Avocado Prices and Volume by Area

By Sedona Munguia and Alex Hayes

Questions:

1. How has avocado consumption and prices changed over time?
    a. We will be answering this by using our avocado dataset and our US cities dataset. We will compute the prices and per capita consumption in the cities from the avocado dataset by using the volume sold from the avocado dataset and the population data from the US cities data set. We will then plot this by week from 2015 to 2018.

2. How is avocado consumption and price based on location within the US?
    a. We will use all three of our data sets for this question. We will compute the per capita consumption by using the volume sold in certain cities throughout the US. We will then use the geographical data provided with the in class data set to help us visualize where in the US the avocado consumption is higher vs. lower.

3. How can we predict the price and volume sold of avocados in various locations in the future?
    a. We will use a subset of our avocado dataset to train a machine learning model that we can then tweak by adjusting the parameter influence so that we can create a model that gives us the highest prediction accuracy.

Question Results:

1. Avocado consumption and price have stayed fairly steady over time, but it appeared seasonal for both.

2. Avocados consumption was more on the west coast, with one outlier on the east coast. Prices did not appear to have a large change across the USA, so location is not a major factor in price.

3. We can predict the volume relatively accurately, depending on the location, and we can predict the average price less accurately, but still fairly accurate in some locations.

Motivation:

The answers to our research questions are of the ***utmost*** importance. Knowing the answers to our research questions will help both producers and consumers of avocados have more information at their disposal before they buy or decide to produce. We can predict the future price of avocados, so producers and consumers can make the most educated guess for when to buy or sell. We can also make our data more readable by general audiences with our data visualization techniques, which will give a larger population greater access to avocado prices and other data. Ultimately, answering our questions will help avocados become the most popular fruit in the world!

Datasets:

1. This is a dataset about avocados that was scraped from the Haas Avocado Board website in 2018. It provides data on avocado prices and volume for various regions across the United States from 2015 through 2018. https://www.kaggle.com/neuromusic/avocado-prices

2. This dataset provides information on geographical locations of cities and their populations within the US. This will help us with our plotting by region. https://simplemaps.com/data/us-cities

3. This is the dataset provided in class that gives the geographical data to map the United States, which will help in our plotting and was not included in our cities dataset. (We end up not needing this dataset because plotly already has the united states map programmed in) https://courses.cs.washington.edu/courses/cse163/20sp/files/data/lecture-readings/gz_2010_us_040_00_5m.json

Challenge Goals:

- Multiple Datasets
    - We want to be able to look at our avocado data per capita over various regions so we are using a second dataset that includes populations for cities and city locations to help us with calculating and plotting. We have a third dataset that will help us plot the full map of the United States so we can show our data in the regions across the US.

- New Library
    - We want to use plotly to help make our graphs/maps. A lot of what we're doing is presenting data in an appealing visual way and plotly can help us

do that. We want to incorporate a level of interactiveness so you can see the data in a visual sense and then the actual numerical data. Also, we potentially would use it to make a sliding bar to emphasize our data change over time.

Challenge Goals Evaluation:

We believe we met both of our challenge goals in this project. We used multiple datasets and spent a considerable amount of time working to merge these datasets together in a way that allowed us to both plot and analyze with machine learning. This allowed us to come up with a more detailed analysis of the avocado dataset by making a plot that has geographical data as well as the volume and average price. For our new library, we went through the Plot.ly documentation to find a good way to visualize our data in a way that was both readable and informative. We decided to add a slider in our plot to help visualize both the change in average price and total volume sold over time. We also added in a machine learning portion, which added another challenge to our project.

Methodology:

First, we will load in our datasets. We will merge the avocado dataset with us-cities dataset by city, so we have population and geographical data about cities with the avocado data. This will allow us to make visualizations of the data over time, as well as per city. We will compute the per capita consumption in the cities from the avocado dataset by using the volume sold from the avocado dataset and the population data from the US cities dataset.

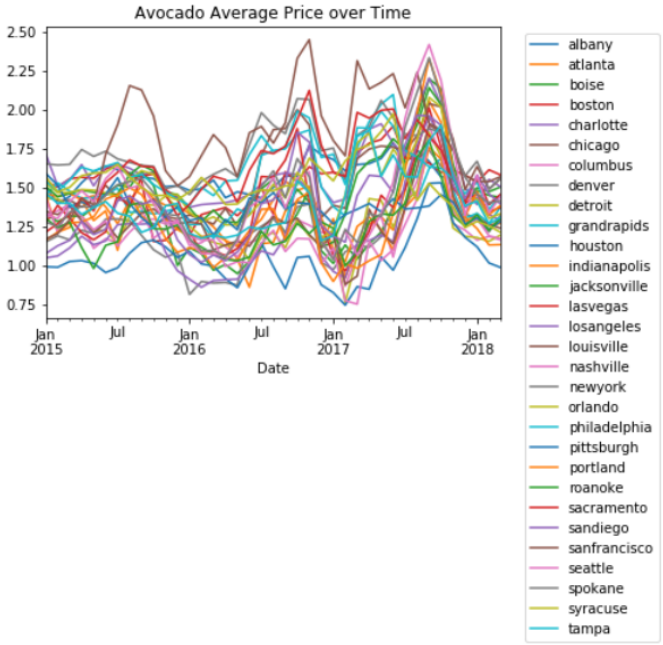
We will then plot this by month from 2015 to 2018. We will then use the geographical data provided with the Plot.ly to help us visualize where in the US the avocado consumption is higher vs. lower. We will do this by using the Plot.ly geographical data to give us a better representation of the United States and where each city falls, to give us more data on how avocado consumption changes by region.

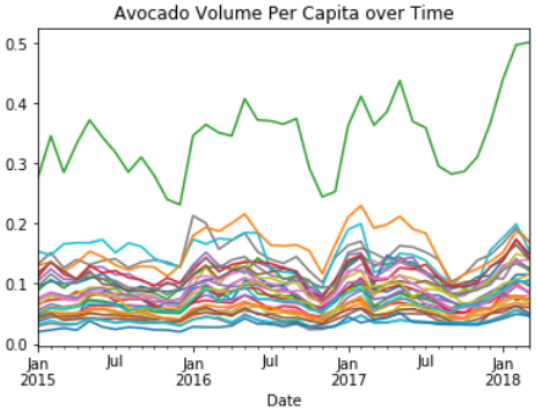
We will be using plotly to create our interactive graph that will showcase avocado consumption over time as well as by region. Using plotly we will make a bubble map over the US with larger bubbles corresponding to larger per capita consumption for the city. This plot will include a sliding bar for time that increases by month and shows the data for that month in that year.

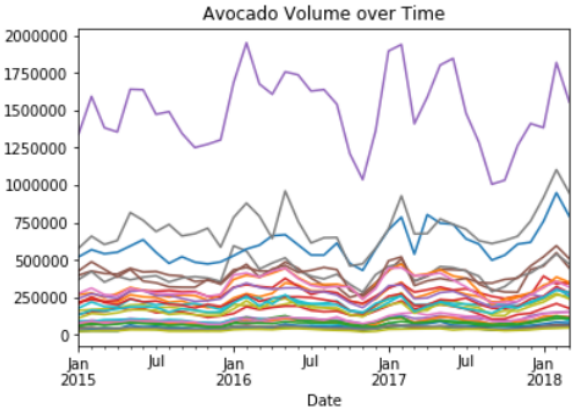
Finally, we will take our data from our merged dataset and use sklearn to train a machine learning model to predict both the price and volume of avocados in the future in various locations. We will have clean data with no missing values, and we will create the model by city. We will use features such as average price, total volume, and others over time to construct a decision tree regressor. We will split the data into 75% for testing and 25% for training. We can then adjust hyperparameters to optimize the accuracy of our model.

Results:

1. There was not a significant upward trend throughout the 3 years. There were definitely peaks, showing that avocado prices and consumption are seasonal. The prices tend to increase in late summer/fall and are lower in the beginning of the year. There is one large peak in Sep 2017, however prices go right back to those of 2015 by Jan 2018. Avocado Volume per capita was the avocados consumed divided by the city population. There are some general dips around early winter, but overall there is very little steady increase over time, both per capita and total volume, which is not what we expected.

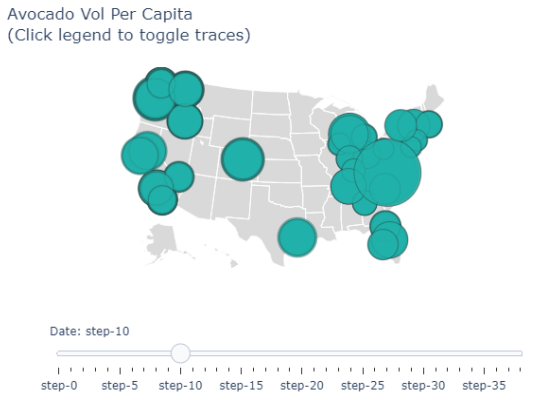


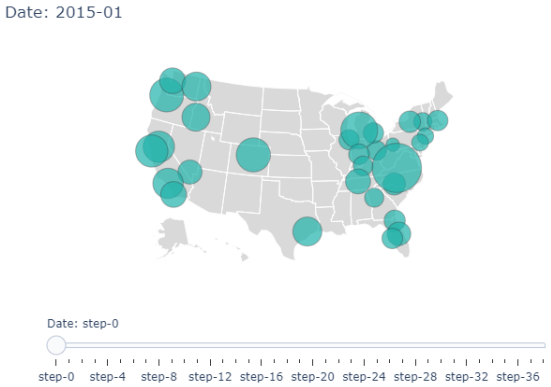


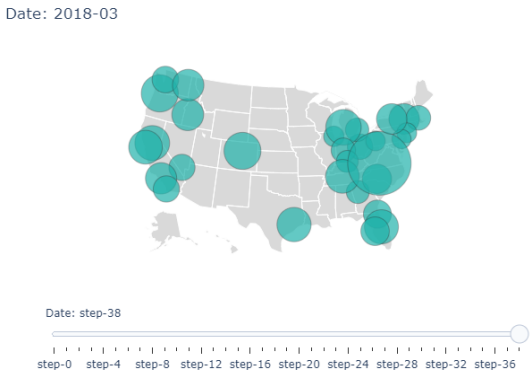


2. A lot of the consumption is based on the west coast. There was one outlier, Roanoke VA, which had the largest avocado consumption per capita than every other city throughout all three years. There is not a significant change of consumption over time, but the maps show that there is an increase in east coast consumption over time. This is shown below with the map from January 2015 and March 2018. The prices were very similar throughout the country, typically ranging about 50 cents. These were not quite the results expected, I thought that there may be more of a dependence of price on location, but the analysis shows that is not the case. Below there are maps from Jan 2015, March 2018, and Sep 2017 because that was when the prices peaked, before returning back to the regular prices.

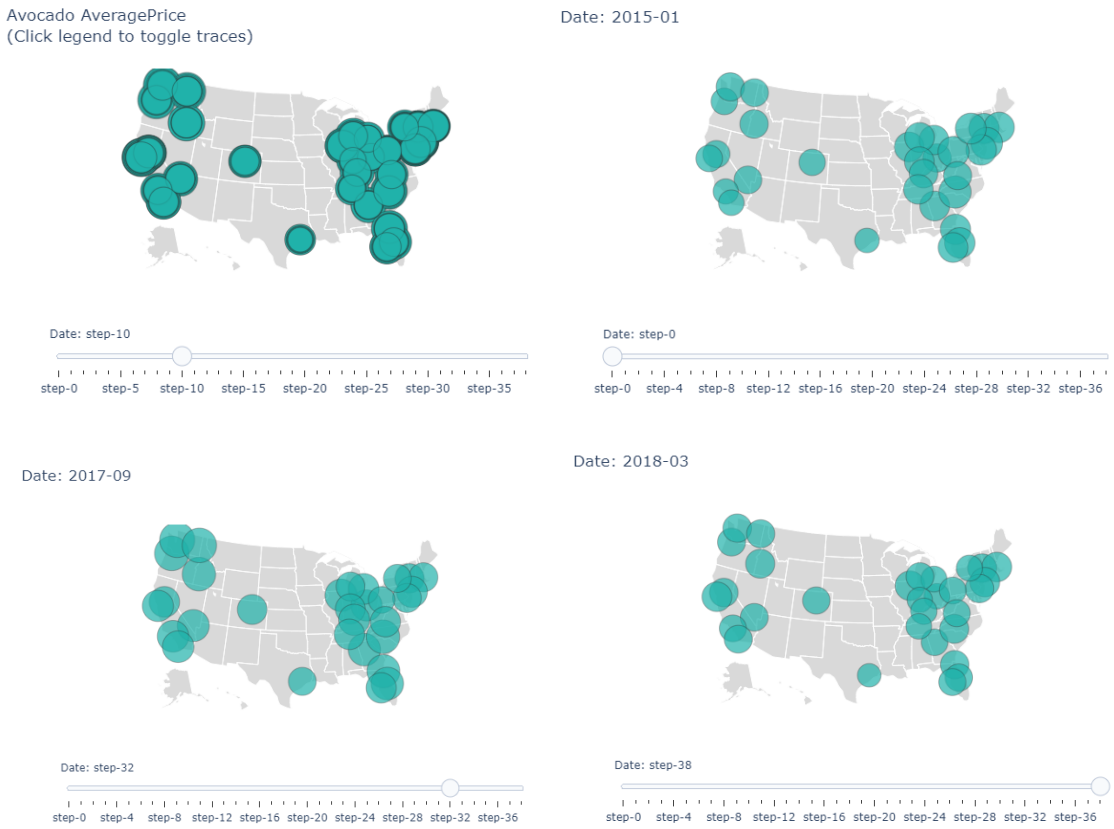
   Avocado consumption per capita: Overall, first, last







