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 graphs or for this question. Just text should be enough.

Answer

Clustering of countries Assignment is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most.

Started off with the necessary data inspection and EDA tasks suitable for this dataset - data cleaning, univariate analysis, bivariate analysis ,and outlier handling.

Then we recalculated the actual value of the columns Export, Import and heath that were given as the percentage of GDPP in the original dataset

Outlier Handling :

For columns such as child_mort, inflation, total_fer, you should not do anything to the upper range outliers but you may deal with the lower range outlier(capping).

But for rest of the columns, you should not do anything for the lower range outliers but you may deal with the upper range outliers(capping)

Scaling:

Scaling was done on columns except Country

Hopkins Test:

Then we passed the data from Hopkins test and got the score >80 on average for 10-Tests.

Finding the Optimal K:

To find the optimal K, for K mean clustering we performed SSD and Silhouette Analysis and observed k=3 to be the optimal K value.

Perfoming the clustering & Cluster profiling:

After selecting k=3, we performed kmean and checked the cluster profiles based on GDPP, child_mort and income.

Cluster with cluster_ID = 0 was such a cluster we were searching for, with 48 countries

Then we sorted the countries wrt GDPP & income in ascending and child_mort in descending, to get the list of top countries that were of concern for AID.

We did Hierarchical Clustering and achieved similar results, here cluster_labels = 0 , was the one with the desired countries and then we performed the Profiling

Following are the countries based on the analysis that are in direst need of aid :

['Burundi',

'Liberia',

'Congo, Dem. Rep.',

'Niger',

'Sierra Leone',

'Madagascar',

'Mozambique',

'Central African Republic',

'Malawi',

'Eritrea']

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

Answer

a) K-means clustering produced clusters of 91, 48 & 28 countries whereas Hierarchical produced clusters of 148, 18 & 1 countries.
b) Step 1: Choose the number of clusters k

Step 2: Select k random points from the data as centroids.

Step 3: Assign all the points to the closest cluster centroid

Step 4: Recompute the centroids of newly formed clusters

Step 5: Repeat steps 3 and 4. Till there is no further assigning of new centroids happen.

c) K in K-mean is selected statistically, by SSD (elbow curve) or Silhouette score

SSD considers the inertia/distortion of the data points and select the cluster number where it reaches the minimum or with a sudden break therefore elbow type curve.

The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. So we select the no. of clusters K with highest score.

In business terms we select the K value based on the business requirements. That means clusters that have low health spending & low GDPP; others where child_mort is low because of low income irrespective of health spending; etc.

d) Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance

e) Single-Linkage

Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage

Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

Average-Linkage

Average-linkage is where the 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. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

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 similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn't arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.