**The Avocado Problem**

**Introduction**: An avocado is a bright green fruit with a large pit and dark leathery skin. Avocadoes are consumed by people heavily in the United States. Given data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. This data will be analysed and make different predictions.

**Conceptual Background of the Domain Problem:** A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

**Review of Literature:** Considerable about of online research is done in order to understand the problem and requirement of ML in solving this problem.

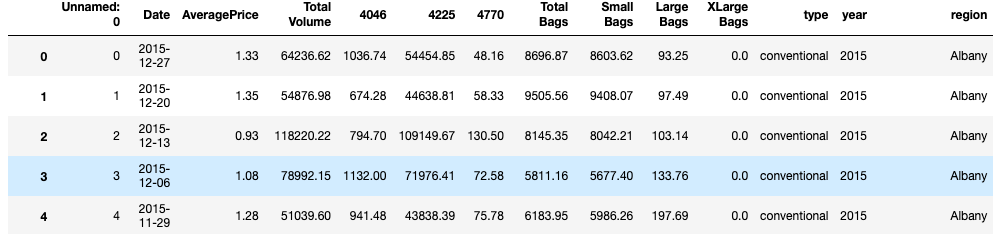
**Problem statement:** The avocado, a tree likely originating from south-central Mexico, is classified as a member of the flowering plant family Lauraceae. The fruit of the plant, also called an avocado, is botanically a large berry containing a single large seed. The given dataset can be seen in two angles to find the region and find the average price .

**Task**: One of Classification and other of Regression

**Data Sources and their formats:** 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 data 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 data are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this data.The dataset consists of both numerical and categorical variables. There is a total of 14 explanatory variables describing every aspect of the analysis. Overview of the 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



**Data Preprocessing Done:**

Fair amount of data cleaning process was also involved. This includes checking for space and null values and then dealing with them. Also our available data has 14 columns ('Unnamed: 0', 'Date', 'AveragePrice', 'Total Volume', '4046', '4225', '4770', 'Total Bags', 'Small Bags', 'Large Bags', 'XLarge Bags', 'type', 'year', 'region'). Outliers were also detected and removed.

**Hardware and Software Requirements and Tools Used:**

Python code was written in Jupiter notebook. Below are the libraries needed in the process.

Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model\_selection.cross\_val\_score, sklearn.linear\_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion\_matrix,accuracy\_score, sklearn.metrics.f1\_score, scipy.stats, seaborn

Identification of possible problem-solving approaches (methods)

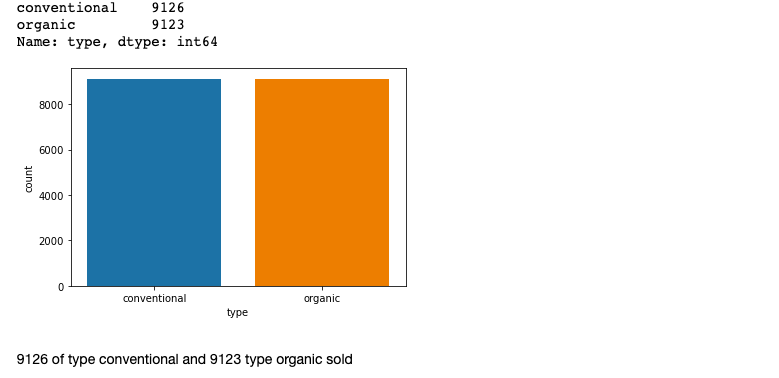
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.
* Skewness was checked using histogram.

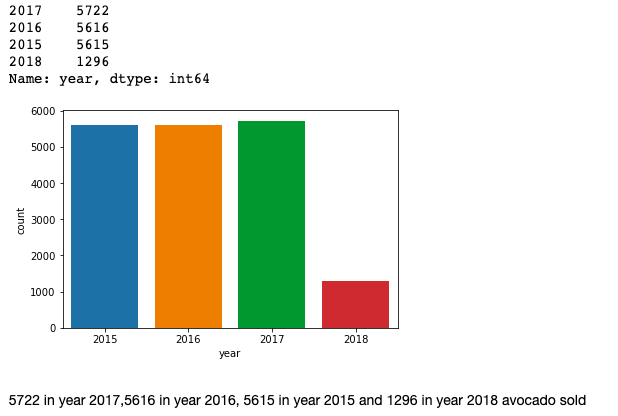
Testing of Identified Approaches (Algorithms):

