Predicting Sales of Hass Avocados

Yunyao Zhu

5/1/2021

1 Introduction

In 2017, a property developer claimed that millennials were spending too much money on avocado toast instead of saving for their first home (CNNMoney 2017). While the millennial avocado toast stereotype is by no means convincing, avocados are indeed a favored fruit in the U.S. According to a market report released by the United States Agency for International Development (USAID 2014), the U.S. is the world's biggest importer of avocados. In recent years, avocados have been marketed as a healthy dietary choice and a good source of beneficial monounsaturated oil (AgMRC 2018). As more health benefits are discovered and promoted, demand for avocados surged even more and so did the price.

From the perspective of a retailer or a consumer, it would be helpful to be able to reasonably forecast the per-unit retail price of avocados. For a grower or a marketer, being able to predict the sales volume of avocados would help inform business strategies. Thus, for this project, our goal is to attempt to forecast avocado sales volume and per-unit price using publicly accessible data.

1.1 Data

The dataset used in this analysis is the publicly available Kaggle Avocado Prices dataset (Kiggins 2018), which credited the Hass Avocado Board for the collection and release of the data. This dataset contains 13 variables encompassing the per-unit prices, total sales volume, regions, and sizes of Hass avocados, a cultivar of avocados. Specifically, the per-unit prices are the average retail price of individual avocados in dollars. The total sales volume is the count number of individual avocados sold. Considering the magnitude of this variable is comparatively large (an 8-digit value), we recoded it as millions of avocados to keep it as a value under 100.

Data from the start of January 2015 to the end of March 2018 are available. Each entry represents one observation from one region in the U.S. during one week. To smooth out the data, we aggregated the weekly raw data into monthly values by taking the mean.

To enrich this dataset, we joined it with selected macroeconomic data. Specifically, we included the unadjusted monthly unemployment rate released by the U.S. Bureau of Labor Statistics (BLS 2021b) and the unadjusted monthly Consumer Price Index (CPI) for all items in the U.S. (assuming Index 2015 = 100) provided by the Federal Reserve Bank (FRED 2021). The rationales for selecting these external data will be explained in the methodology section.

1.2 Research Goals

For this project, we focus on Hass conventional (non-organic) avocados sold in the U.S as a whole. We plan to conduct the forecast from April 2017 to March 2018, which corresponds to the last 12 months in our dataset. We make predictions on a full year because we suspect that there might be some seasonal patterns in avocado sales volume and per-unit price that repeat annually. We choose to use the last 12 months in our dataset because we would like to compare our predicted values with the actual results.

Thus, we specify our research goals as follows:

1. From the perspective of a grower/marketer, we would like to reasonably predict the total number of conventional Hass avocados sold in the U.S. from April 2017 to March 2018.
2. From the perspective of a buyer/consumer, we would like to reasonably predict the per-unit retail prices of conventional Hass avocados in the U.S. from April 2017 to March 2018.

1.3 Paper Outline

The paper is organized as follows:

Sections 2 and 3 will address the two research goals respectively. Each of these two sections will include its corresponding methodology, results, and discussion subsections. Section 4 concludes with a discussion of the strengths and limitations of the project. Section 5 is the Appendix and Section 6 is the Bibliography.

2 Research Goal #1: Predicting Sales Volume

2.1 Variables and EDA

For the first research question, our numerical response variable is the total count number (in millions) of conventional avocados sold in the U.S. The main predictor of interest is the per-unit retail price of avocados in dollars. Intuitively, we might expect higher sales volume when the prices are low and vice versa. The data supports this intuition. As shown in Figure 1, we largely see higher sales volume associated with lower per-unit prices. We observe a potential inverse relationship between the total sales volume and the pre-unit price of avocados[1].

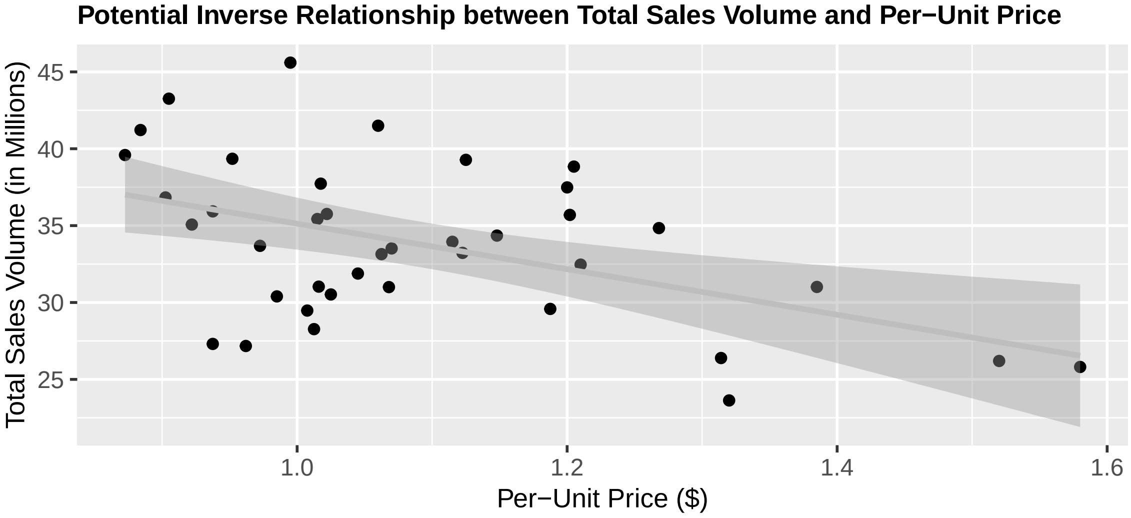

Figure 1: Scatter Plot of Total Volume of Avocados Sold in Millions v.s. Per-Unit Price of Avocados

Another independent variable of interest is unemployment rate. In economics, unemployment rate is defined as the proportion of unemployed individuals in the labor force (Picardo 2020). Since unemployed persons usually lose purchasing power, a high unemployment rate can potentially be associated with a decline in the retail sector (Amadeo 2020). In addition to unemployment rate, the Consumer Price Index (CPI) is also a commonly used macroeconomic indicator. Specifically, CPI is a measure of "the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services"(BLS 2021a). In other words, CPI can be used to assess price changes associated with the cost of living (Fernando 2020). We hope to use CPI to potentially control for some measure of the changes in the price level of consumer goods overall (ICLS 2013).

2.2 Methodology

2.2.1 Issues with a Multiple Regression Model

As a first step, it seems intuitive to use multiple linear regression to model the relationship between the total sales volume and the price, unemployment rate, and CPI. We specify the model as follows:

[1] The gray line in Figure 1 represents the linear model Sales Volume$_i = \beta_0 + \beta_1$ Per-Unit Price$_i + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. The dark gray region shows its 95% confidence interval (ggplot2 2021).

$$\text{Sales Volume}_i = \beta_0 + \beta_1 \text{ Per-Unit Price}_i + \beta_2 \text{ Unemployment Rate}_i + \beta_3 \text{ CPI}_i + \epsilon_i,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

where i represents the index of observation.

An examination of the model assumptions reveal that a few of the assumptions of an OLS model are violated. First, when checked for evidence of multicollinearity, the VIF values are all less than 5 [Page 13, Appendix 5.2.1, Table 3], which seem to fall into the commonly acceptable range (Glen 2015). However, as shown in the correlation plot [Page 14, Appendix 5.2.1, Figure 7], the correlation between unemployment rate and CPI turns out to be -0.824, which is relatively high. As a reference, the frequently used threshold for the indication of potential multicollinearity is a correlation of a magnitude of 0.7 (Wonsuk Yoo and James W. Lillard 2015). This highly negative correlation between unemployment rate and CPI might not be a surprise to economists. CPI can be used as a measure of inflation, and inflation and unemployment rate are traditionally said to be inversely correlated (Picardo 2021). To address the issue of multicollinearity, we have tried removing one of the highly correlated predictors. Removing one predictor indeed changes the OLS model outputs. Specifically, after removing CPI, the estimated coefficient of unemployment rate changes from a positive value to a negative value. This updated result seems to conform to our expectation: intuitively, higher unemployment rate might be associated with lower sales volume. In addition, the magnitude of the correlation between CPI and per-unit price (0.544) is slightly higher than the magnitude of correlation between unemployment rate and per-unit price (-0.523). Thus, we decided to remove CPI as a predictor and attempt the following model:

$$\text{Sales Volume}_i = \beta_0 + \beta_1 \text{ Per-Unit Price}_i + \beta_2 \text{ Unemployment Rate}_i + \epsilon_i,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

However, even in this updated model, some of the model assumptions are still violated. Most notably, the residuals seem to have a non-constant variance [Figure 2(a)] and also seem to be correlated over time, or potentially autocorrelated [Figure 2(b)].

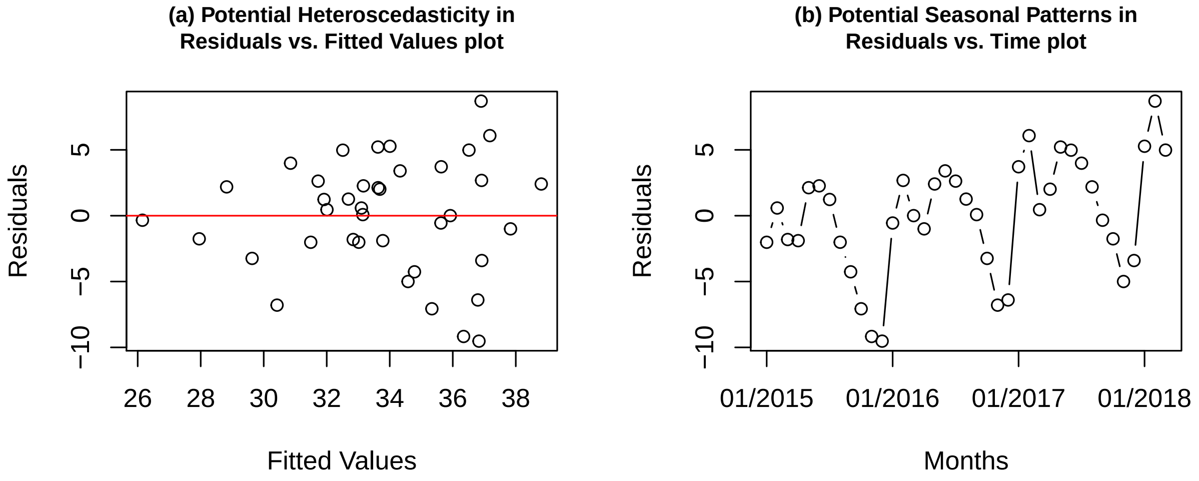

Figure 2: (a) Residuals vs. Fitted Values Plot and (b) Residuals vs. Time Plot

Actually, the patterns we observe in Figure 2(b) are somewhat expected since we are analyzing time series data. A time series is a sequence of values taken at successive equally spaced points in time (Hayes 2021). In this case, the time space is a month. When we use regression to model time series data, the error terms are often autocorrelated (Bossche, Wets, and Brijs 2004). In addition to observing the residual plots, we typically use the Autocorrelation Function (ACF) plots and the Partial Autocorrelation Function (PACF) plots to detect autocorrelation in the residuals (Hyndman and Athanasopoulos 2018a). We can take autocorrelation into account by adding structures such as autoregression (AR), moving average (MA), and/or using techniques such as differencing.

After examining the ACF and PACF plots [Page 15, Appendix 5.2.2, Figure 9], we realized that the patterns we observe in the Figure 2(b) might be evidence of AR or MA structures. In addition, the decomposition plot of the residuals [Page 15, Appendix 5.2.2, Figure 8] shows potential seasonal patterns, suggesting that we might need to incorporate seasonal differencing. To account for these structures, we choose to use a regression with Seasonal Autoregressive Integrated Moving Average (SARIMA) errors (Bossche, Wets, and Brijs 2004; Simon and Heckard 2021b; Hyndman 2021).

2.2.2 Regression with SARIMA Errors

A regression with SARIMA errors is largely similar to a linear regression model except for a different error structure (Hyndman 2010). Specifically, we can define our new model as:

$$\text{Sales Volume}_t = \beta_0 + \beta_1 \text{ Per-Unit Price}_t + \beta_2 \text{ Unemployment Rate}_t + n_t$$

where n_t, or the errors of the regression, is modeled by SARIMA. We use subscript t to index an observation in the time series (i.e. at month t). SARIMA models are usually concisely represented as $(p, d, q) \times (P, D, Q)_s$. The lowercase letters correspond to the non-seasonal components and the uppercase letters correspond to the seasonal components. Specifically, the parameter p is the order of the autoregressive (AR) model, d is the degree of differencing, and q is the order of the moving-average (MA) model. The uppercase P, D, Q denote the autoregressive, differencing, and moving average terms for the seasonal part of the SARIMA model. Lastly, s denotes the time period (SAS 2021). In this case, s is 12, denoting the 12 months in a year in our monthly version of the data. The values of the parameters p, d, q, P, D, Q are determined using principles from the Box–Jenkins method (Rundel 2017b). We choose to describe the errors by a $(1, 0, 0) \times (0, 1, 0)_{12}$ process; details of the parameter selection procedure are described in Appendix 5.2.3 [Pages 16-18].

We can express the error terms as follows:

$$(1 - \phi_1 L)(1 - L^{12})\ n_t = a_t,\ a_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

where ϕ_1 is the non-seasonal order-1 autoregression parameter, a_t is assumed to be white noise (Bossche, Wets, and Brijs 2004; Rundel 2018a), and L is the lag operator (i.e. $L^k\ n_t = n_{t-k},\ k = 1, 2, ...$) (Rundel 2018b). It is worth noting that n_t represents the errors with autocorrelation structures and a_t represents the white noise residuals.

The main assumption for a SARIMA model is that its residuals indeed follows a white noise process. Specifically, the residuals should have a mean of zero, have largely constant variance and also be uncorrelated (Asamoah-Boaheng 2014). As shown in Figure 14 [Page 20, Appendix 5.2.5], these assumptions all seem to be satisfied.

We used the Arima() function in the R forecast package to fit this regression model with SARIMA errors and the predict() function to make the predictions. Further details are included in Appendix 5.1 [Page 13]. The data from January 2015 to March 2017 are used to fit the model, while the forecast is made from April 2017 to March 2018.

2.3 Results

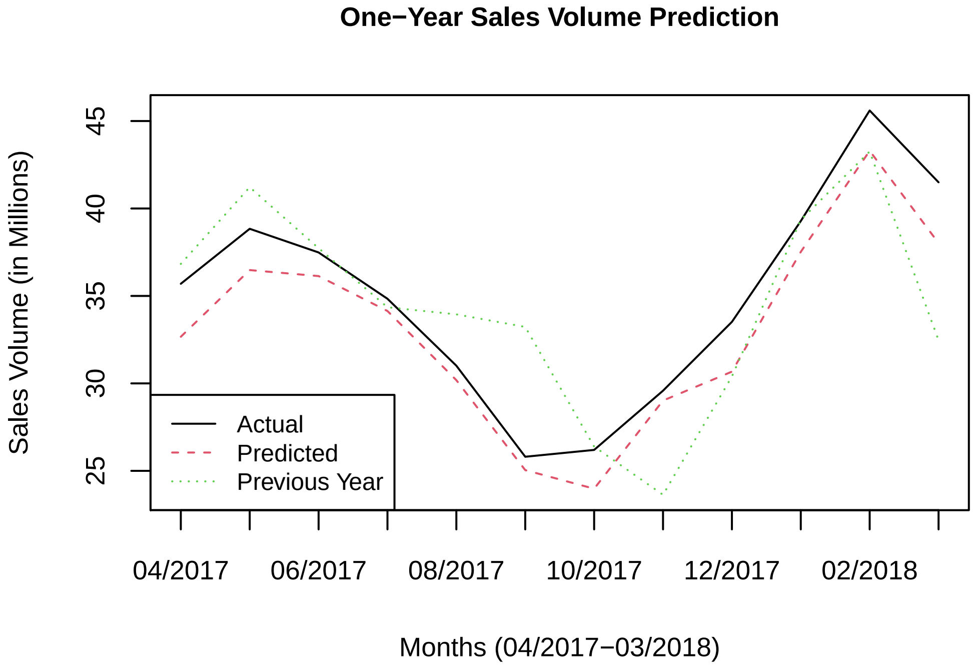

Figure 3: Predicting the Monthly Avocado Sales Volume from April 2017 to March 2018

Table 1: RMSE Comparison

Type	RMSE
Sales Volume Prediction	2.085
Sales Volume from Previous Year	4.115

2.4 Interpretation and Conclusion

Figure 3 shows the prediction results[2]. The red dashed line in Figure 3 shows the predicted sales volume from April 2017 to March 2018 while the black line shows the actual sales volume during the same time period. The green dotted line represents the sales volume in the previous year (i.e. from April 2016 to March 2017). If we do not fit a model and simply use the sales volume from the

[2] The corresponding plot with uncertainty [Page 19, Figure 13] is included in Appendix 5.2.4.

previous year, we could still roughly approximate the sales volume. However, our predicted values seem to match more closely with the actual sales volume. A commonly used metric to evaluate forecast accuracy is the Root Mean Square Error (RMSE) (Hyndman and Athanasopoulos 2018b). RMSE values are always non-negative. In general, we prefer RMSE values that are closer to 0.

The RMSE of our prediction is approximately 2.085, which is smaller than an RMSE of 4.115 if we simply use the previous year's sales volume. In the context of this analysis, this means that 2.085 is the square root of the mean of squared differences between the actual sales volume and our predicted sales volume (Wong 2018). One advantage of the RMSE metric is that it has the same unit as the response variable (Holmes 2000); in this case, the unit is millions of avocados. While an RMSE of 2.085 million avocados is a considerably large quantity, we recall that our response variable is the total number of avocados sold in the U.S. as a whole. In addition, as shown in Figure 3, the actual sales volume in one year can range from around 25 to 45 million avocados. Thus, we consider an RMSE of 2.085 million avocados an acceptable value in this context and conclude that we have made a reasonable prediction using our limited data.

It is worth noting that our predicted values under-predict the sales volume during the entire prediction time period. This might be due to some other factors not accounted for in this model such as the changes in the supply volume and/or the quality of the supply.

3 Research Goal #2: Predicting Per-Unit Price

3.1 Variables and EDA

Our numerical response variable for the second research goal is the average per-unit retail price of conventional avocados in the U.S. It is very tempting to include the total sales volume as the predictor given the potential inverse relationship shown in Figure 1. However, we recognize that we would not know the sales volume when we do not know the retail prices. Thus, it does not make sense to use the sales volume to predict the prices. One counter-argument is that we can potentially use the sales volume as a proxy for demand. While this is not entirely convincing, we created a model using the sales volume as a predictor in a sensitivity analysis in Appendix 5.3.1 [Pages 20-24]. For the main analysis, we choose to proceed without the sales volume variable.

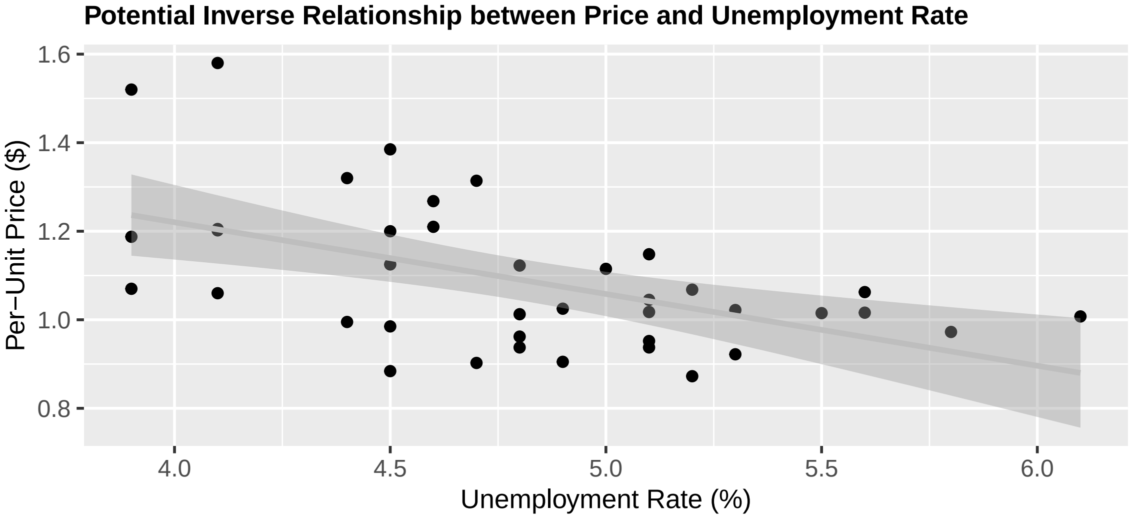

Figure 4: Scatter Plot of Per-Unit Price v.s. Unemployment Rate

Other available predictors are the macroeconomic measures unemployment rate and CPI. In general, we expect higher unemployment rates to be associated with lower prices. As shown in Figure 4, there indeed seems to be a potential inverse relationship between unemployment rate and per-unit prices of avocados. On the other hand, since CPI can be interpreted as a measure of inflation (Fernando 2020), we expect higher CPI values to be associated with high prices. When plotting the prices against the CPI values, we find a potential positive relationship between the two variables as expected.

