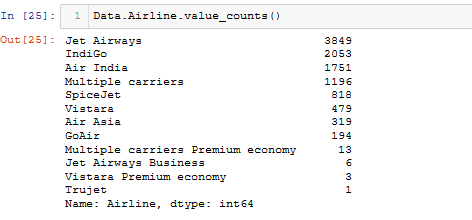


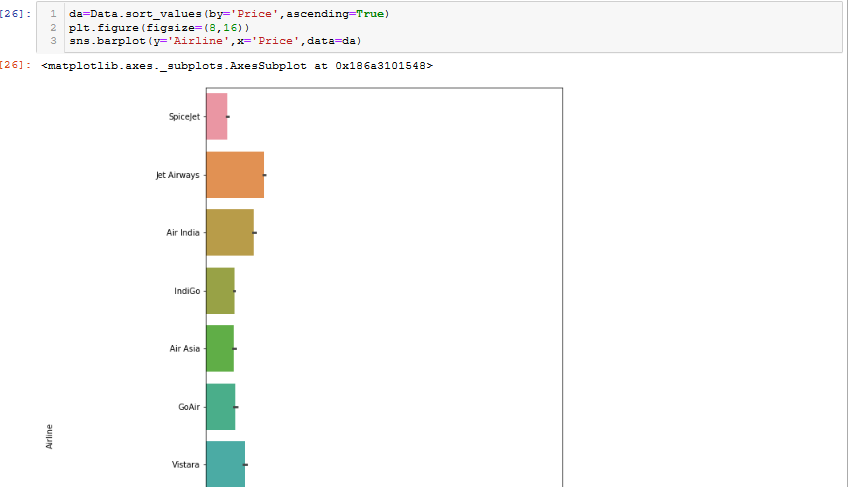
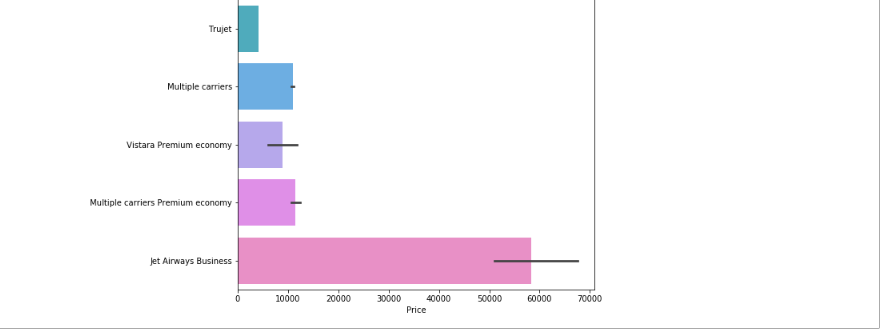
**We take Train\_Data as Data.**

**Univariate and Bivariate analysis:**

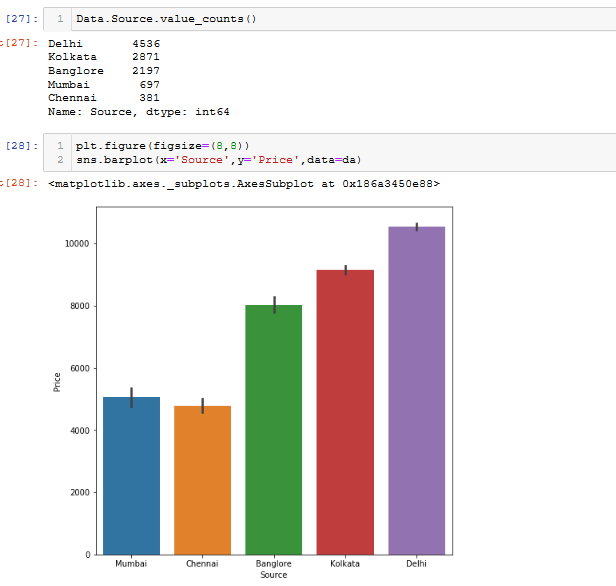
In univariate and bivariate analysis we analyse single features and compare two features so that we can deduce useful information about the features.

**Airllines**



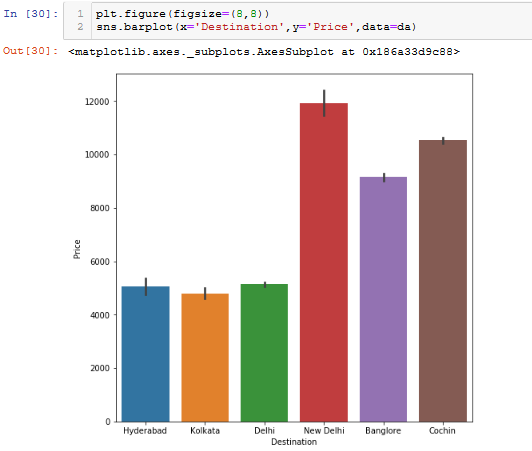


**Source**



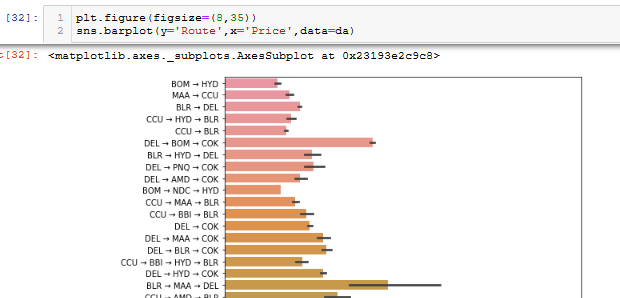
**Destination:**



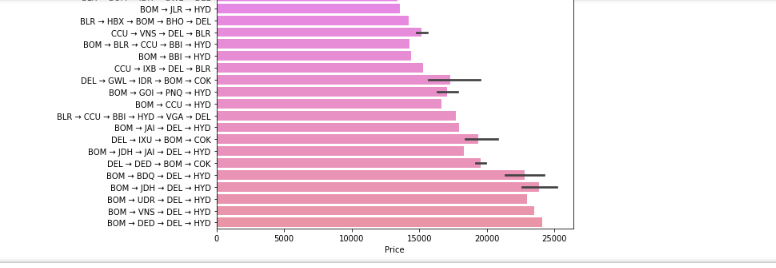


**Route:-**

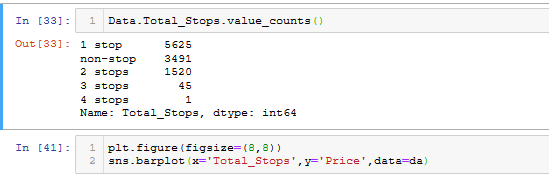
Lower Price Routes

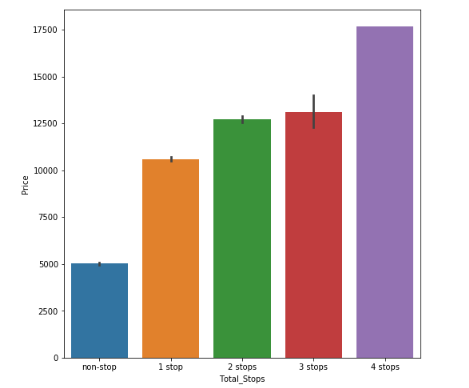


Higher Price Routes

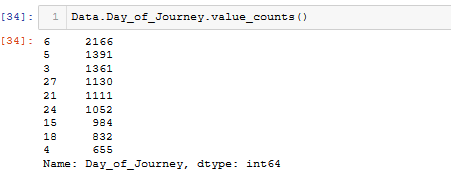


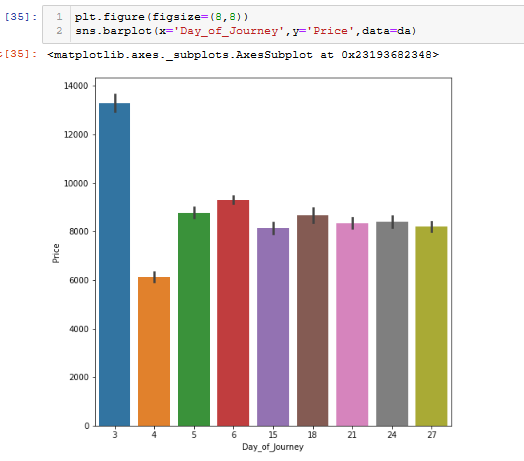
**Stops:-**

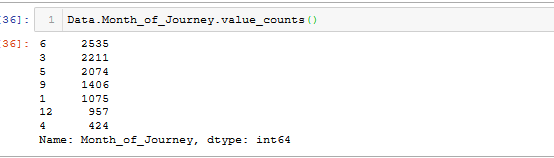


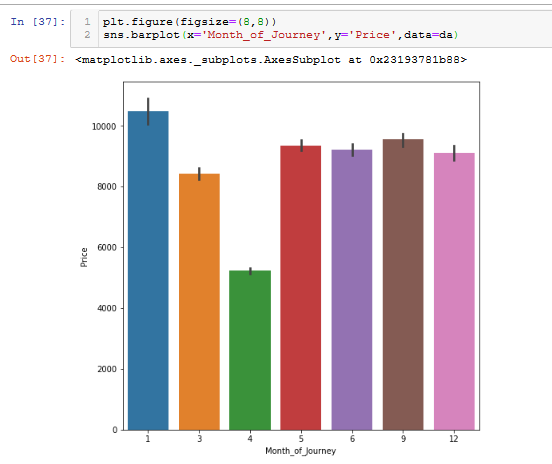


**Day\_of\_Journey:**









EDA Conclusion:-

In the Analysis we found that there are 10682 valid observations are found in the train dataset. Actually it is not the Flight Ticket for various airlines in between March to June 2019 it is the Dataset of January, March, April, May, June, September and December 2019.

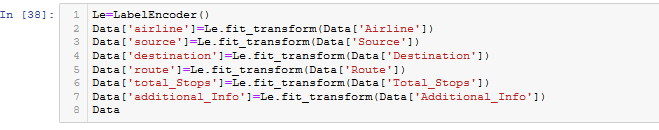
Among the 12 Airlines Jet Airways has highest i.e. 3849 number of Flights among the various cities followed by IndiGo, Air India, Multiple carriers and the Lowest number of flights i.e. 1 has been taken by Trujet but the Price of Jet Airways Business is highest.

Among the 5 Sources highest number of flights are sourced from Delhi i.e. 4536 followed by Kolkata and lowest number is 381 sourced from Chennai and also among the 6 destinations highest number of flights are destinated towards Cochin i.e also 4536 followed by Bangolore and lowest destination is Kolkata but The price of Flight from Delhi is very high as compared to other Sources and Chennai is lowest and if the Destination is New Delhi then the Price is Highest followed by Cochin and Bangalore and lowest is Kolkata.

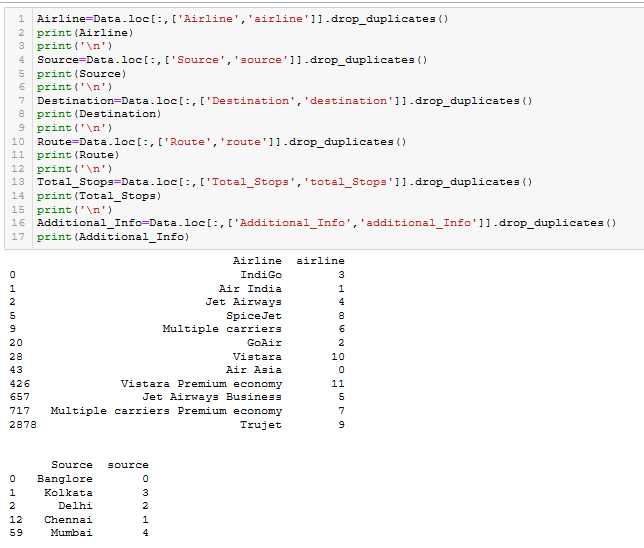
With increase in in stoppages the price of flight ticket increases and with decrease in stoppages ticket price decresases. When the Route is BOM->HYD the price of ticket is lowest and route is BOM->DED->DEL->HYD the price of ticket is highest.

