Forecasting Avocado Prices

07.10.2020

Bei Wang

Overview

Avocado is known as 'Green Gold' in world trade. Avocados are a great source of vitamins C, E, K, and B-6, as well as riboflavin, niacin, folate, pantothenic acid, magnesium, and potassium. They also provide lutein, beta-carotene, and omega-3 fatty acids. Although most of the calories in an avocado come from fat, but that fat is very healthy, especially Hass avocados.

Hass avocados are higher in fat than other varieties, which gives them a richer taste and smoother, creamier texture. The taste, benefit of health and food culture make avocados become an important food/fruit to people around the world. It is the main daily food in South America. In most recent years, the avocado sales market is increasing in Asia as well.

Since the avocados market is important to the world, I'd like to forecast avocado prices in the future. I am going to use the dataset of US sales reports from 2015 to 2018 to implement the research.

Goals

1. Analysis of avocados price changes from 2015 to 2018.
2. Predict the avocados price in the future.

Methodology

This is a time series problem:

Theses are classic time series models:

1. Autoregression (AR)
2. Moving Average (MA)
3. Autoregressive Moving Average (ARMA)
4. Autoregressive Integrated Moving Average (ARIMA)
5. Seasonal Autoregressive Integrated Moving-Average (SARIMA)
6. Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX)
7. Vector Autoregression (VAR)
8. Vector Autoregression Moving-Average (VARMA)
9. Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX)
10. Simple Exponential Smoothing (SES)
11. Holt Winter's Exponential Smoothing (HWES)

12. Exponential Moving Averages

There are more recent approaches to time series and I am gonna use prophet, ARIMA and EMA to compare the result.

Date and description

I use avocados price dataset including the price from 2015 to 2018, Time Series will be the basic technique type in the project.

Data source: https://www.kaggle.com/neuromusic/avocado-prices

Some relevant columns in the dataset:

- Date - The date of the observation
- AveragePrice - the average price of a single avocado
- type - conventional or organic
- year - the year
- Region - the city or region of the observation
- Total Volume - Total number of avocados sold
- 4046 - Total number of avocados with PLU 4046 sold
- 4225 - Total number of avocados with PLU 4225 sold
- 4770 - Total number of avocados with PLU 4770 sold

Data Wrangling

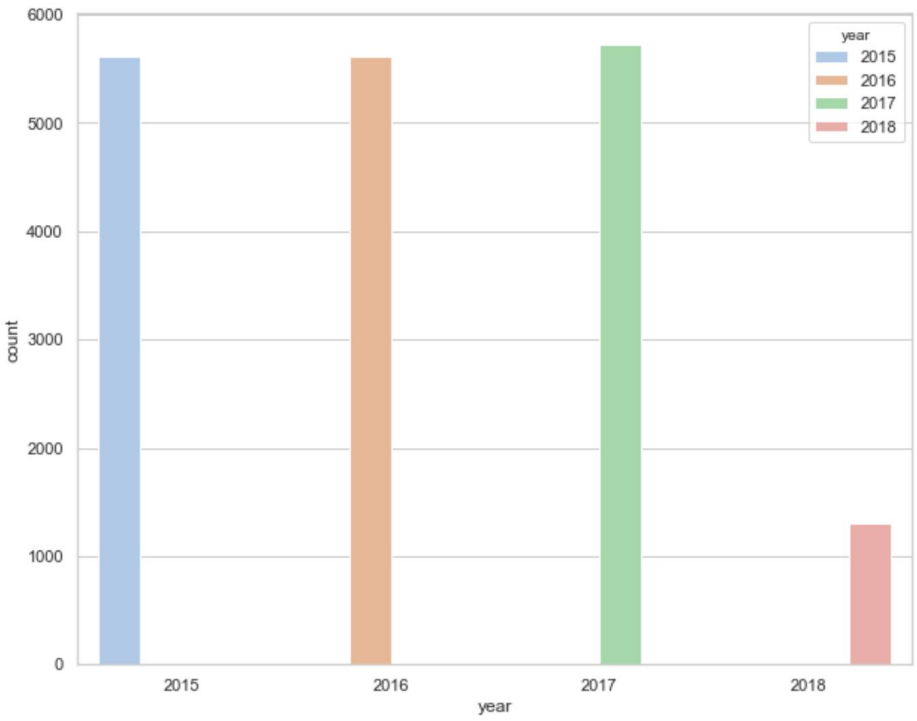
Dataset is quite clear, no null data or missing data.
Sort the dataset in a new csv file.

Exploratory Data Analysis

The data will display the answers of our concerned, and it also will bring more questions after that. So let's see what the data tells us from a macro perspective. And then analyze it in detail at the micro level.

Overview of the Sales

At the beginning, let' s check the distribution of sales in each year:

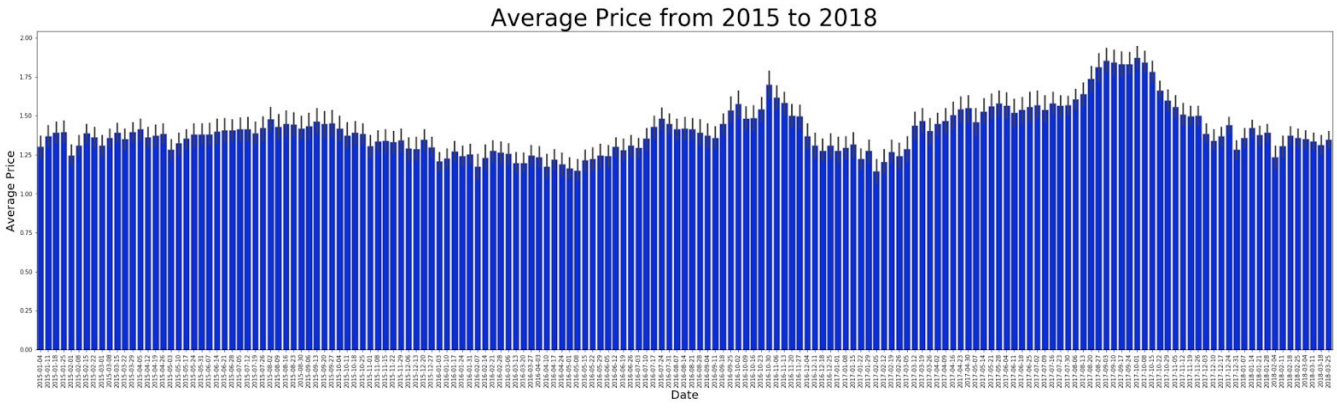


The counts of 2015 to 2016 are similar, 2016 reached the highest point. 2018 only collected 3 months of data, so it's short.

