Data Mining
Assignment – II
<u>Implement Bisecting k-means clustering algorithm</u>

1) Write a program to implement Bisecting k-means clustering algorithm by using Iris data set and find the followings:
i) Clusters of the Iris data set (final clustering solution).
ii) Sum of the Intra-Cluster Distances (SICD) values of the obtained clustering solution.
iii) Graphical representation of the obtained clusters.
iv) Give 500 runs of the implemented Bisecting k-means by changing the initial centroids and see the changes in the clustering solution (SICD values).
v) Compare k-means and Bisecting k-means with respect to iteration vs SICD plot.

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

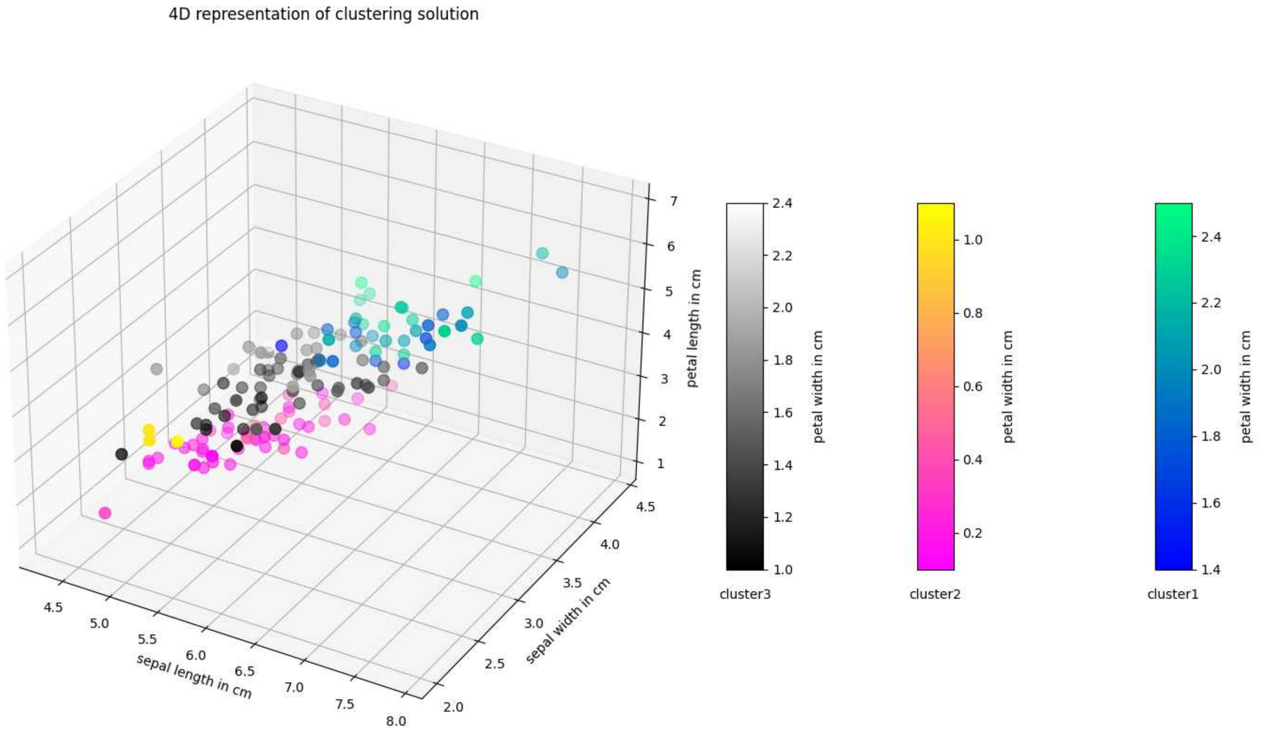
cluster II
[[4.6, 3.2, 1.4, 0.2], [4.9, 3.1, 1.5, 0.1], [5.3, 3.7, 1.5, 0.2], [5.0, 3.3, 1.4, 0.2], [5.0, 3.5, 1.6, 0.6], [4.7, 3.2, 1.6, 0.2], [4.8, 3.0, 1.4, 0.1], [4.9, 2.4, 3.3, 1.0], [5.4, 3.9, 1.7, 0.4], [4.7, 3.2, 1.3, 0.2], [4.8, 3.4, 1.6, 0.2], [5.0, 3.5, 1.3, 0.3], [5.5, 4.2, 1.4, 0.2], [4.4, 3.0, 1.3, 0.2], [5.1, 2.5, 3.0, 1.1], [5.0, 3.2, 1.2, 0.2], [5.7, 4.4, 1.5, 0.4], [4.6, 3.4, 1.4, 0.3], [4.8, 3.0, 1.4, 0.3], [5.1, 3.7, 1.5, 0.4], [5.4, 3.4, 1.5, 0.4], [5.2, 4.1, 1.5, 0.1], [5.1, 3.4, 1.5, 0.2], [4.4, 2.9, 1.4, 0.2], [5.4, 3.9, 1.3, 0.4], [5.2, 3.5, 1.5, 0.2], [5.1, 3.8, 1.6, 0.2], [5.1, 3.8, 1.9, 0.4], [4.5, 2.3, 1.3, 0.3], [5.0, 3.6, 1.4, 0.2], [5.8, 4.0, 1.2, 0.2], [4.6, 3.1, 1.5, 0.2], [5.5, 3.5, 1.3, 0.2], [4.8, 3.4, 1.9, 0.2], [5.0, 2.3, 3.3, 1.0], [5.0, 3.4, 1.5, 0.2], [4.8, 3.1, 1.6, 0.2], [4.9, 3.1, 1.5, 0.1], [5.1, 3.5, 1.4, 0.3], [5.1, 3.8, 1.5, 0.3], [4.3, 3.0, 1.1, 0.1], [5.1, 3.5, 1.4, 0.2], [5.2, 3.4, 1.4, 0.2], [5.4, 3.4, 1.7, 0.2], [4.9, 3.0, 1.4, 0.2], [4.6, 3.6, 1.0, 0.2], [4.4, 3.2, 1.3, 0.2], [4.9, 3.1, 1.5, 0.1], [5.0, 3.4, 1.6, 0.4], [5.4, 3.7, 1.5, 0.2], [5.7, 3.8, 1.7, 0.3], [5.0, 3.0, 1.6, 0.2], [5.1, 3.3, 1.7, 0.5]]

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

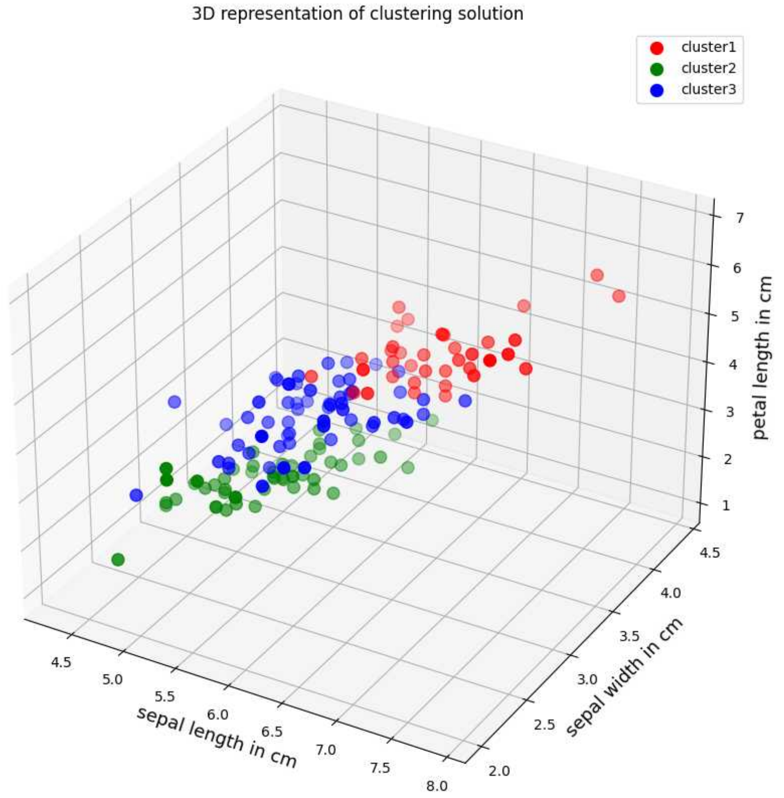
ii) The SSE value obtained for the final clustering solution is 84.22450726272024

iii) Various representation of the clustering solution (4D, 3D, 2D representations) have been shown below. Representations in lower dimensions are given to have a better picture on how the clusters have been made and hence may lack theoretical accuracy

