Title- R Iris Dataset Exploratory Data Analysis (EDA) Report

Date: April 2nd, 2025
Author(s): Cason Henderson

1. **Introduction**

- **Dataset Name:** Iris Dataset
- **Source:** Built-in dataset in R, originally introduced by Ronald A. Fisher in 1936.
- **Objective:** The purpose of this analysis is to explore the physical characteristics of iris flowers across three species being Setosa, Versicolor, and Virginica. We can also understand patterns and visualize trends using various data analysis techniques.

2. **Data Description**

- **Number of Observations:** 150 rows
- **Number of Variables:** 5 columns
- **Brief Description of Key Variables:**

Variable Name	Description	Data Type
Sepal.Length	Length of the sepals in centimeters	Numeric
Sepal.Width	Width of the sepals in centimeters	Numeric
Petal.Length	Length of the petals in centimeters	Numeric
Petal.Width	Width of the petals in centimeters	Numeric
Species	Species of the iris flower	Factor

- **Missing Values:** None because the data set is complete
- **Data Cleaning Steps Taken:** No additional cleaning was required as the iris dataset is pre-cleaned. However, we creating new variables in this set.

3. **Data Transformation**

- **New Variables Created (mutate() used):**

- **Petal.Area** – The product of Petal.Length and Petal.Width, representing the petal's area.
- **Sepal.Area** – The product of Sepal.Length and Sepal.Width, this represents the sepal's area.

- **Grouping & Aggregation (group_by() used):**
  - Grouped the data by **Species**.
  - The summary statistics computed include the mean for **Petal.Area**, **Sepal.Area**, and other key features such as Sepal.Length and Petal.Length.

4. Exploratory Data Analysis (EDA)

4.1 Summary Statistics

- **Mean, Median, Standard Deviation for Key Variables –** An example of the iris data set would be the mean Petal.Length ranging from approximately 1.5 cm (Setosa) to 5.5 cm (Virginica).
- **Outliers detected? -** Box plots of Sepal.Width suggested minor outliers for certain species, yet nothing major.

4.2 Visualizations (10 Plots Required)

1. Histogram of Petal Length by Species
2. Box Plot of Sepal Length by Species
3. Scatter Plot of Petal Length vs Petal Width
4. Bar Chart of Mean Sepal Area by Species
5. Density Plot of Sepal Width by Species
6. Line Plot of Sepal Length vs Sepal Width
7. Correlation Heatmap of numeric variables
8. Pairwise Scatterplot Matrix (Customized visualization)
9. Box Plot of Petal Area by Species
10. Histogram of Sepal Area by Species

5. Findings & Insights

- **Trends Observed:** Petal.Length and Petal.Width are significantly larger for Virginica compared to the other species. Sepal.Area tends to have lower variations across species, while Petal.Area shows noticeable differences.
- **Patterns & Anomalies:** Setosa species have notably smaller Petal.Length and Petal.Width compared to Versicolor and Virginica.
- **Surprising Observations:** A linear relationship exists between Petal.Length and Petal.Width, especially for Versicolor and Virginica species.

6. **Conclusions & Recommendations**

- **Final Summary of Findings:** The iris dataset highlights the physical characteristics of iris flowers, and displays accurate classification into their species. Clear trends in Petal.Area and Sepal.Area provide meaningful insights.
- **Possible Further Analysis:** Perform predictive modeling for data that relates like decision trees and random forests to classify iris species.
- **Potential Business/Research Recommendations:** These insights can help data scientists identify and categorize flower species in real-time using measurements and predictions.

7. **References**

Built-in R dataset of Iris.