

**TABLE OF CONTENTS**

[**1. Introduction** 3](#_Toc70024688)

[**2. Data** 3](#_Toc70024689)

[**2.1. Original Data Set** 3](#_Toc70024690)

[**2.1.1. Data Cleansing** 4](#_Toc70024691)

[**2.2. Data Patterns** 4](#_Toc70024692)

[**3. Methodology** 6](#_Toc70024693)

[**3.1. Forecast Methods** 6](#_Toc70024694)

[**3.1.1. Forecast Horizon** 6](#_Toc70024695)

[**3.1.2. Naïve** 6](#_Toc70024696)

[**3.1.3. Simple average** 6](#_Toc70024697)

[**3.1.4. Moving averages** 6](#_Toc70024698)

[**3.1.5. Holt’s linear exponential smoothing** 7](#_Toc70024699)

[**3.1.6. Multiple regression** 8](#_Toc70024700)

[**3.1.7. Box-Jenkins Model (ARIMA)** 8](#_Toc70024701)

[**3.2. Measurements of forecasting accuracy as a criteria to select the appropriate method** 9](#_Toc70024702)

[**3.2.1. MAD** 9](#_Toc70024703)

[**3.2.2. RMSE** 9](#_Toc70024704)

[**3.2.3. MAPE** 9](#_Toc70024705)

[**4. Analysis** 10](#_Toc70024706)

[**4.1. Naïve** 10](#_Toc70024707)

[**4.2. Simple average** 10](#_Toc70024708)

[**4.3. Moving averages** 12](#_Toc70024709)

[**4.3.1. Three terms in the average** 12](#_Toc70024710)

[**4.3.2. Five terms in the average** 13](#_Toc70024711)

[**4.4. Holt’s linear exponential smoothing** 14](#_Toc70024712)

[**4.5. Multiple regression** 16](#_Toc70024713)

[**4.6. Box-Jenkins Model (ARIMA)** 17](#_Toc70024714)

[**5. Summary** 20](#_Toc70024715)

[**6. Forecasting and conclusion** 21](#_Toc70024716)

[**7. Reference** 22](#_Toc70024717)

[**8. Appendix** 23](#_Toc70024718)

# **1. Introduction**

The goal of the project is to predict the future sales of avocados, find the seasonality, the trend, random variations, and make our predictions based on them. By finding the following, we will be able to help the agricultural industry by providing them accurate forecasts of the fluctuations in avocado demand, which will provide them recommendations on when to harvest more avocados. The demand of avocados affects the price, which in turn would affect the food and restoration industry.

# **2. Data**

## **2.1. Original Data Set**

The original dataset, provided by the Hass Avocado Board from kaggle.com, contains weekly 2018 retail scan data for National retail volume (units) and price.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Index | Sample observed within a given region |
| Date | Observation’s date |
| AveragePrice | Average price of Avocado |
| Type | Conventional or Organic |
| Total\_Volume | Total avocado sold on a given date |
| Region | City or region the avocado was sold |
| 4046 | Small/Medium sized avocado |
| 4225 | Large avocado |
| 4770 | Extra large avocado |
| Total\_Bags | Total bag purchased |
| Small\_Bags | Small bags |
| Large\_Bags | Large bags |
| XLarge\_Bags | Extra large bag |
| Year | Observation’s year |

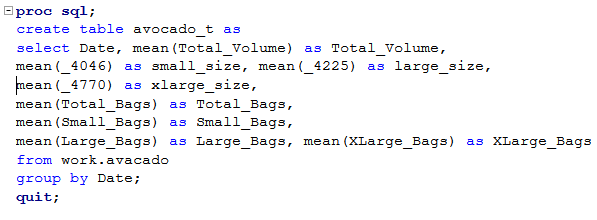
### **2.1.1. Data Cleansing**

In order to cleanse our data, we used the Impute node on SAS Enterprise Miner to make sure that our dataset did not have any missing values. We used the input “median” to input our interval variables. Our dataset did not contain any class variables.



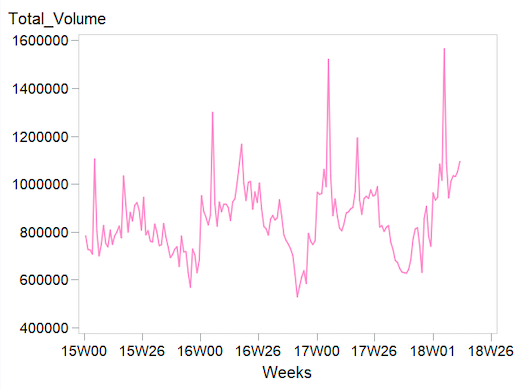
Through Stepwise Selection, we identified that the variables Type, Region, and Year were not useful in predicting the target variable TotalVolume. Therefore, we went ahead and removed these 3 variables from our model. The variables remaining in the model are: Date, AveragePrice, Total\_Volume, 4046(Small/Medium sized avocado), 4225(Large avocado), 4770(Extra large avocado), Total\_Bags, Small\_Bags, Large\_Bags, as well as XLarge\_Bags.

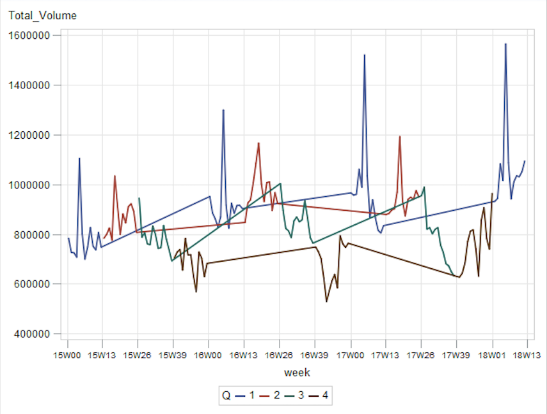
As our dataset contained many observations per singular calendar date, we had to transform the data using PROC SQL and merge all same date observations into one observation. By doing so, we found the mean of the demand of avocado per state, which created the demand for avocados in America.



