AVACADO PROJECT

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. The table 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 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.

Goal

Inspired by the popularities of avocado toasts among millennials, and finding skyrocketed prices on avocadoes at produce sections recently, I wanted to find out which cities in the U.S. provide the most reasonable prices for avocados and understand the market and trends better to hopefully benefit suffering millennials (including myself). I explored the data of prices and volumes of avocados sold in the major metropolitan areas, analyzed the costs at different cities and the correlation between the volume and prices, and built machine learning models to predict prices.

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

The dataset comprises of 18249 observations of 14 columns. Below is a table showing names of all the columns and their description.

The unclear numerical variables terminology is explained in the next section:

|  |  |
| --- | --- |
| **Features** | **Description** |
| ‘Unamed: 0’ | Its just a useless index feature that will be removed later |
| ‘Total Volume’ | Total sales volume of avocados |
| ‘4046’ | Total sales volume of Small/Medium Hass Avocado |
| ‘4225’ | Total sales volume of Large Hass Avocado |
| ‘4770’ | Total sales volume of Extra Large Hass Avocado |
| ‘Total Bags’ | Total number of Bags sold |
| ‘Small Bags’ | Total number of Small Bags sold |
| ‘Large Bags’ | Total number of Large Bags sold |
| ‘XLarge Bags’ | Total number of XLarge Bags sold |

## Problem Statement

* The notebooks explores the basic use of Pandas and will cover the basic commands of (EDA) for analysis purpose.
* In this study, we will try to see if we can predict the Avocado’s Average Price based on different features. The features are different (Total Bags,Date,Type,Year,Region…).

The variables of the dataset are the following:

* Categorical: ‘region’,’type’
* Date: ‘Date’
* Numerical: ’Total Volume’, ‘4046’, ‘4225’, ‘4770’, ‘Total Bags’, ‘Small Bags’,’Large Bags’,’XLarge Bags’, ’Year’
* Target:‘Average Price’

Data processing

The Feature "Unnamed:0" is just a representation of the indexes, so it's useless to keep it, we have remove it in pre-processing .

Well as a first observation we can see that we are lucky, we don’t have any missing values (**18249** complete data) and **13 columns**. So we did some Feature Engineering on the Date Feature in pre-processing later we are able to use the day and the month columns in building our machine learning model later. ( I didn't mention the year because its already there in data frame). Then I have done Pandas Profiling before preprocessing dataset, so we can get initial observations from the dataset in better visual aspects, to find correlation matrix and sample data. File was saved as html file **avocado before preprocessing .html**. Will take a look at the file and see what useful insight you can develop from it.

## Data Analysis:

### 1. Correlation between price and volume

I had a hypothesis of demand and supply, which means, in other words, that if the consumed volumes are higher, then the prices would be lower. The scatter plot I created, using matplotlib in Python libraries displays that it seems there is a trend for that direction. The Pearson correlation coefficient showed a small negative correlation (-0.225) between the average price and average volume consumption. Thus, there is an association between demand and supply, but that cannot explain everything about how the prices are structured. Having some outliers on the right side of the plot where some cities had the highest prices while the consumed volumes are limited.

In the next series of analyses, I would dive deep into the data and testing hypotheses on how the network of the geography of consumption and production impact prices.

### 2. Ranking of cities by volume and price

Based on the below bar chart, surprisingly, Los Angeles consumes twice more than that of the second-highest volume cities, New York. The volume of the consumption of Los Angeles is astonishingly larger than any other city.

According to U.S. Department of Agriculture, in 2017, the world production of avocados was 5.9 million tonnes, led by Mexico with 34% (2.01 million tonnes) of the total. In the U.S., California is the major producer, accounting for 93 percent of U.S. avocado output. However, most of US consumption relies on imports from Mexico - 89% of the imports. Thus, the proximity to these areas could highly likely affect the prices of avocados.

In terms of prices, the below chart shows that the Northeast and Mideast are the most expensive areas for avocados. It is surprising that Hartford-Springfield, Connecticut was the most expensive city. Also, San Francisco is geographically not so far from the production regions, but it ranked very high, even higher than New York and other Northeastern cities.

This brings down to a hypothesis on one factor that drives up prices: transportation costs from main production areas/regions (Mexico, CA, etc.) is the major element to impact prices? This would explain why Harford and other Northeastern cities had the highest costs. Based on the report from the Department of Agriculture, this could be strongly backed by their study for other fresh produce. The results of the "study indicate that transportation costs significantly increase the costs of marketing these produce items and therefore their wholesale price."

