Project: Predictive Analytics Capstone

Using Alteryx 2021.3

Task 1: Determine Store Formats for Existing Stores

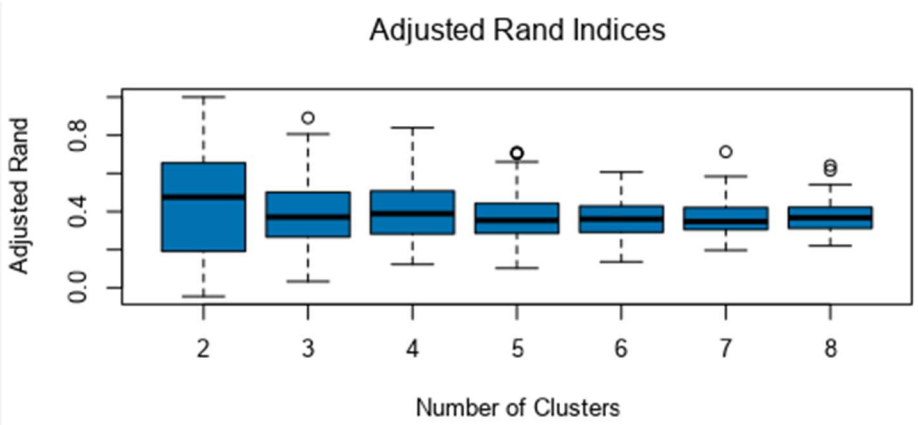
1. **What is the optimal number of store formats? How did you arrive at that number?**

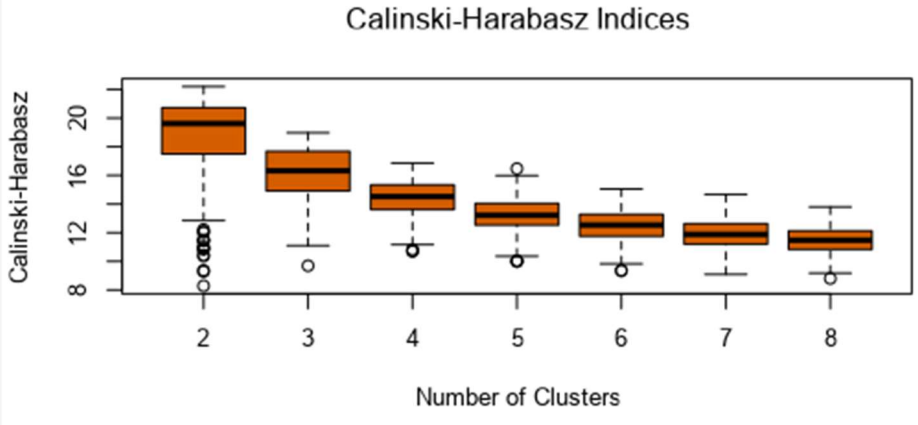
   Optimum number of store formats is 3.

   The project explicitly states to use K-Mean clustering only, so the other models, K-Median and Neural Gas, were not evaluated. The number of clusters explored ranged from 2 to 8 and for each number of clusters, their Adjusted Rand (AR) Indices and Calinski-Harabasz (CH) Indices were evaluated.

   While the number of clusters of 2 may seem to be an appropriate number based on the highest AR and CH indices, the high number of outliers (open circles) at the lower fence of the box plot for the CH index for cluster 2 may indicate high probability for deviations if this number of clusters is chosen. Instead, a safer choice is the next best that is 3 clusters.

   Note: the version of Alteryx and random seeds used can affect the clustering results.





