Exploratory Data Analysis of the Iris Dataset

Marek Homiak

November 14, 2025

1 Introduction

The Iris dataset consists of 150 samples of iris flowers from three species: *setosa*, *versicolor*, and *virginica*. Each sample includes four features: Sepal Length, Sepal Width, Petal Length, and Petal Width. This dataset is widely used for classification studies because *setosa* can be clearly separated from the other species, while *versicolor* and *virginica* partially overlap.

2 Data Overview

The dataset contains 150 rows and 5 columns. All features are numeric except the target variable *Species*. Basic descriptive statistics reveal the ranges, medians, and variations of each feature. The dataset is balanced, with each species represented by 50 samples. Figure 1 illustrates the distribution.

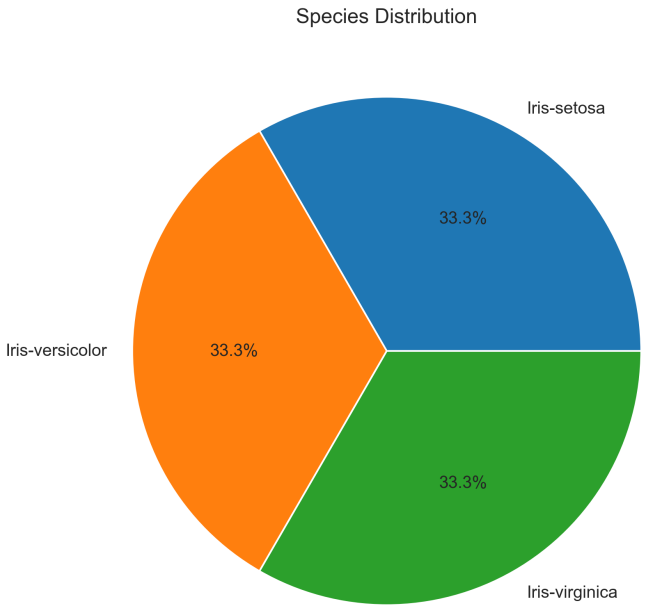

Figure 1: Distribution of Iris species.

3 Data Cleaning

Before performing analysis, the dataset was checked for missing values and duplicates:

- There are no missing values in any of the columns, so no imputation is necessary.
- There are no duplicate entries, confirming that each observation is unique.

This ensures that the dataset is complete and reliable for exploratory analysis.

4 Exploratory Data Analysis

4.1 Univariate Analysis

Univariate analysis examines individual features to understand their distributions and variability. The distributions of all features are summarized in Figure 2.

