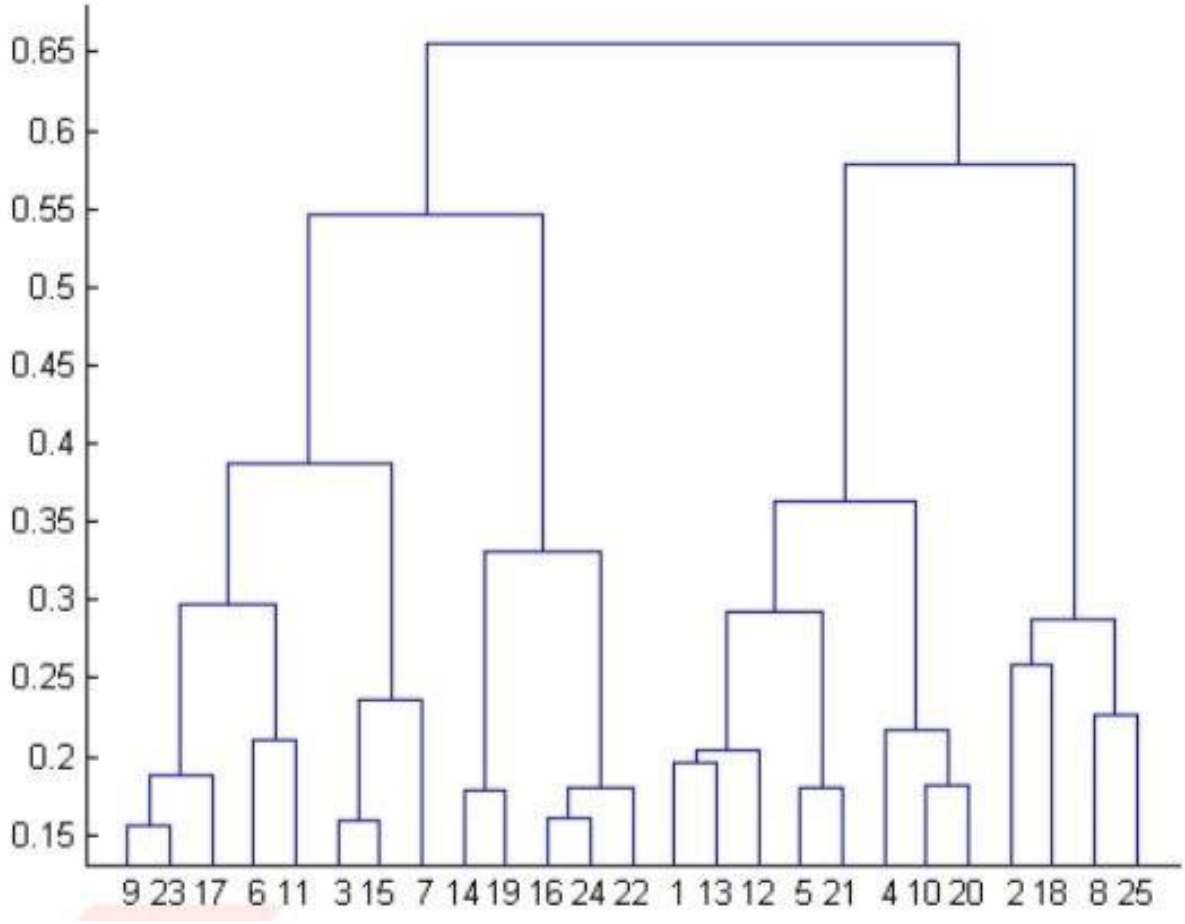
Q1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



a) 2

b) 4

c) 6

d) 8

Ans b) 4

2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers

Q2. Data points with different densities

3. Data points with round shapes

4. Data points with non-convex shapes

Options:

a) 1 and 2

b) 2 and 3

c) 2 and 4

d) 1, 2 and 4

Ans d) 1, 2 and 4

Q3. The most important part of _____ is selecting the variables on which clustering is based.

a) interpreting and profiling clusters

b) selecting a clustering procedure

c) assessing the validity of clustering

d) formulating the clustering problem

Ans d) formulating the clustering problem

Q4. The most commonly used measure of similarity is the or its square.

a) Euclidean distance

b) city-block distance

c) Chebyshev's distance

d) Manhattan distance

Ans a) Euclidean distance

Q5._____ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

a) Non-hierarchical clustering

b) Divisive clustering

c) Agglomerative clustering

d) K-means clustering

Ans b) Divisive clustering

Q6. Which of the following is required by K-means clustering?

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

d) All answers are correct

Ans d) All answers are correct

Q7. The goal of clustering is to

a) Divide the data points into groups

b) Classify the data point into different classes

c) Predict the output values of input data points

d) All of the above

Ans a) Divide the data points into groups

Q8. Clustering is a

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

Ans b) Unsupervised learning

Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a) K- Means clustering

b) Hierarchical clustering

c) Diverse clustering

d) All of the above

Ans d) All of the above

Q10. Which version of the clustering algorithm is most sensitive to outliers?

a) K-means clustering algorithm

b) K-modes clustering algorithm

c) K-medians clustering algorithm

d) None

Ans a) K-means clustering algorithm

Q11. Which of the following is a bad characteristic of a dataset for clustering analysis

a) Data points with outliers

b) Data points with different densities

c) Data points with non-convex shapes

d) All of the above

Ans d) All of the above

Q12. For clustering, we do not require

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

Ans a) Labeled data

Q13. How is cluster analysis calculated?

Ans. **There are two methods of calculating cluster analysis :**

***1.)k*-means clustering:** The first form of classification is the method called *k-means clustering* or the mobile center algorithm. As a reminder, this method aims at partitioning n n observations into k k clusters in which each observation belongs to the cluster with the closest average, serving as a prototype of the cluster. We do not go too much into details about the mathematics. Instead, we focus on how to apply it in R and by hand.

2.)Hierarchical clustering: Remind that the difference with the partition by *k*-means is that for hierarchical clustering, the number of classes is not specified in advance. Hierarchical clustering will help to determine the optimal number of clusters. Before applying hierarchical clustering by hand and in R, let's see how the ascending hierarchical clustering works step by step: 1 It starts by putting every point in its own cluster, so each cluster is a singleton 2 It then merges the 2

points that are closest to each other based on the distances from the distance matrix. The consequence is that there is one less cluster 3 It then recalculates the distances between the new and old clusters and save them in a new distance matrix which will be used in the next step 4 Finally, steps 1 and 2 are repeated until all clusters are merged into one single cluster including all points.

There are two types of Hierarchical clustering :

1.)Solution by hand : For all 3 algorithms, we first need to compute the distance matrix between the 5 points thanks to the Pythagorean theorem. Remind that the distance between point *a* and point *b* is found with: √ (x a − x b) 2 + (y a − y b) 2

2.)Solution in R :To perform the hierarchical clustering with any of the 3 criterion in R, we first need to enter the data (in this case as a matrix format, but it can also be entered as a dataframe): X <- matrix(c(2.03, 0.06, -0.64, -0.10, -0.42, -0.53, -0.36, 0.07, 1.14, 0.37), nrow = 5, byrow = TRUE)

Q14. How is cluster quality measured?

Ans. Measures for Quality of Clustering:

If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations. But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

1. Dissimilarity/Similarity metric: The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.

2. **Cluster completeness:** Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category. Let us consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg) > Q(C1, Cg).

3. Ragbag: In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category. Let us consider a clustering C1 and a cluster $C \in C1$ so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth. Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, $Q(C2, Cg) > Q(C1, Cg)$.

4. Small cluster preservation: If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive.

Suppose clustering C1 has split into three clusters, $C11 = \{d1, \ldots, dn\}$, $C12 = \{dn+1\}$, and $C13 = \{dn+2\}$. Let clustering C2 also split into three clusters, namely $C1 = \{d1, \ldots, dn-1\}$, $C2 = \{dn\}$, and $C3 = \{dn+1, dn+2\}$. As C1 splits the small category of objects and C2 splits the big category which is preferred according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, $Q(C2, Cg) > Q(C1, Cg)$.

Q15. What is cluster analysis and its types?

Ans. Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Types of Cluster Analysis Broadly, there are 2 types of cluster analysis methods. On the basis of the categorization of data sets into a particular cluster, cluster analysis can be divided into 2 types - hard and soft clustering. They are as follows –

Hard Clustering –
This implies that a hard-core classification of datasets is required in order to organize and classify data accordingly. For instance, a clustering algorithm classifies data points in one cluster such that they have the maximum similarity. However, there are no other grounds of similarity with data sets belonging to other clusters.

Soft Clustering –

The second class of cluster analysis is Soft Clustering. Unlike hard clustering that requires a given data point to belong to only a cluster at a time, soft clustering follows a different rule. In the case of soft clustering, a given data point can belong to more than one cluster at a time. This means that a fuzzy classification of datasets characterizes soft clustering.