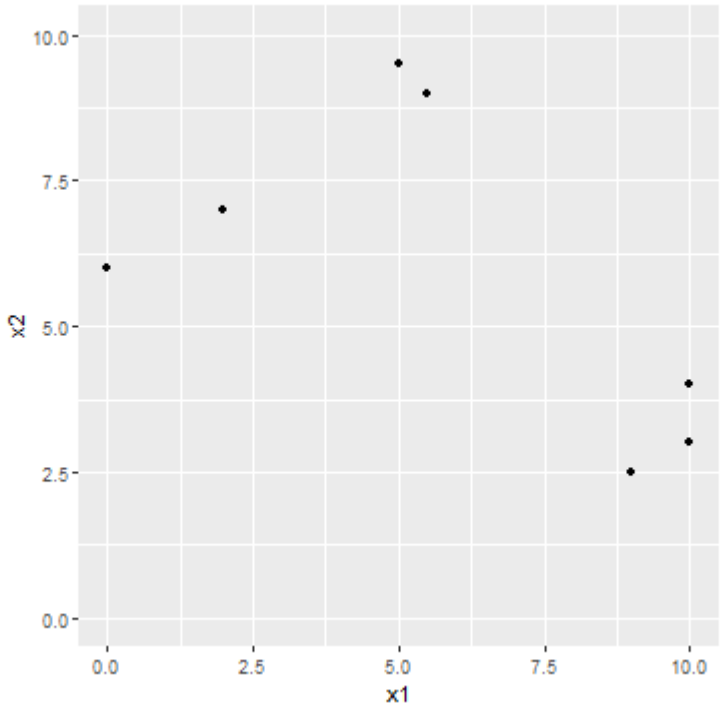
Hierarchical Clustering

Son Nguyen

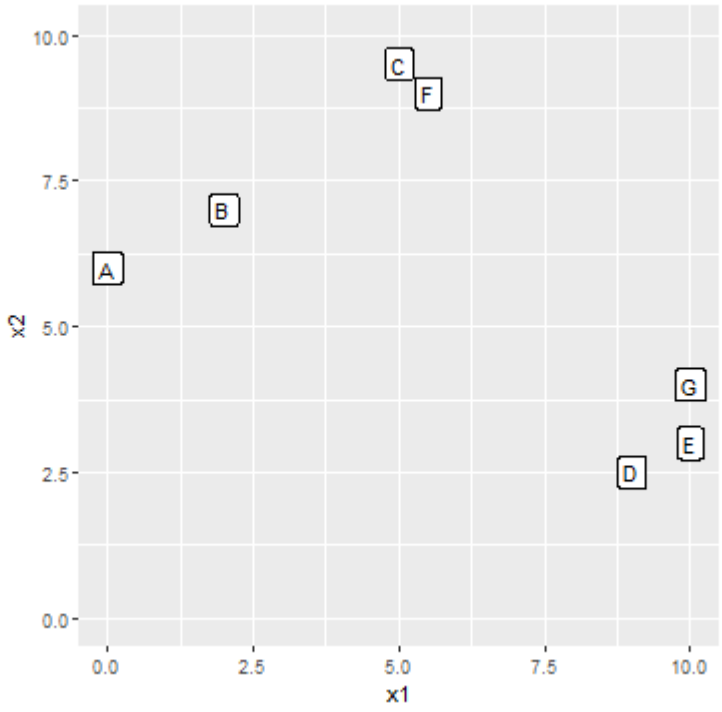
Hierarchical clustering - Centroid Linkage

cluster	x1	x2
A	0.0	6.0
B	2.0	7.0
C	5.0	9.5
D	9.0	2.5
E	10.0	3.0
F	5.5	9.0
G	10.0	4.0

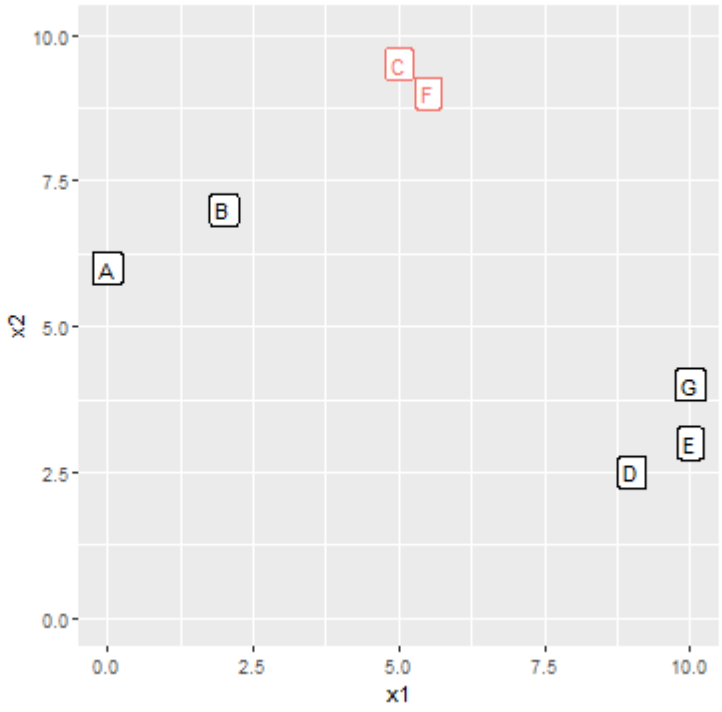
Hierarchical clustering - Centroid Linkage



