

Airline companies use**complex algorithms** to calculate flight prices given various conditions present at that particular time. These methods take financial, marketing, and various social factors into account to predict flight prices. Nowadays, the number of people using flights has increased significantly.

Optimal timing for airline ticket purchasing from the consumer’s perspective is challenging principally because buyers have insufficient information for reasoning about future price movements.

[Why this Project?](https://achyutjoshi.github.io/btp/flightprices#why)

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics which they call "revenue management" or "yield management". The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on -

1. Time of purchase patterns (making sure last-minute purchases are expensive)
2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)



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 travellers 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

#### 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

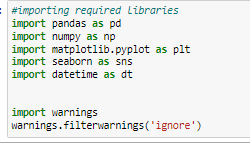
Price: The price of the ticket

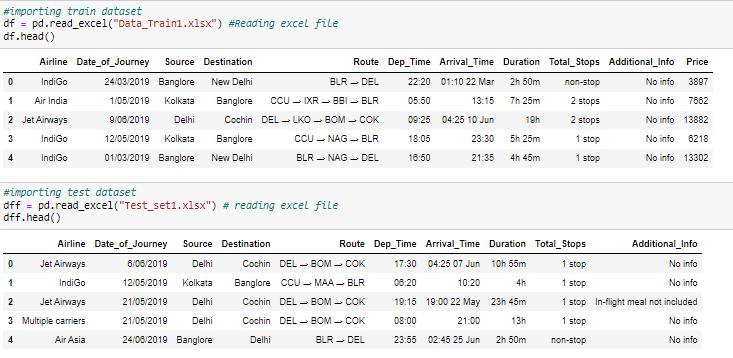
**I have to build a machine learning model to predict the price of the flight ticket.**

**Data Analysis: -**

The process of cleaning, transforming, and extracting data to discover the useful information for business decision making is called data analysis. It plays a role in making decisions more scientific and helping business operate more effectively.

Importing necessary libraries and dataset.

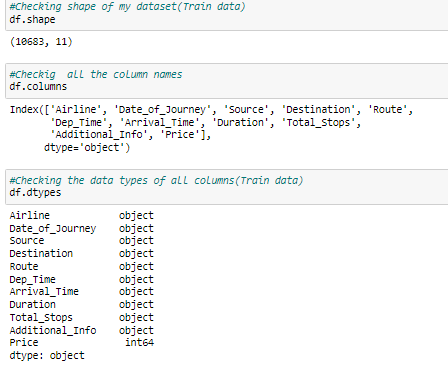




**Above are my train and test datasets and the target is seeming to be price so this particular problem is Regression Problem**

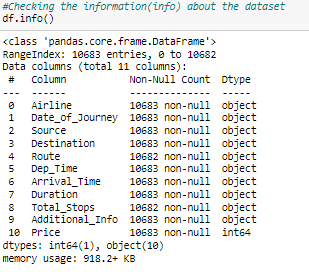
**Pre-Processing and EDA**

1.**Train dataset:**



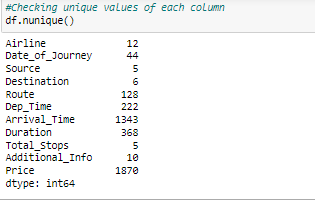
Except Price all other columns are object type data but I have to convert date of journey, dep\_time and arrival time columns from object to datetime type data.

Checking Dataset info.

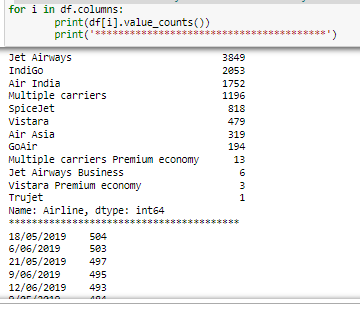


* There are nan values in Route and Total Stops columns.
* There is only one different data type is present in dataset i.e. is Integer datatype remaining all are object data type.

Now, I am checking Unique values in each column.

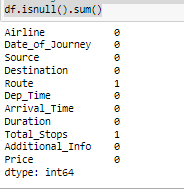


Lets check the value count of each column to see if there are any unexpected and unwanted entries present in the column.



Above are the value counts of each column.

Checking null values in the dataset

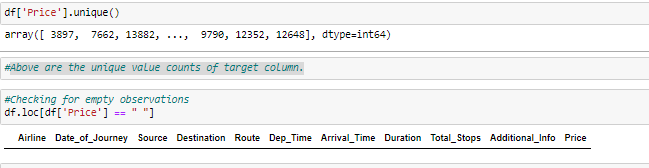


Imputation technique to replace nan values:



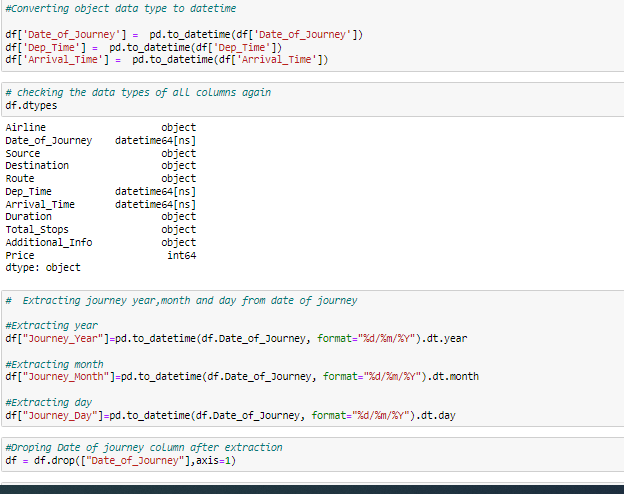
Now all the nan values has been replaced.

Checking Unique values in target column



There is no empty observations in my target columns.

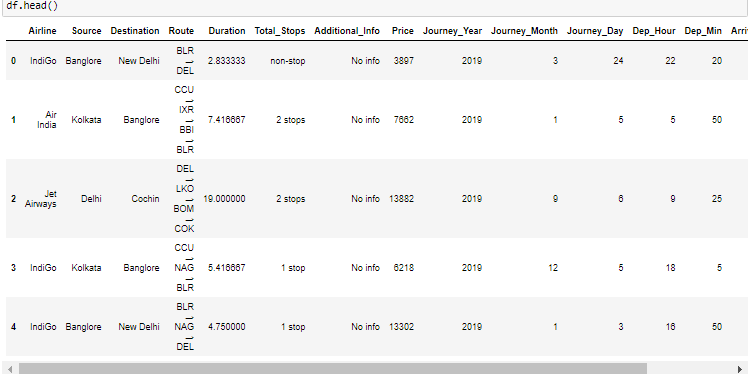
Feature Extraction:



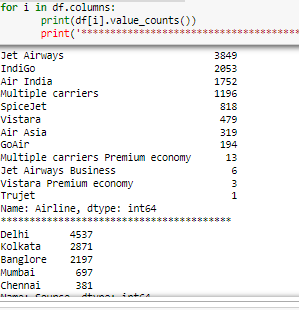
This particular duration column is the difference between Arrival\_time and Dep\_time but the given Duration is not in format (i.e.,-h-m) so let me extract proper duration column from the Arrival\_time and Dep\_time

.

Printing the dataset after extraction

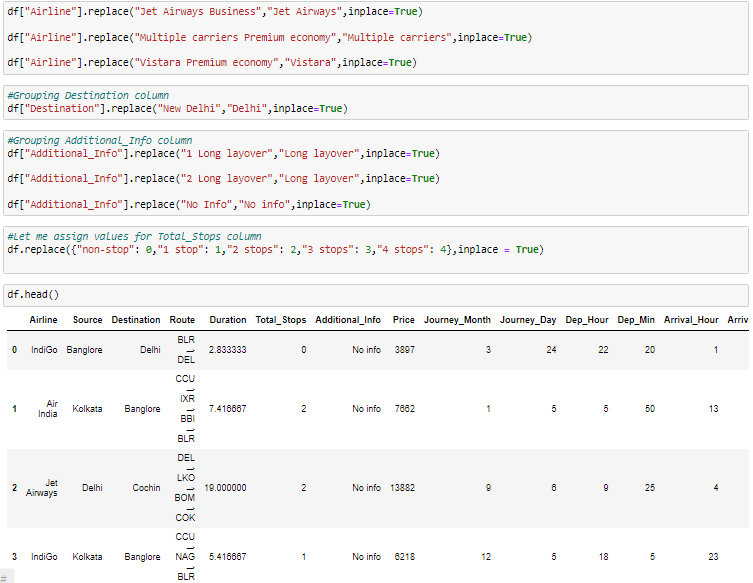


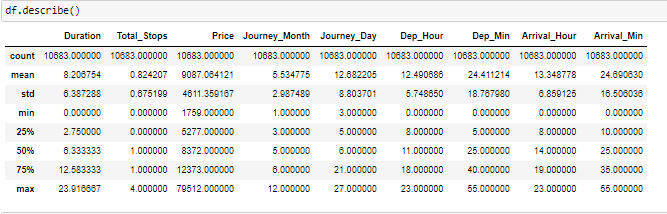
lets check the value counts now





By looking into the categorical columns i can do grouping as follows.



