Project: Predictive Analytics Capstone

Our company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. Our aim here is to provide analytical support to make decisions about store formats and inventory planning.

Task 1: Determine Store Formats for Existing Stores

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

According to my analysis, the optimal number of store formats is 3. We can see that the median is highest for position number 3 for both Adjusted Rand Indices and Calinski-Harabasz Indices.

K-Means Cluster Assessment Report

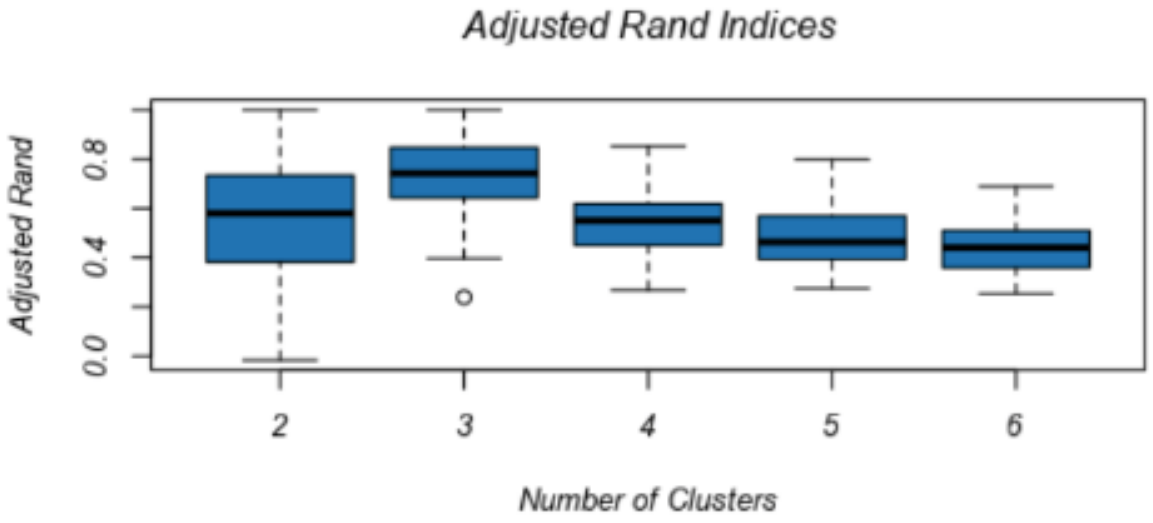
Summary Statistics

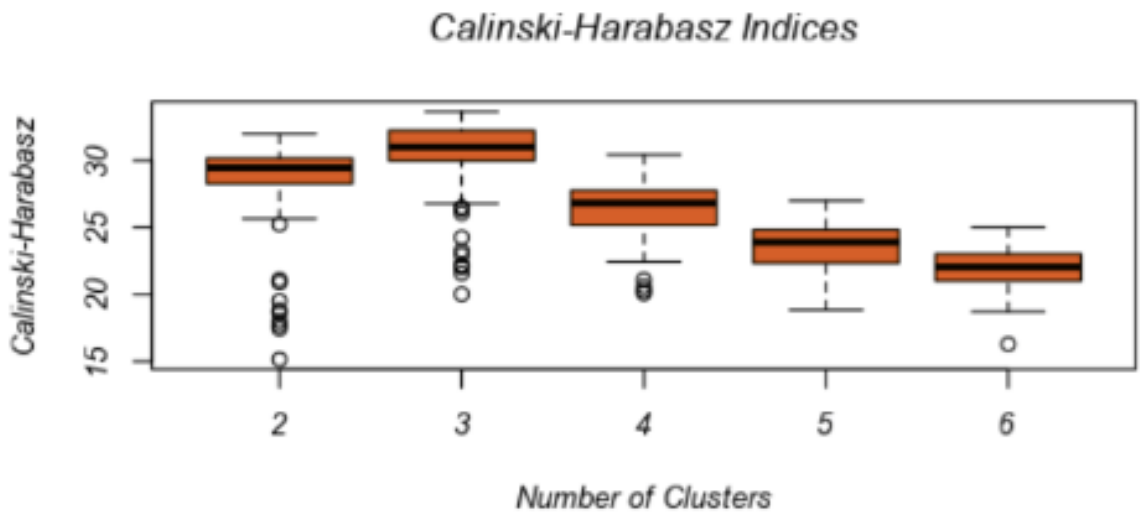
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.238908	0.26746	0.275161	0.254075
1st Quartile	0.389138	0.643526	0.451546	0.393179	0.361002
Median	0.579832	0.742946	0.550094	0.46327	0.440569
Mean	0.538248	0.716946	0.539436	0.480527	0.444128
3rd Quartile	0.734477	0.841627	0.618537	0.564177	0.507959
Maximum	1	1	0.851619	0.798934	0.689104

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.14927	20.01657	20.07469	18.84105	16.28411
1st Quartile	28.27367	30.07272	25.16346	22.35521	21.04521
Median	29.4511	31.00382	26.81884	23.89722	22.0471
Mean	28.40735	30.28555	26.35179	23.56802	21.93001
3rd Quartile	30.16162	32.23534	27.76016	24.82346	22.99673
Maximum	31.9781	33.63781	30.41396	26.97019	25.00769





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	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 sold more General Merchandise compared to more produce in cluster 2 stores. Cluster 1 stores have the highest total sales compared to Cluster 2 and 3 stores.

Summary Report of the K-Means Clustering Solution K

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:

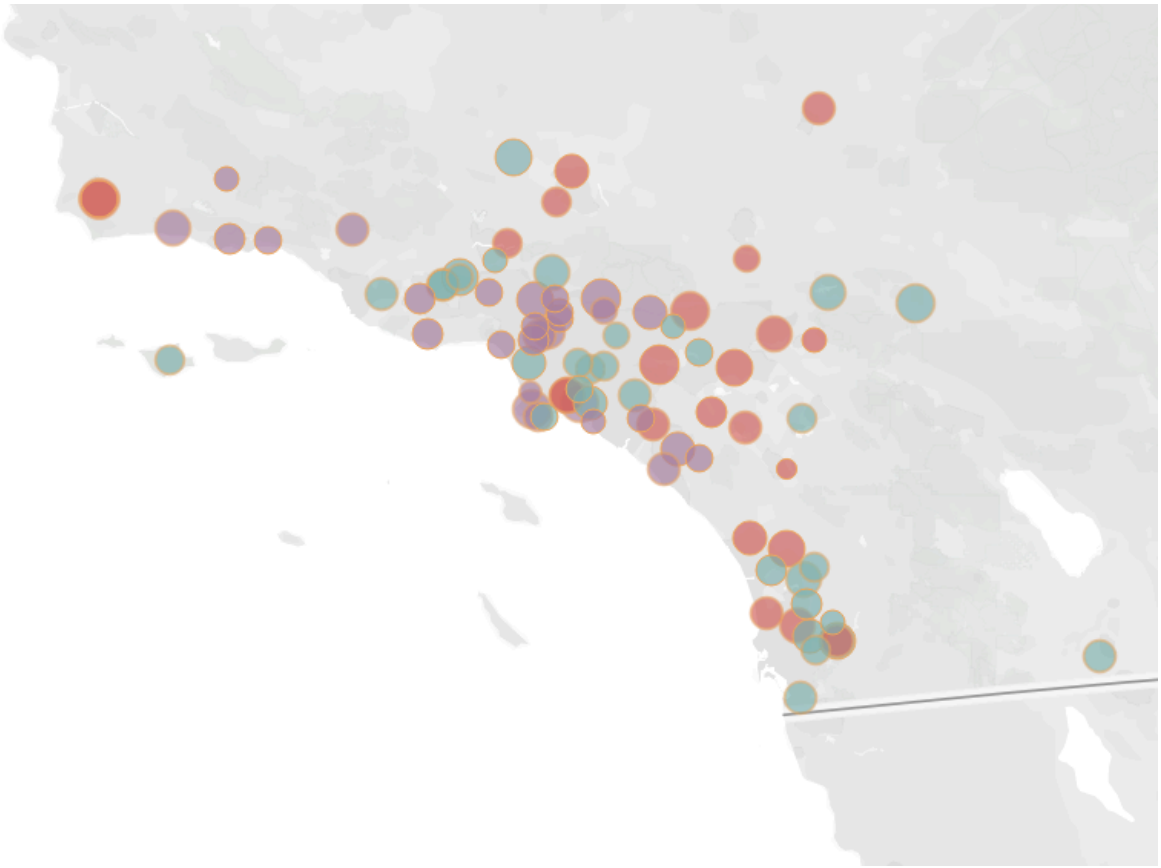
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

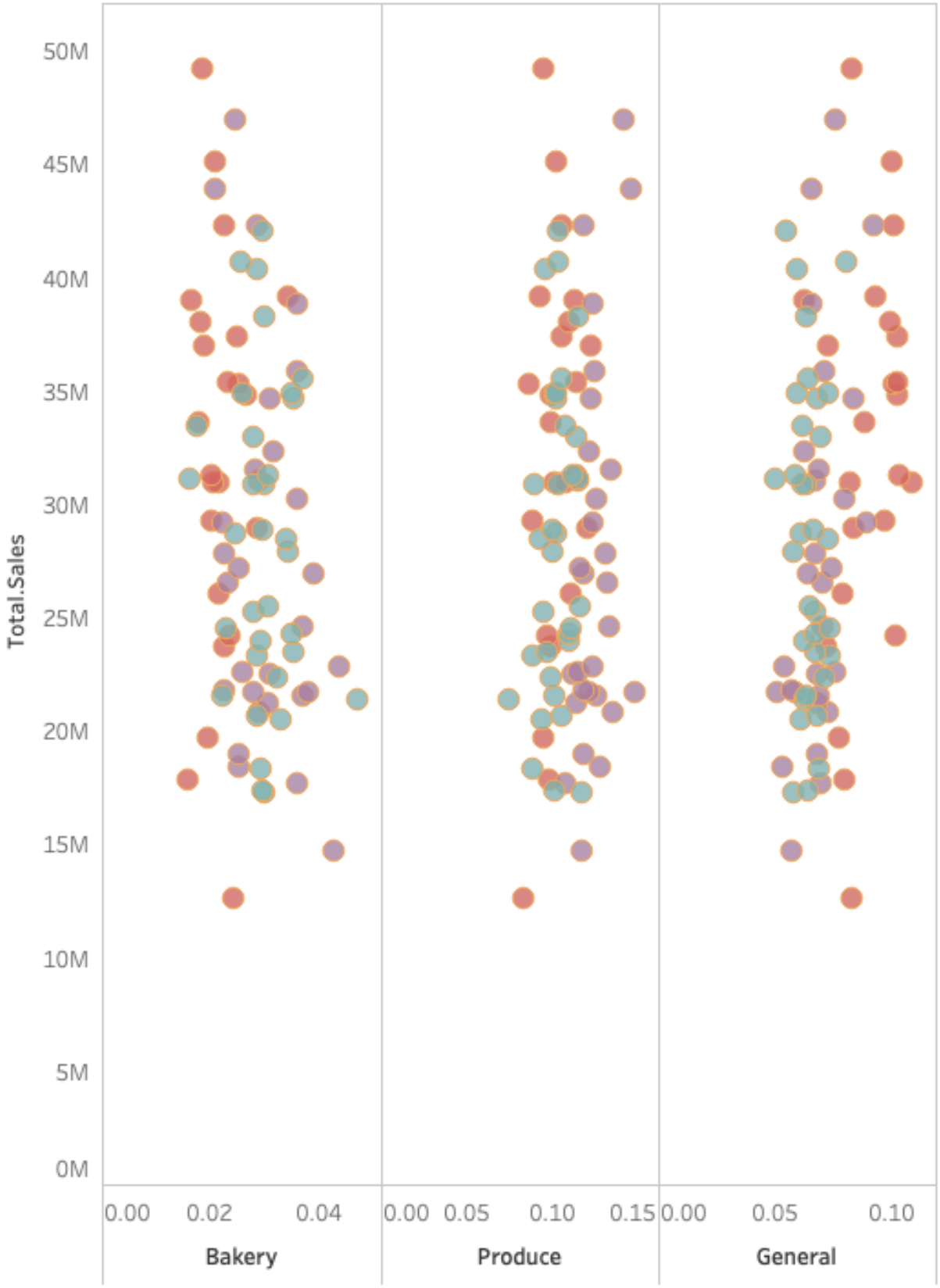
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_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

	Percent_Bakery	Percent_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.





https://public.tableau.com/profile/dikshide#!/vizhome/PredictiveAnalyticsNanodegreeCapstone/Task1

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.)

I used the Model Comparison tool to show comparison between Decision Tree, Forest Model and Boosted Model. We picked Boosted model because it has a higher F1 score despite having the same accuracy as the Forest Model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7327	0.6000	0.6667	0.8333
FM	0.8235	0.8251	0.7500	0.8000	0.8750
BM	0.8235	0.8543	0.8000	0.6667	1.0000

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of DT

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

Confusion matrix of FM

	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	1
S0087	2
S0088	3
S0089	2
S0090	2

S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

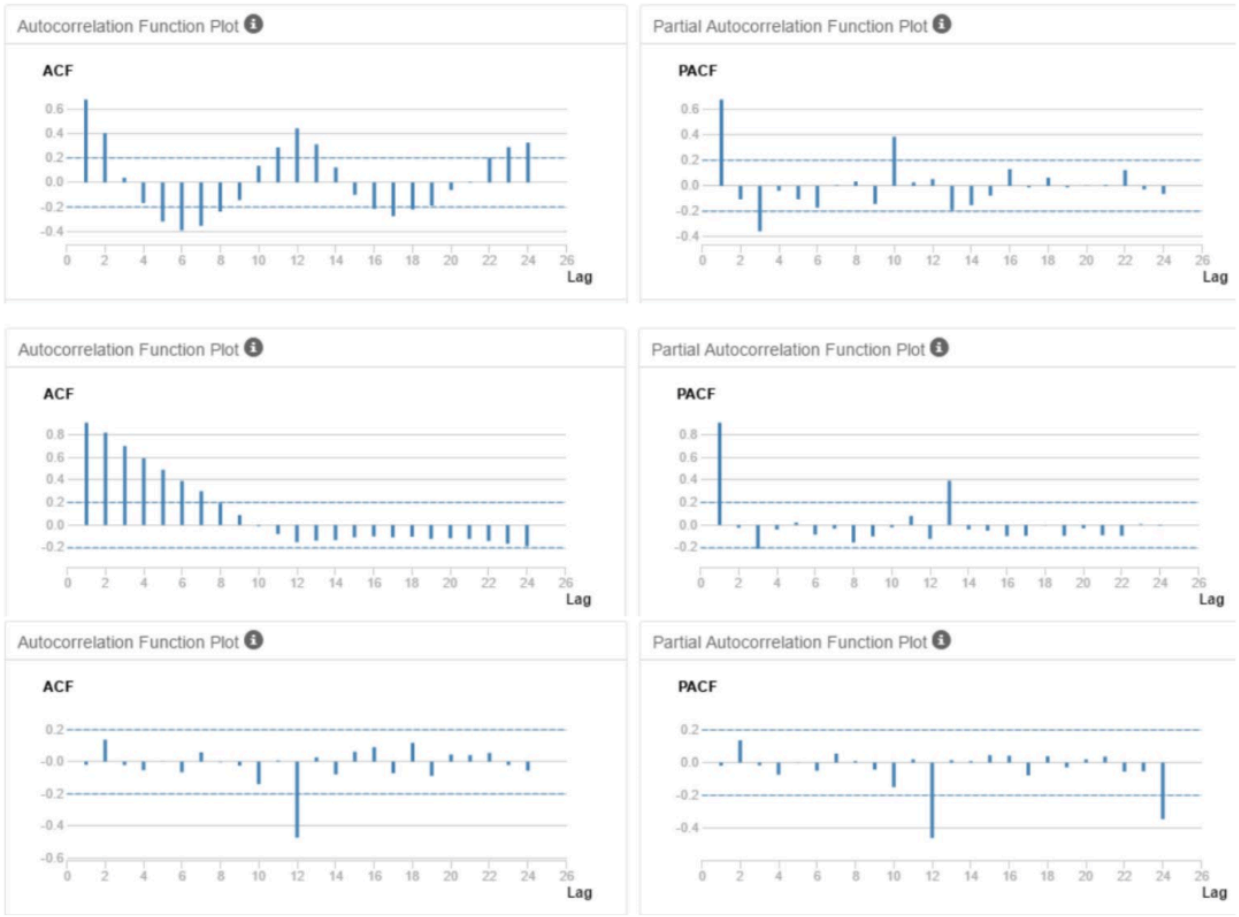
Task 3: Predicting Produce Sales

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?

