FORECASTING AVOCADO SALES

Submitted by: Akanksha Arun, Juhil Ahir, Shubhangi Sharma, Tanvi Vijay

Executive Summary:

This detailed report addresses the critical need for precise sales volume forecasting in a small California-based avocado production business. Covering the period from 2015 to 2018, the project specifically focuses on conventional and organic avocados, tailoring its insights to the unique challenges faced by a small-scale enterprise. The primary objective is to provide actionable recommendations for strategic decision-making, considering the intrinsic seasonality of avocado sales and the potential influence of pricing dynamics. Within the specific context of a small avocado production business, the introduction outlines the necessity for accurate sales predictions. Motivated by the operational realities of a smaller scale, the project seeks to empower the business with insights to optimize production, enhance supply chain efficiency, and formulate effective pricing strategies. The project employs practical data exploration techniques such as time series decomposition, autocorrelation function (ACF) analysis, and visualizations tailored to unravel patterns in historical sales data. Training and testing datasets are meticulously crafted to lay the foundation for subsequent model development. Techniques range from fundamental linear models and naive forecasting to advanced ARIMA models, with a particular focus on incorporating the intricate relationship between avocado prices and sales volumes.

Data Source: Kaggle

Avocado Price Prediction | Kaggle

Dataset contains Avocado Sales for US Markets. A subset of this dataset has been used for the analysis of Avocado sales in California.

Exploratory Data Analysis:

Time Series Visualization:

Conventional Avocados: The time series plot for conventional avocados reveals an overall increasing trend in sales volume. Seasonality is evident, with a notable surge in sales during the second week of February each year. Outliers are observed, indicating potential areas of interest for further investigation. (Figure 1 in Appendix)

Organic Avocados: The time series plot for organic avocados also displays an increasing trend, with a similar seasonal pattern to conventional avocados. The second week of February again stands out as a peak selling period. This visual consistency suggests a shared market behavior for both avocado types. (Figure 2 in Appendix)

Decomposition, Seasonality and Trend:

Conventional Avocados: Decomposition of the time series shows clear seasonality, emphasizing the yearly sales pattern. The autocorrelation function (ACF) further supports this observation, with a pronounced peak at lag 52, indicating a strong yearly seasonality component.

However, this is not the case with trend. The analysis revealed that there is no clear trend in avocado demand from 2015 to 2018. In other words, a consistent upward or downward movement in demand over this period was not observed.

Organic Avocados: Similar to conventional avocados, decomposition and ACF analysis for organic avocados confirm a significant yearly seasonality component. The consistent seasonality patterns in both types suggest a shared market influence or external factors affecting avocado sales in California. Organic Avocados, like Conventional Avocados Exhibit no trend. (Figure 3, 4, 5,6,7 and 8 in Appendix)

Boxplot Analysis:

Conventional Avocados: The boxplot highlights the presence of outliers in the sales data. The median conventional avocado sales is about ~6 Million. (Figure 9 in Appendix)

Organic Avocados: The boxplot does not indicate the presence of outliers in the sales data. The median organic avocado sales is about ~150,000. (Figure 10 in Appendix)

Forecasting Models:

In the process of forecasting avocado sales in California, several models were evaluated to capture the intricate dynamics observed in the dataset. Time series analysis involved both conventional and organic avocados, considering factors such as seasonality, trend, and external influences. Linear models were employed, including simple and log-transformed time series linear models, providing a baseline understanding of the data. Naive and seasonal naive models were explored for their simplicity and effectiveness. Moving average models were tested, acknowledging potential limitations in handling trends and seasonality. Furthermore, ARIMA models, tailored to accommodate the temporal patterns, were deployed. The evaluation extended to incorporating the effect of avocado prices into the models, recognizing the potential impact of pricing dynamics on sales. The final chosen model integrated the strengths of ARIMA with the additional influence of price, presenting a robust approach to forecasting avocado sales in California. The selection process involved thorough consideration of each model's ability to capture the observed seasonality, trend, and respond to external factors contributing to the unique sales patterns of both conventional and organic avocados.

Model 1: Simple Linear Model

Conventional Avocados:

The linear model for conventional avocado sales volume in California incorporates both trend and seasonality components. The model suggests that sales are influenced by a general increasing trend and specific seasonal patterns for each month. However, the residuals from the model exhibit significant serial correlation, implying that some temporal patterns are not adequately captured. The model's accuracy metrics on the training set indicate generally low error rates, with the Mean Absolute Percentage Error (MAPE) at 6.67%. (Figure 11 in Appendix)

Organic Avocados:

The linear model for organic avocado sales volume in California incorporates both trend and seasonality components. The coefficients indicate the presence of a general increasing trend and specific seasonal patterns for each month, with varying degrees of influence. However, the model reveals significant serial correlation in the residuals, suggesting that some temporal patterns remain unaccounted for. Despite this, the model exhibits reasonably low error rates on the training set, with a Mean Absolute Percentage Error (MAPE) of 13.08%. (Figure 12 in Appendix)

Model 2: Log-Transformed Linear Model

Conventional Avocados:

The log-transformed linear model for conventional avocado sales volume in California reveals a positive intercept and coefficients for both trend and seasonality components. The model successfully captures the underlying patterns in the log-transformed data, with season-specific coefficients indicating the influence of each month on sales. However, the residuals exhibit significant serial correlation up to lag 24, indicating the presence of unaccounted temporal patterns. Despite this, the model shows high accuracy on the training set, a Mean Absolute Percentage Error (MAPE) of 0.71%. (Figure 13 in Appendix)

Organic Avocados:

The logarithmically transformed linear model for organic avocado sales volume in California exhibits a positive intercept and coefficients for both trend and seasonality components. The model captures the underlying patterns in the log-transformed data, as reflected in the season-specific coefficients indicating the influence of each month on sales. However, the residuals show significant serial correlation up to lag 56, indicating potential unaccounted temporal patterns. Despite this, the model demonstrates high accuracy on the training set, with a Mean Absolute Percentage Error (MAPE) of 1.08%. (Figure 14 in Appendix)

Model 3: Naive and Seasonal Naive Models

Conventional Avocados:

The Naive Model and Seasonal Naive Model were evaluated for forecasting avocado sales in California. The Naive Model simply forecasts future values based on the last observed value, resulting in a point forecast of 7,329,081 units for each future period. The model exhibits a Mean Absolute Percentage Error (MAPE) of 13.45%, indicating its limitations in capturing the underlying patterns and fluctuations in avocado sales. (Figure 15 in Appendix)

On the other hand, the Seasonal Naive Model takes into account the seasonality component by forecasting future values based on the corresponding season's last observed value. The model provides more nuanced forecasts, considering the seasonal variations in avocado sales. The point forecasts range from 5,597,391 to 9,615,418 units across different periods. The Seasonal Naive

Model demonstrates a MAPE of 14.38%, suggesting improved performance compared to the Naive Model. However, it's crucial to note that both models might be limited in capturing more complex trends and external factors affecting avocado sales. (Figure 16 in Appendix)

Organic Avocados:

The Naive Model and Seasonal Naive Model were applied to forecast avocado sales in the organic category in California. The Naive Model predicts a constant value for future periods, with a point forecast of 164,146.7 units for each upcoming period. The model exhibits a Mean Absolute Percentage Error (MAPE) of 15.35%. (Figure 17 in Appendix)

The Seasonal Naive Model considers the seasonality component and forecasts future values based on the corresponding season's last observed value. The point forecasts range from 135,684.6 to 254,316.6 units across different periods, capturing the seasonal fluctuations in organic avocado sales. The model shows a MAPE of 12.54%, indicating a better performance compared to the Naive Model. However, it's important to consider that both models may have limitations in capturing more intricate trends and external factors influencing organic avocado sales. (Figure 18 in Appendix)

Model 4: Moving Average Models

Conventional Avocados:

After applying the Naive and Seasonal Naive models, we decided to step it a little further, and tried experimenting with another model due to the previous model limitations when it comes to capturing trends and external factors that might be impacting our data. We decided to experiment with a moving-average model. This model plays a crucial role in capturing trends and patterns with sequential data. Unlike the Naive and Seasonal Naive Models, the Moving Average Model considers a broader context by smoothing out fluctuations over a specified window of time, providing a clearer representation of the underlying trends. For our conventional avocados, we obtained a MAPE of 8.76%. This model resulted in better accuracy than the Naive and similar to Seasonal Naive models.

Organic Avocados:

We repeated the same steps for our organic avocados. For our organic avocados, we were not able to achieve around the same accuracy level as for conventional avocados. Our model prediction power decreased. The MAPE value obtained from the moving average model for organic avocado was 23.76%. This moving average model doesn't perform better than previous models for this category of the dataset.

