# Avacado Machine Learning Models

# Problem Definition:

Avocado is a fruit consumed by people heavily in the United States.

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV.

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. greenskins) are not included in this table.

Some relevant columns in the dataset:

Date - The date of the observation

Average Price - 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

# Data Analysis

# About Dataset

# 

# There are 18249 rows and 14 columns in avocado dataset.

# Checking for Null/Missing Values

# 

**EDA (Exploratory data analysis):**

Exploratory Data Analysis (EDA) is an approach of analyzing

data sets to summarize their main characteristics, often with

visual methods, a statistical model can be used or not, but

primarily EDA is for seeing what the data can tell us beyond the

formal modelling or hypothesis testing task. we can say that EDA

is statisticians’ way of storytelling where you explore data, find

patterns and tell insights. EDA is a phenomenon under data

analysis used for gaining a better understanding of data aspects

like: - main features of data variables and relationships that hold

between them identifying which variables are important for our

problem

# 

Large Bags and XLarge Bags may contain outliers since there is a huge difference between 75% and max and their respective means and 50% also are significantly different.

A higher std than mean in Total Bags, Small Bags and, XLarge Bags columns indicates skewness.

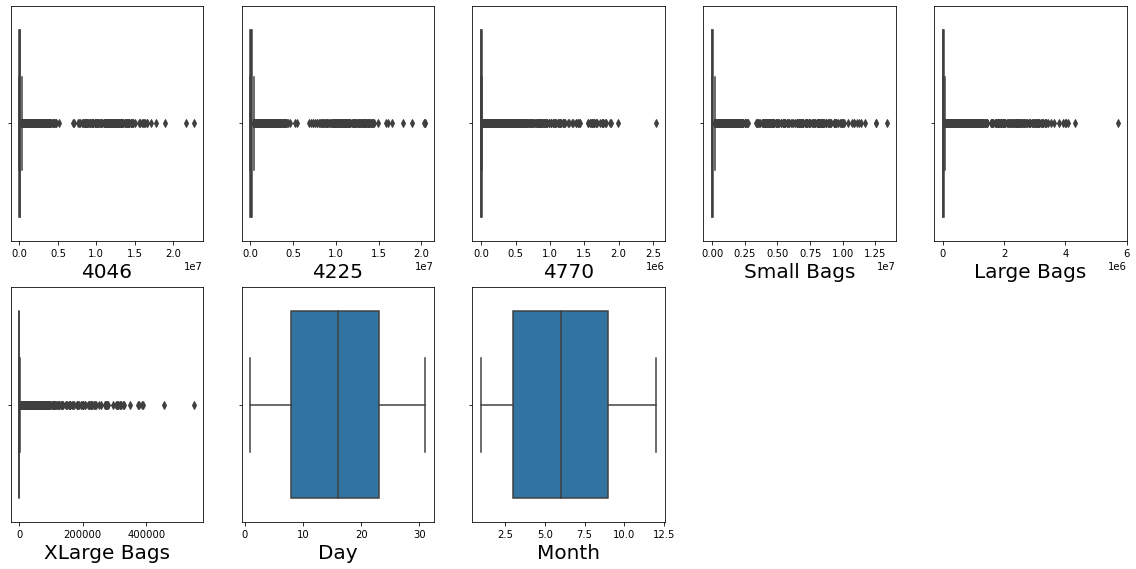
# 

# 

Total Volume, 4046, 4225 show a bimodal data distribution

Total Bags, Large Bags,Small Bags and, XLarge Bags columns have skewed data distribution

**Checking Outliers:**

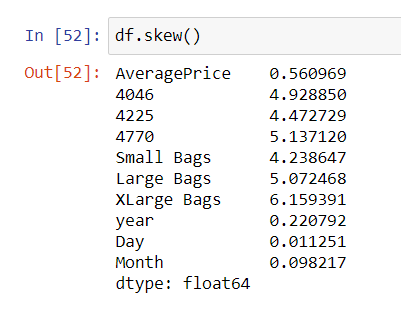


# There are outliers which can be removed by Z SCORE method.

**Skewness**

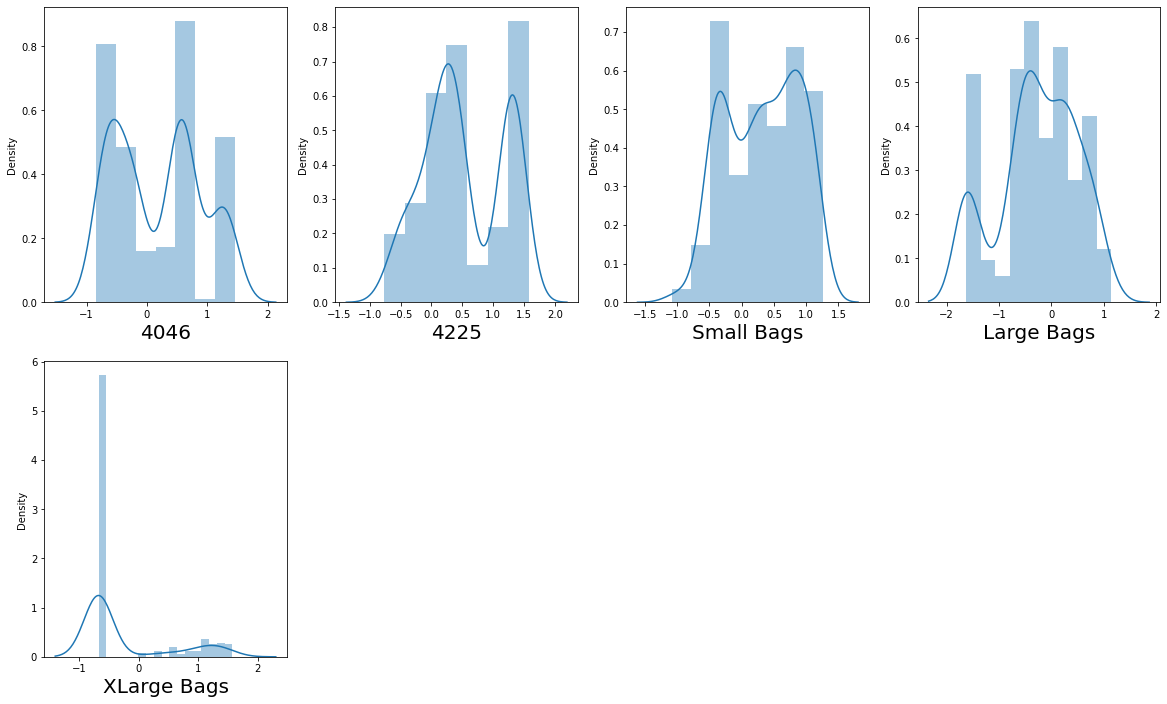
The data can be right skewed or left skewed if the median or mean is high and data is highly spread it can be observed through the skew() method, if the skew score is negative and greater than 5 it means data

is negatively skewed on left side and if the data is more than +5 it means the data is skewed on right side.