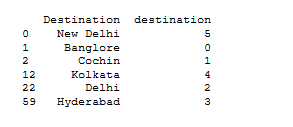
**Step-4**

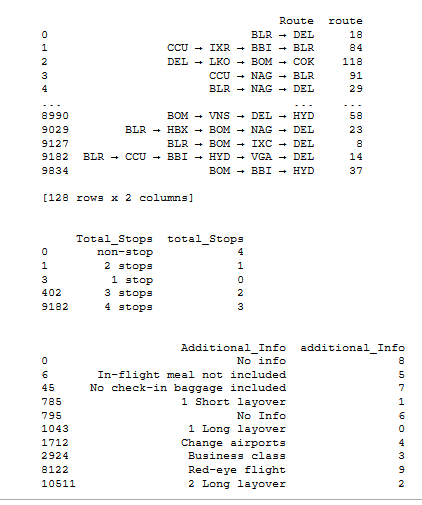
Machine Can understand only numerical values may it is a Regression or Classification Problem. There are some methods to Encode the categorical values to its numeric representation and among them LabelEncoder and OneHotEncoder are highly powerful encoder. So let’s encode the categorical features present in the Datasets.



We are encoding the Categorical values to other columns so that we save the categorical values and their numerical representations for future use.

Saving 

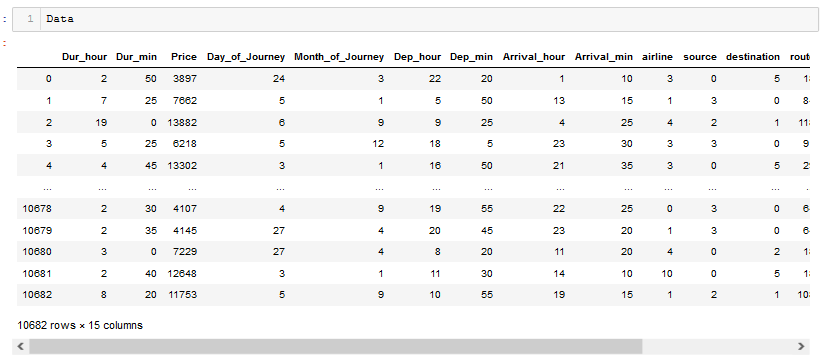


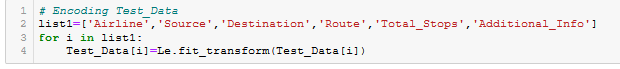


Now dropping the Categorical value columns as they are encoded in other columns.



