Iris Dataset Analysis Report

Prepared by : Ruwan Pathiranage Sanduni Nisansala

Internship ID : CV/A1/45373

Tools Used : Python, pandas, matplotlib, seaborn

Project : Data Cleaning, Exploratory Data Analysis, and Visualization

Table of Contents

1. Introduction

The Iris dataset is one of the most famous and widely used datasets in data science. It contains measurements of sepal length, sepal width, petal length, and petal width for three species of the Iris flower - *Setosa*, *Versicolor*, and *Virginica*.

The main purpose of this project is to:

- Data Cleaning & Preprocessing.
- Exploratory Data Analysis (EDA).
- Basic Data Visualization.

This project demonstrates beginner level foundational skills in data handling, analysis, and visualization using Python libraries such as pandas, matplotlib, and seaborn.

2. Objectives

The objectives of this project in level 1 were divided into three main tasks:

Level 1 – Task 1: Data Cleaning and Preprocessing

- Load the dataset using pandas.
- Identify and handle missing value.
- Remove duplicate rows & Standardize inconsistent data formats.

Level 1 – Task 2: Exploratory Data Analysis (EDA)

- Calculate summary statistics (mean, median, mode, standard deviation).
- Visualize data distributions using histograms, boxplots & scatterplots.
- Find correlations between numerical features.

Level 1 – Task 3: Basic Data Visualization

- Create bar plots, line charts, and scatter plots.
- Customize plot labels, titles, and legends.
- Export plots as images for reports.

3. Tools and Technologies

Tool	Purpose
Python	Programming language used for analysis
pandas	Data cleaning, manipulation, and summary statistics
matplotlib	Basic plotting and data visualization
seaborn	Advanced visualizations and statistical plots
VS Code	Code editor for Python development / Development environment

4. Task 1: Data Cleaning and Preprocessing

Description

The dataset 1)iris.csv was imported using pandas.

The following cleaning steps were applied:

1. Loaded the dataset:

```
df = pd.read_csv("1)iris.csv")
```

2. Checked for missing values and duplicates:

   - No missing values found.
   - Duplicate rows were removed using:

```
df.drop_duplicates(inplace=True)
```

3. Saved the cleaned dataset:

```
df.to_csv("cleaned_iris.csv", index=False)
```

Results

Step	Output
Missing Values	None
Duplicate Rows	Removed
Clean Dataset	cleaned_iris.csv

5. Task 2: Exploratory Data Analysis (EDA)

Summary Statistics

The main summary statistics were calculated using:

```
df.describe()
df.mean(), df.median(), df.mode(), df.std()
```

Summary Statistics for Numerical Columns:

	sepal_length	sepal_width	petal_length	petal_width
count	147.000000	147.000000	147.000000	147.000000
mean	5.856463	3.055782	3.780272	1.208844
std	0.829100	0.437009	1.759111	0.757874
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.400000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Mean values:

sepal_length 5.856463
sepal_width 3.055782
petal_length 3.780272
petal_width 1.208844
dtype: float64

Median values:

sepal_length 5.8
sepal_width 3.0
petal_length 4.4

petal_width 1.3
dtype: float64

Mode values:
sepal_length 5.0
sepal_width 3.0
petal_length 1.4
petal_width 0.2
species versicolor
Name: 0, dtype: object

Standard Deviation:
sepal_length 0.829100
sepal_width 0.437009
petal_length 1.759111
petal_width 0.757874
dtype: float64

Key Observations:

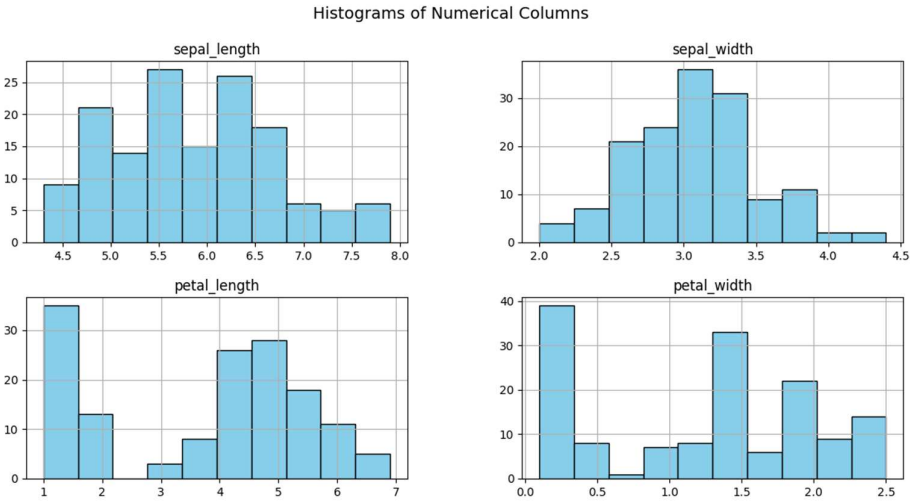
- All measurements are numeric and continuous.
- Petal-related numerical features show larger variation compared to sepal-related numerical features.