   Avocado average price: Overall, first, peak, last



3. Predicting the price and volume of avocados with a decision tree regressor illustrated some interesting results. We were able to more accurately predict the volume than we were able to predict the average price, and the location wildly affected our model. This is because some locations have extreme outliers that skew our data and are difficult to predict. In figure 1, we show a plot of the feature importance as well as the predicted data versus the actual data for total volume in New York. Interestingly, the most important feature by far was '4225', which is a type of avocado. We used the 2015 data as a training set, and were able to fairly accurately predict the 2016-2018 data points. Our model did struggle to predict any large spike in volume, especially a negative deviation. Our model had a tendency to overfit the data, even with hyperparameter tuning. We tuned max_depth, min_impurity_decrease, max_leaf_nodes, and min_samples_leaf. The most influential of these was max depth. Shown in figure 3 is a plot of max_depth vs. MSE for our model. There is a sharp decline, and then it stays mostly constant. This plot varies from city to city, and also varies with the model being for price or volume.

Our model for price was much less accurate. Shown in figure 2, we plot feature importance and prediction accuracy for Seattle. The feature importance was more equally distributed, but our model overfit when working with the training data. Our model works on some cities within the dataset, but for others it is very inaccurate. This could be due to a low amount of samples that our models have to work with.

Overall, our model does a decent job of predicting the future sales and volume, and would be beneficial for both consumers and growers of avocados. There are clear trends in the data, even where our model is not as accurate, and we can use those trends to know when avocados will be more or less expensive.

