Homework Assignment: Understanding Environmental Justice Indexes with Python

Part 1: Data Loading and Initial Exploration

1. **Loading Data**:
   - Write Python code to load a CSV file named `United States.csv` into a Pandas DataFrame and display the first five rows. Explain why it is important to inspect the first few rows of a DataFrame when loading a new dataset.

Part 2: Working with SPL Columns

2. **Filtering SPL Columns**:
   - Write code to filter columns that start with "SPL_" and create a new DataFrame containing only these columns. Display the first five rows. What do the `SPL_` columns represent in the context of environmental justice?
3. **Adding Additional Columns**:
   - Add specific columns (`'STATEFP'`, `'COUNTYFP'`, etc.) to the DataFrame created in the previous step and display the first five rows. What is the purpose of including these columns alongside the `SPL_` columns?

Part 3: Data Cleaning

4. **Checking for Missing Data**:
   - Write code to identify columns with missing values in the `SPL_` DataFrame and display the counts of missing values. Why is it important to check for and address missing data in your analysis?
5. **Removing Missing Data**:
   - Remove rows with missing values from the DataFrame. Print the number of rows before and after cleaning. What impact does removing rows with missing data have on the analysis?

Part 4: Descriptive Statistics

6. **Summary Statistics**:
    - Write Python code to calculate and display the mean, standard deviation, skewness, and kurtosis for each `SPL_` column. How can skewness and kurtosis help you understand the distribution of a dataset?

Part 5: Visualization

7. **Histograms**:
    - Create histograms for all columns that start with "SPL_" in the cleaned DataFrame. Explain what insights you can gain from viewing the histograms of these columns.
8. **Elbow Method for Clustering**:
    - Write code to perform the Elbow Method for determining the optimal number of clusters for K-Means clustering on the `SPL_` columns. Include a plot showing the sum of squared distances for different numbers of clusters (1 to 20). What does the "elbow" in the plot represent?

Part 6: Clustering and PCA

9. **K-Means Clustering**:
    - Run K-Means clustering with an optimal number of clusters (e.g., determined from the elbow plot) on the `SPL_` columns. Print the first five rows of the DataFrame with the assigned cluster labels. What is the significance of clustering in analyzing environmental justice data?
10. **Hierarchical Clustering**:
    - Using the training data, randomly select 5,000 data points and perform divisive (top-down) hierarchical clustering. Plot the dendrogram and decide on the number of clusters. How does hierarchical clustering differ from K-Means in terms of methodology and results?
11. **Principal Component Analysis (PCA)**:
    - Write code to perform PCA on the `SPL_` columns, keeping only the first two principal components. Plot the data points using these two components and mark the centroids for each cluster. How does PCA help in visualizing high-dimensional data?

Part 7: Simulation and Validation

12. **Simulating Clusters**:

    - Write a function to simulate clusters by randomly assigning data points to clusters and calculating the Within-Cluster Sum of Squares (WCSS). Run this function for 10,000 iterations and plot histograms of the WCSS. Why might this simulation be important for validating clustering results?

13. **Hold-Out Data Validation**:

    - Split the original dataset into training and hold-out sets (85% training, 15% hold-out). Run K-Means clustering on the training data and predict the clusters for the hold-out data. Print the counts of each cluster in the hold-out data. Why is it important to test clustering results on hold-out data?