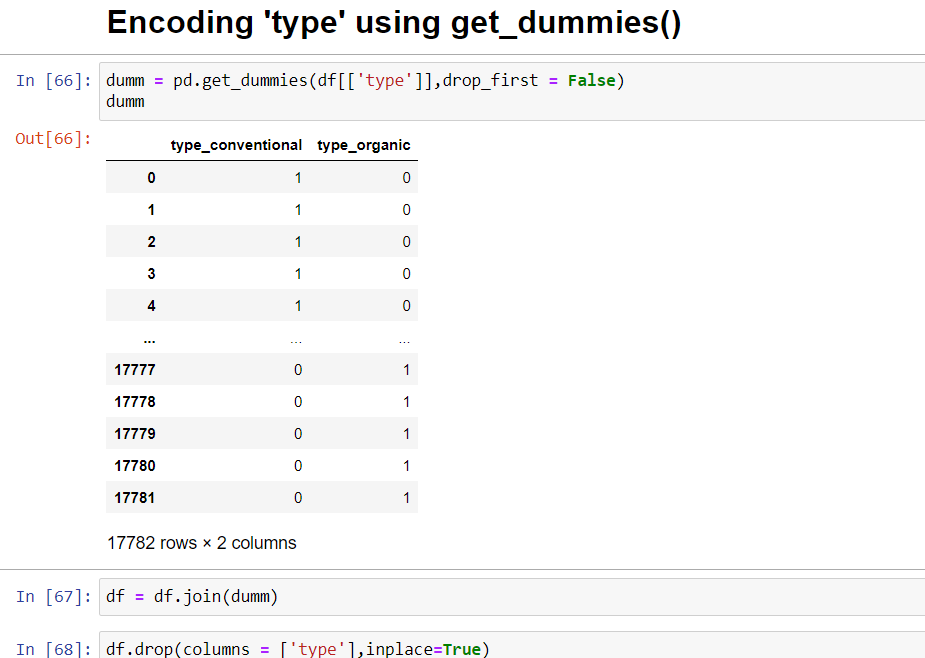
# Normalizing Data Distribution using Power Transformer Method

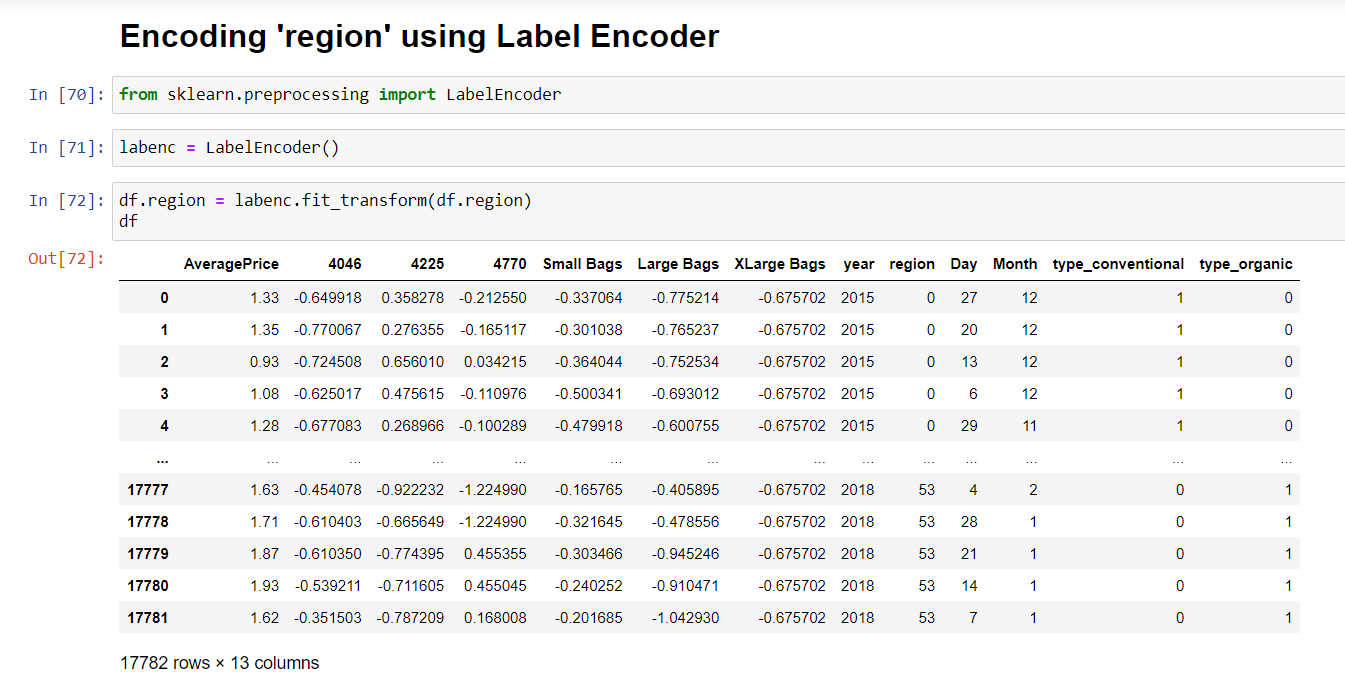




**Handling Categorical Columns:**

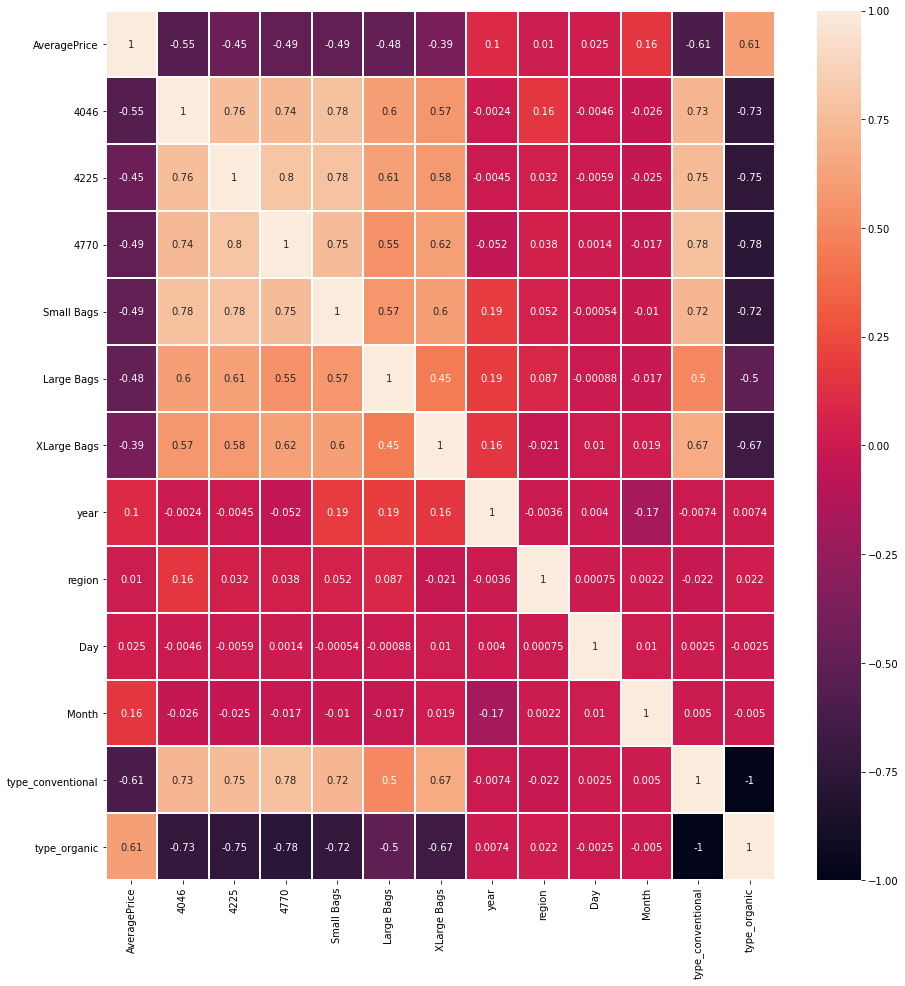
Most of the statistical models cannot take Objects / Strings as input they only takes numbers as inputs. Using Encoding Dummies Technique they only takes numbers as inputs, with LabelEncoder () it is possible to categorize the string into Numbers as 1,2,3 ,they all are converted through LabelEncoder().fit transform function.

****

****

**Finding Correlation:**

Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables.corr() function is used to see the correlation among the dependent variable and independent variable you can see correlation in the following figure

****

**Pre-processing Pipeline:**

Pipelines are the special way to simplify the code, Pipeline is generally used if we have to perform the code repeatedly usually when there is different train and Test data

Here Our EDA process is completed now moving towards next step and building model using Machine Learning.

