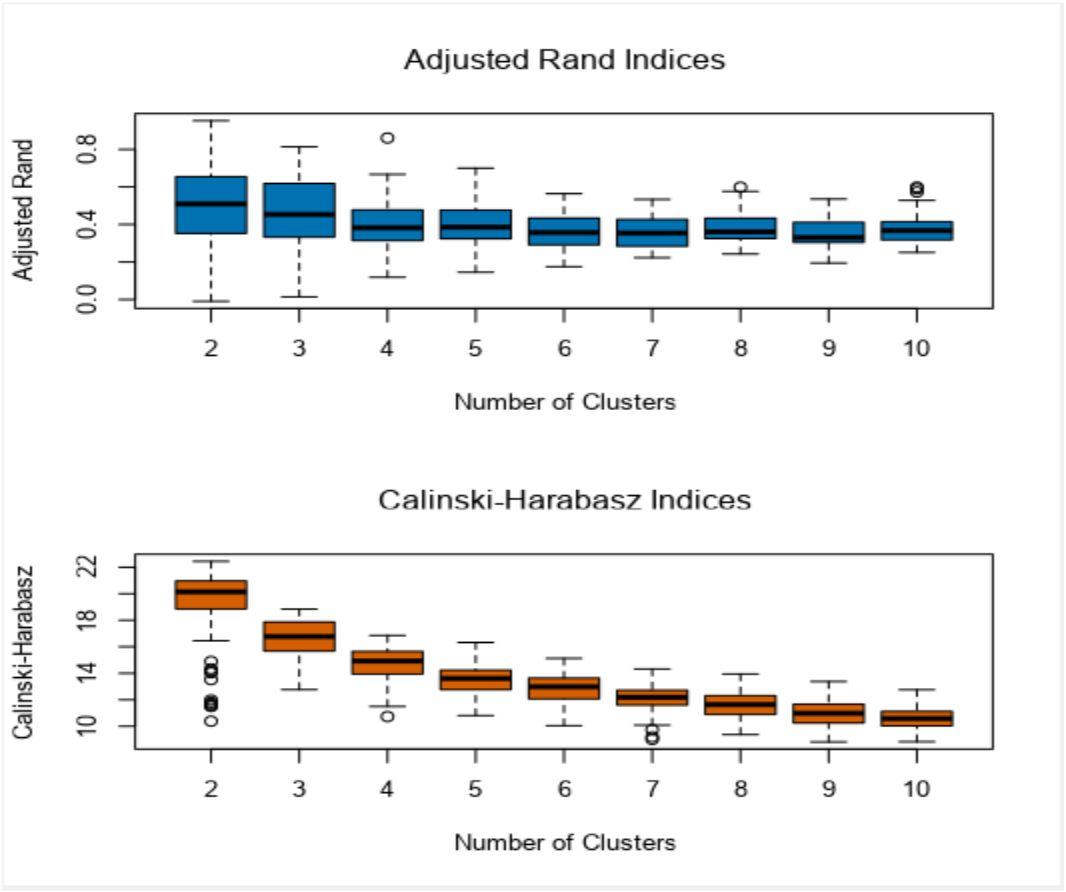
Project: Determining the Format for the New Stores and Forecasting Produce Sales

Summary: In this multi-stage project, I first analyzed the optimal number of store formats for a grocery store chain and then developed a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store. In the last step I prepared a monthly forecast for produce sales for the full year for both existing and new stores.

Task 1: Determining the Format

To determine the optimal number of store formats I evaluated the **K-Centroid Diagnostics** results. As can be seen from the plots below **Adjusted Rand Indices** and **Calinski-Harabasz Indices** indicate that 3 clusters will provide the best segmentation. Several factors are considered in these plots to determine the optimal number of clusters. The X axes in the plots indicate the number of clusters, higher median scores as well as compactness and less dispersion indicate better fit. Taken altogether the plots indicate number 3 as the optimal number of clusters.

