Q1. (5 pts in total)

If $k = 2$, and the initial means are (10,1) and (10,30):

- Cluster-1: $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$, mean=(9.29, 9.57)(1 pt); Cluster-2: $\{x_8\}$, mean=(10, 30). (1 pt)

If $k = 3$ and the initial means are (10,1), (10,30), and (3,10)

- Cluster-1: $\{x_1, x_3, x_5, x_7\}$, mean=(14.5, 9)(1 pt); Cluster-2: $\{x_8\}$, mean=(10, 30)(1 pt); Cluster-3: $\{x_2, x_4, x_6\}$, mean=(2.33, 10.33).(1 pt)

Q2. (5 pts in total)

Advantages: easy to implement and understand. (1 pt)

Disadvantages:

(a) It is difficult to determine the value of k as the number of clusters is not known. (1 pt) Possible fix: we can try different value of k and adopt the most suitable one. (1 pt)

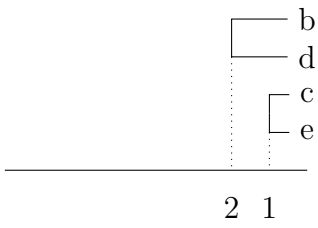
(b) It is sensitive to the initial guess of means. (1 pt) Possible fix: we can try different sets of initial guesses. (1 pt)

Q3. (10 pts in total)

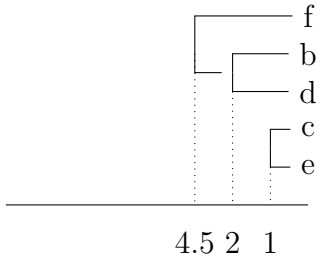
We group data points *w.r.t* the group average linkage:

	a	b	(ce)	d	f	g	h
a	0						
b	11	0					
(ce)	6	15	0				
d	12	2	16	0			
f	13	4	17.5	5	0		
g	9	15	13.5	16	19	0	
h	11	20	14.5	21	22	30	0

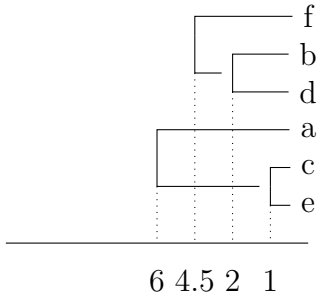
	a	(bd)	(ce)	f	g	h
a	0					
(bd)	11.5	0				
(ce)	6	15.5	0			
f	13	4.5	17.5	0		
g	9	15.5	13.5	19	0	
h	11	20.5	14.5	22	30	0



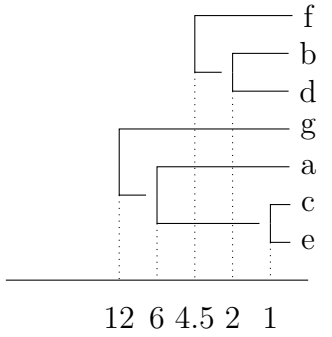
$$\begin{array}{c} \\ a \\ (bdf) \\ (ce) \\ g \\ h \end{array} \begin{array}{c} \begin{array}{ccccc} a & (bdf) & (ce) & g & h \end{array} \\ \begin{pmatrix} 0 & & & & \\ 12 & 0 & & & \\ 6 & 16.17 & 0 & & \\ 9 & 16.67 & 13.5 & 0 & \\ 11 & 21 & 14.5 & 30 & 0 \end{pmatrix} \end{array}$$



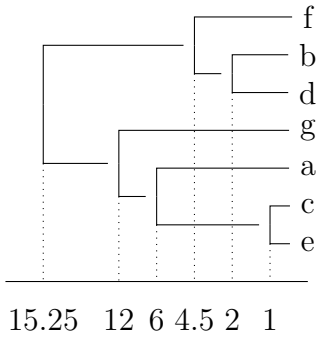
$$\begin{array}{c} \\ (ace) \\ (bdf) \\ g \\ h \end{array} \begin{array}{c} \begin{array}{ccccc} (ace) & (bdf) & g & h \end{array} \\ \begin{pmatrix} 0 & & & \\ 14.78 & 0 & & \\ 12 & 16.67 & 0 & \\ 13.33 & 21 & 30 & 0 \end{pmatrix} \end{array}$$

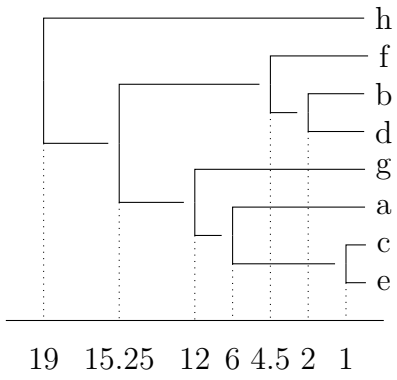


$$\begin{array}{c} \\ (aceg) \\ (bdf) \\ h \end{array} \begin{array}{c} \begin{array}{ccc} (aceg) & (bdf) & h \end{array} \\ \begin{pmatrix} 0 & & \\ 15.25 & 0 & \\ 17.5 & 21 & 0 \end{pmatrix} \end{array}$$



$$\begin{array}{c} \\ (abcdefg) \\ h \end{array} \begin{array}{c} \begin{array}{cc} (abcdefg) & h \end{array} \\ \begin{pmatrix} 0 & \\ 19 & 0 \end{pmatrix} \end{array}$$





$(abcdefgh)$

According to the above dendrogram, the 5 clusters are:

- Cluster 1: $\{b, d, f\}$;
- Cluster 2: $\{c, e\}$;
- Cluster 3: $\{g\}$;
- Cluster 4: $\{h\}$; and
- Cluster 5: $\{a\}$

Suggested Marking Scheme: Students do not have to show the matrix update at each step, but they need to show a correct dendrogram and distance metric in the diagram.

- 7 pts for correct final dendrogram
- Deduct 2 pts if the distance metric of dendrogram is incorrect (at most 2 points will be deducted for this type of error)
- Deduct 1 pt if a cluster output is incorrect (at most 3 points will be deducted for this type of error)