Segmenting The Breakfast Cereal Market

Vincent Pun

Abstract. There is no shortage of product variety when it comes to the breakfast cereals market. Understanding characteristics such as nutritional facts are important for both consumers and manufacturers that are making either purchasing or product development decisions. In this paper, unsupervised machine learning methods such as clustering and multidimensional scaling are used to visualize relationships between 77 commonly available breakfast cereals that are described by 13 variables that are sourced from F&DA food labels in 1993.

Keywords. *Breakfast cereals, segmentation, factor analysis, principal component analysis, k-means algorithm, t-distributed stochastic neighbor embedding (t-SNE), fuzzy clustering, biclustering, multidimensional scaling (MDS)*

Introduction. While there are many important chemical elements and compounds found on food labels, it is inefficient for users to compare each of these values when differentiating between cereal brands. This study is prepared to equip both consumers and manufacturers with visualization tools and segmentation techniques to quickly compare brands to alternative products.

By finding important correlations and understanding the dimensionality of the data, we are able to evaluate traits within the homogenous subgroups that constitute the breakfast cereal market. These findings may also help manufacturers identify imbalances in their product portfolios in addition to gaining competitor insights.

Literature review. Methods such as segmentation and conjoint analysis have been used to learn about consumer expectations, needs, and preferences for both breakfast cereals and nutrition labeling. In a 284-respondent conjoint analysis, researchers found that participants valued high fiber levels and low sugar levels when it came to choosing healthy cereals. Based on market simulations, it was concluded that “increasing the sugar content from 6g to 12g

decreases the market share significantly while an increase from 5% to 15% in the fiber content has the greatest impact on market share (Boivin, Parissier, Alle, Forcier & Langlois, 2014). By applying summary statistics to product clusters in this study, we observe how these findings relate to the specifications found in manufacturers' product offerings.

Fuzzy clustering approach was used in a recent study to identify consumer subgroups based on both understanding and use for nutrition labels. Introduced by Zadeh in 1965, fuzzy clustering algorithms are extremely useful when assigning observations to distinct clusters is challenging (Souiden, Abdelaziz & Faucconier, 2013); observations may belong to more than one subgroup, so it may be considered as a more robust way of segmentation (Xu and Wunsch, 2009).

Methods. Observations with missing values are first removed from the original dataset; after three observations are omitted, Homestead's only product in the source data, Maypo, becomes the only hot cereal remaining. Thus, the "Cold.or.Hot" attribute will not be used for this study given the sparsity of hot cereal examples. Furthermore, the Rating column is removed, as the calculation for this metric is unknown.

Before we are able to apply unsupervised learning methods for both visualization and dimensionality reduction purposes, it is necessary to perform exploratory analysis and normalize the data. In **Figure 1**, a bar plot is used to identify that Kelloggs and General Mills are considered key players in this research, as they offer more than double the number of products compared to their competitors. Also, bivariate distribution charts (**Figure 2**) are helpful in identifying observations that are causing high skewness in variables such as Fiber and Vitamins, but these observations are kept in case they are representing a niche position in the market. Both scatter plots (**Figure 3**) and correlation plots (**Figure 4 and Figure 5**) are able to clearly identify strong positive relationships between variables such as Protein, Fiber, and

Potassium, which will be analyzed subsequently using tools such as principal component analysis and factor analysis.

After normalizing the remaining variables on a common scale from 0 to 1 to prevent variables from having too much influence, our objective is to find variables that are most valuable toward explaining variance within the data. Clustering model goodness of fit results will then be compared between models that used all variables and a “reduced” model that contains fewer variables.

A scree plot (**Figure 6**) shows that four components may be sufficient for describing the data; based on a principal components analysis, it is confirmed that 82% of the variation resides in these components. Similarities are noted between correlation plots and PCA component contributions (**Figure 7),** as we see that variables such as Sugars, Potassium, and Fiber are influential in describing the first two principal components. Also, a PCA biplot (**Figure 8**) can be used to visualize how certain products such as AllBran and 100%Bran strongly influence the second component. Finally, a path diagram (**Figure 9**) derived from a factor analysis with four factors relies heavily on the same variables. Sodium and Vitamins are not valuable in describing the first four latent variables, so these dimensions are removed.

Visualizing results from a variety of clustering algorithms such as K-means, tsne, biclustering, and fuzzy clustering provides consumers and manufacturers with thorough interpretations of product segments. Averages for the K-means and fuzzy clustering results are produced, and manufacturer product allocations across segments are generated too.

As product offerings change over time, manufacturers may want to not only pinpoint product similarities but also analyze the degrees of saturation within the cereal market. Thus, MDS is used to visualize the euclidean distance between observations on both a two-dimensional and

three-dimensional representation in this study, which helps manufacturers see when scenarios such as product cannibalization have occurred.

Results. A K-means model with four clusters is the first clustering algorithm used in this study. By creating models on both the full (9 variables) and reduced (7 variables) data subsets, we use the R-squared (R2) and average silhouette width to evaluate results. The reduced model is ultimately selected for all clustering solutions, as its R2 value (0.59) was considerably higher than that of the full model (0.46); however, it should be noted that the ASW for the reduced model experienced a decrease from 0.43 to 0.35.

Plotting the K-means results (**Figure 10**) confirm that using four clusters may be optimal for the purposes of this research given clear separation between each homogeneous subgroup. Moreover, using t-SNE to plot groups identified by the K-means cluster analysis also verifies these segment distinctions. Containing only 3 out of the 73 products, Group 1 contains the three “outlier observations" that were previously identified when exploring the data. A manufacturer portfolio allocation shows that Kelloggs and Nabisco occupy this niche market (**Table 1**). We believe that this segment targets consumers with specific dietary needs, as these cereals have the highest fiber, protein, and potassium contents, and they have low carbohydrate and sugar levels too (**Table 2**). Group 2 appears to be very nutrient-dense (most protein and second highest potassium levels), which makes these products great family-friendly options. Group 3’s key trait is that it contains cereals with high sugar content at 11.48g per serving. Group 4 contains cereals that are well-balanced (low-sugar, low-fat, and low-calorie), making them suitable for consumers that prefer healthier options.

Alternative clustering and visualization tools are also used to segment the cereal market. For example, a checkerboard plot (3 rows by 3 columns) from biclustering (**Figure 14**) shows that clusters are formed based on two very influential groups of variables, which is similar to the

results found when performing PCA. However, there are no clear patterns found when clustering row segments, which suggests that manufacturers are not heavily concentrated in any particular position in the market.

Fuzzy clustering (**Figure 16**) is also used to fit the data, where observations are configured to belong to at least two clusters. With this soft clustering technique, it became difficult to interpret modeling solutions that used more than two centroids. Thus, the model is fitted with two clusters, and we use a 0.70 cutoff determination to create a third subgroup to contain observations that don't distinctly belong in either of the original clusters; sugar content is the most influential discriminant that separates these three groups (**Table 5**).

When reviewing the fuzzy cluster membership designations **(Figure 15**), we noticed that certain observations had similar or identical values. Visualizing product similarities is performed through the use of metric MDS. While a two-dimensional visualization (**Figure 12**) was challenging to understand due to the presence of high-density clusters, an interactive three-dimensional solution (**Figure 13**) that is colored based on manufacturer brings clarification to these modeling results; now we are able to identify that General Mills has two cereals (Cocoa Puffs and Count Chocula) that are nearly identical, which may be causing product cannibalization.

Conclusions. Throughout this study, we have learned that unsupervised learning methods are helpful for conducting dimensionality reduction, clustering products, and measuring product differentiation. While alternative segmentation models such as fuzzy clustering may offer deeper insight concerning brand positioning, the K-means model already provides clear results that can be used to divide the breakfast cereal market. Finally, we have witnessed that three-dimensional modeling may elucidate model results when observations on a two-dimensional plot appear too compact for interpretation.

Appendix

Table 1 - K-means Clustering - Proportions

	A	B	C	D	E	F	G	H	I	J
1	**Manufacturer**	**1**	**2**	**3**	**4**	**Sum**	**prop.g1**	**prop.g2**	**prop.g3**	**prop.g4**
2	A_HomestatFarm	0	0	0	0	0	NA	NA	NA	NA
3	General Mills	0	5	10	7	22	0	0.23	0.45	0.32
4	Kelloggs	2	6	6	9	23	0.09	0.26	0.26	0.39
5	Nabisco	1	0	0	4	5	0.2	0	0	0.8
6	Post	0	3	3	3	9	0	0.33	0.33	0.33
7	Quaker_Oats	0	2	2	3	7	0	0.29	0.29	0.43
8	Ralston	0	2	0	5	7	0	0.29	0	0.71
9	Sum	3	18	21	31	73	0.04	0.25	0.29	0.42

Table 2 - K-means Clustering - Summary Statistics (**Mean**)

cluster	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins
1	63.33	4	0.67	176.67	11.00	6.67	3.67	310.00	25
2	126.67	3.17	2.06	156.94	3.22	14.11	9.50	155.56	31.94
3	110.95	1.52	1.00	168.57	0.57	12.43	11.48	47.38	25.00
4	97.42	2.61	0.42	165.16	1.87	17.39	3.23	79.68	30.65

Table 3 - Fuzzy Clustering - Membership (Allocation to each cluster)

1	Brand Name	1	2
11	CapNCrun	0.80	0.20
12	Cheerios	0.36	0.64
13	CinTCrun	0.71	0.29
14	Clusters	0.65	0.35
15	CocoPuff	0.81	0.19
16	CornChex	0.26	0.74
17	CornFlak	0.21	0.79
18	CornPops	0.71	0.29
19	ContChoc	0.81	0.19
20	COatBran	0.62	0.38

(please see “3_fuzzymembership.csv” for complete information)

Table 4 - Fuzzy Clustering - Proportions

Manufacturer	1	2	3	Sum	prop.g1	prop.g2	prop.g3
A_HomestatFarm	0	0	0	0	NA	NA	NA
General Mills	10	6	6	22	0.45	0.27	0.27
Kelloggs	5	6	12	23	0.22	0.26	0.52
Nabisco	0	4	1	5	0	0.8	0.2
Post	2	2	5	9	0.22	0.22	0.56
Quaker_Oats	2	0	5	7	0.29	0	0.71
Ralston	0	3	4	7	0	0.43	0.57
Sum	19	21	33	73	0.26	0.29	0.45

Table 5 - Fuzzy Clustering - Summary Statistics (Mean)

cluster	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins
1	111.58	1.74	1.21	157.37	0.74	11.89	11.58	57.37	25.00
2	100.00	2.43	0.38	173.57	1.86	18.43	3.05	77.62	32.14
3	109.09	2.97	1.27	163.03	3.27	13.97	7.24	135.61	29.55

