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?

   Based on the results below, the optimal number is 3.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

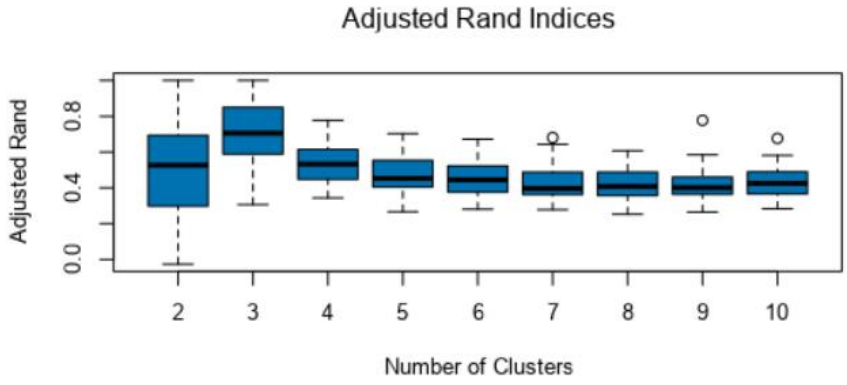
	2	3	4	5	6	7	8
Minimum	-0.025874	0.306842	0.343542	0.266826	0.280496	0.278231	0.252511
1st Quartile	0.297559	0.589274	0.44787	0.40537	0.376447	0.363932	0.358099
Median	0.526763	0.705458	0.532705	0.452807	0.444997	0.396862	0.408013
Mean	0.50356	0.703985	0.537325	0.473925	0.455989	0.434435	0.422232
3rd Quartile	0.694159	0.849878	0.613802	0.550618	0.519704	0.487303	0.486396
Maximum	1	1	0.777747	0.702984	0.672136	0.680929	0.607571

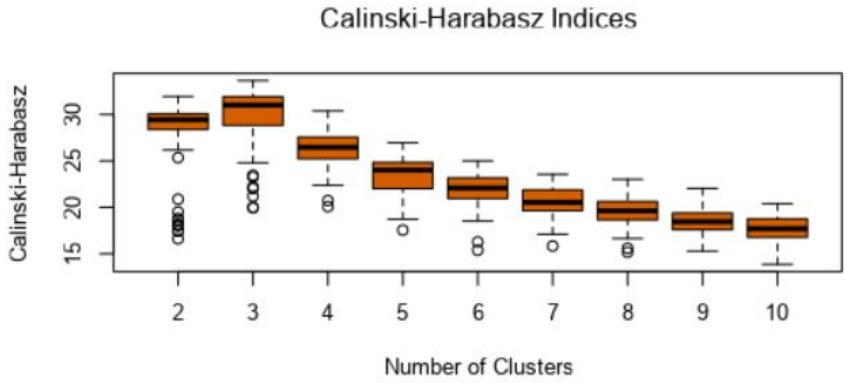
	9	10
Minimum	0.264847	0.283873
1st Quartile	0.364209	0.368813
Median	0.401348	0.425056
Mean	0.416707	0.428736
3rd Quartile	0.458573	0.488903
Maximum	0.777547	0.676411

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.61829	19.94748	20.07469	17.55277	15.38909	15.83815	15.17814
1st Quartile	28.42185	28.92031	25.25672	22.02319	20.95706	19.64379	18.64939
Median	29.44042	31.00639	26.47092	24.01173	22.08642	20.53428	19.61483
Mean	28.36087	29.95459	26.27505	23.41569	21.95238	20.59125	19.59946
3rd Quartile	30.08768	31.91091	27.56085	24.84598	23.17782	21.87173	20.62091
Maximum	31.94034	33.63781	30.37935	26.97019	25.00769	23.55413	23.01808

	9	10
Minimum	15.29594	13.87458
1st Quartile	17.59888	16.78185
Median	18.46684	17.70345
Mean	18.47646	17.69964
3rd Quartile	19.38663	18.75893
Maximum	22.04859	20.40397





2. How many stores fall into each store format?

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

From the results I got, cluster 1 has 23 stores and 29 stores in cluster 2, 33 stores in cluster 3.

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

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

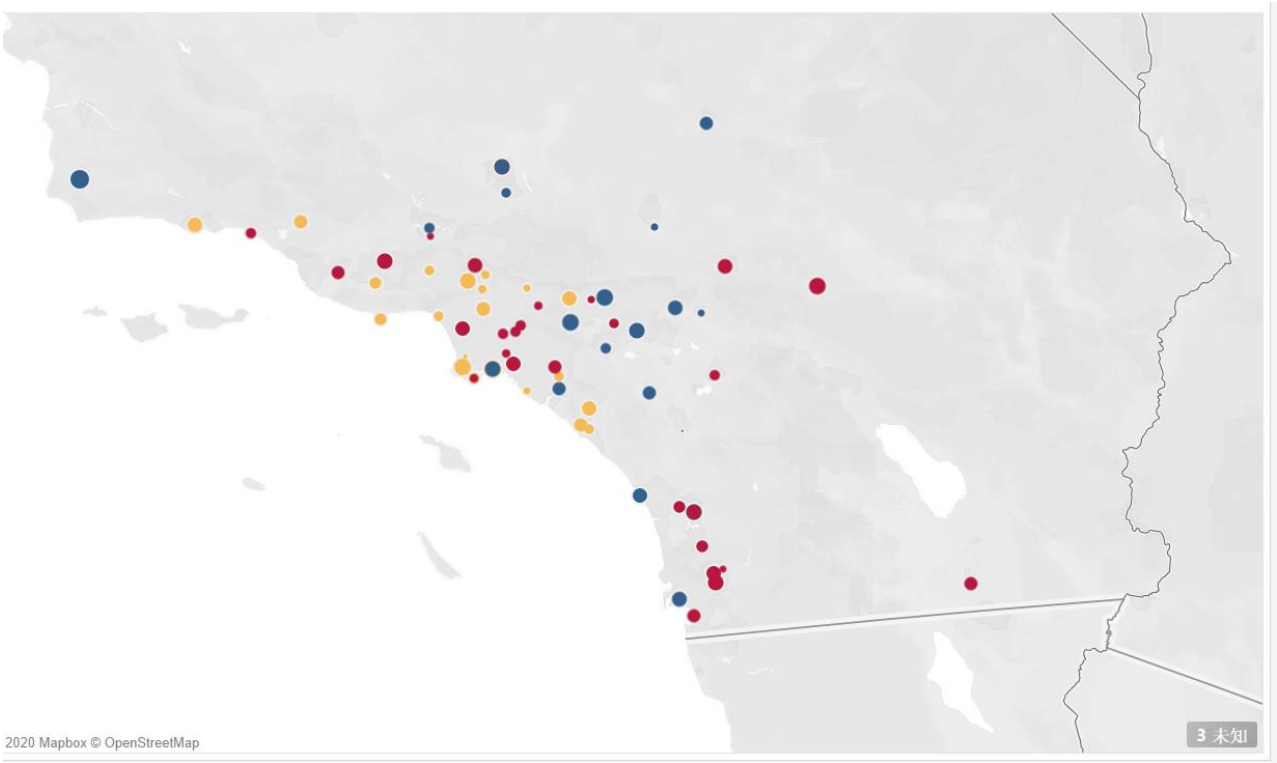
	Sum_Produce	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Floral	Sum_Deli
1	-0.509185	0.327833	-0.761016	-0.389209	-0.086176	-0.301524	-0.23259
2	1.014507	-0.730732	0.702609	0.345898	-0.485804	0.851718	-0.554641
3	-0.53665	0.413669	-0.087039	-0.032704	0.48698	-0.538327	0.64952

	Sum_Bakery	Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Based on the results I got, cluster 1 sold more general_merchandise.

Cluster 2 sold more produce and cluster 3 sold more deli.

