1. **In hierarchical clustering, how is the distance between clusters measured? Explain how this metric is used to decide when to end the iteration.**

* In hierarchical clustering, the distance between clusters is measured using various linkage methods, including:
* Single Linkage: Measures the shortest distance between any two data points in the two clusters.
* Complete Linkage: Measures the longest distance between any two data points in the two clusters.
* Average Linkage: Measures the average distance between all pairs of data points in the two clusters.
* Centroid Linkage: Measures the distance between the centroids (mean points) of the two clusters.
* Ward's Method: Minimizes the increase in the total within-cluster variance when merging clusters.

1. **In the k-means algorithm, how do you recompute the cluster centroids?**

* In the k-means algorithm, you recompute the cluster centroids as follows:
* For each cluster, calculate the mean (average) of the data points within that cluster.
* The mean becomes the new centroid for the cluster.
* Repeat this process for all clusters.
* The new centroids represent the updated cluster centers for the next iteration of the algorithm.
* Iterate until convergence, when centroids no longer change significantly, or a stopping criterion is met.

1. **At the start of the clustering exercise, discuss one method for determining the required number of clusters.**

* One method for determining the required number of clusters at the start of a clustering exercise is the "Elbow Method." It involves plotting the SSE (Sum of Squared Errors) against the number of clusters and looking for the point at which the SSE starts to level off, resembling an "elbow." The number of clusters at this "elbow" point is considered a suitable choice for the dataset.

1. **Discuss the k-means algorithm's advantages and disadvantages.**

* **Advantages of K-Means:**
* Simple and efficient: K-Means is computationally efficient and easy to implement.
* Scalability: It can handle large datasets.
* Versatile: It's applicable to various types of data.
* Interpretable: Clusters are easy to interpret based on centroid characteristics.

**Disadvantages of K-Means:**

* Sensitivity to initial centroids: Results may vary based on initial cluster centers.
* Fixed cluster number: The user must specify the number of clusters (k) in advance.
* Sensitive to outliers: Outliers can significantly impact cluster formation.
* Limited to spherical clusters: Works best when clusters are roughly spherical and equally sized.
* May converge to local optima: It might not find the globally optimal solution.