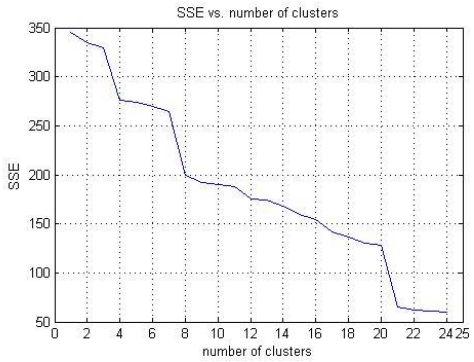
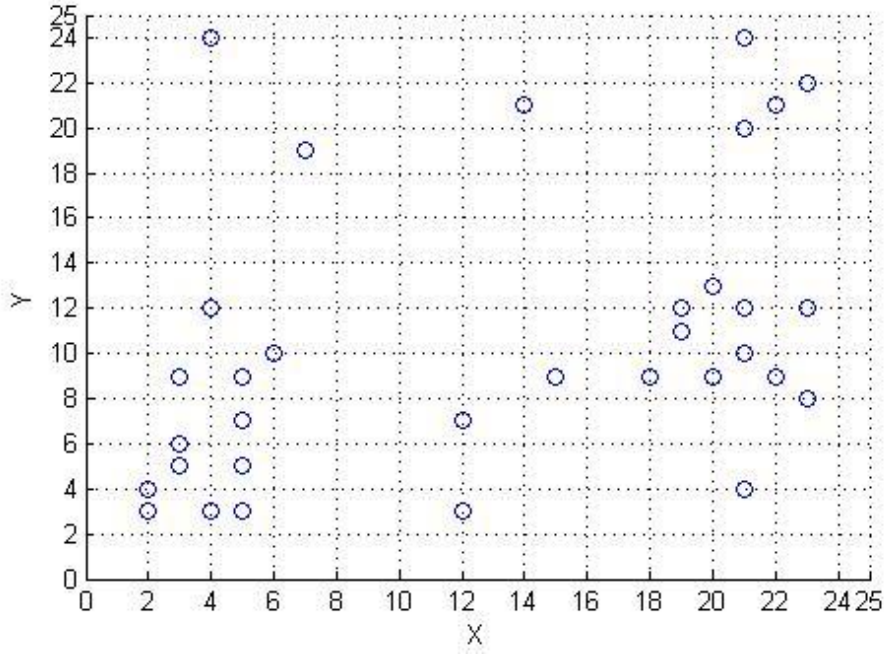
Practice Questions for Intelligent Data Analysis – Set_2

Consider the following data points: (3, 4), (4, 7), (2, 9), (2, 10), (4, 9), (7, 15), (7, -7), (6, 14), (6, -2), (4, 14), (4, -7), (10, -8) for the following questions.

1. Perform hierarchical clustering using the single-linked chain method and show the resulting dendrogram.
2. Perform hierarchical clustering using the complete-linked chain method and show the resulting dendrogram.
3. Use the basic sequential clustering algorithm and create three clusters from the above data by going from left-to-right while scanning the points.
4. Use the basic sequential clustering algorithm and create three clusters from the above data by going from right-to-left while scanning the points.
5. Use Rand index to compare the clusterings obtained in response to questions 3 and 4 above.
6. Consider the plot of the SSE values vs. the number of k-means clusters obtained from a dataset. What can you say about the number of clusters in this dataset?



7. Consider the data points shown in the figure enclosed with this test. Use an epsilon value of 3.0 and number of neighbors to be 4 and show the points that the DBSCAN algorithm will mark as core, noise, and border.



8. Discuss the difference between the WPGMA and the UPGMA clusterings for a dataset. For what types of datasets do you expect to see major difference between the two types of clusterings?

9. How is the Local Outlier Factor of a point is computed? Outline the main steps of the process, giving the main quantities computed along the way to the final LOF value.
10. What is a subspace cluster and why is it important? Give a brief outline of Cheng and Church's algorithm for finding good biclusters. What is meant when we say (In Cheng and Church algorithm) that a bicluster is good?
11.