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)?   Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Ans: The problem statement is an NGO to choose the countries that are in the direst need of aid.

The solution that I followed is: I have prepared my data and Exploratory Data Analysis tasks which include - data cleaning, univariate analysis, bivariate analysis and visualize the data using plots for a better understanding of data.

Plots such as I have used distplot to identify net income versus total health spending, and GDP versus total health spending. Also, I did a pair plot, and to see the correlation between the different attributes heat map is done.

In Outlier Analysis I have performed skewness to interpret outliers. We get to see that in integer data type all columns are highly skewed which giving us a Hint that there would be Outliers. To confirm that we checked the Box Plots and histogram for these Columns. Once identified then we do capping because if we drop them then we may lose countries that are dire need of AID. Similarly, in float data type we follow the same method to treat outlier.

Then I have moved with the clustering procedure but over here I will first check my Hopkins score just to understand whether this data is good for clustering or not and it looks like we have a good cluster that can be formed using this data.

Then I scaled the data using StandardScaler. And then we will find out the value of K through Silhouette Score and Elbow Curve. And we got 3

We perform k-mean and try to visualize through cluster profiling, we are able to determine cluster 2 which satisfied our business requirement and able identify top 5 countries which are in dire need of aid based on some socio-economic and health factors

Though Hierarchical Clustering helps to visualize the clusters from a business standpoint and through complete linkage we can able to identify the right value of K.

Hence, K-means type of Clustering produced a better result and Hierarchical Clustering helps to visualize the clusters better

Question 2: Clustering

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

Ans: By Comparing, K-means handle data well give a better result whereas Hierarchical clustering helps to visualize the clusters better. K-means gave top 5 countries that are urgent need of AID which are Haiti, Central African Republic, Lesotho, Sierra Leone, Mali and Chad.

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

Ans: We find out the value of K through Silhouette Score and Elbow Curve. And we got 3

By considering the number of clusters (k) = 3 we perform k-mean. The cluster centers for each of the K clusters are randomly picked. Each observation n country is assigned to the cluster whose cluster center is the closest to it. The closest cluster center is found using the squared Euclidean distance. Then, again the cluster centers are computed. And algorithm achieves the optimal point or stops when further cluster center will not update.

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

Ans: By statistical method we find out the value of K through Silhouette Score and Elbow Curve.

In case of Silhouette Score, we will use that particular value k for which the silhouette Score is maximum. So, we can see maximum at cluster K=2 but going with k=2 is not a good idea because k=2 basically means that we are just dividing data into two half's and that's why 2 is not always taken as a number of clusters. So, we will go with the next highest value that is 3.

In case of Elbow Curve, goal is to choose small value of k which still has low sum of squared error and the elbow usually represents the point from where the returns start diminishing with increasing values of k

In business aspect we need to determine those countries which are dire need of AID and Country having high child_mort, low income and low GDP. So, I first refer performing Hierarchical clustering and in complete linkage visually get to know value of k

d) **Explain the necessity for scaling/standardization before performing Clustering?**

Ans: When we use squared Euclidean distance between data points. It is important to ensure that attributes with a larger range of values do not out-weight the attributes with a smaller range. Thus, scaling down of all attributes to same normal scale helps in this process.

Scaling helps in making the attributes unit-free and uniform. Standardized scaling, on the other hand, brings all the data points in a normal distribution with mean zero and standard deviation one. It is performed using the formula: X – mean(C)/SD(C).

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?**

Ans: Single linkage: shortest distance between two points in each cluster. It cannot separate clusters if there is noise between clusters but it can separate non-elliptical shapes as long as the gap between two clusters is not small

Complete linkage: longest distance between 2 points in each cluster. It does well in separating clusters if there is noise between clusters and it breaks bigger cluster

Average linkage: average distance between each point in one cluster to every point in the other cluster. It also does well in separating clusters if there is noise but it is biased towards globular structure