1

Using the distances in Example 12.3, cluster the items using the average linkage hierarchical procedure. Draw the dendrogram. Compare the results with those in Examples 12.3 and 12.5.

已知

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

我們選取距離最近的 $3, 5$ 去合併成一組，並根據定義計算 (35) 與其他組間距離

$$d_{(35)1} = \frac{3+11}{2\times 1} = 7$$
$$d_{(35)2} = \frac{7+10}{2\times 1} = 8.5$$
$$d_{(35)4} = \frac{9+8}{2\times 1} = 8.5$$

可得新矩陣爲

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 8.5 & 9 & 0 & \\ 8.5 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

選取距離最近的 $2, 4$ 合併成一組，並根據定義計算組間距離

$$d_{(24)(35)} = \frac{7+10+9+8}{2\times 2} = 8.5$$

$$d_{(24)1} = \frac{9+6}{2\times 1} = 7.5$$

可得新矩陣爲

$$\begin{array}{cccc} & (35) & (24) & 1 \\ (35) & 0 & & \\ (24) & 8.5 & 0 & \\ 1 & 7 & 7.5 & 0 \end{array}$$

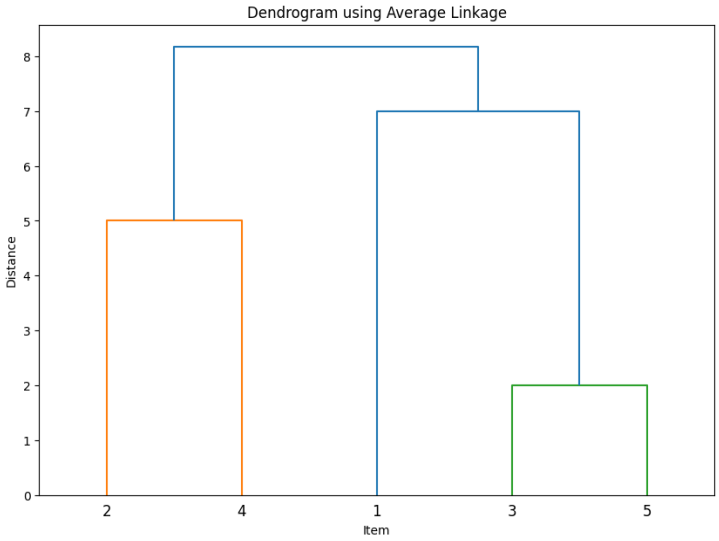
選取距離最近的 $1, (35)$ 合併成一組，並根據定義計算組間距離

$$d_{(135)(24)} = \frac{9+6+7+9+10+8}{3\times 2} \approx 8.167$$

可得新矩陣爲

$$\begin{array}{ccc} & (135) & (24) \\ (135) & 0 & \\ (24) & 8.167 & 0 \end{array}$$

可透過 python 繪製其 dendrogram



透過與講義的範例 12.3 及 12.5 比對可發現其結果與 complete linkage 結果不一樣，但與 single linkage 一樣，唯一不同的是 single linkage 是分完 (35) 後再合併 1，之後再分 (24)；而該題則是分完 (35) 後直接分 (24)，之後再用 (35) 合併 1。

2

The vocabulary "richness" of a text can be quantitatively described by counting the words used once, the words used twice, and so forth. Based on these counts, a linguist proposed the following distances between chapters of the Old Testament book Lamentations (data courtesy of Y. T. Radday and M. A. Pollatschek):

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ .76 & 0 & & & \\ 2.97 & .80 & 0 & & \\ 4.88 & 4.17 & .21 & 0 & \\ 3.86 & 1.92 & 1.51 & .51 & 0 \end{bmatrix} \end{array}$$

Cluster the chapters of Lamentations using the three linkage hierarchical methods we have discussed. Draw the dendrograms and compare the results.

