Avocado Project Final Project

Avocados are fruits we can buy on the markets or in store all year round. But can we predict the volume of sales week after week? Are there events that impact our avocados consumption? Does the price of the avocados influence our consumption?

The question to answer is:

How can I predict the volume of avocados that will be sold ?

This project is useful for all actors of the avocado industry:

- The **producers** can estimate their income, using the retail sale price as an indicator. The trend over the years can help them to prepare for the next harvest.
- The **importers** can order and ship avocados according to this forecast. Most of the imported avocados come by truck from Mexico, then by boat from central and South America. Because of the shipping time, the importers need to anticipate the volume of avocado they will sell.
- The **wholesalers** can anticipate the volume of avocados they will need. Fruits are perishable goods, so the stock volume must be monitored efficiently to avoid waste.
- The **retailers** can adjust their sale price according to the estimation, and be competitive.

The dataset I will use for this project is available to download as excel files from the Hass Avocado Board website (http://www.hassavocadoboard.com). This group was established in 2002 by Domestic Hass avocado producers and importers. Their objective is to promote the consumption of Hass avocados in the United States.

The data contain the actual retail price and volume collected from the retailers' cash registers every week.

The techniques to use in this project will be Time Series modeling and forecasting.
The deliverables will be Jupyter Notebook, a technical report and a presentation of the project.

Table of contents

1. Get the data

I got 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 Hass avocado sales for U.S.A.

First, I read the Excel files and built a dataframe with datetime index.
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

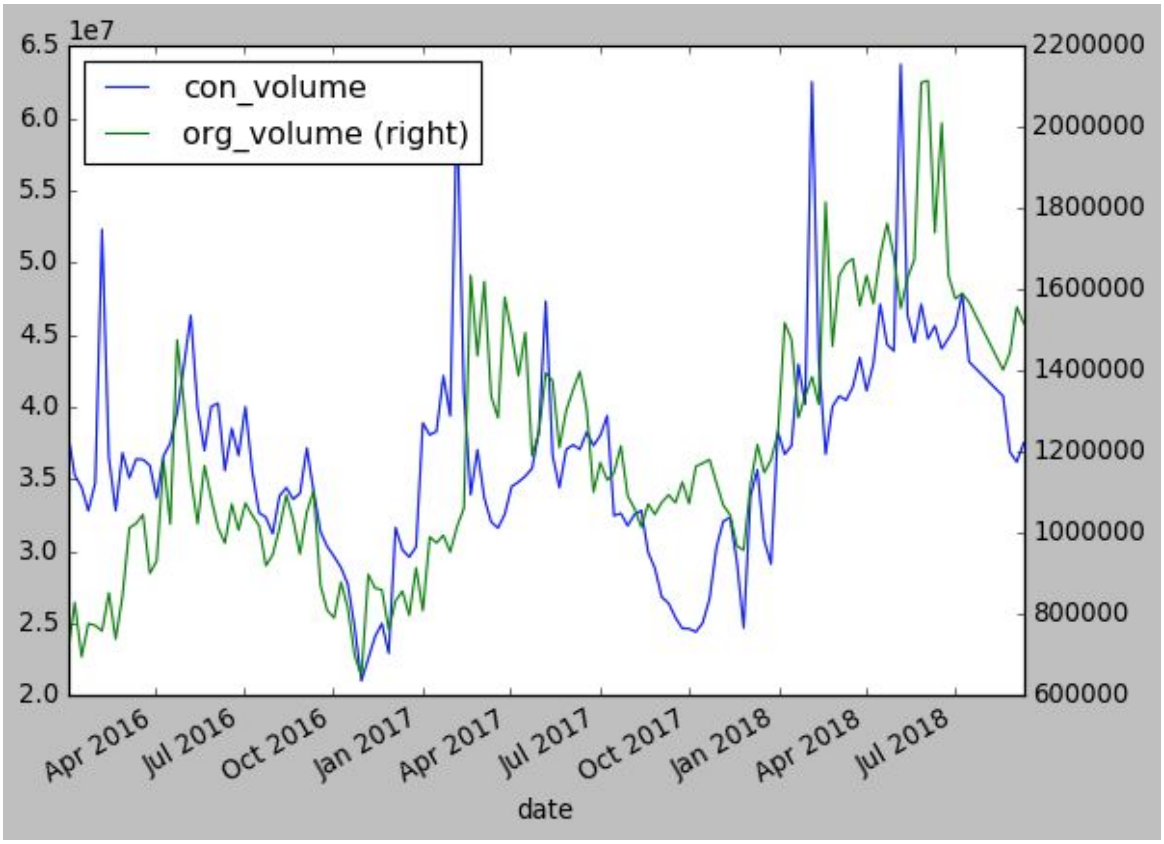
Columns for organic avocado:

- 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.

2. Plot the data

Plot volume for conventional and organic avocados:



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

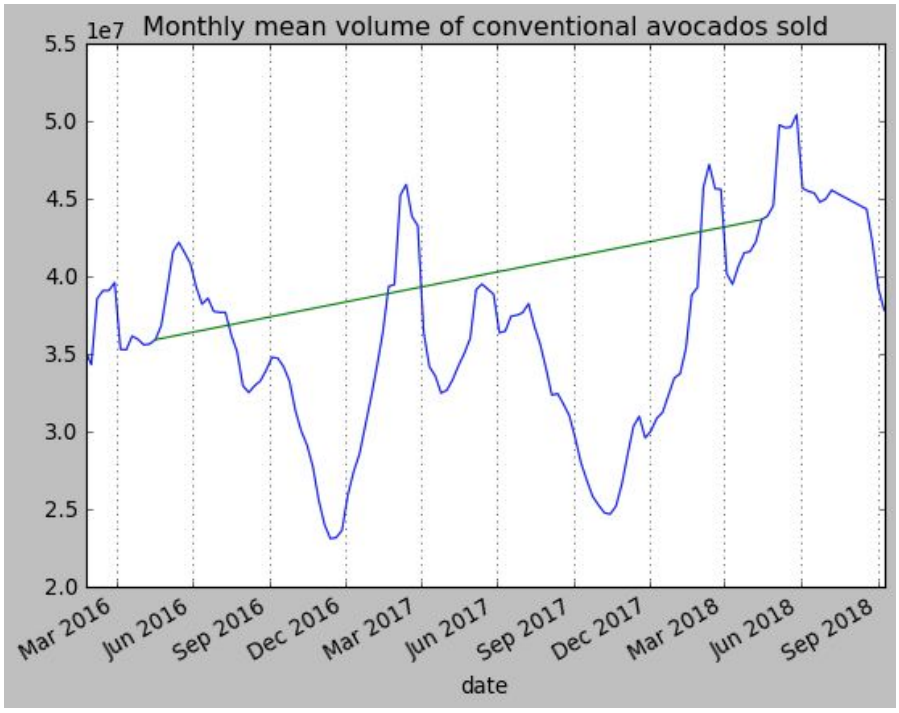
- These peaks appear each year at the same period
- These peaks appear on the plots for each type of avocado

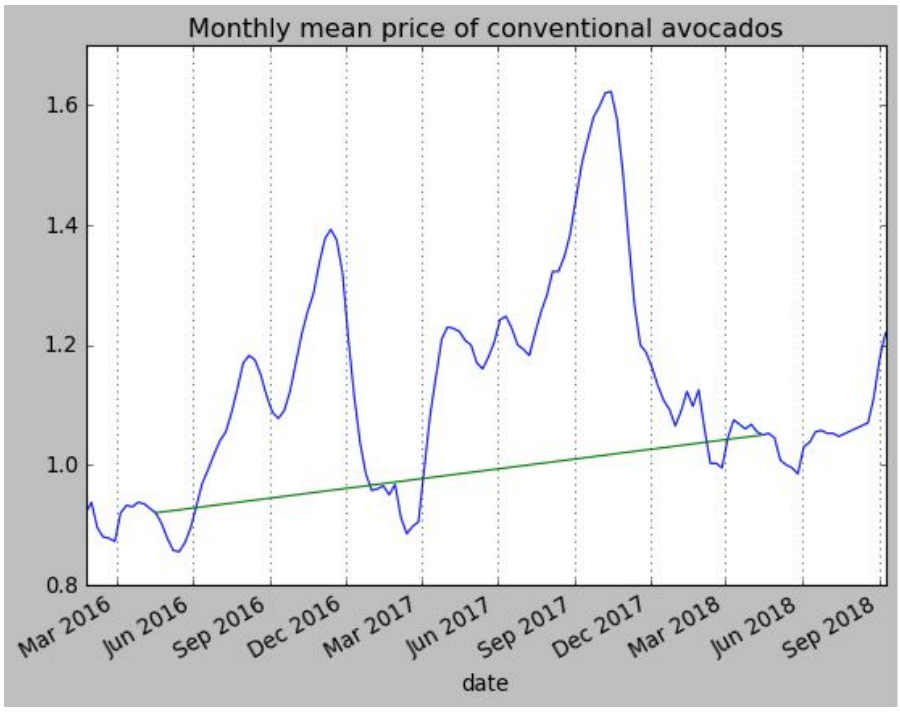
There is a pattern that reproduces every year, with peaks at the beginning of February, and beginning of May.
Are these peaks related to holidays or special events?
Beginning of February is Super Bowl, beginning of May is Cinco de Mayo. However, some holidays don't seem to have the same impact on avocado 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.

I create a holiday 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 event), but the sales are reported at the end of the week, so I think nearest method is more accurate.

Seasonality and trend:

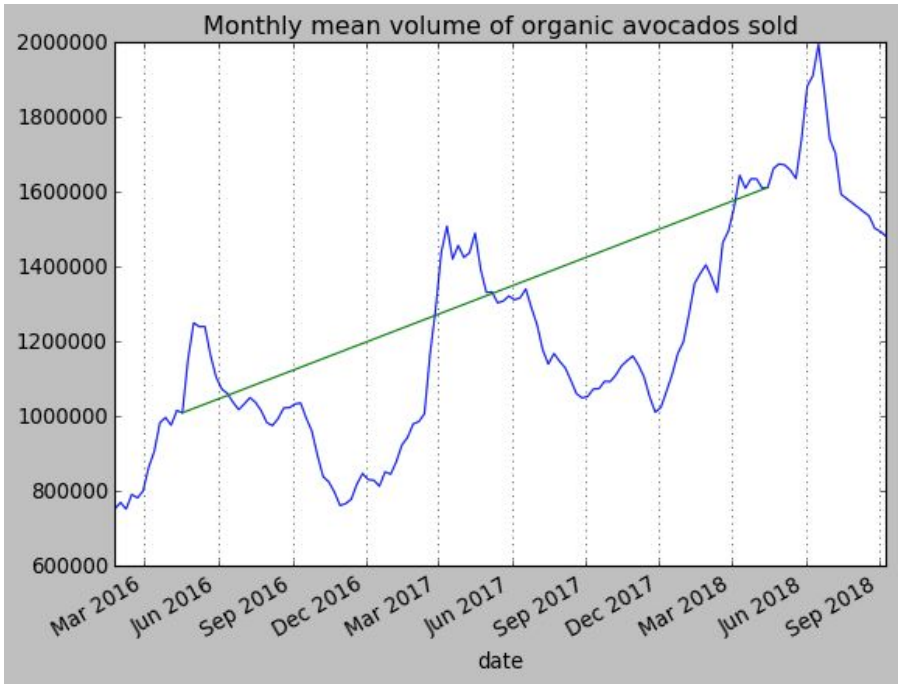
These plots show the monthly averages for volume and price of conventional avocados:
The green line is drawn between the values in April 2016 and April 2018.

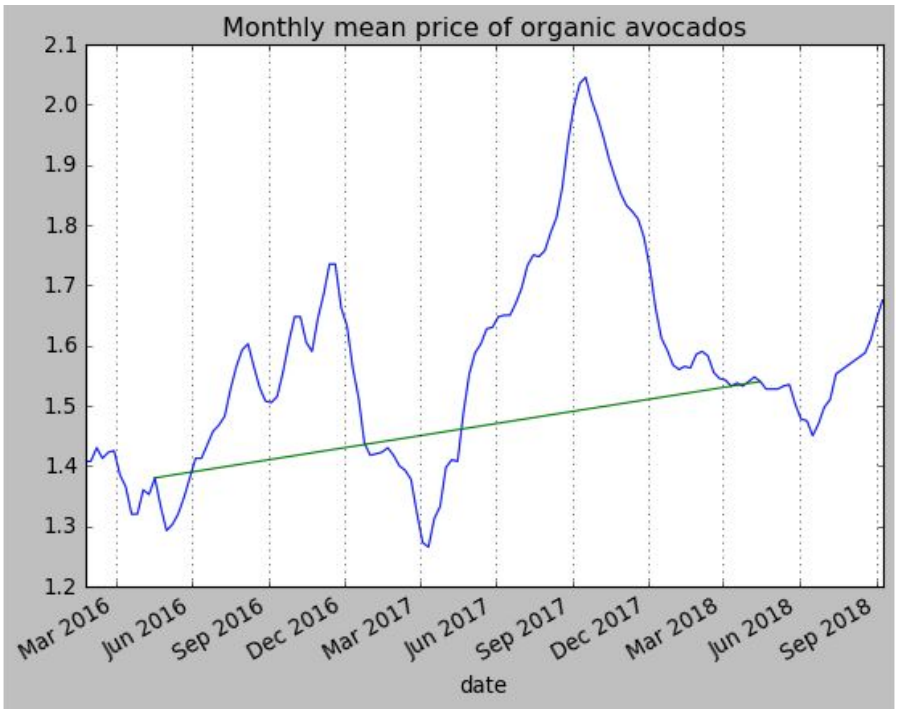




There is seasonality: during winter months, the volume drops and the price raises.
The trend is up, volume and price raise year after year.

Here are the same plots for organic avocados:

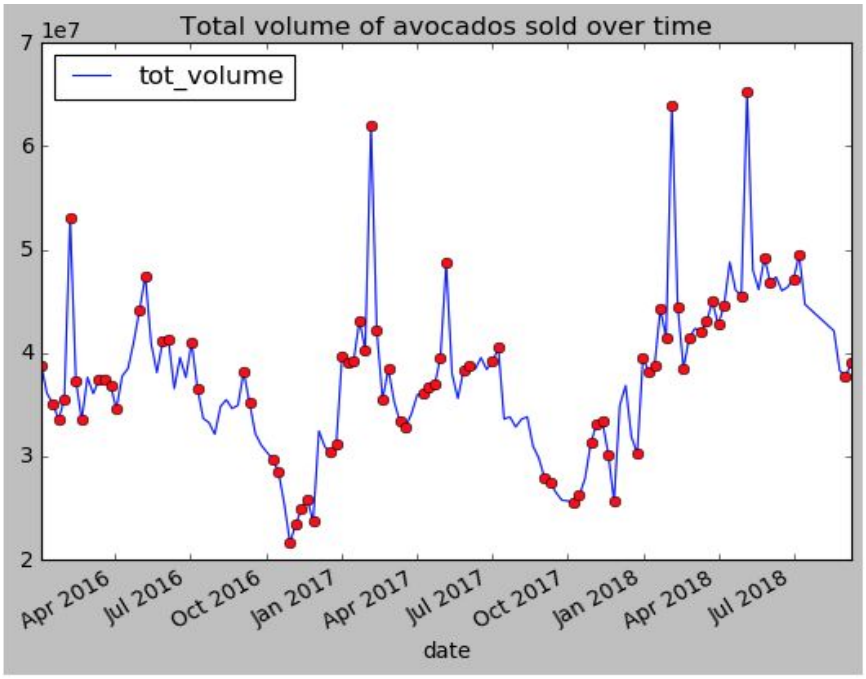




There is seasonality: during winter months, the volume drops and the price raises.
The trend is up, volume and price raise year after year.

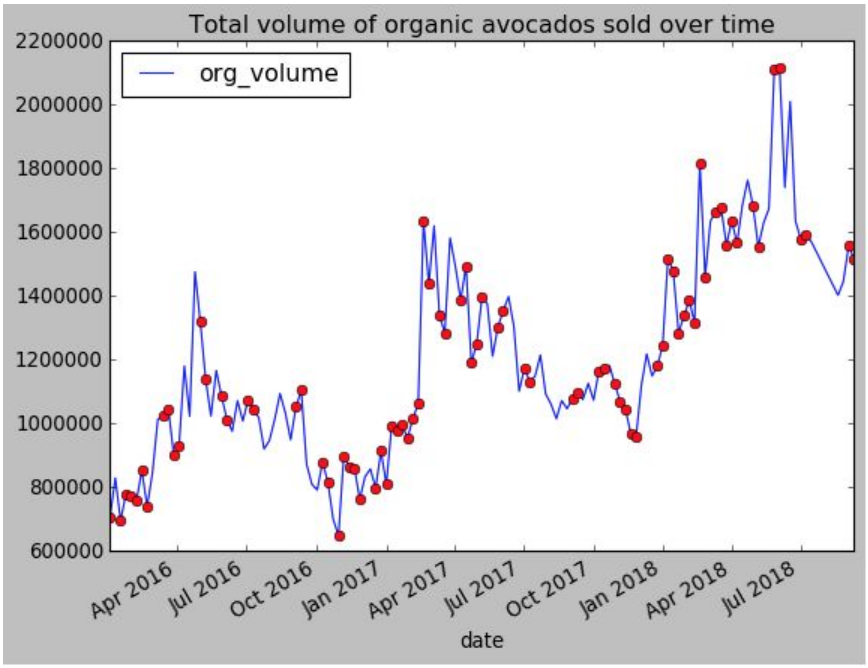
Holiday effect:

In the plot below, the red dots are holidays:



The peaks I see correspond to holidays, so they have an impact on the data. But some holidays seem to have no impact. The highest peaks are Super Bowl, the second highest are Cinco de Mayo.

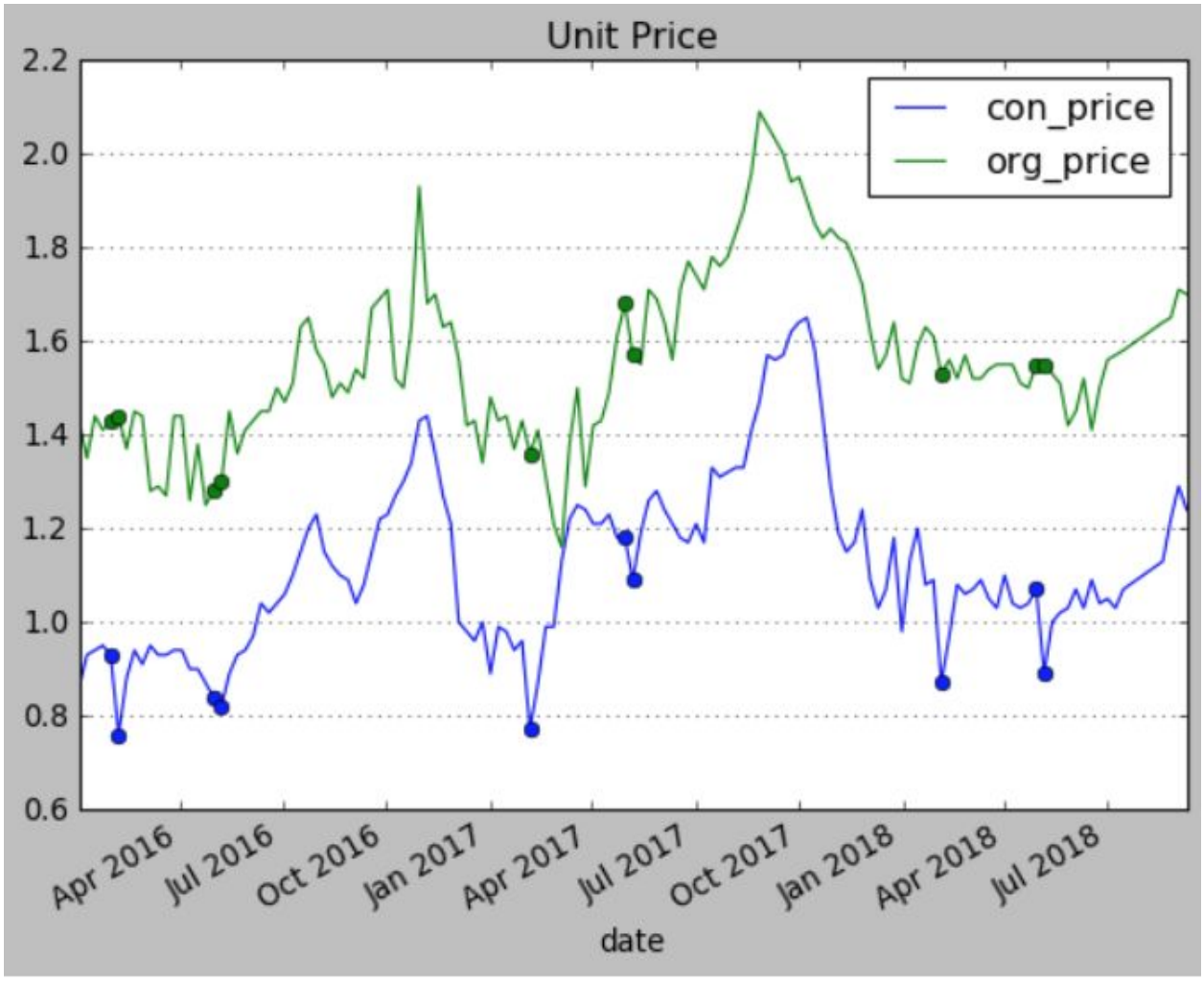
Here is the same plot for organic avocados only:



Organic avocados follow the seasonality (less volume in winter), but they do not follow the holiday pattern as clearly as conventional avocados.

Predicting volume for organic avocados will need a different model from conventional avocados.

Avocado prices:



The price of avocados increases slightly over years, but the price difference between organic and conventional avocados remains consistent over years.

The conventional avocado price is lower for the Super Bowl and Cinco de Mayo, where the volume sold is the highest.

The organic avocado price remains stable.

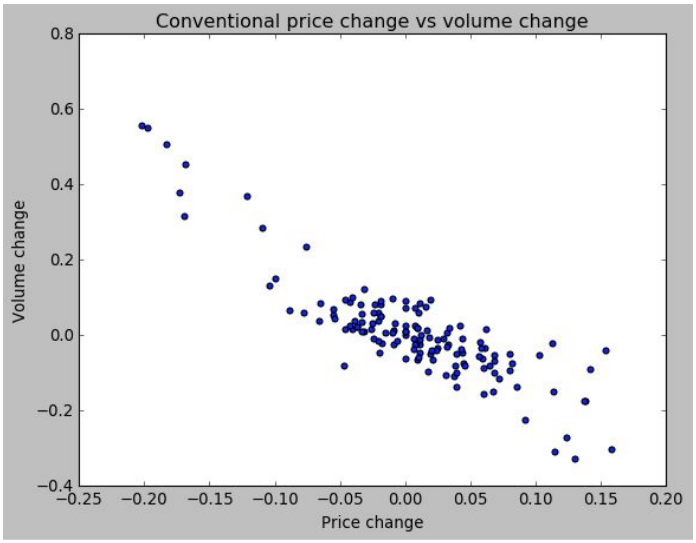
The price of avocados is higher in winter months, this is a good reason for buying less in winter.

3. Analyse Correlations

Price / Volume correlation

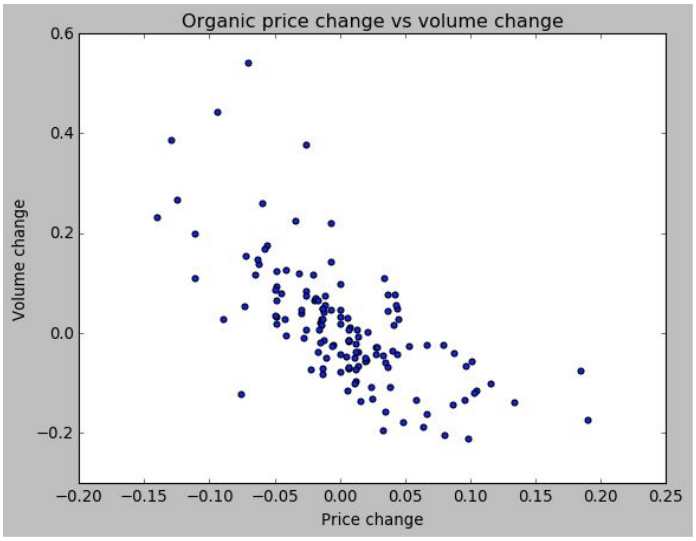
Looking at the previous curves, we can see that price and volume follow the same trend and seasonality.
To validate there is a relationship between them, we analyse the correlation between volume change week to week and price change week to week.

Here is the plot of the change of volume/price for conventional avocados:



The Pearson coefficient is: -0.871, with p_value 3e-43.
The conclusion is that there is strong relationship between volume and price. They vary together in opposite directions.

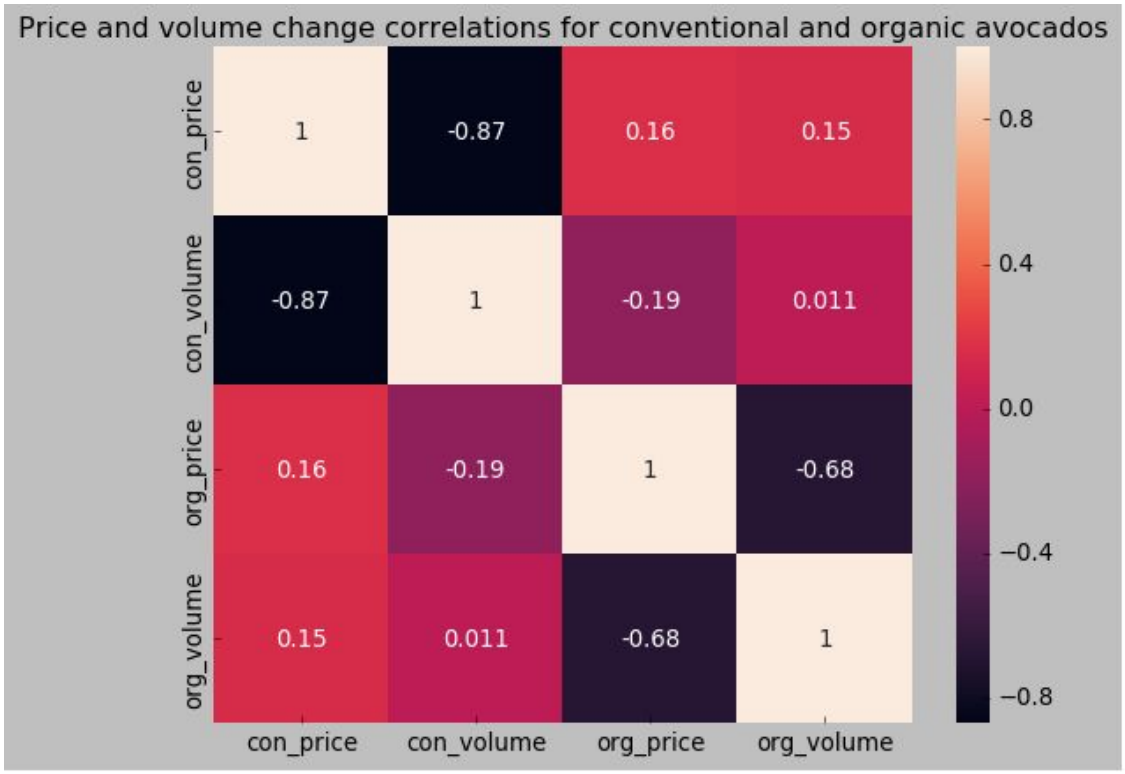
Here is the plot of the change of volume/price for organic avocados:



The Pearson's correlation coefficient is -0.679, with a p_value 1e-19.
This correlation is statistically significant. There is a correlation between volume change and price change for organic avocados. The price and volume vary together in opposite directions.

Conventional/Organic correlation:

Here is the heatmap for price and volume changes correlation:



Although the time series seem to follow the same trend, with the analysis of the changes, we can conclude that there is no evidence that conventional and organic avocados are related.
I think organic and conventional avocados will need different forecasting models.

To conclude:

- Conventional avocados represent the vast majority of avocados sold (36 000 000 in average for conventional and 1 100 000 in average for organic).
- The trend is increasing volume and price, especially for organic avocados volume.
- The volume of avocados sold is influenced by seasons (higher price and less volume in winter), holidays (especially by Super Bowl and Cinco de Mayo).
- Volume and price are related. They vary together in opposite directions.
- Conventional and Organic avocados time series are not related. They will need different forecasting models.

4. Predict the volume of conventional avocados

I will use Facebook Prophet to predict the volume of traditional avocados sold.
FB prophet is a Generalized Additive Model. Its predictions are based on the addition of 3 components:

- trend (models non periodic changes) : we saw that the volume increases over the years, so this evolution should be well modeled by the trend component.
- seasonality (models periodic changes) : we saw that less avocados are sold in winter, so this alternance should be well modeled by the seasonality component
- holidays effect : we saw peaks for some of the holidays and events, especially for the Super Bowl and for Cinco de Mayo. These peaks should be part of the holidays component.

For these reasons, I think FB Prophet should be a good model for my dataset.
Here are the sources for FB Prophet documentation:

https://facebook.github.io/prophet/docs/diagnostics.html

https://peerj.com/preprints/3190.pdf

https://pythondata.com/forecasting-time-series-data-with-prophet-part-1/

I will use years 2015, 2016 and 2017 to train the model, and 2018 to test it.

To evaluate the model, I will use plots and metrics:

- Plot of the predicted and real time series together

- Plot of the 3 FB Prophet components

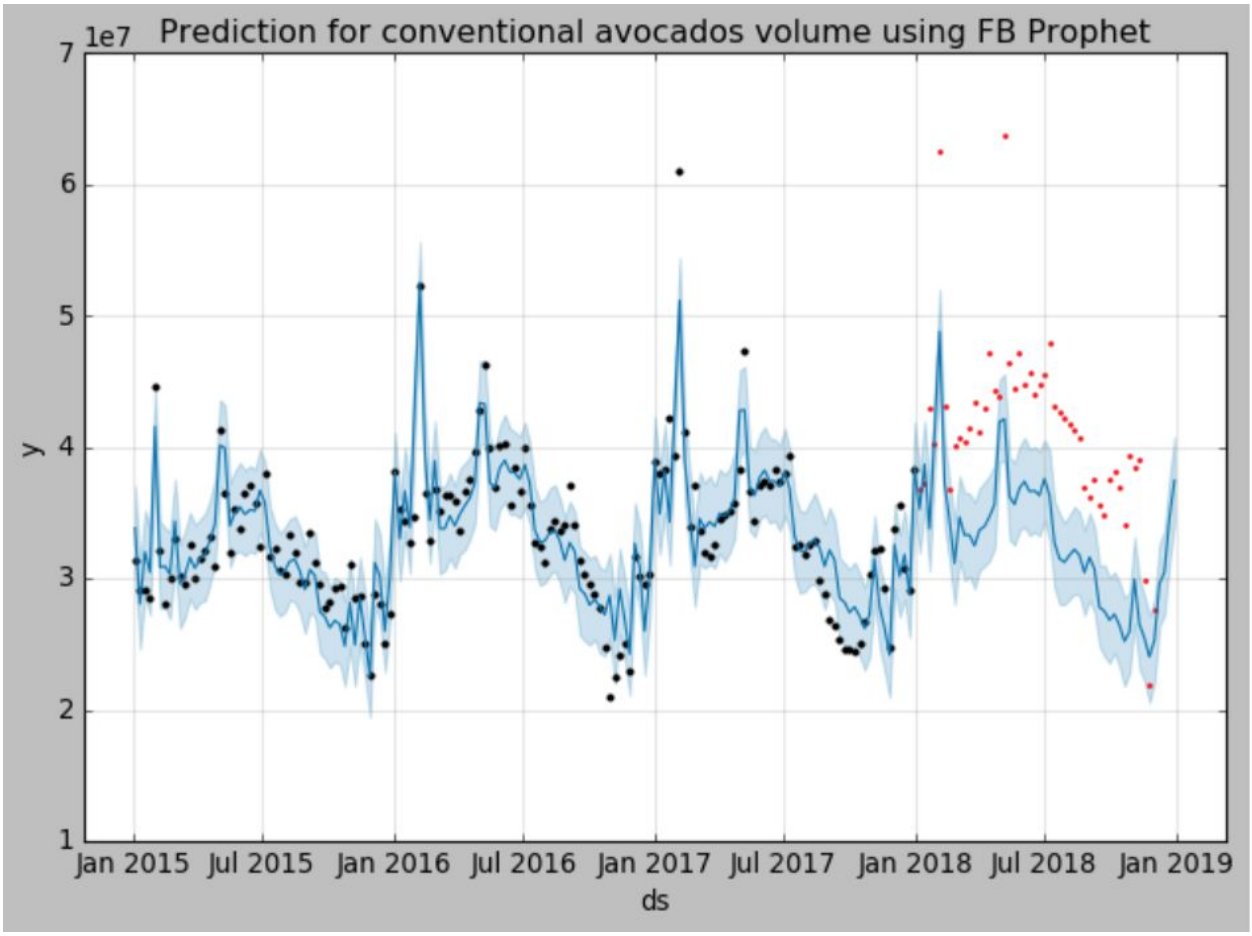
- MAE : Mean Absolute Error = $\frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|$

- MAPE : Mean Absolute Percentage Error = $100 \times \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$

- R-squared

FB Prophet with raw data

First, I tried the FB Prophet model with raw data. Here are the results:



The blue line represents the predicted data with its confidence interval.
The dots represents the real data (black for training data, red for test data).

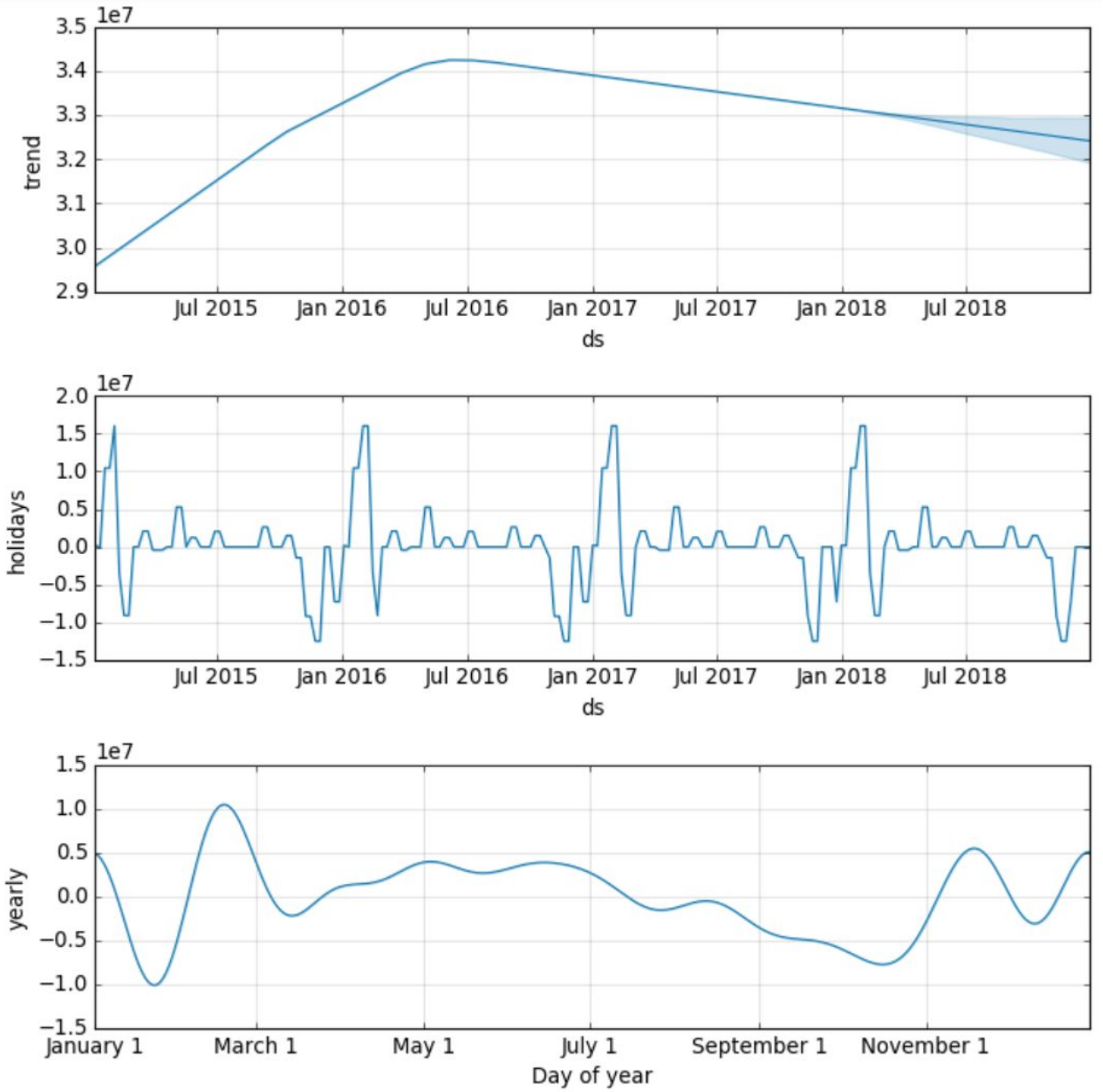
The estimation looks good for the training data, but is under estimated for the test data.
This plot shows a strong increase of the mean volume in 2018, compared with the increases between previous years.
So I would say this is rather an over performance in 2018, than an underestimation.

The metrics are as follow:

	MAE	MAPE(%)	R2
overall score	3.33836e+06	8.96383	0.47739
training score	1.8399e+06	5.71208	0.795903
test score	8.23958e+06	19.5998	-0.787764

The scores reflect what I saw on the plot. The results are good for the training data, but not for the test data.

