Subjective Questions

Question 1:

Answer:

Problem statement: An International NGO, HELP International provides aid to countries that are in dire need of assistance. They also run operational projects time to time to raise awareness as well as for funding purposes. At a recent such programmes, the NGO was successful in raising $ 10 million and the CEO now has to decide as to which all countries need financial assistance. We thus need to analyse the dataset, cluster it to find out which all countries are not so well-to-do and can be helped.

Methodology implemented:

- The dataset was inspected and **no null values** were found so as to perform missing value treatment.
- Few variables, **Income**, **Health** and **Exports** were in percentages and they were converted back to actual numbers.
- Few variables **Exports**, **Imports**, **Health** and **Income** had outliers and they were treated by computing **Mean of the Outliers**.
- For **Child mortality**, **outlier treatment** was **skipped** as the higher values were nothing but the Countries that are in a severe predicament and need assistance.
- The dataset was then scaled using **Standard Scaler** and **Hopkins Statistics** was used to determine if the data showed a higher tendency to get clustered. **Hopkins Statistics score** came as **0.92**.
- **Sum of squared distance** (**SSD**) was measured and an Elbow curve was generated. With the curve we were able to infer that the **SSD** is **extremely high** for **2** clusters and it **decreases significantly** with an additional cluster.
- When we have **3** clusters, the **SSD** did not drop that significantly. I'd therefore choose **3** clusters and to confirm if **3** clusters are optimum.
- To confirm if 3 clusters were optimum, I generated Silhouette score for clusters ranging from 2 to 7. For **3** clusters, the **Silhouette score** was **0.47**.
- I have then made 3 clusters which later produced the mentioned list of Countries that are in dire need of assistance:- **Haiti**, **Sierra Leone**, **Chad**, **Central African Republic**, **Mali**, **Nigeria**, **Niger**, **Angola**, **Congo, Dem. Rep.**, **Burkina Faso**.

Question 2:

1)

- Hierarchical Clustering:
    - Hierarchical clustering is an algorithm that builds hierarchy of clusters.
    - This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm ends up merging when there is only a single cluster left.
    - We use dendrograms to interpret hierarchical clustering.
    - A line is then drawn that cuts the dendrograms and this indicate the number of clusters that we wish to work with.
    - Hierarchical clustering generally produces better clusters, but is more computationally intensive.
- K-Means:
    - For K-means, below steps are followed:
        - Randomly assign each data point to a cluster.
        - Compute cluster centroids, which is nothing but the average of the cluster data points.
        - Re-assign each point to the closest cluster centroid and then re-compute cluster centroids till the point that the centroid won't change.
        - We try to minimise the intra-cluster distance and maximise the inter cluster distance

2)

- We first randomly assign each data point to a cluster.
- Then we compute cluster centroids, which is nothing but the average of the cluster data points.
- We re-assign each point to the closest cluster centroid and then re-compute cluster centroids
- Again the previous step is repeated till the point that the centroid of the clusters won't change.
- In a way, we try to reduce the intra-cluster distance and maximise the inter cluster distance

3)

a. For a certain dataset, we first try to find sum of squared distances for each cluster.

b. We find the optimum number of clusters with the notion behind it being that we SSD does not drop significantly. This is the statistical approach where we calculate SSD and plot an elbow curve.

c. Business approach includes opinions from the functional team, the client. Let's consider a case wherein statistically, i.e. with Elbow curve method, we found that the 4 clusters are optimum. But when conveyed to the client, they suggest it to be 3. Thus even if statistically we prove or we find that certain value for K is optimum, it also depends on the business problem that we are solving, the suggestions or the value suggested by the business.

4) It is extremely important to perform Scaling/Standardization clustering. I'll take an example from the case study. The data format for Imports, Exports, Inflation, GDP per capita are different. If the variables are not scaled to be in same format, then it would affect the cluster distance and therefore mess up the clustering analysis.

5)

a. Single linkage: In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.

b. In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.

c. In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster