**Flight Price Prediction Project**

Submitted by:

Tushar Saini

**Acknowledgement**

I would like to express my sincere gratitude to fliprobo technologies for supporting me in this project. I wish to express my sincere gratitude to my SME, Ms. Sapna Verma, for his enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that helped me tremendously at all times in my project time. Her immense knowledge, profound experience and professional expertise in Data Science has enabled me to complete this research successfully. Without her support and guidance, this project would not have been possible.

I also wish to express my sincere thanks to the DataTrained for helping me to get trained in the field of Data Science. Without them I don’t know anything about Data Science. They help me to learn the Data Science from scratch. The faculty their helped me a lot in every aspect.

I also like to thank websites like geeksforgeeks.com, towardsdatascience.com, etc for providing the information regarding many topics of data science.

1. **Introduction:-**

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on -

* 1. Time of purchase patterns (making sure last-minute purchases are expensive)
  2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

So, we have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

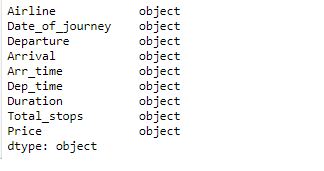
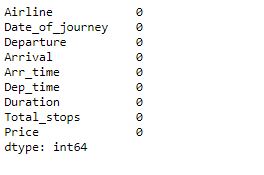
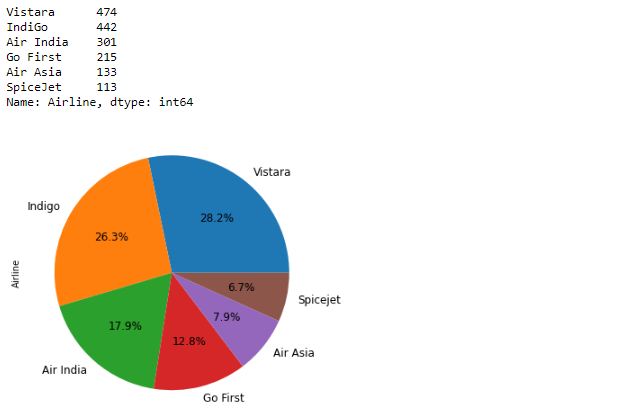
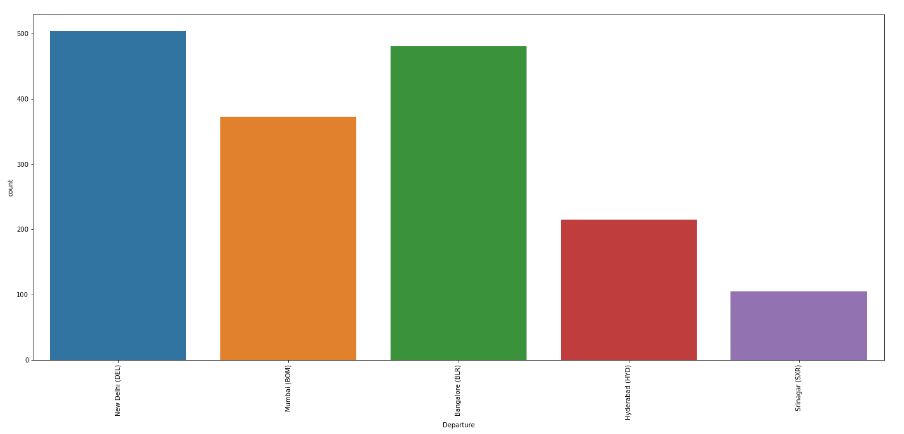
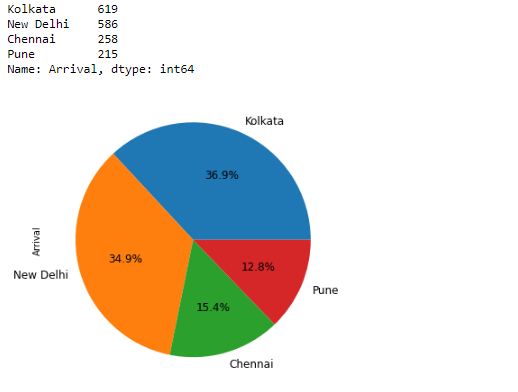
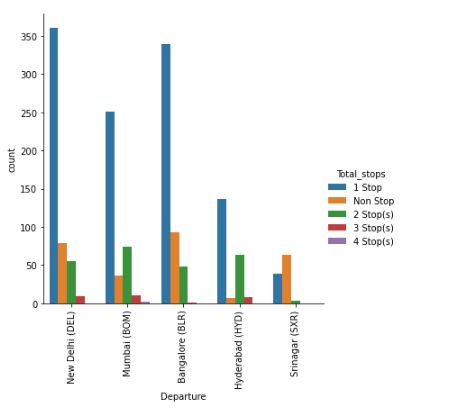
