Group G20
Amal Sony (asony)
Mohd Sharique Khan (mkhan8)
Siddu Madhure Jayanna (smadhur)

1.a)

K-Means clustering with Euclidean distance and k =3.

Data points:

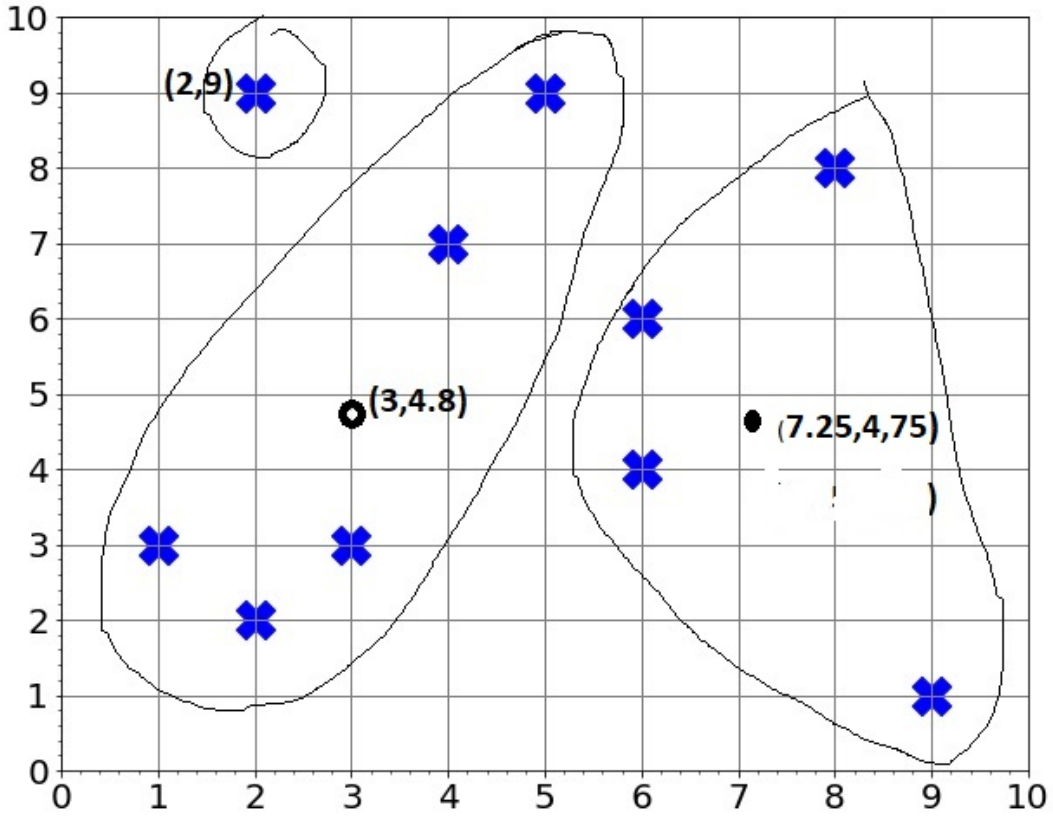
Point	X	Y
A	5	9
B	1	3
C	4	7
D	9	1
E	2	2
F	6	4
G	8	8
H	2	9
I	6	6
J	3	3

Round 1: Centroid C,H,I

	C	H	I	Cluster
A	2.24	3	3.16	C
B	5	6.08	5.83	C
C	0	2.82	2.23	C
D	7.81	10.63	5.83	I
E	5.38	7	5.65	C
F	3.60	6.40	2.0	I
G	4.12	6.08	2.82	I
H	2.82	0	5	H
I	2.24	5	0	I
J	4.12	6.08	4.24	C

C1 → A,B,C,E,J
C2 → D,F,G,I
C3 → H

New Centroids:
C1 → (5+1+4+2+3)/5 , (9+3+7+2+3)/5 →(3,4.8)
C2 →(9+6+8+6)/4 , (1+4+8+6)/4 → (7.25,4.75)
C3 → (2,9)



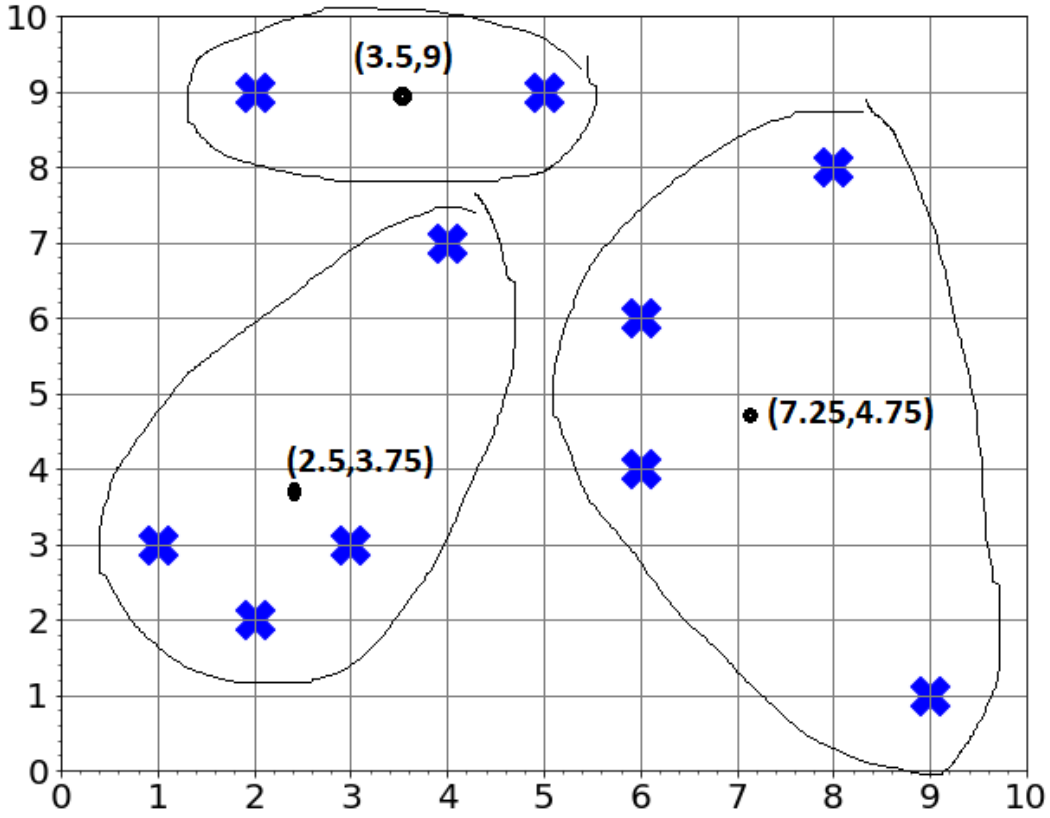
Round 2: Centroid C1,C2,C3

	C1	C2	C3	Cluster
A	4.65	4.80	3	C3
B	2.69	6.49	6.08	C1
C	2.41	3.95	2.82	C1
D	7.1	4.13	10.63	C2
E	2.97	5.92	7	C1
F	3.10	1.45	6.4	C2
G	5.93	3.33	6.08	C2
H	4.31	6.75	0	C3
I	3.23	1.76	5	C2
J	1.79	4.59	6.08	C1

C1 → B,C,E,J
C2 → D,F,G,I
C3 → A,H

New Centroids:

C1 → (1+4+2+3)/4 , (3+7+2+3)/4 →(2.5,3.75)
C2 →(9+6+8+6)/4 , (1+4+8+6)/4 → (7.25,4.75)
C3 → (2+5)/2 , (9+9)/2 → (3.5,9)

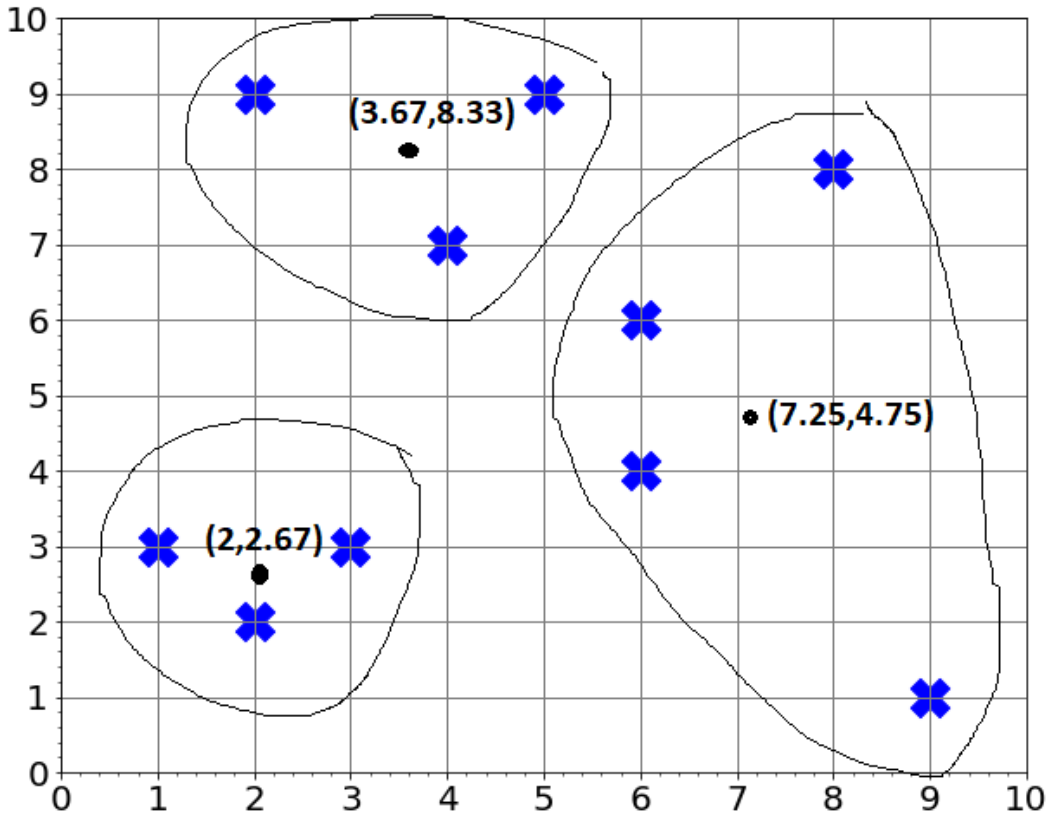


Round 3: Centroid C1,C2,C3

	C1	C2	C3	Cluster
A	5.81	4.80	1.5	C3
B	1.67	6.49	6.5	C1
C	3.57	3.95	2.06	C3
D	7.05	4.13	9.70	C2
E	1.82	5.92	7.15	C1
F	3.508	1.45	5.59	C2
G	6.95	3.33	4.60	C2
H	5.27	6.75	1.5	C3
I	4.16	1.76	3.90	C2
J	0.901	4.59	6.02	C1

C1 → B,E,J
C2 → D,F,G,I
C3 → A,C,H

New Centroids:
C1 → (1+2+3)/3 , (3+2+3)/3 →(2,2.67)
C2 →(9+6+8+6)/4 , (1+4+8+6)/4 → (7.25,4.75)
C3 → (2+5+4)/3 , (9+9+7)/3 → (3.67,8.33)



