FLIP ROBO

FLIGHT PRICE PREDICTION PROJECT

Submitted by:

Priya Rajput

ACKNOWLEDGMENT

I would like to express my gratitude towards Flip-Robo for providing me this opportunity to show case my talent and also for their constant support and guidance. Also It is indeed a pleasure for me to have worked on this project.

I express my deepest thanks to Miss Sapna Verma, for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make my project easier. I choose this moment to acknowledge his contribution gratefully.

TABLE OF CONTENTS

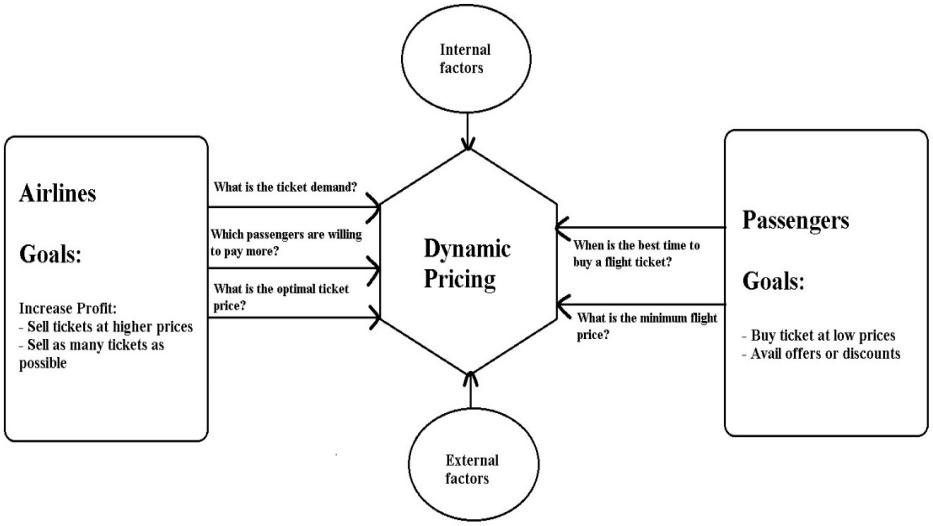
INTRODUCTION

Business Problem Framing

The airline industry is considered as one of the most sophisticated industry in using complex pricing strategies. Nowadays, ticket prices can vary dynamically and significantly for the same flight, even for nearby seats. The ticket price of a specific flight can change up to 7 times a day. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible and maximize their profit. However, mismatches between available seats and passenger demand usually leads to either the customer paying more or the airlines company losing revenue. Airlines companies are generally equipped with advanced tools and capabilities that enable them to control the pricing process. However, customers are also becoming more strategic with the development of various online tools to compare prices across various airline companies. In addition, competition between airlines makes the task of determining optimal pricing is hard for everyone.

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. 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)



Conceptual Background of the Domain Problem

1. Collect the data.
2. Check whether the project is a regression type or a classification type.

3. Check whether our dataset is balanced or imbalanced. If it is an imbalanced one, we will apply sampling techniques to balance the dataset.
4. Model building and find the accuracy of the model.
5. Build a model with good accuracy and also go for hyperparameter tuning.

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. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That's why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

Review of Literature

As per the requirement of client, I have scrapped the data from online sites and based on that data I have did analysis like for based on which feature of my data prices are changing and checked the relationship of flight price with all the feature like what flight he should choose.

Motivation

Our main objective of doing this project is to build a model to predict whether the users are paying the loan within the due date or not. We are going to predict by using Machine Learning algorithms.

The sample data is provided to us from our client database. I have worked on this on the bases of client requirements and followed all the steps till model deployment

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

In our scrapped dataset, our target variable "Price" is a continuous variable. Therefore, we will be handling this modelling problem as regression.

This project is done in three parts:

- Data Collection
- Data Analysis
- Model Building

1. Data Collection

We have to scrape at least 1500 rows of data. You can scrape more data as well, it's up to you, more the data better the model. In this section you have to scrape the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc). The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

2. Data Analysis

After cleaning the data, you have to do some analysis on the data. Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time? What is the best time to buy so that the consumer can save the most by taking the least risk? Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

3. Model Building

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Follow the complete life cycle of data science. Include all the steps like

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

DATA SOURCES AND THEIR FORMATS

Data Source: The read_csv function of the pandas library is used to read the content of a CSV file into the python environment as a pandas DataFrame. The function can read the files from the OS by using proper path to the file.

