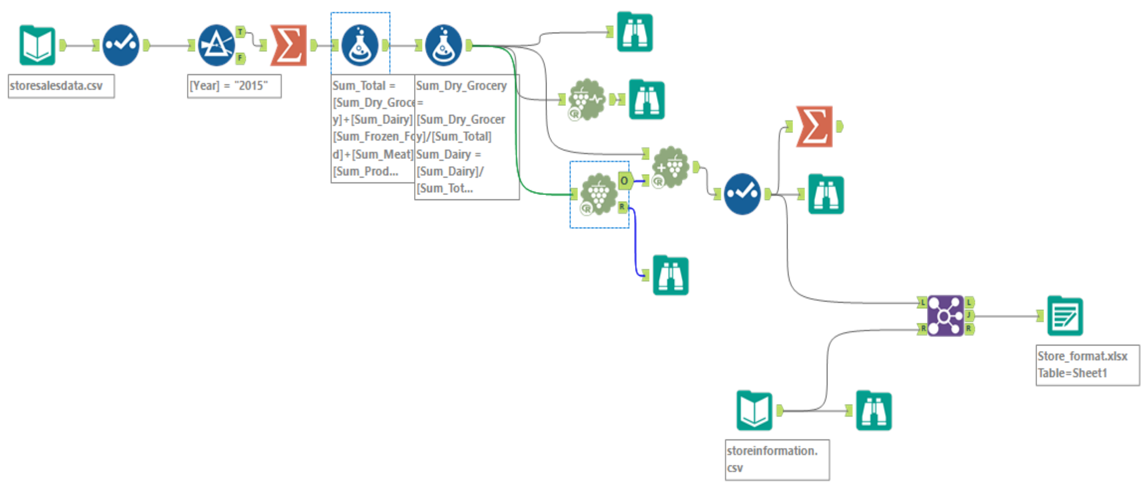
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

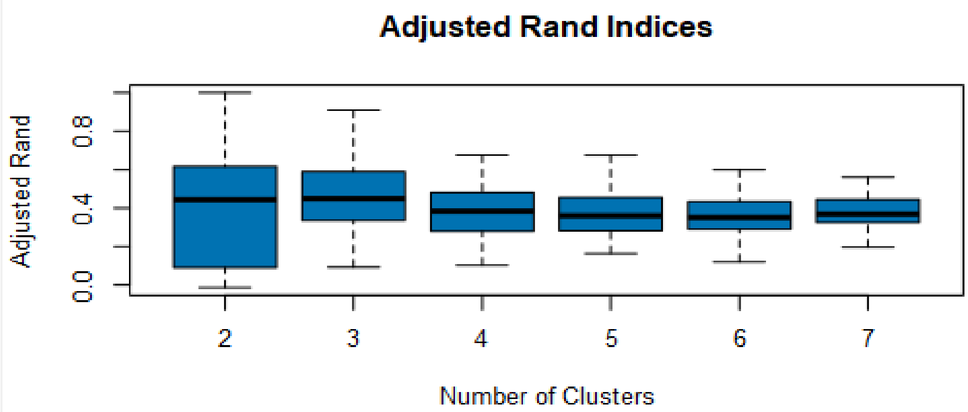
Company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all store use the same store format to sell all their products, which begin to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. I am asked to provide analytical support to make decisions about store formats and inventory planning.

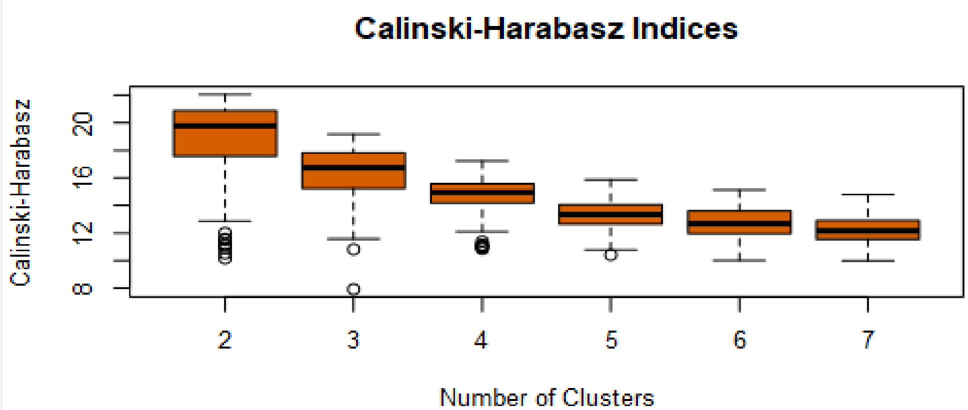


1. What is the optimal number of store formats? How did you arrive at that number?

   I used only 2015 sales data and applied K-means clustering model. Based on K-means cluster assessment report, I picked the final optimal number of store formats to be 3 as it has highest rand in Adjusted Rand Indices chart and highest rand in Calinski_harabasz Indices chart as well.

