

FLIGHT PRICE PREDICTION PROJECT

BY

GIRIJA CHANDRA MOHAN

**ACKNOWLEDGMENT**

I would like to express my gratitude to my guide Khusboo Garg (SME, Flip Robo) for her constant guidance, encouragement and unconditional help towards the development of this Used Car Price Prediction project. She helped me whenever I got stuck somewhere in between. The project would have not been completed without her support and confidence.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1) <https://www.google.com/>

2) <https://www.youtube.com/>

3) <https://scikit-learn.org/stable/user_guide.html>

4) <https://github.com/>

5) <https://www.kaggle.com/>

6) <https://medium.com/>

7) <https://towardsdatascience.com/>

8) <https://www.analyticsvidhya.com/>

9) <https://www.expedia.com/>

**Business Problem Framing**

Machine Learning is a field of technology developing with immense abilities and applications in automating tasks, where neither human intervention is needed nor explicit programming.

The power of ML is great that we can see its applications trending almost everywhere in our day-to-day lives. ML has solved many problems that existed earlier and have made businesses in the world progress to a great extent.

Today, we will go through one such practical problem and build a solution (model) on our own using ML.

We are about to deploy an ML model for Flight price prediction and analysis. This kind of system becomes handy for many people. So, to be clear, this model will provide you will the approximate price for your flight based on the flight name, timing of the flight, destination and source, the number of stops.

**Conceptual Background of the Domain Problem**

Airline 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 the prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

**Review of Literature**

As per the requirement of client, we have scrapped the data from online sites and based on that data analysis was done to find which feature affects the prices and checked the relationship of flight price with all the feature like what flight should be chosen.

**Motivation for the Problem Undertaken**

I have worked on this on the bases of client requirements and followed all the steps to model deployment.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem**

In our scrapped dataset, our target variable "Flight\_Prices " is a continuous variable. Therefore, we will be handling this modelling problem as regression.

This project is done in three parts:

* Data Collection
* Data Analysis
* Model Building

1. Data Collection

You have to scrape at least 1500 rows of data. You can scrape more data as well, it’s up to you, More the data better the model. In this section you 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 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. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

2. Data Analysis

After cleaning the data, you have to do some analysis on the data. 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?

3. Model Building

After collecting the data, you 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.

Follow the complete life cycle of data science. Include all the steps like

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

**Data Sources and their formats**

I have collected from expedia website using web scrapping.

There are more than 5500 observations and 8 features including the target feature fare in dataset.

The dataset is in the form of CSV (Comma Separated Value) format and consists of 8 columns with 5590 number of records as explained below:

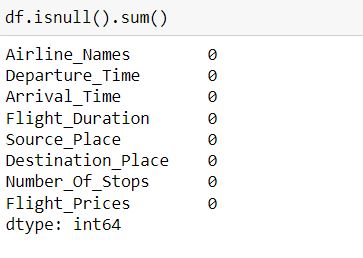
* Airline\_Names : This shows the list of all the Airline Names for which the data got scraped
* Departure\_Time : In this column we have the timings of every flight departure
* Arrival\_Time : Here in this column we have the timings of every flight arrival
* Flight\_Duration : We can see the total duration of a flight that it took to fly from the source to the destination
* Source\_Place : Gives us the name of the source place where the flight journey began
* Destination\_Place : Shows us the name of the destination place where the flight journey ended
* Number\_Of\_Stops : Lists the number of stops the flight is going to take to complete the entire journey
* Flight\_Prices : Finally we have our label column that has the ticket prices for the aircraft journey

We can see our dataset includes a target label "Flight\_Price" column and the remaining feature columns can be used to determine the price of the flights. Since price is a continuous value it makes this a Regression problem.

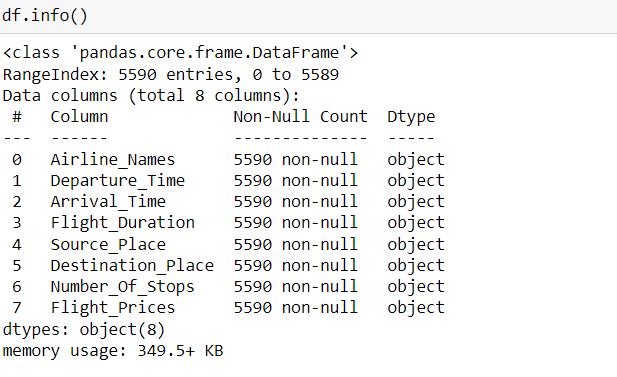


Data Preprocessing Done

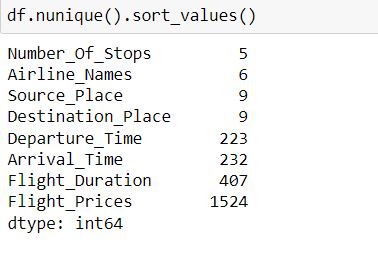
For the data pre-processing step, We checked the dataframe for missing values, imputed records with “-“ using various imputing techniques to handle them.



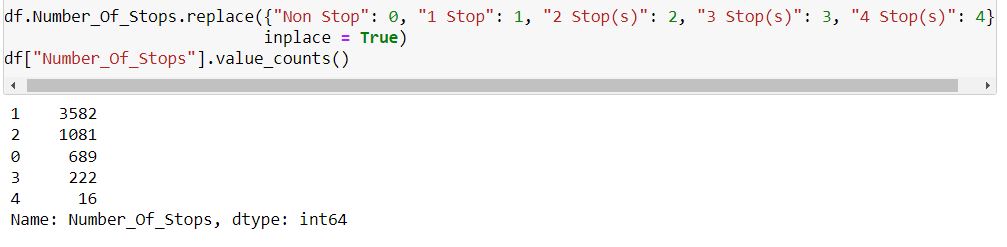
Checked the datatype details for each column to understand the numeric ones and its further conversion process.

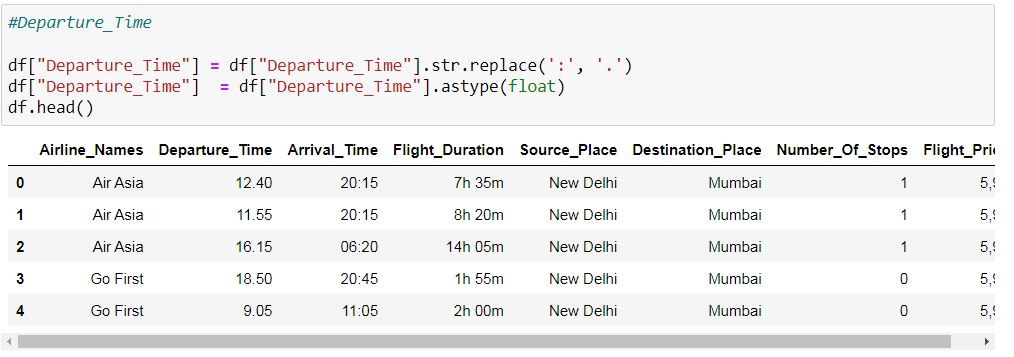


We also took a look at all the unique value present in each of the columns and then decided to deal it with the imputation part accordingly.



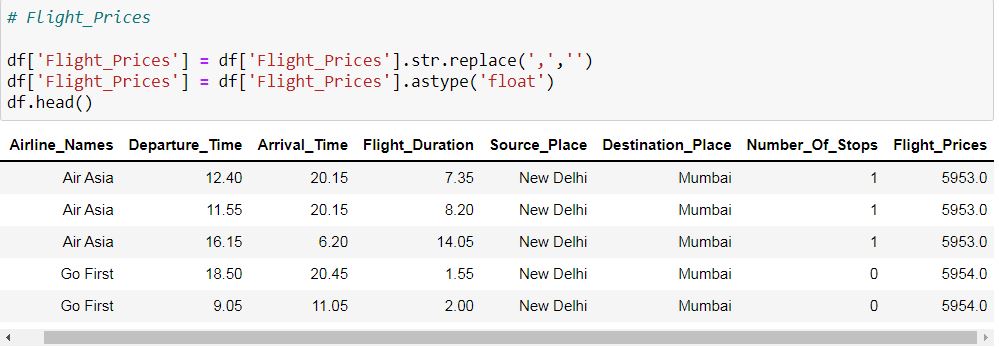
The various data processing performed on the data set is shown below with the code.





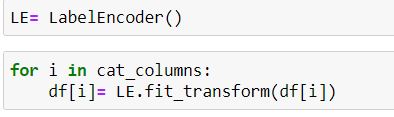




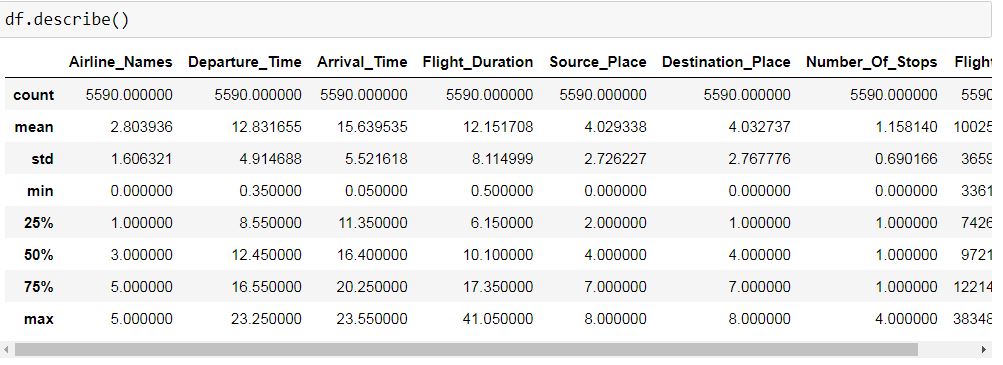


Data Inputs- Logic- Output Relationships

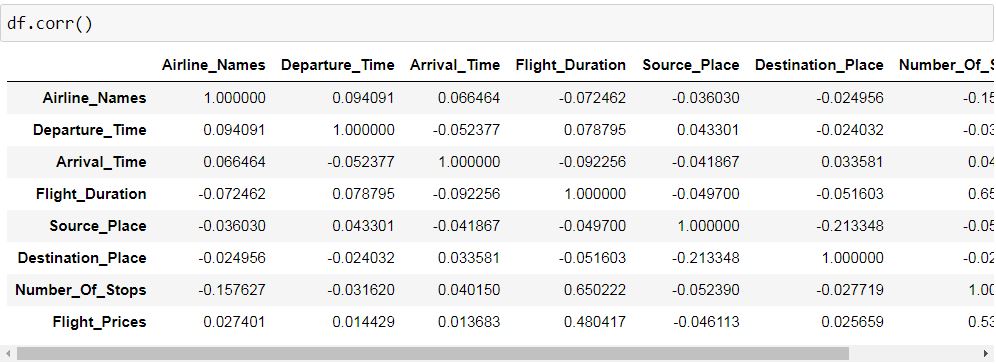
The input data were initially all object datatype so the data is cleaned by removing unwanted information like “km” from Kilometres Driven column and ensured the numeric data are converted accordingly. Then Label Encoding method was used to convert all the categorical feature columns to numeric data type.