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

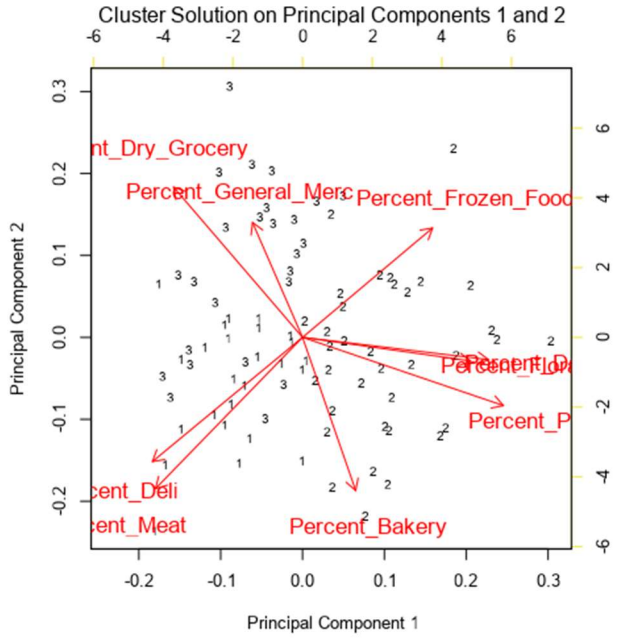
Cluster	Size
1	25
2	35
3	25

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

From the principal component relationships between component 1 and 2 in the figure below, there is a similar co-relationship for PC1 and PC2 in Store Format 1 for Meat and Deli. Also, the median percent for Meat and Deli are higher than for the other 2 Store Formats by about 0.9 and 0.7 basis points respectively

For Store Format 2, Floral, Produce and Dairy are the strong sellers and their median percent sales are higher than the other 2 by 0.2, 2.0 and 0.5 basis points respectively.

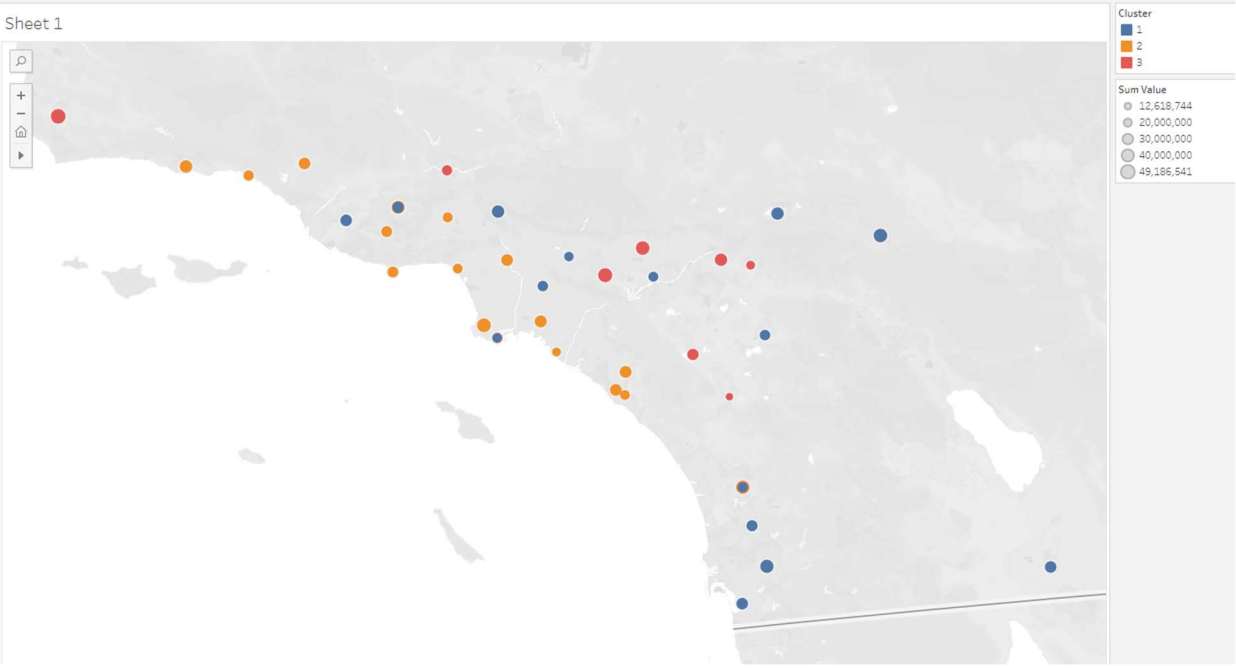
For Store Format 3, General Merchandise are the top seller and their median percent sale are higher than the other by 2.0 basis points.



