1. What is your definition of clustering? What are a few clustering algorithms you might think of?

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.

1. Library Installation
2. Clustering Dataset
3. Affinity Propagation
4. Agglomerative Clustering
5. BIRCH
6. DBSCAN
7. K-Means
8. Mini-Batch K-Means
9. Mean Shift
10. OPTICS
11. Spectral Clustering
12. Gaussian Mixture Model

2. What are some of the most popular clustering algorithm applications?

Clustering analysis is broadly used in many applications such as market research, **pattern recognition, data analysis, and image processing, data analysis**. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.3. When using K-Means, describe two strategies for selecting the appropriate number of clusters.

4. What is mark propagation and how does it work? Why would you do it, and how would you do it?

5. Provide two examples of clustering algorithms that can handle large datasets. And two that look for high-density areas?

6. Can you think of a scenario in which constructive learning will be advantageous? How can you go about putting it into action?

7. How do you tell the difference between anomaly and novelty detection?

In "novelty detection", you have a **data set that contains only good data**, and you're trying to determine whether new observations fit within the existing data set. In "outlier detection", the data may contain outliers, which you want to identify.

The training data contains outliers which are defined as observations that are far from the others. ... In this context an outlier is also called a novelty. Outlier detection and novelty detection are both used for anomaly detection, where one is interested in **detecting abnormal or unusual observations**.

8. What is a Gaussian mixture, and how does it work? What are some of the things you can do about it?

A Gaussian mixture distribution is **a multivariate distribution that consists of multivariate Gaussian distribution components**. Each component is defined by its mean and covariance, and the mixture is defined by a vector of mixing proportions.

Gaussian Mixture Models (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends **to group the data points belonging to a single distribution together**.

Gaussian Mixture models are used for **representing Normally Distributed subpopulations within an overall population**. The advantage of Mixture models is that they do not require which subpopulation a data point belongs to. It allows the model to learn the subpopulations automatically.

Probabilistic mixture models such as Gaussian mixture models (GMM) are used **to resolve point set registration problems in image processing and computer vision fields**.

9. When using a Gaussian mixture model, can you name two techniques for determining the correct number of clusters?

1) Silohuette score

2) Distance Between GMMs

3) Gradient of BIC scores