**Avocado dataset**



**Problem description:**

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. Here’s how the Hass Avocado Board describes the data on their website.

The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (e.g. green skins) are not included in this table.

The avocado, a tree likely originating from south-central Mexico, is classified as a member of the flowering plant family Lauraceae. The fruit of the plant, also called an avocado, is botanically a large berry containing a single large seed.

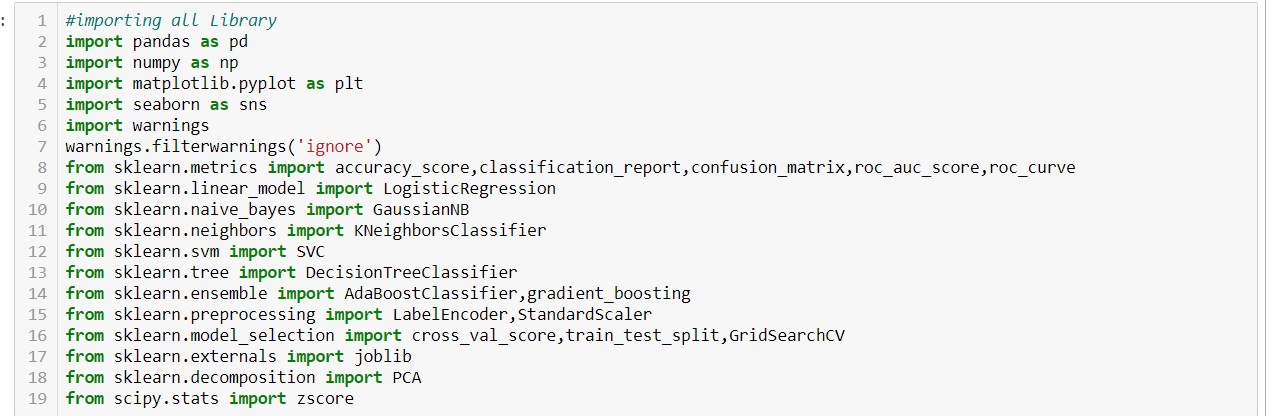
**Inspiration /Label**

The target variable for this dataset will be "AveragePrice". This is what I have to predict

**Fields/Columns:**

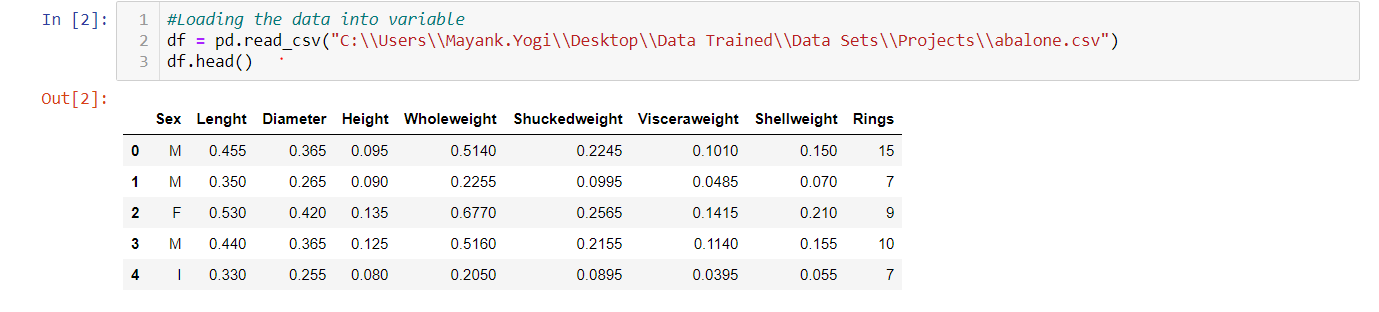
* Date — The date of the observation
* Average Price — the average price of a single avocado
* type — conventional or organic
* year — the year
* Region — the city or region of the observation
* Total Volume — Total number of avocados sold
* 4046 — Total number of avocados with PLU 4046 sold
* 4225 — Total number of avocados with PLU 4225 sold
* 4770 — Total number of avocados with PLU 4770 sold

