

Data Mining
Assignment – 1
National Institute of Technology Calicut
Submitted By: Prabodh T R, B170371CS, CSE

1) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4),
B1(5, 8), B2(7, 5), B3(6, 4),
C1(1, 2), C2(4, 9)

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only

- (a) The three cluster centers after the first round execution
(b) The final three clusters

Ans: (a)

step 1 : find the distance between current centroids and rest of the patterns/objects using the euclidean distance formula :

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

step 2 : for each pattern, add them to the cluster represented by the centroid nearest to them

step 3 : recalculate the centroid for each cluster

step 4 : repeat above steps till the centroid remains the same for all the cluster

	A1	B1	C1
A2	5	4.242641	3.162278
A3	8.485281	5	7.28011
B2	7.071068	3.605551	6.708204
B3	7.211103	4.123106	5.385165
C2	2.236068	1.414214	7.615773

after first iteration, $C_1 = \{ A1 \}$, $C_2 = \{ A3, B1, B2, B3, C2 \}$, $C_3 = \{ A2, C1 \}$ and

cluster center of $C_1(C_1') = (2, 10)$

cluster center of $C_2(C_2') = (6, 6)$

cluster center of $C_3(C_3') = (1.5, 3.5)$

(b).

iteration 2 :

$C_1 = \{ A1, C2 \}$, $C_2 = \{ A3, B1, B2, B3 \}$, $C_3 = \{ A2, C1 \}$

centroids: $C_1' = (3, 9.5)$, $C_2' = (6.5, 5.25)$, $C_3' = (1.5, 3.5)$

iteration 3 :

$C_1 = \{ A1, B1, C2 \}$, $C_2 = \{ A3, B2, B3 \}$, $C_3 = \{ A2, C1 \}$

centroids: $C_1' = (3.66, 9)$, $C_2' = (7, 4.33)$, $C_3' = (1.5, 3.5)$

final clusters are as follows:

$C_1 = \{ A1, B1, C2 \}$, $C_2 = \{ A3, B2, B3 \}$, $C_3 = \{ A2, C1 \}$

2) Write a program to implement k-means clustering algorithm by using Iris data set (available in UCI Machine learning repository) and find the followings:

i) Clusters of the Iris data set (final clustering solution).

ii) Sum of the Intra-Cluster Distances (SICD) or Sum of the Squared Error (SSE) values of the obtained clustering solution.

iii) Graphical representation of the obtained clusters.

- iv) Give 150 iterations of the implemented k-means by changing the initial centroids and see the changes in the clustering solution (SICD values).
v) Plot iteration vs. SICD values.

Ans:

i) cluster – I

[[6.5, 3.2, 5.1, 2.0], [6.9, 3.2, 5.7, 2.3], [6.7, 3.0, 5.2, 2.3], [6.7, 2.5, 5.8, 1.8], [6.4, 2.8, 5.6, 2.1], [6.3, 3.4, 5.6, 2.4], [6.4, 2.7, 5.3, 1.9], [7.7, 3.8, 6.7, 2.2], [6.3, 3.3, 6.0, 2.5], [6.7, 3.1, 5.6, 2.4], [6.5, 3.0, 5.8, 2.2], [7.4, 2.8, 6.1, 1.9], [6.8, 3.2, 5.9, 2.3], [6.5, 3.0, 5.5, 1.8], [7.3, 2.9, 6.3, 1.8], [6.7, 3.3, 5.7, 2.1], [7.2, 3.0, 5.8, 1.6], [7.7, 2.6, 6.9, 2.3], [7.9, 3.8, 6.4, 2.0], [6.7, 3.0, 5.0, 1.7], [6.5, 3.0, 5.2, 2.0], [6.3, 2.9, 5.6, 1.8], [6.9, 3.1, 5.4, 2.1], [7.2, 3.2, 6.0, 1.8], [6.2, 3.4, 5.4, 2.3], [7.7, 3.0, 6.1, 2.3], [6.1, 2.6, 5.6, 1.4], [6.8, 3.0, 5.5, 2.1], [7.7, 2.8, 6.7, 2.0], [6.9, 3.1, 4.9, 1.5], [6.4, 3.1, 5.5, 1.8], [6.9, 3.1, 5.1, 2.3], [6.4, 3.2, 5.3, 2.3], [6.7, 3.3, 5.7, 2.5], [7.2, 3.6, 6.1, 2.5], [6.4, 2.8, 5.6, 2.2], [7.1, 3.0, 5.9, 2.1], [7.6, 3.0, 6.6, 2.1]]

