Avocado Price Analysis

By: Poornima Prakash

Introduction

- This dataset represents weekly retail scan data for National retail volume (units) and price. Thus has Historical data on avocado prices and sales volume in multiple US markets starting from 2013 to current year.
- Data has been collected from IRI / FreshLook Marketing Multi-Outlet (MULO) retail scan data. Multi-Outlet (MULO) reporting reflects retail scan sales across 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.

Cont…

- Dataset contains 18249 rows and 14 columns.
- Below is the table showing names of all the columns and their description.

Column Name	Description
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
Date	The date of the observation

Initial Observations

Summary of data types in this dataset:

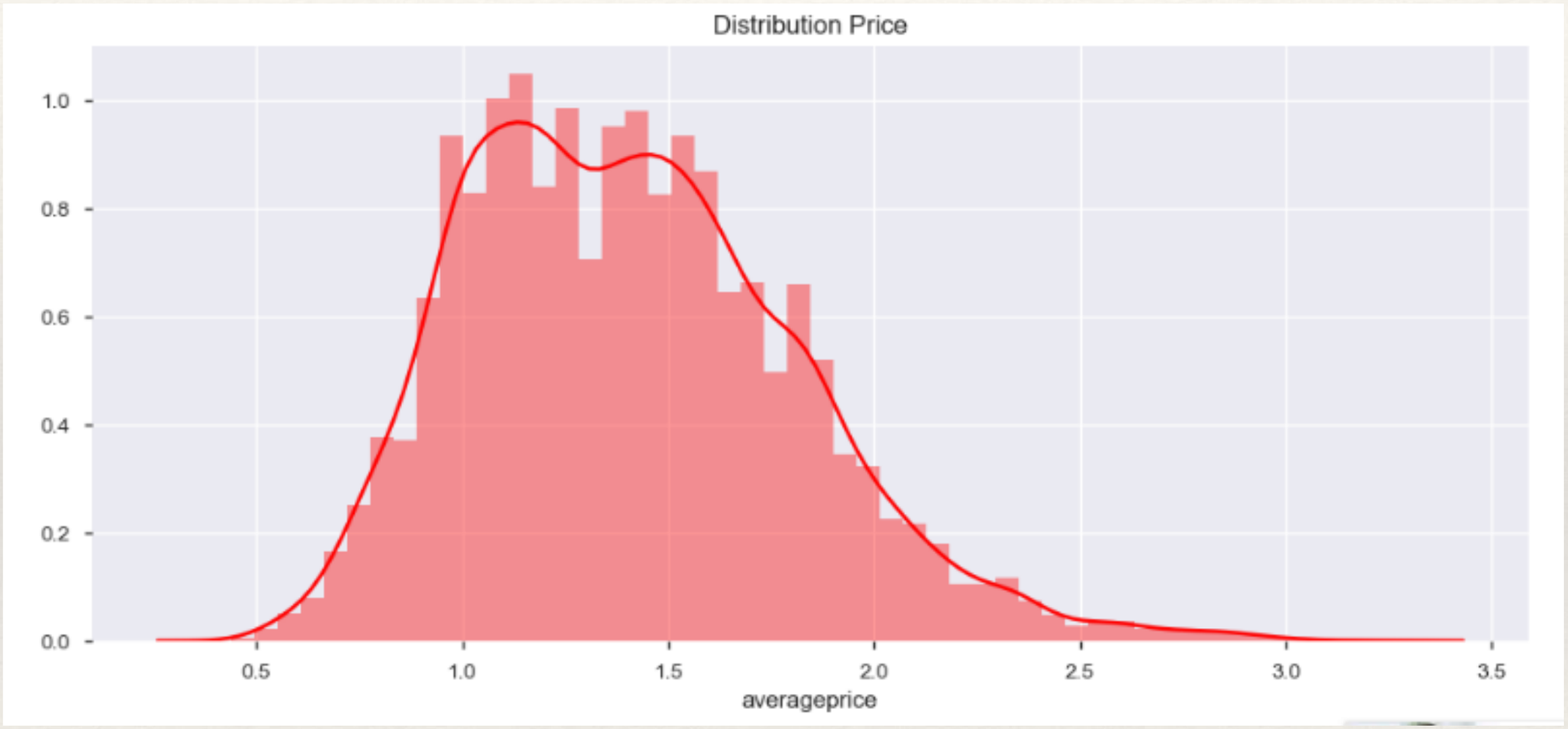
- ***Object_Columns***: **type column**
- ***Numeric_Columns*** : Everything else
- **There are no Nan or missing values in the dataset.**

Data Normalization

- Standardized all the column headers to lower case to avoid any typo errors.
- Convert Date column to pandas readable Datetime type
- Remove the column 'Unnamed: 0', as it is just an index column
- Do below listed Non-numerical data conversions for easy data analysis and model building for prediction.
  - Convert ‘type’ column into dummies by separating it into 2 other columns: organic and conventional.
  - Convert ‘region’ column into a categorical variable, as the data among 54 regions are pretty even.
  - Convert ‘Date’ column to the right scale to be used in the model.

Objectives & Graphs

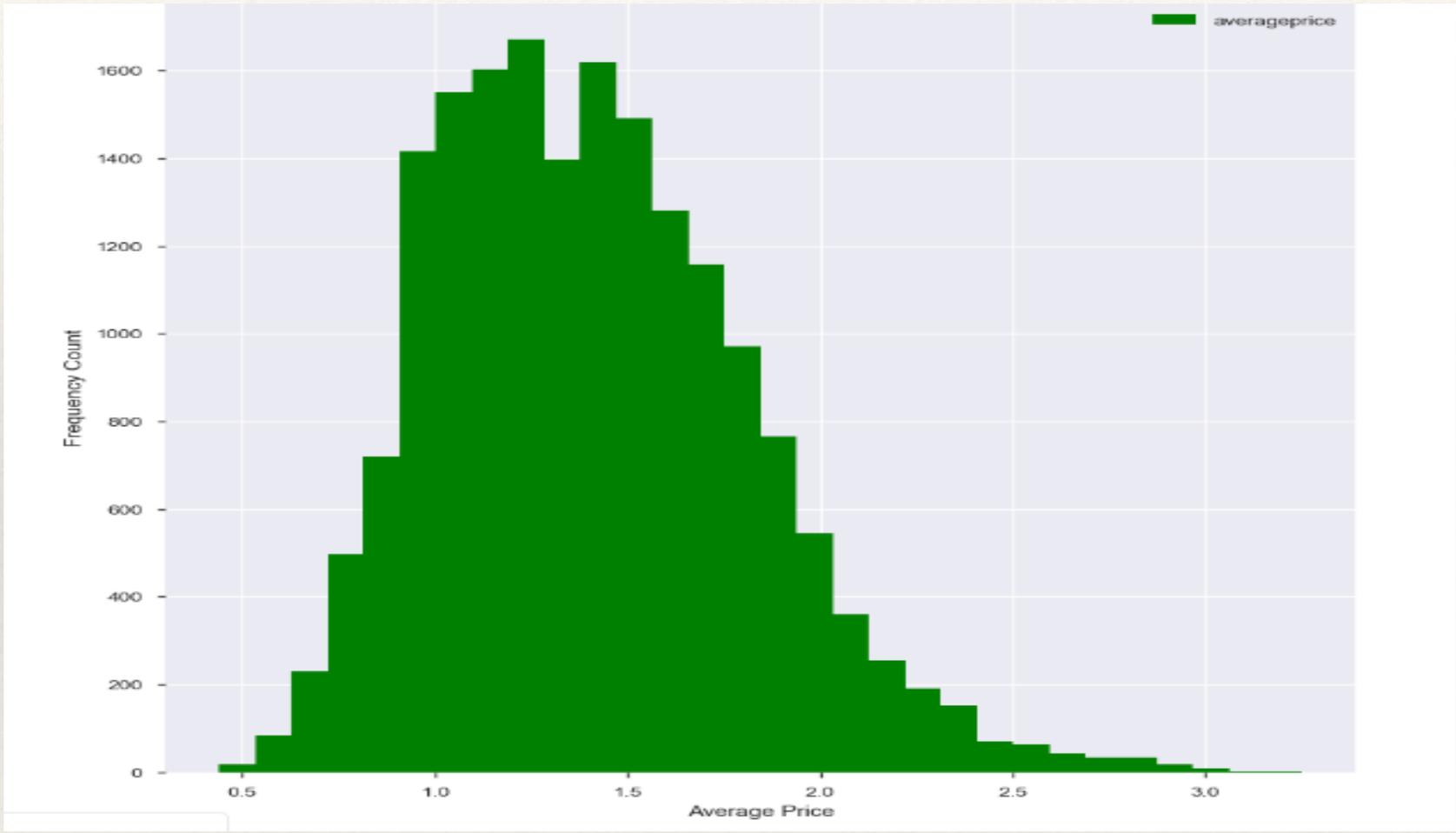
Distribution of average price of Avocado:



Average Price of avocado is normal with the values mostly between 1.0 and 2.0.

