Project:
FLIGHT PRICE PREDICTION

Submitted by:
LALBIAK ZAUVA

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 during 30th September till 8th October 2022 and between various cities.

DATA COLLECTION:

Data is collected by a method of web scraping using selenium from the website : https://www.makemytrip.com/

Total number of rows = 1213
Total number of columns = 11

Airline	= Name of the airline
Airline code	= Unique code for each airline
Date of journey	= The date of the journey
Month of journey	= The month of the journey
Source	= The source city from which the service begins
Destination	= The destination city where the service ends
Departure time	= Time when the journey starts from the source
Arrival time	= Time of arrival at the destination
Flight duration	= Total duration of the flight
Type of flight	= Whether the flight is non-stop or not
Price **(target)**	= Total price of the flight

The scraped data is stored as a .csv file which can then later be imported in the jupyter notebook.

LIBRARIES:

Necessary libraries are imported which are

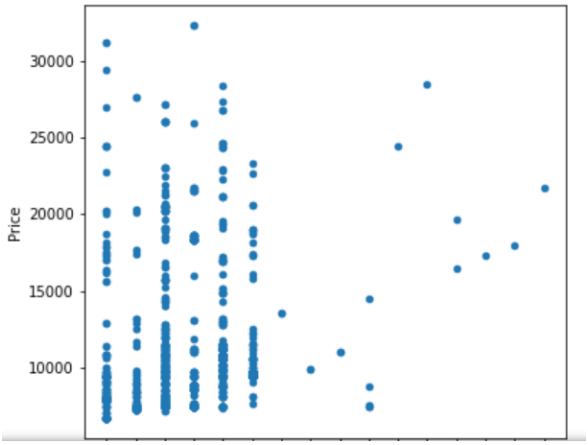
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

The dataframe looks like

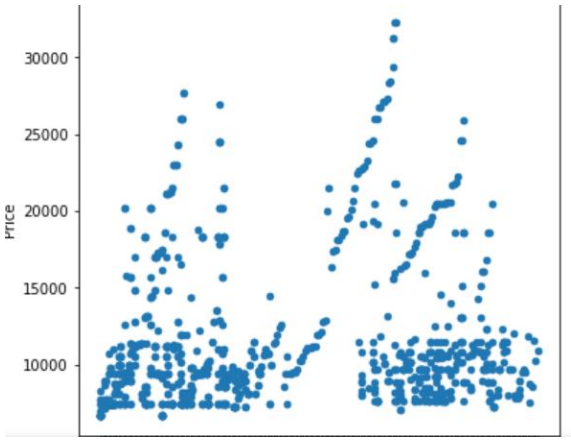
	Airline	Airline code	Date of journey	Month of journey	Source	Destination	Departure time	Arrival time	Flight duration	Type of flight	Price
0	Go First	G8 113	30	9	New Delhi	Bengaluru	05:50	08:35	02 h 45 m	Non stop	7003
1	SpiceJet	SG 191	30	9	New Delhi	Bengaluru	06:05	08:55	02 h 50 m	Non stop	7419
2	IndiGo	6E 2048	30	9	New Delhi	Bengaluru	03:50	06:40	02 h 50 m	Non stop	7424
3	IndiGo	6E 6612	30	9	New Delhi	Bengaluru	06:00	08:35	02 h 35 m	Non stop	7424
4	IndiGo	6E 5009	30	9	New Delhi	Bengaluru	06:55	09:50	02 h 55 m	Non stop	7424

DATA ANALYSIS & EDA:

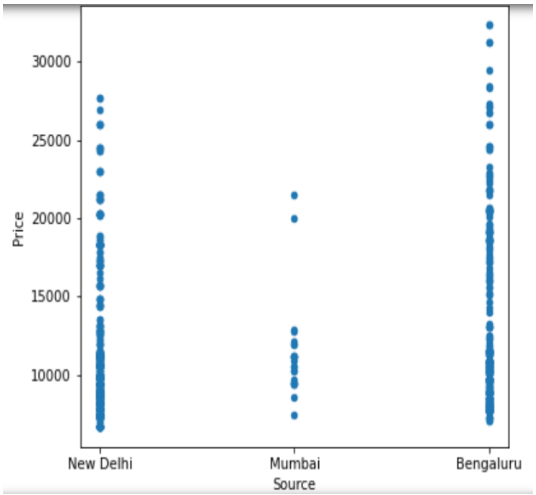
- There are no null values, so we need not fill them.
- There are 8 columns which are of obj data type; while there are 3 columns which are of int64 data type.
- Using .describe() function, the data looks fine in terms of skewness but there might be some skewness, so we need to be sure about this and deal it later.
- A scatter plot is used to see the relation between each feature and the target.

