Rise of the Avacados

A previous DATA 2401 student

Spring 2020

Introduction

Avocados! Some people love them and others can't stand them. I am indifferent about them, but they are great in a Pico de Gallo. In the most recent years people have been more and more concerned about what they put into their bodies. How has this changed the sales of avocados?
The data that I will be using was collected from the Hass Avocado Board. The HAB records the sales of avocados in the United States, but only for Hass products. The data will be enough to give us a good understanding of avocado sales in the U.S. I will be finding different patterns in sales and price fluctuations in different regions. Analyzing these findings will help us see the demand of avocados and I will try to find the role that price has with demand.

```
# Load tidyverse to manipulate data
# Load lubridate to manipulate dates.
# Load ggplot2 for graphing,
library(tidyverse, warn.conflicts = F)
library(lubridate, warn.conflicts = F)
library(ggplot2, warn.conflicts = F)
```

First I will read in the data and save it under a new variable.

```
# Read in avacado data
my_avocado <- read.csv("avocado.csv", stringsAsFactors = FALSE)
glimpse(my_avocado)
```

```
## Rows: 18,249
## Columns: 14
## $ X            <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ Date         <chr> "2015-12-27", "2015-12-20", "2015-12-13", "2015-12-06"...
## $ AveragePrice <dbl> 1.33, 1.35, 0.93, 1.08, 1.28, 1.26, 0.99, 0.98, 1.02, ...
## $ Total.Volume <dbl> 64236.62, 54876.98, 118220.22, 78992.15, 51039.60, 559...
## $ X4046        <dbl> 1036.74, 674.28, 794.70, 1132.00, 941.48, 1184.27, 136...
## $ X4225        <dbl> 54454.85, 44638.81, 109149.67, 71976.41, 43838.39, 480...
## $ X4770        <dbl> 48.16, 58.33, 130.50, 72.58, 75.78, 43.61, 93.26, 80.0...
## $ Total.Bags   <dbl> 8696.87, 9505.56, 8145.35, 5811.16, 6183.95, 6683.91, ...
## $ Small.Bags   <dbl> 8603.62, 9408.07, 8042.21, 5677.40, 5986.26, 6556.47, ...
## $ Large.Bags   <dbl> 93.25, 97.49, 103.14, 133.76, 197.69, 127.44, 122.05, ...
## $ XLarge.Bags  <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, ...
## $ type         <chr> "conventional", "conventional", "conventional", "conve...
## $ year         <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, ...
## $ region       <chr> "Albany", "Albany", "Albany", "Albany", "Albany", "Alb...
```

Clear View

There is so much information about avocados. Let us clear the view. There are 14 variables, some require further analysis. There is a variable which contains a year so, what is the range of our data?

```
#Find the range of years in the data
range(my_avocado$year)
```

```
## [1] 2015 2018
```

Our data spans three years starting in 2015 and ending in 2018. How far into 2018 is the data? and when doe it begin in 2015? This is necessary to know to make sure we are able to make accurate comparisons.

```
#Find the maximum date
max(my_avocado$Date)
```

```
## [1] "2018-03-25"
```

```
#find the minimum date
min(my_avocado$Date)
```

```
## [1] "2015-01-04"
```

We have information starting January of 2015 which means it will have data for the entire year. Unfortunantley we only have data until March of 2018, which means we will not be able to compare it accuratley to the other years. It is missing data. I believe it will be best to disregard that data.

```
#Filter out data from 2018
my_avocado <- my_avocado %>% filter(!year == 2018)
```

I want to take a closer look at the variable total volume. What is the total volume consist of? In order to do this I will find the smallest total volume and look at all of the observations in that row.

```
#Find the row with the minimum total volume.
my_avocado %>% filter(Total.Volume == min(Total.Volume))
```

```
##   X       Date AveragePrice Total.Volume X4046 X4225 X4770 Total.Bags
## 1 7 2015-11-08         1.59        84.56  3.95  3.95     0      76.66
##   Small.Bags Large.Bags XLarge.Bags    type year            region
## 1      73.33       3.33           0 organic 2015 MiamiFtLauderdale
```

This data states that on November 8, 2015, the average price of an organic avocado in the region of MiamiFtLauderdale was $1.59. The total volume sold on that day was of 84.56 avocados. It was a slow day for organic avocados in Miami. The total volume is the addition of all the different sizes of avocados and the avocados sold in bags. The different sizes are lables by their price look-up codes. In order to read the data easily I will rename the variables.

```
#Remane columns for clarity
my_avocado <- my_avocado %>%
  rename(Small.Individual = X4046,
         Large.Individual = X4225,
         XLarge.Individual = X4770)
```

I want to know how many distinct regions there are and if any of them overlap.

```
#Find distinct regions
my_avocado %>%
  distinct(region)
```

```
##                 region
## 1               Albany
## 2              Atlanta
## 3  BaltimoreWashington
## 4                Boise
## 5               Boston
```

```
## 6     BuffaloRochester
## 7           California
## 8            Charlotte
## 9              Chicago
## 10    CincinnatiDayton
## 11            Columbus
## 12       DallasFtWorth
## 13              Denver
## 14             Detroit
## 15         GrandRapids
## 16          GreatLakes
## 17  HarrisburgScranton
## 18 HartfordSpringfield
## 19             Houston
## 20        Indianapolis
## 21        Jacksonville
## 22            LasVegas
## 23          LosAngeles
## 24          Louisville
## 25   MiamiFtLauderdale
## 26            Midsouth
## 27           Nashville
## 28    NewOrleansMobile
## 29             NewYork
## 30           Northeast
## 31  NorthernNewEngland
## 32             Orlando
## 33        Philadelphia
## 34       PhoenixTucson
## 35          Pittsburgh
## 36              Plains
## 37            Portland
## 38   RaleighGreensboro
## 39     RichmondNorfolk
## 40             Roanoke
## 41          Sacramento
## 42            SanDiego
## 43        SanFrancisco
## 44             Seattle
## 45       SouthCarolina
## 46        SouthCentral
## 47           Southeast
## 48             Spokane
## 49             StLouis
## 50            Syracuse
## 51               Tampa
## 52             TotalUS
## 53                West
## 54    WestTexNewMexico
```

We have a total of 54 different regions. Some of them do overlap with each other. Therefore I will filter out the regions that do overlap such as “West” or “Northeast”.

```
#Create a vector containing all broad regions
Broad_Regions <- c("West", "California", "GreatLakes", "Midsouth",
```

```
                  "Northeast", "Plains", "SouthCentral",
                  "Southeast")
#Filter out overlapping regions and save it to a new variable.
new_avocado <- my_avocado %>% filter(!region %in% Broad_Regions)
```

There is one more thing I want to do. The dataframe has an inmense amount of infomration but we do not need all of it. So I will create a new dataframe to isolate the variables that I will be interested in.

```
# Create a new dataframe with only variables of interest.
new_avocado <- new_avocado %>%
  select(Date, AveragePrice, Total.Volume, type, year, region)
```

For our data for it to be ready to answer all of our questions we have to be able to use the dates. Using a library called lubridate, I will create two new variables. One will be for months and the other for day, this will allow me to manupulate the data by date or day.

```
# Format dates using lubridate and create variables for month and day.
new_avocado <- new_avocado %>%
  mutate(Date = ymd(Date)) %>% mutate_at(vars(Date), funs(month, day))
glimpse(new_avocado)
```

```
## Rows: 14,441
## Columns: 8
## $ Date         <date> 2015-12-27, 2015-12-20, 2015-12-13, 2015-12-06, 2015-...
## $ AveragePrice <dbl> 1.33, 1.35, 0.93, 1.08, 1.28, 1.26, 0.99, 0.98, 1.02, ...
## $ Total.Volume <dbl> 64236.62, 54876.98, 118220.22, 78992.15, 51039.60, 559...
## $ type         <chr> "conventional", "conventional", "conventional", "conve...
## $ year         <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, ...
## $ region       <chr> "Albany", "Albany", "Albany", "Albany", "Albany", "Alb...
## $ month        <dbl> 12, 12, 12, 12, 11, 11, 11, 11, 11, 10, 10, 10, 10, 9,...
## $ day          <int> 27, 20, 13, 6, 29, 22, 15, 8, 1, 25, 18, 11, 4, 27, 20...
```

Let us start analyzing this data with a simple overview. I will take a look at the average price and compare that to the median price oof each region. This will show us if there are outliers. If there are we will ttry to find out why. The same thing will be done with the total volume of avocados sold. Then I want to know how much of a difference there is between organic and conventional.

```
# Create an overview of the data
new_avocado %>%

  # group the data by region and type
  group_by(region, type) %>%

  #Find the average and median price per avocado and volume sold
  summarise(avg_price = mean(AveragePrice),
            mid_price = median(AveragePrice),
            avg_vol = mean(Total.Volume),
            mid_vol = median(Total.Volume)) %>%
  #arrange the table by descending average price
  arrange(region)
```

```
## # A tibble: 92 x 6
## # Groups:   region [46]
##    region              type         avg_price mid_price avg_vol mid_vol
##    <chr>               <chr>            <dbl>     <dbl>   <dbl>   <dbl>
##  1 Albany              conventional      1.35      1.35  90514.  92080.
##  2 Albany              organic           1.79      1.83   2007.   1796.
```

```
##  3 Atlanta             conventional      1.07      1.07 500802. 475922.
##  4 Atlanta             organic           1.61      1.62  11132.   9859.
##  5 BaltimoreWashington conventional      1.35      1.29 759227. 744159.
##  6 BaltimoreWashington organic           1.74      1.7   21378.  18004.
##  7 Boise               conventional      1.07      1.04  81641.  80034.
##  8 Boise               organic           1.61      1.54   2426.   2138.
##  9 Boston              conventional      1.30      1.26 551728. 552369.
## 10 Boston              organic           1.75      1.81  12838.  11414.
## # ... with 82 more rows
```

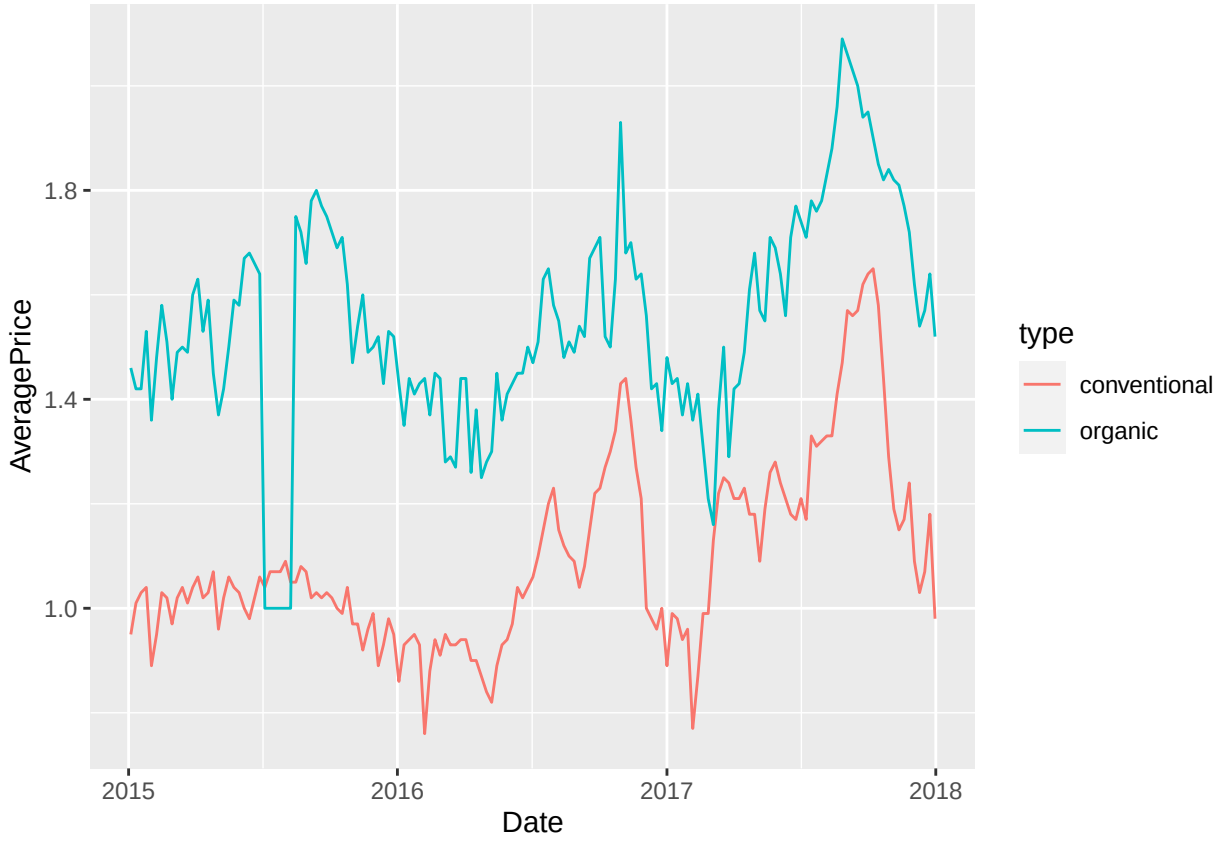
When it comes to price the average and the median are similar for all regions. In general an organic avocado is more expensive than a conventional one per region. In some regions the price of an organic avocado is lower than a conventional one in another region. The total volume of avocados sold shows a big disparity between organic and conventional. In some regions the median and average are very different, which means that there are outliers present.

Question Time

What are the price trends for each year in the entire Unites states?

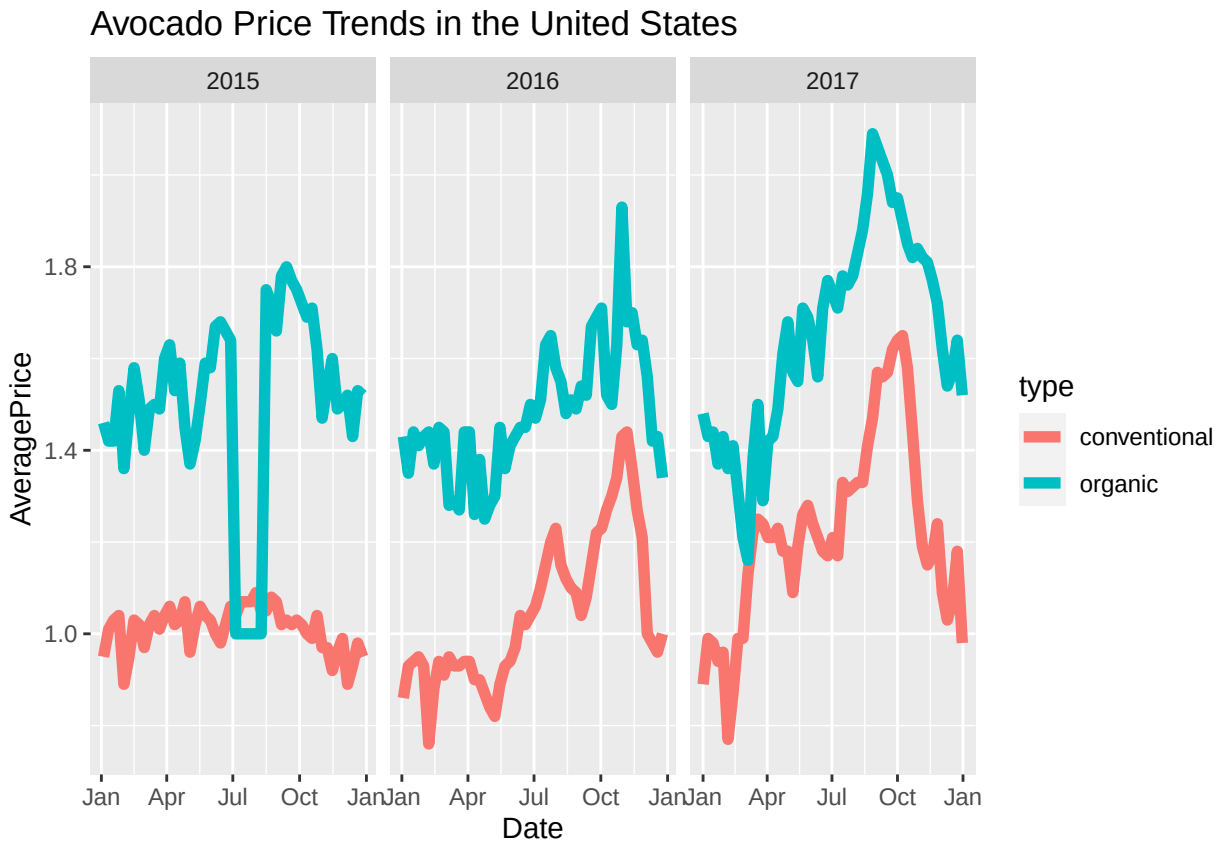
First I will plot the raw data. I will filter the region of TotalUs and set Date as the x-axis. The average price will be the y-axis. Then I will color the graphs based on which type of avocado. Since we want to see the trend as time goes on I will use a line plot.

```
#Set up the variables to plot and save it
US <- ggplot(data = new_avocado %>% filter(region == "TotalUS"),
             aes(y = AveragePrice, x = Date, color = type))
# Add a plot type
US + geom_line()
```



This graph shows us that there is no steady rise in the price of avocados. This is a good thing for the avocado lovers but during the year there is much fluctuation. Lets sepeare this graph by year and make the lines thicker so they can be easier to see. I will als adjust the x-axis tick marks and labels. Lastly I will add a title.

```
US + geom_line(size = 2) + facet_wrap(~year, scales = "free_x") + scale_x_date(date_labels = "%b") +
  ggtitle("Avocado Price Trends in the United States")
```



