Avocado Price Growth

Trish Girmus

August 7, 2020

Introduction

When I began this project, I learned that the popularity of avocados has increased over the past two decades. Consumption per capita in the 90's was 1.6 lbs. From 2014-16 alone that increased to 7.1 lbs. per capita. As avocados have now been deemed a "superfood", I was curious to see the consumer behavior with purchase. If there is enough demand for this health food, can price be increased? I also wanted to see if there was correlation between the variables. The data set I worked with is from 2015-2019 and is the brand Hass avocados. This data contains purchases made in the U.S. only.

Load Data File & Install Packages

```
## Load Data, Read File & Install Packages ##
setwd("C:/Users/Owner/Documents/R Class/DSC520")
avocado<-read.csv("Avocado_Prices_Data.csv", header=TRUE)
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tibble)
library(pander)
```

Clean Data

I began this process by taking the data set and cleaning it. While observing this information, I noticed that there were many overlapping geographic locations where avocados were purchased, and some regions contained the same name but listed differently. I subset the Total U.S. response under the Region variable and removed the rest of the regions for my initial analysis. This was to prevent skewness with the data and to create a smaller sample set to work with. I also changed the formatting of dates or purchase as well as changing Type of avocado and Region variables from characters to factors. Once this was completed, my data was ready for further analysis.

```
## Clean Data ##
names(avocado)[2]<-"Average_Price"
names(avocado)[3]<-"Total_Volume"
names(avocado)[4]<-"Small_Hass"
names(avocado)[5]<-"Large_Hass"
names(avocado)[6]<-"XLarge_Hass"
names(avocado)[7]<-"Total_Bags"
names(avocado)[8]<-"Small_Bags"
names(avocado)[9]<-"Large_Bags"
names(avocado)[10]<-"XLarge_Bags"
names(avocado)[11]<-"Type"
names(avocado)[12]<-"Year"
names(avocado)[13]<-"Region"
names(avocado)
```

```
##  [1] "Date"          "Average_Price" "Total_Volume"  "Small_Hass"
##  [5] "Large_Hass"    "XLarge_Hass"   "Total_Bags"    "Small_Bags"
##  [9] "Large_Bags"    "XLarge_Bags"   "Type"          "Year"
## [13] "Region"
```

```
avocado[1]=lubridate::parse_date_time(avocado$Date, "%m/%d/%Y")

avocado$Region = as.factor(avocado$Region)
levels(avocado$Region)
```

```
##  [1] "Albany"               "Atlanta"              "Baltimore/Washington"
##  [4] "BaltimoreWashington"  "Boise"                "Boston"
##  [7] "Buffalo/Rochester"    "BuffaloRochester"     "California"
## [10] "Charlotte"            "Chicago"              "Cincinnati/Dayton"
## [13] "CincinnatiDayton"     "Columbus"             "Dallas/Ft. Worth"
## [16] "DallasFtWorth"        "Denver"               "Detroit"
## [19] "Grand Rapids"         "GrandRapids"          "Great Lakes"
## [22] "GreatLakes"           "Harrisburg/Scranton"  "HarrisburgScranton"
## [25] "Hartford/Springfield" "HartfordSpringfield"  "Houston"
## [28] "Indianapolis"         "Jacksonville"         "Las Vegas"
## [31] "LasVegas"             "Los Angeles"          "LosAngeles"
## [34] "Louisville"           "Miami/Ft. Lauderdale" "MiamiFtLauderdale"
## [37] "Midsouth"             "Nashville"            "New Orleans/Mobile"
## [40] "New York"             "NewOrleansMobile"     "NewYork"
## [43] "Northeast"            "Northern New England" "NorthernNewEngland"
## [46] "Orlando"              "Philadelphia"         "Phoenix/Tucson"
## [49] "PhoenixTucson"        "Pittsburgh"           "Plains"
## [52] "Portland"             "Raleigh/Greensboro"   "RaleighGreensboro"
## [55] "Richmond/Norfolk"     "RichmondNorfolk"      "Roanoke"
## [58] "Sacramento"           "San Diego"            "San Francisco"
## [61] "SanDiego"             "SanFrancisco"         "Seattle"
## [64] "South Carolina"       "South Central"        "SouthCarolina"
## [67] "SouthCentral"         "Southeast"            "Spokane"
## [70] "St. Louis"            "StLouis"              "Syracuse"
## [73] "Tampa"                "Total U.S."           "TotalUS"
## [76] "West"                 "West Tex/New Mexico"  "WestTexNewMexico"
```

```r
avocado.us=subset(avocado, Region == "Total U.S." | Region == "TotalUS")
levels(avocado$Region) = c(levels(avocado$Region)[1:74], levels(avocado$Region)[74], levels(avocado$Reg
avocado.us = avocado.us[order(avocado.us$Date),]
```

Analysis

```r
## Data Overview ##
str(avocado.us)
```

```
## 'data.frame':    506 obs. of  13 variables:
##  $ 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         : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ Region       : Factor w/ 78 levels "Albany","Atlanta",..: 75 75 75 75 75 75 75 75 75 75 ...
```

I started my analysis by running the structure of the data frame. After cleaning the data, it shows I have 506 observations and 13 variables.

```r
## Data Overview #
summary(avocado.us)
```

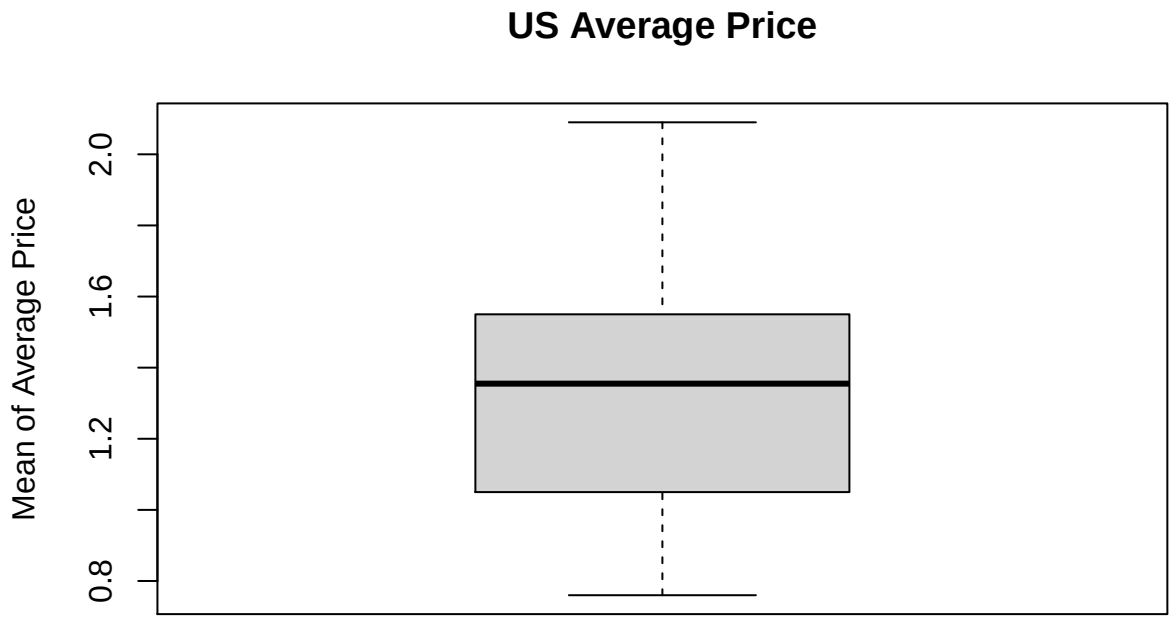
```
##       Date                      Average_Price    Total_Volume
##  Min.   :2015-01-04 00:00:00   Min.   :0.760   Min.   :    3424
##  1st Qu.:2016-03-20 00:00:00   1st Qu.:1.050   1st Qu.: 1120420
##  Median :2017-06-04 00:00:00   Median :1.355   Median : 1835222
##  Mean   :2017-06-09 07:52:24   Mean   :1.331   Mean   :18436425
##  3rd Qu.:2018-08-19 00:00:00   3rd Qu.:1.550   3rd Qu.:36109420
##  Max.   :2019-12-01 00:00:00   Max.   :2.090   Max.   :63716144
##
##    Small_Hass         Large_Hass        XLarge_Hass        Total_Bags
##  Min.   :      23   Min.   :      95   Min.   :      0   Min.   :    3186
##  1st Qu.:  131866   1st Qu.:  274288   1st Qu.:   4146   1st Qu.:  687229
##  Median :  306430   Median :  476348   Median :  18916   Median : 1315223
##  Mean   : 6069724   Mean   : 5778196   Mean   : 444895   Mean   : 6141724
##  3rd Qu.:11940905   3rd Qu.:11249276   3rd Qu.: 816441   3rd Qu.:11105739
##  Max.   :22743616   Max.   :20470573   Max.   :2546439   Max.   :23472989
##
##    Small_Bags         Large_Bags       XLarge_Bags          Type
##  Min.   :    2148   Min.   :      0   Min.   :     0.0   Length:506
##  1st Qu.:  485965   1st Qu.: 160891   1st Qu.:     0.0   Class :character
##  Median : 1119886   Median : 384647   Median :   497.5   Mode  :character
```

```
##  Mean   : 4383689   Mean   :1652892   Mean   :105143.6
##  3rd Qu.: 8489435   3rd Qu.:2748346   3rd Qu.:163579.1
##  Max.   :15436247   Max.   :8378356   Max.   :844929.8
##
##       Year                        Region
##  Min.   :2015   Total U.S.          :402
##  1st Qu.:2016   TotalUS             :104
##  Median :2017   Albany              :  0
##  Mean   :2017   Atlanta             :  0
##  3rd Qu.:2018   Baltimore/Washington:  0
##  Max.   :2019   BaltimoreWashington :  0
##                 (Other)             :  0
```

```
summary(avocado.us$Average_Price)
```

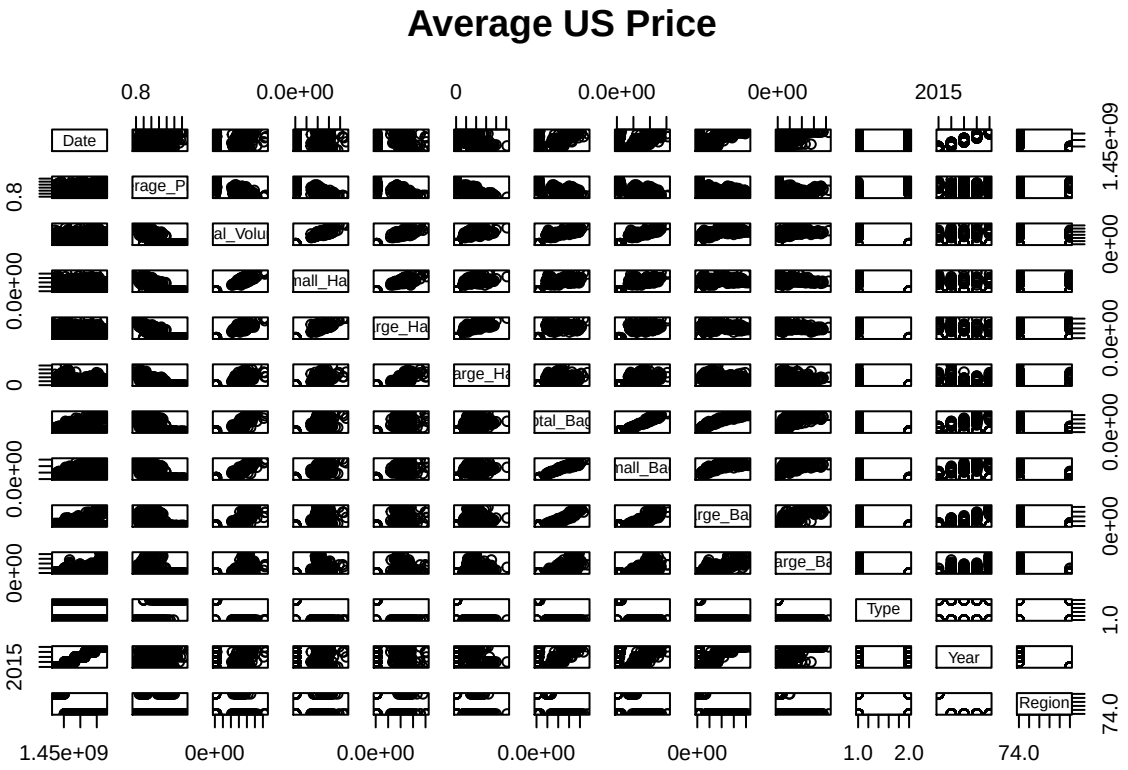
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.760   1.050   1.355   1.331   1.550   2.090
```

```
boxplot(avocado.us$Average_Price, ylab = "Mean of Average Price", main = "US Average Price")
```

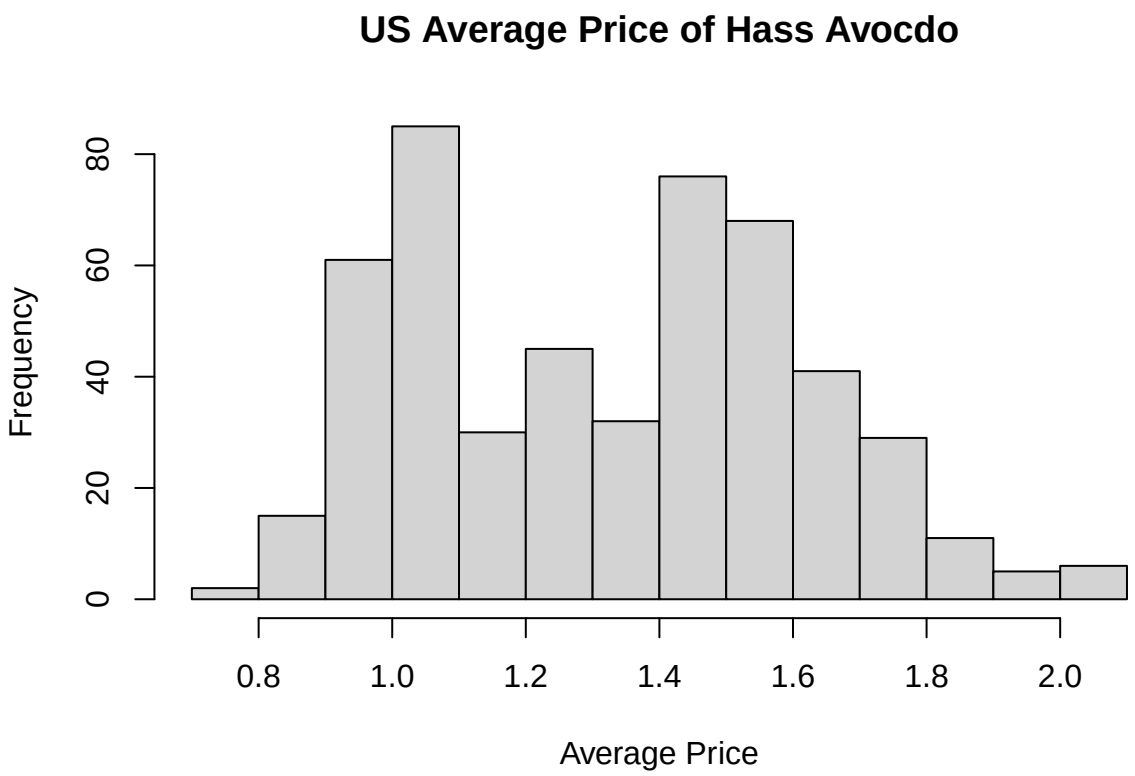


After looking at the structure of the data frame, I ran a summary of the data to see what the average price of an avocado was. It is $1.33. I then plotted a box plot to see what this looked like visually.

```
plot(avocado.us, main = "Average US Price")
```



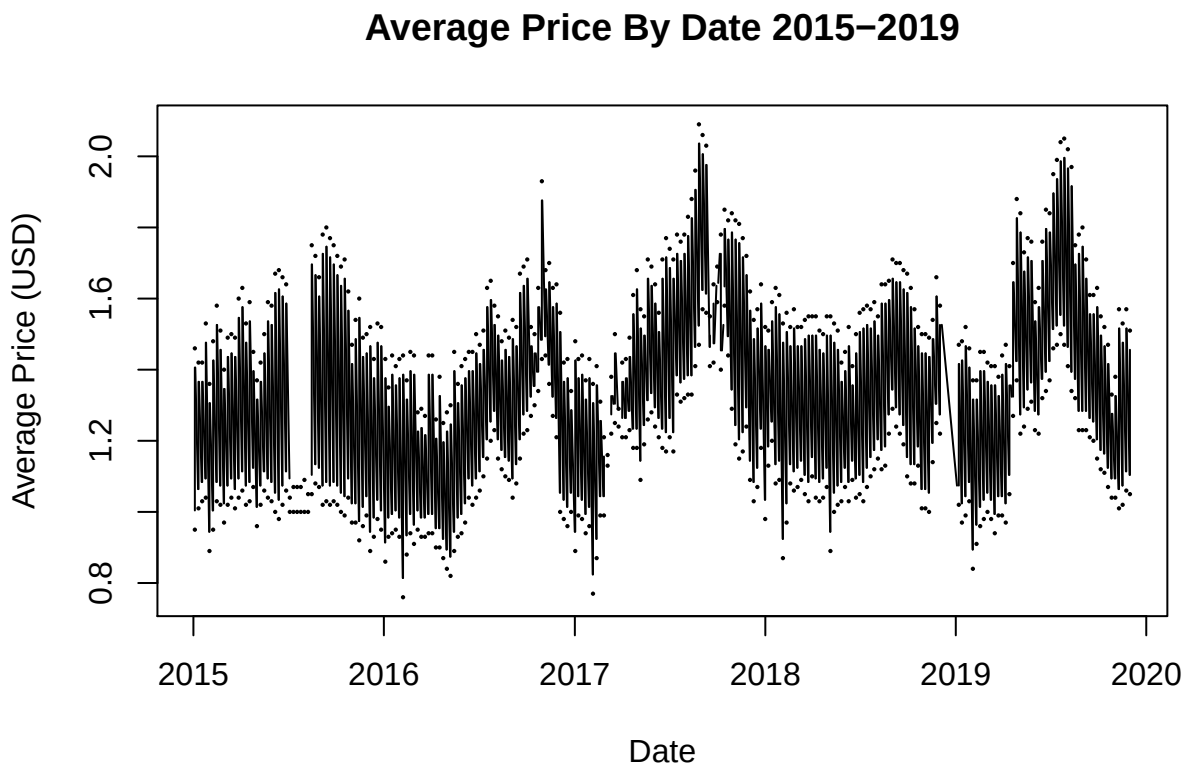

```
hist(avocado.us$Average_Price, xlab = "Average Price", main = "US Average Price of Hass Avocdo")
```



I also plotted a histogram with Average Price. It appears to have a normal distribution. When I ran the histogram prior to cleaning my data, it was showed skewness which alerted me that something was wrong with my data. I also studied the minimum ($0.76) and maximum ($2.09) price. I also looked at Total Volume to see what the minimum (3,424) and maximum (63,716,144) volume of avocados sold showed. I would say we as Americans enjoy eating avocados! Something I discovered while analyzing this information is noticing that by analyzing individual size of avocado sold versus total bags and total volume would cause error and overlapping with my strategy, so I decided to include Date, Average Price, Total Volume, Type and Region. I will go into further detail about this in another section.

Next, I plotted another graph to show the Average Price purchased by Date to see if there was a trend in seasonality with relation to price.

```
plot(avocado.us$Date, avocado.us$Average_Price, type = "b", pch = 16,cex = .3,
     xlab = "Date", ylab = "Average Price (USD)", main = "Average Price By Date 2015-2019")
```



As I suspected and as the graph shows, average price is lower during the winter months, but price slowly increases around spring through summer months where it then peaks. From fall back into winter months the price slowly declines. This buying behavior shows to be consistent throughout the five years of data collected. This story tells me a couple different things. One, when weather is colder the demand is less, and therefore consumers are not buying avocados. When the weather is warm, the average price increases when there is more demand. Weather does affect avocado price!

Linear Regression

My next steps were running simple linear regression models on the individual variables to see if they showed significance with Average Price. Part of this step determined that I could not use the individual size and total bags of avocados as it would affect Total Volume of avocados sold. This is because Total Volume is its own variable.

```
## Linear Regression ##
Total_Volumelm = lm(Average_Price~Total_Volume, data = avocado.us)
summary(Total_Volumelm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Total_Volume, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55925 -0.11239 -0.00590  0.09135  0.53678
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.567e+00  1.064e-02  147.25   <2e-16 ***
## Total_Volume -1.277e-08  4.102e-10  -31.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1683 on 504 degrees of freedom
## Multiple R-squared:  0.658,  Adjusted R-squared:  0.6574
## F-statistic: 969.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Small_Hasslm =lm(Average_Price~Small_Hass, data = avocado.us)
summary(Small_Hasslm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Small_Hass, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55407 -0.10313 -0.00434  0.09597  0.53416
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.560e+00  1.015e-02   153.7   <2e-16 ***
## Small_Hass  -3.773e-08  1.164e-09   -32.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1639 on 504 degrees of freedom
## Multiple R-squared:  0.6757, Adjusted R-squared:  0.675
## F-statistic:  1050 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Large_Hasslm =lm(Average_Price~Large_Hass, data = avocado.us)
summary(Large_Hasslm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Large_Hass, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56024 -0.08942 -0.01367  0.08066  0.52924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.571e+00  9.764e-03  160.91   <2e-16 ***
## Large_Hass  -4.155e-08  1.191e-09  -34.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1558 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.7071, Adjusted R-squared:  0.7065
## F-statistic:  1217 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
XLarge_Hasslm =lm(Average_Price~XLarge_Hass, data = avocado.us)
summary(XLarge_Hasslm)
```

```
##
## Call:
## lm(formula = Average_Price ~ XLarge_Hass, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52617 -0.10628 -0.01148  0.10463  0.56283
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.528e+00  1.014e-02  150.77   <2e-16 ***
## XLarge_Hass -4.429e-07  1.484e-08  -29.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.173 on 504 degrees of freedom
## Multiple R-squared:  0.6387, Adjusted R-squared:  0.638
## F-statistic: 891.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Total_Bagslm =lm(Average_Price~Total_Bags, data = avocado.us)
summary(Total_Bagslm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Total_Bags, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51302 -0.11127  0.02011  0.12046  0.59334
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.517e+00  1.302e-02  116.54   <2e-16 ***
## Total_Bags  -3.031e-08  1.465e-09  -20.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2117 on 504 degrees of freedom
## Multiple R-squared:  0.4593, Adjusted R-squared:  0.4583
## F-statistic: 428.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Small_Bagslm =lm(Average_Price~Small_Bags, data = avocado.us)
summary(Small_Bagslm)
```

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

```
## lm(formula = Average_Price ~ Small_Bags, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53387 -0.11239  0.02016  0.11378  0.58045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.537e+00  1.266e-02  121.41   <2e-16 ***
## Small_Bags  -4.703e-08  2.045e-09  -22.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2011 on 504 degrees of freedom
## Multiple R-squared:  0.512,  Adjusted R-squared:  0.511
## F-statistic: 528.7 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
Large_Bagslm =lm(Average_Price~Large_Bags, data = avocado.us)
summary(Large_Bagslm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Large_Bags, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48359 -0.13425  0.00518  0.14783  0.63412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.463e+00  1.357e-02  107.85   <2e-16 ***
## Large_Bags  -7.991e-08  5.168e-09  -15.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2371 on 504 degrees of freedom
## Multiple R-squared:  0.3218, Adjusted R-squared:  0.3204
## F-statistic: 239.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
XLarge_Bagslm =lm(Average_Price~XLarge_Bags, data = avocado.us)
summary(XLarge_Bagslm)
```

```
##
## Call:
## lm(formula = Average_Price ~ XLarge_Bags, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55564 -0.19078  0.02479  0.16494  0.68479
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.405e+00  1.379e-02   101.9   <2e-16 ***
```

```
## XLarge_Bags -7.050e-07  6.981e-08   -10.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2625 on 504 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.1666
## F-statistic:   102 on 1 and 504 DF,  p-value: < 2.2e-16
```

Since this was my first analysis/projectproject of this sort, I quickly discovered after running the simple linear regression that all variables had significance with Average Price and did not make sense to use this with my prediction. I am including these examples merely to show that the variables all showed significance.

I did review my model on Type which backs up my theory on a graph I plotted before I ran the linear regression models. I originally had this information in another section of my analysis but decided to include it in this section because of topic relevancy.

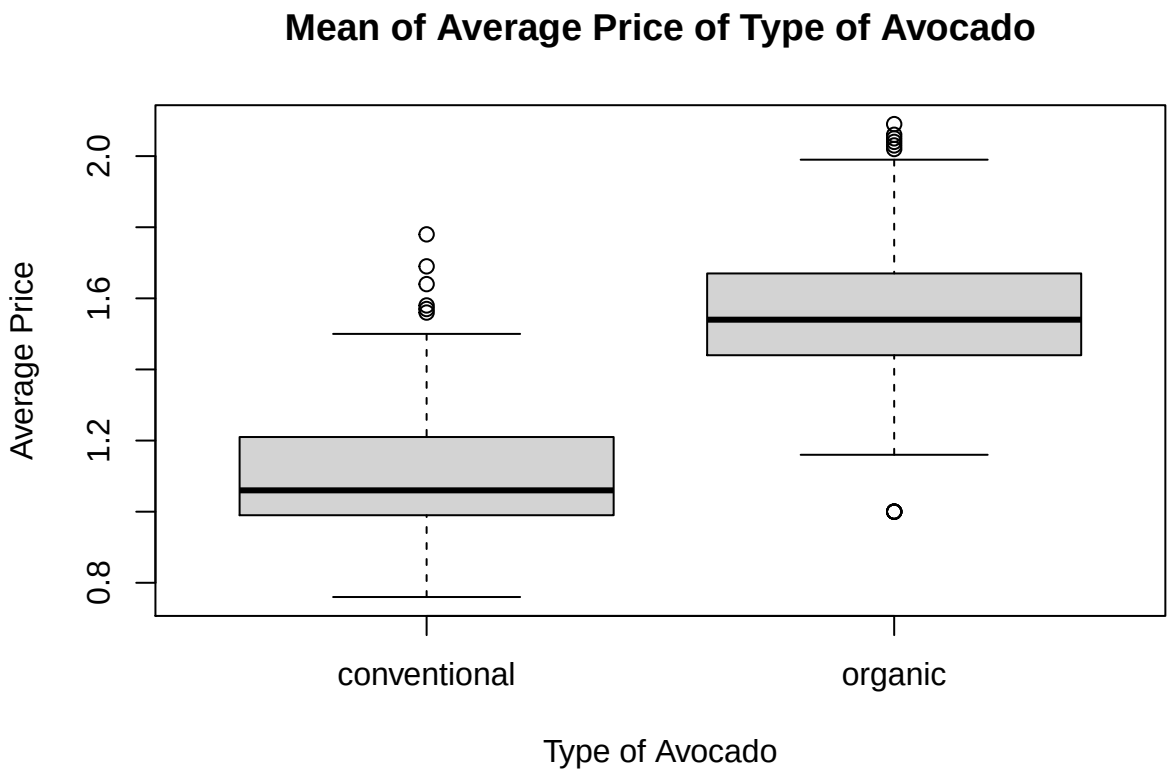
```
avocado.us$Type = as.factor(avocado.us$Type)
Typelm =lm(Average_Price~Type, data = avocado.us)
summary(Typelm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Type, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55731 -0.11486 -0.03486  0.10514  0.67514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.10486    0.01116   99.03   <2e-16 ***
## Typeorganic  0.45245    0.01578   28.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1775 on 504 degrees of freedom
## Multiple R-squared:   0.62,  Adjusted R-squared:  0.6193
## F-statistic: 822.4 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
anova(Typelm)
```

```
## Analysis of Variance Table
##
## Response: Average_Price
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Type        1 25.896 25.8960  822.36 < 2.2e-16 ***
## Residuals 504 15.871  0.0315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

