IMPLEMENTING MACHINE LEARNING ON AVOCADO DATASET

Definition:

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.

Scientific Name: Persea Americana

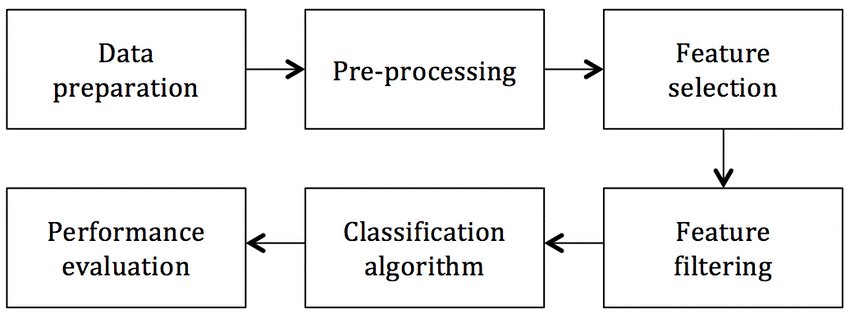
Genus: Persea

Content:

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](http://www.hassavocadoboard.com/retail/volume-and-price-data):

> 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.

Data Source :- <https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects>

**Work Flow:**

* **Problem Definition:**

The goal here is to predict the avearge price which is continuous in nature of different types of avocado and by using the region that in which region they are lying.

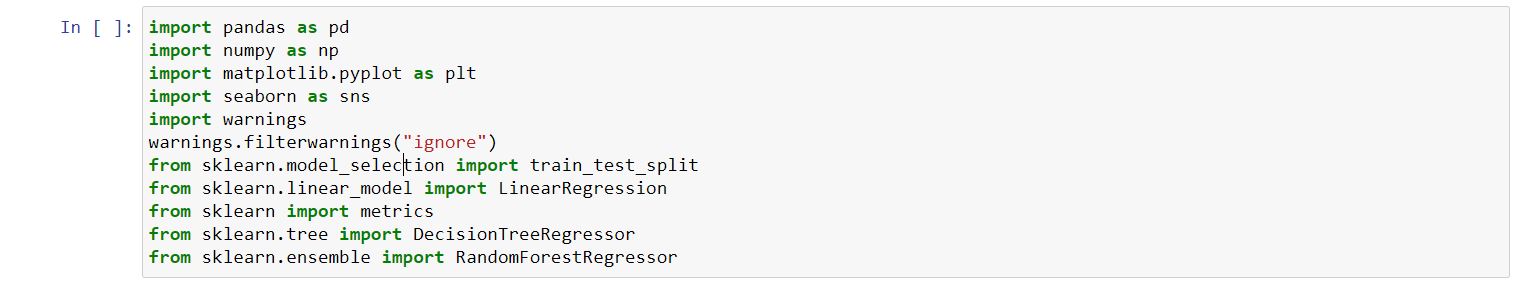
**Inspiration/Label:**

The dataset can be seen in two angles to find the city or region and find the average prices. So, I’m gonna predict the datset in both ways.

**Fields/Columns:**

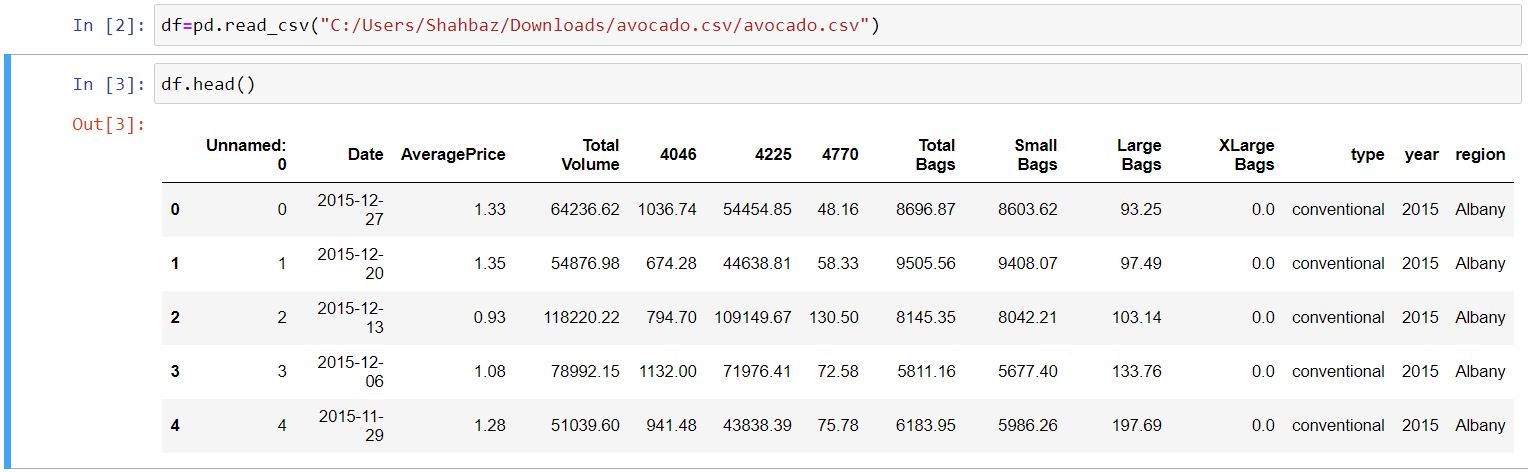
* Date – The date of observation
* Average Price – average price of single avocado
* Year – year
* Type – organic or conventional
* Region – the region or city of observation
* Total Volume – total number of avocados sold
* 4046 – Total number of avocados with PLU 4046 sold
* 4225 – Total number of avocados with PLU 4225 sold
* 4770 – Total number of avocados with PLU 4770 sold
* **Data Analysis**

**Importing library:**



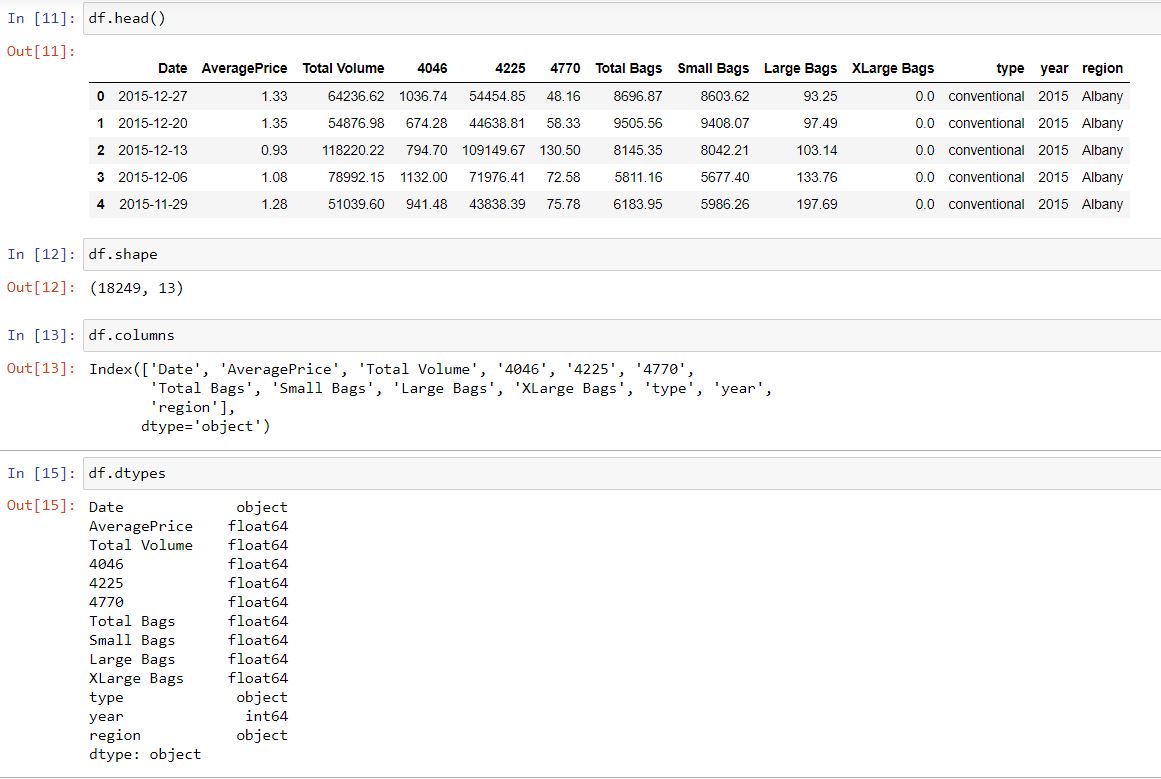
Here I have imported all the libraries which I’ve used in my project that is required for further EDA, Visualization, Prediction and finding all the Matrices. We could find all the importing statement at one place without finding it on whole notebook and can update also.