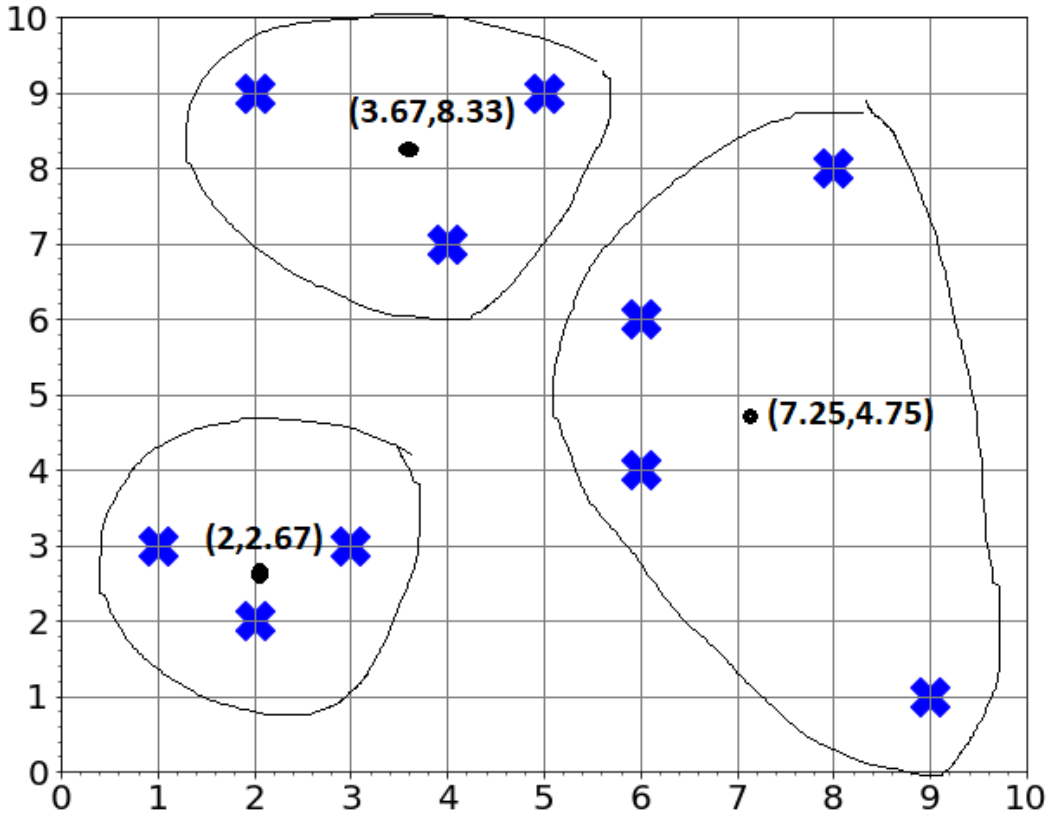
Round 4: Centroid C1,C2,C3

	C1	C2	C3	Cluster
A	7.00	4.80	1.48	C3
B	1.05	6.49	5.96	C1
C	4.76	3.95	1.37	C3
D	7.19	4.13	9.06	C2
E	0.66	5.92	6.54	C1
F	4.21	1.45	4.91	C2
G	8.02	3.33	4.34	C2

H	6.33	6.75	1.79	C3
I	5.20	1.76	3.29	C2
J	1.05	4.59	5.37	C1

C1 → B,E,J
C2 → D,F,G,I
C3 → A,C,H

New Centroids:
C1 → (1+2+3)/3 , (3+2+3)/3 →(2,2.67)
C2 →(9+6+8+6)/4 , (1+4+8+6)/4 → (7.25,4.75)
C3 → (2+5+4)/3 , (9+9+7)/3 → (3.67,8.33)



1.b)

Total 4 iterations were needed for the k means to converge.

2.a)

Initial distance

	1	2	3	4	5	6	7	8	9	10
1	0.00	7.21	2.24	8.94	7.62	5.10	3.16	3.00	3.16	6.32
2		0.00	5.00	8.25	1.41	5.10	8.60	6.08	5.83	2.00
3			0.00	7.81	5.39	3.61	4.12	2.83	2.224	4.12
4				0.00	7.07	4.24	7.07	10.63	5.83	6.32
5					0.00	4.47	8.49	7.00	5.66	1.41
6						0.00	4.47	6.40	2.00	3.16
7							0.00	6.08	2.83	7.07
8								0.00	5.00	6.08
9									0.00	4.24
10										0.00

After first iteration

Shortest distance: 1.41 , points clustered: 2,5,10

Updated distance table:

	(2,5,10)	1	3	4	6	7	8	9
(2,5,10)	0.00	6.32	4.12	6.32	3.16	7.07	6.08	4.24
1		0.00	2.24	8.94	5.10	3.16	3.00	3.16
3			0.00	7.81	3.61	4.12	2.83	2.24
4				0.00	4.24	7.07	10.63	5.83
6					0.00	4.47	6.40	2.00
7						0.00	6.08	2.83
8							0.00	5.00
9								0.00

After second iteration

Shortest distance: 2.00, points clustered: 6,9

Updated distance table:

	(2,5,10)	(6,9)	1	3	4	7	8
(2,5,10)	0.00	3.16	6.32	4.12	6.32	7.07	6.08
(6,9)		0.00	3.16	2.24	4.24	2.83	5.00
1			0.00	2.24	8.94	3.16	3.00
3				0.00	7.81	4.12	2.83
4					0.00	7.07	10.63
7						0.00	6.08
8							0.00

After third iteration:

Shortest distance: 2.24, points clustered: 1,3,6,9

Updated distance table:

	(2,5,10)	(1,3,6,9)	4	7	8
(2,5,10)	0.00	3.16	6.32	7.07	6.08
(1,3,6,9)		0.00	4.24	2.83	2.83
4			0.00	7.07	10.63
7				0.00	6.08
8					0.00

After fourth iteration:

Shortest distance: 2.83, points clustered: 1,3,6,7,8,9

Updated distance table:

	(2,5,10)	(1,3,6,7,8,9)	4
(2,5,10)	0.00	3.16	6.32
(1,3,6,7,8,9)		0.00	4.24
4			0.00

After fifth iteration:

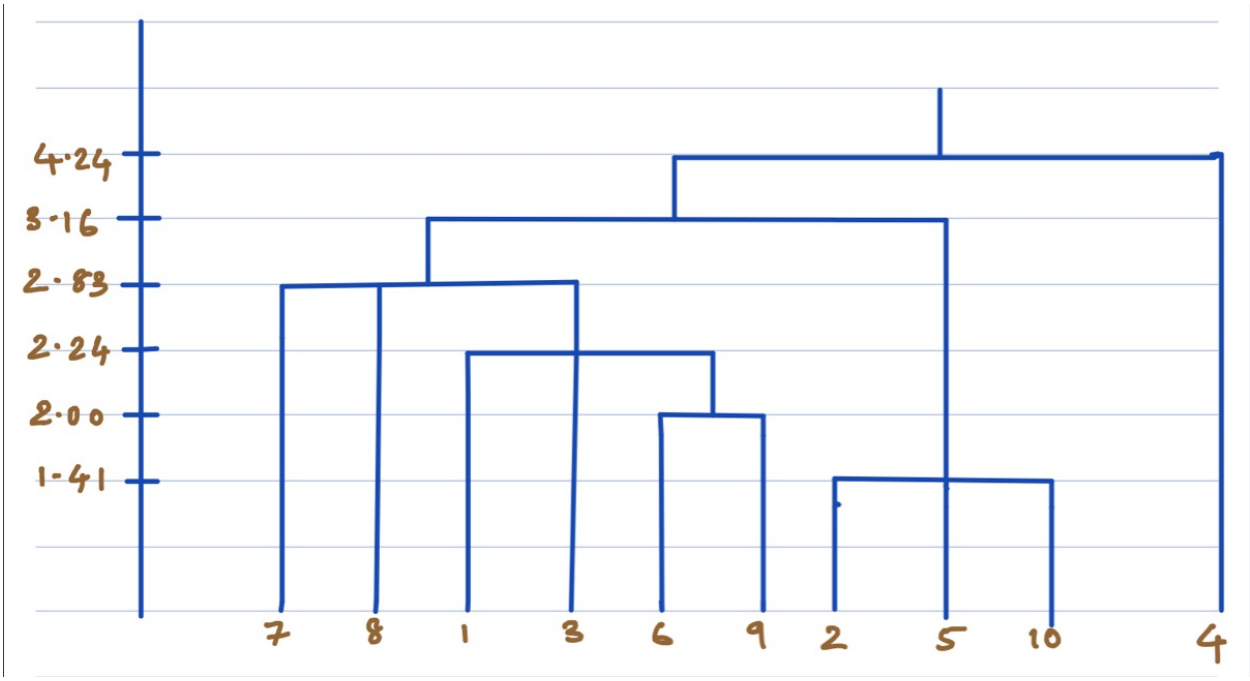
Shortest distance: 3.16, points clustered: 1,2,3,5,6,7,8,9,10

