Part A

Q.no. 1

Q.1.

Given:

$$w(c) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \in C_k} \| x_i - x_{i'} \|^2$$

To prove:

$$w(c) = \sum_{k=1}^{K} n_k \sum_{i \in C_k} \| x_i - \bar{x}_k \|^2$$

Proof: (Adding and Subtracting the mean and then viewing the norm as ~~the~~ an inner product, we get:)

$$w(c) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \in C_k} \| x_i - \bar{x}_k - (x_{i'} - \bar{x}_k) \|^2$$

$$= \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \in C_k} \left(\| x_i - \bar{x}_k \|^2 + \| x_{i'} - \bar{x}_k \|^2 - 2 \langle x_i - \bar{x}_k, x_{i'} - \bar{x}_k \rangle \right)$$

$$= \frac{1}{2} \sum_{k=1}^{K} \Big(N_k \sum_{i \in C_k} \| x_i - \bar{x}_k \|^2 + N_k \sum_{i' \in C_k} \| x_{i'} - \bar{x}_k \|^2$$

$$- 2 \Big\langle \sum_{i' \in C_k} (x_{i'} - \bar{x}_k), \sum_{i' \in C_k} (x_{i'} - \bar{x}_k) \Big\rangle \Big)$$

$$= \sum_{k=1}^{K} N_k \sum_{i \in C_k} \| x_i - \bar{x}_k \|^2$$

Proved

Q.no. 2. A

Q.2.

a. Single linkage hierarchial procedure
Taking the minimum distance as the recalibrated distance in every single step.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Link between 3 and 5 with distance 2.

	1	2	3,5	4
1	0			
2	9	0		
3,5	3	7	0	
4	6	5	8	0

Link between 3,5 and 1 with distance 3

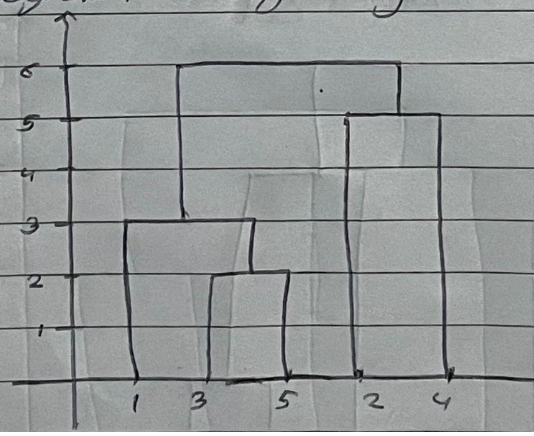
	1,3,5	2	4
1,3,5	0		
2	7	0	
4	6	5	0

Link between 2 and 4 with distance 5

	1,3,5	2,4
1,3,5	0	
2,4	6	0

Link between 1,3,5 and 2,4 with distance 6

Dendogram: Single linkage



Q.no. 2. B

b. Complete linkage hierarchial procedure

Taking the maximum distance as the recalibrated distance in every single step.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Link between 3 and 5 with distance 2