Model 5: Holt - Winter's Model

The dataset revealed evident seasonality, prompting consideration of the Holt-Winters model for time series forecasting. However, it was noted that the applicability of the Holt-Winters model is constrained to datasets with frequencies up to 24. Given that the dataset in question consisted of weekly data, equating to 52 frequencies per year, the utilization of the Holt-Winters model was deemed impractical.

Model 6: ARIMA Models

ARIMA is a model very well known in time series forecasting due to its flexibility and robustness under various circumstances. It is designed to capture both short-term fluctuations and long-term trends in time series data. Its autoregressive (AR) component accounts for past values' influence, while the integrated (I) component incorporates differencing to achieve stationarity, and the moving average (MA) component helps filter out noise.

ARIMA models can adapt to a wide range of time series patterns, including linear and non-linear trends, seasonality, and cyclic behavior. The flexibility to adjust the model's parameters allows it to accommodate different data characteristics.

For our model, we checked at the auto-correlation factor to make sure how to proceed with building the ARIMA. As observed from figure 19 and 20, we generated the ACF and PACF plots to make sure that correlation was tailing off or spiking. Implementing this technique really helps identify what configuration to apply in the ARIMA model. As mentioned earlier, our avocado data showcased a stable pattern, but it does showcased seasonality. Seasonality could also be observed in these ACF and PACF plots. We tried multiple configurations for our ARIMA model from AR(1) all the way to our most promising model ARIMA(1,0,0) x (1,1,0). Our best ARIMA model is composed of the following:

Non-Seasonal Component (ARIMA): It has one autoregressive term (p=1) and no moving average term (q=0) as ACF plot tails off while the PACF plot cuts off. No differencing (d=0) has been applied as the dataset lacks trend.

Seasonal Component: It has one seasonal autoregressive term (P=1) as we can see seasonality in the ACF graph, one seasonal differencing (D=1), and no seasonal moving average term (Q=0).

Conventional Avocados:

As mentioned above our best ARIMA configuration was ARIMA(1,0,0) x (1,1,0). This configuration takes care of the seasonality observed in our dataset. After building the model, we were able to obtain a MAPE value 10.96%. This model performed better than our previous models. Still, we were expecting a significant drop in the MAPE value since ARIMA is a very powerful technique that encompasses multiple data characteristics, hence we decided to utilize an external factor that could impact the volume of the avocado and improve our model performance. In our case, we decided to see the effect price had on the volume. We were able to incorporate the price factor to our model and test the model's performance. As a result, the model's MAPE value went down from 10.96% to 6.89%. We were able to forecast the volume for our conventional avocado as shown in figure 21. Our highest forecasted value is around 11 million units in volume, and our lowest forecasted value is around 4.2 million units in volume.

Organic Avocados:

We repeated the same process for our Organic avocado. Similar to the conventional avocado model, initially we obtained a MAPE of 11.15%, then after incorporating the price factor to our model. Our model's MAPE value decreased from 11.15% to 5.94%. This MAPE value showcases how powerful and useful can ARIMA be in predicting future demand for our product. As shown in figure 22, we forecasted the volume of organic avocado for 2018 and 2019. Our highest forecasted value is around 233 thousand units in volume, and our lowest forecasted value is around 108 thousand units in volume. As expected, organic avocados' forecasted values are lower than conventional because of the price point amongst conventional and organic.

Model 7: Regression Model with External Variable

Conventional Models:

The linear trend model aims to capture the overall trajectory of avocado prices, indicating a weekly decline, though the trend coefficient lacks statistical significance. However, this model exhibits limited predictive capacity, as reflected in a low R-squared value and a relatively high

Mean Absolute Percentage Error (MAPE) of 13.99% on the test set. In contrast, the extended model incorporates the average price as an external variable, enhancing its explanatory power. With both the trend and average price coefficients proving statistically significant, model demonstrates improved accuracy on both training and test sets, boasting a higher R-squared and a lower MAPE of 6.89%. The negative coefficient of price in the model, for every unit increase in price there is a -2,145,623 unit decrease in volume. This signifies that the extended model offers a more nuanced understanding of avocado price dynamics and delivers more accurate predictions compared to the simplistic linear trend model. However, ongoing refinement and exploration of additional variables could further enhance predictive capabilities.

Organic Avocados:

The time series analysis for avocado prices involves the application of two models to predict trends and variations. The linear trend model suggests a positive trend in avocado prices but exhibits limited predictive power, evident in its low R-squared value (0.0556) and a relatively high Mean Absolute Percentage Error (MAPE) of 12.75% on the test set. In contrast, the extended model incorporates the average price as an external variable, resulting in improved accuracy with a significantly higher R-squared value of 0.4525. The coefficients for both the trend and average price are statistically significant, providing a more nuanced understanding of price dynamics. Visualization of the training time series with fitted values shows a close alignment. Evaluation on the test set confirms the enhanced predictive performance of the extended model, with a lower MAPE of 5.82%, indicating better alignment between predicted and actual prices. The negative coefficient of price in the model, for every unit increase in price there is a -100,532 unit decrease in volume. The inclusion of the average price variable proves to be a valuable enhancement in capturing and forecasting avocado price dynamics.

Summary:

After evaluating various forecasting models for avocado sales in California, we have decided to adopt the ARIMA model for our analysis. Among the considered models, ARIMA demonstrated superior performance in accounting for the seasonality observed in our data. Importantly, our dataset does not exhibit a clear trend, making ARIMA an appropriate choice as it focuses on capturing and forecasting based on the observed seasonality patterns.

ARIMA, or Autoregressive Integrated Moving Average, is well-suited for time series data characterized by seasonality and lacks a discernible trend. This model takes into account both autoregressive (AR) and moving average (MA) components, along with differencing to make the time series stationary. By leveraging these features, ARIMA excels in capturing the recurring patterns in our avocado sales data.

By selecting ARIMA, we aim to enhance the accuracy of our sales forecasts, providing valuable insights into the seasonal variations of avocado demand in California. This choice aligns with the nature of our dataset and ensures a robust approach to forecasting without the need to account for a trend component.

Furthermore, the ARIMA model exhibited remarkable accuracy in our forecasting analysis, boasting the lowest Mean Absolute Percentage Error (MAPE) compared to other models. Specifically, for Conventional Avocados, the ARIMA model achieved an impressively low MAPE of 6.89%, indicating its ability to provide forecasts with minimal percentage deviation from the actual sales data. Similarly, for Organic Avocados, the ARIMA model outperformed other models with an even lower MAPE of 5.94%.

The low MAPE values signify the effectiveness of ARIMA in producing forecasts that closely align with the observed sales patterns, reinforcing its suitability for our avocado sales data in

California. This heightened accuracy is crucial for stakeholders and decision-makers, as it ensures more reliable predictions of avocado demand, facilitating informed business planning and resource allocation. The consistent outperformance of ARIMA in both conventional and organic avocado sales further validates its selection as the preferred forecasting model for our analysis.

Insights and Recommendations:

The analysis reveals a robust correlation between avocado demand and price, indicating a crucial market dynamic that can be leveraged for strategic pricing of future demand. This inverse relationship implies that fluctuations in price significantly impact demand, highlighting the importance of understanding and forecasting market dynamics.

Key observations include the highest recorded volume of 11,213,596 units and the lowest price of $0.67. Notably, the same week in February witnesses peak sales for both conventional and organic avocados, reflecting a consistent market trend. This pattern can be strategically capitalized upon in pricing decisions, considering the historical surge in demand during this period.

The avocado harvest season contributes to an abundance of fresh avocados, influencing both price and demand. As production increases, prices tend to decrease, subsequently boosting demand. Conversely, factors such as drought in California and Mexico, worker strikes in Mexico, and a shortage of avocados at the end of 2016 resulted in a price hike, leading to a decrease in demand.

The impact of external events, such as the Super Bowl, is evident in increased demand. Conversely, the end of 2017 saw a decrease in demand due to elevated prices caused by reduced production, attributed to drought conditions in California and Mexico. This multifaceted analysis underscores the intricate interplay between various factors influencing avocado market dynamics, providing valuable insights for effective pricing and resource management strategies.

In conclusion, these insights have direct implications for our avocado farming community. Understanding demand influencers like the Super Bowl and harvest cycles enables strategic planning for production and pricing. The forecasted values provide a roadmap for anticipating market trends and optimizing yield, offering actionable intelligence for informed decision-making. We appreciate your commitment and look forward to applying these insights to foster a thriving future for our avocado farming community.

APPENDIX

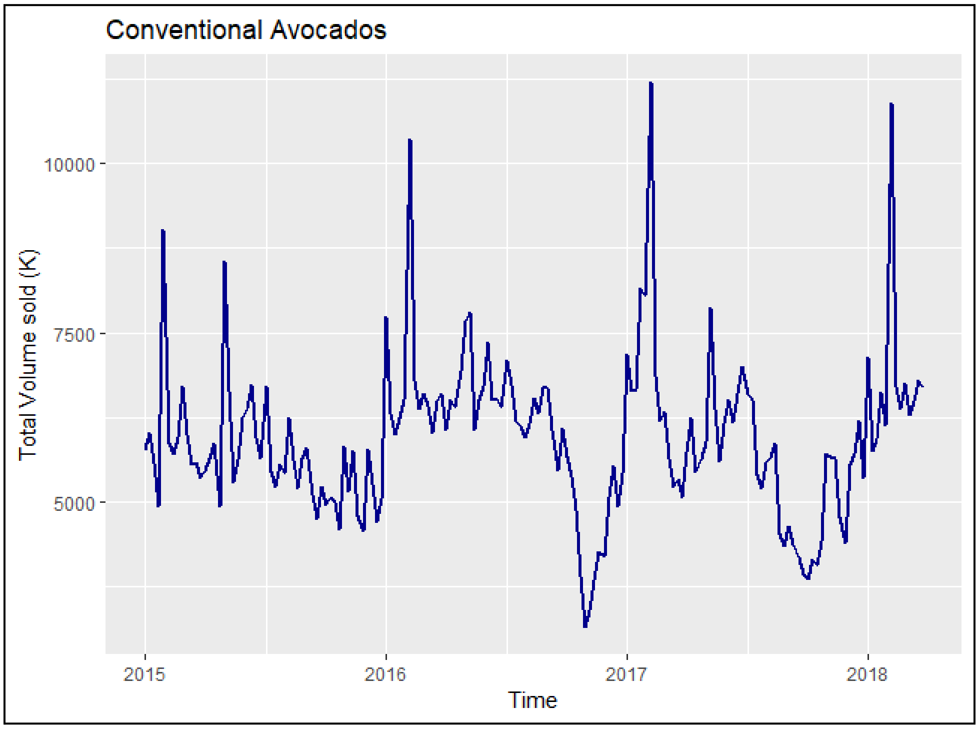

Figure 1

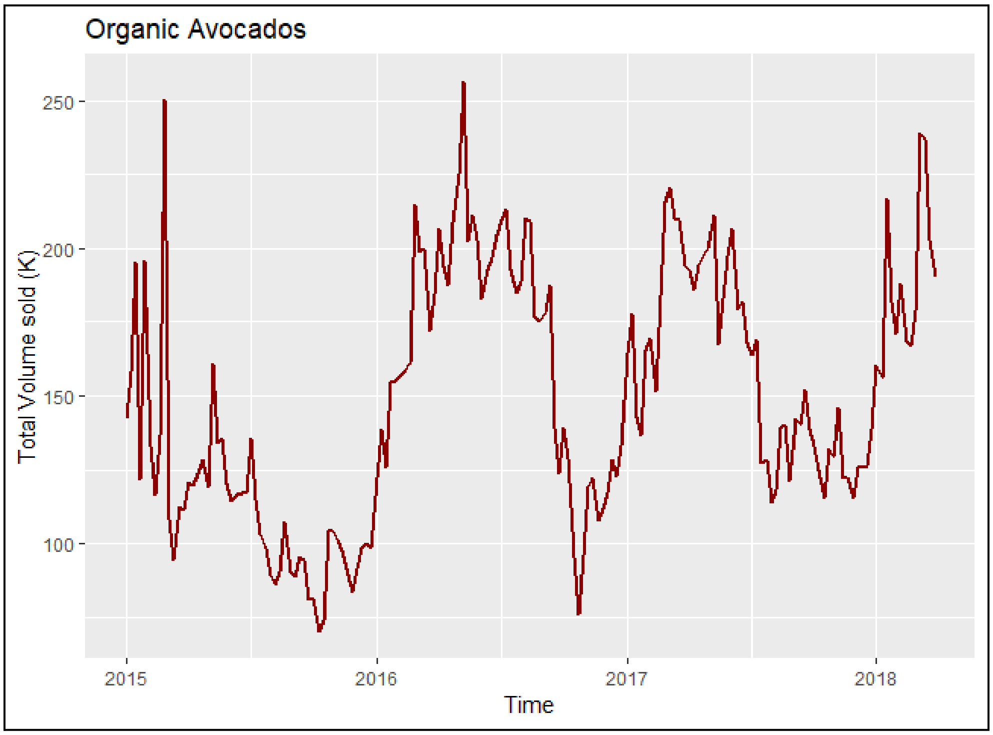

Figure 2

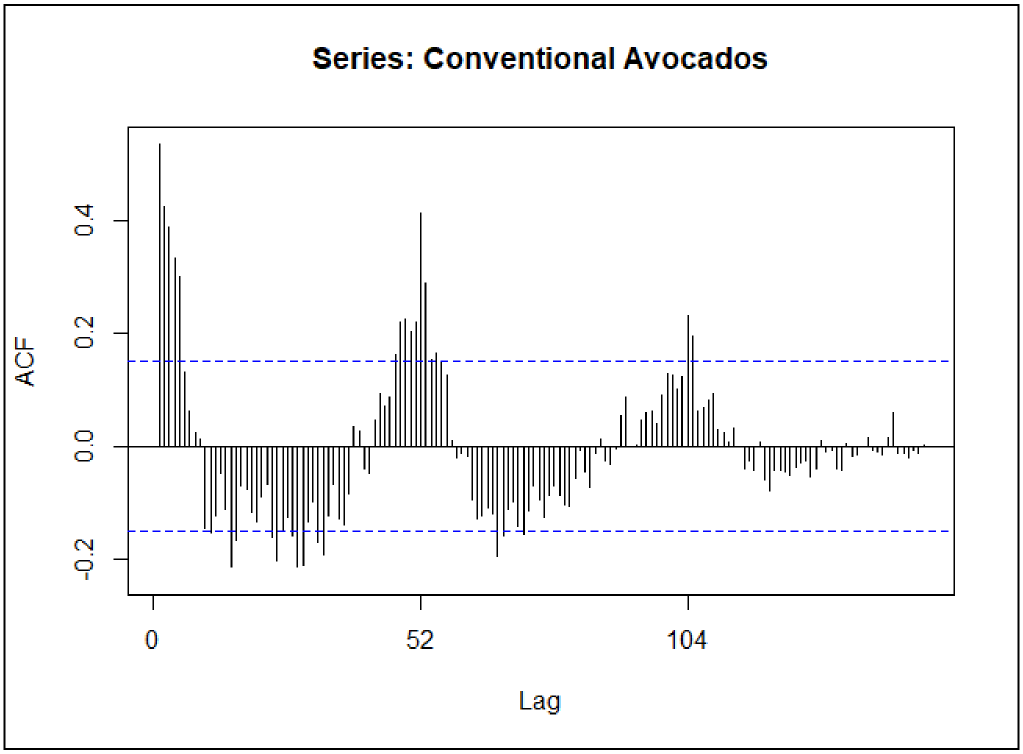

Figure 3

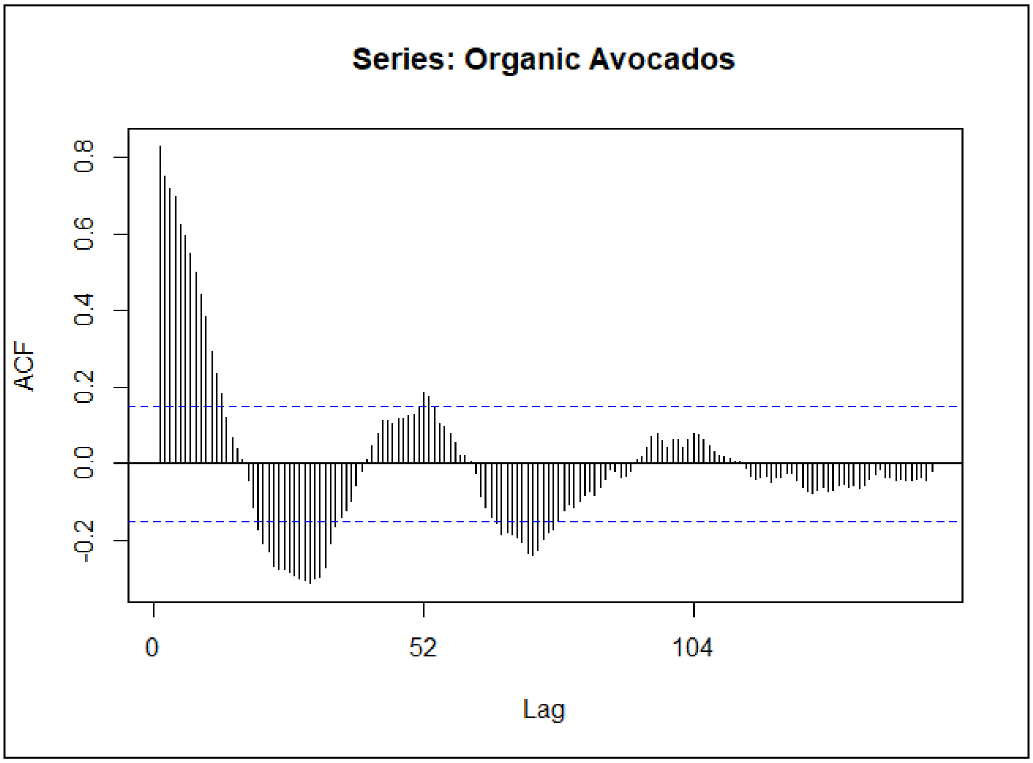

