

**“A PROJECT REPORT ON FLIGHT TICKET PRICE PREDICTION”**



SUBMITTED BY

HIMAJA IJJADA

**ACKNOWLEDGMENT**

I express my sincere gratitude to FlipRobo Technologies for giving me the opportunity to work on “**A PROJECT REPORT ON FLIGHT TICKET PRICE PREDICTION”** using machine learning algorithms. I would also like to thank FlipRobo Technologies for providing me with the requisite knowledge to webscrape the datasets to work with. And I would like to express my gratitude to Mr. Mohd Kashif (SME FlipRobo) and Ms. Sapna Verma (SME FlipRobo) for being of a great help in completion of the project.

Most of the concepts used to predict the Prices of flight tickets project are learned from Data Trained Institute and below documentations.

* https://scikit-learn.org/stable/
* https://seaborn.pydata.org/
* https://www.scipy.org/

**CONTENTS**

* **Introduction**
* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken
* **Analytical problem framing**
* Mathematical/ Analytical Modeling of the Problem
* Data Sources and their formats
* Data Preprocessing Done
* Data Inputs- Logic- Output Relationships
* Assumptions
* Hardware and Software Requirements and Tools Used
* **Model/s Development and evaluation**
* Visualizations
* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration
* Interpretation of the Results
* **Conclusion**
* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**Introduction**

**Business Problem Framing**

Airline industry is one of the most sophisticated in its use of dynamic pricing strategies to maximize revenue, based on proprietary algorithms and hidden variables. That is why the airline companies use complex algorithms to calculate the flight ticket prices. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices. Considering the features such as departure time, arrival time and time of the day it will give the best time to buy the ticket. 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 models 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.

**Business goal:** The main aim of this project is to predict the price of flight tickets based on various features. The purpose of the paper is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices. Then using this information, build a system that can help buyers whether to buy a ticket or not. So, we will deploy Machine Learning model for flight ticket price prediction and analysis. This model will provide the approximate selling price for the flight tickets based on different features.

**Conceptual Background of the Domain Problem**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, and it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable. 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 2 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).

Here we are trying to help the buyers to understand the price of the flight tickets by deploying machine learning models. These models would help the sellers/buyers to understand the flight ticket prices in market and accordingly they would be able to book their tickets.

**Review of Literature**

Literature review covers relevant literature with the aim of gaining insight into the factors that are important to predict the flight ticket prices in the market. In this study, we discuss various applications and methods which inspired us to build our supervised ML techniques to predict the price of flight tickets in different locations. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of data information by doing web scraping from [www.yatra.com](http://www.yatra.com) website which is a web platform where buyers can book their flight tickets. This project is more about data exploration, feature engineering and preprocessing that can be done on this data.

Since we scrape huge amount of data that includes more flight related features, we can do better data exploration and derive some interesting features using the available columns. Different techniques like ensemble techniques, and decision trees have been used to make the predictions. The goal of this project is to build an application which can predict the price of flight tickets with the help of other features. In the long term, this would allow people to better explain and reviewing their purchase in this increasing digital world.

**Motivation for the Problem Undertaken**

Air travel is the fastest mode of transport around, and can cut hours or days off of a trip. But we know how unexpectedly the prices vary due to the **Dynamic pricing**. So, I was interested in Flight Fares Prediction listings to help individuals and find the right fares based on their needs. And also, to get hands on experience and to know that how the data scientist approaches and work in an industry end to end.

**Analytical problem framing**

**Mathematical/ Analytical Modeling of the Problem**

We need to develop an efficient and effective Machine Learning model which predicts the price of flight tickets. So, “Price” is our target variable which is continuous in nature. Clearly it is a Regression problem where we need to use regression algorithms to predict the results. This project is done on three phases:

• Data Collection Phase: I have done web scraping to collect the data of flights from the well-known website www.yatra.com where I found more features of flights compared to other websites and I fetch data for different locations. As per the requirement we need to build the model to predict the prices of flight tickets.

• Data Analysis: After cleaning the data I have done some analysis on the data by using different types of visualizations.

• Model Building Phase: After collecting the data, I built a machine learning model. Before model building, have done all data pre-processing steps. The complete life cycle of data science that I have used in this project are as follows:

* Data Cleaning
* Exploratory Data Analysis
* Data Pre-processing
* Model Building
* Model Evaluation
* Selecting the best model

**Data Sources and their formats**

We have collected the dataset from the website [**www.yatra.com**](http://www.yatra.com)**which is a web platform where** one can book their flight tickets. The data is scrapped using Web scraping technique and the framework used is Selenium. We scrapped approximately 5300 of the data rows and fetched the data for flights between different locations and collected the additional information of different flights and saved the collected data in excel format. The dimension of the dataset is 5303 rows and 9 columns including target variable “Price”.

****

Here I am importing the collected dataset which is in excel format and storing it into data frame (df) for further usage. Here we can observe first 5 and last 5 rows of the dataset. There are 5303 rows and 10 columns in the data frame. The dataset contains both numerical and categorical data. There are both dependent and independent variables present in the data frame. We have our target variable **"Price"** which stores the price of the flight tickets and it is continuous in nature which makes this problem to be a **"Regression Problem"**.

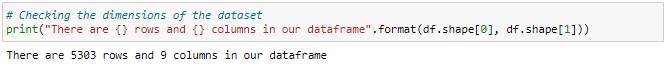
**Features Information:**

* Airline - The Name of airline
* Departure\_time - The time when the journey starts from the source
* Time\_of\_arrival - Time of arrival at the destination
* Duration - Total duration taken by the flight to reach the destination from the source
* Source - The source from which the service begins
* Destination - The destination where the service ends
* Meal\_availability - Availability of meals in the flight
* Number\_of\_stops - Total stops between the source and destination
* Price - The price of the flight ticket

**Data Preprocessing Done**

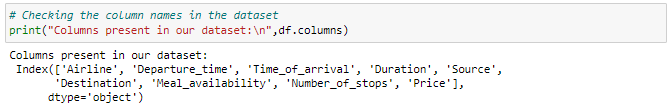
Data pre-processing is the process of converting raw data into a well-readable format to be used by Machine Learning model. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre-process our data before feeding it into our model. I have used following pre-processing steps:

Checking the dimensions of the dataset

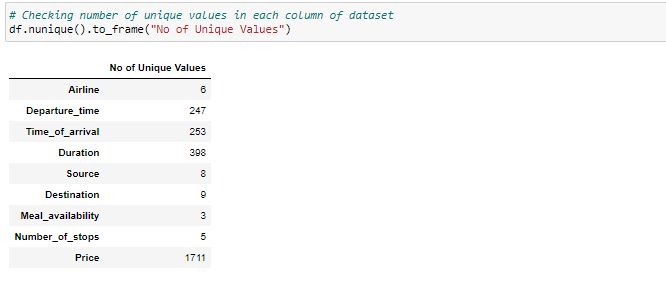
****

The dataset contains 5303 rows and 9 columns. Out of 9 columns 8 are independent variables and remaining one is our target variable "Price" which is dependent variable.

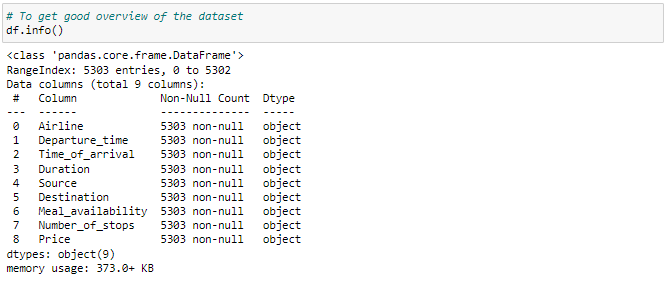
Checking the column names in the dataset

****

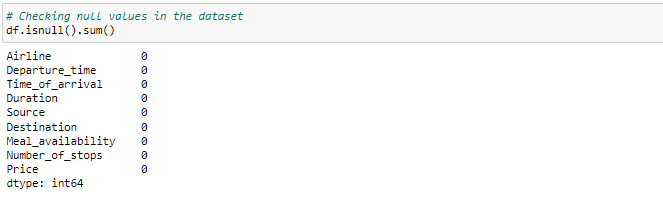
These are the columns present in our dataset.

****

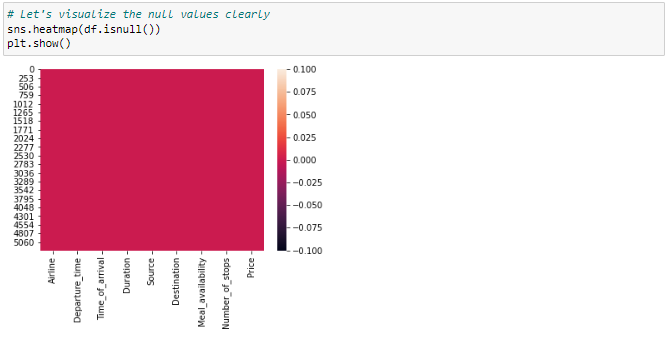
**Above are the number of unique values present in each of the columns present in the dataset.**

****

* This info() method gives the information about the dataset which includes indexing type, column type, no-null values and memory usage.
* The dataset contains object type data. We will encode the object datatypes using appropriate encoding techniques before building machine learning models.
* Since counts of all the columns are same, which means there are no null values present in the dataset.

****

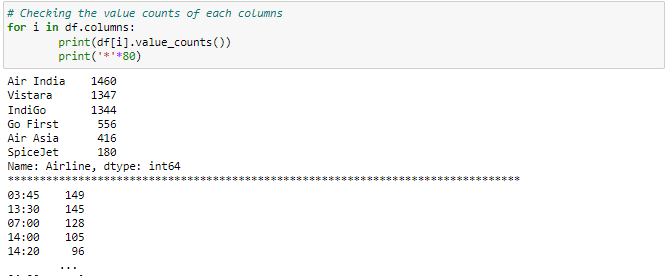
As we can see there are no missing values in any of the columns.

****

Here we can clearly observe there are no missing values.

