Flight : PRICE PREDICTION

Submitted by:

SHREYA JAIN

ACKNOWLEDGMENT

I would like to express my very great appreciation to my SME Mr. Keshav Bansal for his valuable and constructive suggestions during the planning and developmentof this research work. His willingness to give his time so generously has been very much appreciated. Separately, I would like to thank:

- FlipRobo Technologies team
- Data Trained Team

INTRODUCTION

- Flights are one of the necessary need of each and every person around the globe and therefore flight tickets market is the market which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain.

- Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in flight tickets sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for flight tickets selling companies. Our problem is related to one such flight-tickets selling client.

- We are required to model the price of tickets with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

- With the covid 19 impact in the market, we have seen lot of changes in the flight tickets market. Now some flights are in demand hence making them costly and some are not in demand hence cheaper. One of the clients works with small traders, who sell flight tickets. With the change in market due to covid 19 impact, client is facing problems with their previous flight tickets price valuation machine learning models. So, they are looking for new machine learning models from new data.

- For this client wants to know:

  - 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?

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them.

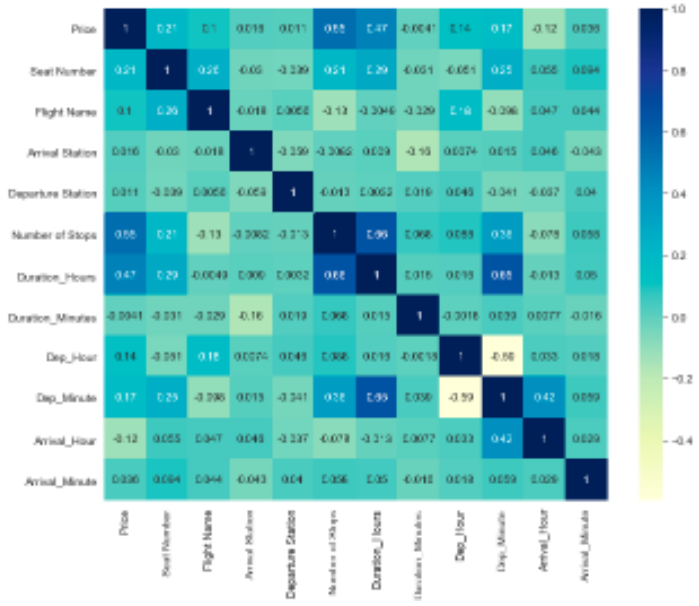
```
In [82]: df.describe()
```

Out[82]:

	Price	Seat Number	Flight Name	Arrival Station	Departure Station	Number of Stops	Duration_Hours	Duration_Minutes	Dep_Hour	Dep_Minute	Arrival_Ho
count	1573.000000	1573.000000	1573.000000	1573.000000	1573.000000	1573.000000	1573.000000	1573.000000	1573.000000	1573.000000	1573.0000
mean	10754.798567	1.956771	2.315321	3.910998	3.839542	0.859504	8.530197	28.490782	15.015257	445.162111	13.0298
std	4529.753777	1.810884	1.680149	2.450209	1.800237	0.819833	7.315057	17.017225	6.781938	852.828880	4.8916
min	4460.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
25%	7391.000000	0.000000	1.000000	2.000000	3.000000	0.000000	2.000000	10.000000	10.000000	20.000000	9.0000
50%	9902.000000	1.000000	3.000000	4.000000	3.000000	1.000000	6.000000	25.000000	17.000000	40.000000	13.0000
75%	13297.000000	3.000000	3.000000	5.000000	5.000000	1.000000	11.000000	40.000000	21.000000	1445.000000	17.0000
max	34408.000000	5.000000	5.000000	8.000000	8.000000	3.000000	38.000000	55.000000	23.000000	1495.000000	23.0000

```
In [83]: # Multivariate Analysis
plt.figure(figsize=(10,8))
sns.heatmap(dfcorr,cmap='YlGnBu',annot=True)
```

```
Out[83]: <matplotlib.axes._subplots.AxesSubplot at 0x202541fb688>
```



Removing the Outliers using Z-score

Removing Outliers

In [86]:
```
from scipy.stats import zscore
z=np.abs(zscore(df))
```

In [87]:
```
z
```

Out[87]:
```
array([[1.06039397, 1.12866389, 1.00301493, ..., 0.60568558, 0.61959186,
        0.80854982],
       [1.06039397, 1.12866389, 1.00301493, ..., 0.6210131 , 1.42534729,
        0.42565578],
       [1.06039397, 1.12866389, 1.00301493, ..., 0.6210131 , 1.01635946,
        0.80854982],
       ...,
       [2.29273061, 0.5285125 , 0.78310864, ..., 0.68232318, 0.21060403,
        1.65986138],
       [2.97090801, 0.57627176, 1.37848316, ..., 0.59802182, 1.64206144,
        1.11710122],
       [3.18091996, 0.57627176, 1.37848316, ..., 0.59802182, 1.64206144,
        1.11710122]])
```

In [88]:
```
threshold=3
print(np.where(z>3))
```

```
(array([ 258,  384,  396,  496,  521,  522,  596,  599,  604,  765,  769,
        774,  794,  965,  966,  967,  969, 1077, 1079, 1080, 1093, 1094,
       1160, 1161, 1218, 1220, 1305, 1306, 1334, 1335, 1336, 1337, 1368,
       1369, 1560, 1572], dtype=int64), array([0, 6, 0, 6, 0, 0, 6, 6, 6, 6, 6, 6, 0, 6, 6, 6, 6, 6, 6, 6, 0, 0,
       0, 0, 6, 6, 0, 0, 0, 0, 0, 0, 5, 5, 6, 0], dtype=int64))
```

In [89]:
```
df_new=df[(z<3).all(axis=1)]
```

In [90]:
```
df_new.shape
```

Out[90]:
```
(1537, 12)
```

In [91]:
```
df.shape
```

Out[91]:
```
(1573, 12)
```

In [92]:
```
((1573-1537)/1573)*100
```

Out[92]:
```
2.288620470438652 4
```

In [93]:
```
df=df_new
```

- ## Data Sources and their formats

The sample data is extracted by using web scraping from selenium. It is stored in csv format and hence we import it using pandas. Then we further checked more about data using info, checked data types using dtypes, shapes using .shape, columns using .columns, null values using .isnull.sum, and further visualize it through heatmap as follows:

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import scipy
        import sklearn
        import warnings
        warnings.filterwarnings('ignore')
        from sklearn.linear_model import LinearRegression
        from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error
        from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
        from sklearn.naive_bayes import GaussianNB
        from sklearn.svm import SVR
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.neighbors import KNeighborsRegressor
        from sklearn.model_selection import GridSearchCV
        from sklearn.model_selection import cross_val_score
        from sklearn.model_selection import train_test_split
        import warnings
        warnings.filterwarnings('ignore')
```

```
In [2]: df=pd.read_csv('flights_data.csv')
        df.head()
```

Out[2]:

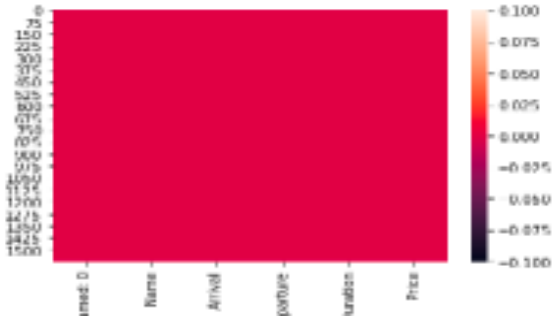
	Unnamed: 0	Name	Arrival	Departure	Duration	Price
0	0	SpiceJet\nSG 8184	10:40\nDelhi	12:50\nMumbai	2h 10m\nNon Stop	5953
1	1	SpiceJet\nSG 507	20:20\nDelhi	22:40\nMumbai	2h 20m\nNon Stop	5953
2	2	SpiceJet\nSG 611	18:40\nDelhi	20:40\nMumbai	2h 00m\nNon Stop	5953
3	3	SpiceJet\nSG 757	16:20\nDelhi	18:30\nMumbai	2h 10m\nNon Stop	5953
4	4	Indigo\n6E 2339	21:55\nDelhi	00:05+1\nMumbai	2h 10m\nNon Stop	5953

