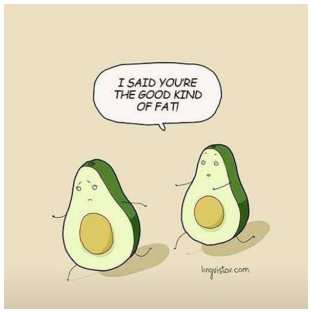
Maxine Nussbaum

Python Programming

Professor Stephen Haag



Python Final: Avocado Mania

The Data

Yearly, the Hass Avocado Board publishes data about avocado sales in the U.S. The data used in the following project contains avocado sales from January, 2015- August, 2020, taken from the Hass Avocado Board data. Two different datasets were used in the project, one with observations only for the entirety of the U.S., and one with data by city in the U.S. Other than the longitude/latitude variables , both datasets are the same. In both datasets, there are the following variables:

Region: (string) For the total U.S. dataset, this only consists of 'TotalUS.' However, for the city dataset, this variable consists of city names.

Date: (datetime) Date of observation (Year and Month variables were extracted from this variable)

Type: (string) Conventional or Organic

Average Price: (float) average price on the date of observation for the region involved

Total Volume: (float) total volume on the date of observation for the region involved

PLU 4046: (float) volume of that PLU sold on the date of observation for the region involved. Small Hass avocado PLU.

PLU 4225: (float) volume of that PLU sold on the date of observation for the region involved. Large Hass avocado PLU.

PLU 4770: (float) volume of that PLU sold on the date of observation for the region involved . Extra Large Hass avocado PLU.

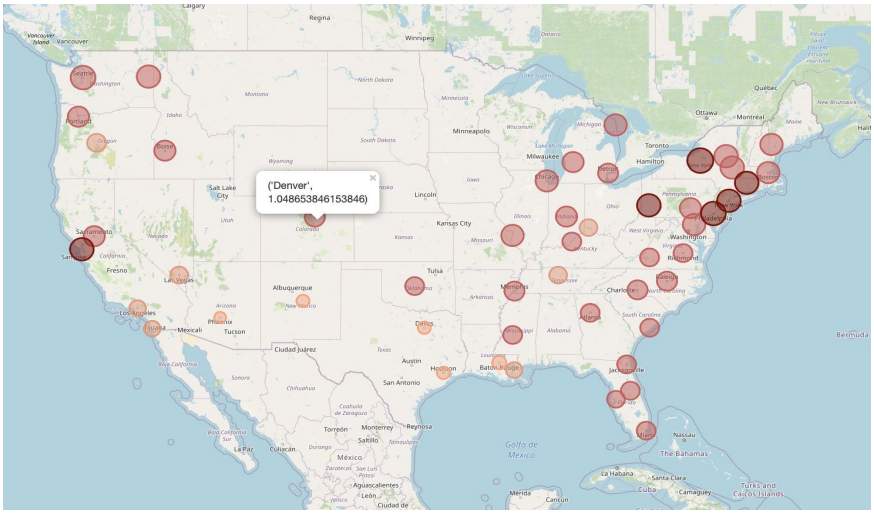
***Lon:** Longitude of region

***Lat:** Latitude of region

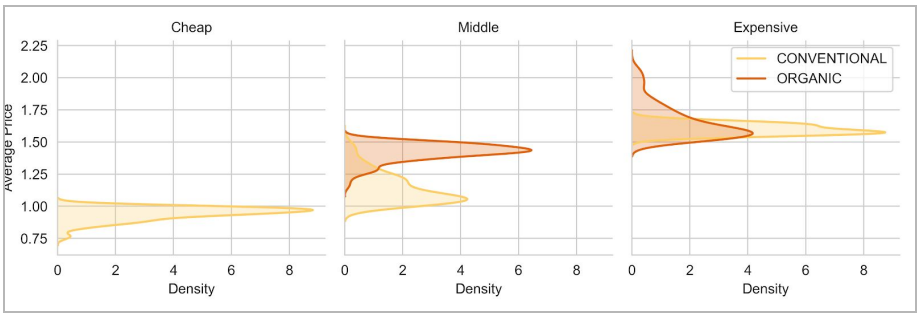
*Variables only in the regional/city dataset

1.Binning

To begin the journey into the avocado market, below are two visualizations of avocado average prices throughout the US, binned into three sections: Cheap, Middle, and Expensive. The map, created using Python's Folium package, is actually interactive HTML, which displays both the city name, and average price in 2015 for conventional avocados, when a bubble is selected - as shown with Denver on the map. Light orange depicts the Cheap bin, light red depicts the Middle bin, and the maroon color depicts the Expensive bin.



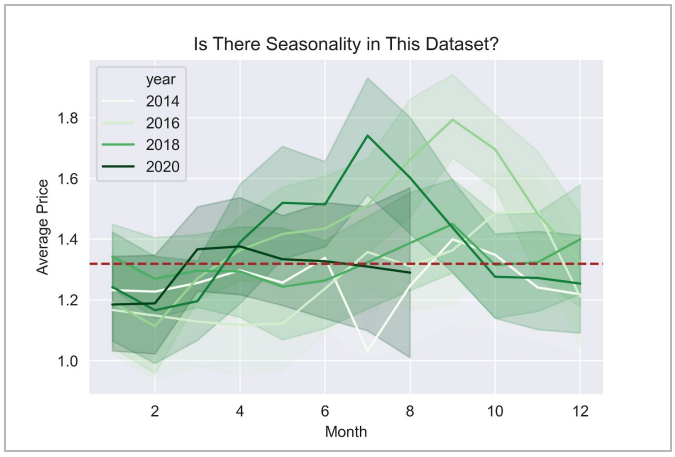
Next is a faceted density plot depicting the density within each bin, by type of avocado. It is clear from this plot that, overall, organic avocados retail at a higher price than conventional avocados.



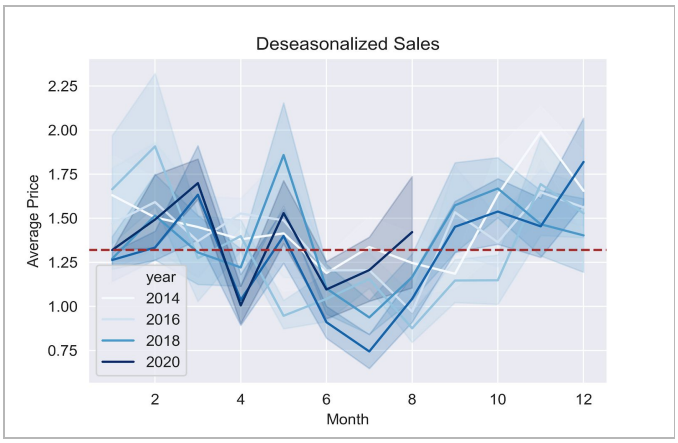
2. Graphing

2a. Summary graphing

In order to better understand the data, examination into its seasonality was essential.The below two visualizations depict both raw data and deseasonalized data, respectively. As somewhat clear from the raw data, there is a general position trend toward August - month 8 - and then a decline in sales following. Perhaps there is not very clear seasonality, but an upward and then downward trend can certainly be inferred.

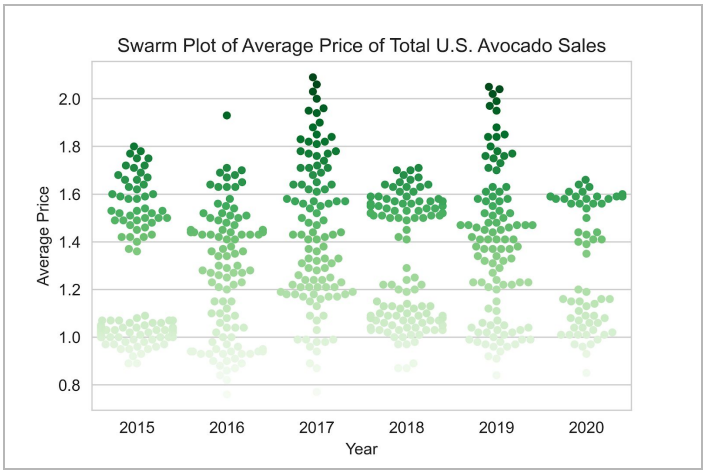


**For 2020, there is only data available until August.

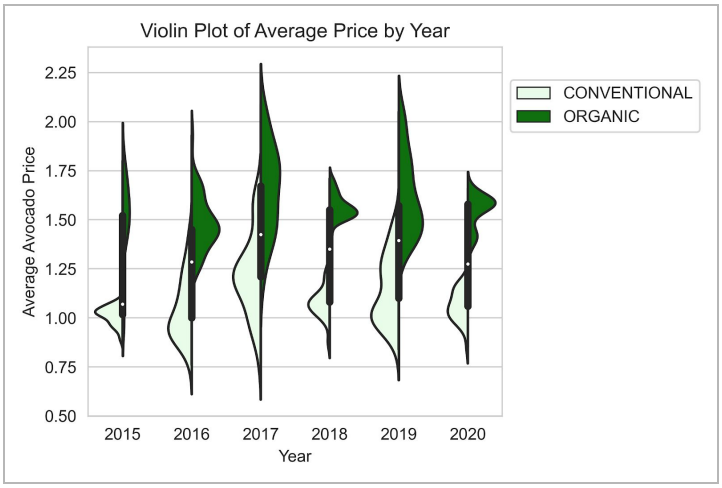


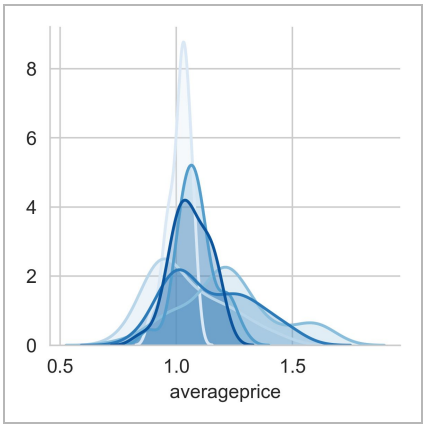
The deseasonalized lineplot to the left depicts the data after it has been analyzed for seasonal indexes. The process this involved was creating a seasonal index by month to determine the distance away from the mean for each month. Then, the original average price is divided by the season indexes for the corresponding month, and thus creating a deseasonalized dataset. In this visualization, there is less of a clear trend by month.

Next is an analysis of sales per year. Both of the to the right, and plot below display similar conclusions, but in different ways. The swarm plot directly to the right displays the average price for both conventional and organic avocados by year.

