Question 1: Assignment Summary

Ans:- In the assignment we are having the problem statement to findout the most top 5 countries where low gdpp, low income and higher child mortality
So this clustering assignment has been performed by using two algo,
K-mean and Hierarchical clustering. And in the final step we got the conclusion of the top 5 countries where low gdpp, low income and high child mortality.

K-mean clustering the cluster 2 has the low gdpp, low income and high child-mortality

Hierarchical clustering the cluster 0 has the low gdpp, low income and high child-mortality

Question 2: Clustering

- a) Compare and contrast K-means Clustering and Hierarchical Clustering.

S.N	K-Mean	Hierarchical
1	k-mean clustering can handle the big amount of data	Hierarchical clustering can't handle big data
2	This is because the time complexity of K Means is linear O(n)	hierarchical clustering time complexity is quadratic O(n2)
3	In K Means clustering, we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.	Results are reproducible in Hierarchical clustering.
4	K Means is found to work well when the shape of the clusters is hyper spherical	
5	K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into.	But you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram
6	K-Mean we can find out the k using Silhouette score and Elbow curve-ssd	Hierarchical clustering initial labels all data points are individual clusters then based on distance we merge them into one by one and make a tree.

- b) Briefly explain the steps of the K-means clustering algorithm.
    - To start with K-mean first find out the analysis columns
    - Perform the EDA analysis
    - Treat the outliers based on EDA box plot visualization
    - Scale the variables
    - Perform the Silhouette score and Elbow curve-ssd to find out the k value
    - Perform the K-mean algo to assign the clutter
    - Merge the data into cluster and check the cluster visualization
    - Find the conclusion

- c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
    - Based on Silhouette score and Elbow curve-ssd we find the value of k in the K-mean clustering algorithm.
    - To choose the best number of K, we see the visualization view of it and check how frequent the graph dropping to the point by point and check out business requirement and decide the k value

- d) Explain the necessity for scaling/standardisation before performing Clustering.
    - Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.
    -
- e) Explain the different linkages used in Hierarchical Clustering.

    - **Single-Linkage**: Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

    - **Complete-Linkage**: Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

- **Average-Linkage:** Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

- **Centroid-Linkage**: Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn't arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.