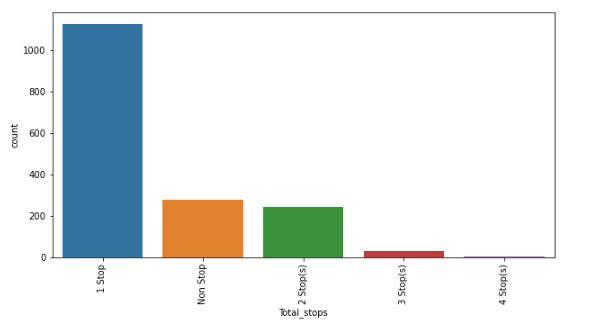
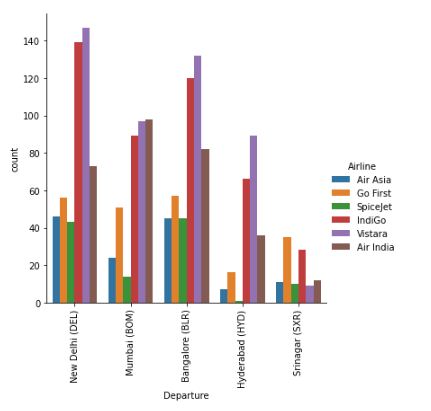
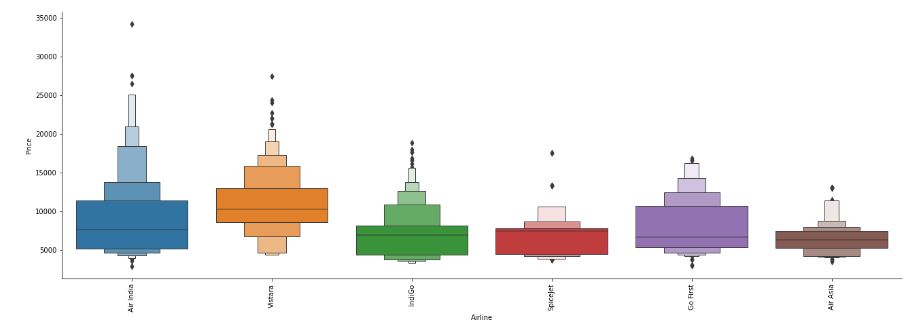
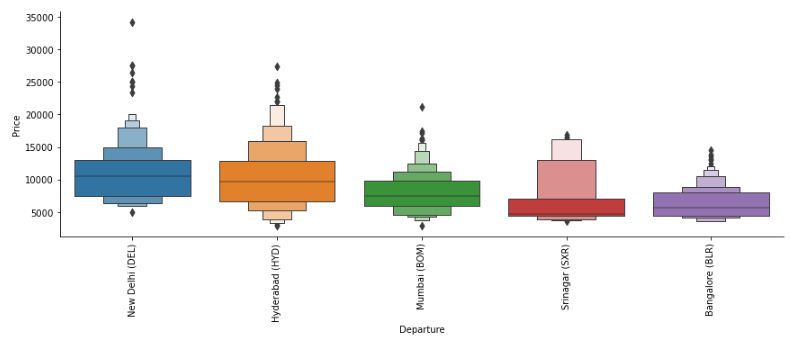
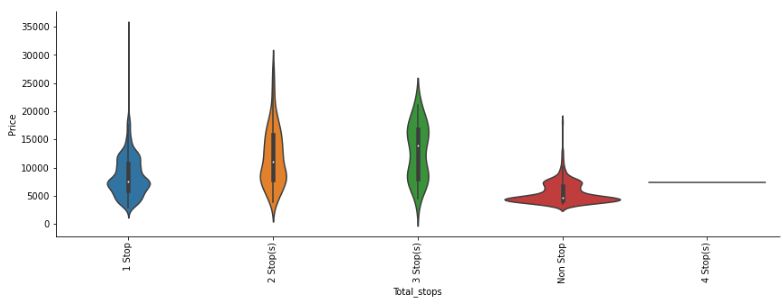
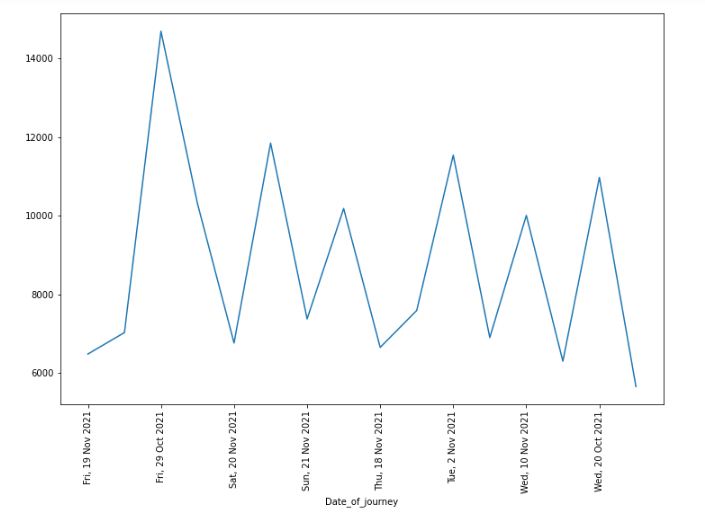
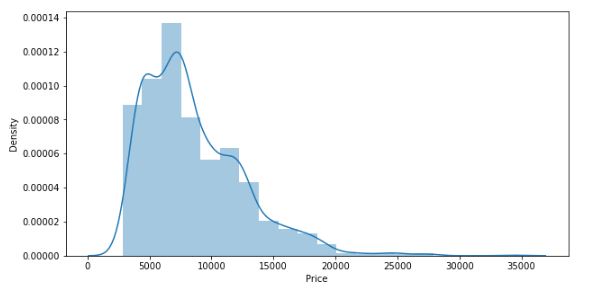
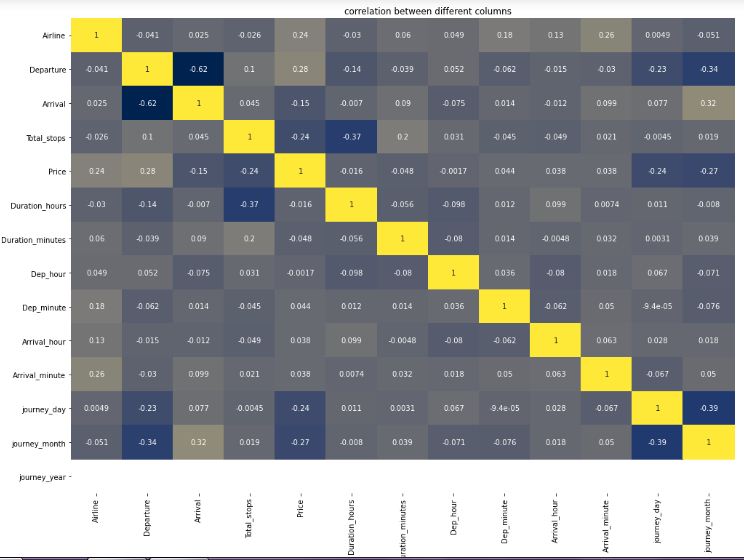
We have to scrape at least 1500 rows of data. I am scrapping the data from 'Yatra.com'. We can scrape more data as well, it’s up to individual, more the data better the model. The number of columns for data doesn’t have limit, it’s up to you and your creativity. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price. We can make changes to it, we can add or we can remove some columns, it completely depends on the website from which we are fetching the data.

After collecting the data, we need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

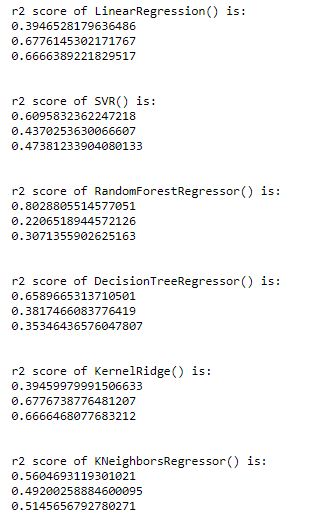
1. **Analytical Problem Framing:-**

* **Data Visualization :-**

By analyzing the data, I have found some important features of that dataset. The analyzed information that I found are given below:-

* There are 1678 rows and 9 columns in the dataset and it doesn’t have any null values.
* I also observe that the datatypes of columns in the dataset are of ‘object’.
* There are no null values in the dataset.
* We can observe that majority of the airlines in the dataset are Vistara and Indigo.
* We can observe that majority of the flights in the dataset are departed from 'Delhi' or 'Bangalore’.
* We can observe that 'Kolkata' and 'New Delhi' is the most arrived city in the dataset.
* We can observe that in all the cities except 'Srinagar', majority of the flights are having 1 stop but majority of the flights from 'Srinagar' are non-stop.
* We can observe that maximum number of flight have 1 stop in their route.
* We can observe that from 'Bangalore', 'Delhi' and ‘Hyderabad’ vistara and indigo have most number of flights. From 'Mumbai', Air India and vistara and indigo have most number of flights.
* We can observe that 'Air India' has the highest price of ticket and 'Air Asia' has the lowest price of ticket.
* We can observe that from 'New Delhi', the ticket price is the highest and from 'Bangalore' the ticket price is the lowest.
* We can observe that the flights which have 1 stop have higher price than any other.
* We can observe that the prices of the flights are higher around date 29 Oct 2021.
* We can observe that the majority of the flights have price range from 3000 to 12000.
* We can observe that 'Airline' and 'Departure' is the most positively correlated column with the 'Price' and 'journey\_month' and 'journey\_day' is the most negatively correlated column with the 'Price'
* **Data Pre-Processing:-**
* We can drop the unnamed column as it is of no use because we made our own index.
* We have converted the price column from object to int32.
* I have separated the hours and minutes of the duration column and make a separate column for both of them and drop the duration column.
* I have separated the hours and minutes of the Dep\_time column and make a separate column for both of them and drop the Dep\_time column.
* I have separated the hours and minutes of the Arr\_time column and make a separate column for both of them and drop the Arr\_time column.
* I have separated the Days, month and year of the Date\_of\_journey column and make a separate column for each of them and drop the Date\_of\_journey column.
* We have encoded the dataset using Ordinal Encoder.
* We can observe that there are no considerable outliers in the dataset.
* We observe that there is some skewness present in the ‘Total\_stops’, ‘Price’, ‘journey\_month’. We have removed the skewness from these columns using ‘Power\_Transform’ except ‘Journey\_month’ as it is a classification column.
* We have use Feature selection technique to find the best features for our Target value.
* **Hardware and Software Requirements and Tools Used:-**
* We used anaconda jupyter notebook to do research on this project.
* I use python to do my coding for the dataset.
* We use different kind of libraries present in the jupyter notebook like pandas, numpy, sklearn, seaborn, matplotlib, joblib, etc.
* I have used dell inspiron 1440 which has 3GB Ram and windows 7.
* I use pandas to import that data make it in a dataframe.
* Sklearn is used to import all the model and methods that are used for pre-processing and model building of the dataset.
* Matplotlib and Seaborn are used to plot various graphs and figures to visualize the data.

1. **Model/s Development and Evaluation:-**

* For building a machine learning model first I split the dataset into independent and target variable.
* Independent variable that is ‘x’ contains all the other columns except ‘Price’ because ‘Price’ is our target value which we have to predict.
* Target variable which is ‘y’ contains the ‘Price’ column.
* Now I find the best random state for one model and I use that random state for other models also.
* First I use ‘Linear Regression’ as it is a Regression problem. I make a loop and take 0-200 random state. That loop will give me the accuracy score for every random state from 0-200.
* At random state 59 the testing accuracy and training accuracy is highest.
* Now I test this random state with other models also.
* I used a total of 6 models which are ‘Linear Regression’, ‘SVR’, ‘RandomForest Regression’, ‘DecisionTree Regression’, ‘KernelRidge’, ‘KNeighbors Regression.
* ‘Random Forest Regression’ gives me the best accuracy score that is 80.28%.
* Now I check whether my models are underfitted or overfitted. For that I use cross validation score and taking the k-fold value of 3.
* Minimum difference in r2\_score and cross validation score is for SVR (26.86). So it is our best model.
* After that I have done hyperparameter tuning of the model to increase the model accuracy score.
* After doing the hyperparameter tuning, our model is 61% accurate.

1. **Conclusion:-**
   * From the above project I find that the majority of the flights have a ticket price range between 3000 to 12000.
   * Flight ticket price depends on how much prior you are booking the ticket and what is your Departure location.
   * People are travelling more from flights nowadays, so airlines make the most profit from this by increasing the flight price according to the demand.
   * I made a machine learning model to check what will be the flight price by taking different parameters of the car. My model is 61% accurate.
   * I have learned that through visualization we can understand the core of the data. Visualization is very important for a data science project as it will give us information about the dataset.
   * I have used 6 algorithms. What I have learned that it might be possible some algorithms can be used for a specific problem. So we have to check what is the problem through which we can choose our algorithm.