```
In [3]: df.shape
```

```
Out[3]: (1573, 6)
```

```
In [4]: sns.heatmap(df.isnull())
```

```
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x1a16b6f8308>
```



- <u>Data Preprocessing Done</u>

First we will determine whether there are any null values and since there were no null values as well as NaN values present in the dataset we proceeded further by Label encoding using label encoder. Then we performed some data visualization in which we observed certain attributes were having skewness and outliers that were plotted using distplot and boxplot. Outliers were removed with the help of Zscore in which 36 rows were removed.

<u>Data Inputs- Logic- Output Relationships</u>

The data consists of 11 inputs and one output-“Price”. Arrival Hours is the least/negatively correlated column with target('Price') variable. Number of Stops is highly correlated column with target variable followed by other attributes.

Hardware and Software Requirements and Tools Used

In this project we have used HP Pavilion PC with 64-bit operating system and have Windows 10 pro. We have used python to develop this project in which we have used various libraries such as numpy, pandas, matplotlib, seaborn for handling data or arrays and their visualization. For statistical purpose we have used zscore from scipy.stats to remove outliers. Lastly, to develop the model we have used various libraries and metrics from sklearn such as train_test_split, Linear Regression, Lasso, Ridge, Elastic Net, SVR, Decision Tree Regressor, KNeighbors Regressor, Random Forest Regressor,AdaBoostRegressor,mean_squared_error, mean_absolute_error and r2_score.

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

We have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them. We have used distplot to find the distribution of all attributes.

