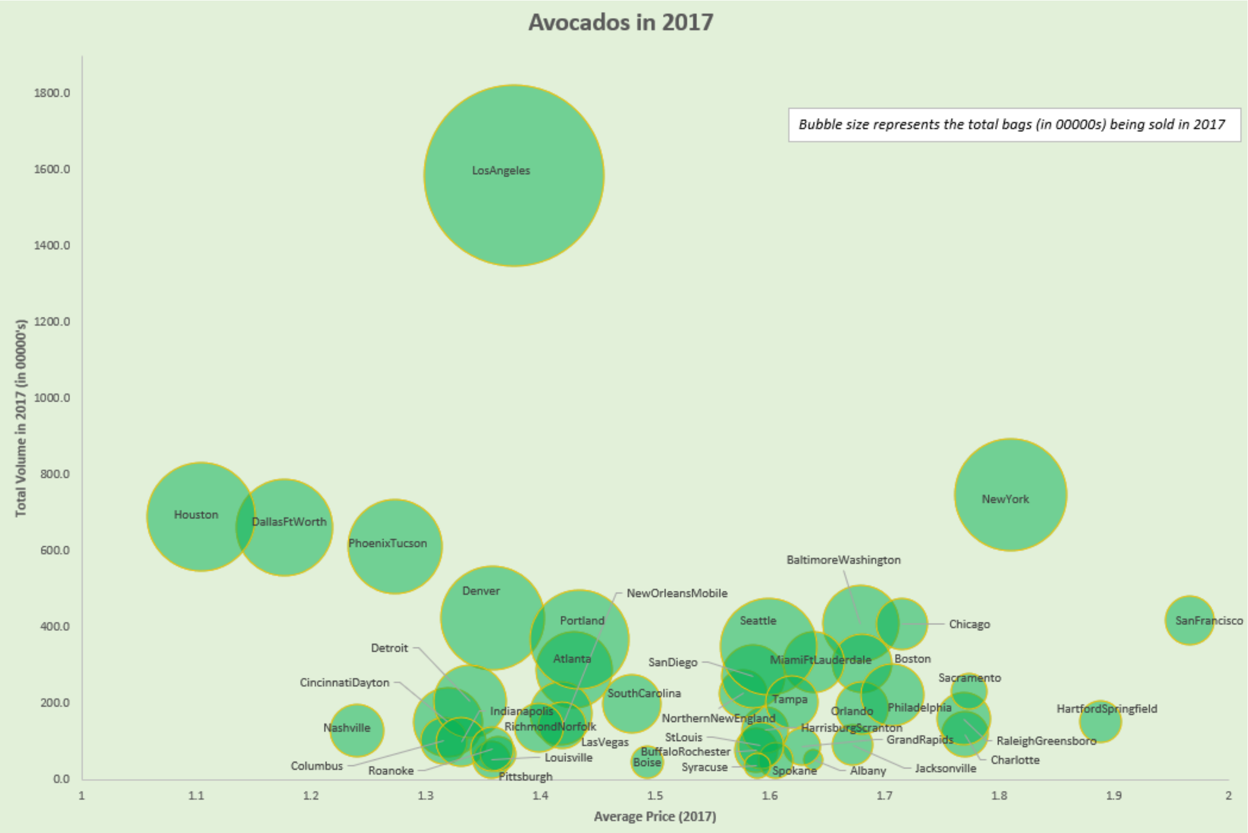
Analysis of Avocado Prices

Nikhar Jain
nikhar.jain@utexas.edu
December 2019

Executive Summary

My main goal of this report is to analyze the average avocado prices over 2015-18 and how each factor has an impact on total sales. I also aim to drill deeper into each of the variable in the dataset and draw meaningful insights to understand the what, how and why.



Notes: The x-axis for average prices has been deliberately made to start with 1 instead of zero, as the chart was not giving a clear picture - this shows that regions have varied prices between $1-$2; The above chart does not include data for California, Great Lakes and Plains as these are regions and cannot be compared to city-level data; Since 2018 data was available for only a few months, I used 2017 to show what was happening to avocados in 2017

Key Findings:

- Over 2015-2018, the average price of avocados has been on an upward trend. It followed a nearly normal distribution, with mean price lying between $1.1-$1.6 per avocado
- Conventional avocados dominate the overall avocado market. However, organic avocados are gaining traction
- In 2017, a nation-wide shortage of avocados due to reduced harvests from major avocado producing regions, led to an increase in the average prices
- Avocado sales are lower in the winter months, typically due to its harvesting season which lasts from January to early November
- Cities in the northeast and northwest regions have higher avocado price

Data Overview

This data was obtained from Kaggle, an online community for data scientists and machine learners with free access to 1000+ open datasets.

Originally, this dataset was downloaded from the Hass Avocado Board website in May 2018. The dataset includes 15,886 observations with 14 columns for the year 2015 - March 2018. Below is a snapshot of the columns in the dataset as well as their description.

Name	Description
Date	Observation Date
AveragePrice	Average price of each avocado
Type	Type of avocado - conventional or organic
Year	Year in consideration
Region	Observed city or state
Total Volume	Total volume purchased
4046	Total number of avocados with PLU code 4046
4225	Total number of avocados with PLU code 4225
4770	Total number of avocados with PLU code 4770
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 extra-large bags sold

Caveats:

- The Average Price (of avocados) in the table reflects cost per avocado, even though multiple avocados are being sold in bags
- The Product Lookup codes (PLU's) in the table are only for Hass avocados and other varieties of avocados (e.g. greenskins) are not included in this dataset.

Data Preparation

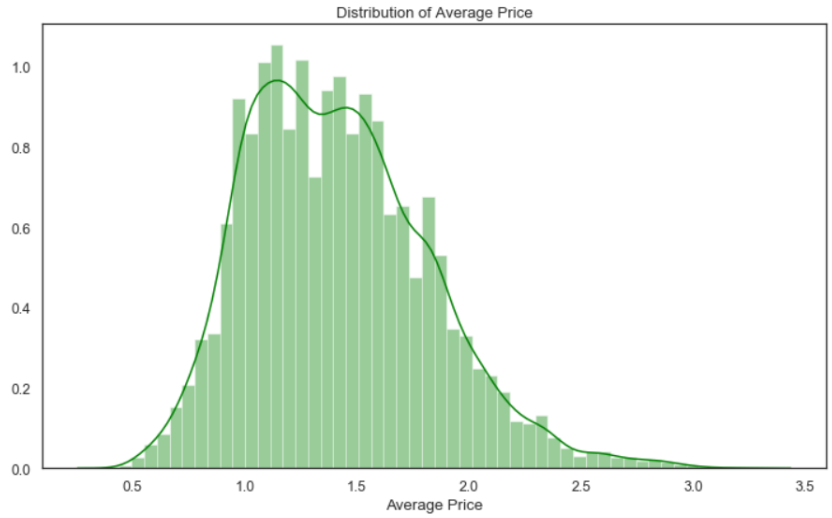
By getting a concise summary of the dataset, I observe that there are no missing values in this dataset. Hence, we do not need to treat them.

Further, breaking down the date column into separate columns for day, month and year, reveals that these observations are weekly – they are recorded for Sunday of every week over 2015-2018.

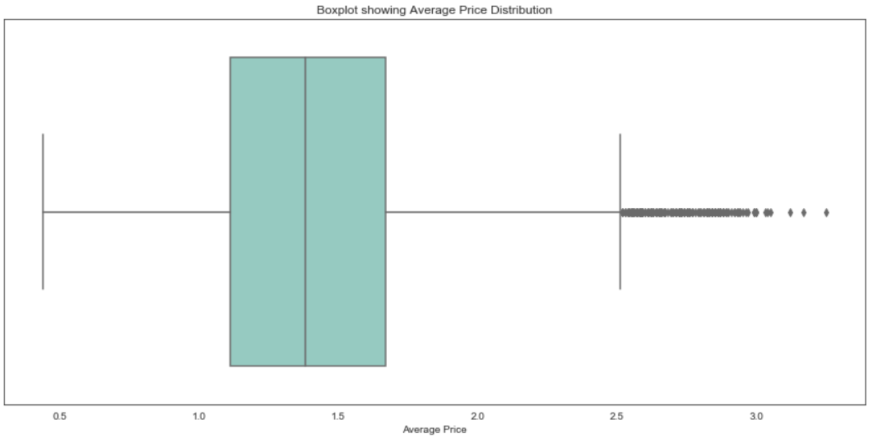
I also added an additional column of Total Sales (Average Price * Volume) to get the dollar value sales.

Analysis

First, let's look at the distribution of average prices over 2015-2018:

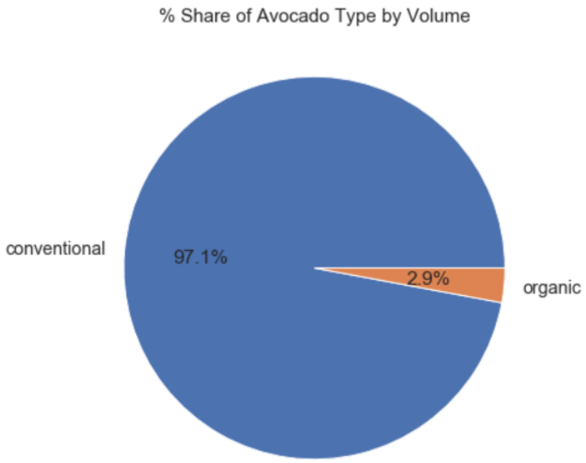


The above histogram shows that the average avocado price is nearly normally distributed with prices ranging between $1.1-$1.6 per avocado. However, there have been some instances when the price increased to more $3.0. Further, let's use a boxplot to understand the spread of average prices and visualize the outliers:

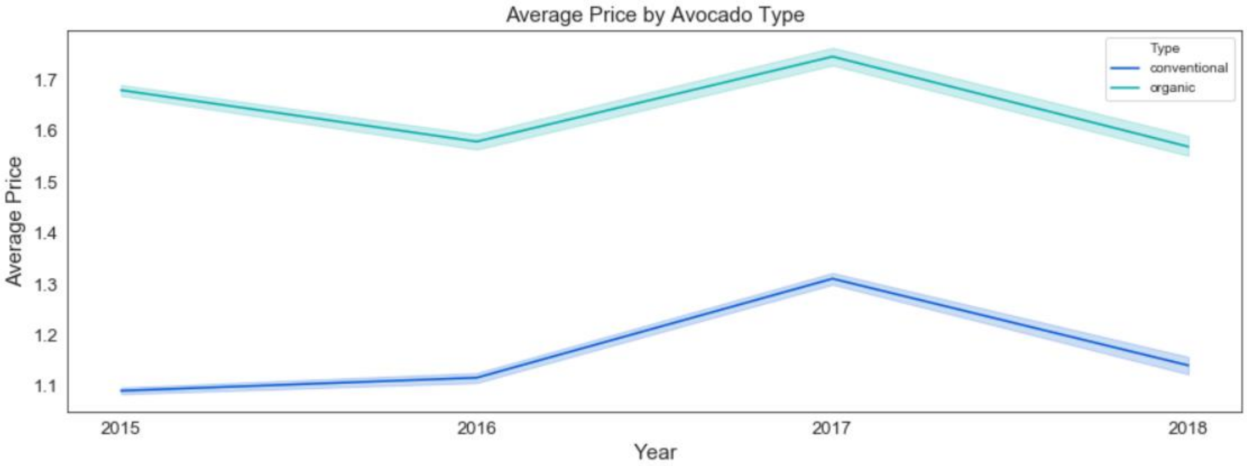


In line with the histogram, the boxplot also follows nearly a normal distribution, with all prices above $2.5 being considered as outliers.

Next, let’s look at the percentage share of avocado sales by type over 2015-2018. Conventional avocados dominate the overall avocado sales accounting for ~97% of the total volume. However, organic avocados, which form only ~3% of the volume sales, are slowly gaining traction as organic agriculture has been on the rise since 2015. A quick google search shows that conventional avocado growers are slowly transitioning to growing organic avocados to meet the strong demand, as its taste and health properties turned it into a major food trend. The main difference between organic and conventional avocados are the chemicals involved during production and processing.