**Value count function**

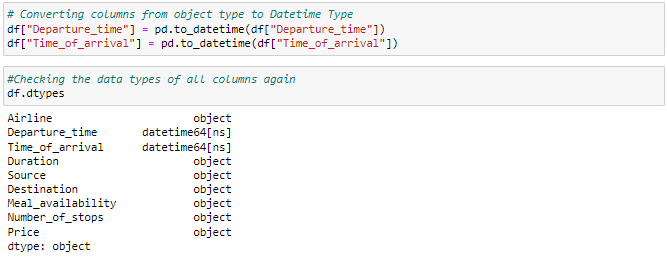
Let's check the list of value counts in each columns to find if there are any unexpected or corrupted entries present in the dataset

****

These are the value counts of the columns present in the dataset.

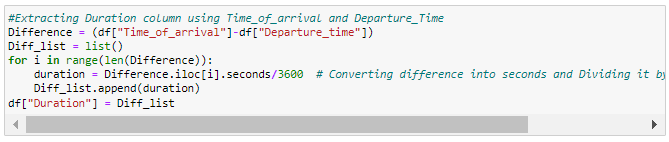
## **Feature Engineering**

The columns Time\_of\_arrival and Departure\_Time showing object data type which means python is not able to understand the type of data in this column due to some string values or categorical signs like ":" which we can observe in the value count function. Therefore, we have to convert this datatype into timestamp (datetime) to use them properly for prediction.

****

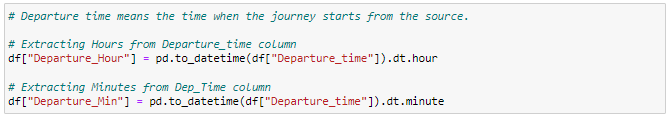
### Duration

The column Duration has values in terms of minutes and hours. Duration means the time taken by the plane to reach the destination and it is the difference between the arrival time and Departure time. Let's extract proper duration time in terms of float data type from Time\_of\_arrival and Departure\_time columns.

****

### Departure\_time

Let's extract values from Departure\_time. Departure time means when a flight leaves the airport and this column contains hours and minutes so we will extract hours and minutes from Departure\_time.

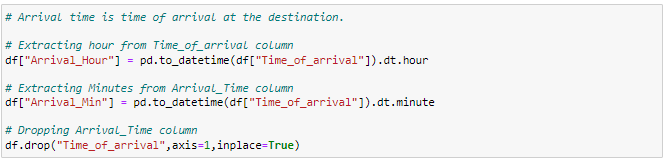
****

Now we have extracted hour and minute from Departure\_time column. Let's drop Departure\_time column as it is of no use now.

****

### Time\_of\_arrival

Similarly we can extract hours and minutes from Time\_of\_arrival column and dropping Time\_of\_arrival column.

****

Now we have extracted required data from the columns.

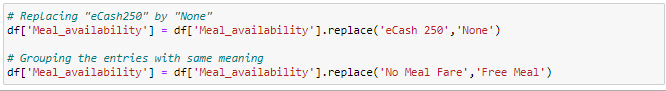
### Price

The target column should be in continuous numeric data type but it is appearing as object data type due to some categorical sign ",". Let's replace this sign by empty space and convert the type into float.

****

### Meal\_availability

From the value count function of Meal\_availability we can observe "eCash 250" entry which does not belongs to meals so we can replace it as "None". Also, the other two entries "No meal fare" and "Free meal" belongs to same category that is they give same meaning so we can group them as well. We can also drop this column, but there are only few features in the dataset so, trying to retain the columns for prediction.

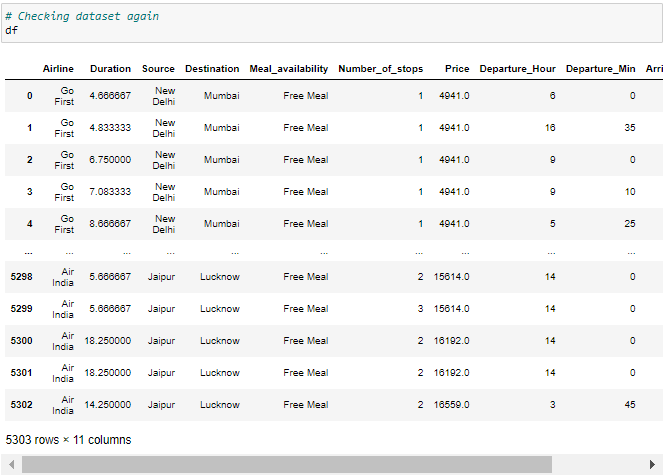


### Number\_of\_stops

From the value count function of Number\_of\_stops we can observe the categorical values, let's replace them with numeric data.

****

Now we have successfully cleaned our data, let's have a look at dataframe.

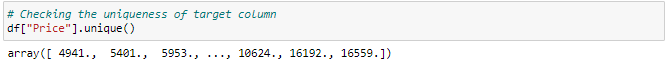
****

****

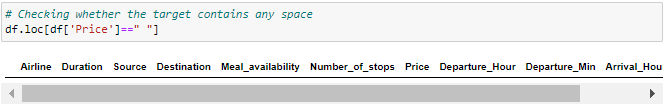
Now the dataset contains 5303 rows and 11 columns.

****

The dataframe has 3 types of data that is object, integer and float data types. We will encode the object data types before building the ML model.

****

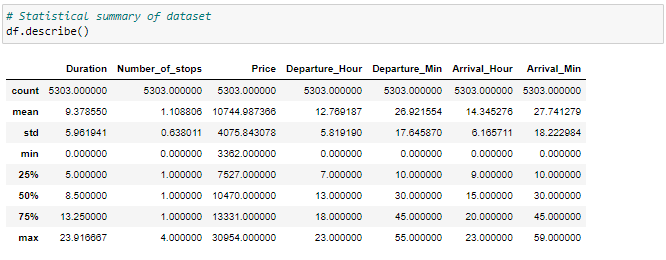
These are the unique values present in the target column.

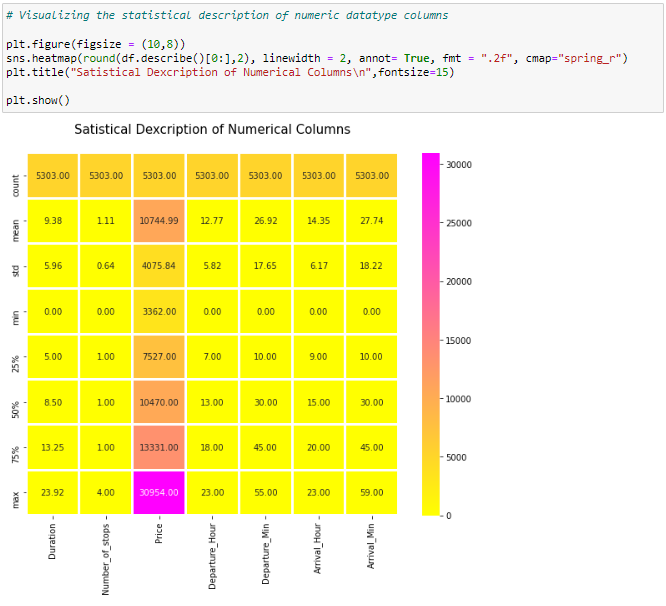
****

There are no any empty spaces in any of the columns.

**Data Inputs- Logic- Output Relationships**

## Description of Dataset

****

****

From the heat map we can observe the statistical summary of the numerical features present in the dataset.

**Assumptions**

This gives the statistical information of the dataset. The summary of this dataset looks perfect since there is no negative/ invalid values present. It gives the summary of numerical data.

From the above description we can observe the following things

* The counts of every column is same which means there are no missing values present in the dataset.
* The mean value is greater than the median (50%) in the columns Duration, Number\_of\_stops, and Price so we can say these columns are skewed to right.
* The median (50%) is bit greater than mean in Departure\_Hour, Departure\_Min, Arrival\_Hour, Arrival\_Hour and Arrival\_Min which means these columns are skewed to left.
* From the description we can say the minimum price of the flight ticket is Rs.3362.00 and maximum price is Rs.30954.00 also the mean is 10744.987366.
* In summarizing the data we can observe that there is huge difference in maximum and 75% percentile in the columns Price, Duration, etc that means huge outliers present in those columns. These differences can also be seen in many other columns. So we need to remove outliers and skewness to get better model and prediction.

**Hardware and Software Requirements and Tools Used**

To build the machine learning projects it is important to have the following hardware and software requirements and tools.

**Hardware required:**

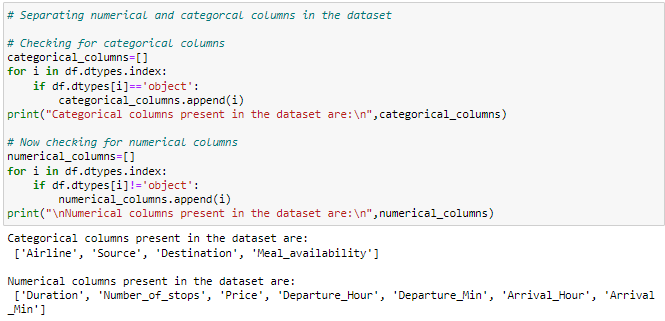
* Processor: core i5 or above
* RAM: 8 GB or above
* ROM/SSD: 250 GB or above

**Software required:**

* Distribution: Anaconda Navigator
* Programming language: Python
* Browser based language shell: Jupyter Notebook
* Chrome: To scrape the data

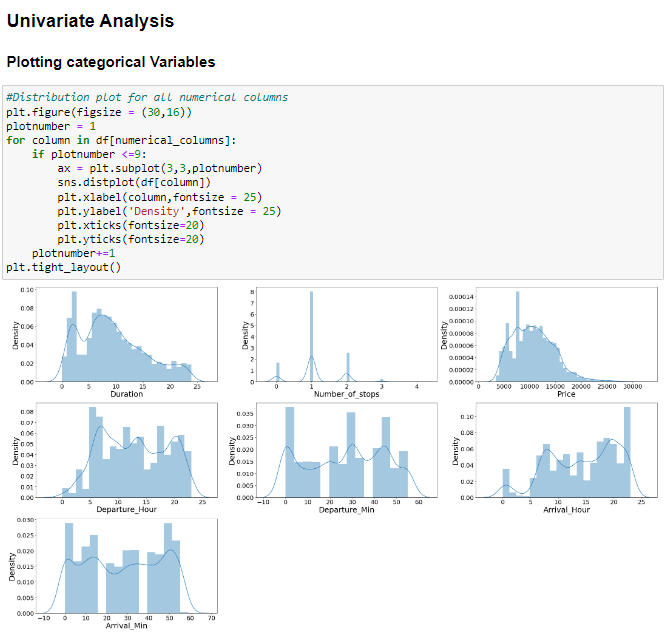
**Model/s Development and evaluation**

**Visualizations**

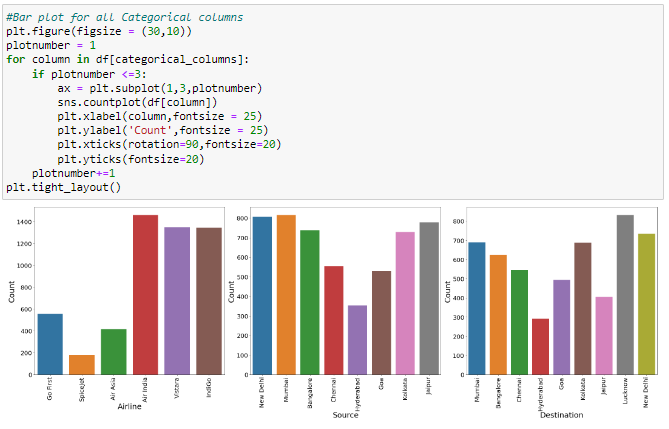
****

# Exploratory Data Analysis (EDA)

# Data Visualization

****

There is no skewness in any of the numerical columns.

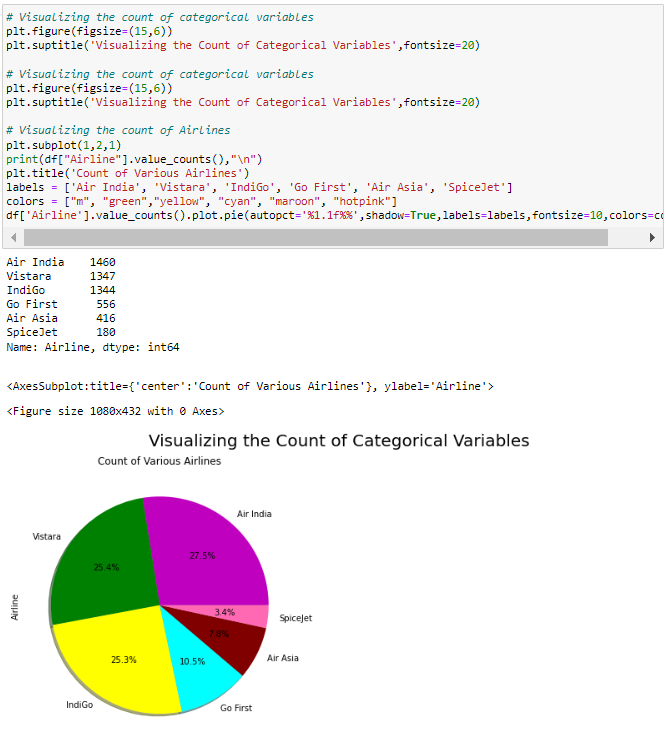
****

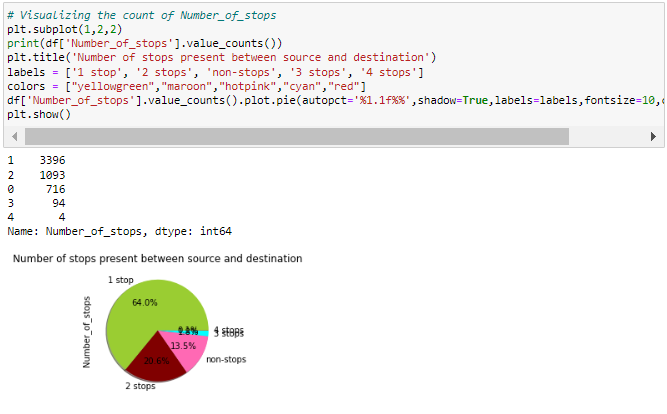
* Indigo has maximum count which means most of the passengers preferred Indigo for there travelling.
* New Delhi has maximum count for source which means maximum passengers are choosing New Delhi as there source.
* New Delhi has maximum count for Destination which means maximum passengers are choosing New Delhi as there Destination.

****

## Observations:

* Flights with 1 stop costs more price compared to other flights.
* At 2PM departure time of every day the flight Prices are high so it looks good to book flights rather than this departure time.
* And Departure minute has less relation with target Price.
* At 7AM to 1PM Arrival time of every day the flight Prices are high so it looks good to book flights rather than this arrival time.
* And Arrival minute has less relation with target Price.

****

****

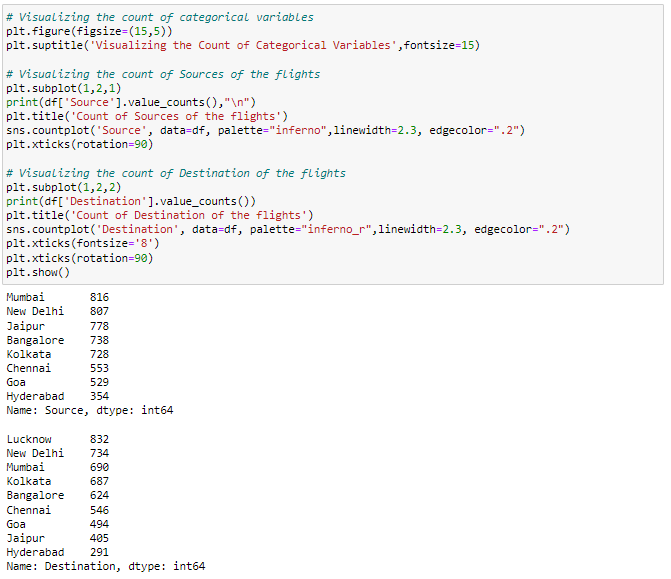
**Observations:**

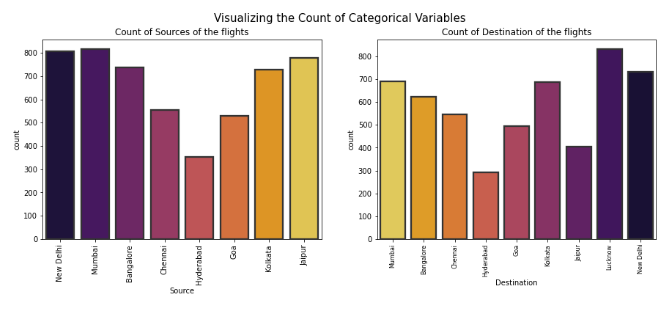
* **Airline:** From the pie plot we can infer that there are more number of flights of "Air India", "Vistara" and "Indigo" compared to others. Also, the count of Spicejet flights are very less.
* **Number\_of\_stops:** From the above pie plot we can infer that 64% of the flights have only 1 stop during the journey and some of the flights (20.6%) have 2 stops where only few flights have 3 and 4 stops.

****

**Observations:**

* **Meal\_availability:** Most of the flights providing free meals and only few flights are not providing any meals.

****

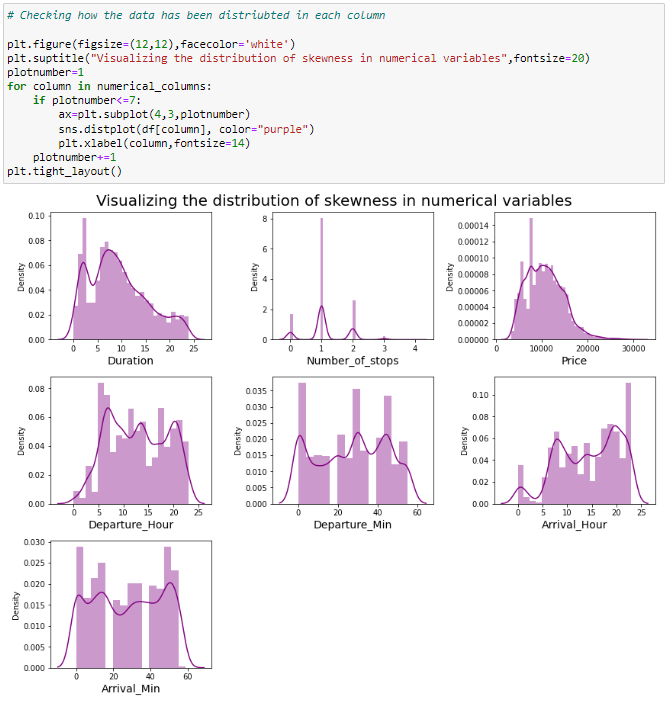
****

**Observations:**

* **Source:** From the count plot we can observe more number of flights are from Mumbai, New Delhi, Jaipur, Kolkata and Bangalore. Only few flights are from Hyderabad.
* **Destination:** More number of flights are heading towards Lucknow, New Delhi and Kolkata. Only few flights are travelling to Hyderabad.

## Distribution of skewness

### Plotting Numerical Variables

****

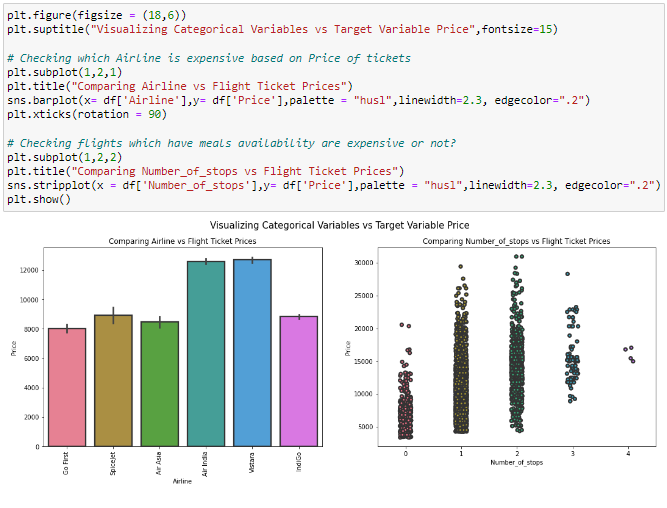
**Observations:**

Above plot shows how the data has been distributed in each of the columns.

* From the distribution plot we can observe the columns are somewhat distributed normally as they have no proper bell shape curve.
* The columns like "Duration", "Number\_of\_stops" and "Price" are skewed to right as the mean value in these columns are much greater than the median(50%).
* Also the data in the column Arrival\_Hour skewed to left since the mean values is less than the median.
* Since there is presence of skewness in the data, we need to remove skewness in the numerical columns to overcome with any kind of data biasness.

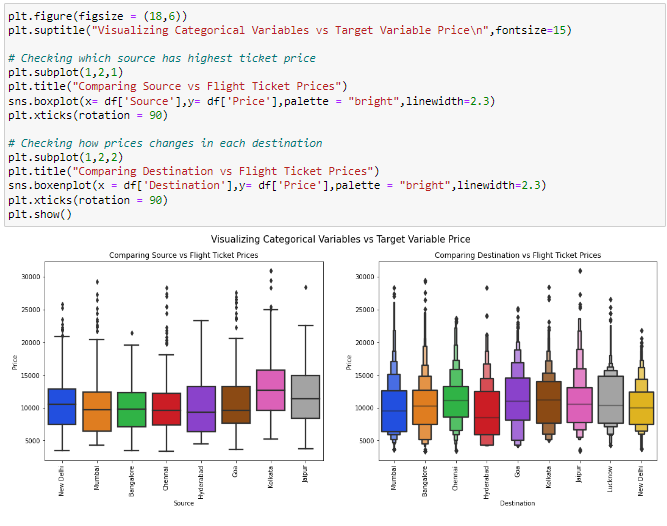
## Bivariate Analysis

### Visualizing Categorical Variables vs Target Variable Price

****

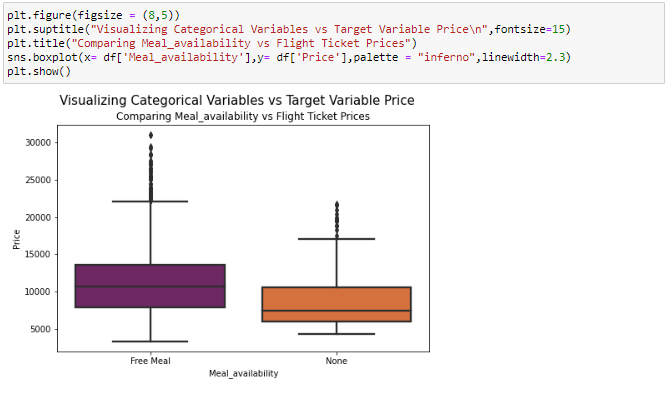
**Observations:**

* **Airline vs Price:** From the bar plot we can notice "Vistara" and "Air India" airlines have highest ticket prices compared to other airlines.
* **Number\_of\_stops vs Price:** From the strip plot we can notice the flights which have 1 and 2 stops between source and destination have highest ticket prices compared to others. The airlines which have 4 stops during the journey have very less ticket price. So we can say as the stops increases, ticket price decreases.

****

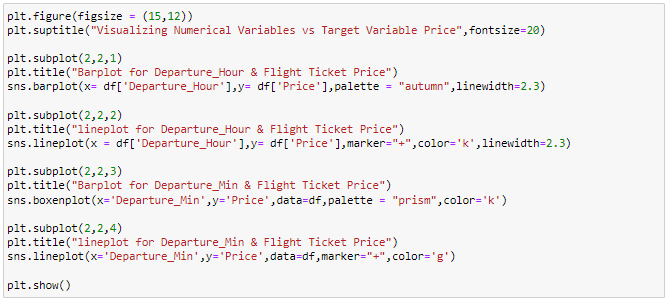
**Observations:**

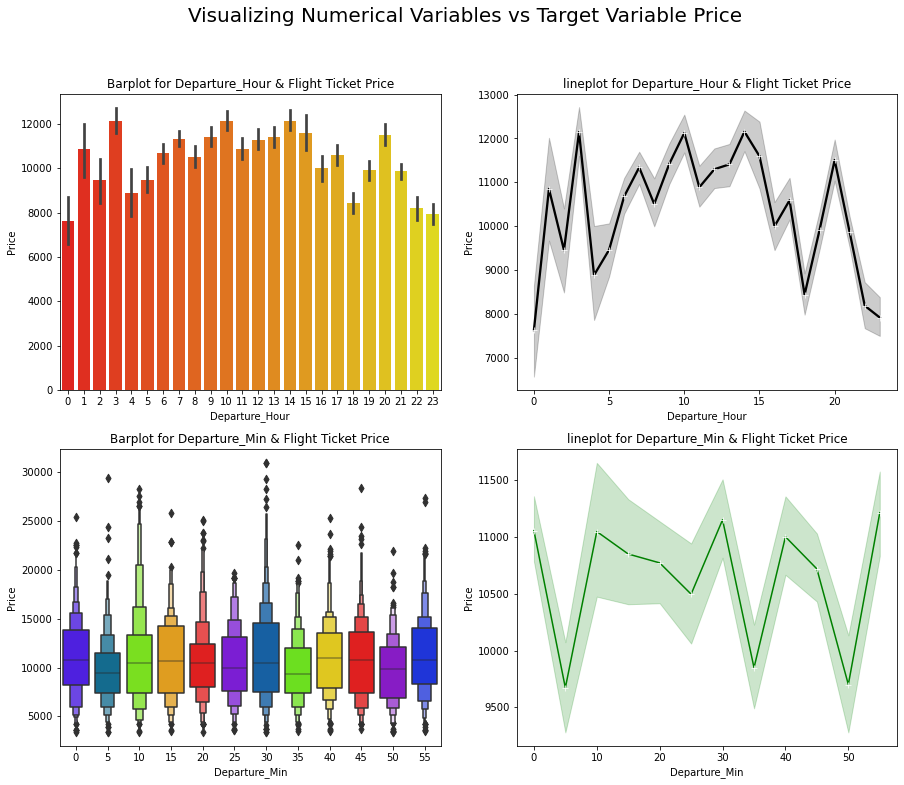
* **Source vs Price:** From the box plot we can observe the flights from Kolkata are having somewhat higher prices compared to other sources.
* **Destination vs Price:** From the boxen plot we can notice that the flights travelling to Goa have higher flight ticket prices.

****

### Observations:

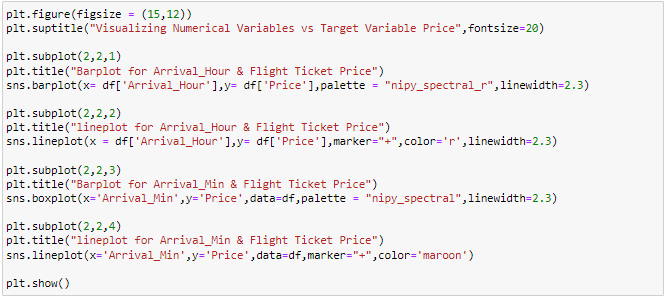
* **Meal\_availability vs Price:** The boxplot shows the flights having Free meal facility have high ticket prices.

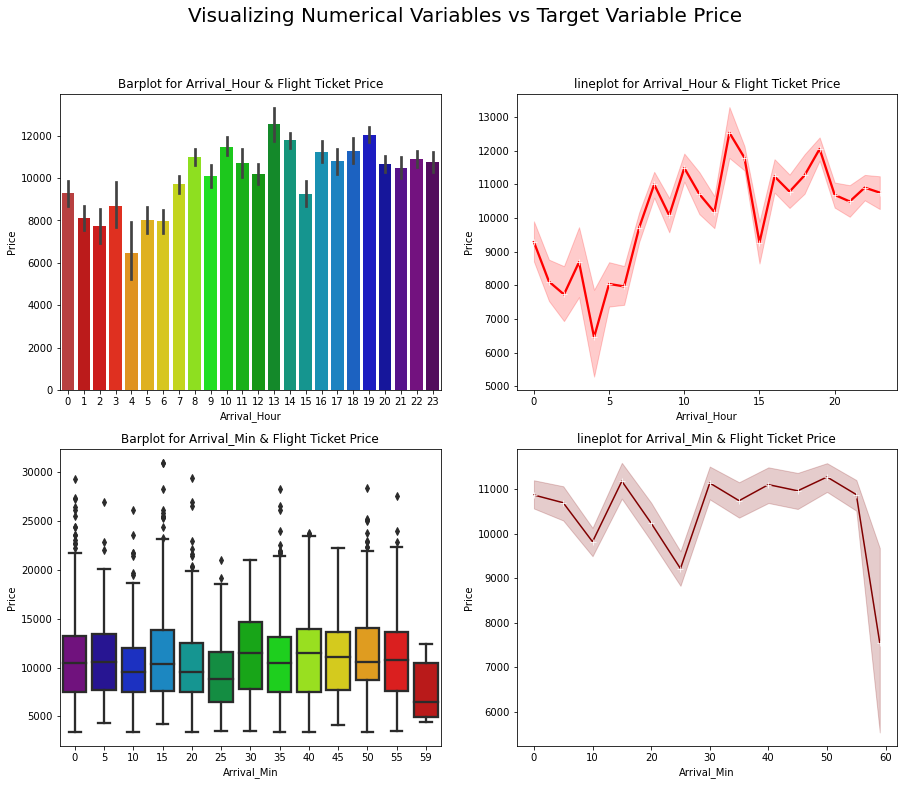
****

****

### Observations:

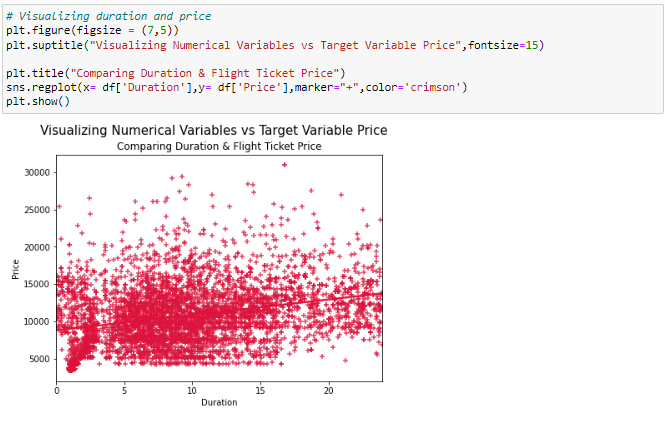
* **Departure\_Hour vs Price:** From the bar plot and line plot we can see that there are some flights departing in the early morning 3 AM having most expensive ticket prices compared to late morning flights. We can also observe the flight ticket prices are higher during afternoon (may fluctuate) and it decreases in the evening.
* **Departure\_Min vs Price:** The boxen plot and line plot gives there is no significant difference between price and departure min.

****

****

### Observations:

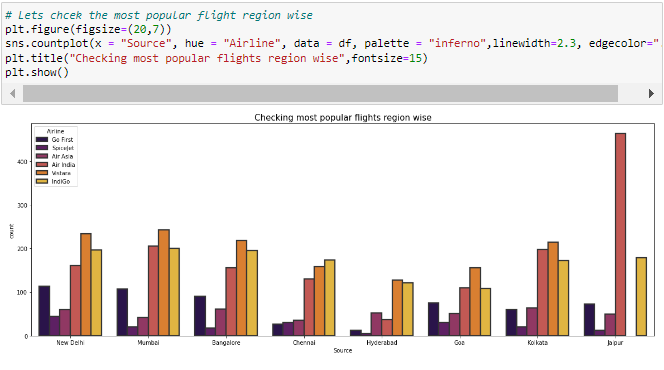
* **Arrival\_Hour vs Price:** From the bar plot and line plot we can observe that very few flights are arriving in the early morning that is 0 to 6 AM they have very less ticket price. Also, the flights which are arriving in the afternoon and evening have somewhat higher price. So, we can conlude this column has some positive correlation with price.
* **Arrival\_Min vs Price:** There is no significant difference between this feature and price. We can say flight ticket prices are not much dependent on the Arrival\_min.

****

### Observations:

* **Duration vs Price:** From the reg plot we can observe some positive linear relation between Duration and Price. Flights having 1-12 hours of duration, they have ticket price of around 15000.

Till now we have checked the relation between the independent variables and dependent variable that is our target column "Price". Now let's check the relation between two independent variables and compare each of them with others.

****

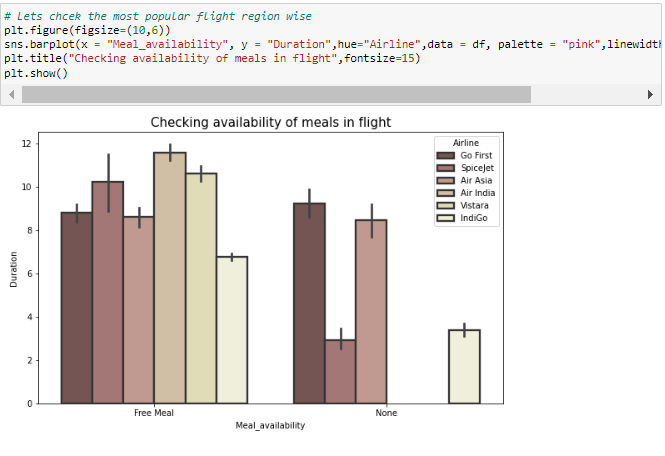
### Observations:

* **Source vs Airline:** The plot showing the region wise count of airlines which tells us that Jaipur source is not having Vistara flights and it has Air India flights in higher count compared to other sources. Other sources have Air India, Vistara and Indigo flights with higher count.

****

### Observations:

* Above plot gives the relation between Airline and Departure hour based on Number of stops. Air India and Air Asia flights are departing in the evening and they have less than 4 stops during the journey.

****

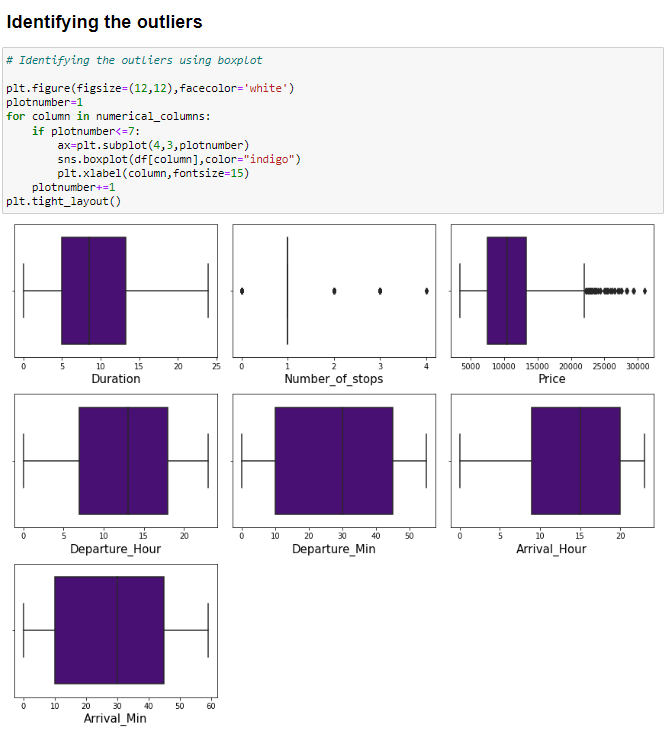
### Observations:

* All the airlines provides free meals during the journey having the duration below 11 hours.

****

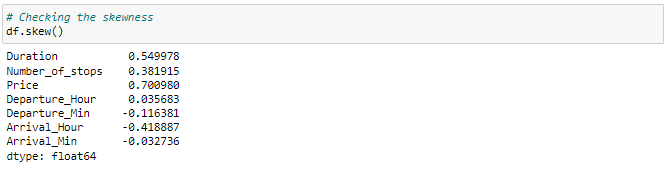
* This pair plot gives the pairwise relation between the columns, we can observe the relation between the features
* Here we can observe the correlation between the features and on the diagonal we can notice the distribution plot which shows whether the column has skewness or not.

**Identification of possible problem-solving approaches (methods)**

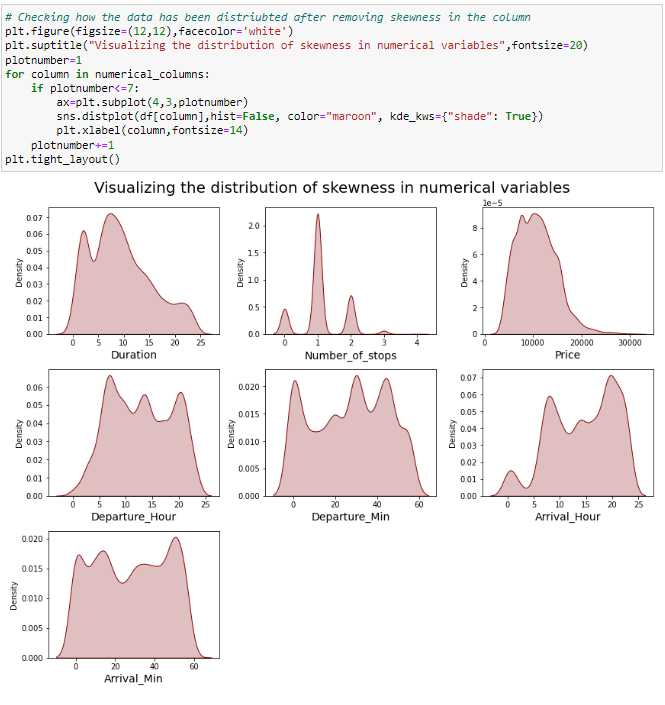
****

* The outliers present in Number\_of\_stops and "Price" columns.
* Since Price is our target column and Number\_of\_stops is our categorical variable so no need to remove outliers in this columns. Finally there is no need to remove outliers in the dataset.

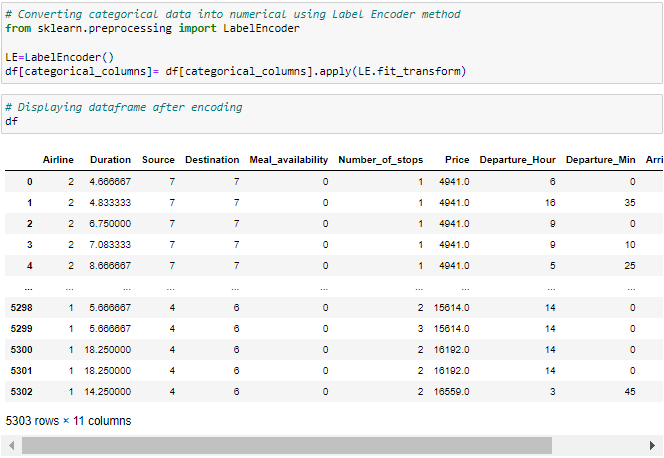
Checking for skewness in the data

****

We can find the skewness in Duration column and Price column. Price is our target variable we should not loose any data so, no need to remove skewness in this column. The skewness in Duration column is also near normal so, let's not remove skewness in Duration column.

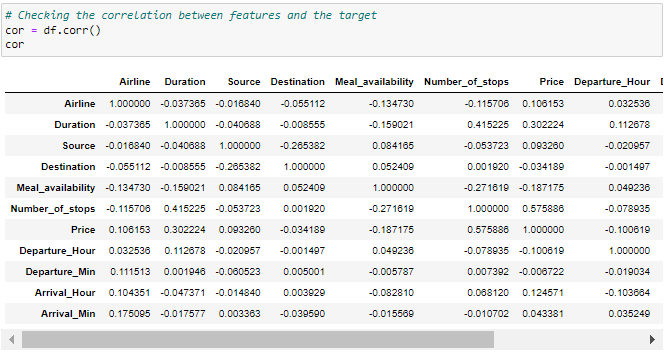
****

## Encoding the categorical columns using Label Encoder Method

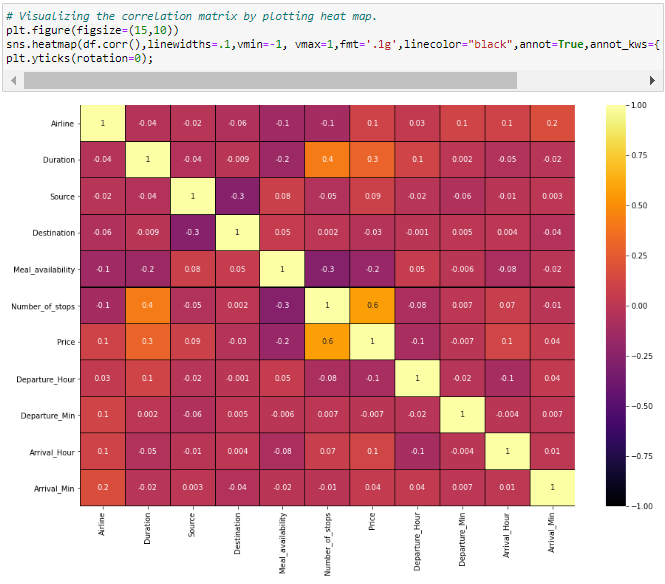
****

Now we have converted the categorical columns into numerical columns using label encoding method.

## Correlation between the target variable and independent variables using HEAT map

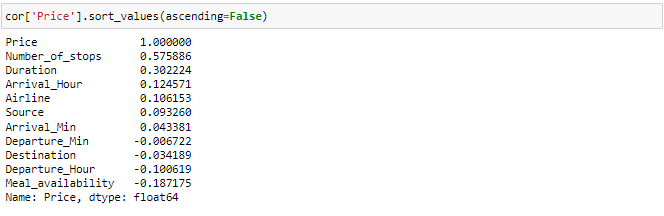
****

This gives the correlation between the dependent and independent variables. We can visualize this by plotting heat map.

****

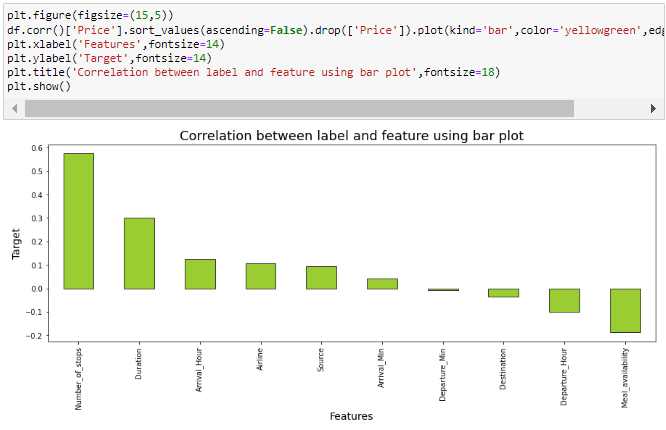
This heatmap shows the correlation matrix by visualizing the data. we can observe the relation between one feature to other.

* This heat map contains both positive and negative correlation.
* The features Number\_of\_stops, Duration Arrival\_Hour and Airline are highly positively correlated with the target column compared to other features.
* The other features have very less correlation with the target column.
* From the map we can also observe there is no multicollinearity issue exists.

****

Here we can notice the positive and negative correlation between features and label in the descending order.

## Visualizing the correlation between label and features using bar plot

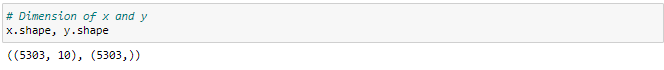
****

From the bar plot we can clearly observe the positive and negative correlation between the label and features. Here the column "Departure\_Min" has less correlation with the label compared to other features, we can drop this column if necessary but for now let's keep it as it is.

## Separating the feature and label into x and y

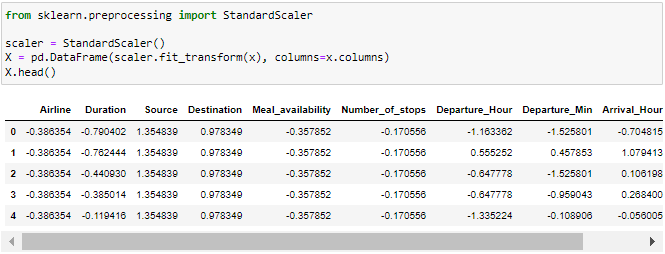
****

We have separated both dependent and independent variables.

****

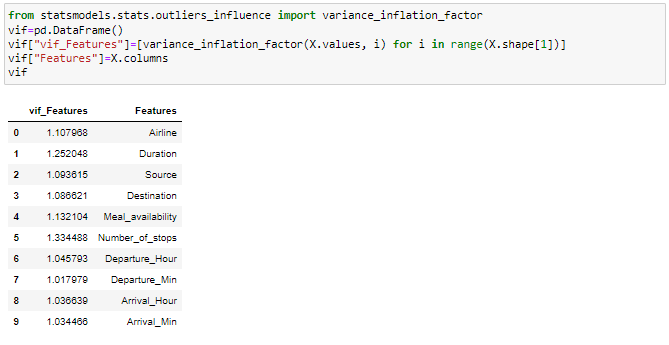
After data cleaning and preprocessing we are left with 10 columns which we are using to train our machine learning model for predicting the ticket price of the flights.

## Feature Scaling Using StandardScaler

****

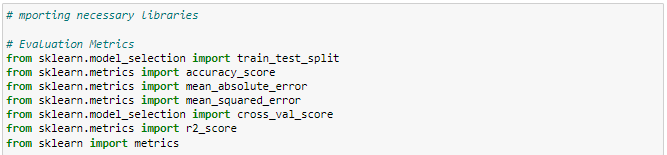
We have scaled the data using StandardScaler method to overcome with the issue of data biasness and displayed the data of independent variables after scaling.

# Checking for multicolinearity issue using VIF:

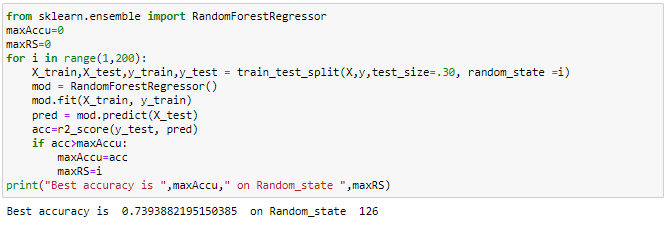
****

There is no multicolinearity issue in this dataset.

# Building Machine Learning Models

****

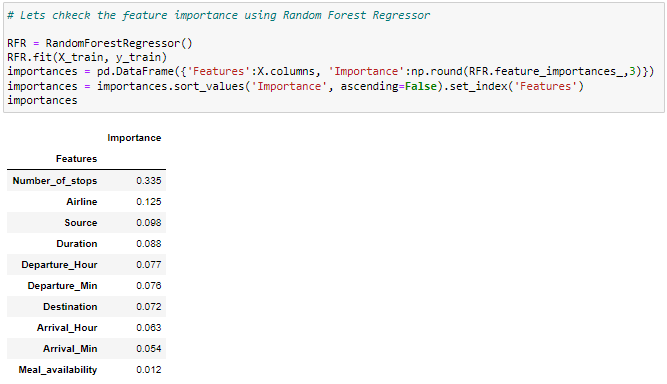
## Finding the Best Random State and Accuracy

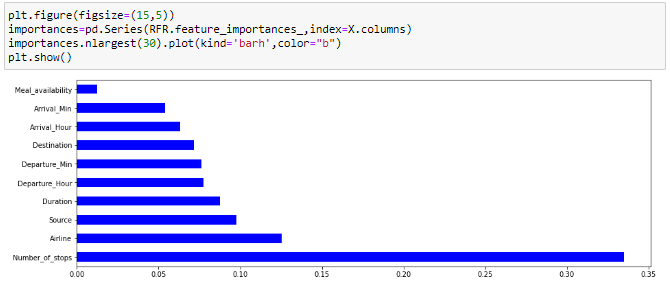
****

Best accuracy is 0.7393882195150385 on Random\_state 126

With the help of random state selection process we have found our random state to be 126 amongst 1-1000 with best accuracy as 73.93% using Random Forest Regressor.

## Feature Importance

****

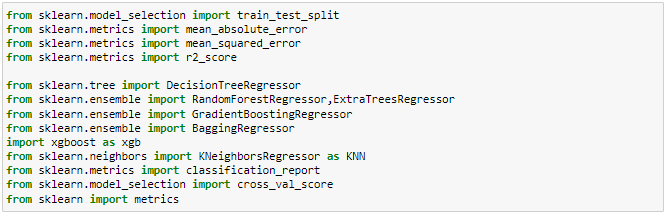
****

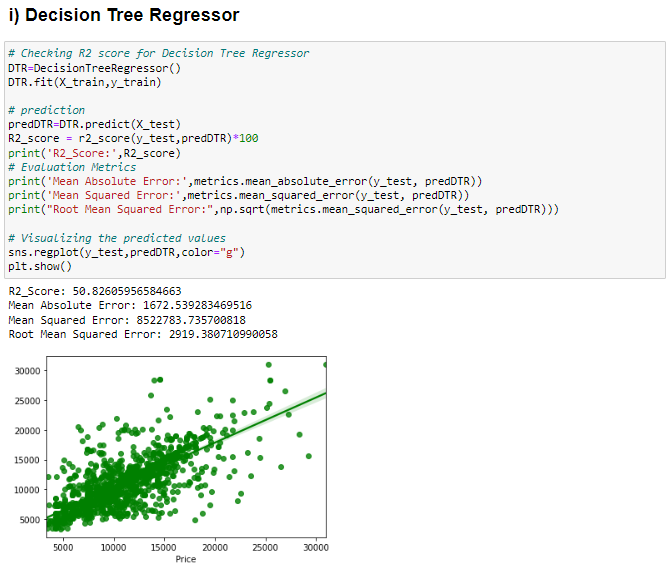
Here with the help of RandomForestRegressor we are able to list down the importance given to a column as per it's involvement in predicting our label. Here the column "Number\_of\_stops", "Airline" and "Source" contributing more for prediction which means these features are important for the predictions.

## Creating new train test split

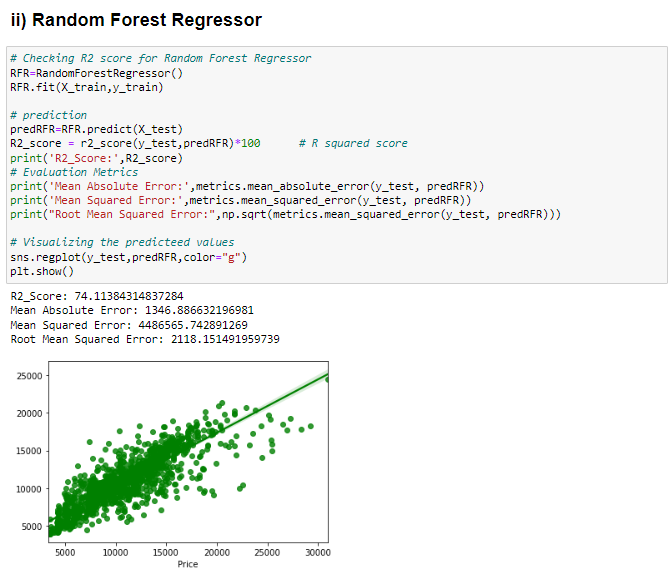
****

I am taking 30 percent of the complete dataset for training purpose and the remaining 70 percent will be used to train the machine learning models using the random state.

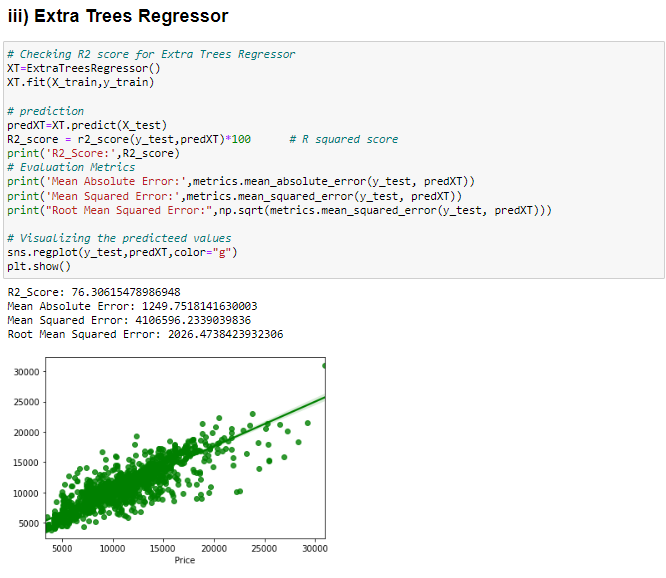
****

****

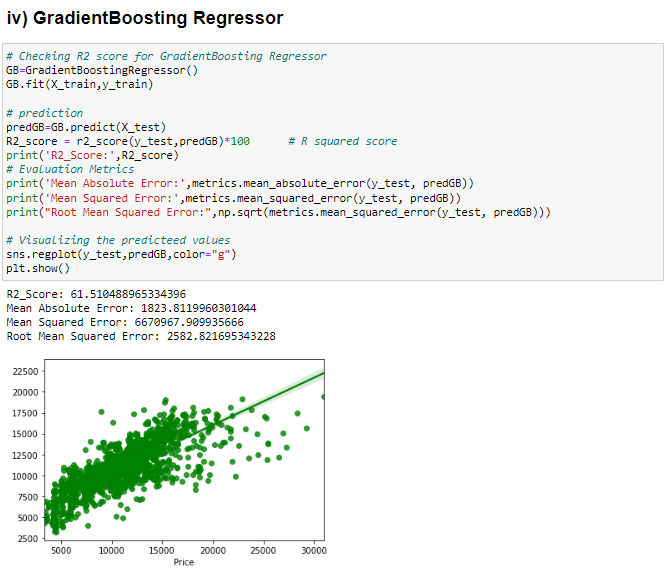
* Created Decision Tree Regressor model and checked for its evaluation metrics. The model is giving R2 score as 50.82%.
* From the graph we can observe how our model is mapping. In the graph we can observe the straight line which is our actual dataset and dots are the predictions that the model has given.

****

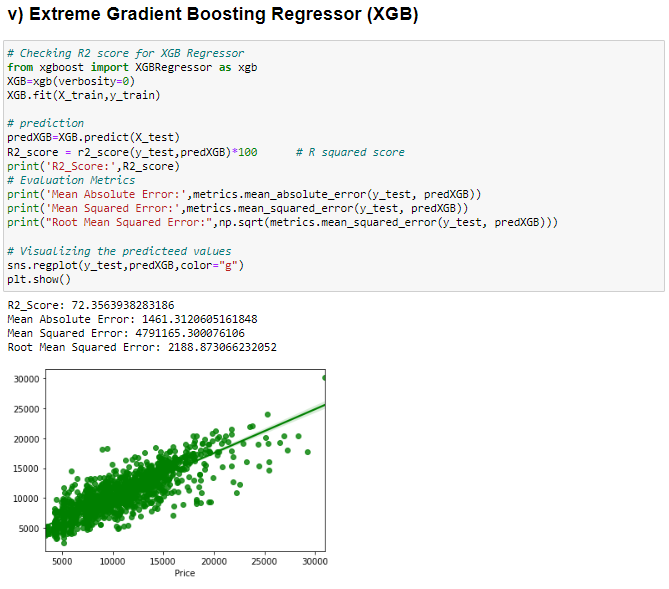
* Created Random Forest Regressor model and checked for it's evaluation metrics. The model is giving R2 score as 74.11%.
* From the graph we can observe how our model is mapping. In the graph we can observe the straight line which is our actual dataset and dots are the predictions that our model has given.

****

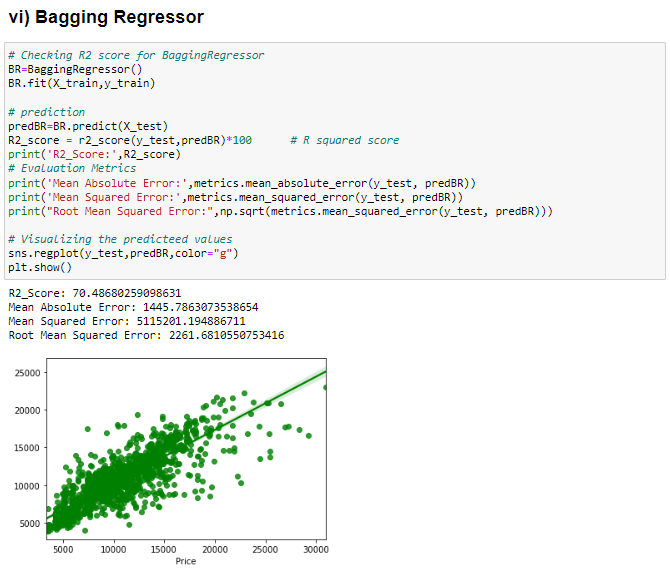
* Created Extra Trees Regressor model and checked for its evaluation metrics. The model is giving R2 score as 76.30%.
* From the graph we can observe how our model is mapping. In the graph we can observe the straight line which is our actual dataset and dots are the predictions that our model has given.

****

* Created GradientBoosting Regressor model and checked for its evaluation metrics. The model is giving R2 score as 61.51%.
* From the graph we can observe how our model is mapping. In the graph we can observe the straight line which is our actual dataset and the dots are the predictions that our model has given.

****

* Created XGB Regressor model and checked for its evaluation metrics. The model is giving R2 score as 72.35%.
* From the graph we can observe how our model is mapping. In the graph we can observe the straight line which is our actual dataset and the dots are the predictions that our model has given.

****

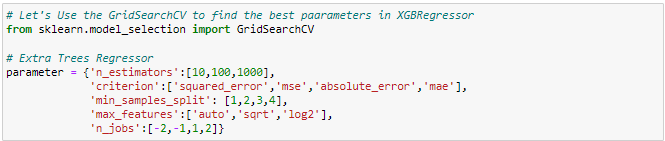
* Created Bagging Regressor model and checked for its evaluation metrics. The model is giving R2 score as 70.48%.
* From the graph we can observe how our model is mapping. In the graph we can observe the straight line which is our actual dataset and the dots are the predictions that our model has given.

## **Model Selection**

**From the above created models, Extra Trees Regressor algorithm has high R2 score and less RMSE value. So, we can conclude that "Extra Trees Regressor" as the best fitting model. Let's try to increase our model score by tuning the best model using different types of hyper parameters.**

**Testing of Identified Approaches (Algorithms)**

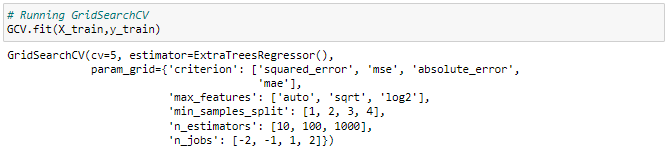
## Hyper Parameter Tuning

****

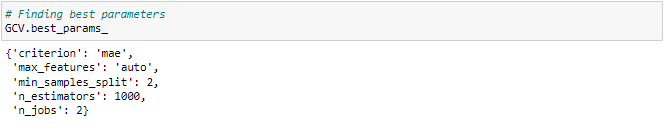
I have used 5 Extra Trees Regressor parameters to be saved under the variable "parameter" that will be used in GridSearchCV for finding the best output.

