**Hass Avocado Dataset**



Avocados are a stone fruit with a creamy texture that grow in warm climates. Their potential health benefits include improving digestion, decreasing risk of depression, and protection against cancer. Avocados are a naturally nutrient-dense food and contain nearly 20 [vitamins](https://www.medicalnewstoday.com/articles/195878.php) and minerals.

Benefits of Avocado

Numerous studies have found that a predominantly plant-based diet that includes foods such as avocados can help to decrease the risk of [obesity](https://www.medicalnewstoday.com/info/obesity/how-much-should-i-weigh.php), [diabetes](https://www.medicalnewstoday.com/info/diabetes/), [heart disease](https://www.medicalnewstoday.com/articles/237191.php), and overall mortality while promoting a healthy complexion and hair, increased energy, and overall lower weight.



* Avocados are nutrient rich
* Healthy for the heart
* Great for vision
* Osteoporosis prevention
* Cancer
* Healthy babies
* Lower risk of depression
* Improved digestion
* Natural detoxification
* Osteoporosis treatment
* Antimicrobial action
* Protection from chronic disease

As sales increase every year, avocados continue to be popular with customers. The great taste and health benefits offered by avocados are flocking to younger consumers. Avocados' versatility makes them a staple ingredient in many kitchens, from sticking them on toast to adding them to salads and rice bowls.

Avocado Popularity

Avocados are still trendy buyers thus prices have grown. Younger shoppers flock to the attractive edges of style and health provided by avocados. Avocados' versatility makes them a staple ingredient in many households, from placing them on toast to adding them to salads and rice bowls.

INSIDE HAB

‘‘*The Hass Avocado Board (HAB) exists to help make avocados America’s most popular fruit.*

*HAB is the only avocado organization that equips the entire global industry for success by collecting, focusing and distributing investments to maintain and expand demand for avocados in the United States.*

*HAB provides the industry with consolidated supply and market data, conducts nutrition research, educates health professionals, and brings people together from all corners of the industry to collectively work towards growth that benefits everyone. The organization also collects and reallocates funds to California and importer associations to benefit specific countries of origin in promoting their avocado brands to customers and consumers across the United States.*’’

Problem Statement.

**Avocado prices**

The data 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.

Some relevant columns in the dataset:

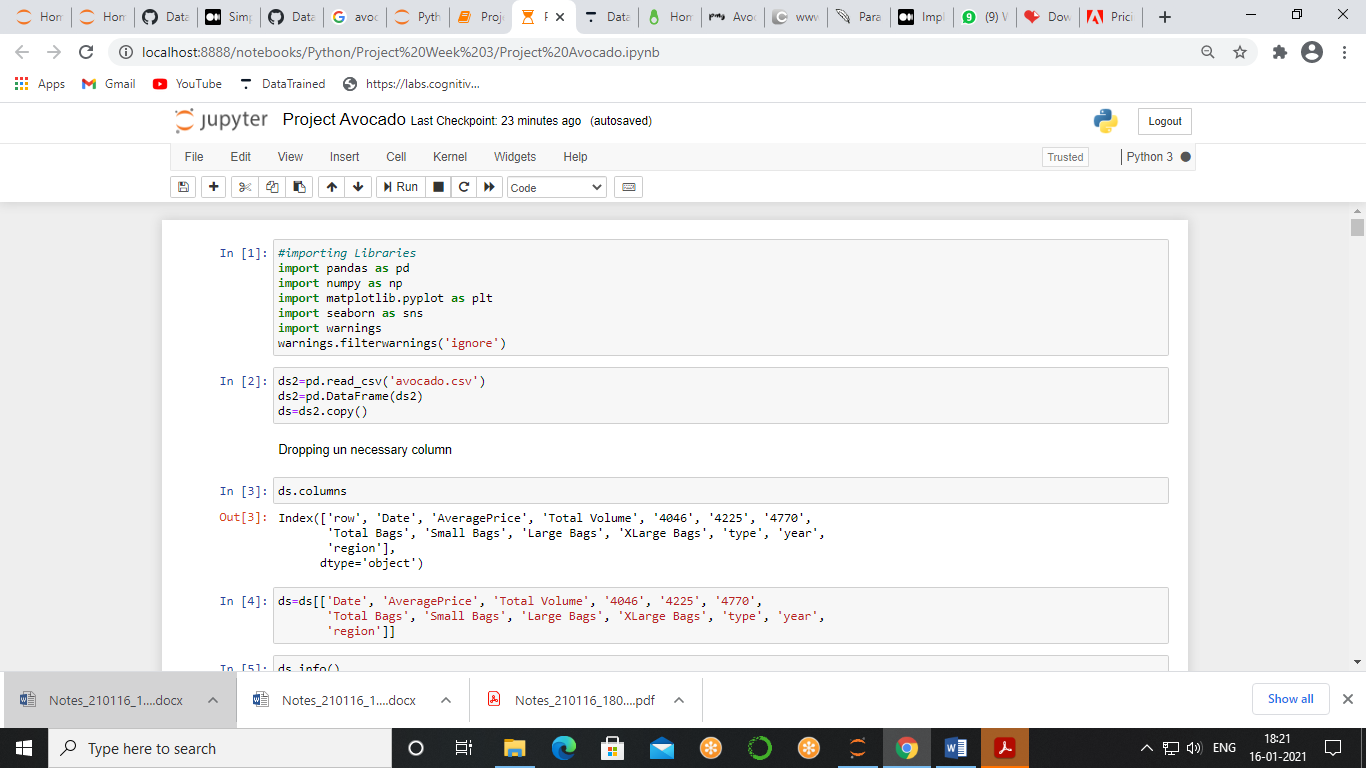
* Date - The date of the observation
* AveragePrice - the average price of a single avocado
* type - conventional or organic
* year - the year
* Region - the city or region of the observation
* Total Volume - Total number of avocados sold
* 4046 - Total number of avocados with PLU 4046 sold
* 4225 -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

Dataset

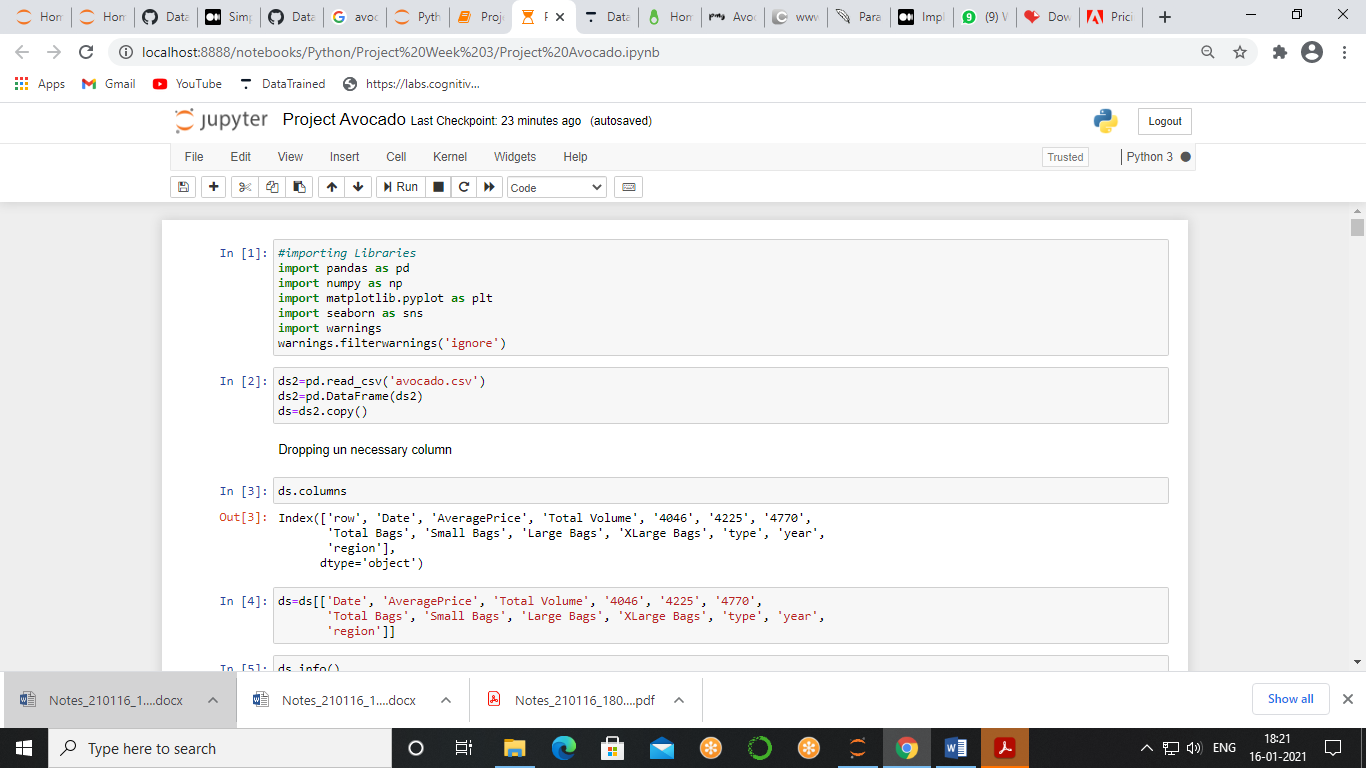
The dataset that is given to us can be seen in two ways by that we mean we can predict two features one is price that will be a regression problem and another is City or region that will be a classiﬁcation problem, so, we will proceed with regression one.

Importing Libraries

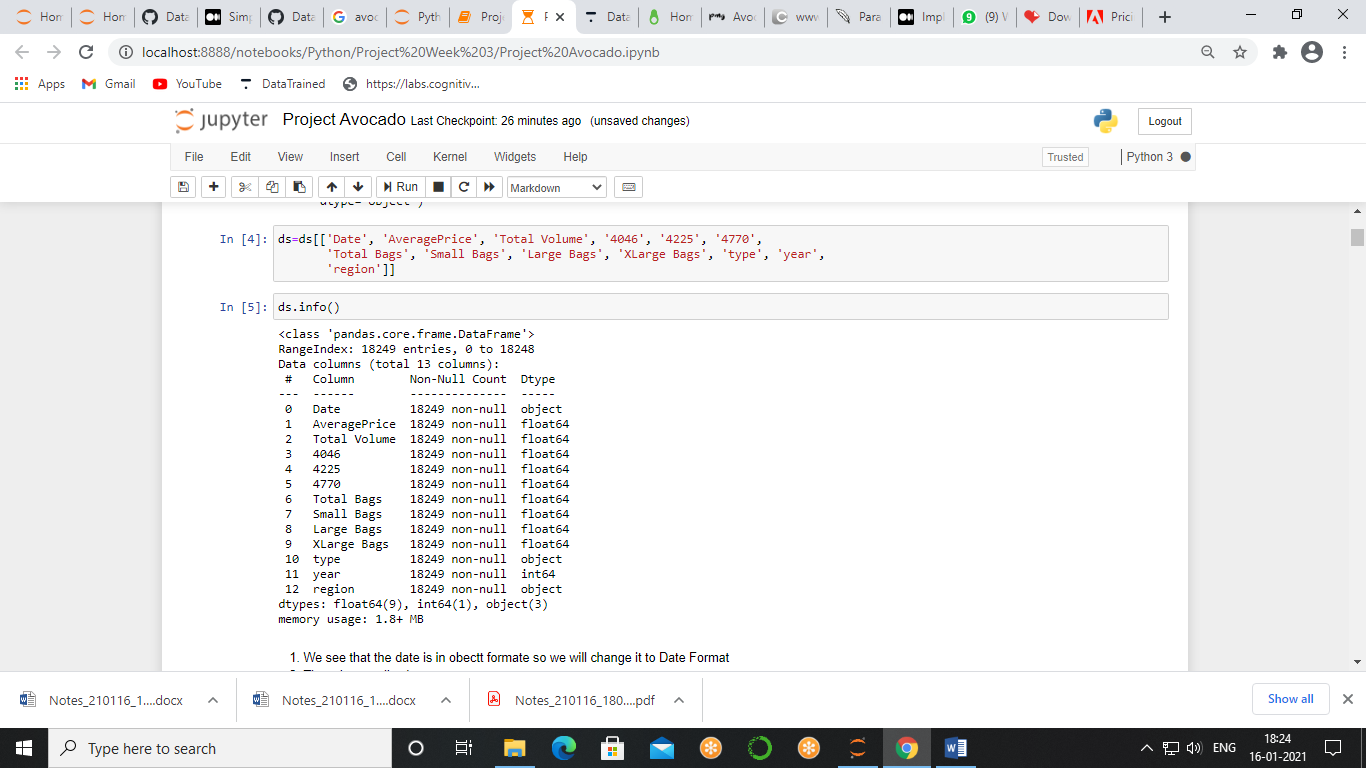
Our ﬁrst task will be important the necessary libraries that we will be using in data processing we might need some other libraries in between we will import them at the time of need currently we are importing three type of libraries 1st is mathematical Numpy, second one is Pandas and third one is visualization library, we are also importing warnings to ignore unnecessary warnings between the process.



