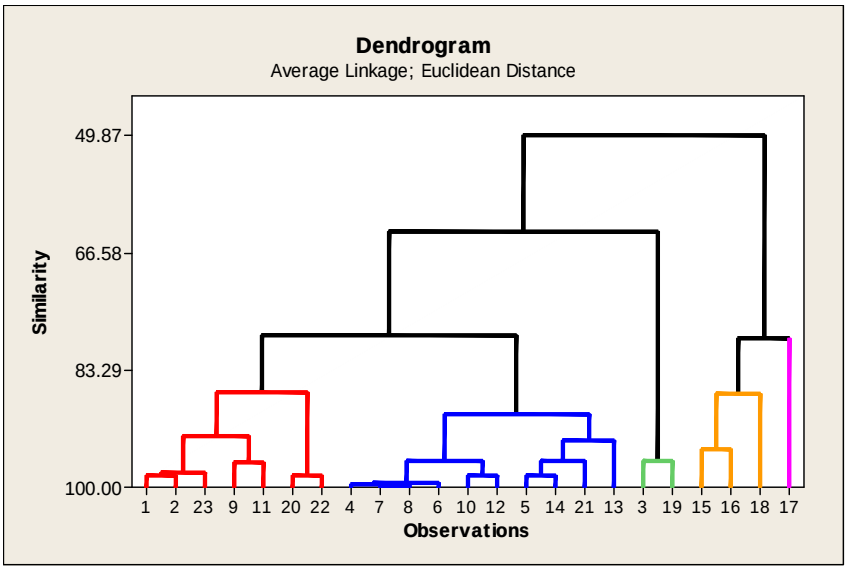
Breakfast Cereal - Solution

The software for cluster analysis typically offers a number of options with respect to the clustering algorithm used. Here is a dendrogram from a cluster analysis (by Minitab) where we have chosen the options Euclidean distance and average linkage directly on the data.



We see here the stepwise clustering, starting with every item for itself (100% similarity). In each stage two clusters are joined, and we may follow how the degree of similarity degreases for each joining, going upwards in the dendrogram until all items are joined in one cluster. The numbering of the items is according to the order of items in the data file.
The details of the clustering process follow below. The columns of the table show, for each step, the number of clusters, the similarity and distance levels, the name of the clusters joined and the name of the new cluster (the smallest number for the two clusters joined), the number of items in the new cluster. This continues until all items are joined together.

Cluster Analysis of Observations: Energy; Protein; Carbo; Fat; Fibres

Euclidean Distance, Average Linkage

Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	22	99.79	1.225	7	8	7	2
2	21	99.60	2.281	4	7	4	3
3	20	99.39	3.481	4	6	4	4
4	19	98.40	9.100	10	12	10	2
5	18	98.39	9.192	5	14	5	2
6	17	98.34	9.449	20	22	20	2
7	16	98.26	9.912	1	2	1	2
8	15	98.01	11.378	1	23	1	3
9	14	96.44	20.303	9	11	9	2
10	13	96.32	21.018	4	10	4	6
11	12	96.31	21.048	3	19	3	2
12	11	96.18	21.780	5	21	5	3
13	10	94.68	30.373	15	16	15	2
14	9	93.30	38.245	5	13	5	4
15	8	92.73	41.463	1	9	1	5
16	7	89.57	59.503	4	5	4	10
17	6	86.76	75.515	15	18	15	3

18	5	86.41	77.523	1	20	1	7
19	4	78.89	120.392	15	17	15	4
20	3	78.34	123.529	1	4	1	17
21	2	63.69	207.096	1	3	1	19
22	1	49.87	285.942	1	15	1	23

We see that the similarity measure decreases little during the first 12 steps, where 11 clusters remain. Then it decreases somewhat more until step 18, where we are left with 5 clusters. At step 19 a larger drop occurs, while the drop at step 20 is small. At this point we are left with 3 clusters. This result indicates that it may be is meaningful to study 3 or 5 clusters.

The dendrogram presents the results more transparent. By horizontal cuts at the similarity levels 66 and 83 we get 3 and 5 clusters respectively. In the situation with three clusters. the clusters are the items (3,19), (15,16,17,18) and (1,2,4,5,6,7,8,9,10,11,12,13,14,20,21,22,23) If we choose to split in five clusters, we keep (3,19), while (17) is removed from the second cluster, and constitutes a cluster for itself. The third cluster is split in two as follows (1,2,9,11,20,22,23) and (4,5,6,7,8,10,12,13,14,21).

After having specified a final partition, the software offers the supplementary tables given below. The first table summarize each cluster by the number of items, the sum of squared distances within the cluster, and the average and maximal distance from the items of the cluster to its centroid. In general, a cluster with small sum of squares will be more compact than one with a larger som og squares. The centroid is the vector in the space of observed variables determined by the averages of the variables for the items within the cluster, which is taken as the midpoint in the cluster. The second table gives the centroids for each cluster, and the third table gives the distance between the cluster centroids. This may help when interpreting each cluster. Note that the clusters are numbered from 1 to 5 according to the lowest numbered item within the cluster.

Final Partition

Number of clusters: 5

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	7	10868.237	32.810	56.169
Cluster2	2	221.500	10.524	10.524
Cluster3	10	10417.709	28.708	63.243
Cluster4	3	4259.000	33.659	50.317
Cluster5	1	0.000	0.000	0.000

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Grand centrd
Energy	1494.0000	1360.0000	1616.9000	1800.0000	1920.0000	1594.2174
Protein	9.9571	10.5000	6.9200	7.5000	7.5000	8.2565
Carbo	73.1143	63.5000	80.7400	69.3333	76.0000	75.2261
Fat	2.7571	2.5000	3.5900	13.6667	18.0000	5.1826
Fibres	6.9714	13.0000	4.1000	6.6667	5.5000	6.1435

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster1	0.0000	134.4810	123.2101	306.2278	426.2920
Cluster2	134.4810	0.0000	257.6588	440.2361	560.4121
Cluster3	123.2101	257.6588	0.0000	183.7503	303.4831
Cluster4	306.2278	440.2361	183.7503	0.0000	120.2688
Cluster5	426.2920	560.4121	303.4831	120.2688	0.0000

We see that the two large clusters (1 and 3) are in the middle for Energy, medium-high on Carbohydrates and low on Fat. What makes them separate is mainly that Cluster 1 is higher than Cluster 3 on Protein and slightly lower on Energy and Carbohydrates. Cluster 2 contains two products that are high on Protein and Fibre, but low on Energy and Carbohydrates. . Cluster 4 contains three products high on Energy and Fat. Cluster 5 contains only one product. It is close to Cluster 4, but became separate since it is slightly different on all characteristics except Protein.

I will be of interest to repeat the cluster analysis using other joining criteria, say single linkage. It turns out that there may be slight differences in the solutions, but they will not overturn the general conclusions. Furthermore, it may be interesting to see if the replacement of Carbohydrates by the two variables Sugar and Starch will affect the clustering process, and furthermore whether the addition of Sodium will change things. It turns out that the solutions become different, apart from some special features that remain the same.

An objection to the analysis above may be that we used it directly on the original data, despite that their magnitude is very different. Cluster software typically has an option to do the analysis on standardized data, or equivalent using correlation (Pearson) distance. If not, it is easy to standardize data yourself by, for each variable, subtract its mean and divide by its standard deviation. A cluster analysis on the standardized variables came out very differently, as seen by comparing the solution with five clusters:

Original data: (1,2,9,11,20,22,23) (3,19) (4,5,6,7,8,10,12,13,14,21) (15,16,18) (17)

Standardized data: (1,4,5,6,7,8,9,10,11,12,13,21,23) (2) (3,19,20,22) (14) (15,16,17,18)

We note here the big cluster of 13 of the 23 products and that two new ones came out as unique. Only the Crusli type products (15 to 18) remain mostly similar. If we increase to six clusters product 3 goes out of its cluster of four. We will not go into the possible interpretation of the standardized analysis here, but invites the reader to perform this analysis and do so. The cluster solution taken to be the most meaningful, will depend in the aim of the analysis, combined with the insight in the role of the variables in the product, for instance from the nutrition point of view. The lesson may be:

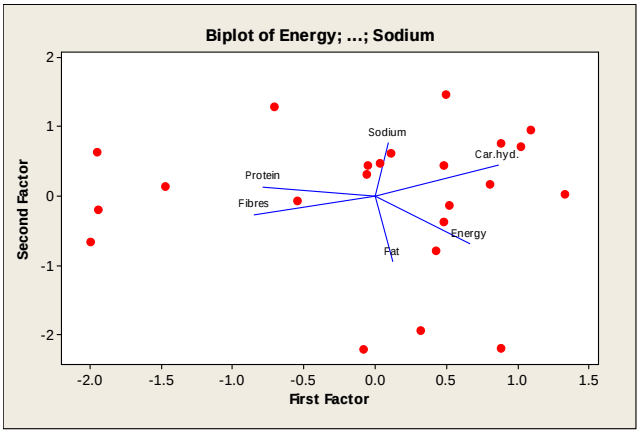
Be aware that the many options of cluster analysis may give different results, and be sceptical to all of them unless they are supported by substantial insights in the area of application!

Our data were limited to 23 products from five different brands. Some brands have several products within the same cluster, and some have no product in a cluster. This may expose possibilities for adding or removing products to certain brands in the shelves. Of course sales statistics will also be of major importance when such decisions are made.

Recall that the data are just observed on the shelves in two stores. We have no information about whether some of the brands have additional products covering a wider spectrum not taken into these specific stores. For the product developer or the advertiser, the complete set of competing products is necessary to perform a meaningful cluster analysis intended to uncover possibilities for positioning.

Factor analysis is often regarded as an alternative to cluster analysis. This is typically done on the correlation matrix. This corresponds to doing it on the standardized variables. Here only two factors turn out significant. The software has a number of choices with respect to how the factors are extracted and whether the solution should be rotated to obtain a solution

more easily interpreted (hopefully). Here we illustrate the unrotated principal component solution by a biplot. This is revealing how the two factors are composed and where each of the 23 observations are located with respect to these in a “map”. We see that the first factor is mainly (Energy, Carbohydrate) versus (Protein, Fibres) and the second factor Sodium versus Fat. In order to match the cluster analysis we should be able to point out clusters in this map. However, clusters are not apparent, except the group of four points on the left and at group of three at the bottom. The rest looks like one big cluster, except possibly the single points on top and one in isolation in the middle.



In the secondary plot below we introduced the product number and see that it conforms fairly well to the clusters from the standardized cluster analysis,

(1,4,5,6,7,8,9,10,11,12,13,21,23) (2) (3,19,20,22) (14) (15,16,17,18)

One exception is product 14 which is in the middle of the large cluster, but will show up separately if we plot the scores of the third factor against the second.