IRIS DATA EXPLORATION

EXPLORING IRIS DATASET FROM SKLEARN

Benjan Adhikari

March 26, 2025

Iris Data Exploration

Table of Contents

1. Executive Summary

In this Exploratory Data Analysis (EDA) project on the Iris dataset, I identified potential outliers using box plots and determined that there are **no true outliers** based on KDE plots. I confirmed that the dataset is free of zero or null values. I also performed data type conversions, ensuring that numerically stored **categorical variables** were properly represented as categorical data.

Through histograms and KDE plots, I observed that the two petal features show **bimodal distribution**. Correlation analysis and pairplots revealed that *sepal width* has a **weak negative correlation** with the other three features—*sepal length, petal length*, and *petal width*.

These observations can be useful for **classification** tasks. Notably, the **Setosa** species has the **largest average *sepal width*** while having the **smallest measurements** in other features, making it a clear distinguishing factor. Among the remaining three features, **Virginica** consistently has the **highest values**, which helps differentiate it from **Versicolor**.

2. Introduction

The Iris dataset, available through the scikit-learn library, was first introduced by R.A. Fisher [1]. It contains information on 150 iris flowers, including measurements of their sepal and petal features, as well as their species classification.

2.1 Problem Statement and Goals

The Iris dataset is widely used for classification tasks, where flowers are categorized into species based on their petal and sepal features. In this project, we are not developing a classification model. Instead, our goal is to explore the dataset to understand its features, identify patterns, detect outliers, and examine relationships between variables.

Specifically, we aim to determine whether each species exhibits distinct sepal and petal characteristics that differentiate it from the others, which could be useful for classification.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a key step in data analysis, focusing on understanding patterns, trends, and relationships through statistical tools and visualizations [2]. This project primarily focuses on EDA to explore the data and uncover meaningful insights.

3.1 Description of the data

The dataset contains 150 observations with 5 variables.

Numerical variables: There are 4 numerical variables:

- *sepal length (cm)*: length of sepal measured in centimeters
- *sepal width (cm):* width of sepal measured in centimeters
- *petal length (cm)*: length of petal measured in centimeters
- *petal width (cm):* width of petal measured in centimeters

Categorical variables: There is 1 categorical variable.

- *species*: species of the individual flower

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

Figure 1: First 5 rows of the Iris dataset

A statistical summary of the numerical features is shown in figure 2. From the figure, we can observe that there are **no 'zero' values** as the *min* for all columns is greater than 0. Similarly, there are no absurdly large values compared to the mean as seen in the *max* and *mean* rows.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Figure 2: Feature statistics

There are three categories in the *species* column and there are 50 instances of each of these species.

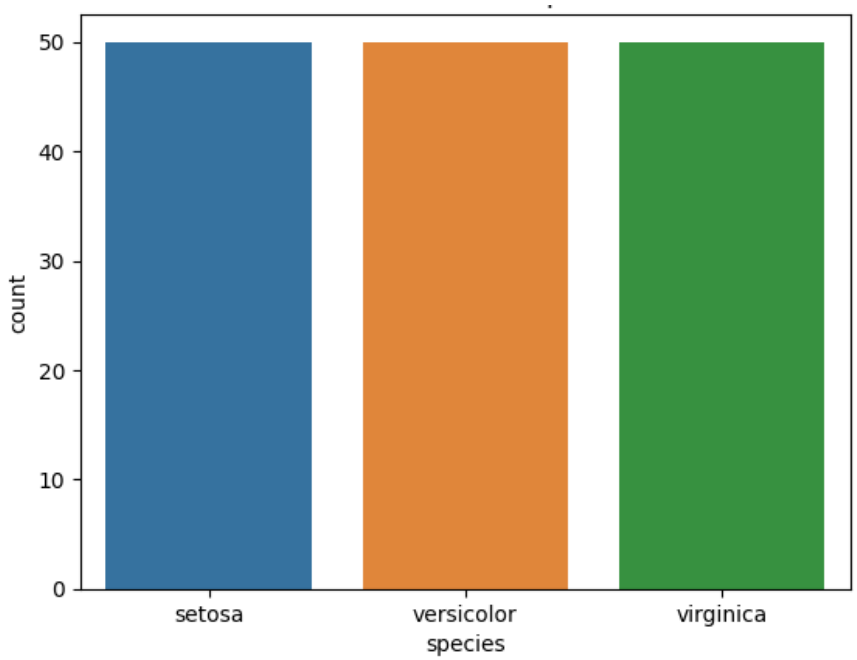

Figure 3: Count of species

3.2 Data Cleaning

The Iris dataset by default is a clean dataset.

- There are no null or zero values in the dataset.
- There are no duplicates.
- The *species* column initially had numerical discrete values {0,1,2} for the different species. While these numerical labels are useful for machine learning models, they are not as intuitive for exploratory data analysis. To improve readability in visualizations and summaries, I mapped these values to their corresponding species names: 0 → setosa, 1 → versicolor, and 2 → virginica.

3.3 Univariate Analysis

Univariate analysis focuses on studying one variable to understand its characteristics. It helps describe the data and find patterns within a single feature [2].

In this section, I use histograms to visualize the frequency distribution of numerical features and boxplots to detect potential outliers.

3.3.1 Histograms

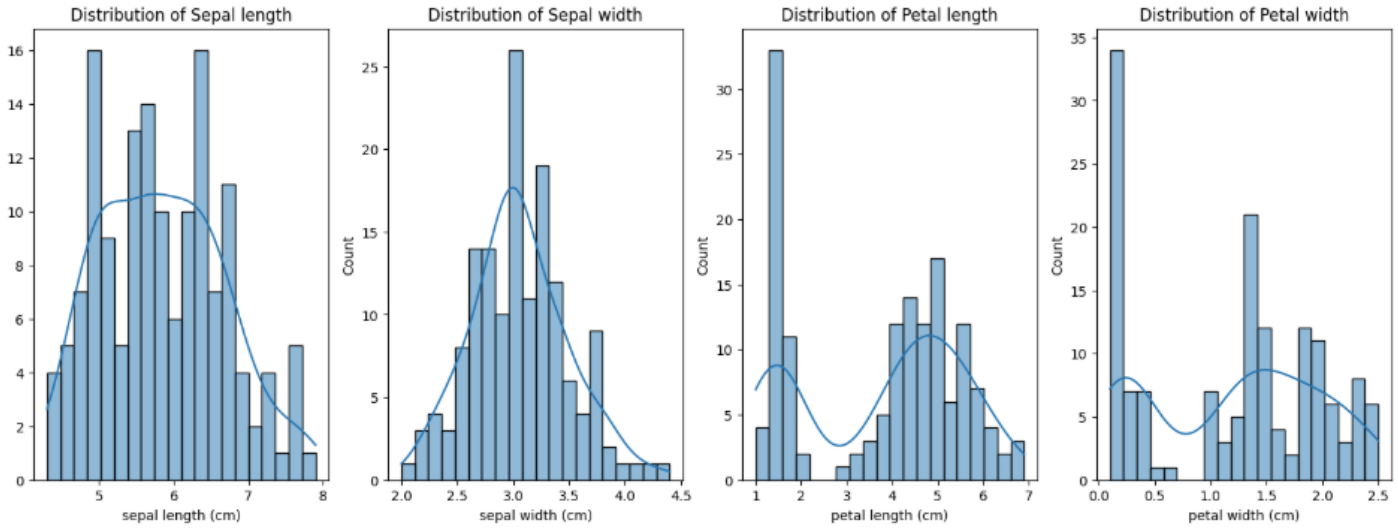

Figure 4: Distribution of the numerical features

As seen in Figure 4, the two petal features (petal length and petal width) show a **bimodal distribution**, meaning there are two clear peaks in the data. This likely happens because different species have different petal sizes. To confirm this, we can look at kernel density estimation (KDE) plots for each species separately to see how their petal measurements are distributed.

