Avocado: what is expecte d the price and consumption of avocado in 2022?

Mireille Feudjio
Springboard Capstone Project

Contents

Summary

Avocado occupies an important place in meeting the concerns of diet, health, nutrition, and food. U.S. consumption of avocados has generally trended upward since 1970, it was 2.21 pounds per capita in 2000 to 7.81 pounds per capita in 2019 (Statista, 2020). Within this context, what are the expected avocado stock prices and quantities to be sold next year that can allow the Company Alpha base in New York to sign a contract with new producers and increase the price of avocado by 2%?

A time series dataset of the commercialization of avocado from 2015 to 2021 has been analyzed. The following section highlight the key results.

- Analysis of the commercialization of avocado showed a clear distinction between organic and conventional avocados,
- the price of the organic avocado is higher than that of the conventional avocado,
- the total volume of organic avocados sold is less than that of conventional avocados.
- New York is among the top 5 regions with high price (and high volume) in organic and conventional avocado (occupied 4th or the 5th rang).
- when total volume increases, price decreases and vice versa, the price of both types decreases in 2018 and decreased considerably from 2019 to 2021. While the volume increased significantly from 2018 to 2021.

8 models have been developed for each type of avocado; the specifications of the best forecasting models were:

- The best forecasting model of conventional avocado price is (1,0,0,0,0,1),52
- The best forecasting model of conventional avocado volume is (8,0,0)
- The best forecasting model of organic avocado price is (1,0,0,1,0,0),52
- The best forecasting model of organic avocado volume is (2,0,0)

Our models did good. The predicted prices of avocados (conventional and organic) are close to real values. the average of the difference between this value (pred – actual) is -0,024, the MAE is 0.063 and the MAPE is 0,050. More, the Pearson correlation is 0,712. However, forecasting volume was not good enough. There is a huge difference between real and predicted values (MAPE is 0.106, MAE is 224116.165, the Pearson correlation is 0,760).

Before covid 19, the price and volume of avocado was constantly increasing. the marketing of avocado was affected between 2019 and 2021. With time series analysis, this had a considerable impact on forecasting the future value of 2022. If considering that everything is back to normal (less impact of covid 19), we can suggest to the company to revise the prices of avocado considering the local trend, and the values of the upper limit of the price and volume.

How can we improve the current model? maybe add exogenous variables indicating if it is a covid period or not, indicating the population, the percentage of people affected by covid for each observation.

I. Introduction

Avocado (Persea americana) is a fruit which have been marketed as a healthy dietary choice and as a good source of beneficial monounsaturated oil. A whole medium avocado contains approximately 15 percent of the FDA's recommended daily amount of saturated fat. In addition, avocados have 60 percent more potassium than bananas. They are also rich in B vitamins, vitamin E, vitamin K and folate. Avocados are also a benefit to a diabetic diet. With diabetes increasing in the United States, avocados can offer a nutritious choice for those following a diabetic diet (CAC).

The value of U.S. avocado production measured $426 million in 2020. The United States produced 206,610 tons (NASS, 2020). U.S. consumption of avocados has followed a variable but generally increasing trend since 1970, increasing significantly from 2.21 pounds per capita in 2000 to 7.81 pounds per capita in 2019 (Statista, 2020). The United States is a net importer of avocados from Mexico. Mexico supplied most of the avocados imported into the United States in 2020. In 2020 the United States imported $2.4 billion in fresh avocados and exported approximately $45,502 in fresh avocados (ERS 2020).

Within this context, company Alpha based in the region of New York, wants to build a partnership with Avocado producers (farmers) to secure the stock of avocado for the next years (2022) and explore the possibility of increasing the price. Before taking any action, the Company seeks a Data scientist's advice who could help to take a decision. The main issue is: What are the expected avocado stock prices and quantities to be sold next year that can allow the Company Alpha base in New York to sign a contract with new producers and increase the price of avocado by 2%?

Based on the given dataset that presents the state of the commercialization of avocado, we assessed and analyzed with the guide of the following questions:

- What is the behavior of avocado price and consumption since 2015?
- What is the current price of avocado? or what was the trend of avocado price over the past 5 years and compare with the current year?
- What is the trend of avocado sold over the past 5 years and compare with the current year?
- In which region the avocado sold is high and how is the price as compared to others?
- In which region the price is high and how is the total avocado sold compared to others?
- Do the type of avocado (organic and conventional) impact the price and the consumption?
- What is the relationship between regions, price, avocado sold, and avocado type?
- What could we expect in the next one years in terms of price and consumption of avocado?

- What is the expected range of the avocado price and sold and where Company Alpha can explore for the definition of the new price and the volume for the next 2022?

The answering of these questions will guide Company Alpha to take a reasonable decision with regards to signing a contract with new producers if the consumption of avocado will increase in 2022.

Data sources

The dataset for this project was downloaded from Kaggle

Some relevant columns in the dataset:

- Date - The date of the observation
- Average price - the average price of a single avocado
- type – 2 types of avocado (conventional or organic)
- year – 2015 to 2021
- Region - the city or region of the observation
- Total Volume - Total number of avocados sold.

The present report outlines the key finding of the project structured in:

- Data Wrangling
- Exploratory data analysis
- Forecasting methodology
- Forecasting results
- Conclusion

II. Data Wrangling

The dataset shape was 41045 rows and 13 columns. There is no missing data, and no duplicate.

Data wrangling involved renaming of incorrect string value types in categorical features (type of avocado and country), indexing date and convert to datetime, checking for duplicate, missing value and data type of all columns. We filtered out subtotals from the dataset (total us, west, midsouth, northeast, south central and southeast).

Outliers' visualization

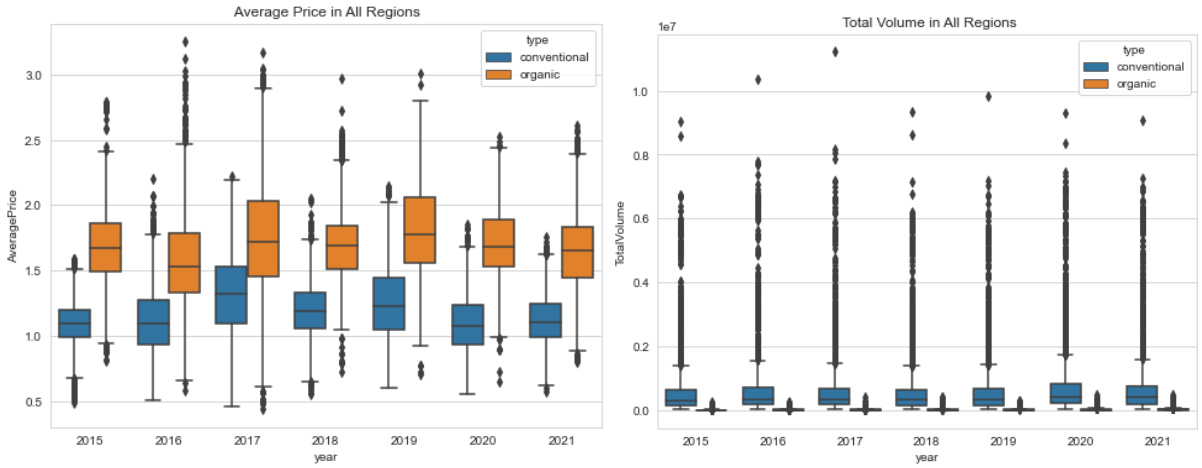


Fig. Outliers visualization for all regions

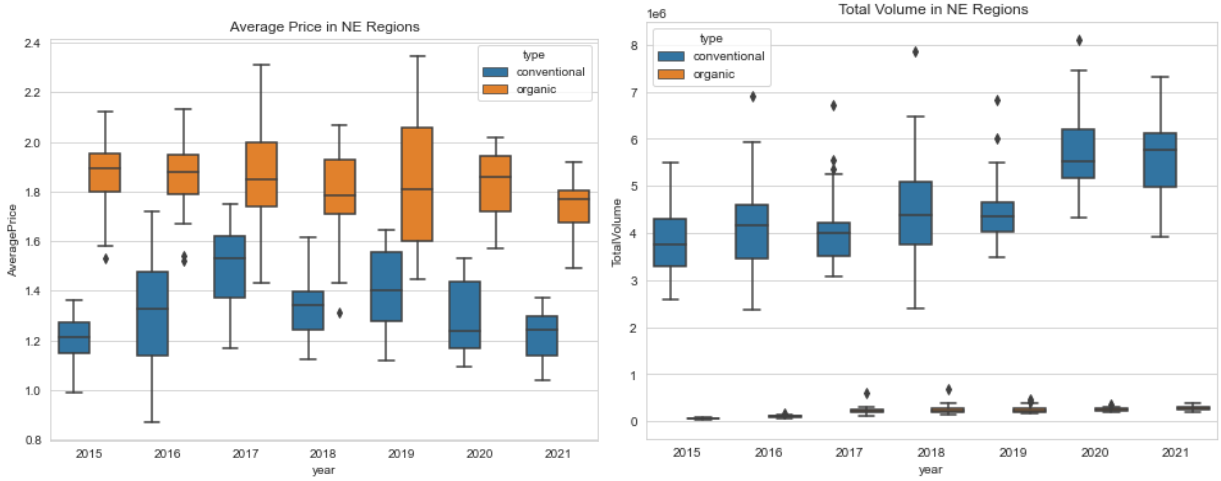


Fig. Outliers visualization for the Northeast regions

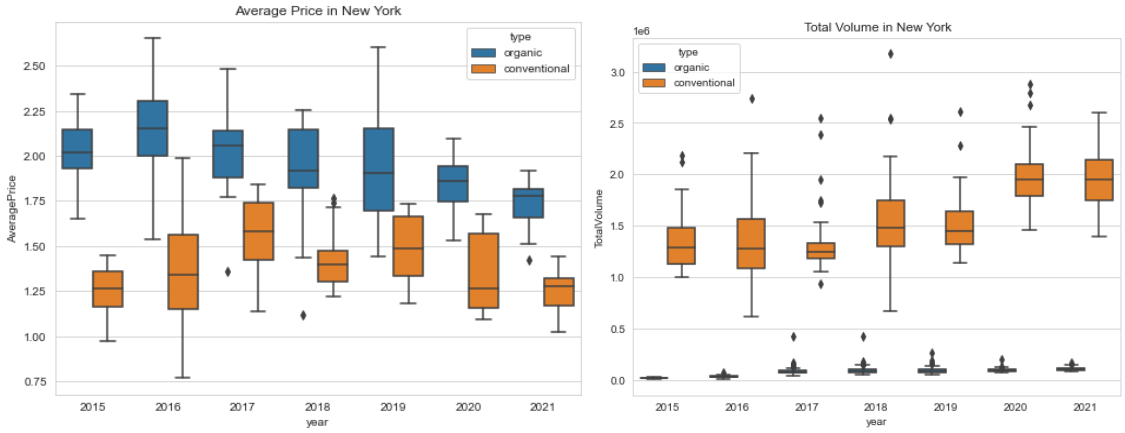


Fig. Outliers visualization New York

The outliers are abundant when considering all the dataset (all regions). This is due to the specificity that exists in each region. By plotting all together, the disparity that exists between countries is visible through the outliers. This is the raison why the outliers are less pronounced in New York dataset when plotting alone.

III. Exploratory data analysis

Conceptual framework for exploratory data analysis attempted to understand how the pattern of the time series look when considering (i) all regions of the dataset, (ii) the Northeast regions where New York belongs and, (iii) the New York city alone. Just like a top-down approach, we analyzed and visualized the avocado price and consumption behave at different scales. Here, we explored answers to the above questions to guide the forecasting section of the project.

NB: Northeast region state in the dataset contain (Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, and Pennsylvania).

1. **What is the price behavior of avocados in all regions?**

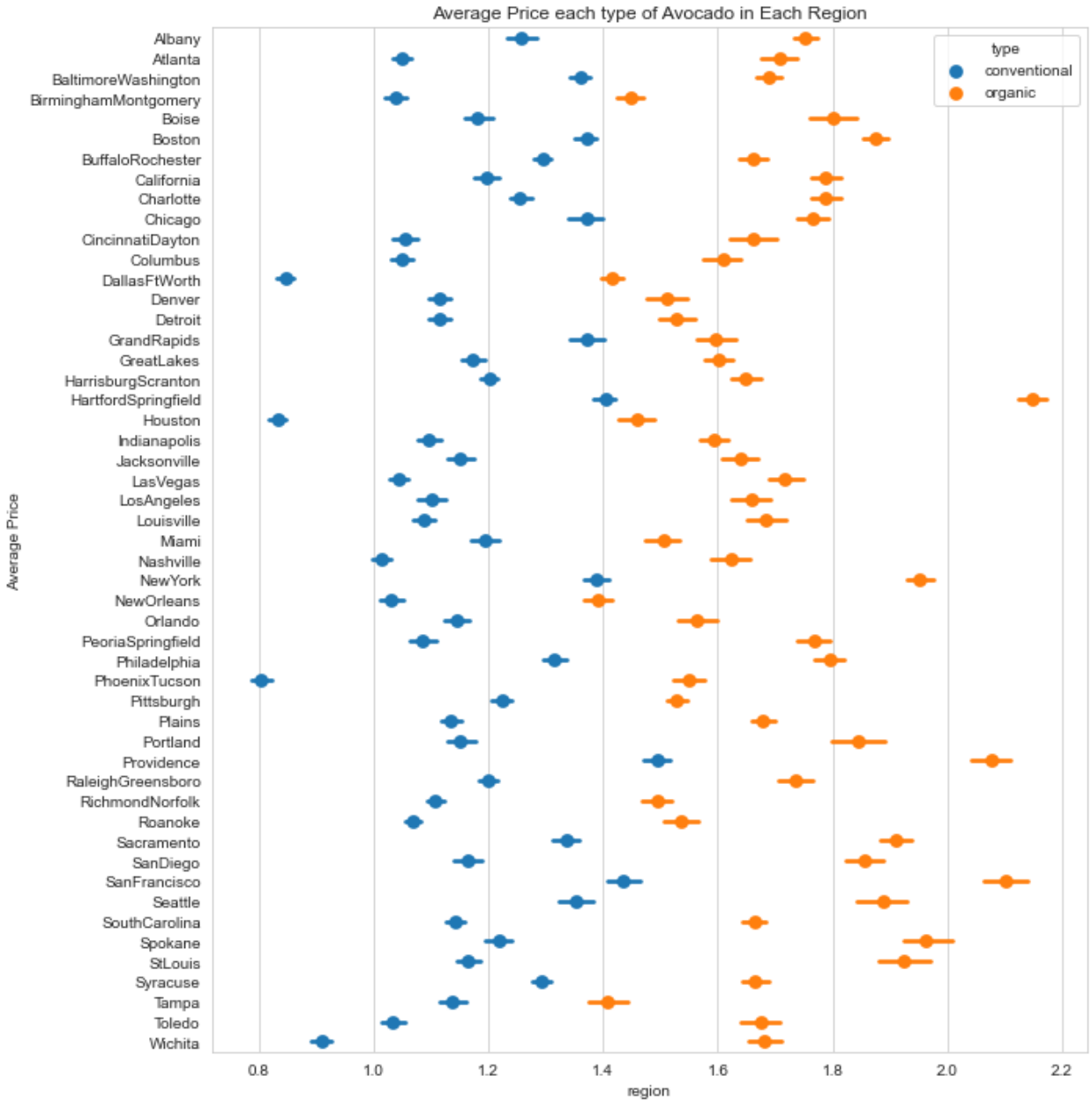
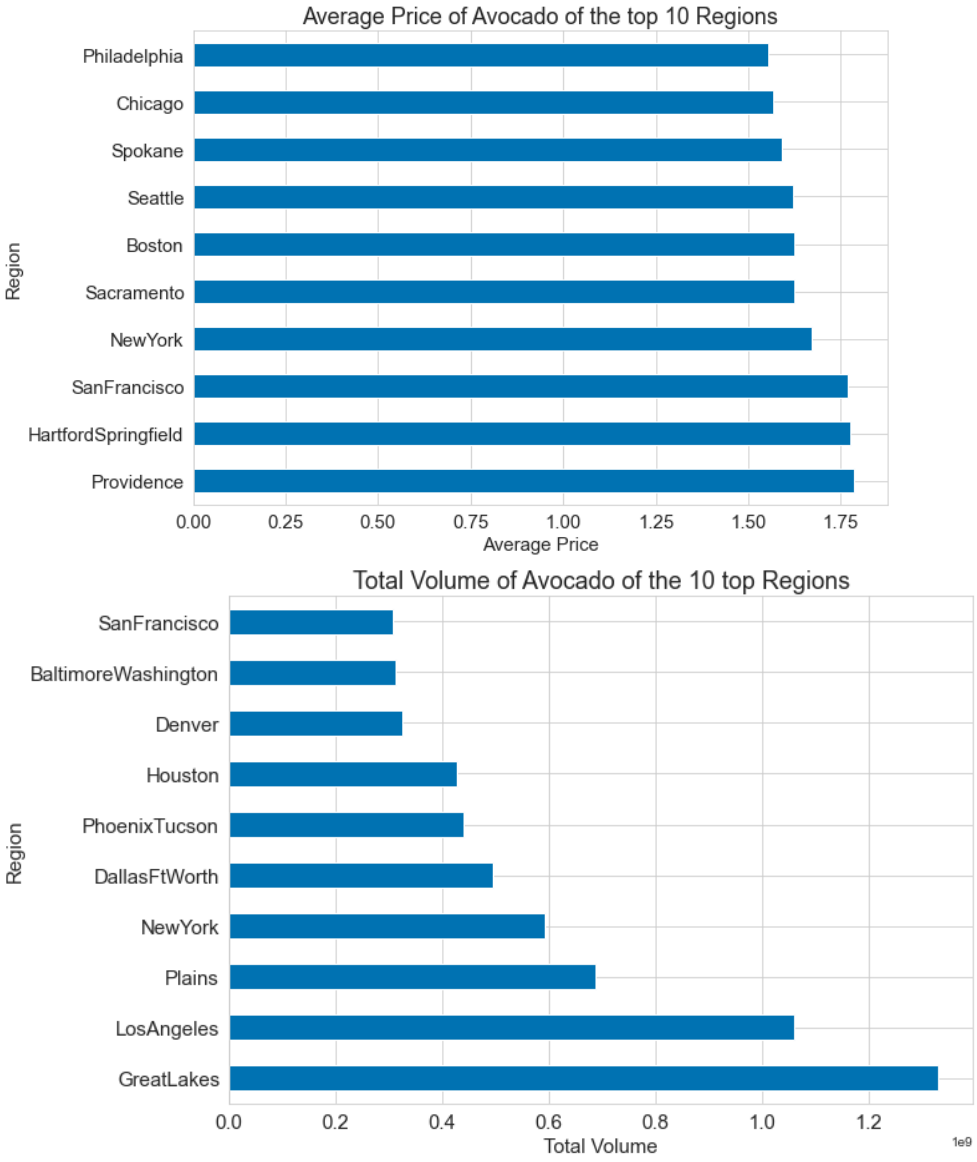


Fig. Average price of Avocado in different regions.

The figure above highlighted a clear distinction between organic and conventional avocado in all regions. In general, the price of organic avocado is higher than the price of conventional avocado.

2. **What are the top 10 regions with high price and high quantities sold of avocados (conventional and organic) since 2015?**

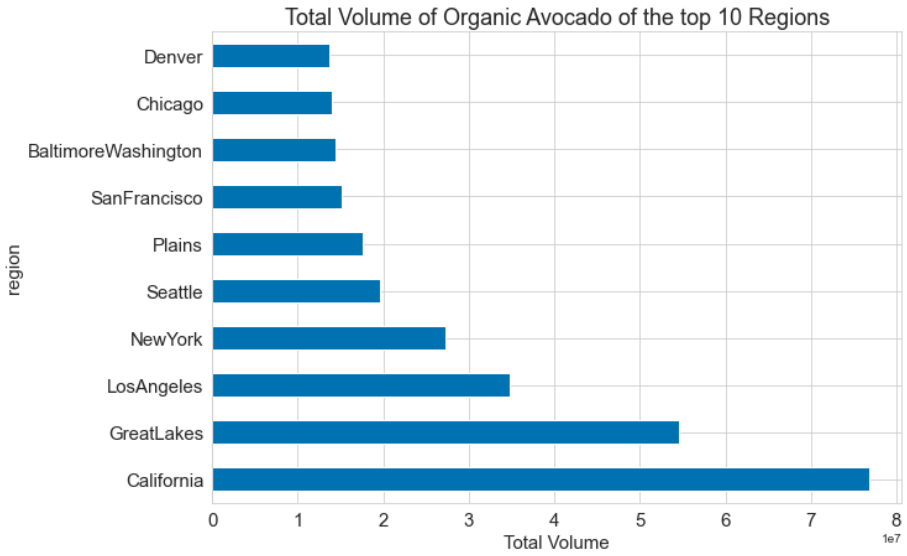


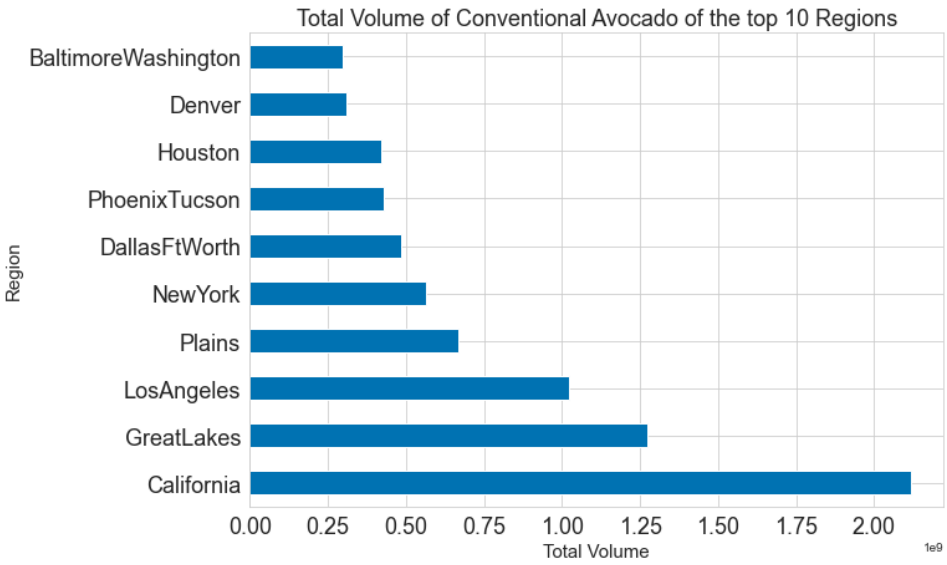
As New York is our target region, he is among the top 10 regions with high price in organic and conventional avocados and, the top 10 regions with total volume sold of organic and conventional avocados.

New York is the 4th region with a high volume of avocado sold, the 1st is Great Lakes, the 2nd Los Angeles, the 3rd is Plains.

New York is also the 4th region with high price of avocados sold, the 1st is Providence, the 2nd is Hartford Springfield, the 3rd San Francisco.

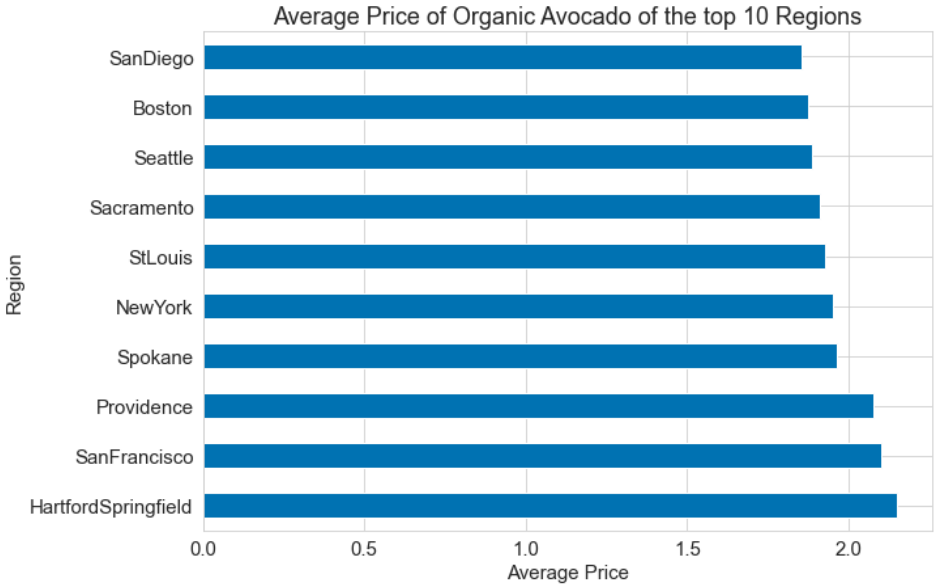
3. **What are the top 10 regions with high quantities of avocado sold since 2015?**

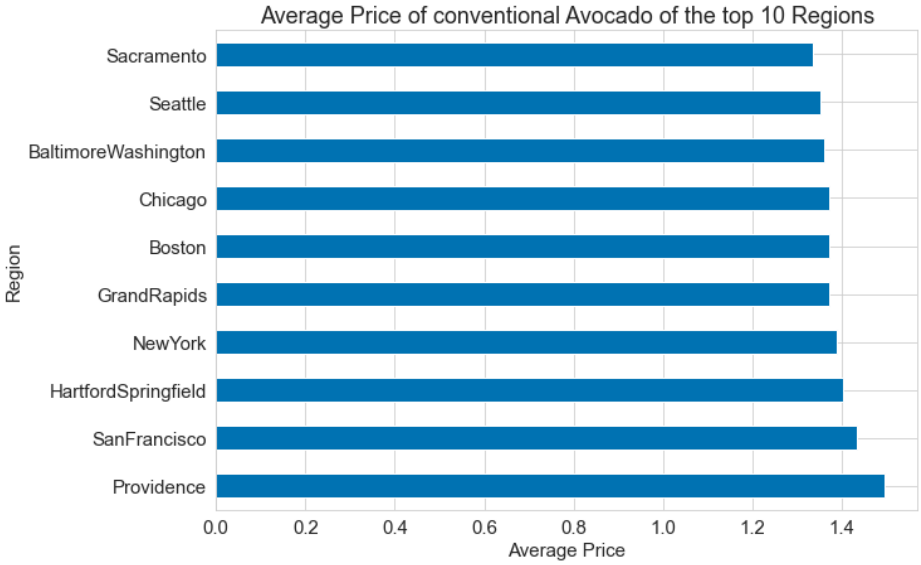




New York is the 4th region with a high volume of organic and the 5th region in conventional avocado, the 1st is California, the 2nd Great Lakes, next is Los Angeles and Plains.

4. What are the top 10 regions with high prices of avocado since 2015?



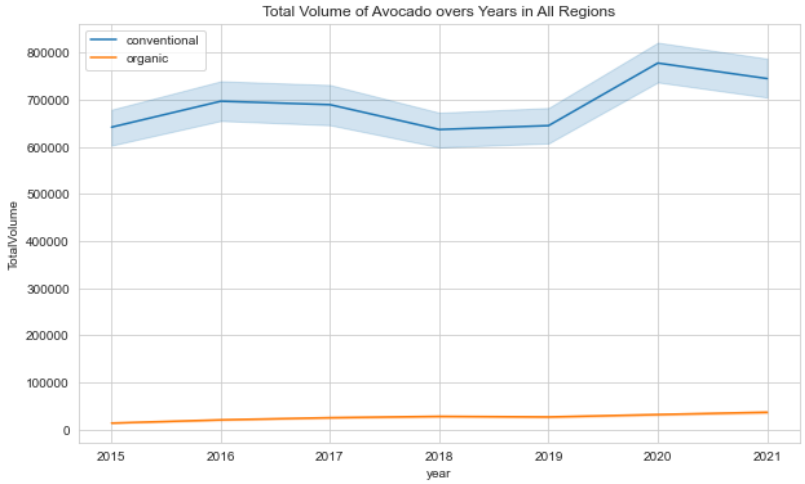


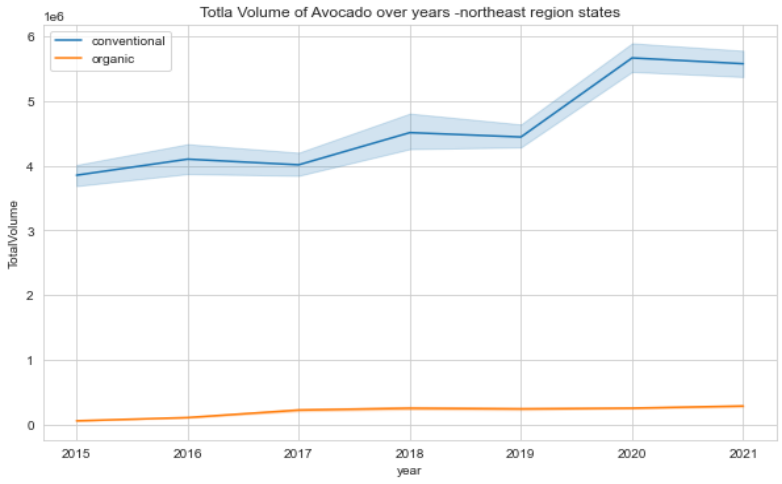
5. What is the price trend of avocados since 2015?

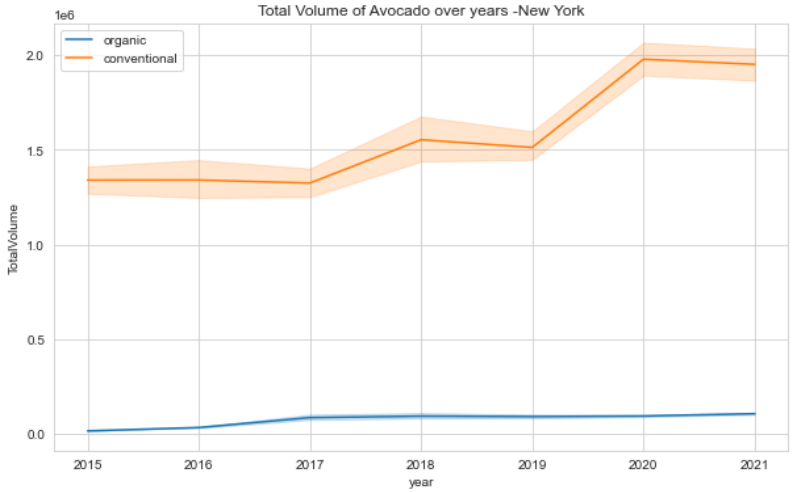
Fig. Price trend of avocados (organic and conventional) from 2015 to 2021

The price of Avocado is different according to the type (conventional or organic). The price of both types dropped down in 2018 and decreased considerably from 2019 to 2021. This observation can be explained with the effect of Covid 19 which starts in 2019, and it is the same in all regions and across scales.

6. What is the trend of avocados sold since 2015?

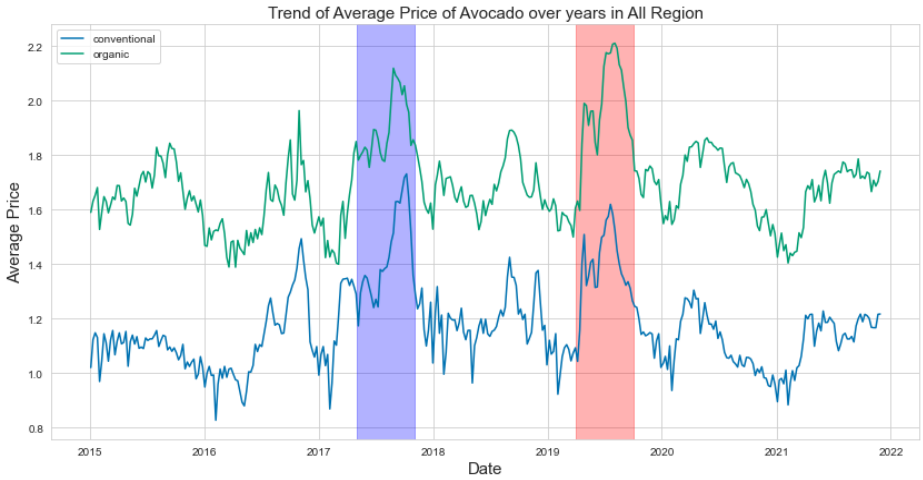


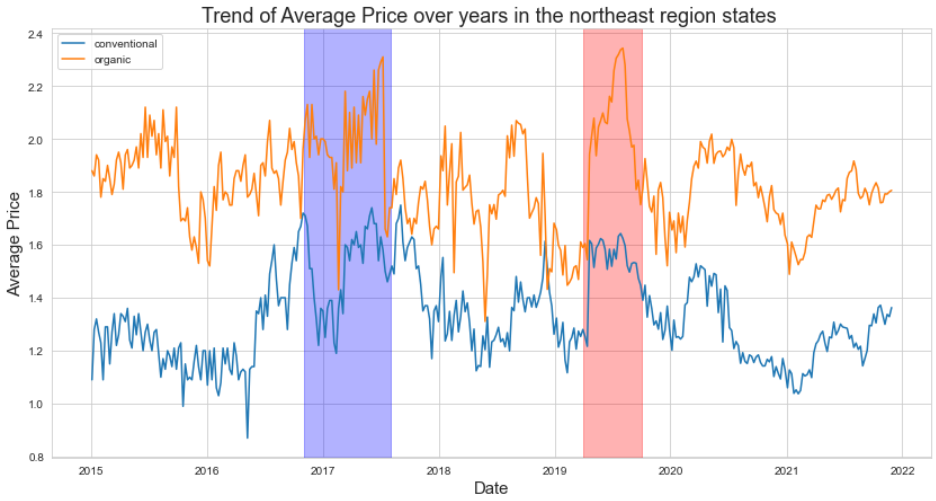


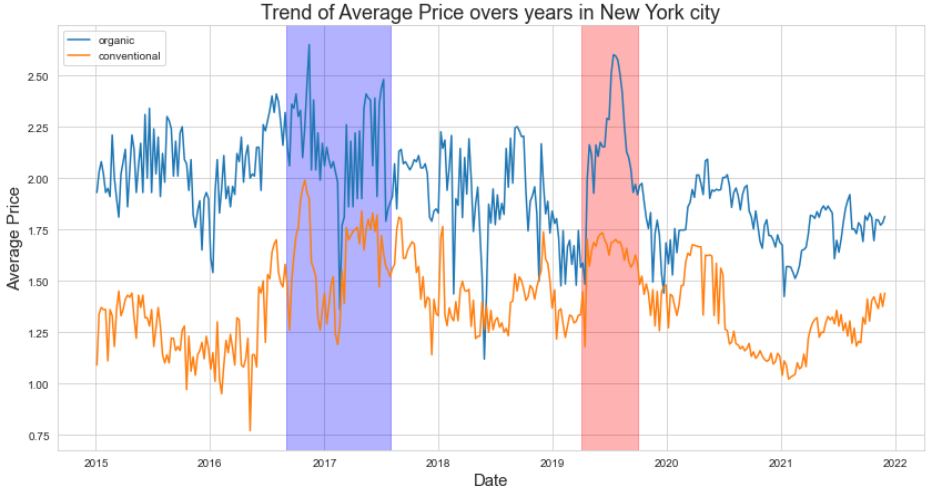


The total volume of organic avocado sold is lower than conventional avocado in all regions. While the price had declined, the volume increased significantly from 2018 to 2021. This finding is consistent across all regions and scales.

7. What was the weekly trend of avocado prices between years?

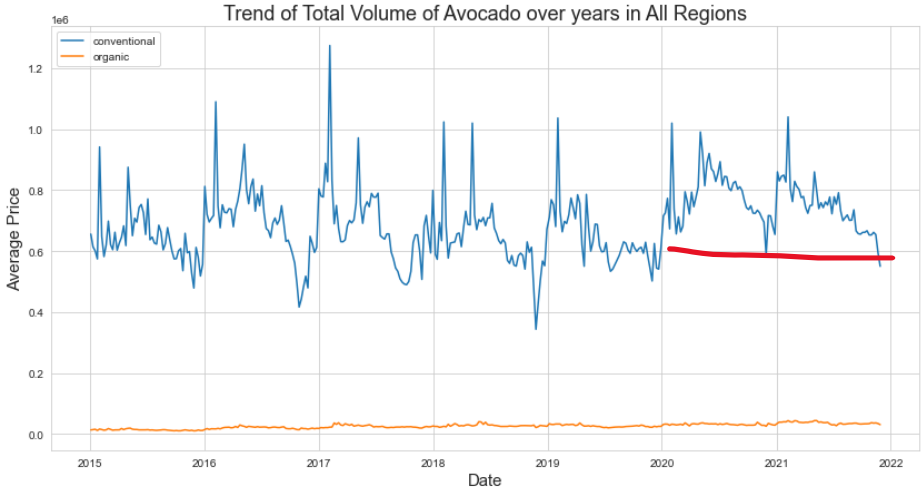


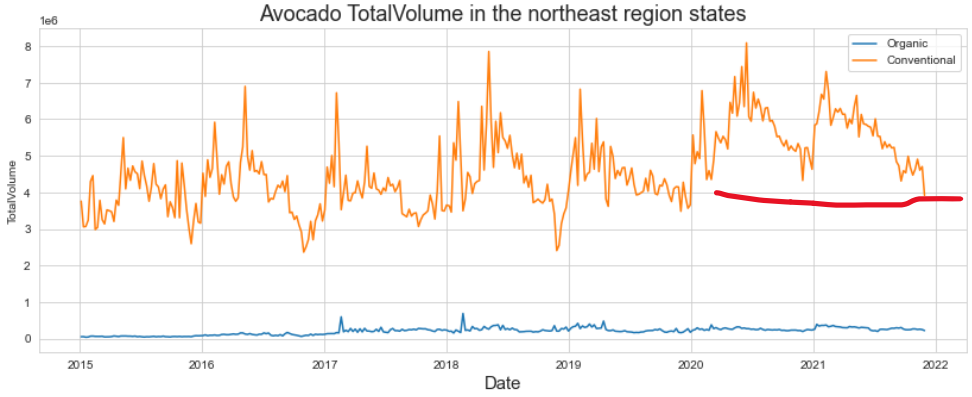


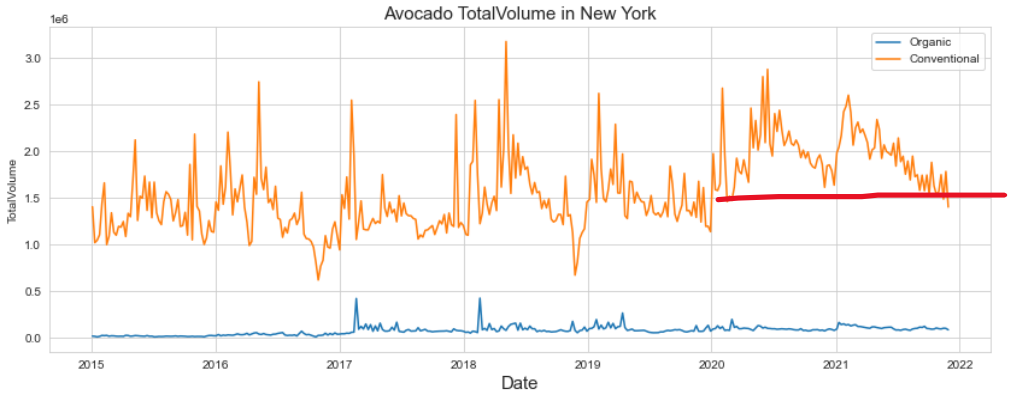


The figures highlighted an important trend in 2019 between March and September. The price of organic and conventional avocado first increased, then decreased until the end of 2019, this decrease has progressed until 2021. Earlier, between the November 2016 and September 2017. the price was high, but instable with many fluctuations.

