

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)

Answer:

Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and heath factors which will show overall development of the country.

So, in order to do so first I analyze the dataset – It has total 167 countries with no missing values, I transformed some of the variables like health, imports and exports from % of GDP to 5 of GDPP as it makes more sense to the dataset. Then I visualized the data by finding the correlation between variables to check the multicollinearity and found that most of the variables are having multicollinearity.

Now I checked the outliers of the dataset and I found there are outliers on all the variables. I checked this by describing the dataset in percentiles and by plotting boxplot. So, after this I made two approaches – one with outliers and another without outliers. By this way I will get an idea about the dataset and will choose the final countries from any one of them.

Again, same thing I did with hierarchical clustering and found different results than k-means. So, from this I have to choose which method gave us correct outcome and I have choose k-means output because even though for less records hierarchical clustering have much advantage than k-means, I have to check the business requirements and more precise information is giving from k-means.

Question 2: Clustering

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

K-Means Clustering	Hierarchical Clustering
We need to have desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights
It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
Works very good in large dataset	Works well in small dataset and not good with large dataset
The main drawback of k-Means is it doesn't evaluate properly outliers.	Outliers are properly explained in hierarchical clustering
K-means only used for numerical.	Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.

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

Step 1: Randomly select K points as initial centroids.

Step 2: All the data points closet to the centroid will create cluster center according to Euclidean distance function.

Step 3: Once we assign all the points to each of k clusters, we need to update the cluster centers or centroid of that cluster created.

Step 4: Repeat 2,3 steps until cluster centers reach convergence.

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

'K' value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of 'k'. for example, we have a dataset of variables like 'pen', 'pencil', 'books', 'notebooks', 'mobiles', 'charger', 'laptop'. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on business aspect, after viewing the dataset we can easily make cluster = 2, one in electronics category and another non-electronics.

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

It is definitely a good idea to do scaling/standardisation because our variables may have units at different scale and as our method stresses more on calculation of direction of

space or distance, so if we have one variable with high scale units then while calculating for k-Means or hierarchical it will create a big difference as the clusters will tend to move with the variables having greater values or variances. By applying standardisation/scaling will increase the performance of our model.

E) Explain the different linkages used in Hierarchical Clustering.

Linkage is a technique used in Agglomerative Clustering.

Linkage helps us to merge two data points into one using below linkage technique.

Single linkage: The distance between two clusters is calculated by the minimum distance between two points from each cluster.

Complete linkage: The distance between two clusters is calculated by the maximum distance between two points from each cluster.

Average linkage: The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.

Ward linkage: The distance between clusters is calculated by the sum of squared differences with all clusters.