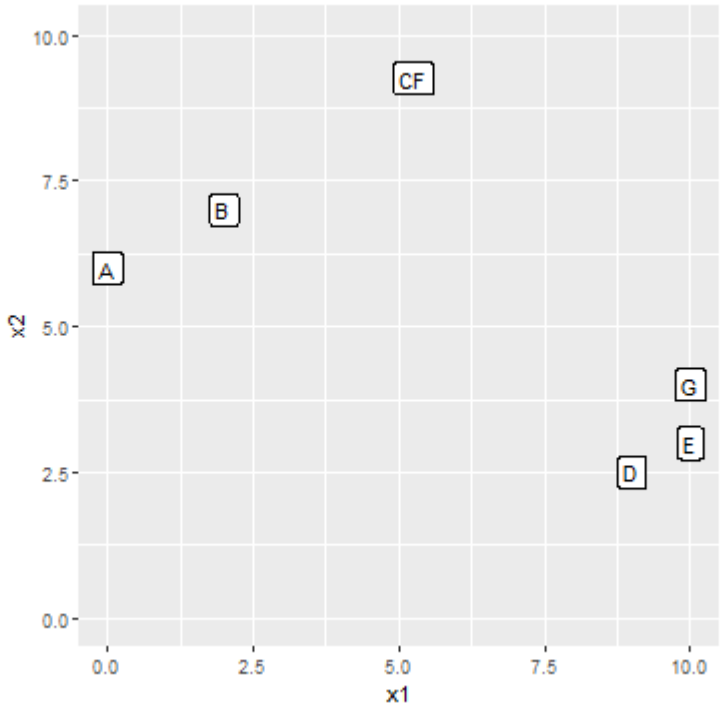
Label the Points



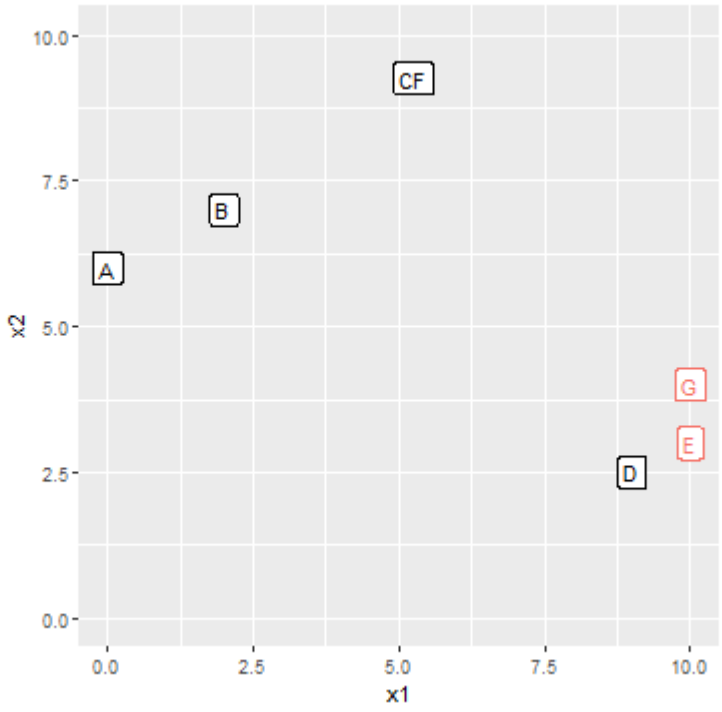
Pair with the smallest distance.