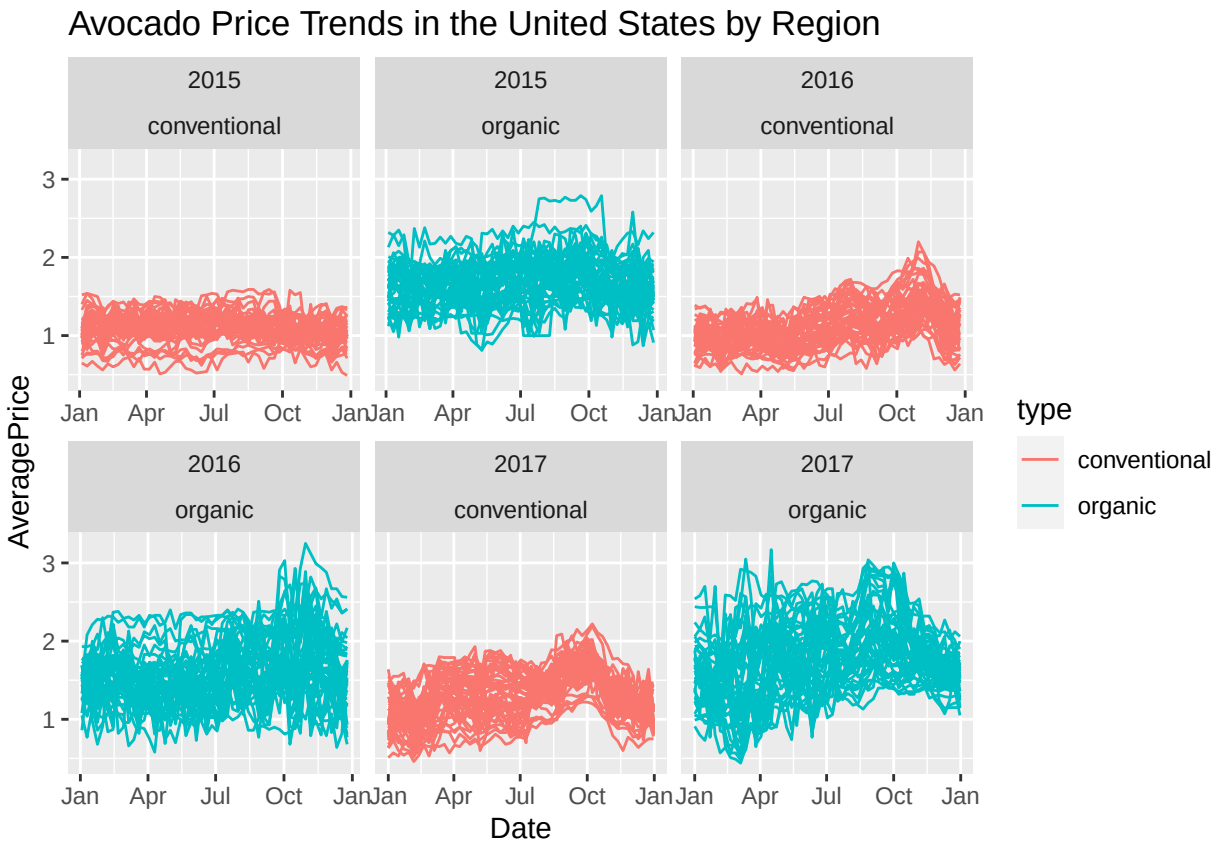
This provides a much clearer picture. 2015 is a very interesting year, in my opinion. The organic avocados took a big drop in price but was able to recover a couple months later. Could this be because demand went back up? Or could it have been a issue with production? I will come back to that later. The conventional avocado, on the other hand, was very steady compared to the all the other years. In 2016 and 2017 the conventional and organic prices almost mirror each other, which is to be expected, since they are almost the same product. Although the prices fluctuate during the year it always ends with a decline in price. This does not necessarily mean that the finishing price is lower than the starting price for a particular year. Let's take a look at the different regions t see if we see the same pattern.

What are the price trends per region? How do they compare to the TotalUS data?

To answer this question, I will plot the same data but group it by region. This will allow us to see the data for each region. It will be overlapping but it will still gives us a big picture view of the price trend.

```
# Set up data to be plotted
ggplot(new_avocado, aes(Date, AveragePrice, color = type)) +
  # Select a type or plot
  geom_line(aes(group = region)) +

  # Make a differend plot for eyery year and type
  facet_wrap(~year + type, scales = "free_x") +
```

```
  # X-axis is scaled by months
  scale_x_date(date_labels = "%b") +

  #add a title
  ggtitle("Avocado Price Trends in the United States by Region")
```



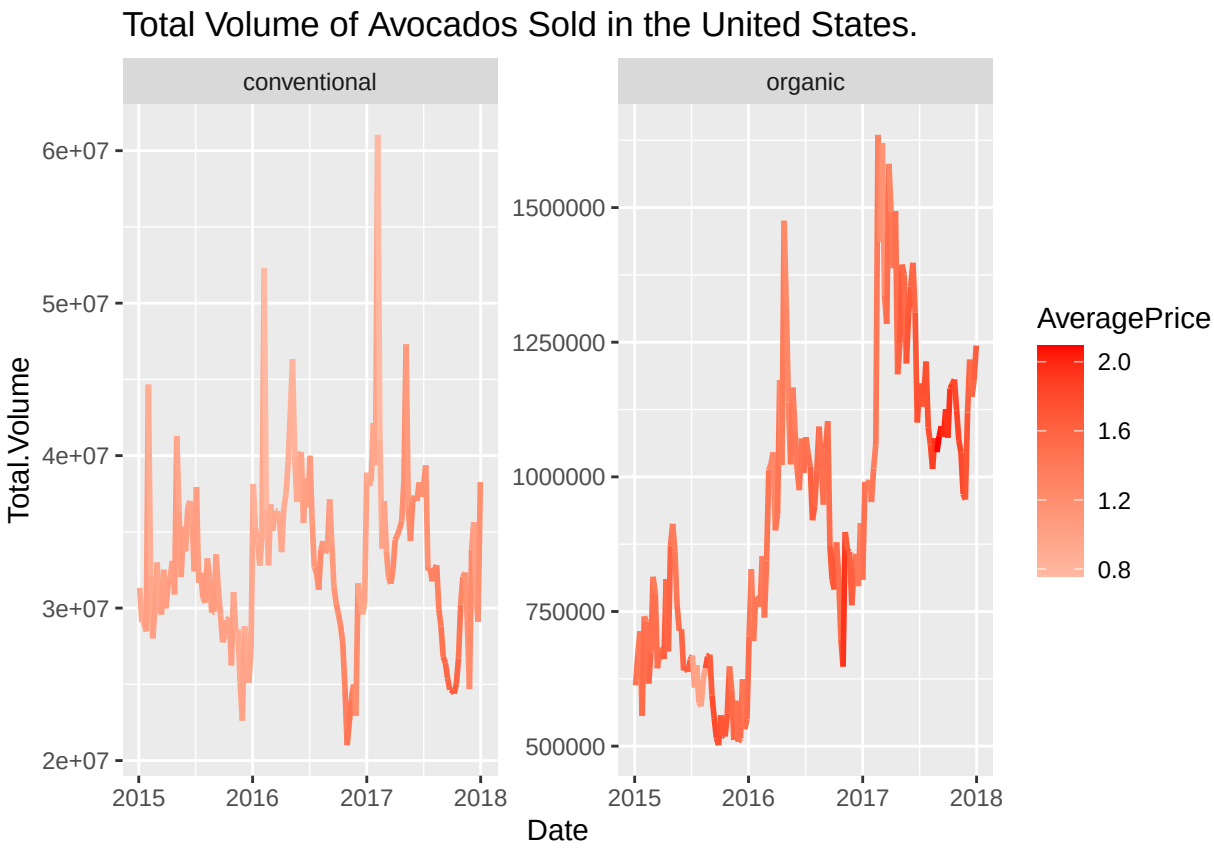
This data shows a different story. Here the price drop that we saw with the who United States does not happen. There are some regions which actually saw a price increase in July of 2015. This means every region sets their prices independently of one another. This is best viewed with the price activity of organic avocados in 2017. Unlike the other graphs which had only some outliers, in this graph it is hard to see a clear pattern especially in the beginning. In some regions it looks like the extra dollar on a salad may not be enough.

What are the sales trends for the entire United States? How do they compare to the prices?

I will plot the sales of avocados to include only the entire United States. Then I will seperate the conventional from the organic due to the volume difference and I will color it based on the average price of an avocado.

```
#select which data to plot and save it to a new variable.
# Color the plot based on the AveragePrice
US2 <- ggplot(data = new_avocado %>% filter(region == "TotalUS"),
              mapping = aes(x = Date, y = Total.Volume))

#Select a plot to use and set color based on Average price
US2 + geom_line(aes(color = AveragePrice), size = 1) +

  # Make different graph for each type and have y-axis
  facet_wrap(~type , scales = "free") +

  # X-axis is ticked by year
```

```
  scale_x_date(date_labels = "%Y") +

  #Scale the color so that darker is higer price
  scale_color_gradient2(low = "white", high = "red") +

  #Add a title
  labs( title = "Total Volume of Avocados Sold in the United States.")
```



Similar to the pattern we saw with the prices there is a constant fluctuation in the sales of avocados. As expected the sales go up when the prices go down. What is surprising is that in 2015 the price per avocado was consistent and yet the sales still had a large variety. Sales seem to always rise during the beginning of the year but it does not last very long. This could be because, after some research, the beginning of the season for avocados in California is in January. Supply is then high and fresh which lowers prices resulting in more sales. Even so the organic sales have been increasing every year. Our answer gets closer and closer.

Do different regions have the same behavior in sales?

This will help us determine if we can use only the region of TotalUS to answer our main question. I want to compare the region with the most amount of sales to the one with the least amount of sales. First let us find out which regions those are.

```
#Create a temporary data frame without "TotalUs" and without rganic type
temp <- new_avocado %>%
  filter(!region == "TotalUS", type == "conventional")

#Find the region with maximum sales
which.max(temp$Total.Volume) %>% temp$region[.]
```

```
## [1] "LosAngeles"
```