****

Assigning a variable to the GridSearchCV function after entering all the necessary inputs.

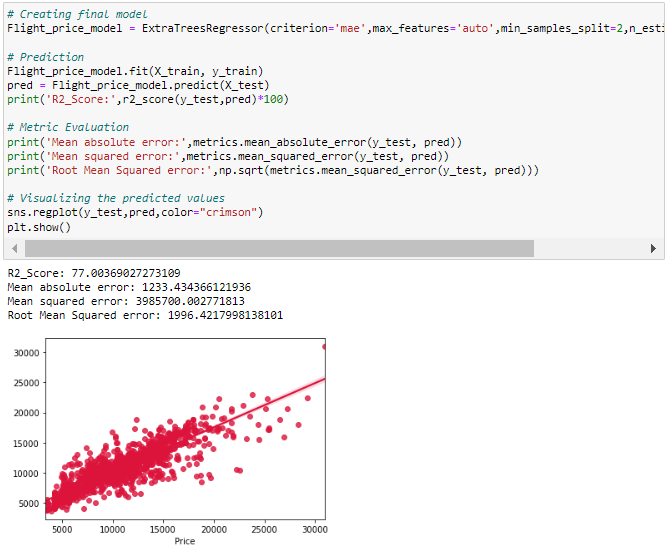
****

Now we use our training data set to make the GridSearchCV aware of all the hyper parameters that needs to be applied on our best model.

****

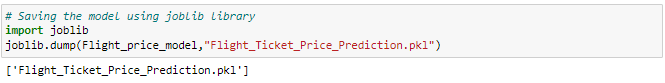
This gives us the list of best parameters which will be used further in our final model creation.

**Run and Evaluate selected models**

****

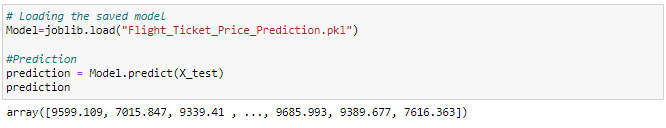
* We have successfully incorporated the hyper parameter tuning using best parameters of Extra Trees Regressor and the R2 score of the model has been increased after hyperparameter tuning and received the R2 score as 77% which is very good.
* From the graph we can observe how our final model is mapping. In the graph we can observe the best fit line which is our actual dataset and the dots are the predictions that our best final model has given.

## Saving the Final model

****

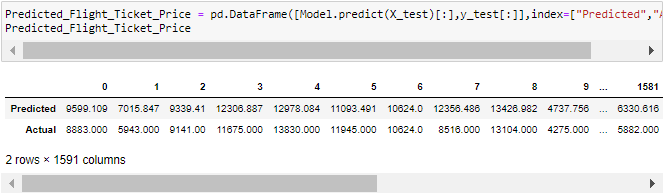
I am using the joblib option to save the final regression model in the form of .pkl.

## Loading the saved model and predicting Flight Ticket Price

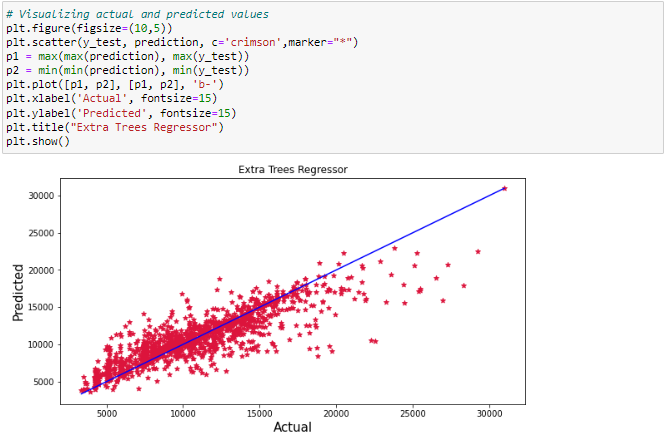
****

These are the predicted price of the flight tickets.

## Creating DataFrame for the predicted values

****

Using regression model, we have got the predicted price of the flight tickets. From the above output we can observe that predicted values are almost near to the actual values.

****

The graph shows how our final model is mapping. The plot gives the linear relation between predicted and actual price of the flight tickets. The blue line is the best fitting line which gives the actual values/data and red dots gives the predicted values/data.

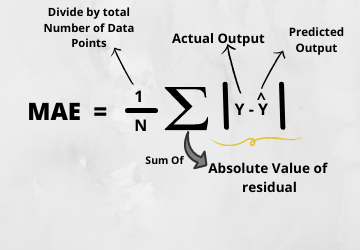
****

We have saved the predicted flight ticket price values in csv file

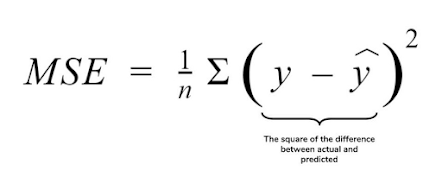
**Key Metrics for success in solving problem under consideration**

The essential step in any machine learning model is to evaluate the accuracy and determine the metrics error of the model. I have used Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R2 Score metrics for my model evaluation:

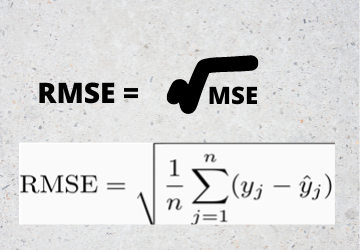
* **Mean Absolute Error (MAE):** MAE is a popular error metric for regression problems which gives magnitude of absolute difference between actual and predicted values. The MAE can be calculated as follows:



* **Mean Squared Error (MSE):** MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value. We perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.



* **Root Mean Squared Error (RMSE):** RMSE is an extension of the mean squared error. The square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.



* **R2 Score:** I have used R2 score which gives the accurate value for the models used. On the basis of R2 score I have created final model.

**Interpretation of the Results**

**Visualizations**: In univariate analysis I have used count plots and pie plots to visualize the counts in categorical variables and distribution plot to visualize the numerical variables. In bivariate analysis I have used bar plots, strip plots, line plots, reg plots, box plots, and box plots to check the relation between label and the features. Used pair plot to check the pairwise relation between the features. The heat map and bar plot helped me to understand the correlation between dependent and independent features. Detected outliers and skewness with the help of box plots and distribution plots respectively. And I found some of the features skewed to right as well as to left. I got to know the count of each column using bar plots.

**Pre**-**processing**: The dataset should be cleaned and scaled to build the ML models to get good predictions. I have performed few processing steps which I have already mentioned in the pre-processing steps where all the important features are present in the dataset and ready for model building.

**Model** **building**: After cleaning and processing data, I performed train test split to build the model. I have built multiple regression models to get the accurate R2 score, and evaluation metrics like MAE, MSE and RMSE. I got Extra Trees Regressor as the best model which gives 76.30%R2score.After tuning the best model, the R2 score of Extra Trees Regressor has been increased to77% and also got low evaluation metrics. Finally, I saved my final model and got the good predictions results for price of flight tickets.

**Conclusion**

**Key Findings and Conclusions of the Study**

The case study aims to give an idea of applying Machine Learning algorithms to predict the price of the flight tickets. After the completion of this project, we got an insight of how to collect data, pre-processing the data, analyze the data, cleaning the data and building a model. In this study, we have used multiple machine learning models to predict the flight ticket price. We have gone through the data analysis by performing feature engineering, finding the relation between features and label through visualizations. And got the important feature and we used these features to predict the car price by building ML models. Performed hyper parameter tuning on the best model and the best model’s R2 score increased and was giving R2 score as 77%. We have also got good prediction results of ticket price.

**Findings**:

* Flight ticket prices change during the morning and evening time of the day. From the distribution plots we came to know that the prices of the flight tickets are going up and down, they are not fixed at a time. Also, from this graph we found prices are increasing in large amounts.
* Some flights are departing in the early morning 3 AM having most expensive ticket prices compared to late morning flights. As the time goes the flight ticket fares increased and midnight flight fares are very less (say after 10 PM). Also, from categorical and numerical plots we found that the prices are tending to go up as the time is approaching from morning to evening.
* From the categorical plots (bar and box) we came to know that early morning and late-night flights are cheaper compared to working hours.
* From the categorical plots we found that the flight ticket prices increaseas the person get near to departure time. That is last minute flights are very expensive.
* From the bar plot we got to know that both Indigo and Spicejet airways almost having same ticket fares.
* Not all flights are expensive during morning, only few flights departing in the early morning 3 AM are expensive. Apart from this the flight ticket fares are less compared to other timing flight fares.

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

While working on this project I learned many things about the features of flights and about the flight ticket selling web platforms and got the idea that how the machine learning models have helped to predict the price of flight tickets. I found that the project was quite interesting as the dataset contains several types of data. I used several types of plotting to visualize the relation between target and features. This graphical representation helped me to understand which features are important and how these features describe price of tickets. Data cleaning was one of the important and crucial things in this project where I dealt with features having string values, features extraction and selection. Finally got Extra Trees Regressor as best model.

The challenges I faced while working on this project was when I was scrapping the real time data from yatra website, it took so much time to gather data. Finally, our aim was achieved by predicting the flight ticket price and built flight price evaluation model that could help the buyers to understand the future flight ticket prices.

**Limitations of this work and Scope for Future Work**

**Limitations:** The main limitation of this study is the low number of records that have been used. In the dataset our data is not properly distributed in some of the columns many of the values in the columns are having string values which I had taken care. Due to some reasons our models may not make the right patterns and the performance of the model also reduces. So these issues need to be taken care.

**Future work:** The greatest shortcoming of this work is the shortage of data. Anyone wishing to expand upon it should seek alternative sources of historical data manually over a period of time. Additionally, a more varied set of flights should be explored, since it is entirely plausible that airlines vary their pricing strategy according to the characteristics of the flight (for example, fares for regional flights out of small airports may behave differently than the major, well flown routes we considered here). Finally, it would be interesting to compare our system's accuracy against that of the commercial systems available today (preferably over a period of time).