The next question here is if living costs could be another major factor, considering some cities such as San Francisco where the proximity to Mexico would not be an issue, also ranked very high. Clearly, San Francisco could be an exception for this "proximity" theory, and yet the cost of living index, paradoxically indicates that since some cities which did not normally register as expensive cities such as Hartford, Philadelphia, and Albany (all Northeastern cities) did rank higher for avocados, confirms the hypothesis that the proximity to Mexico and South America may matter to the pricing.

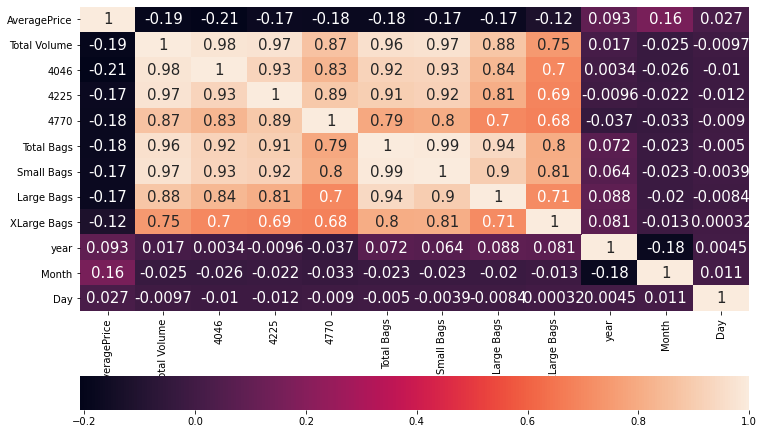
### 3. Volume and price fluctuation (Time Series Analysis)

The below graph shows that there was a huge drop in the price in the summer, 2015. In general, it fluctuates over seasons - in summer, it goes down and in winter, it goes up. However, as a general trend, the plot shows that it is steadily increasing over years. And in 2017 there was a tremendous increase in the prices.

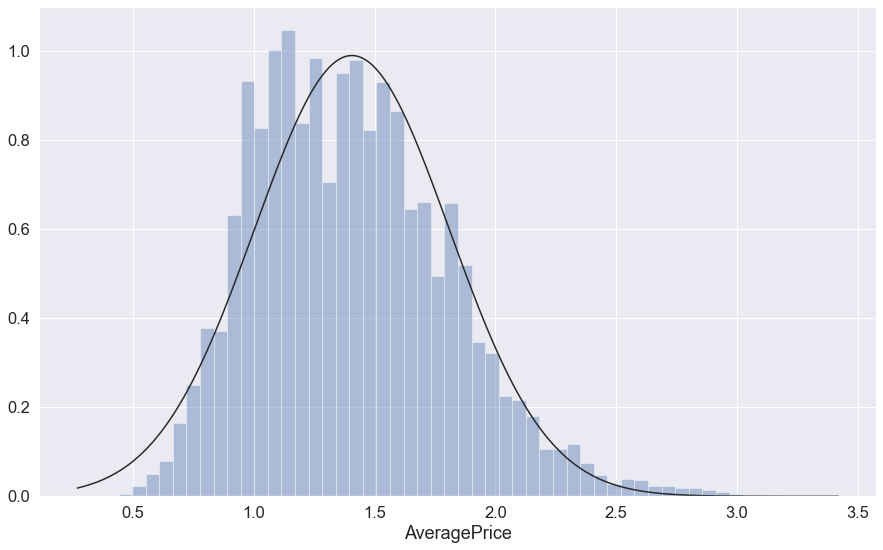
It is interesting to compare these ups and downs with those of the volume. In a sense, the volumes and prices are the flip sides of coins - the volume goes up strikingly high at the beginning of spring, and in summer, there is a smaller uptick and these are reflected on the price chart as downtimes. And that was explained in the first test on the Pearson correlation coefficient. There is a smaller association between the volume and price. Therefore, when there are more avocados in the market, the prices go down, and when there are not enough, the price increases.

