Cluster analysis

It is a technique for grouping objects, cases, entities on the basis of multiple variables. The advantage of the technique is that it is applicable to both metric and non-metric data.

Secondly, the grouping can be done post hoc, i.e. after the primary data survey is over. The technique has wide applications in all branches of management. However, it is most often used for market segmentation analysis.

- Can be used to cluster objects, individuals and entities
- Similarity is based on multiple variables
- Measures proximity between study variables
- Groups that are grouped in one cluster are homogenous as compared to others
- Can be conducted on metric, non-metric as well as mixed data

Applications in Marketing Research

- Market segmentation – customers/potential customers can be split into smaller more homogenous groups by using the method.
- Segmenting industries – the same grouping principle can be applied for industrial consumers.
- Segmenting markets – cities or regions with similar or common traits can be grouped on the basis of climatic or socio-economic conditions.
- Career planning and training analysis – for human resource planning people can be grouped into clusters on the basis of their educational/experience or aptitude and aspirations.
- Segmenting financial sector/instruments – different factors like raw material cost, financial allocations, seasonality and other factors are being used to group sectors together to understand the growth and performance of a group of industries.

For Metric data analysis

$$d_{ij} = \sqrt{\sum_{k=1}^{3} \left(X_{ik} - X_{jk} \right)^2}$$

Where,

- d_{ij} = distance between person *i* and *j*.
- *k* = variable (interval / ratio)
- *i* = object
- *j* = object

For Non-metric data

1. Simple matching coefficient = $\frac{p}{(p+m+n)}$
2. Jaccard coefficient = $\frac{p}{(p+m)}$

Where

- P=positive matches
- N=negative matches
- M=mismatches

For Mixed Data

$$S_{ij} = \frac{\sum_{k=1}^{m} W_k S_{ijk}}{\sum_{k=1}^{m} W_k}$$

Where, S_{ij} = The similarity of objects i and j,

S_{ijk} is the similarity of objects i and j on the k^{th} characteristic, with m characteristics in all. (The value S_{ijk} must be = 0 and = 1.)

With qualitative characters, it is 1 when there is a match and 0 with a mismatch.

With quantitative characters $S_{ijk} = (X_{ik} - X_{jk} |/R_k)$, where X_{ik} and X_{jk} are the values of attribute k for the i^{th} and j^{th} objects,

R_k = The range of character k in the sample,

W_k = The weight attached to the k^{th} attribute.

Key Statistic related to Cluster analysis

Cluster centroid. The cluster centroid is the mean values of the variables for all the cases or objects in a particular cluster.
Cluster centres. The cluster centres are the initial starting points in non-hierarchical clustering.Clusters are built around these centres or seeds.
Cluster membership. Cluster membership indicates the cluster to which each object or case belongs.
Dendrogram. A dendrogram, or tree graph, is a graphical device for displaying clustering results. Vertical lines represent clusters that are joined together. The position of the line on the scale indicates the distances at which clusters were joined. The dendrogram is read from left to right.
Distances between cluster centres. These distances indicate how separated the individual pairs of clusters are. Clusters that are widely separated are distinct and therefore desirable.
Icicle diagram. An icicle diagram is a graphical display of clustering results. It is so called because it resembles a row of icicles hanging from the eaves of a house. The columns correspond to the objects being clustered, and the rows correspond to the number of clusters. An icicle diagram is read from bottom to top.
Similarity/distance coefficient matrix. A similarity/distance coefficient matrix is a lower Triangular matrix containing pairwise distances between objects or cases.

Steps for conducting cluster analysis:

1. **Formulate the problem**

The most important part of formulating the clustering problem is to select the variables on which the clustering is based. Inclusion of even one or two irrelevant variables may mislead an otherwise useful clustering solution. Basically, the set of variables selected should describe the similarity between objects in terms that are relevant to the marketing research problem. The variables should be selected based on past research, theory or the hypotheses being developed or tested. If cluster analysis is used as an exploratory approach, the researcher uses his or her judgment.

