**Avocado Price Prediction Using Machine Learning**

# By Neha Chand



Avocado is a fruit consumed by people heavily in the United States. The data was downloaded from the Hass Avocado Board website in May of 2018 and compiled into a single CSV.

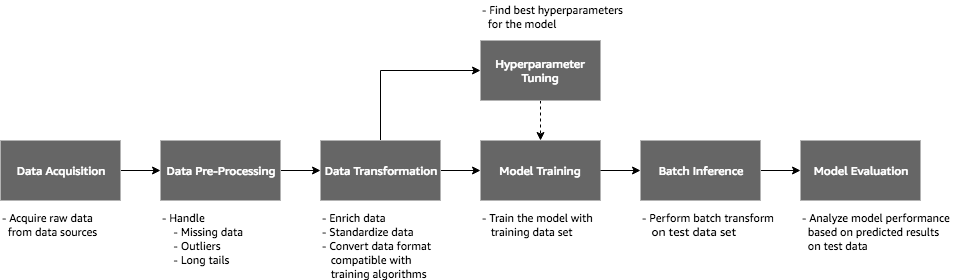
Here’s How the Hass Avocado Board describes the data on their website:

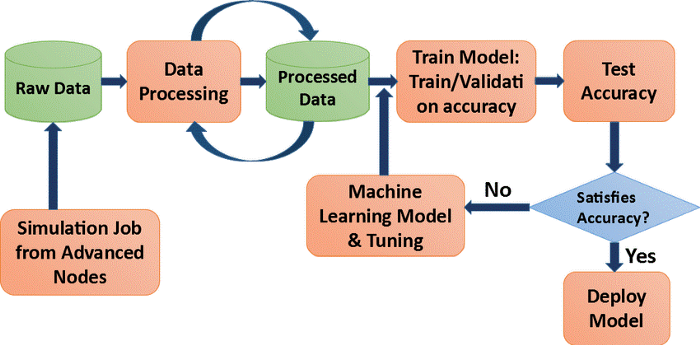
The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects as 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

**Work flow in Machine Learning:**





**Inspiration / Label :**

Your task is to make a mode that can consider the data provided and predict the Average Price.

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

**Problem description:**

The goal is to predict the average price which is continuous in nature of the different type of avocado and using the region that in which region they are lying.

**Importing necessary libraries:**

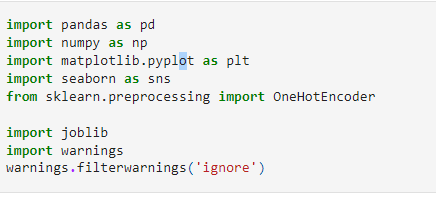
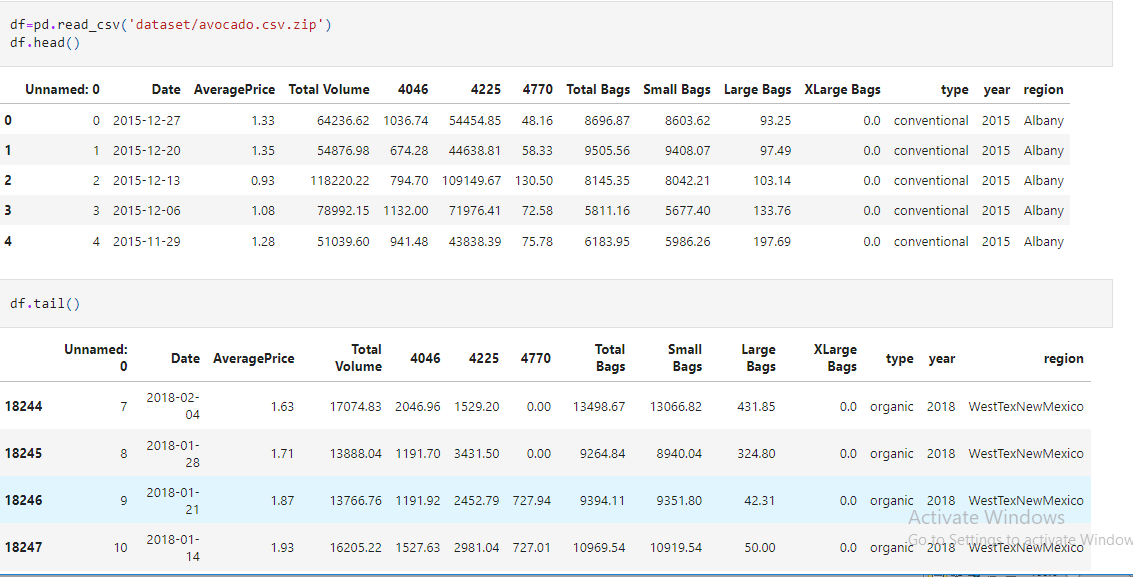


Fig 1

I am here importing the necessary libraries which are required in EDA, visualization and here are many more libraries which were used further for the prediction and finding all matrices and also for the model training, I mentioned the other libraries were I needed to mention them.

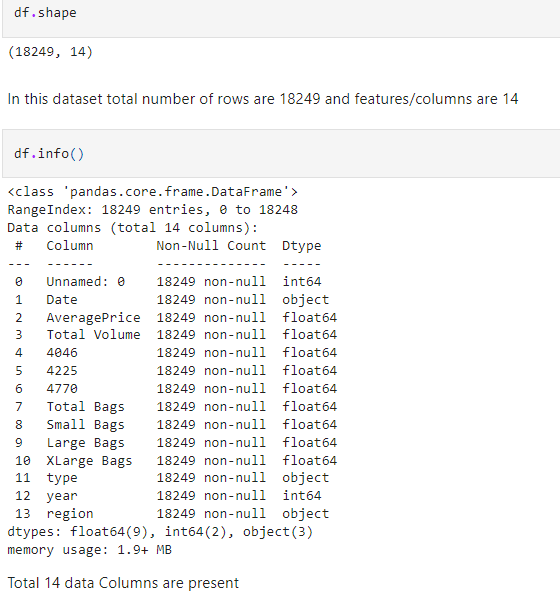
It is very important to load the necessary libraries first because it helps to perform different functions on our dataset.

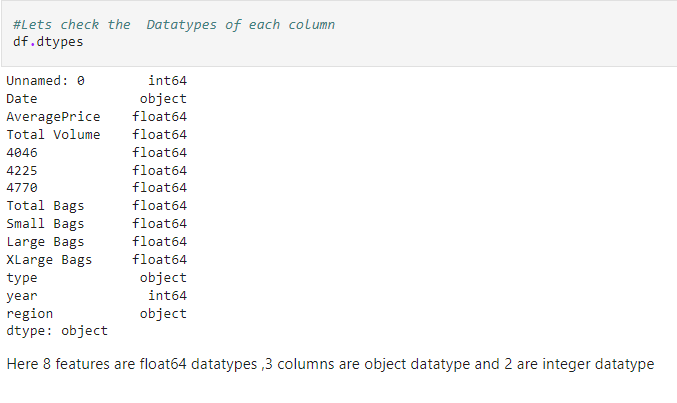
**Loading Data set into variable:**



Here I am loading the data set into the variable i.e.”df ” and also we processing the first 5 and the last 5 rows of the dataset. In this dataset most of the features/columns are float in nature and type and sex are of categorical value.

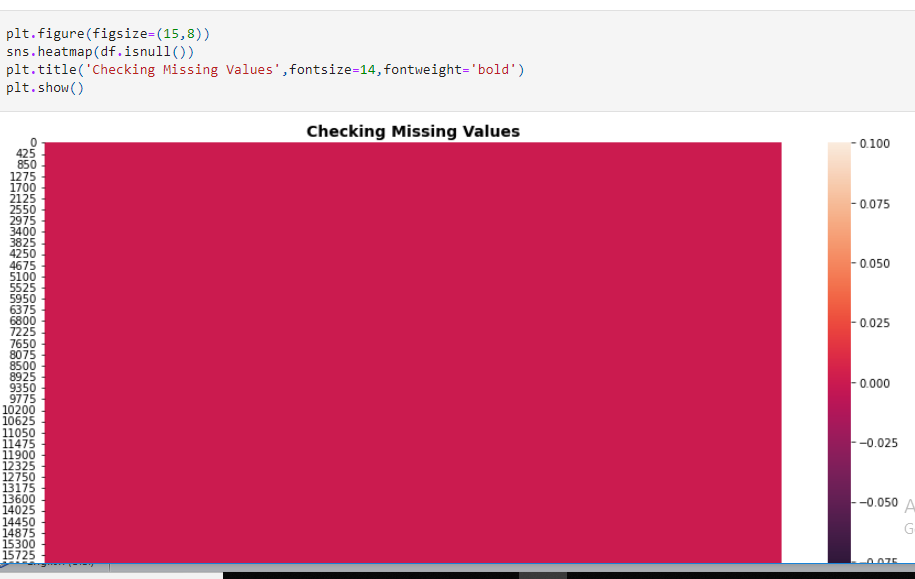
**Exploratory Data Analysis:**

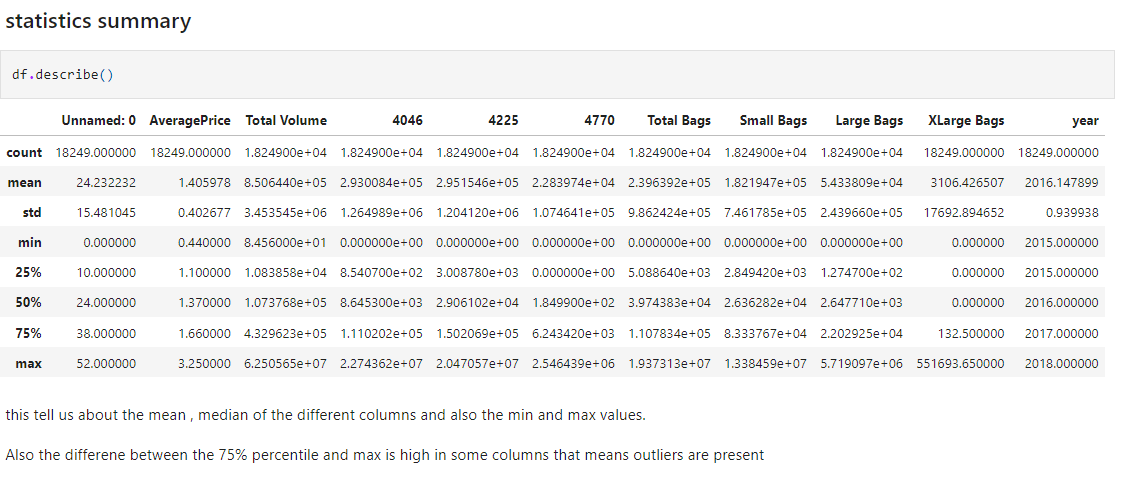




Here, I am checking the shape of the data set as there are 18249 records and 14 columns are present. The Unnamed column is not important for prediction in the price of avocado, so I will remove that column later. Also, most of the columns are of same data type that is float and Date, type and region are of object data type.

**Checking the Null values:**



Above I am checking the null values with the help of seaborn as it shows only one color i.e pink, which means that in our Above statistics data show that their multiple outliers mostly in XL Bags There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction

**Exploratory Data Analysis**

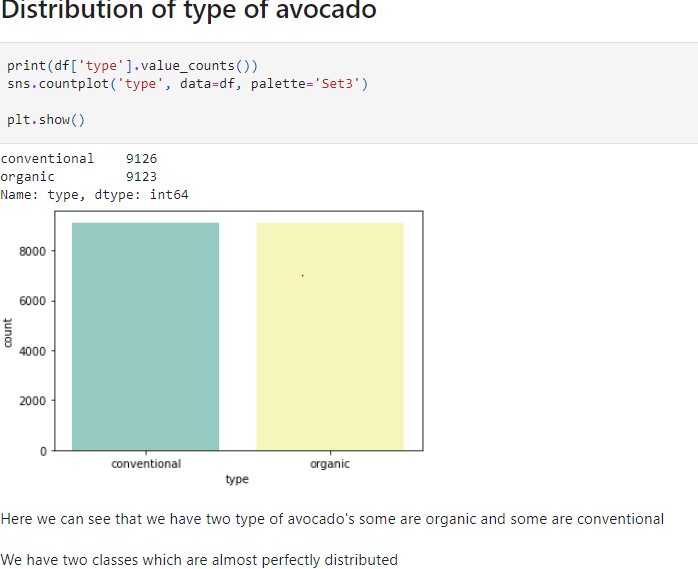
* Also, number of rows in each column are same, means there are no null values in the data set.
* Also, the mean and 50%value of most of the column are same and the STD and mean are very close to each other.
* Most of the column statistics data are near to 0 values.
* By checking the difference between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

**Data visualization:**

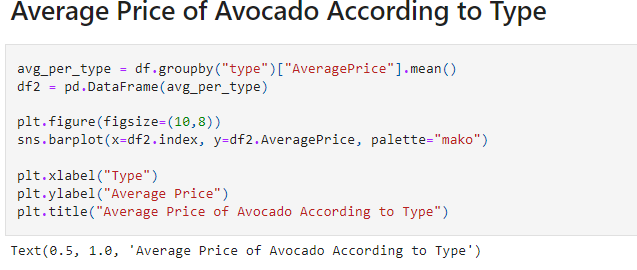
In this portion we can plot different graph using different columns and try to visualize the data using matplotlib and seaborn library.

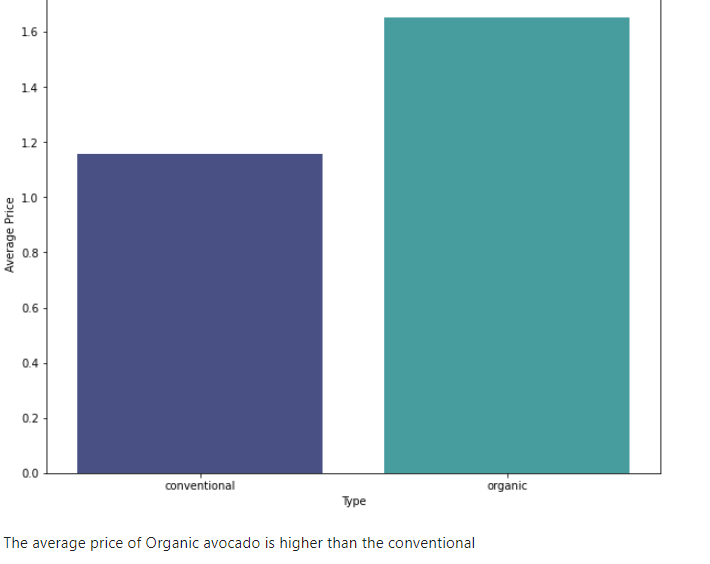
We use different graph include:

* Bar plot
* Count plot
* Line plot and distribution plot
* Histogram and Scatter plot

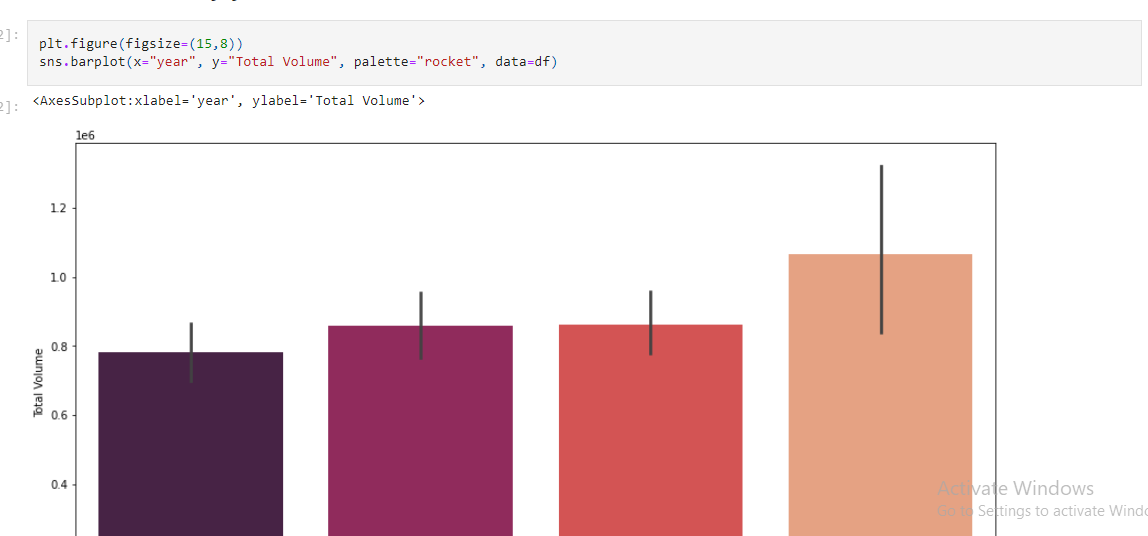


From above graph we came to know that the two type of avocado are present first is Conventional and the other type of avocado are organic.

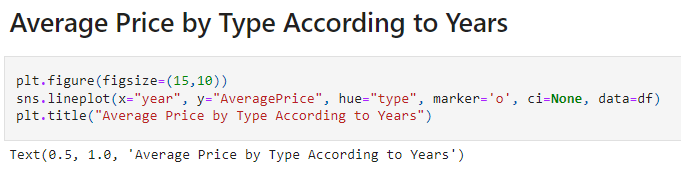


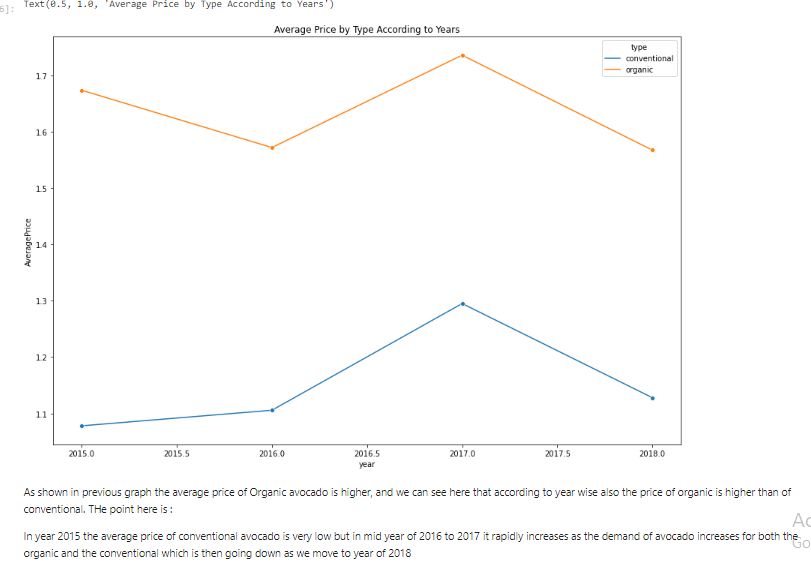


By above image we can see that the According to type of Avocado, the average price of the Organic avocado is higher than the Conventional avocado



Here we can see that in 2018 the volume of avocados sold is very high and in year 2017 and 2018 is almost equal number

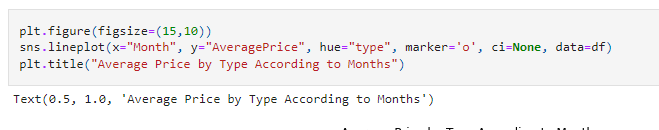


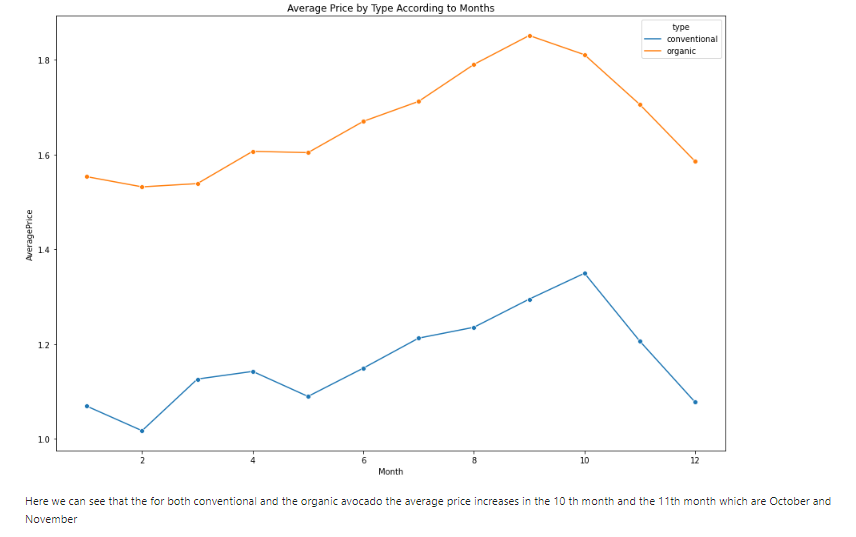


From above we came to know that:

* 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.
* Also the price of Organic Avocado is always higher than the conventional avocado in every year.

## Average Price by Type According to Months



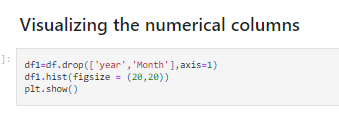


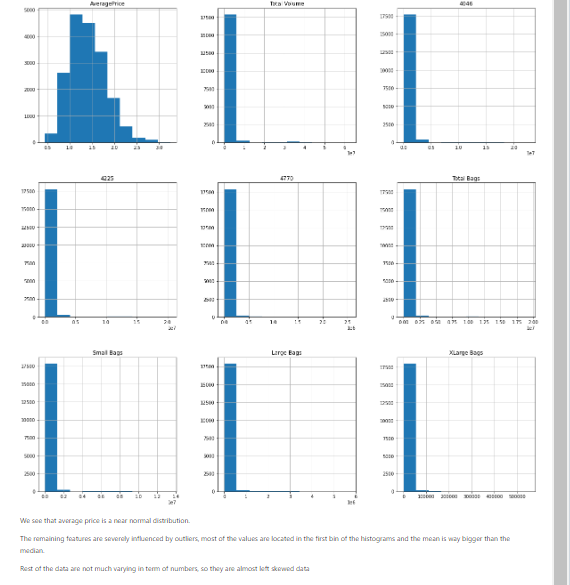
From the above we came to know that:

* 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.
* There is hike between month 8–10 of both type of 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 but then it get decrease and so on.

**Plotting Histogram:**

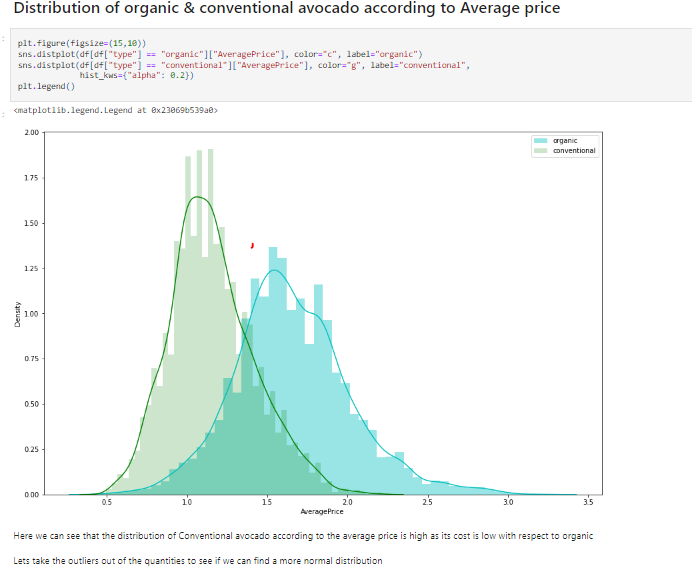
* A **histogram** shows the frequency on the vertical axis and the horizontal axis is another dimension. Usually it has bins, where every bin has a minimum and maximum value. Each bin also has a frequency between x and infinite
* So, in this we can also check whether the graph is right skewed, left skew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



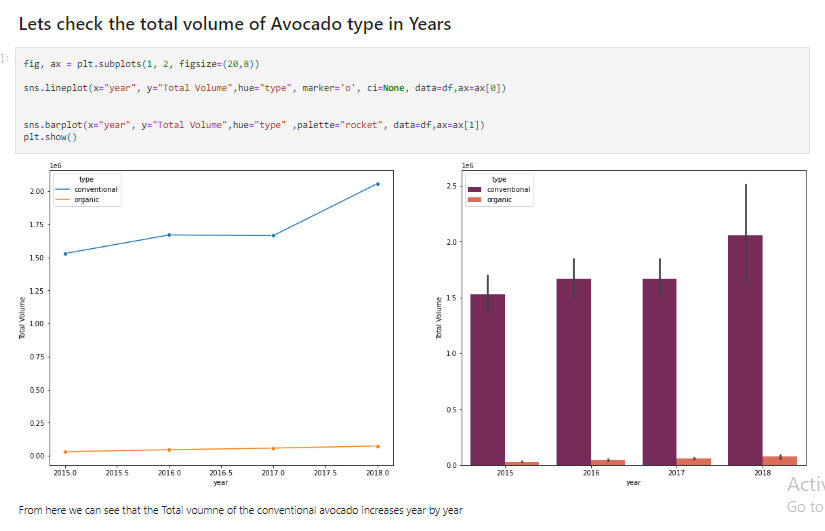


From plotting this histogram, we can see that :-

* 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
* The remaining features are severely influenced by outliers, most of the values are located in the first bin of the histograms and the mean is way bigger than the median.



From above graph we see the Distribution of organic and conventional avocado according to the average price and we can clearly see that the distribution of conventional avocado is very high its density is very high with respect to the organic.

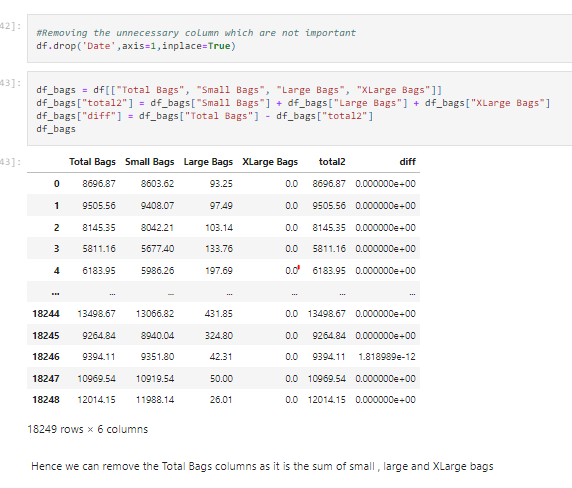


Here I check the Total volume of Avocado in the recent years , i.e. I checked that in recent years which type of avocado produce more and are sold in the market and we can see that the demand of conventional avocado increases year by year and its volume increases very much .

The cost of organic avocado is very high in comparison to the conventional that’s why the middle class and lower class people can’t afford that, that’s why it is the reason the demand of the conventional avocado is very high and its total volume increases every year.

While in case of Organic avocado it is clearly visible that its total volume is increase very slow rate, because the average price of the Organic Avocado is high and people in the United States not prefer that much. The middle class people will give more preference to the conventional avocado.

**Feature Engineering for Model building:**

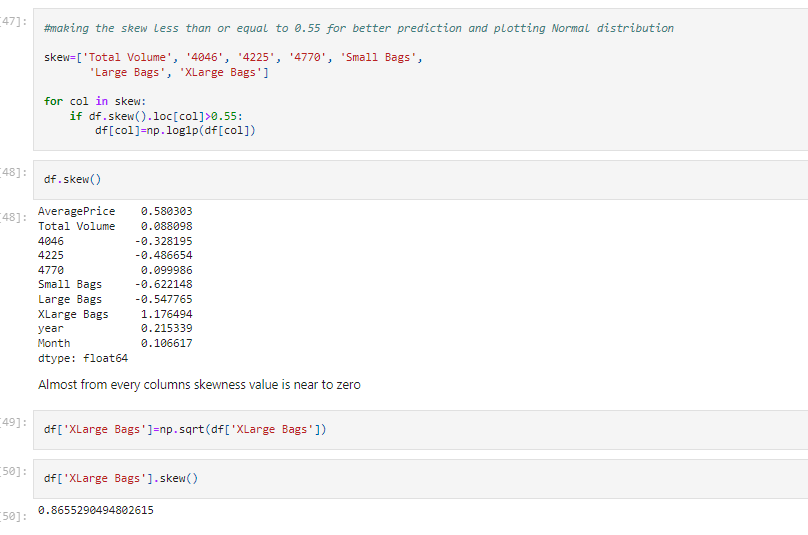




Now I do some feature Engineering to make data more clear for the model training. Here I first drop the date column as it is not important for the future predictions and for the model training.

After that, I check the column with name Bags, as the bags were used to carry the items, so here I check that, there are 3 types of bags are given – Small Bags, Large Bags and the X Large bags and Also One more column is present The Total Bags column, when I add those three columns and calculate the difference with the Total bags , I found out that the Total Bags are the sum of small , large and X Large bags.

And Hence I finally deleted the total bags column as it contain the similar type of data and by deleting it does not affect the main data and further training.



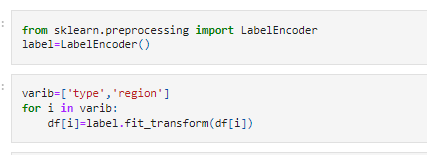
In above image we are first calculating the skew value and some of the column skew value are far from zero so :

* The best skew value for normally distributes is very close to zero, so we are using “log1p” method to make the skew value near to zero
* In the Xlarge Bags 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.
* But in some cases by applying all skewness reduced method but still not skewness is reduced that means it is the nature of that data
* After performing “log1p” method all the columns have skewness value near to zero.

Note: Making the skewness value near to zero will help to get better score.

**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



Converting Region and Type into numeric value by using encoder.

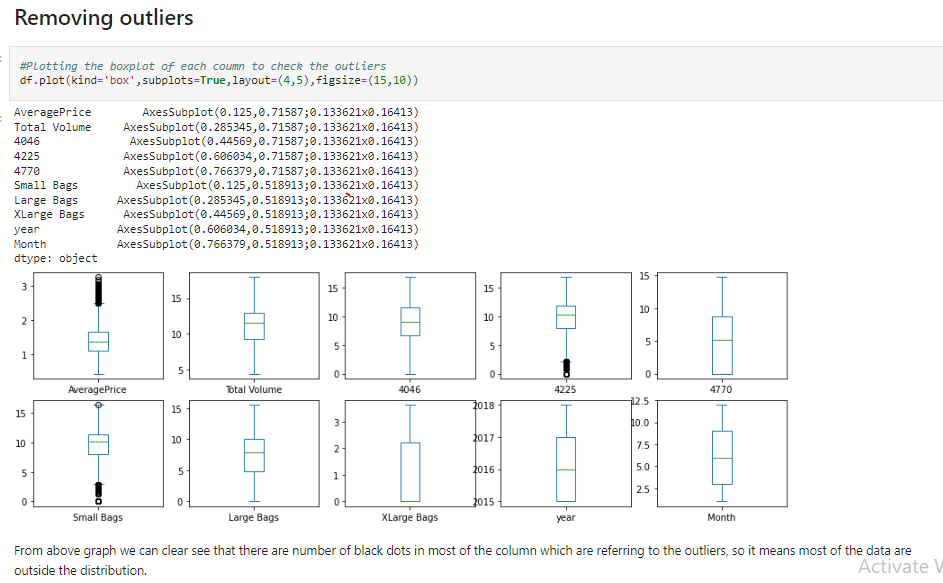
**Outliers:**

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.

Let’s understand by example from above figure we can clear see that the number 3 in the number series is far away from the nature of the series so here “3” is an outlier.

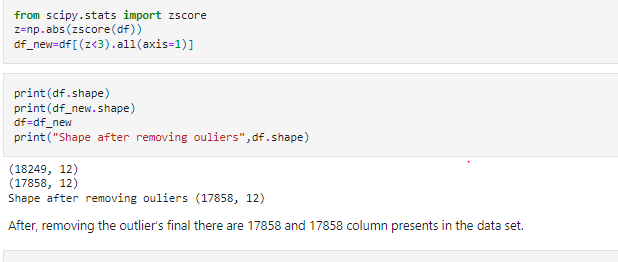
Now that we know outliers can either be a mistake or just variance, how would you decide if they are important or not. Well, it is simple if they are the result of a mistake, then we can ignore them, but if it is just a variance in the data, we would need think a bit further.

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. Here I use the z score method.



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. After, removing the outlier’s final there are 17858 and 12 column presents in the data set.

**Correlation Matrix:**

A **correlation matrix** is simply a table which displays the **correlation**. The measure is best used in variables that demonstrate a linear relationship between each other. Denoted by r, it takes values between -1 and +1.

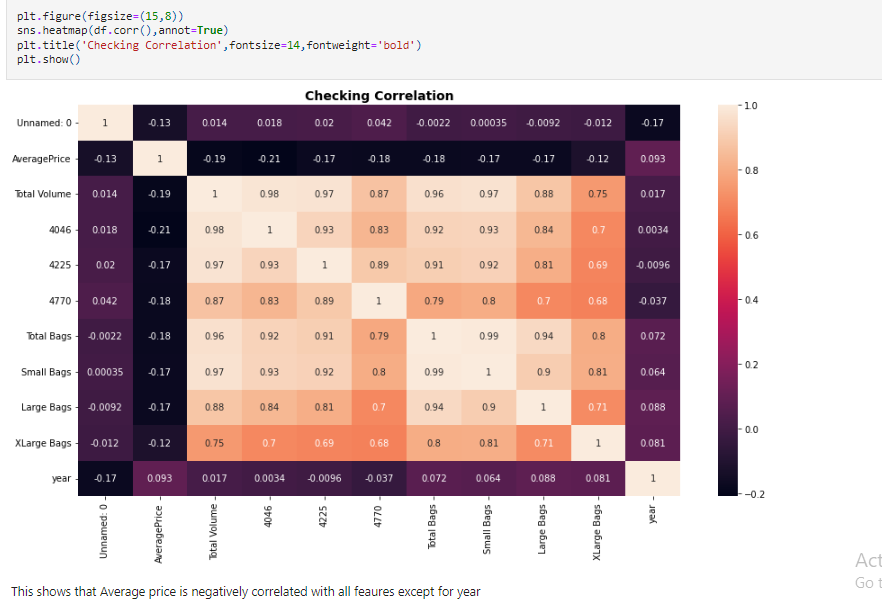
Now I am finding the correlation value of each column, this value is categorized into mainly 2 parts that are:

- Positive correlated value

- Negative correlated value

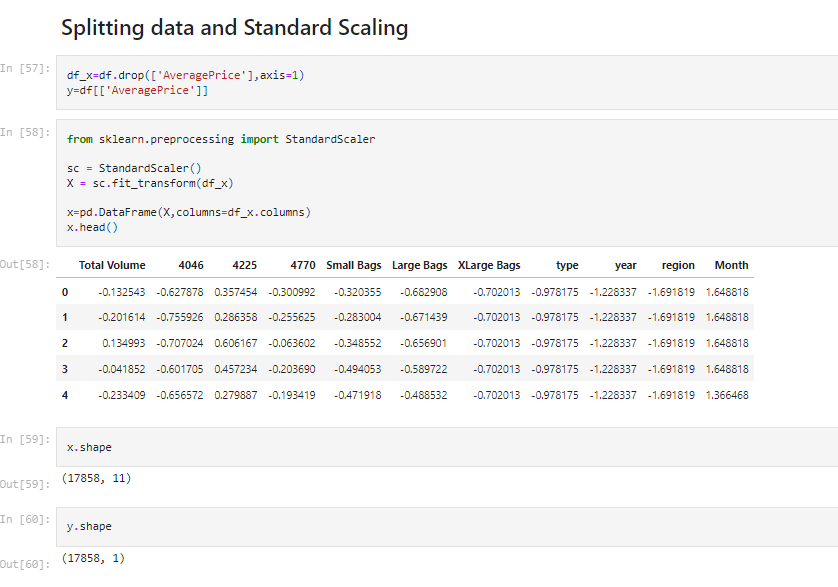
The most the value is positive means that column is much co related and vice versa.

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



Here we can see that Average price is negatively correlated with all features/columns except for year it is positively correlated with that column.

**Splitting and Standard Scalar the data:**



Here I am making two variable x and y where x is having all column except Average Price and y is having only Average price column.

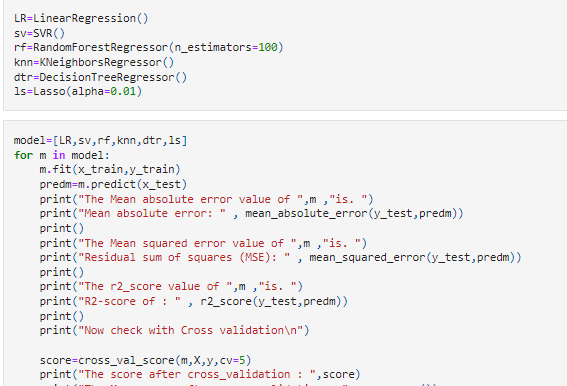
Also, I am using the standard scaling method on x variable.

So finally our data is splitted into x and y where x variable has 17858 rows and 11 columns.

Y variable has the 17858 rows and 1 column.

**Prediction with Average Price:**





Here I first importing all the required necessary libraries from sklearn and different models for the training. After that I used the Max R2 score for selecting the best model, then to make model work more efficiently and correctly the hyperparameter tuning is done.

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.

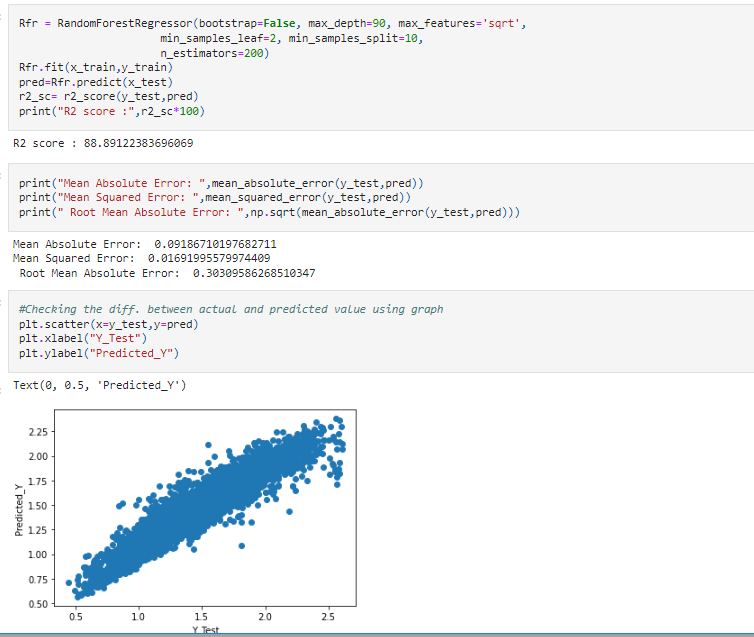
**Hyperparameter Tuning :**

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Two best startegies for hyperparameter tuning are:

* GridSearchCV
* RandomizedSearchCV.



As above we have our best model and to make it more accurate and good for the final training we done the Hyperparameter tuning and here I used the method of RandomizedSearchCV and after this we find the best parameter which give more accurate prediction. After this I trained the model for the one final time with the help of hyper tuned parameters.



**In prediction:**

* I had done this prediction by taking Average price as an output variable which is continuity in nature so that why I’m using the regression technique
* While calculating the best random state the 70 is best state which providing the highest R2 score value for this model.
* After using the RandomizedSeachCV, I can find the best parameters and then I used these parameters for that model.
* After using the best parameters I can get the best R2 score and the model is RandomForestRegressor.
* There are following matrices which I find, and which are providing the best score.
* I also plot the scatter plot graph and we can see that the actual value and predicted values are very close to each other, so the line is best fit line.

**Observation:**

* Here I take price as y variable for predicting.
* Also, I used the Label Encoder to make the categorical data into numeric data i.e. Region and Sex
* Then I calculated the Max R2 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 using the zscore method
* As year is most negative correlated column among all columns
* The demand of Conventional Avocado increases year by year
* In between August to October the price of avocado is much higher as compared to other months
* 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
* At last I save the model as an object, for the model deployment.