**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( why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

**Note**: You don't have to include any images, equations or graphs for this question. Just text should be enough.

**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 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.

And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.  The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

**Solution**

**The Problem Statement has couple of objectives**

1. Categorise the countries using some socio-economic and health factors that determine the overall development of the country.
2. Suggest the CEO the countries which are in dire need of help.

**Approach:**

* The given data has 9 variables ('country', 'child\_mort', 'exports', 'health', 'imports', 'income', 'inflation', 'life\_expec', 'total\_fer', 'gdpp') that describe socio-economic and health factors of each country. There are 167 countries in the world and data is available for all the countries.
* There is no missing data and the data is clean. Exports, Health and Imports are given as % of GDP and hence they are transformed into % of GDPP to be on the same scaling.
* Visualize, Perform EDA to understand the data.
* Since all the data is on different scales, I have tried to use 2 scalers.

1. Model 1 with Normalizer which is robust to Outliers
2. Model 2 with Standard Scaler which is sensitive to Outliers

* **Model-I - Feature Selection using PCA and Clustering using Hierarchical and K-Means**

1. Data is Scaled with Normalizer
2. Perform Dimensionality Reduction using PCA
3. Select no of PCA's which give maximum variance. From Spree plot I have observed that 3 PC's contribute to above 95% of the variance in the data. I take the 95% as cut-off and hence consider to select 3 PC's. The principal components which are contributing to little above 95% variance are **'income', 'gdpp', 'imports\_per\_cap'**
4. Perform a check with Hopkins Statistic if cluster tendency is good or not. Hopkins Statistic gave me score of around 0.84. This indicates that Cluster tendency is good and we can proceed with Clustering mechanism.
5. Proceed with Hierarchical Clustering, cut the clusters (Linkage - Single, Complete and Ward) and visualize the the clusters vs PC's.
6. Single Linkage - No of Clusters selected were 4. When I observe the dendrogram, 4 longer stems forms nice groups.
7. Ward Linkage - No of Clusters selected were 4. When I observe the dendrogram, 4 longer stems forms nice groups.
8. Complete Linkage No of Clusters selected were 4. When I observe the dendrogram, 4 longer stems forms nice groups.
9. Proceed with K-Means Clustering, evaluate different K, select K, and visualize PC's.
10. K-means with some arbitrary k. I selected it as 3
11. Finding the Optimal Number of Clusters
12. - Using SSD or Elbow Curve – At 4, I see nice elbow.
13. - Silhoute Analysis – I observe that 2 has best silhouette score.
14. However, I choose 4 based on the business understanding. With 2 only 2 groups are segregated.
15. Fit and derive the labels using the selected no of clusters

* **Model-I - Feature Selection using PCA and Clustering using Hierarchical and K-Means**

1. Data is Scaled with Normalizer
2. Perform Dimensionality Reduction using PCA –
3. Select no of PCA's which give maximum variance. From Spree plot I have observed that 3 PC's contribute to above 95% of the variance in the data. I take the 95% as cut-off and hence consider to select 3 PC's. The principal components which are contributing to little above 95% variance are **'gdpp', 'child\_mort', 'inflation', 'health\_per\_cap', 'income'**
4. Perform a check with Hopkins Statistic if cluster tendency is good or not. Hopkins Statistic gave me score of around 0.92. This indicates that Cluster tendency is good and we can proceed with Clustering mechanism.
5. Proceed with Hierarchical Clustering, cut the clusters (Linkage - Single, Complete and Ward) and visualize the the clusters vs PC's.
6. Single Linkage - No of Clusters selected were 4. When I observe the dendrogram, 4 longer stems forms nice groups.
7. Ward Linkage - No of Clusters selected were 4. When I observe the dendrogram, 4 longer stems forms nice groups.
8. Complete Linkage No of Clusters selected were 4. When I observe the dendrogram, 4 longer stems forms nice groups.
9. Proceed with K-Means Clustering, evaluate different K, select K, and visualize PC's.
10. K-means with some arbitrary k. I selected it as 3.
11. Finding the Optimal Number of Clusters
12. - Using SSD or Elbow Curve – At 4, I see nice elbow.
13. - Silhouette Analysis – I observe that 2 and 4 have best silhouette scores.
14. However, I choose 4 as per Elbow curve and Silhouette analysis.
15. Fit and derive the labels using the selected no of clusters.