Data after Encoding and dropping the Categorical value columns

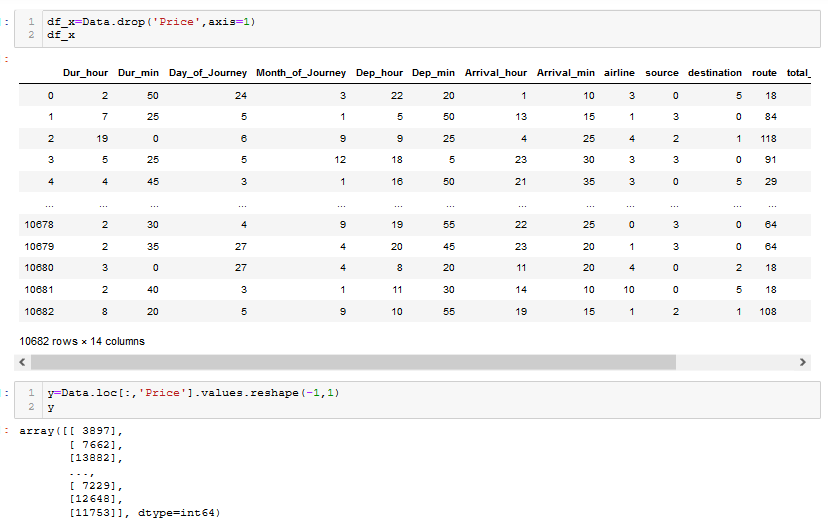




**Step-5**

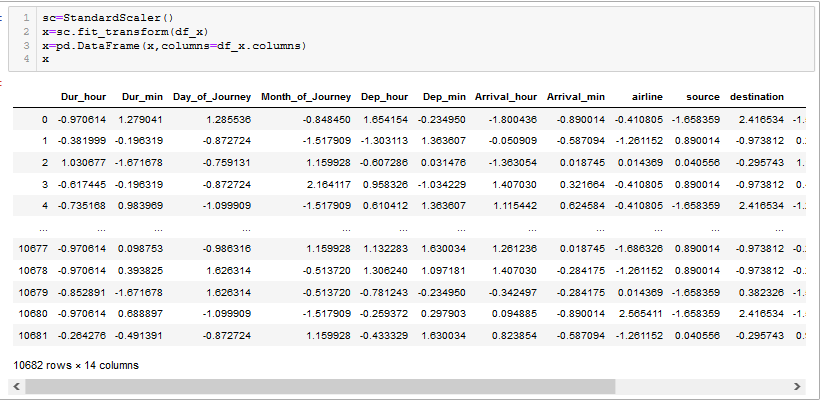
Supervised Machine Learning is the technique of accomplishing a task by providing training input and output pattern to the system. So that the Machine Learn and Predict as per the training.

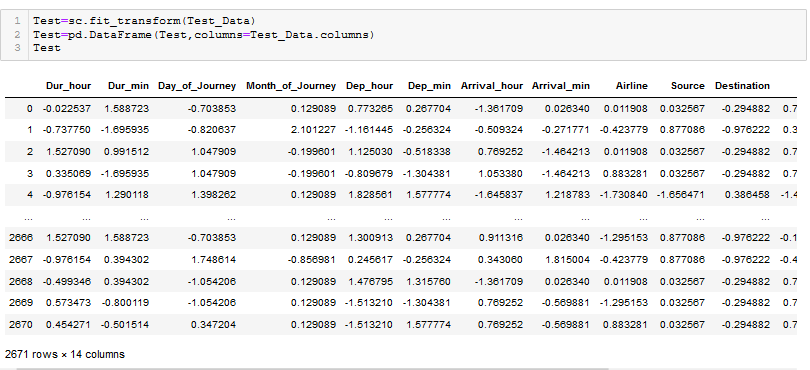
So we split Dataset into independent and dependent variable dataset



**Step-6** (Feature Scaling) –

**Feature scaling** is a method used to normalize the range of independent variables or **features** of data. In data processing, it is also known as data **normalization** and is generally performed during Machine Learning to standardize the data sets. Let’s Scale using StandardScaler



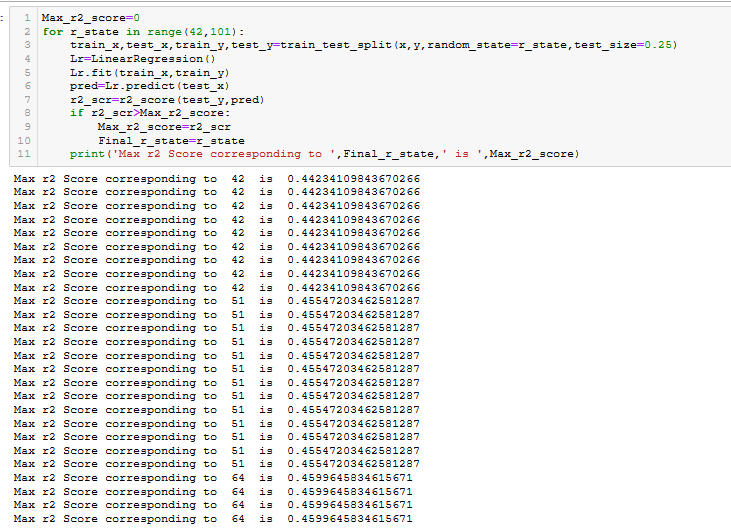
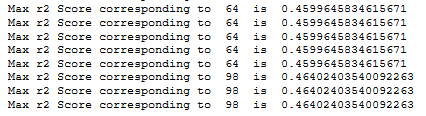


Independent and Dependent variable datasets are ready as x and y.

