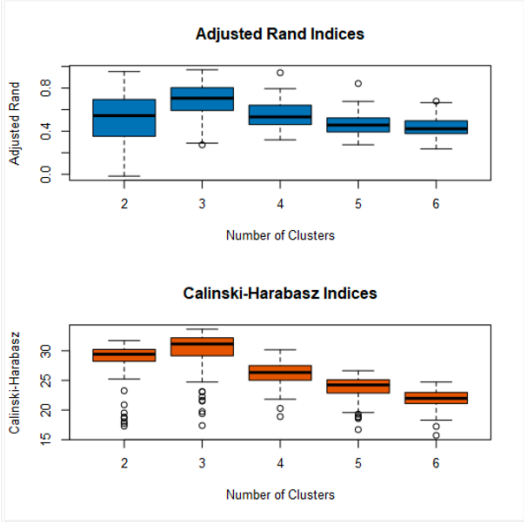
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.281	17.38103	18.89398	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.03471	22.86498	21.10249
Median	29.4157	31.14178	26.33467	24.22188	21.96958
Mean	28.56936	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63781	30.1583	26.63063	24.72038

Based on the following analysis:

1- K-means cluster assessment report
2- Adjusted Rand
3- Calinski-Harabasz

The number of clusters can be chosen and concluded to be **3** clusters.

2. How many stores fall into each store format?

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

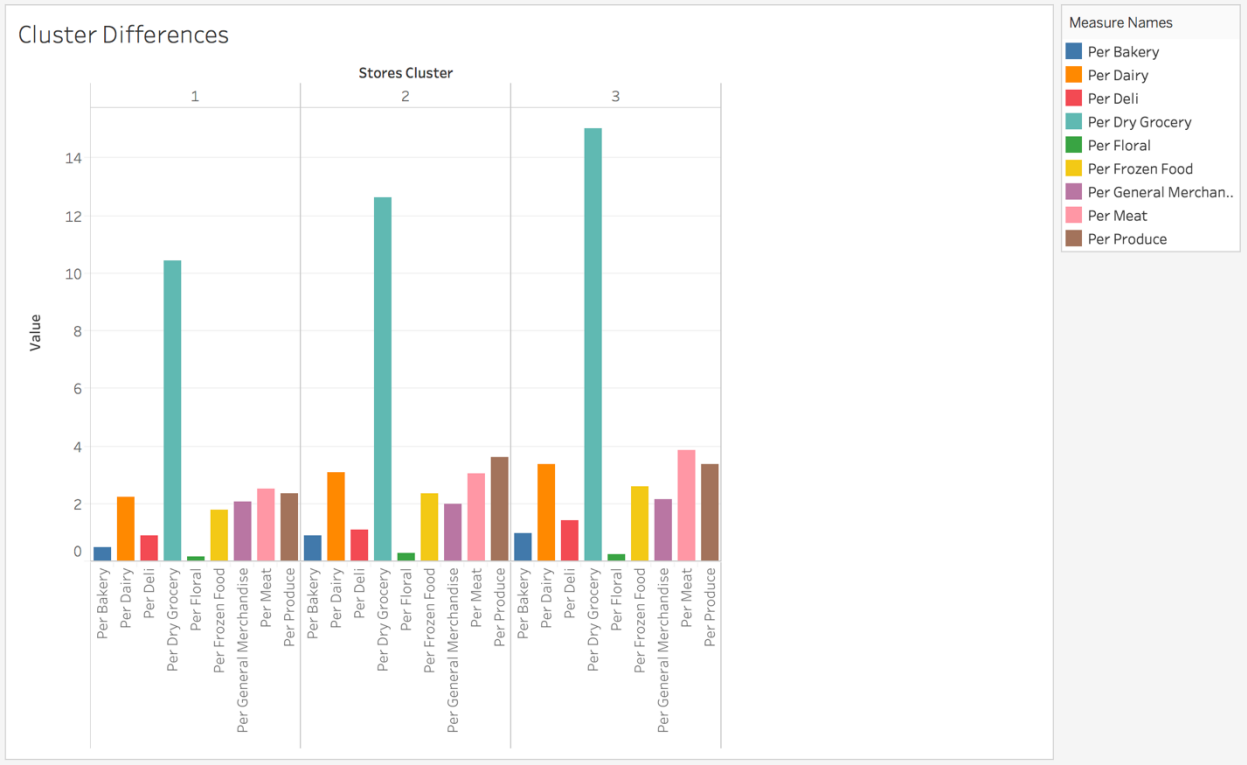
Based on the analysis shown on the table above:

1- Cluster 1: 23 stores
2- Cluster 2: 29 stores
3- Cluster 3: 33 stores

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

	Per_Dry_Grocery	Per_Dairy	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_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

	Per_Bakery	Per_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

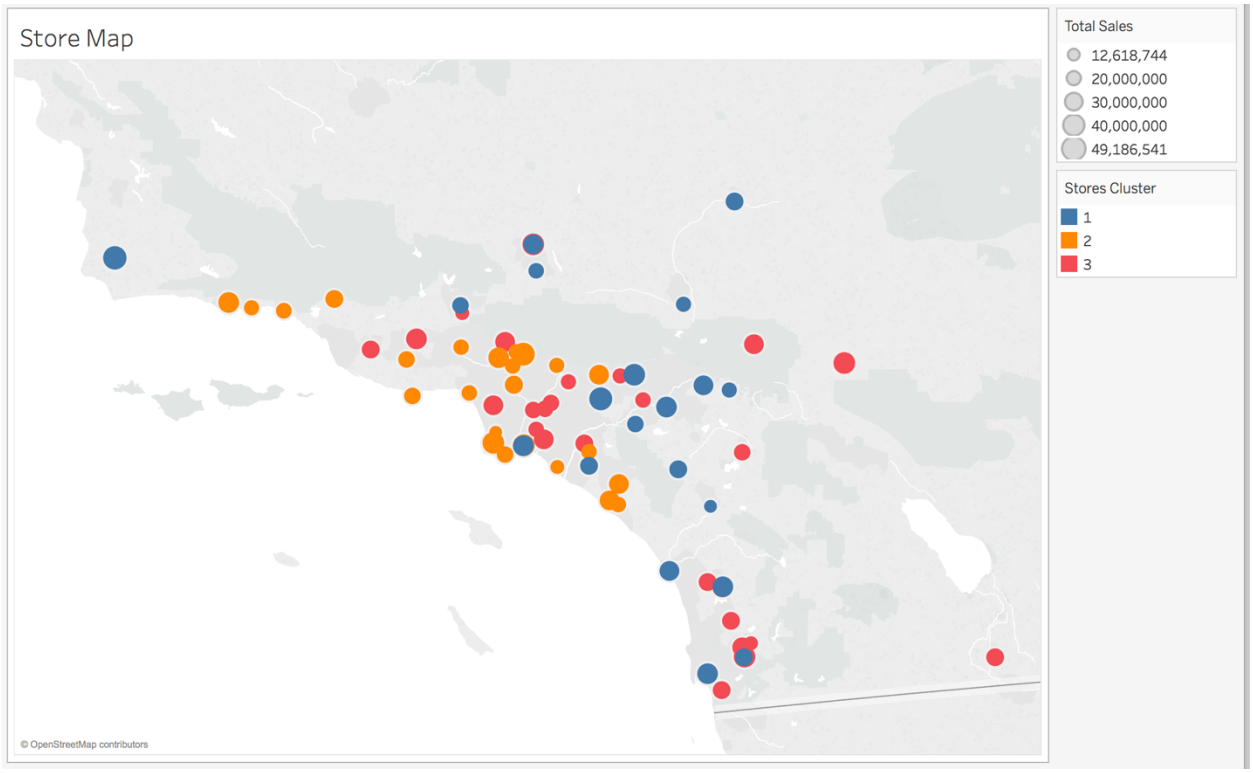


https://public.tableau.com/profile/shaikhah.almahasheer#!/vizhome/ClusterDifferences-Shaikhah/Sheet2

Clusters have many differences:

1- Sales categories: some have a higher value than the others and some has a lower value
2- Meat, Dry Grocery, Deli, diary and bakery sales: significantly higher – in terms of percentages - in cluster 3 than the other clusters
3- Produce and Floral: higher in cluster 2
4- General Merchandise: highest in cluster 1

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.**



https://public.tableau.com/profile/shaikhah.almahasheer#!/vizhome/Task1-Shaikhah/Sheet1

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?**

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778

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

	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_Model

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

Based on the analysis provided by the model comparison report, Forest and the Boosted Model are similar in terms of accuracy. Boosted Model is higher in terms of F1. Therefore, Best model to be used is Boosted Model.

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

Store Number	Cluster
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?**

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

ARIMA (0,1,2)(0,1,0)

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

ETS (M,N,M)

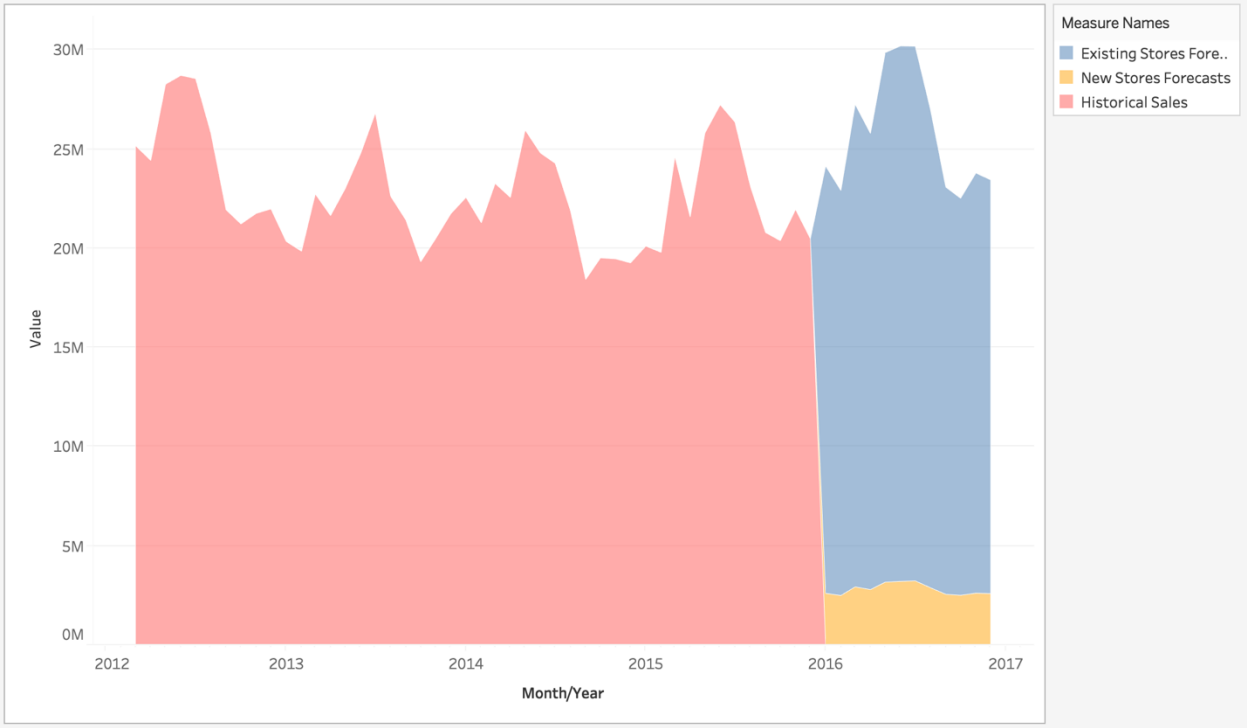
Based on the analysis shown above, ETS is a better measure in terms of RMSE and MASE.

1- Seasonality trend shows an increase and should be applied multiplicatively
2- Nothing clear about the trend and therefore nothing should be applied
3- Error trend is irregular and should be applied multiplicatively

Based on all of the above points, **ETS (M,N,M) with no dampening** is used.

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.

Month	Existing Stores	New Stores
1	21539936.01	2477352.89
2	20413770.6	2913185.24
3	24325953.1	2775745.61
4	22993466.35	3150866.84
5	26691951.42	3188922
6	26989964.01	3214745.65
7	26948630.76	2866348.66
8	24091579.35	2538726.85
9	20523492.41	2488148.29
10	20011748.67	2595270.39
11	21177435.49	2573396.63
12	20855799.11	2587450.85



https://public.tableau.com/profile/shaikhah.almahasheer#!/vizhome/ForecastVisual-Shaikhah/Sheet1