

A step wise approach to create a machine learning model for Avocado dataset

A Classification & Regression model building approach



Table of Contents

[Introduction & Background 2](#_Toc118326619)

[Problem Definition. 3](#_Toc118326620)

[Exploratory Data Analysis. 4](#_Toc118326621)

[Initial Data Analysis 4](#_Toc118326622)

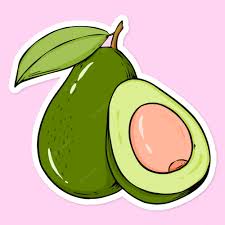
[Descriptive Statistics of numerical data 6](#_Toc118326623)

[Data Visualization 8](#_Toc118326624)

[Pre-Processing Data. 17](#_Toc118326625)

[Building Machine Learning Models. 24](#_Toc118326626)

# Introduction & Background



*It is a well-known fact that Millennials LOVE Avocado Toast. It's also a well-known fact that all Millennials live in their parents’ basements.*

*Clearly, they aren't buying home because they are buying too much Avocado Toast!*

*But maybe there's hope… if a Millennial could find a city with cheap avocados, they could live out the Millennial American Dream.*

The aim of this blog is to help you understand how to create a Machine learning model & what are the necessary steps to predict the regions based on the dataset available at [this link](https://raw.githubusercontent.com/santoshhulbutti/-DataTrained_Evaluation_Projects/main/Avacado%20Project/avocado.csv).

*(This blog is written with the assumption that the reader has preliminary understanding of machine learning terminologies & basic libraries machine learning libraries & intermediate level of understanding of python programming language.)*

We will understand all the necessary steps needed to clean & manipulate data & build ML model. So, lets go & dive into it…

# Problem Definition.

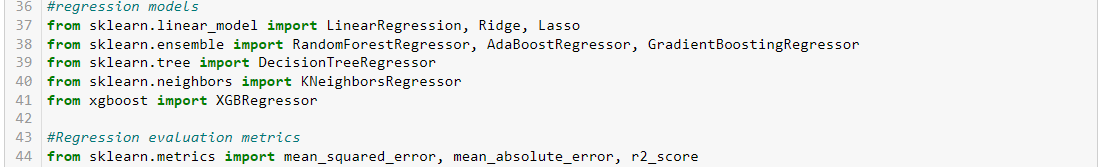
The dataset represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the dataset reflects an expanded, multi-outlet retail data set. Multi-outlet 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 dataset.

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

We will create one Classification model for predicting the region where avocados are sold & one regression model for predicting the average price for each avocado, type, year, volume, bag size and other features in the dataset.

We will import following machine learning libraries necessary for the model creation & validation;



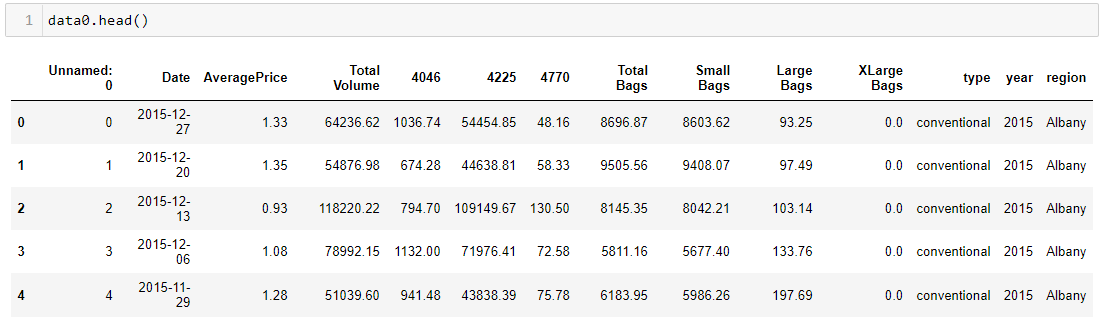
# Exploratory Data Analysis.

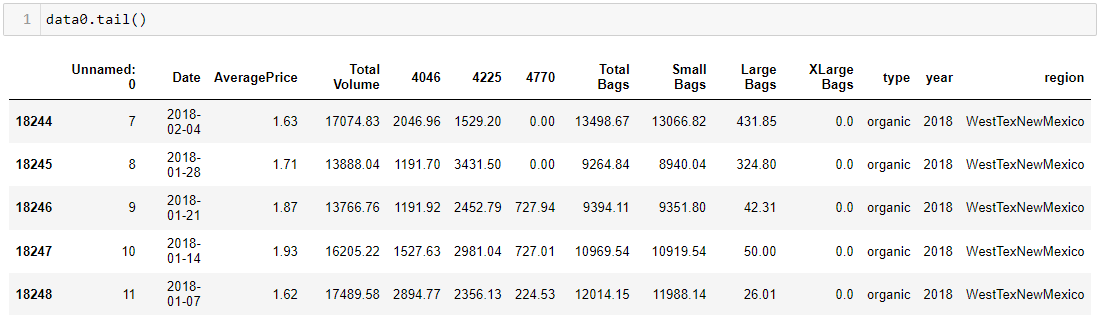
## Initial Data Analysis

* We will load the data using pandas ‘**pd.read\_csv()**’ method.



* Reading the first & last 5 rows of data

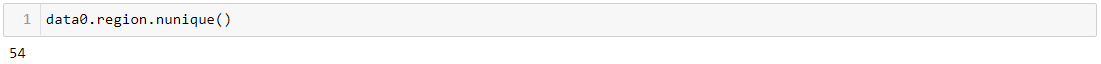




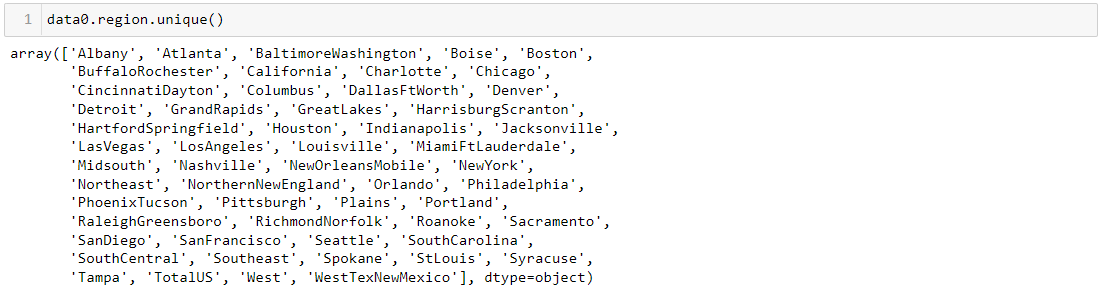
**Observation:**

Here we see there are total 14 columns, out of which ‘region’ is our target/label column or dependent variable. We also see that there is a column named ‘Unnamed:’, it is an observation number from a specific region (you can check this using data0.head(55) & data0.sample(15)). Observations are taken on every 7th day from the last observation. We will remove this column as it brings no meaning our model.

Let’s understand our label column,



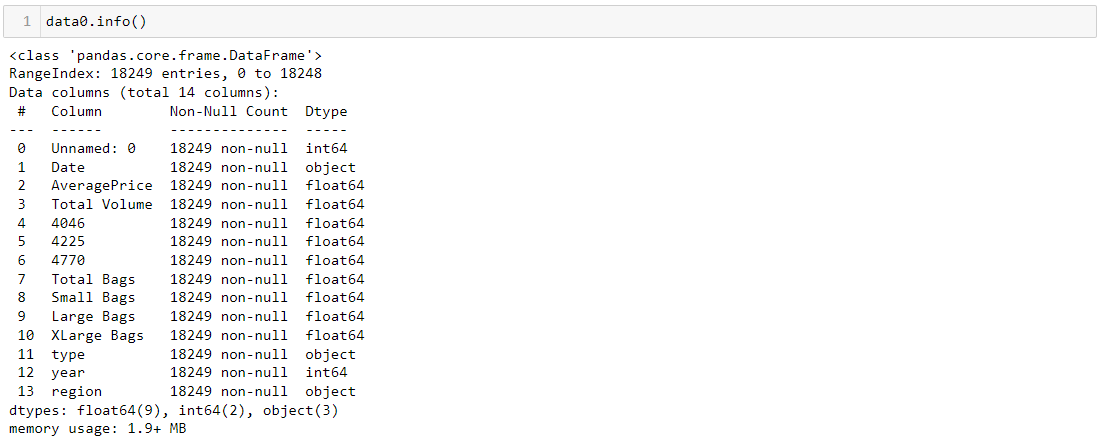
There are total 54 unique entries in target column, following are the values;



**Observation:**

From above we see that, amongst the city name we also see there is an entry as ‘TotalUS’. this means some observations were entered using total of all the regions in US. These entries should be dropped as we are predicting individual regions.

Going ahead with the next step we will check missing entries, data types, total number of unique entries in all features, column names & data size;

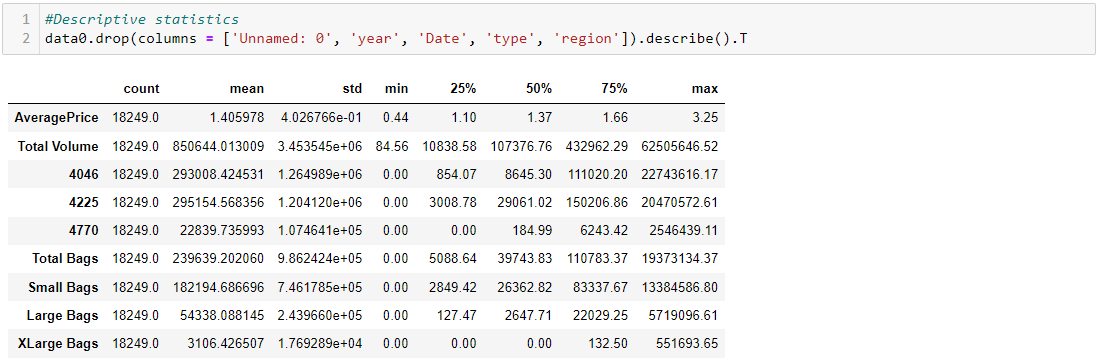




**Observation:**

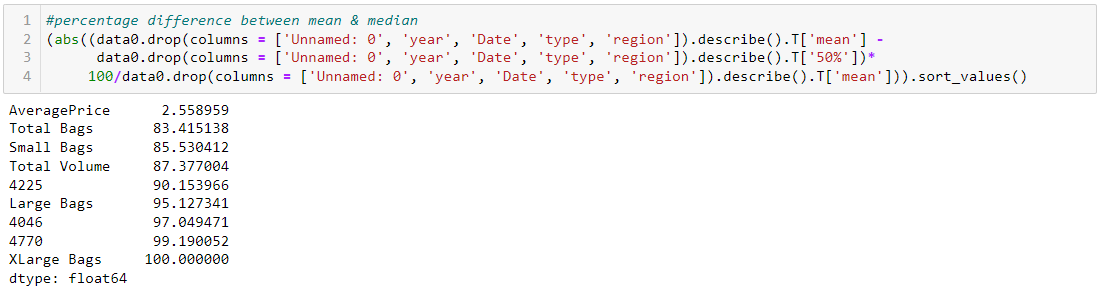
* There are total **18249 observations** in dataset & **14 columns**, & at initial level there are no missing entries.
* **Date** is formatted as object datatype, we will convert it into week number & month number, which will be helpful for visualization. The data was gathered during 169 unique days spanned over **4** **year**s.
* **Average price** is continuous type of numerical data.
* **type** is an object data type & it has 2 unique entries.
* type, year & region are the 3 categorical features from the dataset. from this we will remove year column as date is also there in the dataset.
* except Unnamed: 0, Date, type year & region all other features are float type continuous data.

## Descriptive Statistics of numerical data



**Observation:**

* We see no missing observation from count. every feature has 18249 number of entries.
* There are some entries/observations where the minimum value of features was 0.0.



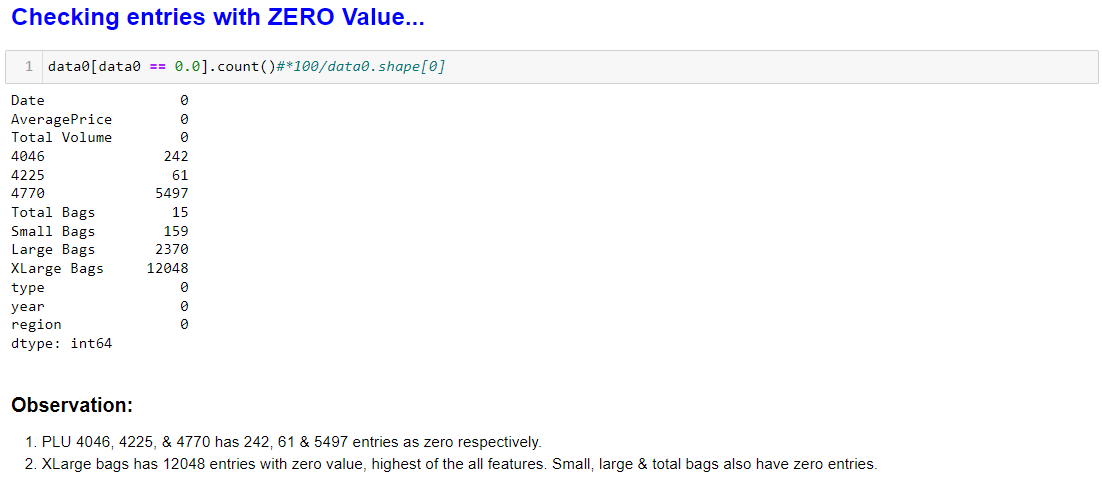


**Observation:**

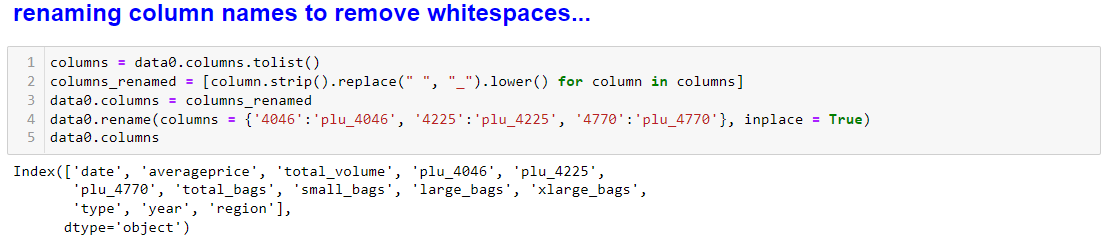
* The mean & median % difference for all columns is more than 50% implying skewness in the dataset, except AveragePrice.
* features with % difference between 75% quantile & maximum value, more than 99% indicate possible outliers.

Dropping the ‘Unnamed: 0’ column & checking for duplicate & checking zero entry count.

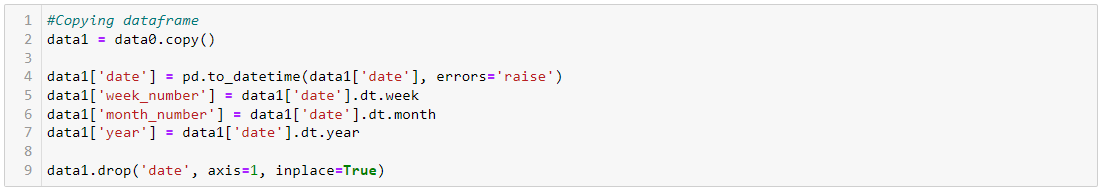




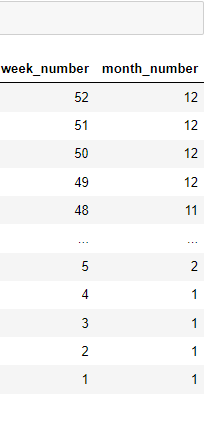
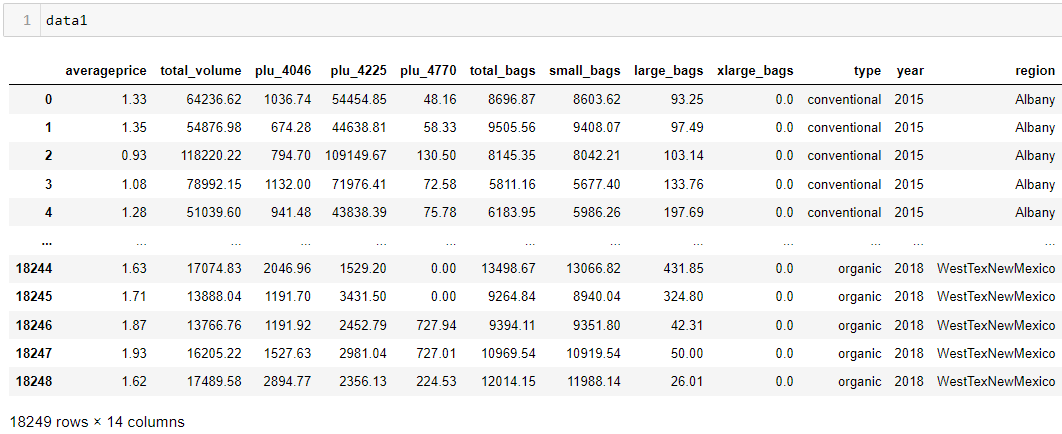
We will rename all the columns to be able to easily understand & remove whitespaces. (This step is not necessary but I like to do it, whenever I see whitespace in column names)



Now we will copy the input data frame & do feature engineering for Date column on the copied dataframe as follows:

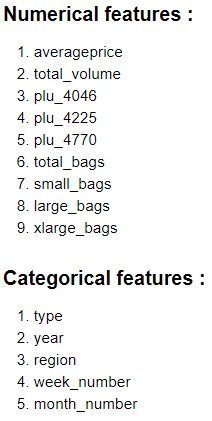


The new dataframe looks like as follows:

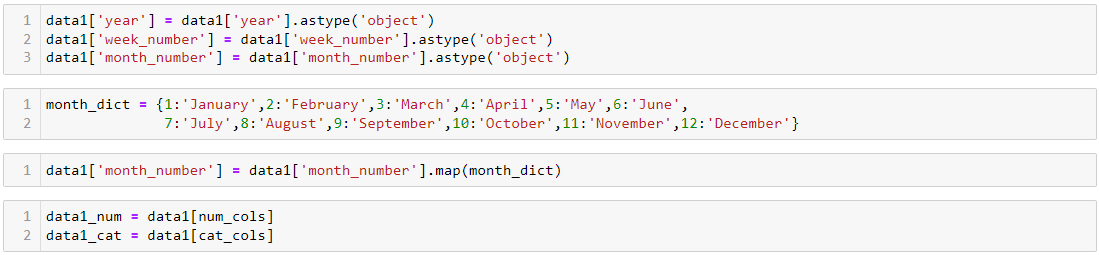


## Data Visualization

We will make two group for data visualization purpose;

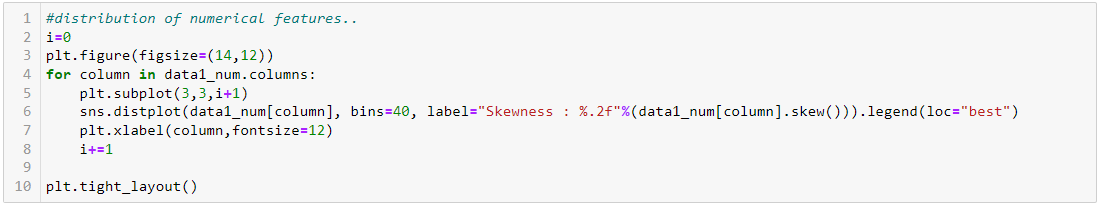


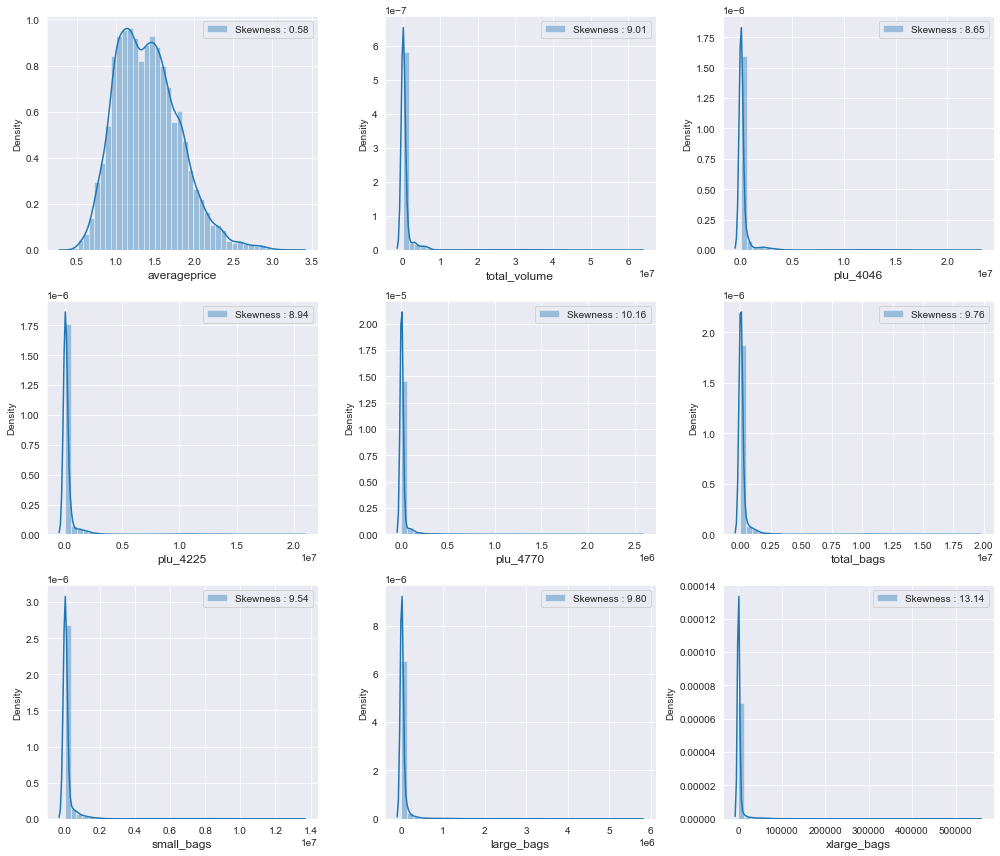
Converting Some of the features in to proper data type for visualization purpose. & Creating two dataframes for numerical features & categorical features;



### Univariate Analysis

For Numerical features:

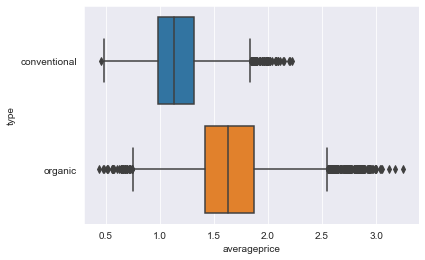




**Observation:**

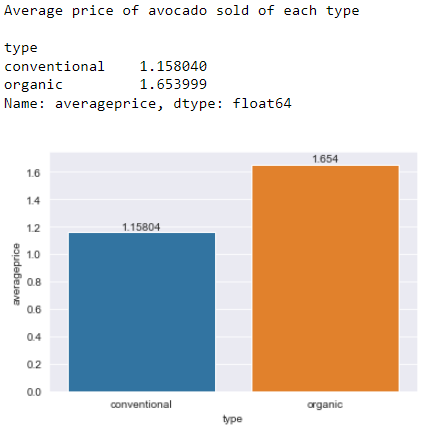
* + All the numerical features except average price are skewed & it is evident from the above distribution plots.

For categorical features:



**Observation:**

* + We can say that Conventional type Avocados are cheaper than the Organic type of Avocado.
  + Organic type of avocado has average price ranging from 0.4 to 3.3, with mean value close to 1.65.
  + Conventional type of avocado has avg price within 0.4 to 2.4, & mean being closer to 1.15.

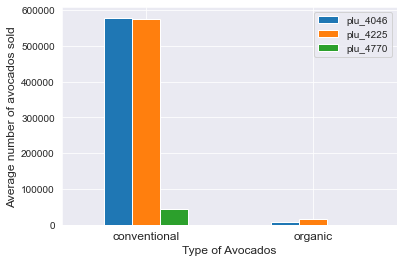


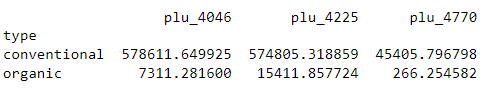
**Observation:**

* + An organic type of Avocado is about 50 cents costlier than conventional type.

### Bivariate Analysis

1. Average number of avocados sold based on product lookup codes of each type;

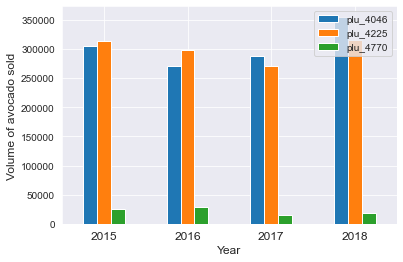




**Observation:**

* + Mostly Conventional type of avocados with package label 4046 & 4225 are sold.

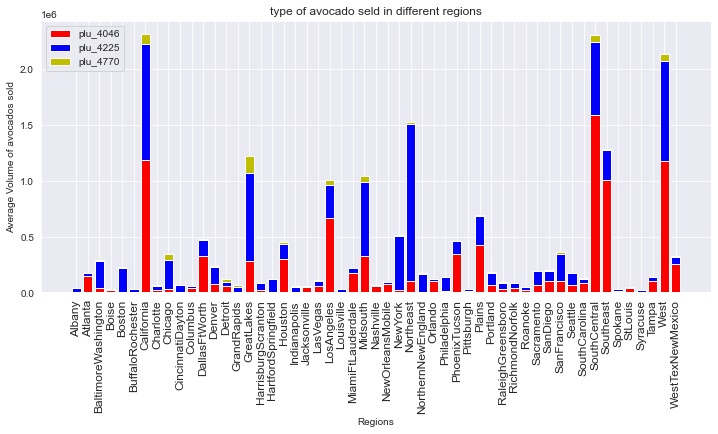
1. Average number of avocados sold based on bags size of each year



**Observation:**

* + In all year of the data available, we see Conventional type of avocados with package label 4046 & 4225 are sold most in terms of volumes.

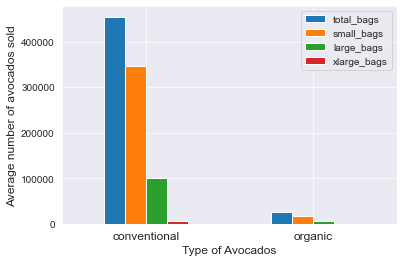
1. Average volume of avocado sold in different regions based on package label type;



**observation:**

* + California, South central, South east & west regions are the regions with most consumption of avocados. i.e., higher demand from these 4 regions.

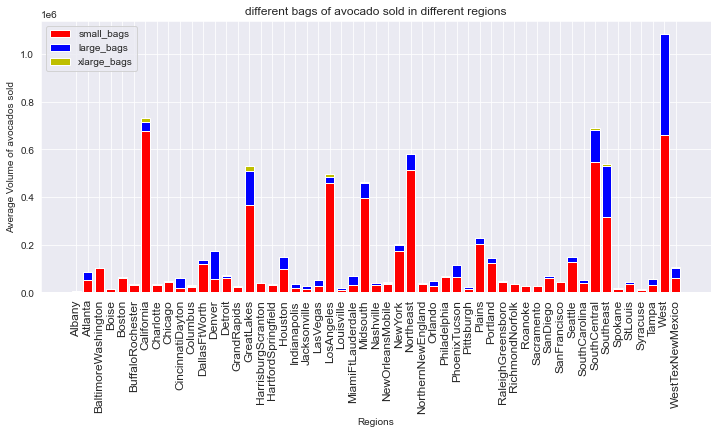
1. Average number of avocados sold based on bags size of each type;



**observation:**

* + Organic type of avocado is very less consumptive but higher price.
  + Conventional type of avocado is sold heavily with higher volume & lower price.
  + Smaller bags of conventional avocado is sold in much bigger volume than other bag sizes.

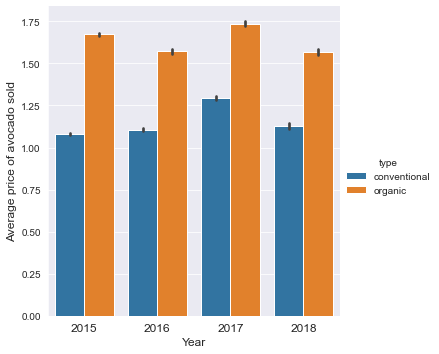
1. Different bags of avocado sold in different regions based on bag type;



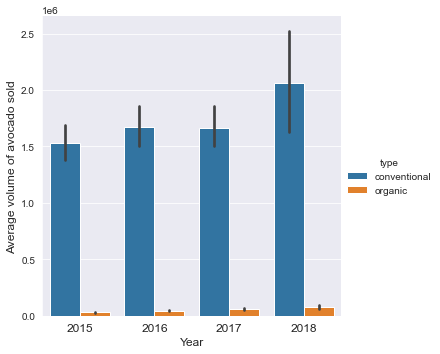
**Observation:**

* + Smaller bags of avocados are sold heavily in all the regions.

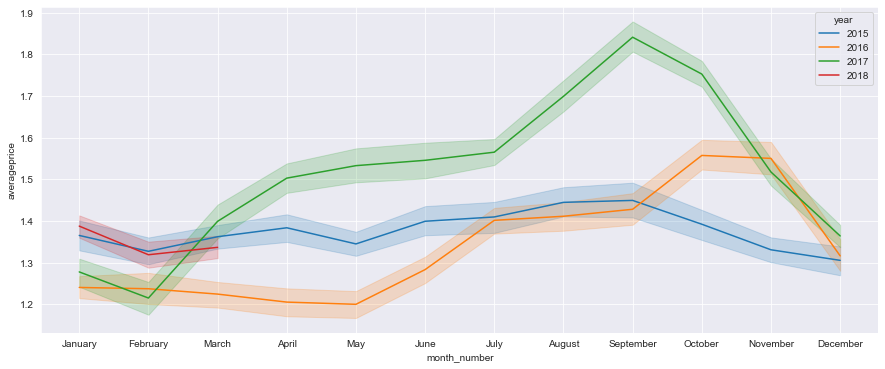
1. Average price of each type of avocado sold each year;



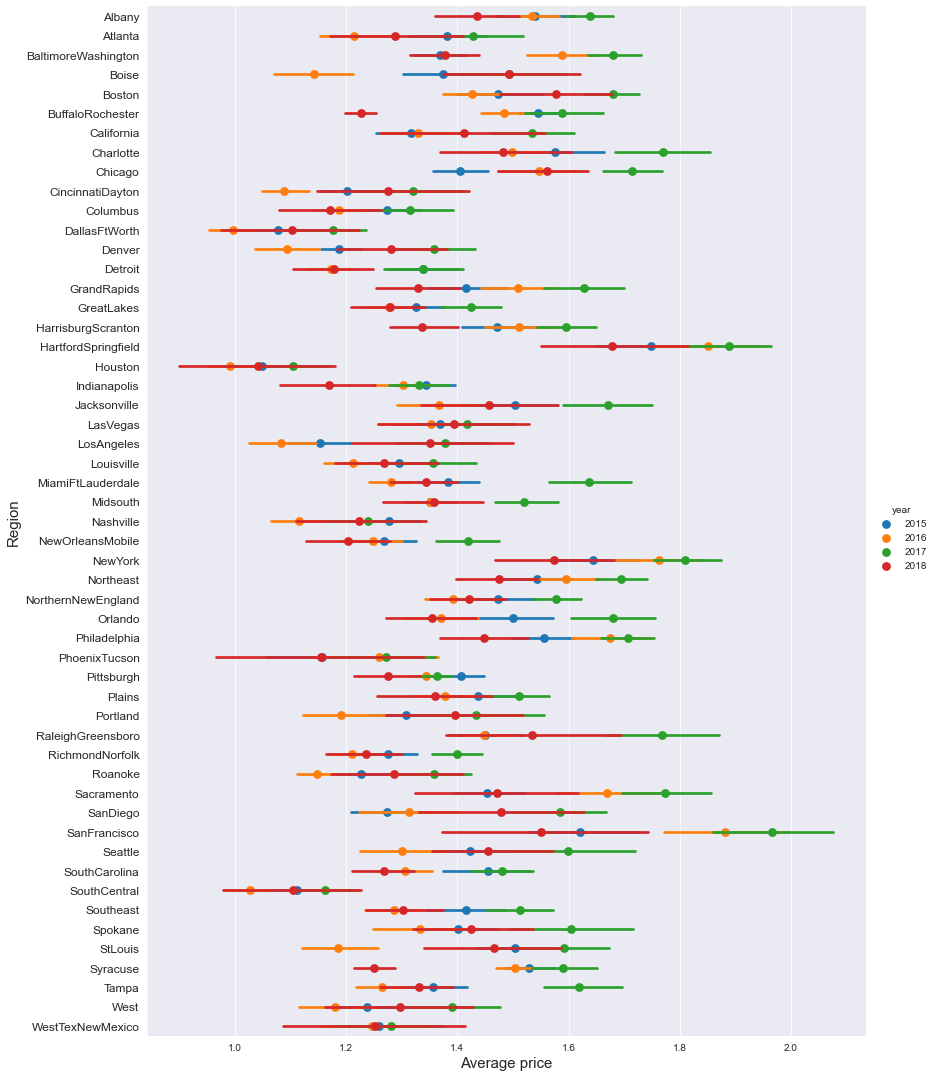
1. Average number of each type of avocado sold each year;



1. Average price variation during each year



1. Average price variation in all regions during each year



In 2015:

* average price of avocado was higher in Hart fort Springfield, New York, San Francisco, charlotte, philly. price was above 1.5 per unit.
* Houston, Dallas. South Central & LA had lower price per unit.
* Volume consumption was higher in California, West & South-central regions

In 2016:

* average price of avocado was higher in Hart fort Springfield, New York, San Francisco, philly, Sacramento & Northeast regions. price was above 1.6 per unit.
* Houston, Dallas. South Central & LA had lower price per unit.
* Volume consumption was higher in California, West & South-central regions

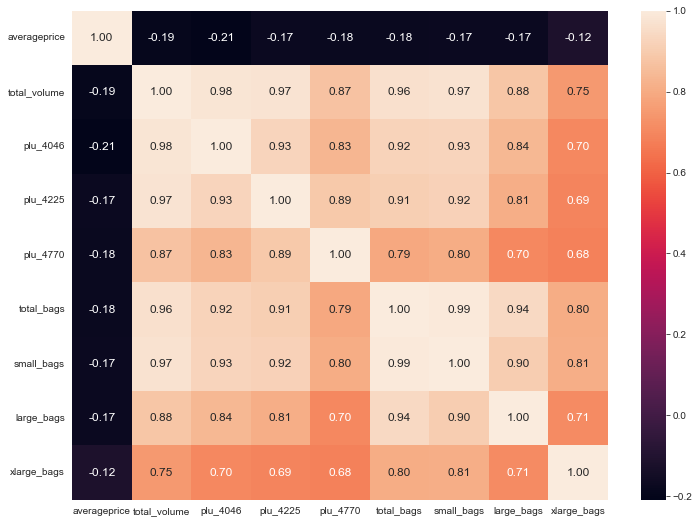
In 2017:

* average price of avocado was higher in hart fort Springfield, New York, San Francisco, charlotte, Sacramento regions. price was above 1.75 per unit.
* Houston, Dallas. South Central & Nashville had lower price per unit.
* Volume consumption was higher in California, West & South-central regions

In 2018:

* Data was only of 3 months, so it won’t make sense to compare with other years data.

1. Correlation matrix of numerical features;



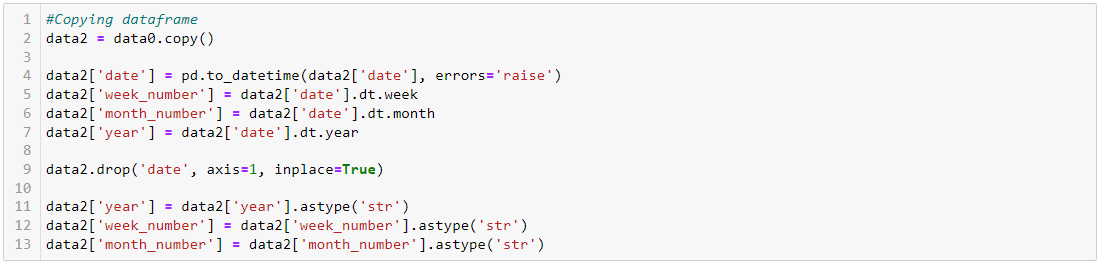
**Observation:**

* + All Numerical features seem to have very low correlation with target variable 'Average Price' & high correlation with each other, hinting about multicollinearity.

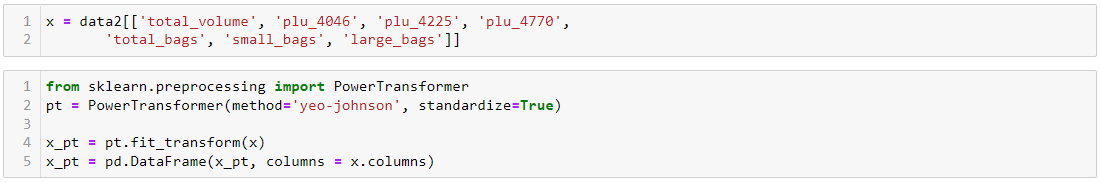
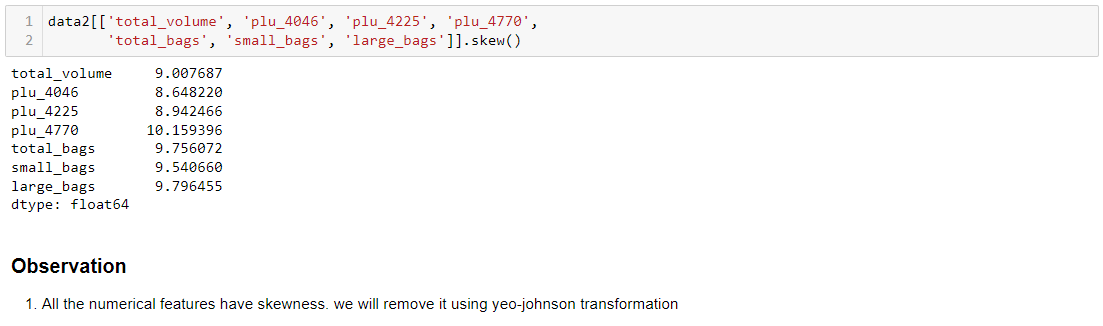
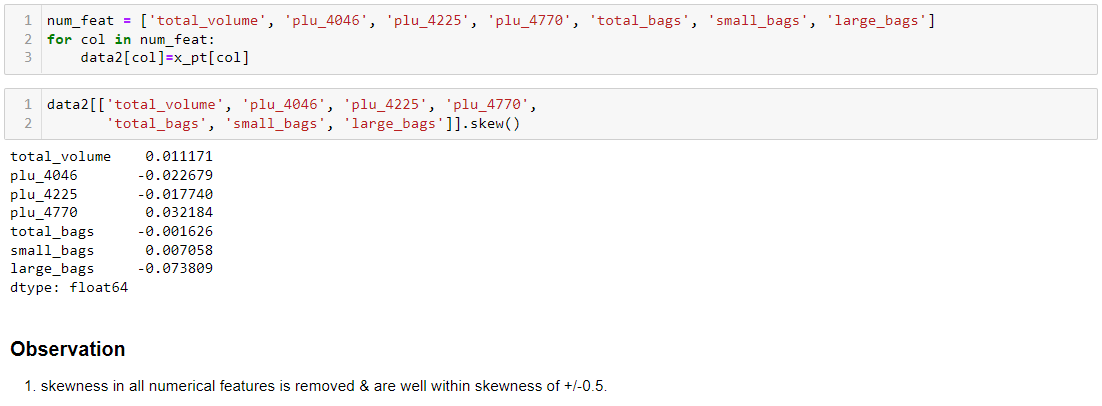
# Pre-Processing Data.

1. In this step we will make week number, month number & year as categorical datatypes.

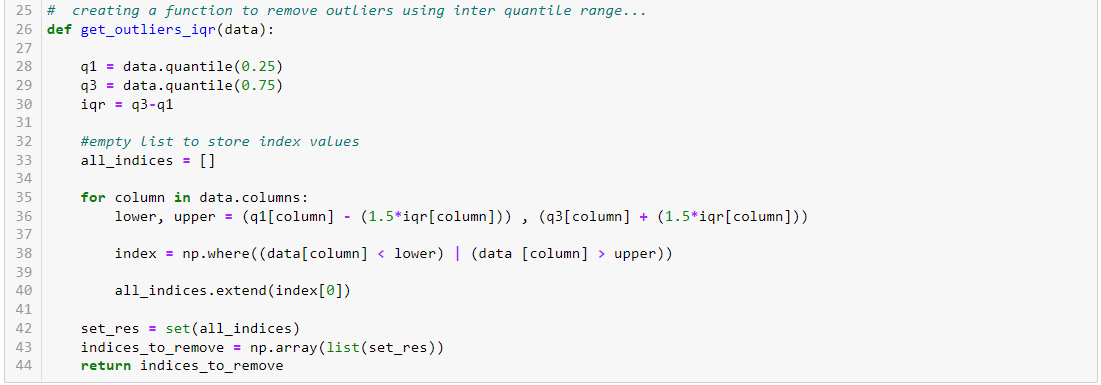
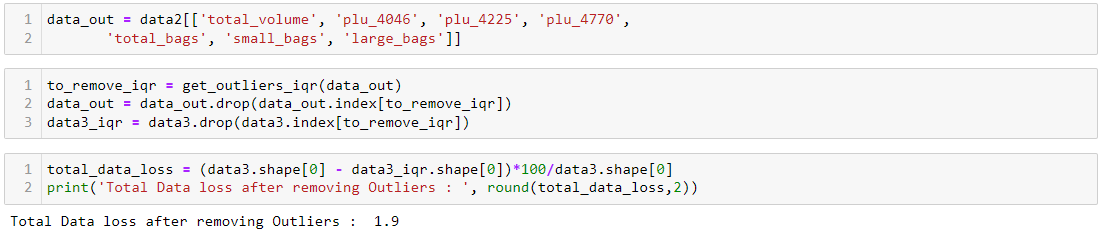
Copying the dataframe & converting all the necessary features in to proper datatypes;



1. Checking Skewness & Removing using Power transformation

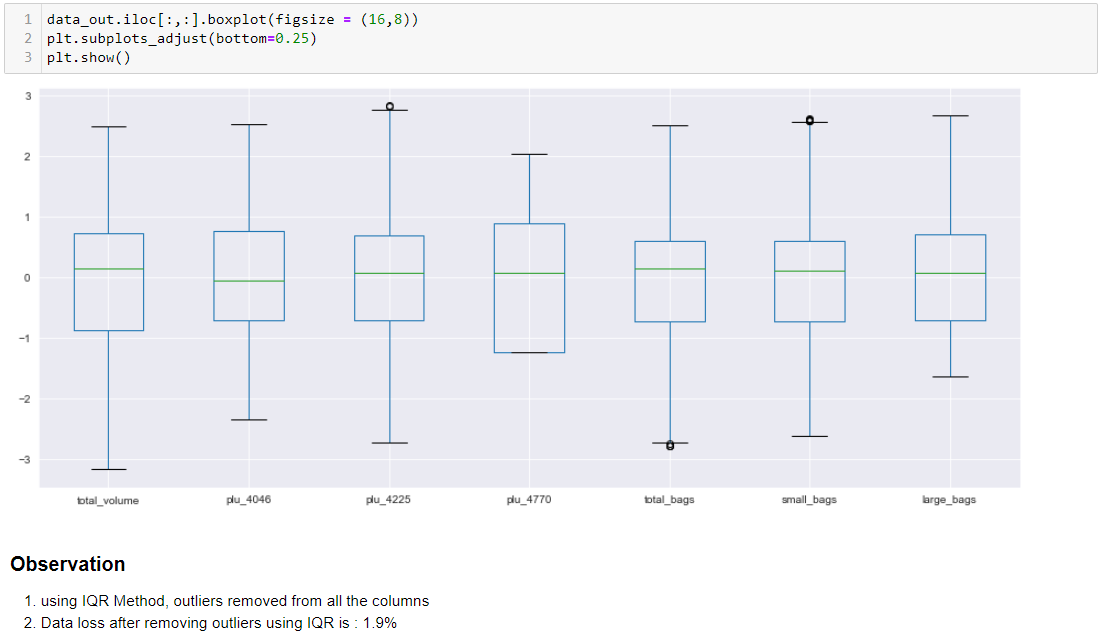
 

1. Checking outliers & removing if any using IQR Method

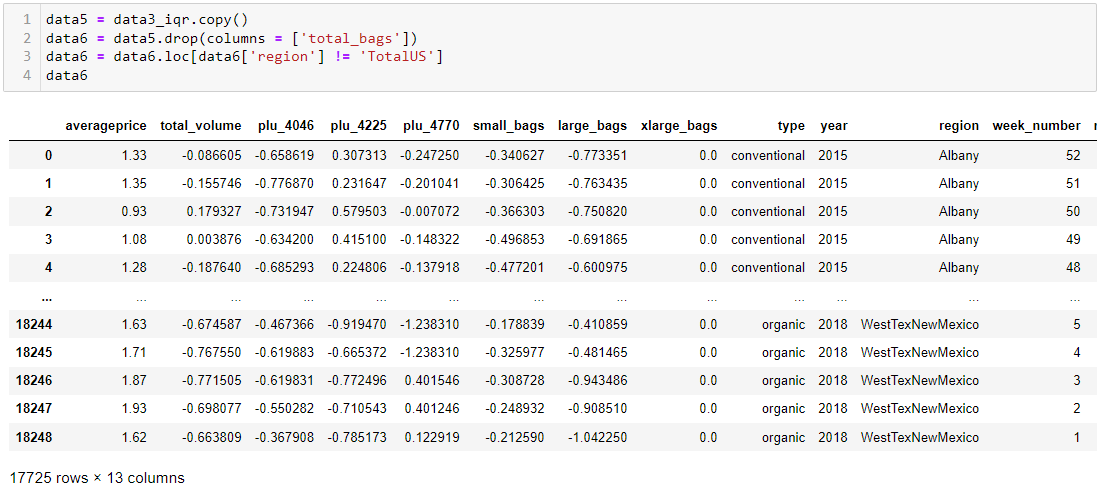
 

**Observation:**

* + We see that total data loss after removing outliers was 1.9%.

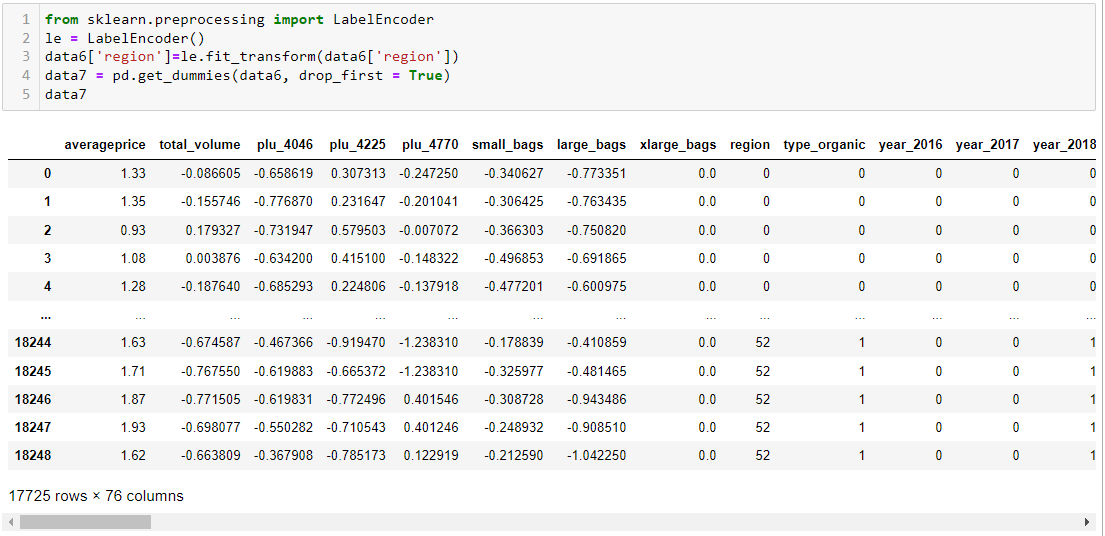


1. We will Remove all the observations where region column has ‘TotalUS’ entry;

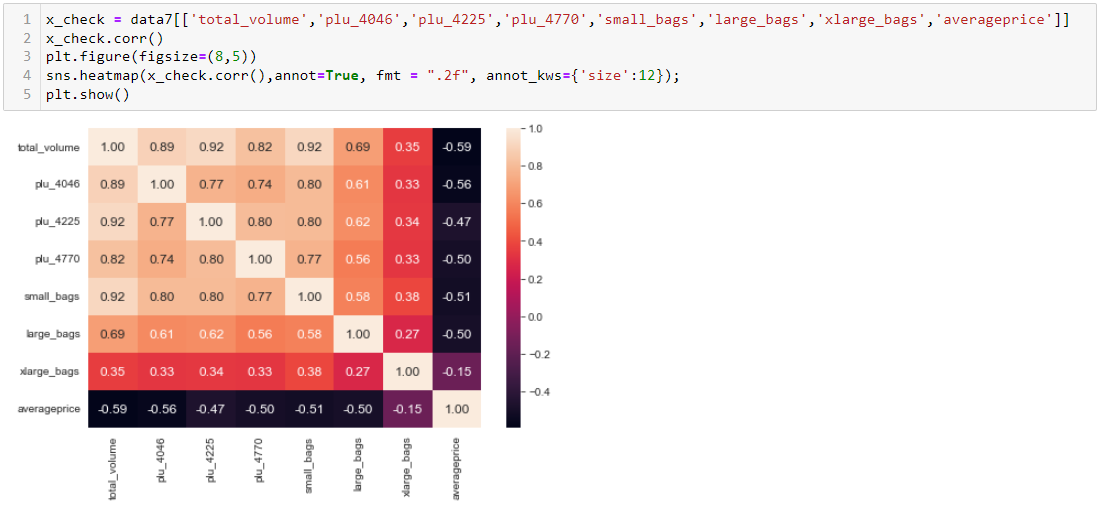


1. Using Label encoder for encoding region column.

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.

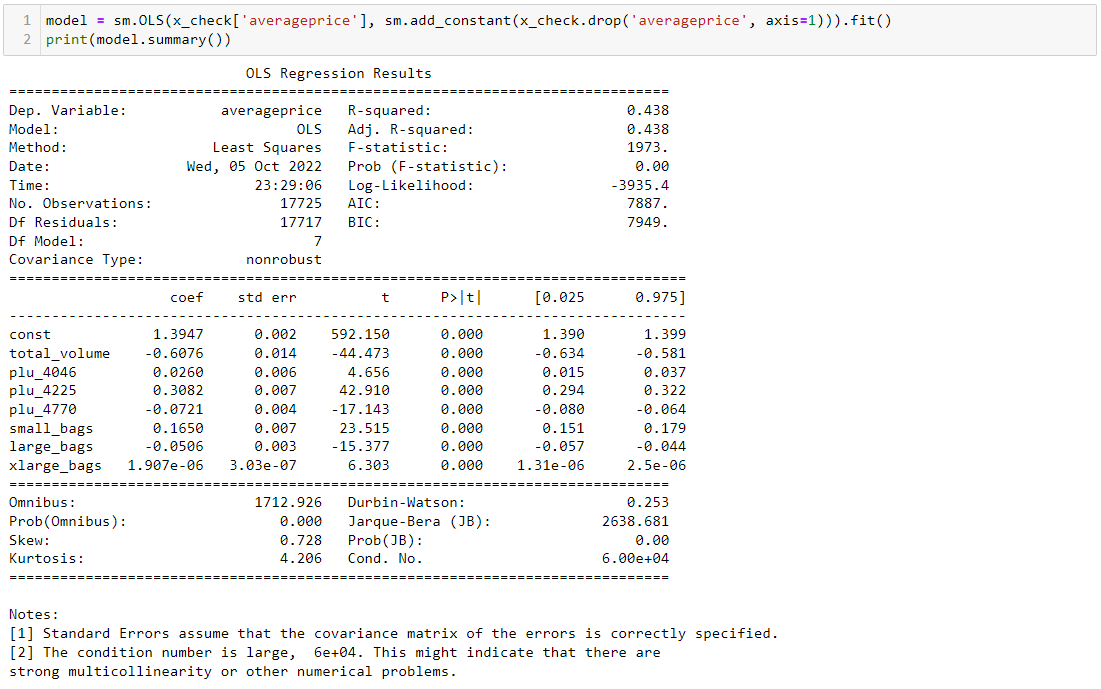


1. We will check for correlation & Multicollinearity using .corr() method & VIF resply.



**Observation:**

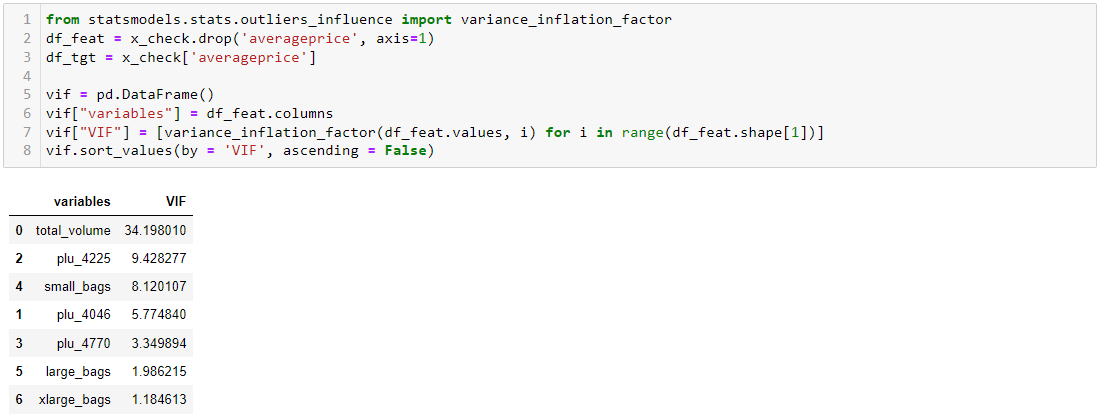
* + We see there are some features with highly correlated with each other, We will confirm the same using Statsmodels’ OLS result as follows.



**Observation:**

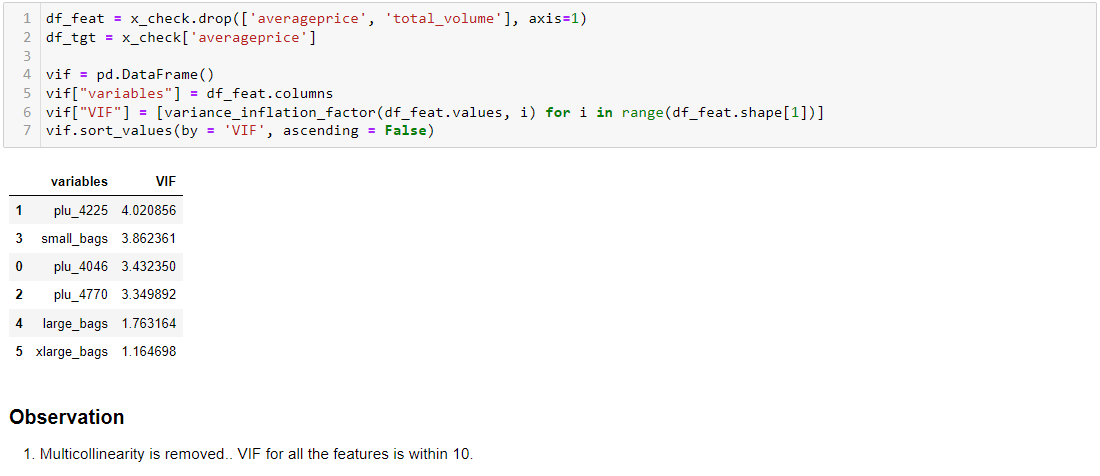
* + Condition number is more than 10, this indicate there is multicollinearity issue.

1. We will remove multicollinearity using Variance inflation factor



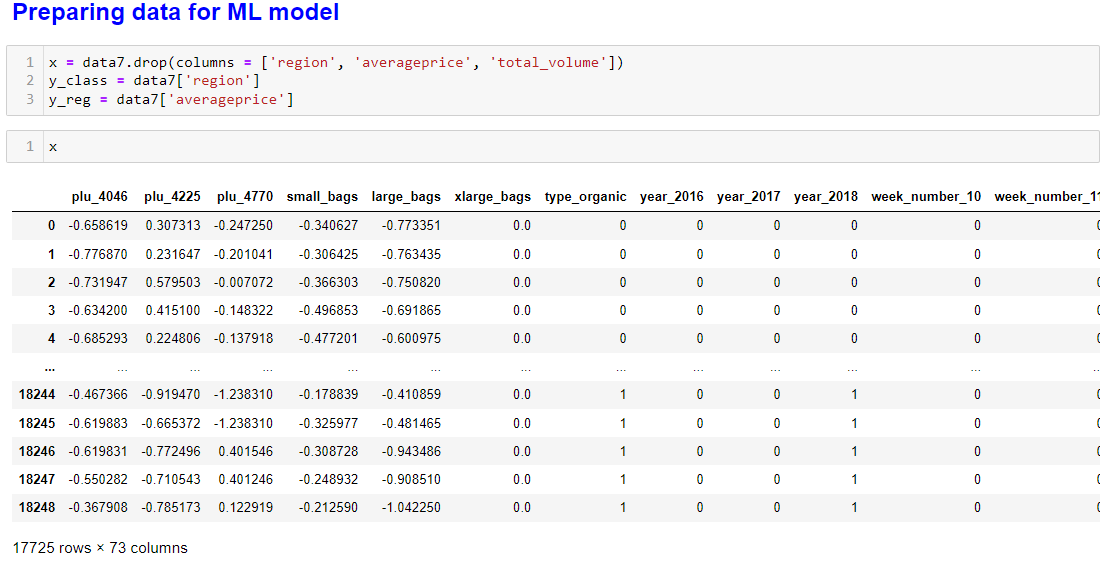
**Observation:**

* + We will remove total volume as it shows VIF more than 10.