Visualizations

All plots were created using matplotlib and seaborn, and exported as PNG files.

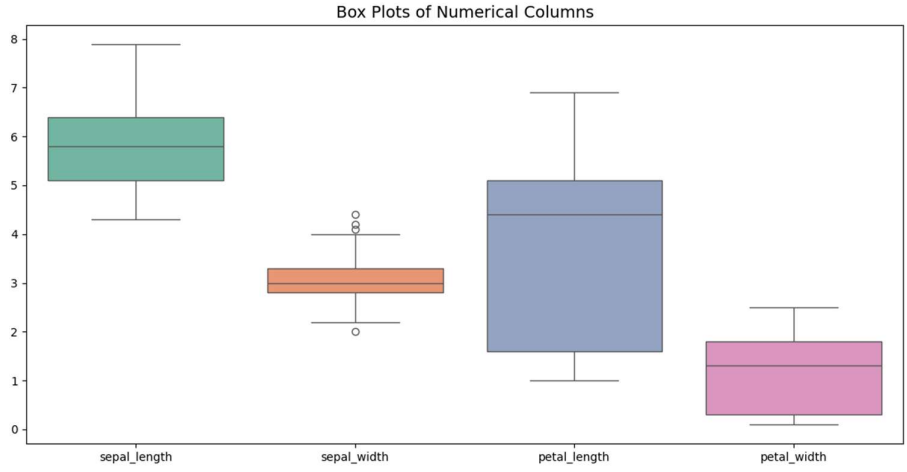
Histograms

- Showed the distribution of each numerical column.
- Sepal width followed a near-normal distribution.



Box Plots

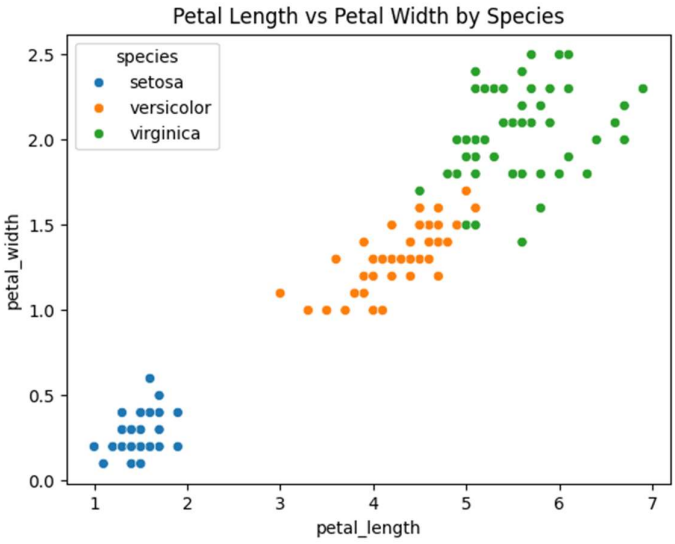
- Helped detect data spread and potential outliers.