**Importing Reqired Library:**



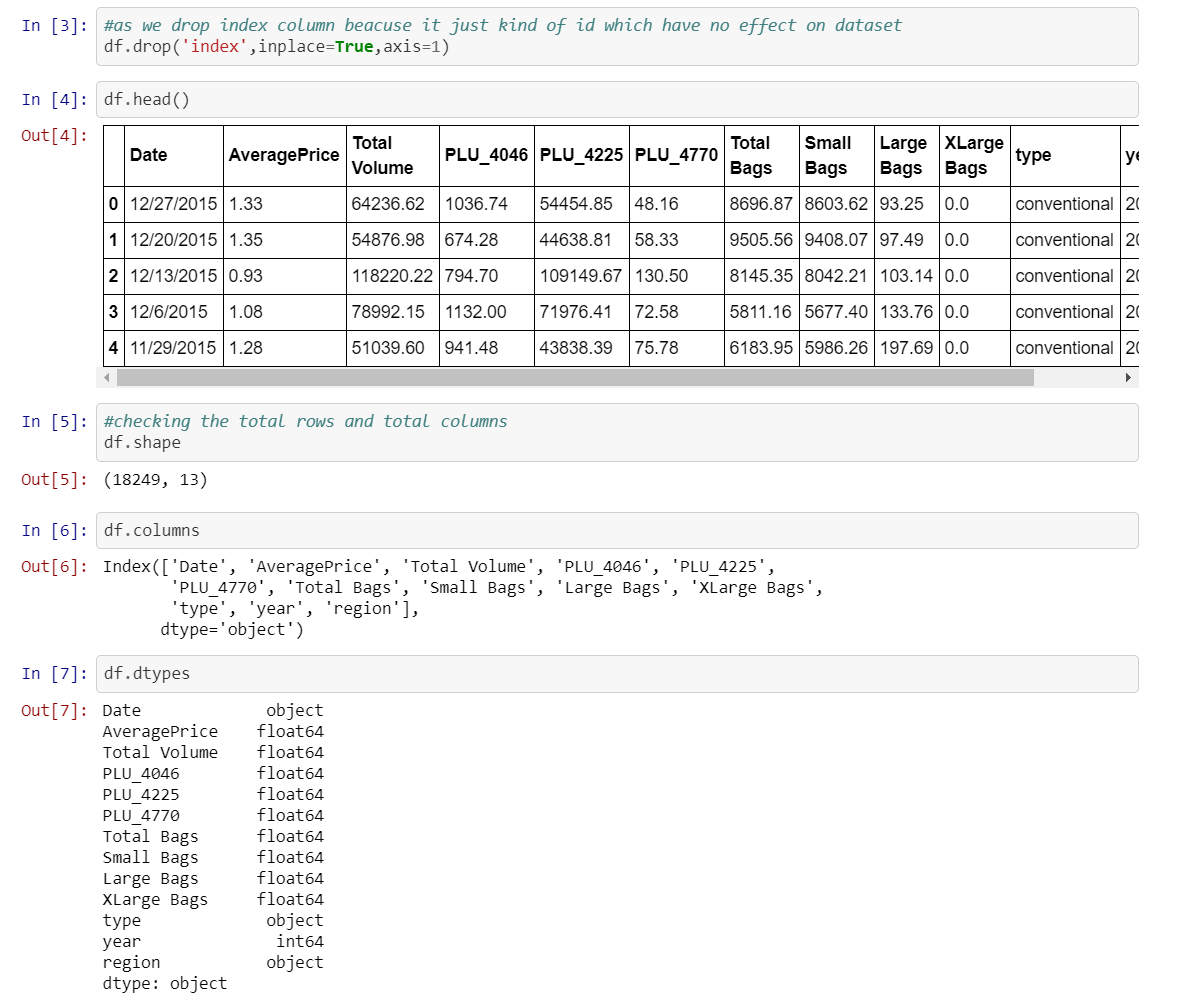
I am importing the all library which I required for EDA, visualization, prediction and finding all matrices. The reason of doing this is that it become easier to use all the import statement at one go and we do not require to import the statement again at each point. We could find all the importing statement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

**Loading Data Set into variable:**

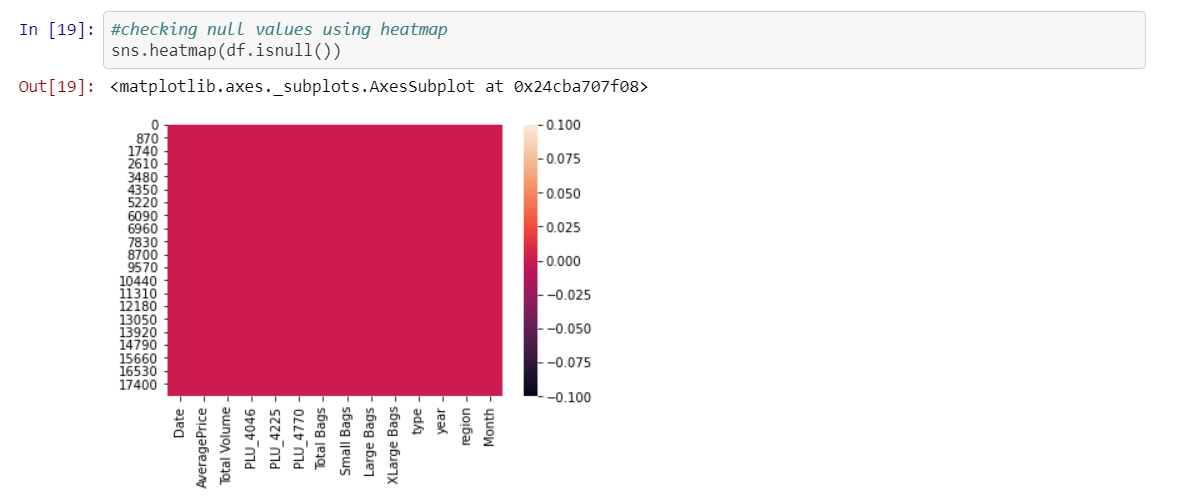


Here I am loading the data set into a variable i.e. “df” and processing the first 5 rows.As in this data set most of the column are float in nature and type and sex is of categorical value.

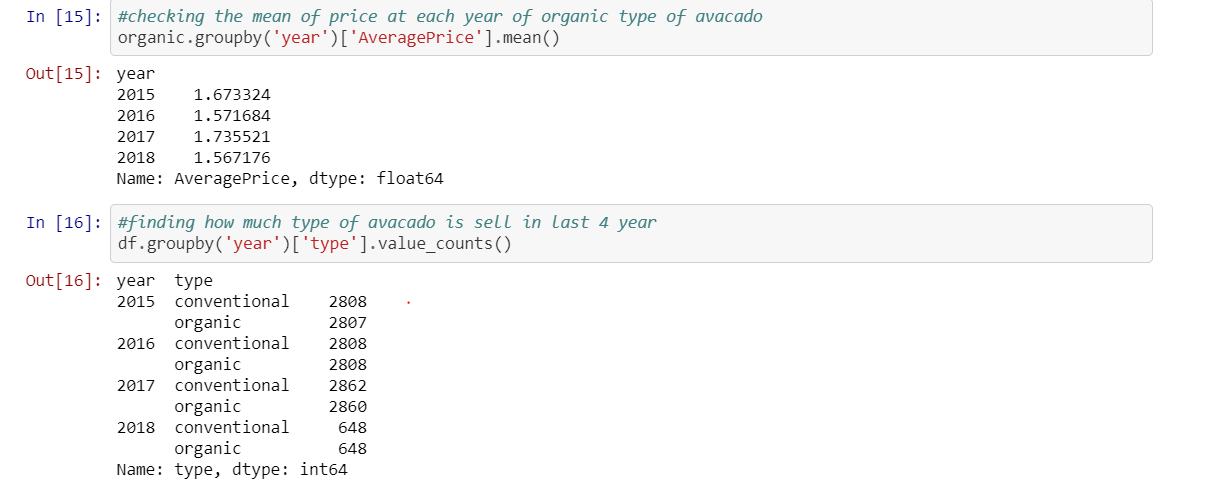
**Data Analysis:**



* As seen in data set there is one ‘Unnamed:0’ column which does not play any important role for prediction in the price of avocado, so I am dropping that column.
* Also, I am checking the shape of the data set as there are 1517 rows and 13 columns after deleting the index column.
* As we there are no null vales present in the dataset.
* Also, most of the column are of same data type that is float type ,region and date are of object data type & year is of int type.



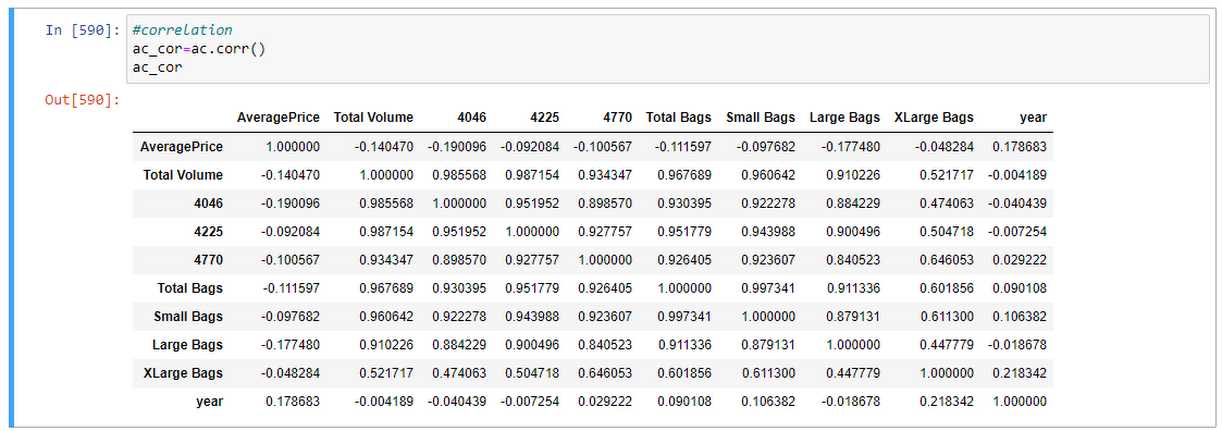
Above I am checking the null values, as find there are no null values in the data set because the red color is distributed equally correspond to each column.



In above, I am finding that year 2017 is aggressive year where avocado price is higher as compared to other year and 2015 is at second number.

Also, I am finding that at each year present in the data set, which type of avocado is has total count, so both type of avocado is present almost in same amount in the data set.

**Stastical summary**



 Above statistics data show that their multiple outliers mostly in XLargeBags There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction

* Also, number of rows in each column are same, means there are no null values in the data set.
* Also, the mean and 50%value of most of the column are same and the STD and mean are very close to each other.
* Most of the column statistics data are near to 0 values.
* By checking the difference between the 75% and max value there are outliers in some of the column, I will check it soon.

**Exploratory Data Analysis:**

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

We use different graph include:

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



From above we came to know that:

* Year 2017 is that year where the price is maximum as compared to other year, and there is less difference among rest of the year.
* September and October are the month where 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.

**Plotting Histogram:**

* A **histogram** shows the frequency on the vertical axis and the horizontal axis is another dimension. Usually it has bins, where every bin has a minimum and maximum value. Each bin also has a frequency between x and infinite
* So, in this we can also check whether the graph is right skewed, left skew or the graph is normally distributed graph.

From plotting this histogram, I used the bin size as 30, we can take any bin size (suited as per as data).

* Average price column is normally distributing over the histogram.
* Rest of the data are not much varying in term of numbers, so they are almost left skewed data
* To make the column as normal distributed we can use different methods, but I am using numPy log to make the skew values as normal distributed.
* **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 (straight-line) association between two quantitative variables. Denoted by r, it takes values between -1 and +1.
* Now I am finding the correlation value of each column, this value is categorized into mainly 2 parts that are:
* - Positive correlated value
* - Negative correlated value
* The most 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

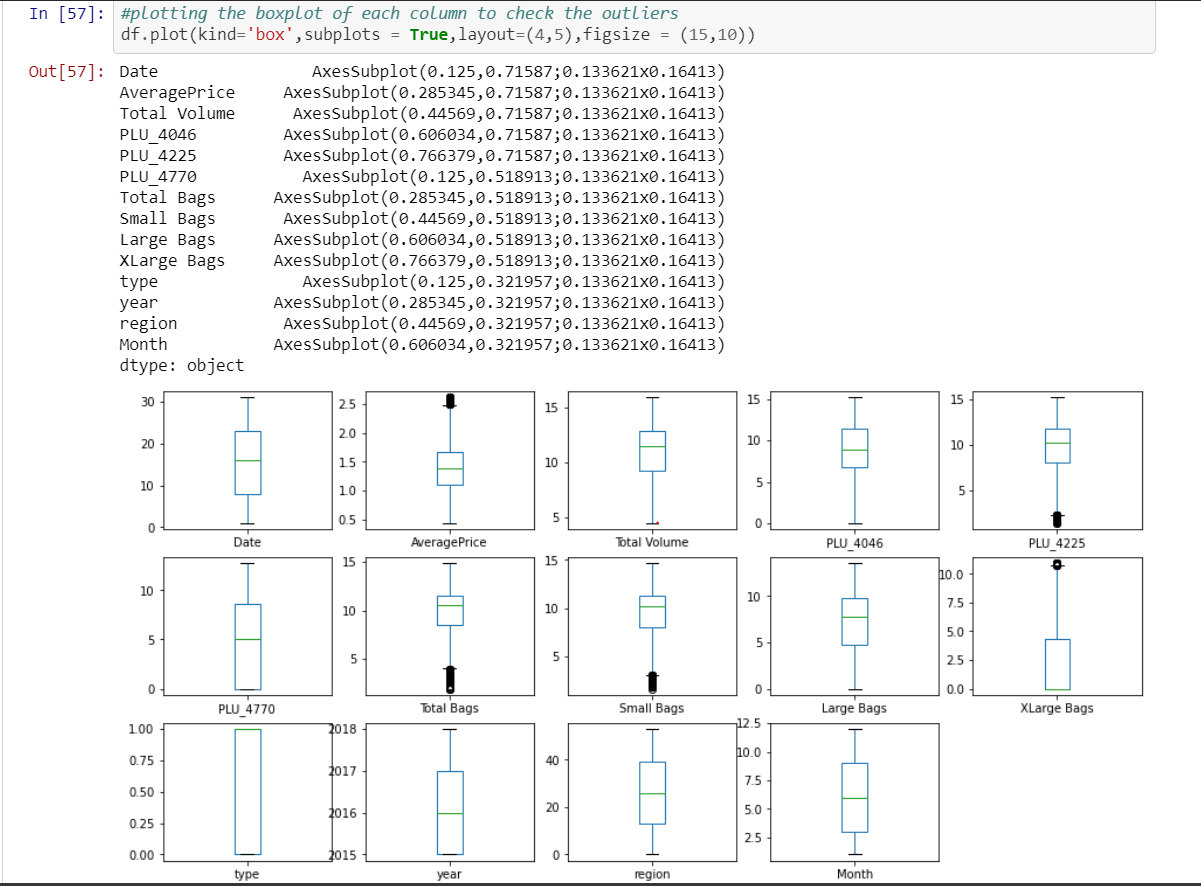
**Checking outliers:**

From above we can say that many outliers are present in cloumns.

An **outlier** is a 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

Now that we know outliers can either be a mistake or just variance, how would you decide if they are important or not. Well, it is simple if they are the result of a mistake, then we can ignore them, but if it is just a variance in the data, we would need think a bit further.

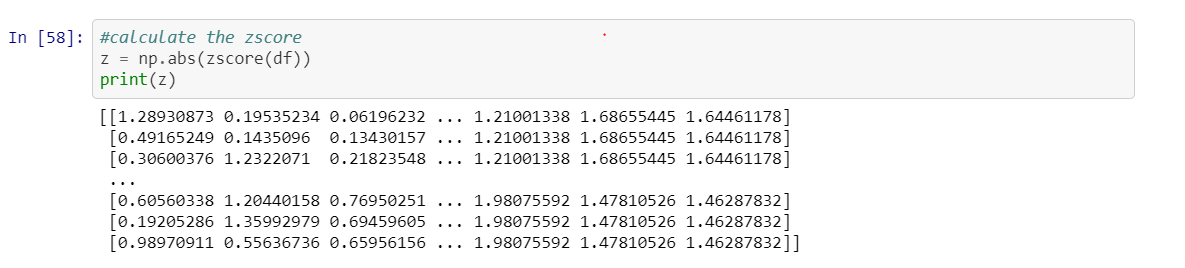
For avocado problem first check the outliers of each column.

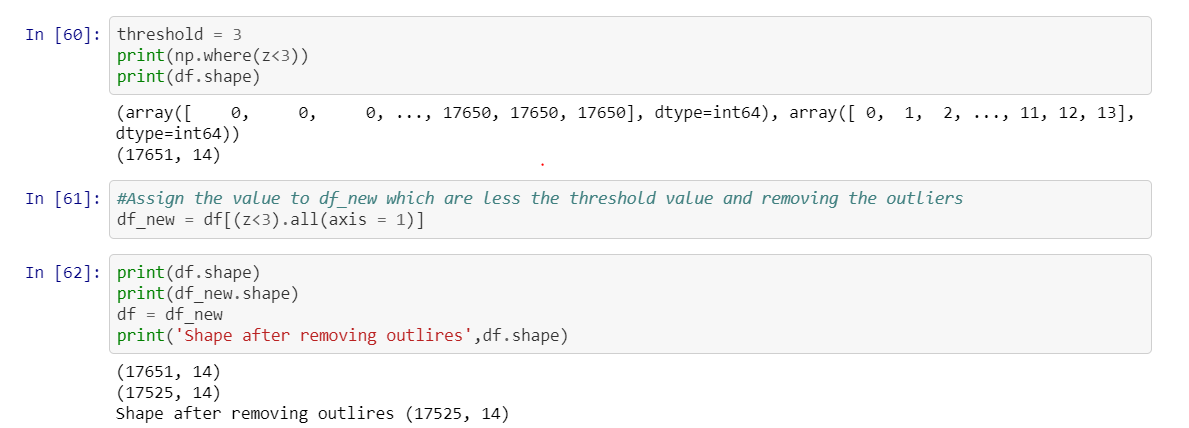


From above image we can clear see that there are number of black dots in most of the column which are referring to the outliers, so it means most of the data are outside the distribution.

So now we detect the outliers now the second step is to remove the outliers, there are different way to remove the outliers that are find the IQR, zscore values.

I am using both zscore value then I again check if there are some of the outliers then I will remove it by replacing the outliers with the mean value of that column.





So, I first find the zscore value and then I decide to make one threshold value as 3 which is standard of industry recommend value and then I remove all the outliers which zscore value is greater than 3.

**Pre Processing Pipeline:**

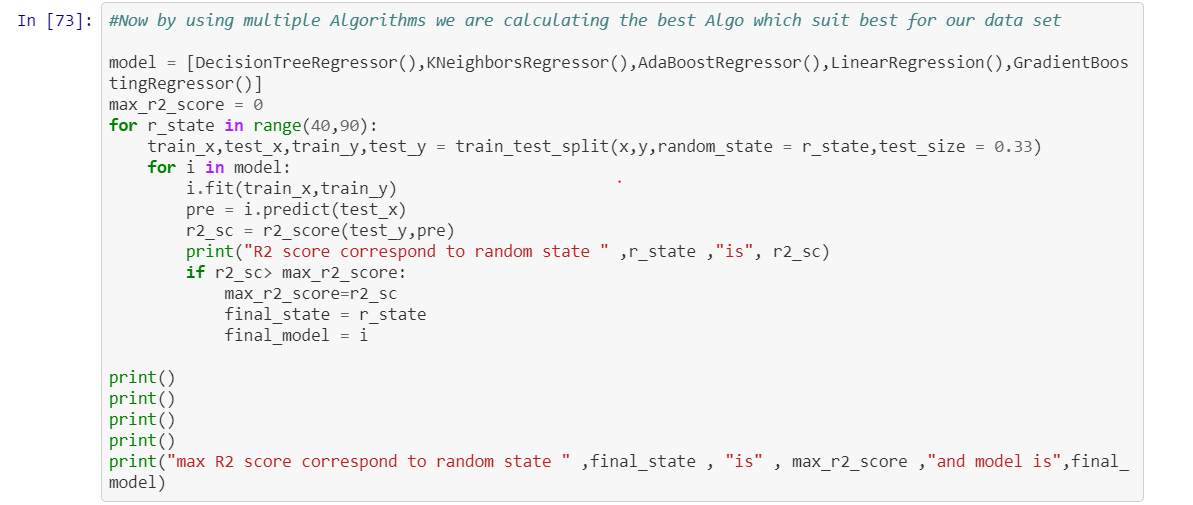
**Separating data variable and independent variable i.e. Average price**

**Drop and Standard Scaler:**

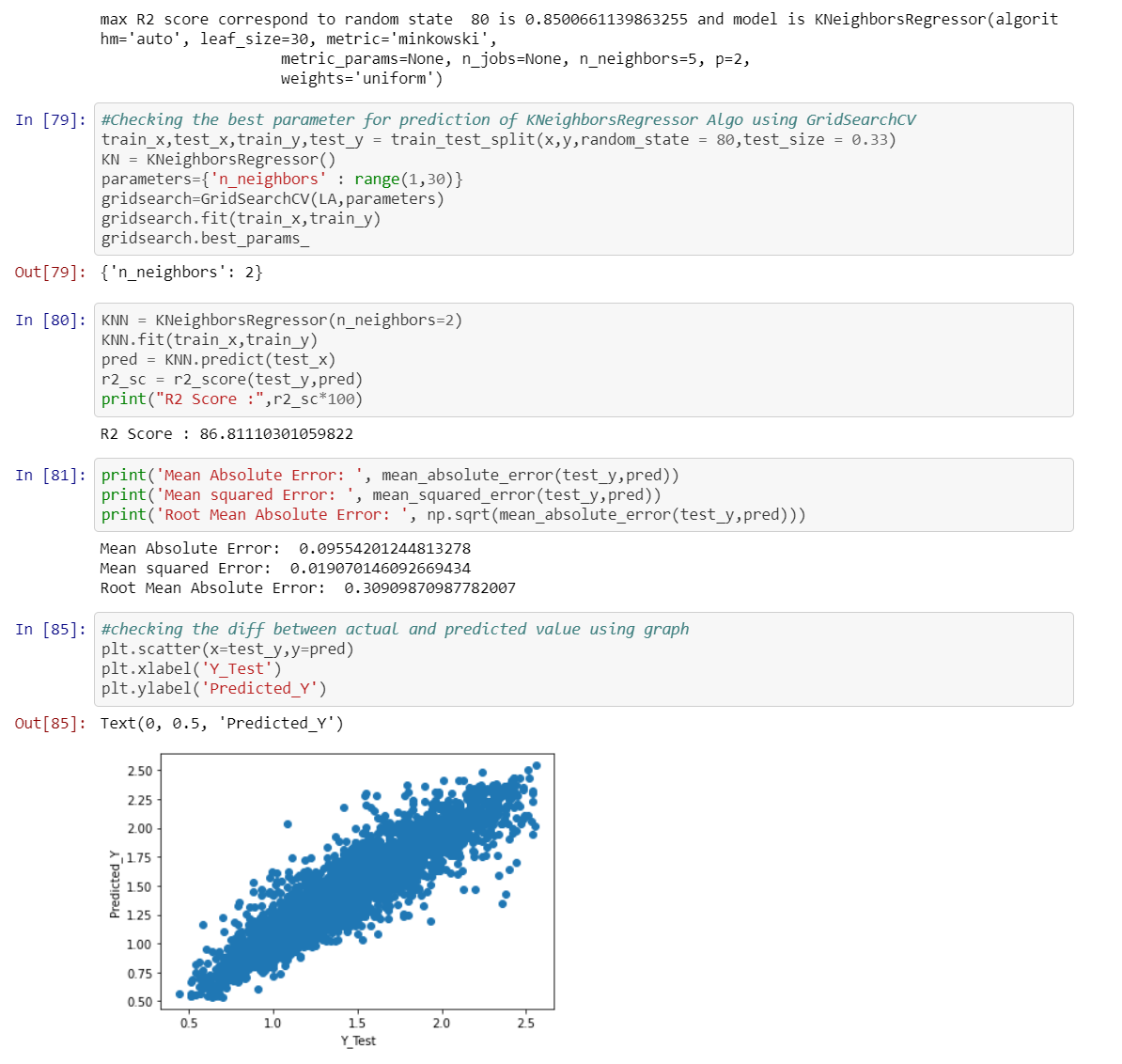
Here I am making two variable x and y where x is having all column except Average Price and Date, we can also drop the Date column, but I kept for EDA purpose and y is having only Rings column.

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

**Building Machine Learning Models:**



Above I am using the for loop which help me to provide the R2 score at each random state and for the best state where R2 score is maximum is come as output value.

