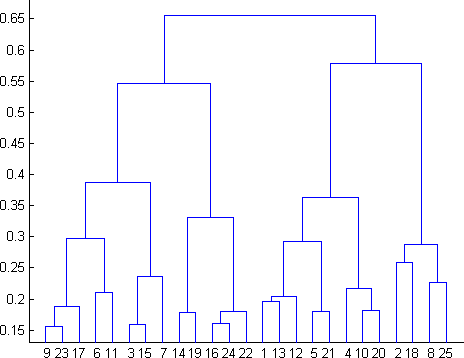
# Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

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



* 1. 2
  2. **4**
  3. 6
  4. 8

1. In which of the following cases will K-Means clustering fail to give good results?
2. Data points with outliers
3. Data points with different densities
4. Data points with round shapes
5. Data points with non-convex shapes Options:
6. 1 and 2
7. 2 and 3
8. 2 and 4
9. **1, 2 and 4**
10. The most important part of is selecting the variables on which clustering is based.
    1. interpreting and profiling clusters
    2. selecting a clustering procedure
    3. assessing the validity of clustering
    4. **formulating the clustering problem**
11. The most commonly used measure of similarity is the or its square.
    1. **Euclidean distance**
    2. city-block distance
    3. Chebyshev’s distance
    4. Manhattan distance
12. 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.
    1. Non-hierarchical clustering
    2. **Divisive clustering**
    3. Agglomerative clustering
    4. K-means clustering
13. Which of the following is required by K-means clustering?
    1. Defined distance metric
    2. Number of clusters
    3. Initial guess as to cluster centroids
    4. **All answers are correct**
14. The goal of clustering is to-
    1. **Divide the data points into groups**
    2. Classify the data point into different classes
    3. Predict the output values of input data points
    4. All of the above
15. Clustering is a-
    1. Supervised learning
    2. **Unsupervised learning**
    3. Reinforcement learning
    4. None
16. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
    1. K- Means clustering
    2. Hierarchical clustering
    3. Diverse clustering
    4. **All of the above**
17. Which version of the clustering algorithm is most sensitive to outliers?
    1. **K-means clustering algorithm**
    2. K-modes clustering algorithm
    3. K-medians clustering algorithm
    4. None
18. Which of the following is a bad characteristic of a dataset for clustering analysis-
    1. Data points with outliers
    2. Data points with different densities
    3. Data points with non-convex shapes
    4. **All of the above**
19. For clustering, we do not require-
    1. **Labeled data**
    2. Unlabeled data
    3. Numerical data
    4. Categorical data

# Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

1. How is cluster analysis calculated?

* Cluster analysis is an exploratory analysis that tries to identify structures within the data
* The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.
* K-means cluster is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

1. How is cluster quality measured?

* To measure the quality of the cluster, we can compute the average silhouette coefficient value of all objects in the cluster.

1. What is cluster analysis and its types?

* Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another.
* The common types of cluster analysis are:
  + Centroid Clustering:
    - In centroid cluster analysis you choose the number of clusters that you want to classify.
  + Density Clustering:
    - Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points, the more related they are.
  + Distribution Clustering:
    - Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.
  + Connectivity Clustering:
    - The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster.