FLIGHT PRICE PREDICTION

Submitted by:

YASH JAISWAL

ACKNOWLEDGMENT

I would like to thank you Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot.

INTRODUCTION

- Business Problem Framing

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 –

So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

- Conceptual Background of the Domain Problem

Flight Price Order and Time of purchase patterns making sure last-minute purchases are expensive 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 This usually happens as an attempt to maximize revenue based on

- Review of Literature

This is a comprehensive summary of the research done on the behalf You 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).

- Motivation for the Problem Undertaken

In today's world of globalization, people travel from one country to another for the purpose of business or leisure. High competition in this field has prompted

airlines across the world to capture the large volume markets. Cheap flight deals give travelers an opportunity to travel to places that otherwise were not excisable due to the high airline ticket prices. Time of purchase patterns making sure last-minute purchases are expensive 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.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

We are building a model in machine learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So this model will help us to determine which variable are import predict the price of variable and also how these variable describe the price of the Flight.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3202 entries, 0 to 125
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      3202 non-null   int64
 1   Airline_name    3202 non-null   object
 2   From            3202 non-null   object
 3   To              3202 non-null   object
 4   Arival_time     3202 non-null   object
 5   Stops           3202 non-null   object
 6   Price           3202 non-null   int32
 7   Depart hour     3202 non-null   int64
 8   Depart min      3202 non-null   int64
 9   Duration_hours  3202 non-null   int64
 10  Duration_mins   3202 non-null   int64
dtypes: int32(1), int64(5), object(5)
memory usage: 287.7+ KB
```

For specific mathematical reason this allows to research to estimate the conditional expectation of the dependent variable when the in independent variable take on a given modelling technique which investigates variable (predictor) this technique is used for d=forecasting time series modelling and finding the causal effect relationship between the variable.

- Data Sources and their formats

Data scraping by yatra.com with useful web driver to scrape the data by yatra.com and airindia.com these columns are airline name,

data of journey, source, destination, route, departure time, arrival time , duration, total stops and price.

```
driver = webdriver.Edge(r"C:/Users/yjjai/Downloads/edgedriver_win64/msedgedriver.exe")
driver.maximize_window()
```

```
url = "https://flight.yatra.com/air-search-ui/dom2/trigger?type=O&viewName=normal&flexi=
driver.get(url)
```

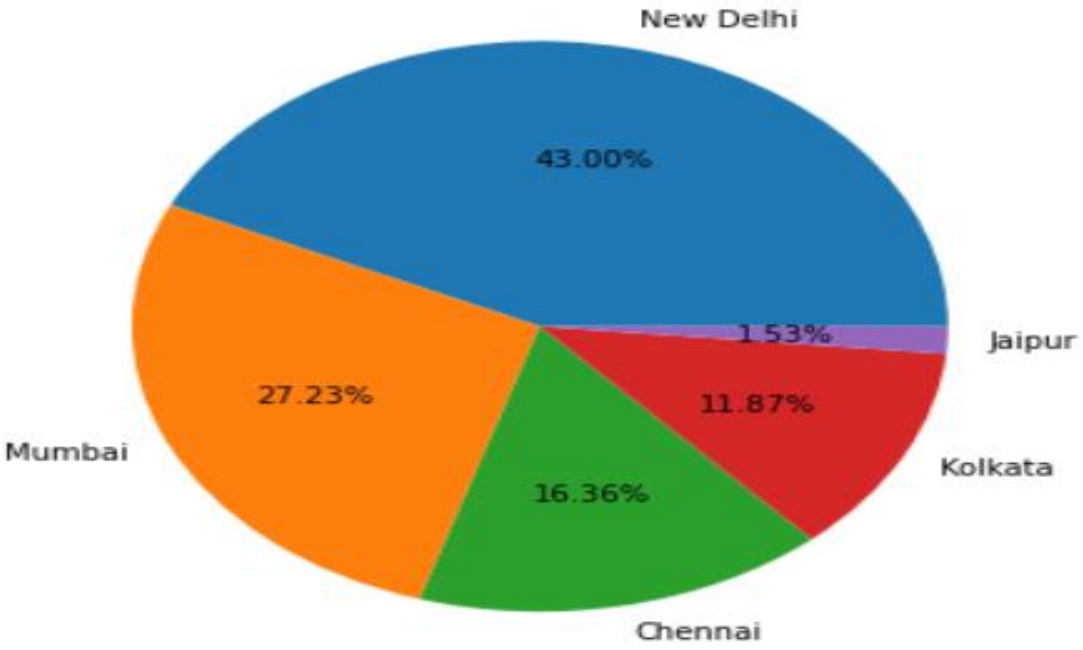
```
#scraping the data of Airline_name
try:
    air_name = driver.find_elements(By.XPATH,'//span[@class="i-b text ellipsis"]')
    for os in air_name:
        Airline_name.append(os.text)
except NoSuchElementException:
    Airline_name.append('-')


#scraping the data of Depart_time
try:
    air_dec = driver.find_elements(By.XPATH,'//div[@class="i-b pr"]')
    for os in air_dec:
        Depart_time.append(os.text)
except NoSuchElementException:
    Depart_time.append('-')


#scraping data of display of the Arival_time
try:
    air_ari = driver.find_elements(By.XPATH,'//p[@class="bold fs-15 mb-2 pr time"]')
    for disp in air_ari:
        Arival_time.append(disp.text)
except NoSuchElementException:
    Arival_time.append("-")
```

Then all columns are join to with method with concat to format of csv (comma separated value) the dimension of data is 3202 rows and 9 columns. These are 30 data set to csv format to join one data set.

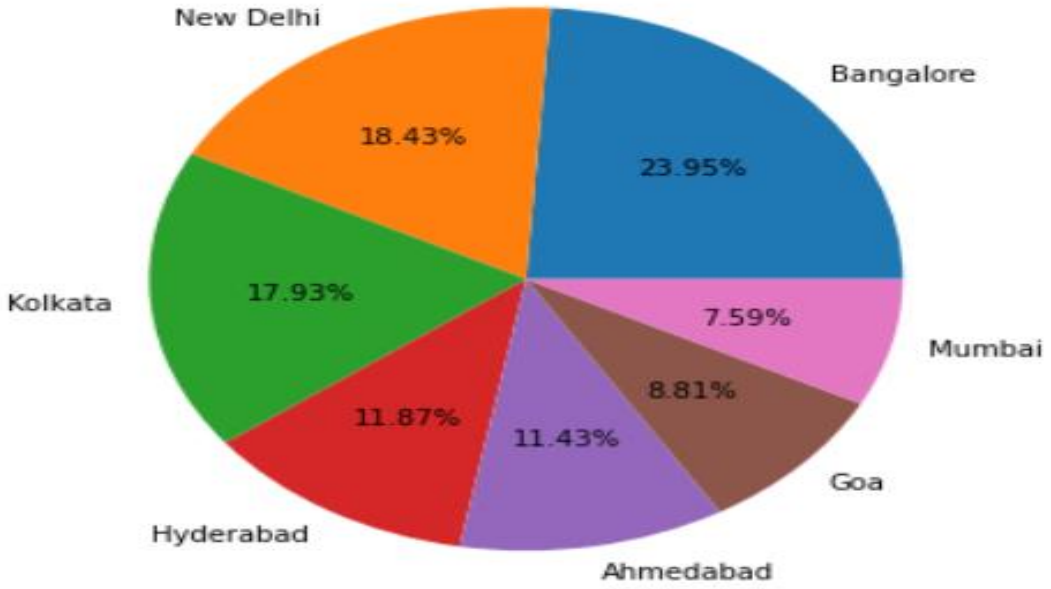
- Data Preprocessing Done

Loading the training data set as a dataframe , used pandas to set display I ensuring we do not see any truncated information , checked the number of rows and columns present