Cluster	Median_Percent_Dry_Grocery	Median_Percent_Dairy	Median_Percent_Frozen_Food	Median_Percent_Meat
1	45.983515	10.075956	7.806358	11.692294
2	43.811699	10.538849	8.182537	10.811252
3	44.638398	9.782225	7.718514	10.812977

Median_Percent_Produce	Median_Percent_Floral	Median_Percent_Deli	Median_Percent_Bakery	Median_Percent_General_Merc
10.12594	0.651293	4.605237	3.017637	6.371459
12.10597	0.929907	3.833571	2.90537	6.839854
10.186055	0.75587	3.982887	2.150065	8.87061

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.)**

   Boosted model is selected to predict the best store format for the new stores.

   Store demographics and clustering out from Task1 were joined by Store and the resulting dataset split into 80% training and 20% validation. The training dataset was used to train the Decision Tree, Random Forest and Boosted models using on the demographics data (double datatype) to predict the Cluster target variable (categorical datatype). The performance of each model was compared. Of the 3 models, the Decision Tree model performed worst with lowest Accuracy and F1 scores, while both the Random Forest and Boosted models performed best and equally based on Accuracy, F1 scores and Confusion Matrix. The equal performances of the 2 models is probably due to the low sample size of the dataset (85 training and 10 validation).

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Forest	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted	0.7059	0.7500	0.5000	1.0000	0.7500

Confusion matrix of Boosted

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

Confusion matrix of Decision_Tree

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	2
Predicted_2	3	5	0
Predicted_3	1	0	2

Confusion matrix of Forest

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

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

Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
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) was selected to forecast for the 3 Store Formats.

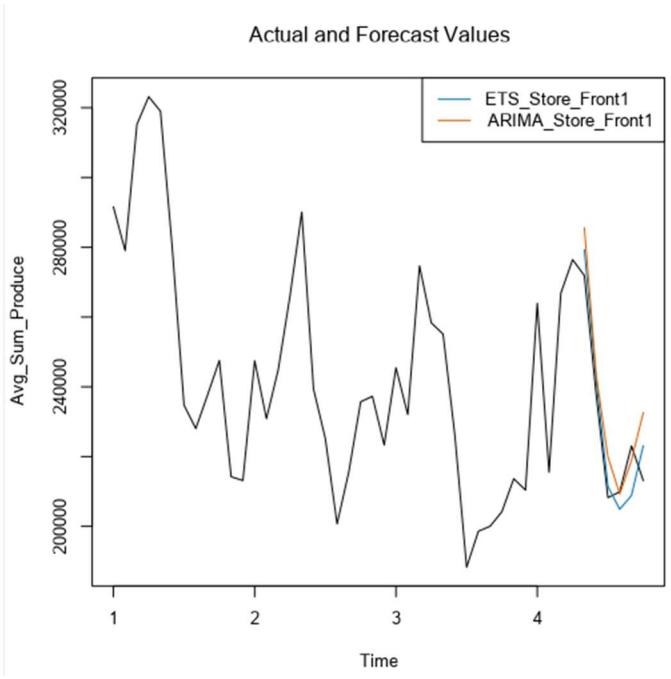
ETS and ARIMA models for the 3 Store Formats were compared using the last 6 months

of data that had been held out. The ETS model for each Store Front cluster consistently perform better than corresponding ARIMA with smaller error metrics including ME, RMSE, MAE, MPE, MAPE and MASE. The plots pf actual vs forecast values for the 3 Store Fronts also showed closer predictions for ETS than ARIMA models.

