CLASS ASSIGNMENT TWO

Insights into Avocado Market Dynamics: A Time Series Analysis and Modeling Approach

Module Title	Applied Financial Analysis
Module Code	B9FT106
Module Lecturer	Lynn Monaghan
Group members	Thi Thuy Trang Cao – 20008109
	Gauri Shingane – 20018204
Submission Date	24 April 2024
Word count	1000 words

1. Introduction

Avocados have become most sought-after for their nutritional values and culinary uses. Hence, it becomes important for suppliers, distributors, and sellers to be able to predict avocado sales. The sales are mainly influenced by seasonal changes, economic factors, and consumer preferences. This looks like a typical time series problem, as we are going to use machine learning algorithms to analyse the historical sales data to predict future sales. We are using prediction analysis that could help optimise inventory, enhance pricing and improve profitability for the Avocado sellers.

2. Data Understanding

The dataset is sourced from Kaggle.com and contains detailed information about Avocado sales, such as Average Price, Total Volume, Total Bags, etc. Certain features need improvement, like the data column can be split into year, month, and day for better analysis of when does the sale of Avocados improve or decline. We have a column named 'type' that shows the different types of avocados based on whether they are organic or conventional. A better view of all features is shown in the table below.

Table 1
The variables in each category

Variables	Non-null	Dtype
1. Date	Non-null	Object
2. AveragePrice	Non-null	float64
3. Total Volume	Non-null	Float64
4. 4046	Non-null	Float64
5. 4225	Non-null	Float64
6. 4770	Non-null	Float64
7. Total Bags	Non-null	Float64
8. Small Bags	Non-null	Float64
9. Large Bags	Non-null	Float64
10. Xlarge Bags	Non-null	Float64
11. Type	Non-null	Object
12. Year	Non-null	In64
13. region	Non-null	Object

3. Components

3.1 Trend

As shown in figure 1, we can observe there has been a consistent pattern throughout time where the organic avocados are more expensive than the conventional avocados.

That could be the primary reason for the dominance of conventional avocados in market as they are cheaper. The volume of conventional avocados sold each have been much higher than the organic ones. Even if the volume of the conventional avocados sold has been more, from figure 1, we can see that there has been a consistent increase in the sales of organic avocados.

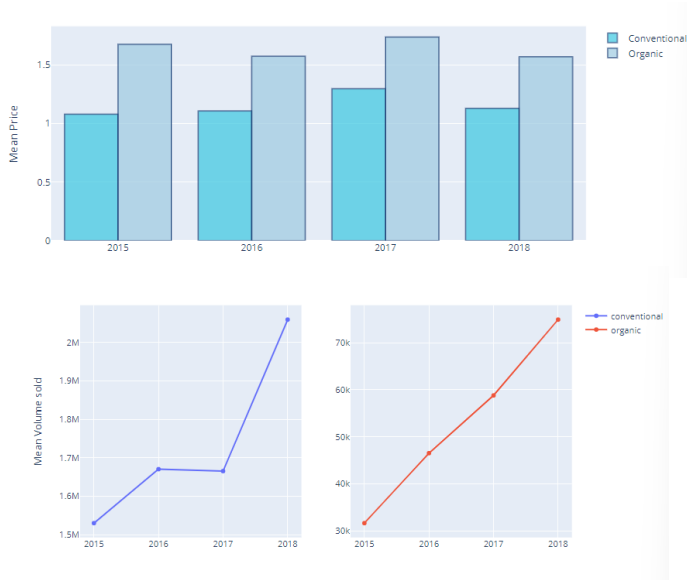


Fig 1: Comparing organic and conventional avocado prices

3.2 Seasonal Patterns

In this section, we will focus on constant patterns that occur frequently from year to year and from month to month in both conventional and organic avocados. As shown in Figure 2, we can observe the seasonal fluctuations.

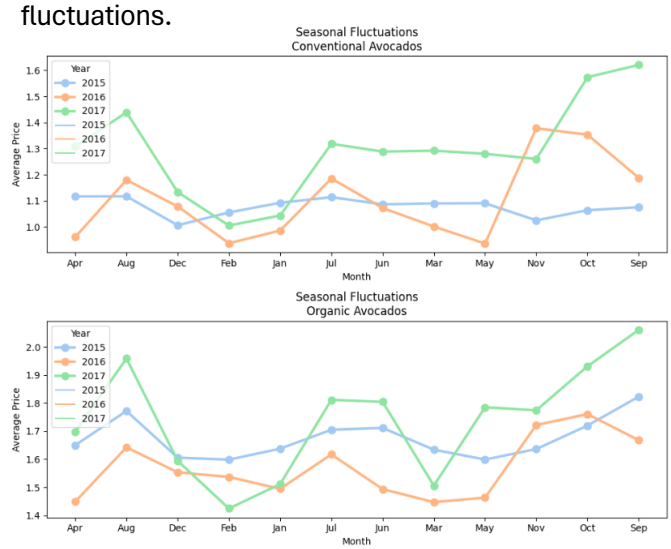


Fig 2: Seasonal Fluctuations

- Distributions per year: It looks like most of the prices in 2015 were $1.00 for conventional avocados. For 2016 and 2017, the prices' density was slightly higher.
- Price peaks per Month: It looks like most price peaks occur for both conventional and organic avocados between the months of September and October.

- Major drop in prices at the end of the year: Interestingly enough, we see that there is a major drop in the price of avocados at the end of the year.
- Standard deviation as a measure of volatility: Standard deviation is just the square root of a variance. We can see that during the year 2017, the avocado market experienced the highest volatility for both conventional and organic avocados.
- Upward Trend: The most significant upward trend happens between June - September for organic avocados and June - October for conventional avocado types. Both types of avocados have a similar trend.
- Most expensive months to buy avocados: Like in the previous analysis, through this visualization, we confirm that the months with the highest average price are September and October however, August is pretty expensive as well. For each month, we have three peaks. That represents the highest peak in each month of each year. Since we are using the prices of 2015, 2016, and 2017, we can see how the prices of each year behave in all twelve months. This will help us see if there are any major seasonal patterns.

3.3 Residual variation

The scatter plot in Figure 3 reveals heteroscedasticity, indicating that as the total avocado sales increase, the residual variance also increases, forming a cone shape. Notably, organic avocados' distribution is more concentrated than conventional avocados, which exhibit a wider spread. This suggests potential differences in sales dynamics between organic and conventional avocados. Understanding these variations is essential for market analysis and decision-making.

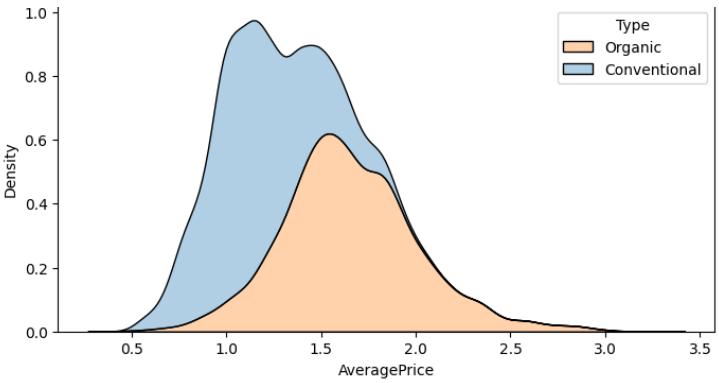


Fig 3: Avocado Price by type

4. Stationarity Analysis

To assess stationarity, we employed the Augmented Dickey-Fuller Test:

- Null Hypothesis: Assumes the time series is non-stationary.
- Alternate Hypothesis: Rejecting the null hypothesis indicates stationary behavior.

