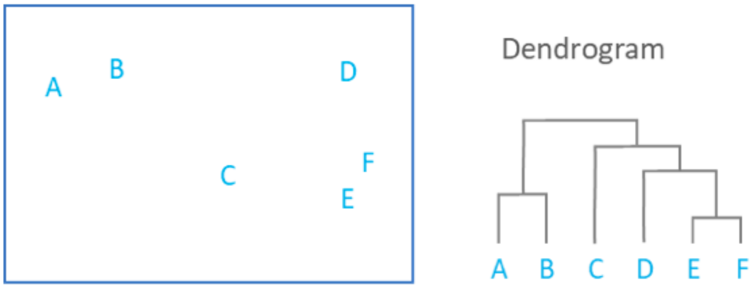
WHAT IS...

What is a Dendrogram?

by Tim Bock

A *dendrogram* is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from *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 to on the *scatterplot* to the left*.* (Dendrogram is often miswritten as dendogram.)



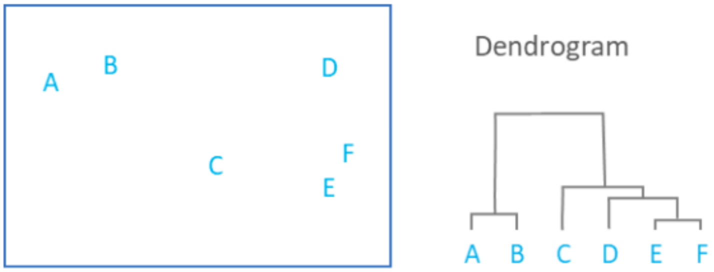
Use Template

To create your own dendrogram using hierarchical clustering, simply click the button above!

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.

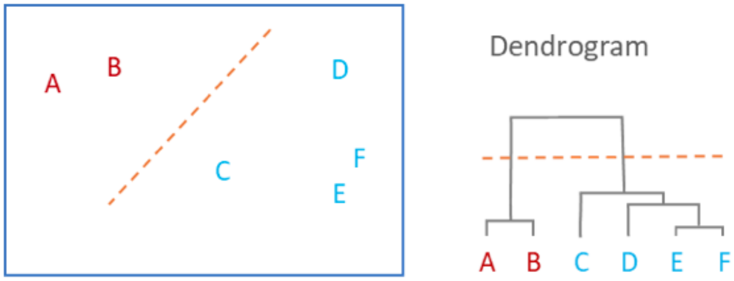


It is important to appreciate that the dendrogram is a summary of the distance matrix, and, as occurs with most summaries, information is lost. For example, the dendrogram suggests that C and D are much closer to each other than is C to B, but the original data (shown in the scatterplot), shows us that 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.

Allocating observations to clusters

Observations are allocated to clusters by drawing a horizontal line through the dendrogram. Observations that are joined together below the line are in clusters. In the example below, we have two clusters, one that combines A and B, and a second combining C, D, E, and F.



Dendrograms cannot tell you how many clusters you should have

A common mistake people make when reading dendrograms is to assume that the shape of the dendrogram gives a clue as to how many clusters exist. In the example above, the (incorrect) interpretation is that the dendrogram shows that there are two clusters, as the distance between the clusters (the vertical segments of the dendrogram) are highest between two and three clusters.

Such an interpretation is justified only when the ultrametric tree inequality holds, which, as mentioned above, is very rare. In general, it is a mistake to use dendrograms as a tool for determining the number of clusters in data. Where there is an obviously “correct” number of clusters this will often be evident in a dendrogram. However, dendrograms often suggest a correct number of clusters when there is no real evidence to support the conclusion.

We hope you're now an expert in dendrograms! If you want to find out more data science terminology, check out our "What is" and "How to" series. Or even better, create your own hierarchical cluster for free!

Use Template