Name: Madhusudhan Anand
Email: maddymaster@gmail.com

Problem Statement: HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, they have been able to **raise around $ 10 million**. Now the CEO of the NGO needs to decide how to use this **money strategically and effectively**. The significant issues that come while making this decision are mostly related to **choosing the countries** that are in the direst need of aid.

My job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Question 1: Assignment Summary:

I took the countries dataset, read and understood the descriptions of the dataset. Performed the following tasks as a part of **data preparation**

1. At first checked that there are any missing values. There were none.
2. Also except columns data, everything else was numerical data which was good.
3. Checked for outliers, at different percentiles. There were outliers in some places.
4. Created box plots to check on suspected outliers.
5. Noticed the RFM that I would be using did not have any outliers. But thought of doing this at a later stage

Preparing data for modelling

1. Created RFM - Recency, Frequency and monetary
2. Recency is the average number of years a new born child would live if the current mortality patterns are to remain the same
3. Frequency is the number of death of children under 5 years of age per 1000 live births
4. Monetary is the GDP per capita, calculated as the total GDP divided by the total population. This is interestingly monetary not income is because this captures the overall monetary value, not just income which is the average of all the incomes.
5. Created monetary, recency and frequency data groups for every country and merged them

Outlier Treatment and rescaling

1. Checked using box plots, then used statistical outliers for RFM.

2. Then performed rescaling, instantiated the group
3. I could choose minmax or standard scaler, but standard scaler looked more relevant, min max compresses everything between 0,1 while StandardScaler removes the mean and scales the data to unit variance. However, the outliers have an influence when computing the empirical mean and standard deviation which shrink the range of the feature values.

Model - KMeans

1. To start with began with Kmeans
2. Started with arbitrary k, set n clusters to 4 and max iter to 50
3. Then performed fit, put out an elbow curve and as per **Silhouette Analysis,** The value of thescore range that lies between -1 to 1, a score closer to 1 indicates that the data point is very similar to other data point in the cluster. And -1 depics that they are not
4. Now that the clusters are created, plotted them in the box.

Understanding:

1. PCA is a linear transformation method and works well in tandem with linear models such as linear regression, logistic regression etc., though it can be used for computational efficiency with non-linear models as well
2. In the K-Means algorithm, you divided the data in the first step itself. In the subsequent steps, you refined our clusters to get the most optimal grouping. In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions/merges take place, which may run from a single cluster containing all objects to n clusters that each contain a single object or vice-versa.

Model - Hierarchial Clustering

Given below are five data points having two attributes x and y:

The distance matrix of the points, indicating the Euclidean distance between points, is as follows:

Label	1	2	3	4	5
1	0.00	3.00	2.24	3.61	5.83
2	3.00	0.00	2.83	3.16	3.61
3	2.24	2.83	0.00	1.41	4.12

4	3.61	3.16	1.41	0.00	3.00
5	5.83	3.61	4.12	3.00	0.00

Take the distance between two clusters as the minimum distance between the points in the two clusters.

Created complete and single linkage dendrograms and it was useful to see that we could cut at 3 clusters based on the complete linkage.

Now that the Hierachial clustering made more sense with the box plot concluding the clustering with a lot of clarity, it was clear that the countries that belonged to cluster label 1 need more attention or help

PCA
Hence we could conclude in PCA we can directly use the PCA function, we could perform fit and transform on the RFM data, using which we could verify the model. In the final calculation also it was verified for the 3 clusters.
PCA doesnt change the total variance of the dataset, it only rearranges them in the direction of maximum variances.
So plotted a correlation matrix, there were no correlations between any two components.

Woohooo - to PCA!

We have effectively removed multicollinearity from our situation and our models will be stable.

Then finally to put it all together, we now know which country is in the **Direst of need of help** grouped the country back with the clusters and plotted them to select cluster 1 and make recommendations to the CEO of HELP!

------------------------thats all folks----