

# 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?

Analysis is performed using the K-Centroid Diagnostics Tool and clustering method as K-means.

Based on the K-means report, Adjusted Rand and Calinski-Harabasz indices, the optimal number of store formats is 3 because both the indices have the highest median value.

### K-Means Cluster Assessment Report

#### Summary Statistics

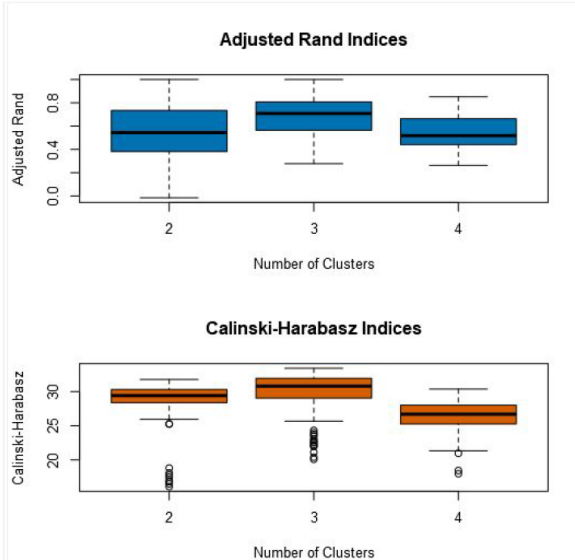
Adjusted Rand Indices:

	2	3	4
Minimum	-0.01556	0.2776	0.2631
1st Quartile	0.3889	0.5697	0.4424
Median	0.544	0.7083	0.518
Mean	0.5099	0.6863	0.5469
3rd Quartile	0.7244	0.8065	0.6618
Maximum	1	1	0.8516

Calinski-Harabasz Indices:

	2	3	4
Minimum	16.1	20.09	17.99
1st Quartile	28.37	29.14	25.28
Median	29.41	30.8	26.69
Mean	28.43	29.69	26.48
3rd Quartile	30.3	31.93	27.98
Maximum	31.78	33.41	30.37

#### Plots



2. How many stores fall into each store format?

Cluster 1 has 23 stores, cluster 2 has 29 stores while cluster 3 has 33 stores.

#### 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 the highest total sales when compared to Cluster 2 stores and Cluster 3 stores.  
Cluster 1 stores sold more General Merchandise in terms of percentage while Cluster 2 stores sold more Produce.

Record Report

1

Summary Report of the K-Means Clustering Solution K

2

Solution Summary

3

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"))

4

Cluster Information:

5

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

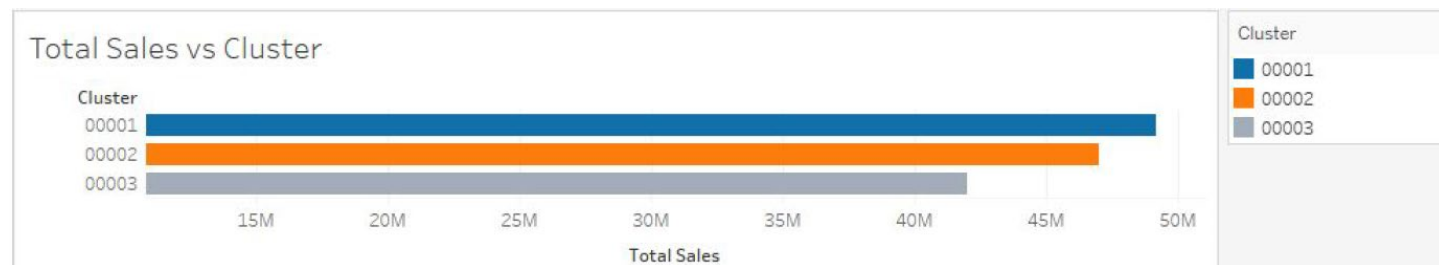
6

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

7

	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					

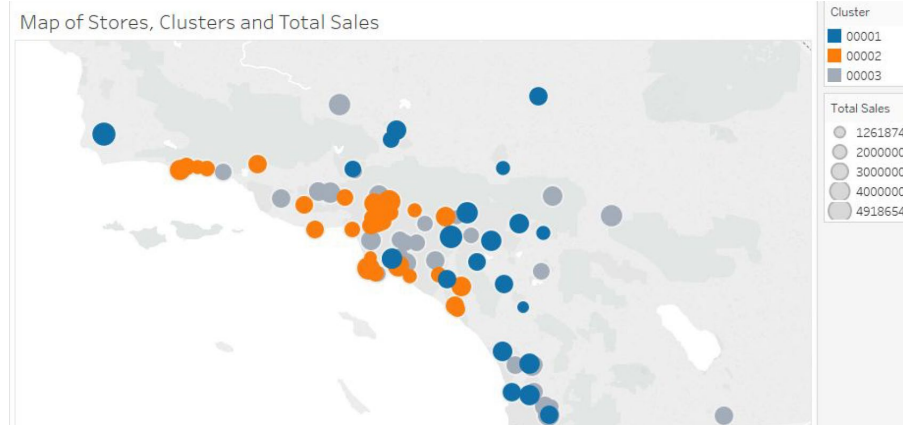
## Tableau Visualization



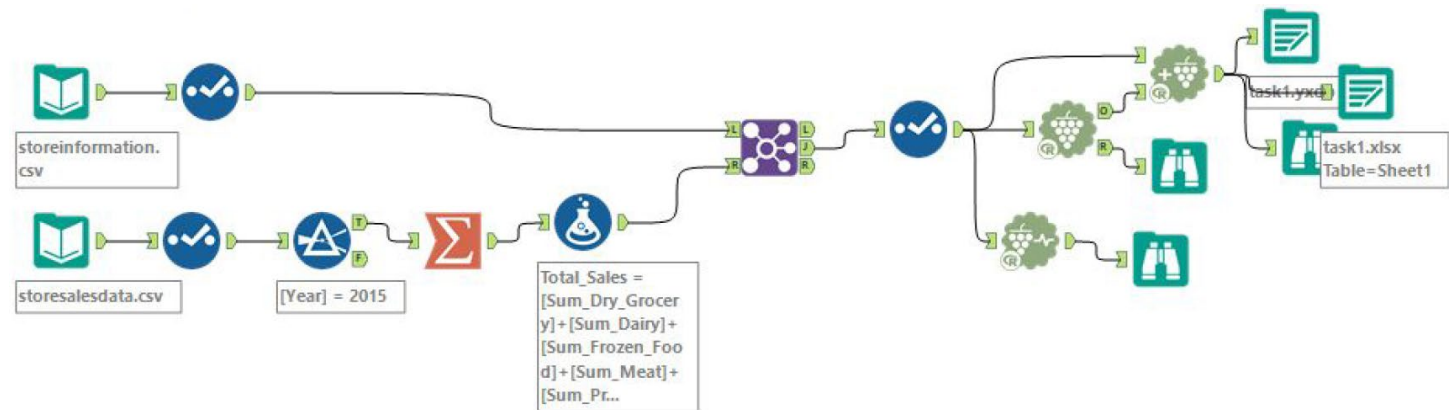
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

Tableau Public Profile

<https://public.tableau.com/profile/sing.min.hong#!/>



Alteryx Workflow



## 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.

All the models have same accuracy but Boosted Model has the highest F1 value.

Boosted Model is chosen.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750
Decision_Tree	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000

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

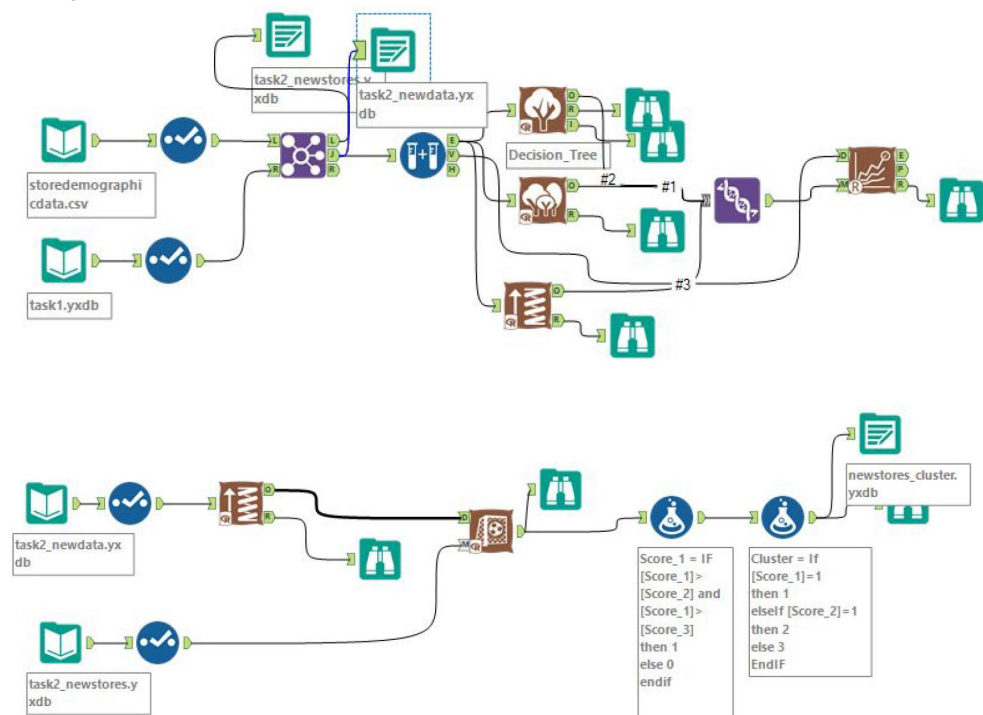
Confusion matrix of Decision_Tree			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

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

## Alteryx Workflow



## 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) with no dampening** is used for ETS model.

The seasonality shows increasing trend and should be applied multiplicatively.

