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)

Ans1)

This is our problem statement:

1. Help the HELP International’s CEO to make the right decision on how to use 10 Mn fund strategically and effectively on which countries.
2. Need to categorise the countries using some socio-economic and health factors that determine the overall development of the country.
3. Need to suggest the countries which the CEO needs to focus on the most.

<u>THE PROCESS</u>

1. First we read the data in python notebook and checked whether the data has null values.

2. Then we get the principal components using PCA and with the help of scree plot,
   And we see that around 88% variance is explained by 3 PCs,so we take principal Components= 3.

3.Then we plot a scatter plot of ‘PC1’ vs ‘PC2’ and see that Hopkins ratio coming out to be .82, so we can say that the data is good for clustering then we drew Silhouette Plot, Elbow Plot we got to know tha the correct numbers of clusters would be4.

4. After creating the principal components we treated for the Outliers Values because this affect the clusters and removed ouliers on gthe higher side i.e more than 95 percentile.

5 Then we have chosen K - Means Clustering and Hierarchical Clustering methods to create different clusters.

6) After when we made the clusters then we check for the countries which are below a certain criteria which we considered as a dire need of financial help.

Which type of clustering produced the better result

- We checked both the methods for clustering i. e Hierarchical Clustering and k-means clustering and both produced the results and as we can see from our analysis that countries pointed out by hierarchical method were present in result given out by k means ,so those countries were in fact common to both so we chose those countries.
- I would say hierarchical clustering is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.
- So,we can say that heirarchical method produced better results.

Question 2

State at least three shortcomings of using Principal Component Analysis.

Shortcomings of Principal Component Analysis.

Answer—

1) Relies on the correlation of the variables with each other, if the variables are not correlated to each other , PCA cannot be used effectively.

2) PCA needs the components to be perpendicular to one another. If its not the case then there's a problem.

3) PCA takes that the components which do not explain much variance are not useful , which cannot be the case in prediction problems.

Question 3--

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer→

With k-Means clustering, you need to have a sense ahead-of-time what your desired number of clusters is (this is the 'k' value). Also, k-means will often give unintuitive results if (a) your data is not well-separated into sphere-like clusters, (b) you pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low, or (c) you have weird initial values for your cluster centroids (one strategy is to run a bunch of k-means algorithms with random starting centroids and take some common clustering result as the final result).

In contrast, hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group. You end up with a 'dendrogram', or a sort of connectivity plot. You can use that plot to decide after the fact of how many clusters your data has, by cutting the dendrogram at different heights. Of course, if you need to pre-decide how many clusters you want (based on some sort of business need) you can do that too. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.