1. **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 travelers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between 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 was 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

In this problem, I am going to predict the price of the flights with the help of Duration, Arrival, and departure time, Total stops, Destination, source

It is a Standard supervised regression problem. There are 11

independents variable and 1 target variable(Price).

1. **Data Analysis**

Training data contains 10683 rows and 11 columns(including target variable price).

First, I will perform EDA to know the relation of different independent variables with the target variable with the help of data visualization.

Then after building the best model will use the same model for predicting

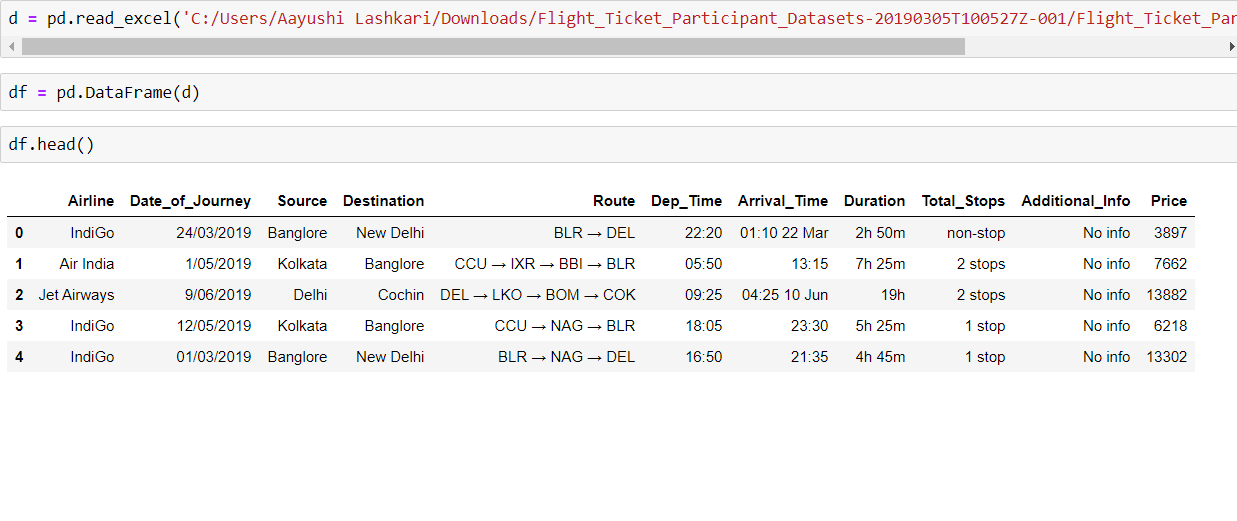
the price for test data.

**2.1 Data Description**

 First, the Dataset is imported from the source location and then converted into the Dataframe for further data analysis.

df.head is showing the first five rows of the data frame.

All the columns except the target variable are in object datatype.



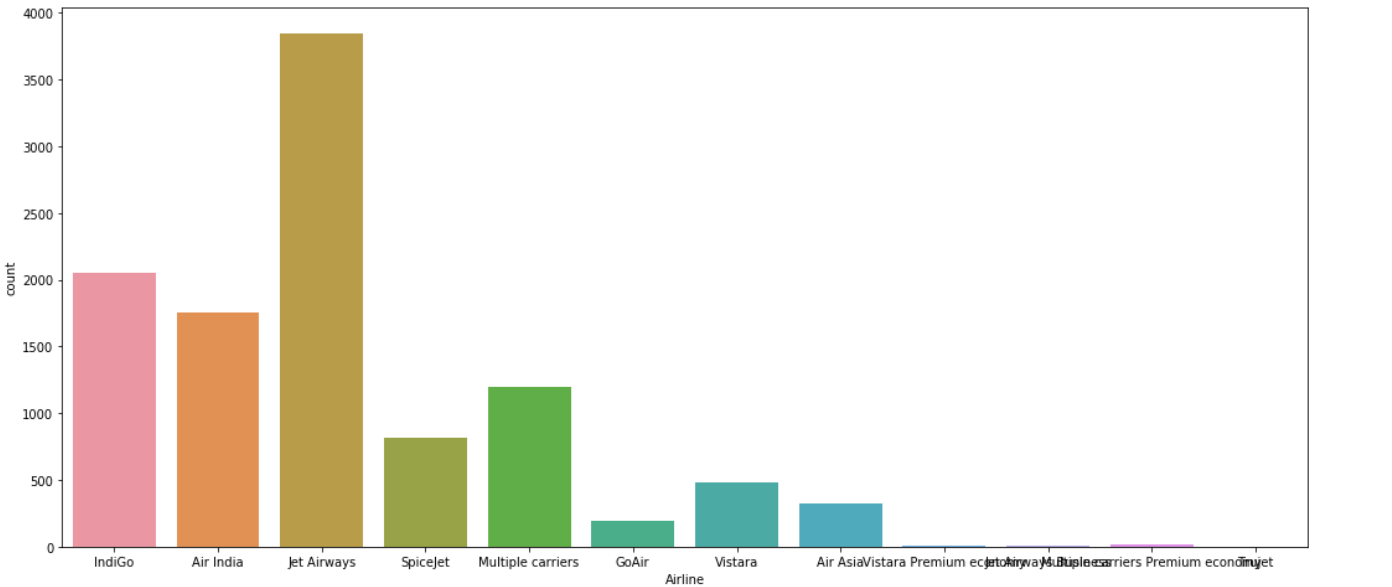
**2.2 Data Visualisation**

In this Section First, we performed univariate analysis

followed by bivariate and multivariate analysis.

**2.2.1 Airlines**

There are a total of 12 airlines. Where JetAirways has the highest number of flights followed by Indigo and AirIndia.

