FLIGHT PRICE PREDICTION PROJECT

Submitted by:

Khanin Deka

ACKNOWLEDGMENT

- https://datascience.stackexchange.com/
- https://stackoverflow.com/
- FlipRobo Technologies
- https://scikit-learn.org/
- https://www.yatra.com/
- https://simpleflying.com/india-air-fare-changes/

INTRODUCTION

Business Problem Framing

In the era of globalization passenger airlines have seen tremendous growth in traffic all over the world. India is the world's third-largest domestic civil aviation market. In India passenger traffic amounted to over 115 million at airports across the country in financial year 2021, out of which over 10 million were international passengers.

In this project we will try to analyze the price of airline tickets based on various factors. 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. We will try to analyze this price variation by analyzing various features associated with it.

Conceptual Background of the problem

Recently, the Government of India has done what many airlines in the country had wanted for quite some time – removal of airfare bands. With the government intervention gone, all the traditional factors are now coming into play to decide how much flights will cost. Demand and supply, competition, route, etc dictates how much passengers have to pay for a ticket.

We will build a machine learning model to predict the price of flight tickets. Since price is a continuous variable, hence we will use regression algorithms to predict it based on various features. We have a dataset of 59707 flights which we will be using for building our model.

Analytical Problem Framing

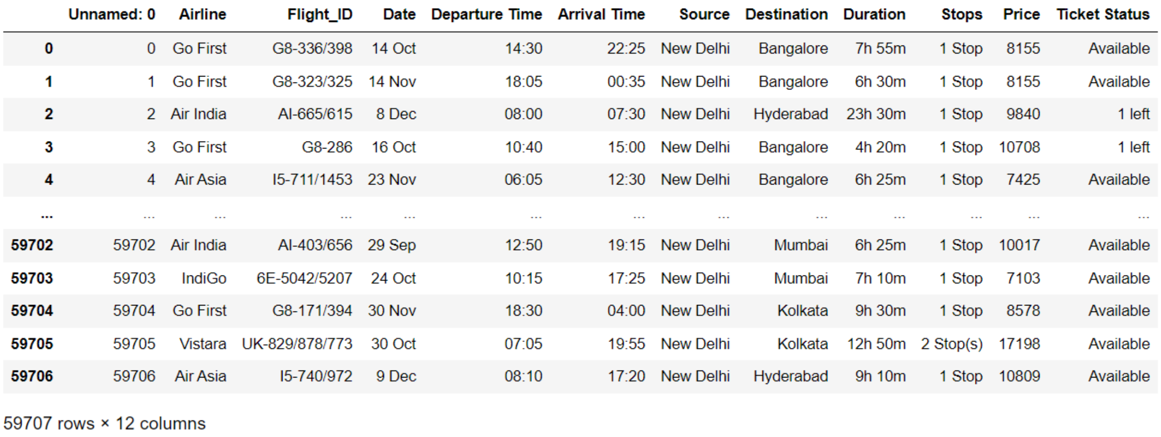
Data Sources and their formats

1. The data is collected from www.yatra.com for four different routes with dates ranging from 22nd September to 31st December of 2022.
2. All the flights have New Delhi as the source city and either Mumbai, Bangalore, Kolkata or Hyderabad as the destination city.
3. All the flight tickets are of economy class.
4. Most of the data is initially in text format.
5. The following details of a flight are collected:
   a) Airline name
   b) Flight ID
   c) Date
   d) Departure time
   e) Arrival time
   f) Source
   g) Destination
   h) Duration
   i) Stops
   j) Ticket status
   k) Price
6. The dataset contains a total of 59707 entries.

Tools and Libraries used

1. This project is built using Jupyter-Notebook of Anaconda Navigator.
2. Python language version 3.9.12 is used.
3. List of libraries include pandas, numpy, sklearn, selenium, and libraries for various regression algorithms.
4. For data visualization matplotlib and seaborn are used.

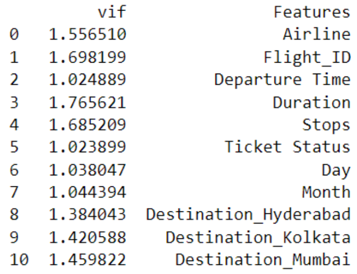
- **Data Preprocessing Done**
  - A snapshot of the dataset:



	Unnamed: 0	Airline	Flight_ID	Date	Departure Time	Arrival Time	Source	Destination	Duration	Stops	Price	Ticket Status
0	0	Go First	G8-336/398	14 Oct	14:30	22:25	New Delhi	Bangalore	7h 55m	1 Stop	8155	Available
1	1	Go First	G8-323/325	14 Nov	18:05	00:35	New Delhi	Bangalore	6h 30m	1 Stop	8155	Available
2	2	Air India	AI-665/615	8 Dec	08:00	07:30	New Delhi	Hyderabad	23h 30m	1 Stop	9840	1 left
3	3	Go First	G8-286	16 Oct	10:40	15:00	New Delhi	Bangalore	4h 20m	1 Stop	10708	1 left
4	4	Air Asia	I5-711/1453	23 Nov	06:05	12:30	New Delhi	Bangalore	6h 25m	1 Stop	7425	Available
...	...	...	...	...	...	...	...	...	...	...	...	...
59702	59702	Air India	AI-403/656	29 Sep	12:50	19:15	New Delhi	Mumbai	6h 25m	1 Stop	10017	Available
59703	59703	IndiGo	6E-5042/5207	24 Oct	10:15	17:25	New Delhi	Mumbai	7h 10m	1 Stop	7103	Available
59704	59704	Go First	G8-171/394	30 Nov	18:30	04:00	New Delhi	Kolkata	9h 30m	1 Stop	8578	Available
59705	59705	Vistara	UK-829/878/773	30 Oct	07:05	19:55	New Delhi	Kolkata	12h 50m	2 Stop(s)	17198	Available
59706	59706	Air Asia	I5-740/972	9 Dec	08:10	17:20	New Delhi	Hyderabad	9h 10m	1 Stop	10809	Available