Figure 1 - Univariate Analysis - Manufacturer Name

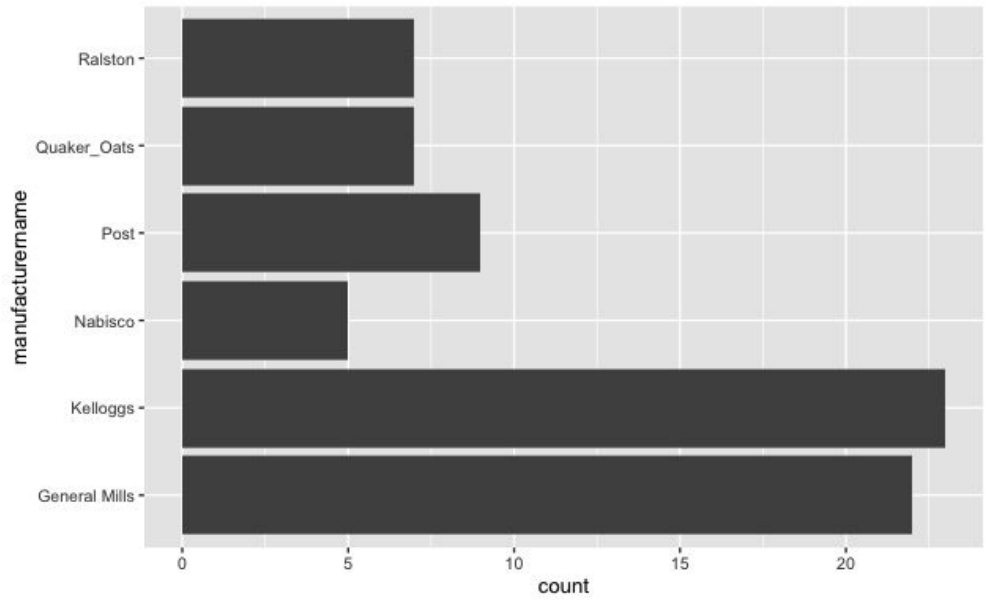


Figure 2 - Fiber Distribution

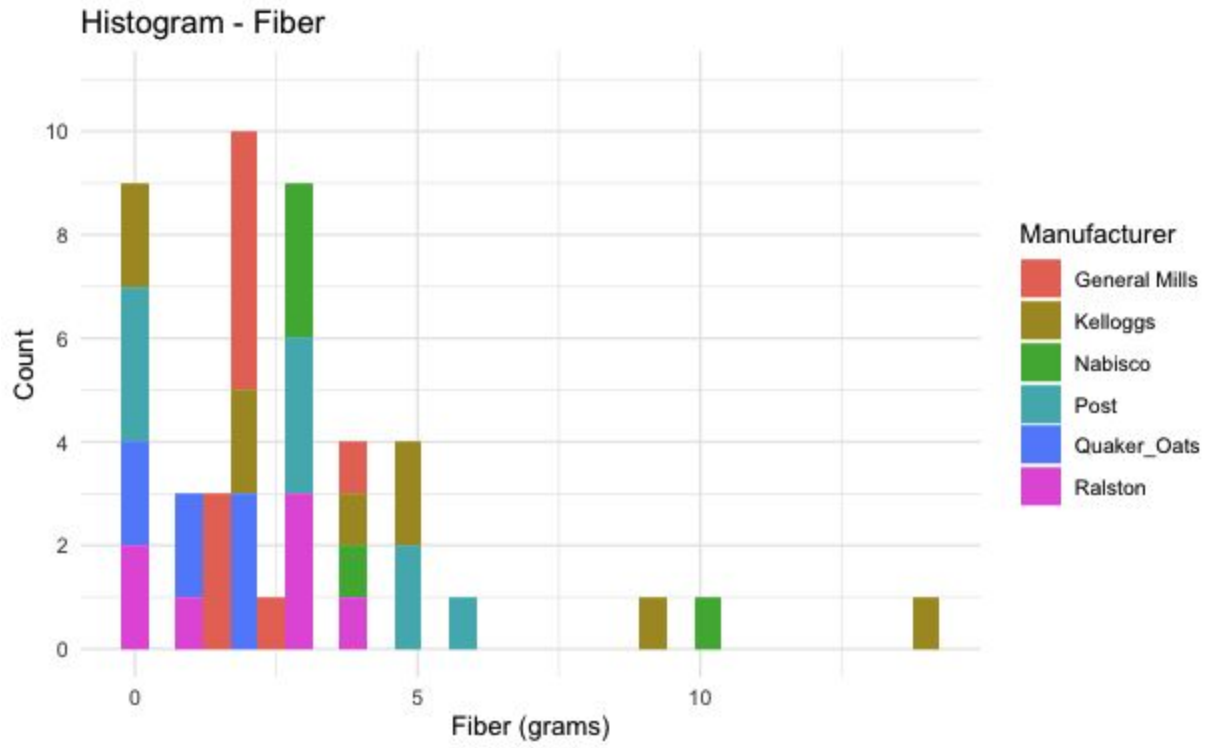


Figure 3 - Bivariate Analysis - Potassium and Fiber