The components are as follow:



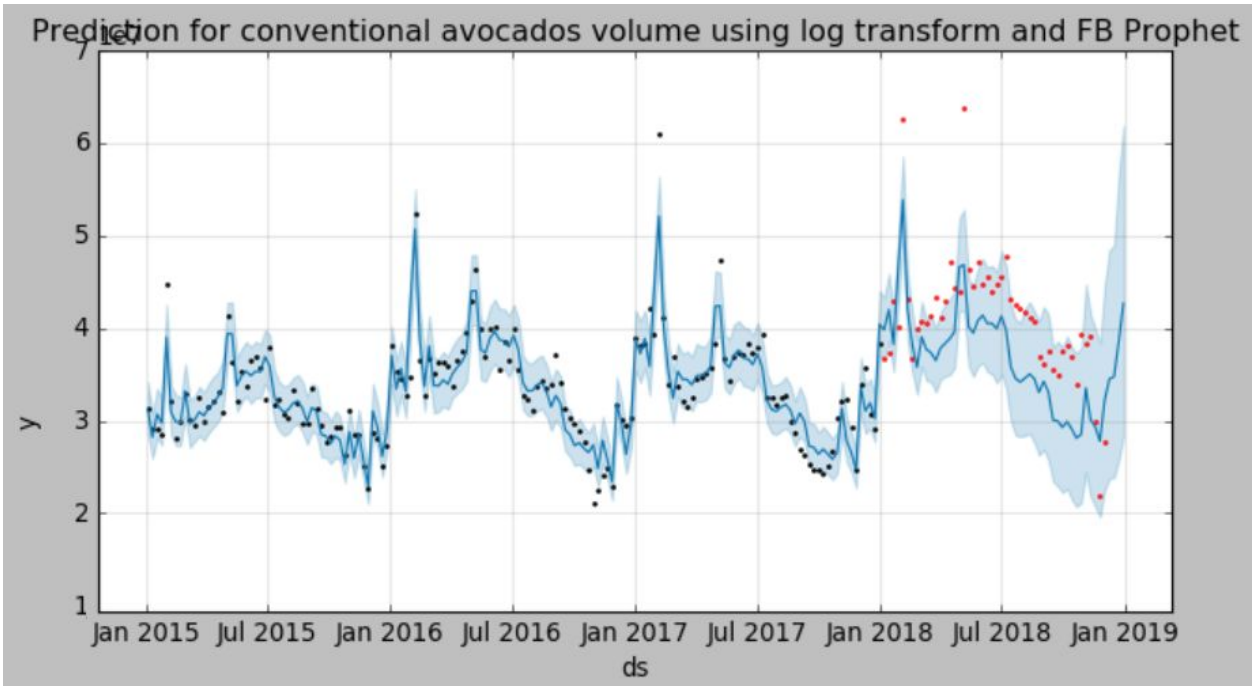
The yearly seasonality shows that the volume goes down in winter.

The holidays shows peaks for Super Bowl and Cinco de Mayo. However, the volume is underestimated for these events. I could try to give more weight to Super Bowl and Cinco de Mayo.

The trend shows that the model did not see the 2018 increase, which is normal because it was not in the training data.

FB Prophet with log transform

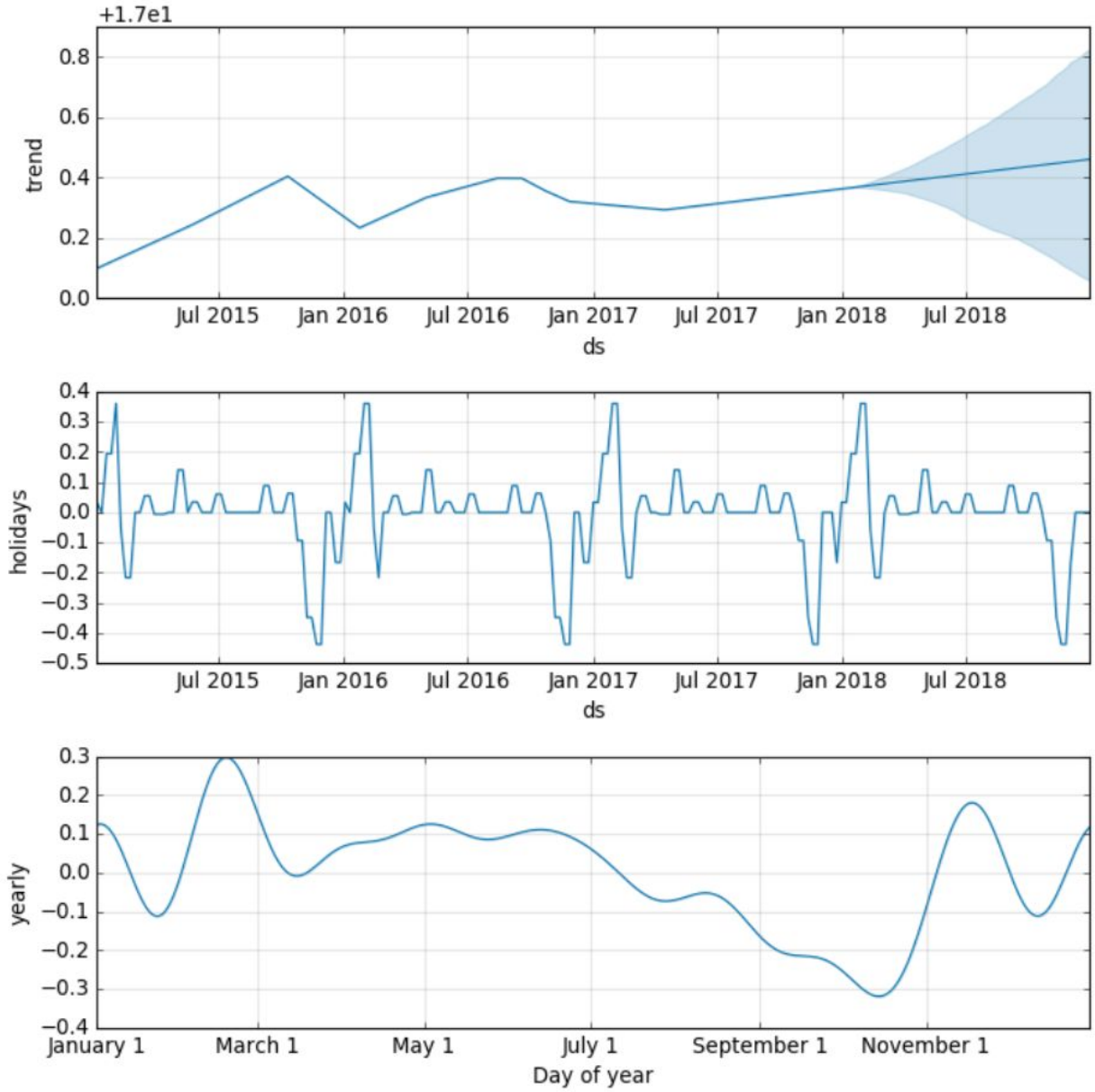
I tested the time serie for stationarity using Augmented Dickey Fuller test and I got ADF score -3 with p_value 0.03.
So I the time serie is stationary.
However, I will log transform the data before sending it to FB Prophet, to model multiplicative changes instead of additive ones.
Here are the results:



The results seem better. The changes in the volume are better predicted with a multiplicative model.
The metrics are as follow:

	MAE	MAPE(%)	R2
log transform overall score	2.54717e+06	6.94874	0.722405
log transform training score	1.63459e+06	4.96443	0.839811
log transform test score	5.53205e+06	13.4391	0.164596

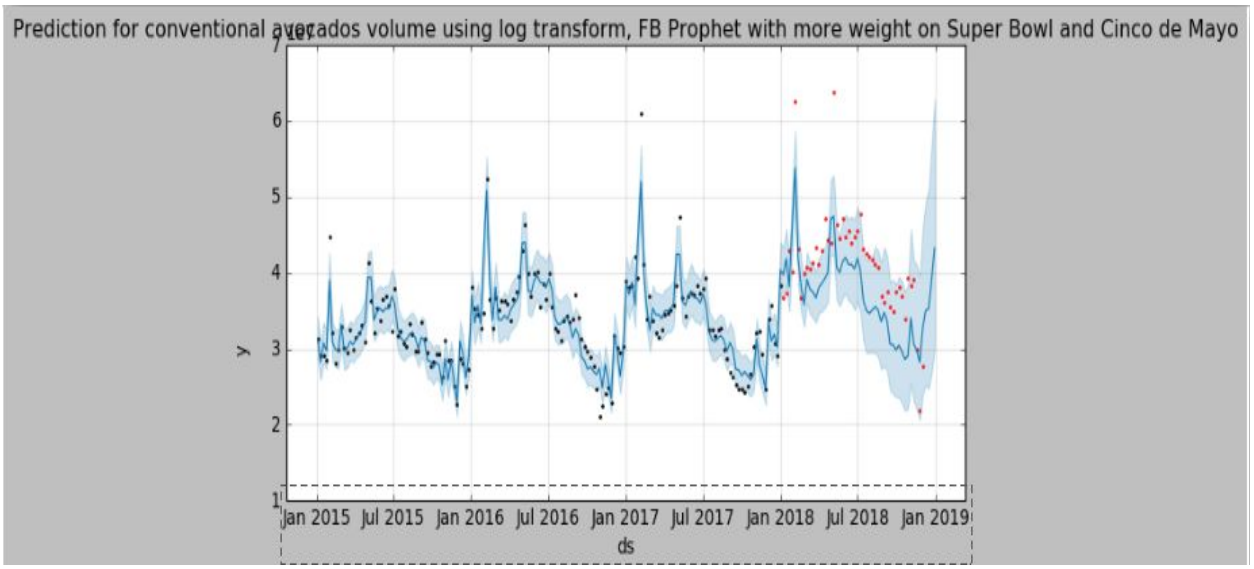
The scores are much better !



The trend component is going up for 2018.

FB Prophet with log transform tuning

I will now try to give more weight to Cinco de Mayo and the Super Bowl.



	MAE	MAPE(%)	R2
holiday prior overall score	2.48101e+06	6.80373	0.74149
holiday prior training score	1.6502e+06	5.01521	0.837426
holiday prior test score	5.19843e+06	12.6537	0.251581

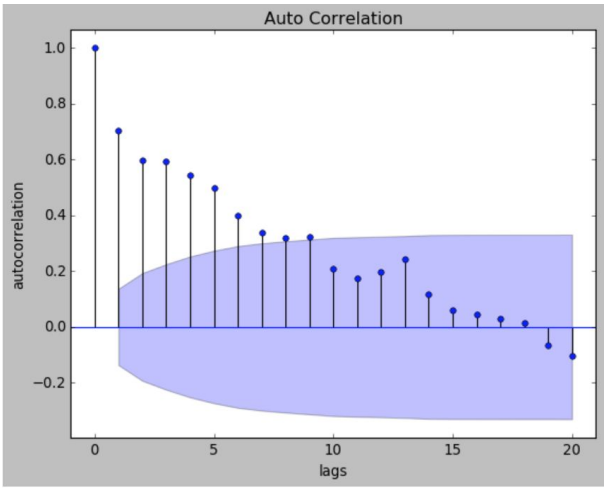
The scores have improve. We now have a MAPE of 12.6% for a horizon of up to 48 weeks.

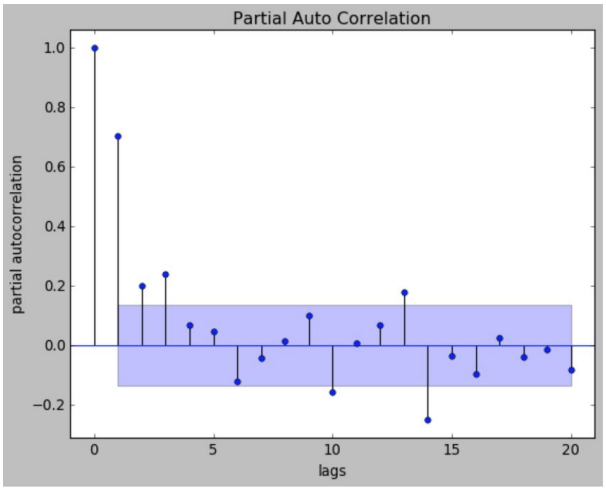
We could continue to fine tune the model to get better results, for example by tuning the Fourier order of the seasonality, tuning the seasonality prior, and investigating the drop in November/December.

Compare with Auto Regressive model

How does this FB Prophet compare with a plain Auto Regressive Moving Average model?

First, I plot the Auto Correlation and Partial Auto Correlation to determine the best parameters.

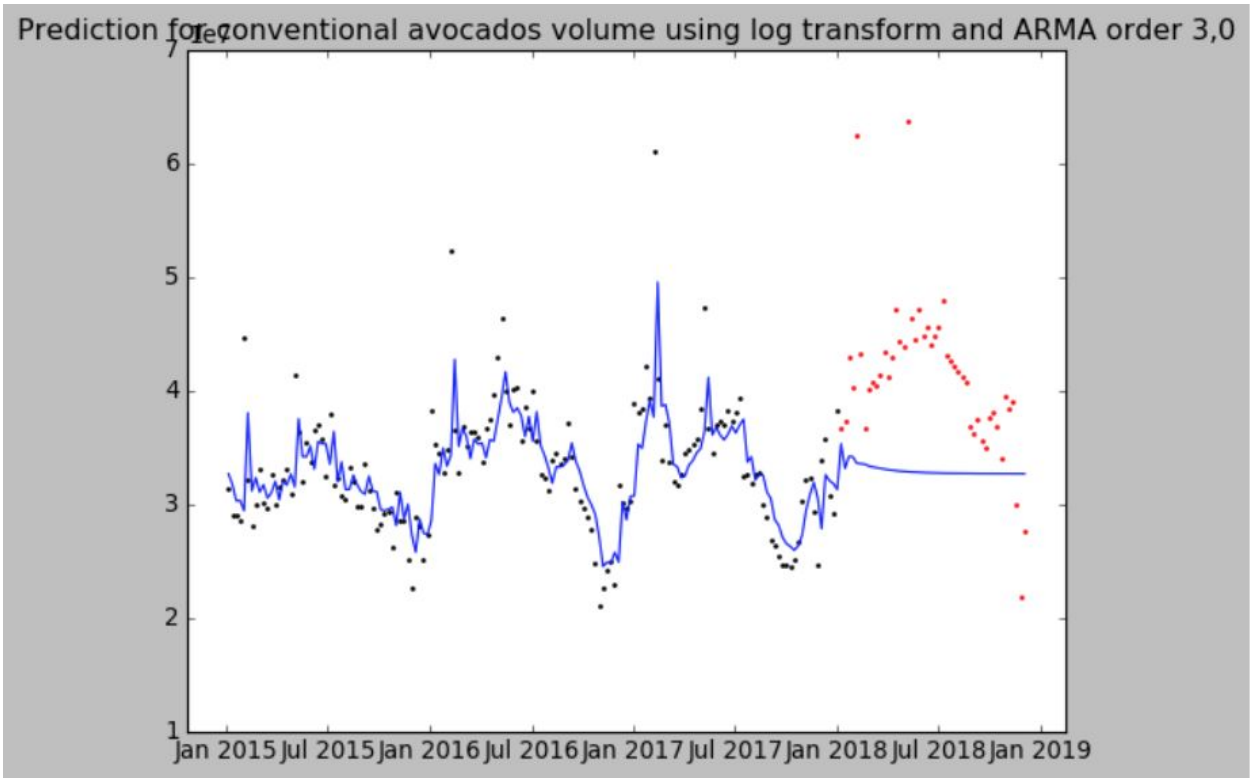




ACF shows a strong correlation up to lag 3, then trailing off. PACF shows a strong correlation up to lag 3, then no correlation. So I will model an Auto Regressive model of order 3.
Here are the metrics:

	MAE	MAPE(%)	R2
AR overall score	4.20479e+06	10.9873	0.149696
AR training score	2.751e+06	7.98258	0.459339
AR test score	8.95991e+06	20.8153	-1.4487

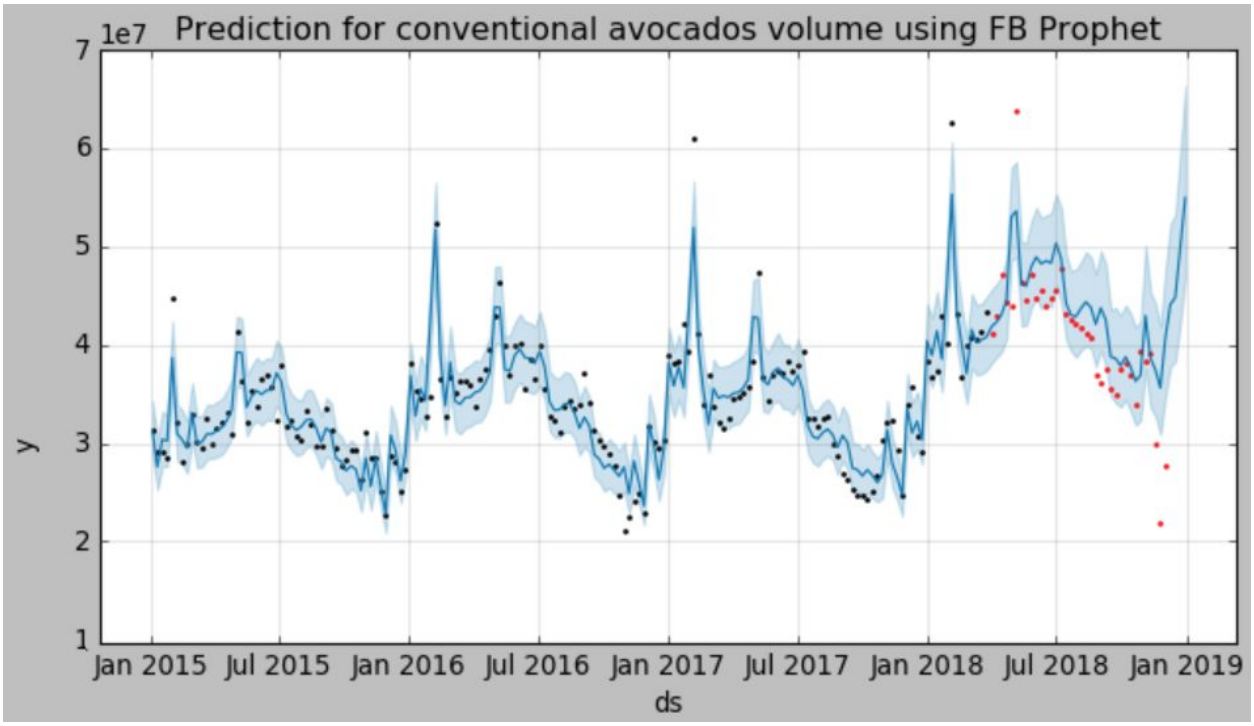
And the plot:



The prediction is not good for long horizon like we can see for test data.

More training data, shorter horizon

I would like to train my model with data points in 2018, because we saw that 2018 over performed.
Was this trend predictable with training data up to the first 3 months of 2018?
I will split the dataset on 2018-03-31.
Here is the plot:



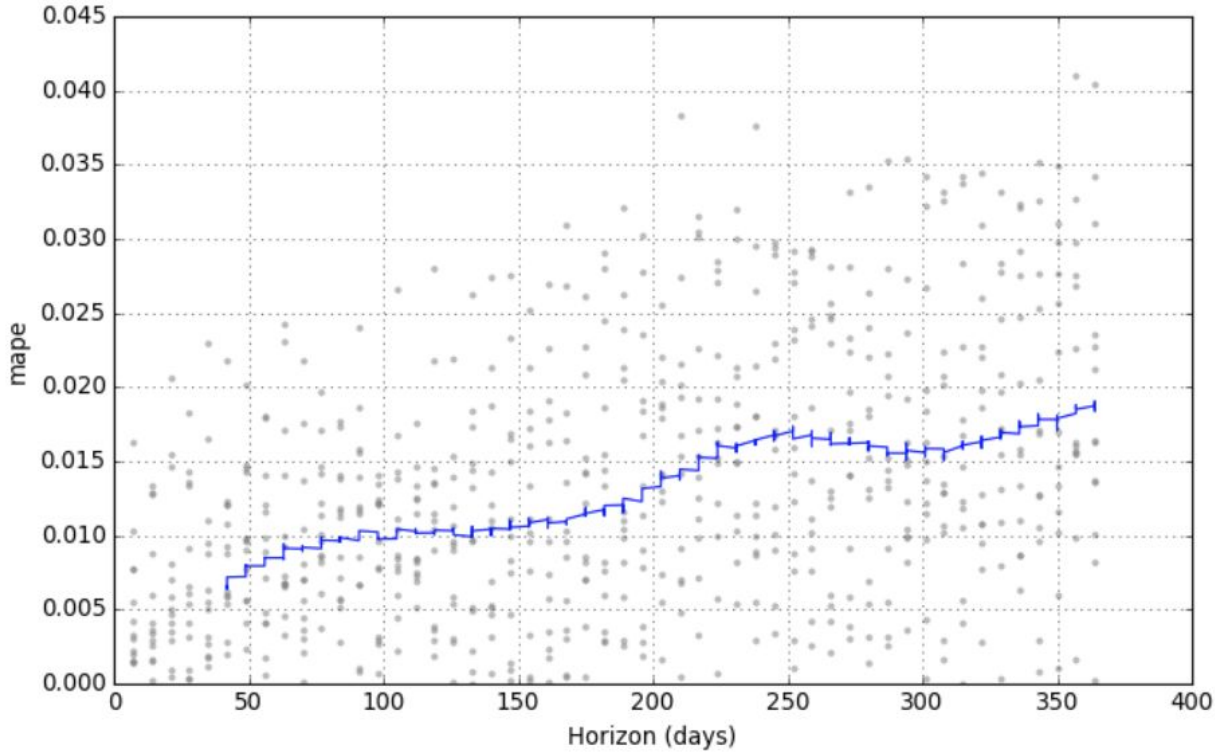
Here are the metrics:

	MAE	MAPE(%)	R2
cut off March overall score	2.13856e+06	6.23084	0.799252
cut off March training score	1.82788e+06	5.4222	0.841601
cut off March test score	3.59704e+06	10.027	0.462865

The scores are much better for the test data.
We have a mean absolute percentage error of 10% with an horizon of up to 36 weeks.

Use Simulated Historical Forecast to plot mape score vs horizon:

I will now use simulated historical forecast and the cross validation tools of FB Prophet to calculate the average scores for different horizons.
I will use an initial training period of 104 weeks (2 years), then make 13 forecasts with cutoffs between 2017-01-01 and 2017-12-03.



The predictions Mean Absolute Percentage error remains under and around 10% up to 22 weeks. The Mean Absolute Percentage Error is less than 20% with a 52 week horizon.

To conclude:

- This study showed that the avocados sales have over performed in 2018, compared to previous years. Congratulations to all avocado professionals!
- FB Prophet is a good model for this time serie. It gives better results when log transforming the data first.
- The model can be more fine tuned by working on seasonality component, and november/december holidays.
- The model error is around 10% up to 22 weeks, and below 20% with a 52 week horizon.
- The model can certainly help avocados retailers, wholesalers and importers to forecast their activity.