Margaret Perez

February 27th, 2024

Comparing what year(s) sold the most Organic vs Conventional Avocados

Photo link https://www.flickr.com/photos/160866001@N07/48880402911/in/photostream/

Business problem solving for:

By comparing these metrics, you can gain insights into the performance and trends of organic and conventional avocado sales over time, which can help in making strategic decisions, understanding market dynamics, and identifying growth opportunities.

The objective of this analysis would be to tell a visual story of what year(s) sold the highest sales of Organic vs Conventional avocados.

Inspiration

The inspirational part of this analyst is to find out what year(s) the most Organic and Conventional avocados were sold in to determine what caused the significant increase in 2017 vs 2016, what is the driving sales of Conventional vs Organic? This is something our Sales team

can further determine, what is causing this and where we can improve on, was it the price of conventional or the demand of that was the factor in these sales.

The challenge that we face is to further analyze, why are organic much higher than conventional avocados, why do they need to be? Is it the care that raising organic costs, is it where the avocados come from or originate from?

We all know the cost of avocados fluctuates, as well as the demand for avocados, from size, location, price, and type of avocado, Organic and Conventional. The price of Organic avocados is a little higher in price, therefore the demand for these types of avocados will depend on the cities, counties, states, and even regions. In this case study, we are focusing on the states and which state sold the most Organic and Conventional avocados from 2015 through partial 2018.

From making the best guacamole for dips to adding this flavoring fruit with a dish, avocados are shared on the biggest days such as Superbowl and Cinco de Mayo and delicious by themselves and providing a healthy fat into our bodies. our bodies.

About the Dataset

Content

Summary of my analysis from the "Price of Avocados", dataset downloaded from the following:

https://www.kaggle.com/datasets/neuromusic/avocado-prices

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV.

http://www.hassavocadoboard.com/retail/volume-and-price-data

The table below 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 below 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.

Conclusion is what we figured out, is in 2016 we sold significantly less Organic Avocados, we will need to dive deeper with the Sales/Marketing teams, what can we do different or figure out how to close the gap between the sales of both Organic and Conventional types of avocados.

I have listed here the urls to google docs, screenshot and copies of the data that was used to determine my findings. I learned a lot through this case study, not only is it important to focus on the task at hand but to thoroughly understand the objective and to complete the studies as quickly as possible and not let too much time pass.

Below are the columns that were used from the dataset to draw and conclude the analyst.

Date - The date of the observation

Average Price - the average price of a single avocado

type - conventional or organic

year - the year

State - all cities for the states combined into one number

Region - the city or region of the observation

Total Volume - Total number of avocados sold, both Conventional and Organic

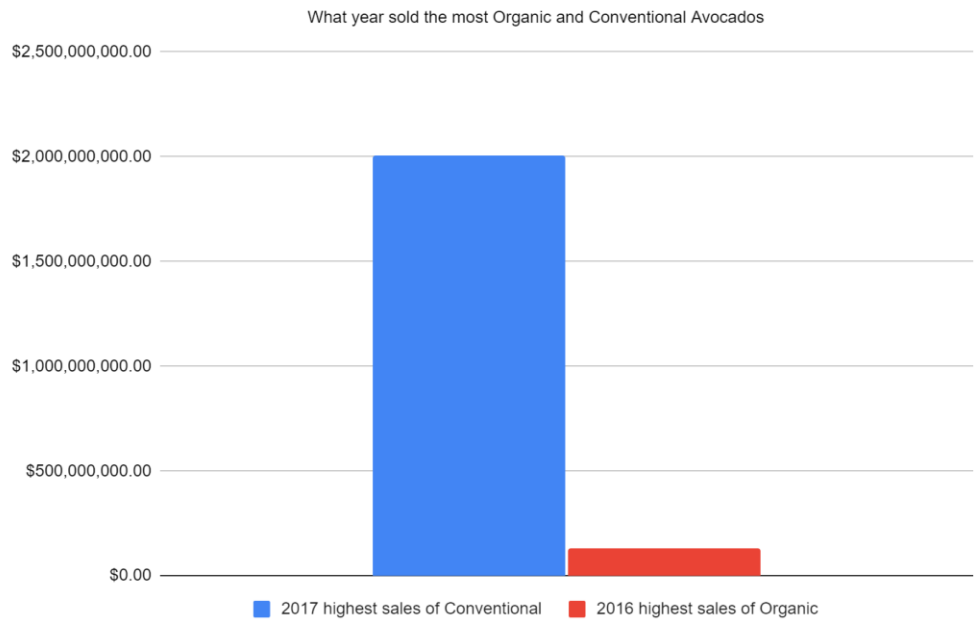
4046 - Total number of avocados with PLU 4046 sold

4225 - Total number of avocados with PLU 4225 sold

4770 - Total number of avocados with PLU 4770 sold

Downloaded the dataset into a csv file, loaded it into google sheets and created a google doc below.

https://docs.google.com/spreadsheets/d/1YSaD5P6QXSxbmixgBy9wjACiZY6BuJRcSBHI5wXB-2U/edit#gid=515318811



Steps used to clean the data:

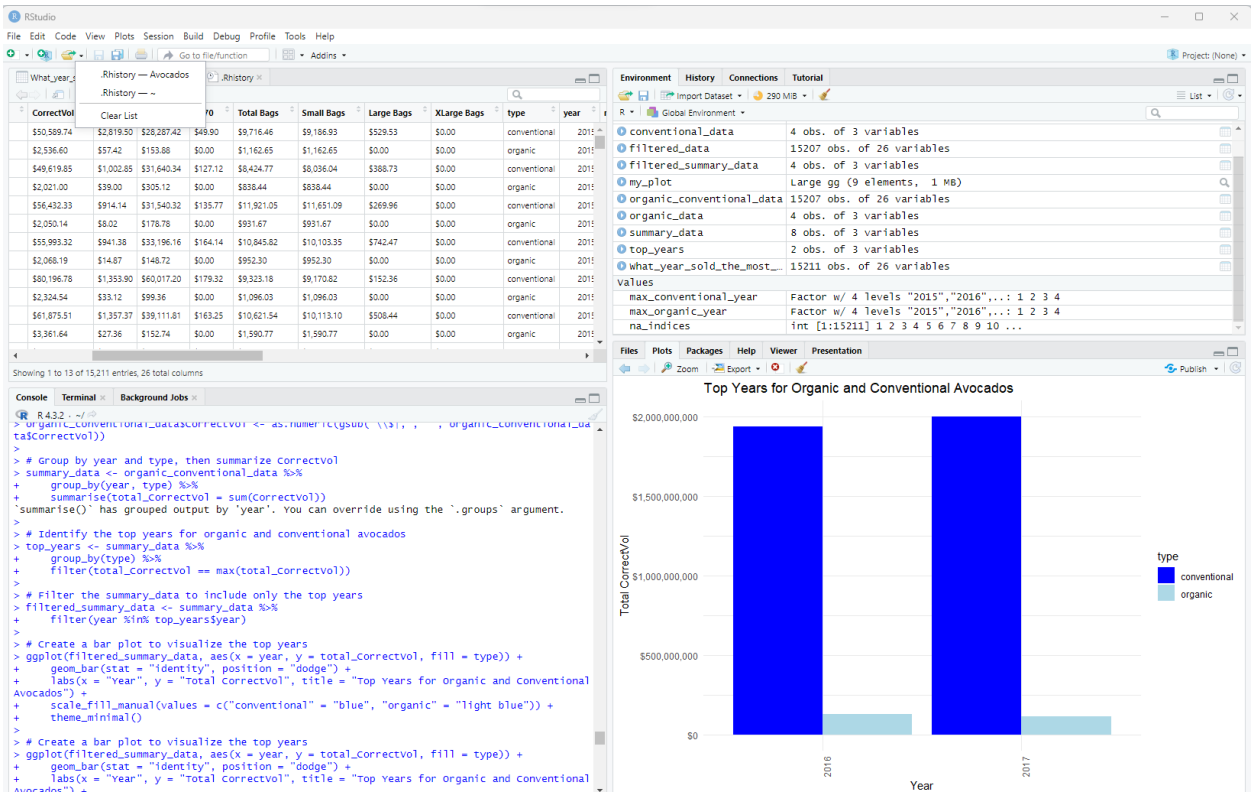
1. Validated the Total Volume was correct, determined this was not the case, steps to correct the total.
    a. Added a column and labeled it “CorrectVol”
    b. Created a =SUM(F2:K2) filtered the equation to the rest of the rows, this gave me the correct total for each row
    c. Added column and labeled it “year” I only want to sort by year
    d. Tallied all regions to the correct states not sure which direction I was going to go with the data

