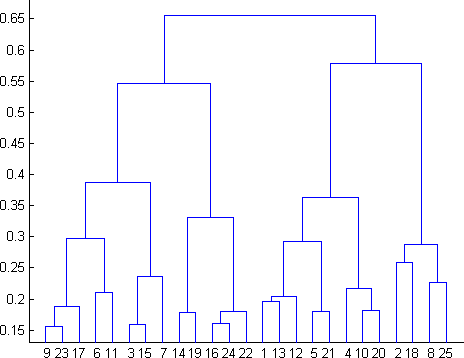
# Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

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?

Ans. [Cluster analysis](https://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/cluster-analysis-2/) is a class of techniques that are used to classify objects or cases into relative groups called clusters.  Cluster analysis is also called classification analysis or numerical taxonomy.  In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Cluster Analysis has been used in marketing for various purposes.  Segmentation of consumers in cluster analysis is used on the basis of benefits sought from the purchase of the product.  It can be used to identify homogeneous groups of buyers.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

The variables on which the cluster analysis is to be done should be selected by keeping past research in mind.  It should also be selected by theory, the hypotheses being tested, and the judgment of the researcher.  An appropriate measure of distance or similarity should be selected; the most commonly used measure is the Euclidean distance or its square.

Clustering procedures in cluster analysis may be hierarchical, non-hierarchical, or a two-step procedure.  A hierarchical procedure in cluster analysis is characterized by the development of a tree like structure.  A hierarchical procedure can be agglomerative or divisive.  Agglomerative methods in cluster analysis consist of linkage methods, variance methods, and centroid methods.  Linkage methods in cluster analysis are comprised of single linkage, complete linkage, and average linkage.

The non-hierarchical methods in cluster analysis are frequently referred to as K means clustering.  The two-step procedure can automatically determine the optimal number of clusters by comparing the values of model choice criteria across different clustering solutions.  The choice of clustering procedure and the choice of distance measure are interrelated.  The relative sizes of clusters in cluster analysis should be meaningful.  The clusters should be interpreted in terms of cluster centroids.

1. How is cluster quality measured?

Ans. The cluster quality is measured by:

SSE: sum of the square error from the items of each cluster.

Inter cluster distance: sum of the square distance between each cluster centroid.

Intra cluster distance for each cluster: sum of the square distance from the items of each cluster to its centroid.

Maximum Radius: largest distance from an instance to its cluster centroid.

Average Radius: sum of the largest distance from an instance to its cluster centroid divided by the number of clusters.

1. What is cluster analysis and its types?

Ans. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Broadly speaking, clustering can be divided into two subgroups :

* **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.
* **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

**3. Types of clustering algorithms**

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the ‘*similarity’* among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly, let’s look at them in detail:

* **Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.
* **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.
* **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.
* **Density Models:**These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.