Assignment: Part II

By: Simranjeet Singh
E-mail: sjsingh1809@gmail.com

Question 1: Assignment Summary

Question: 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:
In this assignment we were supposed to give the list of countries that are in direst need of aid so that the help International's NGO CEO can help in viewing the values in a bigger picture.

So, we started with the normal EDA process required to analyse and get familiar with data. From the data dictionary we found that the variables exports, imports & health were in the percentage of the GDP, hence converted these in their absolute values. Then we checked the outliers as they can affect the formation of groups.

On analysing, we found that there are some outliers present, but they seemed to be valid and not statistically insignificant because there we no hypothetical value. So, we kept those and verified the correlation, but that did not give much insight because the data was skewed so we transformed data to make it less skewed.

After power transformation the correlation became more meaningful and the outlier also reduced so the data was ready for the clustering.

We started with the k-means algo and first thing was to get the optimal number if clusters. For that we used elbow method and Silhouette score and number of clusters came to be 3.

Then we used hierarchical clustering, we used ward method to get the dendrogram and on analysing that we decided to use 3 groups.

For both we classified different countries into 3 groups (0 – low GPA, 1 – mid GDP 2 – high GDP) and the cluster 0 was the one that required attention.

Between these 2, hierarchical clustering gave better result because the data was pretty well grouped (exports were least, gdpp was low & child_mort were high).

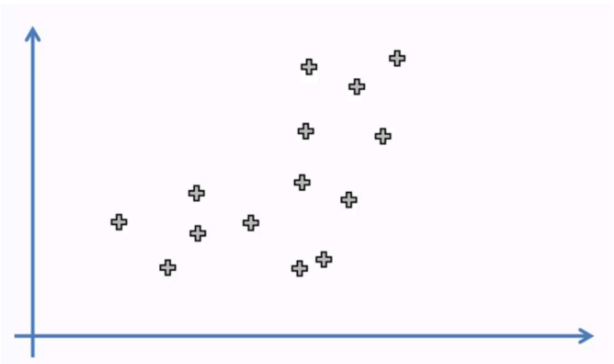
Countries which are in direst need of aid based on the gdpp from the analysis work are given below: Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone, Madagascar, Mozambique, Central African Republic, Malawi, Eritrea.

Question 2: Clustering

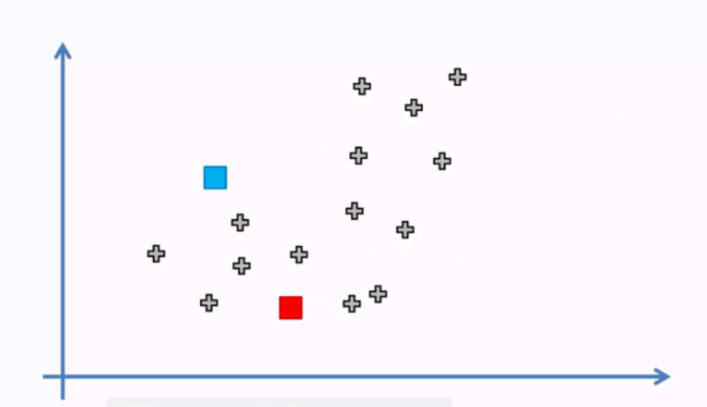
Question: Compare and contrast K-means Clustering and Hierarchical Clustering.
Answer:

K-means Clustering	Hierarchical Clustering
K-means clustering uses a pre-defined number of clusters.	Hierarchical clustering methods can be either divisive or agglomerative.
K-means clustering requires advance knowledge of K i.e. number of clusters one want to divide your data.	Hierarchical clustering one can stop at any number of clusters using the dendrogram
K-means can handle bid data well because the time complexity of K-means is linear (O(n)).	Hierarchical clustering can't handle big data well because the time complexity of hierarchical clustering is quadratic (O(n2)).
In this clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.	In this clustering, results are reproducible.
We use Elbow method to find the optimal number of clusters in K-means.	We use dendrograms in Hierarchical.

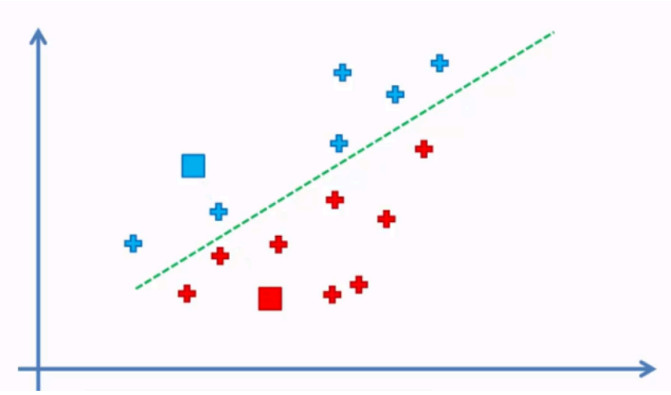
Question: Briefly explain the steps of the K-means clustering algorithm.
Answer:
Steps of the K-means clustering algorithm are:
1. Select 'k' cluster centres either by using Elbow method or business aspect.