Plots



After determining the number of clusters, based on my analysis I decided how many stores fall into each store format. Each cluster has stores as follows.

Cluster 1 = 25 Stores
Cluster 2 = 35 Stores
Cluster 3 = 25 Stores

ord Report

Summary Report of the K-Means Clustering Solution Cluster_Capstone

Solution Summary

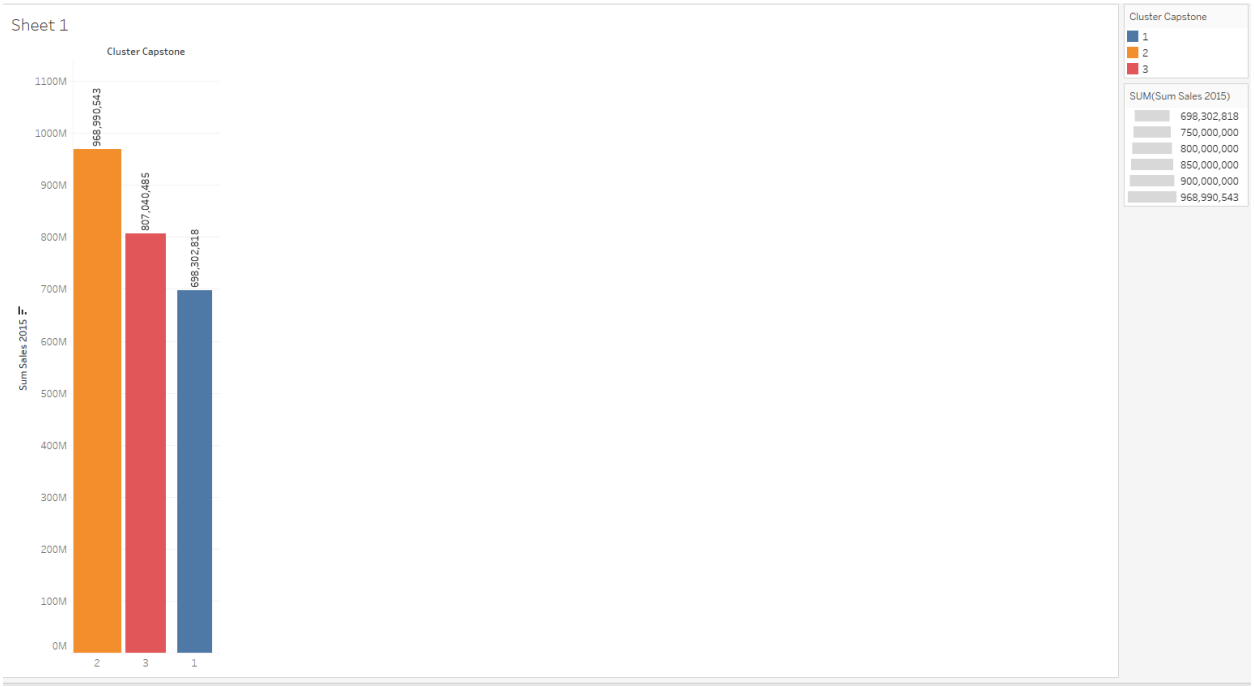
Call:

stepFlexclust(scale(model.matrix(~-1 + Dry_Grocery. + Dairy. + Frozen_Food. + Meat. + Produce. + Floral. + Deli. + Bakery. + General_Merchandise., the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