Overview of the Average Price

Let's start to find stories in the data.

First, I am going to check the average price from 2015 to 2018.

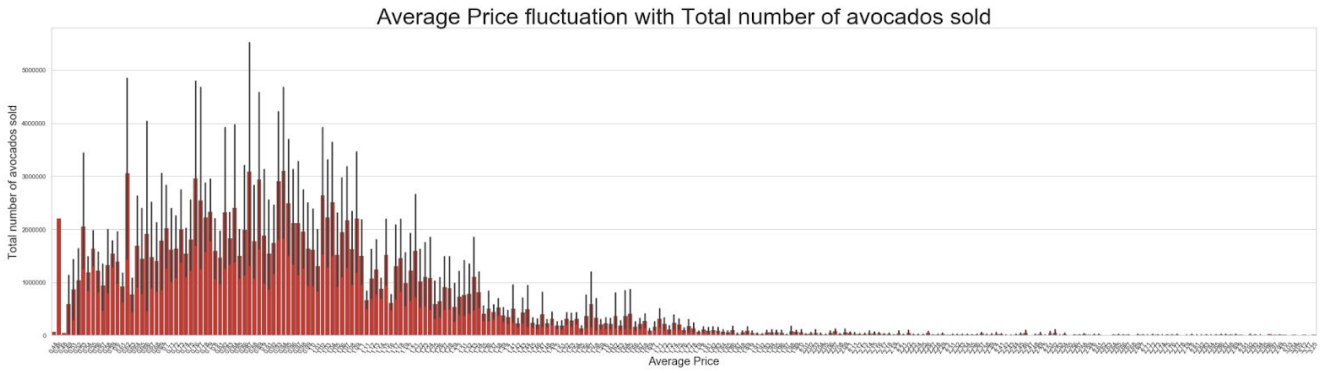


Throughout 2015, prices were relatively stable.

From 2016 to 2018, prices showed an upward trend,especially in September 2017, it reached the peak of four years.

Price Affect Sales Volume

So, does the price fluctuations affect sales volume?



Sales volume does affect price fluctuations: low prices bring high sales, and high prices bring low sales. But there are some exceptions.

For example: price as low as $0.44 and $0.48 cannot bring a higher sales:

	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
12214	2017-03-05	0.44	64057.04	223.84	4748.88	0.0	59084.32	638.68	58445.64	0.0	organic	2017	CincinnatiDayton

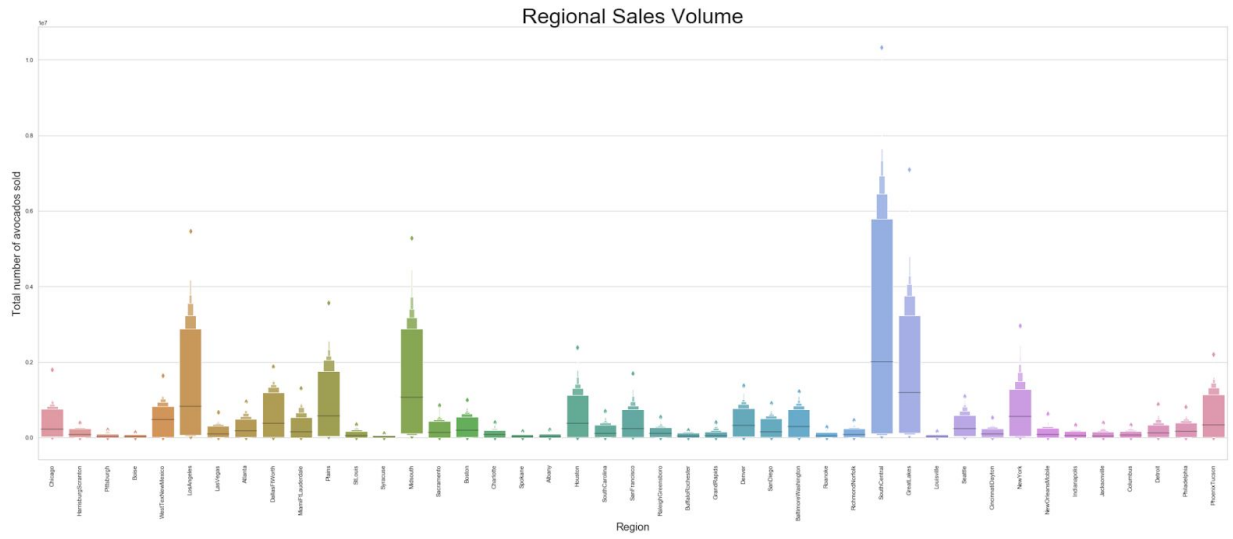
	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
12247	2017-03-05	0.48	50890.73	717.57	4138.84	0.0	46034.32	1385.06	44649.26	0.0	organic	2017	Detroit

Two lower Prices with lower sales occurred in Cincinnati Dayton and Phoenix Tucson, I cannot find any news related to the local avocado retail market, which may be caused by special circumstances.

Regional Sales

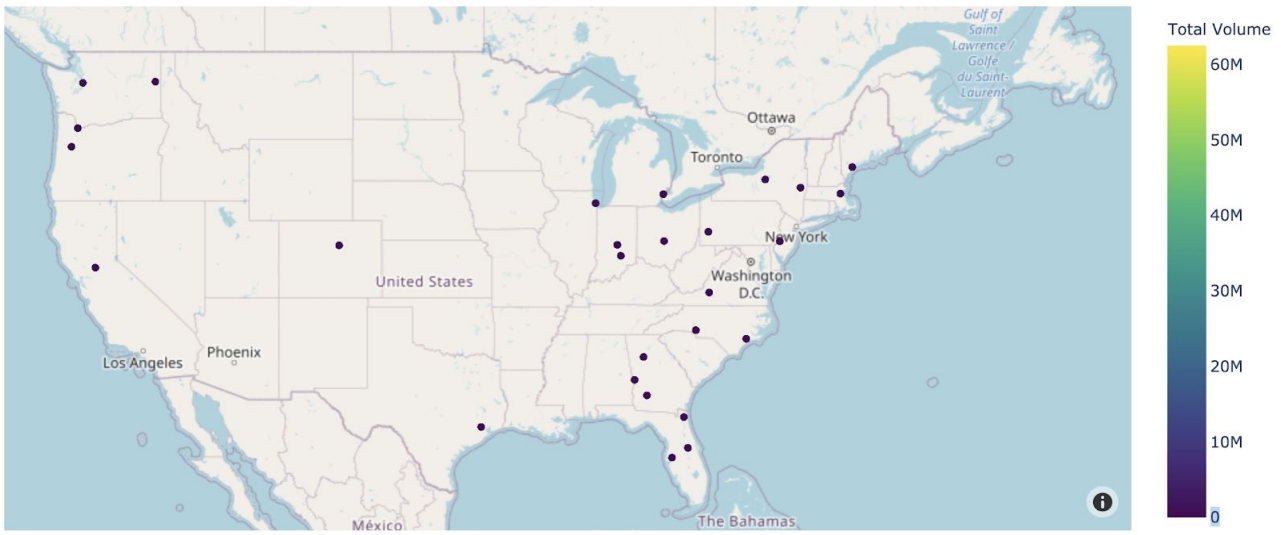
In the dataset, avocados were sold in most cities of the USA. How is the sales volume in each city? Where is the best sales market of avocados?

.



South Central is the topest sales city. Cities with better sales are distributed widely, with a certain proportion in the Midwest and Northeast.

Cities' sales volumes distribution in the map will be clearly:

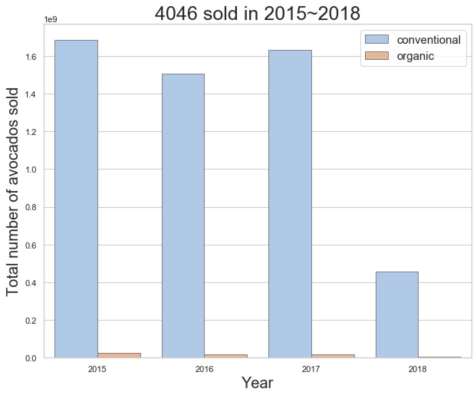


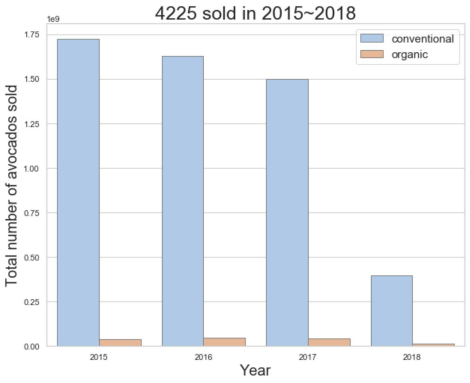
In the map, It can be clearly seen on the map that sales markets are mainly concentrated in the Northeast and West coast. The reason may be related to the popularity of local food and the dense population of eating.

Compare Sales Volume of 3 Avocados of PLU

Next, I am going to check the sales situations of organic and non-organic products in 3 avocado's products with PLU codes: 4046, 4335 and 4770.

Now draw plots to compare them.



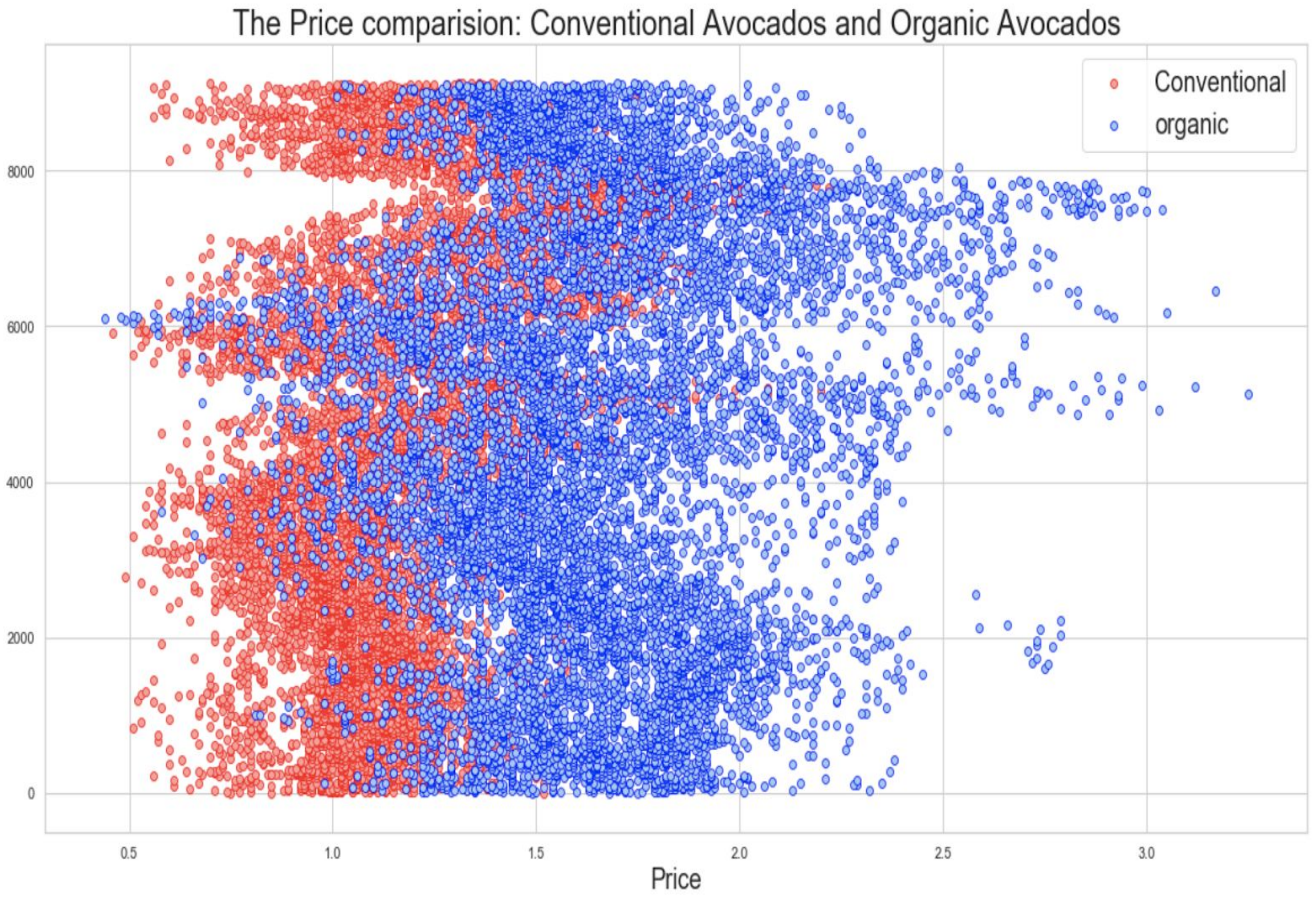


No matter what kind of avocado, the sales volume of conventional(non-organic) avocados is much higher than organic sold.

Compare Price Difference Between Organic and Non-Organic

Is that the price caused by the difference of the sales volume between non-organic and organic?

Let's go dig in the price difference of non-organic and organic.



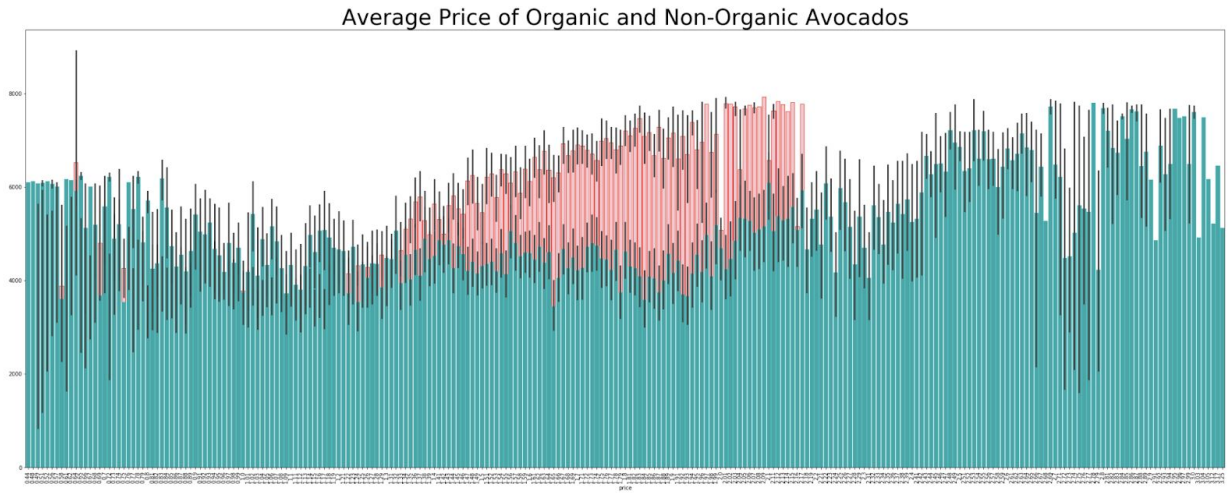
Although the chart shows that the price of most organic avocados is higher than that of non-organic avocados, it is not difficult to find that the prices of the two kinds of avocados have a large overlap, and both have reached the highest and lowest prices.

Therefore, can I determine that is the price a certain factor that leads to product sales?

I will do a hypothesis test for it later.

Hypothesis Testing

Before doing hypothesis testing, I'd like to overview the average price difference of organic and non-organic in bar plot again:



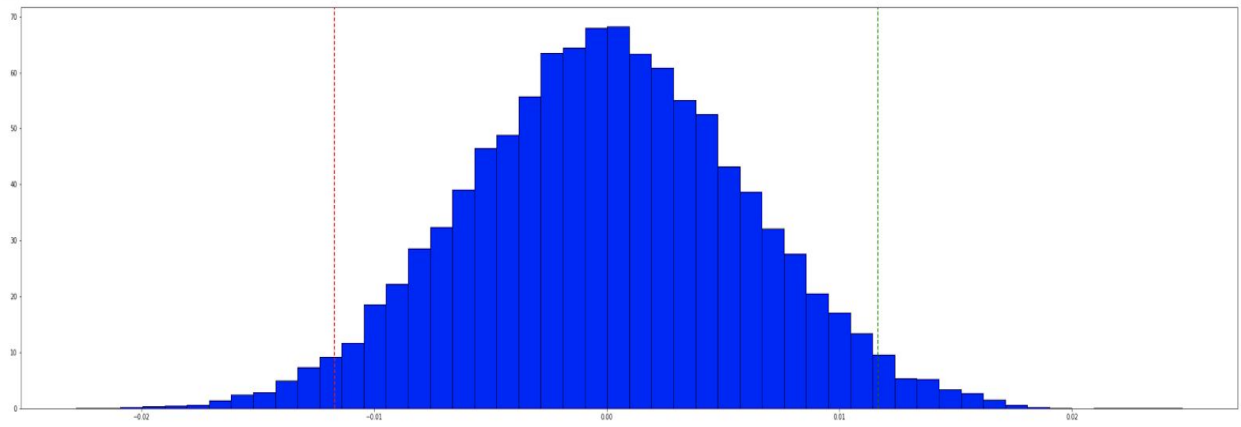
In this chart, it is obvious that both organic avocado and non-organic avocado have reached the highest point in price, so is price still the guiding factor for the two types of sales?

Let me do hypothesis testing to verify it.

H0: There is a difference in average price between organic and conventional.

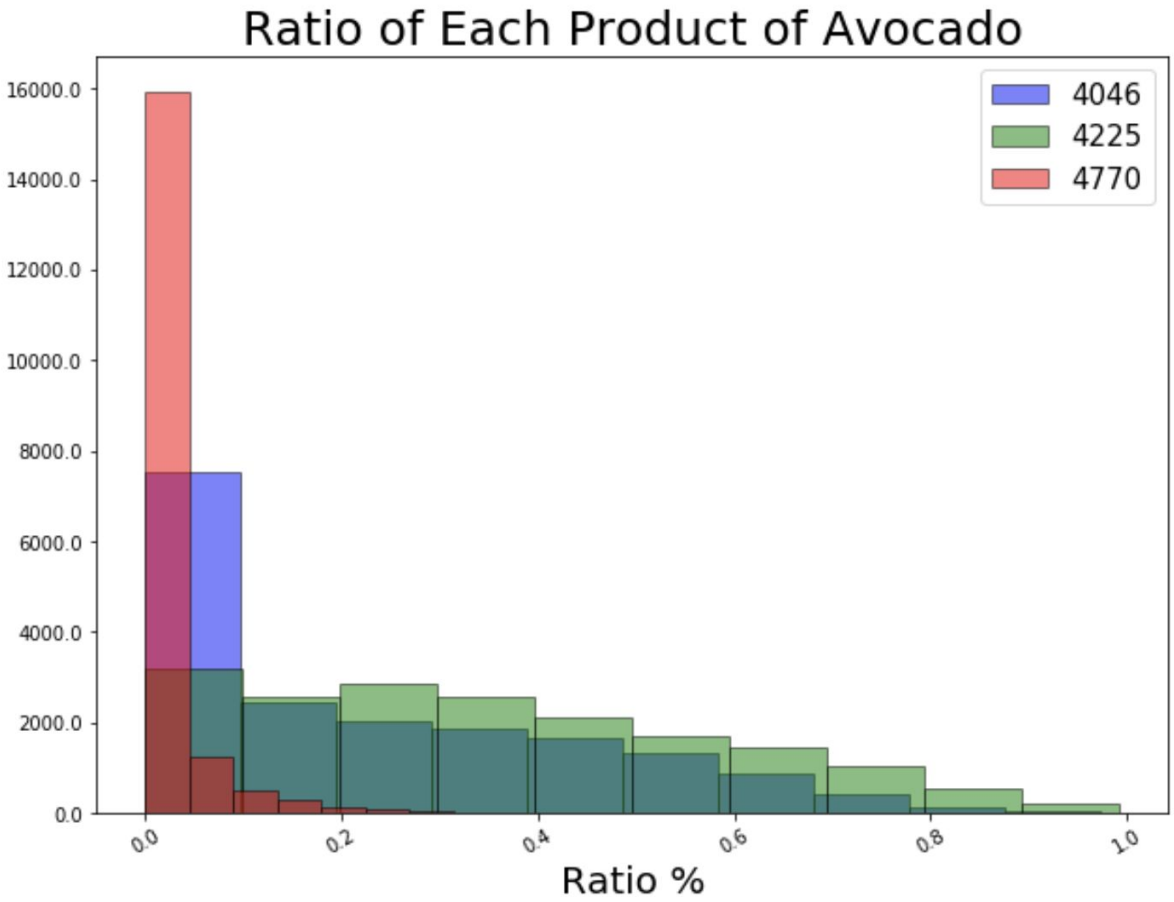
H1: There is no difference in average price between organic and conventional.

After calculation, I got the P value is 0.00, when α=0.05, p = 0.00 indicates that there is no actual difference price between organic and conventional. H0 rejected.



Ratio of 3 Types of Avocados Sale

Think about 3 products of avocados with different PLU, what is the ratio of each product sale of total sales?