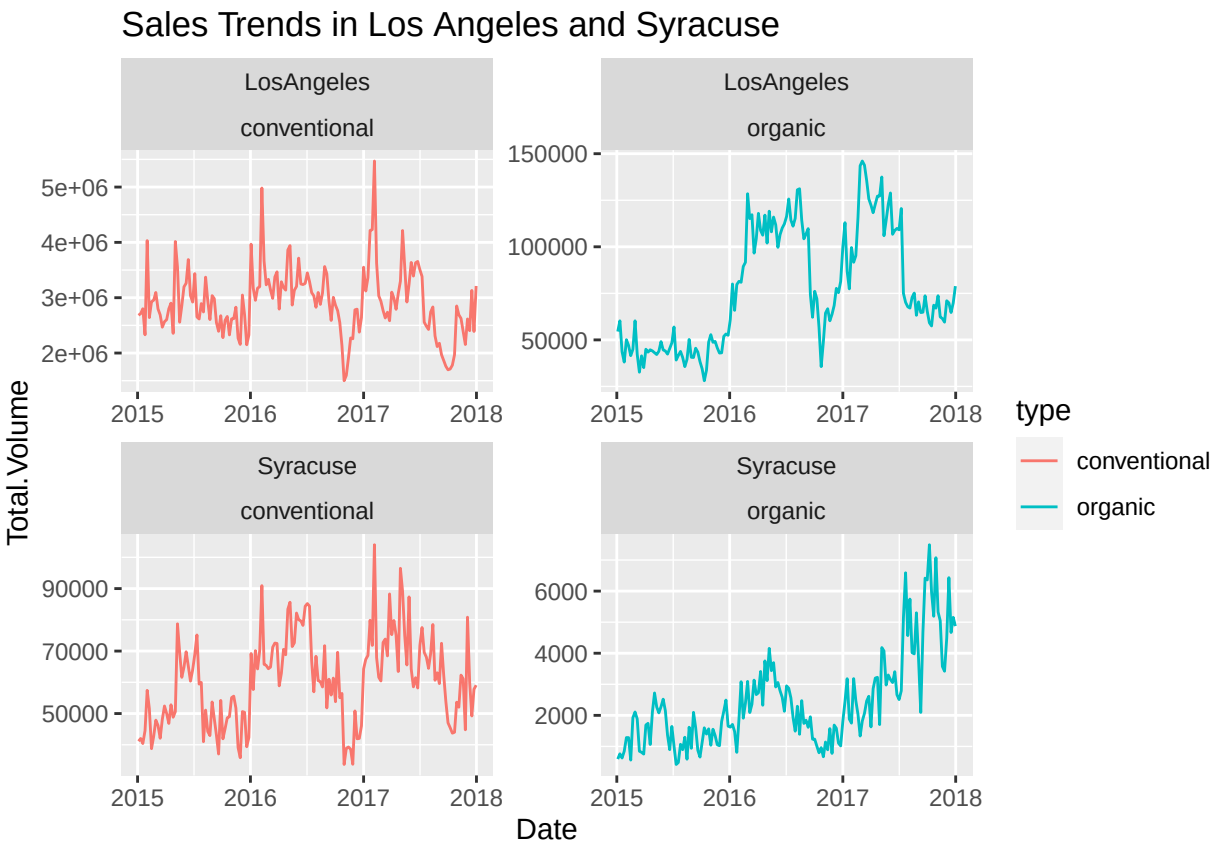
```r
#Find the region with minimum sales
which.min(temp$Total.Volume) %>% temp$region[.]
```

```
## [1] "Syracuse"
```

```r
#Erase temporary data frame
temp <- NULL
```

Let us compare the sales patters of Los Angeles and Syracuse. This will allw us to see if volume is a factor in the pattern of sales.

```r
#Create a variable with plotting data
LH <- ggplot(new_avocado %>%
         filter(region %in% c("LosAngeles", "Syracuse")),
       aes(Date, Total.Volume, color = type))

#Plot using a line graph
LH + geom_line() +

  #Seperate the graph by region and type
  facet_wrap(~region + type, scales = "free") +

  # X-axis is ticked by year
  scale_x_date(date_labels = "%Y") +

  #Add a title
  labs( title = "Sales Trends in Los Angeles and Syracuse")
```



The conventional sales are in par with what we saw with the data for the entire United States. There is big

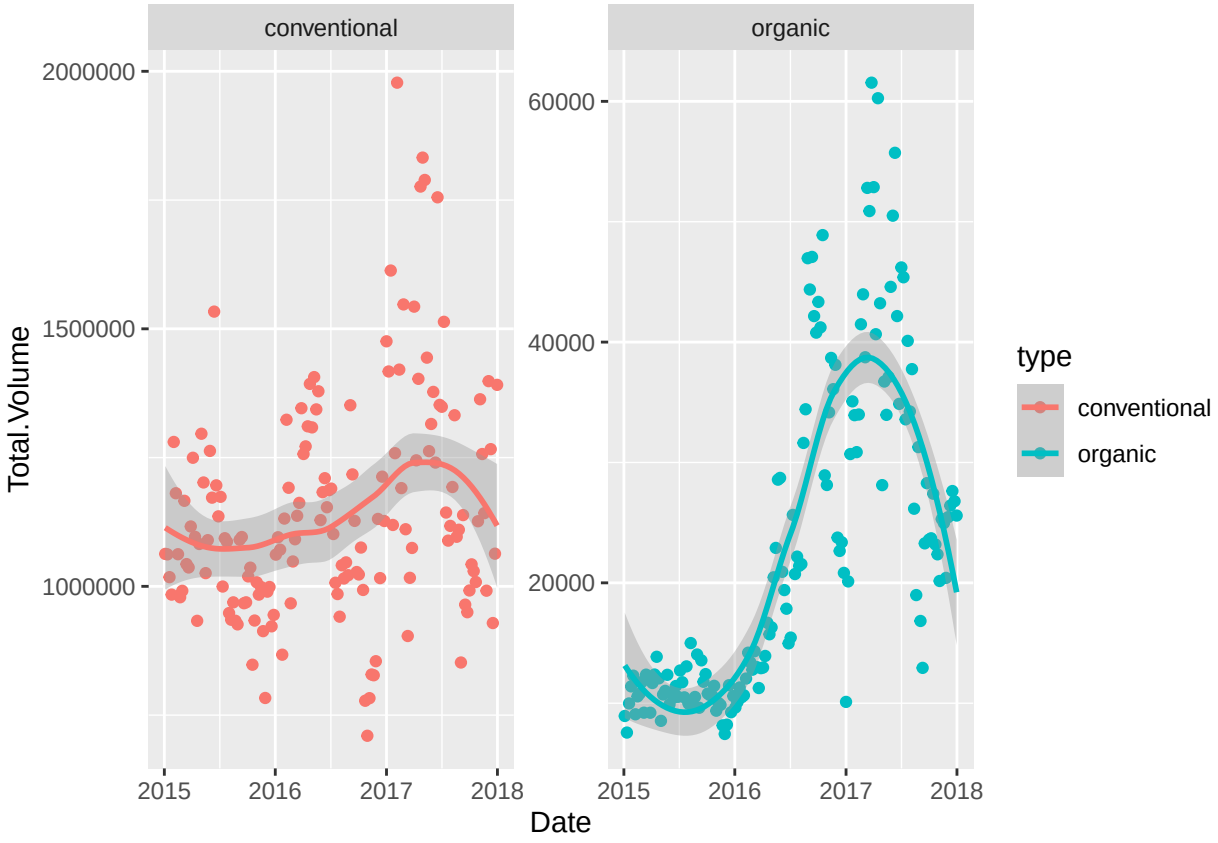
fluctuation throughout the year but nothing out of the ordinary. Once again the organic trends are different. In Syracuse organic avocado sales have been on a steady rise since 2015. Los Angeles, on the other hand, has a more sporadic trend it also is slowly increasing it sales. Even when there was a big dip in sales in 2016 for Los Angeles, it did not drop below the lowest point in the data. Organic avocados are trending now, but will they take over? As far as volume goes there is no significant correlation between the amount sold and the region where it is sold for my purposes.

Are organic sales affecting conventional sales?

In order to see the effects I will be plotting the region of Houston, and analyze the findings.

```
#Create a variable with plotting data
BR <- ggplot(new_avocado %>% filter(region == "Houston"),
             aes(Date, Total.Volume, color = type))

#Plot using a line graph
BR + geom_point() + geom_smooth(method =  "loess") +

  #Seperate the graph by region and type
  facet_wrap(~type, scales = "free") +

  # X-axis is ticked by year
  scale_x_date(date_labels = "%Y")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Both the conventional and organic sales have the same curvature. This means that those who love organic are not lowering the sales of the conventional avocado. We have all we need to answer our main question.

Summary

How has the want for a healthier lifestlye affected the sales of avcados?

I wanted to know how much can cultural norms affect the sale of a product. In this case it was the want of a healthy lifestyle versus avocado sales. First I we took a look at the price trends of the United States in order to rule it out as a factor in the amount of avocados sold. We saw that the price of avocado is very stable. The biggest fluctuations happen during the year but it always returns back to normal. In order to make sure this applied to all regions I plotted the price trends of all regions and they gave the same results. Since we then knew that price was not going to be a big factor we took a look at sales. We used the data labeled "TotalUs" and noticed that, although the numbers were low compare to conventional sales, organic sales were steadily rising. In order to make sure we could rely on this data we compared two different regions, the one with the most sales and the one with the least, to see if there were any abnormalities. Once again our findings were the same as when we used the "TotalUS" data, organic sales were going up and the region did not matter. If organic sales are going up, does that affect conventional sales? In order to find this answer I plotted the region of "Houston" and discovered that organic sales don't affect conventional sales. Both sales trends actually behave very similarly. This means people aren't switching to organic, new healthy lifestylists are buying organic avocados. In conclusion organic avocado sales are going up but it will take a long time for it to reach the same volume of sales as conventional avocados.