Exploratory Data Analysis (EDA) – Iris Dataset

The *Iris Dataset* (also known as Fisher's Iris Dataset) is a classic and widely used dataset in data analysis and machine learning. It includes measurements of four morphological features for three species of iris flowers: Setosa, Versicolor, and Virginica.

Objective: To assess whether it is possible to distinguish iris species based on their morphological features — in particular, the dimensions of petals and sepals.

Basic Information

The dataset contains **150** rows and **5** columns

	sepal_length	sepal_width	petal_length	petal_width	species
82	5.800000	2.700000	3.900000	1.200000	versicolor
133	6.300000	2.800000	5.100000	1.500000	virginica
113	5.700000	2.500000	5.000000	2.000000	virginica
93	5.000000	2.300000	3.300000	1.000000	versicolor
5	5.400000	3.900000	1.700000	0.400000	setosa
114	5.800000	2.800000	5.100000	2.400000	virginica
128	6.400000	2.800000	5.600000	2.100000	virginica
119	6.000000	2.200000	5.000000	1.500000	virginica

The most important column in the dataset is `species`, which contains labels for **3** iris species:

- **setosa**: 50 samples
- **versicolor**: 50 samples
- **virginica**: 50 samples

For each flower, the dataset includes:

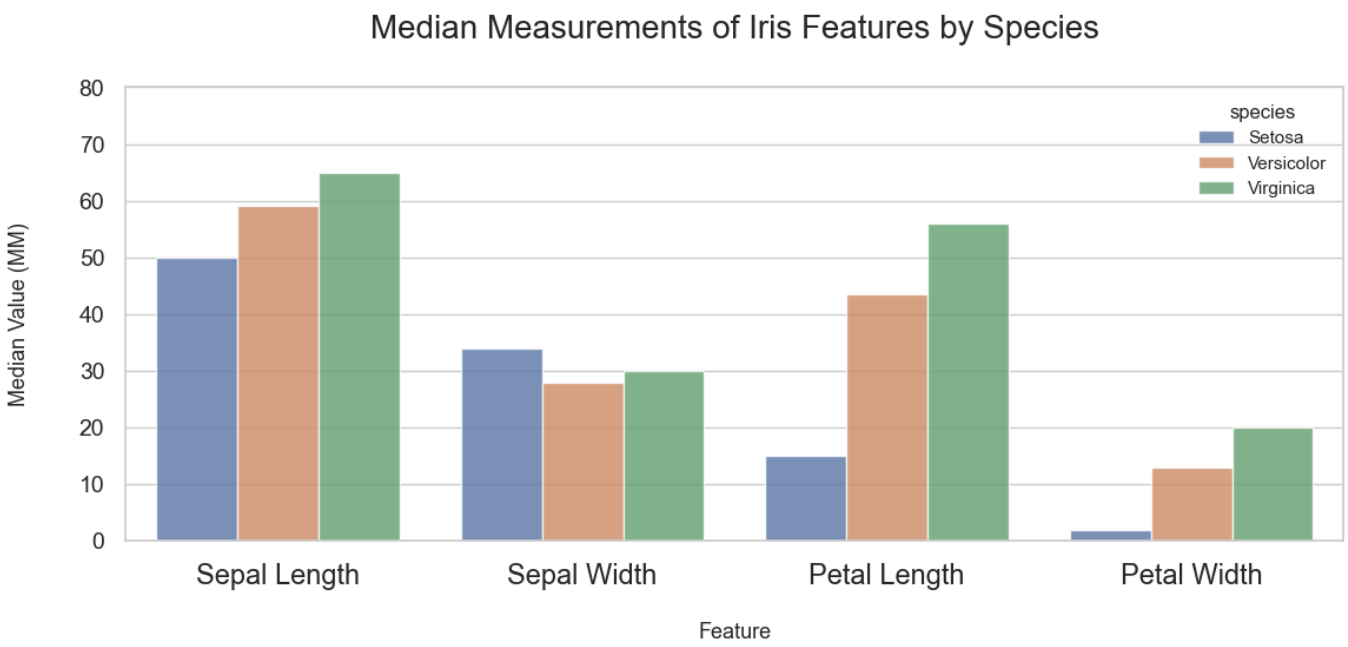
- **Sepal length** (in mm)
- **Sepal width** (in mm)
- **Petal length** (in mm)
- **Petal width** (in mm)

There are **0** missing values in the dataset

Visual Analysis

Feature Medians by Species – Chart 1

This chart presents the median values of all four features (sepal length, sepal width, petal length, and petal width) for each species. The median is the middle value — it represents a typical measurement for that species.

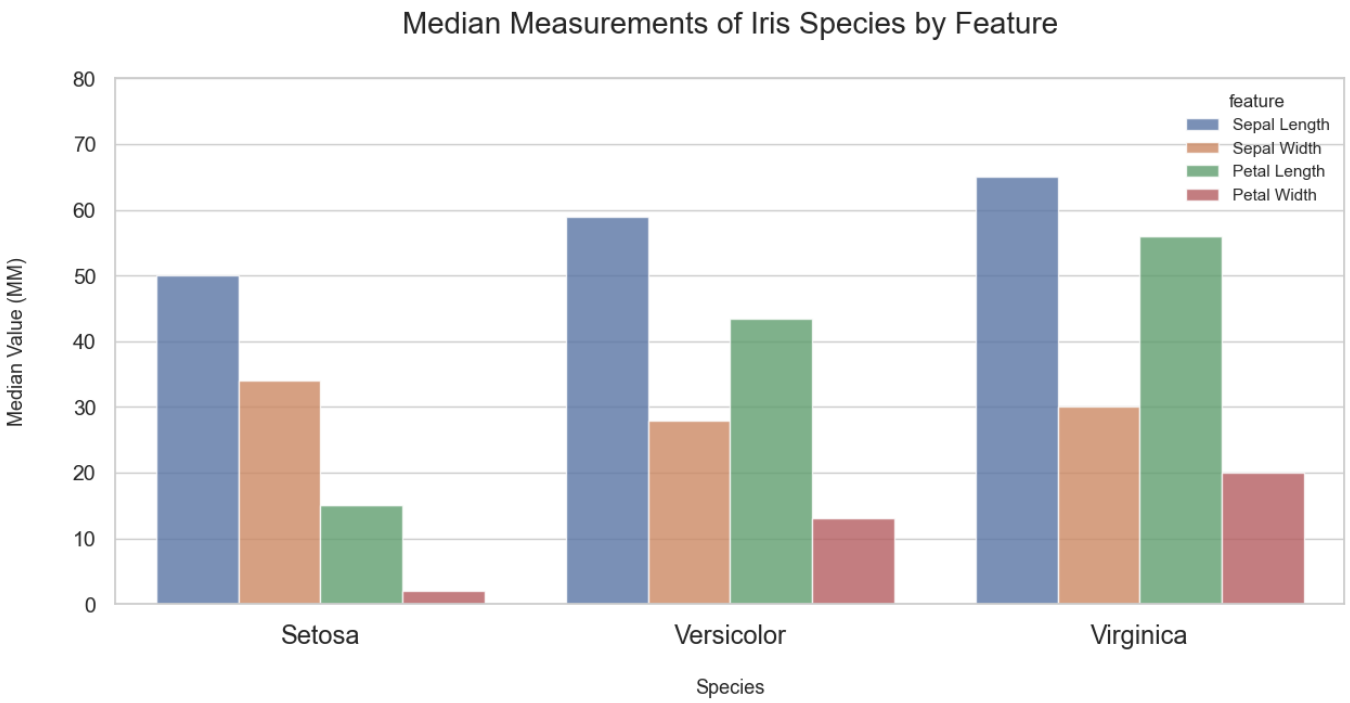


Insights:

- *Iris Setosa* has much shorter and narrower petals than the other two species.
- *Iris Virginica* has the longest petals and sepals overall.
- *Iris Versicolor* lies in between — acting as a transition form.
- Even at this stage, petal length appears to be a strong distinguishing feature between species.

Feature Medians by Trait – Chart 2 (Rotated View)

This chart shows the same data but from a different perspective — for each feature, it compares medians across all three species.



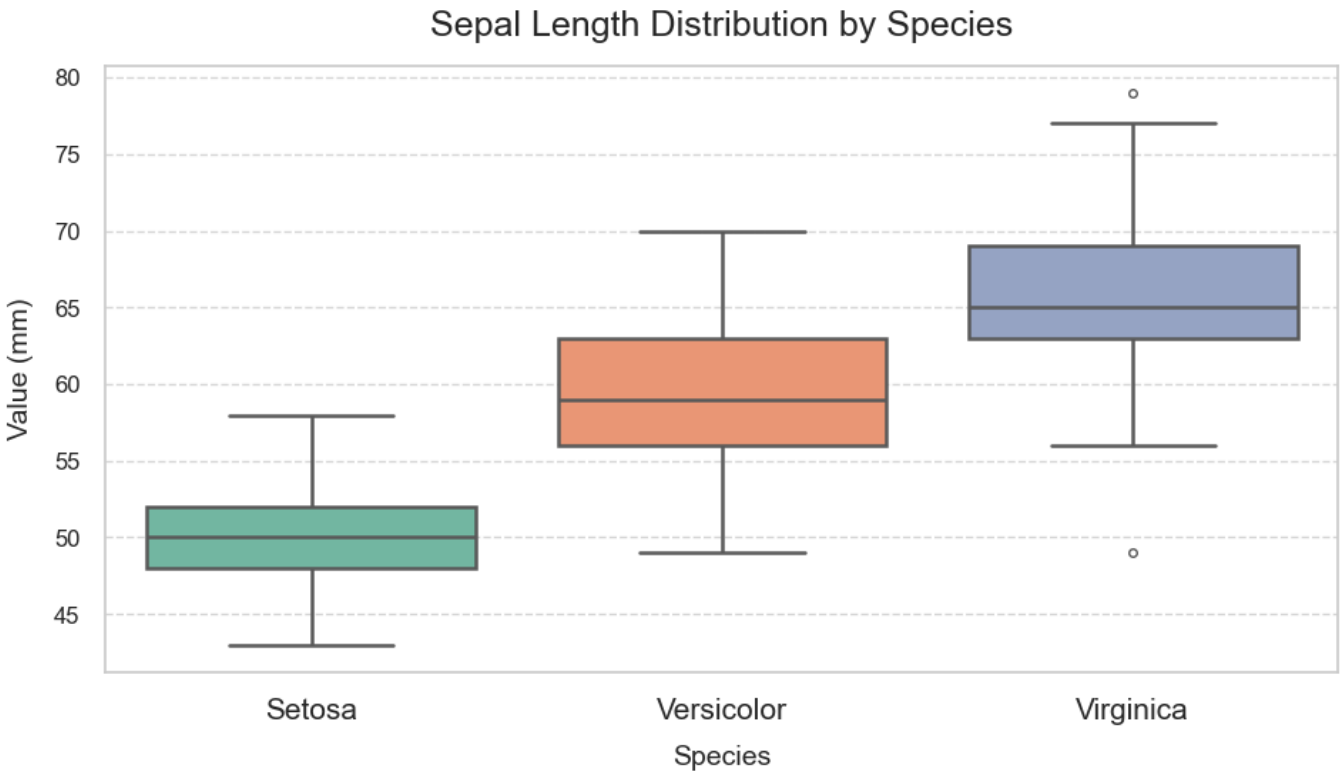
Insights:

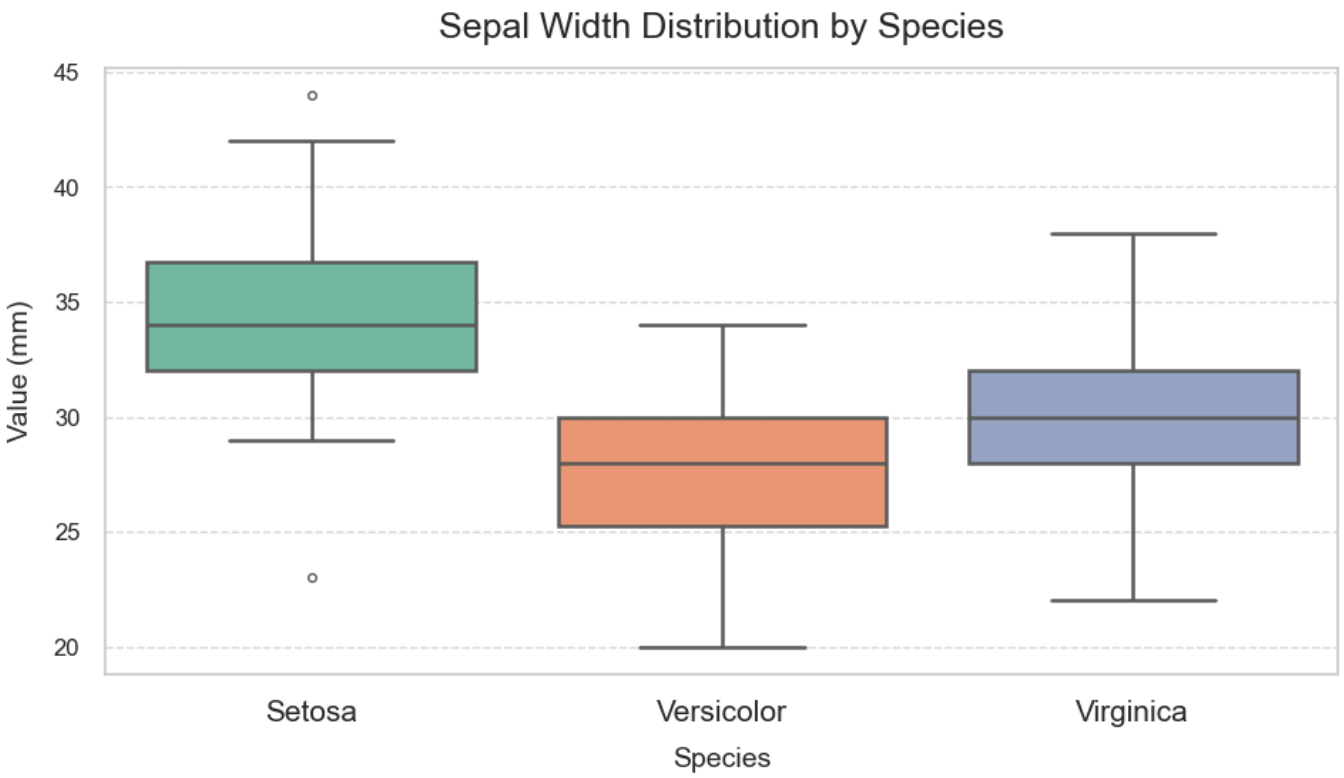
- Petal length and width differ significantly between species.
- Sepal width shows less distinction — especially between Versicolor and Virginica.