Data description: Pandas describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values. When this method is applied to a series of string, it returns a different output which is shown below.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
pd.set_option('display.max_rows',None)
```

```
df=pd.read_excel(r'C:\Users\Admin\Desktop\FlightDataWebscrape1.xls')
df.head()
```

	Unnamed: 0	Airline	Flight_Number	Date_of_Departure	From	To	Duration	Total_Stops	Price
0	0	Indigo	6E-5023	Apr 29\n7319\nApr 30\n7318\nMay 01\n7318\nMay ...	Delhi	Mumbai	02h 05m	non-stop	7319.0
1	1	Indigo	6E-6814	Apr 29\n7319\nApr 30\n7318\nMay 01\n7318\nMay ...	Delhi	Mumbai	02h 05m	non-stop	7319.0
2	2	AirAsia	I5-764	Apr 29\n7319\nApr 30\n7318\nMay 01\n7318\nMay ...	Delhi	Mumbai	02h 10m	non-stop	7319.0
3	3	Indigo	6E-2112	Apr 29\n7319\nApr 30\n7318\nMay 01\n7318\nMay ...	Delhi	Mumbai	02h 10m	non-stop	7319.0
4	4	GO FIRST	G8- 530	Apr 29\n7319\nApr 30\n7318\nMay 01\n7318\nMay ...	Delhi	Mumbai	02h 10m	non-stop	7319.0

Data Pre-processing Done

For the data pre-processing step, I checked through the dataframe for missing values and renamed values that needed a better meaningful name.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1422 entries, 0 to 1421
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Unnamed: 0       1422 non-null   int64  
 1   Airline          1422 non-null   object 
 2   Flight_Number    1422 non-null   object 
 3   Date_of_Departure  1422 non-null   object 
 4   From             1422 non-null   object 
 5   To               1422 non-null   object 
 6   Duration         1422 non-null   object 
 7   Total_Stops      1422 non-null   object 
 8   Price            1197 non-null   float64
dtypes: float64(1), int64(1), object(7)
memory usage: 100.1+ KB
```

```
df.isnull().sum()
```

```
Unnamed: 0           0
Airline              0
Flight_Number        0
Date_of_Departure    0
From                 0
To                   0
Duration             0
Total_Stops          0
Price              225
dtype: int64
```

```
# Replacing categorical values with numeric data
df.Total_Stops.replace({"non-stop": 0,"1-stop": 1,"2+-stop": 2,"1-stop Via IXU": 3,"1-stop Via IDR": 4,"1-stop Via Indore": 5,"1-
```

```
#filling the float missing data with mean
df['Price'].fillna(df['Price'].mean(),inplace=True)
```

```
# Replacing categorical values with numeric data
df.Airline.replace({"Vistara": 0,"Indigo": 1,"Air India": 2,"GO FIRST": 3,"SpiceJet": 4,"AirAsia": 5,"AllianceAir": 6},inplace =
```

```
# Replacing categorical values with numeric data
df.From.replace({"Delhi": 0,"Mumbai": 1,"Kolkata": 2},inplace = True)
```

```
# Replacing categorical values with numeric data
df.To.replace({"Bangalore": 0,"Delhi": 1,"Goa": 2,"Mumbai": 3,"Pune":4,"Kolkata":5,"Patna":6},inplace = True)
```

```
df.describe()
```

	Airline	From	To	Total_Stops	Price	Duration_hour	Duration_min
count	1422.000000	1422.000000	1422.000000	1422.000000	1422.000000	1422.000000	1422.000000
mean	1.776371	0.573840	1.921941	0.917018	15506.781955	10.427567	25.903657
std	1.630741	0.678278	1.673643	0.705965	6081.842455	8.023191	16.839763
min	0.000000	0.000000	0.000000	0.000000	6258.000000	1.000000	0.000000
25%	0.000000	0.000000	1.000000	1.000000	10843.250000	5.000000	10.000000
50%	1.000000	0.000000	2.000000	1.000000	15506.781955	8.000000	25.000000
75%	3.000000	1.000000	3.000000	1.000000	18438.000000	15.000000	40.000000
max	6.000000	2.000000	6.000000	7.000000	51000.000000	47.000000	55.000000