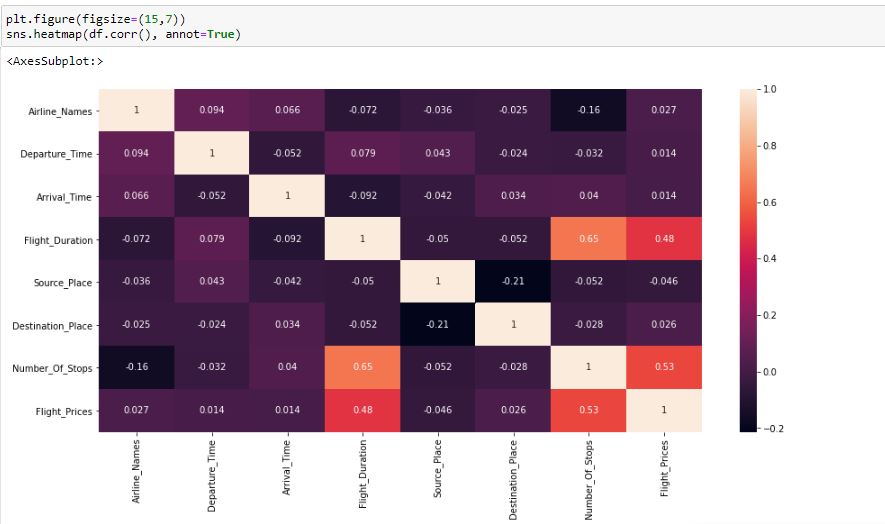


Then we used the “describe” method to check the count, mean, standard deviation, minimum, maximum, 25%, 50% and 75% quartile data.

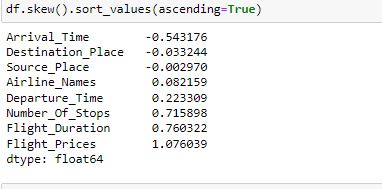


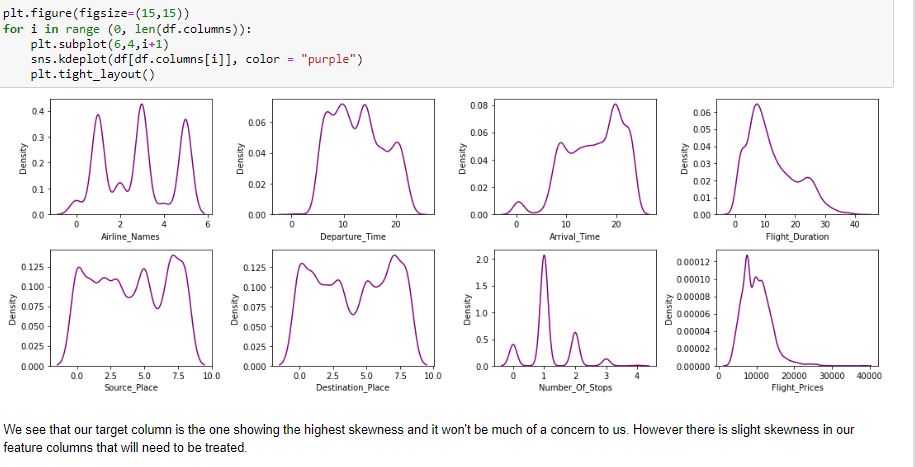
Then we used correlation method to find correlation between columns and with the target variable.





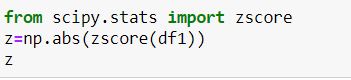
Then we used Power Transform to remove Skewness from the dataset.

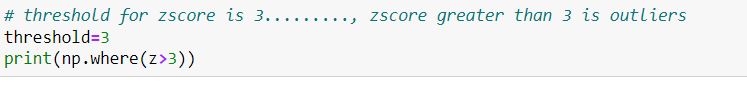




Capture.JPG

Then we used zscore technique to remove outliers.





Capture.JPG

## Hardware and Software Requirements and Tools Used

1. Python 3.8.
2. NumPy.
3. Pandas.
4. Matplotlib.
5. Seaborn.
6. Data science.
7. SciPy
8. Sklearn.
9. Anaconda Environment & Jupyter Notebook

**Hardware and Software Requirements and Tools Used**

**Open source web-application used for programming:**

**1. Jupyter Notebook**

Python Libraries / Packages used were:

1. **Pandas**: pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

We have used pandas to import the csv file using pd.read\_csv all data analysis have been done using the pandas and numpy libraries. The data characteristics have been studied using pandas functions like df.shape(), df .dtypes, df.columns etc.

2. **NumPy:** NumPy is an open-source numerical Python library. NumPy contains a multi-dimensional array and matrix data structures. It can be utilised to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic.

We have used the np.where function many times while dealing with the z-scores.np.abs () function has also been used to find the zscore and some mathematical operations like square root.

3. **Matplotlib**: library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

The Matplotlib libraries pyplot function is used for making plots ,plt.show() ,plt.figure(figsize) that has been used is a part of matplotlib library.

4. **Seaborn**: Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

All the visualizations made are built using the seaborn library. Alias used for seaborn is sns.

Sns.boxplot(), sns.heatmap(), sns.distplot(), sns.scatterplot() ,sns.stripplot,sns.swarmplot ,heatmap are few of the libraries used

5. **SciPy**: SciPy, a scientific library for Python is an open source library for mathematics, science and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. The main reason for building the SciPy library is that, it should work with NumPy arrays.

In this particular project scipy functions such as scipy.stats is used .Zscores are also obtained via scipy.stats library

6. **Sklearn:** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, and clustering and dimensionality reduction via a consistence interface in Python.

All the Machine Learning regression algorithms have been imported from the sklearn package. Simple imputer used is also a part of sklearn

The evaluation metrics, RMSE,MSE,MAE functions are also imported from same.

7. **Python:** is a set of Data analysis tools in python which gives us measures of association for categorical features, Plots features correlation and association for mixed data-sets (categorical and continuous features) in an easy and simple way.

# Model/s Development and Evaluation

## Identification of possible problem-solving approaches (methods).

Considering the business requirement provided, the data was collected to predict the Flight price but there where multiple data of the flight are available. Data like departure time, arrival time, Flight Duration, Source, Destination. But after analysing all these data we have selected the data that have more correlation with the price. Steps followed are listed here.

1. Clean the dataset from unwanted scraped details.

2. Rename values with meaningful information.

3. Encoding the categorical data to get numerical input data.

4. Compare different models and identify the suitable model.

5. R2 score is used as the primary evaluation metric.

6. MSE and RMSE are used as secondary metrics.

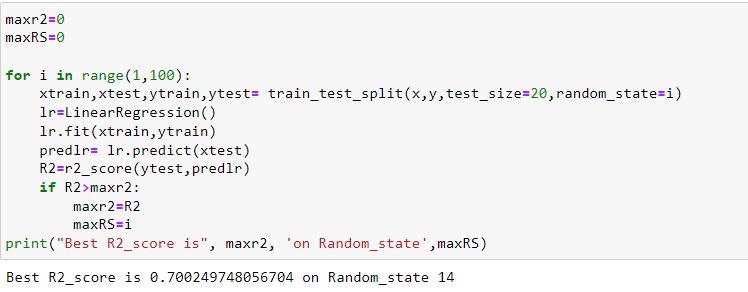
7. Cross Validation Score was used to ensure there are no overfitting our underfitting models.

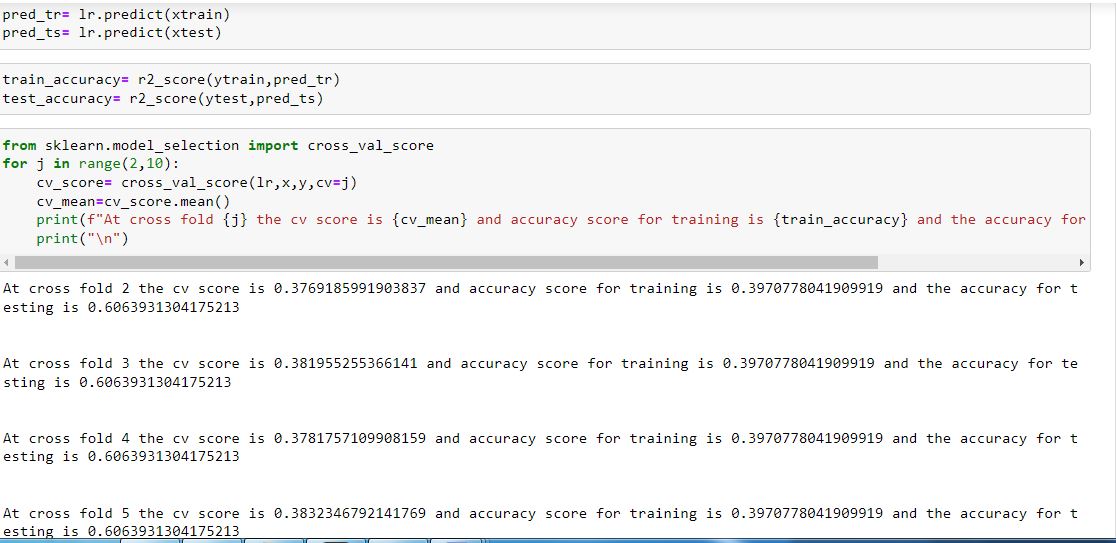
The algorithms used on data are as follows:

* Linear Regression Model
* Ridge Regularization Regression Model
* Lasso Regularization Regression Model
* Support Vector Regression Model
* Decision Tree Regression Model
* Random Forest Regression Model
* K Nearest Neighbours Regression Model

**Run and Evaluate selected models**

We have used a total of 7 Regression Models after choosing the random state amongst 1-100 number. Then it is checked for best fold for CV using the following codes.

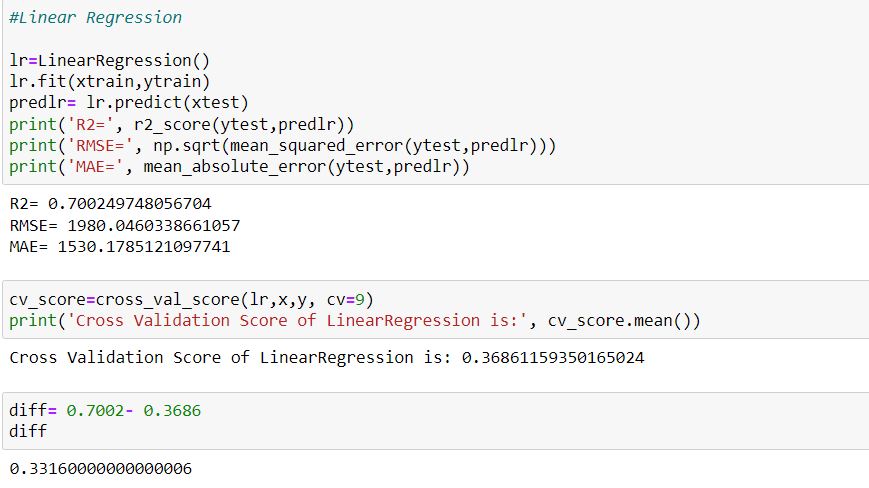




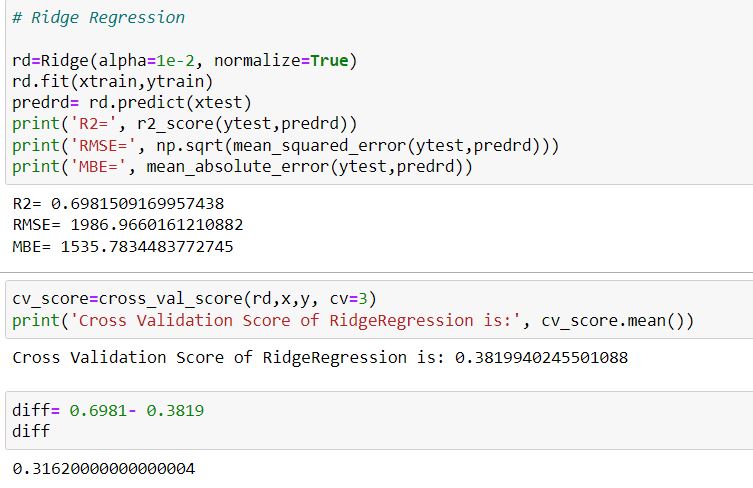
Regression Model Function

Capture.JPG

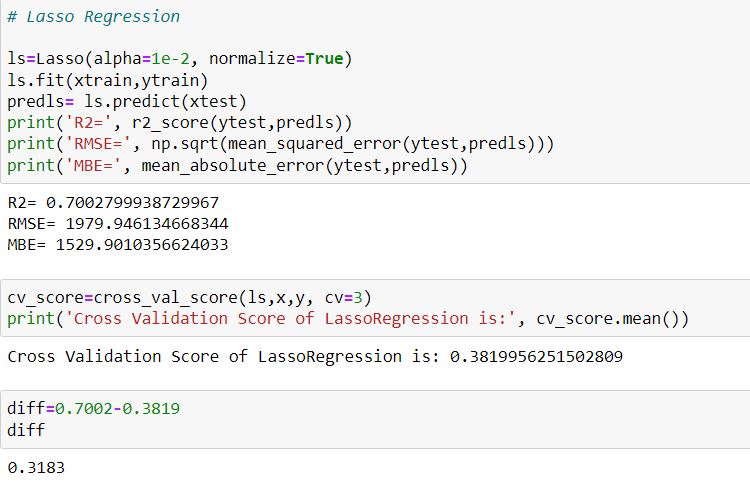
**Linear Regression**



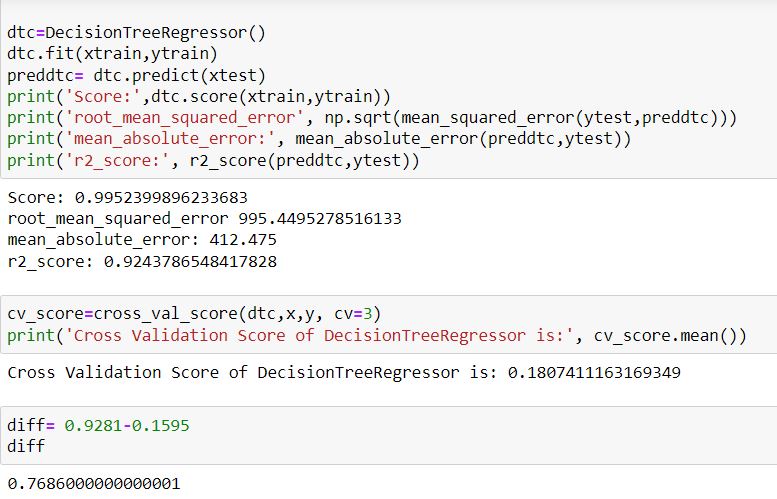
Ridge Regression



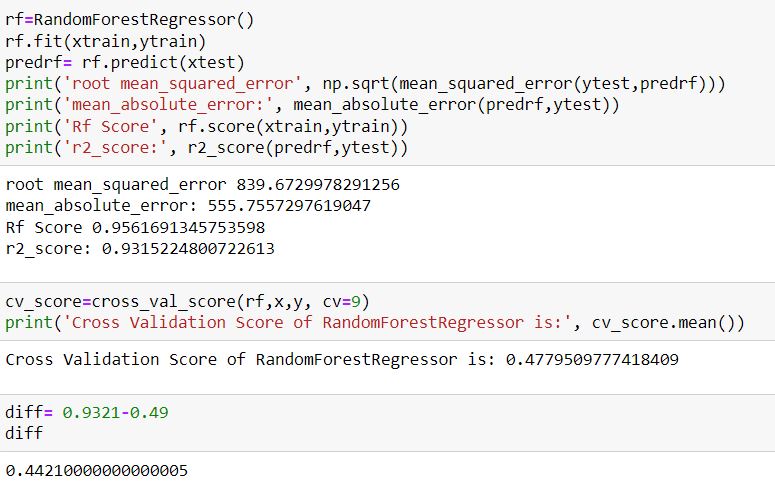
**Lasso Regression**



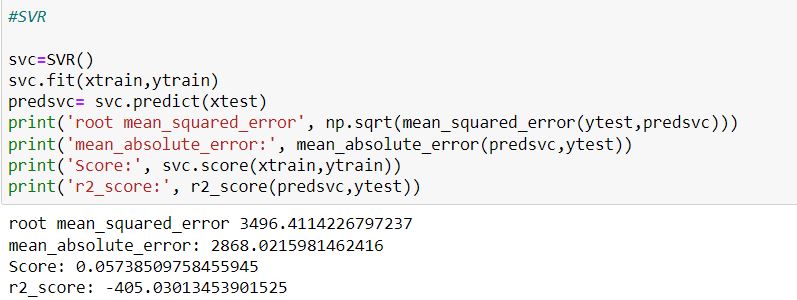
**DecisionTree Regression**



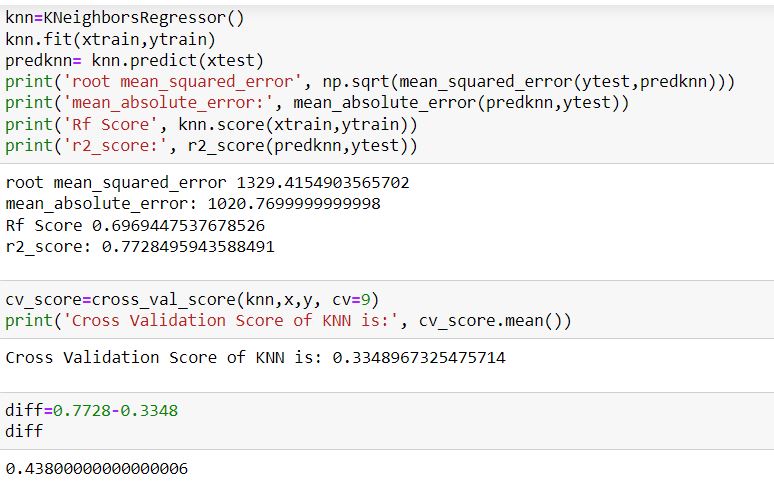
**Random Forest Regression**



**Support Vector Regression**



**KNN**



**Key Metrics for success in solving problem under consideration**

The key metrics used here were r2\_score, cross\_val\_score, MAE, MSE and RMSE. We tried to find out the best parameters and also to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method.

1. Cross Validation:

Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 7 folds, the data will be divided into 7 pieces, where each part being 20% of full dataset. While running the Cross-validation the 1st part (17%) of the 7parts will be kept out as a holdout set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset.

