# Creating Data Visualizations

(HW1)

| CWID | Name | Contribution | Percent Contribution |
|---|---|---|---|
| A20563418 | Hanyu Wang | Write reports and analyze data | 50% |
| A20563432 | Zhengcheng Peng | Collecting data and creating charts | 50% |

## Dataset Description

**Origin and Background**

**Link：** Iris Species

The Iris dataset is one of the most classic datasets in machine learning, collected and published by statistician and biologist R.A. Fisher in 1936. The dataset measures four features of three different species of Iris flowers: Setosa, Versicolor, and Virginica.

**Dataset Specifications**

Number of samples: 150 (50 samples per species)

Number of features: 4 numerical features + 1 categorical label

Feature descriptions:

SepalLengthCm: Sepal length in centimeters

SepalWidthCm: Sepal width in centimeters

PetalLengthCm: Petal length in centimeters

PetalWidthCm: Petal width in centimeters

Target variable: Species (Iris-setosa, Iris-versicolor, Iris-virginica)

**Preprocessing Steps**

The preprocessing steps applied to the dataset were minimal:

1. Data validation: Confirmed there were no missing values in the dataset
2. Species name simplification: Removed the "Iris-" prefix from species names for cleaner visualization
3. Data integrity check: Verified there were no outliers that required special handling

The dataset is well-balanced with exactly 50 samples for each species, making it excellent for comparative analysis and a benchmark for classification algorithms.

# Visualization Methods

**1.Bar Chart: Species Distribution**

**Visualization Type:** Bar Chart

**Libraries Used:** Matplotlib, Seaborn

**Creation Method:**

Used Seaborn's barplot function to create a categorical bar chart;

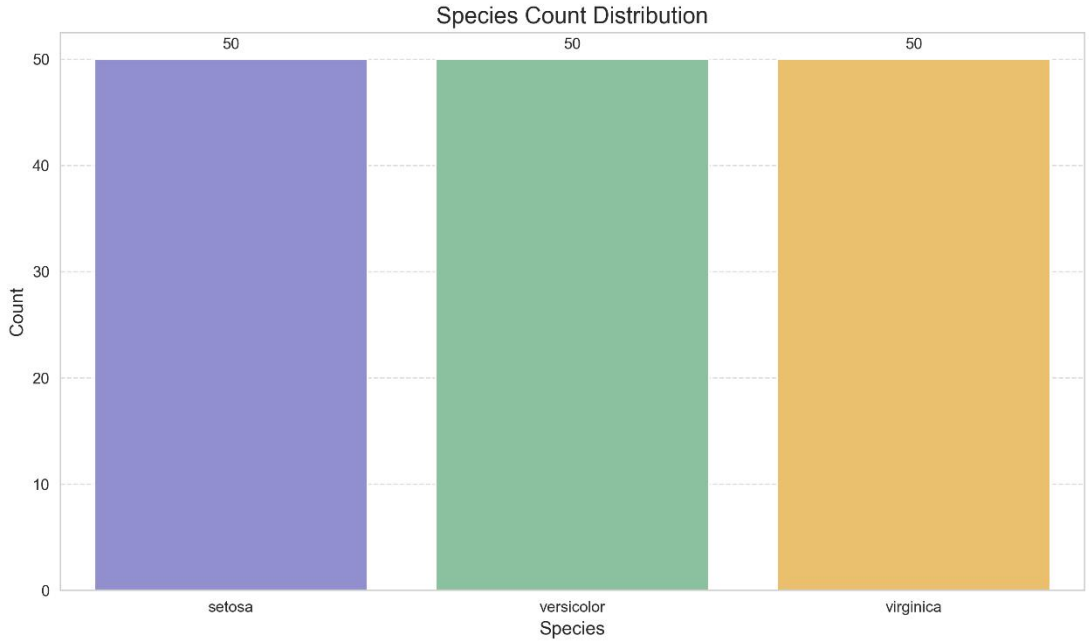
Applied a custom color palette for visual distinction between species;

Added numerical labels above each bar for exact count values;

Implemented grid lines for better readability.

**Analysis of results：**

The bars clearly show the balance of the dataset, with 50 samples for each species (Setosa, Versicolor, Virginica). This balance makes the dataset ideal for categorization tasks, as each category has an equal number of samples, avoiding the problem of category imbalance.



**2．Scatter Plot: Petal Length vs. Width Relationship**

**Visualization Type:** Scatter Plot

**Libraries Used:** Matplotlib

**Creation Method:**

Created a scatter plot using Matplotlib's scatter function;

Separated data points by species using different colors;

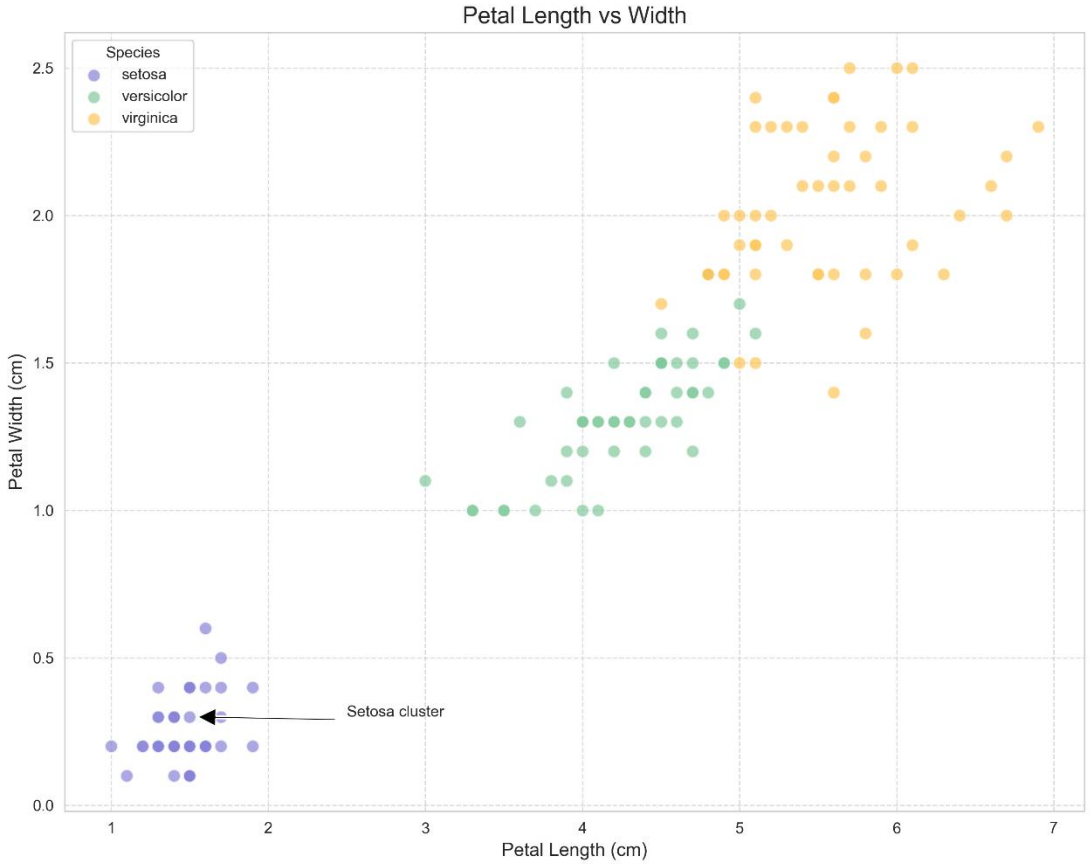
Customized point appearance with size, alpha transparency, and white edges for better visibility;

Added annotations to highlight the Setosa cluster's clear separation

Included a legend to identify each species.

**Analysis of results：**

The scatterplot clearly shows that the petal length and width of the Setosa species are significantly smaller than those of the other two species (Versicolor and Virginica.) Most of the petals of Setosa are less than 2 cm in length and less than 0.6 cm in width, forming a completely separate cluster. This apparent separation makes Setosa easy to distinguish in taxonomic tasks.



**3．Box Plot: Sepal Length Distribution Comparison**

**Visualization Type**: Box Plot with Strip Plot overlay

**Libraries Used**: Seaborn

**Creation Method**:

Used Seaborn's boxplot function to create the main box plot

Added a stripplot overlay to show the actual data points

Applied consistent color coding for species identification

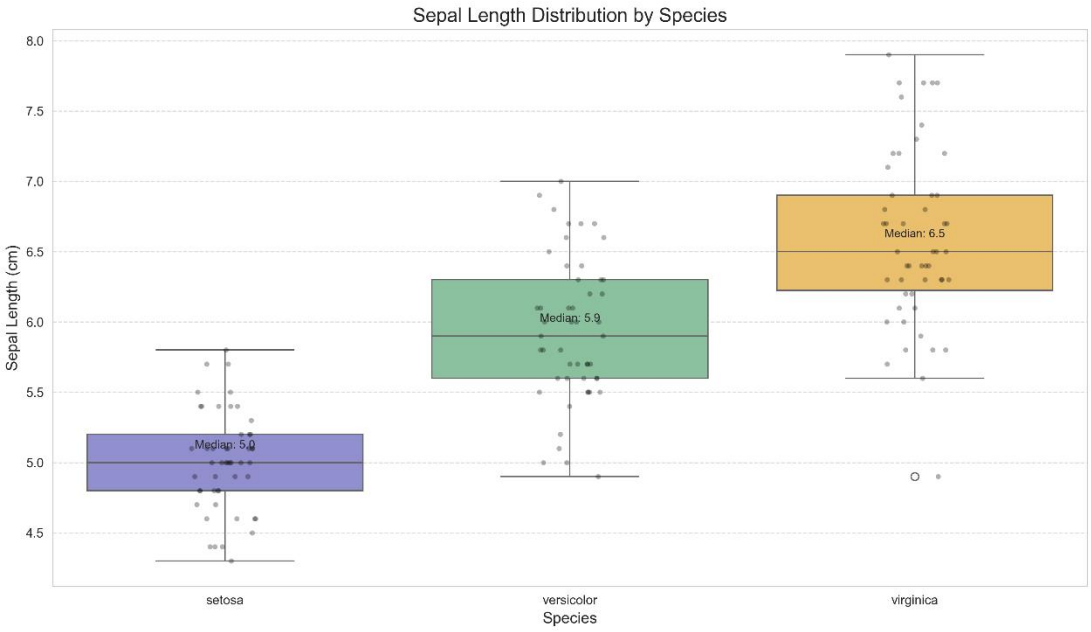
Added median labels for each species

Included grid lines for easier measurement comparison

**Analysis of results：**

The box plot shows that Setosa has the smallest median sepal length, Versicolor the next largest, and Virginica the largest. This trend suggests that sepal length varies significantly among species, especially between Setosa and Virginica;

The narrower range of sepal length distribution in Setosa suggests that sepal lengths are relatively uniform in Setosa. In contrast, Versicolor and Virginica had a wider distribution of sepal lengths, indicating that sepal lengths are more variable in these two species.