Store Format 1:

Accuracy Measures:

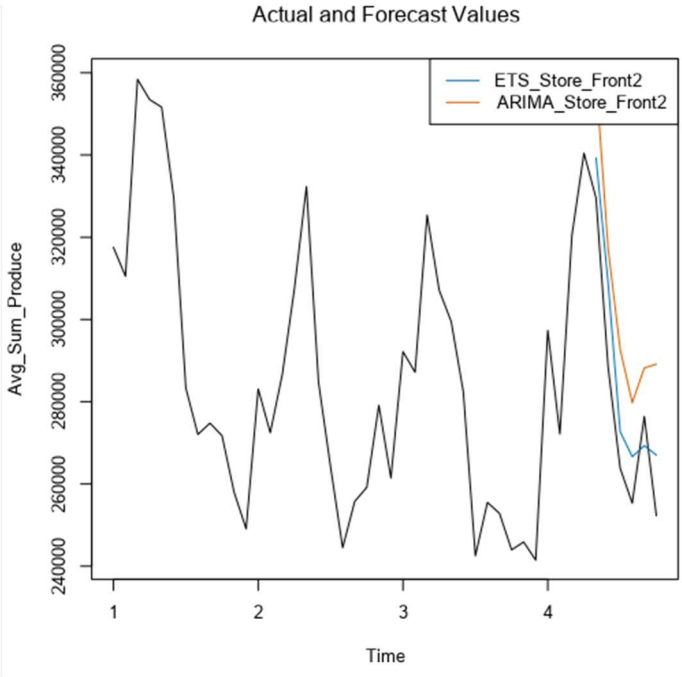
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Store_Front1	-932.0971	8258.898	7335.408	-0.3271	3.2448	0.3522
ARIMA_Store_Front1	-7617.9754	11204.122	9216.161	-3.3323	4.0547	0.4426



Store Format 2:

Accuracy Measures:

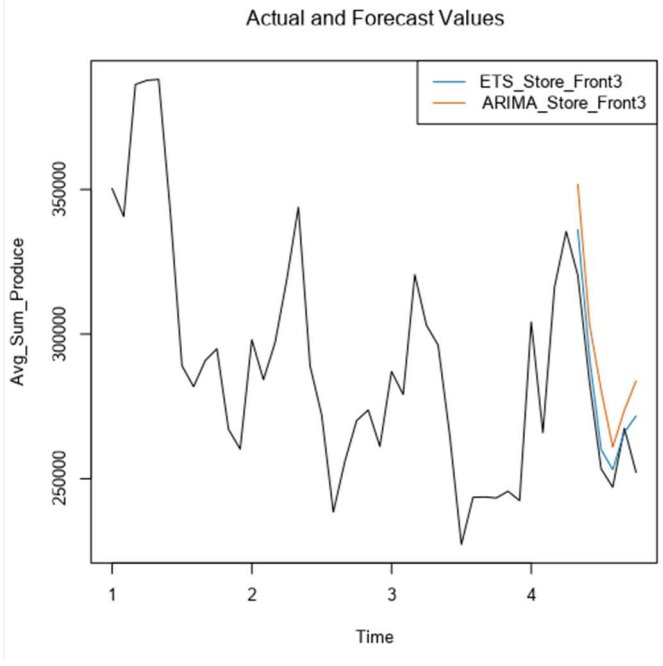
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Store_Front2	-9674.675	12835.96	12048.69	-3.5208	4.3797	0.6253
ARIMA_Store_Front2	-26674.856	27738.74	26674.86	-9.7121	9.7121	1.3843



Store Format 3:

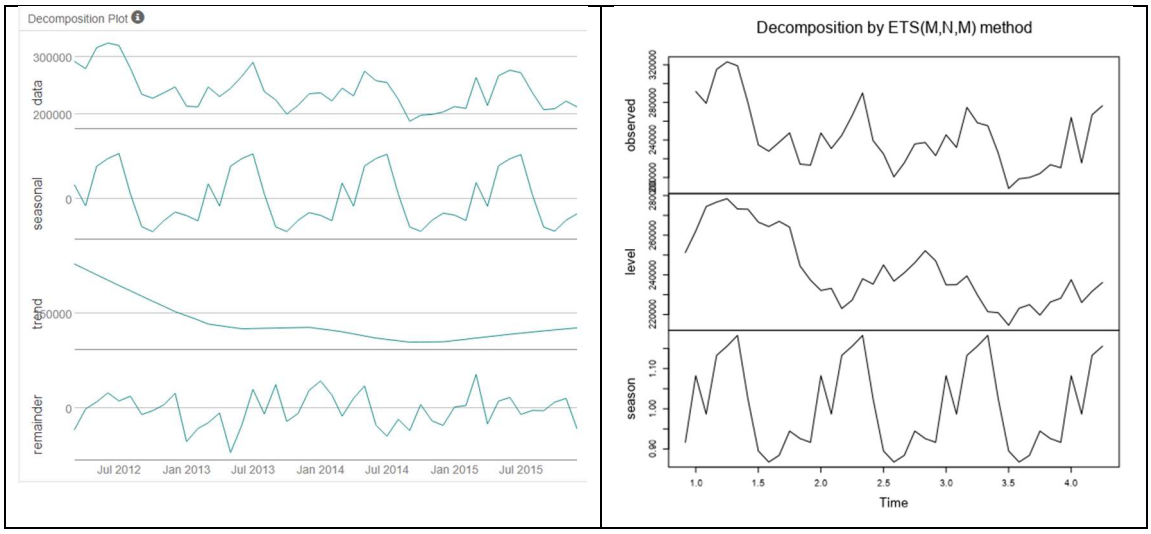
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Store_Front3	-9161.638	11401.14	9618.574	-3.3608	3.5317	0.4488
ARIMA_Store_Front3	-21710.399	23645.81	21710.399	-8.0077	8.0077	1.013

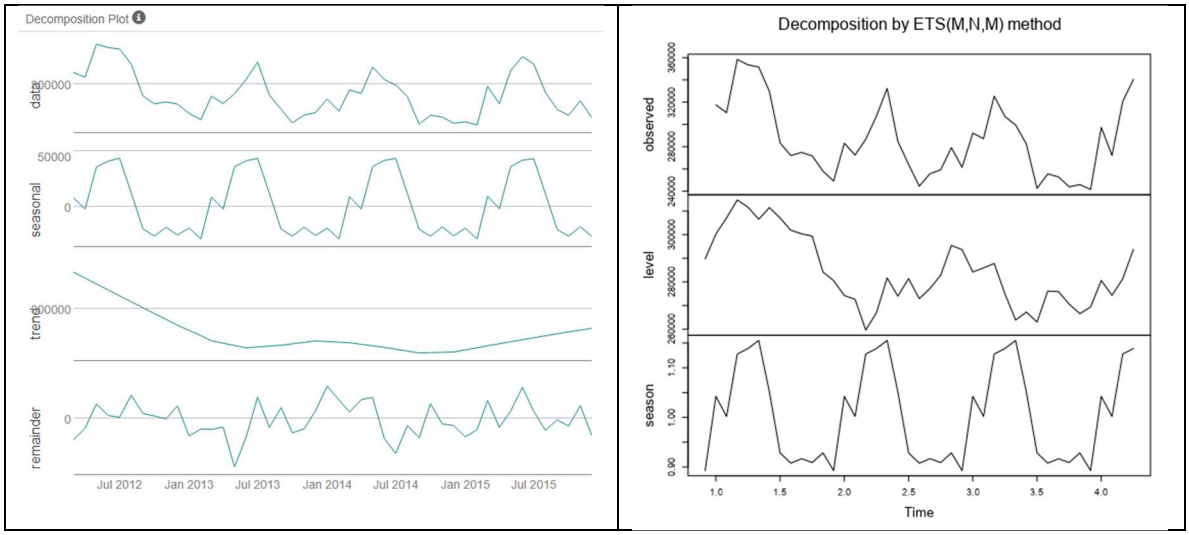


Decomposition plots of Store Format 1 cluster showed growing error, no trend and slight

