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 1:

PROBLEM STATEMENT : Categorising the countries on the basis of socio-economic and health factors that determine the overall development of the country to help the CEO of HELP International decide which countries to provide the raised funds that are in direst need of aid.

ASSUMPTION FOR BUSINESS ASPECT: If GDPP and Income of any country are high, then that country would not be in any dire need of financial aid. So countries whose GDPP and Income are lowest and Child mortality rate at the highest, would be in the most severe need of financial aid to provide for their citizens.

SOLUTION METHODOLOGY :

- First I read the dataset and converted the columns such as Exports, Health and Imports from percentage of GDPP to their absolute values.
- Next, I performed univariate analysis for 3 variables namely GDPP, Income and Child Mortality rate using a distplot and boxplot. This is where I came to know about the outliers in the dataset and thus decided to treat those by capping at 5% and 95%.
- Then I performed the bivariate analysis using the pairplot, correlation matrix and heatmap. Then I calculated the Hopkins statistic score and came to a conclusion that the dataset is good for modelling purposes because of the score of 0.96.
- Next I scaled the data using standard scaler method and began with KMeans clustering algorithm.
- In KMeans clustering, I started with silhouette score analysis and elbow curve analysis wherein I figured out that K=3 is an ideal cluster count to proceed with the analysis because of its readability and strong association with the logic of business aspects as it clearly differentiated the countries into 3 clusters of developed, developing and under developed countries.
- As we needed the top 5 countries to provide aid to, so the final cluster of k=3 consisted of the following countries that were in the direst need of aid:
    - Central African Republic
    - Congo Dem Rep
    - Niger
    - Mozambique
    - Burundi

- Next I performed Hierarchical clustering analysis with single linkage and complete linkage methods. Here at first I came with 3 clusters, but the resulting country with high child mortality rate and low GDPP & low income was only Nigeria which was inappropriate.
- Hence I decided to go for 5 clusters and came up with cluster_label = 0 wherein I received the exact same countries that I received in KMeans clustering.
- Thus, it concluded my analysis because both the clusters in KMeans and Hierarchical the business aspect was being satisfied.

Question 2: Clustering

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

The most important difference between the above 2 clustering methods is that KMeans provides 2 methods which clearly indicate the final clusters i.e. Silhouette score analysis and Elbow curve analysis. Where as the Hierarchical clustering method has only one method which gives a slight clear view of final clusters to be formed i.e. Complete linkage method.

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

STEP 1: Either choose the K arbitrarily or with the help of silhouette score and elbow curve.

STEP 2: Perform value counts for the said clusters.

STEP 3: Perform scatter plot analysis.

STEP 4: Perform cluster analysis and choose the final variables.

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

The value of k is chosen by visualizing the silhouette curve and elbow curve. If the value of KMeans clustering is not showing any major change after the Nth number of iterations, then that preceding value is the value of K. Furthermore, according to the business aspect the value of K has to be clearly readable and logical.

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

The most important reason for performing scaling of the dataset is to bring all the values to the same scale to correctly interpret the results of the analysis. If scaling wouldn't have performed, then the analysis of the plots would not be readable and interpretable and it may also result in the wrong analysis.

e) Explain the different linkages used in Hierarchical Clustering.

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.

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.

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.