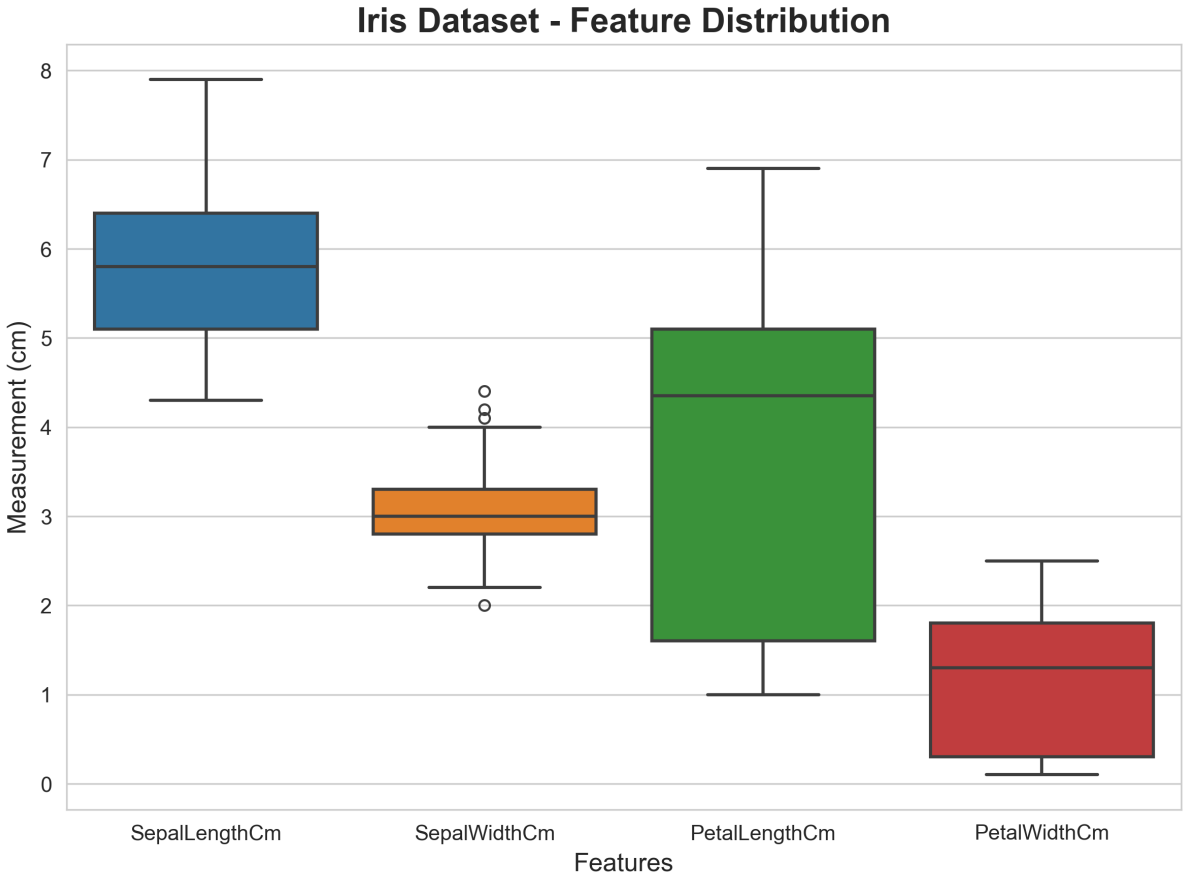

Figure 2: Boxplots showing the distribution of all features.

- **Sepal Length:** The sepal length varies moderately, with most measurements between 5 and 6.5 cm and a median around 5.8 cm, providing limited differentiation among species.
- **Sepal Width:** Sepal width shows the least variation (2.5–3.5 cm, median 3.0 cm), suggesting it is less informative for species separation.

- **Petal Length:** Exhibits the largest variation (1–7 cm, median 4.3 cm), making it highly useful for distinguishing species, particularly *setosa*.
- **Petal Width:** Also shows considerable spread (0.2–2.5 cm, median 1.3 cm) and complements petal length in discriminating between species.

Figure 3 shows the density of petal lengths for each species.

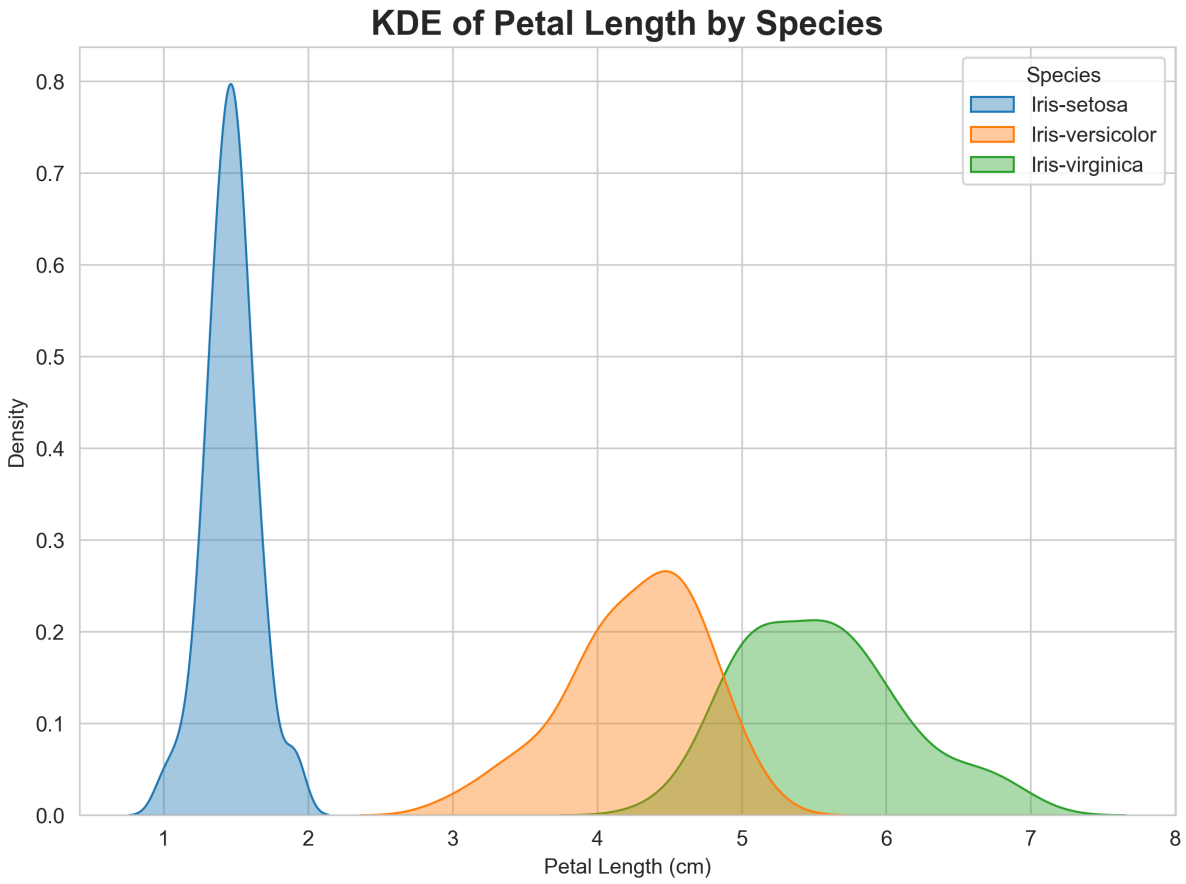

Figure 3: Kernel density estimate of Petal Length by species.

- *Setosa* flowers are concentrated around 1.5 cm, making them easily separable.
- *Versicolor* flowers are mostly between 3.5 and 5 cm, partially overlapping with *virginica*.
- *Virginica* flowers range between 4.5 and 7 cm, with wider spread and lower density.

Examining sepal width (Figure 4):

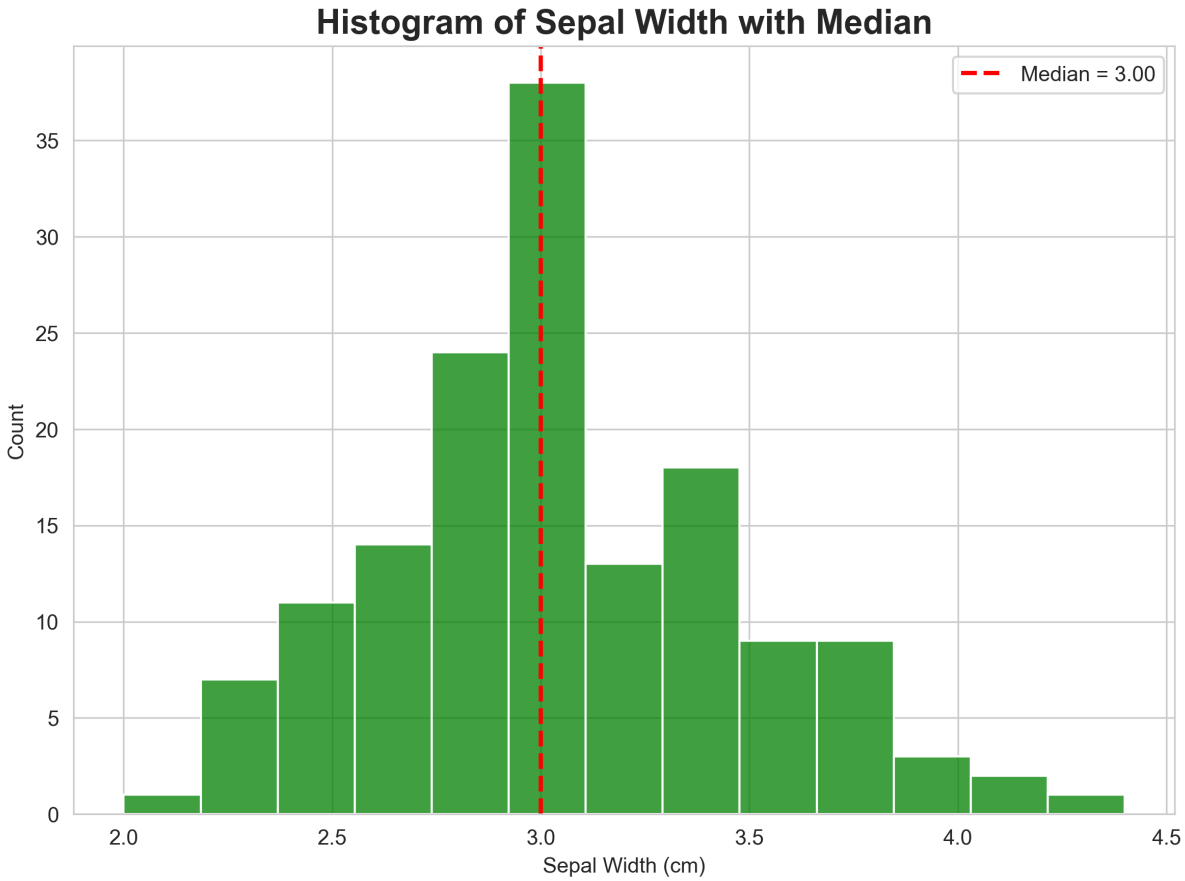

Figure 4: Histogram of Sepal Width with median line.

- Most flowers have a sepal width near 3.0 cm; the symmetric distribution shows low variability.
- A few flowers have extreme values, suggesting minor outliers.

4.2 Bivariate Analysis

Bivariate analysis examines relationships between features. Figure 5 shows the correlation matrix.

