FLIGHT PRICE PREDICTION

Submitted by:

MOHAMMED MINHAJ

ACKNOWLEDGMENT

Wikipedia, Youtube.com, google.com. The data were collected from the used car website cardheko.com. Referred in study materials of machine learning which I am studying.

INTRODUCTION

- Business Problem Framing

As we all know, we can get flight fare for cheap and expensive. When we try to book a flight ticket today, for travelling tomorrow the fare may get higher. While we try to book flight for 2 weeks before travelling we may get cheap rate. So fare may vary in this factor and also if the flight is not direct to our destination it may also varies. According to the airlines also fare may vary. So this is the problem of this project. Now let's build a machine learning model what are the factors may affect in flight fare.

- Conceptual Background of the Domain Problem

As we want to predict the house price within features first of all we want to build a regression model. Try to relate all features with price how they are related to price by visualization.

- Review of Literature

For travelling to country to country we all choose flight for travel. Because it saves time. Recently we can choose flight to travel through other cities or states in our country. For this project, I scrape the data from flight booking website yatra.com. While scraping the data from I can observe that the price is varying in many factors. Different airlines have different fares, may be it will be cheap or low. And also, whether the flight is connected flight or directed flight. It may also vary in flight fare. And another factor I observe that the date. That is, when we try to book the ticket before one or two days before travelling, the fare will be very expensive. But when we try to book within a week or two weeks the fare is very much low. So we can conclude that when demands increases price increases.

- Motivation for the Problem Undertaken

Airline companies are using the seasonal days travelling very expensive fare. International fares may also vary according to the seasonal days. We hear about how the airline companies used the pandemic situation. For quarantine and others they have charge very high price from travellers. In this project, I focused on the domestic fares. The main objective is to find the fare price prediction of the flight, how it changing the price, what are the factors for affecting the price. I check about the economy class. Other classes like business and first class is very expensive. Let us focus on the fare of economy class, what are the factors affecting the fare.

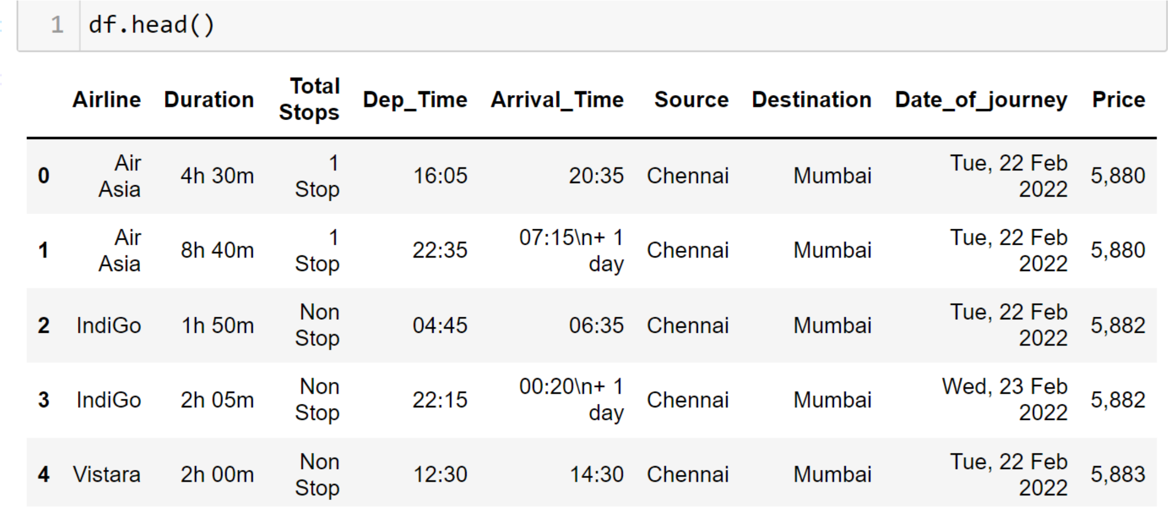
Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

Predictions are done by analysing the data. Here data are collected from the website yatra.com. For this project, we collected the future data. I mean that the fare of flight of the future. Means a day after or a month after. The machine predict the price using these data. By analysing the data, I understand some factors that affecting the price. What are the factors that affects for low price and what are the factors that affect for high price expensive. The data of the project is about the domestic fares and economy class. Different airline country have different prices. For example, the fare in 'Vistara' for same source and destination may be different fare for 'Spicejet' airlines. So the airline company is one of the factor for affecting the price. Another factor is the stops that is connection or direction flight. These may vary the flight fare. And the distance between the source and destination. These all are the basic factors that affects the price I understand by analysing the data.

- ## Data Sources and their formats

The data I mention that I collected from the flight booking website yatra.com. The website is so comfort for scraping it have all required data for our prediction. We want to collect the require feature for prediction, we want to choose wisely. Scraped around 1700 entries with 9 columns. The features are the airline name, source, from where you want to travel, destination, to where you want travel, stops, is the flight is connected or direct flight connected flight have stops but directed flight doesn't have stops. Departure time, when flight departure from the source and arrival time when the flight lands at the destination. At last our target variable price.



```
df.head()
```

	Airline	Duration	Total Stops	Dep_Time	Arrival_Time	Source	Destination	Date_of_journey	Price
0	Air Asia	4h 30m	1 Stop	16:05	20:35	Chennai	Mumbai	Tue, 22 Feb 2022	5,880
1	Air Asia	8h 40m	1 Stop	22:35	07:15\n+ 1 day	Chennai	Mumbai	Tue, 22 Feb 2022	5,880
2	IndiGo	1h 50m	Non Stop	04:45	06:35	Chennai	Mumbai	Tue, 22 Feb 2022	5,882
3	IndiGo	2h 05m	Non Stop	22:15	00:20\n+ 1 day	Chennai	Mumbai	Wed, 23 Feb 2022	5,882
4	Vistara	2h 00m	Non Stop	12:30	14:30	Chennai	Mumbai	Tue, 22 Feb 2022	5,883

- ## Data Pre-processing Done

As we know, the data for prediction may be not clean, before prediction we want to clean the data in proper manner. May be the data type f features that may not be read by machine, so we want to convert it. Here for this project the data are date format, time, categorical features. We want to clean it in appropriate manner. First of all, the features of time that is, departure time, arrival time and duration. Extract hour and minute from the departure time and arrival time, while in duration the time is given like '3h 40m' from that data I extract minute and hour separately and replace 'h' and 'm' from the data. And also in date the format is different so I change the format into 'dd/mm/yyyy' using replace function. And then convert to date time and separate date and month. Year is same so I rejected it. And our target variable is in object data type, replace the comma (,) using replace function and convert to numeric. And at last the encoding, create dummies for airlines and stops. Encode the source and destination using label encoder. These all are the pre-processing steps done for cleaning data.

- ## Data Inputs- Logic- Output Relationships

  For a better prediction machine needs the better features for prediction. Then only machine can understand the data and give a better prediction so in this project we are predicting the flight fare, so for that we need appropriate features. Here after analysing data, I understand that, fare may vary in some features that we extract. The airline name, different company were owned by different authorities. So the fare may different for different. And the source and destination, how much far is between source and destination may also affect the fare. The feature stops, whether the flight is direct or connected flight. Connected flight have stops and direct flight doesn't have, so it is also major feature for price. Next the duration, how much time you travel in flight also impacts the price. When the duration is high the price also increases. so in these ways the feature affects the target price or flight fare.

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

  As I mention that the flight fare may increase when the flight have stops and the duration of travelling increases also may affect the price or flight fare. The duration will indicate the how much far from the source and destination.

- Hardware and Software Requirements and Tools Used

  As usual the dataset is small the memory usage is also small around 133 kb of memory is used for this dataset. For data analysing and visualizing here we use pandas packages, matplotlib and seaborn for visualization technique. The packages for five algorithms tree package for decision tree, linear model package for linear regression, packages for support vector machine, ensemble packages for random forest. Package for Boosting technique xg boost regression. For performing an algorithm first of all we need to split it into train and test model for that here we can use the package model selection. In same package we can use for hyper parameter tuning and to find cross validation score. And metrics packages used for finding mean squared error, mean absolute error and R2 score. For scaling and encoding imported from the package preprocessing. For removing outliers, Zscore is used the package scipy is used for importing.

Model/s Development and Evaluation

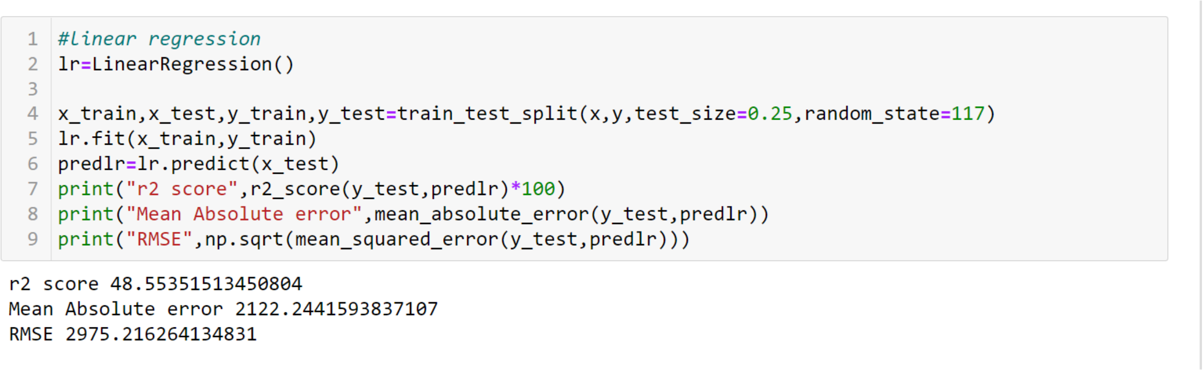
- Identification of possible problem-solving approaches (methods)

  Our objective in this project is to predict the flight price. For the prediction we need some features related to the label. Here for the flight fare is related to some features of the flight. We can analyse the data by visualizations for better understanding of data.

- Testing of Identified Approaches (Algorithms)
  - Decision Tree
  - Linear Regression
  - Support Vector
  - Random Forest
  - XG Boost

- Run and Evaluate selected models

Linear Regression



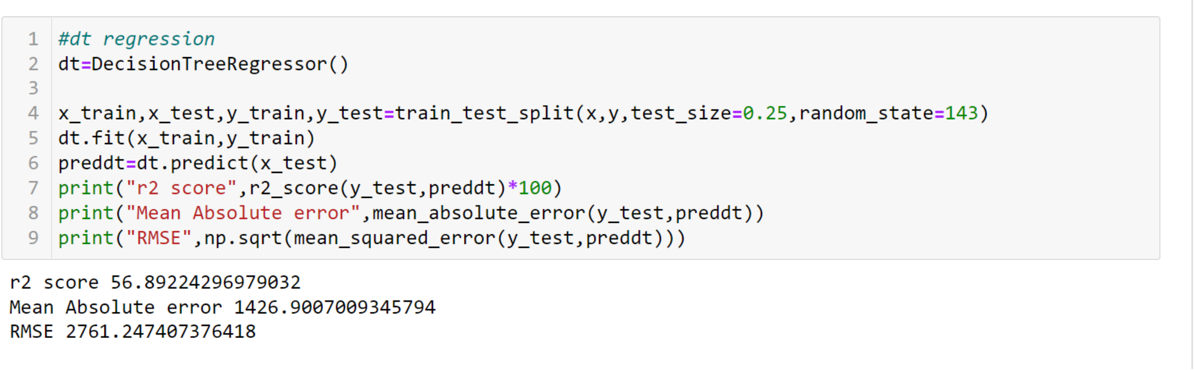
```python
#linear regression
lr=LinearRegression()

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=117)
lr.fit(x_train,y_train)
predlr=lr.predict(x_test)
print("r2 score",r2_score(y_test,predlr)*100)
print("Mean Absolute error",mean_absolute_error(y_test,predlr))
print("RMSE",np.sqrt(mean_squared_error(y_test,predlr)))
```

```
r2 score 48.55351513450804
Mean Absolute error 2122.2441593837107
RMSE 2975.216264134831
```

Model split into train and test data, 25% of data were used for testing and 75% of the data were used for training. R2 score of linear regression is very low. So we can say that linear regression perform badly for this dataset.

Decision Tree



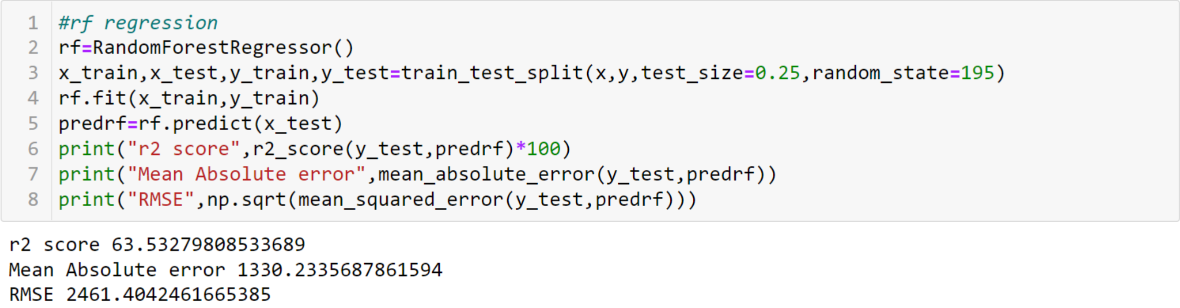
```python
#dt regression
dt=DecisionTreeRegressor()

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=143)
dt.fit(x_train,y_train)
preddt=dt.predict(x_test)
print("r2 score",r2_score(y_test,preddt)*100)
print("Mean Absolute error",mean_absolute_error(y_test,preddt))
print("RMSE",np.sqrt(mean_squared_error(y_test,preddt)))
```

```
r2 score 56.89224296979032
Mean Absolute error 1426.9007009345794
RMSE 2761.247407376418
```

Model split into train and test data, 25% of data were used for testing and 75% of the data were used for training. R2 score of decision tree is very low is about 56%. So we can say that decision tree perform badly for this dataset.

Random Forest

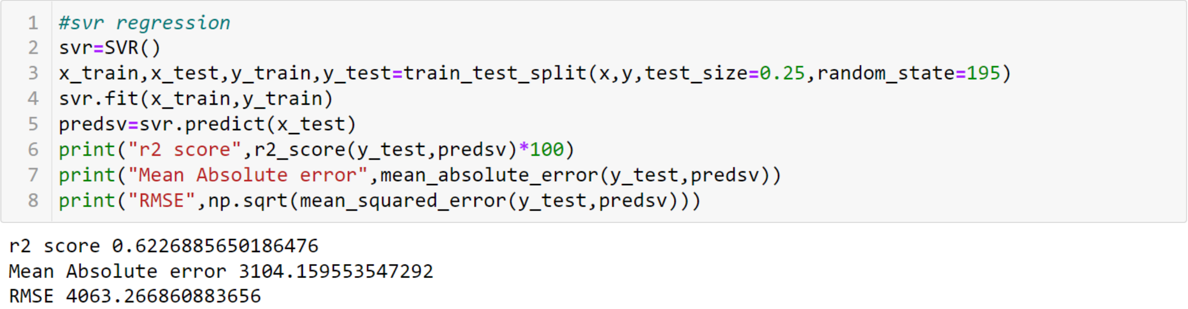


```
#rf regression
rf=RandomForestRegressor()
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=195)
rf.fit(x_train,y_train)
predrf=rf.predict(x_test)
print("r2 score",r2_score(y_test,predrf)*100)
print("Mean Absolute error",mean_absolute_error(y_test,predrf))
print("RMSE",np.sqrt(mean_squared_error(y_test,predrf)))
```

```
r2 score 63.53279808533689
Mean Absolute error 1330.2335687861594
RMSE 2461.4042461665385
```

Model split into train and test data, 25% of data were used for testing and 75% of the data were used for training. R2 score of random forest is about 63%. The dataset is small we can say that the r2 score of random forest is fine for this dataset.

Support Vector

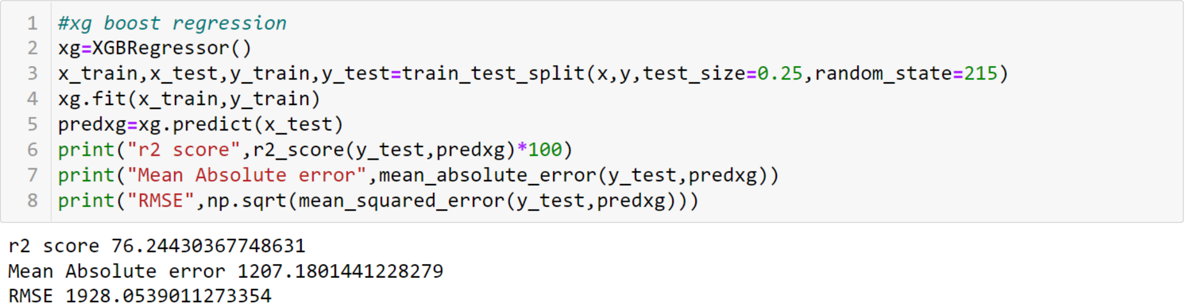


```
#svr regression
svr=SVR()
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=195)
svr.fit(x_train,y_train)
predsv=svr.predict(x_test)
print("r2 score",r2_score(y_test,predsv)*100)
print("Mean Absolute error",mean_absolute_error(y_test,predsv))
print("RMSE",np.sqrt(mean_squared_error(y_test,predsv)))
```

```
r2 score 0.6226885650186476
Mean Absolute error 3104.159553547292
RMSE 4063.266860883656
```

Model split into train and test data, 25% of data were used for testing and 75% of the data were used for training. R2 score of support vector is too low. So we can say that support vector perform badly for this dataset.

XG Boost



```
#xg boost regression
xg=XGBRegressor()
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=215)
xg.fit(x_train,y_train)
predxg=xg.predict(x_test)
print("r2 score",r2_score(y_test,predxg)*100)
print("Mean Absolute error",mean_absolute_error(y_test,predxg))
print("RMSE",np.sqrt(mean_squared_error(y_test,predxg)))
```

```
r2 score 76.24430367748631
Mean Absolute error 1207.1801441228279
RMSE 1928.0539011273354
```

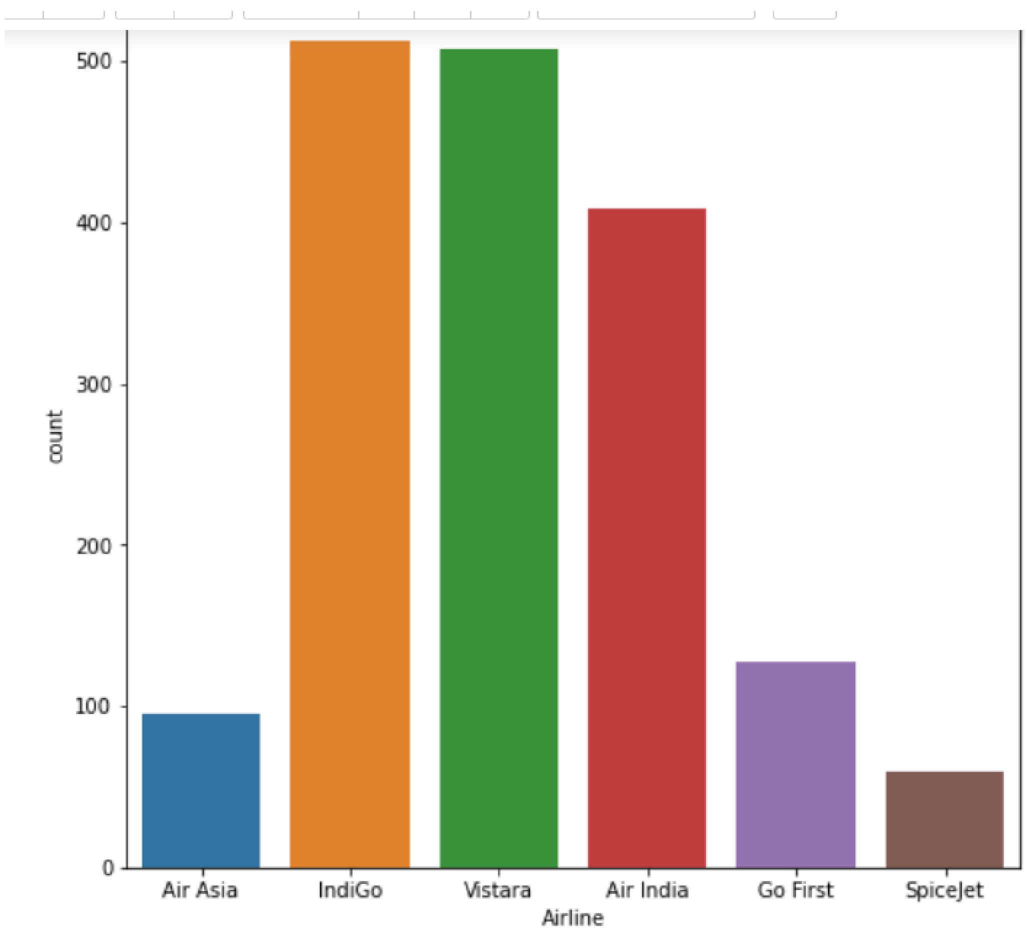
Model split into train and test data, 25% of data were used for testing and 75% of the data were used for training. R2 score of XG Boost is about 76%. Good among other algorithms.

- Key Metrics for success in solving problem under consideration

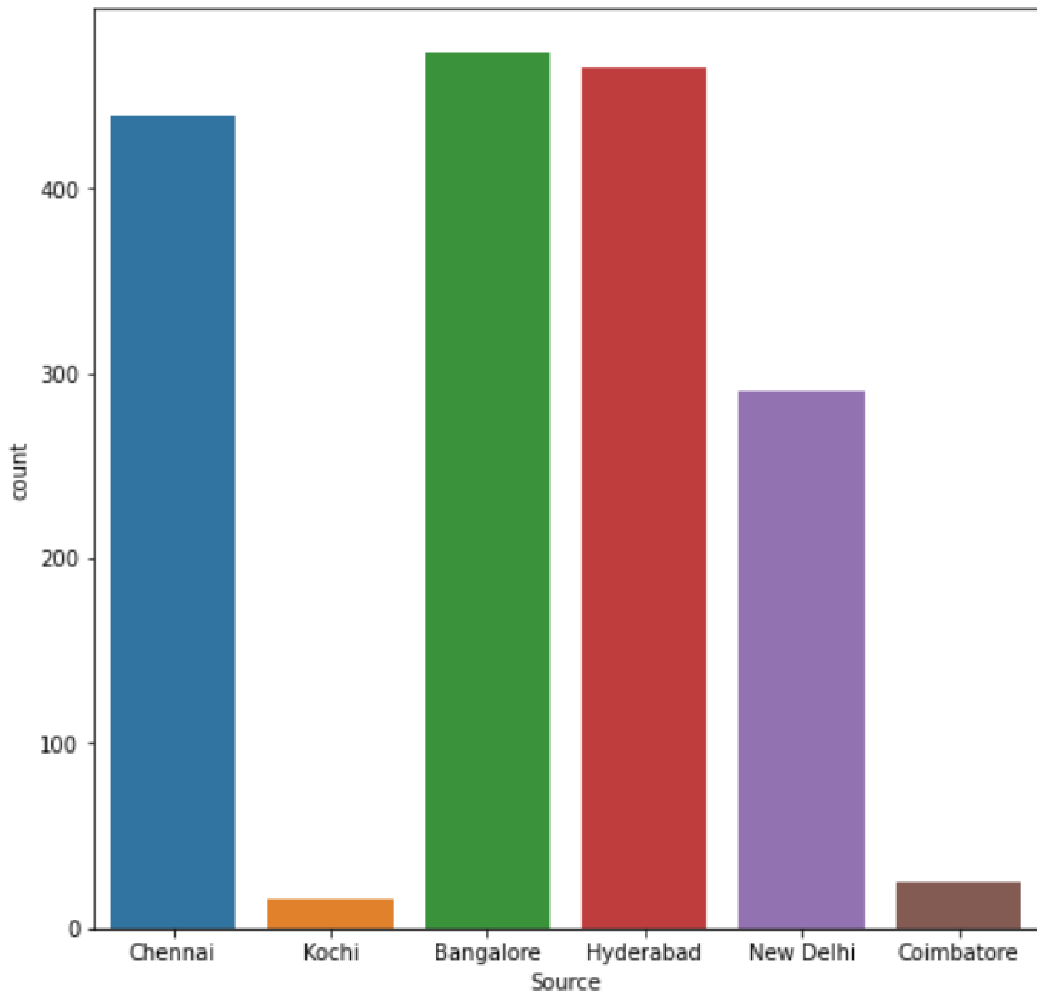
R2 score, Mean absolute error (MSE) and Root Mean Squared error (RMSE) were used in this project. Here the R2 score of the project using XG Boost regression is 74%. As we know the Mean squared error told us that how close the regression line with set of points. Least difference makes a good regression line. While RMSE is the standard deviation of the residuals or errors. These are the key metrics used in this project.

- Visualizations

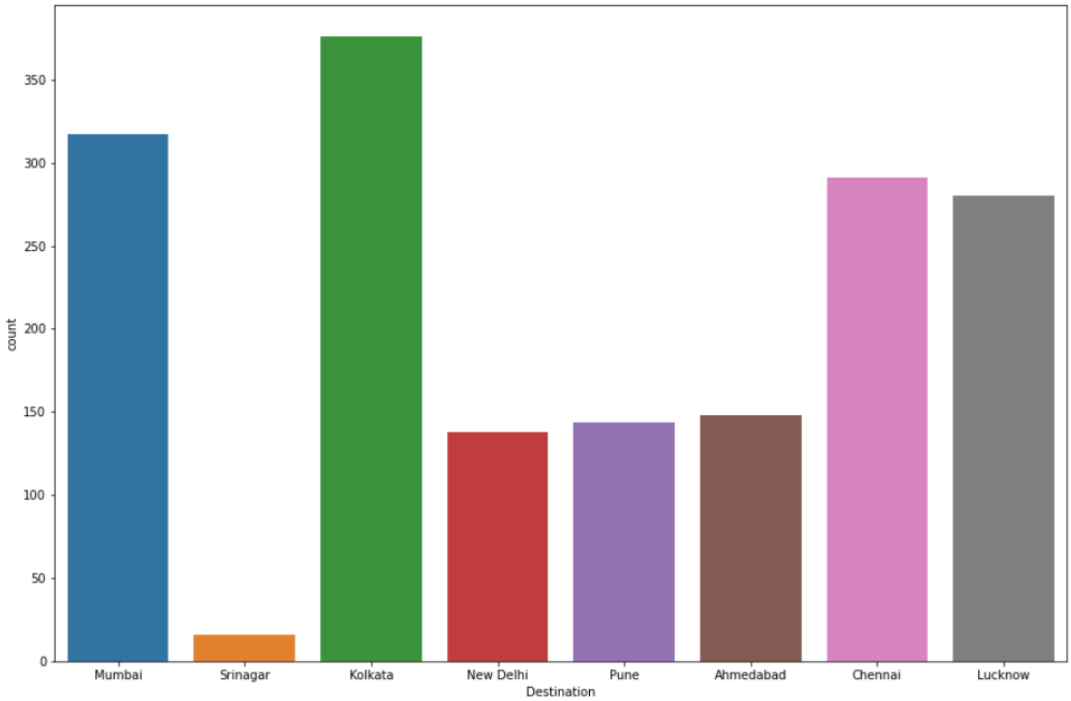
Visualization is one of the best tool for understanding and analysing the data. Here in this project I analyse the categorical and numerical features separately.



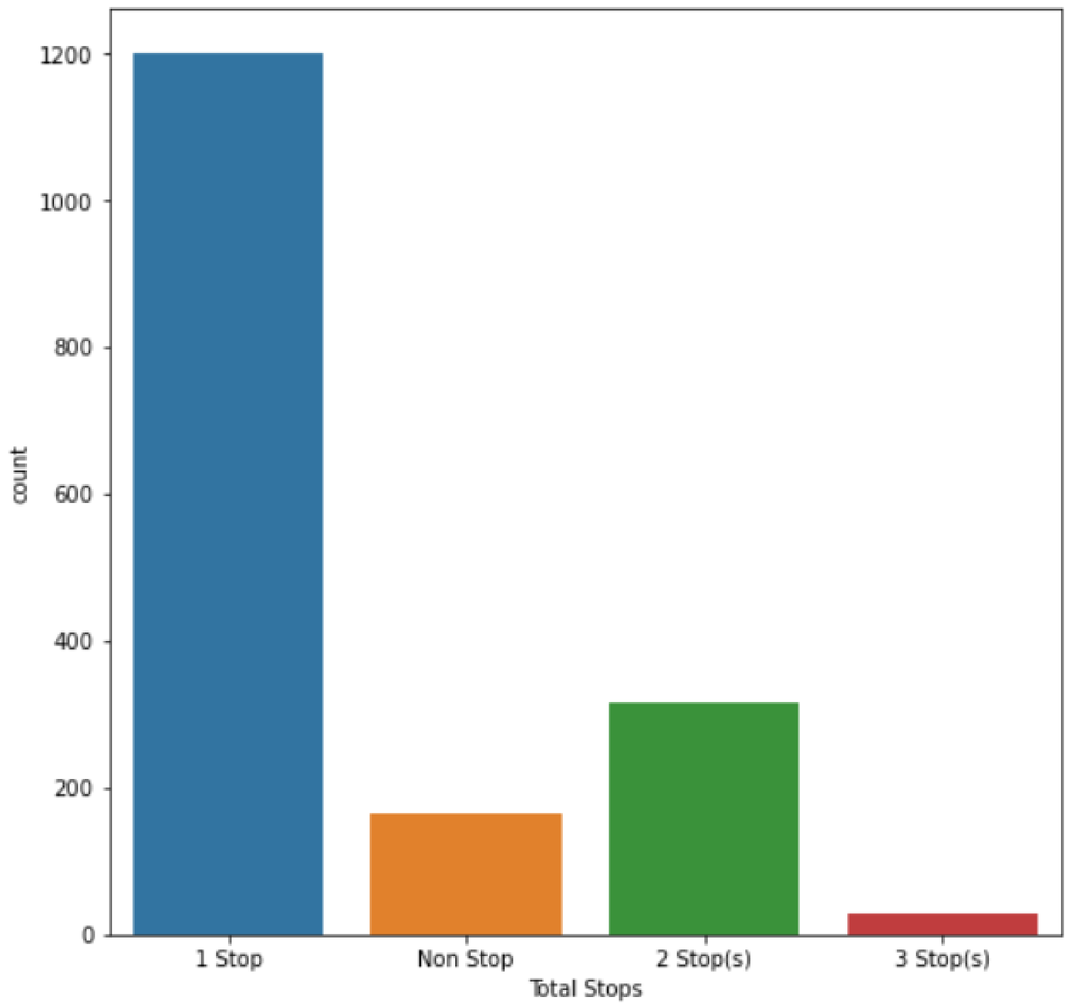
Here we can observe that the most of the data were of vistara airlines and Indigo airlines.



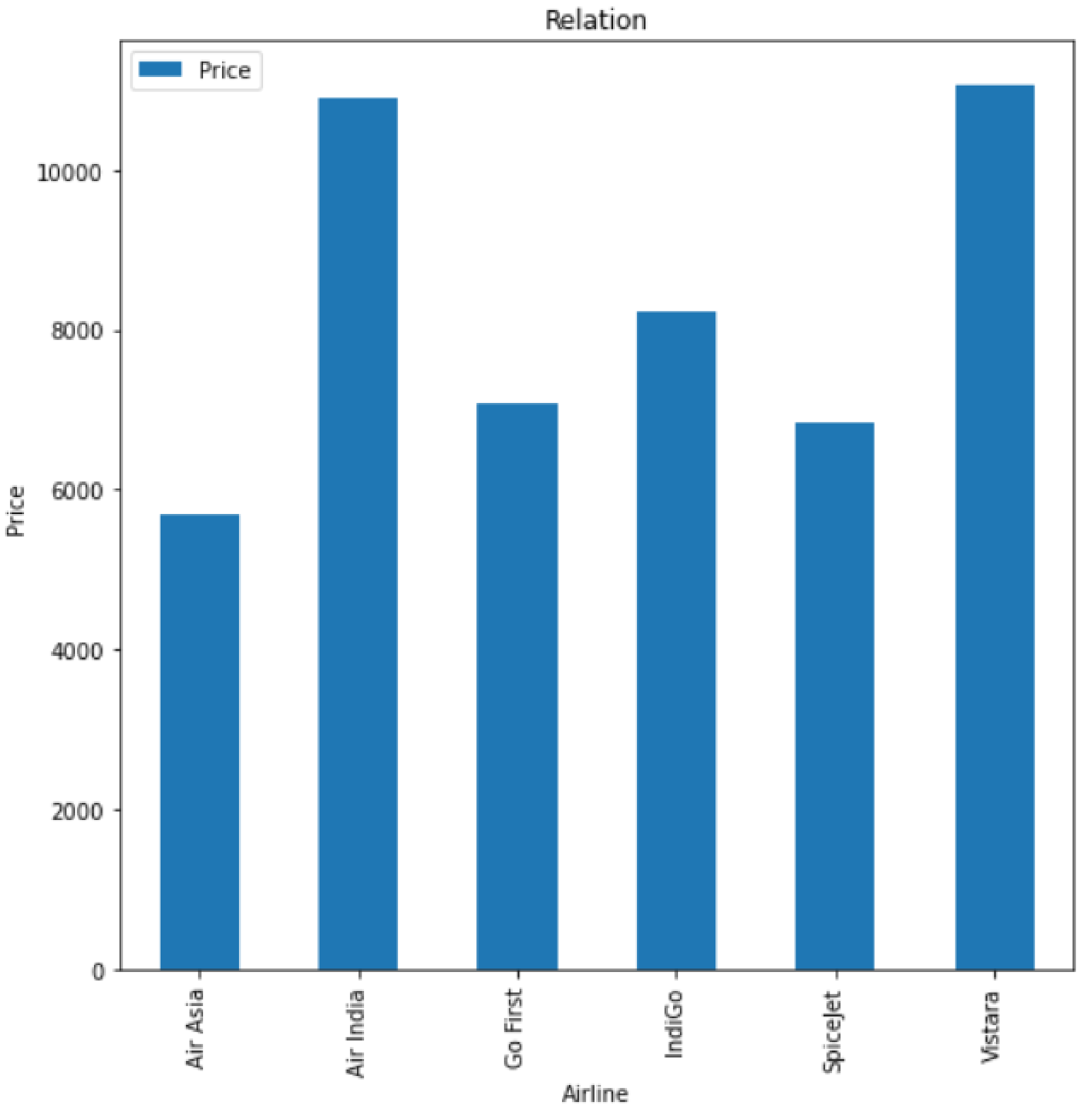
The source, most of the source is from Bangalore and Hyderabad.



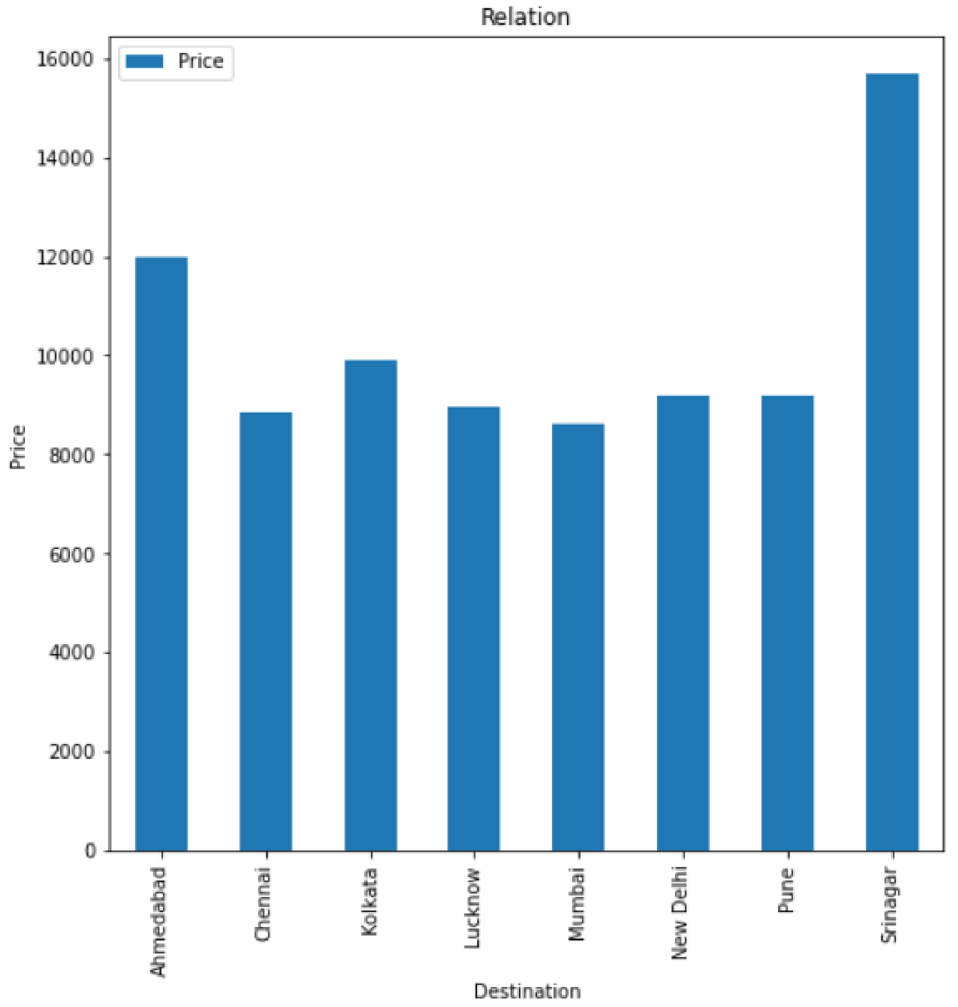
Kolkata is the destination in most of the data.