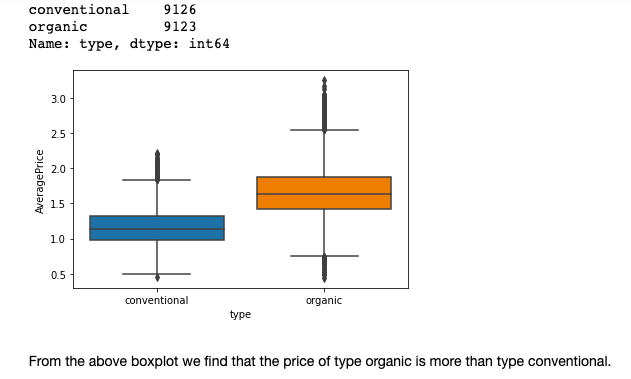
Below algorithms were used for training and testing of data

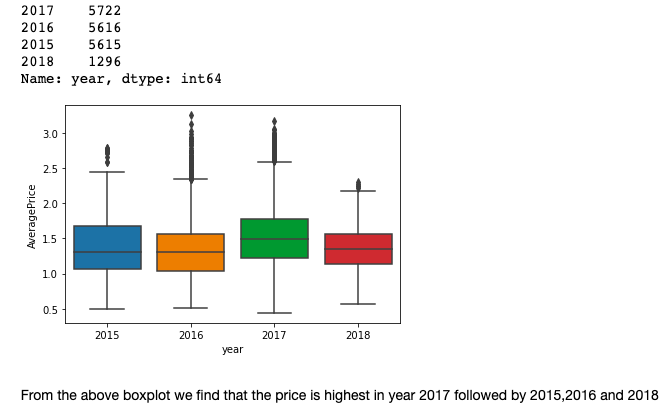
* RandomForestClassifier
* KNeighborsClassifier
* DecisionTreeClassifier
* GradientBoostingClassifier
* Ridge