e. Created filter(s) =SUMIFS(E:E, B:B, ">="&DATE(2015,1,1), B:B, "<="&DATE(2015,12,31), M:M, "conventional") for each year for both Conventional and Organic

2. Once data has been cleaned I created two more queries one for Organic and one for Conventional =SUMIFS(E:E, B:B, ">="&DATE(2017,1,1), B:B, "<="&DATE(2017,12,31), M:M, "conventional") and =SUMIFS(E:E, B:B, ">="&DATE(2016,1,1), B:B, "<="&DATE(2016,12,31), M:M, "organic")

I then moved onto R studio to tell the same story

I downloaded the cleaned googlesheet into R studio and created a graph




```
> # Identify the top years for organic and conventional avocados
> top_years <- summary_data %>%
+     group_by(type) %>%
+     filter(total_CorrectVol == max(total_CorrectVol))
>
> # Filter the summary_data to include only the top years
> filtered_summary_data <- summary_data %>%
+     filter(year %in% top_years$year)
>
> # Create a bar plot to visualize the top years
> ggplot(filtered_summary_data, aes(x = year, y = total_CorrectVol, fill = type)) +
```

```
+    geom_bar(stat = "identity", position = "dodge") +
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal()
>
> # Create a bar plot to visualize the top years
> ggplot(filtered_summary_data, aes(x = year, y = total_CorrectVol, fill = type))
+
+    geom_bar(stat = "identity", position = "dodge") +
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal() +
+    theme(axis.text.x = element_text(angle = 90, hjust = 1))
>
> library(dplyr)
> library(ggplot2)
>
> # Filter the data for organic and conventional avocados
> organic_conventional_data <-
What_year_sold_the_most_Organic_and_Conventional_avocado %>%
+    filter(type %in% c("organic", "conventional"))
>
> # Convert CorrectVol to numeric
> organic_conventional_data$CorrectVol <- as.numeric(gsub("\\$|,", "",
organic_conventional_data$CorrectVol))
>
> # Group by year and type, then summarize CorrectVol
> summary_data <- organic_conventional_data %>%
+    group_by(year, type) %>%
+    summarise(total_CorrectVol = sum(CorrectVol))
`summarise()` has grouped output by 'year'. You can override using the `.groups`
argument.
>
> # Identify the top years for organic and conventional avocados
> top_years <- summary_data %>%
+    group_by(type) %>%
+    filter(total_CorrectVol == max(total_CorrectVol))
>
> # Filter the summary_data to include only the top years
> filtered_summary_data <- summary_data %>%
+    filter(year %in% top_years$year)
>
> # Create a bar plot to visualize the top years
> ggplot(filtered_summary_data, aes(x = year, y = total_CorrectVol, fill = type))
+
+    geom_bar(stat = "identity", position = "dodge") +
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal()
>
> library(dplyr)
> library(ggplot2)
```

```
>
> # Filter the data for organic and conventional avocados
> organic_conventional_data <-
What_year_sold_the_most_Organic_and_Conventional_avocado %>%
+    filter(type %in% c("organic", "conventional"))
>
> # Convert CorrectVol to numeric
> organic_conventional_data$CorrectVol <- as.numeric(gsub("\\$|,", "",
organic_conventional_data$CorrectVol))
>
> # Group by year and type, then summarize CorrectVol
> summary_data <- organic_conventional_data %>%
+    group_by(year, type) %>%
+    summarise(total_CorrectVol = sum(CorrectVol))
`summarise()` has grouped output by 'year'. You can override using the `.groups`
argument.
>
> # Identify the top years for organic and conventional avocados
> top_years <- summary_data %>%
+    group_by(type) %>%
+    filter(total_CorrectVol == max(total_CorrectVol))
>
> # Create a bar plot to visualize the top years
> ggplot(top_years, aes(x = year, y = total_CorrectVol, fill = type)) +
+    geom_bar(stat = "identity", position = "dodge") +
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal()
>
> library(dplyr)
> library(ggplot2)
>
> # Filter the data for organic and conventional avocados
> organic_conventional_data <-
What_year_sold_the_most_Organic_and_Conventional_avocado %>%
+    filter(type %in% c("organic", "conventional"))
>
> # Convert CorrectVol to numeric
> organic_conventional_data$CorrectVol <- as.numeric(gsub("\\$|,", "",
organic_conventional_data$CorrectVol))
>
> # Group by year and type, then summarize CorrectVol
> summary_data <- organic_conventional_data %>%
+    group_by(year, type) %>%
+    summarise(total_CorrectVol = sum(CorrectVol))
`summarise()` has grouped output by 'year'. You can override using the `.groups`
argument.
>
> # Identify the top years for organic and conventional avocados
> top_years <- summary_data %>%
+    group_by(type) %>%
+    filter(total_CorrectVol == max(total_CorrectVol))
>
> # Create a bar plot to visualize the top years
> ggplot(top_years, aes(x = year, y = total_CorrectVol, fill = type)) +
+    geom_bar(stat = "identity", position = "dodge") +
```

```
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal()
>
> # Convert 'year' to factor
> filtered_summary_data$year <- as.factor(filtered_summary_data$year)
>
> # Create a bar plot to visualize the top years
> ggplot(filtered_summary_data, aes(x = year, y = total_CorrectVol, fill = type))
+
+    geom_bar(stat = "identity", position = "dodge") +
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal() +
+    theme(axis.text.x = element_text(angle = 90, hjust = 1))
>
> library(scales)
>
> # Create a bar plot to visualize the top years
> ggplot(filtered_summary_data, aes(x = year, y = total_CorrectVol, fill = type))
+
+    geom_bar(stat = "identity", position = "dodge") +
+    labs(x = "Year", y = "Total CorrectVol", title = "Top Years for Organic and
Conventional Avocados") +
+    scale_fill_manual(values = c("conventional" = "blue", "organic" = "light
blue")) +
+    theme_minimal() +
+    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
+    scale_y_continuous(labels = dollar_format(prefix = "$"))
>
```

Biq Query Stuido

https://console.cloud.google.com/bigquery?project=sonorous-sign-406521&ws=!1m5!1m4!4m3!1ssonorous-sign-406521!2sAvocados!3sAvocado_clean

```sql
SELECT

    CASE

        WHEN type = 'conventional' AND year = 2017 THEN SUM(CorrectVol)

        ELSE 0

    END AS total_sales_conventional_2017,

    CASE

        WHEN type = 'organic' AND year = 2016 THEN SUM(CorrectVol)

        ELSE 0

    END AS total_sales_organic_2016
```

```
FROM
    `sonorous-sign-406521.Avocados.Avocado_clean`
WHERE
    (type = 'conventional' AND year = 2017)
    OR (type = 'organic' AND year = 2016)
GROUP BY
    type, year;
```