**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)

Answer:

The data provided includes all the key indices of countries status measures. By carefully analysing the data we have to segregate the countries into different clusters basis on their performance. There is a high need to figure out which countries in the world are very much in need of support for reviving their current socio-economic status. There is need to provide aid to these nations by analysing their current economic status using their GDPP and INCOME statistics. By the help of mortality rate data and health related expenditure we can access these countries and plan the delivery of needed resources to make these countries perform better.

The solution methodology followed here is:

* Doing an exploratory data analysis to observe the distribution of data for all countries and correlation of indices and briefly checking similar performing countries.
* After the exploratory data analysis, Performed Hopkins statistics to check how much the given data is suitable for cluster analysis. Before doing the Hopkins statistics data has been standardized to bring all indices in same scale. Multiple iterations of Hopkins code have run to finalize the Hopkins statistics score.
* Once we found the data is suitable for clustering, we have tried multiple approaches of clustering algorithms. We have tried K-means clustering, Hierarchical clustering with single linkage and complete linkage.
* The decision of total number of clusters required made using different approaches like ‘elbow method’ and ‘silhouette score’. Based on these results the final cluster of under developed countries identified.
* K-means and Hierarchical clustering with complete linkage given a good result. By filtering the clusters, we found out the most aid needed countries.
* There are total of 11 countries found from the clustering analysis which are in need of help.

**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 A:

K-means clustering and Hierarchical clustering are one of those popular algorithms for clustering analysis. The comparison and contrast of these algorithms are:

* For huge data it is very complex and time taking for hierarchical clustering. K-means clustering performs well on huge data.
* Time complexity of k-means clustering is liner :O(n), for hierarchical clustering the complexity is O(n2). One of reason is: K-means clustering depends on just centroids and measures the distance with centroids. Whereas hierarchical clustering compares with each points time to check the distance.
* K-means clustering needs K value prior to run the model, whereas hierarchical clustering dose not need any pre definition of cluster.

Answer B:

The steps of K-means clustering are:

* By using certain methods like Hopkins statistics, we check this data for feasibility of cluster analysis.
* Based on domain knowledge, silhouette score, elbow method used to find out the number of clusters needed.
* Finalize K, Standardize/Normalise the data to bring all features in same scale.
* Run the k-means algorithm from any available packages (scikit learn). By tuning the parameters like number of iterations etc.

Answer C:

K value chosen by following approach:

* Business/Domain knowledge, like how many segments business wants to segregate its customers. How many segments would best describe the business need.
* Using some mathematical approaches like: Silhouette score, elbow method and figure out best number of cluster suitable for given data.

Answer D:

Scaling: Normalization or Standardization is must perform thing before performing clustering, As clustering is a distance calculation approach hence all the features need to be in same kind of scale to perform the calculation faster and form best clusters.

Answer E:

Single linkage and complete linkage are two algorithms of agglomerative Hierarchical clustering. Different linkage in Hierarchical Clustering are:

* Single linkage: Check distance with the most closed cluster and this process repeats for all points.
* Complete Linkage: Here the distance between the farthest points are taken as distance measure to find the clusters.