Figure 4

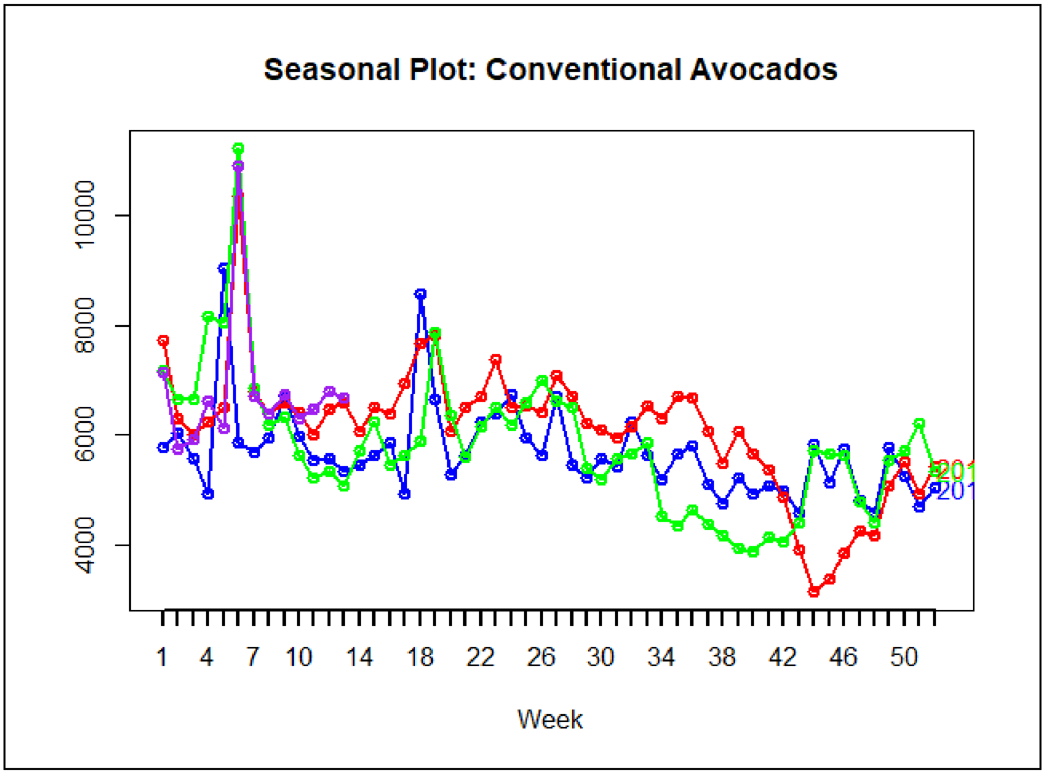

Figure 5

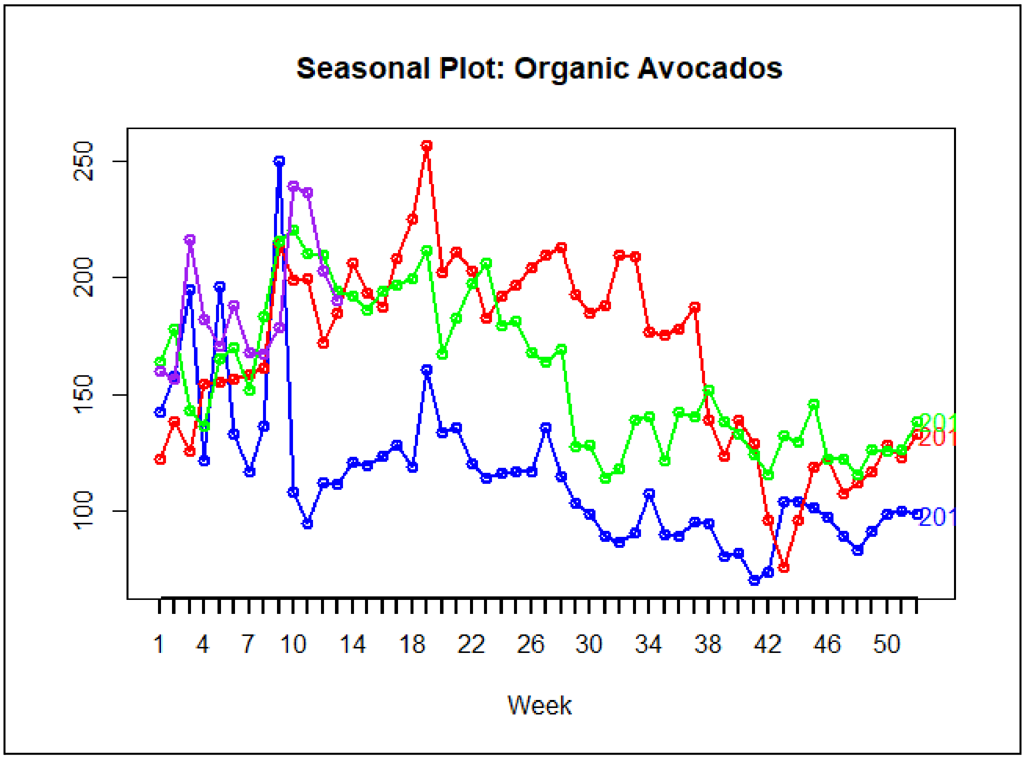

Figure 6

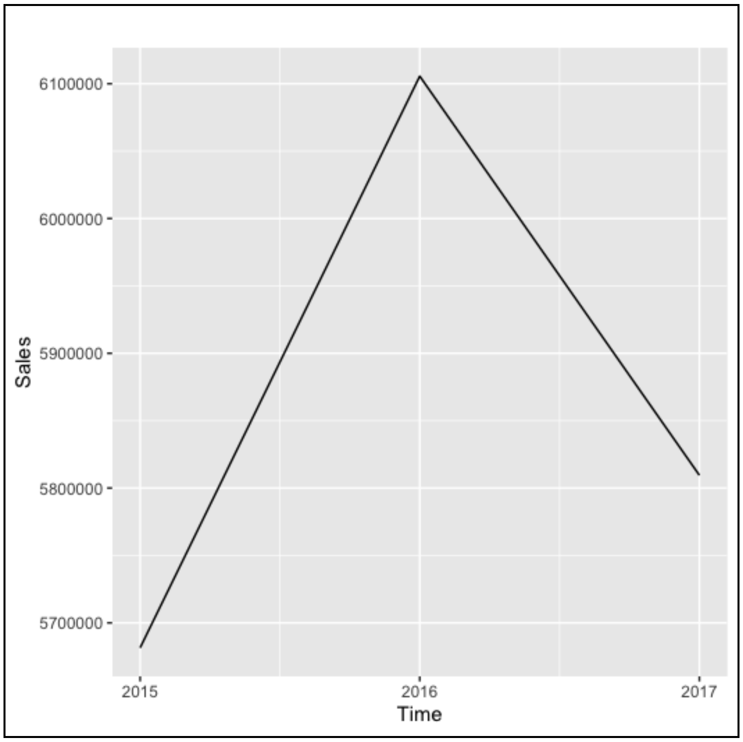

Figure 7

