

**ROLE: DATA ANALYST**

**PROJECT 1: IRIS DATASET BASIC ANALYSIS**

**Deekshitha Dommati**

**INTRODUCTION**

**Project Overview:**

The Iris dataset is a renowned dataset in the fields of machine learning and statistics. It comprises 150 observations of iris flowers from three species: setosa, versicolor, and virginica. Each observation includes measurements of four features: sepal length, sepal width, petal length, and petal width. The dataset is frequently used for benchmarking classification algorithms and for educational purposes in data analysis.

**Objective:**

The primary objective of this project is to perform an exploratory data analysis (EDA) on the Iris dataset using Python. The goal is to uncover insights, explore correlations, and understand the distributions of different features across the species. Subsequently, the insights obtained will be visualized using Power BI or Tableau to represent the identified patterns effectively.

**METHODOLOGY**

**Data Preparation:**

* Import Libraries: Necessary libraries such as pandas, seaborn, and matplotlib were imported to facilitate data manipulation and visualization.
* Load Dataset: The Iris dataset was loaded using seaborn’s built-in function.
* Initial Inspection: The first few rows of the dataset were displayed to understand its structure.
* Basic Data Checks: Performed data checks to understand data types, summary statistics, and to check for missing values.

**Data Visualization in Python:**

* Pairplot: Visualized pairwise relationships between features, with species differentiation using hues.
* Boxplot: Displayed the distribution of each feature.
* Violin Plot: Illustrated the distribution and density of each feature across species.
* Correlation Heatmap: Showed the correlation between numerical features.

**Data Visualization with Tableau:**

1. Save Dataset: Saved the Iris dataset as a CSV file for import into Power BI or Tableau.
2. Load Data: Imported the CSV file into Power BI or Tableau.
3. Create Visualizations:

* Pairwise Scatter Plots: To visualize relationships between feature pairs.
* Box Plots: To compare feature distributions across species.
* Violin Plots: To show feature distributions and densities.
* Correlation Heatmap: To visualize correlations between numerical features





**PATTERNS AND INSIGHTS**

Pairwise Relationships

* Observation: Pairwise scatter plots reveal clear separations between species, particularly when plotting petal length and petal width.
* Insight: Petal dimensions are more effective in distinguishing species compared to sepal dimensions.

