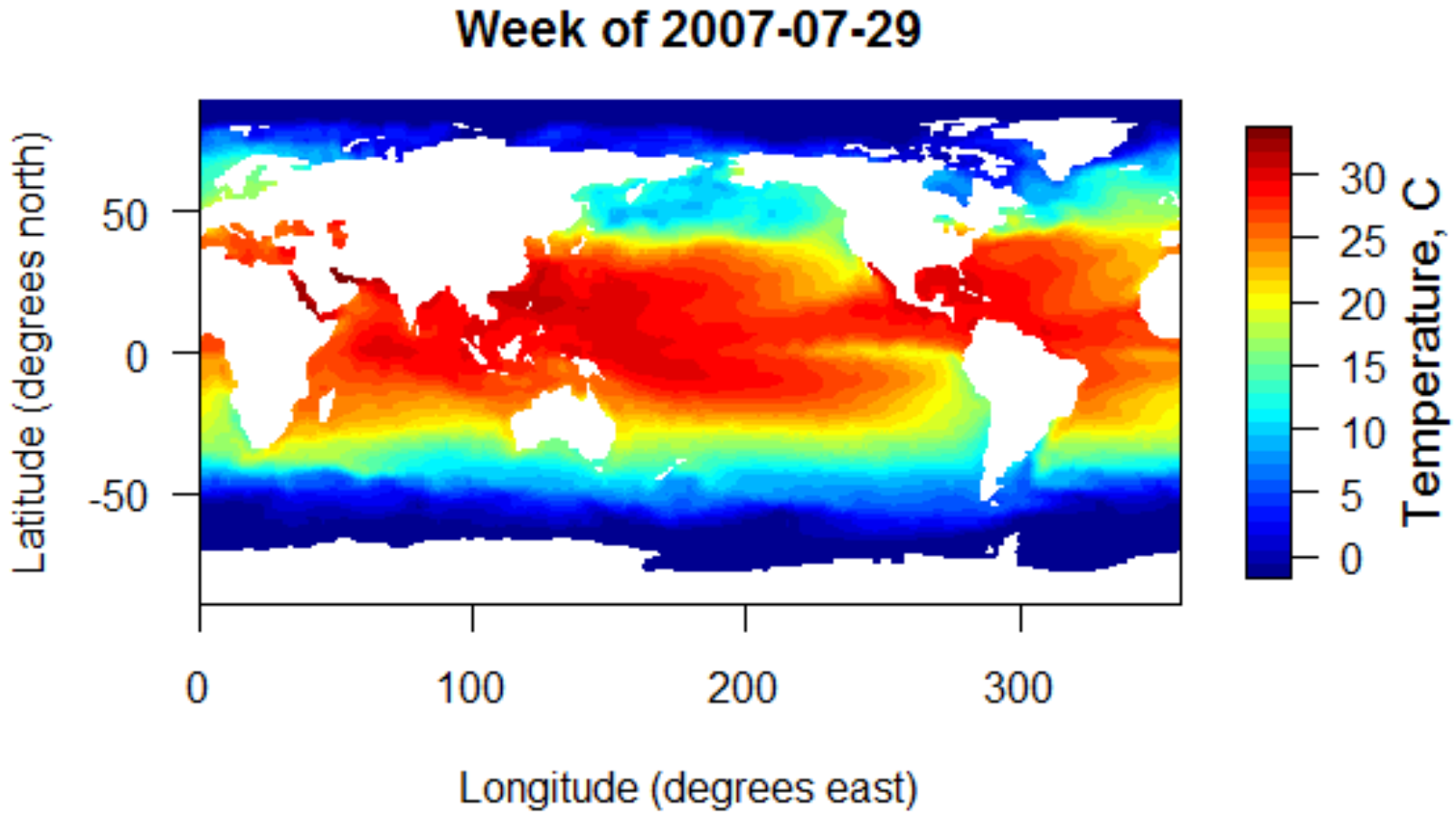
2.6 Data Visualization

- Data Visualization is to present the data in a **visual** or **tabular** format.
- Humans have a well developed ability to analyze visual information.
- By visualization:
  - **Simplify** the complex **quantitative information**.
  - Identify the **relationship** between **datapoints and variables**.
  - Explore **new patterns** and **hidden patterns**.

Example of Data Visualization

- Sea Surface Temperature



General Concepts

- Representation
- Arrangement
- Selection

Representation

- Representation: Mapping Data to Graphical Elements
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - Explicit representation of relationships: graphical elements such as nodes and links
  - Implicit representation of relationships: spatial arrangement or proximity of elements on a plot

Arrangement

- Arrangement: placement of visual elements within a display
- Example: importance of rearranging a table of data

Table 3.5. A table of nine objects (rows) with six binary attributes (columns).

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

Table 3.6. A table of nine objects (rows) with six binary attributes (columns) permuted so that the relationships of the rows and columns are clear.

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Selection

- Selection: elimination or the de-emphasis of certain objects and attributes
- Selection may involve the choosing a **subset of attributes**
  - Consider pairs of attributes
  - Dimensionality reduction: PCA
- Selection may also involve choosing a **subset of objects**
  - Eliminate duplicate or incomplete data
  - Sampling

Data Visualization Demonstration

- Python and Matplotlib
- Iris Dataset
- Plots
  - Histograms
  - Box Plots
  - Pie Charts
  - Scatter Plots
  - Matrix Plots
  - Parallel Coordinates Plots

Python Package -- Matplotlib

Advantages of matplotlib:

- Fast and efficient
- Compatible with various OS
- High-quality graphics and plots
- Full control over graphs and plot styles
- Large community support
- …

Iris Dataset

- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
- From the statistician Douglas Fisher
- Three flower types (classes):
  - Setosa
  - Virginica
  - Versicolour
- Four attributes
  - Sepal width and length
  - Petal width and length

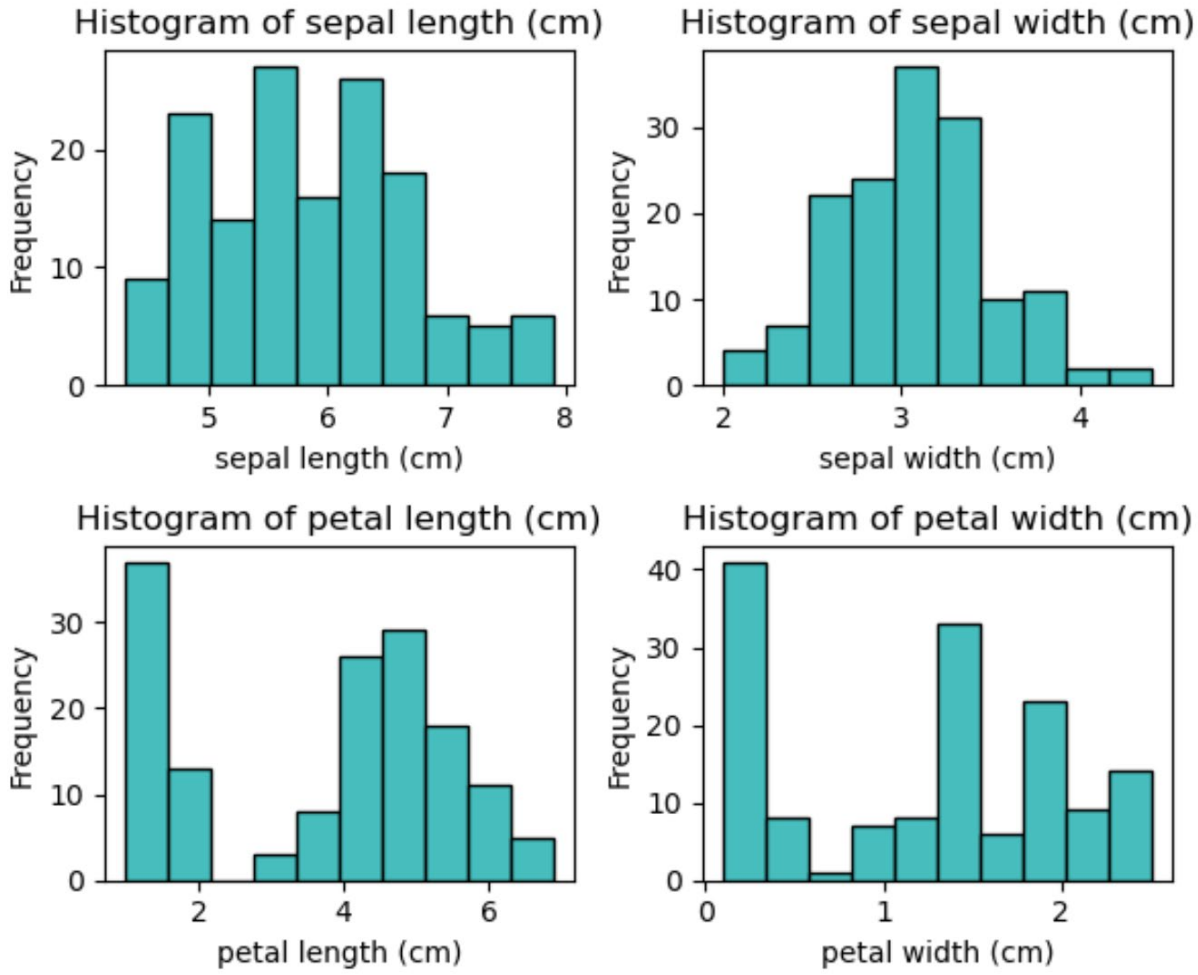
Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Histograms

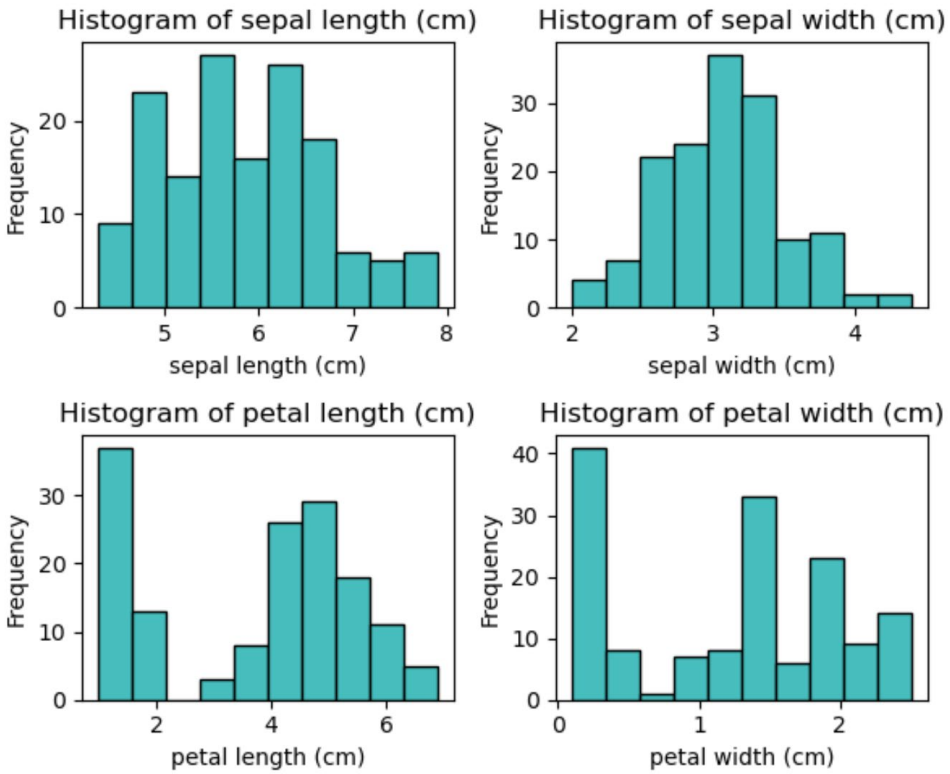
- Usually shows the distribution of values of a **single** variable
- Divide the values into **bins** and show a bar plot of the number of objects in each bin.
- The **height** of each bar indicates the **number** of objects
- Shape of histogram depends on the number of bins

Histograms

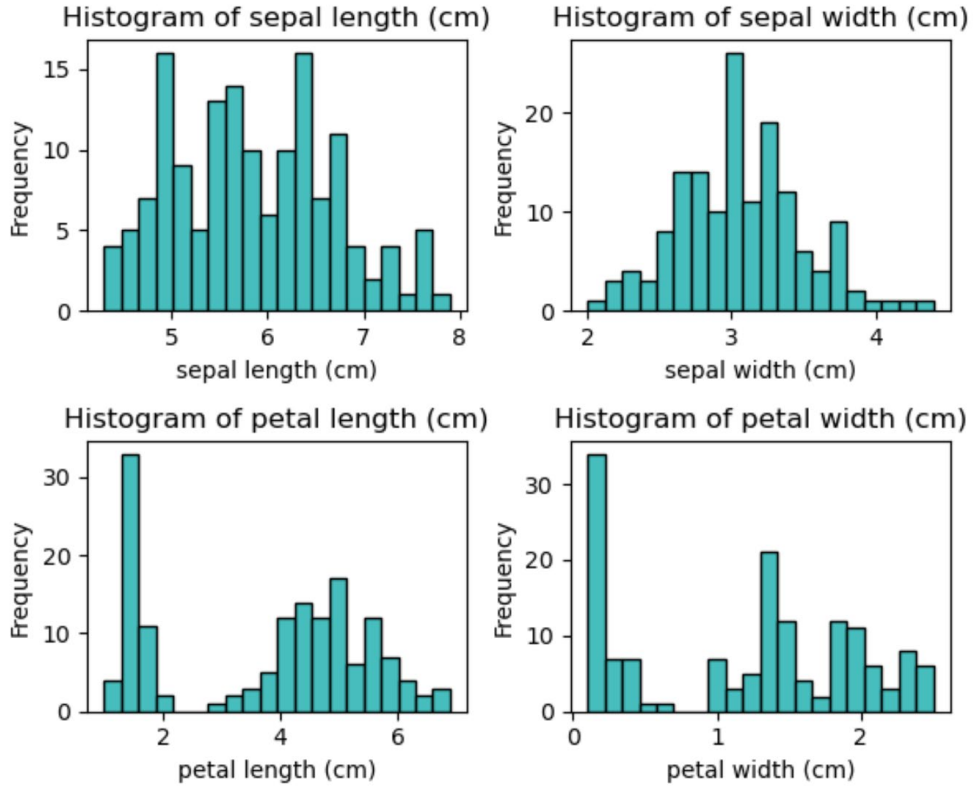

```
# Plot histograms for each feature
for i, ax in enumerate(axs):
    ax.hist(data[:, i], bins=10, color='c',
    ax.set_title(f'Histogram of {feature_na
    ax.set_xlabel(feature_names[i])
    ax.set_ylabel('Frequency')
```



Histograms



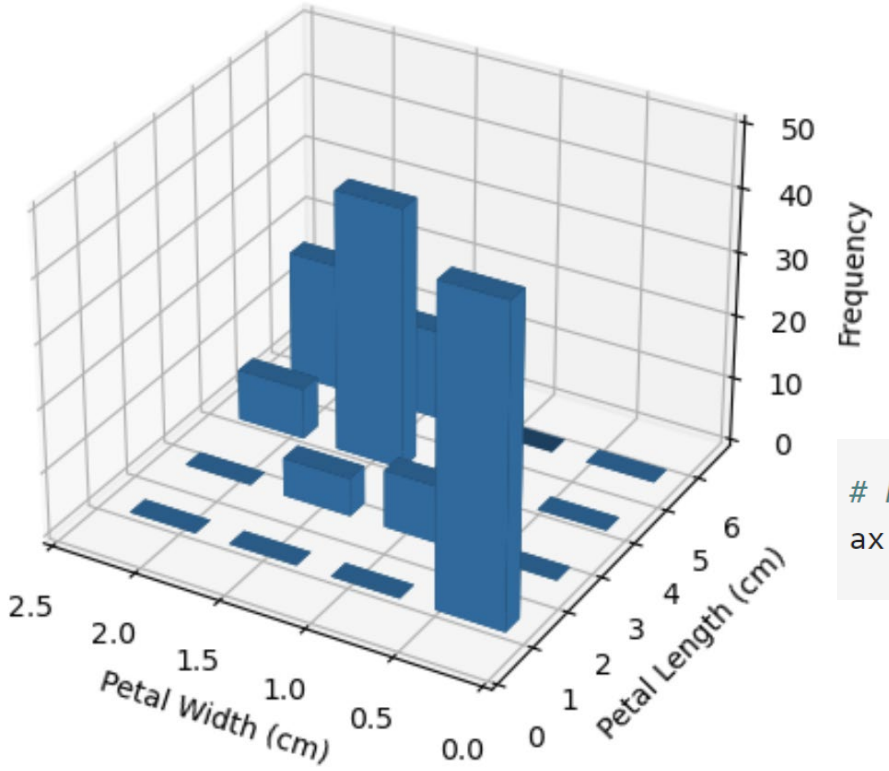
histograms (with 10 bins)



histograms (with 20 bins)

2D Histograms

- Show the **joint distribution** of the values of two attributes.

