Clustering Assignment: Part II

1. Assignment Summary:

Problem Statement: To cluster the countries using some socio-economic and health factors that determine the overall development of the country., and find out top-most countries who are in dire need of aid, so that the CEO of the NGO can decide how to use the received financial package strategically and effectively.

Solution Methodology: In order to cluster the data for optimum numbers, followed below steps.

I. Understand Data, and check for any abnormality in data.
II. Inspection of missing values & outlier, and treat them accordingly.
III. Did not treated upper outlier of child mortality, lower outlier of income and GDP as those are important pints which will give the projection of countries which need aid.
IV. Exploratory Data Analysis, using pair plot and scatter plot as there were only continuous data.
V. Preparing Data: Checked the data, if it is compatible for clustering or not, using HOPKINS score
VI. Hopkins score were calculated 10 times by random selection of data. Conclusion: Each time, it is more than 80 hence data seems to be good for clustering.
VII. Scaling the data, because of Income and GDP columns as values are really huge as compared to the values of rest of the columns. Used Standard Scaler to scale
VIII. Once data is scaled, checked for optimum number of clusters using technique of Silhouette Score and Elbow curve. And result received from both, and as per our business need concluded that optimum number of cluster can be 3.
IX. Finally applied K-Means on the scaled data for K=3, and obtained top 10 countries who are in dire need.
X. Applied Hierarchical clustering (single method and complete method) and it turns out that it shows only 2 clusters that are distinct so revised our number of clusters from 3 to 2 for this modelling technique, and obtained top 10 countries who are in dire need.
XI. Final result: There is no difference in the end result of the countries which we want to select. Hence both method is correct, still if I would prefer, it would be K-Mean method.

2. Clustering:

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

Performed both methods top cluster the data points given, however found no such difference, as results obtained from both datasets are same. Still there are some differences between two.

K-Mean clustering is centroid based or we can say partition based method, however hierarchical clustering is hierarchy based agglomerative. Method to find optimum number of clusters is different in both, as in K-Mean checked with Silhouette score and elbow curve and concluded that optimum number of clusters will be 3, however in case of hierarchical clustering used Dendrogram, and revised assumption for this method from assumed number of clusters of 3 to 2.

2. Briefly explain the steps of the K-means clustering algorithm.

K-means clustering algorithm is used for unsupervised learning and cluster the data-points. K-means clustering aims to find the set of k clusters such that every data point is assigned to the closest centre, and the sum of the distances of all such assignments is minimized.

For that first number of clusters is selected, after that k random points from the data set. Then all the data points are assigned to the closest cluster centroid, then recomputed the centroids of newly formed clusters and this process iterates till a specific number of times which is given by user or until there is no change in the newly formed centroids.

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

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

Value of K is selected after the analysis of Silhouette score and Elbow curve.

Ideally maximum Silhouette score is selected for the number of clusters, but as 2 clusters will not give much of information therefore we select more than 2 and also analyse the Elbow curve, after which point there is no significant change.

Hence, most importantly it depends on the business need, so with the analyses of all the three we select number of clusters.

4. Explain the necessity for scaling/standardisation before performing Clustering.

As, we know K-mean clustering algorithm's operation is performed on the distance between two data points, hence it is dependent on the distance between the points. Therefore there should not be much of difference in the values of different columns, as it will drag the centroid which will disturb the cluster formation.

5. Explain the different linkages used in Hierarchical Clustering

Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster.

In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest maximum pairwise distance

In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest minimum pairwise distance).