

FLIGHT PRICE PREDICITON PROJECT

Submitted by:

Asha Vinod Kumar

# ACKNOWLEDGMENT

Firstly, I would like to thank FlipRobo Technologies for giving me the opportunity to work on this project. Also, I would like to thank the DataTrained team, for providing me the knowledge and guidance which helped me a lot to work on this project.

# INTRODUCTION

* Business Problem Framing

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, you 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.

* Conceptual Background of the Domain Problem

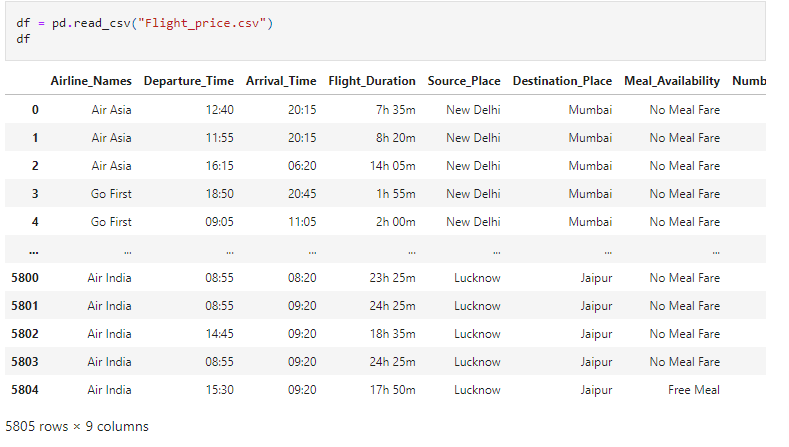
Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time? What is the best time to buy so that the consumer can save the most by taking the least risk? Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

* Motivation for the Problem Undertaken
  1. The objective behind to take this project is to harness the required data science skills.
  2. Improve the analytical thinking.
  3. Get into the real-world problem-solving mechanics.

# Analytical Problem Framing

* Data Sources and their formats

In this section we have to scrape the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc). The number of columns for data doesn’t have limit, it’s up to us and our 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, you can add or you can remove some columns, it completely depends on the website from which we are fetching the data.

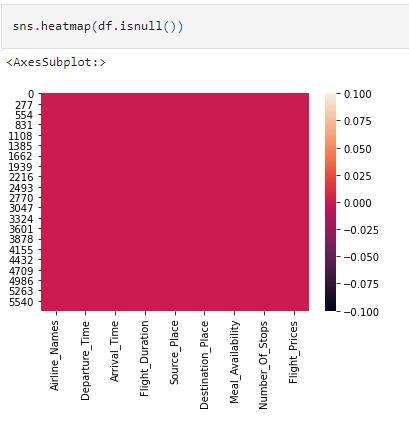
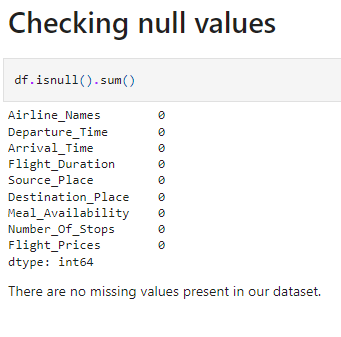


