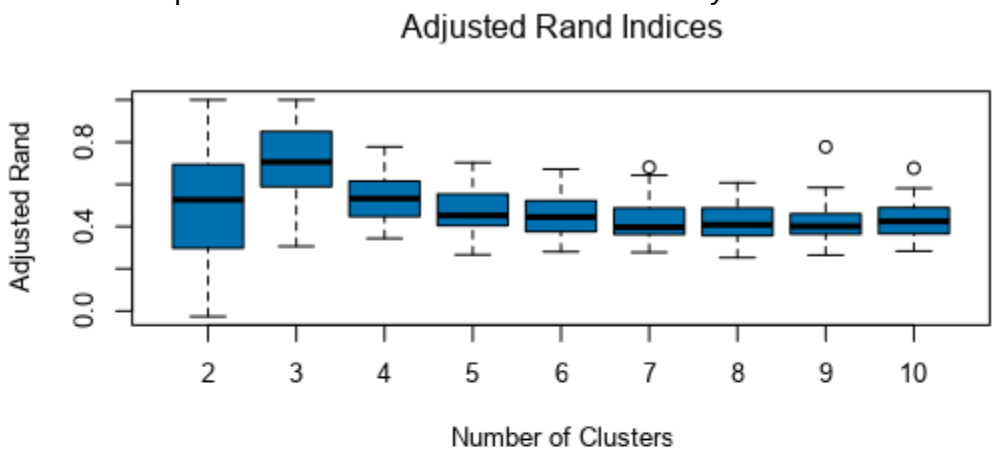
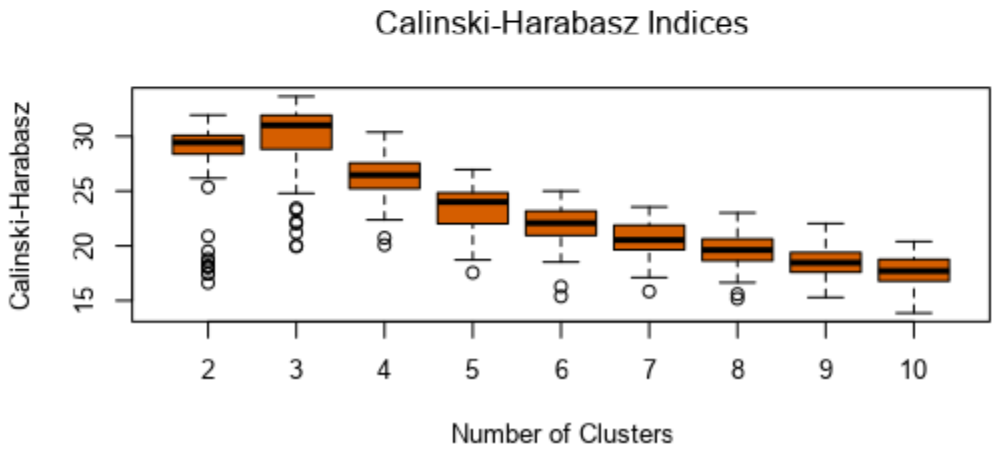
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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





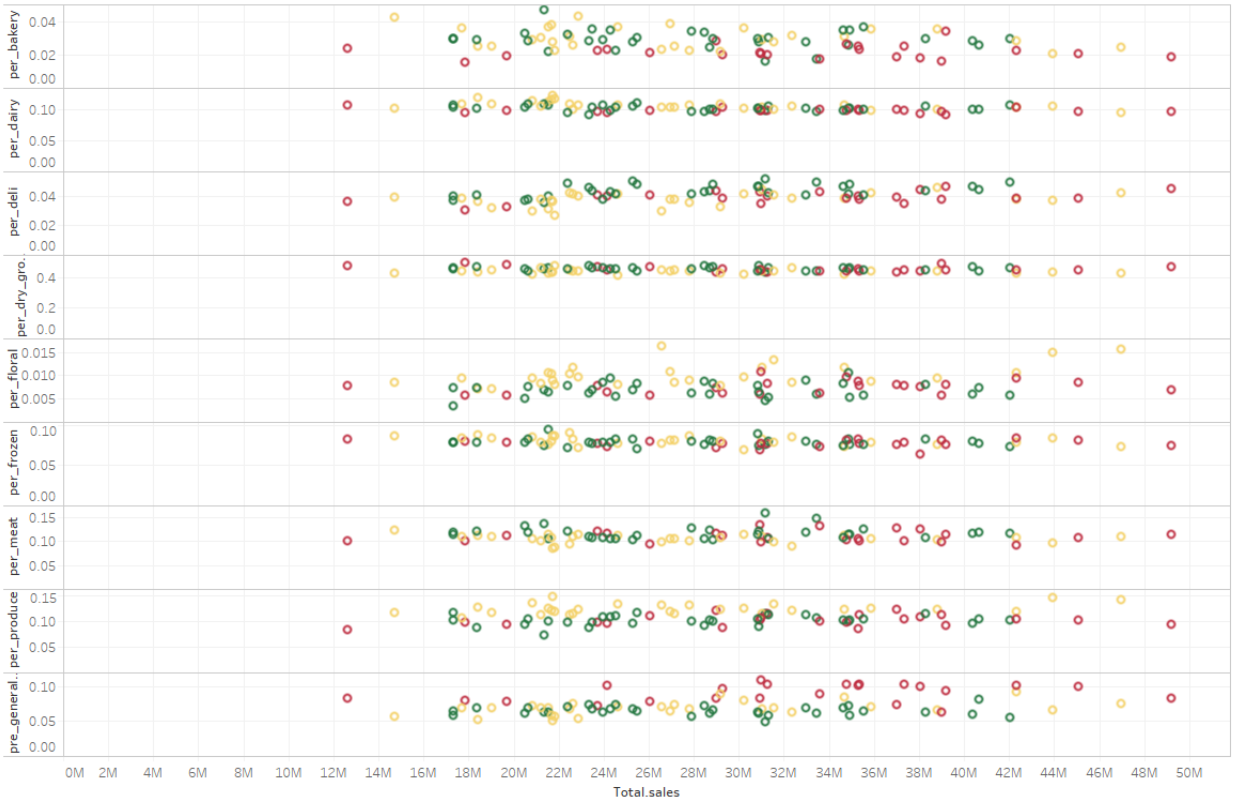
Base on report, Adjusted Rand and Calinski-Harabasz Indices, I chose 3 as the cluster number.

2. How many stores fall into each store format?

Record	Cluster	Count
1	1	23
2	2	29
3	3	33

For cluster 1, total 23 stores. For cluster 2, total 29 stores, for cluster 3, total 33 stores.

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



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

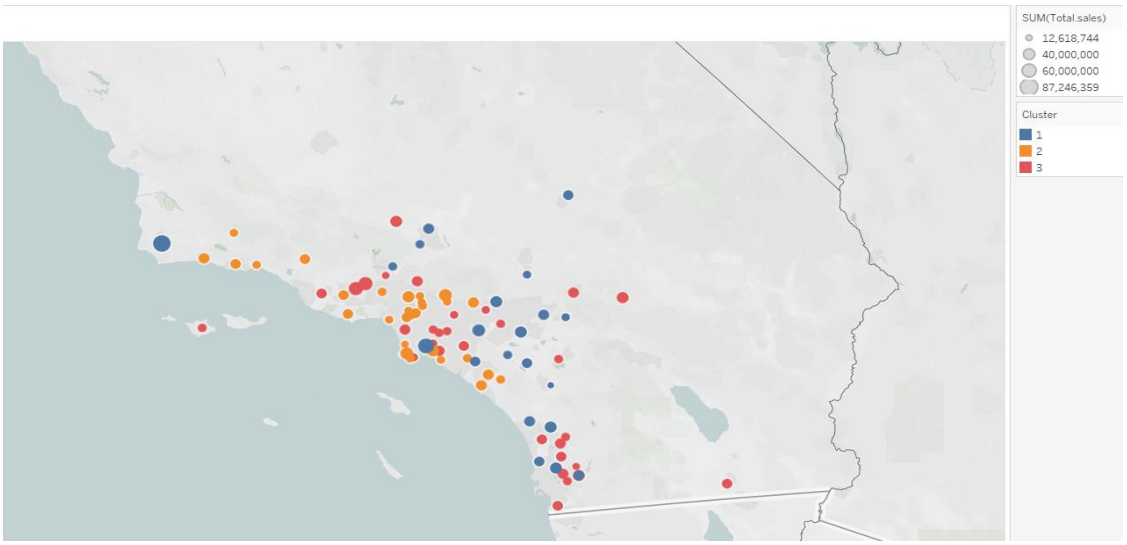
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	per_dry_grocery	per_dairy	per_produce	per_floral	per_deli	per_bakery	pre_general_merchandise
1	0.327833	-0.761016	-0.509185	-0.301524	-0.23259	-0.894261	1.208516
2	-0.730732	0.702609	1.014507	0.851718	-0.554641	0.396923	-0.304862
3	0.413669	-0.087039	-0.53665	-0.538327	0.64952	0.274462	-0.574389

	per_frozen	per_meat
1	-0.389209	-0.086176
2	0.345898	-0.485804
3	-0.032704	0.48698

From category and total sales plot, I can't see too much difference between clusters. But form convergence after 12 iterations. Them more positive value will be more sales for category. So for cluster 1, gen_general_merchandise more sales. For cluster 2, produce more sales. For cluster 3. Deli and meat more 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.



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?

I chose to use decision tree, forest model and boosted model, then to compare which one is the best. As the results shows below, boosted model is used for prediction, because although it's same accuracy with forest model, boosted model has a higher F1 value(F1 Score is the weight average of precision and recall).

