**Flight Price Prediction**

**Submitted by**

**Amrutha Dhondale**

**Problem Statement**

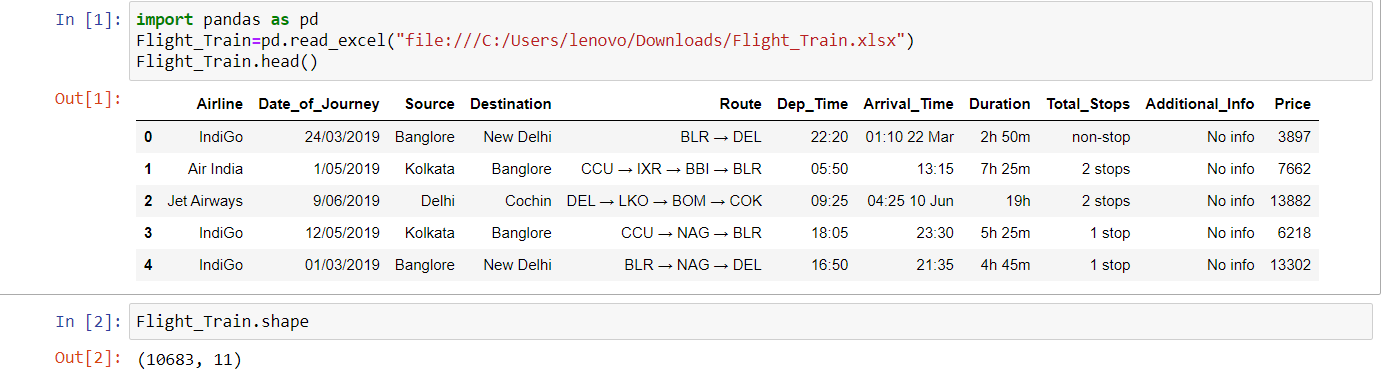
*The fare of a flight ticket is vary based on number of aspects such as duration, number of stops, source, day, airline and many more aspects. Each airline has its own algorithms and rules to decide the fare. Machine Learning helps us to predict the price much closer than actual price. It’s always hard to guess the flight price because prices are dynamic. The purpose of this blog is to predict the fare of the flight ticket, with the help of dataset. The model prediction accuracy is very good i.e., 88.6 R squared Score with dataset. The dataset is from March 2019 to June 2019 that to includes various source and destination.*

**Plan**

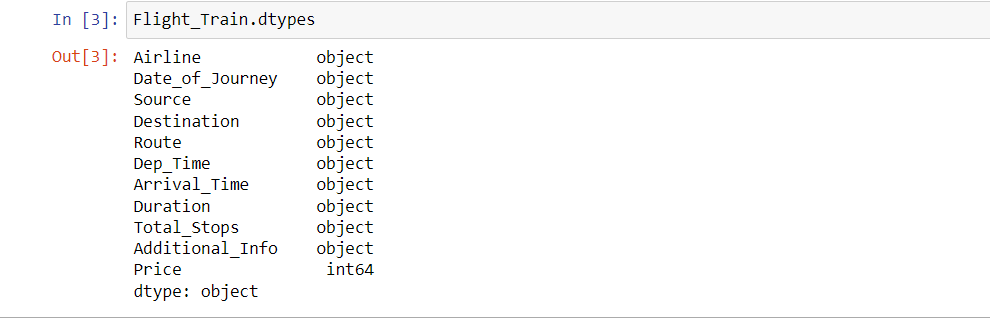
*The dataset which will going to use here is Flight Price Fare took it from Kaggle. Here given dataset will be analyzed based based on all the given information. The target value will be continuous that is Fare of the flight. Though here Regression algorithms will be coming in use with the help of Python. Comparison between different — different algorithm will happen here just to get highest accuracy, which will help to get accurate price of the flight.*

**Data Analysis**

Data Analysis is a procedure for gathering raw data than converting it into useful and informative data that will help for making decisions clear by the user. Data will be collected, analysed to answer the questions.



The Flight dataset contains 10683 rows and 11 unique columns.

****

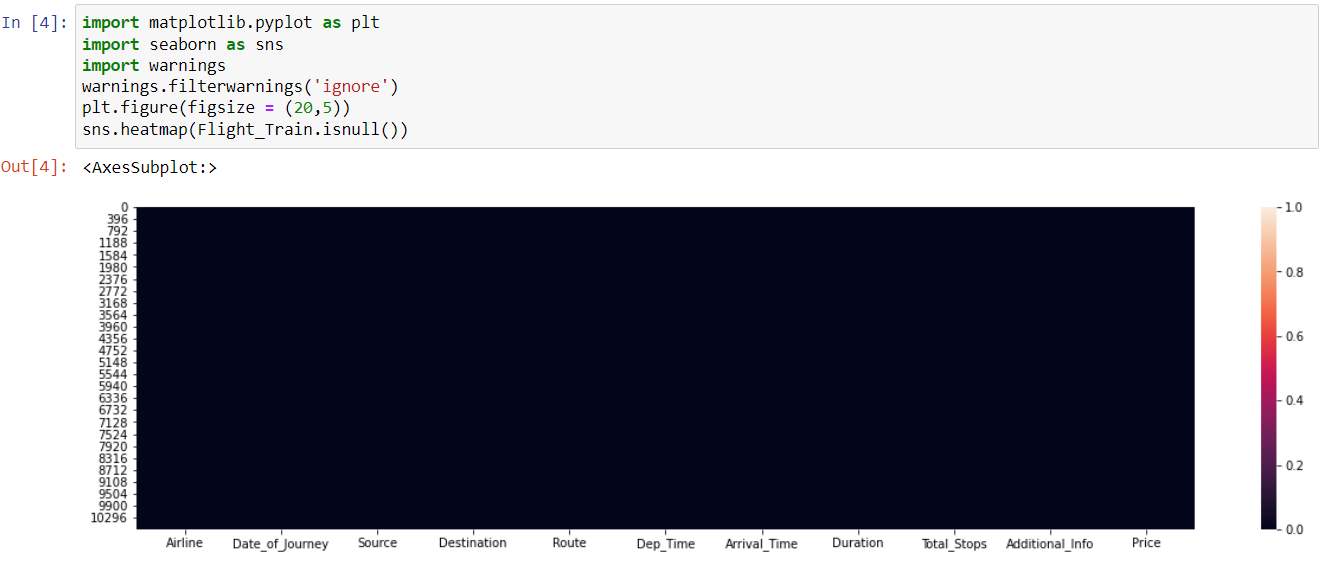
There are 10683 examples, 10 columns + 1 target variable. 10 columns are object type.

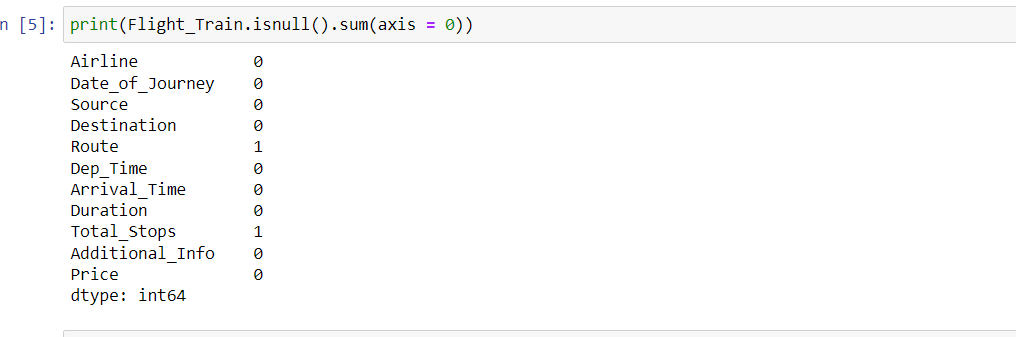
Airline —> Name of the Flight company.  
Date\_of\_Journey —> The date on which flight departure.  
Source —> From where Flight departure.  
Destination —> Arrival place.  
Route —> Route followed by Flight, Where flight stopped in between.  
Dep\_Time -> Time on which Flight Departure.  
Arrival\_Time -> Arrival time of the flight.  
Duration -> Total journey hours.  
Total\_Stops -> How many times flight stopped in between.  
Additional\_Info -> Hand bag allowed, business class etc.  
Price -> Cost of the Journey.