in our training data set , checked for missing data and the number of rows with null values , verified the percentage of missing data in each columns are decide to dicard the once that value more than 50% of null value , dropped all the unwanted columns are duplicated data present in our data frame, separated categorical columns and numeric columns name in separate list variable for ease in visulazation , checked the unique values information in each column to get a gist for categorical data. Used pandas profiling during the visulazing phase along with pie plot count plot scatter plot and the other , with the help of label encoding technique converted all object data type columns to numeric data types.

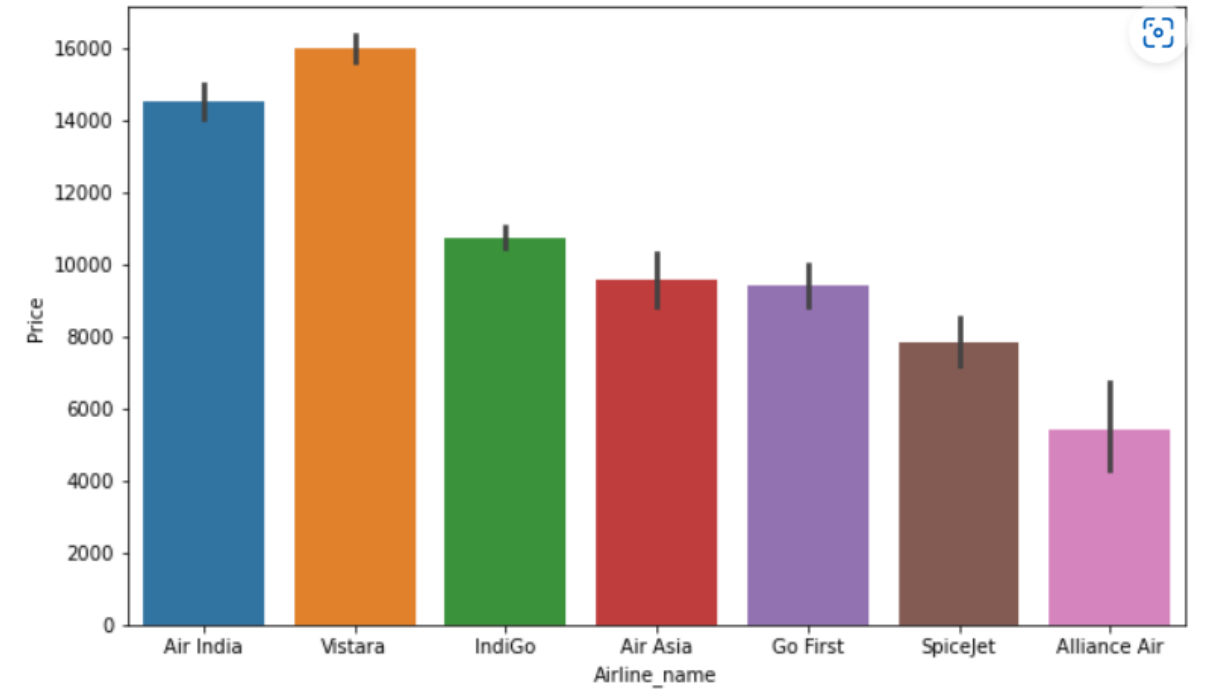
- Data Inputs- Logic- Output Relationships



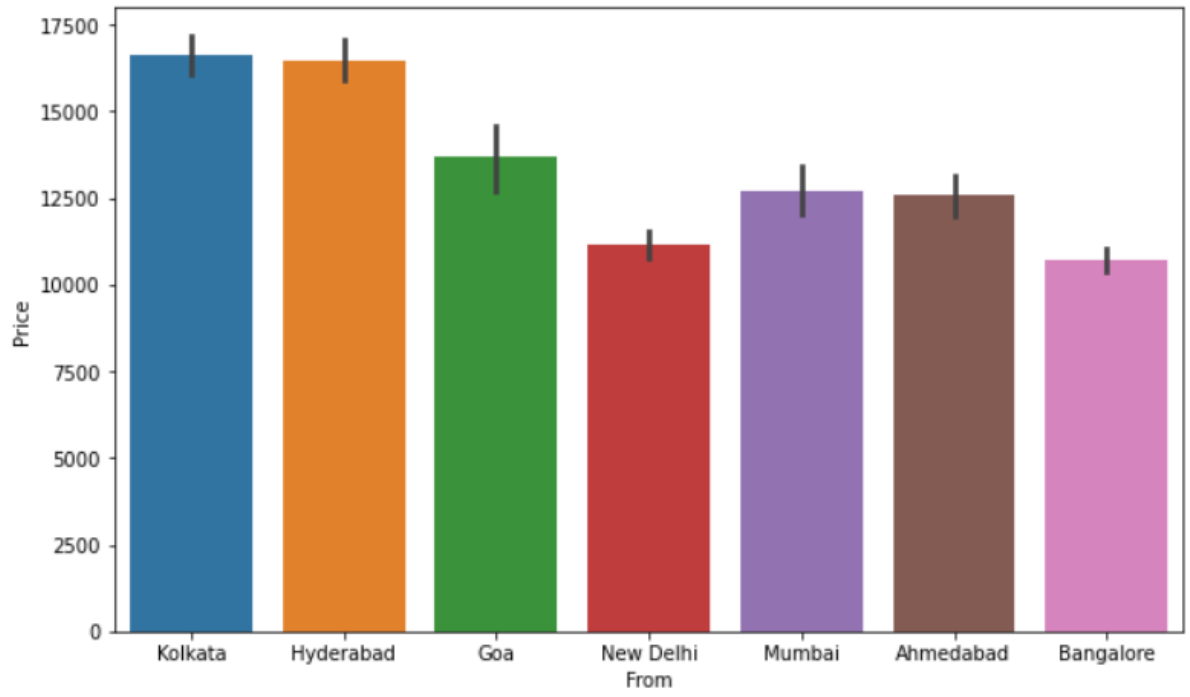
Arrival people are more then high New Delhi to arrive to from anyware 43% are our data scarping to high New Delhi to going any ware.