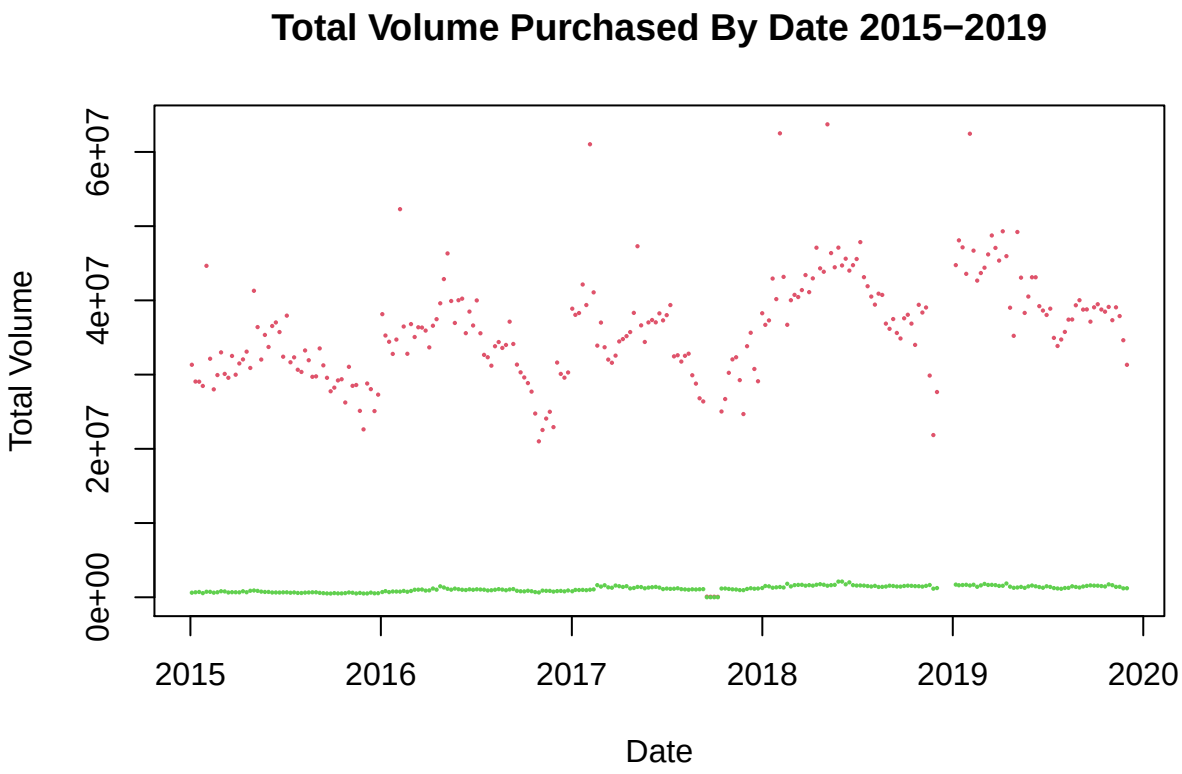
```
plot(avocado.us$Type, avocado.us$Average_Price, ylab = "Average Price", xlab = "Type of Avocado", main =
```



As you see, for a conventional avocado, the average price is $1.10, and shows that the average price for an organic avocado goes up $0.45 higher, or $1.55 when adding the two together.

A graph I plotted earlier in my analysis as discussed above was regarding Type of avocado (in this case conventional, which is red and organic, which is green) and Total Volume Purchased by Date. I wanted to see what type of avocados are purchased and when. I was also curious to see what the trend looks like.

```
plot(avocado.us$Date, avocado.us$Total_Volume, pch = 16,cex = .3,
     xlab = "Date", ylab = "Total Volume", main = "Total Volume Purchased By Date 2015-2019",col = (avo
```



As you can see, the total volume of organic avocado purchases is low. There is also a lower variance with organic volume. Conventional total volume is much higher, and you can also see the dips with time of year and seasonality when purchases are made. This graph tells me a couple of things as well. One, organic avocado purchase/volume is flat throughout the five-year span. I presume this is due to price as organic food is typically more expensive than conventional food. There is also no demand for organic avocados any time of year. Secondly, looking at the data of conventional volume purchased, more volume is sold overall with this type of avocado. Total volume sales occur in the summer and warmer months than they do during the winter months. It also does appear that total volume in 2019 was the highest overall five years. There is an increase in conventional volume sold from 2015 to 2019.

```
Datelm =lm(Average_Price~Date, data = avocado.us)
summary(Datelm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Date, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55955 -0.25102 -0.00476  0.21202  0.74999
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.340e-01  4.186e-01  -1.515     0.13
## Date         1.313e-09  2.795e-10   4.697 3.41e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2818 on 504 degrees of freedom
## Multiple R-squared:  0.04194,    Adjusted R-squared:  0.04004
## F-statistic: 22.06 on 1 and 504 DF,  p-value: 3.409e-06
```

```
Yearlm =lm(Average_Price~Year, data = avocado.us)
summary(Yearlm)
```

```
##
## Call:
## lm(formula = Average_Price ~ Year, data = avocado.us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56254 -0.24849 -0.00254  0.21421  0.75746
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -66.033369  18.104497  -3.647 0.000293 ***
## Year          0.033399   0.008976   3.721 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.284 on 504 degrees of freedom
## Multiple R-squared:  0.02674,    Adjusted R-squared:  0.0248
## F-statistic: 13.84 on 1 and 504 DF,  p-value: 0.0002209
```

I also decided when running the simple linear regression to include regions back into this type analysis, for this part only. It did not make sense to use Total U.S. for this model. I originally wanted to determine what affect Average Price had with Region. As this was somewhat complicated initially due to duplicates names and regions overlapping themselves, I included the responses of Region from the data with Great Lakes, Mid South, Northeast, Plains, South Central, Southeast, and West for my model. This model gave me some good information about Average Price and Region. As you will see, when using Great Lakes at the intercept, most other regions have a higher average price.

