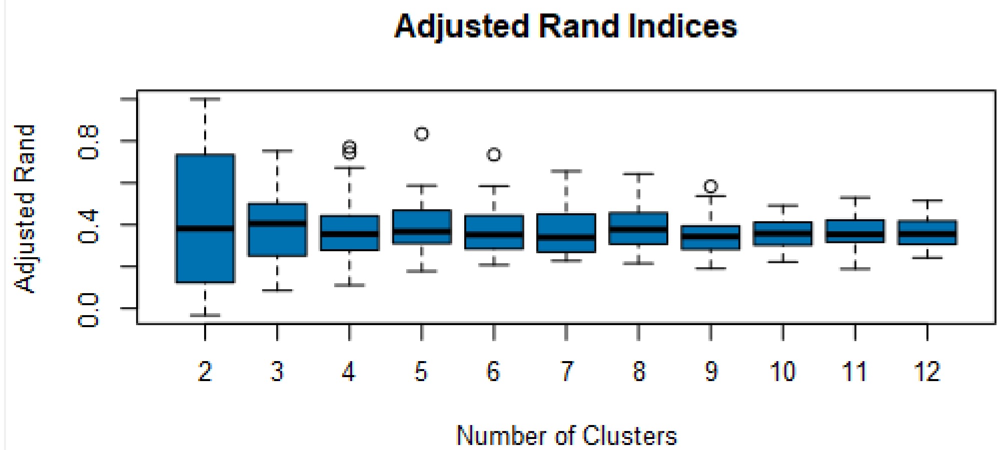
<u>Project: Predictive Analytics Capstone</u>

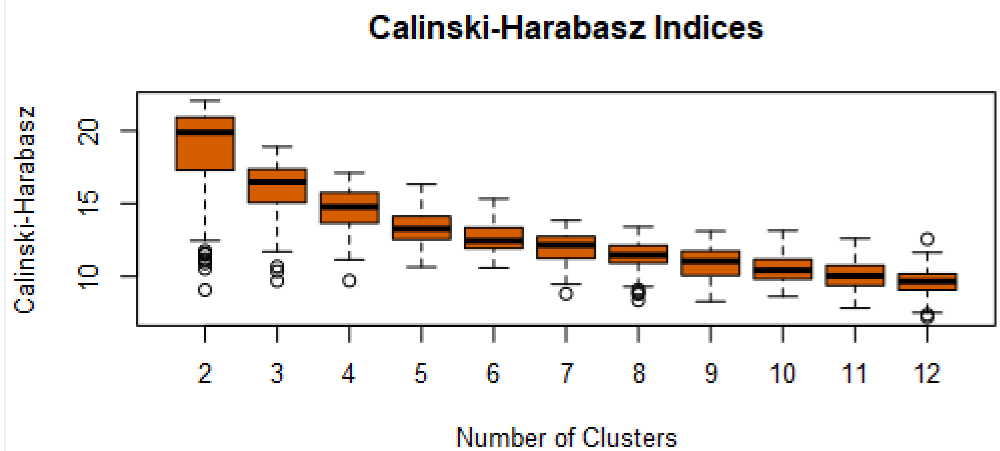
(This was the Capstone project for my Predictive Analytics for Business Nanodegree at Udacity. The project required me to first determine an optimal number of segments, based on product sales, in which to put 85 stores. Then I assigned segments to 10 new stores based on demographic data. Lastly, I forecasted monthly produce sales for all existing and new stores for an entire year.)

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 of formats is three. The AR and CH indices resulting from K-Means assessment are below.





   The three-cluster AR index has the highest median at 0.406. The three-cluster CH index has the second highest median at 16.483, and an interquartile range of 2.3. This IQR is significantly tighter than the two-cluster value of 3.4.

2. How many stores fall into each store format?

   Below is a chart with cluster number and number of stores.

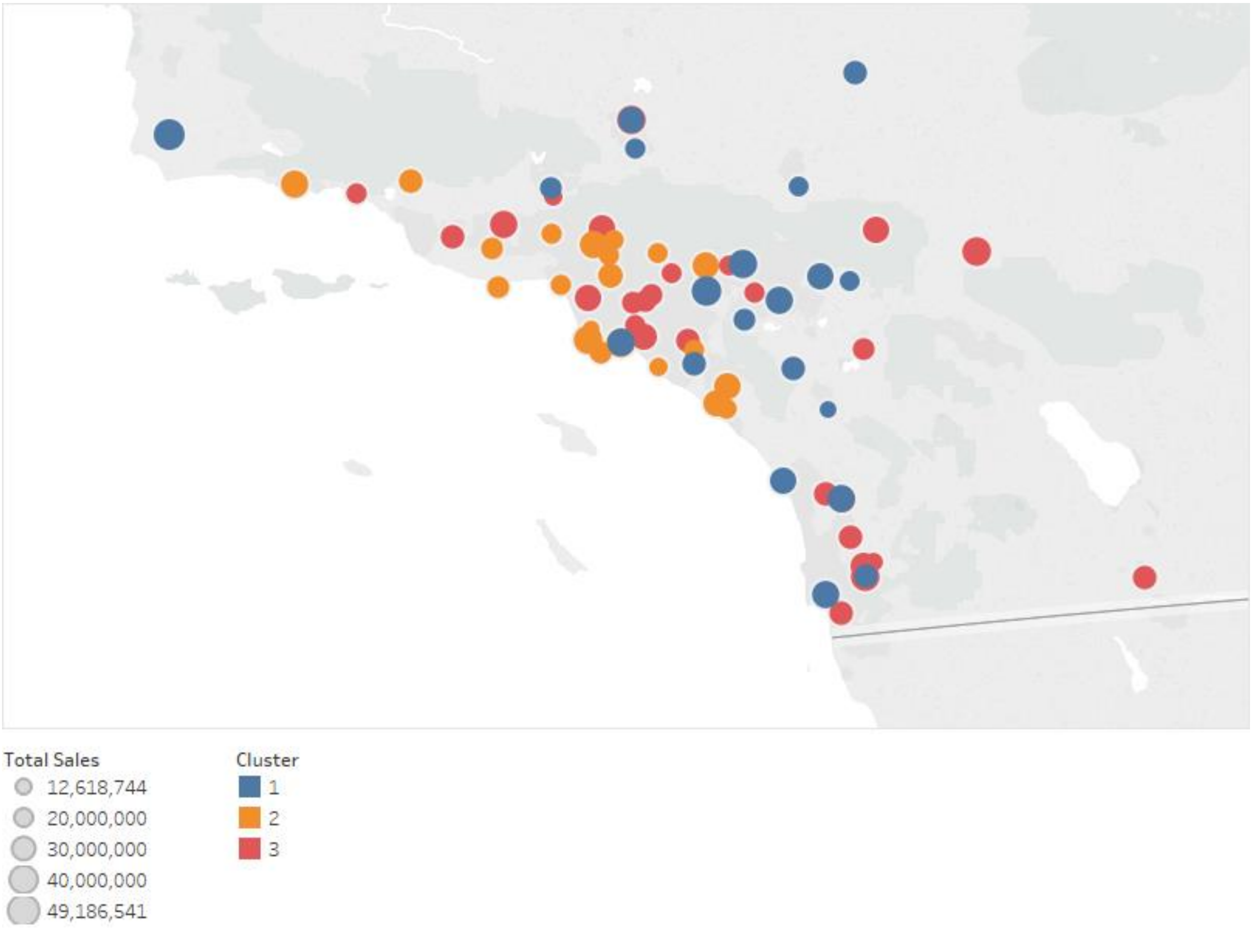
Cluster	Size
1	23
2	29
3	33

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

   Clusters 1 and 2 are farthest apart in their values for the variable Pct_Dairy, indicating that one cluster represents stores with the highest percentage of dairy sales, and the other represents stores with the lowest percentage of dairy sales. The same situation is found in percentage of meat sales in clusters 2 and 3; these clusters fall at the opposite ends of the Pct_Meat variable and represent the stores with the highest and lowest percentage of meat sales.

	Pct_Dry_Grocery	Pct_Dairy	Pct_Frozen_Food	Pct_Meat
1	0.327833	-0.761016	-0.389209	-0.086176
2	-0.730732	0.702609	0.345898	-0.485804
3	0.413669	-0.087039	-0.032704	0.48698

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.



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.)

The Boosted model will be used to predict the best new store formats. While all three models scored an identical 0.8235 in accuracy (please see graphic that follows) the Boosted model had the highest F1, or precision, score at 0.8889. Also, the confusion matrix shows that cluster 1 and cluster 2 stores were correctly identified 100% of the time.

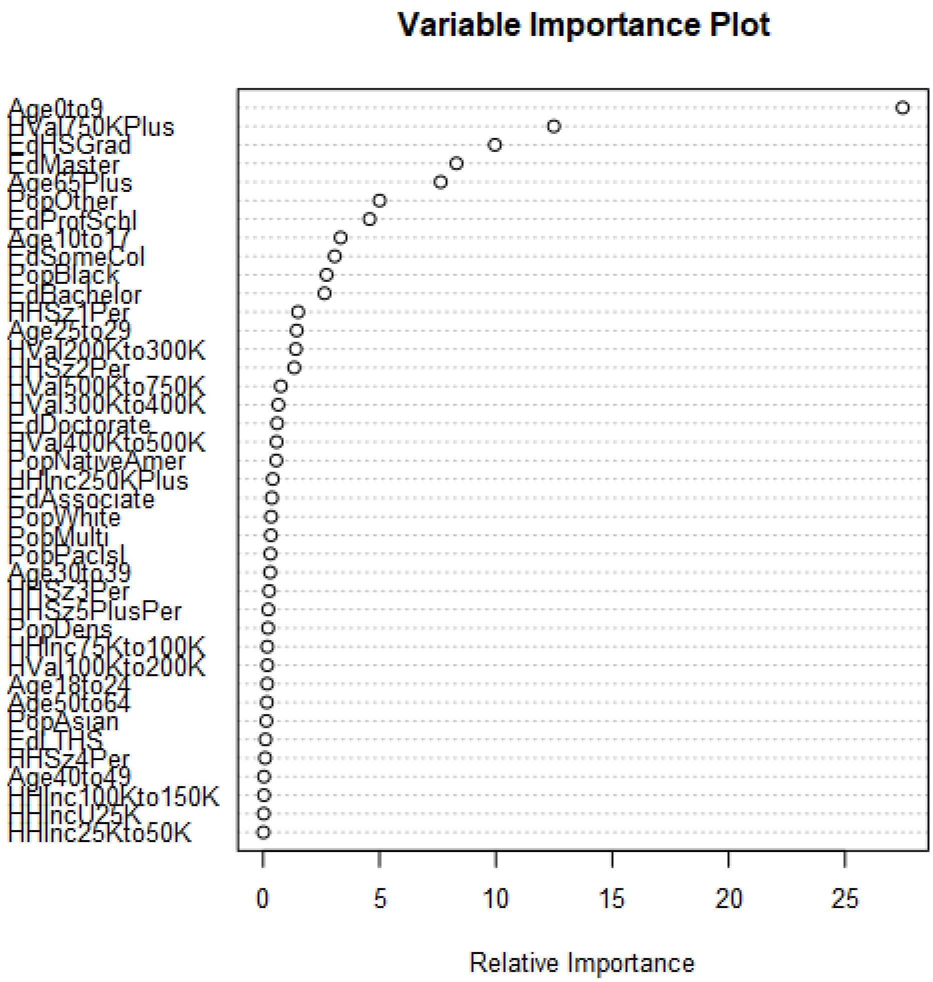
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Tree	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667

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

The three most important variables for explaining the relationship between demographic indicators and store formats are Age0to9, HVal750KPlus, and EdHSGrad, as can be seen at the top of the Variable Importance Plot below.



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

Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2

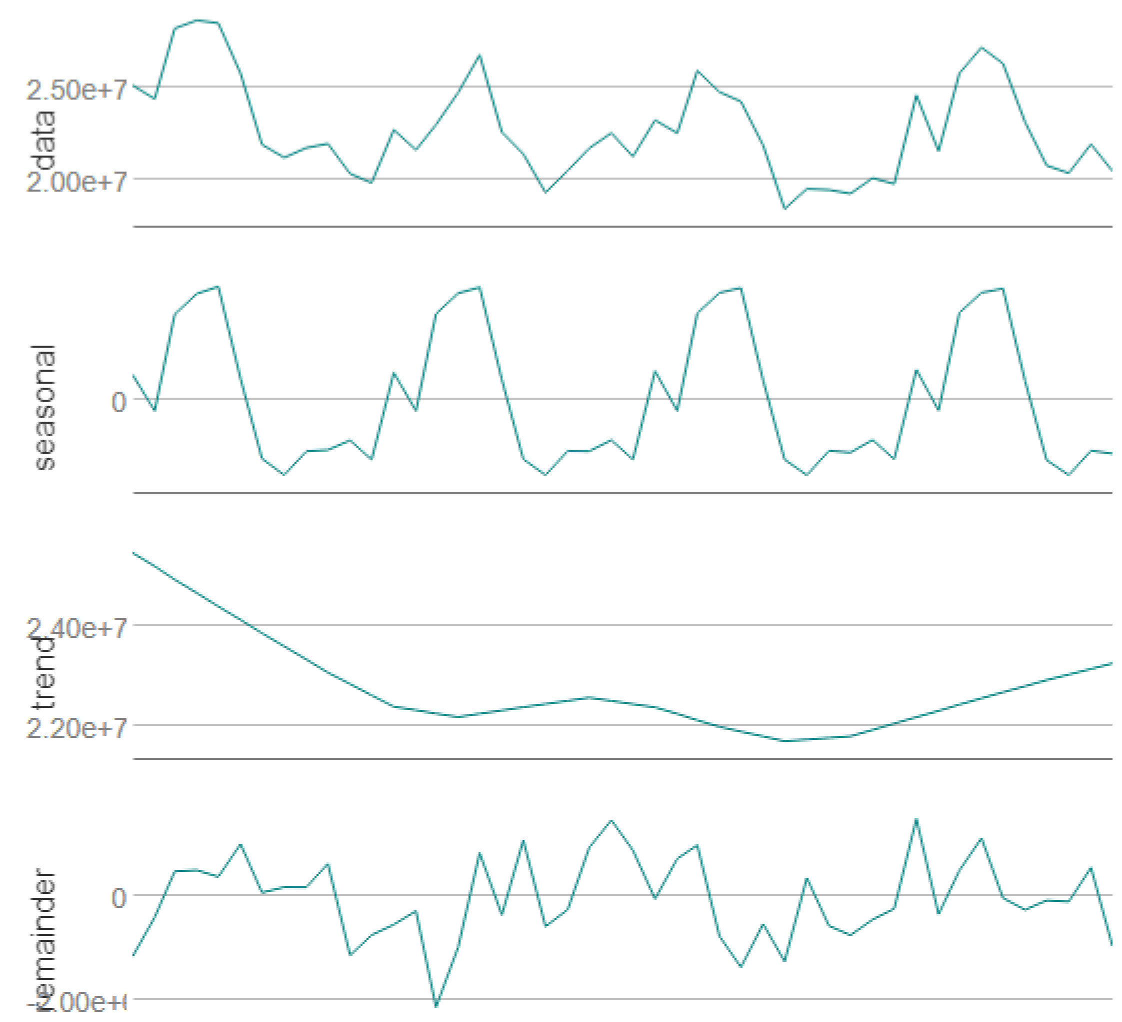
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

