Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

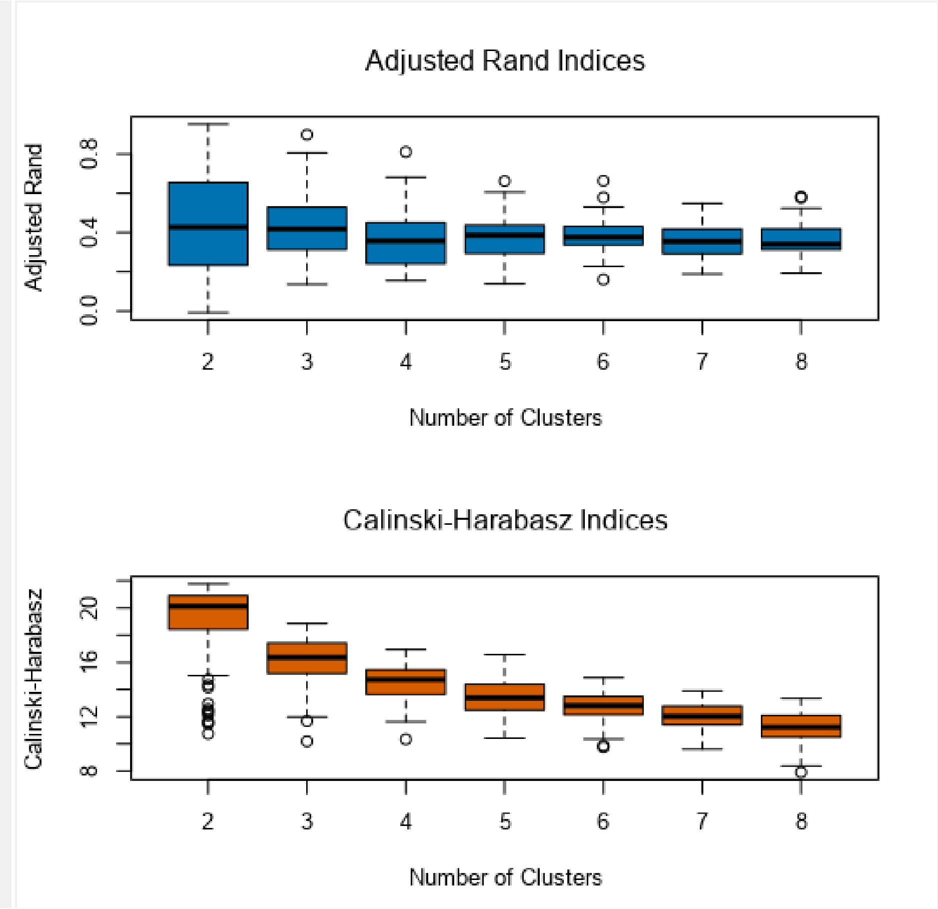
1. **What is the optimal number of store formats? How did you arrive at that number?**

The company store formats had always been similar to one another in terms of the same products and layouts. To achieve the different store formats introduced by the company, a segmentation and clustering analysis will be applied, where each store format will have different products matched to the local demand.

The optimal number of store formats derived is three.

The three store formats was derived using clustering and segmentation analysis. The stores sales data was used to run the analysis in Alteryx. Using the StoreSalesData.csv and StoreInformation.csv files, I summed the sales data by Store ID and Year, used the percentage sales per category per store for the clustering while filtering only the 2015 sales data. The K-means clustering model was used which involves tools like K-centroid diagnostics, K-centroids cluster analysis and the append cluster tool.

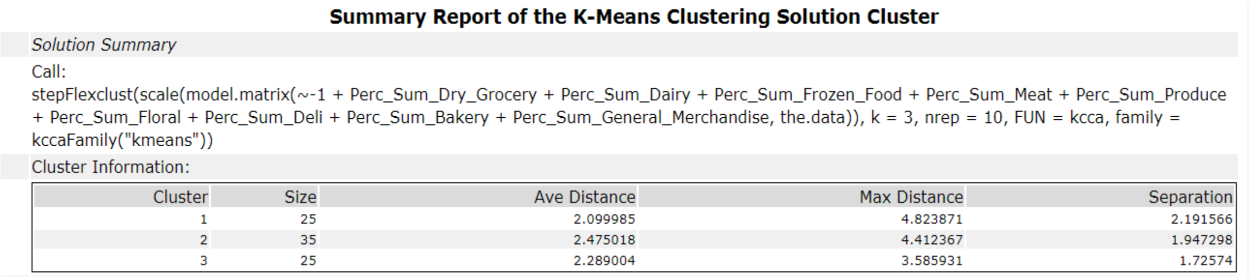
To determine the optimal number of clusters, the Rand and CH Indices was generated using K-centroid diagnostics tool which indicated that 3 clusters are optimal due to the high median/mean value, as well as the compactness, and fewer outliers indicated. See the box plots of the Rand and CH indices below that ascertained the conclusion:



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

For each store format; **25 stores** was is in **cluster 1**, **35 stores** in **cluster 2**, and **25 stores** in **cluster 3**.

