Problem Statement:
An NGO has raised 10 million find resently, and the CFO has given us the responsibility to identify the country in dire need of the funds. The funds raised will be given to countries based on the prediction made by us.

My approach and solution:

1. The following parameter wrt 167 countries were given to us in the data frame for various countries: country name, child_mprtality rate, exports, imports, income, health, inflation, life expectancy, gdpp.
2. I started with the initial data analysis, where I identified the data type of the parameters. Then the number of null data were identified for these columns, there were no columns with null.Then I convert **exports, health, imports** of the data frame that were given as a percentage of the gdpp.
3. I plotted the data distribution of the various parameters and tried to analyse the relations between them. The I did the outlier analysis and capped the data falling greater than 99 percentile and less than 1 percentile.
4. Next I calculated the hopkins statistics that helps us identify if clustering is the right method and if the data can be clustered or not. I found out that the score ~0.91 which signifies there is a high chance of clustering. Then I did scale the data using standard_scalr.fit_transform.
5. The next step involved calculating the silhouette score for various number of cluster starting from 2 to 10., there I was able to draw meaning full insights on the number of clusters that we should go ahead with. The optimal number of clusters chosen was 3. Then we did plot the elbow curve, which represents the distance between the point and its assigned center, a lower value means the clustering is good but as the values decreases the number of clusters increases so we need to decide an optimal number.
6. With K=3 we proceeded and created the cluster ad assigned the clusters after dropping the country name column. I then plotted various scatter plots to find out the relatin that existed between the clusters.based on the graphs plotted I saw that the cluster with cluster id =2 were in dire need of this aid.
7. Based on my final analysis I found 10 countries than needed finacial aid.
8. The I applied hierarchical clustering to the data where I did perform both single and complete linkage clustering. Single linkage is based on the shorted distance between the points where as complete linkage takes into account the longest distance between the points.
9. Again after performing the analysis I was able to get the worst hit countries and these included: Burundi, Liberia, Congo, Niger, Sierra Leone, Madagascar, Mozambique, Central African Republic, Malawi, Eritrea.

Subjective questions on clustering part 2:

1) Compare and contrast K-means Clustering and Hierarchical Clustering.

Kmeans clustering is less process intensive as compared to Hierarchical Clustering, because in the former there are computations which compare the point with every other point in the plane. But we get a benefit here as we can visualize the formation of clusters as in we can cut the dendrogram in between and we can check the number of clusters present at that level. The in practical world makes more sense as e can identify the segment and not pick us any niche segment where it would be difficult for us to track. So after doing this we should identify the number of clusters and than based on the number of clusters we should be running the k means algorithm and find the clusters.

2) Briefly explain the steps of the K-means clustering algorithm.

The process starts by selection of k random points in the population based on the number of k. The next step is calculation of the distance of each and every point to the assigned center and summation. Then the points shift and the distance is recalculated and this process goes on till the formed clusters do not move any further and then we stop return back the result.The process can be visualized using a elbow curve.

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

Selection of a correct value of K is very important, if we select a larger number i.e more segments it would be difficult for us to manage the niche segments and concentrate, similarly if we choose a lower value of k we are actually not able to segregate the number of clusters because a single cluster in itself will contain a variety of population showing different behaviour. We can make use of the silhouette score and the elbo curve.As the curve drops this means the inter cluster distance decreases and the intra cluster distance increases i.e better clustering. We should consider both the value and no of cluster using that.

4) Explain the necessity for scaling/standardisation before performing Clustering.

We need to scale and standardise the data before passing it to the model because if there is a difference in tha scale of data the difference, distance and

the weight calculation would be inversely effected. So we perform min-max, or any other scaling so that the data is limited in the range and all parameters are given equal weightage.

5) Explain the different linkages used in Hierarchical Clustering.
Linkage are of three types single, complete, average. So in hierarchical clustering a point is selected and its compared with every other point in the data set so that we can find out he similarity of the data point to the other data points and can assign them the same cluster. So suppose if we start with 4 data points then in first iteration the point will be compared to all the points in the data frame, same will happen for the second and all 4. After the first iteration some clusters are formed then we need to compare the inter cluster distance than we make use of the above feature where we can consider the min, max and avg distances between the clusters respectively to come up with new cluster.