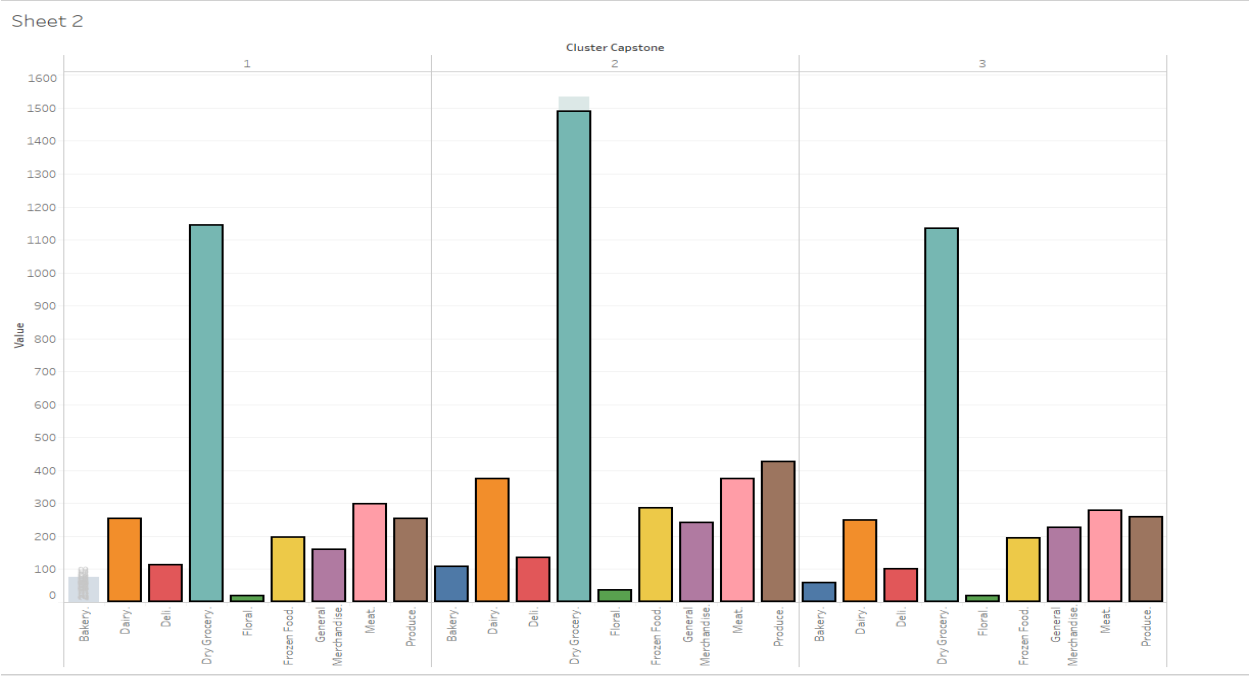
Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

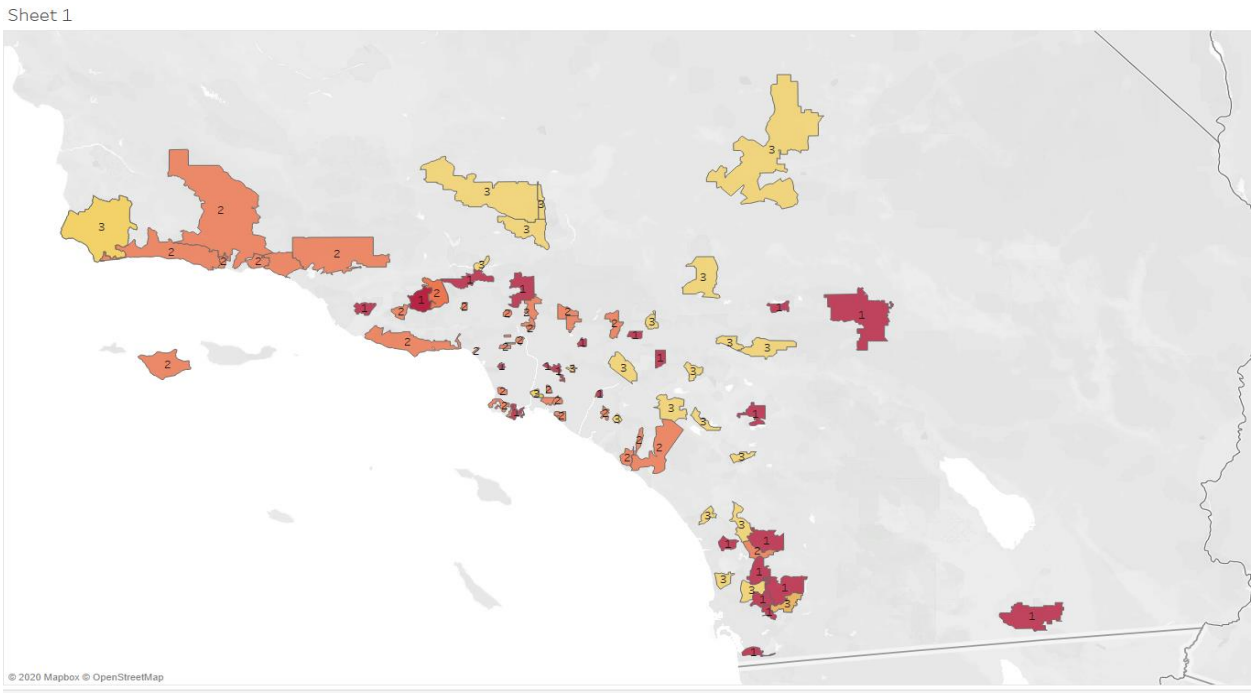
We can safely say that sales of the stores are considered in clustering process. Cluster 2 has the highest sales, then comes cluster 3 and then cluster 1. That makes sense as we used the sales numbers as our clustering values.



