FLIGHT PRICE PREDICTION

**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 and it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable. So here, our task is to analyze the prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

Here are the definitions of the 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 the destination.

**Additional\_info :** Additional information about the flight.

**Price :** The price of the ticket.

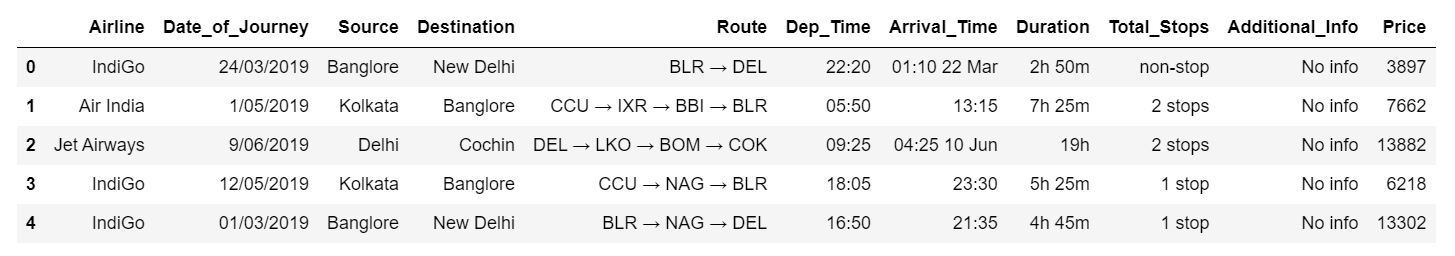
We have to build a machine learning model to **predict the price** of the flight ticket.

Before we jump into the data analysis, one must have a basic understanding of the problem. So, a little bit of domain knowledge would really help in building a good model. In our problem, since flight prices looks unpredictable, that is where machine learning algorithms do their job and predict for us so that we could have an idea how prices vary depending on various factors.

* The prices of flights depends on the **distance** that you wish to travel. The farther the distance, more will be the travel time and more expensive will be the ticket. But this is not the deciding factor as there are other factors we need to consider as well.
* The second thing we need to consider is the **season** when the ticket was booked. Some tourist destinations has their own peak seasons and flight prices will also be more expensive during those seasons. And whether it is a holiday season or not determines the prices as well since most travelers would want to go home for Christmas or go somewhere for a trip. So, prices will go up as demand is increasing in those seasons.
* The **timing of the flight** also plays a major role in deciding the price. Whether you want to travel in the early morning, noon or afternoon could vary the price even for the same airline and the same date. Flights also do have their off-hours, you may be charged extra amount in that case. And booking flights in advance (maybe a couple of months in advance) will give you a comparatively lower price than booking the same flight just a few days before the departure. And booking flights during non-holiday seasons could really save you some extra bucks.
* **Flight type** also determines the price of the ticket. Whether the flight is a direct flight so that you save extra travel hours or if the flight has multiple stops so that you have to compromise the duration of the flight also affects the price. If you do not mind spending extra hours in the airports, then you are likely to save extra money in that case.

**Exploratory Data Analysis (EDA):**

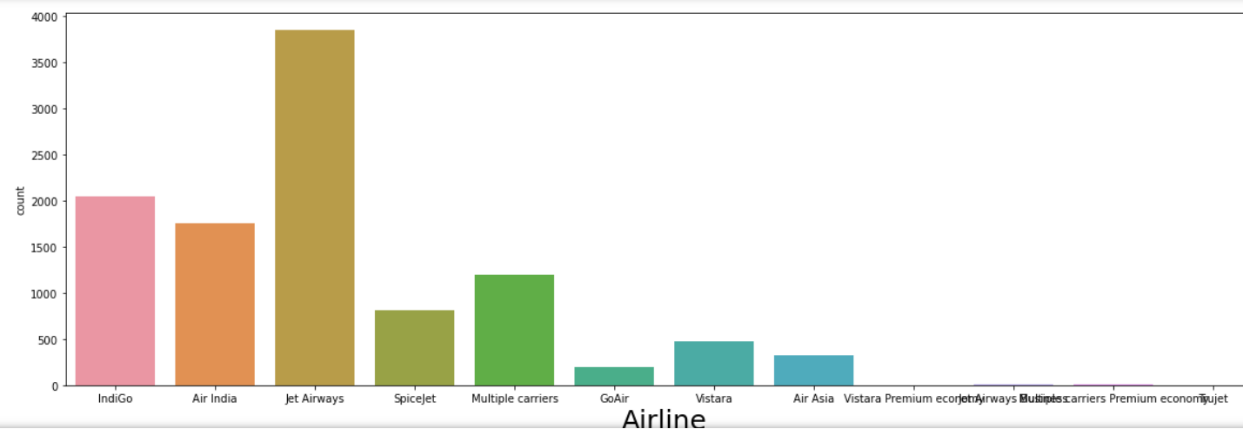
The data provided is the prices of different airlines during March to June 2019 and between various cities in India. There are 10683 rows with 11 columns. The data looks like:



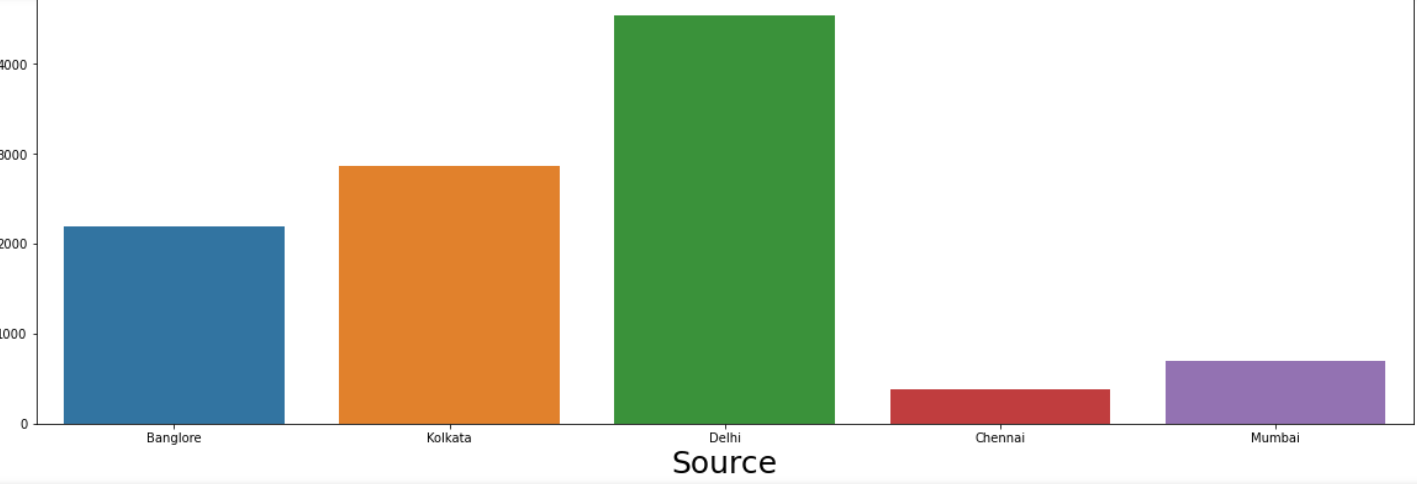
* There is one row having null values and since the data is big enough, the null values are dropped and total rows become 10682.
* All the features are categorical and the target variable is continuous. So, our problem is a regression problem.
* We have to encode the data accordingly so that the machine understands the data.

**Visualization:** We can visualize how the features looks like and how important they are to predict the price. Also we can see how travelers decide in terms of the type of flight, duration, destination etc.

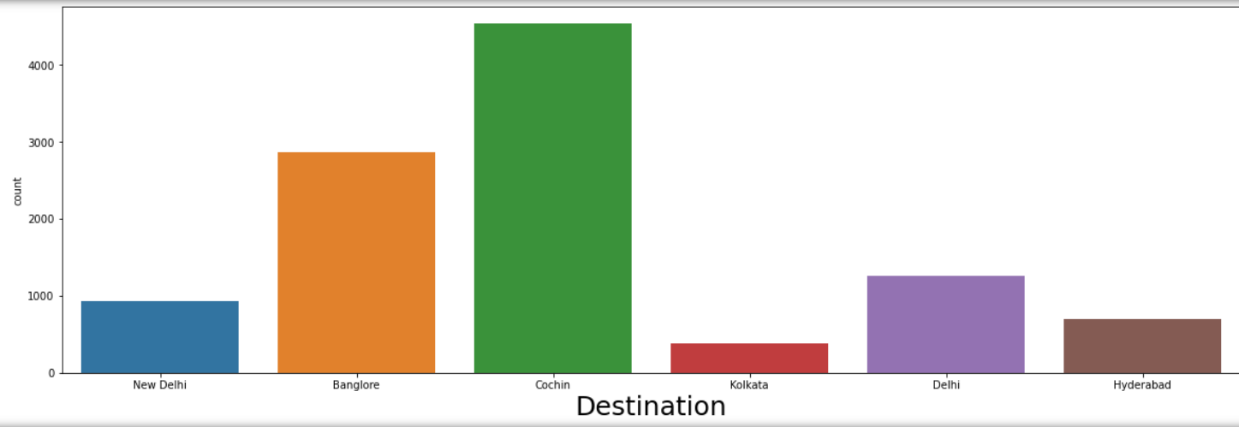
* Of all the airlines, Jet Airways is the most preferred airline by most travelers followed by Indigo and Air India.

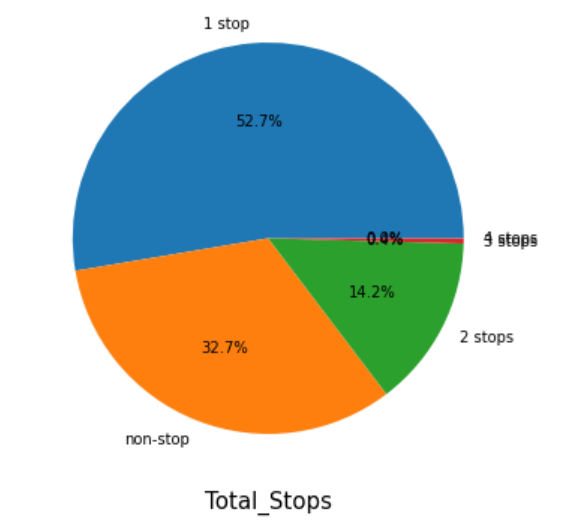
****

* Most of the travelers depart from Delhi followed by Kolkata and Bangalore.

****

* Cochin has the greatest number of destination followed by Bangalore

****

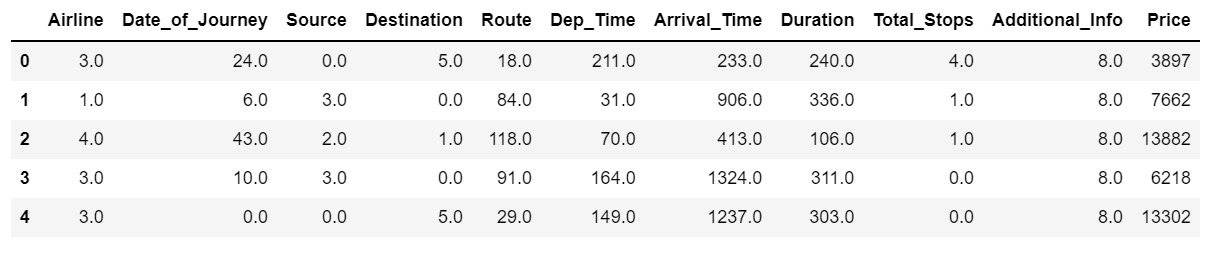
****

* More than half of the travelers prefer 1-stop flight, and around 1/3rd of them prefer non-stop flight. Airlines tend to increase the price when you book a ticket during a short period of time.

When the seating capacity is not yet full, instead of lowering the price to attract people, they tend to increase because they do not want to risk a seat with a low price and also considering some business travelers tend to book tickets at a specified time and book a ticket anyway since they need to be present at a particular location for business meetings. So, they do not mind the expensive ticket rather than late for a meeting.