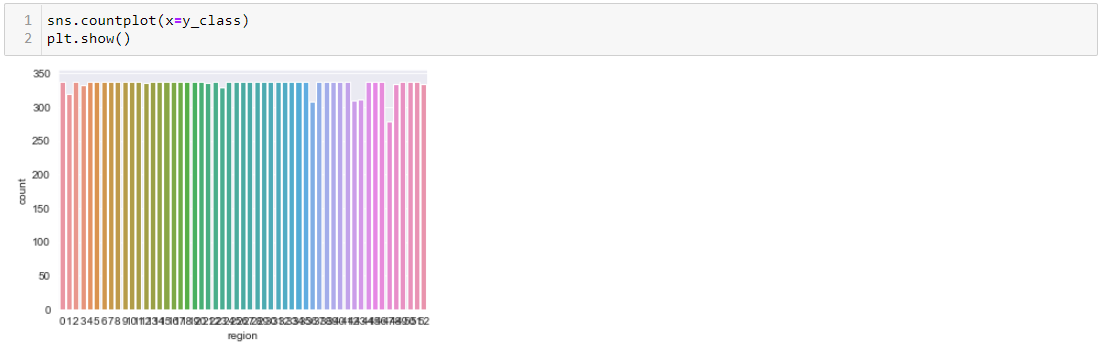


1. Final data preparation for model training.

We will separate dependant & independent features.

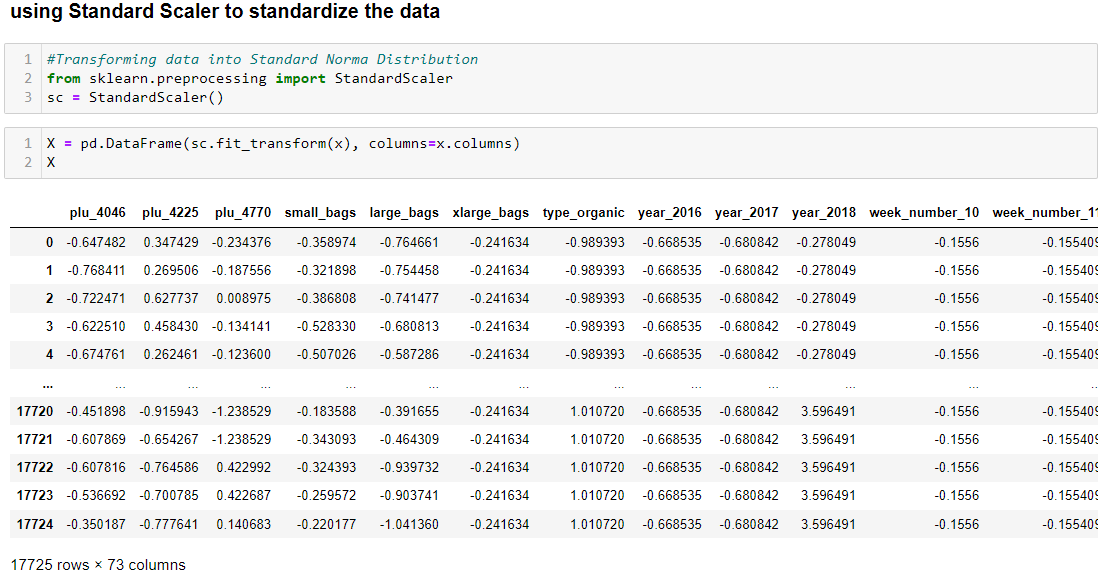


1. Checking data imbalence for classification problem



No need to do use SMOTE method as the data is perfectly balenced.

1. Applying Standard Scaling to Independent features

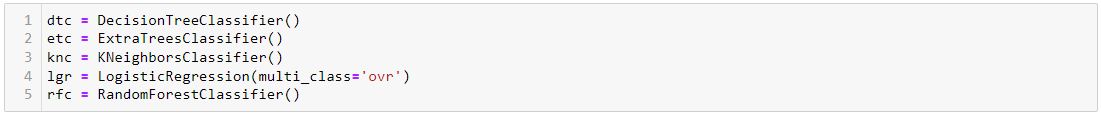
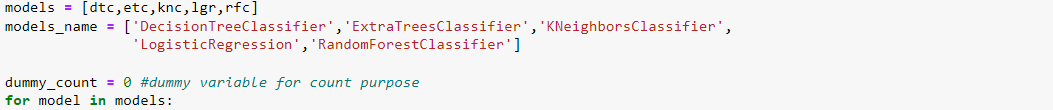
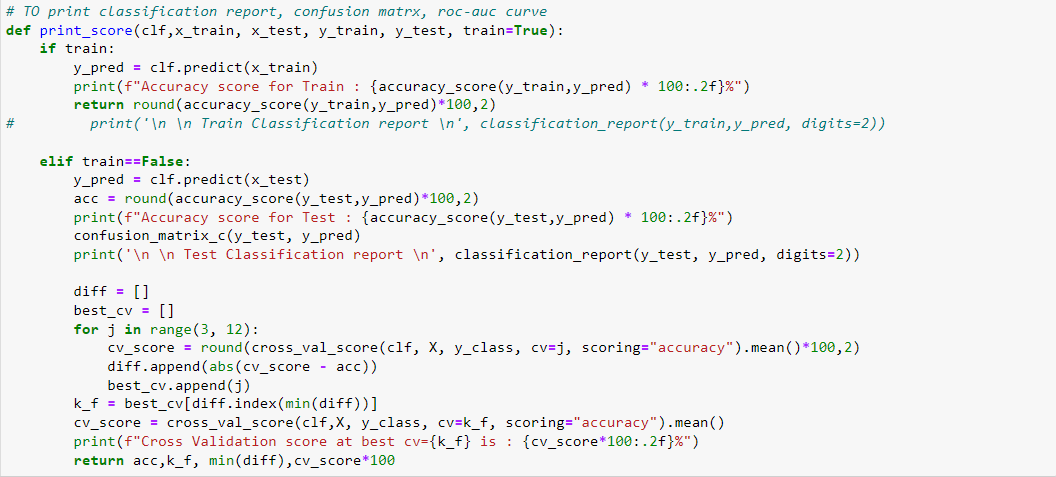
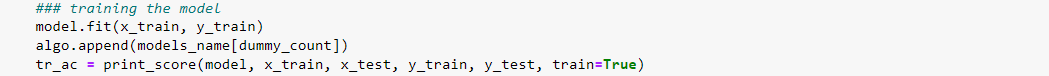


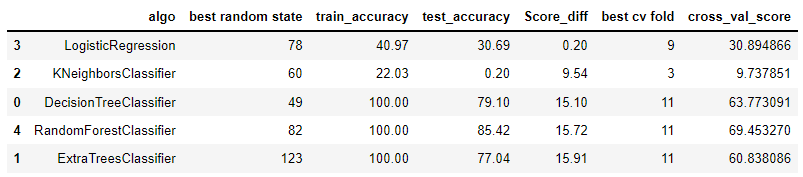
# Building Machine Learning Models.

Please refer the notebook in the [github link](https://github.com/santoshhulbutti/-DataTrained_Evaluation_Projects/blob/main/Avacado%20Project/Avacado%20Project_draft.ipynb) for detailed codes:

For Classification model;

We will test the following models (the detailed code has all algorithm run in a loop to get best random state & best cross validation fold). The code will generate a table consisting best random state & Cross validation fold & all related evaluation metrics of the respective model.

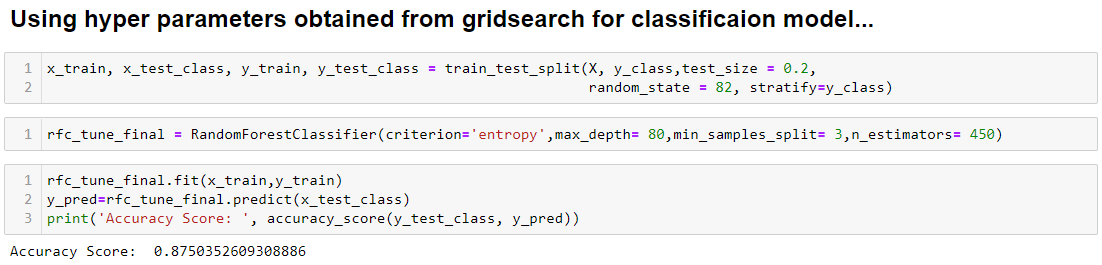
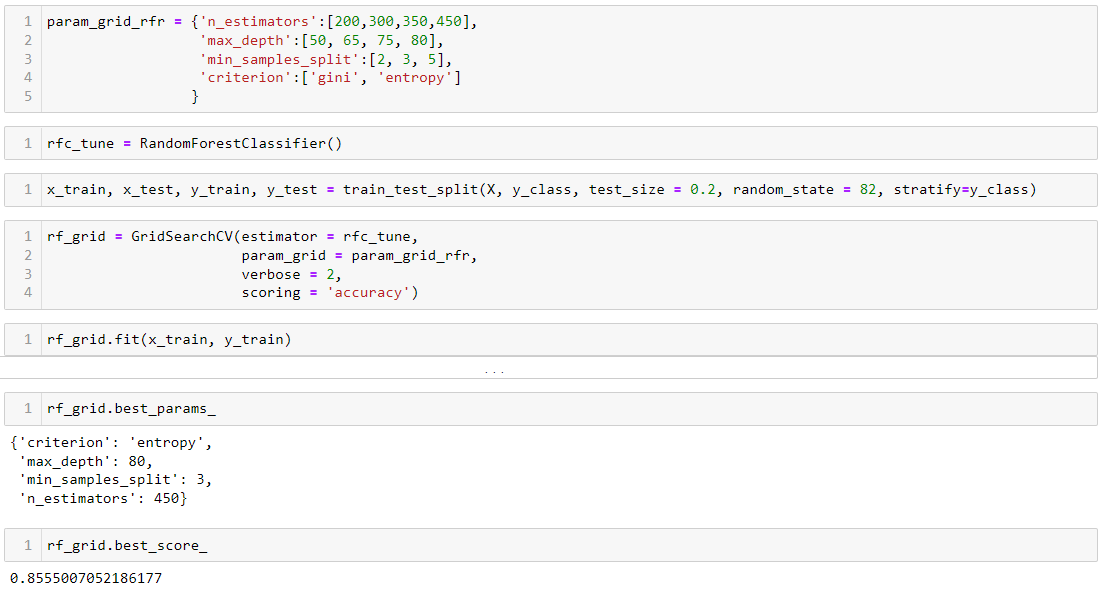
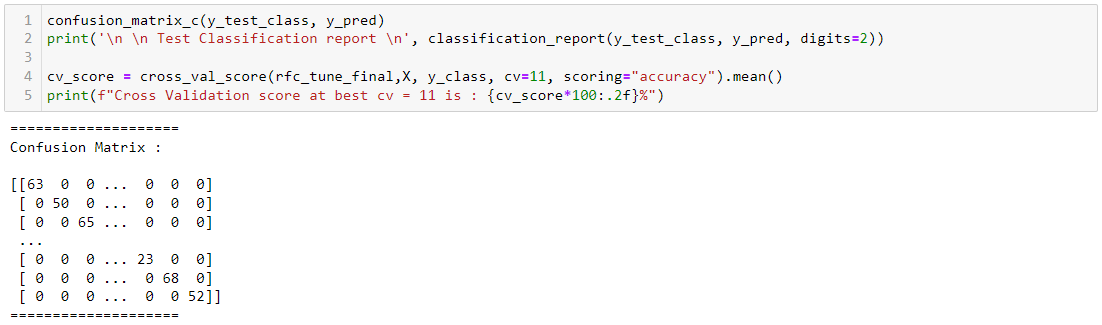
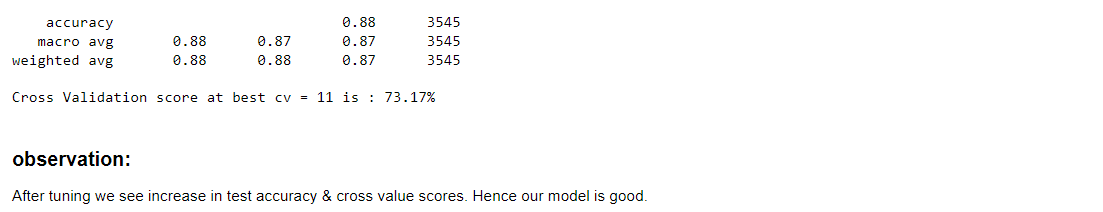
  

