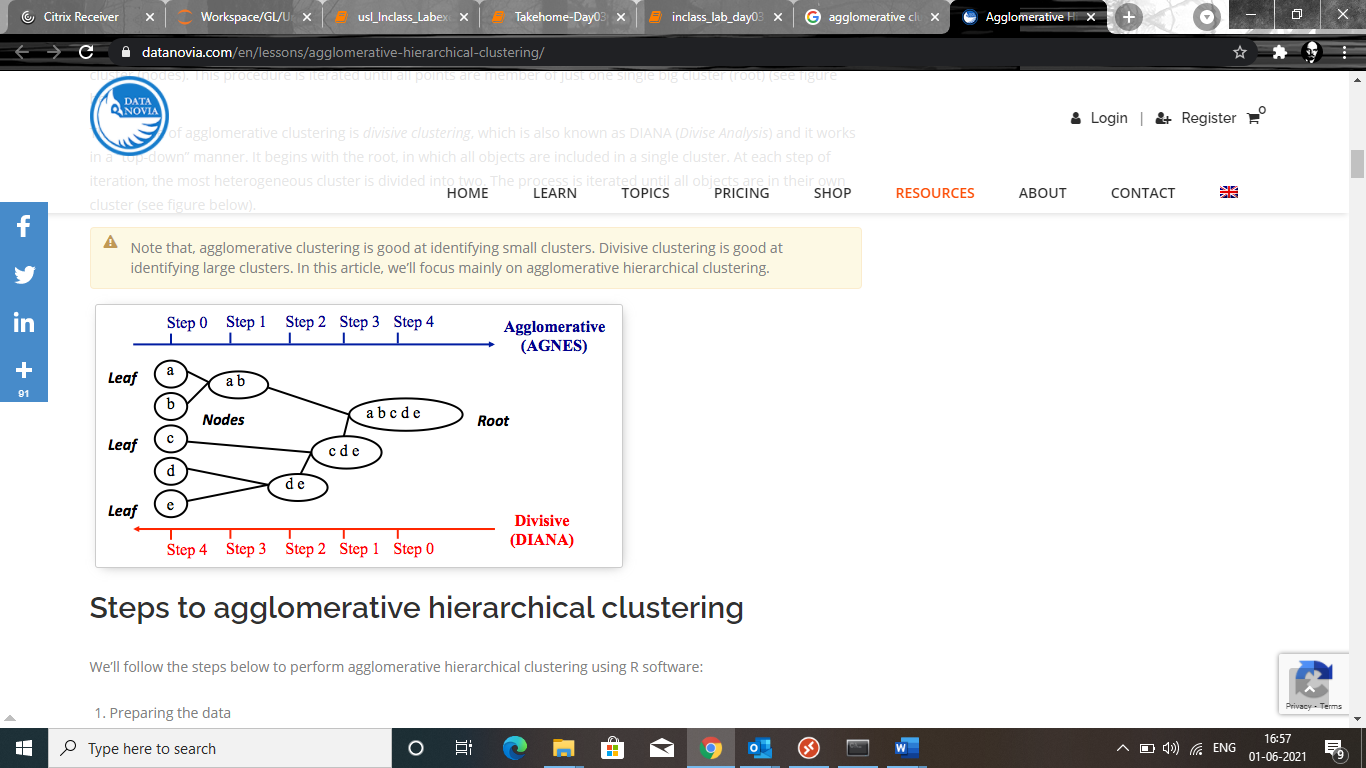
**Agglomerative Hierarchical Clustering**

The **agglomerative clustering** is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It’s also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram.

**Algorithm**

Agglomerative clustering works in a “bottom-up” manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster



## Steps to agglomerative hierarchical clustering

We’ll follow the steps below to perform agglomerative hierarchical clustering using R software:

1. Preparing the data
2. Computing (dis)similarity information between every pair of objects in the data set.
3. Using linkage function to group objects into hierarchical cluster tree, based on the distance information generated at step 1. Objects/clusters that are in close proximity are linked together using the linkage function.
4. Determining where to cut the hierarchical tree into clusters. This creates a partition of the data.

### Similarity measures

In order to decide which objects/clusters should be combined or divided, we need methods for measuring the similarity between objects.

There are many methods to calculate the (dis)similarity information, including [Euclidean and manhattan distances](https://www.datanovia.com/en/lessons/clustering-distance-measures/). In R software, you can use the function dist() to compute the distance between every pair of object in a data set. The results of this computation is known as a distance or dissimilarity matrix.

By default, the function dist() computes the Euclidean distance between objects; however, it’s possible to indicate other metrics using the argument method.

### Linkage

The linkage function takes the distance information, returned by the function dist(), and groups pairs of objects into clusters based on their similarity. Next, these newly formed clusters are linked to each other to create bigger clusters. This process is iterated until all the objects in the original data set are linked together in a hierarchical tree.

For example, given a distance matrix “res.dist” generated by the function dist(), the R base function hclust() can be used to create the hierarchical tree.

Maximum or *complete linkage*: The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.

 Minimum or *single linkage*: The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, “loose” clusters.

 Mean or *average linkage*: The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2.

 *Centroid linkage*: The distance between two clusters is defined as the distance between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2.

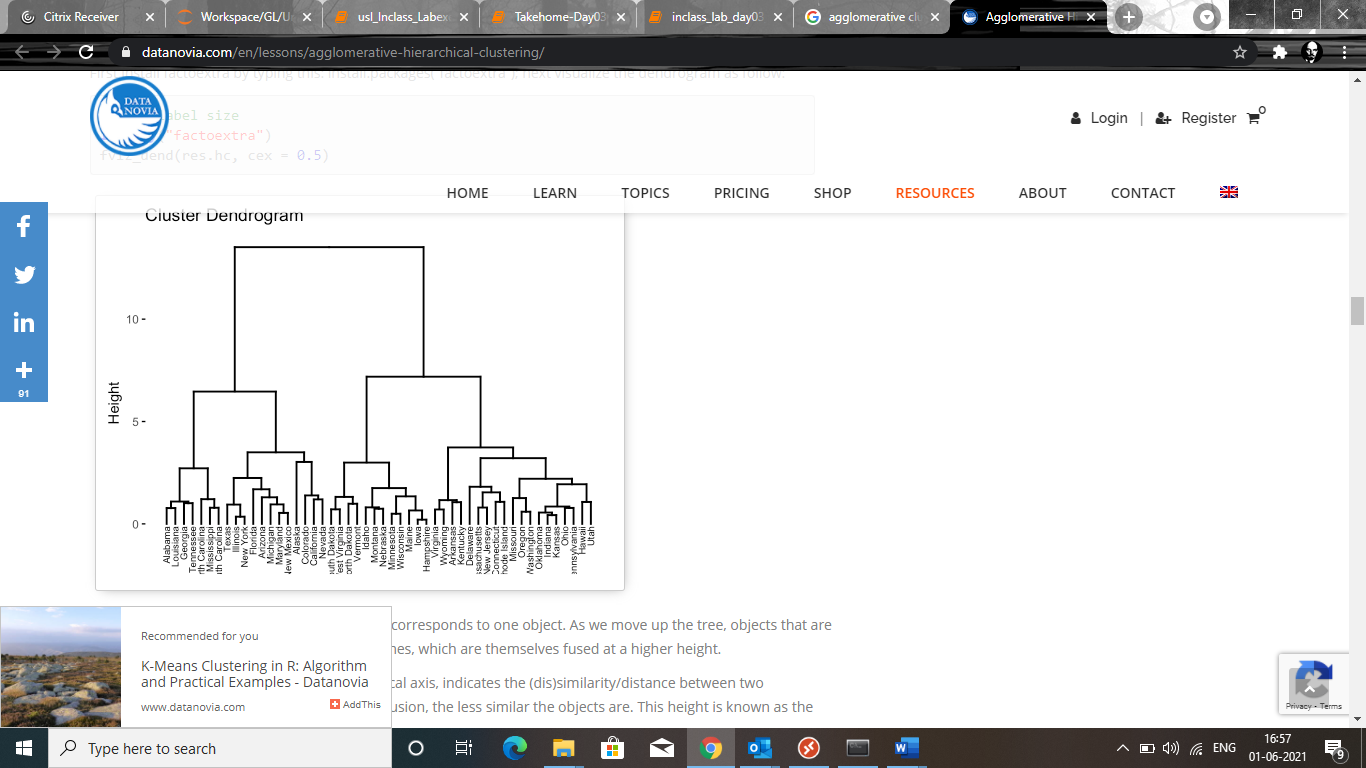
 *Ward’s minimum variance method*: It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

### Dendrogram

Dendrograms correspond to the graphical representation of the hierarchical tree generated by the function

In the dendrogram displayed above, each leaf corresponds to one object. As we move up the tree, objects that are similar to each other are combined into branches, which are themselves fused at a higher height.

The height of the fusion, provided on the vertical axis, indicates the (dis)similarity/distance between two objects/clusters. The higher the height of the fusion, the less similar the objects are. This height is known as the cophenetic distance between the two objects.



## Verify the cluster tree

After linking the objects in a data set into a hierarchical cluster tree, you might want to assess that the distances (i.e., heights) in the tree reflect the original distances accurately.

One way to measure how well the cluster tree generated by the hclust() function reflects your data is to compute the correlation between the cophenetic distances and the original distance data generated by the dist() function. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the original distance matrix.

The closer the value of the correlation coefficient is to 1, the more accurately the clustering solution reflects your data. Values above 0.75 are felt to be good. The “average” linkage method appears to produce high values of this statistic. This may be one reason that it is so popular.

## Cut the dendrogram into different groups

One of the problems with hierarchical clustering is that, it does not tell us how many clusters there are, or where to cut the dendrogram to form clusters.

You can cut the hierarchical tree at a given height in order to partition your data into clusters. The R base function cutree() can be used to cut a tree, generated by the hclust() function, into several groups either by specifying the desired number of groups or the cut height. It returns a vector containing the cluster number of each observation.