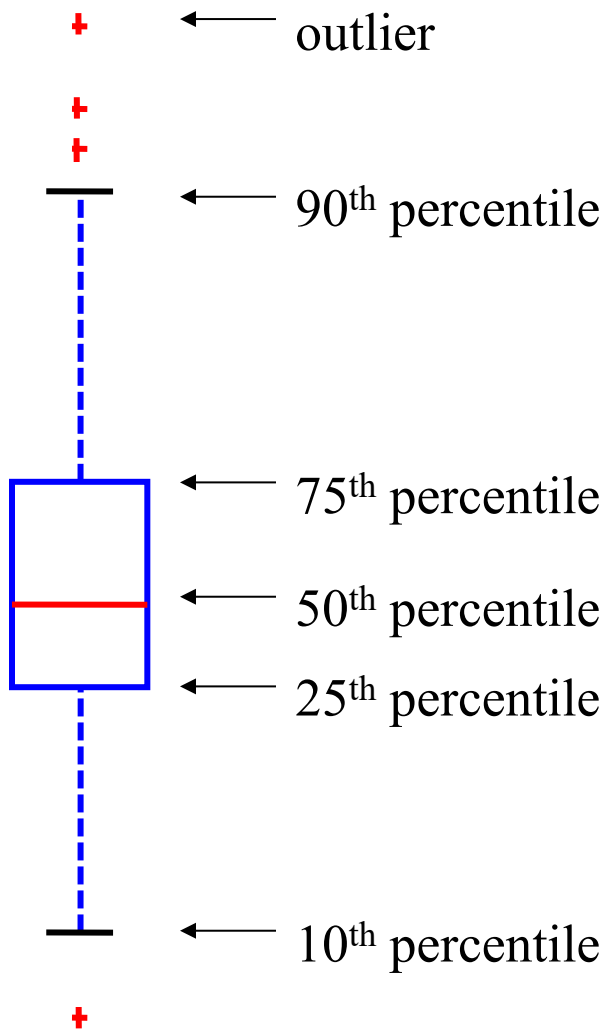



```
np.histogram2d(petal_width, petal_length, bins=[x_bins, y_bins])
```

```
# Plot the bars for petal width (x-axis) and petal length (y-axis)
ax.bar3d(x_pos, y_pos, z_pos, dx, dy, dz, zsort='average', shade=True)
```

Box Plots

- Displays distribution of a **single** variable
- Right figure shows the basic part of a box plot for sepal length.

