Predictive Analytics Capstone

Combining Predictive Techniques

Project Development

Project Rubric, https://review.udacity.com/#!/rubrics/437/view
Nanodegree, https://www.udacity.com/course/predictive-analytics-for-business-nanodegree--nd008

Task 1: Determine Store Formats for Existing Stores

I developed the following activities to solve task 1.

1) understood the objective of segmentation;
2) prepared data;
3) determined the number of clusters;
4) created the cluster model;
5) visualized and validated results.

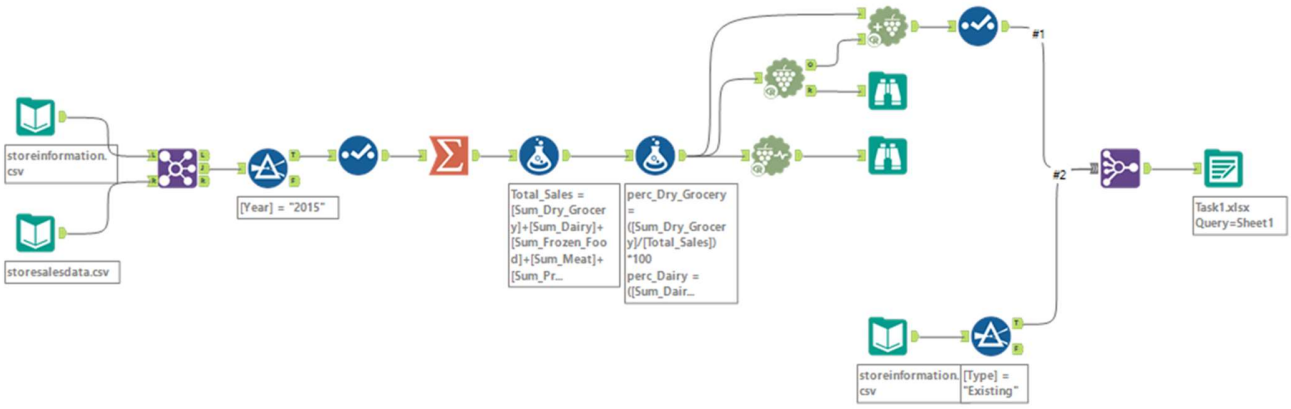

Image 1, Determining Store Formats for Existing Stores using Alteryx

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

   I concluded it would be 2 according to the results of the Alteryx workflow I developed. Although by https://knowledge.udacity.com/questions/450900 explanation and other mentor orientation that I read and understood by the time this project was conceived there was an older version of Alteryx running performing then higher median values related to cluster 3 in AR and C-H plots.

The optimal number of store formats is 3.

Understanding segmentation objective and preparing data.

The data provided had all the information needed to accomplish the analysis. First I understood the objective for the segmentation I was trying to create. Joined two data files, filtered 2015 information, changed Sales fields from VString to Double. After that I summarized data grouping it by *Store Id* and *Year* and summing by each category. Using *Formula Tool* I created a *Total_Sales* field which sum all categories. Next step was the creation of nine different columns each one receiving the sum of a category divided by *Total_Sales* multiplied by 100, in other words. By the end of this specific cycle I obtained 85 records.

Determining the number of clusters.

Proceeded using *K-Centroids Diagnostics Tool*

Customized parameters …

- Standardize the fields = enabled (z-score)
- Clustering method = K-Means
- Minimum number of clusters = 2
- Maximum number of clusters = 10
- Bootstrap replicates = 50
- Number of starting seeds = 10

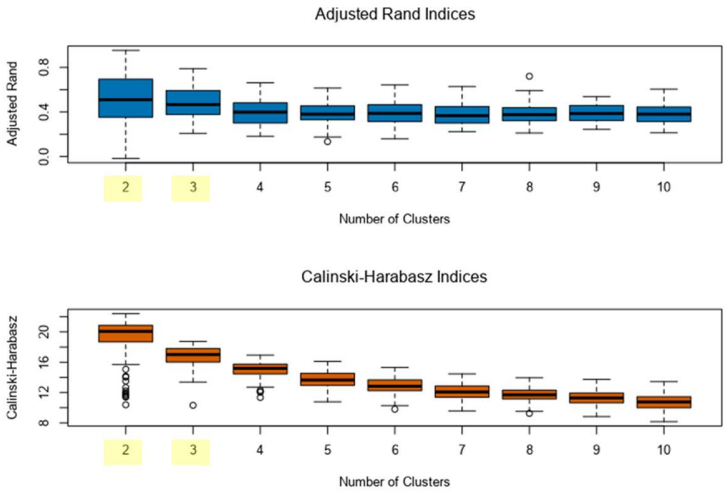

Image 2, AR and C-H Index plots

For both of the following plots AR and C-H indexes Images 1& 2) the y-axis is the value of the index and the x-axis is the number of clusters. For the AR index, the higher the index, the better the stability of the cluster. For the C-H index, the higher the index, the better the distinctness and compactness of the clusters.

	2	3	4	5	6	7	8
Minimum	-0.017586	0.208197	0.181585	0.133772	0.158757	0.222502	0.21093
1st Quartile	0.352613	0.377392	0.302314	0.331809	0.314419	0.299658	0.322749
Median	0.509257	0.466169	0.398104	0.380556	0.387434	0.366279	0.375409
Mean	0.494056	0.479493	0.404888	0.388834	0.39306	0.381404	0.384298
3rd Quartile	0.693746	0.58771	0.481097	0.454895	0.46369	0.447859	0.436717
Maximum	0.952939	0.788895	0.661744	0.614672	0.64242	0.62851	0.720498

	9	10
Minimum	0.244439	0.212783
1st Quartile	0.325103	0.315087
Median	0.386151	0.380127
Mean	0.390303	0.379638
3rd Quartile	0.457811	0.442954
Maximum	0.538277	0.604545

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.38298	10.31461	11.34984	10.77356	9.80353	9.577281	9.253901
1st Quartile	18.69647	16.03968	14.46704	12.9405	12.24542	11.378557	11.166056
Median	20.07012	17.00754	15.19152	13.65142	12.83476	12.07357	11.697797
Mean	19.08577	16.73685	14.98778	13.68998	12.83426	12.156743	11.681178
3rd Quartile	20.87407	17.78773	15.74729	14.53404	13.67175	12.859807	12.311206
Maximum	22.41555	18.73715	16.93911	16.10526	15.30862	14.460893	13.955665

	9	10
Minimum	8.822973	8.153824
1st Quartile	10.648806	10.002731
Median	11.287124	10.760594
Mean	11.359959	10.745482
3rd Quartile	11.937564	11.429852
Maximum	13.731897	13.433832

Image 3, K-Means Cluster Assessment Report

2. How many stores fall into each store format?
   25 stores in cluster 1, 35 stores in cluster 2 and 25 stores in cluster 3.

   None of the clusters I determined have less than 20 stores or more than 40 stores.

Creating cluster model.

Proceeded using *K-Centroids Cluster Analysis Tool*

Customized parameters …

- Standardize the fields = enabled (z-score)
- Clustering method = K-Means
- Number of clusters = 3
- Number of starting seeds = 10

Report

Summary Report of the K-Means Clustering Solution Cluster_Model

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + perc_Dry_Grocery + perc_Dairy + perc_Frozen_Food + perc_Meat + perc_Produce + perc_Floral + perc_Deli + perc_Bakery + perc_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

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	perc_Dry_Grocery	perc_Dairy	perc_Frozen_Food	perc_Meat	perc_Produce	perc_Floral	perc_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	perc_Bakery	perc_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Plots

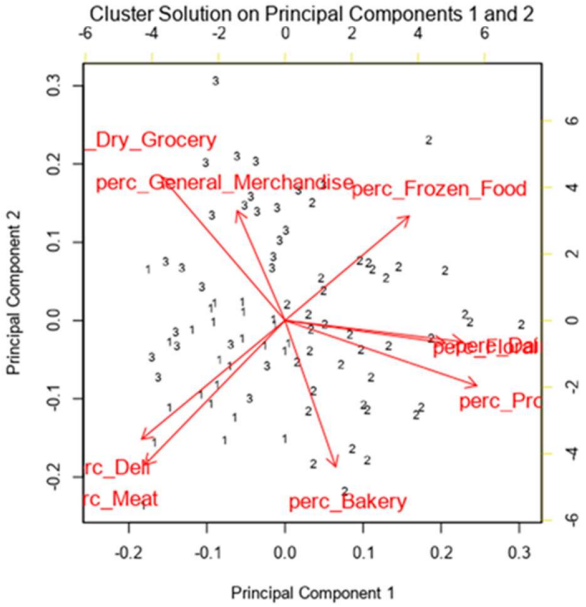

Image 4, K-Centroids Cluster Analysis reporting

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

   The main features that distinguish clusters are *average distance*, *maximum distance* and *separation*.

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

Image 5, Cluster differences | Avg Distange, Max Distance, Separation |

- Cluster 1 is the most compact. And cluster 2 might present more variability.
- Cluster 1 has the highest maximum distance from centroid.
- Cluster 1 is more separated from the other clusters.

I used the chart with the individual variables checking extreme values for each variable. In green the highest and in yellow the lowest values. A high positive versus a high negative might indicate that those two clusters are opposite to each other.

	perc_Dry_Grocery	perc_Dairy	perc_Frozen_Food	perc_Meat	perc_Produce	perc_Floral	perc_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	perc_Bakery	perc_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Image 6, Cluster differences | individual variables |

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.

   https://public.tableau.com/profile/rosana7921#!/vizhome/Task1_16120363339330/Sheet1?publish=yes

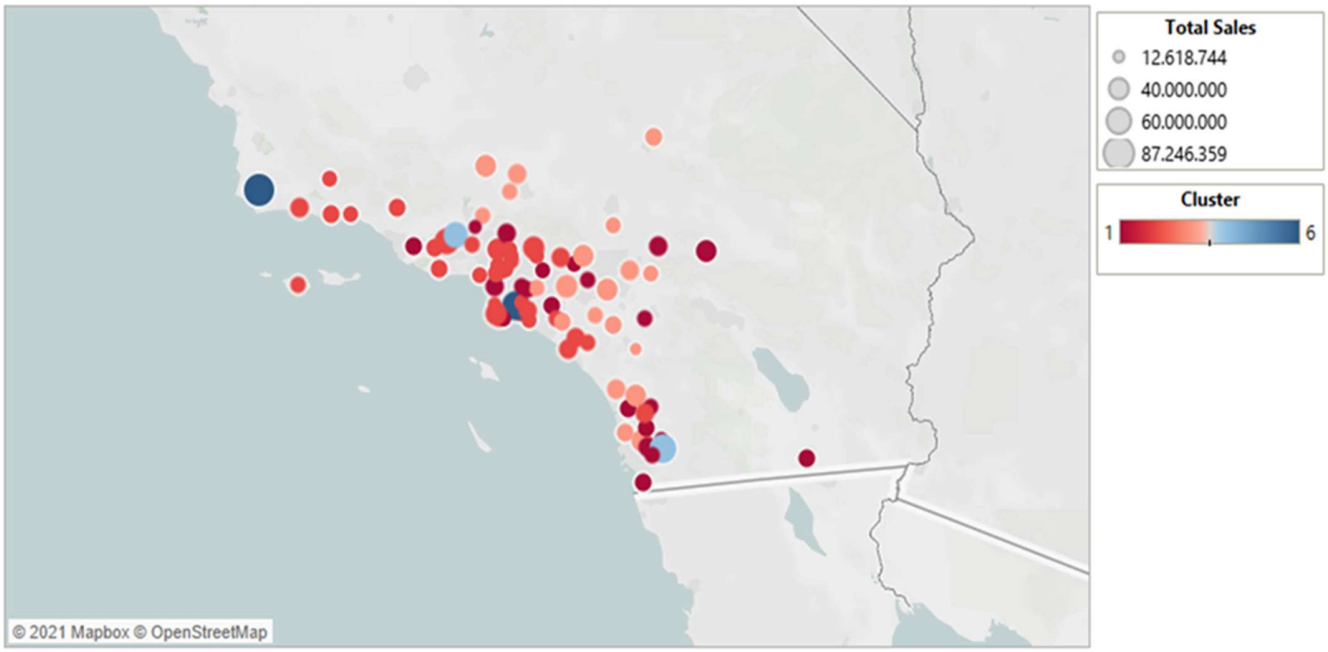

Image 7, Stores location by Total _Sales and coloured by Cluster using Tableau