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).

Solution:

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.

Methodology:

Our main task is to perform the Analysis on the given dataset and to give recommendations to the organization about which countries are socio-economic background and needs an financial assistance to improve their lifestyle. To make the recommendations I followed the following methodology to get the countries whcih requires financial aid:

Step - 1: Importing necessary libraries to perform the various analysis on the given dataset.

Step – 2 - Data Understanding: The given dataset is imported in Jupyter notebook and checked for the following:

1. Check for the length of the data(no. of rows and columns)
2. Whether data set contains any null values
3. As there are three columns(exports, health, and imports) which contain %age of values in GDP and converting them to values.

Step - 3 - Data Visualization: The pair plot is drawn to visualize the relation between the numerical variables. The correlation matrix is created and drawn heat map to find out multicollinearity. Boxplots are drawn to visualize the data for outliers.

Step - 4 - Data Preparation: From step-3, we can say there are some outliers and we need to treat them before proceeding to scale or model building. In

this case, I am using the Zscore method to remove the outliers with zscore > 3. After treating outliers, We checked the missing values, the shape of the data frame.

The country column is dropped to perform the scaling of data set to proceed with further analysis.

Step – 5 - PCA: Principal Component analysis is performed on the scaled data set and calculated the cumulative variance. We have considered 4 principal components which are accounted for almost 90% of data.

Step – 6 – Hopkins Score: Hopkins score is calculated after the PCA. We got Hopkins score as 90.04%, which is a good sign to proceed with Clustering analysis using K-Means and Hierarchical clustering.

Step – 7 – Elbow Curve and silhouette Score : Elbow curve and silhoutte score curves are plotted to find out optimal K-values to proceed further analysis.

We got the 4 values in elbow curve and 4values in silhoutte curve. Hence we are considering 4 values for the K-Means clustering.

Step - 7 – K-Means Clustering: K-Means clustering is performed on the PCA Dataframe which was obtained in Step-5 with k=4. Checked with the no. of clusters in each clusters using bar graph. Plotted box plot for the to visualize each feature and its clustering. Finally checked which country falls in what cluster.

The same process repeated for K=5 in K-Means clustering.

After performing k=4 and k=5, I can see that almost results are similar but very minor changes are there. Hence we considered K=4 will be the optimum clusters for this business requirement.

Step - 8 – Hierarchical Clustering: Hierarchical clustering is performed on the PCA Dataframe which was obtained in Step-5 with k=4. Single linkage and complete linkage methods are performed in this section. After that, I have considered a complete linkage method.

Checked with the no. of clusters in each clusters using bar graph. Plotted box plot for the to visualize each feature and its clustering. Finally checked which country falls in what cluster.

At last added cluster_hierarchicalPCA to the data frame to perform profiling of clustering on this dataset.

Step - 9 – Binning: To perform the profiling, I concatenated the dataframes with PCA and dataframes obtained during step 7 & 8.

Step - 10 – Profiling Clustering: Profilling performed on the data frame using

1. Clustering using K-Means – Scatter plot
2. Clustering Using Hierarchical – Scatter Plot
3. Clustering of each numerical variables with GDPP using K-Means-Scatter plot
4. Clustering of each numerical variables with GDPP using Hierarchical-Scatter plot

Step - 11 – Inferences and Recommendations: From the analysis above, we can draw the following inferences:

Note: These recommendation are after the outlier treatment.

1. Child mortality rate is highest in Chad
2. Life expentancy is lowest in Lesotho
3. Inflation is highest in Mangolia
4. Total fer is highest in mali
5. GDPP is lowest in Congo, Dem, Rep
6. Mayanmar recorded lowest imports and exports
7. Health sector should be focussed by Eritrea

Question 2: Clustering

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

K-Means Clustering	**Hierarchical Clustering**
K- means is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.	In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.
K-Means Algorithm has so many assumptions.	Hierarchical clustering has fewer assumptions - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points.
The k-means clustering is parameterized by the value k, which is the number of clusters that you want to create	Hierarchical clustering, instead, builds clusters incrementally, producing a dendogram.
k-means will often give unintuitive results	Hierarchical clustering can be more computationally expensive

	but usually produces more intuitive results.
In k-means method, we find the mutually exclusive cluster of spherical shape based on distance. In this case, we can use mean or median as a cluster centre to represent each cluster. It is helpful in the small and medium size of data.	In Hierarchical methods, we create hierarchical decomposition of the given set of data. We create hierarchical decomposition in two ways such as from bottom to the top or top to down.

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

Kmeans Algorithm

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**.

Algorithm of K-Means:

1. Initially specify the number of clusters(k=some number) basically by using the elbow curve and silhouette score.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroids for the clusters by taking the average of all data points that belong to each cluster.

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

The K-Means Algorithm works on the pre-chosen k-value. In General, their is no particular method to determine the exact value of K, but an accurate value of K can be estimated through the following method:

Mean distance is one of the commonly used methods to determine the value of k and their centroid.

Since increasing the number of clusters will always reduce the distance to data points, increasing *K* will *always* decrease this metric, to the extreme of reaching zero when *K* is the same as the number of data points.

Elbow curve is another method used to determine K value, where the curve is sharply decreasing and taking into the opposite direction

A number of other techniques exist for validating *K*, including cross-validation, information criteria, the information-theoretic jump method, the silhouette method, and the G-means algorithm. In addition, monitoring the distribution of data points across groups provides insight into how the algorithm is splitting the data for each *K*.

All these methods are used based on the business requirement. If the business requirement is required to cluster into 2 groups then if we got k=4 from any of the above methods, we will be considering K=2 only as business requires to be classified into 2 clusters.

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

Standardizing variables tends to make the training process better behaved by improving the numerical condition of the optimization problem and ensuring that various default values involved in initialization and termination are appropriate. Standardizing targets can also affect the objective function.

In some of the data sets, changing the measurement units may even lead one to see a very different clustering structure.

For example, the age (in years) and height (in centimeters) of four imaginary people. It appears that {A, B) and { C, 0) are two well-separated clusters. On the other hand, when the height is expressed in feet where the obvious clusters are now {A, C} and { B, D}. This partition is completely different from the first because each subject has received another companion.

To avoid this dependence on the choice of measurement units, one has the option of standardizing the data. This converts the original measurements to unitless variables.
On a simple, oranges cannot be compared with apples while clustering the dataset. Hence standardization needs to be performed to make unitless variables to continue with clustering analysis.

e) Explain the different linkages used in Hierarchical Clustering.

Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. Hierarchical clustering is often represented as a dendrogram

In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest **maximum** pairwise distance). The worst case time complexity of complete-link clustering is at most O(n^2 log n).

In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest **minimum** pairwise distance). The time complexity of single-link clustering is O(n^2).

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

There are many applications of PCA.

Image Compression:

Image compression is one of the applications of PCA. Suppose we have an image to recognize which is not part of any of the previous image data set. The PCA performs the difference between the image to be recognized and the previous dataset and process it to Principal Components. Based on this Image to be recognized from the data set based on the principal components and its transformed matrix.

Quantitative Finance:

PCA can be applied directly to the risk management and derivative of portfolios.

Taking the market derivatives and applying PCA to determine let's say 3 to 4 Principal Components which can give more returns in less time and many other categories.

Health – NeuroScience:

PCA used in Neuroscience to identify specific properties of stimulus that increase neuron's probability of generating an action potential.

It will take the covariance matrix of all the features of stimuli that generate an action potential.

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

Basis transformation:

Suppose, We have a dataset composed of a set of properties from cars. These properties describe each car by its size, color, circularity, compactness, radius, number of seats, number of doors, size of the trunk and so on. However, many of these features will measure related properties and so will be redundant. Therefore, we should remove this

redundancy and describe each car with fewer properties. This is exactly what PCA aims to do.

For example, think about the number of wheel as a feature of cars and buses, almost every example from both classes have four wheels, hence we can tell that this feature has a low variance (from four up to six wheels or more in some of the rare buses), so this feature will make bus and cars look the same, but they are actually pretty different from each other. Now, consider the height as a feature, cars, and buses have different values for it, the variance has a great range from the lowest car up to the highest bus. Clearly, the height of these vehicles is a good property to separate them. Recall that PCA does not take the information of classes into account, it just looks at the variance of each feature because it is reasonable assumes that features that present high variance are more likely to have a good split between classes.

Mathematically speaking, PCA performs a linear transformation moving the original set of features to a new space composed by principal component. These new features do not have any real meaning for us, only algebraic, therefore do not think that combining linearly features, you will find new features that you have never thought that it could exist.

Variance as information:

The Covariance matrix of the principal componets in the PCA. Let us consider there are three PCS having covariance matrix. The sum of diagnol elements of the matrix will become variability.

PCA replaces with each of the elements in covariance matrix and forms a new matrix and which is orthogonal.

If we calculate the first element of the element divided by diagonal of all elements then it gives some x% of the information in that matrix which is called as the variance of PC1. If we do the same for similar elements we will get the variance of the remaining PCs.

If we cumulate these variances, then it will become cumulative variances of the PCs. If two PCs account for 90% of the cumulative variance then

we can say that two PCs are account for the 90% of the information given in the data set.

c) State at least three shortcomings of using Principal Component Analysis.

1. Independent variables become less interpretable:

Once implementing PCA on the dataset, the original features will convert into principal Components which are the linear combination of the original features that are not readable and interpretable as original features.

2. Data standardization is a must before PCA:

PCA will not be able to find the optimal Principal components unless you perform scaling/standardization of data before its implementation.

Let's suppose we have a dataset has expressed in the units of Kilogram, Centimeter and millions, the variance scale is huge in the dataset. If we implement PCA on such a feature set without standardization, the result of the features with high variance and it also be large further it leads to false results.

Further, all the categorical variables are required to be converted to be numerical variables before implementing PCA.

3. Information Loss:

It is very essential that selecting the number of Principal Components with care, it may miss some of the information which is required to be business analysis.

4. Orthogonality:

Principal Components are orthogonal to the others. it will become an issue when we would like to find the projections with the highest variance.

5. Large variance implies more structure:

PCA uses variance as the measure of how each feature is important. Due to this, high variance features are treated as principal components, while the others are neglected.