Jovany Gonzalez
CISD41
12/06/2020

Avocado Summary

For this project, Rashmi Chocksey approached Jorge Medina, and myself, to make this project go by quicker, easier on our lives, and get to know one another. We spent 3 days debating and deciding on which dataset we should use. We thought about using COVID, space, video games, ramen, or a mental health dataset before we finally settled on our final dataset with the help of hunger, avocados. We broke up the work to see what kind of data each of us can pull from this dataset and set to meet up at least twice a week outside of lab time. This dataset had a lot of interesting information that we saw that we could manipulate for comparisons or highlight. Did Small sized avocados sell more than the other sizes? Which year sold more between 2015 and 2019? Do people buy XL bags? We had plenty of questions, although some of these questions didn't make the cut due to the inconsistent values that messed with the rest of our data. An example of this was Organic avocado types, we had to drop this because there was little data and plenty of null values ridden with this type. For my hypothesis I chose "2015 did better than 2019 in Average Price." From doing some research we assumed that 2015 had to be high point for avocados because of a threat of an avocado extinction would cause a price hike and I thought there would be a lower demand. Later I tried to find what year did better so when I tried plotting those years out to see if the assumption was right, turns out I was wrong. Avocados rose in sales and even with a threat of extinction avocado sales skyrocketed. I used violinplots, barplots, stripplots, and a jointplot to see when and where the prices did well. I threw in the location, various sizes, bags, and even month to see how much it would change our data output. Every time we met up one of us would find something new and interesting that would help us along the way to answering our questions. For example, as a jump off point I showcased how to combine two separate years into one dataframe using concat. We have used this syntax quite a bit in our code to find and make it easier to find the overall population mean. By the time we were done plotting it was clear that avocado increased 10x in between 2015-2019. It's quite unfortunate we couldn't use the data from 2020 as the dataset is currently incomplete and we were curious on how sales were doing during the pandemic.