Question 1: Assignment Summary

Solution.

As per the requirement, we had to find the list of countries that need the aid most. This can be measured as the country which has less money, so the income or gdp per person should be low.

Now the initial EDA shows that the country with more gdpp or income, invests more in health. Thus they have a low child mortality rate. So we can finally claim that the countries with high child mortality rate(child_mort) and low income and gdpp values are the ones we are looking for.

So I approached the problem as a clustering problem as we are not sure of the final outcome.

As we have the percentage data for imports, exports and health, I multiplied them by the gdpp to get the actual values.

Initially I handled the outliers by capping them for 99 percentile. After handling them, I got a balanced datasetFor clustering, I used 2 approaches.

1. K-Means Clustering
2. Hierarchical Clustering

K-Means Clustering:

Using Elbow Curve as well as Silhouette analysis, I finalised the cluster size as 3. Once we had the cluster size, using KMeans approach, I divided the data into 3 different clusters. Out of these clusters, 1 had the group of countries with low income and gdp per person but high child mortality rate.

Hierarchical Clustering:

Using Complete linkage dendogram, the cluster size came down to 3. On division of clusters, 1 of the clusters had the required low income and gdpp with high child_mort

So on analysing the results from both the methods, following is the list of top 5 countries in the order of increasing income/gdpp and decreasing child_mort

1. Liberia
2. Burundi
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone

Question 2: Clustering

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

Solution

- The time complexity in case of K Means is linear, O(n), while that of hierarchical clustering is quadratic, O(n2)
- For large data, k-means clustering is used. We can’t use Hierarchical clustering.
- For K-Means, output depends on initial choice of clusters. Whereas output of hierarchical clustering is always the same.
- To find the k value for k-means, we need to decide initially the number of clusters. But in case of hierarchical clustering, dendogram is used to find the number of clusters.

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

Solution

Step 1: Initialization

The first thing k-means does, is **randomly** choose K examples (data points) from the dataset as initial centroids and that's simply because it does not know yet where the center of each cluster is. (a centroid is the center of a cluster).

Step 2: Assignment:

Now we calculate the Euclidean distance of each data point from each cluster centroid. All the data points that are closest to a centroid creates one cluster. Then for each cluster a new centroid will be calculated. This will be the mean of the new cluster.

Step 3: Move the centroid

Now we will move the old centroid to new centroids.

Step 4: Repetition

Now we will keep repeating steps 2 and 3 until the centroids stops moving or in other words the K-Means algorithm converges.

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

Solution:

There are different ways in which we can compute the first 'k' for clustering.

1. Elbow Method
    a. We compute clustering algorithm (k-means) for different k values
    b. For each k, compute within cluster sum-of-squares (wss)
    c. Plot the curve of wss vs k values.
    d. The point at which curve takes a bend (knee) is generally chosen as the optimum k value for clustering
2. Average Silhouette method
    a. We compute clustering algorithm (k-means) for different k values
    b. For each k, compute the average silhouette score
    c. Plot the curve of average silhouette vs k values.
    d. The point at which curve takes a bend (knee) is generally chosen as the optimum k value for clustering

Silhouette Score

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

- p is the mean distance to the points in the nearest cluster that the data point is not a part of
- q is the mean intra-cluster distance to all the points in its own cluster.

Business Aspect

There can be multiple bend points in the curve, but at that time we have to look from the business domain point of view at what can be the ideal size of the clusters. Too many clusters is generally not advisable.

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

Solution

Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:

- Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.
- The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.

Solution

There are 3 different type of linkages used in Hierarchical clustering

1. Single Linkage
2. Complete Linkage
3. Average Linkage

Single Linkage

In single linkage, the distance between 2 clusters is defined as the shortest distance between points in the 2 clusters

Complete Linkage

In single linkage, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.

Average Linkage

In single 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.