

## **Data Clustering:**

**Q1)** Consider the following points (p1-p6) whose coordinates are given below which belong to a two-dimensional vector space. The dataset containing these points will be clustered into 2 clusters using the ***K-means (incremental approach)***. P1 and p2 are randomly chosen as the initial centroids. The  $L_1$  norm will be used as the “closeness” measure to determine the centroid each point is “close” to.

P1	P2	P3	P4	P5	P6
(0,3,3)	(0,4,0)	(0,5,8)	(0,6,4)	(1,0,5)	(2,5,6)

A) Find the clusters each point belongs to after the first iteration of the algorithm. Show your work clearly for both cases. (15 points)

B) Calculate the Sum of Squared Error (SSE) for the three clusters found in part (A). (7 pts)  
Note that the centroid of each cluster is found by calculating the average of all points that belong to that cluster.

**Q1)** Consider the following points (p1-p6) whose coordinates are given below which belong to a two-dimensional vector space.

A) Find the clusters each point belongs to using the **single link technique (min)**. Show your work clearly. (17 pts)

P1	P2	P3	P4	P5	P6
(0,3,3)	(0,4,0)	(0,5,8)	(0,6,4)	(1,0,5)	(2,5,6)

B) Draw the clusters dendrogram. (5 pts)