Amy Newman
618 WN 2018 Final Project Report
Title: Avocados Over Time

1. Motivation

For my Master's Thesis, I helped create a large dataset from a Meraki dashboard containing data from wifi routers in European refugee camps, specifically ones located in Serbia and Greece. Included with the router data will be demographic data from UNHCR reports including: population amount, gender, nationality, and adult and child population per refugee camp. This data is being stored in a MySQL database. Due to the thesis data not being complete at the time of the initial project proposal, I decided to find a dataset that would allow for similar analyses and predictions. For this project, I chose the Avocado Prices dataset from Kaggle as it has similar enough type of features.

Four Questions

1. Look at the general trends and time series plots for average price and total average volume sold in all the regions. What are the trends? Is the data normal? Is there stationarity?
2. What is the forecast of total volume of avocados sold for conventional and organic avocados?
3. Does the data demonstrate seasonality?
4. Are sales prices statistically significant between two cities?

2. Data Source

I wanted to find a dataset similar enough to my thesis data that would also be fun to explore. For this project, I utilized the Avocado Prices dataset from Kaggle: https://www.kaggle.com/neuromusic/avocado-prices. I downloaded the .CSV from Kaggle and imported into a Jupyter Notebook. The data is a weekly account of the following variables over a period of four years (2015-2018) and includes numeric (integer and float), string, and datetime data types:

- Numeric data includes: type of avocados sold either by the avocado or by the bag, specifically small, large, extra large, the average price per week, the total volume sold, and the year
- String data includes: the city or region and whether the avocados were organic or conventional
- Datetime data includes: the week

3. Methods

1. Look at the general trends and time series plots for average price and total average volume sold in all the regions. What are the trends? Is the data normal?

- **Data Manipulation and Preparation:** To determine general trends for average price and volume, I first did a pairplot and then a distribution plot to determine if there were any specific trends. I found that there were some correlations with specific columns and also that the overall data looked to be bimodal. Considering that organic and conventional items tend to differ in price, I created a new distribution plot splitting the two types and found that type seemed to play a role in how the data is distributed. From there on, I decided to look at the data in terms of organic and conventional so as

to find more distinct trends. With regards to time, I looked at the data as a whole, and by year to help determine yearly trends and a general trend in the entire dataset.

- **Missing and Incomplete Data**: Fortunately there were no missing data for this dataset. The main issue was transforming the Date column from an object to a datetime type so it could be used for analysis. From there, the Date column had to be reorganized as dates were in descending, rather than ascending order for each year of data.
- **Workflow of Source Code**: Once the data was organized and cleaned, to find trends, first created an overlay of average avocado price per year and average avocado volume sold per year to see if there were any trends. I did this for conventional avocados and their total sales volume is greater than that of organic. From there, I decided to also take a look at the general trends of each variable. Volume sold seemed most interesting to me as there were distinct trends each year, or a few distinct spikes in sales. I looked at the data for rolling mean and standard deviation for both conventional and organic to compare the two trends.
- **Challenges:**I had not encountered time series data before, so it was fun to learn how to manipulate it and create trends. I plan on using similar techniques for my thesis project.

2. What is the forecast of total volume of avocados sold for conventional and organic avocados?

- **Data Manipulation and Preparation:** For this question, I utilized the clean data from question 1. I calculated the mean value of each column per date. The dataset initially included regional data, so there were multiple entries per date, or the Date column had multiple entries with the same date. I did an average by Date column in order to get an average value across each region to come up with a total average price or total average volume sold across the entire dataset.
- **Missing and Incomplete Data**: Please look at question 1.
- **Workflow of Source Code**: I looked at trends for the total average volume of conventional and organic avocados sold, then utilized the Prophet package to forecast future volume sold based on the current data. Once the model was run and predictions were made, I concatenated the current data with the future data to show average total volume sold over time.
- **Challenges:** I initially tried predicting prices, but the prices didn't look quite right, so went with average total volume sold.

3. Are there Seasonal Trends in the Data?

- **Data Manipulation and Preparation:**

1. What is the forecast of total volume of avocados sold for conventional and organic avocados?
    a. how did you manipulate the data to prepare it for analysis?
        i. I utilized the same data from the previous questions, gathering the mean total volume sold and the date.

- **Missing and Incomplete Data**: Same as for the previous questions.
- **Workflow of Source Code**: I imported the statsmodels packages and used it to determine if the data has seasonality for both conventional and organic avocados. Trends showed similar data spikes at around the same time period each year.
- **Challenges:** I had issues with getting the data formatted correctly for statsmodels. It initially worked, then ran into a few errors.

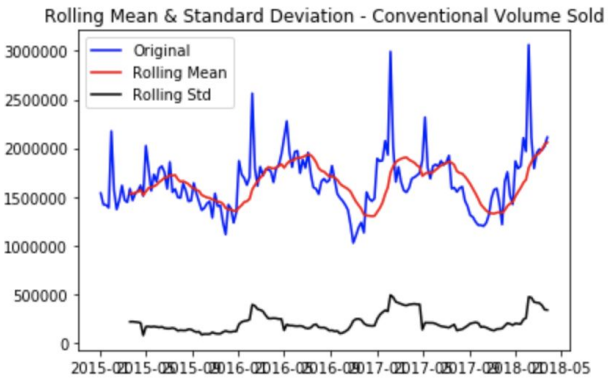
4. Are sales prices statistically significant between two cities?

- **Data Manipulation and Preparation:** I utilized the data from the previous questions and also created a factor plot to determine pricing trends across regions. I reorganized the data to show cities in order with the lowest to highest prices in the year 2017 because it has a full years worth of data and is more recent. I then pared the data down to Portland and Spokane because the cities are in a similar part of the country and I wanted to determine whether their pricing would have a similar relationship.
- **Missing and Incomplete Data**: There was no missing or incomplete data for this task.
- **Workflow of Source Code**: I first visualized the data, then chose to cities to focus on within the visualization. From there, I chose two similar cities and used a t-test to determine whether their average prices were similar.
- **Challenges:** Challenges for this task were determining the proper test to carry out for the data.

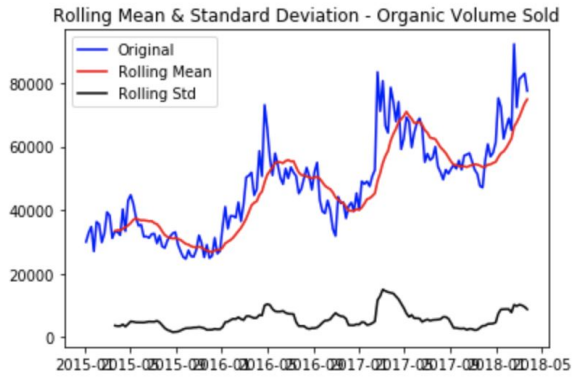
4. Analysis and Results

Question 1: Look at the general trends and time series plots for average price and total average volume sold in all the regions. Is there stationarity in the data?

Interesting Results for Stationarity: When looking for the stationarity of time series data, we're looking to see whether the data demonstrates a constant mean, constant variance, and an autocovariance that does not depend on time. Taking a look at the initial plots for both Conventional and Organic Volume Sold and then comparing to the Rolling Mean and Standard Deviation plots, it does not look either subsets of the data display stationarity. For data to have stationarity, it must now show any trends or seasonality. The data spans four years and there seem to be four larger spikes for each year of data. This demonstrates that there could possibly be a seasonal trend in the data, which would make sense as the data is for an agricultural product. As noted in *Forecasting: Principles and Practice*, "a stationary time series will have no predictable patterns in the long-term." The four larger spikes in the data seem that they could be predictable patterns, leading me to believe that there is no stationarity in the data. **Source: https://otexts.org/fpp2/stationarity.html**



