

FLIGHT PRICE PREDICTION

Submitted by:

KUNAL CHAND

INTERNSHIP 19

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to my teacher

Dr. Deepika Sharma from Data trained as well as Shubham Yadav (SME) from Flip Robo, who gave me the golden opportunity to do this wonderful project on the topic

**FLIGHT PRICE PREDICTION**

, which also helped me in doing a lot of research and I came to know about so many new things on the flight prices when to book and avail maximum benefits. I am thankful to them.  
Secondly, I would also like to thank my parents who helped me a lot in finalizing this project within the limited time frame.

**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 -

Time of purchase patterns (making sure last-minute purchases are expensive)

Keeping the flight as full as they want it (raising prices on a flight which is filling up to reduce sales and hold back inventory for those expensive last-minute expensive purchases) So, you must 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

The conceptual background of the problem depends on the time when will book a flight tickets to avail maximum discounts.

* Review of Literature

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data, and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can analyse data collected by multiple data sources and stored in data warehouses or any scrapped data from websites, an organization must know how many cases are in a data set, what variables are included, how many missing values there are, and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working. We divided the data 8:2 for Training and Testing purposes respectively.

* Motivation for the Problem Undertaken

Every problem of Machine learning gives us chance to enhance and develop problem-solving skills. These Problems do’s the same.

When this real-life problem of predicting the flight prices for the future which time and date is best for avail maximum discounts, all the scraped data is new and predict the future prices and with help of A. I technology we make a completely new model of prediction. As Data scientists it is our role to help and understand the market better with newer data, for constructing real-life helpful models for companies and individual’s.

we 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 must 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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

As for any basic model building, we must understand the type of target variable, the data of the target variable is continued or classified.

Data Analysis is always the difficult part, for better understanding different kinds of bar plots, distribution plots are created with the target Column for finding the insights of the dataset we have.

Analytical Modelling always starts with the target variable we have, and in that case, our target variables Price attribute. first, we must filter them and make data clean then using different analysis tools select the best features which makes our prediction more accurate, we create some distribution plots with the target variable to understand which feature columns help to learn the model best and which feature columns reduce the accuracy of the model.

And after finding the relation and correlation with the target variable we choose either Regression Model or Classification Model. Here in this problem, our target feature column is continuing so we build our Machine Learning model on regression.

* Data Sources and their formats

Data Set Description

The data set contains data of flights of February month, which has approximately 1500 rows of data.

The data set includes:

'AIRLINES', 'DATE OF JOURNEY', 'SOURCE', 'DESTINATION','DEPARTURE TIME', 'ARRIVAL TIME', 'DURATION', 'TOTAL STOPS', 'PRICE'

FEATURES:

Airline: The name of the airline.

Date of Journey: The date of the journey

Source: The source from which the service begins.

Destination: The destination where the service ends.

Dep Time: The time when the journey starts from the source.

Arrival Time: Time of arrival at the destination.

Duration: Total duration of the flight.

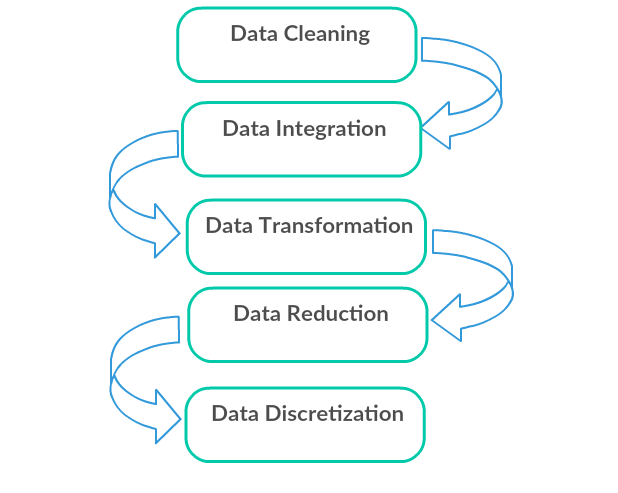
Total Stops: Total stops between the source and destination.

Price: The price of the ticket

This project is more about exploration, feature engineering, and classification that can be done on this data. Since the data set having not much data but we try to find all the insights and answer basic questions as asked in this project.

* Data Preprocessing Done

Data pre-processing can refer to the manipulation or dropping of data before it is used to ensure or enhance performance and is an important step in the data mining process.



1. Data Cleaning: First we clean the data which have no use in prediction like the ID column, and then we drop the data which has a high no of missing percentages.
2. Data Integration: then we do some EDA process for finding out the meaning full insights of the data.
3. Data transformation is the process of changing the format, structure, or values of data; we use a labelled encoder for coding the object data into integer data.
4. Data Reduction: it is the process of finding the most correlated columns and combining them because the machine does not understand which feature columns impact the most on accuracy.
5. Data discretization converts many data values into smaller once, so that data evaluation and data management become very easy, using box plots is makes a clear understanding of the data.

* Data Inputs- Logic- Output Relationships

Regression and classification models are important tools for researchers in various fields. The application of these many-to-one mapping models is two-fold. First, they can be used for prediction. The output value or class of a (new) case can be predicted by applying the inferred mapping to the input variables of the case. Second, they inform us about the relationship between the input and the output. They specify how the input variables are (mathematically) interacting with each other to produce the output variable. The usefulness of the second application is, however, limited by the power of the human intellect. We suggest that the interpretation of these many-to-one mapping models is of utmost, yet undervalued, importance in many research fields.

* State the set of assumptions (if any) related to the problem under consideration

As we understand that many factors affecting the prices of data but here we don’t make ant assumptions because our data set is not big enough.

* Hardware and Software Requirements and Tools Used

Python is widely used in scientific and numeric computing:

SciPy is a collection of packages for mathematics, science, and engineering.

Pandas are data analysis and modelling libraries.

Matplotlib, Plotly are visualization libraries

Libraries Used for this Project include –

* 1. Pandas
* 2. NumPy
* 3. Matplotlib
* 4. Seaborn
* 5. Scikit Learn
* 6. Plotly

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

After analysing the dataset, I observe that many of the feature columns are object type and date of journey is timestamp data type. so first, we must convert them into an integer so that the machine interprets the data and for that, we use label encoding.

Then find the correlation between the columns with target columns and delete the non-related feature columns.

After converting text into int datatype and classes are defined.

The target column is continuous, so we start work on Regression models building.

