CLUSTERING

1. For the given data, compute two clusters using K-means algorithm for clustering where initial cluster centers are (1.0, 1.0) and (5.0, 7.0). Execute for two iterations.

Record Number	A	B
R1	1.0	1.0
R2	1.5	2.0
R3	3.0	4.0
R4	5.0	7.0
R5	3.5	5.0
R6	4.5	5.0
R7	3.5	4.5

Solution:

Initialization: Number of clusters (K) = 2, centroid for cluster1 (C1)= (1.0, 1.0) and centroid for cluster2 (C2) = (5.0, 7.0). We use **Euclidean distance** to find closest point to centroids.

Iteration1:

Record Number	**Close to C1**(1.0, 1.0)	**Close to C2**(5.0, 7.0)	**Assign to cluster**
R1(1.0,1.0)	dist(R1, C1)=0.0	dist(R1, C2)=7.21	Cluster1
R2(1.5,2.0)	dist(R2, C1)=1.12	dist(R2, C2)=6.12	Cluster1
R3(3.0,4.0)	dist(R3, C1)=3.61	dist(R3, C2)=3.61	Cluster1
R4(5.0,7.0)	dist(R4, C1)=7.21	dist(R4, C2)=0.0	Cluster2
R5(3.5,5.0)	dist(R5, C1)=4.12	dist(R5, C2)=2.5	Cluster2
R6(4.5,5.0)	dist(R6, C1)= 5.31	dist(R6, C2)=2.06	Cluster2
R7(3.5,4.5)	dist(R7,C1)=4.30	dist(R7, C2)=2.92	Cluster2

Thus, we obtain two clusters containing:

Cluster1 {R1, R2, R3} and Cluster2 {R4, R5, R6, R7}.

Their new centroids are:

C1 = (1.0+1.5+3.0)/3, (1.0+2.0+4.0)/3 = 5.5/3, 7.0/3 **= 1.83, 2.33**

C2 = (5.0+3.5+4.5+3.5)/4, (7+5+5+4.5)/4 = 16.5/4, 21.5/4 **= 4.12, 5.37**

Iteration2:

Record Number	**Close to C1(**1.83, 2.33**)**	**Close to C2(**4.12, 5.37**)**	**Assign to cluster**
R1(1.0,1.0)	dist(R1, C1)=1.57	dist(R1, C2)=5.37	Cluster1
R2(1.5,2.0)	dist(R2, C1)=0.47	dist(R2, C2)=4.27	Cluster1
R3(3.0,4.0)	dist(R3, C1)=2.04	dist(R3, C2)=1.77	Cluster2
R4(5.0,7.0)	dist(R4, C1)=5.64	dist(R4, C2)=1.85	Cluster2
R5(3.5,5.0)	dist(R5, C1)=3.15	dist(R5, C2)=0.72	Cluster2
R6(4.5,5.0)	dist(R6, C1)=3.78	dist(R6, C2)=0.53	Cluster2
R7(3.5,4.5)	dist(R7,C1)=2.74	dist(R7, C2)=1.07	Cluster2

Therefore, new clusters are:

Cluster1 {R1, R2} and Cluster2 {R3, R4, R5, R6, R7}.

Their new centroids are:

C1 = (1.0+1.5)/2, (1.0+2.0)/2 = 2.50/2,3.0/2 = **1.25,1.5**
C2 = (3.0+5.0+3.5+4.5+3.5)/5, (4+7+5+5+4.5)/5 = 19.5/5, 25.5/5 = **3.9, 5.1**

Final Result after Two Iterations

- **Cluster 1:** {R1, R2} with centroid (1.25, 1.5)
- **Cluster 2:** {R3, R4, R5, R6, R7} with centroid (3.9, 5.1)

2. Use the **distance matrix** in Table1 to perform single link and complete link hierarchical clustering. Show your results by drawing a dendogram. The dendogram should clearly show the order in which the points are merged.

Table 1 Distance matrix

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

Solution:

For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined as the minimum of the distance (maximum of the similarity) between any two points in the two different clusters.

Steps:

- Using graph terminology, start with all points as singleton clusters.
- Add links between points one at a time (shortest links first). - These single links combine the points into clusters.