3.2 Issues with an OLS Model

As discussed in Section 2, we observed a high magnitude of correlation between unemployment rate and CPI, suggesting potential multicollinearity. To address this issue, we also tried removing one of the two predictors. In this case, removing either CPI or unemployment rate does not change the OLS model outputs by much. Proceeding with either model yields very similar results. Thus, we choose to present the following model using unemployment rate as the predictor while noting that we could instead use CPI as the predictor and the subsequent procedure would be roughly the same.

We specify the OLS model as follows:

$$\text{Per-Unit Price}_i = \beta_0 + \beta_1 \text{ Unemployment Rate}_i + \epsilon_i,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

We note that some assumptions of this model are still violated. For instance, the residuals seem to have a non-constant variance [Figure 5(a)] and also seem to be correlated over time [Figure 5(b)]. The decomposition plot shows potential seasonal patterns, and the ACF and PACF plots show evidence of autocorrelation.

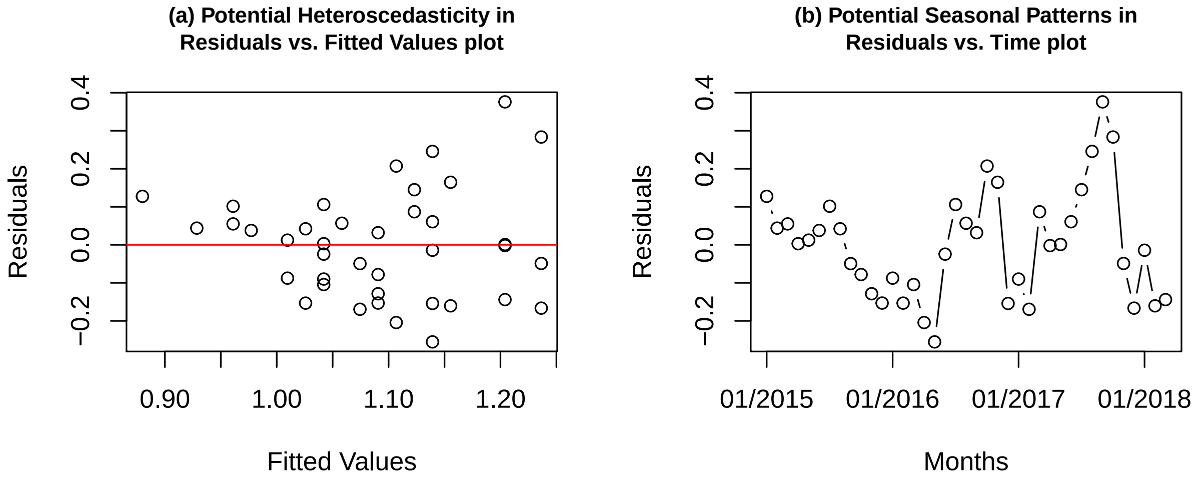

Figure 5: (a) Residuals vs. Fitted Values Plot and (b) Residuals vs. Time Plot

Following a similar argument in Section 2, we would like to account for these time series structures using a regression with SARIMA errors.

3.2.1 Regression with SARIMA Errors

From Section 2, we know that we can express our regression with SARIMA errors as follows:

$$\text{Per-Unit Price}_t = \beta_0 + \beta_1 \text{ Unemployment Rate}_t + n_t$$

where the information remaining in the error terms are modeled by SARIMA $(1,0,0)\times(0,1,0)_{12}$. The values of p,d,q and P,D,Q are selected using the Box-Jenkins method (Rundel 2017b). We use the subscript t to index an observation (i.e. at month t). We thus have:

$$(1-\phi_1 L)(1-L^{12})\ n_t = a_t,\ a_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$$

where ϕ_1 is non-seasonal order-1 autoregression parameter, a_t is assumed to be white noise, and L is the lag operator (Rundel 2018b).

Figure 20 [Page 25] in Appendix 5.3.2 shows that our the residuals of our regression with SARIMA errors model largely resembles a white noise process, satisfying the assumptions of zero mean and uncorrelated residuals (Asamoah-Boaheng 2014).

We note that the data from January 2015 to February 2017 are used to fit the model, while the forecast is made from April 2017 to March 2018.

3.3 Results

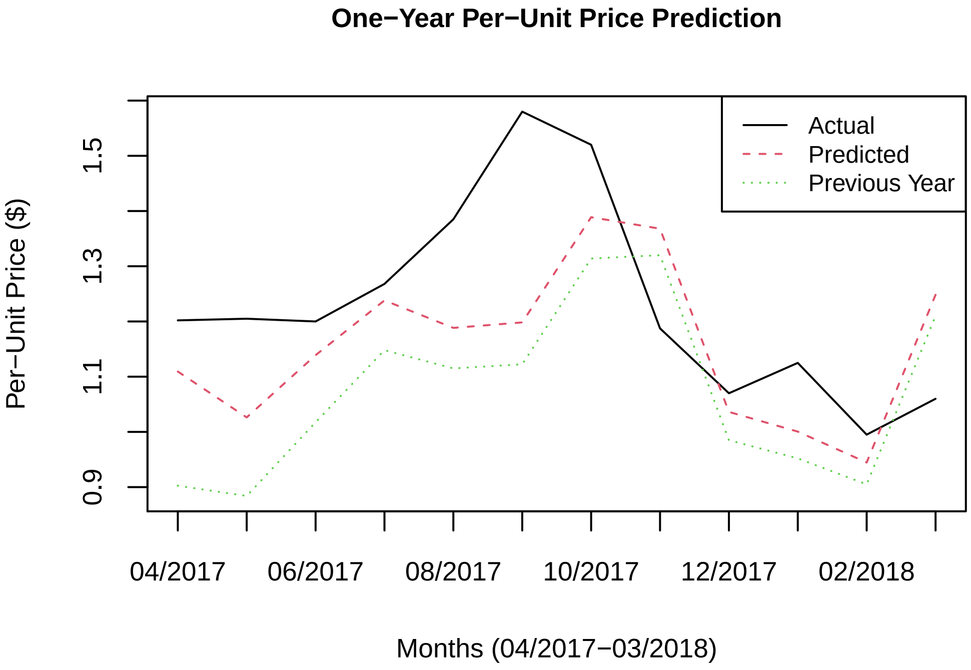

Figure 6: Predicting the Per-Unit Price of Avocados from April 2017 to March 2018

Table 2: RMSE Comparison

Type	RMSE
Per-Unit Price Prediction	0.167
Per-Unit Price from Previous Year	0.233

3.4 Interpretation and Conclusion

The predicted results are shown in Figure 6. The black line shows the actual per-unit retail price of avocados from April 2017 to March 2018, the dashed red line represents the predicted results, and the green dotted line is the per-unit price from the previous year (April 2016 to March 2017).

First, we notice that the predicted values do not seem to match the actual results very closely. The seasonal trend is largely captured, but there seems to be a slight “delay” in the predicted values. For instance, according to the actual values, the per-unit price peaked in September 2017, while our highest predicted value occurs in October 2017. The price drop between September and October 2017 in the actual values is somewhat mirrored in our prediction, although the drop is shifted one month back.

Actually, our predicted values are closer to previous year’s prices than to the actual prices. This is likely due to our lack of external data. Perhaps we have not accounted for some important changes in the production of Hass avocados in our model. For example, it is said that the Hass cultivar has a tendency to bear well in alternate years. After a season with a relatively low yield, due to factors such as cold weather, the Hass avocado trees tend to produce more abundantly during the next season. When the crop depletes the stored nutrients, the yield in the following season would be reduced, thus establishing an alternate bearing pattern (AgMRC 2018). In future analyses, if we would have access to the actual production data and/or data for a longer time period (i.e. more than three years of data), we would potentially be able to better predict the prices.

Nonetheless, given our current data and model, our predicted values are closer to the truth than simply using the prices from the previous year. Specifically, the RMSE of our prediction is 0.167, which is smaller than the RMSE of the prices from last year (0.233). In the context of this analysis, $0.167 is the square root of the mean of the squared differences between the actual per-unit retail prices of avocados and our predicted prices. Given that the actual average per-unit price of avocados can range from around $1.0 to $1.6, we consider an RMSE of $0.167 an acceptable value given our limited data.

4 Discussion

In this project, we used regression models with SARIMA errors to predict the sales of Hass avocados. As discussed in the previous two sections, we believe that we have reasonably addressed our two research goals. Specifically, we are able to predict the total sales volume of conventional avocados in the U.S. from April 2017 to March 2018 with an RMSE of 2.805 million avocados. We then predicted the average per-unit retail price of conventional avocados in the U.S. from April 2017 to March 2018 with an RMSE of $0.167.

One of the main strengths of this analysis is that we have accounted for the autocorrelation structure and the seasonal pattern in the errors of our regression models. In addition, our predictions are somewhat accurate given our very limited data. Our data are limited in two ways. First, we only have a little more than three years (39 months) of data. Given that we have decided to make predictions on an entire year, we only have 27 months to fit the model. Second, even though there are 13 variables in the avocado dataset, many of them are linear combinations of each other and are thus not usable in the model. For instance, the total sales volume is the sum of the sales volume of small, large, and extra large avocados. In addition, the macroeconomic measures unemployment rates and CPI turned out to be highly correlated. In future analysis, we hope to potentially incorporate additional predictors such as supply volume, import/export volume, tariff, and weather to better inform our prediction.

Beyond the limitations in our data, there is another important limitation with a regression model approach: if we want to make predictions about the future (instead of the known 04/2017-03/2018 time period), we would need the future data for our predictors. The actual values are certainly not accessible, but we can potentially obtain plausible predictions. For instance, organizations such as the Federal Reserve releases their forecasts for macroeconomic measures including the unemployment rate (FOMC 2021). We can use these projections to predict the per-unit prices and subsequently use our predicted prices to forecast the sales volume.

An interesting extension of this study would be to collect and use avocado data in the last two years to investigate the avocado retail sales before and during the pandemic. 2019 is a special year because earlier that year, tariffs and transportation-related delays contributed to avocado supply chain disruptions. Avocado prices in the first half of 2019 were unusually high as a consequence (Johnson 2020; Schwartz 2019). On the other hand, the avocado industry has been remarkably resilient during the pandemic. It is said that the demand for avocados surged during the pandemic since people are paying more attention to their health and wellness, and more people are cooking at home (Perez 2020). A previous study (Evans and Nalampang 2009) might be relevant as the researchers were predicting the avocado prices in the 2009-2010 season in the midst of an economic crisis. The study incorporated annual data of per capita avocado consumption and disposable income as predictors in a multiple regression model. Inspired by this approach, we believe that incorporating avocado consumption data as a proxy for demand and accounting for income-related data (e.g. economic impact payments) might help better inform the predictions for 2019 and 2020.

