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.

In market basket analysis, **association rules are used to predict the likelihood of products being purchased together**. Association rules count the frequency of items that occur together, seeking to find associations that occur far more often than expected.

Market Basket Analysis is **modeled on Association rule mining**, i.e., the IF {}, THEN {} construct. For example, IF a customer buys bread, THEN he is likely to buy butter as well.

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

Apriori is an **algorithm for frequent item set mining and association rule learning over relational databases**. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

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

For most common hierarchical clustering software, the default distance measure is **the Euclidean distance**. This is the square root of the sum of the square differences. However, for gene expression, correlation distance is often used. The distance between two vectors is 0 when they are perfectly correlated.

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

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.

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

**The “Elbow” Method**  
  
Probably the most well known method, the elbow method, in which the sum of squares at each number of clusters is calculated and graphed, and the user looks for a change of slope from steep to shallow (an elbow) to determine the optimal number of clusters.

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

**Advantages of k-means**

Relatively simple to implement.

Scales to large data sets.

Guarantees convergence.

Can warm-start the positions of centroids.

Easily adapts to new examples.

Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

**Disadvantages of k-means**

Choosing k manually.

Being dependent on initial values.

Clustering data of varying sizes and density.

Clustering outliers.

Scaling with number of dimensions.

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.