In the similar way further iterations are made for the second 17% of the dataset is held as a holdout set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross-validation process to get the remaining estimate of the model quality.

2. R2 Score:

It is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

3. Mean Squared Error (MSE):

MSE of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. RMSE is the Root Mean Squared Error.

4. Mean Absolute Error (MAE):

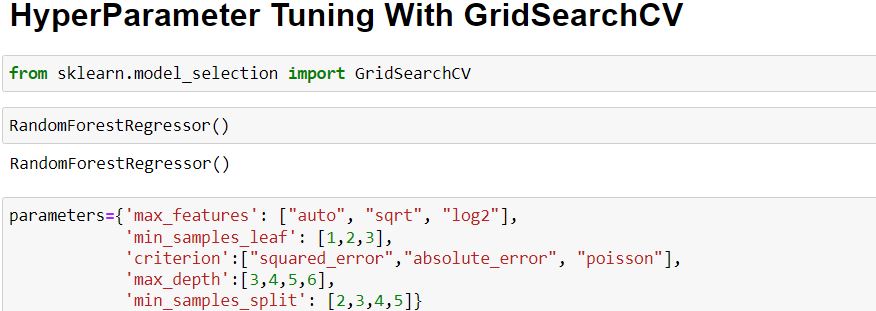
MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

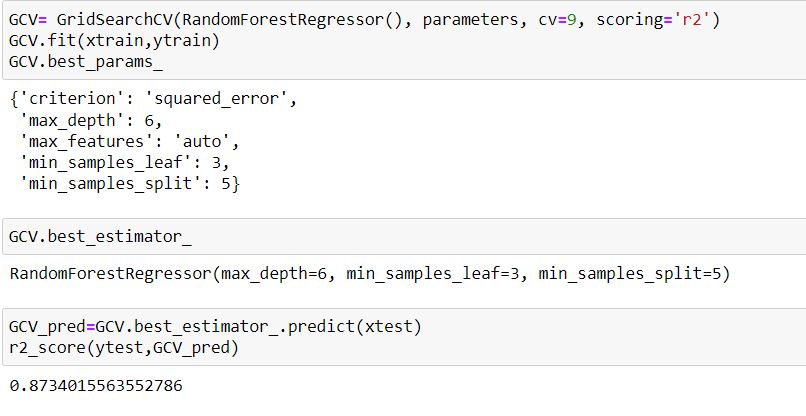
5. Hyperparameter Tuning:

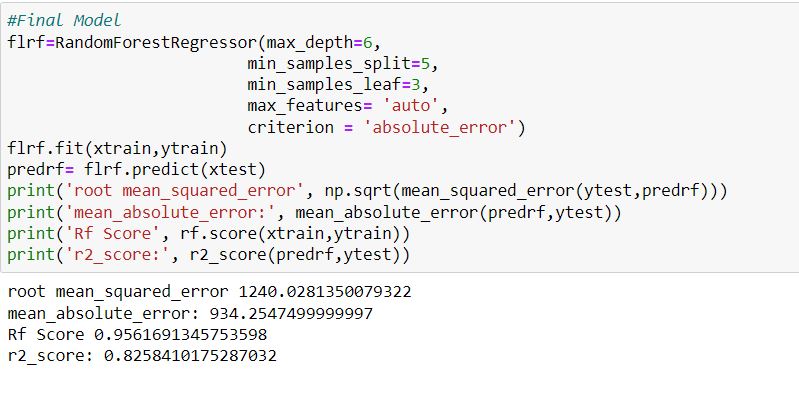
There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as Hyperparameters. These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. You must select from a specific list of hyperparameters for a given model as it varies from model to model.

We are not aware of optimal values for hyperparameters which would generate the best model output. So, what we tell the model is to explore and select the optimal model architecture automatically. This selection procedure for hyperparameter is known as Hyperparameter Tuning. We can do tuning by using GridSearchCV.

GridSearchCV is a function that comes in SK-learn model selection package. This function helps to loop through predefined hyperparameters and fit our estimator (model) on our training set. So, we can select the best parameters from the listed hyperparameters.



