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:

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

Ans: 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.

Algorithms that use association rules include AIS, SETM and Apriori. The Apriori algorithm is commonly cited by data scientists in research articles about market basket analysis and is used to identify frequent items in the database, then evaluate their frequency as the datasets are expanded to larger sizes.

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

Ans: The Amazon website employs a well-known example of market basket analysis. On a product page, Amazon presents users with related products, under the headings of “Frequently bought together” and “Customers who bought this item also bought.” Association rules can be thought of as an IF-THEN relationship. Suppose item **A** is being bought by the customer, then the chances of item **B** being picked by the customer too under the same **Transaction ID** is found out.

[](https://camo.githubusercontent.com/a77ee4346ac4cecab2e18c32af643690c21950b733e5b03654ec7f2f56aeadba/68747470733a2f2f692e696d6775722e636f6d2f54674c4e7245542e706e67)

**There are two elements of these rules:**

**Antecedent (IF)**: This is an item/group of items that are typically found in the Itemsets or Datasets.

**Consequent (THEN)**: This comes along as an item with an Antecedent/group of Antecedents.

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: Clusters are merged based on the distance between them and to calculate the distance between the clusters we have different types of linkages.

**Ward’s linkage**: Minimizes the variance of the clusters being merged. Least increase in total variance around cluster centroids is aimed.

**Average linkage:** Average distance of each data point in two clusters.

**Complete (maximum) linkage:** Maximum distance among all data points in two clusters.

**Single (minimum) linkage:** Maximum distance among all data points in two clusters.

This metric is used to decide when to end the iteration.

* Stop after a number of clusters is reached (**n\_clusters**)
* Set a threshold value for linkage (**distance\_threshold**). If the distance between two clusters are above the threshold, these clusters will not be merged.

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

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

Step-2: Select random K points or centroids. ...

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.

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

Ans: Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well data points are clustered with other data points that are similar to each other. This method can be used to find the optimal value of ‘k’. This score is within the range of [-1,1]. The value of ‘k’ having the silhouette score nearer to 1 can be considered as the ‘right’ number of clusters.

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

## Ans: Advantages of k-means

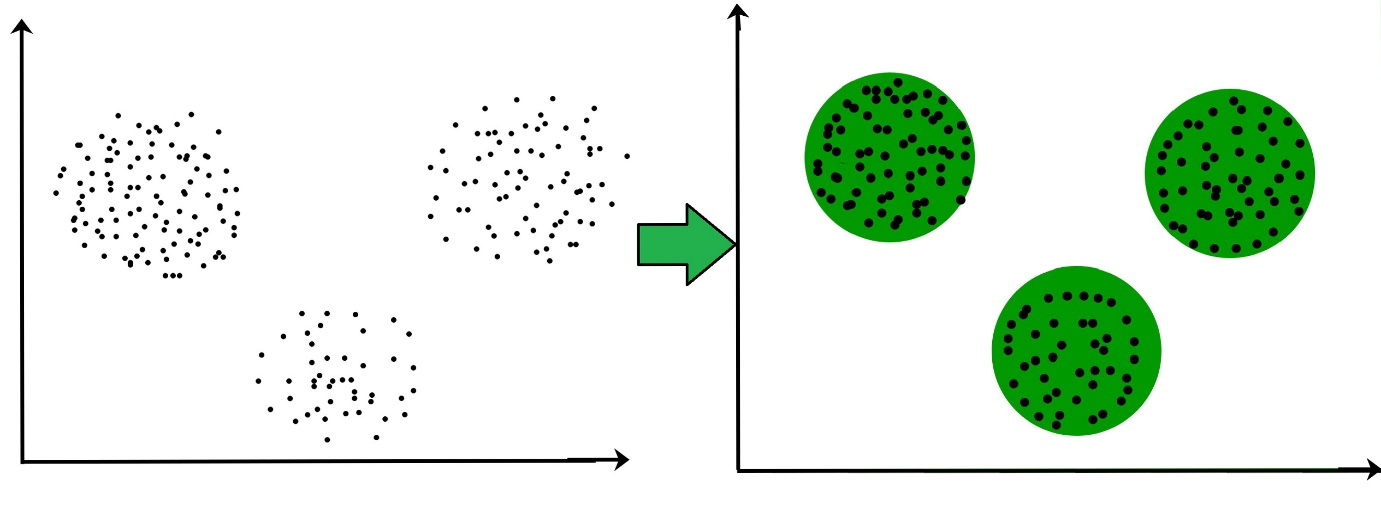
* Relatively simple to implement.
* Scales to large data sets.
* Guarantees convergence.
* Can warm-start the positions of centroids.
* Easily adapts to new examples.
* Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

## Disadvantages of k-means

* Choosing k  manually:Use the “Loss vs. Clusters” plot to find the optimal (k), as discussed in [Interpret Results](https://developers.google.com/machine-learning/clustering/interpret).
* Being dependent on initial values.
* Clustering data of varying sizes and density:k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means as described in the [Advantages](https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages#advantages_of_k-means) section.
* 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.

Ans: It is basically a type of [unsupervised learning method](https://www.geeksforgeeks.org/supervised-unsupervised-learning/). An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. **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. It is basically a collection of objects on the basis of similarity and dissimilarity between them. The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



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:

1.C1: (4,4), C2: (2,2), C3: (7,7)

2. C1: (6,6), C2: (4,4), C3: (9,9)

3. C1: (2,2), C2: (0,0), C3: (5,5)

4. C1: (2,6), C2: (0,4), C3: (5,9)

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.