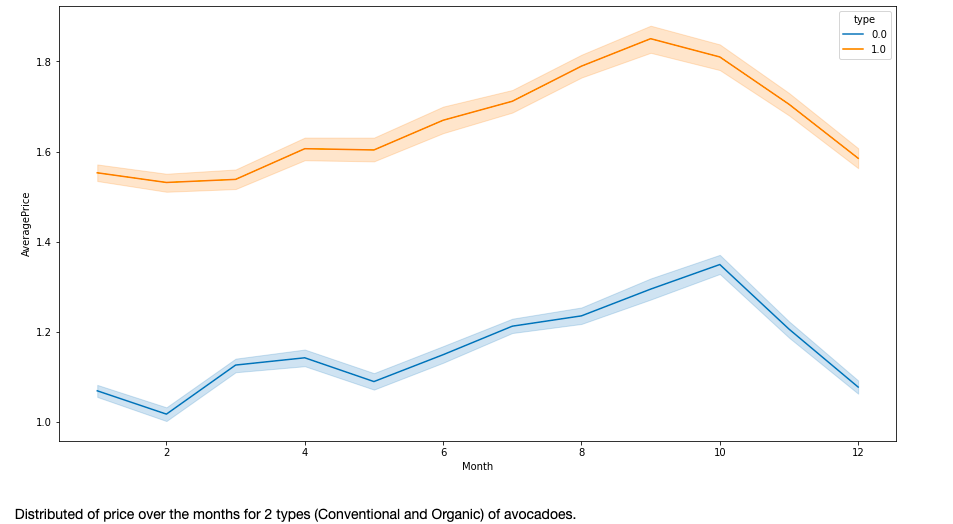
**Visualizations:**

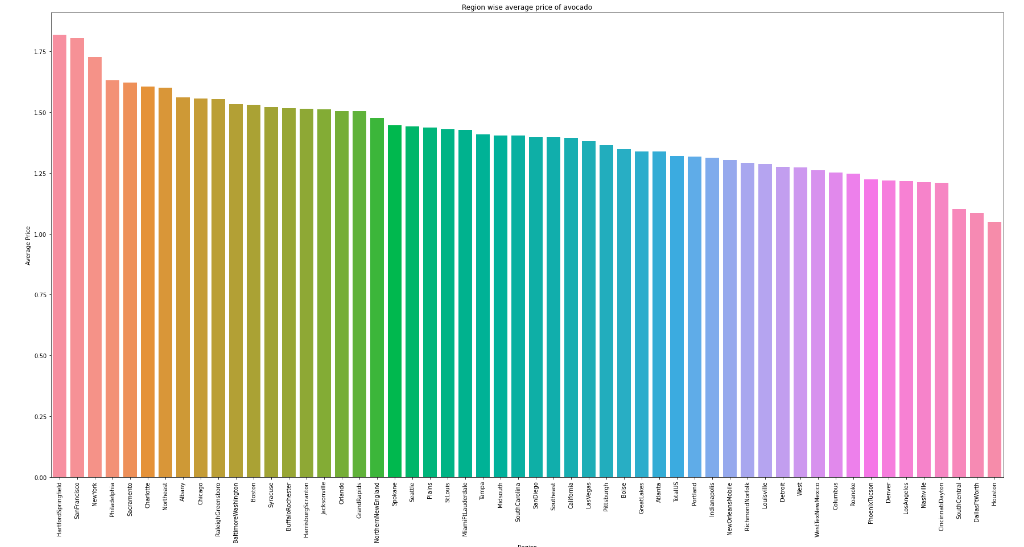


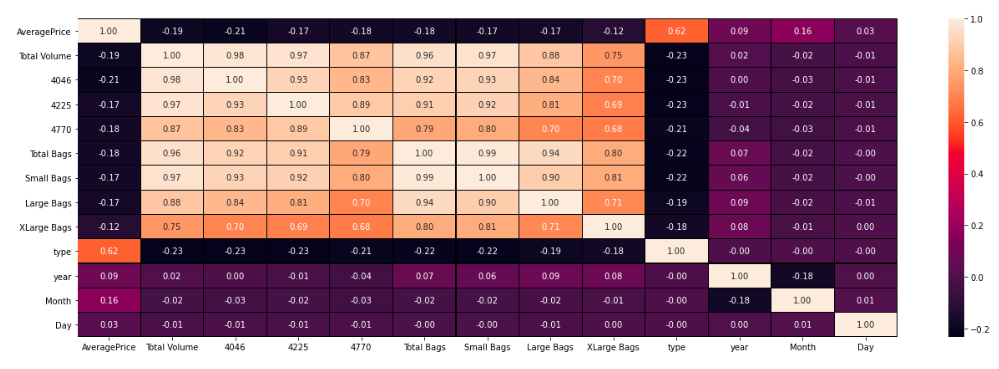


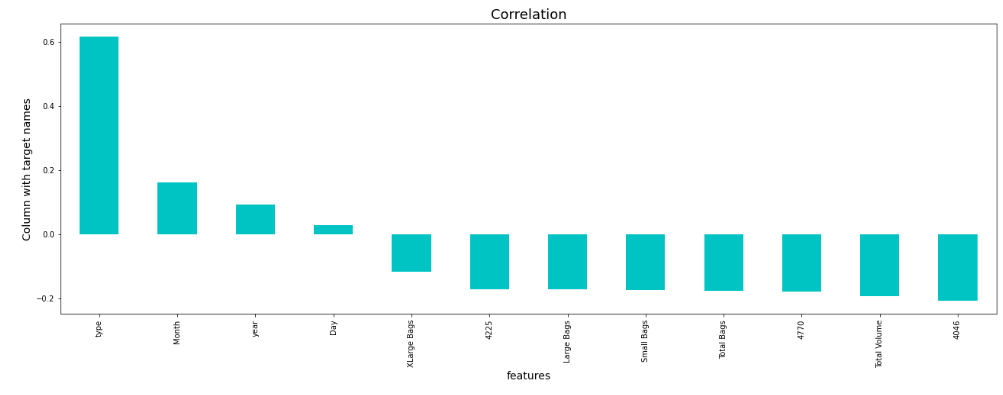


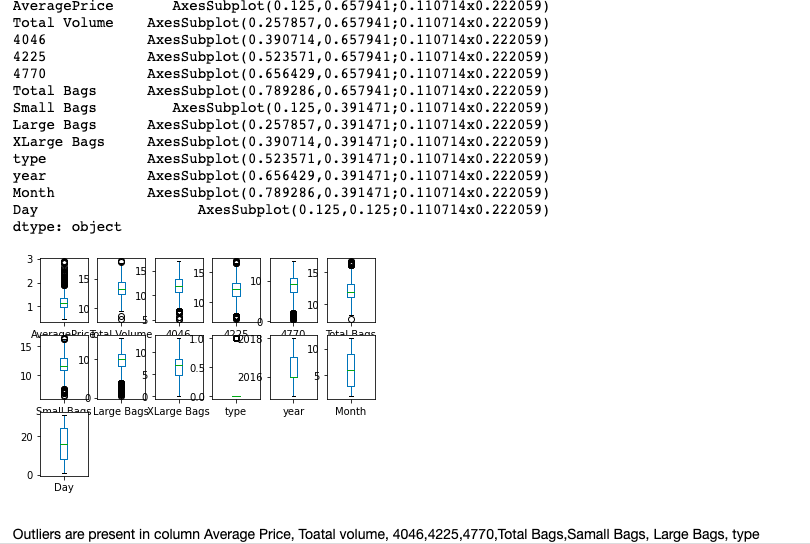












**Key Findings and Conclusions of the Study**

Score was highest in DecisionTreeClassifier and GradientBoostingClassifier followed by KNeighborsClassifier.