Chapter 17 - Cluster analysis

Politics.csv contains the views of 150 respondents on four political issues, in 5 different age groups ranging from 61+ to 18-30. We are interested in differences in age groups. Produce a dendrogram and provide a (tentative) opinion on what it may represent. It is recommended that you reduce the sample size for clarity.



Generally, you can use the code in the Hierarchical clustering section of Chapter 17. A few adaptations are suggested:

 The set-up phase will require "Politics.csv" as the file and "Age" as the 'name' column variable. To reproduce the results below, keep the "Euclidean" and "ward.D2" options.

The second, file-reading, paragraph could have the following:

data = read.csv(file)

library(dplyr)

data = sample\_n(data, 30) # creates a sub-sample of 30 randomly selected cases

The next line is the same:

usedscales = colnames(data) [-1]

The next line of code does not use scaling, as the data in the file has already been transformed:

datamatrix = data[, usedscales]

The final major paragraph remains the same as in the book, producing a dendrogram:



Do note that the random sampling methods used above means that your result may be somewhat different from the result shown here, and will change a little each time you try it. Details on this may be found in Chapter 21, R tutorial 4, Sampling, Random sampling.

You can of course change the size of the sub-sample, and maybe try out different distance and cluster options.

I have the impression that older respondents tend to have different points of view than younger ones. To demonstrate what I think I saw, let's add a couple of red boxes:

rect.hclust(fit, k=2)



People in their 50s and 60s tend to occupy one uber-cluster. Younger people tend to appear in the other. Not that I think that Brexit was that clear-cut!