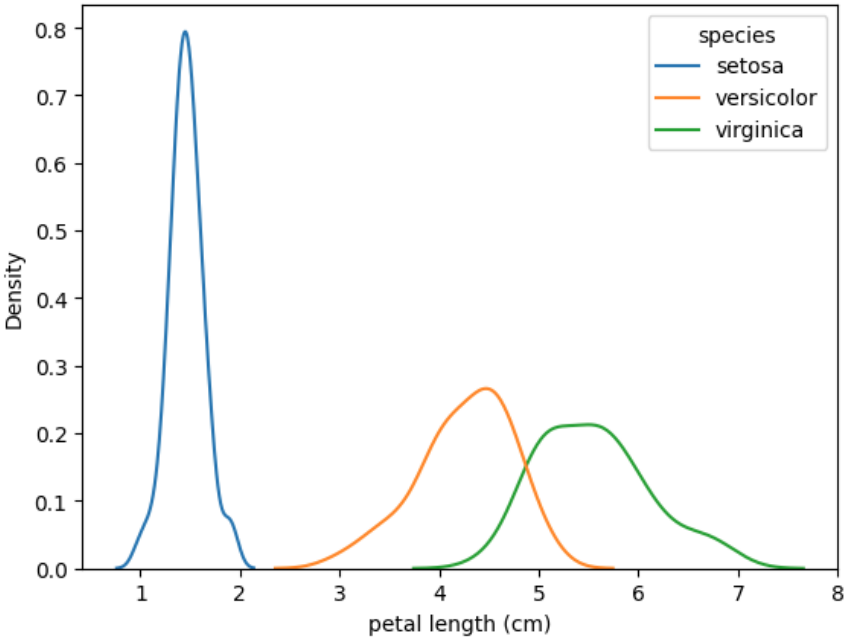

Figure 5: KDE plot for petal length by species

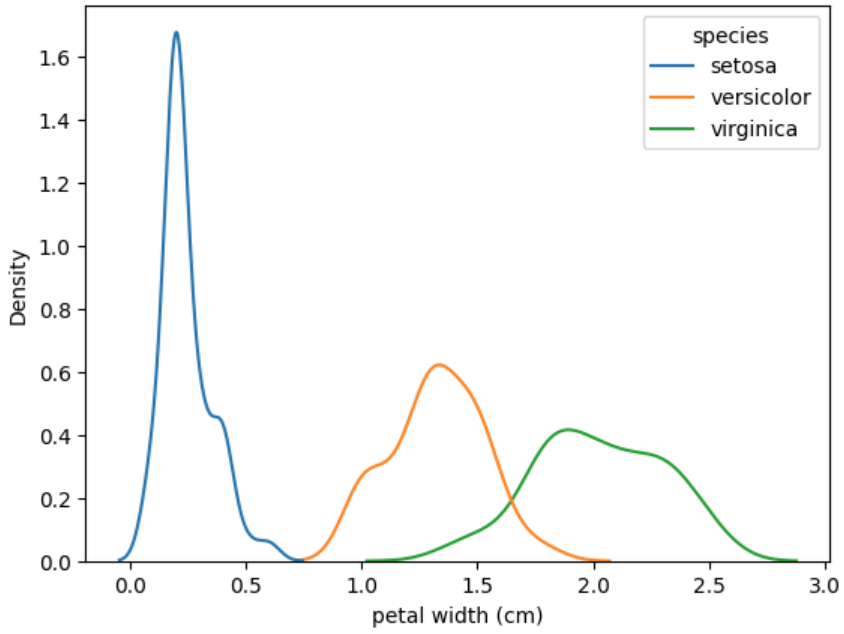

Figure 6: KDE plot for petal width by species

The KDE plots support our idea of bimodal distribution. The petal length and width have **two distinct groups** – setosa (smaller petal size) and, versicolor and virginica (larger petal size).

3.3.2 Boxplots for Outlier Detection

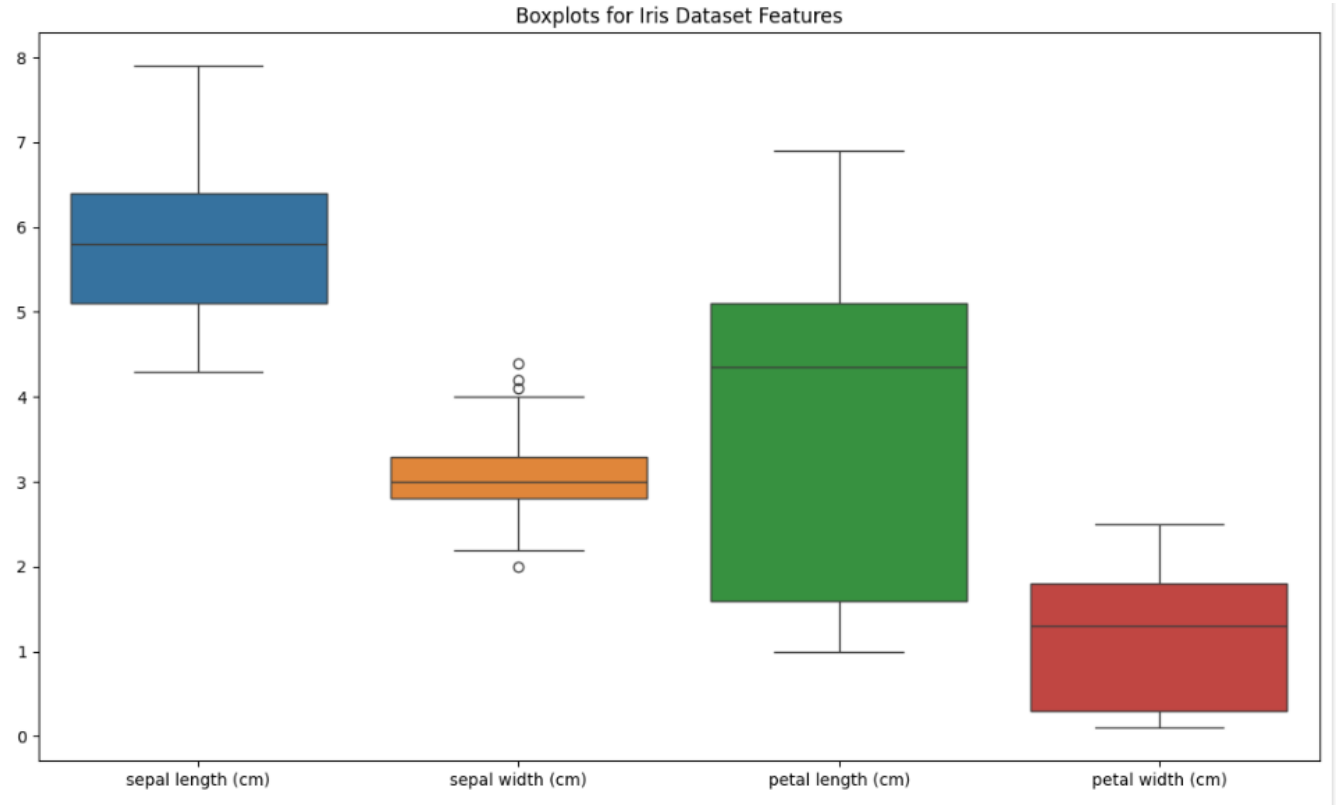

Figure 7: Boxplots for Iris Dataset Numerical Variables

From the boxplot in Figure 7, we observe that only the *sepal width* values extend beyond the whiskers, indicating potential outliers. There are four such points outside the whiskers. The whiskers appear in range between 2 and 4, so we will identify any data points that fall outside this boundary. Next, we will observe these four suspected outlier points in the dataset.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
14	5.8	4.0	1.2	0.2	setosa
15	5.7	4.4	1.5	0.4	setosa
32	5.2	4.1	1.5	0.1	setosa
33	5.5	4.2	1.4	0.2	setosa
60	5.0	2.0	3.5	1.0	versicolor

Figure 8: Suspected Outliers

The observation at **index 60** is the only suspected outlier in the lower range of sepal width, with a value of 2 cm. Since this lies exactly at our outlier boundary, my initial intuition is **not** to classify it as an outlier.

