****

**What is Clustering?**

Clustering is dividing data points into homogeneous classes or clusters:

* Points in the same group are as similar as possible
* Points in different group are as dissimilar as possible

When a collection of objects is given, we put objects into group based on similarity.

**Application of Clustering:**

Clustering is used in almost all the fields. You can infer some ideas from Example 1 to come up with lot of clustering applications that you would have come across.

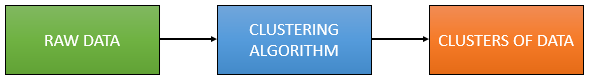
Listed here are few more applications, which would add to what you have learnt.

* Clustering helps **marketers** improve their customer base and work on the target areas. It helps group people (according to different criteria’s such as willingness, purchasing power etc.) based on their similarity in many ways related to the product under consideration.
* Clustering helps in identification of **groups of houses** on the basis of their value, type and geographical locations.
* Clustering is used to study earth-quake. Based on the areas hit by an earthquake in a region, clustering can help analyse the next probable location where earthquake can occur.

**Clustering Algorithms:**

A Clustering Algorithm tries to analyse natural groups of data on the basis of some similarity. It locates the **centroid of the group of data points**. To carry out effective clustering, the algorithm evaluates the **distance between each point** **from the centroid of the cluster.**

The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data.



**Unsupervised:**

**What is K-means Clustering?**

K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K-means clustering is a method of **vector quantization**, originally from signal processing, that is popular for cluster analysis in data mining.

K-Means clustering is a very popular and simple clustering technique. The main objective of K-Means clustering is to group the similar data points into clusters. Here, ‘**K’ means the number of clusters**, which is predefined.

**K-means Clustering – Example 1:**

A pizza chain wants to open its delivery centres across a city. What do you think would be the possible challenges?

* They need to analyse the areas from where the pizza is being ordered frequently.
* They need to understand as to how many pizza stores has to be opened to cover delivery in the area.
* They need to figure out the locations for the pizza stores within all these areas in order to keep the distance between the store and delivery points minimum.

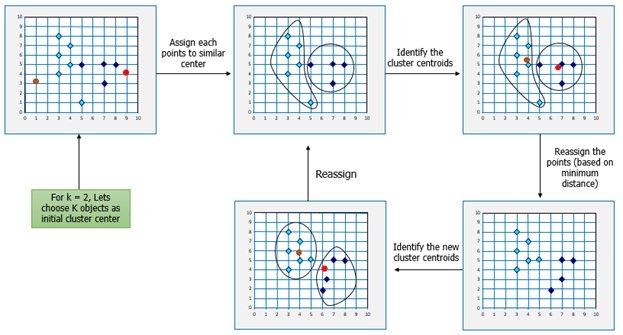
Resolving these challenges includes a lot of analysis and mathematics. We would now learn about how clustering can provide a meaningful and easy method of sorting out such real life challenges. Before that let’s see what clustering is.

**K-means Clustering Method:**

If k is given, the K-means algorithm can be executed in the following steps:

* Partition of objects into k non-empty subsets
* Identifying the cluster centroids (mean point) of the current partition.
* Assigning each point to a specific cluster
* Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
* After re-allotting the points, find the centroid of the new cluster formed.

**The step by step process:**

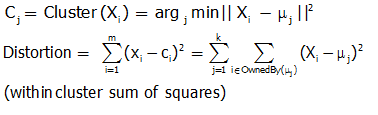


Now, let’s consider the problem in Example 1 and see how we can help the pizza chain to come up with centres based on K-means algorithm.

## ****Mathematical Formulation for K-means Algorithm:****

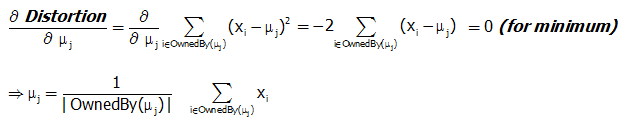
D= {**x1,x2,…,xi,…,xm**} à data set of m records

**xi**= (xi1,xi2,…,xin) à each record is an n-dimensional vector



## ****Finding Cluster Centers that Minimize Distortion:****

Solution can be found by setting the partial derivative of Distortion w.r.t. each cluster center to zero.



For any k clusters, the value of k should be such that even if we increase the value of k from after several levels of clustering the distortion remains constant. The achieved point is called the “Elbow”.

This is the ideal value of k, for the clusters created.

k-means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of ‘K’.

Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.

**Hierarchical Agglomerative andDivisive clustering**

# **Hierarchical Agglomerative Clustering**

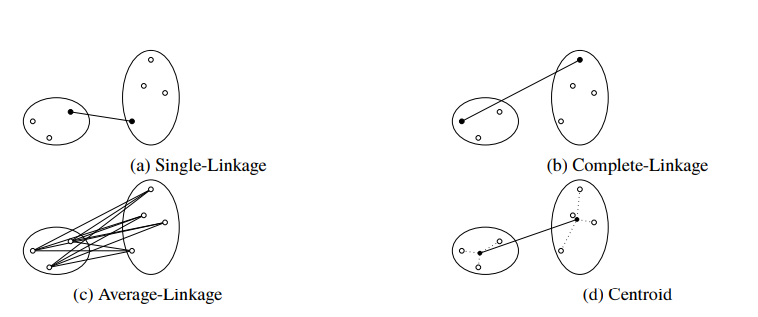
The hierarchical agglomerative clustering uses the bottom-up approaches. In the HAC algorithm starts with every single data point as a single cluster. The similar clusters are successively merged until all clusters have merged into a one cluster and result is represents in tree structure as named dendogram.

**The Single-Linkage Criterion:** The single-linkage criterion for hierarchical clustering merges groups based on the shortest distance over all possible pairs

**The Complete-Linkage Criterion:** Rather than choosing the shortest distance, in completelinkage clustering the distance between two groups is determined by the largest distance over all possible pairs

**The Average-Linkage Criterion:** Rather than the worst or best distances, when using the average linkage criterion we average over all possible pairs between the groups

**The Centroid Criterion:** Another alternative approach to computing the distance between clusters is to look at the difference between their centroids

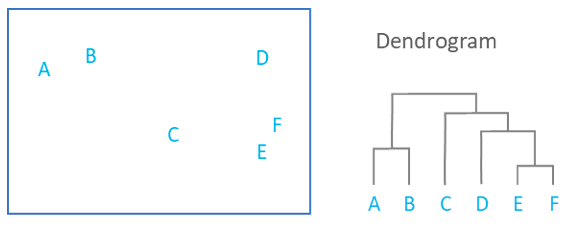


A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from [hierarchical clustering](https://www.displayr.com/what-is-hierarchical-clustering/). The main use of a dendrogram is to work out the best way to allocate objects to clusters. The dendrogram below shows the hierarchical clustering of six observations shown on the scatterplot to the left. (Dendrogram is often miswritten as dendogram.)

How to read a dendrogram

The key to interpreting a dendrogram is to focus on the height at which any two objects are joined together. In the example above, we can see that E and F are most similar, as the height of the link that joins them together is the smallest. The next two most similar objects are A and B.

In the dendrogram above, the height of the dendrogram indicates the order in which the clusters were joined. A more informative dendrogram can be created where the heights reflect the distance between the clusters as is shown below. In this case, the dendrogram shows us that the big difference between clusters is between the cluster of A and B versus that of C, D, E, and F.

[](https://app.displayr.com/Try/Template%20Hierarchical%20Cluster%20Analysis)

It is important to appreciate that the dendrogram is a summary of the distance matrix, and, as occurs with most summaries, information is lost. For example, the dendrogram suggests that C and D are much closer to each other than is C to B, but the original data (shown in the scatterplot), shows us that this is not true. To use some jargon, a dendrogram is only accurate when data satisfies the ultrametric tree inequality, and this is unlikely for any real-world data.

The consequence of the information loss is that the dendrograms are most accurate at the bottom, showing which items are very similar.