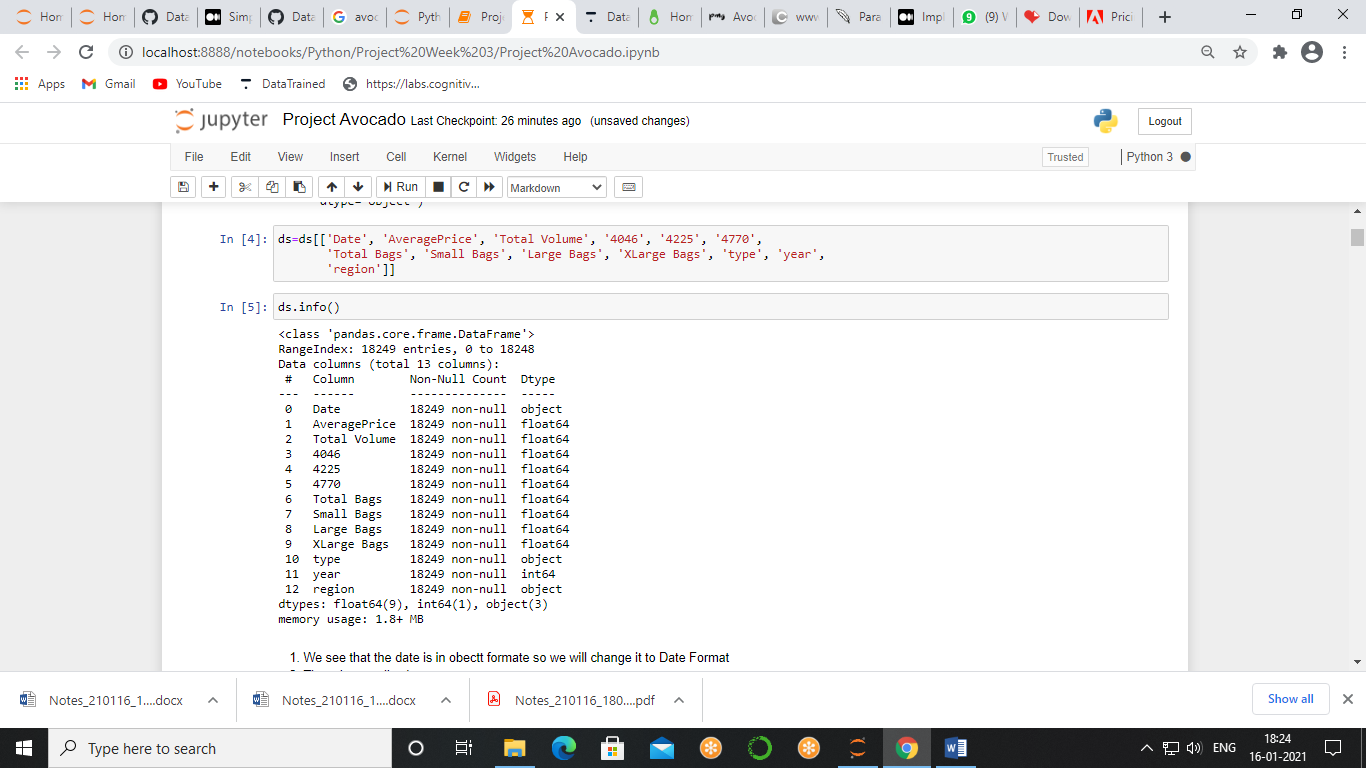
Then we will import our data using csv method and making a copy of a data for the future in case if we need original form of data



When we look at our dataset, we observe that there are 18249 rows and 14 columns and the description of of columns are given below, Now, there is one unnecessary feature which we will not require at any scenario, so it's better to proceed with our work after dropping this column.

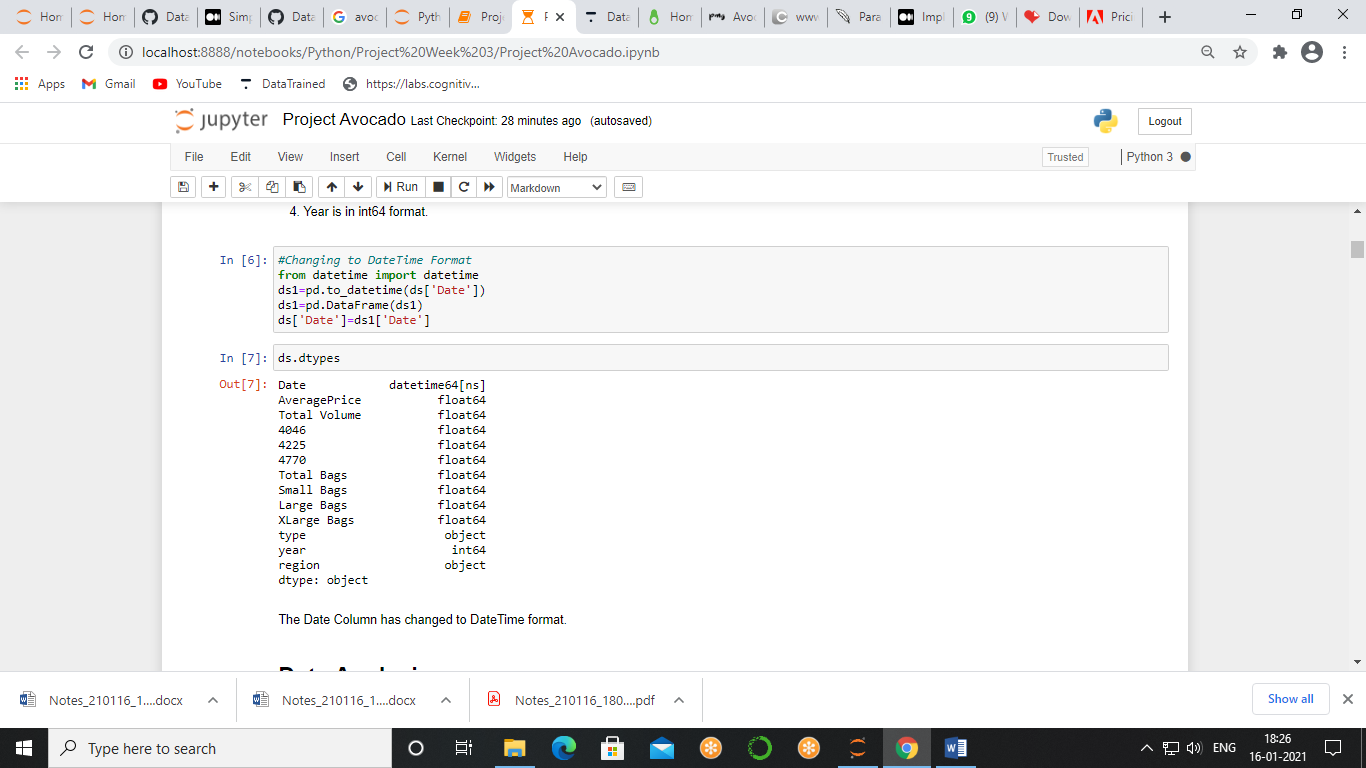


We will now look at basic information of the data that which will give us data types and null values.



After looking at the information table we observe, we see that the date is in object format, so we will change it to Date Format, there is no null values, Type and Region columns are in object format, Year is in int64 format.

To change the date variable into to datetime data type there is one library available which can convert a formatted string into a datetime format, so, it will be easy to us to fetch out day month and year separately.



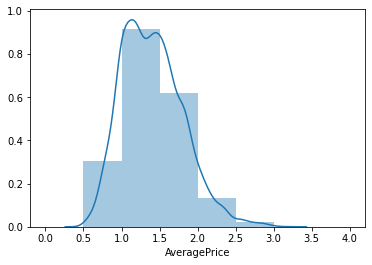
Exploratory data analysis

Exploratory data analysis is the process of analyzing the important features of the data as well as what that I can give to us to achieve our goal, with EDA we can get insight of data and put it in visualization format for our better understanding.

Here we will analyses the distribution of sales data throughout the year as well as for different countries and what are the factors that are affecting the sales and at the end we will make sales strategy based on our observations.

Average Price

We will start by making the distribution curve of our average price that is also our target variable.



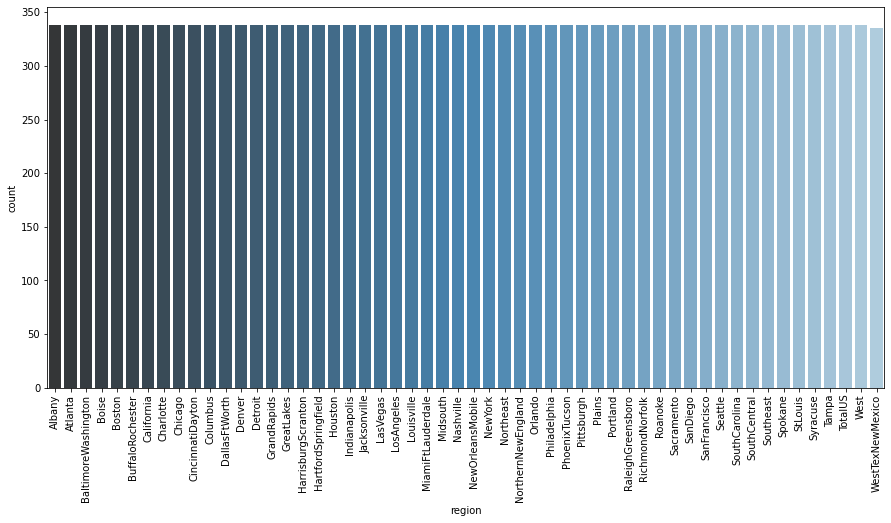
The price of avocado ranges from 0.5 to 3.0 and most of the avocados are sold of the price range

1.0 to 1.5

**Strategy** - Keeping the avocado of price range 1.0 to 1.5 can Hike up the sales

Region

Our next analysis will be of region that is our target variable of classiﬁcation problem.