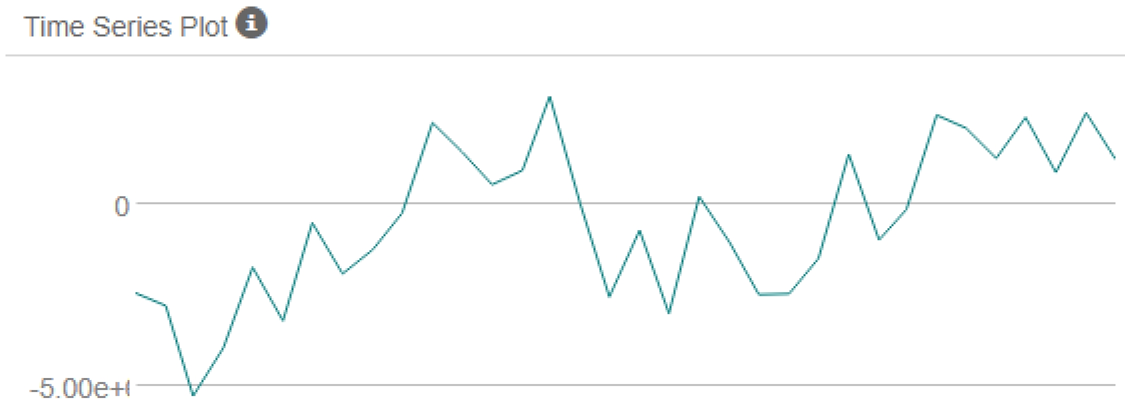
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

Below is the decomposition of the time series plot. The remainder portion has varying magnitude over time, so the error component of the ETS model is applied multiplicatively. The trend moves in both downward and upward directions. As a result, a trend component of none is used. The plot shows seasonality that changes in magnitude over time, and so the seasonality component is applied multiplicatively. Thus, the ETS model is ETS (M,N,M).

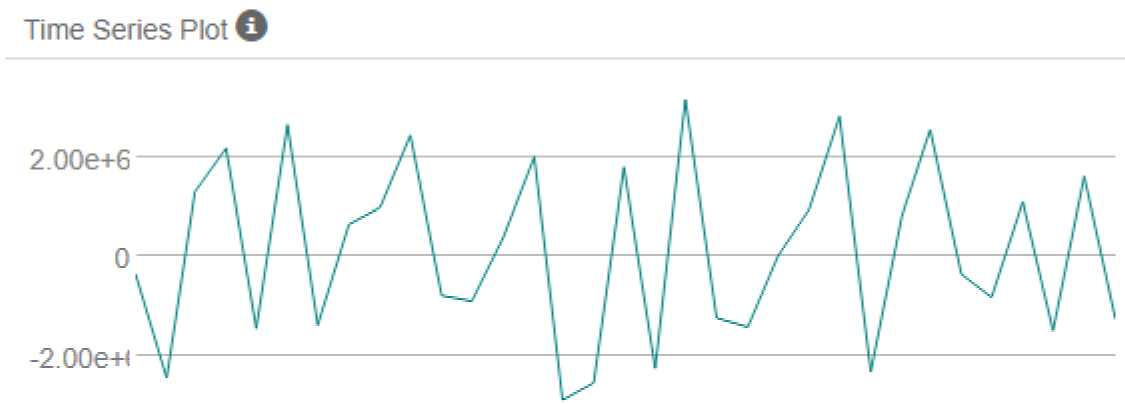
Decomposition Plot



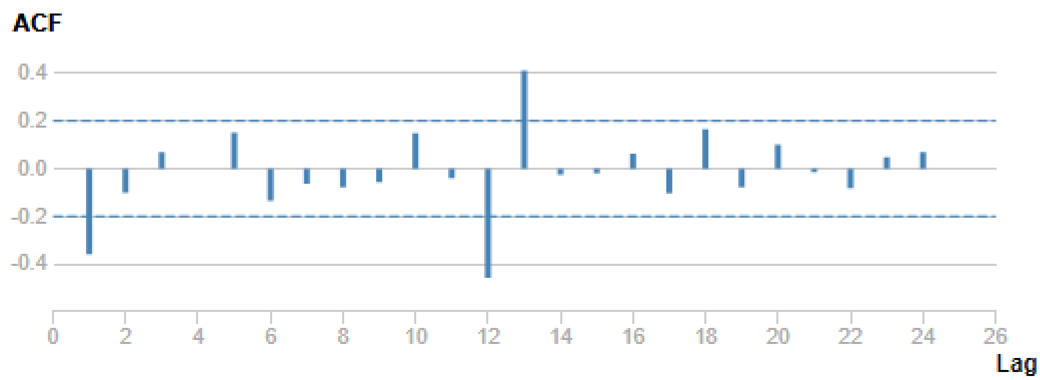
Referring again to the above graphic, seasonality is evident. The ARIMA model will use seasonal differencing to achieve stationarity. Below is the data after seasonal differencing.