Distribution of Features

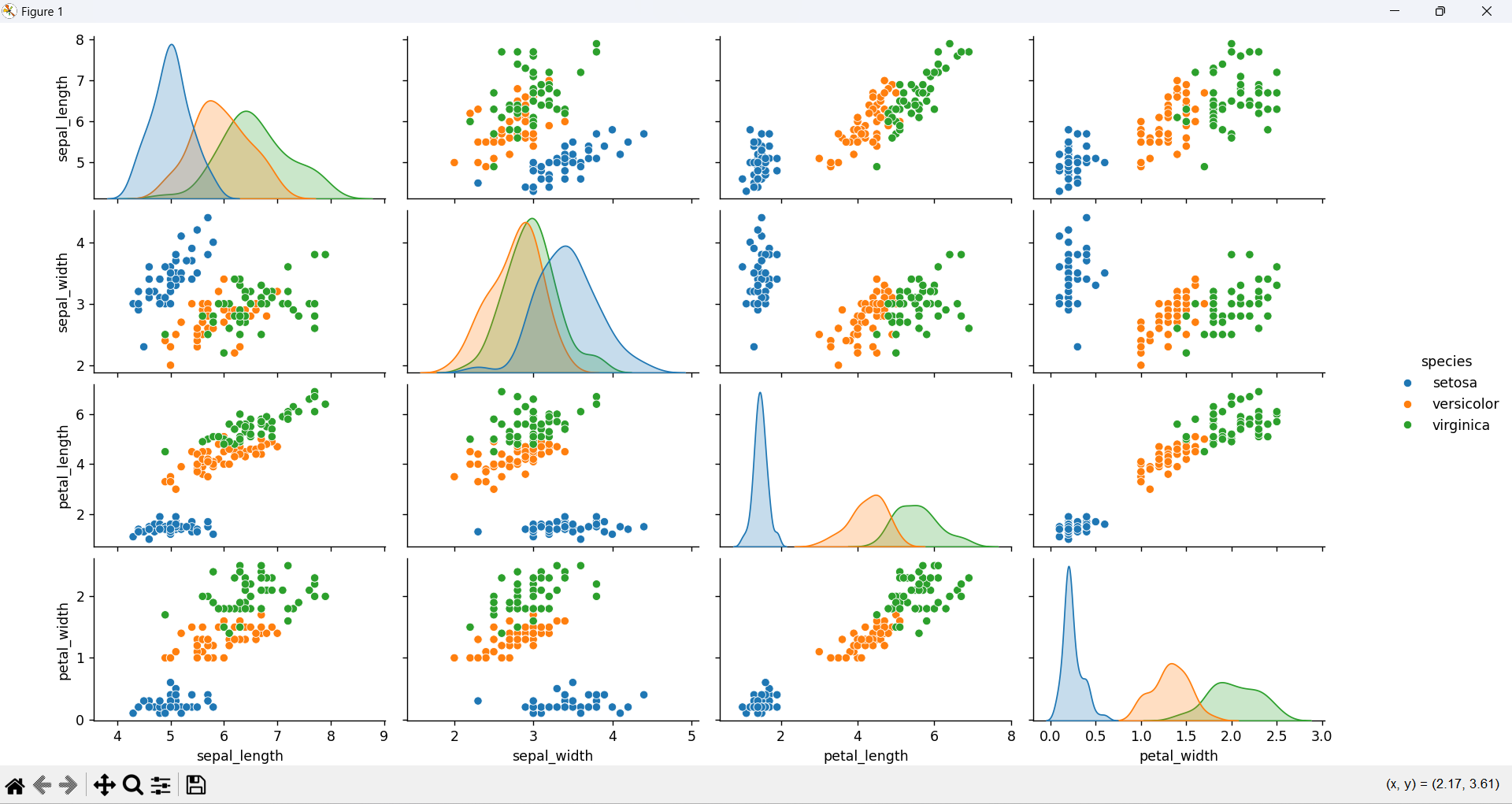
* Observation: Box plots and violin plots show the distribution of sepal and petal lengths and widths across species.
* Insight: Setosa species have smaller petal lengths and widths, while virginica species tend to have larger values for these features.