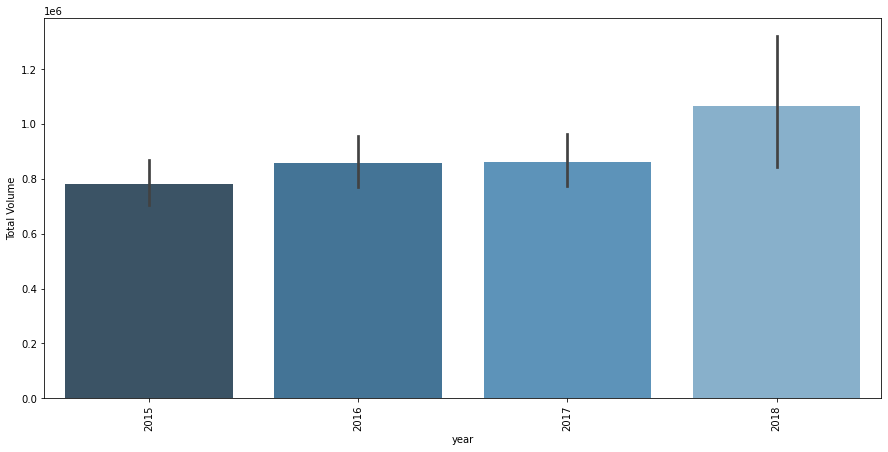


By looking at the graph we see that all the region has same distribution therefore our data is perfectly balanced hence we can use accuracy score as a preferred Matrix when we made machine learning model.

The graph also shows that the data that is given to us have almost same number of observation from all the region.

Year Sales

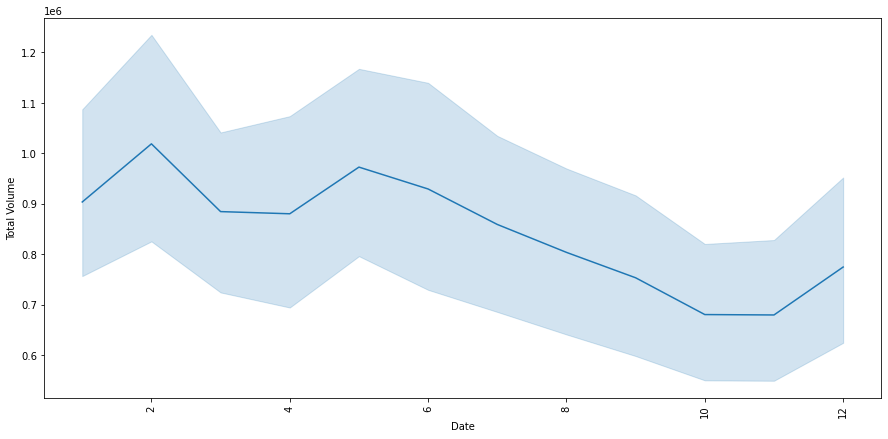
Now when we look at the columns we see there are several columns which show us how much avocado has been sold but there is one column name as total volume which is the summation of all the columns hence we will take this column into consideration and we'll see the sales year wise.



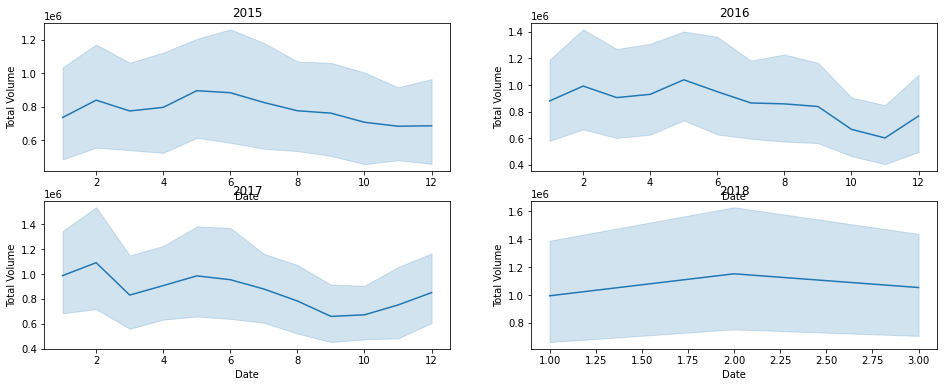
Here we see the growth in year 2016 but in 2017 no growth can be seen whatsoever, but approximately 20% growth can be seen in year 2018, hence we can conclude that demand of avocado is increasing.

Monthly Sales

We see hike in sales as we go forward with the year but what are the monthly factors which are taking part in the increment or decrement of sales, to analyse that we will make a graph of monthly sales of all the 4 year.



when we go forward in the year sales decreases there is a hike sale in summer and in winter holiday season so to get the better and deep understanding we will analyses the monthly performances of each year.



Here we analyses that avocado sale hikes up during summer time but decline as we go through summer to winter, after analyzing monthly sale of all 4 years of data we see that best year was 2018 we see the least decrement in sales as we move forward with the months.

**Seasonal Strategy -**

**Winter**: Holiday parties and gatherings call for chips and dip as well as salads and sandwiches. Avocados can play a role in all of those. Promote them as a nutritious addition to leftover turkey sandwiches and salads for the holiday table as well as their traditional role as a great guacamole dip.

**Spring**: Spring is the time for fresh salads. Promote avocados as a healthy, hearty addition to salads.

**Summer**: Summertime makes consumers think of backyard barbecues and summertime holidays. Include avocados in grilling promotions as they make a tasty and original topping for burgers and sandwiches.

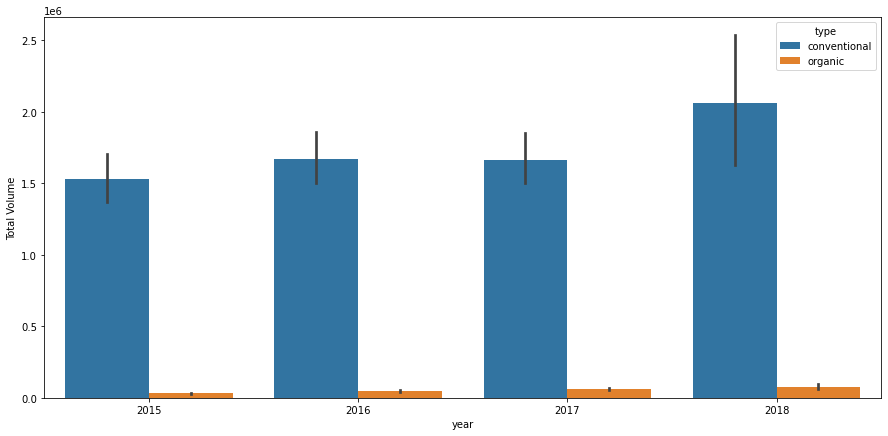
Types of Avocado

There are basically two type of avocados based on their farming they are organic and conventional, the essential difference between organic and conventional farming is that conventional farming relies on chemical intervention to ﬁght pests and weeds and provide plant nutrition. That means synthetic pesticides, herbicides, and fertilizers. Organic farming relies on natural principles like biodiversity and composting instead to produce healthy, abundant food.



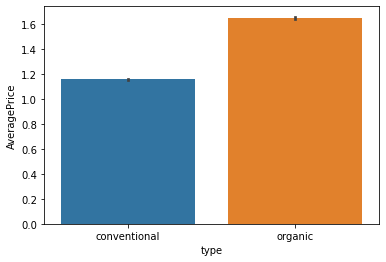
In our dataset both organic and conventional data is given separately and after seeing the above graph we see that its balanced which means that both type of data is given to us equally.

Still question remains the same which type of avocado have more sale and what are the factors affecting it?



It is clearly shown that conventional type avocado has more sales then organic type moreover the sale of organic type is not compare to the conventional type also so we can see hike is more than conventional type but in organic type the hike is not so much.

What is the reason because of which the tale of organic type avocado is much less we know that when it comes to organic products the price is high? so we will analyses price.

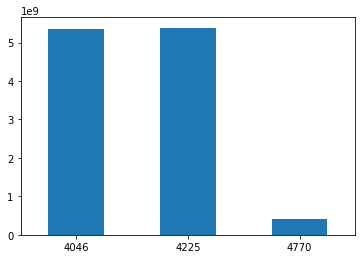


The price of organic avocado is 30% more than that of conventional, we also saw that the conventional type of avocados has more sales so price is more important factor while selecting avocado rather than is origination.

**Strategy**- always have a stock of conventional avocado and if we store want to sell more organic avocado then they have to lower its price or to do an advertisement in such a way showing the beneﬁts of organic avocado.

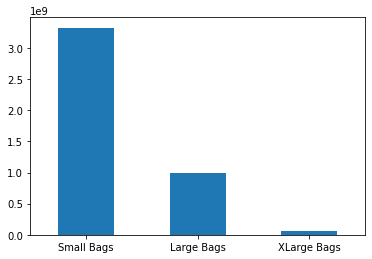
Unit Sales

Other than origination there is also a three type of avocados labelled as different PLUs, we will see their sales.



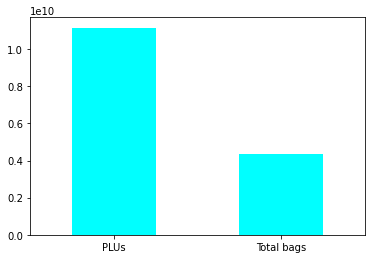
Here we observe that PLU 4040, 4225 has almost equal sale but 4770 has only 10% sale compare to the other PLUs.

Now when avocados is sold as a single unit it mentioned in a PLUs but there is one more way by which avocado is sold and that is bags, they are divided into three category, small, large, and extra-large.



When we look at the sales data of avocado sold in bags we saw that most of the people prefer the small and that are reduced to more than half when we move forward it suggest that people are willing to buy avocado in small amounts rather than buy them at once and store them.

Before making a ﬁnal strategy there is one thing that left and that is to compare the sales between PLUs and bags and to see which type has more sales.



