**Clustering Assignment**

Assignment: Part II

**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.

Answer. HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Our role is to analyze the given data among 167 countries and we need to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Steps for clustering are as follow.

1. Started by collecting and reading the given on the jupyter notebook.
2. Perform EDA on the to fetch some useful insights on the data and to find out if the data is suitable for clustering or not.
3. Preprocessing: In this step we skipped the treatment of outliers as dropping or capping the data was affecting our data as we need to find out the countries which are in dire need on aid.
4. Perform Hopkins test to determine if clusters could form or not.
5. Scaling the data.
6. Finding the optimal K value for clusters with business aspect, silhouette score and elbow curve.
7. Form clusters and do complete analysis and visualization of clusters.
8. Profiling the clusters.
9. After that we perform hierarchal clustering with single linkage and complete linkage and then perform the visualization and profiling on the clusters.
10. After all these steps we get top 10 countries who are in dire need of aid.

**Question 2: Clustering**

1. **Compare and contrast K-means Clustering and Hierarchical Clustering.**

Answer.

1. K-Means Clustering, it needs desired number of clusters. While in Hierarchical Clustering, number of clusters can be decided after the completion of plotting dendrogram , that to by cutting it at different heights.
2. K-means requires a prior knowledge on number of centroid whereas Hierarchical cluster doesn't require any parameters. Instead, cut\_tree() is used.
3. K-Means clustering works on a very large dataset while Hierarchical clustering works on a small dataset.
4. K-Means is only used for numerical while Hierarchical is used when we have variety of data.
5. In comparison to Hierarchical clustering, K- Means is faster.
6. K-Means doesn't evaluate outliers properly in comparison to Hierarchical clustering.
7. **Briefly explain the steps of the K-means clustering algorithm.**

Answer.

1. K points are selected randomly as an initial centroid.
2. All the data points that are closer to the centroid will create cluster centre according to Euclidean distant function.
3. Once all the points are assigned to each of k clusters, the cluster centres or centroid has to be updated.
4. Above second and third steps should be repeated until cluster centres reach convergence.
5. **How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.**

Answer.

As per the statistical aspect, K value is chosen randomly in K-Clustering. And it can be performed using Elbow Curve and Silhoutte Score. If it's about business aspect, comprehension of dataset is important. In that sense, we can decide how many clusters to be taken.

1. **Explain the necessity for scaling/standardisation before performing Clustering.**

Answer.

Though clustering techniques use Euclidean Distance, it will be wise to scale the variables. Ex:- heights in meters and weights in KGs before calculating the distance. It may create a big difference while calculating for K-Means and Hierarchical. This is because the cluster will tend to move variable having greater values or variances. In fact, most clustering algorithms are even highly sensitive to scaling. Rescaling the data can completely ruin the results.

1. **Explain the different linkages used in Hierarchical Clustering.**

**Answer.**

In Agglomerative clustering, linkage plays a vital role in merging two clusters into one.

Different linkages that are used in Hierarchical clustering are as follows:

1. 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.
2. 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.
3. 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.
4. 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.