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 (what EDA you performed, which type of Clustering produced a better result and so on)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Ans: Help International aims to provide help to the people of under-developed countries by the funds they collected and hence they need to decide to use the funds effectively.

Approach:

- Exploratory Data Analysis: This is the first basic step wherein we observe the data carefully and try to find insights from it. Used boxplot(univariate analysis) to observe the outliers, Heatmap and Pairplot(Bivariate analysis) to see the correlation between different variables, Distplot to observe the distribution of data. Handled outliers using capping method. Scaling of data is also done to make the data uniform.
- K-Means Clustering: Next we check the silhouette score and elbow curve to determine the number of clusters(k). We run the KMeans algorithm of the data and the clusters are generated. We analyse the clusters using scatterplot, histogram etc and deduce the socio-economic and health factors of each cluster. Then cluster profiling is done based on GDPP, INCOME, CHILD_MORT and countries needing the aid most is found.
- Hierarchical clustering: Single and complete Linkage is performed to produce the dendogram and then the tree is cut at value 3 to form the clusters. Then the same steps to analyse the cluster is performed as it was done for K-Means followed by cluster profiling.
- The final list of countries is shared to the organization (K-Means result is taken as the clusters are balanced which is not the case with Hierarchical).

Question 2: Clustering

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

Ans:

- K-Means can handle big data well but Hierarchical Clustering cannot as time complexity of k-Means is linear(O(n)) whereas for hierarchical it is quadratic($O(n^2)$)
- In K-Means since the choice of clusters is random, the results obtained by running the algorithm multiple times might differ. But the results are reproducible for Hierarchical clustering.
- K-Means require prior knowledge of k which is not needed in Hierarchical

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

Ans: Steps:

- Select k random clusters.
- Select k random data points as centroid for the clusters.
- Assign each data point to the cluster for which the Euclidean distance is minimum.
- Once the clusters are formed find the centroid and reassign each datapoints to the clusters. Keep iterating this step until the clusters stop changing.

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

Ans: The value of k is chosen based on 2 methods:

- Silhouette coefficient: Silhouette coefficient is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). It is computed using the formula: b(i)-a(i)/max{b(i),a(i)}, where a(i) is average distance from own cluster whereas b(i) is average distance from nearest cluster.In terms of business, the silhouette score for countries in similar conditions would be very high.
- Elbow Curve method: The elbow method involves plotting of the explained variation as a function of the number of clusters and we pick the elbow of the curve as the number of clusters. 'elbow' is the cut-off point, if we take k above that would be case of overfitting.

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

Ans: Scaling controls the variability of the dataset. It converts data into specific range using a linear transformation which generates good quality clusters and also improves the accuracy of clustering algorithms. All distance-based algorithms are highly affected by the scale of the variables. Example in a dataset we have 2 variables: age and salary. The range for both these variables would be entirely different. In such case high value of salary would impact the distance adversely and hence impact the performance of the algorithm.

e) Explain the different linkages used in Hierarchical Clustering.

Ans: Different linkages are:

- Single Linkage: It is the shortest distance between a pair of observations in two clusters.
- Complete Linkage: Distance is measured between the farthest pair of observations in two clusters
- Average Linkage: Distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance.
- Centroid Linkage: Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more like the new larger cluster than to their individual clusters causing an inversion in the dendrogram.