59707 rows × 12 columns

  - 'Unnamed: 0' column is deleted.
  - 914 duplicate entries are found and removed.
  - No null entries found in the dataset.
  - From the 'Date' column the day and month data is extracted to separate columns.
  - The 'Departure Time' column is converted to date-time format using pandas *to_datetime* method. Then we extract the hour and minute, and combine them to make float number which represents the original data. Then we replace the original column with the new column.
  - The 'Duration' column is converted to float data-type from text format.
  - 'Source' column is deleted as all the flights have same source city.
  - 'Arrival Time' column is deleted as we already have the departure time and duration data.
  - **Data Encoding:**
    'Stops' and 'Ticket Status' columns are encoded using 'replace' function. 'Flight_ID' and 'Airline' columns are encoded using 'LabelEncoder' of scikit-learn. The 'Destination' column is encoded using the pandas 'get_dummies' method.
  - **Feature Scaling:**
    The features are scaled using 'StandardScaler' of scikit-learn.
  - **Checking VIF:**
    The variance inflation factor is checked for the features to know about any multicollinearity issues. Below is the screenshot:



```
         vif              Features
0   1.556510               Airline
1   1.698199             Flight_ID
2   1.024889        Departure Time
3   1.765621              Duration
4   1.685209                 Stops
5   1.023899         Ticket Status
6   1.038047                   Day
7   1.044394                 Month
8   1.384043  Destination_Hyderabad
9   1.420588    Destination_Kolkata
10  1.459822     Destination_Mumbai
```

No multicollinearity issues can be seen.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches

  Our task is to predict the price of a pre-owned car which is a regression task. For this multiple algorithmic approaches are tried:

  1) **Multiple Linear Regression**: In this approach relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data**.**
  2) **AdaBoost Regression:** It is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.
  3) **Random Forests Regression**: It is an ensemble learning method for regression that operates by constructing a multitude of decision trees at training time. The mean or average prediction of the individual trees is then returned.
  4) **XGBoost Regression**: Extreme Gradient Boosting or XGBoost is an ensemble machine learning algorithm which uses an objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference

between actual values and predicted values, i.e. how far the model results are from the real values.

5) **KNN Regression**: It approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

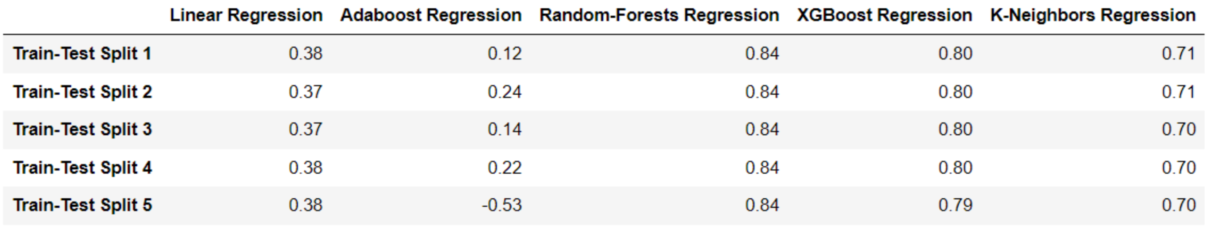
- ## Performance Evaluation

Let’s first use the baseline algorithms and evaluate the performance on different train-test splits. Then we do cross-validation of these models for 5-fold to 10-fold to verify the scores.

We will use the coefficient of determination (R^2) to evaluate the models’ performance. The R^2 score for a model is a useful statistic in regression analysis, as it often describes how ‘good’ that model is at making predictions.

The values for R^2 range from 0 to 1. The closer the score is to 1 better is the model. A model can give a negative R^2 score as well, which indicates that the model is arbitrarily worse than the one that always predicts the mean of the target variable.

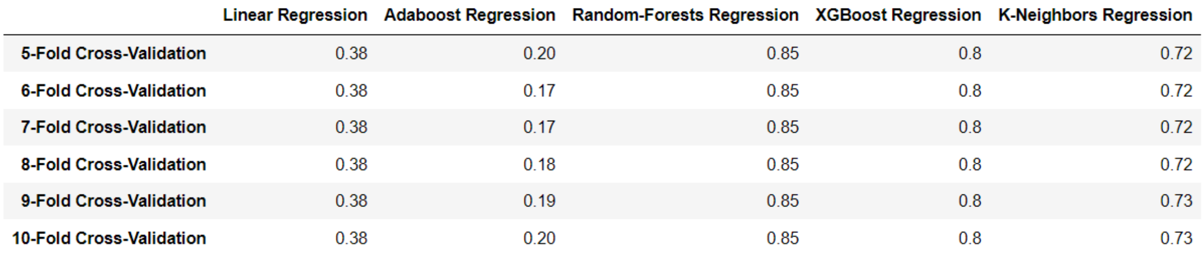
Let’s see the R^2 scores of the 5 different algorithms on 5 different train-test splits:



	Linear Regression	Adaboost Regression	Random-Forests Regression	XGBoost Regression	K-Neighbors Regression
Train-Test Split 1	0.38	0.12	0.84	0.80	0.71
Train-Test Split 2	0.37	0.24	0.84	0.80	0.71
Train-Test Split 3	0.37	0.14	0.84	0.80	0.70
Train-Test Split 4	0.38	0.22	0.84	0.80	0.70
Train-Test Split 5	0.38	-0.53	0.84	0.79	0.70

Random-Forests is giving the best performance among all the algorithms.

Let’s see the cross-validation results:

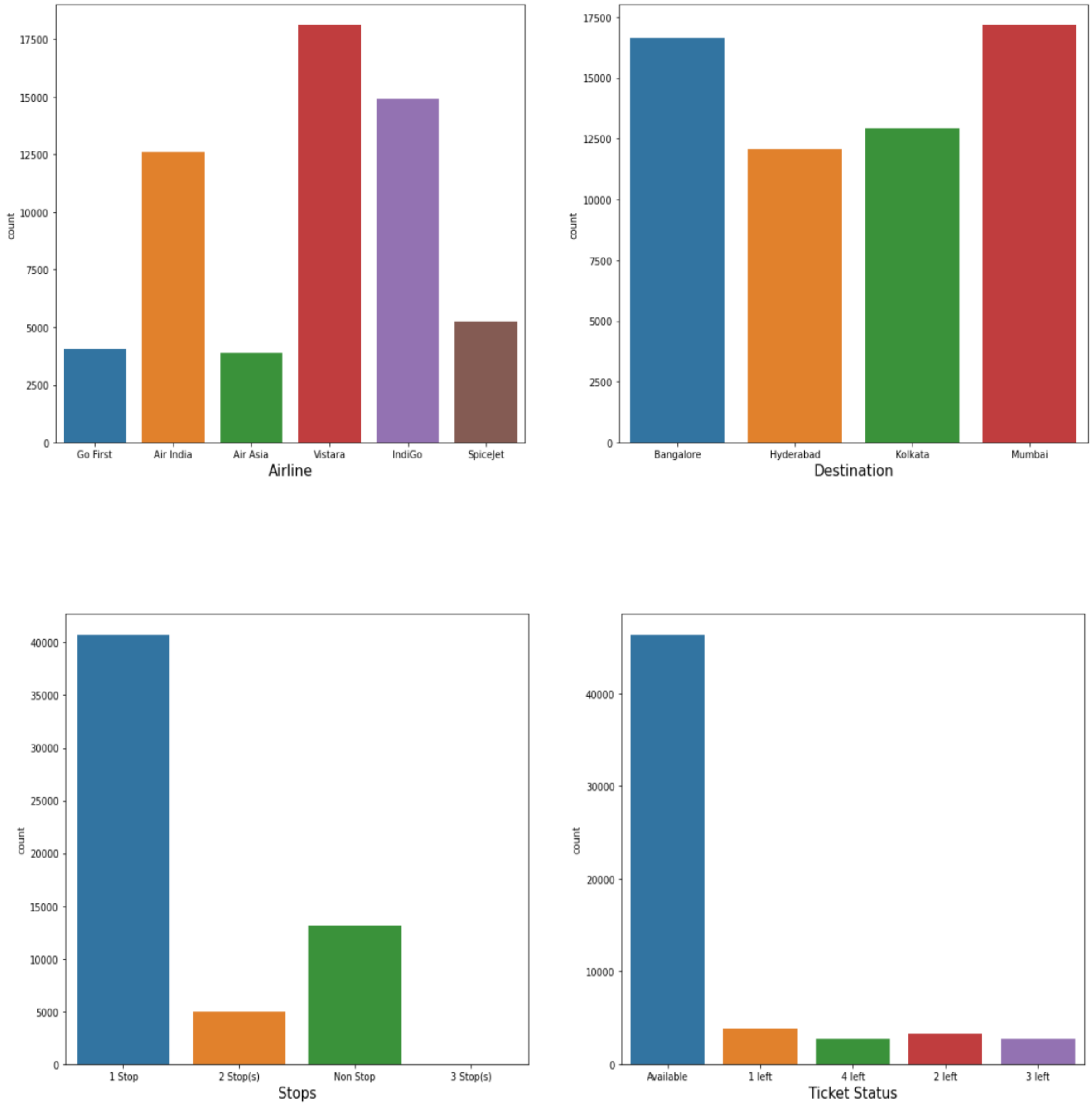


	Linear Regression	Adaboost Regression	Random-Forests Regression	XGBoost Regression	K-Neighbors Regression
5-Fold Cross-Validation	0.38	0.20	0.85	0.8	0.72
6-Fold Cross-Validation	0.38	0.17	0.85	0.8	0.72
7-Fold Cross-Validation	0.38	0.17	0.85	0.8	0.72
8-Fold Cross-Validation	0.38	0.18	0.85	0.8	0.72
9-Fold Cross-Validation	0.38	0.19	0.85	0.8	0.73
10-Fold Cross-Validation	0.38	0.20	0.85	0.8	0.73

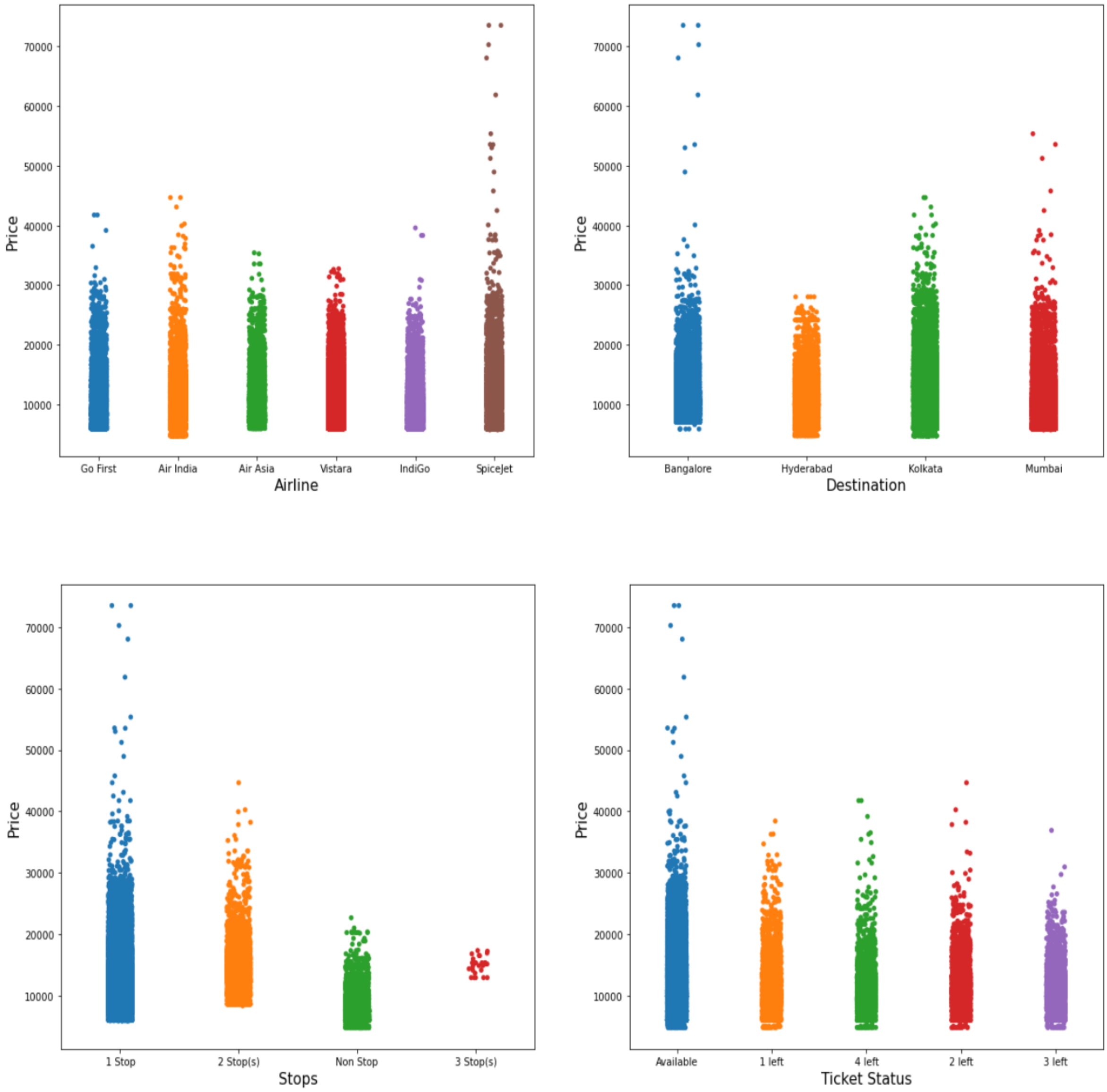
From the cross validation results it can be verified that Random-Forests is the best performing algorithm for our dataset.