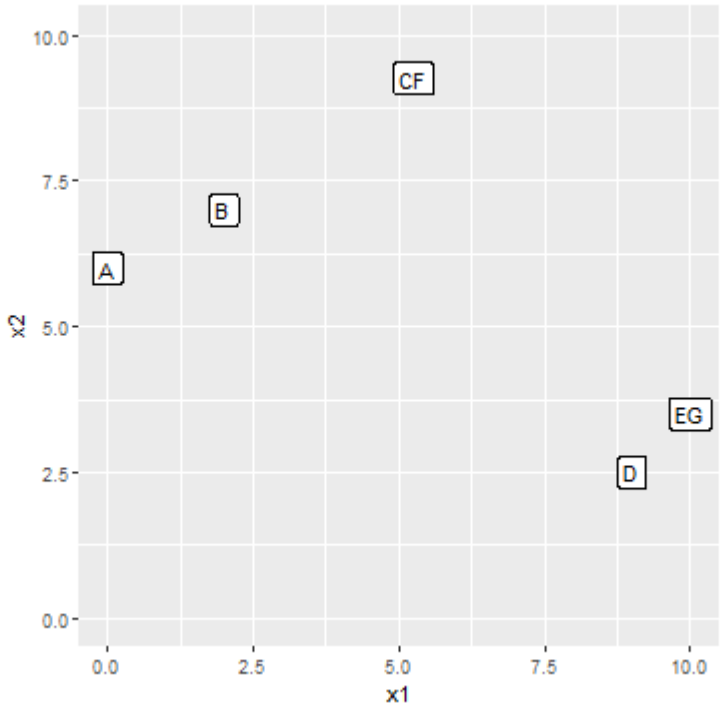
Group the pair



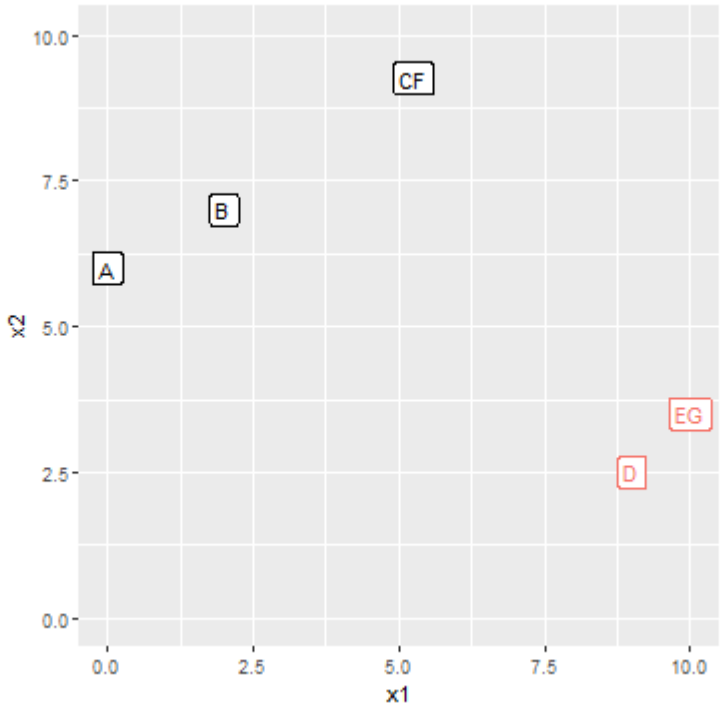
Pair with the smallest distance.



