

**Flight Price Prediction**

Submitted by:

Bhumik P Shah

**ACKNOWLEDGMENT**

While preparing this case study I had gone through a couple of videos on You tube for reference purpose to understand in depth about the flight prices fluctuating based on demand and supply.

**INTRODUCTION**

* Business Problem Framing

Airlines companies use complex algorithms to calculate flight prices given various conditions present at that particular time. These methods take financial, marketing and various social factors into account to predict flight prices. Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That’s why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

Conceptual Background of the Domain Problem

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 flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases).

* Review of Literature

Because I truly think that sharing sources and knowledges allow to help others but also ourselves, the sources of the project are available at the following link:

[**https://www.researchgate.net/publication/337821411\_Predicting\_Flight\_Prices\_in\_India**](https://www.researchgate.net/publication/337821411_Predicting_Flight_Prices_in_India)

Motivation for the Problem Undertaken

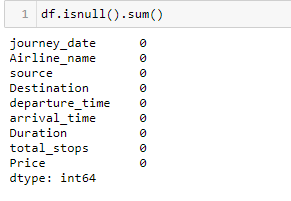
We could inform the travellers with the optimal time to buy their flight tickets based on the historic data and also show them various trends in the airline industry we could help them save money on their travels.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset, including its size, initial patterns in the data and other attributes.

There are no null values present in any of the features in this dataset.



There are outliers present in independent features named as duration, total\_stops, as confirmed by boxplot. Further, I am not removing outliers as total\_stops can be 1 or more depending on the flight connectivity between source and destination i.e whether it is a non stop flight between source and destination or 1 stop flight or 2 stop flight The similar thing is for feature duration i.e time taken to reach destination can be more than 24 hours depending on the total\_stops between source and destination and the layover time. The independent features

Duration and total\_stops are having skewness greater than 0.5 which is treated using power transform method.

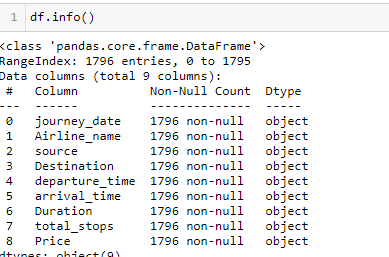
The correlation of every numerical feature with label/target is checked and one of the feature journey\_month is having no correlation with label/target which is price in your case so that feature is dropped from the dataset.

Data Sources and their formats

The data is collected from easemytrip.com for the duration 05.10.2021 to 13.10.2021 for the flights available between Delhi as a source location and Mumbai as a destination location using web scrapping tool selenium.

The dataset contains 1796 records (rows) and 9 features (columns) with price feature being the target attribute.

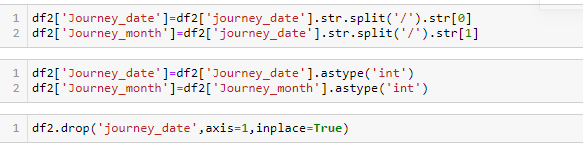
The details of 9 features(columns) are attached below:



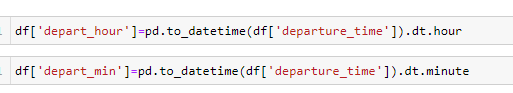
All the features in this dataset are of object datatype.

* Data Pre-processing Done

Extracting out new features named as Journey\_date and Journey\_month from the existing feature journey\_date and finally dropping the feature journey\_date as all the useful information is extracted.

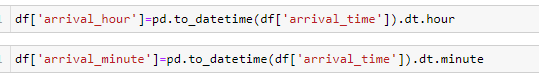


On the similar lines new features depart\_hour and depart\_min are extracted from the existing feature departure\_time and finally dropping the feature departure\_time as all useful information is extracted.





Similarly new features named as arrival\_hour and arrival\_minute are extracted from the existing feature arrival\_time and finally dropping the feature arrival\_time as all useful information is extracted.

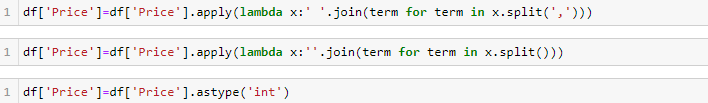




The feature duration i.e time taken to reach from source to destination is having hours and minutes in records so the same is converted entirely into minutes from hours say for example if it is 2 hours and 15 minutes it is converted into 135 minutes for better visualization purpose.

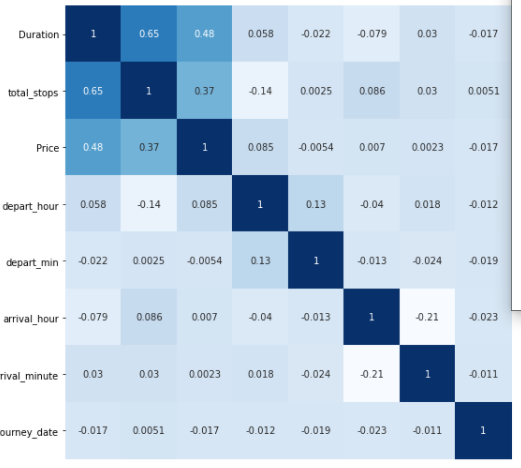


The feature price which is the target attribute in this dataset is having object datatype which is converted into integer datatype as price is always a numerical value.



* Data Inputs- Logic- Output Relationships

As per the correlation details attached below the target attribute price is having positive linear correlation with features Duration and total\_stops The feature journey\_month is dropped from the dataset as the entire data is extracted for the month of October so there is no variation in the data resulting into no correlation with target attribute price.



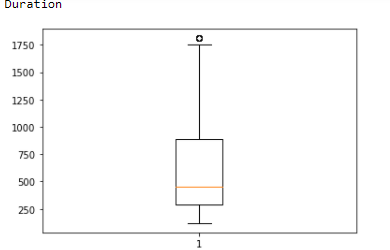
* Hardware and Software Requirements and Tools Used

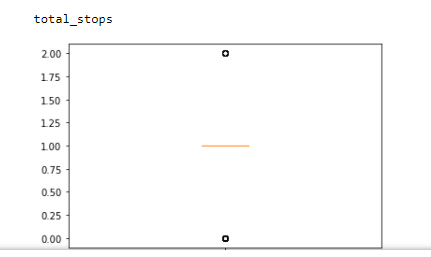
I have used Jupyter Notebook for writing python codes. Further, libraries such as Pandas, Seaborn, Matplotlib, SKlearn, NumPy are used.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

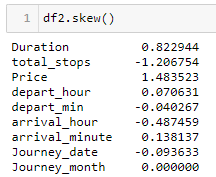
The shape of the dataset, datatypes, null values present in the dataset or not are amongst few things that are first to be known. There are no null values present in this dataset. Further, next step is to go for outlier detection if any using box plot and distribution plot. The features Duration and total\_stops are having outliers as confirmed from the boxplot.

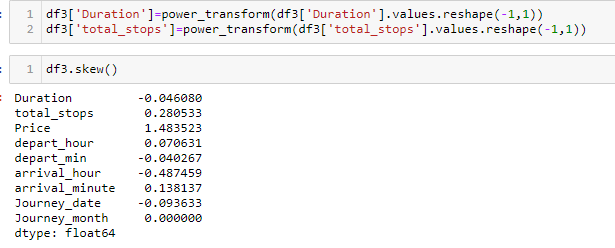




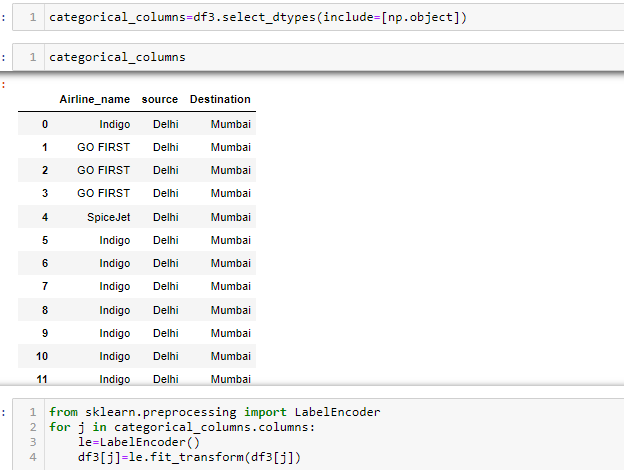
Further, the outliers are not treated as the total\_stops can be 1 or more and the same is for duration feature as the duration to reach from source to destination can be more than 24 hours.

After treating outliers next step is to go for skewness check .The features duration and total\_stops are having skewness greater than 0.5 which is treated using power transform method.





The categorical features are converted into numeric type using label encoder as machine learning models only accepts numerical values.



After completing all the above mentioned steps dataset is splitted into two variables x and y.



Standard Scaling is to be applied first before splitting the data into train\_test\_split.

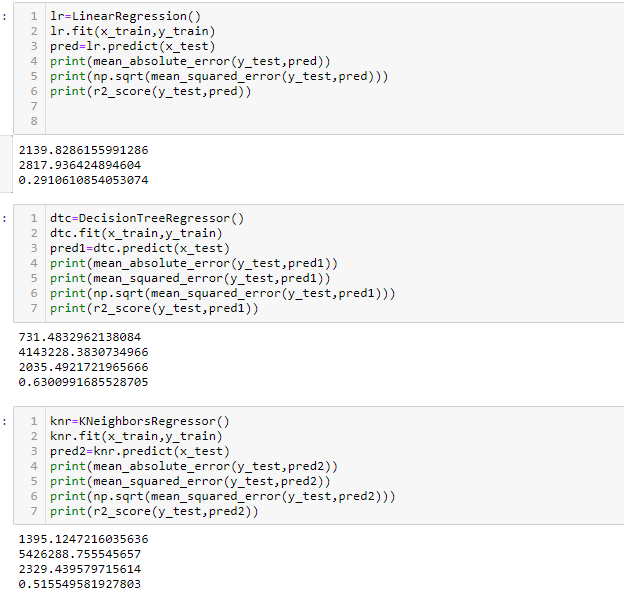


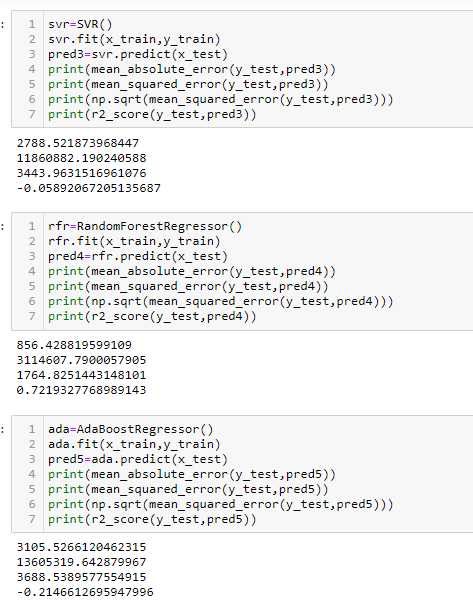
* Testing of Identified Approaches (Algorithms)

The various algorithms like Linear Regression, Decision Tree Regressor, SVR, KNeighborsRegressor, Ada Boost Regressor, Gradient Boosting Regressor, Random Forest Regressor, Lasso , Ridge are used. First, approach is to find r2 score using all mentioned algorithms. Secondly, cross validation score of all the above mentioned algorithms are checked .The difference between r2 score and cross validation score is minimum for KNeighborsRegressor. KNeighborsRegressor is my best model and to obtain better r2\_score hypertuning is carried out.

