Lab: Data Cleaning, Transformation, and Aggregation on the Iris Dataset

Scenario Overview:

You are analyzing the Iris dataset, which contains information on the physical characteristics of three species of iris flowers. Your goal is to clean the dataset, perform transformations, handle missing data, and apply aggregation techniques to summarize the data. Additionally, you'll reshape the data to prepare it for deeper analysis.

Tasks:

Task 1: Identifying and Imputing Missing Data

- **Locate Missing Data:** Examine the dataset to locate any missing values. Identify the columns with missing data and report how many missing values are present in each column.
- **Handle Missing Data in Numerical Columns:** Fill in missing values for numeric columns (i.e., `sepal_length`, `sepal_width`, `petal_length`, `petal_width`) using the median value of each column. Justify why you chose this approach.
- **Handle Missing Data in Categorical Columns:** Identify if there are missing values in the **species** column. If so, impute them with the most frequent value (mode) in the column.

Task 2: Data Integrity and Transformation

- **Remove Duplicate Records:** Review the dataset for duplicate rows (where all values in a row are identical) and remove any duplicates found. Ensure that only one unique record per flower remains in the dataset.
- **Feature Engineering:** Create a new feature called **total_area** by adding the areas of both the sepal and petal. To do this, create separate columns for the sepal area and petal area and then add them to form the `total_area` column.
- **Handling Missing Values Again:** After imputing missing data, inspect the dataset again and drop any rows that still have missing values in any of the columns.

Task 3: Aggregation and Data Transformation

- **Numerical Conversion of Categorical Data:** Convert the **species** column, which is categorical, into a numerical format by assigning each species a unique number (e.g., 0, 1, 2).
- **Apply Grouped Aggregation:** Using the transformed data, group the flowers by species and calculate the total sum of the numeric columns (`sepal_length`, `sepal_width`, `petal_length`, `petal_width`). Present the results in a table that shows the sum of each feature per species.

Task 4: Data Reshaping

- **Reshape the Dataset into a Long Format:** Reshape the dataset so that each flower's attributes (sepal length, sepal width, etc.) are stacked in a single column, with a new column indicating the attribute type (e.g., `sepal_length`, `sepal_width`, etc.).