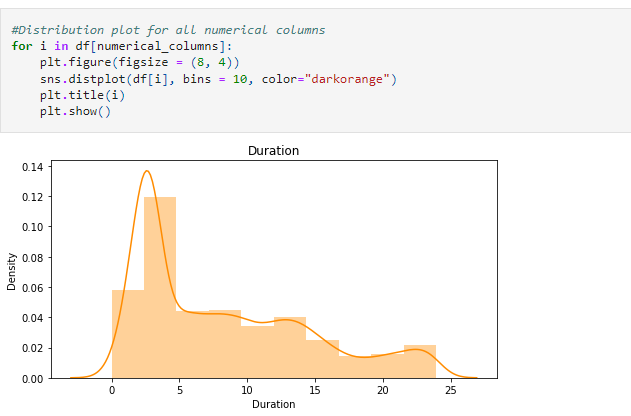


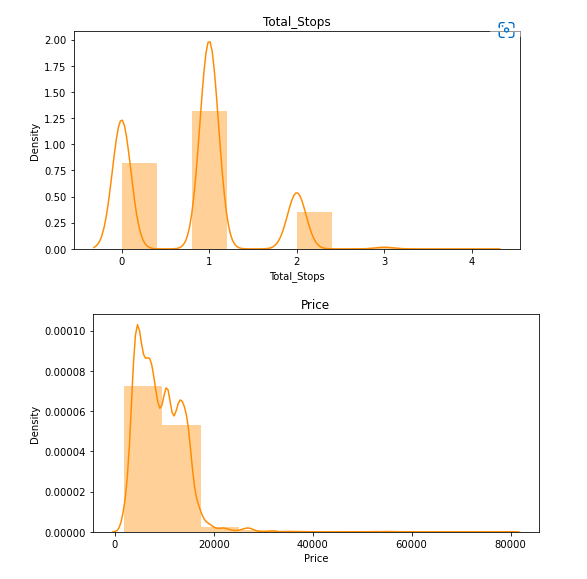
**Data Visualization:**

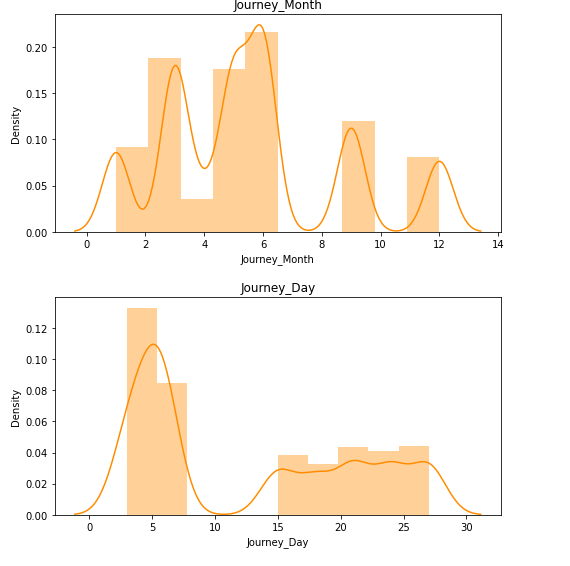
### Univariate Analysis:

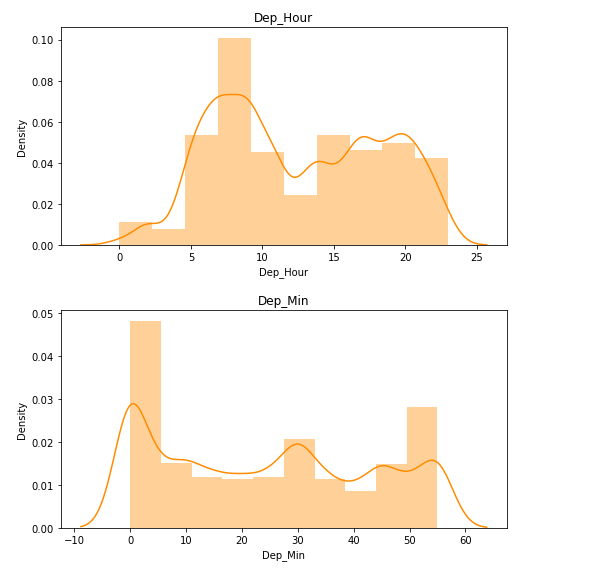


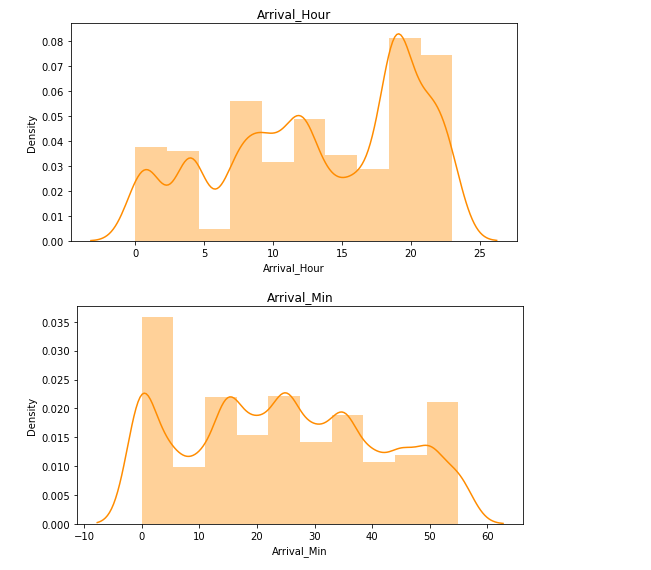
#### Univariate analysis for numerical columns:



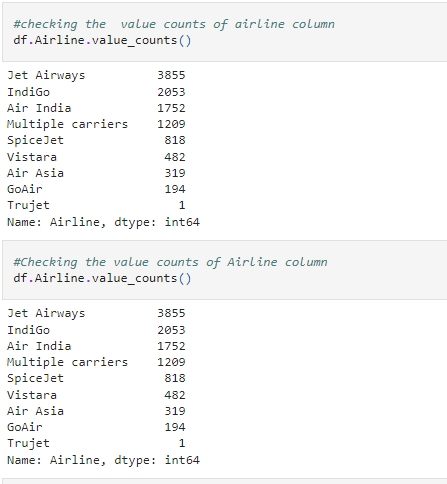




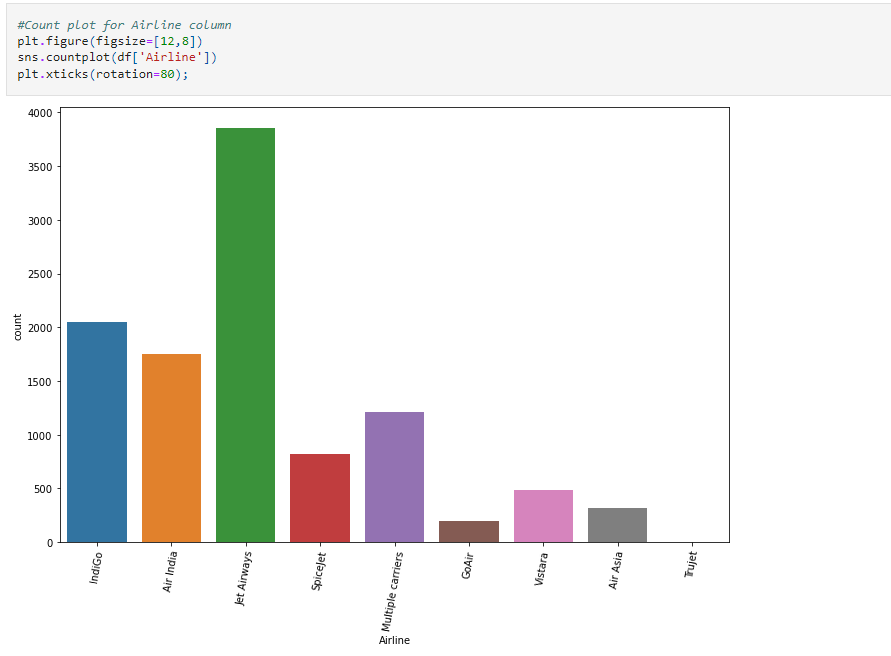




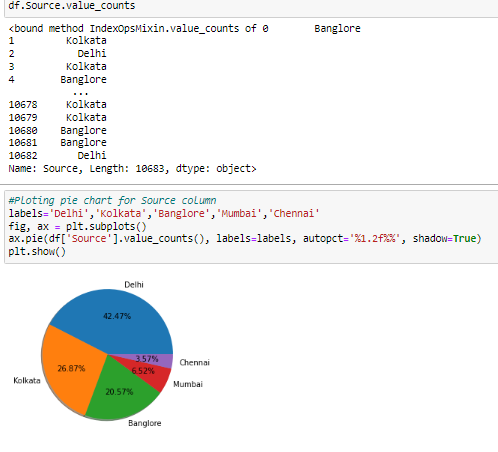
#### Univariate Analysis for categorical columns:

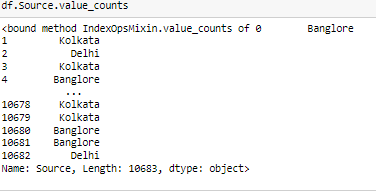


Count plot for Airline column



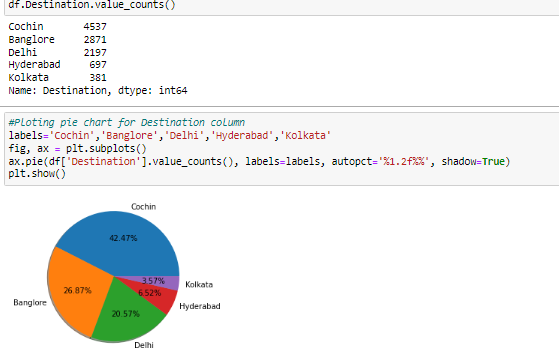
Checking the value-counts of source column





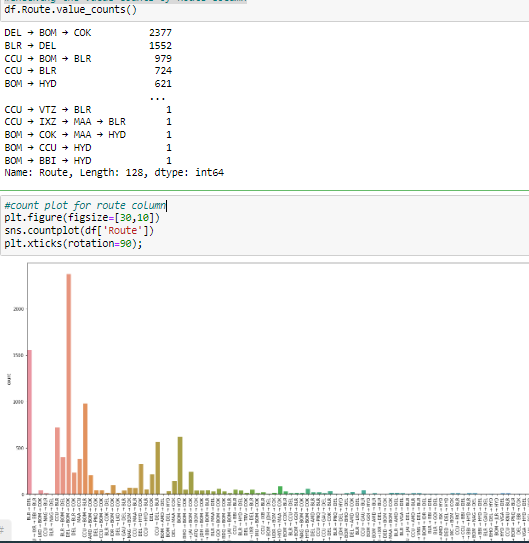
Delhi has maximum count which means Delhi is taken as source in maximum airlines. whereas Kolkata has minimum count in above given labels.

Checking the value counts of Destination column

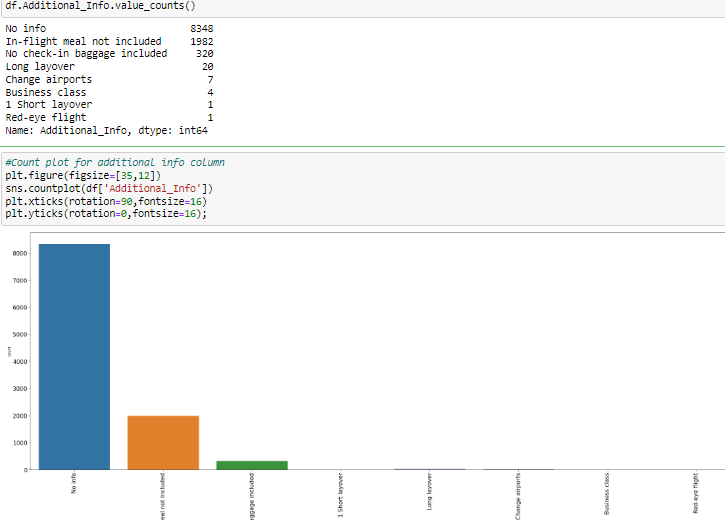


Cochin is maximum airlines for destination and for few airlines Kolkata is the destination.

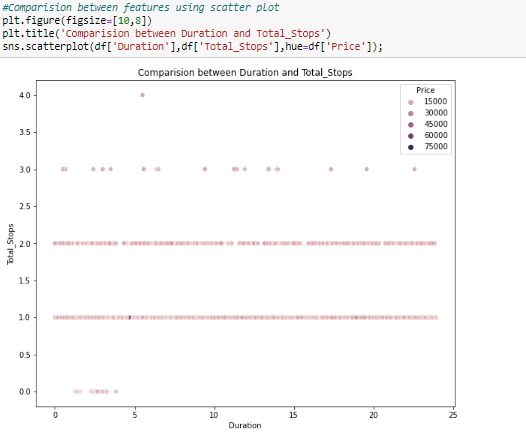
Checking the value counts of Route column

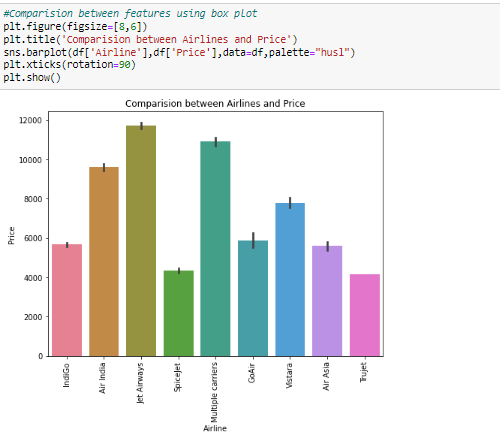


Checking the value counts of additional info column



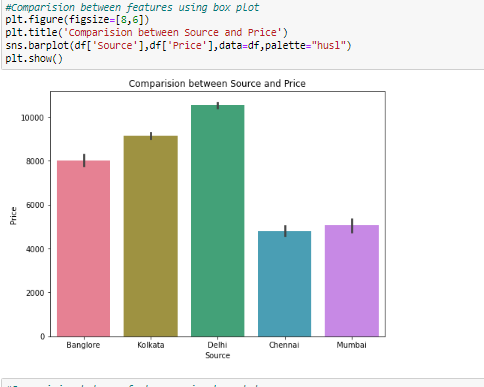
Bivariate Analysis:

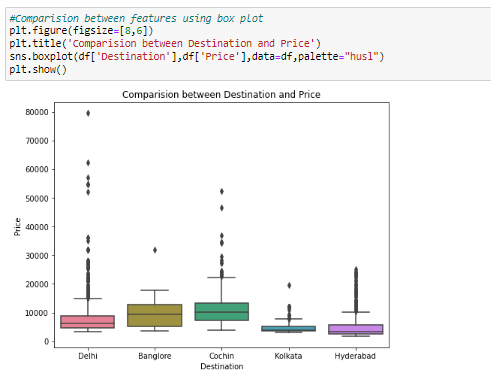




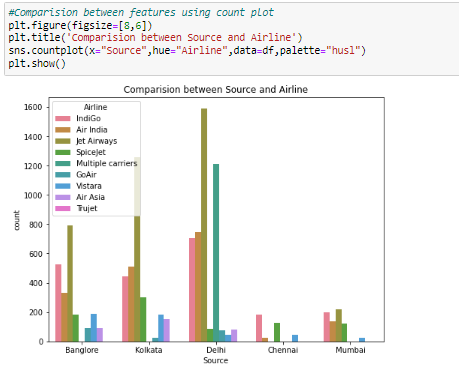
In Air India, Multiple carriers and Jet Airways the price is high they look costly.

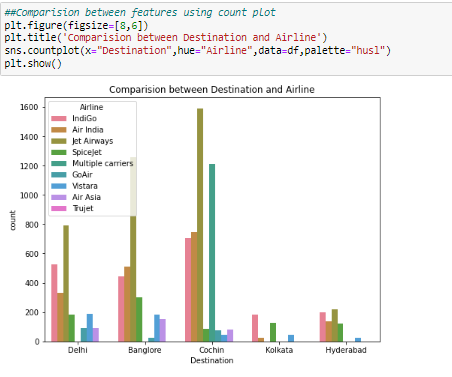
Trujet and Spicejet are cheaper compared to other airlines.

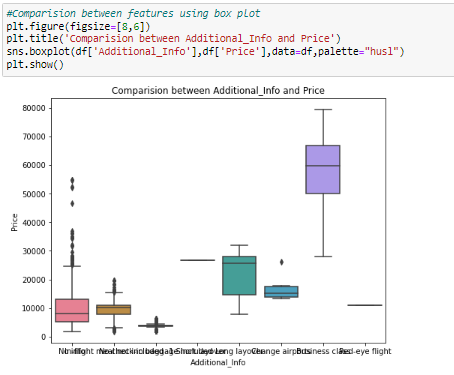


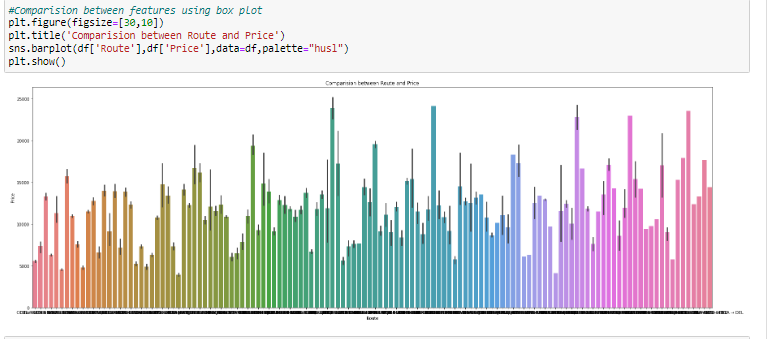


If Cochin is the destination then the price is more compared to other airlines.

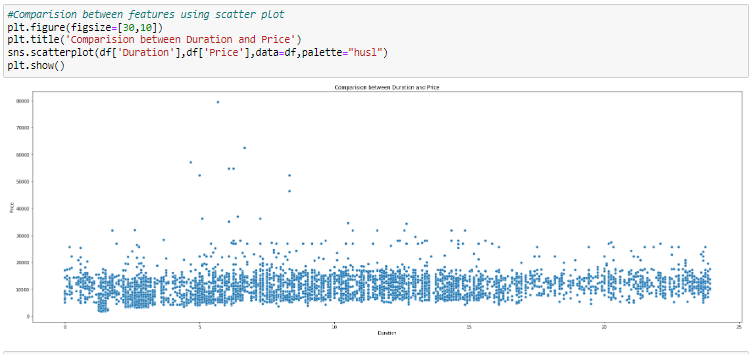


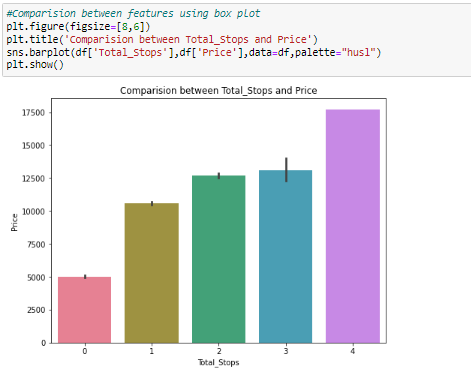


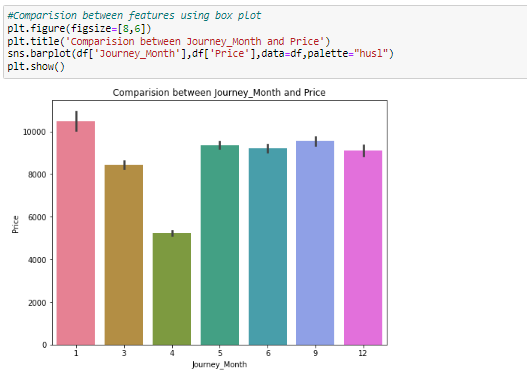


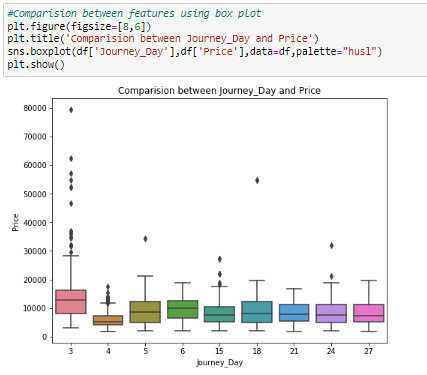


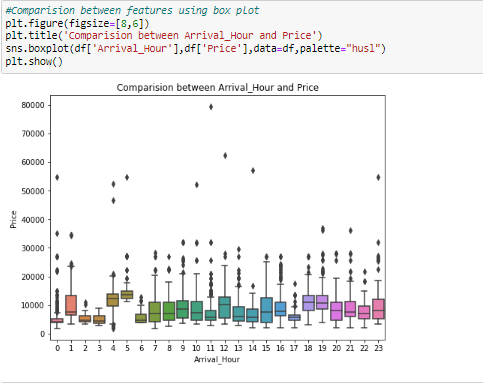
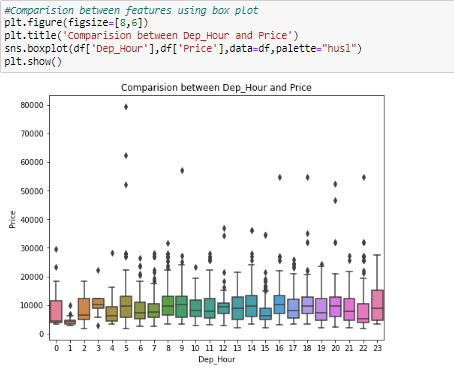
#BOM->JDH->DEL->HYD, BOM->DED->DEL->HYD and BOM->BOQ->DEL->HYD Routes has maximum price compared to other routes.



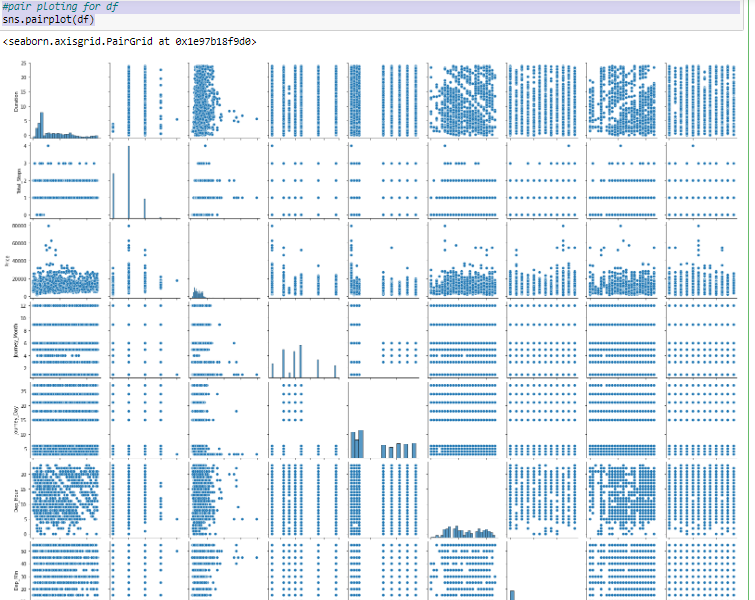








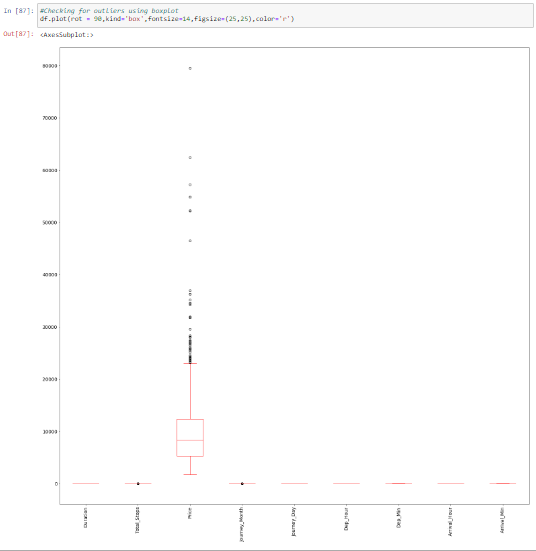
Multivariate Analysis:



Above are the pair plots of each pair of features.we can notice some outliers in each plot.I have to take care of those outliers.

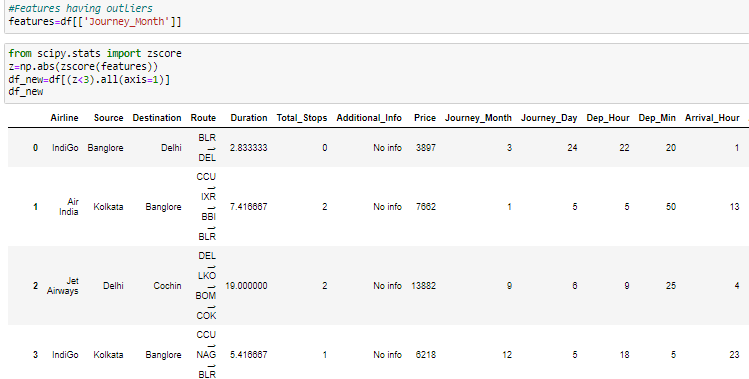
Checking for outliers:



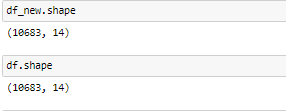


The outliers present in the Total\_stops,Price, and Journey\_Month Inn these i can't remove the outliers because i used every outliers present in these except Journey Month.

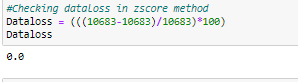
Removing Outliers: ----ZScore method:



10683 rows × 14 columns

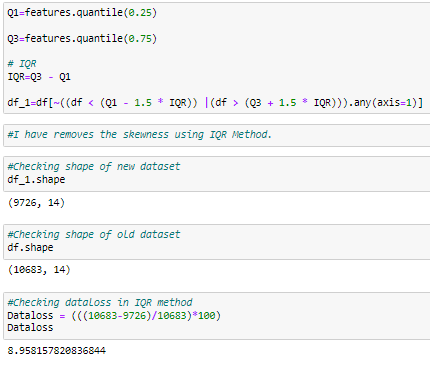


In my both datasets(new and old) have same rows and columns(10683,14).



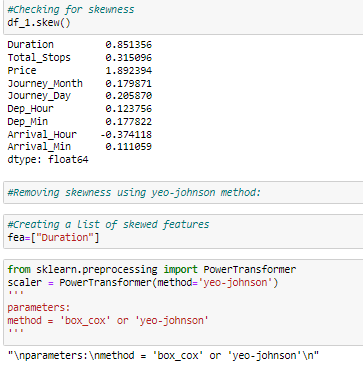
There is no dataloss is present in the zscore method which means it had not removes any outliers present in the datasets.

IQR Method:

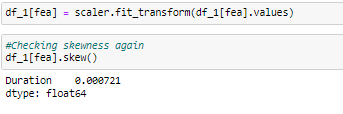


In IQR method the data loss is less than 10% so i can use IQR method to remove outliers.

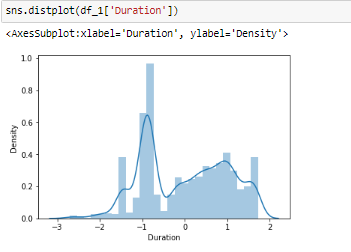
Checking for skewness:



Using yeo\_johnson method I have removed the skewness.

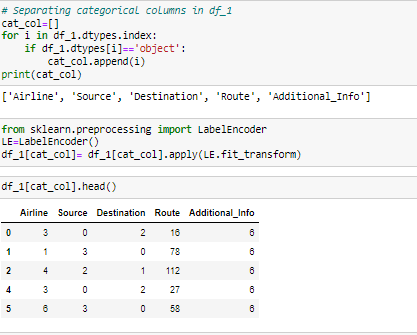


Now the skewness is removed, so let me proceed



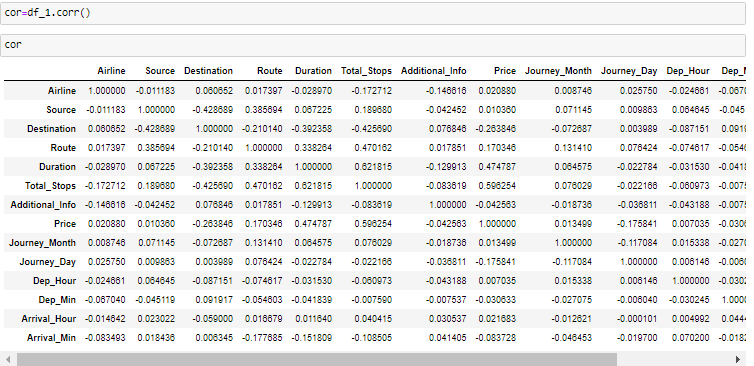
it looks like skewness is reduced.

Label Encoding:



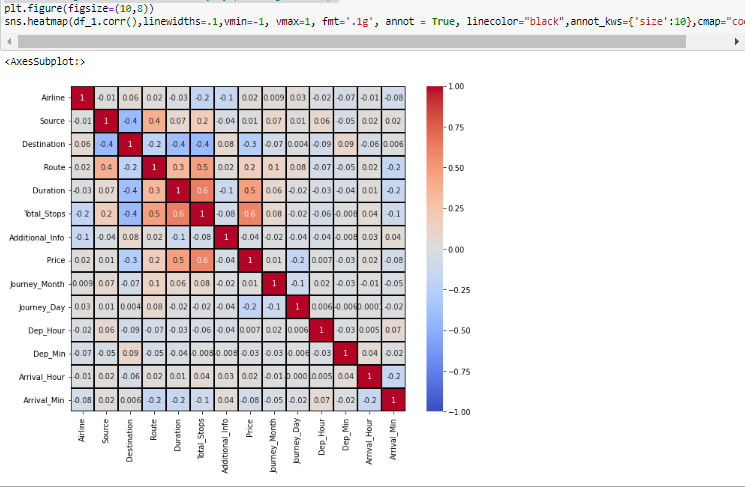
Using label encoder I have encoded the categorical columns.

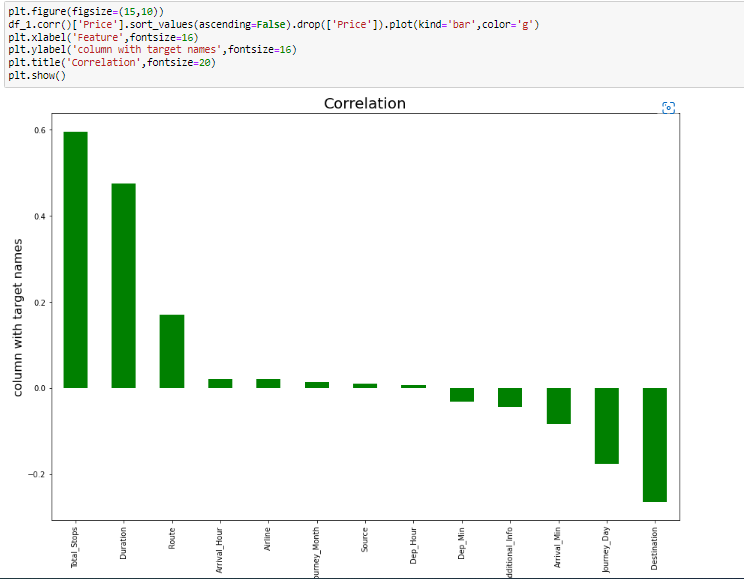
Checking correlation using heat map:



Above all are the correlations of the pair features. let me use the heatmap to know in better manner.

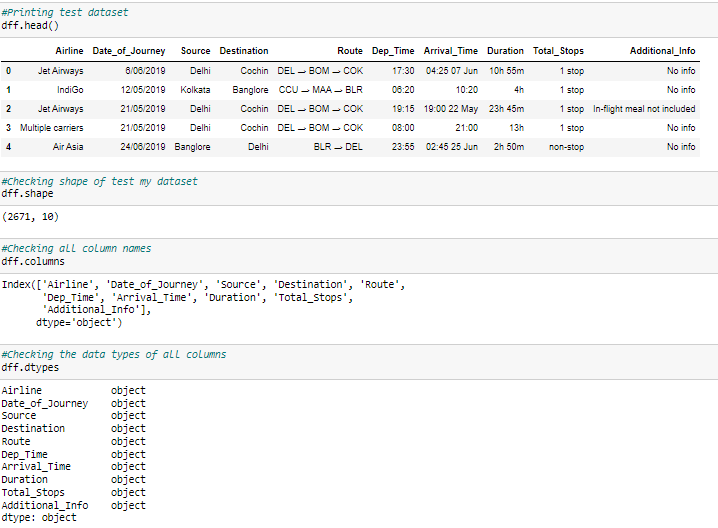
Visualizing the correlation matrix by plotting heat map.

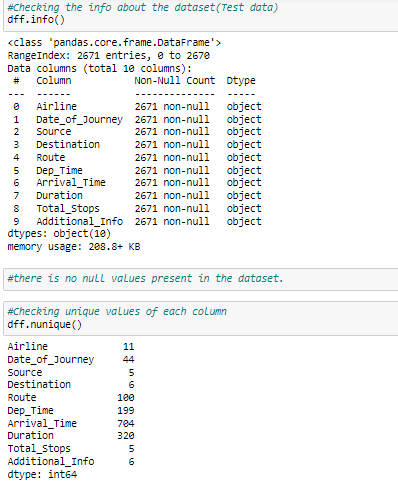


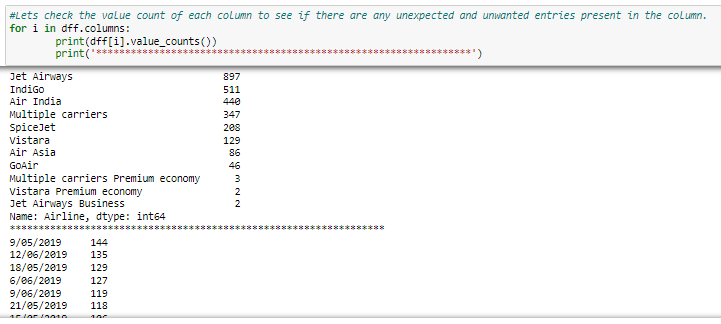


**Pre-Processing and EDA**

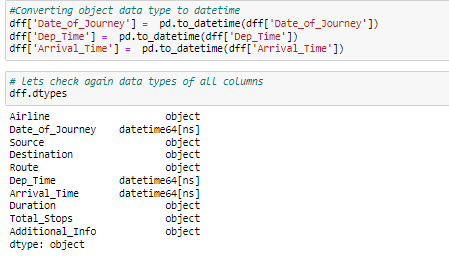
**2.Test dataset:**





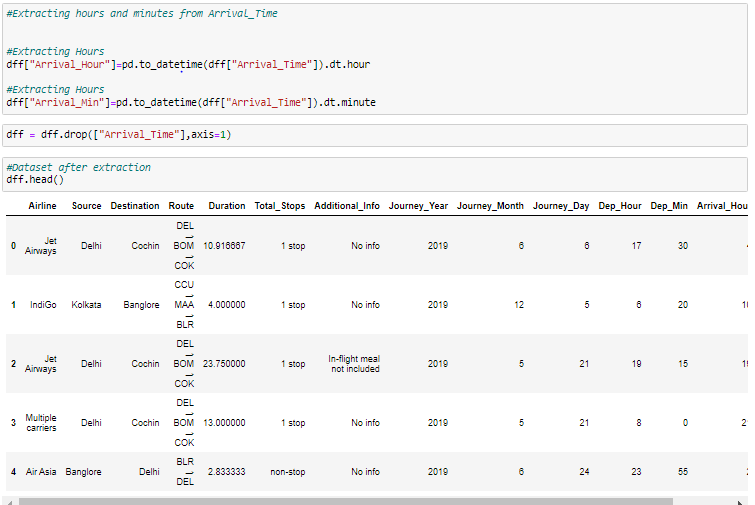


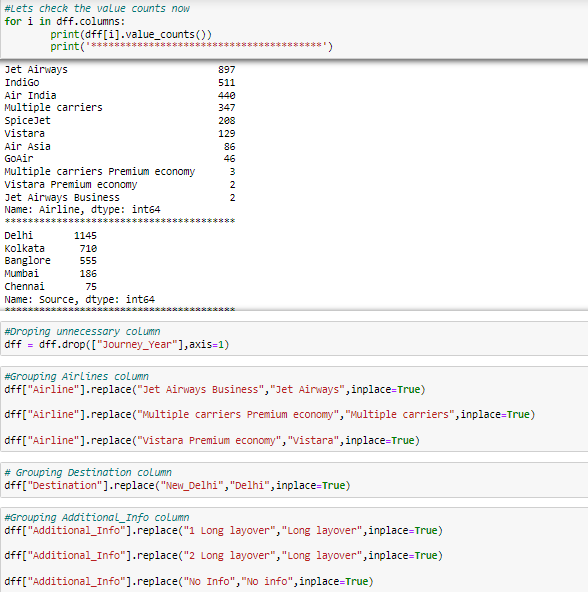
Feature Extraction:

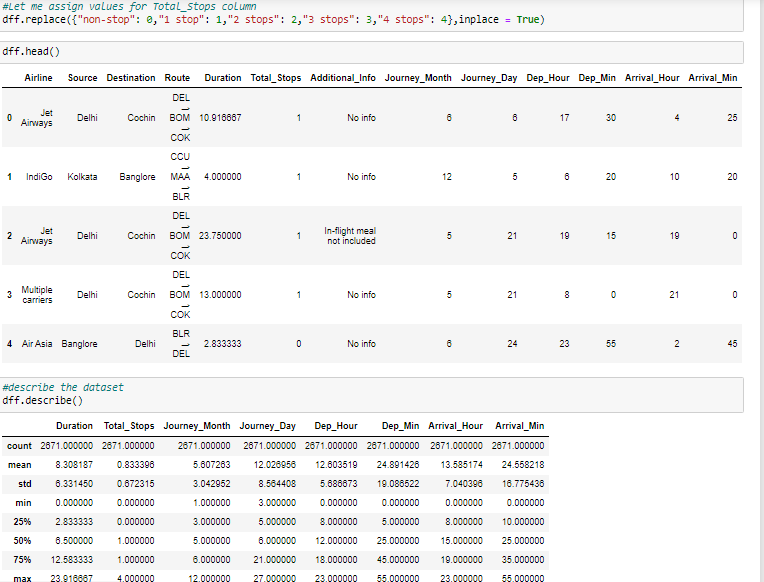


Now the data type has changed and have datetime and object data types.





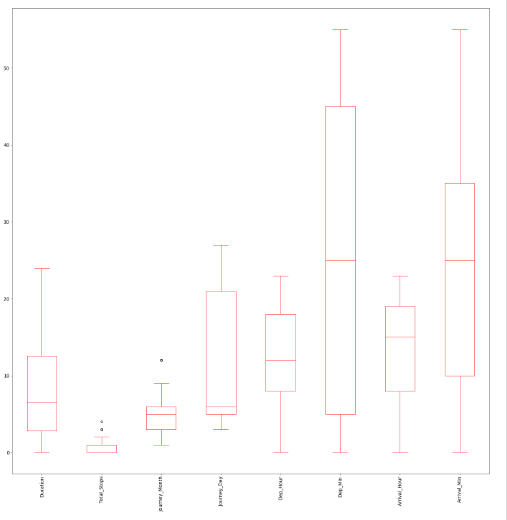




Above is the statistics about the dataset. The mean and the 2nd quantile values are not same so there is extreme outliers in the dataset.

Checking for outliers:

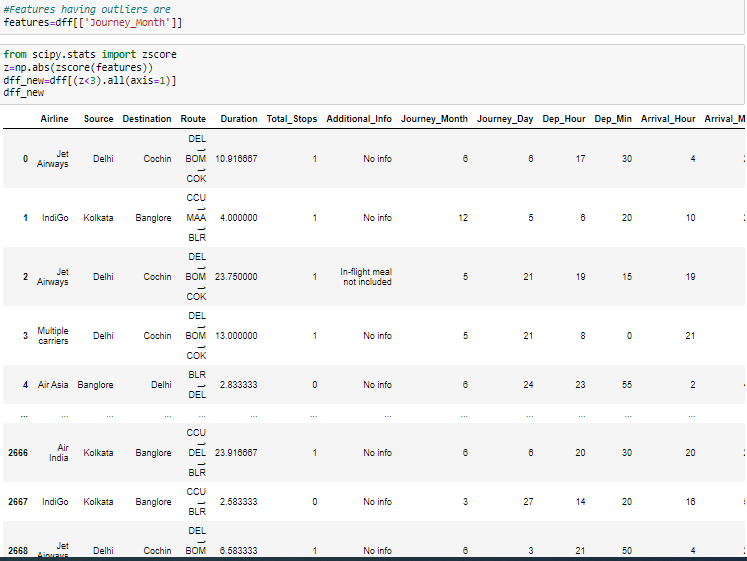




The outliers are present in Total\_stops,journey\_month but we can't remove outliers of total\_stops because it is categorical column .

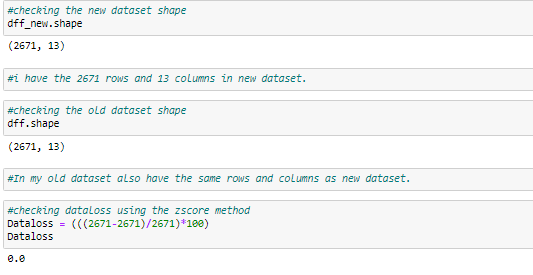
Removing Outliers:

Zscore method:



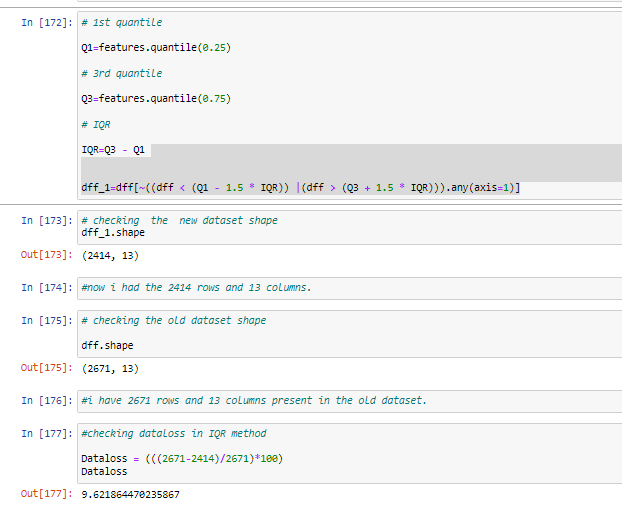
2671 rows × 13 columns

Above skewness is removed by zscore method.



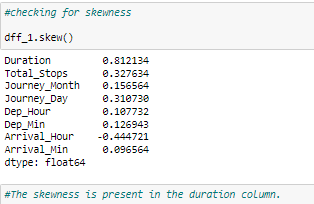
There is no dataloss in zscore method which means it has not removed any outliers

IQR Method:

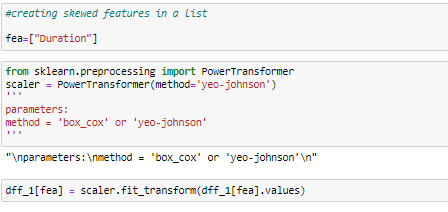


By using the IQR method i have dataloss is less than 10% so i can use the IQR method to remove outliers.

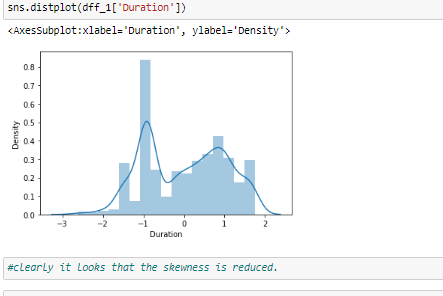
Checking for Skewness;



Removing skewness using yeo-johnson method:

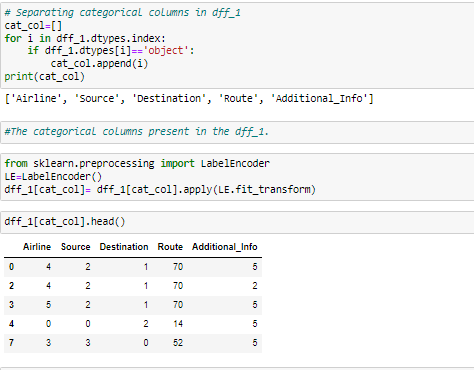


checking the skewness again



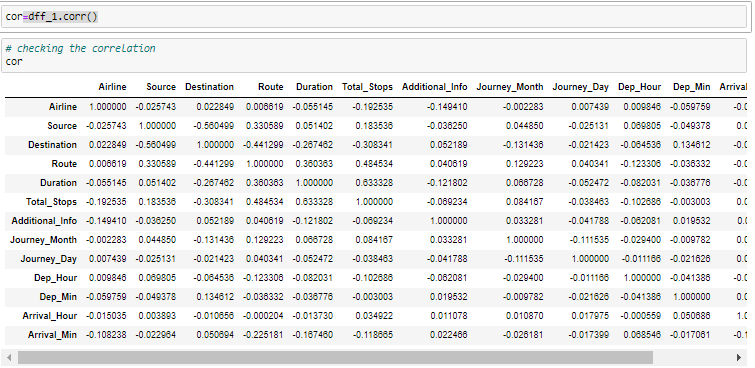
Skewness is removed.

Label Encoding:



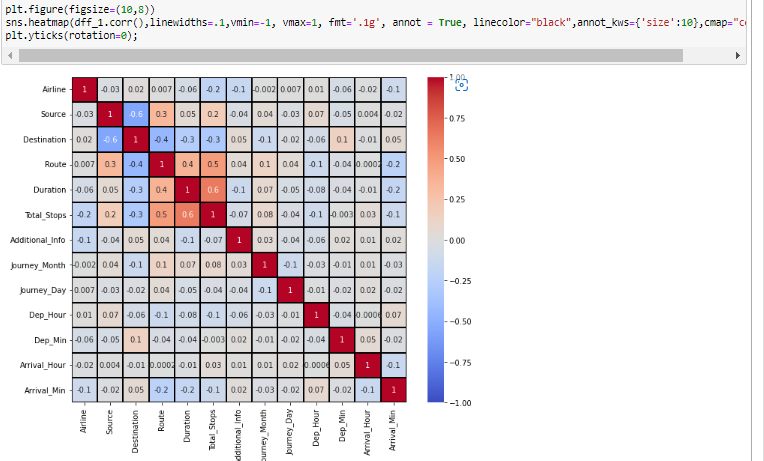
I have encoded the categorical columns using the label encoder.

Checking correlation using heat map:



Above are the correlations of all the pair of features.

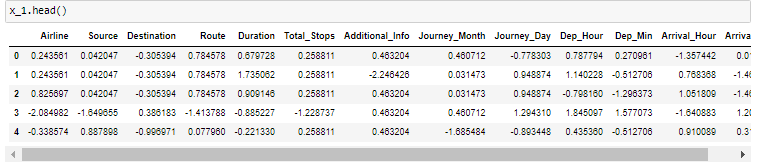
visualizing the correlation matrix by plotting heat map



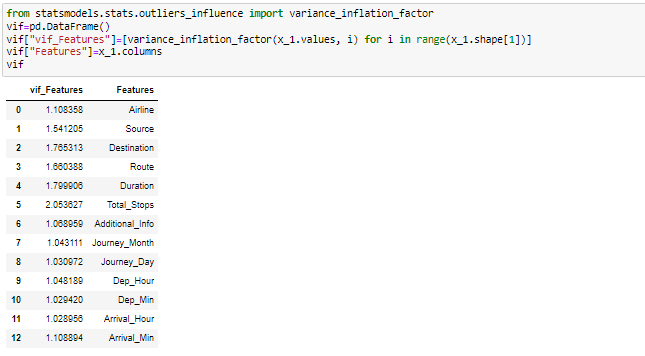
Scaling the data using standard scaler:



I had done the scaled for my data using the standard scaler.

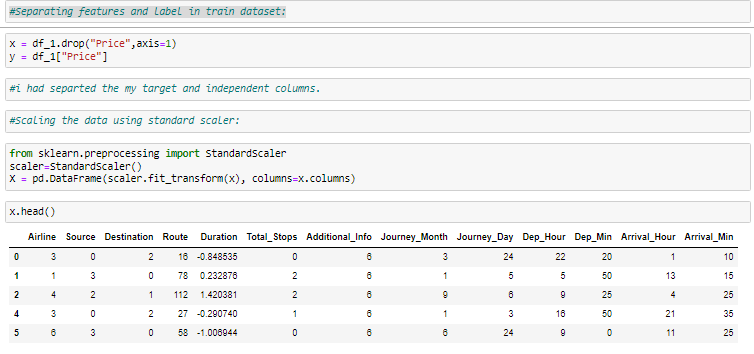


Checking for multicollinearity issue using VIF:

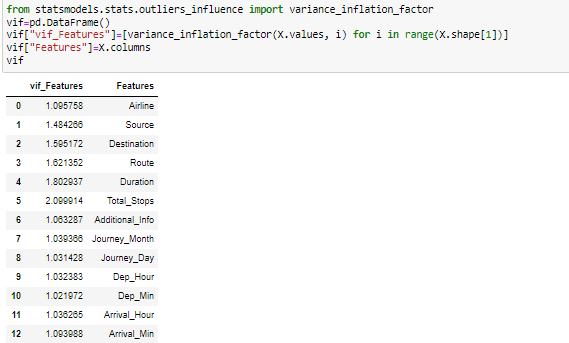


There is no multicollinearity issue in this dataset

#MODEL BUILDING FROM TRAIN DATASET:



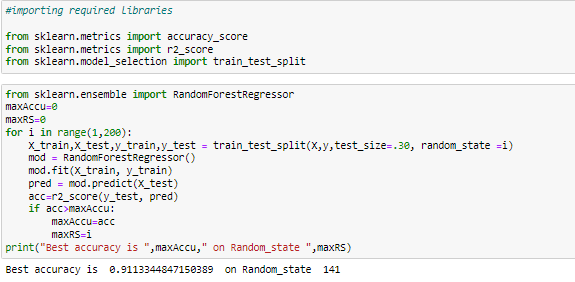
Checking for multicollinearity issue using VIF:



There is no multicollinearity issue in the dataset.

