FLIGHT PRICE PREDICTION

Submitted by: Kundan Patil

ACKNOWLEDGEMENT

In summary, we have completed a flight price prediction project by collecting data from various websites, specifically using yathra.com which provided data on 6 major aviation companies in India. The data collection process was challenging due to the various factors that affect ticket prices such as class, distance, and number of travels. We utilized web scraping techniques using selenium and then used tools such as Tableau, Microsoft Power BI, and Jupyter for data analysis and model building. We also received valuable support and guidance from various sources including websites such as towardsdatascience, geeksforgeeks, and stackoverflow, as well as from the companies Flirobo and Datatrained.

INTRODUCTION

- ## Business Problem Framing

The problem that we are trying to solve is the unpredictability of flight ticket prices. The fluctuation of prices can make it difficult for consumers to plan and budget for their travel, and also make it challenging for airlines to accurately forecast their revenue.

Our goal is to create a model that can predict flight ticket prices based on various factors such as time of purchase, flight itinerary, class of service, and other related information. By utilizing data collected from multiple sources, our model aims to provide a more accurate prediction of flight ticket prices, which can benefit both consumers and airlines.

The project will involve collecting data on flight ticket prices and other relevant features, cleaning and preprocessing the data, building and training a predictive model, and finally evaluating the performance of the model. Additionally, we will also be using visualization tools like tableau and Microsoft power bi for exploratory data analysis.

Overall, this project aims to provide a more accurate and transparent view of flight ticket prices, making it easier for consumers to plan and budget for their travel, and also help airlines make more informed decisions on pricing strategy.

- ## Conceptual Background of Domain Problem

To begin the project, we need to first focus on gathering the data that is required for our model. To do this, we will use web scraping techniques to extract flight ticket prices and other relevant information from various websites.

In order to extract the data, we will need to find the URLs of the flight ticket pages and then use relevant XPaths to extract the required data. We will need to be familiar with web scraping tools such as Selenium, Beautiful Soup, or Scrapy to be able to automate this process.

Once we have collected the data, we will need to preprocess it and clean it up to ensure that it is in a suitable format for our model. This step will involve removing any missing values or outliers, converting data types and variables as necessary, and normalizing the data if required.

To build the predictive model, we will be using a linear regression model, as it is a widely used technique for predicting numerical values. Linear regression is a statistical method that tries to fit a linear equation to the data points. To be able to build the linear regression model, we will have to be familiar with tools such as Python libraries like numpy, pandas and sklearn, which contains libraries related to machine learning.

Lastly, we will be using Jupyter notebook or python IDE to develop our predictive model. We will have to evaluate the performance of the model using appropriate evaluation metrics such as Mean Squared Error or Root Mean Squared Error, to check the accuracy of the model and evaluate its performance.

- ## Literature Review

In order to gather enough relevant data for our project, we will need to crawl through the internet and find suitable websites that have the necessary information. One website that we found to be particularly useful was yathra.com, which provided us with a large amount of data on flight ticket prices from various airlines.

To ensure we were able to find the best possible resources for the project, we also used various other websites such as stackoverflow, towardsdatascience, and geeksforgeeks. These websites provided valuable information and tutorials on different aspects of web scraping, data preprocessing, and model building, which helped us to overcome any obstacles and implement our model successfully.

By using various sources to gather the data, we were able to ensure the quality and accuracy of the data we used for our project, which allowed us to make a more robust model.

- ## Motivation for the Problem Undertaken

  The motivation for this project is twofold. Firstly, our primary goal is to provide our client with a new model that can predict flight ticket prices with greater accuracy. By providing them with a more transparent view of flight ticket prices, we aim to help them make more informed decisions on pricing strategy and revenue forecasting.

  Secondly, this project also serves as an opportunity for me to upgrade my skills and knowledge by solving a real-world problem in a virtual environment. The project will provide me with the opportunity to learn and apply new techniques such as web scraping, data preprocessing, and model building, which are all important skills in today's job market.

  Overall, this project represents a valuable opportunity to make a meaningful impact for our client while also allowing me to enhance my own skills and knowledge, making it a win-win situation.

Analytical Problem Framing

- **Mathematical and Analytical understanding of the problem**

  The data we collected was in in terms of rows and columns, there are 10 columns and 1587 rows out of which 3 columns where of object type and remaining are integer type, so in order to make the machine understands this we have to convert the categorical value to numerical values we accomplish it through label encoding or one hot encoding.

  We did scaling of the dataset in order to have values of variables within certain limits so that machine can perform better, for this purpose we used different types of scaling namely standard scaling, min max scaling and Robust scaler, later we reduce the skewness using various techniques.

Importing the dataset ¶

In [1]:
```
import pandas as pd
df=pd.read_csv('flight_tableau.csv')
df.head()
```

Out[1]:

	Unnamed: 0	Company Name	No of stops	No of days in advanced booked	class	where to where	Route Value	Departure Time	Duration	Price	distance in km
0	0	Air Asia	1	1	Economy	Delhi - Mumbai	3	8.00	6.58	5953	1148
1	1	Air Asia	1	1	Economy	Delhi - Mumbai	3	9.42	6.58	5953	1148
2	2	Air Asia	1	1	Economy	Delhi - Mumbai	3	12.67	7.58	5953	1148
3	3	Air Asia	1	1	Economy	Delhi - Mumbai	3	11.92	8.33	5953	1148
4	4	Air Asia	1	1	Economy	Delhi - Mumbai	3	8.00	8.58	5953	1148

- **Data Source and Their Format**

  The major portion of the data we collected from car trade.com , there are 10 major parameters and rows out which 3 are of categorical type column and remainings are of categorical type columns.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1587 entries, 0 to 1586
Data columns (total 10 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Company Name                  1587 non-null   object
 1   No of stops                   1587 non-null   int64
 2   No of days in advanced booked 1587 non-null   int64
 3   class                         1587 non-null   object
 4   where to where                1587 non-null   object
 5   Route Value                   1587 non-null   int64
 6   Departure Time                1587 non-null   float64
 7   Duration                      1587 non-null   float64
 8   Price                         1587 non-null   int64
 9   distance in km                1587 non-null   int64
dtypes: float64(2), int64(5), object(3)
memory usage: 124.1+ KB
```

- **Data Processing**

  There were enough amount of anomalies in the data, we looked column by column for anomalies and we cleaned each column separately , through the aid of pandas, loops and regular expressions we achieved it.

  1.We imported the collected data to a separate notebook which was meant for data preprocessing .

  2.The ticket price column contains the data which were of strings type and we adopted regular expressions to clean it out.

