**Introduction**

Avocado, a bright green fruit with a large pit and dark leathery skin, has become a fashion among millennials in America. Avocados featured in its unique flavor, rich texture, and high percentage of healthy fat, which is an ideal food for the healthy diet. However, the average price for avocados is not cheap, which made those millennials unable to achieve their “Avocado Freedom” dream. In addition, as an applicant in American master and a fan of avocado, it is of significant interests for me to conduct the explorative data analysis of avocados, hoping to figure out some hidden patterns, like seasonal patterns or cyclical patterns. Furthermore, related factors were also evaluated to suggest the influence on the avocados’ price, which could help these millennials to find the cheaper avocados in a scientific way.

**Methods and Results**

**Collection and preprocessing of the avocado dataset**

The dataset was downloaded from Kaggle (*Avocado Prices (2020) | Kaggle*, 2020), which contains the data from 4 January 2015 to 17 May 2020. Specifically, the dataset directly scanned from retailers’ cash registers based on actual retail sales of Hass avocados. Avocados in this dataset could be seperated into to two categories, either conventional or organic; the corresponding average price and volume of avocados sold differs on types and geographical locations. In this experiment, the dataset does not contain any missing value and I mainly focused on six parameters including date, average price, total volume, year, type and geography, aiming to find the best strategy in buying avocados.

**Explorative data analysis**

**Comparison between different avocado types**

In this dataset, there are two types of avocados, conventional and organic. I would like to compare the difference in both the average price and the volume of these avocados sold. Hence, I firstly plotted a density plot to demonstrate the distribution of average price among two types of avocados. Figure 1 indicates that the average price for organic avocados distributed more evenly and the average price could be more expensive than conventional type. Conventional avocados’ average price centered around 1 dollar while organic centered around 1.5 dollar.

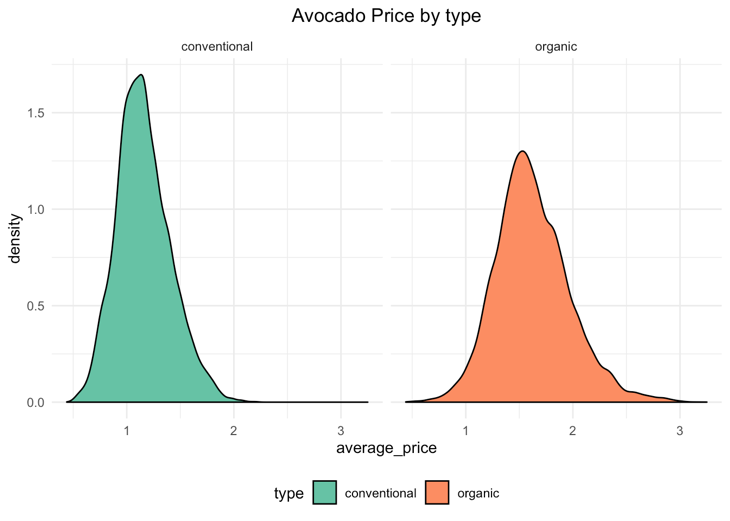


Figure 1: Density plot of avocado price by type

Table 1 was created for better illustration of the correlation among these two types. As can be seen from the table, conventional avocado could account for approximately 97% in the markets and the average price would be 28% lower than organic type. Furthermore, the trends of average price among the two types could be visualized in Figure 2. It is clear that the average price of conventional avocados is continuously lower than organic one since green stands for organic type while red stands for conventional type.

**Table 1: Average volume and price comparison among two types**

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Average volume | Average price | Volume percent |
| Conventional | 1,818,206 (1.8 M) | 1.16 $ | 96.8% |
| Organic | 60,127 (0.06 M) | 1.62 $ | 3.2% |

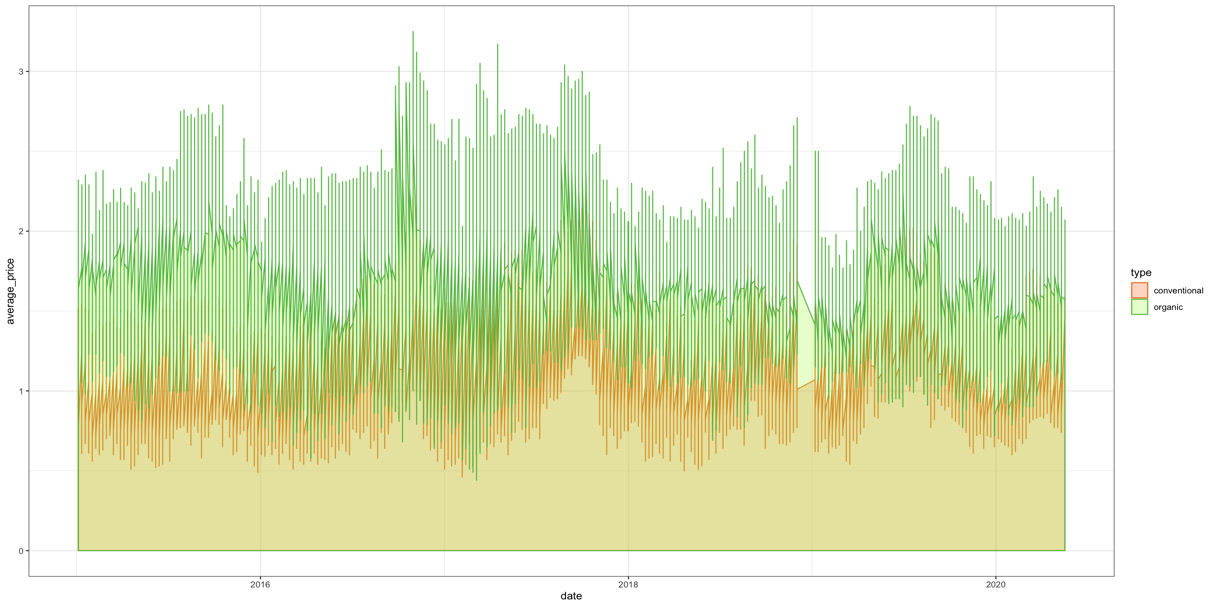


Figure 2: Avocado price with the date in two types of avocados

**Correlation between avocados’ price and volume**

Additionally, it is a common view that the volume and the price could be negative correlated since the customers’ demand would increase as the avocado price decrease. I would like to see whether two types of avocados follow these rules. In this case, Figure 3 illustrates the average price and volume among organic and conventional avocados from 2015 to 2020 in each month. As can be seen from the figure, the overall trend for conventional avocados volume increases steadily while the organic avocados volume starts to decrease after 2019. Furthermore, we could notice that the volume undergoes some periodic change. In other words, the avocado price and volume may have some seasonal patterns.

图形用户界面, 图表

描述已自动生成

Figure 3: Average price and volumes trends of conventional and organic avocados from 2015 to 2020

**Seasonal patterns of price and volume analysis per year**

In order to elucidate the seasonal patterns of avocados, I categorized the samples based on seasons: Spring (Mar-May), Summer (Jun-Aug), Autumn (Sep-Nov), Winter (Dec-Feb), and calculated the average price and volume respectively from 2015 to 2020 (Figure 4). From the perspective of avocado volume sold, Spring would be the highest in either conventional or organic avocados sold, indicating that American people really love to buy avocado in Spring. From the perspective of average price, however, the patterns are not consistent, either winter or spring could be the season with lowest price, which stimulates me to explore the data monthly.

表格

描述已自动生成

Figure 4: Average price and volumes sold of two types avocados by season

**Monthly patterns analysis of price and volumes per year**

图表, 折线图

描述已自动生成Figure 5 compares the average price and volume per month from 2015 to 2020. As can be seen from the figure that the average price would reach the peak during September or October and reach the bottom in next year’s February. Regarding to the average volume sold, American people prefers to buy avocados at next year’s February or May and the willingness achieved lowest in November.

Figure 5: Average price and volumes sold of two types avocados per month

**Geographical analysis of the cheapest avocados**

For the purpose of finding the cities or regions in United States with cheapest avocados, I thus returned the top six places and visualized with the flipped bar chart (Figure 6). Phoenix/Tucson, with the average price of 0.78 dollars, would be the cheapest region in the conventional avocados. Meanwhile, organic avocado could find its lowest average (1.34$) price in Dallas/Ft. Worth.