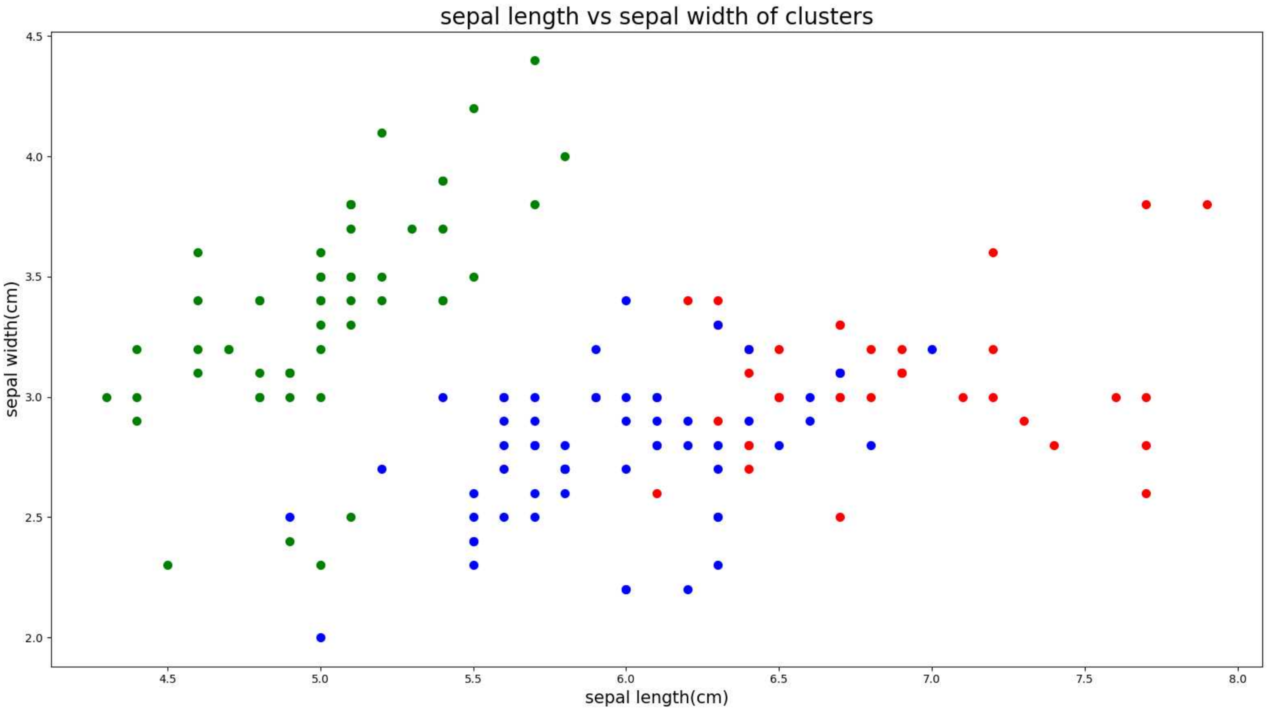
4D representation :