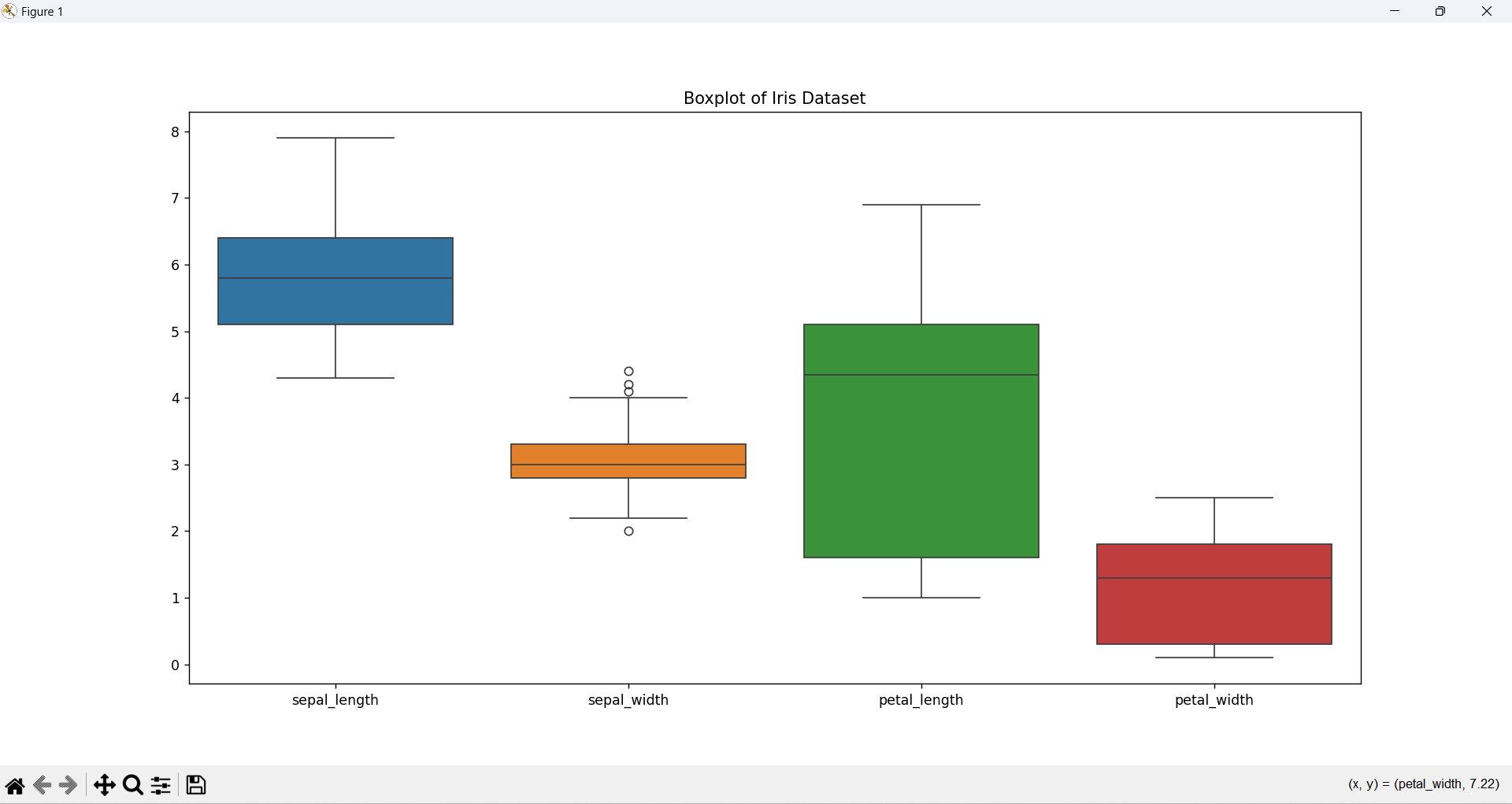
Correlation Analysis:

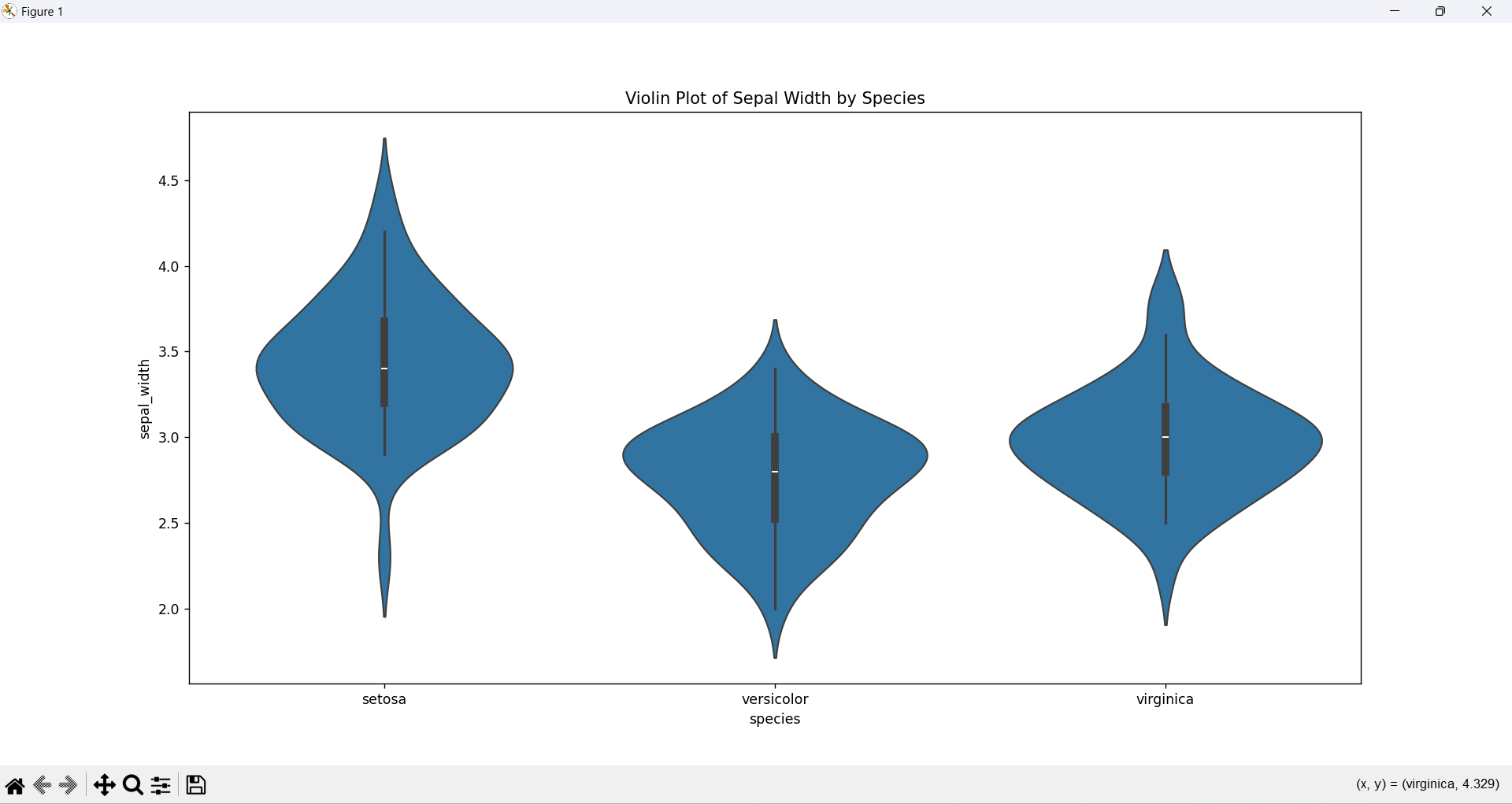
* Observation: The correlation heatmap indicates strong positive correlations between petal length and petal width.
* Insight: High correlation between petal length and width suggests these features grow proportionally and can be used together for species classification.

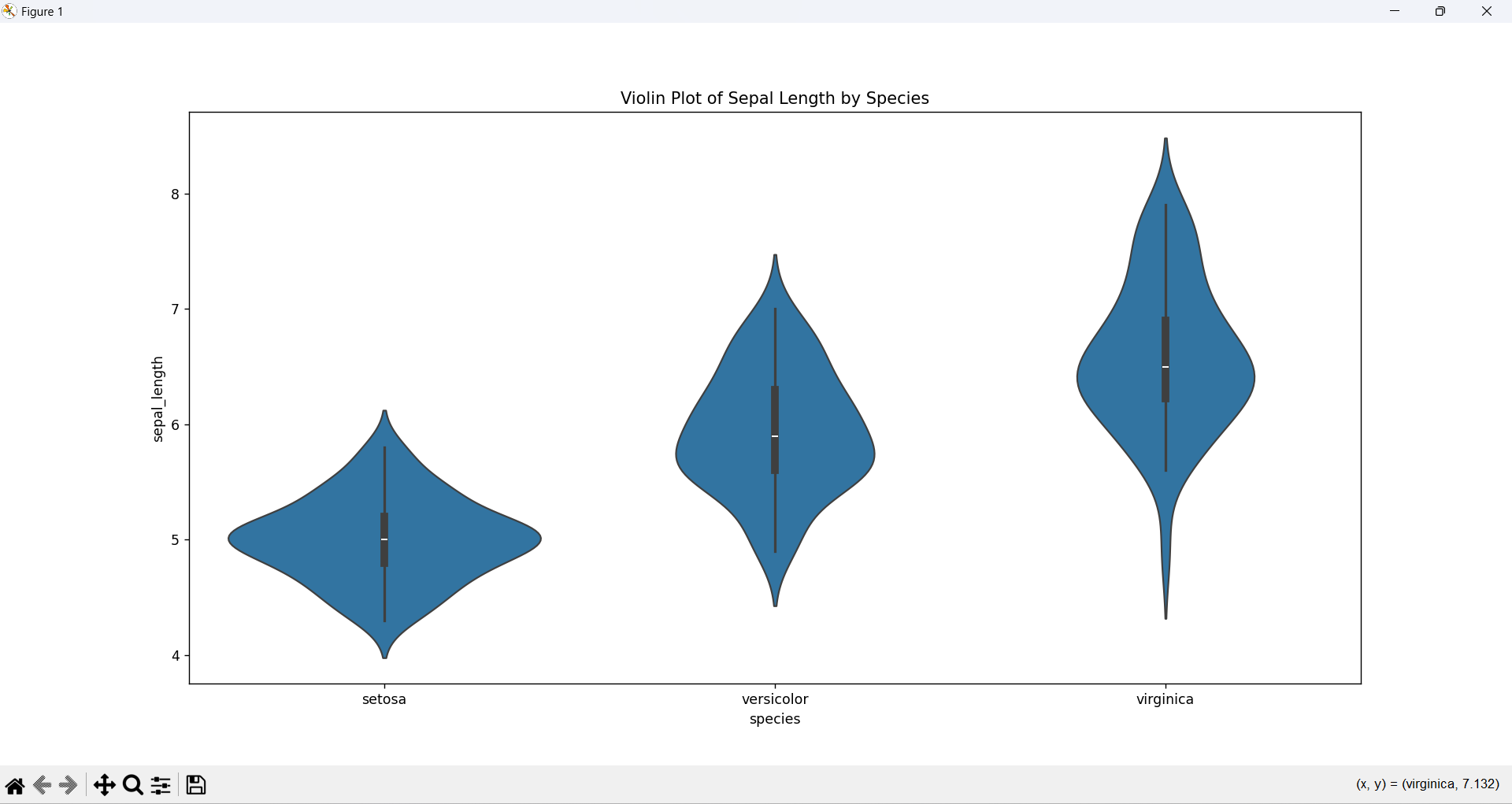
Species Distribution:

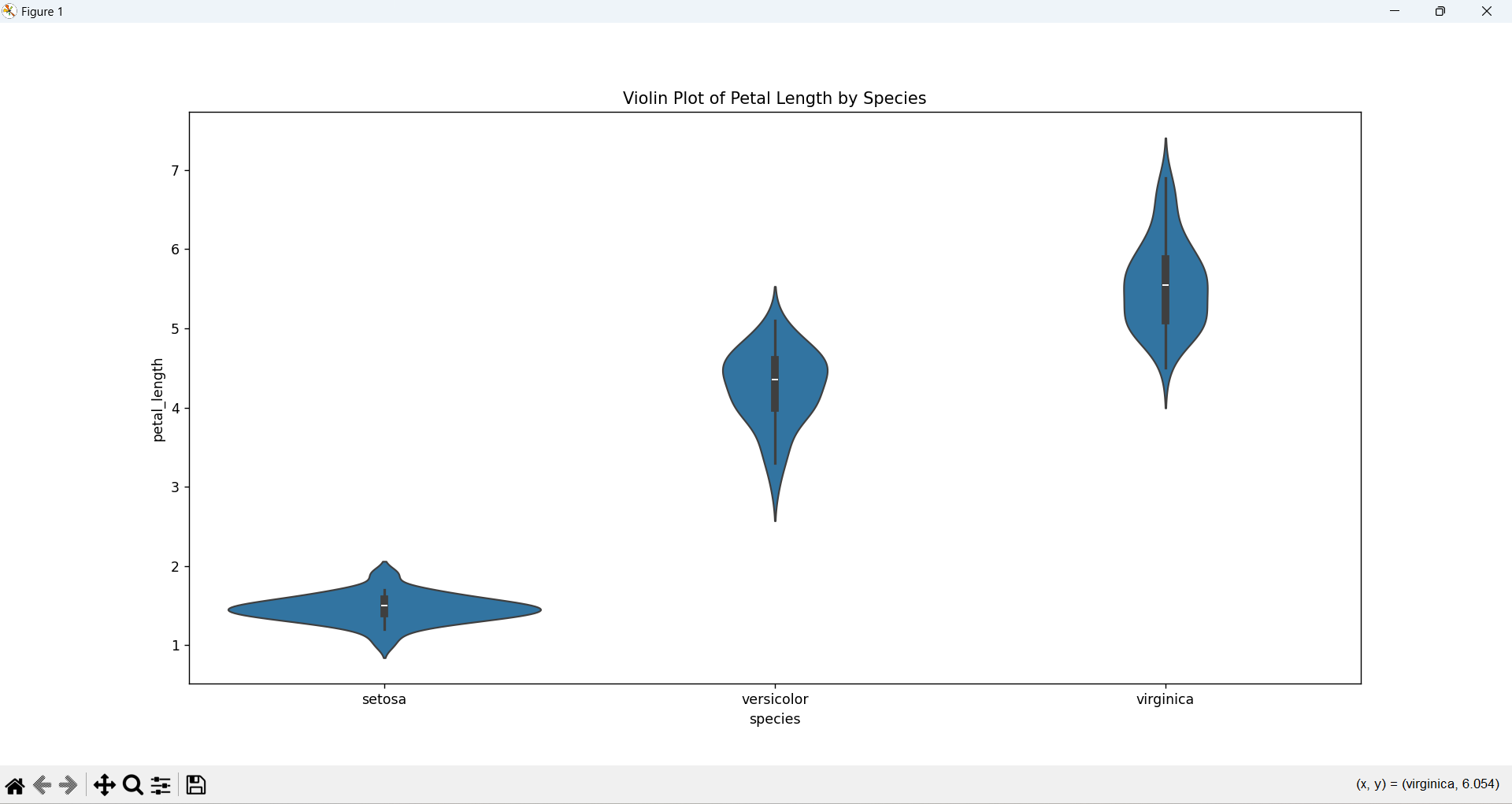
* Observation: The distribution of species is balanced, with each species represented by 50 samples.
* Insight: The balanced dataset ensures that the analysis and any subsequent modeling will not be biased towards any species.

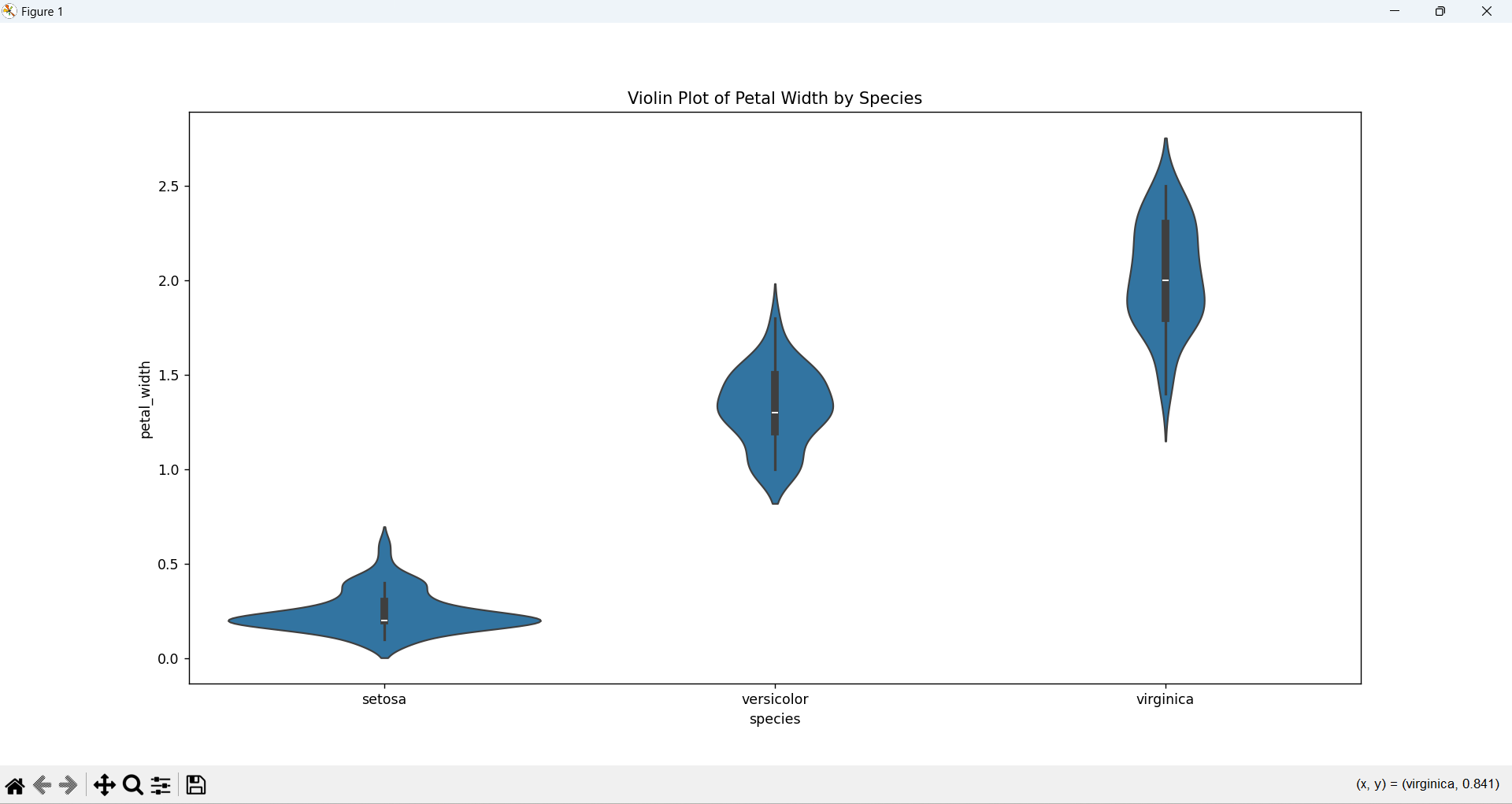


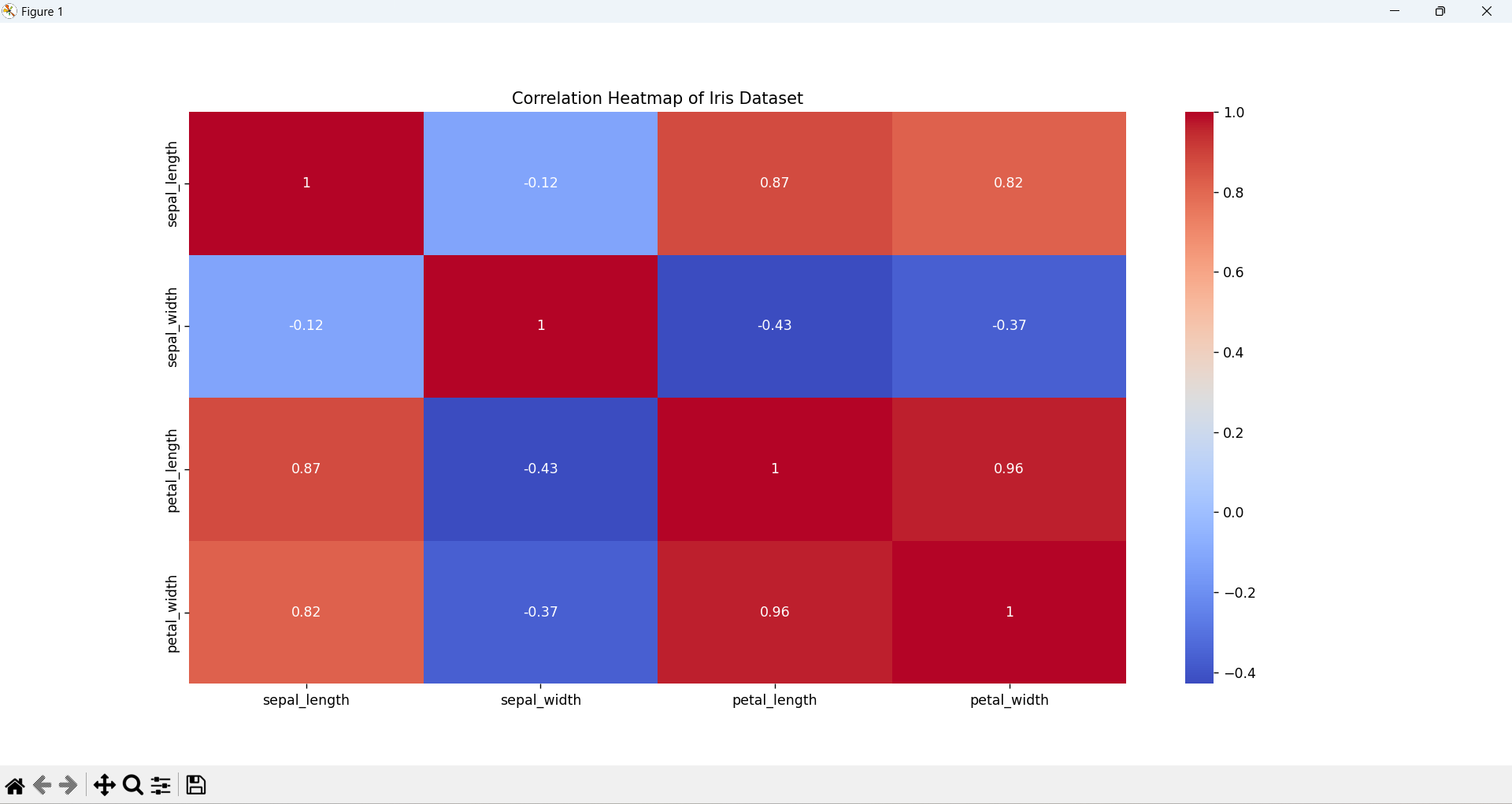












**CONCLUSION**

**Summary:**

The exploratory data analysis (EDA) on the Iris dataset provided significant insights into the relationships and distributions of the four key features: sepal length, sepal width, petal length, and petal width. By utilizing various visualization techniques, we were able to:

* Identify clear separations between the three iris species, particularly in the petal dimensions.
* Highlight the distinct distribution patterns of each feature across the species.
* Observe strong correlations between petal length and petal width, indicating these features' proportional growth.

These insights underscore the value of EDA in understanding the underlying structure of the data and setting the stage for more advanced analysis, such as predictive modeling.

**Future Work:**

Future work could focus on:

* Machine Learning: Applying classification algorithms (e.g., logistic regression, decision trees, random forests) to predict the species of iris flowers based on the observed features.
* Advanced Visualizations: Developing more sophisticated and interactive visualizations using tools like Power BI or Tableau to enable dynamic exploration of the dataset.
* Feature Engineering: Exploring additional derived features or transformations that might further improve classification accuracy.
* Model Evaluation: Evaluating and comparing different models' performance using metrics like accuracy, precision, recall, and F1-score.

By continuing this analysis, we can build robust models that leverage the identified patterns and correlations to achieve accurate species classification, thus demonstrating the practical applicability of EDA in machine learning and data science projects.