* Data Pre-processing Done

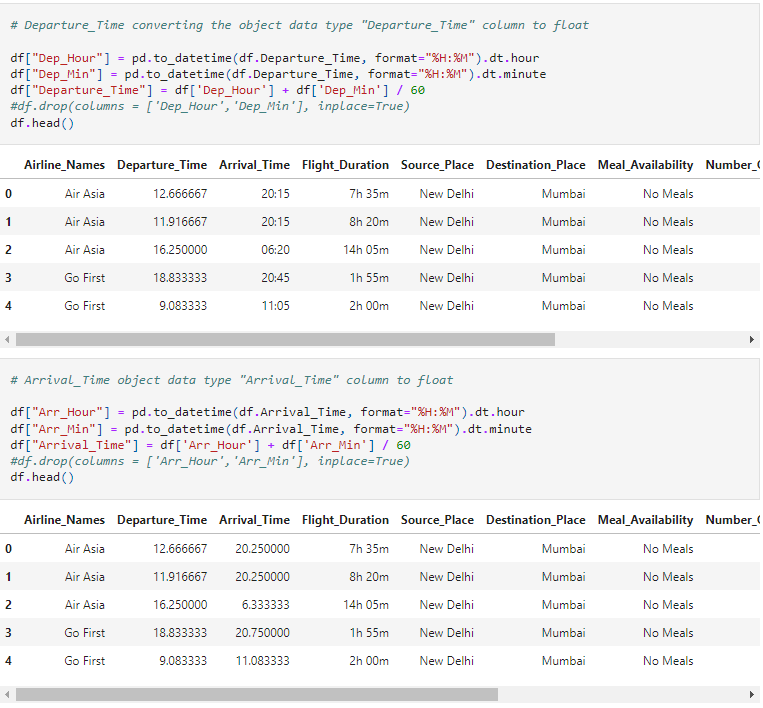
Below are the steps which we have taken in data pre - processing:

* + - Null Values:

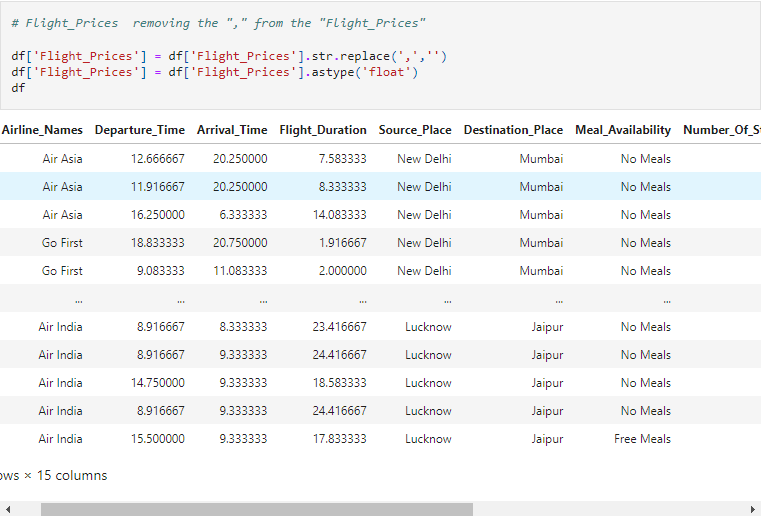
We checked for the null values (missing values) and found that there is no null values in the given dataset.



* + - Data Cleaning:
    1. We are renaming meal availability column for no meal fare as no meals, free meal as free meals and ecash 250 as ecash meals.
    2. Also renaming the column number\_of\_stops values non stop=0, I stop=1, 2 stops=2, 3 stops=3 and 4 stops-4 for easy reading and convenience.
    3. Converting departure time and arrival time to float vales from object values.

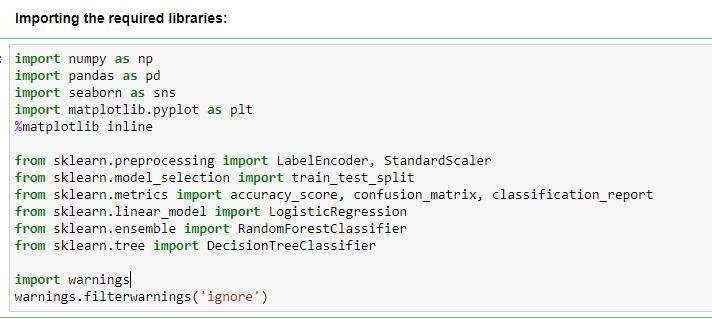


* + 1. Removing ‘,’ from flight\_prices.