- Dropped the column ‘Unnamed:0’ :as it is of no use, it is only giving serial numbers starting from 1, ‘Date of journey’ and ‘flight number’ also dropped as not giving any relevant information in building the model.
- Checked the correlation between dependant and independent variables using heatmap.
- Visualization using distribution plots where I can clearly see some outliers and skewness from the plots.
- Checked outliers using boxplots and also removed them using zscore.
- Splitted the dependant and independent variables into x and y.

Data Inputs- Logic- Output Relationships

EDA was performed by creating valuable insights using various visualization libraries.

Hardware and Software Requirements and Tools Used

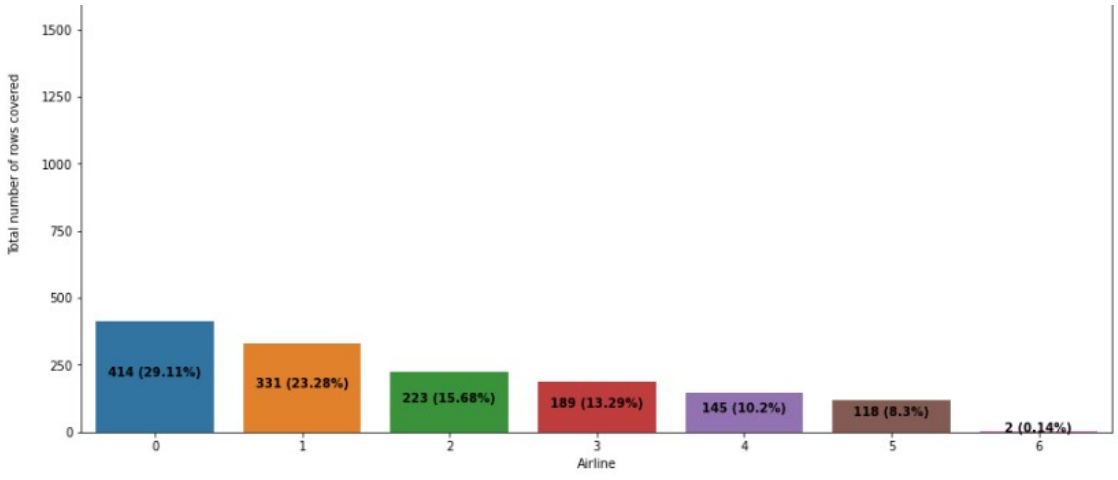
Hardware required:
Processor: core i3
RAM: 8 GB
Software required:
Anaconda 3- language used Python 3
Microsoft Excel

Exploratory Data Analysis

Let us explore our data and visualize

```
try:
    x = 'Airline'
    k=0
    plt.figure(figsize=[15,8])
    axes = sns.countplot(df[x])
    for i in axes.patches:
        ht = i.get_height()
        mr = len(df[x])
        st = f"{ht} ({round(ht*100/mr,2)}%)"
        plt.text(k, ht/2, st, ha='center', fontweight='bold')
        k += 1
    plt.ylim(0,2000)
    plt.title(f'Count Plot for {x} column\n', fontsize = 20)
    plt.ylabel(f'Total number of rows covered\n')
    plt.show()

except Exception as e:
    print("Error:", e)
    pass
```



Observation:

Highest number of airline preferred by people are vistara covering 29.11% of the total record