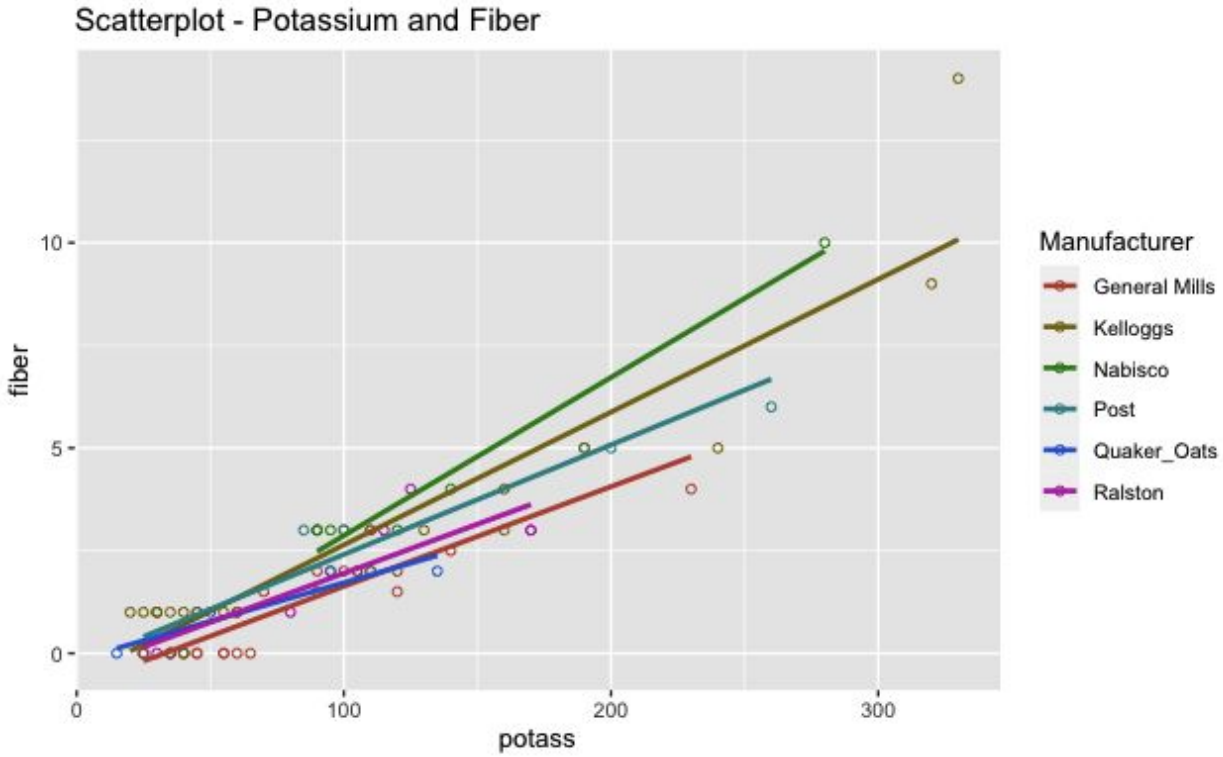


Figure 4 - Correlation Plot

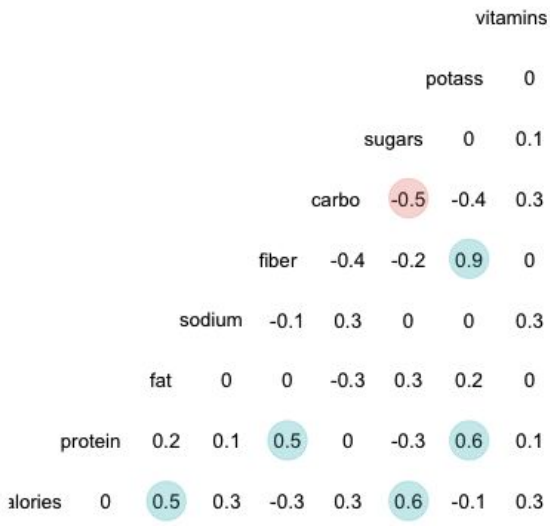


Figure 5 - Correlation Plot (hierarchical clustering order)

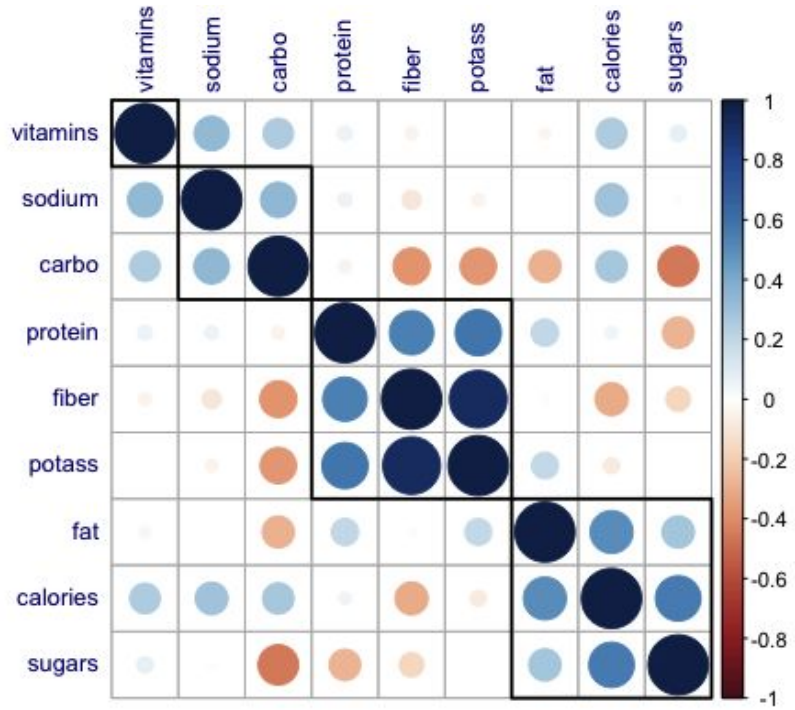
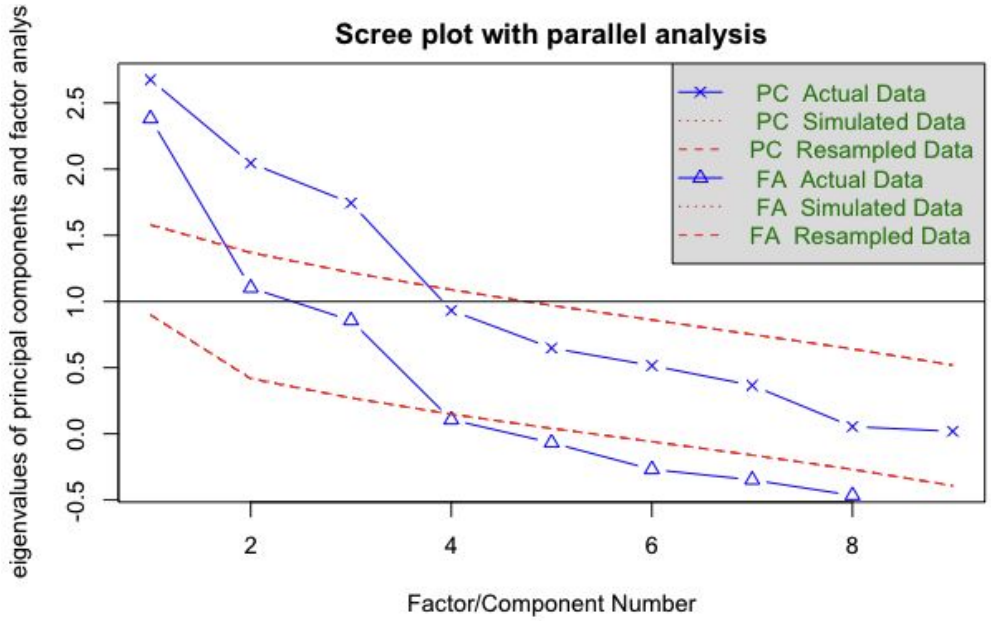


Figure 6 - Scree plot with parallel analysis



PCA Results:

```
In princomp.default(minmaxdf, scale. = TRUE) :
 extra argument 'scale.' will be disregardedImportance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     0.3413399 0.3372470 0.2985635 0.20705988 0.19051167 0.15376198
Proportion of Variance 0.2639304 0.2576390 0.2019243 0.09711968 0.08221641 0.05355663
Cumulative Proportion  0.2639304 0.5215694 0.7234937 0.82061342 0.90282983 0.95638647
                           Comp.7      Comp.8      Comp.9
Standard deviation     0.12804143 0.044834856 0.029130080
Proportion of Variance 0.03713782 0.004553516 0.001922201
Cumulative Proportion  0.99352428 0.998077799 1.000000000

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
calories  0.360         0.283  0.363  0.135  0.348  0.109  0.339  0.624
protein  -0.270 -0.290  0.385  0.288  0.120  0.162 -0.744        -0.130
fat       0.232 -0.232  0.171  0.630        -0.547  0.280        -0.290
sodium           0.306  0.607 -0.172 -0.678 -0.192
fiber    -0.226 -0.378  0.173 -0.215         0.107  0.269  0.747 -0.292
carbo    -0.162  0.500  0.211  0.249  0.155  0.503  0.295 -0.132 -0.485
sugars    0.785 -0.241        -0.226         0.306 -0.147        -0.386
potass   -0.180 -0.530  0.322 -0.147         0.222  0.415 -0.546  0.198
vitamins         0.180  0.443 -0.421  0.691 -0.329

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

Figure 7 - PCA Component Contribution Visualization

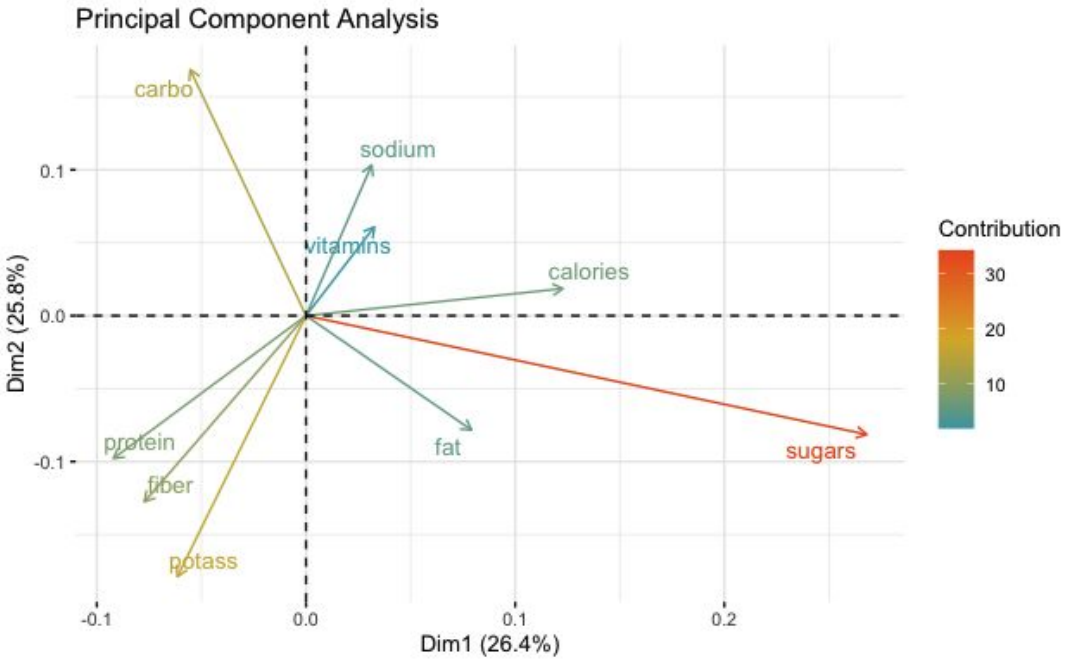
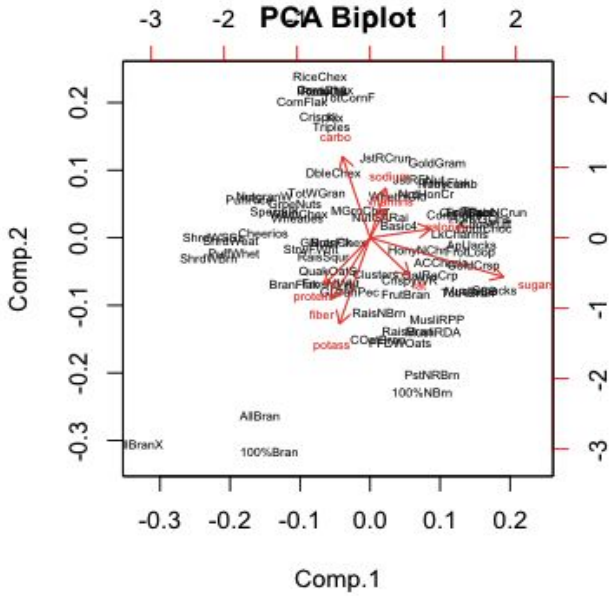


Figure 8 - PCA Biplot



Factor Analysis Results:

```
Factor analysis with Call: fa(r = minmaxdf, nfactors = 4, rotate = "none", fm = "pa")

Test of the hypothesis that 4 factors are sufficient.
The degrees of freedom for the model is 6  and the objective function was  0.85
The number of observations was  73  with Chi Square =  55.46  with prob <  0.00000000037

The root mean square of the residuals (RMSA) is  0.02
The df corrected root mean square of the residuals is  0.06

Tucker Lewis Index of factoring reliability =  0.275
RMSEA index =  0.336  and the 10 % confidence intervals are  0.26 0.423
BIC =  29.72
Loadings:
          PA1    PA2    PA3    PA4
calories -0.374  0.767  0.446
protein   0.547         0.428 -0.144
fat       0.126  0.665        -0.534
sodium   -0.198         0.433  0.189
fiber     0.940         0.148  0.183
carbo    -0.531 -0.351  0.675
sugars   -0.167  0.847 -0.365  0.384
potass    0.903  0.207  0.239  0.136
vitamins -0.141  0.107  0.366  0.249

                PA1   PA2   PA3   PA4
SS loadings    2.523 1.936 1.380 0.606
Proportion Var 0.280 0.215 0.153 0.067
Cumulative Var 0.280 0.495 0.649 0.716
```

Figure 9 - Path Diagram - Latent variable analysis

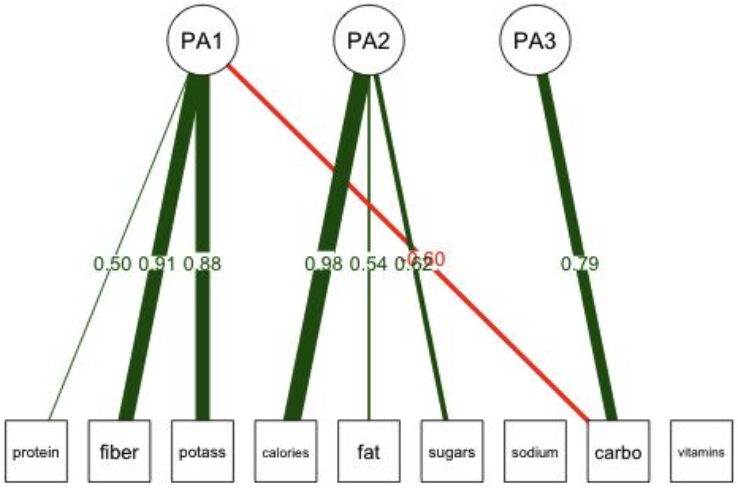


Figure 10 - K-means Plot

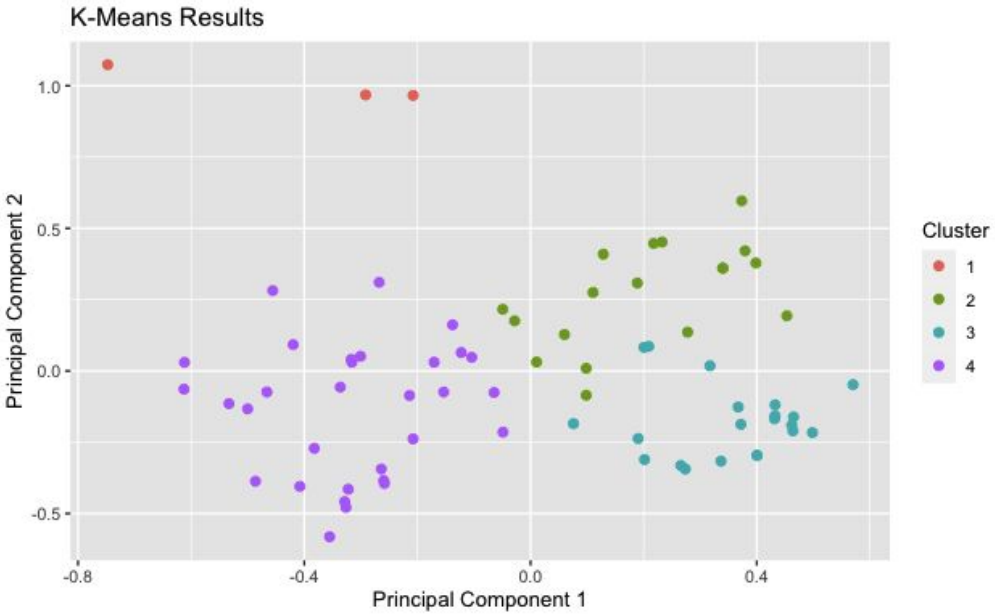


Figure 11 - t-SNE Plot

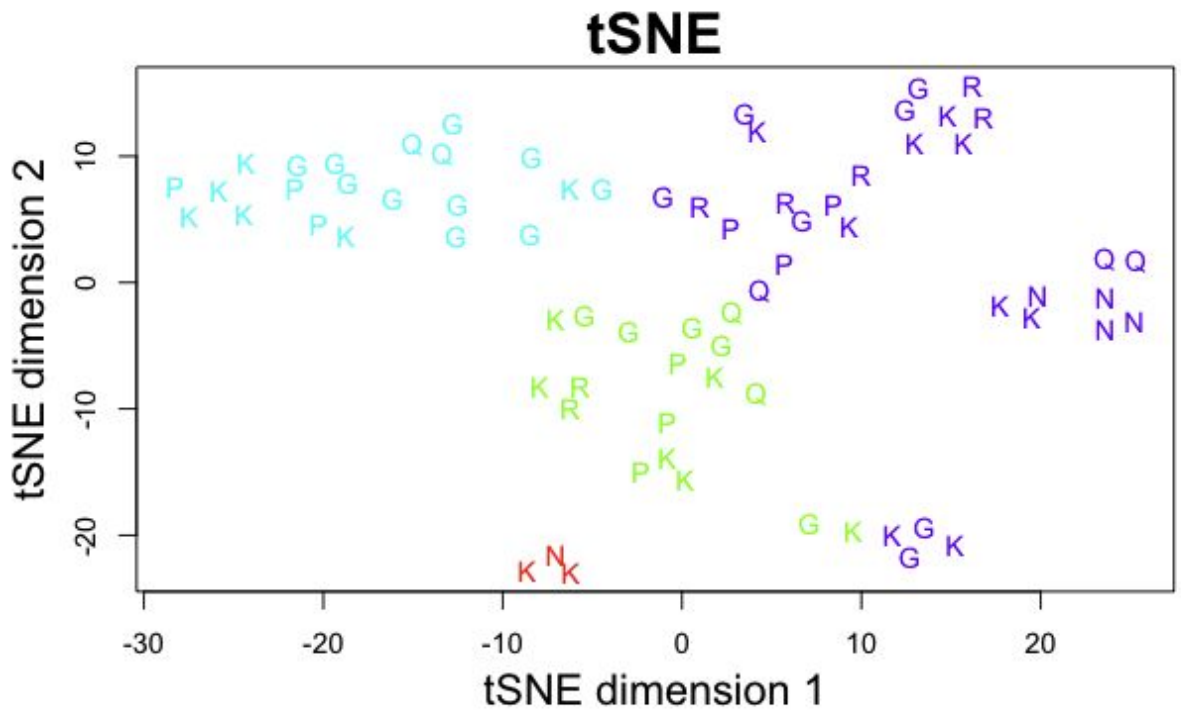


Figure 12 - Metric Multidimensional Scaling

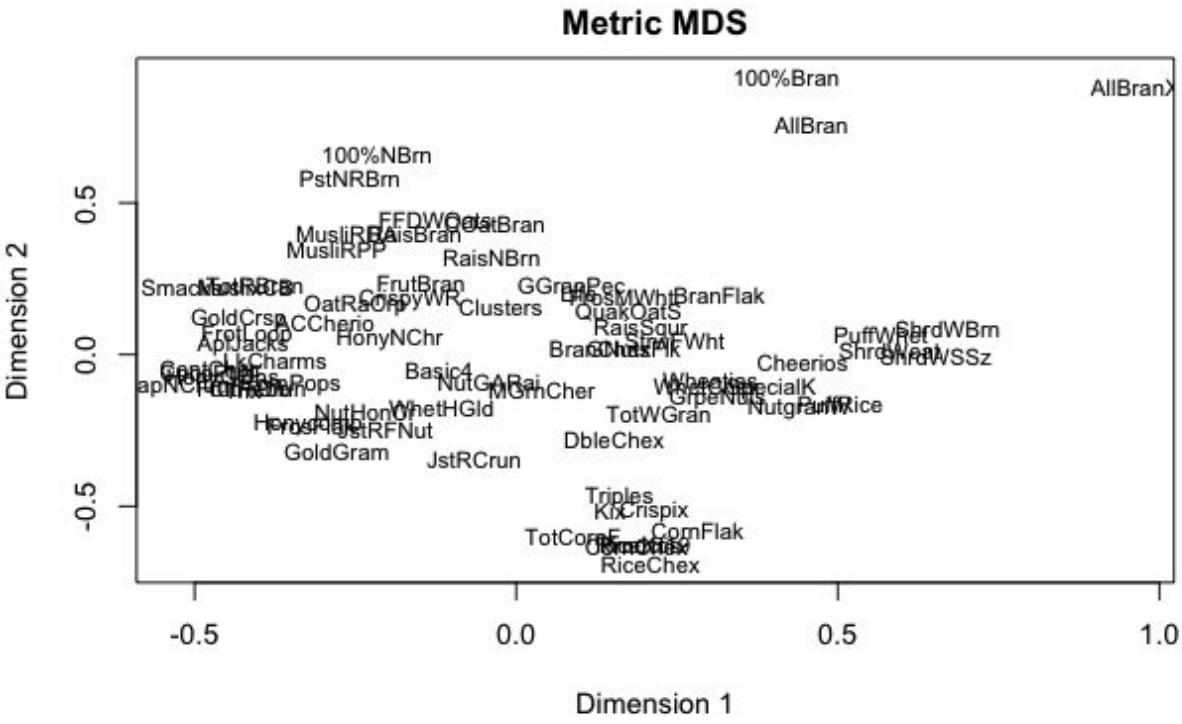


Figure 13 - Metric Multidimensional Scaling (3D)

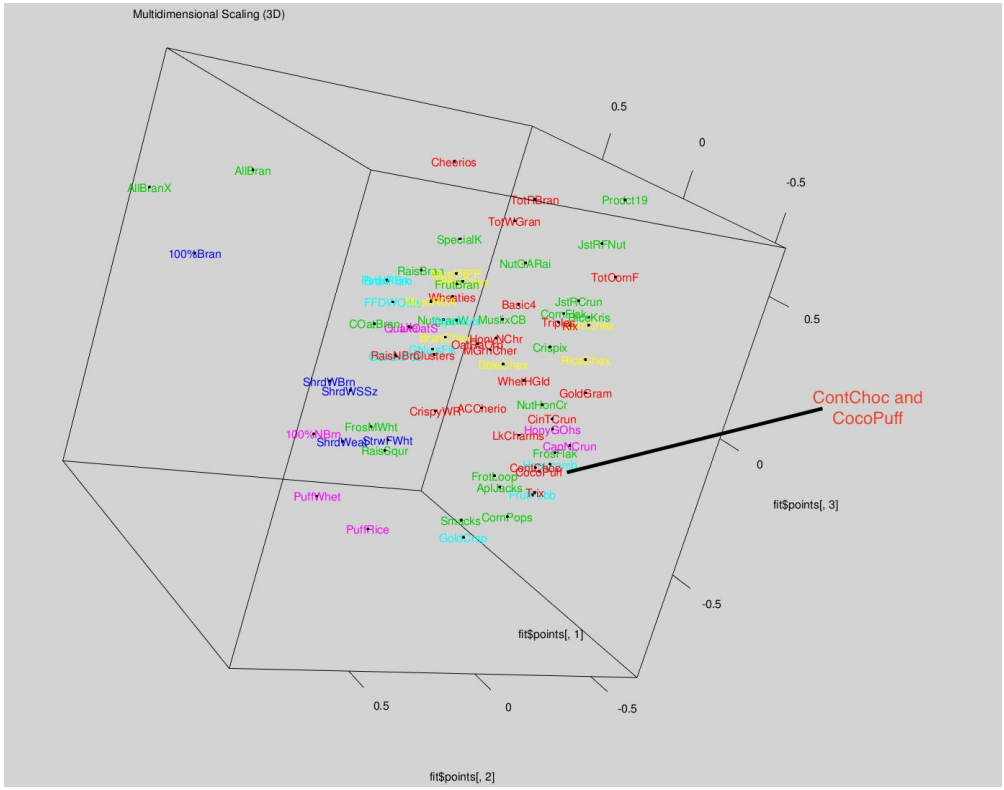


Figure 14 - Biclustering

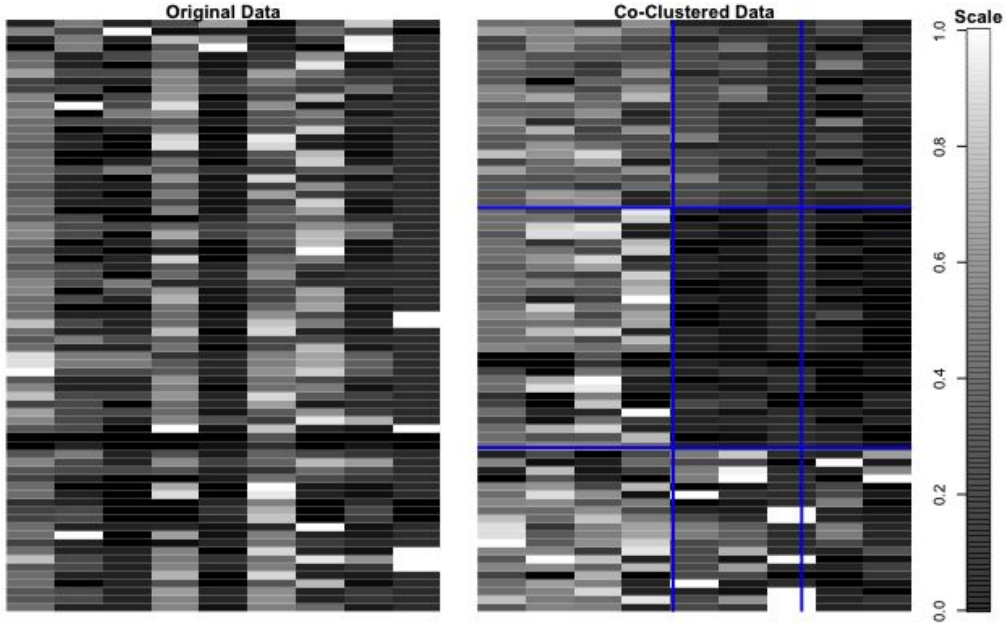
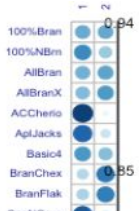
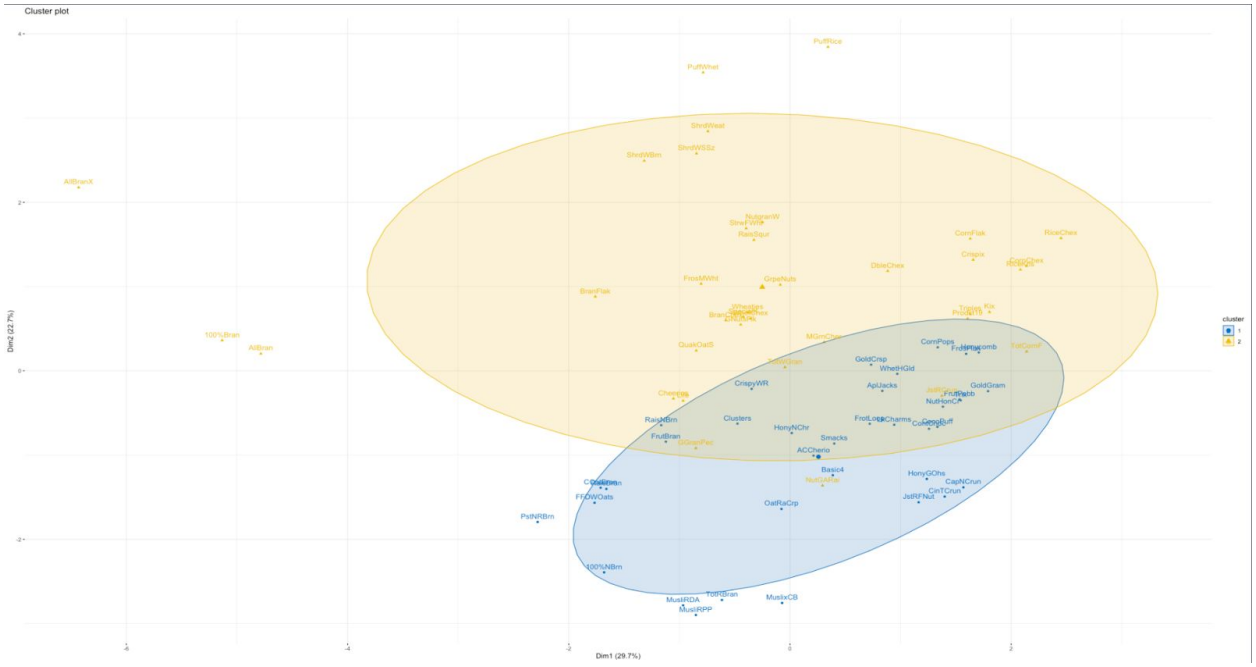


Figure 15 - Fuzzy Cluster Allocations



(please refer to “15_fuzzyclusterallocations.jpg” for non-truncated visualization)

Figure 16 - Fuzzy Clustering Plot



References

