Assignment 4 - Clustering

Instructions:

For the Fourth Assignment, we will be implementing the clustering algorithms: k-Means and Agglomerative/Hierarchical clustering.

Task 1: k-Means

Part 1A: 5 points

- Check if there are any missing values in the provided dataset.
- If missing values are found, fill them using appropriate methods.
- Explain your decision for the chosen action.

Part 1B: 10 points

- Drop the necessary columns that will not be used for clustering.
- Scale the data for further processing.
- Display the processed data.
- Explain your actions.

Part 1C: 15 points

- Use the elbow method to determine the optimal number of clusters for k-means clustering.
- Perform the elbow method for k values ranging from 2 to 12.
- Plot a line chart of the SSE (Sum of Squared Errors) for each value of k.
- Determine the best k (number of clusters) from the graph.
- Interpret the results and explain which k value is most appropriate.

Part 1D: 10 points

- Use the Silhouette score to select the most appropriate value for k.
- Apply Silhouette analysis on the same data for k values ranging from 2 to 12.
- Display the Silhouette score for each k value.
- Explain which k value you selected from the results and why.

Part 1E: 10 points

- Use the best k value obtained from the previous step and apply the k-means algorithm on the data.
- Print out the entire data along with the cluster labels assigned to each row as a new column.

Part 1F: 10 points

- For each formed cluster, display the average value for each column from the dataset.
- Interpret the results and explain the findings.

Part 1G: 5 points

- Display the top 5 countries within each group/cluster.
- If there aren't enough countries within a cluster, display as many as possible.

Part 1H: 5 points

- Plot a scatter plot of "Health" vs "Life Expectancy".
- Shade each point belonging to its respective cluster.

Task 2: Hierarchical Clustering

Part 2A: 5 points

- Plot the co-ordinates from the imported data.
- Guess possible clusters that could be formed based on the plot.
- Explain your guess.

Part 2B: 15 points

- Perform hierarchical clustering with two different distance metrics: Euclidean + Ward and Manhattan + Average.
- Display the clustering results for cluster sizes 4, 5, 6, and 7.
- Provide a title for each plot.
- Interpret the results and suggest the appropriate value of k and the preferred pair of affinity and linkage.
- Explain your selection and compare it with your previous guess.

Part 2C: 10 points

- For your selected value of k, display the dendrogram for the analysis.
- Plot a line where k clusters are formed.
- Provide a proper title for the plot.

Programming Assignment Details:

- You can use the following libraries: NumPy, Pandas, scikit-learn, Matplotlib, and Seaborn.
- Explain each task performed in the code using Markdown cells.
- Comment your code appropriately.

- If using any external resources (books, internet), cite them within the cell.
- Do not rename the dataset files.

Submission Details:

- Fill in the name and ID of each group member in the Jupyter notebook in the provided format.
- Name your submission files using the following format: yourLastName_Last4digitsofyourID.ipynb (e.g., smith_1234_assignment4.ipynb).
- Only one team member should submit