* **Model Selection**

1. Since, I have nearly 6 models which needs to be used, I used following approach to select the best model.
2. Visually how well the points are grouped wrt PC1 and PC2
3. Check the Silhouette Scores
4. Manual exploration of each model.
5. When I have checked the above metrics, I have chosen Model 2 (Data standardized with Standard Scaler) K-Means with k=4 as my model to derive conclusions.
6. This model had higher silhouette score compared to others
7. Clusters are clearly segregated when visualized wrt PC1 and PC2.
8. Minimum and Maximum of each feature across cluster don’t overlap with other and its nice indication that clusters are well grouped.

* **Visualize the Clusters and Analyze with respect to original data**

1. After analysis of the data, it was identified that
2. Cluster with label as 0 has countries whose socio-economic characters are average
3. Cluster with label as 1 has countries whose socio-economic characters are poor
4. Cluster with label as 2 has countries whose socio-economic characters are good
5. Cluster with label as 3 has countries whose socio-economic characters are very good
6. I have selected all the countries with Cluster label 1 which had nearly 30% of the countries whose socio-economic conditions are poor. The total no of countries was 48 in this cluster.
7. Based on following criteria, I have narrowed down to final list of countries which was recommended to derive final list of the countries.
8. Countries having Child mortality above 65 or 6.5%. I selected 65 as cutoff the reason being the cluster with average socio economic and health conditions have highest child mortaltiy as 64 and hence beyong this might not be good criteria.
9. Life Expectancy anything below 60. This cluster has life expectancy ranging from 32 to 71 with mean of 59. Hence I have selected 60 as the criteria.
10. Countries having Income below 2500 and gdpp above below can be considered as low economic conditions.

**The 5 countries which are in need of of help are as follows**

1) Congo Dem Rep - Congo Dem Rep has very low GDPP (334), low income (609), high child mortality rate (11.6%), health per capita spending of 26 and life expectany is around 57 years.

2) Central Africal Republic - Central Africal Republic has very low GDPP (446), low income (888), high child mortality rate (14.9%), health per capita spending of 17 and life expectany is around 45 years.

3) Niger - Niger has very low GDPP (348), low income (814), high child mortality rate (12.3%), health per capita spending of 17 and life expectany is around 57 years.

4) Burundi - Burundi has very low GDPP (231), low income (764), high child mortality rate (9.3%), health per capita spending of 17 and life expectany is around 57 years.

5) Haiti - Haiti has very low GDPP (662), low income (1500), high child mortality rate (20.8%), health per capita spending of 45 and life expectany is around 32 years.

**Question 2: Clustering**

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

**Similarities**

* Both of them Unsupervised Clustering mechanisms and form N data points into K-Clusters

**Differences**

* Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear while that of hierarchical clustering is quadratic.
* In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
* Hierarchical Clustering consumes lot of in memory space while K means uses less utilization of resources.
* K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
* K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

1. Briefly explain the steps of the K-means clustering algorithm.

**K-Means Algorithm** – K-Means algorithm is the process of dividing the N data points into K groups or clusters based on the Euclidian distance. The algorithm has 2 major steps and it starts by choosing K random points as the initial cluster centres.

