**Blog Submission**

**Project-1**

*This is my first machine learning project. Recently I have completed this project. I am here to describe how i solved the case study in a very detailed manner.*



**Submitted by: Krishna Prasad**

**Acknowledgement**

*Many thanks to Data Trained for providing me this project to understand about the Real Time Field work present in Data Science Industry.*

As this project is about prediction the price of flight ticket by using various featues.

**Contents**

1. [**Introduction**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#c099)
2. **Work Steps**
3. [**EDA**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#030a)

1. **[Building Machine Learning Models](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d" \l "8341)**
2. **[Conclusion](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d" \l "13e2)**

**Flight Price Prediction**

**1)Introduction**

### Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here i will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

### 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.

### Route: The route taken by the flight to reach the destination.

### 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.

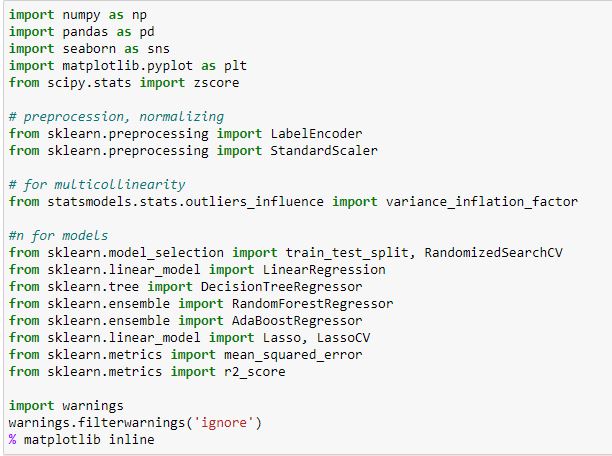
### Total Stops: Total stops between the source and destination.

### Additional Info: Additional information about the flight

### Price: The price of the ticket.[¶](http://localhost:8888/notebooks/Reassignment%20submittion/Flight%20Price%20Prediction.ipynb#Price:-The-price-of-the-ticket.)

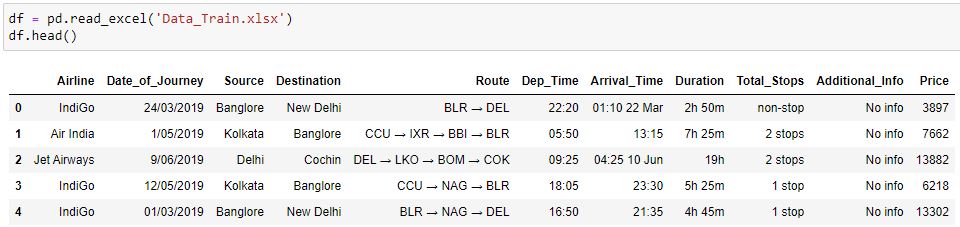
## 2)Work Steps

## Libraries



I have import important libraries for this project in one go so that it will save time.

## Dataset



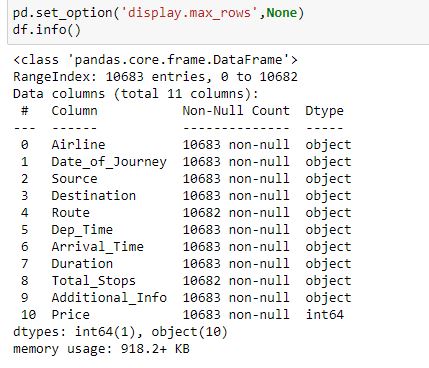
From the dataset i can say that there are few columns in integer and few are in strings. All the columns have vital information so i will not be dropping any columns from the dataset except one column.

## 3)EDA - Exploratory Data Analysis

## Before you start a machine learning project, its important to ensure that the data is ready for modelling work. Exploratory Data Analysis (EDA) ensures the readiness of the data for Machine Learning. In fact, EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. So in this dataset I have used various EDA Techniques to analyse a data in a proper way.

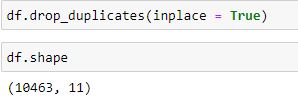


There are 10683 rows and 11 columns in the dataset.



It is a mixed dataset as 10 columns are object type and 1 columns are integers type.

There are 1 missing values in 2 columns which i will fix soon.



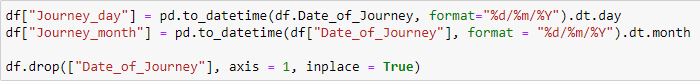
There are few duplicates value in the dataset.

Now there are 10463 rows and 11 columns in the dataset.



There are few columns which are categorical in nature and few columns are continuous in nature.

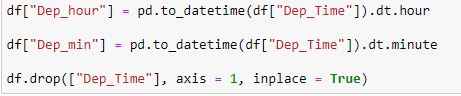
### Date\_of\_Journey



I have converted Date\_of\_Journey column into integers.

Since i have converted Date\_of\_Journey column into integers, Now we can drop as it is of no use.

### Departure time



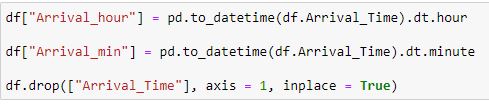
Departure time is when a plane leaves the gate.

Similar to Date\_of\_Journey we can extract values from Dep\_Time.

I have extracted departure hour and departure minute from Departure time column.

Now i can drop Dep\_Time as it is of no use.

### Arrival Time



Arrival time is when the plane pulls up to the gate.

Similar to Date\_of\_Journey we can extract values from arrival Time.

I have extracted Arrival hour and Arrival minute from Departure time column.

Now i can drop arrival Time as it is of no use.

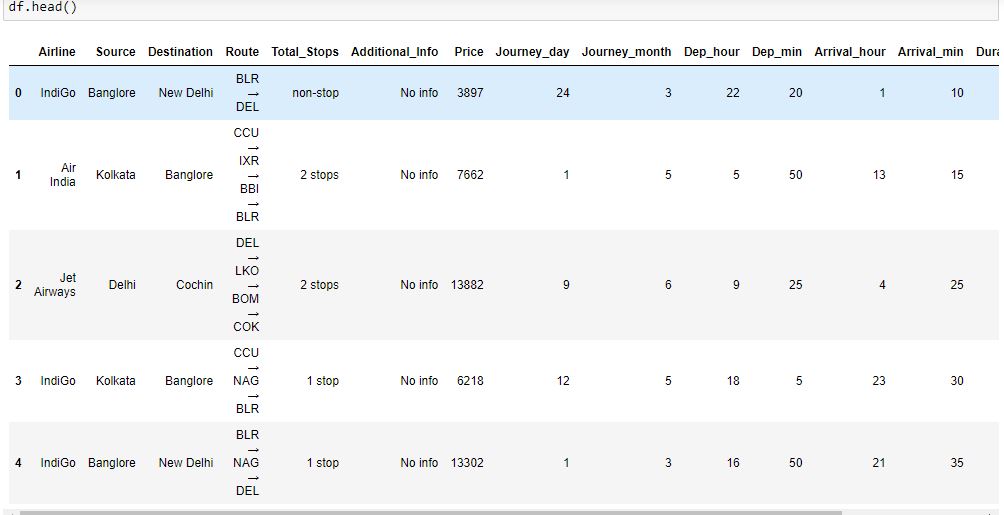
## Duration



I have extract hours from duration.

I have extract minutes from duration.

I will be dropping duration columns now.

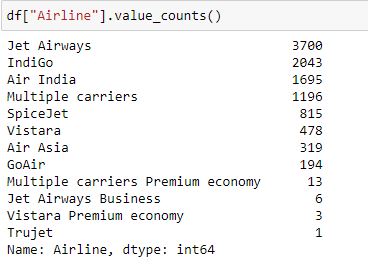


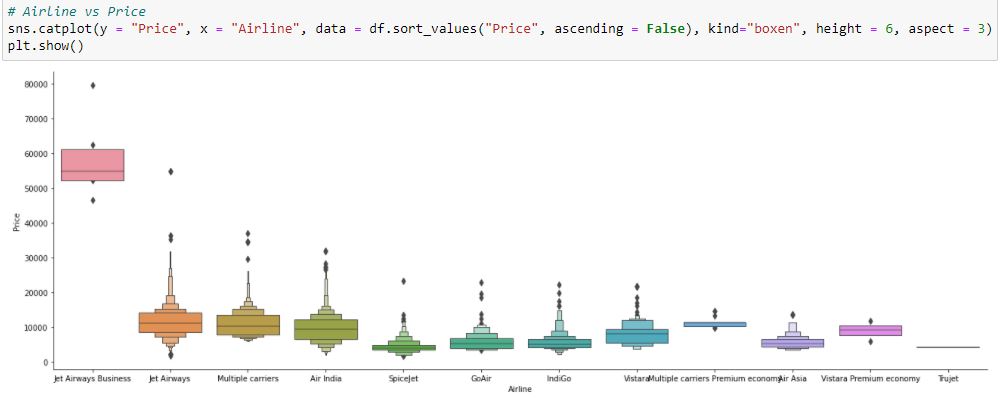
Now the number of columns increased as I have divided the 3 columns into 6 columns.

Now i have columns in the dataset.

## Cat Plot

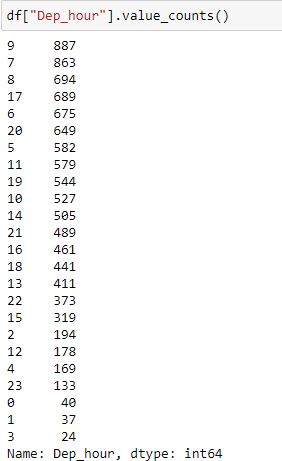
**Airline**

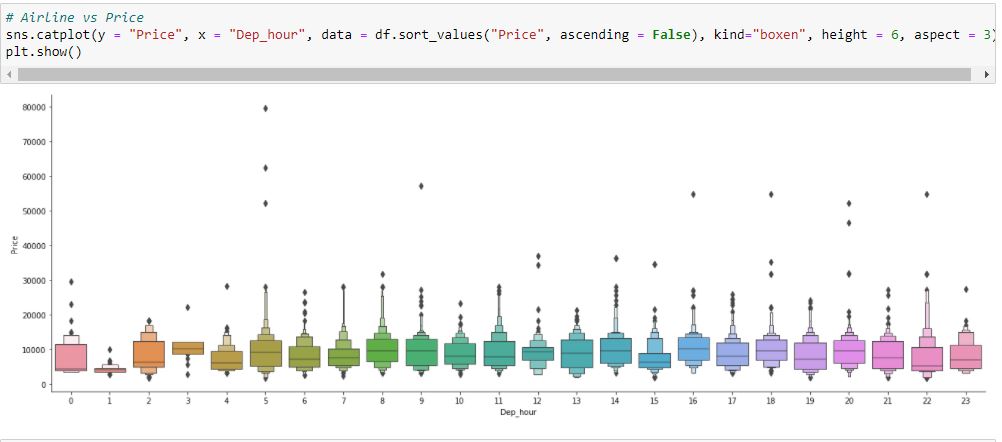




Jet Airways Business flight price is high as compared to other flight service.

### **Dep\_hour**

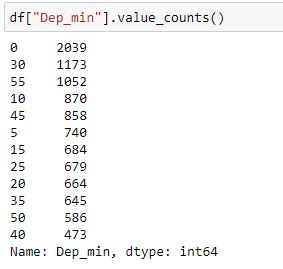


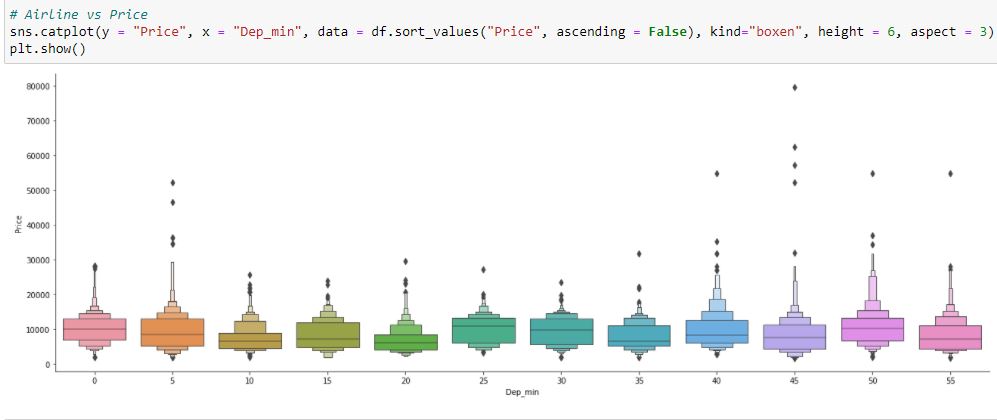


Departure hour doesn't affect flight price.

Irrespective of journey hours price is same for all the flight.

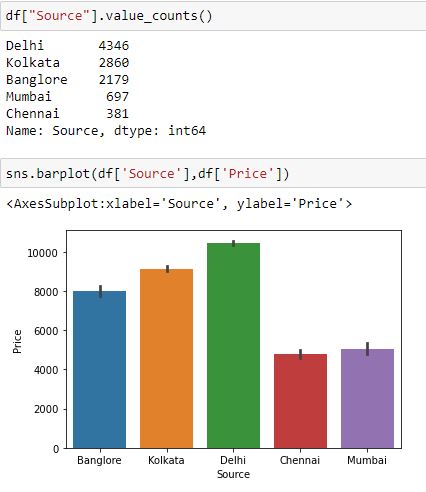
### **Dep\_min**





Departure min doesn't affect flight price. So I can say that price is same irrespective of any time.

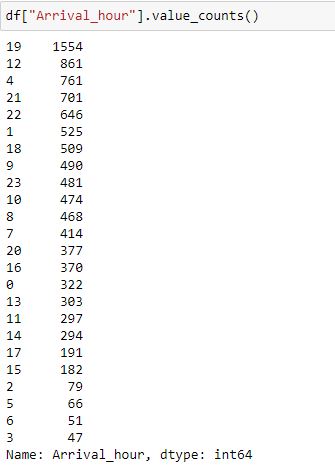
**Source**

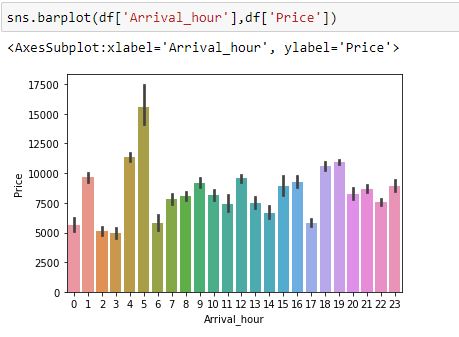


From delhi the flight price is high as compared to other cities.

From plot I can say that Chennai and Mumbai flight price is almost same.

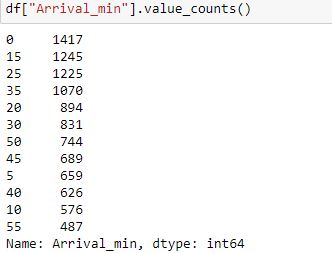
### Arrival\_hour

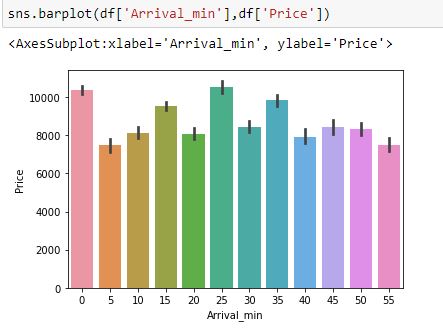




Those flight have 5 hours journey price is high as compared to other flight hours.

### Arrival\_min

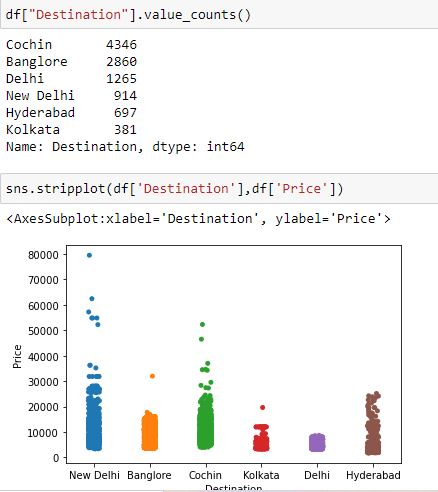




Those flight have 25 mins journey there price is high.

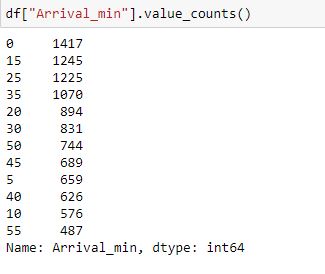
## Strip Plot

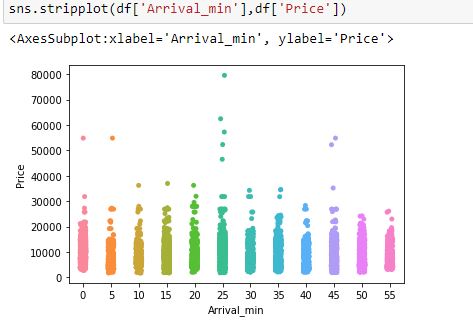
## Destination



Those flight have New Delhi as a destination journey there price is high.

### **Arrival\_min**





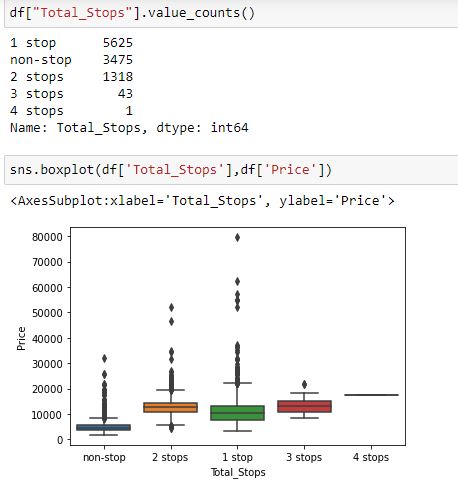
Those flight have 25 mins journey there price is high.

## Dropping columns



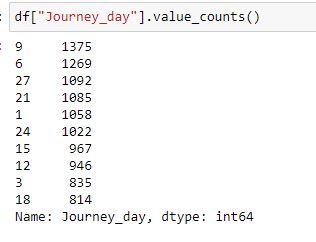
I have drop Route and Additional info columns as it is of no use.

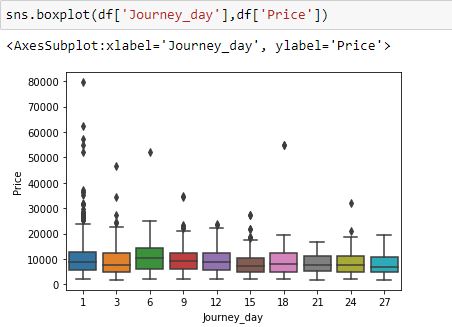
#### **Total\_Stops**



Those flight have 1 stop journey there price is high.

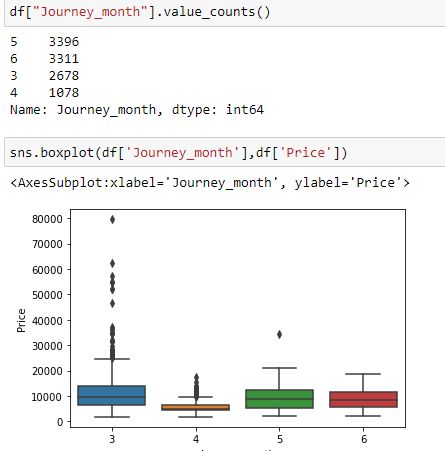
**Journey\_day**





Those flight have 1 day journey there price is high.

### **Journey\_month**



Those people who are travelling for around 3 months are paying high price for flight.

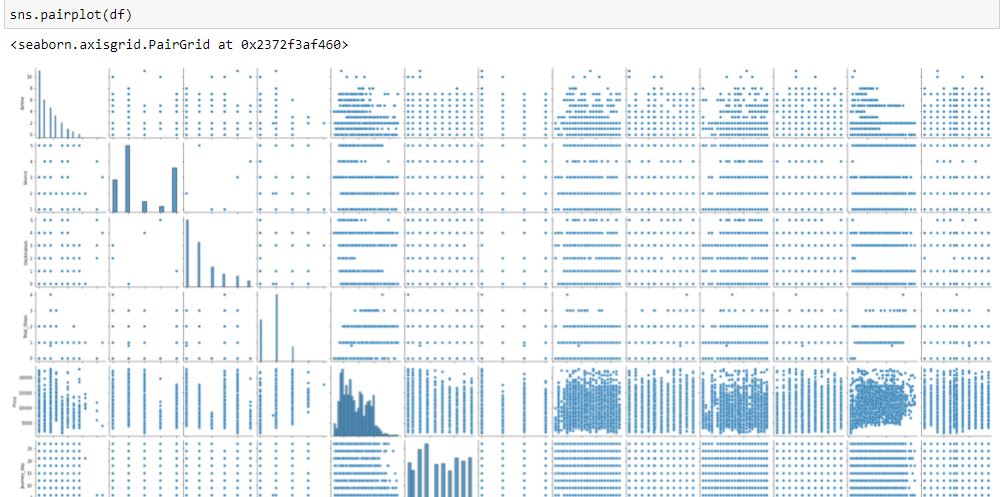
## Label Encoder



I have used label encoder to convert the strings values into integers.

It will help me in model building.

**Pair Plot**



The above pair plot gives us the clear idea that target variable 'Price' are positively corelated with few feature. Only few columns are not correlated.

## Filling the missing values

## 

## There is missing value in Total\_Stops columns which i fill it with mean.

## 

## Total number of counts in each columns is matching as there is no missing values.

## The difference between the mean and 50% is not much.

## There are outliers in the dataset which i will remove it soon.

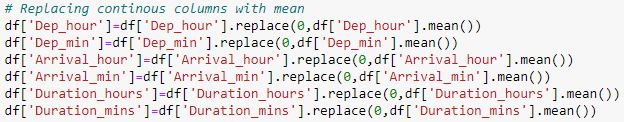
## Distribution Plot

## 

There are skewness in the dataset which i will remove soon.

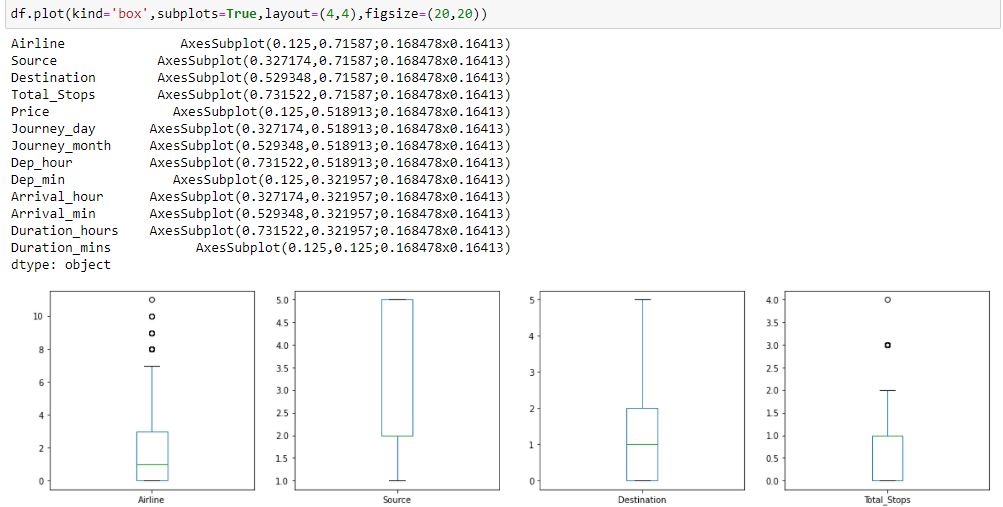
There are few zero values as well that also need to be taken care.

## Replacing all the zero values from different columns



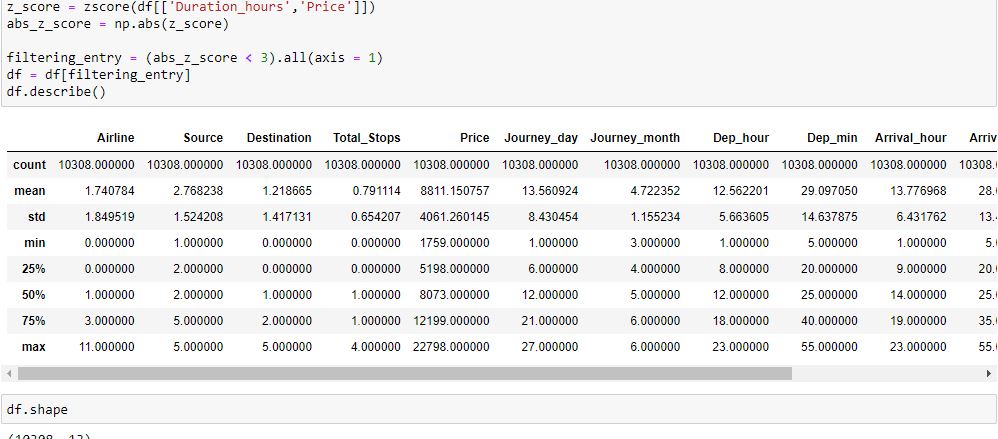
There are few zero values that got replaced with the help of mean.

## Box Plot



There are outliers in the dataset which i will fix it with the help of Z score.

## Z score



I have used the Z Score to remove the outliers from the dataset.

Now there are 10308 rows and 13 columns in the dataset.

## Heatmap



Total stops and price is positively correlated with each other.

Duration hour and price is positively correlated with each other.

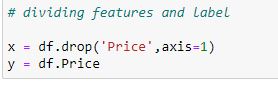
Duration and Total stops is positively correlated with each other.

Total stops and Destination are negatively correlated to each other.

Destination and price services are negatively correlated to each other.

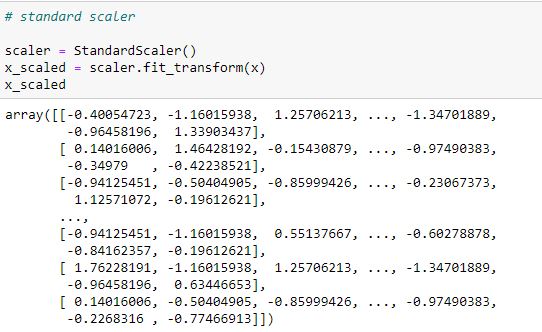
## 4)Machine Learning

*Here’s the role of the label that comes into existence. A label is given to such features to distinguish them from other features. Thus, I have obtain labels as output when provided with features as input.*



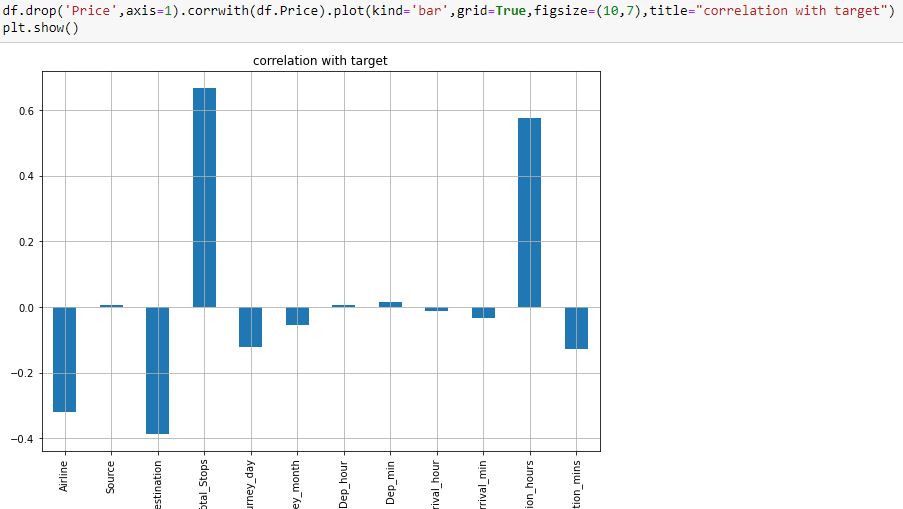
I have divided dataset into feature and label.

## Standard Scaler



Standard scaler is basically scaling the date in one range so that it will be easy for Model building.

## Visualize the correlation



The above plot gives me an clear idea that few columns are positively correlated and few are negatively correlated with label. However i will use all the columns for model prediction.

## VIF - variance inflation factor



VIF is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.

Multicollinearity is a phenomenon when two or more independent variables are highly intercorrelated.

From the above stats i can say that none of the features are highly intercorrelated it means Multicollinearity doesn't exist.

## Model Building

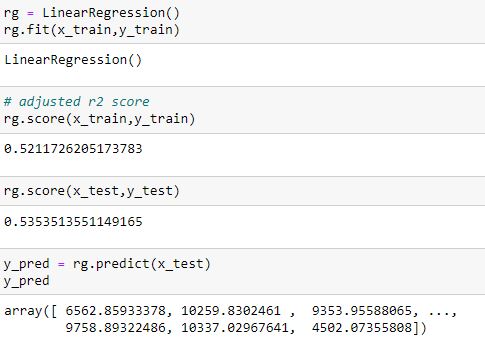


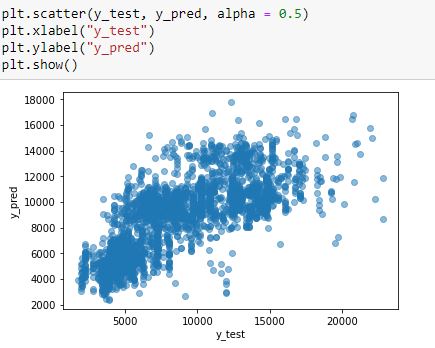
For model prediction i am dividing the dataset into 2 parts.

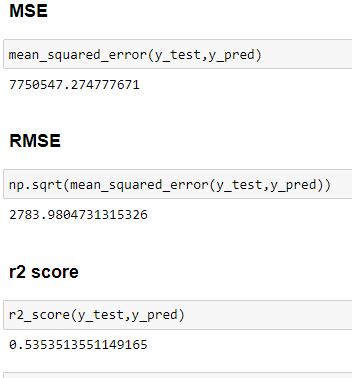
One part is used for training purpose i.e 75% dataset.

Other part is used for testing purpose i.e 25% dataset.