2. **Select a distance measure**

As the objective of clustering is to group similar objects together, some measure is needed to assess how similar or different the objects are. The most common approach is to measure similarity in terms of distance between pairs of objects. Objects with smaller distances between them are more similar to each other than are those at larger distances.

There are several ways to compute the distance between two objects. The most commonly used measure of similarity is the Euclidean distance or its square. The Euclidean distance is the square root of the sum of the squared differences in values for each variable. Other distance measures are also available. The city-block or Manhattan distance between two objects is the sum of the absolute differences in values for each variable. The Chebychev distance between two objects is the maximum absolute difference in values for any variable.
If the variables are measured in vastly different units, the clustering solution will be influenced by the units of measurement. To overcome this, we must standardize the data by rescaling each variable to have a mean of 0 and a standard deviation of 1. Although standardization can remove the influence of the unit of measurement, it can also reduce the differences between groups on variables that may best discriminate groups or clusters. It is also desirable to eliminate outliers. Use of different distance measures may lead to different clustering results. Hence, it is suggested to use different measures and to compare the results. After selecting a distance or similarity measure, we select clustering procedure.

3. Select a clustering procedure

Clustering procedures can be hierarchical or non-hierarchical.
Hierarchical clustering is characterised by the development of a hierarchy or tree like structure. Hierarchical methods can be agglomerative or divisive. **Agglomerative clustering** starts with each object in a separate cluster. Clusters are formed by grouping objects into bigger and bigger clusters. This process is continued until all objects are members of a single cluster. **Divisive clustering** starts with all the objects grouped in a single cluster. Clusters are divided or split until each object is in a separate cluster.

Agglomerative methods are commonly used in marketing research. They consist of linkage methods, error sums of squares or variance methods, and centroid methods.
Linkage methods include single linkage, complete linkage and average linkage. The **single linkage** method is based on minimum distance or the nearest neighbour rule. The first two objects clustered are those that have the smallest distance between them. The next shortest distance is identified, and either the third object is clustered with the first two or a new two-object cluster is formed. At every stage, the distance between two clusters is the distance between their two closest points. Two clusters are merged at any stage by the single shortest link between them. This process is continued until all objects are in one cluster. The single linkage method does not work well when the clusters are poorly defined. The **complete linkage** method is similar to single linkage, except that it is based on the maximum distance or the farthest neighbor approach. In complete linkage, the distance between two clusters is calculated as the distance between their two farthest points. The **average linkage** method works similarly. In this method, the distance between two clusters is defined as the average of the distances between all pairs of objects, where one member of the pair is from each of the clusters. The average linkage method uses information on all pairs of distances, therefore, it is usually preferred to the single and complete linkage methods.

The **variance methods** generate clusters by minimizing the within-cluster variance. A commonly used variance method is **Ward's procedure**. For each cluster, the means for all the variables are computed. Then, for each object, the squared Euclidean distance tothe cluster means is calculated, and these distances are summed for all the objects. At each stage, the two clusters with the smallest increase in the overall sum of squares within cluster distances are combined. In the **centroid method**, the distance between two clusters is the distance between their centroids (means for all the variables). Every time objects are grouped, a new centroid is computed. Out of the hierarchical methods, the average linkage method and Ward's procedure have been reported to be better than the other procedures.

The second type of clustering procedures, the **non-hierarchical clustering** methods, are also referred to as *k*-means clustering. These methods include sequential threshold, parallel threshold and optimising partitioning. In the sequential threshold method, a cluster centre is selected and all objects within a pre-specified threshold value from the centre are grouped together. A new cluster centre or seed is then selected, and the process is repeated for the unclustered points. Once an object is clustered with a seed, it is no longer considered for clustering with subsequent seeds. The **parallel threshold method** operates similarly except that several cluster centres are selected simultaneously and objects within the threshold level are grouped with the nearest centre. The **optimizing partitioning method** differs from the two threshold procedures in that objects can later be reassigned to clusters to optimize an overall criterion, such as average within-cluster distance for a given number of clusters. Two major disadvantages of the non-hierarchical procedures are that the number of clusters must be pre-specified and that the selection of cluster centres is arbitrary. Furthermore, the clustering results may depend on how the centres are selected. Many non-hierarchical programs select the first *k* cases (*k* = number of clusters) without missing values as initial cluster centres. Thus, the clustering results may depend on the order of observations in the data. Yet non-hierarchical clustering is faster than hierarchical methods and has merit when the number of objects or observations is large. It has been suggested that the hierarchical and non-hierarchical methods be used in tandem. First, an initial clustering solution is obtained using a hierarchical procedure, such as average linkage or Ward's. The number of clusters and cluster centroids so obtained are used as inputs to the optimizing partitioning method. The choice of a clustering method and the choice of a distance measure are interrelated. For example, squared Euclidean distances should be used with Ward's and the centroid methods. Several non-hierarchical procedures also use squared Euclidean distances.