For the null hypothesis to be rejected, two conditions must be met:

- Critical Value (5%) > Test Statistic
- p-value < 0.05

The analysis of the Augmented Dickey-Fuller Test outputs revealed (Figure 4):

- The rolling mean is close to 0 with minor variations.
- The rolling standard deviation fluctuates around 0.05.
- Critical Value (5%): -2.88 > Test Statistic: -13.82, indicating stationarity with 99% confidence.
- The p-value (0.00) is less than 0.05, further supporting stationarity.

Based on these results, we reject the Null Hypothesis and accept the Alternate Hypothesis, confirming the stationary nature of the time series.

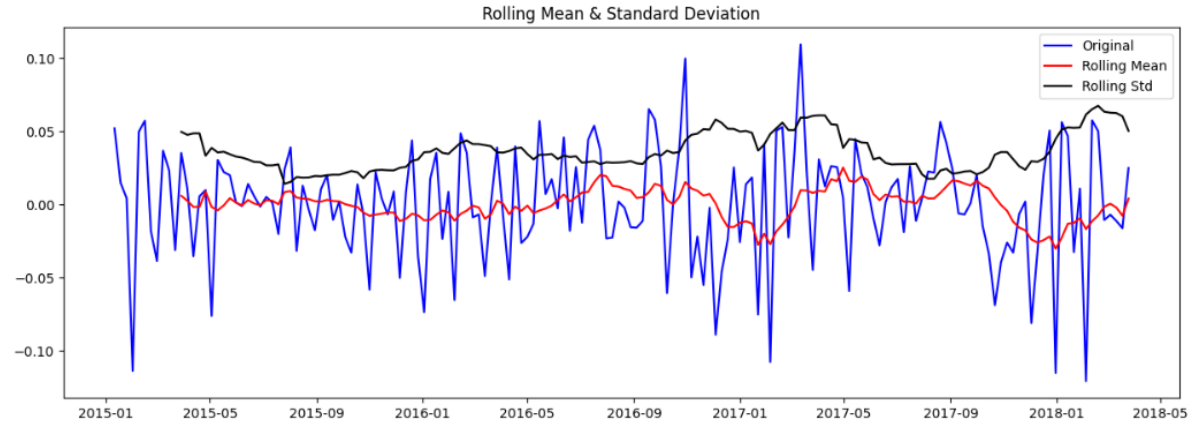
5. Model Selection

5.1 Auto regressive integrated moving average (ARIMA)

The ARIMA model, a prominent method in time-series prediction, merges autoregressive (AR) and moving average (MA) processes to forecast future values (Dubey et al., 2021). Defined by its parameters (p,d,q), where 'p' represents the auto-regressive term, 'q' signifies the moving average term, and 'd' denotes the differencing term, ARIMA is particularly suited for stationary time series data. The AR term involves regression against past values to predict future outcomes, while the MA term utilizes past forecast errors. Differencing is employed to attain stationarity by eliminating trends and seasonality. By integrating these components, ARIMA models capture intricate temporal patterns, making them valuable tools in diverse domains such as finance and meteorology. Their ability to provide accurate forecasts aids decision-making processes, enabling anticipation of future trends with confidence.

5.2 Seasonal ARIMA Mode (SARIMA)

SARIMA model, an extension of the ARIMA framework, is tailored to handle time series data with seasonal patterns. In addition to the parameters (p,d,q) of the ARIMA model, SARIMA introduces three new superparameters—P, D, and Q—alongside an extra seasonal periodic parameter, denoted as 's'. These components, P (seasonal autoregression), D (seasonal integration), Q (seasonal moving average), and s (seasonal period length), augment the ARIMA model's capability to capture and forecast seasonal variations in the data. SARIMA models can effectively model and predict complex seasonal patterns, making them invaluable tools in fields such as economics, retail, where seasonality plays a significant role in data dynamics (Adeyeye et al., 2023)




```
Results of Dickey-Fuller Test:
Test Statistic                -1.381774e+01
p-value                        8.010524e-26
#Lags Used                     0.000000e+00
Number of Observations Used    1.670000e+02
Critical Value (1%)           -3.470126e+00
Critical Value (5%)           -2.879008e+00
Critical Value (10%)          -2.576083e+00
dtype: float64
```

