Project 2: Data Exploration

Introduction:

This project explores the Iris dataset. It includes loading the dataset, Explore its structure, checking for missing values, and visualizing feature distributions using plots and graphs.

Dataset Overview:

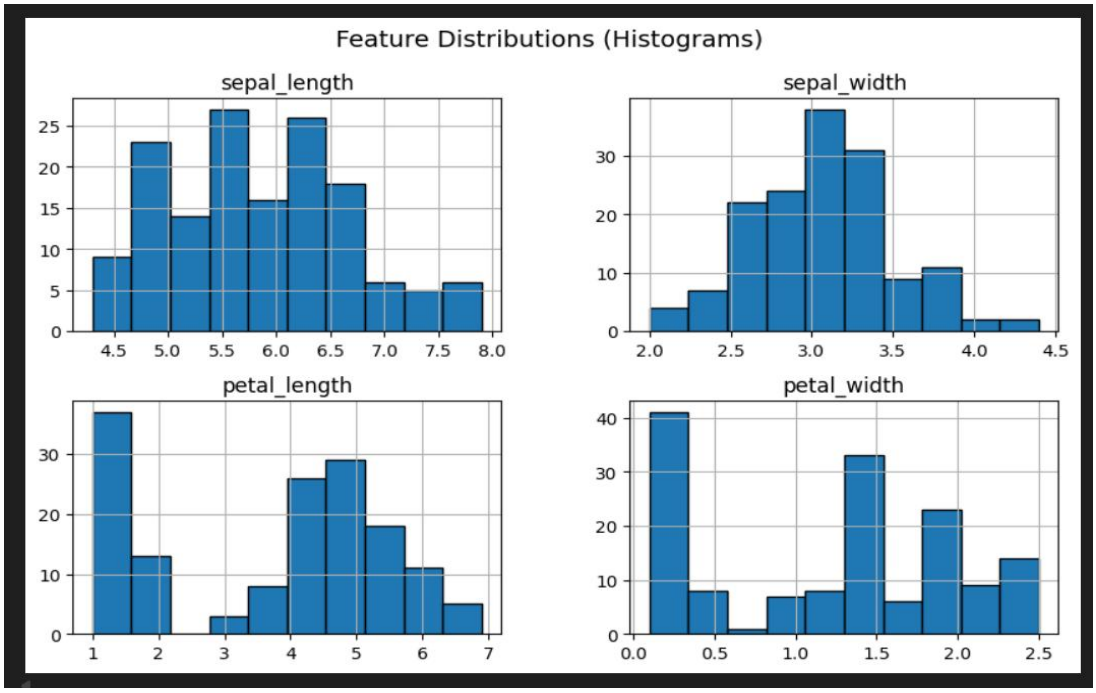
First 5 rows of the dataset:

```
First 5 Rows:
   sepal_length  sepal_width  petal_length  petal_width species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa
```

Missing Values:

sepal_length	0
sepal_width	0
petal_length	0
petal_width	0
species	0
dtype	int64

Feature Distributions:



1. **Sepal Length:**

-Values are spread roughly between **4.3 cm and 7.9 cm**.

-Distribution appears fairly uniform with slight peaks around **5.0–6.5 cm.**

-No extreme skewness, indicating balanced sepal length values across species.

2. **Sepal Width:**

-Values range from **2.0 cm to 4.4 cm**.

-Distribution is slightly left-skewed, with most flowers having sepal widths between **2.8 cm and 3.4 cm**.

-Few samples exist with very low (near 2.0) or very high (above 4.0) sepal widths.

3. Petal Length:

-Clear bimodal distribution.

-One group around **1.0–2.0 cm** (Iris-setosa).

-Another group spread between **3.0–7.0 cm** (Iris-versicolor and Iris-virginica).

-Indicates petal length is a key distinguishing feature among species.

4. **Petal Width:**

-Shows bimodal distribution.

-A cluster near **0.1–0.6 cm** (setosa).

-Another spread from **1.0–2.5 cm** (versicolor & virginica).

-Similar to petal length, petal width is **highly discriminative** for species classification.

