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)

- The problem statement specifies that an international humanitarian NGO wants to identify the countries which are in direst need of aid so that it can channel its collected funds in the right direction. We had to categorise the countries using some socio-economic and health factors that determine the overall development of the country and suggest the countries which the CEO needs to focus on the most.
- I started off with understanding the dataset at hand. Once a fair understanding of the data was established, I moved towards cleaning it where a check was made to see if there were any missing values in the columns and performed other required clean-up activities.
- Further I did some EDA on the data to understand relationship of one column with the others. We also checked for outliers and treated them making sure no valuable data is lost in the process.
- Post analysing the data I checked for the Hopkins score to see if the dataset was good enough for clustering or not. Verified the Hopkins score to be more than 85 for at least 10 runs. Thus making sure that the data is eligible for clustering.
- Data was then scaled using the minmax scaling technique, post which I applied K-Means algorithm to form clusters. With the elbow curve and silhouette score, I figured that the data could be divided effectively in 3 clusters. Using K-Means clustering the data was divided into 3 clusters. While analysing the model, I observed that the cluster labelled 2 that was formed had a section of countries with low income, low gdpp and high child mortality rate, cluster labelled 0 being the right opposite, i.e., countries that have high income, high gdpp and low child mortality rate.
- The data was then sorted by descending order of child mortality, ascending order of income and gdpp to find out the countries in direst need of aid.
- Hierarchical clustering was also performed wherein we checked its single linkage as well as the complete linkage in the dendrograms. .
- The countries shortlisted (by both the clustering algorithms applied) for the funding are: Haiti, Sierra Leone, Chad, Central African Republic, Mali

Question 2: Clustering

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

- Hierarchical clustering cannot handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2). Every iteration or step that the hierarchical process runs it has to hold the similar size of data hence using homogeneous amount of memory. If you are using this on your system with bigger data, it will almost stop running. However, if the data is in the cloud system along with the software that is running the hierarchical process, you will be able to use the clustering. Hence, we can say size is not a problem but the technology at hand is.
- Hierarchical clustering is comparatively more expensive than K-Means clustering.

- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means clustering requires prior knowledge of K i.e. number of clusters you want to divide your data into. However, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

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

- K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps:

    - Mention the desired number of clusters
    - Choose k(no of clusters) initial centroids
    - Assign the points to the nearest centroid.
    - Recompute the centre by calculating the mean of the data points in each cluster.
    - Re-assign each point to the closest cluster centroid
    - Repeat steps 4 and 5 until no improvements are possible

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 can be chosen by using approaches such as Silhouette analysis and elbow curves.

Silhouette analysis:

It is a way to measure how close each point in a cluster is to the points in its neighbouring clusters. Silhouette values lies in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighbouring cluster and very close to the cluster it is assigned to. Similarly, value of -1 indicates that the point is close to its neighbouring cluster than to the cluster it is assigned to. And, a value of 0 means that it is at the boundary of the distance between the two cluster. Higher the value better is the cluster configuration. We want the clusters to have a low intra cluster distance and a high inter cluster distance.

Elbow curve:

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters.

Distortion is calculated as the average of the squared distances from the cluster centres of the respective clusters. Typically, the Euclidean distance metric is used.

Inertia: It is the sum of squared distances of samples to their closest cluster centre.
We iterate the values of k from 1 to 9 and calculate the values of distortions for each value of k and calculate the distortion and inertia for each value of k in the given range. It is possible to visually determine the best value for k. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k.

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

- Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

e) Explain the different linkages used in Hierarchical Clustering.

- Different types of linkages used in hierarchical clustering are:
    - Single linkage: Distance between the clusters is defined as the shortest distance between points in the two clusters.
    - Complete linkage: Distance between the clusters is defined as the maximum distance between any two points in the clusters.
    - Average linkage: Distance between the clusters is defined as the average distance between every point in one cluster to every other point in the other cluster.