Assignment Part -2

Question 1: Assignment Summary

Ans: The main objective of us in this assignment is to find those countries which are in dire need of money because of socio-economic and health factors.

So after reading the data I found out that there are 167 countries and 9 features(columns). I have performed a data quality check to find whether it has any outliers or not, and then performed some EDA on it, changed the Income, Imports and Exports columns into absolute values from percentages of gdpp. Then I have plotted the graphs of all the features using boxplots to find the outliers. Later heatmap was shown to know about the correlation of all the variables. Later Hopkins test was performed on the dataset to check the randomness of the data. Scaling of the data is done in order to standardize all the features.

Now using the K-Means algorithm and using ssd/elbow curve and the silhouette score I have selected the number of clusters as 4. Using k=4 and after clustering them I have done cluster profiling using GDPP, Income, Child Mortality. Using K-means I have listed down the bottom 5 countries.

Now using the Hierarchical clustering by both single and the complete linkage I have clustered the data set. Using the single linkage and by selecting the number of clusters as 2 I have listed down the bottom 5 countries and now using the complete linkage I have listed down the bottom 5 countries.

So after both the analysis I have found out that both the K-Means and Hierarchical(complete) methods gave me the same result.

Question 2: Clustering

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

Ans:

K-Means	Hierarchical
1. We need to assign random number of clusters.	1. We can decide the number of clusters after completion of dendrogram and cutting the the clusters.
2. It works on larger dataset.	2. It works on smaller dataset
3. In this distance between points is calculated and the clusters are formed	3. In this clusters are formed like tree structures and most similar structures falls under one branch.
4. K-means is only used for numerical type of data	4. Hierarchical data is used for any type of data.

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

Ans:

1. Select n points as centroid randomly

2. Data points closest to the centroids forms a cluster based on the Euclidean distance.

3. Once all the points are assigned to n centers, update the centers or centroid of the clusters created

4. Repeat the process until there is no change in the centroid values.

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

Ans: K in K-Means is selected randomly in statistical aspect, after that we will apply elbow curve, silhouette score to find out the optimal number of clusters. But from business point of view we need to understand the data and then decide the optimal K and shouldn't only depend on the statistical results.

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

Ans: Scaling or Standardization is a very good process to fit the data into actual clusters where those belong. If we don't scale there might be a difference of units among variables and because of which the higher unit variables tend to cluster into one group which might be wrong clustering, so if we scale the values all the values comes under same scale and it increases the performance of the models.

e) Explain the different linkages used in Hierarchical Clustering.

Linkage describes the different approaches to measure the distance between two sub clusters.

1. Single Linkage: The distance between each cluster is calculated by minimum distance between two points from each cluster.

2. Complete Linkage: The distance between each cluster is calculated by maximum distance between two points from each cluster.

3. Average Linkage: The distance between two clusters is the average distance from each and every points of both the clusters