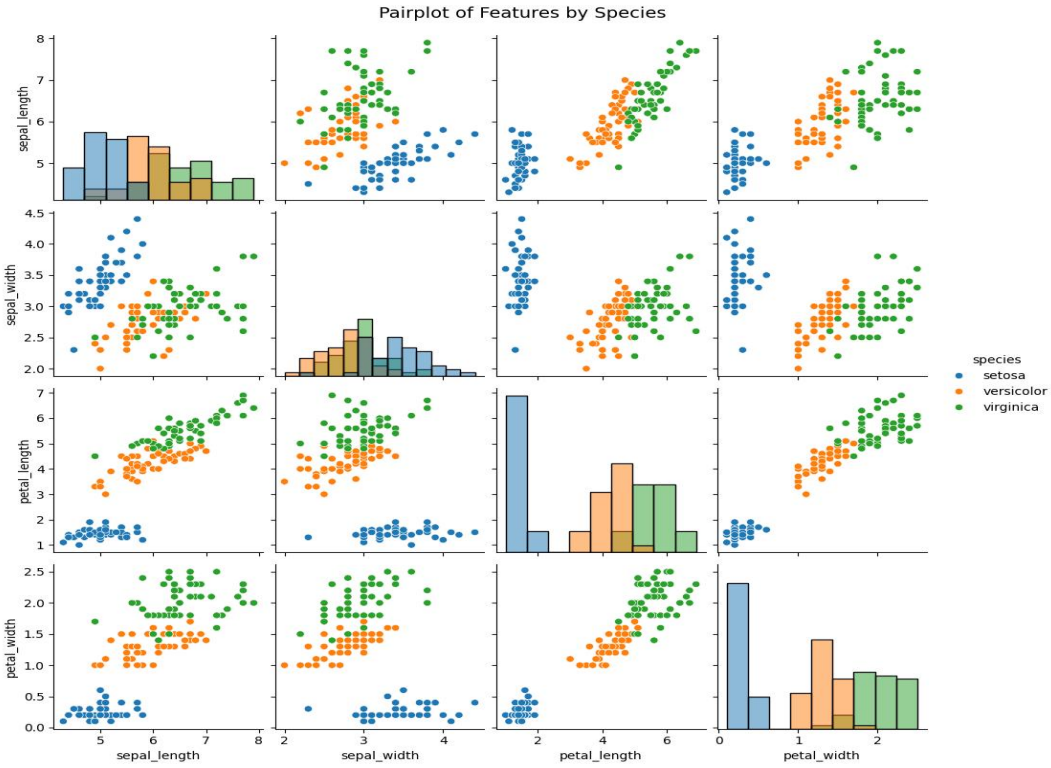
5. **Overall Observation:**

-Sepal features (length, width) show moderate overlap across species.

-Petal features (length, width) provide **clear separation** between Iris-setosa and the other two species.

-This highlights the importance of petal dimensions in species identification within the Iris dataset.

Pairplot of Features by Species:



- **Setosa** is clearly separable from other species using petal length and width.

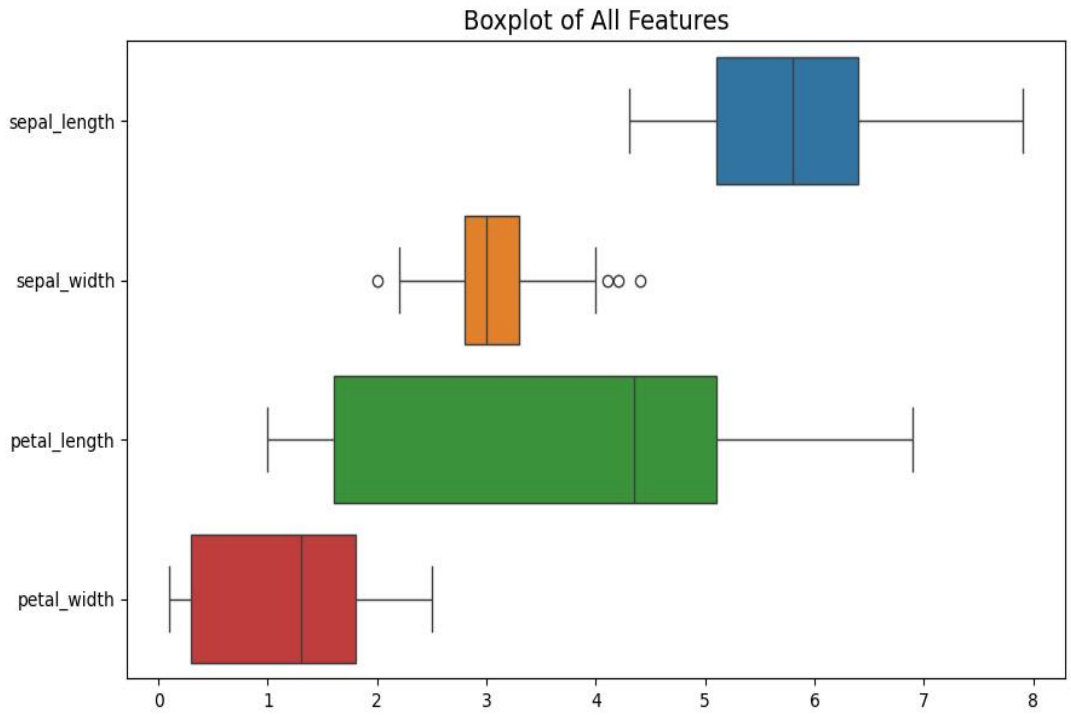
- **Versicolor and Virginica** overlap but can be distinguished with petal features.