**Building Machine Learning Models:**

Finding Best Random State and Accuracy:

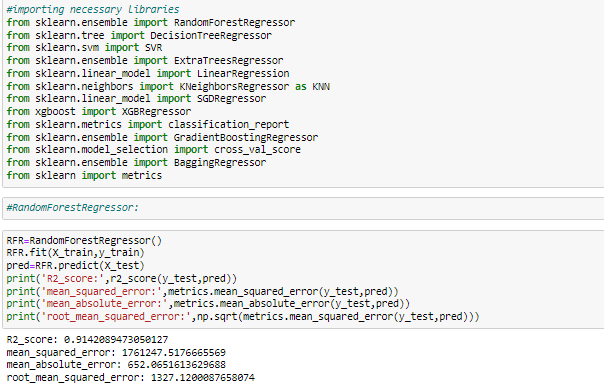


Now I got the best accuracy and random state.

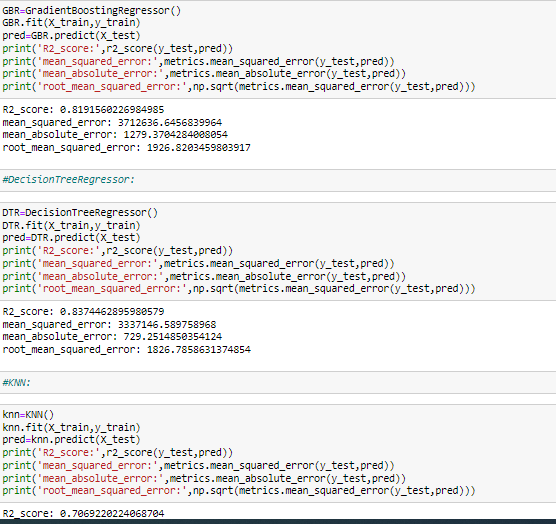
Creating train test split.

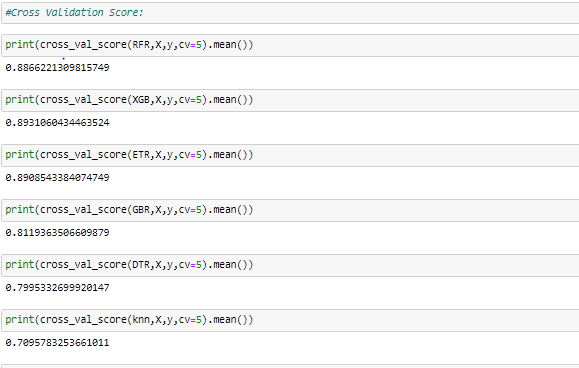


Regression Algorithms:





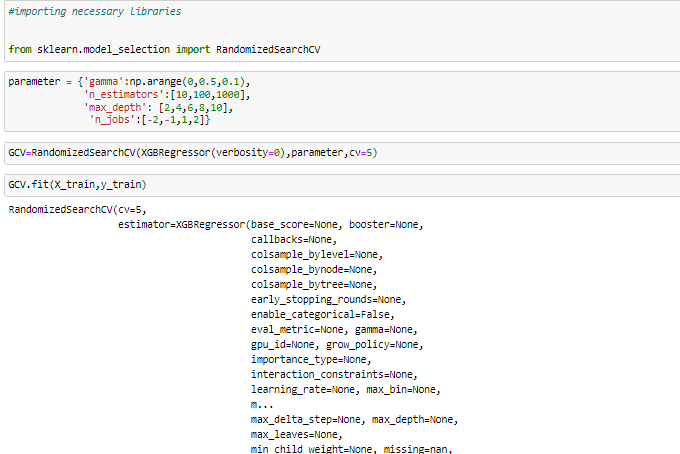


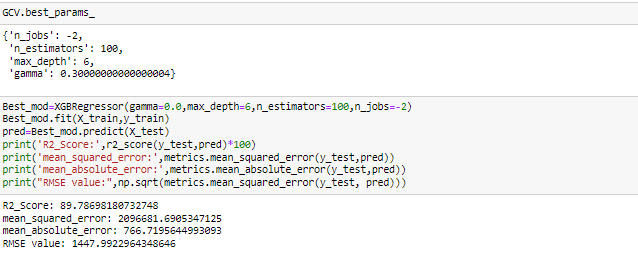


Above all are the cross validation scores of RFR,XGB,ETR,GBR,DTR and KNN.

By looking into the difference of model accuracy and cross validation score i got least difference for XGB.So XGB is my best model and i have to tune it to get better accuracy.

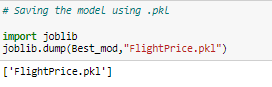
Hyper Parameter Tuning:

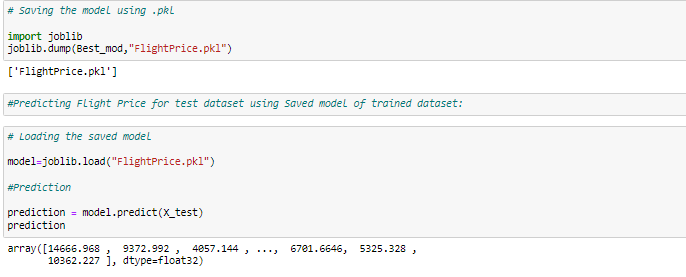


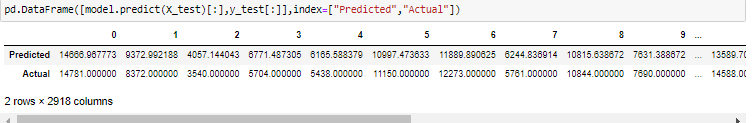


This is my model after tuning.I got 89.79% as r2\_score which is best.But the accuracy has not increased it has remained same.

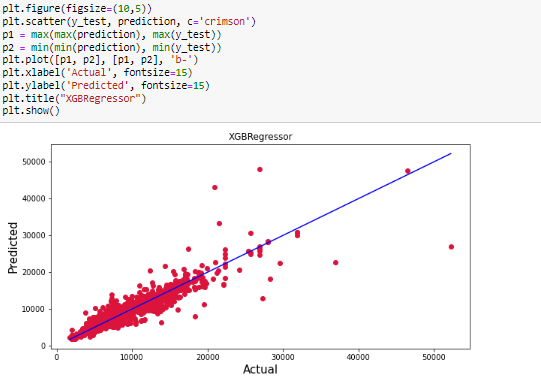
Saving the model:



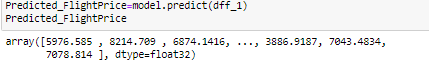




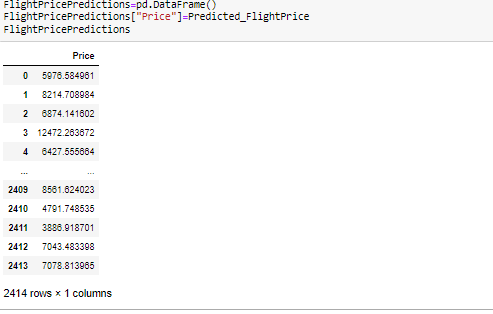
Above are the predicted values and the actual values but they are alomost similar.



Predicting flight price using cleaned test dataset dff\_1



Making dataframe for predicted Price





saves the predicted values in form of csv file.

**Concluding Remarks:**

1.Delhi has maximum count which means Delhi is taken as source in maximum airlines. whereas Kolkata has minimum count .

2.Cochin is maximum airlines for destination and for few airlines Kolkata is the destination.

3.In Air India, Multiple carriers and Jet Airways the price is high they look costly.

4.Trujet and SpiceJet are cheaper compared to other airlines.

5. Departure Hour has no proper relation with price.

6. Predicted values and the actual values of price are almost similar.