AVOCADO PROJECT

Introduction

This Data was collected from Data Trained Projects Platform. The table beneath addresses week after week 2018 retail checks information for National retail volume (units) and cost. Retail filter information comes straightforwardly from retailers' sales registers based on real retail deals of Hass avocados. Beginning in 2013, the table beneath mirrors an extended, multi-outlet retail informational collection. Multi-outlet detailing incorporates a collection of the accompanying channels: staple, mass, club, medication, dollar and military. The Average Price (of avocados) in the table mirrors a for every unit (per avocado) cost, in any event, when different units (avocados) are sold in sacks. The Product Lookup codes (PLU's) in the table are just for Hass avocados. Different assortments of avocados (for example green skins) are excluded from this table.

**In this blog-post, I will go through the entire interaction of making a machine learning model on the celebrated Avocado\_dataset, which is utilized by numerous individuals everywhere on the world**.

**DATA DESCRIPTION:**

The variables of the dataset are the following:

Categorical: ‘region’, ‘type’ Date: ‘Date’ Numerical: ‘Unnamed: 0’,’Total Volume’, ‘4046’, ‘4225’, ‘4770’, ‘Total Bags’, ‘Small Bags’, ‘Large Bags’,’ XLarge Bags’, ’Year’ Target : ‘Average Price’ The unclear numerical variables terminology is explained in the next section:

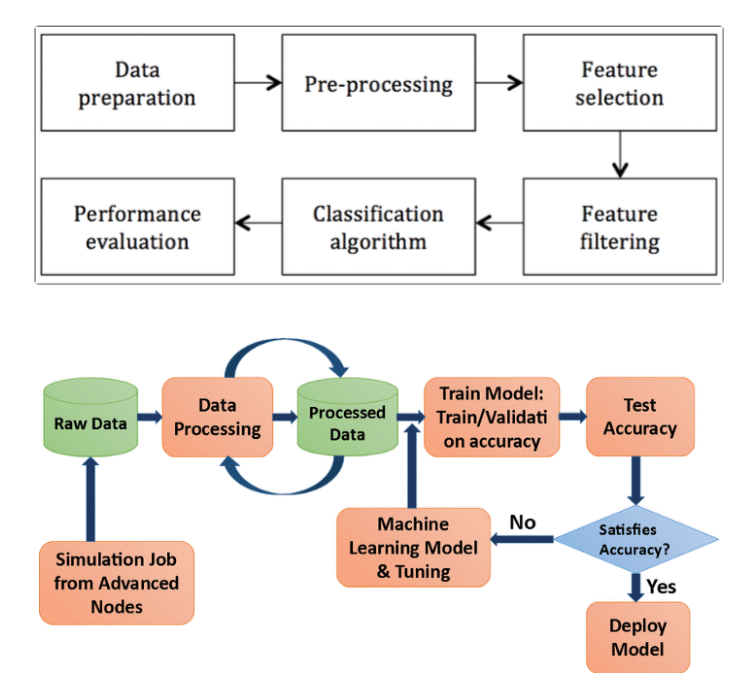
* ‘Unnamed: 0’ : It’s simply a futile index feature that will be removed later
* ’Total Volume’ : Total sales volume of avocados
* ‘4046’ : Total sales volume of Small/Medium Hass Avocado
* ‘4225’ : Total sales volume of Large Hass Avocado
* ‘4770’ : Total sales volume of Extra Large Hass Avocado
* ‘Total Bags’: Total number of Bags sold
* ‘Small Bags’: Total number of Small Bags sold
* Large Bags’: Total number of Large Bags sold
* ‘XLarge Bags’: Total number of XLarge Bags sold
* 'Type' : conventional
* 'Date' : date of sale
* ‘Year’: same as date but we will go through it during Data Cleaning.

PROBLEM STATEMENT:

In this project, we will try to see if we can predict the Avocado’s Average Price based on different features. . The features are different (Total Bags, Date, Type, Year, Region…).

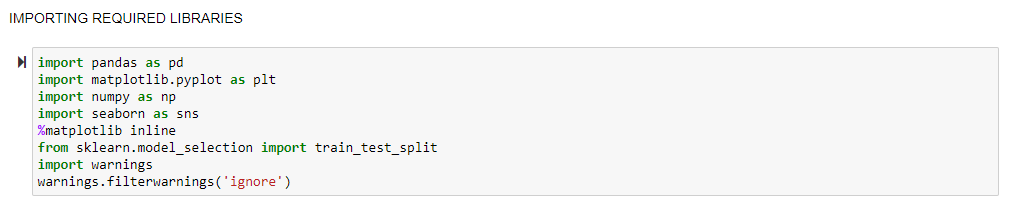
OUR TASK - Is to make a model that can consider the data provided and predict the average price.

## Work Flow:

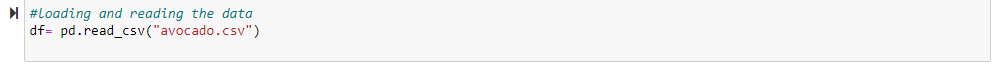


## Exploratory data analysis:

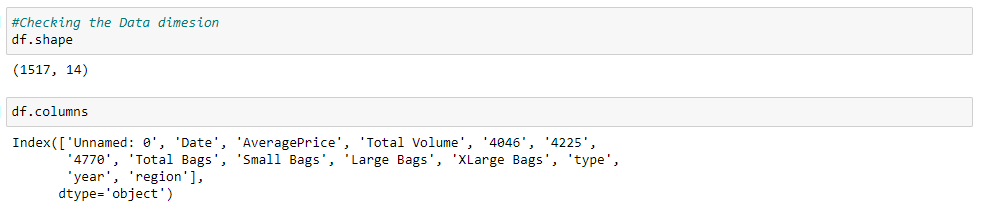
We’ll be using seaborn for visualisation and pandas for data manipulation.



## Loading the Dataset:

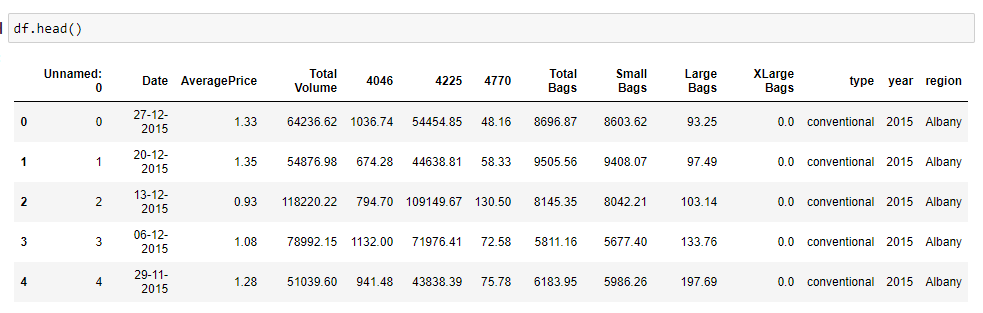
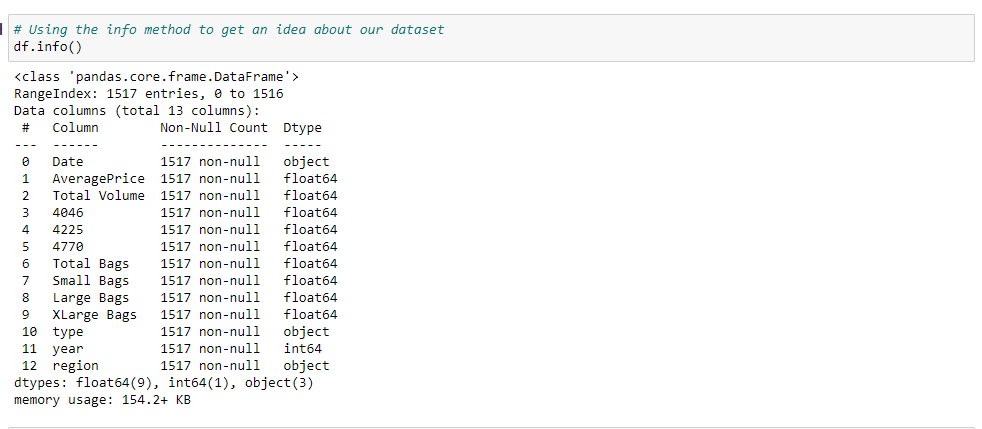


Checking the Data Dimension and the Columns->



As we can see from the above that there are 13 independent variables and 1 target variable i.e. AveragePrice.

## We can look at few top rows using the head function->

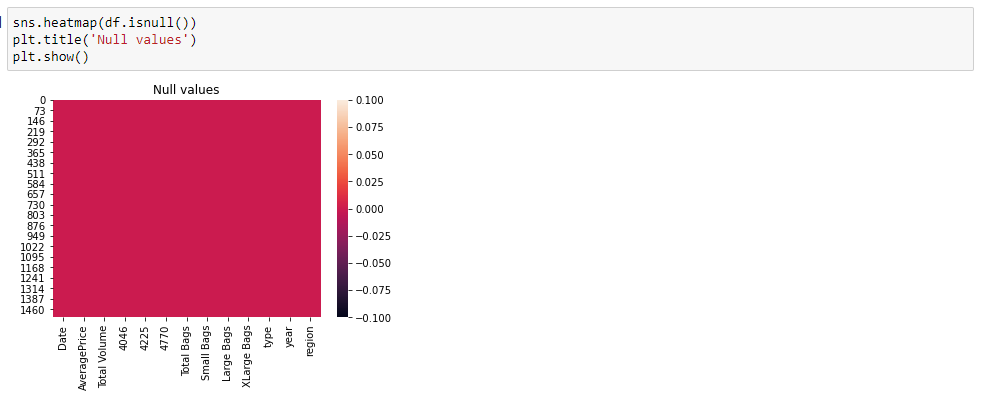


As seen in data set there is one index column which does not play any important role for prediction in the price of avocado, so I am dropping that column.

Also, I am checking the shape of the data set as there are 18249 rows and 13 columns after deleting the index column.

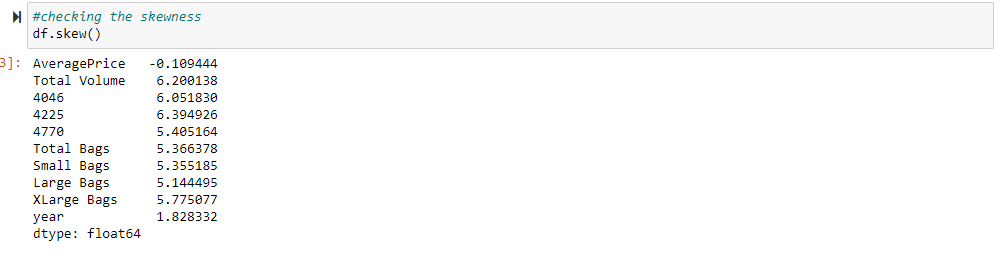
Also, most of the columns are of same data type that is float and Date, type and region is of object data type.

## Let’s check for null values:



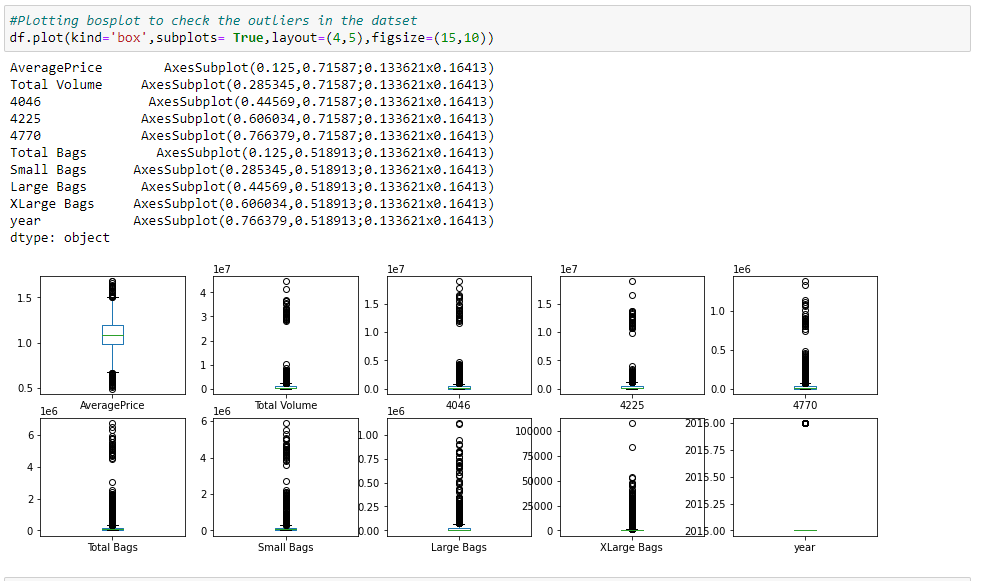
Well at first we can see that we are lucky, we don't have any missing values (18249 complete information) and 13 columns. Presently we should do some Feature Engineering on the Date Feature so we can have the option to utilize the day and the month segments in building our model later. Earlier in info we have seen that Date is Object type not the date type. We have to change its type to date type.

## Checking for Skewness:

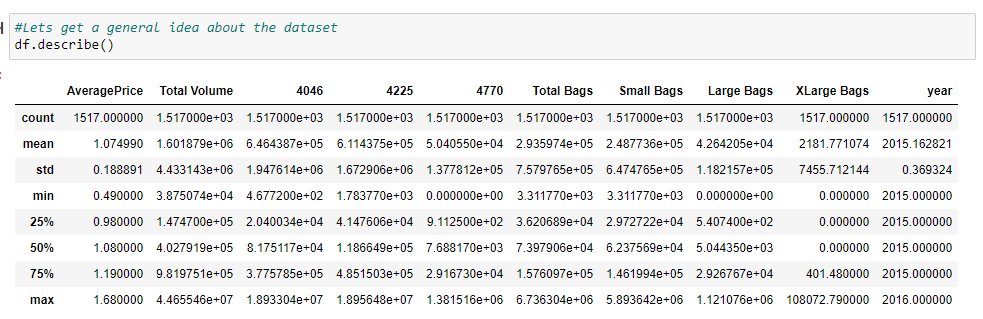


Some kind of skewness can be seen from the above information in the figure.

## Checking for Outliers:



From the above we can see that there are some outlier values in the Dataset which we will deal with later on.

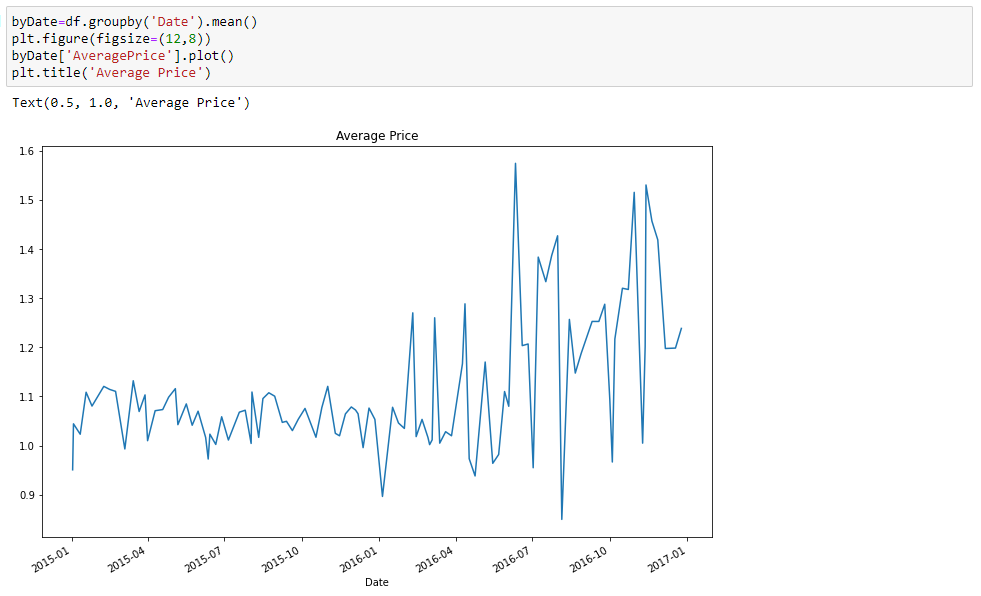


Above statistics data show that their multiple outliers mostly in XLargeBags There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction

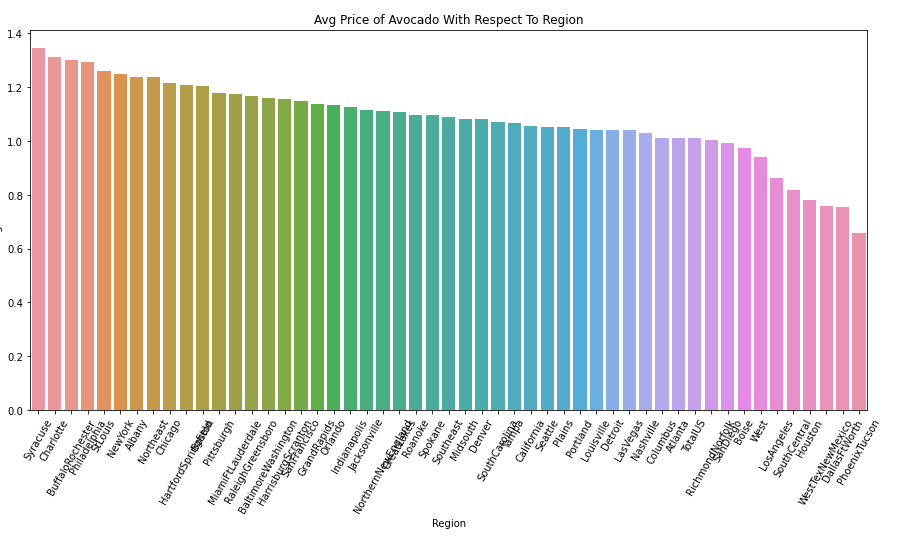
* Also, numbers of rows in each column are same, means there are no null values in the data set.
* Also, the mean and 50%value of most of the column are same and the STD and mean are very close to each other.
* Most of the column statistics data are near to 0 values.
* By checking the difference between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

## Now let’s do some plotting to get a general idea and visualize the data to understand it better:

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

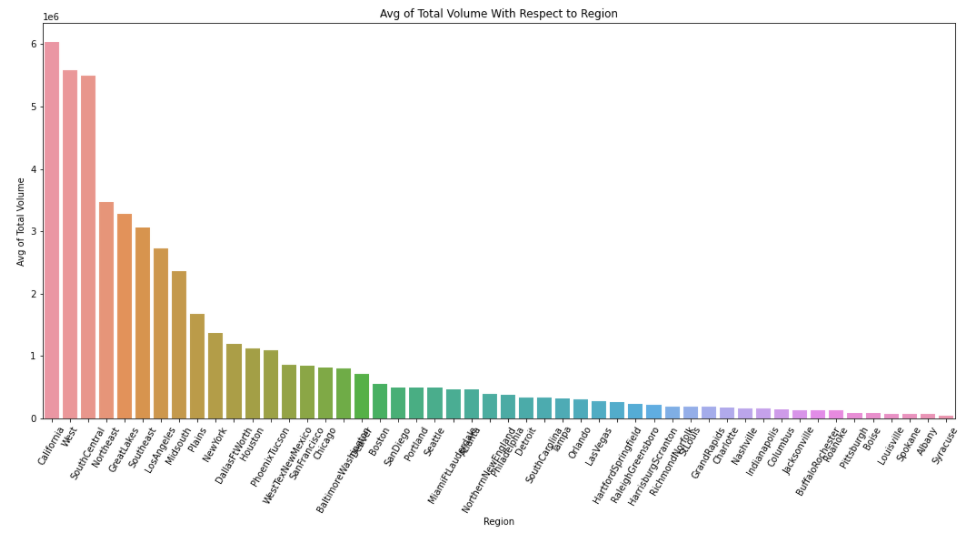


From the above we can see the distribution between AveragePrice with respect to Date. We can clearly see that there was a surge on price from the Year 2016-2017, which is justified because of the high demand by the millennial.