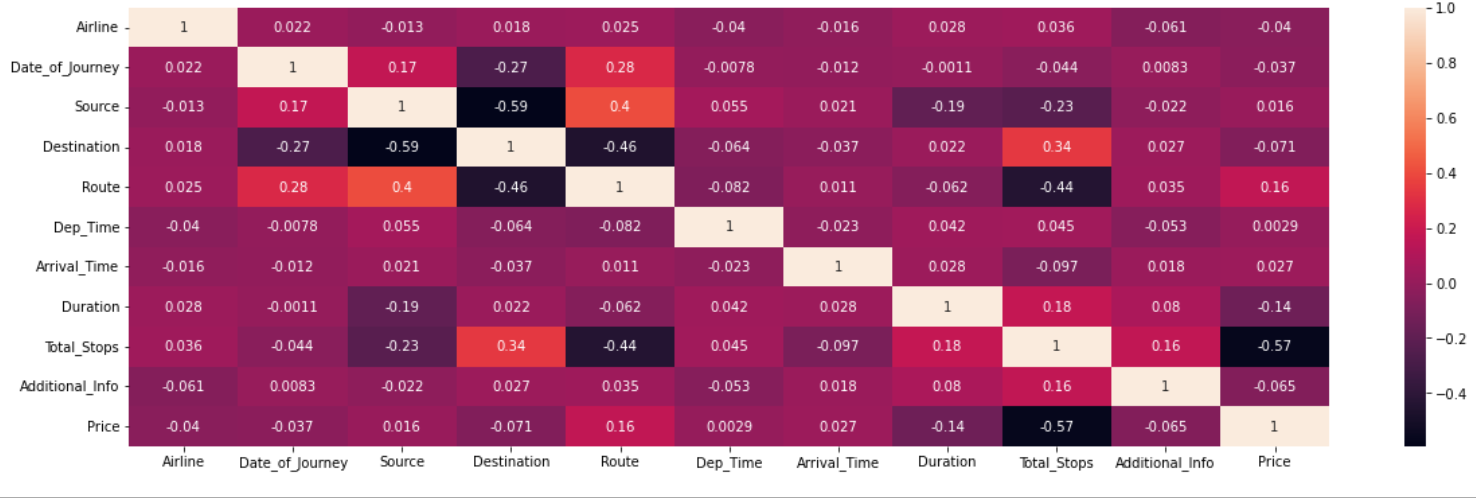
**Encoding the data:**

The categorical data are encoded using Ordinal Encoder so that the machine learning model understands the data. After encoding, the data looks like:



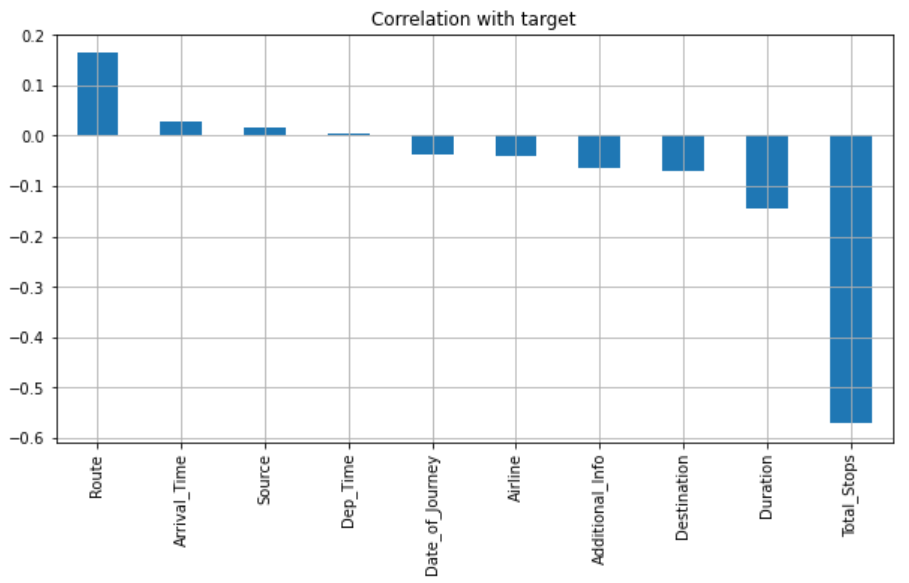
**Checking multicollinearity:**

We will check multicollinearity using heatmap and see how the features are correlated with each other. But our data looks good and there is no multicollinearity. If there is multicollinearity, we can remove the unnecessary features depending on how big the data is.