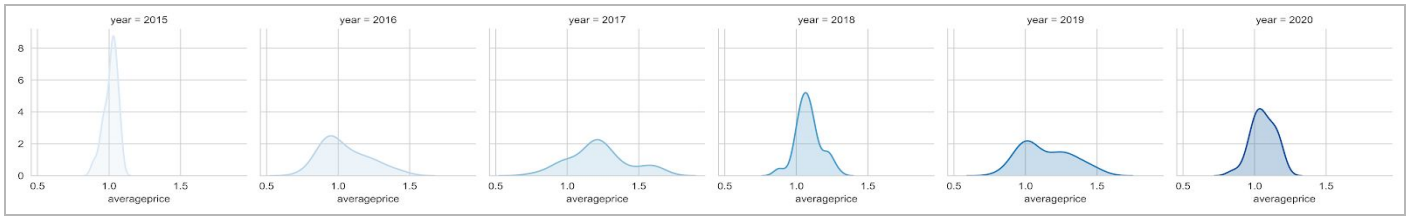


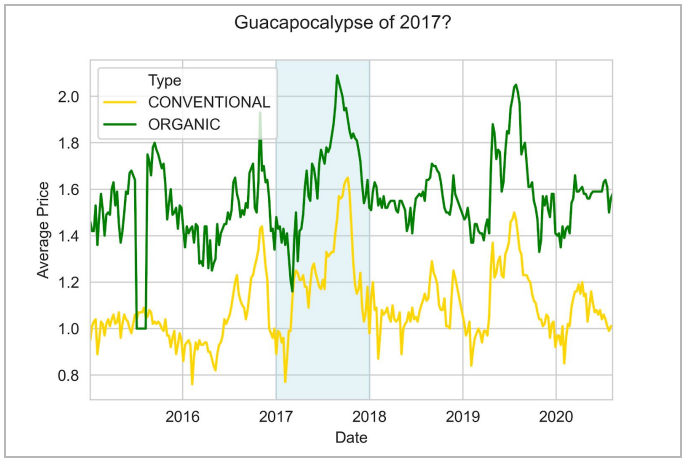
The violin plot to the right displays essentially the same information as the swarm plot above, however this plot splits conventional from organic avocado prices.





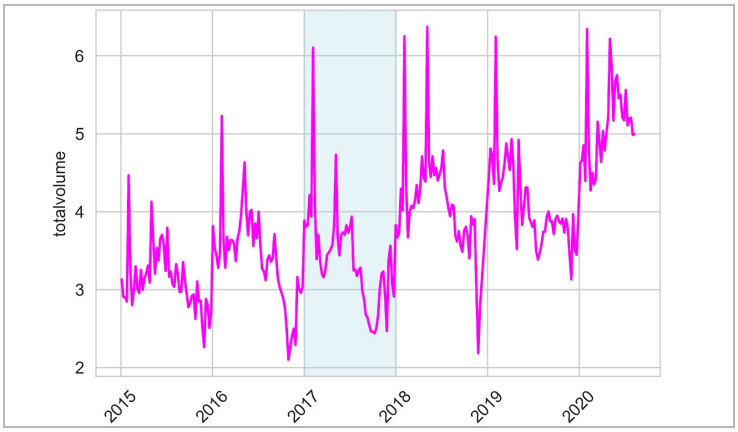
The elongated graph below is a broken down visualization of the graph to the left. Both visualizations display the density of average prive overtime for conventional avocados. Keeping in mind that 2020 only has data until August, it seems as though there is a gradually rising mean in average price overtime.





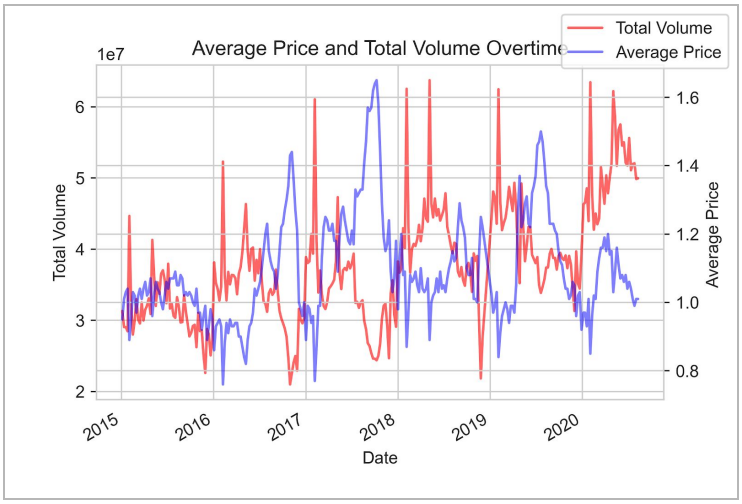
Now, we move to a pivotal question, was there really a guacapocalypse (guac + apocalypse) of 2017?

The year of 2016 showed a steep increase in demand of avocados - see below graph - hence leading to an avocado shortage. Thus creating the supposed-guacapocalypse. In the graph to the left, as well as the swarm and violin plots, 2017 *did* in fact show elevated prices for avocados.

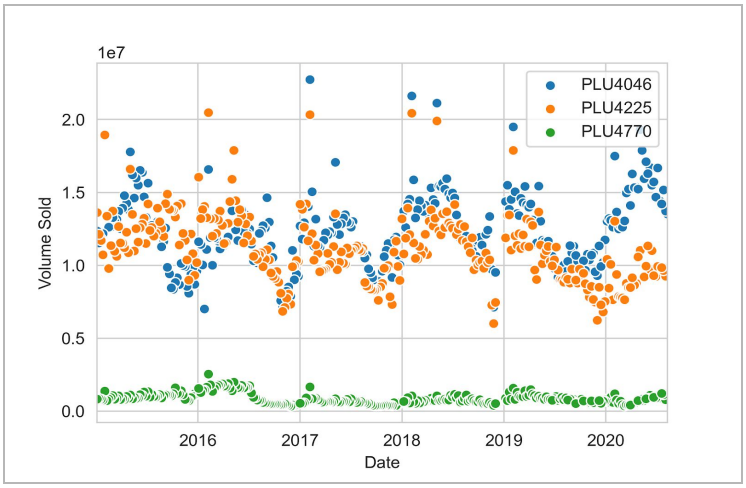


However, after 2017, the pattern of small spikes in avocados prevail, except for a brief stint in 2019 where prices reached near guacapocalyptic levels.

Here is a plot of average price and total volume on a dual axis overtime. Above is the total volume aspect of this plot, and then average price was added on in a dual axis.

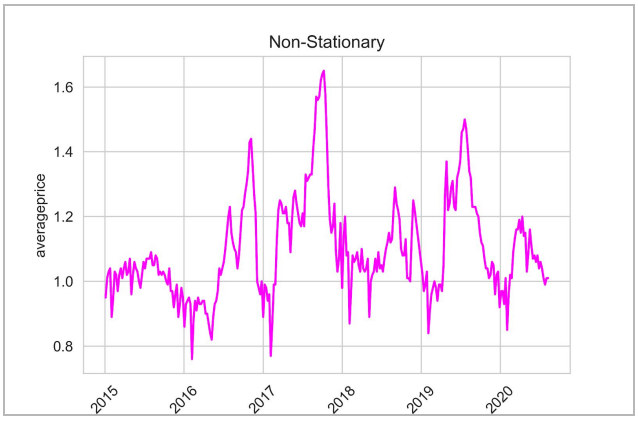


In the regression section, there is a multiple regression that was run on the data in connection with PLU effect on average price. For reference, this is a graph of the different volumes sold of each PLU.

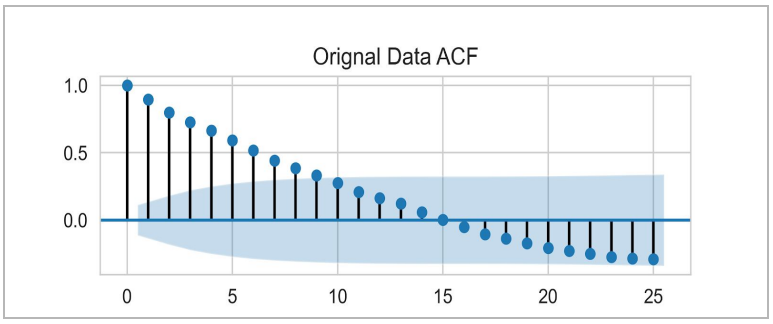


2b. Time-series Analysis

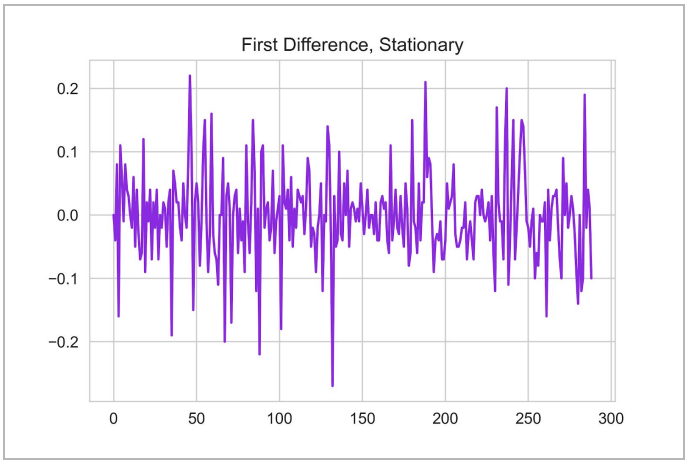
Average Price Time-Series Analysis, Graphing, and Forecasting ARIMA(0,1,0) of Conventional Avocados



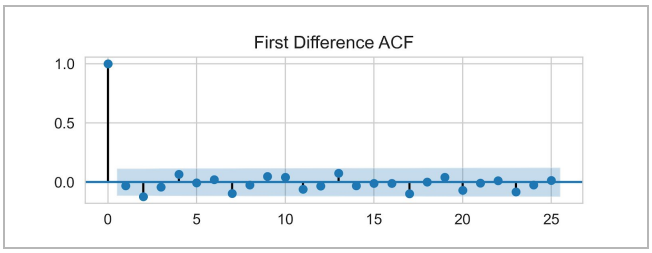
In order to accurately conduct time-series ARIMA analysis, a dataset must be stationary, with a mean of 0. To the left is a line plot of average price overtime; it is clear that the mean is not 0, nor is the mean stationary, it is adrift throughout time.

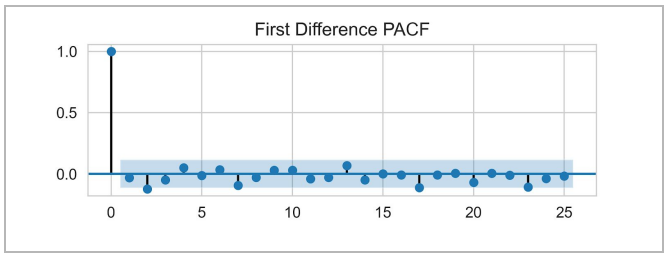


As with the line plot, the ACF (Autocorrelation Factor plot) shows gradual decline, thus also nodding toward non-stationarity. When a Dickey-Fuller test was run on the difference data, the resulting p-value was 0.003052, also indicating non-stationarity.



After a first difference was taken of the dataset, there was a clear mean of 0, thus indicating stationarity, and a q=1 in ARIMA(p,d,q).



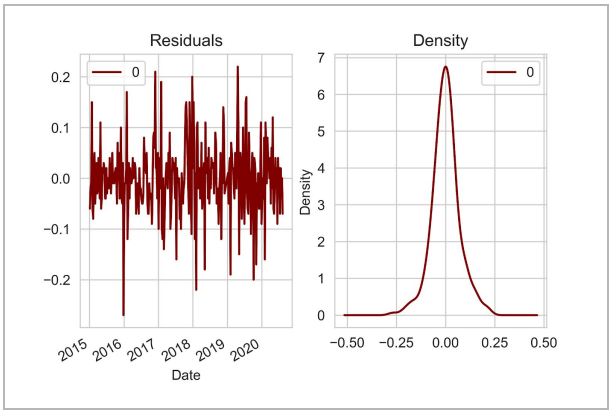


Analysis of the two above graphs - the ACF and PACf (Partial Autocorrelation Factor plot) - indicated a potential of AR(1) and MA(1). However, when looking at the model summary, both ARIMA(1,1,0) and ARIMA(0,1,1), as well as overfitting models, failed to meet required standards. ARIMA(0,1,0) was the best option, and that is what the following graphs proceed from.