**Loading the dataset into variable:**



Here I’ve loaded the dataset into variable i.e. “df”. In this dataset most of the columns are float in nature and type & region is of categorical value.

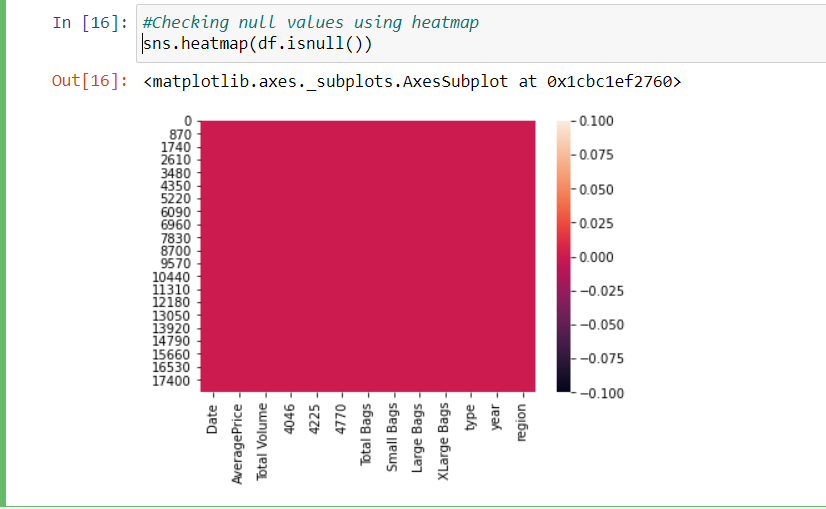
* **Exploratory Data Analysis (EDA)**



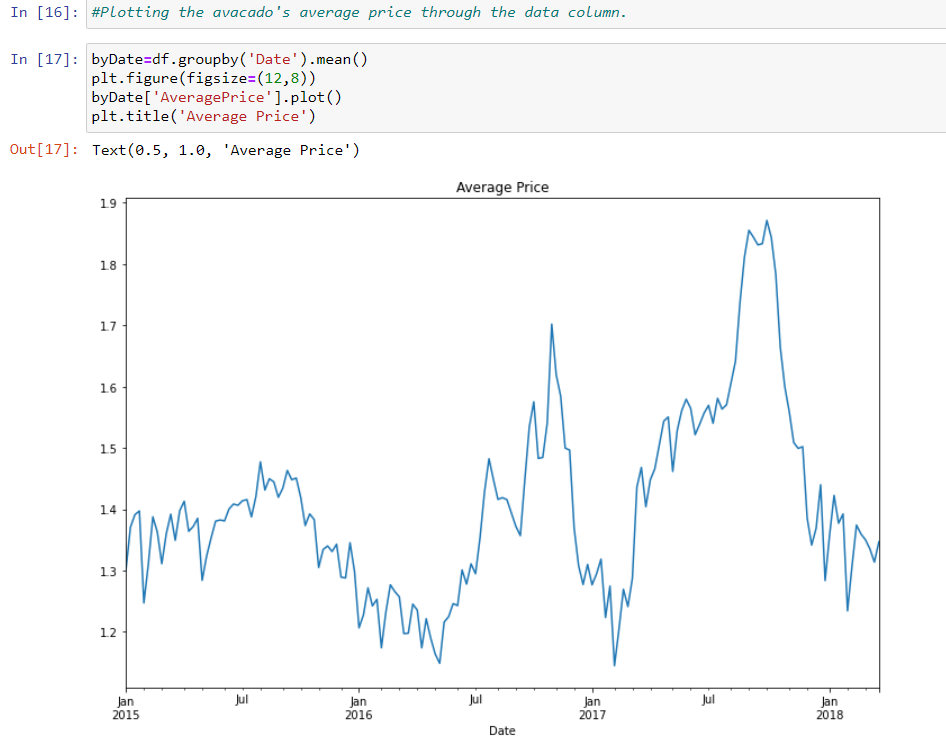
As seen in dataset there is one index column which does not play any role and is not important anywhere hence dropping that column.

Shape of the dataset has 18249 rows and 13 columns after deleting the index column.

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

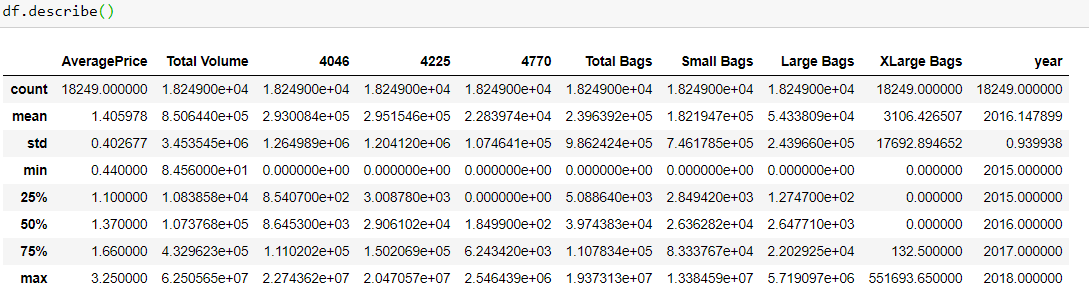


Here, we checked for the null values using heatmap and we observed there are no null values here in dataset since the red color is distributed equally correspond to each column.



Here , I’ve plotted the graph which is showing average prices of the avocado through months with 6 months gaping in between. January to June then June to January respectively. Here, we see from year 2015-2016 there was slight rise in the average price but in mid-2016 it dropped to average price range, and at the end of 2016 the prices goes on peak. In mid-2017 the prices recorded were highest in 3 years span and ultimately dropped at the starting of 2018.

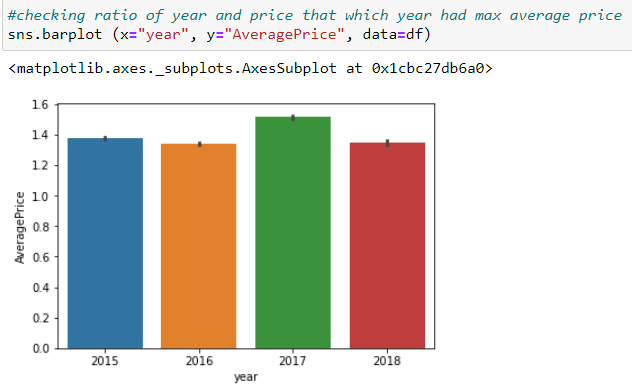
* **Pre-Processing Pipeline**



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 no null values are present.
* Mean and 50% value of most of the column are same and the STD and mean are evry 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.

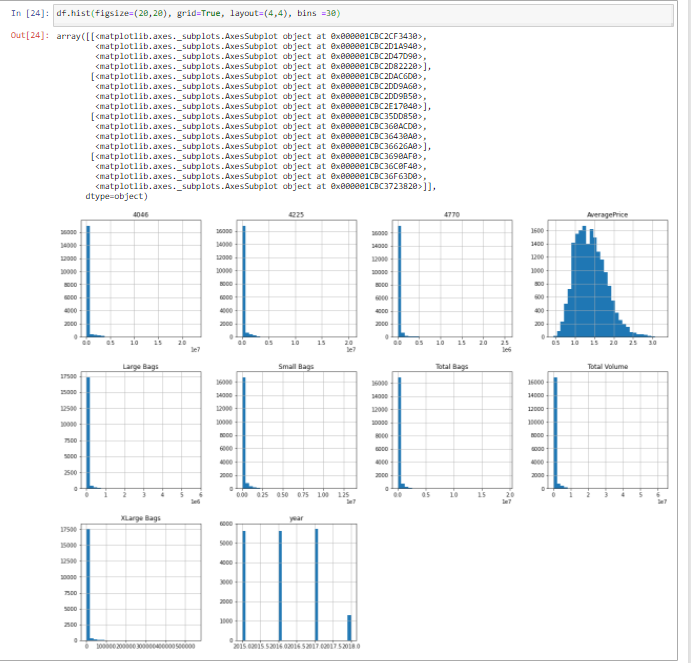
**DATA VISUALISATION**



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

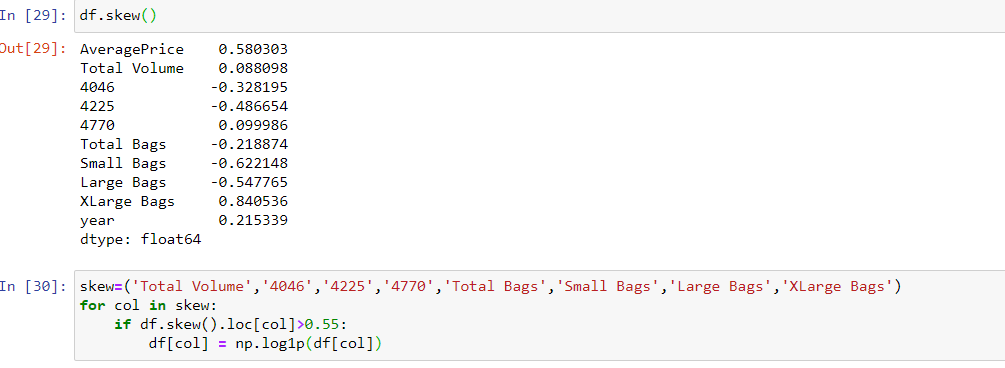
**Plotting Histogram:**

* 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 case we can also check whether the graph is right skewed, left skew or is normally distributed.



From this histogram, we can conclude that:

* Average price column is normally distributing over the histogram.
* Rest of the data are not much varying in terms of numbers, so they are 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.



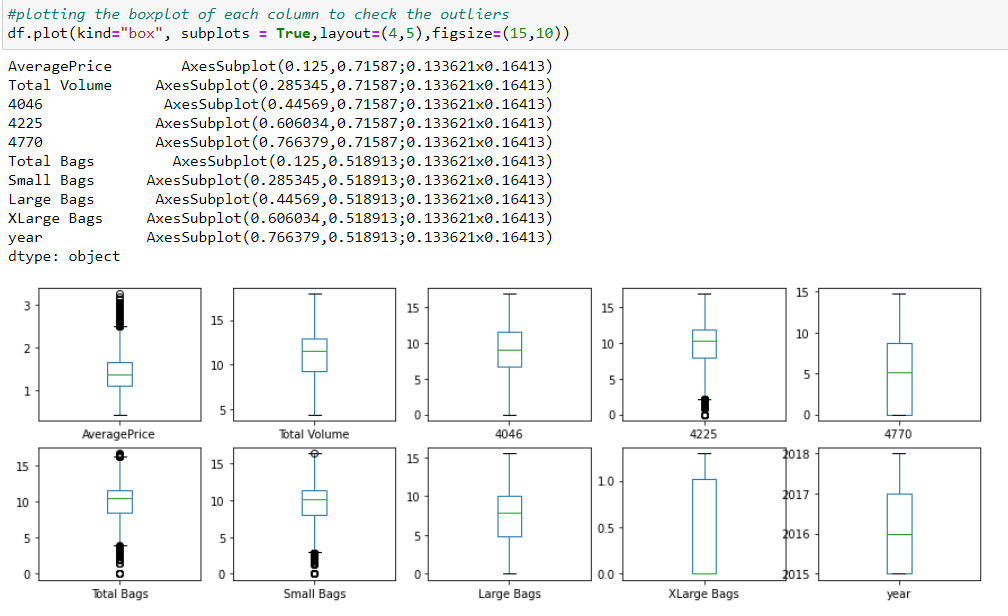
In above image we first calculating the skew values and some of the column skew value are far from zero.

* The best skew value for normally distributes is very close to zero, so we are using “log1p” method to make the skew value near to zero.
* In the last cell, I’m 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.
* Making the skewness value near to zero will help to get better score.
* **Building Machine Learning Models:**

**Outliers:**

An outlier is a data point in a dataset that is distant from all other observations. A data point that lies outside the overall distribution of the datset.

Outliers can either be mistake or just variance, how to 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 to think a bit further.



From the above image we can see that there are number of black dots in most of the columns which are referring to Outliers, which means most of the data is 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.

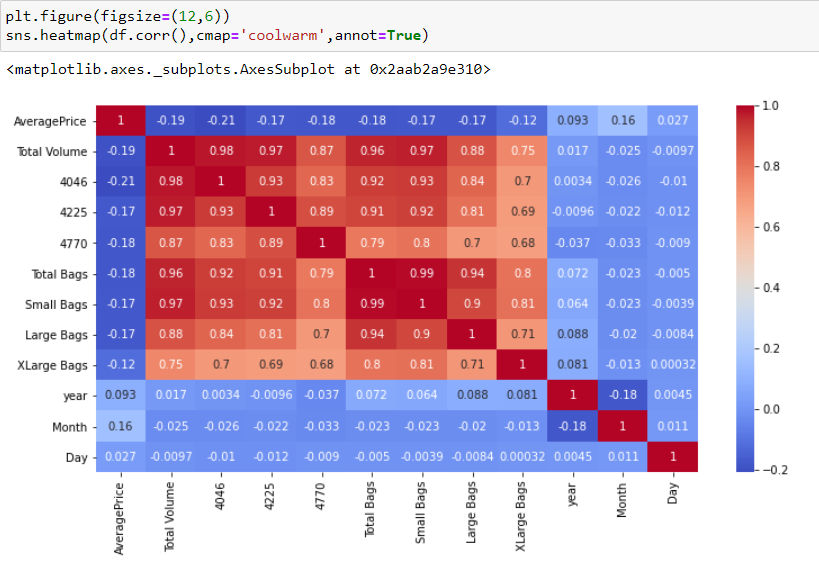
**Correlation Matrix:**

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 & +1.

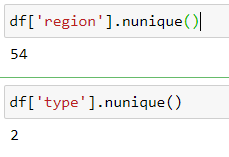
Now I’m finding the correlation value of each column, this value is categorized into mainly 2 parts:

* Positively correlated value.
* Negatively correlated value.

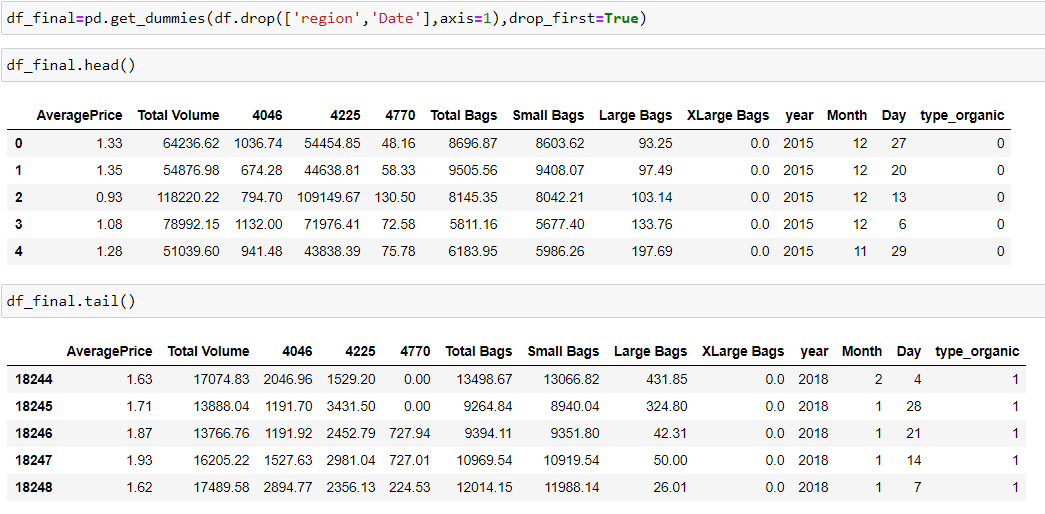
The most the value is positive means that the column is much correlated and vice versa.



Here, we see all the features are not correlated with the Average price column, instead most of them are correlated with each other. This will not give us good model hence will do feature engineering on categorical features: region & type.

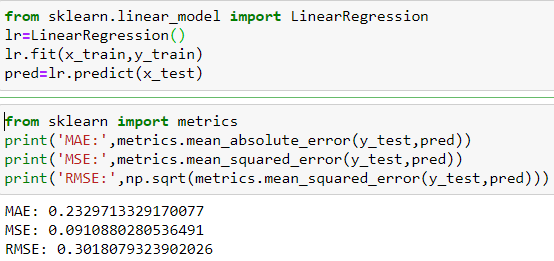


Here , we see there are 54 regions and 2 unique types. It will be easy to transform the type features to dummies but for the region I’ll drop the entire column. I will drop Date feature as well because we already have 3 other columns for Year, Month & Day.

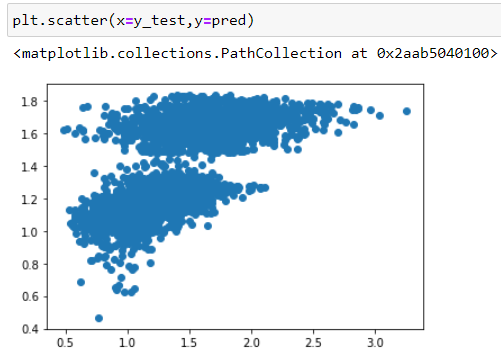


Will apply Linear Regression now since because target variable ‘AveragePrice’ is continuous. We will split up out data into an X array that contains the feature to train on, and a y array with target variable.

**Creating and Training the Model**

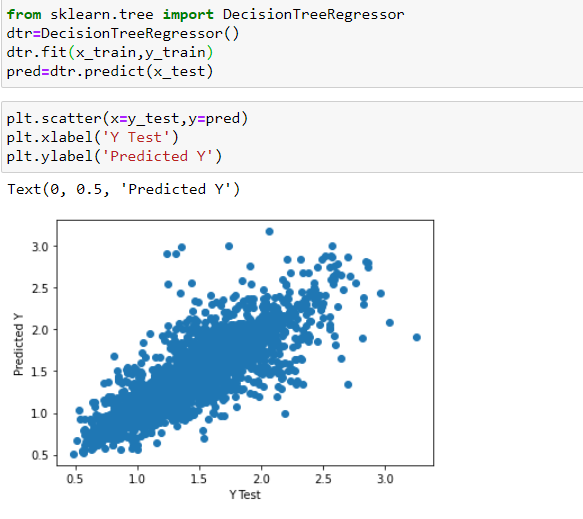


The RMSE is low here so we can conclude that we have a good model, but let’s check to be more sure. Plotting y\_test VS. predictions.



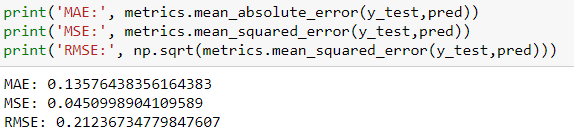
As we see here, it doesn’t have a straight line hence not sure if it is best model for our data.

**Trying DecisionTreeRegressor Model**



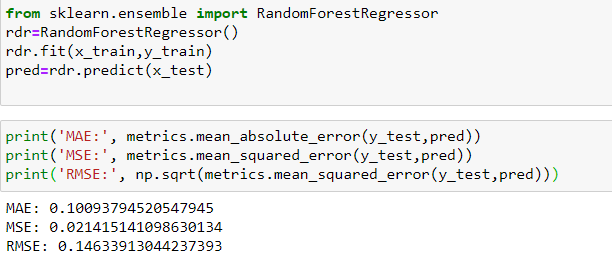
Here we see nearly a straight line hence it is concluded this model is better than previous one ie, Linear regression model.

Now, checking RMSE just to be more sure.

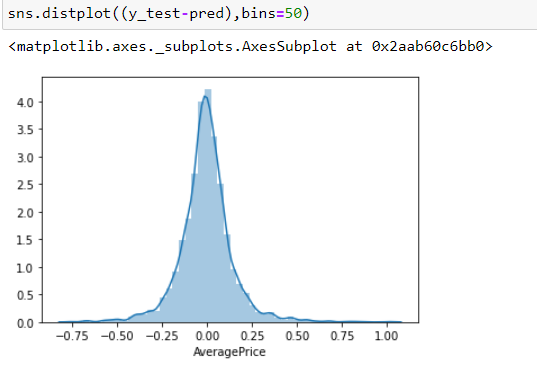


Here we see RMSE is lower than the previous one i.e., Linear Regression model.

Trying one last model here to see if I can improve my predictions for this data which is **RandomForestRegressor.**

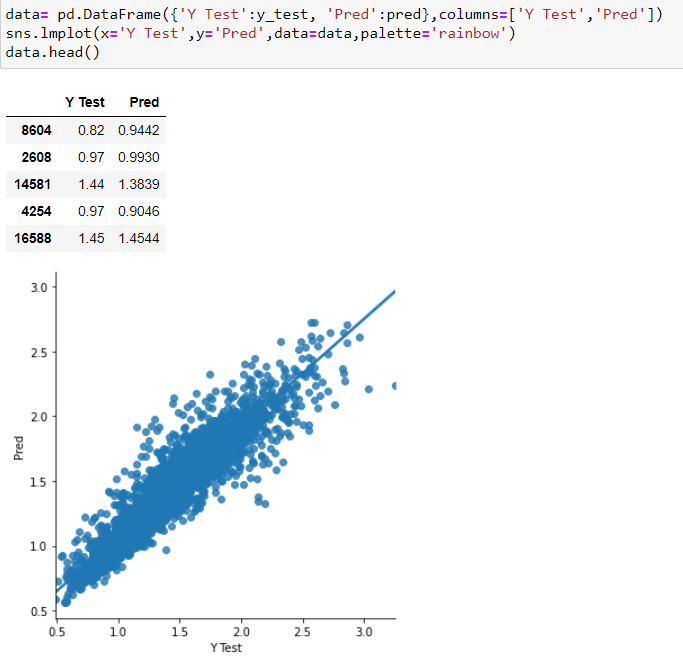


As we see from the image, RMSE is lower than the previous two models, hence we can conclude that RandomForestRegressor is the best model among all.



Notice here that our residuals look normally distributed and hence our choice of Model i.e., **RandomForestRegressor** was correct choice for data.

**Let’s see final Actual Vs Predicted sample.**



* **Final Concluding Remarks**
* **With the help of notebook, I learnt how EDA can be carried out using Pandas and other libraries.**
* **We have seen here how we used the packages like Matplotlib, plotly and seaborn to develop better insights about the data.**
* **I have also seen how preprocessing helps in dealing with missing values and irregularities present in data. Also learnt how to create new features which will in turn help to better predict the survival.**
* **I have seen the impact of columns like type, year/date on Average price increase/decrease rate.**
* **Most important inference drawn from all the analysis is, I got to know the features on which price is highly positively and negatively correlated with.**
* **I came to know through analysis that which model will work with better accuracy with the help of low residual and RMSE scores.**
* **This project also helped in gain insights and how should I go with the flow, which model to choose when and go step by step to attain results with good accuracy. Also, when and where to use Linear, Decision tree and other applicable and require models to fine tune the predictions.**

**This marks the end of this Blog on Avocado Dataset. Hope you like this 😊**