Objectives & Graphs

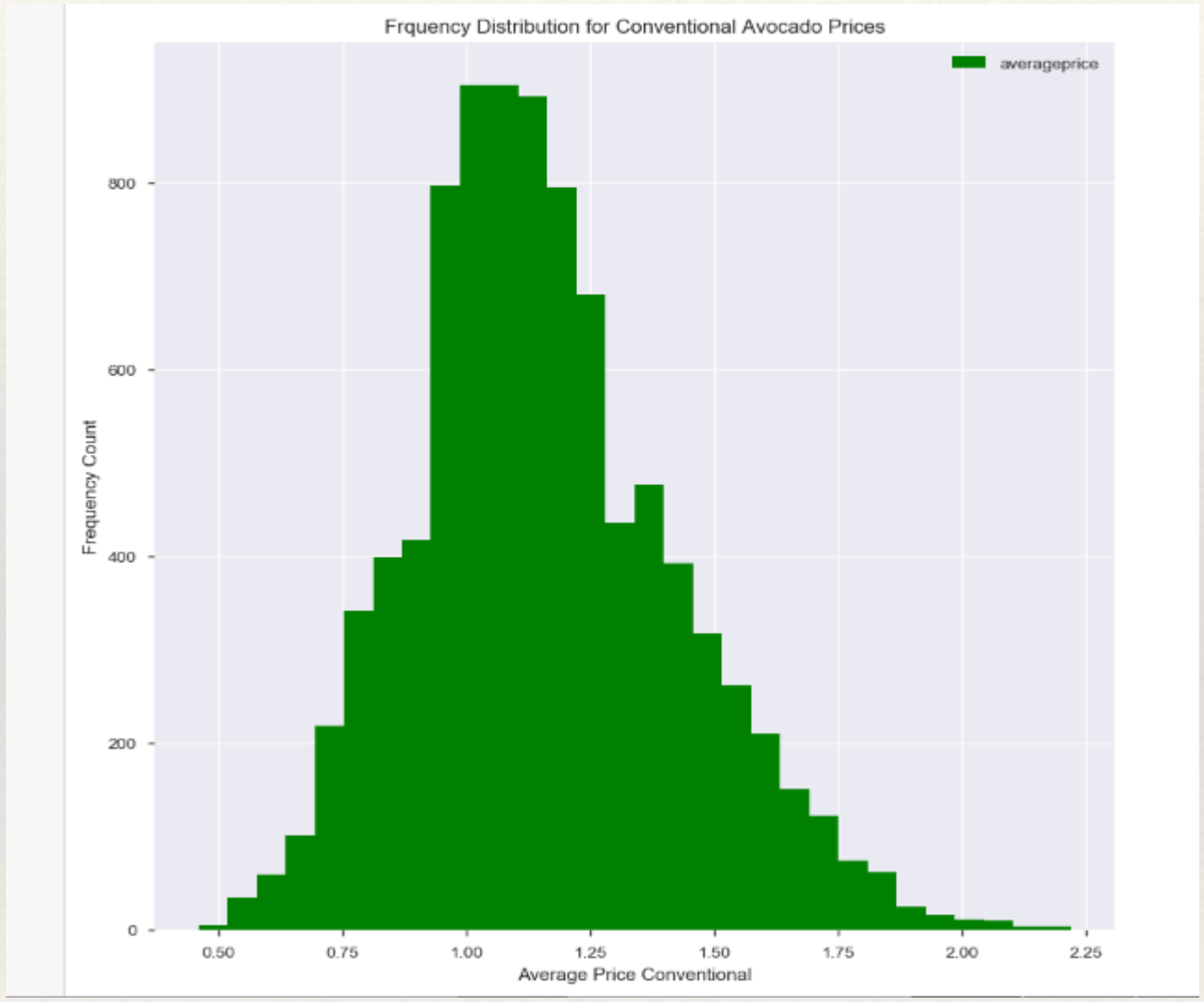
Frequency distribution of Avocado prices:



The price distribution for both types of Avocado is relatively normal.
The avocado price hovers between $1.00 and $1.80.

Objectives & Graphs

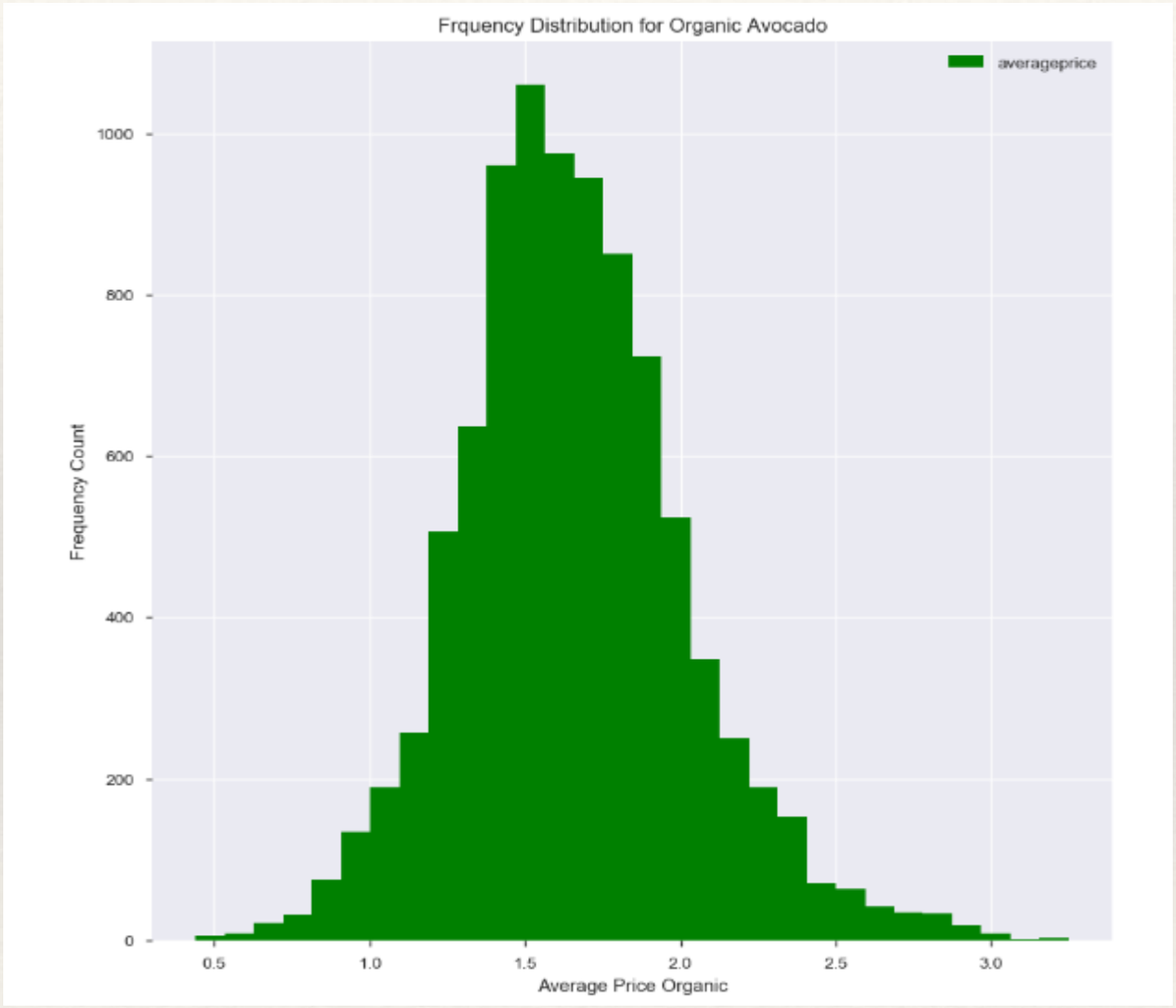
Frequency Distribution for Conventional Avocado Prices :



For conventional avocado's we see the price distribution is relatively normal. The conventional avocado price hovers around $1.00.

Objectives & Graphs

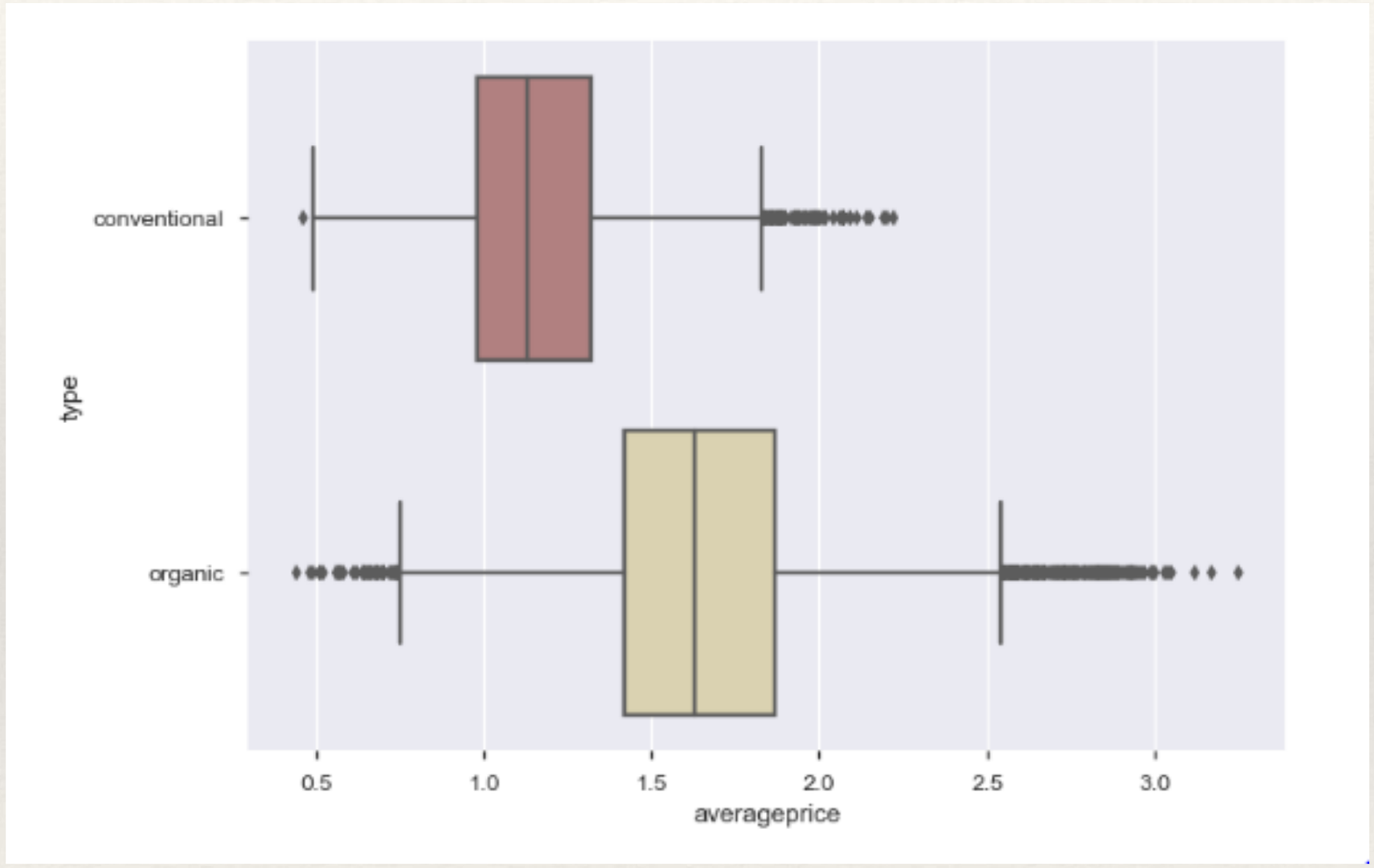
Frequency Distribution for Organic Avocado Prices :



For Organic avocado's we see the price distribution is relatively normal. The Organic avocado price hovers around $1.50.

