Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

Using aggregated 2015 sales data for existing stores, showing the percent sales in each category, I ran a K-Centroids Diagnostics tool with the fields standardized and the clustering method set to K-Means. The minimum number of clusters[1] tested was 2 and the maximum, 10. I used the fields containing the percent sales for each category as my determinant fields. The output (Figure 1) showing the Rand and Calinski-Harabasz (CH) indices for each potential number of clusters was then observed to determine the optimum number of clusters.

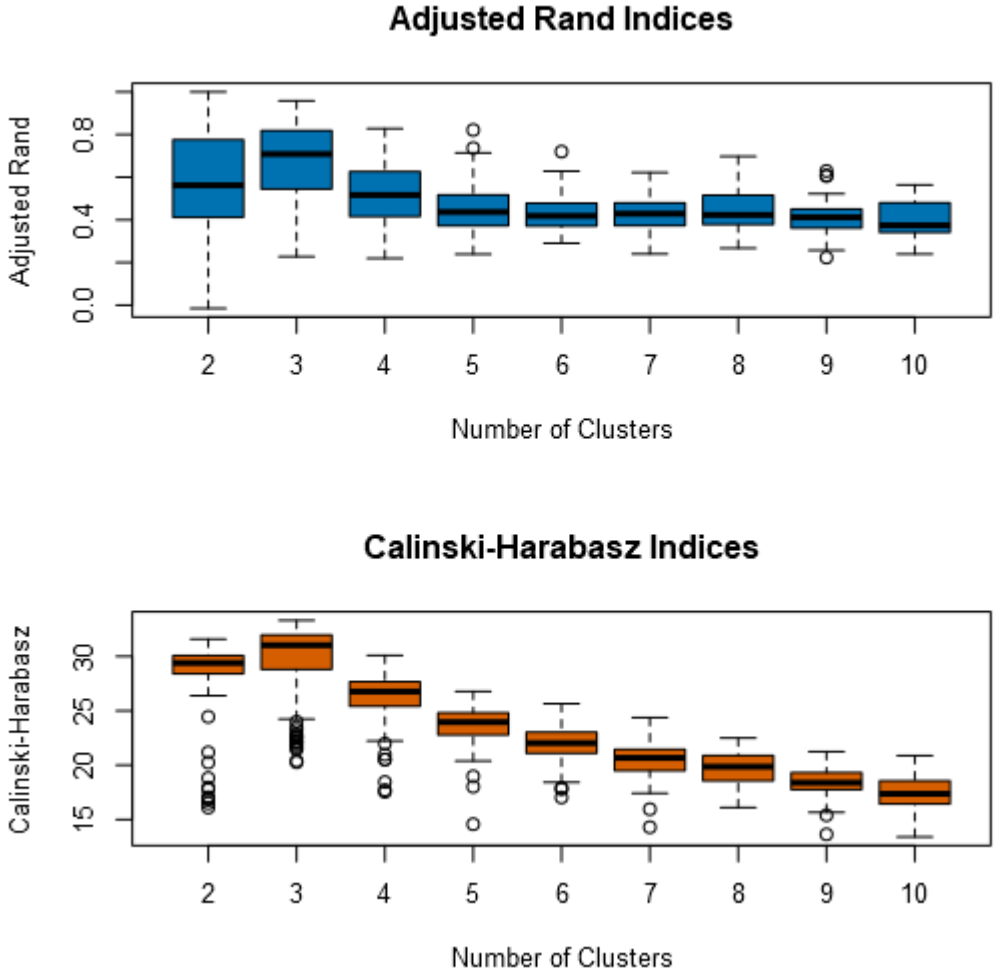

Figure 1: Rand and CH Indices used for Cluster Determination

From the plots, the median Rand and CH indices for 3 clusters is the highest indicating high stability and compactness within each cluster and high distinctness between different clusters. Knowing this, I then passed the data through a K-Centroids Cluster Analysis tool to assign each store to 1 of 3 clusters based on their standardized percent sales in all categories and using the

[1] Clusters here refers to segments, or number of store formats, all 3 may be used interchangeably

K-Means clustering method. The resulting cluster assignment is summarized in Table 1 and is further broken down in Table 2. Figure 2 shows the geographic distribution of the stores.

Table 1: Cluster Information

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Table 2: Further Breakdown of Cluster Information

	Dry Grocery	Dairy	Frozen Food	Meat	Produce	Floral	Deli	Bakery	General Merchandise
1	0.33	-0.76	-0.39	-0.09	-0.51	-0.30	-0.23	-0.89	1.21
2	-0.73	0.70	0.35	-0.49	1.01	0.85	-0.55	0.40	-0.30
3	0.41	-0.09	-0.03	0.49	-0.54	-0.54	0.65	0.27	-0.57

Looking at the further breakdown of the cluster information, it can be observed from the max differences in each category that clusters 1 and 2 differ on their percent sales in dairy, frozen food and bakery; while clusters 1 and 3 differ in their percent sales in general merchandise. Clusters 2 and 3 differ in their percent sales in dry grocery, meat, produce, floral, and deli.

Figure 2: Geographic Distribution of Stores by Segment and Sales

Task 2: Formats for New Stores

Starting from a file that included store demographics for existing stores and their segments, I tested 3 predictive models that handle categorical variables: Decision tree, Forest and Boosted models. I held 20% of the stores as a validation sample to compare the accuracy of the models as summarized in Table 3.

Table 3: Fit and Error Measures of Predictive Models

Model	Accuracy	F1	Accuracy 1	Accuracy 2	Accuracy 3
Forest	0.8235	0.8251	0.7500	0.8000	0.8750
Decision Tree	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000

The resulting comparison report showed similar overall accuracy for all 3 models with the Boosted model having the highest measure of fit. Further observation of the individual segment predictions show that the Boosted Model has the highest prediction accuracy for the segments 1 and 3 and the least for segment 2. The Forest and Decision Tree models have the same fit and accuracy measures. The confusion matrices shown below confirm the findings.

Table 4: Confusion Matrices for all 3 Models

Confusion matrix of Boosted			
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6
Confusion matrix of Decision Tree			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7
Confusion matrix of Forest			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Based on these findings, I decided to use the boosted model for predicting the segments that each of the 10 new stores will fall into. The result is summarized in the table below.

Table 5: New Store Segments

Store Number	Segment
S0086	3
S0087	3
S0088	3
S0089	3
S0090	3
S0091	3
S0092	3
S0093	3
S0094	3
S0095	3

Task 3: Predicting Produce Sales

For the existing stores, I aggregated the produce sales on a monthly basis from March 2012 to December 2015 to produce a dataset that meets the criteria of a time series dataset. It is continuous over the time range, it is categorized by months, and it contains a sales value for each month, measured in sequence across the time range. The most recent 12 months in the time range (2015 sales) will be used as a holdout sample to internally validate the forecasting model(s) to be compared. For the new stores, since all of them were predicted to fall in segment 3, I filtered out the average produce sales data for existing stores in that segment for each month in the same time range (March 2012 to December 2015). The average produce sales was then multiplied by the number of new stores, 10. This dataset, being a subset of the earlier described dataset, also meets the criteria for a time series dataset. I also used the monthly produce sales for the year 2015 to validate the forecasting models. The graphs below show the time series plots of the monthly sales data in both cases.

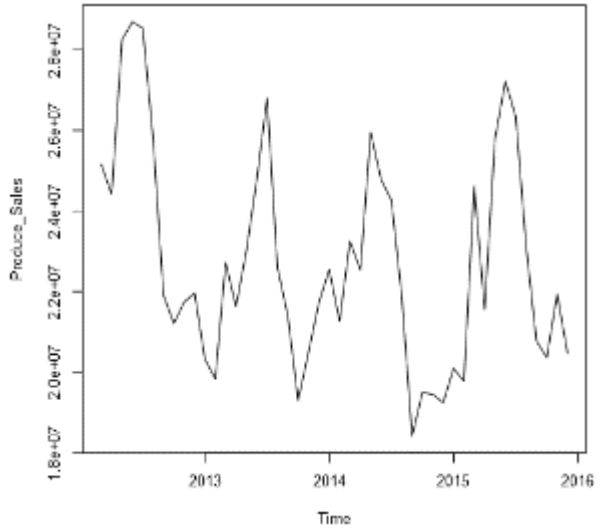

Figure 3: Time Series Plot (Existing Stores)

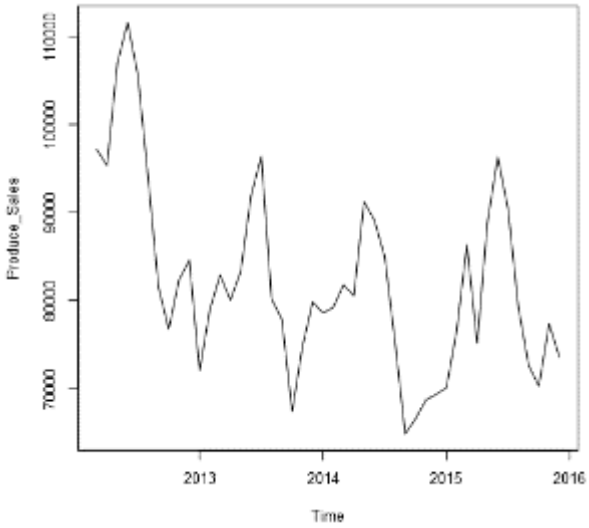

Figure 4: Time Series Plot (New Stores)

Both time series were then decomposed to observe seasonality, trend and error.

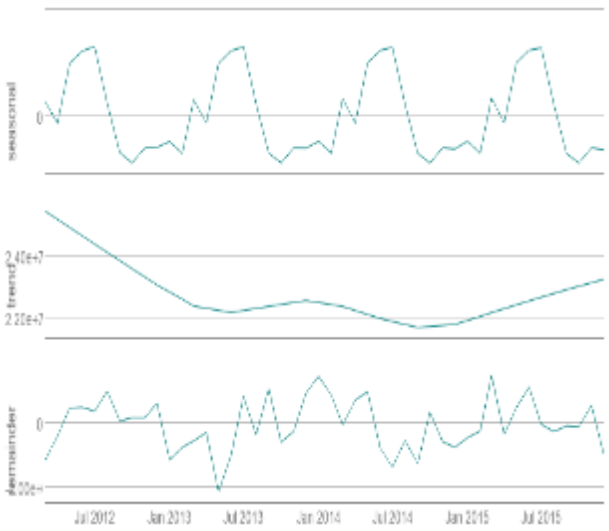

Figure 5: Decomposition Plot (Existing Stores)

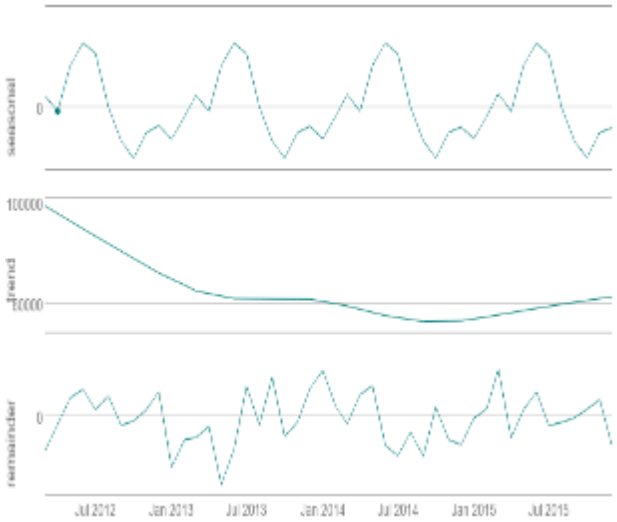

Figure 6: Decomposition Plot (New Stores)

Upon observing both decomposition plots, no obvious trends were observed, seasonality declined very slightly over the years suggesting being a multiplicative component. The errors (remainders) showed changing variances as the plot moved along also suggestive of a multiplicative error component. Therefore, the ETS models considered for both forecasts were ETS MNM. The ACF[2] and PACF[3] plots were then observed to determine the ARIMA terms for the time series and seasonality components. The plots show a positive correlation for the first terms in both ACF and PACF plots with a gradual decay in the ACF but sharp cutoff in PACF, suggesting an AR term of 1 and an MA term of 0 for the time series component of the ARIMA model.

