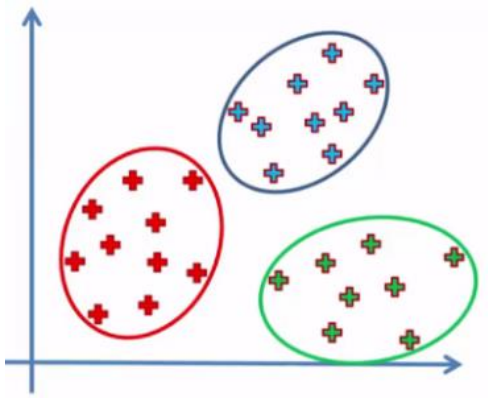
Clustering Assignment: Part II

Question 2: Clustering

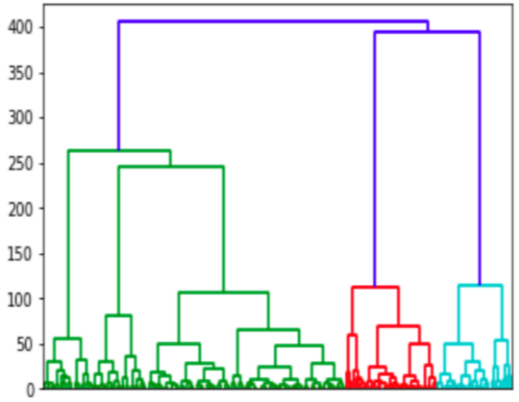
Compare and contrast K-means Clustering and Hierarchical Clustering

	K-means Clustering	Hierarchical Clustering
Big Data	Can Handle	Can’t Handle
Time Complexity	Linear O(n)	Quadratic O(n2)
Values Results	Variable	Reproducible
Running Time	Faster	Slower
Parameters	K (number of clusters)	None
Clusters	Subjective (only a tree is returned)	Exactly K clusters
Partitioning Produces	Produce a single partitioning	Produce different partitioning depending on the level of resolution

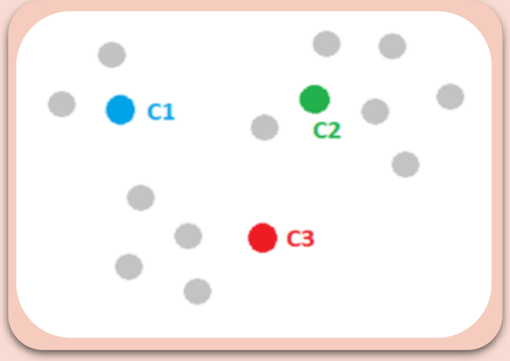
K-means Clustering Example



Hierarchical Clustering Example

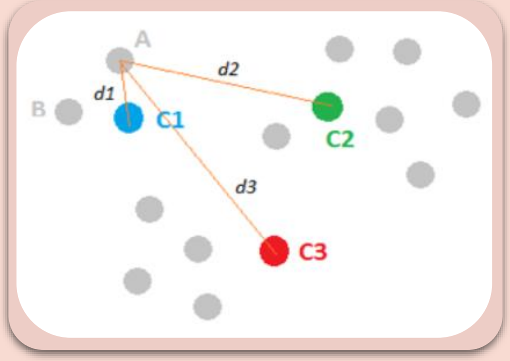


Briefly explain the steps of the K-means clustering algorithm



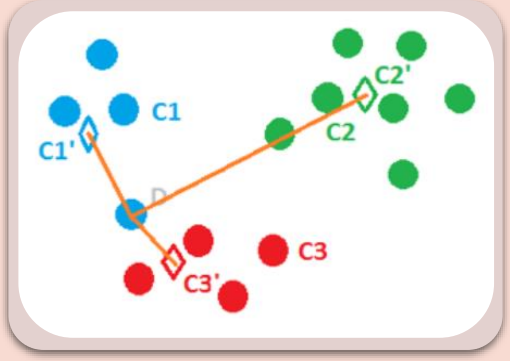
Step – 1: Choose One Center

Choose one center of the data point at random



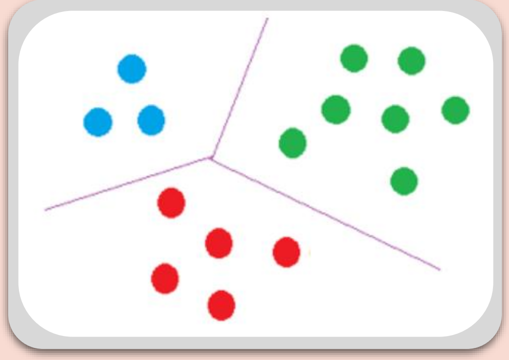
Step – 2: Compute the Distance

For each data point *Xi*, compute the distance between *Xi* and the nearest center that had already been chosen



