Firstly, I am going to implement the Machine Learning on Avocado Data Set.



Some brief explanation about avocado before entering into the project code and description:

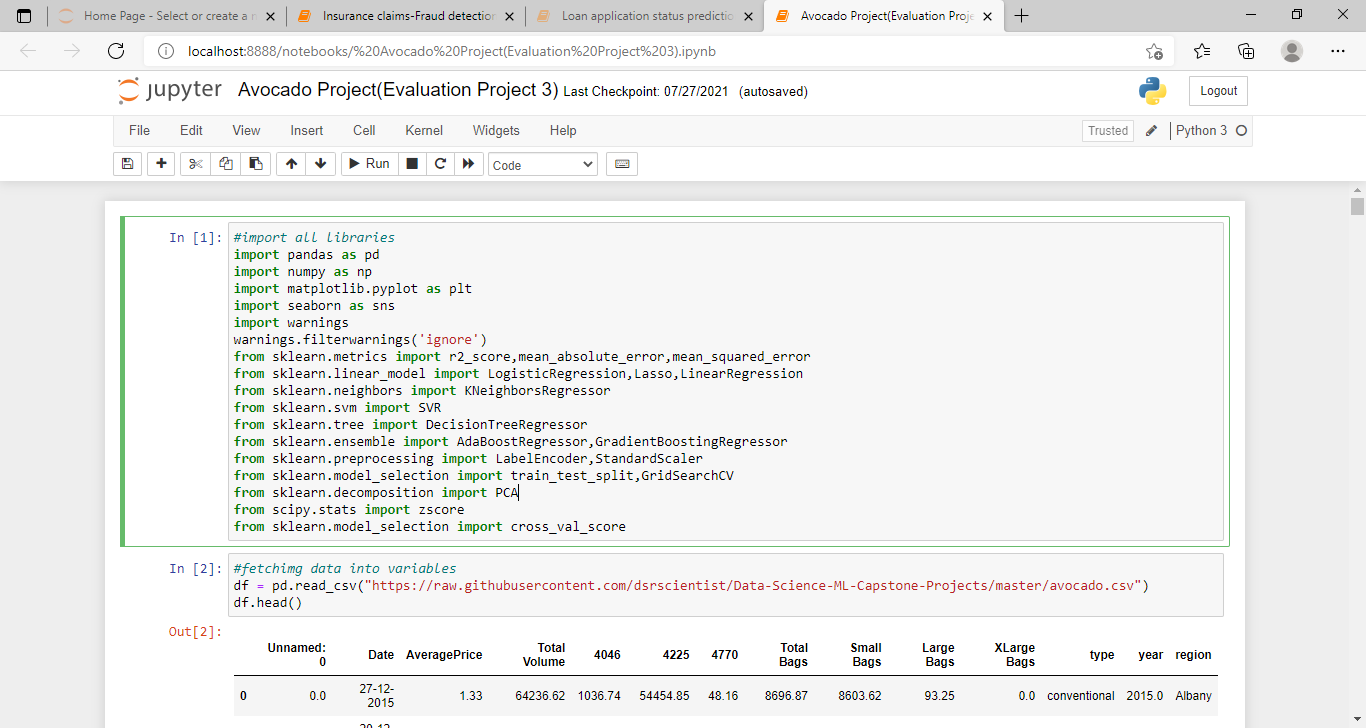
**Avocado,** is also called as **“alligator pear”**, fruit of Persea Americana of family Lauraceae. It is a tree native to the western Hemisphere from Mexico South to the Andean regions. Avocado fruits have greenish flesh or yellowish flesh with a buttery consistency. This is botanically a large berry containing a large single seed.

Data Preparation --🡪 Pre-Processing --🡪Feature Selection -🡪Feature Filtering -🡪 Classification Algorithm -🡪 Performance Evaluation

**PROBLEM DEFINITION:**

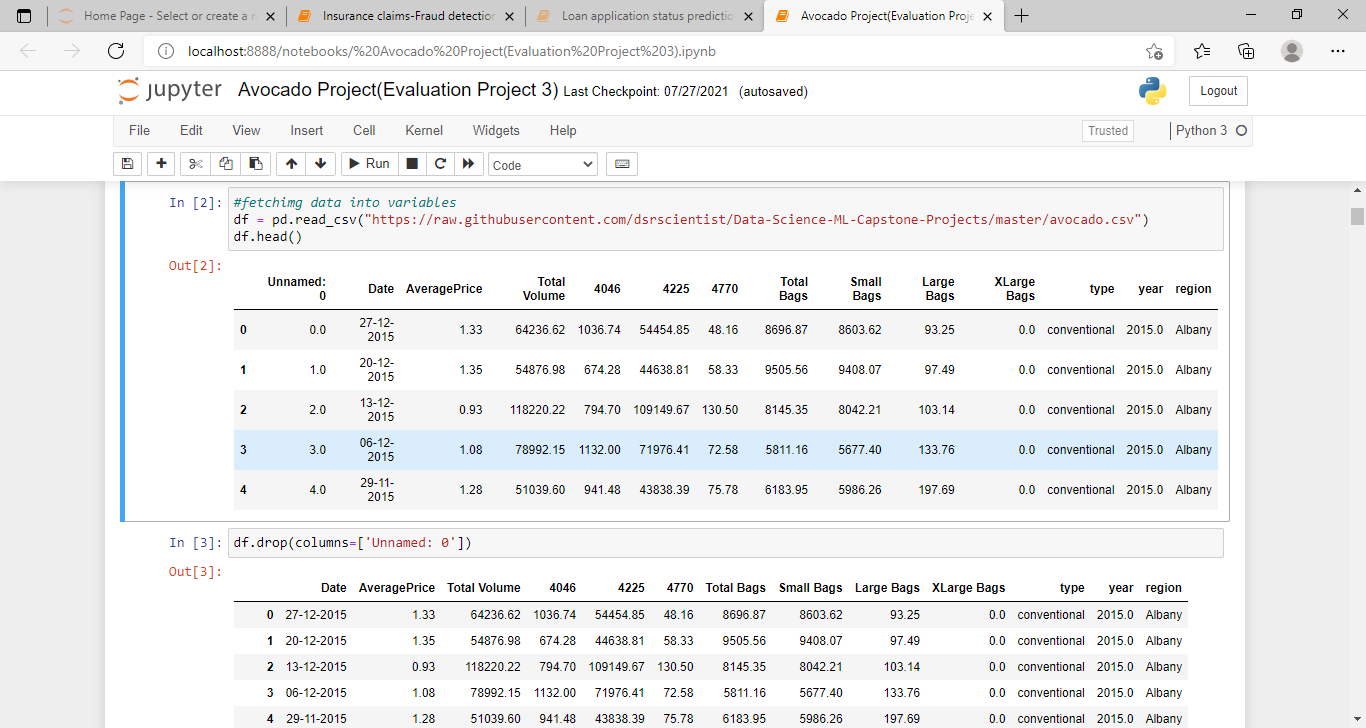
The main goal is to predict the average price which is continuous in nature of different type of avocado and using the region that in which they are lying. We have to import all the libraries to get the exact values for the output.

**Importing Libraries:**

****

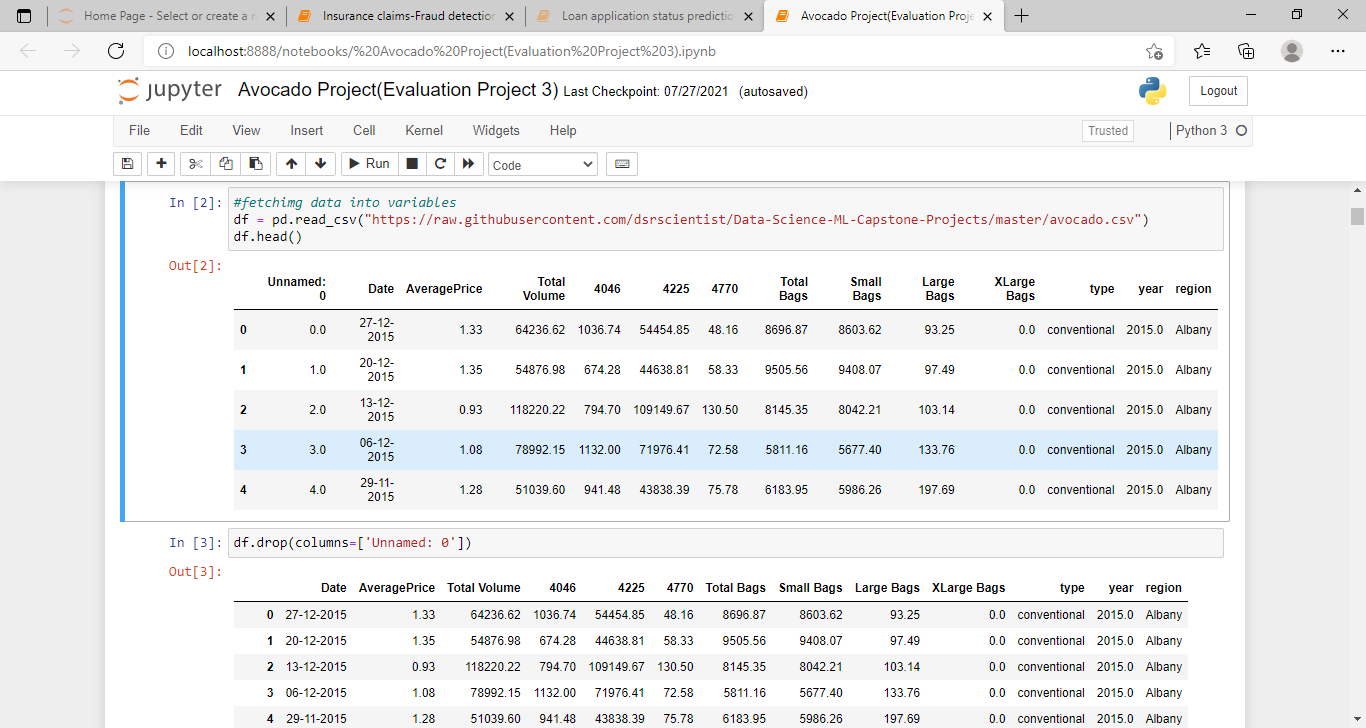
Now, we importing the library which I required for EDA, visualization, prediction and finding all the matrices. The reason of doing this is it become much easier to use all the import statement at on ego and we do not require to import the statement again at each point. We could find all the importing statement at one place without finding it on whole notebook and can update also.

Then, Loading data set into variables:

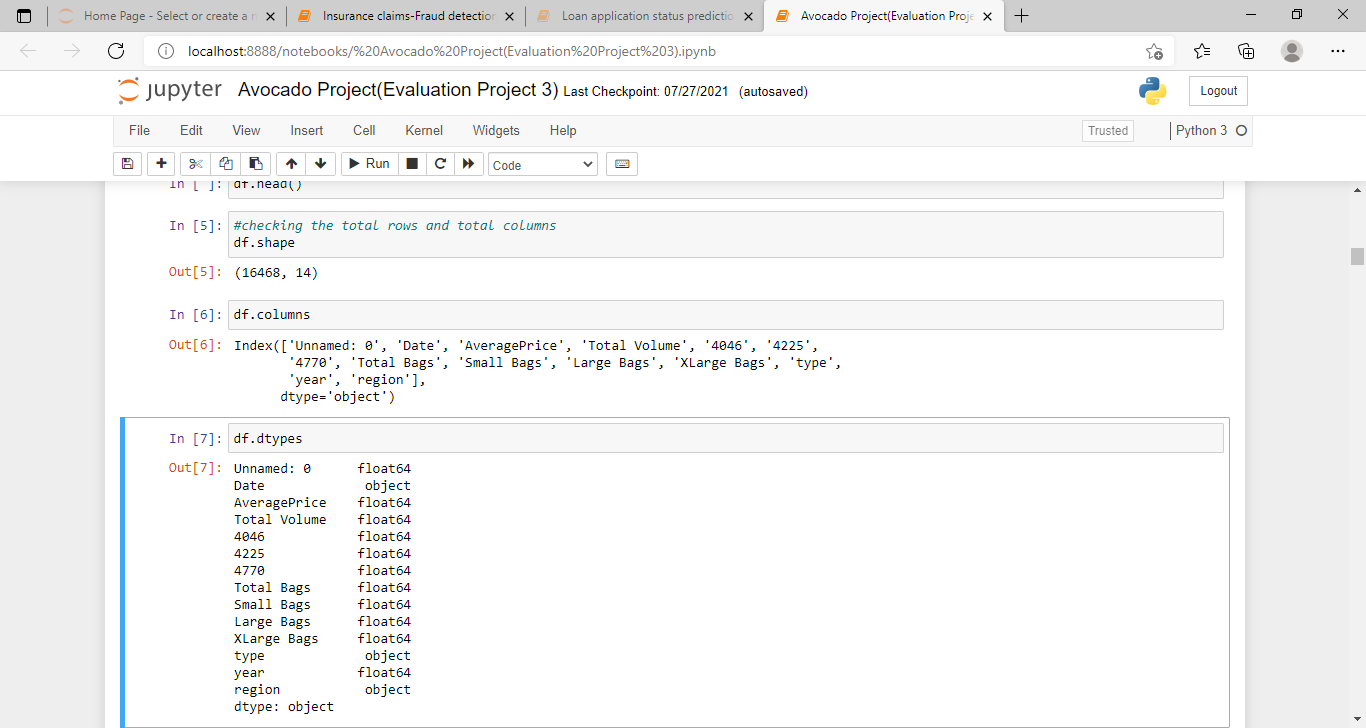


From the above table, we are loading the data set into variable “df”. And this data set most of the column float in the nature and type and sex is a categorical value.

**Then we have exploratory the data analysis:**

****

We have to find the shape, no. of columns and types of data set as below:



As we observed that, in data set there is one index column which does not play any important role for prediction of price in the avocado, because of that I dropping that column.

And also I am checking the shape of the data set there are 16468 rows, 14 columns after deleting the index column.

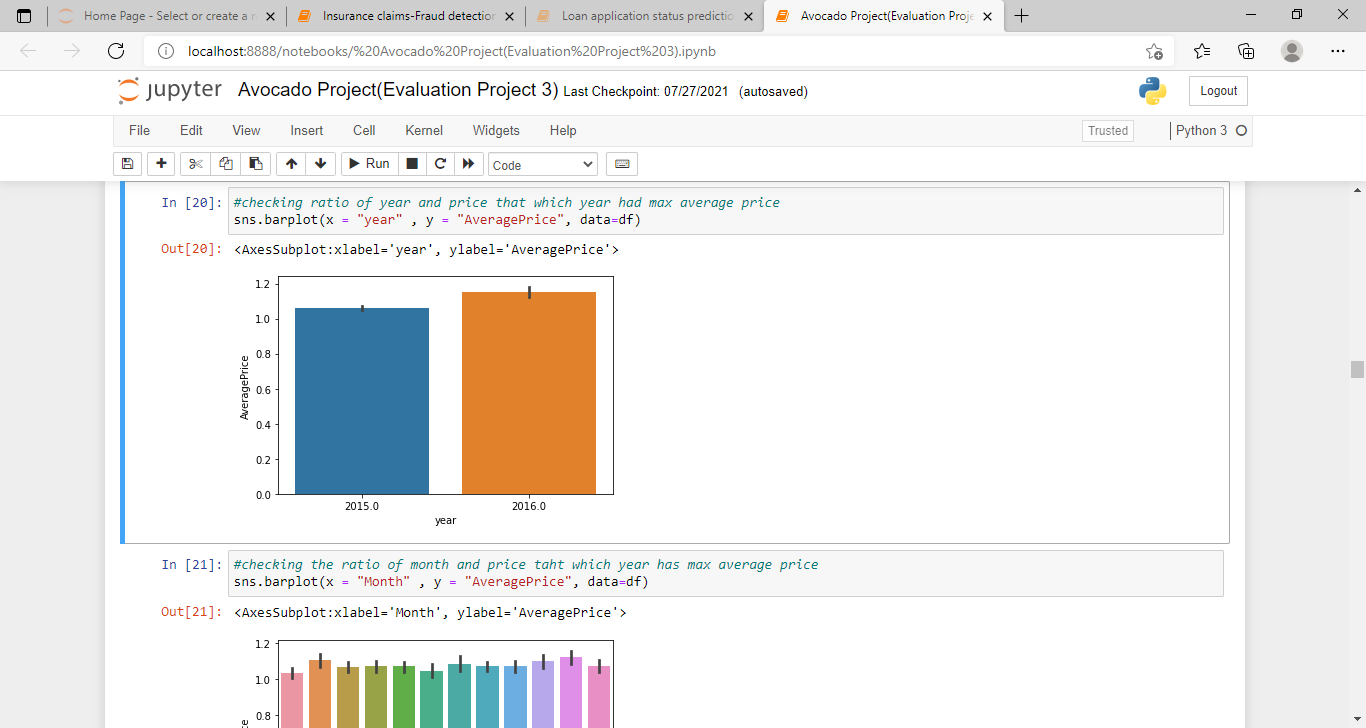
Also, most of the columns are same data type that is float, type, Date and region is of object data type.

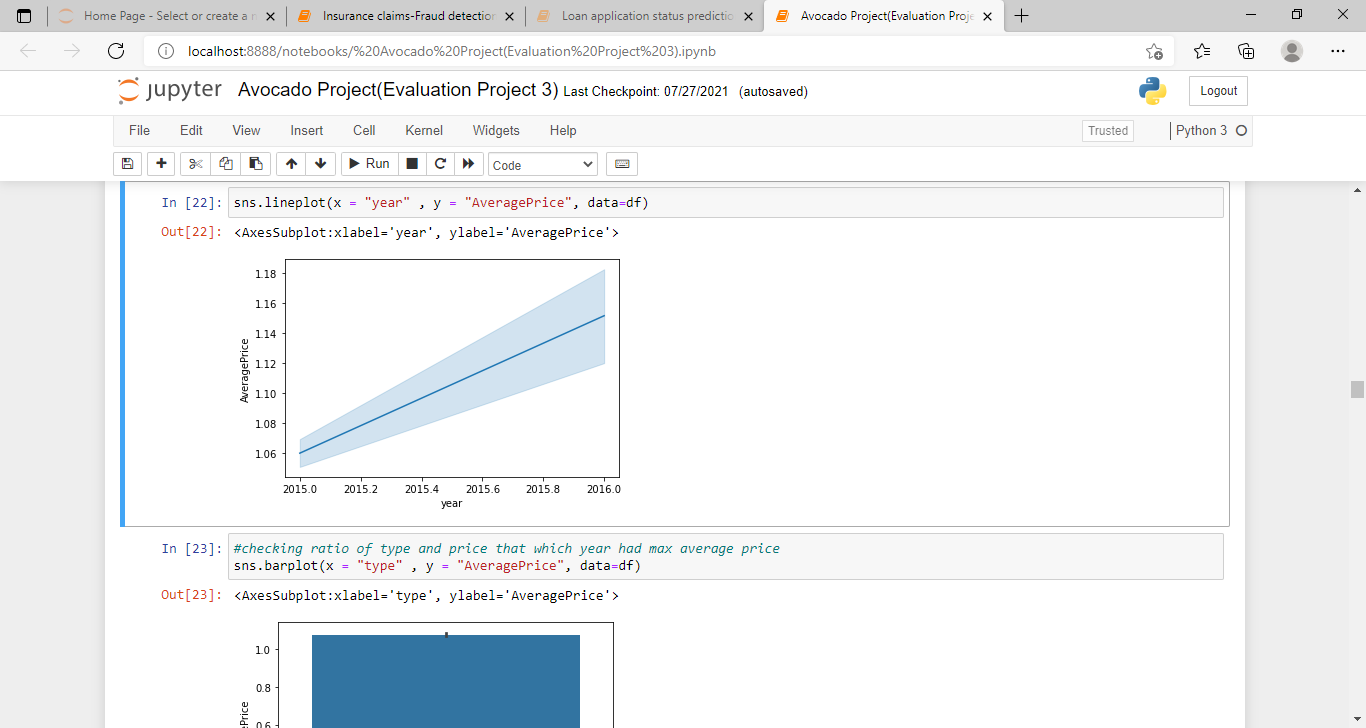
**Data Visualization:**

In this position, we can plot different graph using different columns and try to visualize the data using matplotlib and seaborn library.

We use many different graphs includes:

* Bar plot
* Count Plot
* Line Plot
* Histogram and Pair plot



Fr

From the above two graphs, we came to know that,

From year 2017, is that the year where the price is maximum as compared to other year, and there is less difference among rest of the years.

September and Octobers months have max. no of average price is there, but the thing is almost for whole year the price is almost same for the avocado this prove that there is so much Craze of avocado rather than India.

From the figure of second graph, there is hike between 8-10 both of type avocado both for conventional and organic type of avocado.