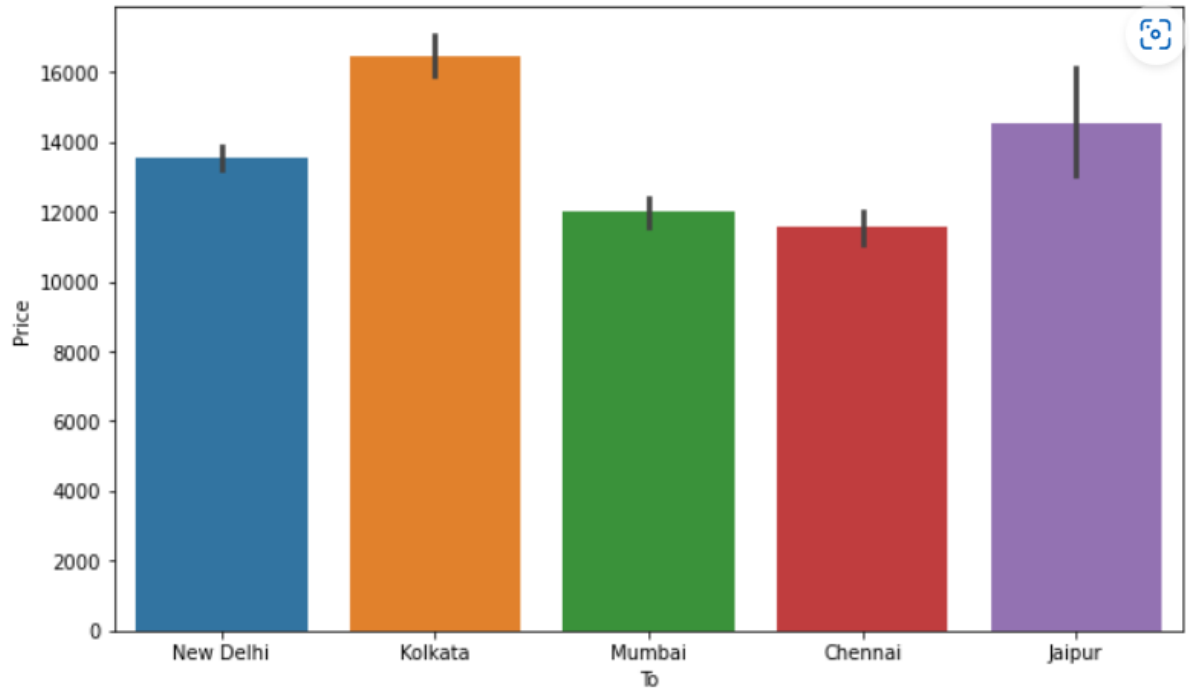
Arrival people are more than going to Bangalore into these data scrape in data scraping to 23.95% people are going to bangalore.



Airline name are Vistara are more 16000 people are likely to going in Vistara airline india also air india are highly .



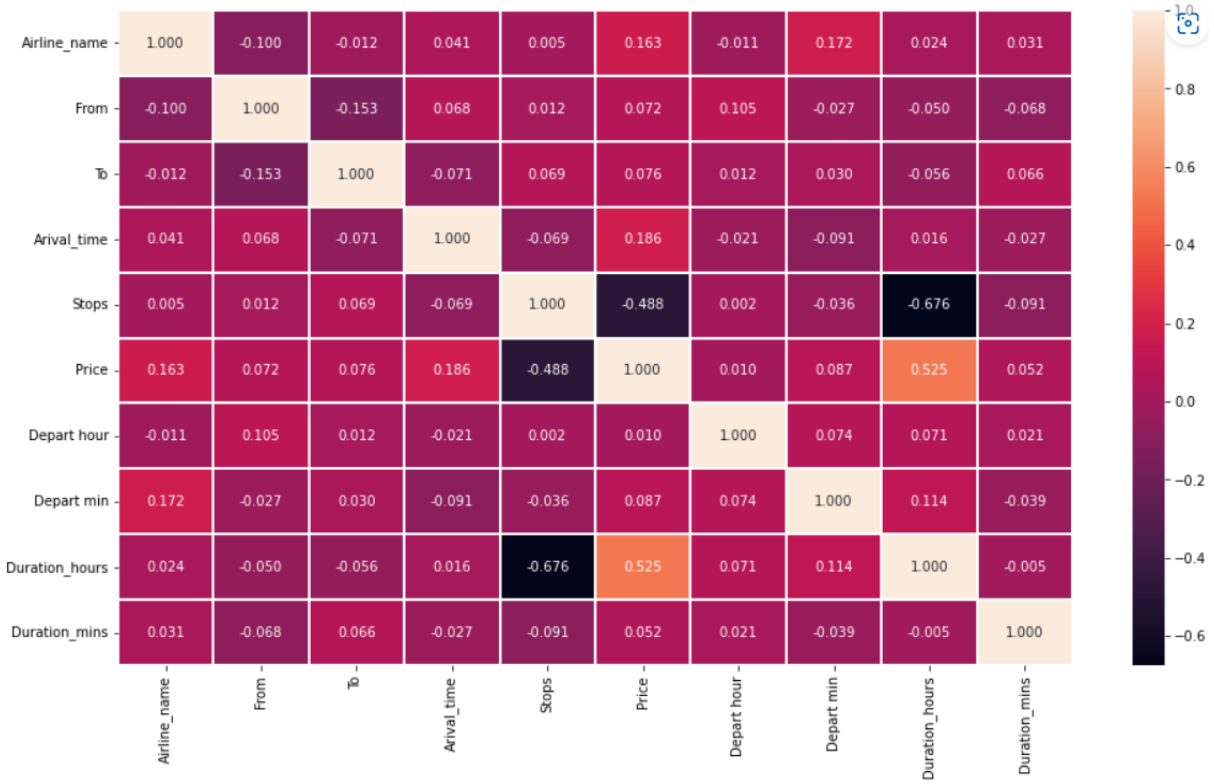
From and price relationship are strong to the Kolkata and Hyderabad are very strong .

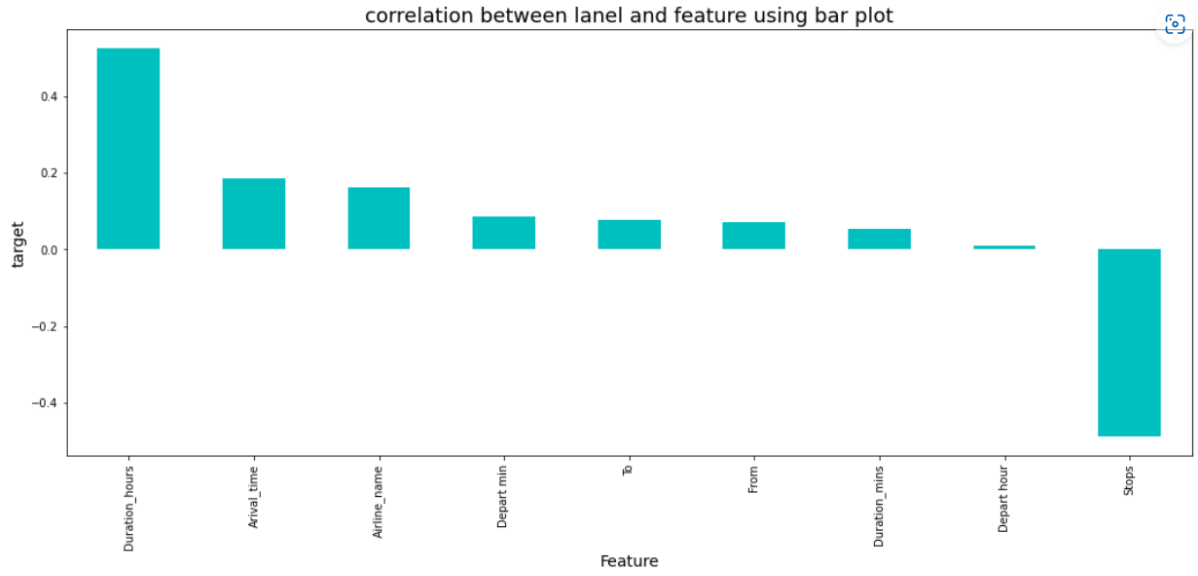


To and price relationship are strong to the new delhi and Kolkata are nice relationship.

- State the set of assumptions (if any) related to the problem under consideration

	Airline_name	From	To	Arival_time	Stops	Price	Depart hour	Depart min	Duration_hours	Duration_mins
Airline_name	1.000000	-0.100111	-0.012070	0.040515	0.005466	0.162884	-0.011497	0.171832	0.024018	0.031329
From	-0.100111	1.000000	-0.152628	0.067805	0.011776	0.071871	0.105231	-0.027017	-0.050326	-0.068390
To	-0.012070	-0.152628	1.000000	-0.070690	0.068881	0.076030	0.012385	0.029914	-0.055921	0.065691
Arival_time	0.040515	0.067805	-0.070690	1.000000	-0.069032	0.185507	-0.021161	-0.091491	0.016038	-0.026670
Stops	0.005466	0.011776	0.068881	-0.069032	1.000000	-0.487563	0.002402	-0.036259	-0.675564	-0.090701
Price	0.162884	0.071871	0.076030	0.185507	-0.487563	1.000000	0.010252	0.086820	0.524558	0.052460
Depart hour	-0.011497	0.105231	0.012385	-0.021161	0.002402	0.010252	1.000000	0.073551	0.071142	0.021301
Depart min	0.171832	-0.027017	0.029914	-0.091491	-0.036259	0.086820	0.073551	1.000000	0.113655	-0.038679
Duration_hours	0.024018	-0.050326	-0.055921	0.016038	-0.675564	0.524558	0.071142	0.113655	1.000000	-0.005009
Duration_mins	0.031329	-0.068390	0.065691	-0.026670	-0.090701	0.052460	0.021301	-0.038679	-0.005009	1.000000





- Hardware and Software Requirements and Tools Used