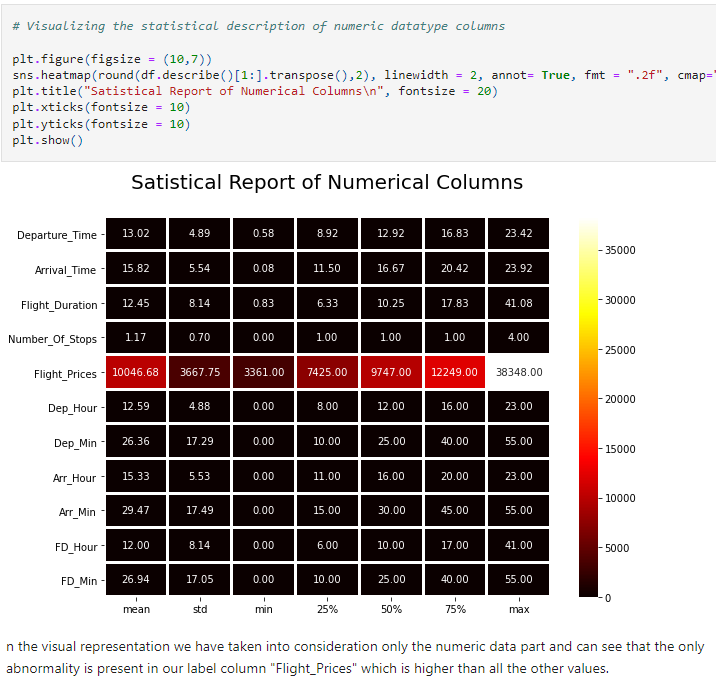


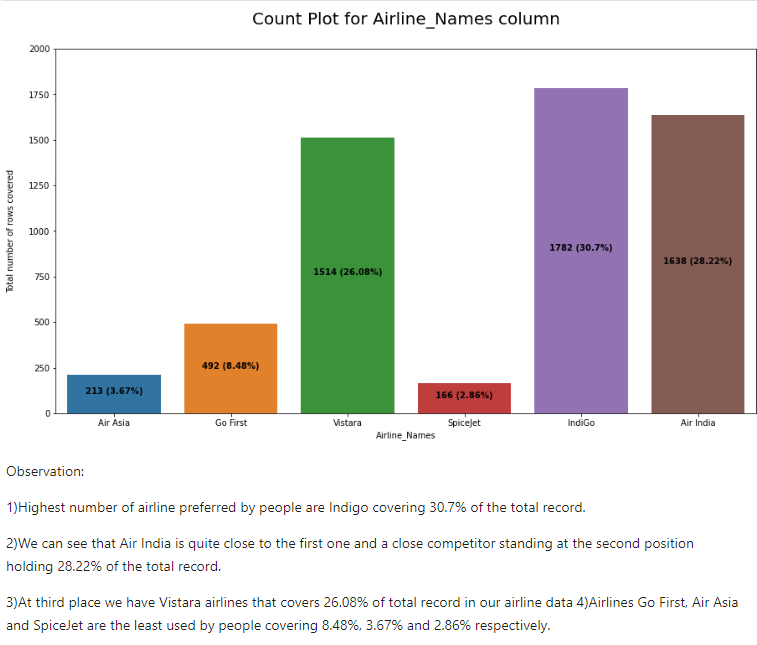
* Data Inputs- Logic- Output Relationships

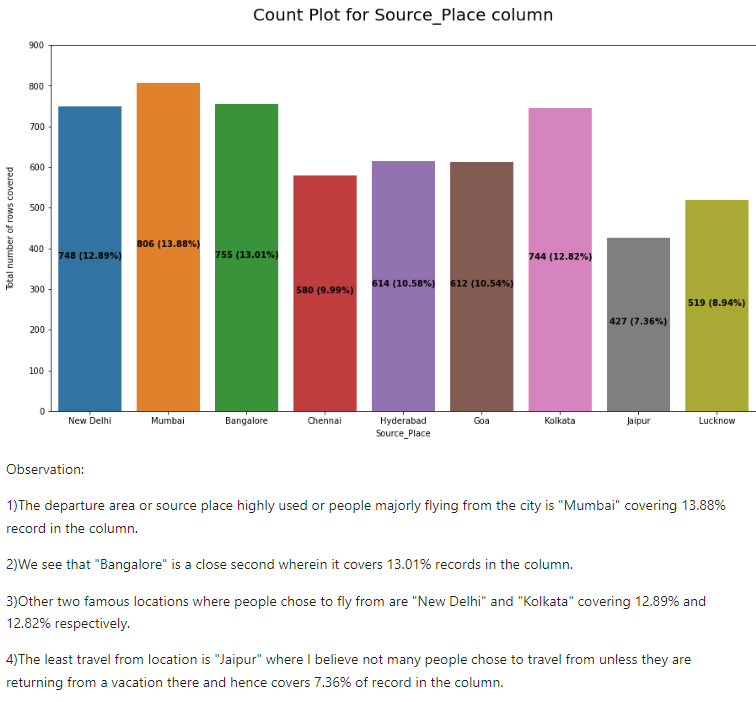
EDA was performed by creating valuable insights using various visualization libraries.

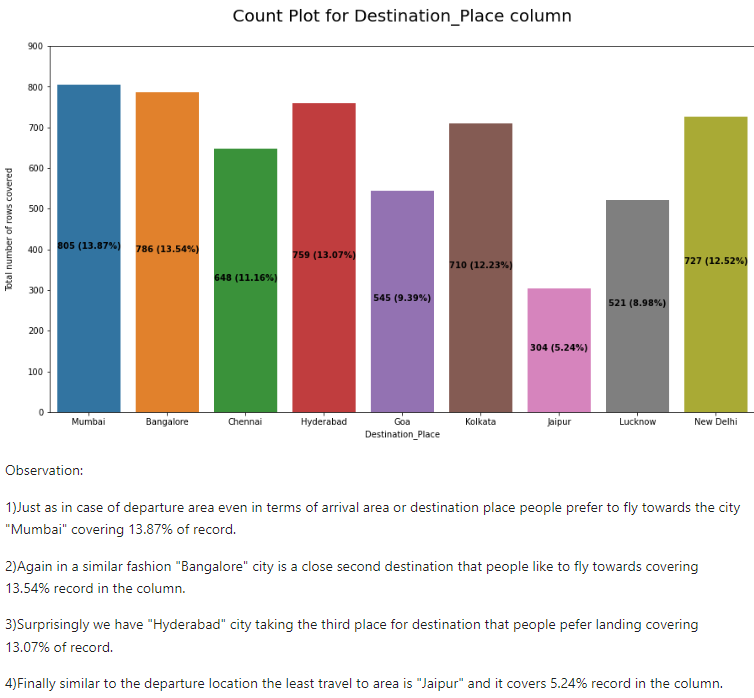


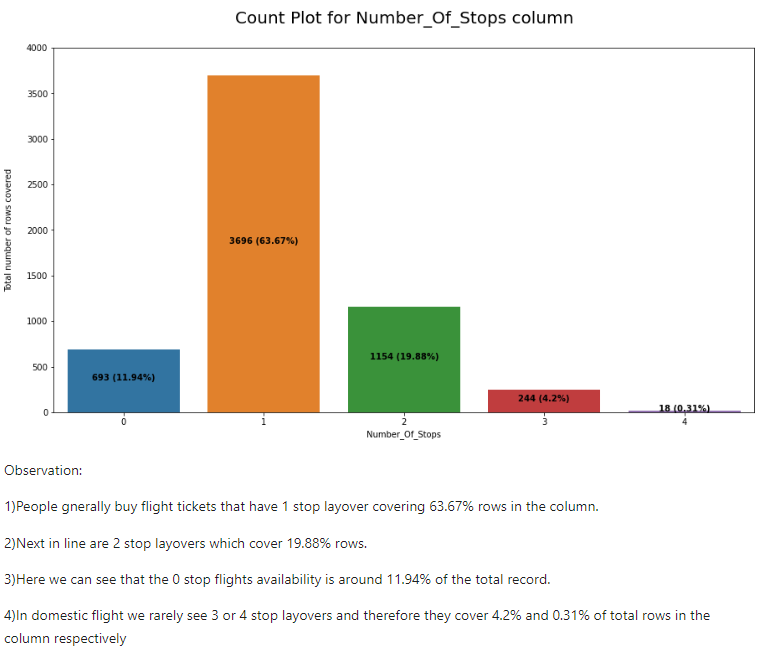
Visualizing Statistical reports of Numerical Columns



