Assignment Part II:

Question 1: Briefly describe the clustering of countries assignment.

The solution to the assignment involved the following steps:

Step 1: Understanding and cleaning the data

- The data from the csv file is read into a data frame and inspected for any missing / duplicate values.
- The exports, health and imports column are listed as percentage of GDP. So, appropriate conversions have been made on those columns.

Step 2: Outlier Treatment

- Outlier treatment have been visualized and treated accordingly.

Step 3: Scaling the data

- Standard scalar has been applied on all the numerical attributes.

Step 4: Correlation Matrix

- A correlation matrix is plotted over a heat map to understand the interdependencies involved between attributes.

Step 5: PCA, Scree plot & Correlation Matrix

- PCA has been applied on the data frame to find the principal components and their corresponding variance ratio.
- A scree plot has been plotted based on which the first 3 principal components have been selected as they explain most of the variance.
- Incremental PCA is applied on the data frame for dimensionality reduction.
- A correlation matrix is plotted and it was observed that no correlations exist among the first 3 principal components.

Step 6: Clustering

- Hopkins statistic is determined.
- Silhouette score analysis is performed. It recommended 3 clusters.
- Elbow curve is determined. It suggested 2 clusters.

Step 6.1: K-means Clustering

- K-means clustering is performed for K=3 & K=2.
- Appropriate visualization have been made based on which it is determined that K=3 enables better clustering of data points.
- A mean aggregation has been performed on all the attributes over the ClusterID.
- The countries in dire need for aid have been identified using the concept of binning and have been determined.

Step 6.2: Hierarchical Clustering

- An agglomerative hierarchical clustering (Dendrogram) has been formed on the PCA dataset using the complete linkage.
- The Dendrogram is cut at a height of 0.7. We get 3 clusters.
- These clusters were analyzed the same way as the clusters of K-means and appropriate aggregation has been made based on the ClusterID.
- The concept of binning has been applied to determine the countries in dire need of aid and have been determined.

Step 7: Reporting the final 5 countries to Help International

- The countries from K-means and Hierarchical clustering are sorted in the decreasing order of child mortality rate.
- The first five countries common to both the clustering have been reported.

Question 2: Shortcoming of PCA (List at least 3)

PCA has 3 major limitations / shortcomings:

1. The principal components have to be linear combination of the original columns
    - The above statement is an assumption made in the derivation of PCA. PCA has the ability to capture linear correlations between the features but fails when the above assumption is violated.
    - Thus, PCA is a linear transformation method and works well on linear models such as linear regression, logistic regression etc.
2. PCA requires PCs to be uncorrelated/orthogonal/perpendicular
    - In cases where the data demands correlated components to represent the data, PCA wouldn't be the right option as it will remove correlations. The alternative for this would be to use Independent Component Analysis (ICA), but is several times slower than PCA.
3. Columns with low variance are not useful
    - PCA assumes that columns with low variance are not useful, which might not be true in prediction setup, especially classification problems with class imbalance. Thus, in supervised learning situations, this can lead to loss of valuable information.

Question 3: Compare and contrast K-means clustering with Hierarchical clustering

Differences are listed below:

K-means Clustering	Hierarchical clustering
1. The choice of initial cluster center has an impact on the final cluster composition as the cluster center calculates attempts to form clusters with those data points close to it.	1. We actually assume each data point as a cluster center and start the clustering process by pairing the closest/most similar clusters into a new cluster.
2. We need to choose the number of clusters K in advance for the clustering to happen.	2. We can choose the number of cluster after the clustering is formed by cutting the dendrogram at an appropriate level.
3. K-means clustering is computationally less time consuming and can handle even large datasets better.	3. Hierarchical clustering is computationally expensive. Hence, it is not recommended to run this clustering on big datasets.

Both K-means clustering and Hierarchical clustering have some similarities as well:

1. Standardization of data:
   We need to convert all data points to represent them on a similar scale. This would ensure that no attribute outweighs other while calculating Euclidean distances.
2. Non-applicability with the categorical data:
   Neither K-means nor Hierarchical clustering can be used when dealing with categorical data as the concept of distance for categorical data wouldn't make much sense.