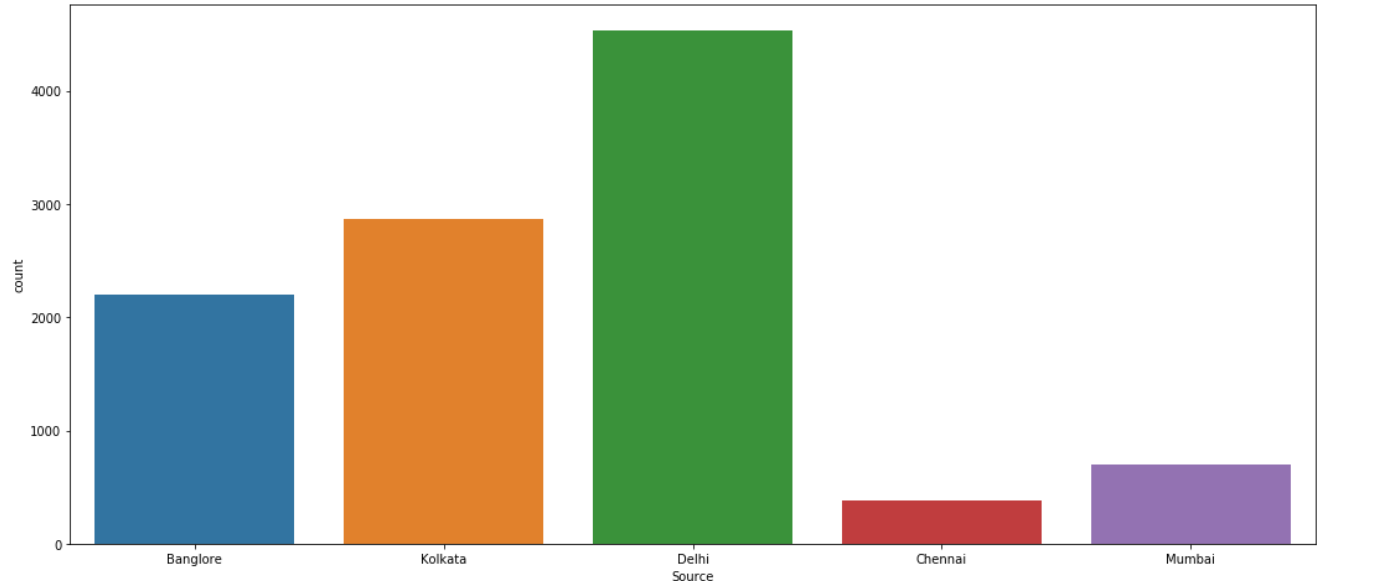


* + 1. **Source**

Source locations of the flights are Mumbai, Delhi, Kolkata, Chennai, Bangalore in the dataset

Most of the flights fly from Delhi followed by Kolkata, Bangalore,

Mumbai and at last Chennai.



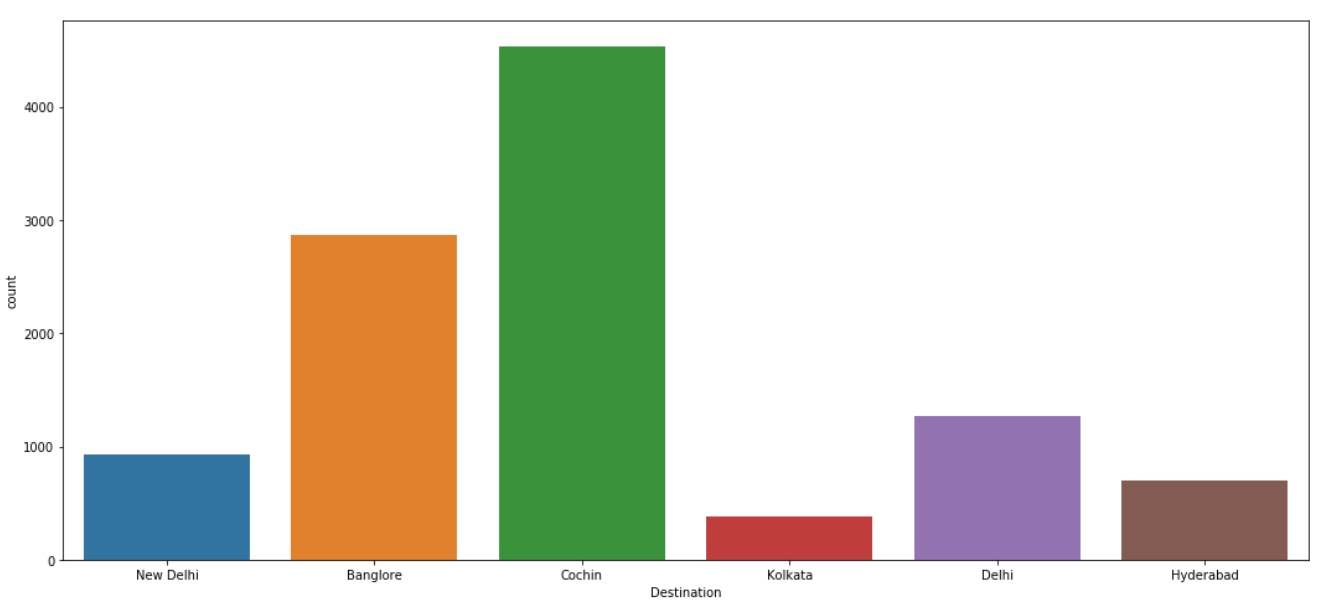
* + 1. **Destination**

Destination locations of the flights are New Delhi, Bangalore, Cochin,

Kolkata, Delhi, Hyderabad.

And most of the flights' lands in Cochin. Least number of flights' lands in

Kolkata.



