## Q1

Difference between Supervised and Unsupervised Learning:

Supervised learning involves training a model on a labeled dataset, where the algorithm learns to make predictions based on input-output pairs. Examples include image classification, spam email detection, and predicting house prices from features.

Unsupervised learning deals with unlabeled data, where the algorithm identifies patterns, structures, or clusters within the data. Examples include customer segmentation, anomaly detection, and topic modeling.

## Q2

Unsupervised Learning Applications:

Unsupervised learning finds applications in various fields:

Clustering: Grouping similar data points, e.g., customer segmentation for targeted marketing.

Dimensionality Reduction: Reducing the number of features while preserving essential information, e.g., PCA in image compression.

Anomaly Detection: Identifying rare or unusual data points, e.g., fraud detection.

Density Estimation: Estimating the probability distribution of data, e.g., in outlier detection.

Recommendation Systems: Suggesting products or content based on user behavior.

## Q3

Three Main Types of Clustering Methods:

Hierarchical Clustering: Builds a tree-like hierarchy of clusters, allowing for different levels of granularity. Characteristics include dendrogram visualization and agglomerative (bottom-up) or divisive (top-down) approaches.

Partitioning Clustering (e.g., k-means): Divides data into non-overlapping clusters, with each data point belonging to a single cluster. It's characterized by specifying the number of clusters (k).

Density-Based Clustering (e.g., DBSCAN): Identifies clusters as dense regions separated by sparser areas. It's characterized by defining clusters based on data point density.

## Q4

K-Means Algorithm for Clustering Consistency:

The k-means algorithm minimizes the sum of squared distances (SSE) between data points and the centroid of their assigned cluster.

Consistency is determined by iterating until the centroids no longer change significantly between iterations, indicating that the clusters have stabilized.

## Q5

Difference Between K-Means and K-Medoids:

K-Means uses cluster centroids (mean of data points) to represent clusters, making it sensitive to outliers. K-Medoids uses the actual data points (medoids) as cluster representatives, making it more robust to outliers.

## Q6

Dendrogram:

A dendrogram is a tree-like diagram used in hierarchical clustering to visualize the arrangement of data points into clusters.

The algorithm starts with each data point as its cluster and then iteratively merges clusters based on similarity, forming a hierarchy.

It helps in understanding the hierarchical structure of data clusters.

## Q7

SSE (Sum of Squared Errors):

SSE is a measure of the within-cluster variance. It calculates the sum of squared distances of data points within each cluster to their cluster's centroid.

In k-means, SSE is minimized to achieve compact and well-separated clusters. Lower SSE indicates better clustering.

## Q8

K-Means Procedure:

Initialize k cluster centroids randomly.

Assign each data point to the nearest centroid.

Recalculate centroids as the mean of data points in each cluster.

Repeat assignment and centroid update until convergence (minimal change in centroids).

## Q9

Single Link vs. Complete Link (Hierarchical Clustering):

Single Link (Minimum Linkage): Measures similarity between clusters by the shortest distance between any two points in different clusters.

Complete Link (Maximum Linkage): Measures similarity by the longest distance between points in different clusters.

## Q10

Apriori in Basket Analysis:

Apriori is an association rule mining algorithm used in market basket analysis.

It reduces measurement overhead by identifying frequent itemsets (combinations of items) and generating association rules.

For example, if customers often buy items A and B together, the algorithm can recommend item B when item A is purchased, reducing the need to track all possible item pairs.