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 of store formats is 3. I arrived this number by using K-Centroids Cluster analysis and K-Centroids Diagnostics Tool with K-Mean Clustering Method. According to K-Mean analysis or below report, both Adjusted Rand Indices and Calinski-Harabasz Indices shows highest mean value at 2 and 3 indicating that the optimal number of stores formats is 3 (but we can also take 2 stores).

K-Means Cluster Assessment Report

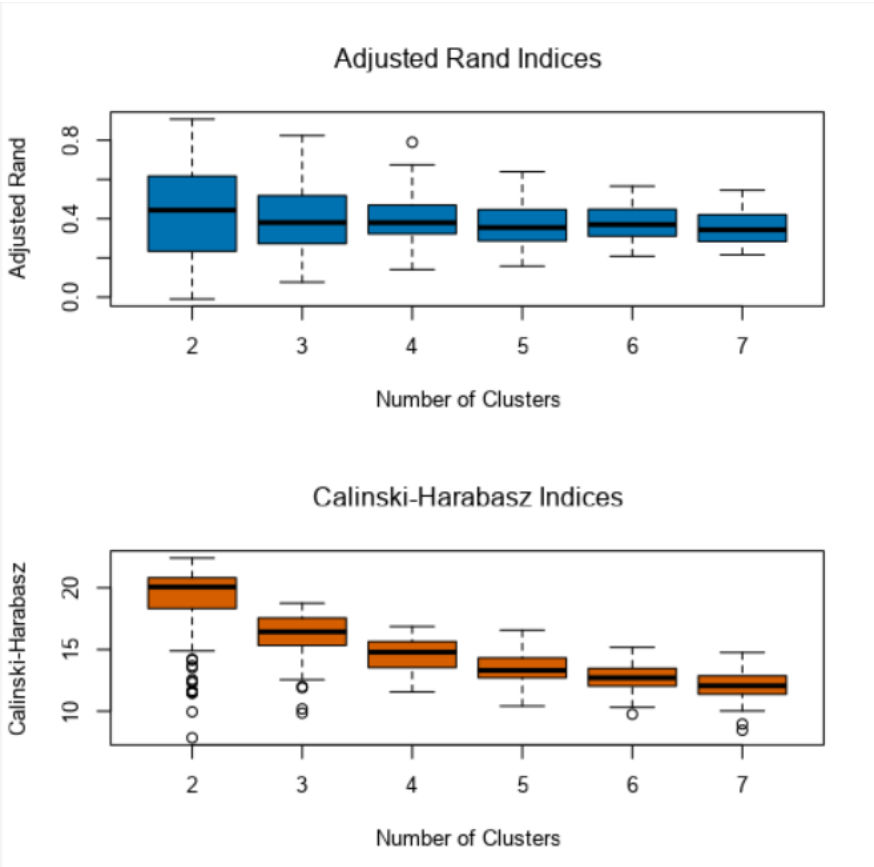
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	-0.009675	0.076235	0.140656	0.157999	0.208442	0.215517
1st Quartile	0.237245	0.273359	0.324062	0.28911	0.310322	0.283793
Median	0.443127	0.379958	0.379205	0.354445	0.369622	0.343121
Mean	0.42889	0.410693	0.396973	0.372638	0.379017	0.355602
3rd Quartile	0.607523	0.513414	0.465973	0.444893	0.445965	0.419453
Maximum	0.907005	0.823811	0.789549	0.639632	0.565878	0.54505

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	7.838511	9.845155	11.56778	10.41516	9.754192	8.452392
1st Quartile	18.329049	15.370633	13.54646	12.70601	12.042498	11.37346
Median	20.072097	16.43124	14.78233	13.31046	12.703326	12.062457
Mean	18.866108	16.214792	14.61573	13.44934	12.742937	12.071297
3rd Quartile	20.790946	17.532122	15.63393	14.31965	13.470937	12.865362
Maximum	22.415549	18.750421	16.86351	16.57168	15.173243	14.756313



2. **How many stores fall into each store format?**

   According to Cluster Information: Cluster 1 has 25 stores; Cluster 2 has 35 stores and Cluster 3 has 25 stores.

Summary Report of the K-Means Clustering Solution Cluster

Solution Summary

Call:

stepFlexclust(scale(model.matrix(~-1 + Percent_Dry_Grocery + Percent_Dairy + Percent_Frozen_Food + Percent_Meat + Percent_Produce + Percent_Floral + Percent_Deli + Percent_Bakery + Percent_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

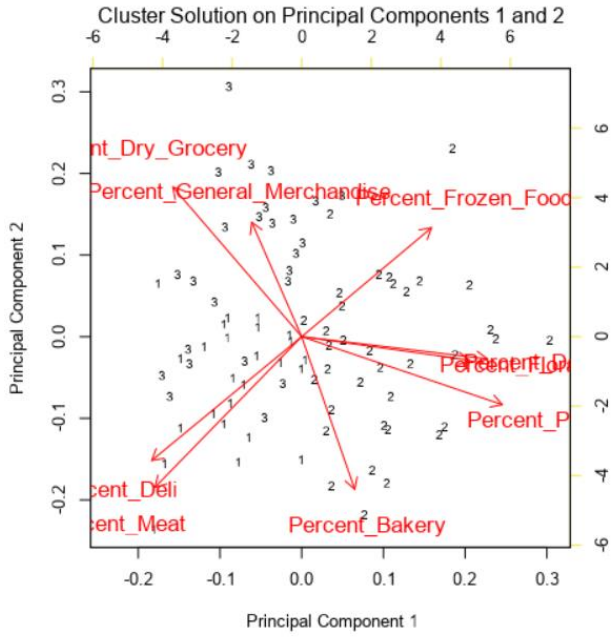
Convergence after 8 iterations.

Sum of within cluster distances: 196.35034.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178481

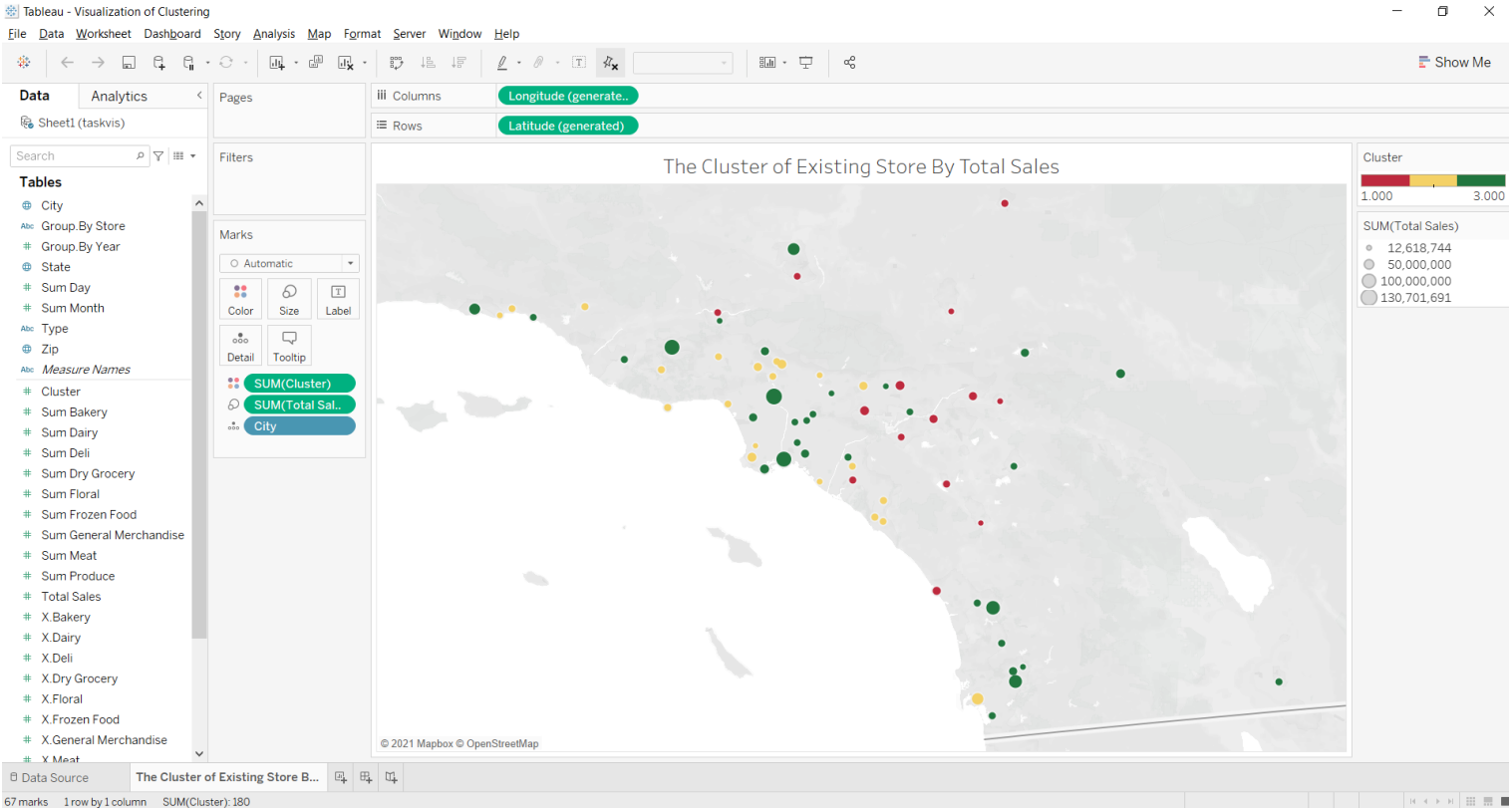
	Percent_Bakery	Percent_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

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



From the report of K-Mean Clustering, we can see the cluster 1 and 3 both of them have same size of 25. Cluster 1 has the highest Max Distance which is 4.82 and have lowest Ave Distance of 2.09 with highest separation of 2.19. Cluster 2 has the largest size of 35 with Highest Ave Distance of 2.47, Max Distance of 4.41 and Separation of 1.94. Similarly, Cluster 3 has Ave Distance of 2.28, Max Distance of 3.58 and Separation of 1.75.

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.)**

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
RF	0.7059	0.7500	0.5000	1.0000	0.7500
BT	0.7059	0.7500	0.5000	1.0000	0.7500
DT	0.6471	0.6667	0.5000	1.0000	0.5000

Confusion matrix of BT

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

Confusion matrix of DT

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

Confusion matrix of RF

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

In any classification problem we will need to set an estimation sample 80 and a

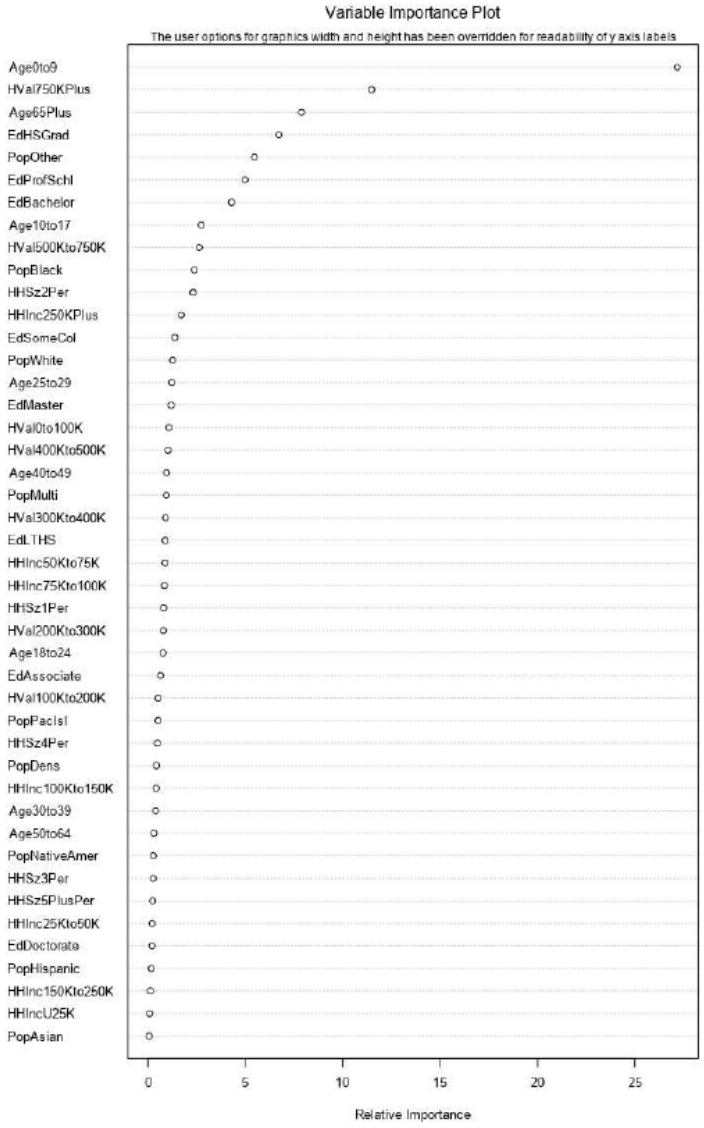
validation sample 20 of my data. This helps us compare different classification models to see which better fit the data. Both Random Model and Boosted Model. I going to use Boosted model because dataset is small in number so that we can run the boosted model in short time. The comparison result made me choose boosted model has the best result in Accuracy = 70.59%, F1 = 75% and Accuracy_1 = 50%.

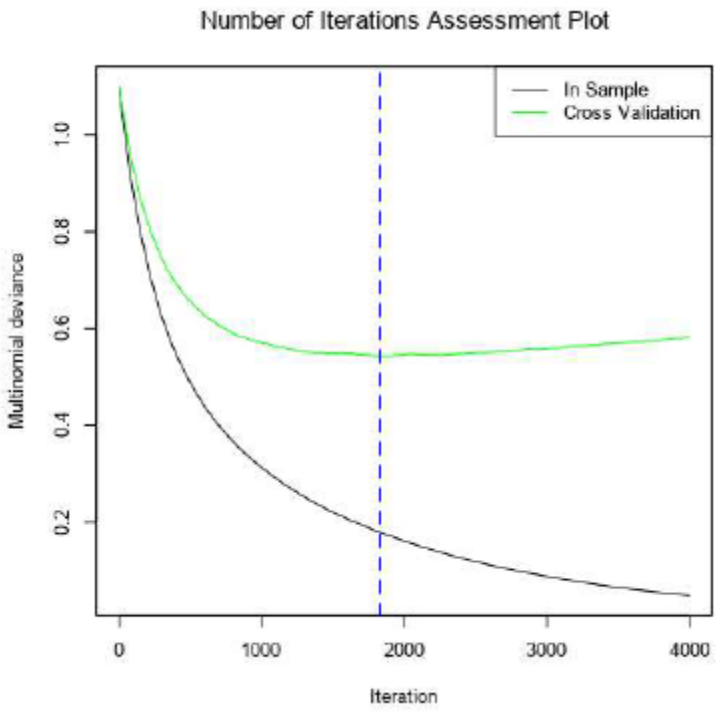
1

Report for Boosted Model BT

Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829





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

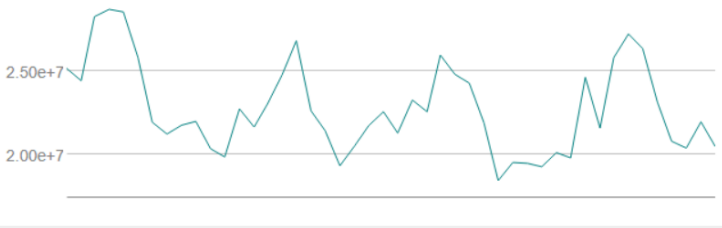
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

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?

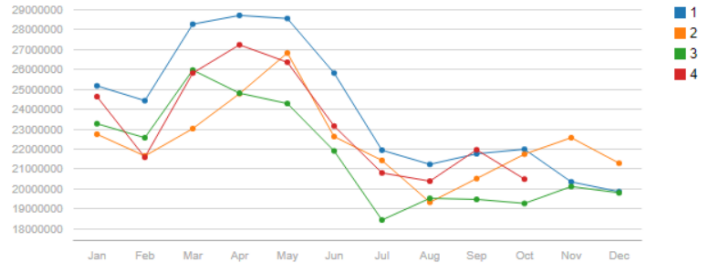
I have used ETS model for forecast. I have come to this decision by comparing between ETS and ARIMA and using TS plot tool. From below Decomposition plot, I have seen the error is multiplicative, the trend is non-exciting and seasonality has an increase trend and multiplicative as the perks change over time. So, I have chosen the ETS model.