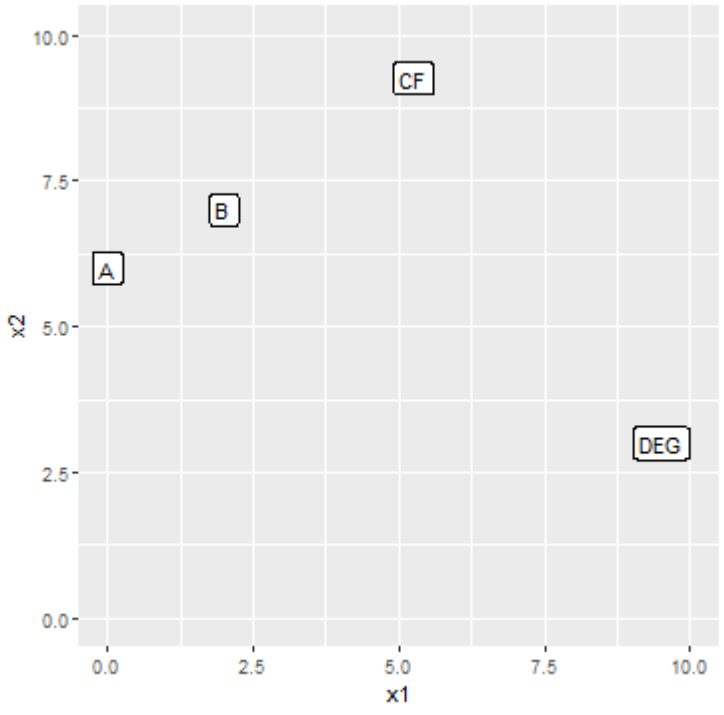
Group the pair



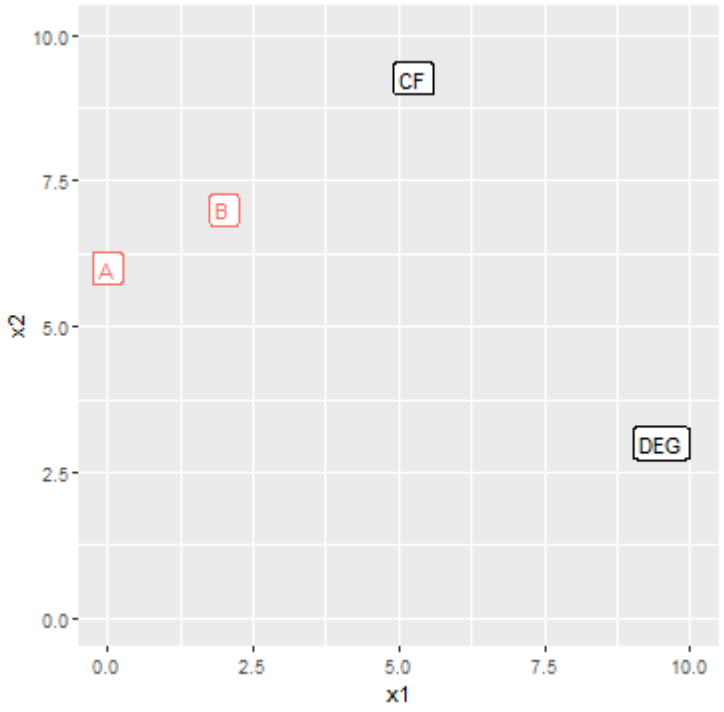
Pair with the smallest distance.



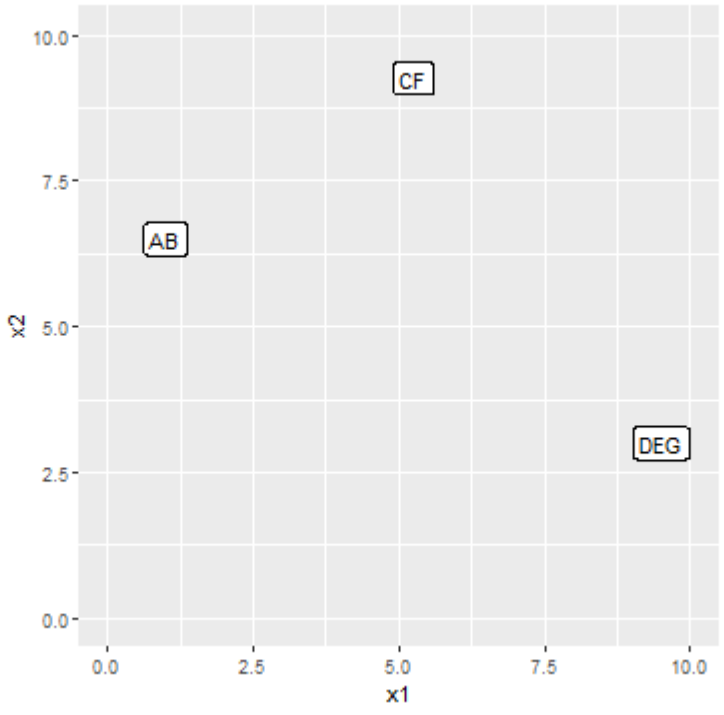
Group the pair



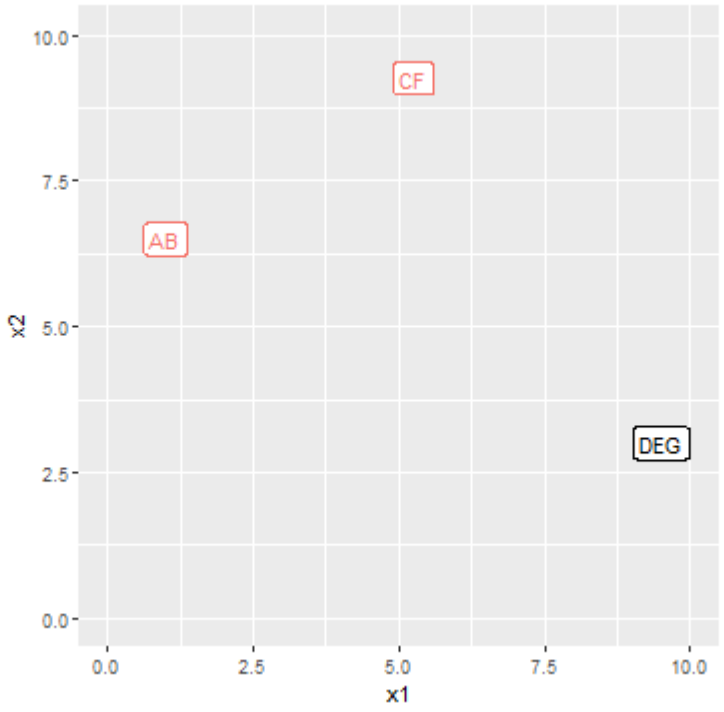
Pair with the smallest distance.