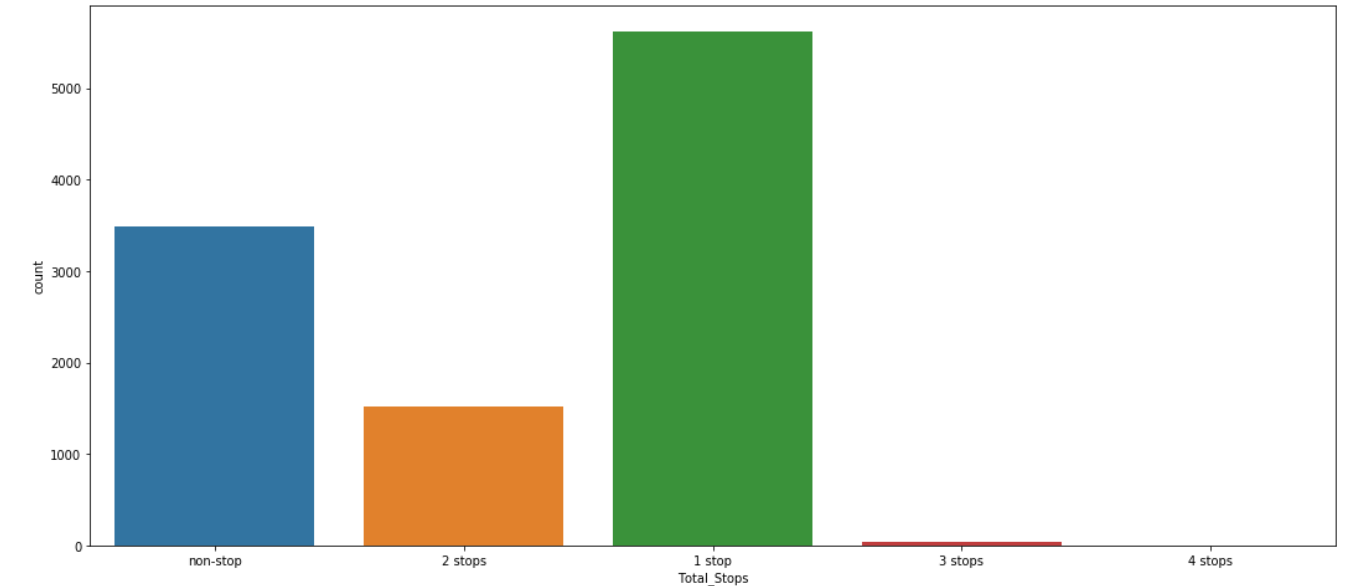
**2.2.4 Total stops**

The number of stops is from 0 to 4.

0 is for non-stop flights.

Most of the flights take one-stop, then non-stop,two-stop, three-stop

and four-stop.



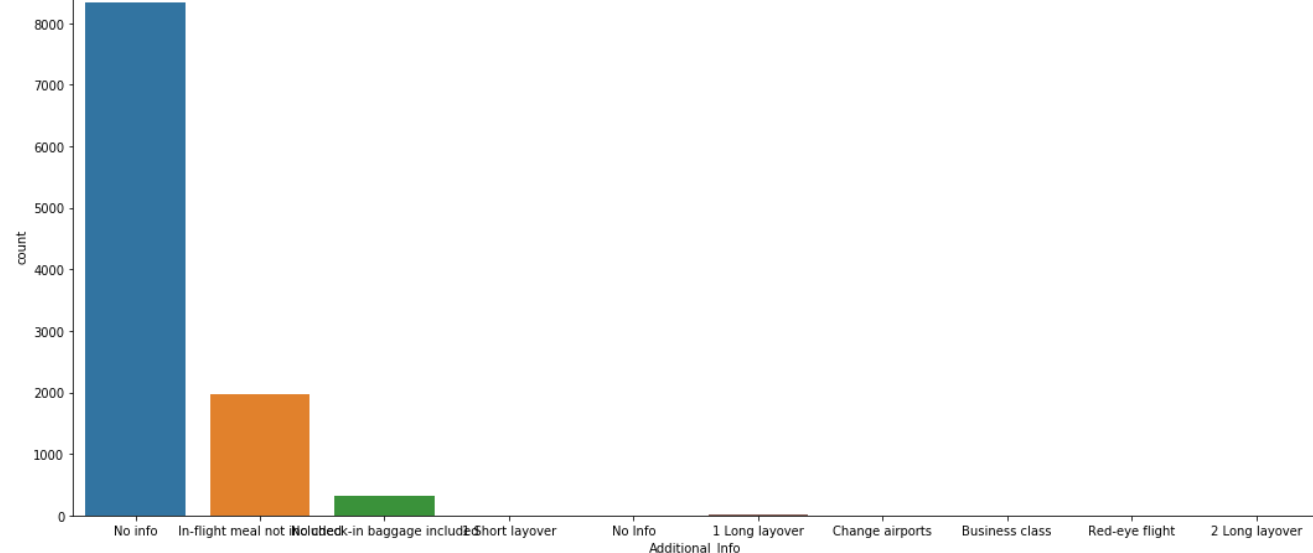
**2.2.5 Additional Info**

Additional info includes No info, In-flight meal not included,

No check-in baggage included, 1 Long layover, Change airports

Business-class, Red-eye flight, 2 Long layovers, 1 Short layover.

