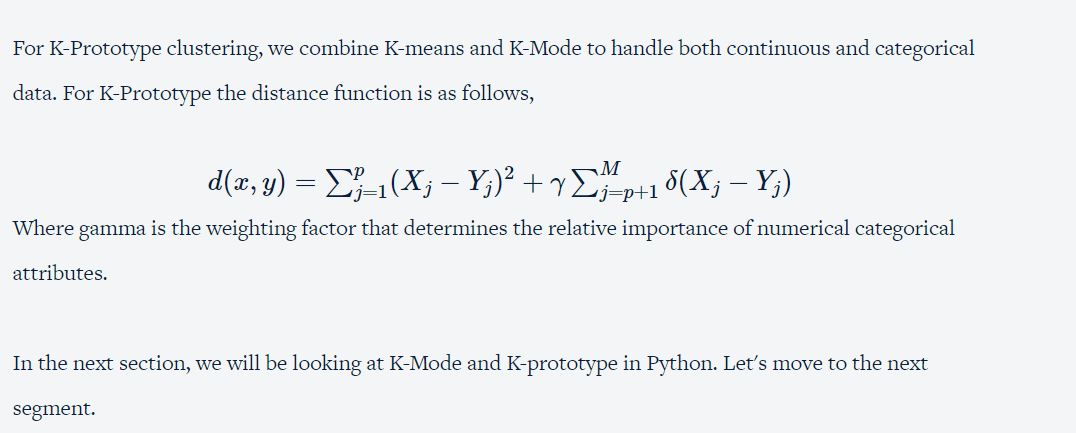
# K-mode Clustering:

The K-Means clustering algorithm is undoubtedly one of the most widely used partitional algorithms **for numerical data or continuous data**, but K-Means **can't handle categorical data**, and the reason is that of the difference in the dissimilarity measure the K-Means uses.

The **K-modes clustering algorithm is based on K-means paradigm but removes the numeric data limitation while preserving its efficiency**.



# K-Prototype in Python

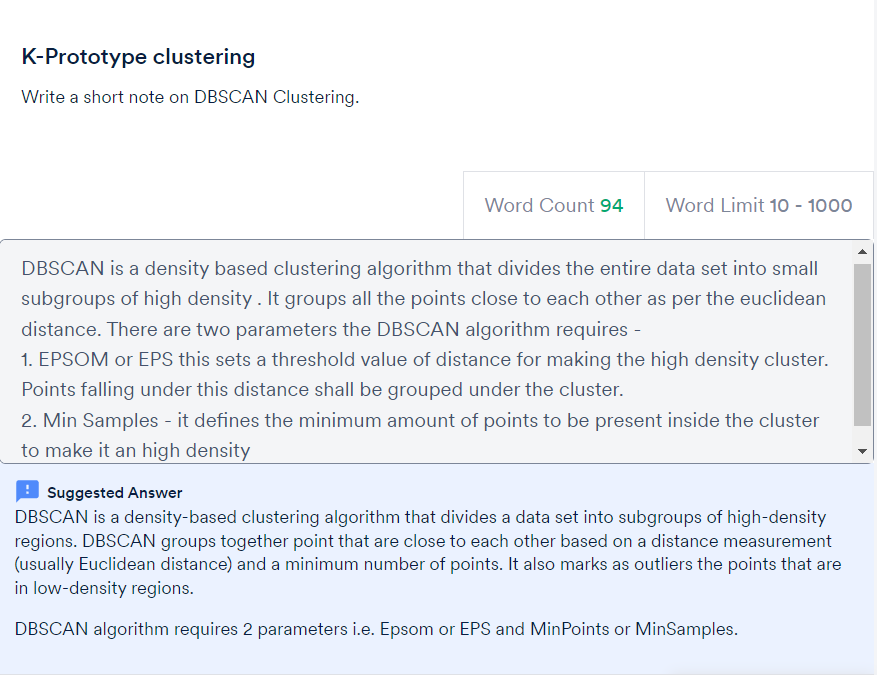
K-Prototype clustering is an algorithm to combine K-Means and K-Modes. K-Prototype can handle both continuous and categorical data to create clusters.

DB scan Clustering:

DBSCAN is a density-based clustering algorithm that divides a data set into subgroups of high-density regions. DBSCAN groups together point that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

DBSCAN algorithm requires 2 parameters:

1. Epsom or EPS
2. MinPoints or MinSamples.



# Summary

In this session, we covered in detail about two algorithms namely K-Mode and K-Prototype clustering.

To summarise, The K-modes clustering algorithm is based on K-means paradigm but removes the numeric data limitation while preserving its efficiency.

K-modes Algorithm uses modes instead of means to form clusters of categorical data.

Steps of the algorithm.

* Randomly assign “K” number of modes.
* Calculate the dissimilarity score between each of the remaining data points from the “K” number of chosen modes.
* Associate the data points to the mode whose score is minimum.
* Repeat from step 2 until there is no reassignment of clusters or when cost function is minimized.

For K-Prototype clustering, we combine K-means and K-Mode to handle both continuous and categorical data. For K-Prototype the distance function is as follows,

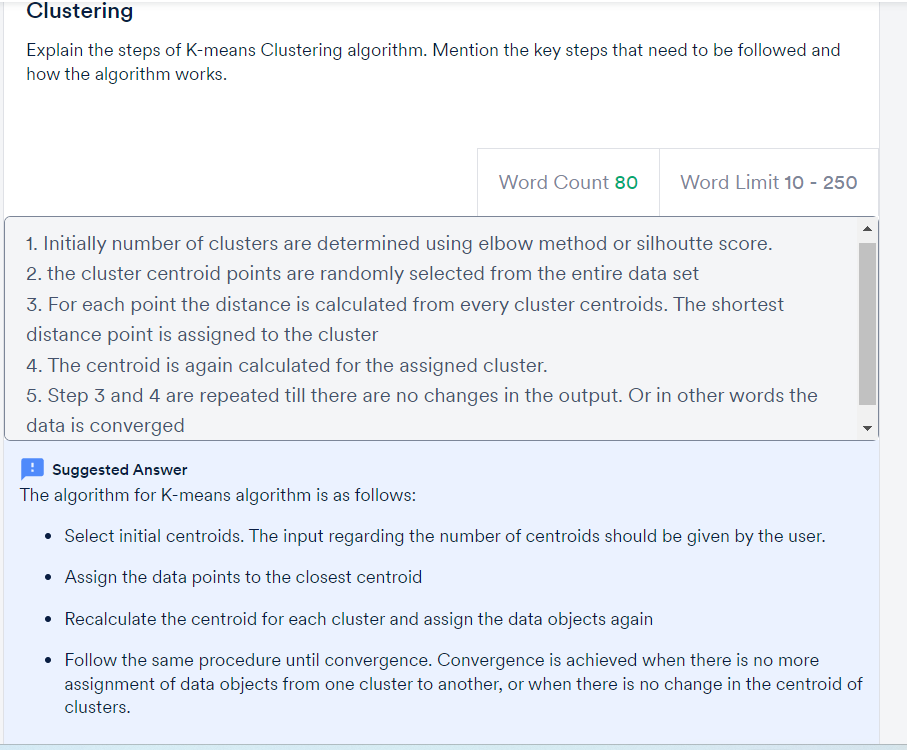
d(x,y)=∑pj=1(Xj−Yj)2+γ∑Mj=p+1δ(Xj−Yj)

Where gamma is the weighting factor that determines the relative importance of numerical categorical attributes.

Steps of the algorithm:

1. Select k.
2. Allocate each data point to a cluster which is done with considering the dissimilarity measure.
3. Retest the similarity of objects against the current prototypes. Update the prototypes.
4. Repeat 3, until no object changes its cluster.

We also talked briefly about the DBSCAN algorithm which is a density-based clustering algorithm that divides a data set into subgroups of high-density regions.



ddd