Testing of Identified Approaches (Algorithms)

We have used following algorithms such as: LinearRegression, Lasso, Ridge, ElasticNet,SVR,DecisionTreeRegressor,KNeighborsRegressor,Random ForestRe gressor and AdaBoostRegressor.

Run and Evaluate selected models

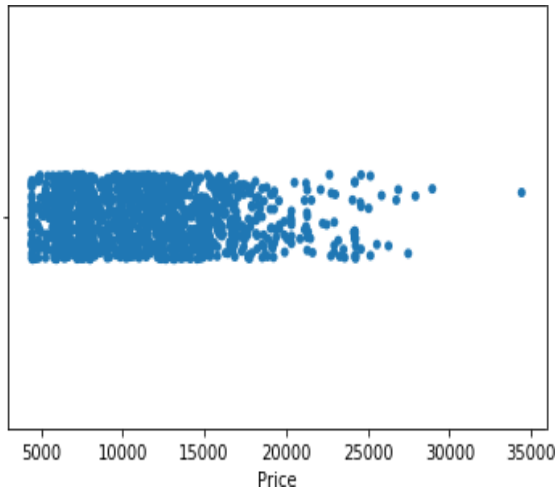
We have formed a loop where all the algorithms will be used one by one and their corresponding Score, Mean Absolute Error, Mean Squared Error, RMSE and r2_score will be evaluated.

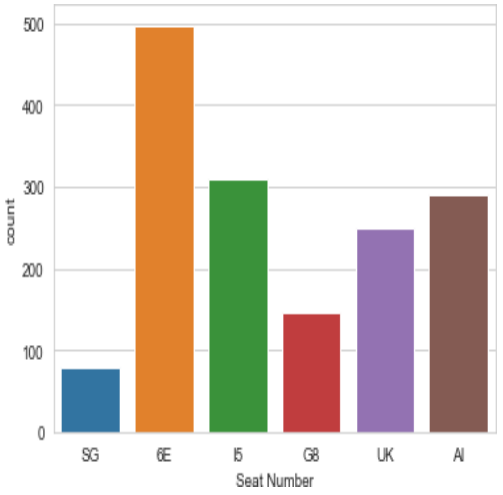
• I chose Decision Tree Regressor as our best model since it's giving us best sco re and it's performing well. It's r2_score is also satisfactory and it shows that o ur model is neither underfitting/overfitting. Then there is no need to perform hyperparamter as it is giving 100% accuracy. Its r2_score is also satisfactory. H ence we saved Decision Tree Regressor as our final model using joblib.

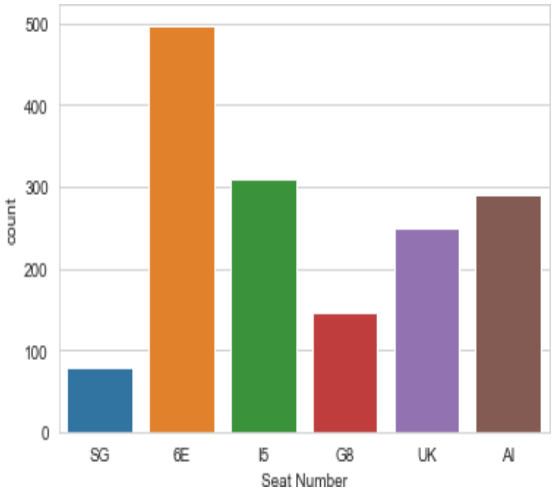
Key Metrics for success in solving problem under consideration

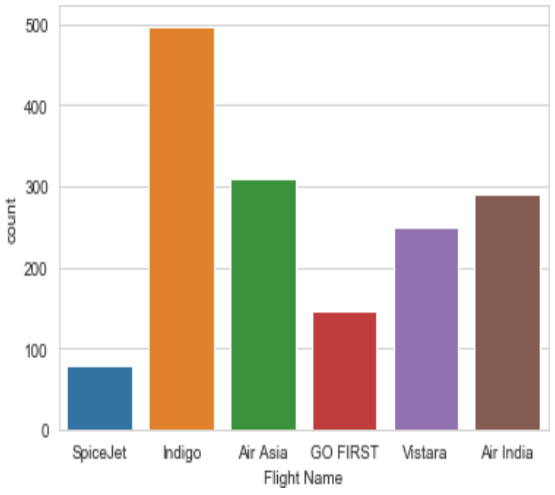
Key metrics used for finalising the model was Score and r2_score. Since in case of Decision Tree Regressor it's giving us good score among all other models and it's performing well. It's r2_score is also satisfactory and it shows that our model is neither underfitting/overfitting .

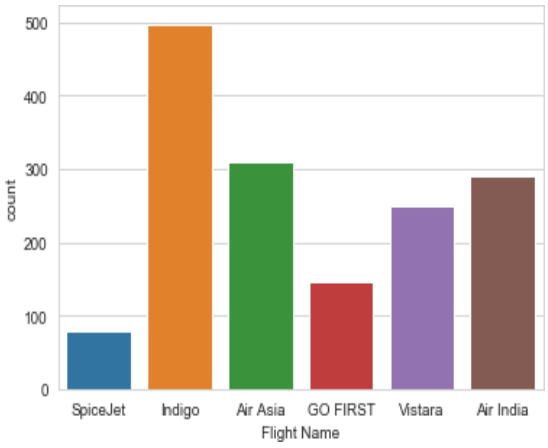
Data Visualization

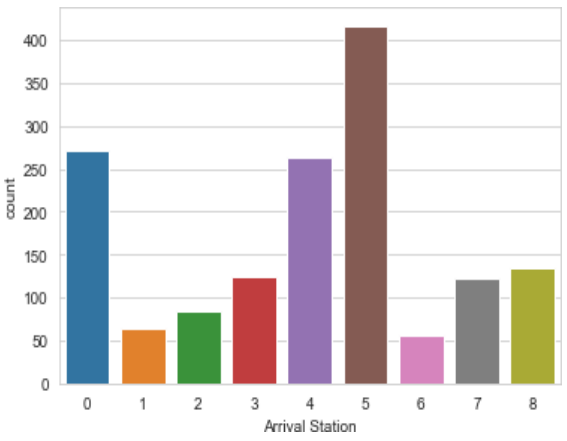


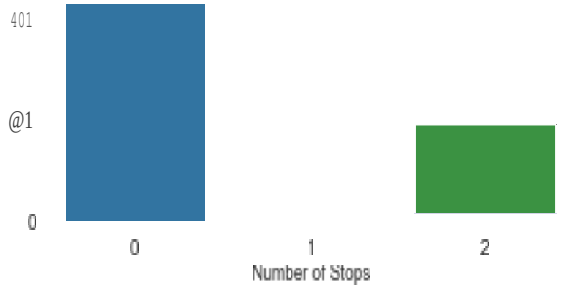


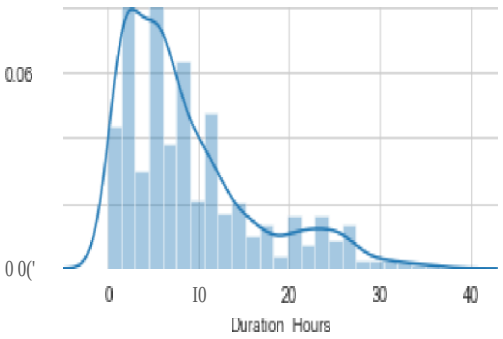


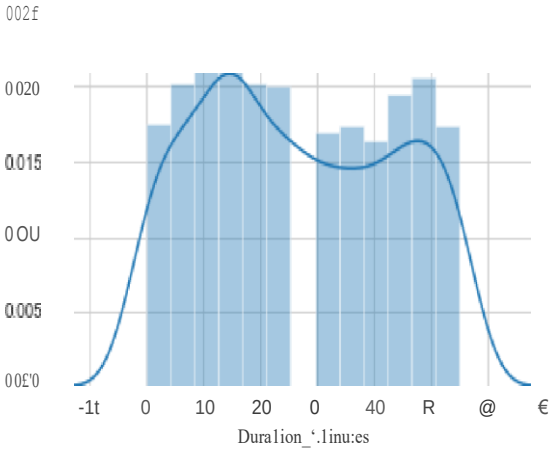


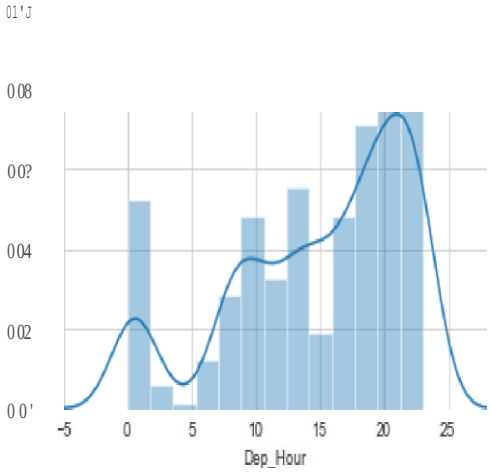


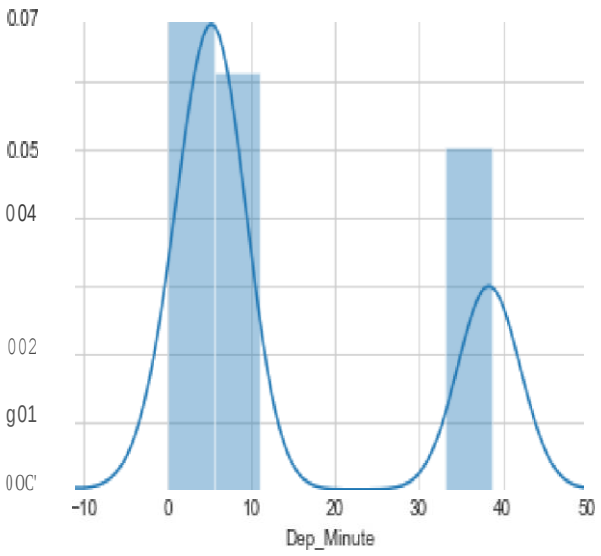


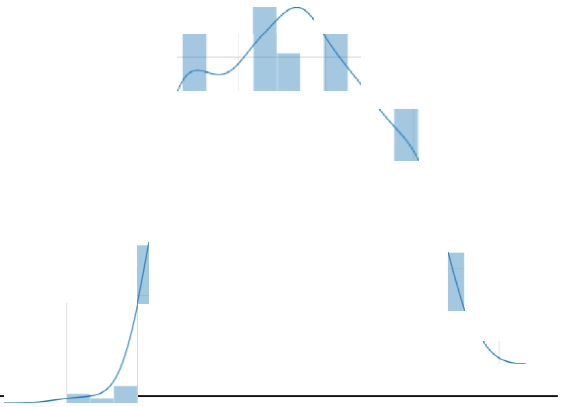

