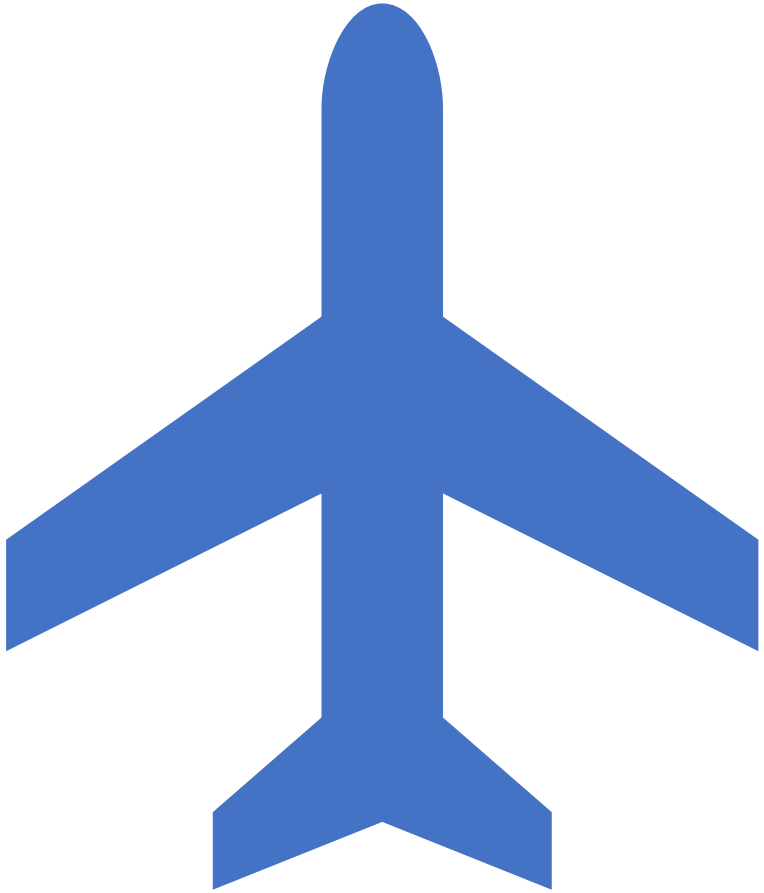
BDSN PROJECT

FLIGHT PRICE PREDICTION

NAME: SHUBHAM BATHWAL

STUDENT ID : A21031



CONTENT

- OBJECTIVE
- RESEARCH QUESTIONS
- ABOUT DATASET
- PROJECT FLOW
  - DATA COLLECTION AND METHODOLOGY
  - LOADING FILES AND SETTING UP ENVIRONMENT
  - DATA CLEANING
  - EXPLORATORY DATA ANALYSIS
  - OUTLIER TREATMENT
  - DATA PREPROCESSING
- MODEL BUILDING
  - LINEAR REGRESSION
  - DECISION TREE REGRESSOR
  - GRADIENT BOOSTING
  - RANDOM FOREST REGRESSOR
  - HYPERTUNING PARAMETERS
- CONCLUSION

OBJECTIVE

The objective of this research is to examine flight booking data in order to determine what factors influence price at the time of booking.

Find out the factors that have more influence on price compared to others. Based on the knowledge find a method to predict the price of the flight given all the other factors.

NEED

A comprehensive examination of the data will assist in the finding of useful information that will be extremely beneficial to passengers.

It will provide customers with information on what things to consider in order to get a better deal.

RESEARCH QUESTIONS:

The aim of our study is to answer the below research questions:

a) Does price vary with Airlines?

b) How is the price affected when tickets are bought in just 1 or 2 days before departure?

c) Does ticket price depend on the departure time and arrival time?

d) How the price changes with change in Source and Destination?

e) How does the ticket price vary between Economy and Business class?

f) Does ticket price depend on the number of stops between the Source and Destination City?

ABOUT DATASET

Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 datapoints and 11 features in the cleaned dataset.

FEATURES

The various features of the cleaned dataset are explained below:

1) **Airline**: The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.

2) **Flight**: Flight stores information regarding the plane's flight code. It is a categorical feature.

3) **Source City**: City from which the flight takes off. It is a categorical feature having 6 unique cities.

4) **Departure Time**: This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels.

FEATURES

5) **Stops**: A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.

6) **Arrival Time**: This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.

7) **Destination City**: City where the flight will land. It is a categorical feature having 6 unique cities.

8) **Class**: A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.

9) **Duration**: A continuous feature that displays the overall amount of time it takes to travel between cities in hours.

10)**Days Left**: This is a derived characteristic that is calculated by subtracting the trip date by the booking date.

11) **Price**: Target variable stores information of the ticket price.

PROJECT FLOW

DATA COLLECTION AND METHODOLOGY

- OctoParse scraping tool was used to extract data from the website. Data was collected in two parts: one for economy class tickets and another for business class tickets. A total of 300261 distinct flight booking options was extracted from the site. Data was collected for 50 days, from February 11th to March 31st, 2022.
- Data source was secondary data and was collected from Ease my trip website.
- Data obtained was in JSON format.
- Data was compressed in zip format and was uploaded in GitHub.

LOADING FILES AND SETTING UP ENVIRONMENT

- Data collected was then loaded in Google Collab Notebook.
- All the necessary libraries were imported.
- MongoDB and Spark were installed and their environment were set.
- Database named Flight was created.
- Two collection Business and Economy were created and the respective JSON files were loaded into it.
- JSON files from the database were loaded as spark data frame.
- Economy and Business data frame were merged to form a single spark data frame.

DATA CLEANING

As the data was scraped from a site their were lots of inconsistency in the data, steps were taken to clean it.

- Arrival Time and Departure Time were mapped into bins/buckets.
- Duration Column had inconsistency in its values which were dealt with and then the duration was converted into hours.
- Two airline companies had very low number of instances in the data (less than 0.01% , so they were removed.
- Consistency of Stop column was maintained.
- Commas were removed from Price

DATA CLEANING

- Feature Extraction :
  - Column Days Left was extracted from Date using SQL query.
  - Character Code and Numerical Code were combined to form Flight Code.
- Feature Selection:
  - Date and Flight Code column were deleted (they were used for EDA and was deleted afterward).
- There were very few null values in our dataset which were dropped.
- 2 duplicate rows were also dropped.
- Data Type of all the column were aligned with its type.

EXPLORATORY DATA ANALYSIS

- For Plotting Pandas, Seaborn and Matplotlib were used.
- Spark data frame was transformed into pandas data frame.

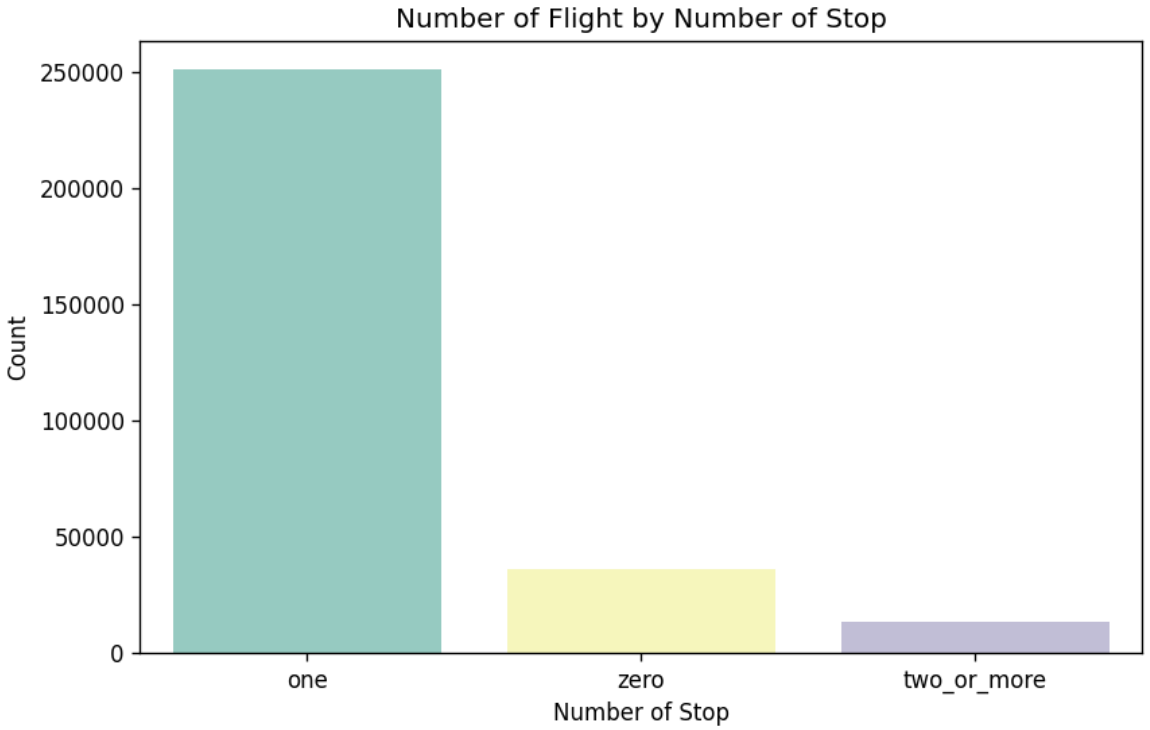
	duration	days_left	price
count	300153.00	300153.00	300153.00
mean	12.22	26.00	20889.66
std	7.19	13.56	22697.77
min	0.83	1.00	1105.00
25%	6.83	15.00	4783.00
50%	11.25	26.00	7425.00
75%	16.17	38.00	42521.00
max	49.83	49.00	123071.00

1. 50% of the flight takes less than 11hrs 15 min, mean duration is 12hrs 12min which is very close to median, so the duration is slightly right skewed.

2. 50% of the flight ticket costs less than Rs7500, mean price is Rs 20889 which is very large compared to median price, so the price is highly right skewed.

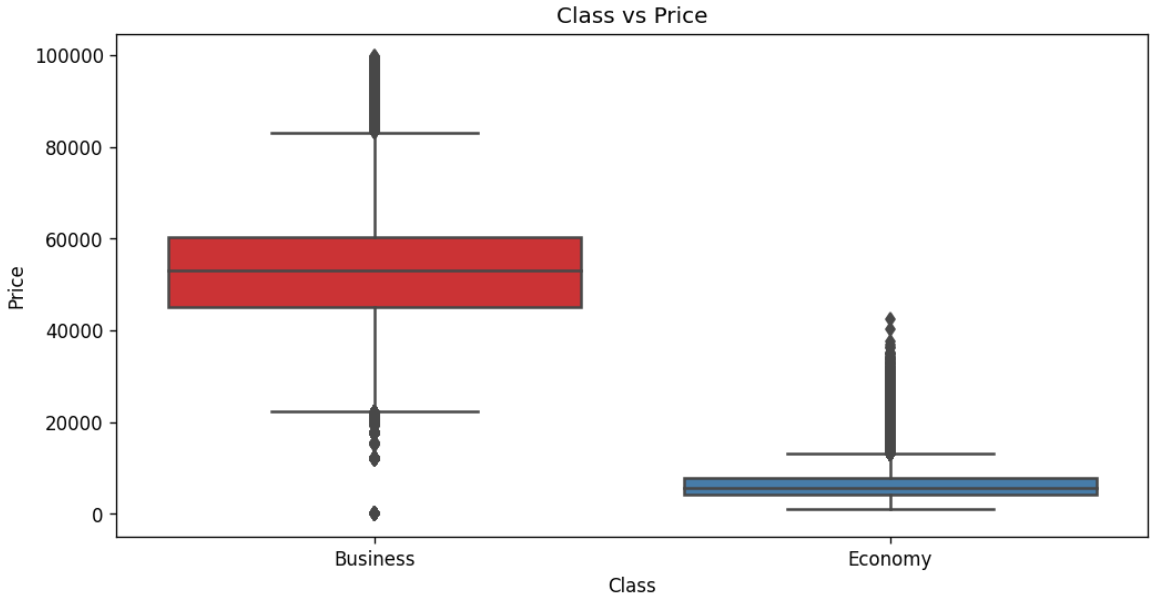
GRAPHICAL ANALYSIS

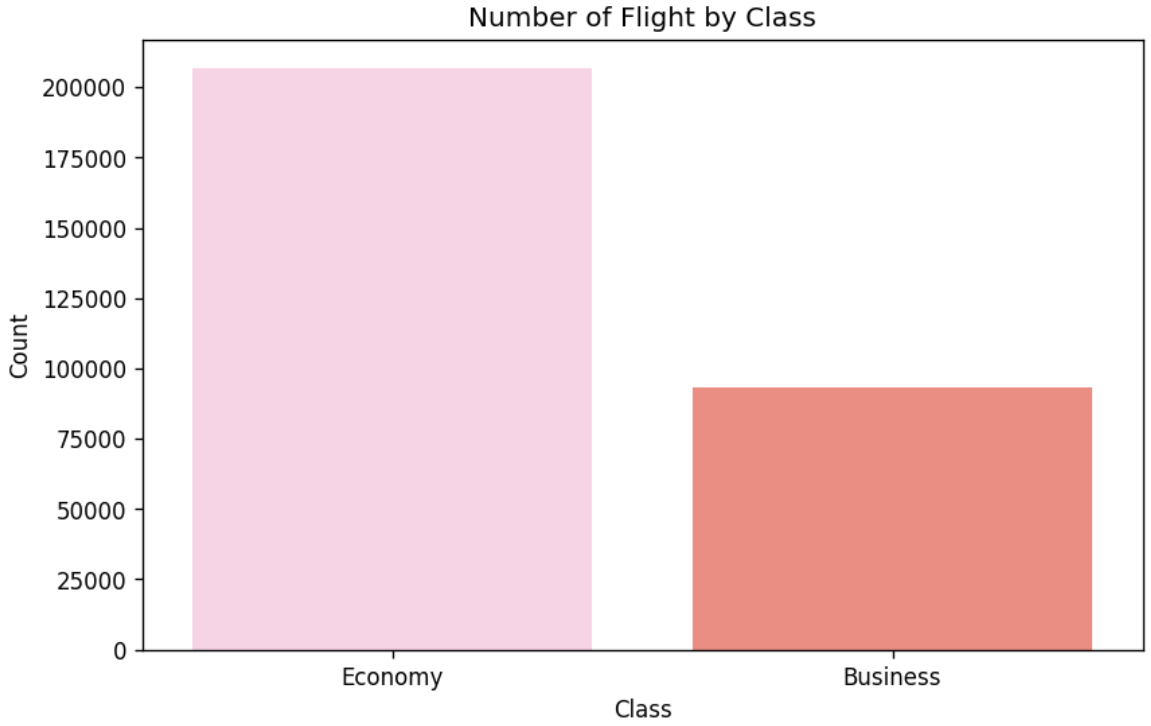
- Between the origin and destination city, most flights make at least one stop.
- The bar plot shows that the mean ticket price of flights with stops is higher than the mean ticket price of flights without any stop.



GRAPHICAL ANALYSIS

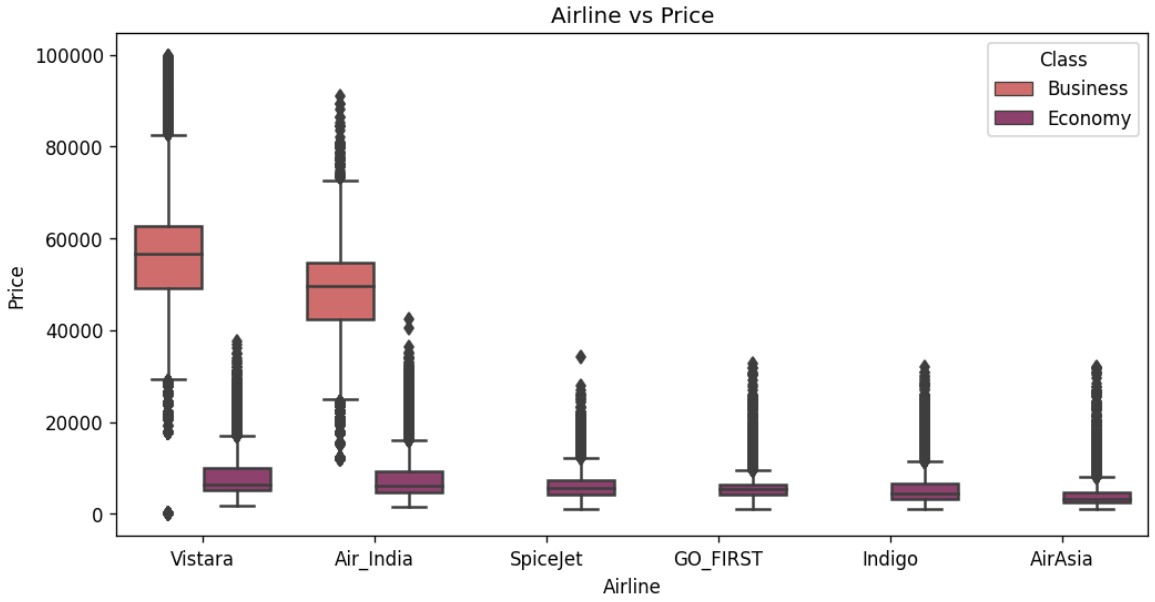
- In India, majority of travel is done in Economy Class.
- Mean ticket price of business class is higher than the mean ticket price of economy class tickets.

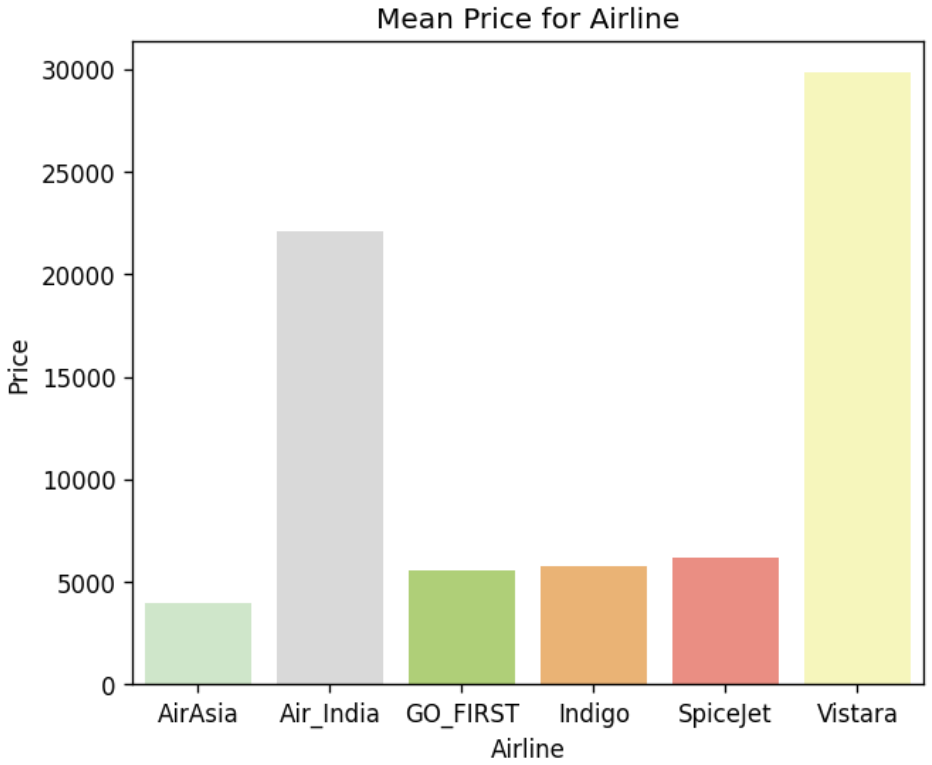




GRAPHICAL ANALYSIS

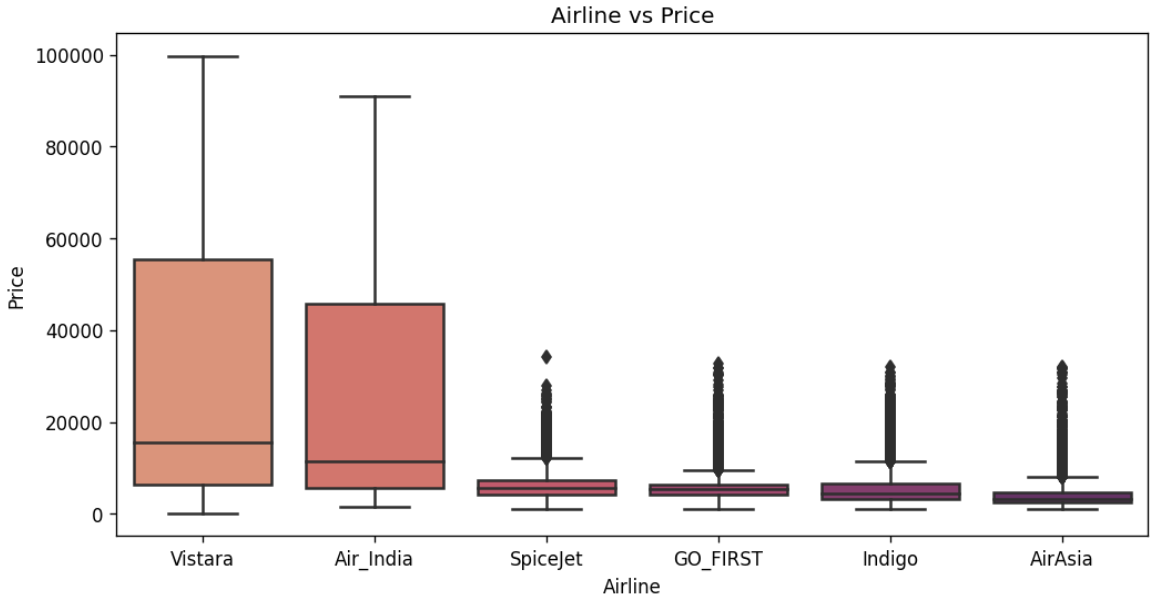
- From the graph, we observe that Vistara and Air India have the highest median prices. Apart from them, practically every airline has comparable median price. Vistara and Air India offers wide range of options in terms of price.
- Higher Price of Vistara and Air India is attributed to the fact they are the airline company that offers Business class services in India.

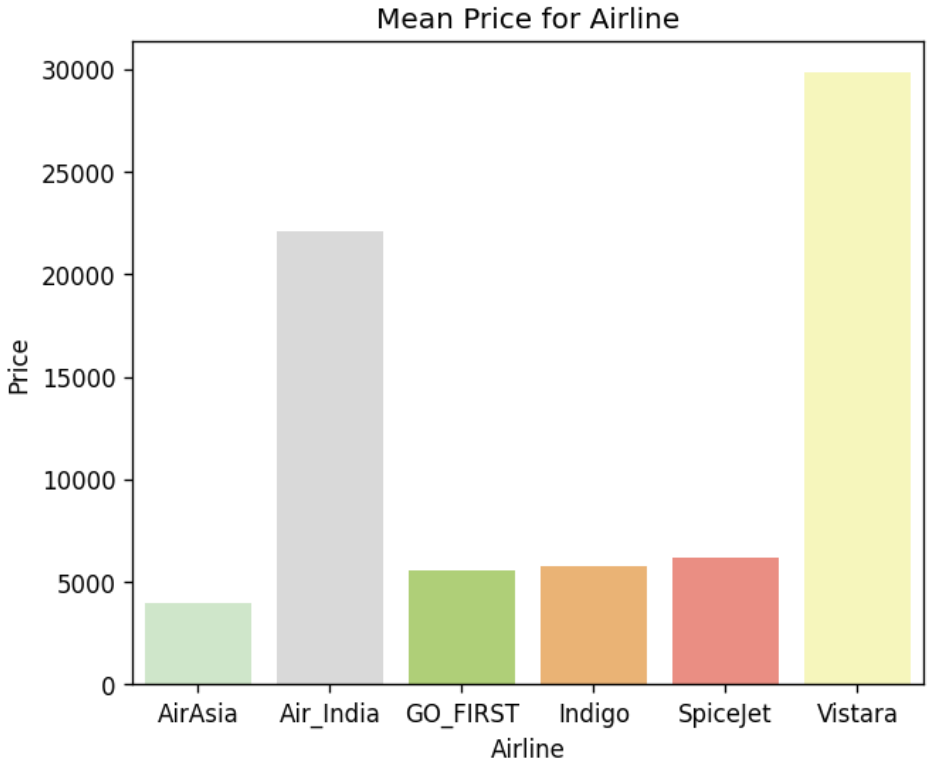




GRAPHICAL ANALYSIS

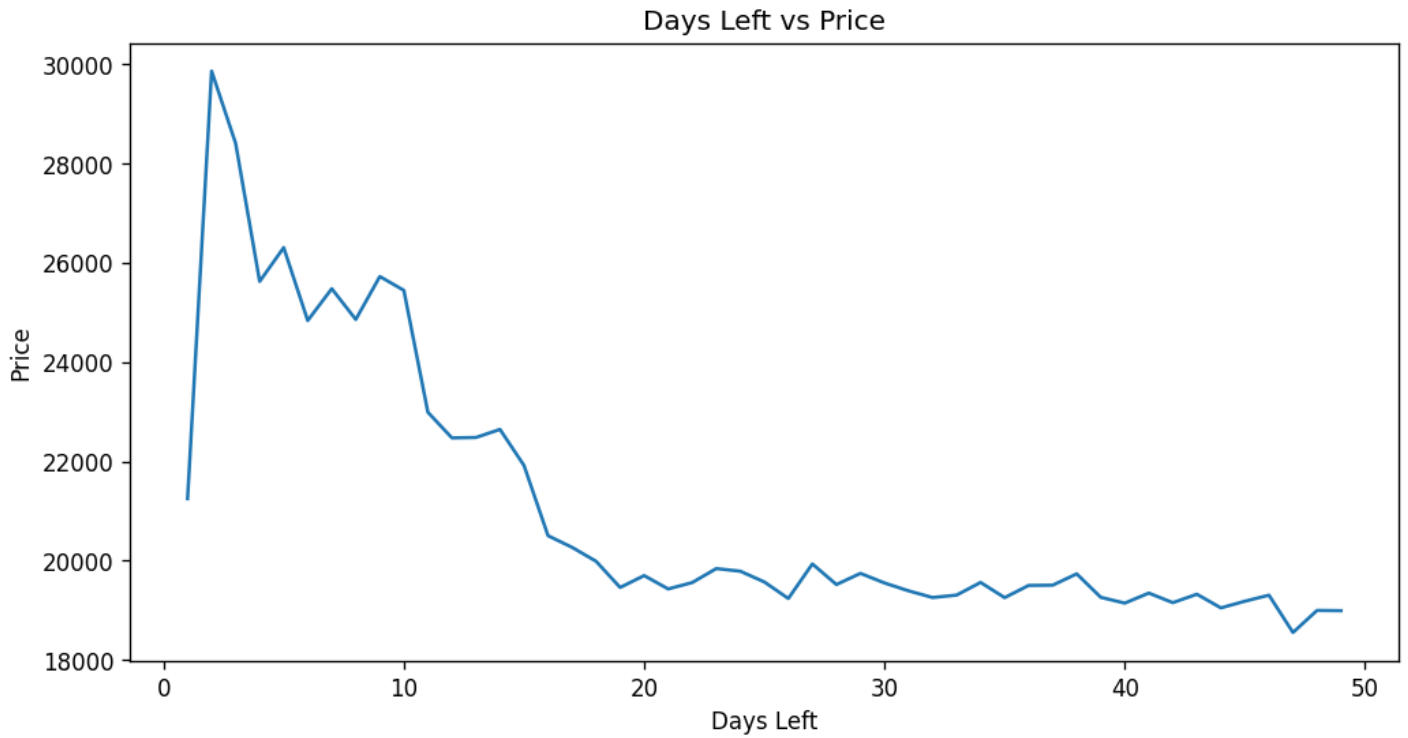
- From the graph, we observe that Vistara and Air India have the highest median prices. Apart from them, practically every airline has comparable median price. Vistara and Air India offers wide range of options in terms of price.
- Higher Price of Vistara and Air India is attributed to the fact they are the airline company that offers Business class services in India.





GRAPHICAL ANALYSIS

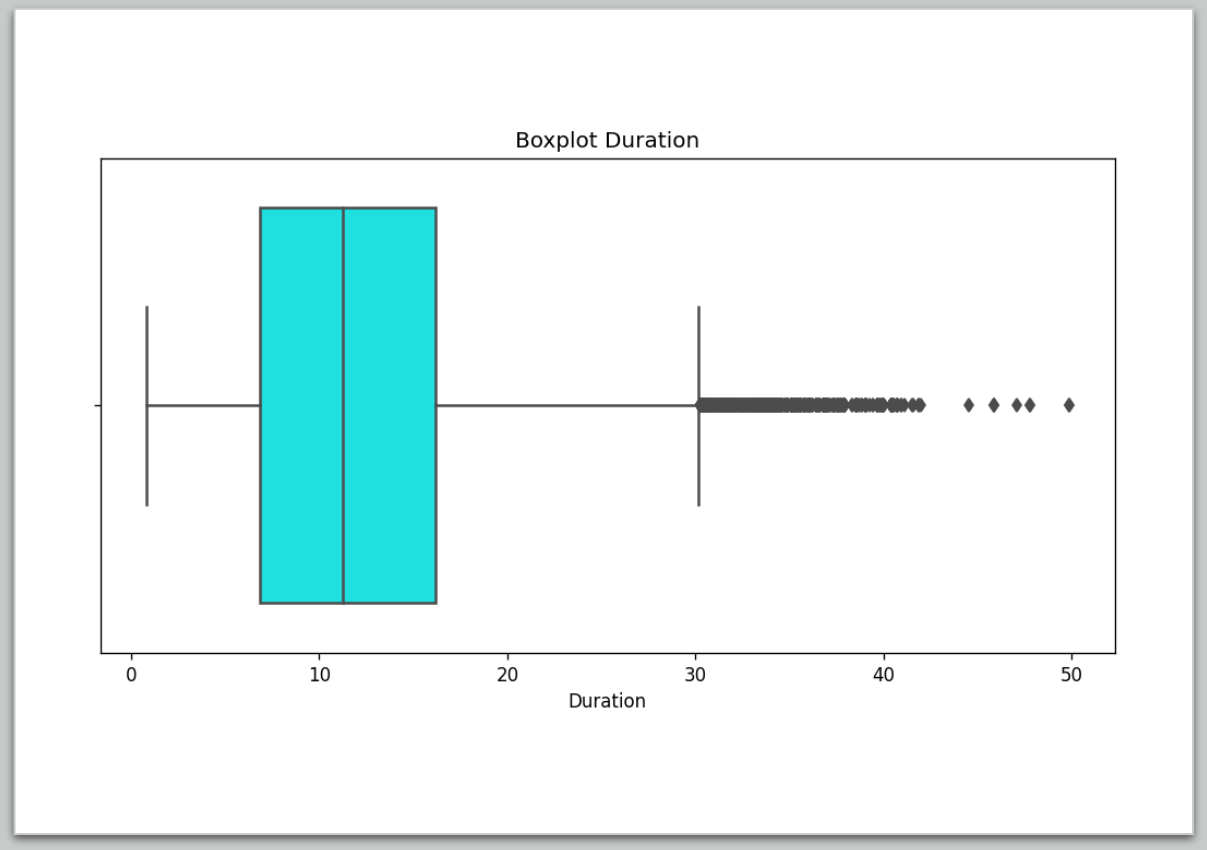
- The cost of a flight varies greatly depending on how early you book it.
- However, after a certain amount of time, it reaches saturation.

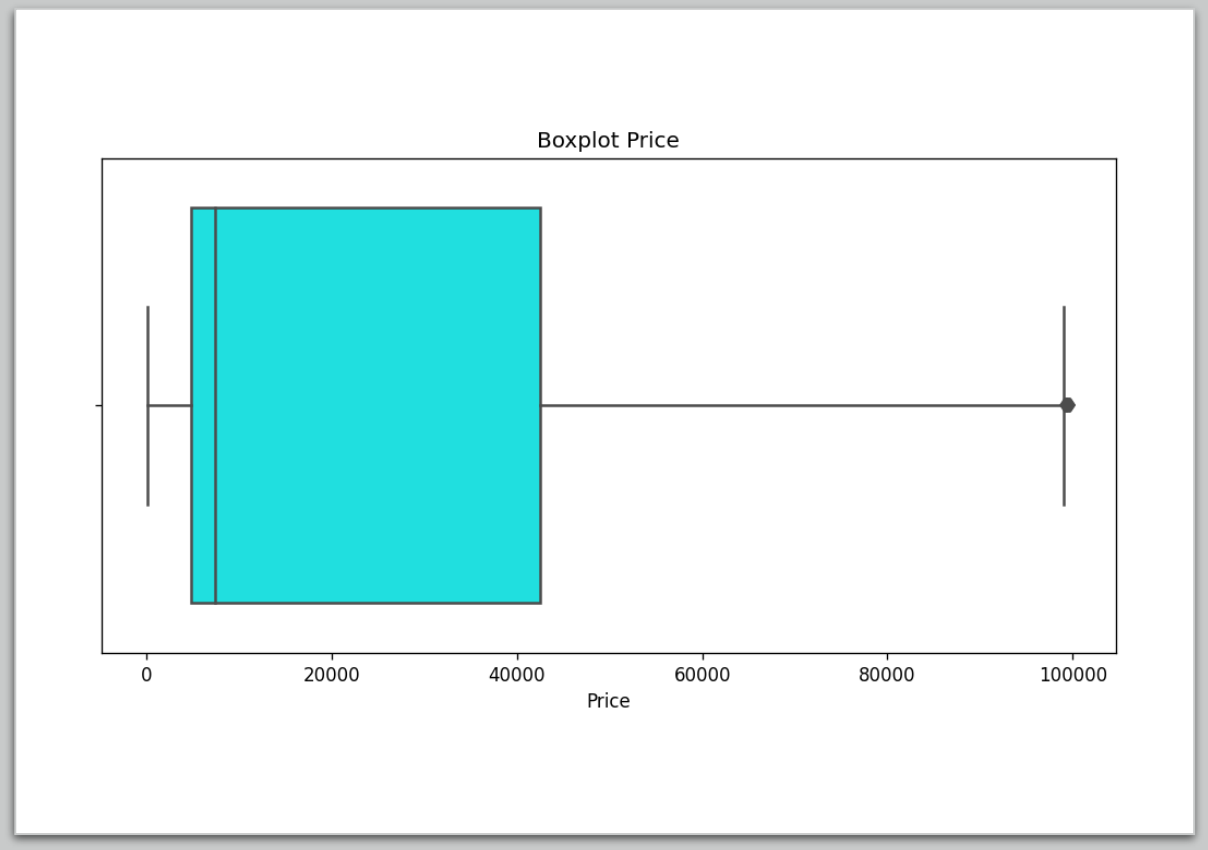


OUTLIER TREATMENT

Price and Duration were found to be highly right skewed; Columns were transformed logarithmically.

Outliers having z scores above 3 or less than 3 were dropped.





DATA PREPROCESSING

- Dummy Variable Encoding:
  - All the categorical columns were encoded as dummies.
  - To avoid falling into dummy trap, one column from each feature were dropped.
- Scaling (Standardisation):
  - Numerical Column Days Remaining and Duration were scaled.
  - Min Max Scaler was also tried but Standardisation gave better result.
- Vector Assembler:
  - All the independent columns were combined into a single vector column (Independent) using Vector Assembler.
- Train Test:
  - Data was randomly split into train (75%) and test (25%).

MODEL BUILDING

LINEAR REGRESSION

- Linear Regression is a supervised machine learning algorithmic which assumes that there exists a linear relationship between the dependent variables and the predictors.
- Linear Regression model was fit into the training set.
- Prediction was made on test set and the result was evaluated.
- Mean Absolute Error : 0.255 i.e. mean difference between the actual and the predicted log price is 0.255.
- Root Mean Squared Error was found out to 0.31.
- R Square : 0.906, telling us that 90.6% variation in price is explained all the independent variables combined.
- Regression Output was found and detailed explanation was posted in google notebook.

Regression Output

Inference Drawn,

1. P-Value of F-statistics is significant saying that the at least one beta coefficient of the independent feature is nonzero and that we may go ahead to analyze the regression output.

2. R-Squared value is 0.904, saying that about 90.4% variance in target variable is explained by all the independent variables.

3. Adjusted R-Squared value is same as R-Squared, telling us that all the independent variables is adding something to explain the target variable. But this inference may not be significant as the dataset has about 3 lakh datapoints.

4. All the variable coefficient is significant as the p-Value is less than 0.05, except departure time morning and early morning

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.904
Model:	OLS	Adj. R-squared:	0.904
Method:	Least Squares	F-statistic:	9.395e+04
Date:	Sat, 05 Mar 2022	Prob (F-statistic):	0.00
Time:	10:14:37	Log-Likelihood:	-1.0739e+05
No. Observations:	300153	AIC:	2.148e+05
Df Residuals:	300122	BIC:	2.152e+05
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.9513	0.013	843.141	0.000	10.926	10.977
Airline_Air_India	0.0514	0.004	13.019	0.000	0.044	0.059
Airline_Indigo	-0.1242	0.004	-29.964	0.000	-0.132	-0.116
Airline_AirAsia	-0.4575	0.005	-98.117	0.000	-0.467	-0.448
Airline_Vistara	0.1736	0.004	44.422	0.000	0.166	0.181
Airline_GO_FIRST	-0.0464	0.004	-10.617	0.000	-0.055	-0.038
Class_Economy	-2.0183	0.002	-1309.064	0.000	-2.021	-2.015
Source_City_Bangalore	0.0869	0.002	36.986	0.000	0.082	0.092
Source_City_Chennai	0.0356	0.002	14.279	0.000	0.031	0.041
Source_City_Mumbai	0.0401	0.002	17.568	0.000	0.036	0.045
Source_City_Kolkata	0.2181	0.002	90.533	0.000	0.213	0.223
Source_City_Delhi	0.0560	0.002	24.538	0.000	0.051	0.060
Departure_Time_Evening	-0.0417	0.010	-4.077	0.000	-0.062	-0.022
Departure_Time_Morning	0.0070	0.010	0.685	0.494	-0.013	0.027
Departure_Time_Early_Morning	-0.0107	0.010	-1.048	0.295	-0.031	0.009
Departure_Time_Afternoon	-0.0209	0.010	-2.038	0.042	-0.041	-0.001
Departure_Time_Night	-0.0442	0.010	-4.309	0.000	-0.064	-0.024
Arrival_Time_Evening	0.0230	0.003	6.810	0.000	0.016	0.030
Arrival_Time_Morning	-0.0347	0.003	-10.176	0.000	-0.041	-0.028
Arrival_Time_Afternoon	-0.0169	0.004	-4.730	0.000	-0.024	-0.010
Arrival_Time_Early_Morning	-0.0862	0.004	-20.111	0.000	-0.095	-0.078
Arrival_Time_Night	0.0150	0.003	4.545	0.000	0.009	0.021
Destination_City_Bangalore	0.0984	0.002	42.200	0.000	0.094	0.103
Destination_City_Chennai	0.0521	0.002	21.276	0.000	0.047	0.057
Destination_City_Mumbai	0.0749	0.002	33.019	0.000	0.070	0.079
Destination_City_Kolkata	0.2001	0.002	85.018	0.000	0.195	0.205
Destination_City_Delhi	0.0757	0.002	32.796	0.000	0.071	0.080
Stop_one	-0.2055	0.003	-64.644	0.000	-0.212	-0.199
Stop_zero	-0.5529	0.005	-119.210	0.000	-0.562	-0.544
Days_Remaining	-0.1919	0.001	-303.048	0.000	-0.193	-0.191
Duration	0.0628	0.001	53.045	0.000	0.060	0.065

Omnibus:	200053.259	Durbin-Watson:	1.470
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28836122.783
Skew:	-2.278	Prob(JB):	0.00
Kurtosis:	50.801	Cond. No.	166.

DECISION TREE REGRESSOR

- Decision Tree Regressor builds regression in the form of a tree structure where the nodes are split several times based on features to give result in the leaf/terminal node.
- Decision Tree Regression model was fit into the training set.
- Prediction was made on test set and the result was evaluated.
- Root Mean Squared Error was found out to 0.29.
- Results obtained were better than Linear Regression.
- Result could be further refined by pruning the tree and tuning hyperparameters.

GRADIENT BOOSTING

- Gradient Boosting is an ensemble technique that combines several biased weak learner sequentially to form unbiased strong learner. Decision Tree with 8 to 32 leaf are used as base learner in GBTRegressor model.
- GBTRegressor model was fit into the training set.
- Prediction was made on test set and the result was evaluated.
- Root Mean Squared Error was found out to 0.26.
- Results obtained were better than Linear Regression and Decision Tree.
- GBTRegressor result can be further improved by tuning its hyperparameter.

RANDOM FOREST REGRESSOR

- Random Forest is also an ensemble learning technique consisting of many decision tree, which uses bagging and feature randomness when building individual tree to make them independent from each other reducing variance error.
- Random Forest Regressor model was fit into the training set.
- Prediction was made on test set and the result was evaluated.
- Root Mean Squared Error was found out to 0.31.
- Results obtained were worst out of all, this is because random forest works best when the number of trees is huge and the trees generated have low bias (full grown or higher depth) but by default pyspark random forest generates only 10 trees.

HYPERTUNING PARAMETERS

- To improve the result of Random Forest Regressor the parameters were tuned.
- Since the dataset size is huge it causes the server to overload and spark context shuts down. So a sample of the dataset was taken.
- Random Forest Regressor model was fit into the training set.
- Pipeline was defined, and the Parameter Grid was build.
- Parameters tuned were number of trees, bootstrap and maximum depth.
- 3 – Fold Cross Validation was used to tune parameters with the help of RMSE score.
- Root Mean Squared Error of the tuned model was found out to 0.23.
- Tuned Random Forest Regressor gave the best result.

CONCLUSION

1. Ticket price depends on various parameters.
2. Tickets should be bought well in advance to get better deal. Tickets purchased 1-2 days before departure date costs much more than the tickets bought more than 30 days.
3. Ticket prices vary depending on departure time. Flights departing late at night are less expensive.
4. Tickets Price costs more if flight originates from or lands in Kolkata. If Hyderabad is the source or destination city, then the ticket price is the lowest.
5. Business class tickets are much more expensive than Economy class tickets.
6. Ticket Price costs less if there is no stop between the source and destination city.

- LINK TO OBJECTIVE
- LINK TO RESEARCH QUESTION

Built a website that predicts flight price.
PLEASE VISIT : https://predict-flights-price.herokuapp.com/