Farida Muhammad

Project: Predictive Analytics Capstone

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?**

   After running the K-Centroids Diagnostics tool in Alteryx to help assess the appropriate number of cluster to use, the K-means report, Adjusted Rand and Calinski-Harabasz indices shown below, indicate that three (3) is the optimal number of store formats. Both, the adjusted rand and the Calinski-Harabasz indices have the highest median values at 3 clusters.

K-Means Cluster Assessment Report

Summary Statistics

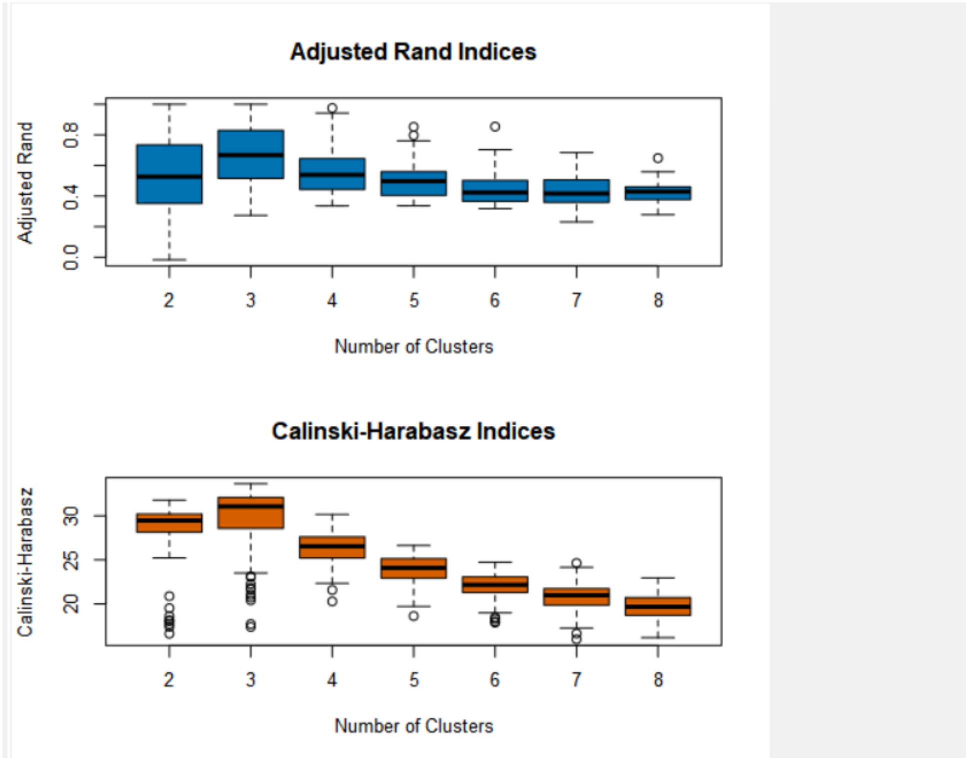
Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.016293	0.27351	0.335359	0.336327	0.318262	0.230196	0.27786
1st Quartile	0.352041	0.515917	0.445826	0.409773	0.366788	0.358895	0.377341
Median	0.526785	0.66768	0.538528	0.497192	0.423541	0.416509	0.428806
Mean	0.53781	0.664773	0.565975	0.50103	0.45115	0.432196	0.421514
3rd Quartile	0.734477	0.826692	0.644691	0.555087	0.499921	0.502931	0.458601
Maximum	1	1	0.975264	0.852076	0.8539	0.683894	0.647983

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.61829	17.38103	20.28456	18.61989	17.8746	15.98702	16.16824
1st Quartile	28.17383	28.57484	25.20913	22.93454	21.30575	19.85155	18.71365
Median	29.46587	31.05384	26.53788	24.086	22.16245	20.97743	19.6662
Mean	28.45131	29.70664	26.41806	23.87003	22.02174	20.77195	19.65973
3rd Quartile	30.17907	32.08726	27.59305	25.10099	23.06602	21.72942	20.7099
Maximum	31.78345	33.63781	30.1583	26.63063	24.72038	24.63982	22.95166

Plots



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

   The K-Centroids Cluster Analysis tool was run in Alteryx. None of the clusters have more than 40 stores in it or less than 20. Cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores.