## Linear Regression model

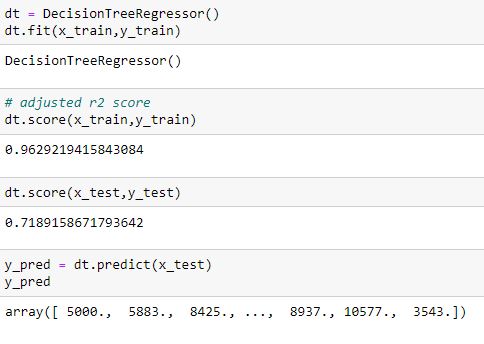


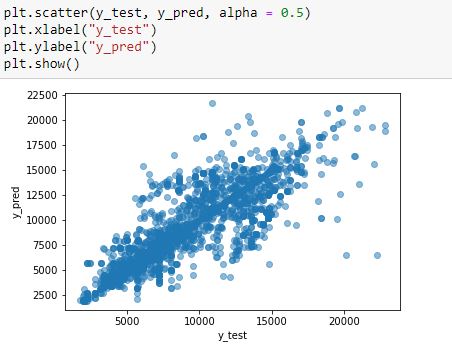


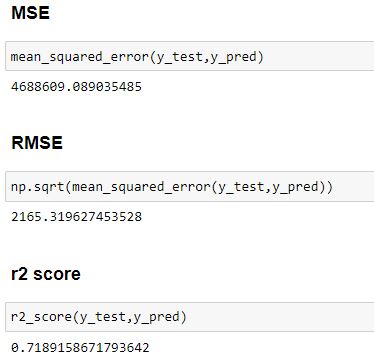


Linear Regression accuracy score 53%

## Decision Tree Regressor Model

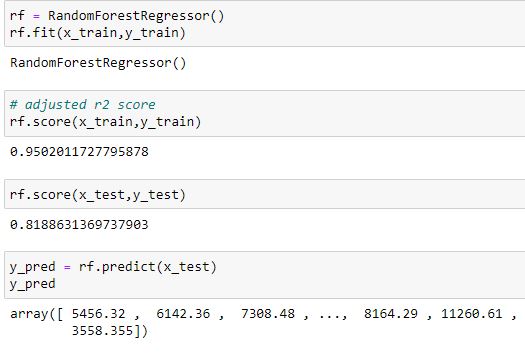


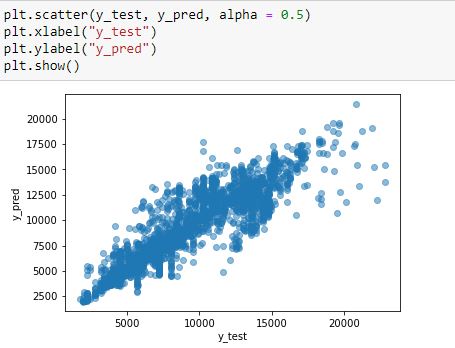


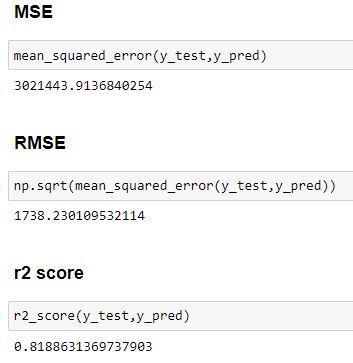


Decision Tree Regression accuracy score 72%

### **Random Forest Regressor Model**

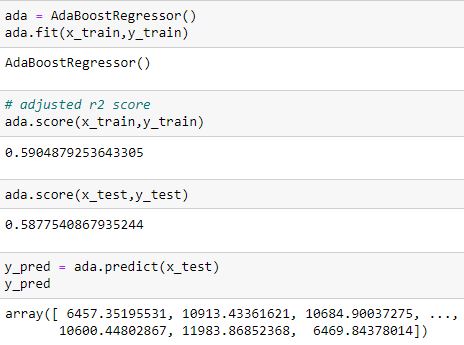


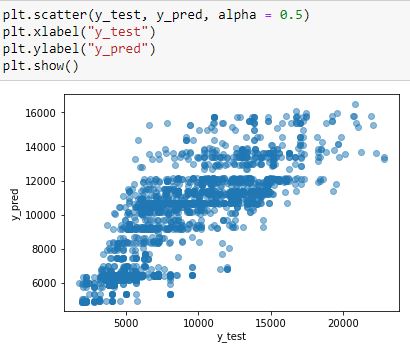


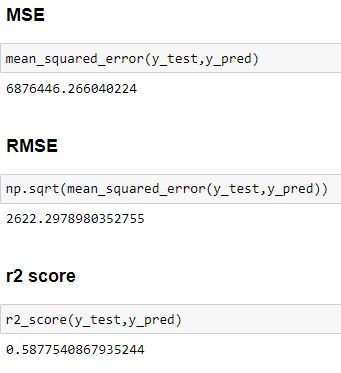


Random Forest Regression accuracy score 82%

## AdaBoost Regressor Model

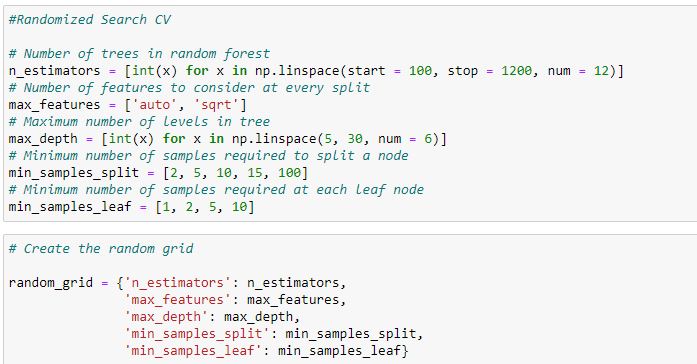


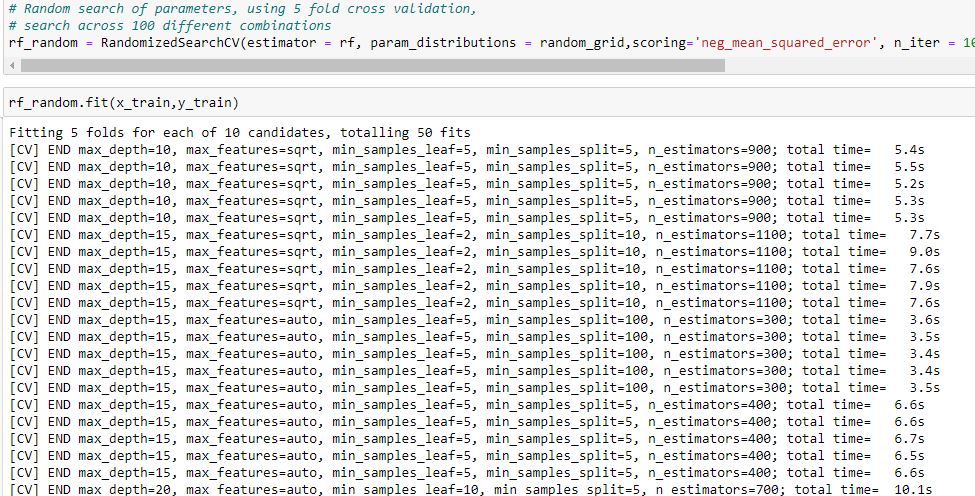
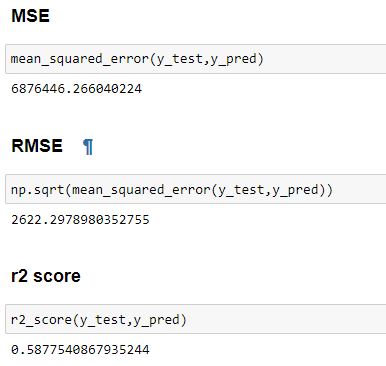




AdaBoost Regression accuracy score 58%

### **Hyperparameter Tuning in Random Forest Regressor Model**

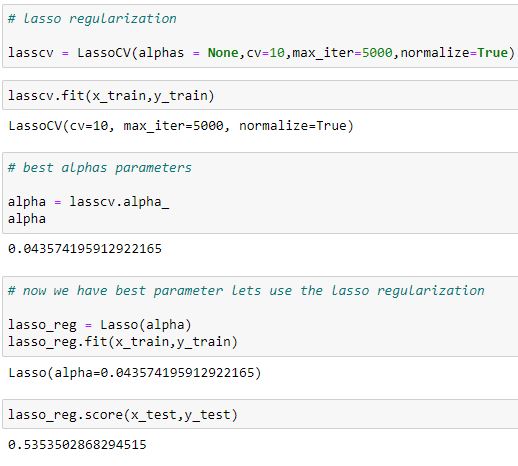


#### Using RandomizedSearchCV method for improving the r2 score.

I have tried to improve the accuracy score by using hyper parameter tuning in random forest algorithm . Hyper parameter is reducing the accuracy score.

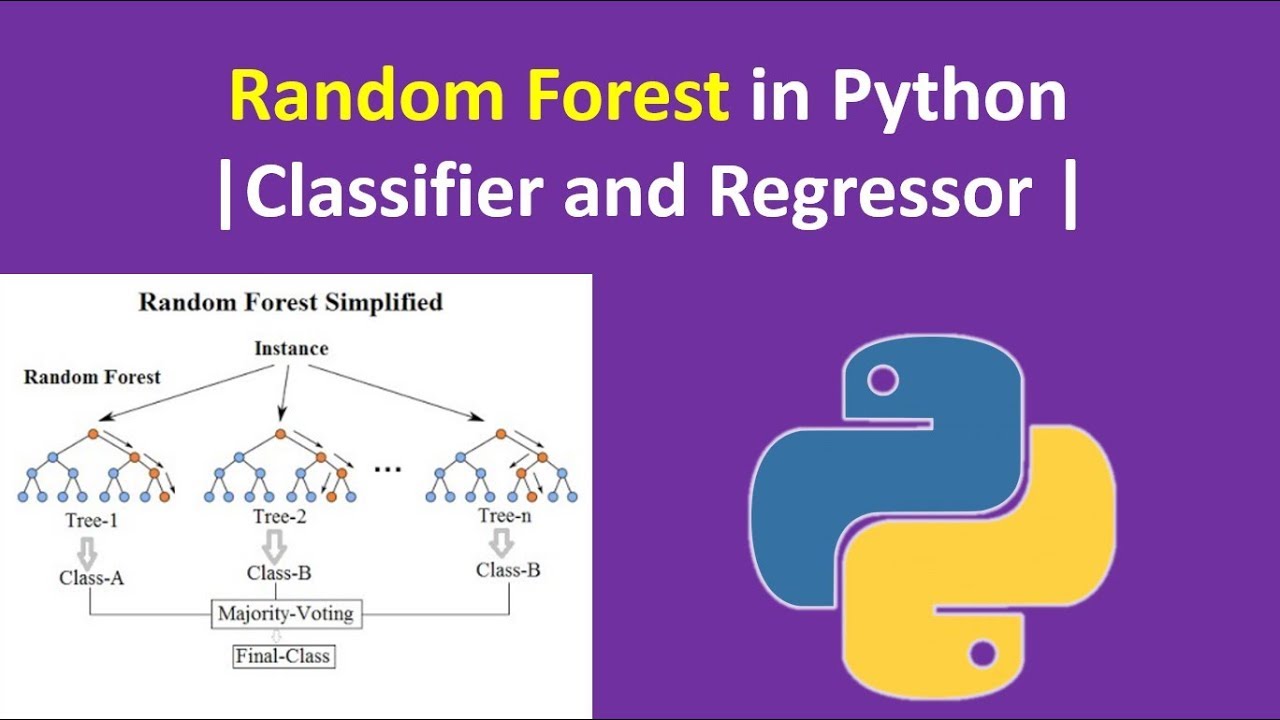
## Regularization



### **Saving the Best Model**



The best model is Random Forest classifier whose accuracy score is 82%.



**5.Conclusion**

I have learned a lot while working on this project.

The price depends on various factor like place , journey time and journey location.

While doing EDA I came across various challenge like how flight price is different for every airline.

The expensive flight is offered by business jet airways.

I have used z score to clean the data and model building was also interesting.

I have used 4 algorithms and the best one is random forest regressor.

**Links References**

https://github.com/krishnaprasad19960726/new-project-resubmission/blob/main/Flight%20Price%20Prediction.ipynb

**Blog Submission**

**Project – 2**

*This is my Second machine learning project. Recently I have completed this project. I am here to describe how i solved the case study in a very detailed manner.*



**Submitted by: Krishna Prasad**

**Acknowledgement**

# *Many thanks to Data Trained for Providing me this project to understand about the Real Time Field work present in Data Science Industry*

*As this project of predicting the price of avocados I gain enough knowledge about the demands and price of avocados.*

# **Avocado Project**

## Introduction

### Avocado is a fruit consumed by people heavily in the United States . Avocados are a good source of B vitamins, which help you fight off disease and infection. They also give you vitamins C and E, plus natural plant chemicals that may help prevent cancer. Avocados are low in sugar. And they contain fiber , which helps you feel full longer.

### Some relevant columns in the dataset:

### Date - The date of the observation

### Average Price - the average price of a single avocado

### type - conventional or organic

### year - the year

### Region - the city or region of the observation

### Total Volume - Total number of avocados sold

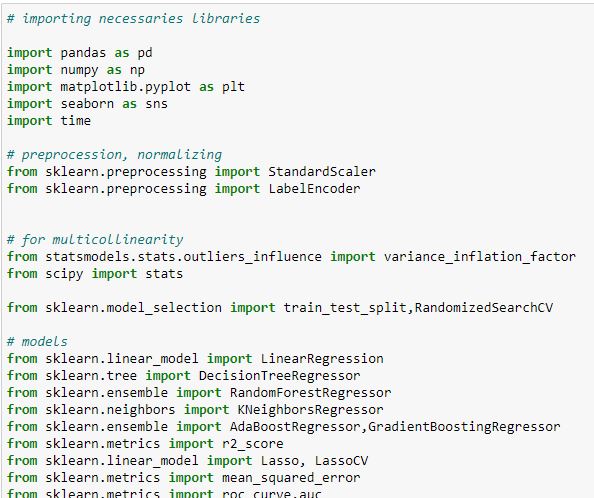
### 4046 - Total number of avocados with PLU 4046 sold

### 4225 - Total number of avocados with PLU 4225 sold

### 4770 - Total number of avocados with PLU 4770 sold

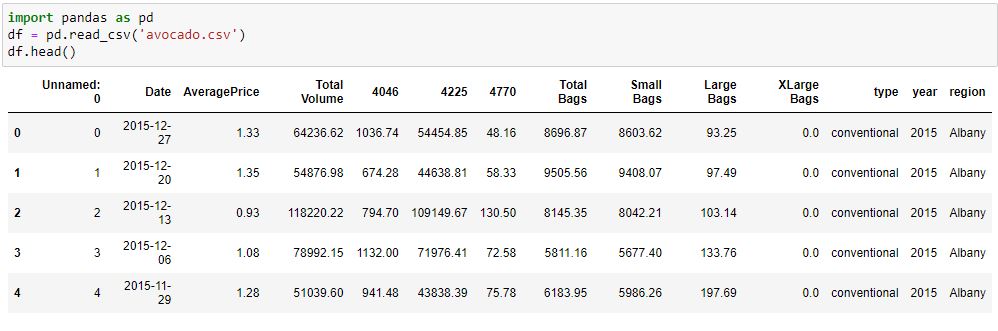
### Aim for this project is just to predict the price of avocado.

## Libraries



I have import important libraries for this project in one go so that it will save time.

## Dataset

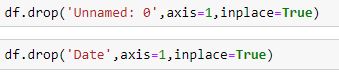


From the dataset i can say that there are few columns in integer and few are in strings.

Surely few columns need to drop as it is of no use. I will do it soon.

## EDA - Exploratory Data Analysis

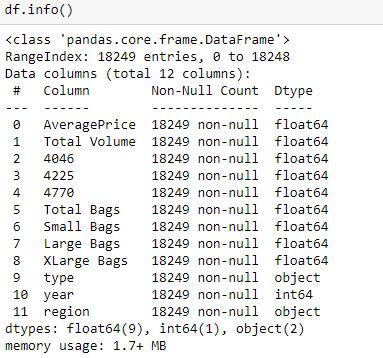
**Dropping the columns**



I am dropping Unnamed:0 and date columns as it is of no use.

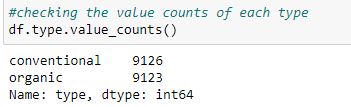


There are 18249 rows and 12 columns in the dataset.



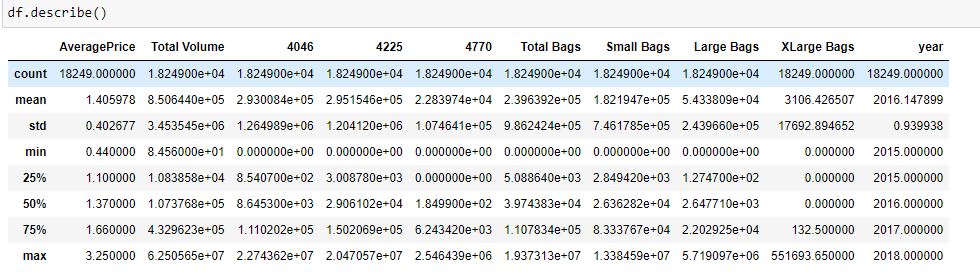
It is a mixed dataset as 9 columns are float type, 2 columns are integers type and 1 columns are object type.

There is no missing values in any columns as count is 18249 for every columns.



There are 2 types avocados in this dataset one is conventional and another is organic.

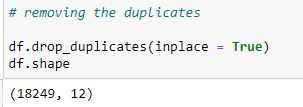
Conventional avocados are 9126 and organic avocados are 9123 in numbers.



Total number of counts in each columns is matching as there is no missing values.

# The difference between the mean and 50% is not much.

# There are outliers in the dataset which i will remove it soon.

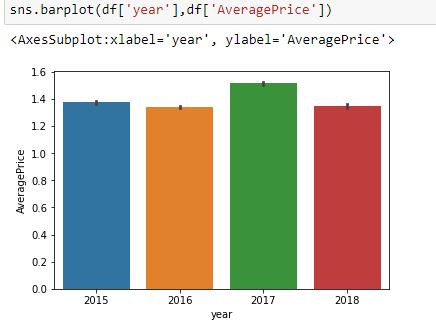


There are no duplicates in the dataset.

**Pair Plot**

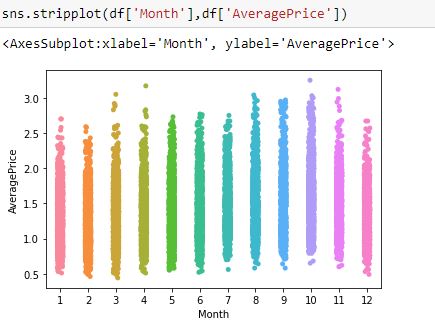


## Bar Plot



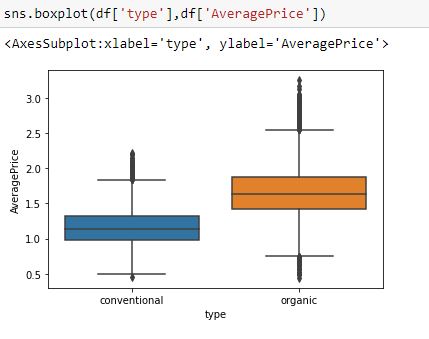
The Average Price of avocado is highest in 2017.

## Strip Plot



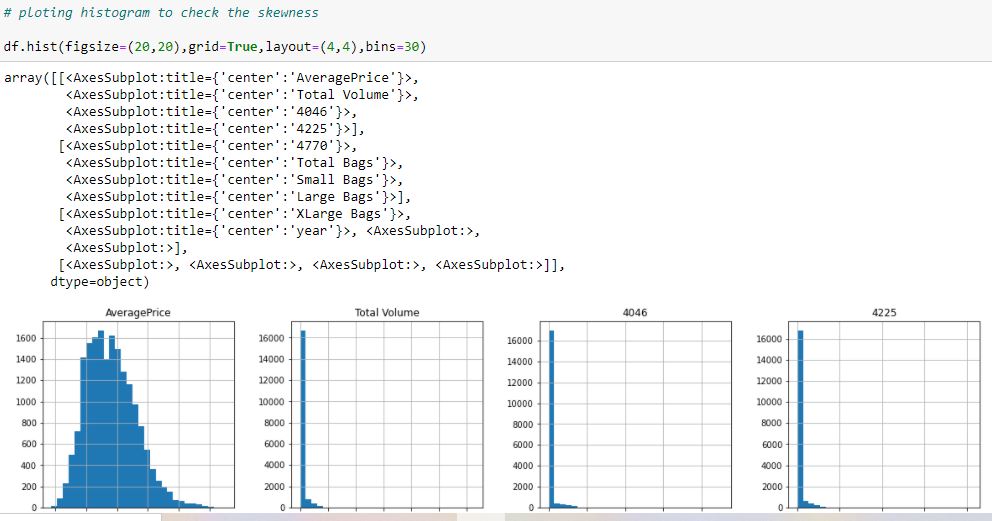
The Average Price of avocado is highest in 10th month.

## Boxplot



he average price of organic avocado is higher than conventional avocado.

## Histogram Plot



There are skewness in the dataset which i will remove soon.

There are few zero values as well that also need to be taken care.

## Removing Skewness



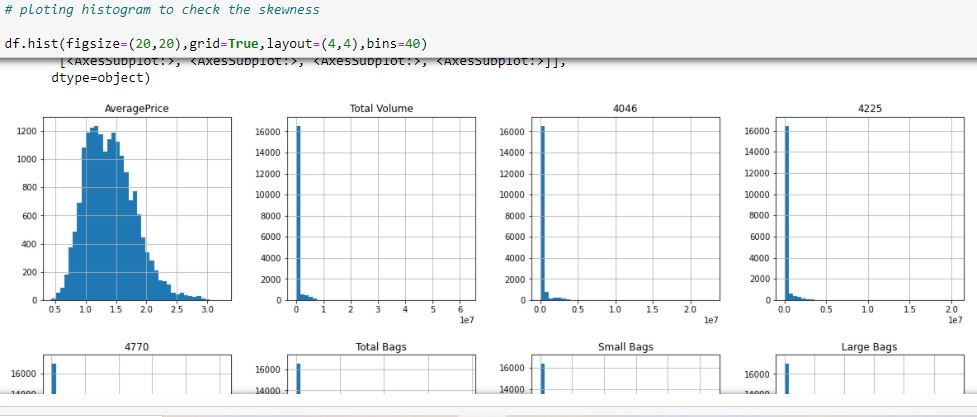
I have tried to remove the skewness from the above columns.

## Replacing all the zero values in different columns

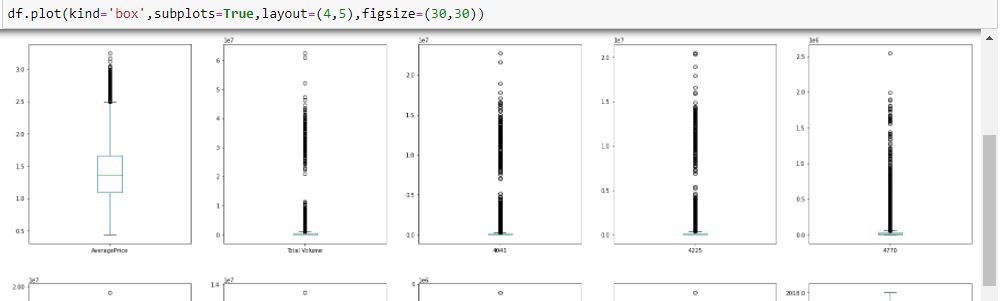


I have removed few zero values from different columns with the help of mean and mode.

## Histogram Plot

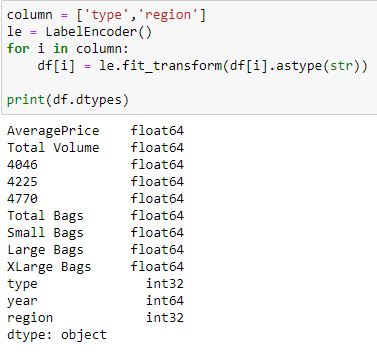


Now from the histogram plot i can say that skewness is removed.



There are outliers in few dataset which i will remove it from z score.

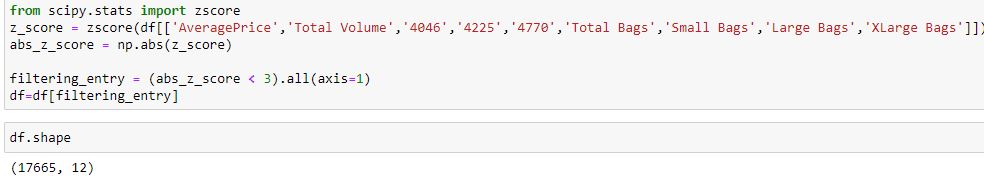
**Making the categorical data into numerical for better EDA**



There are 2 columns type and region both are in strings.

I have converted both of them into integers so that model prediction will be easy.

## Z Score



After removing outliers there are 17665 rows and 12 columns in the dataset.

## Heatmap



XLarge Bags and Large Bags is positively correlated with each other.

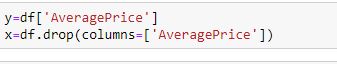
Large Bags and Total Bags is positively correlated with each other.

small bags and 4046 is positively correlated with each other.

small bags and average price are negatively correlated to each other.

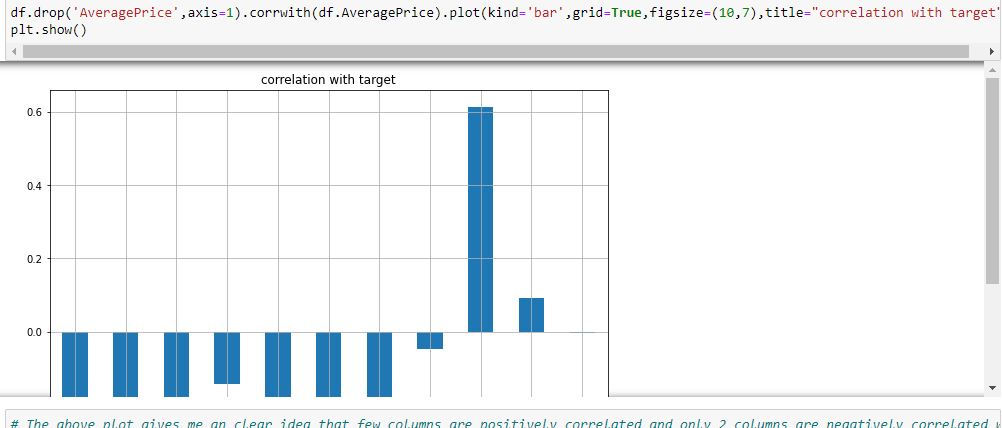
Total volumes and average price are negatively correlated to each other.

## Machine Learning



I have divided dataset into feature and label.

## Visualize the correlation



The above plot gives me an clear idea that few columns are positively correlated and few are negatively correlated with label.

# However i will use all the columns for model prediction.

## Standard Scaler

## 

## Standard scaler is basically scaling the date in one range so that it will be easy for Model building.

## VIF - variance inflation factor

## 

## VIF is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.

## Multicollinearity is a phenomenon when two or more independent variables are highly intercorrelated.

## From the above stats i can say that few features are highly intercorrelated it means Multicollinearity does exist.

## Model Building

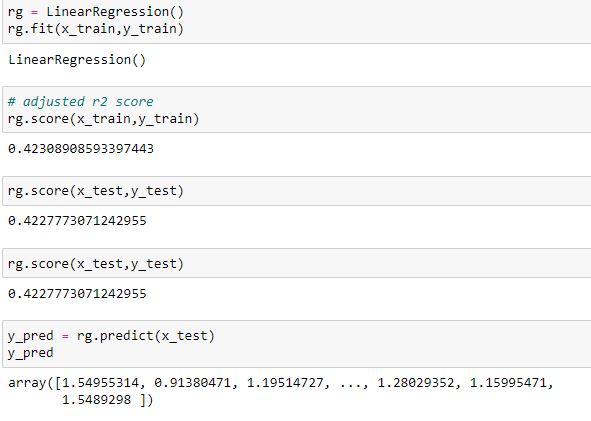


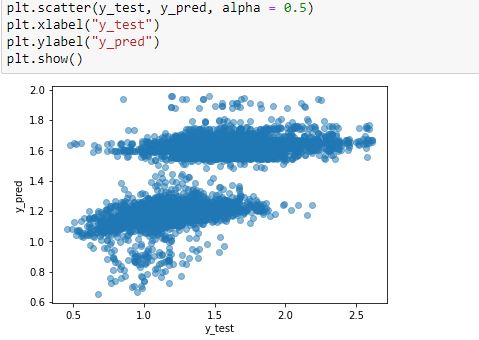
For model prediction i am dividing the dataset into 2 parts.

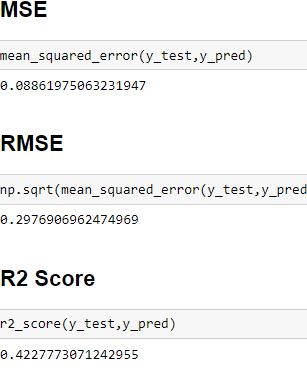
One part is used for training purpose i.e 75% dataset.

Other part is used for testing purpose i.e 25% dataset.

## Linear Regression model

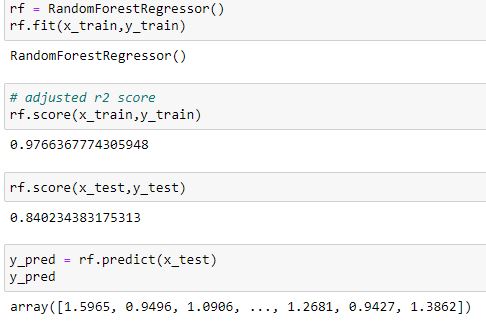


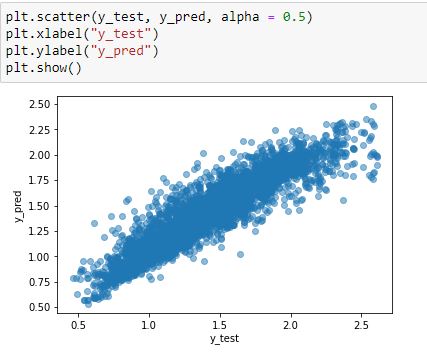


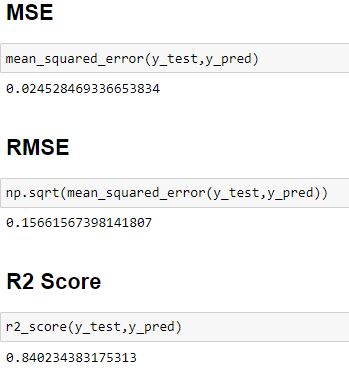


Linear Regression accuracy score 42%

## Random Forest Regressor Model

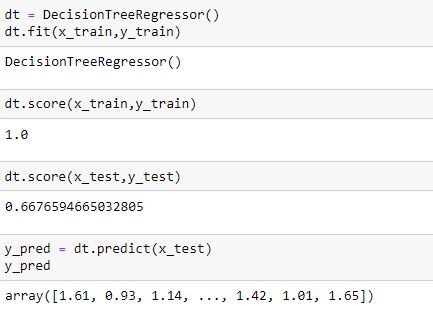


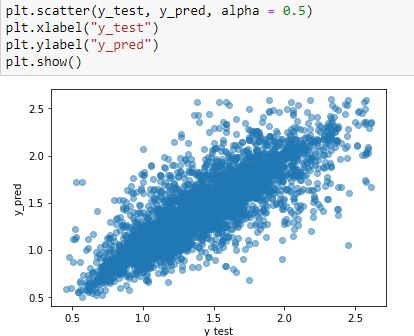


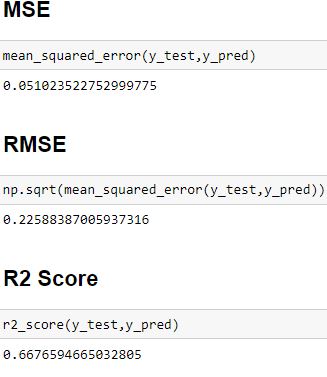


Random Forest Regressor accuracy score 84%

## Decision Tree Regressor Model

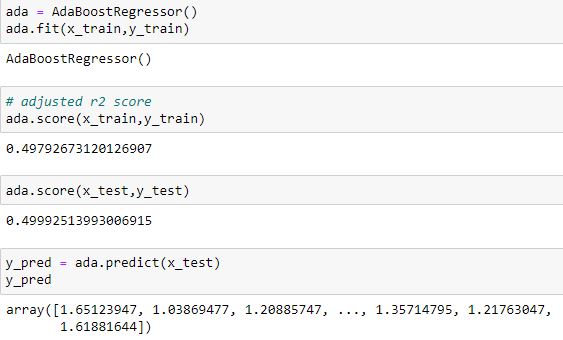


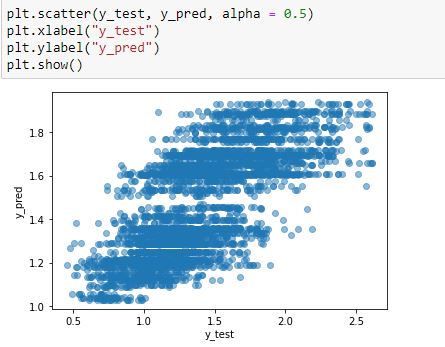




Decision Tree Regressor accuracy score 67%

## AdaBoost Regressor



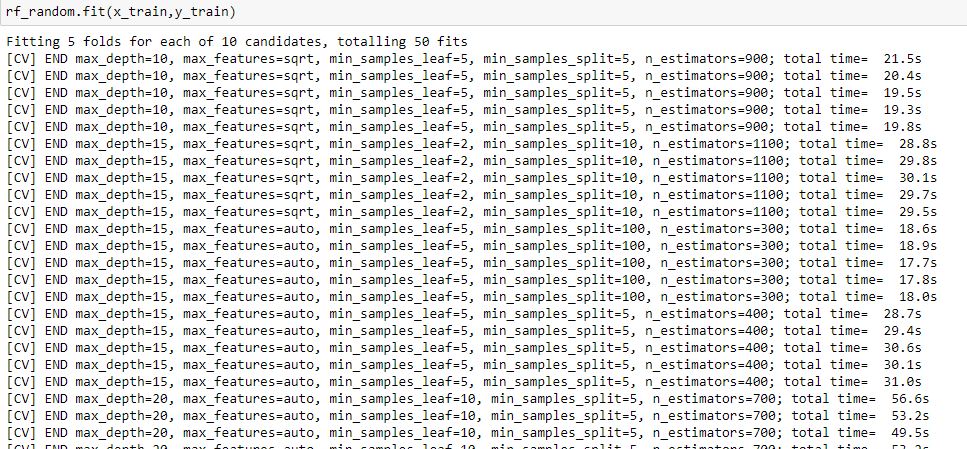


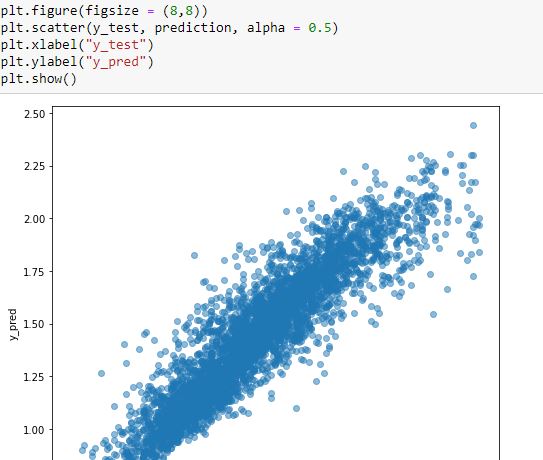


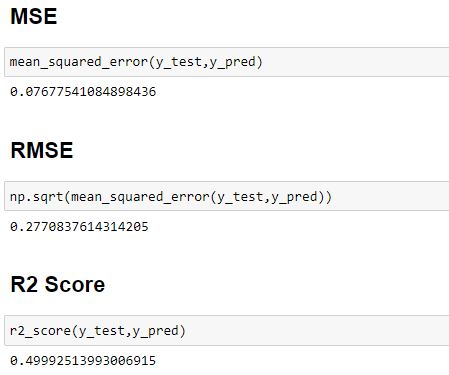
AdaBoost Regressor accuracy score 50%

## Hyperparameter Tuning in Random Forest Regressor Model









I have tried to improve the accuracy score for random forest regressor model but it is reducing the accuracy score.

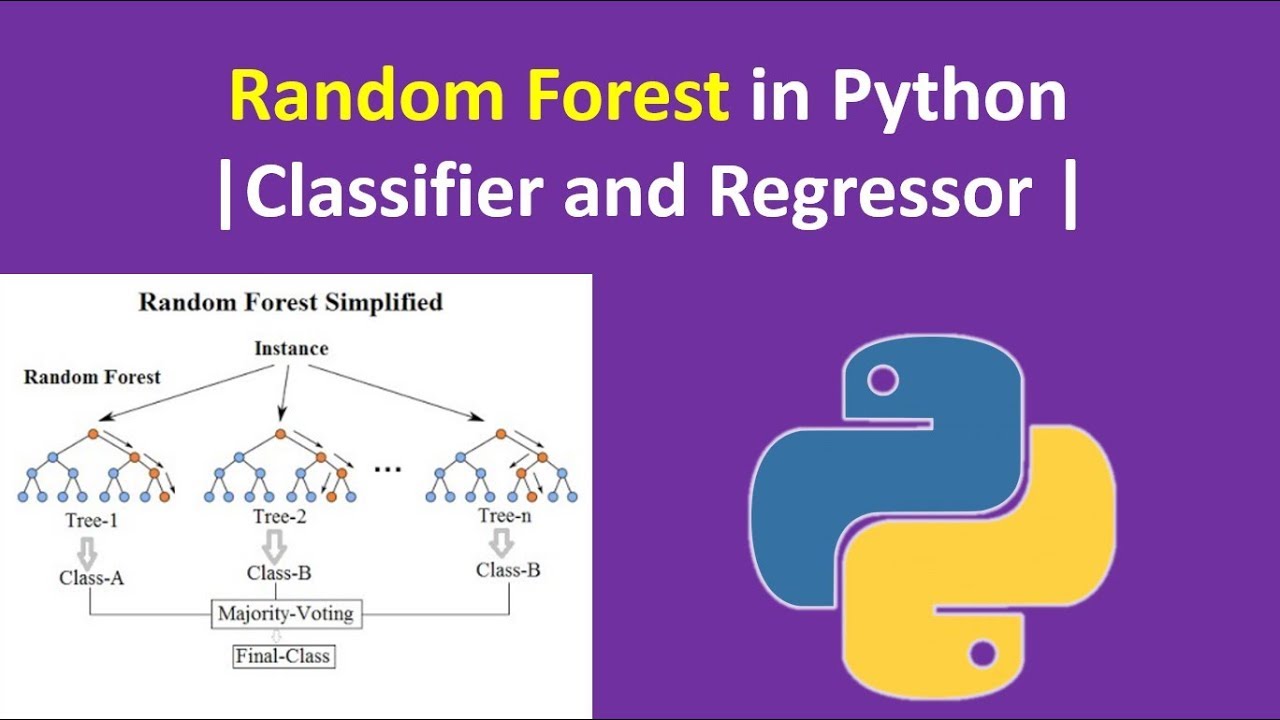
## Regularization



## Saving the Best Model



The best model is Random Forest Regressor whose accuracy score is 84%



**Conclusion**

This project gives me an clear idea what is the demand of avocados.

The price depends on various factor like place and season.

While doing EDA I came across various challenge like how avocados price is different for region

I have used z score to clean the data and model building was also interesting.

I have used 4 algorithms and the best one is random forest regressor.

**Links References**

<https://github.com/krishnaprasad19960726/new-project-resubmission/blob/main/Avocado%20Fruit%20Analysis.ipynb>