Fig 4: Augmented Dickey-Fuller Test Result

6. Results and Discussion

6.1 ARIMA Model

The ARIMA model is constructed based on the selection criteria derived from the partial autocorrelation function (PACF) graph analysis, where the 1st lag appears to be the most significant. With p=1, d=1, and q=2, the model is fitted to the Log_AveragePrice data, which initially lacks differencing and stationarity. The series is transformed to achieve stationarity through the differencing process facilitated by the ARIMA model's parameter 'd'. Figure 6 shows that the series is not stationary. Figure 7 represents ACF and PACF for a first-order difference and a first-seasonal difference time series. The model fitting involves comparing the fitted values with the original AveragePrice data, utilizing the Augmented Dickey-Fuller (ADF) test to confirm stationarity. The resulting fitted values, obtained through cumulative summation and exponentiation, represent the original series supplied by the ARIMA model, facilitating accurate forecasting and prediction.

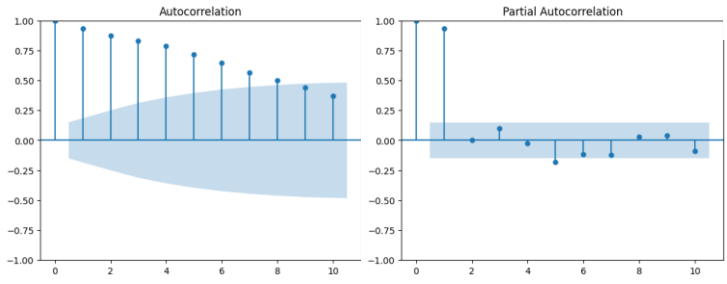


Fig 6: Autocorrelation and partial autocorrelation plots: The original time series - ARIMA model

For the above trained ARIMA model, values generated by the forecast_function and predict function are found to be nearly identical, differing only by a few decimal points. However, despite its accuracy in capturing the overall trend and short-term fluctuations,

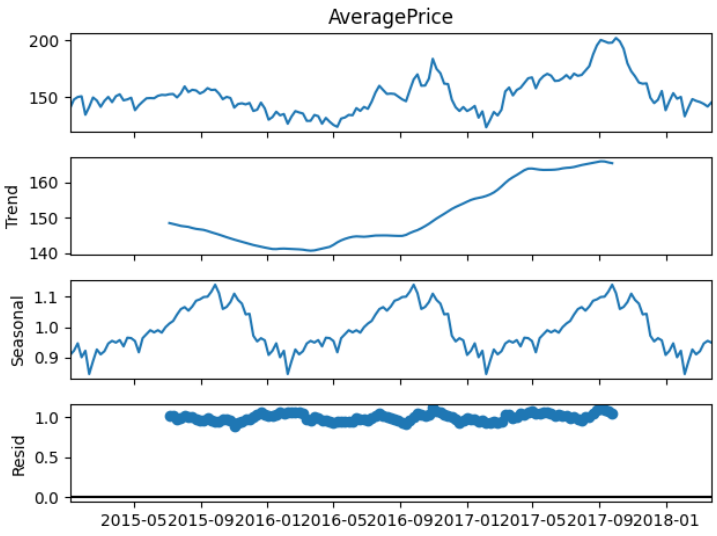


Fig 5: Seasonal decomposition of Avocado price

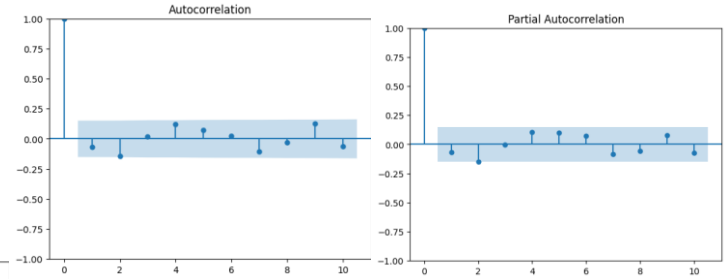


Fig 7: Autocorrelation and partial autocorrelation plots: Log_AveragePrice data- ARIMA model

the model evidently struggled to capture the seasonal patterns inherent in the data. This limitation suggests the need for further refinement or alternative modelling techniques to improve the model's predictive performance, particularly in forecasting seasonal variations accurately (Figure 8).

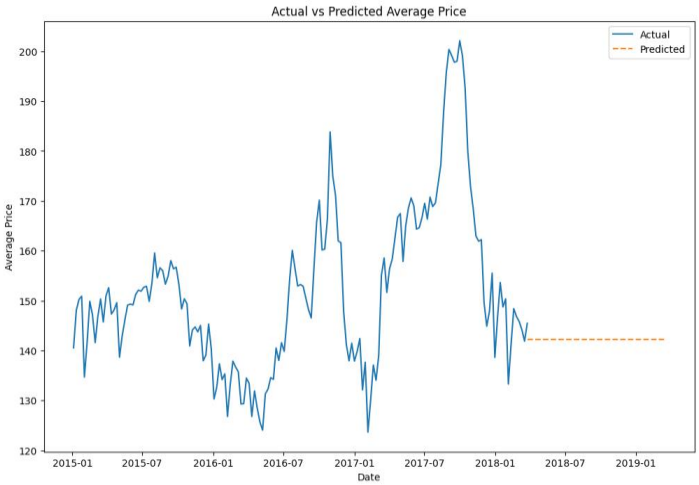


Fig 8: Predict average price – ARIMA model

6.2 SARIMA Model

The SARIMA model offers a comprehensive approach by considering both seasonal and non-seasonal components of the data, resulting in accurate forecasting and prediction capabilities. As depicted in Figure 9, the seasonal decomposition of the training dataset using SARIMA reveals distinct patterns of seasonality, trend, and residual components.

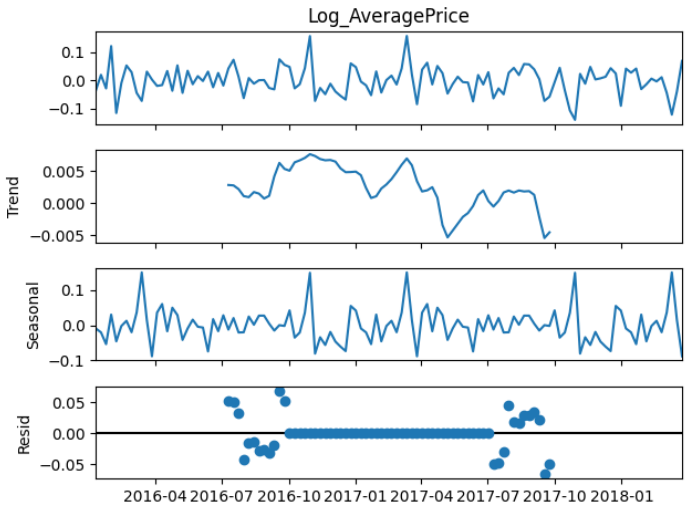


Fig 9: Seasonal decomposition of Log Average Price

These components serve as crucial inputs for the forecasting process, enhancing the accuracy of future predictions. Analysis of the ACF and PACF further confirms the presence of clear seasonal patterns. Figure 10 illustrates ACF and PACF for a first-order difference and a first seasonal difference time series, aiding in determining the appropriate differencing terms for achieving stationarity. The Augmented Dickey-Fuller Test outputs affirm stationarity, with the rolling mean and standard deviation exhibiting minimal variations close to zero. The test statistic falls below the critical value (5%). It yields a p-value of 0.00, allowing rejection of the null hypothesis in favor of the alternate hypothesis, thus confirming the stationary nature of the time series.