Updated distance table:

	(1,2,3,5,6,7,8,9,10)	4
(1,2,3,5,6,7,8,9,10)	0.00	4.24
4		0.00

After sixth iteration

Shortest distance: 4.24, points clustered: 1,2,3,4,5,6,7,8,9,10



2.b

Initial distance

	1	2	3	4	5	6	7	8	9	10
1	0.00	7.21	2.24	8.94	7.62	5.10	3.16	3.00	3.16	6.32
2		0.00	5.00	8.25	1.41	5.10	8.60	6.08	5.83	2.00
3			0.00	7.81	5.39	3.61	4.12	2.83	2.224	4.12
4				0.00	7.07	4.24	7.07	10.63	5.83	6.32
5					0.00	4.47	8.49	7.00	5.66	1.41
6						0.00	4.47	6.40	2.00	3.16
7							0.00	6.08	2.83	7.07
8								0.00	5.00	6.08
9									0.00	4.24
10										0.00

After first iteration

Shortest distance: 1.41 , points clustered: 5,10

Updated distance table:

	(5,10)	1	2	3	4	6	7	8	9
(5,10)	0.00	7.62	2.00	5.39	7.07	4.47	8.49	7.00	5.66
1		0.00	7.21	2.24	8.94	5.10	3.16	3.00	3.16
2			0.00	5.00	8.25	5.10	8.60	6.08	5.83
3				0.00	7.81	3.61	4.12	2.83	2.24
4					0.00	4.24	7.07	10.63	5.83
6						0.00	4.47	6.40	2.00
7							0.00	6.08	2.83
8								0.00	5.00
9									0.00

After second iteration

Shortest distance: 2.00, points clustered (2,5,10)

Updated distance table:

	(2,5,10)	1	3	4	6	7	8	9
(2,5,10)	0.00	7.62	5.39	8.25	5.10	8.60	7.00	5.83
1		0.00	2.24	8.94	5.10	3.16	3.00	3.16
3			0.00	7.81	3.61	4.12	2.83	2.24
4				0.00	4.24	7.07	10.63	5.83
6					0.00	4.47	6.40	2.00
7						0.00	6.08	2.83
8							0.00	5.00
9								0.00

After third iteration

Shortest distance: 2.00, points clustered (6,9)

Updated distance table:

	(2,5,10)	(6,9)	1	3	4	7	8
(2,5,10)	0.00	5.83	7.62	5.39	8.25	8.60	7.00
(6,9)		0.00	5.10	3.61	5.83	4.47	6.40
1			0.00	2.24	8.94	3.16	3.00
3				0.00	7.81	4.12	2.83
4					0.00	7.07	10.63
7						0.00	6.08
8							0.00

After fourth iteration

Shortest distance: 2.24, points clustered (1,3)

Updated distance table:

	(2,5,10)	(6,9)	(1,3)	4	7	8
(2,5,10)	0.00	5.83	7.62	8.25	8.60	7.00
(6,9)		0.00	6.40	5.83	4.47	6.40
(1,3)			0.00	8.94	4.12	3.00
4				0.00	7.07	10.63
7					0.00	6.08
8						0.00

After fifth iteration

Shortest distance: 3.00, points clustered (1,3,8)

Updated distance table:

	(2,5,10)	(6,9)	(1,3,8)	4	7
(2,5,10)	0.00	5.83	7.62	8.25	8.60
(6,9)		0.00	6.40	5.83	4.47
(1,3,8)			0.00	10.63	6.08
4				0.00	7.07
7					0.00

After sixth iteration

Shortest distance: 4.47, points clustered (6,7,9)

Updated distance table:

	(2,5,10)	(6,7,9)	(1,3,8)	4
(2,5,10)	0.00	8.60	7.62	8.25
(6,7,9)		0.00	6.40	7.07
(1,3,8)			0.00	10.63
4				0.00

After seventh iteration

Shortest distance: 6.40, points clustered (1,3,6,7,8,9)

Updated distance table:

	(2,5,10)	(1,3,6,7,8,9)	4
(2,5,10)	0.00	8.60	8.25
(1,3,6,7,8,9)		0.00	10.63
4			0.00

After seventh iteration

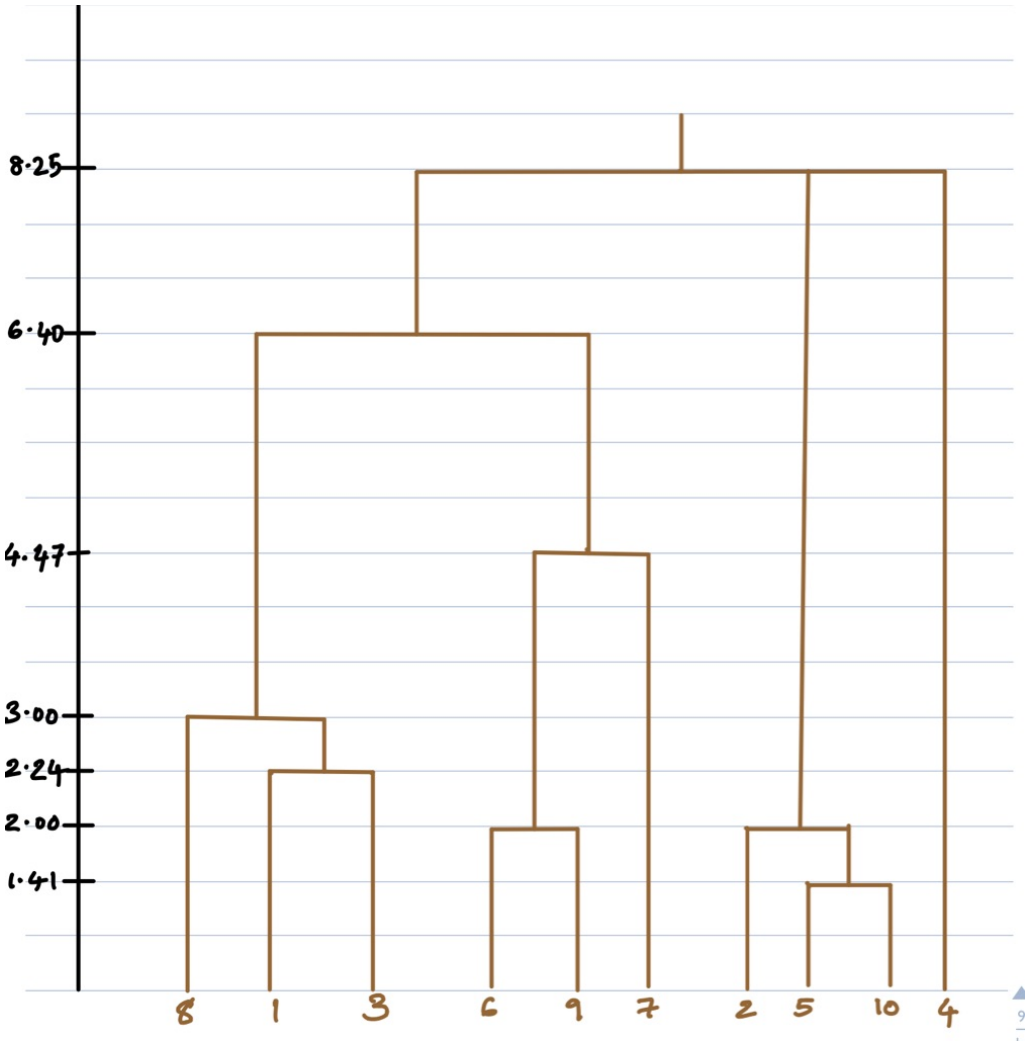
Shortest distance: 8.25, points clustered (2,4,5,10)