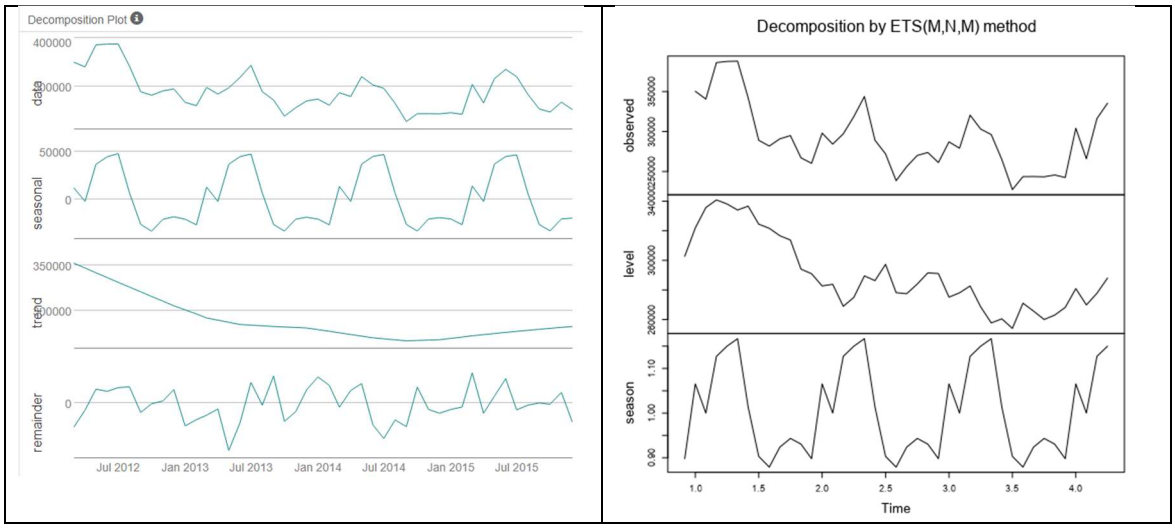
increasing seasonality for M, N and M notation.
Store Format 1:



Decomposition plots of Store Format 2 cluster also showed growing error, no trend and slight increasing seasonality for M, N and M notation.
Store Format 2:



Likewise decomposition plots of Store Format 3 cluster showed growing error, no trend and slight increasing seasonality for M, N and M notation.
Store Format 3:



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.

Forecast models for each Store Front predicts the average produce sales monthly in 2016. Multiplying each with the total number of existing stores in each Store Front cluster and summing monthly gives the monthly forecasts for the 85 existing stores.

Repeating the calculation for the new stores (1 in Store Front 1, 6 in Store Front 2, 3 in Store Front 3) results in monthly forecasts for the 10 new stores.

Table below shows the forecasts of produce sales from the 10 new stores and 85 existing stores for each month in 2016.

Month	New Stores	Existing Stores
Jan-16	$ 2,910,944.15	$ 24,078,058.16
Feb-16	$ 2,764,881.87	$ 22,670,735.53
Mar-16	$ 3,141,305.87	$ 25,858,187.53
Apr-16	$ 3,195,054.20	$ 26,288,436.90
May-16	$ 3,212,390.95	$ 26,501,400.91
Jun-16	$ 2,852,385.77	$ 23,303,548.46
Jul-16	$ 2,521,697.19	$ 20,583,812.16
Aug-16	$ 2,466,750.89	$ 20,160,031.58
Sep-16	$ 2,557,744.59	$ 20,888,455.26
Oct-16	$ 2,530,510.81	$ 20,891,395.24
Nov-16	$ 2,563,357.91	$ 21,057,160.62
Dec-16	$ 2,483,924.73	$ 20,415,891.84

Figure below illustrates the forecasts of produce sales from the new stores and existing

stores for each month in 2016. Given there are 10 new stores compared to 85 existing stores, the forecasted produce sales for new stores are much less than that of the 85 existing stores.