Subjective Questions

Question 1.

Answer:

Problem Statement: HELP International is a global NGO which helps people who are in dire need of help, now the CEO of the NGO needs to decide to spend the money strategically to make most of it. As a data Analyst we have to find the countries which are be helped right away from the data available.

Steps to approach the solution:

1. Look into the Data available
2. Read the Data
3. Clean the Data if required and make necessary corrections.
4. Data Modelling
5. Clustering

As we readily have the data we directly had a quick glance to understand, and the data available is of countries, child mortality, income, GDP per capita etc. after understanding we can cluster the countries into different groups for further analysis in such scenario we are going to proceed with clustering techniques.

Firstly, reading the data and checking for null values info and shape of the dataset. As there was no null values and information available was correct with that checked for outliers using boxplot as the data has outliers and are skewed towards right. Removing the outliers is not a solution as we lose important data (each row of the data gives information of a country). So skipped outliers treatment.

Continued scaling the data via MinMax Scaler to handle the outliers and skewness of the data, After all the steps, continued doing KMeans clustering with max number of clusters with 4 and after plotting the elbow curve found final number of optimal clusters =3.

Now with Hierarchical clustering, plotted dendrogram for both single and complete linkages we got the optimal clusters = 3. Plotted scatter plots and to the clusters formed for visualization between the variable pairs.

The following are the conclusions obtained to from the plots:

Cluster-0: If a country has high mortality rate, there is low income and low GDPp.

Cluster-1: If a country has avg mortality rate, there is avg income and avg GDPp.

Cluster-2: If a country has low mortality rate, there is high income and high GDPp.

Finally sorted the top five countries (based on income, gdpp, child_mort) which are in dire need.

Question 2:

a. Compare and contrast KMeans clustering and Hierarchical Clustering.

KMeans Clustering	Hierarchical Clustering
1. KMeans is a method of cluster analysis using a pre-specified number of clusters. It requires advance knowledge of K.	1. Hierarchical clustering analysis is a method of clustering which seeks to build a number of clusters without having fixed before.

2. Using a pre-specified number of clusters, the method assigns records each cluster to find mutually exclusive cluster of spherical shape based on distance.	2. Hierarchical methods can be either divisive or agglomerative.
3. One can median or mean as a cluster centre to represent each cluster.	3. Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
4. K-Means clustering a simply a division of the set of data objects into non-overlapping such that each data object is exactly one subset.	4. A hierarchical clustering is a set of nested clusters that are arranged as a tree.
5. Convergence is guaranteed	5. Easy of handling, consequently, applicability to any attributes types.

b. Briefly explain the steps of KMeans clustering Algorithm.

KMeans is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a prior. The main idea is to define k centers, one for each cluster.
It consists of 3 main steps.
Step 1: Initialization
The first thing k-means does, is randomly choose K examples (data points) from the dataset (the 4 green points) as initial centroids and that's simply because it does not know yet where the center of each cluster is. (a centroid is the center of a cluster).
Step 2: Cluster Assignment
Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid.
Step 3: Move the centroid
Now, we have new clusters that need centers. A centroid's new value is going to be the mean of all the examples in a cluster.
We'll keep repeating step 2 and 3 until the centroids stop moving, in other words, K-means algorithm is converged.
K-means is a fast and efficient method, because the complexity of one iteration is k*n*d where k (number of clusters), n (number of examples), and d (time of computing the Euclidian distance between 2 points).

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

Elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm.
The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.
So average distortion will decrease. The lesser number of elements means closer to the centroid.
So, the point where this distortion declines the most is the elbow point.

Silhouette Method value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

Business Aspect:
Both the above methods give us a Statistical aspect of selecting optimal number of clusters. Sometimes, it is up to the Business type or domain which could be used to decide the number of clusters.
For example: In customer segmentation, banking team may decide upon previous data to decide upon the number of customer groups.

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

It completely depends on the data. Data standardization is about making sure that data is internally consistent; that is, each data type has the same content and format.
If you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 100 or 1000).
This importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical calculations.

e. Explain the different linkages used in Hierarchical Clustering.

There are 3 types of linkages in hierarchical clustering.

1. **Single-Linkage**

Single-linkage) 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.

2. **Complete-Linkage**

Complete-linkage 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.

3. **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.