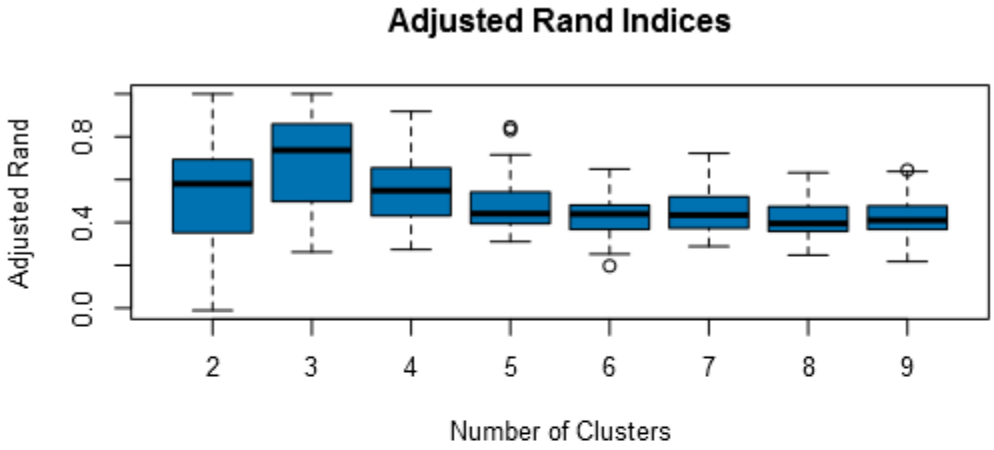
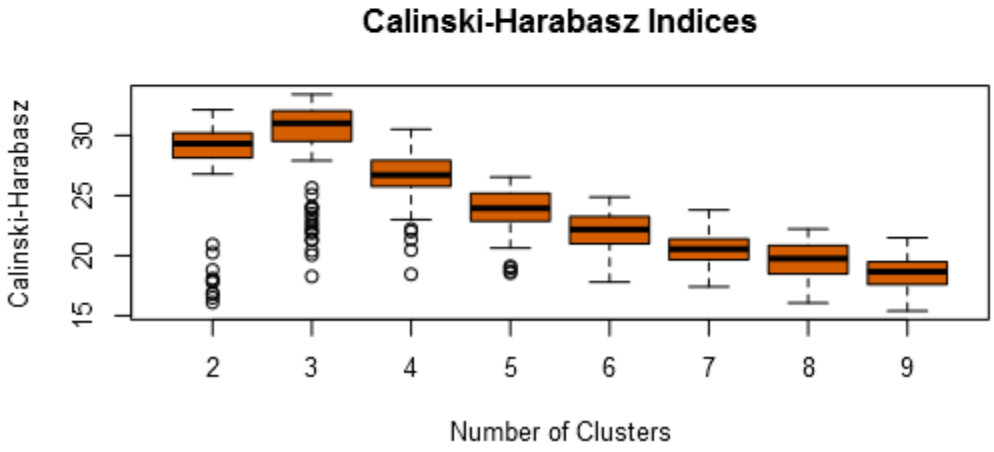
Wilson

Combo Proj – Segmentation, Classification, Forecasting

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

The optimal number is 3 Clusters. I selected all %categories to sale and used K-Centroid Diagnostics to determine the optimal number of clusters.





Based on the AR and CH indices graph above, 3 clusters have the highest mean and median across both indexes.

2. How many stores fall into each store format?

Cluster 1 – 23 Stores

Cluster 2 – 29 Stores

Cluster 3 – 33 Stores

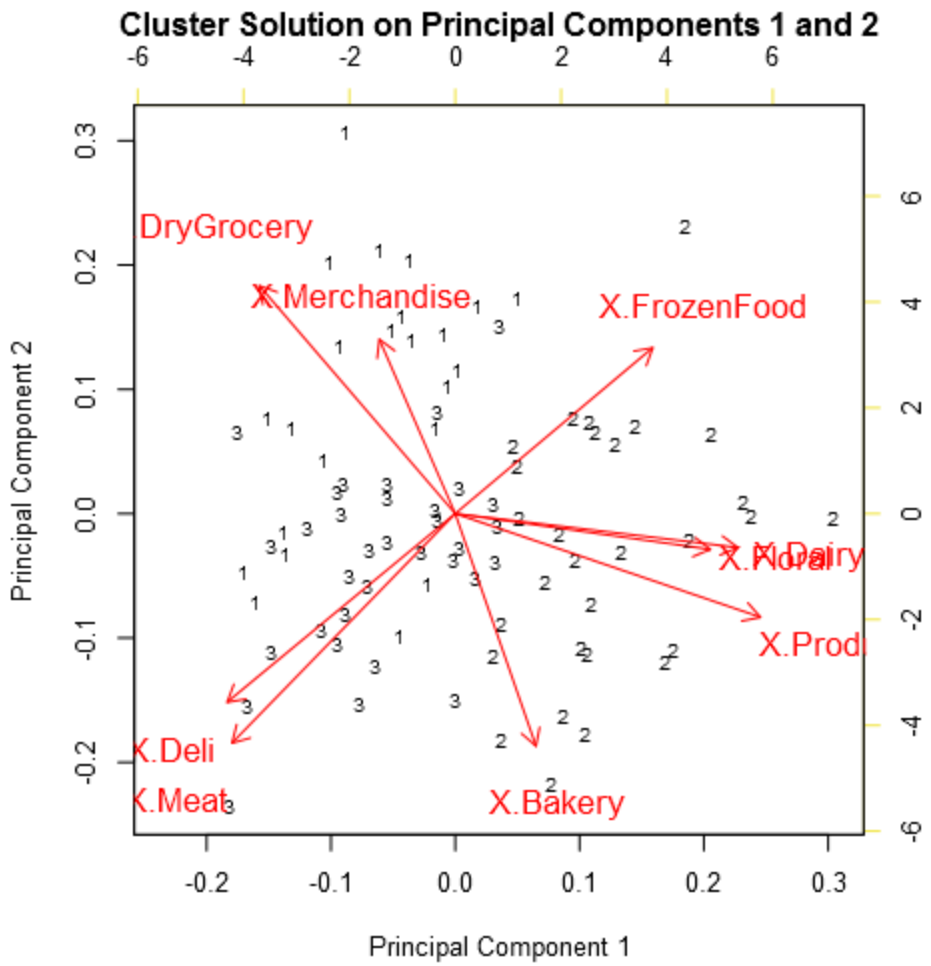
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

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.

	X.DryGrocery	X.Dairy	X.FrozenFood	X.Meat	X.Produce	X.Floral	X.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

	X.Bakery	X.Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389



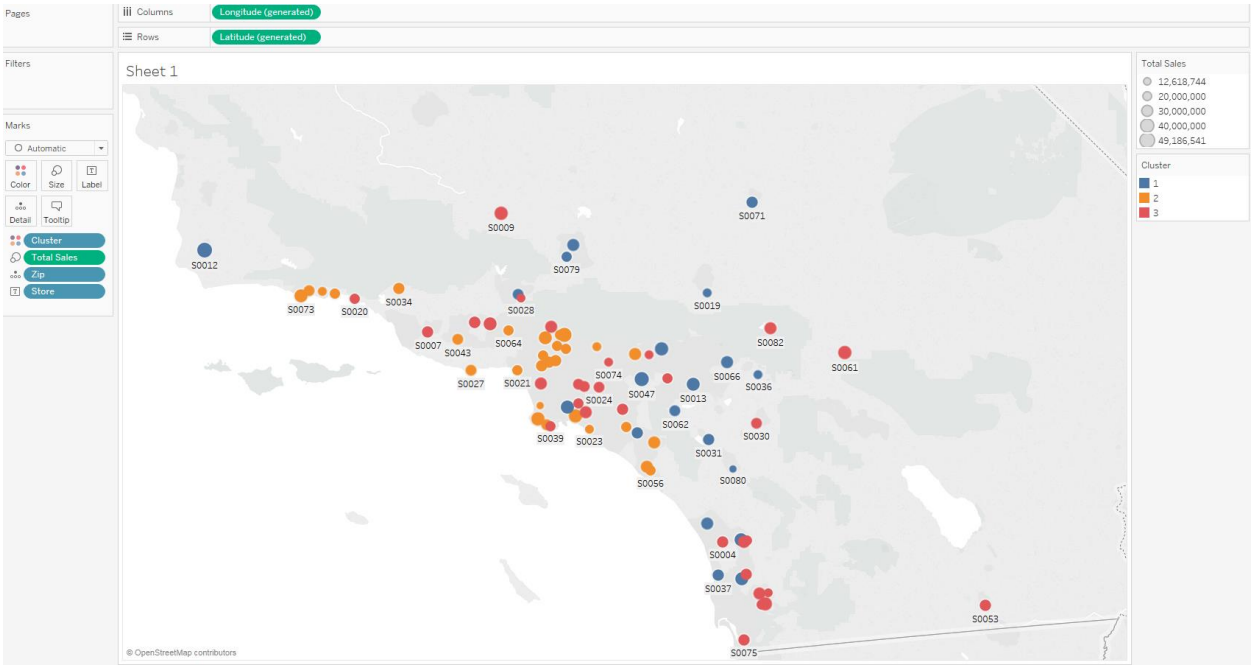
Cluster 1 is selling more general merchandise compared to other two.

Cluster 2 is selling more bakery compared to other two.

Cluster 3 is selling less general merchandise and bakery items compared to cluster 1 and 2.

4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.

https://public.tableau.com/views/FinalProj_Cluster/Sheet1?:embed=y&:display_count=yes&publish=yes



5. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?

Fit and error measures

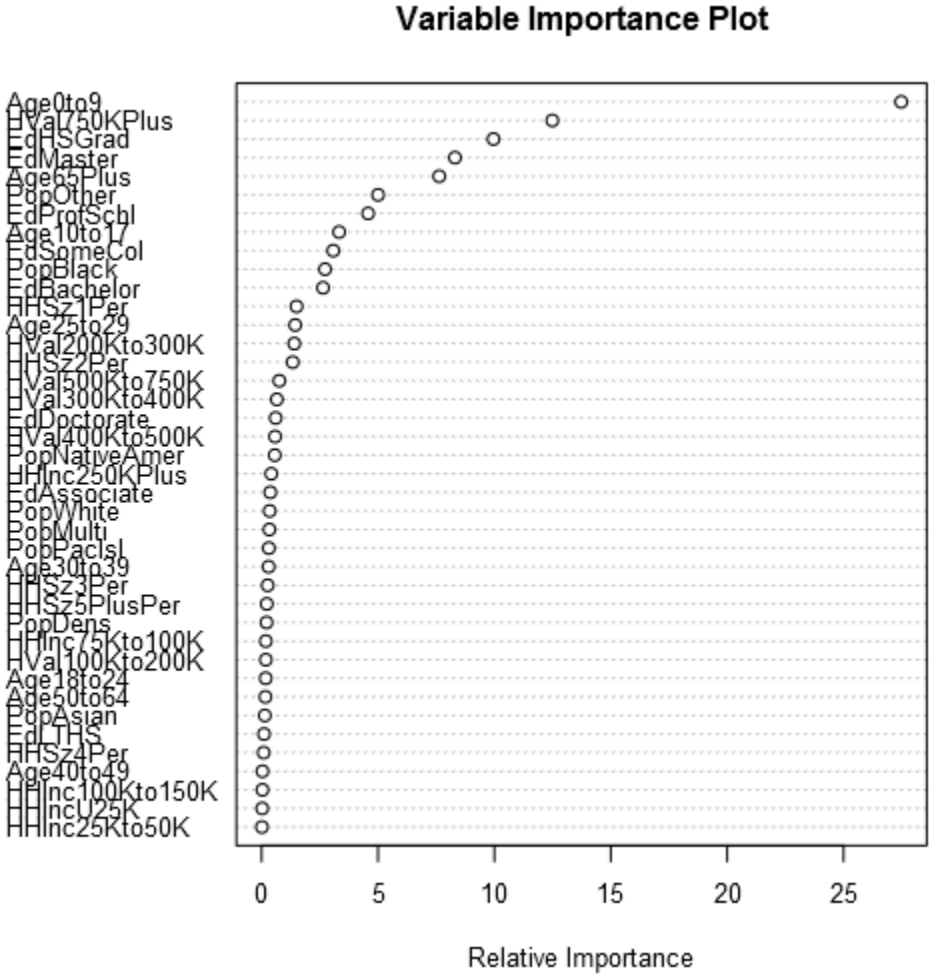
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Cluster	0.8235	0.8251	0.7500	0.8000	0.8750
FM_Cluster	0.8235	0.8251	0.7500	0.8000	0.8750
BM_Cluster	0.8235	0.8543	0.8000	0.6667	1.0000

Confusion matrix of BM_Cluster

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

The best model to classify the clusters is Boosted Model. It predicted 80% accuracy for Cluster 1, 66.7% accuracy for Cluster 2 and 100% accuracy for Cluster 3. Even though 3 models have the same 82.3% overall accuracy Boosted Model takes the edge with 85.4% F1 Score. F1 takes into account classification errors.

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



Based on Boosted Model, Top 3 variables are Age0to9, HVal750KPlus and EdHSGrad

7. What format do each of the 10 new stores fall into? Please provide a data table.

Store	Cluster
S0086	3

S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

8. 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 Existing Store Forecast I used ETS(M,N,M) model.

   I used TS plot to see the errors, trend and seasonality patterns over time.

From the graph above, Error is multiplicative because it's peaks and valleys changes over time. There seems to be no trend. Seasonality is multiplicative as the peaks are slightly growing each season.

ETS (M,N,M) in-sample measures

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

ARIMA(1,0,0)(1,1,0)[12] in-sample measures

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

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

ETS has a higher RMSE and MASE by 10%. Higher AIC than ARIMA by a few hundred.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS41	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA41	-604232.3	1050239	928412	-2.6156	4.0942	0.5463	NA

Validation accuracy ETS performed slightly better with MASE and RMSE. Less deviance than ARIMA. ETS validation sample performed better than ARIMA validation sample. Therefore I chose ETS model for Existing Store forecast.

9. 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	New Store	Existing Store
2016-01	2642246.94	21539936.01
2016-02	2475383.17	20413770.60
2016-03	2818081.15	24325953.1
2016-04	2618570.48	22993466.35
2016-05	2959074.66	26691951.42
2016-06	3039521.59	26989964.01
2016-07	3065160.45	26948630.76
2016-08	2700216.72	24091579.35
2016-09	2403658.78	20523492.41

2016-10	2347203.87	20011748.67
2016-11	2489364.86	21177435.49
2016-12	2535307.21	20855799.11

https://public.tableau.com/views/FProject_Forecast/Sheet1?:embed=y&:display_count=yes&publish=yes