Q1. Assignment Summary:

Answer: The NGO has raised money to supply aid to most needed Countries, and there are multiple data points collected to perform an analysis to assist the most critical Countries, Countries which are in most need of Aid. In process of doing so the data has been checked and was found that there are certain features which are really high in numbers and those high numbers contribute to the goodness of the Countries. In our analysis, such data points are irrelevant as they do not add any value to the analysis of the countries which has really low GDP, High Mortality Rate, Low Income, high Inflation etc. This, I have decided to cape the extreme values that represent the goodness of the countries, there are features, for which, the extreme values point to a worse status of the country, features like these are not caped at their upper limit, and feature like these didn't have any outliers at the lower limit which would have indicate to goodness of the countries.

After the above mentioned EDA process, I decided to check the Hopkin's Statistics for the dataset which tells us how different is our data set from a random sample, and thus, if the returned number from this method is high then it means that the data is a good fit for clustering, I got a value of approximately 89% which indicates that this is good dataset for clustering.

After this, I used the Inertia and the Silhouette Score to determine the number of clusters, and found out that K=3 was a good count of clusters for the data. Going forward with that number, I created 3 clusters using KMeans, and Hierarchical Clustering on the dataset and then sorted the data from the cluster that had datapoints with very low income, GDP, high mortality rare etc.

Then I filtered out top 10 countries from each of the algorithm and found out that all 3 methods (KMeans, Hierarchical Clustering using single linkage and Hierarchical Clustering using complete linkage) returned the same top 10 countries after sorting them.

These countries are then mentioned in the presentation with detail analysis.

Q2. Clustering:

1. Compare and Construct K-Means Clustering and Hierarchical Clustering.

Answer: There is a basic difference in between the above mentioned methods, the K-Means clustering requires the number of clusters to be constructed at time of building the model which makes it less recourse hungry and this is why it is really good when clustering huge datasets.

On the other hand, Hierarchical clustering constructs trees based on Agglomerative or Divisive method, one is bottom up and the other one is the top down method consecutively. For this reason the clusters are not predefined and the algorithm itself is resource hungry, the clusters are created using a 'cut tree' method after creating dendrograms.

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

Answer: In this method, we have to determine the number of clusters(k) we need using the busyness need or the elbow method, after that the algo randomly or using K-Means++, initializes k number of datapoints and then, distance from each and every other point and each of this k number of point is calculated using Euclidean distance method. The new point is associated with the closest point and then the centroid is recalculated using the average distance between the points in a cluster, this process continues until the centroids doesn't converge any more or the maximum iteration limit is reached, whichever happens earlier. This is how the final clusters are formed.

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

   Answer: The value of K in KMeans clustering is chosen using the elbow rule. When we plot inertia in respect to number of clusters, we find that after a certain point the slope becomes less steep and creates an elbow shaped graph, this tells us, there is not much change in inertia after the point where the elbow is created, the value of K at that point is usually the best count of clusters for that particular data set, but there might be a certain scenarios where that number doesn't meet the busyness requirement, if the statical value of K is 5 for a dataset where only 3 types of data points(shirt size of Small, Medium & large) makes sense to busyness then we have to use K =3 though it doesn't make proper sense from a statistical point of view.

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

   Answer: Scaling is necessary before performing clustering as clustering algorithm rely on the calculation of distance(Euclidean or Manhattan or some other method) between datapoints to verify if those data points are similar in nature or not, if the data points are in different scale then the algorithm doesn't understand it as it takes the numeric value only, so, if a data is at an higher scale then it will have a lower numeric value though it cannot be taken as that the value is actually low. This confusion can be easily avoided if all the values are in same scale.

5. **Explain the different linkages used in Hierarchical Clustering.**

   Answer: In Hierarchical clustering there are total of three types of Linkage used.

   a. ***Single Linkage:*** Uses the minimum distance. What it means is that after calculating the distance from each of the data points from one cluster to each of the data points of another cluster, it chooses to use the lowest value as the measurement of distance.

   b. ***Average Linkage:*** Uses the average distance. What it means is that after calculating the distance from each of the data points from one cluster to each of the data points of another cluster, it calculates the average of all the values and then uses that value as the measurement of distance.

   c. ***Complete Linkage:*** Uses the Maximum distance. What it means is that after calculating the distance from each of the data points from one cluster to each of the data points of another cluster, it chooses to use the highest value as the measurement of distance.