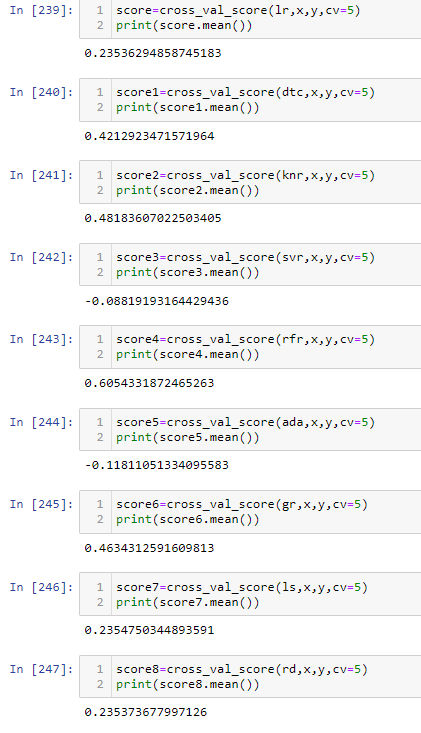
* Run and Evaluate selected models

Kindly find attached below the snapshots of various algorithms used

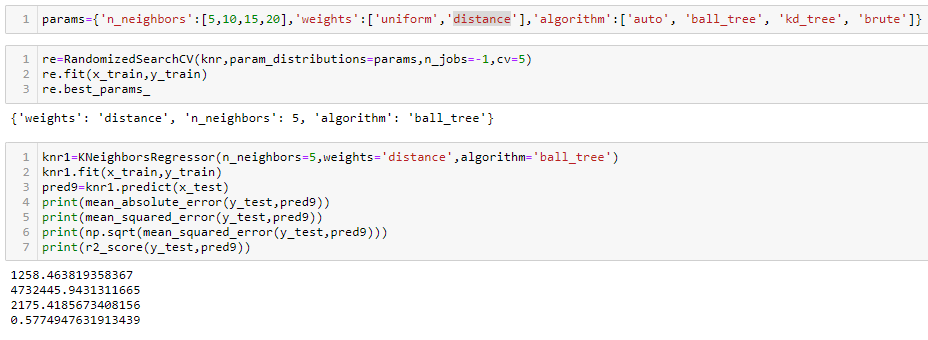








Hyper tuning of KNeighborsRegressor, is carried out to find out if the r2\_score can be improved or not.



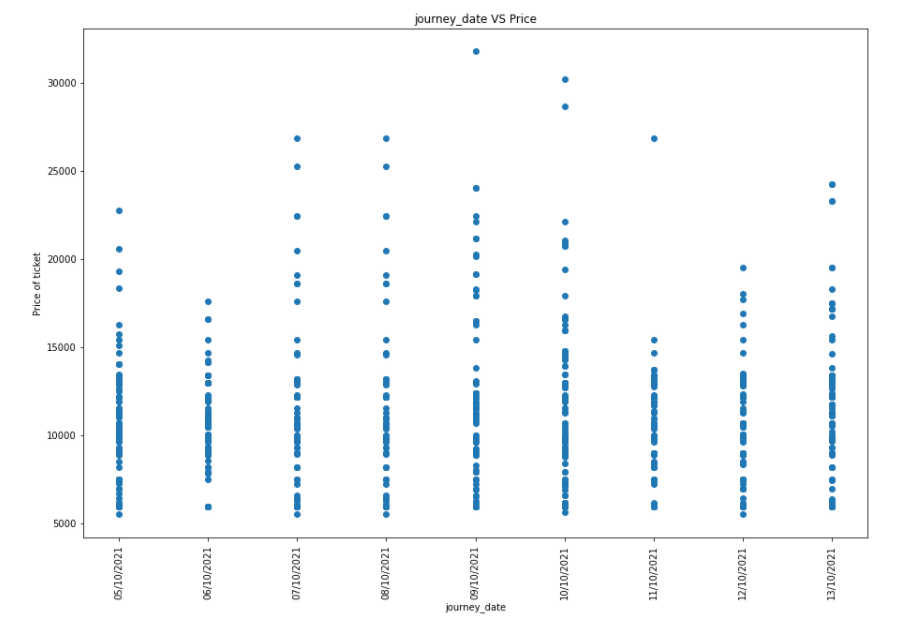
**KNeighborsRegressor is my best model with r2\_Score of 58%.**

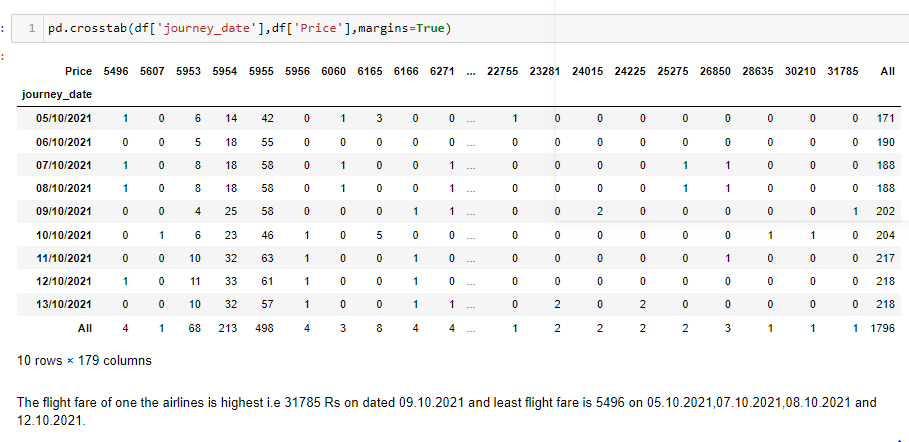
Key Metrics for success in solving problem under consideration

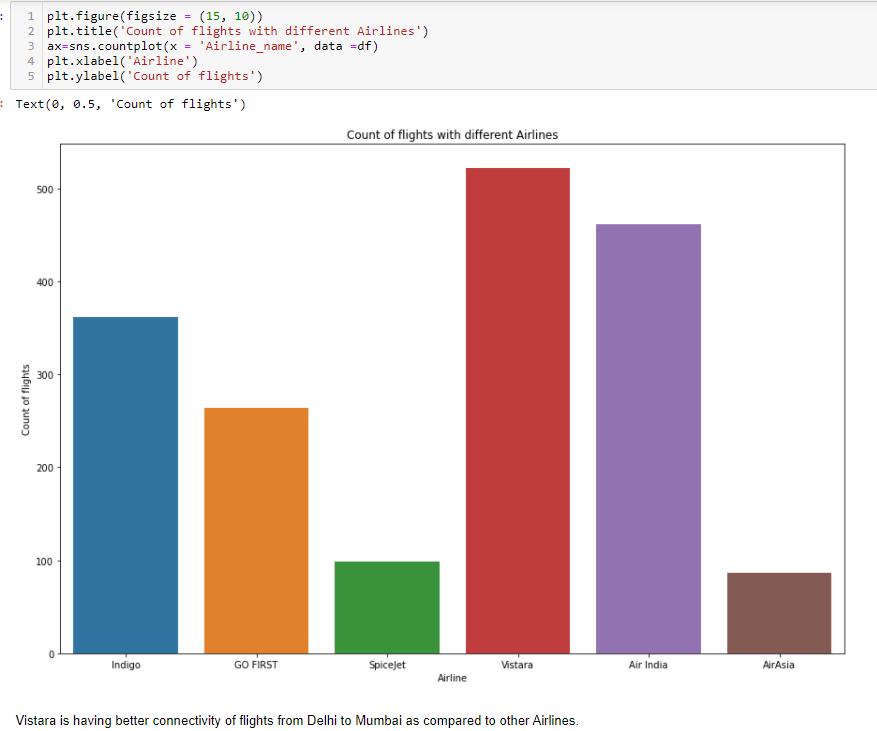
The feature duration is converted into minutes from hours so that better visualization can be carried out.