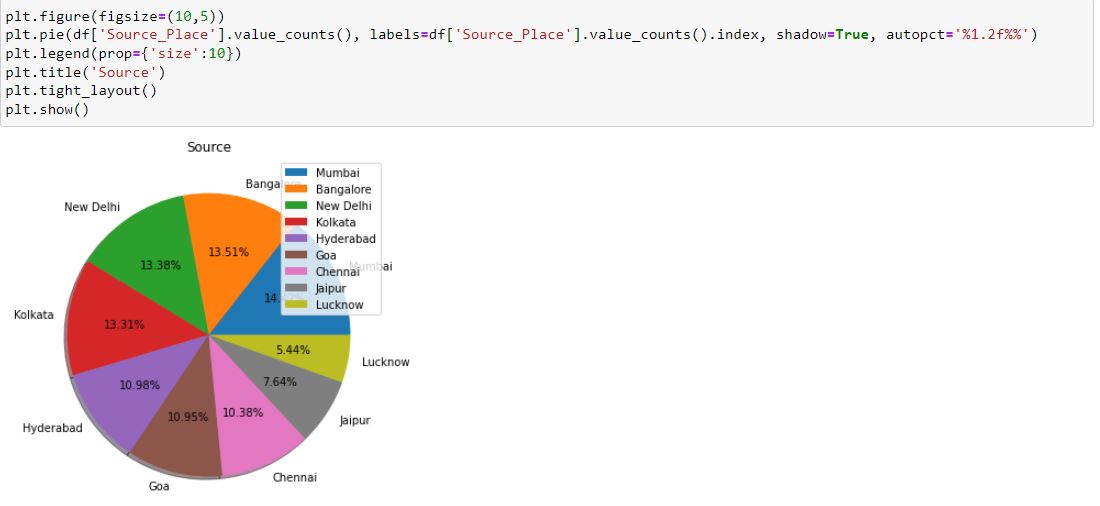


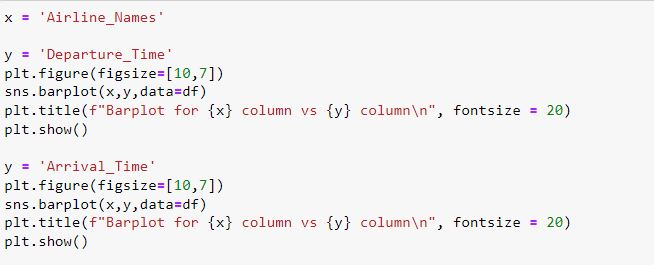


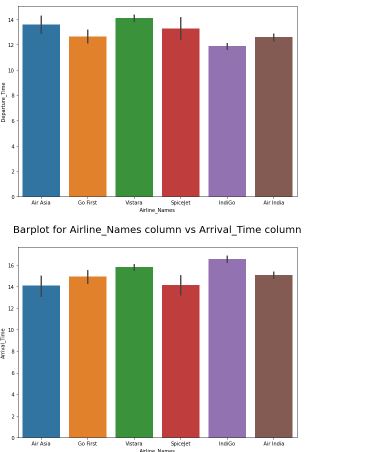
It is possible that there are times when the default parameters perform better than the parameters list obtained from the tuning and it only indicates that there are more permutations and combinations that one needs to go through for obtaining better results.

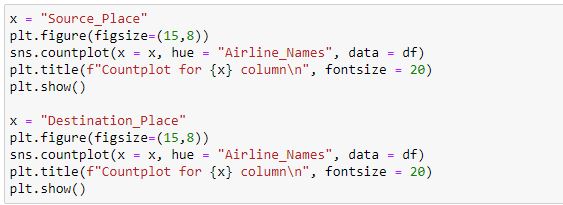
**Visualizations**

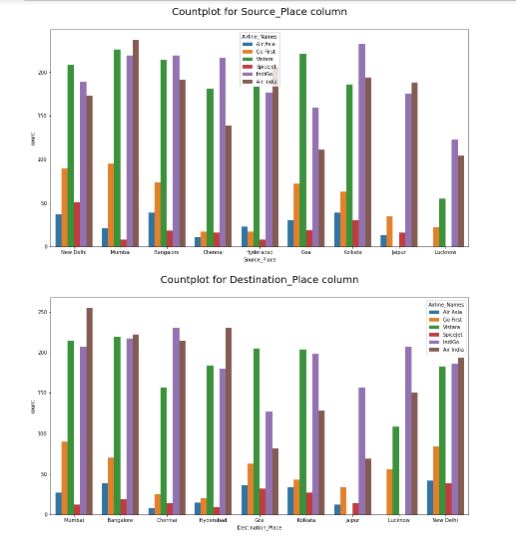
I have done some visualizations to understand the input output logic of the data collected.