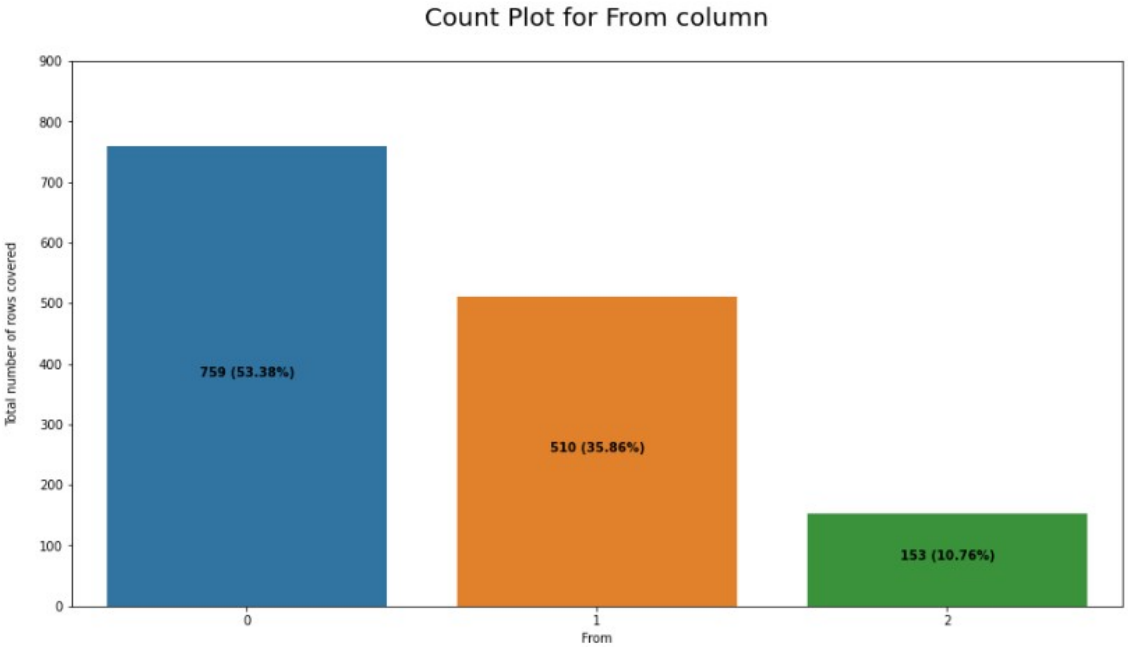
We can see that Indigo is close to the first one

At third place we have Air India airlines that covers 15.68% of total record in our airline data

Airlines Go First, Air Asia and SpiceJet are the least used by people covering 10.2%, 8.3% and 0.14% respectively

```
try:
    x = 'From'
    k=0
    plt.figure(figsize=[15,8])
    axes = sns.countplot(df[x])
    for i in axes.patches:
        ht = i.get_height()
        mr = len(df[x])
        st = f"{ht} ({round(ht*100/mr,2)}%)"
        plt.text(k, ht/2, st, ha='center', fontweight='bold'
        k += 1
    plt.ylim(0,900)
    plt.title(f'Count Plot for {x} column\n', fontsize = 20)
    plt.ylabel(f'Total number of rows covered\n')
    plt.show()

except Exception as e:
    print("Error:", e)
    pass
```



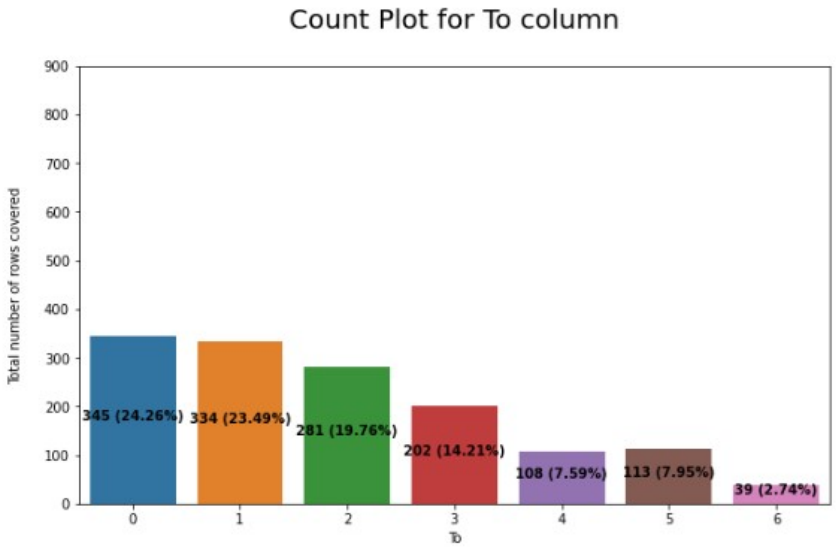
Observation:

The departure area or source place highly used or people majorly flying from the city is "Delhi" covering 53.38% record in the column

on second we have Mumbai with 35..86% records in the column and Kolkata have 10.76%

```
try:
    x = 'To'
    k=0
    plt.figure(figsize=[10,6])
    axes = sns.countplot(df[x])
    for i in axes.patches:
        ht = i.get_height()
        mr = len(df[x])
        st = f"{ht} ({round(ht*100/mr,2)}%)"
        plt.text(k, ht/2, st, ha='center', fontweight='bold')
        k += 1
    plt.ylim(0,900)
    plt.title(f'Count Plot for {x} column\n', fontsize = 20)
    plt.ylabel(f'Total number of rows covered\n')
    plt.show()

except Exception as e:
    print("Error:", e)
    pass
```

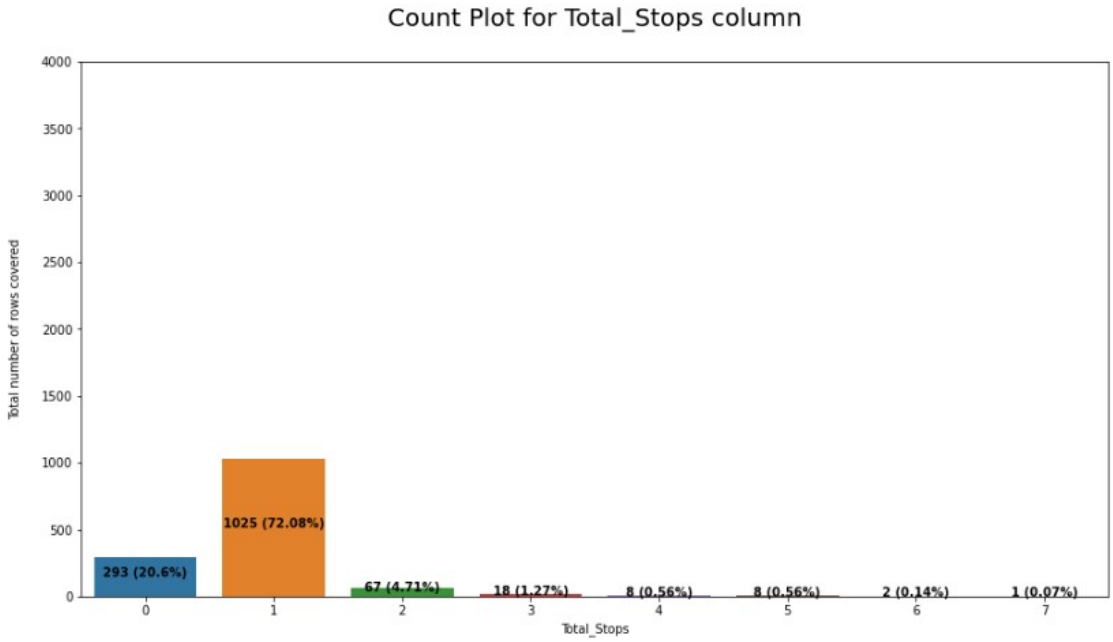


Observation: destination place people prefer to fly towards the city "Bangalore" covering 24.26% of record

Again in a similar fashion "Delhi" city is a close second destination that people like to fly towards covering 23.49% record in the column