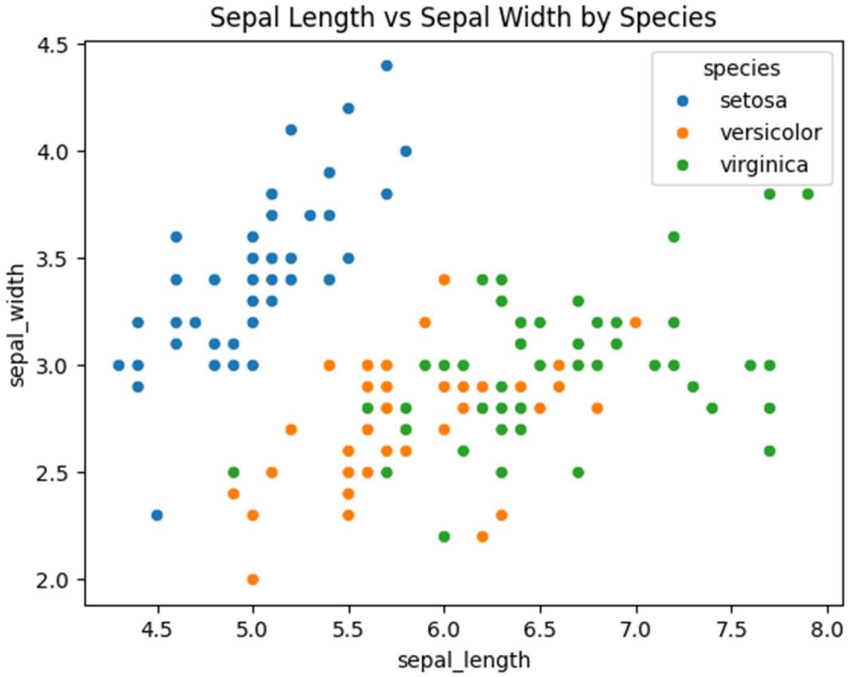
- The outliers can be seen in sepal width.

Scatter Plots

- Distribution between species based on petal length and width.

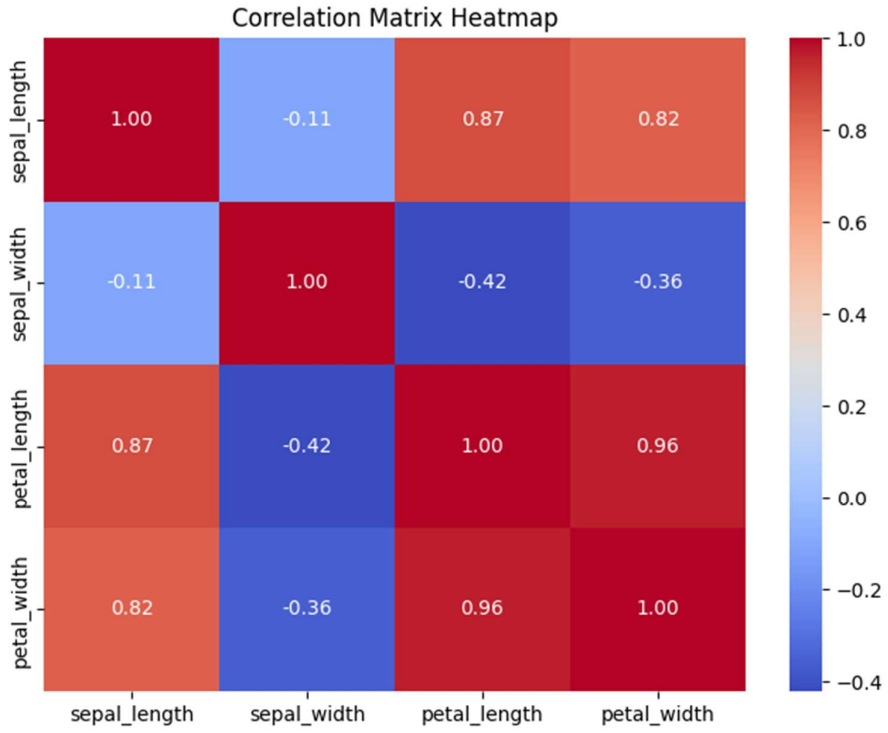


- Distribution between species based on petal length and width.



Correlation Heatmap

- Petal length and petal width have a strong positive correlation (0.96).
- Sepal width has a weak correlation with other features.



