1)What are the different types of clustering algorithms, and how do they differ in terms of their approach and underlying assumptions?

Ans- Clustering methods are mainly divided into density-based, partition-based, hierarchical, and model-based clustering

Centroid-based clustering organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering defined below. k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.

2) What is K-means clustering, and how does it work?

Ans- K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset.

Scales to large data sets. Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

3) What are some advantages and limitations of K-means clustering compared to other clustering techniques?

Ans- Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Limitations:-

It requires to specify the number of clusters (k) in advance.

It can not handle noisy data and outliers.

It is not suitable to identify clusters with non-convex shapes.

4) How do you determine the optimal number of clusters in K-means clustering, and what are some common methods for doing so?

Ans- The silhouette coefficient may provide a more objective means to determine the optimal number of clusters. This is done by simply calculating the silhouette coefficient over a range of k, & identifying the peak as optimum K.

Probably the most well known method, the elbow method, in which the sum of squares at each number of clusters is calculated and graphed, and the user looks for a change of slope from steep to shallow (an elbow) to determine the optimal number of clusters.

5) What are some applications of K-means clustering in real-world scenarios, and how has it been used to solve specific problems?

Ans- Real-life examples include spam detection, sentiment analysis, scorecard prediction of exams, etc. 2) Regression Models – Regression models are used for problems where the output variable is a real value such as a unique number, dollars, salary, weight or pressure, for example.

1. Step-1: Select the number K to decide the number of clusters.
2. Step-2: Select random K points or centroids. ...
3. Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
4. Step-4: Calculate the variance and place a new centroid of each cluster.

6) How do you interpret the output of a K-means clustering algorithm, and what insights can you derive from the resulting clusters?

Ans- Interpreting the meaning of k-means clusters boils down to characterizing the clusters. A Parallel Coordinates Plot allows us to see how individual data points sit across all variables. By looking at how the values for each variable compare across clusters, we can get a sense of what each cluster represents.

7) What are some common challenges in implementing K-means clustering, and how can you address them?

Ans- k-Means doesn't perform well if the clusters have varying sizes, different densities, or non-spherical shapes. Has to be run for a certain amount of iteration or it would produce a suboptimal result. Computationally expensive as distance is to be calculated from each centroid to all data points.

k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means as described in the Advantages section. Clustering outliers. Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored.