**AVOCADO**

Avocado is a green, pear-shaped fruit often called an “alligator pear.” Commercially valuable with production increasing worldwide over 2018-26 at 10% per year, avocados are cultivated in tropical and Mediterranean climates of many countries.

Avocado is consumed highly in United States. In 1985, domestic consumption in US was a meager 436 million pounds. That number has increased six-fold to over 2.6 billion pounds of the berry being consumed by Americans in 2020. Avocado prices have rocketed in recent years by up to 129%. Heatwaves, droughts, and water scarcity can have devastating effects on avocado production.

In this article, we will be working with a dataset in avocado in US.

The dataset we are using has been gathered from [Hass Avocado Board website](https://hassavocadoboard.com/) in May of 2018.

It 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

**PROBLEM STATEMENT**

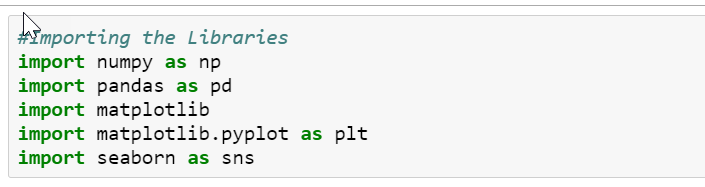
As mentioned earlier, the production and cost of avocado in US has been increasing in recent years. It becomes important that the factors related to the increase in cost must be studied.

Our goal is to create a machine learning model which can predict price of avocado in US based on the information we have at our hand with utmost precision.

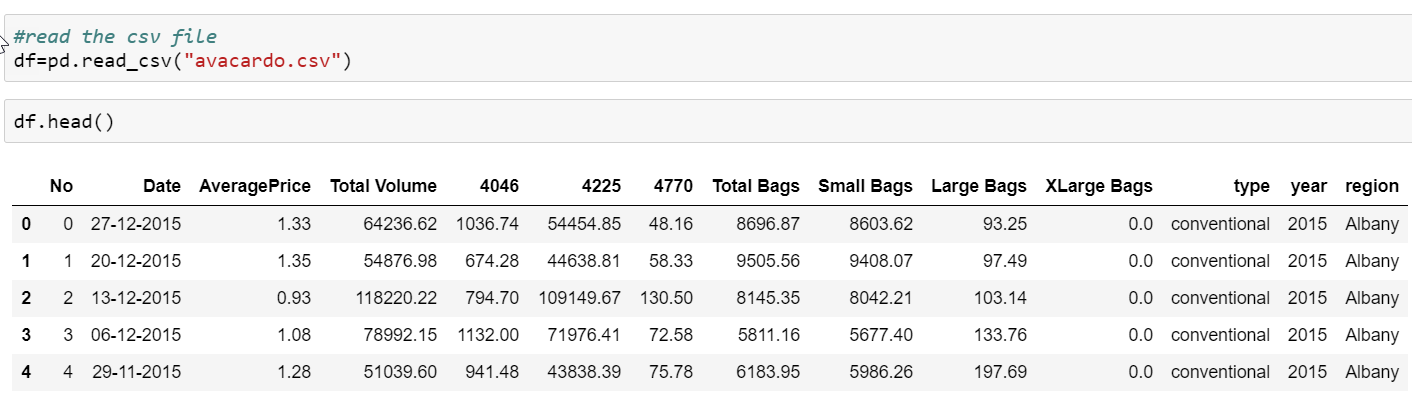
We will also be creating another model which will predict the city/region of the avocado from the observation data we have.

Two models for a fruit generating revenue in billions sounds reasonable.

We will begin with loading libraries which may be utilized at different stages.



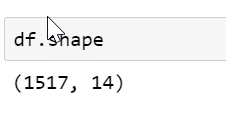
Loading the data...



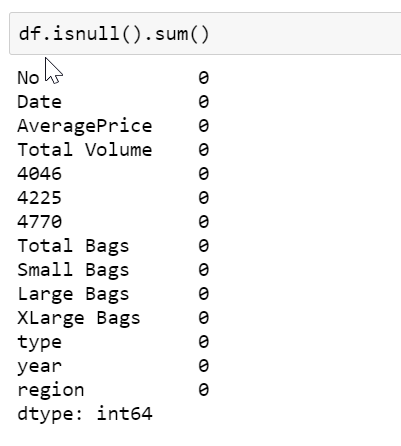
**DATA ANALYSIS**

Let’s start with looking into the data and understanding the data we have at our hands.

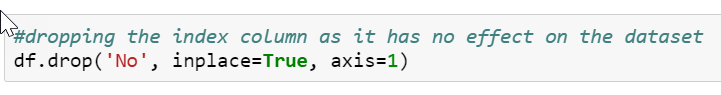
* Checking the dimensions of the data:



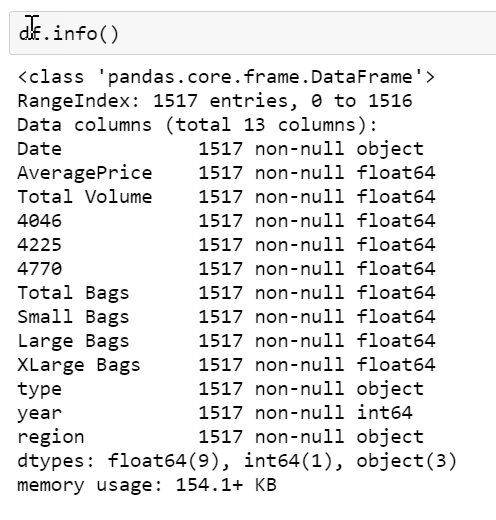
* Checking if there are any null values in the dataset:



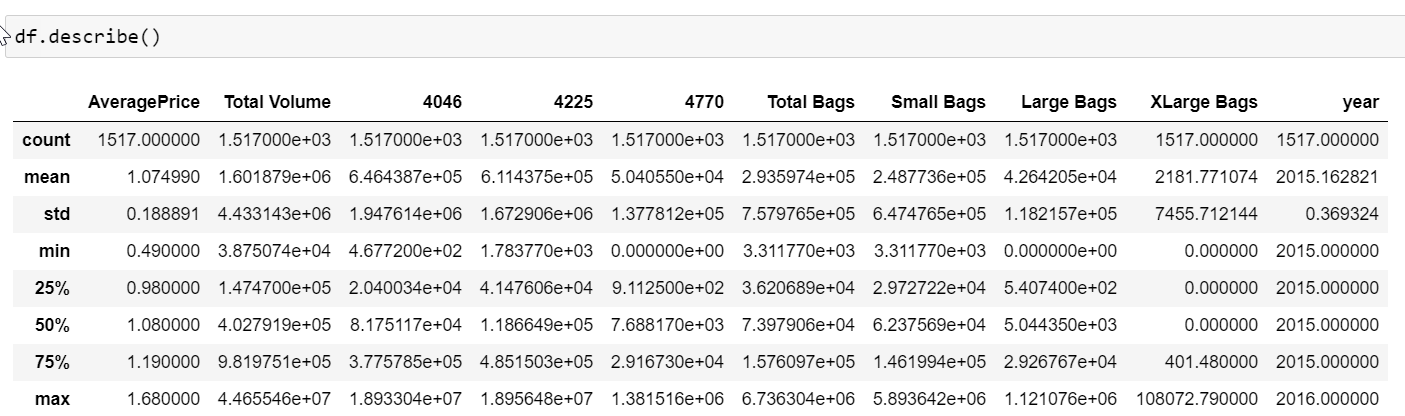
We will be dropping the index column as it is of no use in our dataset for creating the desired machine learning models.



* Let’s look at a concise summary of the dataset:

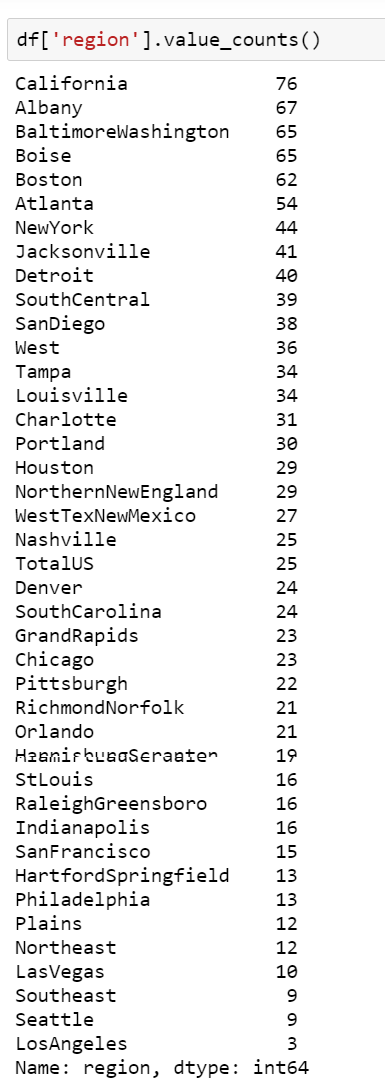


* We will have a look at the basic statistical details like mean, percentile, count, etc. of the columns.

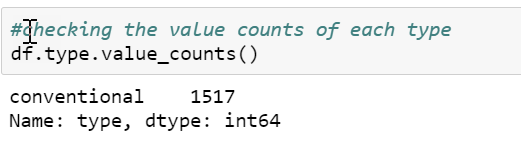


Now we will look at different columns and analyze them either individually or with other columns to draw conclusions on their relativity.

* Let’s look at the region column and identify the unique values in this column and the count of its occurrence.

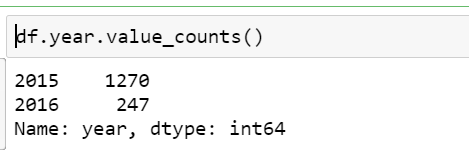


* Let’s verify the type of avocado’s we have

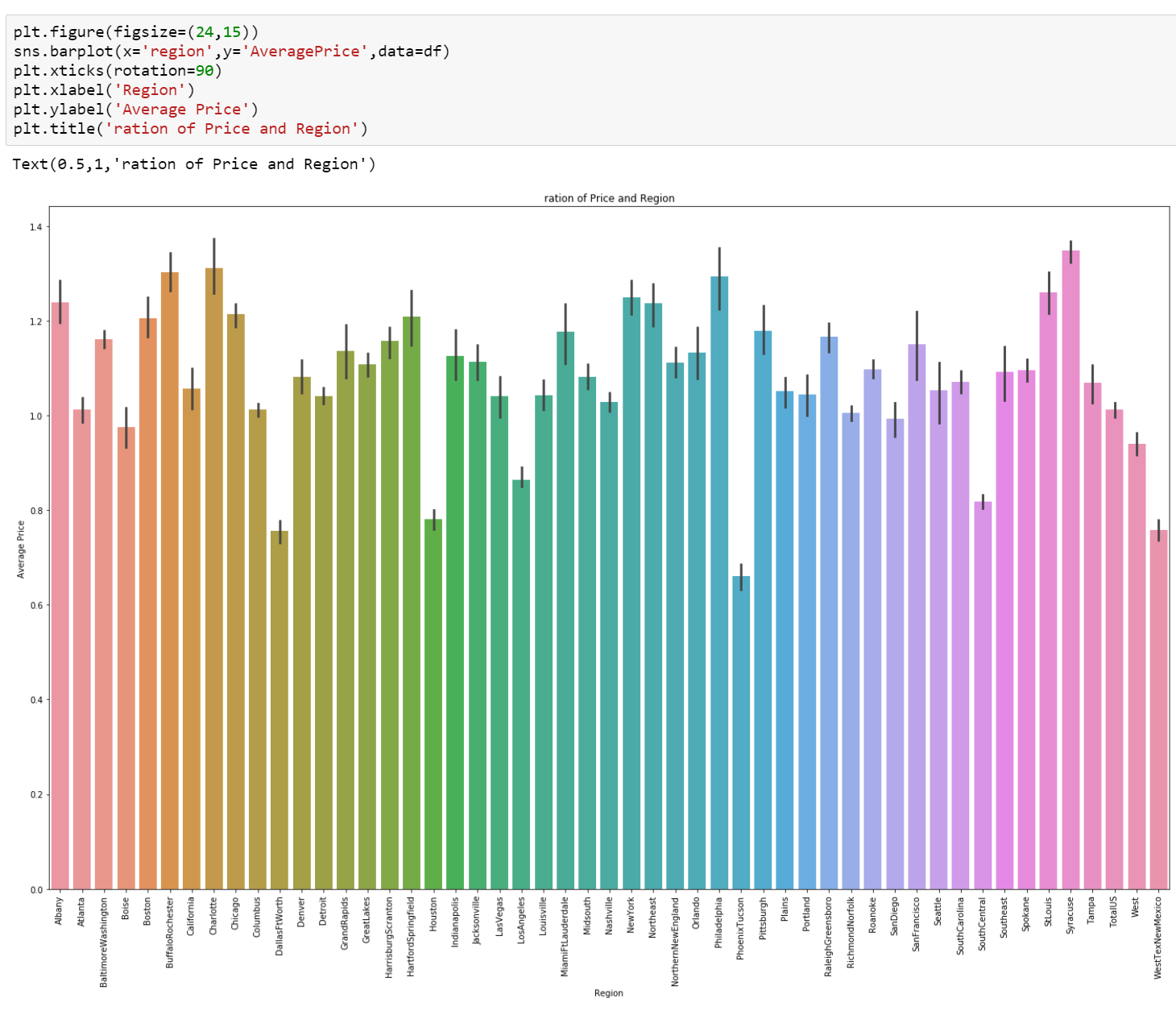


We can remove this column as all the avocado are of same type and will not help in our prediction.

* We will check the observation year and the count of observations in each year

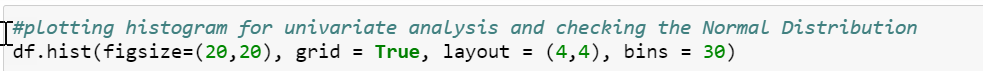


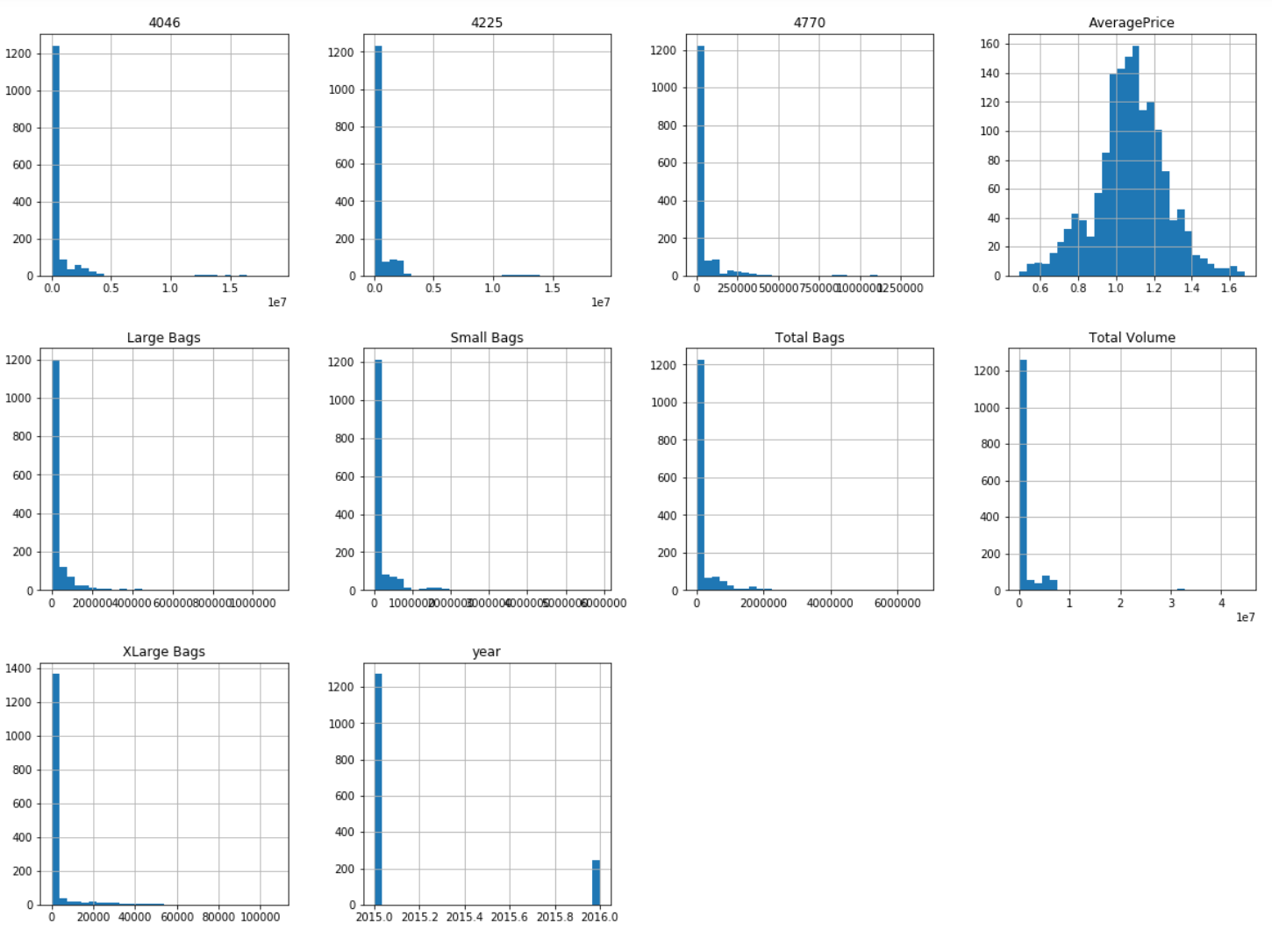
Let’s plot our two target variables Average Price and Region with each other and understand the relation between them.



From the above graph we can clearly infer that the Hartford Springfield, San Francisco and New York have higher average price as compared to other regions.

* We will be plotting a histogram to do a Univariate analysis of each column and check the normal distribution.





**EDA Concluding Remarks**

Our data analysis shows us that the data has no missing values and is complete even though we will be checking for outliers and removing it in our pre-processing.

It also gives us information about the data type in each column. We will have to work on some of the columns to convert them into integer type in our next step.

Univariate analysis of columns shows us the normalization on Average price column.

We were also able to identify city/region with high average prices and low average prices. Cities with high cost of living such as New York and San Francisco took the top spot as one can expect.

Love move to the next step of processing our data where we will be converting, omitting, adding data/columns as deemed fit.

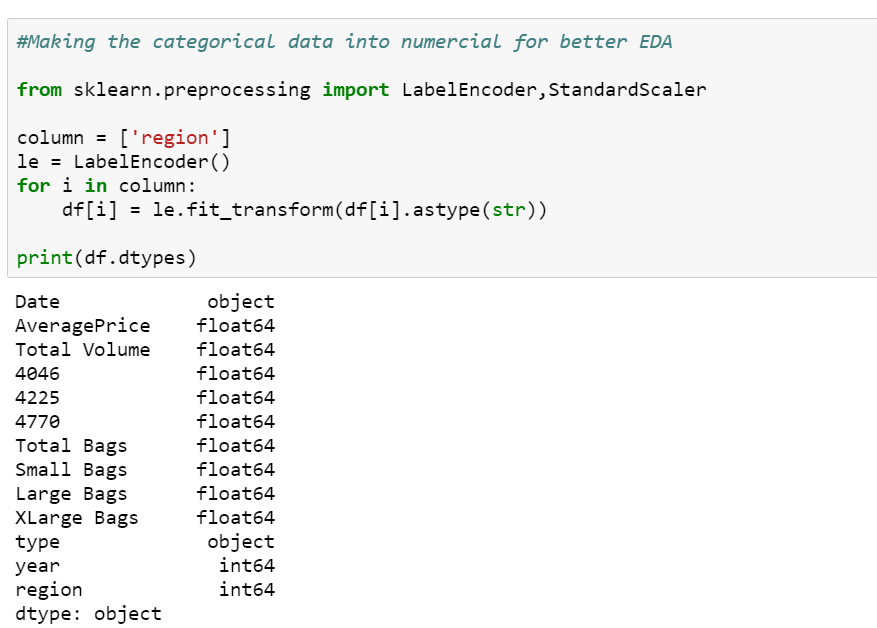
**Pre-Processing Pipeline**

There are 3 columns with object datatype, region, date and type.