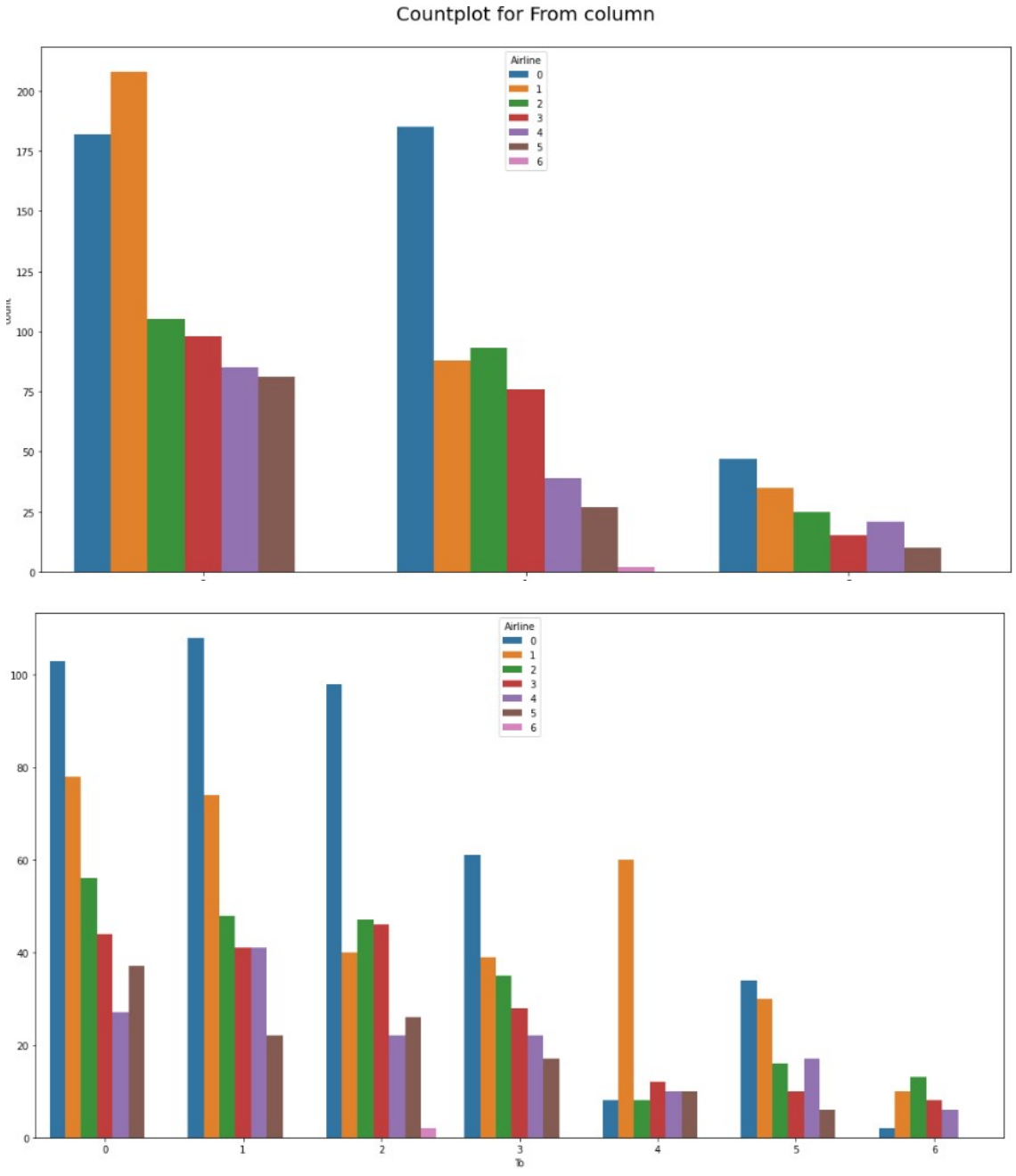
```
try:
    x = 'Total_Stops'
    k=0
    plt.figure(figsize=[15,8])
    axes = sns.countplot(df[x])
    for i in axes.patches:
        ht = i.get_height()
        mr = len(df[x])
        st = f"{ht} ({round(ht*100/mr,2)}%)"
        plt.text(k, ht/2, st, ha='center', fontweight='bold')
        k += 1
    plt.ylim(0,4000)
    plt.title(f'Count Plot for {x} column\n', fontsize = 20)
    plt.ylabel(f'Total number of rows covered\n')
    plt.show()

except Exception as e:
    print("Error:", e)
    pass
```




```
x = "From"
plt.figure(figsize=(18,10))
sns.countplot(x = x, hue = "Airline", data = df)
plt.title(f"Countplot for {x} column\n", fontsize = 20)
plt.show()

x = "To"
plt.figure(figsize=(18,10))
sns.countplot(x = x, hue = "Airline", data = df)
plt.title(f"Countplot for {x} column\n", fontsize = 20)
plt.show()
```

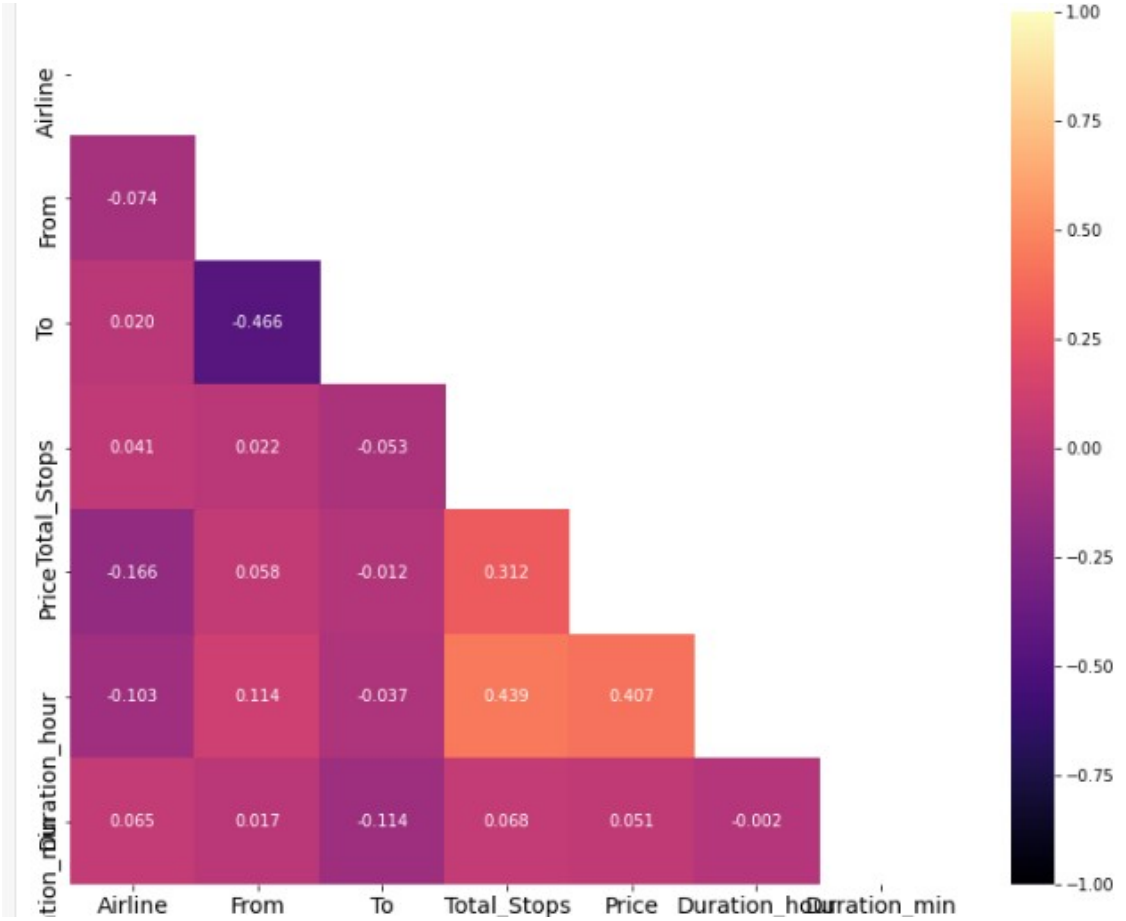


Observation:

Checking out the Source place details for each and every airline we can see that Bengluru city has the highest number of departure flights for Vistara airlines

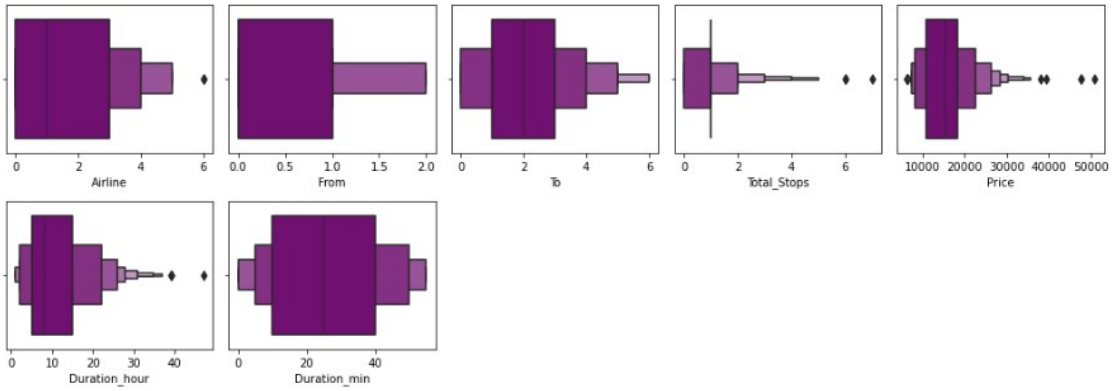
Indigo and Air India are the airlines that are used in almost all the cities to depart while the other airlines do not cover some or the other city

Heatmap



This is the heatmap where I have checked the correlation between the data and also got to know that there are columns or independent variables which are 0% correlated with the target variable.

```
plt.figure(figsize=(14,7))
outl_df = df.columns.values
for i in range(0, len(outl_df)):
    plt.subplot(3, 5, i+1)
    ax = sns.boxenplot(df[outl_df[i]], color='purple')
    plt.tight_layout()
```



Interpretation of the Results

In the visualization part, we have seen how my data looks like using heatmap, boxplot, distribution plots. In the pre-processing part, we have cleaned data using zscore.

In the modeling part, we have designed our model using algorithms like Random Forest Classifier. The accuracy score, confusion_matrix, classification_report are achieved for each model.

```
# Choosing Extra Trees Regressor

fmod_param = {'n_estimators' : [100, 200, 300],
              'criterion' : ['squared_error', 'mse', 'absolute_error', 'mae'],
              'n_jobs' : [-2, -1, 1]
             }

GSCV = GridSearchCV(ExtraTreesRegressor(), fmod_param, cv=5)
GSCV.fit(x_train,y_train)
```

```
GridSearchCV(cv=5, estimator=ExtraTreesRegressor(),
             param_grid={'criterion': ['squared_error', 'mse', 'absolute_error',
                                       'mae'],
                         'n_estimators': [100, 200, 300],
                         'n_jobs': [-2, -1, 1]})
```

CONCLUSION

Key Findings and Conclusions of the Study

In this project we have scraped the flight data from airline webpages. Then the comma separated value file is loaded into a data frame. Looking at the data set we understand that there are some features needs to be processed like converting the data types and get the actual value from the string entries from the time related columns. After the data is been processed, I have done some EDA to understand the relation among features and the target variable. Features like flight duration, number of stops during the journey and the availability of meals are playing major role in predicting the prices of the flights. As we have seen, the prediction is showing a similar relationship with the actual price from the scrapped data set. This means the model predicted correctly and it could help airlines by predicting what prices they can maintain. It could also help customers to predict future flight prices and plan the journey accordingly because it is difficult for airlines to maintain prices since it changes dynamically due to different conditions. Hence by using Machine Learning techniques we can solve this problem. The above research will help our client to study the latest flight price market and with the help of the model built he can easily predict the price ranges of the flight, and also will helps him to understand Based on what factors the fight price is decided.

Learning Outcomes of the Study in respect of Data Science

Visualization part helped me to understand the data as it provides graphical representation of huge data. It assisted me to understand the feature importance, outliers/skewness detection and to compare the independent-dependent features. Data cleaning is the most important part of model building and therefore before model building, I made sure the data is cleaned. I have generated multiple regression machine learning models to get the best model wherein I found Extra Trees Regressor Model being the best based on the metrics I have used.

Limitations of this work and Scope for Future Work

As looking at the features we came to know that the numbers of features are very less, due to which we are getting somewhat lower R2 scores. Some algorithms are facing over-fitting problem which may be because of a smaller number of features in our dataset. We can get a better R2 score than now by fetching some more features from the web scraping by that we may also reduce the over fitting problem in our models. Another limitation of the study is that in the volatile changing market we have taken the data, to be more precise we have taken the data at the time of pandemic and recent data, so when the pandemic ends the market correction might happen slowly. So, based on that again the deciding factors of it may change and we have shortlisted and taken these data from the important cities across India. If the customer is from the different country our model might fail to predict the accuracy prize of that flight.