Retail Sales Data Forecasting Models

Dubstech Datathon 2020

Alexander Van Roijen, Ashley Batchelor

Introduction

We created models to forecast sales for 12 weeks (~3 months) into the future for a small grocery store in Australia. The models included total sales, and sales for the top five selling categories, based on average annual sales. The top five categories included Tomatoes, Citrus, Apples, Potatoes, and "Other Vegies." We aggregated total sales for each category (or the grand total sales) over weekly periods.

Seasonal decomposition of each of the sales data sets showed that the total sales and the sales for each category showed a downward trendline. Sales are declining. The was a distinct annual seasonal component, which we approximated as 52 weeks.

We used the auto_arima feature from the Python pmdarima library to determine the best parameters for an ARIMA type of forecasting model. We used the last year of data as test data and the remainder as training data. We determined the best model was a SARIMA(0,1,1)x(1,0,1,52). We created one exogenous feature using Australian national holidays by labelling a count of the number of holidays in a week. This gave an improvement in each model. For example, our total sales had a RMSE value of 1551.00 without the exogenous variable, which improved to 1473.68 with the exogenous variable. Our final model was then a SARIMAX(0,1,1)x(1,0,1,52) with holidays as an exogenous factor.

All of the code used to create the models may be accessed in the github repository below.

https://github.com/aibatchelor22/dubstech-2020

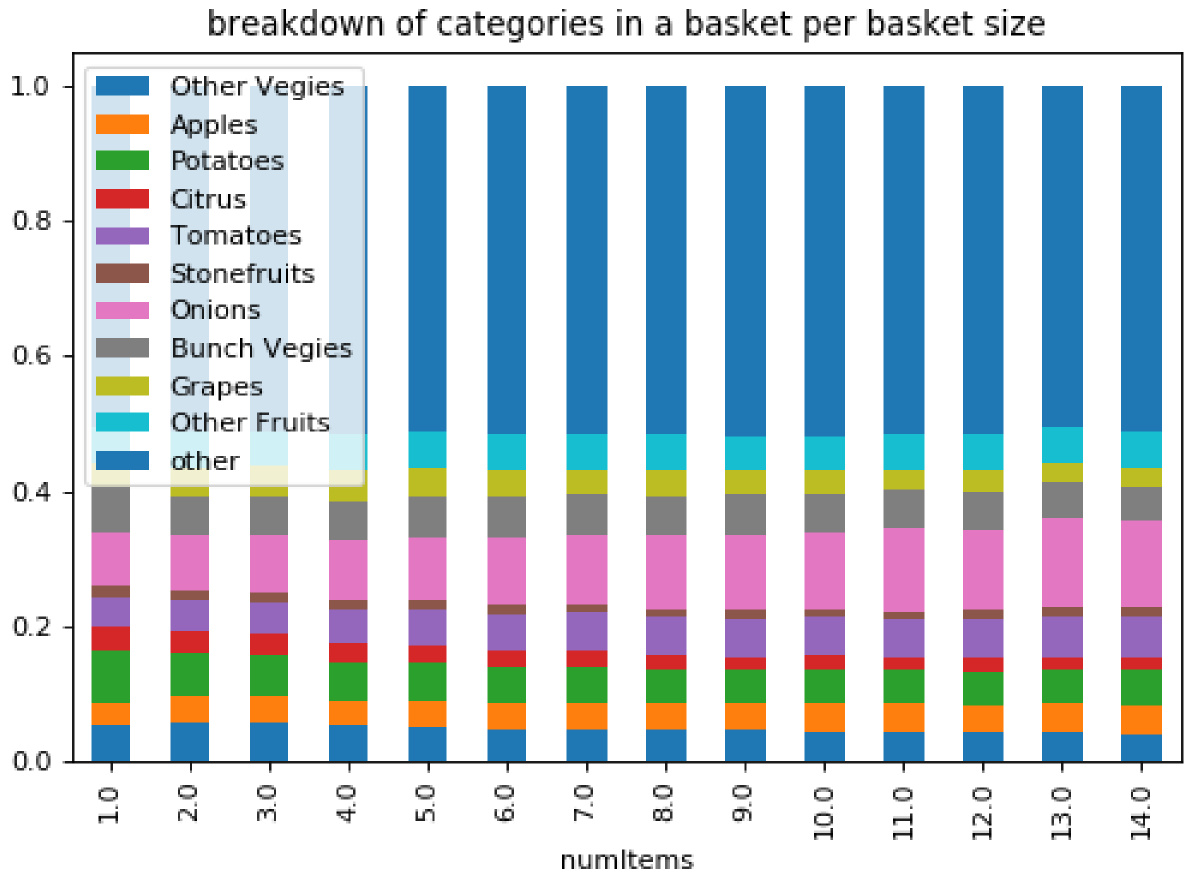
Data Cleaning/Features

All formatting of the data can be found in the dataCleaning folder within the github repository. Highlights include adding a isHoliday feature, that indicates whether a given day is a holiday within Australia. This is rolled up into the weekly counts (meaning during some weeks with holidays in short succession you would have a 2 in this column). Days with zeros were not removed. The format of the data thus looks as follows

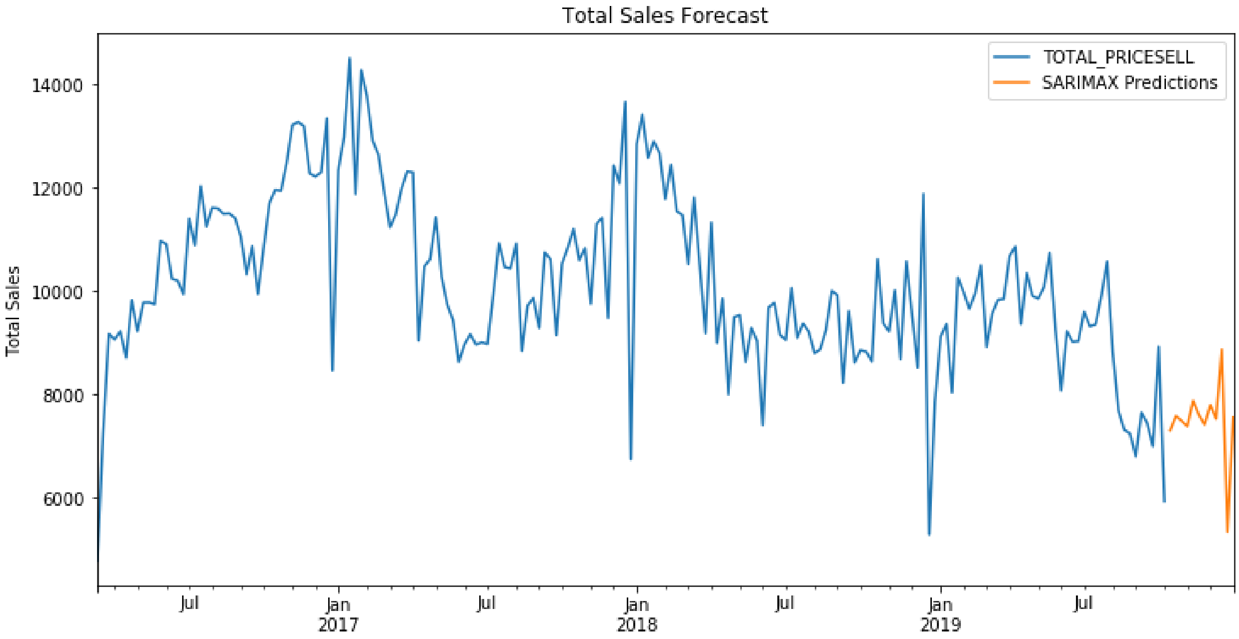
Week	Total_sellprice	isHoliday
12/23/18	4659.32	1
12/30/18	4123.33	1
…	…	…
4/21/19	3856.78	0

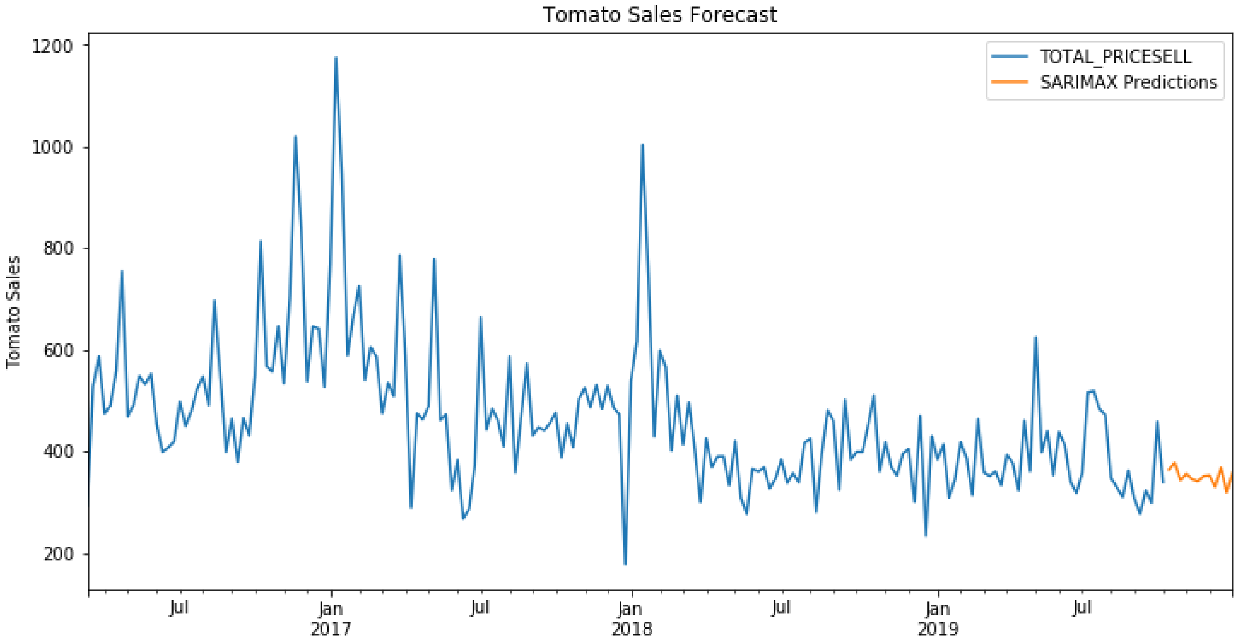
Market Basket Analysis

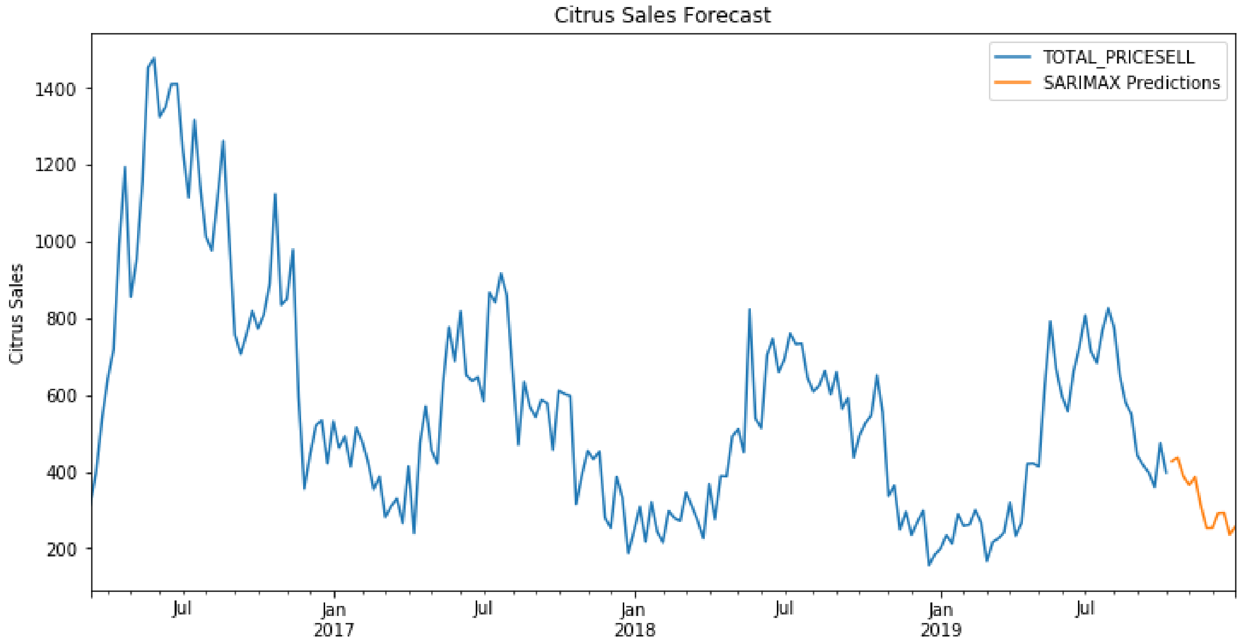
Along side our forecasting, we also studied what composes the average persons market basket, which is often of high interest in determining what products are being sold for how much and when. The table below was generated using the marketBasketAnalysis.py script located in the github repository. The beauty is that we can easily manipulate it to include the top 5,10, or X number of categories as well as cap the number of unique items in the basket. For example, we discovered that as we see an increasing number of unique items in the basket, a larger percentage of them are of the onion family, and less so are potatoes which dominate in the small purchases category. Perhaps its very common for people simply to come in and buy a few potatoes for dinner and then leave? Maybe we should run a promotion to get buyers to purchase less often bought items along with discounted potatoes. Or perhaps this high foot traffic means we should introduce other produce nearby that better compliment purchases with potatoes.

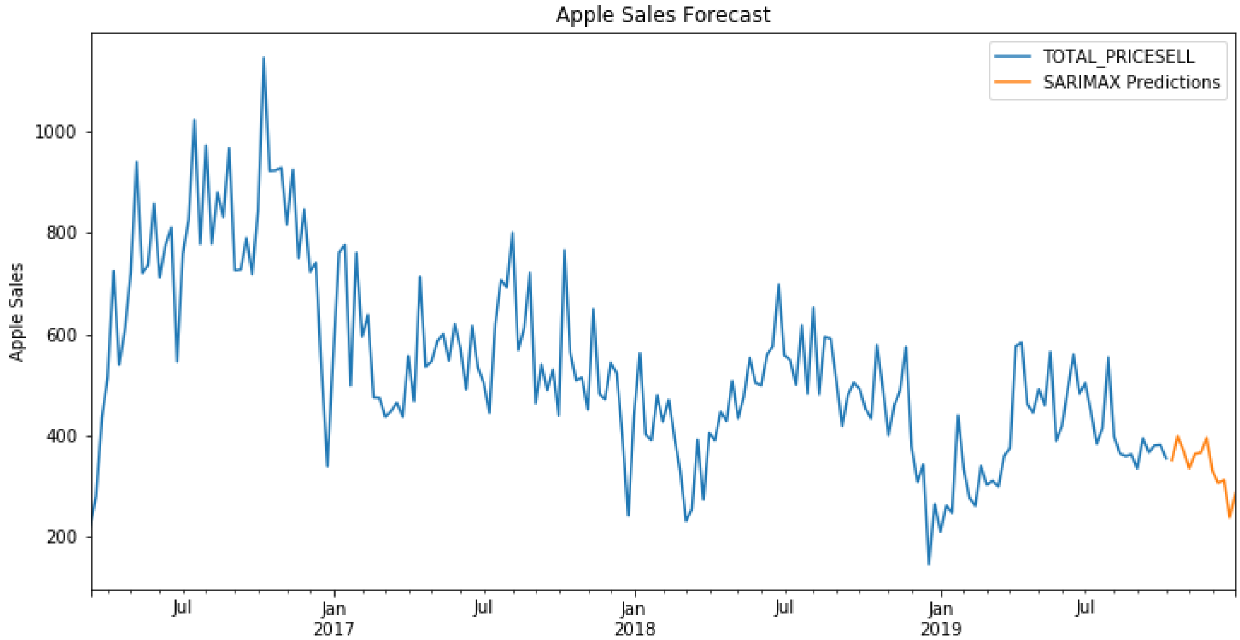


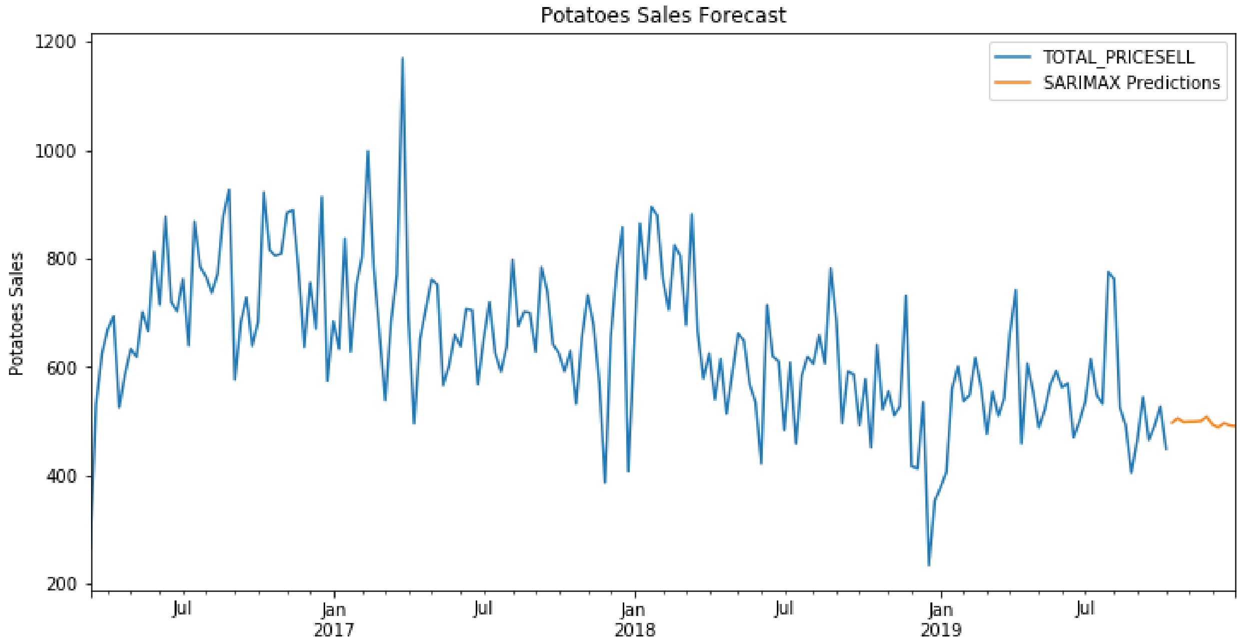
Forecasts

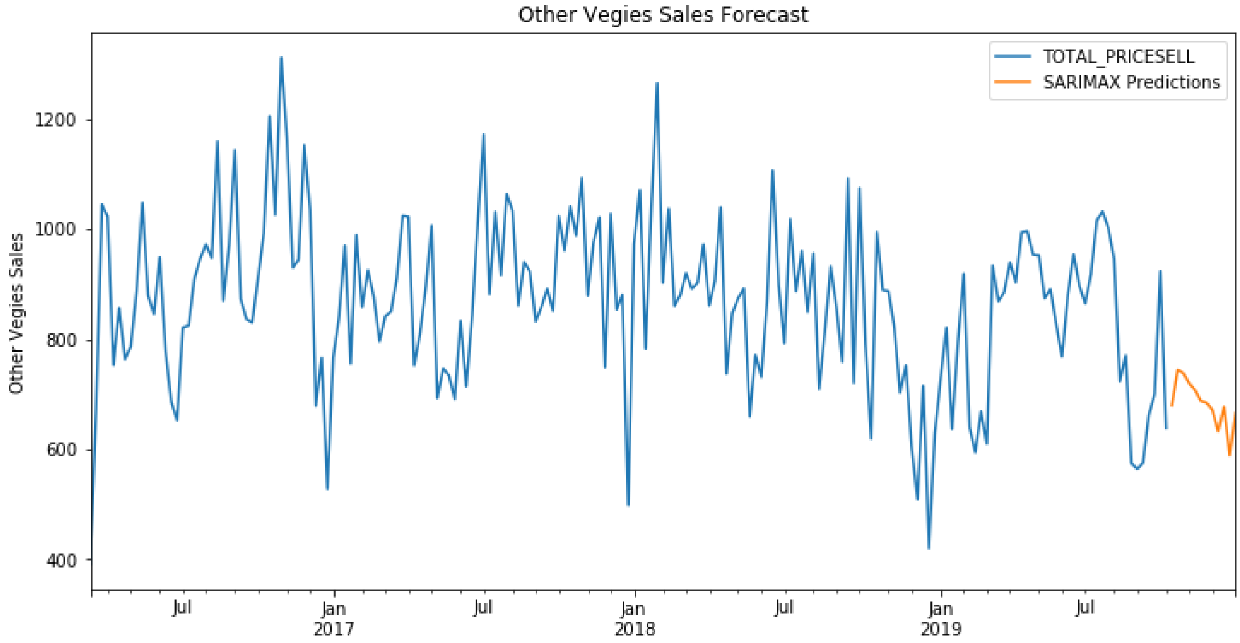

Total Sales Forecast
TOTAL_PRICESELL
SARIMAX Predictions
Total Sales
14000
12000
10000
8000
6000
Jul
Jan
2017
Jul
Jan
2018
Jul
Jan
2019
Jul


Tomato Sales Forecast
TOTAL_PRICESELL
SARIMAX Predictions
Tomato Sales
1200
1000
800
600
400
200
Jul
Jan
2017
Jul
Jan
2018
Jul
Jan
2019
Jul


Citrus Sales Forecast
TOTAL_PRICESELL
SARIMAX Predictions
Citrus Sales
1400
1200
1000
800
600
400
200
Jul
Jan
2017
Jul
Jan
2018
Jul
Jan
2019
Jul


Apple Sales Forecast
TOTAL_PRICESELL
SARIMAX Predictions
Apple Sales
1000
800
600
400
200
Jul
Jan
2017
Jul
Jan
2018
Jul
Jan
2019
Jul





Validation Summary

Total Sales SARIMAX RMSE: 1473.677226

Tomatoes SARIMAX RMSE: 81.07202963

Citrus SARIMAX RMSE: 148.9479082

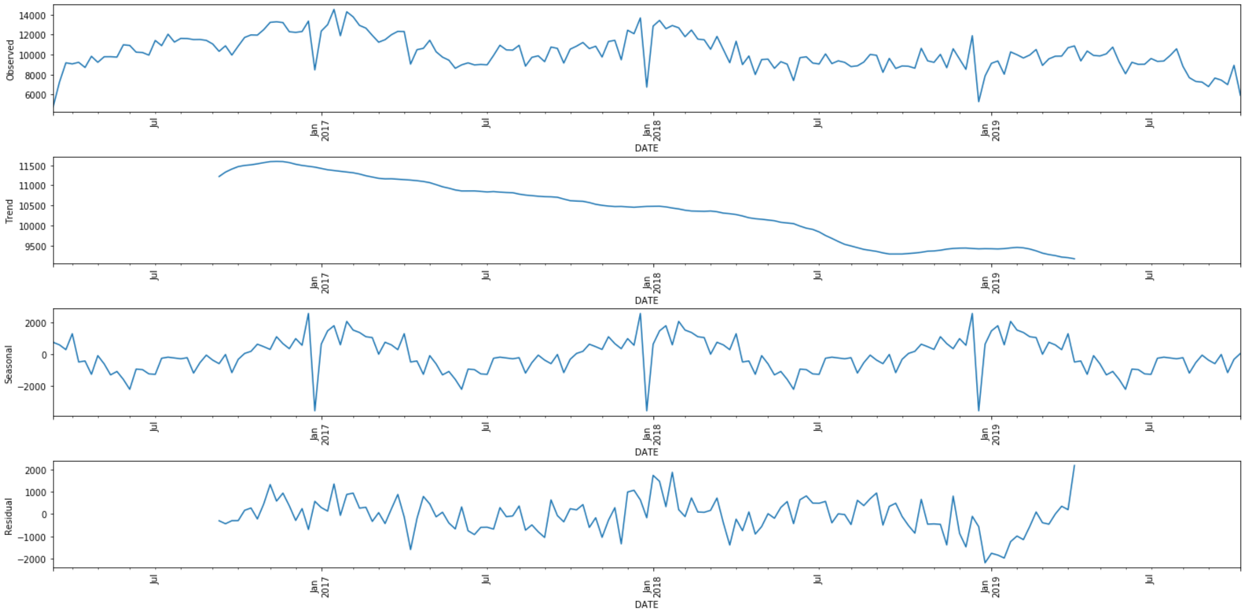
Apples SARIMAX RMSE: 165.0971303

Potatoes SARIMAX RMSE: 115.6674342

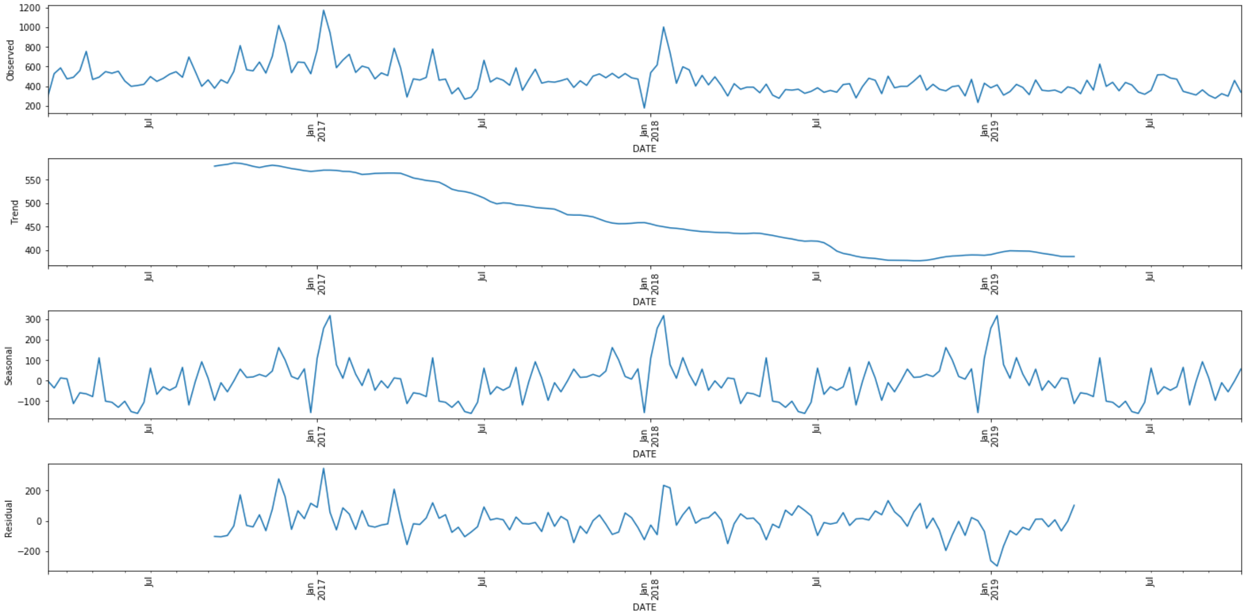
Other Vegies SARIMAX RMSE: 178.2398837

Trend Data

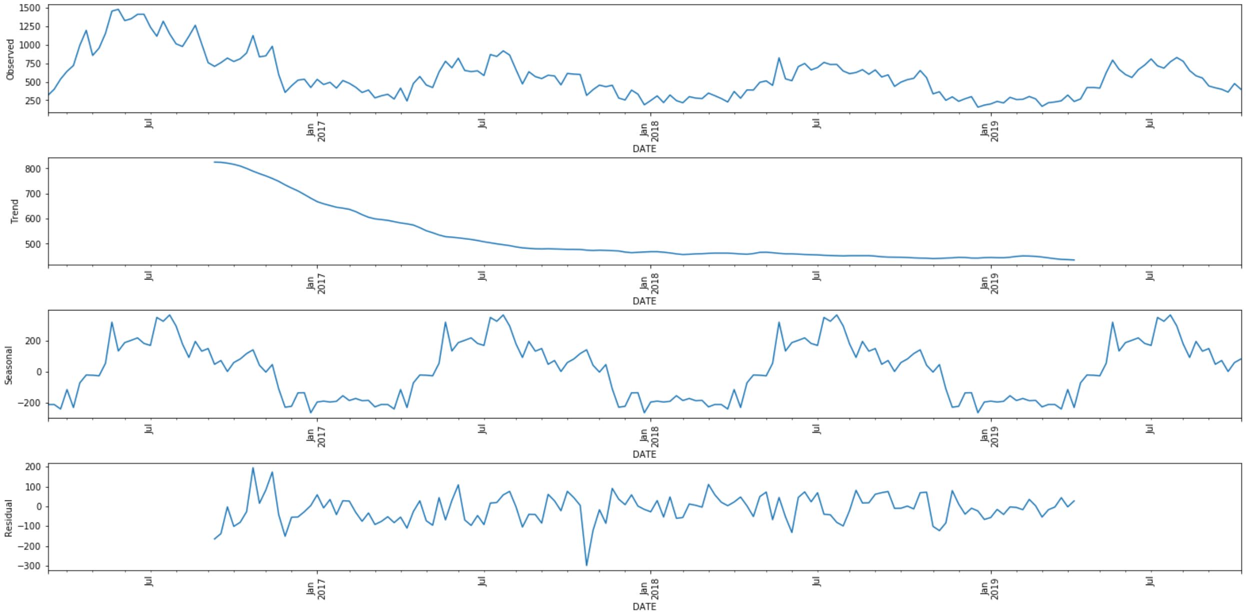
Total Sales



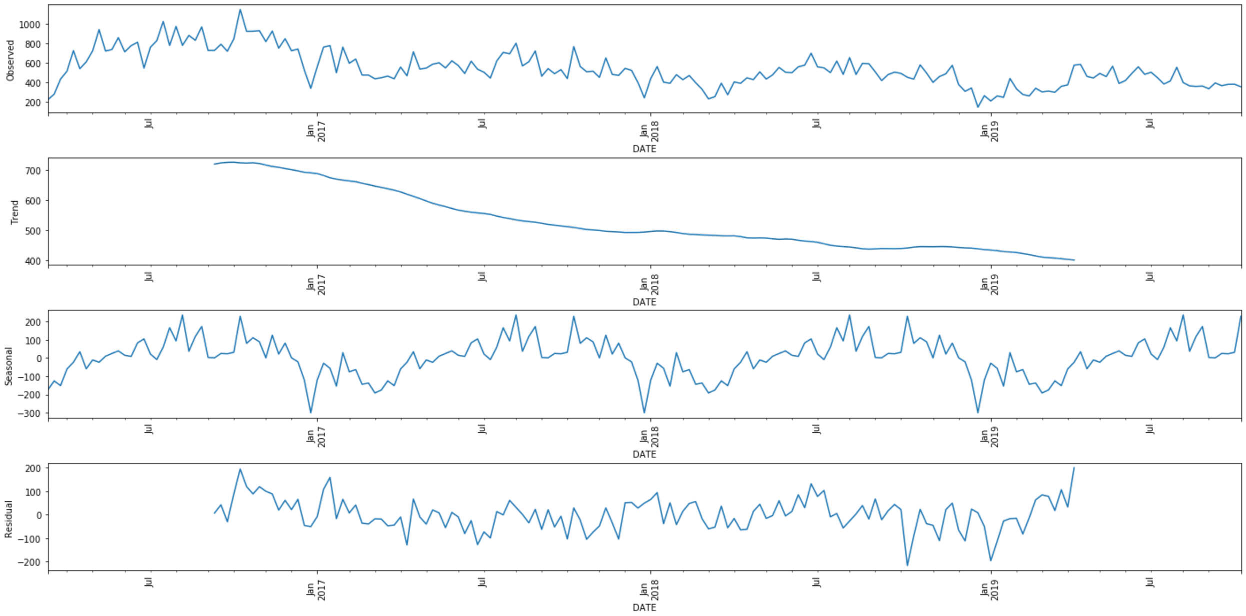
Tomatoes



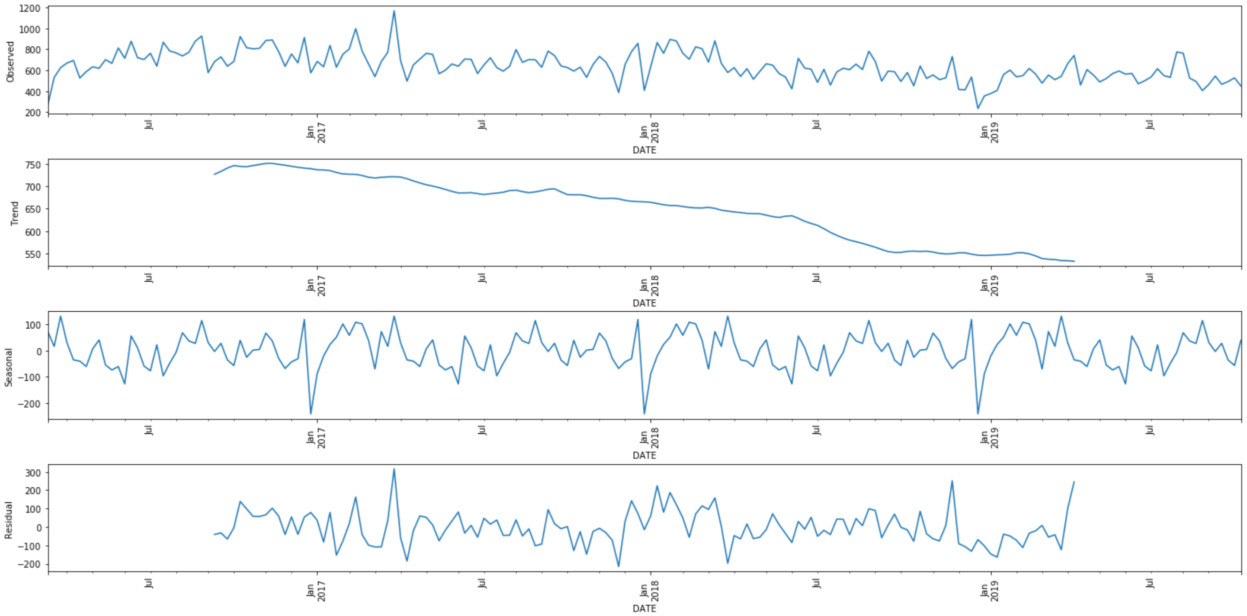
Citrus