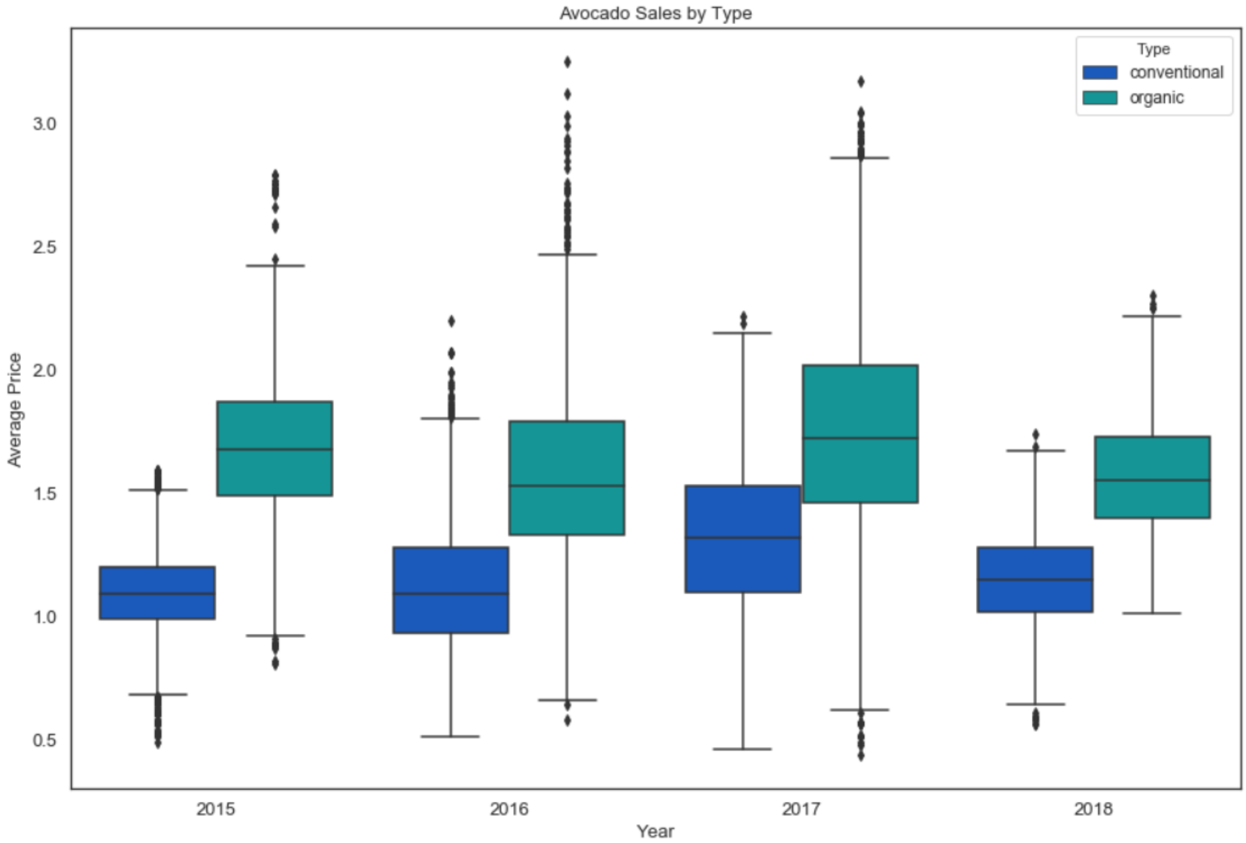


Diving deeper into these types, lets analyze how their prices fluctuated over 2015-2018.



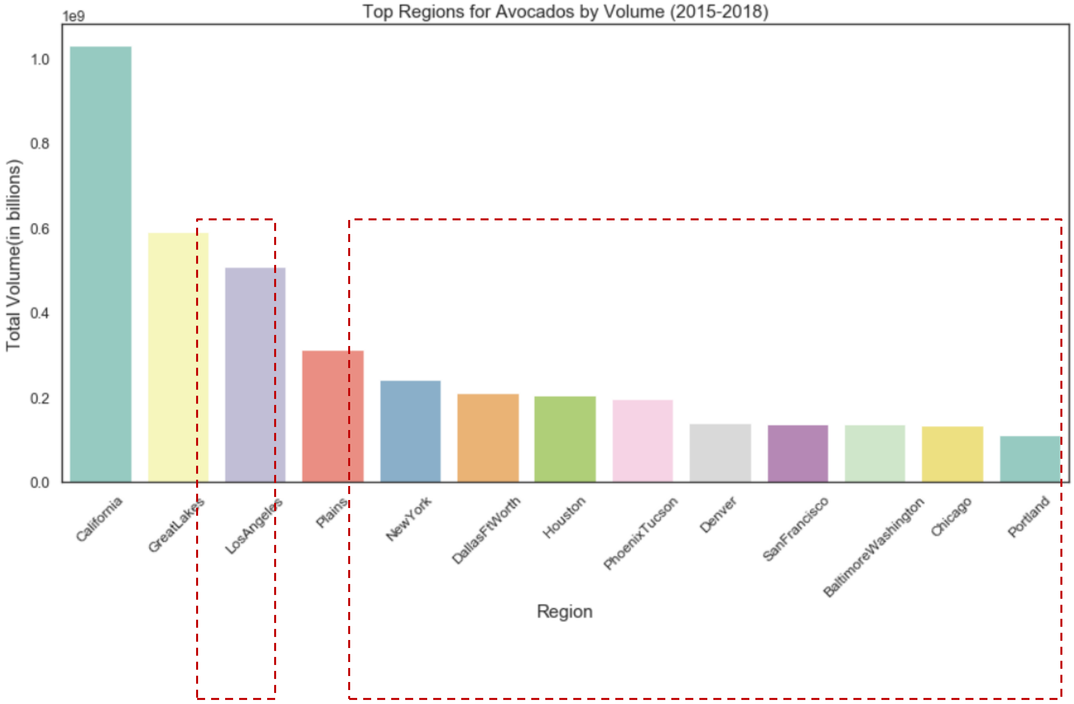
Seeing the above figure, it is clear that organic avocados are expensive than the conventional ones. Also, irrespective of the type, there was a peak in the avocado prices in 2017. This was due to the surging global demand and reduced harvests from major avocado producing regions including Mexico and California – farmers strike in Mexico and a major drought in California in 2016 led to a severe supply crunch, elevating avocado prices in 2017. In 2018, these prices are seen to be reaching their normal levels.

The below boxplot also shows the same trend as discussed above.



Top 10 Avocado Loving Cities in the US:

Looking at the region-wise price dynamics, let’s first see which cities had the highest avocado sales during the period 2015-2018.

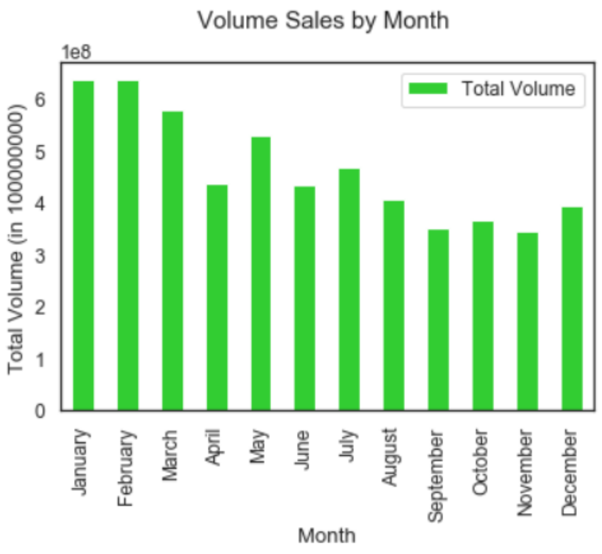


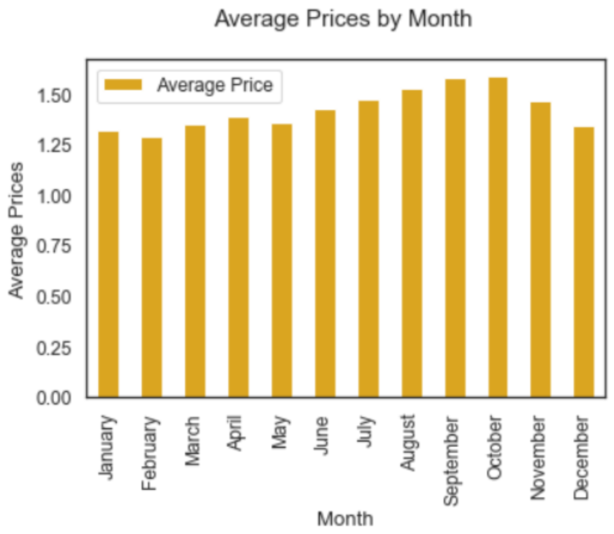
Considering cities only, the above graph illustrates that the Los Angeles, New York and Dallas were the top three avocado loving cities in America over the past four years.

Los Angeles had the highest volume of avocados being sold, which is partly due to the demographics of the city. It is the second most populated city in the US, with a population of ~4 million in 2017. Also, California is the leading producer of Hass avocados and home to ~90% of the nation's crop, which makes Los Angeles as the most avocado consuming city.

An interesting thing to note here is that 2 major cities of California – Los Angeles and San Francisco, are among the top 10 cities that love avocados, owing to the fact that Hass avocados are native to California.

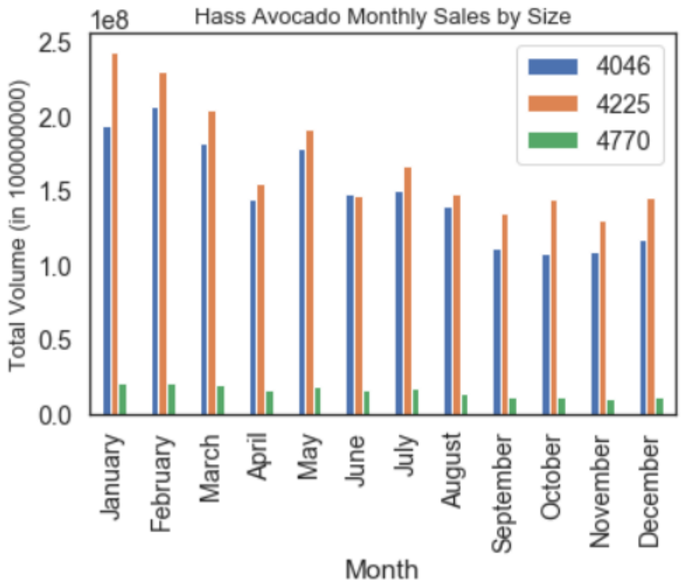
Relationship between Average Price and Volume



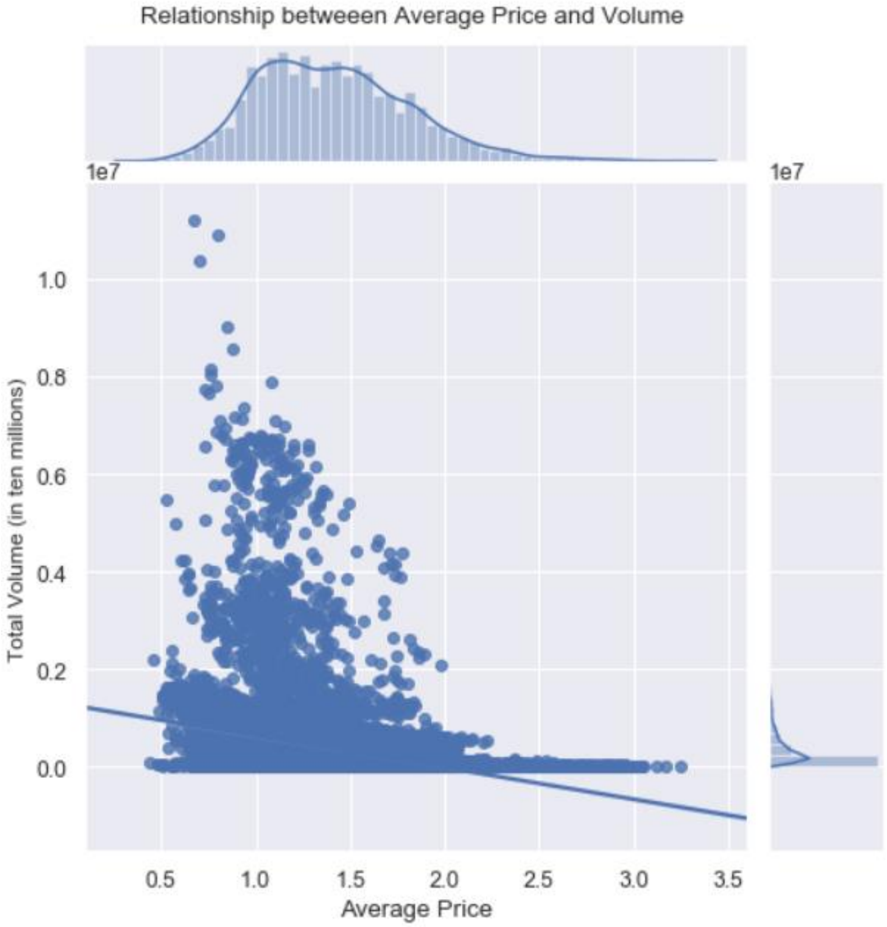


In the above chart depicting total volume over 2015-2018, the months of January, February and May witnessed the highest sales whereas the ending months from September - December witnessed the least sales. The monthly sales of Hass avocados (chart on the right) also depict the same behavior Some further research indicates that this may be due to the Hass avocado growing season, which generally starts from January to the end of October or early November.

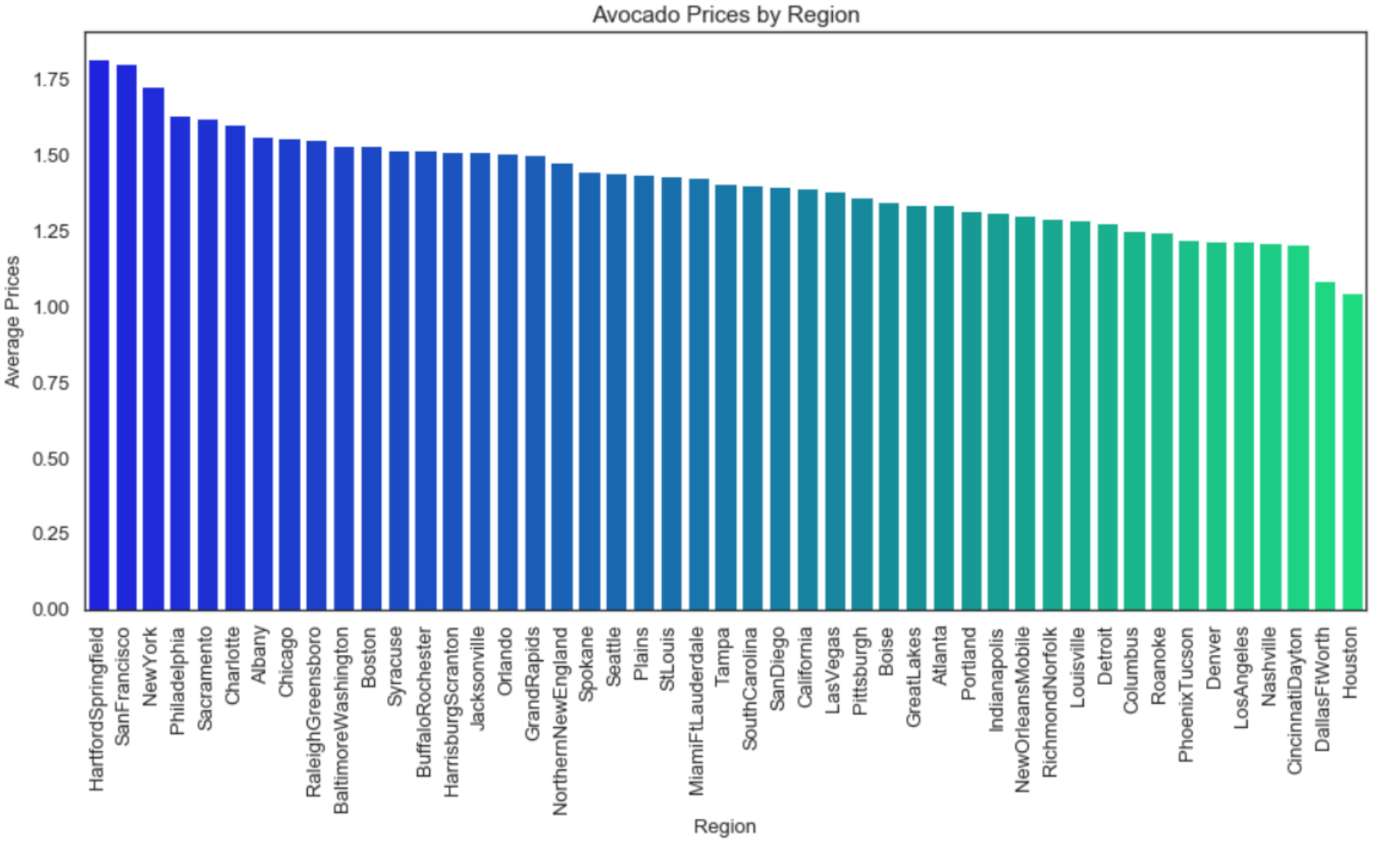
On the other hand, the average price per avocado has been the highest during the ending months of September – December.



This shows that as the prices rise, sales of avocados seemed to increase and when the prices fall, the volume sold increased. This shows that there is an inverse relationship between them. Also, this relation can be displayed by the jointplot as shown below:



Cities which have the most expensive avocados:



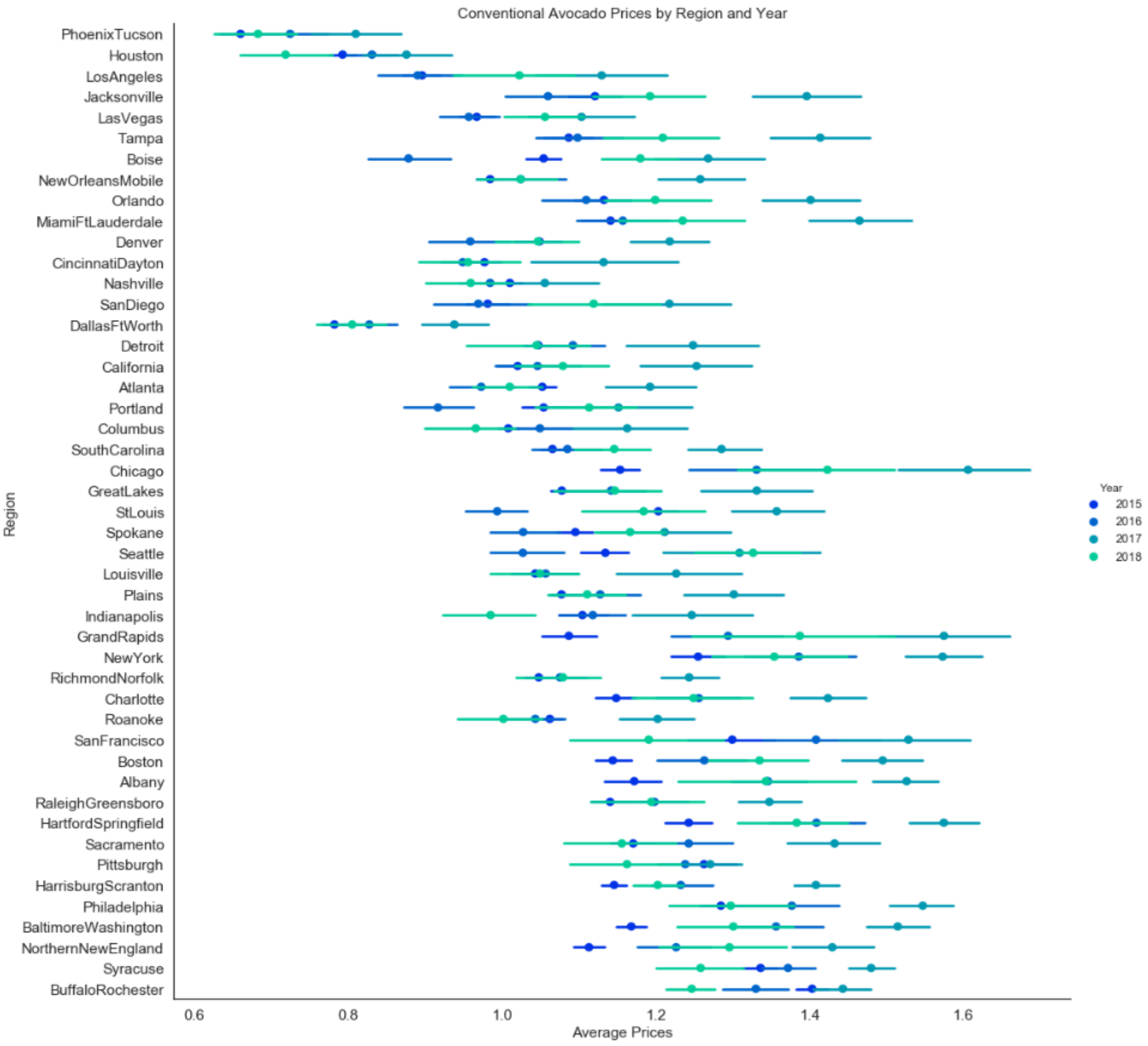
The above graph shows that avocado prices vary greatly across cities. Considering only the cities, San Francisco was the most expensive to buy avocados in, followed by New York, Philadelphia, Sacramento and Charlotte. It can also be seen that the prices seem higher in the north eastern and north western states.