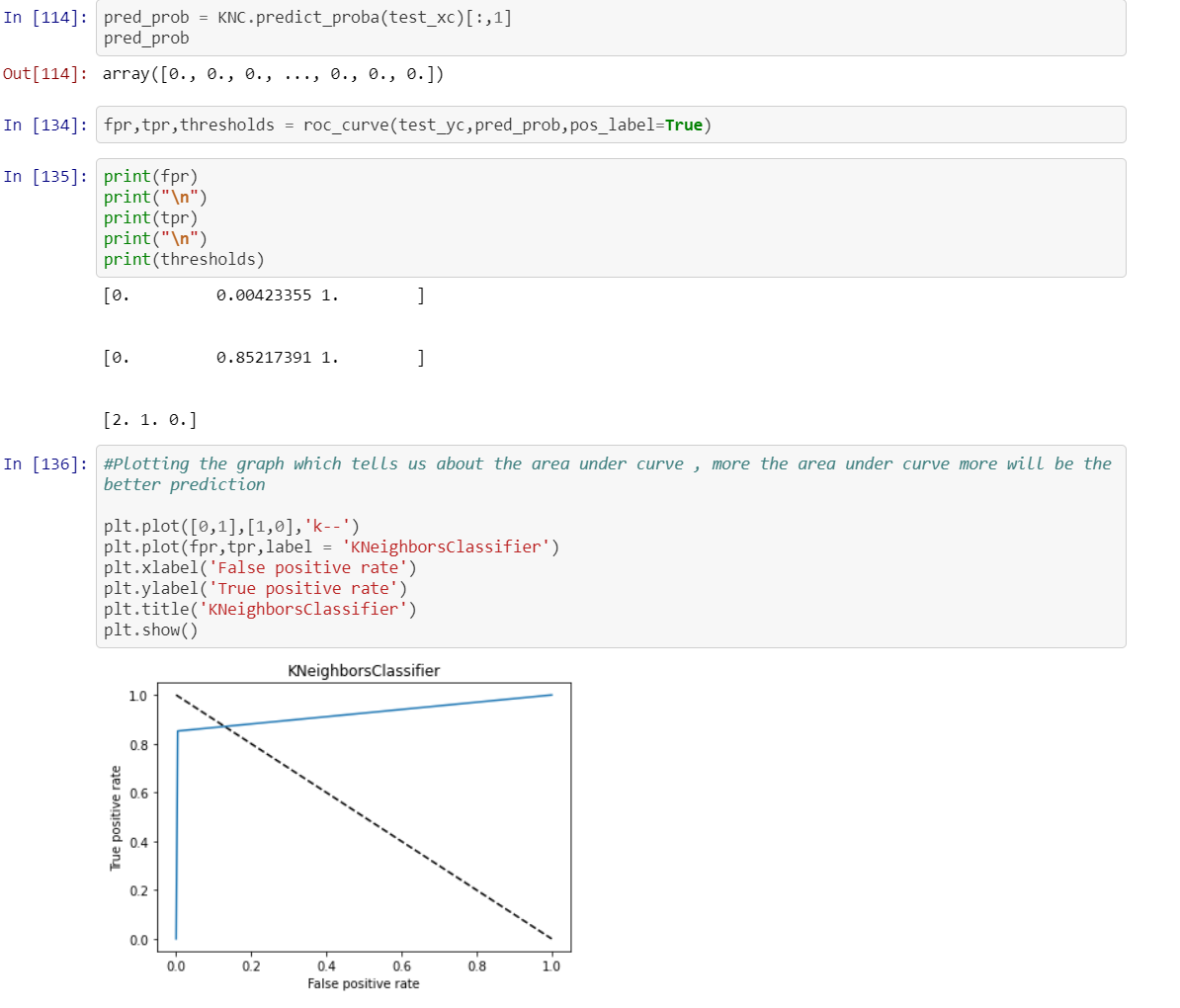


**Roc Curve:**

A useful tool when predicting the probability of a binary outcome is the [Receiver Operating Characteristic curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic), or ROC curve.

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

The true positive rate is calculated as the number of true positives divided by the sum of the number of true positives and the number of false negatives. It describes how good the model is at predicting the positive class when the actual outcome is positive.



**Concluding Remarks**

* R2 score value is also greater then 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 most negative co related column among all columns
* In between August to October the price of avocado is much higher as compared to other months
* Date 28,29 and 30 the price of avocado is high
* Hartford Springfield, San Francisco 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 Average price .
* So in this data set I am using regression technique for making this model.