DISADVANTAGES OF CLUSTERING

1) K Means Clustering –
   - Choosing K manually.
   - Clustering data of varying sizes and density
   - Clustering Outliers.
   - Scaling with number of dimensions (Curse of dimensionality)
   - With global clusters, it didn't work well (Global means it can subdivided into two/three more cluster).
   - Works only good for Circular shape data
   - Distance based model
   - Don't work for Non-Circular shape data.
   - Lack of flexibility in clusts shape
   - Lack of probablistic clust assignment

2) Hierarchical Clustering –
   - If we have a large dataset, it become difficult to determine the correct number of clusters by dendogram.
   - Sensitivity to noise and outliers.
   - Breaking large clusters become difficult.

3) DBScan Clustering –
   - Does not work well when dealing with clusters of varying densities. While DBScan is great at seperating high density clusters from low density clusters, DBScan struggles with cluster of similar density.
   - Struggle with high dimensionality data. DBScan suffers badly with high dimension.

4) Gaussian (EM) Clustering –
   - Algorithim is very complex in nature.
   - Algorithim simply would not work for datasets where objects do not follow Gaussian distribution.
   - Distribution based model.

CHOOSING RIGHT CLUSTERING ALGORITHM IN DATASET

→ Clustering based on computation of distances between the objects of the whole dataset is called connectivity based or hierarchical. Depending on the direction of algorithm, it can unite or inversely divide the information i.e, agglomerative and divisive.

Most prominent example of connectivity based clusterization is classification of plants. The "tree" dataset starts with particular species and end with few category of plants.

→ Centroid based clustering, aimed at classifying each object of dataset to a particular cluster.

↪ When it is a spherical shaped, well seperated data then go for K means clustering.

When it is not so spherical or non-spherical data and we want probability of each data point to a cluster then use GMM clustering. But GMM will not work if data do not follow Gaussian distribution.

→ Density based clustering, DBScan best used when we have outliers in the data. DBScan can be used when data are in arbitrary shapes, and they are extremely accurate. Beside this algorithm doesn't need number of clusters from outside - it is determined automatically

→ K means clustering is also known as partition based clustering.