Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

   The optimal number is 3 store formats because Adjusted Rand (AR) and Calinski-Harabasz (CH) indices have higher median (AR: 0.313552, CH: 16.34166) than for clusters 2 and 4 to 6 and the boxplots are compact. For the AR index, the higher the value index, the better the stability of the cluster. A higher CH index denotes higher compactness and distinctness of the clusters (formats).

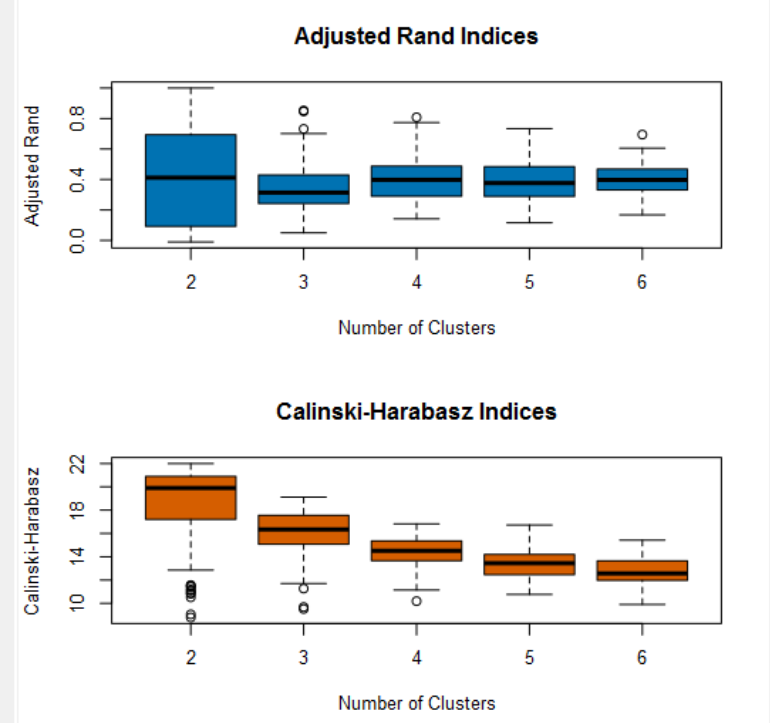

Figure 1. Adjusted Rand (AR) and Calinski-Harabasz (CH) indices to classify the stores into 2 to 6 clusters according to the percentage of sales per category per store.

2. How many stores fall into each store format?

   Table 1 contains the number of stores in each store format: 23 in store format (or cluster) 1, 29 stores in format 2 and 33 stores in format 3.

Table 1. Number of stores in each of the three formats.

Format	Number of stores
1	23
2	29
3	33

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   As shown in the boxplots of Figure 2, one way that the clusters differ from each other is in the median and interquartile range of different measures such as sum of deli, sum of floral, sum of frozen food and sum of general merchandise. Cluster or format 1 in shown in blue, cluster 2 in orange and cluster 3 in red.

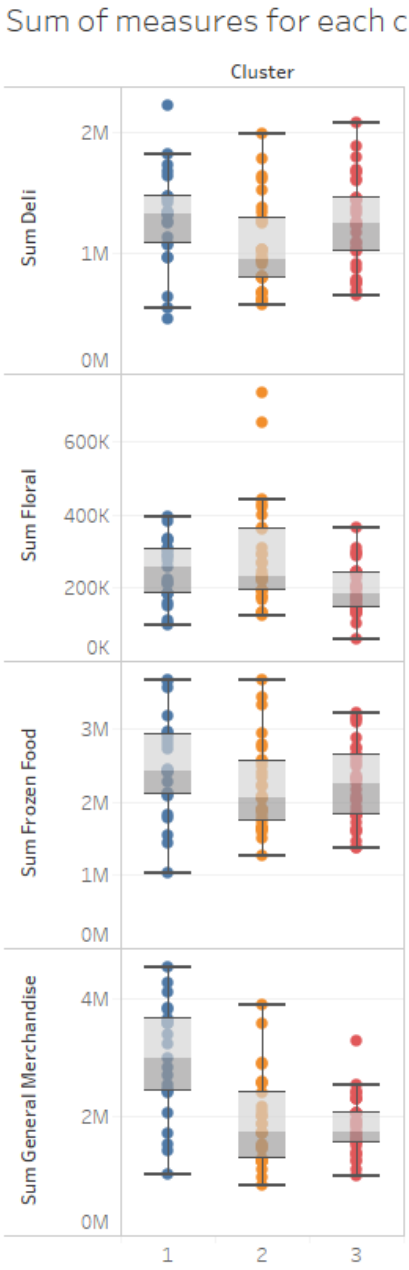

Figure 2. Different measures than the ones used for clustering to see different median and interquartile ranges in the three clusters/formats.

4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

Figure 3 contains a Tableau visualization of the three store formats, using color to show cluster and size to quantify total sales.

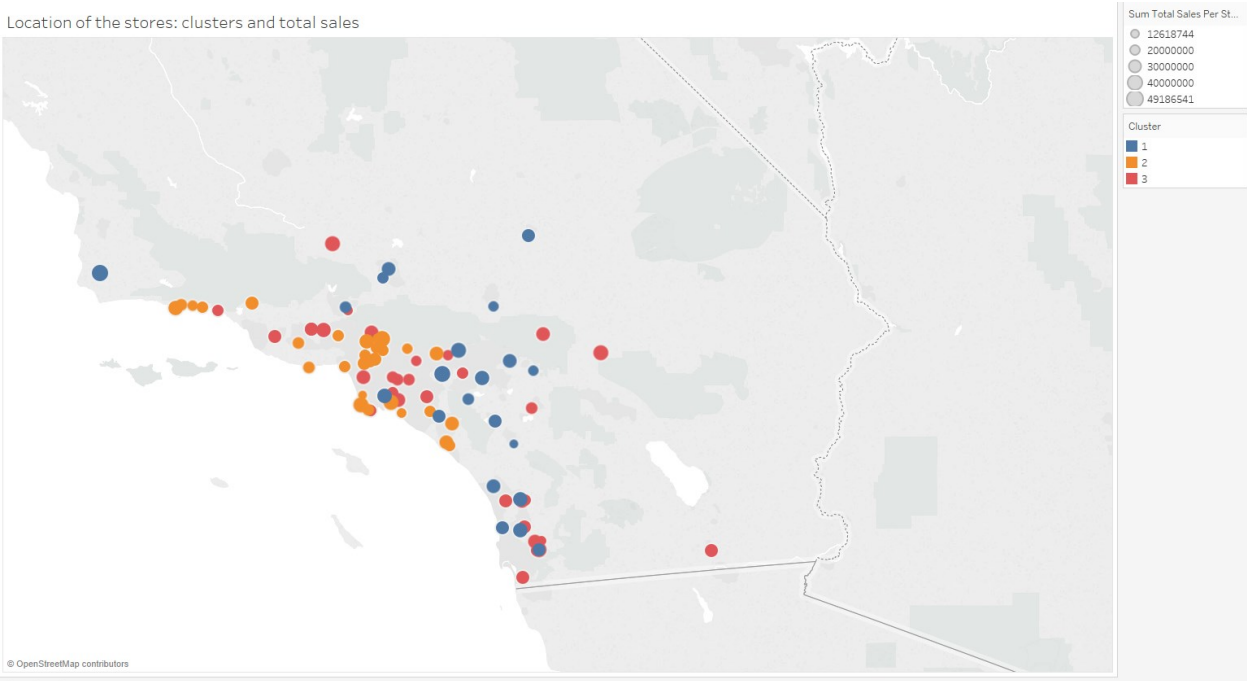

Figure 3. Eighty-five stores classified in three store formats in California. The three colors identify the three different store types and the size of the circles show the amount of total sales.

Task 2: Formats for New Stores

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   Cluster was used as the target variable and all the demographics variables (except for the Store field) were used as predictor variables. As shown in Table 2, forest model and boosted model had the same overall accuracy (0.8235) but Boosted model had higher F1 score (0.8889 versus 0.8426) so it was selected as the best methodology.

Table 2. Model comparison measures for decision tree, random forest and boosted models.

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Random_Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

2. What format do each of the 10 new stores fall into? Please fill in the table below.

   Table 3 shows the number of format where stores S0086 to S0095 fall into.

Table 3. Classification of the 10 new stores into formats.

Store Number	Format
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

The model for each forecast was ETS(M,N,M), which compared to ARIMA(1,0,0)(1,1,0)[12] in terms of forecast error measurements against the holdout sample obtained from the TS Compare tool, the results were better (especially lower RMSE and MASE as indicated in Tables 4 for ETS model and 5 for ARIMA model). Root Mean Squared Error (RMSE) represents the sample standard deviation of the differences between predicted values and observed values. Mean Absolute Scaled Error (MASE) is defined as the mean absolute error of the model divided by the mean absolute value of the first difference of the series.

Table 4. Forecast error measurements against the holdout sample for ETS(M,N,M) model.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_model	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

Table 5. Forecast error measurements against the holdout sample for ARIMA(1,0,0)(1,1,0)[12] model.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_model	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

The data and three main components — seasonality, trend and error —are shown in Figure 4.

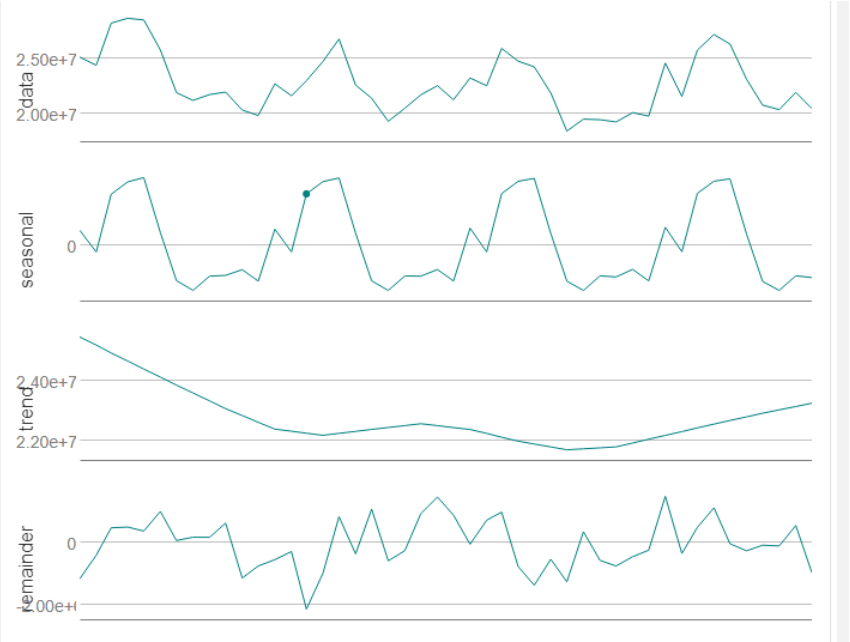

Figure 4. Time series decomposition plot. From top to bottom: data, seasonal patterns, general tendencies and error.

For ETS model on one hand, the error increases variance as the time increases; therefore we should apply the error part multiplicatively. The trend is almost a line with zero slope (the value ranges approximately from 2.20e+7 to 2.40e+7) therefore we should apply no trend (None). The seasonality usually increases in volume each seasonal period, so we should apply seasonality in a multiplicative manner. The model is **ETS(M,N,M)**.

To get the optimal parameters for ARIMA model, it is necessary to plot the autocorrelation function (ACF) and partial correlation function (PACF) in different steps. Figure 5 contains the ACF and PACF of the original time series, and Figure 6 after seasonal difference.