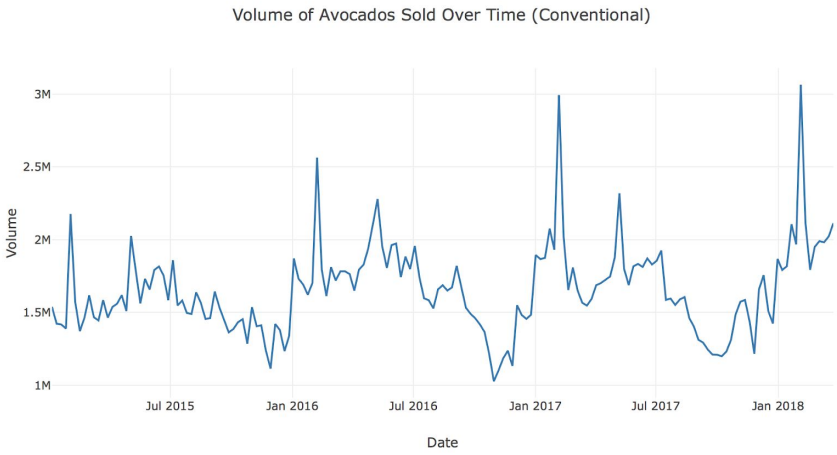
Trying to Determine Stationarity (Conventional)



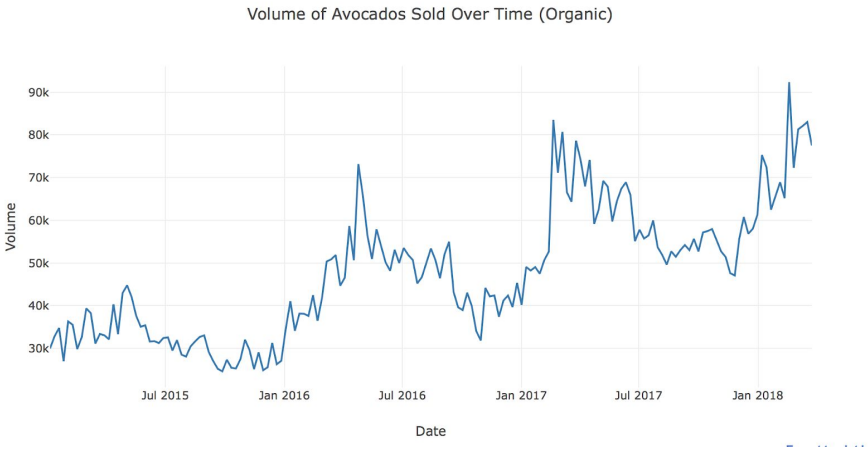
Trying to Determine Stationarity (Organic)

Question 2: What is the forecast of total volume of avocados sold for conventional and organic avocados?

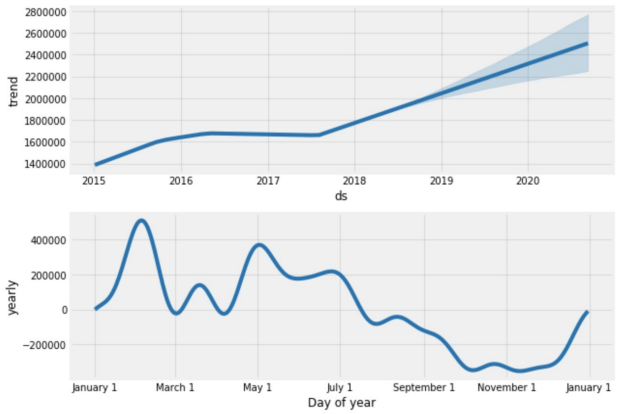
Interesting Results for Forecasting Avocado Sales: For the forecast of total volume of avocados sold, I noticed a difference between the conventional and organic avocados. The data for the conventional avocados showed a linear increasing trend from 2018 past 2020. Looking at the data for volume of conventional avocados sold, while there does seem to be a general increase in volume sold over time, the prediction doesn't seem like it should be as linear as it is. The trend line also does not look as precise, there is more room for variance. There seems to be less certainty in the forecast for average total volume sold of conventional versus organic avocados sold over time. When looking at the total average volume of organic avocados sold over time, there seems to be a general increase in volume sold between 2015 and 2018. It also looks to be more of a constant increase than that seen in the conventional data.



