Project Report on

# Flight Fare Prediction

Prepared By

Rajan Patel

## logo for flight fare prediction

**Importing Data:**

### Importing data is a fundamental task in data analysis and can be performed in various ways depending on the source and format of the data. We have used the method of importing the data from csv files and excel files

### CSV files: Comma-separated values files are one of the most common formats. They can be imported using libraries like pandas in Python (pd.read\_csv()).

### Excel files: Microsoft Excel files can be imported using libraries like pandas (pd.read\_excel()).

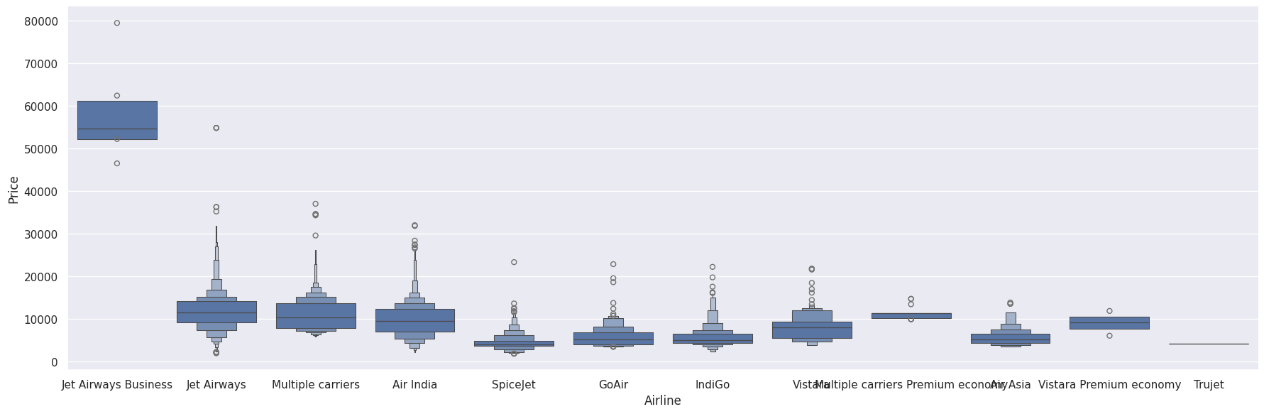
### Text files: Plain text files with structured or semi-structured data can be read using basic file handling techniques or libraries like csv in Python.

## Data Preprocessing:

Initially, all features except for Price were categorized as objects and required processing. Categorical attributes like Airline, Source, Destination, Total\_Stops, and Additional\_Info were identified, while others initially treated as objects held continuous values. The Route feature was redundant as Total\_Stops already conveyed this information, prompting its removal. Additionally, Date of Journey, Departure Time, and Arrival Time, originally stored as objects, were converted to datetime for clarity. Date of Journey was further divided into Journey day and Journey month, and Departure Time and Arrival Time were segmented into Hours and Minutes. Post conversion, Date of Journey, Departure Time, and Arrival Time were eliminated from the dataset.

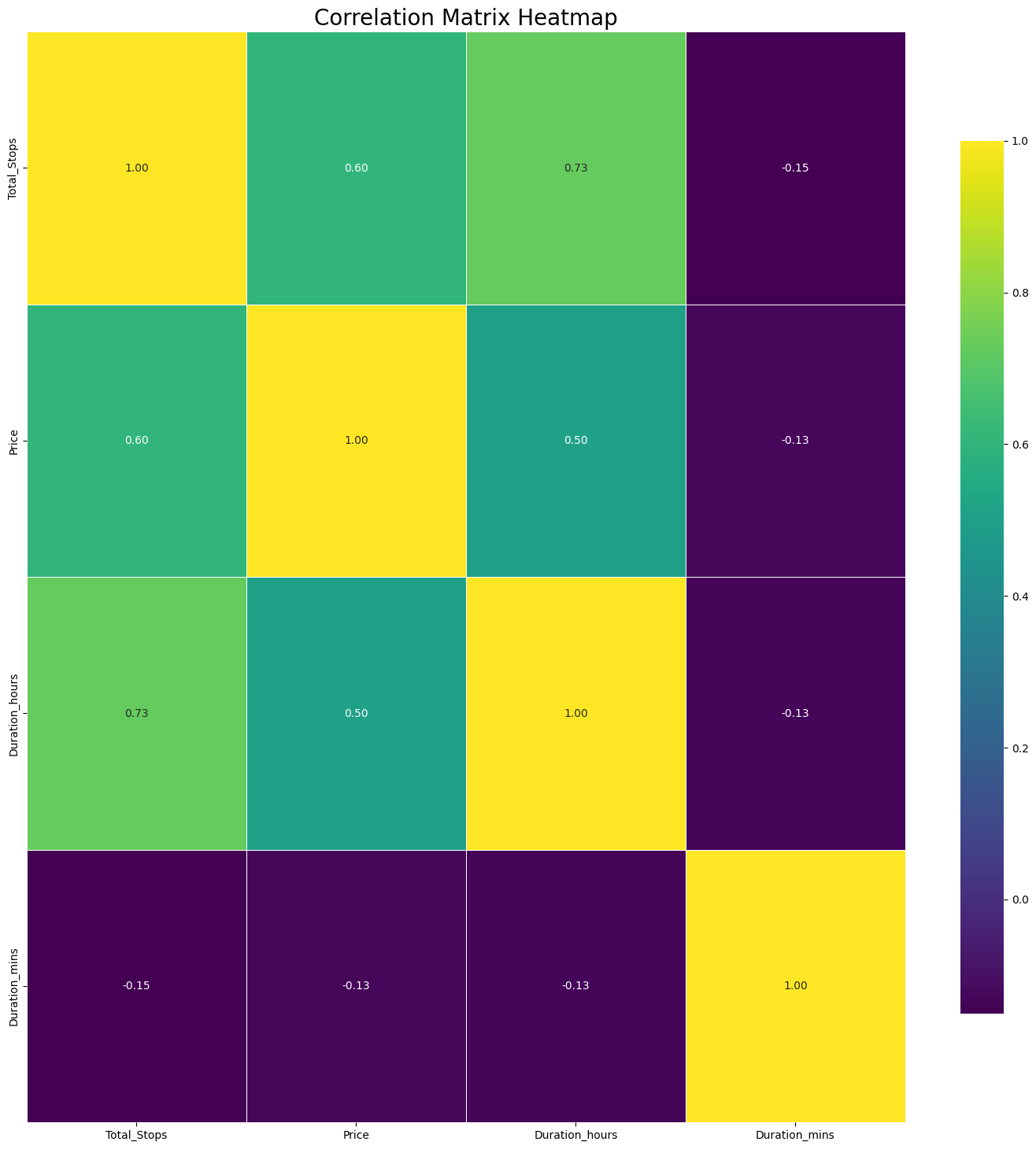
## Exploratory Data Analysis

### **violin plot**



The box plot indicates that flights originating from Bangalore and Delhi tend to have higher median prices and greater price variability, evidenced by numerous high outliers. In contrast, flights from Kolkata, Mumbai, and Chennai exhibit lower and more uniform price distributions. Chennai stands out with the most tightly clustered price range and the lowest median flight prices among the cities compared.

**Heatmap:**

****

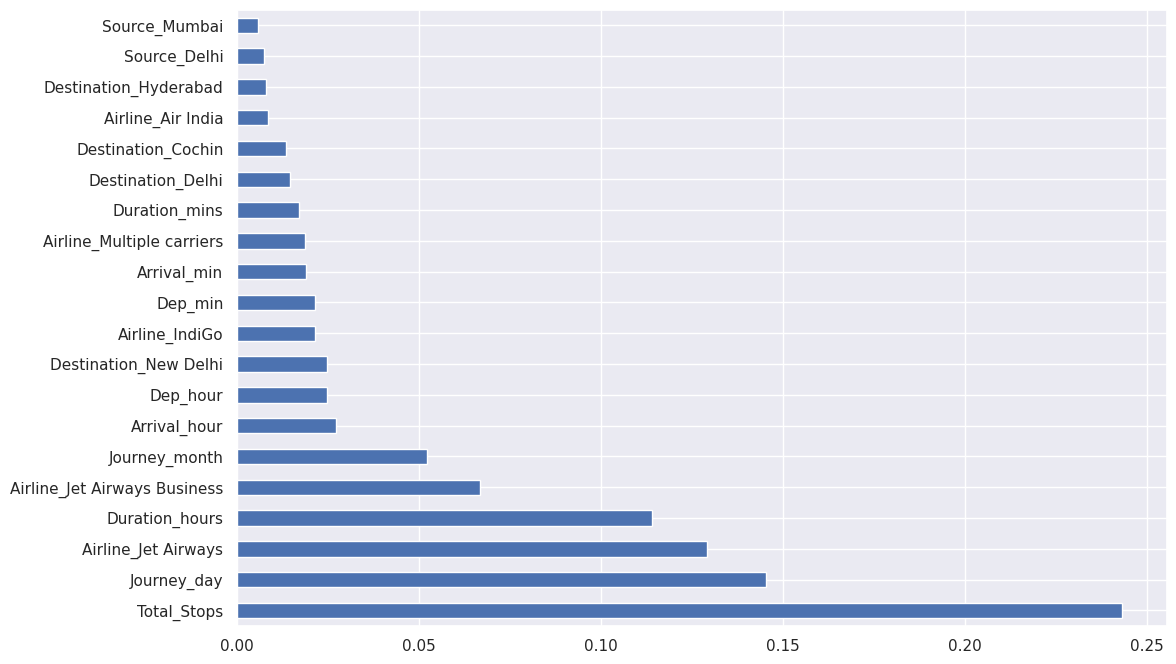
The correlation coefficient, ranging from -1 to 1, indicates the degree of relationship between two variables. A coefficient of 1 signifies a perfect positive correlation, where as one variable increases, so does the other. A coefficient of -1 indicates a perfect negative correlation, where as one variable increases, the other decreases. A coefficient of 0 implies no correlation between the variables.

For instance, if the correlation coefficient between Total Stops and Price is 0.6, it indicates a positive correlation between them, but not a perfect one. This means more expensive stores generally tend to have more stops, although there are exceptions with expensive stores having fewer stops and cheaper stores having more stops.

**Model Building:**

This problem is of regression which can be solved by a Regression model. Models such as Random Forest Regressor, Decision Tree Regressor, and Linear Regression are used to build and predict airplane ticket prices.

First of all, Random Forest Regressor is built with every feature to compute the feature importance. Based on feature importance the features that have more importance are utilized to build Linear Regression and Decision Tree Regressor models.



Important features:

Airline\_multiple carriers premium economy, Source\_mumbai, Destination\_hyderabad, Airline\_spicejet, Additional\_Info\_business class, Source\_banglore, Source\_delhi, Airline\_vistara, Destination\_cochin, Destination\_new delhi, Airline\_indigo, Airline\_air india, Additional\_Info\_no info, Arrival\_min, Airline\_multiple carriers, Departure\_min, Departure\_hour, Arrival\_hour, Total\_Stops, Additional\_Info\_in-flight meal not included, Airline\_jet airways business, Journey\_month, Airline\_jet airways, Journey\_day, Duration.

## Model Evaluation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | MAE | MSE | RMSE | R2 SCORE |
| Linear Regression | 1998.899686 | 8.678835e+06 | 2945.986260 | 0.583754 |
| Random Forest | 1183.562687 | 3.942891e+06 | 1985.671459 | 0.810895 |
| XG Boost | 1117.677118 | 3.255040e+06 | 1804.173067 | 0.843885 |

Random Forest has the least MSE, and MAE among all the models which is followed by Decision Tree. Also, the accuracy of Random Forest is 89% and 87% for Decision Tree.

However, the Linear Regression model has an accuracy score of only 67% also it has high MSE and MAE in comparison to other models.

**Conclusion**

This project effectively demonstrates that machine learning techniques can predict flight ticket prices with a high level of accuracy. By meticulously preprocessing data, engineering features, and applying various regression models, we identified significant factors influencing flight costs, such as airline choice, travel dates, departure city, destination, and number of stops. Among the models tested, the XGBoost Regressor performed exceptionally well, highlighting its ability to capture intricate data relationships.

This analysis highlights the potential of data science in addressing seemingly unpredictable airline pricing strategies, providing valuable insights for both travelers and airlines. Future improvements could involve integrating additional data sources, exploring alternative models and optimizing their parameters, and developing a user-friendly interface for enhanced user interaction.