Plots





2. How many stores fall into each store format?

   Cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores.

Cluster	Count
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?
   Cluster 1 has more general merchandise sales and less Dairy sales
   Cluster 2 has more dairy, frozen food, produce, floral sales and less dry grocery, meat sales.
   Cluster 3 has more meat and deli sales and less general merchandise sales.

Summary Report of the K-Means Clustering Solution CLusterKmean

Solution Summary

Call:

```
stepFlexclust(scale(model.matrix(~-1 + Sum_Dry_Grocery + Sum_Dairy + Sum_Frozen_Food + Sum_Meat + Sum_Produce + Sum_Floral + Sum_Deli + Sum_Bakery + Sum_General_Merchandise,
the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))
```

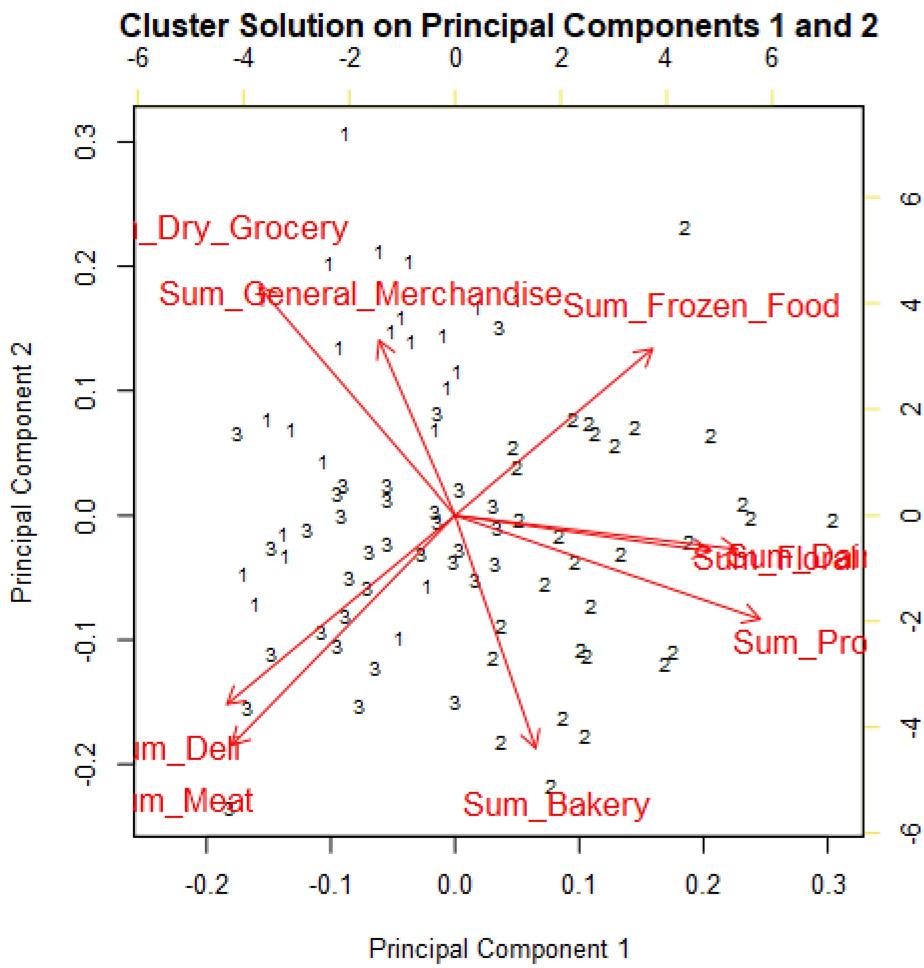
Cluster Information:

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

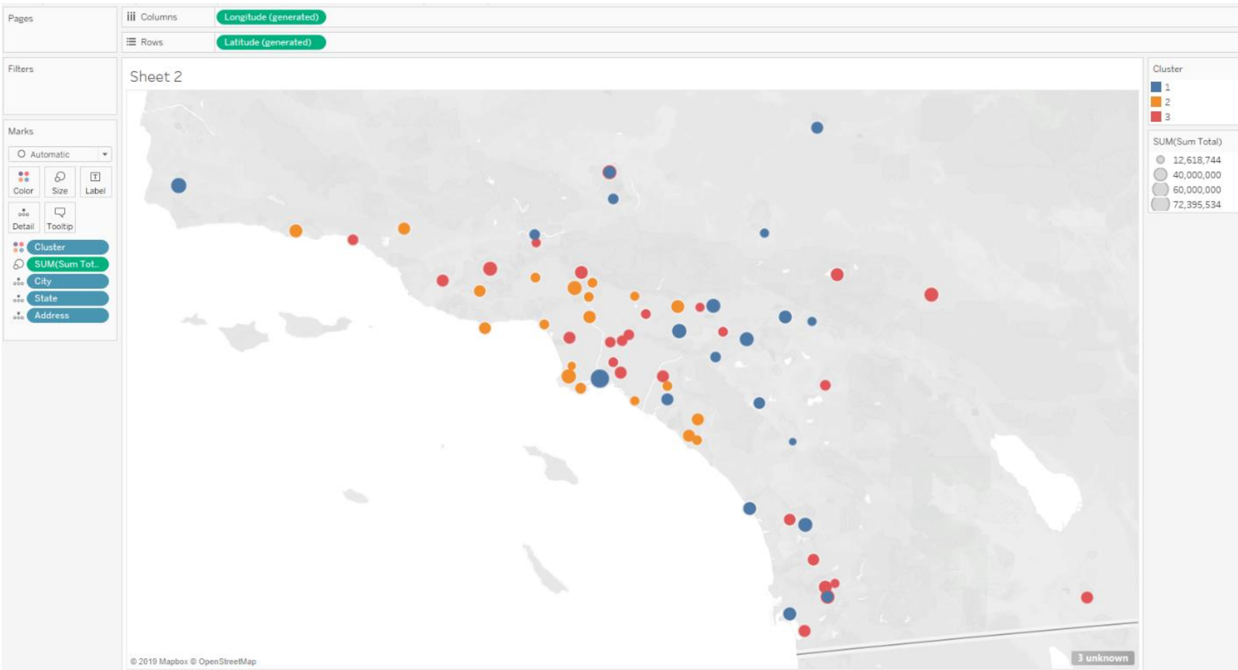
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Produce	Sum_Floral	Sum_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Sum_Bakery	Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389



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

The grocery store chain has 10 new stores opening up at the beginning of the year. I am asked to determine which store format each of the new stores should. Since I don't have sales data for these new stores, we need to determine the format based on demographic data.

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 model comparison report below shows the comparison matrix of Decision Tree, Forest Model and Boosted Model. From the report below, we can see that Boosted Model has the highest accuracy and F1 score. Thus, Boosted model is the best model to apply for the prediction.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Storeformat	0.7059	0.7685	0.7500	1.0000	0.5556
Forest_Storeformat	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Storeformat	0.8235	0.8889	1.0000	1.0000	0.6667

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_Storeformat

	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_Storeformat

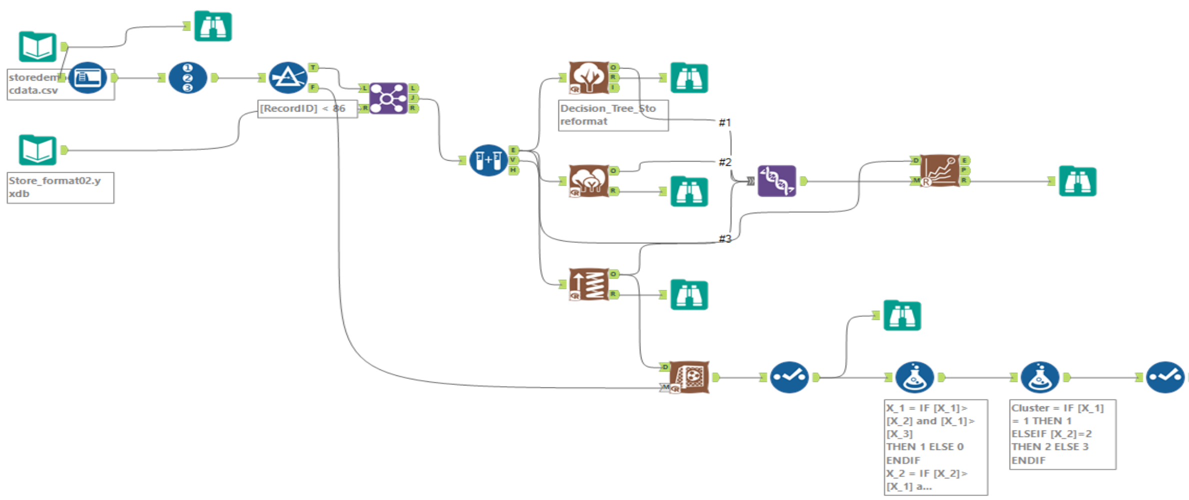
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest_Storeformat

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

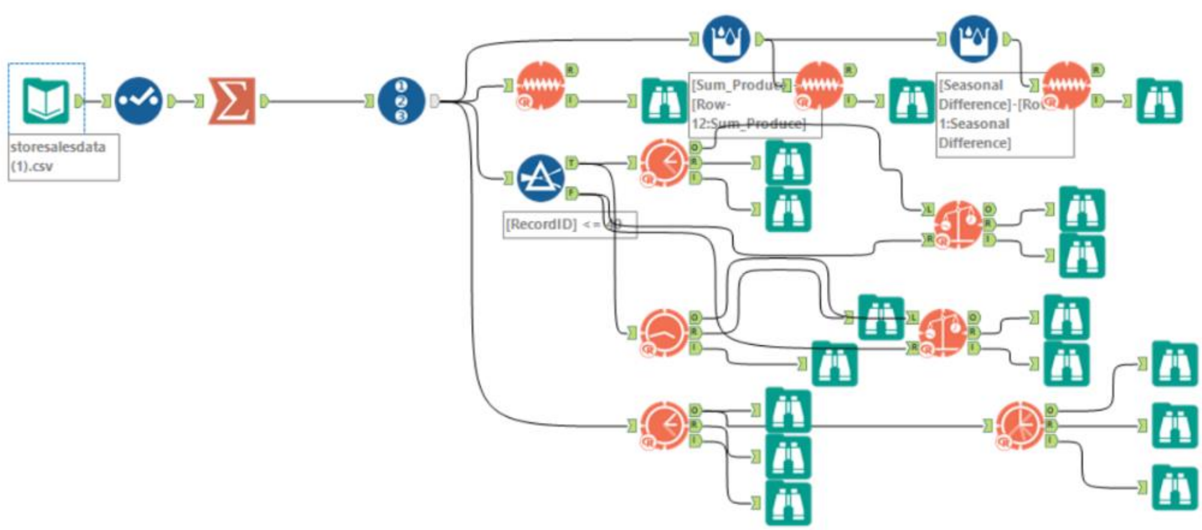
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