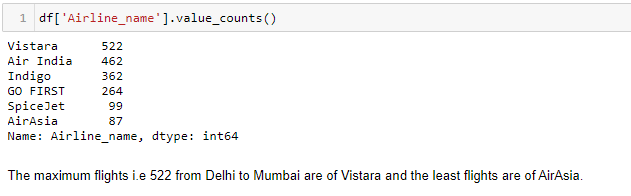
* Visualizations

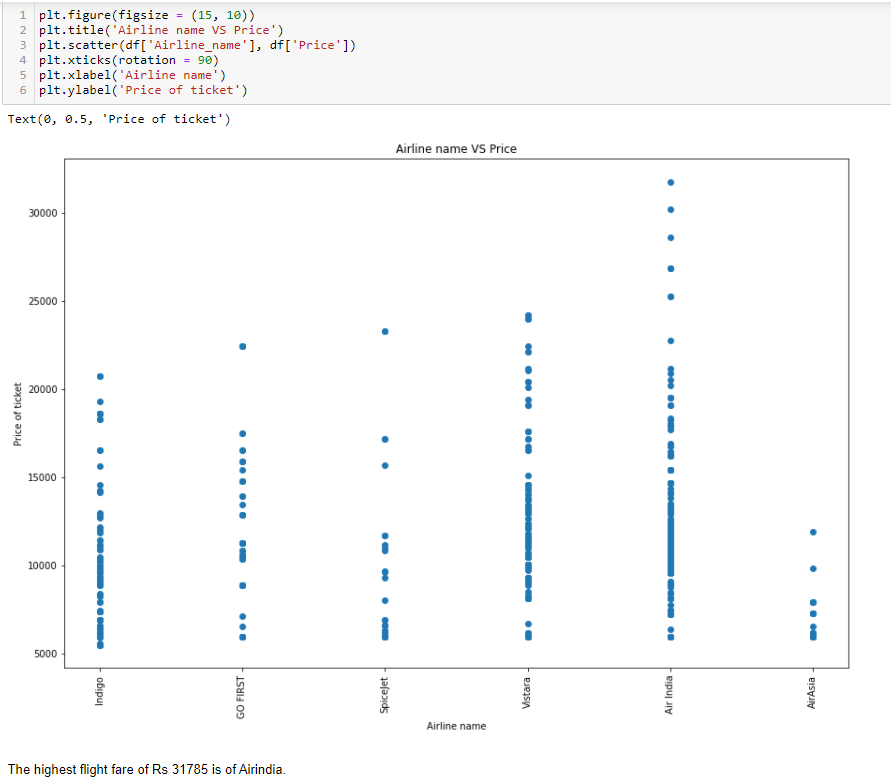
Data Visualization is the graphical representation of the information and data. By using visual elements like charts, graphs and maps data visualization tools provide an accessible way to see and understand trends, outliers and patterns in data.

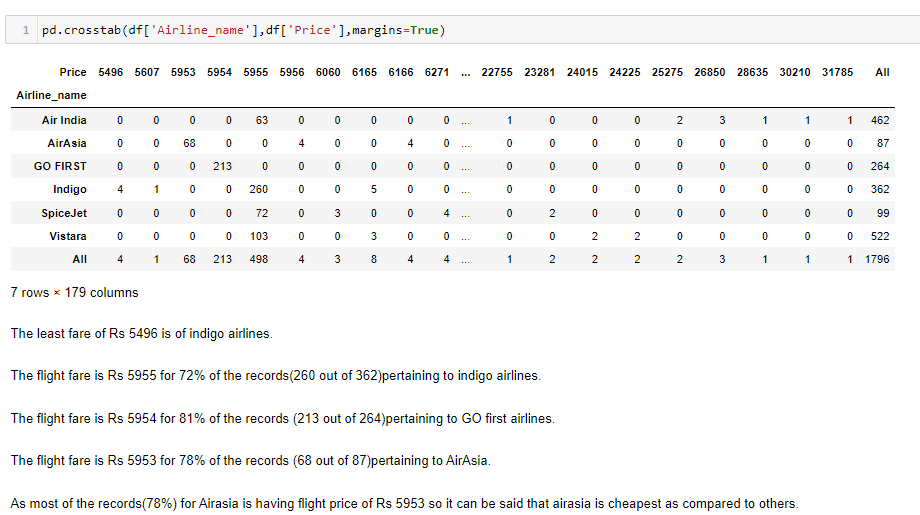


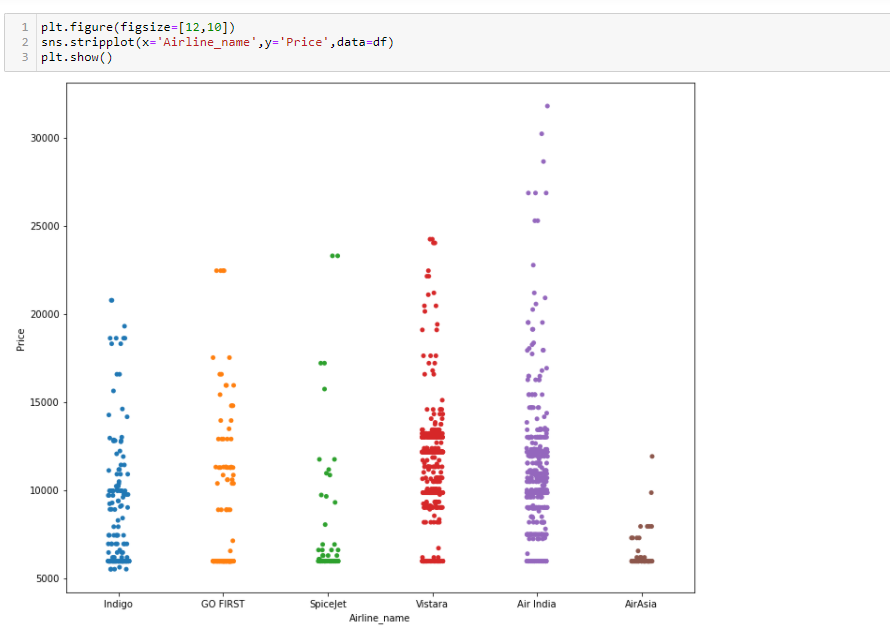


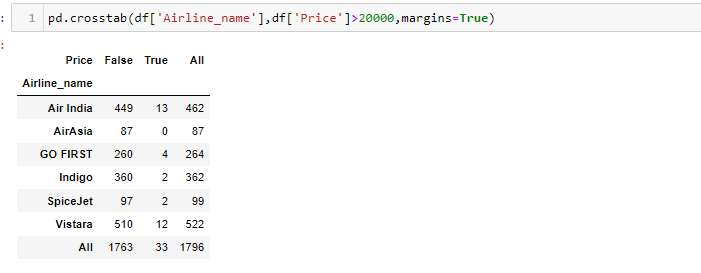


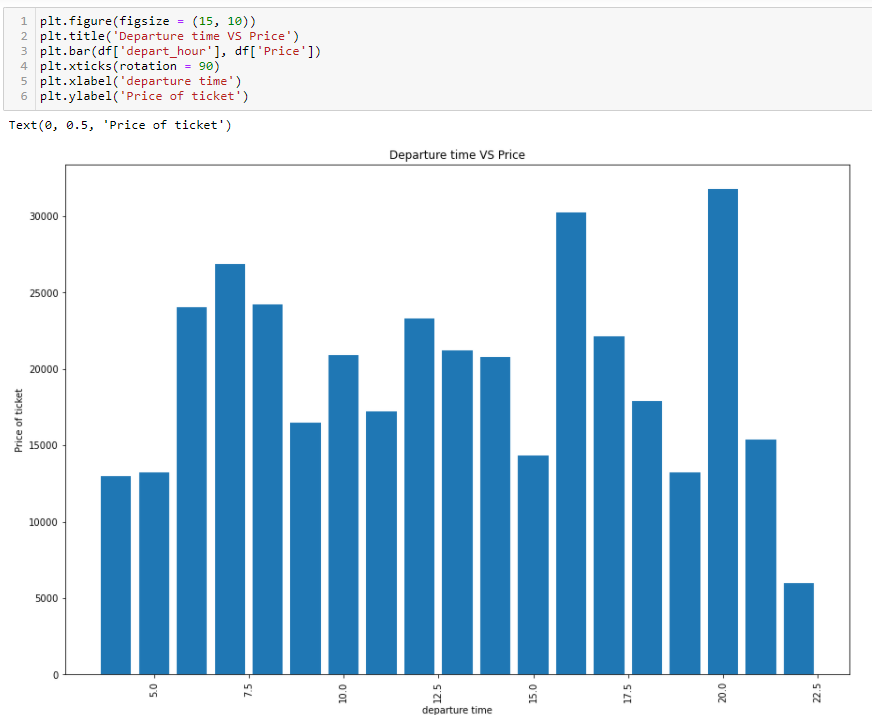


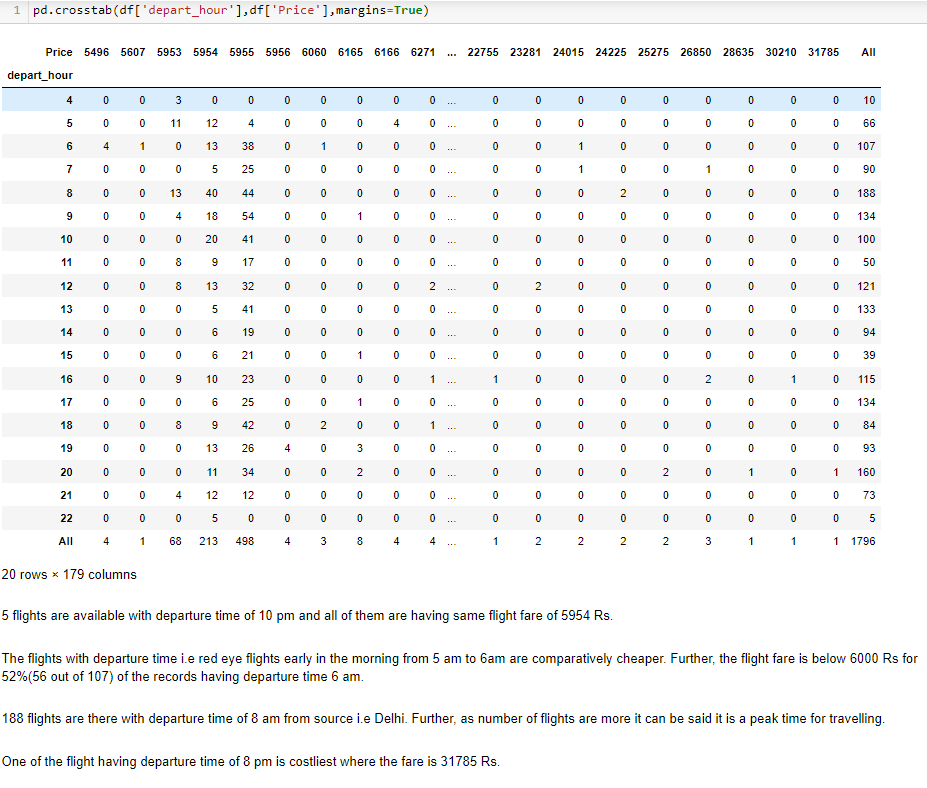


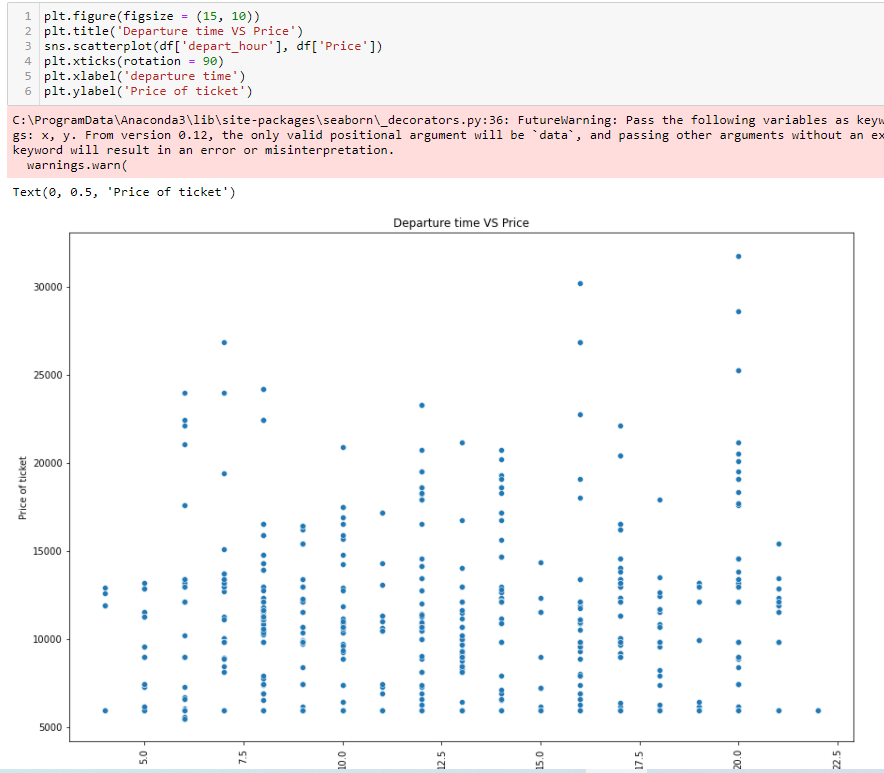


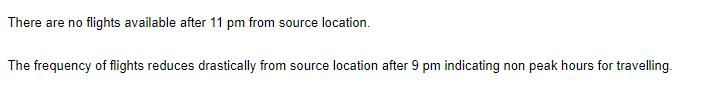


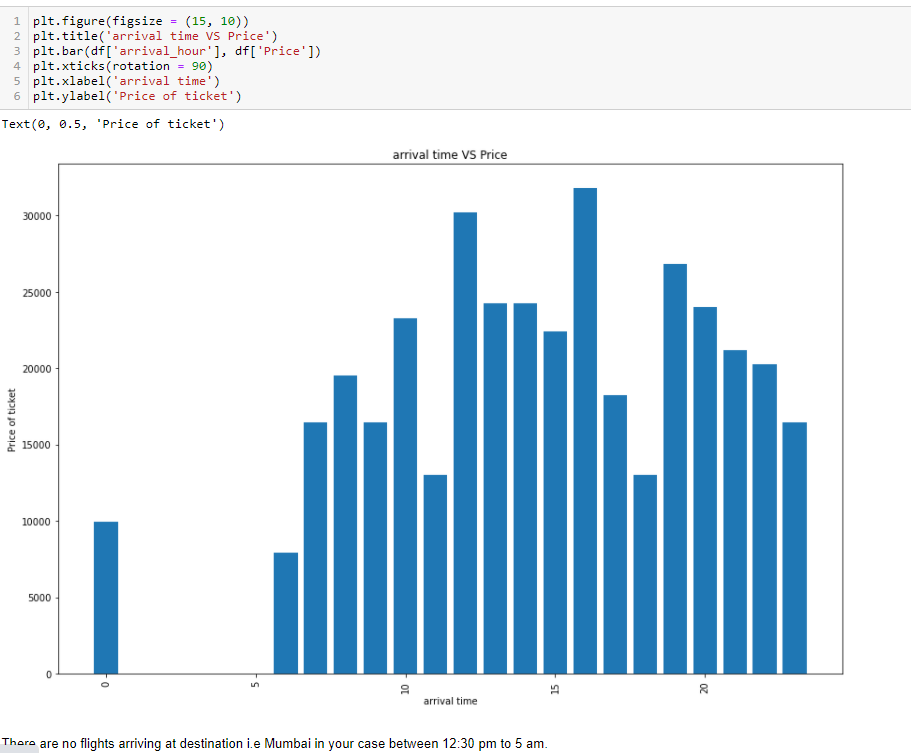


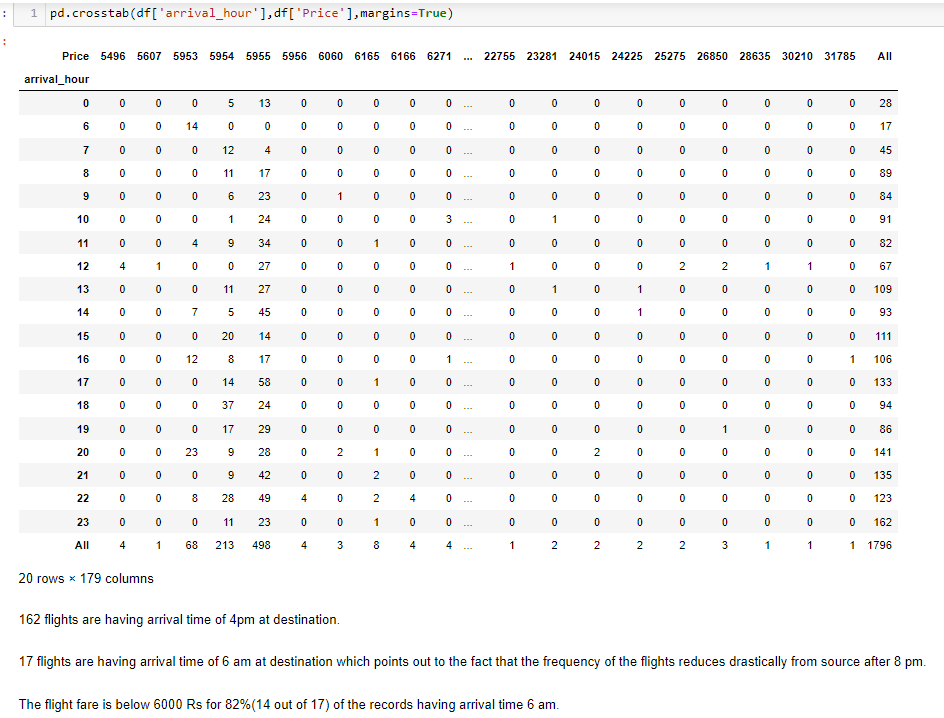


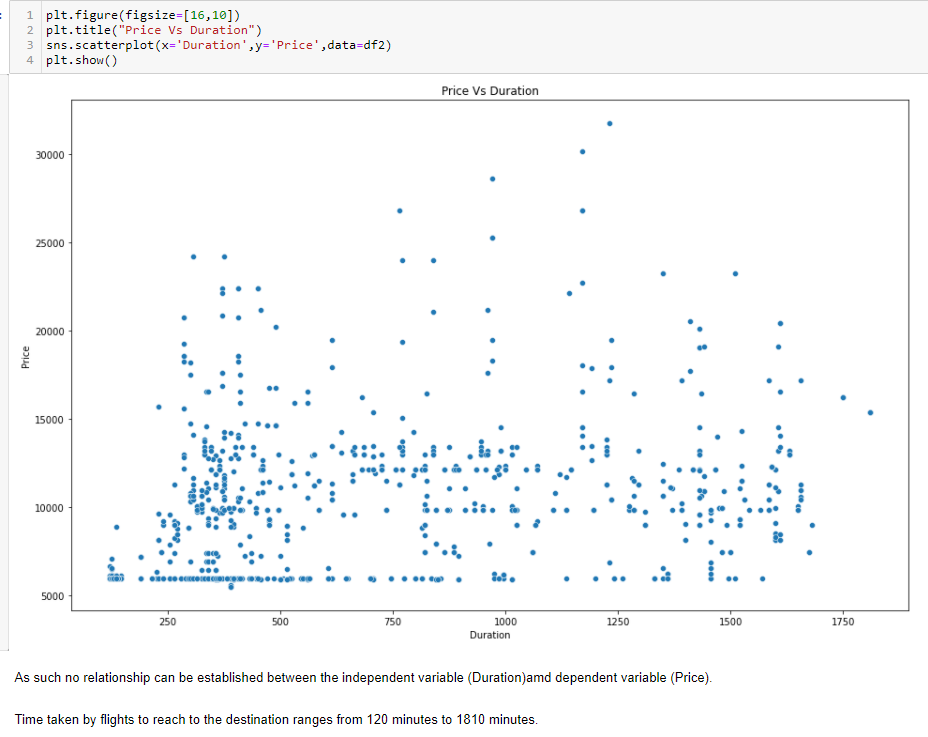


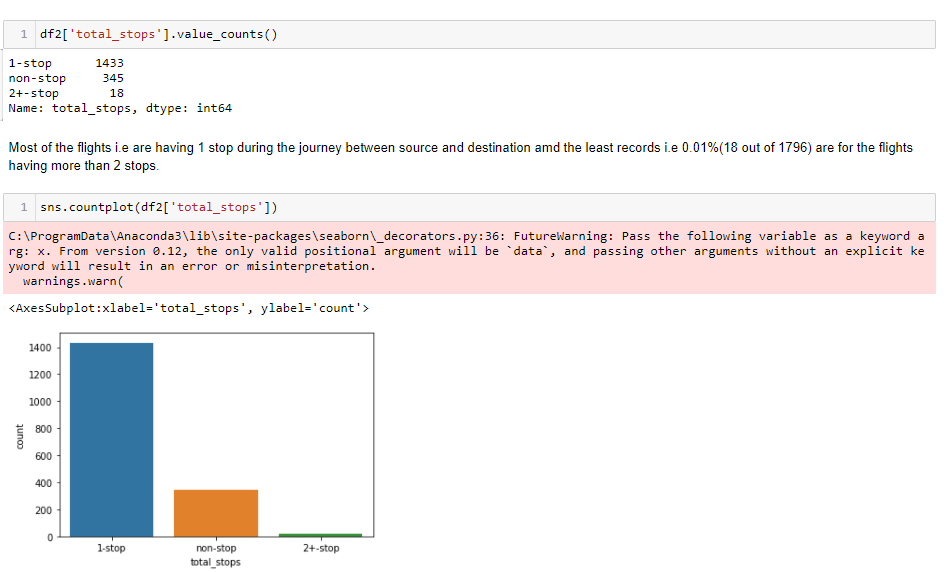


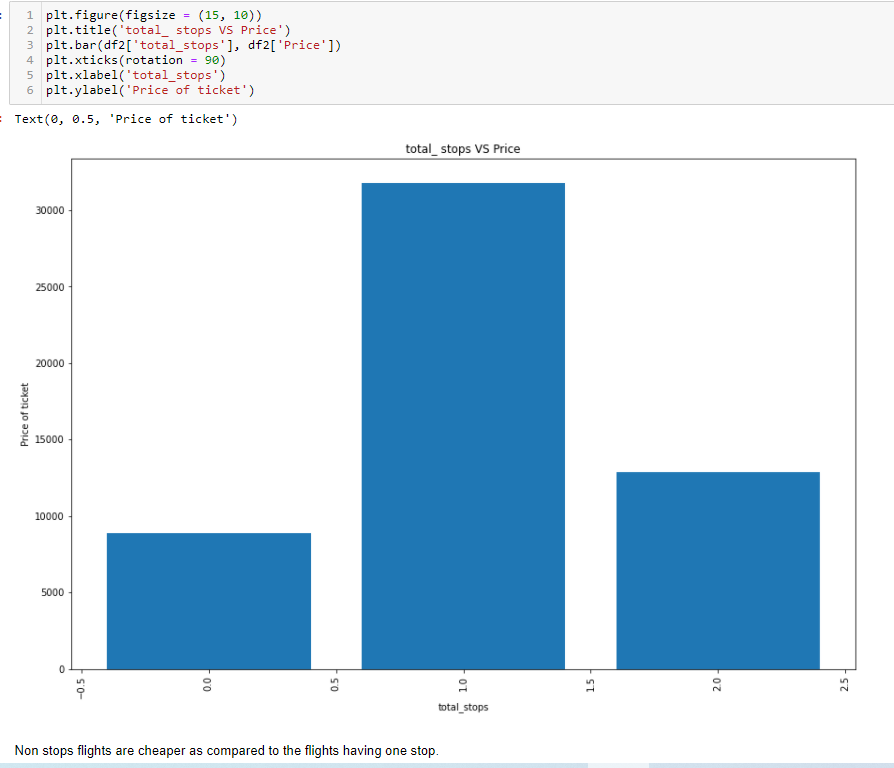


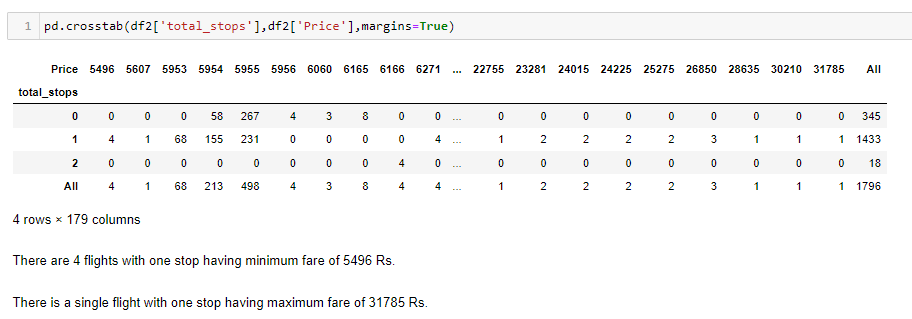


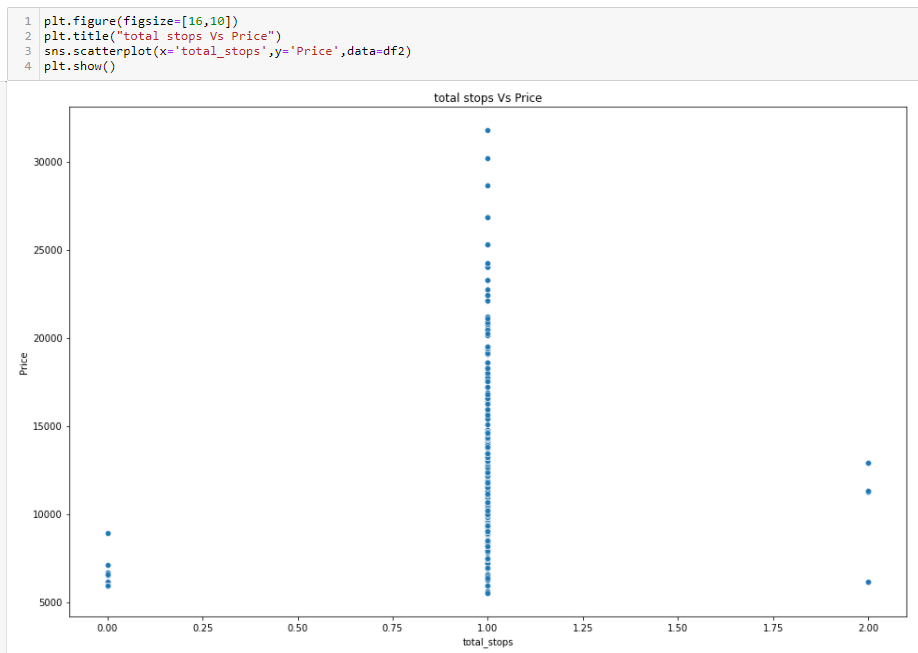


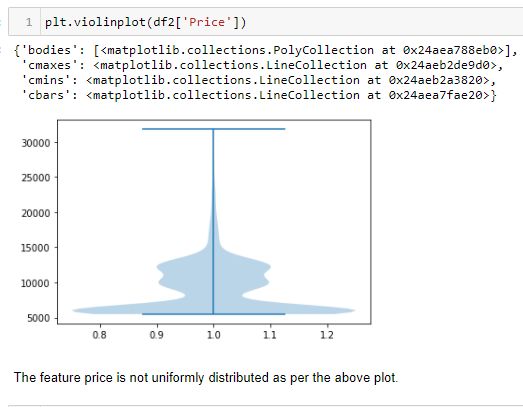


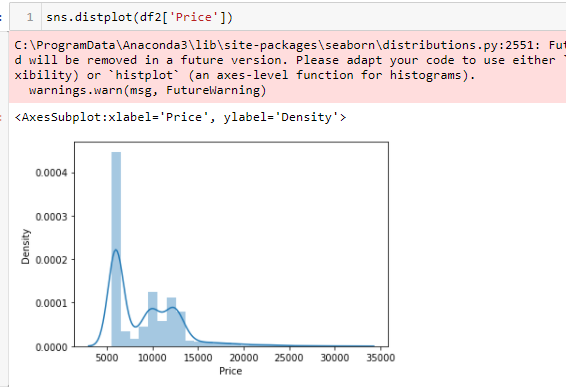














* Interpretation of the Results

From visualization the key findings like skewness, outliers present in the dataset were known. Further, the heat map was drawn showing the correlation with target feature and based on it the one’s having less/no correlation with the target feature can be known and if the need arises can be dropped.

Vistara is having better connectivity of flights( 522) from Delhi to Mumbai as compared to other Airlines and the highest flight fare of Rs 31785 is of AirIndia.

As most of the records(78%) for Airasia is having fare of Rs 5953 so it can be said that airasia is cheapest as compared to others.

The least ticket fare is of Rs 5496 for one of the flight of indigo airlines. Further, there are in total 8 records for Air india having fare of Rs 25000 and above.

The flights with departure time i.e red eye flights early in the morning from 5 am to 6am are comparatively cheaper. Further, the flight fare is below 6000 Rs for 52%(56 out of 107) of the records having departure time 6 am.

There are no flights available after 11pm from source location and the frequency of flights reduces drastically from source location after 9 pm.

The flight fare is below 6000 Rs for 82%(14 out of 17) of the records having arrival time 6 am.

The non stop flights are cheaper as compared to one stop flights.

**CONCLUSION**

* Key Findings and Conclusions of the Study

KNeighborsClassifier is my best model with r2 score of 58%

.

* Learning Outcomes of the Study in respect of Data Science

This case helped me how to hyper tune the parameters using RandomizedSeachCV to get the best accuracy from the model.

* Limitations of this work and Scope for Future Work

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 . This will train the model better giving better accuracies and more savings.

More rules can be added in the Rule based learning based on our understanding of the industry, also incorporating the offer periods given by the airline.

Develop a more user friendly interface for various routes giving more flexibility to the users.