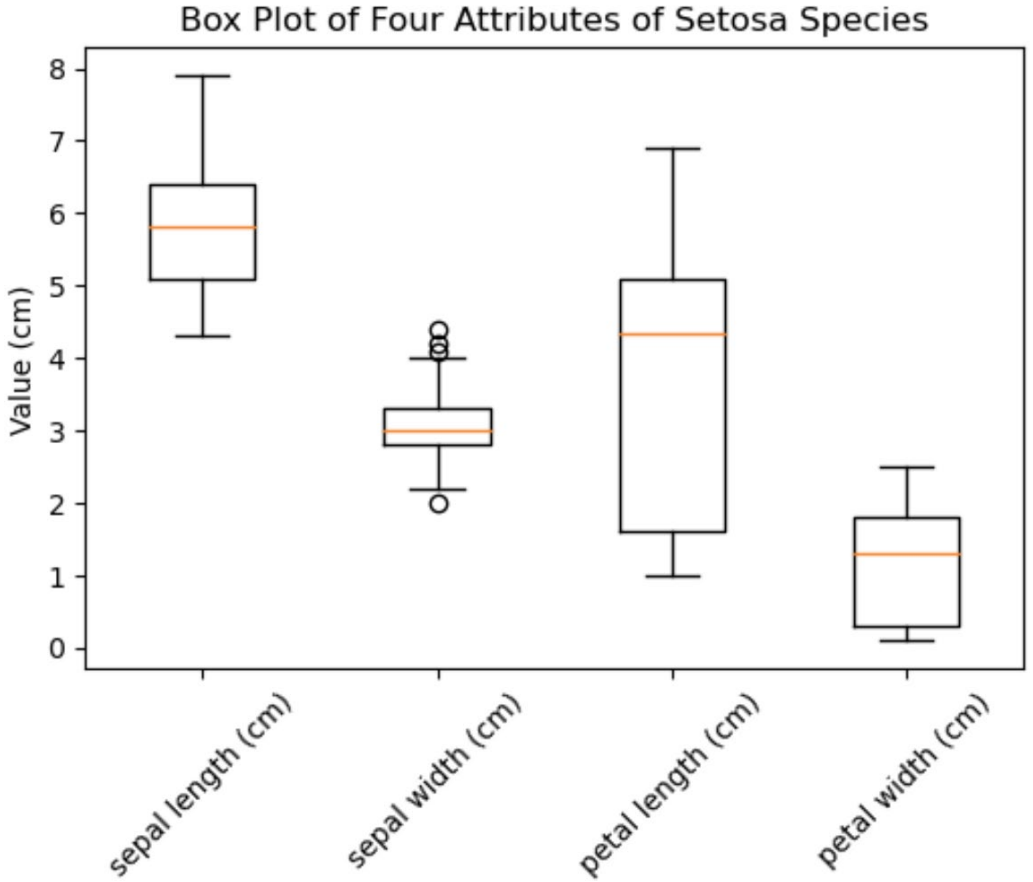


Box Plots

- The box plots for the four attributes of the Iris data set

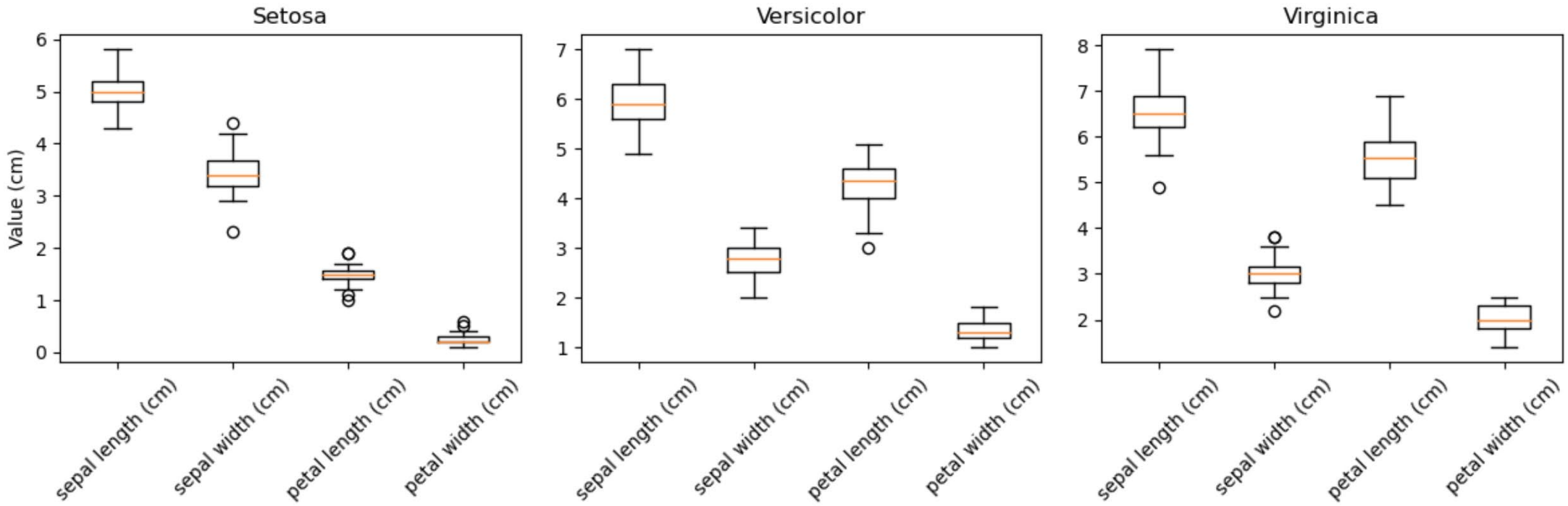
```
# Create a box plot for the four attributes of Setosa
plt.figure(figsize=(6, 4))
plt.boxplot(data, labels=iris.feature_names)

# Add title and labels
plt.title('Box Plot of Four Attributes of Setosa Species')
plt.ylabel('Value (cm)')
plt.xticks(rotation=45)
plt.show()
```



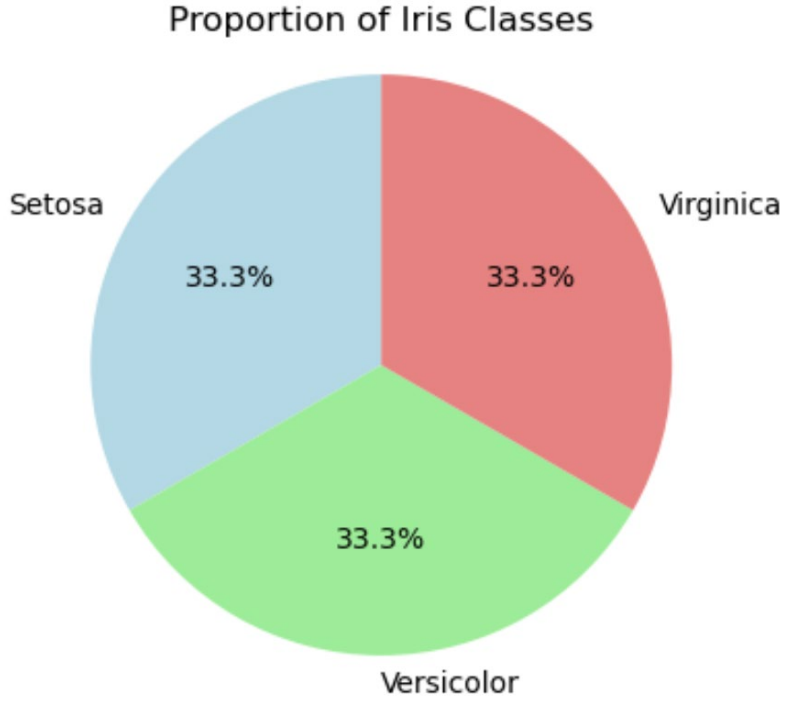
Box Plots

- Compare how attributes vary between different classes



Pie Chart

- Typically used with categorical attributes
- Use relative area of a circle to indicate relative frequency
- Is used **less frequently** in technical publications because the size of relative areas can be **hard to judge**.

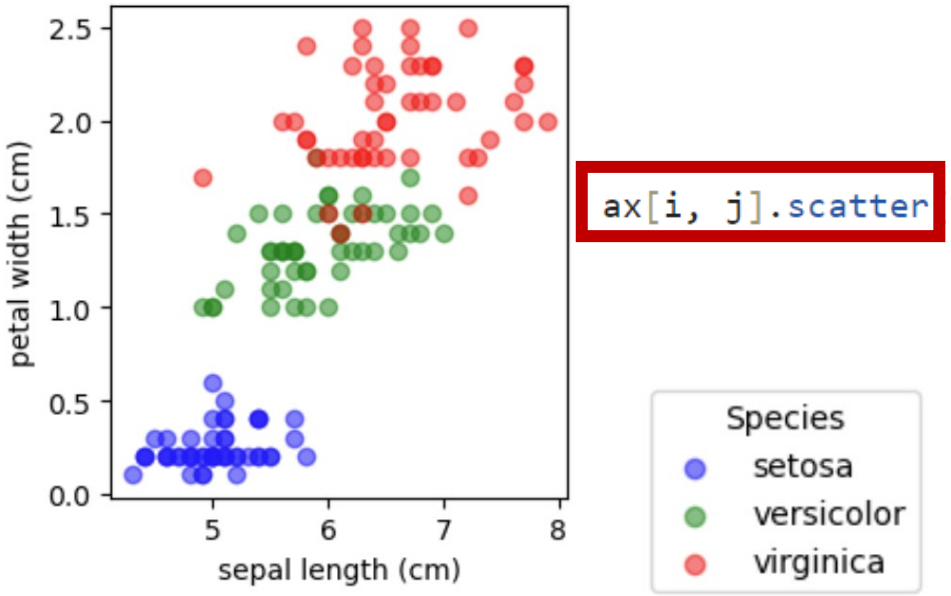


Scatter Plots

- Attributes values determine the position
- 2D scatter plots most common, but can have 3D scatter plots
- Additional attributes can be displayed: size, shape, and color of the markers
- Purpose:
  - Visualize the relationship between two attributes
  - Assess how well two attributes distinguish between classes (with class labels)