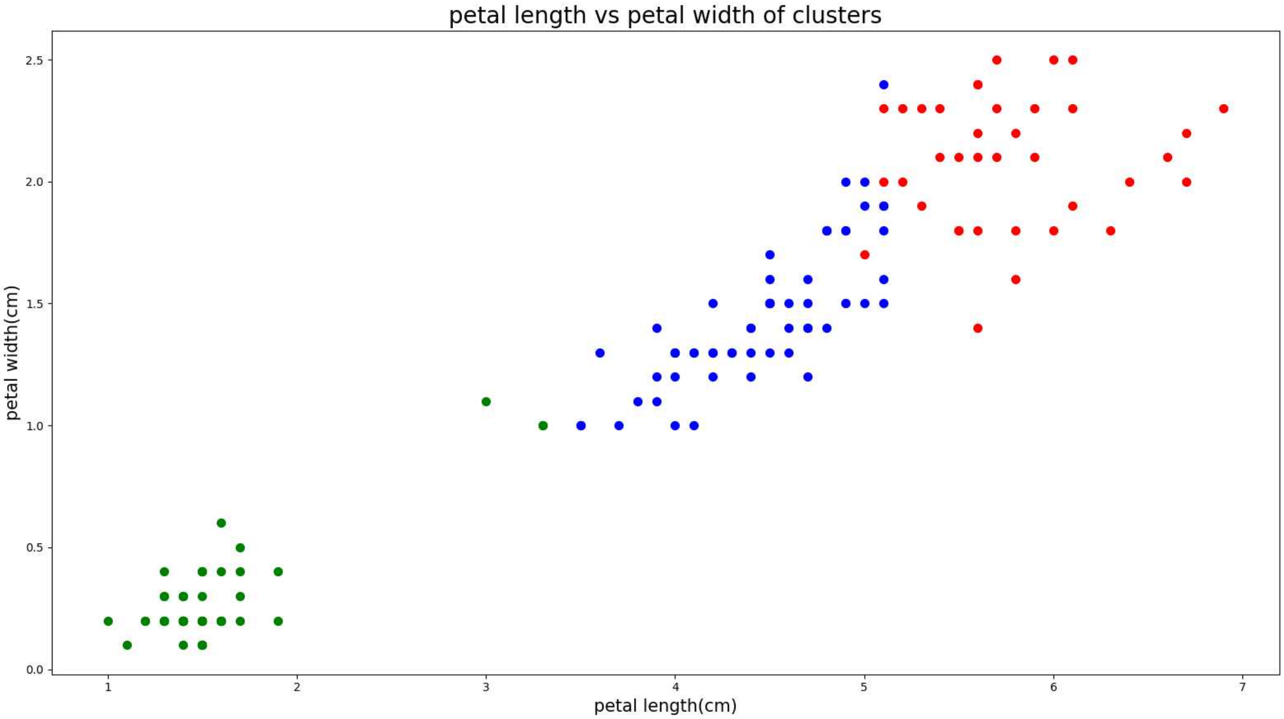


3D Representation:

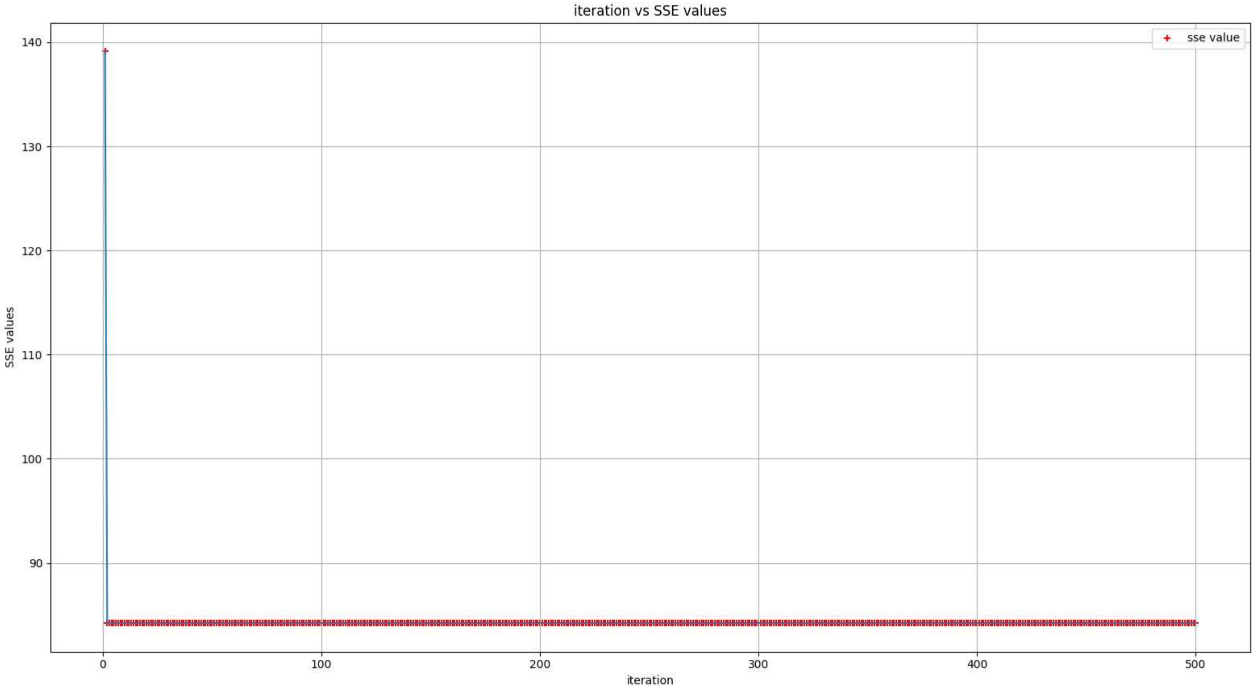


2D Representations:

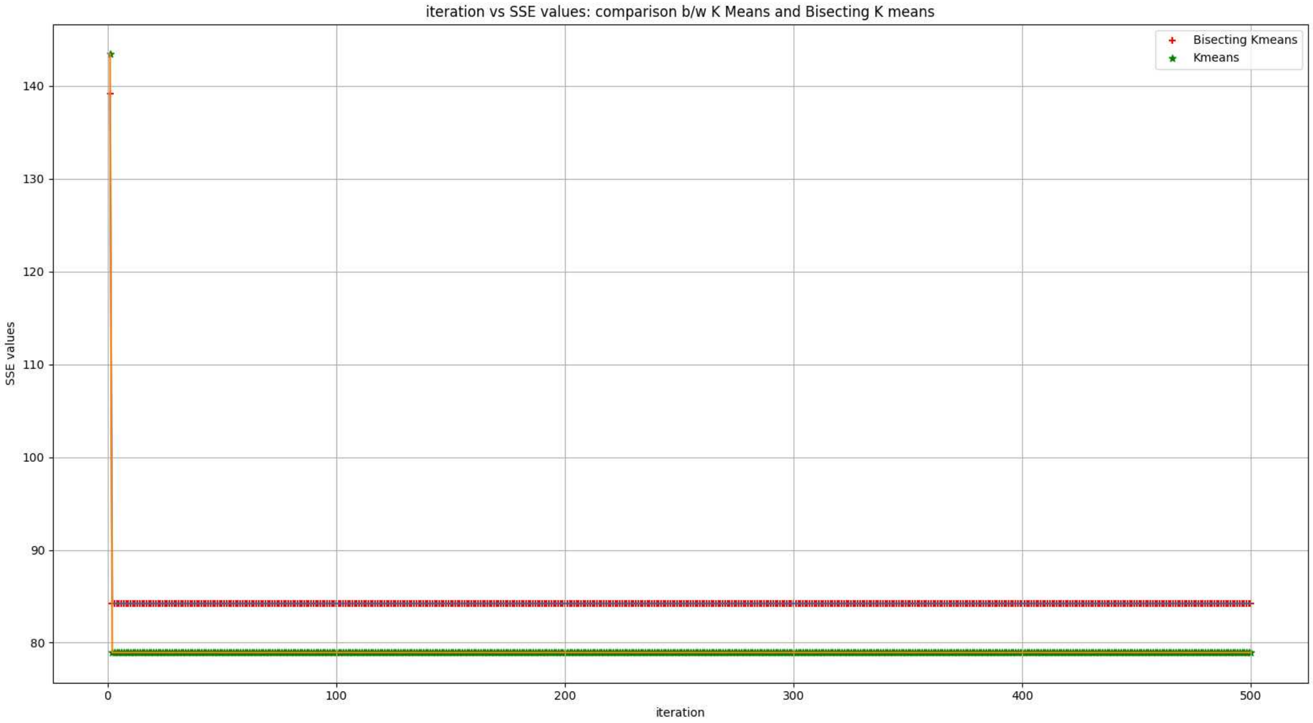




iv) The change in SSE values while running 500 iterations are as shown:



v) Comparing the SSE values of K-Means Algorithm to that of bisecting k-means shows that k-means give a better clustering solution with a lower SSE value of 78.94084142614598


iteration vs SSE values: comparison b/w K Means and Bisecting K means
Bisecting Kmeans
Kmeans
SSE values
iteration