Question-1:

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(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Answer:

Problem statement is to help the CEO of HELP International to find the countries which are the most underdeveloped and which need the direst of help among the countries based on certain factors and decide how to use the money strategically and effectively.
The objective of this analysis is to categorize the countries using some socio-economic and health factors that determine the overall development of the country.
The methodology that has been used is **Unsupervised learning**, to be specific **Clustering.**
To help with reduction of predictor variables we used Principal Component Analysis.
After cleaning data and removing outliers from the dataset we used PCA on the dataset.
Then Visualized the components using Scree plots. From the Scree plot it was observed that 93% of the variance in the data could be explained using just 5 principal components. Thus, we chose **5 principal components** to perform PCA. We then used Hopkin’s statistic to find if the data can be clustered. We then used two different methods to find out the number of clusters that were optimal for the dataset. We chose Silhouette method and Elbow method. Both the methods resulted in the same answer. Thus, we took **2 clusters** and used both K means , Hierarchical Clustering. Upon analysis of the clusters we found that K means provided better clusters and thus decided to go with that clustering algorithm. We visualized using original variables and were able to find that **Cluster 0** contained the countries which were underdeveloped. We finally concluded based on Gdp per capita that **Madagascar,Mozambique,Malawi,Eritrea,Togo** are the countries that need the most help.We also observed countries that need the most help are African countries.

Question-2:

State at least three shortcomings of using Principal Component Analysis.

Answer:

Principal Component Analysis is a really good method to reduce the number of predictor variables. However it does have a few shortcomings ,explained as follows:

- Linearity : One of the limitation is that PCA by default assumes that the principle components are a linear combination of the predictor variables .It's not mandatory that principle components should form a linear combination and in such a case PCA will not provide the correct results.

- Large variance implies more structure : Another limitation of is that it uses variance as a measure to calculate the value or rank of each dimension. Thus it gives the variables with higher variance preference over those that have lesser variance. It uses covariance ratio to provide the principal components.

- Orthogonality: PCA assumes that all principle components need to be orthogonal that is at a $\mathbf{90^0}$ and that they are independent of each other. This is not always the case ,sometimes two variables may have a bit of collinearity but be required for the final model which will be automatically ignored by PCA

Question-3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

K-means Clustering uses the concept of centroid which has least distance from other points from that cluster. At first, we chose random k centers. We then find the distance from each of the point and based on the distance assign each point a cluster. We then take mean of the distance and change the point of the centroid and repeat the same process till the centroid doesn’t change. The advantage of K means is that It can be applied on a huge dataset. It’s easy to understand and efficient. However, its disadvantage is that it can be used on only numerical data, also it cannot handle outliers and number of clusters need to be decided beforehand.

Hierarchical Clustering on the other hand uses nearest distance to group points pairwise, then finds nearest distance from the pairs and so on until a hierarchy like structure is formed which is called a dendrogram. Its advantages are that it is flexible, can be used on categorical data no need to know the number of clusters to be used beforehand. However, it cannot be used in cases where there is large data as dendrogram will not be clear.