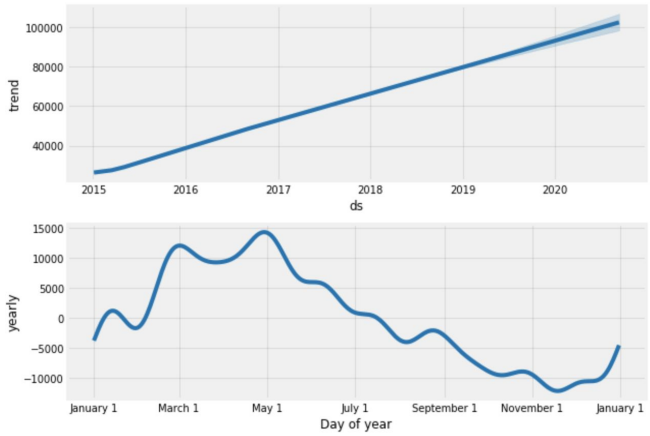
Average Volume of Conventional Avocados Sold from 2015-2018



Average Volume of Organic Avocados Sold from 2015-2018



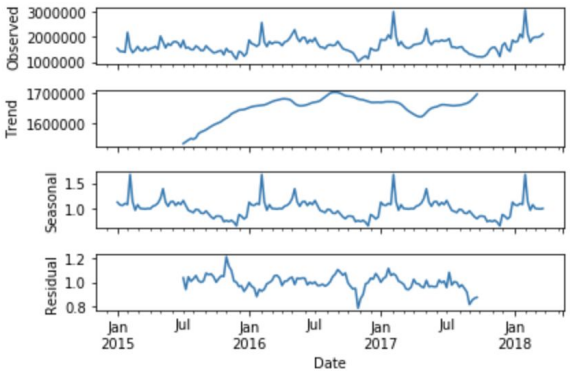
Forecasting Trend for Total Avg Vol Avocados Sold(Conventional)



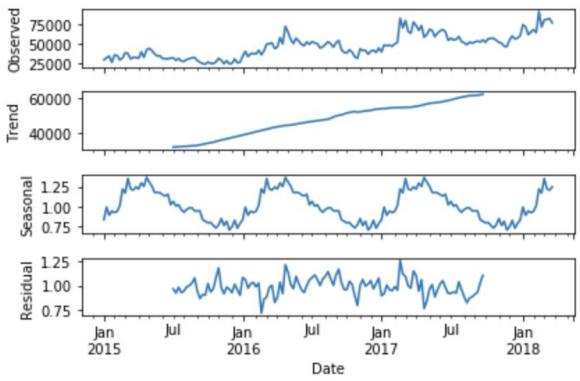
Forecasting Trend for Total Avg Vol of Avocados Sold (Organic)

Question 3: **Does the data demonstrate seasonality?**

Interesting Results for Determining Seasonality: At first glance of both the conventional and organic total volume of avocados sold, there seem to be spikes in the data each year at roughly the same point in time. Taking a closer look, the larger spikes all seem to occur the first week of February, while a few smaller spikes occur the first week of May. A look at the price data will show that the week after the larger spikes in February, or the second week of February, the prices show a small decline. Before doing any analysis on the data, I was hoping to see a relationship between the amount of avocados sold and the Super Bowl. With a quick Google search, I found the dates of the Super Bowl between 2015 and 2018. The dates for the Super Bowl each year match the exact weeks where the data shows a large spike each year, often the largest spike per year. To further test this hypothesis, I used statsmodels to determine whether there was seasonality in the data. In the Seasonal portion of the data, we can see a somewhat distinct seasonal trend in the conventional data and what looks to be a more distinct seasonal trend with the organic data.



