FIT5147 Project Proposal and Data Exploration Project

Name: Jiaming Ren, Student ID: 217218863, Tutor: Rita Hoang, Vaibhavi Bhardwaj, Tutorial 1

Table of Contents

1. Introduction

1.1 Problem Description

Find meaningful insights about how avocado sales is determined in various factors.

Notes: The Hass Avocado Board divide the U.S. into **eight** geographical regions (California, West, Plains, South Central, Great Lakes, Northeast, Midsouth, Southeast). States for each geographical region:

- California: California
- West: Washington, Oregon, Nevada, Utah, Arizona, Idaho, Montana, Wyoming, Colorado, New Mexico
- Plains: North Dakota, South Dakota, Minnesota, Iowa, Missouri, Kansas, Nebraska.
- Great Lakes: Wisconsin, Michigan, Illinois, Indiana, Ohio
- Northeast: Maine, New York, Pennsylvania, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New Jersey.
- Midsouth: Kentucky, Tennessee, North Carolina, Virginia, Delaware, Maryland, West Virginia, District of Columbia.
- Southeast: Mississippi, Alabama, Georgie, Florida, South Carolina, North Carolina
- Southcentral: Texas, Oklahoma, Arkansas, Louisiana

1.2 Question

I. Is there any noticeable change of price in different area? Is there a geographic factor?
II. What area consume more avocados? Any relationship between consumption and average price (single avocado)?
III. Do projections affect average price?
IV. Do GDP affect average price?
V. What is the seasonal pattern for consumptions and average price?

1.3 Motivation

Avocado is healthy and delicious; I love eating avocado, and I know that American loves Avocados. About 400million pounds of avocados are harvested each year in California alone, which are lots of avocados and needs to sell somewhere. I am curious how these avocados are distributed. Therefore, I want to find out more about the factors that drove to sell more avocados.

2. Data Wrangling

2.1 Data Sources

A. Hass Avocado Board data for the United State from 2015 – 2020 (https://www.kaggle.com/timmate/avocado-prices-2020)
B. Bureau of Economic Analysis U.S. DEPARTMENT OF COMMERCE for the America GDP per state (https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1&acrdn=1)
C. United States Census Bureau State Population Totals (2010-2020) (https://www.census.gov/data/datasets/time-series/demo/popest/2010s-statetotal.html#par_textimage_500989927)
D. United States Census Bureau State Population by Characteristics (2010-2020) (https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluationestimates/2020-evaluation-estimates/2010s-state-detail.html)
E. NOAA national centres for environmental information state-wide average temperature per month for each state (1895-2021). (https://www.ncdc.noaa.gov/cag/statewide/mapping)
F. Hass Avocado Board data for Weekly Volume Data & Projections (2019-2021) (https://hassavocadoboard.com/volume-data-projections/?weight=kgs)

G. Us-state-capitals-longitude-latitude (https://github.com/jasperdebie/VisInfo/blob/master/us-state-capitals.csv)
H. us-cities-top-1k (https://github.com/plotly/datasets/blob/master/us-cities-top-1k.csv)

2.2 Tools for Wrangling

- R studio Version 1.4.1103
- Microsoft Excel
- Python Version 3.8.5

2.3 Wrangling Steps

- **Data Source A** – 1. Skip first four rows (heading). 2. Remove all rows with null values. 3. Combine left join 6 files by States name. 4. Wide table to long table. 5. Rename date column. 6. Clean date variable meaningless symbol. 7. Reformat date variables to date type. 8. Set factors
- **Data Source B** – 1. Remove all unnecessary columns. 2. Group data. 3. Rename factors name. 4. Set factors
- **Data Source C** – 1. Remove meaningless columns. 2. Row bind three data frames. 3. Reformat date format. 4. Set factors
- **Data Source D** – 1. Remove meaning less columns. 2. Skip first three rows (heading). 3. Reformat date format. 4. Set factors
- **Data Source E** – 1. Set factors
- **Data Source F** – 1. Skip first three rows (heading). 2. Rename column. 3. Select columns and remove all rows with null value. 4. Clean variable meaningless symbols. 5. Set factors
- **Data Source G** – 1. Remove meaningless characters. 2. Set factors
- **Data Source H** – 1. Set factors

