BUAN 6356 (Johnston)
Homework 5A(20231101)
Due: 4 November 2023 (6PM)

Points available: 70

This assignment is about clustering a data file using both kmeans() and hclust(). The data file provided is named "HW_5.csv" and available through the UTDbox>data folder. You may want the data.table package for this assignment.

As always, the first commands of your code MUST include:

setwd("c:/data/BUAN6356/HW_5"); source("prep.txt", echo=T)

and the last command of your code MUST include:

source("validate.txt", echo=T)

The required code CAN be set up for conditional execution. (E.g.: set a Boolean variable and then use it in an if() to execute these statements.)

Be careful with the quote characters as they must ALL be the same at the beginning and end of a string. (Use the single or double quote character from the key next to "Enter".) Inclusion of these lines is required BEFORE your code will be tested. I hope that most of you understand this by now … :-)

Seed the RNG with 427409920 . Run scenarios for 1 through 10 (inclusive) clusters. Use a nstart= parameter value of 5. Be sure to both center and scale the data via scale() before use.

Your outcome deliverables are listed later in this document.

Submit the code to eLearning as an ASCII file which can be copied directly into R. (That is, the same way you have done this process for the earlier homework assignments.)

You may submit this assignment as many times as needed until you get full credit.

Deliverables (all names are case sensitive; models are result of fit functions):

1. seed (vector) random number seed
2. minClust (vector) the minimum number of clusters
3. maxClust (vector) the maximum number of clusters
4. nst (vector) number of starting iterations in kmeans()
5. wk (data.frame) working data as read
6. kWSS (vector) vector of total within-sum-of-squares values for kmeans() scenarios
7. hWSS (vector) vector of total within-sum-of-squares values for hclust() scenarios

Notes/Hints:

- See the source files 06a_kmeans_boston_v6_20221026.txt and 06b_hclust_boston_v6_20221026.txt from UTDbox>demos for examples of running these analyses.
- Deliverables #6 and #7 cover the initial scenarios and report the total within-sum-of-squares for each scenario rather than the within-group-sum-of-squares for a particular scenario.
- Selection of a “best” number of clusters is not a deliverable for this assignment.
- Plots are not part of the deliverables for this assignment.