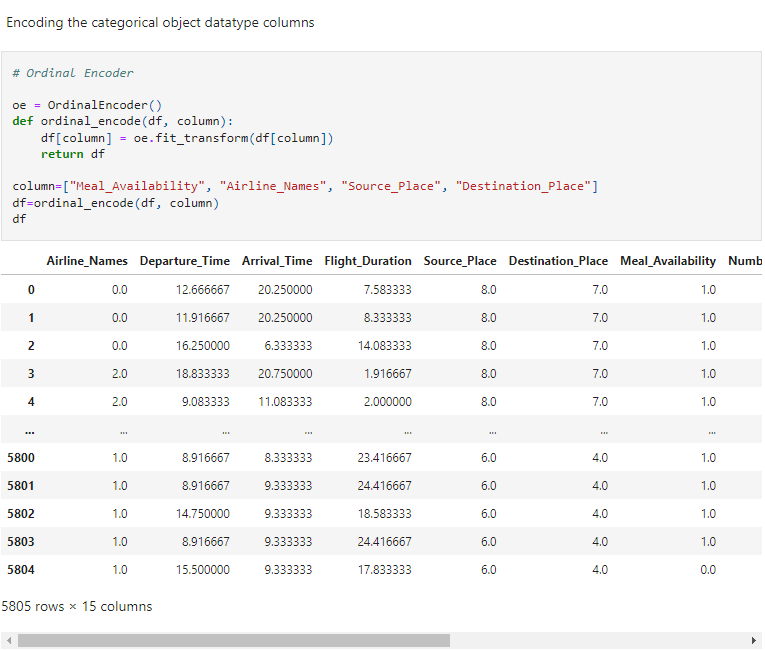




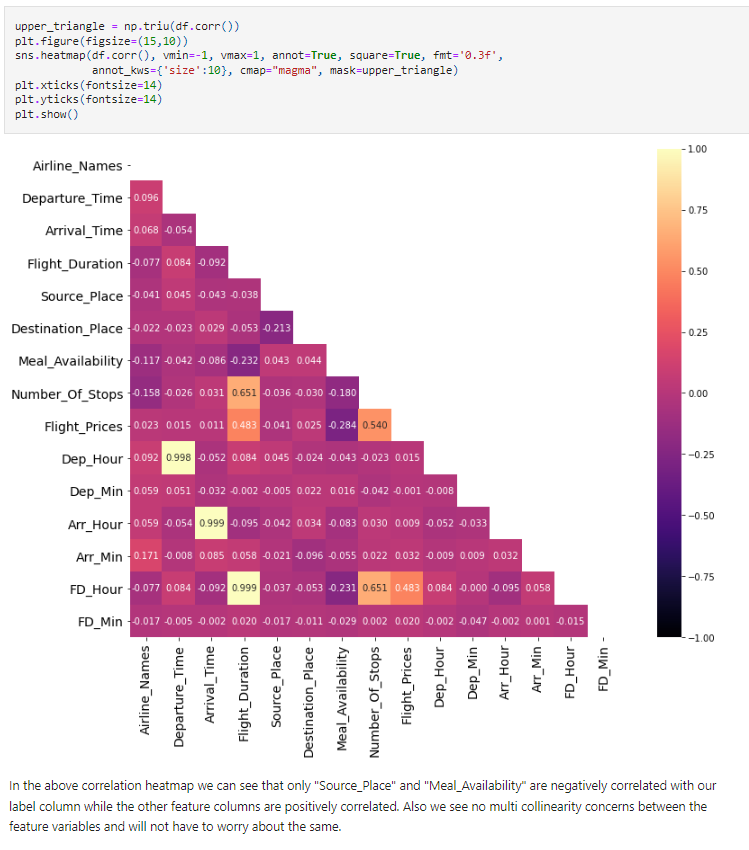


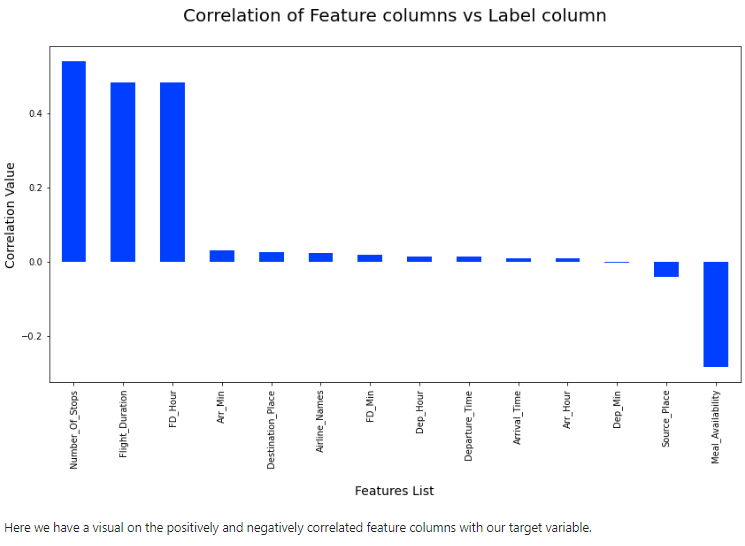


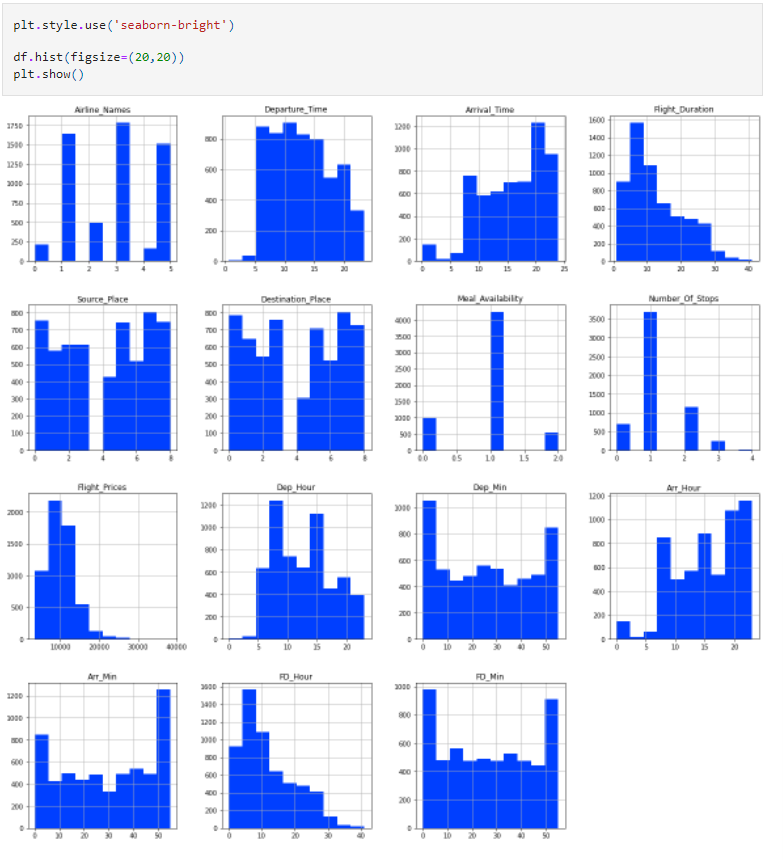
Encoding Categorical Column



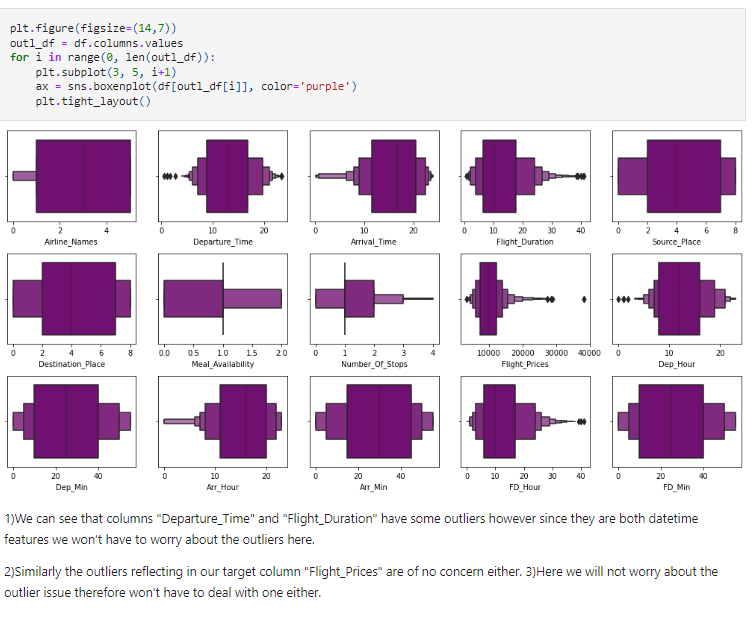
Correlation using heatmap



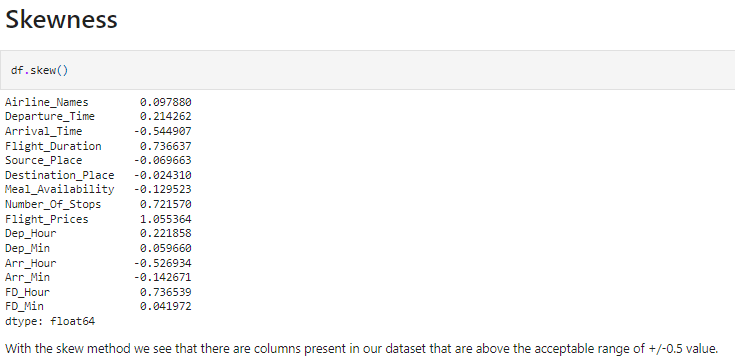




Identifying Outliers:



Data Skewness:



* Hardware and Software Requirements and Tools Used

**Hardware Configuration:**

**Operating System:** Windows 10

**System Type:** 64-bit operating system, x64-based processor

**Processor:** I have used i3 processor with 4GB RAM as hardware.

**Software & Tools:**

* + 1. Jupyter Notebook (used as a notebook to code)
