Question 1: Assignment Summary

Problem statement: HELP International, a global NGO able to raise fund around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. For that they hire a data analyst whose job is to categorise the countries using their socio-economic and health factors that determine the overall development of the respective country. Then suggest the countries which the HELP CEO needs to focus most for providing required Aid.

Solution Methodology: As an analyst we have to first understand the data from the given data dictionary. The data set comprises gdpp, child mortality rate, income etc attributes against each country. First job is the data preparation task.

- ***Data Preparation***:
    - As part of data preparation we have to load the data in pandas library, check the size,shape and other basic statistical references by using df.shape, df.info(), df.describe()
    - Treat the missing values, Null entries along rows or columns
    - Since this is a unsupervised exercise to perform categorization or clustering on the given data set we have get rid-off categorical variable or transform the categorical variables into dummy variables. In this case, only 'country' is categorical variable among all the others variables and we needed a ID column, we separated out 'country' attribute from the rest of the variables.
    - A new dataframe without 'country' column is produced to perform the clustering
    - Standardize the data using normalization method so that all the variables value are in same scale
        - This is required since in clustering internally the clustering algorithms uses Euclidean distance mechanism to calculate the distance of cluster centroids from data points within clusters and/or across clusters
- ***PCA***:
    - Ran PCA to find the principal components for dimensionality reduction using "Single value decomposing" technique using sklearn PCA libarary
        - pca = PCA(svd_solver = 'randomized', random_state=50)
    - Analyzed the variance ratios of principal components by choosing the correct number of PC's (Principal components) by visualizing the Scree Plot
    - Once the final PCA dataframe we produce then we did *Outlier Analysis* by using 1.5*IQR principal
- ***Clustering***:
    - We ran KMeans and Hierarchical clustering one-by-one
    - Plotted BoxPlots for each clusters against various socio-economic aspects like 'gdpp', 'income' and 'child mortality rate'

- Did Cluster Proofing and compared both the cluster to find which are countries needed the most Aid
- Both KMeans and Hierrarchical clustering kind of produces similar output and by comparing the clusters we can conclude that top 10 countries which needed the Aid most are Haiti,Sierra Leone,Chad,Central African Republic,Mali,Niger,Angola,Burkina Faso,Congo,Guinea-Bissau.

Question 2: Clustering

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

a. In KMeans clustering I have chosen k=4 because Silhouette score plot and SSD i.e. Elbow curve suggested the optimal cluster of the given data set could have been either 4 or 5.
b. In Hierarchical Clustering I have chosen to cut the tree at 6. Looking at the Dendogram height the no of cluster we can get between threshold 6 to 8. If we cut the tree at threshold 6, we no of cluster=6, where as if we cut the tree at threshold 8 we get no of cluster = 3.
c. The clustering data for which most needy countries for Aid comes out to be as follows

--- output from KMeans Clustering -----

	country	child_mort	gdpp	income
66	Haiti	208.0	662	1500
32	Chad	150.0	897	1930
31	Central African Republic	149.0	446	888
97	Mali	137.0	708	1870
112	Niger	123.0	348	814
3	Angola	119.0	3530	5900
25	Burkina Faso	116.0	575	1430
37	Congo, Dem. Rep.	116.0	334	609
64	Guinea-Bissau	114.0	547	1390
17	Benin	111.0	758	1820

--- output from Hierarchical Clustering----

	country	child_mort	gdpp	income
66	Haiti	208.0	662	1500
131	Sierra Leone	160.0	399	1220
32	Chad	150.0	897	1930
31	Central African Republic	149.0	446	888
97	Mali	137.0	708	1870
112	Niger	123.0	348	814
3	Angola	119.0	3530	5900
25	Burkina Faso	116.0	575	1430
37	Congo, Dem. Rep.	116.0	334	609
64	Guinea-Bissau	114.0	547	1390

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

a. In K-Means we have to choose the initial K i.e. no of clusters predefined
b. We also need to choose a random cluster centroid or cluster centers
c. The centers at first are randomly selected, and the euclidean distance between the points and the centers are calculated.
d. Based on the distance with the centers, the points are grouped into different clusters
e. Once the grouping is done, the centroid value of all the points belonging to clusters are calculated
f. Once the centers of both clusters are calculated, the euclidean distance is calculated again for all points and steps are repeated again till we have the same values for consecutive iterations.
g. The algorithm runs in two phase assignment phase and optimization phase

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

- In K-Means the value K can be chosen statistically by implementing Silhouette score or SSD (Sum of Squared distances) score or Elbow plots.

- Also by looking into the business aspects sometimes we can judge the optimal K by analyzing the data and the business statement

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

- If variables are on a different scale (e.g. fractions and millions), then PCA (while trying to maximise the variance) will give higher importance to the variables with high variance simply because of scale.

For example, if we change one variable from km to cm (increasing its variance), it may go from having little impact to dominating the first.

e) Explain the different linkages used in Hierarchical Clustering.

- **Single Linkage:** Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
- **Complete Linkage:** Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
- **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.

Question 3: Principal Component Analysis

a) **Give at least three applications of using PCA**
   a. PCA is a dimensionality reduction technique for mainly correlated features where the data set has multicollinearity. PCA can be used various applications such as

      i. For data visualisation and EDA

      ii. For creating uncorrelated features that can be input to a prediction model: With a smaller number of uncorrelated features, the modelling process is faster and more stable as well.

      iii. Finding latent themes in the data: If you have a data set containing the ratings given to different movies by Netflix users, PCA would be able to find latent themes like genre and, consequently, the ratings that users give to a particular genre.

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

PCA transform the data into linear combination of basis vector, thus creating smaller set of variables. The basis vectors are certain set of vectors whose linear combination is able to explain any other vector in that space.

Variance in statistics (sigma variance) is simply sum of the square's deviations from the mean. If a column has more variance, then this column will contain more information. PCA calculates the best direction which provides the maximum variance in the data set reducing multicollinearity. It changes the basis vector such a way that the new basis vectors capture maximum variance or information.

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

- PCA is a linear transformation method. In non-linear model it may not be used as efficiently as in Linear regression, logistic regression, linear model clustering

- PCA always go for high variance columns in data set. It assumes columns with low variance are not useful. This assumption may not hold true always.

- PCA needs the components to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use **Independent Components Analysis.**