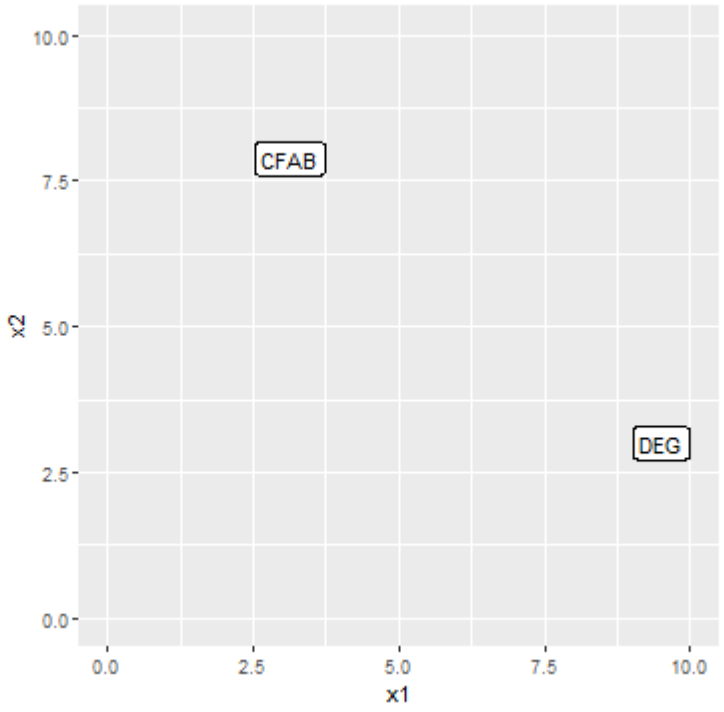
Group the pair



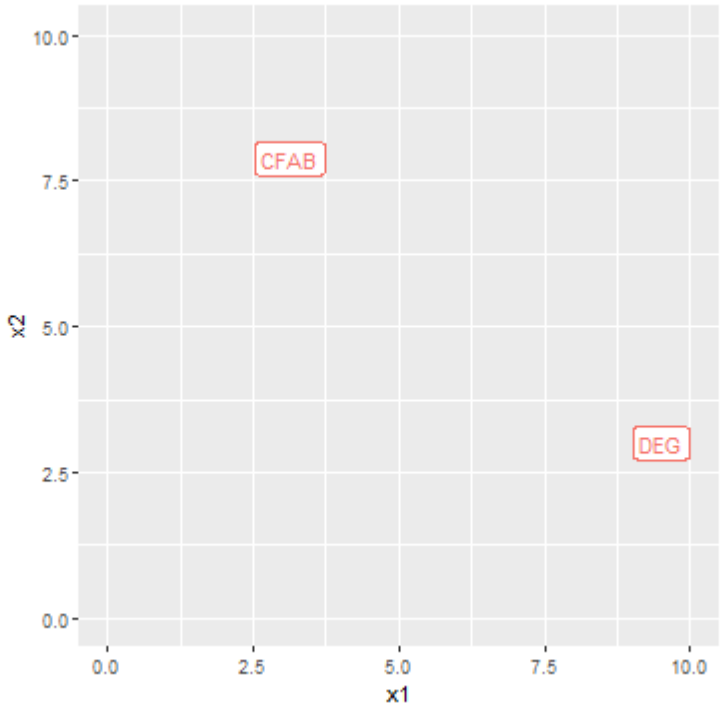
Pair with the smallest distance.



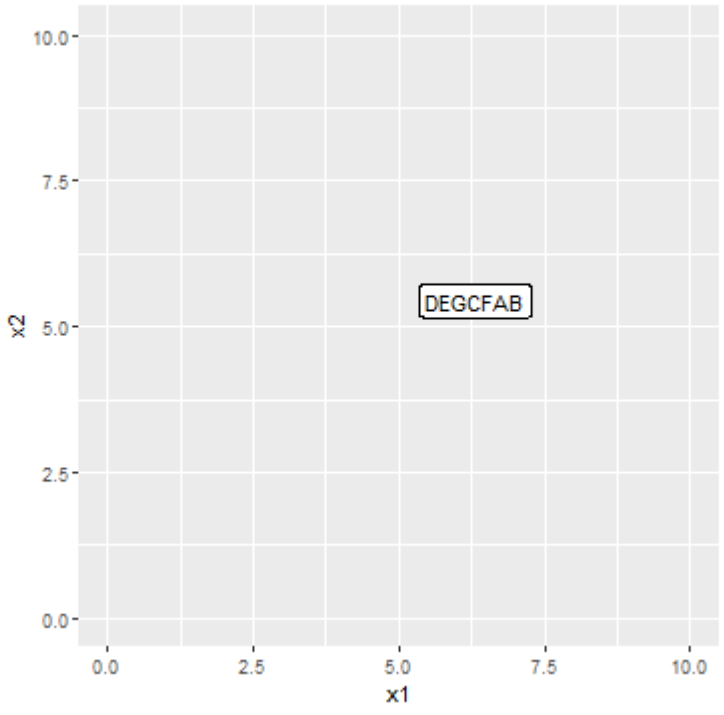
Group the pair



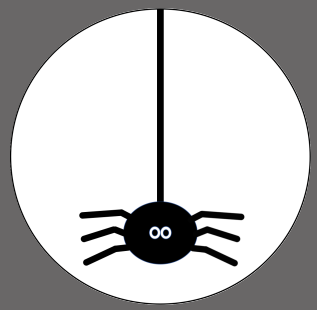
Pair with the smallest distance.



Group the pair



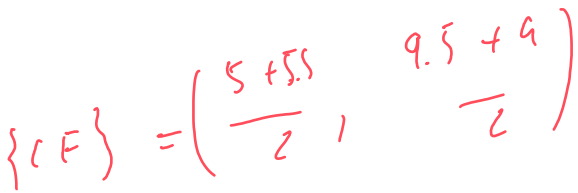
Detail Calculation



Detail Calculation

- The original data

cluster	x1	x2
A	0.0	6.0
B	2.0	7.0
C	5.0	9.5
D	9.0	2.5
E	10.0	3.0
F	5.5	9.0
G	10.0	4.0



Detail Calculation

- Calculate all the possible pair distances

	A	B	C	D	E	F	G
A	0.00	2.24	6.10	9.66	10.44	6.26	10.20
B	2.24	0.00	3.91	8.32	8.94	4.03	8.54
C	6.10	3.91	0.00	8.06	8.20	0.71	7.43
D	9.66	8.32	8.06	0.00	1.12	7.38	1.80
E	10.44	8.94	8.20	1.12	0.00	7.50	1.00
F	6.26	4.03	0.71	7.38	7.50	0.00	6.73
G	10.20	8.54	7.43	1.80	1.00	6.73	0.00

- Example: $AB = \sqrt{(0-2)^2 + (6-7)^2} = 2.24$
- The minimum distance is 0.71
- CF has the minimum distance
- Thus, group CF into one cluster

Detail Calculation

- The updated data after merging C and F

cluster	x1	x2
A	0.00	6.00
B	2.00	7.00
D	9.00	2.50
E	10.00	3.00
G	10.00	4.00
CF	5.25	9.25

- Notice that $CF = \left(\frac{5+5.5}{2}, \frac{9.5+9.0}{2}\right) = (5.25, 9.25)$

Detail Calculation

- Calculate all the possible pair distances

	A	B	D	E	G	CF
A	0.00	2.24	9.66	10.44	10.20	6.17
B	2.24	0.00	8.32	8.94	8.54	3.95
D	9.66	8.32	0.00	1.12	1.80	7.72
E	10.44	8.94	1.12	0.00	1.00	7.85
G	10.20	8.54	1.80	1.00	0.00	7.08
CF	6.17	3.95	7.72	7.85	7.08	0.00

- The minimum distance is 1
- EG has the minimum distance
- Thus, group EG together

Detail Calculation

- The updated data after merging EG together

cluster	x1	x2
A	0.00	6.00
B	2.00	7.00
D	9.00	2.50
CF	5.25	9.25
EG	10.00	3.50

Detail Calculation

- Calculate all the possible pair distances

	A	B	D	CF	EG
A	0.00	2.24	9.66	6.17	10.31
B	2.24	0.00	8.32	3.95	8.73
D	9.66	8.32	0.00	7.72	1.41
CF	6.17	3.95	7.72	0.00	7.46
EG	10.31	8.73	1.41	7.46	0.00

- The minimum distance is 1.41
- D - EG has the minimum distance
- Thus, group D-EG together

Detail Calculation

- The updated data after merging D-EG together

cluster	x1	x2
A	0.00	6.00
B	2.00	7.00
CF	5.25	9.25
DEG	9.50	3.00

Detail Calculation

- Calculate all the possible pair distances

	A	B	CF	DEG
A	0.00	2.24	6.17	9.96
B	2.24	0.00	3.95	8.50
CF	6.17	3.95	0.00	7.56
DEG	9.96	8.50	7.56	0.00

- The minimum distance is 2.24
- AB has the minimum distance
- Thus, group AB together

Detail Calculation

- The updated data after merging AB together

cluster	x1	x2
CF	5.25	9.25
DEG	9.50	3.00
AB	1.00	6.50

Detail Calculation

- Calculate all the possible pair distances

	CF	DEG	AB
CF	0.00	7.56	5.06
DEG	7.56	0.00	9.19
AB	5.06	9.19	0.00

- The minimum distance is 5.06
- CF-AB has the minimum distance
- Thus, group CF-AB together

Detail Calculation

- The updated data after merging CF-AB together

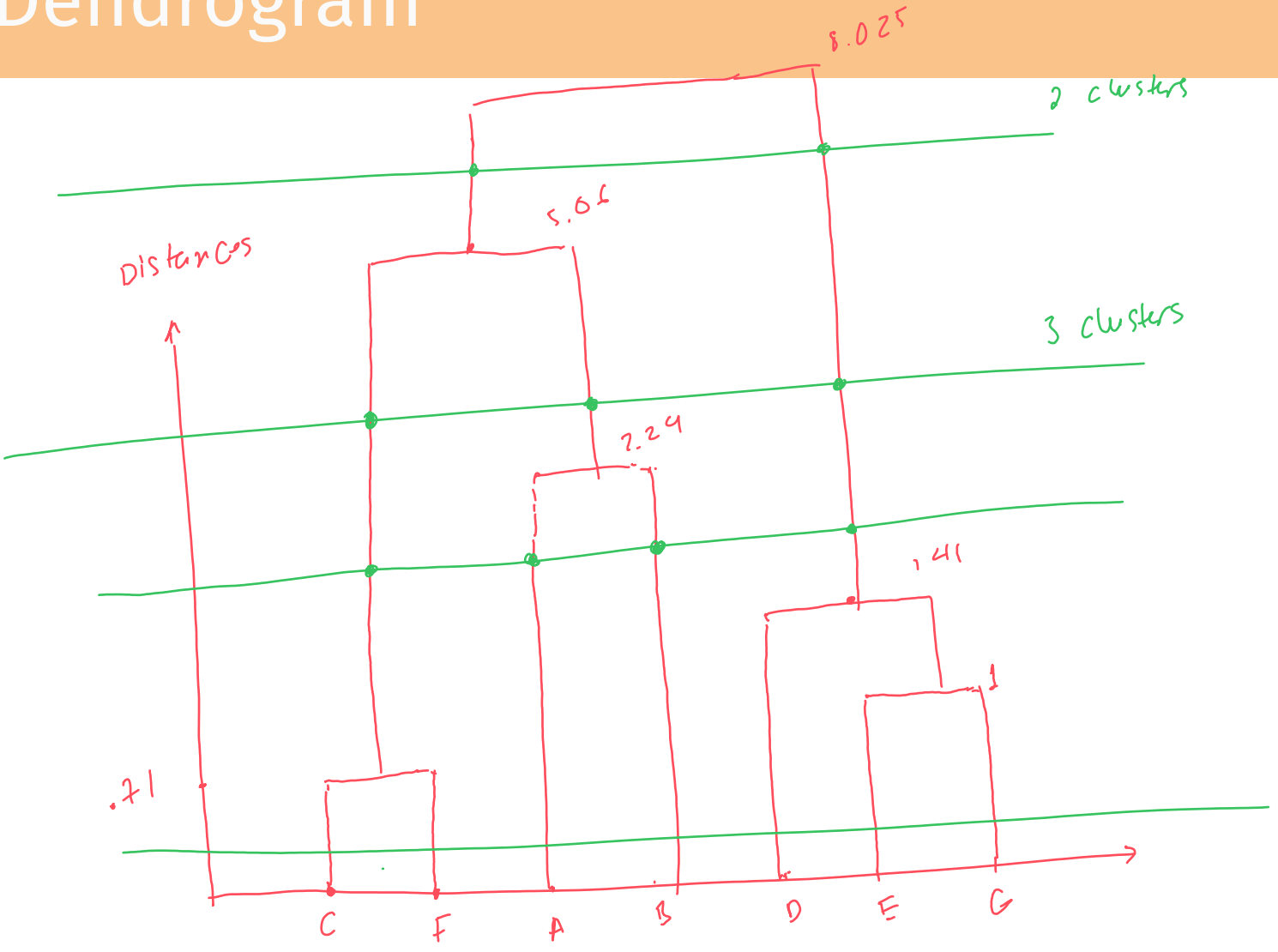
cluster	x1	x2
DEG	9.500	3.000
CFAB	3.125	7.875

Detail Calculation

- There are only two clusters left
- Just need to group them together

cluster	x1	x2
DEGCFAB	6.3125	5.4375

Dendrogram


8.025
2 clusters
5.05
Distances
3 clusters
2.24
1.41
1
.71
C
F
A
B
D
E
G

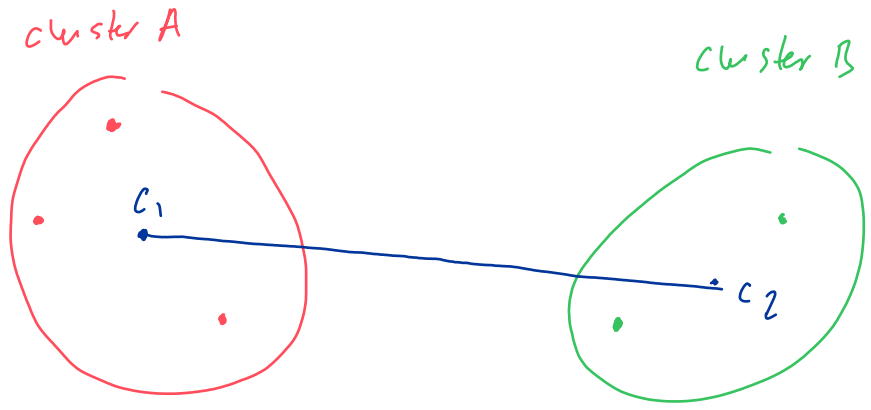
Problem

Point	x1	x2
A	0	6
B	2	7
C	5	9
D	9	2
E	1	3

- Calculate H-Clustering for the data and plot the Dendrogram

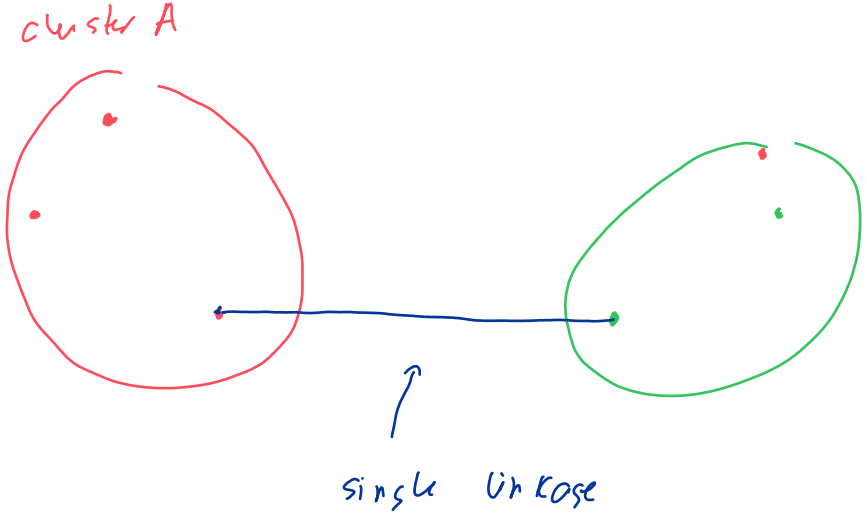
Linkages

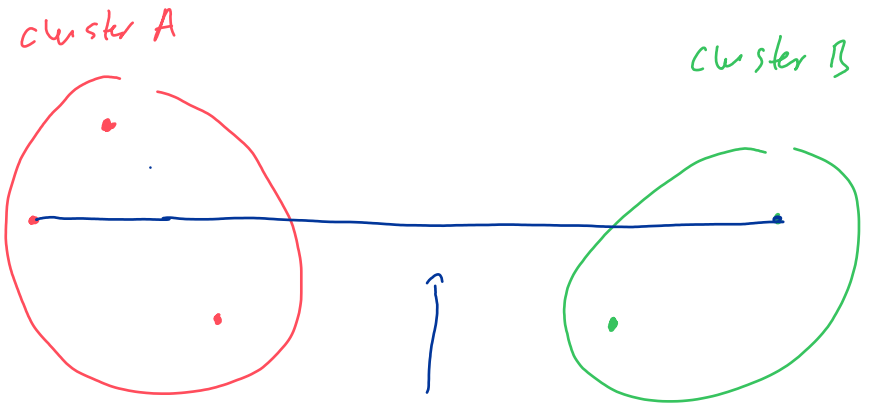
Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.



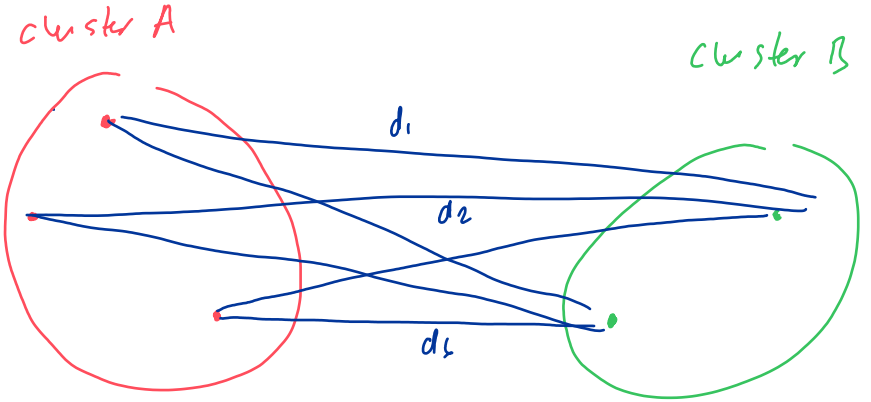
① Centroid Linkage : Distance (A, B) = $C_1 C_2$ cluster A

②





complete
Linkage



Average Linkage $= \frac{d_1 + \dots + d_L}{6}$