Boivin, Caroline, Catherine Parissier, Alexandre Alle, Penelope Asselin Forcier, and Simon Langlois. “Healthy Breakfast Cereals: What Do Consumers Want?” Journal of Foodservice Business Research 17, no. 1 (2014): 48–55. https://doi.org/10.1080/15378020.2014.886914.

Lee, C.M, H.R Moskowitz, and S-Y Lee. “Expectations, Need, and Segmentation of Healthy Breakfast Cereal Consumers.” Journal of Sensory Studies 22, no. 5 (2007): 587–607. https://doi.org/10.1111/j.1745-459X.2007.00127.x.

Xu, R. and Wunsch, D. (2009) Clustering . IEEE Press Series on Computational Intelligence. Hoboken, NJ: John Wiley Sons.

Souiden, Nizar, Fouad Ben Abdelaziz, and Audrey Fauconnier. “Nutrition Labelling: Employing Consumer Segmentation to Enhance Usefulness.” The Journal of Brand Management 20, no. 4 (2013): 267–82. https://doi.org/10.1057/bm.2012.14.

Burke. “Competitive Positioning Strength: Market Measurement.” Journal of Strategic Marketing. 19, no. 5 (n.d.): 421–28. https://doi.org/info:doi/.

Dolnicar, Sara, Sebastian Kaiser, Katie Lazarevski, and Friedrich Leisch. “Biclustering: Overcoming Data Dimensionality Problems in Market Segmentation.” Journal of Travel Research 51, no. 1 (2012): 41–49. https://doi.org/10.1177/0047287510394192.

Izenman, Alan Julian. 2013. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. New York, NY: Springer. [ISBN-13: 978-0-387-78189-1] Chapter 12: Cluster Analysis, pages 407-420. Available from the Springer collection: https://link-springer-com.turing.library.northwestern.edu/book/10.1007%2F978-0-387-78189-1Links to an external site.

Chapman, Chris. & Feit, E. R For Marketing Research and Analytics. Cham: Springer, 2015.

“cereals: Cereal nutrition data.” Statlib CMU, 1993.

About the Author

Vincent Pun is a Master’s candidate in the data science program at Northwestern University, experienced in financial services process automation, tax reporting, and strategy consulting.