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.

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

* Apriori Algorithm is one of the algorithm used for transaction data in Association Rule Learning. It allows us to mine the frequent itemset in order to generate association rule between them.
* Example: list of items purchased by customers, details of website which are frequently visited etc.

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

* The linkage methods work by calculating the distances or similarities between all objects. Then the closest pair of clusters are combined into a single cluster, reducing the number of clusters remaining. The process is then repeated until there is only a single cluster left.

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

* Recalculate the cluster centers by taking the average/mean of all points in a cluster. These might or might not be in the data set. Calculate the sum of distances of the new centers from old ones and lets call it error.One can follow these steps:

Step 1: Choose the number of clusters k.

Step 2: Select k random points from the data as centroids.

Step 3: Assign all the points to the closest cluster centroid.

Step 4: Recompute the centroids of newly formed clusters.

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

* 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

**Being dependent on initial values.**

For a low k, you can mitigate this dependence by running k-means several times with different initial values and picking the best result. As k increases, you need advanced versions of k-means to pick better values of the initial centroids (called **k-means seeding**).

**Clustering outliers.**

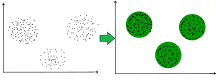
Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.

**Scaling with number of dimensions.**

As the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples. Reduce dimensionality either by using [**PCA**](https://wikipedia.org/wiki/Principal_component_analysis) on the feature data, or by using “spectral clustering” to modify the clustering algorithm as explained below.

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

* Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.



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.