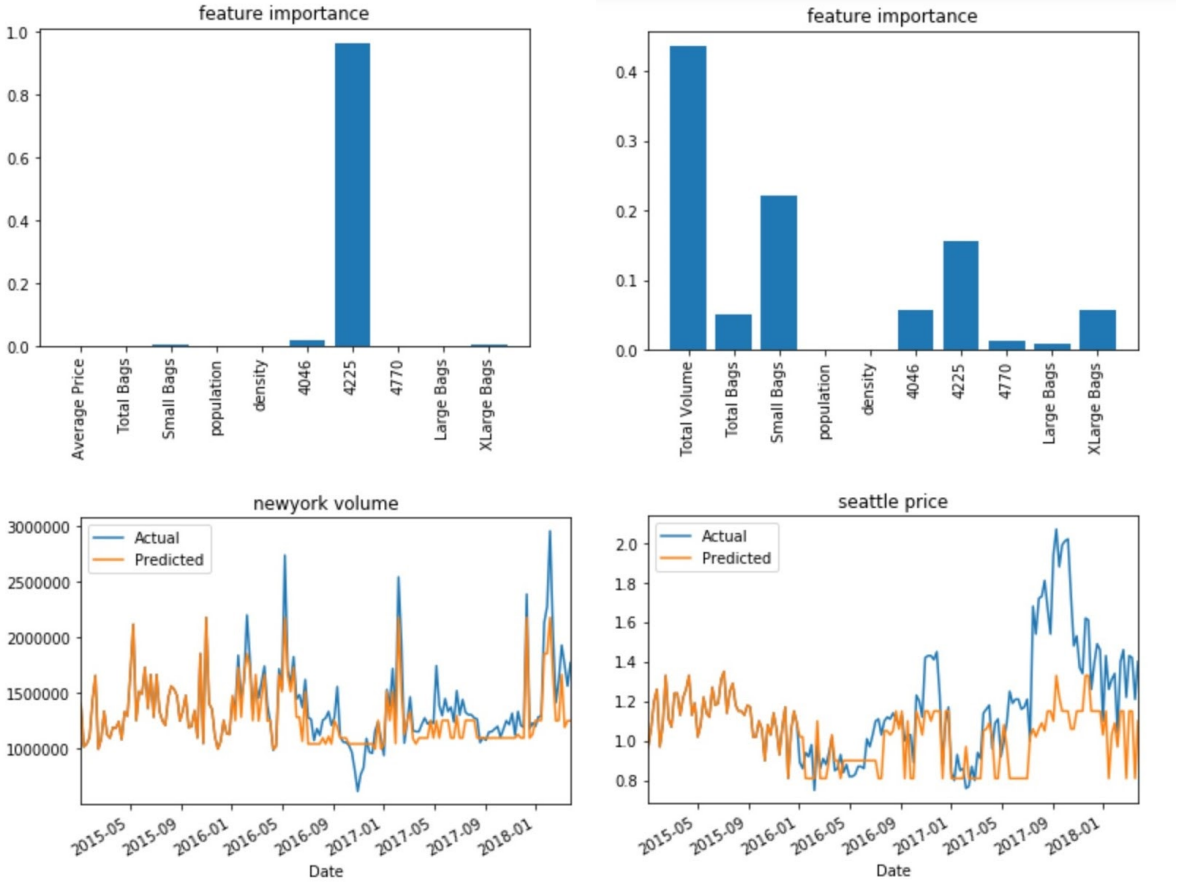


Figure 1

Figure 2

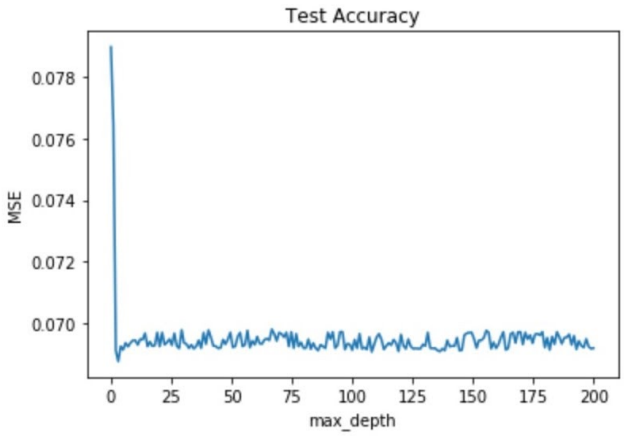


Figure 3

Work Plan:

For this project, we will be using Github to develop and test our code. We will be working together on the project. We will likely split up smaller tasks within our larger work plan points while keeping in close communication so we can ask questions if one person is struggling.

1. Initial data processing - 2 hours
    a. We will merge our datasets and begin initial computations to help us with plotting and machine learning.
2. Plotting - 6 hours
    a. Create our map that shows avocado consumption per capita over time
    b. Create another map showing avocado prices per capita over time
3. Machine learning - 10 hours
    a. Train a model to predict volume and price of avocados in the future for different cities using a decision tree regressor
4. Writing Report / Final edits - 3 hours
    a. Put together our final report and make any final changes to our code so that everything is running smoothly

Work Plan Evaluation:

1. Initial data processing - 2 hours
    a. This took about two hours, a lot of the time was taken up by thinking of ways to clean our data to what we wanted, and to get rid of a lot of the superfluous information that would make our plots more difficult to create. We ended up not needing the in class dataset for our program to plot the

United States, we used it to initially confirm that our cities plotted correctly, but Plot.ly has a U.S map built in that we used instead.

2. Plotting - 6 hours
    a. This definitely took longer than 6, probably about 10 hours. It took longer than expected to learn how to use plotly. Setting up the plot using our specific data required figuring out how to format in order to get that bubble map working.

3. Machine learning - 10 hours
    a. This took a little more than 10 hours, about 13. We were able to set up a somewhat accurate model pretty quickly, but a lot of time was spent figuring out which parameters were most influential and then fine tuning them. We also spent some time creating visualizations of our model and the results, so that took up some time.

4. Writing Report / Final edits - 3 hours
    a. This took about 5 hours, probably a little more to get everything working on Visual Studio.

Testing:

Much of our program was visualizing data, so to test our data and make sure that it was reasonable, we mostly looked at our plots to make sure that the data plotted matched the data from our dataframe. For our part on machine learning, we used the data from 2015 as the training set, so plotting both the training predictions and test predictions vs. the actual data gave us a good idea of how our model was working.

Collaboration:

We did not collaborate with anyone. We used online documentation for plot.ly, pandas, and scikit-learn.