Question 1

Briefly describe the "Clustering of Countries" assignment.

The focus of the assignment is how to do Principal Component Analysis on a Clustering Predictive Analysis method.

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. To identify this, the top 5 most impoverished countries are selected and are divided the $10 million.

Solution Approach:

1. Remove Outliers before PCA so that there is no huge data loss.

2. **Principal Component Analysis:**

   Features are transformed into components as linear combinations of the original variables.

   While performing PCA, the K components of K-means algorithm is arrived at 4 and the its total variance is explained by 95%.

3. **K-Means Clustering:**
   Hopkins Statistics – Measure to check whether clustering is suitable for the data set, obtained value is 0.7 which is greater than 0.5. Hence, we can implement clustering on this data set.

   Ideal Value of K is found by Silhouette Score and Elbow Curve:
   With the output of these two, it is concluded that 4 is an optimum number for clustering, these are confirmed with the equal distribution of clusters also.

4. **Hierarchical Clustering:**

   The number of clusters are determined by the point on y-axis which cuts through maximum number of clusters.

Conclusion:

From the output of the clusters, the most needed cluster is selected and is sorted by high mortality rate, low exports, low health rate, high imports, low income, high inflation, less life expectancy, less fertility rate and less GDP per capita and the top 5 countries in this cluster is presented as the most needful country.

Question 2
State at least three shortcomings of using Principal Component Analysis.

1.Independent variables become less interpretable: PCA transforms the original variables into linear combinations, with this the original interpretation is lost and the PCA components are less readable to derive any insights directly to make business sense.

2. Data standardization is heavy strain in PCA: All the categorical variables must be one hot encoded and numeric variables must be standardized to be in comparable lengths for PCA.

3. Information Loss: Although As a rule of thumb we take cumulative variance >=80, still those features which have lesser variance but more business relevant may be lost.

Question 3

Compare K-means Clustering and Hierarchical Clustering:

K-Means Clustering	Hierarchical Clustering
Work non-linearly, selects some random points forms clusters and optimizes it, converges faster and stops converging after the centroids no more get updated.	Algorithm works linearly, crashes when there is huge set of data and no cloud space to support hierarchical structure as it maps step by step