根據題意，我們分別使用三種 linkage hierarchical methods

(a) Single linkage

選取 $3, 4$ 合併成一組，並根據定義計算組間距離

$$\begin{aligned} d_{(34)1} &= \min\{2.97, 4.88\} = 2.97 \\ d_{(34)2} &= \min\{0.8, 4.17\} = 0.8 \\ d_{(34)5} &= \min\{1.51, 0.51\} = 0.51 \end{aligned}$$

可得新矩陣爲

	(34)	1	2	5
(34)	0			
1	2.97	0		
2	0.8	0.76	0	
5	0.51	3.86	1.92	0

選取 $(34), 5$ 合併成一組，並根據定義計算組間距離

$$d_{(345)1} = \min\{2.97, 3.86\} = 2.97$$
$$d_{(345)2} = \min\{0.8, 1.92\} = 0.8$$

可得新矩陣爲

	(345)	1	2
(345)	0		
1	2.97	0	
2	0.8	0.76	0

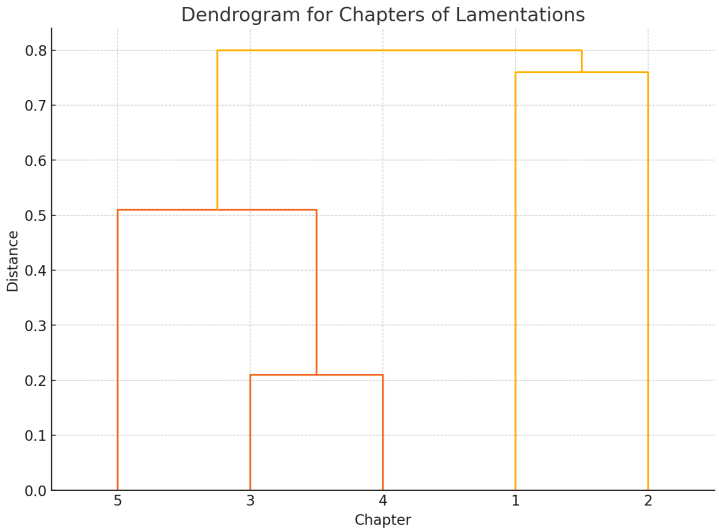
選取 $1, 2$ 合併成一組，並根據定義計算組間距離

$$d_{(12)(345)} = \min\{2.97, 0.8\} = 0.8$$

可得新矩陣爲

	(345)	(12)
(345)	0	
(12)	0.8	0

並且可根據結果繪製 dendrogram



(b) Complete linkage

選取 $3, 4$ 合併成一組，並根據定義計算組間距離

$$d_{(34)1} = \max\{2.97, 4.88\} = 4.88$$
$$d_{(34)2} = \max\{0.8, 4.17\} = 4.17$$
$$d_{(34)5} = \max\{1.51, 0.51\} = 1.51$$

可得新矩陣爲

	(34)	1	2	5
(34)	0			
1	4.88	0		
2	4.17	0.76	0	
5	1.51	3.86	1.92	0

選取 $1, 2$ 合併成一組，並根據定義計算組間距離

$$d_{(12)(34)} = \max\{4.88, 4.17\} = 4.88$$
$$d_{(12)5} = \max\{3.86, 1.92\} = 3.86$$