Around 80% of rows are showing No info

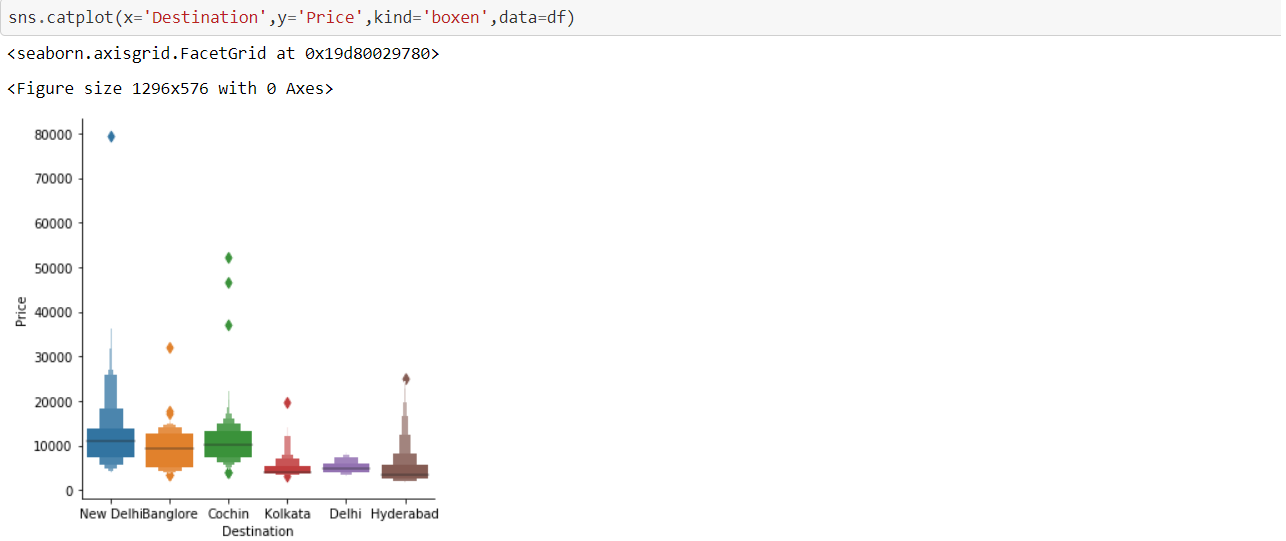


**2.2.6 Prices concerning source and destination**

**1. Destination**

The price of the flights to Cochin is the highest as compare to other cities

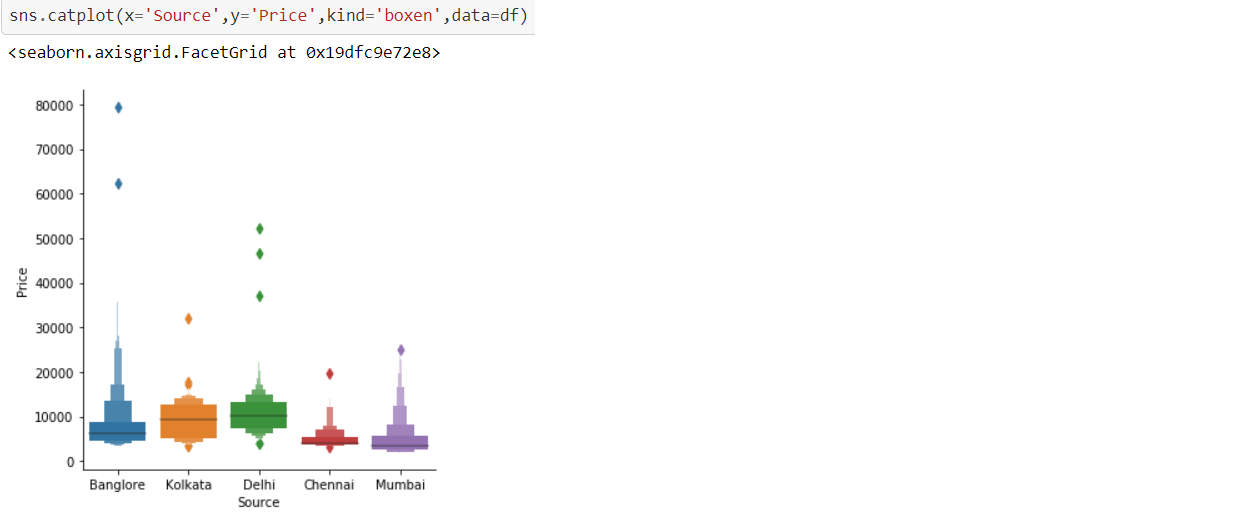
Only Delhi has one flight with an abruptly high price.



**2.Source**

Flights from Bangalore has the highest Price than from Delhi, Kolkata

Mumbai and Chennai.



2.2.6 Prices and Total-Stops

Flights with one-stop are most expensive and flights that take three stops are the least expensive.

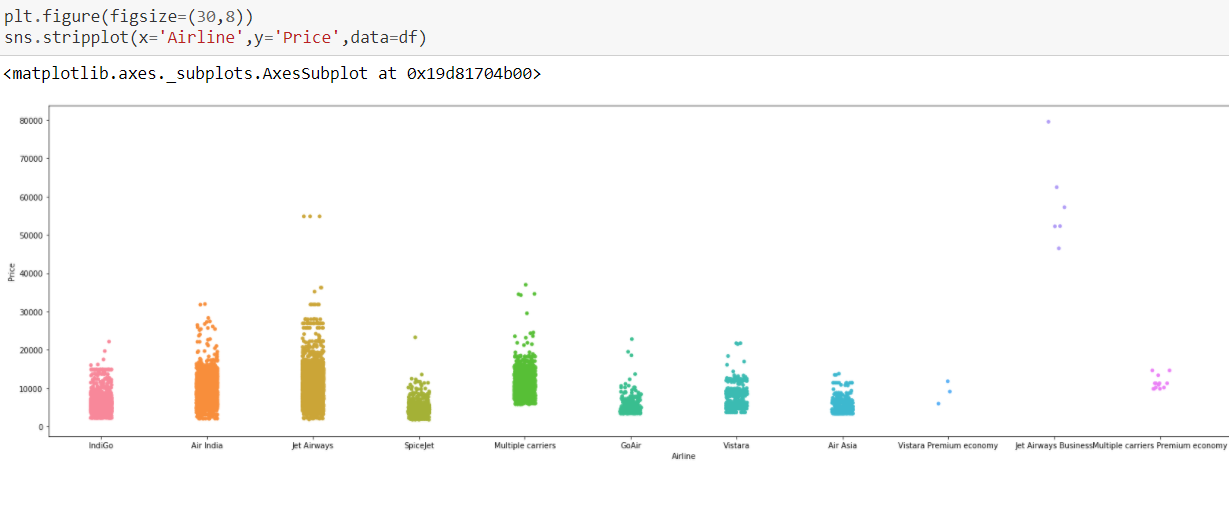


2.2.7 Prices and AirLines

12 AirLines in the Dataset are Jet Airways

IndiGo, Air India, Multiple carriers, SpiceJet, Vistara, Air Asia, GoAir, Multiple carriers Premium economy, Jet Airways Business, Vistara Premium economy, Trujet

Prices of JetAirways Business are most expensive followed by JetAirways, Multiple Carriers, Air India.

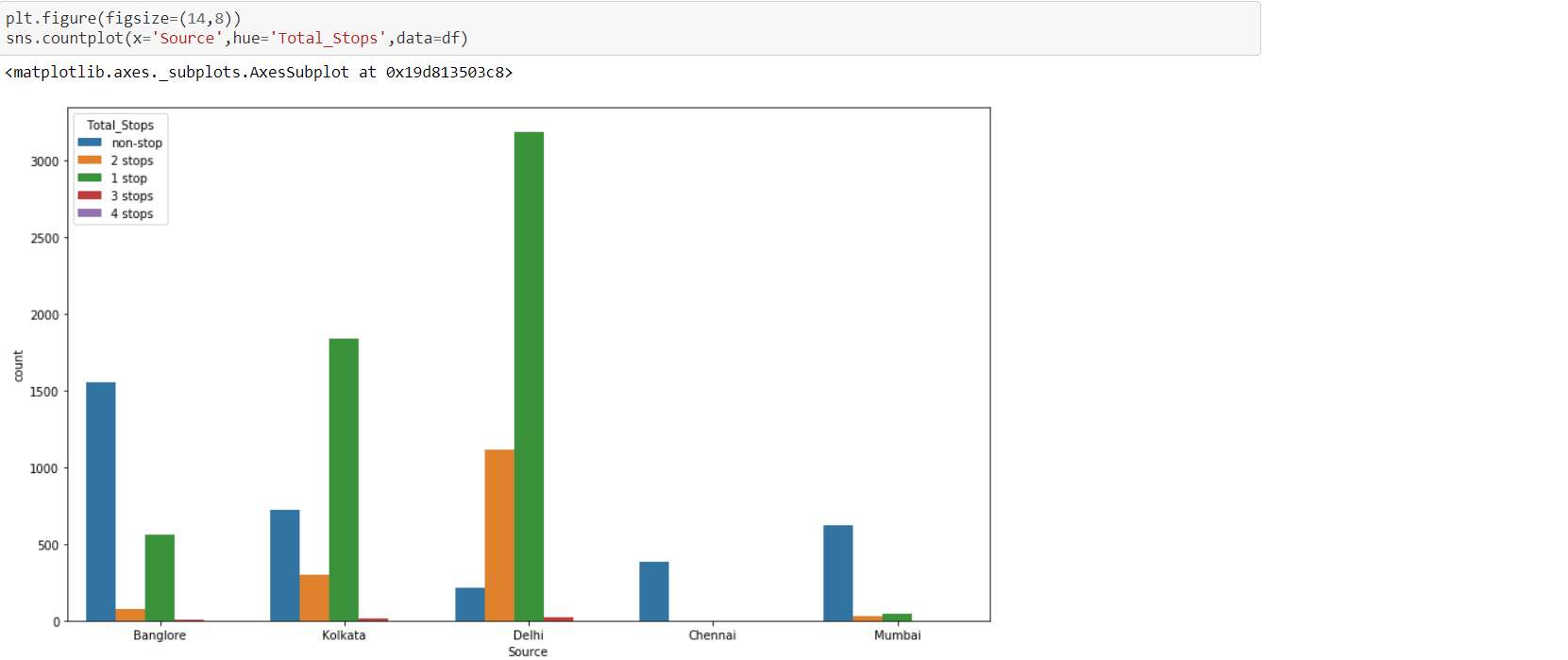


2.2.8 Prices with Total Stops

1.Source

From Delhi and Kolkata, most of the flights take one-stop

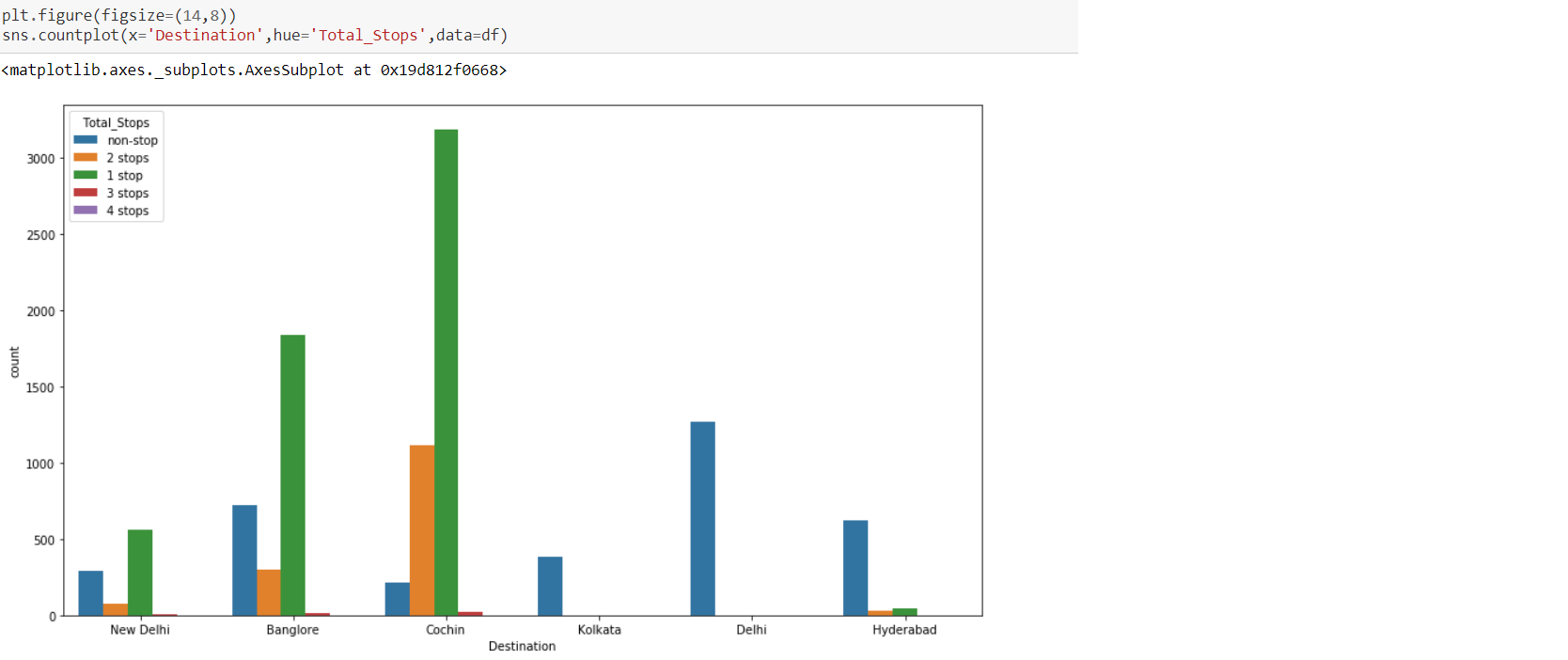
And from other locations, most of the flights are non-stop



2.Destination

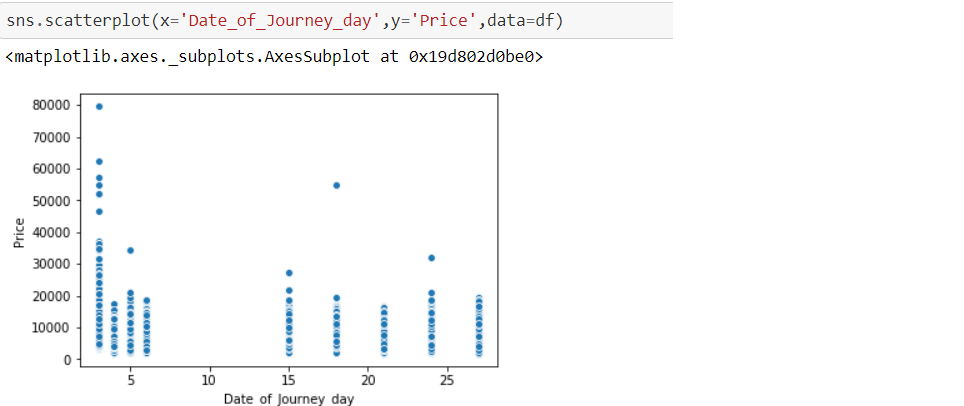
To Cochin, Bangalore, New Delhi most of the flight take one-stop

And to other locations, most of the flights are non-stop



2.2.9Date of journey and Price

In the first week of the month, prices are high and in the middle of the month there is no data given in the data set, In the end of the month, prices are from less to moderate.



3

EDA Concluding Remarks

1. One null value was there in route and Total-Stops feature which was replaced by a mode of the feature.
2. The dataset is imbalanced for the majority of the independent columns.
3. From Chennai to Kolkata and Chennai to Delhi only non-stop flights are there.
4. From Delhi to Cochin most of the flights take one-stop.
5. The price of the Flights with one-stop is high than with two-stops.
6. Flights from Bangalore with Jetways Business class with one-stop are one of the expensive flights.
7. The Prices of Flights to Cochin with one-stop are high.
8. No need to go for outliers and skewness because no independent variable is numerical.

4

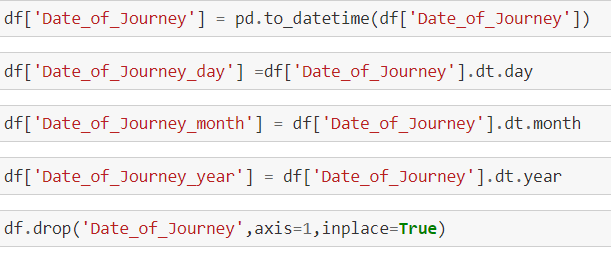
Preprocessing pipelines

Preprocessing pipelines include the essential steps to make the dataset ready for the Machine learning model and eventually predict the target variable.

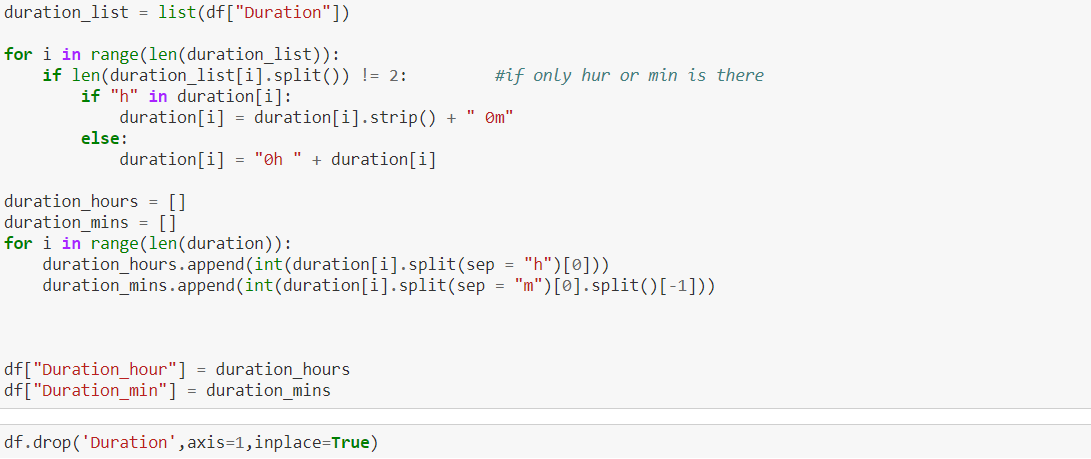
4.1 Correcting DataTypes

Features like Arrival\_Time, Dept\_Time, Duration, Date\_of\_Journey are in the wrong datatype object.

I have changed the datatype of these features and separated hours and minute from Arrival and Dept Time also Month, Day, and Year from Date\_of\_journey



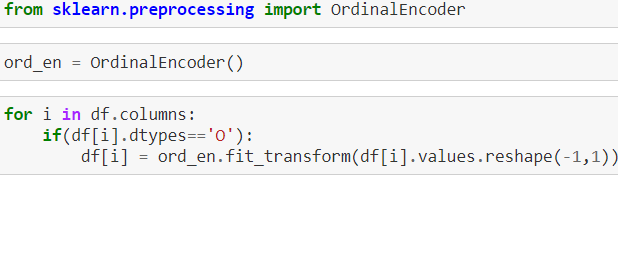




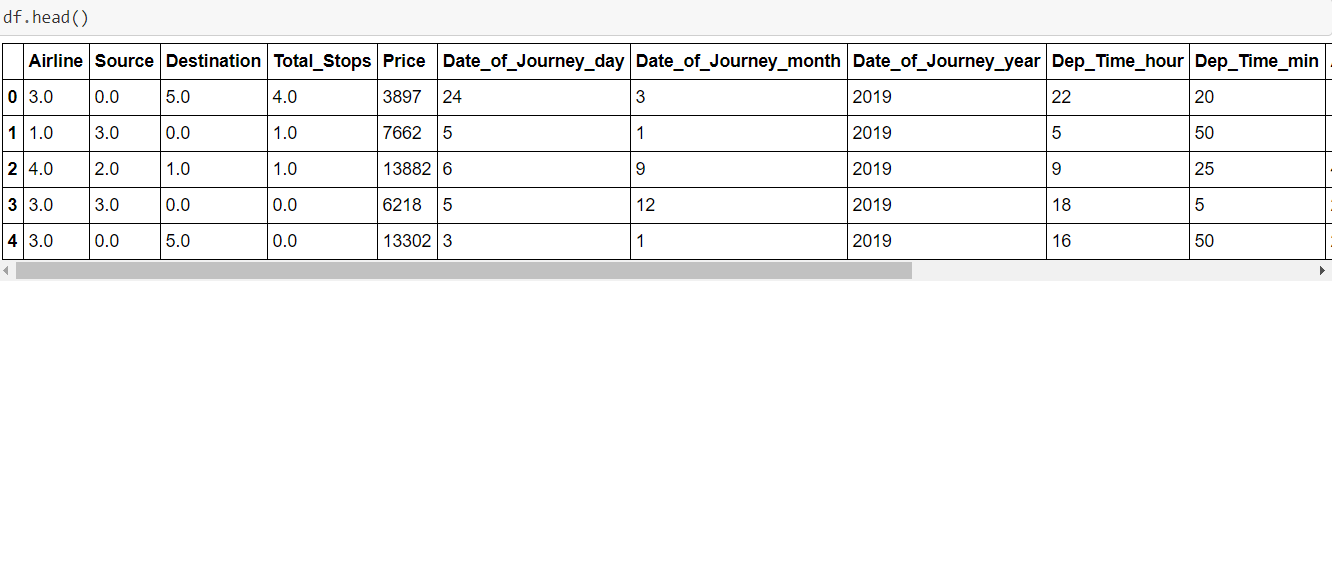
4.2 Encoding

For machine learning algorithms each feature needs to be in numerical form. And for this dataset, every independent feature needs encoding.

For the feature encoding, I used Ordinal Encoder.

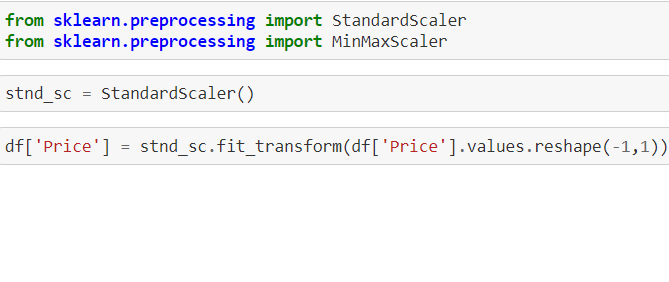


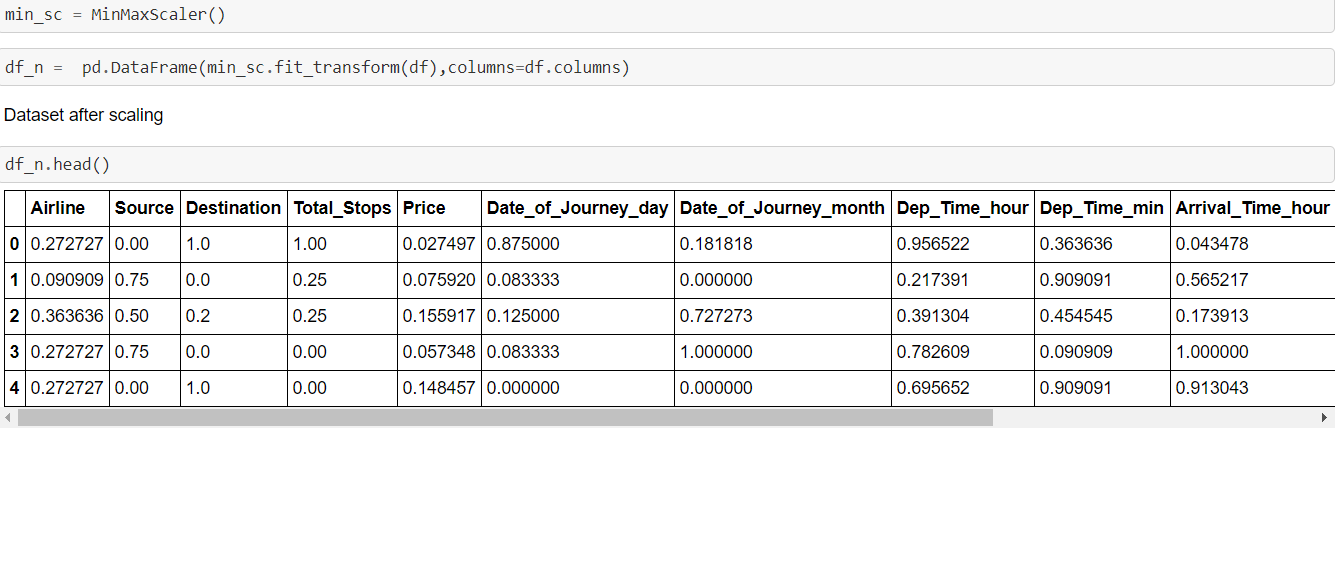
Data set after encoding:

4.3 Scaling

I have used both MinMax Scaler and Standard Scaler.

Standard Scaler for Price as the Price ranges far from other features and MinMaxScaler for other independent features. To shrink all the columns in a similar range for better accuracy.

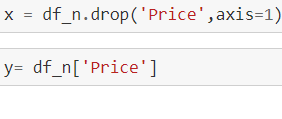


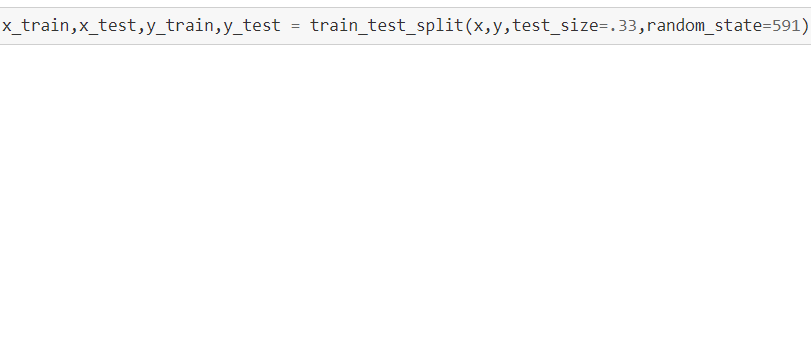


4.4 Splitting Data

Before applying any machine learning algorithm we need to split the data into train data and test data.

We use train data to fit the model and test data to predict and validate the model.





5

Building Machine Learning Model

5.1 Standard Algorithms

Let us use some of the standard algorithms for predicting the target variable.

Basic Standard algorithms that I have used are Linear Regression, Support Vector Regressor, Kneighbors Regressor, Decision Tree Regressor.

Each algorithm is evaluated based on Mean Squared Error(MSE)

and Root Mean Squared Error(RMS).

Mean Squared Error is the average square value of the difference between predicted and actual values.

Less the MSE and RMS more efficient the model is.



5.2 Ensemble Techniques

Ensemble Algorithms builds several models and combines the results to give better accuracy.

I have used three ensemble techniques :

Random Forest Regressor, Ada Boost Regressor, Gradient Boosting Regressor.

The evaluated result is based on the difference between the cross-validation score and r2 score.

Gradient Boosting Regressor gave the least difference between cross-validation score and r2 score.

r2 score is the proportion of the variance calculated by the model to the total variance.

Least the difference between cross-validation and r2 score indicates the model is not overfitted and can give better results for testing data also.



5.3 Hyper Parametric Tuning

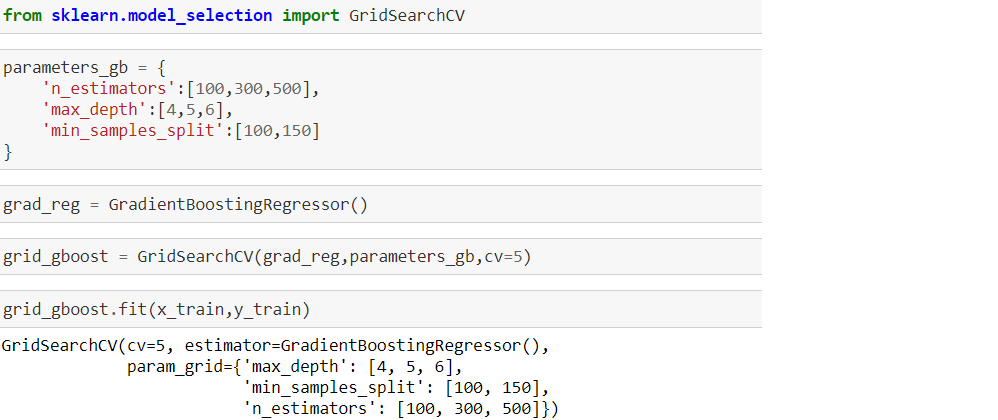
Moving with the Hyper Parametric Tuning of the Gradient Boosting Regressor for the more optimized result.

Hyper Parameters that I have used are :

n\_estimators: Number of sub-models.

max\_depth: Maximum depth of each estimator

min\_samples\_split: Minimum number of samples to split

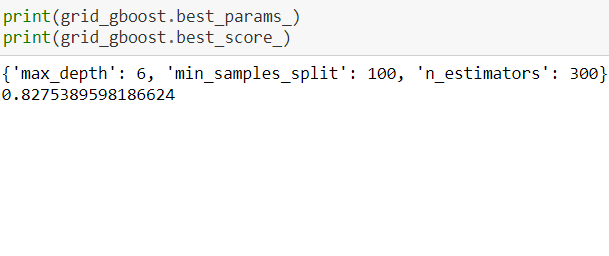


Best Parameters Given:

'max\_depth': 6, 'min\_samples\_split': 100, 'n\_estimators': 300

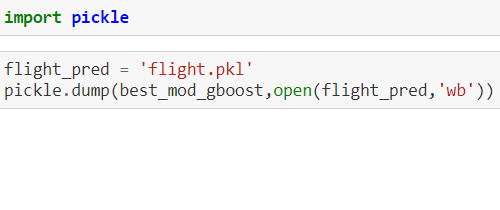
Best Score:

0.82



5.4 Saving the Model

Saving the best model for future use.



5.5 Comparing result with actual values



5.6 Using the Saved Model

Here I have loaded the saved model for predicting the Price of the test Data.



6

Concluding Remarks

* This type of Data requires more focus on features engineering because of the columns like Date\_of\_Journey, Arrival Time, Dept\_Time, and Duration.
* As most of the features are imbalanced so adding more data may increase the accuracy of the model.
* More Work can be done with routes and additional info
* Flights from Bangalore of JetAirways Business Airlines which takes one-stop and fly in the first week of the month are one of the expensive flights.
* Flights from Chennai are non-stop and one of the cheapest flights based on the dataset. It indicates that the data is imbalanced and more data can build a better model.
* All the preprocessing steps are also performed on the test data same as on training data.