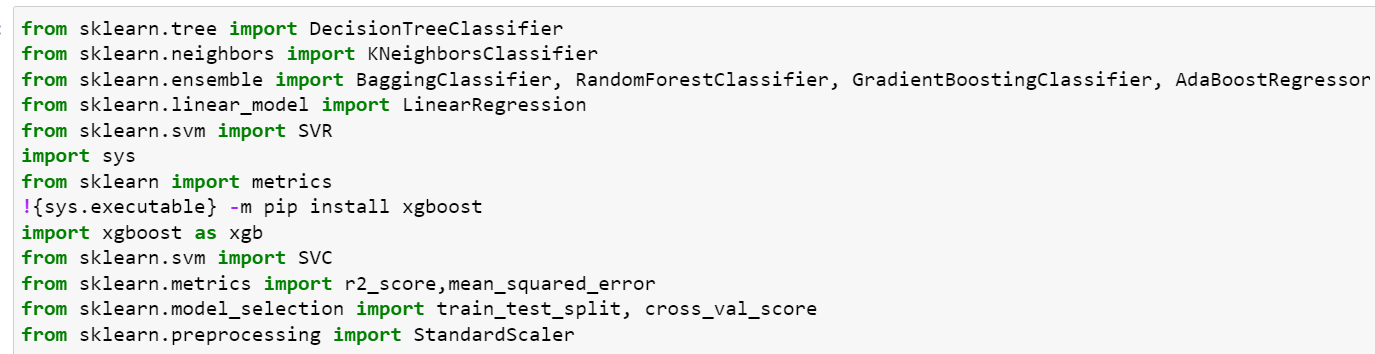
**Checking correlation of features with the target:**

**Total stops** and **route** has the largest correlation with the target variable followed by **duration**.

****

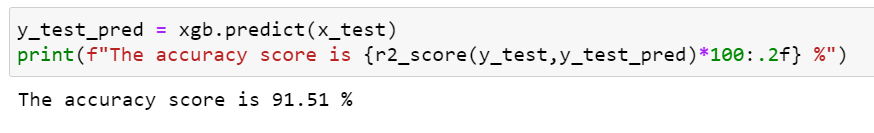
**Important libraries to build the model:**

We need to predict the price using machine learning model. And we must work on different algorithms and build model accordingly and compare the accuracies, and select the algorithm that gives the best accuracy. Also since our problem is a regression problem, we need R2\_score and some metrics to evaluate our model. We need to cross validate the result to see if our model overfits or not. The important libraries imported are as shown:



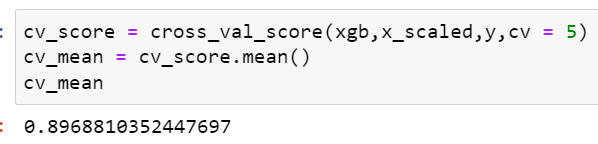
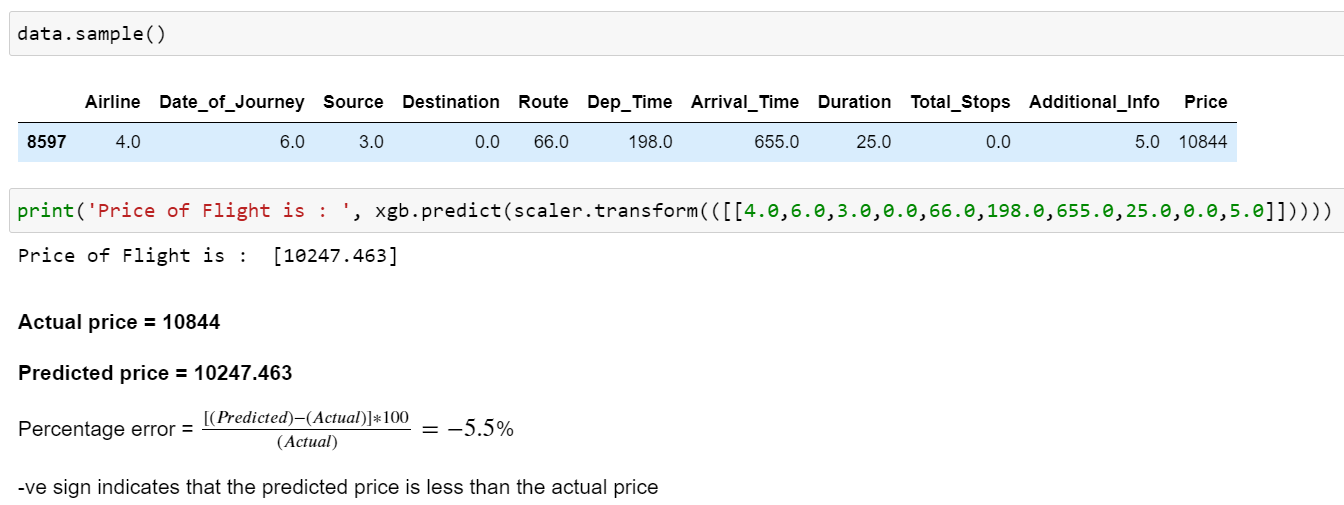
**Building Machine Learning Model:**

* The data needs to be split into train and test dataset so that the machine learns from the train data, and then we test the performance of the model using the test dataset.
* Different algorithms are tried to see which model gives the best accuracy.
* XGBoost Regressor gives the best accuracy out of different algorithms.



Cross Validation is done and it gives 0.89 which shows that our model is not overfitting and looks good.

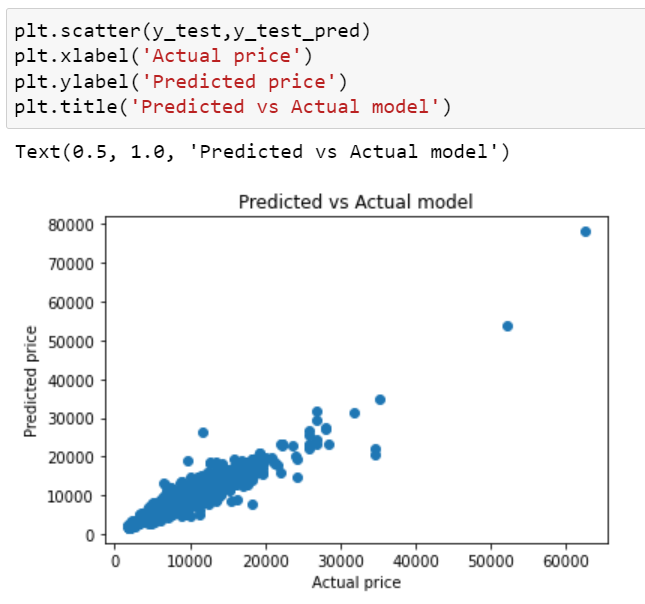
**Now, let’s try to predict the price using the actual data and compare the result with the actual price.**



Here, the negative sign indicates that the predicted price is less than the actual price.

**Let’s see a scatter plot of the predicted price vs the actual price:**

We can see a linear relation between the predicted and the actual price and they do not differ much. The slope is almost linear which shows that our model performs well in predicting the price for most of the data. Most of the prices lie between 5,000 – 20,000. We only see very few prices which lies above 30,000.



**Conclusion:**

Flight prices differ from airline to airline and we see that some features are more important than the others. The route it takes and the number of stops it made has a high relation with the price. The duration of the flight, the distance of the flight and some minor factors affects the price of the flight. The destinations also play an important role since it could be a great place to go for vacation, and that too adds to the price of the flight. So, one must consider different factors to save money. Some travelers might be business persons and some might be flying away from their homes just to spend their leisure time and go for a family trip, with their friends or their life partners.

Business persons would rather be on time to their destinations even if it costs them extra money. But for people who just travel for a vacation, they do not need to reach their destination on a specified period of time. So, they are more flexible in that case and they would plan their trip way before the actual date of flight and in that way, they would be able to save extra money because flight tickets are cheaper if booked months before the actual departure.

As we can observe from the heatmap above, there is no multicollinearity which shows that each feature in the dataset is uniquely important and one must consider all those factors while trying to predict the price of the flight. If there are multiple stops in between the journey, the cost is likely to decrease than those flights which has non-stop journey. So, while booking a ticket, if we want to save extra bucks, we are likely to compromise the duration of the flight and we would reach our destination at a later time. So, time is the cost we have to pay if we want to save money on flights.