- **Visualizations**

  - Count-plots of some of the categorical features:



- Strip-plots of some of the categorical features with respect to the target variable i.e. 'Price':

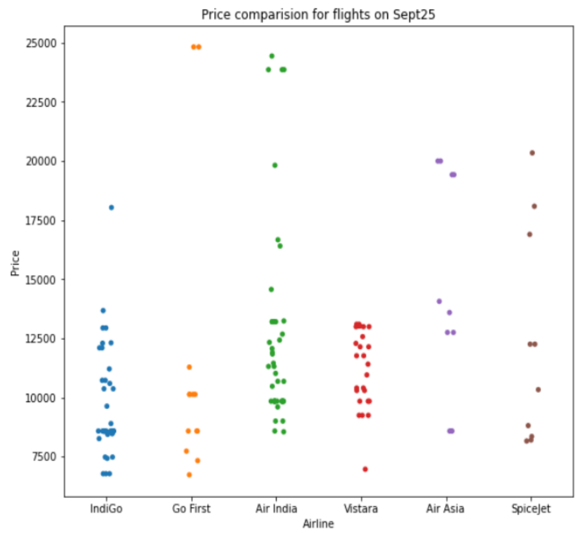


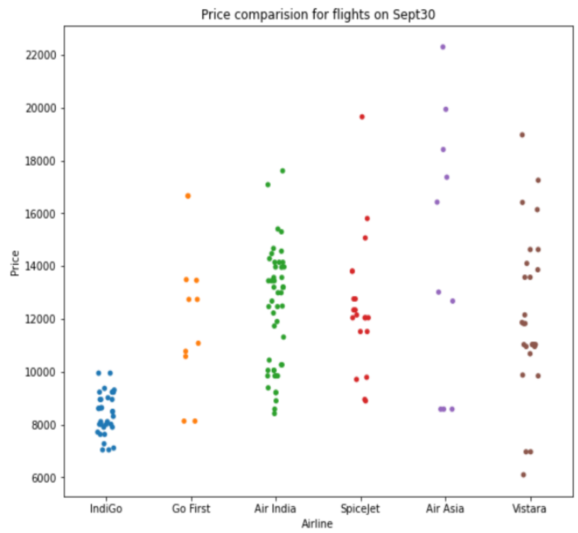
Observations:

1. Flights with more stops have higher starting fares.

- Plots to compare price among different airlines on different dates in the month of September.

  These plots are for New Delhi-Mumbai route having 1 stop during the journey.