Updated distance table:

	(2,4,5,10)	(1,3,6,7,8,9)
(2,4,5,10)	0.00	8.60
(1,3,6,7,8,9)		0.00

After eight iteration

Shortest distance: 8.25, points clustered (1,2,3,4,5,6,7,8,9,10)



2.c

Complete link clustering will be a better approach if we assume two clusters because the points are distributed more evenly among the clusters. In the case of the single link clustering, the second cluster will just have a single point.

2.d

Single link hierarchical clustering

	a^i	b^i	$s(i)$
1	3.332	7.05	0.527
2	1.705	6.303	0.729
3	3.008	4.837	0.378
4			0 (single point cluster)
5	1.410	6.438	0.781
6	4.316	4.24	-0.018
7	4.132	7.07	0.416
8	4.662	6.387	0.27
9	3.046	5.243	0.419
10	1.705	5.165	0.670

Silhouette coefficient = $avg(s(i))$ = 0.4172

K-Means clustering

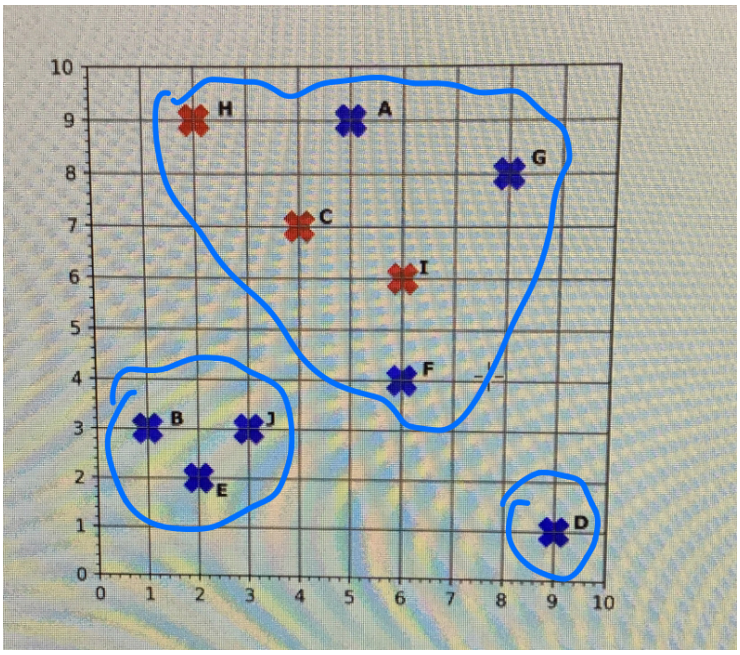
	a^i	b^i	$s(i)$
A	2.62	5.09	0.485
B	1.705	6.097	0.72
C	2.535	4.445	0.430
D	5.713	7.213	0.208
E	1.41	6.423	0.78
F	3.57	4.243	0.159
G	4.79	4.453	-0.07
H	2.915	6.387	0.544
I	3.553	3.467	-0.024
J	1.705	5.197	0.672

Silhouette coefficient = $avg(s(i))$ = 0.3904

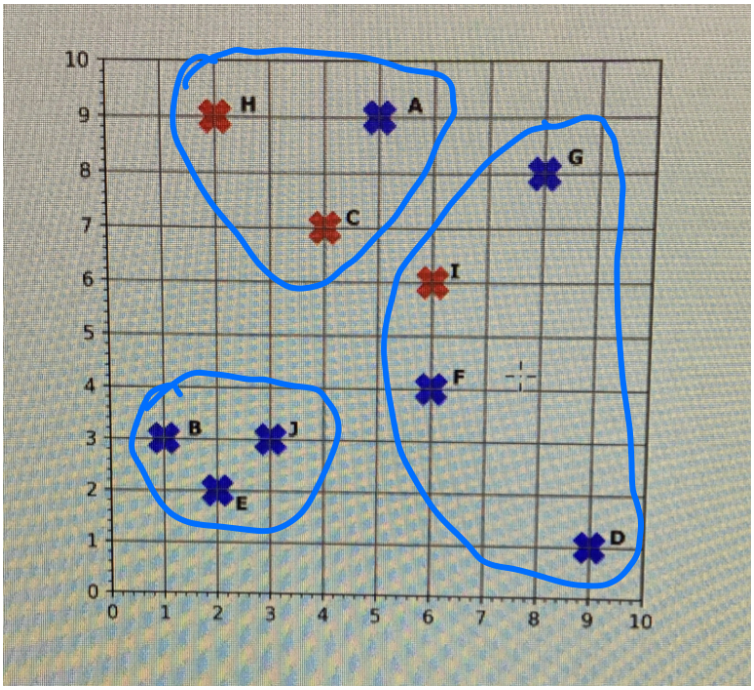
Based on silhouette coefficient, single link clustering is better because it has a higher value. A higher value indicates that the data point is well matched to its own cluster and poorly matched

to neighboring clusters. I agree with this assessment because as could be seen from the figures below, single link clustering gives a more compact clustering among the points than kmeans.

Single link clustering



KMeans clustering



3. Association Rule Mining

(a) What is the maximum number of unique itemsets that can be extracted from this data set (only including itemsets that have ≥ 1 support)?