Time Series Plot



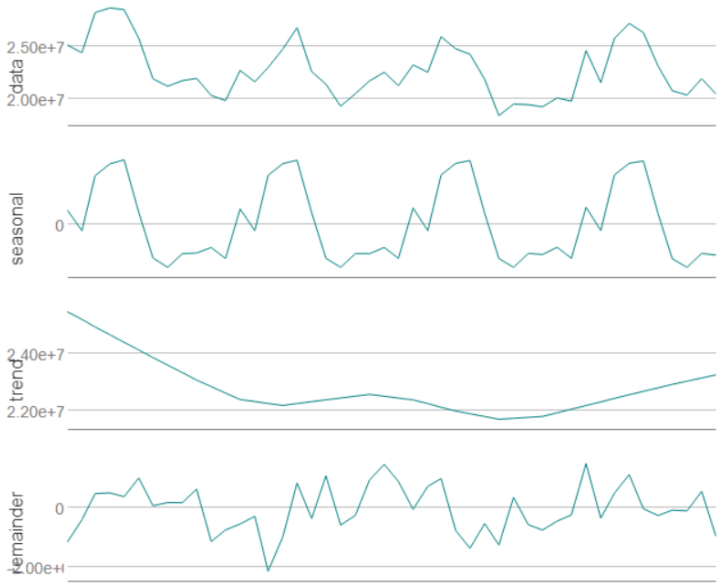
This is a time series plot

Seasonplot



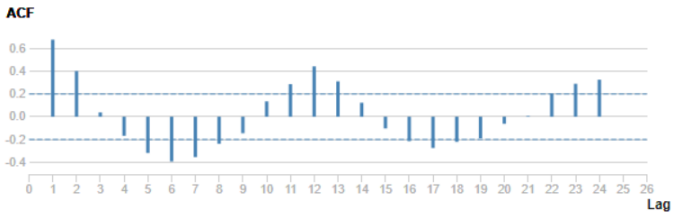
This is a season plot

Decomposition Plot



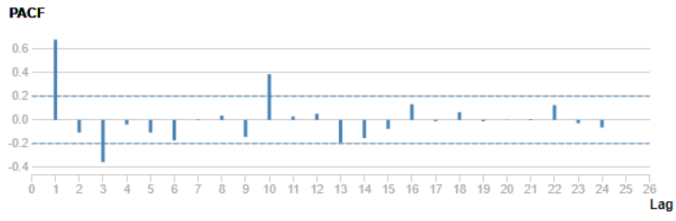
This is a decomposition plot

Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot



This is an partial autocorrelation plot

Summary of ARIMA Model ARIMA

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

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

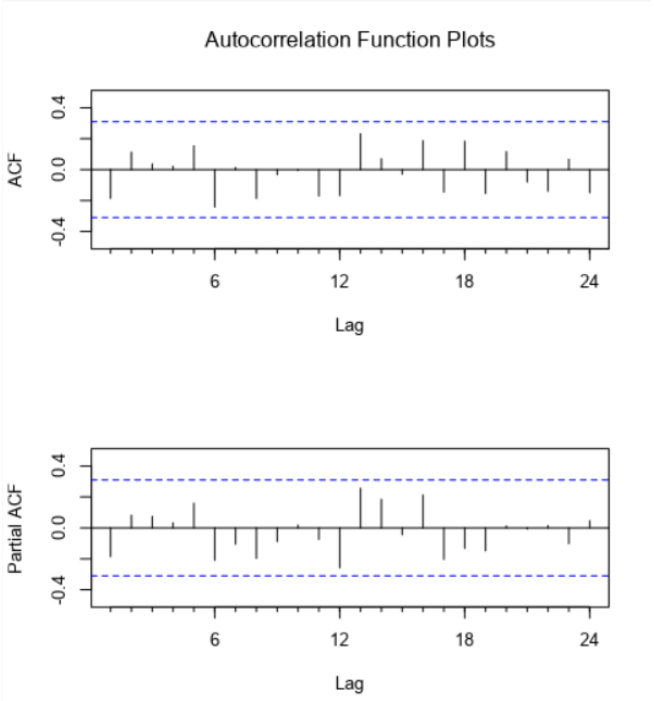
Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616



Summary of Time Series Exponential Smoothing Model ETS1

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-14783.6612202	1044018.8940828	809742.8924252	-0.2664397	3.5527937	0.4555978	0.3283229

Information criteria:

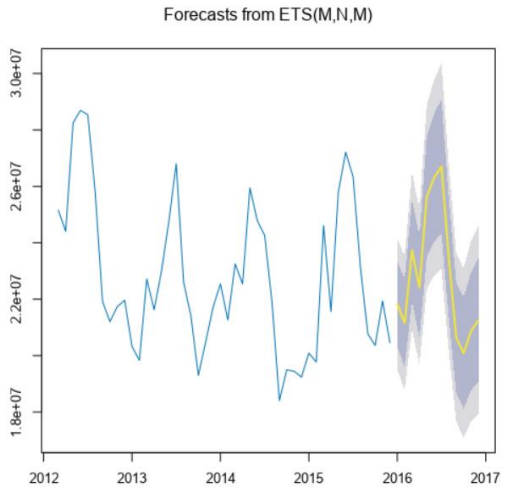
AIC	AICc	BIC
1479.4048	1495.4048	1506.8344

Smoothing parameters:

Parameter	Value
alpha	0.327727
gamma	0.001656

Initial states:

State	Value
l	23159664.744847
s0	0.926093
s1	0.956024
s2	0.930877
s3	0.91335
s4	0.879554
s5	0.903808
s6	1.02648
s7	1.169472
s8	1.151996
s9	1.121918
s10	0.981225



The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

Record Report

1

Comparison of Time Series Models

2

Actual and Forecast Values:

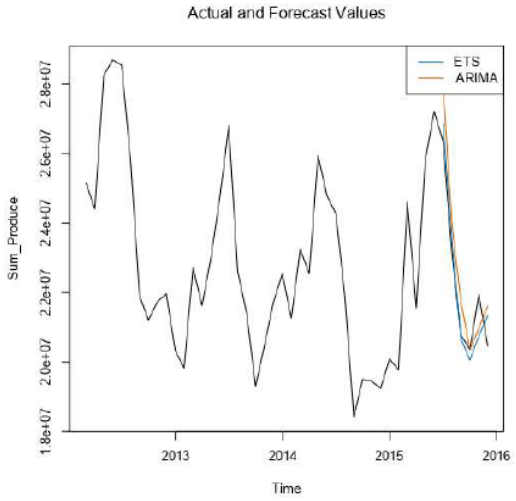
Actual	ETS	ARIMA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

3

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

4



From the Actual vs. Forecast Values for Arima and ETS plots above, I can see the forecast values by the ETS model is most near to the actual values than the forecast values by the Arima model.

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.

The forecasted values for produce, monthly in 2016 for new and existing stores, table down shows the historical data together with these forecasts.

Record	Year	Month	Forecast_sales_existing_stores	Forecast_Sales_new_Stores	Total_Forecast_sales	Date
1	2016	1	21829060.031666	2491319.093207	24320379.124873	2016-1
2	2016	2	21146329.631982	2408384.783604	23554714.415586	2016-2
3	2016	3	23735686.93879	2833157.321387	26568844.260177	2016-3
4	2016	4	22409515.284474	2679433.371626	25088948.6561	2016-4
5	2016	5	25621828.725097	3054885.876482	28676714.601579	2016-5
6	2016	6	26307858.040046	3106151.779247	29414009.819294	2016-6
7	2016	7	26705092.556349	3132699.144598	29837791.700947	2016-7
8	2016	8	23440761.329527	2776154.195458	26216915.524985	2016-8
9	2016	9	20640047.319971	2451565.941438	23091613.261409	2016-9
10	2016	10	20086270.462075	2401771.574835	22488042.03691	2016-10
11	2016	11	20858119.95754	2477301.916348	23335421.873888	2016-11
12	2016	12	21255190.244976	2452170.069396	23707360.314372	2016-12