	1	2	3,5	4
1	0			
2	9	0		
3,5	11	10	0	
4	6	5	9	0

Link between 2 and 4 with distance 5

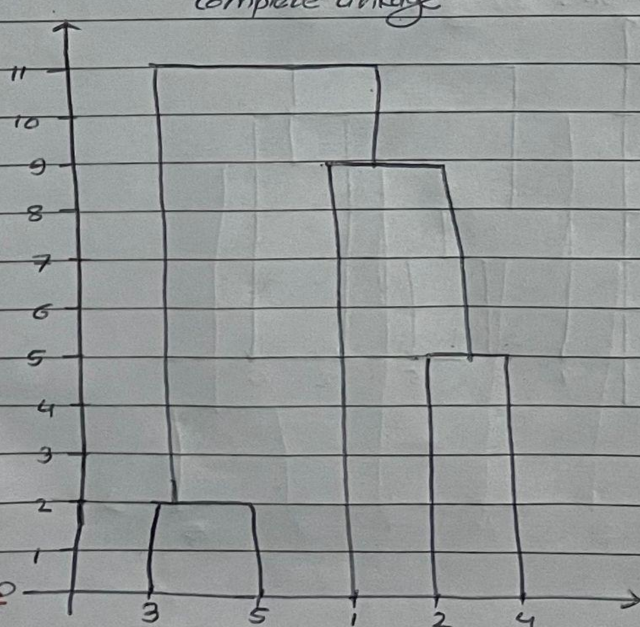
	1	2,4	3,5
1	0		
2,4	9	0	
3,5	11	10	0

Link between 2,4 and 1 with distance 9

	1,2,4	3,5
1,2,4	0	
3,5	11	0

Link between 1,2,4 and 3,5 with distance 11

Complete Linkage



Q.no. 2. C

c) Average linkage hierarchial procedure

Taking the average distance as the recalibrated distance in every single step.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Link between 3 and 5 with distance 2.

~~Avg(~~

	1	2	3,5	4
1	0			
2	9	0		
3,5	7	8.5	0	
4	6	5	8.5	0

$Avg((3,5),1) = \frac{3+11}{2} = \frac{14}{2} = 7$

$Avg((3,5),2) = \frac{7+10}{2} = 8.5$

$Avg((3,5),4) = \frac{9+8}{2} = 8.5$

Link between 2 and 4 with distance 5

	1	2,4	3,5
1	0		
2,4	7.5	0	
3,5	7	8.5	0

$Avg((2,4),1) = \frac{9+6}{2} = 7.5$

$Avg((2,4),(3,5)) = \frac{8.5+8.5}{2} = 8.5$

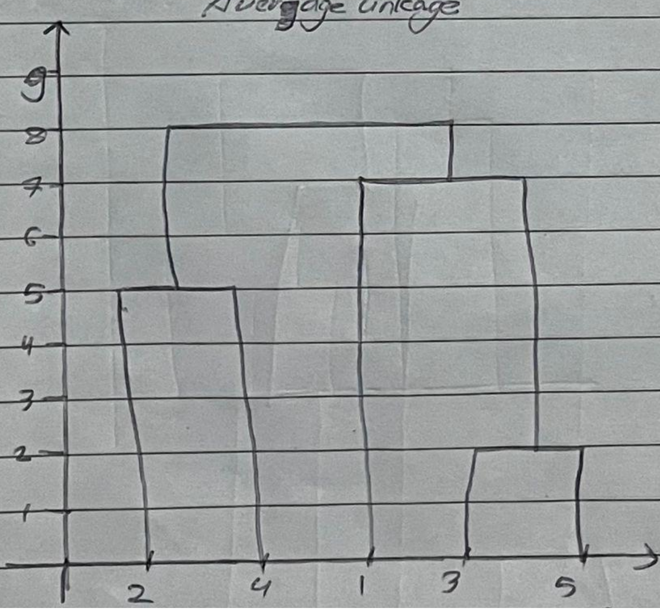
Link between 3,5 and 1 with distance 7

	1,3,5	2,4
1,3,5	0	
2,4	8	0

$Avg((1,3,5),(2,4)) = \frac{7.5+8.5}{2} = 8$

Link between 1,3,5 and 2,4 with distance 8

Average linkage



Due to a relatively small number of observations (only 5), the typical difference between dendrograms obtained from single linkage and the same from complete and/or average linkage - which is single linkage dendrogram yielding extended clusters to which single leaves are fused one by one and complete/average linkage dendrograms yielding evenly sized clusters - is not clearly observable in the above figures. However, it is evident that the single linkage dendrogram has the maximum distance as just 6 between clusters while the same for complete and average linkage dendrogram are relatively higher - 11 and 8 respectively. As such, at a distance of cut off 5.5, single linkage dendrogram leads to only two clusters and complete and average linkage dendrograms result in three clusters.