Assignment: Part II

Q1: Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of Clustering produced a better result and so on)?

Answer:

Main objective for this assignment is to identify the countries that are in dire need of financial aid. As a data analyst our role is to find the countries using socio-economic and health factors that determine the overall development of the country. Then my job is to suggest the countries which the CEO needs to focus on.

First task is to do some data inspection check the shape of the data frame, data type of the data frame and missing values in the dataset. Then perform data quality check like converting some of the columns given in percentage to its absolute values.

Next step is to perform EDA: univariate and bivariate analysis. Like checking the distribution of the data using distplot. Checking for multicollinearity among the numeric variables using correlation matrix. to identify if there is any linear relationship between the variables plot the pair plots and identify the same.

Next step is to identify the outliers now for this business model it is important to understand the outlier treatment of certain variables. there is two ways to do the outlier treatment ie. Soft capping and hard capping. For our assessment it is better to do some soft capping as we can see a lot of outliers in all the columns and we don't want some key information getting lost. Columns such income, gdpp cannot be hard capped and columns like child_mort,health shouldn't be treated at all as these variables are very important for the analysis. For our analysis we have performed soft capping of 99% value for gdpp as we wont lose much data which is required for the analysis.

Perform both kmeans and hierarchical clustering on the dataset to create the clusters.

Single linkage will not give a proper structure for this we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchical clustering.

We saw that final complete linkage hierarchical clustering gave us 3 cluster as the ideal value. Here we were able to get the right cluster size and were able to segment the countries in different groups the right way. But the only problem we had using this approach was that about 89% of the child_mort data was in cluster0 which meant imbalanced data distribution which will not give a good model so we have to go for the kmeans approach.

In k means we calculated the silhouette score and SSD elbow curve and Hopkins test to give us an optimum k value between k=3 to k=5. After performing the iterations, we came to a

conclusion that k=3 gives the best solution as it was able to segregate the countries in the right cluster. Then our next task was to do a cluster profiling with respect to gdpp, income, child_mort, where we need to scatter plot and identify the patterns among the 3 variables and ideally, we want to identify that one cluster where we have low income, low gdpp, high child_mort value. Which would give us those countries that would need the financial aid. Finally, we have to filter and sort the values such that we identify the top 10 countries that need financial aid from the final selected k means cluster method.

Question 2: Clustering

a) **Compare and contrast K-means Clustering and Hierarchical Clustering.**

Answer:

K-Means Clustering	**Hierarchical Clustering**
1.need to decide desired number of clusters ahead of time.	1.We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights
2. It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	2.Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
3. Works well in large dataset	3. works well in small datset.
4. outliers are not treated properly in k-means	4. outliers are properly identified in hierarchical clustering
5. kmeans is valid for numerical values.	5. it works very well with both numerical and categorical data.

b) **Briefly explain the steps of the K-means clustering algorithm.**

Answer: 1. Specify number of clusters K.

2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

3.Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

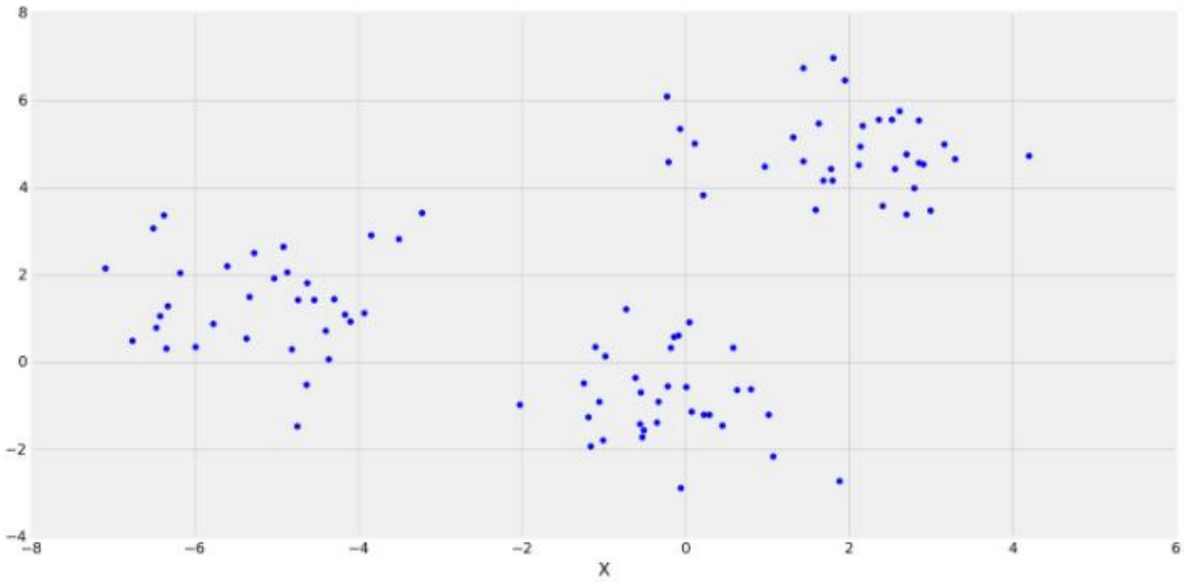
4.Compute the sum of the squared distance between data points and all centroids.

5.Assign each data point to the closest cluster (centroid).

6.Compute the centroids for the clusters by taking the average of the all-data points that belong to each cluster.

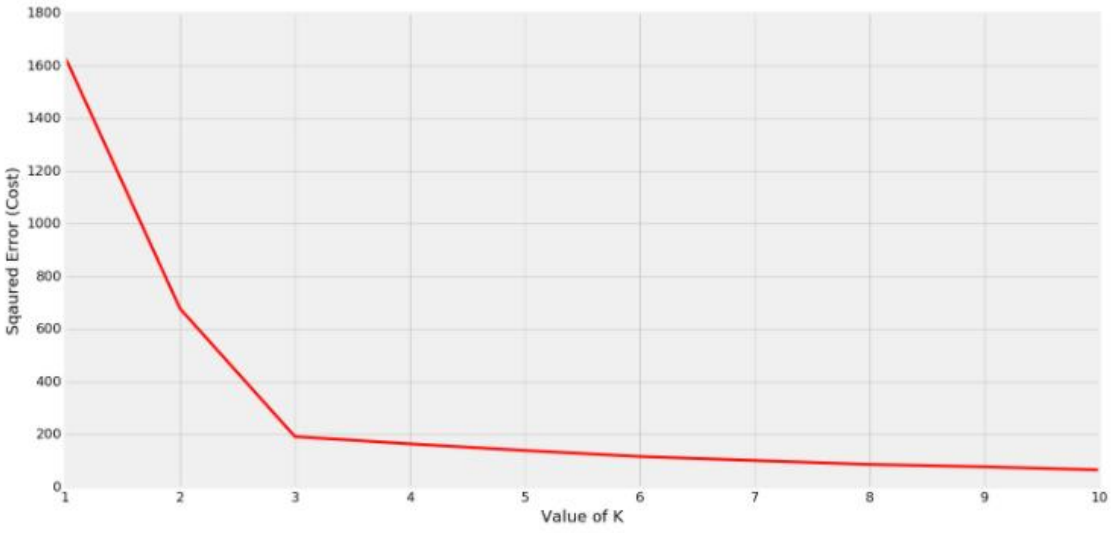
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Answer: There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point



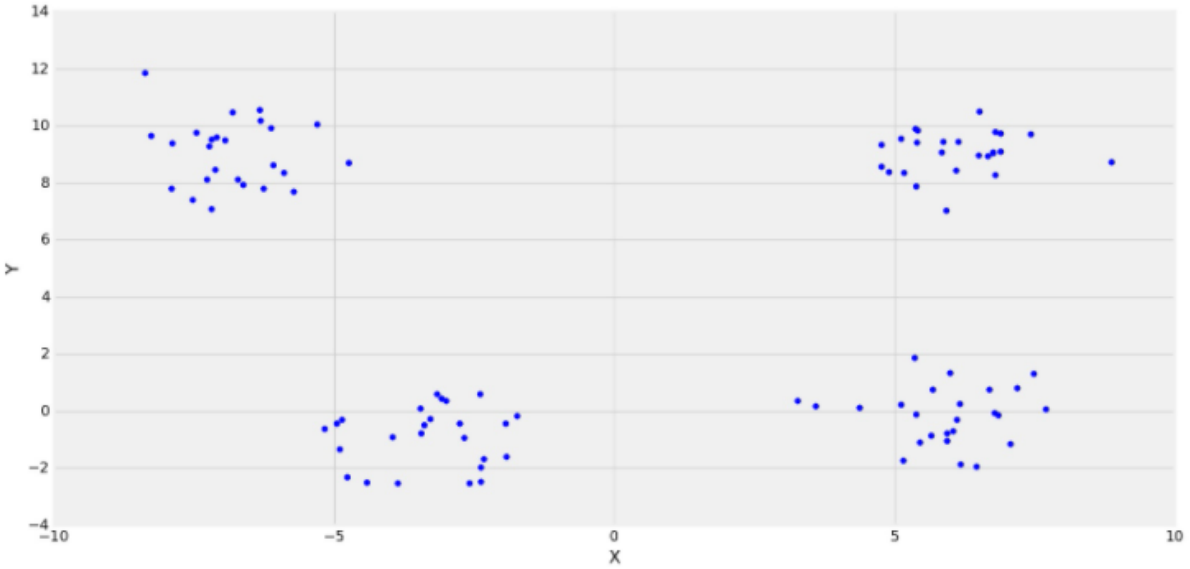
3 clusters are forming

In the above figure, its clearly observed that the distribution of points are forming 3 clusters. Now, let's see the plot for the squared error (Cost) for different values of K.



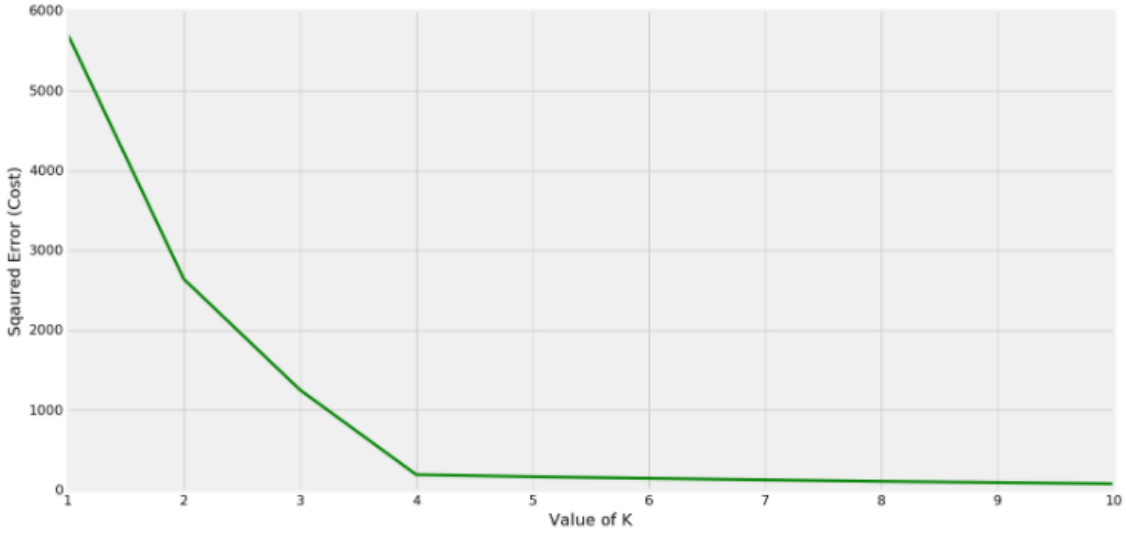
Elbow is forming at K=3

Clearly the elbow is forming at K=3. So, the optimal value will be 3 for performing K-Means.
Another Example with 4 clusters.



4-clusters

Corresponding Cost graph



Elbow is forming at K=4

In this case the optimal value for k would be 4. (Observable from the scattered points).

d) Explain the necessity for scaling/standardisation before performing Clustering.
Answer:

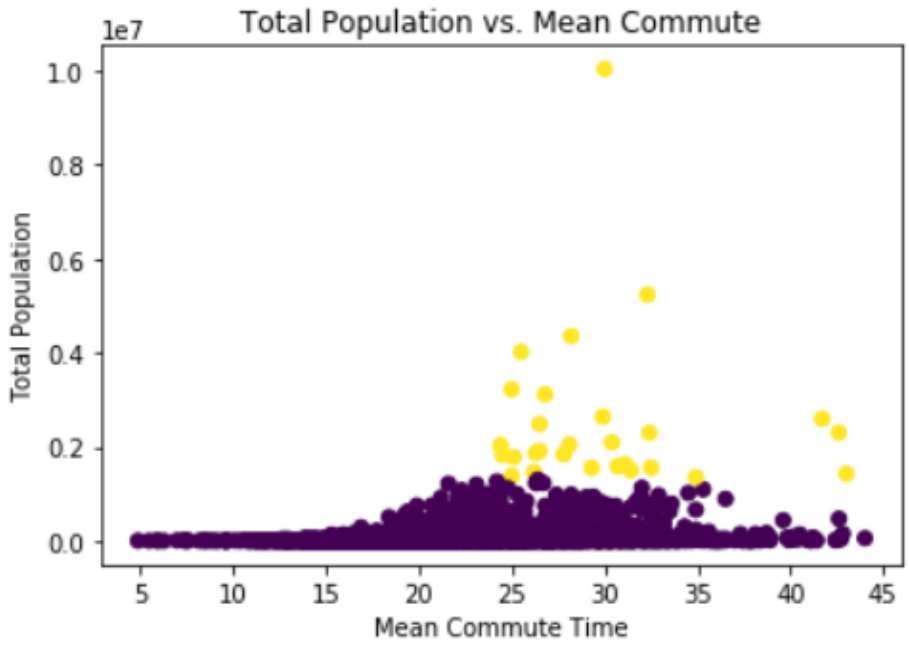
In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

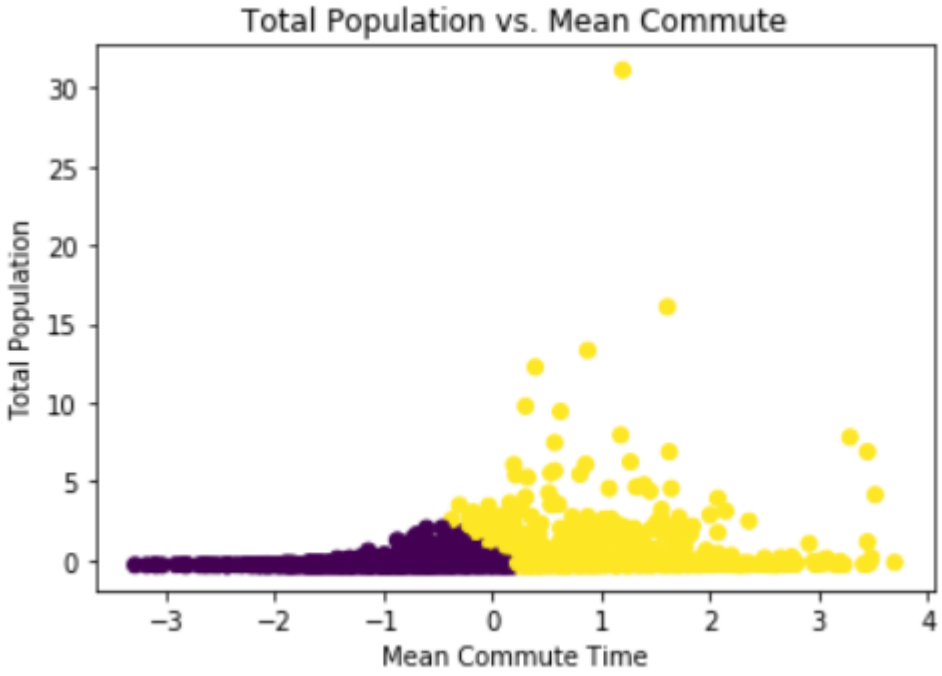
What follows is a couple examples demonstrating how standardization may impact a clustering solution, using the 2015 US Census Demographic Dataset, downloaded from Kaggle. This dataset includes different demographic variables for counties in the United States, including population, race, income, poverty, commute distance, commute method, as well as variables describing employment.

In our first example, we are interested in performing cluster analysis on Total Population and Mean Commute Time. We would like to use these two variables to split all of the counties into two groups. The units (number of people vs. minutes) and the range of values (85 - 10038388 people vs. 5 - 45 minutes) of these attributes are very different. It is also worth noting that Total Population is a sum, and Mean Commute Time is an average.

When we create clusters with the raw data, we see that Total Population is the primary driver of dividing these two groups. There is an apparent population threshold used to divide the data into two clusters:

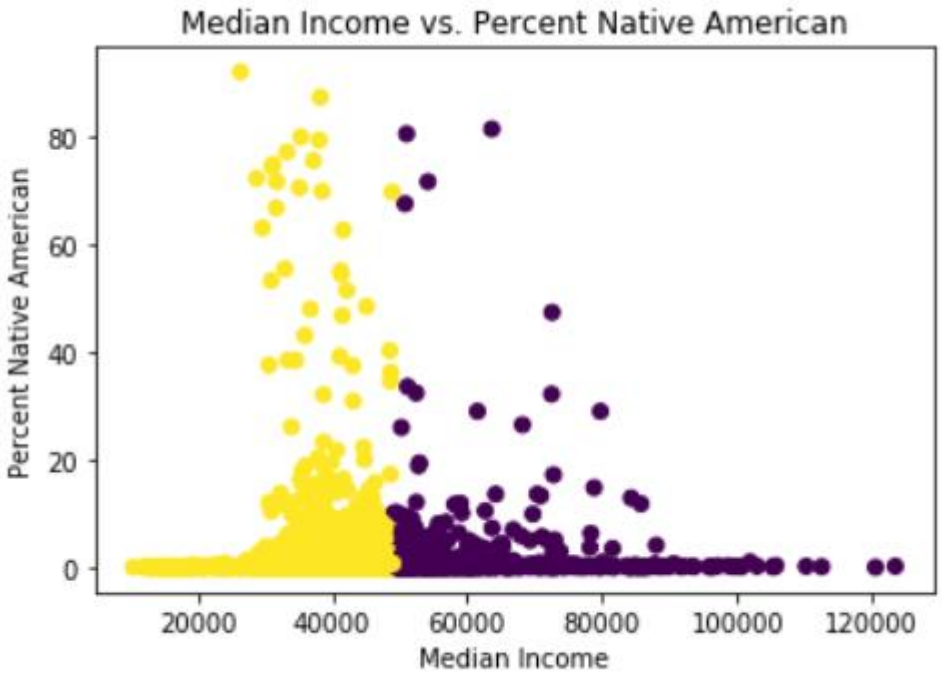


However, after standardization, both Total Population and Mean Commute seem to have an influence on how the clusters are defined.



In this next example, we are interested in clustering on Median Income and Percent of the Population that is Native American (by county). Median Income is measured in dollars and represents the "middle" income for a household in a given county, and Native American is a percentage of the total population for that county. Again, the units and ranges of these variables are very different from one another.

When we perform cluster analysis with these two variables without first standardizing, we see that the clusters are primarily split on Income. Income, being measured in dollars, has greater separation in points than percentages, therefore it is the dominant variable.



When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters.

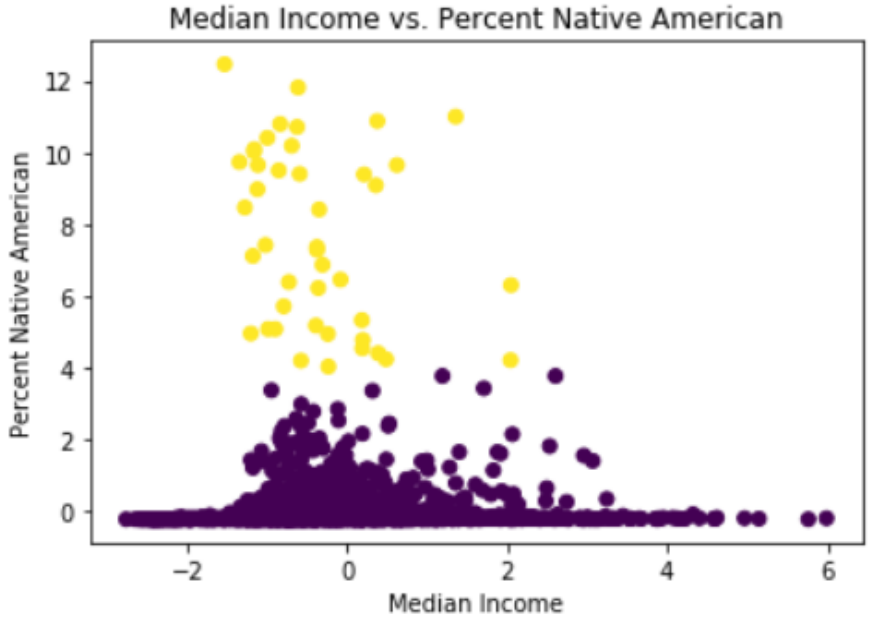
Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

There are a few different options for standardization, but two of the most frequently used are z-score and unit interval:

Z-score transforms data by subtracting the mean value for each field from the values of the file and then dividing by the standard deviation of the field, resulting in data with a mean of zero and a standard deviation of one.
Unit interval is calculated by subtracting the minimum value of the field and then dividing by the range of the field (maximum minus minimum) which results in a field with values ranging from 0 to 1.

Although standardization is considered best practice for cluster analysis, there are circumstances where standardization may not be appropriate for your data (e.g., Latitude and Longitude). If you'd like to read more there are a few great discussions on this topic on the Statistics and Data Science forums of Stack Exchange, as well as this academic article on Standardization and Its Effects on K-Means Clustering Algorithm by Ismail Bin Mohamad and Dauda Usman.

As always, the golden rule is to know thy data. Only you will know if standardization is right for your use case.



e) Explain the different linkages used in Hierarchical Clustering.
Answer: the different types of linkages.

Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
Average Linkage: Here, 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.
Divisive method:
In divisive or top-down clustering method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some circumstances but is conceptually more complex.

Agglomerative method:

In agglomerative or bottom-up clustering method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left.