We will be converting region to int datatype using Label encoder.

Label encoding is simply converting each value in a column to number.

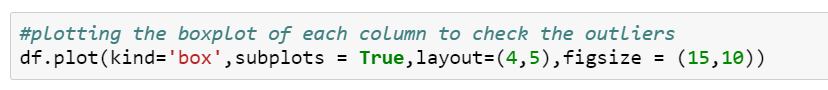
We will use label encoder to convert the columns and check the dataset after applying the same.

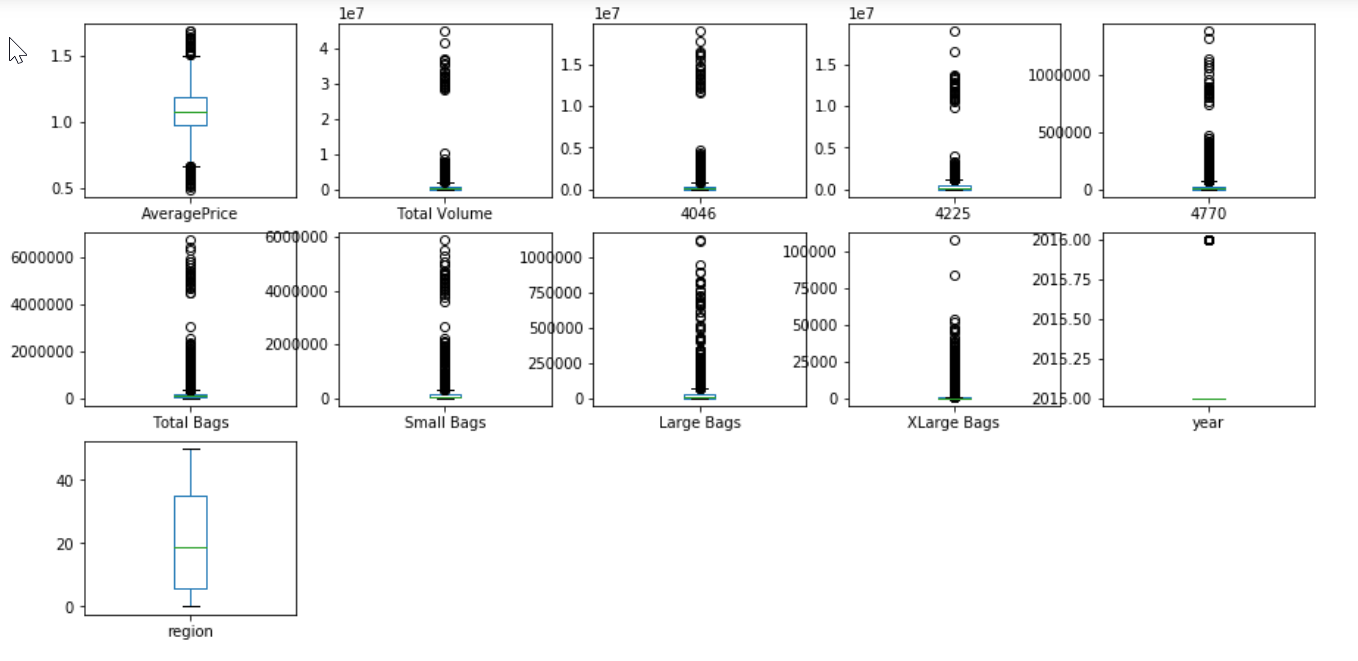


* We can drop the type and date dataset as we have seen in our data analysis that they do not add value to the model we are planning to create for predicting price or region.



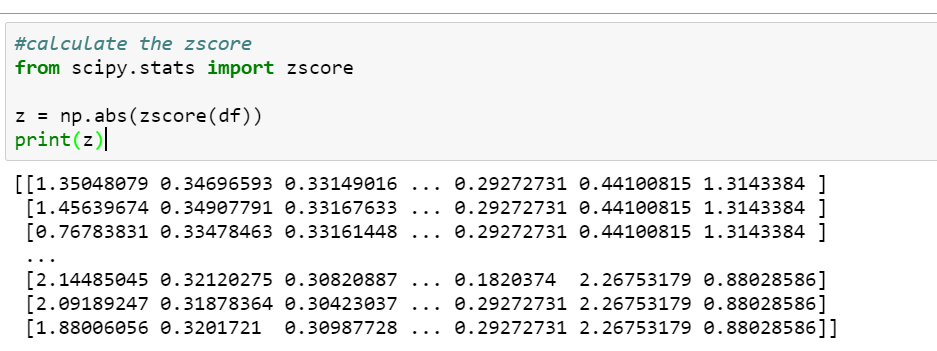
* Now let’s to plotting of each column using boxplot to identify outliers in our dataset.