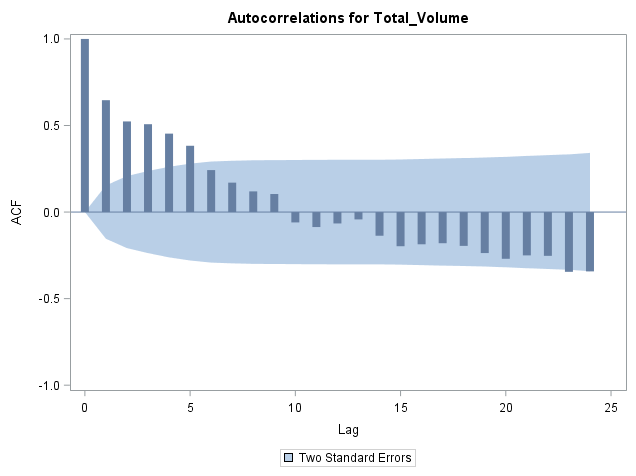
## **2.2. Data Patterns**

Exploring the data in SAS EG exposed patterns of trend and seasonality over time. Peak periods of total avocado sold happens in Q1 while the lowest occurs during Q4. According to the Weaver Street Market, peak season for avocados happens in January through March. It is during this time that they have developed higher oil content, resulting in better flavor and texture, which explains the results in the graph. It can be observed that sales have increased in level over time and annual seasonality exists.

Time plot: Seasonal Plot:



ACF plot:



# **3. Methodology**

## **3.1. Forecast Methods**

### **3.1.1. Forecast Horizon**

As the dataset contains approximately 3 years of observations, we only have enough observations to properly forecast one year, or 52 weeks, into the future. Trying to forecast anything beyond 52 weeks would result in a SAS error expressing the lack of observations.

### **3.1.2. Naïve**

Naive Forecasting is based on the most recent data available. It sets all forecasts to be the value of the last observation. A forecast that is equal to the latest observation within a time series. In other words, the results from last month will be the same results for this month. The equation of this methods is yT+h|T=yT. The major issue of this method is that it does not consider trend and/or seasonality. However, equations such as the seasonal naïve method and the drift method are variations of the naïve method that allow the forecast to follow high seasonality and to increase or decrease over time, respectively.

### **3.1.3. Simple average**

Simple average is a basic forecasting technique that uses the mean of all relevant past

observations to forecast the next period. All the future values are equal to the average of the historical data. The forecast calculation is written as yT+h|T=¯y=(y1+⋯+yT)/T. This technique is useful for long-range forecasting purposes when the setting is stabilized and unchanged. Using this technique on this dataset may not result in an adequate forecast as the range is short.

### **3.1.4. Moving averages**

#### **3.1.4.1. Moving Average Order 3**

Similar to simple averages, a moving average is a statistical forecasting technique that uses past observations to predict future data. In the case of a moving average of order 3, it will take the average of the three previous observations to predict the subsequent observation. This forecasting technique is very sensitive, as a spike in either of the 3 past observations could substantially affect the forecasted value.

#### **3.1.4.2. Moving Average Order 5**

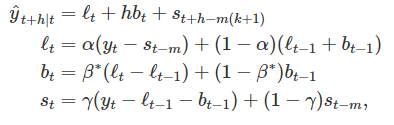
Moving average with order 5 is a more smoothed out version of moving average with order 3. Instead of accounting for the 3 previous observations, 5-MA incorporates the 5 previous observations in its arithmetic. As the order of the moving average impacts the smoothness of the trend-cycle, a larger order will result in a smoother trend line.

### **3.1.5. Holt’s linear exponential smoothing**

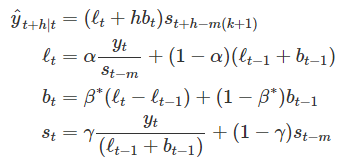
Holt’s linear model is similar to Holt-Winters’ method in regard to trend patterns of a time

series, but it also captures seasonality. This technique “comprises the forecast equation and three smoothing equations — one for the level ℓtℓt, one for the trend btbt, and one for the seasonal component st, with corresponding smoothing parameters α, β∗ and γ (Hyndman & Athanasopoulos, 2018). This method can be performed with two variations that effect the seasonal component. The multiplicative method expresses the seasonal component in percentages. This variant is preferred when the seasonal variations are fluctuating with the level of the time series. On the other hand, the additive method is expressed in absolute terms in the scale of the time series. The series is “seasonally adjusted by dividing... by the seasonal component,” so that “within each year the seasonal component adds up to m.” (Hyndman & Athanasopoulos, 2018). The following equations form the two variations:

Additive method:



Multiplicative method:



### **3.1.6. Multiple regression**

A multiple regression is a linear model that contains two or more predictor variables. The equation for a multiple regression is yt=β0+β1x1,t+β2x2,t+⋯+βkxk,t+εt . Y is the variable to be forecasted, and x1, x2, …, xk are k predictor variables. The βk coefficients measure the effect of each predictor variable after accounting for the effects of all other predictor variables in the model. We must note that there are certain assumptions that must be fulfilled in order to create a proper multiple regression model. First, the forecast variable and the predictor variables represent a reasonable representation of reality. And second, the following assumptions about the error terms must be met: (1) error terms have mean zero, (2) error terms are not autocorrelated. (3) error terms are unrelated to the predictor variables, (4) error terms are normally distributed with constant variance σ2. Multiple regression is a commonly used forecasting technique as it is pretty straight forward; once you create the optimal linear equation, you can forecast different points with a combination of different values.

### **3.1.7. Box-Jenkins Model (ARIMA)**

Autoregressive Integrated Moving Average (ARIMA) is the last method that we will be using for this case. This model can analyze both stationary and nonstationary time series data. ARIMA runs multiple models simultaneously and combines the results in a single model by using the existing data and information within the time series. ARIMA’s model repeat the procedure and run new models until an appropriate model is found whenever when a specific model does not fit. Autocorrelation is an important factor to select the appropriate model. Two types of models exist: the autoregressive model which requires an autoregressive term (AR), and the moving average model, which requires a moving average term (MA). The autoregression model forecast variables of interest using a linear combination of past values of the variable. The model can be written as: yt=c+ϕ1yt−1+ϕ2yt−2+⋯+ϕpyt−p+εt. Whereas the moving average model uses past forecast errors in type of regression model. The model can be written as: yt=c+εt+θ1εt−1+θ2εt−2+⋯+θqεt−q.

## **3.2. Measurements of forecasting accuracy as a criteria to select the appropriate method**

### **3.2.1. MAD**

Mean Absolute Deviation (MAD) is the average of the sum of the mean absolute errors of the

model. This measure is used to describe the variability in the data and identify the spreading of the values. A high MAD indicates high variability and dispersion in the data, a low MAD is the inverse. This measurement is not sensitive to very large errors.

### **3.2.2. RMSE**

Root Mean Squared Error (RMSE) is the square root of the Mean Squared Error (MSE), denoted

by the equation:  . The square root of the average of the sum of the squared errors. This measure is very sensitive to very large errors. If possible, the value of this measure should be as close to 0 as possible indicating little to no error and 100% model predictive accuracy. Since this is almost impossible, aiming for the lowest should be the goal. The advantage of using RMSE over MSE is that the measure is on the same scale as the data itself.

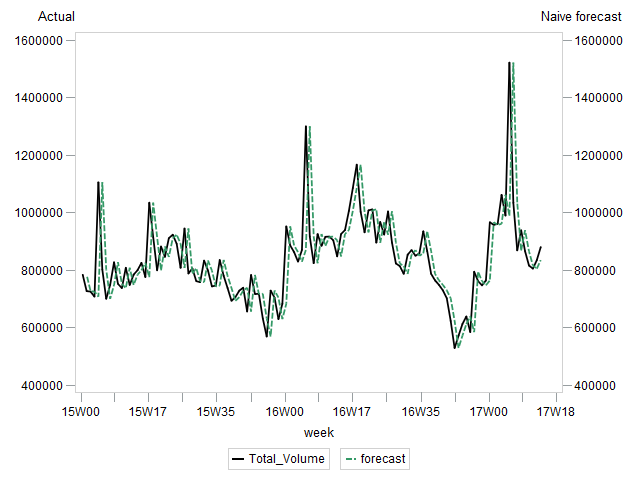
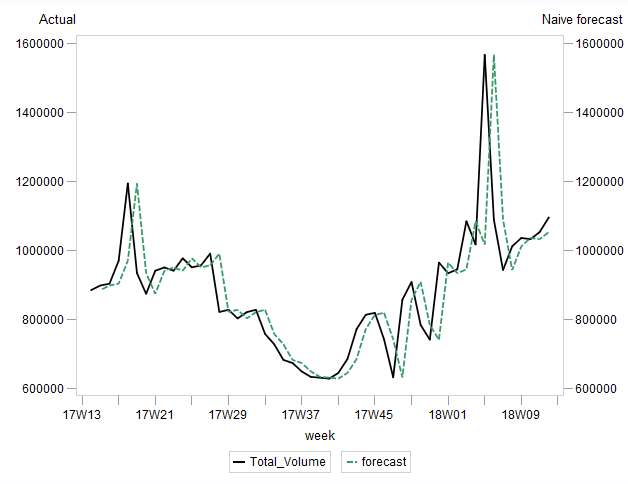
### **3.2.3. MAPE**

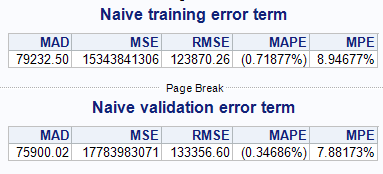
Mean Absolute Percentage Error (MAPE) is one of the most commonly used measure to compare forecast performance between data sets. It’s a measure that look at the error as a percentage of the value rather than an amount. The equation is: . All the errors are summed to evaluate the overall model error. An advantage of this measure is it measure both the magnitude as a percentage as well as the direction of the error. However, it penalizes positive errors more and is no symmetric in under and over-forecasting.

# **4. Analysis**

## **4.1. Naïve**

The graphs below show that the naïve approach follows closely the actual data. However, there is a consistent underestimation and overestimation in both sets, over time. Looking at the accuracy measures of the training and validation set, we can see that the validation results are generally lower than the training set, but still showing high measures suggesting this model may not be appropriate for forecasting purposes. Still, this method will be used as a benchmark going further.

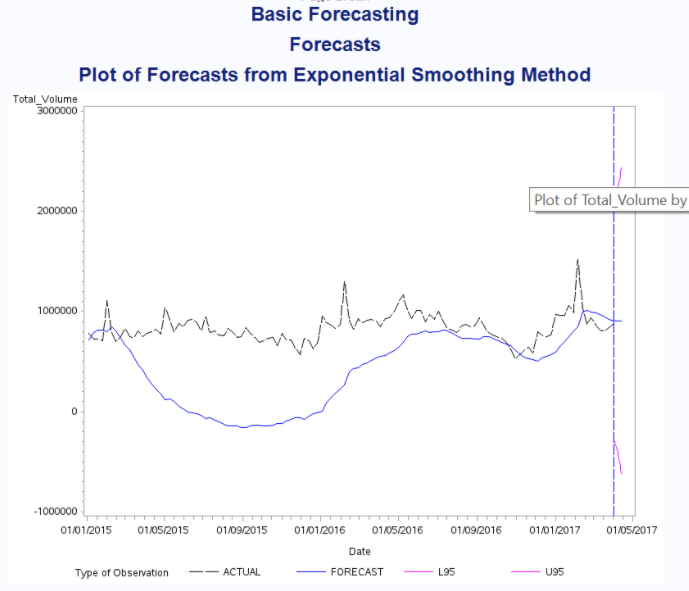
Training set Validation set



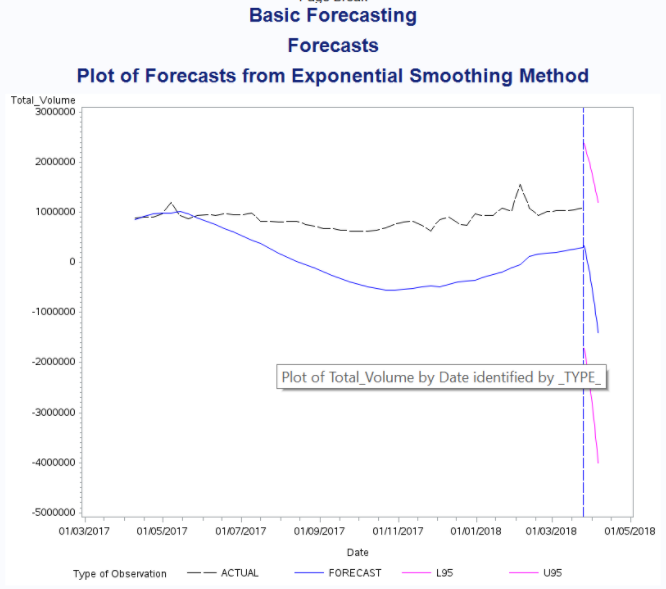
## **4.2. Simple average**

In both the training and validation sets, comparing actual and forecasted values in the graphs shows that the simple average method does not forecast accurately. The forecasted values underestimate over a large period and did not help reveal the level and trend of the series. Mid-2016 through mid-2017 shows a better fit but drastic underestimation is shown in the following periods. Apart from the MAPE measurement, the MAD and RMSE values improved compared to the naïve method but the validation set is showing very high results compared to the training set. These accuracy measures are still relatively high and indicate that this model is not a good fit for forecasting purposes.

*Training set*

*Validation Set*

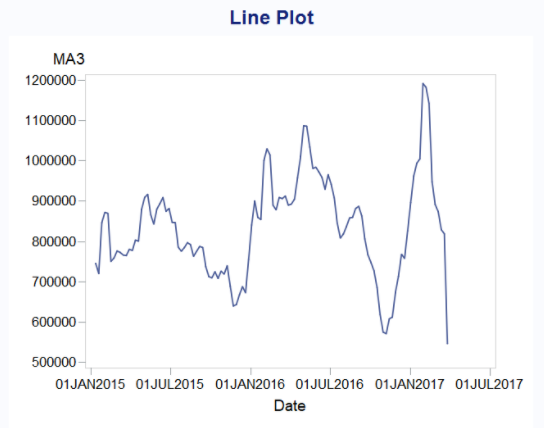
 

## **4.3. Moving averages**

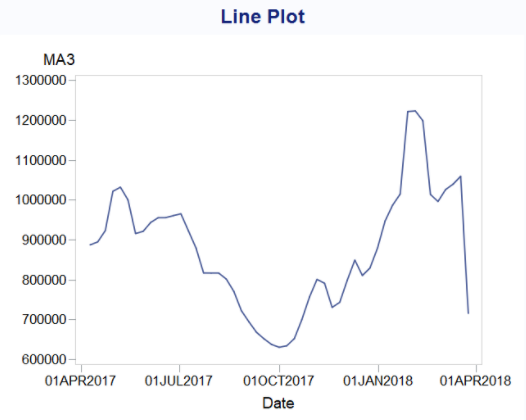
### **4.3.1. Three terms in the average**

Results from the moving averages seem to be encouraging at first sight. The trailing MA graph shows a smoothing behaviour very similar to the actual data. It captures the main movement of the time series without the abrupt fluctuations and helps to visualize the trends in the time periods. When examining the accuracy measures, only the MAPE improved as it decreased by 3.71% in the training set and decreased by 16.53% in the validation set. This improvement is negated by the strong increase of the MAD and RMSE measures. This model may seem like an improvement, but the errors are still very high.

Training Set:

Validation set:

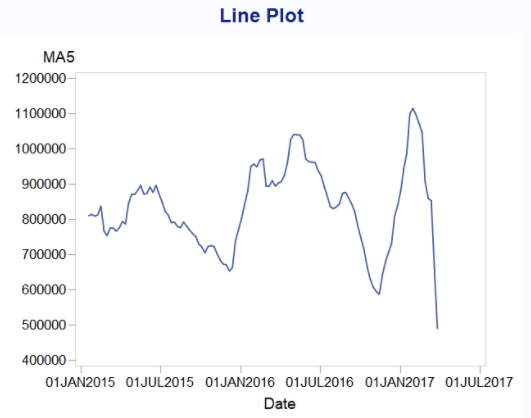




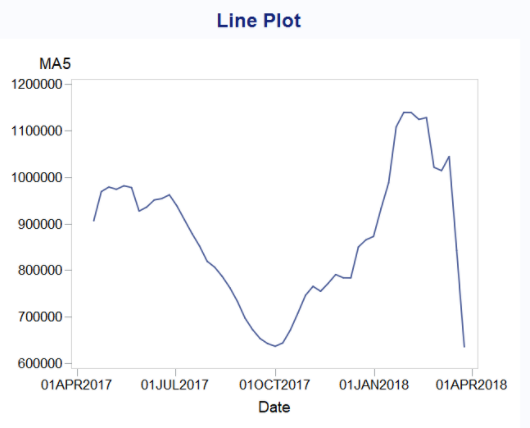
### **4.3.2. Five terms in the average**

As a larger order was performed, the curve got smoother. Improvement was made in terms of the accuracy measure as all the measurements has decreased, making this method as the most accurate up to now.

Training set:

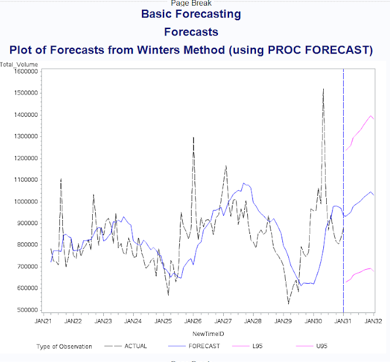


Validation set:



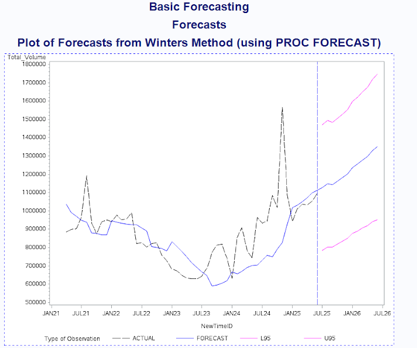
## **4.4. Holt’s linear exponential smoothing**

As the time-series exhibits both trend and seasonality, Holt’s linear method was used as an advanced smoothing technique. There are two variations of Holt-Winter’s Method which are additive and multiplicative. Here, the multiplicative method is more appropriate because the seasonal variations change proportionally to the series. They become bigger and bigger; therefore Holt-Winter’s Multiplicative Method was performed. Accuracy measures recorded are the lowest compared to the previous models which may suggest this may be a better model for the medium-term forecast. The analysis of the graphs shows that this model follows the data, but peak seasonality fluctuations are not as high as the actual data. Moreover, the difference between the training set measures and the validation set is lower compared with the previous models.

Training Set:



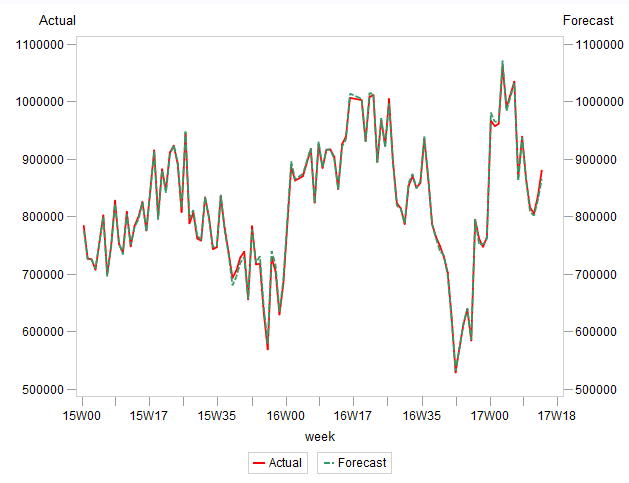
Validation set:





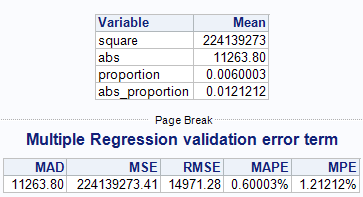
## **4.5. Multiple regression**

In order to create a regression model, we first had to determine if any of our variables were correlated. After doing a Pearson Correlation analysis, we determined that we needed to remove the variables Small\_Bags, Large\_Bags, XLarge\_Bags because it causes multiple correliality issues with Total\_Bags. We applied a log to the remaining variables to remove extreme values and make the model a better fit. This model has lower accuracy measures than Holt’s Linear Exponential Smoothing, making it the best forecasting model thus far.

Training set

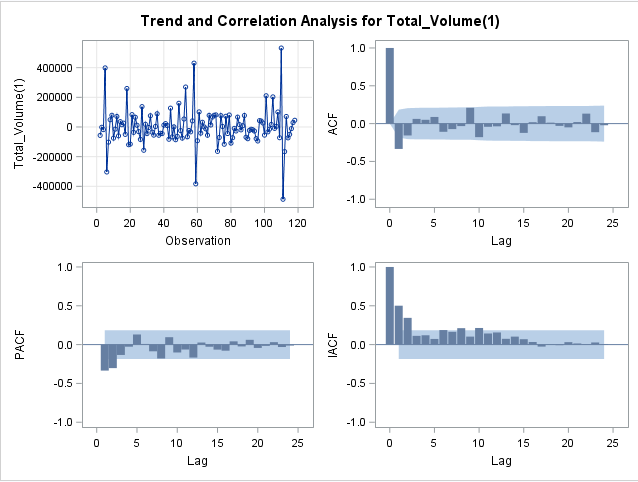


Validation set



## **4.6. Box-Jenkins Model (ARIMA)**

For the last model, the ACF plot shows an exponential decay, it indicates that the time series is not stationary.  Thus, we must run a differentiation of order 1, to convert the time series to stationary. And, as the PACF plot shows moderate negative autocorrelation, we may apply MA(1).

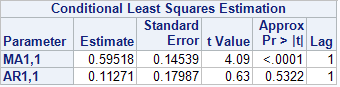
  
Applying differencing of order (1) fix the ACF Plot. Now, the time series is stationary.

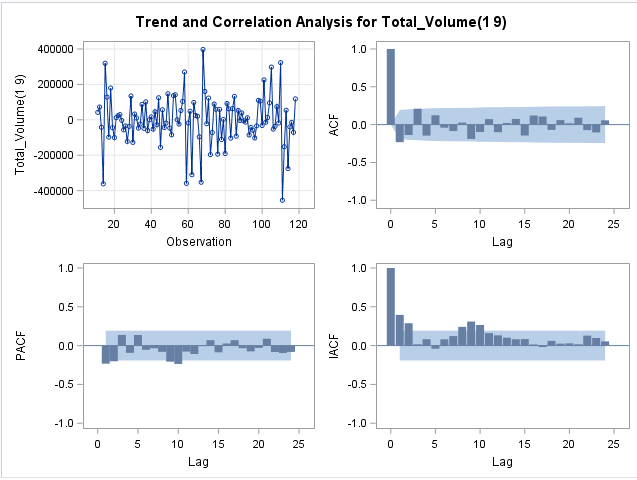
Then, we apply MA(1) on the ARIMA model. MA(1) at lag1 is significant.

ACF Plot demonstrates **weak autocorrelation.** We try to apply AR(1) to the ARIMA model.

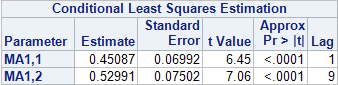
Conditional Least Square estimation indicates the AR(1) has an insignificant effect on the model.

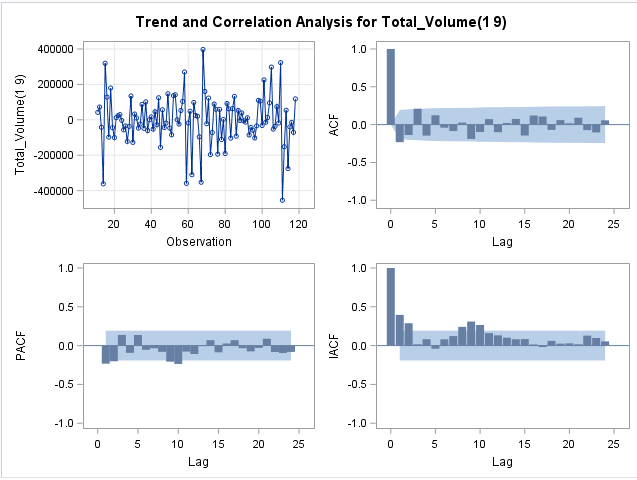
So the fitted ARIMA (0,1,1)

Now, we use ARIMA (0,1,1) on the train data to plot the non-seasonal component.

Suppose we observed ACF plot and PACF plot, excluding a spike a lag(1), both plots experience a significant spike at lag(9). It indicates a seasonal pattern is observed at lag(9).  As a result, we apply a seasonal component to the ARIMA forecast. We apply differentiation at lag(9). The ARIMA adjusts to ARIMA(0,1,1)(0,1,0)9.

In the PACF plot there is **moderate negative autocorrelation.** Thus, we apply MA(1,1)9 on the ARIMA model.

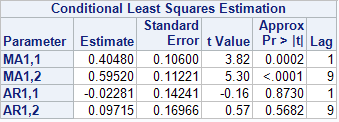
Using the Conditional Least Square estimation, we confirm that MA(1)9  is significant on the model.

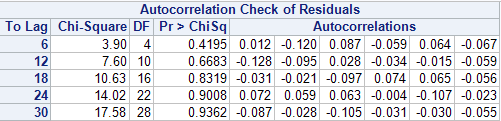
Then, observing the ARIMA (0,1,1)(0,1,0)9’s  ACF plot, we found a **weak autocorrelation**. We apply AR(1)9 to observe the significance effect on the model.

Nonetheless, after application, AR(1)9 remains insignificant.

So the fitted model is ARIMA(0,1,1)(0,1,1)9.

Now, we use ARIMA (0,1,1)(0,1,1)9 on the train data to plot the seasonal components.





Residual Diagnostic of the ARIMA

Using the Autocorrelation Check of Residual, we confirm there is no autocorrelation in the residual at lag(n) as P-value is insignificant.

**Hypothesis:**

Ho: Residual does not have autocorrelation

Ha: Residual has autocorrelation

**Decision rule:**

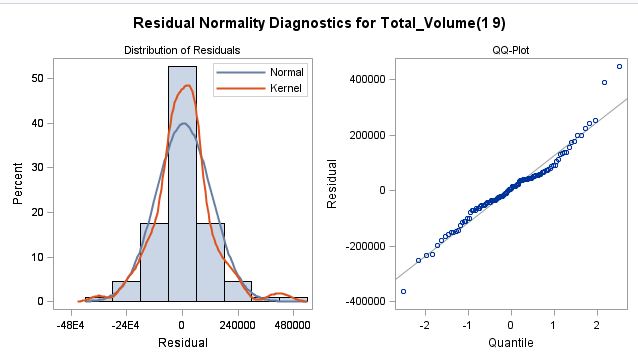
Reject Ho if:  p-value at lag(n) < = 0.05

Do not reject if Ho if:  :  p-value at lag(n) > = 0.05

**P-value:**  lag(6) p-value = 0.4195; lag(12) p-value = 0.6683; lag(18) p-value = 0.8319; lag(24) p-value = 0.9008; lag(30) p-value = 0.9008

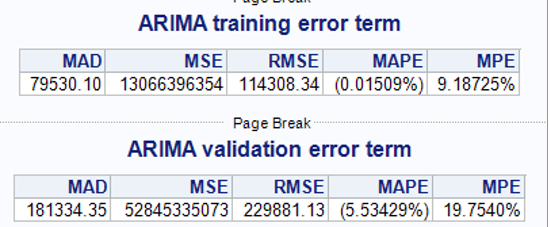
**Decision:** Do not reject Ho   p-value at lag(6) , lag (12), lag(18), lag(24), lag(30) > = 0.05

**Conclusion:** P-values within lag(n) are not significant. Thus, the time series residual does not have any autocorrelation.



Residual normality**:**

Using the Residual Normality Diagnostic, we confirm the residual assume a normal bell-curve.

Ultimaltely, when looking at the accuray measures, We can conclude that this model won’t be selected. As the Multiple regression (MR) is showing much better values, the ARIMA model actualy places itself outside of the top three models. Having the second largest MAPE, and MAPE and RMSE values much higher than MR, this model is not adequate to pursue our analysis.

# **5. Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Forecast Method** | **MAD** | **MAPE** | **RMSE** |
| **Naïve** | 75900.02 | 0.34686% | 133356.60 |
| **Simple Average** | 24972.85 | 6.67% | 35107.64 |
| **Moving average (3)** | 40235.27 | 4.68% | 70618.02 |
| **Moving average (5)** | 31006.34 | 3.72% | 52575.07 |
| **Holt-Winter’s** | 24019.72 | 2.78% | 32150.34 |
| **Multiple Linear Regression** | 11263.80 | 0.60003% | 14971.28 |
| **ARIMA** | 181334.35 | 5.534% | 229881.13 |

# **6. Forecasting and conclusion**

In conclusion, the model for consideration with the best accuracy measure is the Multiple linear regression model. This model had the lowest errors based on MAD, MAPE, and RMSE among all models. Indeed, through process of elimination, we can conclude that the business problem revolving around the sales of avocados can be answered using multiple regression. In order to use the other time series forecasting models, we would need a larger sample size and a longer time horizon. Given the sample size, we cannot accurately determine seasonal or trend pattern in the time series forecast, so we confirm that avocado demand is not dependent on past data.

# **7. Reference**

Kiggins, J. (2018, June 06). *Avocado Prices*. Retrieved from https://www.kaggle.com/neuromusic/avocadoprices?fbclid=IwAR0nPspy\_aWItqKz9d5wPOivdYsKrovAx9jdyHbPYX40tPMuLHJs4i kfQ9w

Fresh Avocado (2021). *How to identify Hass Avocados*. Retrieved from https://loveonetoday.com/how-to/identify-hass-avocados/

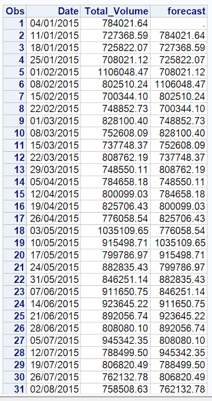
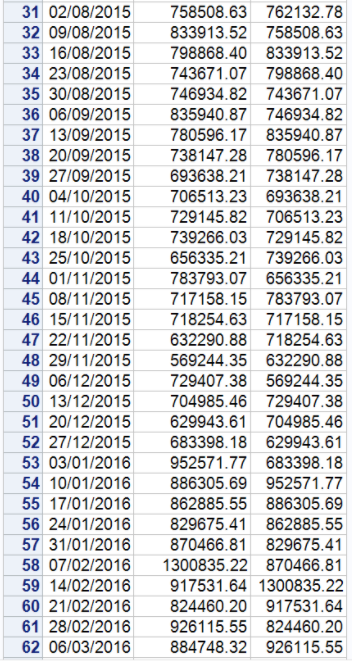
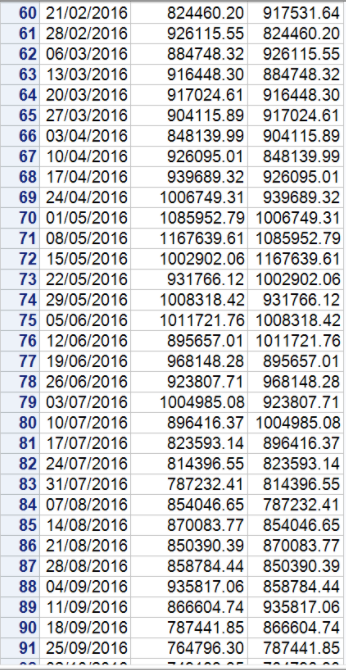
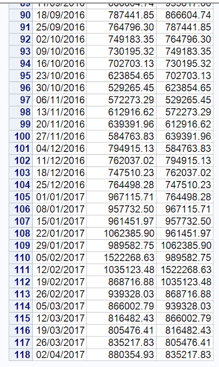
Weaver Street Market (2018). Peak Season for Avocados. Retrieved from https://www.weaverstreetmarket.coop/peak-season-for-avocados/#:~:text=Avocados%20are%20available%20year%20round,texture%20that%20we%20all%20love.

Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: principles and practice, 2nd edition,

OTexts: Melbourne, Australia. Retrieved from: OTexts.com/fpp2.

# **8. Appendix**

**a) APPENDIX 1: NAIVE TABLES**

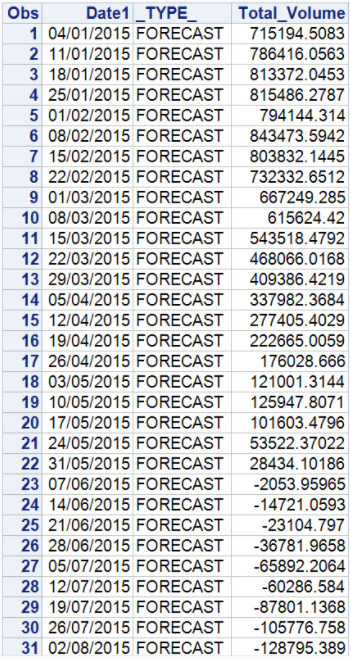
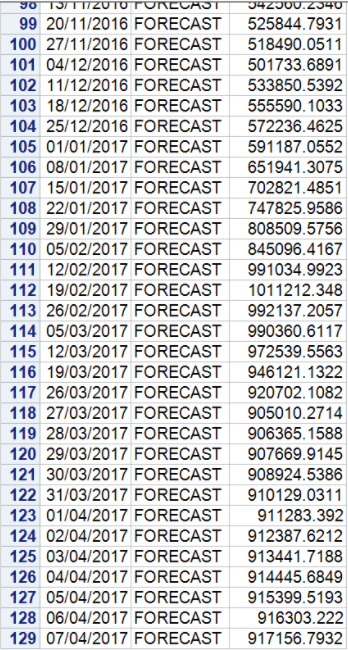
Training Set

Validation set:

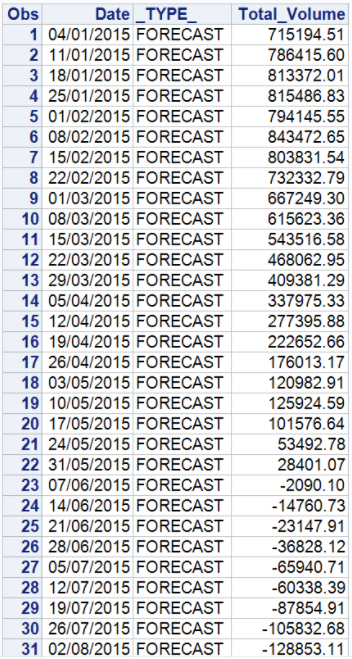


**b) APPENDIX 2: SIMPLE AVERAGE TABLES**

Training set :