表格

描述已自动生成

Figure 6: Six cities or regions with lowest average price in United States

**Future average price of avocados trend prediction**

In addition to the explorative data analysis (EDA), it is of great significance to predict the future price. Therefore, I utilized the R package “prophet” to conduct the predictive data analysis.(Taylor and Letham, 2018). Figure 7 is the combination of both history avocado price and the predicted price indicated by dashed plot. The predicted price of avocado would continue to grow but the increase rate would slow. Figure 8 includes the trend of avocado average price from daily, weekly, monthly and yearly. The positive trend value indicates the price increase while the negative value indicates the decrease. From the subplot A, the average price of avocados continues to grow and reaches the peak in 2017, with a speed higher than 1.3. Regarding to the future prediction, the range is about [1.20-1.34]. Furthermore, the subplot B reconfirms the previous research that starting from May, the average price continues to grow until October. The subplot C suggests the average price in weekends were relative higher than weekdays. The subplot D indicates the time per day, indicating the price change occurred at midnight.

图表

描述已自动生成

Figure 7: Average avocado price trend in next year

图表, 折线图

描述已自动生成

Figure 8: Average avocado price trend per day/month/week/year

**Discussion and Conclusion**

The explorative-based and prediction-based data analysis indicates the avocado market situation in United States. There are two major types of avocados, either organic or conventional. Conventional avocados with average 1.16$ (28% lower than organic type) account for almost 97% of the market, suggesting most American people are sensitive to price, they would choose conventional avocado compared with organic type under the same condition. Moreover, the average price and the volume of avocado sold indicates some negative correlation, which obeys the common view that customers’ demand would decrease as the price increase. Besides, avocados’ price and volume indicate some seasonal patterns which normally reflected in the fruits. The highest price of the whole year would during September or October while the highest volume sold were in February or May. The research indicated that avocado would ripe in approximate September per year (*Is that avocado ready for harvest? Here’s what to look for in your garden | Home/Garden | nola.com*, 2017). With the necessary conservation and transportation, avocados would appear on the market during October. At that time, the previous year’s remaining avocados would reach the minimum and this situation would trigger the increase of price. As a result, with the consistent supply of avocados, the price starts to decrease until February. Simultaneously, since the price reached the button, the avocado volume sold would reach the peak. In addition, the top six regions with the cheapest avocado price were returned from history data. Except for the Nashville, the top five cheapest regions of conventional avocado are all in the middle west of America, near to Mexico, where avocados originated. Based on the research, these places are the centers of avocado planting which therefore achieving such a low price compared with others (*What Types of Fruit Trees Grow in Florida? | Hunker*, no date). Apart from the explorative data analysis, the future price trend on avocados were predicted to conduct the predictive data analysis. The avocado price in the following year would continue to grow, but with a lower increasing rate, ranging from [1.20-1.34]. Furthermore, with the strike of Covid 19 pandemic, people’s consumption of avocados would decrease concomitantly. In this case, the following year’s avocado price increasing rate would be about 1.20.

It is unavoidable that the research could contain some limitations and need to be improved in the future analysis, specifically, the prediction model. In this experiment, I utilized the existing R package developed by Facebook to conduct the avocado price prediction, however, the internal mechanisms of this package are difficult to elucidate because of its intrinsic complexity. In other words, the price prediction model was a “black box problem”. Therefore, in the future, I would build the time-series prediction model individually and emphasis on the interpretability of the model, for instance, the Smoothing Moving Average (SMA), Seasonal Naive Method and Drift Method.

**Reference**

*Avocado Prices (2020) | Kaggle* (no date). Available at: https://www.kaggle.com/timmate/avocado-prices-2020 (Accessed: 4 December 2020).

*Is that avocado ready for harvest? Here’s what to look for in your garden | Home/Garden | nola.com* (no date). Available at: https://www.nola.com/entertainment\_life/home\_garden/article\_c44ca588-7f9c-5402-8338-a12f9c531fcd.html (Accessed: 16 December 2020).

Taylor, S. J. and Letham, B. (2018) ‘Forecasting at Scale’, *American Statistician*, 72(1), pp. 37–45. doi: 10.1080/00031305.2017.1380080.

*What Types of Fruit Trees Grow in Florida? | Hunker* (no date). Available at: https://www.hunker.com/12312954/what-types-of-fruit-trees-grow-in-florida (Accessed: 16 December 2020).