可得新矩陣爲

	(34)	(12)	5
(34)	0		
(12)	4.88	0	
5	1.51	3.86	0

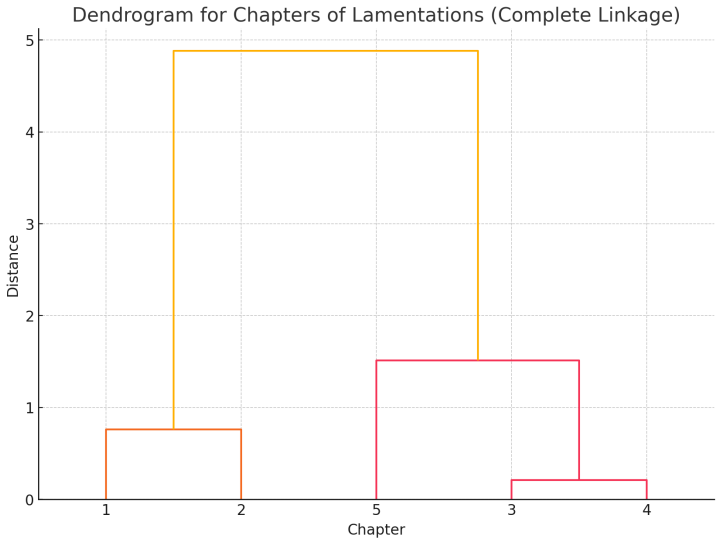
選取 $(34), 5$ 合併成一組，並根據定義計算組間距離

$$d_{(345)(12)} = \max\{4.88, 3.86\} = 4.88$$

可得新矩陣爲

	(345)	(12)
(345)	0	
(12)	4.88	0

並且可根據結果繪製 dendrogram



(c) Average linkage

選取 $3, 4$ 合併成一組，並根據定義計算組間距離

$$d_{(34)1} = \frac{2.97 + 4.88}{2 \times 1} = 3.925$$
$$d_{(34)2} = \frac{0.8 + 4.17}{2 \times 1} = 2.485$$
$$d_{(34)5} = \frac{1.51 + 0.51}{2 \times 1} = 1.01$$

可得新矩陣爲

	(34)	1	2	5
(34)	0			
1	3.925	0		
2	2.485	0.76	0	
5	1.01	3.86	1.92	0

選取 $1, 2$ 合併成一組，並根據定義計算組間距離

$$d_{(12)(34)} = \frac{2.97 + 4.88 + 0.8 + 4.17}{2 \times 2} = 3.205$$
$$d_{(12)5} = \frac{3.86 + 1.92}{2 \times 1} = 2.89$$

可得新矩陣爲

	(34)	(12)	5
(34)	0		
(12)	3.205	0	
5	1.01	2.89	0

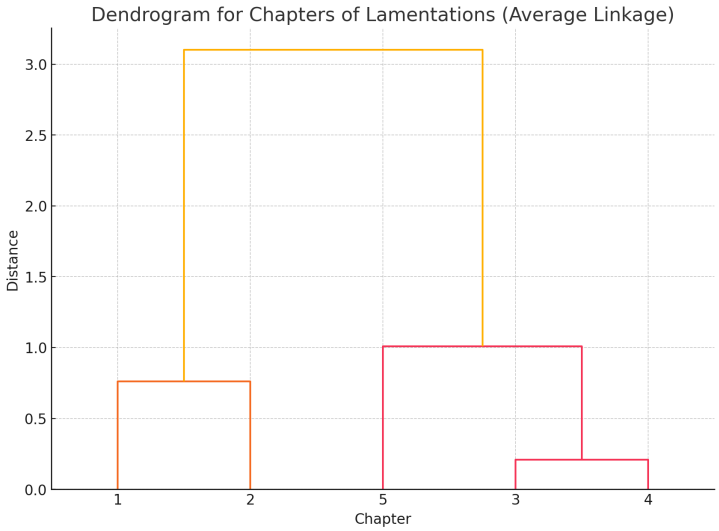
選取 $(34), 5$ 合併成一組，並根據定義計算組間距離

$$d_{(345)(12)} = \frac{2.97 + 0.8 + 4.88 + 4.17 + 3.86 + 1.92}{3 \times 2} = 3.1$$

可得新矩陣爲

	(345)	(12)
(345)	0	
(12)	3.1	0

並且可根據結果繪製 dendrogram



透過觀察可發現三種方法不僅得到的結果相同，甚至連合併的順序也很類似 (除了 Complete linkage 合併順序稍有不同)。

3

Repeat Example 12.11, but start at the bottom of the list of items, and proceed up in the order D, C, B, A. Begin with the initial groups (AB) and (CD). [The first potential reassignment will be based on the distances $d^2(D,(AB))$ and $d^2(D,(CD))$.] Compare your solution with the solution in the example. Are they the same? Should they be the same?

根據 Example 12.1 我們可得

Cluster	$\overline{x_1}$	$\overline{x_2}$
(AB)	2	-1
(CD)	-1	-2

接著分別計算

D 到兩個 cluster 間的距離:

$$d^2(\text{D}, (\text{AB})) = (-3-2)^2 + (-2-2)^2 = 25 + 16 = 41$$
$$d^2(\text{D}, (\text{CD})) = (-3+1)^2 + (-2+2)^2 = 4$$

⇒ 將 D 分到 (CD)

C 到兩個 cluster 間的距離:

$$d^2(\text{C}, (\text{AB})) = (1-2)^2 + (-2-2)^2 = 1 + 16 = 17$$
$$d^2(\text{C}, (\text{CD})) = (1+1)^2 + (-2+2)^2 = 4 + 0 = 4$$

⇒ 將 C 分到 (CD)

B 到兩個 cluster 間的距離:

$$d^2(\text{B}, (\text{AB})) = (-1-2)^2 + (1-2)^2 = 9 + 1 = 10$$
$$d^2(\text{B}, (\text{CD})) = (-1+1)^2 + (1+2)^2 = 9$$

⇒ 將 B 分到 (CD)

A 到兩個 cluster 間的距離:

$$d^2(\text{A}, (\text{AB})) = (5-2)^2 + (3-2)^2 = 9 + 1 = 10$$
$$d^2(\text{A}, (\text{CD})) = (5+1)^2 + (3+2)^2 = 36 + 25 = 61$$

⇒ 將 A 分到 (AB)

可得新的 clusters 爲

Cluster	$\overline{x_1}$	$\overline{x_2}$
A	5	3
(BCD)	-1	-1

D 到兩個 cluster 間的距離:

$$d^2(\text{D},\text{A}) = (-3-5)^2 + (-2-3)^2 = 64 + 25 = 89$$
$$d^2(\text{D},(\text{BCD})) = (-3+1)^2 + (-2+1)^2 = 4 + 1 = 5$$

⇒ 將 D 分到 (BCD)

C 到兩個 cluster 間的距離:

$$d^2(\text{C},\text{A}) = (1-5)^2 + (-2-3)^2 = 16 + 25 = 41$$
$$d^2(\text{C},(\text{BCD})) = (1+1)^2 + (-2+1)^2 = 4 + 1 = 5$$

⇒ 將 C 分到 (BCD)

B 到兩個 cluster 間的距離:

$$d^2(\text{B},\text{A}) = (-1-5)^2 + (1-3)^2 = 36 + 4 = 40$$
$$d^2(\text{B},(\text{BCD})) = (-1+1)^2 + (1+1)^2 = 0$$

⇒ 將 B 分到 (BCD)

A 到兩個 cluster 間的距離:

$$d^2(\text{A},\text{A}) = 0$$
$$d^2(\text{A},(\text{BCD})) = (5+1)^2 + (3+1)^2 = 36 + 16 = 52$$

⇒ 將 A 分到 A

由於結果不再更動，可得最終分組爲 A, (BCD)

4

In Assignment 2, we have performed principal components analysis on `FOODP.DAT` which gives the average price in cents per pound of five food items in 24 U.S. cities. Now perform cluster analysis on the principal components scores to group the cities. Assume that "two" principal components are obtained based on the analysis of correlation matrix

(a) Use single, complete, average, Wards methods to cluster the principal components scores.
   (i) Use a table (format as shown below) to list the R-square values versus Number-of-clusters (from 10 to 1) for all hierarchal clustering methods
   (ii) Plot the R-square versus Number-of-clusters (from 10 to 1) for all hierarchal clustering methods
   (iii) How many clusters do you recommend?

(b) Using the means of the clusters identified by all hierarchal clustering methods as the initial cluster centers, perform K-means method with four clusters. Present the data in a table format as below.

(c) Write down the results of the highest overall R-square in (b)
   (i) Show the final 4 cluster centers as identified by K-means method (with the highest overall R-square).
   (ii) Describe the characteristics of each cluster.

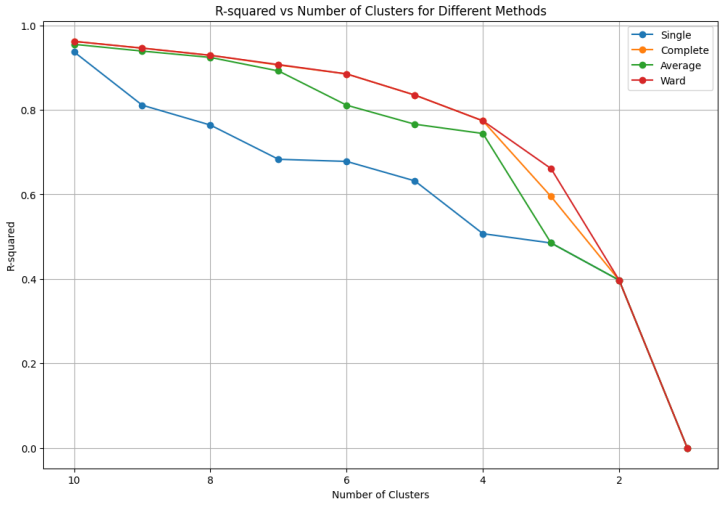
根據主成分分析，我們可得到資料的主成分分數爲

City	ID	Factor1	Factor2
Anchorage	1	4.674532	-0.539971
Atlanta	2	0.323536	0.081827
Baltimore	3	0.366146	1.125827
Boston	4	0.586307	1.652933
Buffalo	5	-2.691778	1.193800
Chicago	6	0.519218	-1.320803
Cincinnati	7	0.271076	0.804306
Cleveland	8	-0.753021	-0.280745
Dallas	9	0.390528	-0.943408
Detroit	10	-0.202286	1.563685
Honolulu	11	2.856135	-0.929948
Houston	12	-0.471509	-1.369421
Kansas City	13	-0.427955	0.403235
Los Angeles	14	-1.579078	-1.198470
Milwaukee	15	-1.311458	-0.555591
Minneapolis	16	-0.743544	0.616150
New York	17	1.316738	0.938323
Philadelphia	18	2.146621	0.812629
Pittsburgh	19	-0.326972	1.107534
St Louis	20	-1.279764	-0.020950
San Diego	21	-2.168506	-0.693957
San Francisco	22	-0.728630	-1.348053
Seattle	23	-1.023051	-0.344750
Washington	24	0.256718	-0.754180

(a) (i) 根據題意，下表呈現各分群方法 R-square 對應的 cluster 數量。

	R^2			
群集數目	Single	Complete	Average	Ward
10	.937	.962	.955	.962
9	.811	.946	.939	.946
8	.764	.929	.924	.929
7	.683	.906	.892	.907
6	.678	.885	.811	.885
5	.632	.835	.766	.835
4	.507	.774	.744	.774
3	.485	.595	.485	.661
2	.397	.397	.397	.397
1	.000	.000	.000	.000

(ii) 根據題意，使用 Python 畫出四種 Scree Plot 尋找 Elbow Point



(iii) 各方法的 Elbow Point 對應的 Cluster 數量如下表:

	Single	Complete	Average	Ward
Num. of Clusters	5	4	4	5

(b) 根據題意分別對四種 hierarchal 結果使用 k-means 進行分組可得下表:

Methods for Initial Seeds	Cluster No.				Overall R-Square
	Cluster #1	Cluster #2	Cluster #3	Cluster #4	
Single	2,3,4,6 7,9,10,13 16,19,24	1	11,17,18	5,8,12 14,15,20 21,22,23	0.691
Complete	6,8,9 12,14,22 23,24	5 15,20 21	2,3,4,7 10,13,16 17,18,19	1,11	0.757977
Average	6,8,9,12 14,15,20,21 22,23,24	2,3,4,7 10,13,16 17,18,19	1,11	5	0.749735
Ward	6,8,9 12,14,22 23,24	5 15,20 21	2,3,4 7,10,13,16 17,18,19	1,11	0.757977

(c) 根據 (b) 的結果可得由 Complete 與 Wards methods 結果得出的 R square 最大，其值爲 0.757977

(i) 可得四個 cluster 的中心點爲

集群	$\overline{X_1}$	$\overline{X_2}$
1	-0.4236	-0.9450
2	-1.8629	-0.0192
3	0.3310	0.9106
4	3.7653	-0.7350

(ii) Cluster #1:

- Cities: Chicago, Dallas, Milwaukee
- Characteristics: Central US cities grouped together, indicating Ward's method's focus on minimizing variance within clusters. Also, this cluster consists of cities with moderate to slightly above-average prices for the given food items.

Cluster #2:

- City: Buffalo
- Characteristics: Buffalo stands alone due to its distinct coordinates and geographical location. It also stands out with below-average prices for most food items, particularly for apples and tomatoes.

Cluster #3:

- Cities: Atlanta, Baltimore, Boston
- Characteristics: Major East Coast cities, grouped based on minimized variance. They are grouped together due to their moderate to high prices and wide range of cost for food items.

Cluster #4:

- Cities: Anchorage, Honolulu, Los Angeles, New York, Philadelphia, Pittsburgh, St. Louis, San Francisco, Seattle, Washington
- Characteristics: This cluster includes geographically diverse and major cities. Anchorage and Honolulu are characterized by significantly higher prices for all food items compared to other cities. This cluster represents cities with the highest cost of living for food.