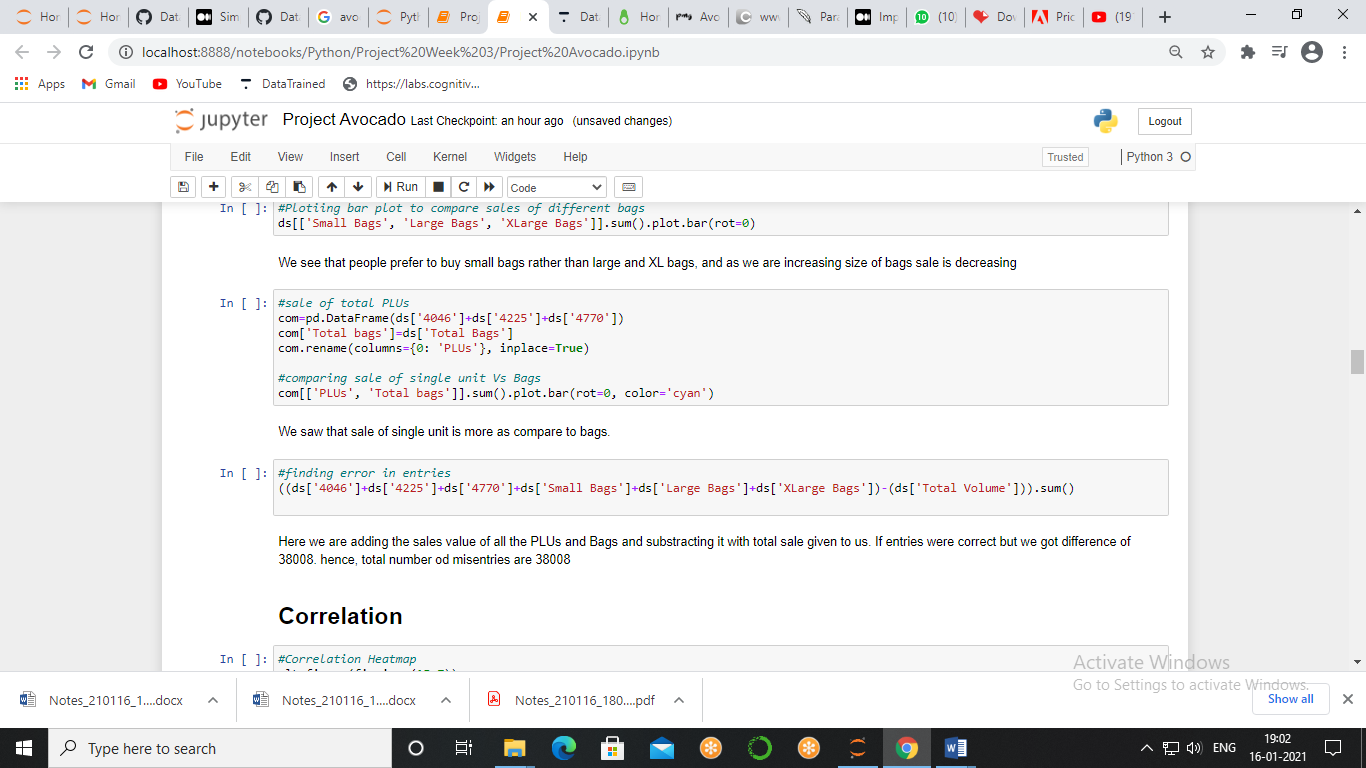
The sale of PL use is approximately 60% more than sale of total bags which clearly show us that people are buying more avocados in the form of single units.

**Strategy**- people are buying avocados in Less numbers maybe because of the economic conditions they don't want to store the avocado the only one the fresh ones, now to increase the sales stores should increase the availability of avocado that are sold in form of single unit and if they want to sell the avocados in form of bags then there should be more availability of small bags.

Error in the entries.

We don't know how data is collected ETA in form of machine or by human but in both the cases there is a chance of error and we want to ﬁnd that error the number of unit which are not mentioned.

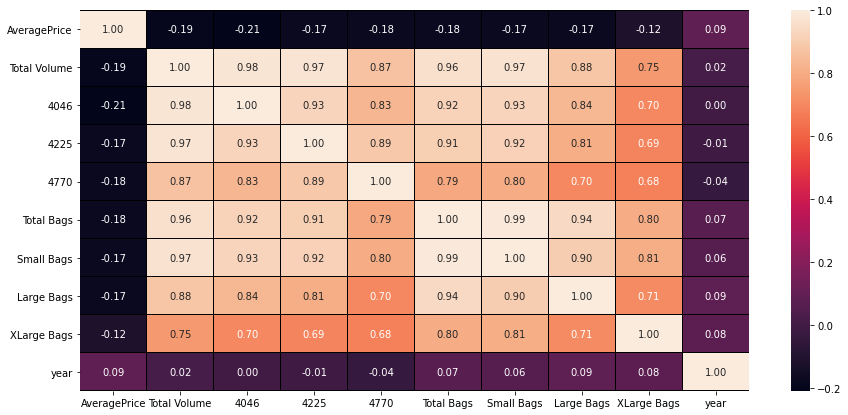
To do this, we know that the Total Volume feature is sum of the different PLUs and different bags, addition of all the features and subtracting it from total volume will give us the number of errors.



If entries were correct but we got difference of 38008. hence, total number of fake entries are 38008.

Correlation

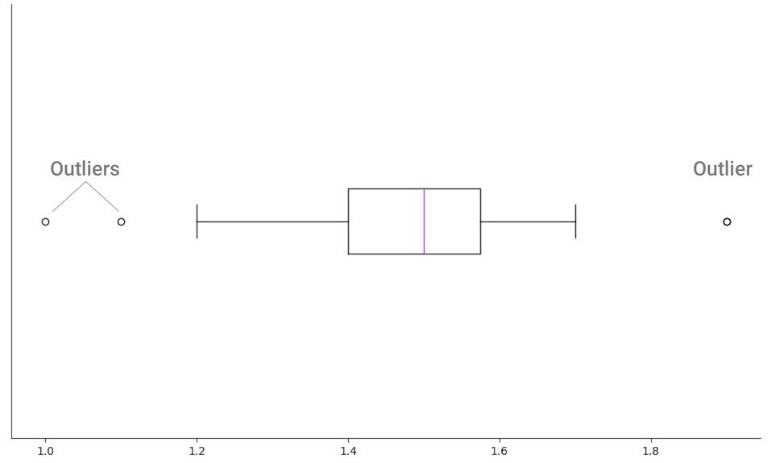
Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It’s a common tool for describing simple relationships without making a statement about cause and effect. We will make heatmap which will tell us the amount by which two factors are how much related to each other.



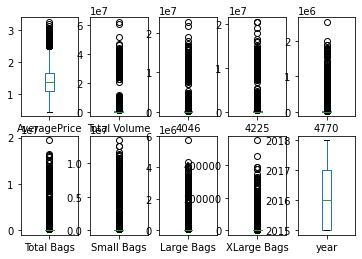
* High Correlation with sales = PLUs and Bags
* Mild Correlation with sales = Year
* Low Correlation with sales = Price

Outlier

An outlier is a piece of data that is an abnormal distance from other points. In other words, it’s data that lies **outside the other values** in the set. If you had Pinocchio in a class of children, the length of his nose compared to the other children would be an outlier.

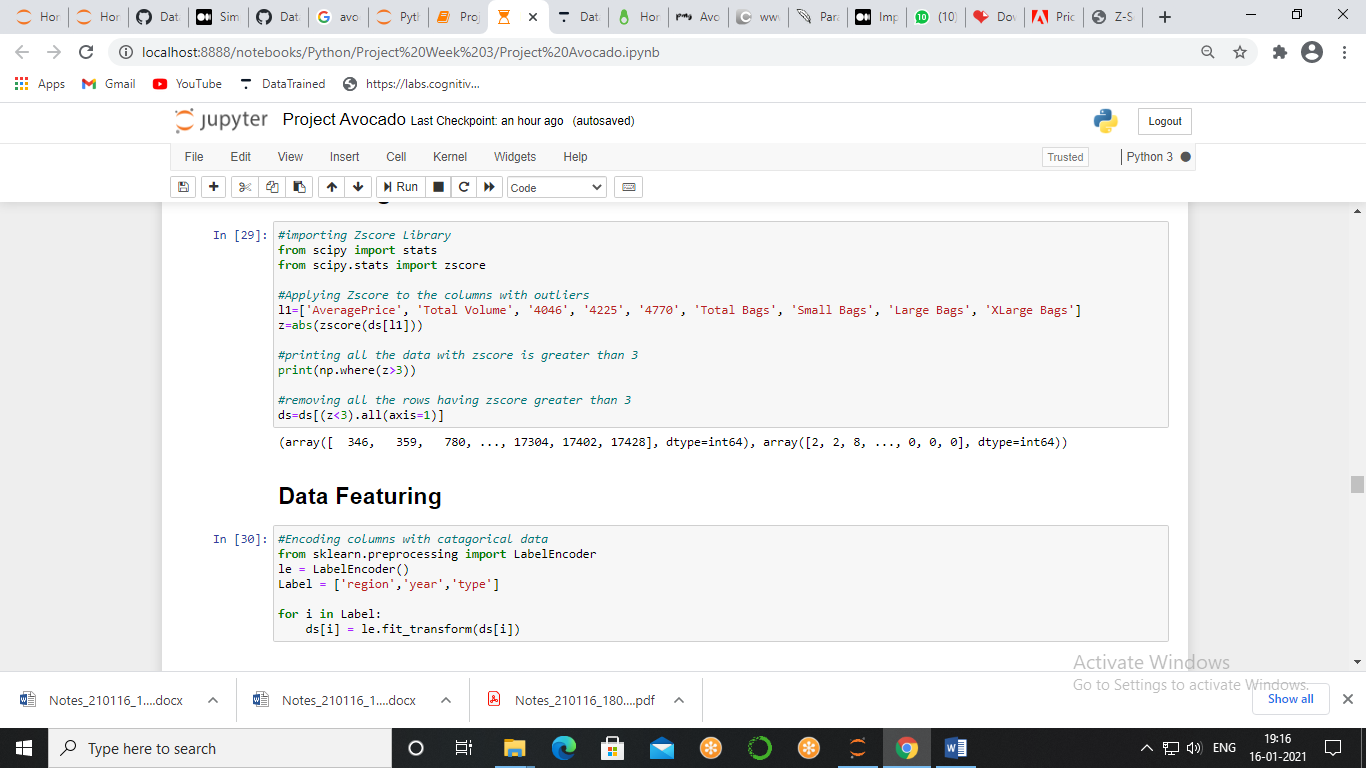


We are making boxplot to see the outliers



we see that columns AveragePrice, Total Volume, 4046, 4225, 4770, Total Bags, Small Bags, Large Bags, XLarge Bags Have Huge Outliers.

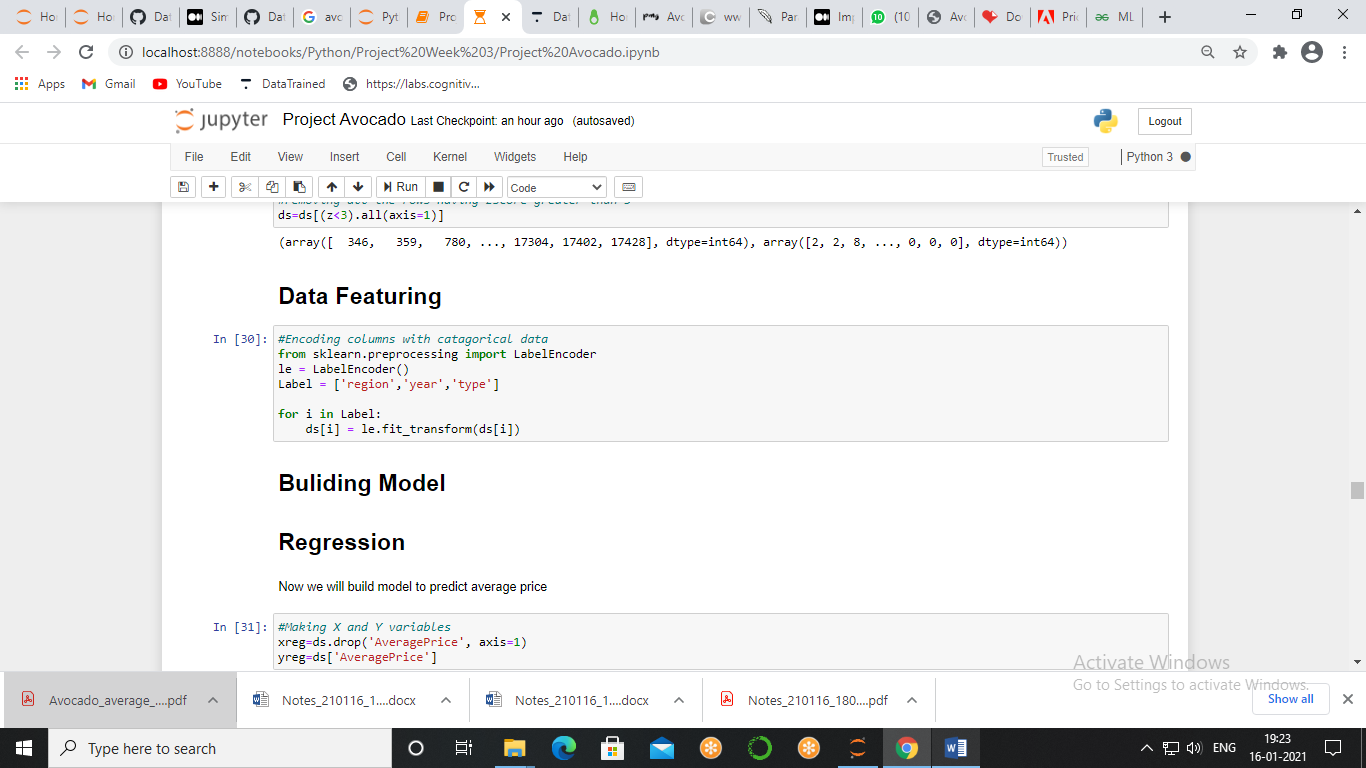
We are removing outliers with the help of z-score, it describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units, the threshold unit is considering as -3 and +3, outside these boundaries lies outliers.



Encoding

**Label Encoding** refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated.

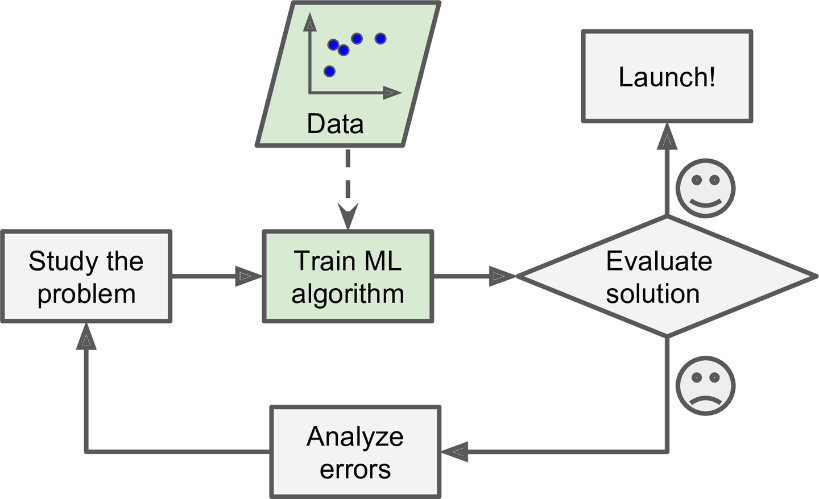
Here we have 3 feature that are in object format and need to convert in numerical value.



Machine Learning Model

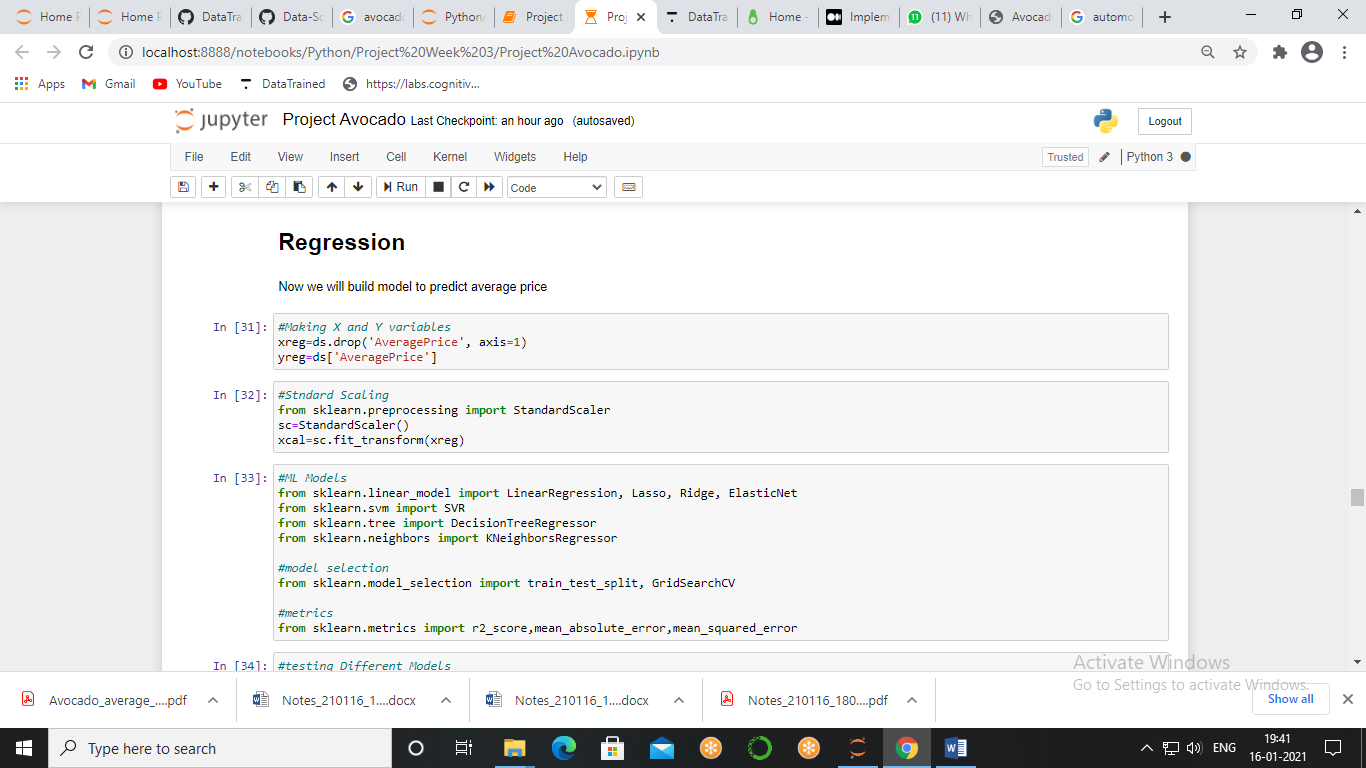
**Average Price Prediction**

We have to predict the average price of a avocado and Average Price data type is continuous type data type so we will go with regression.



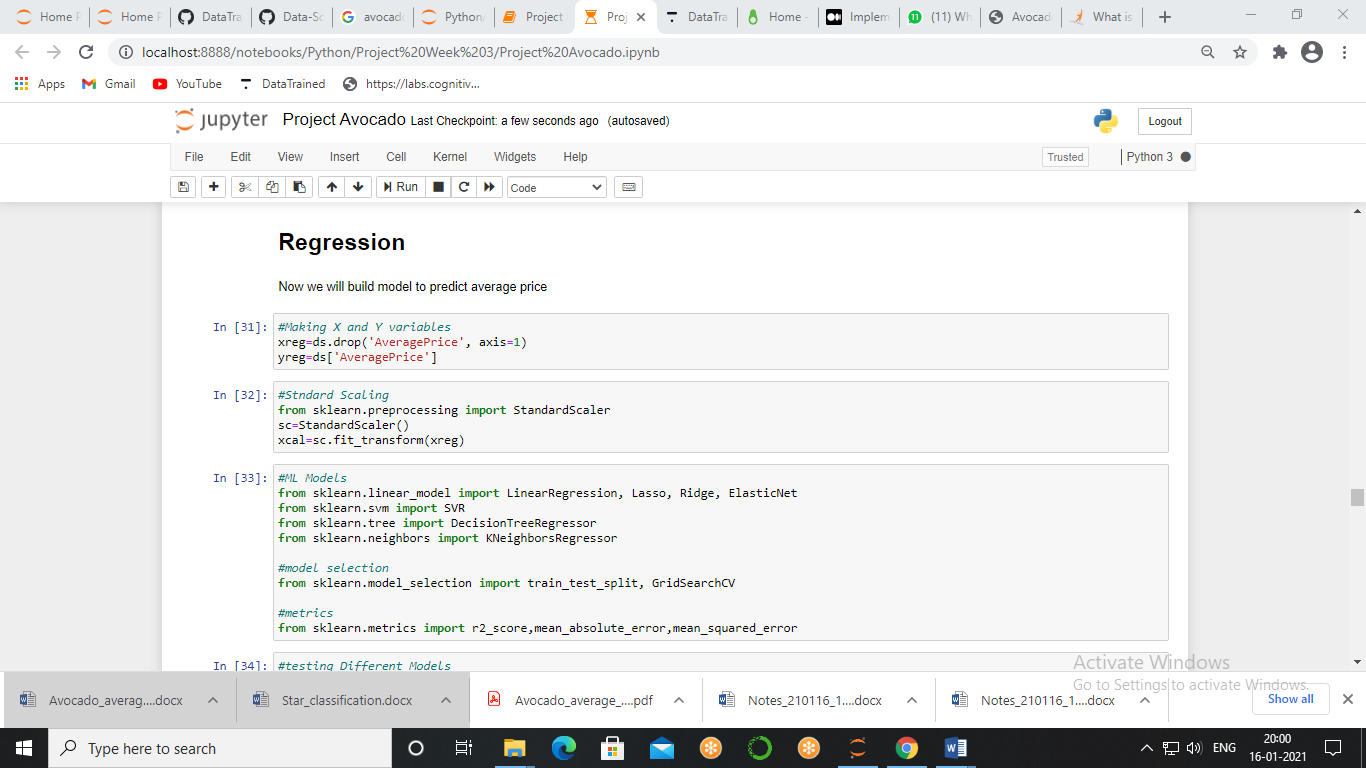
**Spiliting Data**

Now we will split dataframe into two parts X and Y where X will be a dataset by which we will predict and Y will be our target variable we will save them into xreg and yreg variable.



