Avocado Price Prediction - Exploratory Data Analysis

Ragunath Gunasekaran

2020-11-18

Introduction

When I started this project, I learned that Avocado became America's new favorite fruit and It is a "superfood". Hass Avocado Board (HAB) helps to increase the consumption of Avocados. The study shows that Avocado Consumption per capita in the 90's was 1.6 lbs but per capita is increased, as of 2017, Avocado Consumption per capita is 7.1 lbs.

My Project intends to thoroughly exploratory data analysis of the avocado prices increase along with customer behavior. The Data comes from Kaggle which is provided by Hass Avocado Board website compiled into a one CSV file and the data contains from 2015 to 2018 Avocado Purchases in the USA. The Project also tries to analyze the price elasticity of demand and find the comparison of conventional and organic avocados since Organic consumption is increasing in recent days.

I have chosen the fruit Avocado, but this approach can be extended to other food exploratory data analysis by using the Purchases/Sales Data.

The below are research questions which we target to find the answers from the outcome of the EDA of this project.

Is there any Linear relationship between volume and price? Which prices are more either Organic or conventional? How does the season impact the Sales of Avocado Sales? Can we meet the Supply-demand approach in economics? If there is enough demand for Avocado food, can prices be increased? How the price prediction impacts the Region? (Example, near Port Access) How are the purchase increases by date? What is the USA Average price?

Data Preparation and Clean Up

At first step, I have loaded all the required libraries for this project (ggplot2, tidyverse, etc) and then loaded the source data from .csv file data to Dataset. I changed the formatting of data since this is important for all the Data Visualization and Calculation. while performing the Data Analysis, I noticed that Regions where avocados were sold/purchased were overlapped, and some regions contained the same name but listed differently. I also subset the Total U.S responded under the Region Variable. (Region variables from characters to factors)

From the Date, I created a variable called Season - Spring, Summer, Fall, Winter and also a Month, Year variable.

The Below are steps followed as part of Data Preparation

Renamed the column Names - Average_Price, Total_Volume

Date formatted for Visualization and linear methods

Removed the white spaces in the Text fields – Region, Type, Type, etc

Find any missing values exists in the Data frame - Date

Avoid the missing values if any

Unite the fields into one field wherever required.

Used toupper/tolower make the columns into Upper if required.

Added new Columns - Year, Month, Season

Also before cleaning the data, I noticed that skewness existed.

Data Overview

Metadata Details

The below is the Structure of avocado_dataset_usa dataset which explains the detail of metadata.

```
## Structure of avocado_dataset_usa ##
str(avocado_dataset_usa)
```

```
## ’data.frame’:    338 obs. of  16 variables:
##  $ X            : int  51 51 50 50 49 49 48 48 47 47 ...
##  $ Date         : POSIXct, format: "2015-01-04" "2015-01-04" ...
##  $ Average_Price: num  0.95 1.46 1.01 1.42 1.03 1.42 1.04 1.53 0.89 1.36 ...
##  $ Total_Volume : num  31324278 612910 29063543 669529 29043459 ...
##  $ Small_Hass   : num  12357161 233286 11544811 270967 11858139 ...
##  $ Large_Hass   : num  13624083 216611 12134773 260972 11701948 ...
##  $ XLarge_Hass  : num  844093 4371 866575 3830 831302 ...
##  $ Total_Bags   : num  4498940 158642 4517384 133760 4652070 ...
##  $ Small_Bags   : num  3585322 115069 3783261 106844 3873041 ...
##  $ Large_Bags   : num  894946 43573 718334 26916 771093 ...
##  $ XLarge_Bags  : num  18673 0 15789 0 7935 ...
##  $ Type         : chr  "conventional" "organic" "conventional" "organic" ...
##  $ Year         : chr  "2015" "2015" "2015" "2015" ...
##  $ Region       : Factor w/ 54 levels "Albany","Atlanta",..: 52 52 52 52 52 52 52 52 52 52 ...
##  $ Month        : chr  "01" "01" "01" "01" ...
##  $ Season       : chr  "Winter" "Winter" "Winter" "Winter" ...
```

```
dim(avocado_dataset)
```

```
## [1] 18249    14
```

After cleaning the data, the Dataset contains 18,249 rows and 14 Variables.

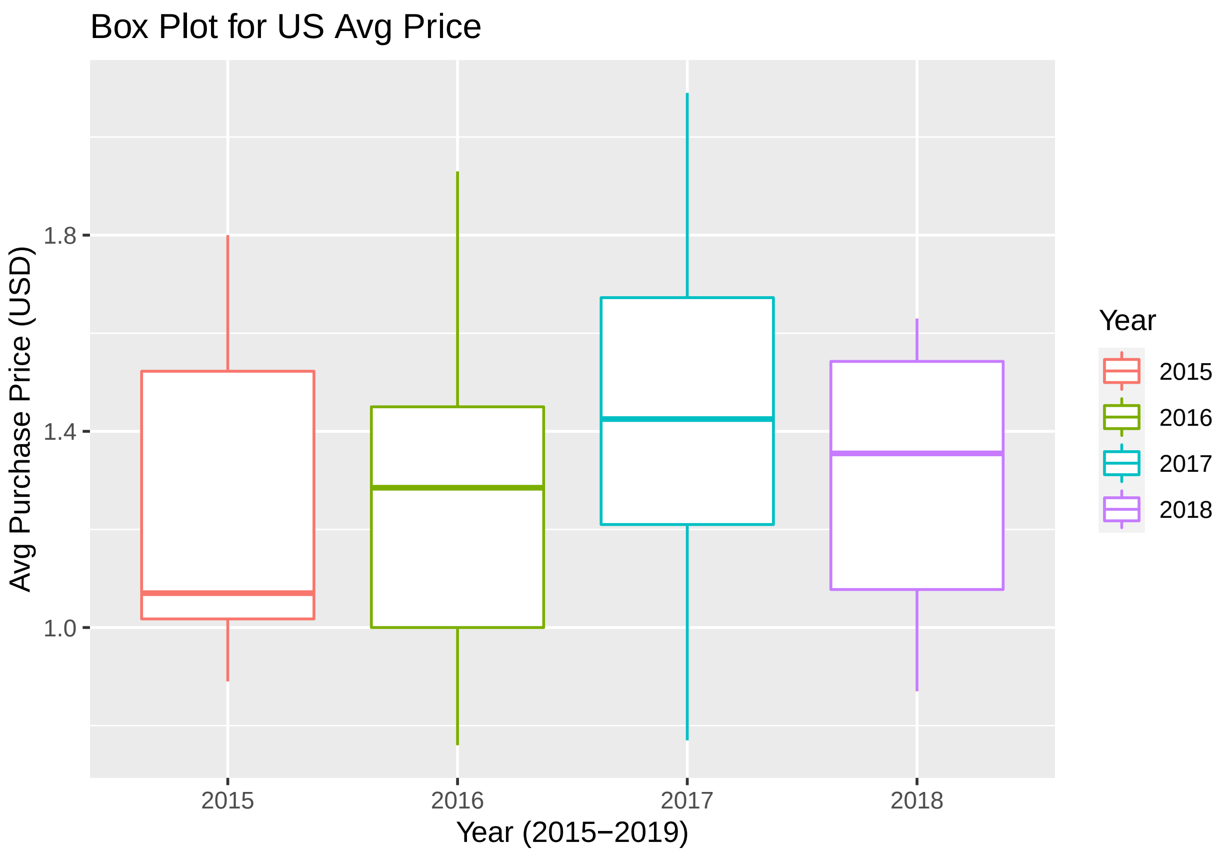
Data Explanation

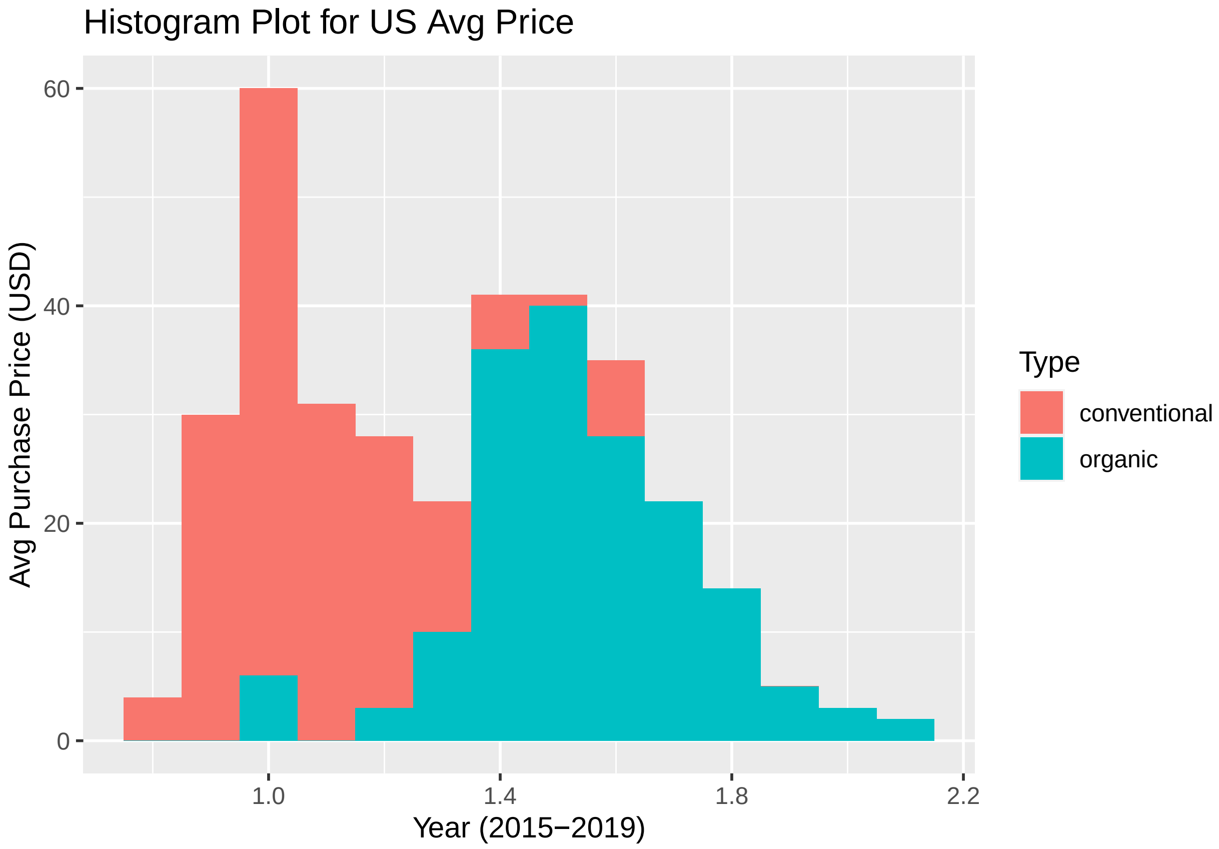
The Overview of Dataset, I displayed only 10 rows as Sample.

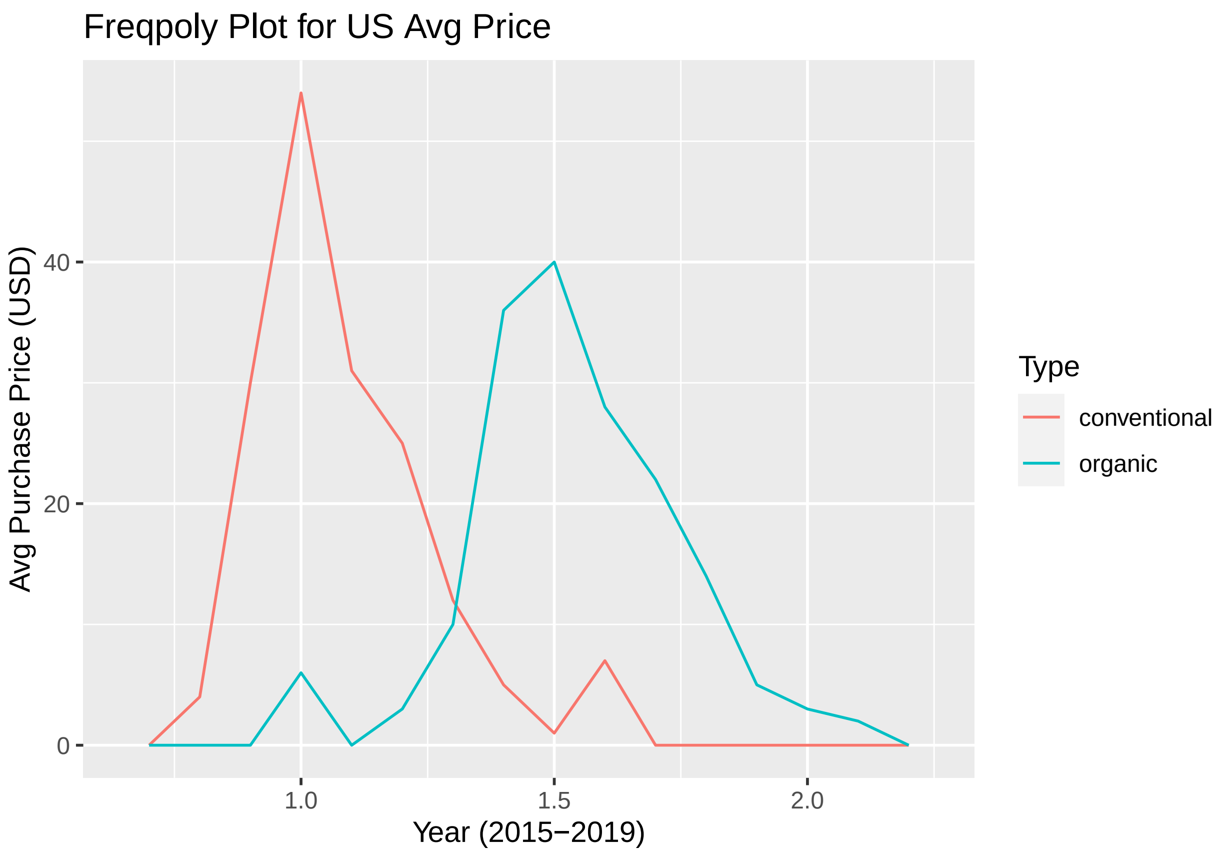
```
## Data Overview ##
head(avocado_dataset_usa,10)
```

```
##          X       Date Average_Price Total_Volume Small_Hass Large_Hass
## 2704  51 2015-01-04          0.95   31324277.7 12357161.3 13624083.1
## 11830 51 2015-01-04          1.46     612910.2   233286.1   216611.2
## 2703  50 2015-01-11          1.01   29063542.8 11544810.5 12134773.4
## 11829 50 2015-01-11          1.42     669528.9   270966.7   260971.6
## 2702  49 2015-01-18          1.03   29043458.9 11858139.3 11701947.8
## 11828 49 2015-01-18          1.42     713120.0   254319.6   311811.0
## 2701  48 2015-01-25          1.04   28470310.8 12167445.0 10734652.8
## 11827 48 2015-01-25          1.53     556368.9   207494.9   212312.0
## 2700  47 2015-02-01          0.89   44655461.5 18933038.0 18956479.7
## 11826 47 2015-02-01          1.36     740897.0   302561.5   259286.4
##       XLarge_Hass Total_Bags Small_Bags Large_Bags XLarge_Bags         Type
## 2704    844093.32  4498940.0 3585321.58  894945.63    18672.81 conventional
## 11830     4370.99   158641.8  115068.71   43573.12        0.00      organic
## 2703    866574.66  4517384.2 3783261.16  718333.87    15789.15 conventional
## 11829     3830.42   133760.1  106844.49   26915.63        0.00      organic
## 2702    831301.90  4652069.8 3873041.26  771093.20     7935.35 conventional
## 11828     4020.85   142968.6  101850.23   41118.33        0.00      organic
## 2701    768020.05  4800192.9 3978636.90  812924.73     8631.31 conventional
## 11827     4753.87   131808.1   95964.83   35843.27        0.00      organic
## 2700   1381516.11  5384427.6 4216452.03 1121076.47    46899.12 conventional
## 11826     5852.28   173196.8  129953.15   43243.63        0.00      organic
##       Year  Region Month Season
## 2704  2015 TotalUS    01 Winter
## 11830 2015 TotalUS    01 Winter
## 2703  2015 TotalUS    01 Winter
## 11829 2015 TotalUS    01 Winter
## 2702  2015 TotalUS    01 Winter
## 11828 2015 TotalUS    01 Winter
## 2701  2015 TotalUS    01 Winter
## 11827 2015 TotalUS    01 Winter
## 2700  2015 TotalUS    02 Winter
## 11826 2015 TotalUS    02 Winter
```

Box, Histogram, freqpoly plot for US Average Price


Box Plot for US Avg Price
Avg Purchase Price (USD)
1.8
1.4
1.0
2015
2016
2017
2018
Year (2015−2019)
Year
2015
2016
2017
2018


Histogram Plot for US Avg Price
Avg Purchase Price (USD)
60
40
20
0
1.0
1.4
1.8
2.2
Year (2015−2019)
Type
conventional
organic



Based on the above Boxplot, we can conclude that the average price of Avocados in United was 1.33 USD.

This is higher level analysis and there is a need to deep further into the Analysis and present the results.

Exploratory Data Analysis:

Exploratory Data Analysis is an important process in Data Science since EDA deals with investigations or Analysis on given dataset to find the insight about Variables, patterns, Relationships, etc.

1. Avocado Purchases Analysis at Type and Season Level

Avocados Types : conventional, organic

Seasons : Spring, Summer, Fall, Winter

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)

## ‘summarise()‘ regrouping output by ’Type’ (override with ‘.groups‘ argument)

##
##
## |     Type     | Mean_Volume | percentage_Volume |
## |:------------:|:-----------:|:-----------------:|
## | conventional |  33735039   |       97.21       |
## |   organic    |   967566    |       2.788       |
##
##
## |     Type     | Season | Mean_Volume | percentage_Volume |
## |:------------:|:------:|:-----------:|:-----------------:|
## | conventional |  Fall  |  27767818   |       20.7        |
## | conventional | Spring |  36097057   |       26.91       |
## | conventional | Summer |  34255623   |       25.53       |
## | conventional | Winter |  36043338   |       26.87       |
## |   organic    |  Fall  |   829472    |       21.54       |
## |   organic    | Spring |   1121772   |       29.13       |
## |   organic    | Summer |   936757    |       24.32       |
## |   organic    | Winter |   963355    |       25.01       |
```

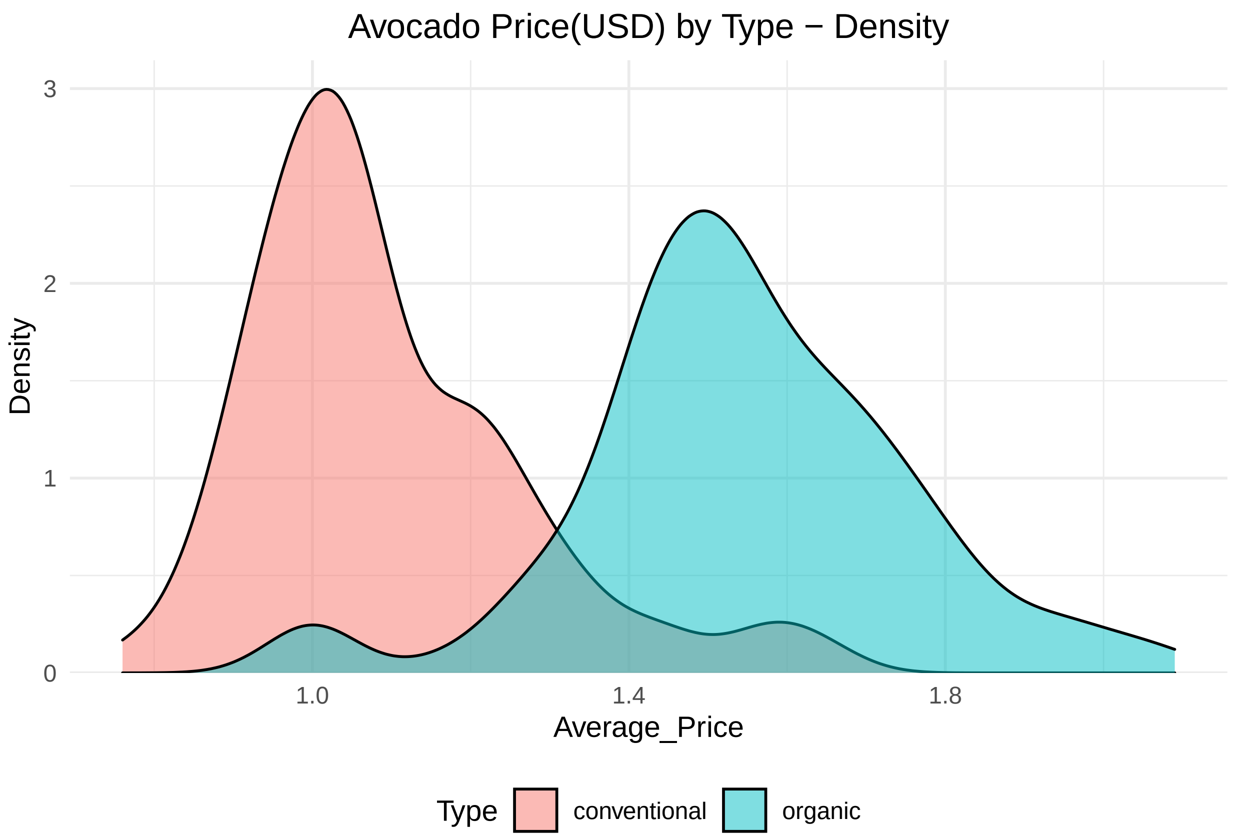
Summary :

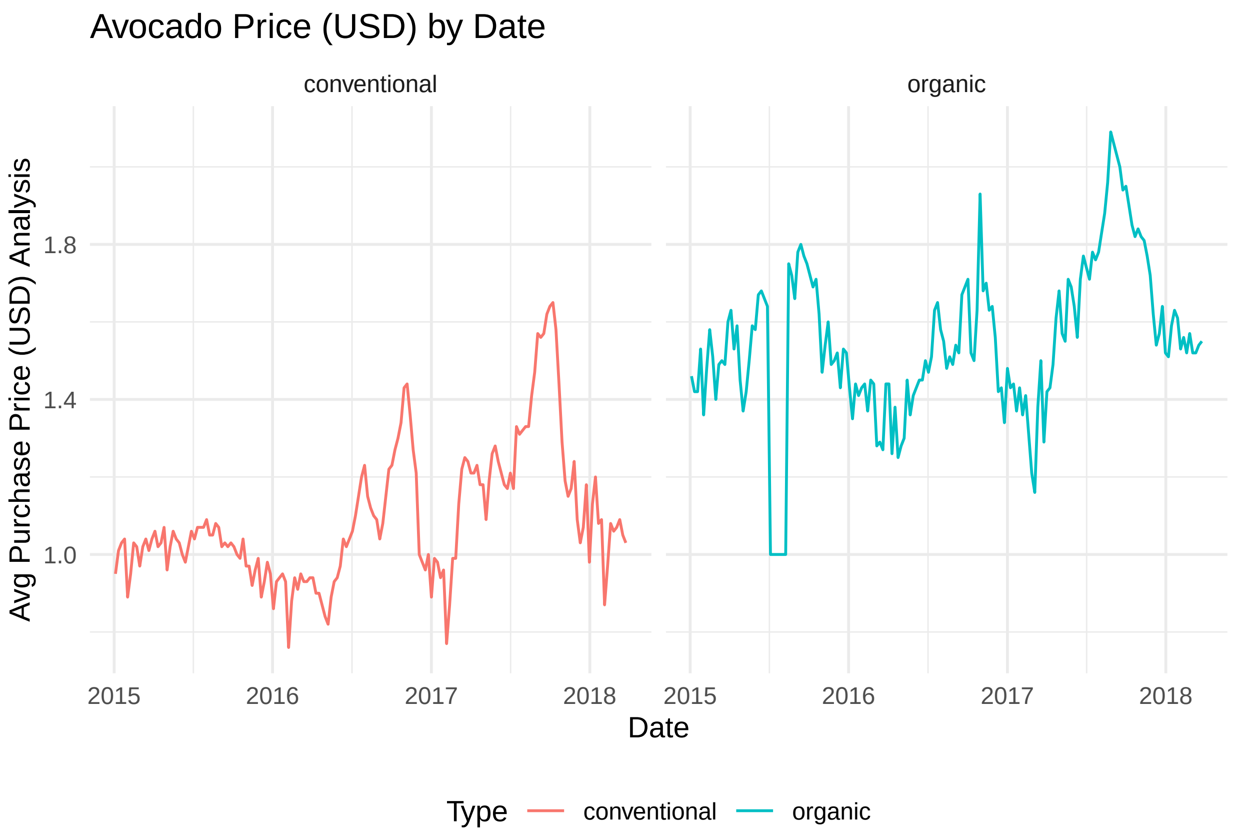
```
    1. conventional Avocados Purchased (97.21%) more than organic (2.78%)
    2. Avocados Purchase was more in Spring Season ( Both Organic and conventional Types)
    3. The conventional Type of Avocados purchase in Winter (26.86%) was very close to Spring (26.90%
    4. Organic Type of Avocados were purchased very less in Fall (21.53%)
```

2. Price Analysis of Avocados at Type and Date Level

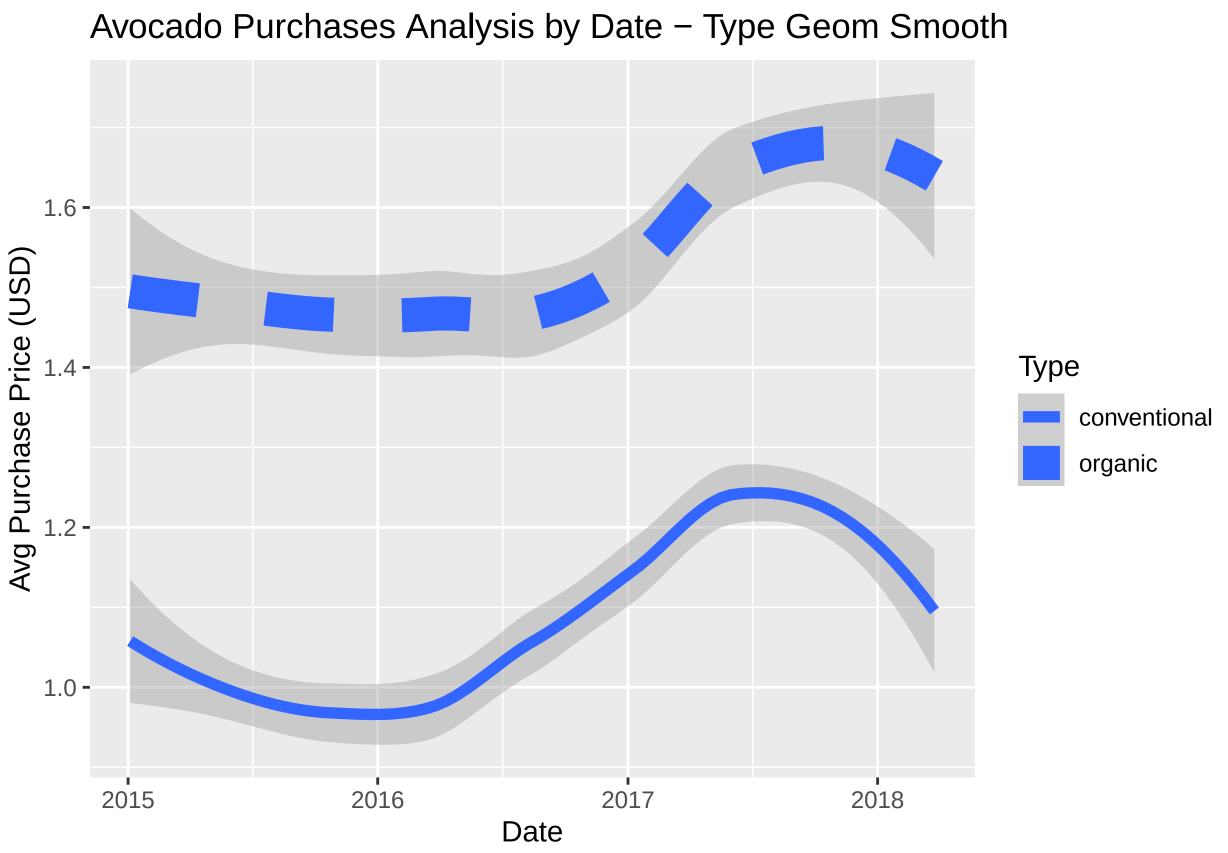
In this Section, I want to analysis Average Price(USD) by including Date, Season,Type and Region.

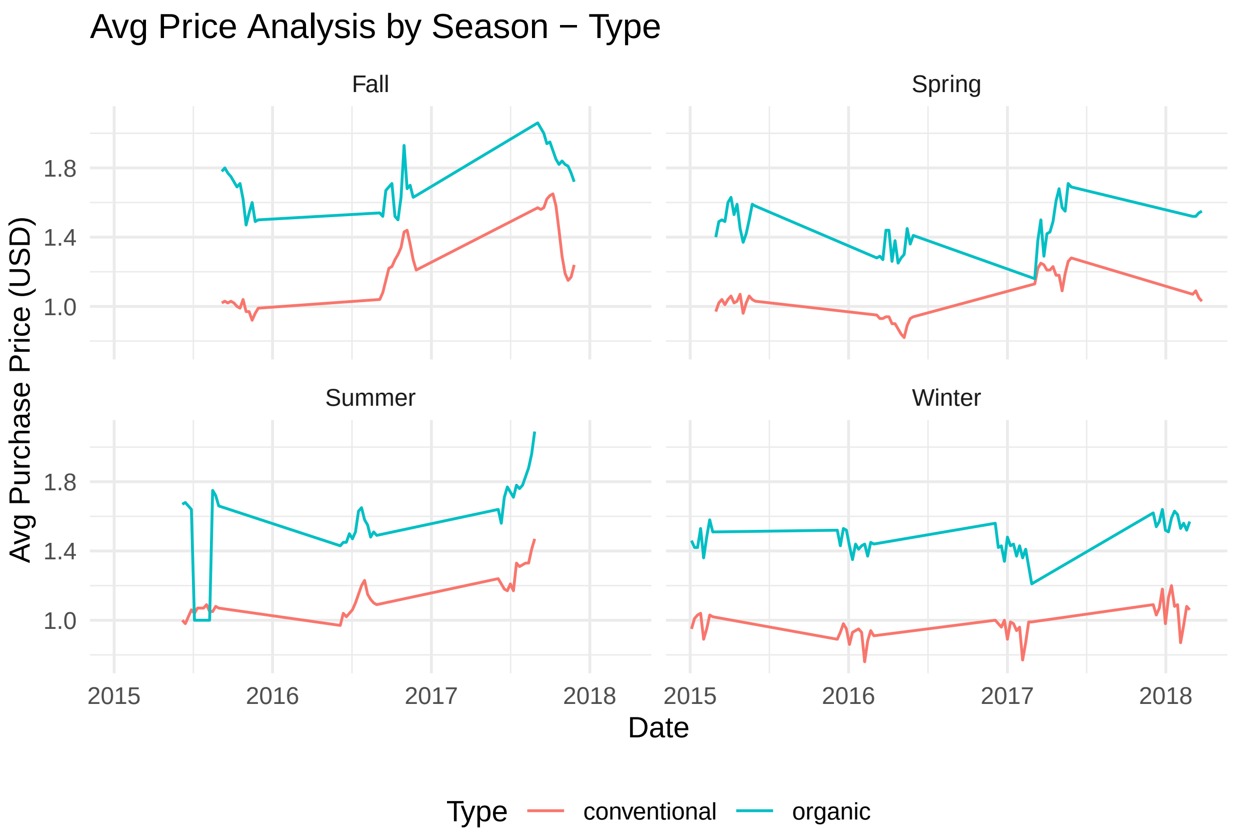
1. Price Analysis by Type by using density plot
2. Price Analysis by Date (2015-2019) along with Type by using line plot
3. Price Analysis by Date (2015-2019) along with Type by Geom_smooth plot
4. Price Analysis by Date (2015-2019) along with Season - Type by using line plot

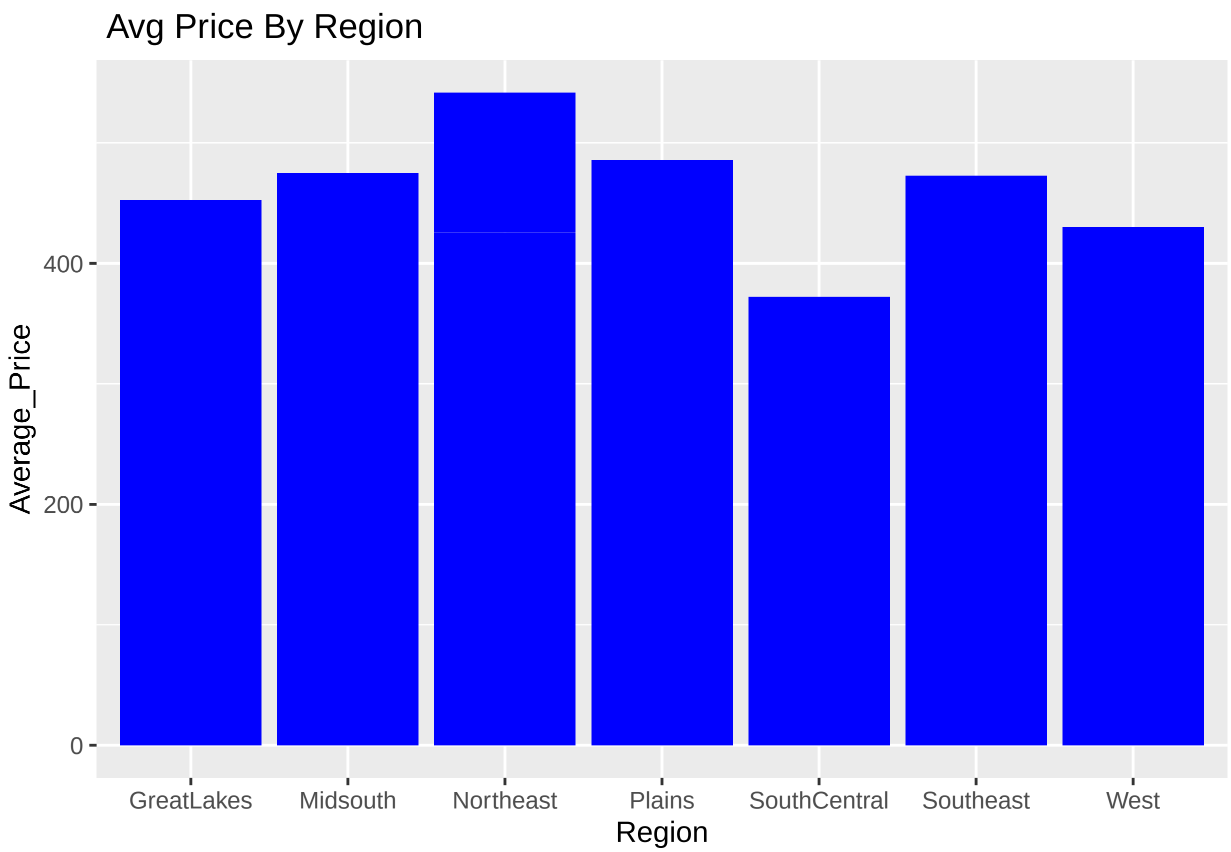

Avocado Price(USD) by Type − Density
Density
3
2
1
0
1.0
1.4
1.8
Average_Price
Type
conventional
organic



```
## Warning: Using size for a discrete variable is not advised.

## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```


Avocado Purchases Analysis by Date − Type Geom Smooth
Avg Purchase Price (USD)
1.6
1.4
1.2
1.0
2015
2016
2017
2018
Date
Type
conventional
organic


Avg Price Analysis by Season − Type
Fall
Spring
Summer
Winter
Avg Purchase Price (USD)
1.0
1.4
1.8
2015
2016
2017
2018
Date
Type
conventional
organic



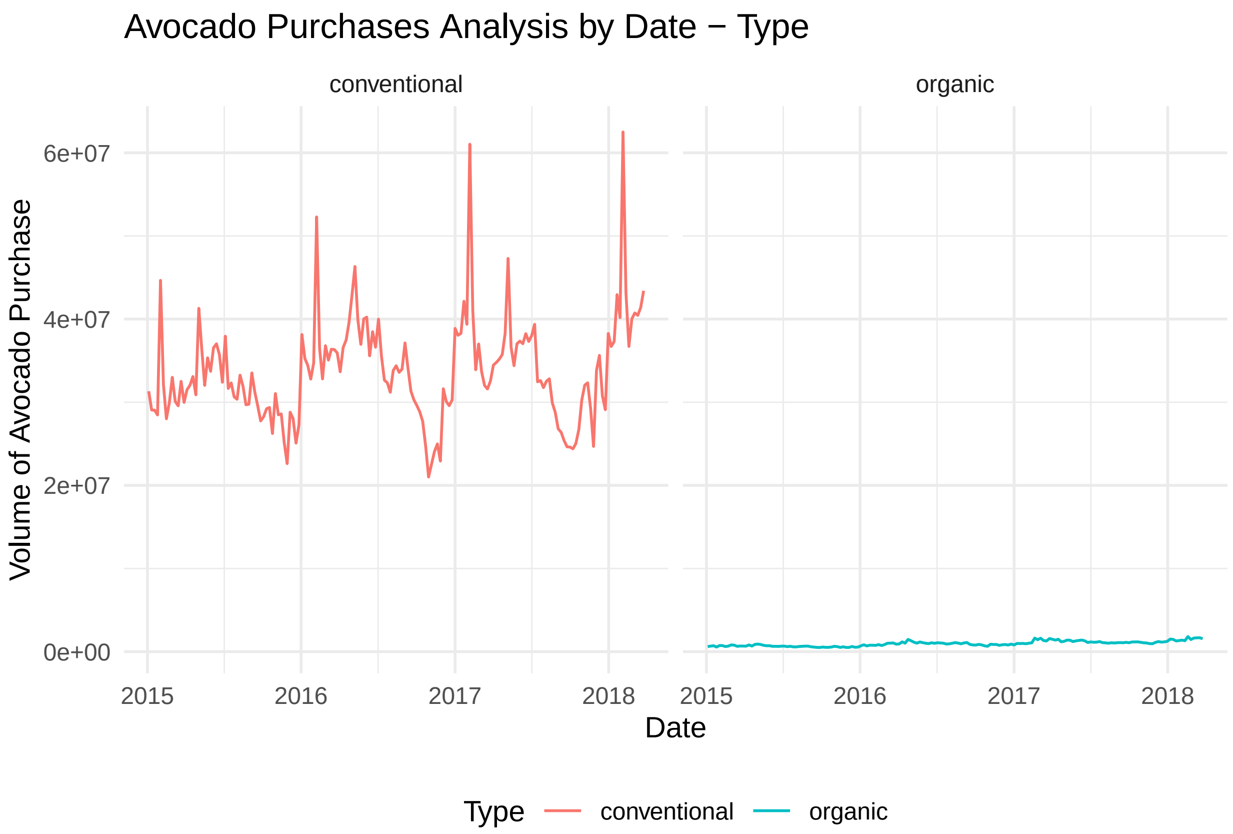
Summary :

```
1. Price - Organic Avocados were more expensive than conventional Avocados
2. The minimum (0.76 USD) and maximum (2.09 USD) price
3. As expected, The average price is lower in winter Season, but price slowly increases in spring
4. From Fall back into the winter Season the price slowly declines
5. Am concluding based on the above study, Season or weather impact the avocado price
```

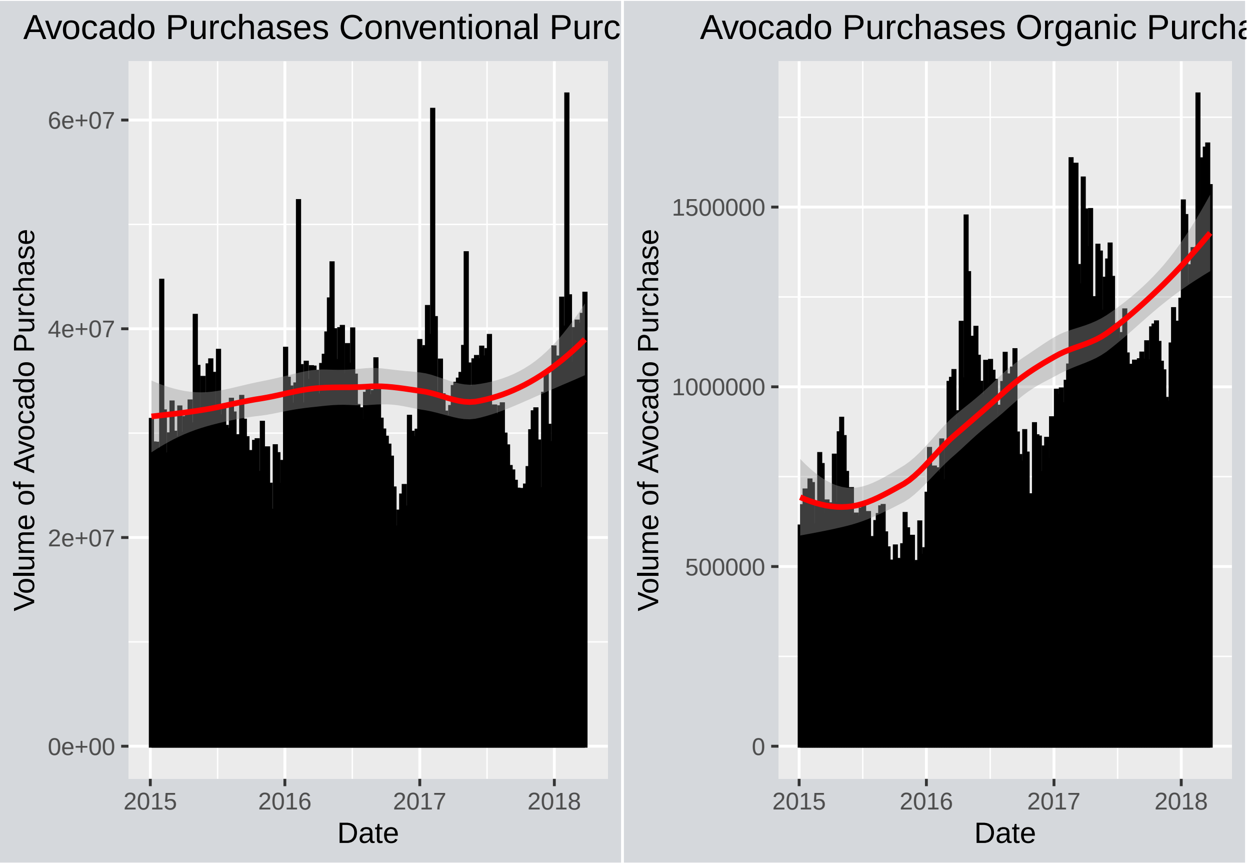
3. Organic and conventional Avocados Purchase Analysis

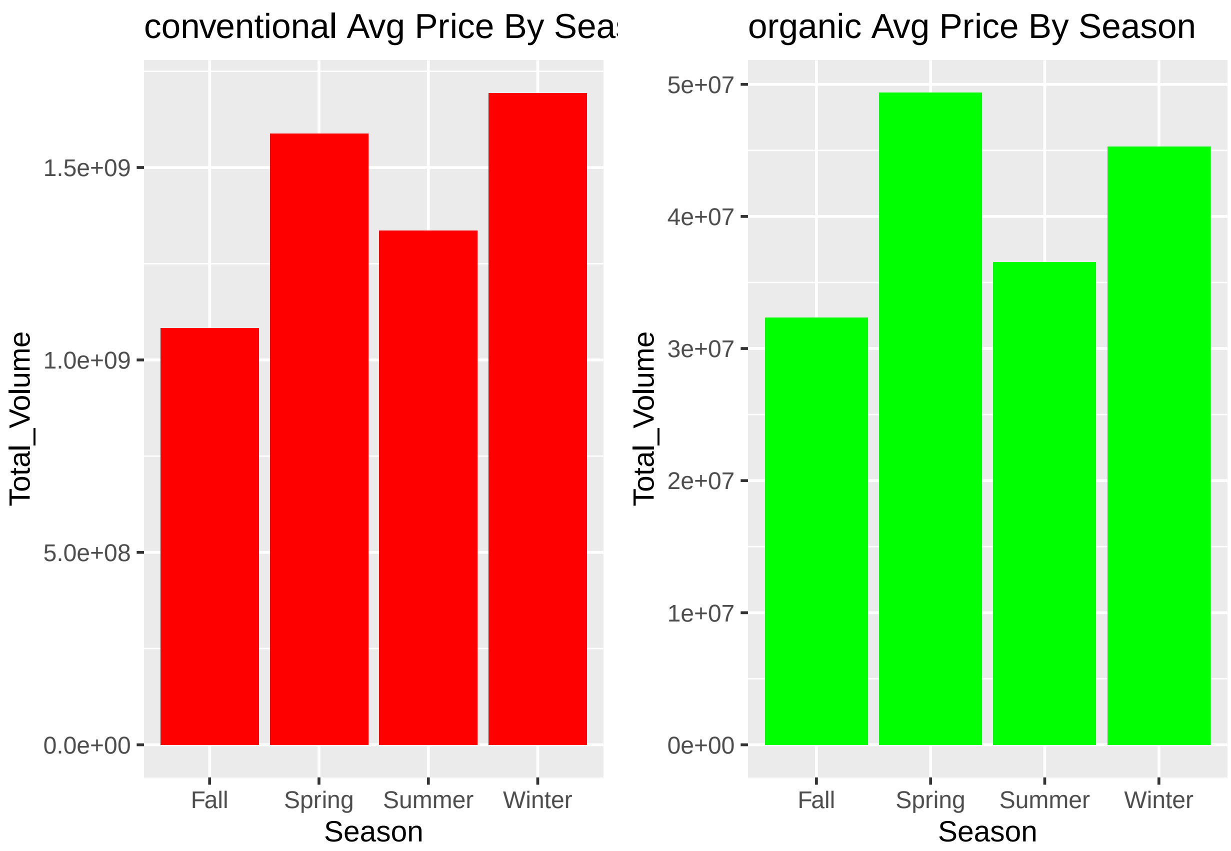
In this Section, I want to analysis Avocados Purchase by including Date, Season,Type and Region.

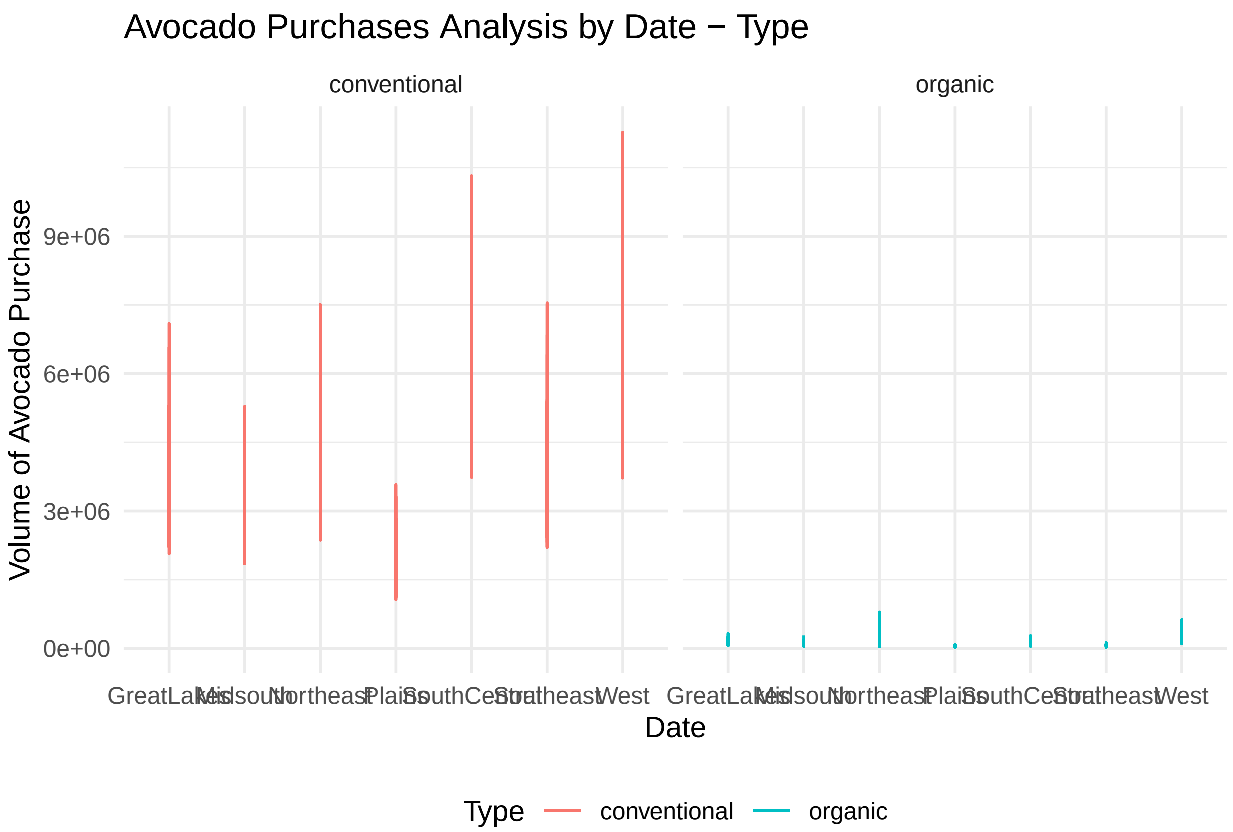
1. Avocados Purchase Volume by Date & Type
2. Avocados Purchase Volume by Date & Type with geom_line
3. Avocados Purchase Volume by Season

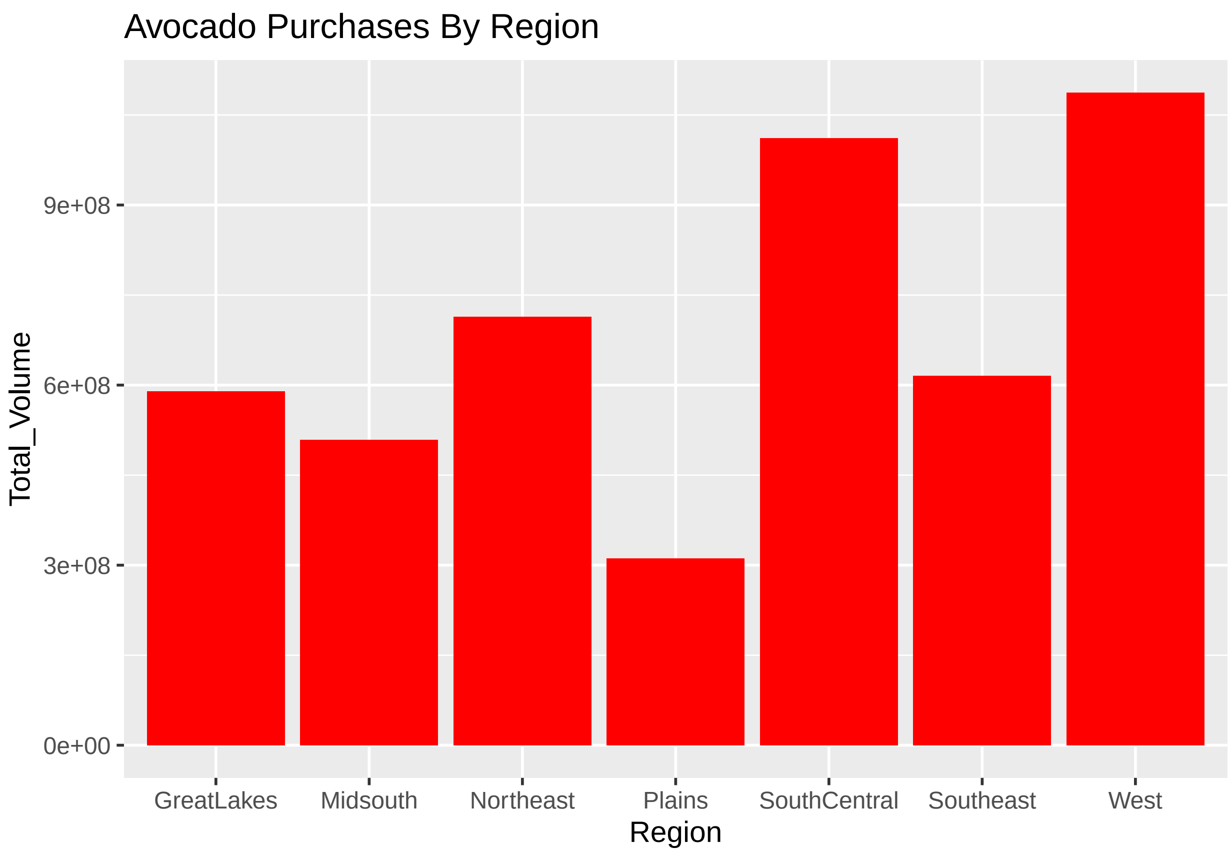


```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

Avocado Purchases Conventional Purc
Volume of Avocado Purchase
6e+07
4e+07
2e+07
0e+00
2015
2016
2017
2018
Date
Avocado Purchases Organic Purcha
Volume of Avocado Purchase
1500000
1000000
500000
0
2015
2016
2017
2018
Date


conventional Avg Price By Sea
Total_Volume
1.5e+09
1.0e+09
5.0e+08
0.0e+00
Fall
Spring
Summer
Winter
Season
organic Avg Price By Season
Total_Volume
5e+07
4e+07
3e+07
2e+07
1e+07
0e+00
Fall
Spring
Summer
Winter
Season


Avocado Purchases Analysis by Date − Type
conventional
organic
Volume of Avocado Purchase
9e+06
6e+06
3e+06
0e+00
GreatLakes
Midsouth
Northeast
Plains
SouthCentral
Southeast
West
Date
Type
conventional
organic



Summary :

```
1. conventional Avocados were more purchased than Organic Avocados
2. Spring Season was the highest volume of purchases in conventional Avocados
3. The purchase rate was more in Organic when compared to conventional
4. The minimum (3,424) and maximum (63,716,144) volume of avocados sold showed
5. 2015-2019 Year shows that customer purchase behaviour is consistent.
```

4. Linear Regression - Avocados Price significance

After the detailed analysis of Avocados Price, Volume Analysis, I am curious to understand the significance of Avocados price with other variables by considering the linear regression model.

```
## Linear Regression ##
TotalVolume_lm = lm(Average_Price~Total_Volume, data = avocado_dataset_usa)
summary(TotalVolume_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Total_Volume, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55212 -0.11937 -0.02607  0.11494  0.54443
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.560e+00  1.389e-02  112.35   <2e-16 ***
## Total_Volume -1.389e-08  5.726e-10  -24.27   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1784 on 336 degrees of freedom
## Multiple R-squared:  0.6367, Adjusted R-squared:  0.6356
## F-statistic: 588.8 on 1 and 336 DF,  p-value: < 2.2e-16
```

```
SmallHass_lm =lm(Average_Price~Small_Hass, data = avocado_dataset_usa)
summary(SmallHass_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Small_Hass, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54232 -0.11357 -0.00633  0.09941  0.54591
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.548e+00  1.383e-02  111.95   <2e-16 ***
## Small_Hass  -3.772e-08  1.595e-09  -23.65   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1813 on 336 degrees of freedom
## Multiple R-squared:  0.6246, Adjusted R-squared:  0.6235
## F-statistic: 559.2 on 1 and 336 DF,  p-value: < 2.2e-16
```

```
LargeHass_lm =lm(Average_Price~Large_Hass, data = avocado_dataset_usa)
summary(LargeHass_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Large_Hass, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55416 -0.09204 -0.02025  0.08177  0.53532
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.565e+00  1.296e-02   120.8   <2e-16 ***
## Large_Hass  -4.126e-08  1.546e-09   -26.7   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 0.1675 on 336 degrees of freedom
## Multiple R-squared:  0.6796, Adjusted R-squared:  0.6787 
## F-statistic: 712.8 on 1 and 336 DF,  p-value: < 2.2e-16
```

```
XLargeHass_lm =lm(Average_Price~XLarge_Hass, data = avocado_dataset_usa)
summary(XLargeHass_lm)
```

```
## 
## Call:
## lm(formula = Average_Price ~ XLarge_Hass, data = avocado_dataset_usa)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.51399 -0.09702 -0.01276  0.11406  0.57504 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.516e+00  1.286e-02  117.91   <2e-16 ***
## XLarge_Hass -4.260e-07  1.793e-08  -23.76   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.1807 on 336 degrees of freedom
## Multiple R-squared:  0.6269, Adjusted R-squared:  0.6258 
## F-statistic: 564.5 on 1 and 336 DF,  p-value: < 2.2e-16
```

```
TotalBags_lm =lm(Average_Price~Total_Bags, data = avocado_dataset_usa)
summary(TotalBags_lm)
```

```
## 
## Call:
## lm(formula = Average_Price ~ Total_Bags, data = avocado_dataset_usa)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.5137 -0.1779  0.0046  0.1561  0.5986 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.520e+00  1.654e-02   91.87   <2e-16 ***
## Total_Bags  -4.136e-08  2.404e-09  -17.20   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.2158 on 336 degrees of freedom
## Multiple R-squared:  0.4683, Adjusted R-squared:  0.4667 
## F-statistic:   296 on 1 and 336 DF,  p-value: < 2.2e-16
```

```
SmallBags_lm =lm(Average_Price~Small_Bags, data = avocado_dataset_usa)
summary(SmallBags_lm)
```

```
## 
```

```
## Call:
## lm(formula = Average_Price ~ Small_Bags, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52136 -0.16880  0.00224  0.15060  0.59764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.525e+00  1.631e-02   93.50   <2e-16 ***
## Small_Bags  -5.609e-08  3.138e-09  -17.87   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2119 on 336 degrees of freedom
## Multiple R-squared:  0.4873, Adjusted R-squared:  0.4858
## F-statistic: 319.4 on 1 and 336 DF,  p-value: < 2.2e-16
```

```r
LargeBags_lm =lm(Average_Price~Large_Bags, data = avocado_dataset_usa)
summary(LargeBags_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Large_Bags, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48256 -0.16598  0.01499  0.17260  0.61411
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.490e+00  1.727e-02   86.26   <2e-16 ***
## Large_Bags  -1.547e-07  1.067e-08  -14.50   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2321 on 336 degrees of freedom
## Multiple R-squared:  0.3849, Adjusted R-squared:  0.3831
## F-statistic: 210.3 on 1 and 336 DF,  p-value: < 2.2e-16
```

```r
XLargeBags_lm =lm(Average_Price~XLarge_Bags, data = avocado_dataset_usa)
summary(XLargeBags_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ XLarge_Bags, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48340 -0.26115  0.02664  0.19414  0.68665
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.403e+00  1.719e-02  81.628   <2e-16 ***
## XLarge_Bags -1.329e-06  1.471e-07  -9.034   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2654 on 336 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.193
## F-statistic: 81.61 on 1 and 336 DF,  p-value: < 2.2e-16
```

Based on the Linear Regression Analysis, I have found that all variables which I mentioned above had significance with Average Price. This result surprised me. I wanted to go another level to verify how the Type and Date will influence the price.

```
avocado_dataset_usa$Type = as.factor(avocado_dataset_usa$Type)
Type_lm =lm(Average_Price~Type, data = avocado_dataset_usa)
summary(Type_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Type, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54604 -0.11604 -0.02902  0.11247  0.55799
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.09201    0.01454   75.12   <2e-16 ***
## Typeorganic  0.45402    0.02056   22.08   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.189 on 336 degrees of freedom
## Multiple R-squared:  0.5921, Adjusted R-squared:  0.5909
## F-statistic: 487.7 on 1 and 336 DF,  p-value: < 2.2e-16
```

```
anova(Type_lm)
```

```
## Analysis of Variance Table
##
## Response: Average_Price
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Type        1 17.419 17.4186  487.71 < 2.2e-16 ***
## Residuals 336 12.000  0.0357
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Date_lm =lm(Average_Price~Date, data = avocado_dataset_usa)
summary(Date_lm)
```

```
##
## Call:
```

```
## lm(formula = Average_Price ~ Date, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59061 -0.22687 -0.01235  0.21838  0.68115
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.727e+00  7.717e-01  -3.534 0.000466 ***
## Date         2.750e-09  5.244e-10   5.245 2.78e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2845 on 336 degrees of freedom
## Multiple R-squared:  0.07567,    Adjusted R-squared:  0.07292
## F-statistic: 27.51 on 1 and 336 DF,  p-value: 2.778e-07
```

```
anova(Date_lm)
```

```
## Analysis of Variance Table
##
## Response: Average_Price
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Date        1  2.2261 2.22605  27.506 2.778e-07 ***
## Residuals 336 27.1929 0.08093
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Year_lm =lm(Average_Price~Year, data = avocado_dataset_usa)
summary(Year_lm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Year, data = avocado_dataset_usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66660 -0.24202 -0.01008  0.23644  0.66644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.257404   0.028001  44.905  < 2e-16 ***
## Year2016    0.006154   0.039600   0.155    0.877
## Year2017    0.179200   0.039412   4.547 7.63e-06 ***
## Year2018    0.049679   0.064666   0.768    0.443
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2856 on 334 degrees of freedom
## Multiple R-squared:  0.07423,    Adjusted R-squared:  0.06591
## F-statistic: 8.927 on 3 and 334 DF,  p-value: 1.051e-05
```

```
anova(Year_lm)
```

```
## Analysis of Variance Table
##
## Response: Average_Price
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Year        3  2.1837 0.72791  8.9267 1.051e-05 ***
## Residuals 334 27.2353 0.08154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we have seen in the previous section on Analysis, for a conventional avocado, the average price is 1.10 USD but the average price for an organic avocado is 1.55 USD which is 0.45 USD more.Hence the Type, Date also has significance with the price of Avocado.

As we have seen conventional avocado purchases were more than Organic, in fact Organic purchase was 2.1%. Also There is no demand for organic avocados any time of year. More avocados were purchased in the summer and warmer months than colder months which reflects in the linear model regression output. Hence Date and Season also had significance with the Price of Avocado.

```
avocado_dataset_region = subset(avocado_dataset, Region == "Great Lakes" | Region == "GreatLakes" | Reg
for(i in 1:nrow(avocado_dataset_region)){
  if(avocado_dataset_region$Region[i]=="Great Lakes"){
    avocado_dataset_region$Region[i]="GreatLakes"
  }else if(avocado_dataset_region$Region[i]=="South Central"){
    avocado_dataset_region$Region[i]="SouthCentral"
  }
}
Regionlm =lm(Average_Price~Region, data = avocado_dataset_region)
summary(Regionlm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Region, data = avocado_dataset_region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77802 -0.27124 -0.00349  0.24524  1.24778
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.33855    0.01700  78.729  < 2e-16 ***
## RegionMidsouth       0.06621    0.02404   2.754  0.00594 **
## RegionNortheast      0.26337    0.02404  10.954  < 2e-16 ***
## RegionPlains         0.09796    0.02404   4.074 4.77e-05 ***
## RegionSouthCentral -0.23731    0.02404  -9.869  < 2e-16 ***
## RegionSoutheast      0.05947    0.02404   2.473  0.01346 *
## RegionWest          -0.06633    0.02404  -2.759  0.00585 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3126 on 2359 degrees of freedom
## Multiple R-squared:  0.1731, Adjusted R-squared:  0.171
## F-statistic: 82.33 on 6 and 2359 DF,  p-value: < 2.2e-16
```

```r
anova(Regionlm)
```

```
## Analysis of Variance Table
##
## Response: Average_Price
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Region       6  48.264  8.0440  82.328 < 2.2e-16 ***
## Residuals 2359 230.489  0.0977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The only Analysis which I felt less was related to Region and I believe that Region makes difference with Avocado prices even though same Date/Season.

If I compare Northeast with Great Lake region, 0.23 USD more in Northeast but South central price was cheaper (0.20 USD) than Great Lake. P-Values for Southeast and West were higher than 0.05, Hence Southeast and West did not show significance.This shows that average prices of avocados would be higher in that region, but maybe due to warmth in West, avocados can be grown and sold cheaper. when we look at the anova table, the categorical variables show that Region is also significant.

5. Multiple Regression - Avocados Price significance

In the Previous Section, I ran the simple linear model for all variables with prices of Avocado. The last step is to run all the variables against the price of Avocado through Multiple Regression.

```r
## Multiple Regression ##
multiRegmodel = lm(Average_Price~Total_Volume+Type+Region, data = avocado_dataset_region)
multiRegmodel_output = step(multiRegmodel,direction="backward")
```

```
## Start:  AIC=-7595.91
## Average_Price ~ Total_Volume + Type + Region
##
##                Df Sum of Sq     RSS     AIC
## <none>                       94.718 -7595.9
## - Total_Volume  1     6.421 101.139 -7442.7
## - Type          1     7.030 101.749 -7428.5
## - Region        6    33.973 128.691 -6882.7
```

```r
summary(multiRegmodel_output)
```

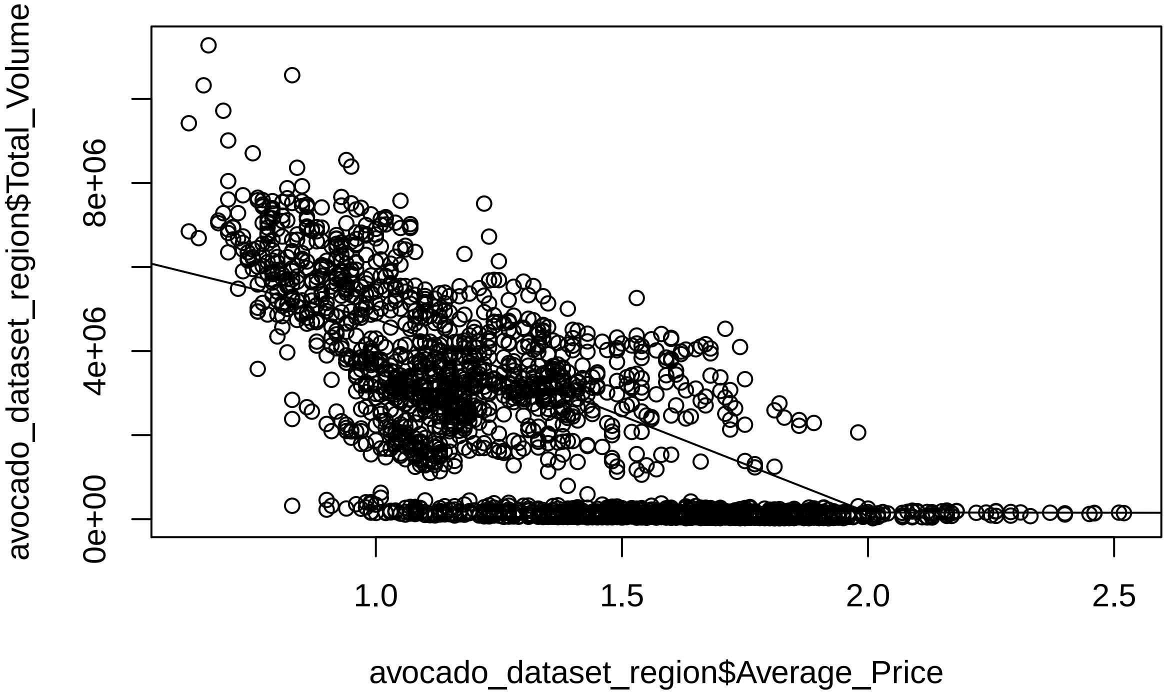
```
##
## Call:
## lm(formula = Average_Price ~ Total_Volume + Type + Region, data = avocado_dataset_region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71422 -0.13211 -0.01367  0.11420  0.94970
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.311e+00  2.004e-02  65.400  < 2e-16 ***
## Total_Volume       -5.614e-08  4.441e-09 -12.641  < 2e-16 ***
## Typeorganic         2.513e-01  1.900e-02  13.227  < 2e-16 ***
## RegionMidsouth      5.271e-02  1.546e-02   3.410  0.00066 ***
## RegionNortheast     2.839e-01  1.551e-02  18.310  < 2e-16 ***
## RegionPlains        5.171e-02  1.585e-02   3.263  0.00112 **
## RegionSouthCentral -1.673e-01  1.639e-02 -10.209  < 2e-16 ***
## RegionSoutheast     6.372e-02  1.542e-02   4.131 3.74e-05 ***
## RegionWest          1.623e-02  1.675e-02   0.969  0.33243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2005 on 2357 degrees of freedom
## Multiple R-squared:  0.6602, Adjusted R-squared:  0.6591
## F-statistic: 572.4 on 8 and 2357 DF,  p-value: < 2.2e-16
```

When we look at the Multiple R-squared was 0.6602 which means 66.02% variability with prices. Hence this model is a good fit one.

6. KNN Regression - Avocados Price significance

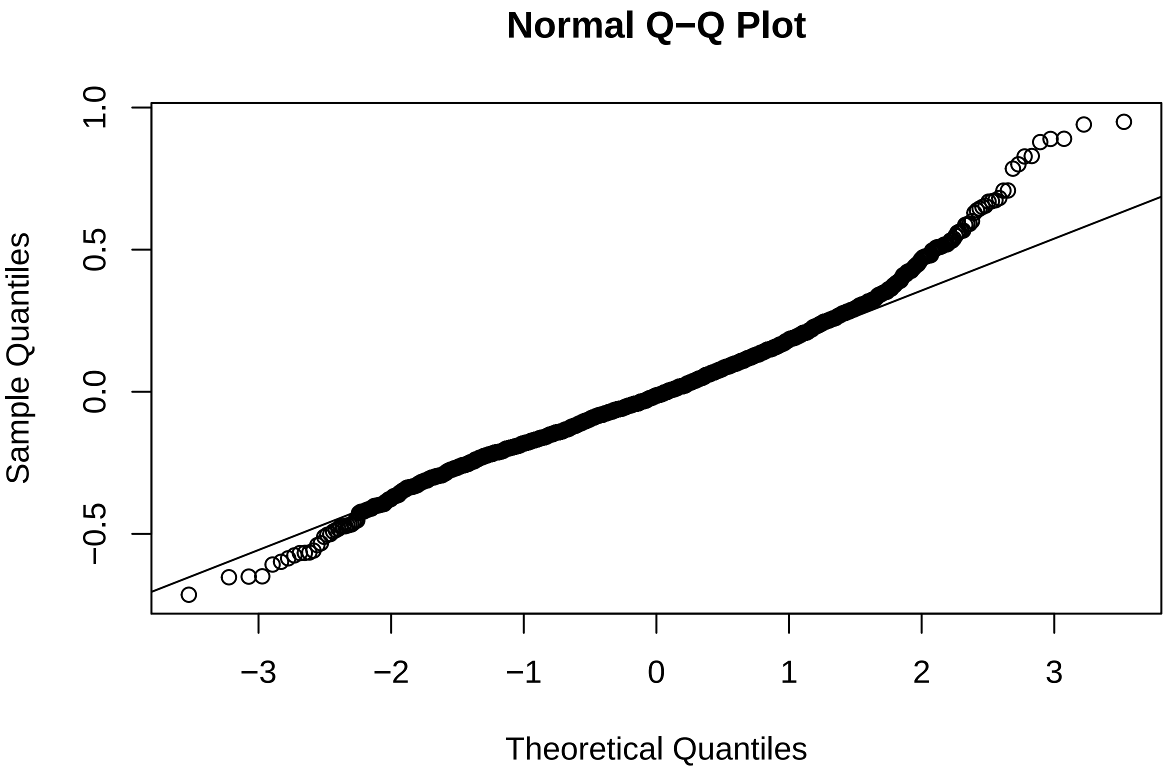
I wanted to try KNN Regression to see how Avocados Price significant with Volume. This Regression helps to undersand the nearest average price for certain Volume Groups.

```
## KNN Regression ##

#install.packages("FNN")
set.seed(1974)

knn3.avocado_dataset <- knn.reg(train=avocado_dataset_region[c("Average_Price")],
                      y=avocado_dataset_region$Total_Volume,
                      test= data.frame(Average_Price=seq(0,3)),
                      k=3)

plot(avocado_dataset_region$Average_Price, avocado_dataset_region$Total_Volume) #adding the scatter for
lines(seq(0,3), knn3.avocado_dataset$pred)
```



Verify the Model - qqnorm and qqline

```
qqnorm(resid(multiRegmodel_output))
qqline(resid(multiRegmodel_output))
```



I wanted to verify the model was good fit by using qqnorm and qqline method. We can see the plots which again confirms that the model is a good fit.

Conclusion - Summary of Analysis

I had seen that there was a linear relationship between Total Volume and Average Price. Also other entities like Date,Region, etc. But we can not increase the price for Revenue growth, since other variables impact the price determination.

As per Law of economics, supply and demand determines the price but in our case, we had seen there were more supply and less demand, This is one of the study points to handle the price fix of Avocado.

This project empowered me to study the entire Avocado business model along with R Project technical expertise and this analysis gave me the insight about Sales Analytics which we can extend to other food items or entities.

We can leverage this project analysis further to find the Competition around Avocado and Avocado products and increase the predictive model. Also we can extend this Analysis to entire world (my project focus only USA)

Finally I wanted to conclude that Organic Avocado is costlier than conventional and consistently price was increasing when Fall starts. So the best time for Avacodo purchase is before Fall Season.

References:

1. R for Everyone by by Jared P. Lander, Pearson Education, 2017. 2nd Edition

2. R for Data Science by Hadley Wickham, Garrett Grolemund, O'Reilly Publisher(2016). - ISBN: 9781491910399 https://r4ds.had.co.nz/

3. Avocado Prices by Justin Kiggins, Kaggle https://www.kaggle.com/neuromusic/avocado-prices

4. Avocado Market Research by R-Bloggers https://www.r-bloggers.com/2018/09/avocado-market-research/

5. Millennials' Favorite Fruit: Forecasting Avocado Prices with ARIMA Models by Sean Holland, Medium https://towardsdatascience.com/millennials-favorite-fruit-forecasting-avocado-prices-with-arima-models-5b46e4e0e914

6. Avocado Prices: Pattern Recognition Analysis by Janio Martinez, Kaggle https://www.kaggle.com/janiobachmann/price-of-avocados-pattern-recognition-analysis

7. Predicting avocado prices_Kaggle dataset by Joan Claverol, Rpubs https://rpubs.com/JoanClaverol/532659

8. MLR_Avacado by Abhishek, Rpubs https://rpubs.com/abhiisinghh/423833

9. What is Exploratory Data Analysis? by Prasad Patil, Medium https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

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