# Exercise 2: Visual Exploratory Data Analysis

# Objective

To understand the role of visual exploratory data analysis (EDA) in identifying class separability and feature importance using paired plots, box plots, violin plots, and correlation heatmaps.

# Introduction

Visual EDA helps in understanding data distribution, relationships, and class separability. Before applying machine learning models, visualizations help identify:  
- Which features best separate the classes.  
- Which features have high within-class variability or overlap.  
- Correlation between features to avoid redundancy.  
  
Visualizations used:  
1. Pair Plots: Show pairwise relationships and class clusters.  
2. Box Plots: Display spread (IQR) and outliers.  
3. Violin Plots: Show distribution density and spread.  
4. Correlation Heatmaps: Summarize linear relationships between features.

# Dataset Used

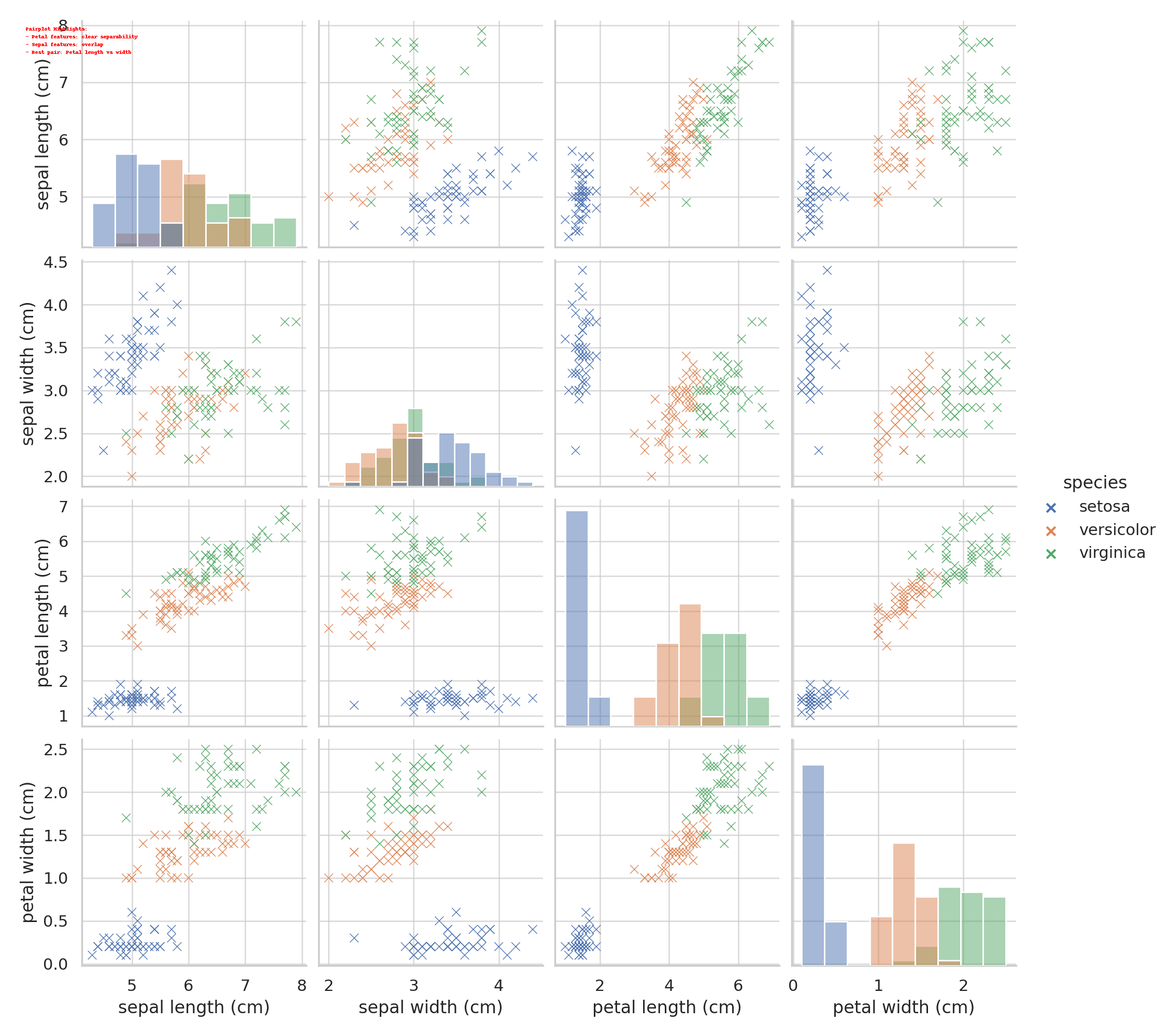
Dataset: Iris Dataset  
Features: Sepal length (cm), Sepal width (cm), Petal length (cm), Petal width (cm)  
Target: Species (Setosa, Versicolor, Virginica)

# Python Program

# Load Dataset  
from sklearn.datasets import load\_iris  
import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt  
  
iris = load\_iris()  
df = pd.DataFrame(iris.data, columns=iris.feature\_names)  
df["species"] = pd.Categorical.from\_codes(iris.target, iris.target\_names)  
  
# Pair Plot  
sns.pairplot(df, hue="species", diag\_kind="hist")  
plt.show()  
  
# Box Plots  
for col in iris.feature\_names:  
 sns.boxplot(x="species", y=col, data=df)  
 plt.title(f"Boxplot: {col}")  
 plt.show()  
  
# Violin Plots  
for col in iris.feature\_names:  
 sns.violinplot(x="species", y=col, data=df)  
 plt.title(f"Violin Plot: {col}")  
 plt.show()  
  
# Correlation Heatmap  
sns.heatmap(df.corr(numeric\_only=True), annot=True, cmap="coolwarm")  
plt.title("Correlation Heatmap")  
plt.show()

# Annotated Visualizations

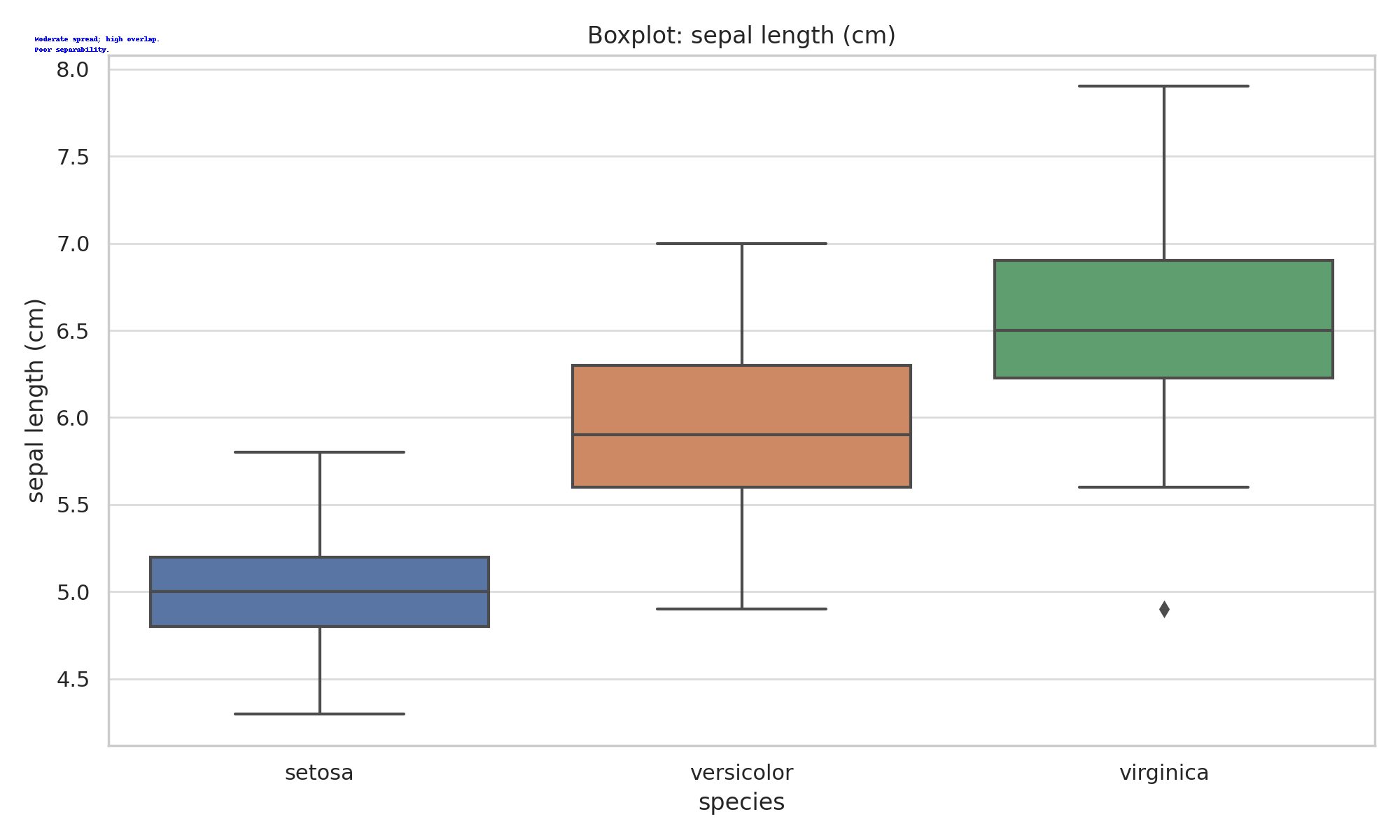
## Pair Plot (Annotated)



Highlights: Petal features show clear class separation; Sepal features overlap. Best pair: Petal length vs Petal width.

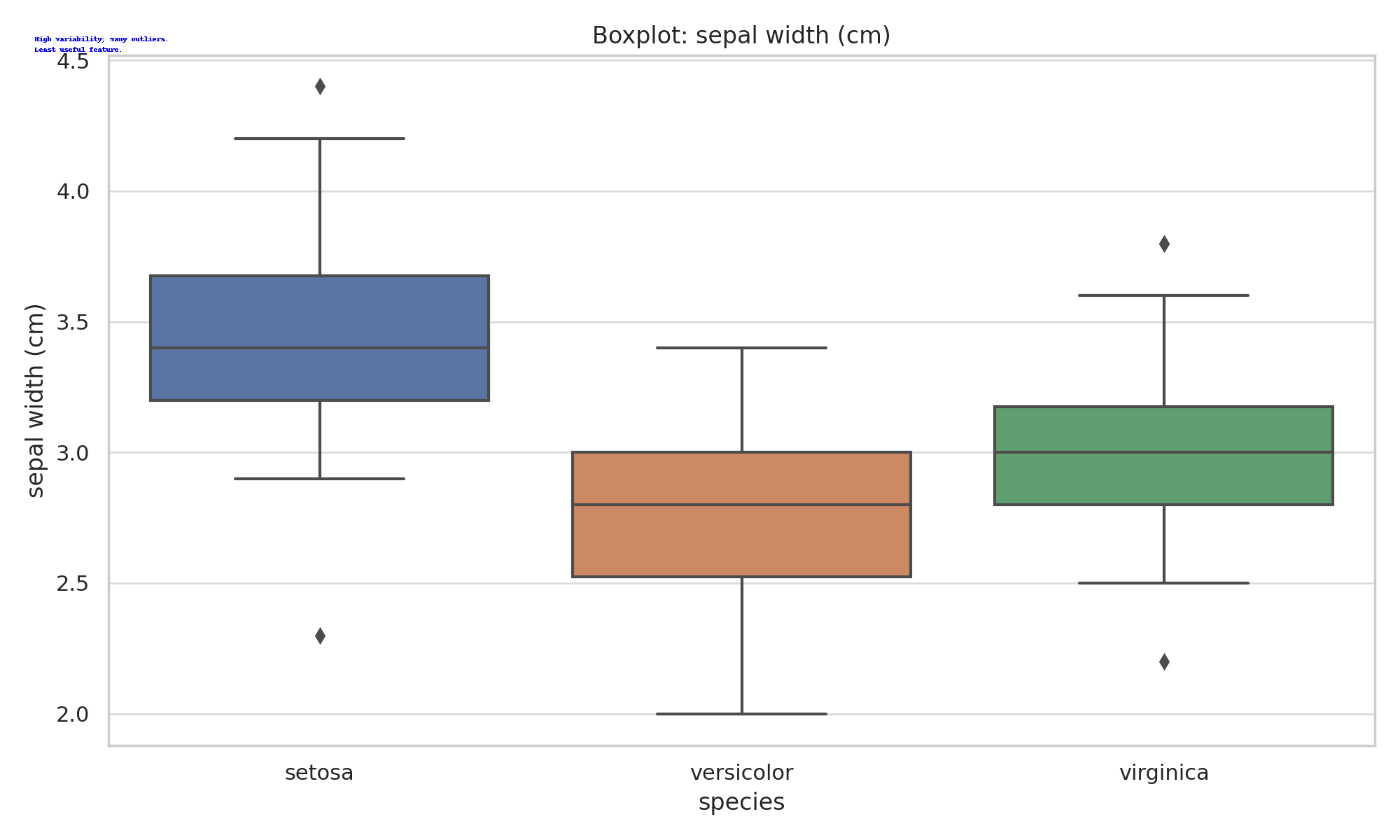
# Box Plot Analysis (Annotated)

## Sepal Length (Cm)



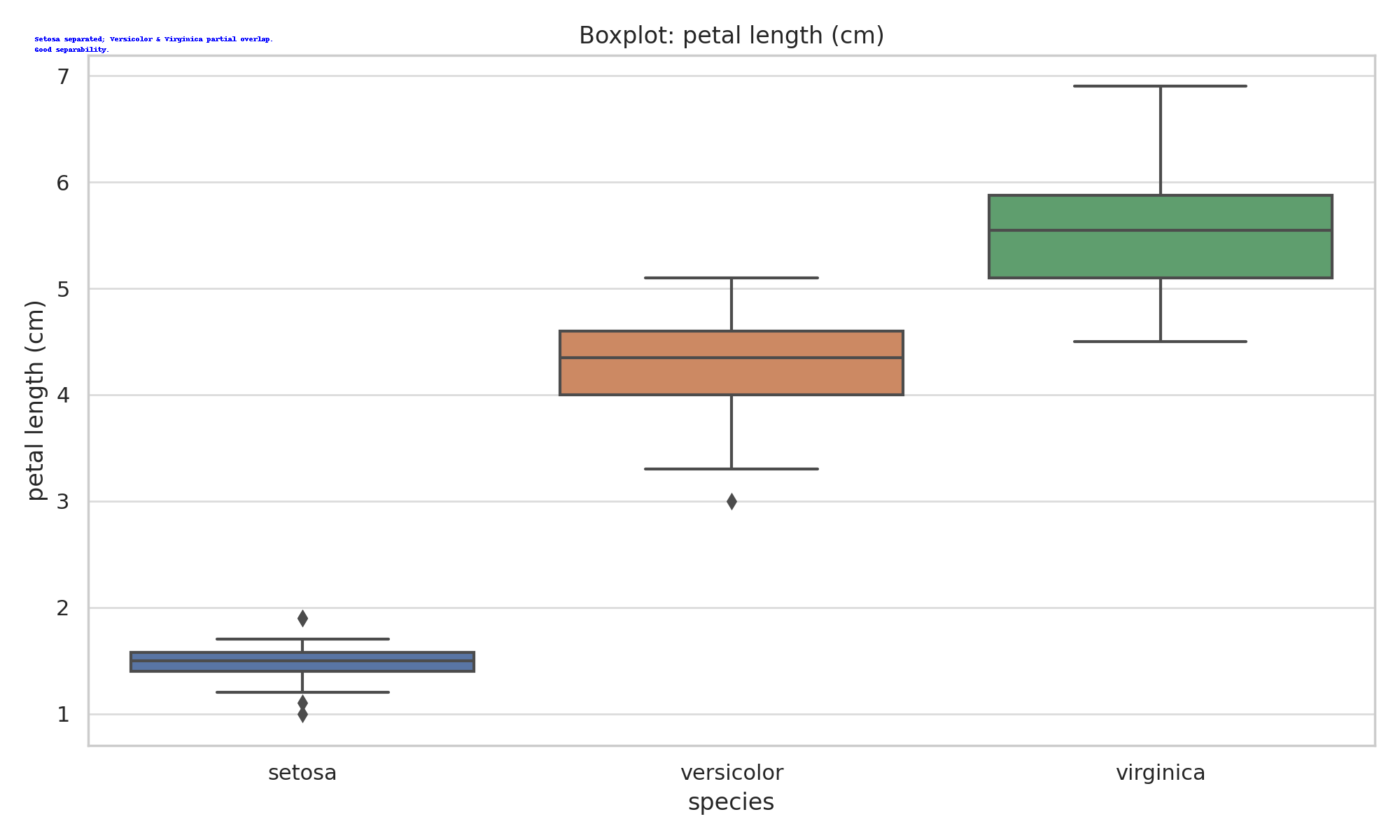
Moderate spread; high overlap.  
Poor separability.

## Sepal Width (Cm)



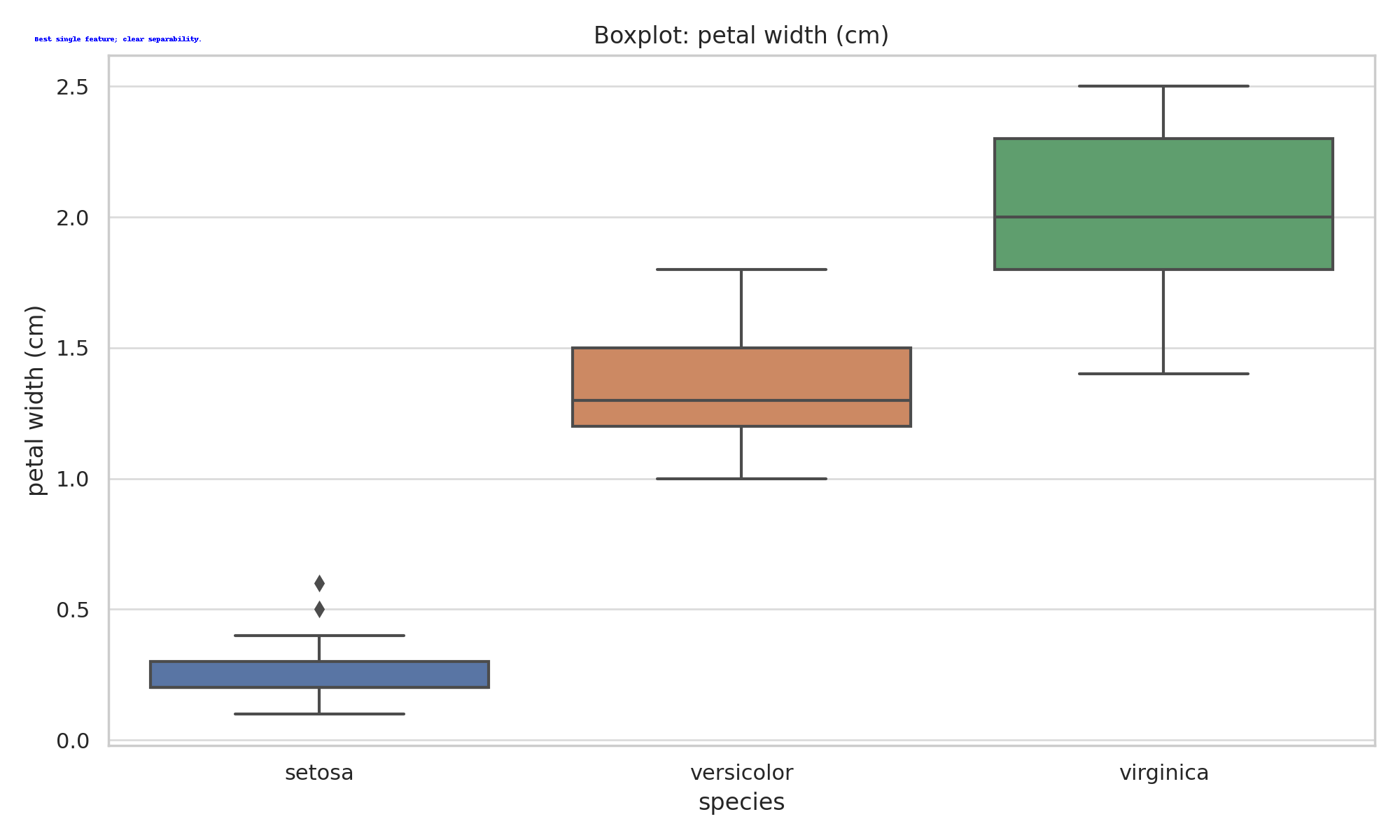
High variability; many outliers.  
Least useful feature.

## Petal Length (Cm)



Setosa separated; Versicolor & Virginica partial overlap.  
Good separability.

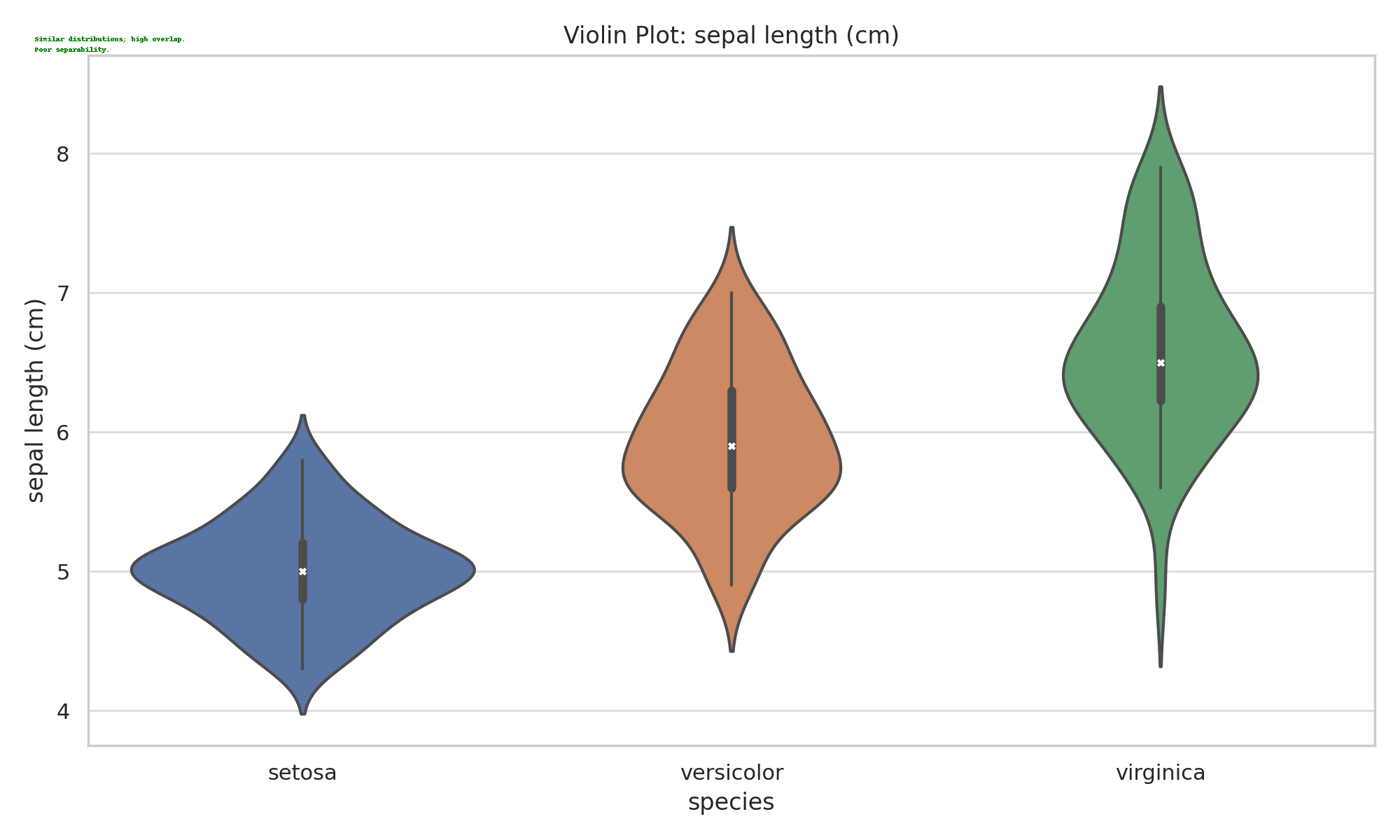
## Petal Width (Cm)



Best single feature; clear separability.

