**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) For each set of centroid values, calculate the SSE.**

Ans-a) Using the given k-means method and initial centroids, two clusters can be created as follows:

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

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

b) The SSE (sum of squared errors) for each set of centroid values can be calculated as follows:

Set 1: SSE = (5-15)^2 + (10-15)^2 + (15-15)^2 + (20-15)^2 + (25-15)^2 + (30-32)^2 + (35-32)^2 = 200

Set 2: SSE = (5-12)^2 + (10-12)^2 + (15-12)^2 + (20-30)^2 + (25-30)^2 + (30-30)^2 + (35-30)^2 = 310

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

Ans-Market Basket Analysis is a technique that uses association rules to find relationships between items that are frequently bought together. It analyzes customers' purchase patterns to identify items that are frequently purchased together and to find the most common patterns. Association analysis concepts, such as support, confidence, and lift, are used to identify the most significant rules and help businesses make informed decisions about marketing, inventory management, and product placement.

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

Ans-The Apriori algorithm is a classic algorithm for learning association rules. Here is an example of the algorithm:

1. Given a set of transactions T, and a minimum support threshold min\_sup.
2. Generate all frequent itemsets of size k by scanning T once.
3. Use the frequent itemsets of size k to generate candidate itemsets of size k+1.
4. Scan T again to count the support of each candidate itemset.
5. Prune the candidate itemsets that do not meet the minimum support threshold.
6. Repeat steps 3-5 until no more frequent itemsets can be generated.

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

Ans-In hierarchical clustering, the distance between clusters is measured using various metrics, such as Euclidean distance, Manhattan distance, and cosine similarity. The metric used determines the similarity between clusters and can be used to decide when to end the iteration. One common stopping criterion is to stop the iteration when the distance between the clusters exceeds a predefined threshold or when the desired number of clusters has been reached.

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

Ans-In the k-means algorithm, the cluster centroids are recomputed as follows:

For each cluster, compute the mean of all the data points in the cluster.

Set the mean as the new centroid for the cluster.

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

Ans-One method for determining the required number of clusters is the Elbow Method. This method involves plotting the SSE against the number of clusters and selecting the number of clusters at the "elbow" of the curve, where the rate of decrease in SSE slows down significantly. This method helps to find a balance between the number of clusters and the level of intra-cluster similarity.

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

Ans-Advantages of the k-means algorithm include its simplicity, efficiency, and effectiveness in handling large datasets. However, it is sensitive to the initial random centroids and may converge to a local minimum instead of the global minimum. Additionally, the algorithm requires the number of clusters to be specified beforehand.

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

Ans-A diagram to demonstrate the principle of clustering could be a scatter plot with data points representing different observations. The points are then grouped into clusters based on their similarity, with each cluster represented by a different colour or shape.

**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?**

Ans-After the first iteration, the clusters C1 and C2 contain the same data points, and C3 contains two data points. Therefore, if a second iteration were to be run, the new cluster centroids would be:

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: (5.5,5.5), (9,9)

The SSE for this clustering can be calculated by summing the squared distances between each data point and its cluster centroid, as follows:

SSE = (0-2)^2 + (4-4)^2 + (0-6)^2 + (4-2)^2 + (0-4)^2 + (4-6)^2 + (0-5.5)^2 + (4-5.5)^2 + (0-9)^2

= 70.5

**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.**

Ans-K-means algorithm is an unsupervised machine learning algorithm used for clustering similar data points. In this case, the team has used the k-means algorithm to cluster 20 defect data points into 5 clusters based on the similarity of their defect details.

Once the clusters have been formed, any new defect discovered during testing must be added to one of the five clusters that have already been identified. This ensures that there are no new clusters formed and that all defects are classified into one of the five clusters.

To represent this visually, we can use a simple diagram. The 20 defect data points can be represented as 20 dots on a graph, and the five clusters can be represented as different coloured circles around the dots that belong to each cluster. Any new defect discovered during testing can be added to the cluster that is closest to it based on its defect details.