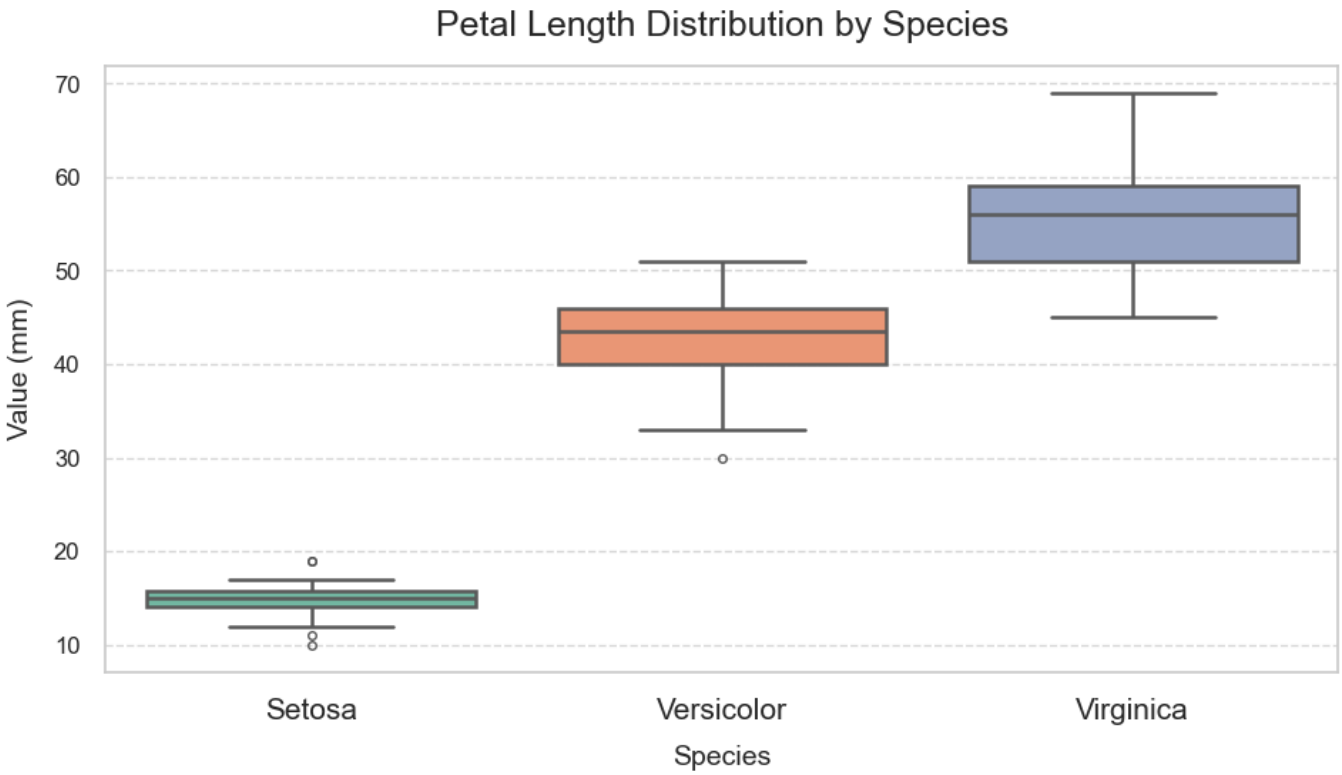
Feature Distributions and Outlier Analysis

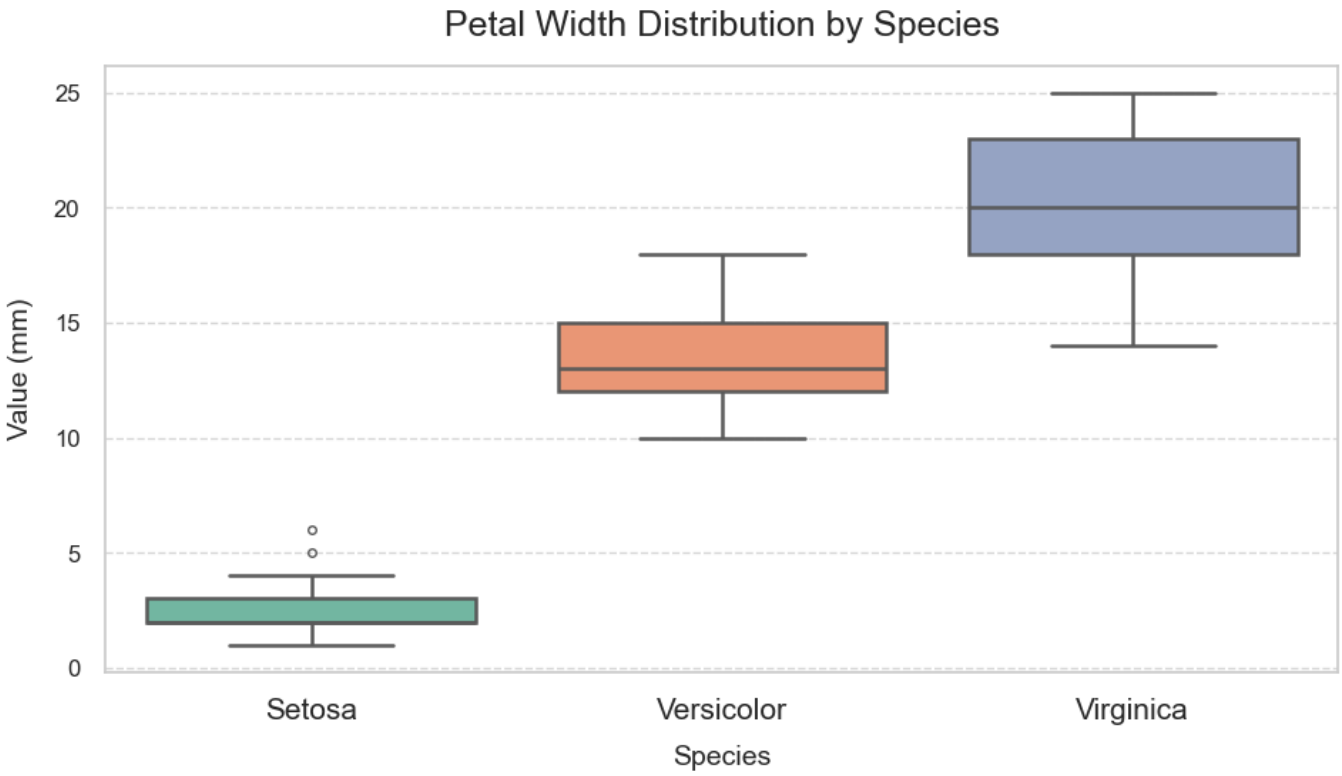
Boxplots – Feature Distribution by Species