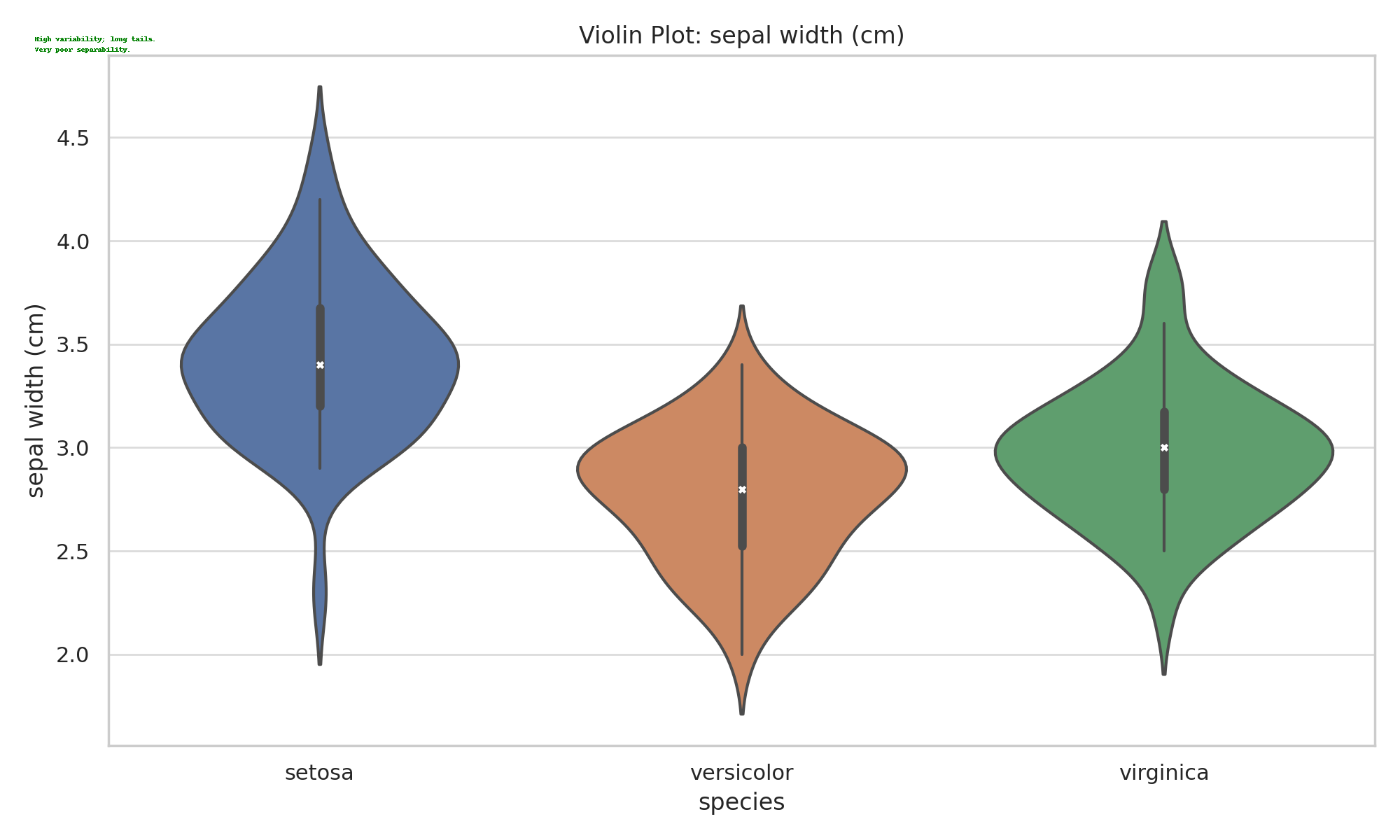
# Violin Plot Analysis (Annotated)

## Sepal Length (Cm)



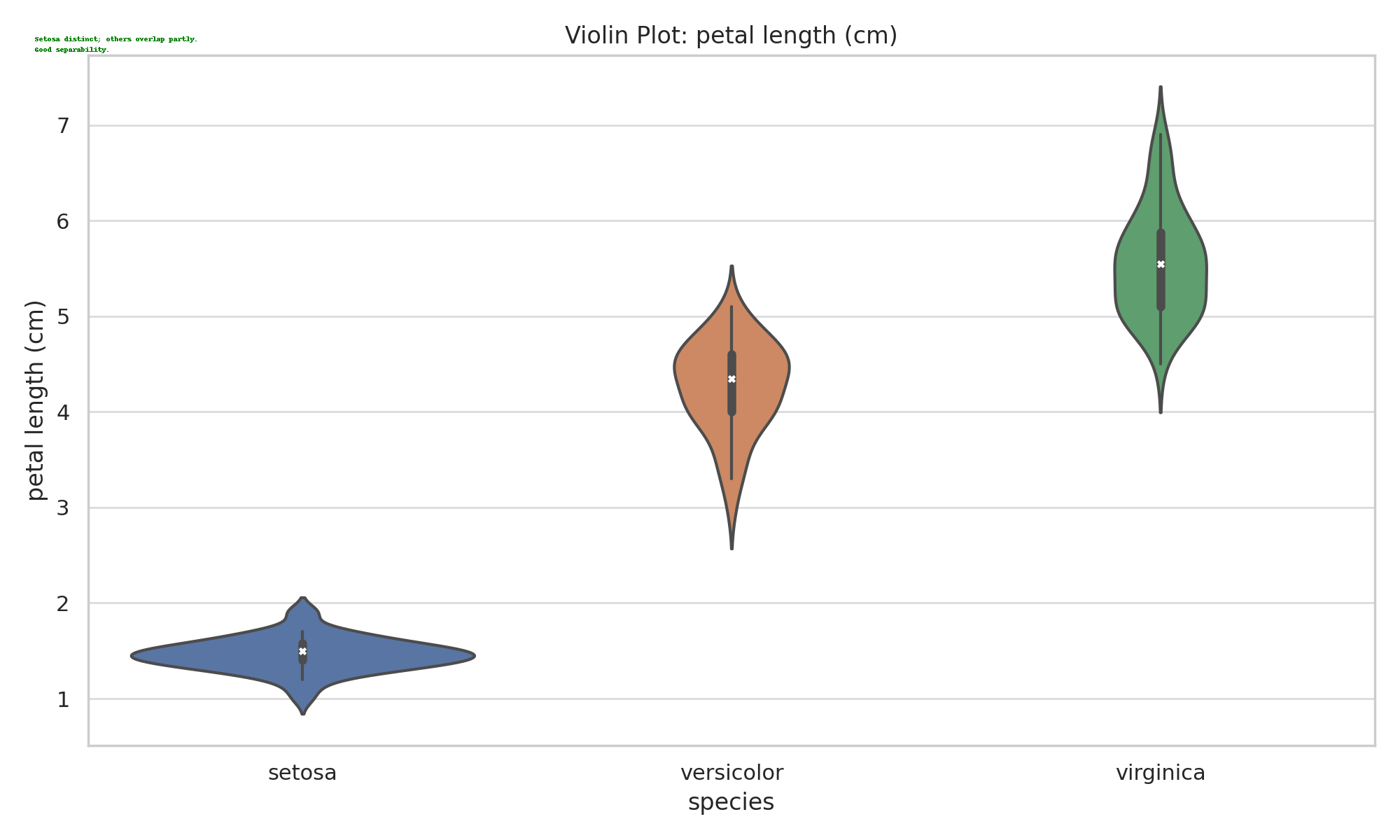
Similar distributions; high overlap.  
Poor separability.

## Sepal Width (Cm)



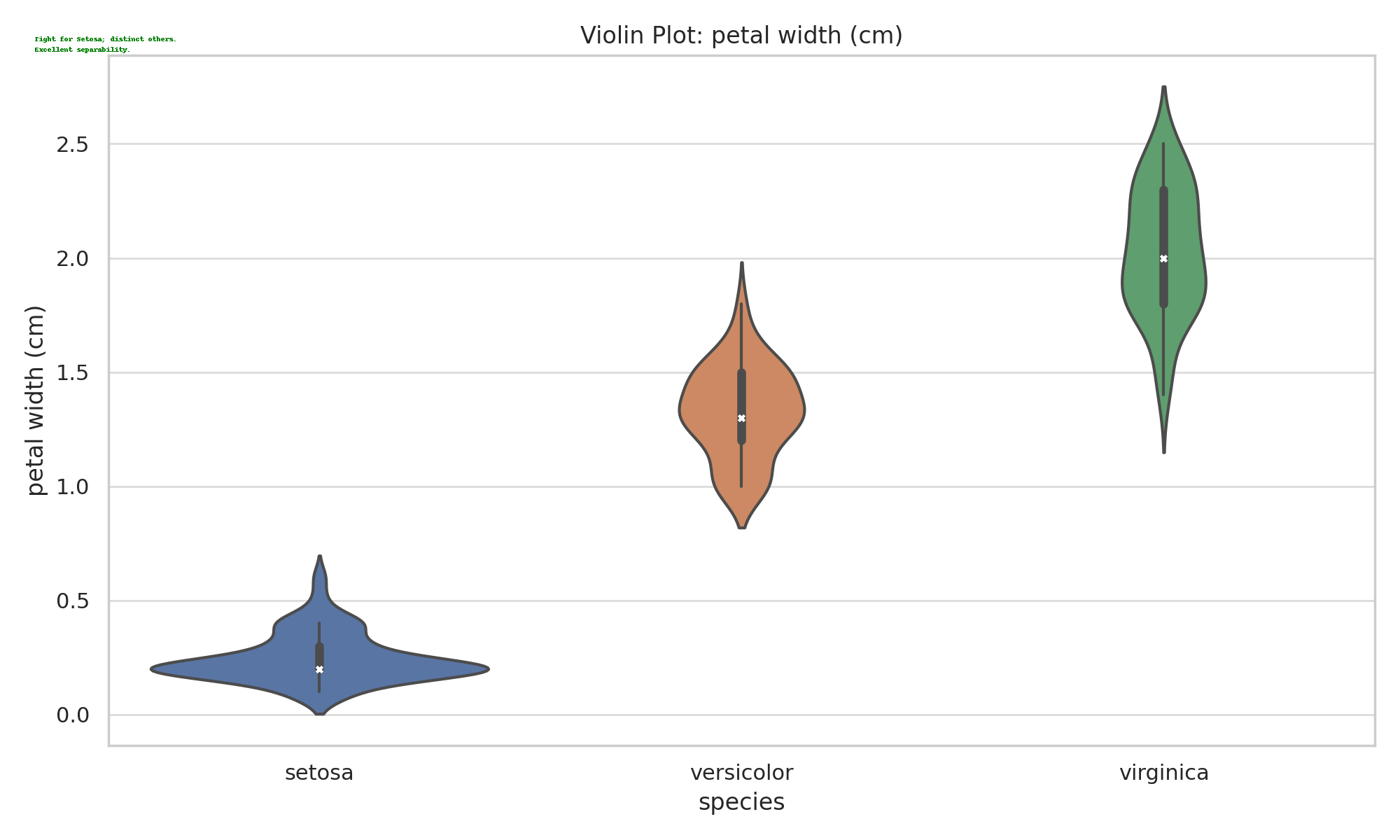
High variability; long tails.  
Very poor separability.