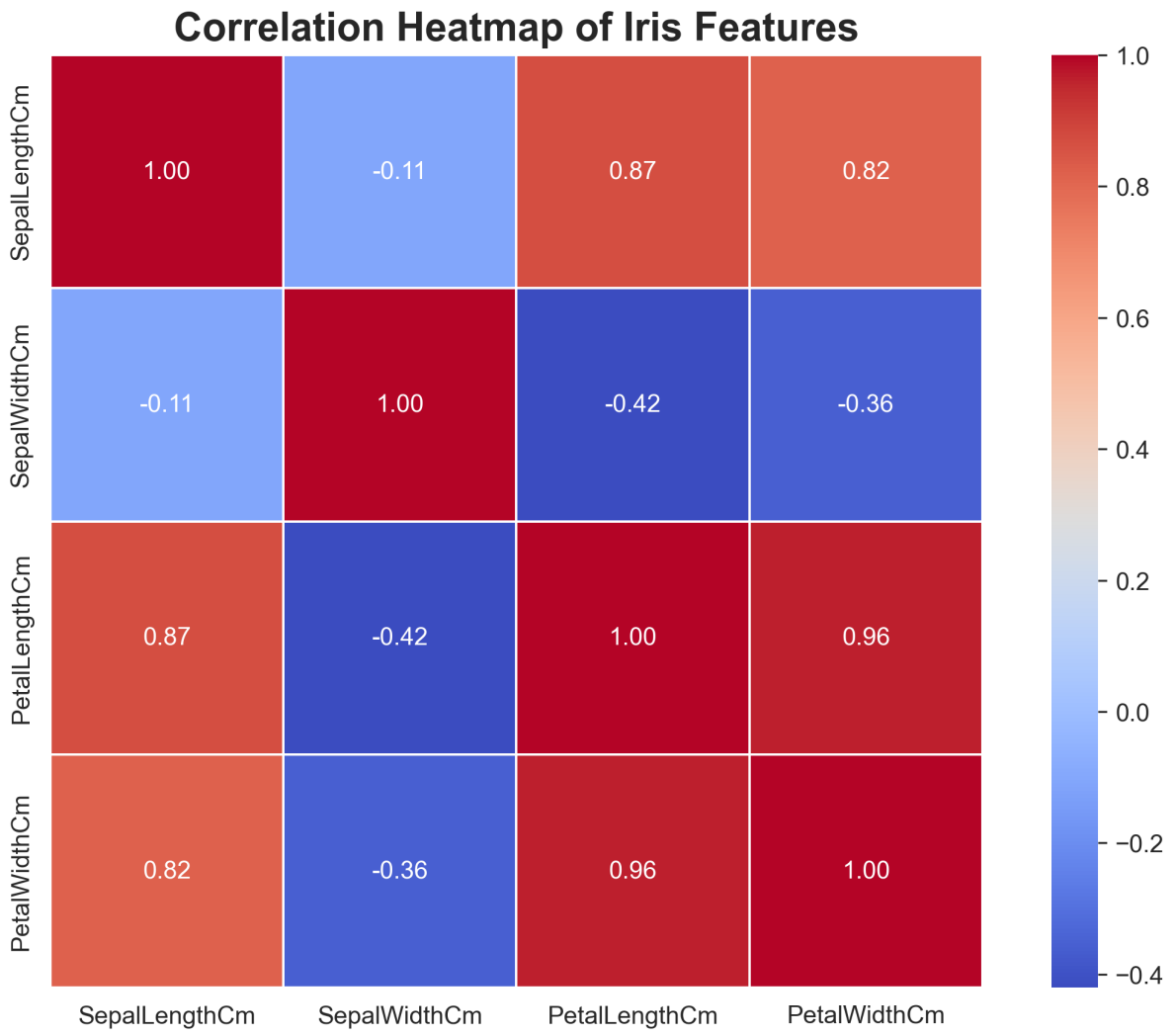

Figure 5: Correlation heatmap of features.

- Petal length and petal width are strongly correlated (0.96), meaning longer petals tend to be wider.
- Petal length and sepal length are moderately correlated (0.87), reflecting that larger flowers have longer petals and sepals.
- Sepal width shows weak correlation (-0.11) and provides unique but less predictive information.

Scatterplots further illustrate these relationships:

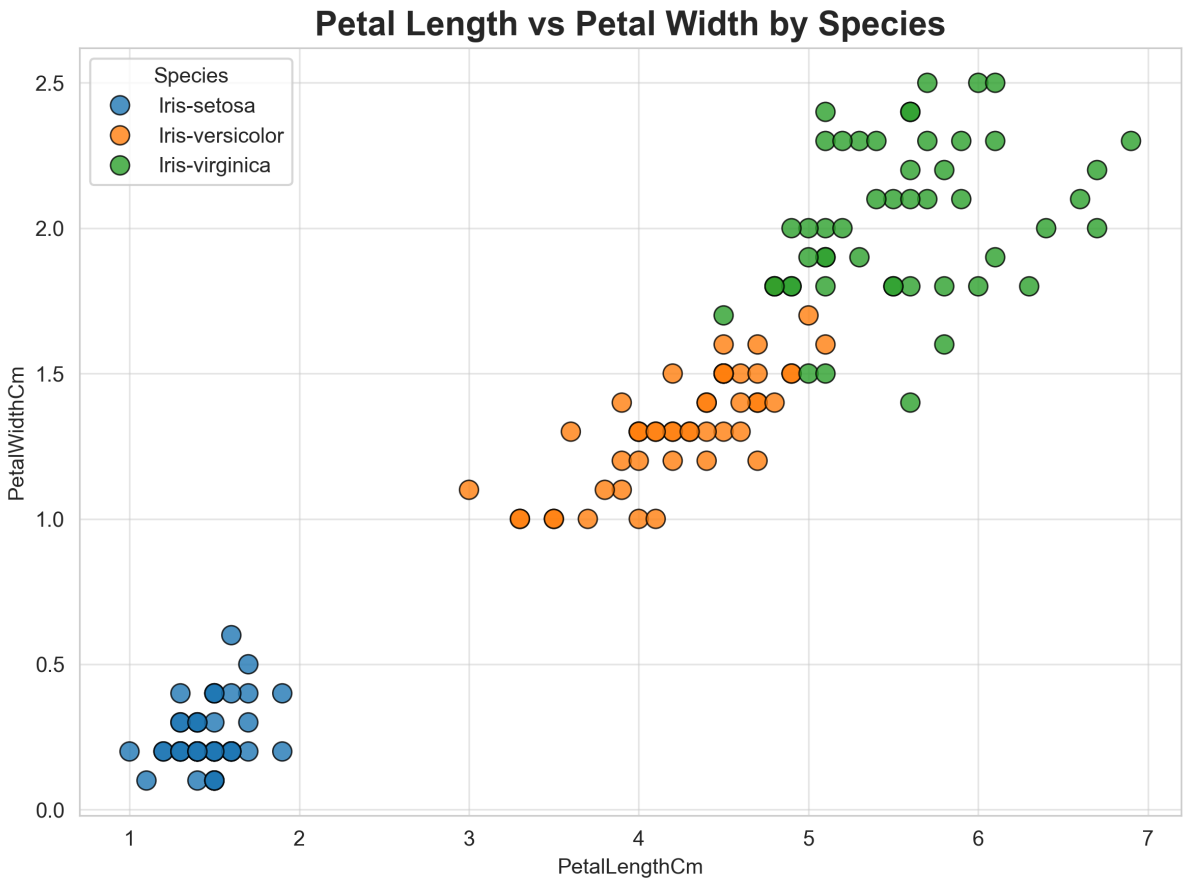

Figure 6: Petal Length versus Petal Width for all species.

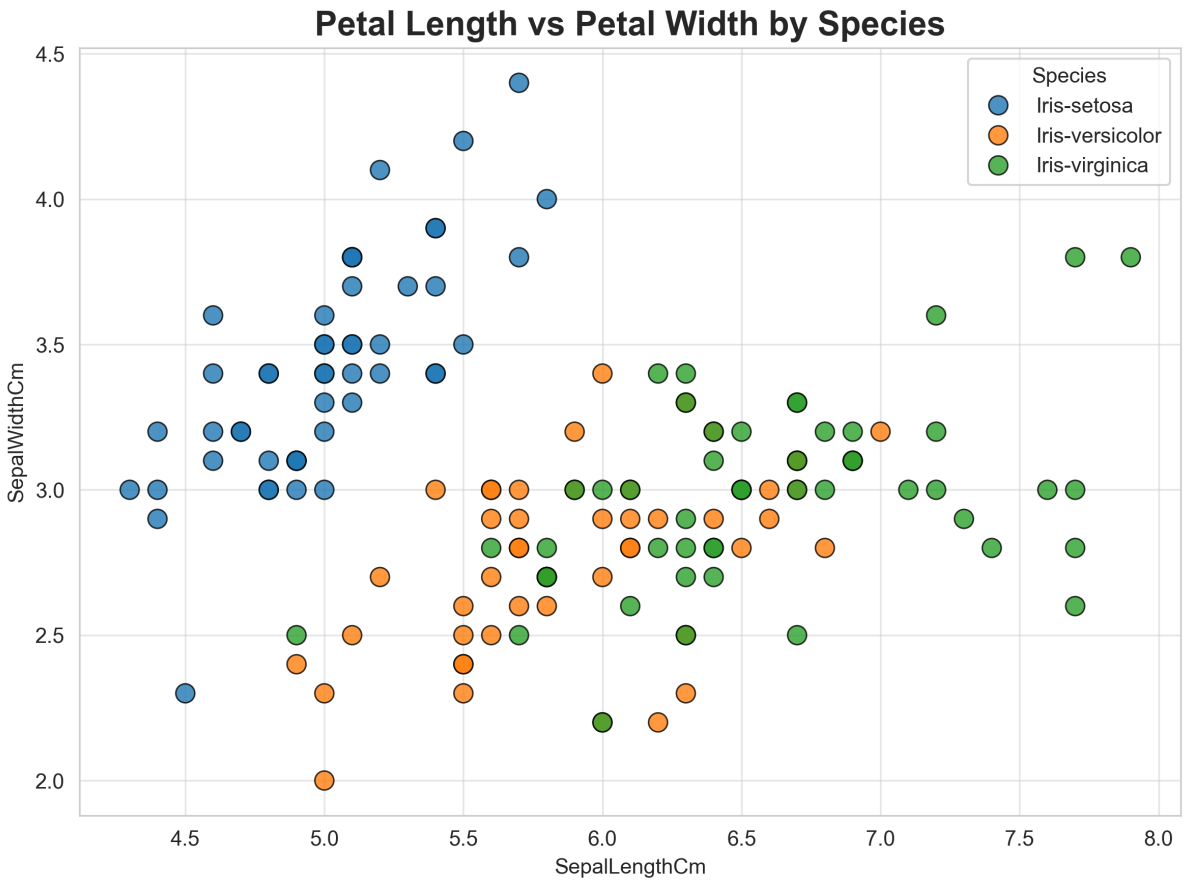

Figure 7: Sepal Length versus Sepal Width for all species.

- In petal length versus width, *setosa* forms a distinct cluster; *versicolor* and *virginica* overlap partially.
- In sepal length versus width, Versicolor and Virginica overlap, showing weaker separation.

Figure 8 shows petal distributions by species:

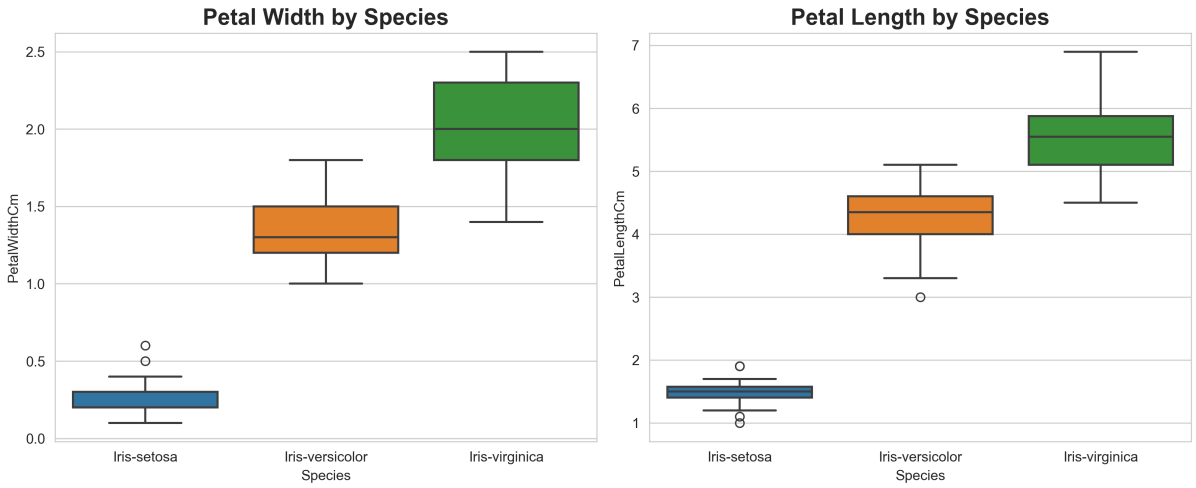

Figure 8: Petal Length and Width distributions by species.

- *Setosa* petals are small and uniform.
- *Versicolor* and *Virginica* show more variability and partial overlap.

Figure 9 presents all feature relationships together:

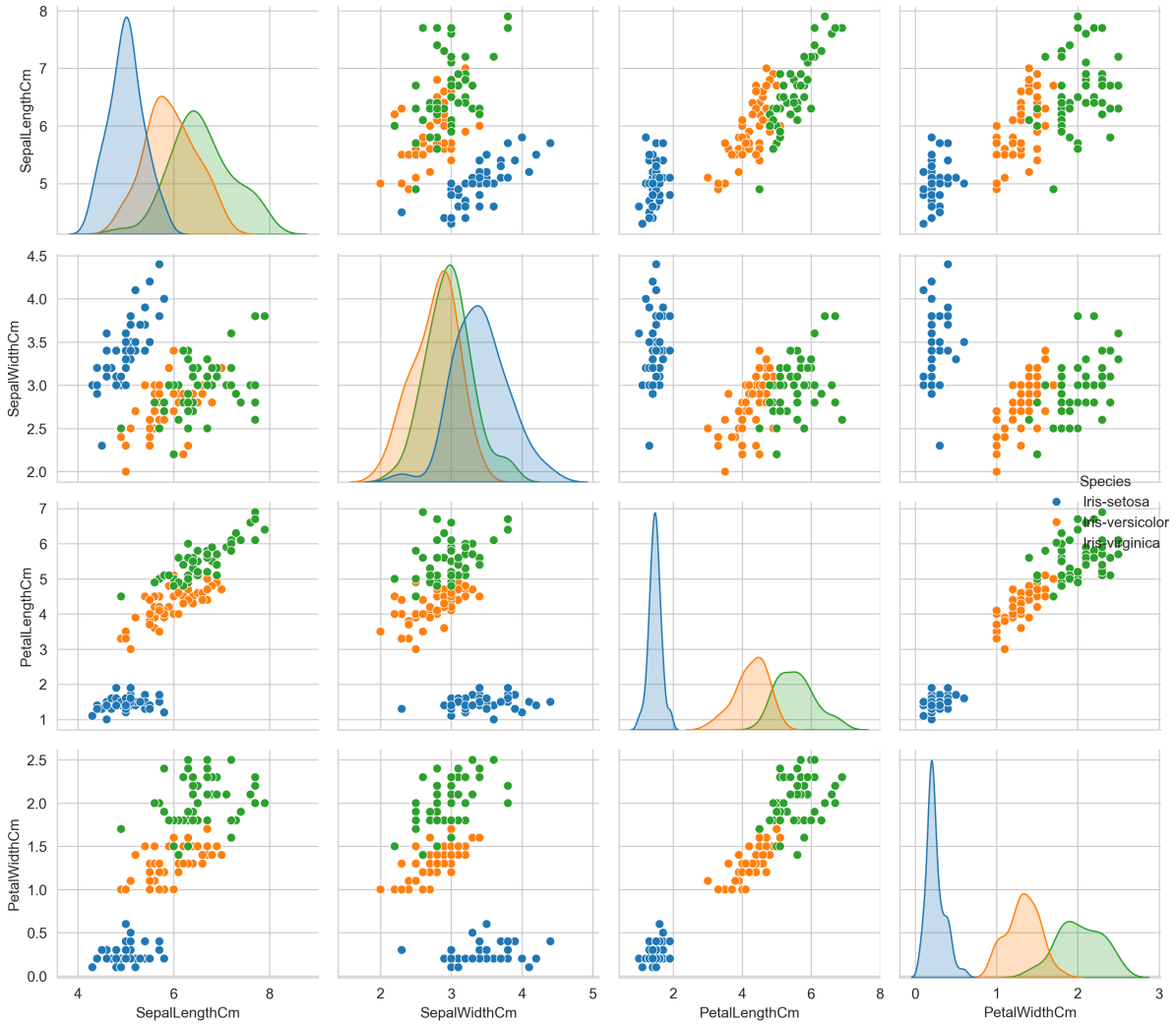

Figure 9: Pairplot of all features colored by species.

5 Insights and Conclusions

- Petal length and width are the most informative features for classification. *Setosa* is fully separable; Versicolor and Virginica partially overlap.
- Sepal length has moderate discriminative power; sepal width contributes the least.
- Strong correlation between petal features indicates redundancy; one could be omitted in predictive models.
- Overall, petal dimensions dominate species separation, while sepal measurements provide supplementary information.