Each boxplot shows how a specific feature is distributed across the three species. The line in the middle of the box indicates the median; the box spans typical values, and dots represent potential outliers.


Sepal Length Distribution by Species
Value (mm)
80
75
70
65
60
55
50
45
Setosa
Versicolor
Virginica
Species


Sepal Width Distribution by Species
Value (mm)
45
40
35
30
25
20
Setosa
Versicolor
Virginica
Species





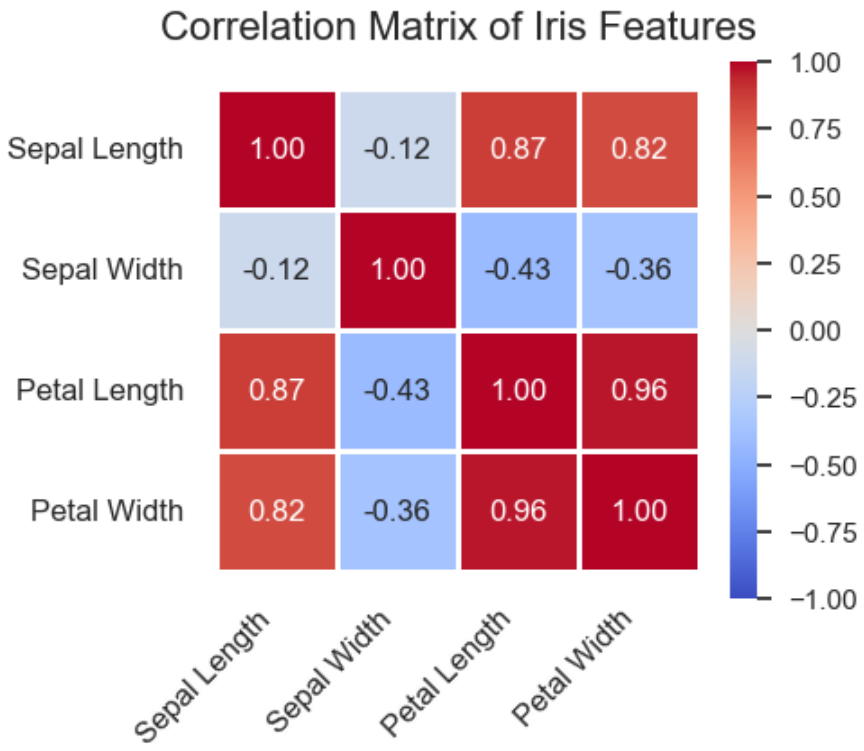
Insights:

- Petal length and petal width are highly discriminative — Setosa's values do not overlap with the other two species.
- Sepal width shows overlapping distributions — it may be less useful for classification.
- A few outliers are present, but overall the data appears clean and well-structured.

Relationships Between Features

Correlation Matrix

A correlation matrix shows how strongly the features are related to one another (values range from -1 to 1).



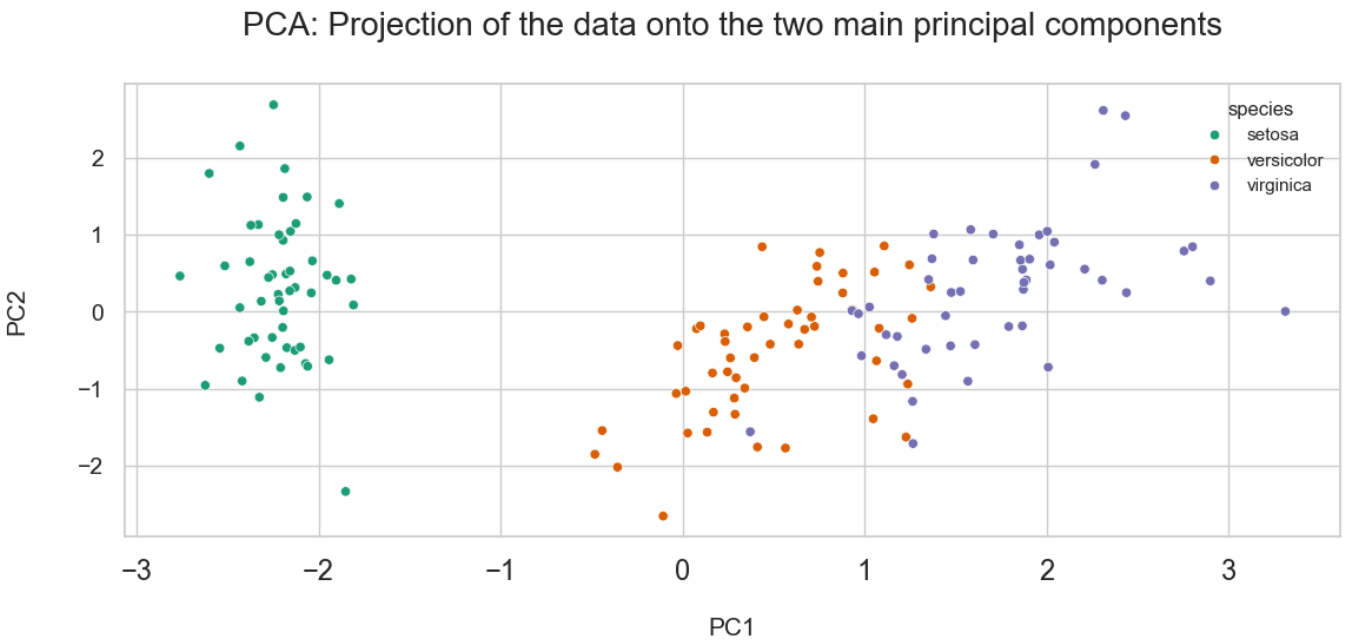
Insights:

- The strongest correlation is between petal length and petal width (`0.96`) — these grow together.
- Sepal length is moderately correlated with petal length (`0.87`), suggesting flowers with longer sepals tend to have longer petals.
- There's also a notable correlation between sepal length and petal width (`0.82`).

Dimensionality Reduction

PCA – Principal Component Analysis

Principal Component Analysis (PCA) reduces the dataset to two dimensions while preserving as much variance as possible. Each point in the plot represents a flower, colored by species.