Support: Fraction of transactions that contain an itemset.
Condition: need to consider only those itemsets whose support is >= 1
Consider all 1-itemsets:
{Bread}, {Milk}, {Butter}, {Eggs}, {Beer}, {Cola}
The support value of all these single items is < 1 as none of the items appear in all the transactions.
Therefore, the support count of their super sets will also be < 1.
So, the maximum number of unique itemsets with support >= 1 is zero.

(b) What is the maximum number of association rules that can be extracted from this data set (including rules that have zero support)?

Given d items, there are 2^d possible candidate itemsets.
In the given problem, the value of 'd' is 6.
Using the formula for the number of association rules calculation,

$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$= 3^d - 2^{d+1} + 1$$

We get R = 602 rules.

(c) Compute the support of the itemset: {Eggs,Cola}

Support = fractions of transactions that contain both Eggs and Cola
Support = 2/10 = 0.2

(d) Compute the support and confidence of association rule: {Bread}→{Butter}?

Support = fractions of transactions that contain both Bread and Beer
Support = 3/10 = 0.3
Confidence = $\frac{\sigma(\text{Bread,Butter})}{\sigma(\text{Bread})} = \frac{3}{6} = 0.5$

(e) Given min support = 0.3 and min confidence = 0.6, identify all valid association rules of the form {A,B}→{C}

Given 6 items, itemsets of size 3 are:
{bread, milk, butter}, {bread, milk, eggs}, {bread, milk, beer},

{bread, milk, cola}, {milk, butter, eggs}, {milk, butter, beer},
{milk, butter, cola}, {bread, beer, cola}, {bread, eggs, beer},
{bread, eggs, cola}, {bread, butter, eggs}, {bread, butter, beer},
{bread, butter, cola}, {butter, beer, cola}, {butter, eggs, cola},
{eggs, beer, cola}, {butter, eggs, beer}, {milk, eggs, beer},
{milk, beer, cola}, {milk, eggs, cola}

Total 20 items and only {bread, cola, milk} has support >= 0.3
Different combinations of form **{A,B}→{C}** are:
{Bread, Cola} -> {Milk} This rule has a support of 0.3 and confidence of 1
{Bread, Milk} -> {Cola} This rule has a support of 0.3 and confidence of 0.75
{Milk, Cola} -> {Bread} This rule has a support of 0.3 and confidence of 0.75
Therefore, there are 3 valid association rules of the form {A,B}→{C} with min support = 0.3 and min confidence = 0.6

(f) In a different dataset, the support of the rule {a} → {b} is 0.46, and the support of the rule {a,c} → {b,d} is 0.23. What can we say for sure about the support of the rule {a} → {b,d}.
The set {a, c, d} is a super set of {a, b} and subset of {a, b, c, d}
Therefore, the support 'S' of the rule {a} → {b,d} will be,
0.23 <= S <= 0.46

4. Association Rule Mining

(a) Compute each step of frequent itemset generation process using the Apriori algorithm with a minimum support count of 3

Step 1: Database with transaction IDs and Items list

TID	Items
T1	A,B,C,D
T2	A,B,D,E
T3	A,B
T4	A,C,D
T5	A,C,E
T6	B,C
T7	C,D
T8	C,D,E

Step 2: Scan database and populate support count for itemset of size 1

Itemset	Sup
{A}	5
{B}	4
{C}	6
{D}	5
{E}	3

Step 3: The support count for all the items is >= 3. So, no dropping.

Step 4: Construct itemsets of size 2 using the results of step 3 and populate their support count

Itemset	Sup
{A,B}	3
{A,C}	3
{A,D}	3
{A,E}	2
{B,C}	2
{B,D}	1
{B,E}	1
{C,D}	4
{C,E}	2
{D,E}	2

Step 5: Drop all itemsets whose support count is < 3

Itemset	Sup
{A,B}	3
{A,C}	3
{A,D}	3
{C,D}	4

Step 6:
a) Construct itemsets of size 3 using the results of step 5 and populate their support count.
b) Prune – test all these itemsets of size 3 for the eexistence of 2-item subsets within the entries of the resultant table of step 5 and prune whenever required.

Itemset	Sup
{A,C,D}	2

Step 7: Drop all itemsets whose support count is < 3
Result:

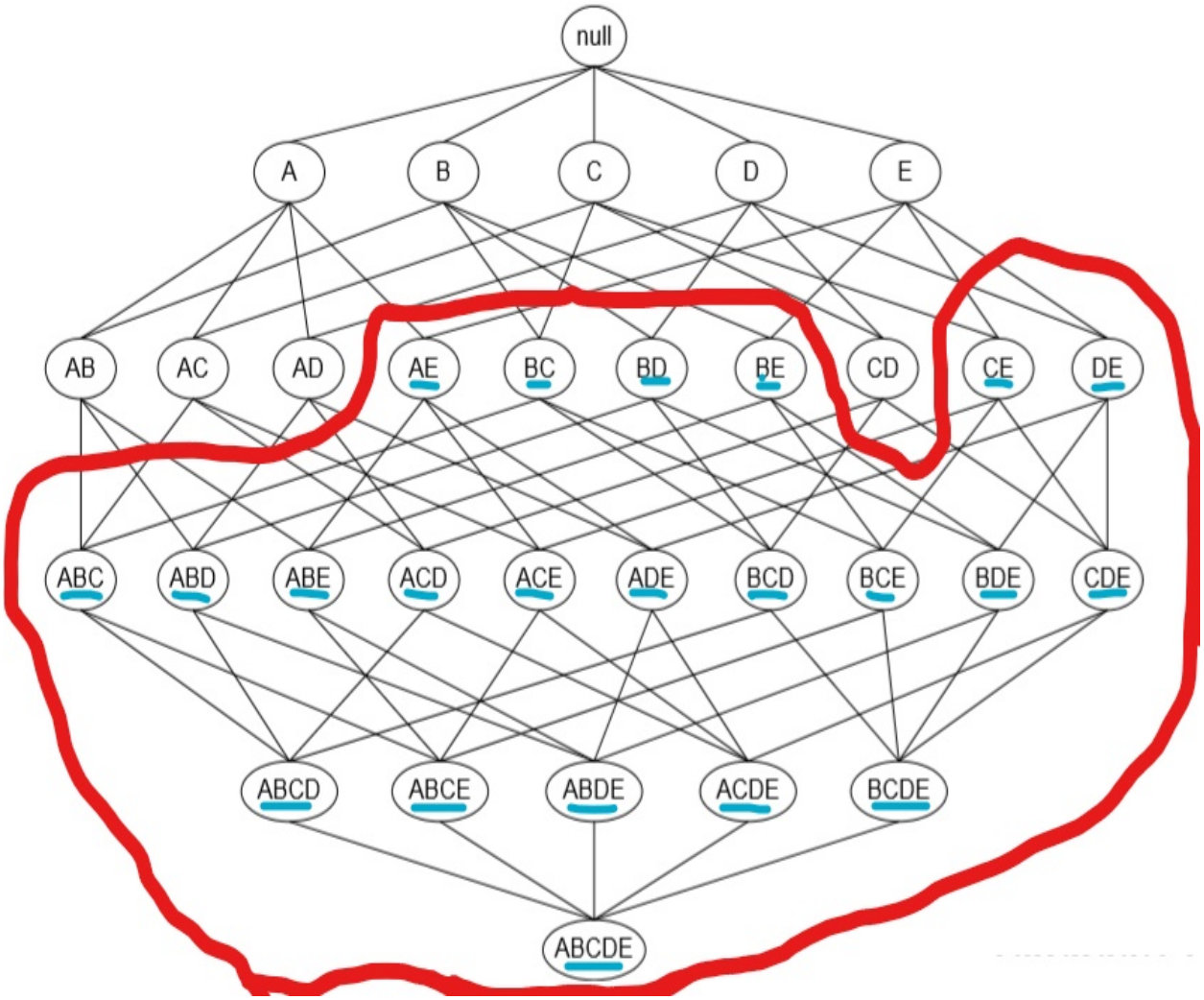
Itemset	Sup
Empty	

(b) Show the lattice structure for the data given in table above and mark the pruned branches if any.

The Red region is the pruned part and all the supersets inside this area are pruned because of infrequent items like AE, BC, BD, BE, CE, DE and these pruned supersets are marked/underlined in blue color.

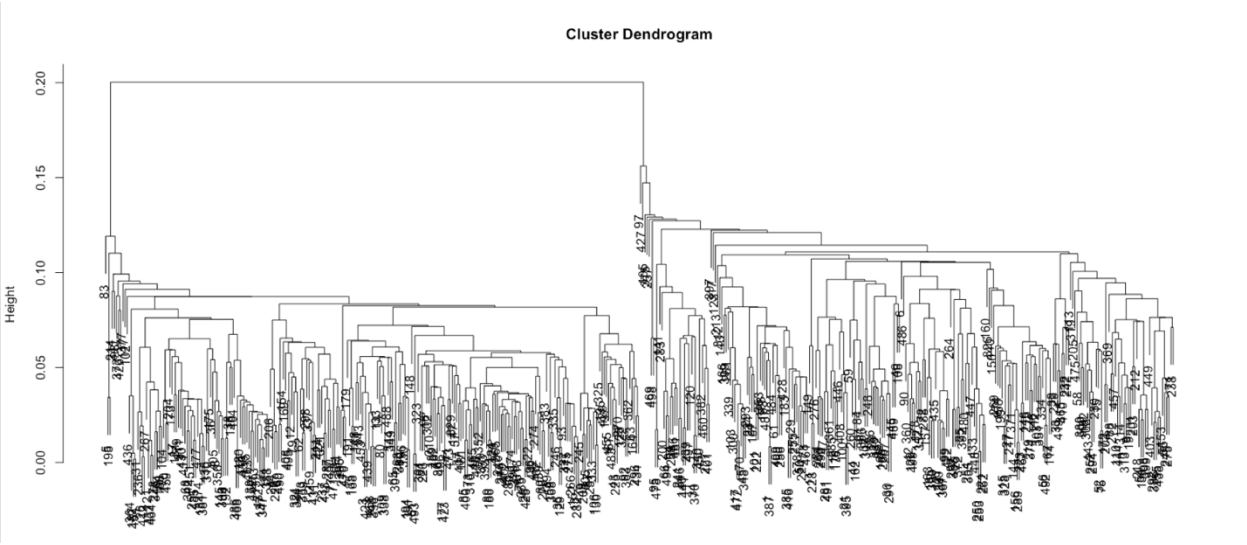
In the previous question, we observed that only {A,B}, {A,C}, {A,D} and {C,D} have support count >= 0.3.

So, all other itemsets of size 2 and their supersets will be pruned as shown below.

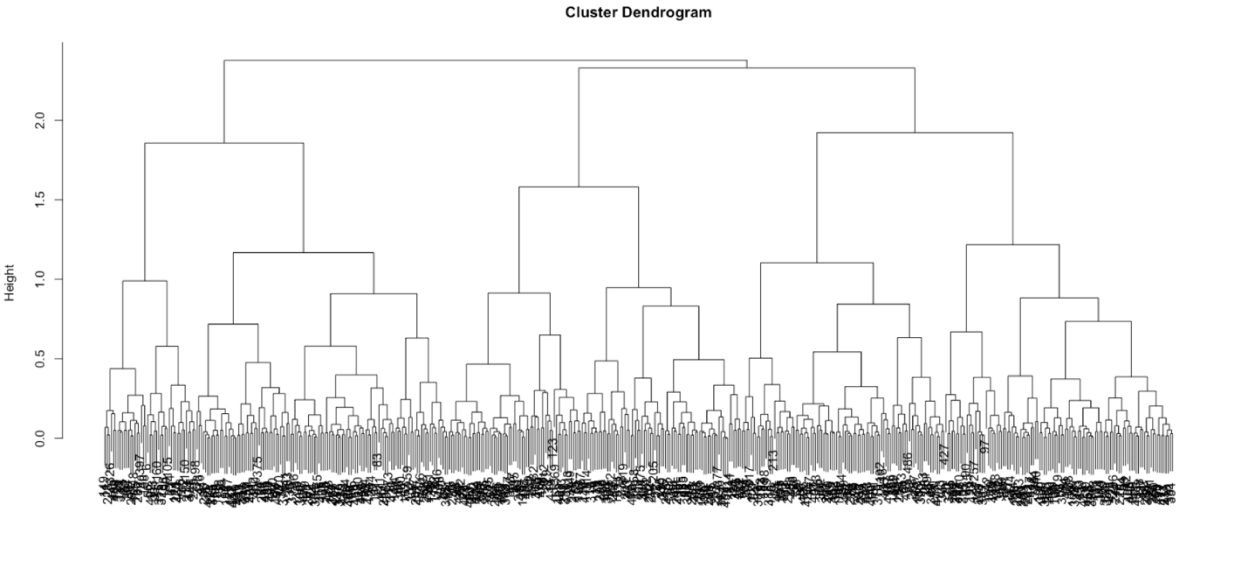


5.iii.A)

Dendrogram for single link hierarchical clustering.