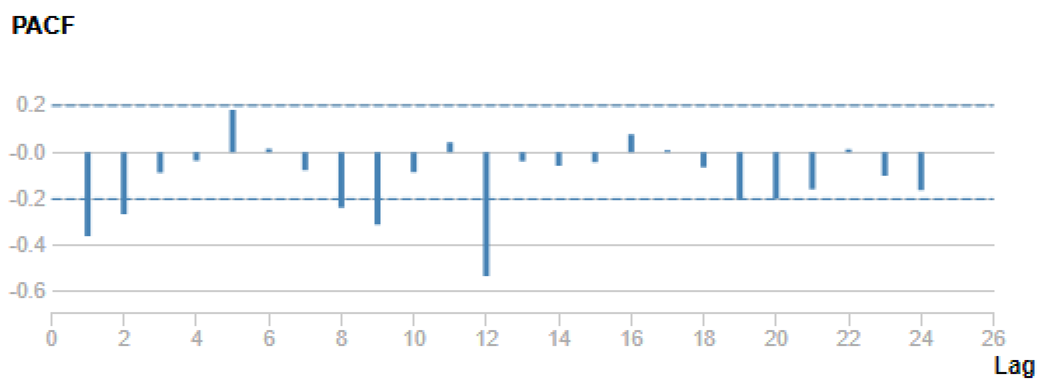


The data is still not stationary. The ARIMA model will difference the data once (please see graphic that follows).



After taking the seasonal difference and first difference, the data is stationary.
The ACF and PACF plots (please find these plots below) provide indicators to complete the ARIMA model. Negative spikes at Lag 1 of the ACF indicate using non-seasonal MA terms. This is confirmed in the PACF. Negative spikes at Lag 12 of the ACF indicate using seasonal MA terms. This is also confirmed by the PACF. The ARIMA model will therefore be ARIMA (0,1,1)(0,1,1).





The ARIMA model has superior internal validation scores versus the ETS model, with a lower RMSE, 935292, lower MASE, 0.351, and a lower AIC, 849.829. Please see the following graphics.

Method: ARIMA(0,1,1)(0,1,1)[12]

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
150815.8641194	935292.1712234	628801.7029024	0.6312352	2.7761535	0.3510153	-0.0469226

Information Criteria:

AIC	AICc	BIC
849.8292	850.8727	853.7167

And the same metrics for the ETS model:

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

External validation, however, tells a different story. The ETS model is superior when compared against the actual data of the holdout sample (please find measurements below). The ETS MASE score is a very good 0.38. In RMSE, a measure of standard deviation, the ETS model beats the ARIMA by nearly 32,000. The ETS (M,N,M) model will be used for the forecast.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_0_1_1__0_1_1_	-492238.8273	792197.3417	735878.1606	-2.1992	3.3098	0.433
ETS_M_N_M_	210494.412	760267.329	649540.846	1.0288	2.9678	0.3822

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.

Monthly produce sales forecasts for 2016, for both existing and new stores, are provided in the graphic below.

Month	New Stores	Existing Stores
Jan-16	2587450.85	21539936.01
Feb-16	2477352.89	20413770.60
Mar-16	2913185.24	24325953.10
Apr-16	2775745.61	22993466.35
May-16	3150866.84	26691951.42
Jun-16	3188922.00	26989964.01
Jul-16	3214745.65	26948630.76
Aug-16	2866348.66	24091579.35
Sep-16	2538726.85	20523492.41
Oct-16	2488148.29	20011748.67
Nov-16	2595270.39	21177435.49
Dec-16	2573396.63	20855799.11

A visualization showing historic produce sales, as well as sales forecasts for existing stores and the 10 new stores, is below.