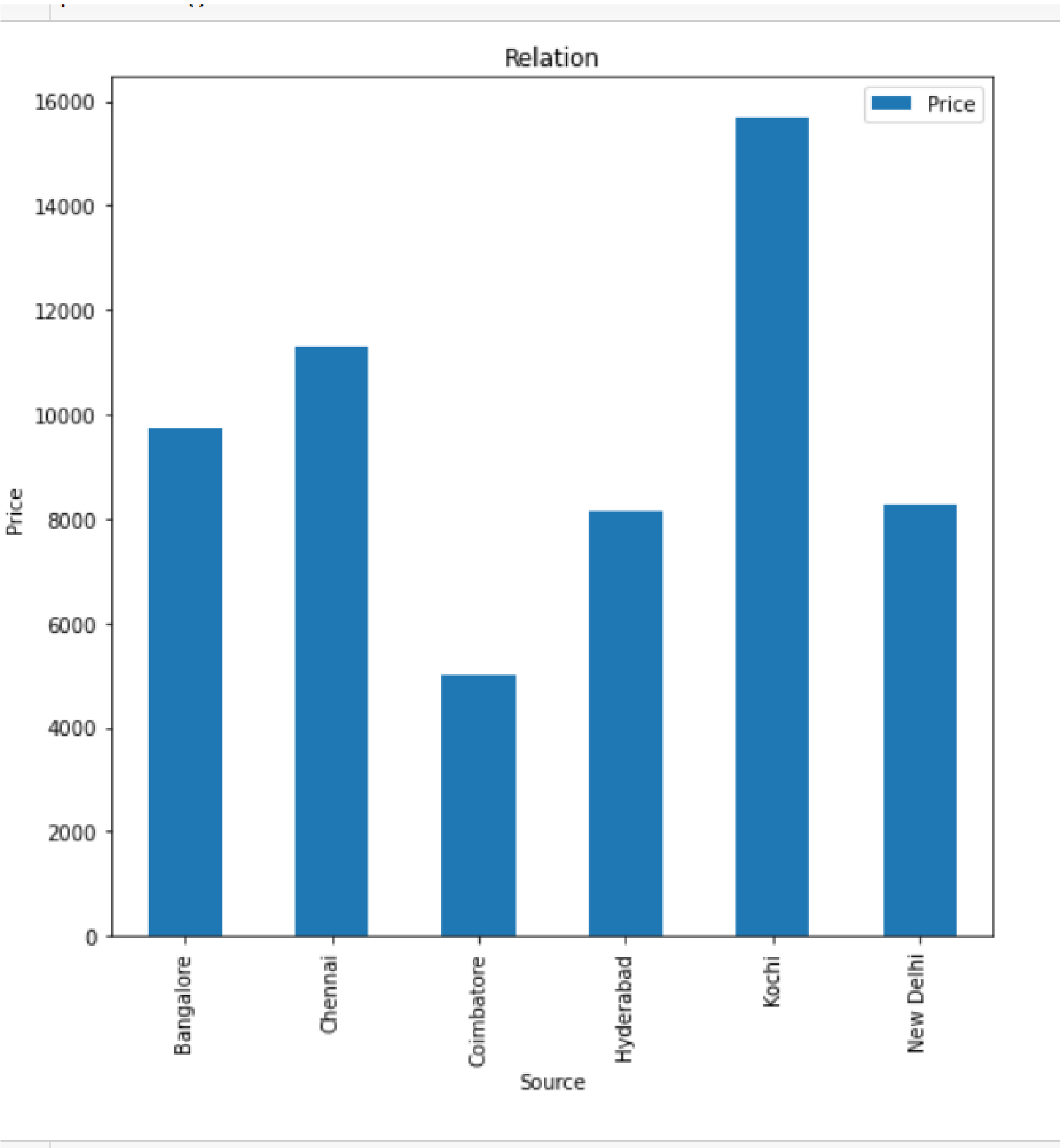
Most of the flights were direct flight.



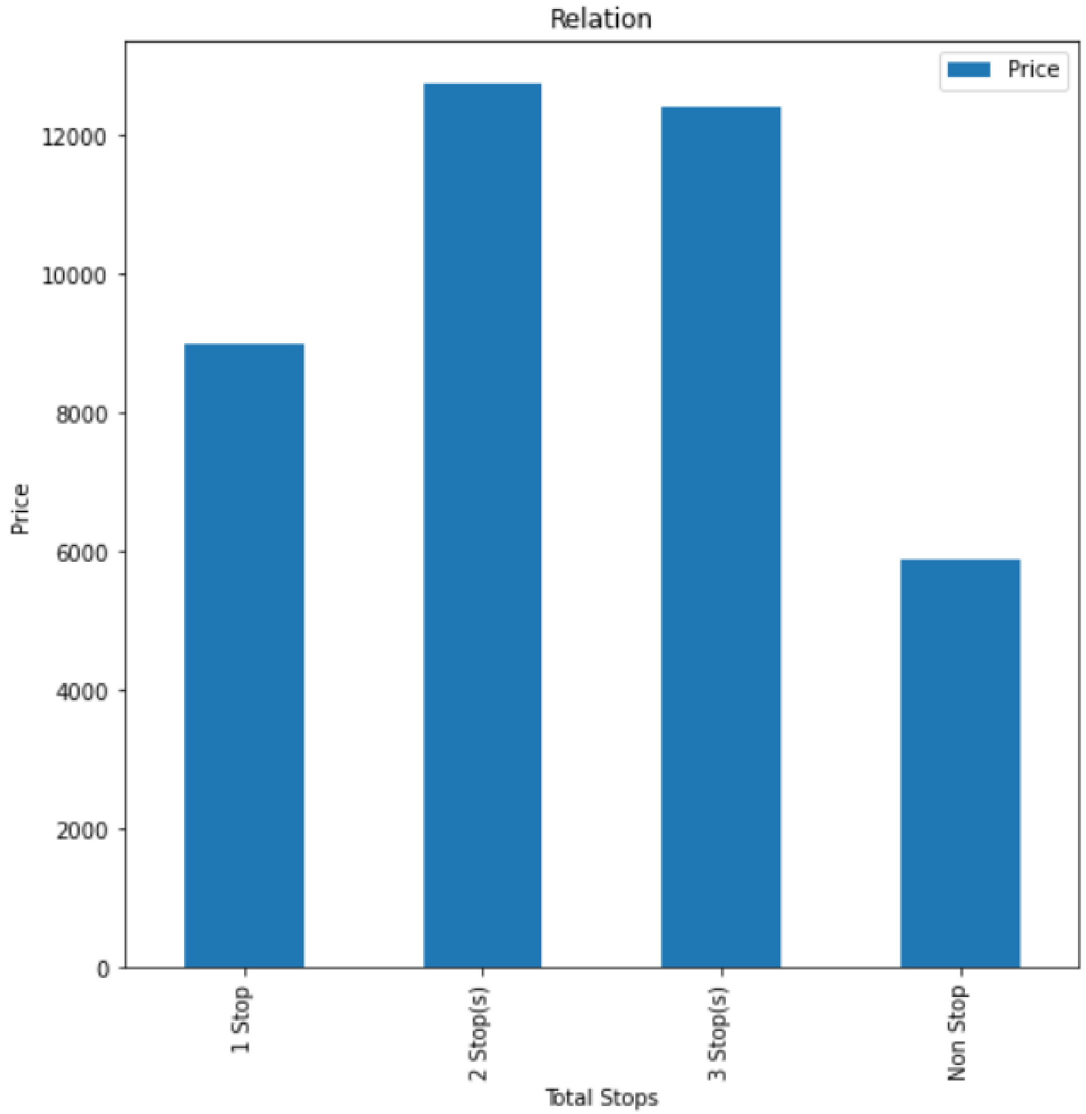
While we comparing the price and airline names, air India and vistara airline have high flight fare.



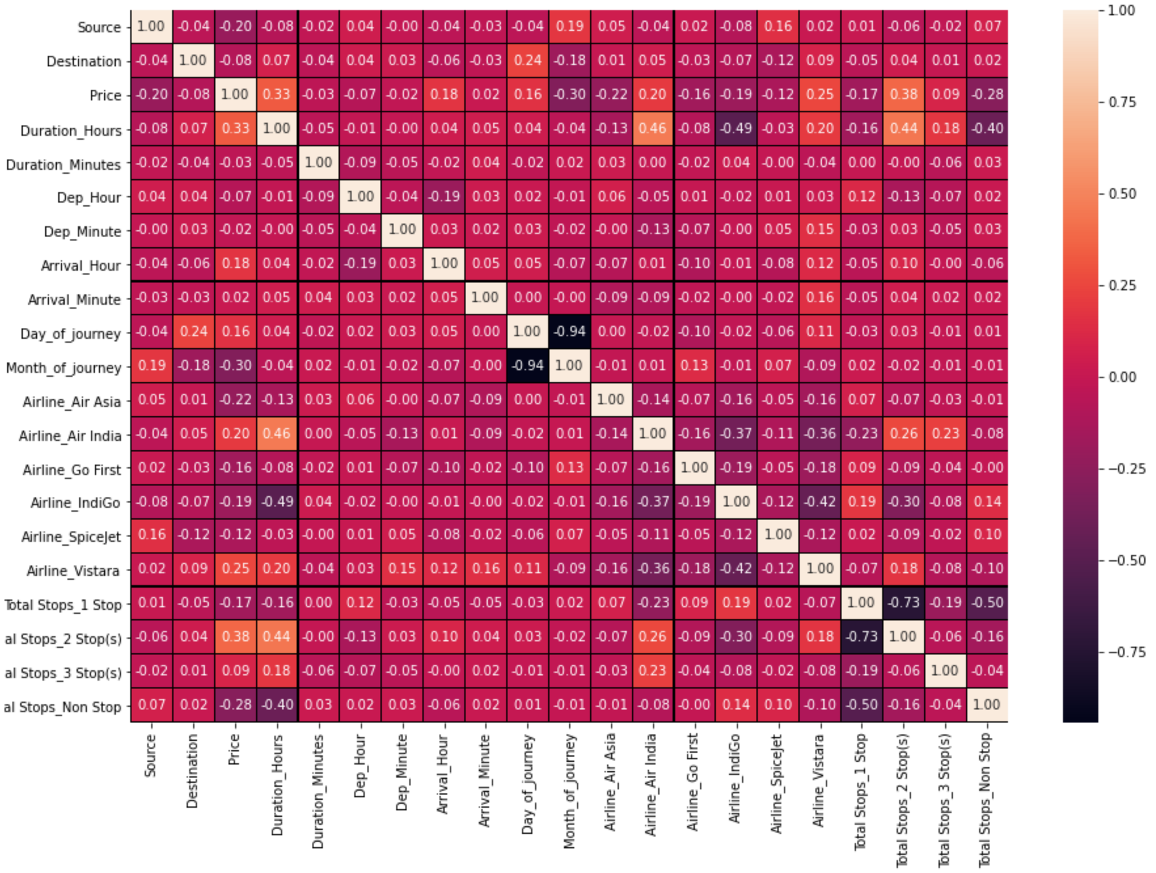
While we comparing the destination, Srinagar have high flight fare.



While comparing price and source Cochin have high price.



While comparing price and stops, connected flight have high price or flight fare. We can observe that flight with 2 stops and 3 stops have high flight fare.



Correlation

In the above figure we can observe the correlation. How the features impacts the label price. Duration hours and flight with two stops are highly correlated to the price.

- ## Interpretation of the Results

After a better understanding and analysing the data, I understand that which features affects the label price. The duration hours that is the distance between the source and destination. And also when the flight have stops the duration may increase, there is no matter about the distance. And also the flight with two stops have high fare. Besides that, while scrapping the data, I understand that the flight fare is high when I try to book for the flight one day before travelling. While in other case when we try to book the flight before one or two weeks of travelling, the flight fare is cheap. So I can conclude that besides the travel date, booking date also impacts the price.

CONCLUSION

- Key Findings and Conclusions of the Study

Travelling in flight from one place to another benefits in many ways. The major benefit is time. We can save time by travelling in flight. So every employees working in different field use the benefit of flight. So in this project we try to find the price of the flight what are the major features affects the price. So I can conclude that duration and stops are highly correlated to price and it impacts price.

- Learning Outcomes of the Study in respect of Data Science

Visualization gives a good understanding of data how the features are related to the label and how features are related to each other. So such basic and detailed understanding of data is given by the visualization tools. In this projects I used five different algorithms. Linear regression perform very badly. It is because we can assume that the data are not linearly scattered. While this problem solved by boosting technique XG Boost. It gives high performance on the project. After tuning with best parameters it gives 74% of accuracy. So it is common for a data scientist occurrence of over fitting and under fitting of data so it is overcome by many techniques, which one is chosen by the data scientist.

- Limitations of this work and Scope for Future Work

For getting cheap price or flight fare we want to select the flights which have non-stops which may affects the flight fare. Here in this project, we have small dataset, which gives low accuracy. Machine wants more data to learn. Then only it give a good accuracy and prediction. More data gives more features. We have different algorithms with different concepts. Every algorithm not performs in every data. Different data can have different best algorithms. So we want to perform multiple algorithms and choose wisely which algorithm is best.