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

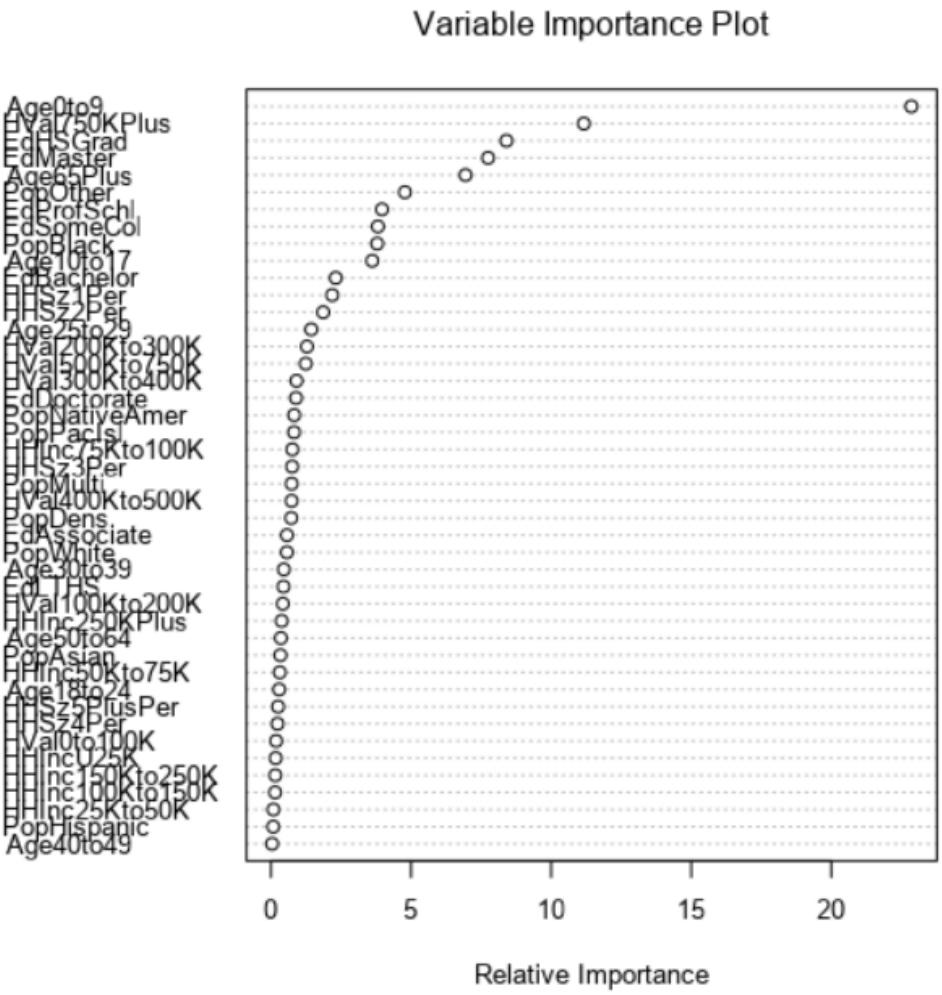
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

Model Comparison Report

Fit and error measures

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

Based on the model comparison results, the best model is the boosted model with 0.8235 accuracy, 0.8889 of F1 and 1.0 accuracy_1.



And Age0to9 is the most important variable.

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

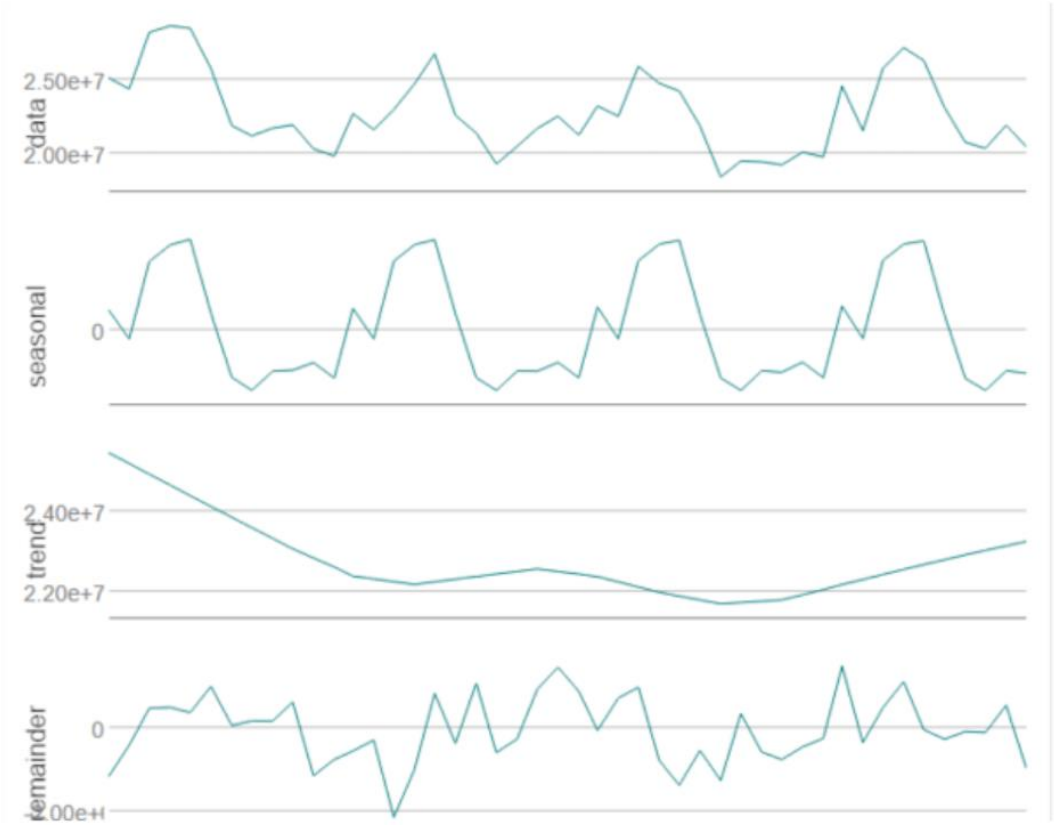
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1