## Petal Length (Cm)



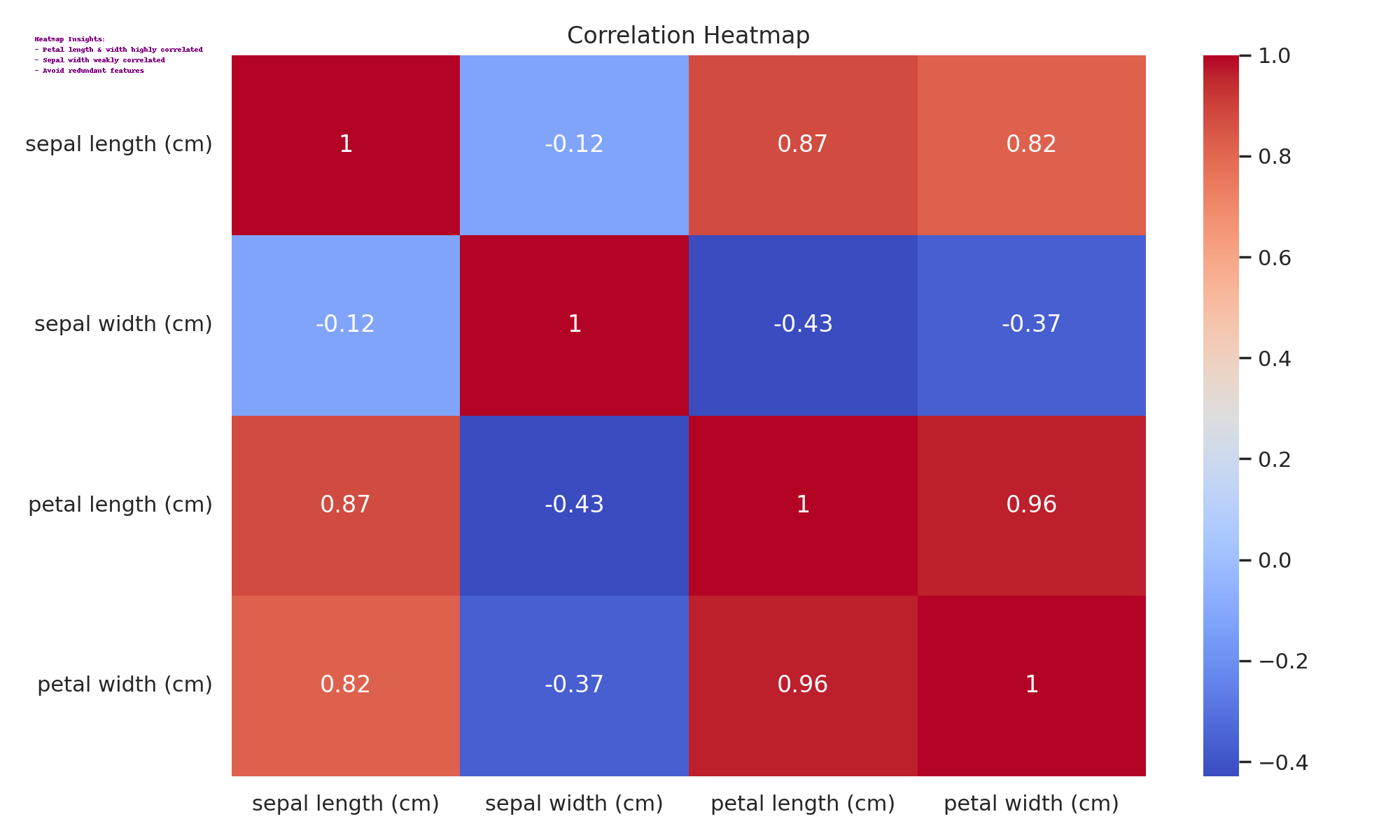
Setosa distinct; others overlap partly.  
Good separability.

## Petal Width (Cm)



Tight for Setosa; distinct others.  
Excellent separability.

# Correlation Heatmap (Annotated)



Petal length and petal width are highly correlated; sepal width shows weak correlation. Remove redundant features if necessary.

1. **Task 1** [CO1] [BTL 2] [3 marks]  
   Implement the above task in colab and share the link here (give public access). Go through the libraries and the functions called specifically for different plots. In the Iris dataset, which features are most discriminative and why?
2. **Task 2** [CO1] [BTL 3] [1 marks]

Take any mobile app classification dataset having 10 features. Concisely explain how would you proceed to find the best feature combinations to yield highest classification performance.