Avocado Project Data Wrangling

The question to answer is:

How can I predict the volume of avocados that will be sold, and their price?

Step 1: Import years 2016, 2017 and 2018 as DataFrames

Get Excel data for years 2016, 2017 and 2018, for both conventional and organic avocados from the website: http://www.hassavocadoboard.com/retail/volume-and-price-data

This data contains the avocado sales for U.S.A.

First, I read the Excel files and append them to two dataframes for conventional and organic avocados.

I set the names of the columns to have different names for conventional and organic, and to have easy to manipulate column names.
I cast the date column of both dataframes to datetime type, then I set the index as date.

I can now concatenate both dataframes.

Step 2: Look at the data

Run df.head(), df.tail(), df.describe() and df.info().
The data spans from 01/03/2016 to 09/09/2018
There are 137 rows, no null values.
There are 18 columns, all of float64 type.
Columns for conventional avocado:

- con_price
- con_volume = con_4046 + con_4225 + con_4770 + con_bags
- con_4046
- con_4225
- con_4770
- con_bags = con_s_bags + con_l_bags + con_xl_bags
- con_s_bags
- con_l_bags

- con_xl_bags

Columns for organic avocado:

- org_price
- org_volume = org_4046 + org_4225 + org_4770 + org_bags
- org_4046
- org_4225
- org_4770
- org_bags = org_s_bags + org_l_bags + org_xl_bags
- org_s_bags
- org_l_bags
- org_xl_bags
-

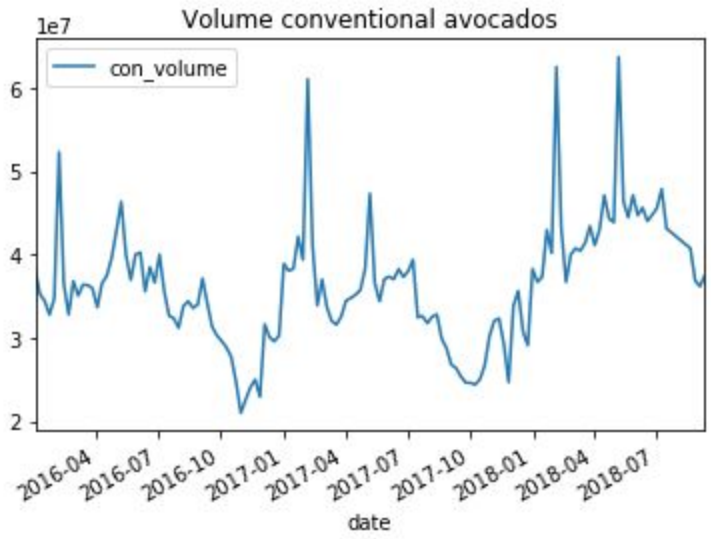
4046 identify small/medium Hass avocado, 4225 identify large Hass avocado, 4770 identify extra large avocado. You can see this product lookup code (PLU) on the avocado sticker.

The max for conventional avocado volume is 30 times the one for organic avocado, so I will have to plot them separately.

The price is $0.76 to $2.09 so I can plot the unit price for conventional and organic avocados together.

Step 3: Plot the data

Volume for conventional avocados:

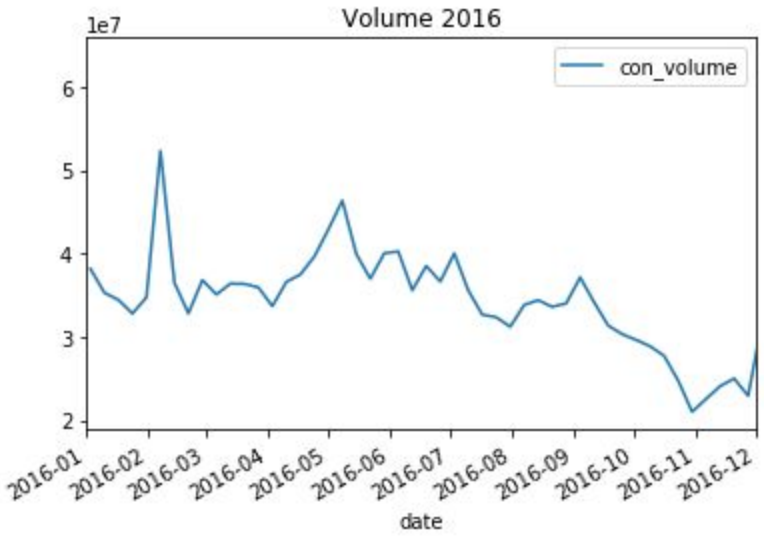


I can see some peaks in the plot. Are these outliers?
I don't think so because:

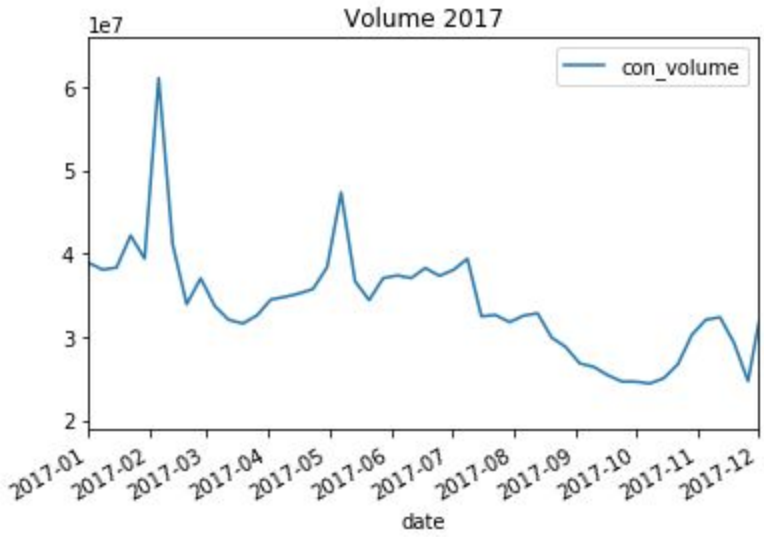
- These peaks appear each year at the same period

- These peaks appear on the plots for each type of avocado

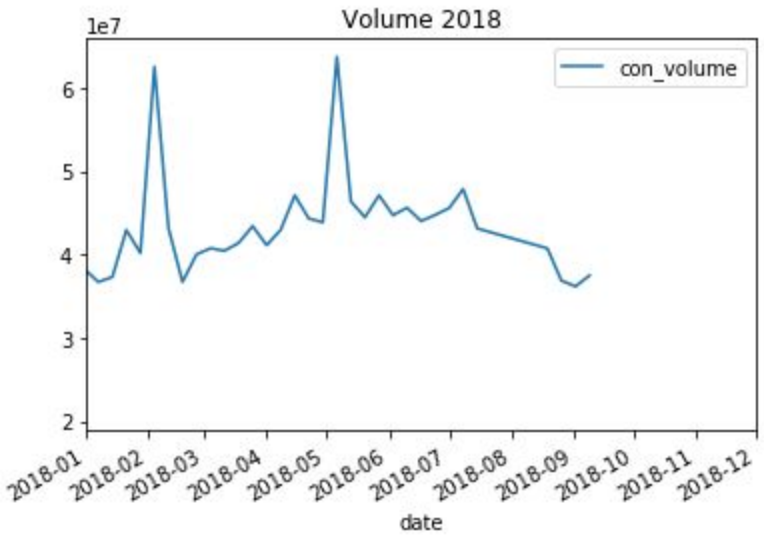
Volume for conventional avocados - year 2016:



Volume for conventional avocados - year 2017:

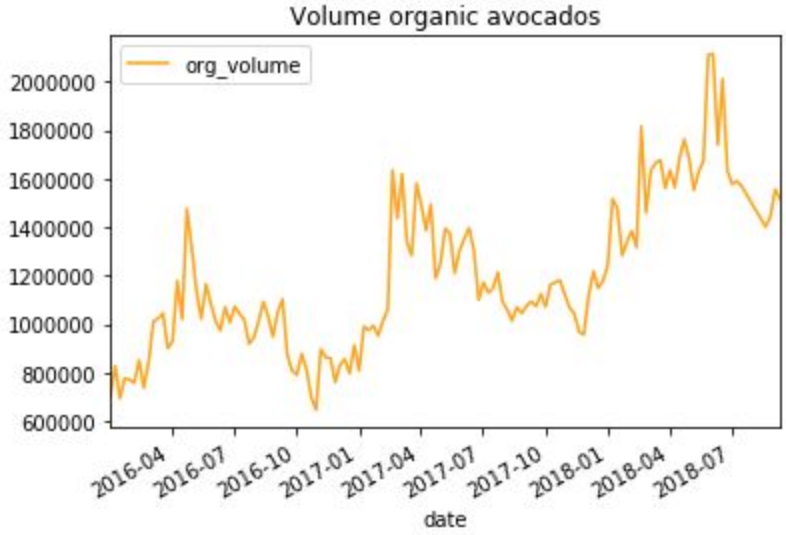


Volume for conventional avocado - year 2018:



There is a pattern that reproduces every year, with peaks at the beginning of February, and beginning of May.
Are these peaks related to holidays?
Maybe, beginning of February is Mexican Constitution Day, beginning of May is Cinco de Mayo. However, some holidays don't seem to have the same impact on avocado sales: Halloween, Thanksgiving.
My first idea was to flag the dates as holiday/not holiday, but after this preview of the data, I will mark the dates with specific holidays.

<u>Volume for organic avocados:</u>

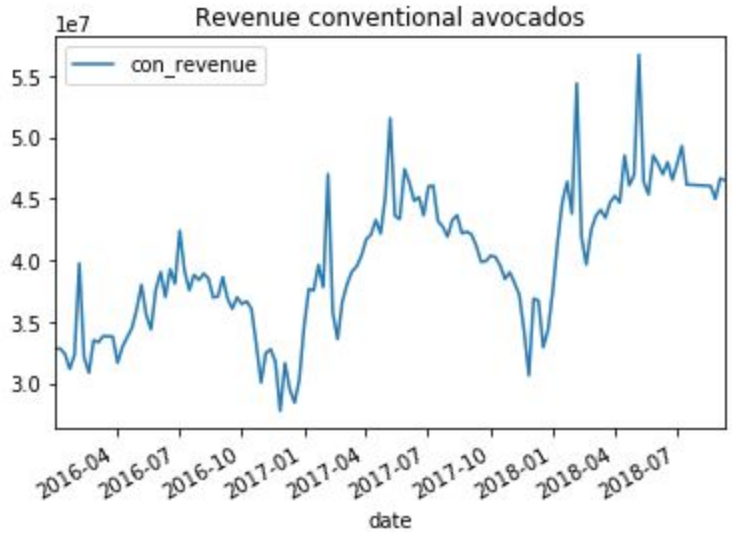


The organic avocados follow the same pattern as conventional avocados, but the peaks are smoother.

Unit price:

It seems there is a negative correlation between price and volume: high volume corresponding to low price.
It could be interesting to plot the revenue, to see if it's stable.

Revenue for conventional avocados:



This plot reflects the same highs and lows as the volume plot.

Step 3: Merge with the holidays

I define the holidays as lists for each year. I will use the most important U.S holidays and add the Mexico constitution day and Cinco de Mayo.

Then, I create a dataframe with date as datetime index.

I reindex this dataframe with the dates of the avocado dataframe, with nearest fill limited to 1. I hesitated with backward fill (because you shop before the holiday), but the sales are reported at the end of the week, so I think nearest method is more accurate.

Finally, I concatenate the holiday dataframe, adding a column ‘holiday’ to the avocado dataframe.