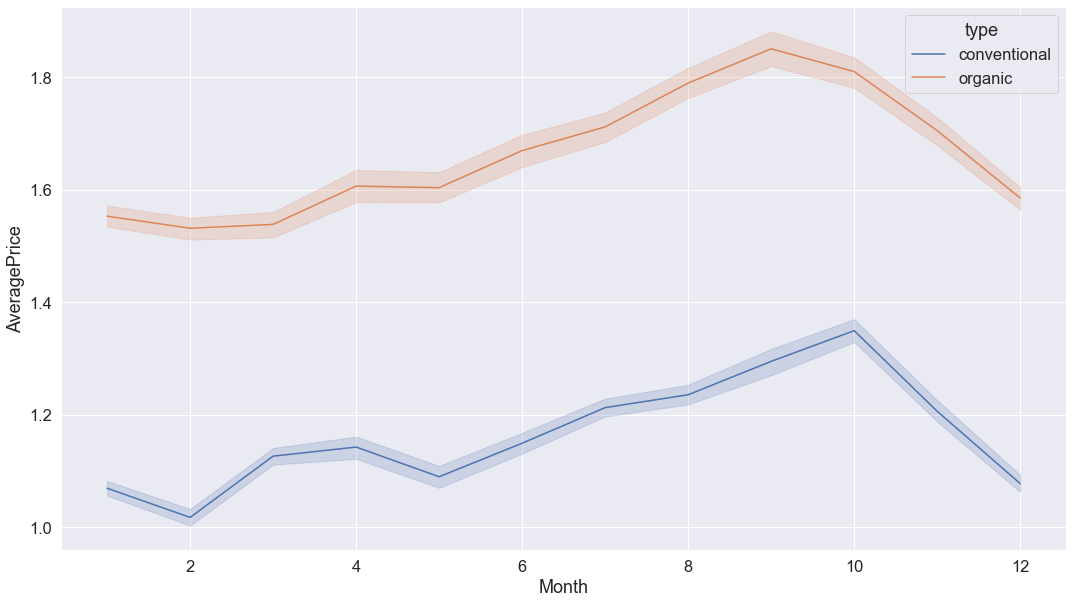
### 4. Conventional vs Organic

Expectedly, the prices of organic avocados usually run higher than conventional ones. For some cities and regions such as Phoenix-Tuscon, West Texas, and New Mexico, compared with other cities, organic avocados are priced so much higher than the conventional ones.

Below graph is the heatmap , which shows the relation between the columns and target variable. all the Features are not correleted with the Average Price column, instead most of them are correlated with each other. 

Below graph shows, Average Price distribution shows that for most cases price of avocado is between 1.1, 1.4.



In below graph,there is a hike between months 8 – 10 for both Conventional and Organic type of Avocados price

--- Below graph shows that these region are where price is very high

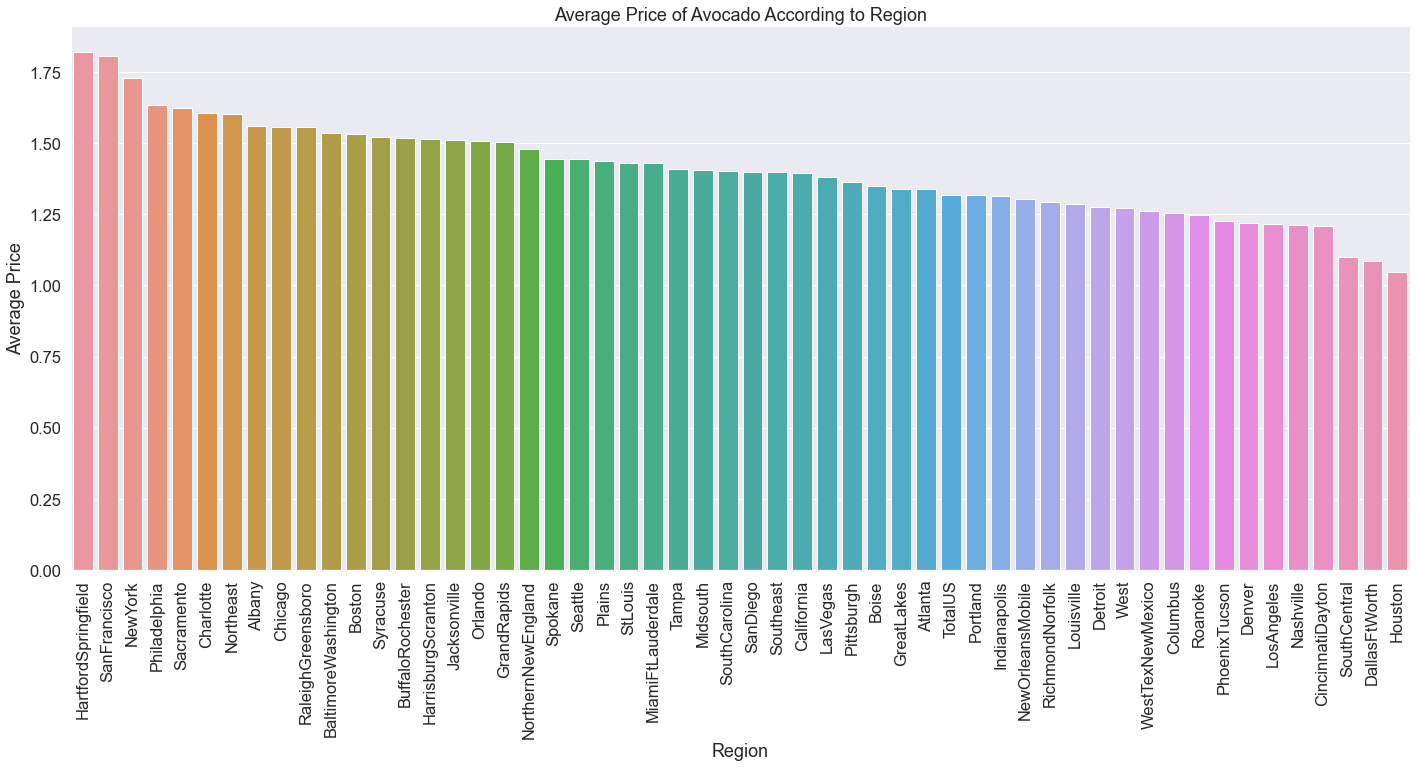
HartfordSpringfield,

SanFrancisco,

NewYork,

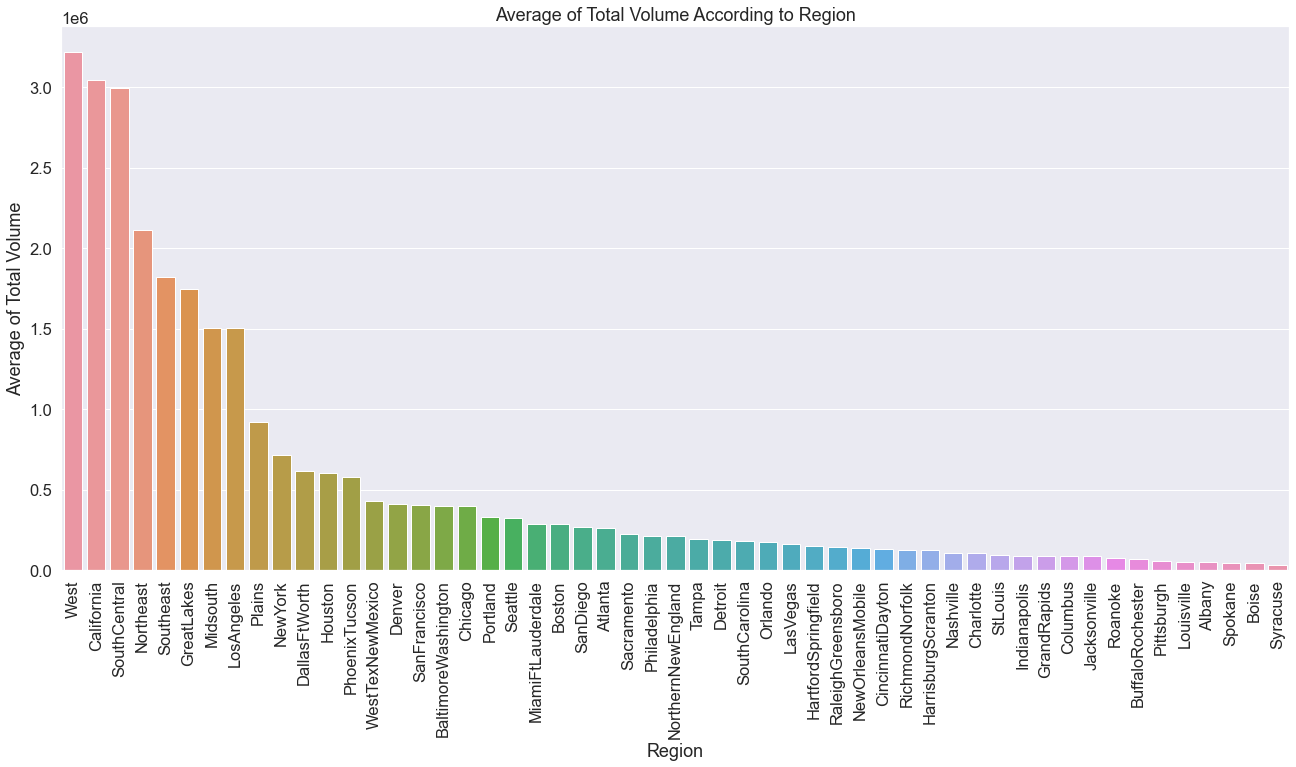
Philadelphia,

Sacramento.

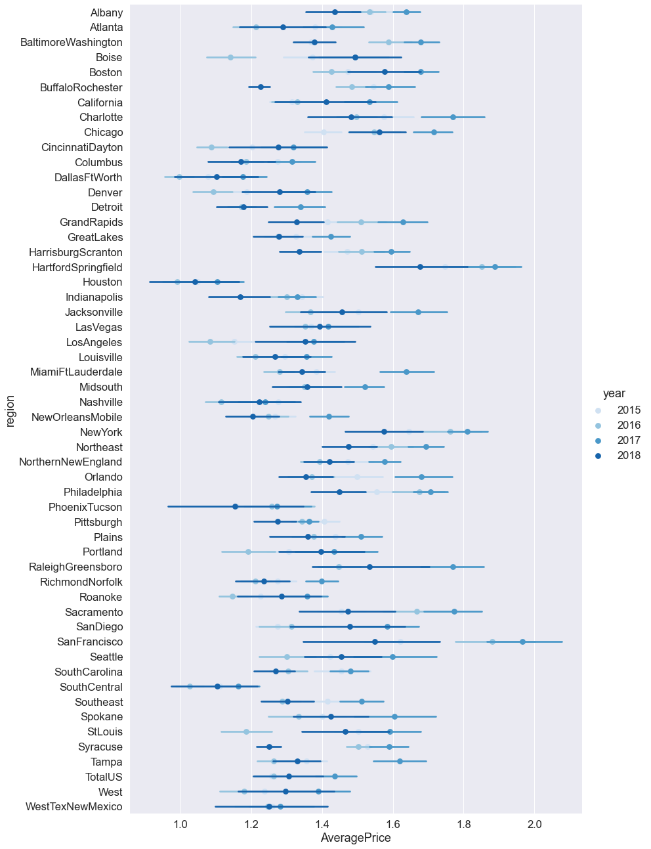


Below graph show that these region are where Consumption is very high

West, California, SouthCentral, Northeast, Southeast.

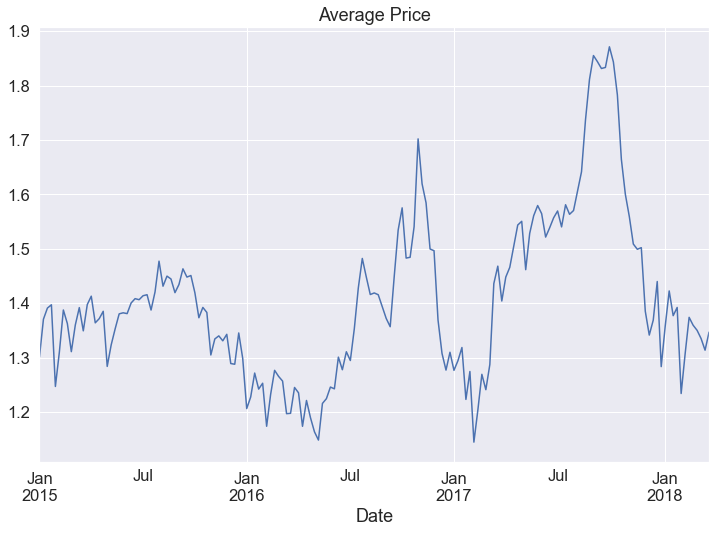


there is a huge increase in Avocado prices as the demand was little high in Year 2017 in SanFranciso region



Feature engeneering

54 regions and 2 unique types, so it's going to be easy to to transform the type feature to dummies, but for the region its going to be a bit complex, so I decided to drop the entire column.I have drop the Date Feature as well because I already have 3 other columns for the Year, Month and Day.



Model building

For the predictive models, I used the following three models to train the datasets, test the predicted prices, and compared them against actual prices to score the accuracies.

* Ensemble Model; Combined the following 3 models
* 2 Random Forest Models; Used different parameters respectively for version 1 and 2
* Linear Regression

Accuracy scores

Based on the below chart, one Random Forest model did quite better than other models including the ensemble model in which I combined all four models' outputs and averaged them out. I used mean squared error, root mean squared error, and R2 (coefficients of determinations) to compare the accuracies and in all three measures, Random Forest 1 did the best.

Hyper parameter tuning

We use parameters n\_estimators and criterion 'mse', 'mae'. We get as a result RandomForestRegressor(criterion='mse', n\_estimators= 9)

Conclusion

This study showed that there are some associations between prices and volumes. Additionally, the proximity to the major production region (Mexico) and seasons can be strong elements that affect the pricing. However, there are exceptions too. For example, you could move to Texas to buy cheaper avocados, but there is a possibility that organic avocados may not be as reasonably priced as conventional avocados.