Contents

1 K-Means Clustering

- Hierarchical clustering requires a distance or similarity matrix between all pairs of cases. That's an extremely large matrix if you have tens of thousands of cases in your data file.
- A clustering method that doesn't require computation of all possible distances is k-means clustering. It differs from hierarchical clustering in several ways. You have to know in advance the number of clusters you want. You can't get solutions for a range of cluster numbers unless you rerun the analysis for each different number of clusters.
- The algorithm repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during the analysis. In agglomerative hierarchical clustering, on the other hand, cases are added only to existing clusters. They are forever captive in their cluster, with a widening circle of "neighbours".
- The algorithm is called **k-means**, where **k** is the number of clusters you want, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest.
- The k-means algorithm follows an entirely different concept than the hierarchical methods discussed before. This algorithm is not based on distance measures such as Euclidean distance or city-block distance, but uses the ***within-cluster variation*** as a measure to form homogenous clusters. Specifically, the procedure aims at segmenting the data in such away that the within-cluster variation isminimized.Consequently,we do not need to decide on a distance measure in the first step of the analysis.
- The action in the algorithm centers around finding the k-means. You start out with an initial set of means and classify cases based on their distances to the centers.
- Next, you compute the cluster means again, using the cases that are assigned to the cluster; then, you reclassify all cases based on the new set of means. You keep repeating this step until cluster means don't change much between successive steps.
- Finally, you calculate the means of the clusters once again and assign the cases to their permanent clusters.

1.1 Initial Cluster Centres

The first step in k-means clustering is finding the k centres. This is done iteratively. You start with an initial set of centres and then modify them until the change between two iterations is small enough.

If you have good guesses for the centres, you can use those as initial starting points; otherwise, you can let SPSS find k cases that are well separated and use these values as initial cluster centers. (i.e. The clustering process starts by randomly assigning objects to a number of clusters).

K-means clustering is very sensitive to outliers, since they will usually be selected as initial cluster centers. This will result in outliers forming clusters with small numbers of cases. Before you start a cluster analysis, screen the data for outliers and remove them from the initial analysis. The solution may also depend on the order of the cases in the data.

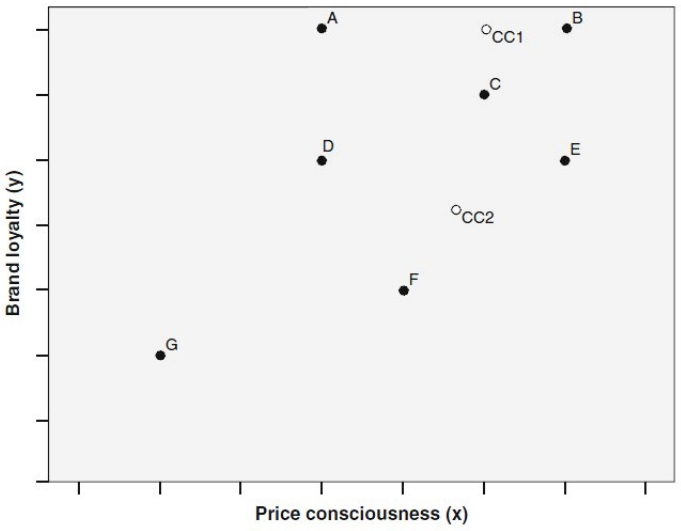
After the initial cluster centers have been selected, each case is assigned to the closest cluster, based on its distance from the cluster centers. After all of the cases have been assigned to clusters, the cluster centers are recomputed, based on all of the cases in the cluster.

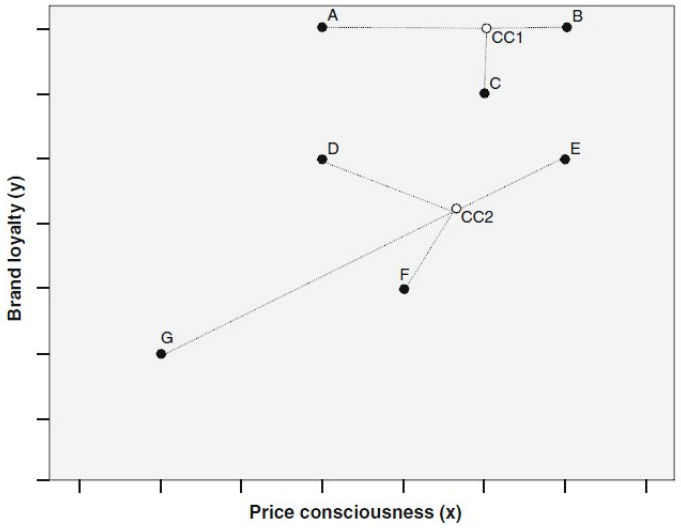
The cases are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an case to another cluster decreases the within-cluster variation, this case is reassigned to that cluster.

Case assignment is done again, using these updated cluster centers. You keep assigning cases and recomputing the cluster centers until no cluster center changes appreciably or the maximum number of iterations (10 by default) is reached.

1.2 Demonstration of k-means

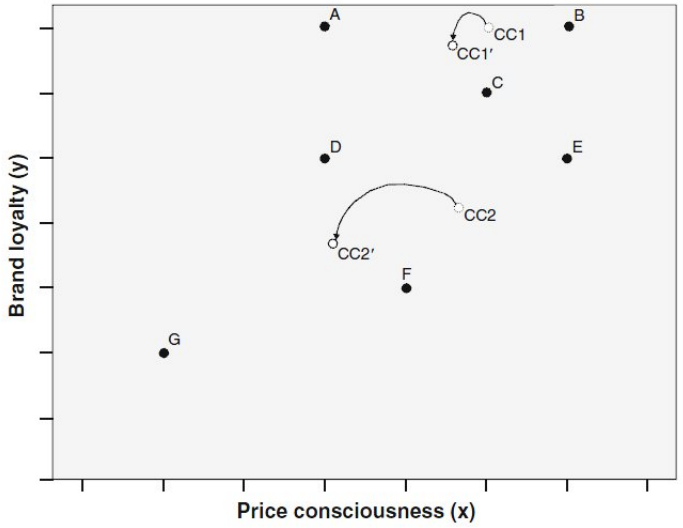
In this example, two cluster centers are randomly initiated, which CC1 (first cluster) and CC2 (second cluster).

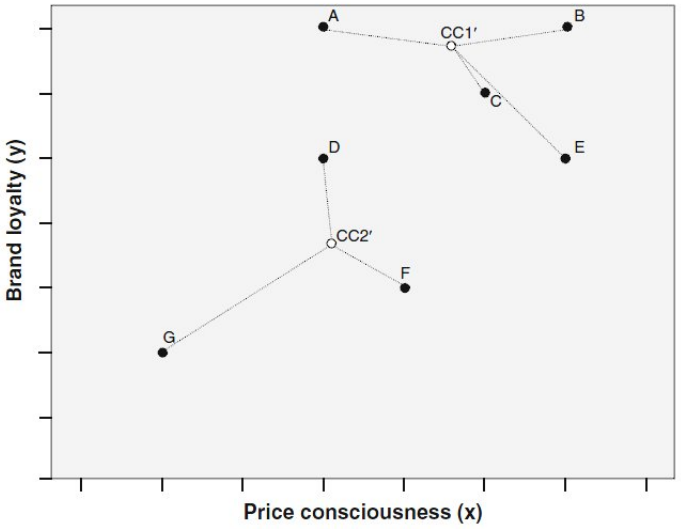




- Euclidean distances are computed from the cluster centers to every single object. Each object is then assigned to the cluster center with the shortest distance to it.
- In this example, objects A, B, and C are assigned to the first cluster, whereas objects D, E, F, and G are assigned to the second. We now have our initial partitioning of the objects into two clusters.
- Based on this initial partition, each clusters geometric center (i.e., its centroid) is computed (third step).

- This is done by computing the mean values of the objects contained in the cluster (e.g., A, B, C in the first cluster) regarding each of the variables (in this example: price consciousness and brand loyalty).





- Both clusterscenters now shift into new positions (CC1 for the first and CC2 for the second cluster).
- In the fourth step, the distances from each object to the newly located cluster centers are computed and objects are again assigned to a certain cluster on the basis of their minimum distance to other cluster centers (CC1 and CC2).

- Since the cluster centers position changed with respect to the initial situation in the first step, this could lead to a different cluster solution. This is also true of our example, as object E is now unlike in the initial partition closer to the first cluster center (CC1) than to the second (CC2). Consequently, this object is now assigned to the first cluster.
- The k-means procedure now repeats the third step and re-computes the cluster centers of the newly formed clusters, and so on. In other words, steps 3 and 4 are repeated until a predetermined number of iterations are reached, or convergence is achieved (i.e., there is no change in the cluster affiliations).

1.3 Performance of k-means clustering

Generally, k-means is superior to hierarchical methods as it is less affected by outliers and the presence of irrelevant clustering variables. Furthermore, k-means can be applied to very large data sets, as the procedure is less computationally demanding than hierarchical methods. In fact, we suggest definitely using k-means for sample sizes above 500, especially if many clustering variables are used. From a strictly statistical viewpoint, k-means should only be used on interval or ratio-scaled data as the procedure relies on Euclidean distances. However, the procedure is routinely used on ordinal data as well, even though there might be some distortions.

One problem associated with the application of k-means relates to the fact that the researcher has to pre-specify the number of clusters to retain from the data. This makes k-means less attractive to some and still hinders its routine application in practice.