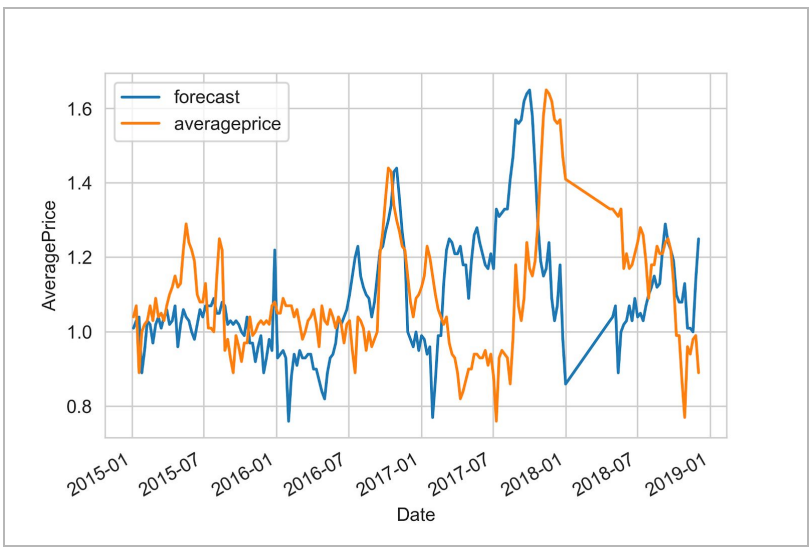
ARIMA Model Results

Dep. Variable:	D.averageprice	No. Observations:	301
Model:	ARIMA(0, 1, 0)	Log Likelihood	366.884
Method:	css	S.D. of innovations	0.072
Date:	Tue, 17 Nov 2020	AIC	-729.768
Time:	15:04:36	BIC	-722.353
Sample:	1	HQIC	-726.801

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0005	0.004	0.129	0.897	-0.008	0.009

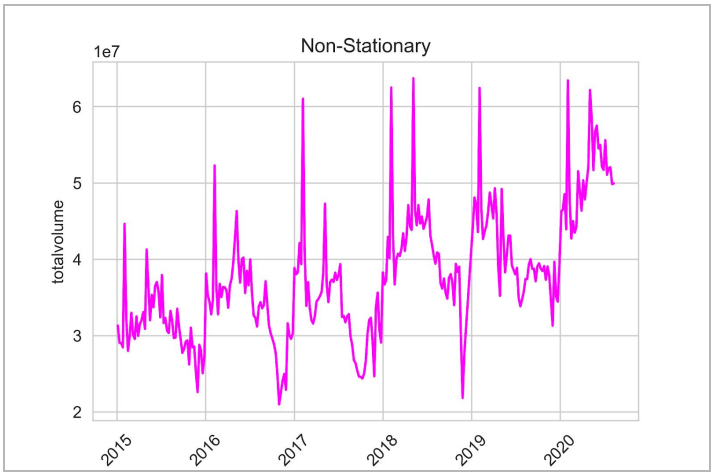


Residual diagnostics of this model resulted in normally distributed whitenoise residuals. Indicating the validity of the model chosen.

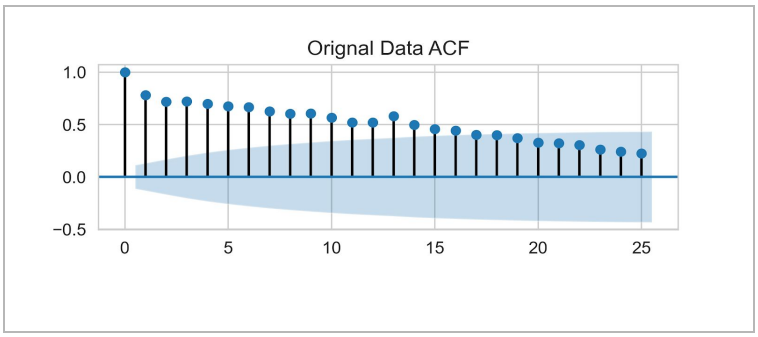


Displayed here is the forecasted predictions from the model vs the actual values of average price. While the model does not perfectly fit the actual values, it comes close to predicting spikes and average trend.

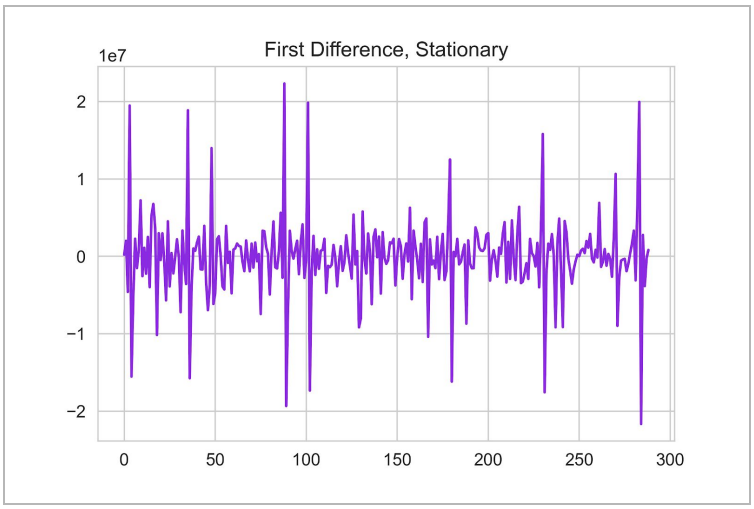
Total Volume Time-Series Analysis, Graphing, and Forecasting ARIMA(0,1,2) of Conventional Avocados



As with average price, and perhaps even more so, total volume is not a stationary line plot. Hence a first difference was taken. When a Dickey-Fuller test was run, the resultant p-value was 0.243424, indicating non-stationarity as well.

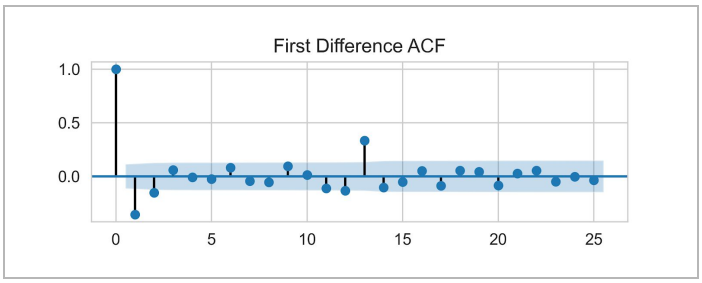


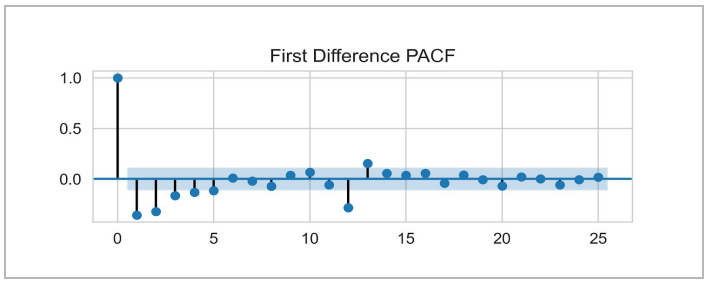
As with average price, the initial ACF plot of the raw data indicated non-stationarity as a result of a gradually decreasing trend in the plot.



With the first difference plot, a mean of 0 is clear, thus indicating stationarity.

Resulting ACF and PACF plots from the first differenced dataset indicated both stationarity and the possibility for MA(1 or 2) and AR (1,2,3, of 4).





After analysis of multiple models, the best model came to be ARIMA(0,1,2) - summary displayed below.

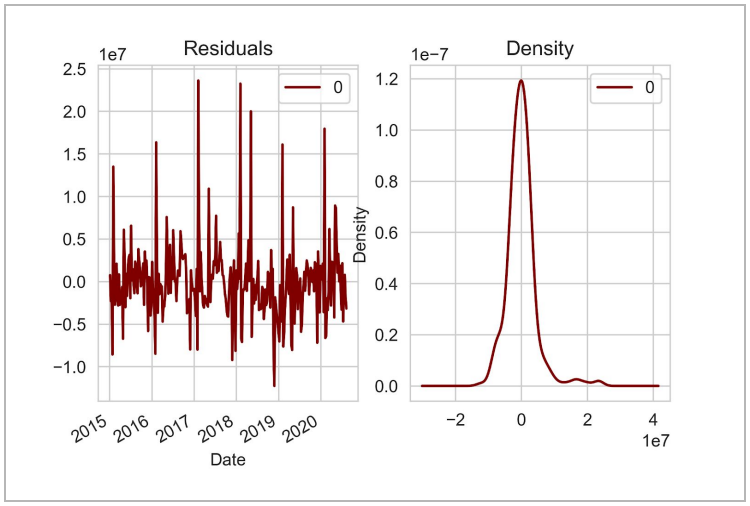
ARIMA Model Results

Dep. Variable:	D.totalvolume	No. Observations:	301
Model:	ARIMA(0, 1, 2)	Log Likelihood	-5046.897
Method:	css-mle	S.D. of innovations	4626145.282
Date:	Tue, 17 Nov 2020	AIC	10101.793
Time:	14:56:43	BIC	10116.622
Sample:	1	HQIC	10107.727

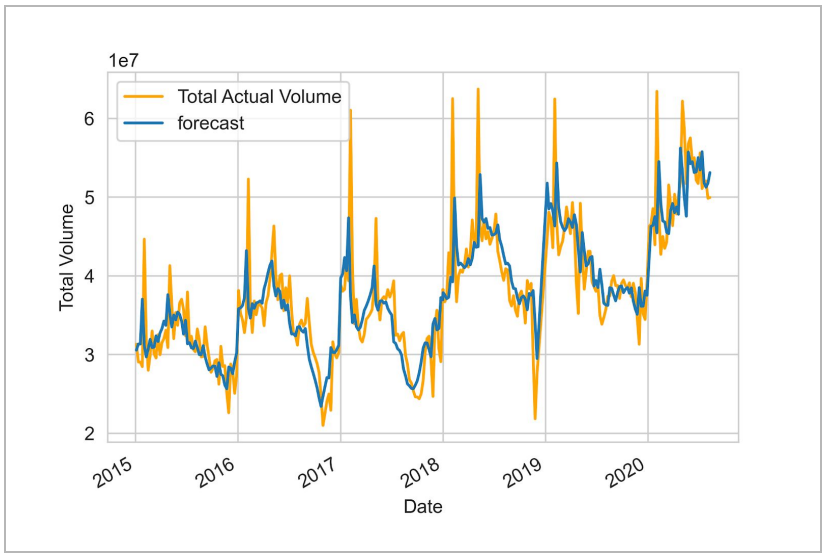
	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.192e+04	8.48e+04	-0.376	0.707	-1.98e+05	1.34e+05
ma.L1.D.totalvolume	-0.5576	0.061	-9.212	0.000	-0.676	-0.439
ma.L2.D.totalvolume	-0.1271	0.060	-2.101	0.036	-0.246	-0.009

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.3672	+0.0000j	1.3672	0.0000
MA.2	-5.7550	+0.0000j	5.7550	0.5000



Residual diagnostics resulted in whitenoise, normally-distributed residuals, therefore indicating a good model.



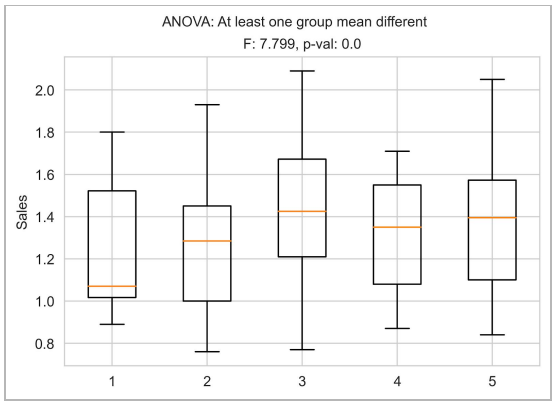
Displayed here is the forecasted predictions from the model vs the actual values of total price. Unlike the average price model, this model fits more closely to the actual values of total volume.