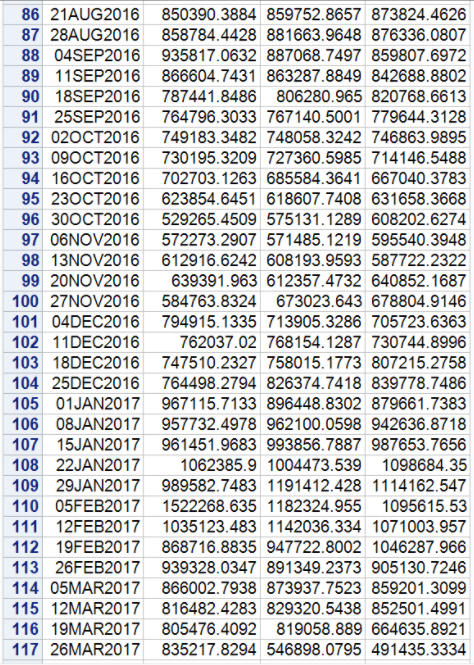
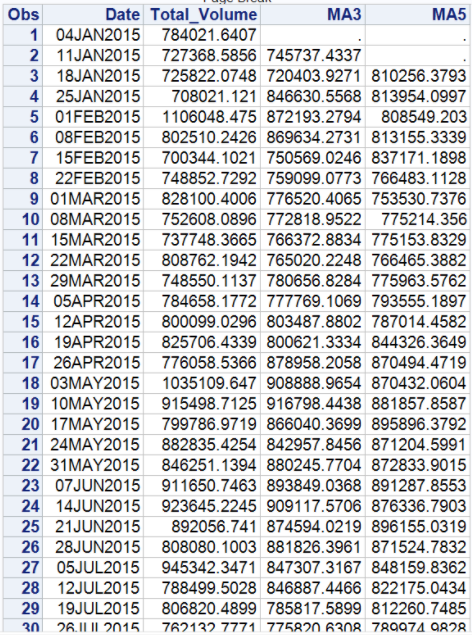
 

Validation set :

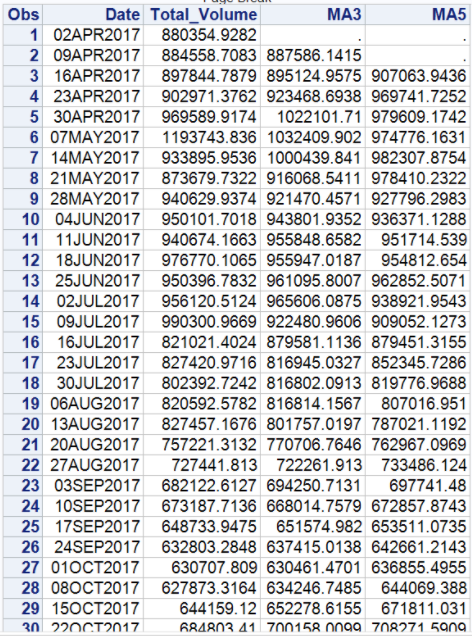
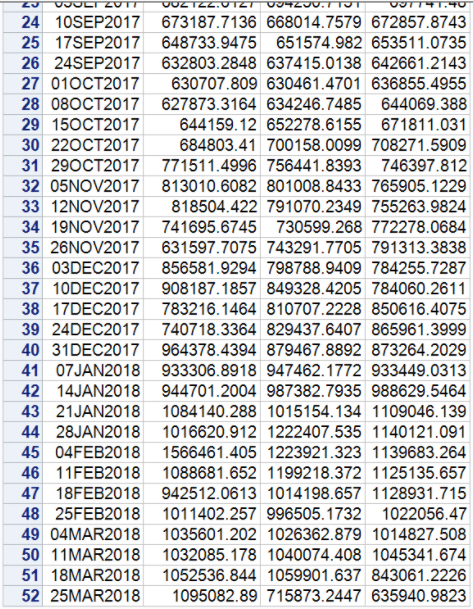
 

**c) APPENDIX 3: TRAILING MOVING AVERAGE TABLES**

Training set

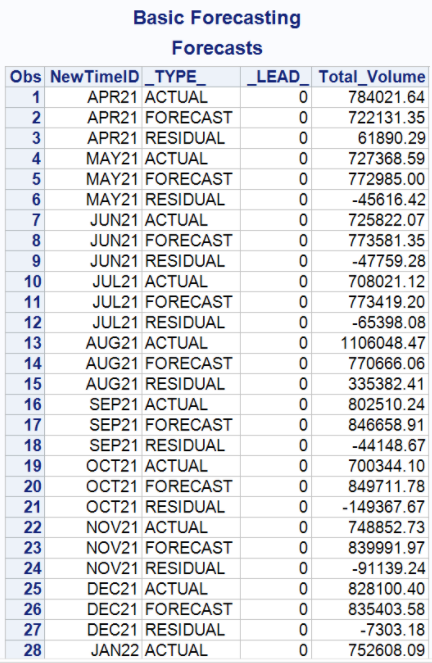


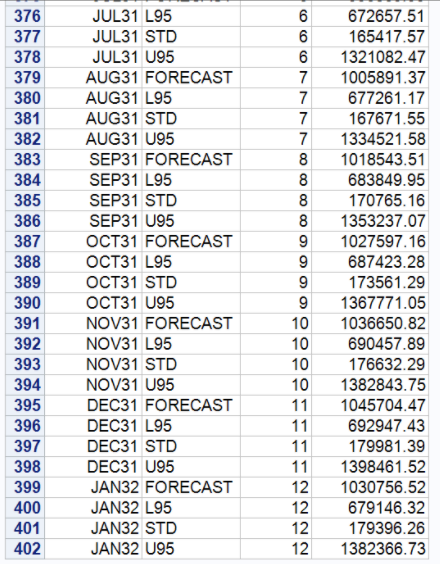
Validation set:

**d) APPENDIX 4: HOLT-WINTER’S TABLES**

Training set:





Validation set:

