Question 1: Assignment Summary

Answer:

Clustering of Countries:

Problem Statement:

Present 5 countries with direst need to NGO, so that they can spend the money strategically on these countries.

Approach:

1. Load the data and understand the data using the dictionary and convert export, import and health to their absolute value from percentage.
2. Visualized the data using boxplots and bar graphs . This helped in visualizing outliers and also whether data can be grouped or not in clusters
3. Performed outlier treatment using capping method as dropping data in a smaller dataset (167 countries) seemed unreasonable.
4. Applied Min-MAX scaling on numerical data
5. Started creating cluster using Kmeans by finding number of cluster first. Based on the outcome of SSD elbow curve and Silhouette curve found out number of cluster are 3.
6. Formed 3 clusters using Kmeans
7. Formed cluster using hierarchical method(both single and complete).
8. After performing the visualization of dendograms and also Country bar graph for Income, gdpp and child mortality, the Hierarchical clustering with complete linkage had the meaningful outcome.
9. We found the countries with max child mortality and low gdpp and low income as the countries with direst need from NGO such Help International.

Report:

The 5 top countries with direst need because they have high child mortality and low income and low gdpp of help from NGO:

1. Congo Dem Rep.
2. Liberia.
3. Burundi.
4. Niger.
5. Central African republic.

This cluster has 59 countries is this country.

Question 2: Clustering

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

In K-Means algorithm, we divide N data points into K clusters. Steps of the algorithm are:
1. Choose K random points as initial cluster centers.
2. Assign each data point to their nearest cluster center.
3. Compute new cluster center by calculating the mean of all cluster members.
4. Using the new cluster centers, reassign all data points to nearest cluster center
5. Keep iterating through step 3 & 4 until there are no further changes possible.

The steps in the hierarchical clustering are for N data points are :
1. Calculate the distance of each data point from the other and create a NXN matrix
2. Consider each N items as one cluster. So, N clusters each containing just one item.
3. Find the closest pair of clusters and merge them into a single cluster.
4. Compute distances between the new cluster and each of the old clusters.
5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.

The above steps describe the difference in steps between Kmeans and Hierarchical.
Few more key difference are

KMEANS	**HIERARCHICAL**
Preferred for larger data set	Preferred for smaller dataset
Need to specify number of cluster as input	**NO** need to specify number of cluster as input
Computationally less expensive on memory, as only new centroid to be stored.	Computationally expensive on memory, as iteration save data for all cluster

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

In K-Means algorithm, we divide N data points into K clusters. Steps of the algorithm are:
1. Choose K random points as initial cluster centers.
2. Assign each data point to their nearest cluster center.
3. Compute new cluster center by calculating the mean of all cluster members.
4. Using the new cluster centers, reassign all data points to nearest cluster center
5. Keep iterating through step 3 & 4 until there are no further changes possible.

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

Statistical Analyses
We can use Silhouette score or the elbow curve to find the optimal value of K (number of cluster).
The silhouette value measures similarity of a data point to its own cluster and dissimilarity to rest of clusters. The silhouette ranges from −1 to +1, high value indicates the object is well matched to its own cluster and absolutely different to its neighboring clusters.

The elbow curve method uses the sum of square distances(SSD)of all data points to their closest cluster center to find the most optimal clusters at the point where there is not much change in SSD and elbow is formed.

Business Aspect

In this scenario the business understanding of the problem governs as to how many clusters need to be created. For example, if audio a streaming wants to launch 2 separate platform from existing customer database and wants to catch 2 groups of people and cater their needs. Then we need to make 2 cluster and follow 2 business insight.

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

Converting all the columns in a dataset to a comparable scale before applying clustering algorithm is known as Standardization. This is done to offset the effect of very large-scale variables on those having low scale.

For example, if you have variable 'A' which in range 10,000 – 50,000 in a column and another variable 'B' in the range 0-10 . Now if clustering algorithms is applied and since it uses Euclidean distance to compute the similarity between two points, this measure would be heavily influenced by the 'A' column because of its larger scale. Therefore, for proper clustering, we use standardization - to bring the mean of each column to 0 and standard deviation 1.

We can use any other form of normalization as well to bring all variables in same scale, example Min-Max scaler.

e) Explain the different linkages used in Hierarchical Clustering.

Single Linkage: The distance between 2 clusters is defined as the shortest distance between points in the two clusters.

Complete Linkage: The distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.

Average Linkage: 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.