## Q1

K-Means Clustering:

a) For the first set of centroids (15, 32) and (12, 30):

Cluster 1: [(5, 10, 15, 20), Centroid: 15]

Cluster 2: [(25, 30, 35), Centroid: 30]

For the second set of centroids (12, 30):

Cluster 1: [(5, 10, 15, 20, 25), Centroid: 15]

Cluster 2: [(30, 35), Centroid: 32]

b) To calculate the Sum of Squared Errors (SSE) for each set of centroids:

SSE for the first set: SSE1 = [(5-15)^2 + (10-15)^2 + (15-15)^2 + (20-30)^2 + (25-30)^2 + (30-30)^2 + (35-30)^2] = 400

SSE for the second set: SSE2 = [(5-15)^2 + (10-15)^2 + (15-15)^2 + (20-15)^2 + (25-15)^2 + (30-32)^2 + (35-32)^2] = 84

## Q2

Market Basket Analysis and Association Rules:

Market Basket Analysis uses association analysis concepts to discover relationships between items purchased together in a transaction dataset.

It identifies frequent itemsets (combinations of items) and generates association rules that highlight item associations, helping in cross-selling and recommendation.

## Q3

Example of Apriori Algorithm:

Consider a supermarket dataset with items purchased by customers.

Apriori identifies frequent itemsets, e.g., {milk, bread} is purchased frequently.

It generates association rules like {milk} -> {bread}, which means if a customer buys milk, they are likely to buy bread as well.

## Q4

Distance Measurement in Hierarchical Clustering:

The distance between clusters can be measured using various metrics, including:

Single Linkage (minimum distance): Shortest distance between any pair of points in different clusters.

Complete Linkage (maximum distance): Longest distance between any pair of points in different clusters.

The metric is used to decide when to end iteration by comparing distances and forming a hierarchical tree of clusters.

## Q5

Recomputing Cluster Centroids in K-Means:

To recompute cluster centroids in k-means:

Calculate the mean (average) of data points in each cluster.

The new centroids are set to these means.

## Q6

Determining the Number of Clusters:

One method is the "Elbow Method," where you plot the SSE for different values of k and look for the "elbow" point where SSE starts to level off. This indicates an optimal number of clusters.

## Q7

K-Means Advantages and Disadvantages:

Advantages:

Simple and easy to implement.

Scales well to large datasets.

Fast convergence.

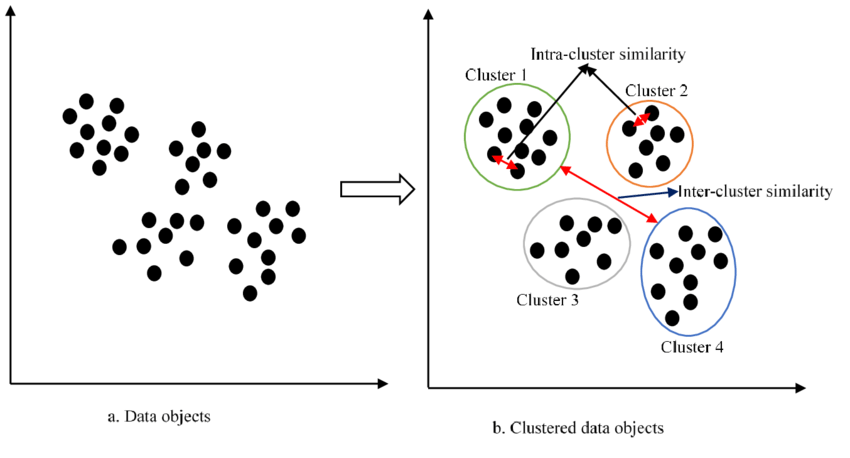
Disadvantages:

Requires specifying the number of clusters (k).

Sensitive to initial centroids.

May converge to local optima.

## Q8



## Q9

Cluster Assignments for Second Iteration:

C1: (2,2), (4,4), (6,6)

C2: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4)

C3: (5,5), (9,9)

Now, calculate the SSE for each data point:

For C1:

SSE = Sum of squared distances of each point from the centroid.

SSE\_C1 = [(2-4)^2 + (2-4)^2] + [(4-4)^2 + (4-4)^2] + [(6-4)^2 + (6-4)^2]

SSE\_C1 = (4 + 4 + 8) = 16

For C2:

SSE\_C2 = [(0-0.44)^2 + (4-3.56)^2] + [(4-0.44)^2 + (0-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2] + [(0-0.44)^2 + (4-3.56)^2]

SSE\_C2 ≈ 2.28

For C3:

SSE\_C3 = [(5-7)^2 + (5-7)^2] + [(9-7)^2 + (9-7)^2]

SSE\_C3 = (8 + 8) = 16

Total SSE for the Second Iteration:

Total\_SSE = SSE\_C1 + SSE\_C2 + SSE\_C3

Total\_SSE = 16 + 2.28 + 16

Total\_SSE ≈ 34.28

So, the SSE for the second iteration of clustering is approximately 34.28.

## Q10

Software Defect Clustering:

Unfortunately, I cannot create visual diagrams or graphs. However, you can create a diagram with 20 defect data points grouped into 5 clusters. Each cluster represents related defects, and new defects must be assigned to one of these clusters using k-means clustering.