*CYBR 520  
Lab 1: Introduction to Exploratory Data Analysis (Python version)*

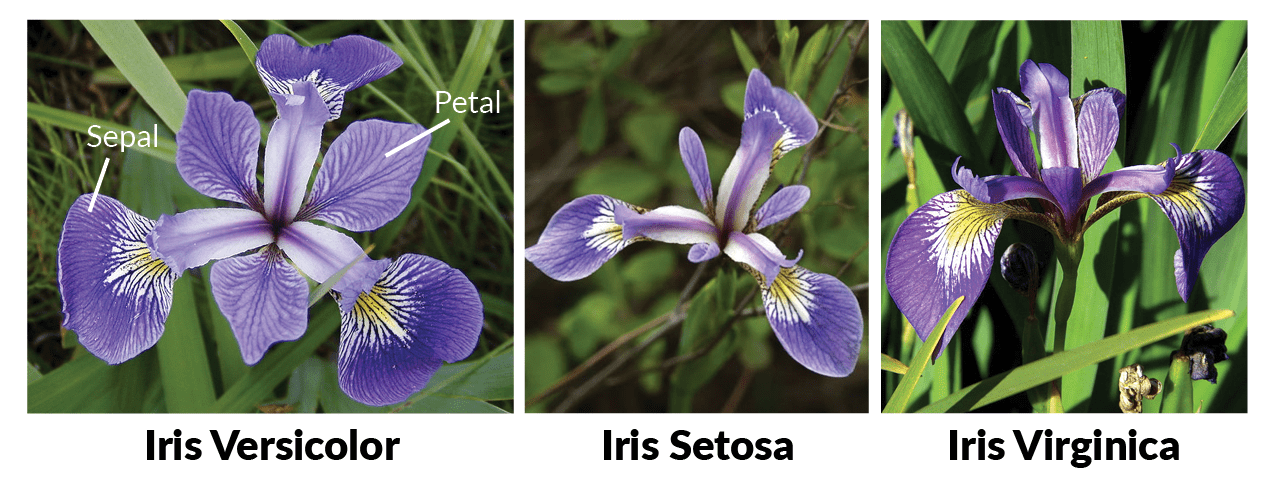


**Dataset:** Iris

**Topics:** Summary statistics, data visualization, variable relationships, feature distributions, and outlier detection

**Python Libraries:** pandas, matplotlib, seaborn, sklearn

**Background:** The iris dataset is a classic example dataset used for machine learning in Python. Most, if not all flowers have a sepal and a petal. The sepal (Figure 1) [[1]](#footnote-0) functions as a protector for the flower and also support the petals when the flower is in bloom:

The Iris dataset contains measurements (in cm) of petal length and width, and sepal length and width for 3 species of the iris flower: *iris versicolor, iris setosa,* and *iris virginica* (Figure 2)[[2]](#footnote-1). There are exactly 50 observations for each species type, bringing the total number of observations to 150. Each observation is also assigned membership to a specific species as well. The membership (aka “class label”) is assigned based upon the name of the species “versicolor”, “setosa”, “virginica”.

# **Lab Structure**

The first part of this lab is a walkthrough with basic questions. Note that a companion walkthrough video is available on eCampus. Please answer each of the questions after the associated section of the walkthrough.

The second part of the lab, is where you will explore the data and test your analytics capabilities. Answer each question fully, and present all code and plots associated with answering each question.

**Reproducibility of your code and your analysis to answer the questions are important, so include those as part of your companion walkthrough or README file with your lab submission.**

**REMEMBER:** Take your time with this. Don’t just rush through to complete the lab. Understand what you are doing, and how to leverage analytical techniques and software to solve cybersecurity problems. This lab is meant to expand your knowledge and skillset, not just for a grade…

# **Setup and Data Loading**

Launch VS Code. Start a regular terminal, and save your python file. Click the “run” arrow button every time you want to run your code.

Install the following packages:

A screen shot of a computer program

AI-generated content may be incorrect.

Now we need to load the data. Type:

A computer screen with text and symbols

AI-generated content may be incorrect.

Lets now take a look at this dataset by typing:



*Question 1. What are the variable names and their data types? How many rows and columns does the dataset have?*

**The variable names are Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. Everything except species is a float64, and the species is an object. The dataset has 5 rows and 5 columns.**

# **Exploratory Data Analysis – Iris Dataset**

Now lets do some exploratory data analysis. Exploratory data analysis is a process of performing an initially investigation of the dataset using analysis and visualization techniques to discover patterns, spot outliers and anomalies, and relationships between variables in the data. This is an often skipped, but critical step in any analysis. It can also help you develop a more informed hypothesis for scientific inquiries. In essence, we’re going to try to gain some insight on this dataset.

*Question 2. What can you tell about the range and variability of each measurement?*

A screen shot of a computer code

AI-generated content may be incorrect.

**Based on the output of the dataset, the range and variability of each measurement differs between the four features. The Sepal length and width displays moderate variation along with some overlap between the species, which makes it less effective to distinguish among them. However, the petal length and width shows a much wider range as well as a greater variability, which can help with separating the species. Overall, based on the results, we can assess that the petal measurements vary the most, while the sepal measurements remain more consistent across the dataset.**

Lets now look at some univariate (single variable) visualizations

A computer code on a black background

AI-generated content may be incorrect.

*Question 3. Which features show overlapping distributions between species? Which are more distinct?*

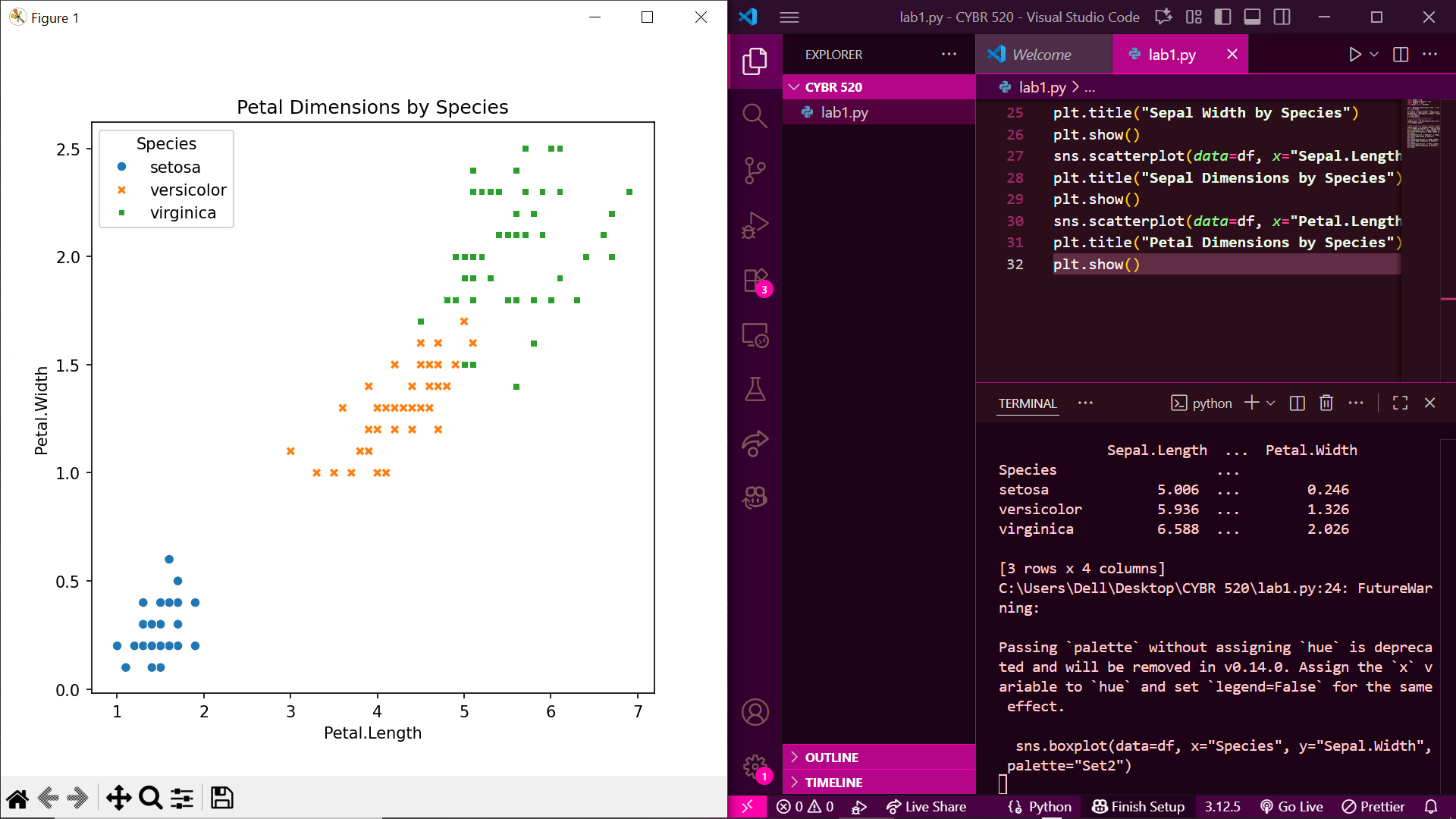
***The sepal length by species showed the most overlapping distributions between species, specifically between the versicolor and virginica species. The more distinct is the sepal width by species between Setosa and virginica. Overall, Setosa and Virginica have the most distinct features, while versicolor and virginica have the most overlapping features.***

Now lets try some bivariate relationships!

A computer screen shot of text

AI-generated content may be incorrect.

*Question 4. Which pair (sepal or petal) shows clearer separation between species?*

***Petal show clearer separation between species.*** 

Finally, let’s explore some correlations. If you have too many plots, use the left and right arrows in the visualization pane to go back and forth. These plots may also be very small, so use the zoom button as well which will bring up a pop up that you can stretch out!

A computer code with colorful text

AI-generated content may be incorrect.

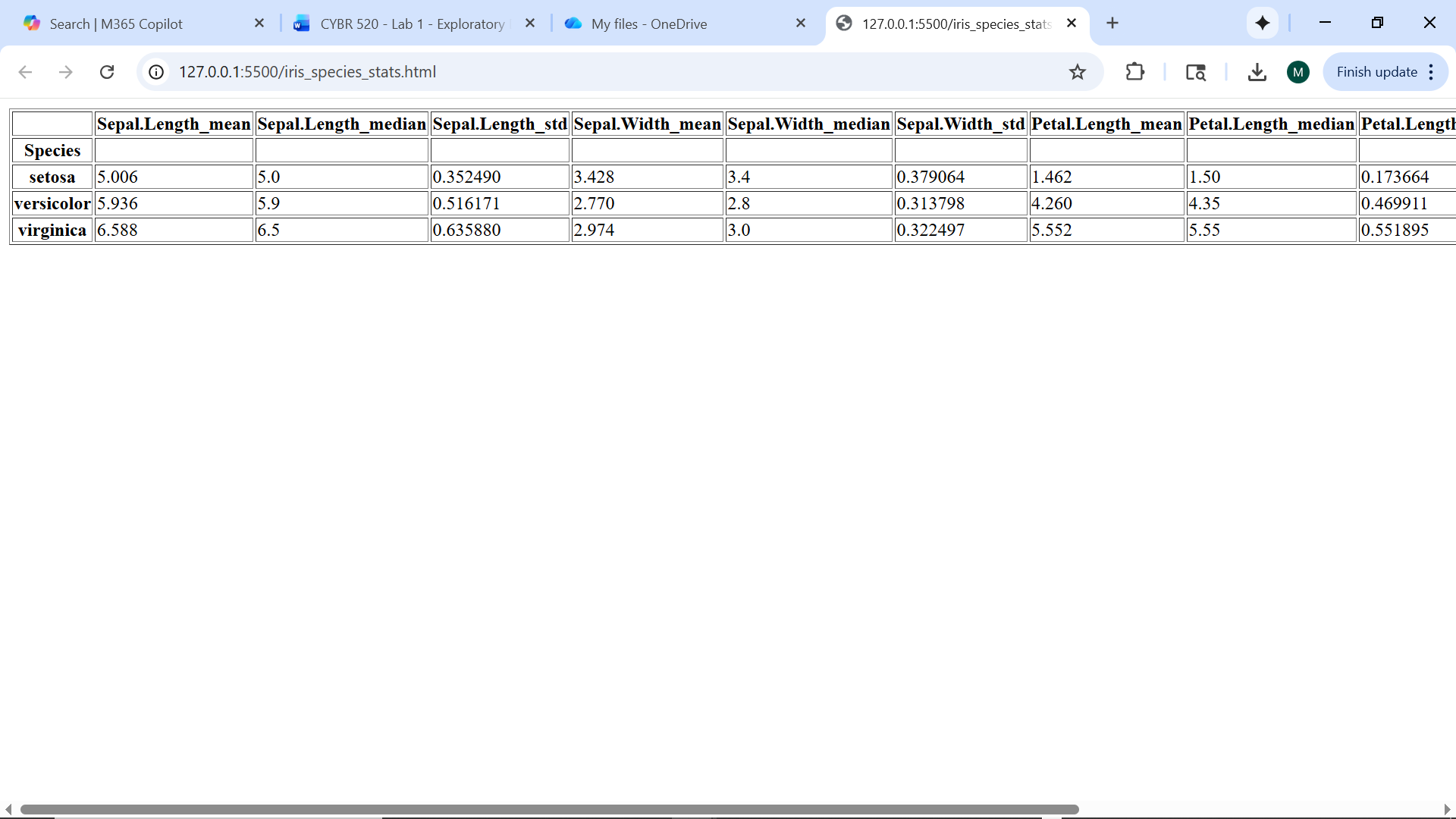
A screen shot of a computer

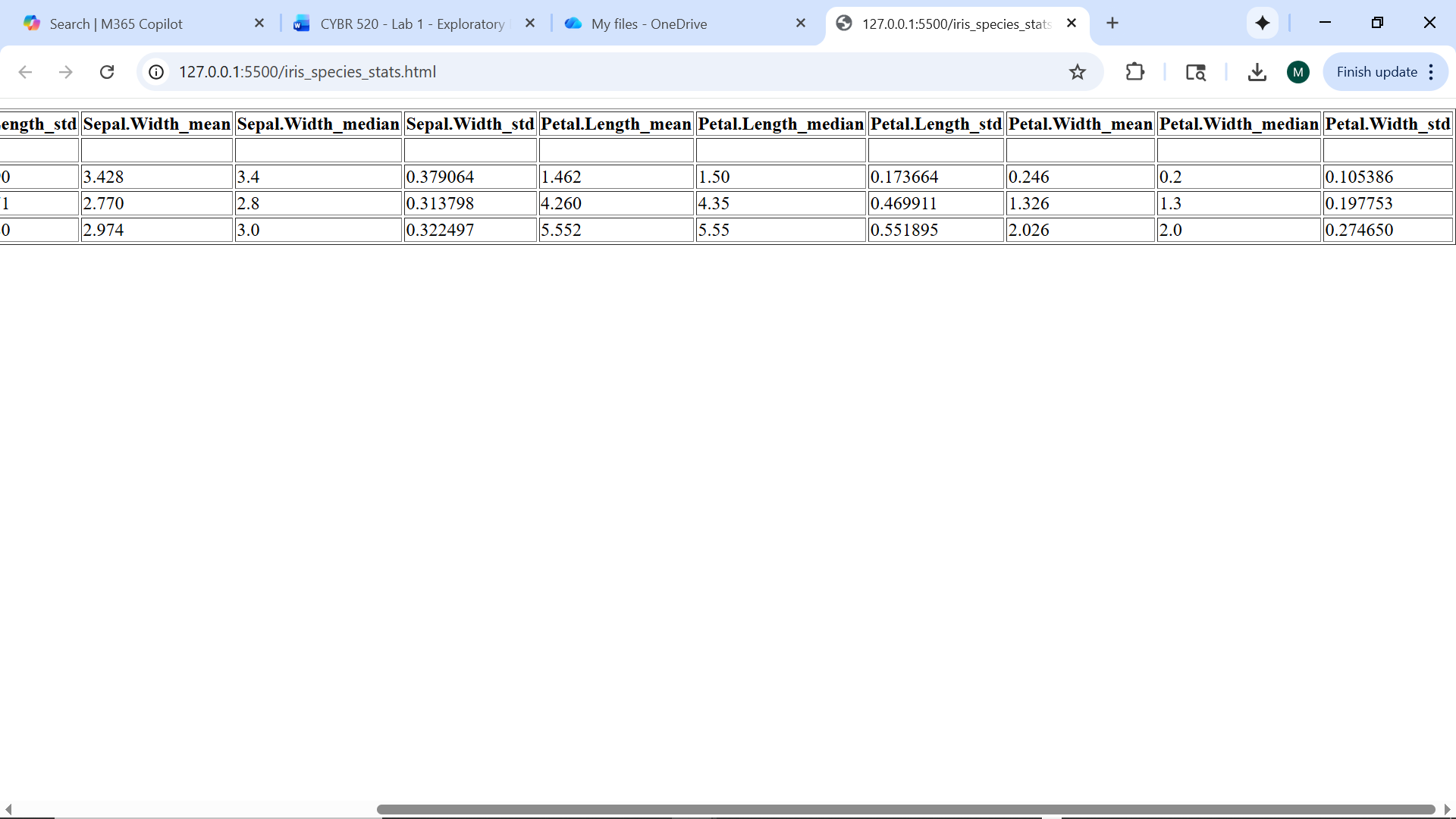
AI-generated content may be incorrect.

# **Exploratory Section (Independent EDA)**

Use any pandas,matplotlib, sklearn functions, or any other python library to answer these questions using the Iris dataset. Include your code and explanations within a readme file!

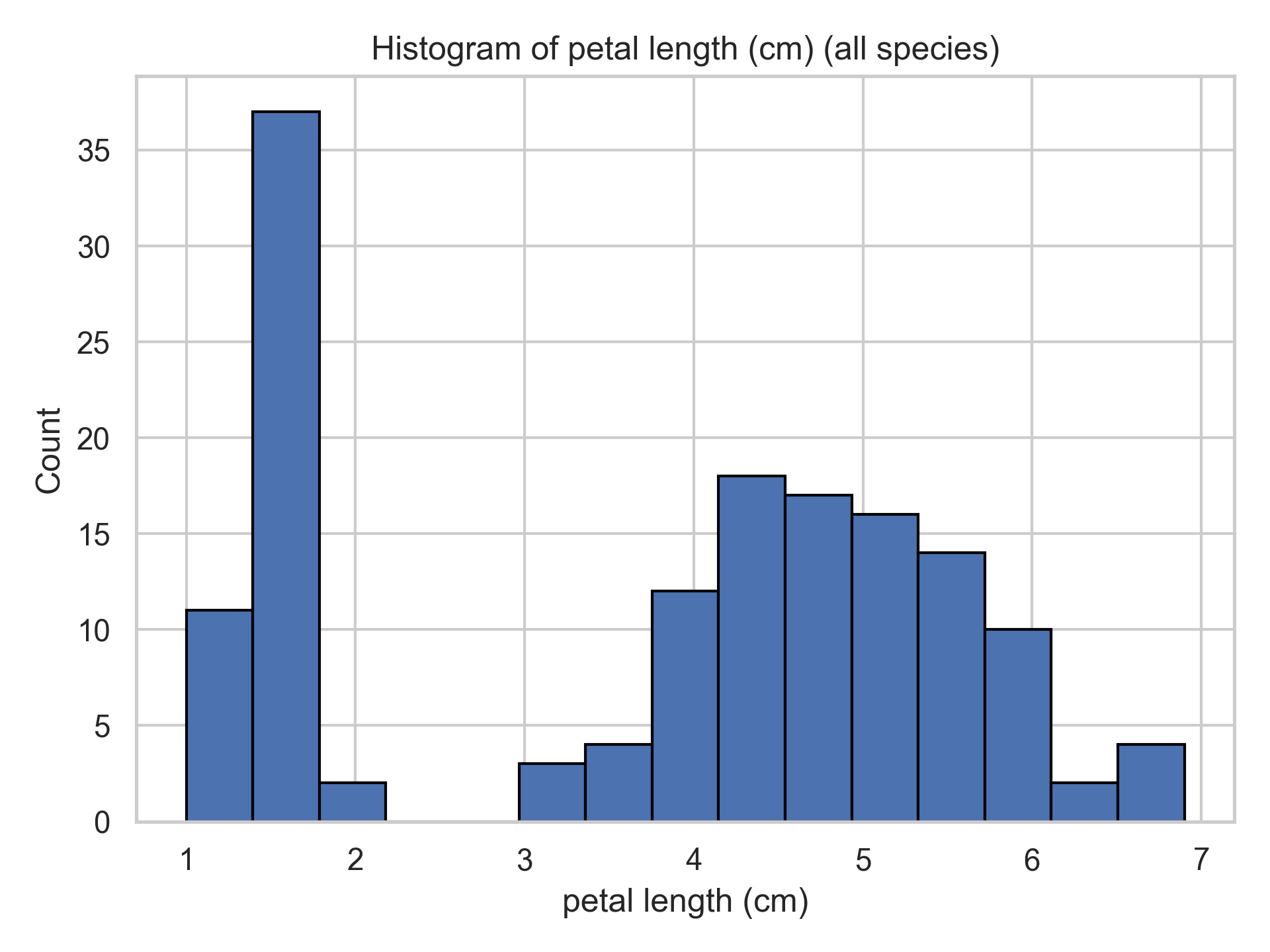
*Question 5. Compute mean, median, and standard deviation for each numeric variable by species.*

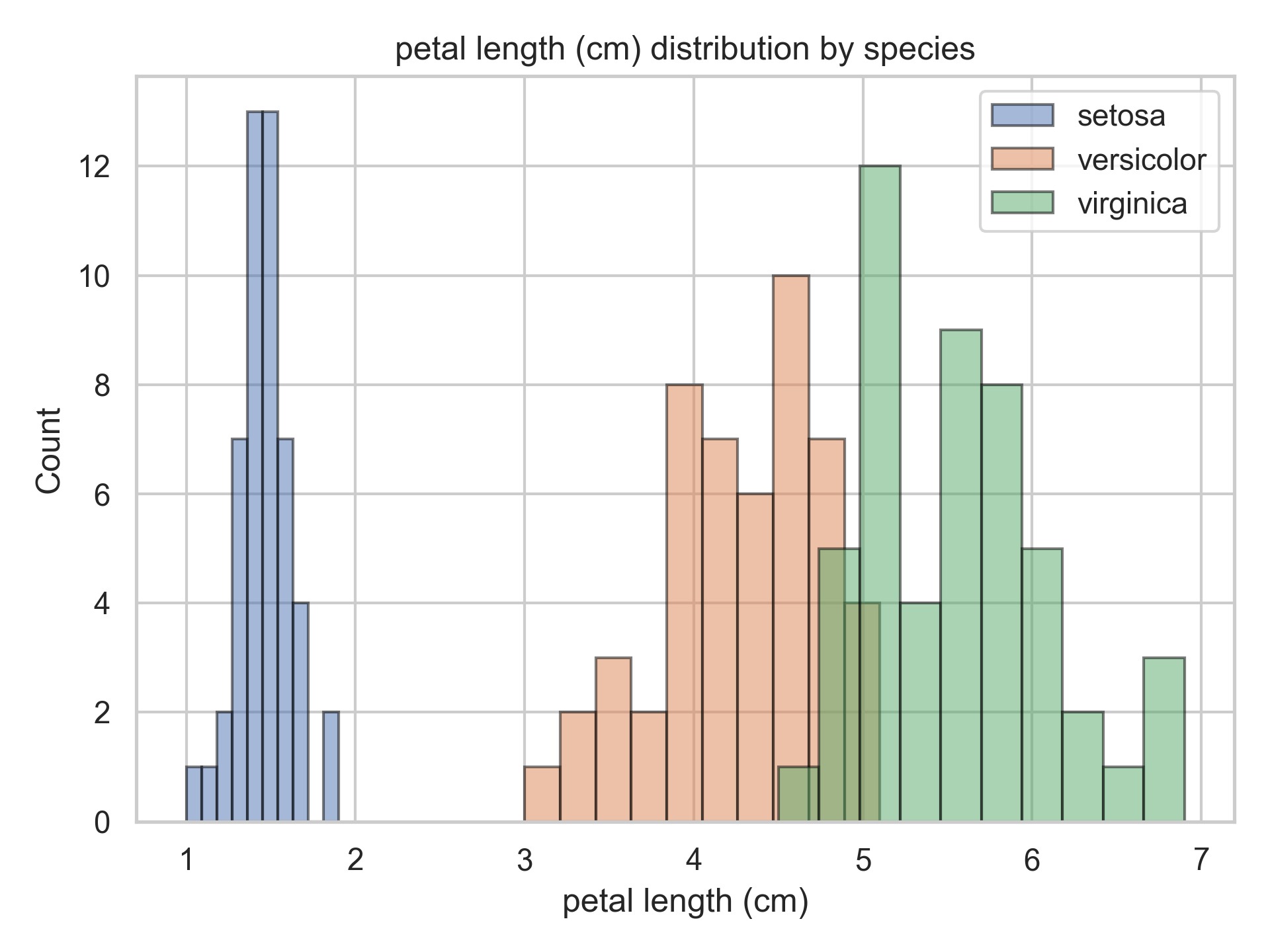
**

**

*Question 6: Plot a histogram for each numeric variable. Comment on shape (skew, symmetry).*

**The overall histograms demonstrate that petal length and width are positively skewed, whereas sepal values are nearly symmetric. The by-species histograms show three different peaks, demonstrating that each species has its unique range of petal sizes. Petal features exhibit better separation, whereas sepal features overlap more, implying that petals are stronger distinguishing variables.**

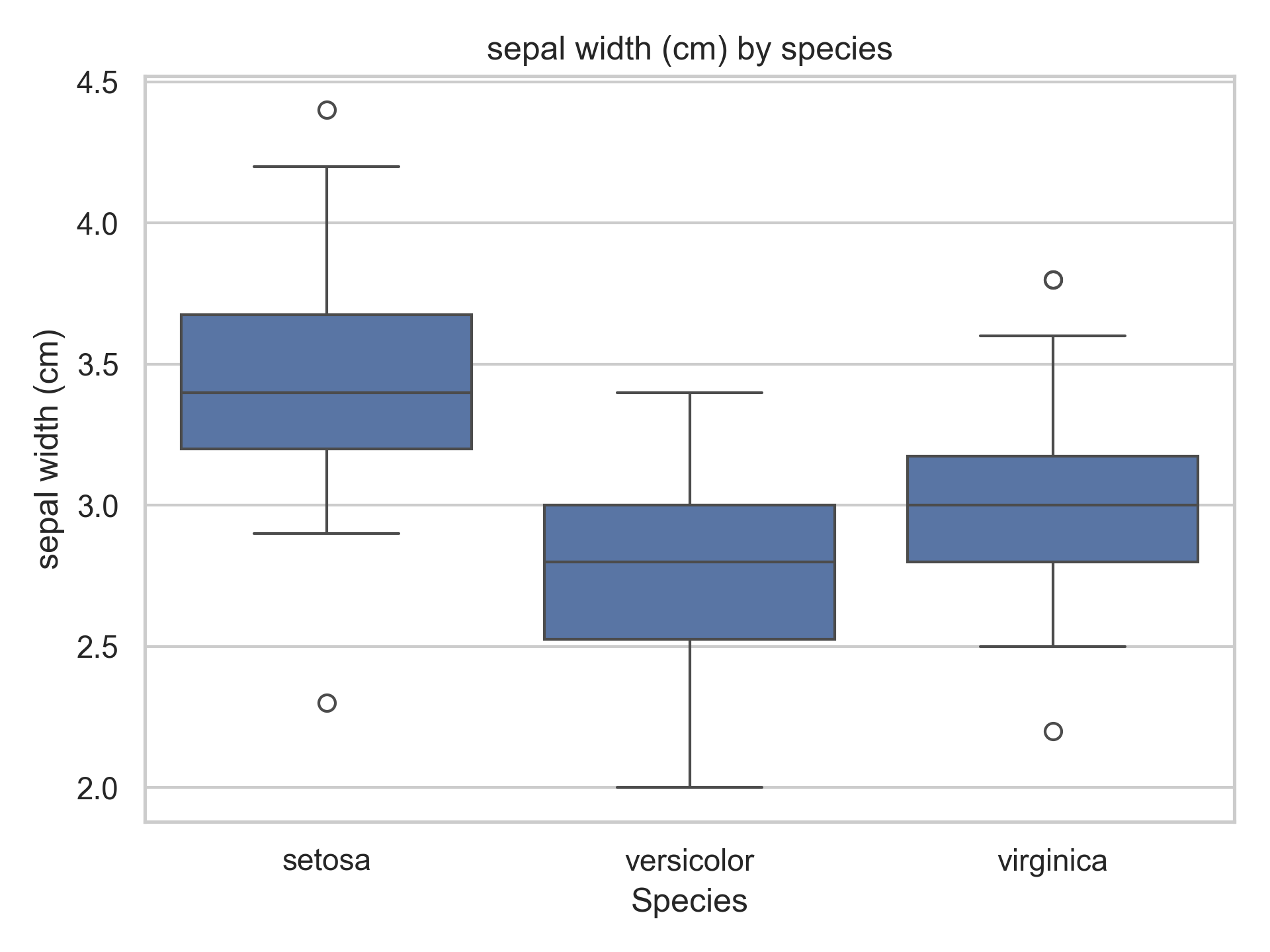
**

**

*Question 7: Create boxplots for each variable grouped by species. Identify any outliers.*

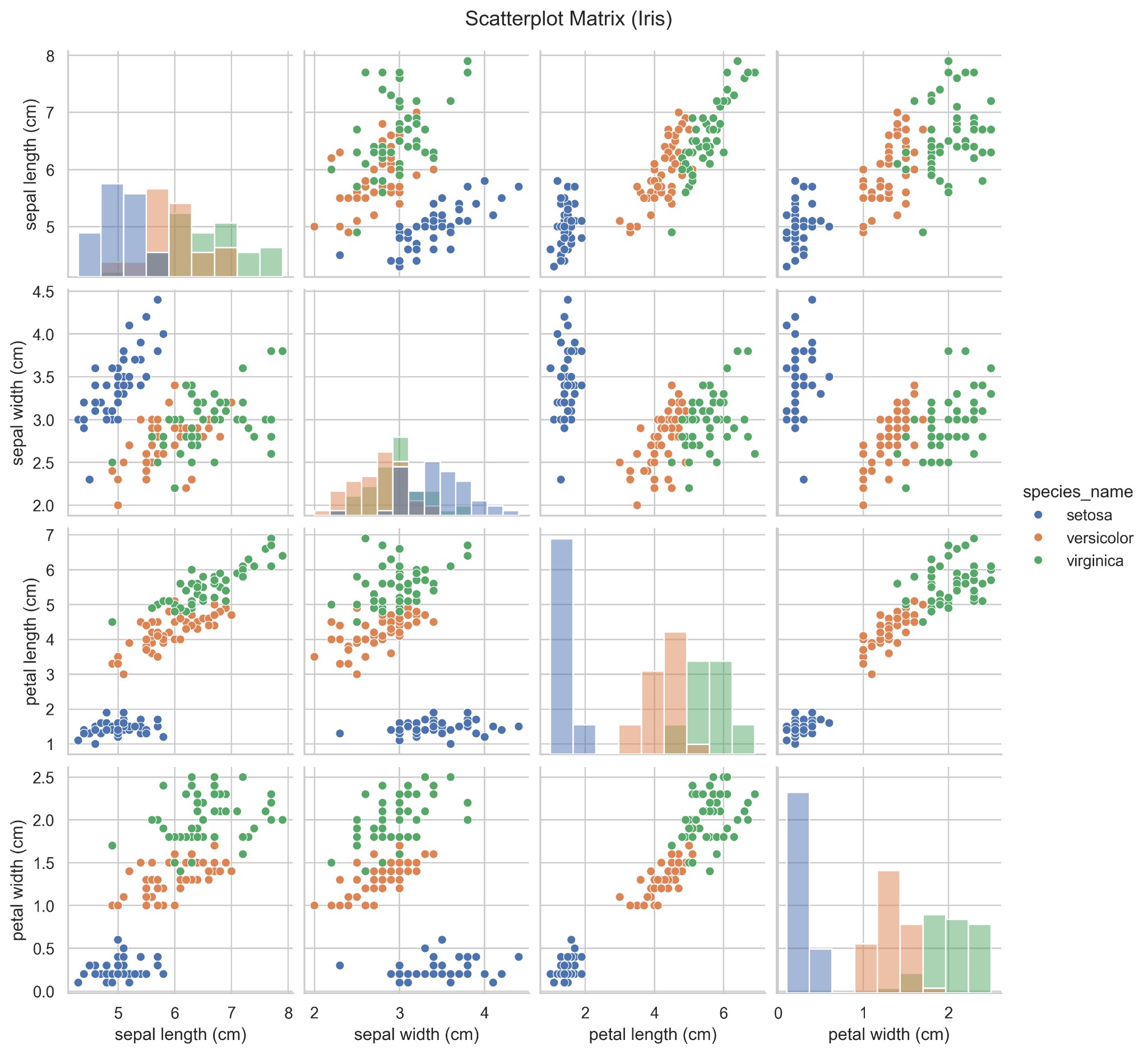
**The boxplots reveal that sepal width contains multiple potential outliers, primarily within the setosa species, although sepal length and petal dimensions are more consistent. Petal length and width have discrete, non-overlapping ranges among species, with no obvious outliers. These findings indicate that there is more diversity in sepal breadth, whereas petal traits remain stable and highly discriminative.**

**

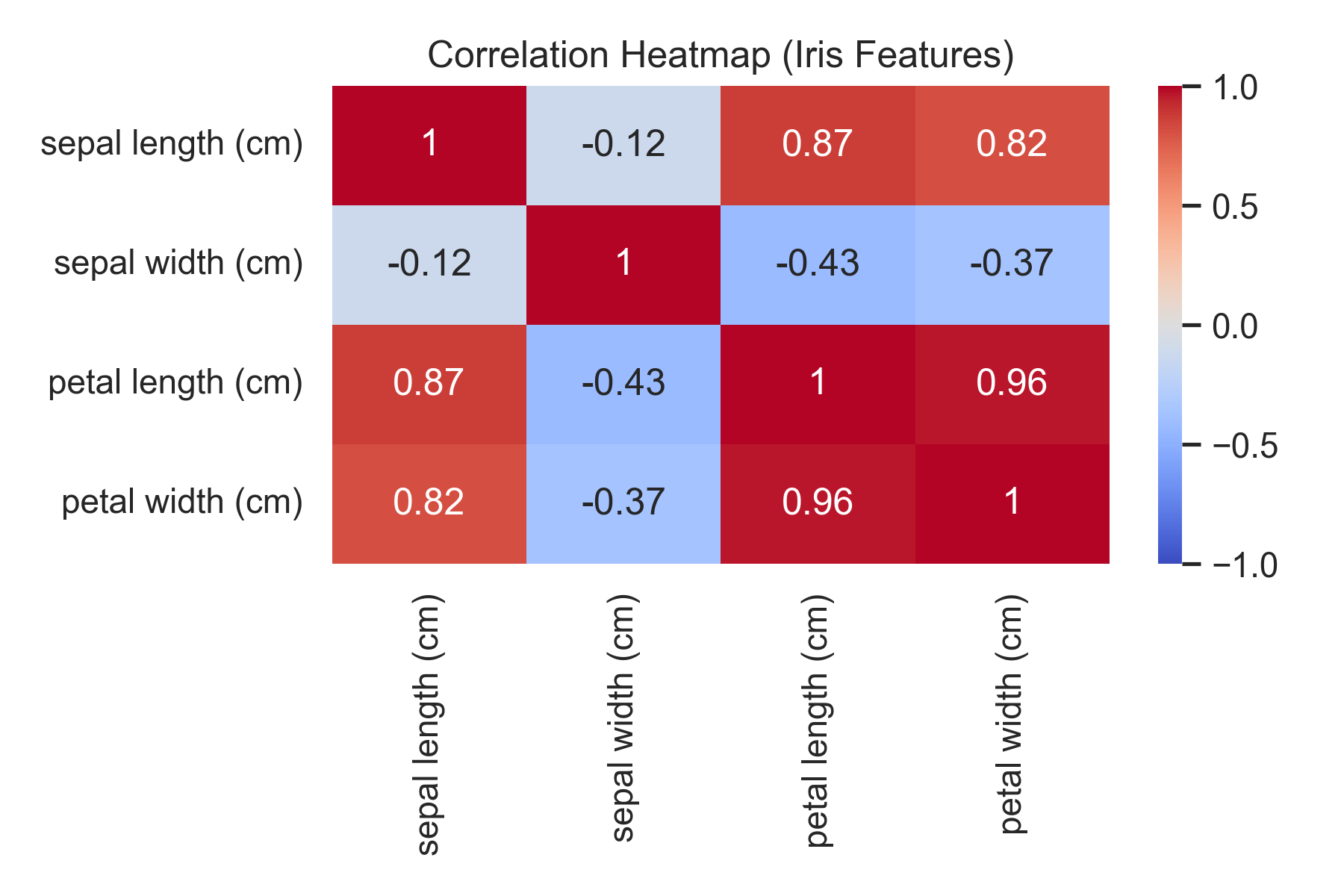
**

*Question 8: Create a scatterplot matrix (or ggpairs). Which variables are most related?*

**The scatterplot matrix reveals that petal length and width are the most closely associated variables, with a clear positive linear association. Sepal length and width are less linked and more distributed. Petal characteristics also enable the best species distinction, with setosa forming a separate cluster from versicolor and virginica.**

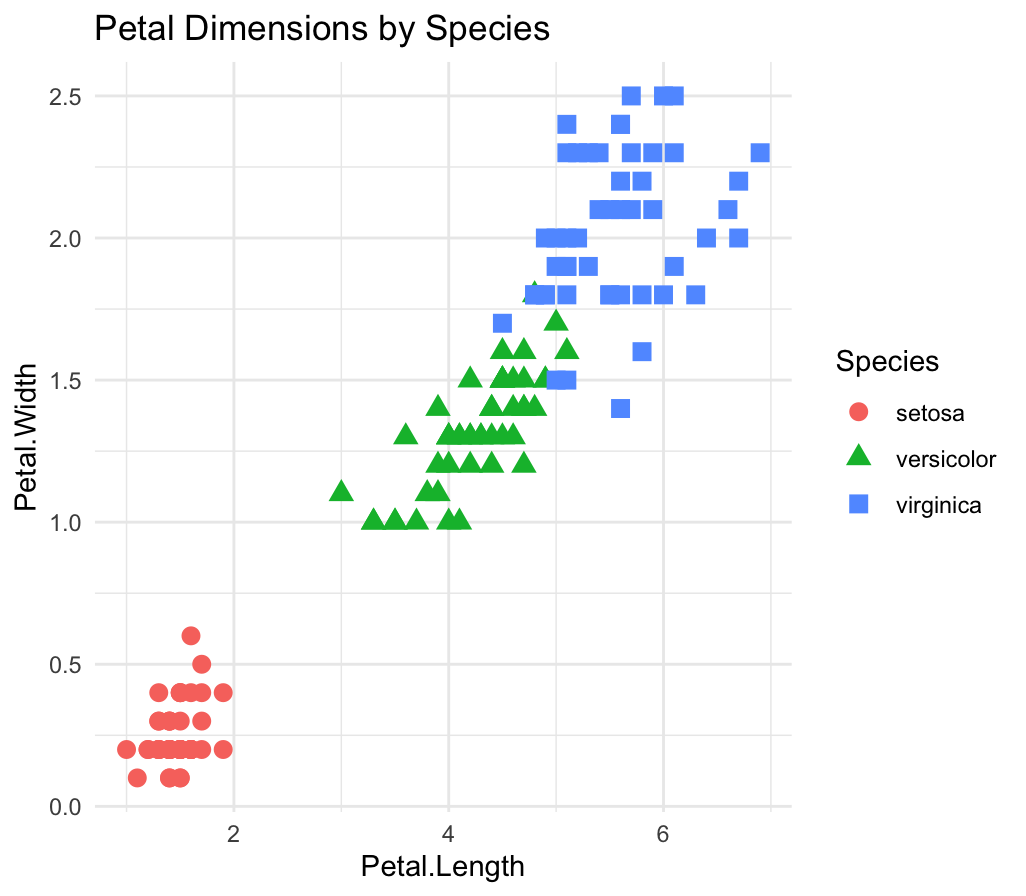
**

*Question 9: Compute and visualize correlations as a heatmap.*

**

*Question 10: Which species shows the greatest variability in petal dimensions?*

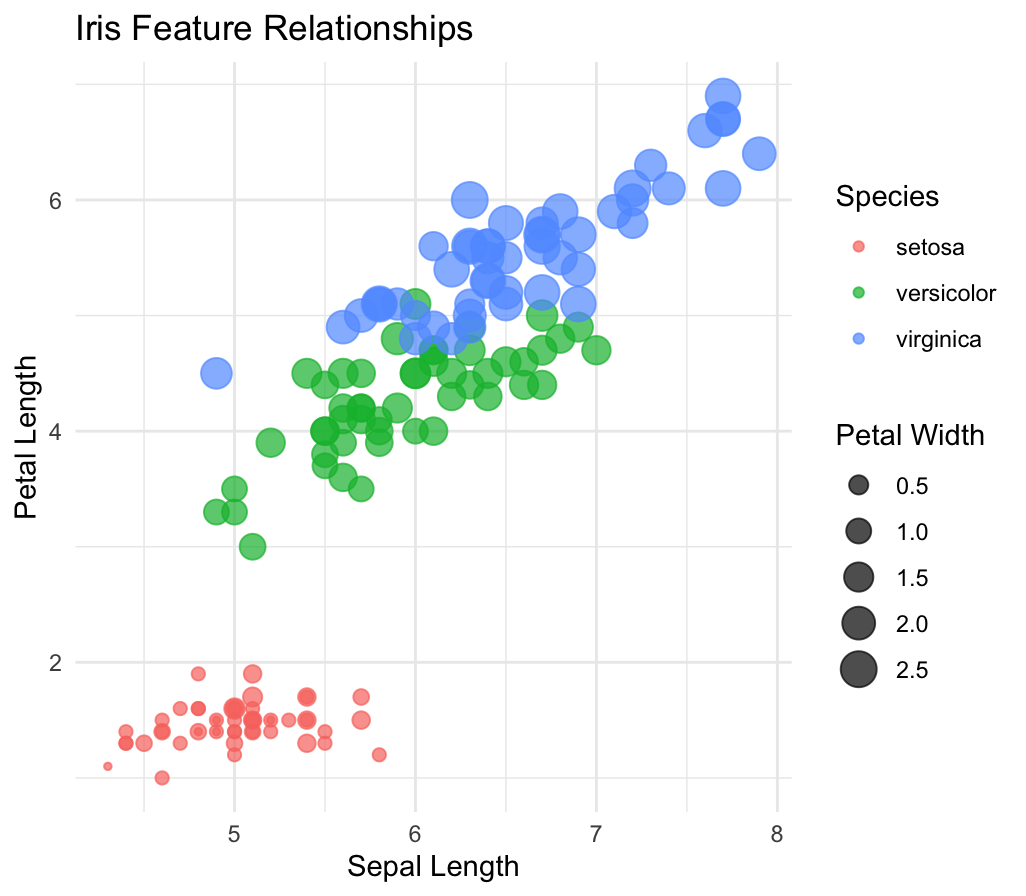
The species that has the greatest variability in petal dimensions would be virginica.

**

*Question 11: Compute and plot the ratio Petal.Length / Petal.Width by species.*

**

*Question 12: Identify any relationships between sepal and petal measurements.*

**

*Question 13: Discuss which features seem most useful for classifying species (qualitatively).*

**The two features that seem the most important for classification of the two different species are the petal length and the Sepal length of the iris.**

*Question 14: Summarize your main insights in at least 1–2 paragraphs.*

**My main insights on this lab having very little experience with using Python is that it was a very good introductory assignment as it was challenging at points, but not to the point where it was impossible to figure out. I look forward to learning more about python and progressing further into more challenging labs and projects. Probably the coolest thing about the lab was the different plots and heatmaps and using them to interpret the data.**

*Question 15: What did you learn most with this lab (1-2 paragraphs maximum)*

**What we learned most with this lab is how to get a basic understanding of Python, by loading datasets and analyzing information from them. It was interesting how using different scatter plots or pair plots can give us different patterns in the species of iris. Also I feel as well this lab gave us a basic knowledge of machine learning. Lastly, getting practice with such tools as pandas, matplotlib, and seaborn will prove to very useful going forward with this course.**

1. <https://en.wikipedia.org/wiki/Petal#/media/File:Petal-sepal.jpg> [↑](#footnote-ref-0)
2. <https://www.datacamp.com/community/tutorials/machine-learning-in-r> [↑](#footnote-ref-1)