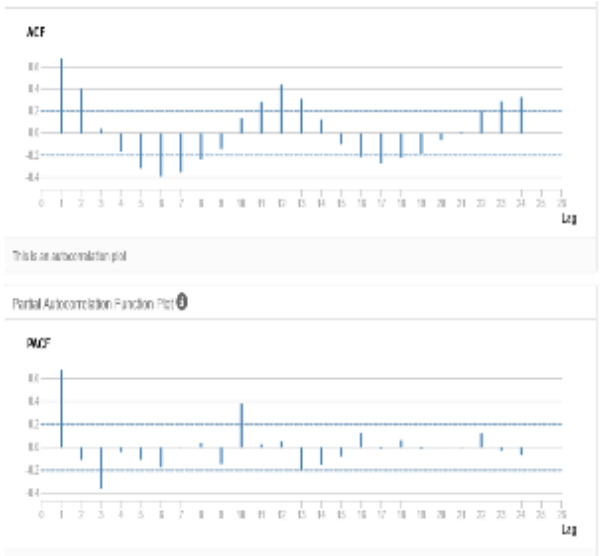

Figure 7: ACF and PACF Plots (Existing Stores)

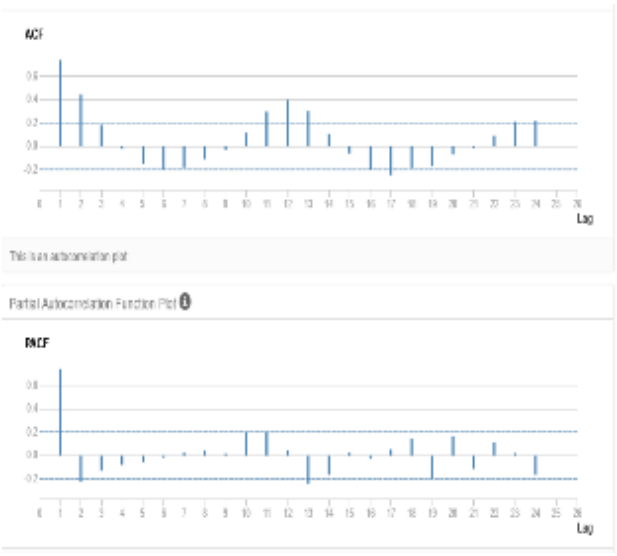

Figure 8: ACF and PACF Plots (New Stores)

[2] ACF refers to Auto-Correlation Function
[3] PACF refers to Partial ACF[2]

The ACF plots gradually decay to 0 with increases at the seasonal lags (12 and 24). Since serial correlation is high, the series had to be seasonally differenced. The plots below show the time series, ACF and PACF plots of the seasonally differenced series.

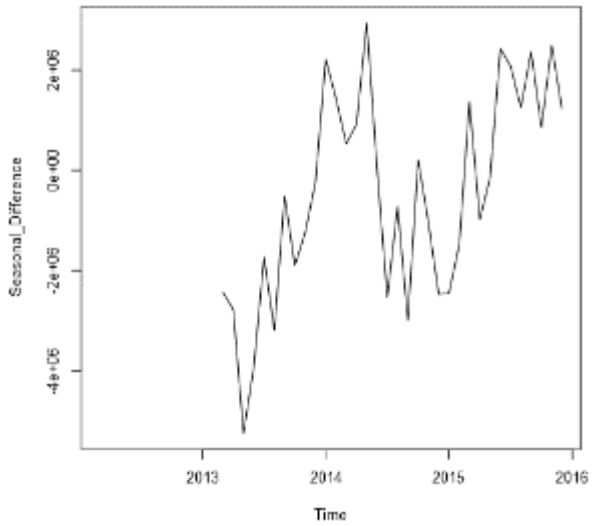

Figure 9: Time Series Plot (Existing Stores, Seasonally Differenced)

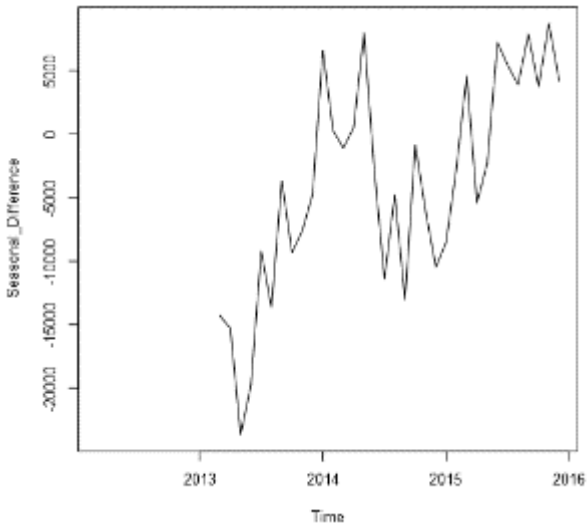

Figure 10: Time Series Plot (New Stores, Seasonally Differenced)

From the time series plots, it can be observed that stationarity has not been achieved.

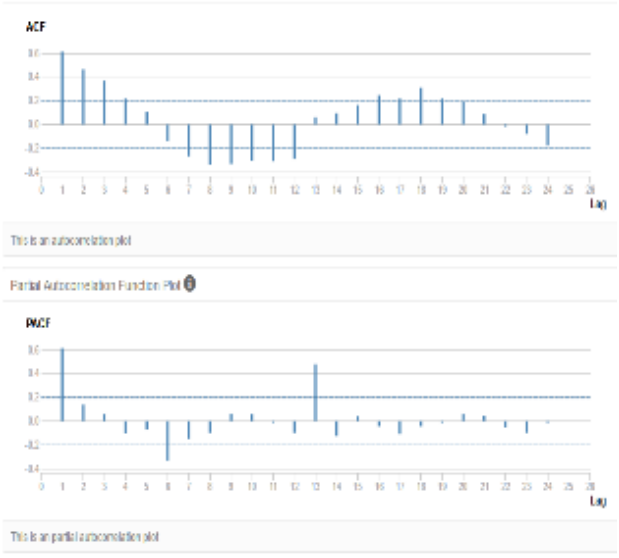

Figure 11: ACF and PACF Plots (Existing Stores, Seasonally Differenced)

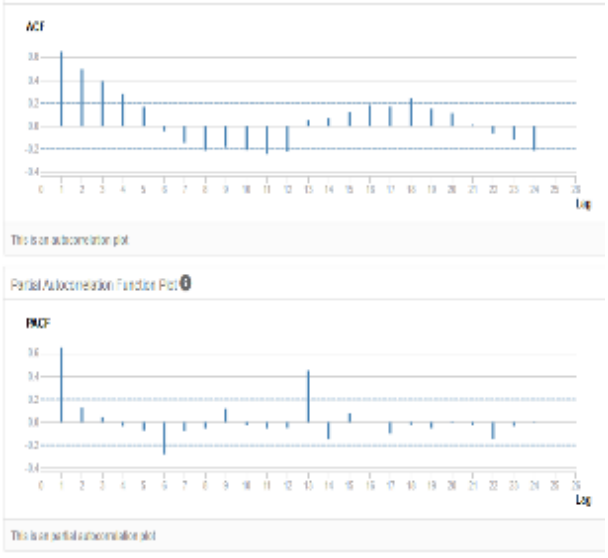

Figure 12: ACF and PACF Plots (New Stores, Seasonally Differenced)

Although less correlated, the ACF and PACF plots for the seasonal difference exhibit similar patterns as those in the plots prior to differencing. To further capture the correlation in in this series, the first difference was then taken, after which stationarity was achieved.

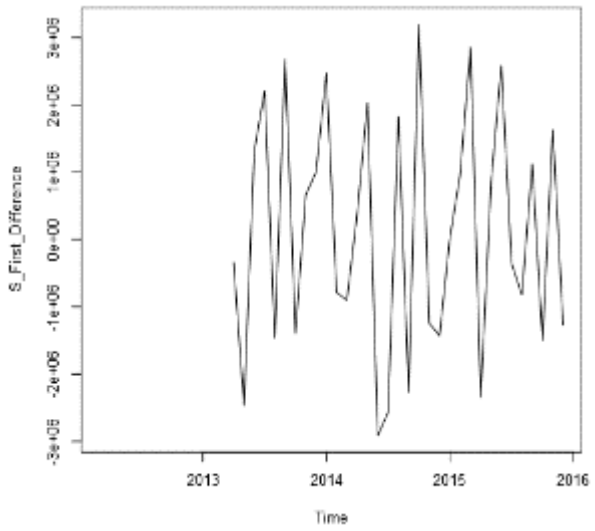

Figure 13: Time Series Plot (Existing Stores, Seasonal First Difference)

