**Avocado Dataset Analysis**

**Problem Definition:**

In this model we are going to predict the average price of an avocado, the dataset represents the weekly retail scan data for national retail volume and price. 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 dataset 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.

**Data Analysis:**

The given dataset contains 16468 rows and 14 columns, some of the relevant columns in the dataset are:

* 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

Now we can proceed with the EDA (Exploratory Data Analysis) process to grab some insights from the dataset.

**EDA Concluding Remarks:**

We have gone through the EDA process to train the dataset, in which we have separated the column “Average Price” (renamed as target) since it is the target variable which needs to be predicted. We have also dropped all the null values which doesn’t help in this analysis. Moving forward we have addressed the outliers and replaced categorical variables into numerical through label encoding technique and also the outliers has been removed using z score method. Finally, through correlation visualization of the dataset we noticed that column “Type” is not correlated with the target hence column “Type” has been dropped. Now the dataset is ready to build and train our model to predict the target.

**Pre-processing Pipeline:**

In this process we have to analyze the dataset which we have taken into action to predict the target such as required libraries, methods which we have followed to make sense of the given data which provides a highly sensible insights to build and train the machine learning model.

**Building Machine Learning Model:**

Hence, we are ready to build a sensible machine learning model which pulls out the exact prices of the avocados. We have used Linear Regression model since the price of an avocado is highly positively correlated with most of the columns in the dataset. Finally we got great numbers on r2 score, mean squared error and mean absolute error which are shown below.

Mean Absolute Error: 0.1282032923050958

Mean Squared Error: 0.02751741344208727

Root Mean Squared Error: 0.16588373471225945

R2 Score: 0.2064270028617825