5 Appendix

5.1 Using the `forecast` package

When using the `Arima()` function, we can specify the non-seasonal component using the `order` argument and the seasonal component using the `seasonal` argument. We can incorporate our predictors as a $n \times p$ matrix in the `xreg` argument where n is the number of the observations and p is the number of predictors.

When using the `predict()` function, we can specify the values of the predictors during the forecast period in the `newxreg` argument as a $n' \times p$ matrix where now the value of n' is 12 months as we are predicting the sales volume in the last 12 months.

5.2 Supplementary Materials for Research Goal #1: Predicting Sales Volume

5.2.1 OLS Model with Per-Unit Price, Unemployment Rate, and CPI as Predictors

Model Formulation:

$$\text{Sales Volume}_i = \beta_0 + \beta_1 \text{ Per-Unit Price}_i + \beta_2 \text{ Unemployment Rate}_i + \beta_3 \text{ CPI}_i + \epsilon_i,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Table 3: First OLS Model VIF Values

Variable	VIF Value
Per-Unit Price	1.457
Unemployment Rate	3.188
CPI	3.291

The VIF values for all three predictors are within the commonly acceptable range (i.e. less than 5) (Glen 2015). However, according to the the correlation plot [Figure 7], the unemployment rate seems to be highly negatively correlated with the CPI.

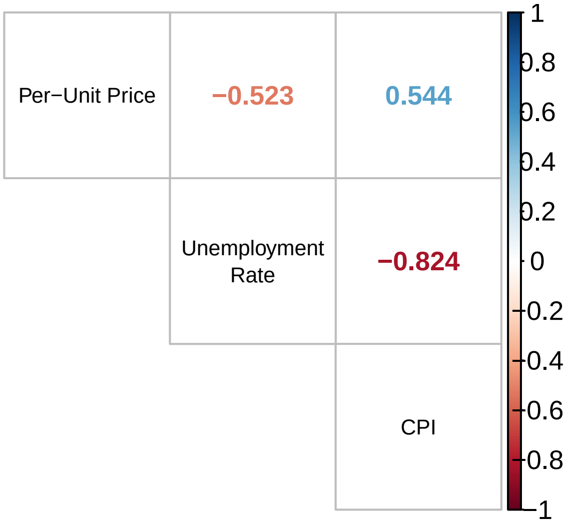

Figure 7: Unemployment Rate and CPI Highly Correlated

5.2.2 OLS Model with Per-Unit Price and Unemployment Rate as Predictors

Model Formulation:

$$\text{Sales Volume}_i = \beta_0 + \beta_1 \text{ Per-Unit Price}_i + \beta_2 \text{ Unemployment Rate}_i + \epsilon_i,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Table 4: Updated Model VIF Values

Variable	VIF Value
Per-Unit Price	1.457
Unemployment Rate	3.188

The VIF values per-unit price and unemployment rate are still small (< 5) in this updated model [Table 4]. Furthermore, the magnitude of the correlation between the two predictors is $0.523 < 0.7$, suggesting no severe multicollinearity (Wonsuk Yoo and James W. Lillard 2015).

One of the commonly used visualization tools for a time series is the decomposition plot. The decomposition plot is constructed based on the decomposition model, which reduces a time series into three components: trend, seasonal effects, and random errors (E. E. Holmes and Ward 2021).

An additive decomposition model for a time series x_t can be represented as

$$x_t = m_t + s_t + e_t$$

where at time t, m_t is the trend, s_t is the seasonal effect, and e_t is a random error assumed to have a mean of zero and correlated over time (E. E. Holmes and Ward 2021).

Below, we use the `decompose()` function in R to compute and plot the seasonal effects. According to the function description, the seasonal effects are computed by “averaging, for each time unit, over all periods” (RDocumentation 2021).

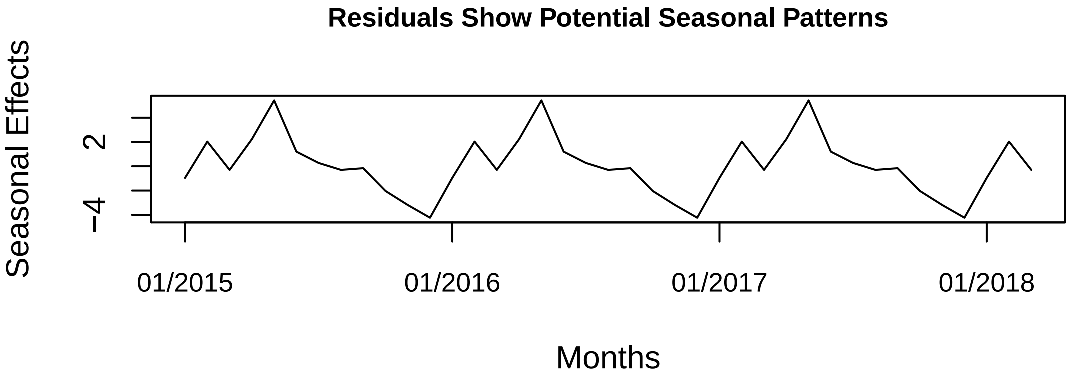

Figure 8: Seasonal Panel of Decomposition Plot

We observe a potential seasonal pattern repeated annually according to Figure 8.

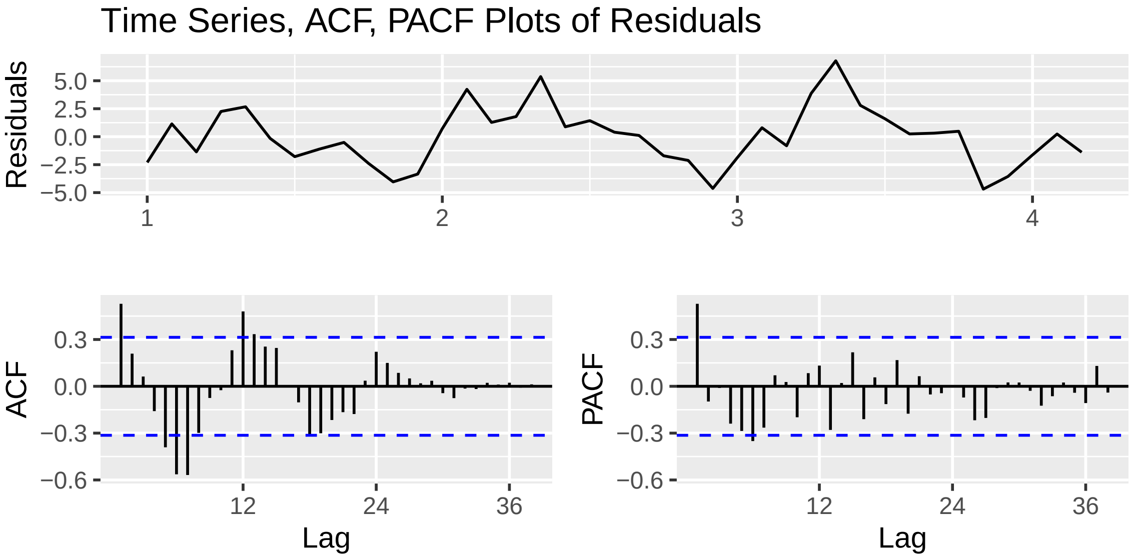

Figure 9: Time Series, ACF, PACF Plots of Residuals

According to the ACF and PACF plots in Figure 9, there seems to be evidence of autocorrelation in the residuals.

5.2.3 Parameter Selection for the SARIMA Process

Referring back to Figure 9, we observe a pattern similar to a sinusoidal oscillation in the ACF plot, which is suggestive of a seasonal component and an autoregressive structure (Nau 2020). We also notice the spikes at lag 12 and 24. Thus, we would like to check if we should use seasonal differencing.

We first fit two regressions with SARIMA errors $(0,0,0) \times (0,0,0)_{12}$ and $(0,0,0) \times (0,1,0)_{12}$, respectively. We fit these models using total sales volume as the response variable, per-unit price and unemployment as the predictors, and the data from January 2015 to March 2017.

Table 5: Lower Model Selection Criterion Values for SARIMA(0,0,0)(0,1,0)[12]

p	d	q	P	D	Q	AIC	AICc	BIC
0	0	0	0	0	0	153.739	154.783	157.627
0	0	0	0	1	0	83.520	84.520	84.936

As shown in Table 5, including a seasoning differencing greatly reduces the AIC, AICc, and BIC values. AIC (Akaike Information Criterion), AICc (corrected Akaike Information Criterion), and BIC (Schwarz's Bayesian Information Criterion) are commonly used model selection criteria. Compared to AIC, BIC tends to favor more parsimonious models (Tackett 2019), while AICc is said to be preferred when the sample size is small (Burnham and Anderson 2004). In this case, all three criteria agree. Thus, we proceed with the model with a seasonal difference (i.e. $D = 1$).

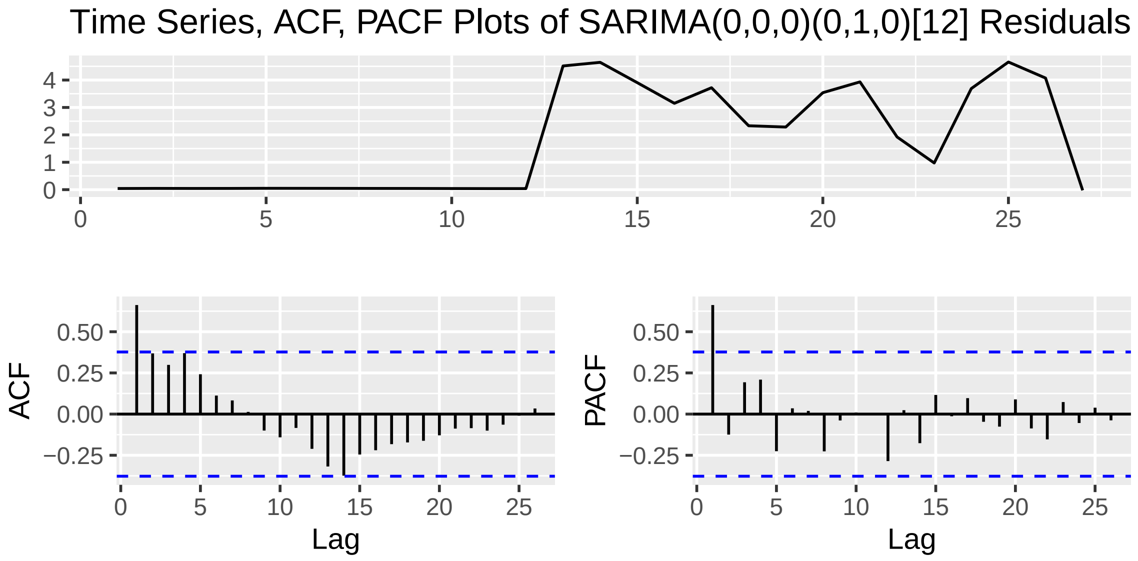

Figure 10: Time Series, ACF, PACF Plots of Residuals

We check the residuals of the regression with SARIMA(0,0,0)(0,1,0)[12] errors. Ideally, these residuals should resemble white noise. For a white noise series, we usually expect the spikes in the ACF and PACF plots to roughly lie within $\pm\frac{2}{\sqrt{T}}$, where T is the length of the time series (Hyndman and Athanasopoulos 2018c). The blue dashed lines represent these bounds, in this case

$\pm\frac{2}{\sqrt{27}} \approx \pm 0.385$. However, observing the ACF and PACF plots in Figure 10, there still seem to be AR or MA structures in the residuals.

