Clustering Assignment

1. Suppose the following table shows the age of loyal customers who regularly visit an ice cream shop near the University.

Customer ID	Age
1	23
2	5
3	**22**
4	40
5	**24**

   a. Use the "complete" linkage hierarchical clustering method to cluster these customers. You need to show every step of your calculations and the distance matrices after each step.
   b. Draw a dendrogram showing the clustering process afterwards.

2. Suppose you have two attributes and their values of each data points are given as follows.

OBS#	X1	X2
1	2	22
2	25	30
3	16	7
4	4	27
5	6	24
6	18	4
7	16	2
8	29	29
9	6	30
10	27	26

   a. Plot all the points (x1, x2) in a 2-dimensional space and try identifying clusters visually. Draw a circle for each cluster so that member points are within the circle.
   b. Create three clusters using the k-means algorithm with the Euclidean distance until there are no more changes in the centroids. Pick three observations which you think are the best candidates for initial centroids. This part has to be done by hand.
   c. Create two clusters using the k-means algorithm. This part can be done with R, SAS Enterprise Miner/Enterprise Guide, or SPSS Statistics/Modeler, but use **at least two tools**, including "by hand" (manual calculations). Do you see any differences in cluster membership when you use different tools or change the options (e.g., changing the random seed/algorithm)?

3. Hands-on assignment: Use any analytics software (e.g., R, SAS Enterprise Guide/Enterprise Miner, IBM SPSS Statistic/Modeler) to complete the following exercise.

The **DUNGAREE** data set (Use dungaree.csv unless you are using SAS) gives the number of pairs of four different types of dungarees that were sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans that were sold.

Name	Model Role	Measurement Level	Description
STOREID	ID	Nominal	Identification number of the store
FASHION	Input	Interval	Number of pairs of fashion jeans sold at the store
LEISURE	Input	Interval	Number of pairs of leisure jeans sold at the store
STRETCH	Input	Interval	Number of pairs of stretch jeans sold at the store
ORIGINAL	Input	Interval	Number of pairs of original jeans sold at the store
SALESTOT	Rejected	Interval	Total number of pairs of jeans sold (the sum of **FASHION**, **LEISURE**, **STRETCH**, and **ORIGINAL**)

   a. Apply k-means algorithm to this data and present its results. At the minimum, the cluster membership should be provided. How did you determine the number of clusters (k)? Provide a justification of why the number you picked is the best.

   b. Apply "complete linkage" or "Ward" hierarchical clustering method to the same clustering problem. Compare and contrast two algorithms, i.e., k-means and hierarchical, in particular, on the quality of clusters (e.g., measured by average silhouette)