Analysis on Avocado Prices

Authors: Katie Blau, George Jiang, Benson Yu

1. Description of Topic Area

The topic we chose to explore is avocado prices in the US markets. We thought this would be an interesting topic to explore because in recent years, the popularity of avocados has risen because of foods like guacamole, using avocados as additions to smoothies and salads, and using them as a replacement for oil, butter, and cream in recipes for desserts. We also all like eating avocado toast – an extremely popular food item that has helped this rise in popularity. Avocados are also known to fluctuate in price because of availability and importing concerns, so examining the prices helps provide us with some insight on any price trends that may arise (for example, depending on time of year).

From this dataset, we analyzed avocados on a variety of different factors including type of avocado, region, seasonality, PLU (price look-up code), and the overall number of avocados sold. There are also several questions we explored in our analysis. More specifically, we questioned if the type of avocado (conventional or organic) impacted the average price of avocados. We also asked, do certain regions have different demand levels for avocados overall? Were avocados sold in the summer months more expensive than avocados sold in winter months? How did the PLU impact the number of avocados sold?

2. Description of Data Set

To get started on the investigation, the data needed to be read from the .csv file and converted into a suitable structure that can be explored, cleaned, and analyzed according to the needs of answering our hypotheses. The .csv file was read into RStudio named avocado and the first few rows are shown in **Figure 1.** The full size of the data set is 18,249.

...1	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.00	conventional	2015	Albany
1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.00	conventional	2015	Albany
2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.00	conventional	2015	Albany
3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.00	conventional	2015	Albany
4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.00	conventional	2015	Albany
5	2015-11-22	1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0.00	conventional	2015	Albany
6	2015-11-15	0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0.00	conventional	2015	Albany
7	2015-11-08	0.98	109428.33	703.75	101815.36	80.00	6829.22	6266.85	562.37	0.00	conventional	2015	Albany

Figure 1. The first 8 columns of data from the avocado DataFrame.

In the data set, there were a variety of quantitative and qualitative variables including:

- Date: The date of the observation (Quantitative)
- Average Price: the average price of a single avocado (Quantitative)
- Type: conventional or organic (Qualitative)
- Year: the year (Quantitative)
- Region: the city or region of the observation (Qualitative)
- Total Volume: Total number of avocados sold (Quantitative)
- #4046: Total number of avocados with PLU 4046 sold (Quantitative)
    - Non-organic small/medium Hass Avocados (~3-5 oz)
- #4225: Total number of avocados with PLU 4225 sold (Quantitative)
    - Non-organic large Hass Avocados (~8-10 oz)

- #4770: Total number of avocados with PLU 4770 sold (Quantitative)
    - Non-organic extra large Hass Avocados (~10-15 oz)

In addition to these columns, there were four other qualitative variables that were not as relevant to our analysis: Total bags, Small bags, Large bags, and X-Large bags. There was not enough information provided to determine what exactly these variables were. We thought the size of the bag most likely corresponded to the number of avocados in each bag, but it was not clear as to whether the numbers collected were the number of bags sold or possibly the number of bags made, so we did not use these variables in our analysis.

The data set “Avocado Prices” came from Hass Avocado Board (HAB). The data set was downloaded from the website in May of 2018 and compiled into a single csv file. Hass Avocado Board is dedicated to making avocados America's most popular fruit. While Hass Avocado Board may have a bias towards emphasizing avocados and potentially increasing the number of avocados sold, the data was gathered directly from retailers’ cash registers based on actual retail sales of Hass avocados. This means that it was up to the retailers who kept track of the number of avocados sold. Therefore, it was unlikely that the retailers had an incentive to markup the number of avocados sold, and thus the risk of bias in this data set was lower. However, all of the data was collected from 2015-2018, which did pose the risk of not having a broad enough data set over multiple years to see larger trends in avocado data.

In terms of data cleaning, there were not any missing data entries or any duplicates. This meant that the data was carefully organized and recorded. However, for the column “Date”, the data is written in the format *year-month-day*. Since we intended to find out whether the day of the week had an impact on the number of avocados sold, we changed the

formatting to represent a day of the week. To do that we created a new column in the dataframe, avocado, and used the function “weekdays” which automatically converts the *year-month-day* to the corresponding day of the week as shown in **Figure 2.** Our results are then shown in **Figure 3.**

```
> avocado$weekday <- weekdays(avocado$Date)
```

Figure 2. Using the function “weekdays.”

	...1	Date	weekday	AveragePrice	Total Volume	4046	4225
1	0	2015-12-27	Sunday	1.33	64236.62	1036.74	54454.85
2	1	2015-12-20	Sunday	1.35	54876.98	674.28	44638.81
3	2	2015-12-13	Sunday	0.93	118220.22	794.70	109149.67
4	3	2015-12-06	Sunday	1.08	78992.15	1132.00	71976.41
5	4	2015-11-29	Sunday	1.28	51039.60	941.48	43838.39
6	5	2015-11-22	Sunday	1.26	55979.78	1184.27	48067.99
7	6	2015-11-15	Sunday	0.99	83453.76	1368.92	73672.72
8	7	2015-11-08	Sunday	0.98	109428.33	703.75	101815.36

Figure 3. The results of Figure 2 on the DataFrame.

As for outliers in the data set, there were none. Data that might seem like they are outliers in the Average Price column were not seen as such because of our analysis on price fluctuations. Lastly, because we did not use the columns “Total bags, Small bags, Large bags, and X-Large bags,” we deleted these columns from the data set to get them out of the way for our analysis.

3. Hypotheses

Below are the questions that we answered in our data analysis. We wrote them in a way to do hypothesis testing.

1. Are organic avocados more or less expensive than conventional avocados?

- Null: the mean price of conventional avocados is the same as the mean price of organic avocados
    - $H_0 : \mu_{\text{conventional price}} = \mu_{\text{organic price}}$
- Alternative: the mean price of conventional avocados is less than the mean price of organic avocados
    - $H_a : \mu_{\text{conventional price}} < \mu_{\text{organic price}}$

2. Do certain regions have different demand levels of avocado overall?

- Null: the proportion of total volume of avocados for a specific region equals the proportion of total volume of avocados for another specific region
    - Example: $H_0 : p_{\text{atlanta}} = p_{\text{albany}}$
- Alternative: the proportion of total volume of avocados for a specific region does not equal the proportion of total volume of avocados for another specific region
    - Example: $H_a : p_{\text{atlanta}} \neq p_{\text{albany}}$

3. Are avocados sold in the summer months more expensive than avocados sold in winter months?

- Null: the mean price of avocados sold during the summer is the same as the mean price of avocados sold during winter
    - $H_0 : \mu_{\text{summer price}} = \mu_{\text{winter price}}$
- Alternative: the mean price of avocados sold during the summer is more expensive than the mean price of avocados sold during winter
    - $H_a : \mu_{\text{summer price}} > \mu_{\text{winter price}}$

4. How does the size of avocados impact the number of avocados sold?

- Null: the mean number of small/medium avocados sold is the same as the mean number of extra large avocados sold
    - $H_0 : \mu_{\text{number of small/medium avocados sold}} = \mu_{\text{number of extra large avocados sold}}$
- Alternative: the mean number of small/medium avocados sold is not equal to the mean number of extra large avocados sold
    - $H_a : \mu_{\text{number of small/medium avocados sold}} \neq \mu_{\text{number of extra large avocados sold}}$

4. Data Visualization

1. Are organic avocados more or less expensive than conventional avocados?

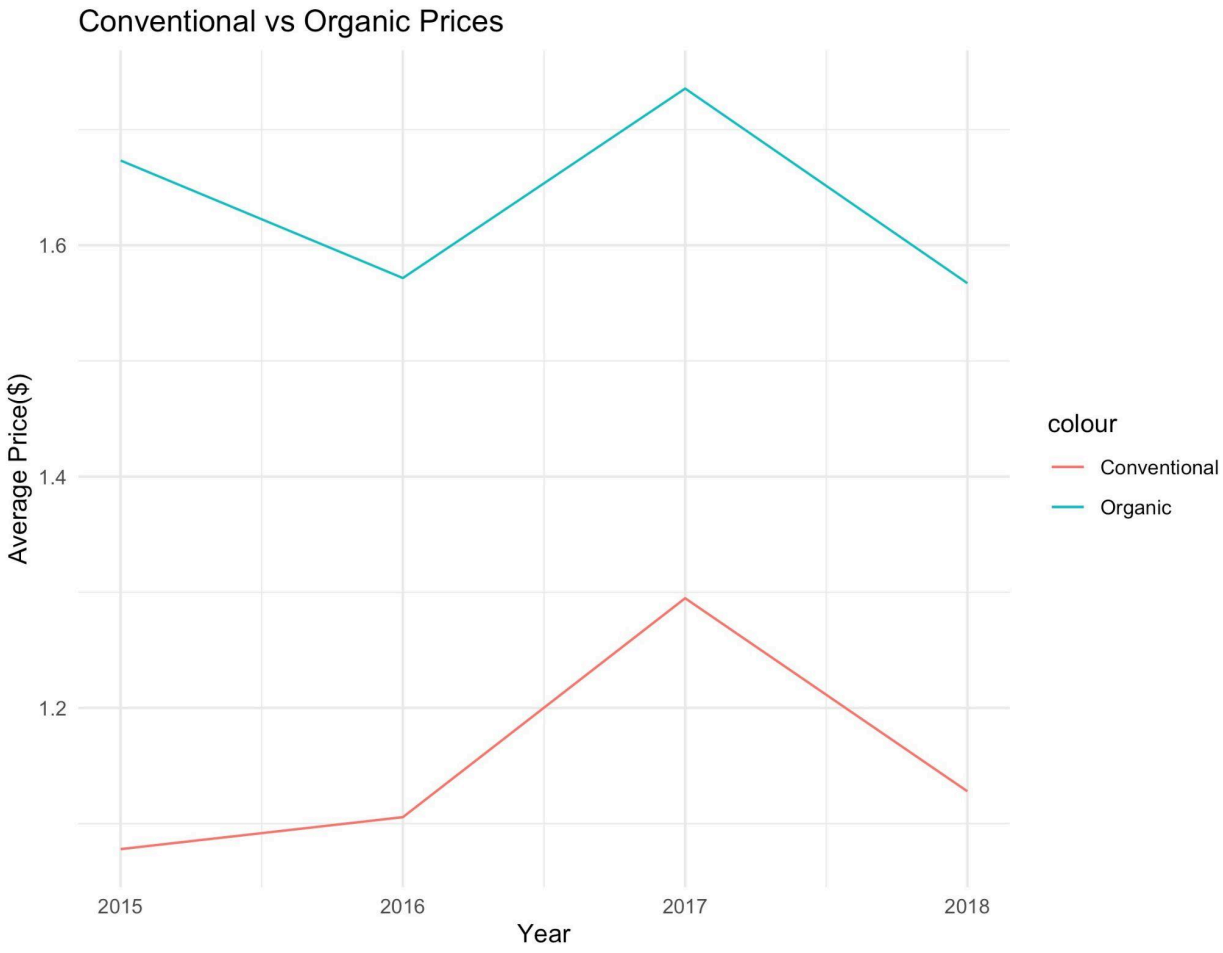

Figure 4. Line charts of conventional vs. organic avocado prices plotted against years.

This line chart above shows how the average prices of both conventional and organic avocados change over time. The red line represents the conventional avocados and the blue line represents the organic avocados. Both lines representing conventional and organic

avocados show a similar fluctuation in price over the years 2015 to 2018; however, the average price of the organic avocados are consistently higher than the average price of conventional avocados over time. In 2017, both types of avocados also show a spike in average price before they both go back down in average price in 2018. This line chart gives evidence to show that organic avocados are more expensive than conventional avocados, which supports our hypothesis.

2. Do certain regions have different demand levels of avocado overall?

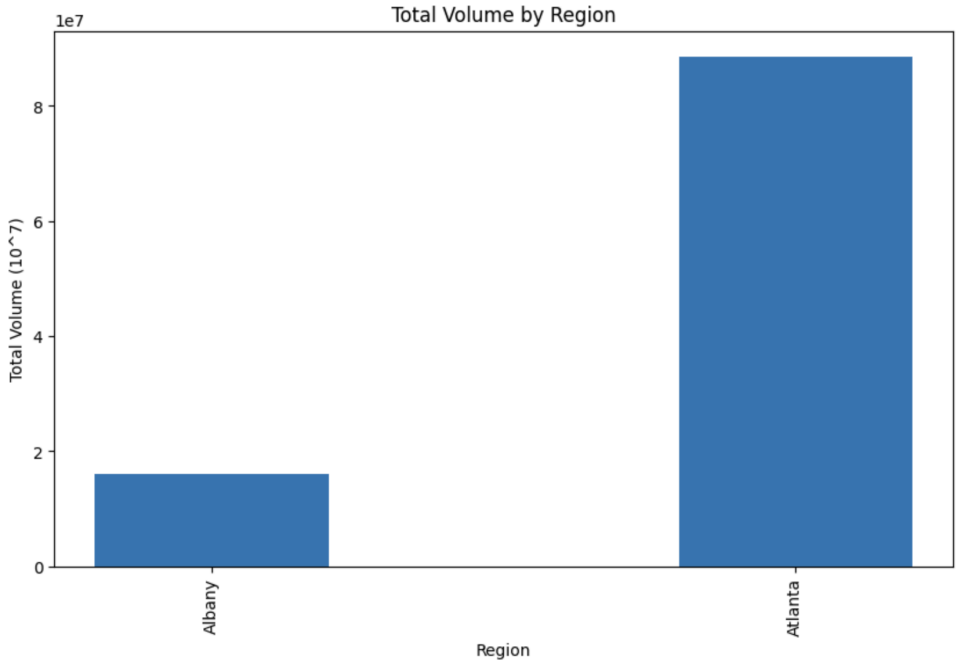

Figure 5: Bar graph of the total volume of avocados in Albany and Atlanta.

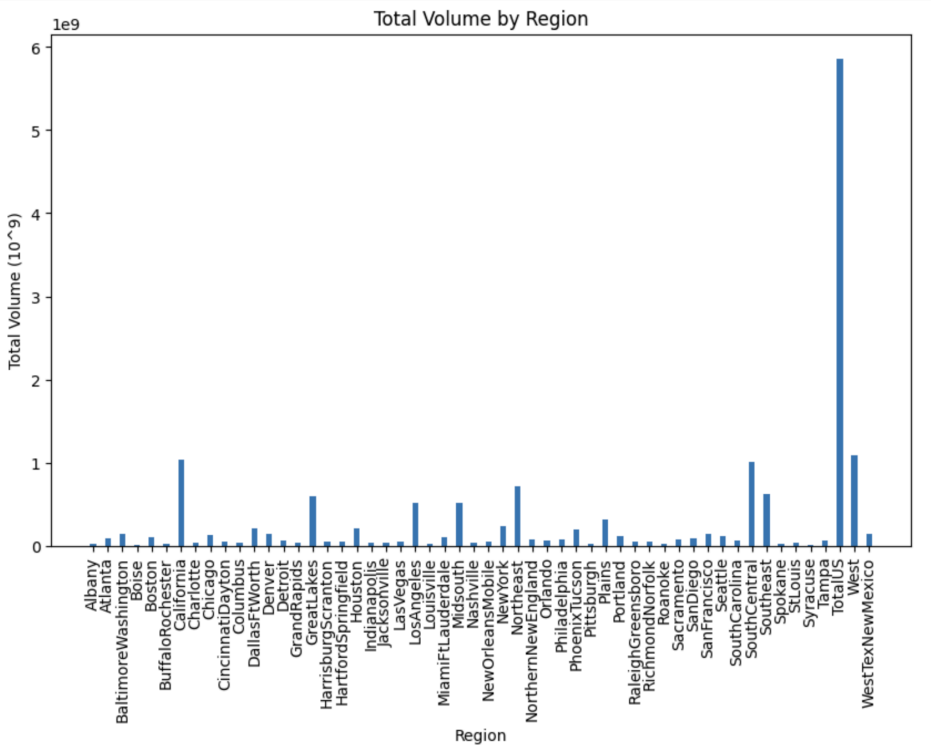

Figure 6: Bar graph of the total volume of avocados in each region of the data set.

Figure 6 shows the total volume of avocados in Albany and Atlanta and Figure 6 shows the total volume of avocados of all the regions in the data set. We wanted to know if certain regions had different demand levels of avocados, so showing each region's total volume sold next to each other helps to see the differences in demand. Places such as Los Angeles, San Francisco, and the West region have the highest total volume of avocados sold, while other locations such as Biose, Pittsburgh, and Syracuse have the lowest total volume sold. This could show a pattern that locations in the West of the US have higher demand for avocados than other regions in the US.

3. Are avocados sold in the summer months more expensive than avocados sold in winter months?

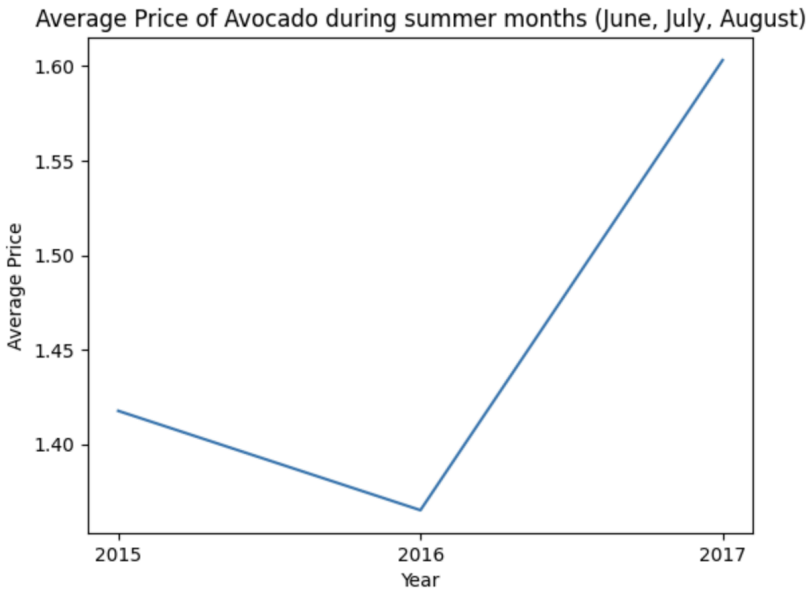

Figure 7: Line chart showing the average price of avocados in only the summer months.

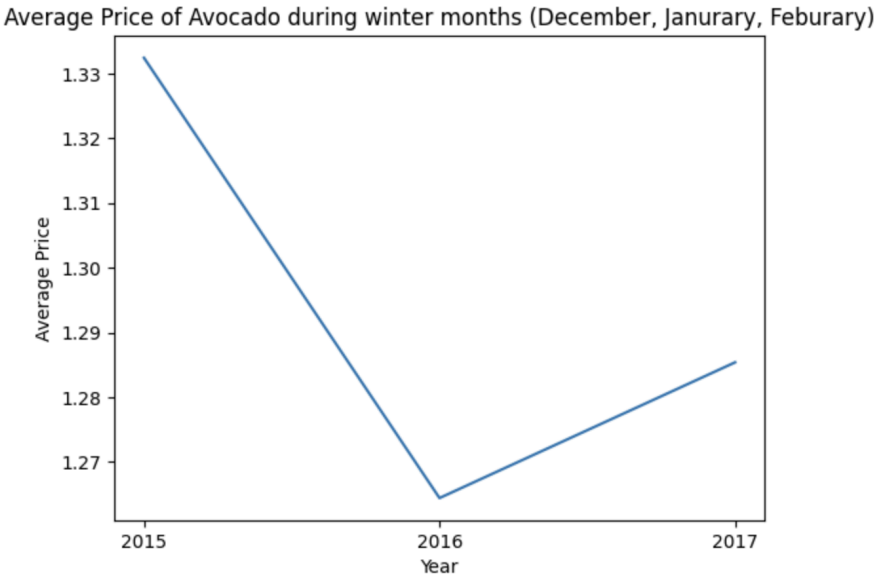

Figure 8: Line chart showing the average price of avocados in only the winter months.

Figure 7 and 8 show the average price of avocados in the summer and winter months respectively. In both charts, the year 2018 cannot be shown in the line charts because it is

missing data for the months of December. Even though the charts show different months, it is interesting to note that the average price of avocados drops in 2016, no matter what. Looking at the Y-axis, the average price of avocados, we see that the averages are higher (from around $1.40 - $1.60) in the summer months chart, and the prices are lower in the winter months chart (from around $1.27 - $1.33). This pattern shows that the average prices of avocados are higher in the summer months and are lower in the winter months. This supports our hypothesis that the mean price of avocados sold during the summer is higher than the mean price of avocados sold during the winter.

4. How does the PLU (price look-up code) impact the number of avocados sold?

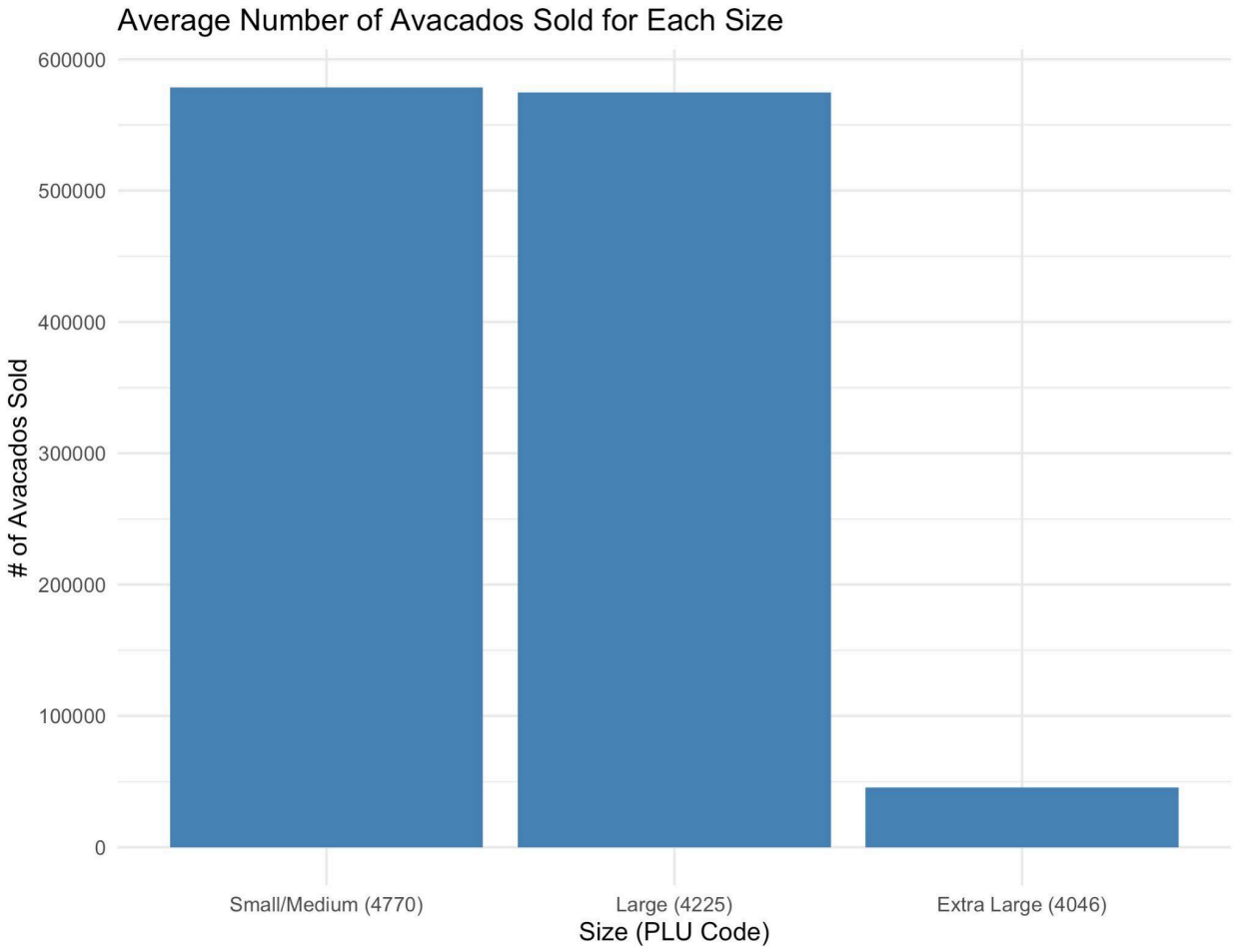

Figure 9: Bar graph showing the average of each PLU sold

This bar graph shows the average of each size of avocados sold.. As a reminder the PLU is the price look-up code, where 4046 corresponds to non-organic small/medium Hass avocados, 4225 corresponds to non-organic large Hass avocados, and 4770 corresponds to non-organic extra large Hass avocados. The bars show that the average number of small/medium avocados sold is just slightly larger than the average number of large avocados sold. The bars also show that the average number of extra large avocados sold is abundantly lower than the average number of small/medium or large avocados sold. This does support our hypothesis that the average number of extra large avocados sold is not equal to the average number of small/medium avocados sold. This shows that the sample preferred the small/medium and large avocados as opposed to the extra large avocados.

5. Hypothesis Testing

1. Are organic avocados more or less expensive than conventional avocados?

- Null: the mean price of conventional avocados is the same as the mean price of organic avocados
    - $H_0 : \mu_{\text{conventional price}} = \mu_{\text{organic price}}$
- Alternative: the mean price of conventional avocados is less than the mean price of organic avocados
    - $H_a : \mu_{\text{conventional price}} < \mu_{\text{organic price}}$

Two sample t test results:

- Alpha = .05
- Test statistic = 105.59
- Used sample mean as estimate of population mean
- Assumptions = sample size is large enough
- P-value < 0.0001

Analysis:

At the 5% significance level, there is sufficient evidence to reject the null hypothesis that the mean price of conventional avocados is the same as the mean price of organic avocados. Therefore, we can support the claim that the mean price of conventional avocados is less than the mean price of organic avocados.

Our first hypothesis to test was whether organic avocados were more expensive than conventional avocados. The null hypothesis was that there is no significant difference in the mean prices of these two types of avocados, while the alternative hypothesis states that organic avocados are more expensive. In order to complete the hypothesis testing for this null

hypothesis, the 2 Sample T-test will be used with an α value of 0.05. After perofming the test, the results show that the p-value for the test was extremely low, being less than 0.0001. Using the α value of 0.05, we can reject the null hypothesis since the p value was less than the α value of 0.05. Furthermore, there is sufficient evidence to support the claim that the mean price of conventional avocados is less than the mean price of organic avocados.

2. Do certain regions have different demand levels of avocado overall?

- Null: the proportion of total volume of avocados for a specific region equals the proportion of total volume of avocados for another specific region
    - Example: $H_0 : p_{atlanta} = p_{albany}$
- Alternative: the proportion of total volume of avocados for a specific region does not equal the proportion of total volume of avocados for another specific region
    - Example: $H_a : p_{atlanta} \neq p_{albany}$

Two proportion Z test results:

- Alpha = .05
- Test statistic = 1.04795
- Assumptions = each proportion sample is greater than 15
- p-value = 0.294659

Analysis:

At the 5% significance level, there is not enough evidence to support the claim that the proportion of total volume of avocados for Atlanta does not equal the proportion of total volume of avocados for Albany. Therefore, we fail to reject the null hypothesis that the proportion of total volume of avocados for Atlanta equals the proportion of total volume of avocados for Albany.

3. Are avocados sold in the summer months more expensive than avocados sold in winter months?

- Null: the mean price of avocados sold during the summer is the same as the mean price of avocados sold during winter
    - $H_0 : \mu_{\text{summer price}} = \mu_{\text{winter price}}$
- Alternative: the mean price of avocados sold during the summer is more expensive than the mean price of avocados sold during winter
    - $H_a : \mu_{\text{summer price}} > \mu_{\text{winter price}}$

T-Test result:

- Alpha = .05
- Test statistic = 2.2412
- Assumptions: Samples are large
- P-value = 0.1372

Analysis:

At the 5% significance level, there is not enough evidence to support the claim that the mean price of avocados sold during the summer is more expensive than the mean price of avocados sold during the winter. We fail to reject the null hypothesis that the mean prices are equal.

4. How does the PLU (price look-up code) impact the number of avocados sold?

- Null: the mean number of small/medium avocados sold is the same as the mean number of extra large avocados sold
    - $H_0 : \mu_{\text{number of small/medium avocados sold}} = \mu_{\text{number of extra large avocados sold}}$

- Alternative: the mean number of small/medium avocados sold is not equal to the mean number of extra large avocados sold
    - $H_a : \mu_{\text{number of small/medium avocados sold}} \neq \mu_{\text{number of extra large avocados sold}}$

Two sample T-test result:

- Alpha = .05
- Test statistic = 29.1266
- Assumptions: samples sizes are sufficiently large, using the sample mean as the test statistic
- P-value < 0.0001

Analysis:

At the 5% significance level, there is sufficient evidence to reject the null hypothesis that the mean number of small/medium avocados sold is equal to the mean number of extra large avocados sold. Therefore, we can support the claim that the mean number of small/medium avocados sold is not equal to the mean number of extra large avocados sold.