Aabha Desai
Isha Khasgiwala
Christopher Kupovics
Neetika Saxena
Puneet Sidhu

Project Report 2 - FA2020 Group 1
BUAN 6356.002

1. All the hypotheses clearly spelled out.

1. Avocado pricing shows more correlation to season rather than region.
2. Over time, consumers across all regions become less concerned with fluctuations in avocado pricing.
3. There is an increase in Avocado prices during May through October when we have seasonal California wildfires.

2. All the EDA (I am assuming you will be done with most of them at this point.

The main components of exploring data which we covered:

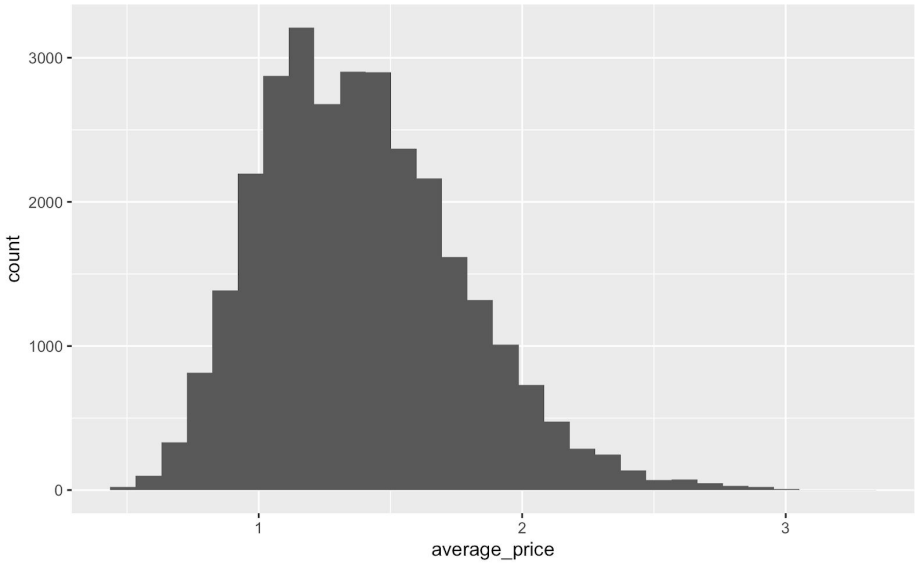
1. Understanding Variables - To formulate the assumption and hypothesis of our modelling we first tried to understand our data and import/install the relevant packages required to run the successful prediction model. Using the Snagit package the output shows that we have around 30021 observations and 13 variables. No Null values but some of the data type should be changed. So in our preliminary processing we changed the data types of the Date variable by setting date 'Date' as "ymd"(av$date = ymd(av$date) format and changing it from character to 'Date' data type.
2. Finding Missing Values and Duplicates - While finding missing and duplicate values, we noticed that the sum of missing and duplicate values in the data is 0. So the functions sum(duplicated(av)) and sum(is.na(av)) gave 0 records.
3. Understanding Numerical and Categorical Variables - During further analysis, we focused on understanding the categorical variable and their unique values with corresponding counts. In the avacado_updated_2020 file, the variable "type" indicated the organic and conventional varieties of avocados in the market. The second categorical variable "geography" indicates the different US locations where the avocado sales data is collected.

The Main Y(Independent Variable) in our prediction model will be **average_price.**
Some relevant columns categorical and Numerical variable in the dataset:
Date - The date of the observation
average_price- the average price of a single avocado
type - conventional or organic
year - the year
Region - the city or region of the observation
Total Volume - Total number of avocados sold
4046 - Total number of avocados with PLU 4046 sold
4225 - Total number of avocados with PLU 4225 sold
4770 - Total number of avocados with PLU 4770 sold

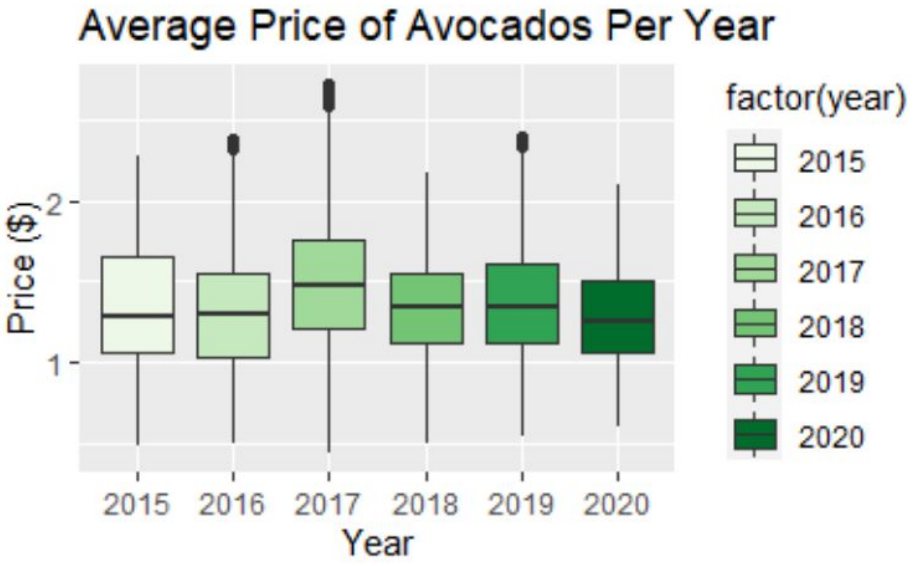
Aabha Desai
Isha Khasgiwala
Christopher Kupovics
Neetika Saxena
Puneet Sidhu

3. Any preliminary analysis you have tried at this point?

The Histogram of Average_Price shows that the dataset is slightly Right-skewed. This also appears to be a bimodal dataset with the mode closer to the left of the graph and smaller than either the mean or the median. The mean ($1.39) appears to be greater than the median ($1.35). The shape indicates that there are some data points,perhaps outliers,that are greater than the mode.

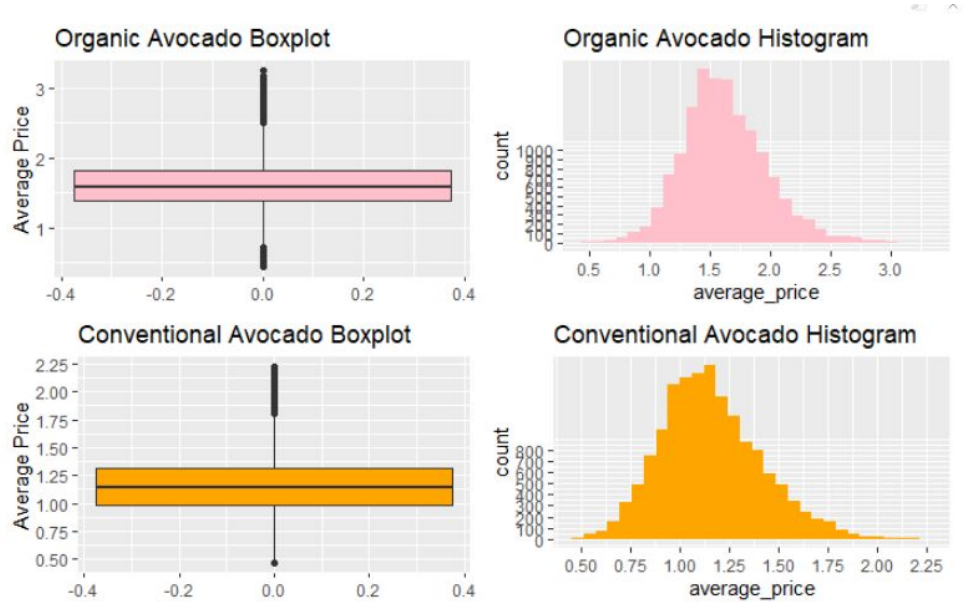


A box plot comparing the average price of avocados per year from 2015 to 2020. The year 2017 appears to have the highest prices. 2020 shows the lowest prices but the data only goes up to May 2020 so this is misleading. 2017 was one of the worst years for California wildfires, so this could provide support for the hypothesis that there is an increase in avocado prices due to the California wildfires.

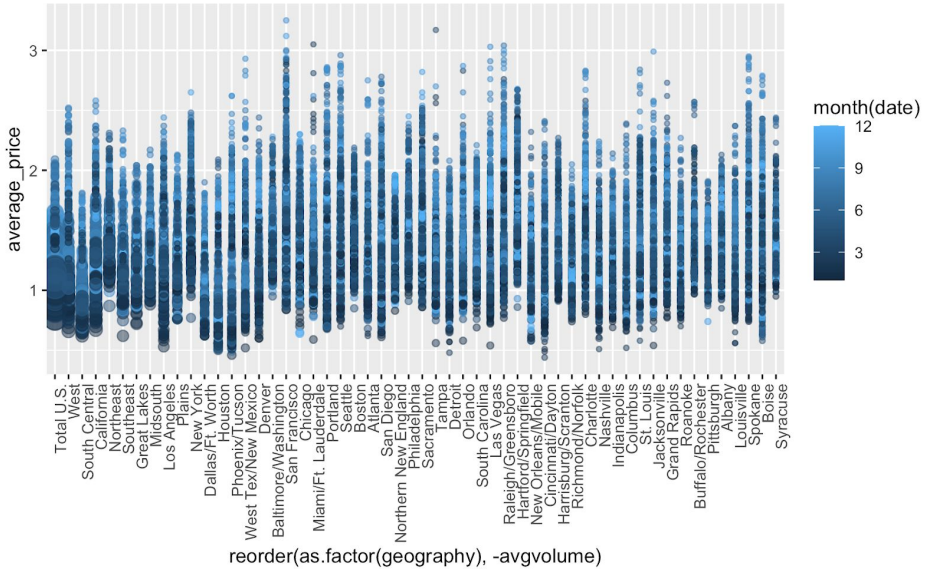


Aabha Desai
Isha Khasgiwala
Christopher Kupovics
Neetika Saxena
Puneet Sidhu

The Average price range for Organic Avocados appears to be less than that for Conventional Avocados. Also, we observe more outliers for Organic Avocados than that for Conventional avocados.The histogram for average prices of organic avocados appears to be bell shaped implying the data to be normally distributed. However, it also shows many outliers at both ends.The mean,mode and median appears to be pretty close and at around $1.5. The histogram for average prices of Conventional Avocados is right skewed showing more outliers on the right side of the graph. The median average price for conventional avocados appears to be in between $1 - $1.25. This shows that in general, the average price for organic avocados is more than that of the conventional ones.



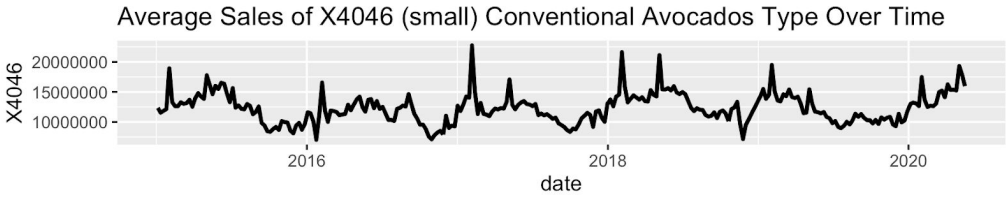
The plot below shows the relationship between average price and month for every geographic region. For most regions it seems that prices increase toward the end of the year.

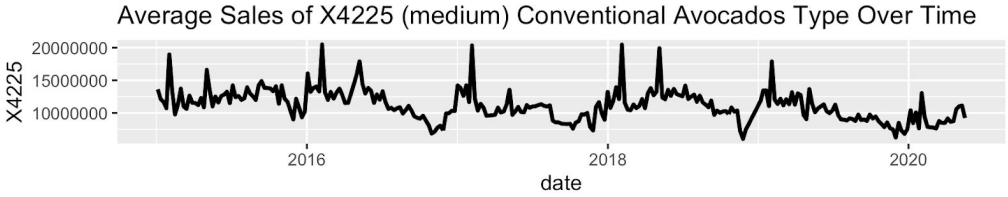


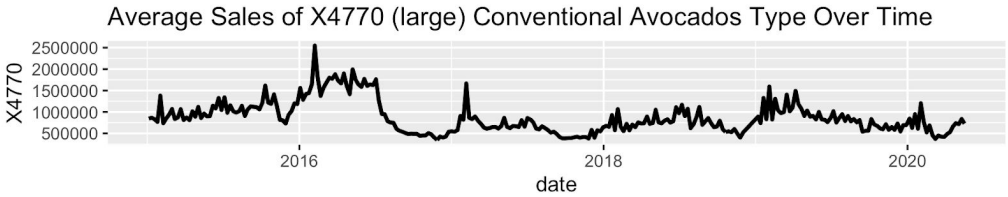
Aabha Desai
Isha Khasgiwala
Christopher Kupovics
Neetika Saxena
Puneet Sidhu

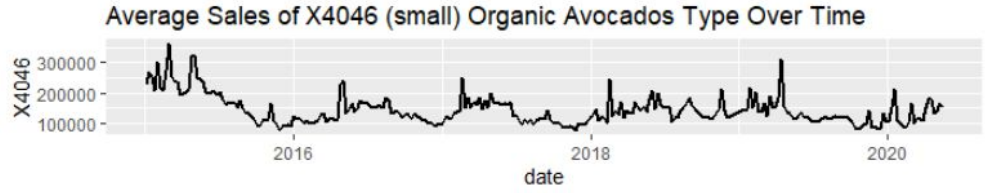
The following charts show the average sales for three sizes of conventional and organic avocados over time. Small and medium conventional avocados show a similar trend, whereas large conventional avocados show a large drop in average sales after 2016. All three conventional sizes saw spikes in sales in 2016 and 2017. Most spikes in sales happen early in each year. Small size conventional avocados shows more of an upward trend toward the end.

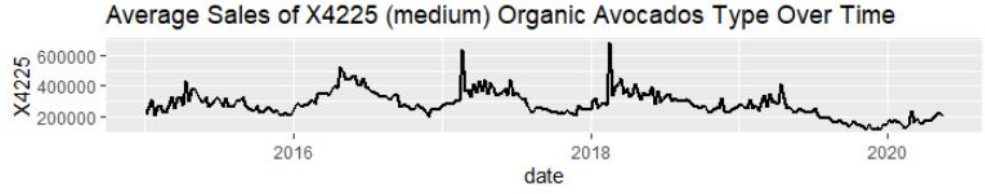
The charts for the organic avocados are much different, overall sales are lower with yearly peaks being less pronounced. Large organic avocados were very unpopular in 2017 and 2018, with a quick increase in 2019 before dropping off again.

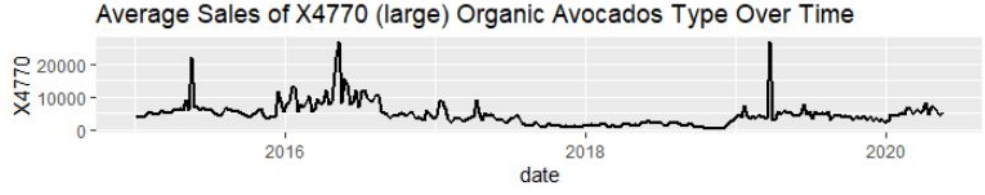






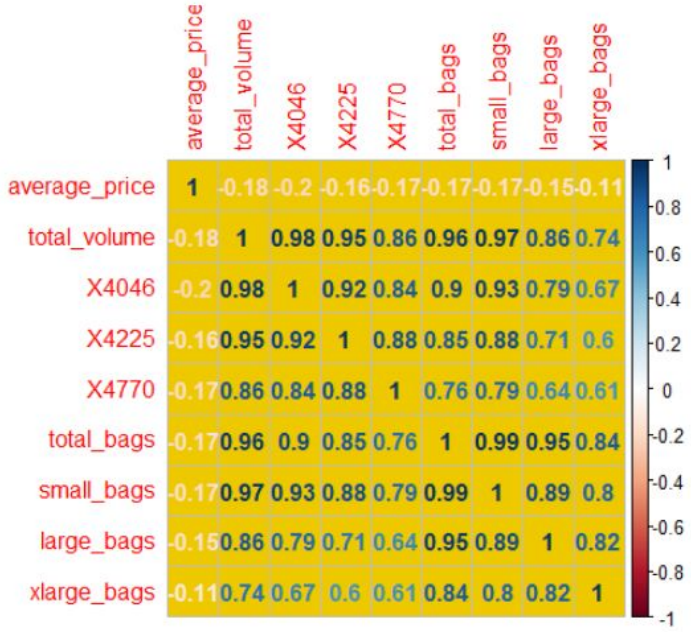






Aabha Desai
Isha Khasgiwala
Christopher Kupovics
Neetika Saxena
Puneet Sidhu

We created a correlation matrix to check if there are variables having strong correlations but it didn't show any useful correlations.



	average_price	total_volume	X4046	X4225	X4770	total_bags	small_bags	large_bags	xlarge_bags
average_price	1	-0.18	-0.2	-0.16	-0.17	-0.17	-0.17	-0.15	-0.11
total_volume	-0.18	1	0.98	0.95	0.86	0.96	0.97	0.86	0.74
X4046	-0.2	0.98	1	0.92	0.84	0.9	0.93	0.79	0.67
X4225	-0.16	0.95	0.92	1	0.88	0.85	0.88	0.71	0.6
X4770	-0.17	0.86	0.84	0.88	1	0.76	0.79	0.64	0.61
total_bags	-0.17	0.96	0.9	0.85	0.76	1	0.99	0.95	0.84
small_bags	-0.17	0.97	0.93	0.88	0.79	0.99	1	0.89	0.8
large_bags	-0.15	0.86	0.79	0.71	0.64	0.95	0.89	1	0.82
xlarge_bags	-0.11	0.74	0.67	0.6	0.61	0.84	0.8	0.82	1

4. **Any challenges you are experiencing (e.g. missing observations, outliers, big data) and how do you plan to deal with them?**

   One of the challenges in the data set is the presence of outliers as we can see in the histograms and boxplots. To deal with the outliers,we are thinking of using a logarithmic transformation to make the dataset distribution normal.

   The geography variable in the dataset is not as per state or region which makes it difficult to use for analysis. We are thinking of categorizing/arranging as per states to make it easy for analysis.

5. **What else will be included in the final report -**

   In the final report, we will go into more detail and perform more tests to prove each of the hypotheses that we have listed above. The tests that we will include are as follows:

   Hypothesis 1: Avocado pricing shows more correlation to season rather than region.

   - Test 1: ANOVA Analysis - this will be done to prove the analysis between the categorical values of the regions (such as Albany, Dallas, etc,) and the continuous variable of price. This will also be repeated using the categorical variables of seasons (such as Winter, Spring, Summer, Fall) This test was chosen as it compares the differences in mean

prices among the groups which will allow us to show which category has a higher variance among means.

Hypothesis 2: Over time, across all region's consumers become less concerned with fluctuations in avocado pricing.

- Test 1: Pearson's correlation - this will be done to prove the correlation between the prices and the demand of each given year.
- Test 2: Linear Regression - this will be used to identify the slope of the line between price and demand for each given year. We will then compare these slopes to determine whether the relationship between price and demand has changed or not

Hypothesis 3: There is an increase in Avocado prices during May-October during the seasonal California wildfires.

- Test 1: Pearson's correlation - we will compare month and price in California to determine whether there is a correlation between the two. This will identify whether the price actually changes based on what date it is. We will then see if the price value is higher in the months of May-October (during seasonal california wildfires).

Lastly, we would like to include a section where we will list some areas where this analysis could be improved and taken further in the future. It will list different opportunities that are available with the given data that we have calculated.