Section 2: Data Mining – Task 1: Data Preprocessing and Exploration (15 Marks)

1. Dataset Loading

The Iris dataset was loaded using `sklearn.datasets.load_iris()`.
It contains 150 samples with four numerical features:

- Sepal length
- Sepal width
- Petal length
- Petal width
- Species (class label)

A pandas DataFrame was created for easier manipulation and exploration.

2. Preprocessing Steps

- **Missing Values:** Checked using `df.isnull().sum()`.

  No missing values were found, but the script demonstrates how they would be handled if present.

- **Normalization:** Applied **Min-Max Scaling** to all four numerical features to bring values into the 0–1 range.

- **Encoding:** The species label was encoded using **One-Hot Encoding** to prepare the dataset for models requiring numerical inputs.

3. Exploratory Data Analysis (EDA)

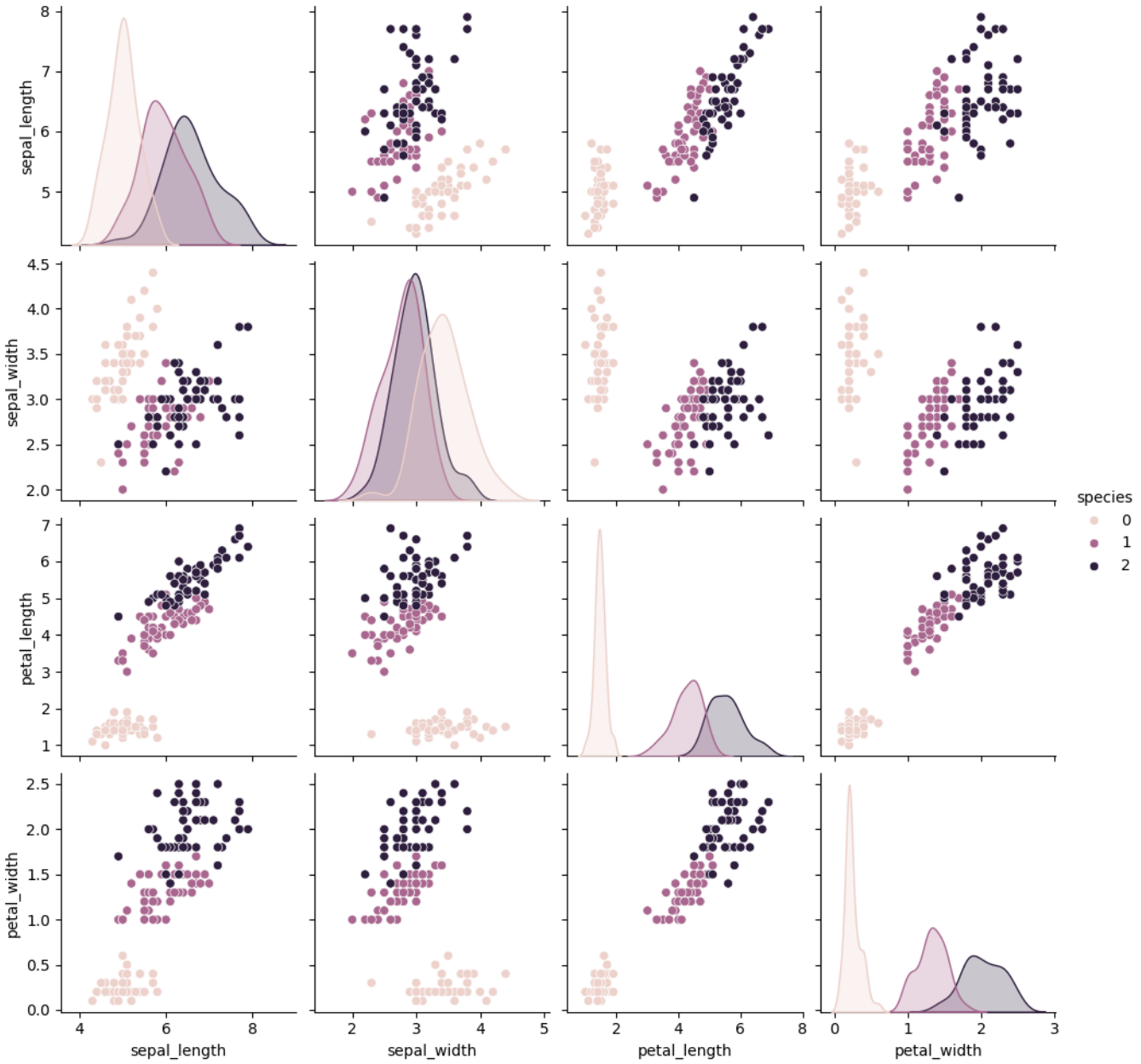
- **Summary Statistics:** Computed using `df.describe()` to observe means, standard deviations, and feature ranges.

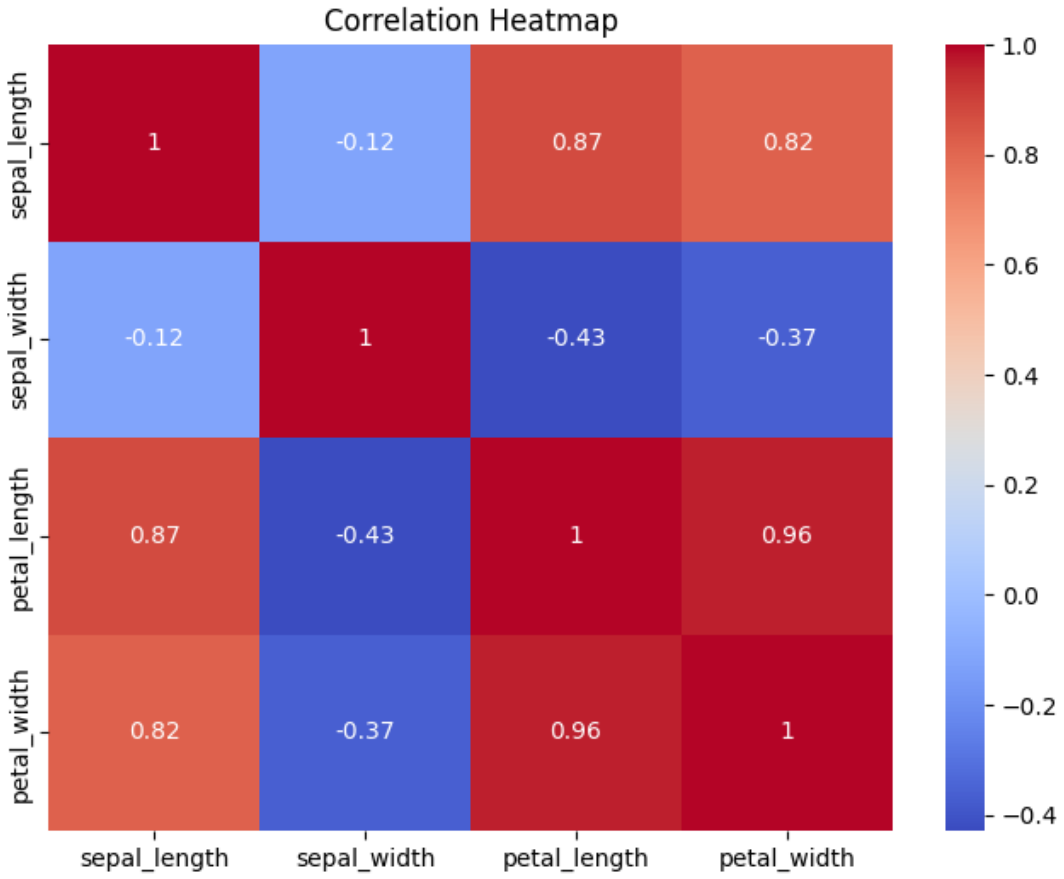
- **Pairplot:** A seaborn pairplot was generated to visualize relationships between features and species separation.

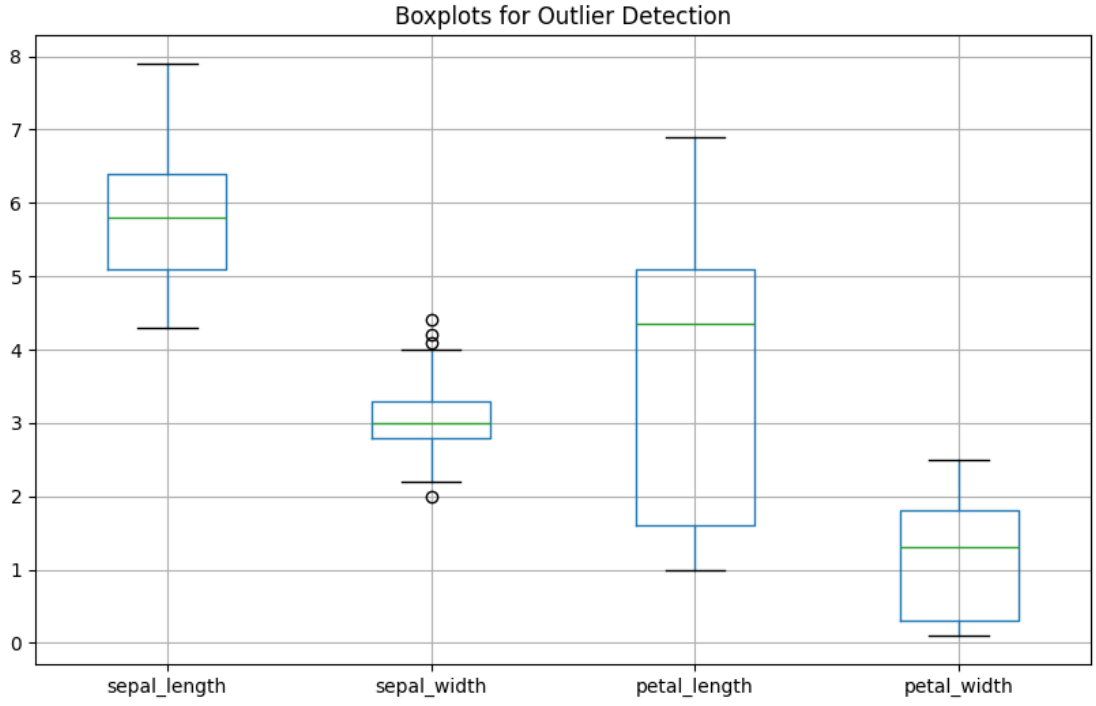
- **Correlation Heatmap:** A heatmap was created to show correlations between numerical features.

- **Outlier Detection:** Boxplots were used to identify potential outliers across all four features.

- **Visualizations**


sepal_length
sepal_width
petal_length
petal_width
species
0
1
2





4. Train/Test Split Function

A custom function `split_data()` was implemented to split the dataset into:

- **80% training data**
- **20% testing data**

This ensures reproducibility and prepares the dataset for clustering and classification tasks in later sections.