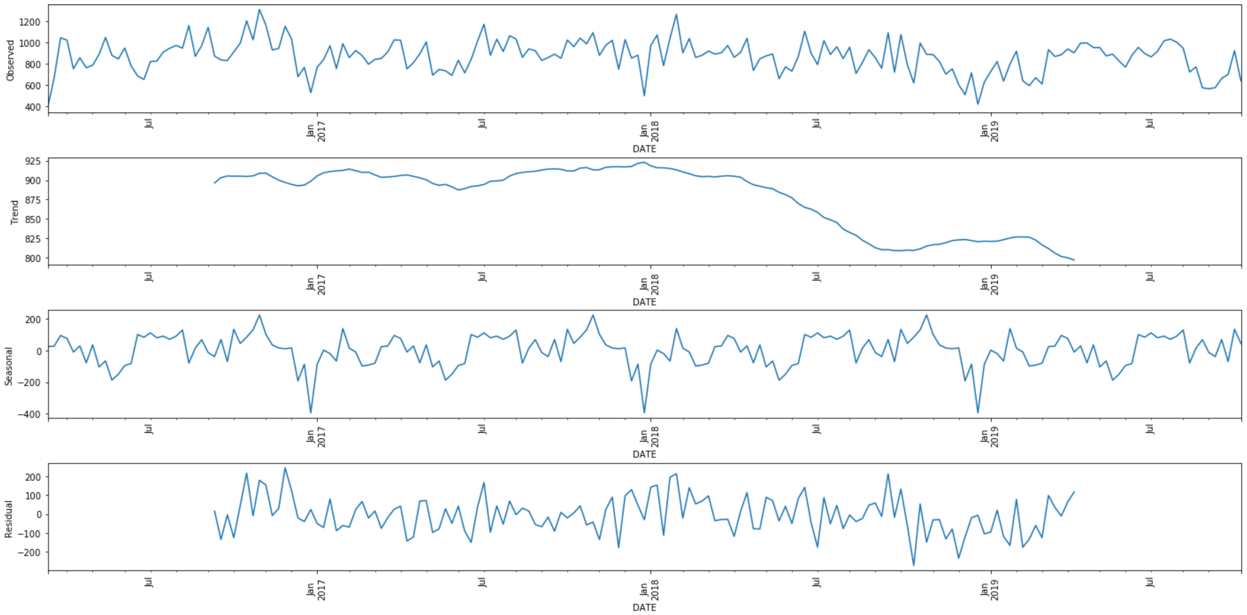
Apples



Potatoes



Other Vegies


Observed
Jul
Jan
Jul
Jan
Jul
Jan
Jul
DATE
Trend
Jul
Jan
Jul
Jan
Jul
Jan
Jul
DATE
Seasonal
Jul
Jan
Jul
Jan
Jul
Jan
Jul
DATE
Residual
Jul
Jan
Jul
Jan
Jul
Jan
Jul
DATE