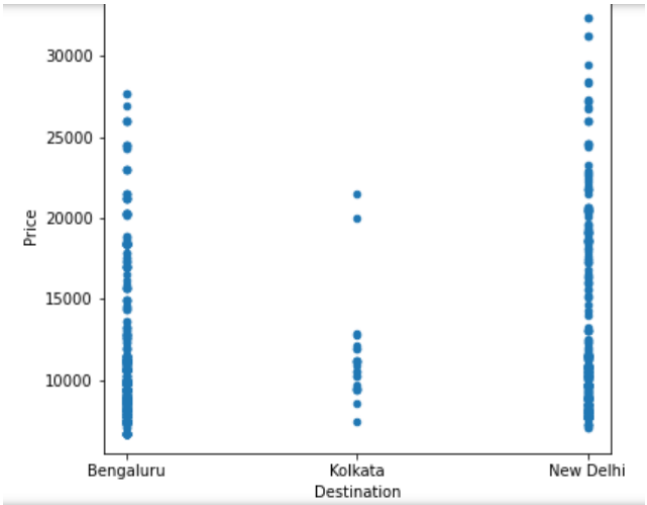


Airline



Airline code





- Flights which offer lower prices are chosen by passengers more frequently
- If source is Delhi, then price is cheap and this may be the reason there are so many flights for passengers since people are attracted by cheaper prices.
- Flight duration which are longer tend to have higher price.

DATA PREPROCESSING

ENCODING THE DATA:

The categorical columns needs encoding since machine only understand numbers and not object data type. So, they are encoded using Ordinal Encoder

```
from sklearn.preprocessing import OrdinalEncoder

oe = OrdinalEncoder()

for i in x.columns:
    if data[i].dtypes == 'object':
        data[i] = oe.fit_transform(data[i].values.reshape(-1,1))
```

```
data.sample()
```

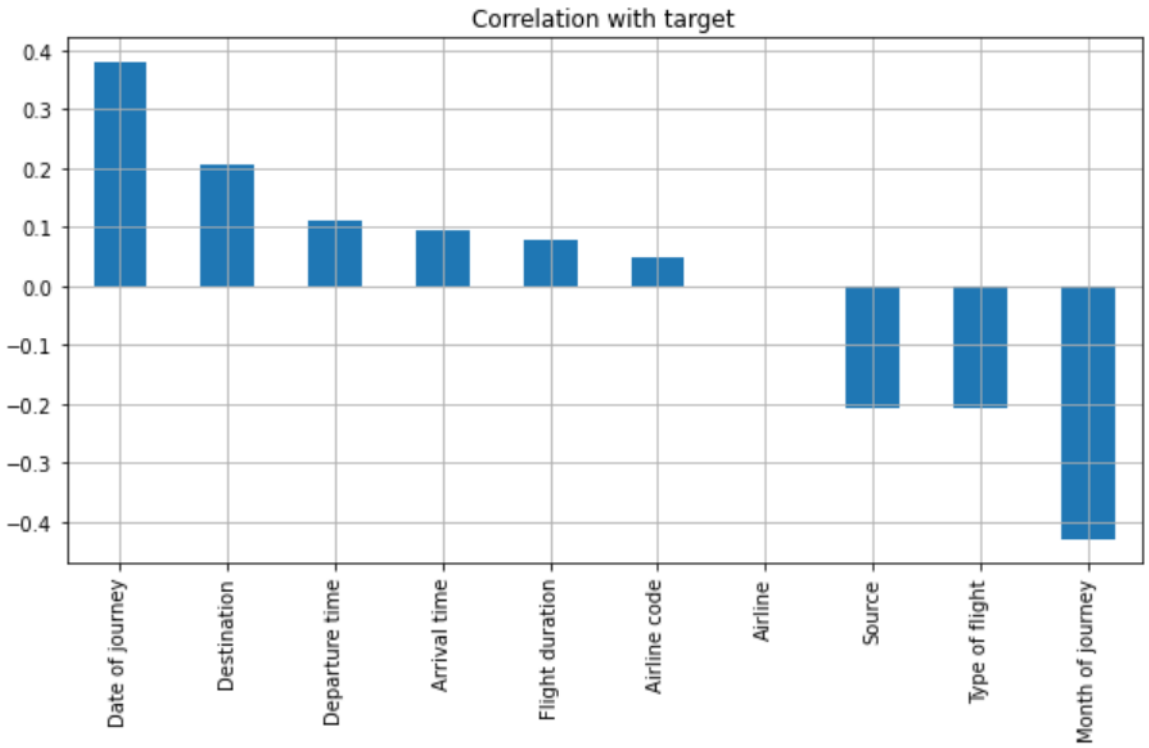
	Airline	Airline code	Date of journey	Month of journey	Source	Destination	Departure time	Arrival time	Flight duration	Type of flight	Price
583	15.0	201.0	3	10	2.0	0.0	69.0	83.0	4.0	24.0	9945

We can see that all the data are numbers and so the machine would understand and we can send it to the machine learning algorithm to build our model.