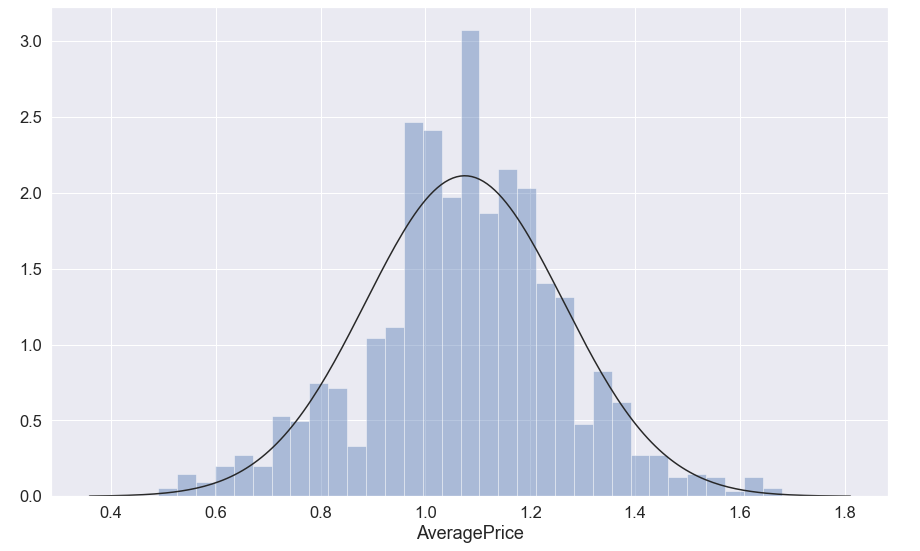
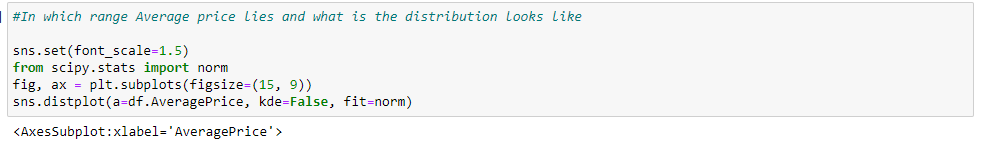
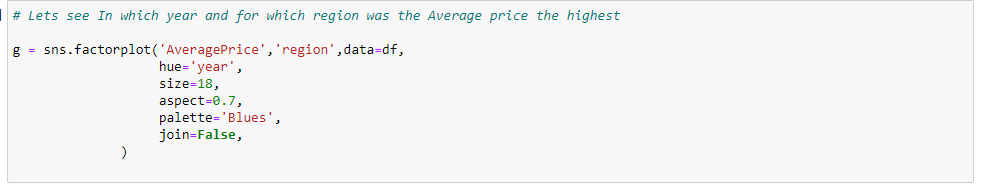
Looks like these regions are where price is very high:

* Syracuse
* Charlotte
* Buffalo Rochester
* Philadelphia
* St Louis
* New York



Looks like these regions are where Consumption is very high:

* California
* West
* South central
* North-east
* Great Lakes
* South East



From the above we can see that the Average Price lies between 1.0 to 1.2 and the distribution curve seems to be normal.

## ****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.

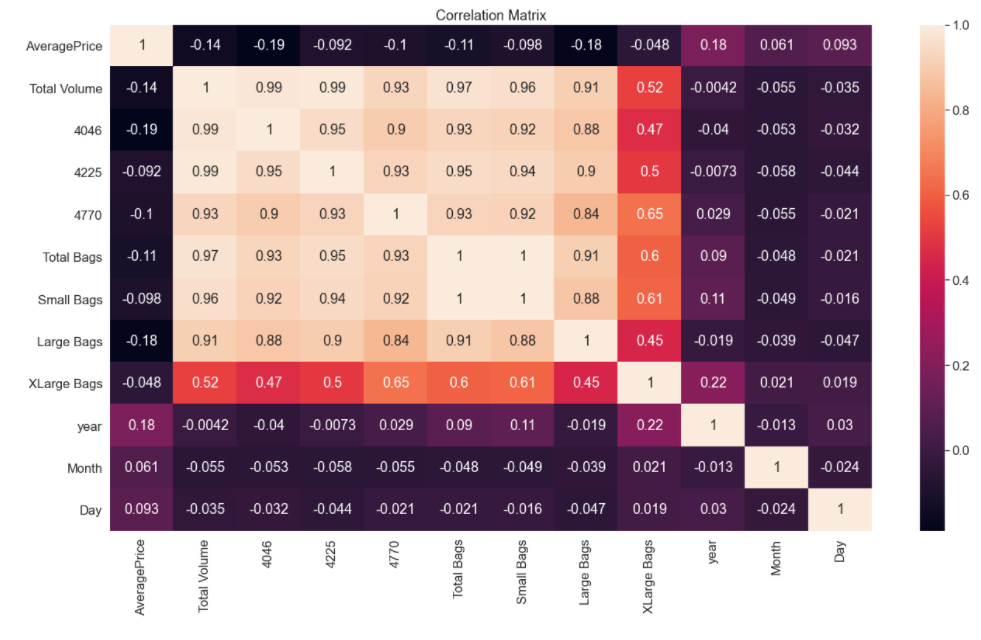
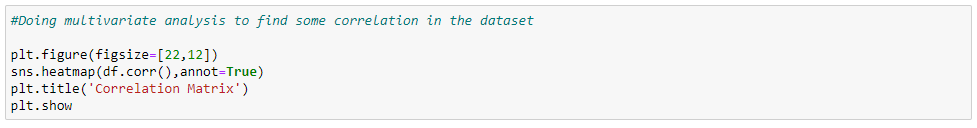
Now I am finding the correlation value of each column, this value is categorized into mainly 2 parts that are:

- Positive correlated value

- Negative correlated value

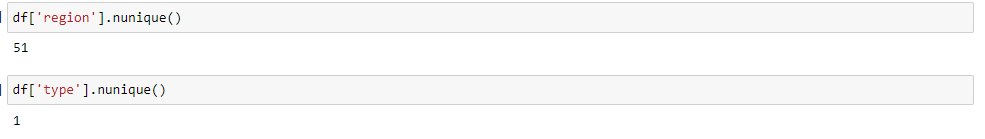
The most the value is positive means that column is much co related and vice versa.

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

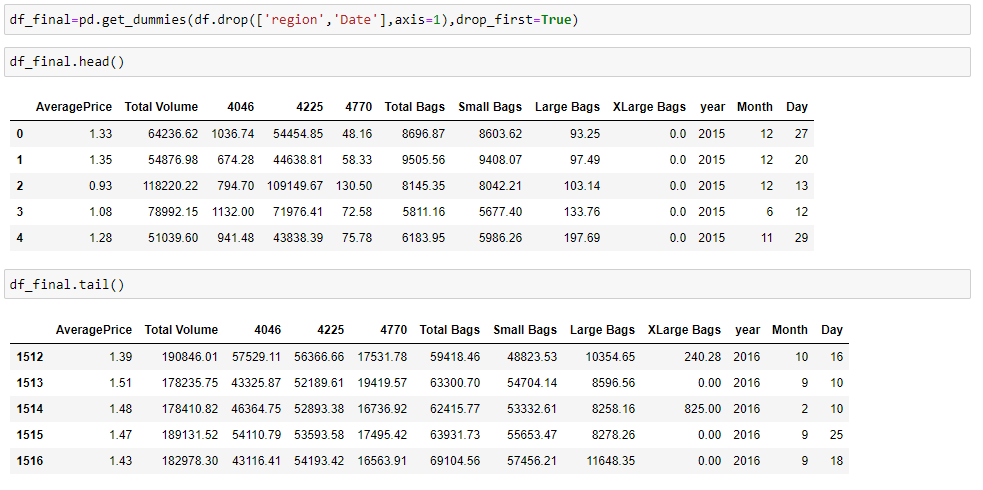


As we can from the heatmap above, every one of the Features is not correlated with the Average Price segment, all things being equal, the majority of them are correlated with one another.

## Let’s do some Feature Engineering on the categorical Features: 'region' and 'type' 🡪

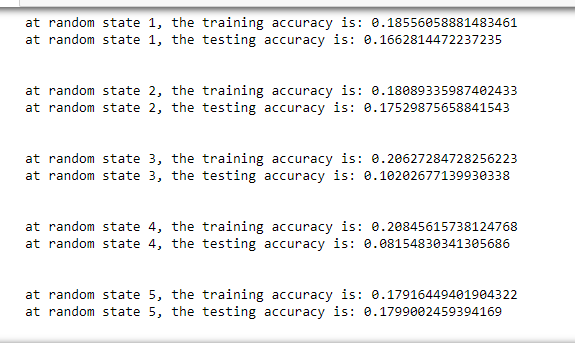
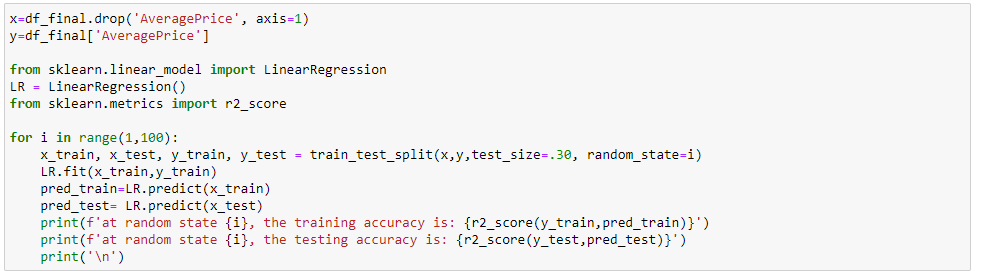


As we can see we have 51 regions and 1 unique type, so we don't really require the Type column, however for the Region it will be somewhat unpredictable so it’s better to drop the whole column as it won't impact much in my model. I will drop the Date Feature too in light of the fact that I as of now have 3 different columns for the Year, Month and Day.



Now our Dataset is ready for Model Building and Random State Selection.

## Selecting the Best Random State:



So from the above we found the best random state for out model is 3.

## MODEL SELECTION:

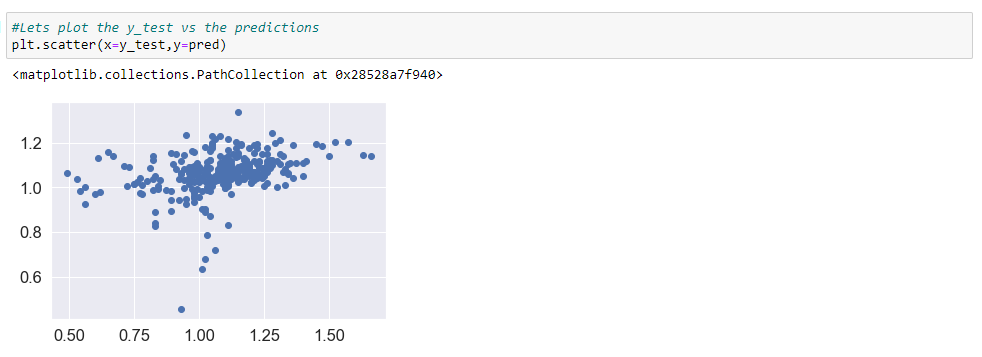
Now our data is ready! Let’s apply our model which is going to be the Linear Regression because our Target variable 'AveragePrice’ is continuous. Let's now begin to train out regression model! We will need to first split up our data into an X array that contains the features to train on, and a y array with the target variable.

FIRST MODEL USING LINEAR REGRESSION:

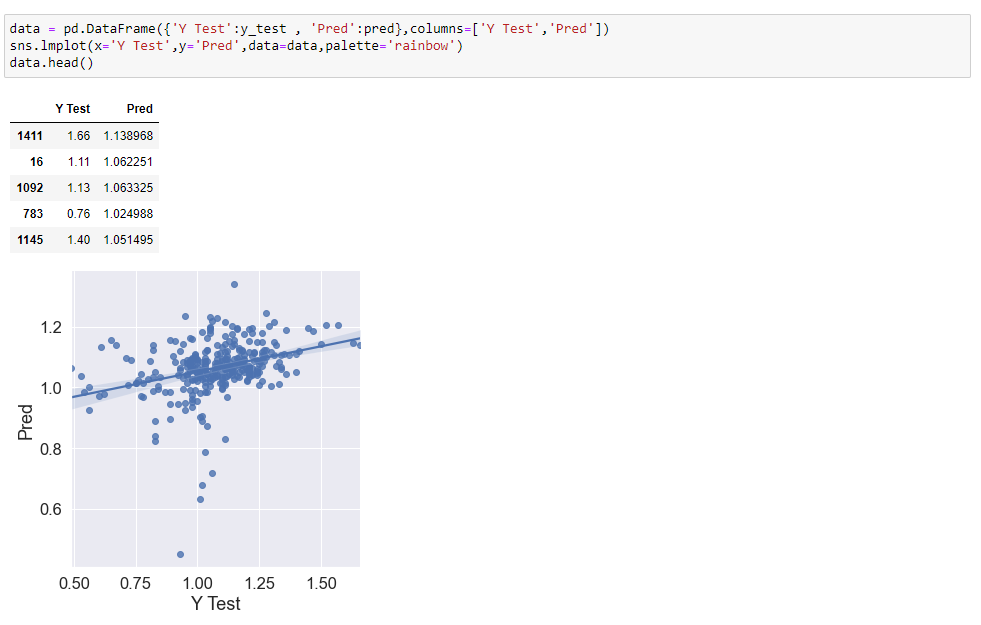


From the above we can see the MAE, MSE and RMSE values.

Let’s plot the predictor values.



As we can see that we kind of have a straight line so I am not 100% sure that this is the best model we can apply on our data.

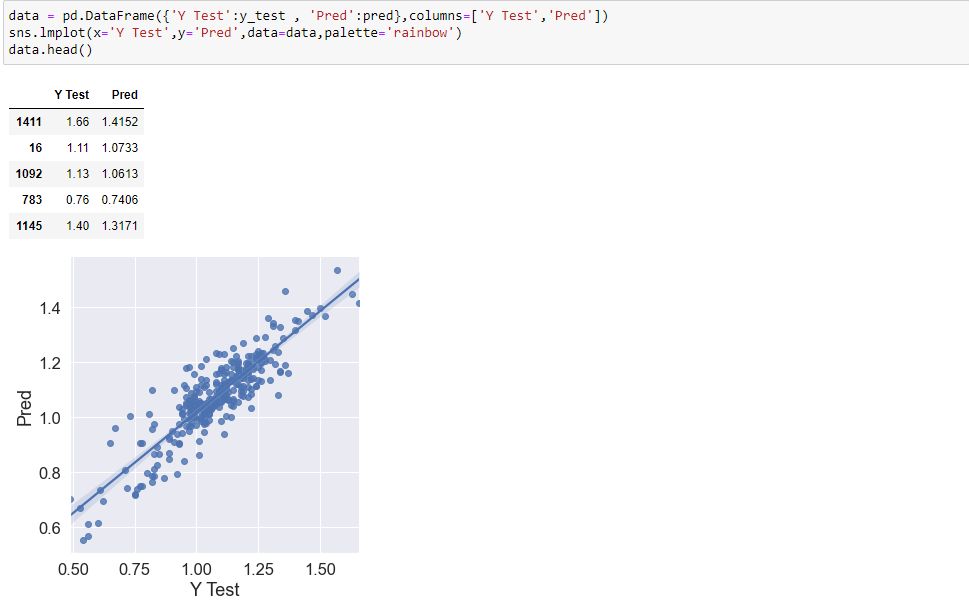


SECOND MODEL USING RANDOM FOREST REGRESSOR:

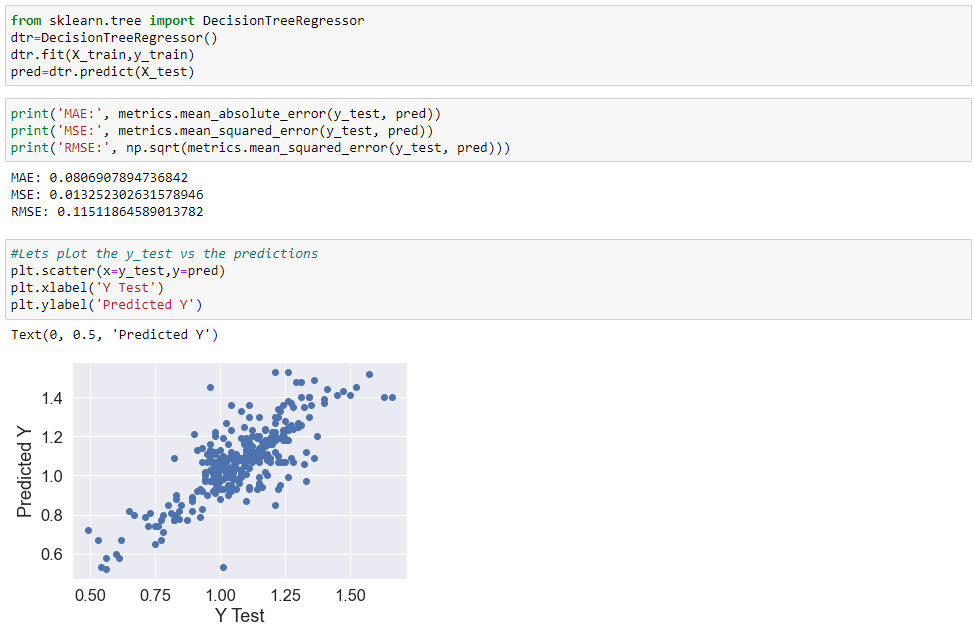


From the above we can see the MAE, MSE and RMSE values.

Let’s plot the Actual vs. Predicted value:

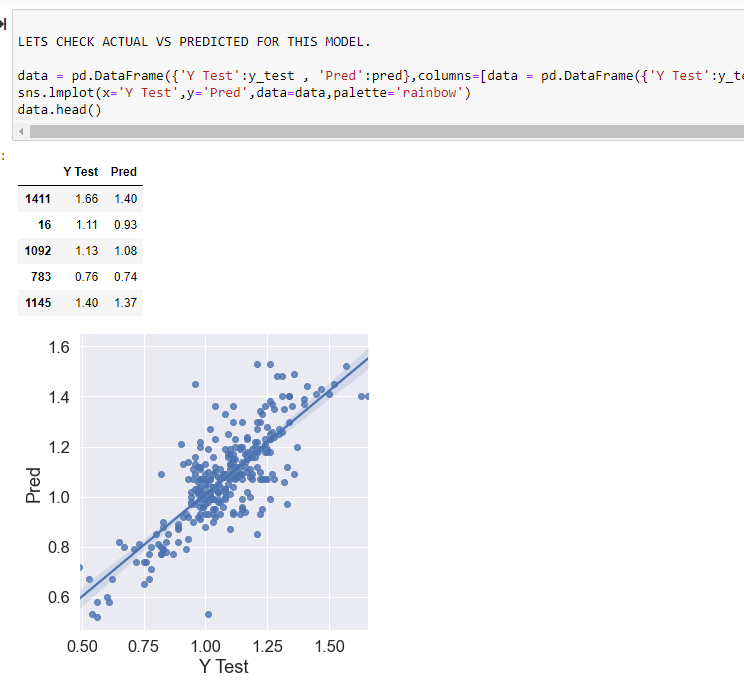


THIRD MODEL USING DECISION TREE REGRESSION:

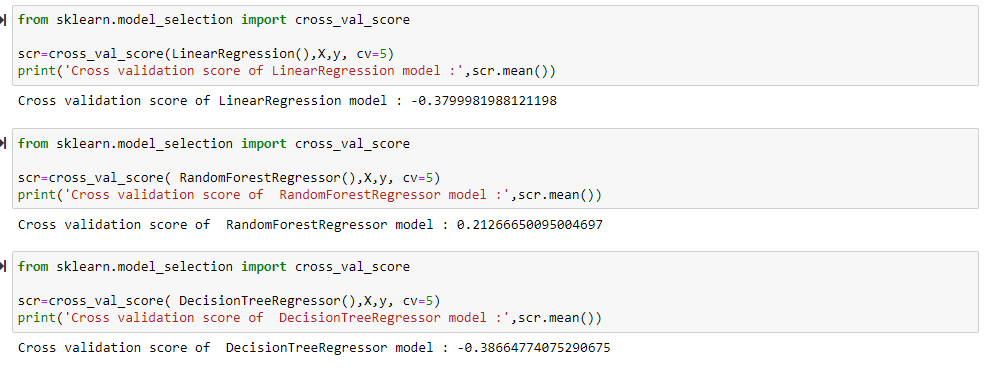


From the above we can see the MAE, MSE and RMSE values.

Let’s plot the Actual vs. Predicted value:



## Let’s Check the Cross Validation Score:



FROM THE ABOVE WE CAN CONCLUDE THAT:

* Random Forest Regressor has the lowest RMSE value which means it was a good fit for our data.
* BY calculating the difference between the actual and predicted value we can see the Random Forest Regressor was the best fit for our data.
* And also by the above graph representation the model predicted by Random forest Regressor has less scattered value from the line.

## GRIDSEARCH CV FOR HYPERPARAMETER TUNING:



## Saving the best model using pkl:



## CONCLUSIONS:

* I have seen the effect of sections like sort, year/date on the Average cost increment/decline rate.
* The main derivation drawn from this examination is, I become acquainted with what are the highlights on which cost is profoundly decidedly and adversely correlated with.
* I came to know through examination which model will be work with better precision with the assistance of low lingering and RMSE scores.
* This task assisted me with acquiring experiences and how I should go with stream, which model to pick first and go bit by bit to accomplish results with great precision. Additionally become more acquainted with where to utilize Linear, Decision Tree and other material and expected models to tweak the forecasts.