Thus, we would like to first check for AR or MA structures in the seasonal component. We fit two additional regressions with SARIMA errors $(0,0,0)\times(1,1,0)_{12}$ and $(0,0,0)\times(0,1,1)_{12}$, respectively.

Table 6: Lowest Model Selection Criterion Values for SARIMA(0,0,0)(1,1,0)[12]

p	d	q	P	D	Q	AIC	AICc	BIC
0	0	0	0	1	0	83.520	84.520	84.936
0	0	0	1	1	0	78.960	81.142	81.084
0	0	0	0	1	1	82.628	84.810	84.752

In this case, according to Table 6, the criterion values for the three models are relatively similar, with the $(0,0,0)\times(1,1,0)_{12}$ model having slightly smaller values compared to the other two models.

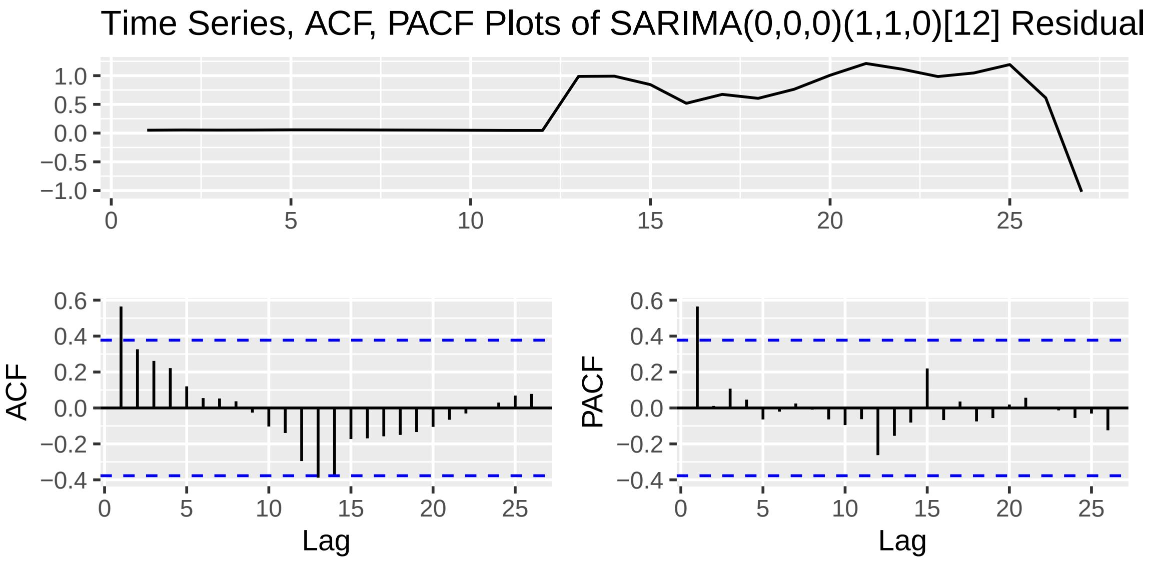

Figure 11: Time Series, ACF, PACF Plots of Residuals

The residual plots of the three models are very similar. Here, we included the residual plots of the regression with SARIMA(0,0,0)(1,1,0)[12] errors in Figure 11 for reference. We notice that even after accounting for seasonal AR or MA components, there still seem to be autocorrelation in the residuals. Specifically, we observe a spike at lag 1 that exceeds the dashed blue line in the PACF plot. This suggests a possible non-seasonal AR(1) component (i.e. $p = 1$) (Simon and Heckard 2021a).

As a next step, we would like to fit three additional regressions with SARIMA errors $(1,0,0)\times(0,1,0)_{12}$, $(1,0,0)\times(1,1,0)_{12}$, and $(1,0,0)\times(0,1,1)_{12}$, respectively.

Table 7: Lowest Model Selection Criterion Values for SARIMA(1,0,0)(0,1,0)[12]

p	d	q	P	D	Q	AIC	AICc	BIC
1	0	0	0	1	0	44.281	46.462	46.405
1	0	0	1	1	0	44.691	48.691	47.523
1	0	0	0	1	1	44.971	48.971	47.803

According to Table 7, the criterion values for the three models are almost half of the previous values. The values among the models are very similar, with the $(1, 0, 0) \times (0, 1, 0)_{12}$ model having slightly smaller values compared to the other two models and being the most parsimonious.

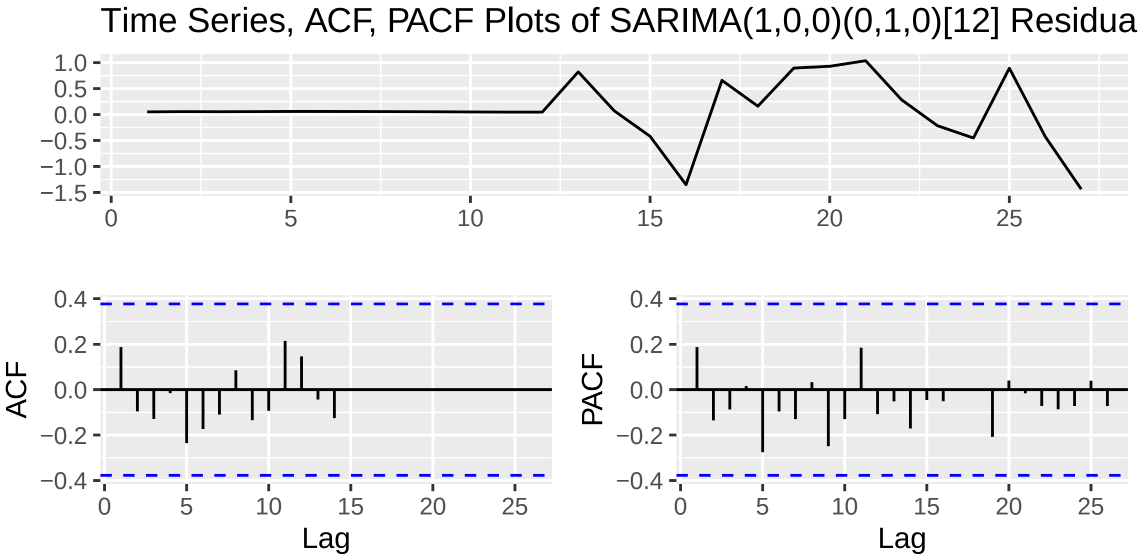

Figure 12: Time Series, ACF, PACF Plots of Residuals

We examine the residual plots of the regression with SARIMA(1,0,0)(0,1,0)[12] errors in Figure 12. This time, the residuals indeed seem like white noise. We feel comfortable proceeding with this model.

5.2.4 Prediction with Uncertainty

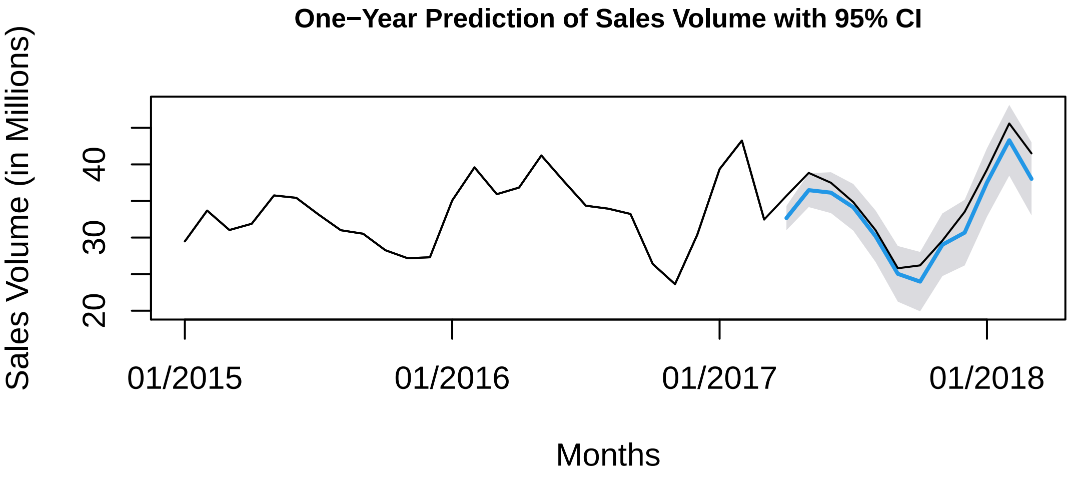

Figure 13: Sales Volume Prediction with Uncertainty

Figure 13 shows the actual sales volume from January 2015 to March 2018 in black and the predicted sales volume from April 2017 to March 2018 in blue. The shaded region represents the 95% confidence interval of the prediction. We notice that the actual sales volume from April 2017 to March 2018 is largely covered by the shaded region.

5.2.5 Model Assumptions

As shown in Figure 14, the plot in the upper panel is the residuals vs. time plot. The first 12 residuals seem to be 0; this is due to the 12-month seasonal differencing in the seasonal component of the SARIMA model. The rest of the residuals roughly have a mean of zero and constant variance as the residuals as mostly between -1 and 1. The ACF plot of the residuals show that the autocorrelation of the residuals are very close to zero (between -0.2 and 0.2), suggesting that the residuals are not very correlated. We observe that all the spikes in the ACF plot are between the blue dashed lines, suggesting no severe autocorrelation. Lastly, the residuals seem to be roughly normally distributed. The spike of residuals at 0 are due to the seasonal differencing. We conclude that our selected SARIMA model for the error terms $(1, 0, 0) \times (0, 1, 0)_{12}$ satisfies all the model assumptions (Asamoah-Boaheng 2014).