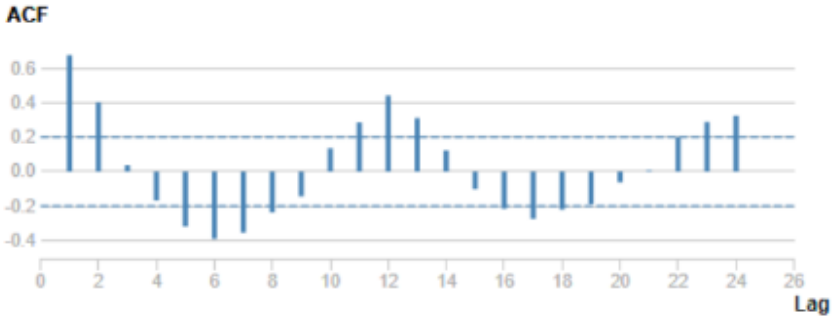
S0094	2
S0095	2

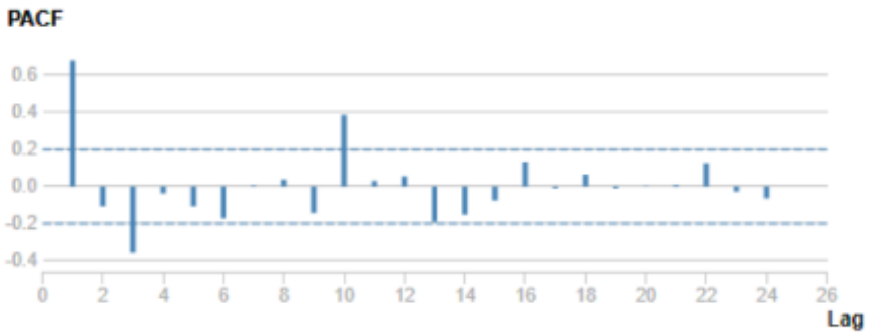
Task 3: Predicting Produce Sales

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?

ETS (M,N,M) without dampening is used for ETS model.







Based on the TS plots, I have discovered the parameters as (0,1,2)(0,1,0).
Now, we can compare two predictions:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

ETS has better accuracy overall.

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	Sub_Period	New Store Sales Forecast	Existing Store Sales Forecast
2016	1	2 588 356.56	21 829 060.03
2016	2	2 498 567.17	21 146 329.63
2016	3	2 919 067.02	23 735 686.94
2016	4	2 797 280.08	22 409 515.28
2016	5	3 163 764.86	25 621 828.73
2016	6	3 202 813.29	26 307 858.04
2016	7	3 228 212.24	26 705 092.56
2016	8	2 868 914.81	23 440 761.33
2016	9	2 538 372.27	20 640 047.32
2016	10	2 485 732.28	20 086 270.46
2016	11	2 583 447.59	20 858 119.96
2016	12	2 562 181.70	21 255 190.24