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

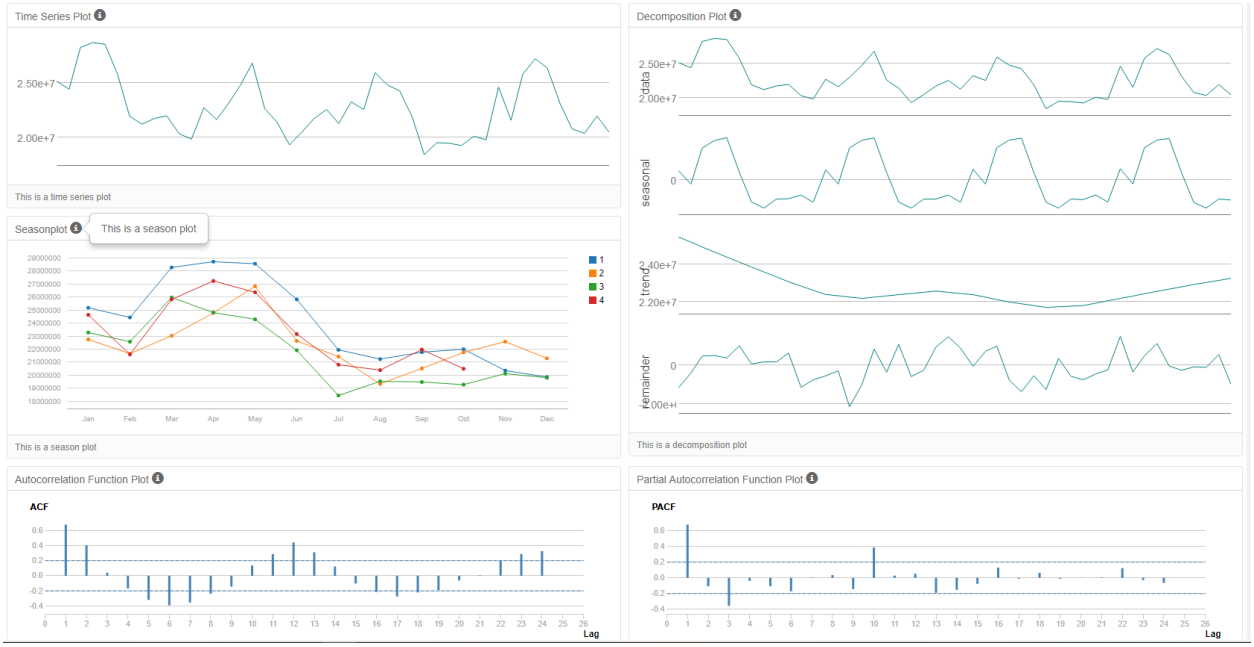
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

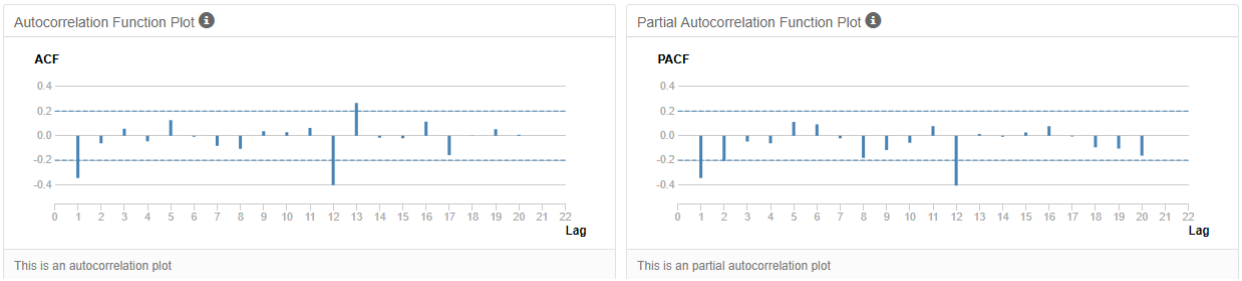
Task 3: Predicting Produce Sales

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?

I'd like to use both ETS and ARIMA models then compare the performance which is better.
Firstly, I use ETS model. As I see from decomposition plot, the error is looks randomly distributed .so I use M. for trend, N for no trend. For seasonally, it's not constant, so M. lastly, I set trend dampening to auto.
ETS Model to use (M, N, M)



Then I use ARIAM to ues (0,1,2)(0,1,0)12 based on below ACF and PACF.



Now it's time to compare the ETS model and ARIMA model.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	2539816	2992649	2647354	10.9832	11.5185	1.6938
ETS	1761302	1978476	1761302	7.5704	7.5704	1.1269

The ETS model's MASE, RMSE value is lower than ARIMA. We should use ETS model to forecast.

3. 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 it's the table includes exist stores and new stores forecast.

Year	Month	Exist stores	New stores
2016	1	21829060.031666	2652563.160464
2016	2	21146329.631982	2537099.6406574
2016	3	23735686.93879	2945431.74602387
2016	4	22409515.284474	2838360.72892148
2016	5	25621828.725097	3203949.43173604
2016	6	26307858.040046	3267698.32835245
2016	7	26705092.556349	3281573.7866757
2016	8	23440761.329527	2892667.4278735
2016	9	20640047.319971	2569171.0357474
2016	10	20086270.462075	2500063.1172564
2016	11	20858119.95754	2584992.31693966
2016	12	21255190.244976	2577896.22824041

The chart below shows all data about exists and new opening store produce sales information.