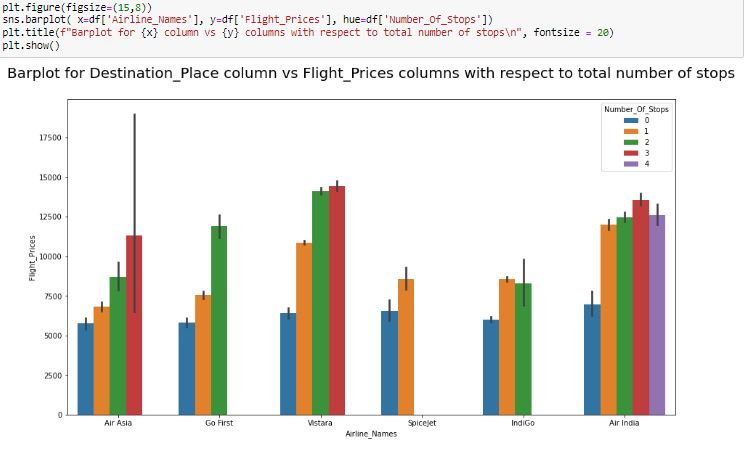


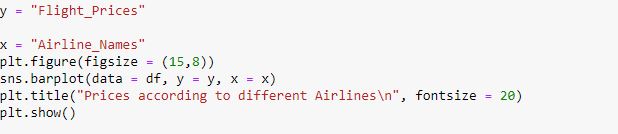


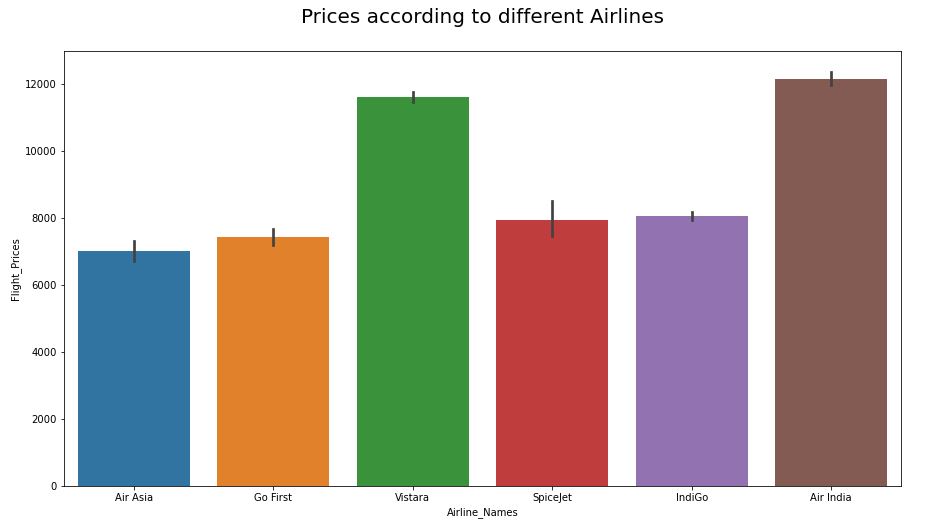


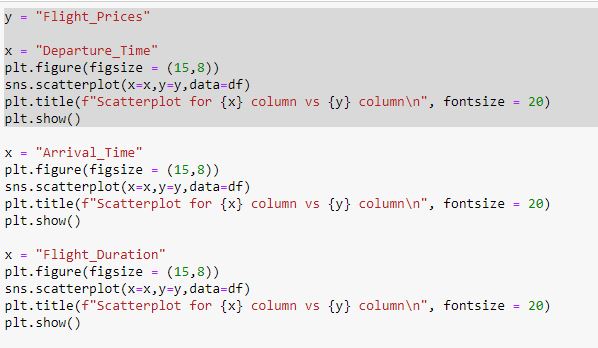


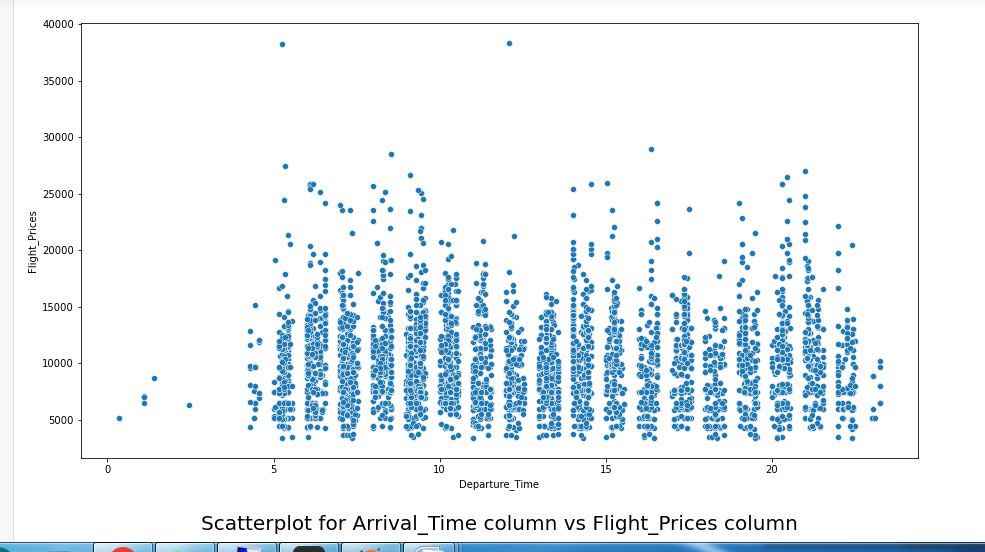


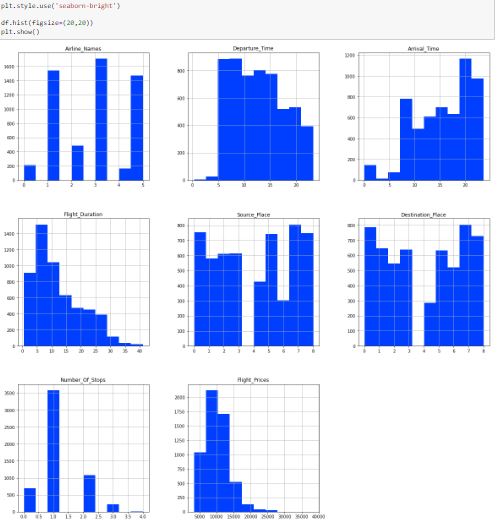


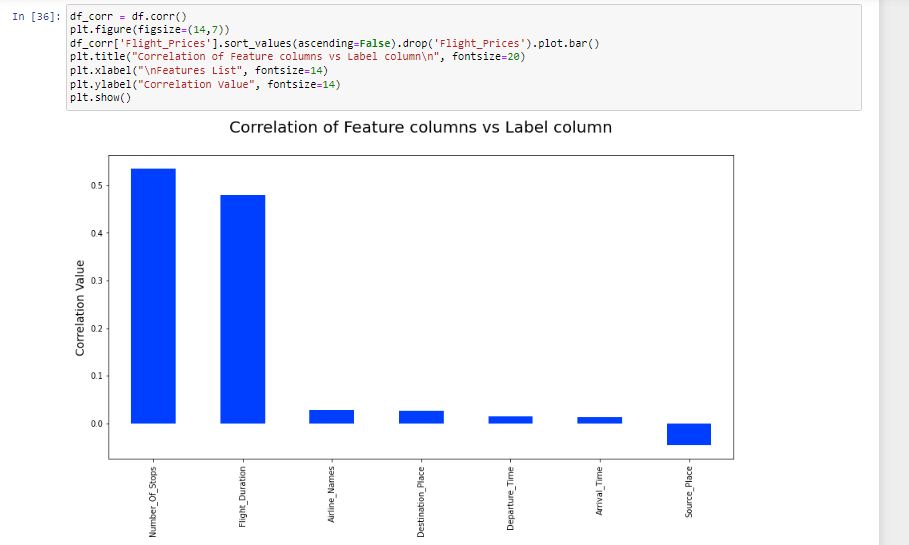












## Interpretation of the Results:

From the above EDA we can easily understand the relationship between features and we can even see which features are affecting the price of flights. 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 have tried 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.

Conclusions of the Study

From this dataset we get to know that each feature plays a very import role to understand the data. Data format plays a very important role in the visualization and applying the models and algorithms.

**Learning Outcomes of the Study in respect of Data Science**

Visualization part helped me to understand the data as it provides graphical representation of huge data. It assisted me to understand the feature importance, outliers/skewness detection and to compare the independent-dependent features. Data cleaning is the most important part of model building and therefore before model building, I made sure the data is cleaned. I have generated multiple regression machine learning models to get the best model wherein I found Random Forest Regressor Model being the best based on the metrics I have used.

**Limitations of this work**

R2scores could have increased with more features. Some algorithms are facing over-fitting problem which may be because of a smaller number of features in our dataset.

Another limitation of the study is that in the volatile changing market we have taken the data, to be more precise we have taken the data after pandemic, it may change once stabilized and we have shortlisted and taken these data from the important cities across India. If the customer is from the different country our model might fail to predict the accuracy prize of that flight.