The trend is not clear and nothing should be applied.

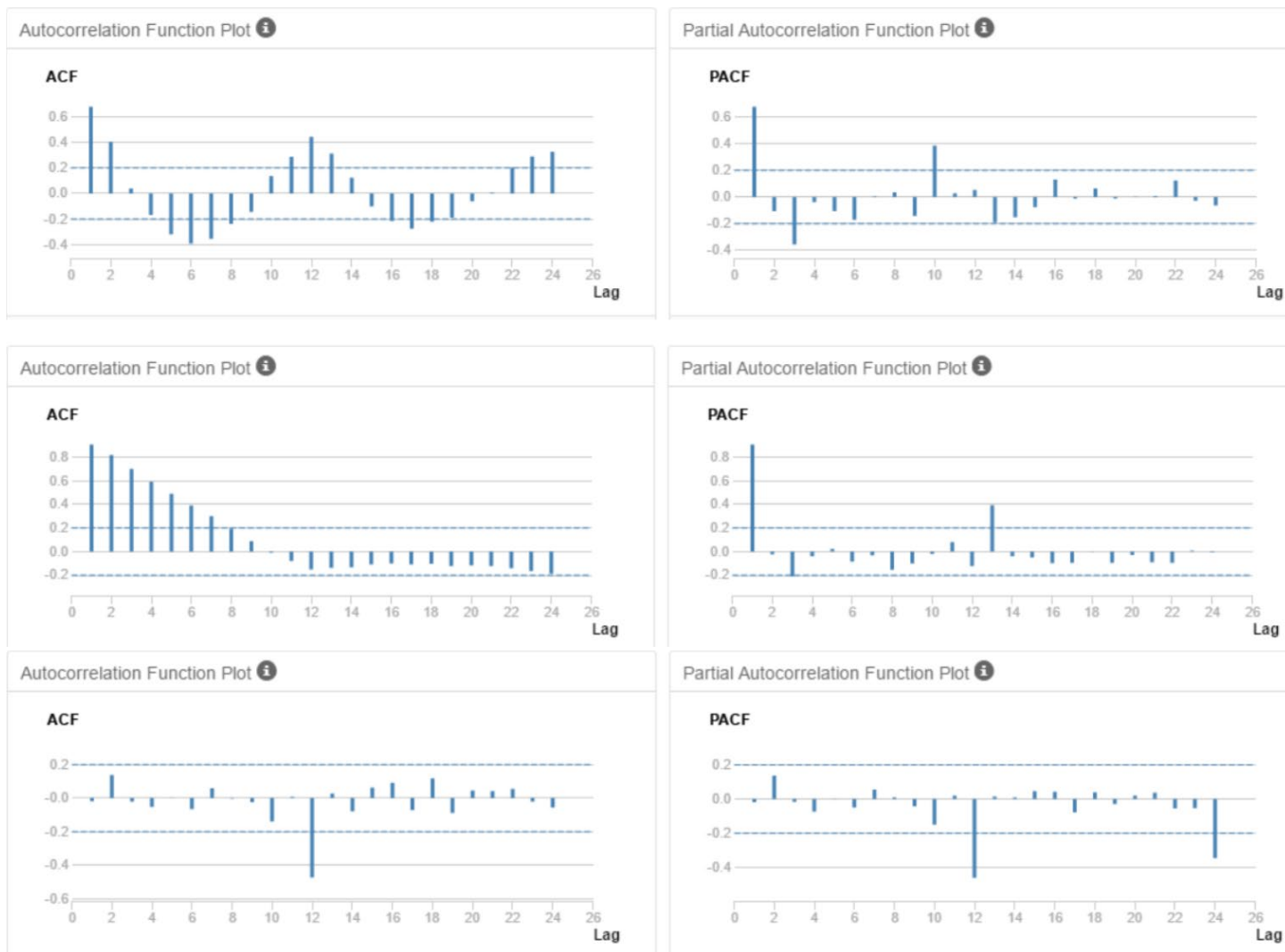
Its error is irregular and should be applied multiplicatively.

ETS Model (M,N,M) is chosen.

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.



**ETS model's accuracy is higher** when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of **1,020,597** is lower than ARIMA's **1,429,296** while its MASE is **0.45** compared to ARIMA's **0.53**. ETS also has a higher AIC at **1,283** while ARIMA's AIC is **859**.

Method:  
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

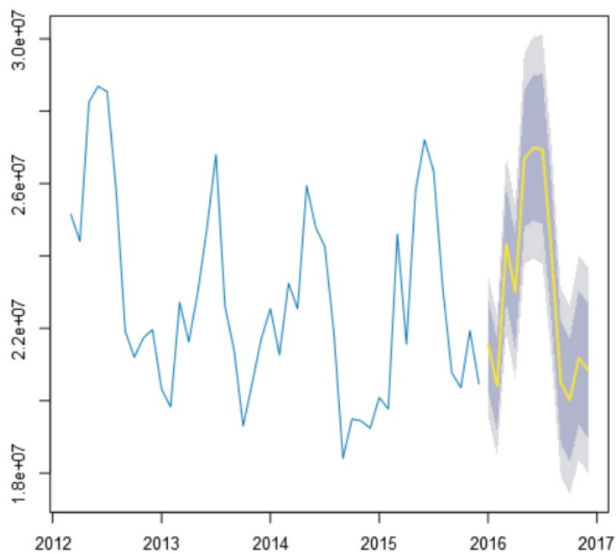
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.5311117	-0.0260961

The graph and table below shows actual and forecast value with 80% & 95% confidence level interval.



Forecasts from ETS



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

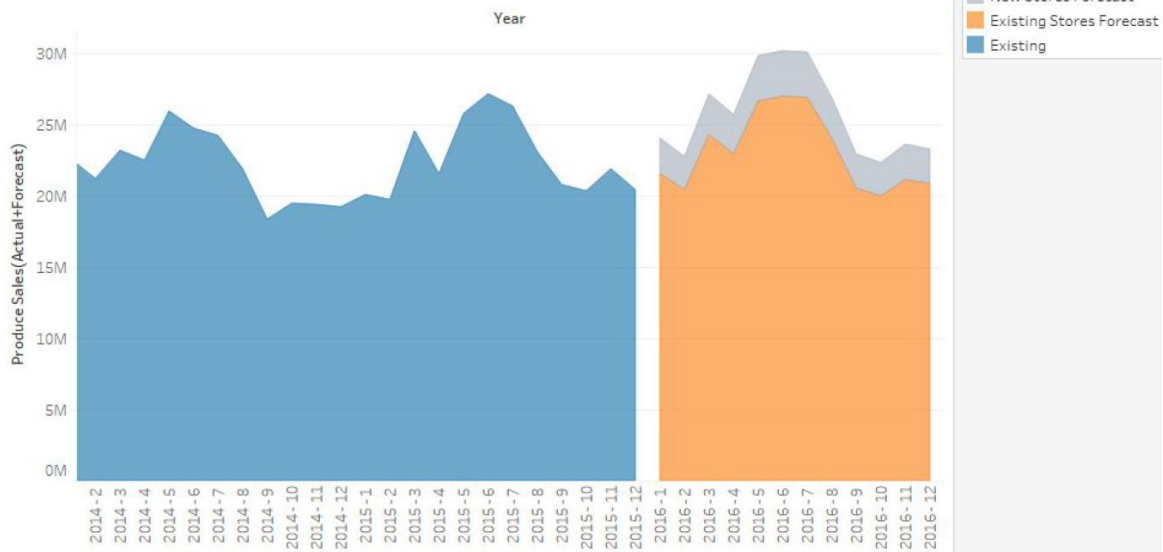
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Period	Month	Existing Stores Sales	New Stores Sales	Total Produce Sales
2016	1	21539936.01	2534110.119	24074046.13
2016	2	20413770.6	2401620.071	22815390.67
2016	3	24325953.1	2861876.835	27187829.93
2016	4	22993466.35	2705113.688	25698580.04
2016	5	26691951.42	3140229.579	29832181
2016	6	26989964.01	3175289.884	30165253.89
2016	7	26948630.76	3170427.149	30119057.91
2016	8	24091579.35	2834303.453	26925882.8
2016	9	20523492.41	2414528.519	22938020.93
2016	10	20011748.67	2354323.373	22366072.04
2016	11	21177435.49	2491462.998	23668898.48
2016	12	20855799.11	2453623.425	23309422.53

Tableau Visualization



Produce Sales vs Year



<https://public.tableau.com/profile/sing.min.hong#!/>

## Alteryx Workflow

