Clustering Assignment Subjective

1. Assignment Summary

   Answer:

   Problem Statement: HELP humanitarian NGO has 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.
   Job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries, which the CEO needs to focus on the most

   a. After reading problem statement we analyzed the data provided. We checked for any missing values but it was not present
   b. Then we checked for outliers, there were few but we did not remove or cap them as we thought removing them might remove countries which might need funding
   c. Then we went for modeling, by Hopkins statistics we got result around 0.9 which suggested that the data was good for clustering
   d. Then we used k-mean algorithm, using elbow curve we found 3 and 4 clusters were good enough cluster size for the data which was confirmed by silhouette score
   e. Using boxplots we were able confirm our finding where we could see how some cluster had low gdpp and income and high child mortality and some with the other way round
   f. We did same with hierarchical clustering too and found 3 clusters were good for the data
   g. Now we took the cluster 2 and analyzed that. We sorted them according gdpp, income and child mortality
   h. We then tried to describe the data, found max, min and mean data were very different. So we took median and used it to get gdpp and income below median with child mortality above median. This was done as some countries with good gdpp and income but bad child mortality need less funding than the countries with less gdpp and income with high child mortality
   i. This helped us to get our list of poor countries and from that top 5 countries for funding

2. Clustering
   a. Compare and contrast K-means Clustering and Hierarchical Clustering.
      In K-mean algorithm we predecide the number of cluster and work towards it but in hierarchical we create dendrograms and cut it using horizontal line to get the cluster number
      K-mean algorithm is best for big data size with inferior hardware, but hierarchical clustering is best for small data size and requires superior hardware

b. Briefly explain the steps of the K-means clustering algorithm.
   1. Select no. of cluster and randomly select point to be centre of the cluster
   2. Mark distance of each point from centroid
   3. we again try to find the centroid by finding the point equidistance from all points.
   4. Keeping doing above steps till we find a centroids do not move anymore

c. How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
   K in K-mean clustering is chosen by intuitation.But there are ways to get it too. One being elbow curve method and other being using silhouette score. Sometimes it better to ask business and understand their domains to get the no. of clusters

d. Explain the necessity for scaling/standardisation before performing Clustering.
   Scaling and standardization is necessary in order to bring all the variables to same level. Else there might be scenario where different units are causing issues in calculation E.g earning of a person in rupees where are family income in dollar will cause issue in caluclations.

e. Explain the different linkages used in Hierarchical Clustering.
   There are 3 types of linkage in hierarchical clustering –
   1. Single – here minimum distance between cluster is used for creating graphs
   2. Complete – here maximum distance between cluster is used for creating graph
   3. Average – here average distance between all the points in both clusters are calculated for creating graph.