    2. Python (used for scientific computation)
    3. Pandas (used for scientific computation)
    4. Numpy (used for scientific computation)
    5. Matplotlib (used for visualization)
    6. Seaborn (used for visualization)
    7. Scikit-learn (used as algorithmic libraries)

**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods)

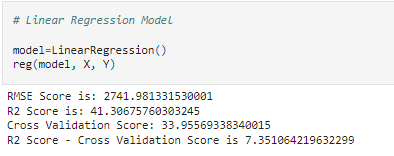
* + - Performed EDA (Exploratory Data Analysis).
    - Data Cleaning and dropping the columns which were not contributing to the dataset.
    - Checked for the outliers and tried to remove the outliers of the dataset.
    - Checked for the skewness in the dataset and removed the skewness for better model building.
    - Train- Test the dataset into independent and dependent variables.
    - Model Building.
    - Cross validation score to check if the model is over-fitted.
* Testing of Identified Approaches (Algorithms)

Below are the algorithms used for the training and testing:

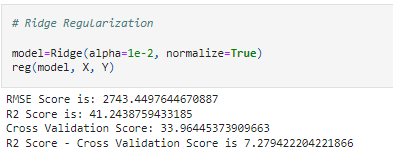
* + 1. Logistic Regression.
    2. Ridge Classifier.
    3. Lasso Classifier.
    4. Decision Tree Classifier.
    5. Random forest Regressor.
    6. K Neighbours.
    7. Gradient Boosting Regressor.
    8. ADA Boost Regressor.

