Ques.1 A set of one-dimensional data points is given to you: 5, 10, 15, 20, 25, 30, 35. Assume that k = 2 and that the first set of random centroid is 15, 32, and that the second set is 12, 30.

a) Using the k-means method, create two clusters for each set of centroid described above.
b) b) For each set of centroid values, calculate the SSE.

Using the first set of centroids: Centroid 1: 15, Centroid 2: 32

- Cluster 1: {5, 10, 15, 20, 25}
- Cluster 2: {30, 35}

Calculating SSE for the first set of centroids:

- SSE = (5-15)^2 + (10-15)^2 + (15-15)^2 + (20-15)^2 + (25-15)^2 + (30-32)^2 + (35-32)^2
- SSE = 50 + 25 + 0 + 25 + 100 + 4 + 9 SSE = 213

Using the second set of centroids: Centroid 1: 12, Centroid 2: 30

- Cluster 1: {5, 10, 15}
- Cluster 2: {20, 25, 30, 35}

Calculating SSE for the second set of centroids:

- SSE = (5-12)^2 + (10-12)^2 + (15-12)^2 + (20-30)^2 + (25-30)^2 + (30-30)^2 + (35-30)^2
- SSE = 49 + 4 + 9 + 100 + 25 + 0 + 25 SSE = 212

2.Describe how the Market Basket Research makes use of association analysis concepts.

Market Basket Research makes use of association analysis concepts to discover patterns and relationships between items in a transactional dataset. It aims to identify associations or co-occurrence of items that are frequently purchased together.

Association analysis is based on the concept of frequent itemsets and association rules. Frequent itemsets are sets of items that appear together in a significant number of transactions. Association rules indicate the relationships between items and can be used to make predictions or recommendations.

For example, in a supermarket, if the association analysis identifies that customers who buy bread also tend to buy butter with a high confidence level, the supermarket can place bread and butter close to each other to encourage additional sales.

3.Give an example of the Apriori algorithm for learning association rules.

Example of the Apriori algorithm for learning association rules: Consider a transactional dataset with the following items: Transaction 1: {A, B, C, D} Transaction 2: {A, C, D, E} Transaction 3: {B, D, E} Transaction 4: {A, C, E}

Step 1: Generate frequent itemsets of size 1: Frequent 1-itemsets: {A}, {B}, {C}, {D}, {E}

Step 2: Generate frequent itemsets of size 2: Join and prune step: Candidate 2-itemsets: {A, B}, {A, C}, {A, D}, {A, E}, {B, C}, {B, D}, {B, E}, {C, D}, {C, E}, {D, E} Prune step: Remove candidate itemsets that have subsets that are not frequent.

Step 3: Generate frequent itemsets of size 3: Join and prune step: Candidate 3-itemsets: {A, B, C}, {A, B, D}, {A, B, E}, {A, C, D}, {A, C, E}, {A, D, E}, {B, C, D}, {B, C, E}, {B, D, E}, {C, D, E} Prune step: Remove candidate itemsets that have subsets that are not frequent.

Step 4: Generate association rules: From the frequent itemsets, we can generate association rules by setting minimum support and confidence thresholds. For example, if we set a minimum support of 2 and a minimum confidence of 0.5, we can generate the following association rules: {A, B} -> {C}, {A, C} -> {D}, {A, C} -> {E}, {A, D} -> {C}, {A, D} -> {E}, {B, D} -> {C}, {B, D} -> {E}, {C, D} -> {E}

4. In hierarchical clustering, how is the distance between clusters measured? Explain how this metric is used to decide when to end the iteration.

In hierarchical clustering, the distance between clusters is measured using different metrics, such as Euclidean distance, Manhattan distance, or correlation distance. The choice of distance metric depends on the nature of the data and the clustering objective.

The distance metric is used to decide when to end the iteration by defining a stopping criterion. One common stopping criterion is to use a predetermined threshold distance value. If the distance between the clusters being merged exceeds the threshold, the iteration stops, and the clusters are considered separate. Another criterion is to specify the desired number of clusters in advance, and the iteration stops when that number is reached.

The specific stopping criterion and the distance metric used can significantly affect the resulting clustering structure and should be chosen based on the characteristics of the data and the clustering goals.

5.In the k-means algorithm, how do you recompute the cluster centroids?

