1)What is hierarchical clustering, and how is it different from other clustering techniques?

Ans- Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

The difference between Kmeans and hierarchical clustering is that in Kmeans clustering, the number of clusters is pre-defined and is denoted by “K”, but in hierarchical clustering, the number of sets is either one or similar to the number of data observations.

2) What are the two main types of hierarchical clustering algorithms? Describe each in brief.

Ans- There are two types of hierarchical clustering: divisive (top-down) and agglomerative (bottom-up).

Agglomerative: Initially, each object is considered to be its own cluster. According to a particular procedure, the clusters are then merged step by step until a single cluster remains. ...

Divisive: The Divisive method is the opposite of the Agglomerative method.

3) How do you determine the distance between two clusters in hierarchical clustering, and what are the common distance metrics used?

Ans- For most common hierarchical clustering software, the default distance measure is the Euclidean distance. This is the square root of the sum of the square differences. However, for gene expression, correlation distance is often used.

Euclidean Distance.

Mahalanobis Distance.

Minkowski Distance.

4) How do you determine the optimal number of clusters in hierarchical clustering, and what are some

common methods used for this purpose?

Ans- To get the optimal number of clusters for hierarchical clustering, we make use a dendrogram which is tree-like chart that shows the sequences of merges or splits of clusters. If two clusters are merged, the dendrogram will join them in a graph and the height of the join will be the distance between those clusters.

5) What are dendrograms in hierarchical clustering, and how are they useful in analyzing the results?

Ans- A dendrogram is a tree-like structure that explains the relationship between all the data points in the system. However, like a regular family tree, a dendrogram need not branch out at regular intervals from top to bottom as the vertical direction (y-axis) in it represents the distance between clusters in some metric.

The main use of a dendrogram is to work out the best way to allocate objects to clusters.

6) Can hierarchical clustering be used for both numerical and categorical data? If yes, how are the distance metrics different for each type of data?

Ans- Most distance metrics, and hence the hierarchical clustering methods, work either with continuous-only or categorical-only data. In applications, however, observations are often described by a combination of both continuous and categorical variables.

7) How can you use hierarchical clustering to identify outliers or anomalies in your data?

Ans- In K-Means clustering outliers are found by distance based approach and cluster based approach. In case of hierarchical clustering, by using dendrogram outliers are found. The goal of the project is to detect the outlier and remove the outliers to make the clustering more reliable.