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)

Problem Statement:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmers, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Objective:

- To categories the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

Method followed:

Data Processing:

- It was found that there were no null values
- There were also no duplicate values for country
- There were a few outliers and they were treated later on during PCA
- The data was standardized for Principal Component Analysis

Scree plot:

Three components are good enough to get the 90% variance as it is explained by 3 principal components, So we build the data frame using those 3 components only. So PC is selected to be 3.

Clustering:

- Both the methods K means and Hierarchical Clustering was used on the 3 PCA components
- For K means , K= 5 was taken using the elbow dip and silhouette analysis .
- While doing the Hopkins Statistics a value of 0.78 was attained.
    - If the Hopkins Statistics values are:
        - 0.3 : Low chase of clustering
        - around 0.5 : Random
        - 0.7 - 0.99 : High chance of clustering

Finally using all these values clusters of 3 were formed and the countries are split into clusters.

Question 2: Clustering

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

Ans: The main difference between k-means and Hierarchical Clustering is

- K Means needs a prior knowledge of number of centroid (K) whereas hierarchical cluster do not need these kinds of parameters. Cut-tree () function is used to create the number of clusters of any choice.
- In K Means clustering the algorithm will calculate the centroid each time.
- K Means is fast compare to hierarchical clustering
- Hierarchical clusters need more ram to run.

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

Ana: K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

This is how the algorithm works:

1. K centroids are created randomly (based on the predefined value of K)
2. K-means allocates every data point in the dataset to the nearest centroid (minimizing Euclidean distances between them), meaning that a data point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid
3. Then K-means recalculates the centroids by taking the mean of all data points assigned to that centroid's cluster, hence reducing the total intra-cluster variance in relation to the previous step. The "means" in the K-means refers to averaging the data and finding the new centroid
4. The algorithm iterates between steps 2 and 3 until some criteria is met (e.g. the sum of distances between the data points and their corresponding centroid is minimized, a maximum number of iterations is reached, no changes in centroids value or no data points change clusters)

Mathematical Formulation for K-means Algorithm:

D= $\{\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_i},\ldots,\mathbf{x_m}\}$ à data set of m records

$\mathbf{x_i}= (x_{i1},x_{i2},\ldots,x_{in})$ à each record is an n-dimensional vector

$$C_j = \text{Cluster}(X_i) = \arg_j \min || X_i - \mu_j ||^2$$

$$\text{Distortion} = \sum_{i=1}^{m}(x_i - c_i)^2 = \sum_{j=1}^{k} \sum_{i \in \text{OwnedBy}(\mu_j)} (X_i - \mu_j)^2$$

(within cluster sum of squares)

Finding Cluster Centers that Minimize Distortion:

Solution can be found by setting the partial derivative of Distortion w.r.t. each cluster center to zero

$$\frac{\partial \textit{ Distortion}}{\partial \mu_j} = \frac{\partial}{\partial \mu_j} \sum_{i \in \text{OwnedBy}(\mu_j)} (x_i - \mu_j)^2 = -2 \sum_{i \in \text{OwnedBy}(\mu_j)} (x_i - \mu_j) = 0 \textit{ (for minimum)}$$

$$\Rightarrow \mu_j = \frac{1}{|\text{OwnedBy}(\mu_j)|} \sum_{i \in \text{OwnedBy}(\mu_j)} x_i$$

For any k clusters, the value of k should be such that even if we increase the value of k from after several levels of clustering the distortion remains constant. The achieved point is called the “Elbow”.

This is the ideal value of k, for the clusters created.

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

Ans: 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**.

Business Uses

The **K**-**means clustering algorithm** is used to find groups which have not been explicitly labeled in the data. This can be used to confirm **business** assumptions about what types of groups exist or to identify unknown groups in complex data sets.

Statistical

The **k**-**means algorithm** is well known for its efficiency in **clustering** large data set. With a large number of variables **k**-**means** may be computationally faster than other algorithms (if **k** is small) and also may produce tighter **clusters** compared to other algorithms.

d) Explain the necessity for scaling/standardization before performing Clustering.

Ans: When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales,

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:

1. **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.

2. **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).**

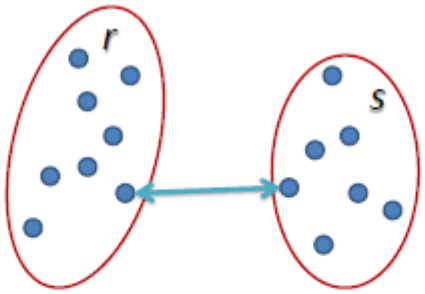
e) Explain the different linkages used in Hierarchical Clustering.

Ans: Hierarchical Clustering

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering,

- Divisive
- Agglomerative

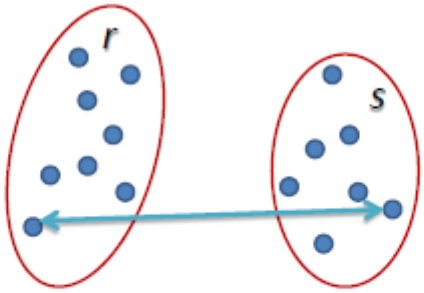
Single Linkage: In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$