Objectives & Graphs

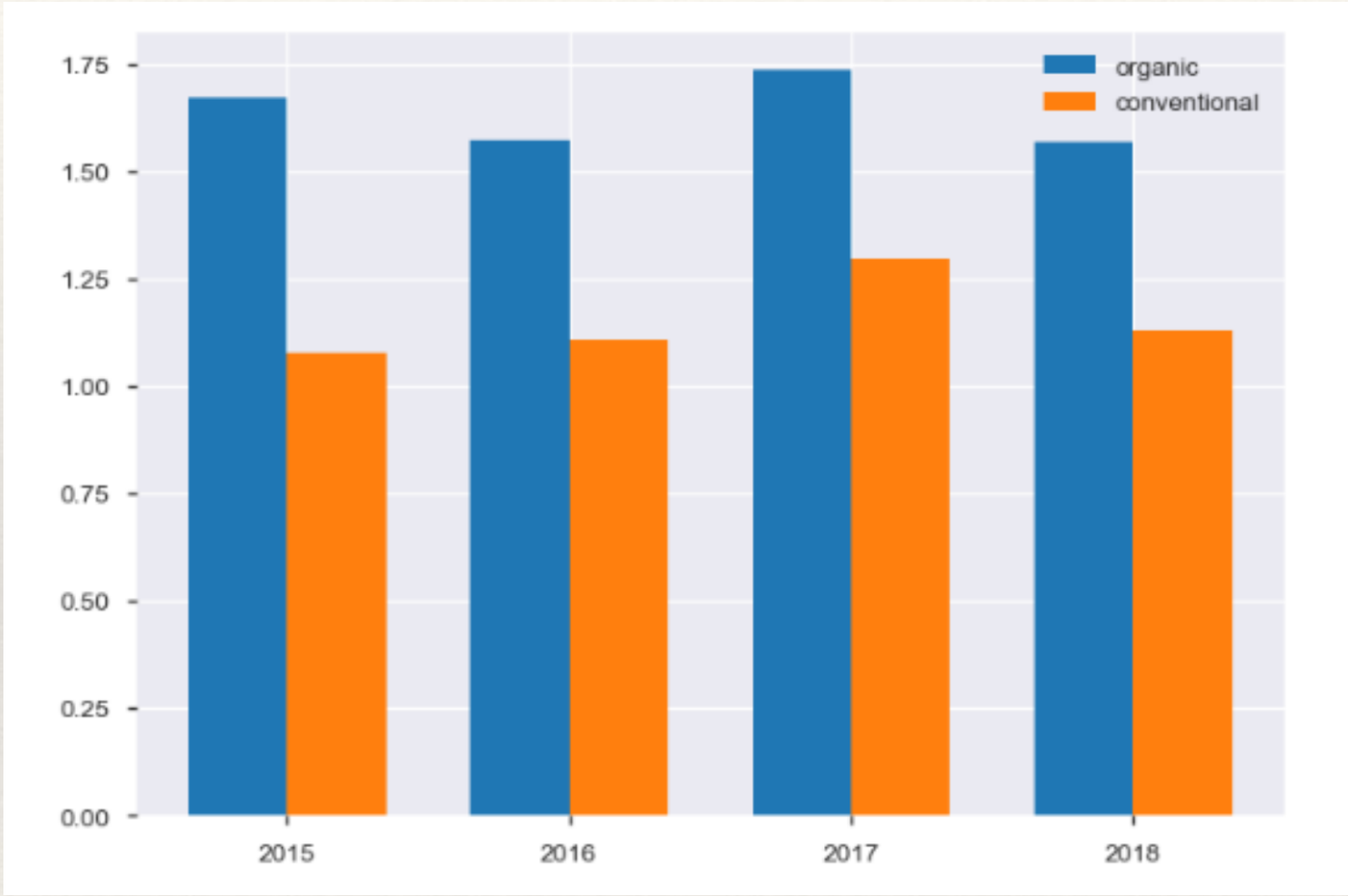
Average Price Vs Avocado type:



Organic avocados are a little more expensive than conventional types of avocados. On average the price difference between Organic and Conventional Avocados is roughly .50 cents.