Report

Summary Report of the K-Means Clustering Solution ClusterByStore

Solution Summary

Call:

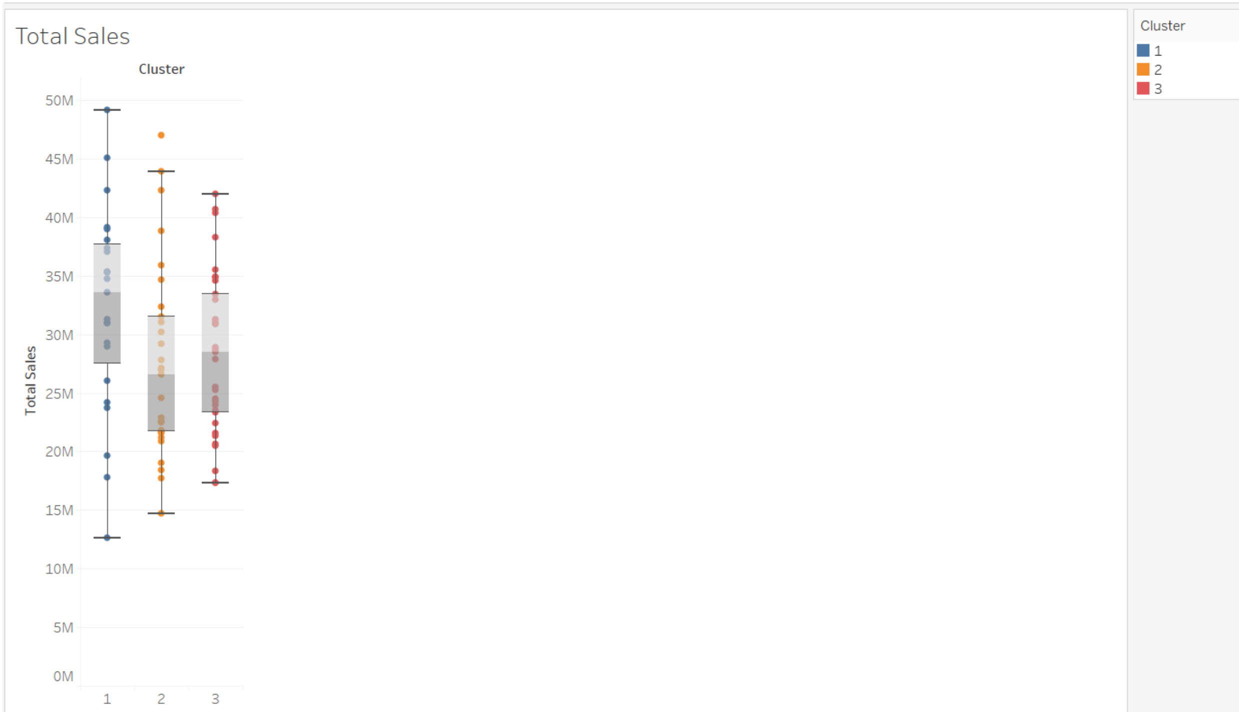
stepFlexclust(scale(model.matrix(~-1 + Dry_Grocery_Percentage + Dairy_Percentage + Frozen_Food_Percentage + Meat_Percentage + Produce_Percentage + Floral_Percentage + Deli_Percentage + Bakery_Percentage + General_Merchandise_Percentage, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

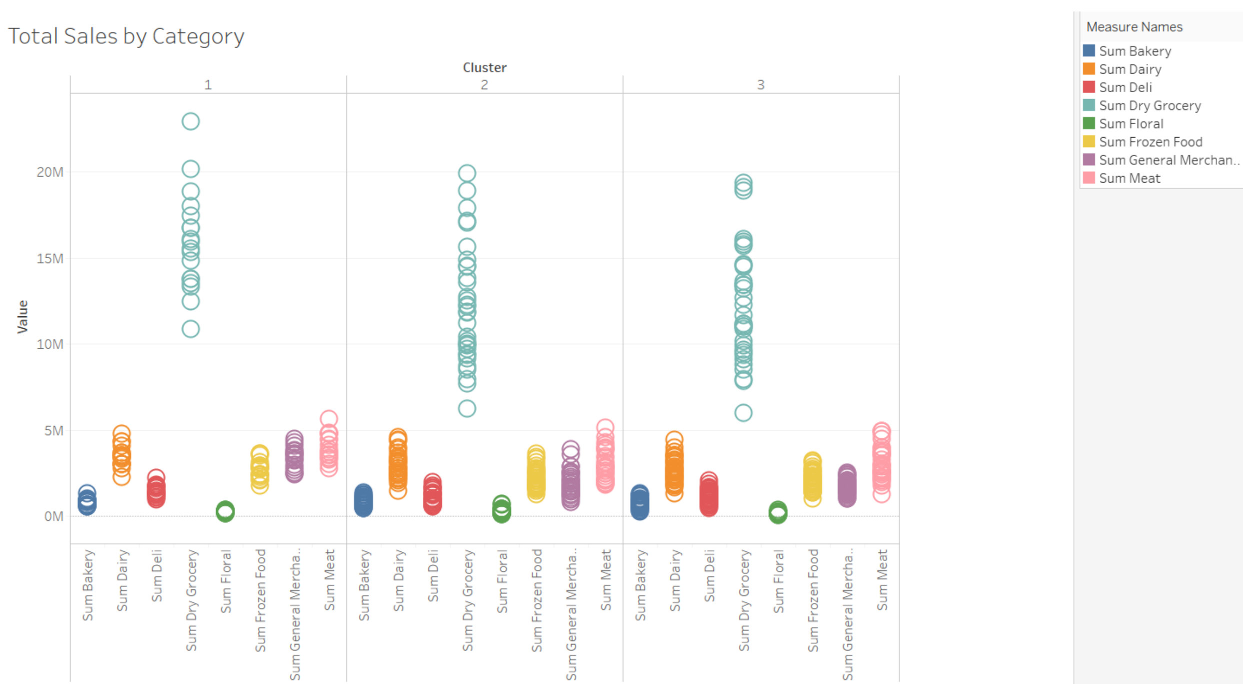
Cluster Information:

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

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

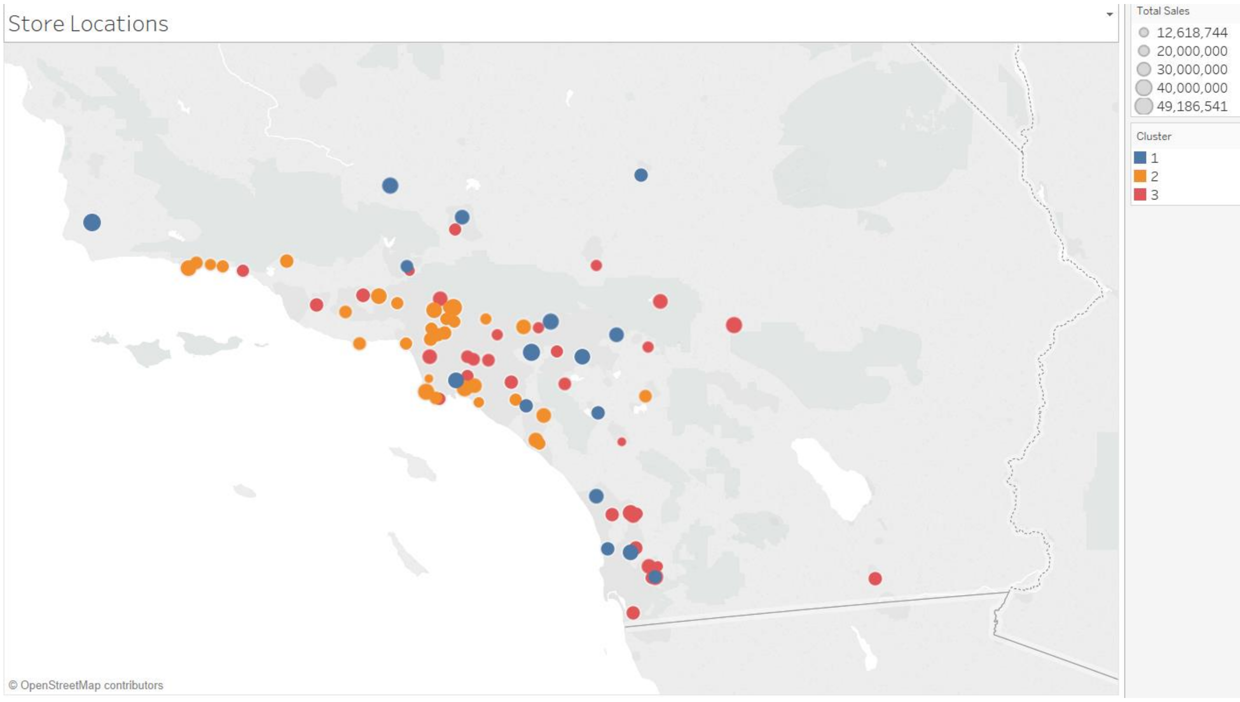
   Cluster 1 stores have highest total sales when compared to the other 2. It also had the highest sales in most categories: Dairy, Deli, Dry Grocery, Frozen Food, General Merchandise, and Meat. Cluster 2 had the highest sales in Bakery, Floral, and Produce. Cluster 3 stores are more compact and have similar sales.





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**.

Store Locations



https://public.tableau.com/profile/farida2860#!/vizhome/Task1StoreLocations_0/StoreLocations

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 model comparison report below shows comparison matrix of Decision Tree, Forest Model and Boosted Model. The accuracy measure for each model is 82.35%. The **Boosted Model** was chosen due to higher precision measure 88.89% (F1 value). The precision measure, which is the 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.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_14	0.8235	0.8426	0.7500	1.0000	0.7778
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	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_Model

	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_14

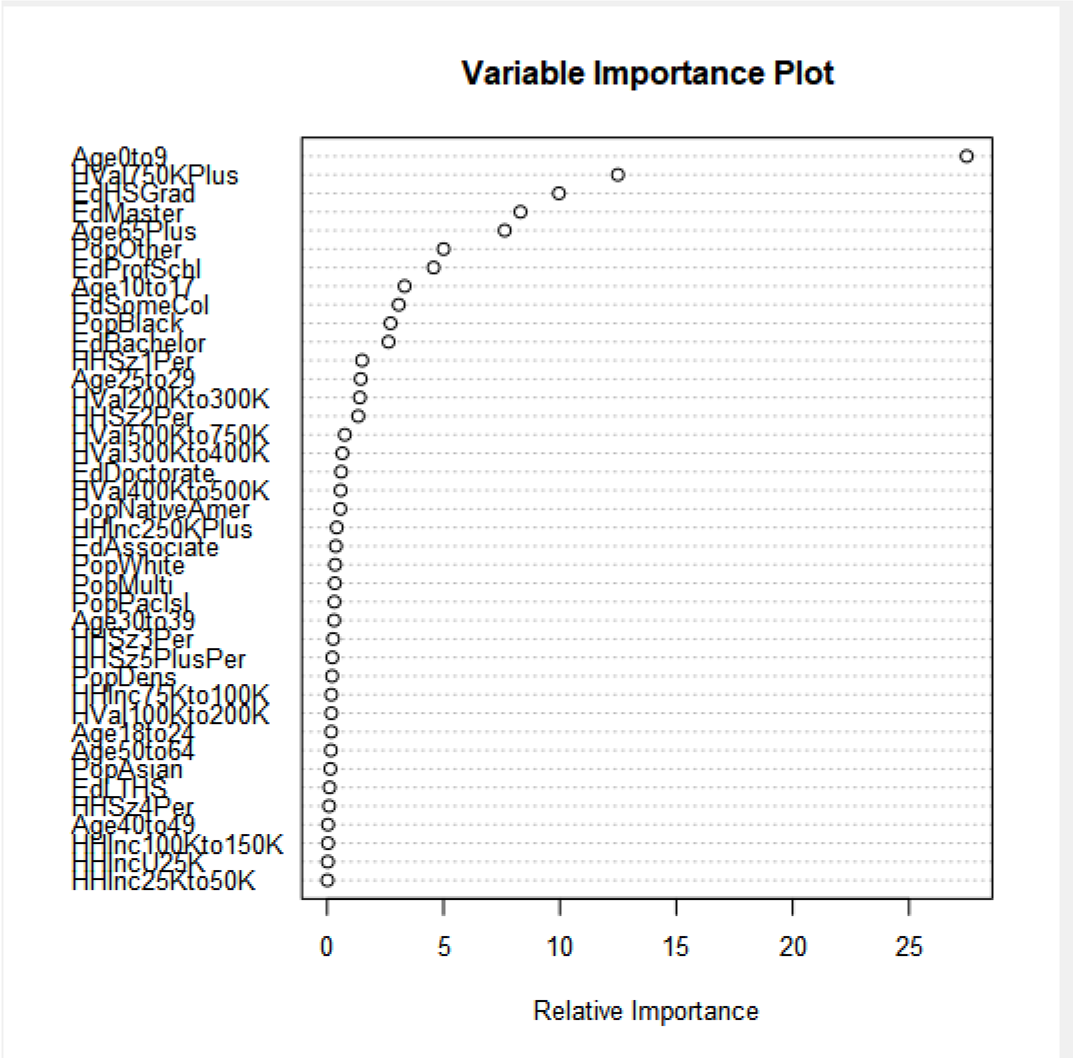
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Confusion matrix of Forest_Model

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

2. **What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization**.

   The three most important variables in the Boosted Model are**:** Age0to9, HVAL750KPlus, and EdHSGrad.



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

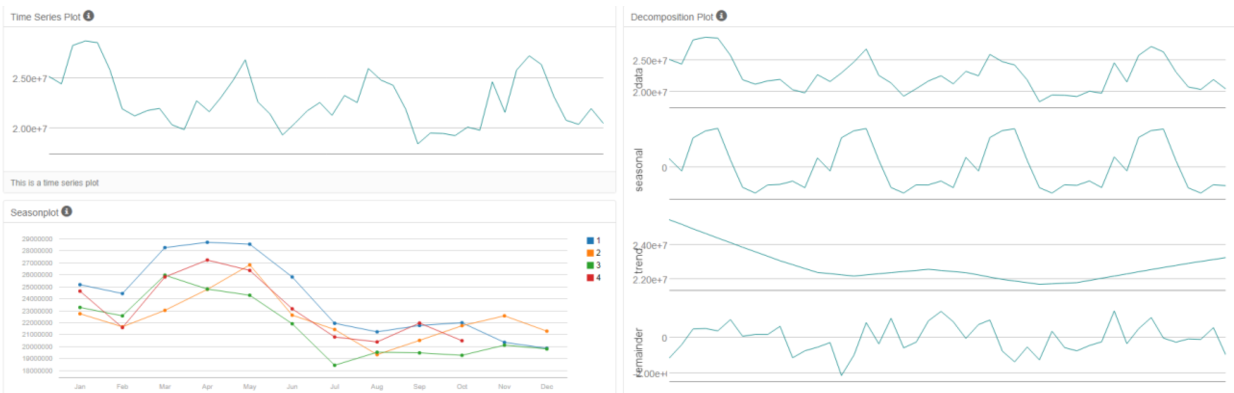
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?

For the ETS Model, I used ETS(M,N,M) with no dampening**.** ETS(MNM) has lower forecasting errors, therefore ETS(MNM) will be used to forecast the sales for the new and existing stores.

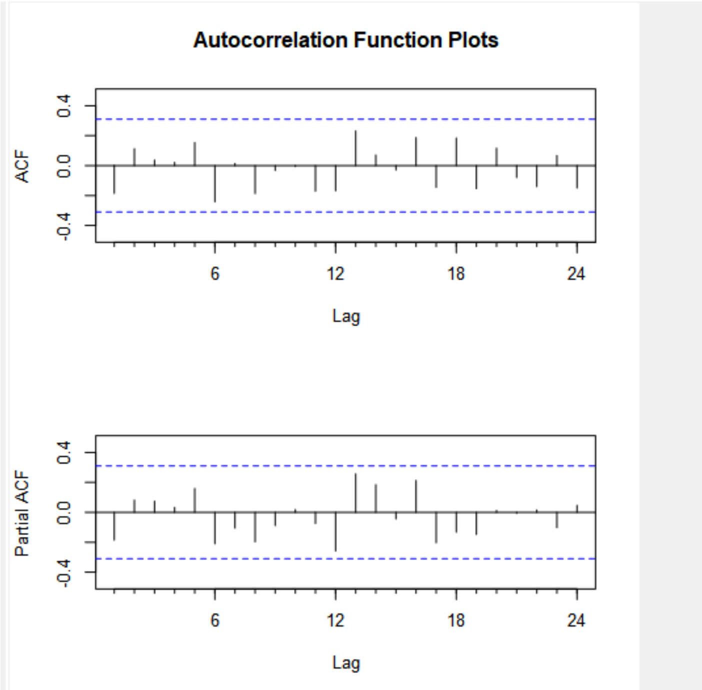
The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. Its error is irregular and should be applied multiplicatively.



For the ARIMA Model, I used ARIMA(1,0,0)(1,1,0)[12].
Just by adding one SAR, stationarized the series, eliminating all of the significant lags and performing much better on the validation set.

After reviewing the ACF and PACF charts, it can be concluded that no additional AR or MR terms are needed since neither chart show significant correlation. The bars are also between the dashed lines.

Plots



- For forecasting, I chose the **ETS(M,N,M) with no dampening model**. After analyzing the data, the investigation shows that the **ETS model's accuracy is higher** when compared to ARIMA model. A holdout sample of 6 months data was used.
    - That the MAPE of the ETS model is lower than the ARIMA. This suggests that, on average, the ETS model misses its forecast by a lesser amount.
    - Also, the RMSE for ETS is lower at 760267.3 compared to RMSE for ARIMA at 1050239. The MASE value for the ETS model is 0.3822**,** which is also lower than MASE value for the ARIMA model at 0.5463.

ETS(M,N,M) with no dampening

Actual	ETS
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

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

ARIMA(1,0,0)(1,1,0)[12].

Actual and Forecast Values:

Actual	ARIMA
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

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

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 the forecast sales for new and existing stores. New store sales is obtained by using **ETS(M,N,M)** analysis with all the 3 individual cluster to obtain the average sales per store. The average sales value (cluster 1 x3, cluster 2 x6, cluster 3 x1) are added up produce New Store Sales.

Month/Year	New Stores	Existing Stores
January 2016	2,587,451	21,539,936
February 2016	2,477,353	20,413,771
March 2016	2,913,185	24,325,953
April 2016	2,775,746	22,993,466
May 2016	3,150,867	26,691,951
June 2016	3,188,922	26,989,964
July 2016	3,214,746	26,948,631
August 2016	2,866,349	24,091,579
September 2016	2,538,727	20,523,492
October 2016	2,488,148	20,011,749
November 2016	2,595,270	21,177,435
December 2016	2,573,397	20,855,799

VISUALIZATION – TOTAL PRODUCE SALES