* Testing of Identified Approaches (Algorithms)

1. Logistic Regression
2. Lasso regression
3. Ridge regression
4. Decision Tree
5. Random forest Regression.
6. Support vector regression
7. Gradient boosting
8. K-nearest Neighbours

# 

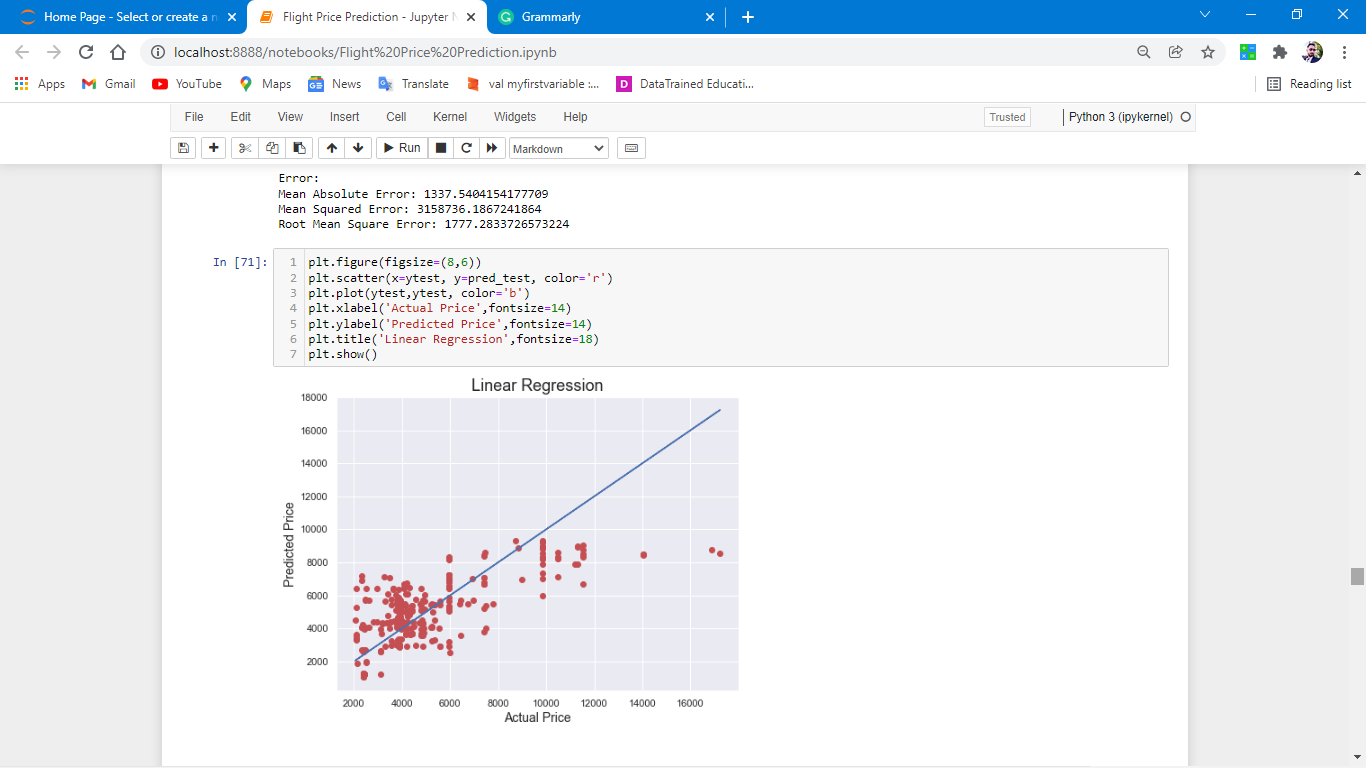
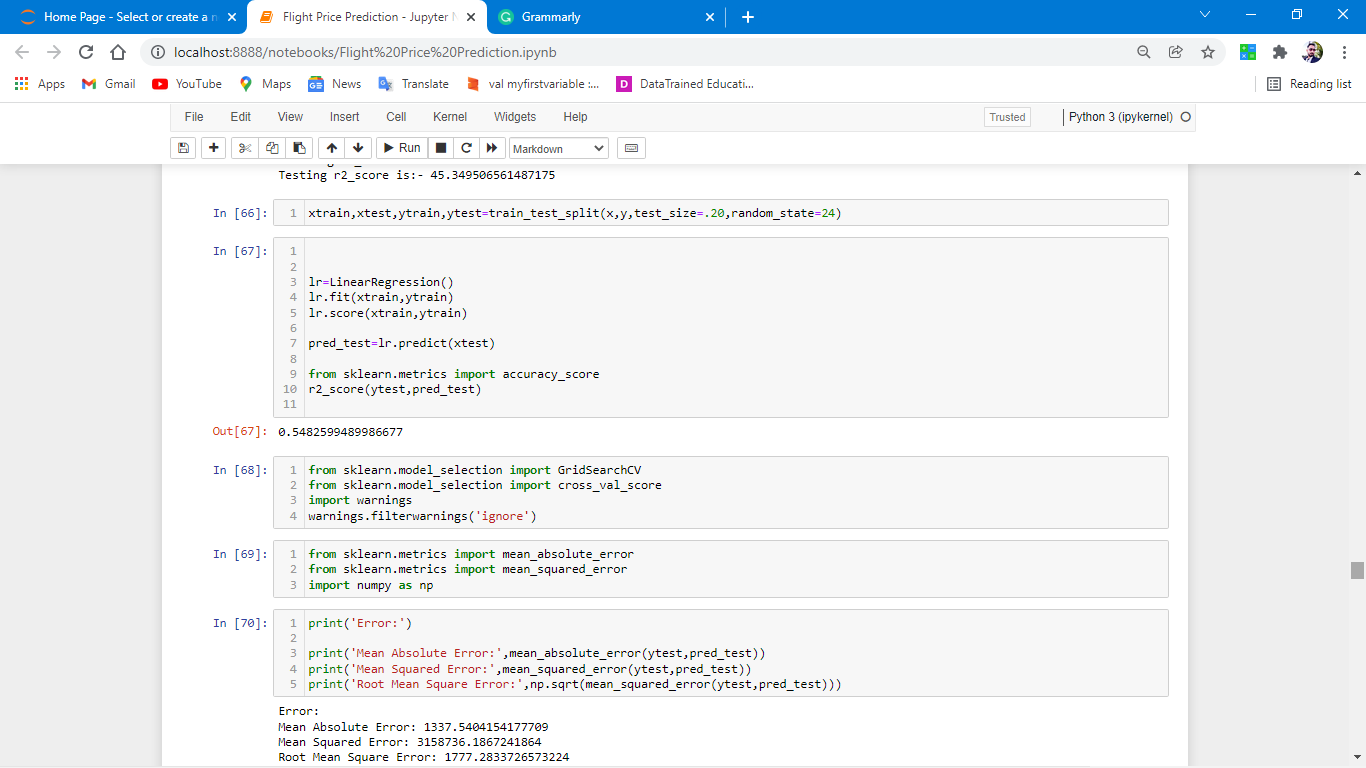
* Run and Evaluate selected models

1.Linear regression

Linear Regression is a machine learning algorithm based on supervised learning.

It performs a regression task. Regression models a target prediction value based on independent variables.

It is mostly used for finding out the relationship between variables and forecasting.





Error:

Mean Absolute Error: 1337.5404154177709

Mean Squared Error: 3158736.1867241864

Root Mean Square Error: 1777.2833726573224

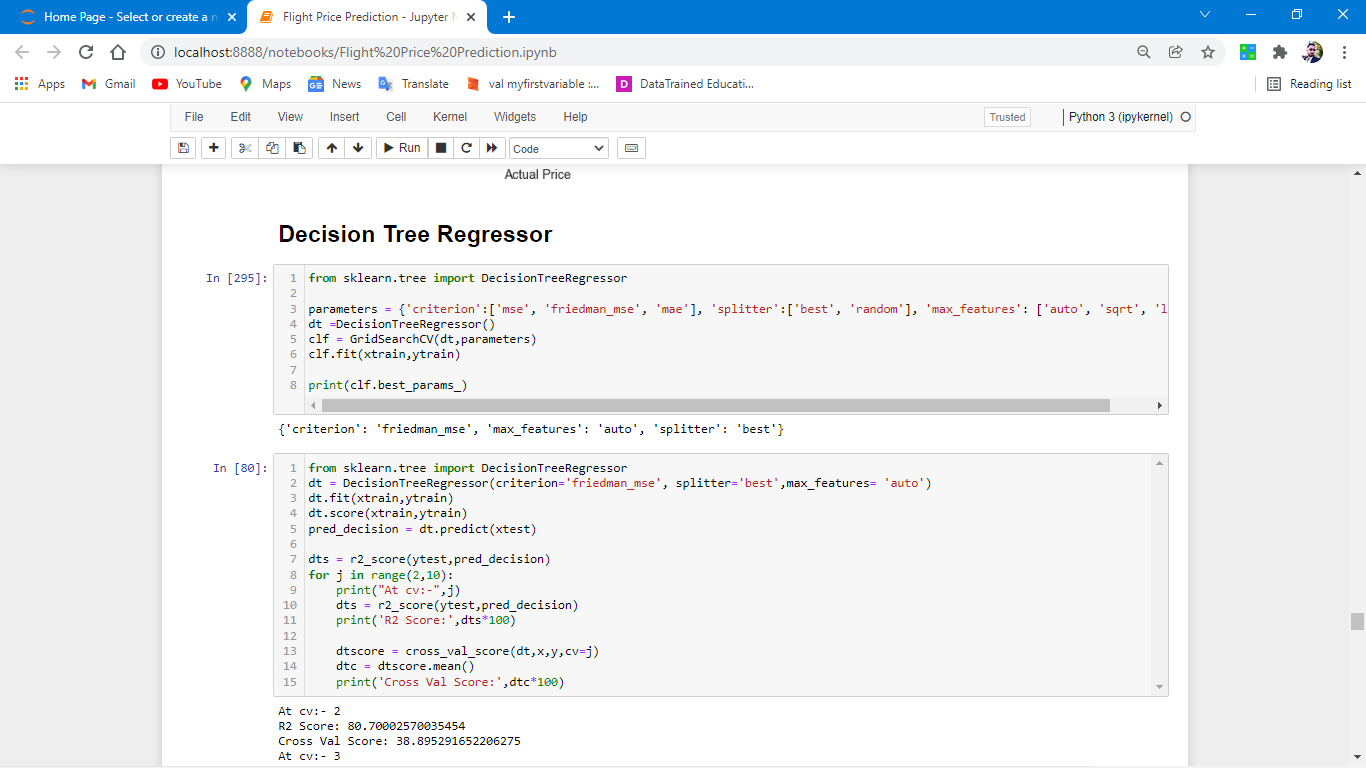
**Observations:**

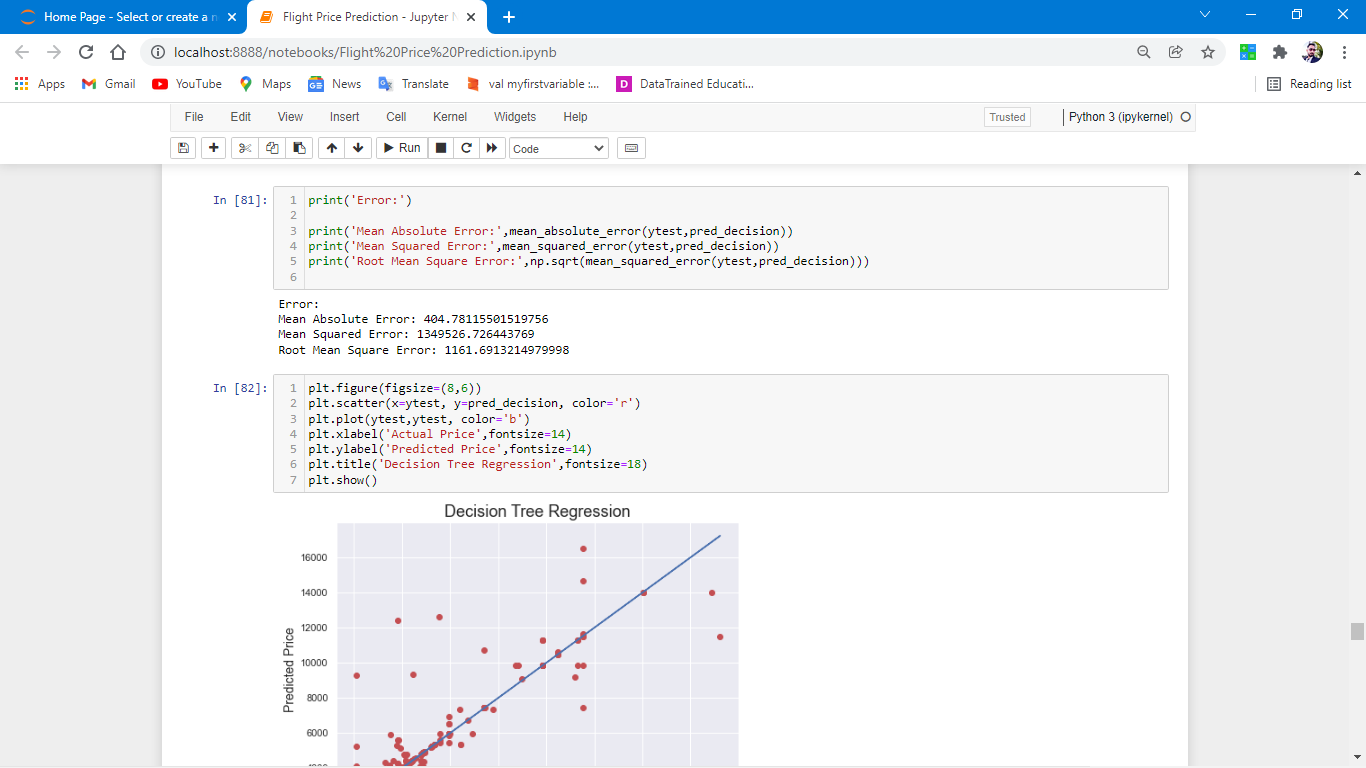
1. Linear regression gives us 50% accuracy which is very less in terms of prediction of flights prices.
2. Data is distributed near best fit line as shown is image.

Decision Tree Regression

A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

The result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g. Normal, Abnormal, and Family), each representing values for the attribute tested. Leaf node (e.g., present or not-present) represents a decision on the numerical target. The topmost decision node in a tree that corresponds to the best predictor is called the **root node**. Decision trees can handle both categorical and numerical data.







Error:

Mean Absolute Error: 404.78115501519756

Mean Squared Error: 1349526.726443769

Root Mean Square Error: 1161.6913214979998

OBSERVATION:

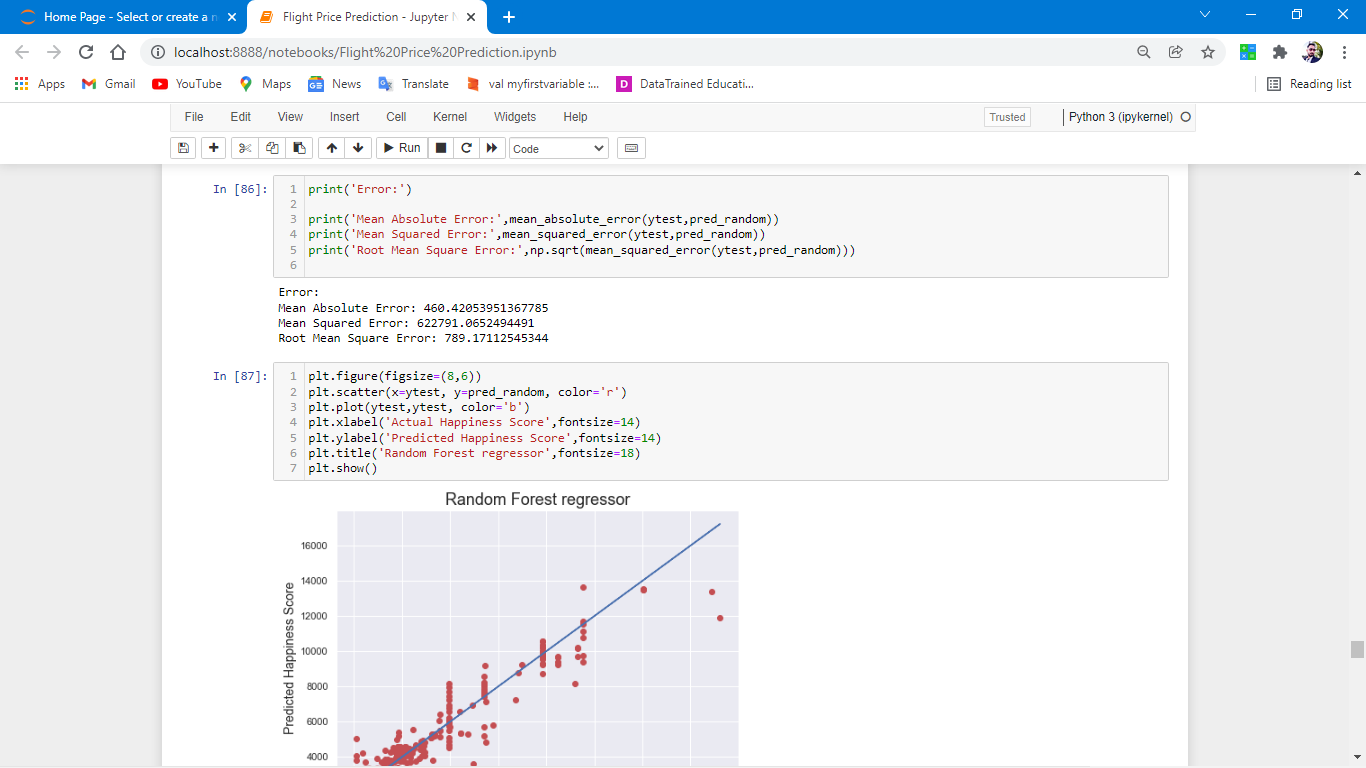
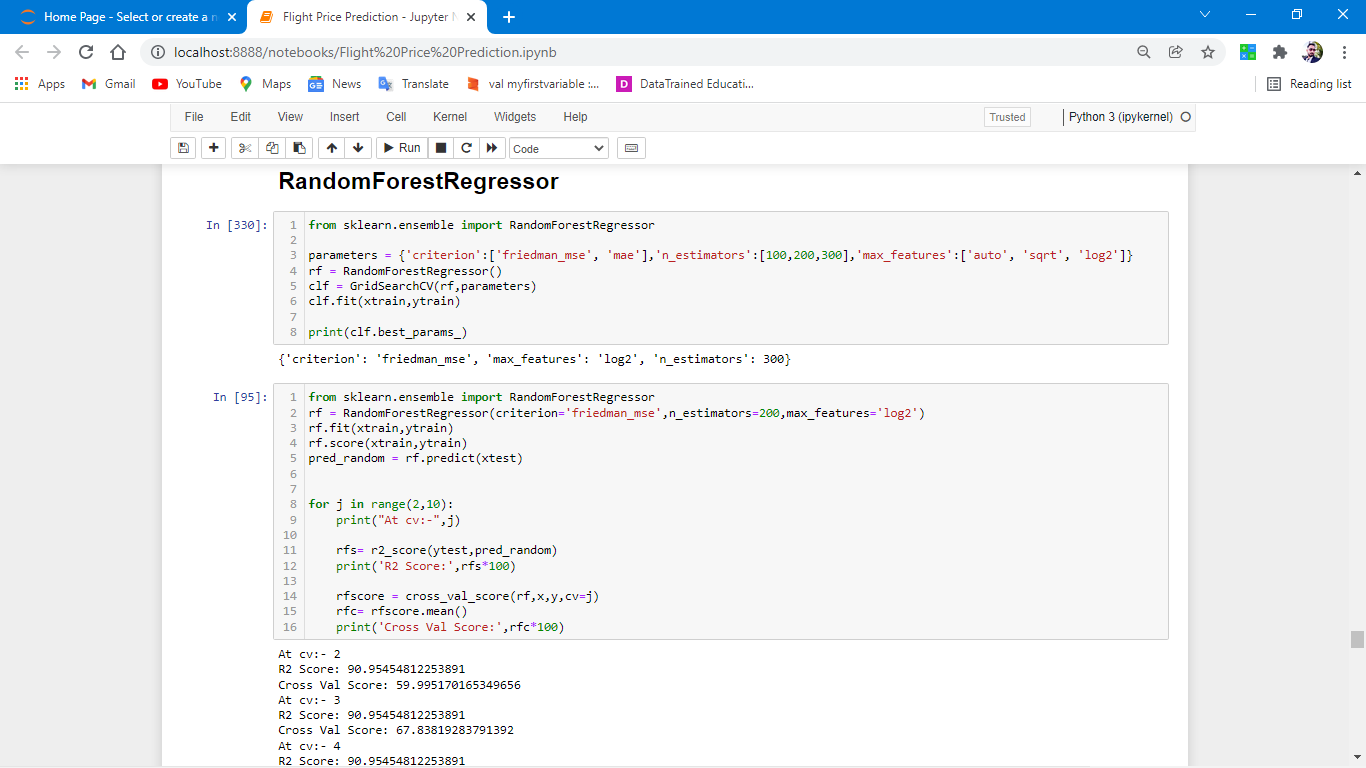
1. R2 score is increased which is near by 80.
2. More number of dots are present on best fit line.

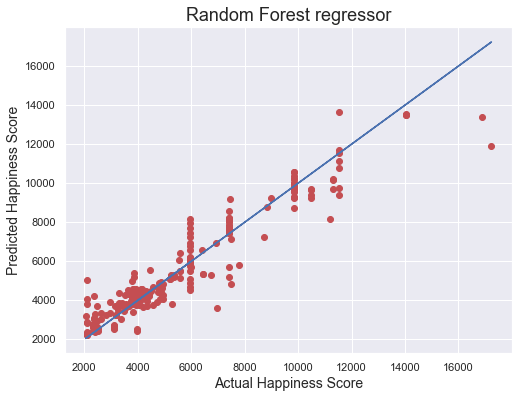
**Random Forest Regression Model**

1. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.

2. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.

3. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.





Error:

Mean Absolute Error: 460.42053951367785

Mean Squared Error: 622791.0652494491

Root Mean Square Error: 789.17112545344

Observation:

1. This model performs best comparison from the rest.
2. R2 score is over 90% which is good for this dataset.
3. More no of dots are present or near by the best fit line.