Scatter Plots

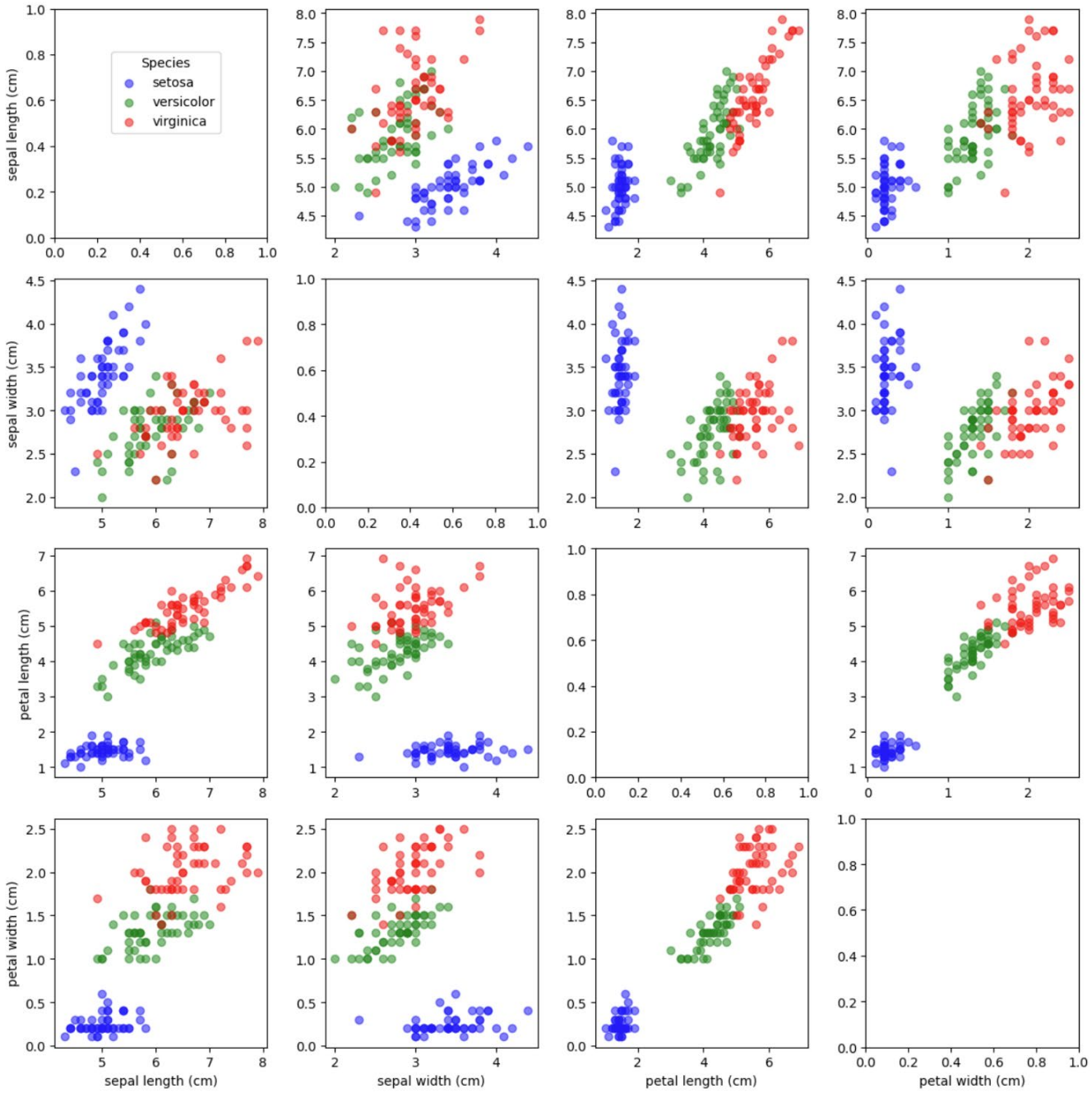
- scatter plot – sepal length vs petal width




```
ax[i, j].scatter data[target == idx, j], data[target == idx, i], c=color,
label=label if i == n_features - 1 and j == 0 else "", alpha=0.5)
```