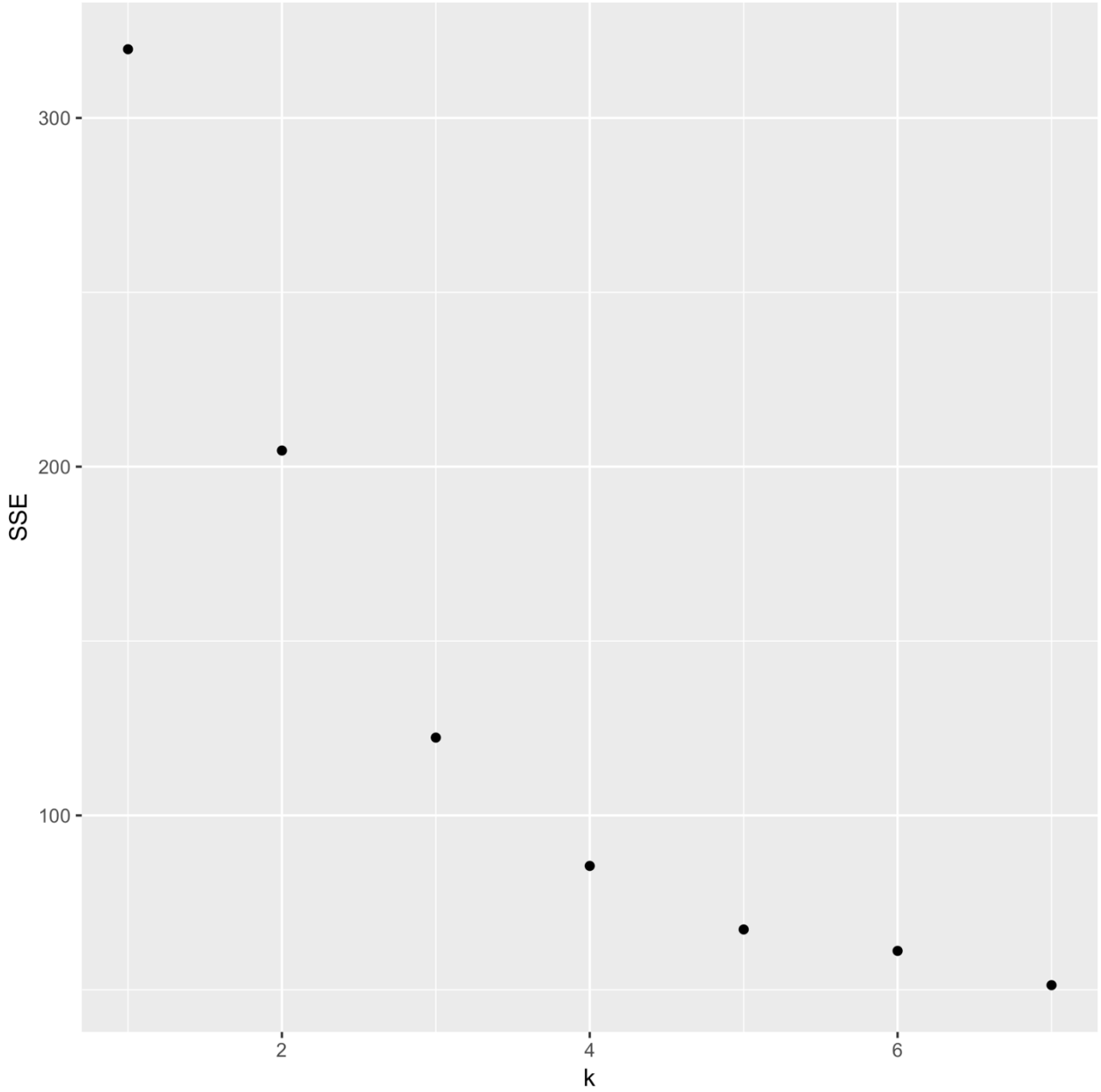


Dendrogram for complete link hierarchical clustering



5.iii.C)

Elbow plot



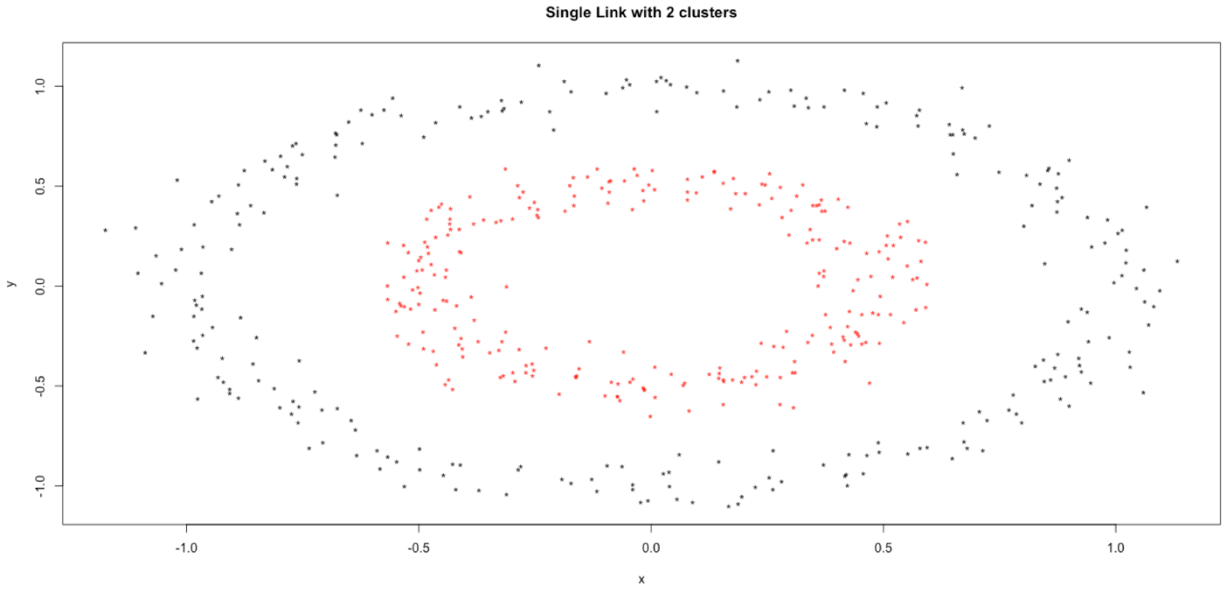
5.iii.D)

KMeans SSE for given params = 204.626515194937
Single link SSE for given params = 319.693456992432
Complete link SSE for given params = 221.041974498579

Based on purely SSE, KMeans clustering is the best since it has the lowest SSE.

5.iii.E)

Based on the visualization plots, single link hierarchical clustering is the best because it nicely and logically clusters the points, the inner ring and the outer ring.



5.iii.F)

No, it is not the same. Single link clustering, even though having a higher SSE then kmeans, does a more logical clustering. The inner ring and the outer ring are put in to separate clusters. Thus it makes sense to conclude that numerical measures alone are not enough to judge the quality of clustering.