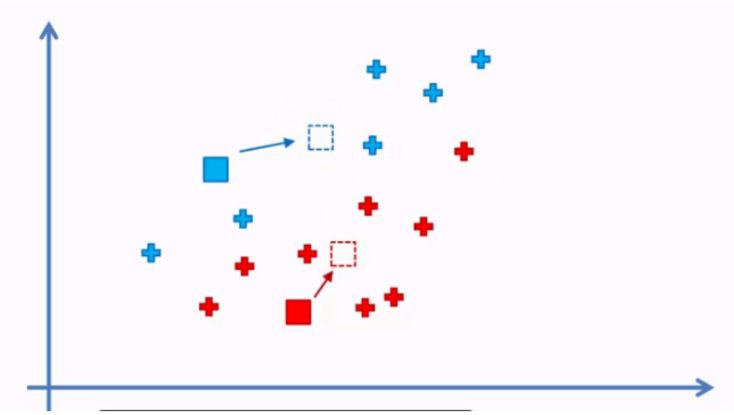
2. Calculated the distance between each data points and cluster centers.



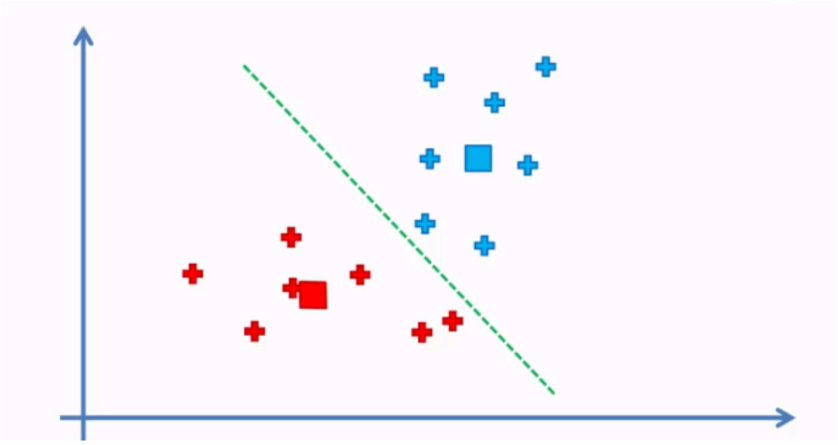
3. Assign the data points to the cluster center whose distance from the cluster center is minimum.



4. Recalculate the distance between each data point and new obtained cluster centers.

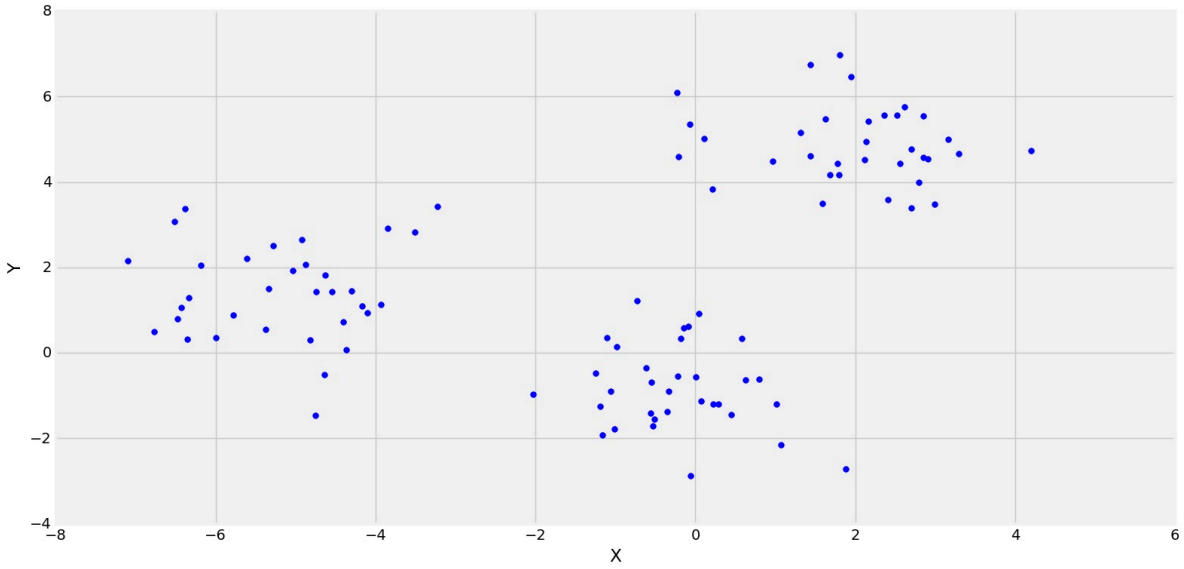


5. Repeat step 3 to step 5 until no data points was re-assigned.

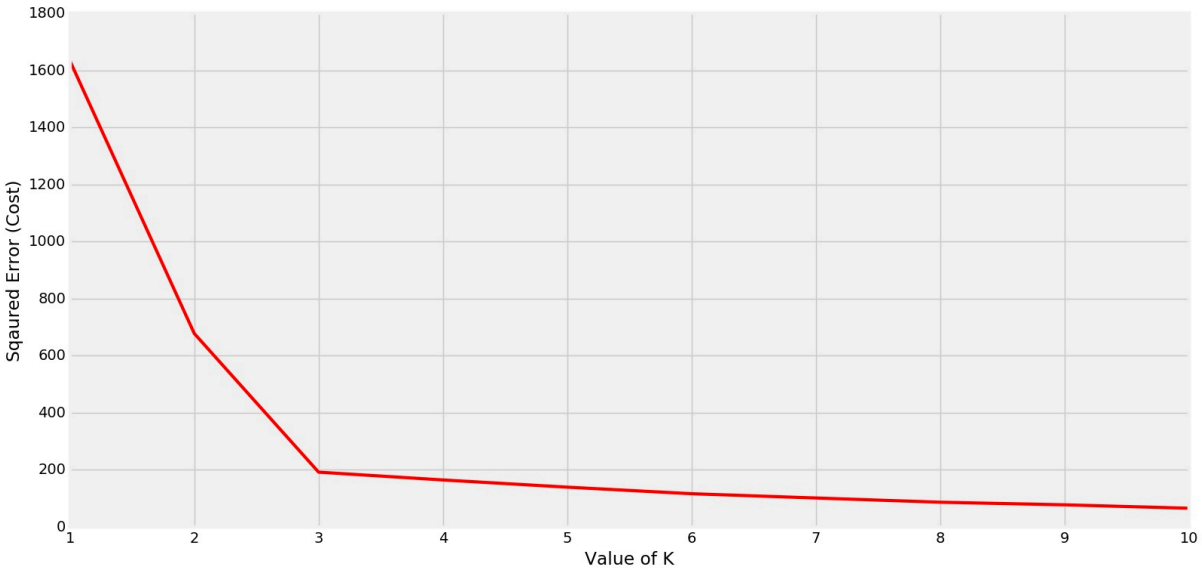


Question: How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
Answer:
The numbers of clusters can be obtained by following methods:

1. Elbow Method: The value of ‘k’ chosen in K-means clustering with the help of Elbow method. 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.
For ex:- Below is the data points. We can clearly see that there are 3 cluster. So, let’s look at the Elbow graph for the above data set.



Here we can clearly the elbow is forming at K=3. So the optimal value will be 3 for performing K-Means.



2. Silhouette Method: In this Silhouette value for every datapoint is calculated, the mean of which is used to find the optimal number of clusters. The silhouette value represents how similar a datapoint is to its own cluster when compared to all the other clusters or cluster centroids. The value range from -1 to +1. A higher silhouette value implies that the datapoint is matched well to its own centroid/cluster and is not so well matched with other clusters. If the mean of the silhouette value measured for all the datapoints is considerably high, then it can be said that the number of clusters are at its optimal value, or in other words, the clustering structure is appropriate. On the other hand, if the mean silhouette value turns out to be very less or negative, then it means that the cluster structure is not proper, and it may be having either more or lesser number of clusters than the optimal value.

NOTE: But sometimes we need to consider the business requirement also and then choose the number of clusters. For example, if using all these methods we found that the optima number of clusters came to be 5 but business does not require or wants more than 3 clusters then we need to accept that and use 3 and the total number of clusters.

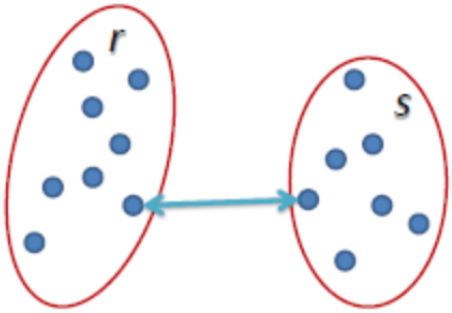
———

Question: Explain the necessity for scaling/standardisation before performing Clustering.
Answer:
In clustering, standardisation 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, standardisation may be important if you are working with data where each variable has a different unit (e.g., inches, meters, kilograms, etc), or where the scales of each of your variables are very different from one another (e.g. 0-1 vs 10-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

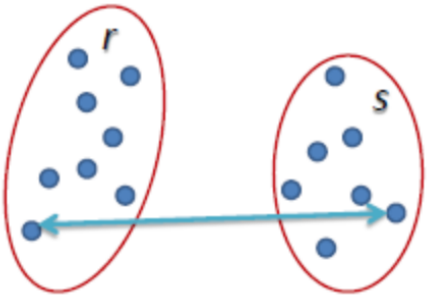
When you are working with data where each variable means something different, (e.g., age, height and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field 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. Standardisation helps to make the relative weight of each variable equal by converting each variable to a unit-less measure or relative distance.

Question: Explain the different linkages used in Hierarchical Clustering.
Answer:
The different linkages used in Hierarchical Clustering are:

1. Single Linkage: In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.



2. Complete Linkage: In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.



3. Average Linkage: In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.