STAT401 Homework #5
Spring 2023

Q1. Suppose that we measure two variables X_1 and X_2 for five subjects, A to E. The data are given as follows:

	Variable	
Subjects	X_1	X_2
A	3	2
B	5	5
C	4	7
D	4	4
E	2	4

(a) Calculate the distance matrix *D* using the squared Euclidean distance.

(b) Cluster these five subjects using the single linkage clustering and draw a dendrogram.

(c) Cluster these five subjects using the complete linkage clustering and draw a dendrogram.

(d) Cluster these five subjects using the group average clustering and draw a dendrogram.

(e) Compare the results of (b), (c) and (d) and recommend a clustering result. By plotting data, justify your choice of clustering.

Q2. Sample correlations for five stocks were given as follows:

$$\begin{array}{l|ccccc|}
 & \text{JP Morgan} & \text{Citibank} & \text{Wells Fargo} & \text{Shell} & \text{Exxon Mobil} \\
\text{JP Morgan} & 1.00 & & & & \\
\text{Citibank} & 0.63 & 1.00 & & & \\
\text{Wells Fargo} & 0.51 & 0.57 & 1.00 & & \\
\text{Shell} & 0.12 & 0.32 & 0.18 & 1.00 & \\
\text{Exxon Mobil} & 0.16 & 0.21 & 0.15 & 0.68 & 1.00
\end{array}$$

Treat the sample correlations as similarity measures and answer the following questions.

(a) Cluster the stocks using the single linkage hierarchical procedures. Construct dendrogram.

(b) Cluster the stocks using the complete linkage hierarchical procedures.

Construct dendrogram.

(c) Cluster the stocks using the group average hierarchical procedures. Construct dendrogram.

(d) After comparing the results of (a), (b) and (c), recommend a clustering result. Justify your choice of clustering.

Q3. Suppose we measure two variables X_1 and X_2 for four individuals A, B, C, and D. The data are given as follows:

	Observations	
Individual	x_1	X_2
A	1	-1
B	3	1
C	1	3
D	4	5

(a) Use the *K*-means clustering technique to divide the individuals into *K* = 2 clusters. Start with the initial groups (AB) and (CD).

(b) Repeat the *K*-means clustering technique in (a) with the initial groups (AD) and (BC). Compare your solution with the solution in (a). Are they the same? Graph the individuals in terms of their (x_1, x_2) coordinates, and comment on the solution.

Q4. The data("CERIAL_R.DAT" is attached with this file) include measurements on 8 variables (Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, and Potassium) for 26 breakfast cereals. Note that the first line of the data includes names of the variables.

(a) Cluster the cereals using the single linkage, complete linkage, and group average hierarchical procedures. Construct dendrograms and compare the results.

(b) Cluster the cereals into *K* = 2, 3 4 using the *K*-means cluster procedure. Compare the results with those in (a).

(c) Based on the results in (a) and (b), recommend a clustering result.