8. What was the weekly trend of avocado sold (volume) between years?

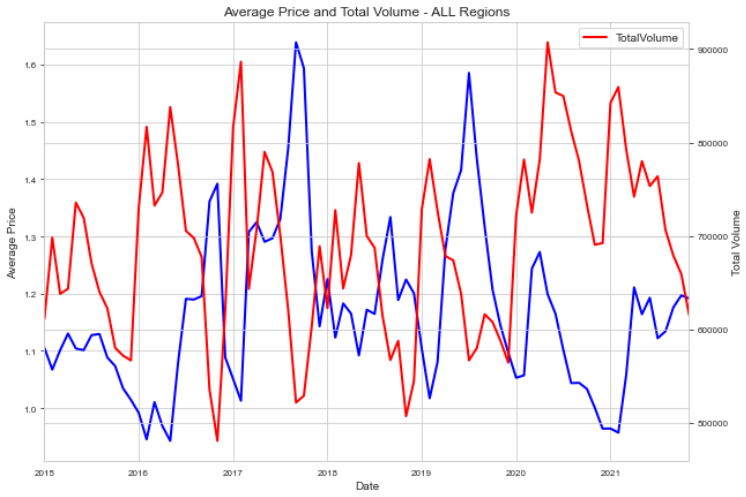


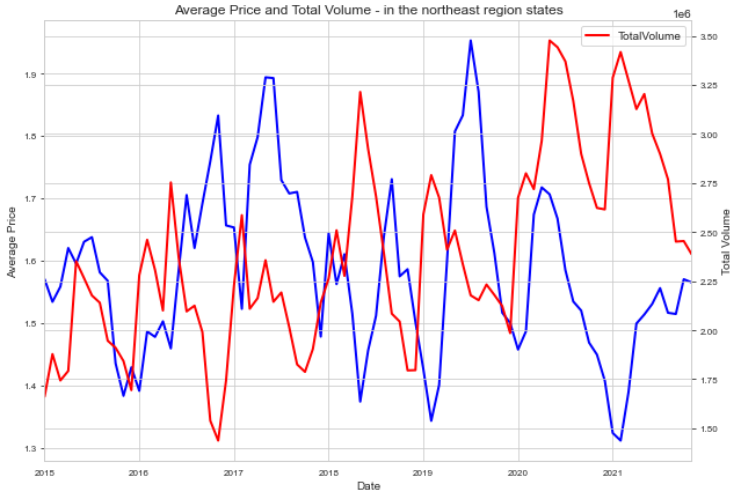


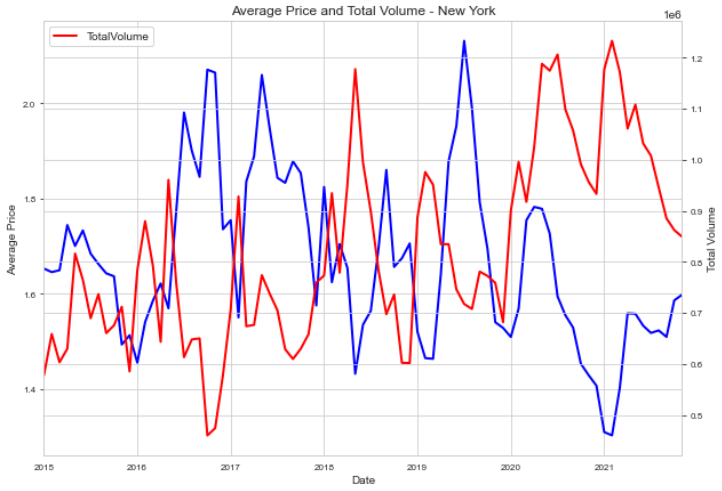


In general, we observe a high consumption peak around the month of January every year and, an increase in consumption in 2020 and 2021.

9. What was the trend of prices and volume trend of avocado over years?







Overall, we observed that when total volume increases, price decreases and when price increases, total volume decreases. The same trend is observed in all regions and across scales (from local scale to upper scale).

10. **Do the type of avocado (organic and conventional) impact the price and the consumption?**

From all the observations above, there is a clear distinction between conventional and organic avocados price and volume. The price of organic avocados is higher than the price of conventional avocados. However, the total volume of conventional avocado sold is higher than the one of organic avocado. From the summary statistic, we grasped the following results:

Table. Summary statistic of Avocado prices and volume in All regions

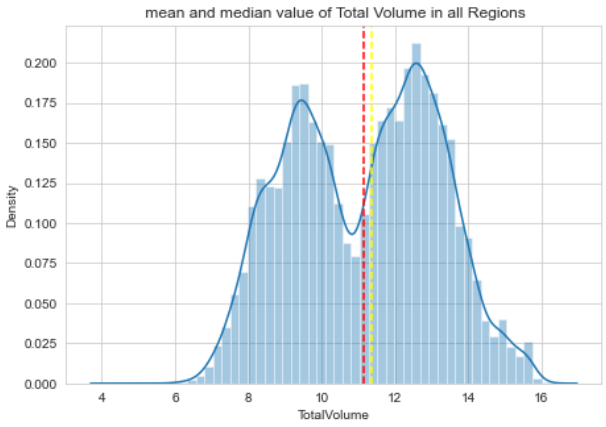
	conventional Avocado		**Organic** Avocado	
	Average Price	**Total Volume**	**Average Price**	**Total Volume**
count	17626.000	17626.000	17626.000	17626.000
mean	1.170	690419.532	1.695	26142.093
std	0.250	1031068.488	0.343	42112.993
min	0.460	33699.680	0.440	84.560
25%	0.995	176106.093	1.475	5545.983
50%	1.143	348225.030	1.678	12508.720
75%	1.325	716016.118	1.896	26719.443
max	2.220	11213596.290	3.250	501557.150

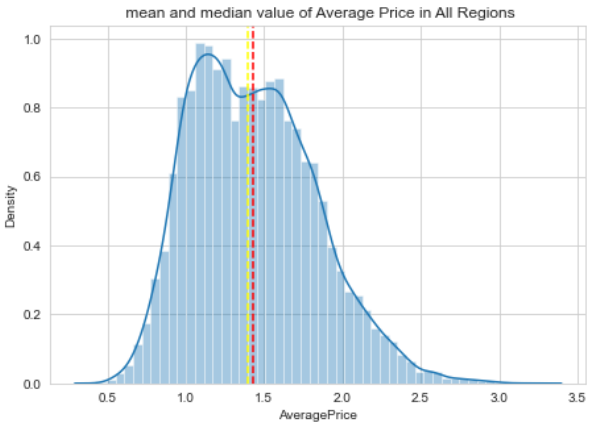
The difference of the average price of conventional versus organic is **0.525,** difference of the average minimum price is **0.020** and max price is **-1,030.**

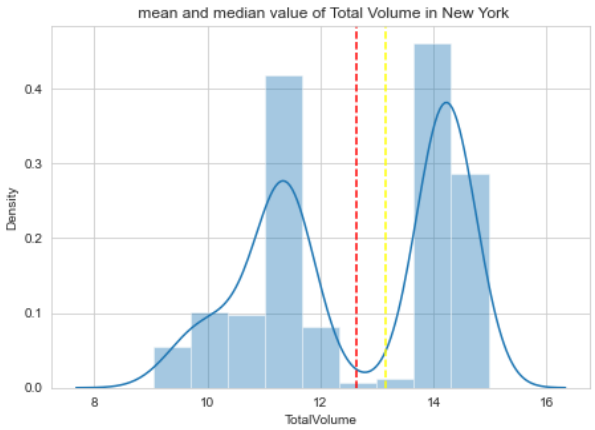
Table. Summary statistic of Avocado prices and volume in New York

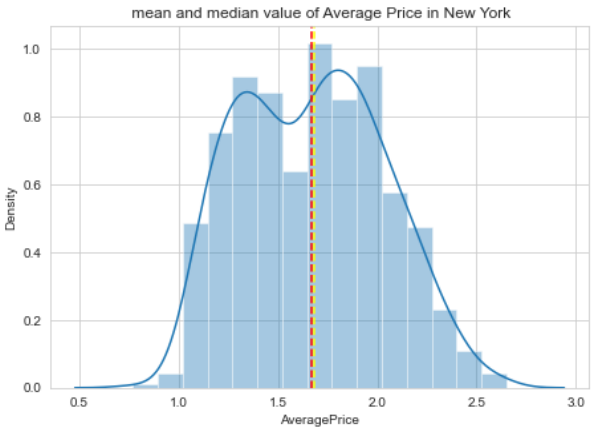
	conventional Avocado		Organic Avocado	
	Average Price	**Total Volume**	**Average Price**	**Total Volume**
count	361.000	361.000	361.000	361.000
mean	1.389	1566999.169	1.952	75484.313
std	0.213	414175.263	0.242	47249.091
min	0.770	618279.770	1.119	8442.790
25%	1.230	1255552.680	1.798	40280.960
50%	1.360	1481997.420	1.931	75479.880
75%	1.540	1838475.070	2.110	97048.630
max	1.990	3172572.870	2.650	424186.840

In New York, the difference between average prices of organic versus conventional is **0.563**, and the difference between volume of organic versus conventional is **1491514.856.**



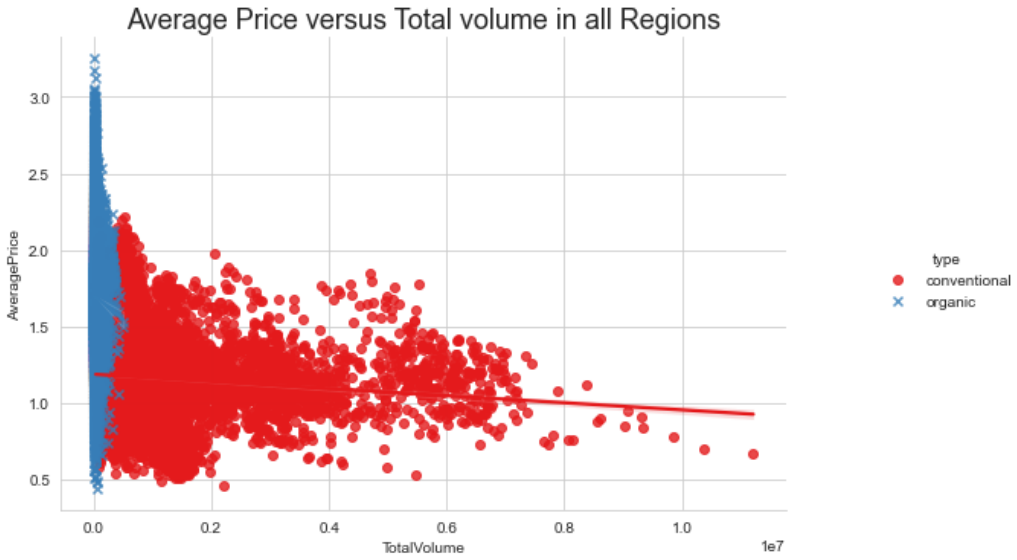


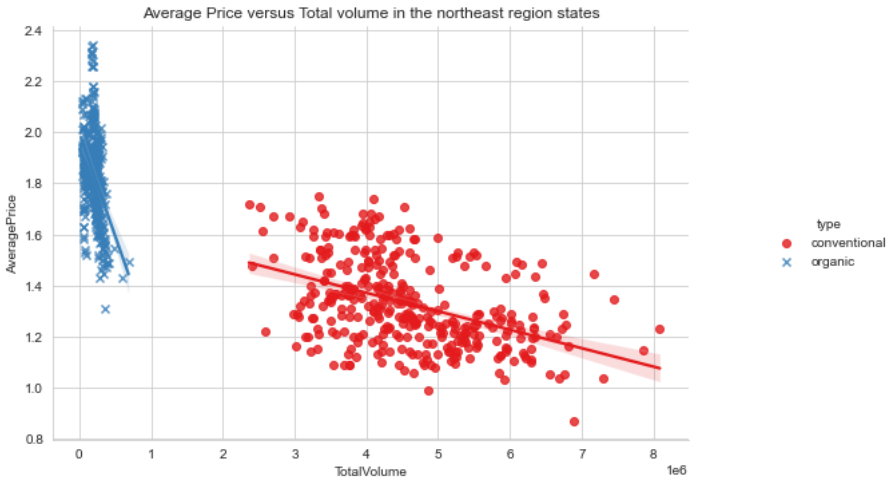


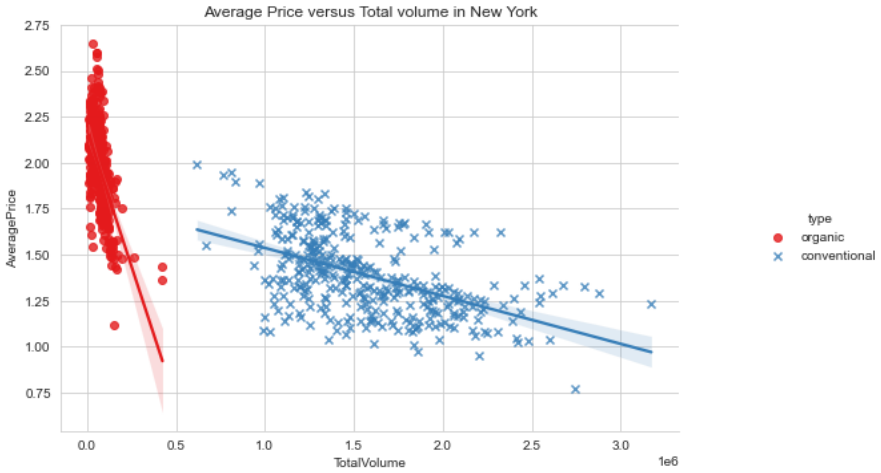


The volume and price distribution are bimodal due to the type of avocado that has different price and consumption. This is an important observation that requires us to split the dataset for forecasting.

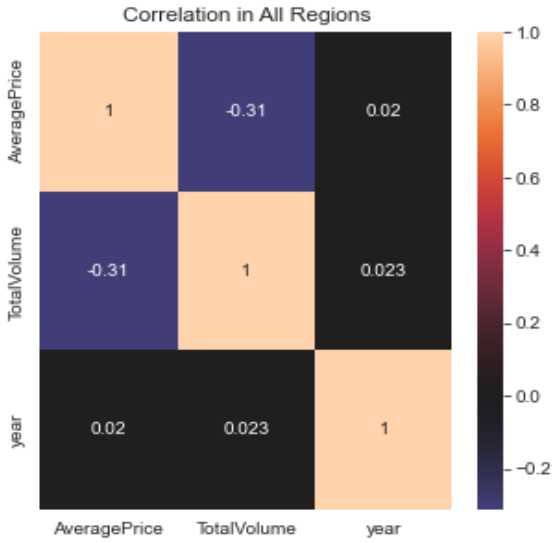
11. What is the relation between the average price, total volume, and avocados type?







From local scale (New York) to upper scale (All regions), the correlation between price and total volume has the same direction (negatively correlated) in both organic and conventional avocados price.



The heatmap reveals that the correlation between price and total volume is -0.31 in all regions, it is -0.44 in Northeast regions and 0,58 in New York. We realized that the correction becomes strong when moving from upper scale to local scale.