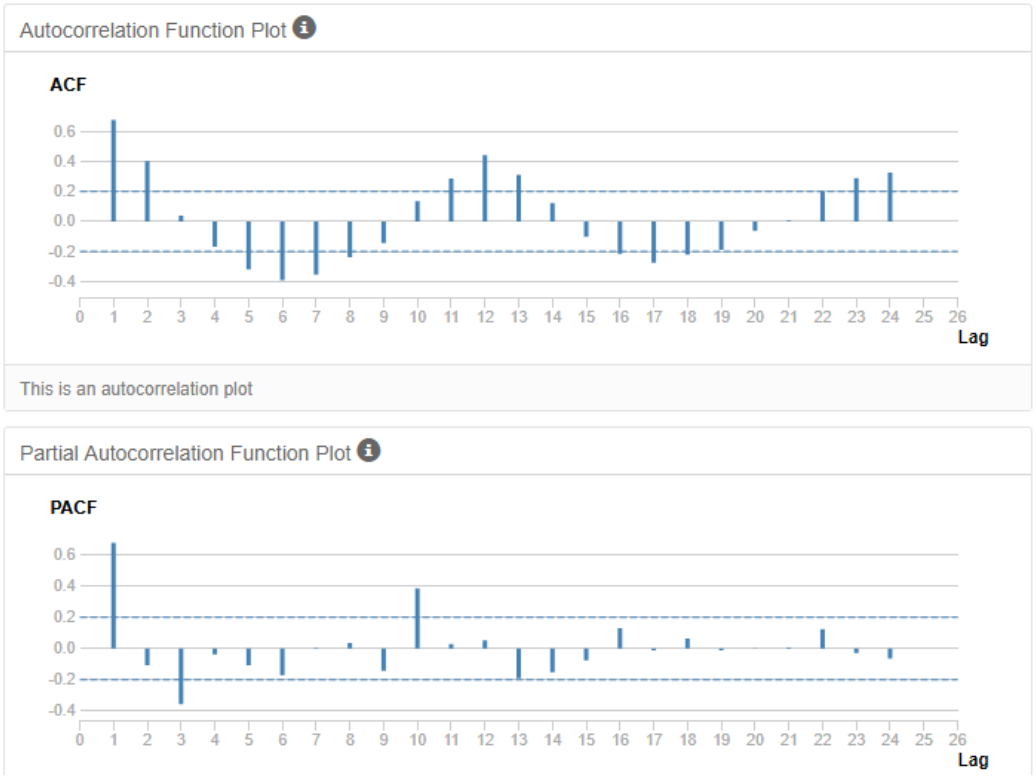

Figure 5. Autocorrelation plots from the original time series.

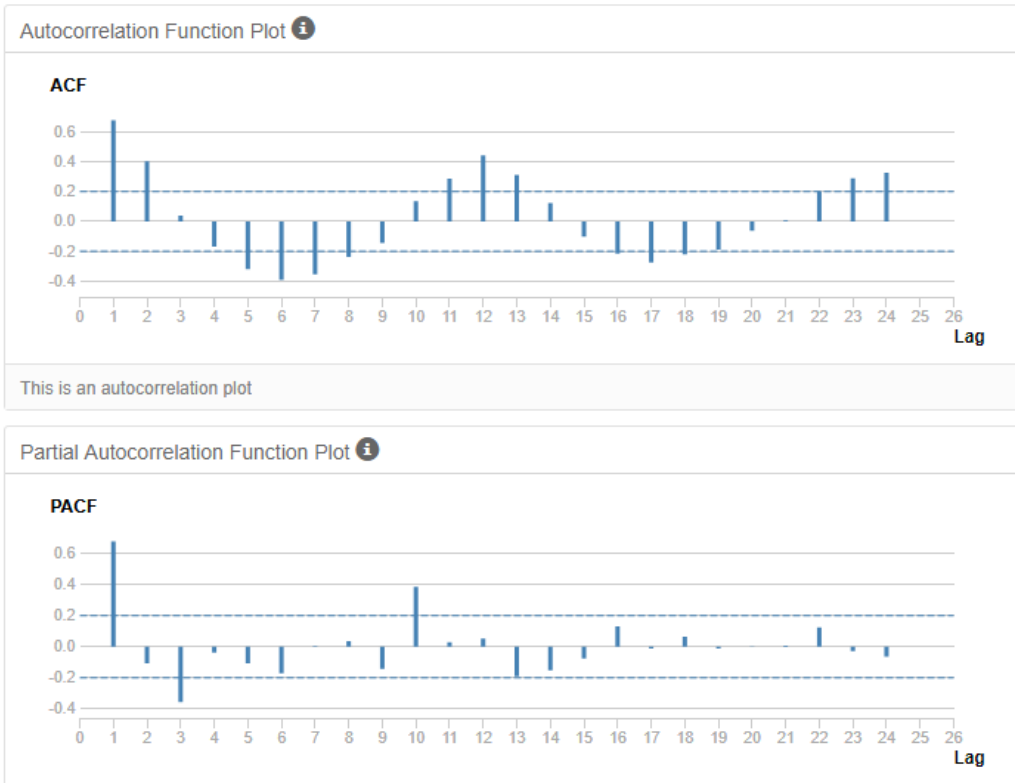

Figure 6. Autocorrelation plots after seasonal difference.

The model needs an AR term for the seasonal part (positive ACF and PACF in lag 1). Figure 7 shows the autocorrelation plots after adding the aforementioned AR term.

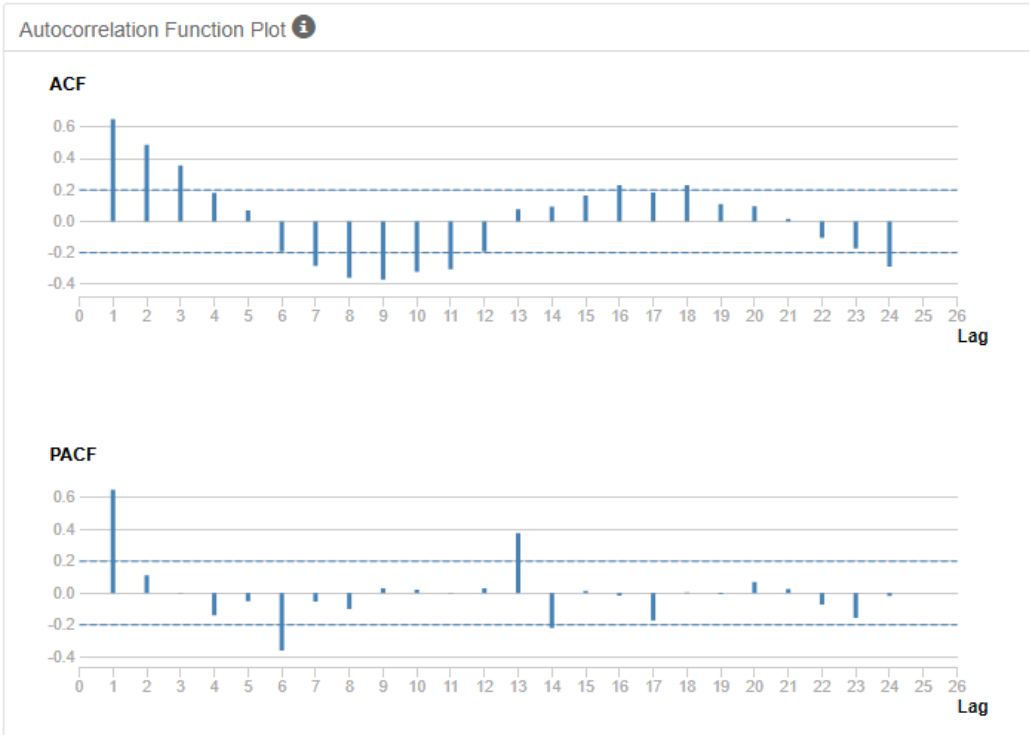

Figure 7. Autocorrelation plots after adding the AR term in the seasonal part of the model.

The plots in Figure 7 indicate that ARIMA needs an AR term for the non-seasonal part. The model also needs and upper case D of 1 since we took a seasonal difference and the AR term in the seasonal part. The final model is **ARIMA(1,0,0)(1,1,0)[12],** whose autocorrelation plots indicate that the time series are stationarized (Figure 8).

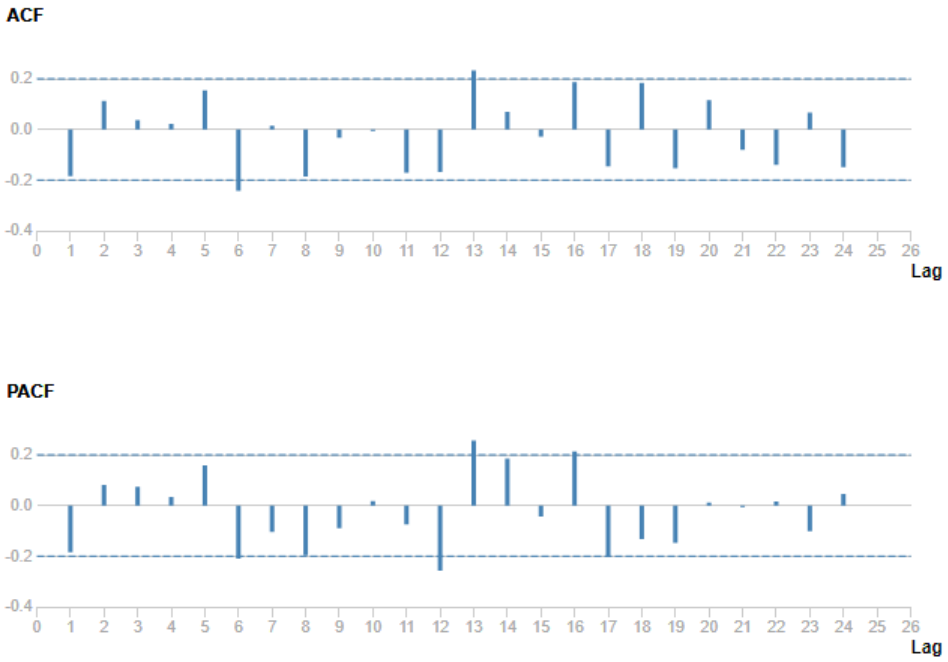

Figure 8. Autocorrelation plots after applying the ARIMA(1,0,0)(1,1,0)[12] model.

3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

   Table 6 contains the forecasted values for produce, monthly in 2016 for new and existing stores, and Figure 9 shows the historical data together with these forecasts.

Table 6. Forecast values for produce in new and existing stores in 2016.

Month	Produce Forecast New Stores ($)	Produce Forecast Existing Stores ($)
Jan-16	2,587,451	21,539,936
Feb-16	2,477,353	20,413,771
Mar-16	2,913,185	24,325,953
Apr-16	2,775,746	22,993,466
May-16	3,150,867	26,691,951
Jun-16	3,188,922	26,989,964
Jul-16	3,214,746	26,948,631
Aug-16	2,866,349	24,091,579
Sep-16	2,538,727	20,523,492
Oct-16	2,488,148	20,011,749
Nov-16	2,595,270	21,177,435
Dec-16	2,573,397	20,855,799

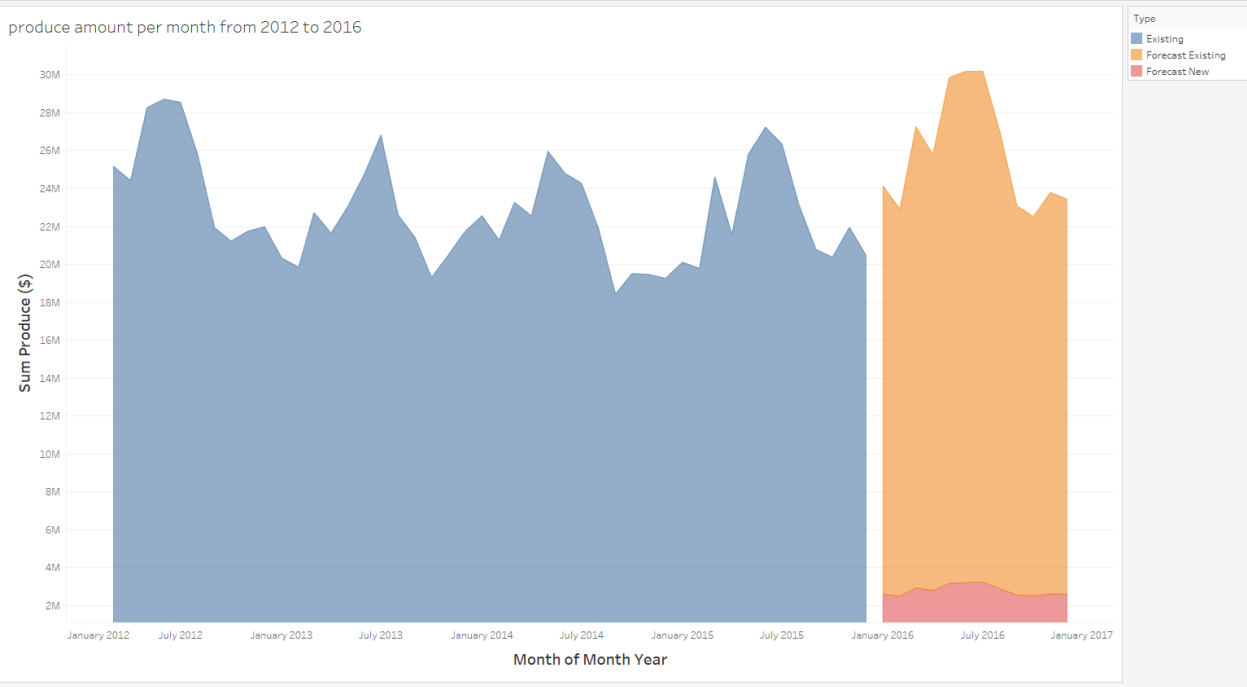

Figure 9. Produce values for existing stores from 2012 to 2015 and forecasted produce values for existing and new stores monthly in 2016.