ASSIGNMENT DATA ENGINEER INTERN - MRIKAL SOLUTIONS

Objective:

The objective of this assignment is to assess your ability to perform data manipulation, cleaning, transformation, and analysis using Pandas in Python. You will be working with a dataset to derive insights and perform various tasks.

Dataset:

You can use any publicly available dataset, but for this assignment, let's use the "Iris" dataset. You can download the dataset from here or use the one provided by the `seaborn` library.

Tasks:

1. Load the Dataset:
   - Load the dataset into a Pandas DataFrame.
   - Display the first 5 rows of the DataFrame.
   - Display the summary statistics of the dataset.
2. Data Cleaning:
   - Check for missing values and handle them appropriately.
   - Ensure that all the column names are in a consistent format (e.g., all lowercase).
3. Data Transformation:
   - Add a new column `sepal_area` which is the product of `sepal_length` and `sepal_width`.
   - Add a new column `petal_area` which is the product of `petal_length` and `petal_width`.
   - Normalize the values in `sepal_length`, `sepal_width`, `petal_length`, and `petal_width`.
4. Data Aggregation and Grouping:
   - Group the dataset by `species` and calculate the mean, median, and standard deviation for `sepal_length`, `sepal_width`, `petal_length`, and `petal_width`.
   - Create a pivot table that shows the mean `sepal_area` and `petal_area` for each species.
5. Data Visualization:
   - Plot the distribution of `sepal_length` for each species.
   - Create a scatter plot of `sepal_length` vs. `sepal_width` colored by species.
   - Create a heatmap to visualize the correlation matrix of the dataset.
6. Advanced Data Analysis:
   - Identify the top 10 rows with the highest `sepal_area`.

- Filter the dataset to only include rows where `petal_length` is greater than the median `petal_length` of the dataset.
- Perform any additional analysis you find interesting or relevant.

Deliverables:

1. A Jupyter Notebook containing your code and results.
2. Explanatory markdown cells that describe what you are doing at each step and the insights you derive from the data.
3. The original dataset in CSV format that you fetched and loaded into your DataFrame.