12. Seasonal decomposition of the time series

The time series data may have multiplicative or additive components. Decomposition helps us to better analyze the data and explore different ways to build the forecasting model. Time series decomposition involves thinking of a series as a combination of components defined as follows:

- level: the average value in the series.
- trend: the increasing or decreasing value in the series.
- seasonality: the repeating short-term cycle in the series.
- noise: the random variation in the series

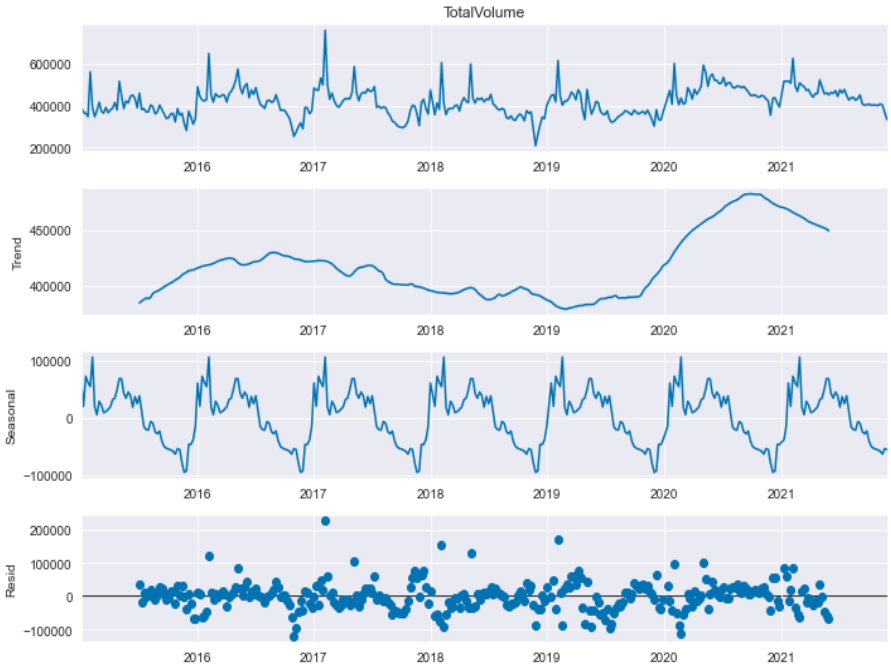


Fig. seasonal decomposition of total volume (all regions)

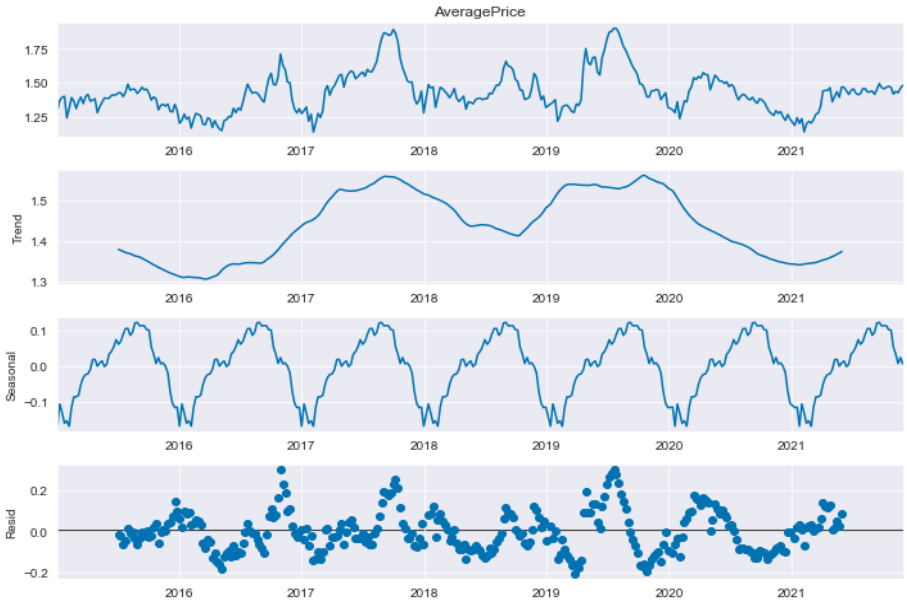


Fig. seasonal decomposition of price (all regions)

The trend shows the decrease in total volume from 2017 to March 2019 and, the increase from March 2019 to 2021. The price was high between 2017 and 2018, and between 2019 and 2020 where it started to decrease until 2021. The seasonal trend shows a pattern that repeats every year.

IV. Forecasting methodology

From the exploratory data analysis, we realized that there is a difference in price of (0.563) and total volume of (-1491514.856) between organic and conventional avocado in New York city. From a business perspective, to have a good idea of avocado price and volume for 2022, the avocado types (organic and conventional) have been considered when forecasting price and volume of avocado. To this end, we forecasted the price of conventional and organic avocado, and forecasted the total volume of conventional and organic avocado.

The time series analysis and forecasting were implemented using the Box-Jenkins Method. For each model, we applied the following steps:

1. Check the stationarity of the time series and define the order of the differencing (d)

A Stationary series is one whose statistical properties like mean, variance, covariance does not vary or change over time, or these stats properties are not the function of time and, the autocorrelation are constant over time. Making the time series stationary is important when forecasting or predicting the future. Just like, regression analysis is good when the variables(predictors) are not correlated again each other, so make the time series stationary, it removes any persistent autocorrelation. In fact, most time series models assume that each point is independent of one another, and it is easier for statistical models to predict effectively and precisely.

We assessed the stationarity of the time series using Augmented Dickey-Fuller Test (ADF test) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests.

Augmented Dickey-Fuller Test is a common statistical test used to test whether a given Time series is stationary or not. Null Hypothesis is that the Time Series is not stationary. It gives a time-dependent trend. Alternate Hypothesis is that the Time Series is stationary. In another term, the series doesn't depend on time.

Interpretation of result: ADF or t Statistic < critical values: Accept the alternative hypothesis. Time series is stationary.

ADF or t Statistic > critical values: Failed to reject the null hypothesis. The time series is non-stationary

if p-value is not less than 0.05 (depend on the alpha), we fail to reject the null hypothesis. This means the time series is non-stationary

Unlike the KPSS test, the null hypothesis is that the series is stationary. Based upon the significance level of 0.05 and the p-value of the KPSS test, if $P < 0.05$, there is evidence for rejecting the null hypothesis in

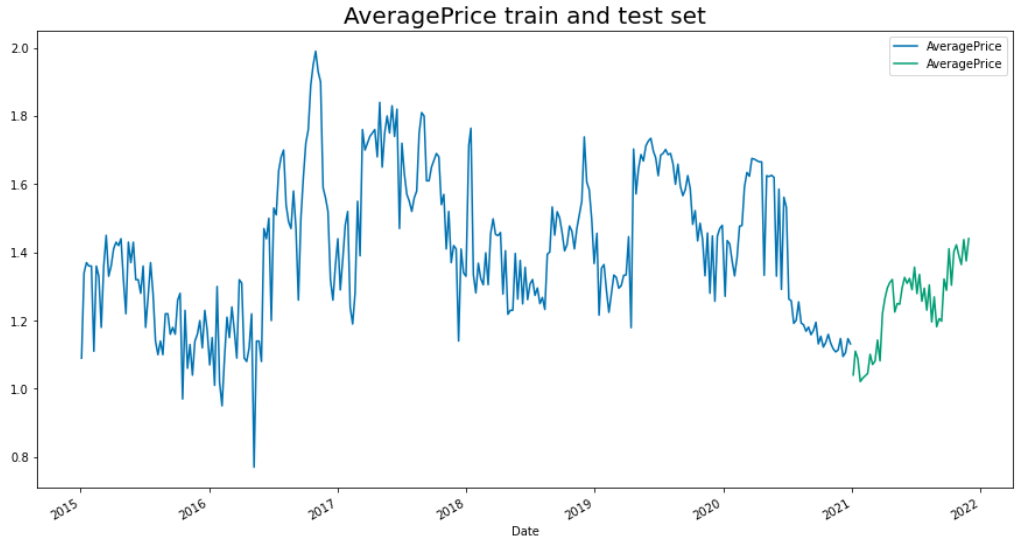
favor of the alternative. Hence, the series is non-stationary as per the KPSS test. Otherwise, the series is stationary.

It is always better to apply both tests, so that it can be ensured that the series is truly stationary. When we are sure that data is stationary, we move to the identification of AR and MA orders.

When the time series is stationary, you can move forward. If it is not, you must make the time series stationary. We use the method of differencing the time series (others method is take the log of ts) which is subtracting the next value by the current value. This can be done more than one time depending on the data. The number of differencing is the order of (d) in the model.

2. Get the train and test set

The dataset was divided into a train set (which considers values from 2015 to 2020) and the test set which was the data of 2021 year.



3. identification of AR and MA orders

The next step was to find the order of the AR term (p) with partial autocorrelation function(pacf), and find the order of the MA term (q) with autocorrelation function (acf). **The basic guideline for interpreting the ACF and PACF plots are as** following**:**

1. Look for tail off pattern in either ACF or PACF.
2. If tail off at ACF ⟶ AR model ⟶ Cut off at PACF will provide order p for AR(p).
3. If tail off at PACF ⟶ MA model ⟶ Cut off at ACF will provide order q for MA(q).
4. Tail of at both ACF and PACF ⟶ ARMA model.

It was not evident to get a clear result from the autocorrelation and partial autocorrelation. Instead, it gives an idea of the range of values of p and q to search on. We searched over a range of values where we computed the 'Akaike's Information Criteria' AIC and 'Bayesian Information Criteria' BIC to get the best indication of p, q, and d. In fact, it is important to notice that:

- BIC favors simpler models than AIC,
- Lower the AIC, better the model and lower BIC indicates a better model,
- AIC is better at choosing predictive models,
- BIC is better at choosing good explanatory model,
- In the context of this project, we will look for lower AIC --- for best predictive model,

Next, we explored the seasonal part of the time series with the use of pmadarima and its auto_arima function to identify the P,Q,D for the seasonal part. We considered 52 as the seasonal order.

4. Train de model

The identified p, q and d were used with SARIMAX and ARIMA algorithms. We computed the evaluation of the model by plotting the residual (result from plot diagnostic function is 4 plots, **standardized residual**, **histogram plus estimated density, Normal Q-Q, correlogram**) and analyzing result of the summary statistic. From the summary table, **Prob(Q)** is the p-value associated with the null hypothesis that residuals have no correlation structure. **Prob(JB**) is the p-value associated with the null hypothesis that residuals are Guassian normally distributed. If either p-value is less than 0.05 we reject that hypothesis.

5. Evaluation of the model

After the prediction, we compute the evaluation on predicted and actual price (or volume) of avocado (test set). Mean Absolute Percentage Error (MAPE), Mean Error (ME), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Correlation (Corr) were computed.

Based on the weakness and strength of each **metric** , we considered MAE and MAPE to compare models . MAE is a scale dependent metrics, robust to outliers, and we worked with time series of the same scale across all models, and MAPE is scale-independency metric and easily interpretable.

Mean Absolute Error (MAE) = np.mean(np.abs(forecast - actual))

Mean Absolute Percent Error (MAPE) = 100 * np.mean(np.abs(forecast - actual)/np.abs(actual))

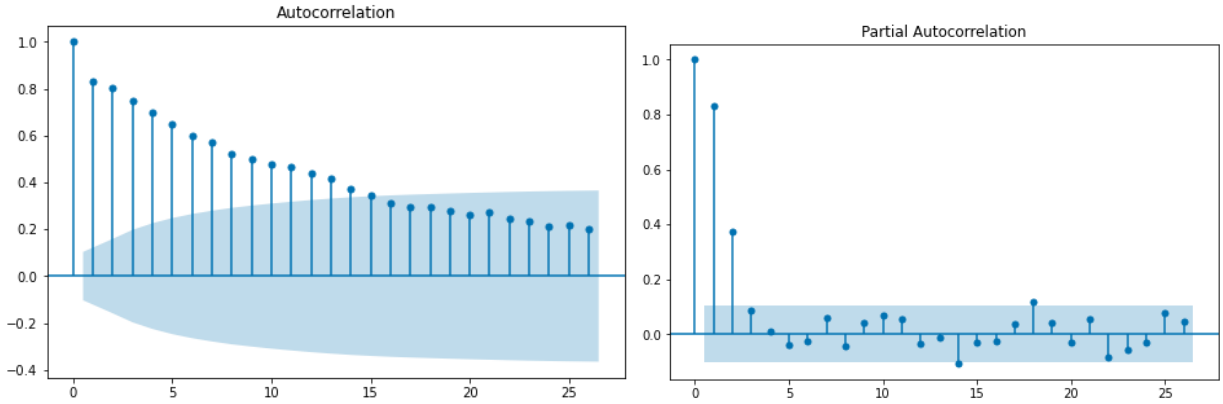
6. Forecasting the 2022 values

Price and total volume for 2022 was forecasted using the best model. We plot the result to allow the company to take a decision accordingly.

7. Application of forecasting methodology for the model development: Case of the avocado conventional price

1.Stationarity: According to ADF test (Average Price : P-Value = 0.004 => Stationary)and KPSS test (Average Price : P-Value = 0.1 => stationary)

2. **Find the order of the AR term (p) -pacf /find the order of the MA term (q)-acf**



From the plot of ACF and PACF above, it suggested that model is a AR(2) , with p =2 and q = 0

3. **Evaluate different p, d, and q** values - choice of best model according to AIC and BIC

We loop over a range of value form (0 to 3) of p and q, compute the Sanimax model on price and evaluate. We sorted the result to get lower AIC.

	p	q	d	AIC	BIC
1st	1	1	0	-438.909479	-427.670869
2nd	3	1	0	-438.894161	-420.163145

According to this result, the best is an ARMA model with p= 1 and q =1

We explored the seasonal with auto_arima function to identify the P,Q,D

4.**Train and evaluate the model** with monthly time series and monthly time series dataset

V. Results of the forecasting of price and volume of Avocado in the city of New York

The following results present the forecasting price and volume sold of avocado in New York City. Because of the difference that exists between conventional and organic avocado, 8 models were developed for: (i) the price of organic avocado, (ii) the price of conventional avocado, (iii) the volume of conventional avocado and, (iv) the volume of organic avocado. The results are presented in the following order:

- First, the table gathered all the models we explored to forecast the price and volume of avocados to be sold.
- Next, the table presented the results of model evaluations with different metrics.
- Next, we plotted the residual of the best model with plot diagnostic.
- Next, we predicted and plotted the prediction, the train and test sets.
- Finally, we presented the predicted values in a table

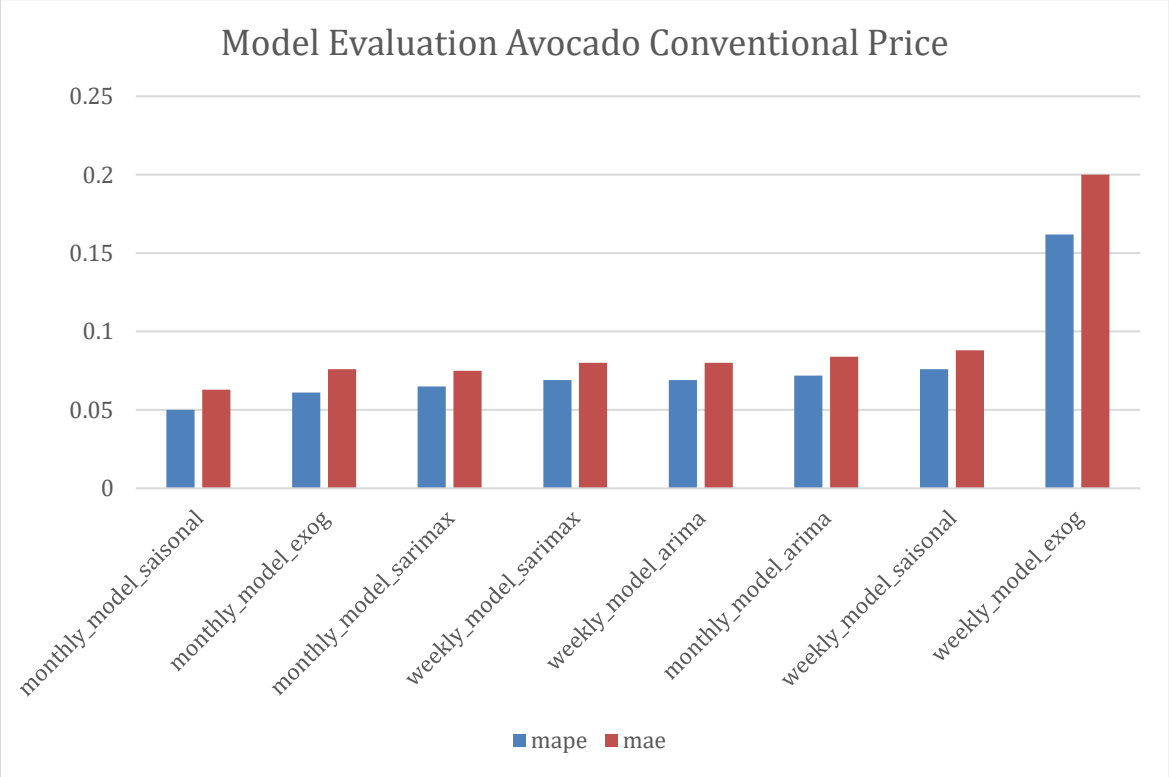
1. Conventional Avocado price

Table. New York conventional avocado price models descriptions

Model name	p,d,q non seasonal part	P,Q,D seasonal part	Used of exogenous variable or not	Name of results metric
results_1	3,0,1	-		monthly_model_sarimax
results_2	1,0,1	-		weekly_model_sarimax
results_3	1,0,1	-		weekly_model_arima
results_4	8,0,1	-		monthly_model_arima
results_s	3,0,0	0,0,0,52		weekly_model_saisonal
results_s2	1,0,0	0,0,1,52		monthly_model_saisonal
results_ex	1,0,1	0,0,1,52	with volume as exogenous	weekly_model_exog
results_ex2	1,0,1	0,0,0,52	with volume as exogenous	monthly_model_exog

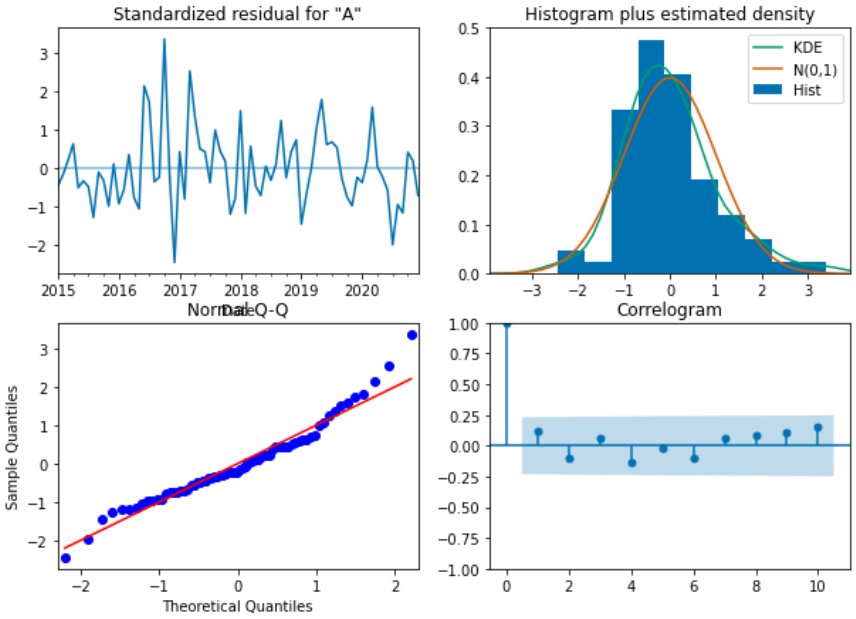
Table. New York conventional avocado price models Evaluation

	mape	me	mae	mpe	rmse	corr
monthly_model_saisonal	0.050	-0.024	0.063	-0.015	0.082	0.712
monthly_model_exog	0.061	-0.026	0.076	-0.017	0.094	0.618
monthly_model_sarimax	0.065	0.067	0.075	0.058	0.094	0.882
weekly_model_sarimax	0.069	0.066	0.080	0.059	0.100	0.834
weekly_model_arima	0.069	0.066	0.080	0.059	0.100	0.834
monthly_model_arima	0.072	0.073	0.084	0.064	0.102	0.905
weekly_model_saisonal	0.076	0.078	0.088	0.069	0.109	0.831
weekly_model_exog	0.162	0.062	0.200	0.060	0.242	-0.093



From the evaluation above, the best model is (1,0,0, 0,0,1),52

New York conventional avocado price - Best Model - Residual Evaluation and plotting of forecasting values
Model specification (1,0,0,0,0,1),52



Result from plot diagnostic function is 4 plots, standardized residual, density plot, Normal Q-Q, correlogram

- The line plot of residual showed the residual error spread around zero, the variance is uniform.
- The density plot of residual is center around zero, that is good, forecast errors is normally distributed around a zero mean.
- The residual Q_Q plot (quantile plot) can be used to quickly check the normality of the distribution of residual errors. The values are ordered and compared to an idealized Gaussian distribution. Q-Q plot showed that the distribution is seemingly normal with a few bumps and outliers.
- The residual autocorrelation Plot express strength of the relationship between an observation and observations at prior time steps. The correlogram showed that the residual errors are not autocorrelated, in fact, any autocorrelation would imply that there is some pattern in the residual errors which are not explained in the model.

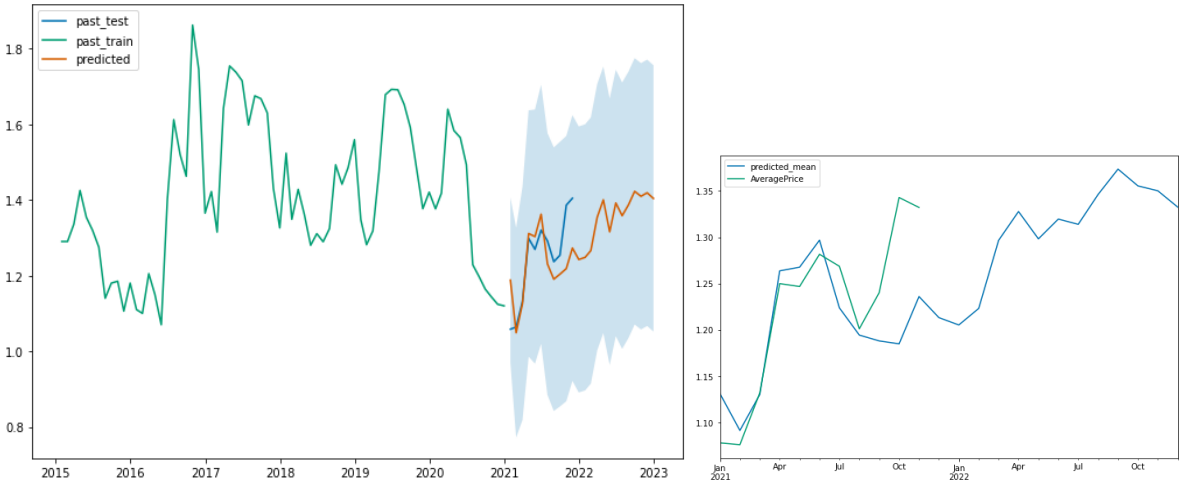


Fig. Plotting of train, test, predicted and forecasted value of conventional avocado price

Table. New_york conventional avocado price - Best Model – values of forecasting price

	lower AveragePrice	upper AveragePrice	pred_mean_AveragePrice	AveragePrice	Diff (pred – actual)
2021-01-31 00:00:00	0.970	1.406	1.188	1.058	0.130
2021-02-28 00:00:00	0.772	1.326	1.049	1.064	-0.015
2021-03-31 00:00:00	0.818	1.434	1.126	1.132	-0.006
2021-04-30 00:00:00	0.986	1.637	1.311	1.298	0.013
2021-05-31 00:00:00	0.967	1.639	1.303	1.269	0.034
2021-06-30 00:00:00	1.020	1.704	1.362	1.320	0.042
2021-07-31 00:00:00	0.885	1.577	1.231	1.292	-0.061
2021-08-31 00:00:00	0.842	1.539	1.190	1.237	-0.046
2021-09-30 00:00:00	0.854	1.553	1.204	1.254	-0.050
2021-10-31 00:00:00	0.868	1.569	1.218	1.386	-0.168
2021-11-30 00:00:00	0.922	1.624	1.273	1.404	-0.132
2021-12-31 00:00:00	0.891	1.594	1.242		
2022-01-31 00:00:00	0.896	1.600	1.248		
2022-02-28 00:00:00	0.914	1.618	1.266		
2022-03-31 00:00:00	1.002	1.706	1.354		
2022-04-30 00:00:00	1.048	1.752	1.400		
2022-05-31 00:00:00	0.964	1.668	1.316		
2022-06-30 00:00:00	1.040	1.744	1.392		
2022-07-31 00:00:00	1.006	1.710	1.358		
2022-08-31 00:00:00	1.035	1.739	1.387		
2022-09-30 00:00:00	1.071	1.775	1.423		
2022-10-31 00:00:00	1.058	1.761	1.409		
2022-11-30 00:00:00	1.067	1.771	1.419		
2022-12-31 00:00:00	1.052	1.756	1.404		

From the result, we found that our model did good. The forecasted prices are close to real values, the average of the difference between this value (pred – actual) is -0,024, the MAE is 0.063 and the MAPE is 0,050. More, the Pearson correlation is 0,712. The model suggested that price in January 2022 is (lower =0.89, mean=1.2, upper=1.59).

2. Conventional Avocado volume

Table. New York conventional avocado volume models descriptions

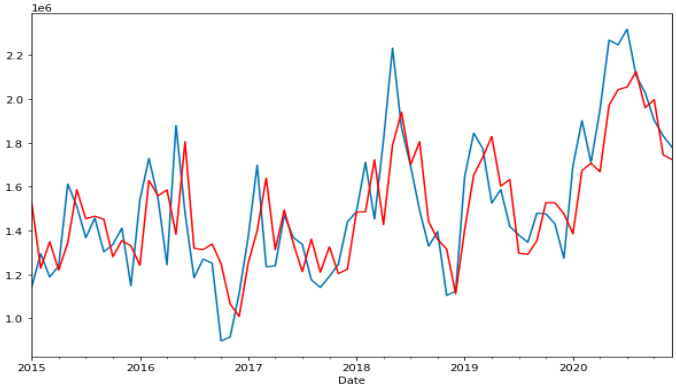
Model name	P,d,q value non seasonal part	P,Q,D for seasonal part	Used of exogenous variable or not	Name of results metric
results_1	8,0,0	-		monthly_model_sarimax
results_2	4,0,1	-		weekly_model_sarimax
results_3	8,0,0	-		weekly_model_arima
results_4	4,0,1	-		monthly_model_arima
results_s	3,0,2	3,0,1,52		weekly_model_saisonal
results_s2	1,0,0	0,0,0,52		monthly_model_saisonal
results_ex	2,0,1	3,0,0,52	with volume as exogenous	weekly_model_exog
results_ex2	1,0,0	0,0,0,52	with volume as exogenous	monthly_model_exog

Table. New_york conventional avocado volume models evaluation

	mape	**me**	**mae**	**mpe**	**rmse**	**corr**
monthly_model_sarimax	0.106	-191532.395	224116.165	-0.086	283305.086	0.760
weekly_model_saisonal	0.110	-213656.709	231941.163	-0.098	305496.983	0.625
monthly_model_arima	0.112	-223441.367	237966.027	-0.103	303116.595	0.827
weekly_model_exog	0.118	-189511.939	247017.485	-0.081	322500.748	0.351
weekly_model_arima	0.178	-364651.611	371772.309	-0.173	436066.129	0.782
monthly_model_saisonal	0.184	-378599.327	378599.327	-0.184	426120.831	0.856
weekly_model_sarimax	0.202	-415257.134	420889.767	-0.198	488505.476	0.566
monthly _model_exog	0.219	-445971.849	445971.849	-0.219	491902.224	0.684

New_york conventional avocado volume - Best Model - Residual Evaluation and plotting of forecasting values

Specification of the Best Model (8,0,0)



Fig, plotting of the prediction again the actual value of the training set

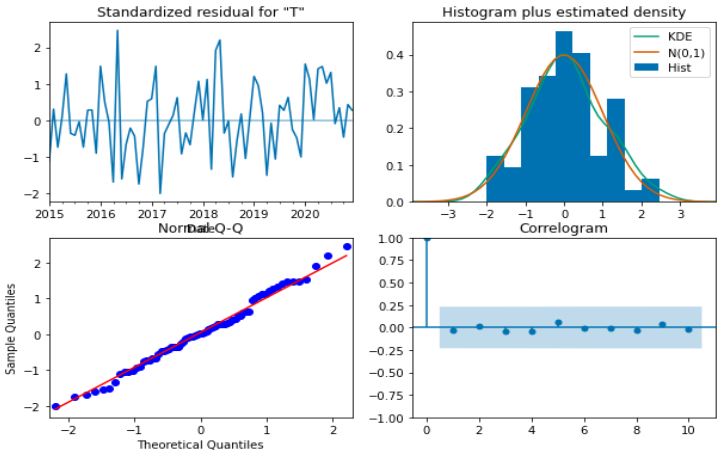


Fig. plot diagnostic result.