Objectives & Graphs

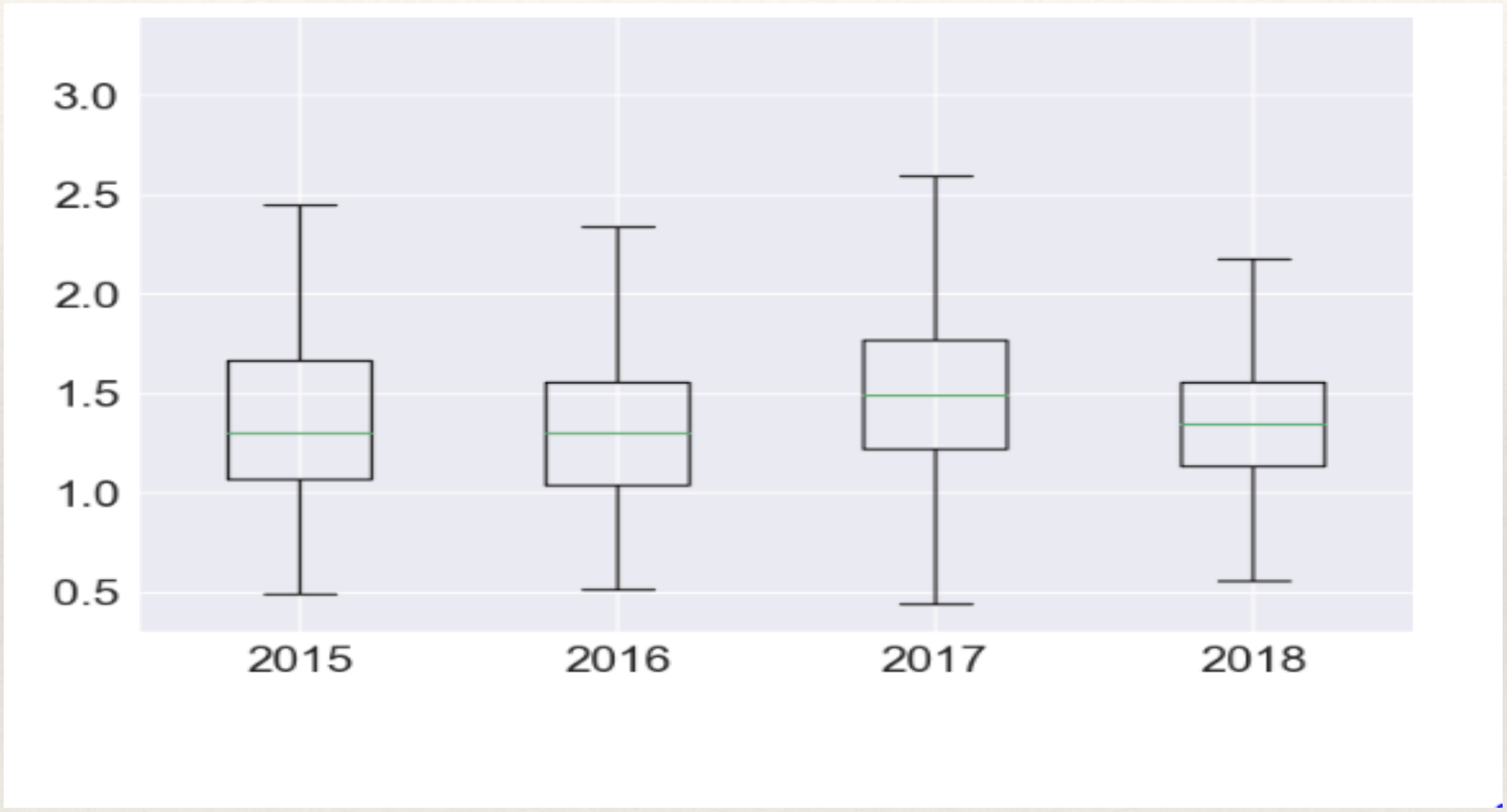
Price Vs Type and Year:



Average price of organic avocados has been always greater than the conventional ones. Year 2017 records the highest price for both types of the avocados.