ETS(M,N,M) model with no dampening is used for ETS model. The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and should be applied. It's error is irregular and should be applied multiplicatively.

ETS model is run with a holdout sample of 12 months.

ARIMA(0,1,2)(0,1,0) is used as seasonal difference and seasonal first difference were performed. There is a lag-2.



ARIMA model used is ARIMA(1,0,0)(1,1,0)[12]. Seasonal difference and seasonal first difference were performed to make the series stationary.

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

Accuracy Measures:

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

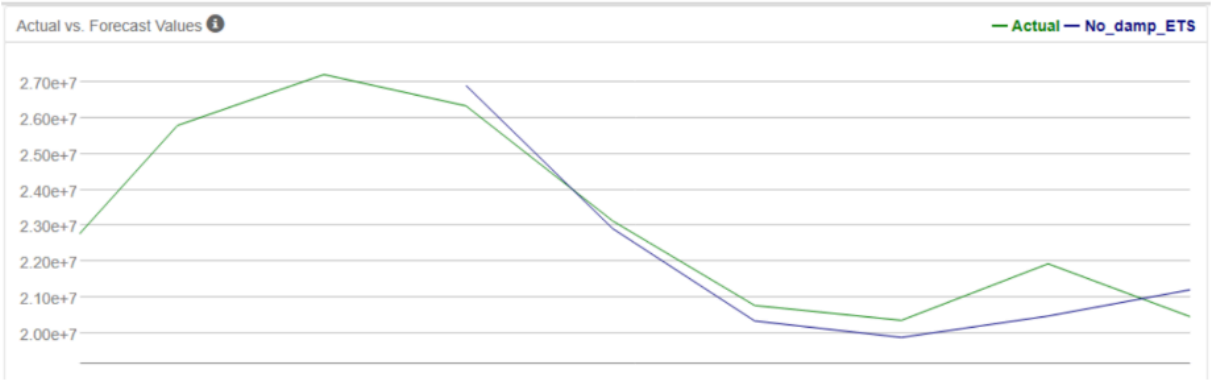
AIC	AICc	BIC
880.4445	881.4445	884.4411

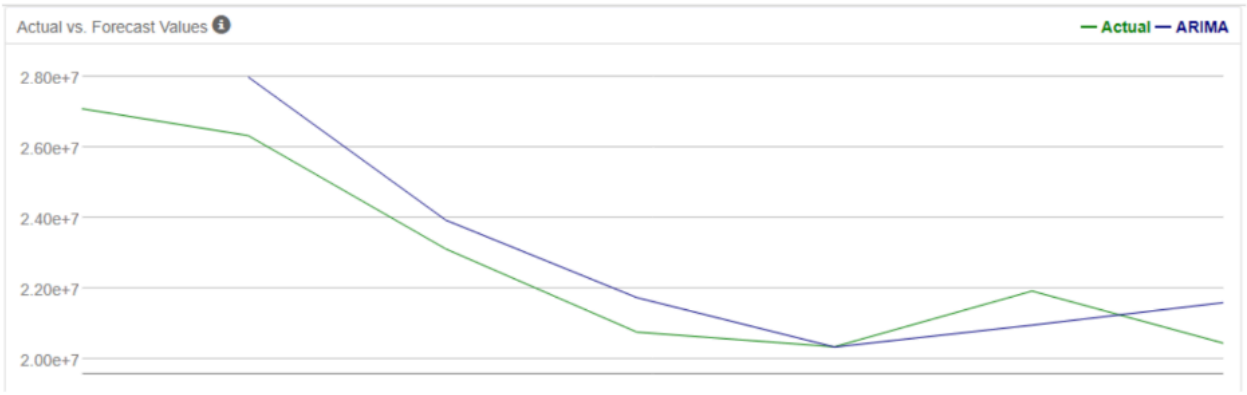
It is clear that ETS is actually closer to actual compared to ARIMA. So taking into account in-sample error measures, accuracy measures and actual vs forecast plots, I think ETS model is a better choice.