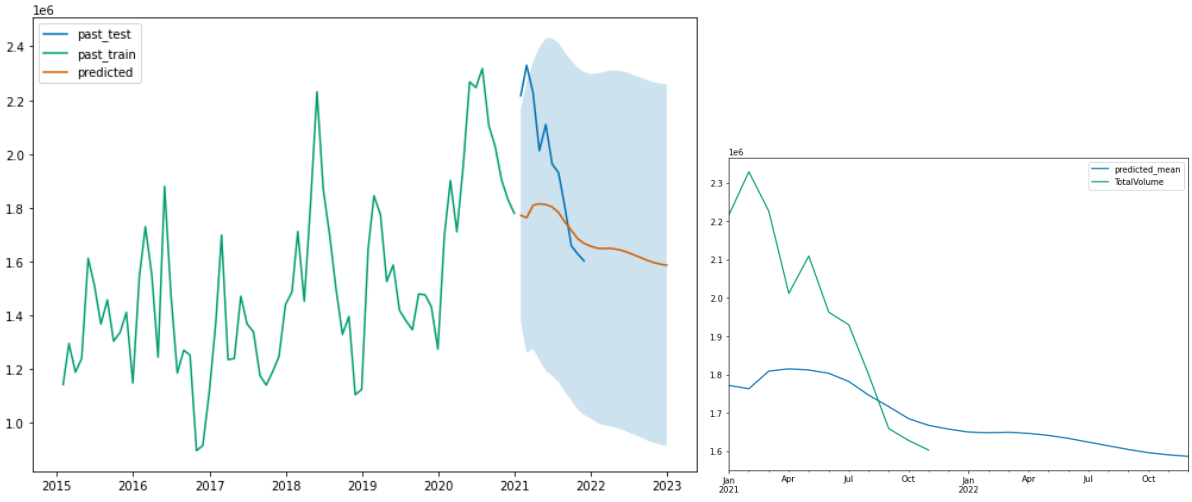


Fig. plotting of train, test, predicted and forecasted value of conventional avocado volume

Table. New York conventional avocado volume- Best Model – values of forecasting volume

	lower TotalVolume	upper TotalVolume	pred_mean_Total Volume	TotalVolume	DIFF (pred – actual)
2021-01-31 00:00:00	1377615	2165795	1771705	2216830	-445125
2021-02-28 00:00:00	1260417	2265373	1762895	2329957	-567062
2021-03-31 00:00:00	1275547	2342736	1809141	2226937	-417796
2021-04-30 00:00:00	1234655	2395020	1814838	2012080	-197242
2021-05-31 00:00:00	1193795	2430216	1812006	2109843	-297837
2021-06-30 00:00:00	1175840	2430665	1803252	1963124	-159872
2021-07-31 00:00:00	1151582	2412515	1782049	1929996	-147947
2021-08-31 00:00:00	1115699	2376650	1746174	1799359	-53185.2
2021-09-30 00:00:00	1084636	2348101	1716368	1658860	57508.19
2021-10-31 00:00:00	1049011	2320348	1684679	1627687	56992.81
2021-11-30 00:00:00	1029969	2304619	1667294	1602584	64709.73
2021-12-31 00:00:00	1016893	2297723	1657308		
2022-01-31 00:00:00	1001010	2298757	1649883		
2022-02-28 00:00:00	992011.6	2303782	1647897		
2022-03-31 00:00:00	988454.6	2310240	1649347		
2022-04-30 00:00:00	981965.4	2310419	1646192		
2022-05-31 00:00:00	974733.7	2307491	1641112		
2022-06-30 00:00:00	965104.7	2301232	1633168		
2022-07-31 00:00:00	954550.7	2292315	1623433		
2022-08-31 00:00:00	944542	2282985	1613763		
2022-09-30 00:00:00	934400	2274017	1604209		
2022-10-31 00:00:00	925189.5	2266368	1595779		
2022-11-30 00:00:00	919018	2261635	1590326		
2022-12-31 00:00:00	914250.7	2258454	1586353		

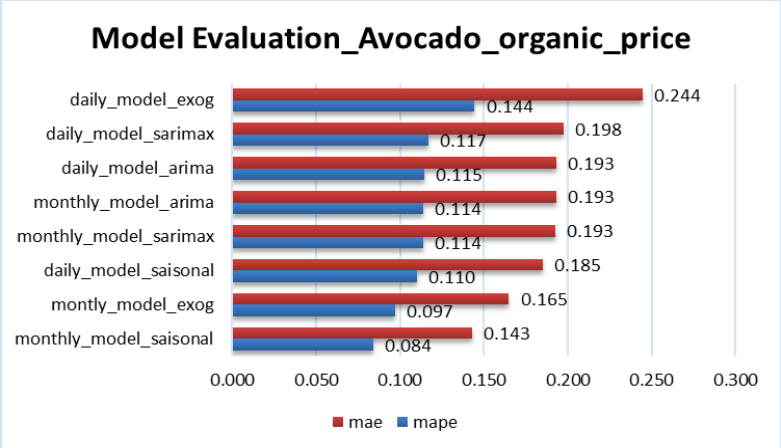
3. Organic Avocado price

Table. New York organic avocado price models descriptions

Model name	p,d,q non seasonal part	P,Q,D seasonal part	Used of exogenous variable or not	Name of results metric
results_1	8,0,0	-		monthly_model_sarimax
results_2	6,0,1	-		weekly_model_sarimax
results_3	6,0,1	-		weekly_model_arima
results_3	8,0,0	-		monthly_model_arima
results_s	2,0,3	0,0,0,52		weekly_model_saisonal
results_s2	1,0,0	1,0,0,52		monthly_model_saisonal
results_ex	3,0,2	2,0,0,52	with volume as exogenous	weekly_model_exog
results_ex2	2,0,0	0,0,0,52	with volume as exogenous	monthly_model_exog

Table. New York organic avocado price models evaluation

	mape	me	mae	mpe	rmse	corr
monthly_model_saisonal	0.084	0.143	0.143	0.084	0.169	0.594
monthly_model_exog	0.097	0.165	0.165	0.097	0.178	0.729
weekly_model_saisonal	0.110	0.185	0.185	0.110	0.202	0.724
monthly_model_sarimax	0.114	0.193	0.193	0.114	0.208	0.575
monthly_model_arima	0.114	0.193	0.193	0.114	0.209	0.574
weekly_model_arima	0.115	0.193	0.193	0.115	0.209	0.730
weekly_model_sarimax	0.117	0.198	0.198	0.117	0.213	0.735
weekly_model_exog	0.144	0.244	0.244	0.144	0.259	0.679



New York organic avocado price - Best Model - Residual Evaluation and plotting of forecasting values
Specification of the Best Model (1,0,0 1,0,0)52

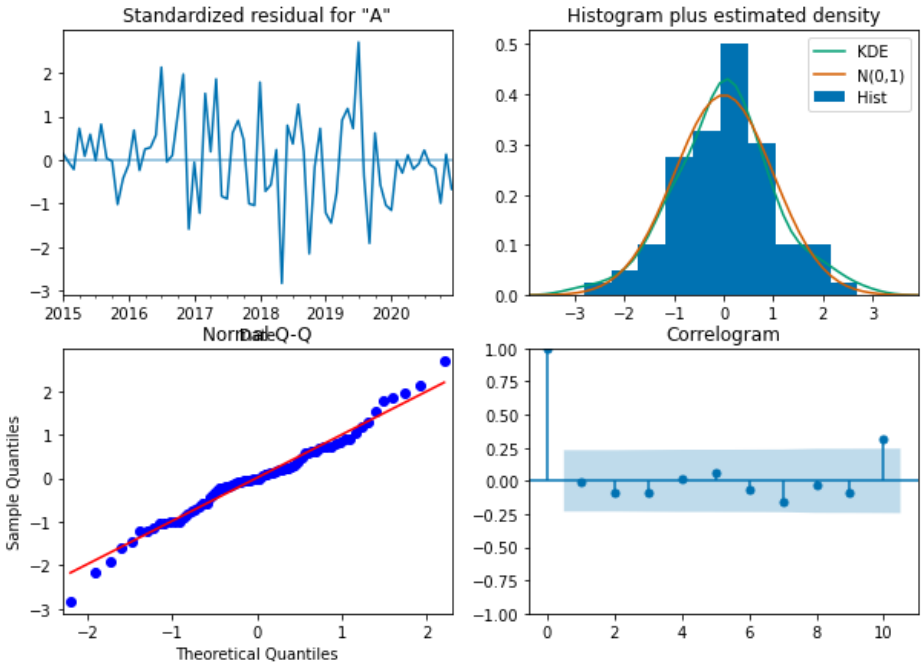


Fig. plot diagnostic result.

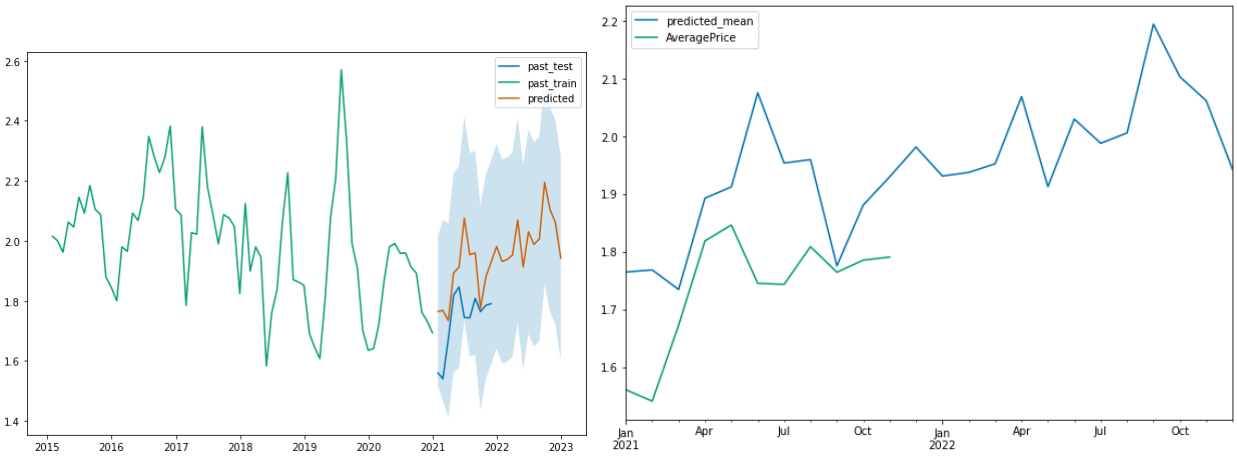


Fig. plotting of train, test, predicted and forecasted value of organic avocado price

Table. New York organic avocado price- Best Model – values of forecasting price

	lower AveragePrice	upper AveragePrice	pred_mean_AveragePrice	AveragePrice	DIFF (pred – actual)
2021-01-31 00:00:00	1.517	2.013	1.765	1.560	0.204
2021-02-28 00:00:00	1.468	2.069	1.768	1.541	0.228
2021-03-31 00:00:00	1.412	2.056	1.734	1.672	0.062
2021-04-30 00:00:00	1.561	2.225	1.893	1.819	0.074
2021-05-31 00:00:00	1.577	2.248	1.912	1.846	0.066
2021-06-30 00:00:00	1.738	2.414	2.076	1.745	0.331
2021-07-31 00:00:00	1.615	2.293	1.954	1.743	0.211
2021-08-31 00:00:00	1.620	2.299	1.960	1.809	0.151
2021-09-30 00:00:00	1.436	2.115	1.775	1.764	0.011
2021-10-31 00:00:00	1.541	2.221	1.881	1.785	0.096
2021-11-30 00:00:00	1.590	2.270	1.930	1.791	0.139
2021-12-31 00:00:00	1.642	2.321	1.982		
2022-01-31 00:00:00	1.592	2.271	1.931		
2022-02-28 00:00:00	1.598	2.277	1.938		
2022-03-31 00:00:00	1.613	2.292	1.953		
2022-04-30 00:00:00	1.729	2.409	2.069		
2022-05-31 00:00:00	1.573	2.253	1.913		
2022-06-30 00:00:00	1.691	2.370	2.030		
2022-07-31 00:00:00	1.648	2.328	1.988		
2022-08-31 00:00:00	1.666	2.346	2.006		
2022-09-30 00:00:00	1.855	2.535	2.195		
2022-10-31 00:00:00	1.764	2.443	2.103		
2022-11-30 00:00:00	1.722	2.402	2.062		
2022-12-31 00:00:00	1.603	2.282	1.943		

4. Organic Avocado volume

Table. New York Organic avocado volume models descriptions

Model name	p,d,q non seasonal part	P,Q,D seasonal part	Used of exogenous variable or not	Name of results metric
results_1	2,0,0	-		monthly_model_sarimax
results_2	5,0,1	-		weekly_model_sarimax
results_3	5,0,1	-		weekly_model_arima
results_3	2,0,0	-		monthly_model_arima
results_s	5,0,0	0,0,1,52		weekly_model_saisonal
results_s2	2,0,0	1,0,1,52		monthly_model_saisonal
results_ex	2,0,2	0,0,1,52	with volume as exogenous	weekly_model_exog
results_ex2	2,0,0	1,0,0,52	with volume as exogenous	monthly_model_exog

Table. New York Organic avocado volume models evaluation

	mape	**me**	**mae**	**mpe**	**rmse**	**corr**
weekly_model_exog	0.238	-27491.991	27491.991	-0.238	32114.376	0.505
weekly_model_arima	0.259	-29932.473	29932.473	-0.259	34560.746	0.715
weekly_model_saisonal	0.263	-30301.889	30397.341	-0.262	36002.681	0.172
monthly_model_sarimax	0.271	-30524.872	30524.872	-0.271	33386.010	0.863
monthly_model_arima	0.277	-31116.373	31116.373	-0.277	33811.776	0.848
weekly_model_sarimax	0.283	-32325.793	32325.793	-0.283	36474.761	0.763
monthly_model_exog	0.418	-44496.038	44496.038	-0.418	46270.271	0.732
monthly_model_saisonal	0.462	-48874.795	48874.795	-0.462	50782.078	0.737

New York organic avocado volume- Best Model - Residual Evaluation and plotting of forecasting values

Best Model is a model with exogeneous variable (price) done with weekly time series

Specification of the Best Model (2,0,2,0,0,1)52

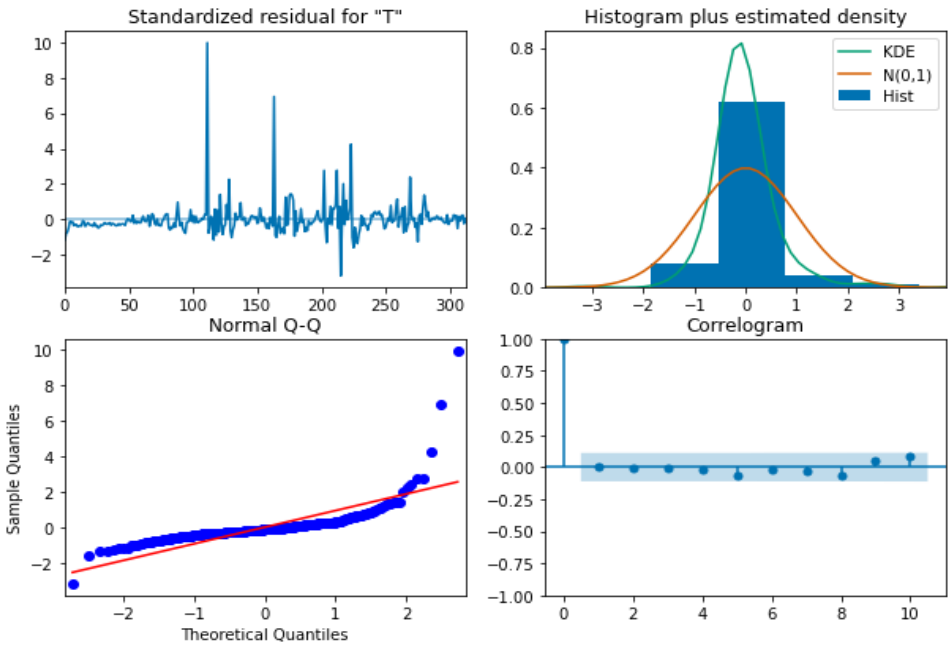


Fig. plot diagnostic result.

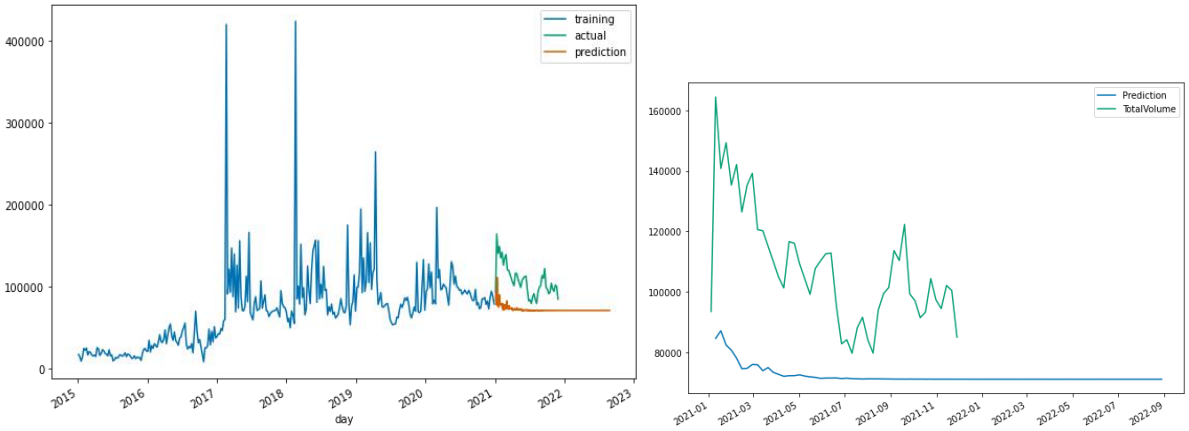


Fig. plotting of train, test, predicted and forecasted value of organic avocado volume

Best model with monthly time series data
Model specification is (2,0,0)

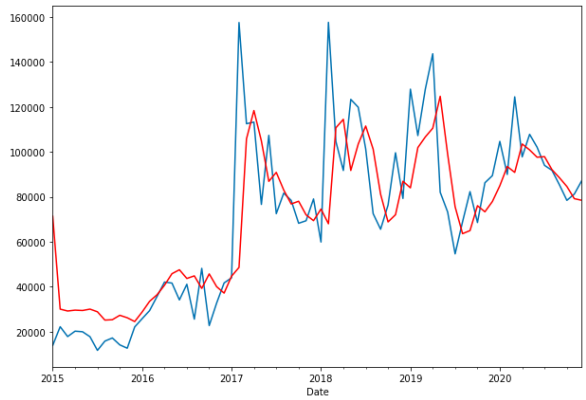


fig. plotting of actual (blue) and predicted values(red)

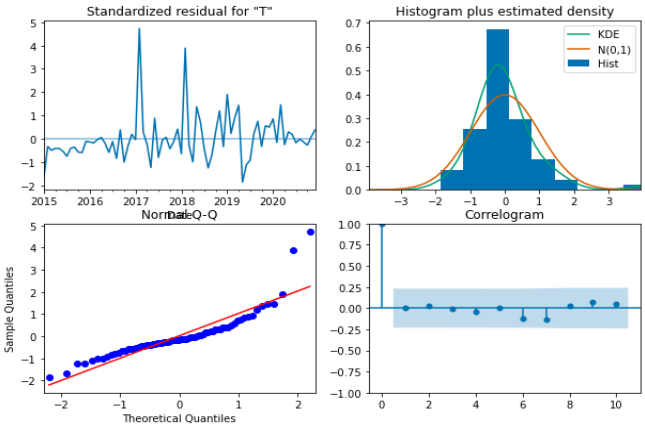
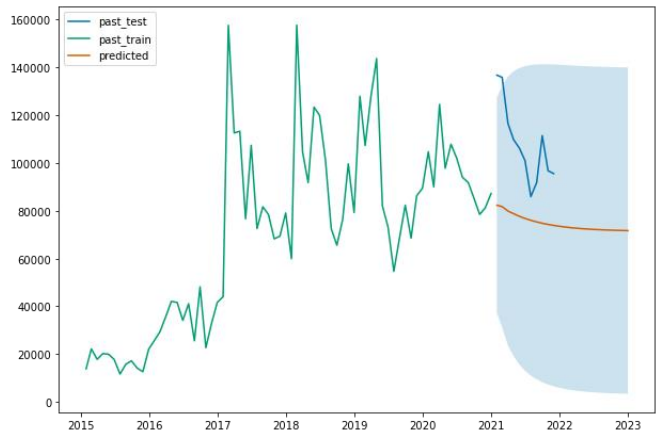


Fig. plot diagnostic result.



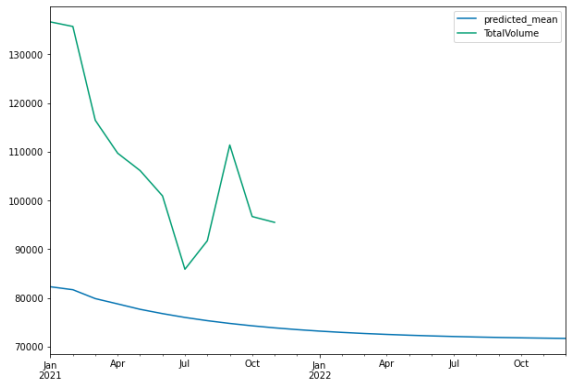


Fig. plotting of train, test, predicted and forecasted value of organic avocado volume

Table. New York organic avocado volume- Best Model – values of forecasting volume

	lower TotalVolume	upper TotalVolume	pred_mean_TotalVolume	TotalVolume	Diff (pred – actual)
2021-01-31 00:00:00	37173.936	127387.485	82280.710	136671.818	-54391.108
2021-02-28 00:00:00	31294.196	132050.375	81672.286	135737.805	-54065.519
2021-03-31 00:00:00	23582.170	136117.250	79849.710	116469.463	-36619.752
2021-04-30 00:00:00	19244.144	138267.204	78755.674	109706.755	-30951.081
2021-05-31 00:00:00	15656.850	139642.682	77649.766	106145.416	-28495.650
2021-06-30 00:00:00	13071.340	140453.673	76762.507	100941.860	-24179.353
2021-07-31 00:00:00	11051.229	140909.448	75980.339	85873.988	-9893.649
2021-08-31 00:00:00	9500.060	141135.606	75317.833	91739.626	-16421.793
2021-09-30 00:00:00	8282.565	141211.907	74747.236	111398.993	-36651.756
2021-10-31 00:00:00	7323.947	141194.523	74259.235	96699.134	-22439.899
2021-11-30 00:00:00	6561.627	141119.661	73840.644	95504.678	-21664.034
2021-12-31 00:00:00	5951.717	141012.349	73482.033		
2022-01-31 00:00:00	5460.337	140888.962	73174.649		
2022-02-28 00:00:00	5062.099	140760.365	72911.232		
2022-03-31 00:00:00	4737.484	140633.459	72685.471		
2022-04-30 00:00:00	4471.490	140512.495	72491.992		
2022-05-31 00:00:00	4252.461	140399.891	72326.176		
2022-06-30 00:00:00	4071.296	140296.840	72184.068		
2022-07-31 00:00:00	3920.833	140203.723	72062.278		
2022-08-31 00:00:00	3795.405	140120.398	71957.902		
2022-09-30 00:00:00	3690.494	140046.404	71868.449		
2022-10-31 00:00:00	3602.480	139981.093	71791.786		
2022-11-30 00:00:00	3528.442	139923.727	71726.085		
2022-12-31 00:00:00	3466.012	139873.542	71669.777		

Conclusion

Exploratory analysis of avocados time series data showed a clear distinction between organic and conventional avocados. In fact, the price of the organic avocado is higher than that of the conventional avocado and the total volume of organic avocados sold is less than that of conventional avocados.

New York is among the top 5 regions with high price in organic and conventional avocado (occupied 4th or the 5th rang), but also high total volume of avocado sold. Across scale and most cities, when total volume increases, price decreases and when price increases, total volume decreases. However, the price of both types decreases in 2018 and decreased considerably from 2019 to 2021. While the volume increased significantly from 2018 to 2021. This observation can be explained with the effect of Covid 19 which starts in 2019. Moreover, price and volume are negatively correlated, and correlation is around 0.58 in New York city.

8 models have been developed for each type of avocado (conventional price, conventional volume, organic price, organic volume), the specifications of the best forecasting models were:

- The best forecasting model of conventional avocado price is (1,0,0,0,0,1),52
- The best forecasting model of conventional avocado volume is (8,0,0)
- The best forecasting model of organic avocado price is (1,0,0,1,0,0),52
- The best forecasting model of organic avocado volume is (2,0,0)

Our models did good. The predicted prices of avocados (conventional and organic) are close to real values. However, forecasting volume was not good. There is a huge difference between real and predicted values.

Before covid 19, the price and volume of avocado was constantly increasing. the marketing of avocado was affected between 2019 and 2021. With time series analysis, this had a considerable impact on forecasting the future value of 2022. If considering that everything is back to normal (less impact of covid 19), we can suggest to the company to revise the prices of avocado considering the local trend, and the values of the upper limit of the price and volume. How can we improve the current model? maybe add exogenous variables indicating if it is a covid period or not, indicating the population, the percentage of people affected by covid for each observation.

Sources

California Avocado Commission 2020 (CAC).
Agricultural Marketing Resource Center 2020 (AgMRC).
National Agricultural Statistic Service (NASS), 2020.
Statista, 2020.
Economic Research Service (ERS), 2020.