ML Assignment 1: Iris Dataset Preprocessing and Missing Value Handling

Objective:

- To familiarize yourself with basic data preprocessing techniques in Python.
- To learn how to identify and handle missing values in a dataset.
- To apply these techniques to the well-known Iris dataset.

Dataset:

- You will be working with the Iris dataset, which contains measurements of sepal length, sepal width, petal length, and petal width for three species of iris flowers (setosa, versicolor,virginica).
- Dataset Link: https://drive.google.com/file/d/1tuqcDirRDgNdaeThuvUcM1QeX2jCYCzg/view?usp=sharing

Tasks:

1. **Load the Dataset:**
   - Use Python libraries (e.g., `pandas`, `scikit-learn`, `seaborn`) to load the Iris dataset into a pandas DataFrame.
   - Ensure you can display the first few rows of the DataFrame to verify successful loading.
2. **Initial Data Exploration:**
   - Use `pandas` functions (e.g., `info()`, `describe()`, `head()`, `tail()`) to understand the structure and basic statistics of the dataset.
   - Identify the data types of each column.
   - Determine the number of rows and columns.
3. **Missing Value Detection:**
   - Check for missing values in the dataset.
   - Use `pandas` methods like `isnull().sum()` or `isna().sum()` to identify the number of missing values in each column.
   - If missing values are present, document which columns contain them and the extent of the missing data.
4. **Missing Value Handling (If Applicable):**
   - If missing values are found, implement appropriate strategies for handling them. Consider and implement at least two of the following strategies:
     - **Deletion:** Remove rows or columns containing missing values. Be mindful of the potential loss of information.
     - **Imputation:** Fill in missing values with estimated values. Common imputation methods include:
       - Mean imputation: Replace missing values with the mean of the column.
       - Median imputation: Replace missing values with the median of the column.
       - Mode imputation: Replace missing values with the mode of the column.

   - Justify your choice of imputation method.
5. **Verification:**
   - After handling missing values, verify that there are no remaining missing values in the dataset.
   - Display the dataframe again to verify that the changes have been made.
6. **Basic Statistical Analysis:**
   - After the missing values are handled, perform a basic statistical analysis on the processed dataset.
   - Calculate the mean, median, standard deviation, and range for each numerical feature.
   - Use boxplots or histograms to visually represent the distribution of each feature.
7. **Identify the features and the dependent variable**
8. **Create the matrix of features (X) and the dependent variable vector (y)**
9. **Print X and y**
10. **Documentation:**
    - Provide clear and concise comments within your code, explaining each step of the process.
    - Include a brief written summary of your findings, including:
      - The number of missing values found (if any).
      - The methods used to handle missing values.
      - Any observations or insights gained from the statistical analysis.

Libraries to Use:

- `pandas`
- `scikit-learn` (for loading the dataset)
- `seaborn` (for loading the dataset, and for visualization)
- `matplotlib` (for visualization)
- `numpy`

Submission:

- Submit a Jupyter Notebook (.ipynb) containing your code and documentation,saved by your name example: (Vishvesh_ML_Assignment_1.ipynb) on the link below:

  https://drive.google.com/file/d/1tuqcDirRDgNdaeThuvUcM1QeX2jCYCzg/view?usp=sharing