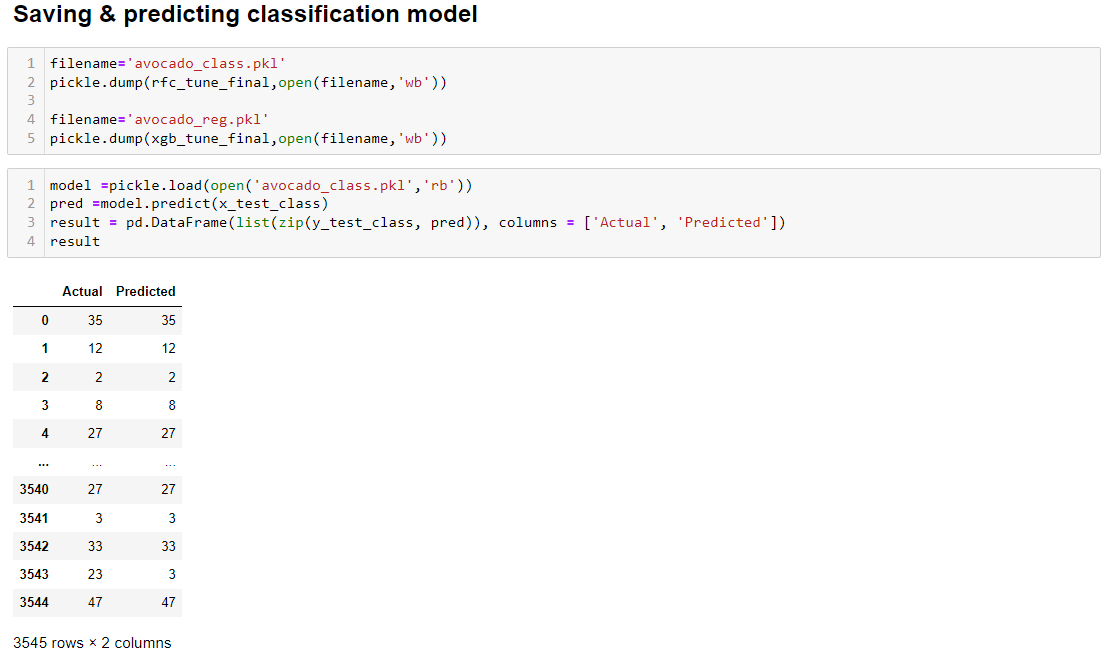


**Observation:**

* + we will select Random Forest classifier as it has higher test accuracy & lower difference between test accuracy & Cross value score amongst all models.
  + It also has higher test accuracy of all the models tested.

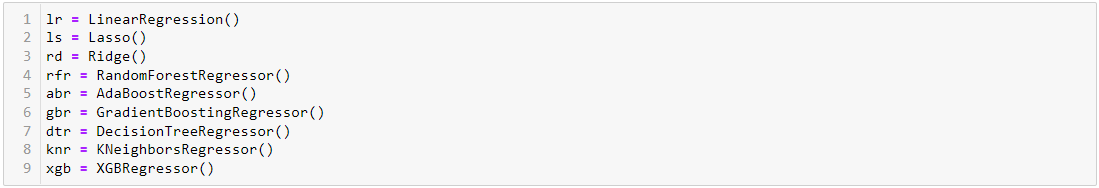
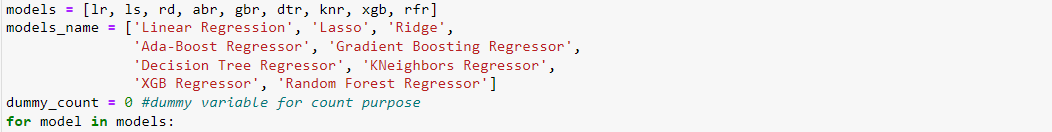
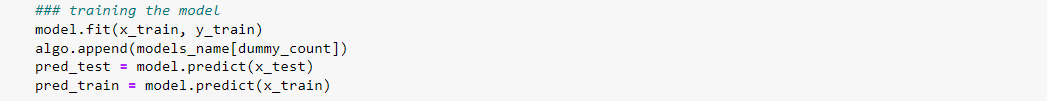
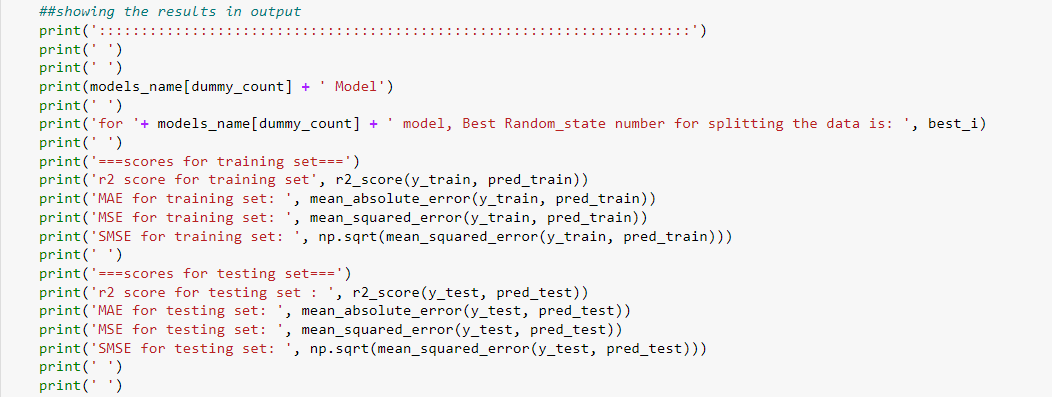
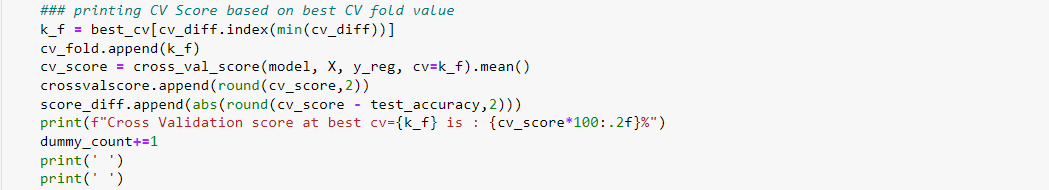
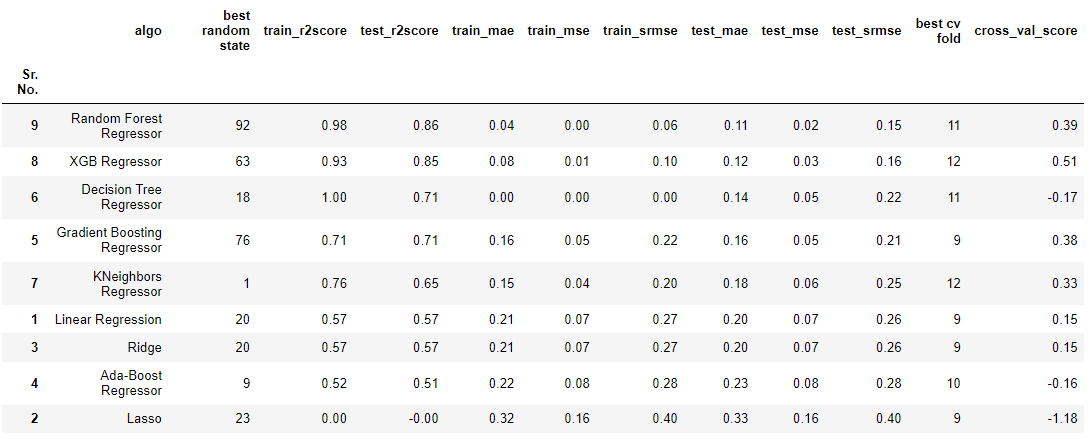
Hyper parameter tuning for classification model;



For Regression model;

We will test the following models (the detailed code has all algorithm run in a loop to get best random state & best cross validation fold). The code will generate a table consisting best random state & Cross validation fold & all related evaluation metrics of the respective model.

**Observation:**

* + we will select XGB Regressor as it has higher test & Cross validation scores, also it has lower MAE value.
  + Random forest has higher test score but cross value score is very less. Hence, we will use XGB Regressor.

Hyper parameter tuning for regression model;