```
import selenium
from selenium import webdriver
import pandas as pd
from selenium.webdriver.common.by import By
import warnings
warnings.filterwarnings('ignore')
import time

from selenium.common.exceptions import StaleElementReferenceException, NoSuchElementException
```

import necessary library

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

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

```
x=df.drop(columns='Price')#Feature
y=df.Price#Target
```

```
#Lets import standardscaler
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
x_scaled=scaler.fit_transform(x)
x_scaled
```

```
vif
```

	vif	feature
0	1.047792	Airline_name
1	1.060827	From
2	1.039293	To
3	1.027788	Arival_time
4	1.902860	Stops
5	1.029966	Depart hour
6	1.064046	Depart min
7	1.910302	Duration_hours
8	1.031427	Duration_mins

No **multicollinearity problem**

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly inculd the pre-processing of the data and EDA to check the correlation of independent and dependent features also before building the model I made sure that the point data was cleaned and claed before it was fed into machine learing models for this project we need

to predict the price flight meaning our target columns is continue so this is a regression problem I have used various regression algorthim I have sekect random forest regressor as the best suitalble algorithm for our final models as it is giving a good r score and least different in r2 score and cv-score among all the algorithm used other regressor algorithm are also giving accuracy but some are over fitting and some are under fitting the result which may be because of less price in flight performance as well as accuracy and to check my model from overfitting and under fitting I have made use of the k fold cross validing and then hyper tuning the final model once I was able to get my desired final model I ensured to save that models i before i loaded the testing data and stared performance the data as training data set and obtaining the predicted price out of the Regression machine learning model.

- Testing of Identified Approaches (Algorithms)

1)Logistic Regression Model

2) Decision Tree Regression Model

3) Random Forest Tree Regression Model

4) XGBOOST Regression Model

5) Gradient Boosting Regression Model

Here we select Random Forest Tree Regression for the model building.