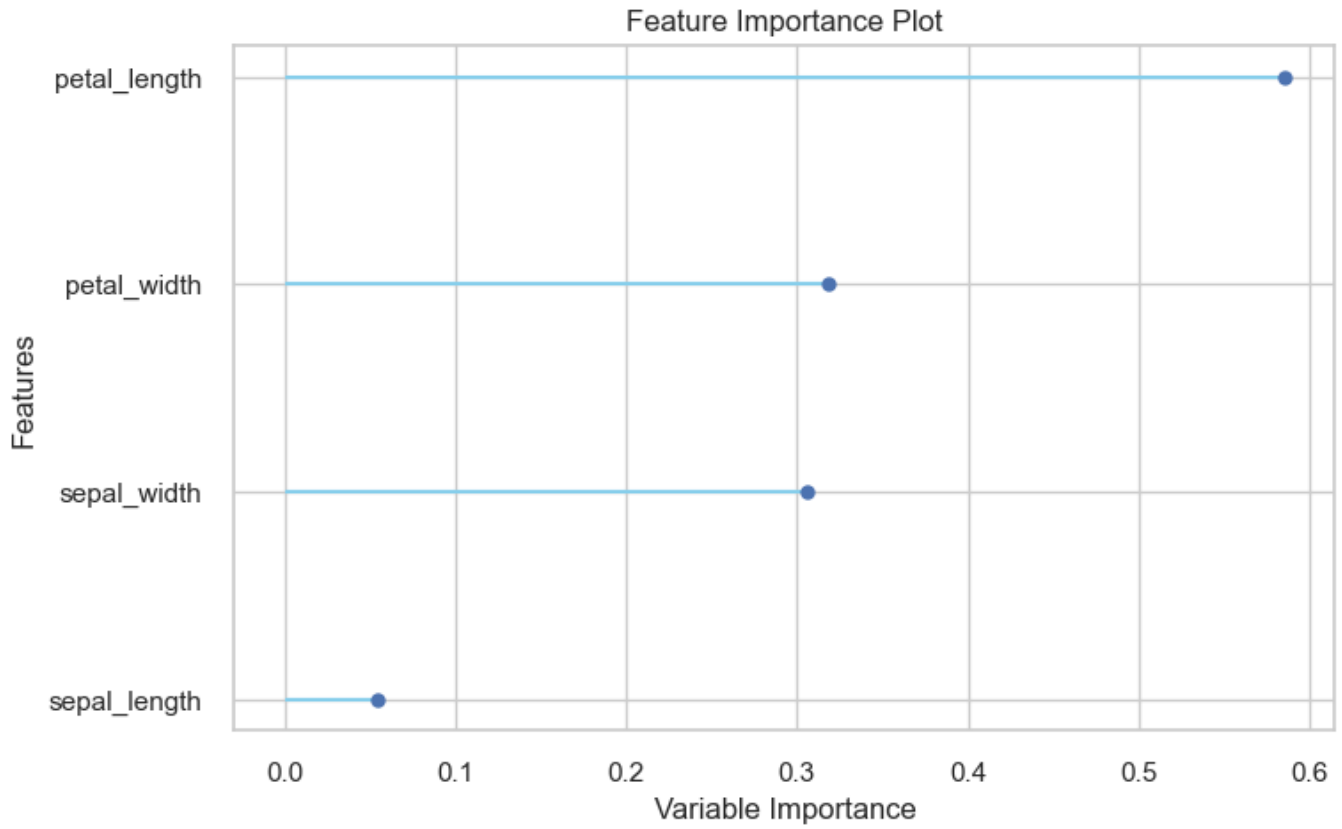
Insights:

- The three species form distinct clusters — especially *Setosa*, which is clearly separated.
- *Versicolor* and *Virginica* are closer together but still partially separable.
- PCA confirms that the dataset is well-structured and likely suitable for accurate classification.

Feature Importance

Which Features Matter Most?

To objectively evaluate which features are most important for distinguishing between iris species, a classification model was trained and the resulting feature importances were analyzed.



Insights:

- **Petal Length (0.58)** is the most important feature for classification.
- **Petal Width** and **Sepal Width** also contribute meaningfully, with similar importance scores.
- **Sepal Length (0.06)** has minimal importance — it adds little to distinguishing the species.

Final Conclusions

- Petal measurements — especially **petal length** — are the most useful features for species classification.
- *Iris Setosa* is the most distinct species, characterized by its short, narrow petals and relatively wide sepals.
- *Iris Virginica* and *Versicolor* are more similar but can still be separated based on petal size.
- Sepal width is the only feature where Setosa scores higher than the other species, making it a unique trait.
- High correlations between petal length and width suggest these dimensions grow proportionally.
- A machine learning model confirmed that petal dimensions are the most important features for species identification.
- With just four simple measurements, we can accurately distinguish between three iris species — a testament to how effectively nature encodes species differences into physical traits.