Objectives & Graphs

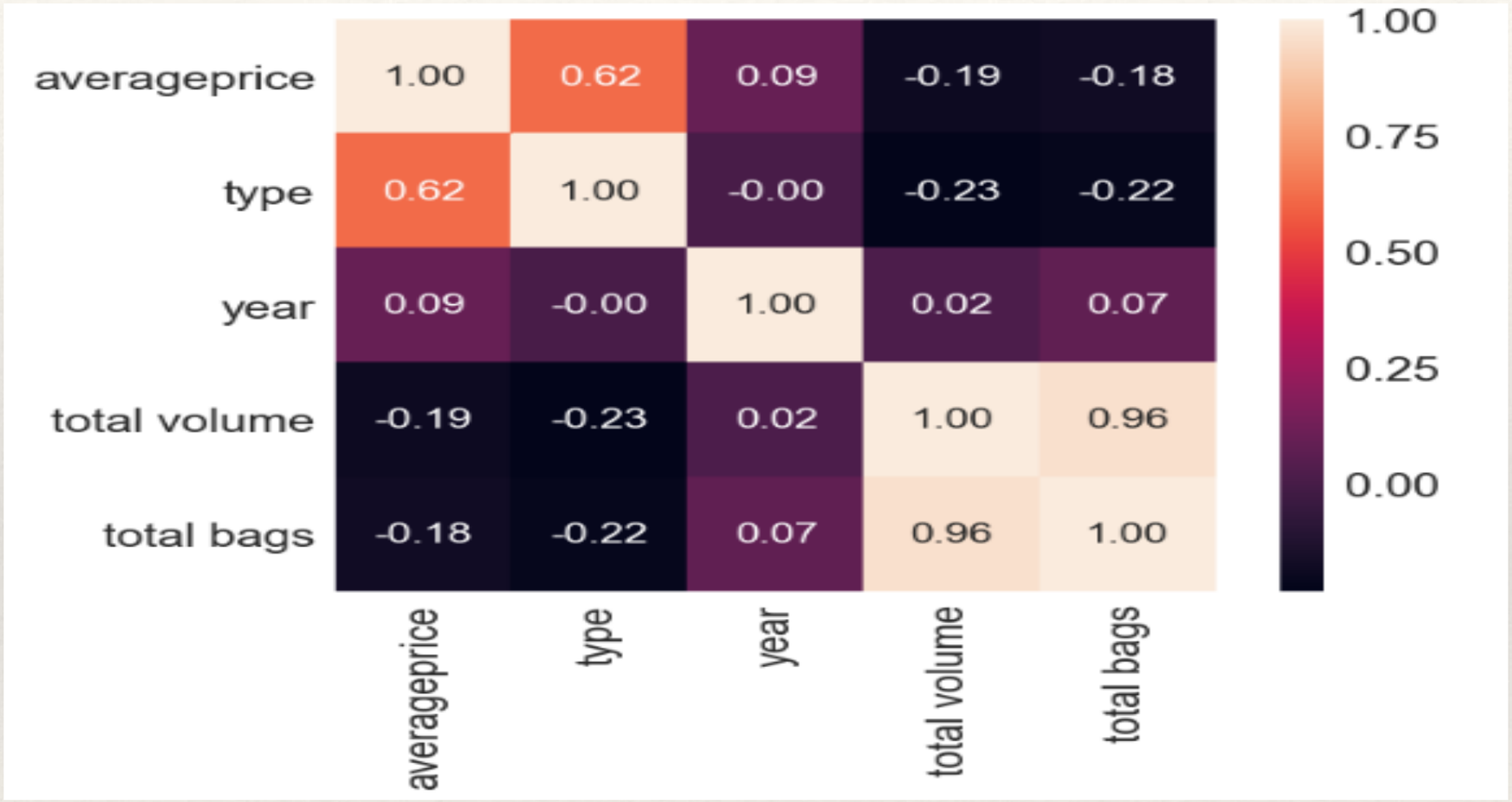
Price Vs Year:



Avocados prices have suffered a price spikes in 2017, but the prices have dropped comparatively in 2018.

Objectives & Graphs

Correlation Heat Map:



The price of avocado is influenced by the column 'type'. Also, there is a strong correlation between the features: "Total Bags" and "Total Volume". features Small Bags, Large Bag are also strongly correlated.

Model Creation and Evaluation

Regression Model with all features:

MAE: 0.23252019140625357
MSE: 0.09060791128181109
RMSE: 0.30101148031563696

Regression Model with selected features:

MAE: 0.23713058607891768
MSE: 0.09344406705311693
RMSE: 0.3056862231980972 4

Model Creation and Evaluation

Decision Tree Model :

MAE: 0.2203021550459430 3
MSE: 0.0827052951890327 3
RMSE: 0.287585283331801 83

Random Forest Model:

MAE: 0.2203866215218266 4
MSE: 0.0827833196690666
RMSE: 0.287720905860291 3

Model Creation and Evaluation

Regression Model with XGBoost :

MAE: 0.0935959514792963 4
MSE: 0.017717366153367 24
RMSE: 0.133106596956601 62
R^2 of XGBoost: 0.890910509620863 8

The RMSE value from the XGBoost regression model is 0.133, obviously lowest compared to the previous models. Also, the model also has the higher R square value. Hence, we chose the XGBoost regression model as our solution.

Conclusion

- The Average Price of both organic and conventional avocados was getting more expensive from 2015 to 2018. Also the price varies from region to region.

- Organic avocados are more expensive than conventional ones.

- The Average Price of avocados is affected by years, regions, types.

- On evaluating different models to predict the avocado prices, found that the XGBoost regression model gave the best results with least RMSE(0.13) value.