401
@1
0
0
1
2
Number of Stops

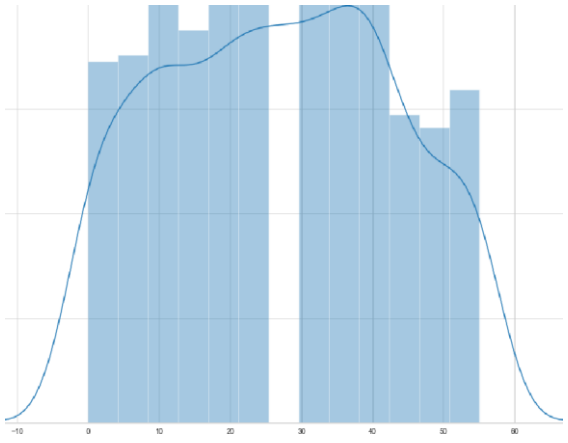

0.06
0 0('
0
I0
20
30
40
Duration_Hours

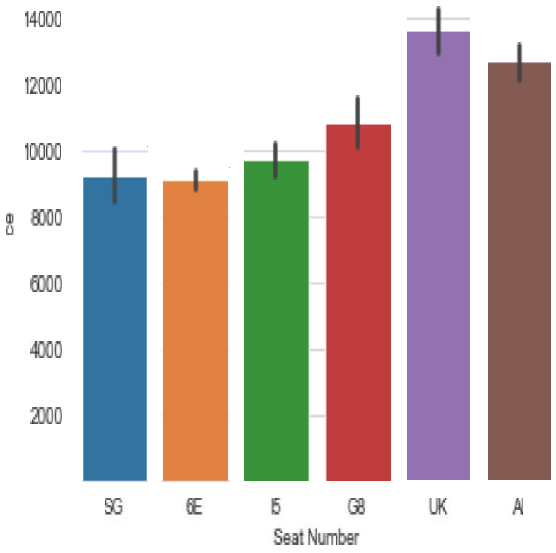

002£
0020
0.015
0 OU
0.005
00£'0
-1t
0
10
20
0
40
R
@
€
Dura1ion_'.1inu:es

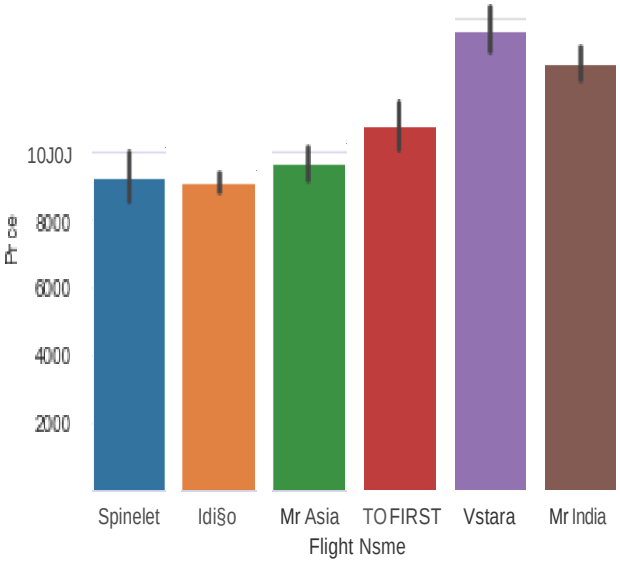

01'J
0 08
00?
004
002
0 0'
-5
0
5
10
15
20
25
Dep_Hour

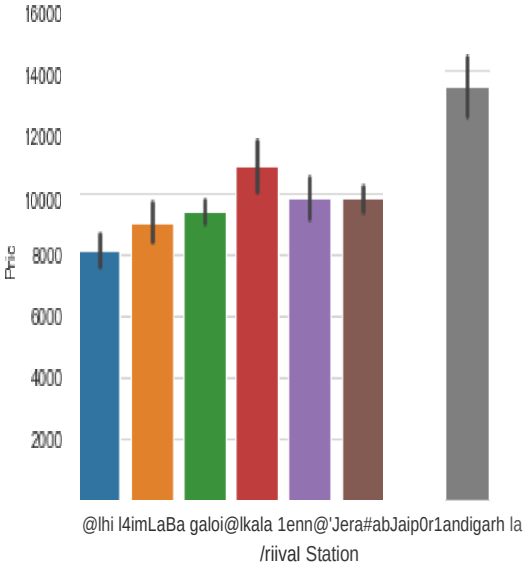

0.07
0.05
004
002
g01
0 OC'
-10
0
10
20
30
40
50
Dep_Minute

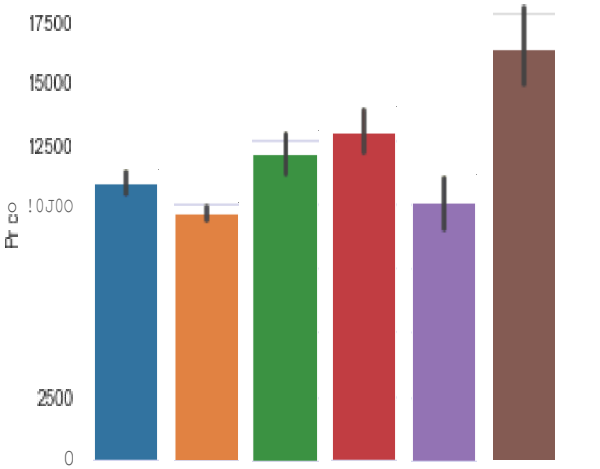


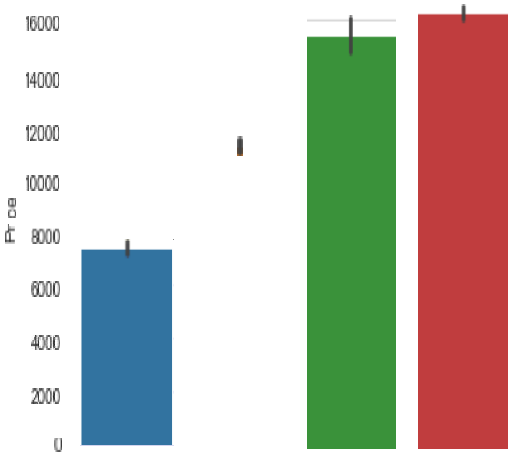


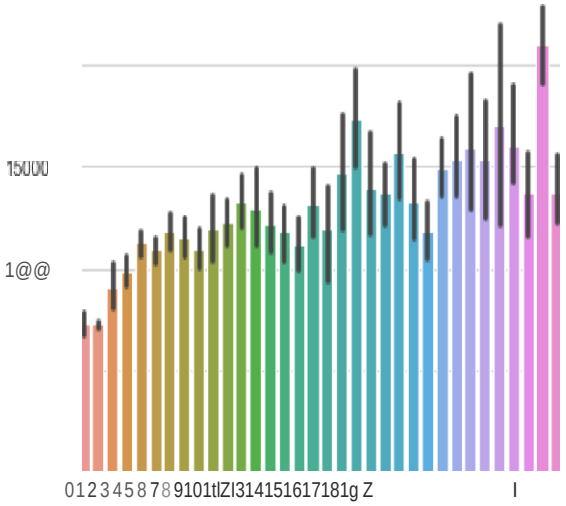

14000
12000
10000
8000
6000
4000
2000
SG
6E
I5
G8
UK
AI
Seat Number

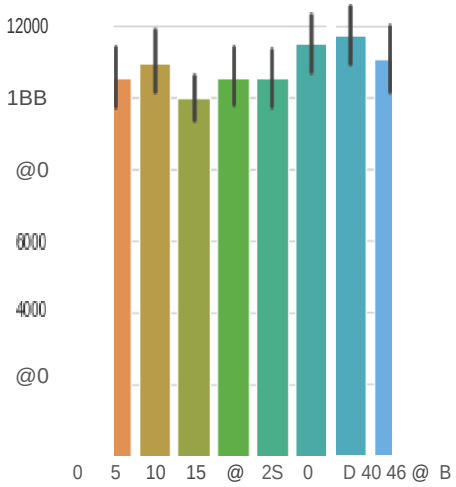

10J0J
8000
6000
4000
2000
Spinelet
Idi§o
Mr Asia
TO FIRST
Vstara
Mr India
Flight Nsme

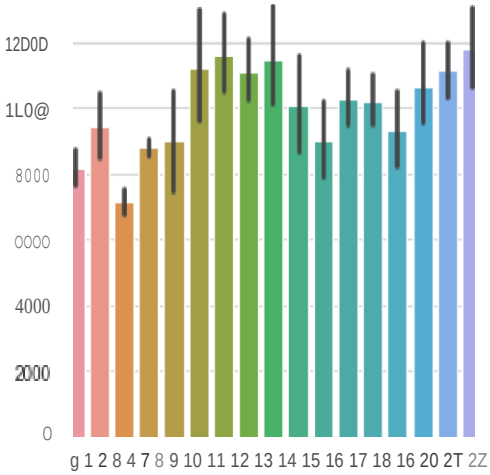

16000
14000
12000
10000
8000
6000
4000
2000
@lhi l4imLaBa galoi@lkala 1enn@'Jera#abJaip0r1andigarh la
/riival Station

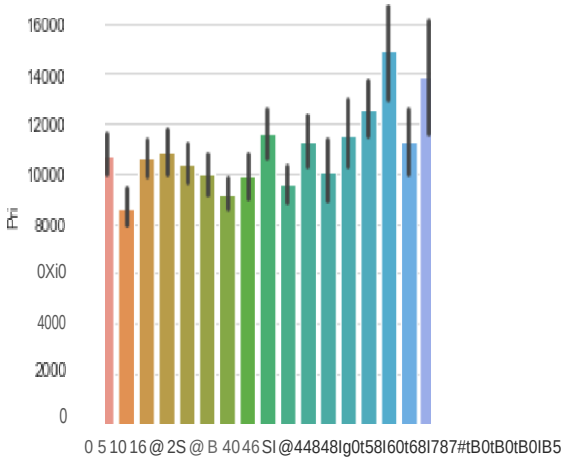

17500
15000
12500
!0J00
2500
0
lñJmLai
@lhi
Bangalore
Kolkata
Chennai
landigarh
Jaipur
DBpartur+ Station

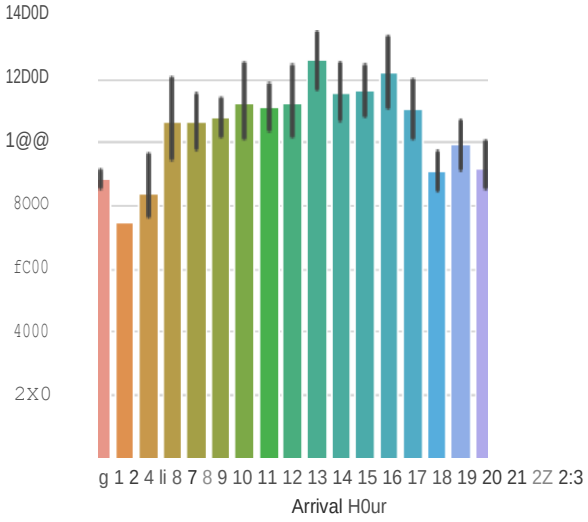

16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
rmJef of Slops

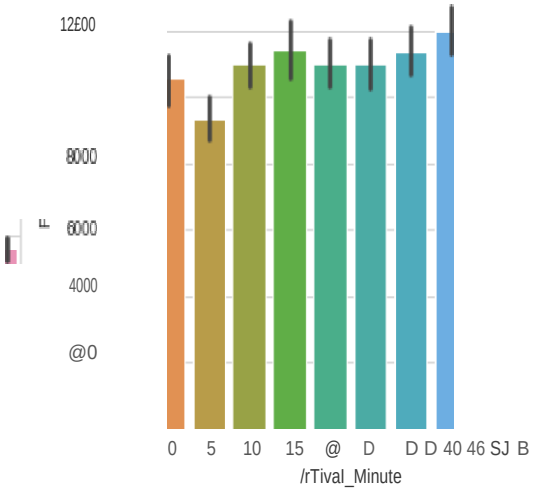

15000
1@@
0 1 2 3 4 5 8 7 8 9 10 1 t l Z I 3 14 15 16 17 18 1 g Z
I


12000
1BB
@0
6000
4000
@0
0 5 10 15 @ 2S 0 D 40 46 @ B


12D0D
1L0@
8000
0000
4000
2000
0
g 1 2 8 4 7 8 9 10 11 12 13 14 15 16 17 18 16 20 2T 2Z D


16000
14000
12000
10000
Phi
8000
0Xi0
4000
2000
0
0 5 10 16 @ 2S @ B 40 46 SI @44848Ig0t58I60t68I787#tB0tB0tB0IB5


14D0D
12D0D
1@@
8000
fC00
4000
2XO
g 1 2 4 li 8 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 2Z 2:3
Arrival H0ur


12E00
8000
F
6000
4000
@0
0 5 10 15 @ D D D 40 46 SJ B
/rTival_Minute