**Scalaing**

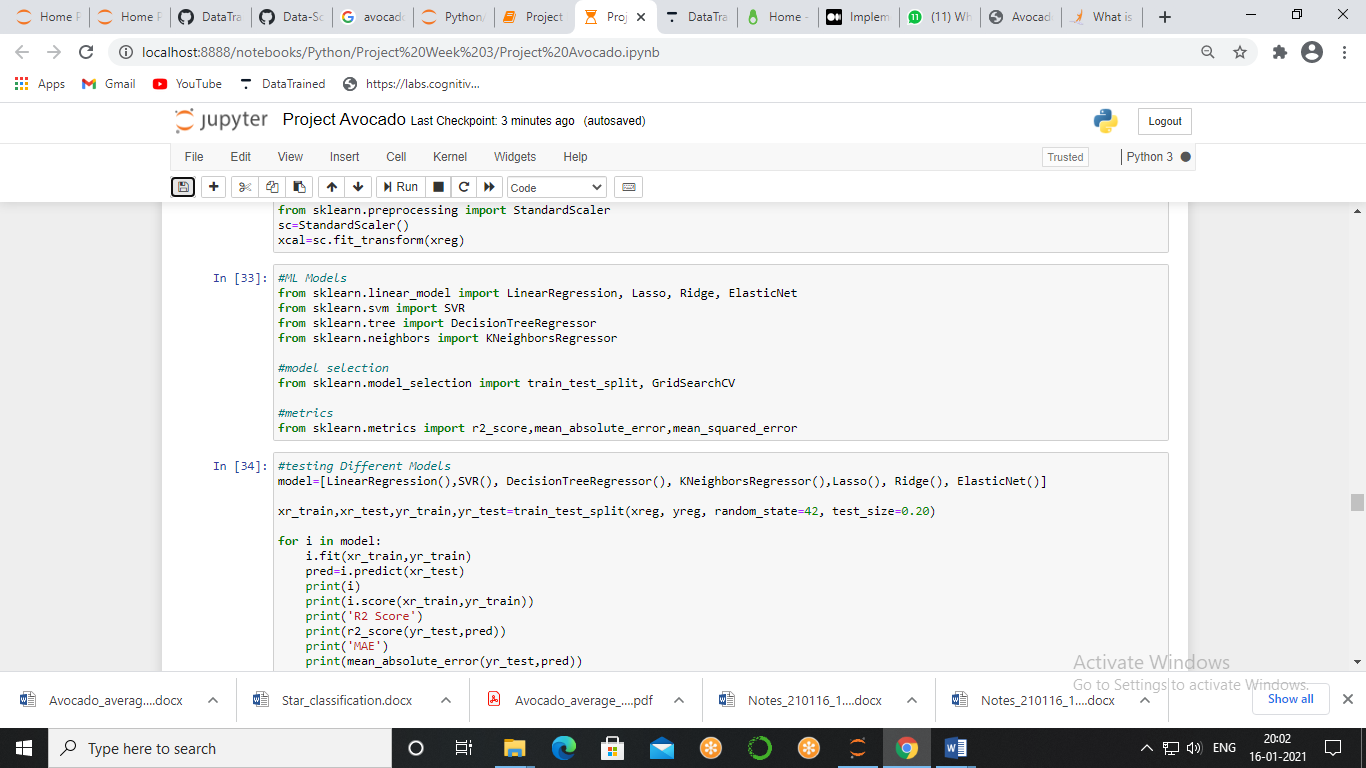
There are model available which can be applied in a regression type dataset to check different model first we will import their libraries then we will import our matrix which will show us performance of our model.



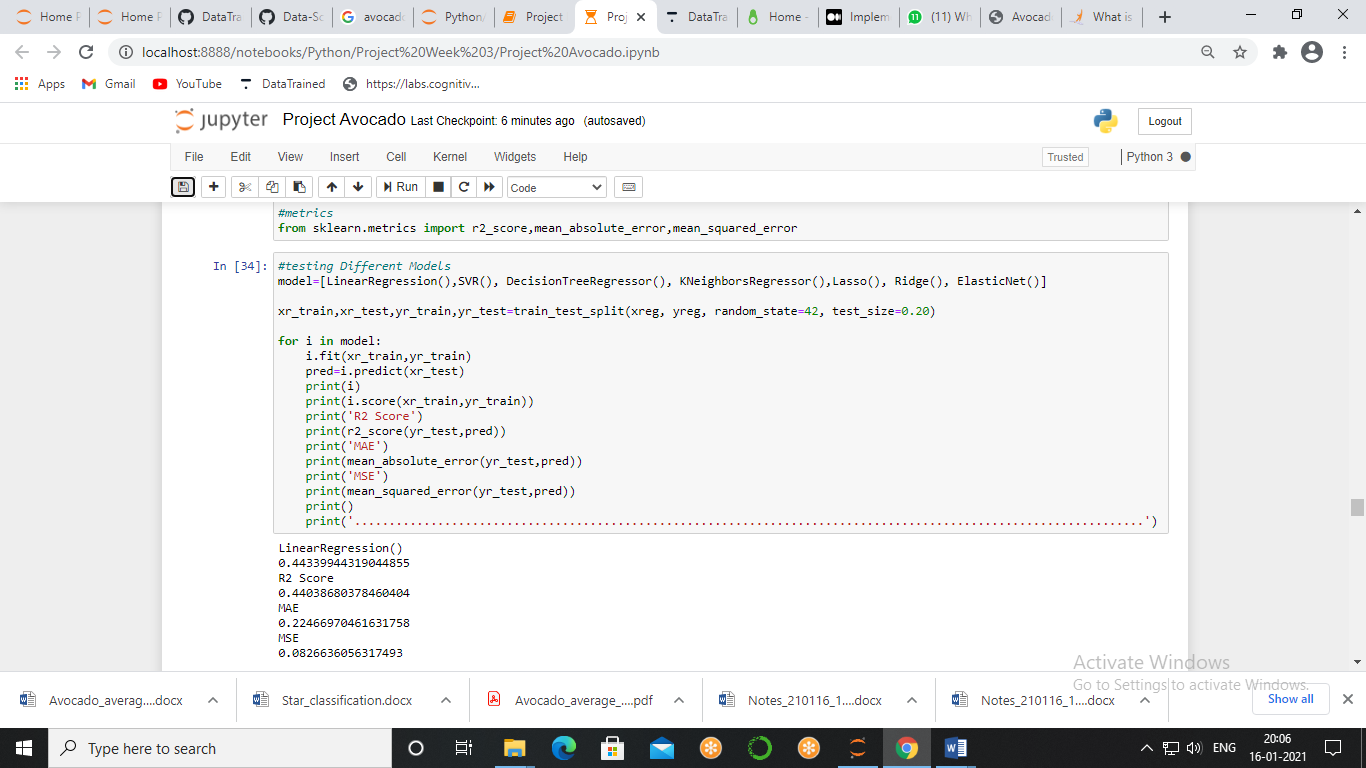
**Building Model**

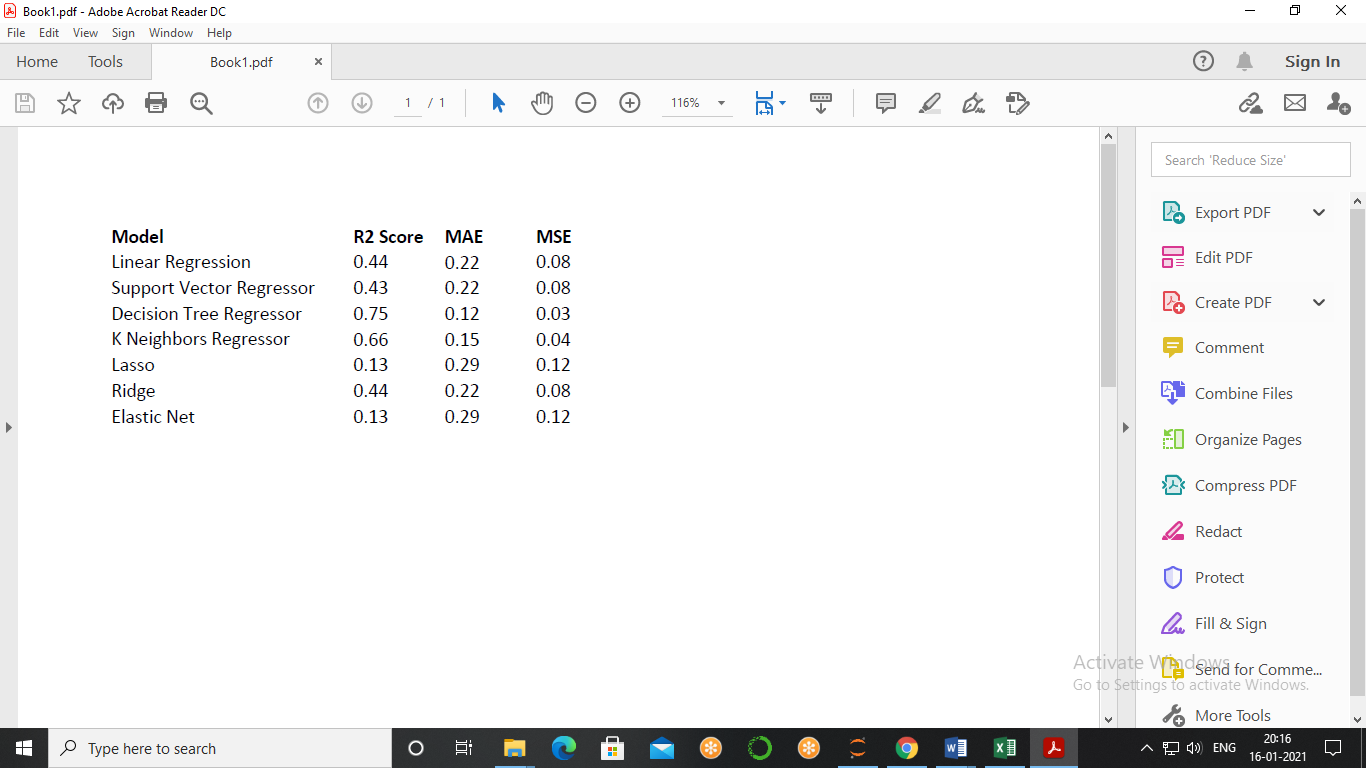
There are model available which can be applied in a regression type dataset to check different model first we will import their libraries then we will import our matrix which will show us performance of our model.

*Matrics* – our prefered matrix will be R2 Score and our aim will be high R2 Score and less MSE and MAE



To check all models at once we made a list of callable name of the models and pass it to the for loop result of which we will get the best of all the models which will make us easy to choose the best model.