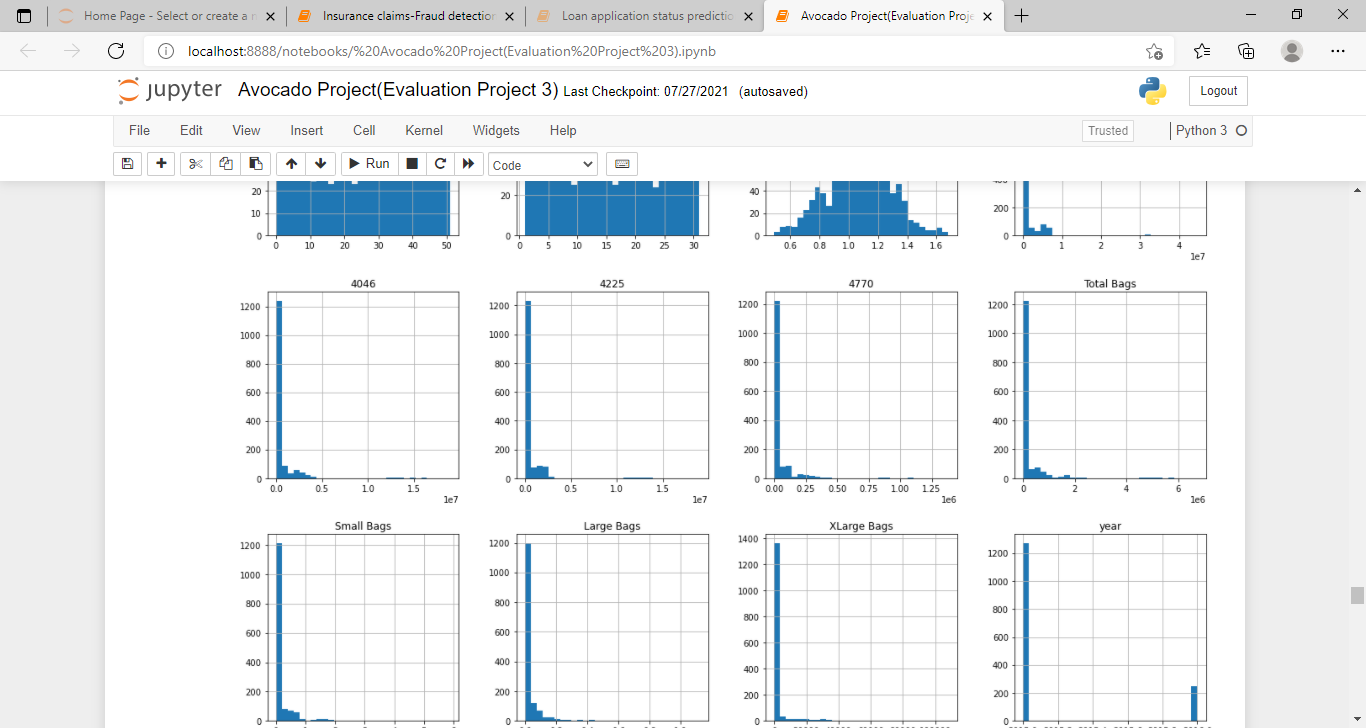
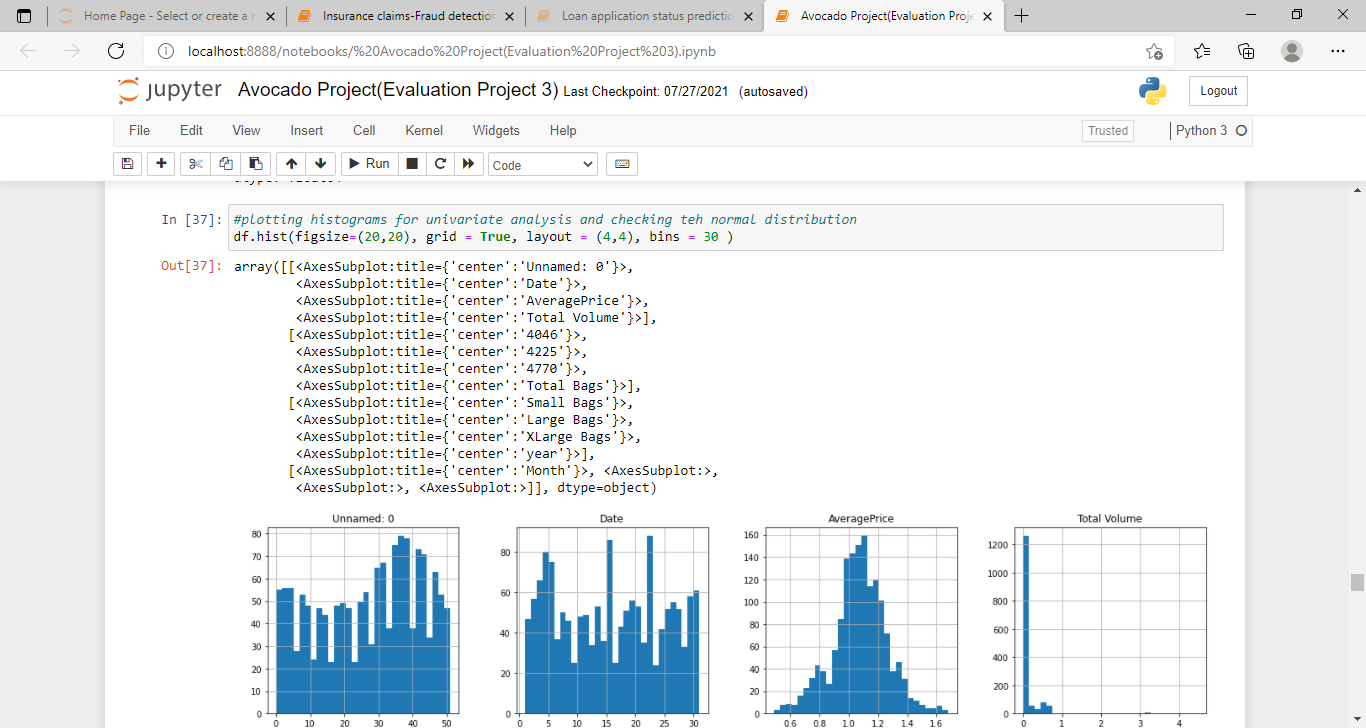
Also, the conventional type of avocado is varying in term of price as seen in line plot because in starting the price is high, then it decrease and so on.

**Plotting the Histogram:**

The Histogram, shows the frequency on the vertical axis and the horizontal axis, is of another dimension. Usually, it has bins, where as every bin has a minimum and maximum values.

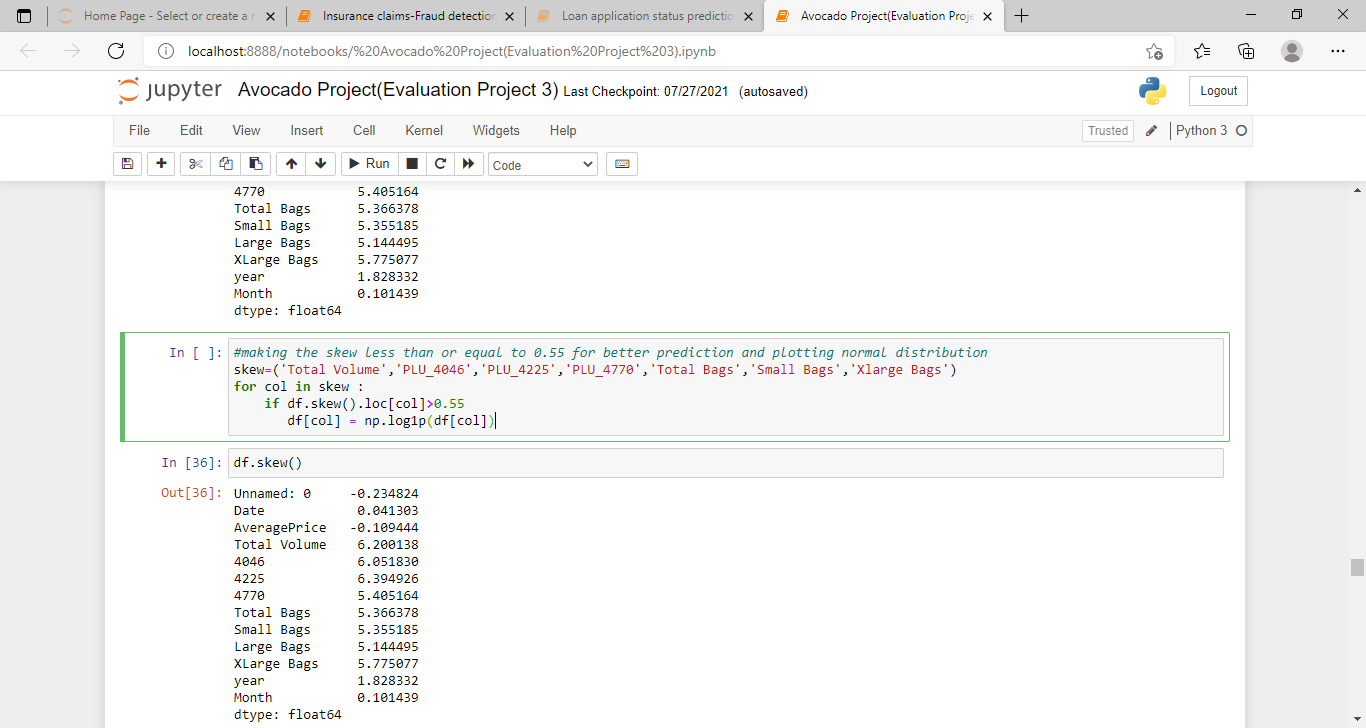
Each bin also has the frequency between x and infinite.

So, in this we can also check the graph whether the graph is right skewed, left skew or the graph is normally distributed graph.



From, plotting this histogram graph I used the bin size as 30, or we can take any bin size which can be suited as per as data.

* Average price column is normally distributing over the histogram graph.
* Rest of the data are not much varying in term of numbers, so they are almost left side skewed data.
* To make the column as normal distributed we can use different methods, but I am using the numPy log to make the skew values as normally distributed.



From the above figure, we are first calculating the skew value and some of the column skew value are far from zero.

* The best skew value for normally distributes is very close to zero, so we are using “log1p” method to make the skew value is near to zero.
* In the last cell I am again checking the skewness value and there is difference between the first skewness value and second, now the skewness value of each column is near to zero.

We make the note as, Making the skewness value near to zero will help to get better score.

**From Label Encoding:**

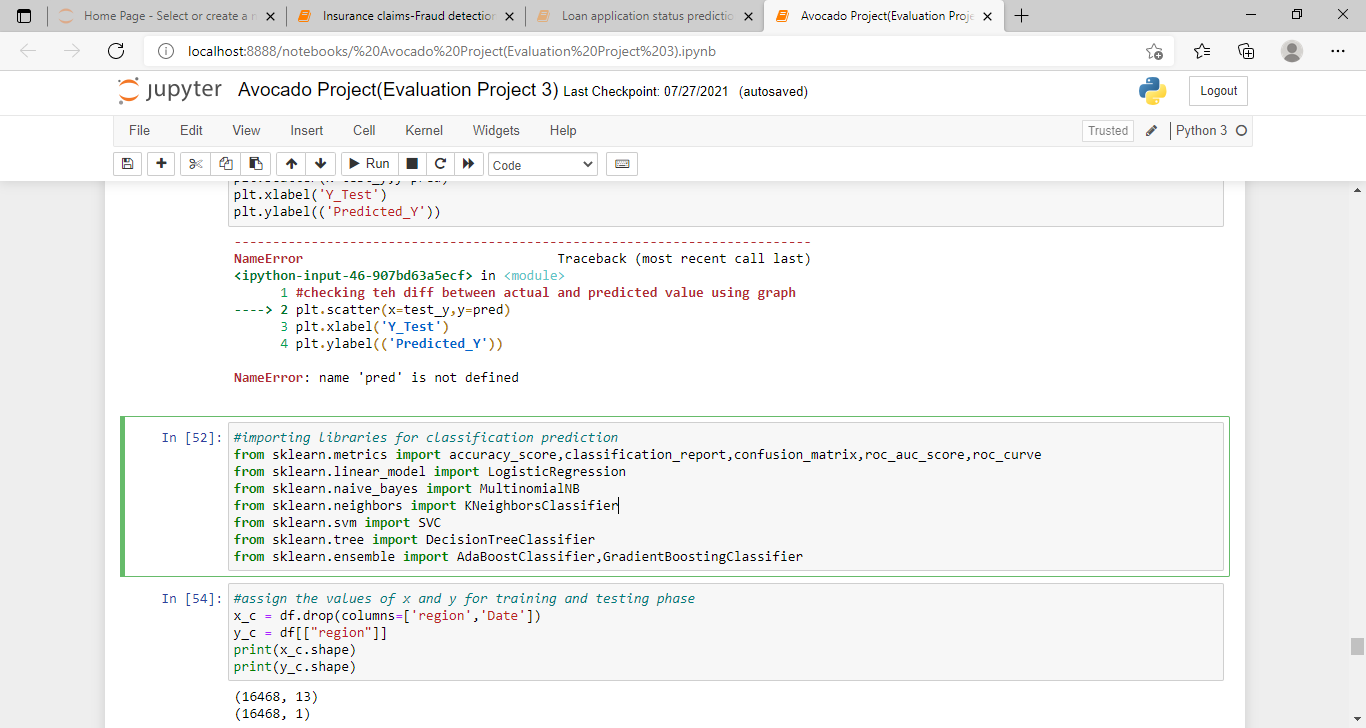
Sklearn provides a very efficient tool for encoding the levels of categorical features into numeric values.

**Label Encoder Encode Labels** with a value between 0 and n-classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier.

Convert region and type into numeric value by using encoder.

**Outliers:**

From the outliers, An outlier is an data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the data set.



**Correlation Matrix:**

**Correlation matrix** is basically a covariance matrix. A summary measure called the correlation describes the strength of the linear association.

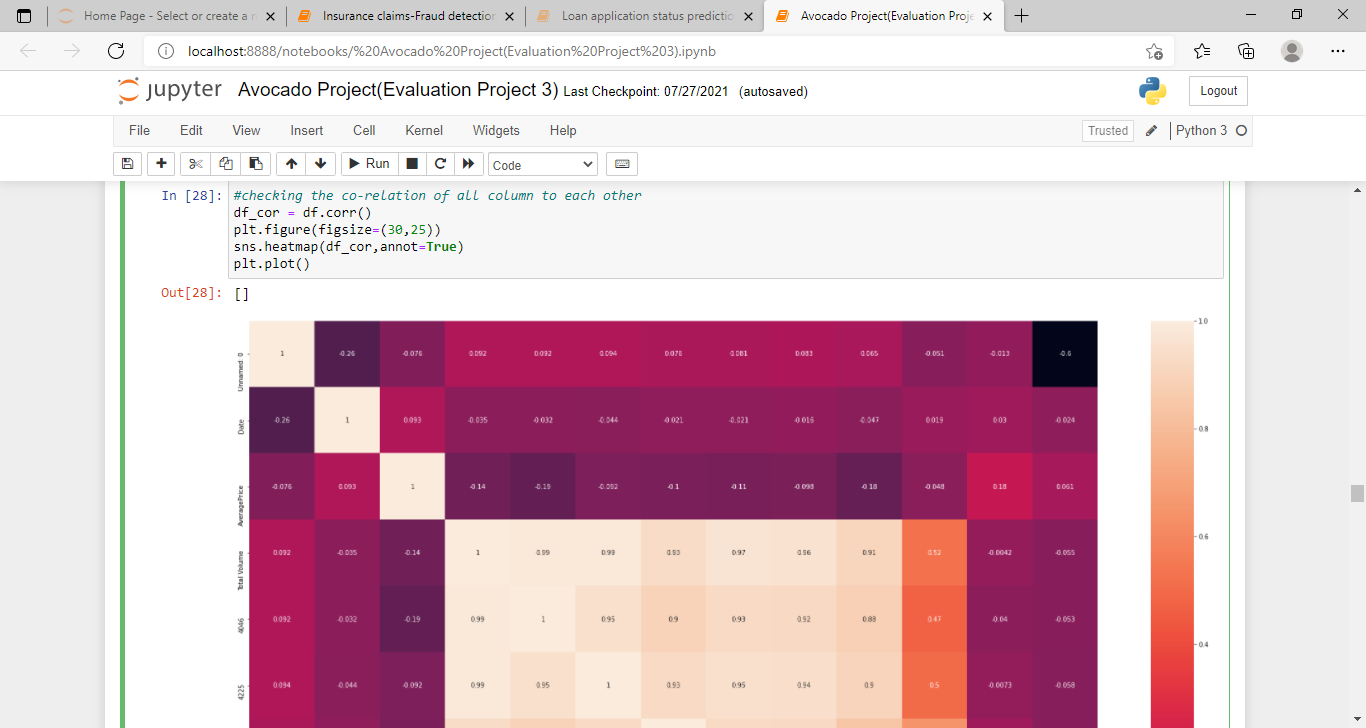
**Correlation** summarizes the strength and direction of the linear, association between two quantitative variables. This variables is denoted by ‘r’ it takes values between -1 to +1.

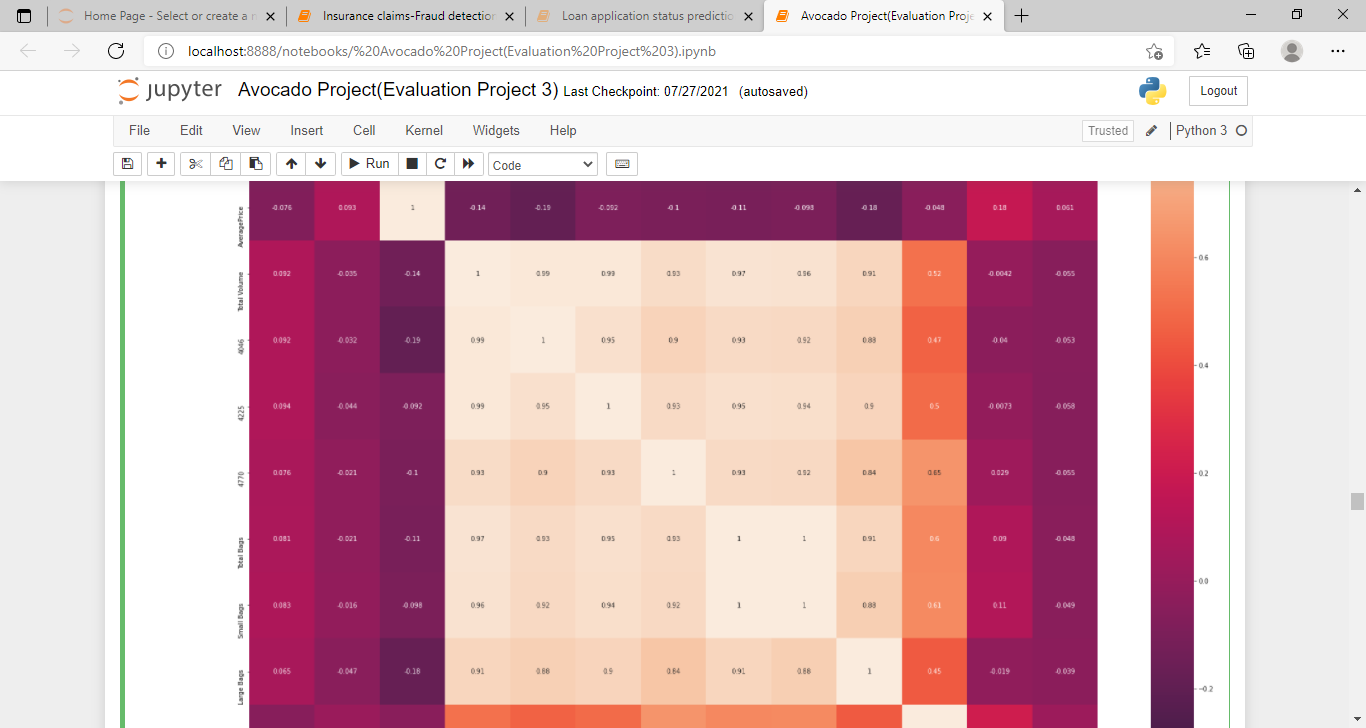
Now, I am finding the correlation values of each column, this value is categorized into mainly 2 parts. Mainly those are:

* Positive correlated value.
* Negative correlated value.

The most of the value is positive means that column is much co related and vice versa.

I am using seaborn heatmap to plot the correlated matrix and plot the corr value in the heatmap graph.



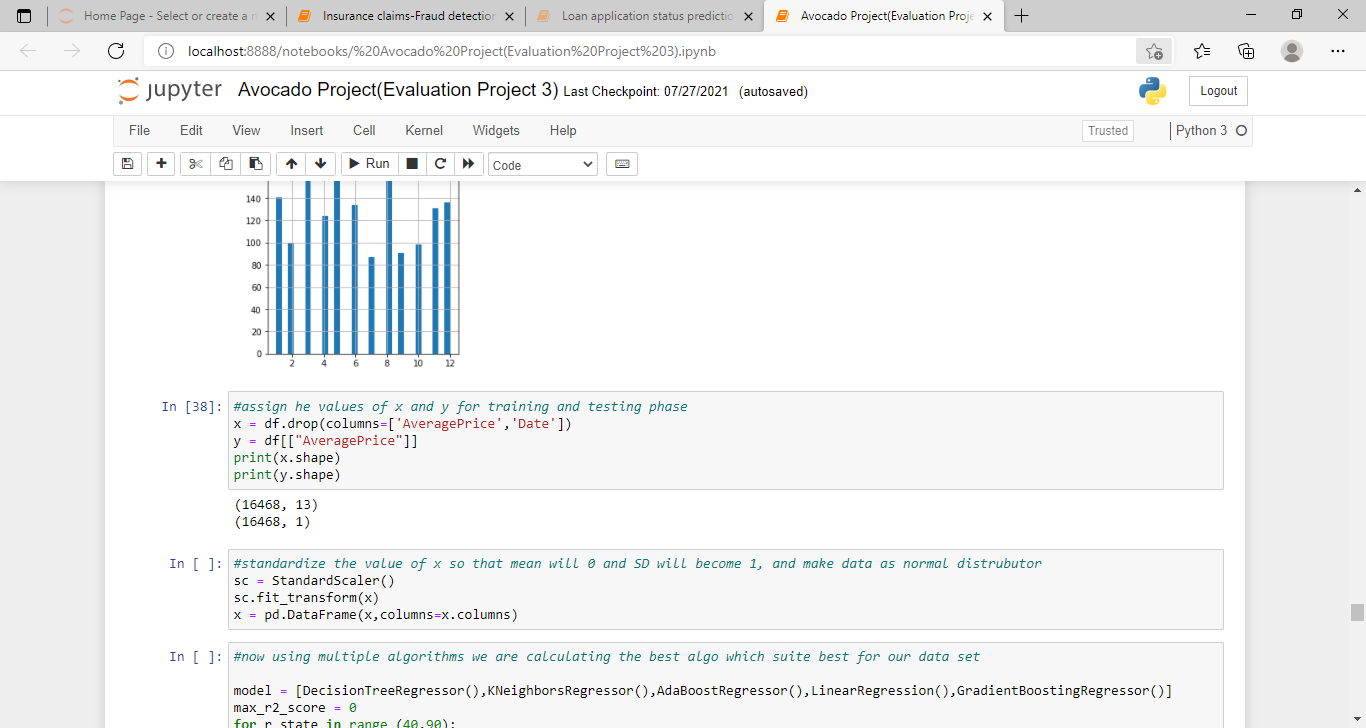


After correlation matrix, we can do:

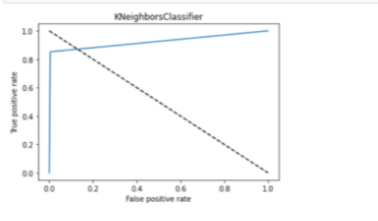
**Drop and Standard Scaler:**

Here I am making two variables x and y , where x is having all the columns except Average price and Date , we can also drop the Date column, but I kept for EDA purpose and y is having only Rings column.

Also, I am using the standard scaling method on x variable.



**Predicition with Region:**



In the prediction,

Here the random state that is occurring is 76 which provide the best accuracy score for the model which is 81%.

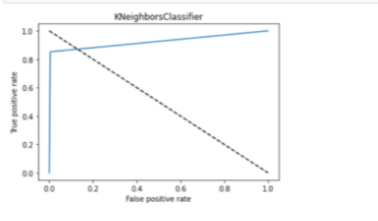
Also, by using the GridSearchCV, I am able to find the best param and then find the best accuracy score that is 85%.

**ROC Curve:**

A useful tool when predicting the probability of a binary outcome is the Receiver Operating Characteristic Curve or ROC Curve.

It is a plot of the false positive rate of x-axis, versus the true positive rate of y-axis, for a number of different candidate threshold values between 0.0 and 1.0. Put on another way, it plots the false alarm rate versus the hit rate.

The true positive rate is calculated as the no of true positives divided by the sum of the number of true positives and the number of false negatives. It describes how the good model is predicting at the positive class when the actual outcome is positive.



From, the above observation

* Taking price as y variable is predicting well for this model as compare to region.
* Also , I used the label Encoder to make the categorical data into numeric data i.e ; Region and Sex.
* Also , R2 score value is also greater than accuracy score.
* Average Price, total bags and total volume is well normally distributed data among all other column.
* There are no outliers in the data set after replacing it through mean value.
* As year is the most negative co related column among all the columns.
* In between August to October the price of avocado is much higher as compared to the other months.
* Date 28, 29 and 30 the price of avocado is high.
* Hartford spring field, San Fransisco and New York are having more average price as compared to another region.
* There is hike between month 8-10 of both type of avocado.
* As organic type of avocado is having the more price per unit then conventional.
* I had done prediction using region and price but using price the prediction score is high as compared to region.
* So, in this data set I am using both regression and classification technique for making this model.