- **Sepal features** (length, width) show high overlap and are less effective for separation.

- **Petal length vs petal width** provides the best species separation.

- **Strong correlation** exists between petal length and petal width.

Boxplots:



- Sepal Length: Median around 5.8; wide spread (≈4.3 to 7.9) with no major outliers.

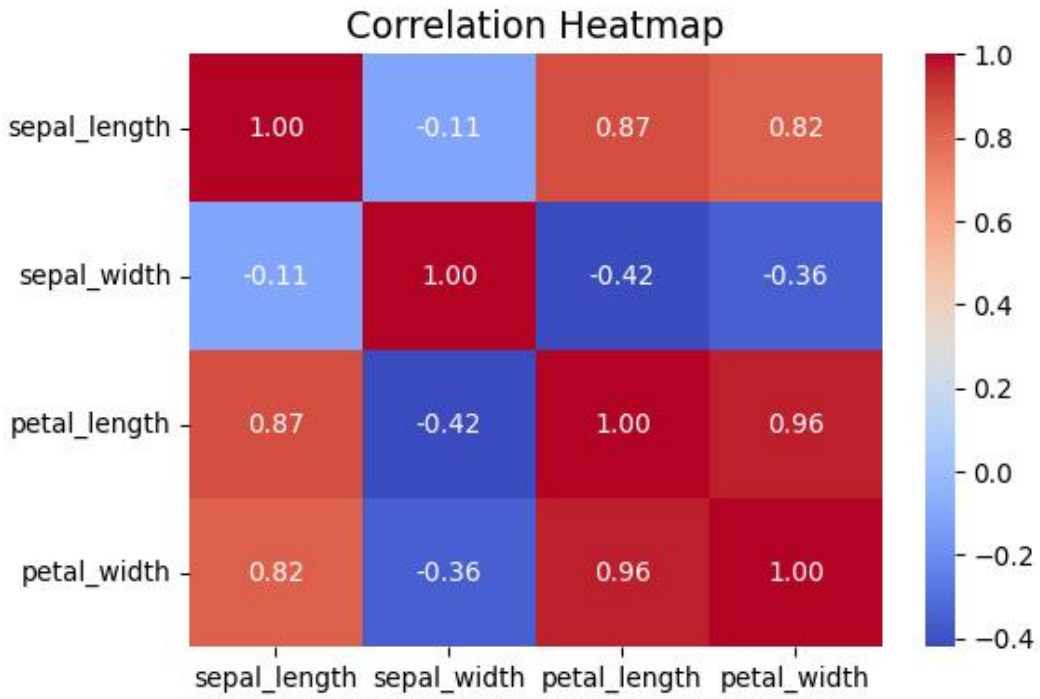
- Sepal Width: Median ≈3.0; smaller spread (≈2.0 to 4.4) with some outliers present.

- Petal Length: Median ≈4.3; widest range (≈1.0 to 6.9), showing high variability.

- Petal Width: Median ≈1.3; range ≈0.1 to 2.5, moderate spread with no extreme outliers.

-Overall: Petal features (length, width) show higher variability than sepal features, and sepal width contains noticeable outliers.

Correlation Heatmap:



1. **Sepal Length:**

-Strong positive correlation with **Petal Length (0.87)** and **Petal Width (0.82)**.

-Very weak negative correlation with **Sepal Width (-0.11)**.

2. **Sepal Width:**

-Negative correlation with **Petal Length (-0.42)** and **Petal Width (-0.36)**.

-Weak relationship with Sepal Length.

3. **Petal Length:**

-Very strong positive correlation with **Petal Width (0.96)**.

-Strong positive correlation with **Sepal Length (0.87)**.

4.Petal Width:

-Strongly correlated with **Petal Length (0.96)** and also with **Sepal Length (0.82)**.

5. Overall:

-Petal Length and Petal Width are the most highly correlated features.

-Sepal Width behaves differently, showing weak or negative correlation with other features.

Conclusion:

The Iris dataset contains no missing values. Feature visualizations show clear differences among species, especially in petal dimensions. Correlation analysis indicates strong relationships between petal length and petal width.