3. Data Checking

3.1 Steps

- **Determine Data sample** – I found not all columns are useful for later exploration. Thus, I removed those columns (e.g., ID columns that cannot help when join tables).
- **Validate the Database** – check the number of columns and records
- **Validate Data format** – Change data to appropriate data type (e.g., reformat data type)

3.2 Errors Found

a) Invalid date types and format
b) Meaningless columns.
c) Inappropriate column names.
d) Heading and null values.
e) Inappropriate factor names
f) Meaningless symbols

3.3 Method and Justification

a) All original date types are char. Thus, I change date types and to appropriate data formats
b) Not all columns are useful for later explorations (e.g., ID). Thus, I remove all meaningless columns
c) Some columns' names are hard to understand. Thus, I change to appropriate column names.
d) Headings don't provide meaningful data, and some rows has too many null values. Thus, I skip heading and remove columns with too many null values
e) Some factors are numbers, and it is hard to understand. Thus, I change to appropriate names to better understanding.
f) Some symbols are meaningless. Thus, I remove all meaningless symbols

3.4 Tools for Checking

- R studio Version 1.4.1103
- Microsoft Excel

4. Data Exploration

4.1 Geographic Factor

Problem Descriptions:

Is there any noticeable change of price in different area? Is there a geographic factor?

Notes: there are two types of avocados (organic and conventional)

Data Source used:

- Data source A

Visualizations Process:

1.filter data, find all geographical regions

2.average price density between geographical regions (in 5 years)

3.Timelines of how average price changed over 5 years for all geographical regions

Statistical Test:

1.Statistical Significance of total U.S. avocado average price by type:

```
data:  average_price by type
t = -34.297, df = 601.17, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4960150 -0.4422856
sample estimates:
mean in group conventional      mean in group organic
                  1.089216                   1.558366
```

Figure 1Statistical Significance of total U.S. avocado average price by type

Result: avocado average price by type has **significant differences** in the mean values.

2.Statistical Significance of total U.S. avocado average price by regions and types (California and Northeast) :

```
data:  average_price by City
t = -10.2, df = 573.33, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1909655 -0.1292959
sample estimates:
mean in group California  mean in group Northeast
                1.152386                 1.312516
```

Figure 2 California and Northeast significance differences (conventional)

```
data:  average_price by City
t = -1.419, df = 563.26, p-value = 0.1565
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06147566  0.00990703
sample estimates:
mean in group California  mean in group Northeast
                1.739444                 1.765229
```

Figure 3 California and Northeast significance differences (organic)

Result: avocado average price by regions and types **has significant differences** in conventional type the mean values. However, the organic type **does not have significant differences**.

Discover:

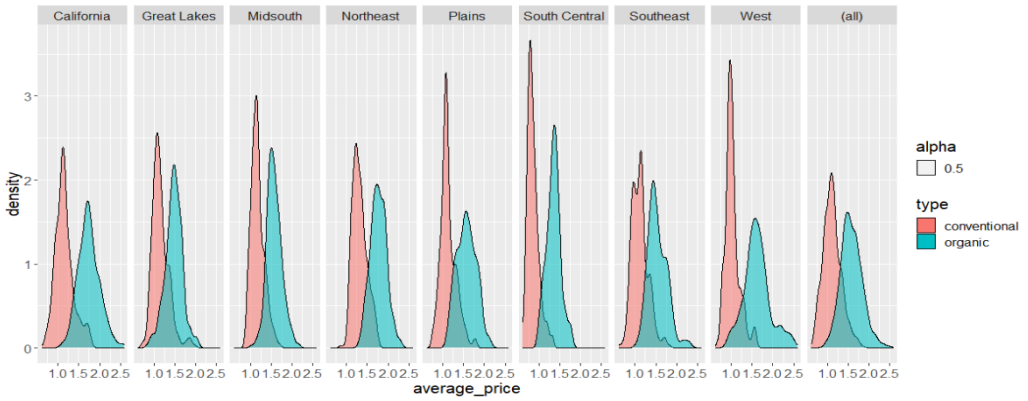

Figure 4 Geographical Regions average price for type density graph in 5 years

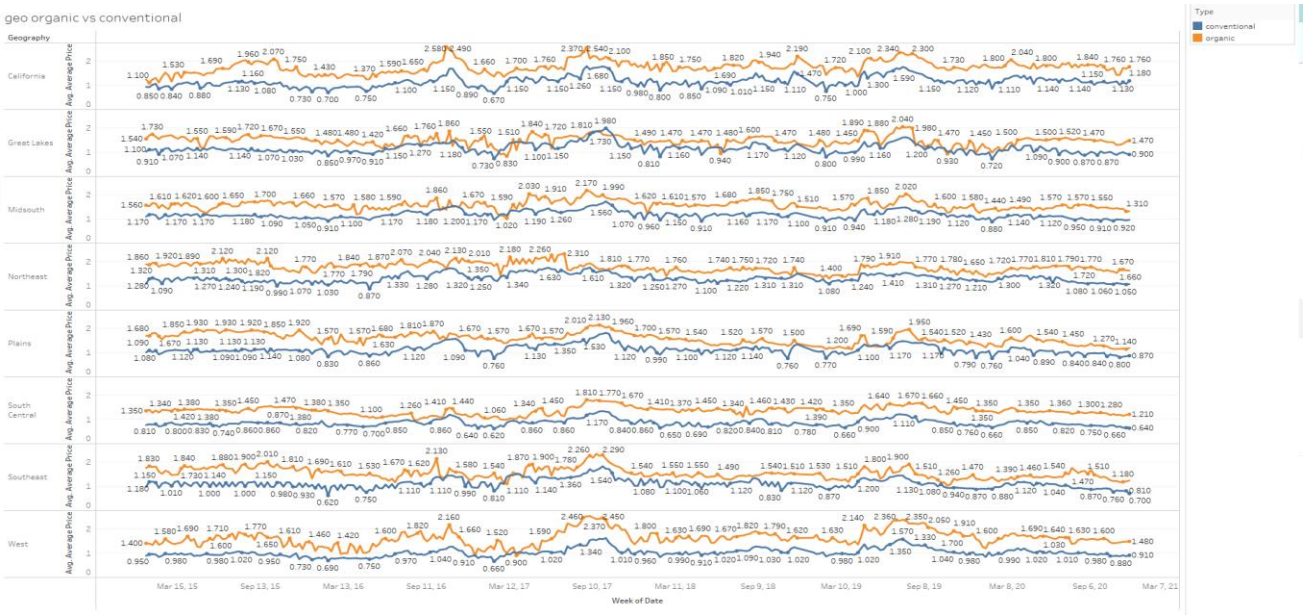

Figure 5 Geographical Regions average price for type timeline graph in 5 years (weekly)

From Figures 4 and 5, I can confirm that for all geographical regions of all time, that organic avocado has a higher average price than conventional avocado.

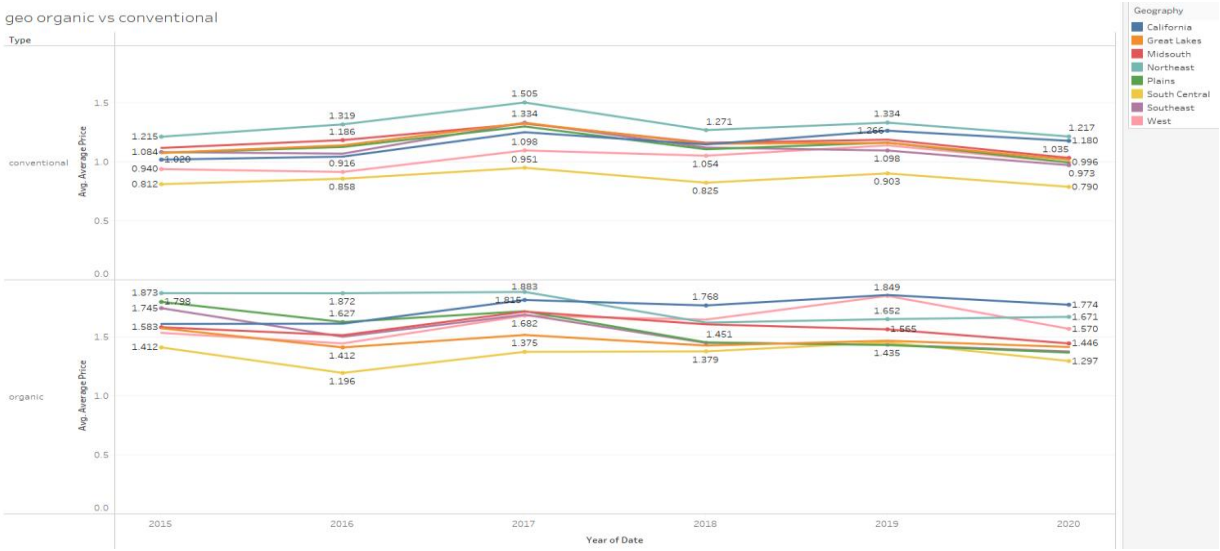

Figure 6 Geographical Regions average price for type timeline graph in 5 years (yearly)

From figure 6, if we dive deeper, we can observe that northeast region has highest average price all time for conventional type. California has the highest average price for organic avocado since mid of 2017, which is interesting because most of the avocados produced in the United States are grown in California (Statista 2021). Southcentral region has the lowest average price in both organic and conventional type in all time.

4.2 Consumption and Average price

Problem Descriptions:

What area consume more avocados? Any relationship between consumption and average price?

Data Source used:

- Data source A

Visualizations Process:

1. Timelines of how average price and consumptions changed over 5 years in total U.S.

Statistical Test:

1.Spearman's correlation coefficients test of the relationship between Total U.S. total volume and average price in five years.

```
        Spearman's rank correlation rho

data:  p$total_volume and p$average_price
S = 66692404, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.7457197
```

Figure 7 Spearman's correlation coefficients test of the relationship between Total U.S. total volume and average price in five years

Result: The average price tends to decrease when the total volume increase.

Discover:

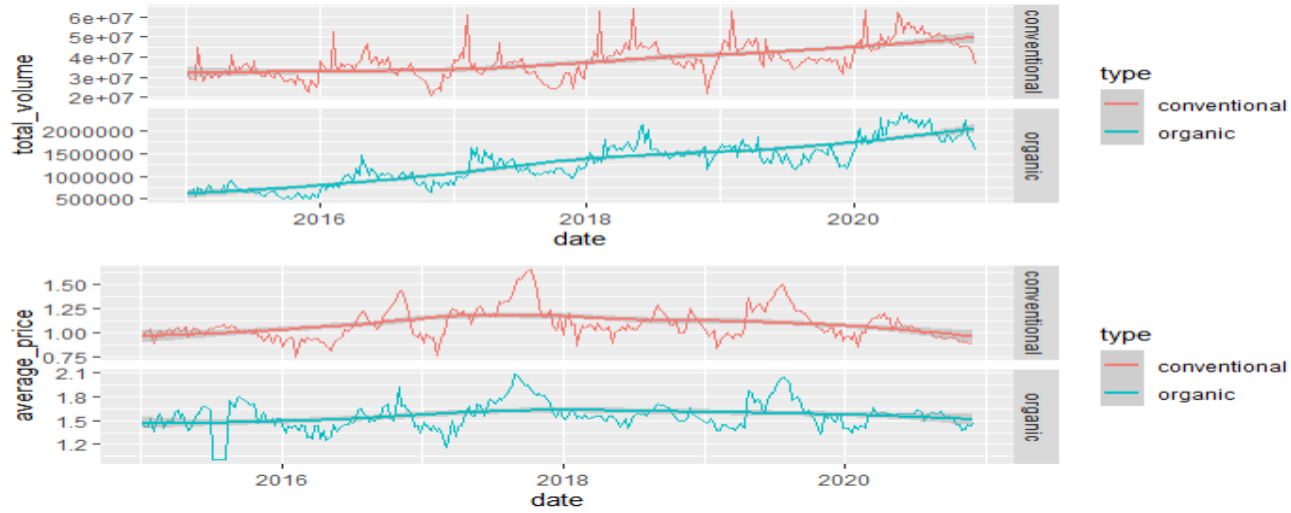

Figure 8 Total U.S. avocado consumption and average price trend

From figure 8, I can observe that overall, the total volume of consumption is increasing for both types. However, the average price is slightly dropping in recent years. Also, people consume much more conventional type than organic.

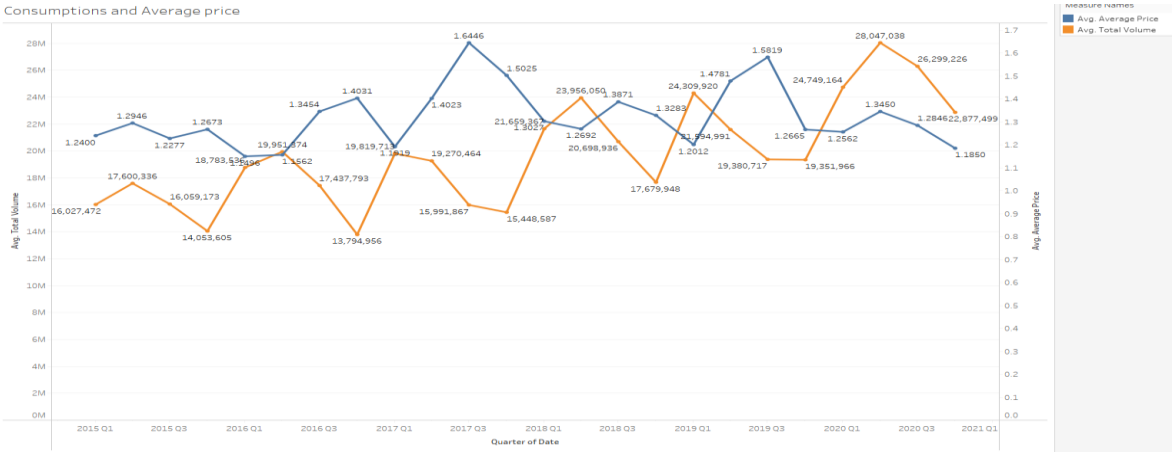

Figure 9 Total U.S. avocado consumption and average price trend (quarterly)

From figure 9, we can observe a pattern that when the total volume is high, the average is low, and when the total volume is low, the average price is high. Another seasonal pattern I can find in this graph is that the lowest total volume for each year is at quarter 4.

4.3 Projections and average price

Problem Descriptions:

Do projections affect average price?

Data Source used:

- Data source A
- Data source F

Visualizations Process:

2. Timelines of how average price and projections changed in 2019 and 2020

Statistical Test:

1. Spearman's correlation coefficients test of the relationship between Total U.S. total projection volume and average price in two years.

```
        Spearman's rank correlation rho

data:  avocado_volume2019_2020_us$Total_Volume and
avocado_volume2019_2020_us$average_price
S = 2514, p-value = 0.2645
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2420949
```

Figure 10 Spearman's correlation coefficients test (projection vs average price)

Result: there is a **weak** relationship between total projection volume and average price

Discover:

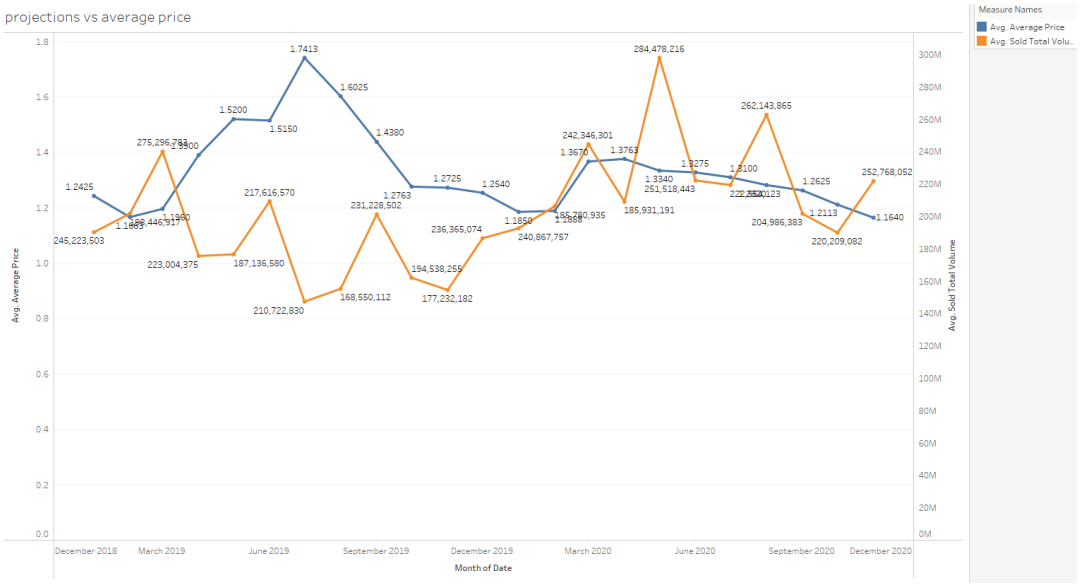

Figure 11 projections vs average price

From figure 11, we can find a weak correlation that when the projections are high, the average price tend to lower.

4.4 Season Pattern for consumption and average price

Problem Descriptions:

What is the seasonal pattern for consumption and average price?

Note: America is a big country, and an exploration is a huge amount of work. Thus, I pick the California and New York states as object of observations. There are great geographical differences between these two states.

Data Source used:

- Data source A
- Data source E

Visualizations Process:

1. Timeline of average temperature for each month in 2015 - 2020
2. Scatter plots of average price vs temperature and consumption vs temperature in California in 2015 – 2020
3. Scatter plots of average price vs temperature and consumption vs temperature in New York in 2015 – 2020
4. Timeline of compare average price for New York and California in 2015 - 2020

Statistical Test:

1.Spearman's correlation coefficients test of the relationship between **New York** consumption and temperature in 2015 -2020 years.

```
        Spearman's rank correlation rho

data:  r_NewYork$total_volume and r_NewYork$temperature
S = 55847, p-value = 0.5982
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06359986
```

Figure 12 Spearman's correlation coefficients test of the relationship between New York consumption and temperature in 2015 -2020 years.

Results: from the result, I can say there is **no** relationship between consumption and temperature

2. Spearman's correlation coefficients test of the relationship between **New York** average price and temperature in 2015 -2020 years.

```
data:  r_NewYork$average_price and r_NewYork$temperature
S = 38547, p-value = 0.002481
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3536671
```

Figure 13 Spearman's correlation coefficients test of the relationship between New York average price and temperature in 2015 -2020 years.

Results: from the result, I can say there is **weak** relationship between average price and temperature

3. Spearman's correlation coefficients test of the relationship between California consumption and temperature in 2015 -2020 years.

```
        Spearman's rank correlation rho

data:  r_California$total_volume and r_California$temperature
S = 60691, p-value = 0.884
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01762314
```

Figure 14 Spearman's correlation coefficients test of the relationship between California consumption and temperature in 2015 -2020 years.

Result: from the result I can say there is **no** relationship between consumption and temperature

4. Spearman's correlation coefficients test of the relationship between California average price and temperature in 2015 -2020 years.

```
        Spearman's rank correlation rho

data:  r_California$average_price and r_California$temperature
S = 32424, p-value = 6.343e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4563351
```

Figure 15 Spearman's correlation coefficients test of the relationship between California average price and temperature in 2015 -2020 years.

Result: from the result I can say there is a **moderate** relationship between average price and temperature

Discover:

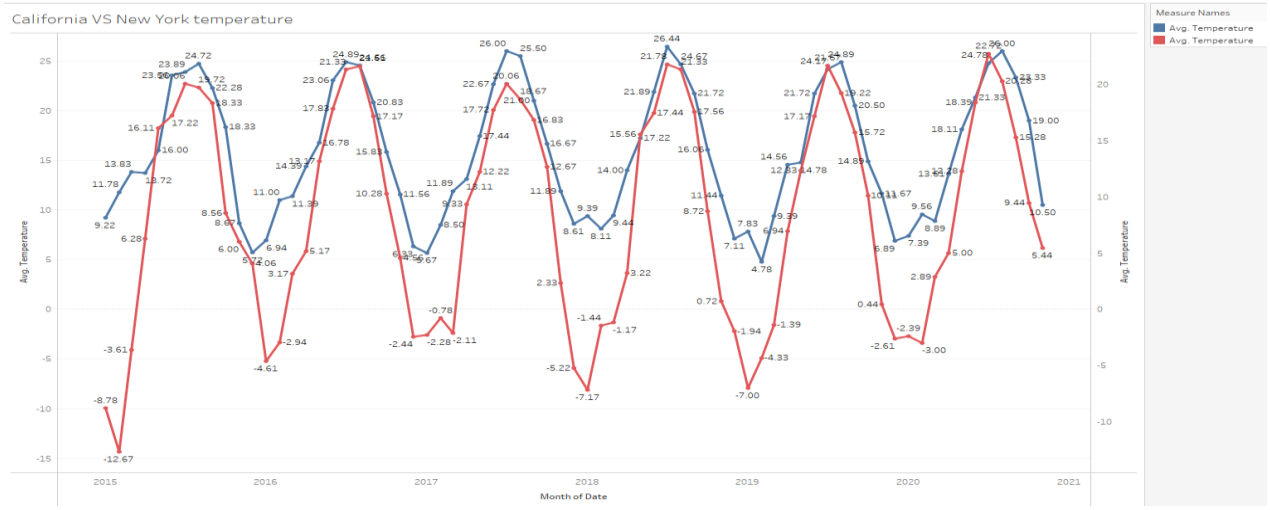

Figure 16 Timeline of average temperature for each month in 2015 - 2020

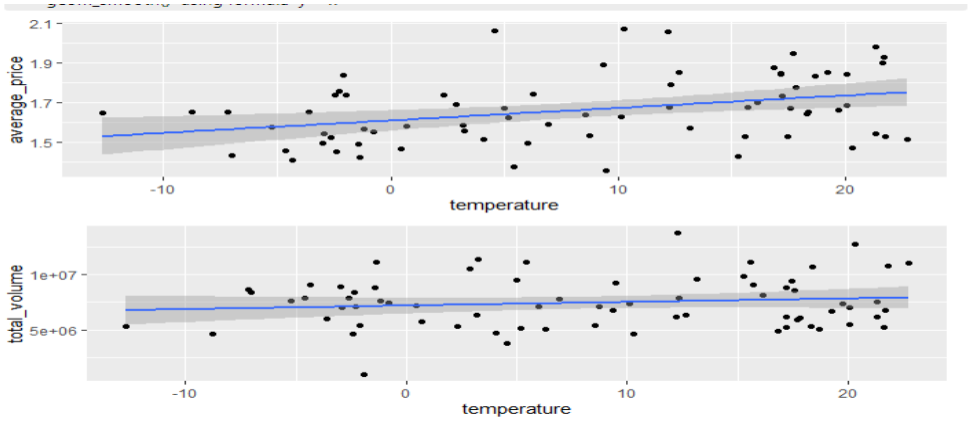

Figure 17 scatter plot of average price and temperature in New York in 2015 – 2020

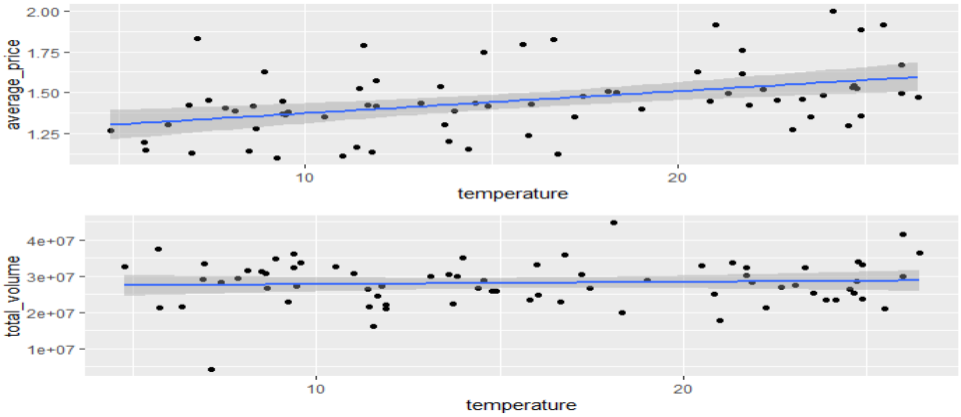

Figure 18scatter plot of average price and temperature in California in 2015 – 2020

From Figures 17 and 18, I can observe there is a moderate relationship between average price and temperature. When the temperature is high, the average price is high, and when the temperature is lower the price will go down. Also, there is no obvious relationship between total consumption and temperature in both states.

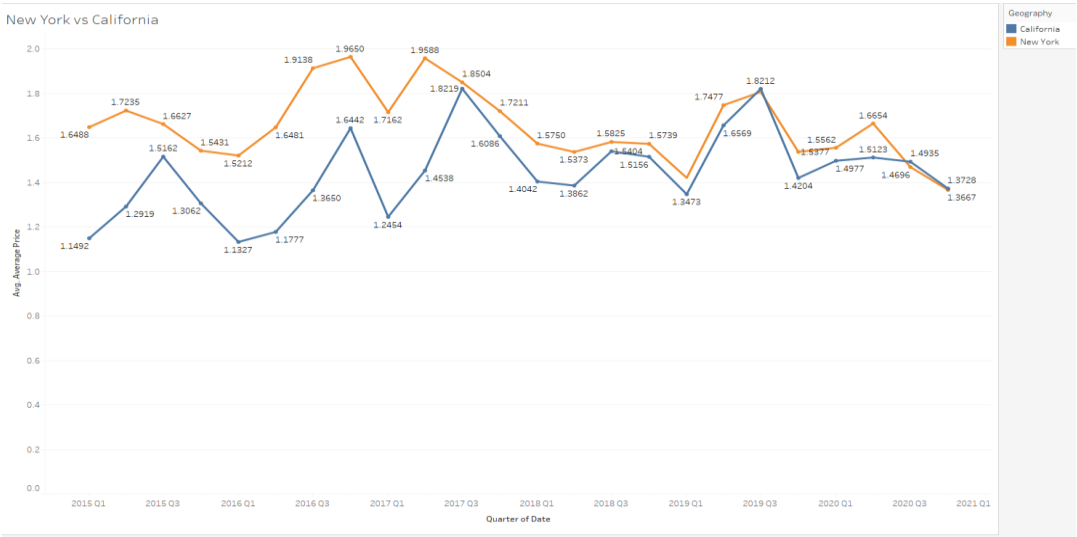

Figure 19 California and New York average price comparison

From figures 16 and 19, I can observe that temperature and average price in New York and California states share a similar pattern, even though there are great geographical differences between these two states.

4.5 Tools for Visualization

- R studio Version 1.4.1103
- Tableau

5. Conclusion

Geographic Factor:

I choose the California and Northeast geographical regions as observations. Overall, I find that avocado average price by regions and types has significant differences in conventional type the mean values. However, the organic type does not have significant differences. Furthermore, the Northeast region has the highest average price all time for conventional type. California has the highest average price for organic type since 2017

Relationship between consumption and Average price:

The average price tends to decrease when the total volume increase. I also find that people consume much more conventional type than organic.

Relationship between projections and average price:

I find there is a weak relationship between total projection volume and average price, when projection volume is higher, the average price will be lower.

Season Pattern for consumption and average price:

I found there is no obvious relationship between consumption and temperature in both states (California and New York). However, there is a weak relationship between average price and temperature in New York and a moderate relationship in California, that when the temperature is higher the price tends to be higher, and when the temperature is lower the price tends to be lower.

6. Reflection

In this exploration, I have enhanced my data checking, wrangling, and explorations skills. Next time I wouldn't try to find too much data that I won't use at all.

7. Bibliography

1. Statista (2021). Production of avocados in the United States in 2020, by state (in 1,000 tons)*. Retrieved from (https://www.statista.com/statistics/610460/production-avocados-us-by-state/)
2. Timeanddate (2021). By Konstantin Bikos and Aparna Kher. Seasons: Meteorological and Astronomical. Retrieved from (https://www.timeanddate.com/calendar/aboutseasons.html)