```
parameter={'criterion':['mse','mae'],
           'max_features':['auto','sqrt','log2'],
           'max_depth':range(10,15),
           'min_samples_split':range(9,10),
           'min_samples_leaf':range(5,6)}
```

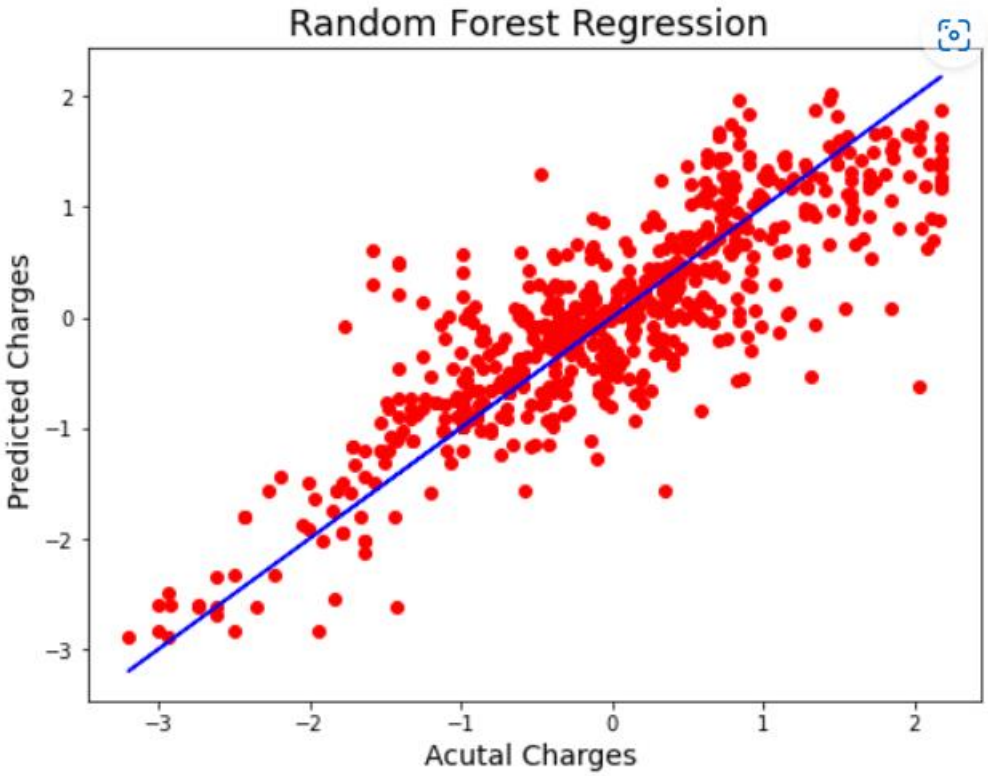
```
rf = RandomForestRegressor()
clf=GridSearchCV(rf,parameter)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

```
{'criterion': 'mse', 'max_depth': 12, 'max_features': 'auto', 'min_samples_leaf': 5, 'min_samples_split': 9}
```

```
rf=RandomForestRegressor(criterion='mse',max_features='log2',max_depth=12,min_samples_leaf=5,min_samples_split=9)
rf.fit(x_train,y_train)
rf.score(x_train,y_train)
pred_decision=rf.predict(x_test)
rfs=r2_score(y_test,pred_decision)
print("R2 Score",rfs*100)
rfs_score=cross_val_score(rf,x_scaled,y,cv=13)
rfc=rfs_score.mean()
print('Cross Val Score',rfc*100)
```

```
R2 Score 71.56197773363478
Cross Val Score 41.55605111502132
```

- Visualizations



CONCLUSION

- Key Findings and Conclusions of the Study

Post models building and choosing the appropriate model I want ahead and scrape the data and join the dataset. After applying all the data pre processing steps as the dataset I was then able to get the predicted price result. Once the dataset with feature columns are predicted label was format I exported the value in a comma separated value file to be accessed as needed.

Conclusion

```
loaded_model=pickle.load(open('Flight','rb'))
result=loaded_model.score(x_test,y_test)
print(result*100)
```

```
71.56197773363478
```