Question 1: Assignment Summary
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(what EDA you performed, which type of Clustering produced a better result and so on)
Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer 1:
Problem Statement:
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmes, they have 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.
Objective:
The requisite is:

- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

Method followed:
-Data Processing:

- It was found that there were no null values
- There were also no duplicate values for country
- There were a few outliers and they were treated later

-Clustering:

- Both the methods K means and Hierarchical Clustering was used
- For K means , K= 3 was taken using the elbow dip and silhouette analysis .
- While doing the Hopkins Statistics a value of 0.95 was attained.
- If the Hopkins Statistics values are:
- 0.3 : Low chase of clustering
- around 0.5 : Random
- 0.7 - 0.99 : High chance of clustering

-Finally using all these values clusters of 3 were formed and the countries are split into clusters.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

Answer 2 (a):

- K Means needs a prior knowledge of number of centroid (K) whereas hierarchical cluster do not need this kinds of parameters. cut_tree () function is used to create the number of clusters of any choice.
- In K Means clustering the algorithm will calculate the centroid each time.
- K Means is fast compare to hierarchical clustering
- Hierarchical clusters need more ram to run.

Answer 2 (b):

1. Initialize cluster centers
2. Assign observations to the closest cluster center
3. Revise cluster centers as mean of assigned observations
4. Repeat step 2 and step 3 until convergence

Answer 2 (c):
There is a method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.

Answer 2 (d):
Standardizing data is recommended because otherwise the range of values in each feature will act as a weight when determining how to cluster data, which is typically undesired.
For example consider the standard metric for most clustering algorithms (including DBSCAN in sci-kit learn) -- euclidean, otherwise known as the L2 norm. If one of your features has a range of values much larger than the others, clustering will be completely dominated by that one feature.

Answer 2 (e):
Single Linkage:
In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.
Complete Linkage:
In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.