```
In [1]: import pandas as pd
        import numpy as np
```

```
In [2]: # Creating the dataframes
        dataframes=[]
        for i in range(1,73):
            k1='df'
            k2=str(i)
            k=k1+k2
            dataframes.append(k)
```

```
In [3]: # Creating the datafiles name
        flight=[]
        for i in range(1,73):
            k1='flight_data_'
            k2=str(i)
            k3='.csv'
            k=k1+k2+k3
            flight.append(k)
```

```
In [4]: #importing all the dataframe
        k=0
        for i in dataframes:
            locals()[i]=pd.read_csv(flight[k])
            k=k+1
```

```
In [5]: df_list=[]
        for i in dataframes:
            df_list.append(locals()[i])
```

```
In [6]: df=pd.DataFrame()
        df = df.append(df_list)
```

3.The Name of Company column contained a string out of which we have to extract only name of the just name of the company, it was a challenging one and we accomplished it through regular expressions.

- **Data Input-Logic-Output Relationships**

  Data Input : It was a precise dataset, which has been scaled having low skewness and very minimal outliers.

  Logic: The logic here is linear regression algorithm which predict the response variable, the linear regression algorithm we used in this case are

  1. Linear Regression
  2,Lasso Regression
  3.Ridge Regression
  4.ElasticNet
  5.Ransac Regressor
  6.Support vector Regressor
  7.Random Forest Regressor

  Data output : We got a model predicting the price of second hand car

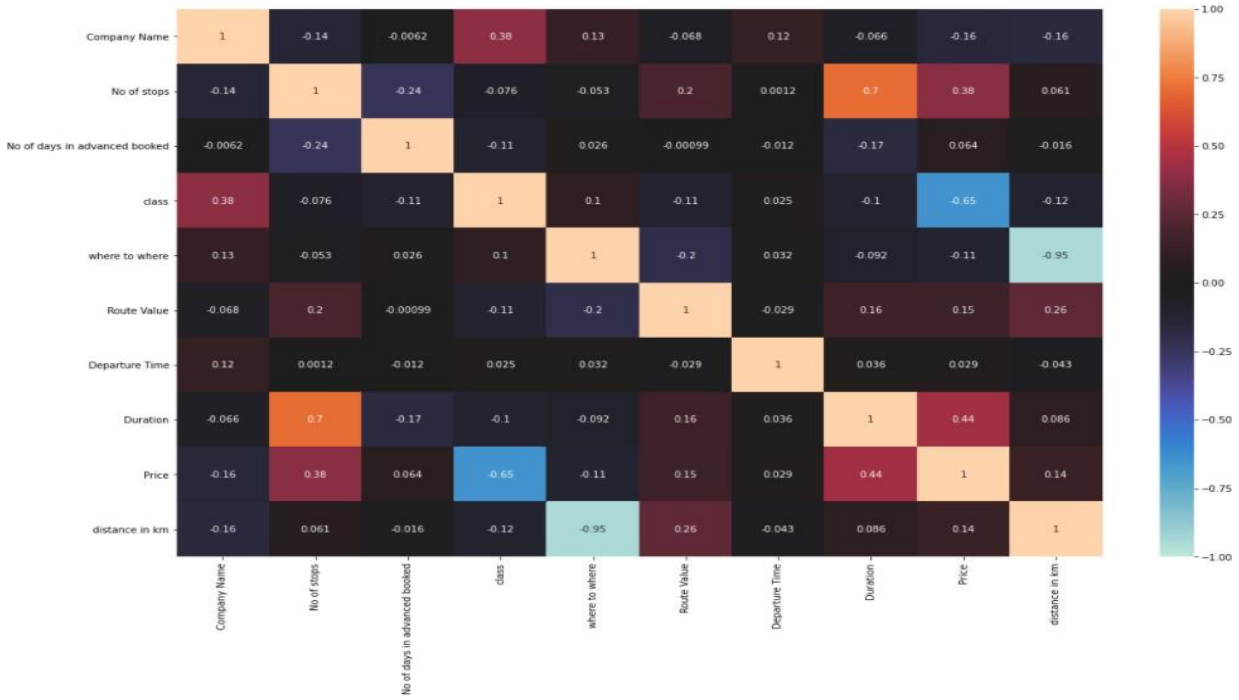
- **Hardware and software tools used**

  Hardware : We used the hardware of 8GB RAM,1Tb ROM and i5 processor.

  Software : For better visualization we used Tableau Public, Jupyter notebook from anaconda navigator for coding and webscrapping and Microsoft word and Power Point Presentation for creating the report and making the presentation respectively.

Exploratory Data Analysis

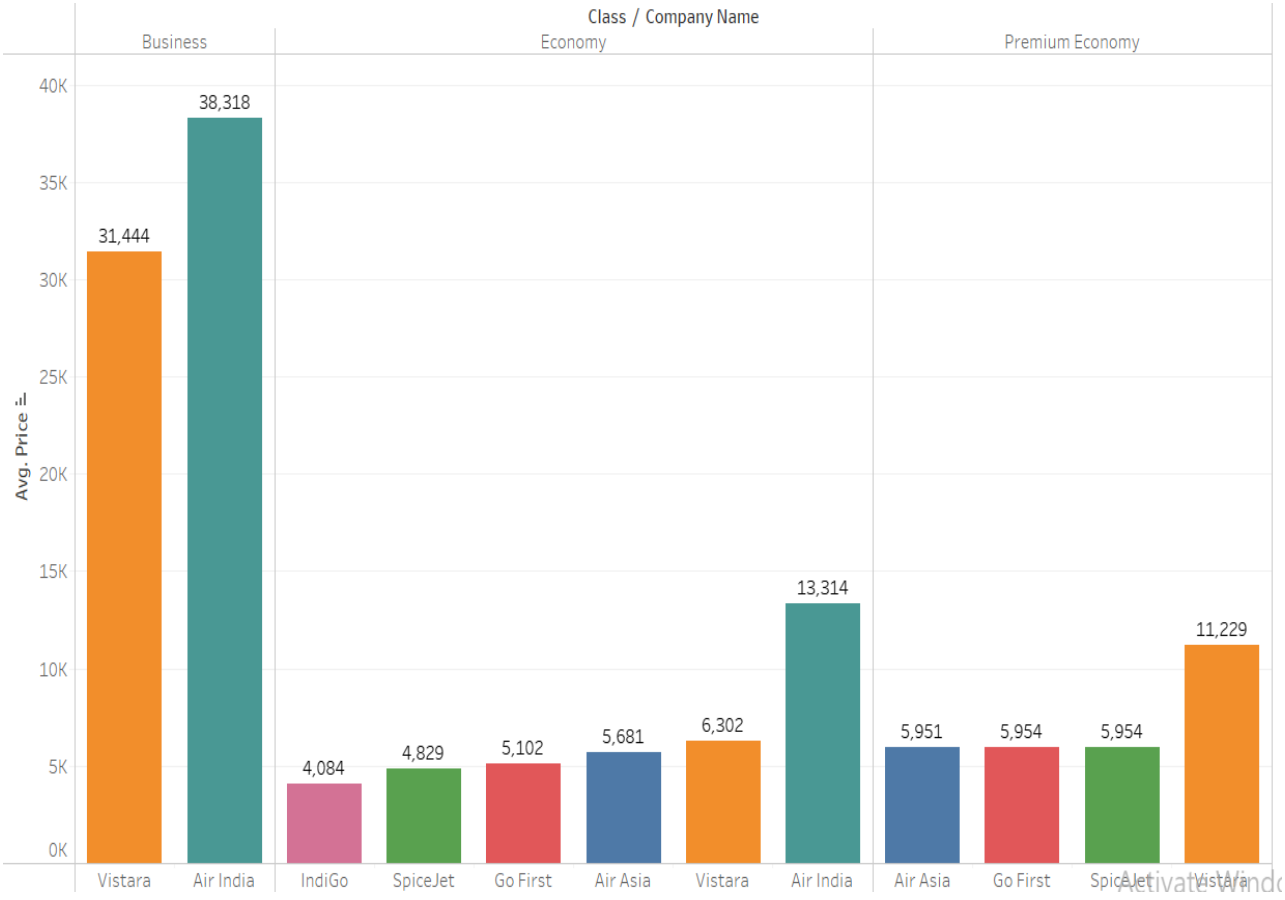
The following are the interpretation we got from data analysis, For analyzing the data we took the help of data visualization libraries like seaborn, Matplotlib, plotly,Tableau software, First we drew the correlation and heatmaps from which we came to know which are the significant variables which decides the flight price.



From the above heatmap we came to know about those independent variable which has got significant effect over the response variable, But here we have a problem in analyzing suppose say we want to know how the advanced booking decides the price but there are factors due to which we can't find directly because we have to give class as hue as well as destination, but providing two hue values and analyzing is bit difficult in python so we did whole analysis part in Tableau.

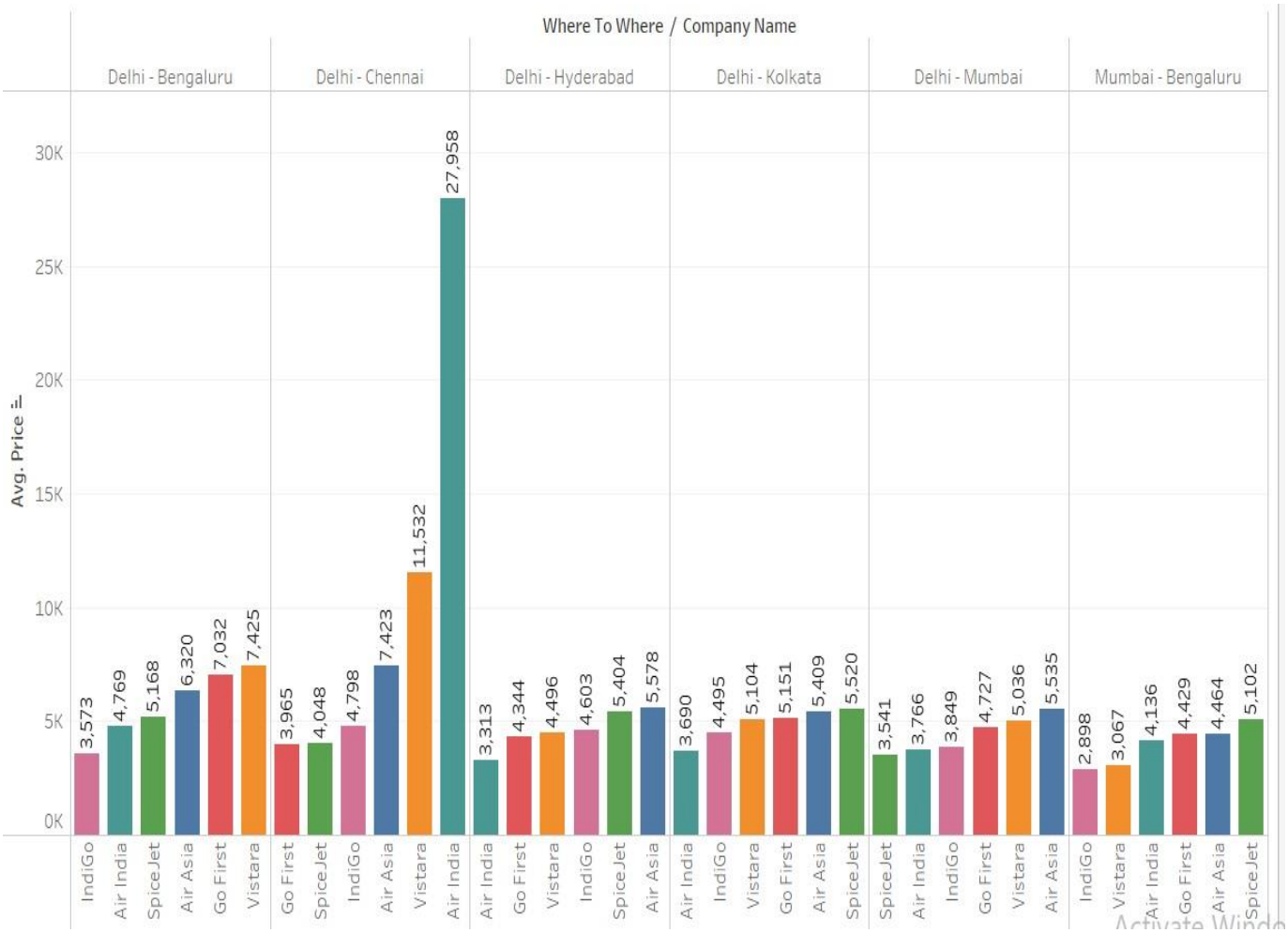
To analyze in tableau first we have prepared set of questions and we tried to find answers for those in tableau.

1. **Average of each class price based on company name**



The above graphs clearly indicates that if you want to travel in Business class choose vistara it has least average price,and for Economy class choose IndiGo, and for Premium Economy go for Air Asia.

2. What's the least price to travel from Delhi and Mumbai to other locations in Economy class

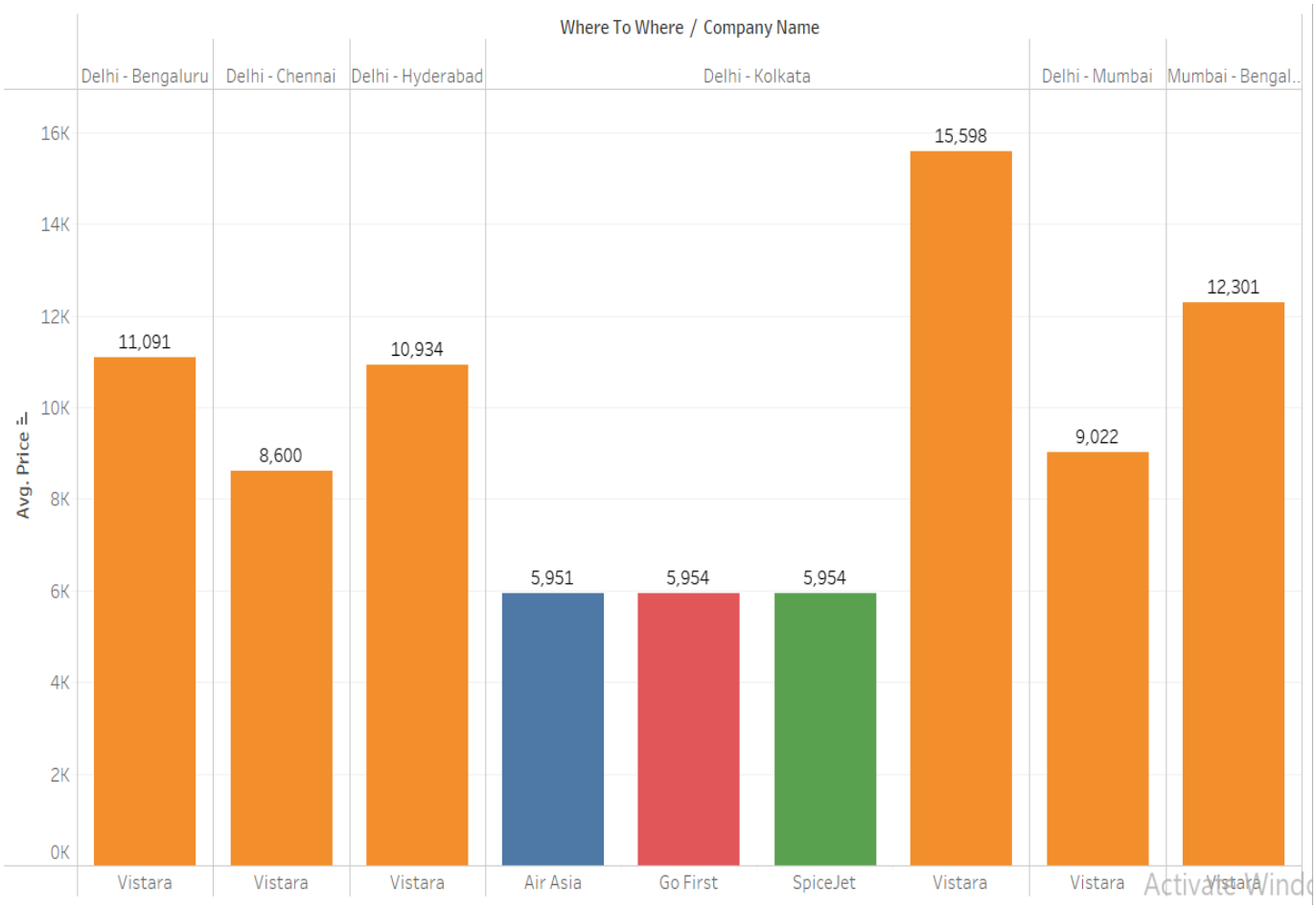


- If you want to travel from Delhi to Bengaluru in Economy class then IndiGo will offer least price followed by Air India and SpaceJet and Vistara is the costly one.
- If you want to travel from Delhi to Chennai in Economy class then Go First will offer least price followed by SpaceJet and IndiGo and Air India is the costly one.
- If you want to travel from Delhi to Hyderabad in Economy class then Air India will offer least price followed by GoFirst and Vistara and Air Asia is the costly one.
- If you want to travel from Delhi to Kolkata in Economy class then Air India will offer least price followed by IndiGO and Vistara and SpaceJet is the costly one.
- If you want to travel from Delhi to Mumbai in Economy class then SpaceJet will offer least price followed by Air India and IndiGo and Air Asia is the costly one.
- If you want to travel from Mumbai to Bengaluru in Economy class then IndiGo will offer least price followed by Vistara and Air India and SpaceJet is the costly one.

3.What's the least price to travel from Delhi and Mumbai to other locations in Premium Economy class

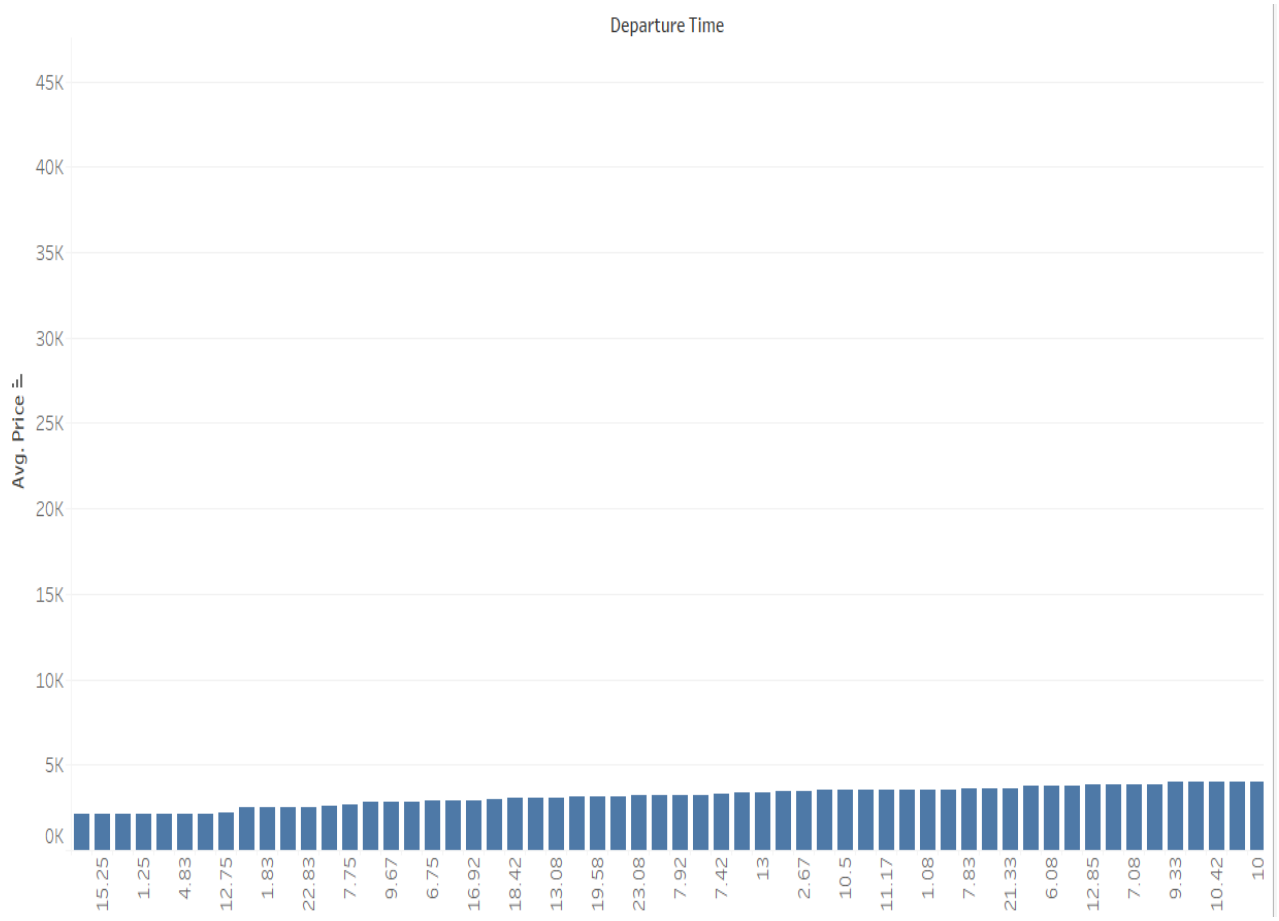
- If you want to travel from Delhi to Bengaluru in PremiumEconomy class then Vistara is the best choice
- If you want to travel from Delhi to Chennai in Premium Economy class then Vistara is the best choice
- If you want to travel from Delhi to Hyderabad in Premium Economy class then Vistara
- If you want to travel from Delhi to Kolkata in Premium Economy class then Air Asia is the cheapest one followed by GoFirst and SpiceJet
- If you want to travel from Delhi to Mumbai and Mumbai to Bengaluru in Premium Economy class then Vistara is the best choice.

4.What's the least price to travel from Delhi and Mumbai to other locations in Business class



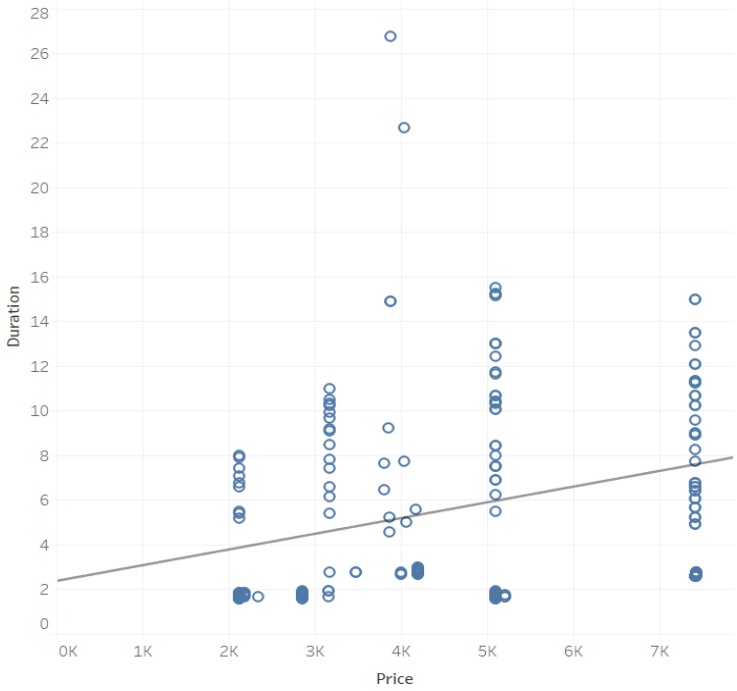
- If you want to go from Delhi to Bengaluru Chennai or Hyderabad or Mumbai or Mumbai to Bengaluru then vistara is the only available as per this data at this moment, It may vary on coming days.
- If you want to go from Delhi to Kolkata in Business class then Air Asia is the cheapest one followed by GoFirst and SpaceJet and Vistara is the costliest one according to this data.

5. What's the general departure time at which Economy class will have low price comparatively?



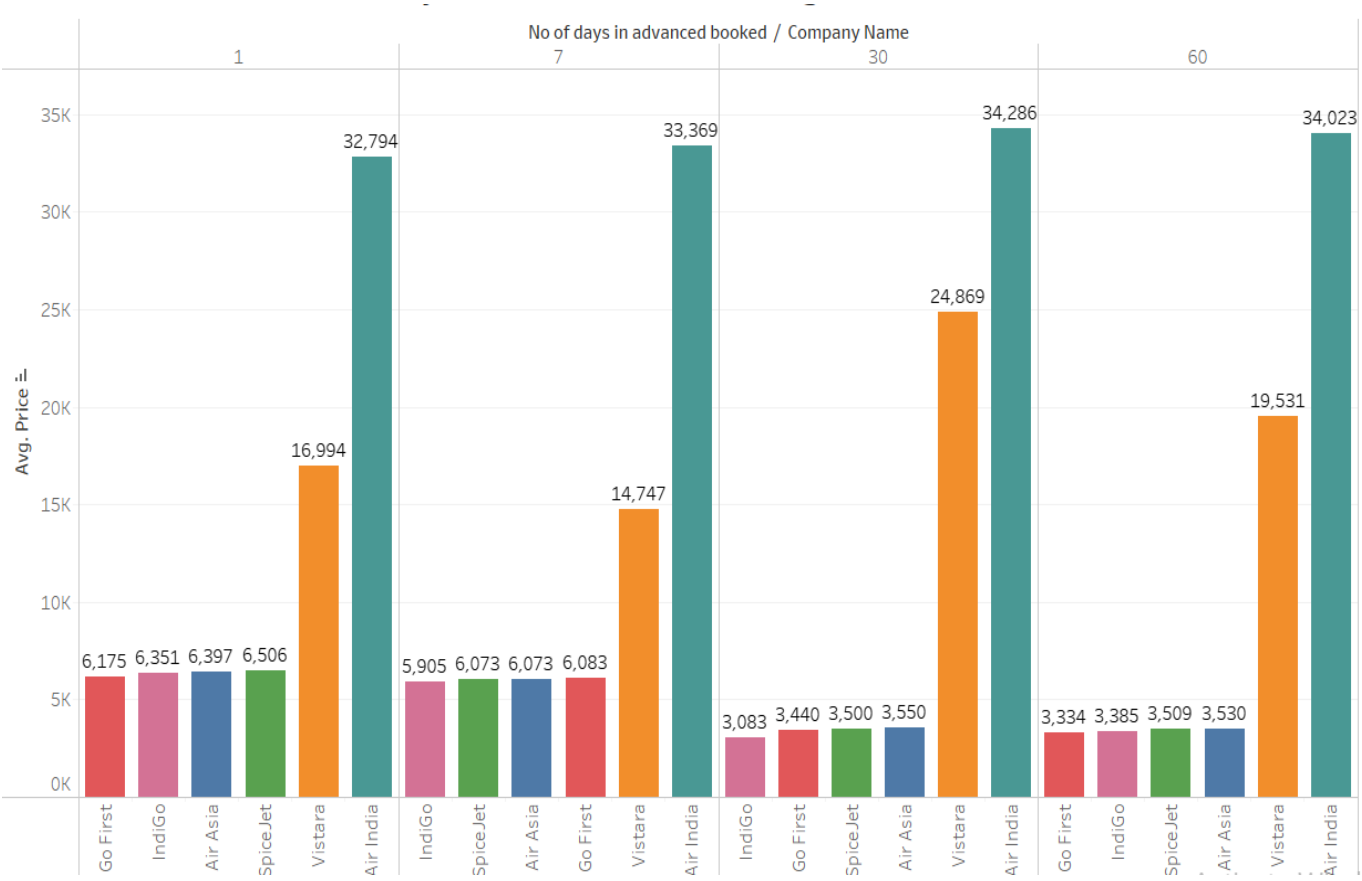
- The above graph depicts that if you want to book for Economy class then if you choose 4am-5am in morning you have maximum chance that your booking price will be lowest in that whole day.
- The highest price is generally observed for Economy class is around 3-5pm.
- For premium economy class the least price observed around 11am to 12pm and costly price is observed between 3pm -4 pm
- For Business class the least price observed around 11am to 12pm and costly price is observed between 3pm -4 pm

6.How duration and price are related based on class and destination?



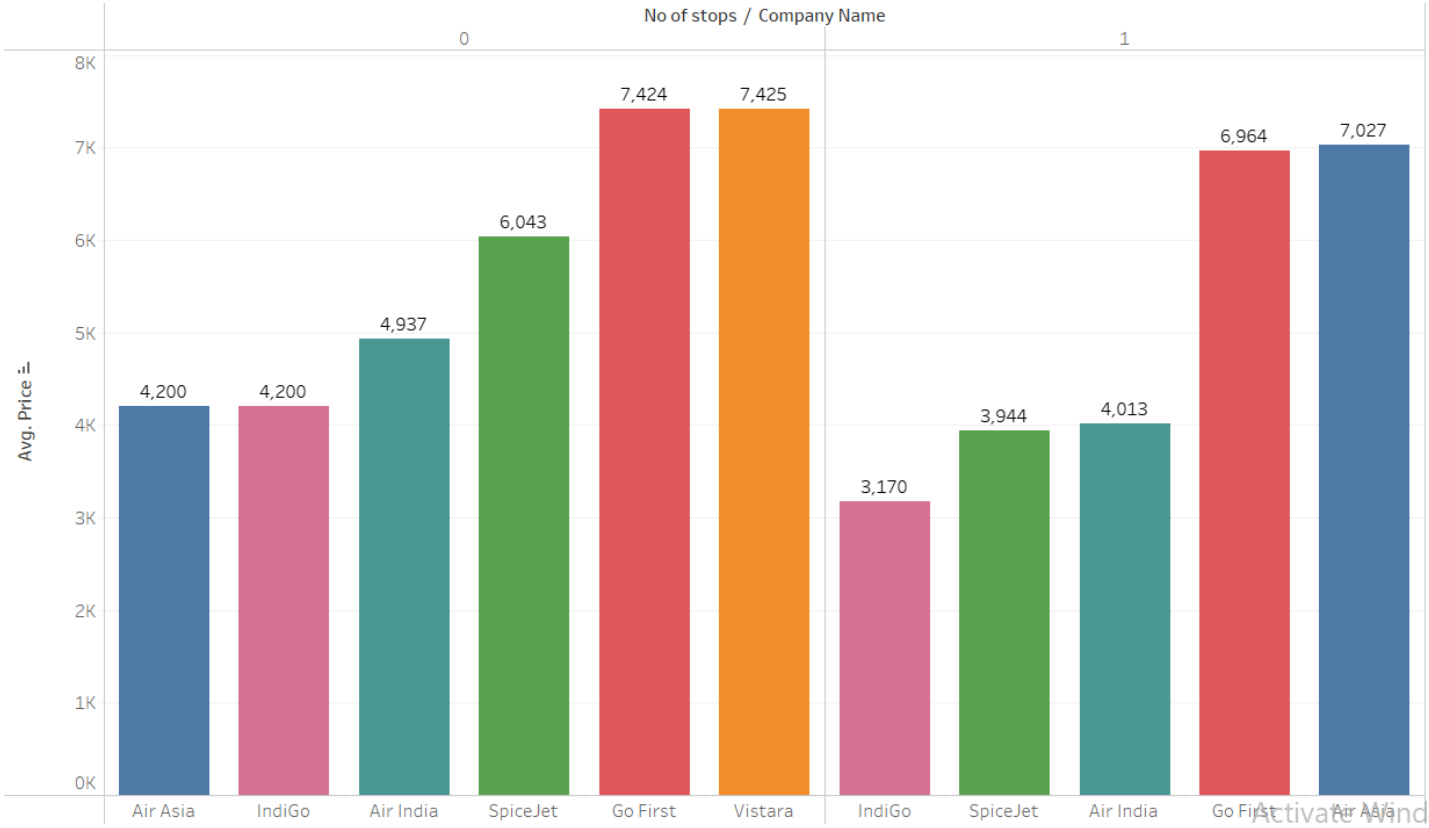
- From the above graph we can clearly see that as duration increases the price is also going to be increased.

7. How advanced booking affects the price of the ticket?



- If you book any ticket in an advance of 7 days you will be profited by 4.37% of cost of booking one day in advance
- If you book any ticket in an advance of 30 days you will be profited by 51% of cost of booking one day in advance
- There won't be much difference if you book 30 or 60 days in advanced.

8. How number of stops varies the price of the ticket to the same destination?



- On an average if there is one stops the ticket price will drop by 24% of the cost of ticket without any stop.

Model Building

We have used several linear regression models to evaluate and finalize the best models, The major models we have used as follows.

1. Linear Regression