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

Combine P1 and P2:

$$dist(\{P1, P2\}, \{P3\}) = min(dist(P1, P3), dist(P2, P3))$$
$$= min(0.41, 0.64)$$
$$= 0.41$$

$$dist(\{P1, P2\}, \{P4\}) = min(dist(P1, P4), dist(P2, P5))$$
$$= min(0.55, 0.98)$$
$$= 0.55$$

$$dist(\{P1, P2\}, \{P5\}) = min(dist(P1, P5), dist(P2, P5))$$
$$= min(0.35, 0.98)$$
$$= 0.35$$

	P12	P3	P4	P5
P12	0.00	0.41	0.55	0.35
P3	0.41	0.00	0.44	0.85
P4	0.55	0.44	0.00	0.76
P5	0.35	0.85	0.76	0.00

Combine P12 and P5:

$$dist(\{P12, P5\}, \{P3\}) = min(dist(P12, P3), dist(P5, P3))$$
$$= min(0.41, 0.85)$$

$$= 0.41$$

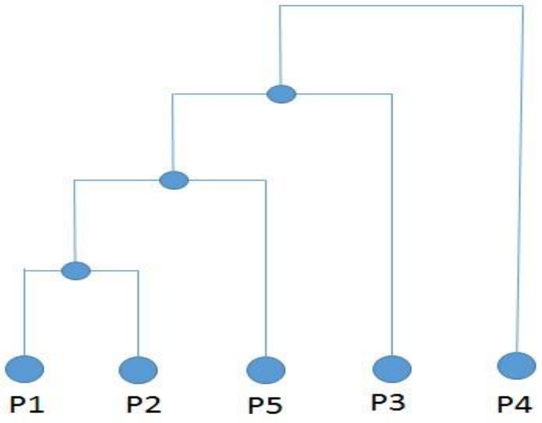
$$dist(\{P12, P5\}, \{P4\}) = min(dist(P12, P4), dist(P5, P4))$$
$$= min(0.55, 0.76)$$
$$= 0.55$$

	P125	P3	P4
P125	0.00	0.41	0.55
P3	0.41	0.00	0.44
P4	0.55	0.44	0.00

Combine P125 and P3:

$$dist(\{P125, P3\}, \{P4\}) = min(dist(P125, P4), dist(P3, P4))$$
$$= min(0.55, 0.44)$$
$$= 0.44$$

	P1235	P4
P1235	0.00	0.44
P4	0.44	0.00



Single Link Dendogram

For the complete link or MAX version of hierarchical clustering, the proximity of two clusters is defined as the maximum of the distance (minimum of the similarity) between any two points in the two different clusters.

Steps:

- Using graph terminology, start with all points as singleton clusters.
- Add links between points one at a time (shortest links first).
- Group points until all the points are completely linked, i.e., clique.

-	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

Combine P1 and P2:

$$dist(\{P1, P2\}, \{P3\}) = max(dist(P1, P3), dist(P2, P3))$$
$$= max(0.41, 0.64)$$
$$= 0.64$$

$$dist(\{P1, P2\}, \{P4\}) = min(dist(P1, P4), dist(P2, P5))$$
$$= min(0.55, 0.98)$$
$$= 0.98$$

$$dist(\{P1, P2\}, \{P5\}) = min(dist(P1, P5), dist(P2, P5))$$
$$= min(0.35, 0.98)$$
$$= 0.98$$

	P12	P3	P4	P5
P12	0.00	0.64	0.98	0.98
P3	0.64	0.00	0.44	0.85
P4	0.98	0.44	0.00	0.76
P5	0.98	0.85	0.76	0.00

Combine P3 and P4:

$$dist(\{P3, P4\}, \{P12\}) = min(dist(P3, P12), dist(P4, P12))$$
$$= max(0.64, 0.98)$$
$$= 0.98$$

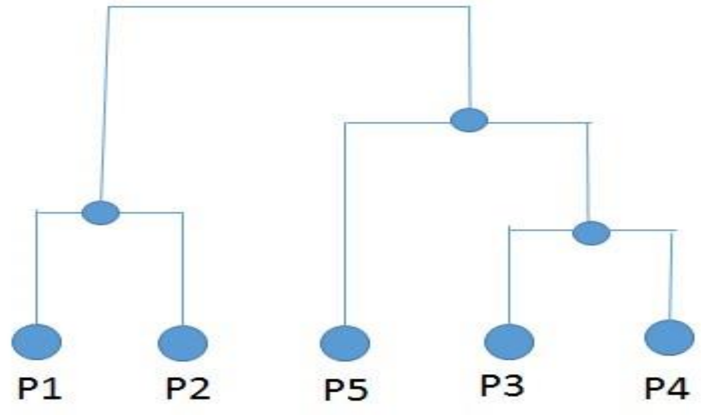
$$dist(\{P3, P4\}, \{P5\}) = min(dist(P3, P5), dist(P4, P5))$$
$$= max(0.85, 0.76)$$
$$= 0.85$$

	P12	P34	P5
P12	0.00	0.98	0.98
P34	0.98	0.00	0.85
P5	0.98	0.85	0.00

Combine P34 and P5:

$$dist(\{P34, P5\}, \{P12\}) = max(dist(P34, P12), dist(P5, P12))$$
$$= max(0.98, 0.98)$$
$$= 0.98$$

	P12	P345
P12	0.00	0.98
P345	0.98	0.00



Complete Link Dendogram