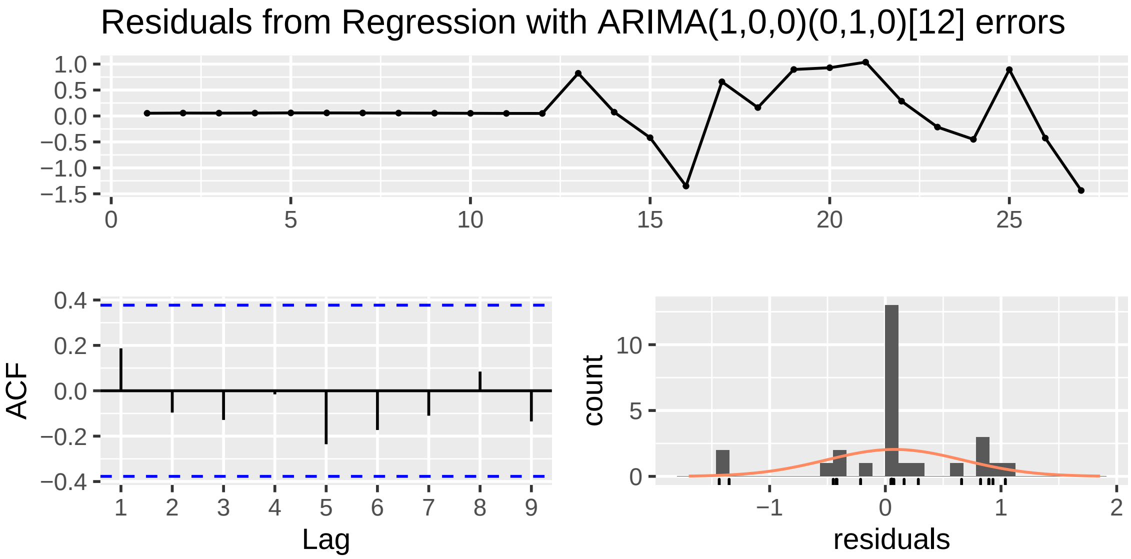

Figure 14: Residuals Check of the SARIMA Model for the Error Terms

5.3 Supplementary Materials for Research Goal #2: Predicting Per-Unit Price

5.3.1 Sensitivity Analysis with Sales Volume as a Predictor

As mentioned in Section 3.1, we would like to potentially use the sales volume as a proxy for demand to predict the per-unit price. The other possible predictors are unemployment rate and CPI. As previously shown, the variables unemployment rate and CPI are highly negatively correlated. To address this potential multicollinearity, we compared the model outputs of the full model with two other models with either unemployment rate or CPI removed. Similar to our observation in Section 2.2.1, after we remove CPI from the model, the estimated coefficient changed from a positive value (0.113) to a negative value (-0.159). Its corresponding p-value also decreases from 0.033 a value less than 0.001. Since we observe a potential negative association between per-unit price and unemployment rate, the outputs of the updated model seem to meet our expectation that higher prices are associated with lower unemployment rates. Thus, we decided to try to remove CPI and proceed with the following model.

$$\text{Per-Unit Price}_i = \beta_0 + \beta_1 \text{ Sales Volume}_i + \beta_2 \text{ Unemployment Rate}_i + \epsilon_i,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

As we check for OLS model assumptions, we once again noticed some patterns in the residual plots. In Figure 15a, we noticed a fanning pattern in the residuals vs. fitted values plot, suggesting potential heteroscedasticity. In Figure 15b, the residuals seem to be correlated over time.

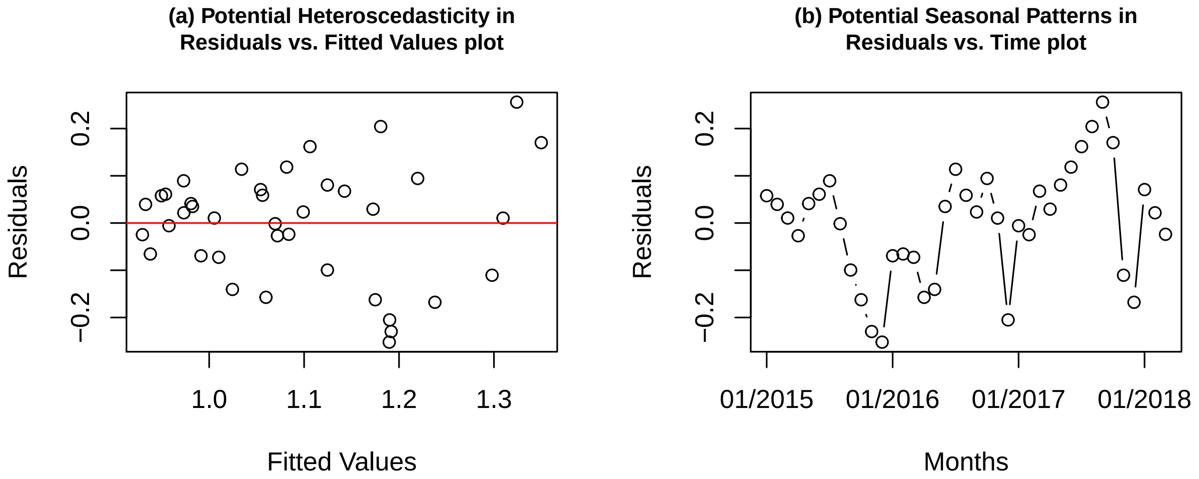

Figure 15: (a) Residuals vs. Fitted Values Plot and (b) Residuals vs. Time Plot

To further investigate these patterns in the residuals, we use the `decompose()` function in R to compute and plot the seasonal effects. As presented in Figure 16, there seems to be a seasonal pattern that is repeated yearly.

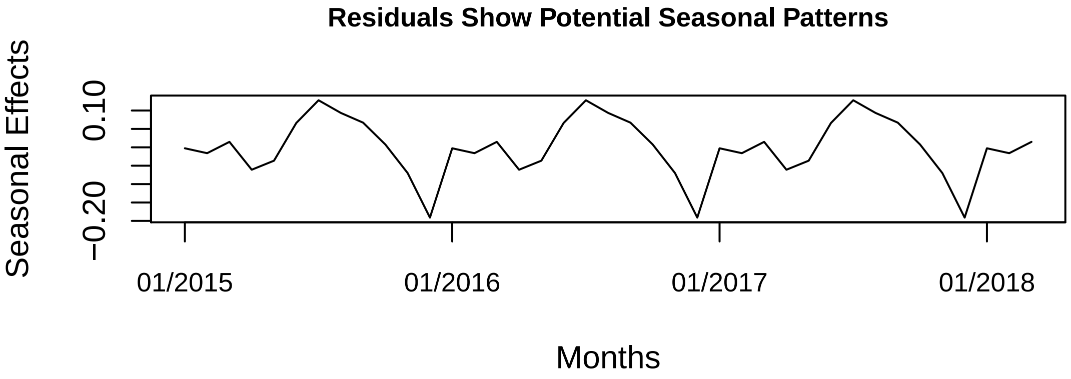

Figure 16: Seasonal Panel of Decomposition Plot

The ACF and PACF plots show slight evidence of autocorrelation in the residuals. Specifically, the lag 1 autocorrelation and partial autocorrelation both exceed the dashed blue line. There also seems to exist a pattern similar to a sinusoidal oscillation in the ACF plot, suggesting potential seasonal components and/or an autoregressive structure (Nau 2020).

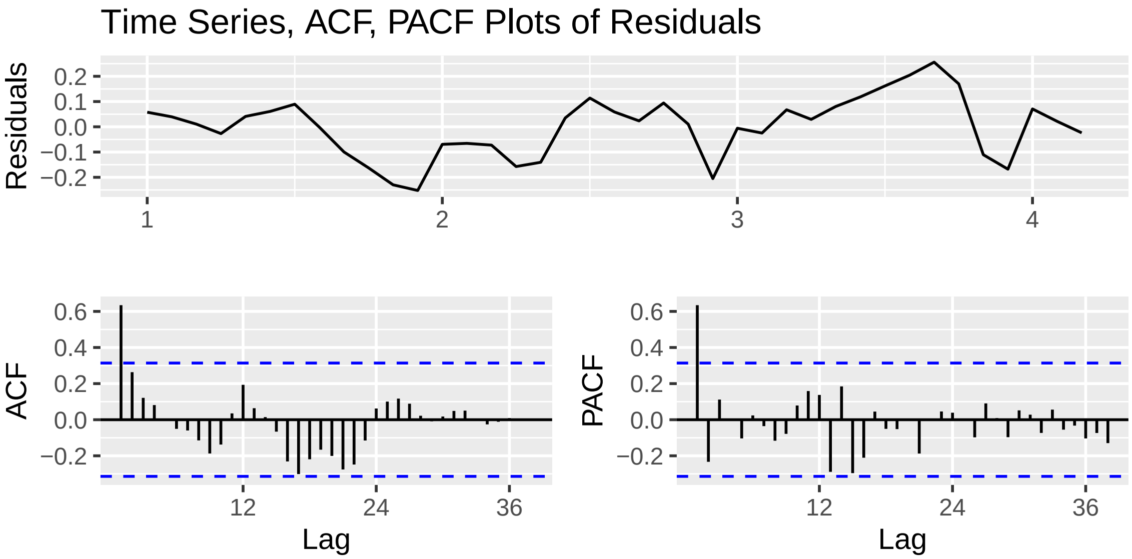

Figure 17: Time Series, ACF, PACF Plots of Residuals

Table 8: Comparing Models

p	d	q	P	D	Q	AIC	AICc	BIC
0	0	0	0	1	0	-13.009	-12.009	-11.593
1	0	0	0	1	0	-51.147	-48.965	-49.022
1	0	0	1	1	0	-51.475	-47.475	-48.643
1	0	0	0	1	1	-50.421	-46.421	-47.589
1	1	0	0	1	0	-50.518	-48.118	-48.600
1	1	0	1	1	0	-49.990	-45.545	-47.433
1	1	0	0	1	1	-49.328	-44.884	-46.772

Here we follow a similar parameter selection procedure as detailed in Section 5.2.3. The selected models and their corresponding model selection criterion values are shown in Table 8. Interestingly, the AIC, AICc, and BIC values for these models are very similar (except for the first model). The two models with comparatively smaller criterion values are regression models with SARIMA errors $(1,0,0)\times(0,1,0)_{12}$ and $(1,0,0)\times(1,1,0)_{12}$. We attempt both and compare their prediction results.