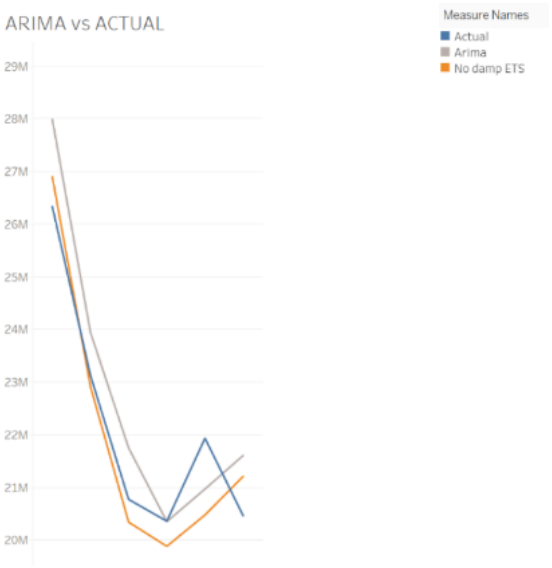
ETS model RMSE accuracy is 760267.3 vs ARIMA model RMSE accuracy of 1050239.
ETS model MASE accuracy is 0.3822 vs ARIMA model MASE accuracy of 0.5463.


Actual vs. Forecast Values
— Actual — ARIMA
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7


Actual vs. Forecast Values
— Actual — No_damp_ETS
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7

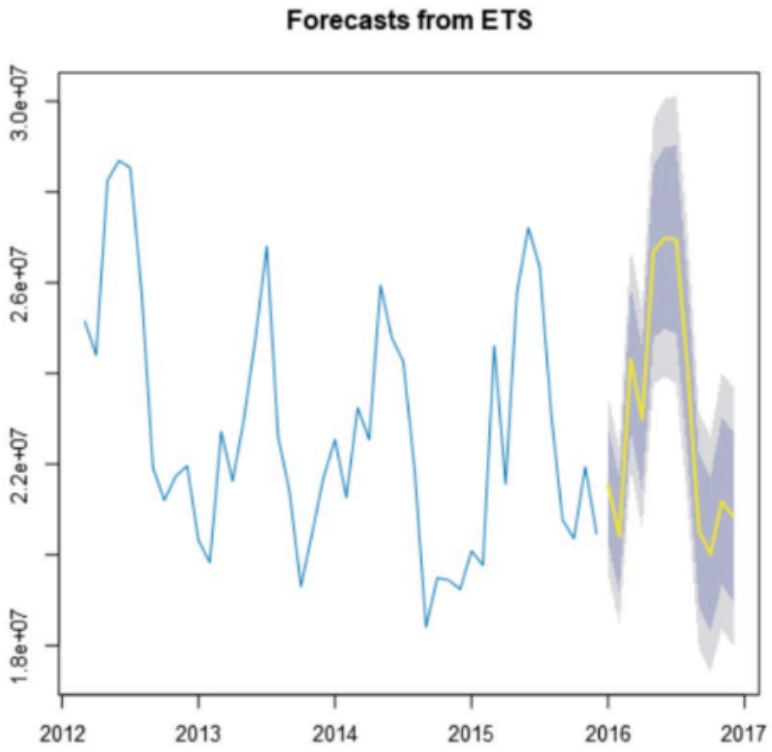

Actual vs. Forecast Values
— Actual — No_damp_ETS
2.70e+7
2.60e+7
2.50e+7
2.40e+7
2.30e+7
2.20e+7
2.10e+7
2.00e+7





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.**

Below graph shows actual forecast value with 80% and 95% confidence level interval.



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

Table below shows the forecast sales for existing stores and new 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 (x3 cluster 1, x6 cluster 2, x1 cluster 3) are added up produce New Store Sales.

Year	Month	New Store Sales	Existing Store Sales
2016	1	2,626,198	21,539,936
2016	2	2,529,186	20,413,771
2016	3	2,940,264	24,325,953

Year	Month	New Store Sales	Existing Store Sales
2016	4	2,774,135	22,993,466
2016	5	3,165,320	26,691,951
2016	6	3,203,286	26,989,964
2016	7	3,244,464	26,948,631
2016	8	2,871,488	24,091,579
2016	9	2,552,418	20,523,492
2016	10	2,482,837	20,011,749
2016	11	2,597,780	21,177,435
2016	12	2,591,815	20,855,799