**4．Heatmap: Feature Correlation Analysis**

**Visualization Type:** Correlation Heatmap

**Libraries Used:** Seaborn, NumPy

**Creation Method:**

Calculated the correlation matrix using pandas' corr() method

Created a mask to show only the lower triangle of the matrix (avoiding redundancy)

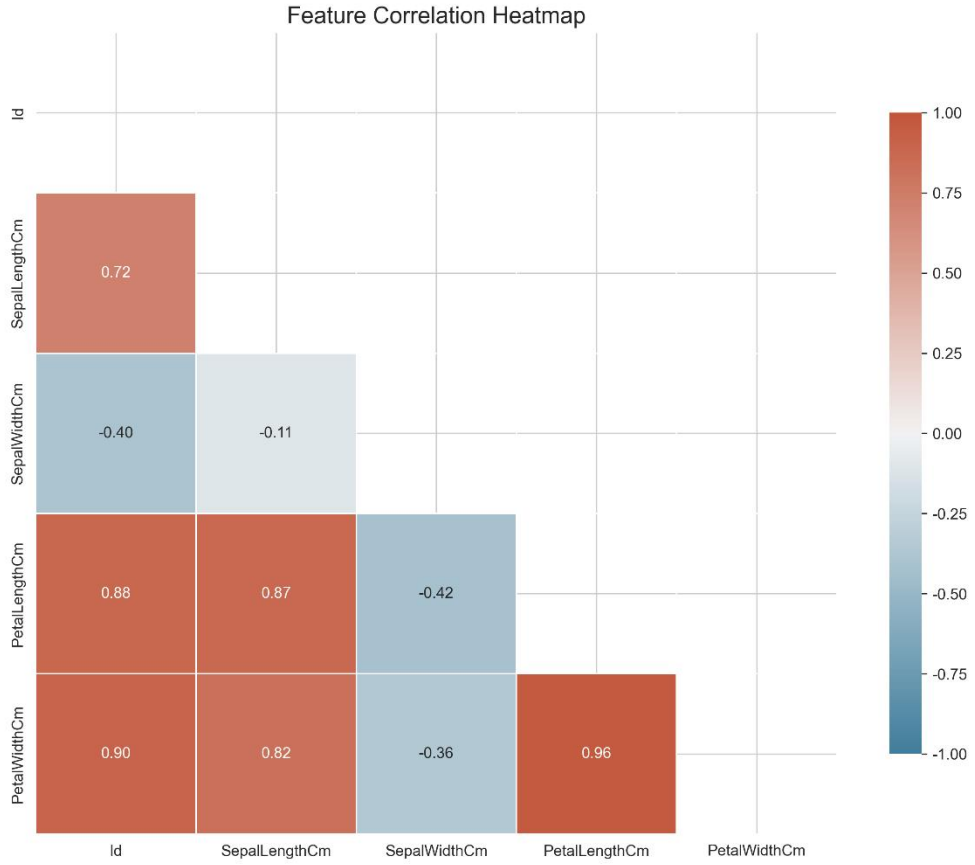
Used Seaborn's heatmap function with a diverging color palette

Added numerical annotations for exact correlation values

Applied custom styling for better readability

**Analysis of results：**

The heatmap shows the strongest correlation between petal length and petal width (r = 0.96), indicating that these two features almost always increase or decrease simultaneously. This strong correlation implies that only one of the features may be needed in a classification task, as both features provide almost the same information.



# Key Insights from Visualizations

**Species Distribution (Bar Chart)**

The dataset is perfectly balanced, with exactly 50 samples for each of the three species (Setosa, Versicolor, and Virginica)

**Petal Characteristics (Scatter Plot)**

Setosa Distinctiveness: Iris Setosa forms a completely separate cluster from the other two species, with significantly smaller petals (length < 2cm, width < 0.6cm)

Linear Relationship: Within each species, petal length and width show a strong positive correlation, suggesting these features change proportionally

**Sepal Length Variation (Box Plot)**

The median sepal length increases from Setosa to Versicolor to Virginica

**Feature Correlations (Heatmap)**

Petal length and petal width show the strongest positive correlation (r=0.96), indicating they almost always increase or decrease together

Sepal length correlates strongly with both petal length (r=0.87) and petal width (r=0.82)

Sepal width shows negative correlations with all other features, most notably with petal length (r=-0.42)