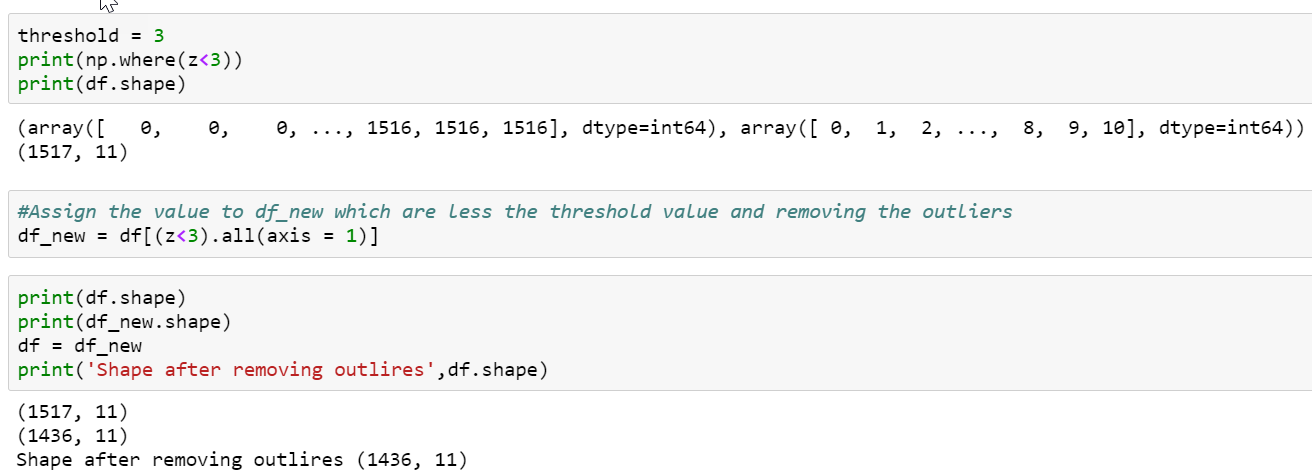


As we can see, there are a few outliers in the columns.

* Let’s look at the z-scores.



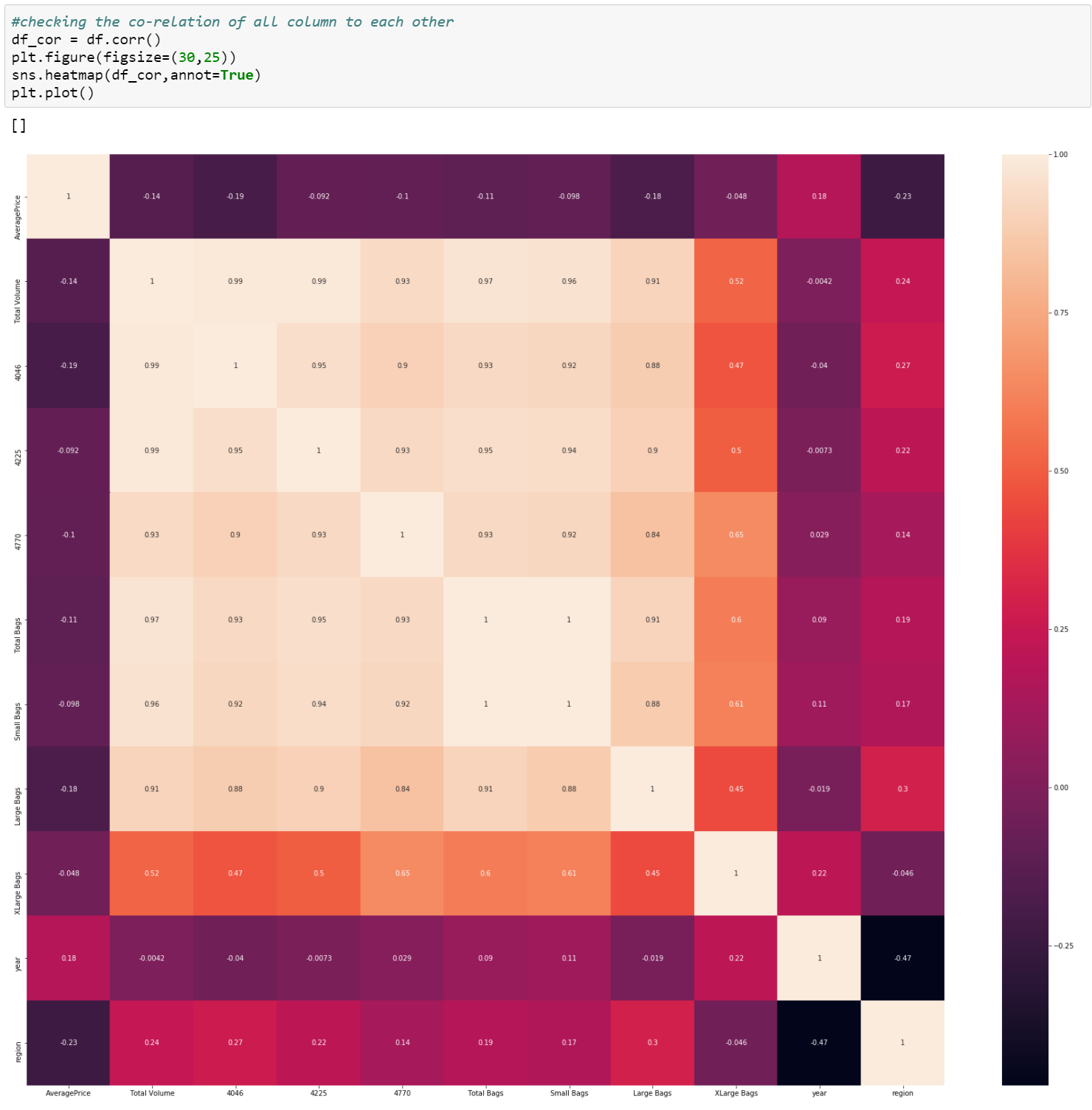
We will check for values with z-score less than three and keep them in our dataset. The rest of the values will be dropped from our dataset.



Now we will find the mean values of all the columns after the outlier values are removed from the dataset. We will use the mean values to replace the outlier values in our original dataset.



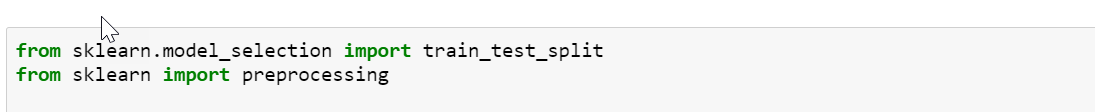
* Now we will look at the correlation of columns with our target columns region and average price.



Both region and AveragePrice column have good correlation with all the columns. Hence, we can work with all the columns as seen above.

**Building Machine Learning Model**

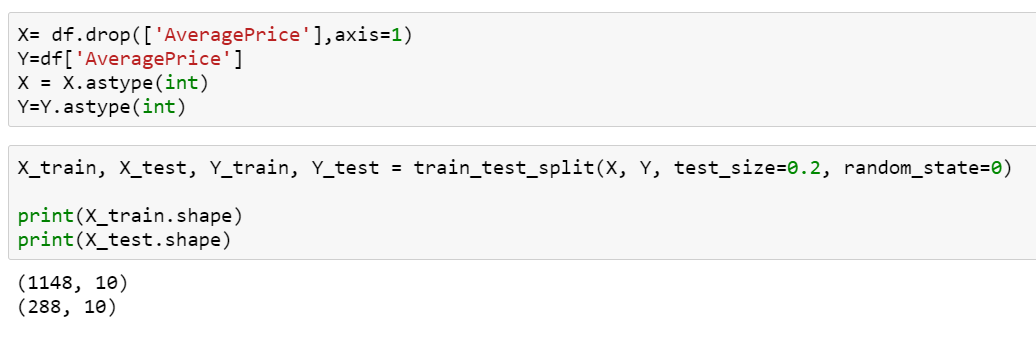
We will be importing the libraries required for creating train and test data and for creating our machine learning model.



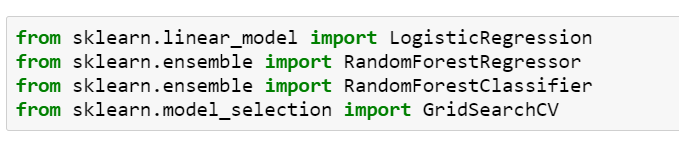
As already mentioned, we will be creating two models for this dataset, one for predicting price and the other for predicting region.

Let us first work on creating a model for predicting price.

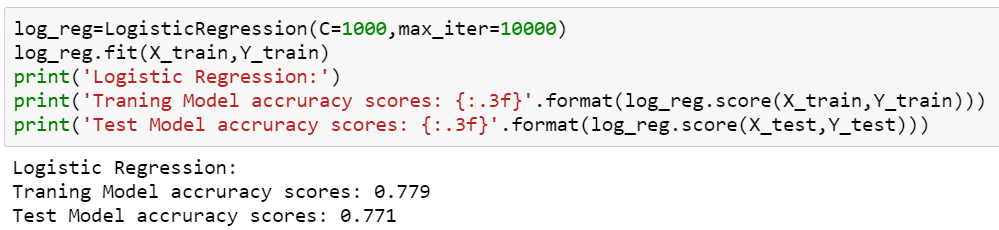
* Let’s split the dataset into training and testing data



Importing Machine learning libraries

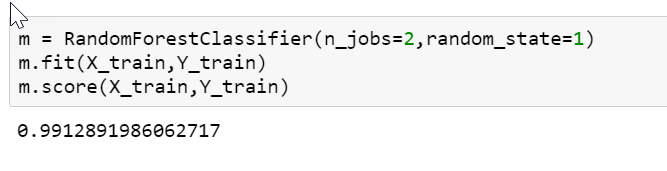


Creating a model using Logistic Regression

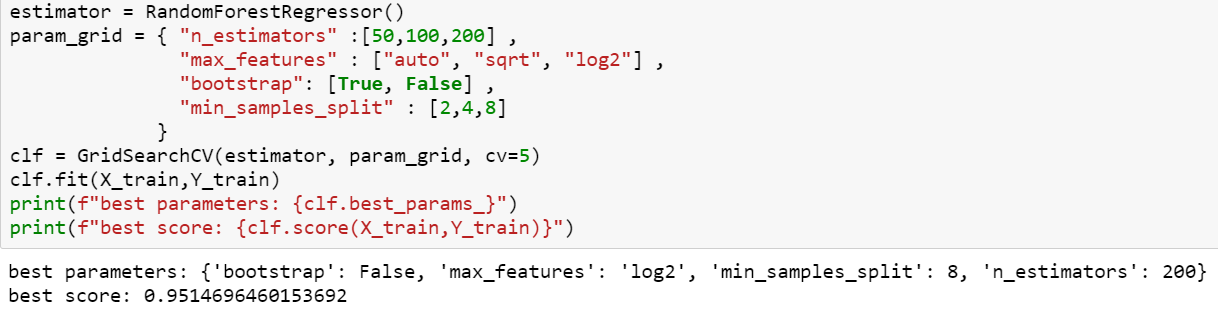


77% accuracy. Not bad for our first model but let us try creating a model through other methods and see if they do better than this.

Creating a model using Random Forest Classifier

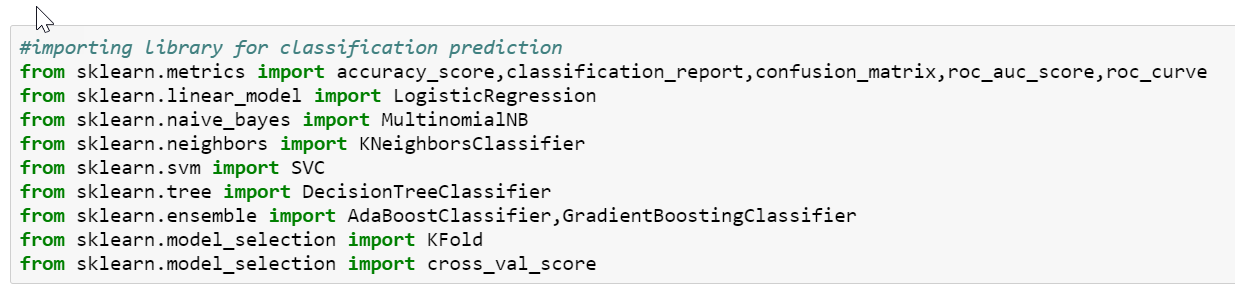


99% accuracy!!! That’s a good one. But to avoid overfitting of our model, we will perform hyper tuning using Grid Search CV.

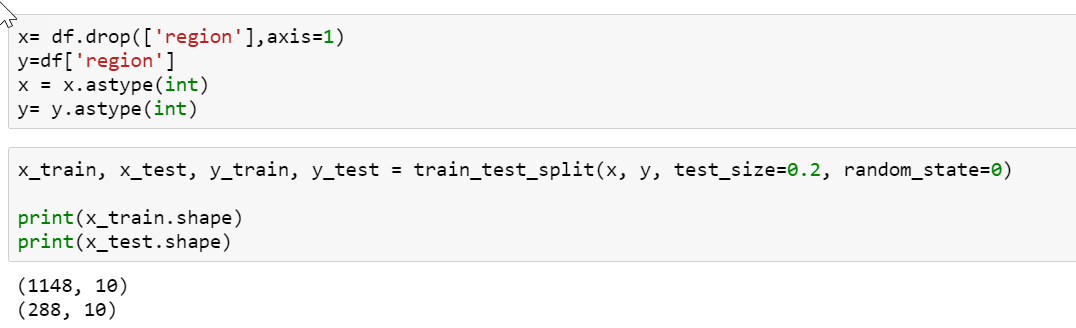


Let’s create our second model now for predicting the region or avocado.

* Importing libraries for creating our machine learning model for predicting region



* Let’s split the dataset into training and testing data



* We will now create multiple models at once and verify their accuracy at same time. This will save time and is a better way for testing out multiple models when you are not sure of which model to utilize for your dataset.



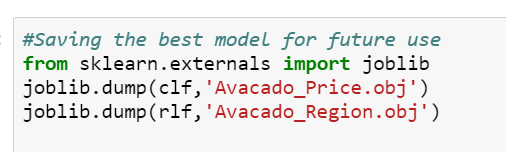
The clear winner here is the model created using Random Forest Classifier method with accuracy of 91%.

Now let’s hyper tune our model using Grid Search CV to avoid overfitting or underfitting.



**Saving the Models**

We will be using joblib library to save our models which can be used later with any other data or production environment.



**Concluding Remarks**

We have successfully created two models –

1. Model for predicting prices – with an accuracy of 95%.
2. Model for predicting region – with an accuracy of 100%.

Prediction of price, region and other factors for Avocado is of huge impact specially in a county like United States where the revenue generated by this fruit is so impactful on the economy. Machine learning will help the companies in planning for storage, selecting locations and setting competitive pricing and stay ahead of their business in the long run.