Subjective Questions

Question 1: Assignment Summary

Problem Statement: As a Data Scientist, we need to find the countries in direct need and help. CEO of HELP International in using the fund money to reach right countries.

Solution:

As we have the Data of countries like child mortality rate, GDP Per Capita, Income etc., we can use Clustering to segregate the countries into different groups. In the data provided, all the features are right-skewed which indicates us that it contains Outliers. Removal of Outliers is not a feasible solution as we will lose data. We used Power Transformation to handle the skewness and also scaling.

After the pre-processing steps, used **Elbow Curve** to find the optimal number of clusters.

Found that 3 is the optimal clusters to be formed. Used **KMeans** and **Hierarchical clustering**

(Both *Single* and *Complete linkages)* to model and predict the labels. After plotting the clusters formed – scatter plots and box plots. Labelled 3 clusters as:

Cluster - 0: High Child Mortality, Low Income and Low GDP

Cluster - 1: Average Child Mortality, Average Income and Average GDP

Cluster - 2: Low Child Mortality, High Income and High GDP

We have got a list of 50 poor countries. We sorted the list based on - **income, gdpp, child_mort** and selected top 5 countries as countries in direct need.

Question 2: Clustering

A. Compare and contrast KMeans Clustering and Hierarchical Clustering

KMeans	Hierarchical
k-means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of 'K'.	Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.

k-means, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance.	Hierarchical methods can be either divisive or agglomerative.
One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
KMeans clustering can handle big data because the complexity is linear O(n)	Hierarchical cannot handle big data as the complexity is Quadratic i.e. $O(n^2)$

K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ	results are reproducible in Hierarchical clustering as it is upto us to decide number of clusters based on cutting the dendrogram

B. Briefly explain the steps of KMeans Clustering algorithm

k-means is one of the simplest unsupervised learning algorithms that solve the wellknown 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 appropriate.

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, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.

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, Kmeans 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 K-means clustering? Explain both the statistical as well as the business aspect of it.

There is a popular method known as **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.

We have another method called "**Silhouette** Method". The silhouette 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.

Both the above methods give us a Statistical aspect of selecting optimal number of clusters.

Sometimes, it is up to the Business teams to decide the number of clusters. For example, in customer segmentation, marketing team may decide upon previous data to decide upon the number of customer segments.

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

In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important 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., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space. When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters.

Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

Although standardization is considered best practice for cluster analysis, there are circumstances where standardization may not be appropriate for your data (e.g., Latitude and Longitude). Whether to use Standardization or not, it depends on the data.

E. Explain the different linkages used in Hierarchical Clustering.

The process of Hierarchical Clustering involves either clustering sub-clusters(data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster

into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two

sub-clusters of data points. The different types of linkages are: -

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 averagelinkage, 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.

Centroid-Linkage

Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn't arise in the other linkage methods because the clusters
being merged will always be more similar to themselves than to the new larger cluster