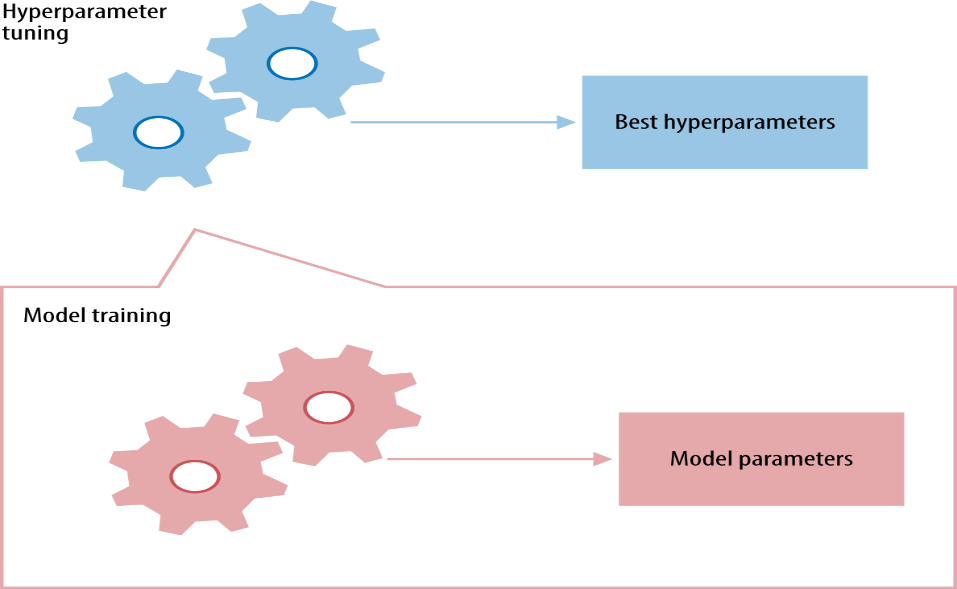




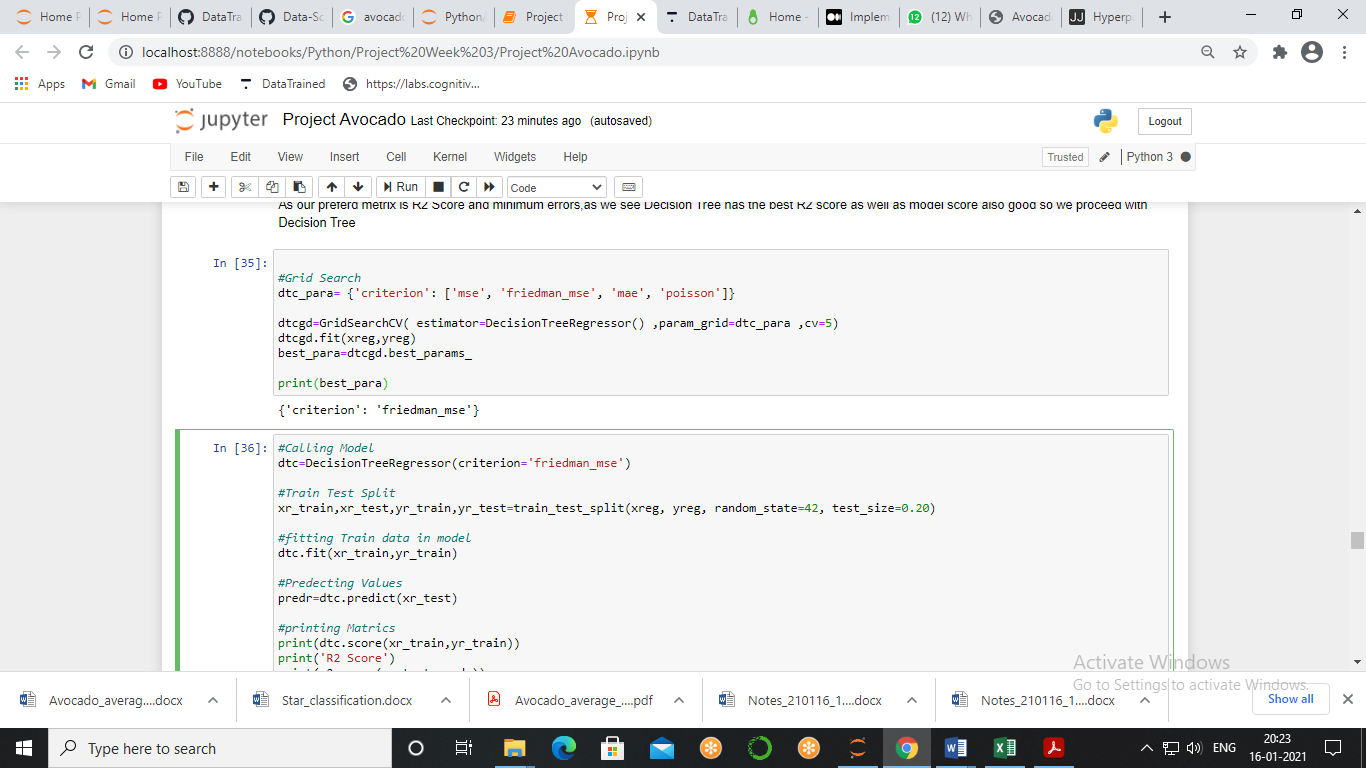
As our prefer metric is R2 Score and minimum errors, as we see **Decision Tree** has the best R2 score as well as model score also good so we proceed with **Decision Tree.**

Hyper Parameter Tuning

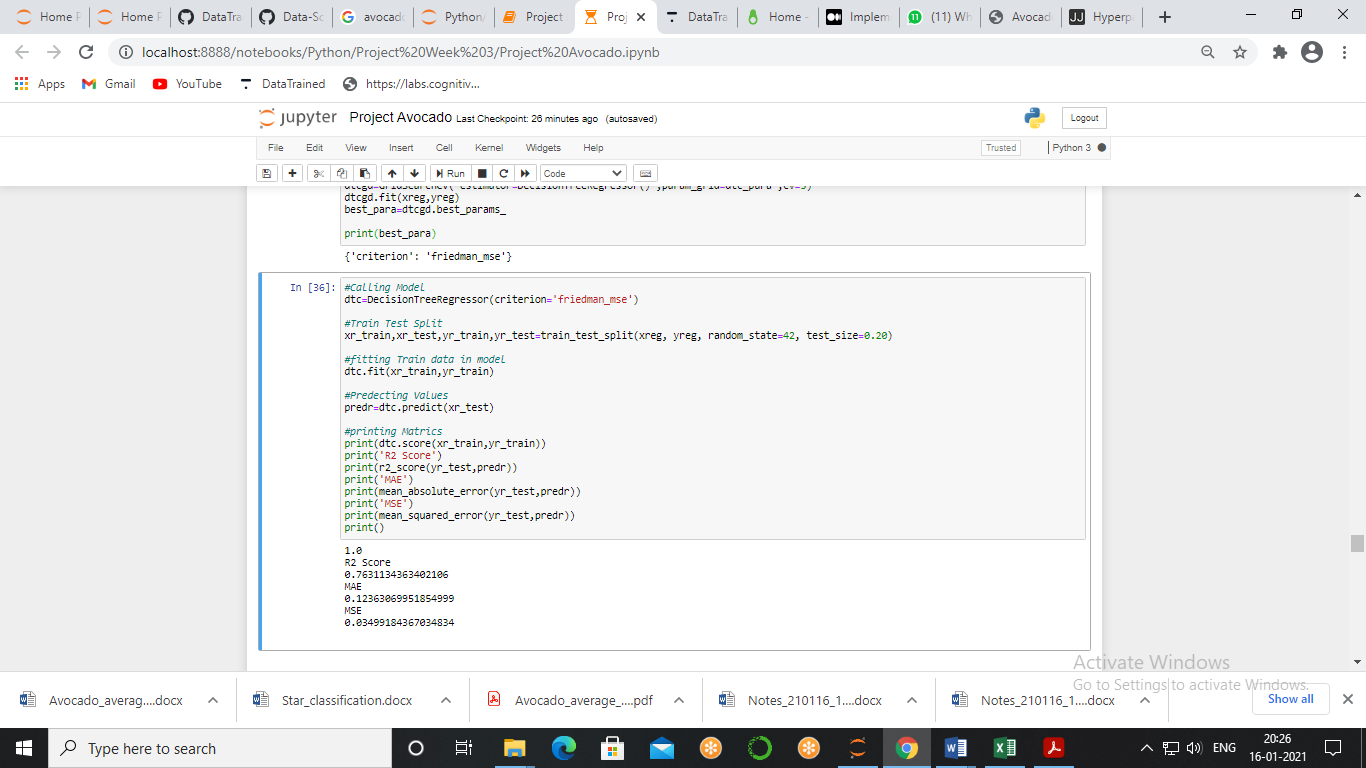
Parameters which define the model architecture are referred to as **HyperParameters** and thus this process of searching for the ideal model architecture is referred to as **HyperParameter** tuning.



we will use grid search to find the perfect parameter.



We got Criterion – Friedman mse as our best parameter so we will make our final model with this parameter.