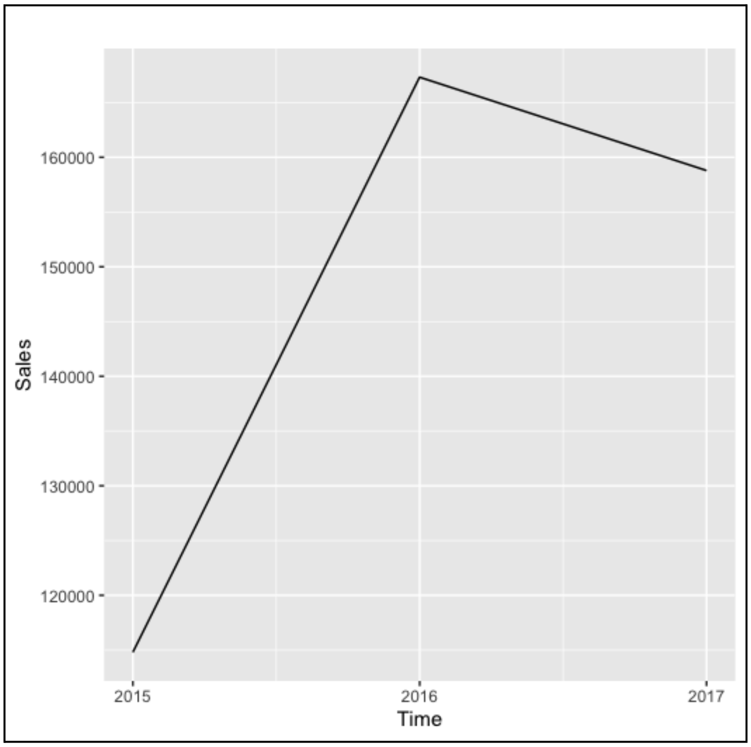

Figure 8

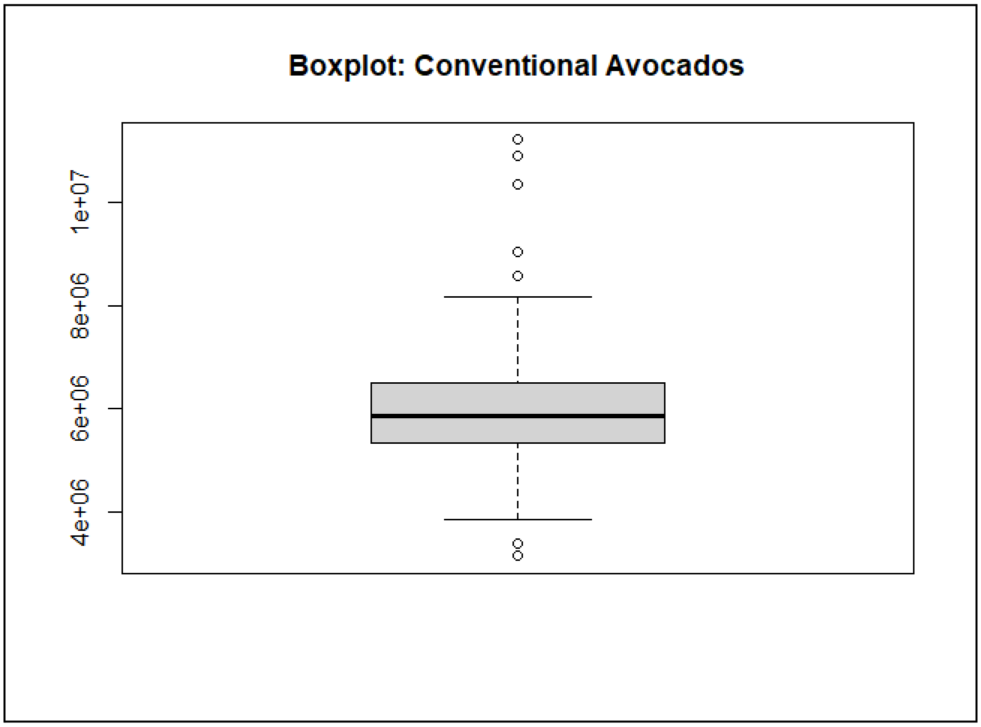

Figure 9

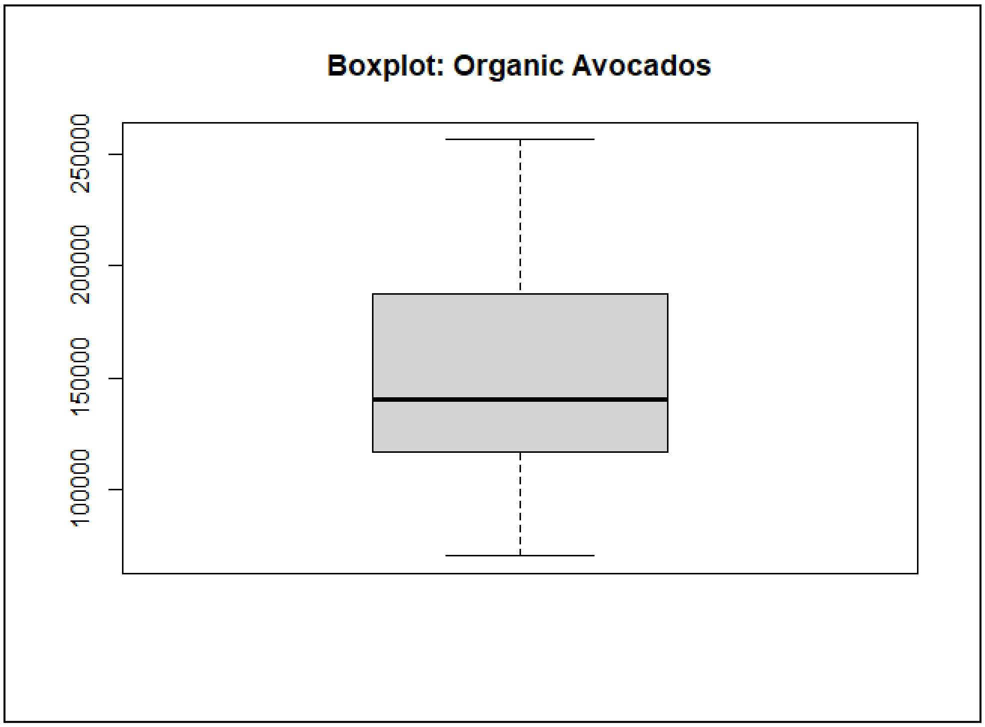

Figure 10

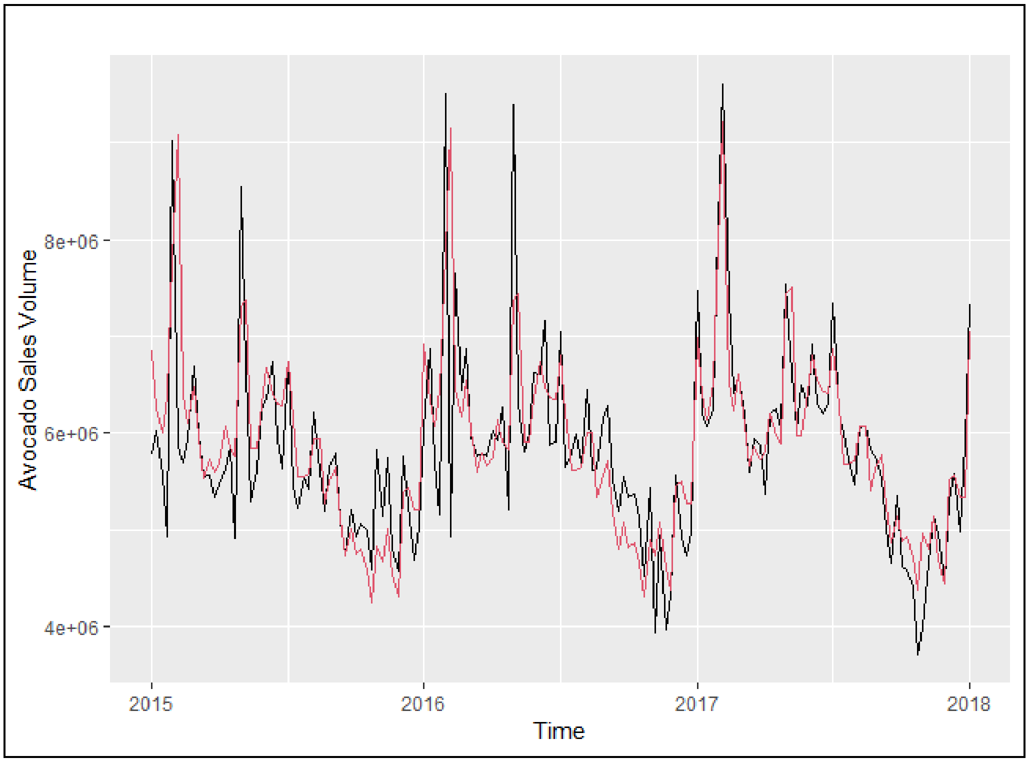

Figure 11

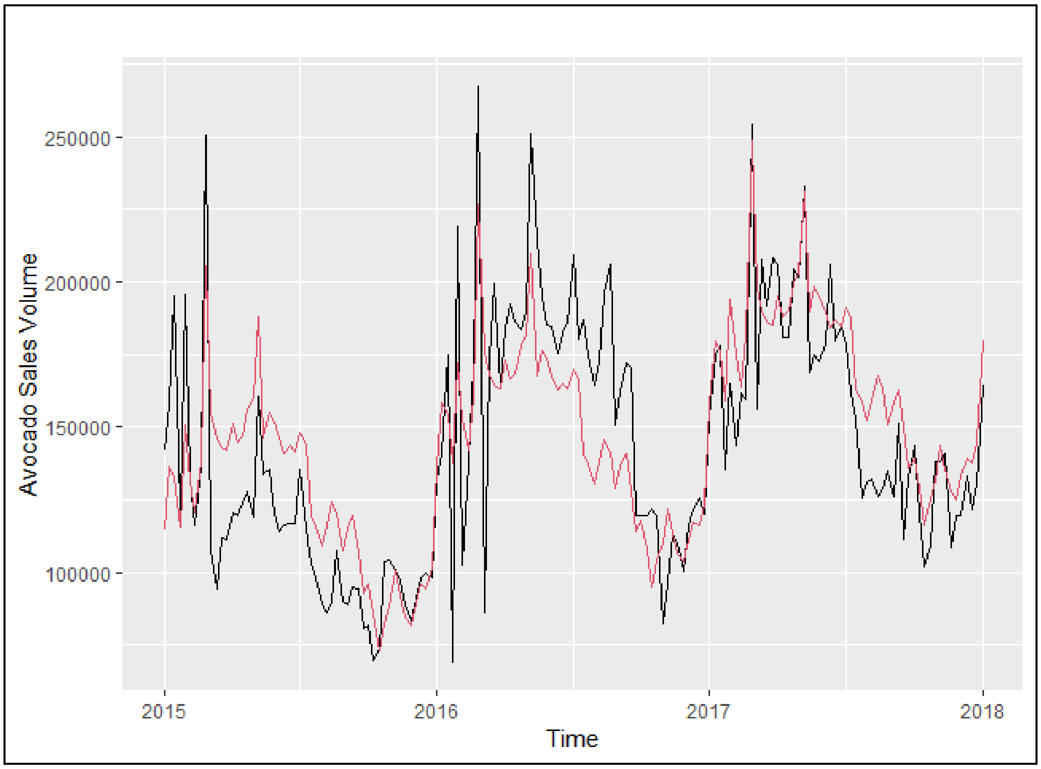

Figure 12

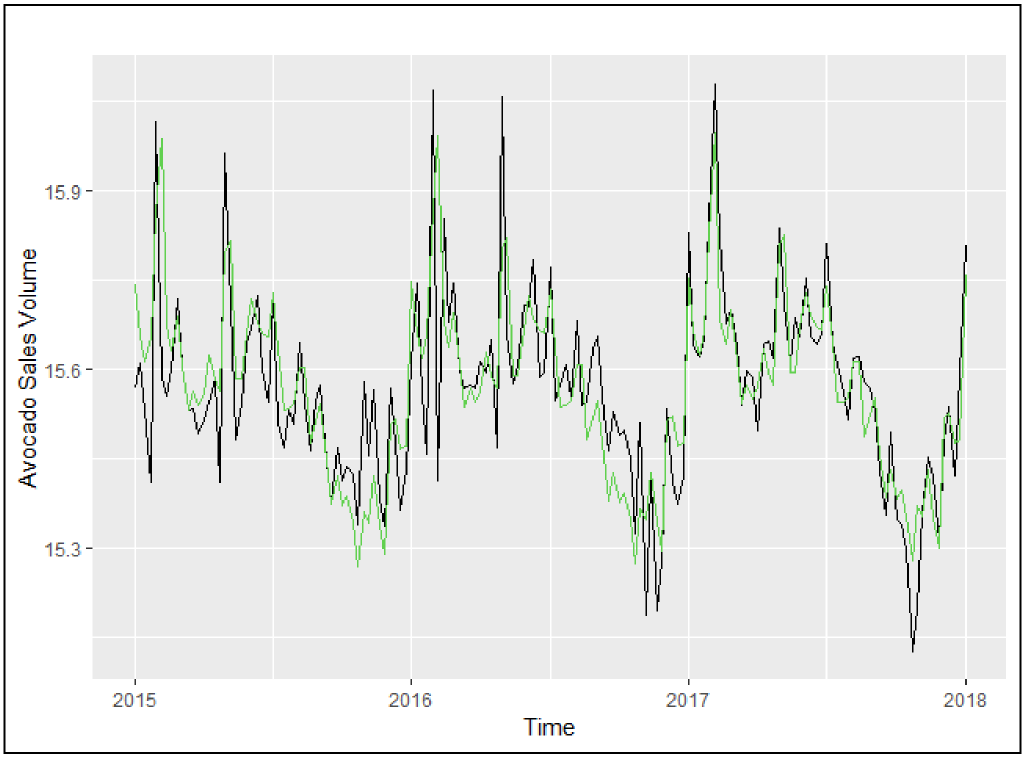

Figure 13

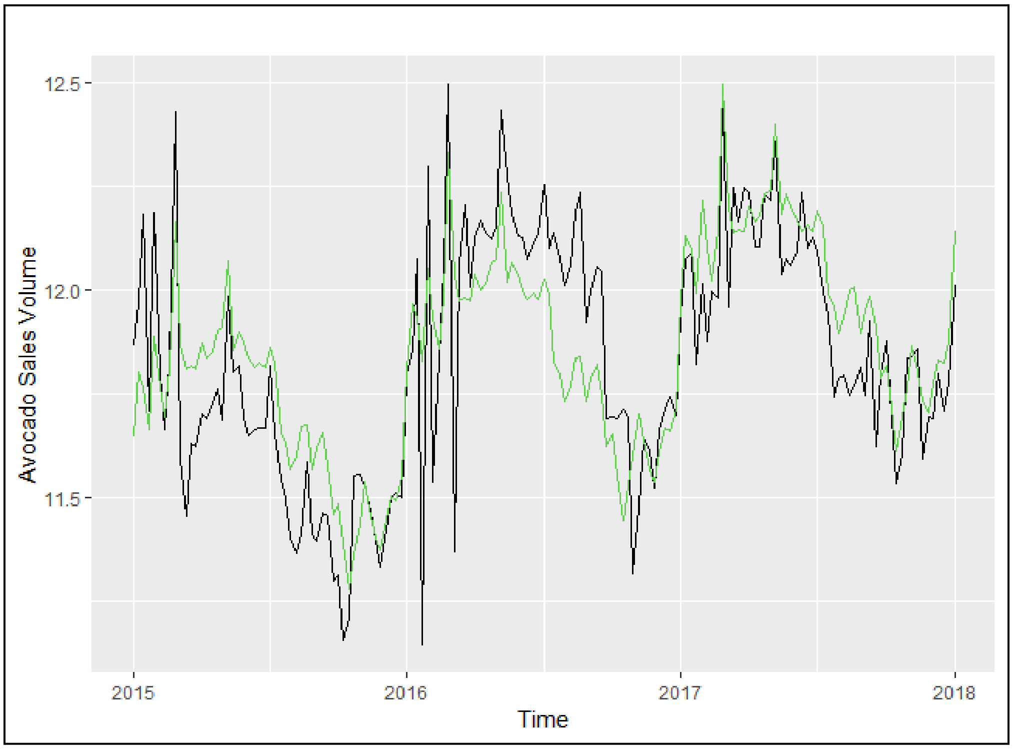

Figure 14

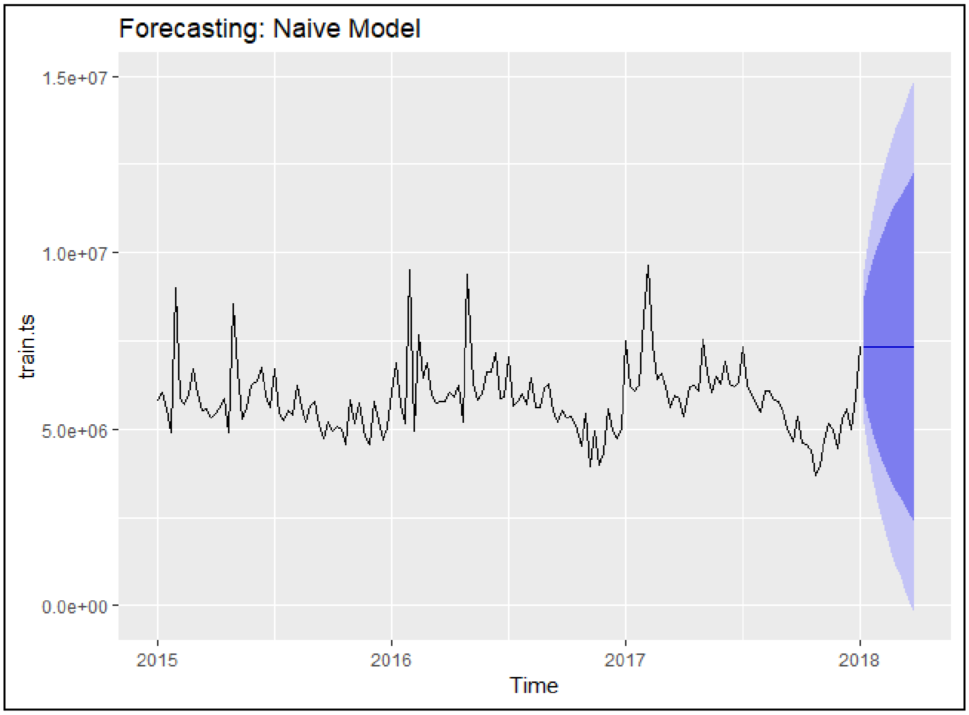

Figure 15

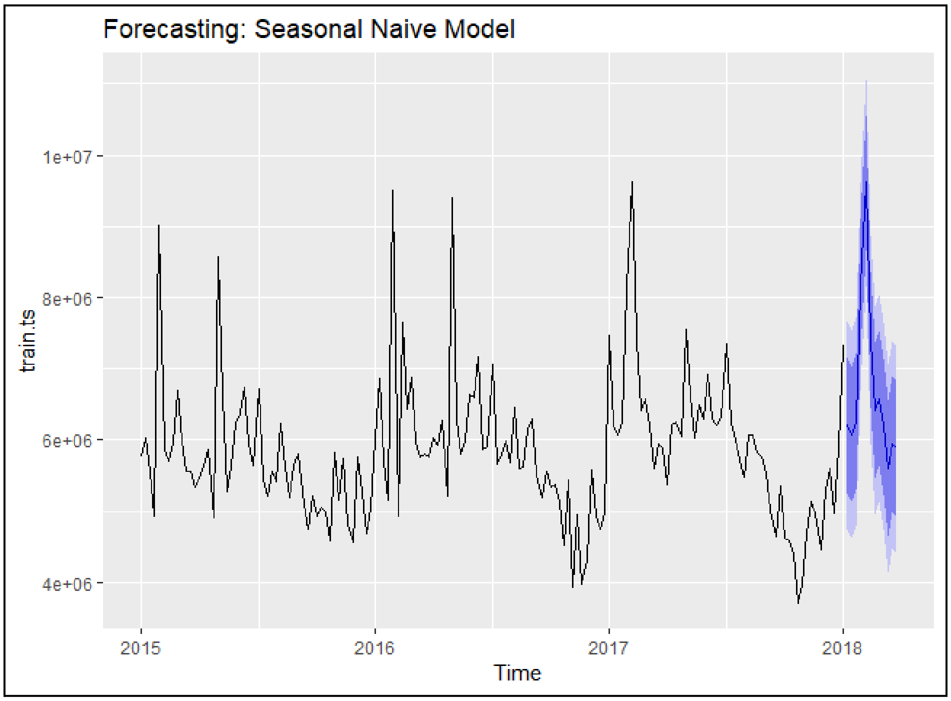

Figure 16

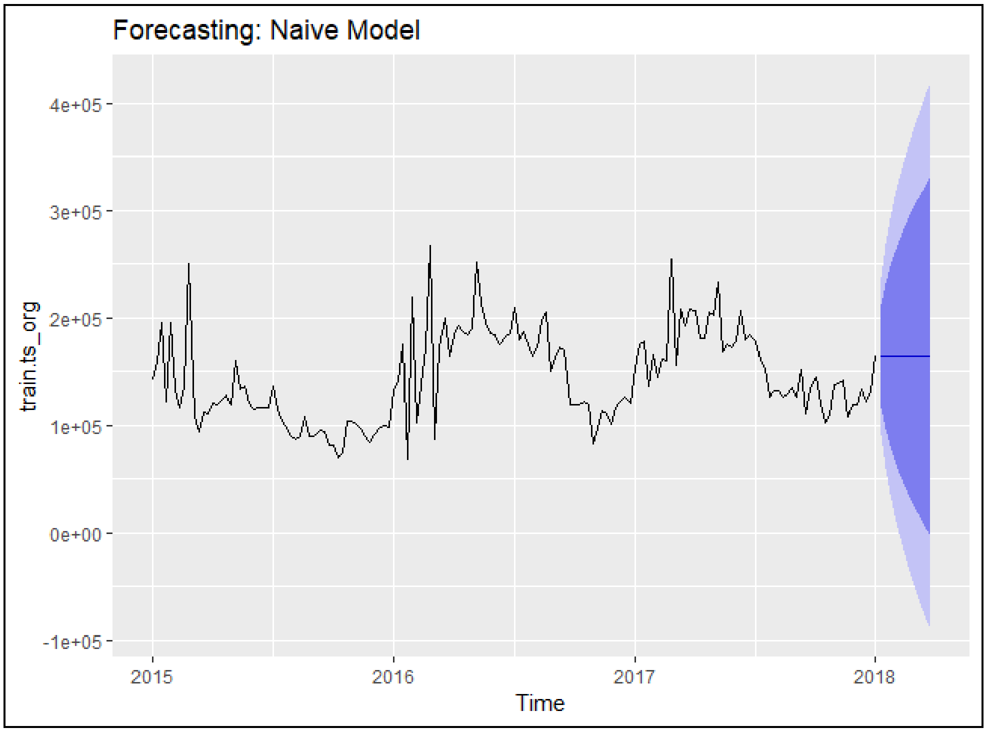

Figure 17

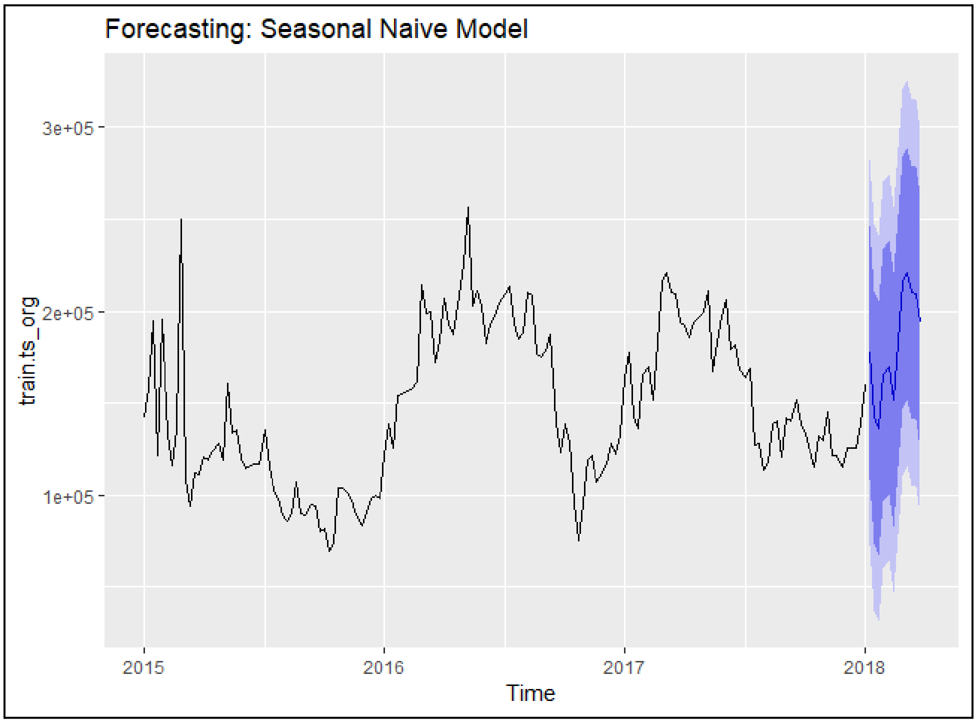

Figure 18

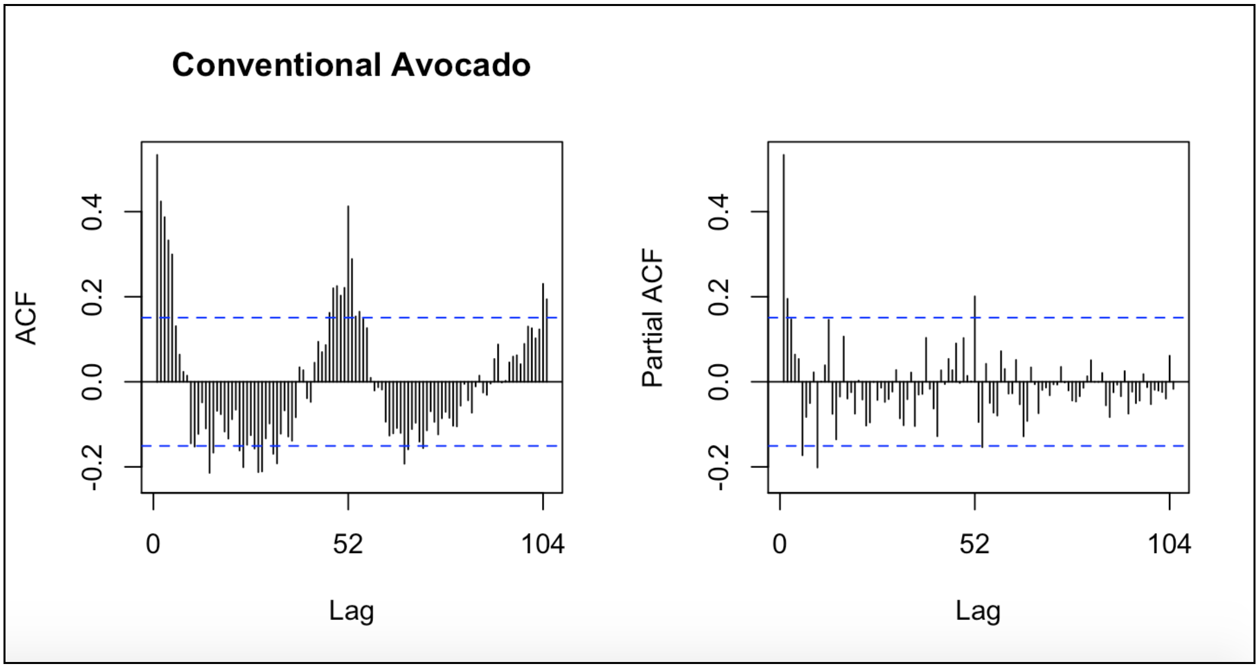

Figure 19

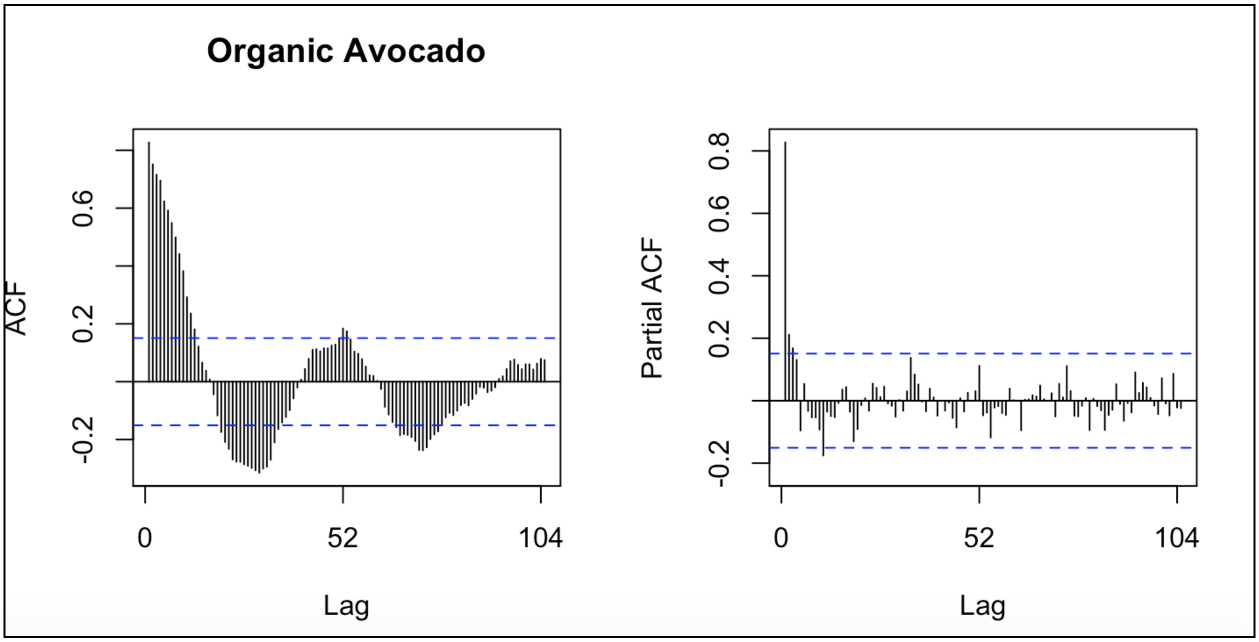

Figure 20

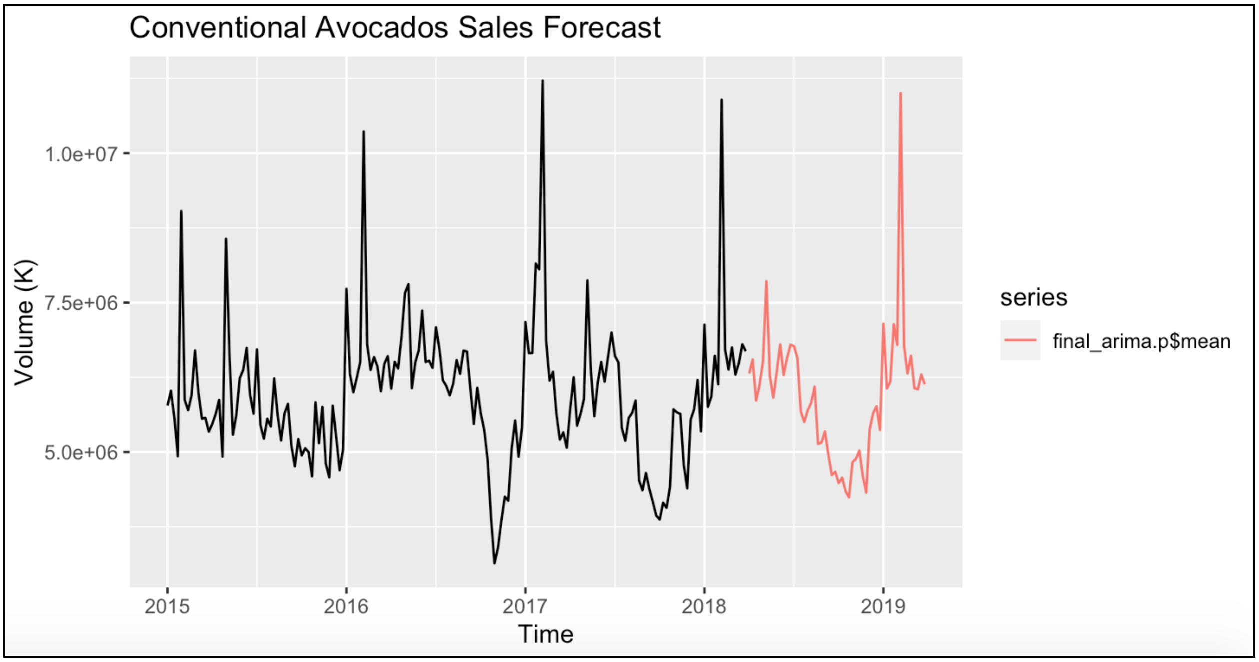

Figure 21

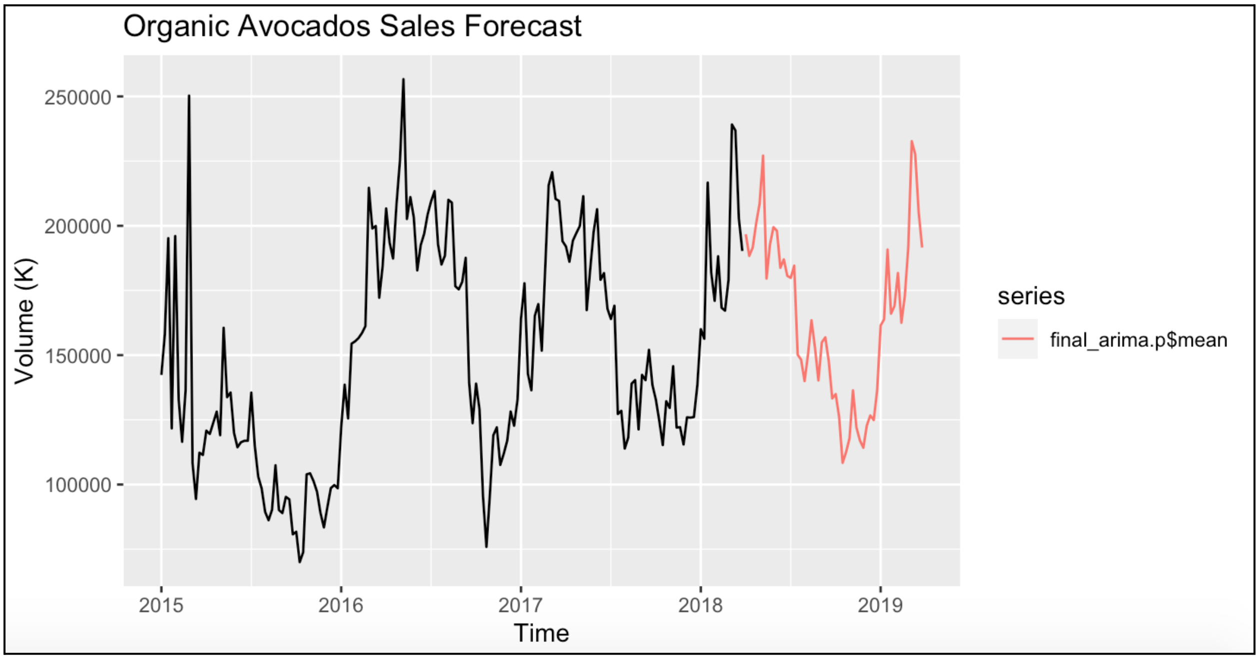

Figure 22