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

```
##
## Call:
## lm(formula = Average_Price ~ Region, data = avocado.region)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74415 -0.26251 -0.01993  0.22350  1.58585
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.32650    0.01389  95.530  < 2e-16 ***
## RegionMidsouth      0.07601    0.01964   3.871 0.000111 ***
## RegionNortheast     0.23190    0.01964  11.809  < 2e-16 ***
## RegionPlains        0.06648    0.01964   3.385 0.000718 ***
## RegionSouthCentral -0.20211    0.01964 -10.292  < 2e-16 ***
## RegionSoutheast     0.03765    0.01964   1.917 0.055296 .
## RegionWest          0.00496    0.01964   0.253 0.800589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3124 on 3535 degrees of freedom
## Multiple R-squared:  0.1276, Adjusted R-squared:  0.1261
## F-statistic: 86.15 on 6 and 3535 DF,  p-value: < 2.2e-16
```

```
anova(Regionlm)
```

```
## Analysis of Variance Table
##
## Response: Average_Price
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Region       6  50.43  8.4055  86.154 < 2.2e-16 ***
## Residuals 3535 344.89  0.0976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For example, the Region Northeast's average price is $1.56 compared to Great Lake's average price of $1.33. Region South Central's price is even cheaper at $1.13. The Coefficients Mid South, Northeast, Plains and South Central all showed significance. The Coefficients Southeast and West did not show significance as the P-Values for those respectively were higher than 0.05. It was also surprising to see that there was no difference between Average Price with Great Lakes and the West Regions. I would automatically assume that average prices of avocados would be higher in that region, but maybe due to weather avocados can be grown and sold cheaper in the West. Also, by running an anova table (since these are categorical variables) it shows Region is significant. So, for my original question I can conclude that Region makes a difference with Average Price!

Multiple Regression

Once I ran and analyzed the models I built from simple linear regression, I decided to build a model with multiple regression. As I mentioned earlier, I chose the variables Total Volume, Type and Region along with Average Price. I learned when running my models for linear regression that looking at individual size of avocado and total bags sold would skew the data as I only want to see what volume was sold.

```
## Multiple Regression ##
full.model = lm(Average_Price~Total_Volume+Type+Region, data = avocado.region)
reduced.model = step(full.model,direction="backward")
```

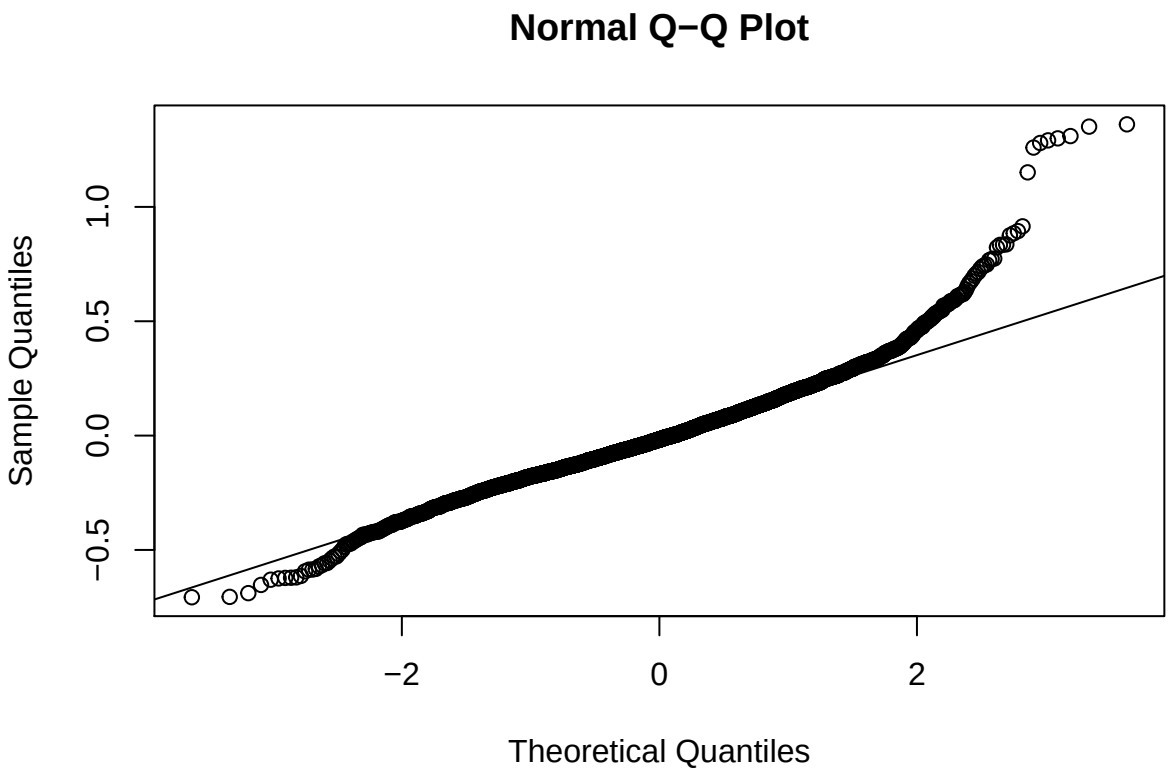
```
## Start:  AIC=-11155.96
## Average_Price ~ Total_Volume + Type + Region
##
##                Df Sum of Sq    RSS    AIC
## <none>                      151.07 -11156
## - Type          1     7.042 158.11 -10997
## - Total_Volume  1    14.570 165.64 -10832
## - Region        6    40.028 191.10 -10335
```

```
summary(reduced.model)
```

```
##
## Call:
## lm(formula = Average_Price ~ Total_Volume + Type + Region, data = avocado.region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70677 -0.13062 -0.01718  0.11276  1.36090
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.341e+00  1.629e-02  82.349  < 2e-16 ***
## Total_Volume        -6.132e-08  3.322e-09 -18.459  < 2e-16 ***
## Typeorganic          1.971e-01  1.536e-02  12.833  < 2e-16 ***
## RegionMidsouth       6.341e-02  1.302e-02   4.871 1.16e-06 ***
## RegionNortheast      2.610e-01  1.310e-02  19.927  < 2e-16 ***
## RegionPlains         1.297e-02  1.332e-02   0.974     0.33
## RegionSouthCentral -1.190e-01  1.376e-02  -8.646  < 2e-16 ***
## RegionSoutheast      5.083e-02  1.302e-02   3.904 9.65e-05 ***
## RegionWest           9.635e-02  1.391e-02   6.926 5.12e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2068 on 3533 degrees of freedom
## Multiple R-squared:  0.6179, Adjusted R-squared:  0.617
## F-statistic:   714 on 8 and 3533 DF,  p-value: < 2.2e-16
```

This model showed significance with all Coefficients except for the Region Plains. When looking at R-squared, it shows that with a 62% variability in the Average Price, it is explained by this model which means this is a good model.

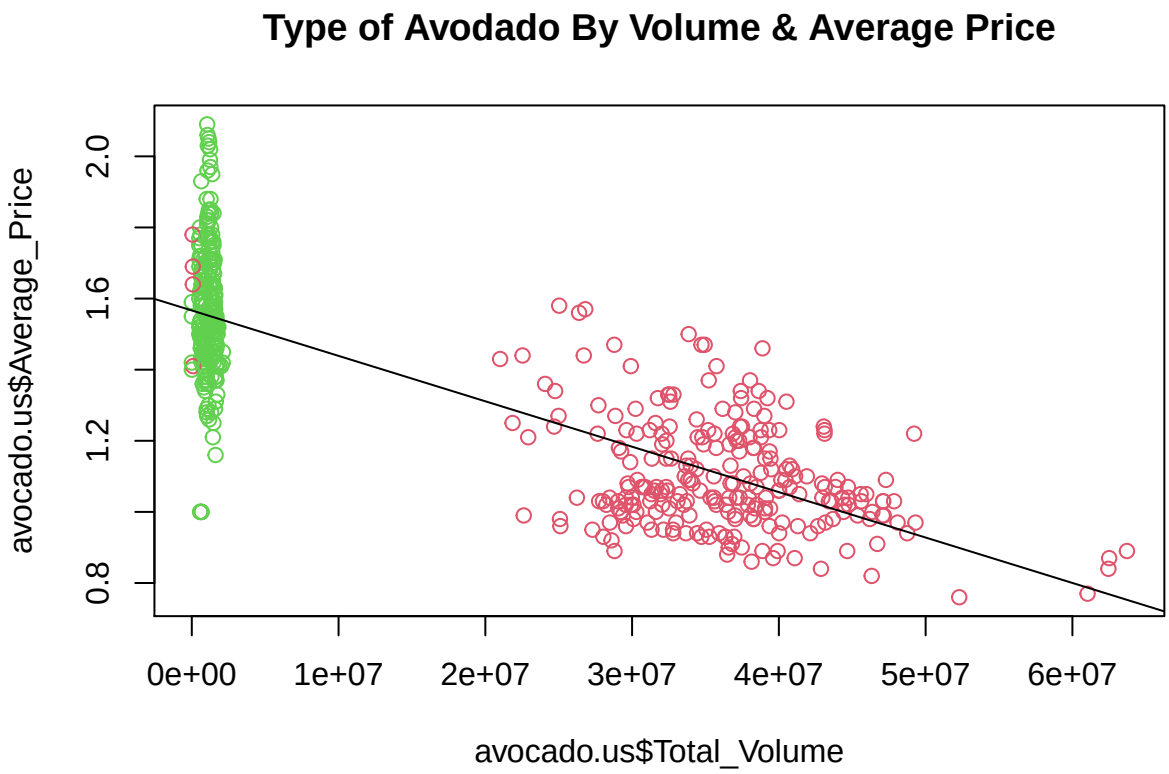
```
qqnorm(resid(reduced.model))
qqline(resid(reduced.model))
```



I also tested this by plotting a qqnorm and qqline to see if this model is adequate. The plot confirms this was a good model.

I also made another plot to show the Total Volume and Average Price based on Avocado Type.

```
plot(avocado.us$Total_Volume, avocado.us$Average_Price, col = (avocado.us$Type=="organic")+2, main = "Ty
abline(lm(Average_Price~Total_Volume,data = avocado.us))
```



The total volume is less when buying organic and shows that a higher price is paid. For conventional avocados, more volume is sold at a lower average price. This also shows that there is a linear relationship between volume and price. As total volume goes up, average price goes down.

Conclusion from Analysis

Based on this model, I have determined that although there is a linear relationship between Total Volume and Average Price, I cannot increase price to grow revenue. This conclusion also results from the rest of my analysis. The law of economics explains supply and demand. In this case, there is too much supply and not enough demand. I am disappointed that I cannot increase price. I also determined that I might have been looking at too many aspects of data to answer my initial question, which may or may not have addressed my question at all. I was eager to slice and dice the data and see what I could draw conclusions from. I also learned through this project as much as I love to get into the weeds with analysis, this research could go on for days, possibly longer depending on what kind of information is analyzed. Something else I learned from this project is that sometimes you might be asking the wrong questions. The data can tell the story you're looking for, but you may have to shift the strategy in order to find the solution you're looking for.

As I was optimistic to increase price, I put together a prediction model based on Total Volume, Type of avocado and Region. The Region does include the handful of geographical locations I looked at in an earlier section and not Total U.S. When looking at the summary of data to start, I looked at minimum total volume sold and picked the Type of avocado (conventional or organic) and a Region to predict price for a single prediction. I also included a confidence interval which gives the mean of the observation. I used the lowest total volume sold to show a comparison of conventional and organic avocados in the West region. Once again, it shows that organic average price is higher than conventional.

Prediction Model for Conventional Avocado

```
##        fit     lwr     upr
## 1 1.437465 1.03006 1.84487
```

```
##        fit      lwr      upr
## 1 1.437465 1.397378 1.477551
```

The predicted & confidence value is $1.44 for a conventional type of avocado

Prediction Model for Organic Avocado

```
##        fit     lwr      upr
## 1 1.634561 1.22859 2.040532
```

```
##        fit      lwr     upr
## 1 1.634561 1.613581 1.65554
```

The predicted & confidence value is $1.63 for an organic type of avocado

This was a fun project and really enjoyed it. It was nice to apply our skills from what we have learned this semester and be able to provide the "why" behind the "how". Thank you to everyone for a great semester!

Sources:

https://hassavocadoboard.com/

https://hassavocadoboard.com/wp-content/uploads/2019/03/hab-latest-independent-economic-evaluation-2018.pdf