Question 1: Assignment Summary

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: We have used PCA above to reduce the variables involved and then done the clustering of countries based on those Principal components and then later we identified few factors like child mortality, income etc which plays a vital role in deciding the development status of the country and builded clusters of countries based on that. Based on those clusters we have identified the below list of countries which are in dire need of aid. The list of countries are subject to change as it is based on the few factors like Number of components chosen, Number of Clusters chosen, Clustering method used etc.which we have used to build the model.

Question 2: Clustering

1. Compare and contrast K-means Clustering and Hierarchical Clustering?

Answer: There are a number of important differences between k-means and hierarchical clustering, ranging from how the algorithms are implemented to how you can interpret the results.

The K-means algorithms is a parameterized by the value k, which is the number of clusters that you want to create. As the animation below illustrates, the algorithm begins by creating k centroids.it then iterates between as assign step(where each sample is assigned to closest centroid) and an update step(where each centroid is updated to become the mean of all the samples that are assigned to it. this iteration continues until some stopping criteria is met; for example ,if no sample is re-assigned to a different centroid.

Hierarchal clustering instead, builds cluster incrementally, producing a dendogram. the algorithm the algorithm begin by assigning each sample to it own cluster. at each step, the two clusters that are the most similar are merged the algorithm continues until all the clusters have been merged. Unlike K-means, you don't need to specify a k parameter once the dendogram has been produced, you can navigate the layers of the tree to see which number of clusters makes the most sense to your particular applications.

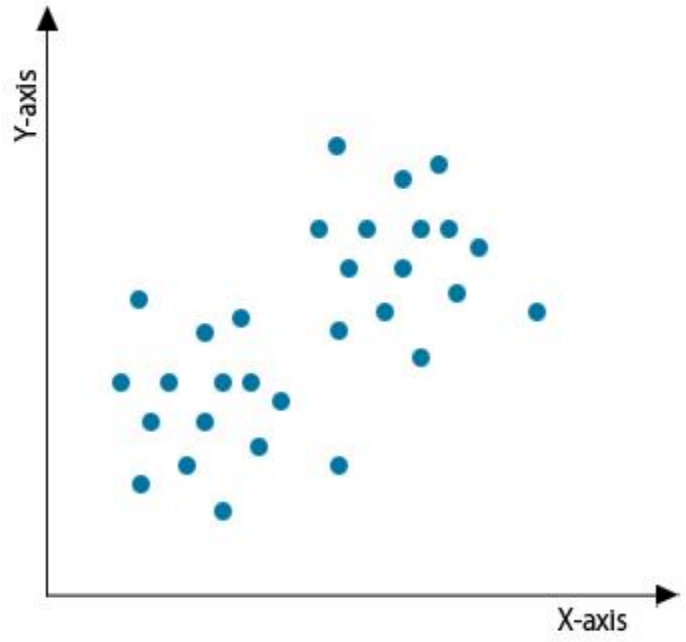
b) Briefly explain the steps of the K-means clustering algorithm.

Answer: This algorithm is an iterative algorithm that partitions the dataset according to their features into K number of predefined non- overlapping distinct clusters or subgroups. It makes the data points of inter clusters as similar as possible and also tries to keep the clusters as far as possible. It allocates the data points to a cluster if the sum of the squared distance between the cluster's centroid and the data points is at a minimum where the cluster's centroid is the arithmetic mean of the data points that are in the cluster. A less variation in the cluster results in similar or homogeneous data points within the cluster.

K- Means Clustering Algorithm needs the following inputs:

K = number of subgroups or clusters

Sample or Training Set = $\{x_1, x_2, x_3, \ldots\ldots x_n\}$



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: 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**.

Specify the number of clusters (K) to be created (by the analyst)

Select randomly k objects from the data set as the initial cluster centers or means.

Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid.

For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a Kth cluster is a vector oflength *p* containing the means of all variables for the observations in the kth cluster; *p* is the number of variables.

Iteratively minimize the total within sum of square (Eq. 7). That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations

Questions: Explain the necessity for scaling/standardization before performing Clustering?

Answer: Clustering models are distance based algorithms, in order to measure similarities between observations and form clusters they use a distance metric. So, features with high ranges will have a bigger influence on the clustering. Therefore, standardization is required before building a clustering model

Questions: Explain the different linkages used in Hierarchical Clustering?

Answer:Single Linkage: For two clusters like R and S, the single linkages returns the minimum distance between two points i and j such that i belong to R and j belong to S.

Complete Linkage: For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.

Average Linkage: For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.