See the table below:



Summary Report of the K-Means Clustering Solution Cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Perc_Sum_Dry_Grocery + Perc_Sum_Dairy + Perc_Sum_Frozen_Food + Perc_Sum_Meat + Perc_Sum_Produce + Perc_Sum_Floral + Perc_Sum_Deli + Perc_Sum_Bakery + Perc_Sum_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

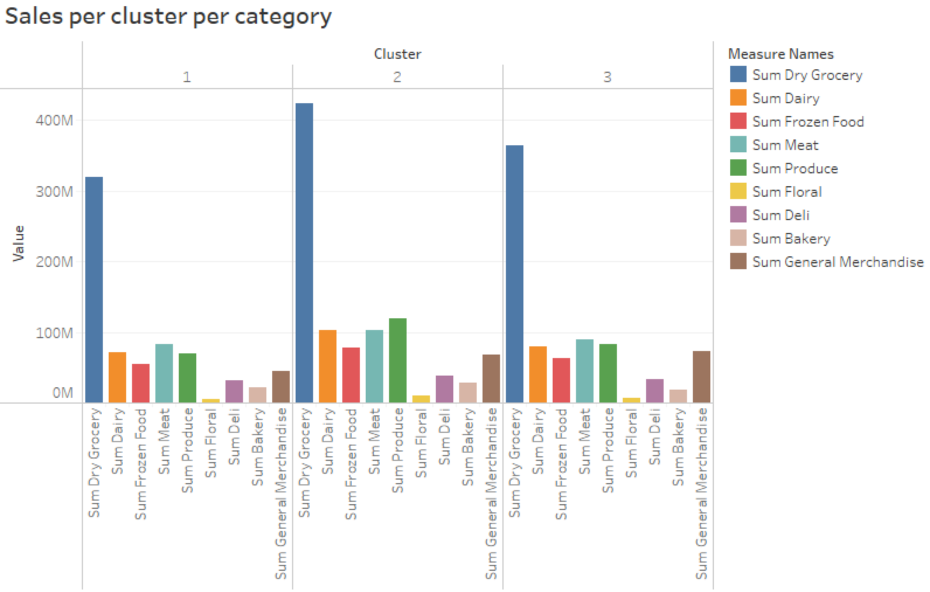
Cluster Information:

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

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

Based on the observations from the result of the model, cluster 1 has more of stores with Dry grocery, meat, deli and bakery products. Cluster 2 has a larger number of stores in it, with major products in dairy, frozen food, produce and floral. Cluster 3 has more stores that contain general merchandise.
Also with the distribution of the cluster sizes, we can infer that cluster two store had higher number of sales compared to the other two clusters.

