NORTH AMERICAN UNIVERSITY
COMPUTER SCIENCE DEPARTMENT
CLASS: COMP 4353 DATA MINING

TOPIC: AVOCADO PRICES AND SALES IN US MARKET

BY: HUSNBONU AVGHONOVA

ABSTRACT

AVOCADO SALES AND PRICES IN US MARKET

Nowadays avocado toast became the main hit so everyone is obsessed with it, especially millennials. No wonder people are consuming up to 6 million avocados per week! The last 5 years the demand for avocado. Increased tremendously and I don't think it's going to decrease anytime soon.

The dataset that used for this project is provided by Hass Avocado Board (HAB).

Hass Avocado Board (HAB) is a trade organization that provides data on avocados sales and prices around the world for research. For this project I used RStudio as my tool. The data set contains information about two types of avocados: conventional and organic, information about different prices and the average price, the amount of different type of avocados sold, the total volume of sold avocados by regions, and last information about different dates, and year of all these sales and demands.

1.1 DATASET

In this project I chose a dataset from Kaggle where it talks about price sales of Hass avocados from January 2015 till March 2018 and it has approximately 18,300 entries. The following dataset can be found in this website: https://www.kaggle.com/neuromusic/avocado-prices. This data represents weekly retail scan which is about 170 weeks. For this period the dataset provides with average price, total volume for conventional and organic avocados in 54 regions of the United States. Now let's see how the data was used to gain information and how the graphs were modeled.

First, as I mentioned I used RStudio for this project so I imported the data into RStudio and to make sure that data is the right one I checked the top only:

```
>  avocados_csv<-import("~/Desktop/avocado.csv")
> head(avocados_csv)
  V1       Date AveragePrice Total Volume    4046      4225   4770 Total Bags Small Bags Large Bags XLarge Bags
1  0 2015-12-27         1.33     64236.62 1036.74  54454.85  48.16    8696.87    8603.62      93.25           0
2  1 2015-12-20         1.35     54876.98  674.28  44638.81  58.33    9505.56    9408.07      97.49           0
3  2 2015-12-13         0.93    118220.22  794.70 109149.67 130.50    8145.35    8042.21     103.14           0
4  3 2015-12-06         1.08     78992.15 1132.00  71976.41  72.58    5811.16    5677.40     133.76           0
5  4 2015-11-29         1.28     51039.60  941.48  43838.39  75.78    6183.95    5986.26     197.69           0
6  5 2015-11-22         1.26     55979.78 1184.27  48067.99  43.61    6683.91    6556.47     127.44           0
         type year region
1 conventional 2015 Albany
2 conventional 2015 Albany
3 conventional 2015 Albany
4 conventional 2015 Albany
5 conventional 2015 Albany
6 conventional 2015 Albany
> avocados<-avocados_csv[  c(1:4, 8:14)]
> head(avocados)
  V1       Date AveragePrice Total Volume Total Bags Small Bags Large Bags XLarge Bags         type year region
1  0 2015-12-27         1.33     64236.62    8696.87    8603.62      93.25           0 conventional 2015 Albany
2  1 2015-12-20         1.35     54876.98    9505.56    9408.07      97.49           0 conventional 2015 Albany
3  2 2015-12-13         0.93    118220.22    8145.35    8042.21     103.14           0 conventional 2015 Albany
4  3 2015-12-06         1.08     78992.15    5811.16    5677.40     133.76           0 conventional 2015 Albany
5  4 2015-11-29         1.28     51039.60    6183.95    5986.26     197.69           0 conventional 2015 Albany
6  5 2015-11-22         1.26     55979.78    6683.91    6556.47     127.44           0 conventional 2015 Albany
```

(Figure 1)

After importing and checking my data I wanted to rename it so it would be easier for further research (Figure 1). Before going right into modeling, I wanted first to make sure that my data is clean and there's no missing data.

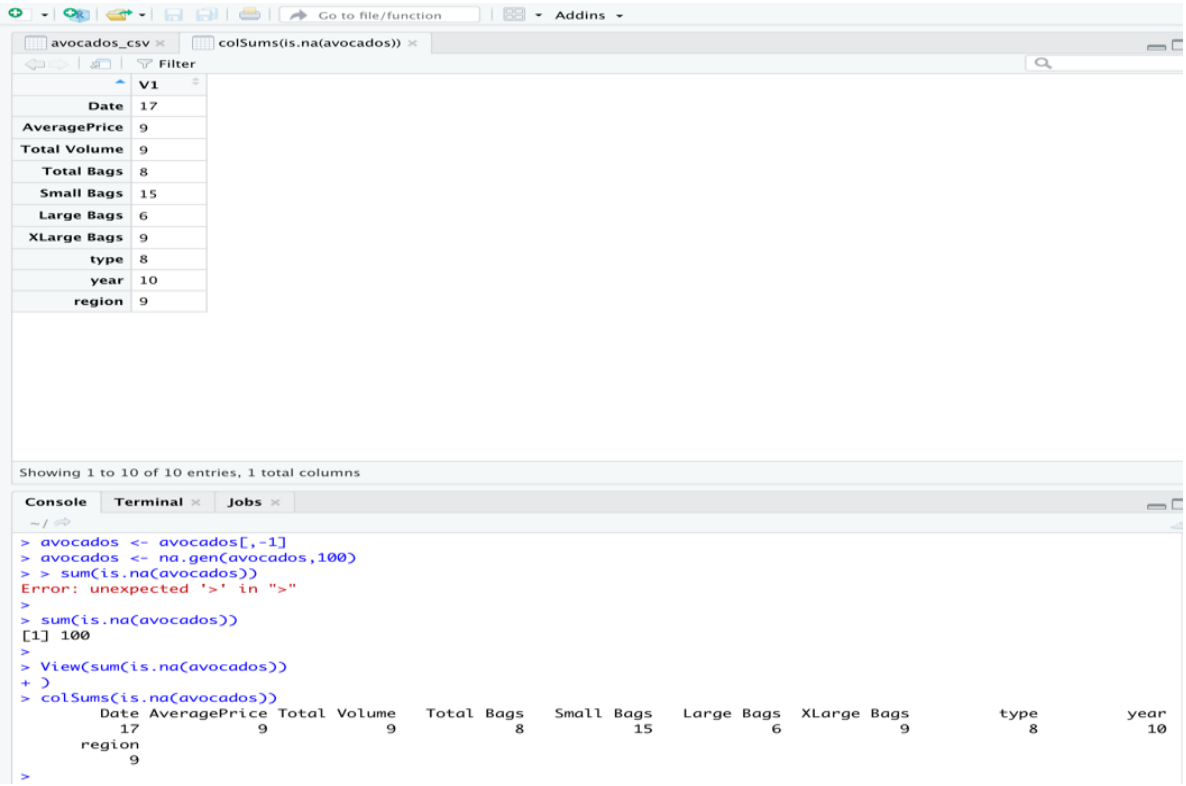
```
> sum(is.na(avocados))
[1] 0
>
```

As a summary it didn't show any missing data's so I just wanted to take little further step and checked for detailed data. Just for my information and to make sure my graphs are as clean as possible. In the figure below (Figure 2) I called for NA (not applicable) generator, so it will generate each data to find the missing one.

```
> na.gen <- function(data,n) {
+     i <- 1
+     while (i < n+1) {
+         idx1 <- sample(1:nrow(data), 1)
+         idx2 <- sample(1:ncol(data), 1)
+         data[idx1,idx2] <- NA #missing data (NotApplicable)
+         i = i+1
+     }
+     return(data)
+ }
```

Figure 2

Now after generating my dataset we can see the missing data. We have 100 missing data and they are all in different categories. We have 17 missing data in date category, 15 missing data in. small bag categories and etc.



After retrieving our missing data, I just simply cleaned it by omitting my avocado and rechecked the data again to make sure my data got clean.

```
> missingdata <- avocados[!complete.cases(avocados), ]
>
> sum(is.na(missingdata))
[1] 100
>
> clean.dt<- na.omit(avocados)
> sum(is.na(clean.dt))
[1] 0
```

2.1 PREPROCESSING

After importing the data and cleaning the missing information, the next step is preprocessing. In my data I have three numerical categories and names with space between which might become an issue in the future so that why I changed their names. First, I selected the categories that doesn't need changes and are currently irrelevant like X1 (the number of rows divided), categories with word "Bags", year and region. The names that were renamed were:

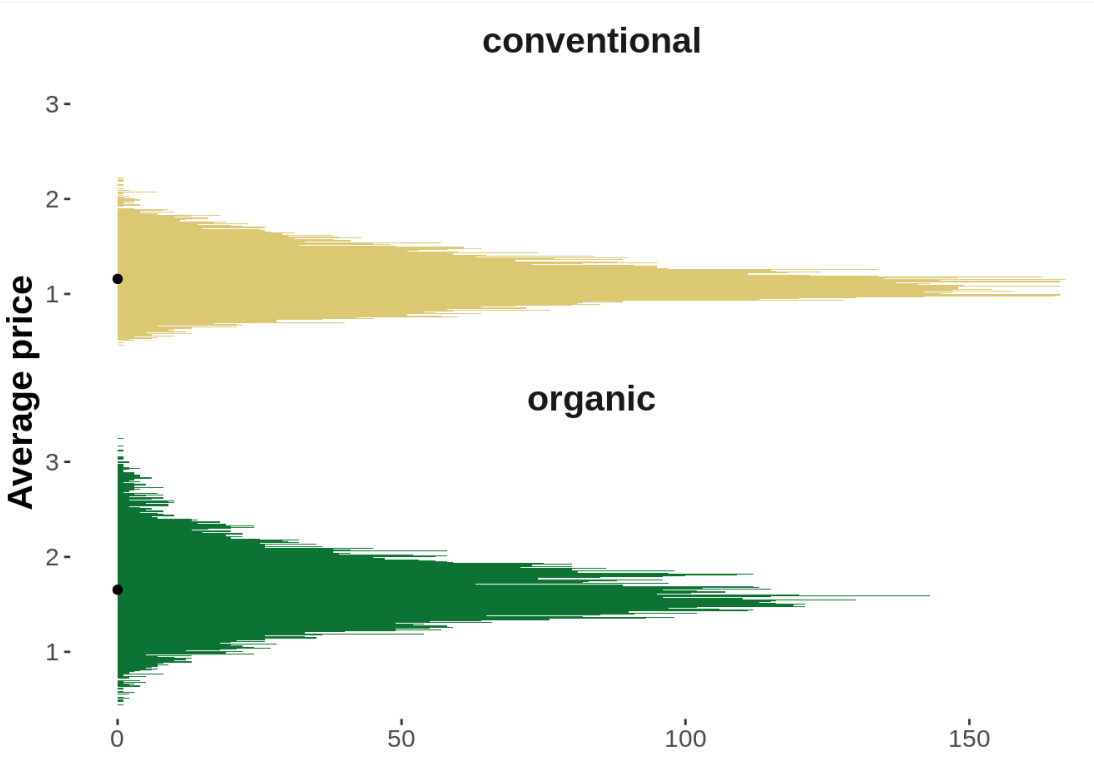
total_value_sold= “Total Volume”, average_price= “Average Price”, small= ‘4046’, medium=’4225’, and large = ‘4770’. These numbers were the PLUs so, we just changed them into easier and understandable format. After preprocessing the data by commanding **glimpse,** I checked my data to see if names were renamed and it worked, they all were changed.

```
Observations: 18,249
Variables: 7
$ Date              <date> 2015-12-27, 2015-12-20, 2015-12-13, 2015-12-06, 201…
$ average_price     <dbl> 1.33, 1.35, 0.93, 1.08, 1.28, 1.26, 0.99, 0.98, 1.02…
$ total_volume_sold <dbl> 64236.62, 54876.98, 118220.22, 78992.15, 51039.60, 5…
$ small             <dbl> 1036.74, 674.28, 794.70, 1132.00, 941.48, 1184.27, 1…
$ medium            <dbl> 54454.85, 44638.81, 109149.67, 71976.41, 43838.39, 4…
$ large             <dbl> 48.16, 58.33, 130.50, 72.58, 75.78, 43.61, 93.26, 80…
$ type              <chr> "conventional", "conventional", "conventional", "con…
>
```

2.2 EXPLORATORY ANALYSIS

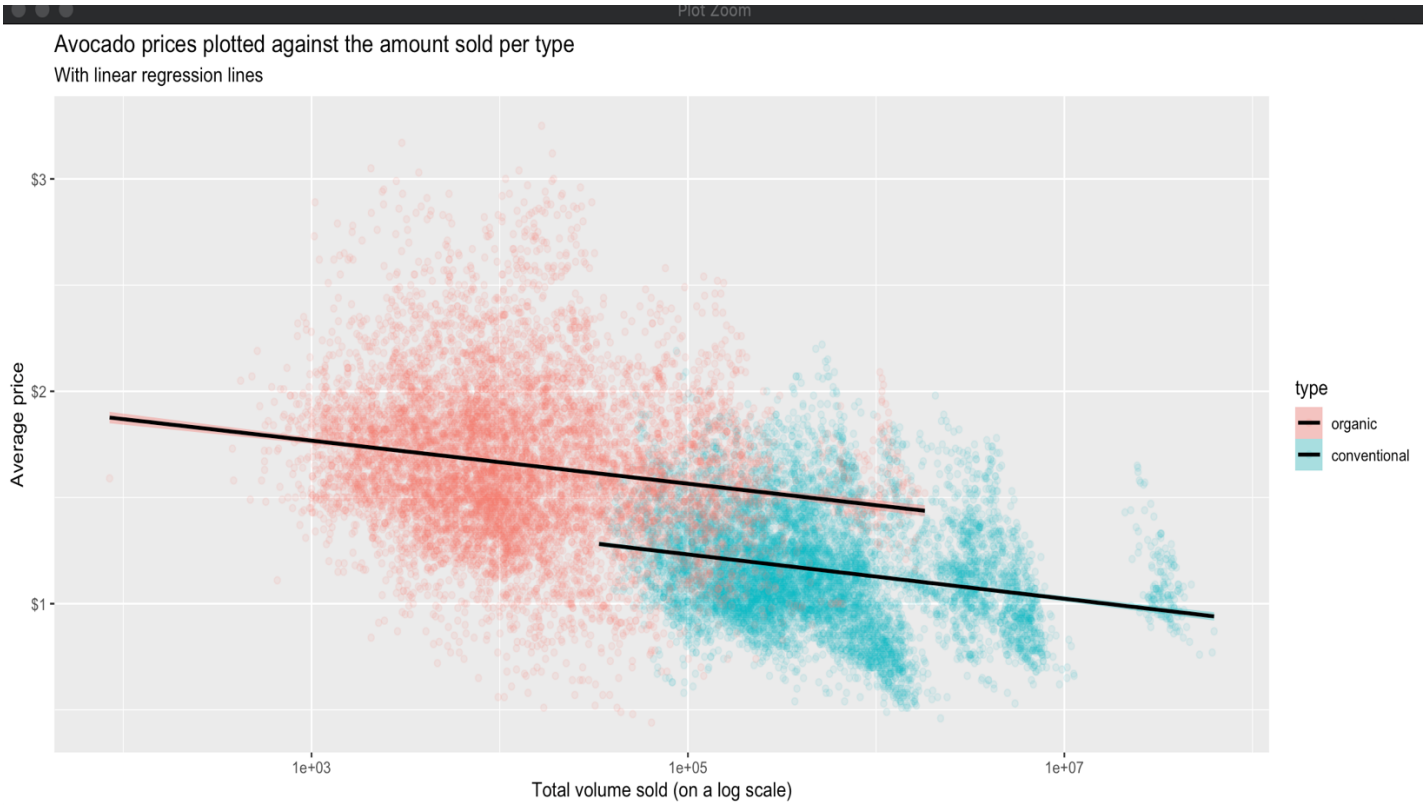
Now that the data has been imported, cleaned from missing information and preprocessed, the next step which was modeling has been applied. These models and graphs revealed a very interesting observation. Here is how it looked like.

Just by looking at this plot (figure 1) it can be concluded that organic avocados are higher on average price and more volatile than conventional avocados.



(Figure 1: Average price of conventional and organic avocados.)

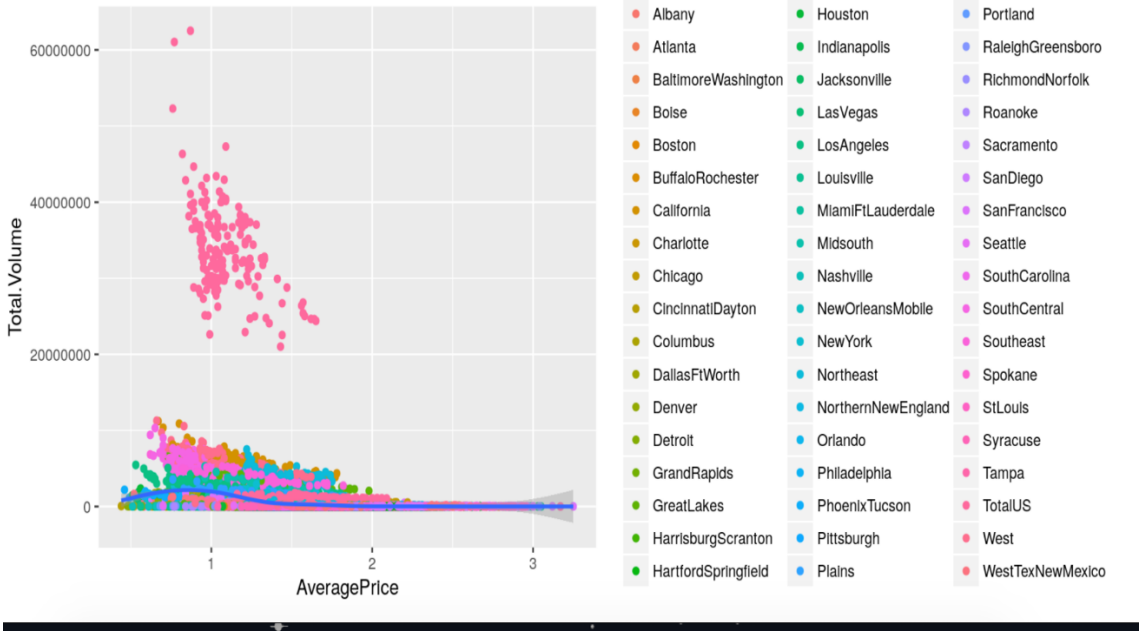
According to my statistic summary the average price for avocados in general is 1.40$, for conventional avocados the average price is 1.15$ and standard deviation is 0.26$ and for organic avocados average price is 1.65$ and standard deviation is 0.36$.



(Figure 2: Avocado prices/amount sold per type)

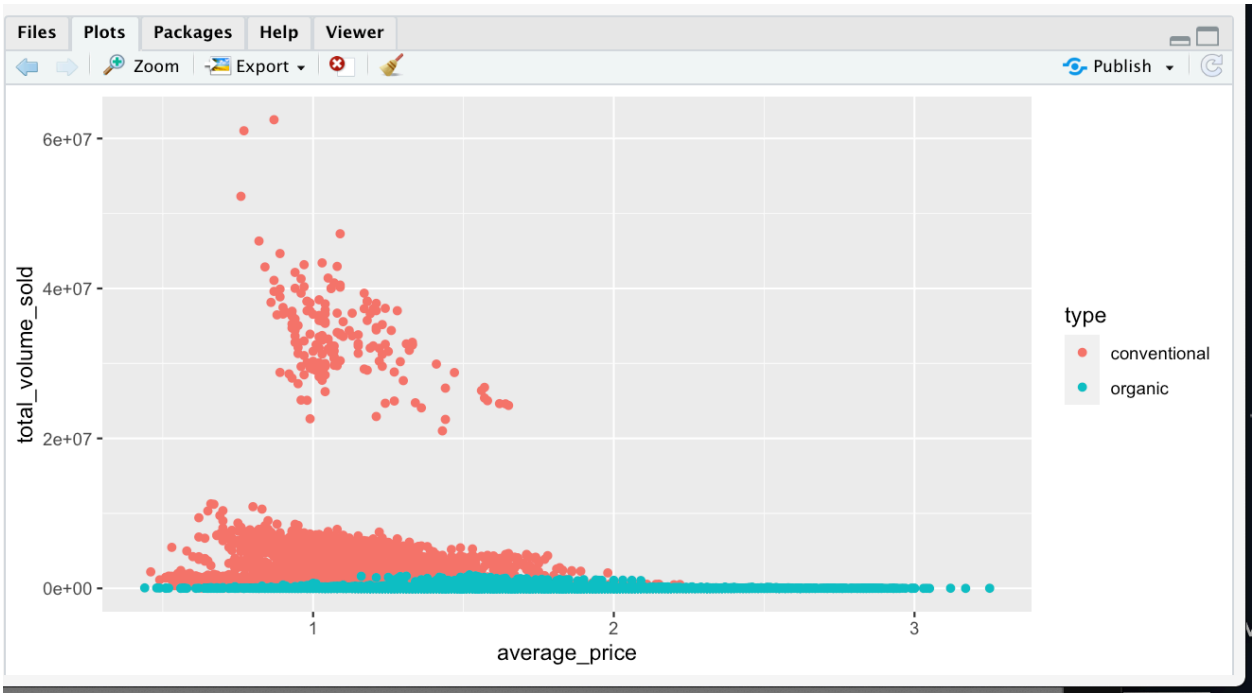
As we can see in the plot above (figure 2) conventional avocados demonstrated a moderate negative correlation between average price and volume sold, which means nothing special happened. Just a basic economic rule where price goes down demand goes up. But organic avocados' demand is flat and scattered all over the plot, that means no matter how the price changes the consumers are not planning to increase their demand.

In figure 3 the plot is shown the total volume sold by region. It becomes obvious that Northeasterners and people around Washington pay the highest rates for organic and conventional avocados in the country.



(figure 3: comparison by regions)

They pay around 1.34$ for conventional avocados and 1.86$ for organic avocados, even though the national average price is 1.09$ for conventional and 1.55$ for organic.



(figure 4: volume comparison)

In terms of volume we can see that Western regions consume the most avocados in the country, both conventional and organic ones (figure 4). As we can see they consumed about ~ six million conventional avocados and 220,000 organic avocados per week!

Finally, I just played around and wanted to see the plot using logarithm and it look. Pretty all around the field scattered and line drastically changed.