Assignment 3

1. Consider the following dataset:

(2, 6)
(3, 4)
(3, 8)
(4, 7)
(6, 2)
(6, 4)
(7, 3)
(7, 4)
(8, 5)
(7, 6)

   a. Perform K-Means Clustering where $K = 3$
   b. Perform K-Medoid Clustering where $K = 2$, $C_1 = x_2$ and $C_2 = x_8$

2. Given 2 objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
   a) Compute the Euclidean distance between two objects
   b) Compute the Manhattan distance between two objects
   c) Compute the Minkowski distance between two objects where $q = 3$

3. Construct the dendrogram, draw the nested clusters and show all the steps for single link and complete link hierarchical clustering for the distance matrix given below:

P1	(1, 1)
P2	(1.5, 1.5)
P3	(5, 5)
P4	(3, 4)
P5	(4, 4)
P6	(3, 3.5)

4. Explain Hadoop ecosystem with a diagram
5. How does MapReduce help in Big Data?

ASSIGNMENT 3

1.

a)

Data	distance from $m_1=(2,6)$	distance from $m_2=(3,4)$	distance from $m_3=(3,8)$	Cluster
(2,6)	0	2.236	2.236	C1
(3,4)	2.236	0	4	C2
(3,8)	2.236	4	0	C3
(4,7)	2.236	3.162	1.414	C3
(6,2)	5.657	3.606	6.708	C2
(6,4)	4.472	3	5	C2
(7,3)	5.831	4.123	6.403	C2
(7,4)	5.385	4	5.657	C2
(8,5)	6.083	5.099	5.831	C2
(7,6)	5	4.472	4.472	C2

$m_1=(2,6)$

$m_2=\left(\frac{3+6+6+7+7+8+7}{7}, \frac{4+2+4+3+4+5+6}{7}\right)$

$=\frac{44}{7}, \frac{28}{7}=(6.286, 4)$

$m_3=\left(\frac{3+4}{2}, \frac{8+7}{2}\right)=(3.5, 7.5)$

Data	distance from $m_1=(2,6)$	distance from $m_2=(6.286,4)$	distance from $m_3=(3.5,7.5)$	Cluster
(2,6)	0	4.729	2.121	C1
(3,4)	2.236	3.286	3.536	C1
(3,8)	2.236	5.177	0.707	C3
(4,7)	2.236	3.772	0.707	C3
(6,2)	5.657	2.02	6.042	C2
(6,4)	4.472	0.286	4.301	C2
(7,3)	5.831	1.229	5.7	C2
(7,4)	5.385	0.714	4.949	C2
(8,5)	6.083	1.984	5.148	C2
(7,6)	5	2.124	3.808	C2

$m_1=\left(\frac{2+3}{2}, \frac{6+4}{2}\right)=(2.5, 5)$

$m_2=\left(\frac{6+6+7+7+8+7}{6}, \frac{2+4+3+4+5+6}{2}\right)$

$=\left(\frac{41}{6}, \frac{24}{6}\right)=(6.83, 4)$

$m_3=(3.5, 7.5)$

Data	distance from $m_1 = (2.5, 5)$	distance from $m_2 = (6.83, 4)$	distance from $m_3 = (3.5, 7.5)$	Cluster
(2,6)	1.118	5.228	2.121	C1
(3,4)	1.118	3.83	3.536	C1
(3,8)	3.041	5.538	0.707	C3
(4,7)	2.5	4.124	0.707	C3
(6,2)	4.609	2.165	6.042	C2
(6,4)	3.64	0.83	4.301	C2
(7,3)	4.924	1.014	5.7	C2
(7,4)	4.609	0.17	4.949	C2
(8,5)	5.5	1.539	5.148	C2
(7,6)	4.609	2.007	3.808	C2

∵ no change in cluster assignment

C1 : (2,6), (3,4)

C2 : (6,2), (6,4) (7,3) (7,4) (8,5) (7,6)

C3 : (3,8), (4,7)

b)

Data	distance from $m_1 = (2,6)$	distance from $m_2 = (3,4)$	Cluster
(2,6)	–	–	C1
(3,4)	~~2.236~~ –	–	C2
(3,8)	2.236	4	C1
(4,7)	2.236	3.162	C1
(6,2)	5.657	3.606	C2
(6,4)	4.472	3	C2
(7,3)	5.831	4.123	C2
(7,4)	5.385	4	C2
(8,5)	6.083	5.099	C2
(7,6)	5	4.472	C2

~~cost = (4 + 3.162)~~

+

cost = (2.236 + 2.236)

+ (3.606 + 3 + 4.123 + 4

+ 5.099 + 4.472)

= 4.472 + 24.3[illegible]

Lets choose new medoid $m_{\blacksquare} = (3,8)$

Data	distance from $m_1=(3,8)$	distance from $m_2=(3,4)$	Cluster
(2,6)	2.236	2.236	C1
(3,4)	—	—	C2
(3,8)	—	—	C1
(4,7)	1.414	3.162	C1
(6,2)	6.708	3.606	C2
(6,4)	5	3	C2
(7,3)	6.403	4.123	C2
(7,4)	5.657	4	C2
(8,5)	5.831	5.099	C2
(7,6)	4.472	4.472	C2

Cost = (2.236 + 1.414 ~~+ 3.162~~)
+ (3.606 + 3 + 4.123 + 4
+ 5.099 + 4.472)
= 3.65 + 24.3 = 27.95

∵ it's less than the previous cost, we choose $m_1=(4,7)$

Data	distance from $m_1=(4,7)$	distance from $m_2=(3,4)$	cluster
(2,6)	2.236	2.236	C1
(3,4)	—	—	C2
(3,8)	1.414	4	C1
(4,7)	—	—	C1
(6,2)	5.385	3.606	C2
(6,4)	3.606	3	C2
(7,3)	5	4.123	C2
(7,4)	4.243	4	C2
(8,5)	4.472	5.099	C1
(7,6)	3.162	4.472	C1

Cost = (2.236 + 1.414 + 3.162)
+ (3.606 + 3 + 4.123 + 4)
= 6.812 + 14.729
= 21.541

Repeat till cost is greater than previous iteration

2. a) Euclidean distance between (x_1, y_1) and (x_2, y_2)

$$= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

2 a) Euclidean distance between $(x_1, x_2, \ldots x_n)$ and $(y_1, y_2, \ldots y_n)$

$$= \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}$$

∴ Euclidean distance between (22, 1, 42, 10) and (22, 0, 36, 8)

$$= \sqrt{(22-22)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{0+1+36+4} = \sqrt{41} =$$

b) Manhattan distance between $(x_1, x_2, \ldots x_n)$ and $(y_1, y_2, \ldots y_n)$

$$= \sum_{i=1}^{n} |x_i - y_i|$$

∴ Manhattan distance between (22, 1, 42, 10) and (22, 0, 36, 8)

$$= |22-22| + |1-0| + |42-36| + |10-8| = 9$$

c) Minkowski distance between ~~(22, 1, 42, 0)~~ $(x_1, x_2 \ldots x_n)$ and $(y_1, y_2, \ldots y_n)$

when $q = 3 = \left(\sum_{i=1}^{n} |x_i - y_i|^3 \right)^{1/3}$

∴ Minkowski distance between (22, 1, 42, 10) and (22, 0, 36, 8) when $q = 3$

$$= \sqrt[3]{(|22-22|)^3 + (|1-0|)^3 + (|42-36|)^3 + (|10-8|)^3} = \sqrt[3]{0+1+216+8}$$

$$= 15$$

3 a) ~~Single link hierarchical clustering~~

Euclidean distance matrix

	P1 (1,1)	P2 (1.5,1.5)	P3 (5,5)	P4 (3,4)	P5 (4,4)	P6 (3,3.5)
P1 (1,1)	—	0.707	5.657	3.606	4.243	3.201
P2 (1.5,1.5)	0.707	—	4.949	2.915	3.536	2.5
P3 (5,5)	5.657	4.949	—	2.236	1.414	2.5
P4 (3,4)	3.606	2.915	2.236	—	1	0.5
P5 (4,4)	4.243	3.536	1.414	1	—	1.118
P6 (3,3.5)	3.201	2.5	2.5	0.5	1.118	—

a) Single link clustering

minimum distance between P4 & P6

	P1	P2	P3	P46	P5
P1	–	0.707	5.657	3.201	4.243
P2	0.707	–	4.949	2.5	3.536
P3	5.657	4.949	–	2.236	1.414
P46	3.201	2.5	2.236	–	1
P5	4.243	3.536	1.414	1	–

minimum distance is between P1 and P2

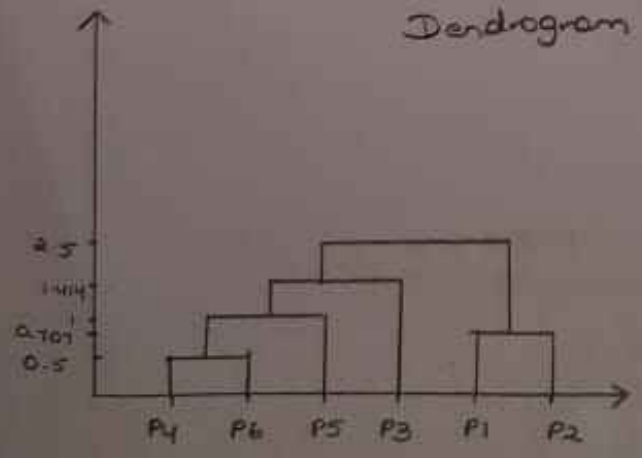
	P12	P3	P46	P5
P12	–	4.949	2.5	3.536
P3	4.949	–	2.236	1.414
P46	2.5	2.236	–	1
P5	3.536	1.414	1	–

minimum distance is between P46 & P5

	P12	P3	P465
P12	–	4.949	2.5
P3	4.949	–	1.414
P465	2.5	1.414	–

minimum distance is between P3 & P465

	P12	P3465
P12	–	2.5
P3465	2.5	–



b) Complete link hierarchical clustering

From Euclidean distance matrix, we know that the minimum distance is between P4 & P6

	P1	P2	P3	P46	P5
P1	–	0.707	5.657	3.606	4.243
P2	0.707	–	4.949	2.915	3.536
P3	5.657	4.949	–	2.5	1.414
P46	3.606	2.915	2.5	–	1.118
P5	4.243	3.536	1.414	1.118	–

minimum distance is between P1 & P2

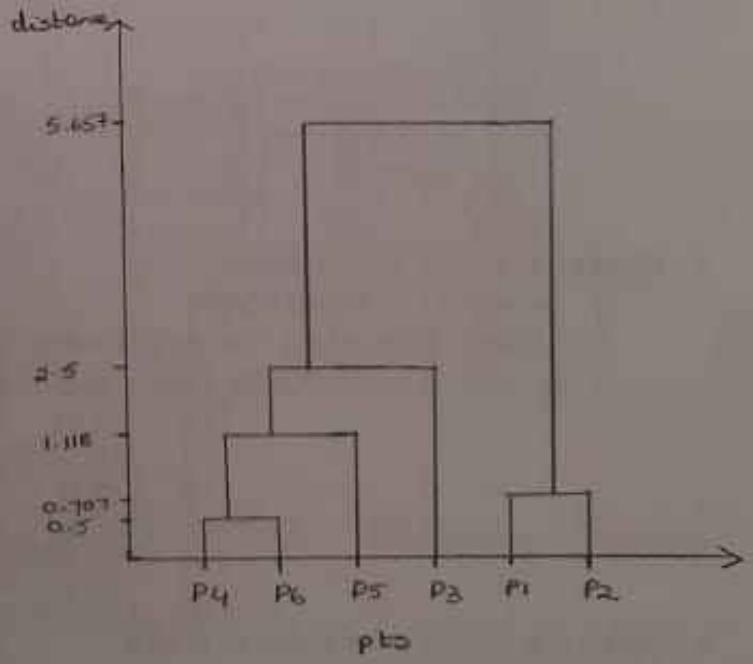
	P12	P3	P46	P5
P12	–	5.657	3.606	4.243
P3	5.657	–	2.5	1.414
P46	3.606	2.5	–	1.118
P5	4.243	1.414	1.118	–

minimum distance is between P46 & P5

	P12	P3	P465
P12	–	5.657	4.243
P3	5.657	–	2.5
P465	4.243	2.5	–

minimum distance is between P3 & P465

	P12	P3465
P12	–	5.657
P3465	5.657	–



pto