➢ Interpretation of the Results

- Flight Prices are majorly in the range of 5000-20000.
- 6E seat number is present majorly in flights.
- SG seat number is present in the least number in flights.
- I5 and AI seat numbers are also present in high number in flights.
- Indigo Airline is present majorly as flight name.
- Spice Jet Airline is present in the least number as flight name.
- Air Asia and Air India are also present in high number as flights name.
- Hyderabad is majorly present as arrival station for flights.
- Jaipur is minorly present as arrival station for flights.
- Goa and Banglore are also present in high number as arrival station for flights.
- One stop is present majorly for flights.
- Three stop are present in the least number for flights.
- Zero Stop is also present in high number for flights.
- Duration hours are majorly in the range of 0-10 for flights.
- Duration hours are slightly in the range of 30-40 for flights.
- 2 hours duration is present majorly for flights.
- 38, 37 and 0 hours duration are present in the least number for flights.
- 5, 6, 1, 7, 8, 4, 9, 11 and 10 hours duration are also present in high number for flights.
- 20, 26, 22, 24, 27, 18, 19, 28, 31, 30, 34, 35, 29, 32 and 33 hours duration are present in low number for flights.
- Duration minutes are majorly in the range of 0-30 minutes for flights.
- Duration minutes are slightly in the range of 30-40 minutes for flights.
- Duration minutes are also in high number in the range of 30-40 minutes for flights.
- Departure hours are majorly in the range of 18-23 hours for flights.
- Departure hours are slightly in the range of 3-5 hours for flights.
- Duration minutes are also in high number in the range of 30-40 minutes for flights.