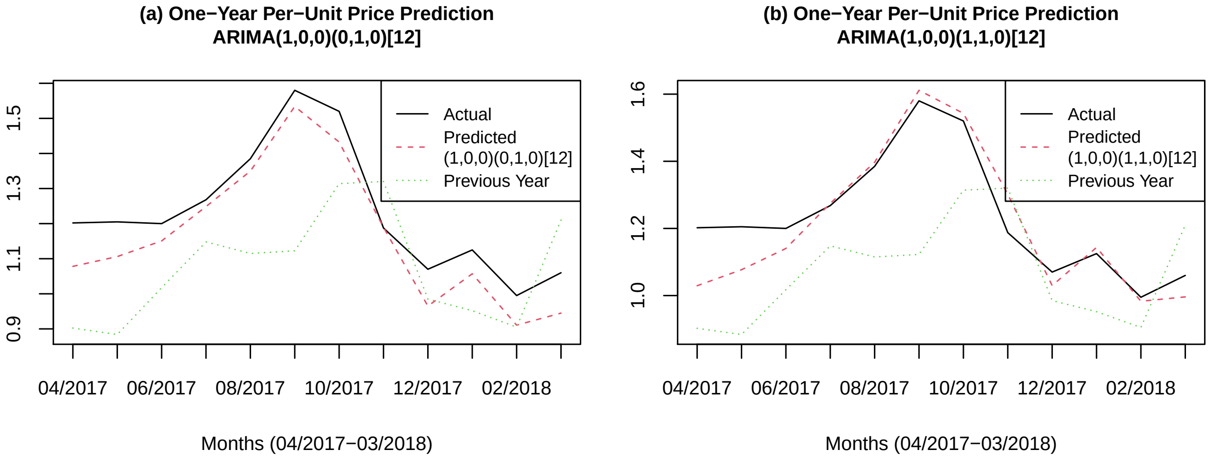

Figure 18: Predicting Avocado Per-Unit Price from April 2017 to March 2018

We notice that the prediction made by the regression with SARIMA errors $(1, 0, 0) \times (0, 1, 0)_{12}$ [Figure 18(b)] seems to be closer to the actual values than that made by regression with SARIMA errors $(1, 0, 0) \times (1, 1, 0)_{12}$ [Figure 18(a)]. Specifically, the prediction shown in Figure 18(a) seem to always under-predict the per-unit price. Nonetheless, the two predictions seem to be very similar, and both seem to be much closer to the actual values than simply using the last year's results.

Table 9: RMSE Comparison

Type	RMSE
Per-Unit Price Prediction (1,0,0)(0,1,0)[12]	0.079
Per-Unit Price Prediction (1,0,0)(1,1,0)[12]	0.077
Per-Unit Price from Previous Year	0.233

As shown in Table 9, the regression with SARIMA errors $(1, 0, 0) \times (1, 1, 0)_{12}$ indeed seem to have the smallest RMSE value, though the RMSE value of the regression with SARIMA errors $(1, 0, 0) \times (0, 1, 0)_{12}$ is very similar. Both of these models have lower RMSE values than the model in Section 3, which yields an RMSE value of 0.167. This suggests that some information about the volume or demand for avocados might potentially improve the prediction. As mentioned in Section 3.1, the problem with this approach is that we would not know the sales volume when we do not know the price. Thus, such a prediction might not be feasible in reality. However, if we have access to a plausible projection of the demand for avocados or the actual production volume data, we might use them to better predict the per-unit price.

Now we check for model assumptions. In Figure 19, we see the residual plots for the regression with SARIMA errors $(1, 0, 0) \times (1, 1, 0)_{12}$. The corresponding residual plots for the regression with SARIMA errors $(1, 0, 0) \times (0, 1, 0)_{12}$ are largely similar. In the upper panel of Figure 19 is the residuals vs. time plot. Besides the first 12 residuals being zero due to seasonal differencing, the remaining residuals seem to roughly have a mean of zero and constant variance. There seems to be no severe autocorrelation among the residuals, since the spikes in the ACF plot are all between the blue dashed lines. In addition, the residuals seem to be largely normally distributed as shown in the bottom right plot. Thus, we conclude that the regression model with SARIMA errors

$(1, 0, 0) \times (1, 1, 0)_{12}$ satisfies the model assumptions (Asamoah-Boaheng 2014).

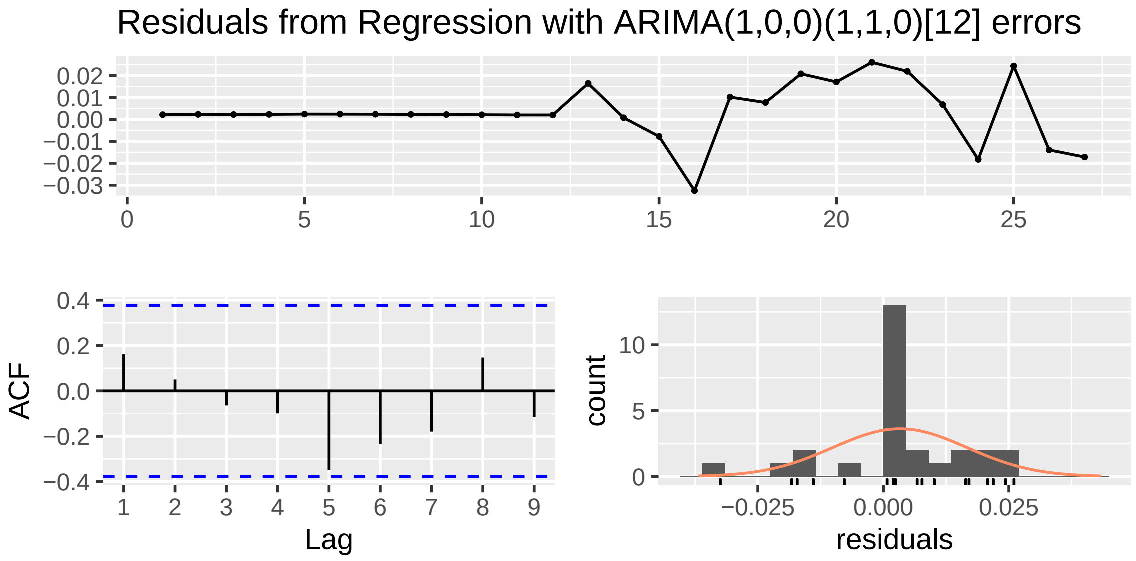

Figure 19: Residuals Check of the SARIMA Model for the Error Terms

5.3.2 Model Assumptions for the Selected Model in Section 3

We observe the residual plots for the regression with unemployment rate as the predictor and with SARIMA errors $(1,0,0) \times (0,1,0)_{12}$. In the residuals vs. time plot, besides the first 12 residuals, the rest of the residuals seem to largely have a mean of zero. Interestingly, the variance of the residuals seem to increase slightly with time, although it is difficult to conclude whether this is a major concern given that we only have 27 months to fit the model. In the ACF plot, there seem to be no severe autocorrelation as the spikes are all between the blue dashed lines. From the bottom right plot, the residuals seem to be largely normally distributed. Thus, the model seems to largely satisfy the SARIMA assumptions (Asamoah-Boaheng 2014).

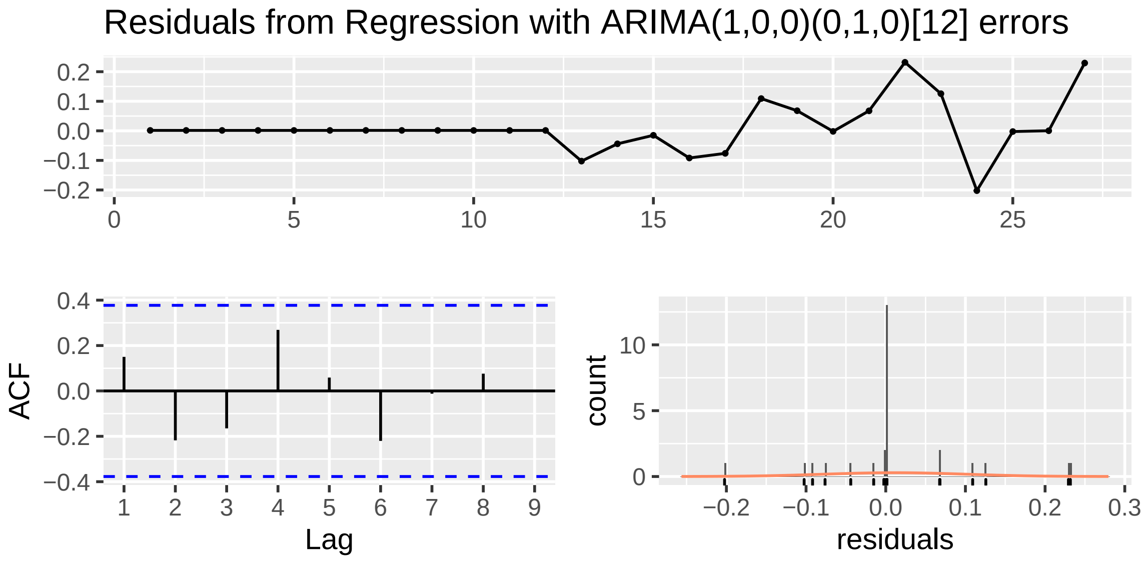

Figure 20: Residuals Check of the SARIMA Model for the Error Terms

5.4 Example of Analysis on the Weekly Version of the Data

We have conducted similar analyses on the weekly version of the avocado data. Since the external macroeconomic data are collected on a monthly basis, we did not include them as predictors. This means that, for instance, if we are predicting the weekly sales volume, we are only using the weekly per-unit price as the predictor.

Figure 21 shows the sales volume prediction from April 2017 to March 2018 using the regression with SARIMA $(1,1,1) \times (0,1,1)_{52}$. The model is specified as follows:

$$\text{Sales Volume}_t = \beta_0 + \beta_1 \text{ Per-Unit Price}_t + n_t$$

where

$$(1-\phi_1 L)(1-L^{52})(1-L)\; n_t = (1+\Theta_1 L)(1+\theta_1 L)\; a_t,\; a_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$$

.

ϕ_1 denotes the non-seasonal order-1 autoregressive parameter, θ_1 represents the non-seasonal moving average parameter, and Θ_1 denotes the seasonal order-1 moving average parameter. L is the lag operator and a_t is the white noise.

Interestingly, although the weekly data seem to be slightly noisier, the prediction is fairly close to the actual results, with an RMSE of 1.460. $1.460 is the square root of the mean of the squared differences between the actual per-unit weekly retail prices of avocados and our predicted weekly prices.