9. Extra Trees Regressor.

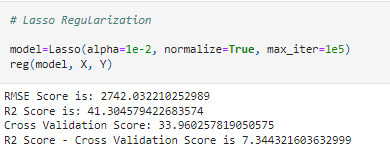
* Run and Evaluate selected models
  1. **Logistic Regression:**



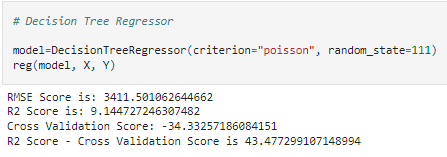
**2.Ridge Classifier:**



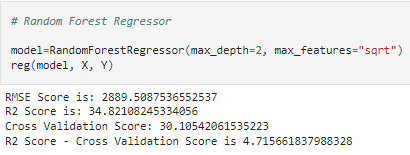
**3.Lasso Classifier:**



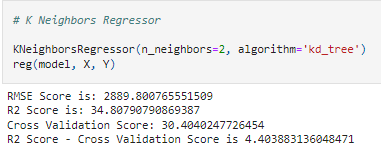
**4.Decision Tree Classifier:**



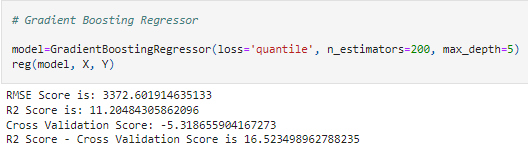
**5.Random Forest Regressor:**



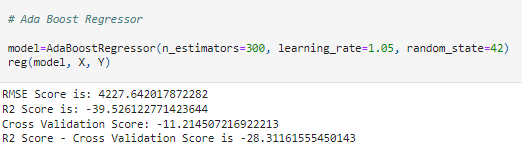
**6. K Neighbours:**



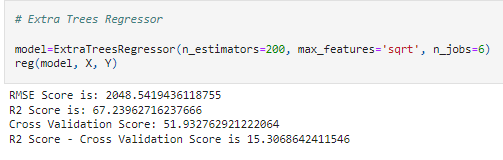
**7. Gradient Boosting Regressor:**



**8. ADA Boost Regressor:**



**9.Extra Tress Regressor:**

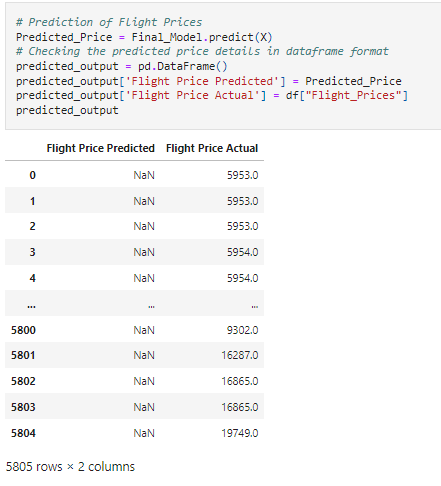
****

* Key Metrics for success in solving problem under consideration

The key metrics used are as follows:

* + 1. RMSE Score
    2. Confusion Matrix
    3. Classification Report
    4. R2 Score
    5. Precision & Recall
    6. Cross validation score

# CONCLUSION

* Key Findings and Conclusions of the Study
* 
* Learning Outcomes of the Study in respect of Data

Science

This project helped me to work on the real time industrial data, which helped me to gain the real time experience. In the project I got to work on the different type of algorithms and fitting the best model based on the accuracy score and cross validation score.

After comparing all the regression models I have selected Extra Trees Regressor as my best model and have listed down it's parameters above referring the sklearn webpage. Here I am using the Grid Search CV method for hyper parameter tuning my best model. And I have trained the Grid Search CV with the list of parameters I feel it should check for best possible outcomes.

Here the Grid Search CV has provided me with the best parameters list out of all the combinations it used to train the model.