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?**

The optimal number of store formats is 3. I arrived at this number by using a K-Means Cluster with past data of stores. From the Adjusted Rand and Calinski-Harabasz Indices below, you can see that 3 store formats had the highest median value.

K-Means Cluster Assessment Report

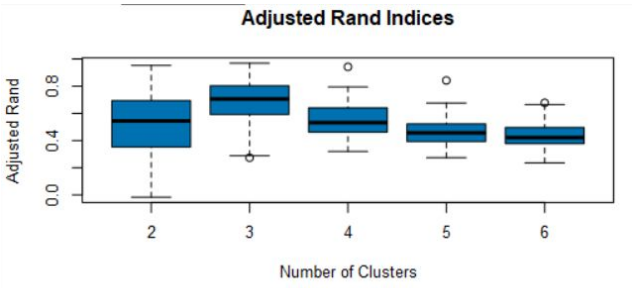
Summary Statistics

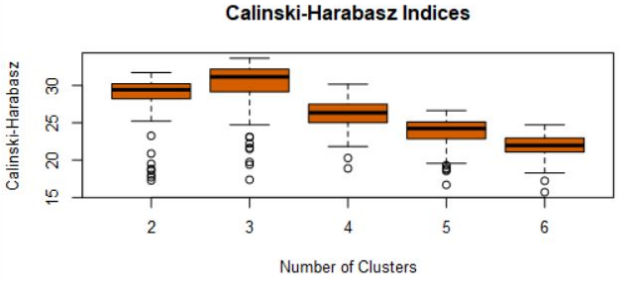
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.281	17.38103	18.89398	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.03471	22.86498	21.10249
Median	29.4157	31.14178	26.33467	24.22188	21.96958
Mean	28.56936	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63781	30.1583	26.63063	24.72038





2. **How many stores fall into each store format?**

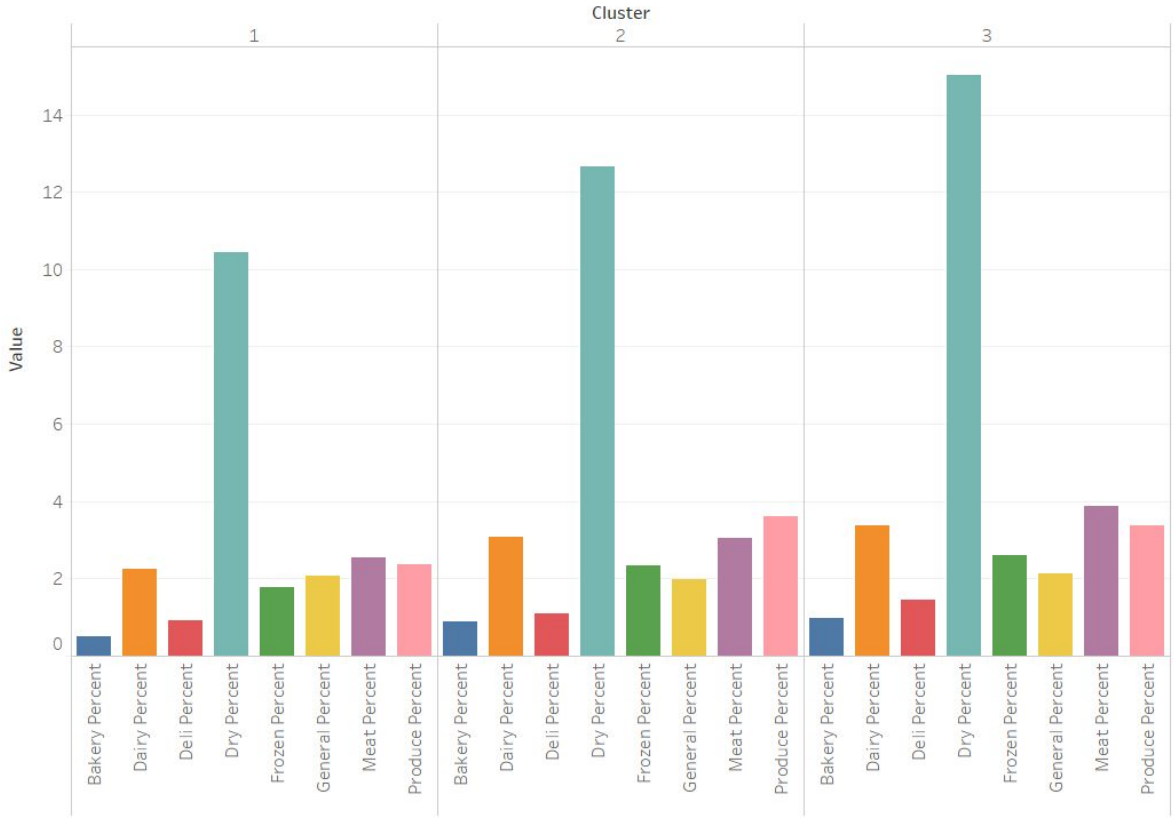
Cluster Information:

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

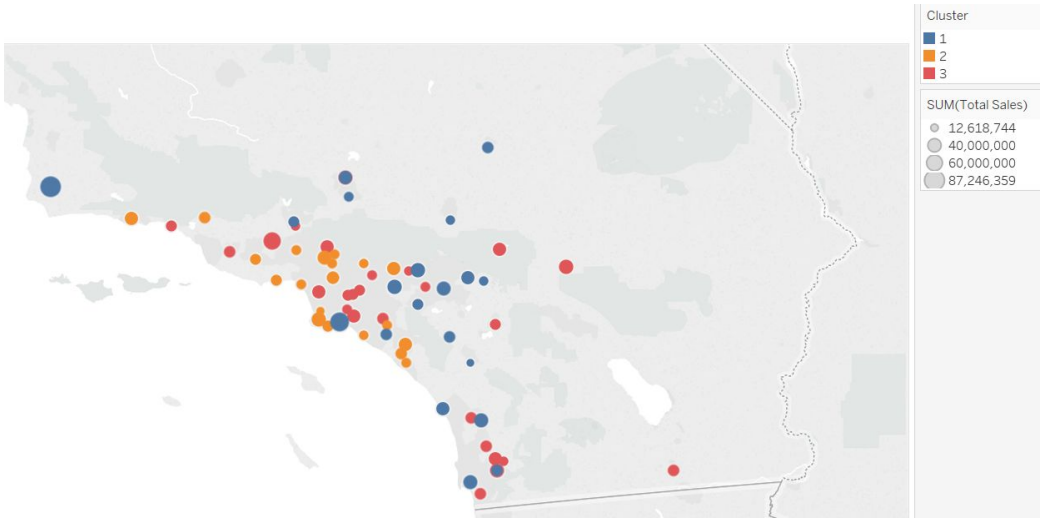
23 stores fall into Cluster 1, 29 stores fall into Cluster 2 and 33 stores fall into Cluster 3.

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

From the graph below, you cans see that stores in Cluster 3 sold more Dry Food than stores in the other clusters, while stores in Cluster 2 sold more Produce than the others. Additionally, Cluster 1 sold less of all categories besides General Merchandise when compared to both Clusters 2 and 3.



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.)**

   After comparing the Decision Tree, Forest and Boosted models, I decided to use the Boosted model to predict the best store format for new stores. In the model comparison below, the Forest Model and Boosted Model both had the same accuracy, however the Boosted Model's higher F1 accuracy made it the better choice.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_8	0.7059	0.7685	0.7500	1.0000	0.5556
FM_cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Cluster	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.**

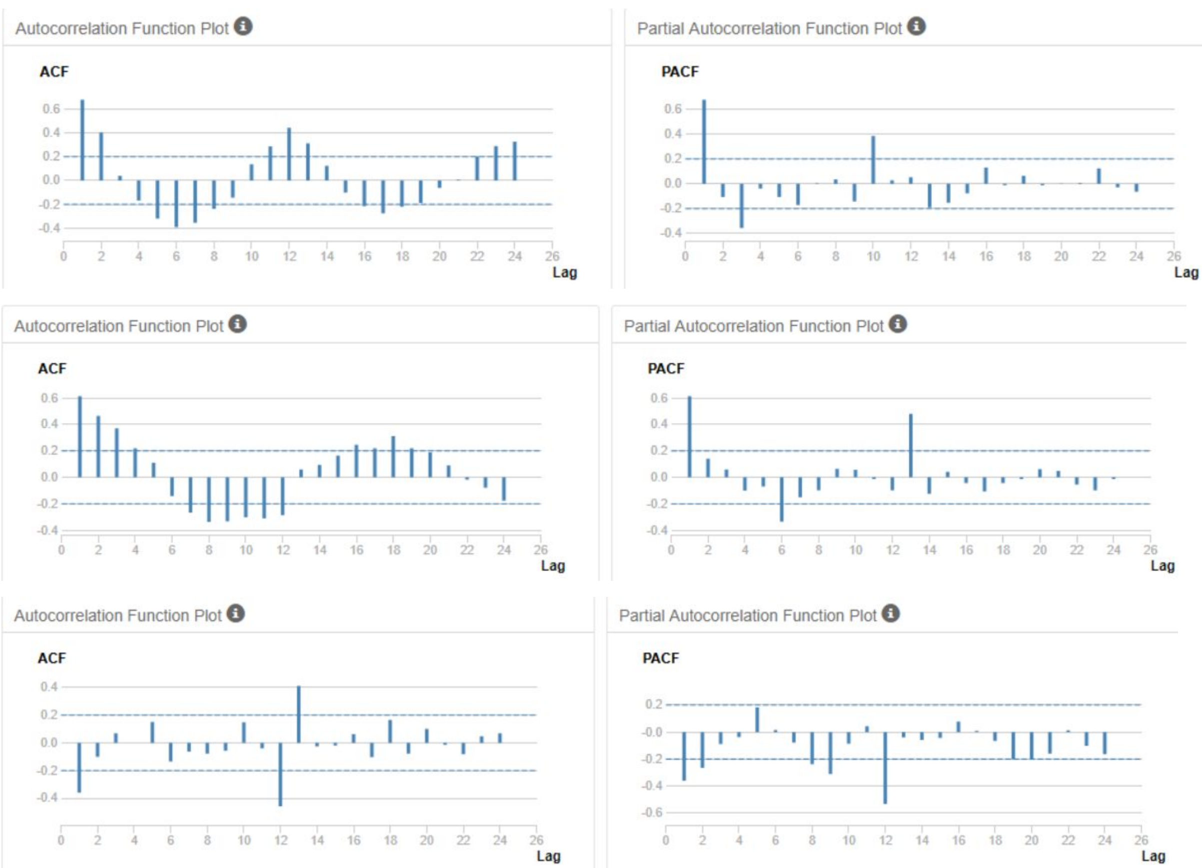
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?

For the ETS model I used an ETS(M,N,M) as the seasonality increased and decreased with the level of time series, there was no trend and the error fluctuates over time. Therefore the seasonality is Multiplicative, the trend is None and the error is Multiplicative.

For the ARIMA model I used ARIMA(1,0,0)(1,1,0)



In the end, I decided to use the ETS(M,N,M) model rather than the ARIMA(1,0,0)(1,1,0). The ETS model has lower RMSE, MASE and MAPE scores, making it the better choice between the two.

Accuracy Measures:

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

[ETS error]

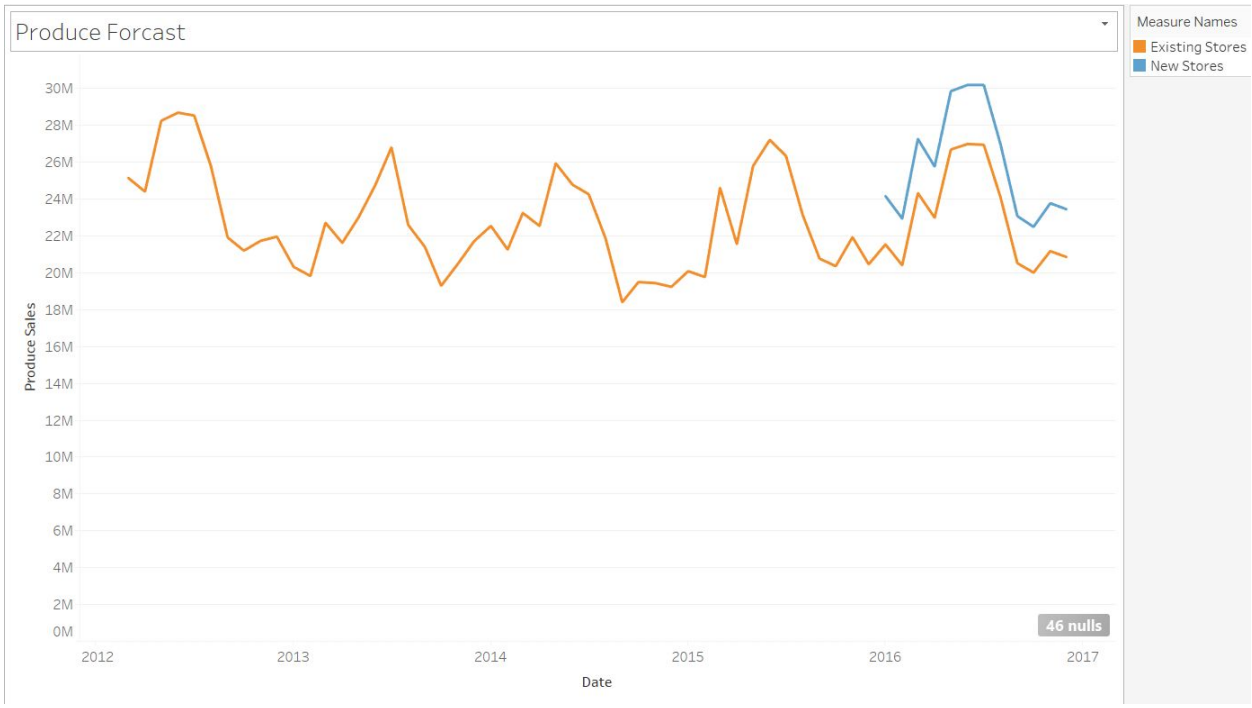
Accuracy Measures:

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

[ARIMA Error]

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.

Month	New Stores	Existing Stores
Jan-16	2,626,198	21,539,936
Feb-16	2,529,186	20,413,770
Mar-16	2,940,264	24,325,953
Apr-16	2,774,135	22,993,466
May-16	3,165,320	26,691,951
Jun-16	3,204,286	26,989,964
Jul-16	3,244,464	26,948,630
Aug-16	2,871,488	24,091,579
Sep-16	2,552,418	20,523,492
Oct-16	2,482,837	20,011,748
Nov-16	2,597,780	21,177,435
Dec-16	2,591,815	20,855,799



Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.