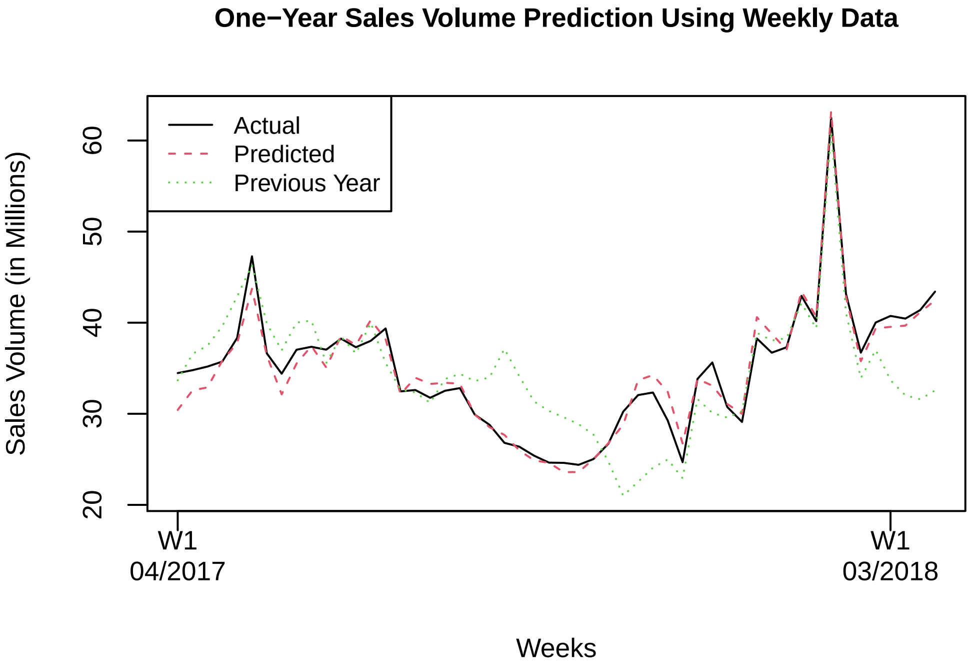

Figure 21: Predicting the Per-Unit Price of Avocados from April 2017 to March 2018

Table 10: RMSE Comparison

Type	RMSE
Sales Volume Prediction	1.460
Sales Volume from Previous Year	4.605

Bibliography

AgMRC. 2018. "Avocados." *Iowa State University*. https://www.agmrc.org/commodities-products/fruits/avocados.

Amadeo, Kimberly. 2020. "Unemployment Rate, Effect, and Trends." *The Balance: US Economy*. https://www.thebalance.com/unemployment-rate-3305744.

Asamoah-Boaheng, Michael. 2014. "Using Sarima to Forecast Monthly Mean Surface Air Temperature in the Ashanti Region of Ghana." *International Journal of Statistics and Applications*. http://article.sapub.org/10.5923.j.statistics.20140406.06.html.

BLS. 2021a. "Consumer Price Index." *U.S. Bureau of Labor Statistics*. https://www.bls.gov/cpi/.

———. 2021b. "(Unadj) Unemployment Rate." *United States Department of Labor*. https://beta.bls.gov/dataViewer/view/timeseries/LNU04000000.

Bossche, Filip Van den, Geert Wets, and Tom Brijs. 2004. "A Regression Model with Arima Errors to Investigate the Frequency and Severity of Road Traffic Accidents." *Steunpunt Verkeersveiligheid*. https://www.researchgate.net/publication/265674721_A_Regression_Model_with_ARIMA_Errors_to_Investigate_the_Frequency_and_Severity_of_Road_Traffic_Accidents.

Burnham, Kenneth P., and David R. Anderson. 2004. "Information and Likelihood Theory: A Basis for Model Selection and Inference." *Model Selection and Multi-Model Inference*. https://link.springer.com/book/10.1007/b97636.

CNNMoney. 2017. "Millionaire to Millennials: Lay Off the Avocado Toast If You Want a House." *Cable News Network*. https://money.cnn.com/2017/05/15/news/millennials-home-buying-avocado-toast/index.html.

E. E. Holmes, M. D. Scheuerell, and E. J. Ward. 2021. "4.2 Decomposition of Time Series." *Applied Time Series Analysis for Fisheries and Environmental Sciences*. https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-decomposition-of-time-series.html.

Evans, Edward A., and Sikavas Nalampang. 2009. "Forecasting Price Trends in the U.s. Avocado (Persea Americana Mill.) Market." *Food Distribution Research Society, Vol. 40(2), Pages 1-10*. https://doi.org/10.1016/S0140-6736(17)32812-X.

Fernando, Jason. 2020. "Consumer Price Index (Cpi)." *Investopedia Economics*. https://www.investopedia.com/terms/c/consumerpriceindex.asp.

FOMC. 2021. "FOMC Projections Materials, Accessible Version." *Federal Reserve*. https://www.federalreserve.gov/monetarypolicy/fomcprojtabl20210317.htm.

FRED. 2021. "Consumer Price Index: All Items for the United States." *Federal Reserve Economic Data*. https://fred.stlouisfed.org/series/USACPIALLMINMEI.

ggplot2. 2021. "Smoothed Conditional Means." *Tidyverse*. https://ggplot2.tidyverse.org/reference/geom_smooth.html.

Glen, Stephanie. 2015. "Variance Inflation Factor." *Statistics How To*. https://www.statisticshowto.com/variance-inflation-factor/.

———. 2021. "RMSE: Root Mean Square Error." *Statistics How To*. https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/.

Hayes, Adam. 2021. "What Is a Time Series?" *Investopedia Fundamental Analysis.* https://www.investopedia.com/terms/t/timeseries.asp.

Holmes, Susan. 2000. "RMS Error." *Regression Effect and Regression.* http://statweb.stanford.edu/~susan/courses/s60/split/node60.html.

Hyndman, Rob J. 2010. "The Arimax Model Muddle." *Hyndsight Blog.* https://robjhyndman.com/hyndsight/arimax/.

———. 2021. "Regression with Arima Errors." *Forecasting Using R.* https://robjhyndman.com/talks/RevolutionR/11-Dynamic-Regression.pdf.

Hyndman, Rob J, and George Athanasopoulos. 2018a. "3.3 Residual Diagnostics." *Forecasting: Principles and Practice.* https://otexts.com/fpp2/residuals.html.

———. 2018b. "3.4 Evaluating Forecast Accuracy." *Forecasting: Principles and Practice.* https://otexts.com/fpp2/accuracy.html.

———. 2018c. "White Noise." *Forecasting: Principles and Practice.* https://otexts.com/fpp3/wn.html#wn.

ICLS. 2013. "Consumer Price Index." *Organisation for Economic Co-Operation and Development (OECD).* https://stats.oecd.org/glossary/detail.asp?ID=427.

Johnson, Greg. 2020. "Avocados Rebound from Pandemic Hit." *Blue Book Services.* https://www.producebluebook.com/2020/05/07/avocados-rebound-from-pandemic-hit/.

Kiggins, Justin. 2018. "Avocado Prices." *Kaggle.* https://www.kaggle.com/neuromusic/avocado-prices.

Lamstein, Ari. 2021. "Mapping Census Bureau Data in R with Choroplethr." *United States Census Bureau.* https://www.census.gov/data/academy/courses/choroplethr.html.

Nau, Robert. 2020. "Introduction to Arima." *ARIMA Models for Time Series Forecasting.* https://people.duke.edu/~rnau/411arim.htm.

Perez, Marvin G. 2020. "Avocados Are the 'Pandemic-Proof' Crop in Lockdown Health Craze." *Bloomberg.* https://www.bloomberg.com/news/articles/2020-12-04/avocados-are-the-pandemic-proof-crop-in-lockdown-health-craze/.

Picardo, Elvis. 2020. "How the Unemployment Rate Affects Everybody." *Investopedia Macroeconomics.* https://www.investopedia.com/articles/economics/10/unemployment-rate-get-real.asp.

———. 2021. "How Inflation and Unemployment Are Related." *Investopedia Macroeconomics.* https://www.investopedia.com/articles/markets/081515/how-inflation-and-unemployment-are-related.asp.

Prabhakaran, Selva. 2021. "ARIMA Model – Complete Guide to Time Series Forecasting in Python." *Machine Learning Plus.* https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python.

RDocumentation. 2021. "Decompose: Classical Seasonal Decomposition." *Stats (Version 3.6.2).* https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/decompose.

Rundel, Colin. 2017a. "Lecture 11 - Seasonal Arima." *STA 444/644.* http://www2.stat.duke.edu/~cr173/Sta444_Sp17/slides/Lec11.pdf.

———. 2017b. *STA 444/644.* http://www2.stat.duke.edu/~cr173/Sta444_Sp17/slides/Lec9.pdf.

———. 2018a. "Lecture 6 - Discrete Time Series." *STA 444/644*. http://www2.stat.duke.edu/~cr173/Sta444_Fa18/slides/Lec06/Lec06.pdf.

———. 2018b. "Lecture 8 - Ar, Ma, and Arma Models." *STA 444/644*. http://www2.stat.duke.edu/~cr173/Sta444_Fa18/slides/Lec08/Lec08.pdf.

SAS. 2021. "Notation for Arima Models." *SAS/ETS(R) 9.3 User's Guide*. https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_tffordet_sect016.htm.

Schwartz, Elaine. 2019. "An Avocado Update." *Blue Book Services*. https://econlife.com/2019/06/avocado-prices-2/.

Simon, Laura, and Robert Heckard. 2021a. "2.2 Partial Autocorrelation Function (Pacf)." *Penn State STAT 501 Regression Methods*. https://online.stat.psu.edu/stat510/lesson/2/2.2.

———. 2021b. "Lesson 8: Regression with Arima Errors." *Penn State STAT 501 Regression Methods*. https://online.stat.psu.edu/stat510/book/export/html/669.

Tackett, Maria. 2019. "14 Model Selection." *STA 210*. https://www2.stat.duke.edu/courses/Fall19/sta210.001/slides/lec-slides/14-model-selection.pdf.

USAID. 2014. "The Us Market for Avocado." *ACCESO*. https://pdf.usaid.gov/pdf_docs/PA00KP28.pdf.

Wong, Ted. 2018. "Inpretreting the Root Mean Square Error (Rmse)." *DataScience Stack-Exchange*. https://datascience.stackexchange.com/questions/36945/interpreting-the-root-mean-squared-error-rmse.

Wonsuk Yoo, Sejong Bae, Robert Mayberry, and Jr. James W. Lillard. 2015. "A Study of Effects of Multicollinearity in the Multivariable Analysis." *International Journal of Applied Science and Technology*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318006/.