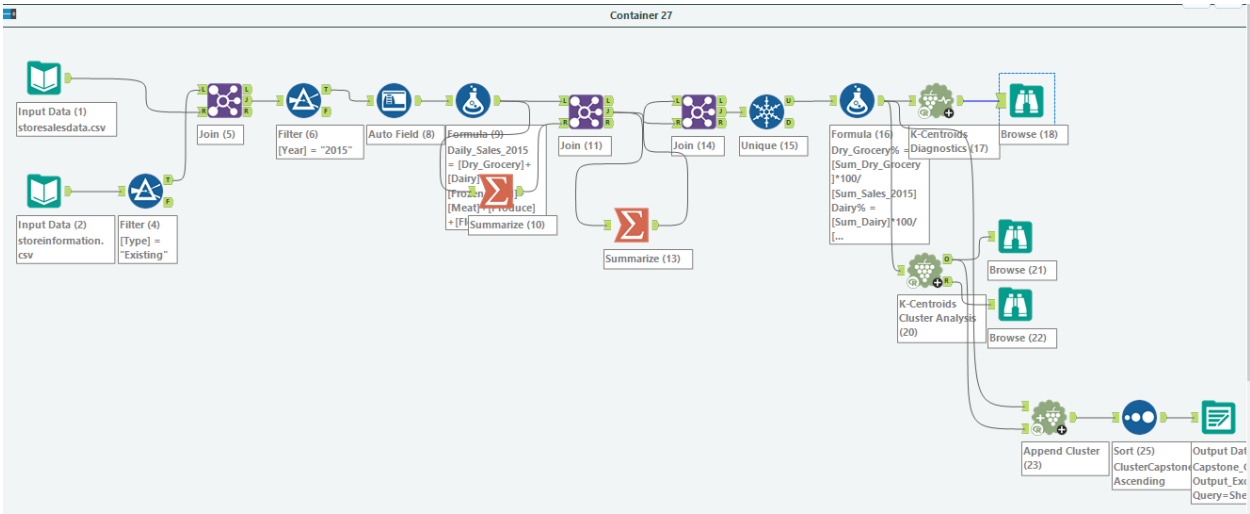
The tableau analysis indicated that; in terms of the sales by the **types of the products** there is not significant difference among the clusters. For example, as shown below, **dry grocery** is the highest selling item in all three clusters. Flora is the lowing selling product in all three clusters.



The Tableau visualization below shows the location of the stores, uses color to show cluster, and size to show total sales.



Below is the Alteryx workflow for Task 1



Kadir AKYUZ, Ph.D.