1. **Problem Definition:**

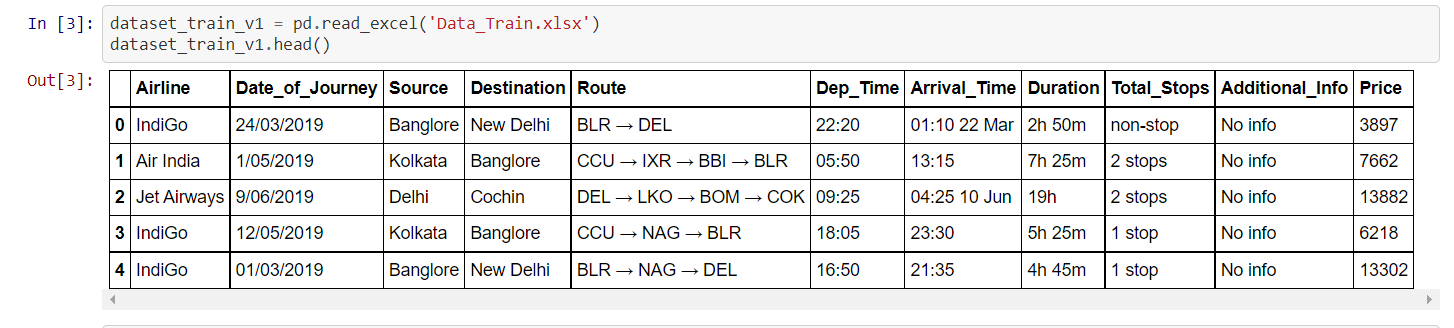
We need to predict the price of Flight based on the available prices of various airlines between the month of March and June 2019 for the applicable cities

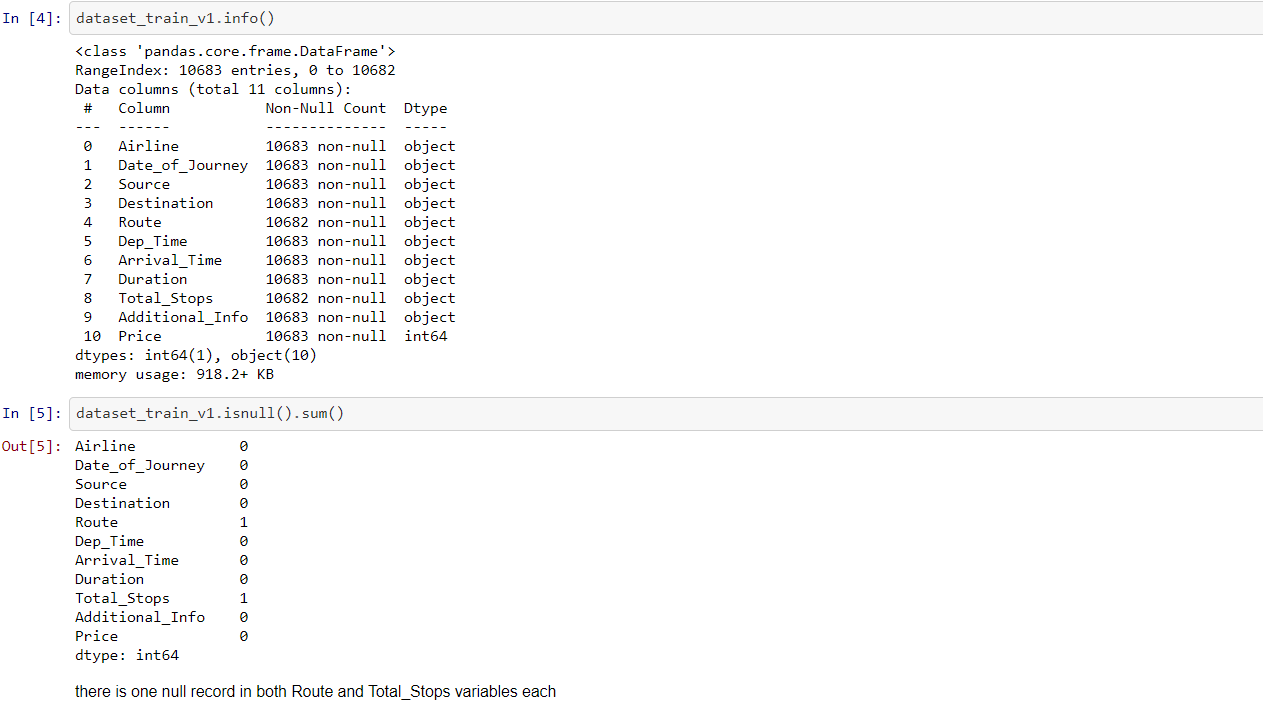
1. **Data Analysis:**

Importing the basic necessary libraries required for performing the EDA Shape, rectangle

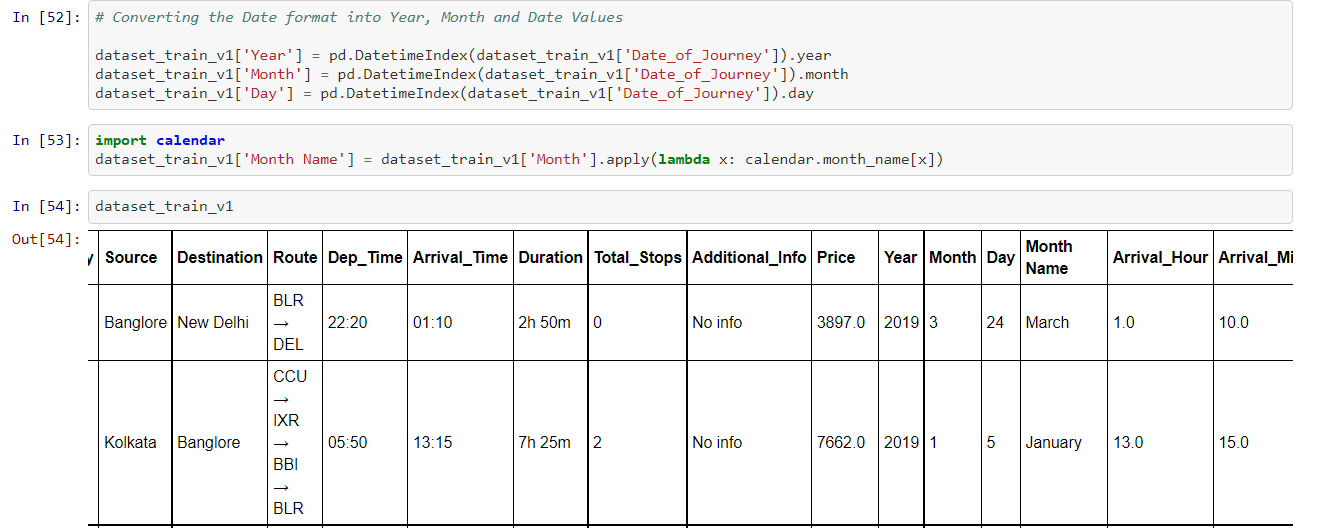
Description automatically generated with medium confidence

Loading the dataset

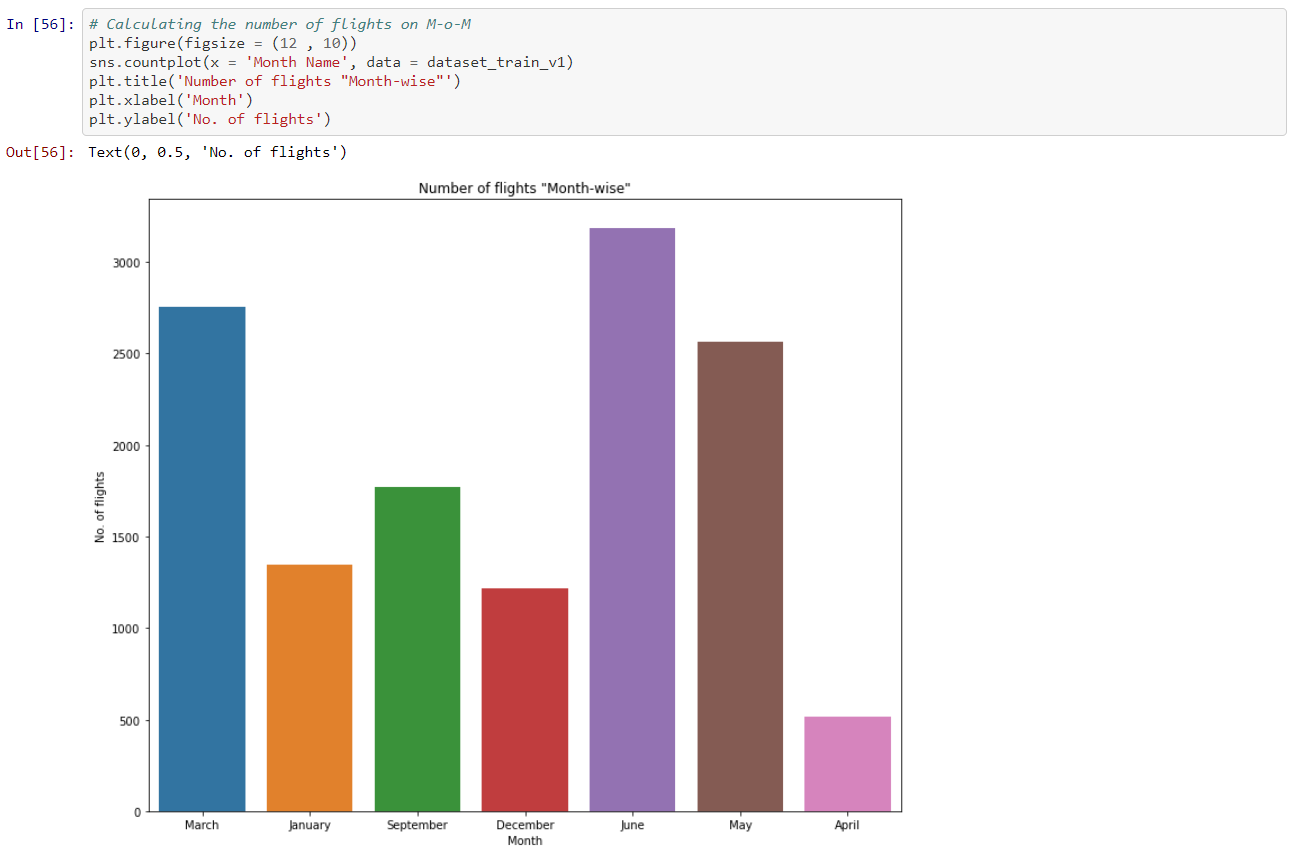




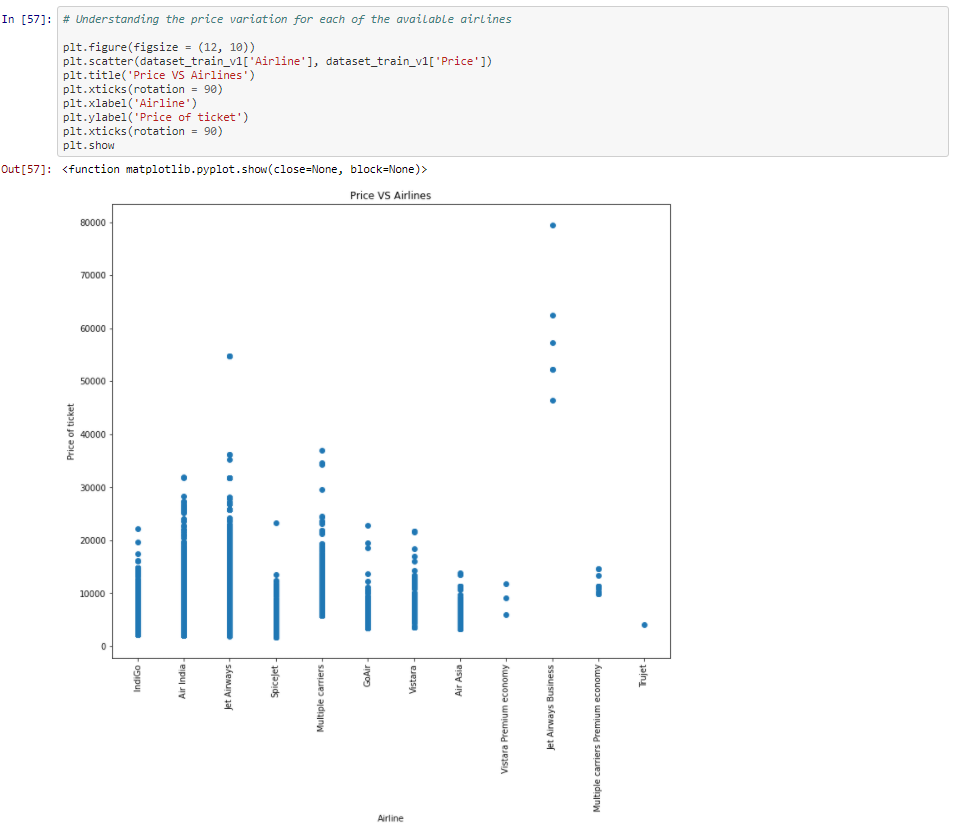
From the above analysis, there are one null record each for Route and Total\_Stops variables. This null data is negligible; hence, we can drop these records from the dataset.



Since, the target variable is a continuous value, we need to understand the change in price over the month and year. Hence, the month and year has been extracted from the date format



June had the highest number of flights followed by March and May. While April had the lowest number of active flights.

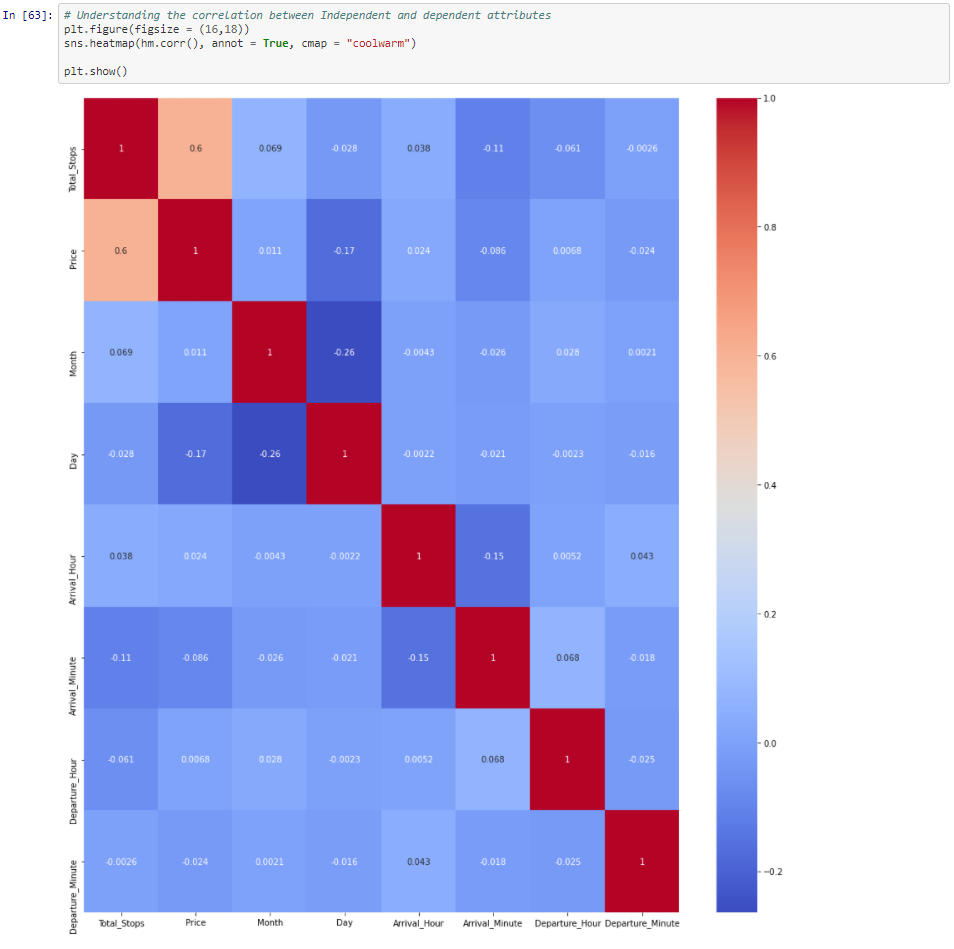


Above scatterplot shows the variation of price for each of the airlines:

* Jet Airways Business has the highest price, followed by Jet Airways.
* Spicejet has the lowest starting price.
* Trujet has single prices for all of their available flights.
* Also, Jet Airways has the maximum variation between starting price and highest price.



From the above barplot, Jet Airways has the highest number of flights, followed by Indigo and Air India.

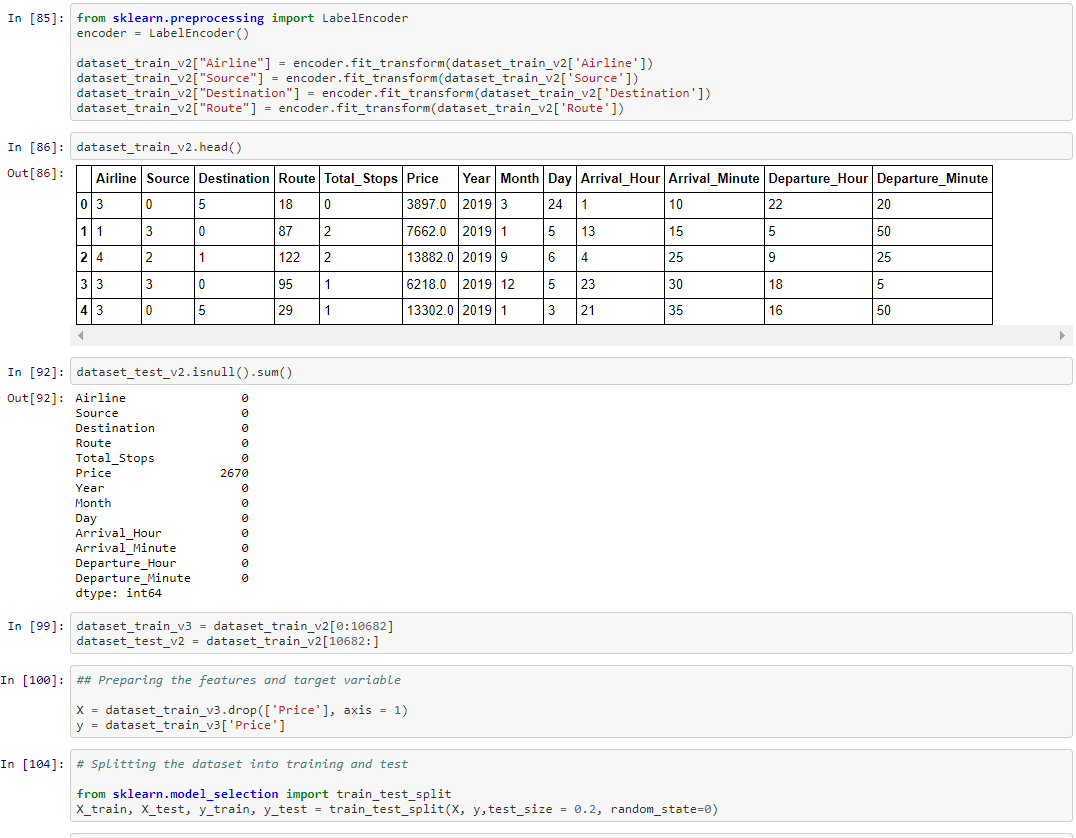


From the above heatmap, there is strong correlation between “Total\_Stops” and “Price”

1. **EDA Concluding Remarks:**

* June had the highest number of flights followed by March and May. While April had the lowest number of active flights
* Jet Airways Business has the highest price, followed by Jet Airways.
* Spicejet has the lowest starting price.
* Trujet has single prices for all of their available flights.
* Also, Jet Airways has the maximum variation between starting price and highest price.
* Jet Airways has the highest number of flights, followed by Indigo and Air India.
* There is strong correlation between “Total\_Stops” and “Price”

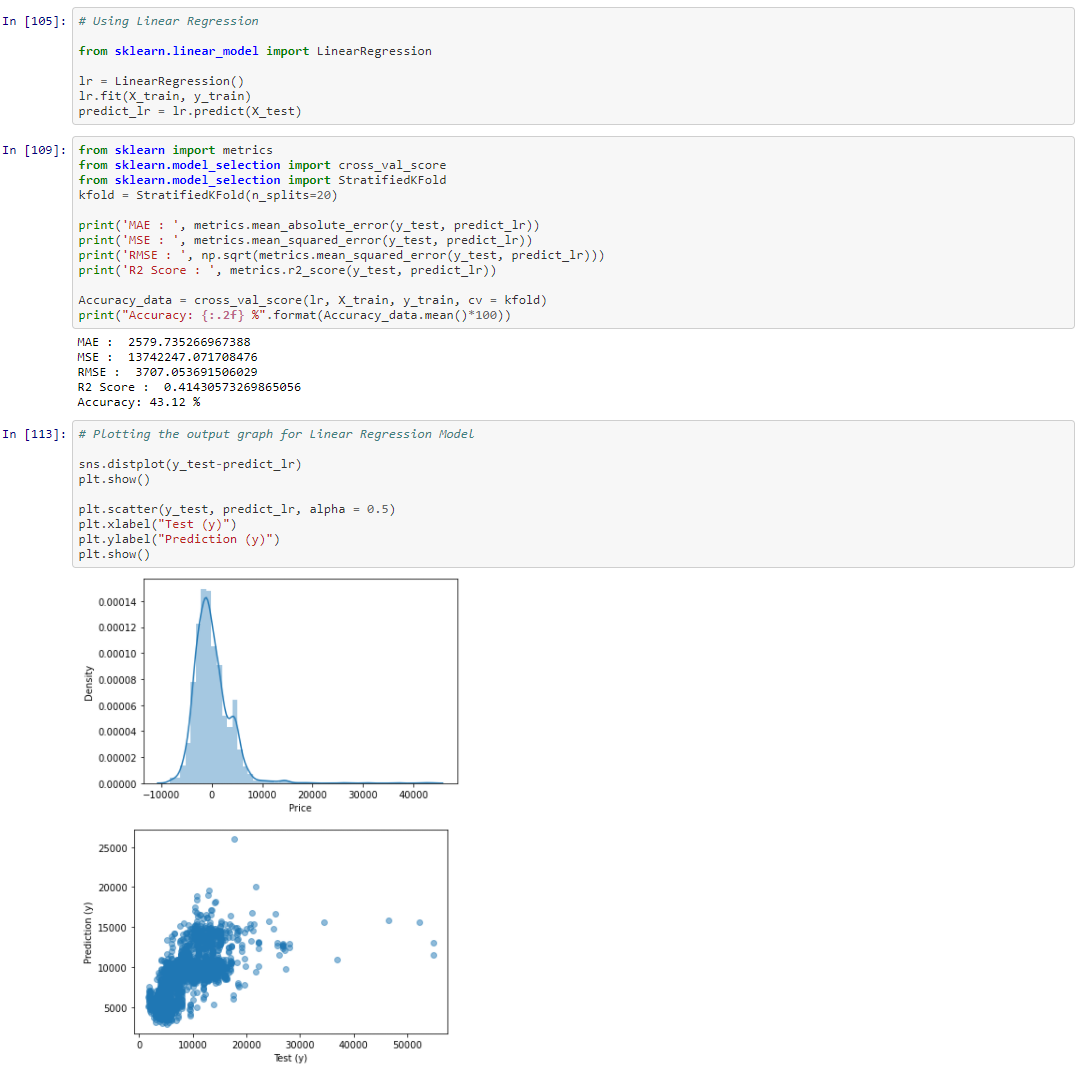
1. **Pre-Processing Pipeline:**



All the object datatypes has been transformed using label encoder and the dataset are divided into feature and target variables for both the test and training dataset.

1. **Building Machine Learning Models:**

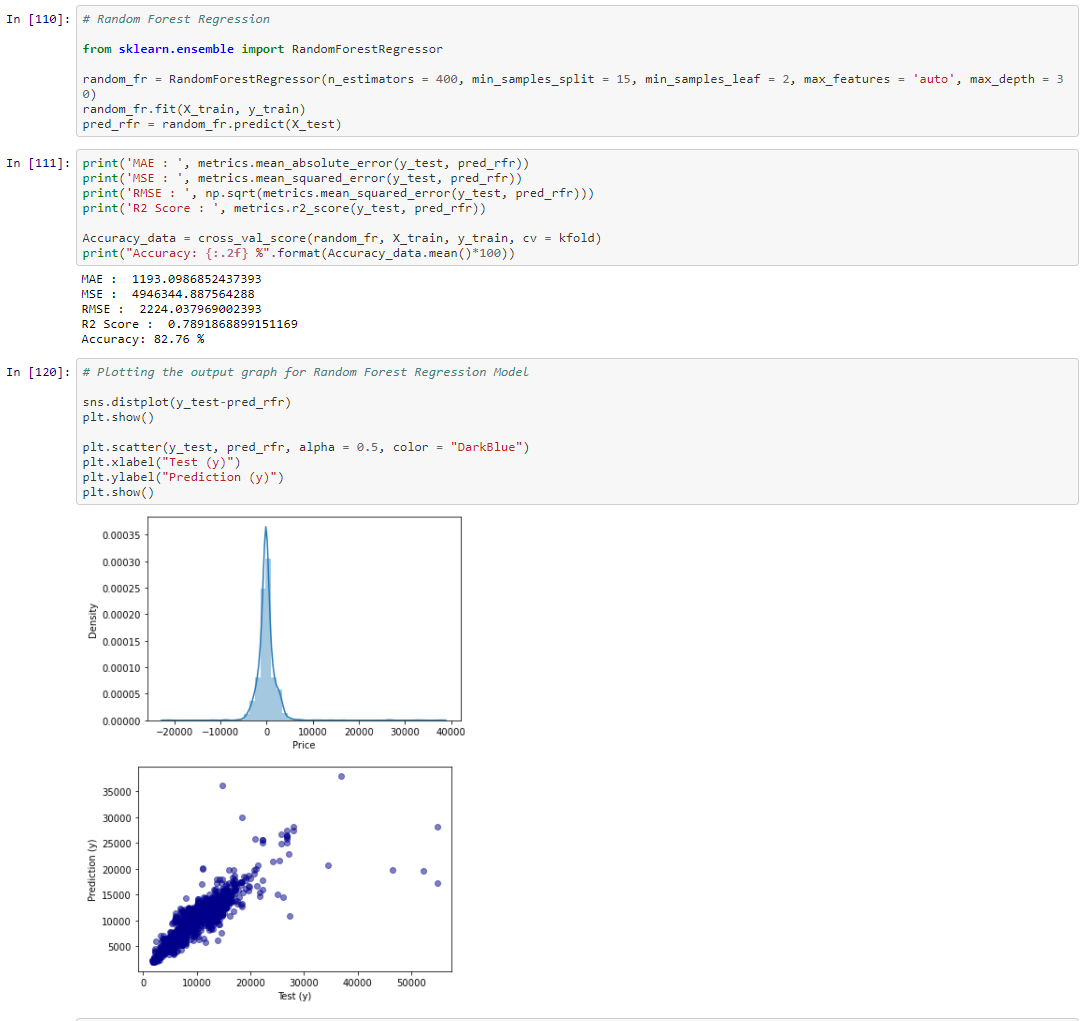
Linear Regression is one of the models selected for predicting the target variable

****

The evaluation metrics for Linear Regression is as mentioned below:

* It has the accuracy rate of 43 %
* While the R2 Score is 0.414

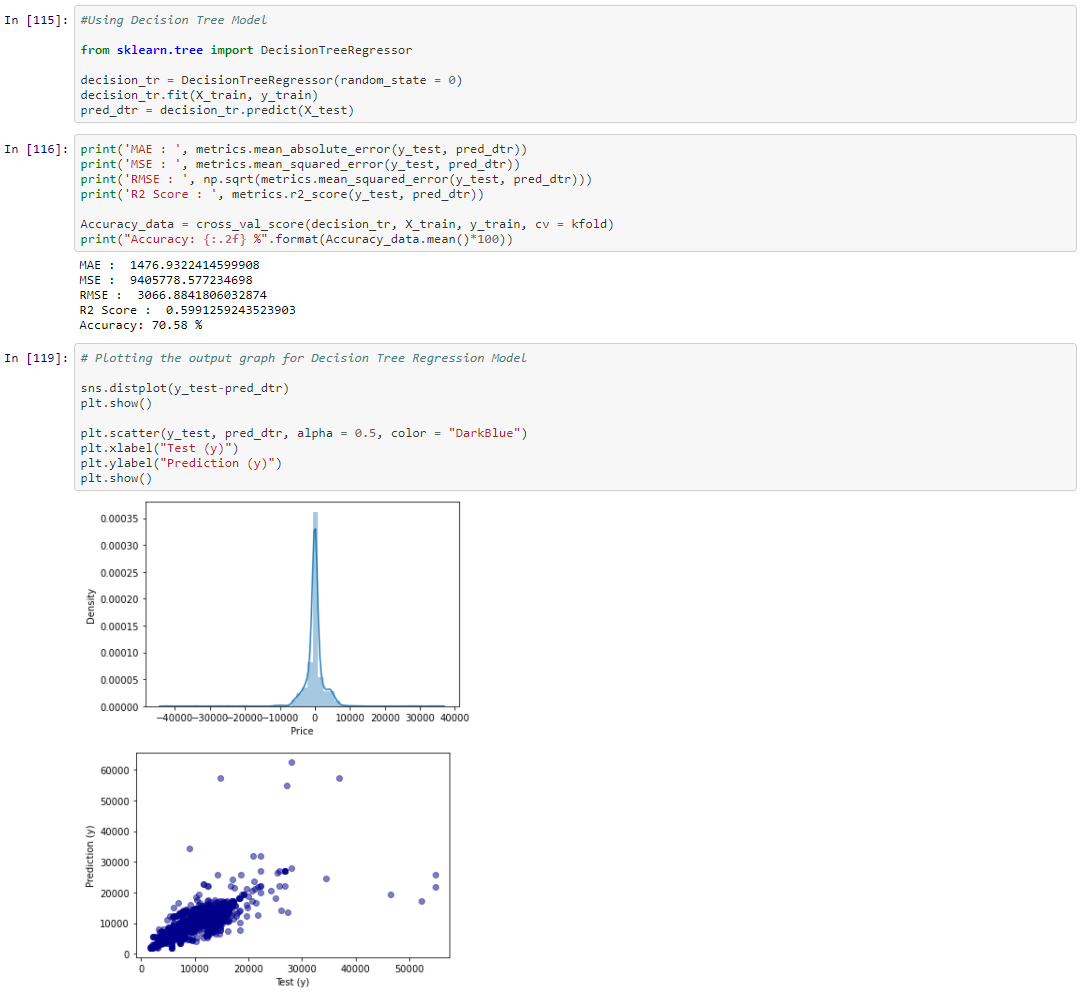
Another machine learning model used here is RandomForest Regression as mentioned below



The evaluation metrics for Random Forest Regression is as mentioned below:

* It has the accuracy rate of 83 %
* While the R2 Score is 0.789

The third machine learning model used here is Decision Tree Regression as mentioned below



The evaluation metrics for Decision Tree is as mentioned below:

* It has the accuracy rate of 71 %
* While the R2 Score is 0.599

1. **Concluding Remarks:**

|  |  |  |
| --- | --- | --- |
| **Machine Learning Models** | **R2 Score** | **Accuracy (%)** |
| Linear Regression | **0.414** | **43 %** |
| Random Forest Regression | **0.789** | **83 %** |
| Decision Tree | **0.599** | **71 %** |

Hence, as per the table, we can conclude that Random Forest Regression is the best model for the prediction of Flight price.