- Departure minutes are majorly in the range of 18-23 hours for flights.
- Departure hours are slightly in the range of 3-5 hours for flights.
- Duration minutes are also in high number in the range of 30-40

minutes for flights.

- Arrival hours are majorly in the range of 12-16 hours for flights.
- Arrival hours are slightly in the range of 0-4 hours for flights.
- Arrival hours are also in high number in the range of 5-9 hours for flights.
- Arrival minutes are majorly in the range of 20-40 minutes for flights.
- Arrival minutes are slightly in the range of 49-53 minutes for flights.
- Arrival minutes are also in high number in the range of 0-10 minutes for flights.
- Uk seat number have the highest prices for flights.
- 6E seat number have the lowest prices for flights.
- AI seat number have also the high prices for flights.
- Vistara airlines have the the highest prices for flights.
- Indigo airlines have the lowest prices for flights.
- Air India airlines have also high prices for flights.
- Spice jet airlines have also low prices for flights.
- Arrival station of Goa have the highest prices for flights.
- Arrival station of Delhi have the lowest price for flights.
- Arrival station of Chandigarh and Jaipur have also high prices for flights.
- Arrival station of Banglore and Mumbai have also low prices for flights.
- Departure station of Chandigarh have the highest prices for flights.
- Departure station of Delhi have the lowest prices for flights.
- Departure station of Kolkata and Jaipur have also high prices for flights.
- Departure station of Chennai have also low prices for flights.
- Flights with three number of stops have the highest prices for flights.
- Flights with zero number of stops have the lowest prices for flights.