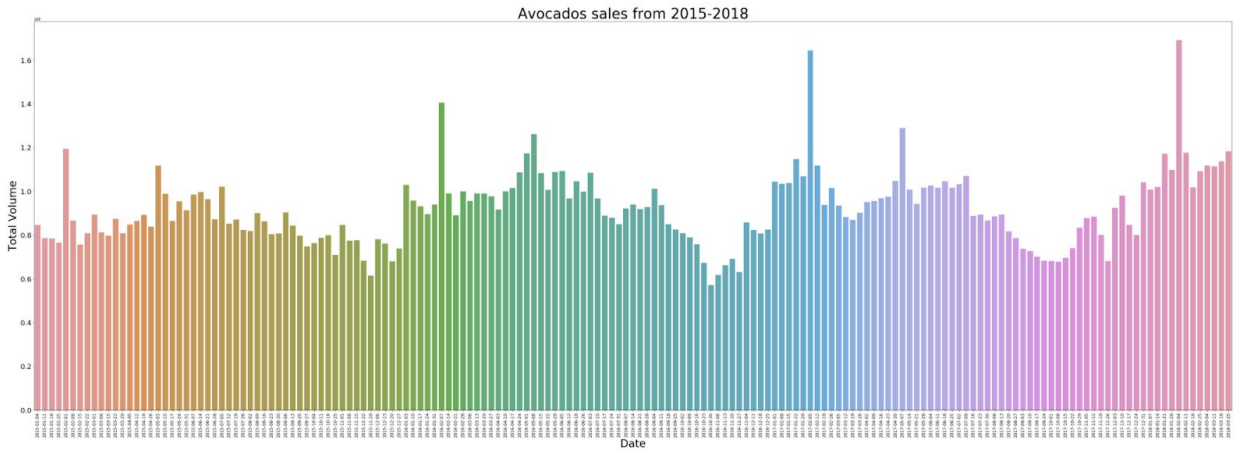
In most of time, 4770 accounted for 0.5% of total sales.

4046 and 4225 accounted average sales of total sales, ranging from 0.1% to 0.89%.

Sales During the Specified Period

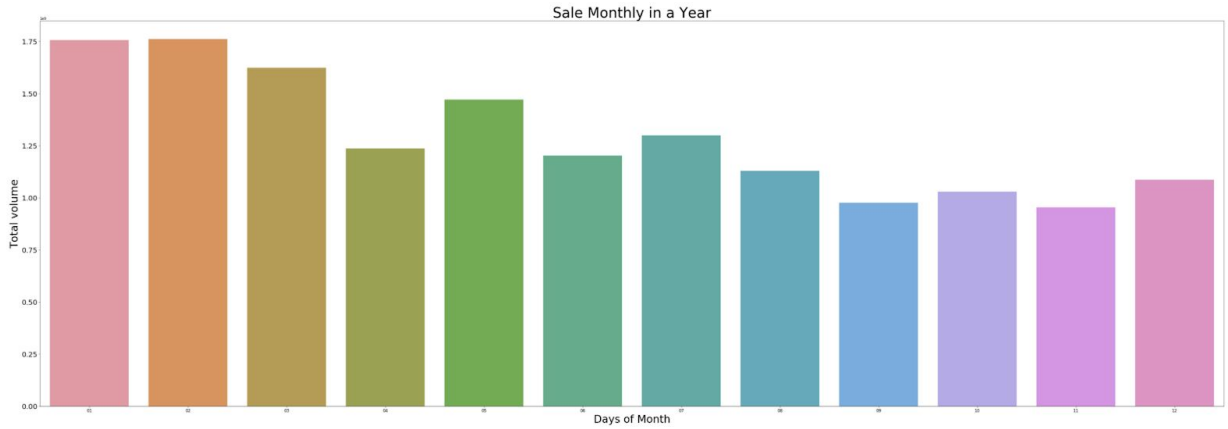
If avocado sales fluctuate by seasonal? To achieve this, I am going to do research by monthly sale during a year and daily sale during a month. Since all data collected by each Sunday, I will not do weekly sales analysis here.

Overview the sale



In this plot,total sales are relatively average except for a few days in February.

Sales during a month



In the monthly plot, we can easy to find the highest sales volume happen in first 3 months of the year, while yearly sales are slowly declining in waves.

In the above, I have obtained corresponding analysis results by analyzing prices and sales volume, sales volume of products by category, organic and non-organic sales volume and price, and combining four-year data to sales volume within one year.

Forecasting Result

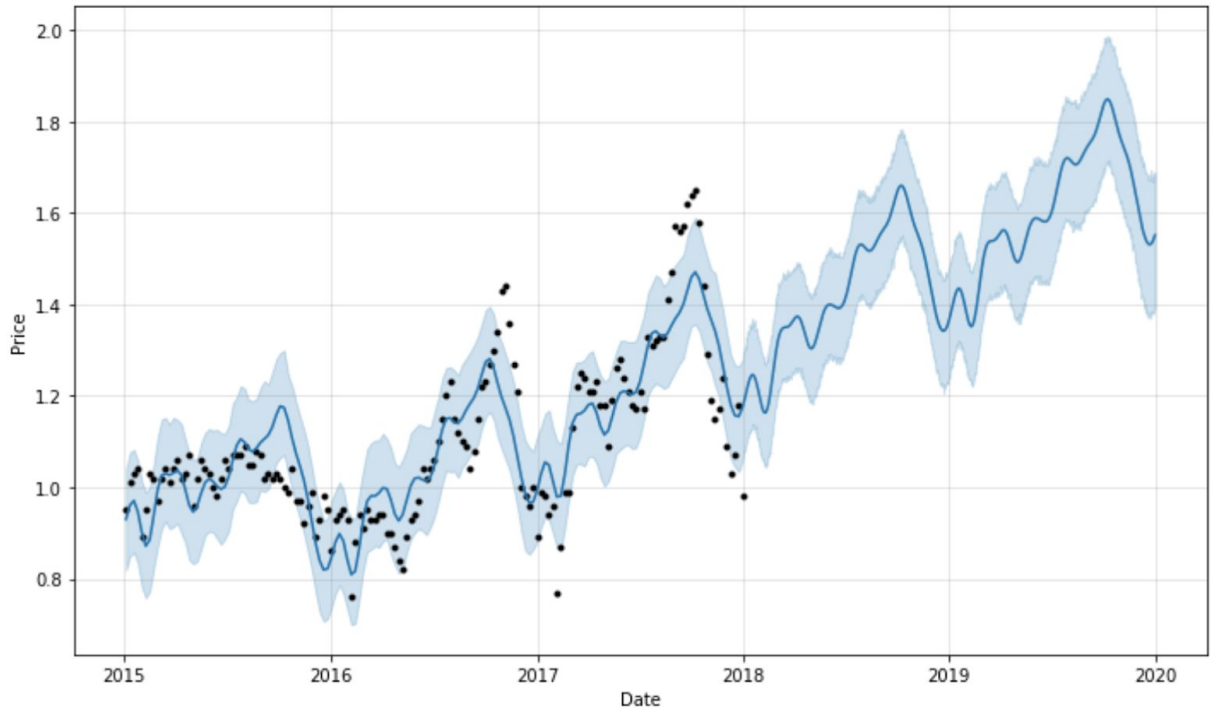
There are two types of avocado: conventional and organic, in this part, I will focus on conventional avocado's price in the total US. Forecast conventional avocado's price from 3 ways to compare the result:

1. Prophet
2. ARIMA
3. EMA

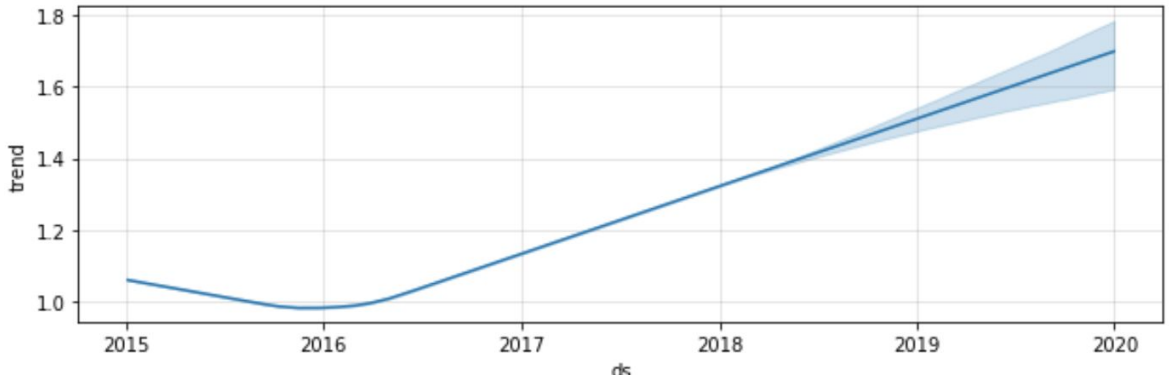
Prophet

Let's start from the prophet.

1. Calculate Prophet Prediction



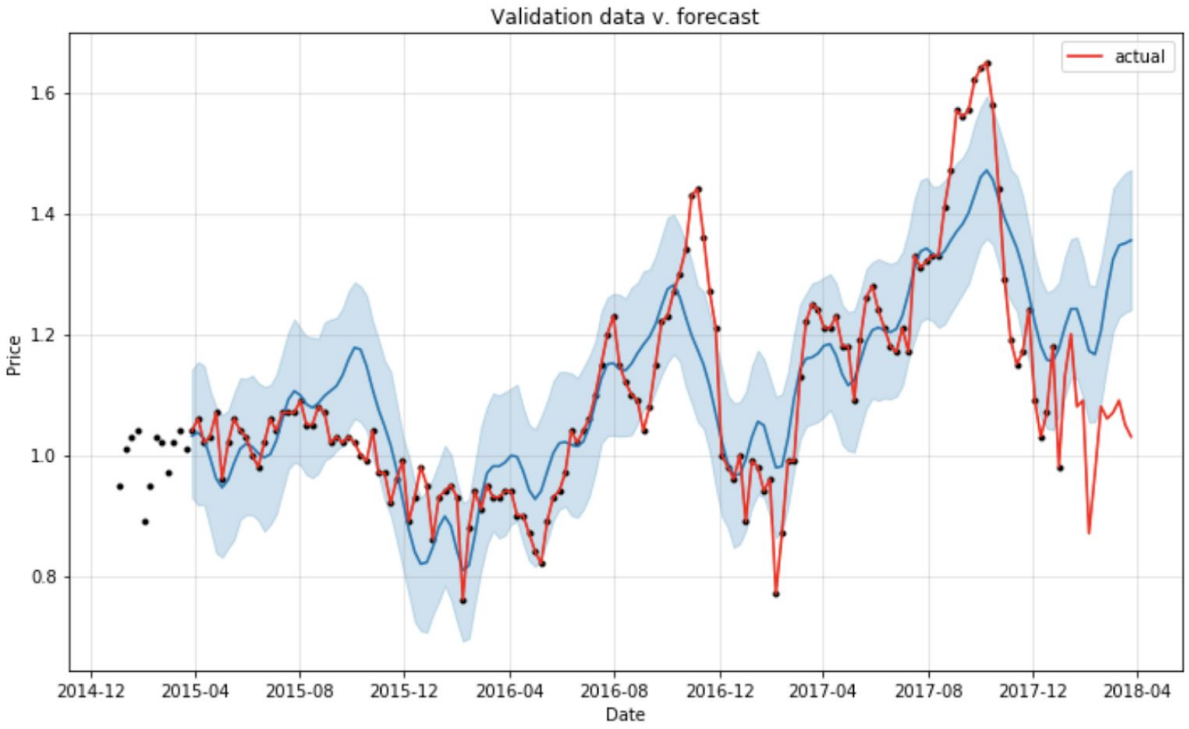
We had the price of avocado from 2015 to 2018, it shows as the black spots part in the graph. The future price after 2018-03 in the graph indicates the price will frequency oscillation go up.



Plotting the estimated components of the model shows the same result.

2. Valuation

Take original data of 2018 as a valuation as below:



The actual price and predication price in the same plot. Two price trend shapes are pretty similar but do not match very well.

Let's do R^2 valuation:

Coefficient of determination R^2 value is 0.9486081370449679, R^2 close to 1, means the fit difference is really small, and the model fits very well.

The result of prophet prediction can be a reference.

ARIMA

I use the AIC calculator engine to find the best model to define p, d, q values. The result is: AIC:-214.528 is the lowest, ARIMA(0, 1, 0)x(0, 0, 0) is the best model.

```
Performing stepwise search to minimize aic
Fit ARIMA(1,1,1)x(0,0,0,0) [intercept=True]; AIC=-408.780, BIC=-396.581, Time=0.092 seconds
Fit ARIMA(0,1,0)x(0,0,0,0) [intercept=True]; AIC=-412.530, BIC=-406.430, Time=0.038 seconds
Fit ARIMA(1,1,0)x(0,0,0,0) [intercept=True]; AIC=-410.775, BIC=-401.626, Time=0.051 seconds
Fit ARIMA(0,1,1)x(0,0,0,0) [intercept=True]; AIC=-410.786, BIC=-401.637, Time=0.173 seconds
Fit ARIMA(0,1,0)x(0,0,0,0) [intercept=False]; AIC=-414.528, BIC=-411.479, Time=0.010 seconds
Total fit time: 0.381 seconds
                           Statespace Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  157
Model:               SARIMAX(0, 1, 0)   Log Likelihood                 208.264
Date:                Wed, 29 Jul 2020   AIC                           -414.528
Time:                        19:30:54   BIC                           -411.479
Sample:                             0   HQIC                          -413.290
                                  - 157
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.0041      0.000     11.341      0.000       0.003       0.005
===================================================================================
Ljung-Box (Q):                       29.27   Jarque-Bera (JB):                24.66
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               2.90   Skew:                            -0.72
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.31
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

After model fit, state space model results table shows each values as below:

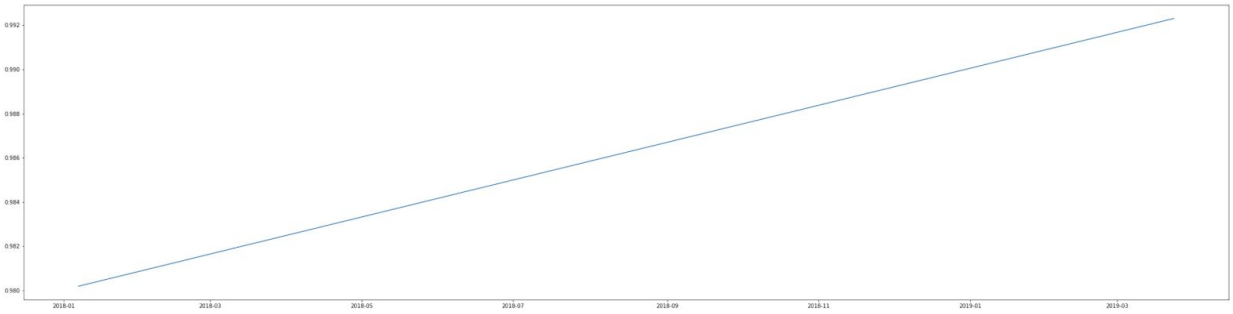
Statespace Model Results

Dep. Variable:	AveragePrice	**No. Observations:**	157
Model:	SARIMAX(0, 1, 0)	**Log Likelihood**	208.265
Date:	Wed, 29 Jul 2020	**AIC**	-412.530
Time:	19:31:16	**BIC**	-406.430
Sample:	01-04-2015	**HQIC**	-410.052
	- 12-31-2017		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0002	0.006	0.035	0.972	-0.011	0.011
sigma2	0.0041	0.000	10.422	0.000	0.003	0.005

Ljung-Box (Q):	29.27	**Jarque-Bera (JB):**	24.66
Prob(Q):	0.89	**Prob(JB):**	0.00
Heteroskedasticity (H):	2.90	**Skew:**	-0.72
Prob(H) (two-sided):	0.00	**Kurtosis:**	4.31

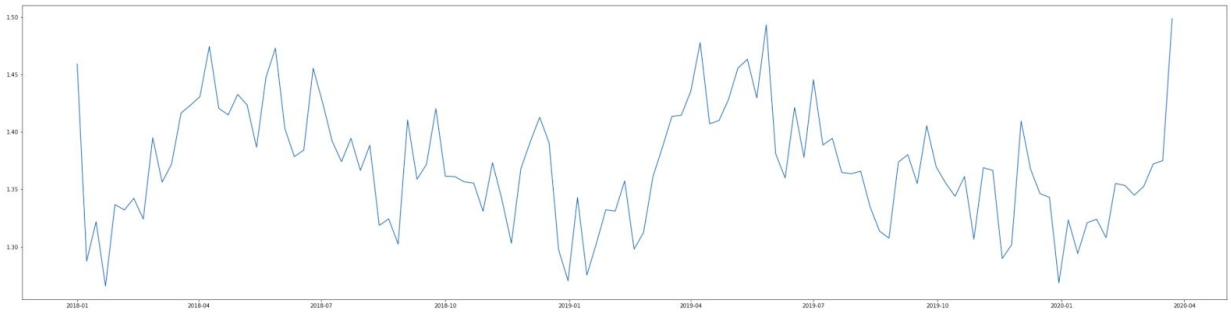
The range of average price in the future is $0.98 to close to $1.00.



ARIMA predicts the future price after 2018-03-25 before 2020-03-24, in the graph indicates the price will go up.

EMA

Exponential Moving Average shows the price in the future will be frequency oscillation, but cannot tell the trend is go up or go down.



Conclusion

Three of the predictions show the similar result. EMA responds quickly to any factor's change, better than ARIMA in this project, Prophet shows more details, clearest average price trend. In this project, the Prophet gives the best forecasting.

Combining three models above, the average price of conventional avocados will be frequency oscillation go up in the future.