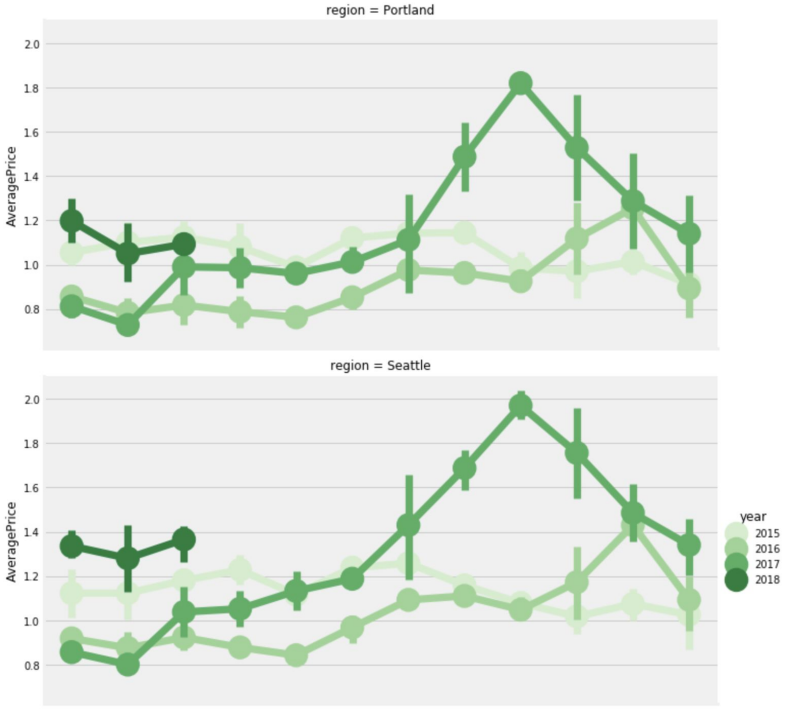
Conventional Avocado Total Volume Sold



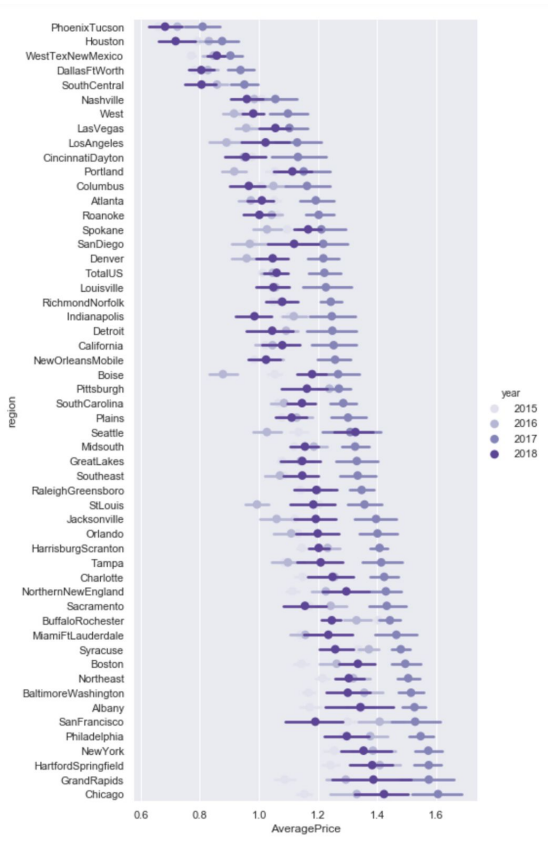
Organic Avocado Total Volume Sold

Question 4: Are sales prices statistically significant between two cities?

Interesting Results for Determining Price Differences: From the visualizations of prices over time for each city, it seems like they could be statistically similar, but the results from the t-test suggest otherwise. The p-value for the t-test was 0.0005, which is smaller than 0.5 and shows that the average avocado prices for these two cities are not statistically similar. This makes sense because even though population sizes may be similar for cities that are fairly geographically close distance-wise, it does not mean that the cities may have much in common. The type of people, based upon personal interests, may differ greatly, as well as typical age range. These factors could play a role in the supply and demand for avocados in each region, which could also lead to price differences.



Avg Avocado Price/ Year (Portland and Seattle)



Avg Avocado Price/Year (all Regions in Dataset)