4. Decide on the number of clusters

A major issue in cluster analysis is deciding on the number of clusters. Although there are no hard and fast rules, some guidelines are available:

- Theoretical, conceptual or practical considerations may suggest a certain number of clusters.

- In hierarchical clustering, the distances at which clusters are combined can be used as criteria. This information can be obtained from the agglomeration schedule or from the dendrogram.

- In non-hierarchical clustering, the ratio of total within-group variance to between-group variance can be plotted against the number of clusters. The point at which an elbow or a sharp bend occurs indicates an appropriate number of clusters. Increasing the number of clusters beyond this point is usually not worthwhile.
- The relative sizes of the clusters should be meaningful

5. Interpret and profile clusters

Interpreting and profiling clusters involves examining the cluster centroids. The centroids represent the mean values of the objects contained in the cluster on each of the variables. The centroids enable us to describe each cluster by assigning it a name or label. It is often helpful to profile the clusters in terms of variables that were not used for clustering, such as demographic, psychographic, product usage, media usage or other variables. For example, the clusters may have been derived based on benefits sought. Further profiling may be done in terms of demographic and psychographic variables to target marketing efforts for each cluster. The variables that significantly differentiate between clusters can be identified via discriminant analysis and one-way analysis of variance.

6. Assess the reliability and validity

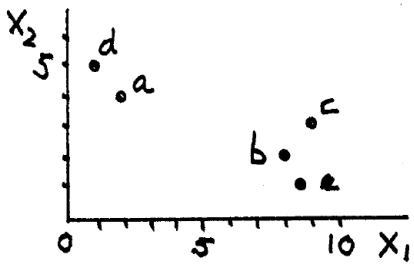
The following procedures provide adequate checks on the quality of clustering results.

1 Perform cluster analysis on the same data using different distance measures. Compare the results across measures to determine the stability of the solutions.
2 Use different methods of clustering and compare the results.
3 Split the data randomly into halves. Perform clustering separately on each half. Compare cluster centroids across the two subsamples.
4 Delete variables randomly. Perform clustering based on the reduced set of variables. Compare the results with those obtained by clustering based on the entire set of variables.
5 In non-hierarchical clustering, the solution may depend on the order of cases in the dataset. Make multiple runs using different order of cases until the solution stabilizes.

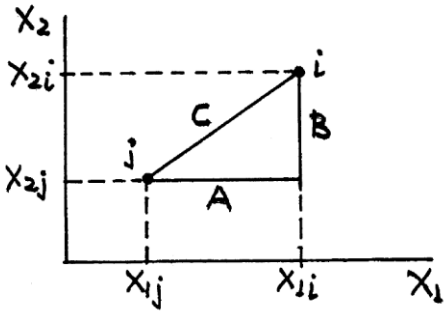
Example: The daily expenditure on food (X_1) and clothing (X_2) of five persons are given in following table:

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

First we plot the data:



For clustering, we use distance measures. For two points *i* and *j* with coordinates (X_{1i}, X_{2i}) and (X_{1j}, X_{2j})



The *Euclidean distance* between the two points is the hypotenuse of the triangle ABC:

$$D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}.$$

An observation i is declared to be closer (more similar) to j than to observation k if $D(i,j) < D(i,k)$.

An alternative measure is the *squared Euclidean distance*.

$$D_2(i,j) = A^2 + B^2 = (X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2.$$

Yet another measure is the *city block distance*, defined as

$$D_3(i,j) = |A| + |B| = |X_{1i} - X_{1j}| + |X_{2i} - X_{2j}|.$$

As the name suggests, it is the distance one would travel if the points i and j were located at opposite corners of a city block.

The distance measures can be extended to more than two variables. For example, the Euclidean distance between an observation $(X_{1i}, X_{2i}, \ldots, X_{ki})$ and another $(X_{1j}, X_{2j}, \ldots, X_{kj})$ is

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}.$$

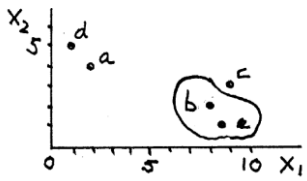
All three measures of distance depend on the units in which X_1 and X_2 are measured, and are influenced by whichever variable takes numerically larger values. For this reason, the variables are often standardized so that they have mean 0 and variance 1 before cluster analysis is applied. Alternatively, weights w_1, w_2, ..., w_k reflecting the importance of the variables could be used and a weighted measure of distance calculated. For example,

$$D(i,j) = \sqrt{w_1(X_{1i} - X_{1j})^2 + w_2(X_{2i} - X_{2j})^2 + \cdots + w_k(X_{ki} - X_{kj})^2}.$$

For the given data, the Euclidean distance between persons is given by

Cluster	a	b	c	d	e
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	1.118
c			0	8.246	2.062
d				0	8.500
e					0

As person b and e are nearest to each other so they are put in same cluster (Using single linkage method) as shown below:



Assuming the nearest neighbor method is used, the distance between the cluster (be) and another observation is the smaller of the distances between that observation, on the one hand, and b and e, on the other. For

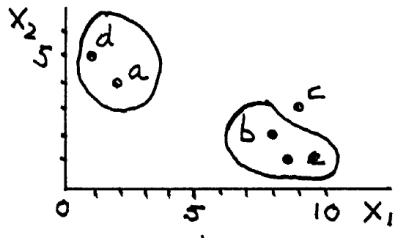
example,

$$D(be, a) = \min\{D(b, a), D(e, a)\} = \min\{6.325, 7.159\} = 6.325.$$

The distance between these clusters is given by:

Cluster	(be)	a	c	d
(be)	0	6.325	1.414	7.614
a		0	7.071	1.414
c			0	8.246
d				0

\



The distance between (be) and (ad) is

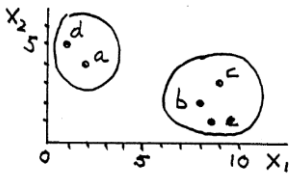
$$D(be, ad) = \min\{D(be, a), D(be, d)\} = \min\{6.325, 7.616\} = 6.325,$$

while that between c and (ad) is

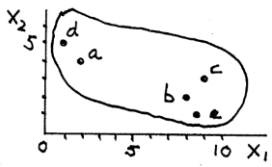
$$D(c, ad) = \min\{D(c, a), D(c, d)\} = \min\{7.071, 8.246\} = 7.071.$$

Using nearest neighbor method, we get:

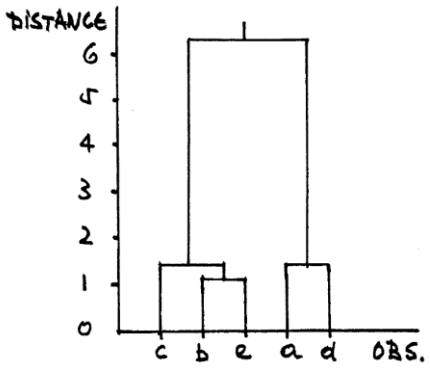
Cluster	(be)	(ad)	c
(be)	0	6.325	1.414
(ad)		0	7.071
c			0



Cluster	(bce)	(ad)
(bce)	0	6.325
(ad)		0



The dendogram is given by


DISTANCE
6
5
4
3
2
1
0
c
b
e
a
d
OBS.