- Flights with 33 hours of duration have the highest prices for flights.
- Flights with 0 hours of duration have the lowest prices for flights.
- Flights with 35, 20, 30, 31, 27, 28, 29 and 23 hours of duration have also high prices for flights.
- Flights with 1, 2, 3 and 4 hours of duration have also low prices for flights.
- Flights with 35 minutes of duration have the highest prices for flights.
- Flights with 15 minutes of duration have the lowest prices for flights.
- Flights with 0 and 20 minutes of duration have also high

prices for flights.

- Flights with 45 and 50 minutes of duration have also low prices for

flights.

- Flights with 23rd hour as departure have the highest prices for flights.
- Flights with 2nd hour as departure have the lowest prices for flights.
- Flights with 18th, 19th, 8th and 10th hour as departure have also high prices for flights.
- Flights with 11th hour as arrival have the highest prices for flights.
- Flights with 23rd hour as arrival have the lowest prices for flights.
- Flights with 14th and 20th hour as arrival have also high prices for flights.
- Flights with 0th and 2nd hour as arrival have also low prices for flights.

Learning Outcomes of the Study in respect of Data Science

With the help of visualization tools such as matplotlib and seaborn we have visualized the impact of each attributes on our target variable. For cleaning the data and plotting outliers we have used distplot and boxplot and for removing outliers we have used zscore which is a statistical tool. At last we got Decision Tree Regressor as our best model.

Limitations of this work and Scope for Future Work

The model is working well and we have performed hyperparameter tuning and we have concluded our project by choosing Decision tree Regressor as our best model.