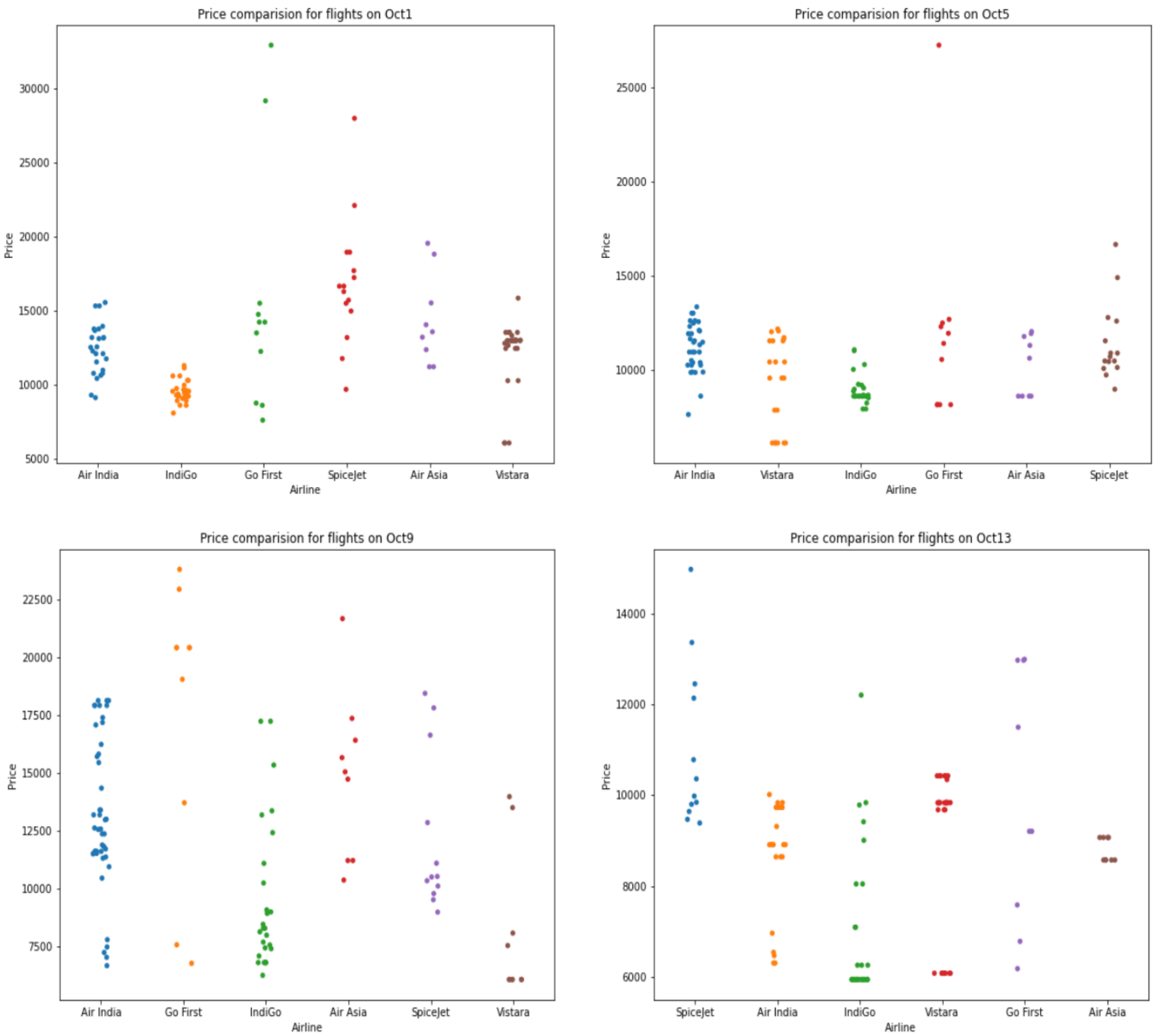




Observations:

1. Indigo and Go First airlines have the lowest starting fares for September 25.
2. Vistara airlines have the lowest starting fare for September 30 but the price range varies a lot. Indigo airlines have a slightly higher starting price but the range is limited near 10000 rupees.
3. There is a lot of price variation within the flights of same airline. Only Indigo airlines have a relatively less variation of price on September 5.

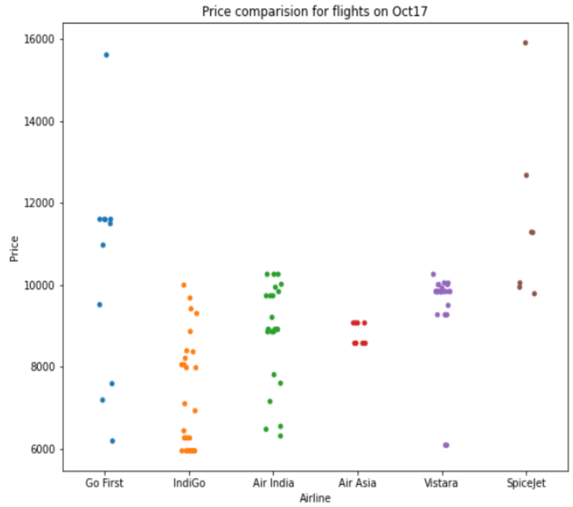
- Plots to compare price among different airlines on different dates in the month of October. These plots are for New Delhi-Mumbai route having 1 stop during the journey.

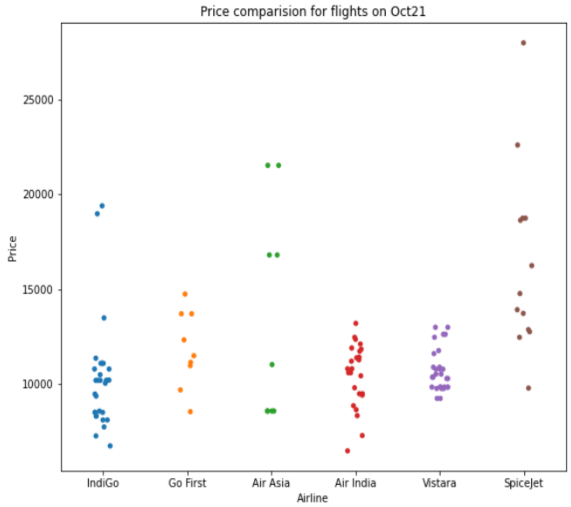


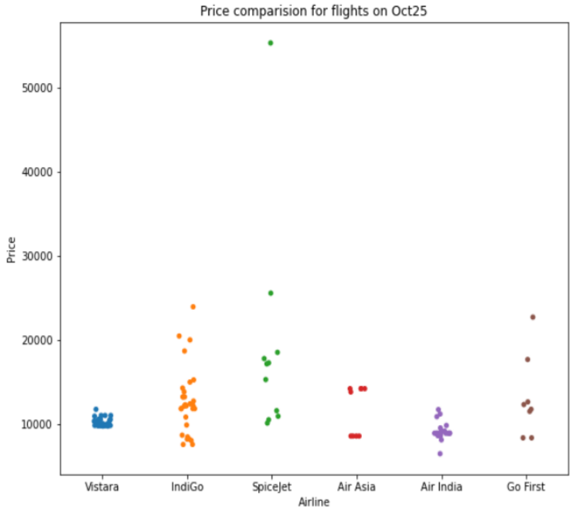
Observations:

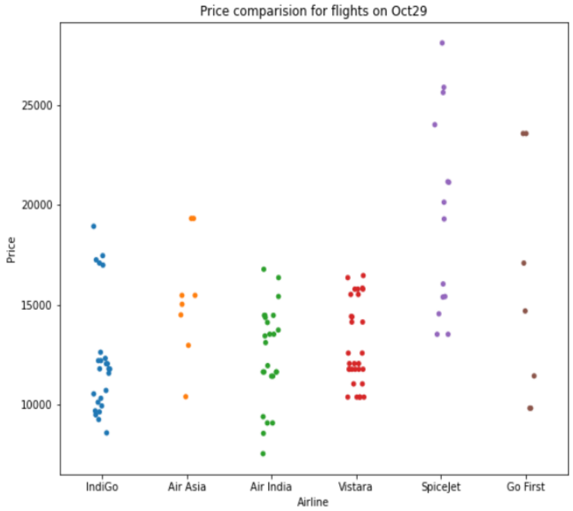
1. Vistara has the lowest starting fare for flights on October 1 and 5.
2. On October 9 and 13, Vistara and Indigo have almost same starting fares.
3. AirAsia and Spicejet have a higher starting price compared to others.
4. Indigo airlines have lesser variation of price as compared to others.

- Plots to compare price among different airlines on different dates in the month of October. These plots are for New Delhi-Mumbai route having 1 stop during the journey.







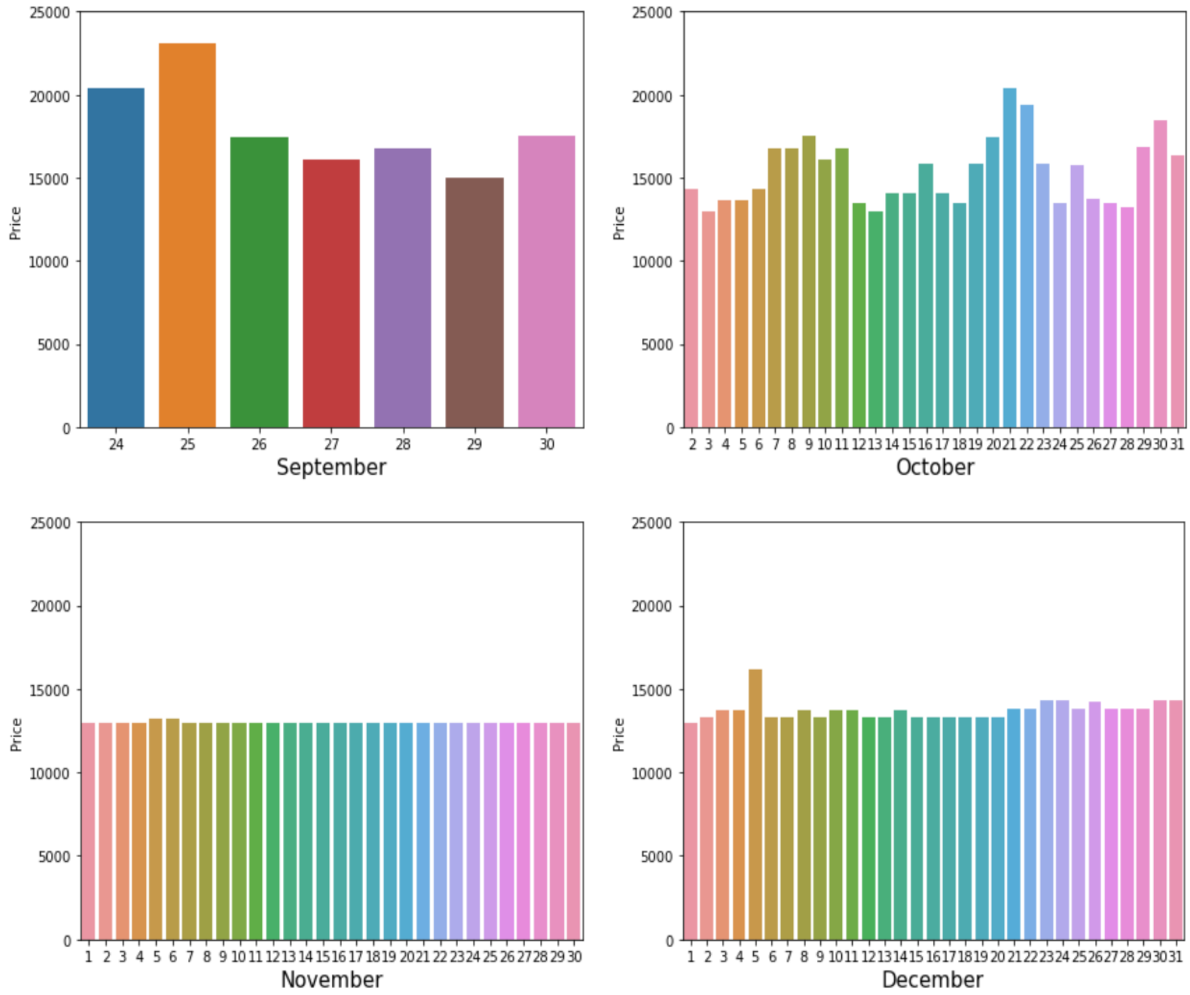


Observations:

1. Indigo airlines have the lowest starting fare on October 17. On October 21, 25 and 29 Air India have the lowest starting fare, Indigo have slightly higher price.
2. Spicejet seems to have a higher starting price as compared to others.

Plots to analyze the variation of fare of a flight over time

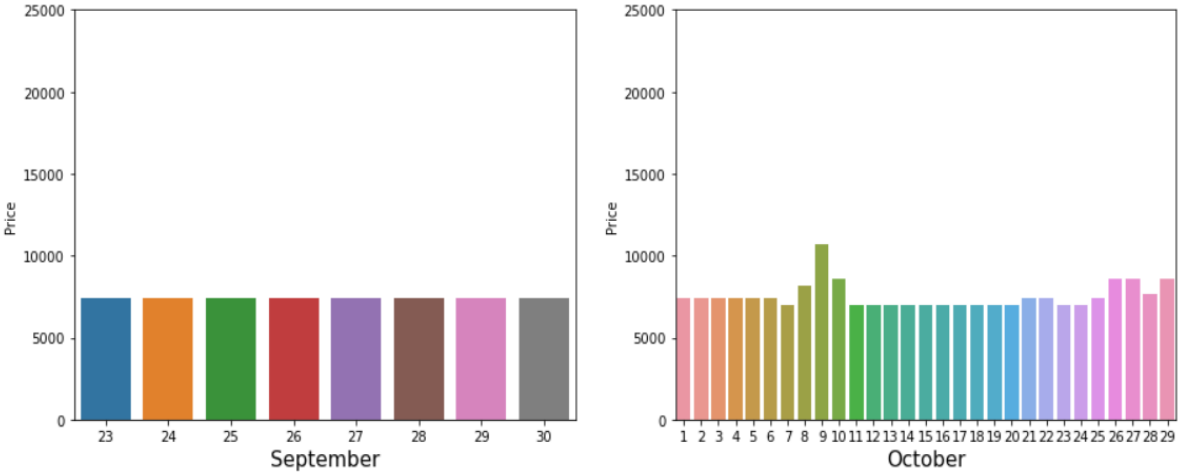
- Let’s see for the 'UK-993/773' flight. This flight is from Delhi to Kolkata with a stop at Mumbai. This is a flight of Vistara Airlines.



Observations:

1. It can be seen that in month of September and October the prices are higher compared to November and December. So, when the date of journey is near fares are higher.
2. The prices are fluctuating more in the month of September and October whereas it’s almost constant for the month of November and December. Hence, when the date of journey is near there is more fluctuation of prices.

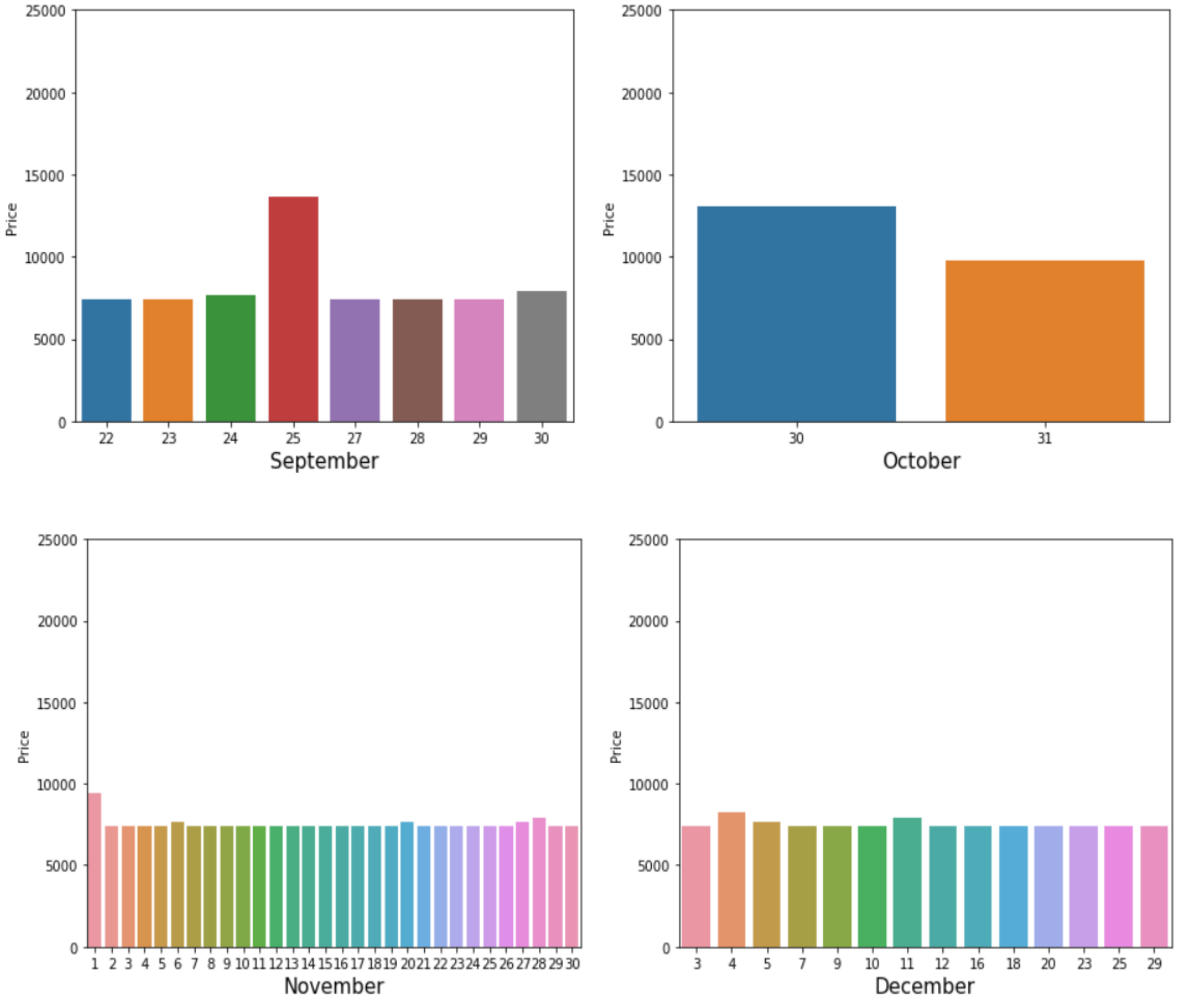
- Let’s see for the ‘6E-2048’ flight. This flight is from Delhi to Bangalore of Indigo Airlines. Data is available for September and October only.



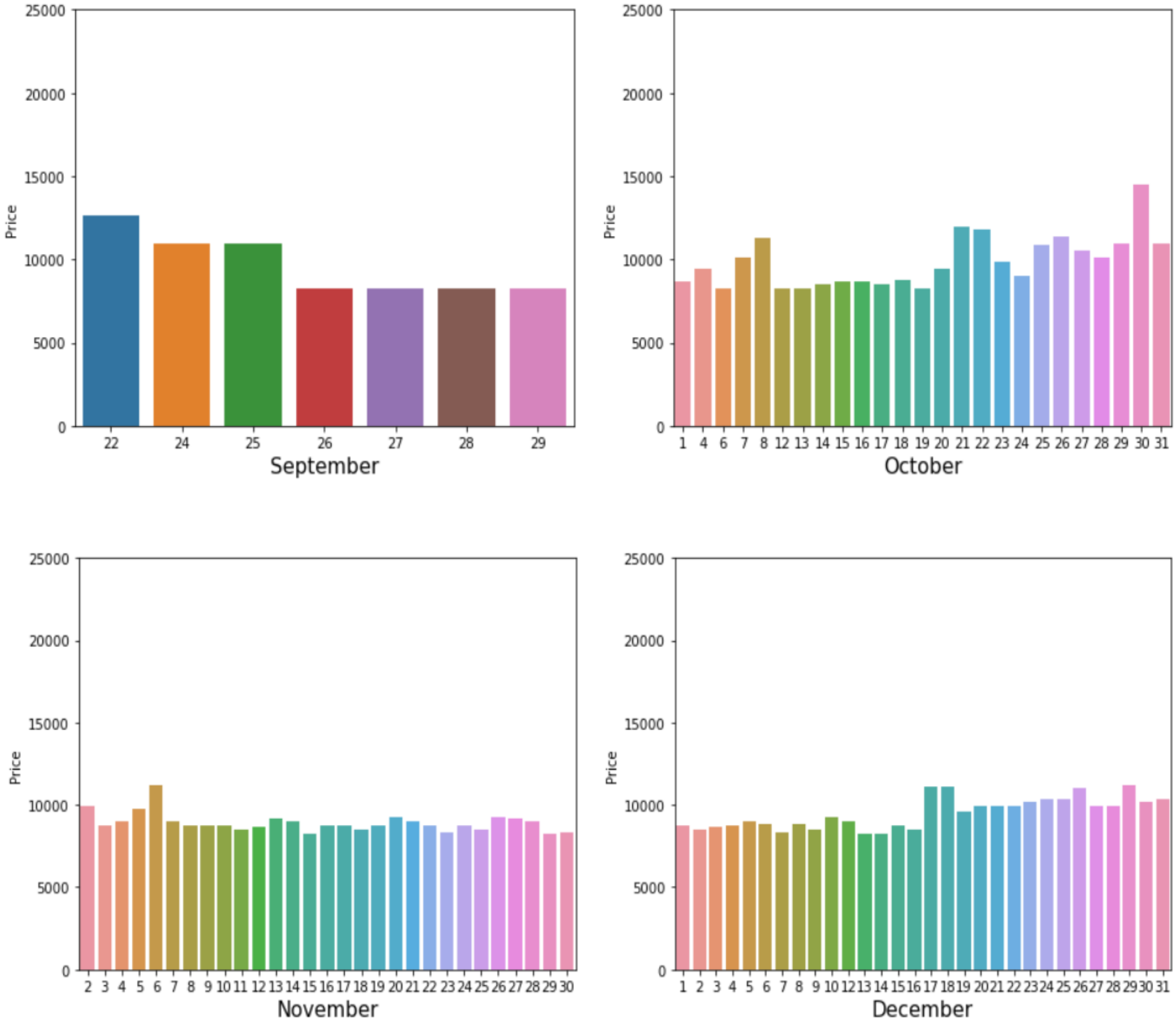
Observations:

1. The fare is quite consistent over the duration of around 40 days.

Let’s see another Indigo flight- ‘6E-781’- Delhi to Bangalore



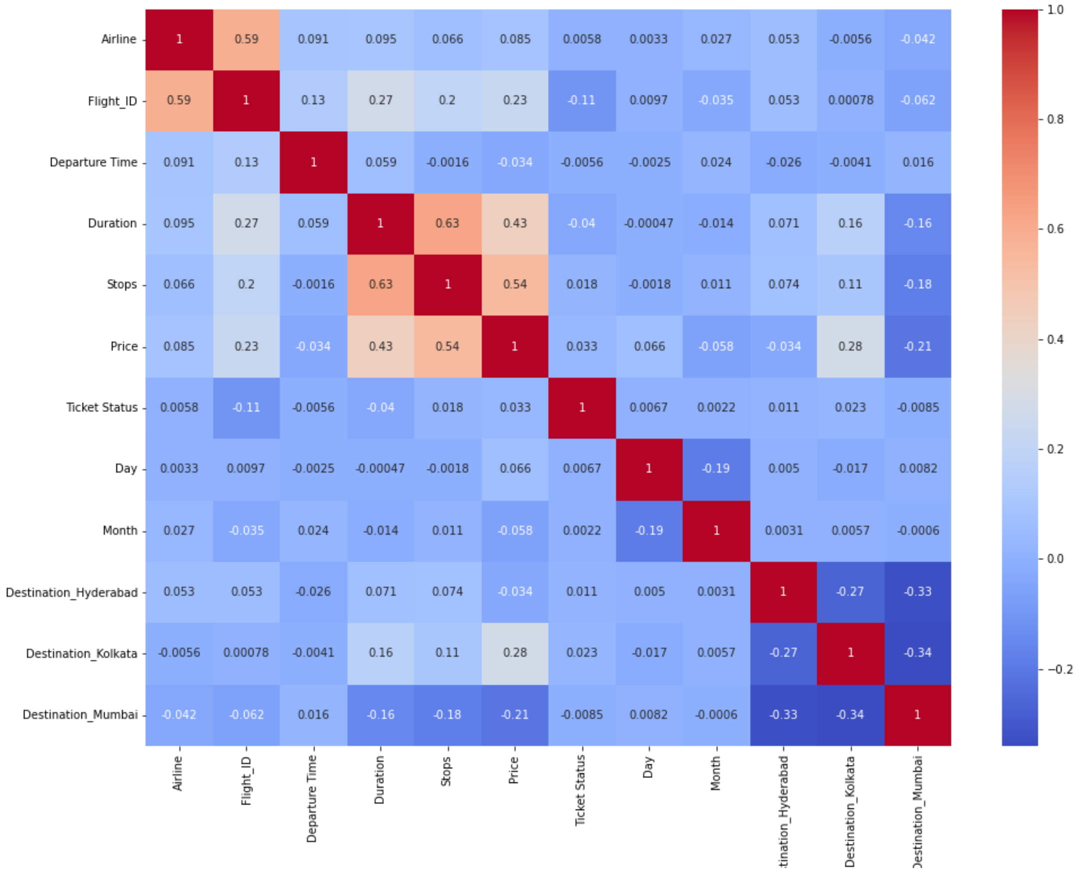
- Let’s see for flight ‘AI-678/641’. This flight is from Delhi to Bangalore with a 1 stop. This is an Air India flight.



Observations:

1. The ticket fare is fairly uniform with some fluctuations in the month of October and December.

- Correlation Heatmap:



Observations:

1. Number of stops and flight duration have the highest correlation with the price.
2. No multicollinearity issues can be seen.

CONCLUSION

Key Findings and Conclusions of the Study

1) Number of stops and flight duration are the features which affect the price most.
2) Indigo Airlines have relatively less price variations as compared to others.
3) In most of the cases it is beneficial to purchase a flight ticket at least a month before travel. The variation of price is lesser in this case.
4) Random Forests Regression Algorithm is giving the best results for prediction of price.

Limitations of this work and Scope for Future Work

1) The hyper-parameters of our final model can be further tuned.
2) In data collection phase we collected data of 10 features and the price. Some more feature data can also be collected. For example, name of the cities where the flight have stops, in-flight miscellaneous services like meals, etc.
3) In this project we worked on only 5 algorithms. There are numerous other regression algorithms with which we can try to make a better model.