CHECKING MULTICOLLINEARITY:

We now will check for multicollinearity which means that if there are correlation between the features, we will remove one of them.

It seems like there is no multicollinearity.
We will check correlation between features and target and see which features are more important than the others.



Let's use log transformation to remove the skewness in the data for continuous data

```
data['Date of journey'] = np.log(data['Date of journey'])
data['Month of journey'] = np.log(data['Month of journey'])
```

Since data cleaning process is done and we have done the analysis, let's build our machine learning model to predict the prices of flights.

MODEL BUILDING:

First let's import necessary libraries

```
from sklearn.linear_model import LinearRegression
import sys
from sklearn import metrics
!{sys.executable} -m pip install xgboost
import xgboost as xgb
from sklearn.metrics import r2_score,mean_squared_error
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler
```

We need to split the data into **train** and **test** data and then test for the best machine learning algorithm which gives the best accuracy.

Since ours is a regression problem, we will calculate the R2 score.

Out of different algorithms, XBGRegressor works best which gives an accuracy of 88.85 %

```
y_test_pred = xgb.predict(x_test)
print(f"The accuracy score is {r2_score(y_test,y_test_pred)*100:.2f} %")
```

```
The accuracy score is 88.85 %
```

After tuning the parameters, the accuracy slightly improved to 89.11 %. Here RandomizedSearchCV is used.

Testing the model:
Now, we take one sample data and try to predict the price and then compare with the actual price and see how close we get.

	Airline	Airline code	Date of journey	Month of journey	Source	Destination	Departure time	Arrival time	Flight duration	Type of flight	Price
346	0.0	94.0	1.94591	2.302585	2.0	0.0	90.0	102.0	6.0	24.0	9420

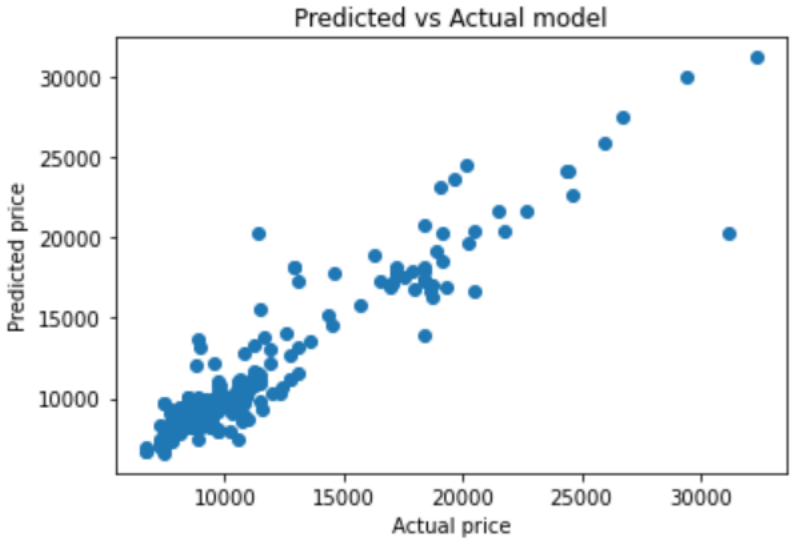
```
print('Price of Flight is : ', xgb.predict(scaler.transform([[0.0,94.0,1.94591,2.302585,2.0,0.0,90.0,102.0,6.0,24.0]])))
```

```
Price of Flight is :  [9865.034]
```

```
print(f'Percentage error is {((9865.034-9420))/9420)*100:.2f} %')
```

```
Percentage error is 4.72 %
```

The scatterplot of predicted and actual price is plotted



Finally, the model is saved in a pickle format for later purpose.

```
import pickle
import joblib

joblib.dump(xgb,'Flight-price.pkl')
```

```
['Flight-price.pkl']
```

CONCLUSION AND FURTHER IMPROVEMENTS:

We now know that flight prices depends on various factors, and they changes frequently. The route also plays an important role in the price of the flight. Longer routes costs higher than shorter routes.

Since I have collected only 1213 records, the model could perform much better if there are more data to train to. The range of the month from which the data is collected could be increased, and also more cities could be added given the time.