Target variable (Price) is int64 i.e., continuous, So Regression will be used to learn the model.

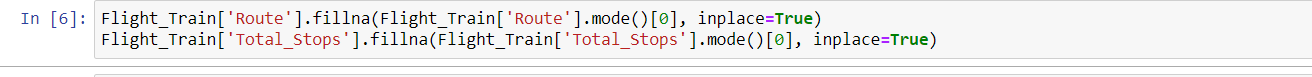
**EDA Concluding Remark**

It’s clearly visible that NULL values present in the dataset

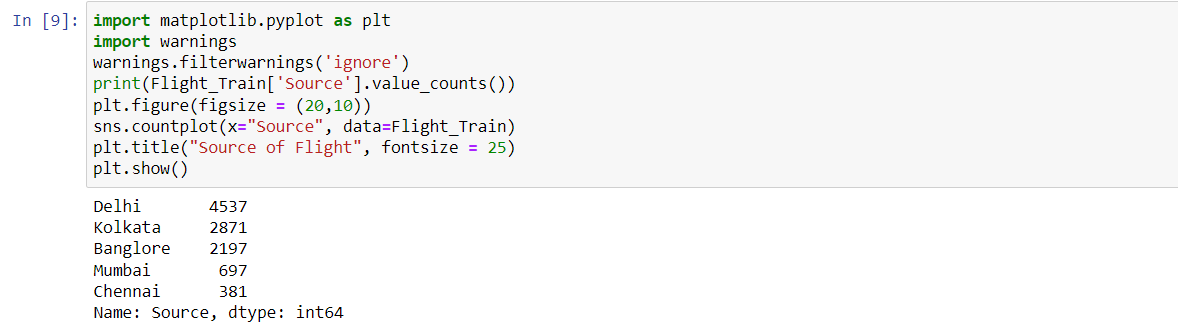


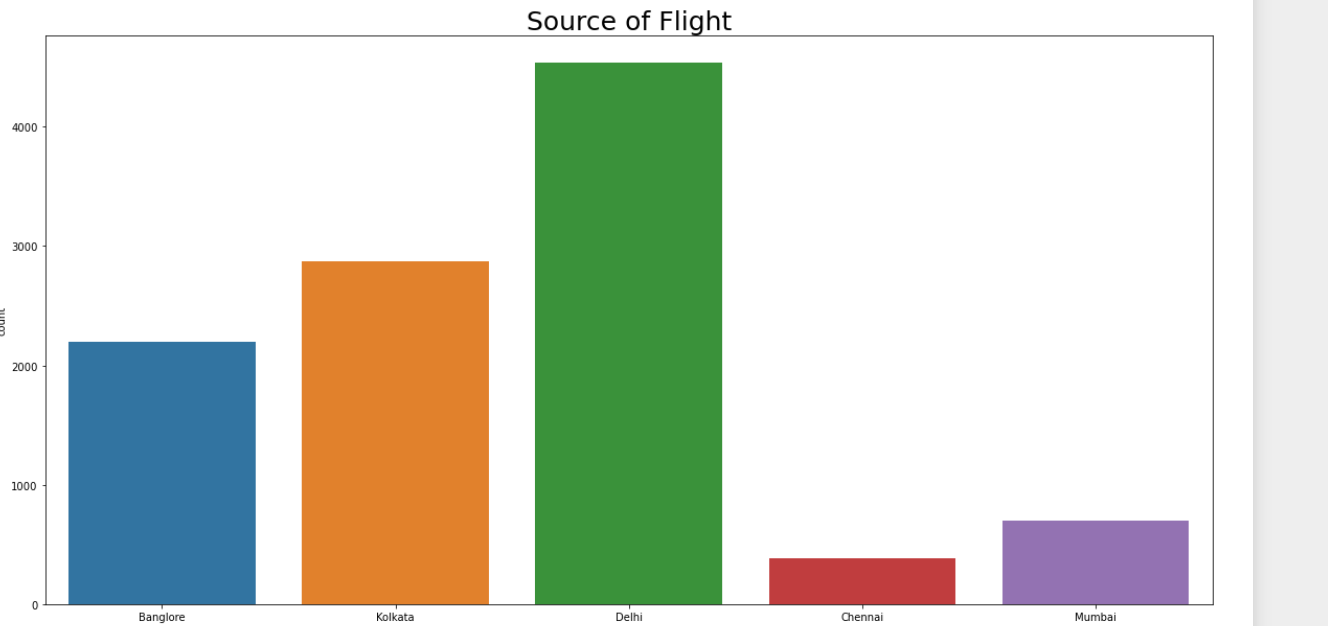


Column Route and Total Stops have 1–1 NULL values. There are 2 ways to get rid of these NULL values. 1st is drop null values but it’s not a good choice to go with. 2nd is replaced NUL values with mean, median or mode. In this case type of variable is int. We can use mean, median or mode but in case of object we have to use mode.

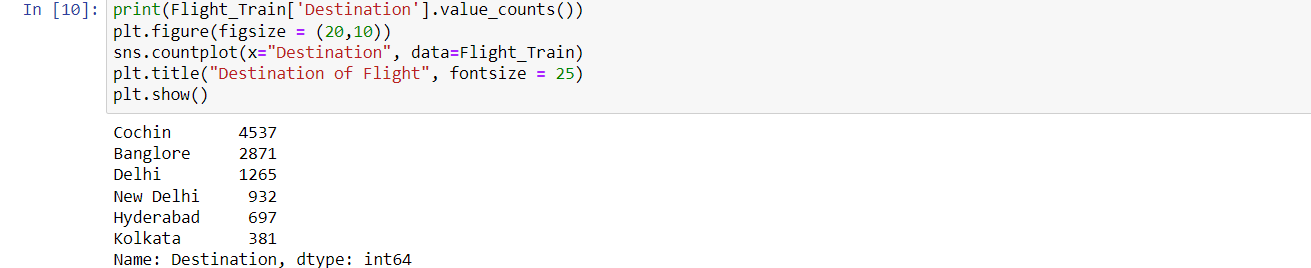
****Replaced NULL values with mode. Now no more NULL values present in the dataset.

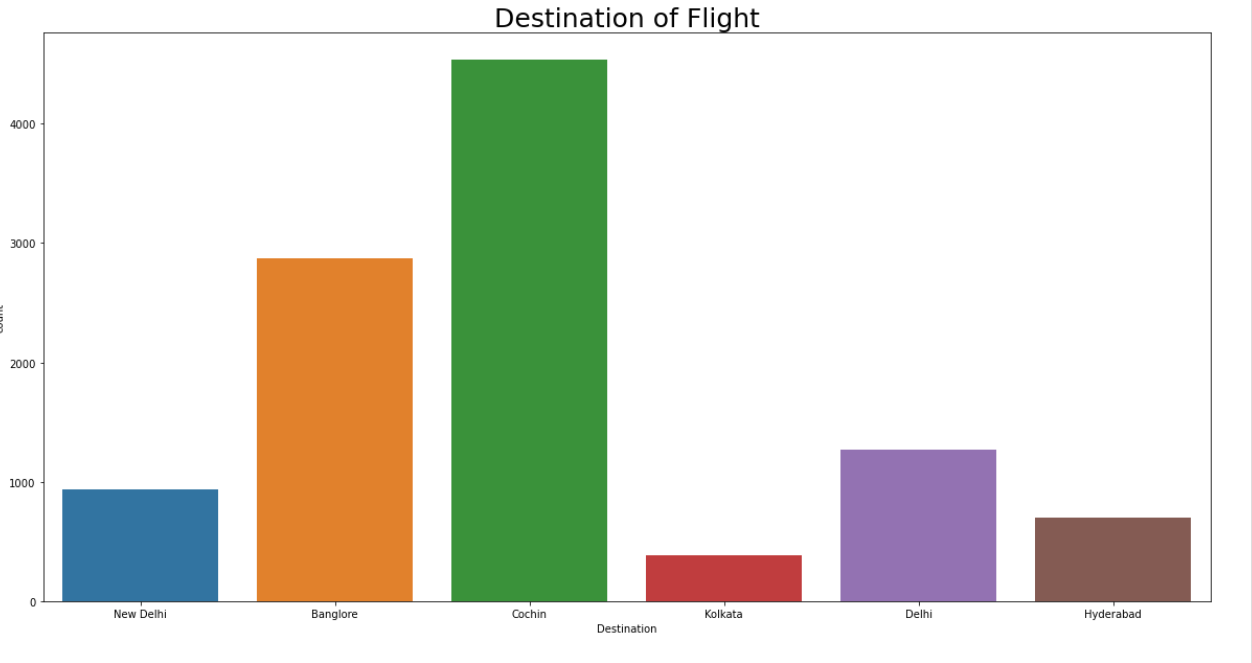
**Univariate Analysis**

****

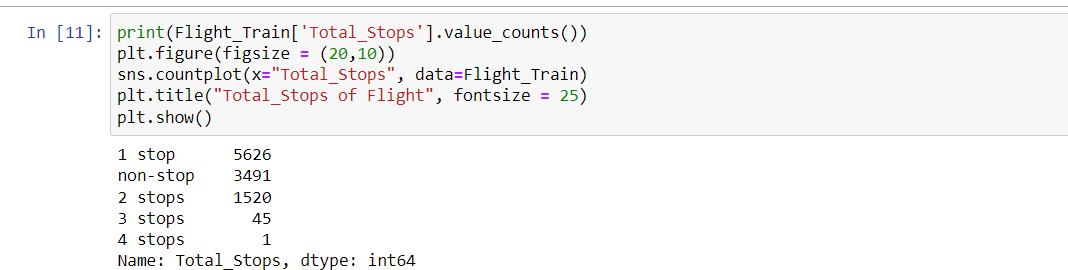
****

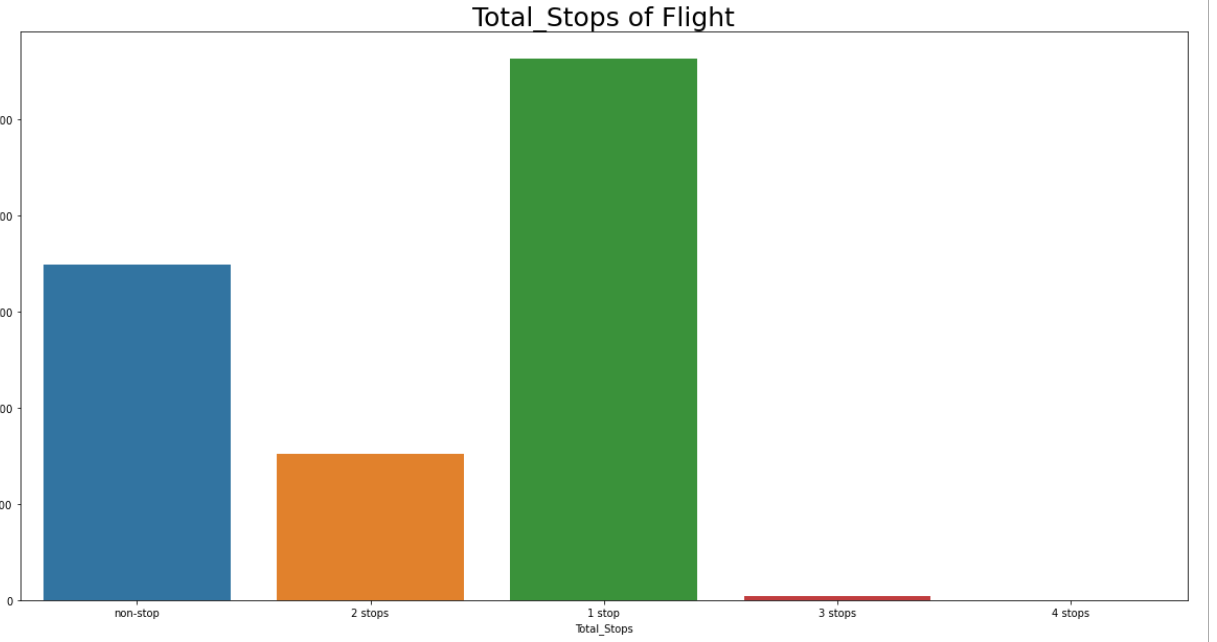
We can clearly see that Maximum number of Filght Source was Delhi or we can say 4537 Flights source was Delhi.

****

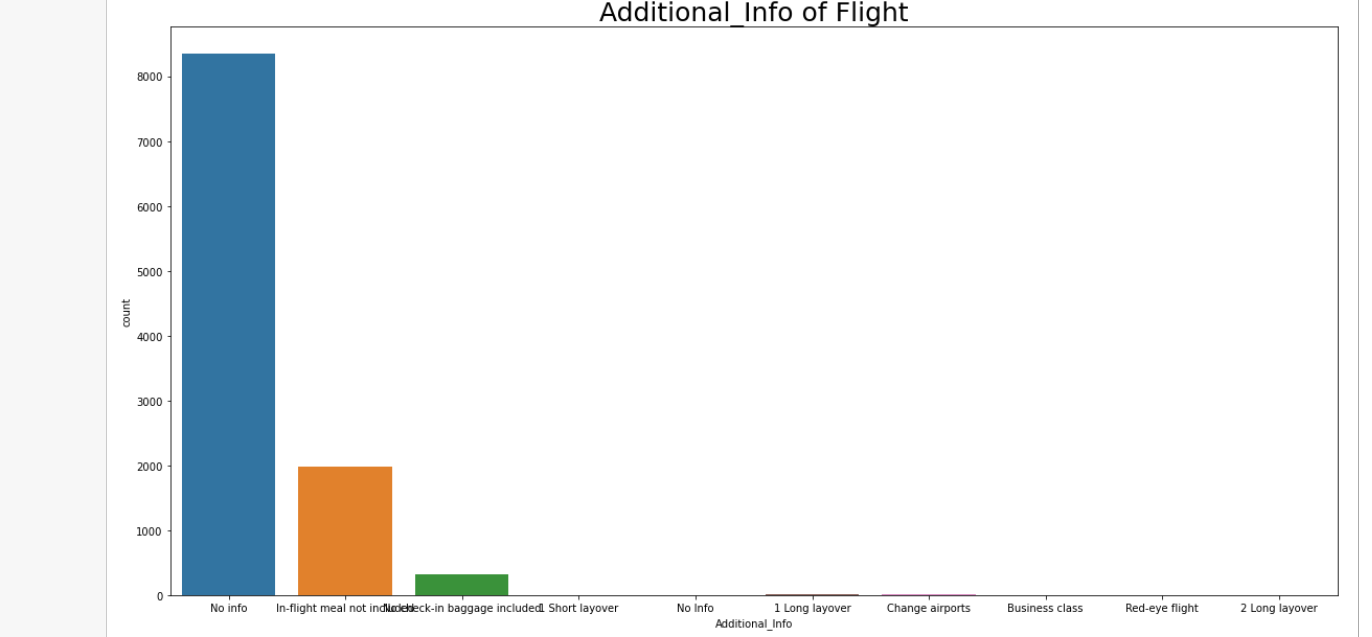
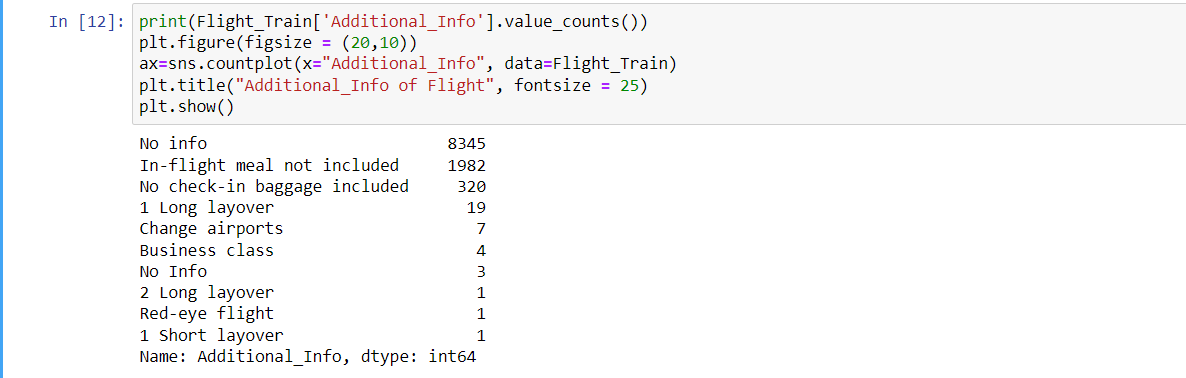
****