1. **Assignment -** Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
2. **Optimization** - For each cluster, compute the new cluster centre which will be the mean of all cluster members.
3. Now re-assign all the data points to the diffrent clusters by taking into account the new cluster centres.
4. Keep iterating through the step 3 & 4 until there are no further changes possible.
5. Keep repeating Assignment and Optimization till the convergence or till there is no change in the no of clusters that are being formed.
6. At this point, we arrive at the optimal clusters. For each of the cluster, the algorithm assigns a label.
7. How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well  as the business aspect of it.
8. There are a number of pointers that can help us decide the K for our K-means algorithm:-
9. Elbow method:-

* Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters. For each k, calculate the total within-cluster sum of square (wss).
* Plot the curve of wss according to the number of clusters k.
* The location of a bend (knee) in the plot is generally considered as an indicator of the

appropriate number of clusters.

1. Average silhouette Method

* Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,
* by varying k from 1 to 10 clusters.
* For each k, calculate the average silhouette of observations (avg.sil).
* Plot the curve of avg.sil according to the number of clusters k.
* The location of the maximum is considered as the appropriate number of clusters.

1. Business View: Often in the business aspects we already know how many clusters we want as per the need or business constraints. For example a telecom service provider who wants to roll out offers only to their

valued customers. In such a case it would want to identify only 2 clusters of valued and non-valued

customers and roll out offer only to the valued customers group. In such cases we do not need statistical measure to compute the value ok k, but use the business domain constraints to fix the value of k.

1. Explain the necessity for scaling/standardisation before performing Clustering.

Most of the times, our data will contain features highly varying in magnitudes, units and range. Since Clustering uses Eucledian distance between two data points in their computations, this is a problem.

1. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. This is important for 2 reasons in K-Means algorithm:

* Since we need to compute the Euclidean distance between the data points, it is important to

ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.

* The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

1. Explain the different linkages used in Hierarchical Clustering.

There are several linkage methods in Hierarchial Clustering.

**Hierarchal Clustering Alogrithm**

1. Calculate the NxN distance (similarity) matrix, which calculates the distance of each data point from
2. the other. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item.
3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
4. Compute distances (similarities) between the new cluster and each of the old clusters.
5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.
6. Using dendrogram cut the clusters at optimal level to see nice well segregated clusters.
7. **Single Linkage -** Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters.
8. **Complete Linkage -** Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the Clusters.
9. **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.
10. **Ward Linkage -** Here, the distance between 2 clusters is defined as the average distance between The distance (D) between to clusters is defined as the error function of the unified cluster minus the error functions of the individual clusters.
11. **Centroid Linkage -** Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to centroid of every other cluster.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Question 3: Principal Component Analysis**

1. Give at least three applications of using PCA.

* Dimensionality Reduction or Feature Elimination
* Reducing Multicollinearity among the features.
* Better Visualization of Data with reduced dimensions

1. Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

* Basis Transformation - This means that we need to transform our data from some high

dimensionality to another lower dimensions. Thus the data representation would change in

different dimension. For example if we have data in 2dimensions plotted against 2 axis f1

and f2, we want to replot the same data without losing much information into 1 dimension

f1’. This is the process of Basis Transformation. Matematically it is a matrix operation where

the representaions in new basis are calculated as :

New Basis = M \* Old Basis

Here M is the matrix multiplicative factor used for transformation.

* Variance as information - We know the more the data varies from each other the more the

information is contained in the data. While moving from higher to lower dimensions we do

not want to lose information. Therefore to capture the maximum information, we need to

retain the maximum variance. In the below example where we want to use PCA to

transform data from 2D to 1D, we would want to rotate our axis f1 through some angle

theta to land at f1’ such that it capture the maximum spread in our data. Also f2’ should be

orthogonal to f1’ and should become an axis along which the variance is minimum so that

we can drop f’ without losing much information.

1. State at least three shortcomings of using Principal Component Analysis.
2. PCA is a linear method, however, in some situations, non-linear methods can produce better results.
3. PCA also requires the data to be highly correlated for it to create reasonable results.
4. PCA produces components which are orthogonal and uncorrelated, but sometimes, correlated components can be the better choice.
5. PCA assumes that lower types of variances aren't useful. Hence it may lead to loss of valuable classes or variables in supervised learning procedures.