```
#linear model
ln=LinearRegression()
ln.fit(x_train,y_train)
predln=ln.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predln)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predln)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predln)),3))
```

```
r2 score is : 0.813
RMSE: 7111.926
mean absolute error: 5103.609
```

2. Lasso Regression model

```
#lasso model
ls=Lasso(alpha=9)
ls.fit(x_train,y_train)
predls=ls.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predls)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predls)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predls)),3))
```

```
r2 score is : 0.813
RMSE: 7104.568
mean absolute error: 5078.614
```

3. Ridge Regression

```
#Ridge model
rd=Ridge(alpha=16)
rd.fit(x_train,y_train)
predrd=rd.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predrd)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predrd)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predrd)),3))
```

```
r2 score is : 0.814
RMSE: 7099.817
mean absolute error: 5045.134
```

4. Elasticnet Regression

```
#ElasticNet model
enr=ElasticNet(alpha=0.001)
enr.fit(x_train,y_train)
predenr=enr.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predenr)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predenr)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predenr)),3))
```

```
r2 score is : 0.813
RMSE: 7110.851
mean absolute error: 5100.703
```

5.Ransac Regressor

```
ran = RANSACRegressor(base_estimator=LinearRegression(), max_trials=100)
ran.fit(x_train, y_train)
predran=ran.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predran)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predran)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predran)),3))
```

```
r2 score is : 0.762
RMSE: 8019.396
mean absolute error: 4890.685
```

6.Support Vector Regressor

```
svr=SVR()
svr.fit(x_train, y_train)
predpoly=svr.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predpoly)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predpoly)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predpoly)),3))
```

```
r2 score is : -0.169
RMSE: 17787.638
mean absolute error: 13066.372
```

7.Random Forest Regressor

```
rf = RandomForestRegressor(n_estimators=100)
rf.fit(x_train, y_train)
predrf=rf.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predrf)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predrf)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predrf)),3))
```

```
r2 score is : 0.976
RMSE: 2537.227
mean absolute error: 1190.071
```

8.AdaBoost Regressor

```
ada = AdaBoostRegressor()
ada.fit(x_train, y_train)
predada=ada.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predada)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predada)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predada)),3))
```

```
r2 score is : 0.847
RMSE: 6424.581
mean absolute error: 5423.27
```

Cross Validation Score

```
models=[ln,ls,rd,enr,ran,svr,rf,ada]
for m in models:

    score=cross_val_score(m,x,y,cv=5)
    print(m,'score is:')
    print(round((score.mean()),3))
    print('\n')
```

```
LinearRegression() score is:
0.672


Lasso(alpha=9) score is:
0.676


Ridge(alpha=16) score is:
0.688


ElasticNet(alpha=0.001) score is:
0.673


RANSACRegressor(base_estimator=LinearRegression()) score is:
0.688


SVR() score is:
-0.241


RandomForestRegressor() score is:
0.772


AdaBoostRegressor() score is:
0.687
```

The difference between accuracy and cross validation is less for random forest regressor, so it is the best model

Hyper Parameter Tuning

```
rf=RandomForestRegressor()
grid_param={
    'criterion':['mse','mae'],

    'max_depth':[10,20,30,40,50],
    'max_features':['auto', 'sqrt', 'log2'],
    'min_samples_split':[2,5,10,15,20],
    'bootstrap':[True,False]
}

gd_sr=GridSearchCV(estimator=rf,
                   param_grid=grid_param,
                   scoring='r2',
                   cv=5)

gd_sr.fit(x,y)

best_parameters=gd_sr.best_params_
print(best_parameters)
best_result=gd_sr.best_score_
print(best_result)
```

```
{'bootstrap': True, 'criterion': 'mse', 'max_depth': 40, 'max_features': 'auto', 'min_samples_split': 20}
0.7991795802183211
```

Final Accuracy

```
rf = RandomForestRegressor(n_estimators=100)
rf.fit(x_train, y_train)
predrf=rf.predict(x_test)
print('r2 score is :',round((r2_score(y_test,predrf)),3))
print('RMSE:',round(np.sqrt(mean_squared_error(y_test,predrf)),3))
print('mean absolute error:',round((mean_absolute_error(y_test,predrf)),3))
```

```
r2 score is : 0.976
RMSE: 2537.227
mean absolute error: 1190.071
```

Conclusions

- **Key findings and the conclusions of the study**

  Based on the above key findings we would like to give some prescription for our client
  if you want to travel in Business class choose vistara it has least average price,and for Economy class choose IndiGo, and for Premium Economy go for Air Asia.
  • If you want to travel from Delhi to Hyderabad in Economy class then Air India will offer least price followed by GoFirst and Vistara and Air Asia is the costly one.
  • If you want to travel from Delhi to Kolkata in Economy class then Air India will offer least price followed by IndiGO and Vistara and SpaceJet is the costly one.
  • If you want to travel from Delhi to Mumbai in Economy class then SpaceJet will offer least price followed by Air India and IndiGo and Air Asia is the costly one.
  • If you want to travel from Mumbai to Bengaluru in Economy class then IndiGo will offer least price followed by Vistara and Air India and SpaceJet is the costly one.
  • If you want to travel from Delhi to Bengaluru in PremiumEconomy class then Vistara is the best choice
  • If you want to travel from Delhi to Chennai in Premium Economy class then Vistara is the best choice
  • If you want to travel from Delhi to Hyderabad in Premium Economy class then Vistara
  • If you want to travel from Delhi to Kolkata in Premium Economy class then Air Asia is the cheapest one followed by GoFirst and SpiceJet
  • If you want to travel from Delhi to Mumbai and Mumbai to Bengaluru in Premium Economy class then Vistara is the best choice.
  • If you want to go from Delhi to Bengaluru Chennai or Hyderabad or Mumbai or Mumbai to Bengaluru then vistara is the only available as per this data at this moment, It may vary on coming days.
  • If you want to go from Delhi to Kolkata in Business class then Air Asia is the cheapest one followed by GoFirst and SpaceJet and Vistara is the costliest one according to this data.
  • The above graph depicts that if you want to book for Economy class then if you choose 4am-5am in morning you have maximum chance that your booking price will be lowest in that whole day.
  • The highest price is generally observed for Economy class is around 3-5pm.
  • For premium economy class the least price observed around 11am to 12pm and costly price is observed between 3pm -4 pm
  • For Business class the least price observed around 11am to 12pm and costly price is observed between 3pm -4 pm
  • From the above graph we can clearly see that as duration increases the price is also going to be increased.

- **Learning outcomes of the study in respect to data science**

1. How to deal with XPaths and obtain URLs - XPaths are an essential part of web scraping as they are used to identify and extract specific elements from a web page. Through this project, I have learned how to navigate through the HTML structure of a web page and find the right XPaths to extract the data we need.
2. The importance of having a clear vision for the project - Before starting the project, it is crucial to have a foresight of its complete picture and a clear understanding of the final outcome. This helps in planning and organizing the project and in identifying the necessary steps to achieve the goal.
3. Importance of understanding data types - I have learned that when working with data, it is crucial to understand the data types of different variables. In this project, I encountered issues with checking whether a variable was of object or integer type, and I learned that we should not include "int" in double inverted commas because it is a keyword.
4. Overall, the project has been a great learning experience and has enriched my knowledge in data science. It has helped me to apply my skills in a real-world problem and has provided me with a deeper understanding of the data science process.

!! Thank You !!