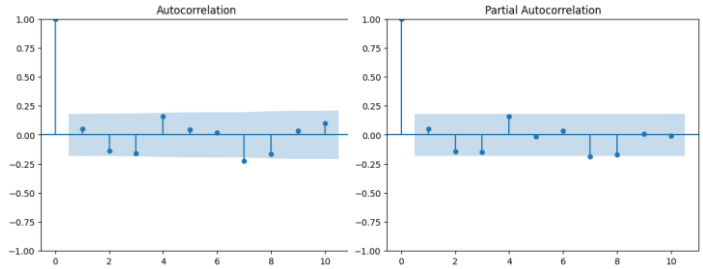


Fig 10: Autocorrelation and partial autocorrelation plots SARIMA model

For the above trained SARIMA model, values generated by forecast_function and predict function are identical or differ by a few decimal points. SARIMA was much better at capturing seasonal patterns of data (Figure 11)

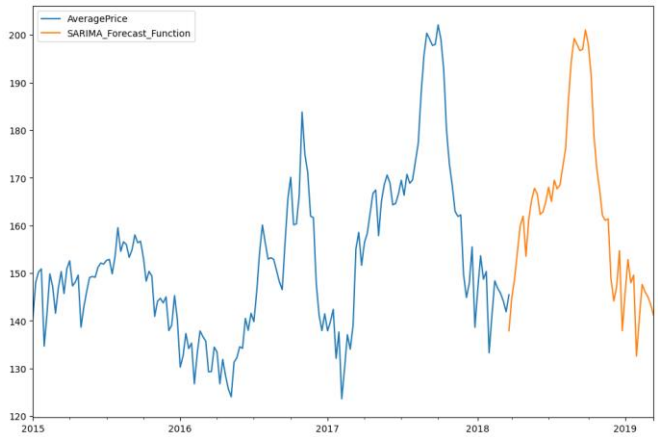


Fig 11: Predict average price - SARIMA model

7. Conclusion

The comprehensive analysis of avocado price dynamics offers valuable insights into the behavior of the avocado market over time. Through examination of distributions per year, price peaks per month, volatility measured by standard deviation, and upward trends, a holistic understanding of the factors influencing avocado prices is gained. These insights empower market participants to anticipate trends, mitigate risks, and capitalize on opportunities in the dynamic avocado market.

Moreover, the analysis conducted on the time series data using ARIMA and SARIMA models has provided further valuable insights into the underlying patterns and dynamics of the dataset. While both models demonstrated proficiency in capturing the overall trend and short-term fluctuations, SARIMA exhibited superior performance in capturing seasonal patterns. Incorporating seasonal parameters in the SARIMA model enabled it to more adeptly model and forecast time series data characterized by seasonal variations. The findings from this analysis offer valuable information for decision-making and planning in fields such as finance, economics, and retail, where accurate forecasting of time series data is crucial. Future research could focus on refining the models or further exploring ensemble methods to enhance predictive accuracy.

REFERENCES

1. Dubey, A., Kumar, A., Garcia-Diaz, V., Sharma, A. & Kanhaiya, K. Study and analysis of SARIMA and LSTM in forecasting time series data. Sustainable Energy Technologies And Assessments. 47 pp. 101474 (2021)
2. Adeyeye, J. & Nkemnole, E. Predicting Malaria Incident Using Hybrid SARIMA-LSTMModel.International Journal Of Mathematical Sciences And Optimization: Theory And Applications. 9 (2023)
3. Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. Holden-Day.
4. Riaz, M., Hussain Sial, M., Sharif, S., Mehmood, Q. & Others Epidemiological Forecasting Models Using ARIMA, SARIMA, and Holt–Winter Multiplicative Approach for Pakistan. Journal Of Environmental And Public Health. 2023 (2023)
5. Aryal, S., Nadarajah, D., Rupasinghe, P., Jayawardena, C. &Kasthurirathna, D. Comparative analysis of deep learning models for multi-step prediction of financial time series. Journal Of Computer Science. 16, 1401-1416 (2020)