**Building Machine Learning Models**

**Separating Features and Target column**:

It is necessary to separate the independent/Features column into a variable (x) and target column into a variable (y).

**Splitting the Data for Training and Testing**

In ML the separated data is split into 4 parts for Training and Testing of features (x) and for Training and Testing of Target (y) like x\_train, x\_test, y\_train, y\_test. It is possible through a inbuilt library of sklearn’s train\_test\_model.

**Training the Models**

To find the best model it is necessary to train 3-4 models, In the same way I have trained rf = Random Forest Regressor

dt = Decision Tree Regressor

xg = XGB Regressor

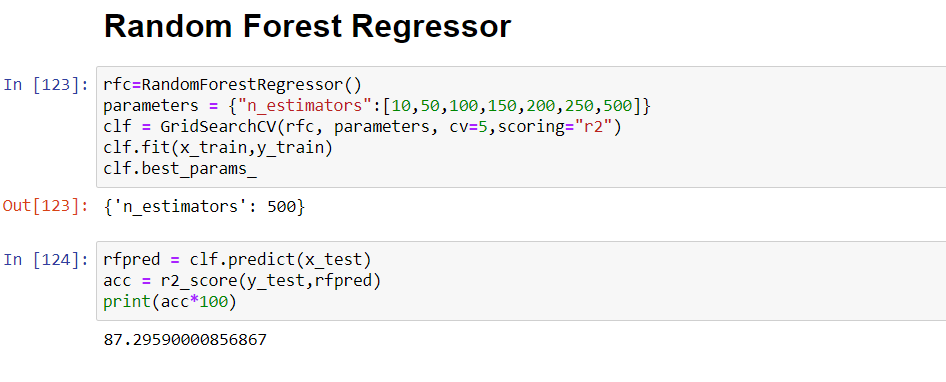
SV= SVR

r=Ridge()

# Based on comparing Accuracy Score results with Cross Validation results, it is determined that Random Forest Regressor is the best model. It also has the lowest Root Mean Squared Error score.

**Hyper Parameter Tuning:**

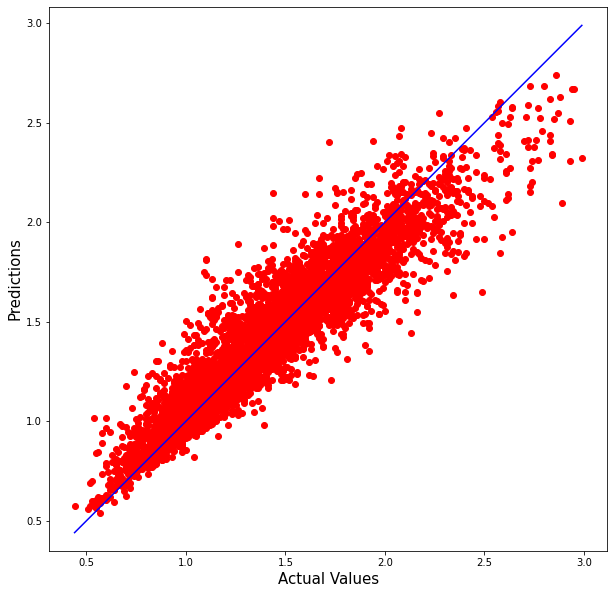
In order to increase the accuracy score of the model, we use hyper parameter tuning of the best model in order to find best parameters by using- GridSearchCV() in the following commands :-



**Data Deployment:**

# In order to dump the model which we have developed so that we can use it to make predictions in future, we have saved or dumped the best model ,i.e., Random Forest Regressor by using- joblib.dump() in the following commands :-

# 



**Concluding Remarks:**

# Accuracy of Random Forest Regressor has changed from 87.14% to 87.29% by hyperparameter tuning.

In this last stage, we deploy the model to production environment so that we can make the models available to users for making predictions and important business decisions.