In the k-means algorithm, the cluster centroids are recomputed by calculating the mean of the data points within each cluster. The steps to recompute the cluster centroids are as follows:

1. Assign each data point to the nearest cluster centroid based on the Euclidean distance.
2. For each cluster, calculate the mean of the coordinates of the data points assigned to that cluster.
3. Update the cluster centroids with the newly calculated means.

4. Repeat steps 1-3 until convergence, where the cluster assignments and centroids no longer change significantly or a maximum number of iterations is reached.

By recomputing the cluster centroids based on the mean of the data points, the algorithm aims to find the center of each cluster, minimizing the distance between the data points and their respective centroids.

6. At the start of the clustering exercise, discuss one method for determining the required number of clusters.

Determining the required number of clusters in a clustering exercise can be challenging. One method to estimate the optimal number of clusters is the "elbow method."

The elbow method involves plotting the number of clusters against a measure of the clustering quality, such as the within-cluster sum of squares (WCSS) or the average distance between data points and their centroid. The plot forms an elbow-like shape, and the optimal number of clusters is typically located at the "elbow" point, where the improvement in clustering quality starts to diminish significantly.

To determine the required number of clusters using the elbow method, you would calculate the clustering quality measure for different numbers of clusters (e.g., 2, 3, 4, ...) and observe the plot. The number of clusters at the elbow point is often chosen as the optimal number for clustering. However, it is important to consider domain knowledge and interpretability of the results when deciding on the final number of clusters.

7.Discuss the k-means algorithm's advantages and disadvantages.

Advantages of the k-means algorithm:

- Relatively easy to understand and implement.
- Efficient and scalable for large datasets.
- Provides well-separated and compact clusters.
- Works well when the clusters have a roughly spherical shape.

Disadvantages of the k-means algorithm:

- Requires the number of clusters to be predefined.
- Sensitive to the initial placement of cluster centroids, which can lead to different results.
- May converge to suboptimal solutions or get stuck in local optima.
- Not suitable for datasets with irregularly shaped or non-linearly separable clusters.
- Affected by outliers, which can distort cluster boundaries.

8. Draw a diagram to demonstrate the principle of clustering.

9. During your study, you discovered seven findings, which are listed in the data points below. Using the K-means algorithm, you want to build three clusters from these observations. The clusters C1, C2, and C3 have the following findings after the first iteration:

C1: (2,2), (4,4), (6,6); C2: (2,2), (4,4), (6,6); C3: (2,2), (4,4),

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

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

What would the cluster centroids be if you were to run a second iteration? What would this clustering's SSE be?

In the diagram, we have four data points represented as squares. Clustering aims to group similar data points together. The clusters can be visualized as enclosed regions or by

assigning colors to the data points belonging to the same cluster. The goal is to find clusters that minimize the distances or dissimilarities within clusters and maximize the distances between clusters.

9. After the first iteration:

- Cluster C1: (2,2), (4,4), (6,6)
- Cluster C2: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4), (0
- Cluster C3: (5,5) and (9,9)

For the second iteration, we recompute the cluster centroids by calculating the mean of the data points within each cluster:

- Cluster C1 centroid: (2+4+6)/3, (2+4+6)/3 = (4, 4)
- Cluster C2 centroid: (0+4+0+0+0+0+0+0+0)/9, (4+0+4+4+4+4+4+4+4)/9 = (1.33, 3.56)
- Cluster C3 centroid: (5+9)/2, (5+9)/2 = (7, 7)

The SSE for this clustering can be calculated as the sum of the squared Euclidean distances between each data point and its assigned centroid:

SSE = (2-4)^2 + (4-4)^2 + (6-4)^2 + (0-1.33)^2 + (4-1.33)^2 + (0-1.33)^2 + (0-1.33)^2 + (0-1.33)^2 + (0-1.33)^2 + (0-1.33)^2 + (0-1.33)^2 + (0-1.33)^2 + (5-7)^2 + (9-7)^2

SSE = 8 + 0 + 8 + 1.775 + 6.55 + 1.775 + 1.775 + 1.775 + 1.775 + 1.775 + 1.775 + 1.775 + 1.775 + 1.775 + 4 + 4 SSE = 48.275

10. In a software project, the team is attempting to determine if software flaws discovered during testing are identical. Based on the text analytics of the defect details, they decided to build 5 clusters of related defects. Any new defect formed after the 5 clusters of defects have been identified must be listed as one of the forms identified by clustering. A simple diagram can be used to explain this process. Assume you have 20 defect data points that are clustered into 5 clusters and you used the k-means algorithm.