CLUSTERING: SUBJECTIVE ASSIGNMENT

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:

The business problem is to categorize countries for the client (Help International) based on overall development using some socio-economic and health factors and then suggest the countries that are in the direst need of aid that the CEO can focus on. EDA was conducted on the data which included Data Inspection to learn the data types, dimensions, null values and statistical descriptions of numerical variables. Univariate & Bivariate Analysis was then carried out for further understanding trends in data by using visualizations that included scatterplots, boxplots, histograms and a correlation matrix based heatmap. This was followed by preparing the data for clustering which included Outlier Analysis & Treatment as well scaling of numeric variables.

Prior to clustering, Hopkins test was run a few times to check for tendency of clustering in the data. Two clustering techniques were employed – k-means and Hierarchical clustering. With k-means, elbow curve and silhouette score were used to arrive at an ideal number of clusters using a statistical approach, in our case – **k = 3**. The analysis of the 3 clusters indicated that there was one cluster that clearly stood out by poorly performing across key indicators (gdpp, child_mort & income). I then moved on to Hierarchical clustering where I employed 'Complete Linkage' as a method of distance measuring between clusters. Here, based on the dendrogram and using external references (for lack of business inputs), I arrived at 5 clusters - **4 primary clusters** with the 5th one having only one country.

The results of both clustering methods were compared and it was concluded that results from Hierarchical Clustering will be used for the following reasons:-

1. Hierarchical Clustering is known to be more accurate than K Means Clustering.
2. The results from the Hierarchical Clustering conformed to the external research which indicated that, in terms of development, countries were divided into 4 segments.

The final list of 5 countries were derived from 2 Clusters (as one had only 3 countries) by sorting them basis the important indicators – gdpp,child_mort and income.

Question 2: Clustering

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

Answer:

Comparisons between the 2 clustering methods are as follows:-

a. K-means clustering is done with clusters pre decided while in Hierarchical clustering, the number of clusters can be decided post interpretation of dendrogram.
b. K-means is suited for larger datasets as its less time consuming unlike Hierarchical clustering.
c. Hierarchical clustering results are reproducible while in K-means the initial set of clusters are random and therefore the clusters can differ a bit.

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

Answer:

Following are the steps of K-means algorithm:-

a. The number of clusters, K, is decided (using business inputs, silhouette score, elbow curve etc.)
b. K-means ++ algorithm is used to select the initial set of clusters.
c. Optimization of centroid: A new centroid is then calculated within each cluster by calculating mean of all data points within respective clusters.
d. Reassignment of centroid: New clusters are formed basis the newly formed centroids. This is done by calculating the Euclidean distance of each datapoint to all new centroids. Datapoints get assigned to the closest centroid and become new clusters.
e. Step c and d are repeated until there is little to no difference in the clusters formed.

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

Answer:

1. From a statistical approach, we usually use both the Silhouette Score and the Elbow Curve to arrive at an optimal number. Both methods test the effectiveness of k clusters by iterating through several number of clusters (i.e k = 2,3,4,5,6,7...). Elbow Curve uses the measure of Sum of Squared Distances between datapoints and clusters while the Silhouette score takes into account intra cluster distance between datapoints and inter cluster distances.
2. From a business standpoint, the number of clusters should make business sense. For eg – 7 clusters being statistically ideal may not make sense to business strategy or business categorizations. Therefore, the business aspect is of high importance when deciding k.

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

Answer:

Variables in the data could be of different magnitudes. This could result in the clustering algorithm developing bias towards variables of higher magnitude. Scaling/Standarization solves this problem by bringing all numeric variables to a comparable scale. One of the popular methods to do this is Standard Scaler.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:

There are 3 types of linkages that are used to form clusters in Hierarchical Clustering :-

a. **Single Linkage:** Used to calculate the shortest distance between a pair of observations in 2 clusters. Clusters with the shortest distance are merged to form a new cluster. This method can result in observations within different clusters that are closer than observations within the clusters itself. Such clusters can seem spread out.
b. **Complete Linkage**: Here, the distance between clusters are calculated using the farthest points between pairs of clusters. Clusters with the shortest distance are then paired. This method can result in tighter clusters.

c. **Average Linkage:** Here, the distance between two clusters is defined as the average of distances between all pairs of datapoints, where each pair is made up of data point from each cluster. The distances are then added up and then divided by the number of pairs to get an inter cluster distance. The clusters at shortest distance are then merged to form a new cluster.