FLIGHT PRICE PREDICTION ANALYSIS

By: LAKSHITA KAIN

DATA SCIENTIST(INTERN)

FLIPROBO TECHNOLOGIES

Introduction

In India, there are over 400 airports and airstrips, while 153 were operational. Passenger traffic amounted to over 115 million at airports across India in financial year 2021, out of which over 10 million were international passengers. Airline companies use various algorithms to predict flight prices on the basis of dynamically changing financial,marketing and social aspects.

Anyone who've booked an airplane ticket online knows that prices on a particular ticket is always constantly varying. Main objective is to analyse and build a dynamic machine learning model that can predict the flight prices on the basis of information of the flight provided by the airlines.

PROBLEM STATEMENT

Main objective is to analyse and build a dynamic machine learning model that can predict the flight prices on the basis of information of the flight provided by the airlines.

DATA COLLECTION

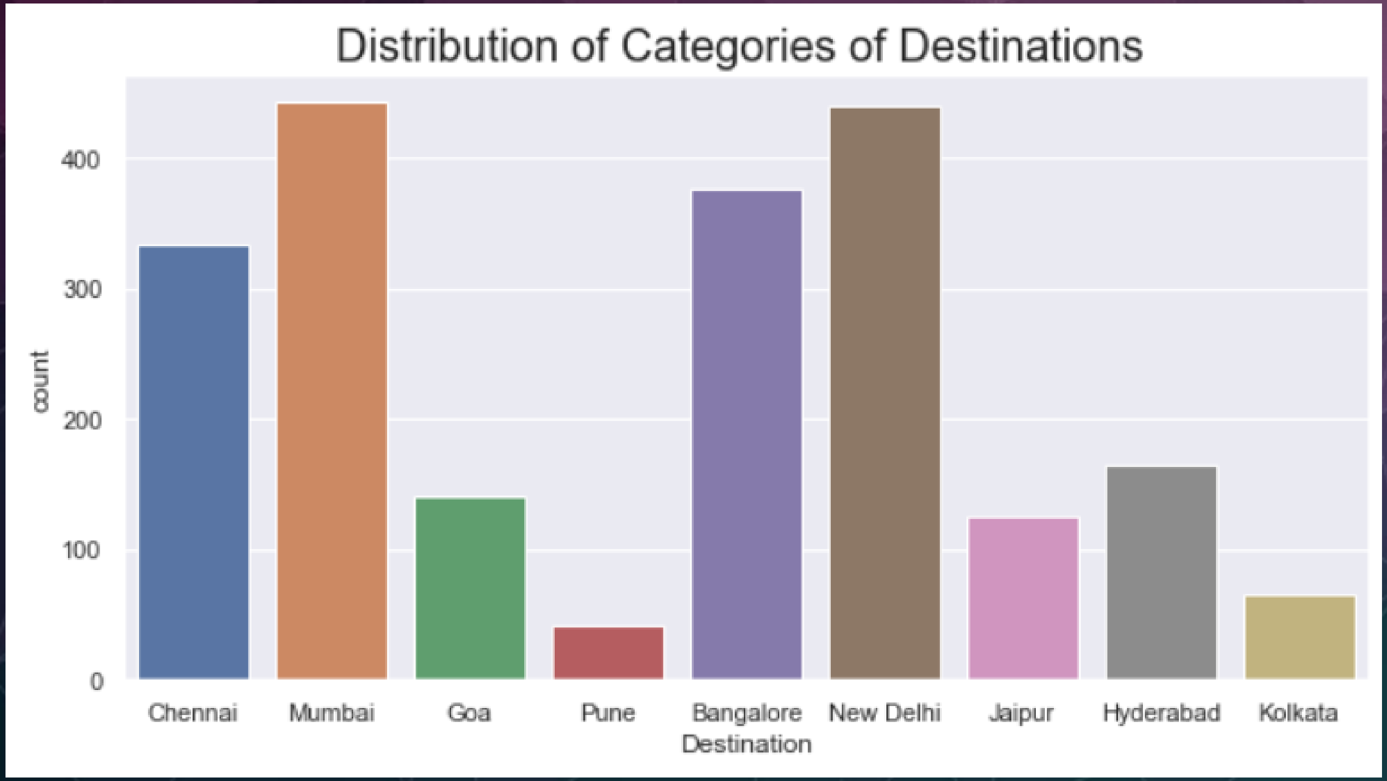
Data has been scrapped using Selenium webdriver with python. Selenium is a powerful tool for controlling web browsers through programs and performing browser automation.

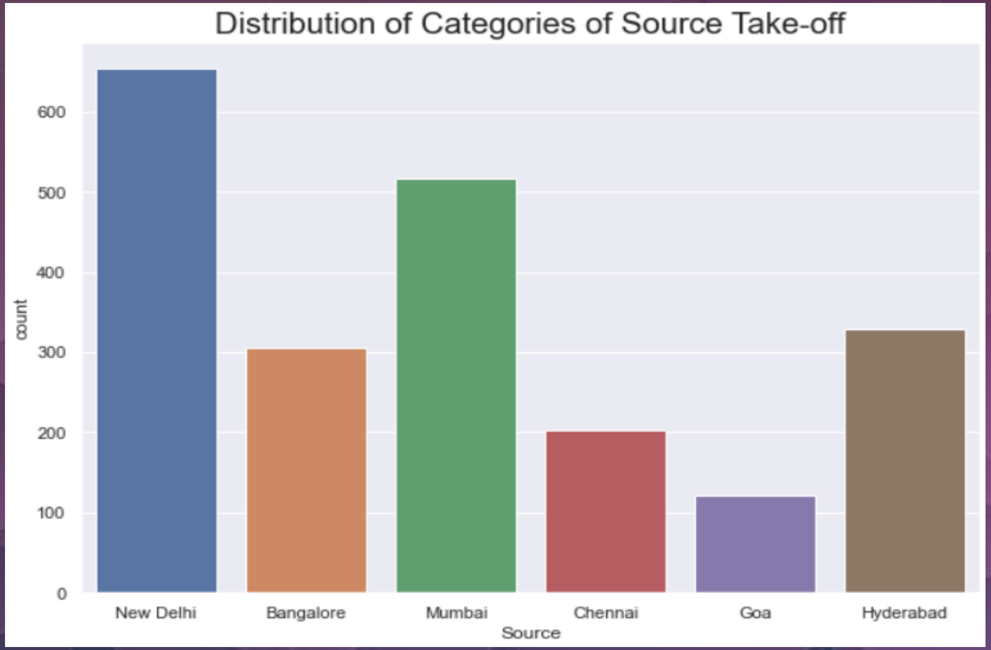
Our data has been retrieved from https://www.yatra.com which consists of Airlines Flight information over the span of 16 days (from 01-12-2021 to 16-12-2021) such as Airlines Name, Date of Departure, Time of Departure, Time of Arrival, Duration of Flight, Number of Stops, Meals included and Price of the ticket.

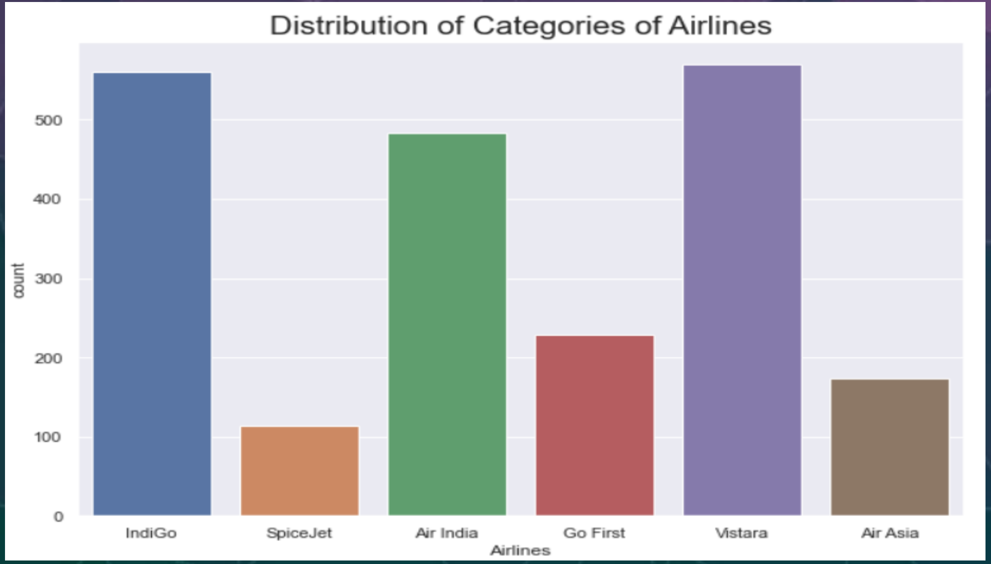
EXPLORATORY DATA ANALYSIS

- Shape of our dataset : records - 2250, features - 10
- There are no null values
- There are 119 duplicated fields or attributes.
- "Unnamed : 0" is an necessary column for our analysis therefore we drop it.
- There are 10 object datatype columns.
- "Duration", "Price" have object datatype which needs to be changed for further analysis.