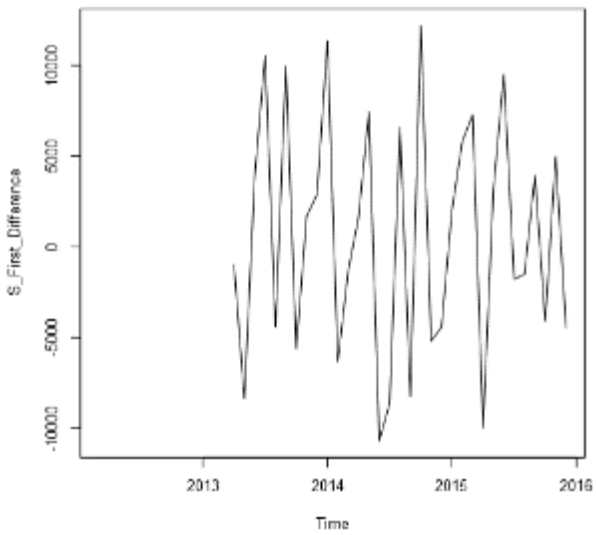

Figure 14: Time Series Plot (New Stores, Seasonal First Difference)

The differencing or I term of the model for both the time series and seasonal components will be 1 as both seasonal and first differences were taken to remove significant lags from the ACF and PACF plots shown below. The ACF and PACF plots suggest MA terms and no AR terms for the time series and seasonal components of the ARIMA model. This is because the ACF plot cuts off to zero after lag 1 while the PACF gradually decays to zero, and both plots show a strong negative correlation for lag 1 and lag 12. 1 MA term was used for the time series component and 2 for the seasonal component as the correlation at lag 12 was pretty significant.

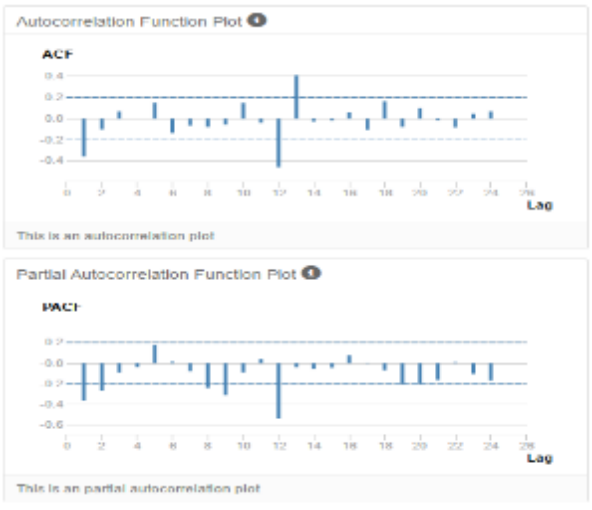

Figure 15: ACF and PACF Plots (Existing Stores, Seasonal First Difference)

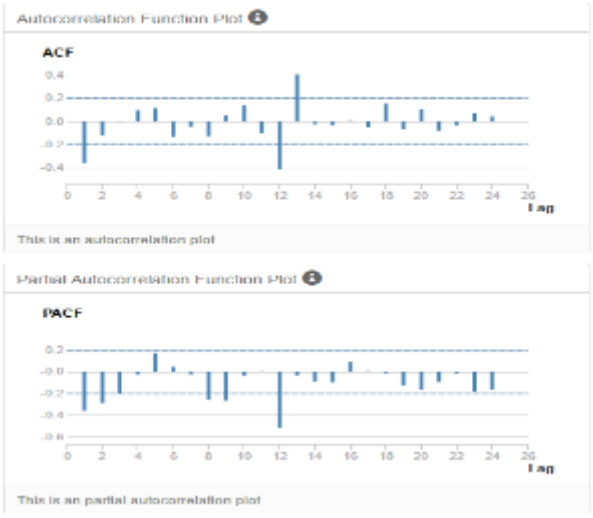

Figure 16: ACF and PACF Plots (New Stores, Seasonal First Difference)

Therefore, the ARIMA model used was the ARIMA 011 012. This was compared against the ETS MNM model to determine the best model to be used in forecasting produce sales.

To compare both models, in-sample error measurements as well as forecast errors from a holdout sample of the 12 months sales in 2015 were used. The ARIMA models captured most of the

correlation in the series, with just a couple barely significant correlations at lags 12 and 13 as shown in the plots below.

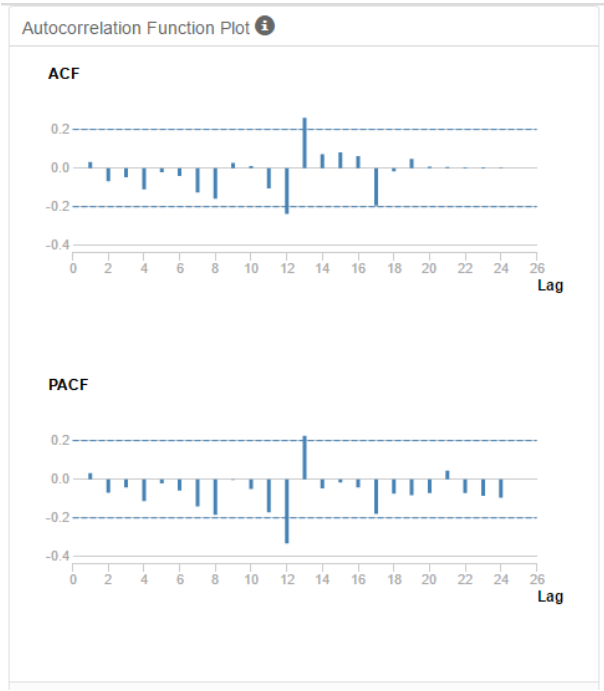

Figure 17: ACF and PACF Plots (Existing Stores, ARIMA 011 012)

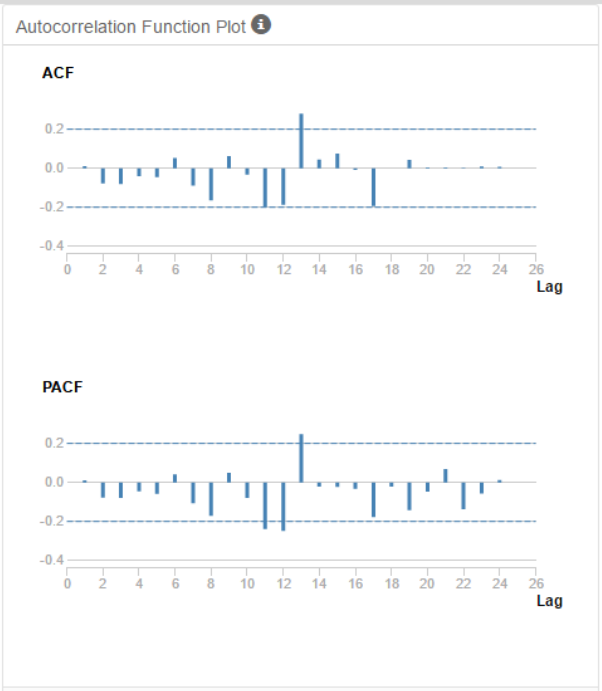

Figure 18: ACF and PACF Plots (New Stores, ARIMA 011 012)

The in-sample and forecast error measurements are summarized in the table below. While the ARIMA models had lower in-sample errors, the ETS models were better at making forecasts.

Table 6: In-sample and Forecast Error Measurements

MODEL	IN-SAMPLE			FORECAST	
	RMSE	MASE	AIC	RMSE	MASE
EXISTING STORES					
ETS MNM	1015013	0.47	1089.7	2226513	1.2691
ARIMA 011 012	622761.55	0.2	660.9	3035191	1.8355
NEW STORES					
ETS MNM	4560.44	0.46	722.8	7355.64	1.1481
ARIMA 011 012	2369.76	0.16	427.1	13605.88	2.2526

To predict 2016 monthly produce sales for new and existing stores, I used the ETS MNM models. The results are shown in a Tableau Public Dashboard here.