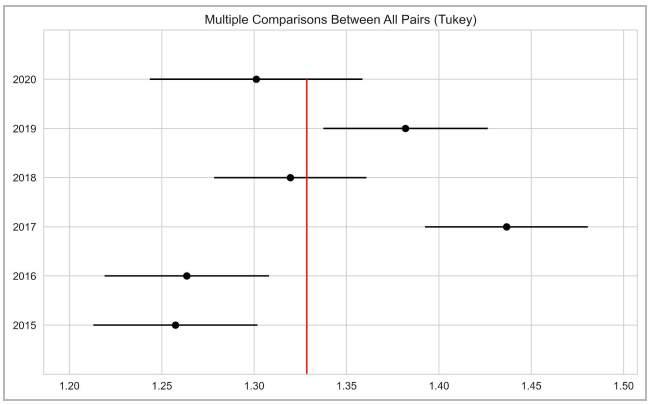
3. ANOVA

An ANOVA test was run on average price between the Eastern United States, and the Western United States. From the test, there was a p-value greater than α, indicating to fail to reject the mean.

```
Conclusion: Fail to Reject Ho: We can't reject that the means are the same
Multiple Comparison of Means - Tukey HSD, FWER=0.20
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
  East   West  -0.0693 0.1617 -0.1328 -0.0059   True
-----------------------------------------------------
```

A second ANOVA test was run on Average Price mean by Year.





After the test was run, the resultant p-value was essentially 0, therefore indicating the reject the null hypothesis, and therefore conclude that at least one group mean in different. In the two visualizations above, this is evident clearly with 2015, 2016, and 2017.

4. Regression

4a. OLS Regression

The first OLS Regression run of this dataset was that of total volume's effect on average price for conventional avocados. Below is the summary output for this regression. Low P-Values, and a higher r-squared values indicate that total volume has *some* effect on average price for conventional avocados.

OLS Regression Results

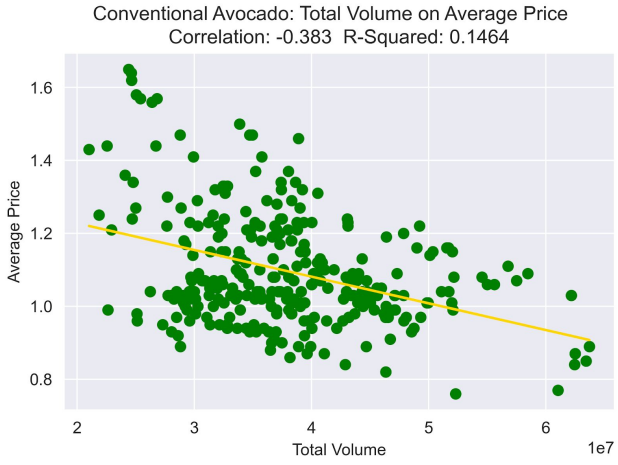
Dep. Variable:	y	R-squared:	0.637
Model:	OLS	Adj. R-squared:	0.636
Method:	Least Squares	F-statistic:	588.8
Date:	Tue, 17 Nov 2020	Prob (F-statistic):	7.29e-76
Time:	09:36:17	Log-Likelihood:	-5934.6
No. Observations:	338	AIC:	1.187e+04
Df Residuals:	336	BIC:	1.188e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.78e+07	2.55e+06	30.478	0.000	7.28e+07	8.28e+07
x1	-4.583e+07	1.89e+06	-24.266	0.000	-4.95e+07	-4.21e+07

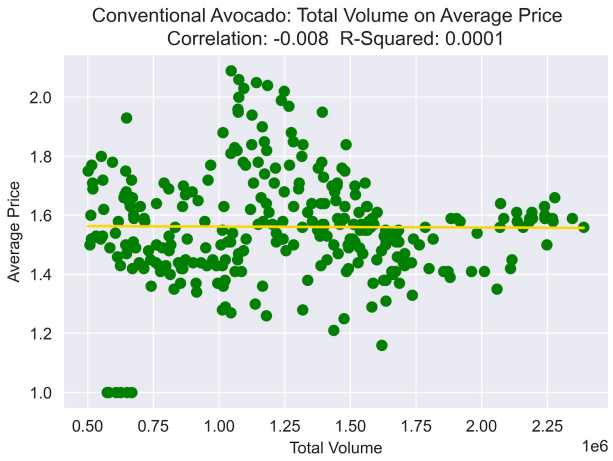
Omnibus:	6.935	Durbin-Watson:	0.188
Prob(Omnibus):	0.031	Jarque-Bera (JB):	6.826
Skew:	-0.300	Prob(JB):	0.0329
Kurtosis:	3.354	Cond. No.	9.48

4b. Simple Linear Regression

A simple linear regression was then run on the same data, and came up with a significantly different answer for the r-squared value. Below is the output plot demonstrating the trend line when plotted against Total Volume v Average Price. Interestingly too, this graph illustrates the demand line in the supply/demand graph prevalent in neo-classical economic theory.



The same simple regression was then run with organic avocado values. Here there is even less of a clear correlation between Total Volume and Average Price.



4c. Multiple Regression

A multiple regression was run of the effect of average price by PLUs of only conventional avocados. In simpler terms, does each different PLU (three separate columns in this dataset) effect average price differently. PLU 4770 are extra large Hass avocados; PLU 4225 are large Hass avocados; and PLU 4046 are small Hass avocados.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           AveragePrice   R-squared:                       0.564
Model:                            OLS   Adj. R-squared:                  0.557
Method:                 Least Squares   F-statistic:                     71.29
Date:                Wed, 11 Nov 2020   Prob (F-statistic):           1.28e-29
Time:                        15:08:53   Log-Likelihood:                 126.35
No. Observations:                 169   AIC:                            -244.7
Df Residuals:                     165   BIC:                            -232.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.6625      0.051     32.758      0.000       1.562       1.763
PLU4046    -5.223e-09   4.79e-09     -1.090      0.277   -1.47e-08    4.24e-09
PLU4225    -3.388e-08   6.81e-09     -4.977      0.000   -4.73e-08   -2.04e-08
PLU4770    -1.243e-07   3.07e-08     -4.054      0.000   -1.85e-07   -6.38e-08
==============================================================================
Omnibus:                       14.997   Durbin-Watson:                   0.433
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               17.501
Skew:                           0.624   Prob(JB):                     0.000158
Kurtosis:                       3.963   Cond. No.                     9.68e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.68e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

5. Seaborn Graphs (See Section 2:Graphing)

All graphs in section 2 have a Seaborn graphing aspect except for the residual plots in the time-series section.

6. Pivot Tables

Type	PLU4046	PLU4225	PLU4770	averageprice	totalvolume
CONVENTIONAL	1.251631e+07	1.117968e+07	867632.900232	1.096192	3.795737e+07
ORGANIC	1.433086e+05	2.795709e+05	4684.335728	1.560894	1.264534e+06

This pivot table displays the mean, by type of avocado, of five variables. It is clear that PLU 4770 has the lowest volume of sales compared to the other two PLUs.

Type	month	averageprice	totalvolume
CONVENTIONAL	1	1.008710	3.907565e+07
	2	0.946071	4.376813e+07
	3	1.057419	3.958603e+07
	4	1.091538	4.009608e+07
	5	1.067037	4.442915e+07
	6	1.116000	4.148711e+07
	7	1.174074	3.891193e+07
	8	1.186087	3.576112e+07
	9	1.225909	3.335037e+07
	10	1.220909	3.091727e+07
	11	1.112857	2.984537e+07
	12	1.012632	3.189859e+07
ORGANIC	1	1.476452	1.211537e+06
	2	1.449643	1.304747e+06
	3	1.462903	1.438721e+06
	4	1.525000	1.447127e+06
	5	1.547037	1.434918e+06
	6	1.602800	1.411825e+06
	7	1.588519	1.281985e+06
	8	1.654783	1.156494e+06
	9	1.744091	1.150647e+06
	10	1.658182	1.093250e+06
	11	1.600476	1.033458e+06
	12	1.514737	1.013334e+06

Means for average price and total volume are shown in this pivot table by type. For Organic avocados, seasonality is somewhat evident, in that the price mean rises from January to October, and then begins to decrease through the rest of the year. Also clear from this pivot table is the distinct difference in total volume between conventional and organic avocados.

Type	month	averageprice	totalvolume
CONVENTIONAL	1	0.083336	5.628429e+06
	2	0.086937	1.038891e+07
	3	0.091832	6.405695e+06
	4	0.122497	6.576468e+06
	5	0.139610	8.986943e+06
	6	0.116404	6.258828e+06
	7	0.157999	7.429780e+06
	8	0.139992	5.729390e+06
	9	0.189628	4.925280e+06
	10	0.213115	5.972665e+06
	11	0.144158	6.021632e+06
	12	0.083589	4.236534e+06
ORGANIC	1	0.080645	4.348056e+05
	2	0.089669	4.184004e+05
	3	0.115187	4.329765e+05
	4	0.132974	4.580021e+05
	5	0.143173	5.097279e+05
	6	0.123576	5.257199e+05
	7	0.295774	4.728756e+05
	8	0.270351	3.945287e+05
	9	0.147798	3.754198e+05
	10	0.168569	4.078523e+05
	11	0.118679	3.684123e+05
	12	0.078766	3.008944e+05

As with the previous pivot table, rising prices from January toward the end of the year are shown. Further, both conventional and organic have similar deviations in average price.

Type	month	averageprice	totalvolume
CONVENTIONAL	1	31.27	1.211345e+09
	2	26.49	1.225508e+09
	3	32.78	1.227167e+09
	4	28.38	1.042498e+09
	5	28.81	1.199587e+09
	6	27.90	1.037178e+09
	7	31.70	1.050622e+09
	8	27.28	8.225058e+08
	9	26.97	7.337080e+08
	10	26.86	6.801799e+08
	11	23.37	6.267527e+08
	12	19.24	6.060731e+08
ORGANIC	1	45.77	3.755764e+07
	2	40.59	3.653290e+07
	3	45.35	4.460035e+07
	4	39.65	3.762530e+07
	5	41.77	3.874279e+07
	6	40.07	3.529564e+07
	7	42.89	3.461360e+07
	8	38.06	2.659936e+07
	9	38.37	2.531424e+07
	10	36.48	2.405151e+07
	11	33.61	2.170262e+07
	12	28.78	1.925335e+07

A sum aggregation function was run in this pivot table for average price and total volume. Within conventional, July has the highest sum price for sales. For organic, January has the highest sum price for sales. Likely, this is due to there only being sales in 2020 for the first eight months of the year. Further, the difference in volumes between conventional and organic show how many more conventional avocados are sold than organic avocados.

Conclusion:

From the analysis performed on this data, a few conclusions can be drawn. Total volume of avocados sold is likely to continue to rise for conventional avocados. However, average price does not show a distinct trend either positive or negative.

A distinct difference between values for conventional versus organic avocados is clear, though. Organic avocados have high average prices, and low total volume sold. Conventional avocados have lower average prices, and significantly high volumes sold.

From the time-series analysis of average price, it is clear that there is not a real distinct trend. Spikes in price over time led to difficulty in forecasting with the ARIMA analysis. However, the time-series forecasting for total volume provided a much clearer upward trend, allowing more ease in predicting and forecasting.