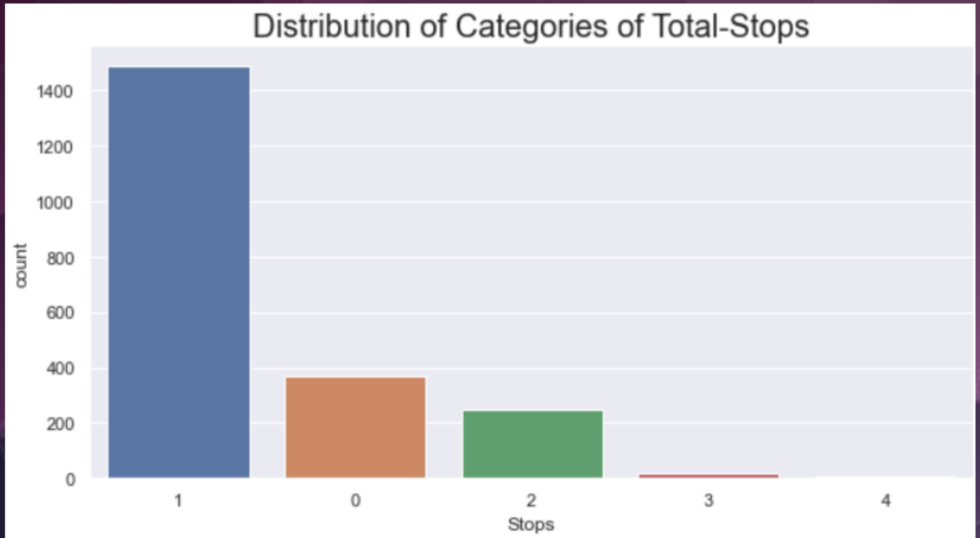
UNIVARIATE ANALYSIS

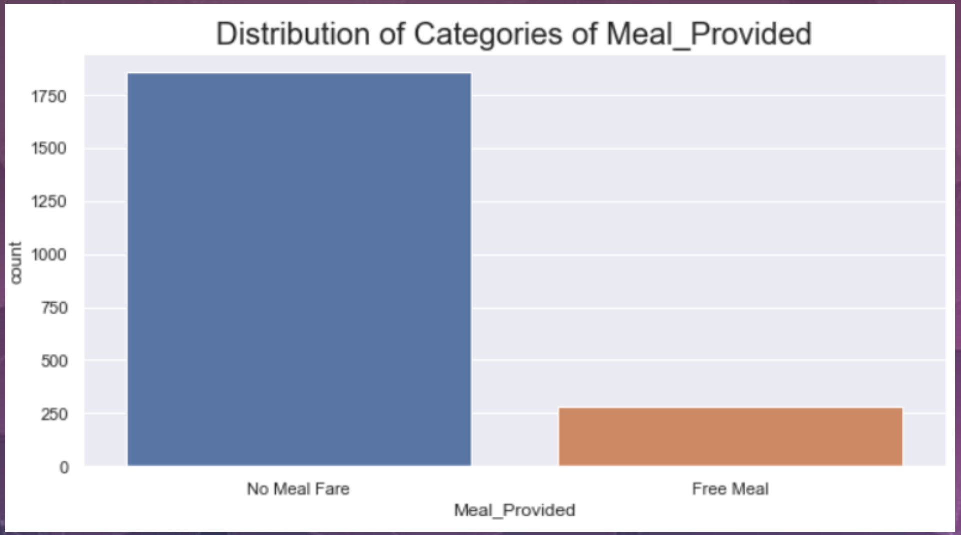


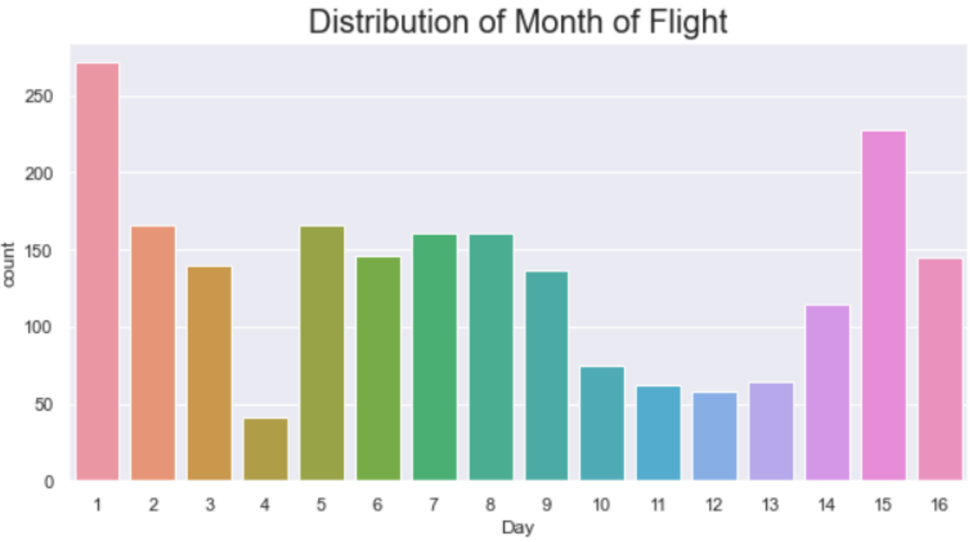


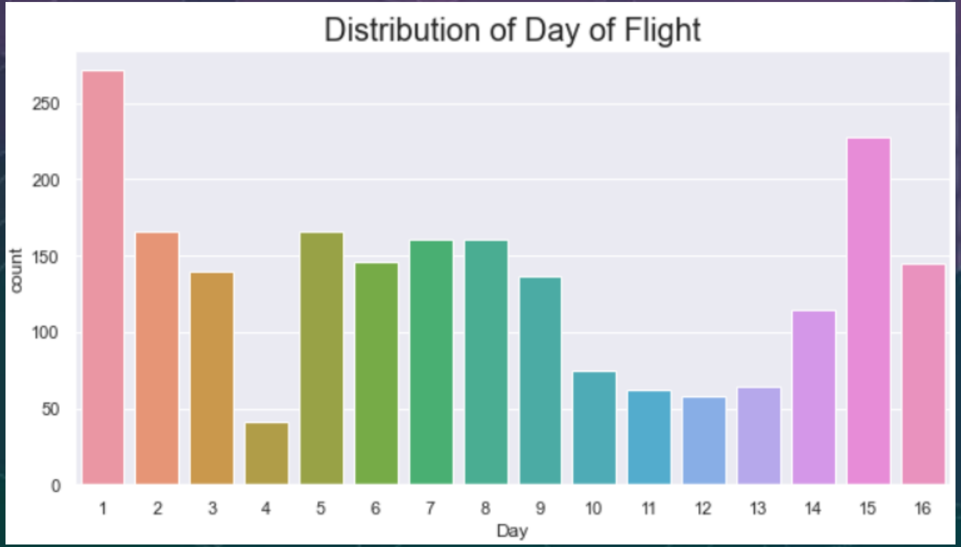


UNIVARIATE ANALYSIS

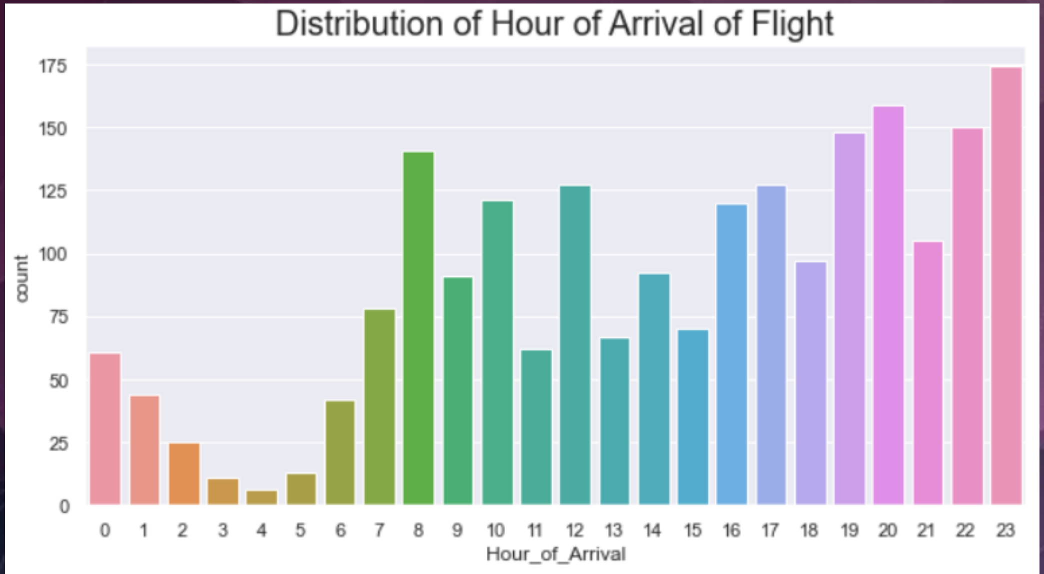


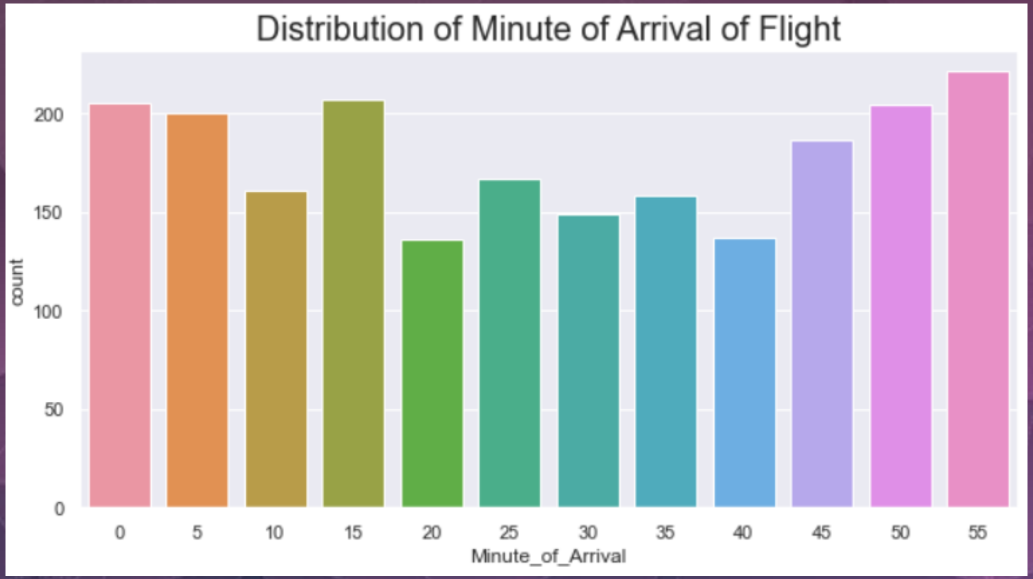


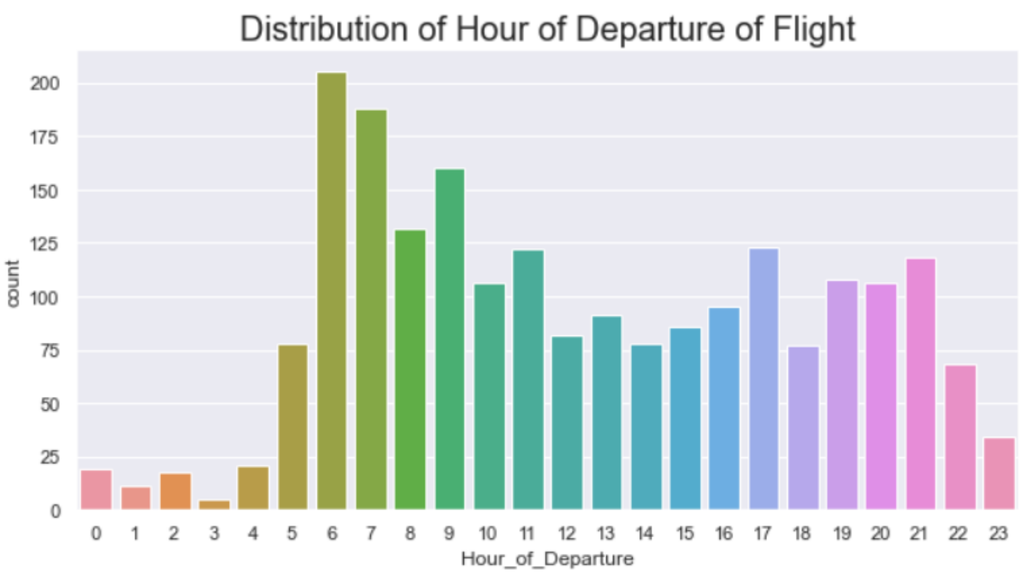


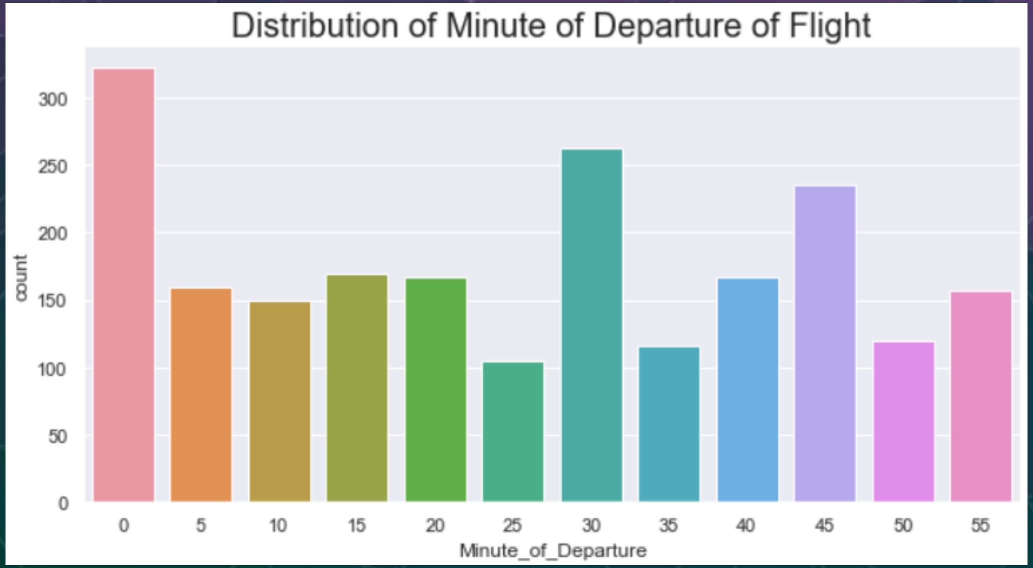


UNIVARIATE ANALYSIS

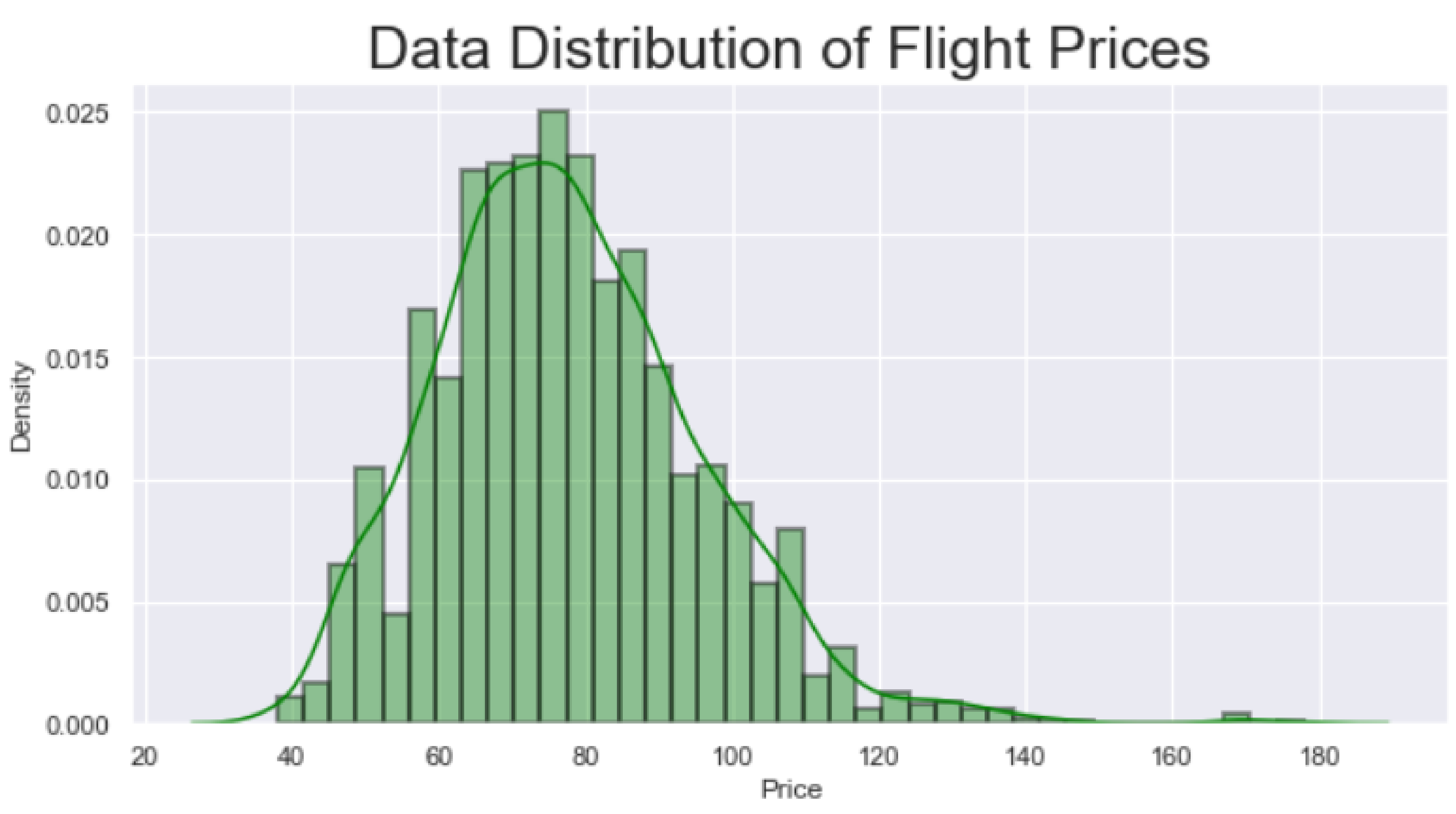








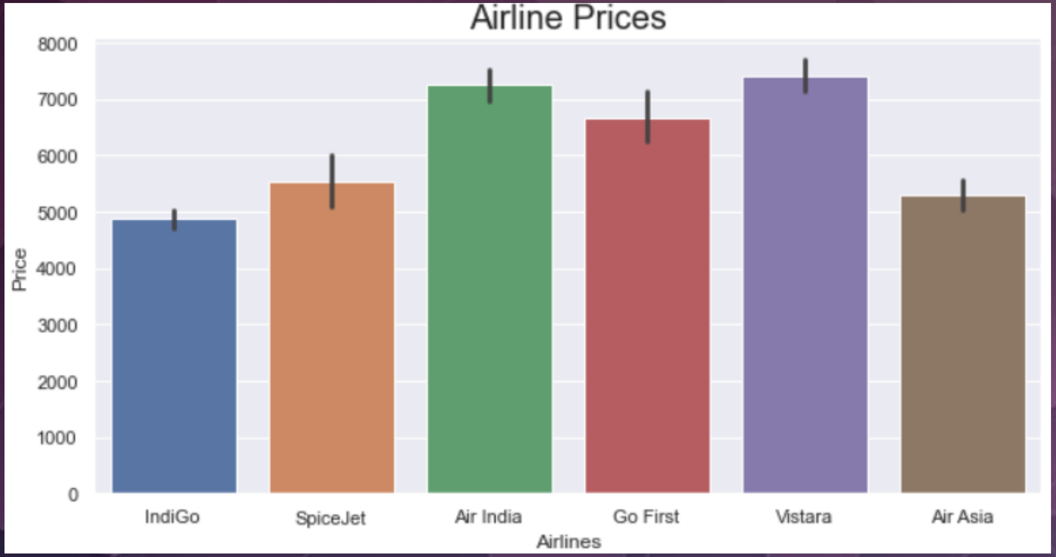
UNIVARIATE ANALYSIS



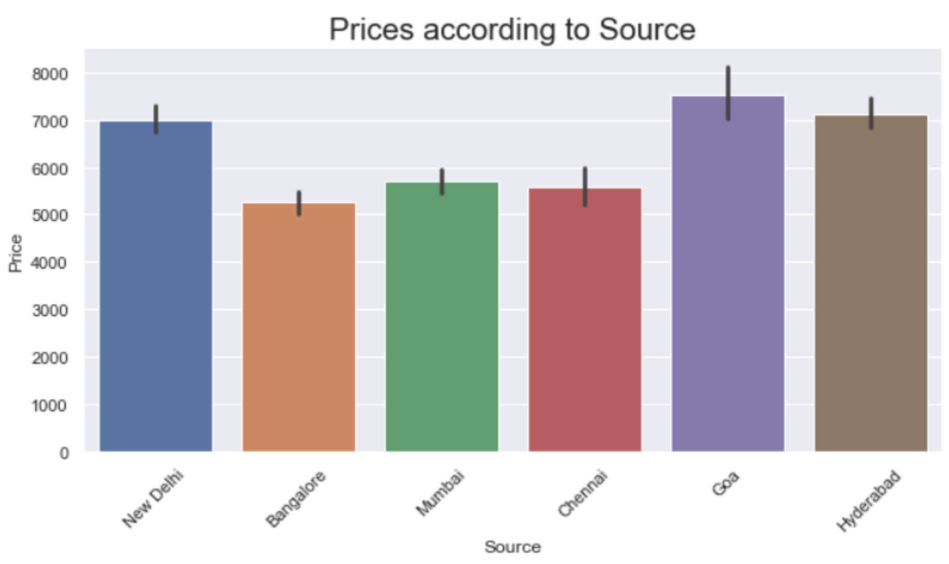
INSIGHTS

- Indigo and Vistara offers more number of flights.
- New Delhi and Mumbai are top destinations searched.
- Most flights are took off from New Delhi
- Majority of Flights have atleast 1 stop in between.
- More Flight tickets don't consists of free meals.
- Most Flights are from 1st of December.
- Most Flights scheduled to take off early morning and land by late at night.
- Price is rightly skewed.

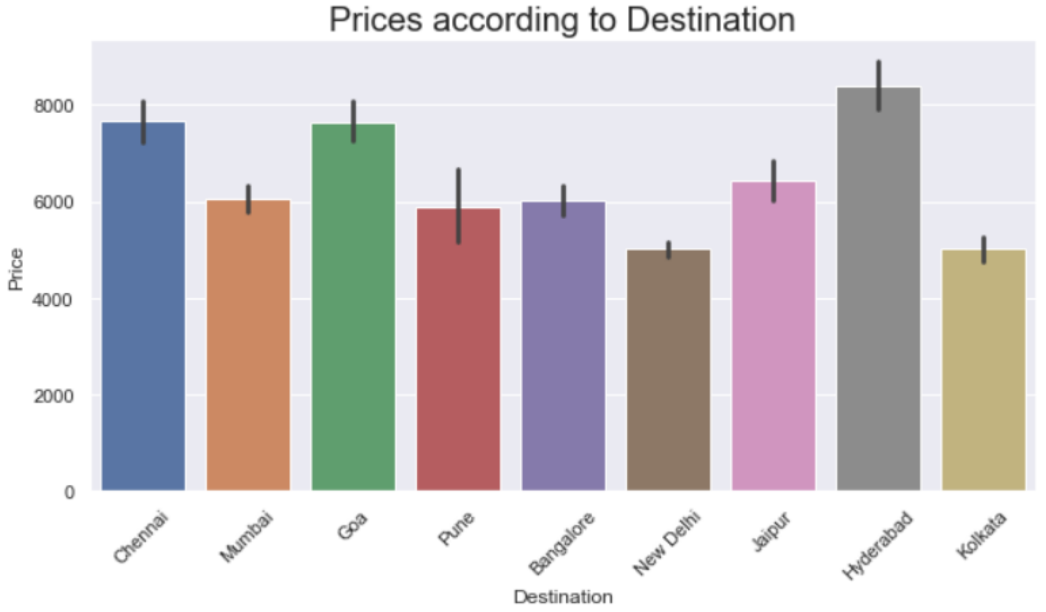
BIVARIATE ANALYSIS



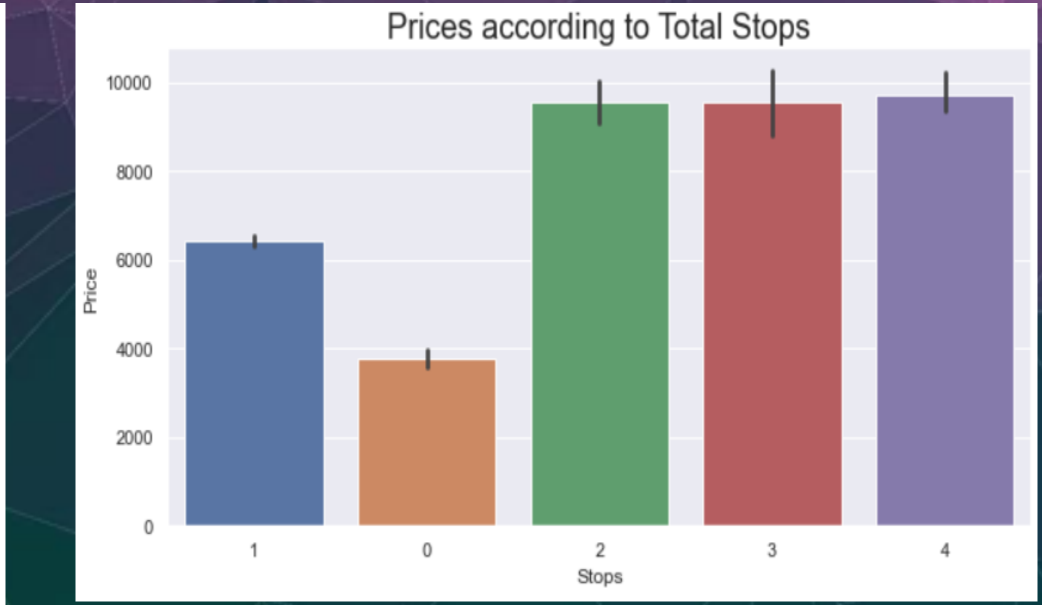
Airline Prices



Prices according to Source

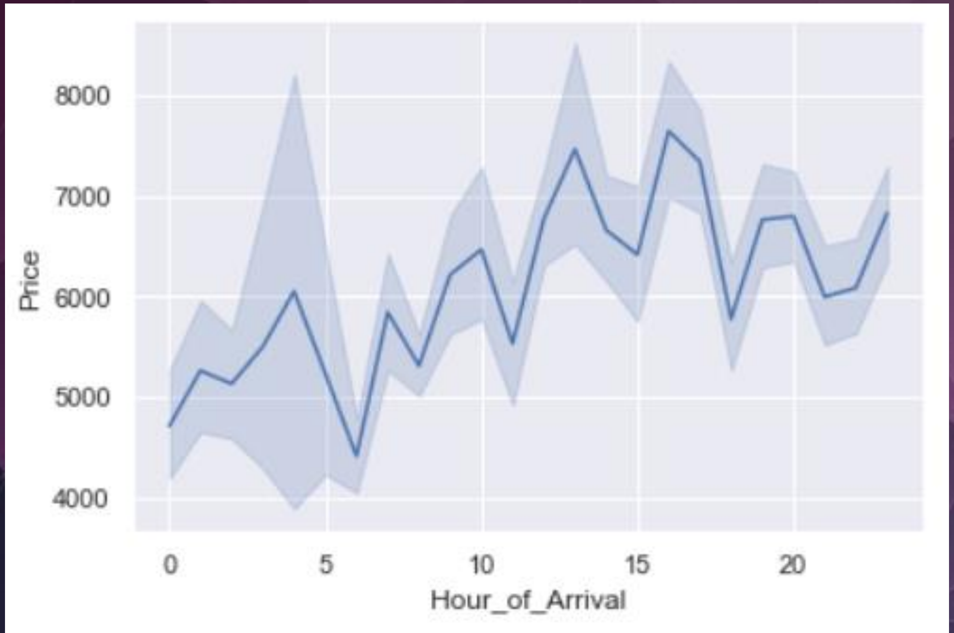


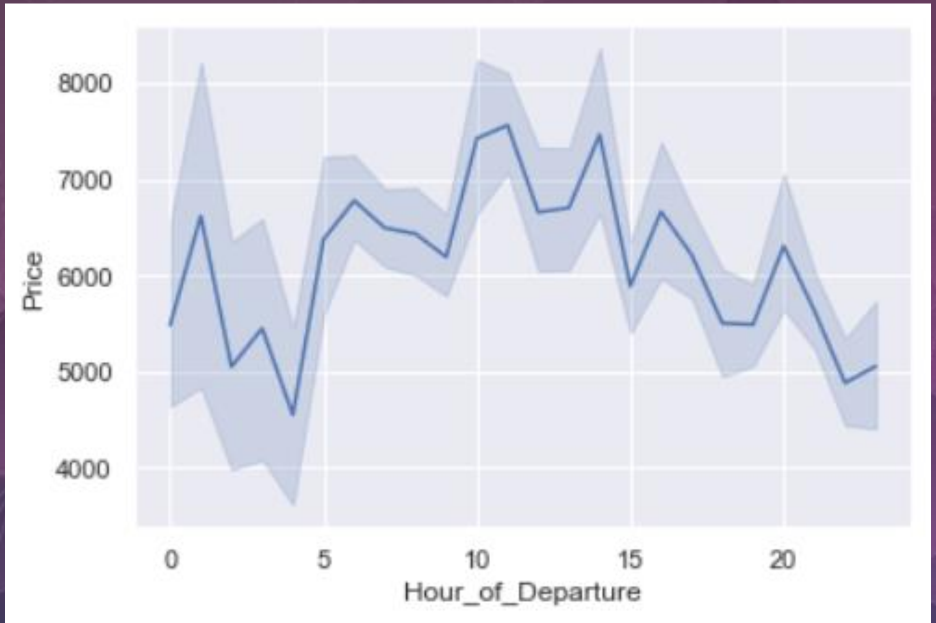
Prices according to Destination

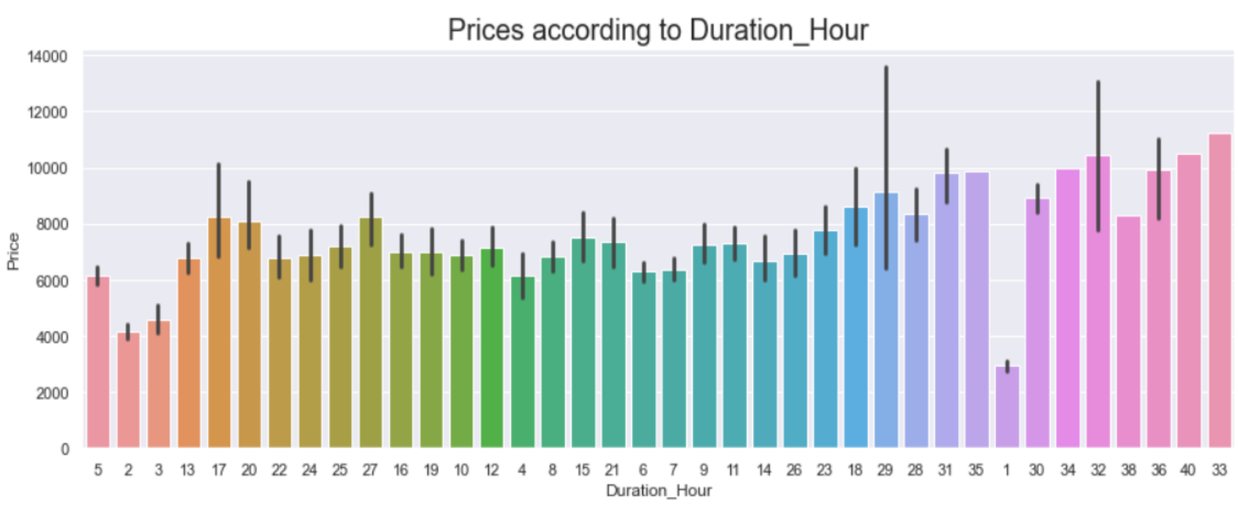


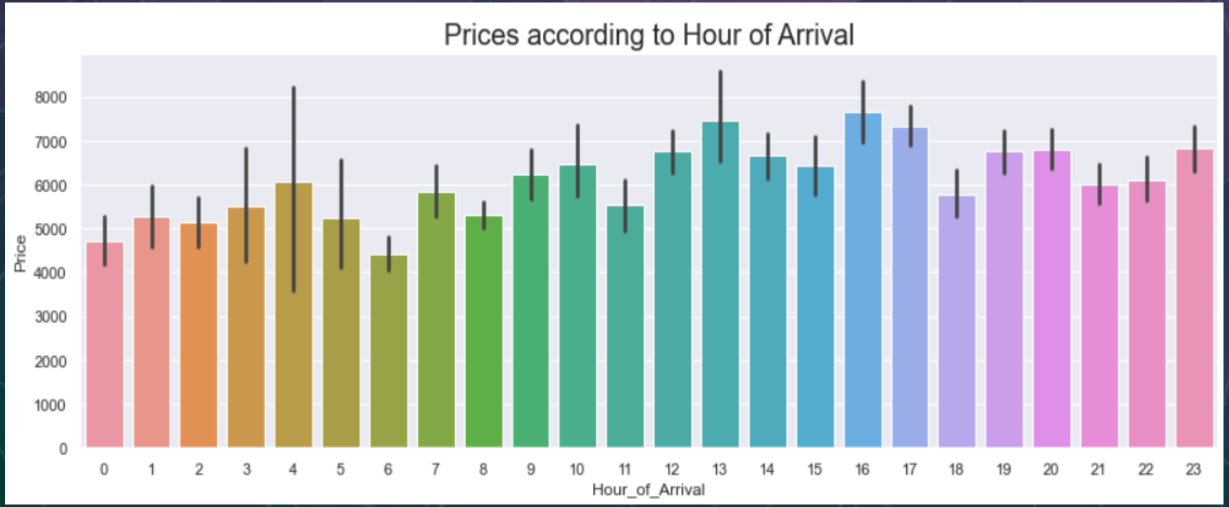
Prices according to Total Stops

BIVARIATE ANALYSIS





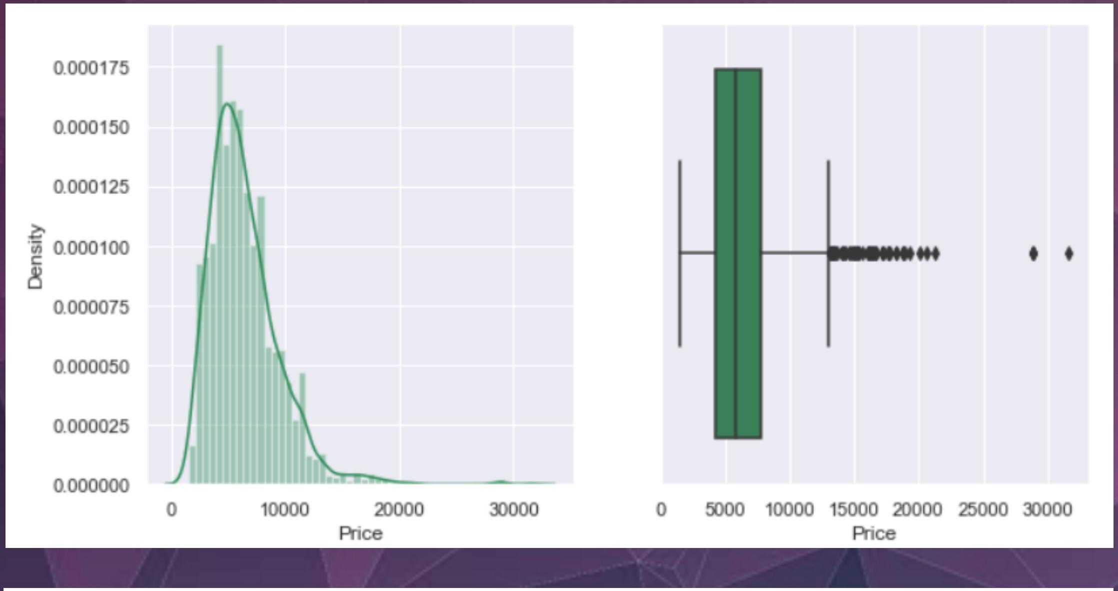


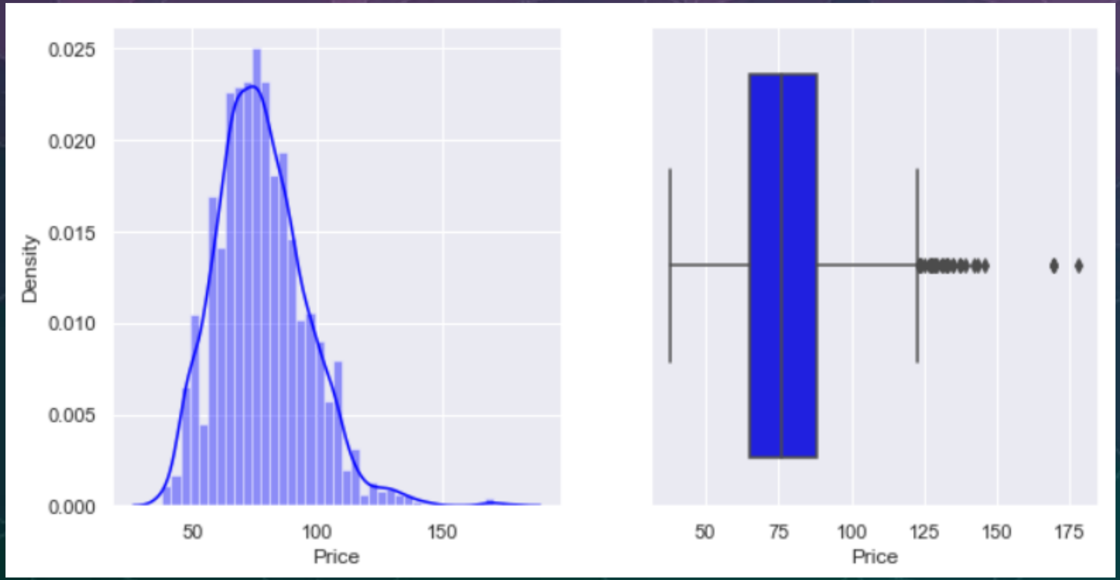


INSIGHTS

- Air India and Vistara offers expensive flights tickets whereas, indigo is most budget friendly.
- Hyderabad flights cost more during december and Kolkata costs least.
- Flight tickets from goa costs more and flight tickets from bangalore costs least.
- Flights with more than 1 stops costs more
- Flights with included meal costs more.
- Flight Price increases as the Flight duration increases.
- Flights at night costs more than flights in morning.
- Flights with departure from 10am to 3pm costs more

SKEWNESS REMOVAL USING SQUARE ROOT TRANSFORMARTIONS

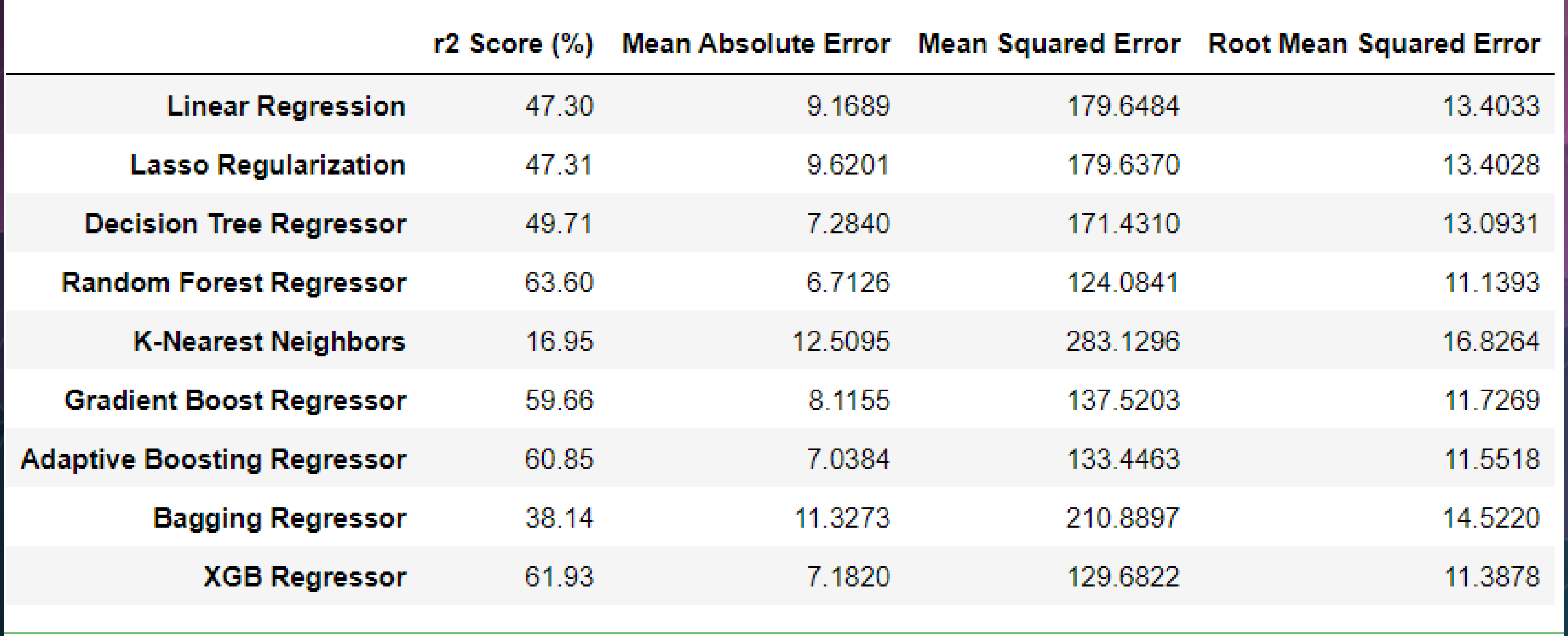




PREPARING DATASET FOR MODELLING

- Categorical features of our dataset were encoded using dummy predictors.
- Dataset was split into two parts 20% for testing and 80% for training.

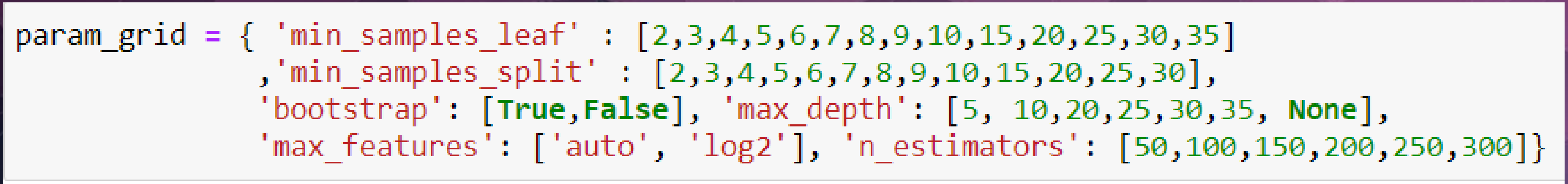
MODELS USED AND THEIR EVALUATION METRICS



	r2 Score (%)	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	47.30	9.1689	179.6484	13.4033
Lasso Regularization	47.31	9.6201	179.6370	13.4028
Decision Tree Regressor	49.71	7.2840	171.4310	13.0931
Random Forest Regressor	63.60	6.7126	124.0841	11.1393
K-Nearest Neighbors	16.95	12.5095	283.1296	16.8264
Gradient Boost Regressor	59.66	8.1155	137.5203	11.7269
Adaptive Boosting Regressor	60.85	7.0384	133.4463	11.5518
Bagging Regressor	38.14	11.3273	210.8897	14.5220
XGB Regressor	61.93	7.1820	129.6822	11.3878

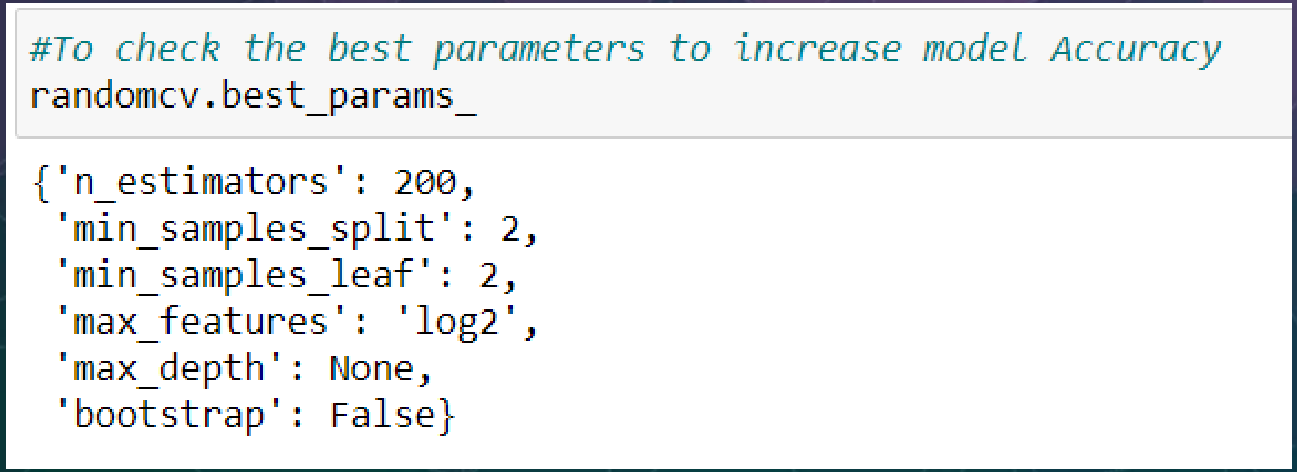
HYPERPARAMETER TUNING

Randomized Search CV was used to tune our Random Forest Regression model using the following parameters.



```
param_grid = { 'min_samples_leaf' : [2,3,4,5,6,7,8,9,10,15,20,25,30,35]
              ,'min_samples_split' : [2,3,4,5,6,7,8,9,10,15,20,25,30],
              'bootstrap': [True,False], 'max_depth': [5, 10,20,25,30,35, None],
              'max_features': ['auto', 'log2'], 'n_estimators': [50,100,150,200,250,300]}
```

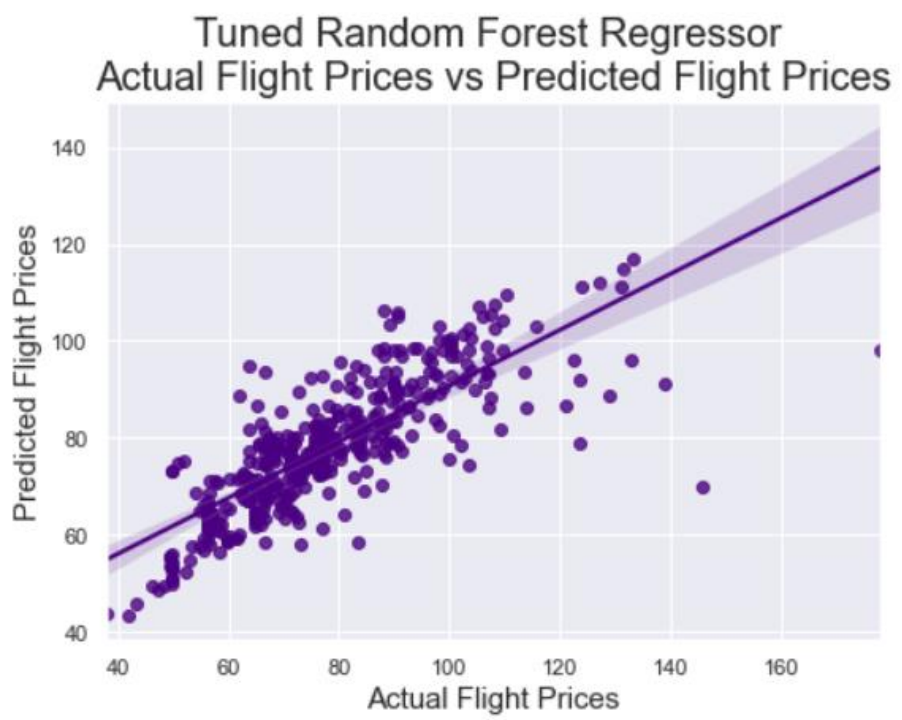
We've deployed Randomized Search CV for Hyperparameter Tuning of our model with 3 folds for each of 300 candidates, totalling 900 fits.

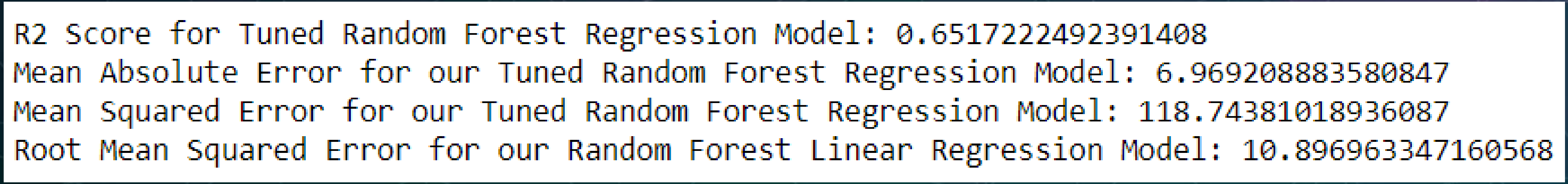


```
#To check the best parameters to increase model Accuracy
randomcv.best_params_
```

```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 2,
 'max_features': 'log2',
 'max_depth': None,
 'bootstrap': False}
```

FINAL TUNED MODEL





```
R2 Score for Tuned Random Forest Regression Model: 0.6517222492391408
Mean Absolute Error for our Tuned Random Forest Regression Model: 6.969208883580847
Mean Squared Error for our Tuned Random Forest Regression Model: 118.74381018936087
Root Mean Squared Error for our Random Forest Linear Regression Model: 10.896963347160568
```

CONCLUSION

There are two groups of airlines: `the economical group and the luxurious group`. Spicejet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the other. Vistara has a more spread out trend.

More routes can be added and the same analysis can be expanded to major airports and travel routes in India.

The analysis can be done by increasing the data points and increasing the historical data used. That will train the model better giving better accuracies and more savings.

Thank You!