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.

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

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

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

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

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

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

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?

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.

* **K-Means Clustering:** a. **Centroids:** Set 1: {15, 32}, Set 2: {12, 30} b. **Clusters and SSE:**
* Set 1 Clusters: C1 (15), C2 (32)
* Set 2 Clusters: C1 (15), C2 (30)
* **SSE Calculation:**
* Set 1 SSE: *(15−15)2+(32−15)2=0+172=289*(15−15)2+(32−15)2=0+172=289
* Set 2 SSE: *(15−15)2+(30−15)2=0+152=225*(15−15)2+(30−15)2=0+152=225
* **Market Basket Research and Association Analysis:**
* **Concepts Used:**
* **Support:** Measures the frequency of co-occurrence of items in transactions.
* **Confidence:** Measures the likelihood that the presence of one item implies the presence of another.
* **Lift:** Measures the extent to which the presence of one item influences the presence of another, compared to random chance.
* **Example of Apriori Algorithm:**
* Consider a dataset of customer transactions with items {A, B, C, D}.
* **Support:** Minimum support is set (e.g., 0.2).
* **Frequent Itemsets:** Frequent itemsets are generated (e.g., {A}, {B}, {C}).
* **Association Rules:** Rules are generated based on confidence (e.g., {A} => {B} with confidence > 0.5).
* **Hierarchical Clustering:**
* **Distance Between Clusters:** Common metrics include Euclidean distance or linkage methods like Ward's method.
* **Iteration Stopping Criterion:** Decide when to end iteration based on a predetermined number of clusters or when the distance between clusters exceeds a threshold.
* **Recomputing K-Means Cluster Centroids:**
* For each cluster, compute the mean of data points along each dimension.
* New centroids are the means of data points in their respective clusters.
* **Determining Required Number of Clusters:**
* Methods include the Elbow Method or Silhouette Analysis.
* **Elbow Method:** Plot SSE against the number of clusters and identify the "elbow" point where SSE starts decreasing at a slower rate.
* **Advantages and Disadvantages of K-Means:**
* **Advantages:**
* Simple and easy to implement.
* Scales well to large datasets.
* **Disadvantages:**
* Sensitive to initial centroids.
* May converge to local minima.
* Assumes clusters of similar size and variance.
* **Diagram Illustrating Clustering Principle:**
* Cluster Diagram
* **K-Means Second Iteration:**
* Centroids:
* C1: (4,4)
* C2: (0,2)
* C3: (7,7)
* SSE: Sum of squared distances within clusters.
* **Software Defect Clustering Process:**
* Cluster Diagram
* Represents 20 defect data points clustered into 5 clusters using the k-means algorithm.