Scatter Plots

- Scatter plots matrix



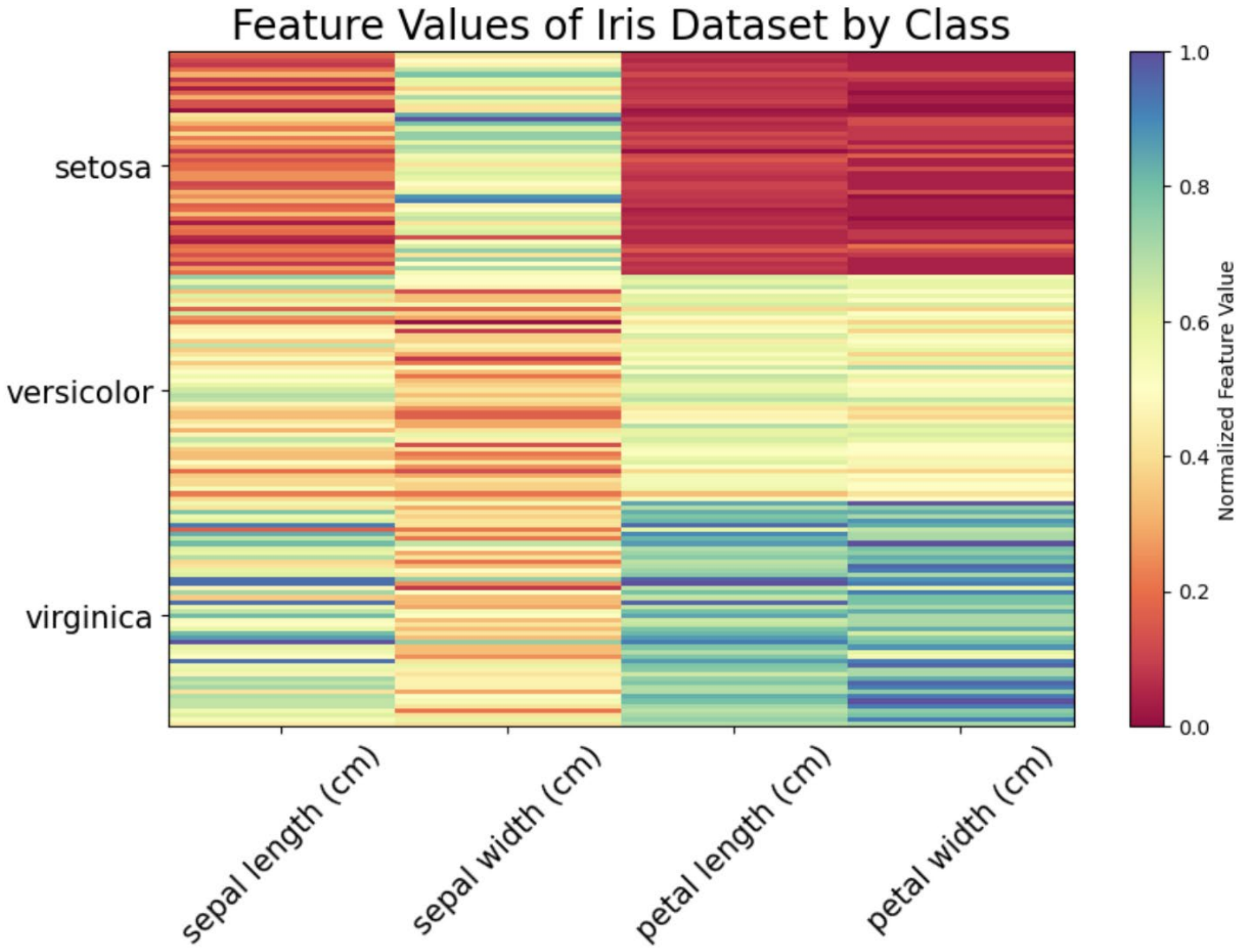
Matrix Plots

- Can plot the data matrix
- This can be useful when objects are sorted according to class
- Typically, the attributes are normalized to prevent one attribute from dominating the plot
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects

Visualization of the Iris Data Matrix

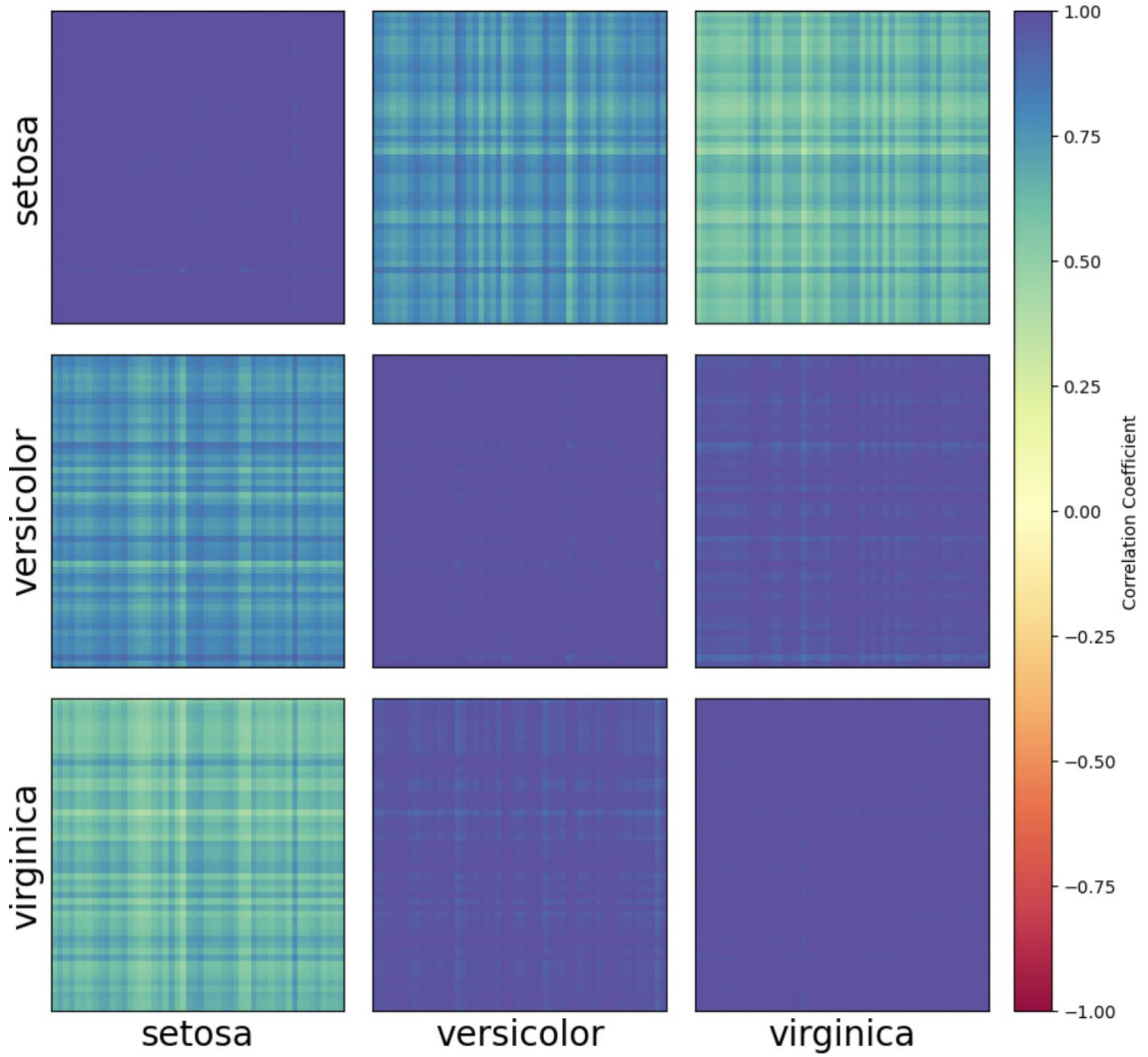
- Each entry of the data matrix

- A pixel in the image



Visualization of the Iris Correlation Matrix

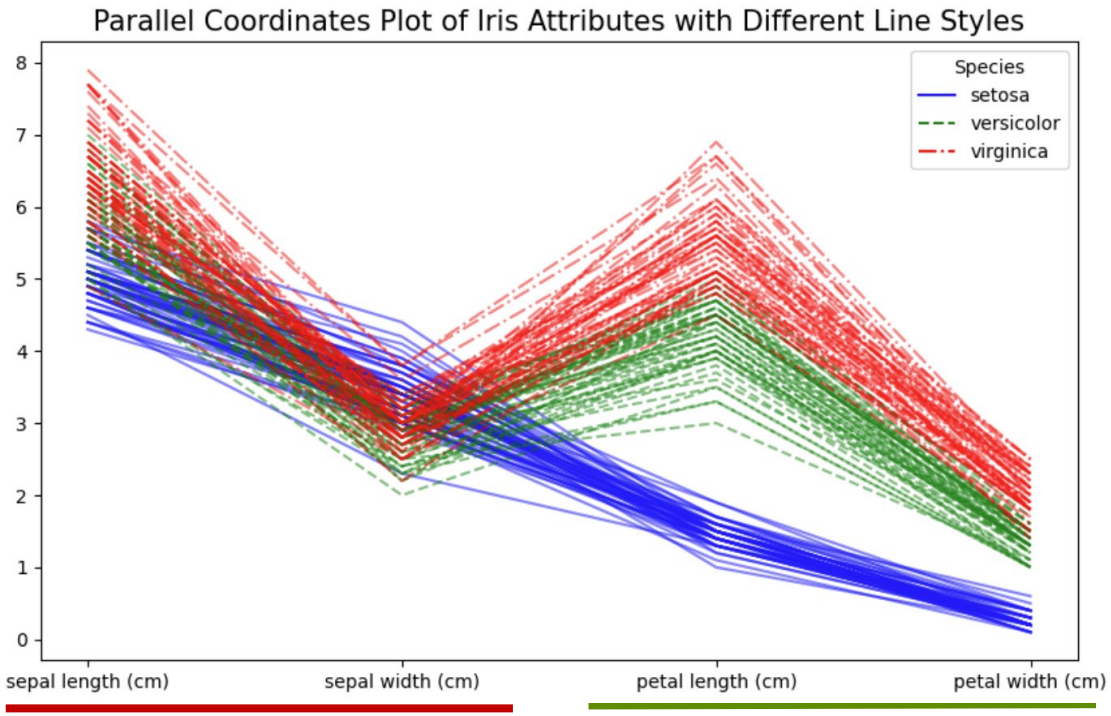
- Pearson correlation between two sample vectors
- Each cell shows the sample-wise correlation between class pairs.



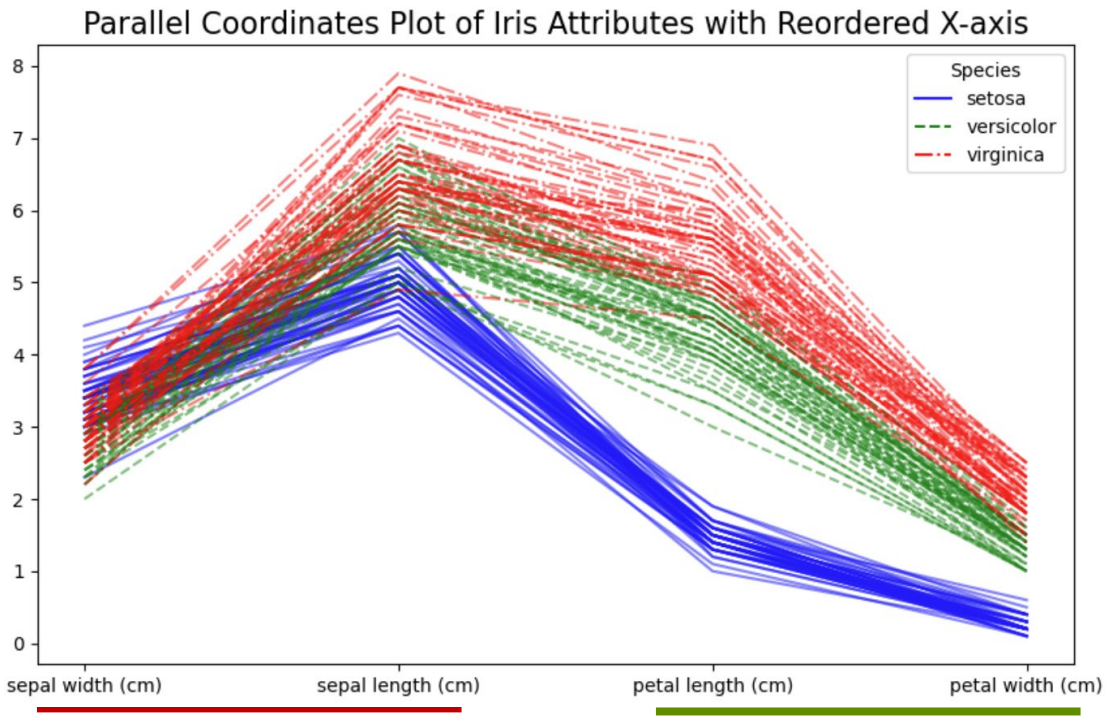
Parallel Coordinates

- Used to plot the attribute values of **high-dimensional** data
- Instead of using perpendicular axes, use a set of **parallel axes**
- The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
- Thus, each object is represented as a line
- Often, the lines representing a distinct class of objects group together, at least for some attributes
- Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data



A parallel coordinates plot of the four Iris attributes.



A parallel coordinates plot with the attributes reordered to emphasize similarities and dissimilarities of groups.