Correlation Matrix:

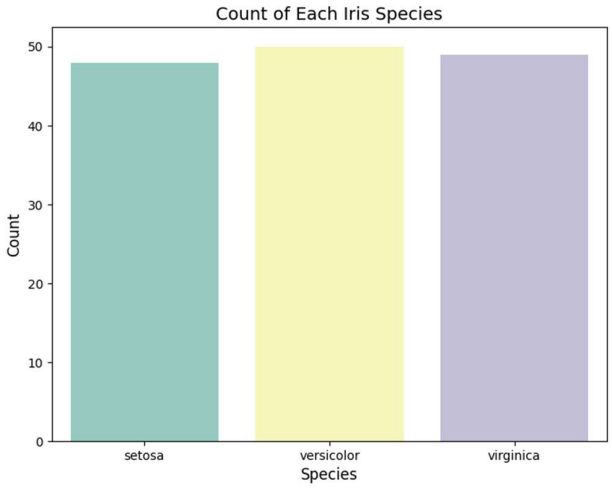
	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.109321	0.871305	0.817058
sepal_width	-0.109321	1.000000	-0.421057	-0.356376
petal_length	0.871305	-0.421057	1.000000	0.961883
petal_width	0.817058	-0.356376	0.961883	1.000000

6. Task 3: Basic Data Visualization

Bar Plot

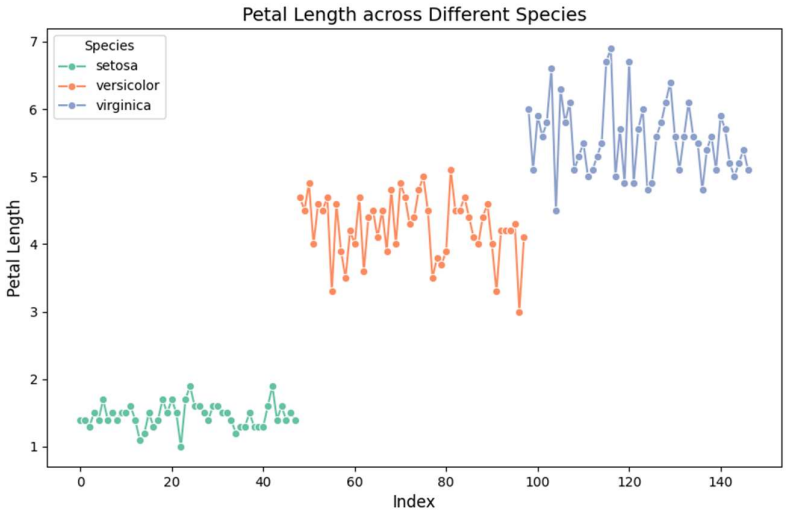
Displayed the count of each Iris species.

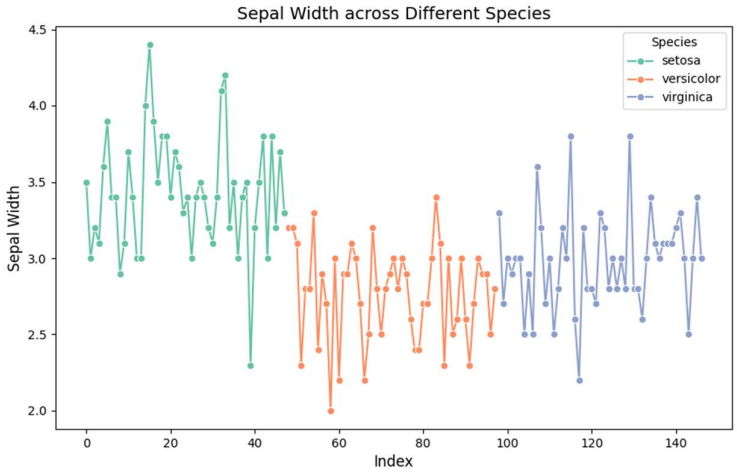
Saved as: species_count_barplot.png

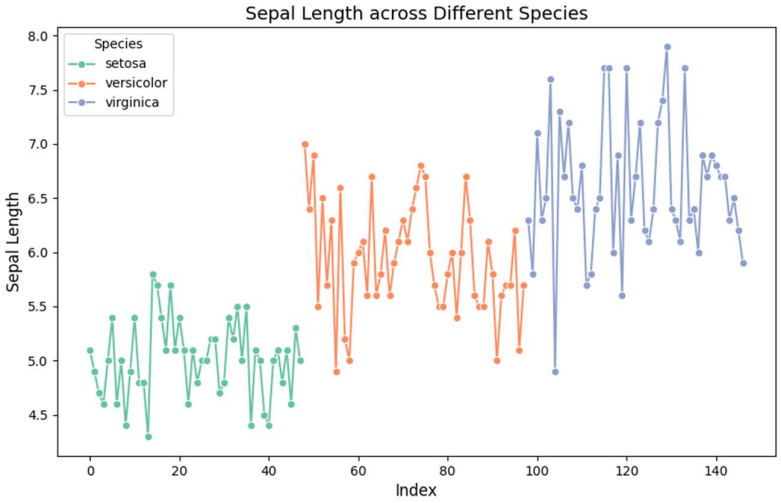


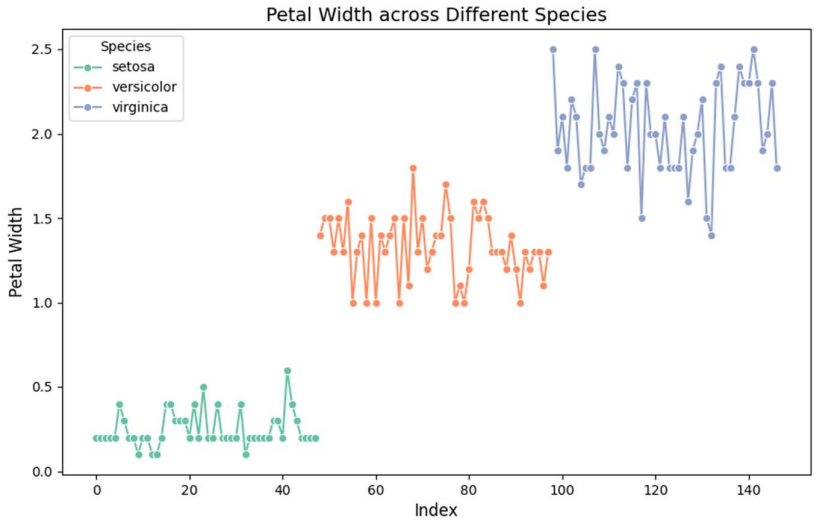
Line Charts

- Visualized the variation of sepal and petal measurements across dataset samples.



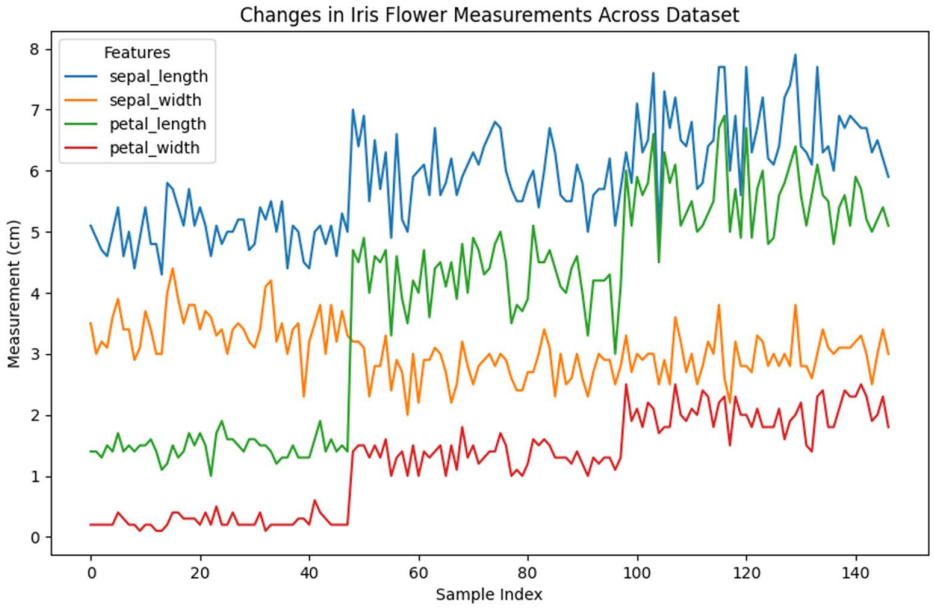

Sepal Width across Different Species
Species
setosa
versicolor
virginica
Sepal Width
Index


Sepal Length across Different Species
Species
setosa
versicolor
virginica
Sepal Length
Index


Petal Width across Different Species
Species
setosa
versicolor
virginica
Petal Width
Index

- Combined chart included all four numeric features.

  Saved as: combined_line_chart.png

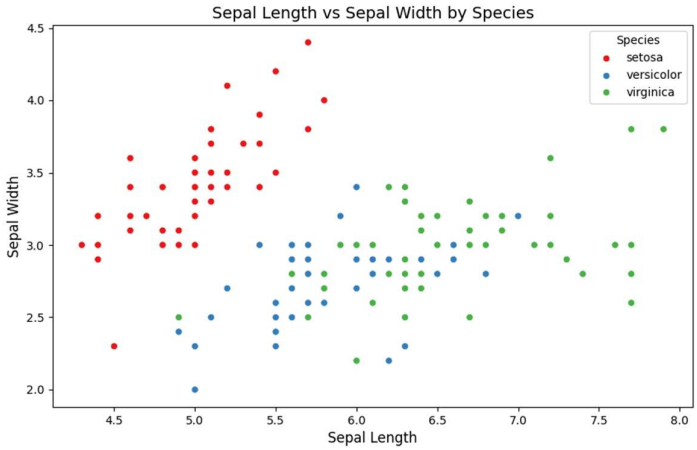


Scatter Plots

- Highlighted relationships between sepal and petal dimensions.

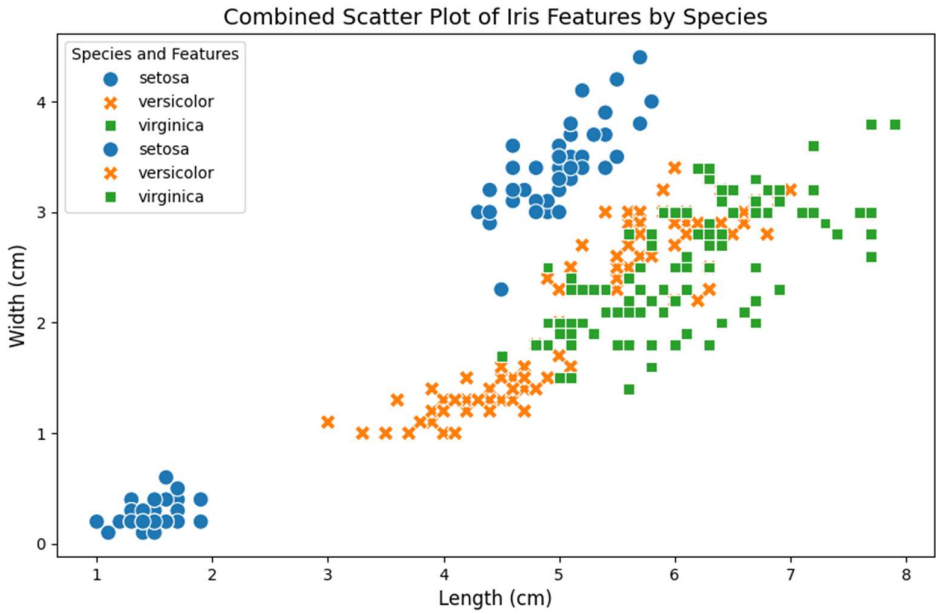
  Saved as:

  - sepal_scatter_plot.png



- petal_scatter_plot.png

- combined_scatter_plot.png



7. Results and Key Findings

Feature Relationship	Observation
Petal Length & Width	Strong positive correlation
Sepal Length & Width	Weak correlation
Species Distribution	Setosa distinctly separated from others
Data Quality	Clean and consistent, no missing values

Insight Summary

- *Setosa* flowers have the smallest petal and sepal sizes.
- *Versicolor* and *Virginica* show gradual increases in petal size.
- Petal-related numerical features are most useful for differentiating species.

8. Conclusion

This project successfully demonstrated the full data analysis process:

1. Cleaning and preparing data.
2. Performing exploratory data analysis.
3. Visualizing insights through meaningful charts.

The Iris dataset proved ideal for practicing basic data science concepts such as descriptive statistics, correlation analysis, and visualization. These steps form the foundation for more advanced tasks like machine learning and predictive modeling.

9. List of Figures

Figure	File Name
Histogram of Numerical Columns	Histogram.png
Box Plots of Numerical Columns	BoxPlots.png
Scatter Plot (Sepal)	sepal_scatter_plot.png
Scatter Plot (Petal)	petal_scatter_plot.png
Combined Line Chart	combined_line_chart.png
Bar Plot (Species Count)	species_count_barplot.png
Correlation Heatmap	correlation_matrix_heatmap.png

End of the Report