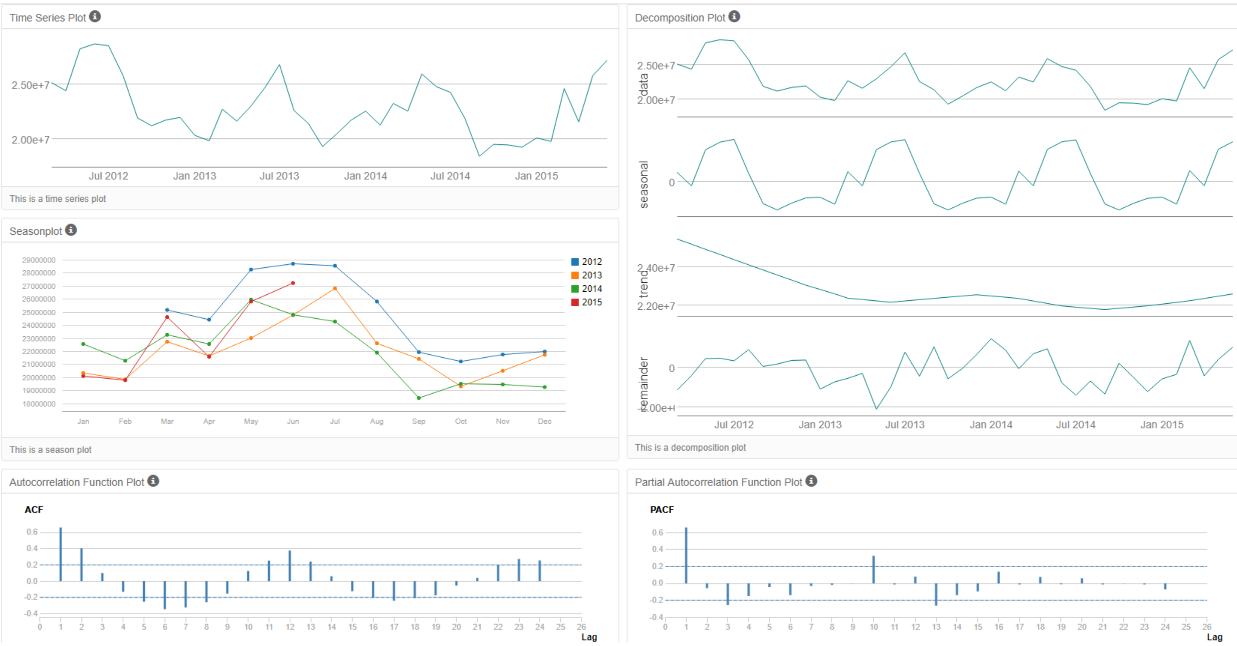


Task 3: Predicting Produce Sales

I am asked to predict the accurate monthly sale forecast.



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?



ETS(M,N,M) with no dampening should be used for ETS model.

The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. The error chart is irregular and should be applied multiplicatively as well.

ARIMA(0,1,2)(0,1,0) is set to calculate elements automatically.

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-139306.2170395	1015013.0030859	880603.2984812	-0.8016736	3.8853672	0.4692243	0.142915

Information criteria:

AIC	AICc	BIC
1089.6723	1116.3389	1112.5677

Summary of ARIMA Model ARIMA_Produce

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

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1
Value	0.663132
Std Err	0.15945

sigma^2 estimated as 3109287776725.33: log likelihood = -347.41299

Information Criteria:

AIC	AICc	BIC
698.826	699.4576	701.0081

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-266968.7825838	1385800.2923691	961223.1598628	-1.2966978	4.3808852	0.5121821	-0.1664469

ETS model's accuracy is higher when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of 1,020,597 is lower than ARIMA's 1,429,296 while its MASE is 0.45 compared to ARIMA's 0.53. ETS also has a higher AIC at 1,283 while ARIMA's AIC is 859.

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:

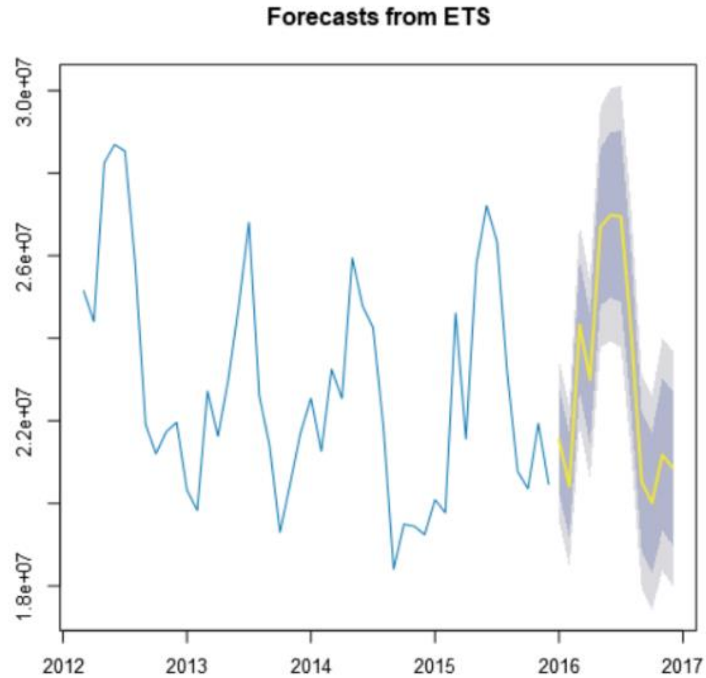
AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

2 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.

Please see below forecast table for existing and new stores.

Period	Month	Existing stores Sales	New stores Sales	Total Produce Sales
2016	1	21539936.01	2534110.119	24074046.13
2016	2	20413770.6	2401620.071	22815390.67
2016	3	24325953.1	2861876.835	27187829.93
2016	4	22993466.35	2705113.688	25698580.04
2016	5	26691951.42	3140229.579	29832181
2016	6	26989964.01	3175289.884	30165253.89
2016	7	26948630.76	3170427.149	30119057.91
2016	8	24091579.35	2834303.453	26925882.8
2016	9	20523492.41	2414528.519	22938020.93
2016	10	20011748.67	2354323.373	22366072.04
2016	11	21177435.49	2491462.998	23668898.48
2016	12	20855799.11	2453623.425	23309422.53

Tableau Visualization