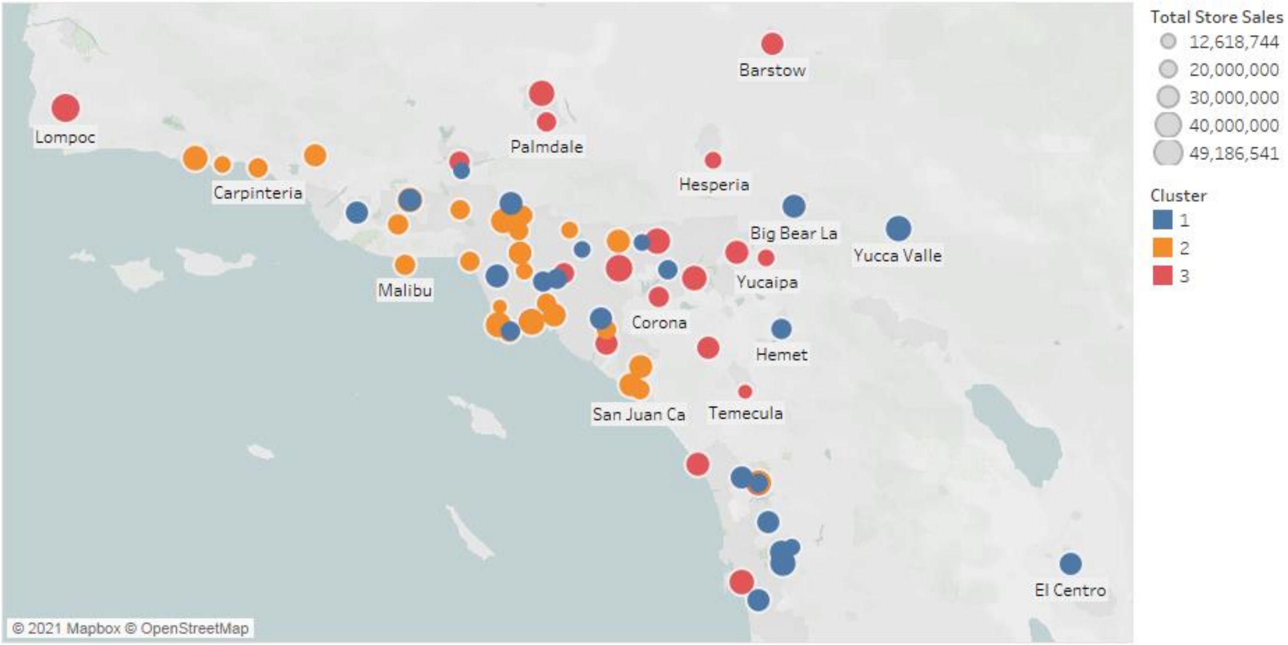
The graph plot below shows the sales of each cluster per category indicating difference in cluster sales:



Sum Bakery, Sum Dairy, Sum Deli, Sum Dry Grocery, Sum Floral, Sum Frozen Food, Sum General Merchandise, Sum Meat and Sum Produce for each Cluster. Color shows details about Sum Bakery, Sum Dairy, Sum Deli, Sum Dry Grocery, Sum Floral, Sum Frozen Food, Sum General Merchandise, Sum Meat and Sum Produce.

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 1 - The location of the stores in each cluster, size, and total sales



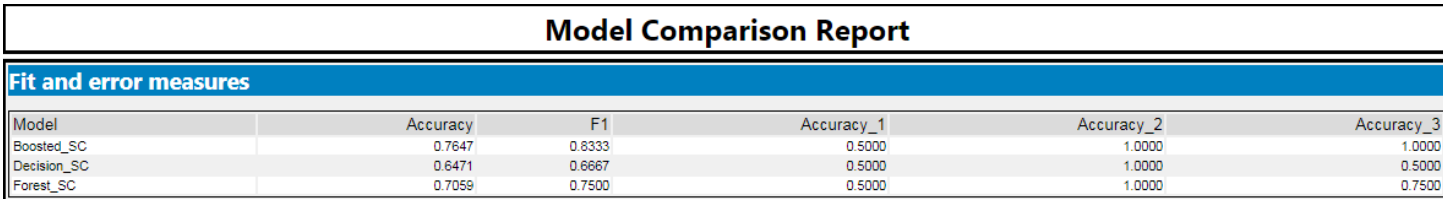
Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows Total Store Sales. The marks are labeled by City. Details are shown for Zip. The view is filtered on Cluster, which keeps 1, 2 and 3.

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**

To predict the best store format for the ten new stores a non-binary classification model was performed using the *StoreDemographicData.csv* and the output of segmented existing stores in Task 1. The models used to test where Boosted Model, Decision Tree, and Forest Model; the accuracy of the models was tested with a 20% validation sample and a random seed = 3. Boosted model had the best performance with a higher F1 accuracy at predicting the best store format for the new stores.

The table below shows a report of the tested models:



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_SC	0.7647	0.8333	0.5000	1.0000	1.0000
Decision_SC	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_SC	0.7059	0.7500	0.5000	1.0000	0.7500

From the table above the Boosted Model had an Accuracy **(0.76)** and F1 Score **(0.83)** higher than the Decision and Forest Models. The boosted model was able to predict the best store format for the new stores as seen in the table in question two.

2. What format do each of the 10 new stores fall into?

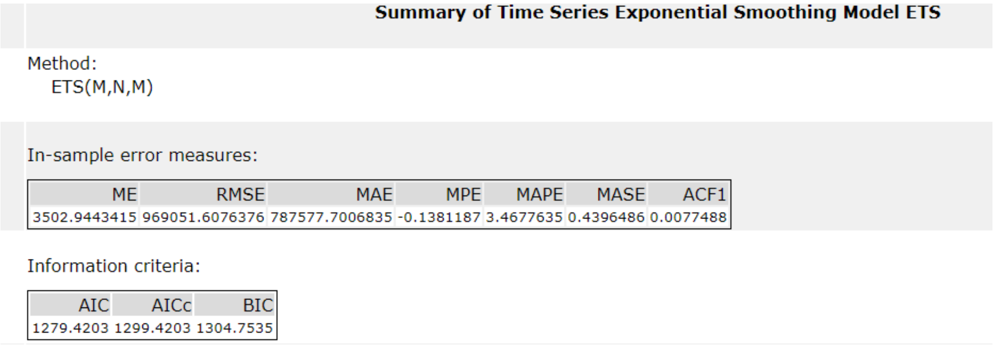
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Task 3: Predicting Produce Sales

1. What type of ETS or ARIMA model did you use for each forecast? How did you come to that decision?

A monthly forecast of produce sales for the full year of 2016 for both existing and new stores was done. To determine the best forecasting model, ETS and ARIMA model was tested on the sum of produce sales in the StoreSalesData.csv and validated using the forecast error measurements against the holdout sample.

Below is a table of the ETS Model Forecast Error Measurement:



Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

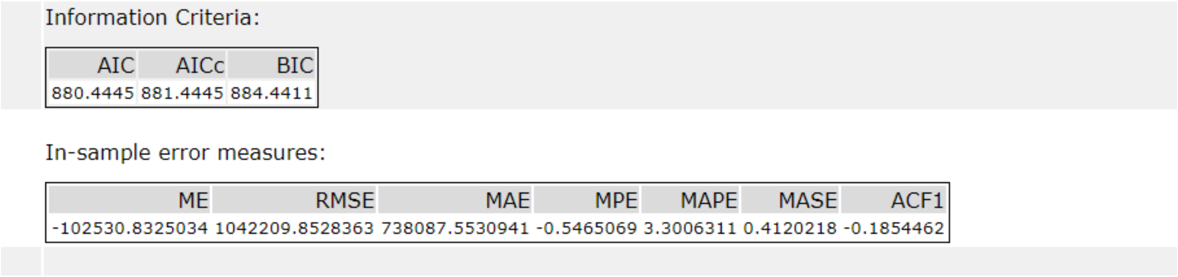
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

A table of the ARIMA Model Forecast Error Measurement:



Information Criteria:

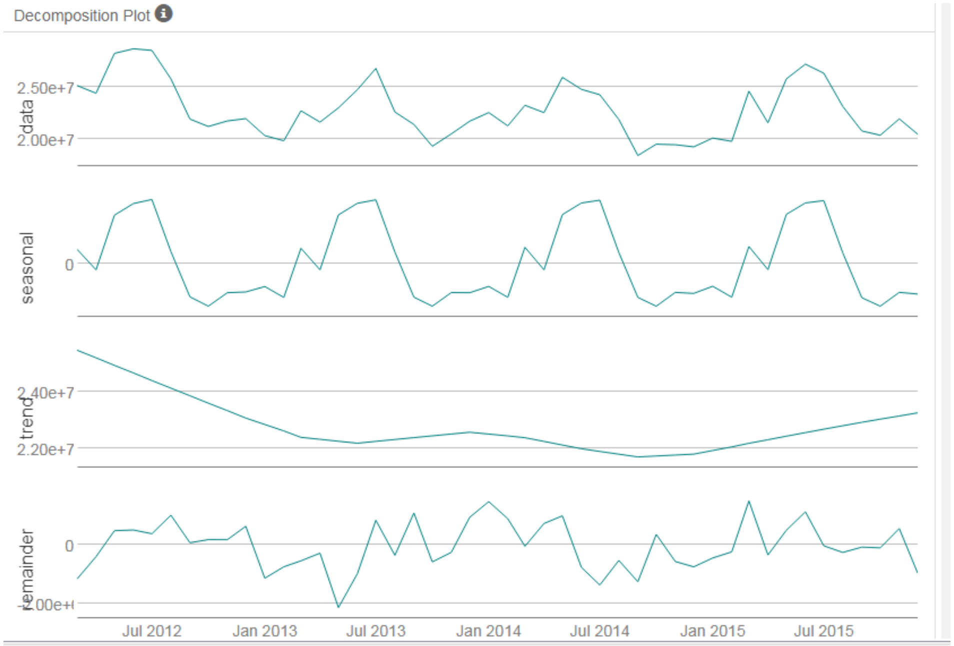
AIC	AICc	BIC
880.4445	881.4445	884.4411

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

The ETS model used for the forecast is the ETS (M,N,M) model, which was based on the results of the decomposition tool derived from running the TS plot tool on the aggregated store sales data.

Below is the Decomposition plot line graph:



From the Decomposition plot above the ETS (M,N,M) model was optimal for the model forecast because:

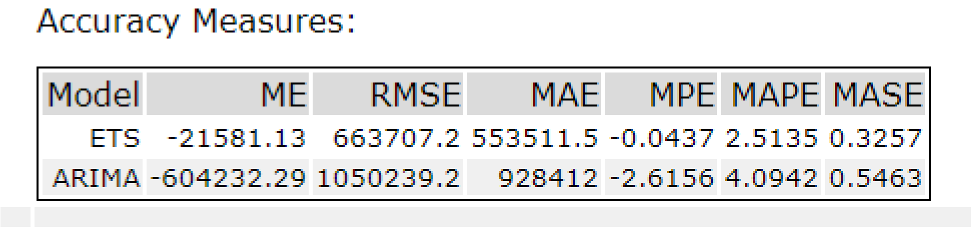
The seasonal data trend was increases and decreases exponentially indicating the use of multiplication **(M)**.

Trend plot had no trend since the trend goes down then up across the plot indicating none **(N)** and

The remainder/error plot fluctuates between the large and small errors across plot indicating multiplication **(M).** The ETS (M,N,M) model was thus selected for the forecast.

The model comparison result showed that the ETS Model was the best model for forecasting the produce sales of the new and existing stores for the next 12 months. The decision was determined based on the measure of accuracy result from comparing the ARIMA and ETS model. The accuracy measure showed that ETS model had a lower forecast measurement error as compared to ARIMA model.

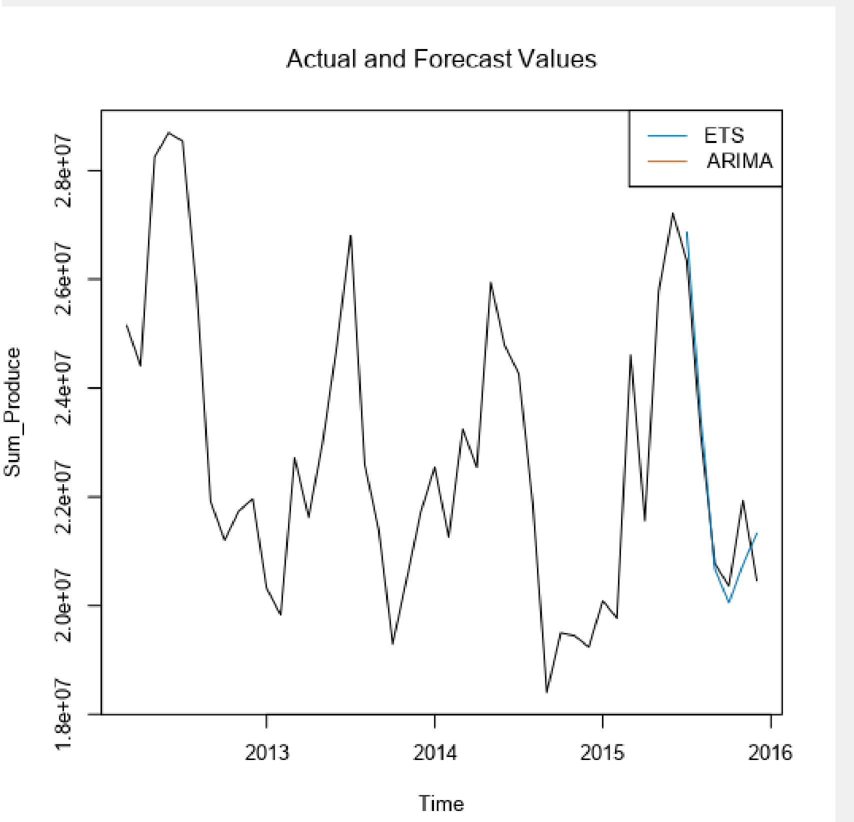
This can be seen in the MASE, ME and RMSE results in the accuracy measure table below:



Accuracy Measures:

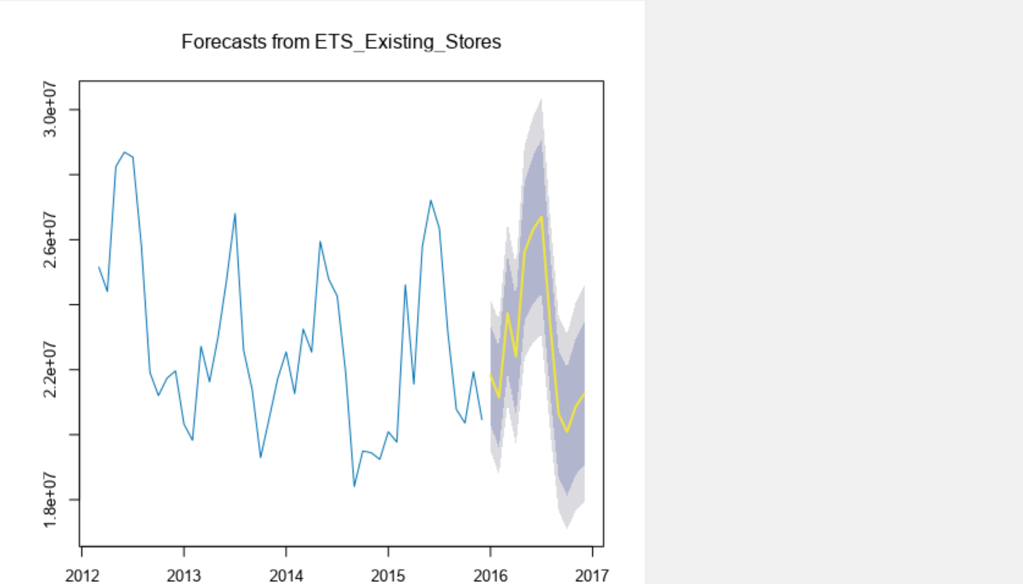
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

Below is a line plot that showed ETS model forecast, much closer to the actual data of the sum of produce in the StoreSalesData.csv.

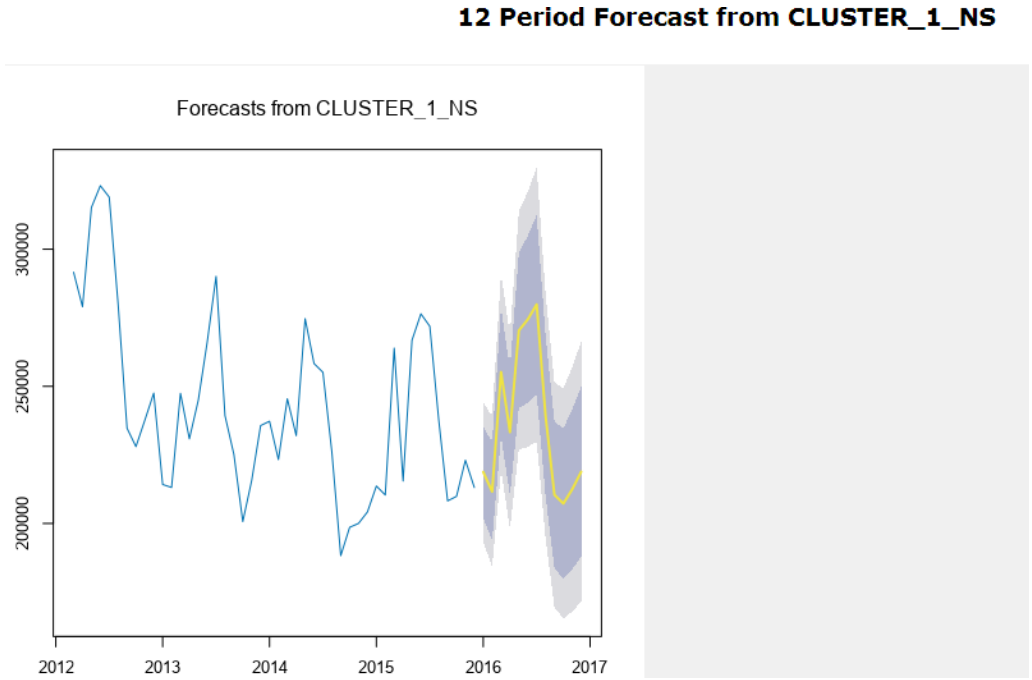


Below forecast graph of the ETS Model on Existing Stores:

12 Period Forecast from ETS_Existing_Stores

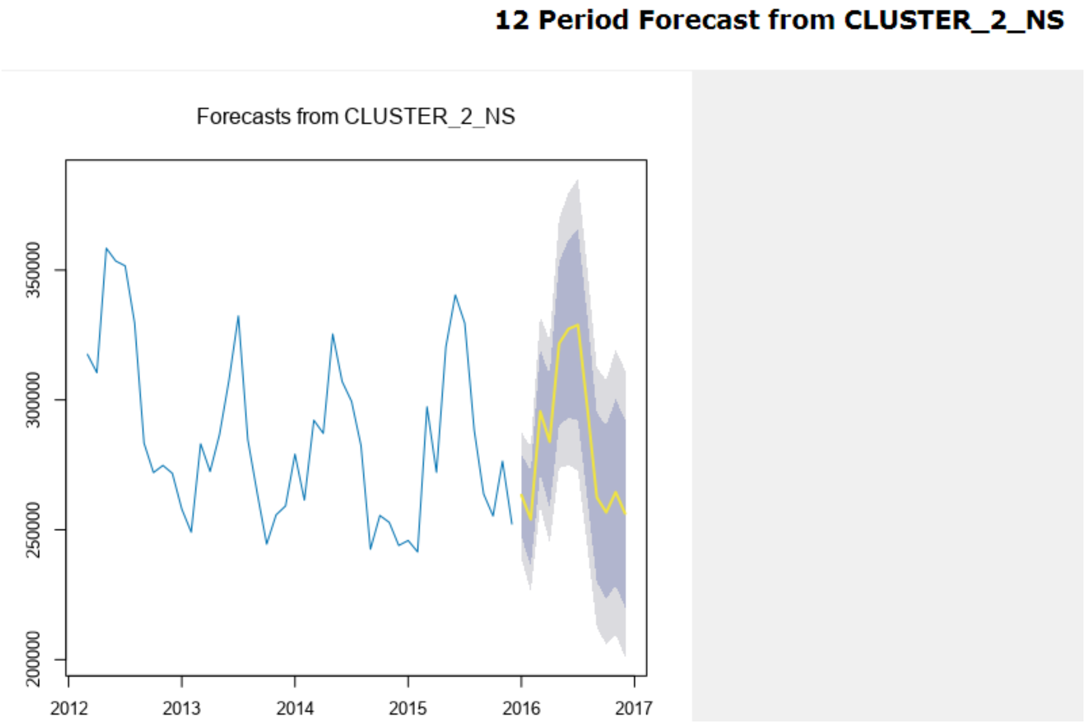


Below forecast graph of the ETS Model on New Store Cluster 1:



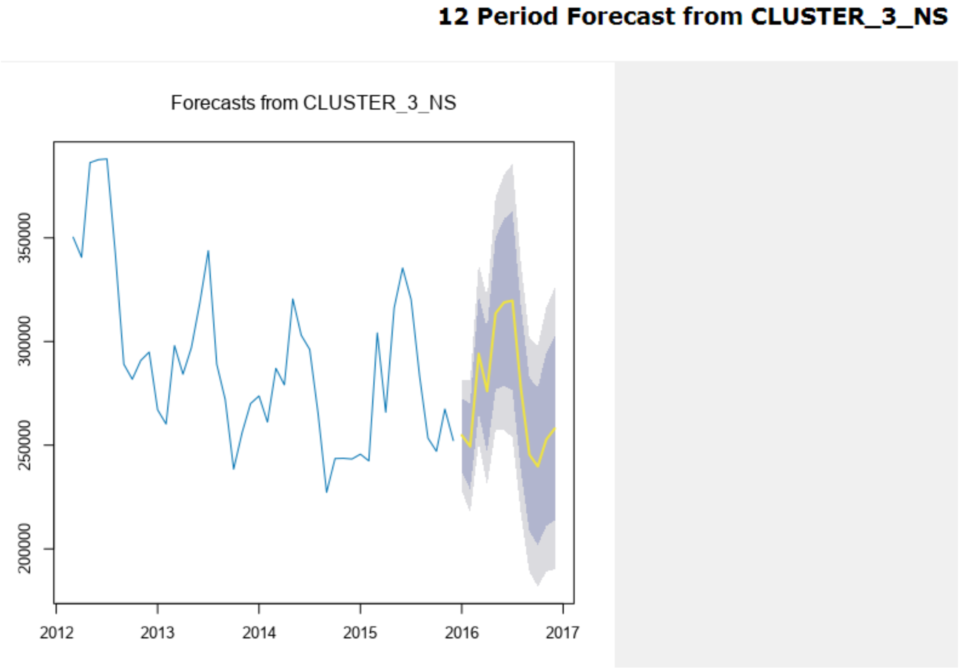
12 Period Forecast from CLUSTER_1_NS

Below forecast graph of the ETS Model on New Store Cluster 2:



12 Period Forecast from CLUSTER_2_NS

Below forecast graph of the ETS Model on New Store Cluster 3:



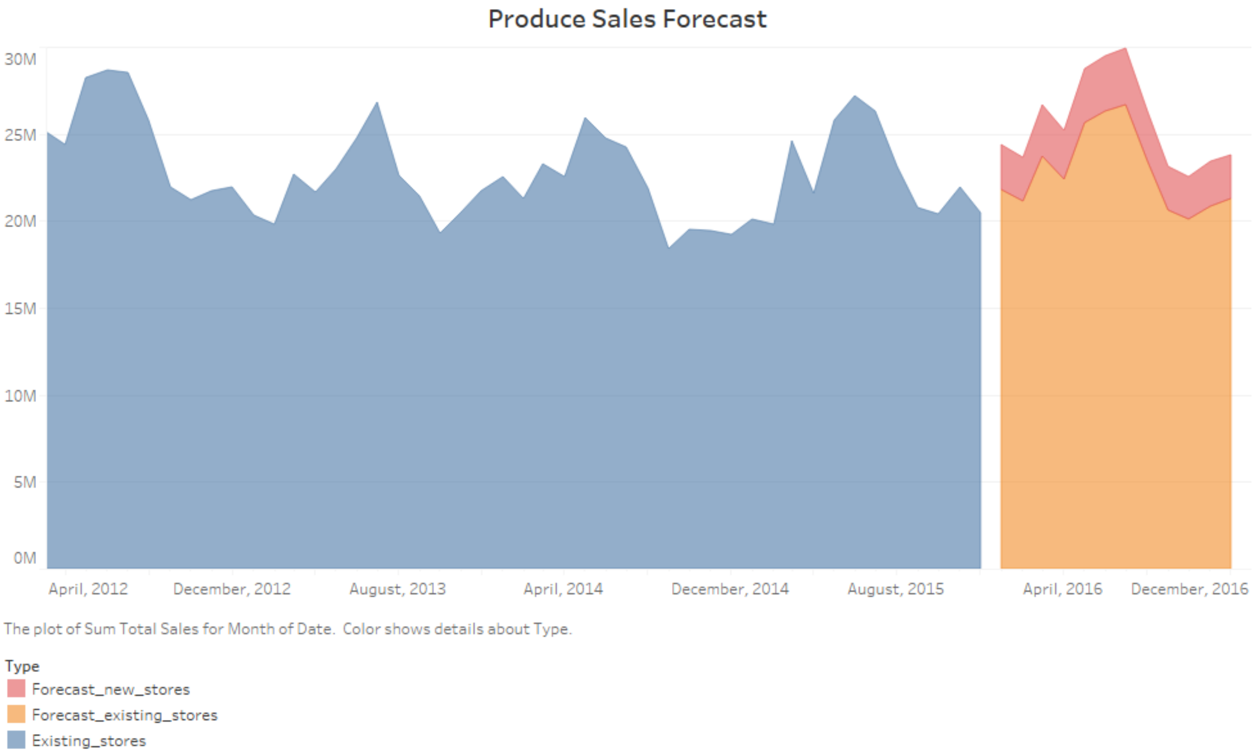
12 Period Forecast from CLUSTER_3_NS

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 table below shows a 12 months forecast of the existing and new stores:

Month	New Stores	Existing Stores
Jan-16	2563357.91	21829060.03
Feb-16	2483924.728	21146329.63
Mar-16	2910944.146	23735686.94
Apr-16	2764881.87	22409515.28
May-16	3141305.867	25621828.73
Jul-16	3195054.204	26307858.04
Jun-16	3212390.954	26705092.56
Aug-16	2852385.769	23440761.33
Sep-16	2521697.187	20640047.32
Oct-16	2466750.894	20086270.46
Nov-16	2557744.588	20858119.96
Dec-16	2530510.805	21255190.24

The visualization below shows a forecast of the historical data of the produce and 12 months forecast of the existing stores, and new stores:



The plot of Sum Total Sales for Month of Date. Color shows details about Type.