One possible explanation for the prices being high in these regions is their location. Since the majority of avocados are produced and harvested in Mexico and California, the cost of transportation is lower for transporting avocados to the southern cities such as Dallas, Houston and Phoenix, as compared to the north eastern and north western cities.

Additionally, the prices of avocados are somewhat directly related to the cost of living in those particular cities. For instance, the cost of living is higher in cities such as Boston, Seattle or New York as compared to Houston, Denver and Dallas - that is the reason why avocado prices are also higher in these cities when compared to Houston.

Now, let’s dive into how the price dynamics differ across regions as well as type.

Regional level price variations Conventional avocados

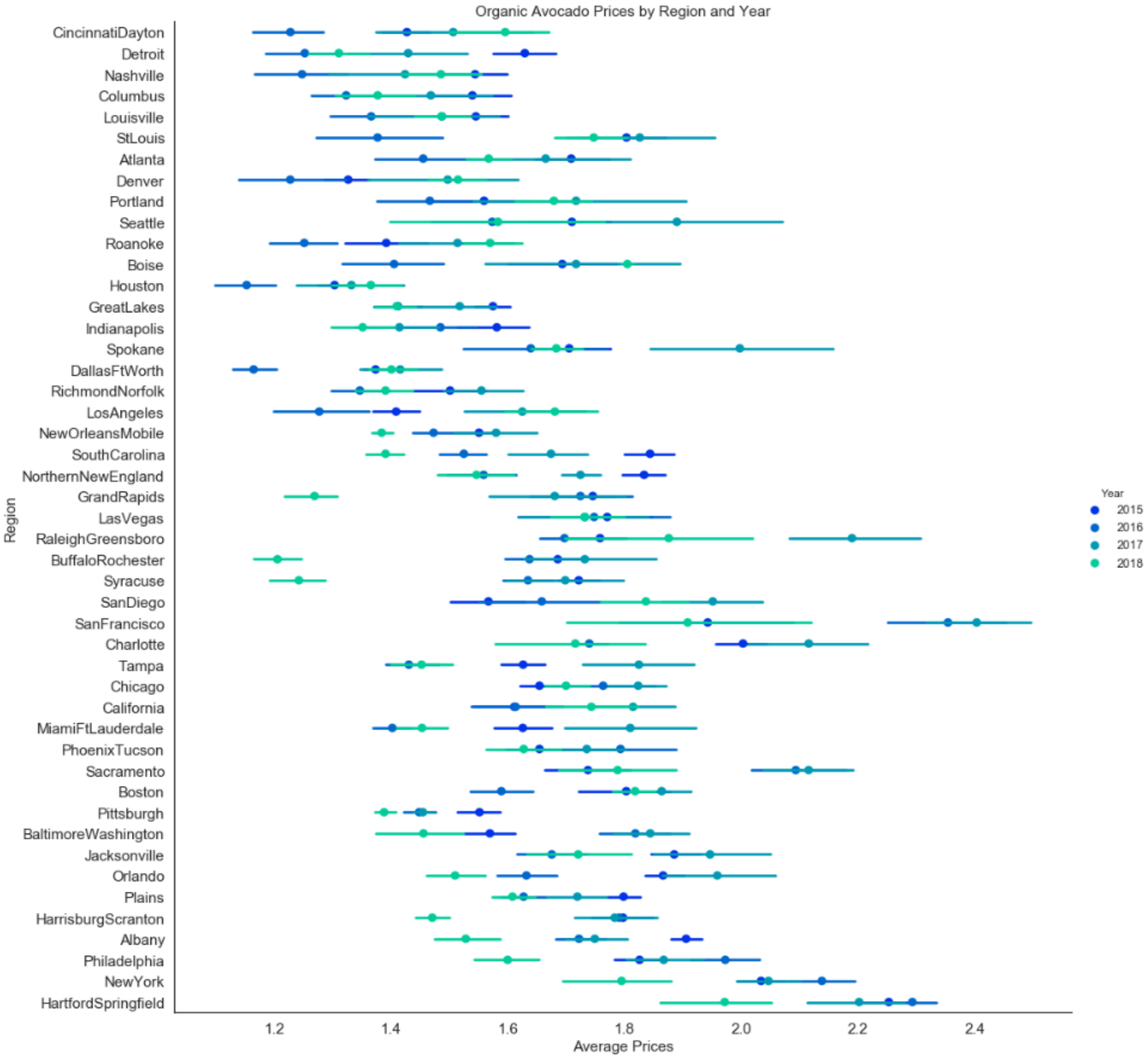


The above chart shows that PhoenixTucson has the cheapest conventional avocados over the period 2015 to 2018. On the other hand, Buffalo-Rochester has the most AveragePrice of conventional avocados

Grand Rapids region had the most variance in AveragePrice through 2015 to 2016, followed by Chicago.

This graph also tells us that owing to consumers becoming more health conscious, Conventional avocados had a strong market demand and they were becoming more expensive from 2015 to 2018 regardless of the regions.

Regional level price variations Organic avocados

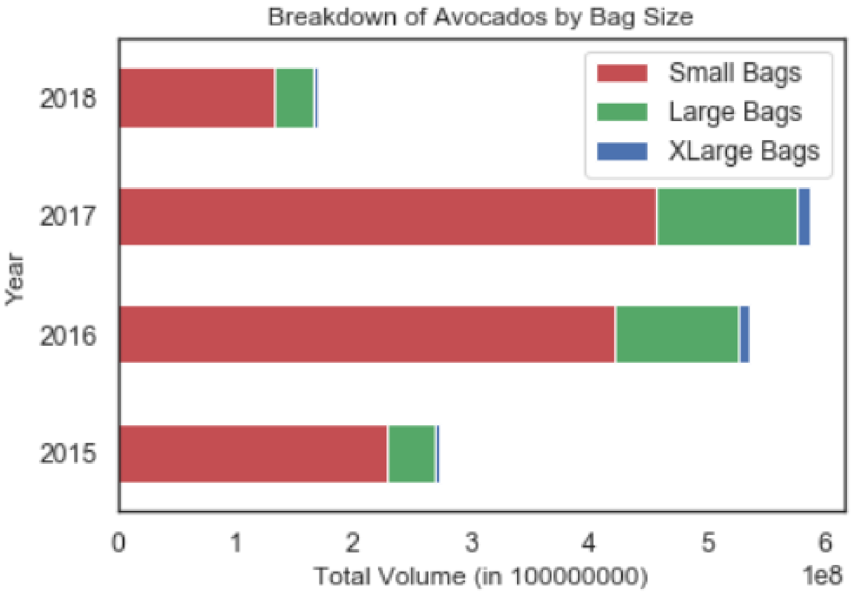


The above chart shows that DallasFtWorth has the cheapest organic avocados over the period 2015 to 2018. On the other hand, San Francisco has the most AveragePrice of organic avocados.

Further, San Francisco also had the most variance in AveragePrice through 2015 to 2018, followed by BaltimoreWashington.

Overall, both the above graphs show that the average price varies across regions, and thus it can be inferred that region plays a critical role in predicting the average prices of avocados.

Bag size dynamics over the years



The above chart shows that consumers prefer buying Small Bags of avocados as compared to Large and XLarge Bags. However, as avocados gained popularity amongst millennials, we can see that the share of Large bags has also increased. From 2015 to 2016, the number of bags sold almost doubled, with a similar growth being registered by the Large bags. In 2017, the growth seemed to have been low due to the abnormally high prices (as discussed above). Although we only have data for the first 3 months of 2018, but a snapshot of this affirms that this trend continued in 2018 as well.

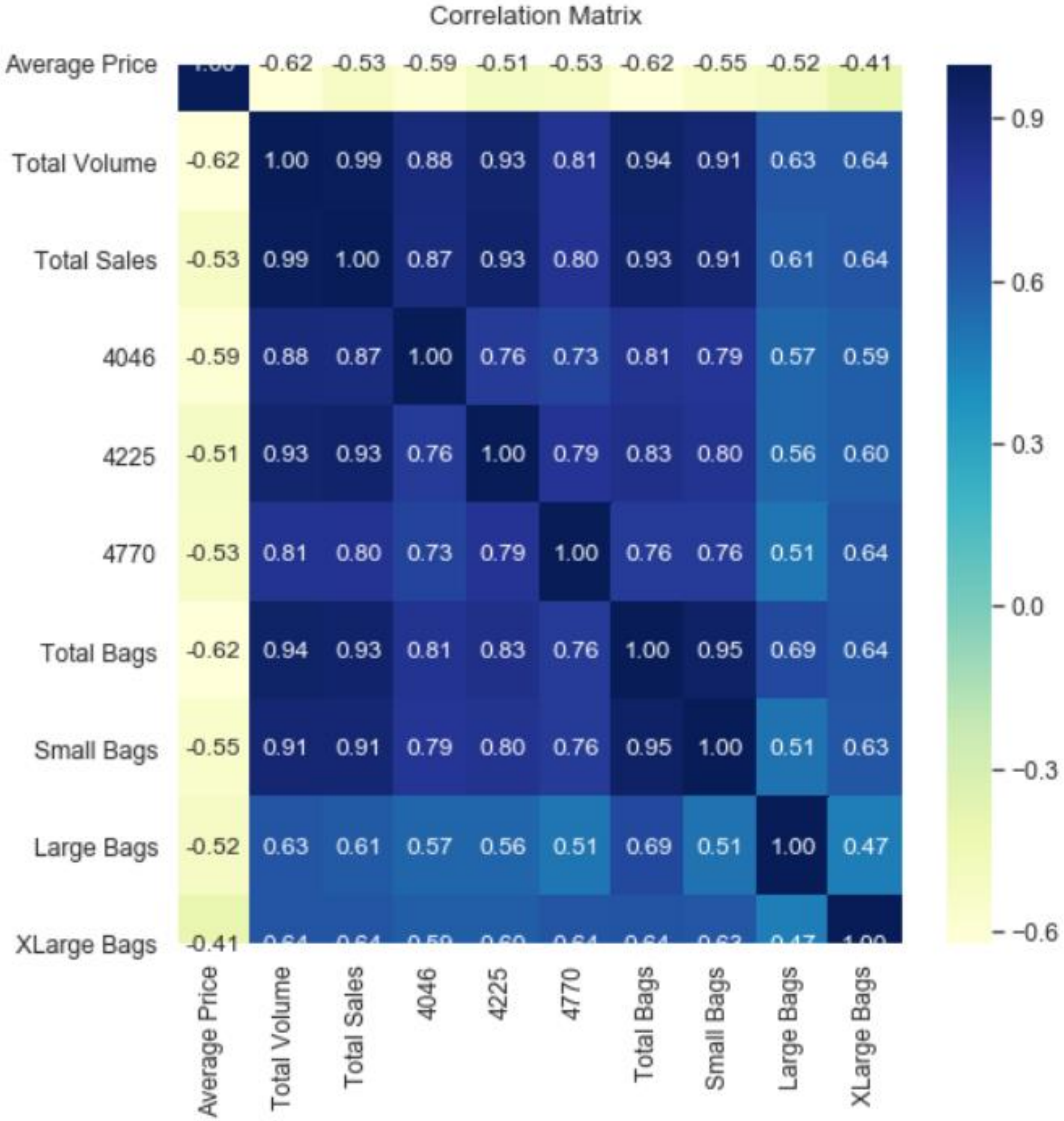
Now, since we have drilled down the data and understood the distributions as well as what was happening with the avocado prices over 2015-2018, let's see how each of the above factors have a bearing on each other.

For this, I will make a correlation matrix to see how these factors correlate amongst themselves.

Just a quick reminder, the reason why we check the correlations among columns is to imply the hints of the relationship of variables. However, correlation does not always mean causation. If two variables have high correlation, we may just conclude that there may be some relationship between the two.

Below is the correlation matrix.

Correlation Matrix



Correlation Matrix

	Average Price	Total Volume	Total Sales	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags
Average Price	1.00	-0.62	-0.53	-0.59	-0.51	-0.53	-0.62	-0.55	-0.52	-0.41
Total Volume	-0.62	1.00	0.99	0.88	0.93	0.81	0.94	0.91	0.63	0.64
Total Sales	-0.53	0.99	1.00	0.87	0.93	0.80	0.93	0.91	0.61	0.64
4046	-0.59	0.88	0.87	1.00	0.76	0.73	0.81	0.79	0.57	0.59
4225	-0.51	0.93	0.93	0.76	1.00	0.79	0.83	0.80	0.56	0.60
4770	-0.53	0.81	0.80	0.73	0.79	1.00	0.76	0.76	0.51	0.64
Total Bags	-0.62	0.94	0.93	0.81	0.83	0.76	1.00	0.95	0.69	0.64
Small Bags	-0.55	0.91	0.91	0.79	0.80	0.76	0.95	1.00	0.51	0.63
Large Bags	-0.52	0.63	0.61	0.57	0.56	0.51	0.69	0.51	1.00	0.47
XLarge Bags	-0.41	0.64	0.64	0.59	0.60	0.64	0.64	0.63	0.47	1.00

