Clustering Assignment

Problem Statement :

We are provided with a dataset of 167 countries. We have to categorize them on the basis of socio-economic factors. And finally identify 5 countries which are in dire need of aid.

We perform below steps to achieve our goal :

1. Data Quality Check

We update the imports, exports and health according to the gdp of the country by multiplying it with gpd and then divide by 100.

2. Univariate and Bivariate

We can see that there is high correlation (both positive and negative) between some columns. The countries where income, health, imports, exports, life_expectancy, gdp are high, there child_mort, inflation and total_fer are less, and vice-versa.

3. Outlier

For columns child_mort, inflation, total_fer,we have to cap the lower range outliers.

For the rest of the cols, we have to deal with the upper range outliers.

Because our focus is on countries with high child_mort, inflation, total_fer. And with less income, health, imports, exports, life_expectancy, gdp

4. Scaling

We perform standardized scaling to scale all the columns before performing clustering.

5. Hopkin's Test

We perform Hopkin's test to check how different our data is from the randomly scattered data. The value we get is approx 0.95 which shows that data can be easily divided in clusters

6. Find the best value of k using SSD Elbow, Silhoutte Score

For performing the k-means clustering we should know the number of clusters in advance. To check optimum no of clusters we performe two tests :

I. SSD elbow

II. Silhoutte Score

Then we plot both of them and find that the number of clusters equal to 3 seems ncie solution.

7. Using the final value of k=3, perform the kMeans analysis

Then we run the k-means model on the dataset provided and assign the labels to output.

We found 3 dataset here. Our requirement is the cluster with low gdp, low income and high child_mort.

We found such cluster with no countries equal to 48.

We plot the results using scatter plot and box plots.

We need to find the countries which are in dire need of aid.

So, we sort them in the descending order of importance.

```
c1.sort_values(by=['gdpp','income','child_mort'], ascending= [True, True, False]).head(10)
```

And get the top 5 counties as our output.

A. Burundi
B. Liberia
C. Congo, Dem. Rep.
D. Niger
E. Sierra Leone

10. Hierarchical Clustering: Single linkage, Complete Linkage

Then we perform the Hierarchical clustering. In Hierarchical clustering, there is no need to identify no of clusters in advance like in k-means.

First we performed a single linkage, but didn't get any output.

Then we move on to complete linkage. We get some results.

We chose no of clusters equal to 4 and cut the tree accordingly.

Our requirement is the cluster with low gdp, low income and high child_mort.

We found such cluster with no countries equal to 34.

There is one cluster with only one country, this is the reason why select number of clusters as 4 in Hierarchical clustering, but no of clusters equal to 3 in k-means clustering.

We plot the results using box plots.

We need to find the countries which are in dire need of aid.

So, we sort them in the descending order of importance.

```
c1.sort_values(by=['gdpp','income','child_mort'], ascending= [True, True, False]).head(10)
```

And get the top 5 counties as our output.

A. Burundi
B. Liberia
C. Congo, Dem. Rep.
D. Niger
E. Sierra Leone