To further investigate, we compared its values across the other three columns with their respective means and standard deviations. We found that all observed values were within one standard deviation of their column means.

Based on these observations, I decided that **index 60 is not an outlier**, and we will keep it in the dataset as is.

	Observed value	Mean	Standard deviation	Z-score
sepal length	5.0	5.84	0.83	-1.00
petal length	3.5	3.75	0.44	-0.57
petal width	1.0	1.19	0.76	-0.25

Table 1: Column values for suspected outlier index 60

The remaining suspected outliers in the higher range all belong to the species **setosa** (Figure 8). When examining the KDE plot for *sepal width* across different species, we observe that the setosa distribution is centered further to the **right** compared to the other two species (Figure 9).

This supports our observations, what seemed like outliers at first are a result of setosa's distinct distribution. Therefore, these points are not true outliers but rather a reflection of the species' natural data pattern.

So, we decide to keep all points as is and decide that there are **no true outliers in the dataset.**

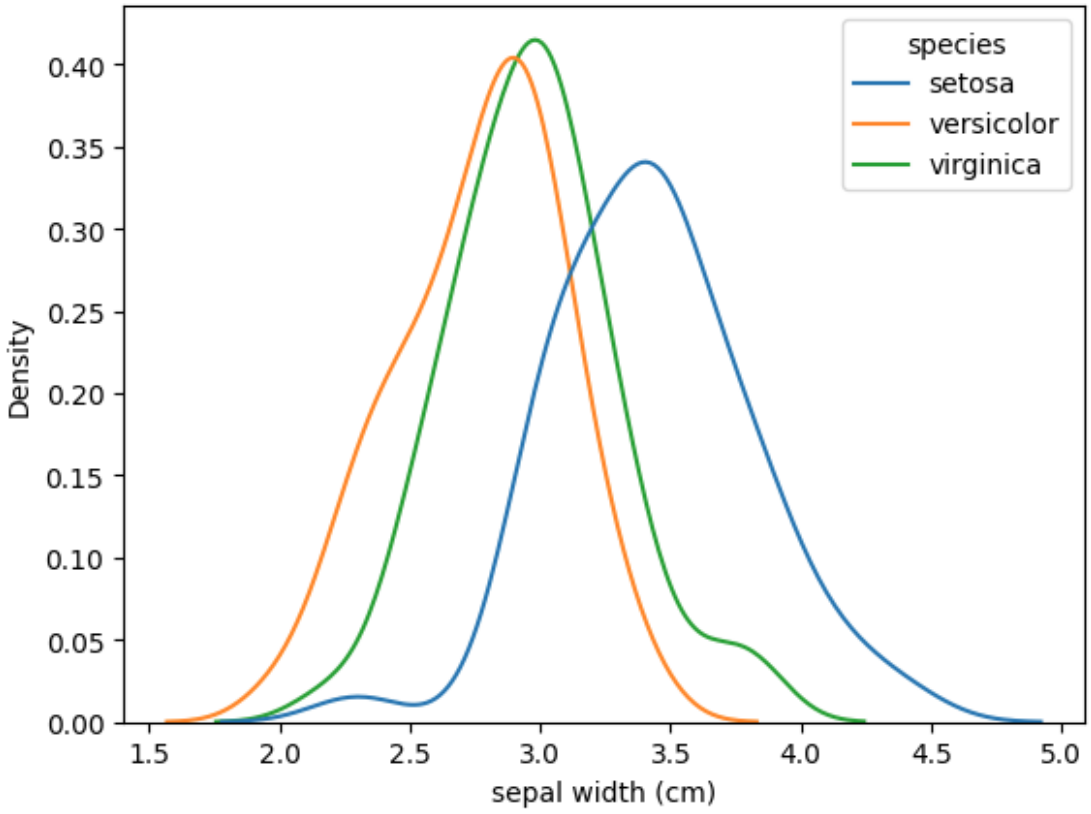

Figure 9: KDE plot for sepal width by species

3.4 Bivariate & Multivariate Analysis

In this section, we will explore the relationship between the different variables using pairplots and correlation heatmap.

3.4.1 Scatterplots & Pairplots

The pairplots for the numerical features of the dataset can be seen below in Figure 10. There are 6 scatterplots, each depicting the relationship between any two of the 4 features. The scatterplots involving *sepal width* are circled in the figure.

We can observe that the scatterplots involving *sepal width* are scattered, thus sepal width does not seem to have any strong relationship with the other features. While the other three features show a clear positive relationship with each other, evident from the upward trend of the points.

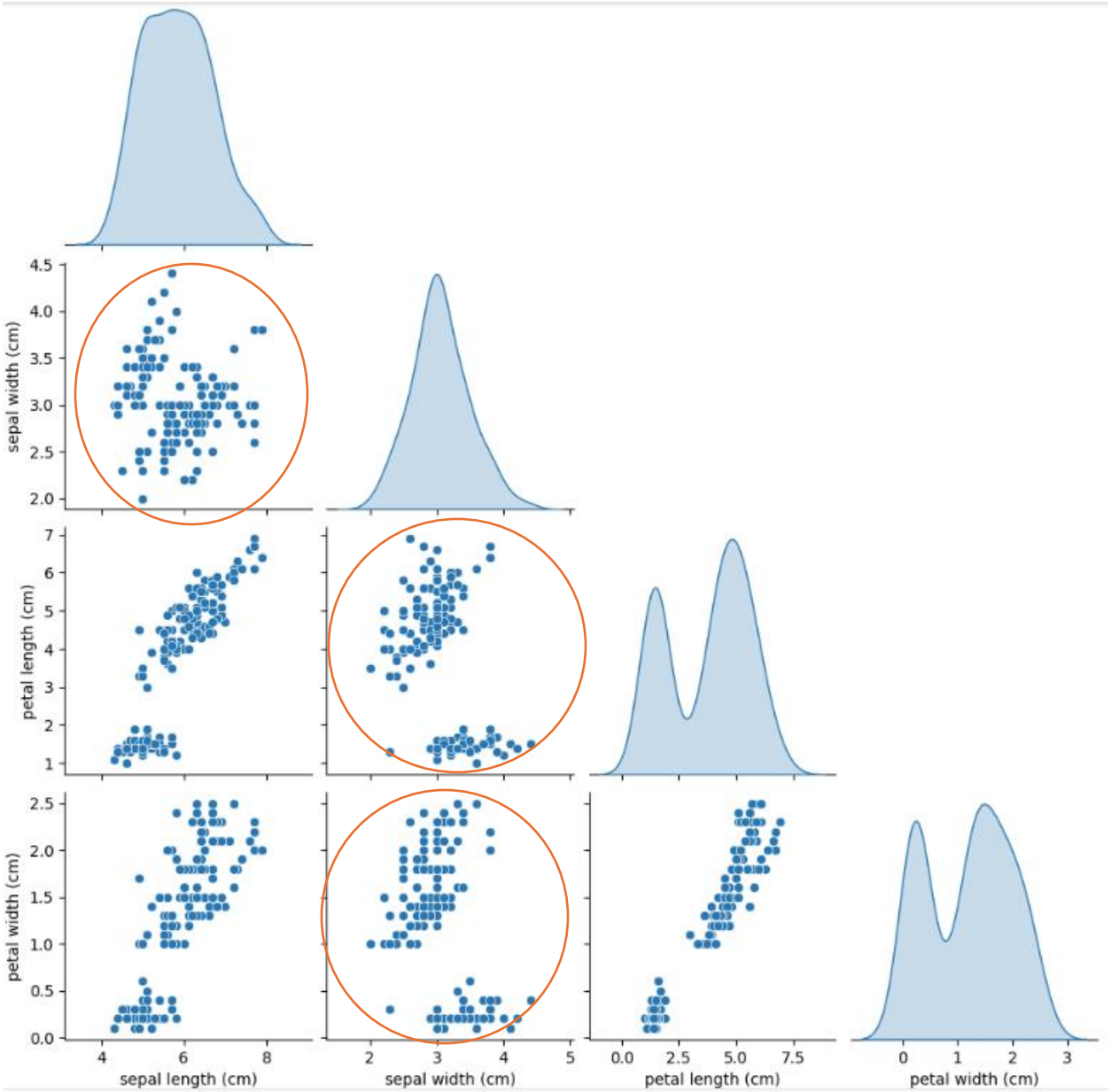

Figure 10: Pairplots for numerical features

3.4.2 Correlation Analysis

The correlation heatmap in Figure 11 further supports our observation from the pairplots. The *sepal width* feature has weak negative correlation with the other three features. While the remaining three features have a high positive correlation (close to +1) with one another.

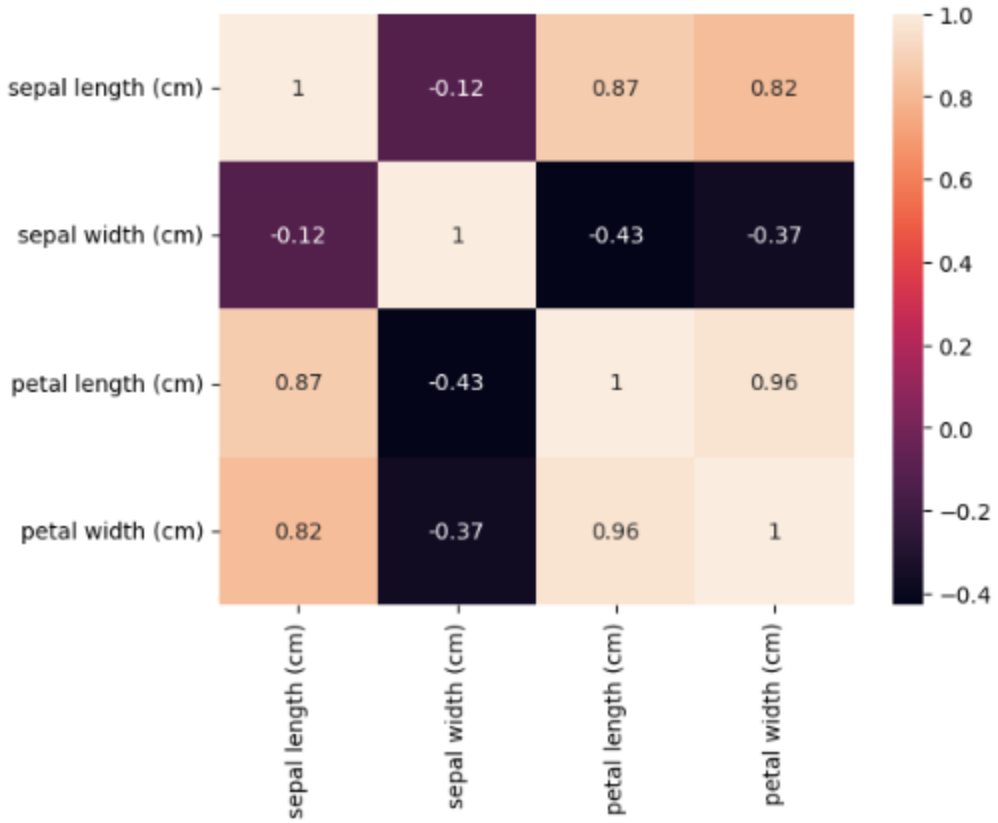

Figure 11: Correlation heatmap of numerical features

The observations from the pairplots and correlation heatmap suggest that *sepal width* may not be a strong predictor for the remaining features as it doesn't have a strong relationship with the remaining features.

Additionally, the strong positive correlation among sepal length, petal length, and petal width indicates potential **multicollinearity** issue, which could affect predictive modeling. While dimensionality reduction methods like Principal Component Analysis (PCA) or feature selection could help address this, they are **not necessary for our project**. Given the small number of features, we will retain all measures to preserve the interpretability of the data.

3.5 Species

The Iris dataset is primarily used for classifying the flower into respective species. Through the KDE plots for separate species in the previous sections, we have noticed some key differences in the feature distributions across species.

In this section, we will further explore these differences by examining the average values of each feature, grouped by species. This will help us gain deeper insights into how the species vary based on their numerical attributes.

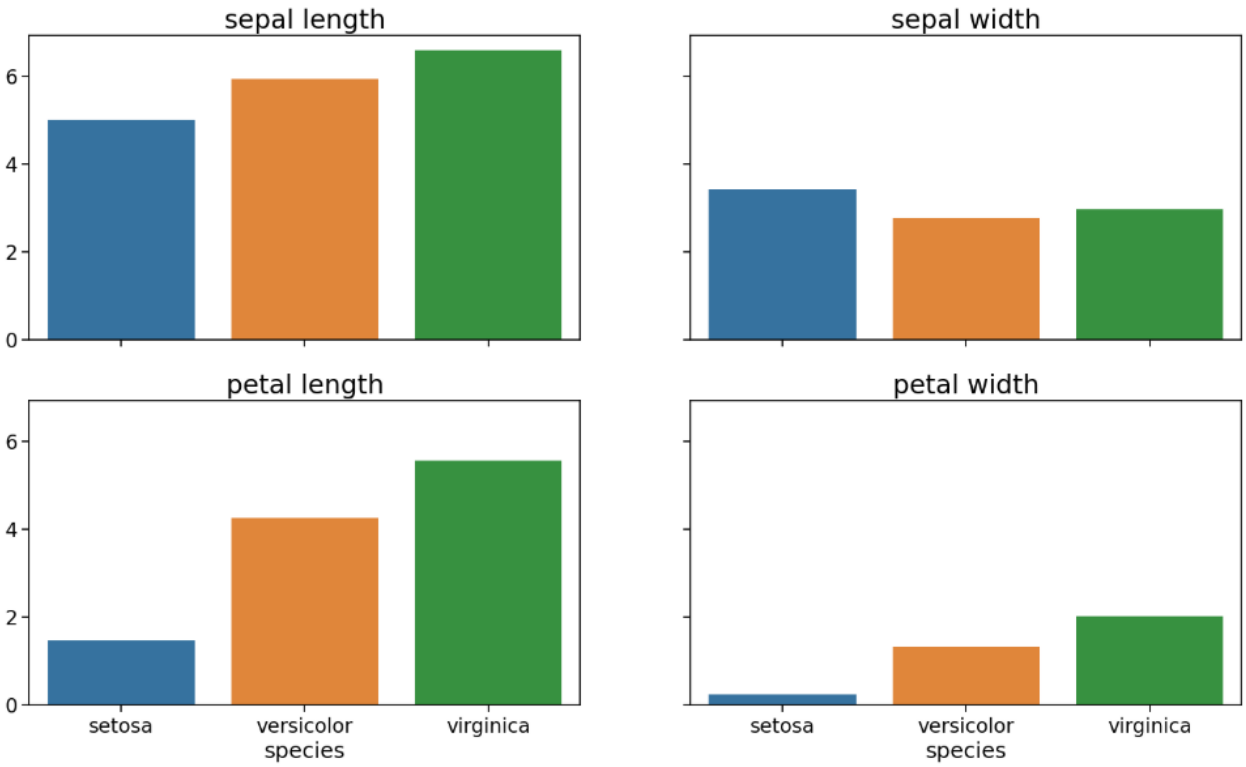

Figure 12: Average feature measurement across species

There are a few things we can notice from the plots:

- The length features *sepal length* and *petal length* seem to be larger on average compared to the width features.
- There are similar trends for the different features. For the three features: *sepal length*, *petal length* and *petal width*, setosa is the smallest, versicolor is in the middle and virginica is the largest. This trend across the three features makes sense as there was high correlation between the three features.
- In terms of sepal width, setosa species is the largest which is in complete contrast to their performance in other three features. This does make sense in terms of our previous observation of it being weak negatively correlated with other features.

4. References

[1] https://scikit-learn.org/stable/datasets/toy_dataset.html#iris-dataset

[2] https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/