Question 1: Assignment Summary

Problem Statement: Categorise all the countries using the given socio-economic and health parameters and determine which group of countries require a financial assistance.

Tools: Clustering algorithm (k-means and hierarchical clustering)

Approach:

- The raw data is subjected to cleaning and pre-processing before getting it ready for clustering. The steps include:
  - Checking for null-values (no null values found).
  - Changing column format (import, export and health converted to absolute values from relative)
- Outliers in the data set are treated as per merit.
  - The outliers belonging to income and GDP per capita are left without much interference as the outliers are very distinct and are likely to form a separate cluster of their own.
  - For the remaining columns, those outliers which are extremely beyond the normal range are completely eliminated and for the same columns the outliers which are close enough are capped. Thus, maximum possible data is retained.
  - Data is scaled using a standard scaler and then subjected to PCA.
  - As per the curve obtained from the explained variance ratio, 3 PCA components explain more than 90% of the variance.
  - The silhouette score and the elbow curve denote that the optimal number of clusters would be 3.
  - Using the 3 PCA components and number of clusters as 3 we perform the K Means algorithm.
  - Similarly, we perform hierarchical clustering algorithm and take number of clusters as 3 by cutting the dendrogram.
  - We obtain very similar results from both approaches and we obtain a list of countries which are separated by their characteristic features.
  - We look into the characteristics of each cluster and present out analysis accordingly.

Question 2: Clustering

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

K Means Clustering	Hierarchical Clustering
The number of clusters in which the data is intended to be divided into should be know and is a pre-requisite to the algorithm.	There is a flexibility to choose the number of clusters by cutting the dendrogram at any location.
The results obtained after every run of the algorithm might vary significantly.	The result of any run of the algorithm might be repeated in a subsequent run.
It is comparatively faster algorithm with a time complexity of O(n)	It is a comparatively slower algorithm with a time complexity of O(n^2)
It consumes more RAM.	It consumes less RAM.
It assigns a data point based on the nearest cluster centroid.	It can be done via single, average or complete linkage.
It is a partitioning algorithm.	It can be done via an agglomerative or divisive approach.

b) Briefly explain the steps of the K Means clustering algorithm.

K Means clustering is an unsupervised learning algorithm that involves the following steps:

- Decide the initial number of clusters (say n) and select ‘n’ points on the scatterplot of all the data points. These ‘n’ points form the initial cluster centres (centroid). Note that the number of clusters (or cluster centre) selected and their placement may impact the final outcome to some extent.
- Assign all the points to the nearest cluster centre. Use Euclidean distance to find the closest cluster centre to a point.
- Now you have ‘n’ different clusters. Calculate the centroid (centre of mass) of these clusters individually. This can be done by taking a summation of all the data points in that cluster divided by the total number of data points in that cluster. We now get updated cluster centres.

- Repeat the last two steps until the cluster centres do know update or any data point does not change its parent cluster.

[Image Source: medium]

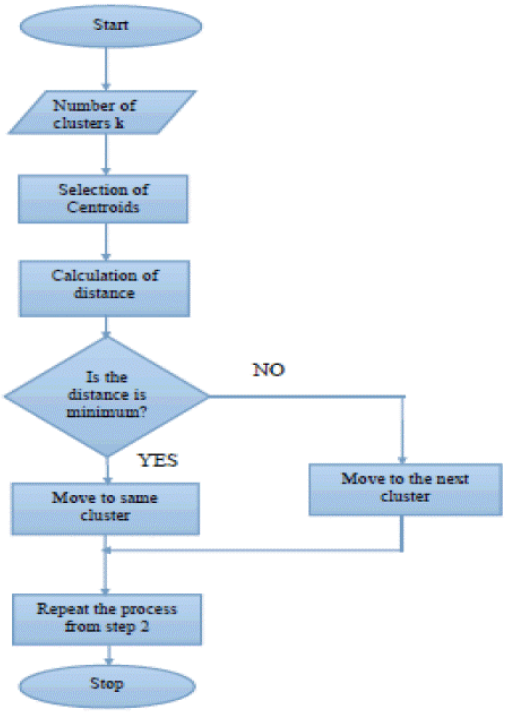

FIG. 1. **Flowchart for K-means clustering.**

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

K means algorithm requires the value of k as a pre-requisite. There are various ways of determining the optimal 'k'. It depends on various statistical and / or business parameters.

Statistics: The value of k can be determines using techniques like

- Elbow Method: For different values of 'k', we determine Within Cluster Sum of Squared Errors (WSS). We choose the value of 'k' for which this value starts to diminish.
- Silhouette Method: It works on the concepts that how similar is the point to the others in the same cluster and how different is it from the points in other clusters.

Business: Sometimes though statistics may represent an entirely different picture, the number of clusters may be decided as per the business use case. For example: An e-commerce company targeting certain categories of consumers. The number of categories may vary upon the company's goals.

d) Explain the importance of scaling / standardisation before clustering.
Scaling or standardization is extremely important to do before clustering. Let us take an example and try to understand.

Let us say that there are two columns in a data set which represent the height and weight of a person. The rate of increase of height may be completely different to that of weight. Also, the unit of measure of both quantities is different and hence aren't comparable in real life scenario. If we do not scale the quantities, the clustering algorithm or the PCA might falsely interpret as one quantity being more important than the other. Also, due to the difference in variation between quantities, the algorithm might encounter large Euclidian distances between heights and low between weights or vice versa which is a wrong basis for cluster formation.

Thus scaling / standardization is extremely important to maintain consistency in the data.

e) Explain the different linkages used in hierarchical clustering.
There are 4 types of linkages in hierarchical clustering.

- **Single Linkage:** It is the shortest distance between a pair of observations belonging to two separate clusters. It is also known as nearest neighbour linkage.
- **Complete Linkage:** It is the longest distance between a pair of observations belonging to two separate clusters. It is also known as farthest neighbour linkage.
- **Average Linkage:** It is the summation of every pair of observations between the two clusters divided by the total number of such pairs.
- **Centroid Linkage:** It is the distance between centroids of two clusters.

Question 3: Principal Component Analysis

a) Give at least 3 applications of using PCA.
PCA is used for dimensionality reduction and can be used in a variety of situations.
Image recognition: If we are given a similar image or an image of the same object, we can check the difference between the target image and the principal components. In the process we can also leave out some information which isn't useful to us for recognition.
Finance: It can be used in risk management, equity portfolio and fund diversification. Given the large number of investment buckets available we can bring them down to a few which might cater the needs and the risk appetite.
Medicine and Biotechnology: In the field of neuroscience, it can be used to identify the stimuli which will increase the neuron's probability of generating an action potential. It can be applied in determining the proportion of the chemical compounds that can be used in the preparation of a drug for best effective results.

b) Briefly discuss the two important building blocks of PCA – Basis transformation and variance as information.
Basis Transformation: When PCA takes the points expressed in standard basis and converts them into the same observations expressed in eigenvector basis.
Variance as information: PCA takes into consideration the variance explained by a particular class variable. Variables explaining higher variance are moved into new basis and are retained and those with low variance are regarded as noise and discarded.

c) State at least 3 shortcomings of using Principal Component Analysis.
The following are certain shortcomings that we might face while leveraging PCA for clustering.
    - PCA works on the principle that the property which is exhibiting the highest variance in the data set will be most useful for data point separation. However, this may not hold true for many cases. PCA assumes large variance to be principle components and small variance to be noise.

- PCA will not give optimal solution with unscaled or non-standardized data. Hence, even categorical variables need to be converted to numerical.
- Selecting the number of principle components carefully is very important as we may miss out on some useful information.
- PCA components are not interpretable and we cannot derive any useful information out of them just by looking at them.
- If the principal components are not a linear combination of the original features, the results of PCA will not be valid.