## EXPERIMENT NO. 6

**Title:**

Identifying and Visualizing Relationships between Variables in the Iris Dataset.

**Objective:**

To identify and visualize relationships between different variables in the Iris dataset using various types of plots such as scatter plots, pair plots, and correlation matrices.

**Brief Theory:**

**Introduction:**

Understanding relationships between variables is a key aspect of exploratory data analysis (EDA). Visualizing these relationships helps in uncovering patterns, correlations, and dependencies that may not be obvious from raw data. This experiment involves using scatter plots, pair plots, and correlation matrices to analyze the Iris dataset.

**Techniques**

- **Scatter Plot:** A plot that shows the relationship between two variables by displaying data points on a two-dimensional plane.
- **Pair Plot:** A grid of scatter plots for all pairs of variables in the dataset, providing a comprehensive view of relationships.
- **Correlation Matrix:** A table showing the correlation coefficients between pairs of variables, which quantifies the strength and direction of their relationships.

**Steps**

- **Step 1: Import Necessary Libraries**

  First, import the libraries required for data manipulation and visualization.

  Code: -

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
```

- **Step 2: Load the Iris Dataset**

Load the Iris dataset using the load_iris function from scikit-learn and convert it into a pandas DataFrame.

Code: -

```
# Load the iris dataset
iris_data = load_iris()
iris = pd.DataFrame(data=iris_data.data, columns=iris_data.feature_names)
iris['species'] = iris_data.target
```

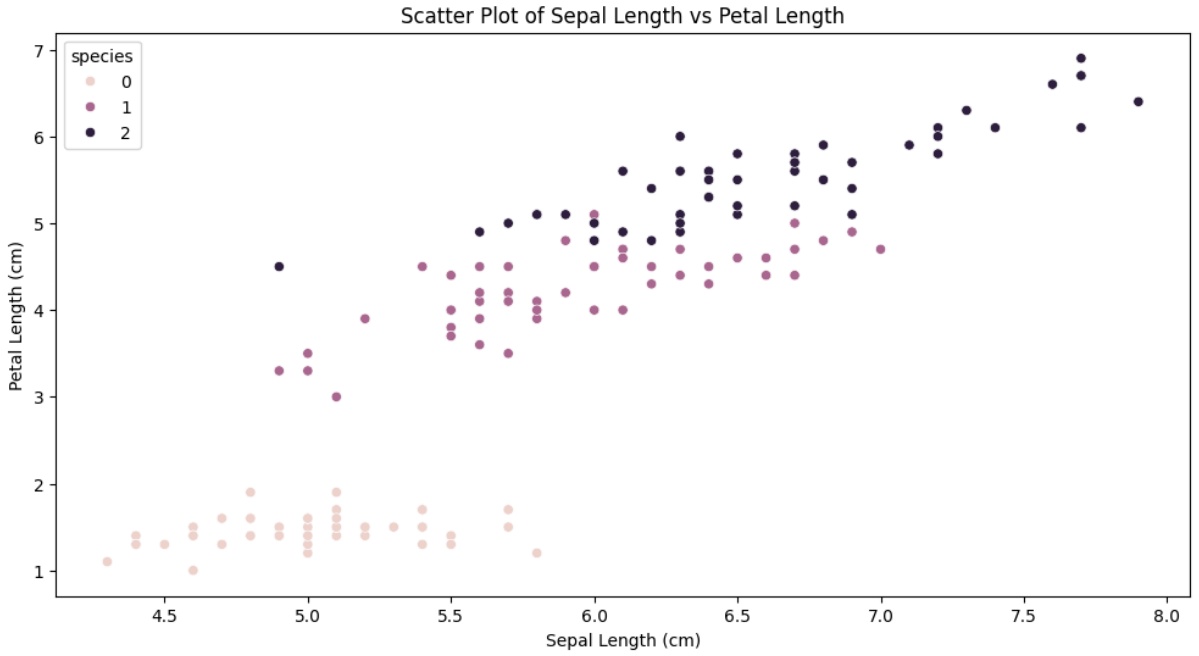
***The load_iris() function loads the dataset, which is then converted to a DataFrame with appropriate column names.***

- **Step 3: Scatter Plot**

Create scatter plots to visualize the relationship between different pairs of variables.

Code: -

```
plt.figure(figsize=(12, 6))
sns.scatterplot(x='sepal length (cm)', y='petal length (cm)', hue='species', data=iris)
plt.title('Scatter Plot of Sepal Length vs Petal Length')
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Petal Length (cm)')
plt.show()
```



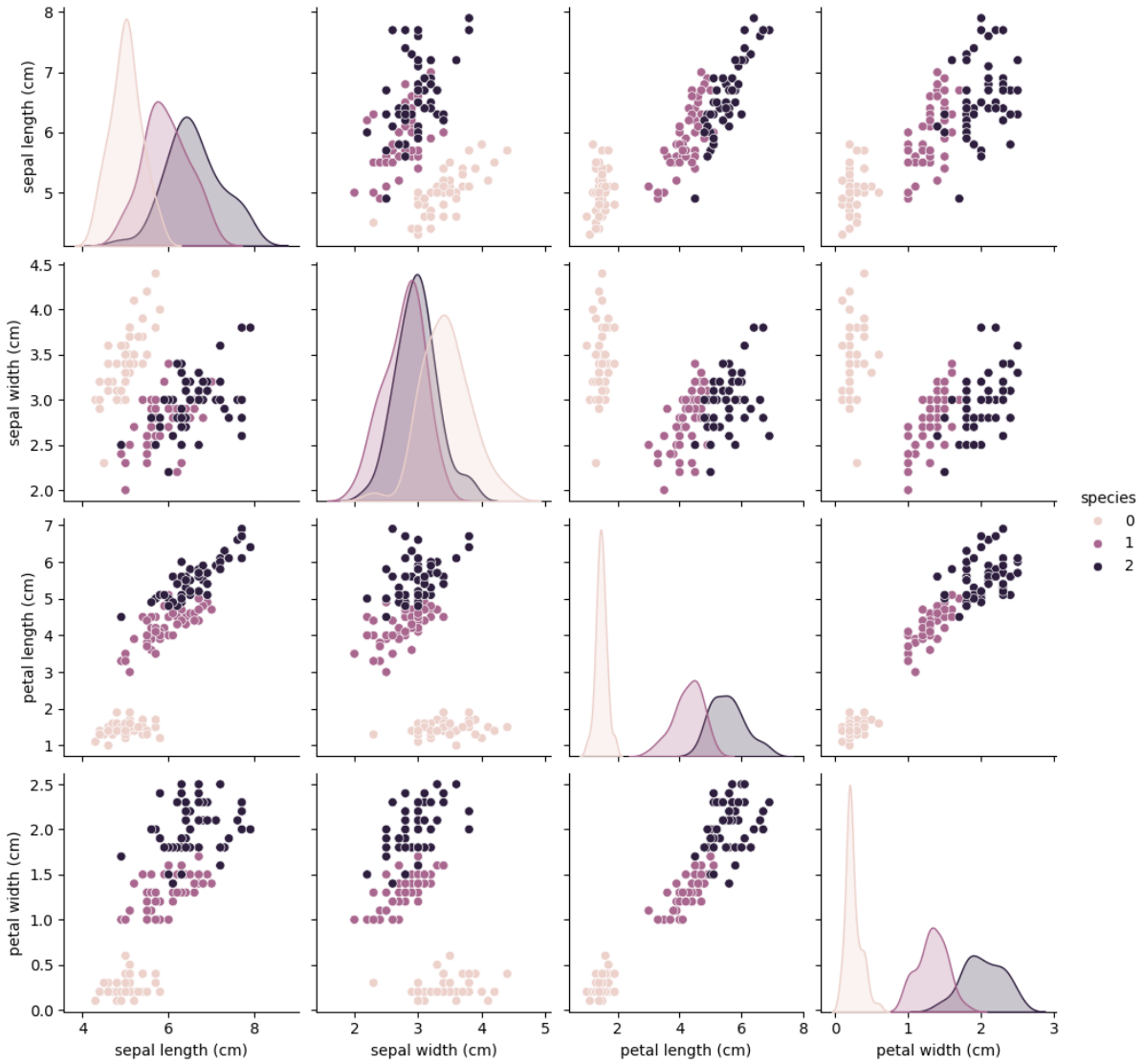

***A scatter plot is created to visualize the relationship between sepal length and petal length, with points color-coded by species.***

- **Step 4: Pair Plot**

  Create a pair plot to visualize the relationships between all pairs of variables in the dataset.

  Code: -

```
sns.pairplot(iris, hue='species')
plt.show()
```




***The pair plot provides a grid of scatter plots for all pairs of variables, giving a comprehensive view of the relationships between different features.***

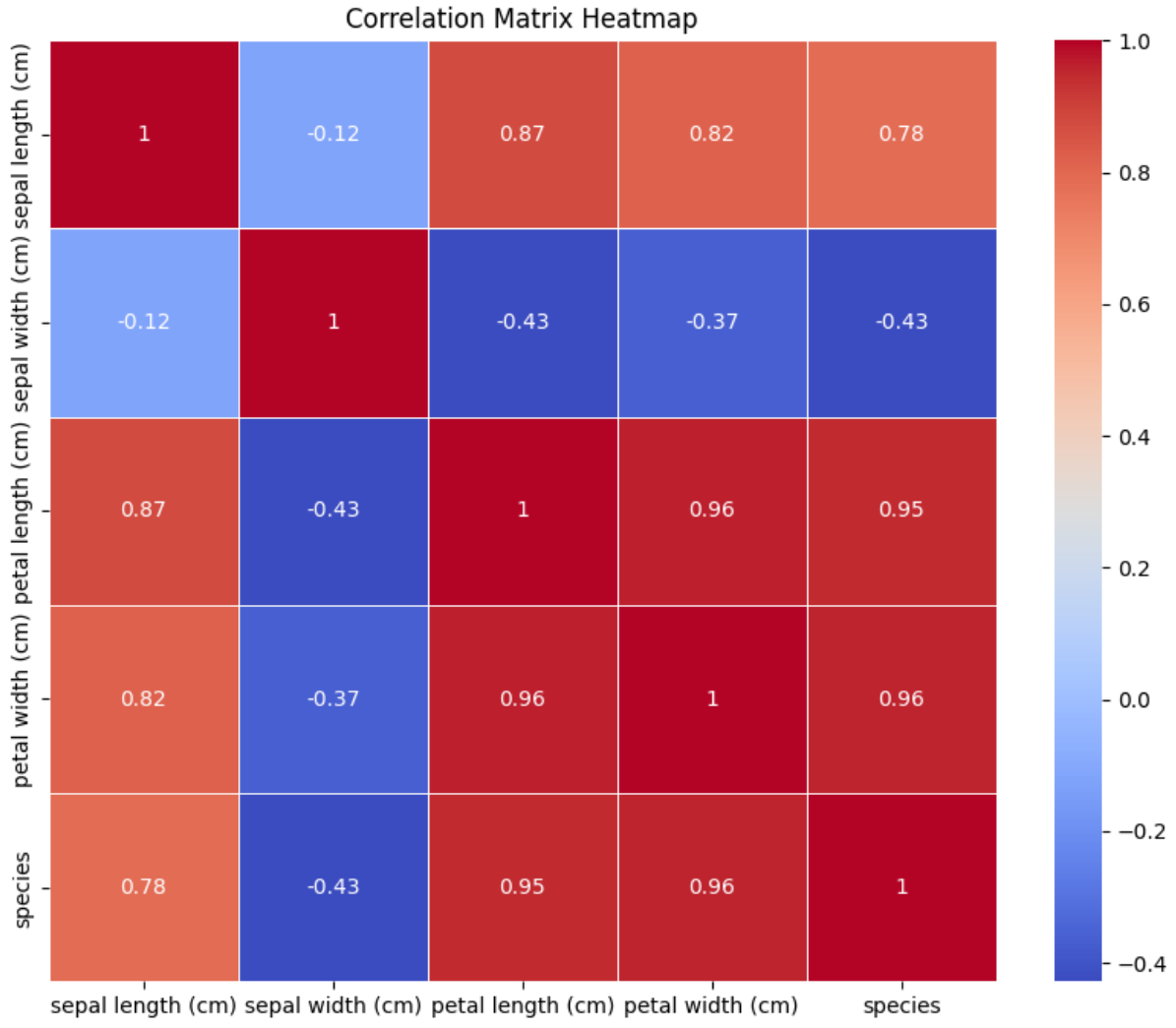
- **Step 5: Correlation Matrix**

  Calculate the correlation matrix and create a heatmap to visualize it.

  Code: -

```
# Calculate the correlation matrix
corr_matrix = iris.corr()

# Create a heatmap of the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix Heatmap')
plt.show()
```




***The correlation matrix is calculated using the corr() method, and a heatmap is created to visualize the correlation coefficients between pairs of variables.***

**Conclusion:**

In this experiment, we identified and visualized relationships between different variables in the Iris dataset. Scatter plots provided a view of individual relationships, pair plots offered a comprehensive view of all variable pairs, and the correlation matrix quantified the strength and direction of relationships. These visualizations and calculations are essential for understanding the interdependencies between variables in the dataset.

**Practice Questions:**

1. Perform an initial exploration of the Iris dataset. Summarize the dataset by identifying the number of observations, features, and classes. What are the basic descriptive statistics (mean, median, min, max) for each feature?
2. Use a pairplot to visualize the relationships between all features of the Iris dataset. How do the features **`sepal length`**, **`sepal width`**, **`petal length`**, and **`petal width`** correlate with each other across different species?
3. Calculate the correlation matrix for the features in the Iris dataset. Which pair of features shows the strongest positive correlation? Provide a heatmap to support your analysis.
4. Create a scatterplot for **`petal length`** **versus** **`petal width`** and differentiate the points by species using colors. What trends do you observe between these two features across different species?
5. Plot the distribution of **`sepal length`** for each species using boxplots. How does the distribution vary between the three species?

**Expected Oral Questions**

1. What is the Iris dataset, and why is it widely used in machine learning and data visualization?
2. Can you describe the four main features of the Iris dataset? How do they relate to the classification of different species?
3. What are the data types of the features in the Iris dataset? Why is this important for analysis?
4. How many classes (species) are present in the Iris dataset, and what are their respective distributions?
5. Which visualization techniques are most suitable for exploring relationships between variables in the Iris dataset? Why?
6. What is the significance of correlation in this experiment? How can you interpret a high positive or negative correlation?
7. When you generate a pairplot for the Iris dataset, what key insights can you draw from it regarding the separability of species?
8. How can the techniques used in this experiment be applied to other real-world datasets? Provide an example.