Step – 3: Weighted probability distribution

Choose the next cluster center using weighted probability distribution where point *X* is chosen with probability proportional *d* (*X*)2



Step – 4: Repeat the steps 2 and 3

Repeat the steps 2 and 3 until *K* centers have chosen

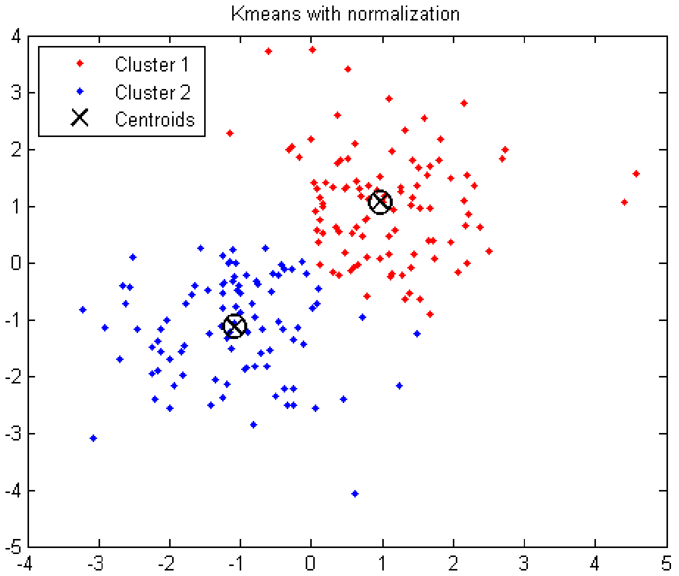
How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it

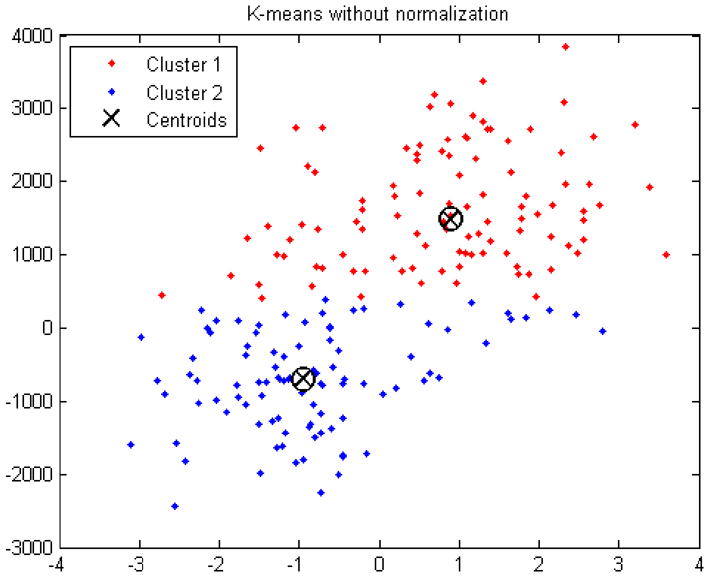
K-Means Clustering belongs to Partitioning Class(of Clustering), the principle rule is to keep the within-cluster variation or total within-cluster sum of square to be minimum. This actually means that, the variation within the cluster should be minimum. This can be done if we choose the right/optimal number of clusters.

- **Statistical Object –** These measures are often applied in combination with probabilistic clustering approaches. They are calculated with certain assumptions about the underlying distribution of the data.
- **Business Object – T**he customers in a particular cluster are not similar to each other, their requirements might vary. If the bank gives them the same offer, they might not like it and their interest in the bank might reduce.

Explain the necessity for scaling/ standardization before performing Clustering

- It controls the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms, check out the link below to view its effects on k-means analysis
- If you have mixed numerical data, where each attribute is something entirely different (say, shoe size and weight), has different units attached (lb, tons, m, kg ...) then these values aren't really comparable anyway; z-standardizing them is a best-practise to give equal weight to them.





Explain the different linkages used in Hierarchical Clustering

There are 03 types of linkages used in Hierarchical Clustering which is as follows:

1. **Single Linkage:** The distance between 2 clusters is defined as the shortest distance between points in the two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.
2. **Complete Linkage:** The distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together.
3. **Average Linkage:** The distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering

Usually, single linkage type will produce dendrograms which are not structured properly, whereas complete or average linkage will produce clusters which have a proper tree-like structure.