4537 Flight Destination was Cochin i.e., Maximum

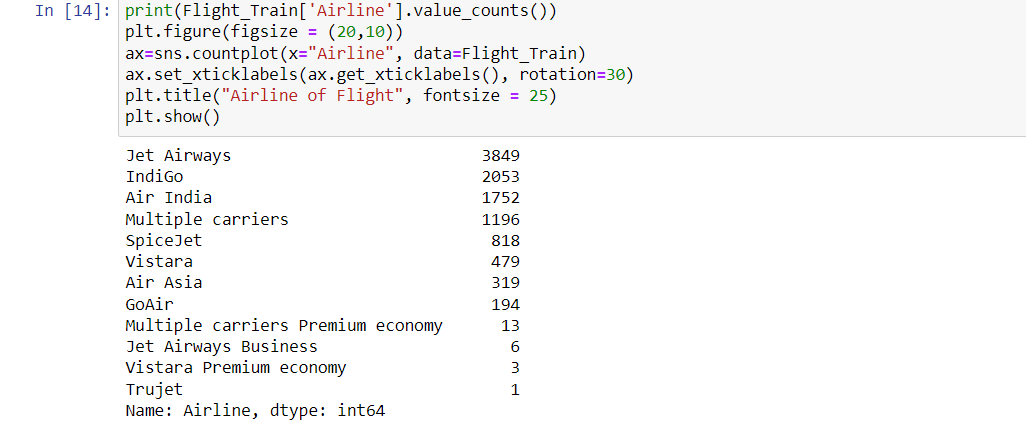
****

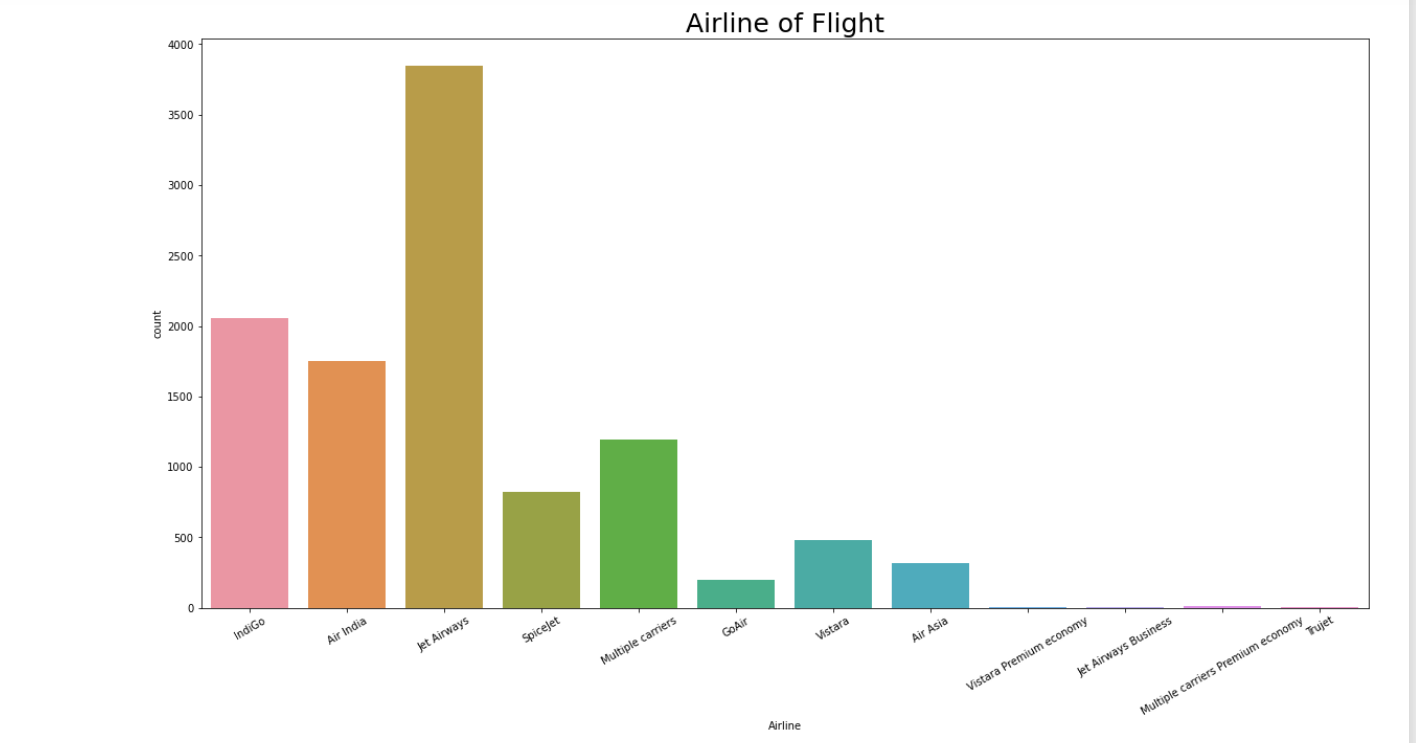
****

Most of the flights were having 1 stop is 5626 or 0 stops is 3491

****

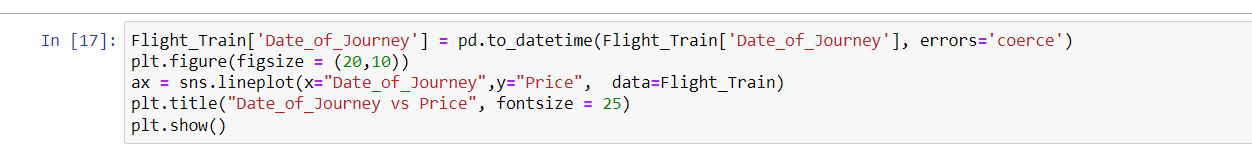
Generally, there is no information in flight.

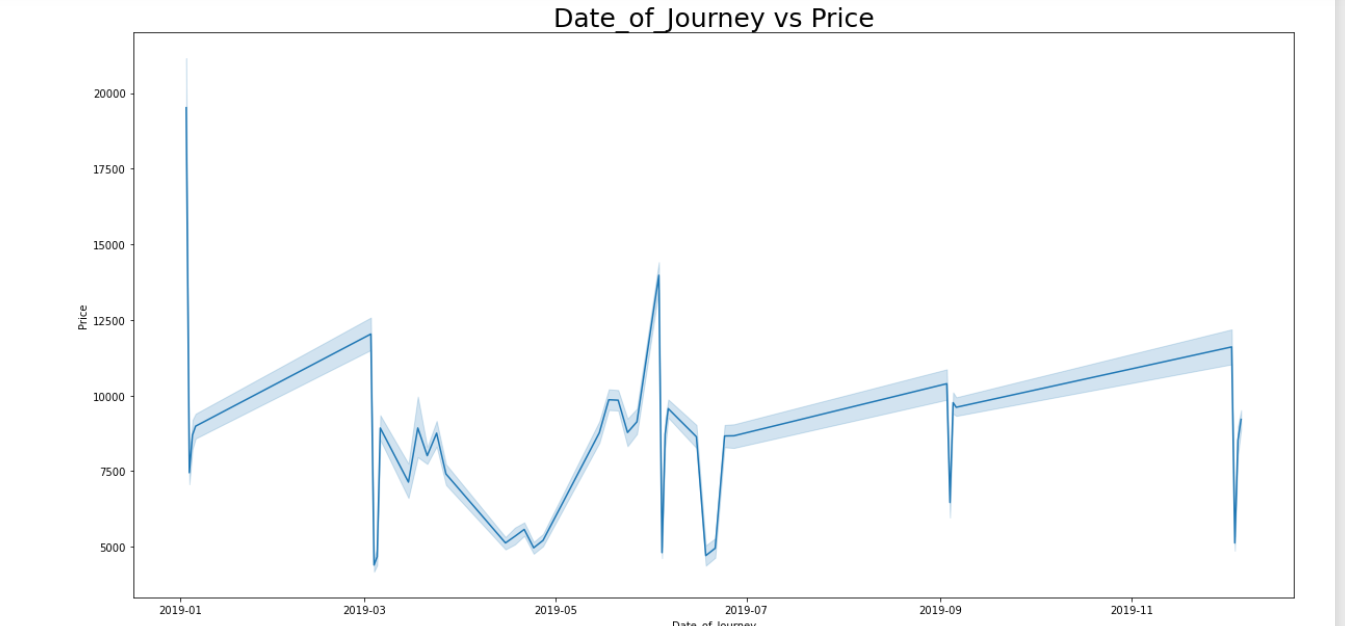
****

****

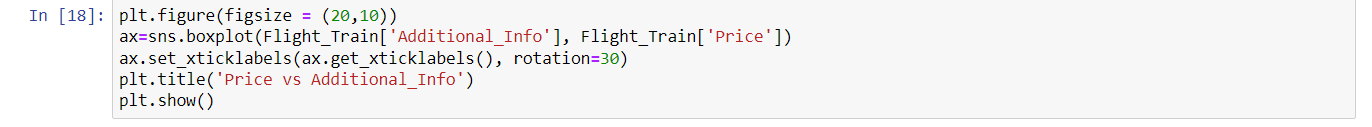
Mostly Jet Airways Airlines Filghts fly in sky.

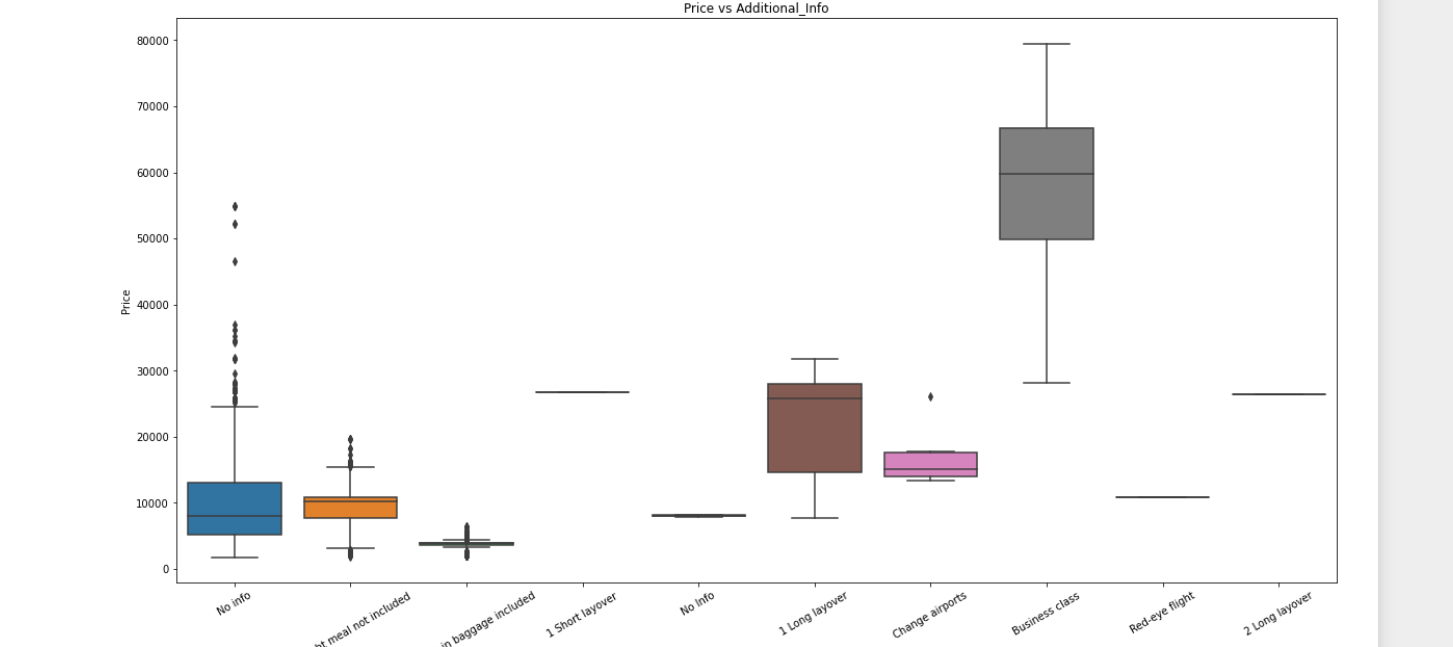
# Bi-variate Analysis[¶](http://localhost:8888/notebooks/Flight%20Price%20Prediction%20Projects.ipynb#Bi-variate-Analysis)

****

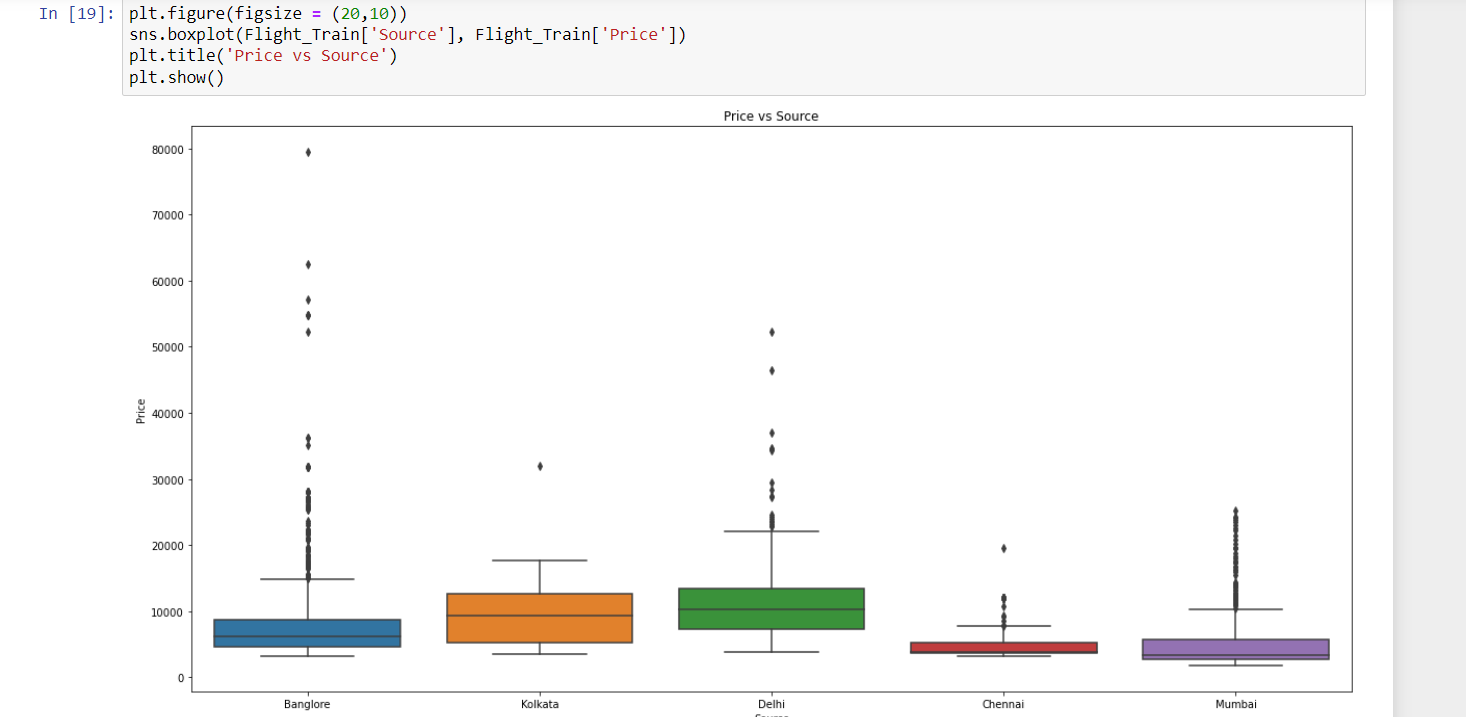
****

In January starting Flight prices were high as 20,000 than suddenly in dropped to 7500, then after that Flight prices slightly increasing from Jun to march 8000 to 12500 and in March and April flight prices are slightly decreasing and increasing multiple times. From May to June prices are going high, June and July prices are going up and down, July till September price increasing suddenly price down and then again, it's going high till mid of November.

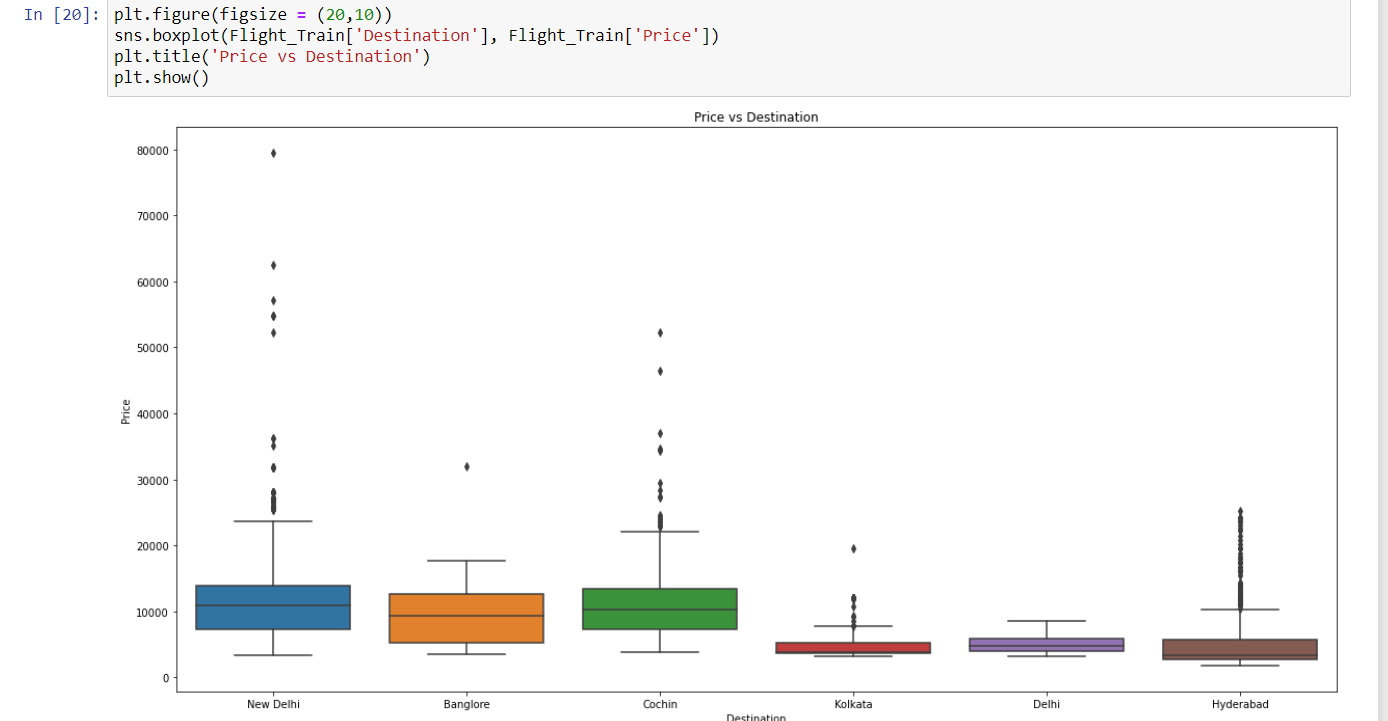
****

****

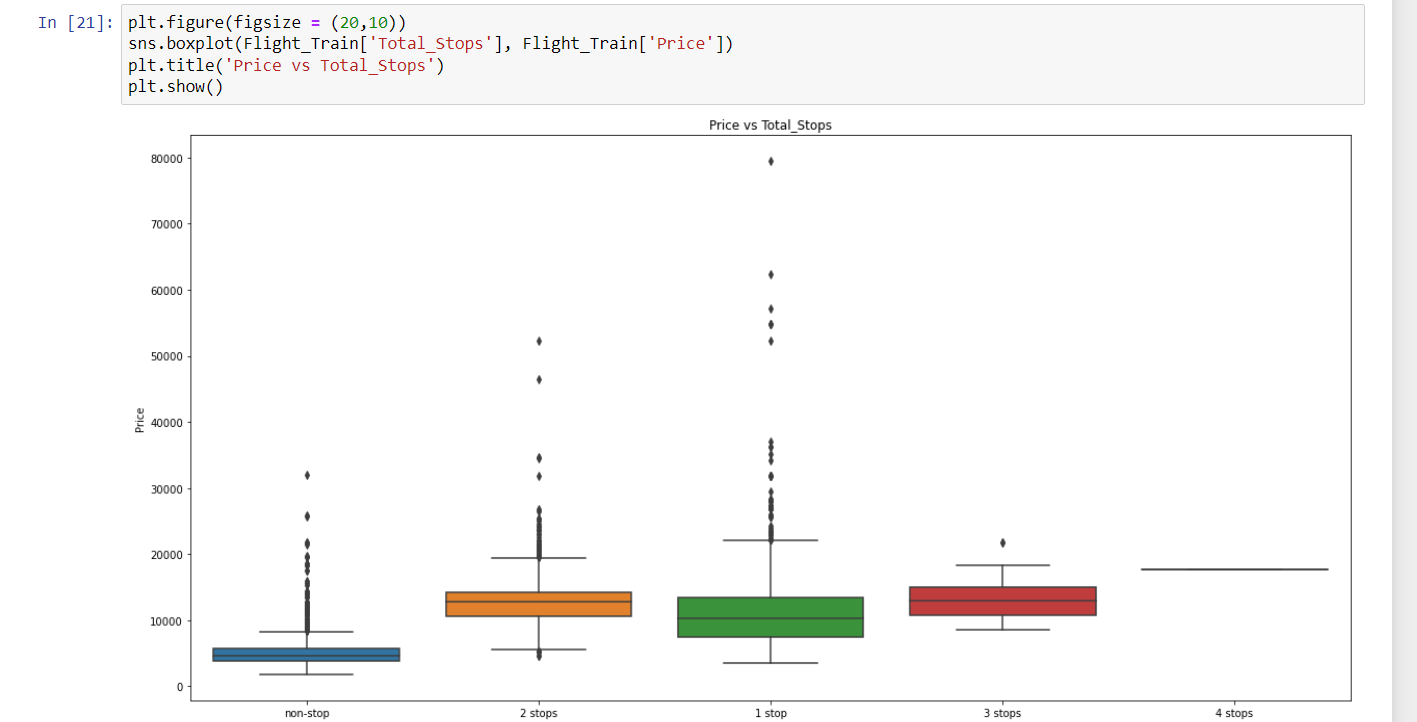
we can notice - The Flight price is too low when No check-in baggage were allowed. When customer choose Business Class that time Price goes too high. When No meal provided in Flight that time flight prices are always lesser than 20,000.

****

From Chennai and Mumbai, the flights are cheaper as compared with other Sources

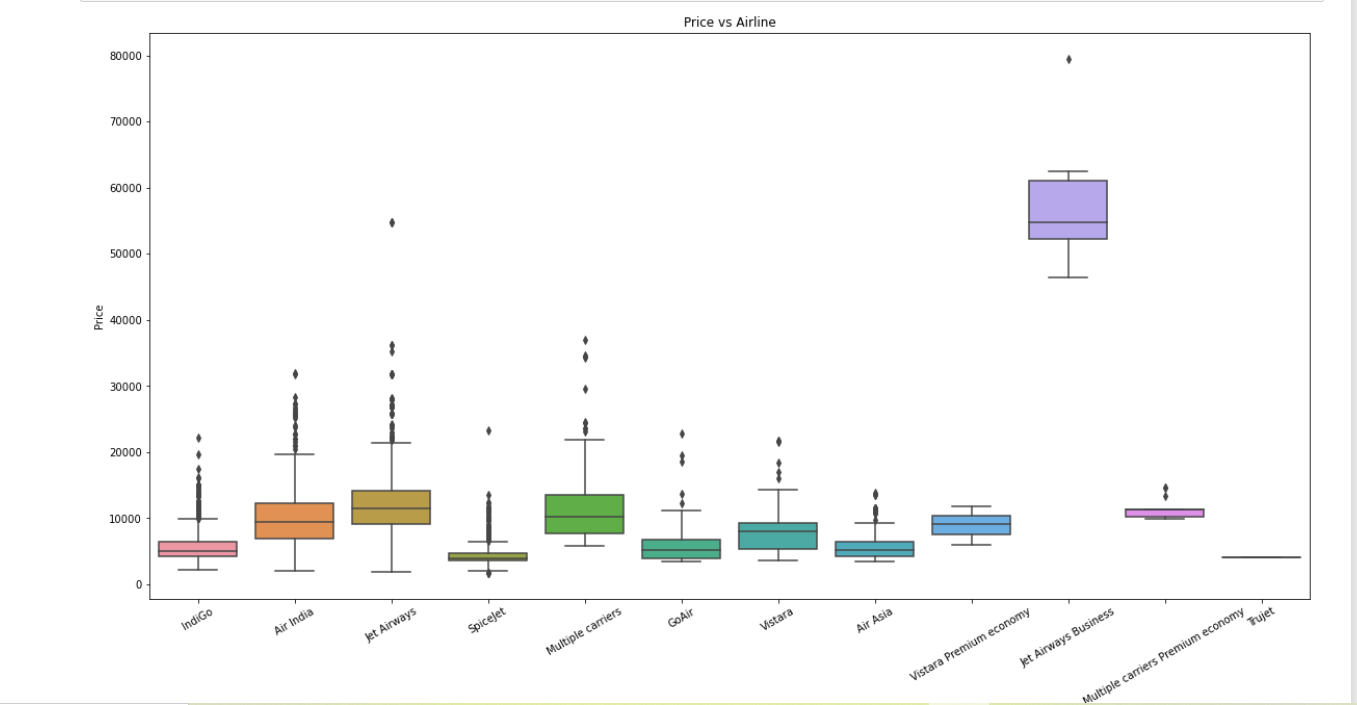
****

Kolkata, Delhi and Hyderabad flights are cheaper than other Destination price

****

Non-Stops flights are cheaper

****

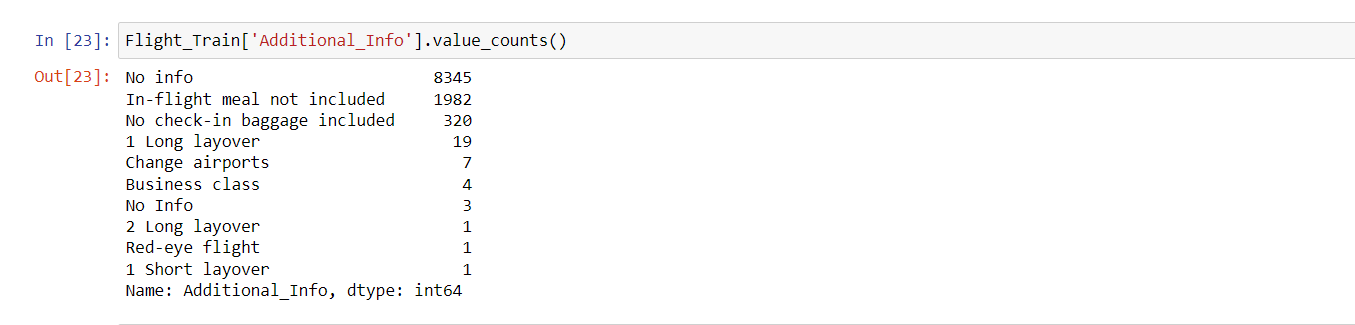
****

Jet Airways flights are always expensive and SpiceJet flights are cheaper.

Further EDA can be seen after Pre-Processing and Data Cleaning.

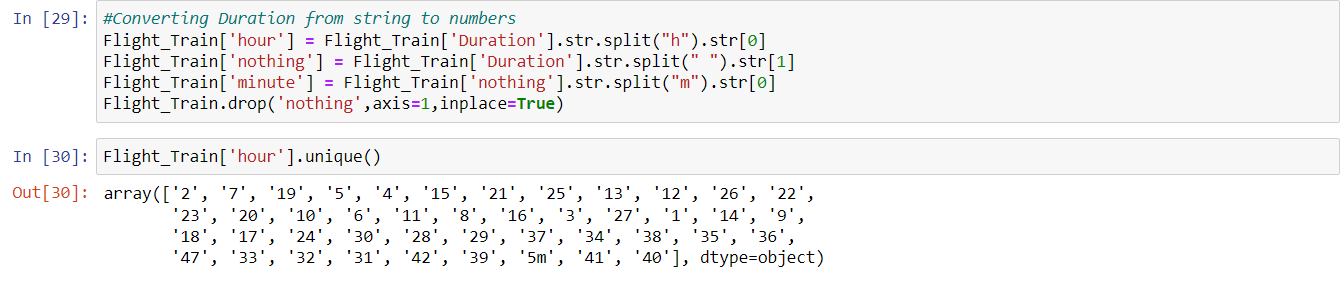
**Pre-Processing Pipeline**

Here will do extraction of the information from all the columns which will help to do better and accurate prediction. In Additional Info when some information is given that highly reflect to flight price because when a person travel on business class that is additional information, fare of the flight goes too high and when No check-in baggage included fare prices goes too low. Total number of stops will also help to predict the fare of flight, when flights are non-stop, fares are too low, when 1 stop is there is journey price increases and so on. Duration of the journey is one the biggest factor which can help to predict the fare of flight. When the duration goes high, cost of travel increases that impact on flight fare directly.

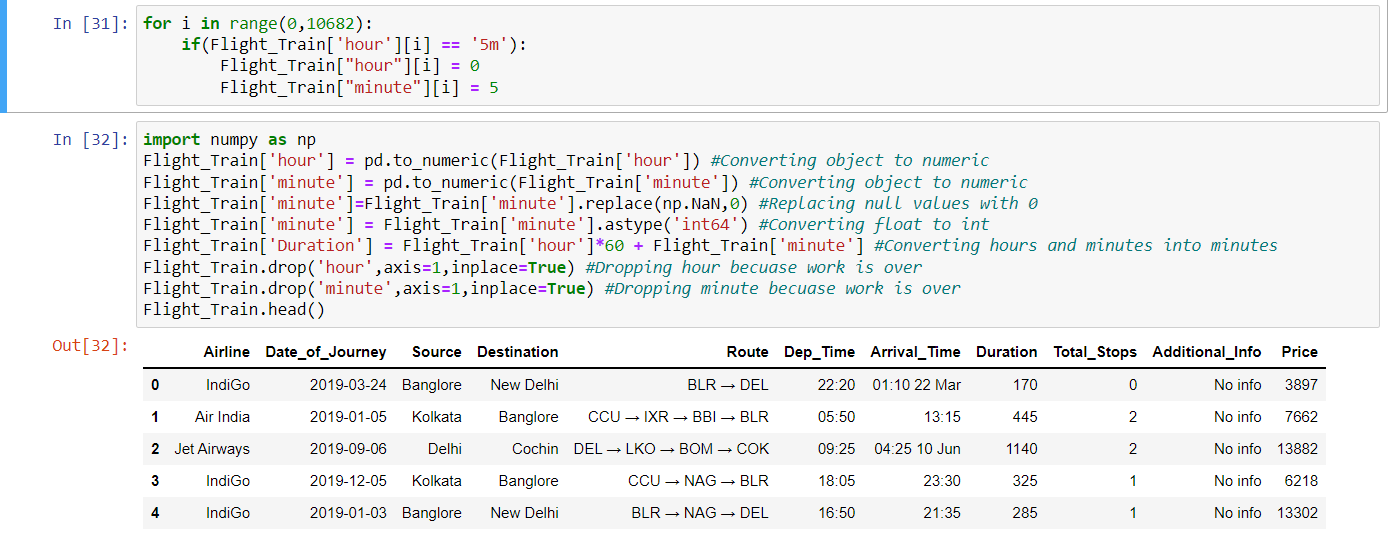




In order to do good and accurate prediction, need to get Duration of journey in minutes from hours and minutes. That will give us journey time on one scale and it’ll boost up the accuracy of the project.

****

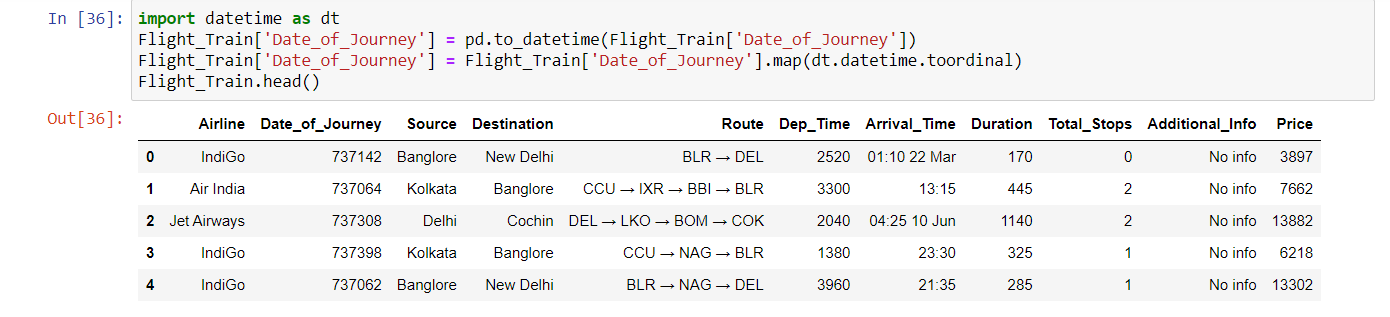
In order to get journey duration correctly had to choose this long way. There are few inbuild function which will give us journey duration but that’s failing in few scenarios.

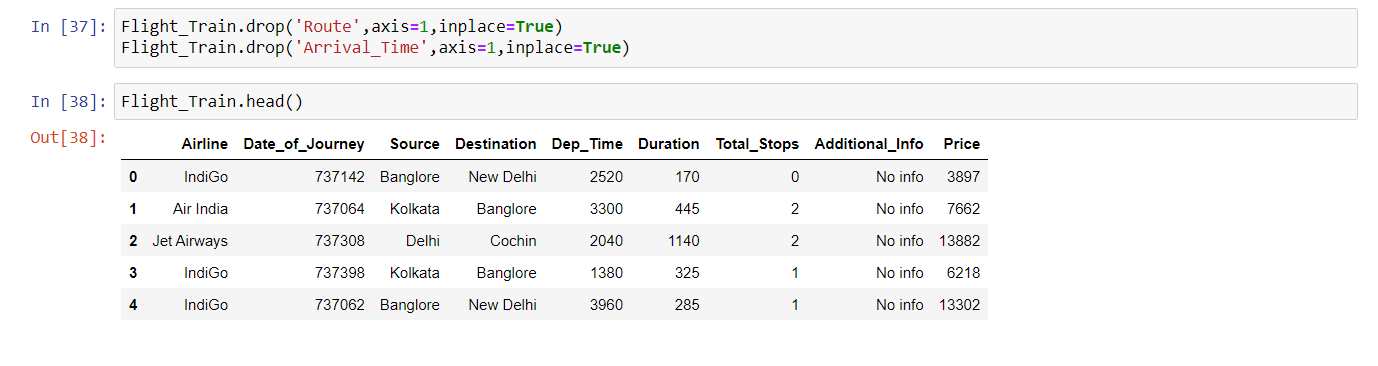
****

*#Converting Departure time into minutes*

*Flight Train[‘Dep\_Time’] = pd.to\_datetime(Flight\_Train[‘Dep\_Time’],format = ‘%H:%M’).dt.hour\*60 + pd.to\_datetime(Flight\_Train[‘Dep\_Time’],format = ‘%H:%M’).dt.minute*

Dep\_Time converted into minutes that will help for better prediction





**Label Encoding**

#Label encoding for the object columns*label\_list=list (Flight\_Train.select\_dtypes([‘object’]). columns)****from******sklearn.preprocessing******import****LabelEncoder le=LabelEncoder()*#Initlize LabelEncoder to le ***for****i****in****label\_list: Flight\_Train[i] = le.fit\_transform(Flight\_Train[i])*

Converted all the non-numeric columns into numbers

Looking into Correlation of other variables with respect to Price i.e., target variable.

# Multivariate Analysis

# 

# 

# Price is highly corelated with Duration and Total\_Stops

# 

# 

# As we saw earlier Duration and Total\_Stops are highly correalted with Price

# Removing Skewness and Outliers

# 

# 

# We lost almost 2% of our data

# Checking of Skewness

# 

# Removing skewness

# 

**Building Machine Learning Models**

**Separating Input and Output Variables**

*y = new\_Flight\_Train[“Price”]*

*x = new\_Flight\_Train.drop([“Price”], axis=1)*

Separated x and y variables to train and testthe data

**Scaling**

Scaling is required because there is huge difference in values of each column.

# 

**Finding Best Random State**

Here Linear Regression is used to get the Random state on which model is working more accurate.

# 

# ****Train Test Split****

# 

# ****Finding Best Algorithm****

# Using Random Forest, Decision Tree, K Neighbours, Gradient Boosting, Ridge and SVR to get best algorithm out of them.

# 

# Output

# 

# 

# 

# Random Forest Regressor model have highest accuracy i.e., 88.59% with 88.77% cross validation score which is good and the difference is too less.

# ****Hyper Parameter Tuning****

# 

# Now the model learnt almost 89%, which is a good score

# ****Saving the model****

# 

# 

**Concluding Remarks**

In this case study, a Machine Learning model is developed to predict the airlines fare. Here several features were mined from the dataset and combined together with the help of Machine Leaning, to do the flight price prediction. With the help of the above techniques, proposed model is able to predict the flight fare with an adjusted R squared score of 88.59%. However, there is still ways to do improvement in this model.

In the future, our model can be predicting the flight fare more accurately, if we get some of information such as seat location, when ticket was booked, special occasion on departure date etc.