Hierarchical Clustering Exercises

Grouping objects into clusters is a frequent task in data analysis. In this set of exercises we will use **hierarchical clustering** to cluster European capitals based on their latitude and longitude. Before trying out this exercise please make sure that you are familiar with the following functions: dist, hlcust, cutree, rect.hclust

We will be using a custom-made dataset. Before starting the exercise please run the following code to obtain the capital locations for Europe (note that you will need to have ggmap library installed):

```
library(ggmap)
capitals <- c("Albania, Tirana", "Andorra, Andorra la Vella","Armenia, Yerevan", "Austria,
Vienna", "Azerbaijan, Baku","Belarus, Minsk", "Belgium, Brussels", "Bosnia and Herzegovina,
Sarajevo", "Bulgaria, Sofia", "Croatia, Zagreb", "Cyprus, Nicosia", "Czech Republic, Prague",
"Denmark, Copenhagen", "Estonia, Tallinn", "Finland, Helsinki", "France, Paris", "Germany,
Berlin", "Greece, Athens", "Georgia, Tbilisi", "Hungary, Budapest", "Iceland, Reykjavik", "Italy,
Rome", "Latvia, Riga", "Kazakhstan, Astana", "Liechtenstein, Vaduz", "Lithuania, Vilnius",
"Luxembourg, Luxembourg", "Macedonia, Skopje", "Malta, Valletta", "Moldova, Chişinău",
"Monaco, Monaco-Ville", "Montenegro, Podgorica", "Netherlands, Amsterdam", "Norway,
Oslo", "Poland, Warsaw", "Portugal, Lisbon", "Republic of Ireland, Dublin", "Romania,
Bucharest", "Russia, Moscow", "San Marino, San Marino", "Serbia, Belgrade", "Slovakia,
Bratislava", "Slovenia, Ljubljana", "Spain, Madrid", "Sweden, Stockholm", "Switzerland, Bern",
"Turkey, Ankara", "Ukraine, Kiev", "United Kingdom, London", "Vatican City, Vatican City")
theData <- geocode(capitals)
rownames(theData) <- capitals
```

Exercise 1

Calculate the Euclidean latitude/longitude distances between all pairs of capital cities.

Exercise 2

Use the obtained distances to produce the hierarchical clustering dendrogram object. Use all the default parameters. NOTE: By default the clusters will be merged together using the maximum possible distance between all pairs of their elements (this fact will be useful later).

Exercise 3

Visualize the obtained hierarchical clustering dendrogram.

Exercise 4

In the previous step the leaves of our dendrogram were placed at different heights. Let’s redo the plot so that all capital names are written at the same level.

Exercise 5

Hierarchical clustering procedure builds a hierarchy of clusters. One advantage of this method is that we can use the same dendrogram to obtain different numbers of groups. Cluster the European capitals into 3 groups.

Exercise 6

Instead of specifying the wanted number of groups we can select the dendrogram height where the tree will be partitioned into groups. Since we used the maximum linkage function (default in exercise 2) this height has a useful interpretation – it ensures that all elements within one cluster are not more than the selected distance apart.

a) Cluster the European capitals by cutting the tree at height=20.

b) Plot the dendrogram and visualize the height at which the tree was cut into groups using a line.

Exercise 7

Now visualize the clustering solution obtained in the 5th exercise on the dendrogram plot. This should be done by drawing a rectangle around all capitals that fall in the same group. Use different colors for different groups.

Exercise 8

Visualize the dendrogram again but this time present both cluster versions obtained in exercise 5 and exercise 6 on the same plot. Use red color to represent exercise 5 clusters and blue to represent clusters from exercise 6.

Exercise 9

The hclust function has 8 implemented different linkage methods – methods used to merge two clusters when building the dendrogram. We want to experiment with all of them. Produce a dendrogram, obtain 5 groups and vizualize them using different color rectangles. Repeat this for all available linkage methods.

Exercise 10

Design your own clustering solution based on what you learned in this exercise and visualize it as a map. Plot capital coordinates with longitude on the x-axis and latitude on the y-axis and color them based on the groups obtained using your hierarchical clustering version.