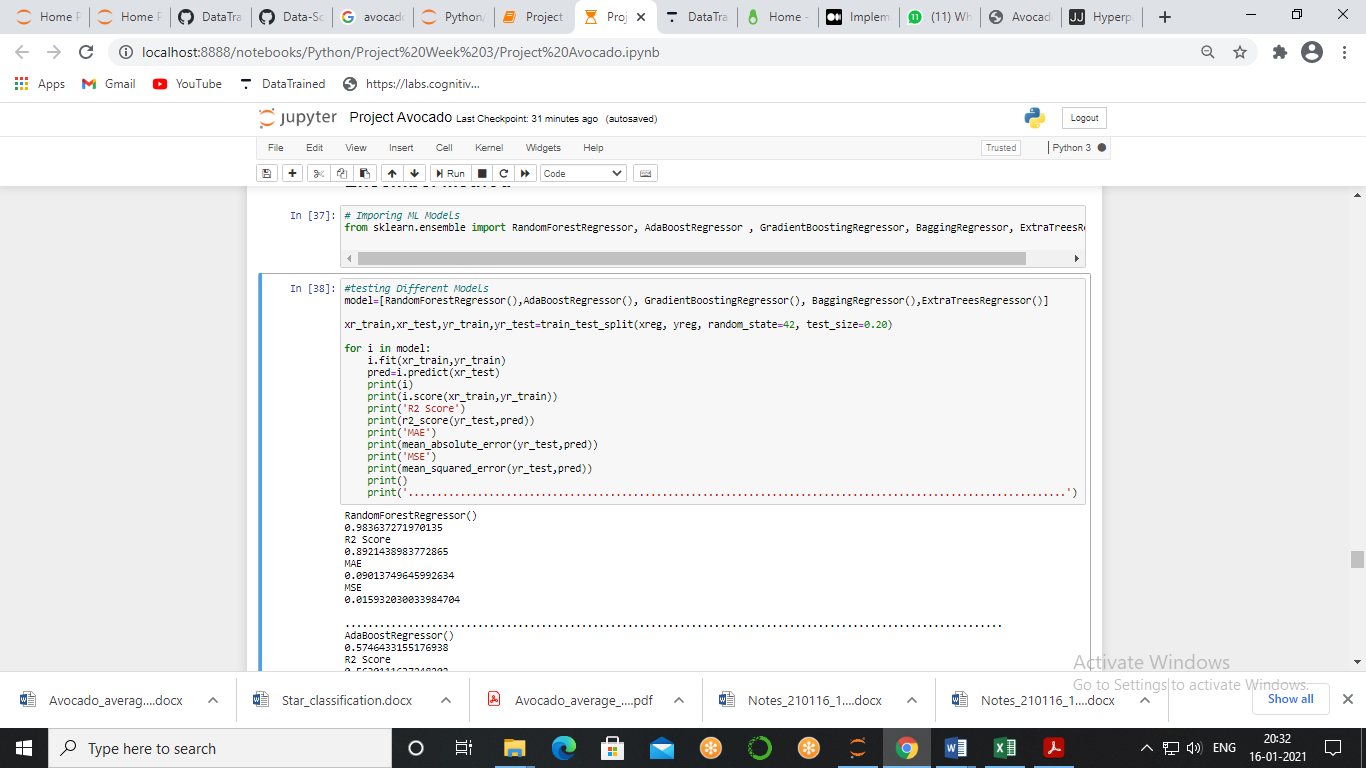


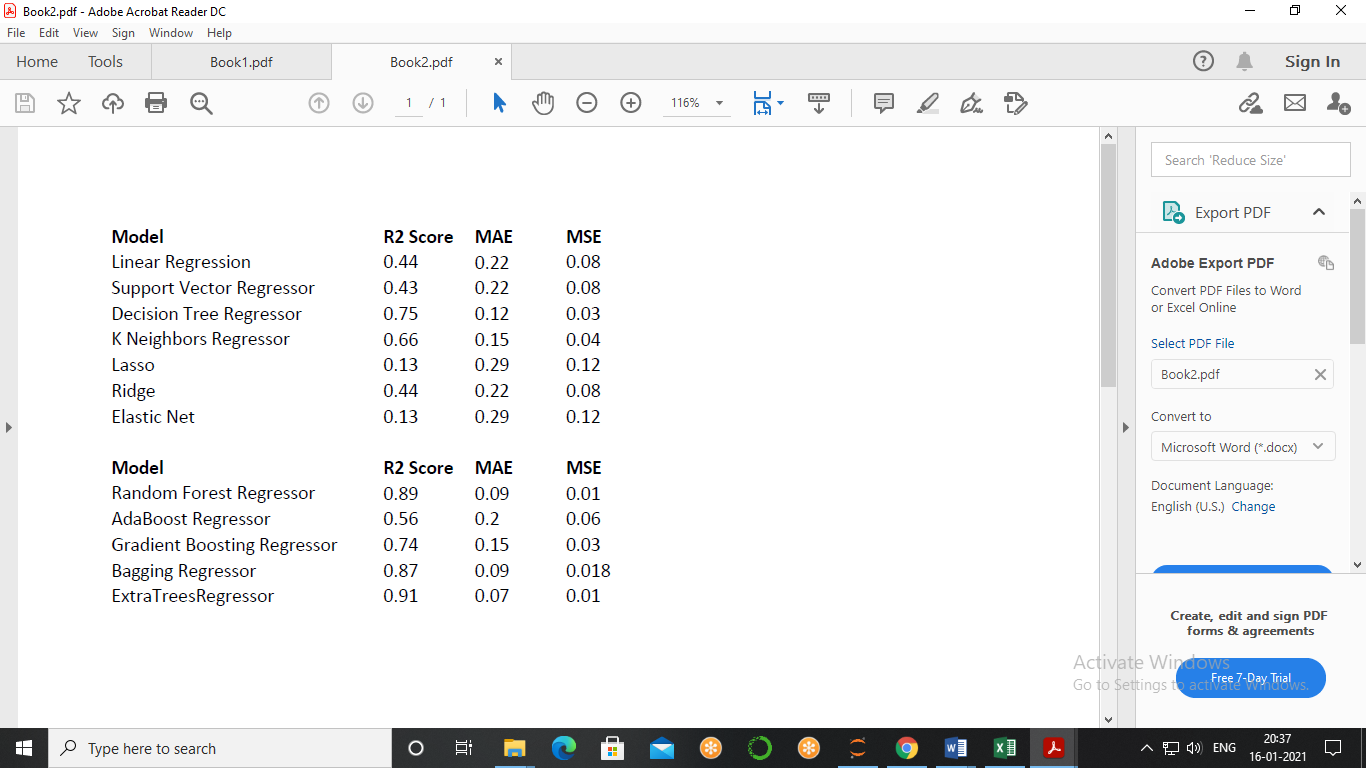
We got our model with **R2 Score = 76%, MAE=0.12, and MSE = 0.03**, it is a good model but we can further improve its performace by using emsembel methods.

Ensembel Method

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

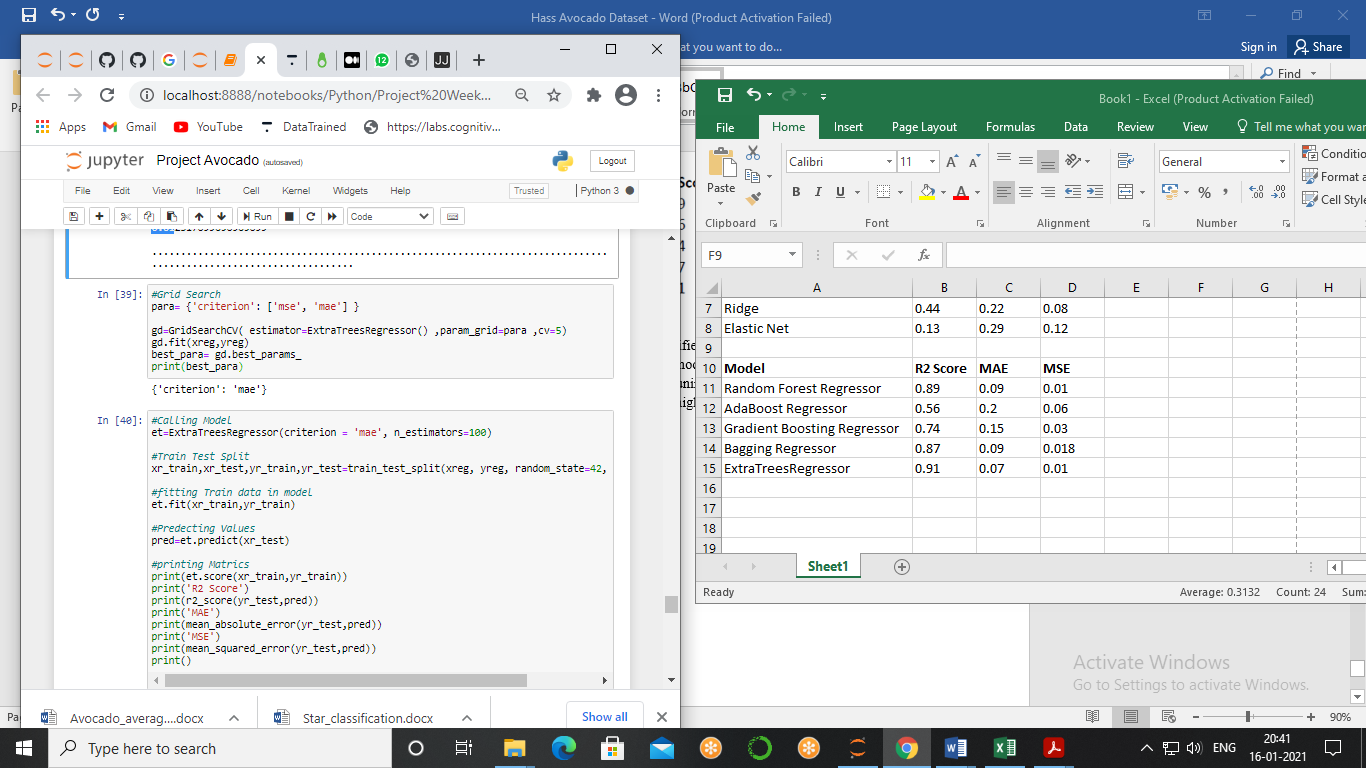
We will test different methods using for on our data and will compare the matrices



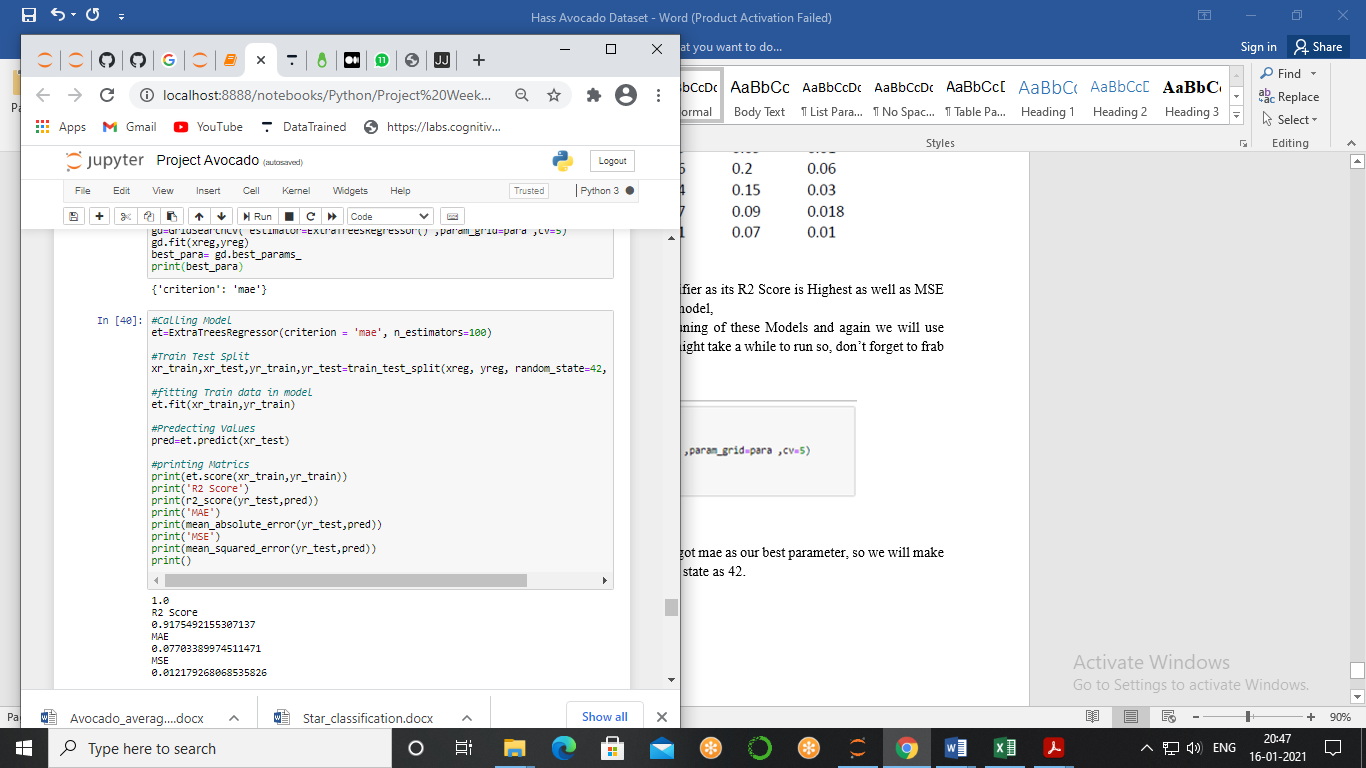


Here we see that best model is **Extra Tree Classifier** as its R2 Score is Highest as well as MSE & MAE is lowest so we will proceed with this model,

Now, we again have to do HyperParameter Tuning of these Models and again we will use GridSearchCV method keep in mind this code might take a while to run so, don’t forget to frab a popcorn.



After some time you will get result in which we got mae as our best parameter, so we will make our final model with this parameter and random state as 42.



So, this is our final model to predict the average price, with **R2 score = 91%** which is quite high and it is good score, **MAE – 0.07 & MSE = 0.01,** which is almost nigleble.

So, if anyone now will pass all the feature to this model he/she will get average price of avacado with 91% accuacy.

Conclusion –

The demand for avocados is increasing globally due to their health benefits as it increases the vision, prevents heart related diseases, and helps in improving digestion, and to fulfill a demand and increase the sales of avocado we did analysis and meet different strategies to increase the sales as well as what are the factors that are affecting the sales.

Price ,origination, freshness and how they are sold are the main factor which affects the sales , controlling and improving them can really hike up the performance.

To take a perfect decision and make a perfect strategy we made a machine learning model ll so that before taking any decision we can enter the values and get the data.