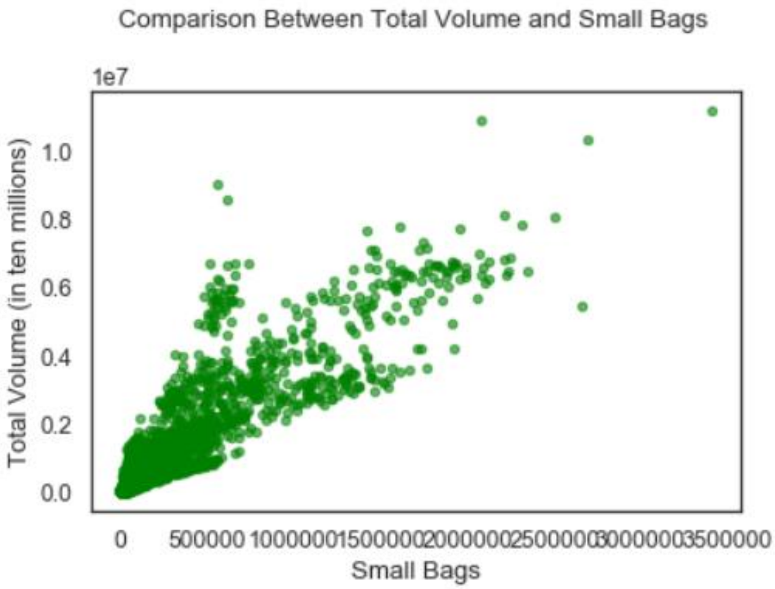
From the above matrix, we see that darker boxes imply a high and strong correlation.

As seen before, there is a negative correlation of -0.62 between average prices and total volume. This means that as prices increase, the demand for avocados or the volume of avocados sold decreases, and vice versa.

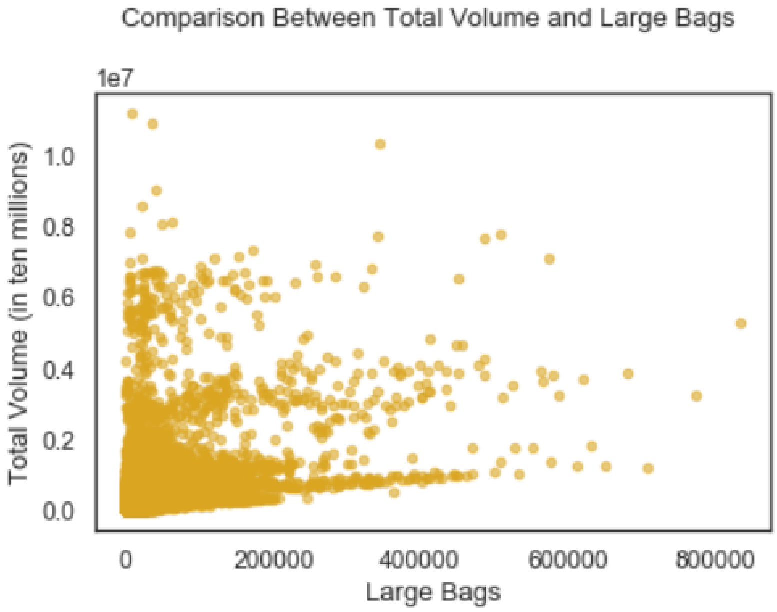
Total Volume is also more strongly related to the PLU 4225 (large Hass avocados) as compared to other variants of Hass avocados.

We can easily observe that Average Price has nothing to do with the type or number of bags. But we can say that as Total Volume gets a higher value, the number of Small Bags increases faster than Large Bags and XLarge Bags.

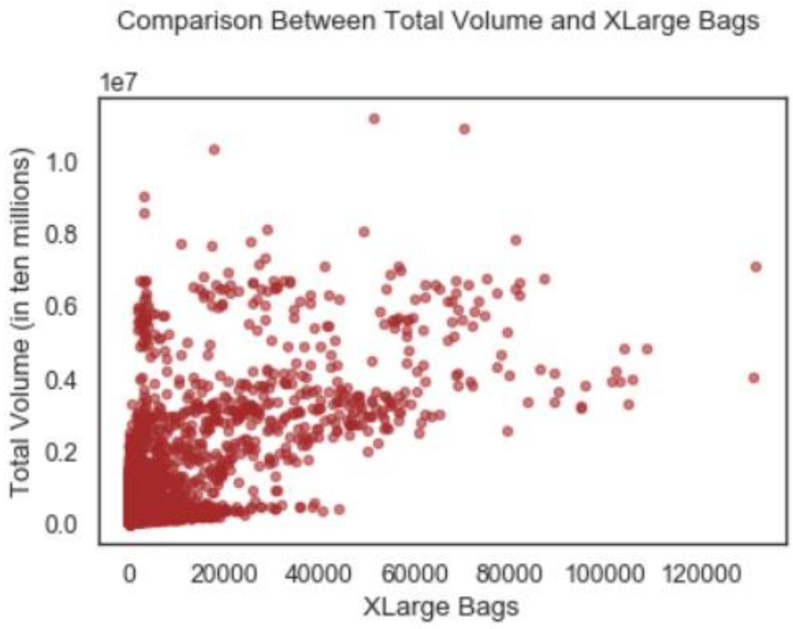
Further, lets visualize this correlation a little better by using a scatter plot and examine the correlation between the volume of a single avocado (Total Volume) and the type of bag (Small Bags, etc)



The scatter plot shows how there is a strong linear relationship between the total volume and small bags



The scatter plot shows that there is a linear relationship between total volume and large bags. However, this correlation is not as strong as for small bags



Again, the scatter plot here also shows that there is a linear relationship between total volume and XLarge bags. However, this correlation is not as strong as for small bags and large bags

The above information might be helpful for retailers who are selling avocado bags to figure out which bags consumers prefer and which ones they should stock more on the shelves.

Next Steps

As next steps, I plan to continue working on the project and run a multiple linear regression model to see which of these variables significantly contribute to the average prices of an avocado. Additionally, I also plan to use other machine learning models to predict the average prices.

References

- Hass Avocado Board (https://hassavocadoboard.com/)

- Kaggle Dataset https://www.kaggle.com/neuromusic/avocado-prices

- Correlation vs Causation https://www.statisticshowto.datasciencecentral.com/causation/

- https://www.msn.com/en-us/foodanddrink/foodnews/millennials-not-alone-in-driving-up-us-avocado-consumption/ar-AAFpVv1

- https://www.washingtonpost.com/news/food/wp/2017/05/02/avocado-prices-are-at-a-record-high-just-in-time-for-cinco-de-mayo/

- https://www.bloomberg.com/news/articles/2017-04-28/guacamole-costs-to-jump-as-avocado-shortage-sparks-record-prices

- https://www.bbc.com/news/business-39768480