* Key Metrics for success in solving problem under consideration

Mean Absolute Error

Mean Squared Error

Root Mean Square Error

* Visualizations

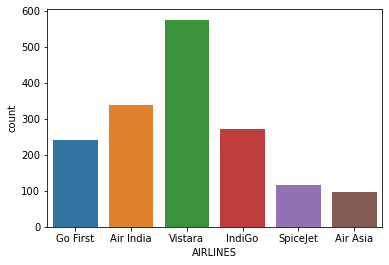
Data visualization is the graphical representation of information and data. By using charts, plots, and graphs data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

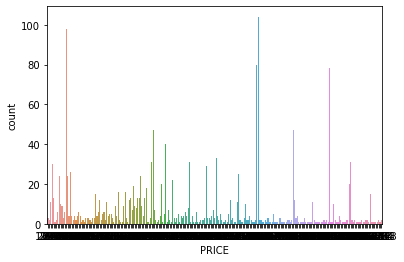
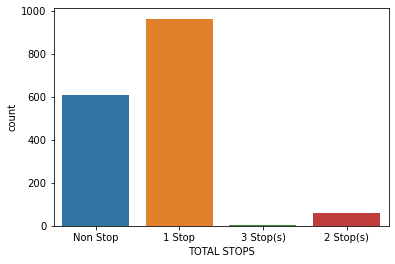
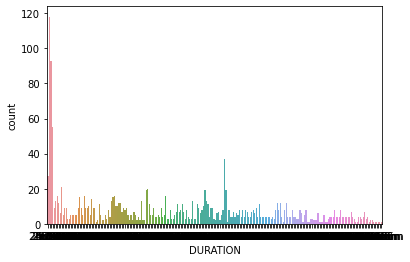
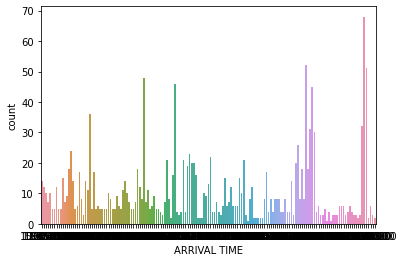
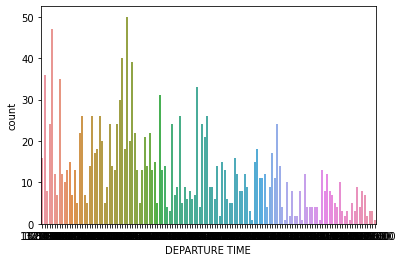
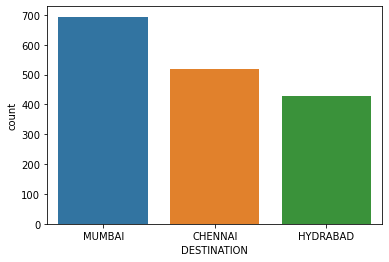
In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions

**Univariate analysis.**

Univariate analysis is the simplest form of analyzing data.

#let’s start with Distribution plotting.





# OBSERVATIONS:

1. This are the plotting’s of all the features of the dataset.

2. from this we can easily observe each future selection easily, let’s start.

3. Vistara, Air India and indigo are among the most common airlines with most no of operation are there.

4. Morning and late evening flights having most counts.

5. Most no of flights having only 2-3 hours of duration time from new Delhi to top routs of India.

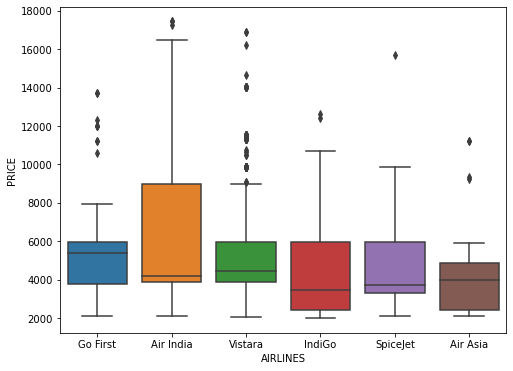
6. 1 stop flights are having more counts and 3 stops flights having the least count.

7. Major price band is 2500 to 1000 almost all the flights in

India are having this price band.

