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.

Answer: -

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.

Problem Statement: - After the recent funding programmes, 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.

Solution Methodology: -

- **Data Collection and Cleaning**
  - Import the gathered data
  - Checking the Data Quality and Cleaning the data for further steps
- **Outlier Analysis and Removal**
  - Removing the outliers where ever required after properly understanding the Business problem and deciding the expected result with the final model
- **Visualization of the data prior to applying the ML Algorithms**
  - Analyzing and plotting the graphs of the cleaned data to make sure that the data is properly prepared for further steps
  - Finding out any remaining outliers or random variances in the data
- **Scaling the Data**
  - Standardizing all the continuous variables so that the clustering algorithm can run efficiently
- **Performing PCA**
  - Principle Component Analysis is done on the cleaned data
  - Variance ratios are determined
  - Plotting the cumulative variances with the number of components identified
  - Performing the Dimensionality Reduction of the data
  - Reducing the correlation among the variables if present
- **Applying the K-Means Clustering Algorithm**
  - Initially identifying the ‘K’ using the Silhouette Analysis and sum of squared distances
  - Creating the ‘K’ clusters on the cleaned and prepared data
  - Visualizing the created clusters and analyzing them to check if the clusters are well segregated
  - Identifying the list of Countries from the clusters formed which needs the Aid from the NGO

- **Hierarchical Clustering**
    - Creating the dendrogram on the cleaned and prepared data
    - Identifying the ‘n’ number of clusters from the dendogram plotted and cutting the dendrogram at a required height to obtain the desired number of clusters
    - Visualizing the clusters obtained
    - Analyzing the clusters to ensure well segregation
    - Identifying the list of Countries from the clusters formed which needs the Aid from the NGO
- **Decision Making/Conclusion**
    - Based on the clusters formed using K-Means or Hierarchical clustering, the countries that need the most Aid from the NGO was identified and the list of such countries were prepared.

Conclusion: -

- 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 built 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

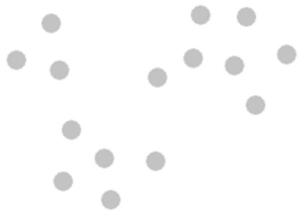
a) Compare and contrast K-means Clustering and Hierarchical Clustering.

- If there is a specific number of clusters in the dataset, but the group they belong to is unknown, choose K-means
- If the distinguishes are based on prior beliefs, hierarchical clustering should be used to know the number of clusters
- With a large number of variables, K-means compute faster
- The result of K-means is unstructured, but that of hierarchal is more interpretable and informative
- It is easier to determine the number of clusters by hierarchical clustering's dendrogram

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

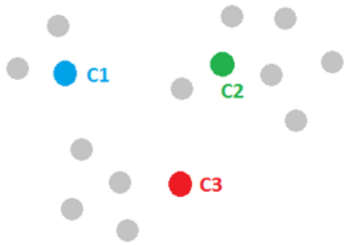
Among all the unsupervised learning algorithms, clustering via k-means might be one of the simplest and most widely used algorithms. Briefly speaking, k-means clustering aims to find the set of k clusters such that every data point is assigned to the closest centre, and the sum of the distances of all such assignments is minimized.

Let's walk through a simple 2D example to better understand the idea. Imaging we have these grey points in the following figure and want to assign them into three clusters. K-means follows the four steps listed below.

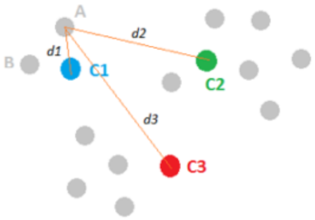


Step one: Initialize cluster centres

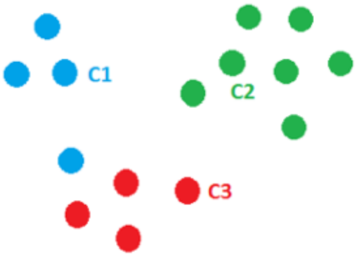
We randomly pick three points C1, C2 and C3, and label them with blue, green and red colour separately to represent the cluster centres.



Step two: Assign observations to the closest cluster centre

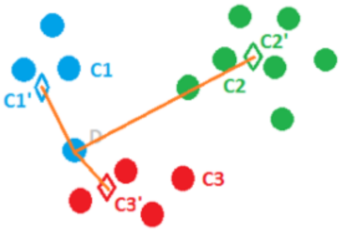


Once we have these cluster centres, we can assign each point to the clusters based on the minimum distance to the cluster centre. For the grey point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of *d1*, *d2* and *d3*, we figure out that *d1* is the smallest, therefore, we assign point A to the blue cluster and label it with blue. We then move to point B and follow the same procedure. This process can assign all the points and leads to the following figure.



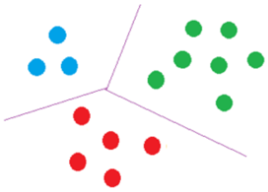
Step three: Revise cluster centres as mean of assigned observations

Now we've assigned all the points based on which cluster centre they were closest to. Next, we need to update the cluster centres based on the points assigned to them. For instance, we can find the centre mass of the blue cluster by summing over all the blue points and dividing by the total number of points, which is four here. And the resulted centre mass C1', represented by a blue diamond, is our new centre for the blue cluster. Similarly, we can find the new centres C2' and C3' for the green and red clusters.



Step four: Repeat step 2 and step 3 until convergence

The last step of k-means is just to repeat the above two steps. For example, in this case, once C1', C2' and C3' are assigned as the new cluster centres, point D becomes closer to C3' and thus can be assigned to the red cluster. We keep on iterating between assigning points to cluster centres, and updating the cluster centres until convergence. Finally, we may get a solution like the following figure. Well done!



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

Elbow Method can used to determine the optimal value of K to perform the K-Means Clustering Algorithm. In this method, various values of cost with changing k are plotted. As the value of K increases, there will be fewer elements in the cluster. Thus, the average distortion will decrease and the lesser the number of elements, shorter is the distance from to the centroid. The point where this distortion declines the most is the elbow point.
Apart from the Elbow Method, Silhouette Analysis can be performed to determine the value of 'K'. The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. The average silhouette method computes the average silhouette of observations for different values of *k*. The optimal number of clusters *k* is the one that maximizes the average silhouette over a range of possible values for *k*

From Statistical View Point, optimal number of clusters is a prerequisite because if the number of clusters given as input to the K-Means algorithm is fewer than the optimal value then in such a case, the algorithm will produce a result that does not capture the important aspects or the essence of the underlying data. In contrast, if the assumed K value is greater than the optimal value, then the model built will represent unnecessary associations between data points.
From Business view point, the value of the 'K' is determined based on the final requirement/proposed outcome as deemed fit by the Business team of the particular company/project. One should not blindly follow the optimal value of 'K' obtained from either Elbow or Silhouette Analysis, rather should also take into account the requirement of the concerned team and the Business Problem in hand to determine the optimal value of 'k'. For example, the Business team might want to classify its customers in 3 categories (Frequent Buyer, Average Buyer, Lazy Buyer) where as the optimal value of 'K' obtained from statistical analysis might come as '4'. Here the business requirement gets more priority and the clustering model should be built with 3 clusters.

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

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.

e) Explain the different linkages used in Hierarchical Clustering.

Linkage used in Hierarchical Clustering

In Hierarchical Cluster Analysis the problem that arises when a cluster contains more than one case is that the squared Euclidean distance can only be calculated between a pair of scores at a time and cannot take into account three or more scores simultaneously. In line with the proximity matrix, the goal is still to calculate the difference in scores between pairs of clusters, however in this case the clusters do not contain one single value per variable. This suggests that one must find the best way to calculate an accurate distance measure between pairs of clusters for each variable when one or both of the clusters contains more than one case. Once again, the goal is to find the two clusters that are nearest to each other in order to merge them together. There exist many different linkage measures that define the distance between pairs of clusters in their own way. Some measures define the distance between two clusters based on the smallest or largest distance that can be found between pairs of cases (single and complete linkage, respectively) in which each case is from a different cluster (Mazzocchi, 2008). Average linkage averages all distance values between pairs of cases from different clusters. Single linkage, complete linkage, and average linkage will each be fully detailed in turn.

Single linkage. Also referred to as nearest neighbour or minimum method. This measure defines the distance between two clusters as the minimum distance found between one case from the first cluster and one case from the second cluster (Florek, Lukaszewiez, Perkal, Steinhaus, & Zubrzchi, 1951; Sneath, 1957). For example, if cluster 1 contains cases a and b, and cluster 2 contains cases c, d, and e, then the distance between cluster 1 and cluster 2 would be the smallest distance found between the following pairs of cases: (a, c), (a, d), (a, e), (b, c), (b, d), and (b, e). A concern of using single linkage is that it can sometimes produce chaining amongst the clusters. This means that several clusters may be joined together simply because one of their cases is within close proximity of case from a separate cluster. This problem is specific to single linkage due to the fact that the smallest distance between pairs is the only value taken into consideration. Because the steps in agglomerative hierarchical clustering are irreversible, this chaining effect can have disastrous effects on the cluster solution.

Complete linkage. Also referred to as furthest neighbour or maximum method. This measure is similar to the single linkage measure described above, but instead of searching for the minimum distance between pairs of cases, it considers the furthest distance between pairs of cases (Sokal & Michener, 1958). Although this solves the problem of chaining, it creates another problem. Imagine that in the above example cases a, b, c, and d are within close proximity to one another based upon the pre-established set of variables; however, if case e differs considerably from the rest, then cluster 1 and cluster 2 may no longer be joined together because of the difference in scores between (a, e) and (b, e). In complete linkage, outlying cases prevent close clusters to merge together because the measure of the furthest neighbour exacerbates the effects of outlying data. Average linkage. Also referred to as the Unweighted Pair-Group Method using Arithmetic averages (UPGMA). To overcome the limitations of single and complete linkage, Sokal and Michener (1958) proposed taking an average of the distance values between pairs of cases. This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters. For average linkage, the distances between each case in the first cluster and every case in the second cluster are calculated and then averaged.

This means that in the previous example, the distance between cluster 1 and cluster 2 would be the average of all distances between the pairs of cases listed above: (a, c), (a, d), (a, e), (b, c), (b, d), and (b, e). Incorporating information about the variance of the distances renders the average distance value a more accurate reflection of the distance between two clusters of cases.

Each linkage measure defines the distance between two clusters in a unique way. The selected linkage measure will have a direct impact on the clustering procedure and the way in which clusters are merged together (Mazzocchi, 2008). This will subsequently impact the final cluster solution. In the next section, a hierarchical cluster analysis will be performed on a previously published dataset using SPSS.