cluster – II

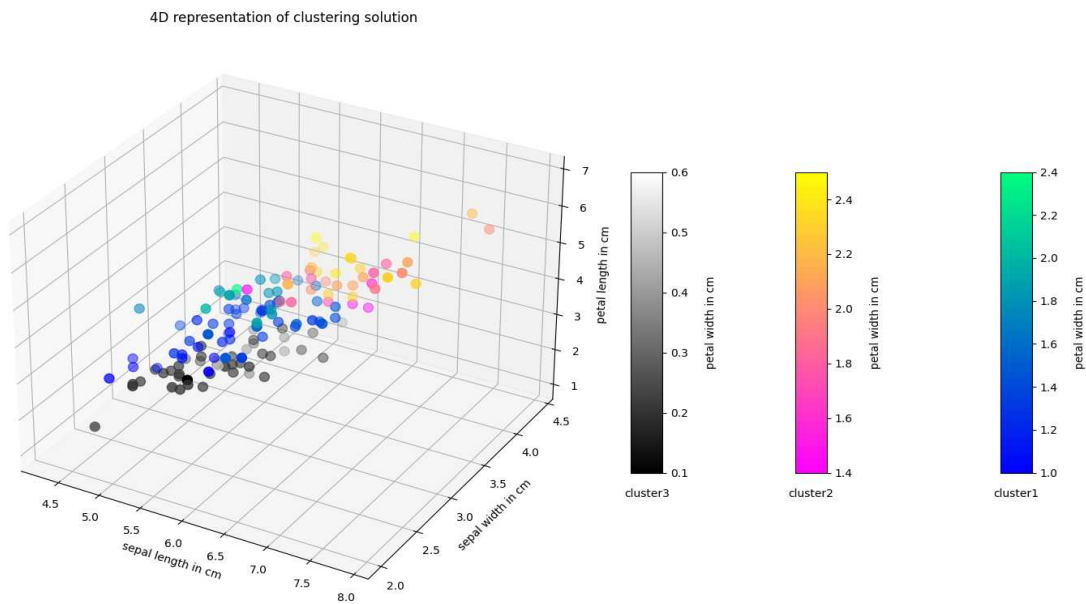
[[6.5, 2.8, 4.6, 1.5], [6.0, 2.2, 4.0, 1.0], [6.2, 2.8, 4.8, 1.8], [5.8, 2.7, 5.1, 1.9], [5.7, 2.9, 4.2, 1.3], [5.4, 3.0, 4.5, 1.5], [6.3, 2.5, 4.9, 1.5], [7.0, 3.2, 4.7, 1.4], [5.7, 2.6, 3.5, 1.0], [5.6, 3.0, 4.5, 1.5], [5.6, 3.0, 4.1, 1.3], [6.1, 3.0, 4.6, 1.4], [5.6, 2.7, 4.2, 1.3], [5.9, 3.0, 4.2, 1.5], [5.8, 2.6, 4.0, 1.2], [5.7, 2.5, 5.0, 2.0], [5.9, 3.0, 5.1, 1.8], [6.2, 2.9, 4.3, 1.3], [5.8, 2.7, 5.1, 1.9], [5.5, 2.5, 4.0, 1.3], [6.7, 3.1, 4.4, 1.4], [6.0, 3.4, 4.5, 1.6], [6.3, 2.5, 5.0, 1.9], [6.0, 3.0, 4.8, 1.8], [6.3, 2.8, 5.1, 1.5], [6.0, 2.2, 5.0, 1.5], [6.7, 3.1, 4.7, 1.5], [6.3, 2.3, 4.4, 1.3], [5.9, 3.2, 4.8, 1.8], [5.5, 2.3, 4.0, 1.3], [4.9, 2.4, 3.3, 1.0], [5.0, 2.3, 3.3, 1.0], [5.8, 2.7, 3.9, 1.2], [5.7, 3.0, 4.2, 1.2], [6.0, 2.9, 4.5, 1.5], [6.4, 2.9, 4.3, 1.3], [6.1, 2.9, 4.7, 1.4], [6.8, 2.8, 4.8, 1.4], [5.5, 2.4, 3.8, 1.1], [5.8, 2.7, 4.1, 1.0], [5.0, 2.0, 3.5, 1.0], [5.1, 2.5, 3.0, 1.1], [5.7, 2.8, 4.1, 1.3], [6.6, 2.9, 4.6, 1.3], [6.3, 2.7, 4.9, 1.8], [5.2, 2.7, 3.9, 1.4], [4.9, 2.5, 4.5, 1.7], [5.6, 2.5, 3.9, 1.1], [6.1, 2.8, 4.0, 1.3], [5.5, 2.4, 3.7, 1.0], [6.6, 3.0, 4.4, 1.4], [5.5, 2.6, 4.4, 1.2], [6.4, 3.2, 4.5, 1.5], [5.6, 2.9, 3.6, 1.3], [6.0, 2.7, 5.1, 1.6], [6.1, 3.0, 4.9, 1.8], [5.8, 2.8, 5.1, 2.4], [6.2, 2.2, 4.5, 1.5], [6.3, 3.3, 4.7, 1.6], [5.6, 2.8, 4.9, 2.0], [5.7, 2.8, 4.5, 1.3], [6.1, 2.8, 4.7, 1.2]]

cluster – III

[[4.9, 3.1, 1.5, 0.1], [5.4, 3.9, 1.7, 0.4], [5.0, 3.6, 1.4, 0.2], [4.4, 3.2, 1.3, 0.2], [5.7, 4.4, 1.5, 0.4], [4.5, 2.3, 1.3, 0.3], [5.5, 4.2, 1.4, 0.2], [4.4, 3.0, 1.3, 0.2], [5.1, 3.3, 1.7, 0.5], [4.6, 3.6, 1.0, 0.2], [5.7, 3.8, 1.7, 0.3], [5.0, 3.0, 1.6, 0.2], [5.2, 3.4, 1.4, 0.2], [5.2, 3.5, 1.5, 0.2], [5.1, 3.8, 1.9, 0.4], [5.0, 3.5, 1.3, 0.3], [5.1, 3.4, 1.5, 0.2], [4.8, 3.0, 1.4, 0.3], [4.4, 2.9, 1.4, 0.2], [5.8, 4.0, 1.2, 0.2], [4.3, 3.0, 1.1, 0.1], [5.4, 3.9, 1.3, 0.4], [5.1, 3.8, 1.6, 0.2], [5.5, 3.5, 1.3, 0.2], [5.3, 3.7, 1.5, 0.2], [4.8, 3.1, 1.6, 0.2], [4.8, 3.4, 1.6, 0.2], [4.9, 3.0, 1.4, 0.2], [5.0, 3.5, 1.6, 0.6], [5.2, 4.1, 1.5, 0.1], [5.4, 3.4, 1.7, 0.2], [4.7, 3.2, 1.3, 0.2], [4.9, 3.1, 1.5, 0.1], [4.7, 3.2, 1.6, 0.2], [4.9, 3.1, 1.5, 0.1], [5.1, 3.5, 1.4, 0.2], [5.4, 3.7, 1.5, 0.2], [5.0, 3.3, 1.4, 0.2], [4.6, 3.4, 1.4, 0.3], [5.0, 3.4, 1.6, 0.4], [4.8, 3.4, 1.9, 0.2], [5.4, 3.4, 1.5, 0.4], [5.0, 3.4, 1.5, 0.2], [5.1, 3.8, 1.5, 0.3], [4.8, 3.0, 1.4, 0.1], [5.1, 3.5, 1.4, 0.3], [4.6, 3.2, 1.4, 0.2], [4.6, 3.1, 1.5, 0.2], [5.1, 3.7, 1.5, 0.4], [5.0, 3.2, 1.2, 0.2]]

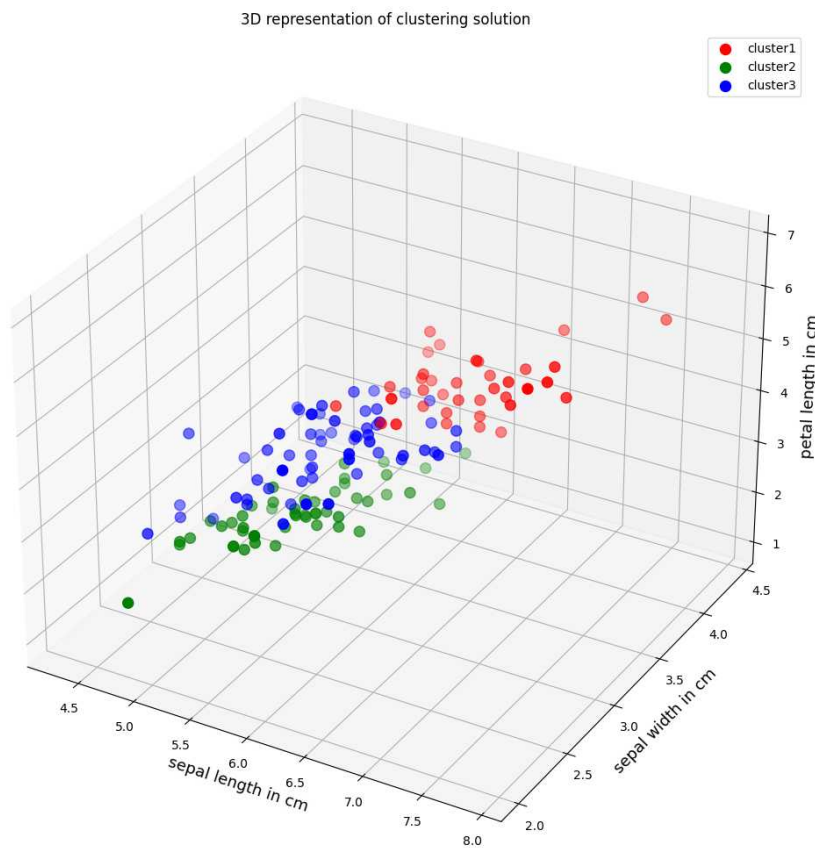
ii) SICD value of the best clustering solution = 78.94084142614598

iii) Graphical representation in **4D view**:

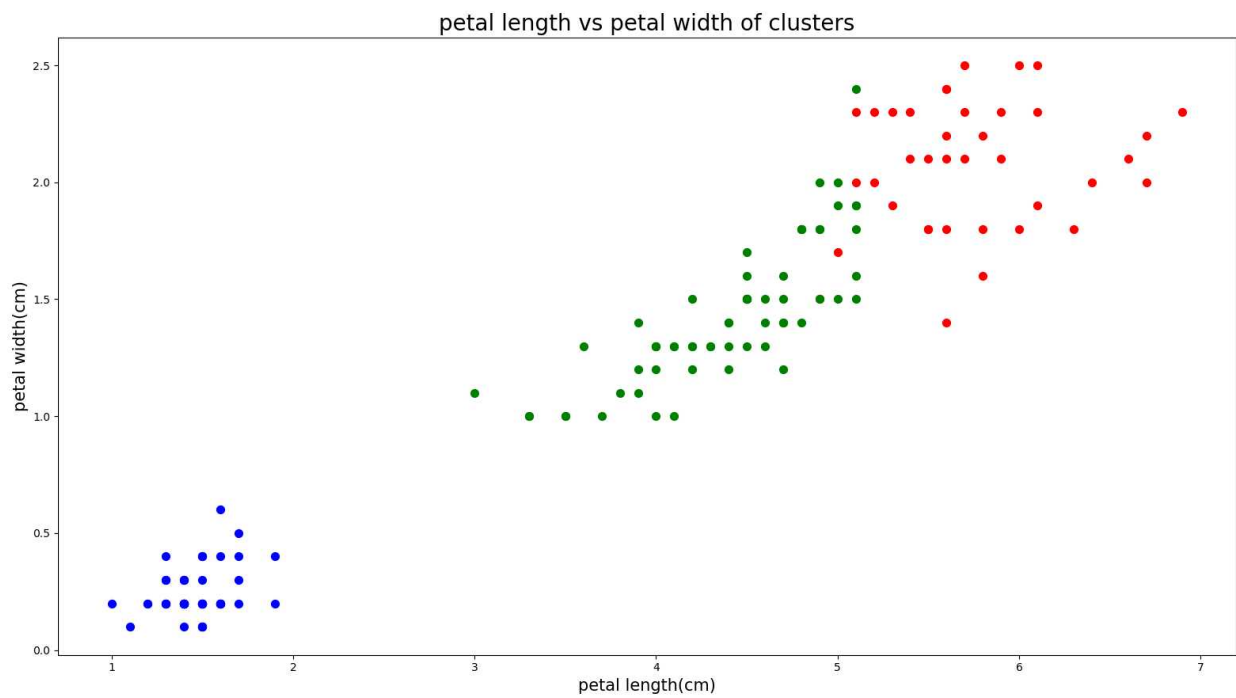
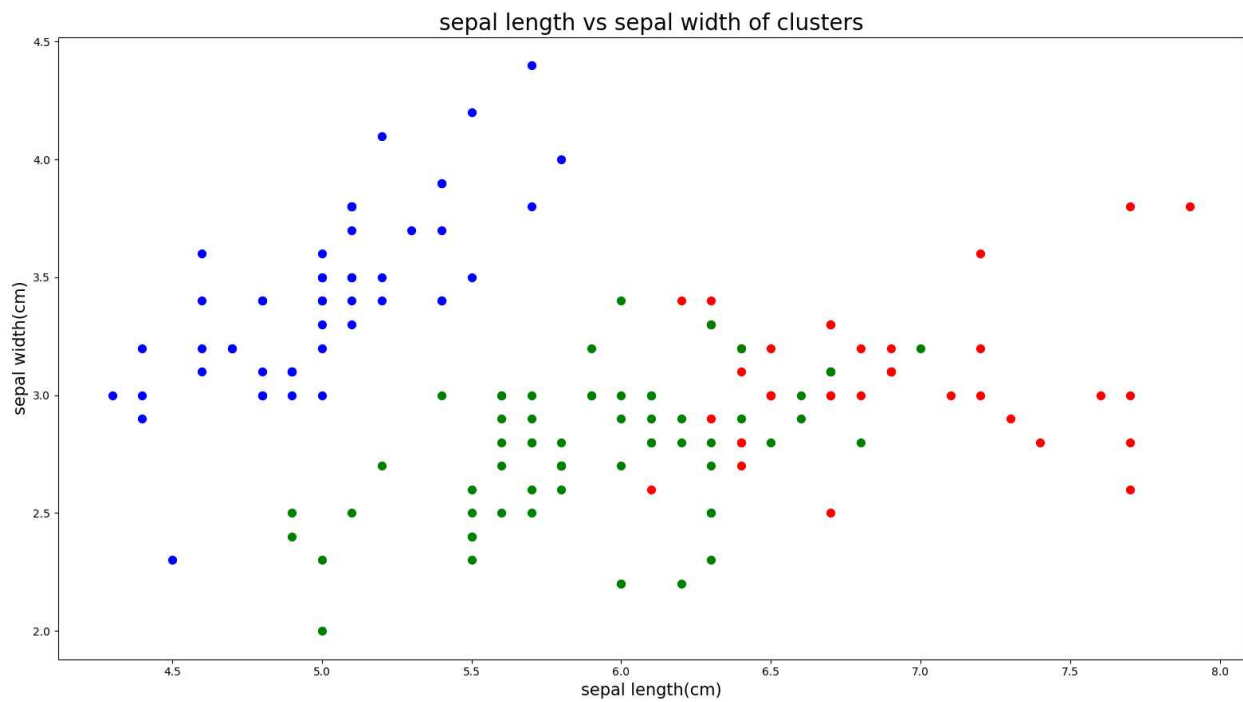


Inorder to visualize the clustering in a much more effective way, the following images of clustering in lower dimensions are used

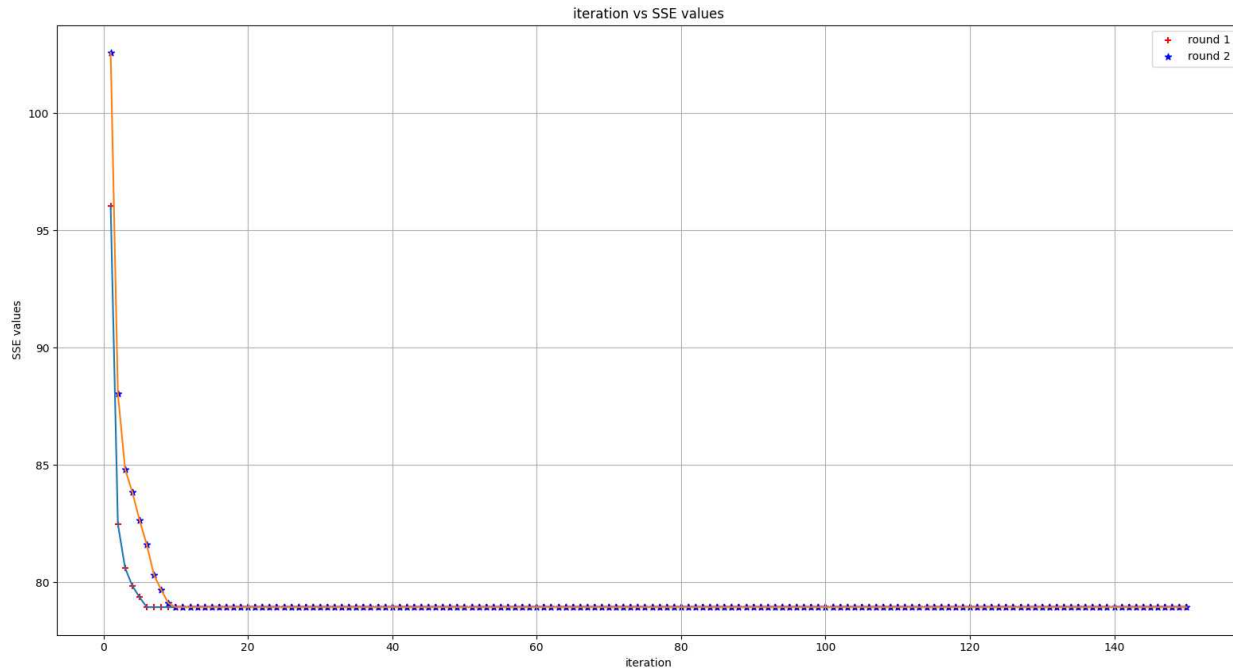
In 3D view:



In 2D view:



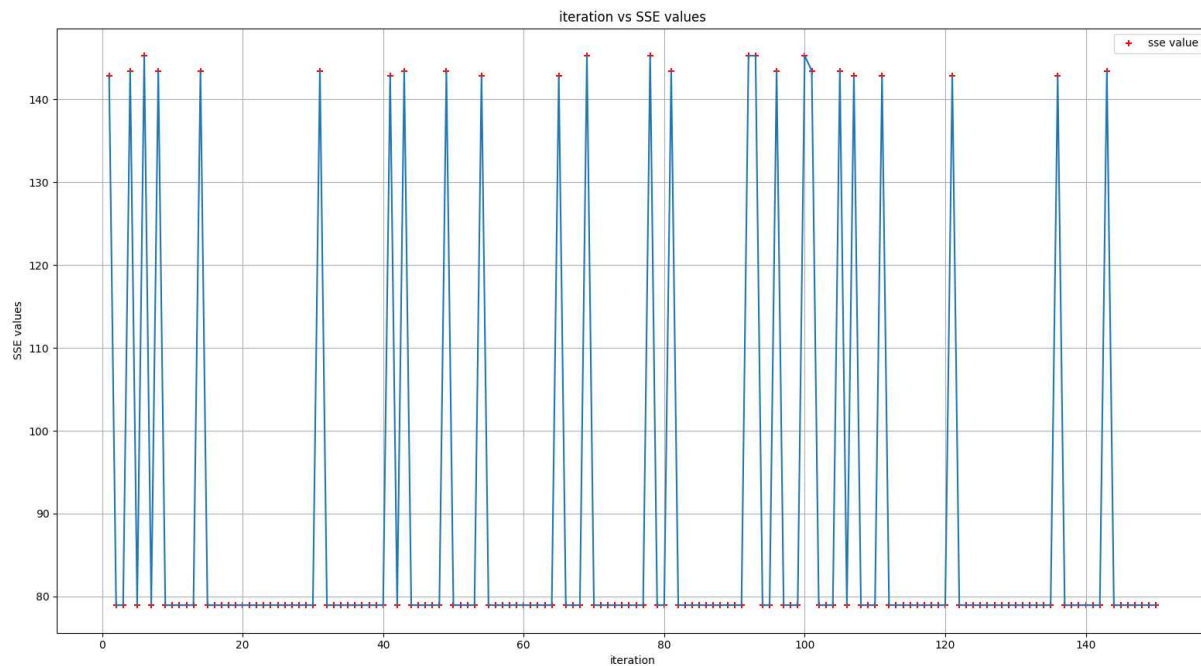
iv) The variation of SSE values on a different initial centroid values are best depicted with the help of a graph as follows:



As we can see, though the clustering followed a different path when started with a different set of centroid values the final clustering as well as the SSE value found are almost the same

Also it can be seen that multiple execution of the program with varying initial centroids, gives clustering solutions that are almost similar to each other but have slightly different SICD values. Higher the SICD value, lower the quality of clustering solution

v) iteration vs SSE graph for 150 clustering solutions are shown below:



The minimal SICD got = 78.94084142614598 and the cluster corresponding to it is the best solution as provided in prior answer