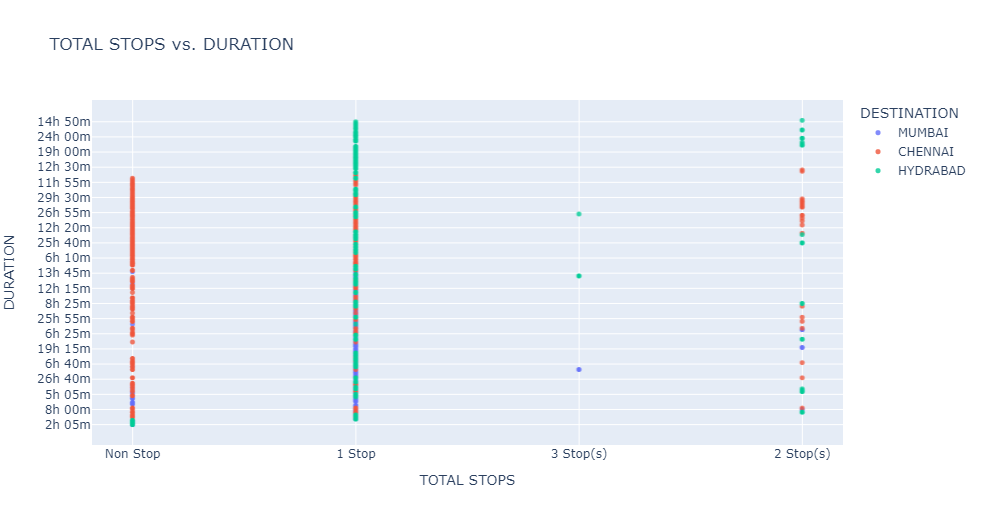
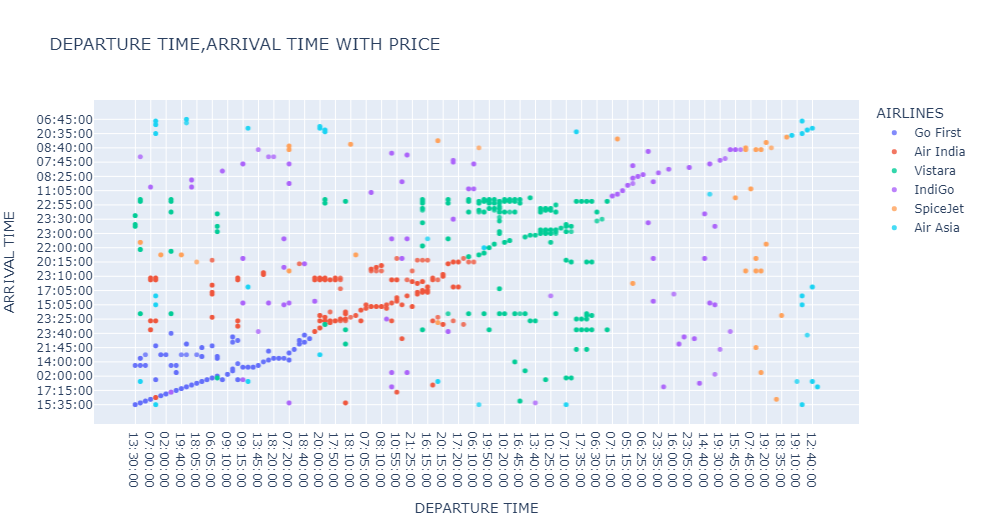
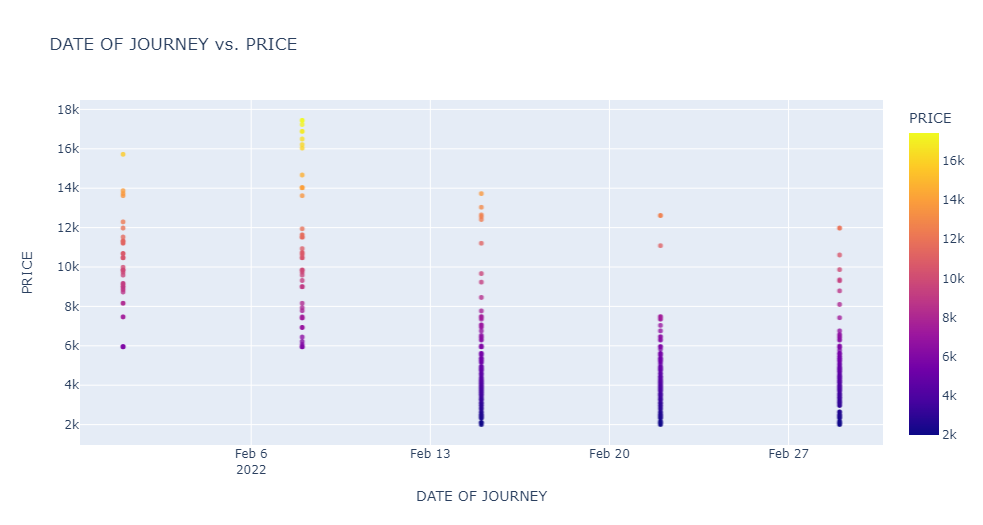
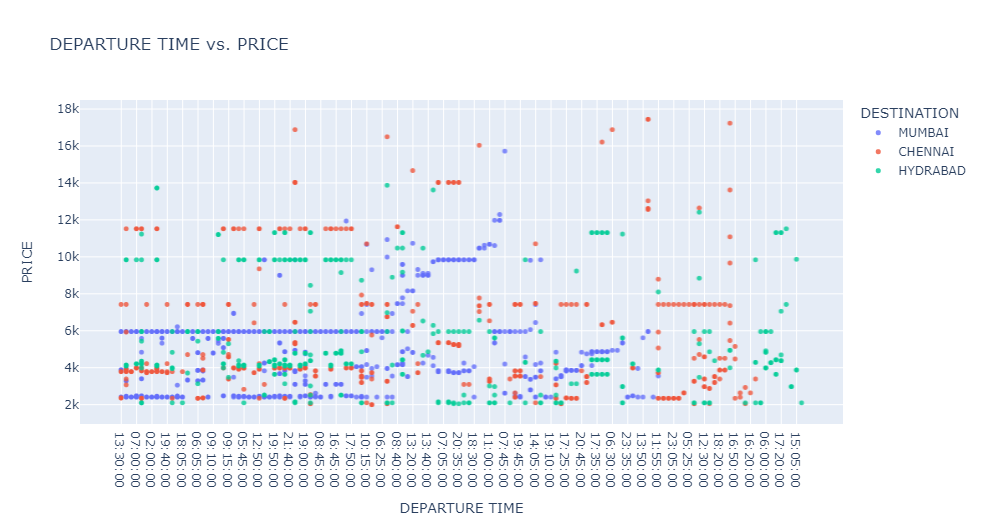
# **Multivariate Analysis**

**Multivariate analysis is a set of techniques used for analysis of data sets that contain more than one variable**



Observations:

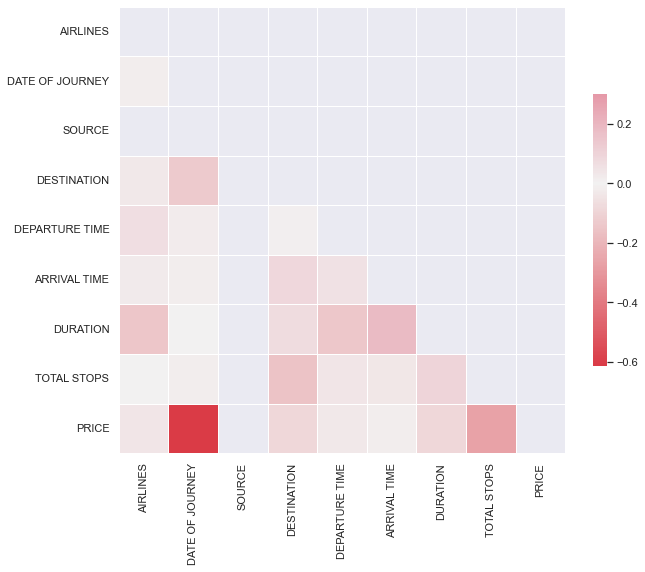
clearly find that maximum no of flights are from VISTARA.

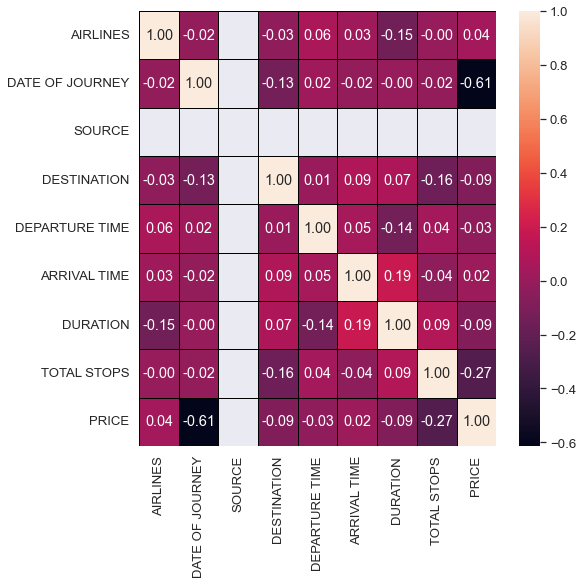


# Observations of data analysis and answers of some common questions.

1. Airfares changes frequently mostly high on 3weeks of departure date.
2. large jumps are present in fares last minutes tickets are high up to 17000 as comparison to book a ticket 5 weeks before the departure date.
3. yes, they go up and down on holidays or in special occasions or events in country.
4. for maximum saving on air ticket consumer can book a nonstop flight 5 weeks earlier for maximum befit from the airlines.
5. yes, prices are high near departure dates.
6. spice jet is cheaper from all the flights present in dataset.
7. As compare to any other time morning flights are very expensive and afternoon flights are cheaper from all.

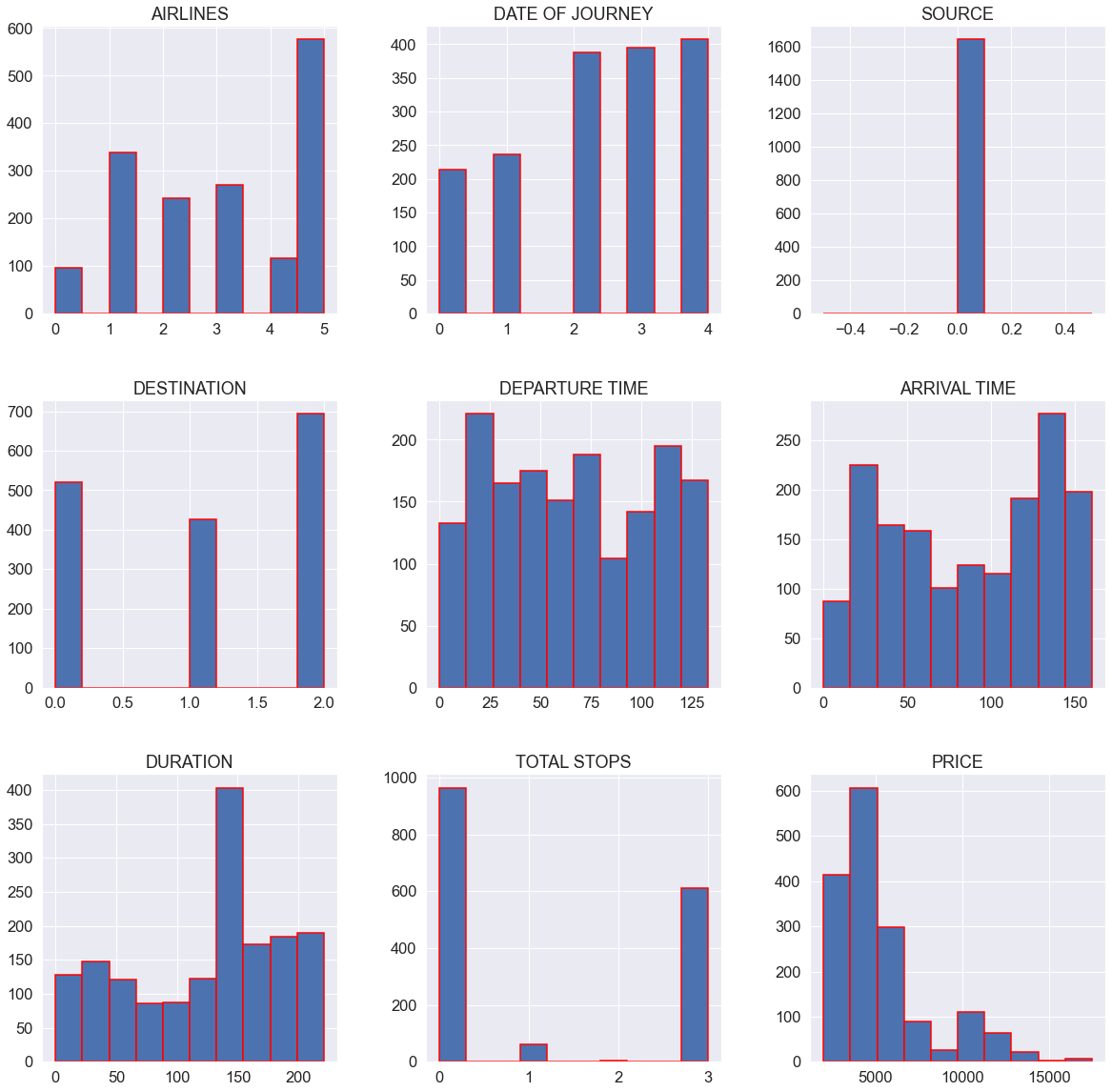
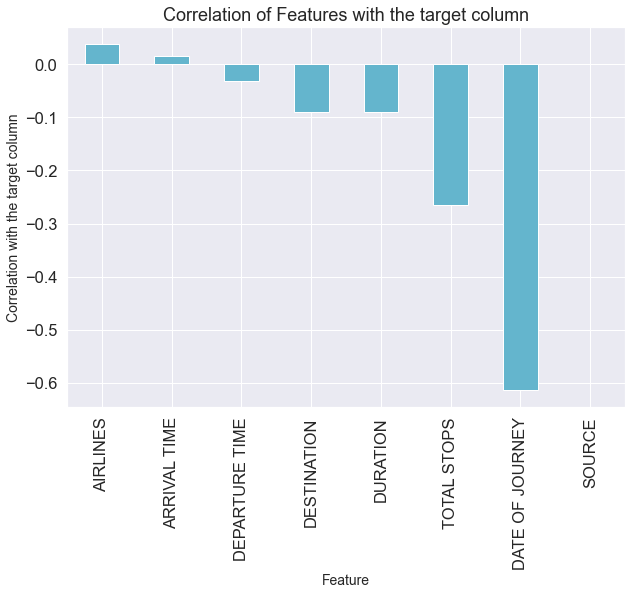
**Correlation of the feature columns.**

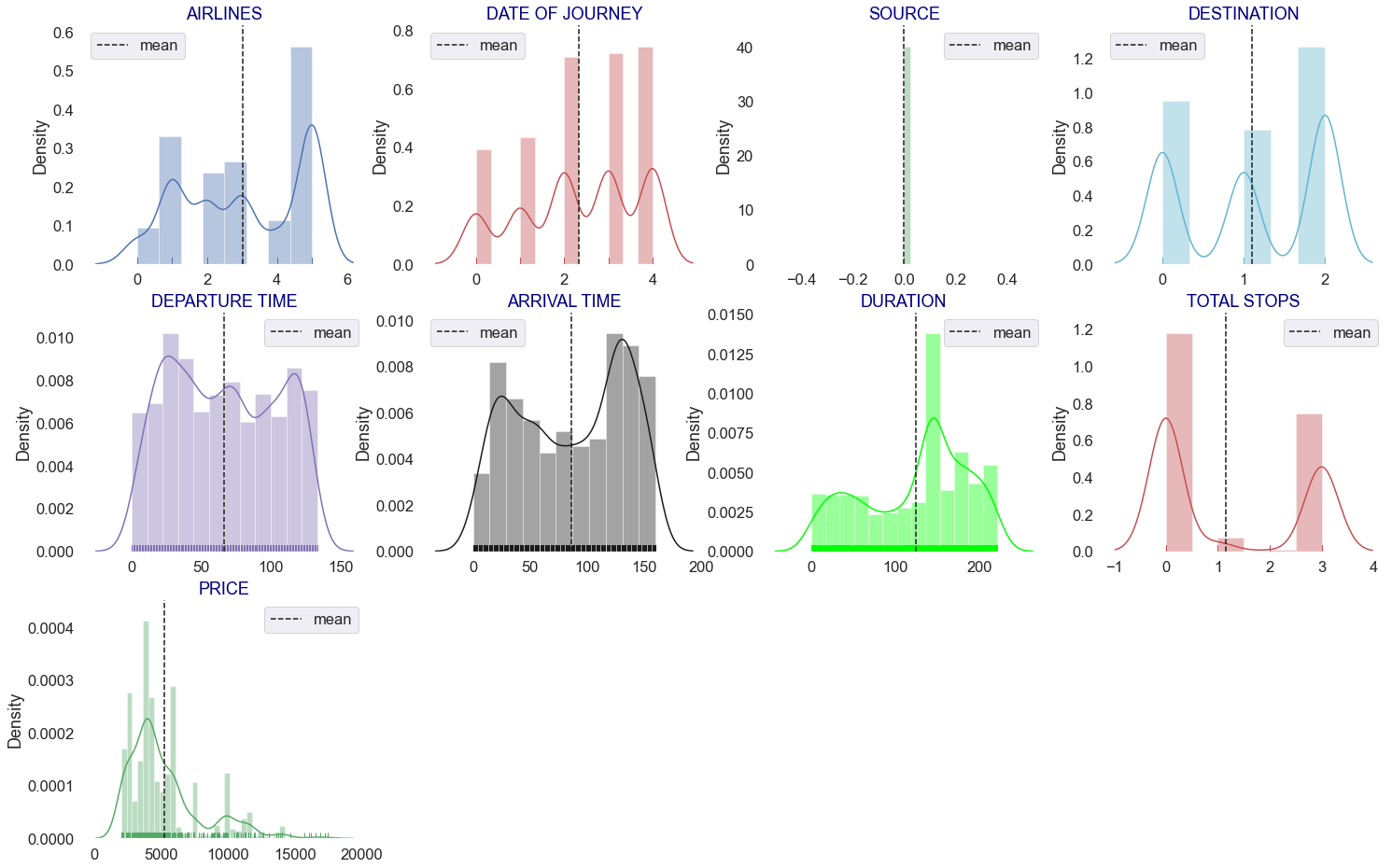




Observations:

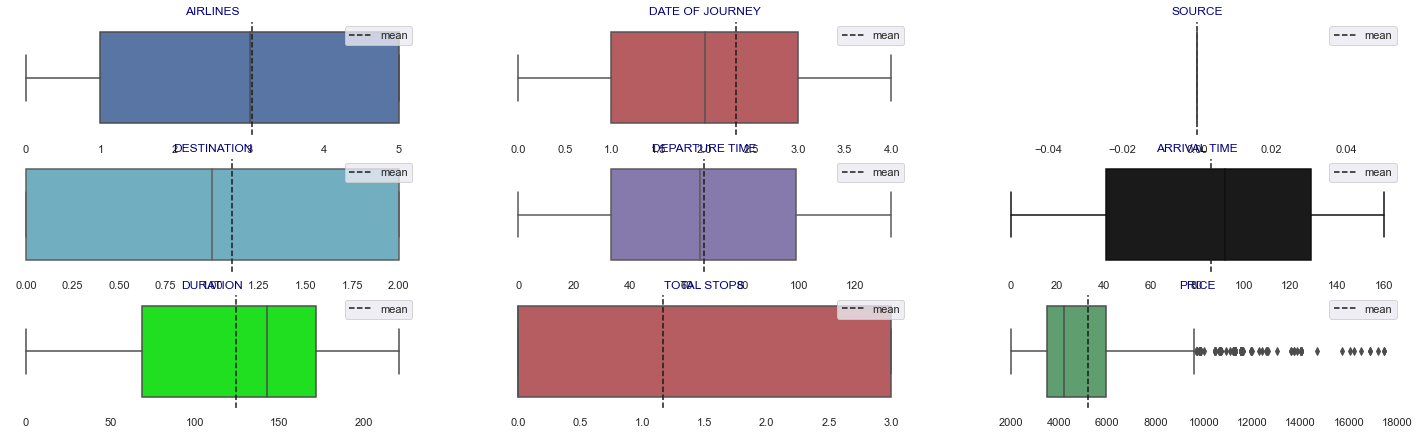
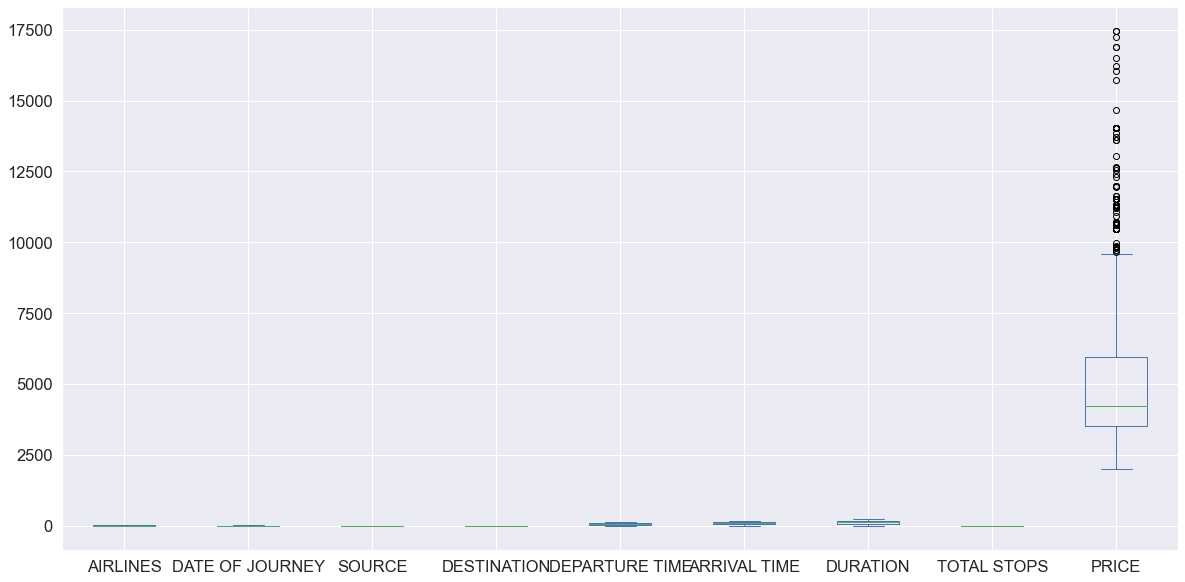
1.Feature columns are not much collinear.





observations:

1. From above plotting of distribution plot, we see that some features columns are not normally distributed.
2. some columns are skewed towards right.
3. Building blocks are out of the normal curve hence outliers are present.



# Outliers are present in data set but in small quantities, and their threshold value is very less, also maximum outliers are present in price column which is our target variable

* Interpretation of the Results

# List of accuracy scores of different regression models.

logistic Regression: - 0.5482599489986677

lasso regression: - 0.548840503584967

ridge regression: - 0.548254166625884

Decision Tree regression: - 0.8070002570035454

Random Forest regression: - 0.909545481225389

gradient boosting: - 0.8711657434274676

support vector: - 0.19788834886242013

K-nearest Neighbors: - 0.7875334974807822

**CONCLUSION**

* Key Findings and Conclusions of the Study

So, our Aim is achieved as we have successfully ticked

all our parameters as mentioned in our Aim Column. It is seen

that Plotly libraries help us to find the outcomes.

The best model is Random forest Regressor. Since the difference between the percentage score of cross validation and r2\_score is optimum.

At cv: - 3

R2 Score: 91.03642190833001

Cross Val Score: 67.81870798851165

Our Model performs with Accuracy 91%... Saving the best models with initials...using pickle library.

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

That's it! We reached the end of our exercise. Throughout this kernel, we put into practice many of the strategies for predicting whether prices are high near departure date which flights are cheaper etc. We philosophized about the variables, we analyzed Price' alone, and with the most correlated variables, we tested some of the fundamental statistical assumptions and we even transformed text data into numeric type using label encoder. That's a lot of work that Python helped us make easier.

* Limitations of this work and Scope for Future Work

Data scraped for this project is only 1500 rows and extract only next 5 weeks of data which is not enough for making big decisions as an organization.

Thank you…