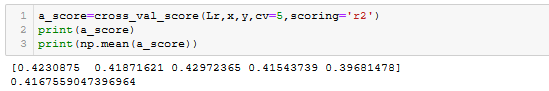
**Step-7-**(Machine Learning)

We will do machine learning to predict the price of Flight Ticket by the possible models and predict the Price by the model which will perform best.

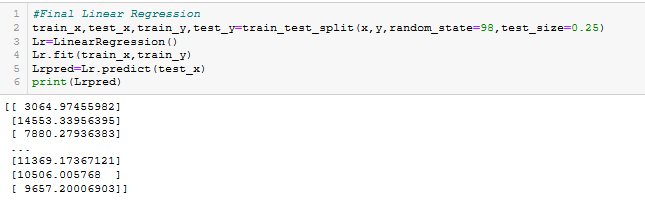
Now we will split the x and y datasets taking test size=25% and see which random\_state performs best result so that we take the random state for final model



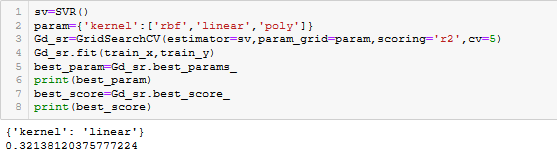
Checking if the upper model is Under or Overfitting and regularizing them



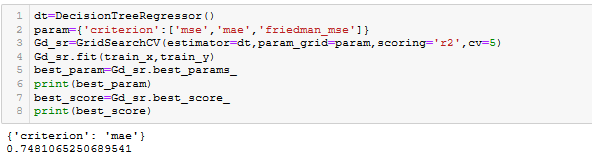
Now we Finalize Linear Regression taking the seed value 98 as it has the highest r2 score



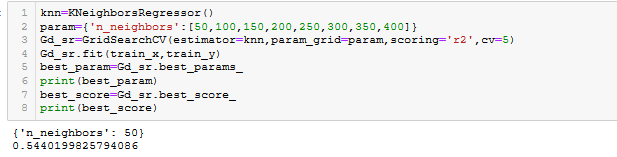
**Support Vector Machine**



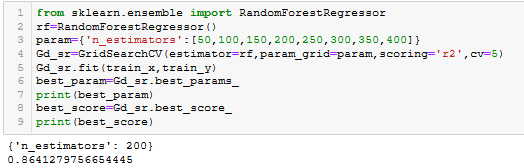
**Decision Tree Regressor**



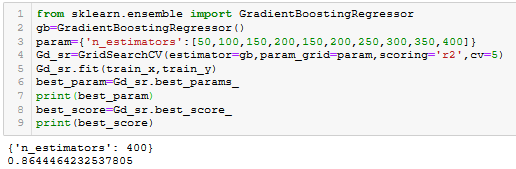
**KNeighborsRegressor**



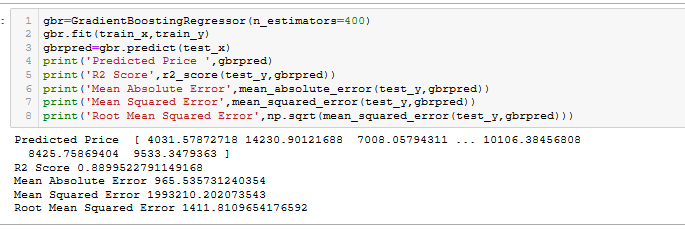
**RandomForestRegressor**



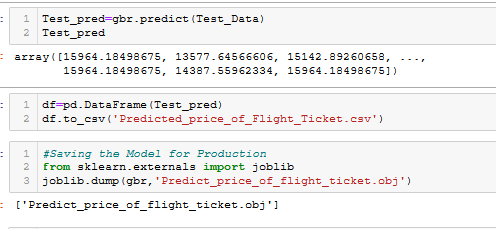
**GradientBoostingRegressor**



Making Final GradientBoosting Regressor



Predicting The Test\_Data and Save the Prediction:



Conclusion:

From all Machine Learning models GradientBoostingRegressor has highest r2 score i.e. 0.8644464232537805(86.44%) Score so we predicted the Price of Flight Ticket using GradientBoostingRegressor model.