Homework 7

due on 04/14/2023 by 5pm

For this HW you will use the USArrests data.

1. In our class we mentioned the use of correlation-based distance and Euclidean distance as dissimilarity measures for hierarchical clustering. It turns out that these two measures are almost equivalent. Assume each observation has been centered to have mean zero and standard deviation one, and let r_{ij} denote the correlation between the ith and jth observations. Then the quantity $1 - r_{ij}$ is proportional to the squared Euclidean distance between the ith and jth observations. Using the data, show that this proportionality holds.

2. Section 3.3 on page 65 gives a formula for calculating the proportion of the total variation (PTV) explained by the principal components. We also saw that the PTV can be obtained using the `sdev` output of the `prcomp` function. Calculate the PTV using these two approaches – they should deliver the same results.

3. We aim at performing hierarchical clustering on the states.

   (a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

   (b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

   (c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

   (d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.