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)

Answer:
The main objective of this assignment is to find the countries that are in the direst need of aid. The countries found using socio-economic and health factors reflect the overall development of country.

Whole process was split across 10 steps.

Step 1: Reading and understanding the data.
The dataset and data dictionary were loaded in their dataframe.

Step 2: Data Cleansing.
I performed basic cleansing steps to check including missing values, any duplicates and then outliers on gdpp, income and inflation

Step 3: Data Visualization.
A heatmap and a pairplot were generated to understand attributes dependency(multicollinearity), from which could infer

- child_mortality and life_expentency have high negative correlation
- child_mortality and total_fertility have high positive correlation`
- imports and exports have high positive correlation.
- life_expentency and total_fertility have high negative correlation.

Step 4: Data Preparation.
Variables like exports, imports and health were converted from percentages to absolute values and column country was removed to include only data columns in dataframe. To perform standardization of data, Standardization technique was used for scaling.

Step 5: PCA Application.
I did PCA (Principal Component Analysis) because I wanted to remove the redundancies in the data and find the most important directions where the data was aligned. After PCA application, I got 9 features or components, as PCA creates components equal to or less than the variables present in the dataset. Then Scree plot was created to choose number of components at 93%, the ideal number of components turn out to 3. So, in the rest of process, the number of principal components equal to 3. Then the pair plot was Principal Component 2 & 1 and Principal Component 3 &1. The first pair plot covered life expectancy, income, gdpp, health through PC1, import and export via PC1 and PC2, child mortality and total fertility via PC2. The second pair helped us infer that inflation is explained as part of PC3.
After this I performed outlier analysis and boxplot were plotted with and without outliers.

Step 6: Hopkins Statistics Test.
I used this test to measure the cluster tendency. Since the value was between 0.77 and 0.9, there's high chance of clustering and it was confirmed with the score.

Step 7: Model Building
I used both K-Means and Hierarchical approach to build the model. I used Silhouette Analysis, Elbow curve to decide the number of cluster to be between 4 or 5 for K-Means clustering. First I did my analysis, with 4 clusters and then I switched to 5 clusters. Similarly for Hierarchical Clustering, I cut the dendrogram at 5 cluster with both single linkage and complete linkage.

Step 8: Final Analysis using Hierarchical Clustering
The analysis was performed using the dataframe created during model building of Hierarchical clustering model building process. The final list of countries turned out to be 48 which are in direst need of aid.

Step 9: Final Analysis using K-Means Clustering.
The analysis was performed using the dataframe created during model building of K-Means clustering model building process. The final list of countries turned out to be 48 which are in direst need of aid.

Step 10: Closing Statement
Final list of countries was published and few analysis were performed.

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/ standardization before performing Clustering.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:

a)

K-Means Clustering	Hierarchical Clustering
Need to decide on desired number of clusters ahead of process	Decide the number of clusters after completion of plotting of dendrogram by cutting dendrogram at different heights
K-Means Clustering produces a single partitioning	Hierarchical Clustering can give different partitioning depending on level of resolution we are looking at
Works good with large dataset	works well in small dataset and not good with large dataset
it doesn't evaluate outliers	outliers are properly explained in Hierarchical clustering
K-Means Clustering is only used for numerical	Hierarchical Clustering can be used when we have variety of data

b)

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here the steps of the algorithm are:

1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all cluster members.
4. Now re-assign all the data points to the diffrent clusters by taking into account the new cluster centres.
5. Keep iterating through the step 3 & 4 until there are no further changes possible.

At this point, you arrive at the optimal clusters.

c)

'K' value can be chosen randomly in K-Means clustering based on statistical aspect. From business aspect, understanding of domain and dataset is helpful and helps decide number of 'K'. We can use below methods to find K in K-Means:

- The Elbow Method
- The Silhouette Method

d)

Standardization is an important step of Data processing. It controls the variability of the dataset, it converts data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms. Because our variables may have units at different scale and 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.

e)

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.