Analysis of the Iris Dataset in RStudio

S.A.Abdulla - 22000021

Laboratry 2 Lab Sheet 14

February 27, 2025

Contents

1 Introduction

The Iris dataset is a well-known dataset in machine learning and statistics, containing 150 observations of iris flowers with features such as Sepal Length, Sepal Width, Petal Length, and Petal Width. The objective of this analysis is to explore the dataset, visualize key features, and perform hypothesis testing.

In this study, we will first conduct an exploratory data analysis (EDA) to understand the structure of the dataset, including summary statistics and visualizations. We will then analyze species distribution using various graphical techniques such as pie charts and bar plots. Additionally, histograms and scatter plots will be used to investigate feature distributions and correlations.

Furthermore, we will conduct hypothesis testing to examine specific statistical claims about the dataset. These tests will help us determine whether certain assumptions about Sepal Length, Sepal Width, and Petal Length hold true with statistical significance.

Beyond the scope of this analysis, the Iris dataset can be used for classification tasks in machine learning. Models such as decision trees, support vector machines, and neural networks can be applied to classify species based on their attributes. This makes the dataset a valuable resource for both statistical and machine learning applications.

Following is the codes and the process used to analyse the dataset.

2 Methodology

The following steps were taken to analyse the iris dataset with precision in R language

- Loaded the Iris dataset in RStudio.
- Viewed the loaded dataset
- Explored its structure using head, unique, str.
- Performed summary statistics to understand the dataset.
- Created visualizations such as pie charts, bar plots, histograms, and scatter plots.
- Conducted hypothesis tests lower-tail, upper-tail, and two-tailed tests on sepal and petal measurements.

3 Results

In this section, we separate the results into subgroups and analyze the dataset based on its specifications, one by one.

3.1 Summary Statistics

The dataset provides information on various features of the Iris flowers, including Sepal Length, Sepal Width, Petal Length, and Petal Width, along with the corresponding species of each flower. The summary statistics give us an overview of the distribution of these variables, which will help us understand the general characteristics of the data.

Below is an example of the dataset, which includes the first few observations:

Index	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**	**Species**
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Table 1: Example Structure of the Iris Dataset

The dataset consists of five columns: the measurements for Sepal Length, Sepal Width, Petal Length, and Petal Width and the Species (a categorical variable indicating the flower species). The summary statistics for these variables, as detailed earlier, provide insights into the central tendency (mean and median), distribution (min, max, quartiles), and variability of the data.

This table helps visualize the data points, and further analysis can be performed on these variables individually or in relation to one another. Understanding the nature of these attributes is essential for subsequent analyses, including classification and clustering tasks.

Given below is the summary of the dataset with Min, Max, Median

Statistic	Sepal Length	Sepal Width	Petal Length	Petal Width
Min.	4.300	2.000	1.000	0.100
1st Qu.	5.100	2.800	1.600	0.300
Median	5.800	3.000	4.350	1.300
Mean	5.843	3.057	3.758	1.199
3rd Qu.	6.400	3.300	5.100	1.800
Max.	7.900	4.400	6.900	2.500

Table 2: Summary Statistics of the Iris Dataset

```r
# Extract unique species from the iris dataset
m <- unique(iris$Species)
print(m)  # This will display the unique species
```

The output from the code is as follows:

```
[1] setosa versicolor virginica
```

```r
# Count the number of unique species
n <- length(unique(iris$Species))
print(paste("Number of unique species =", n))
```

```
[1] "Number of unique species = 3"
```

3.2 Visualizations

```r
library(ggplot2)

species_count <- table(iris$Species)
par(bg = "white")
# Background color for dark theme

pie(
    species_count,
    labels = names(species_count),
    col = rainbow(length(species_count)),
    main = "Species Distribution in Dataset"
)
```

The output of the code is the pie chart shown below:

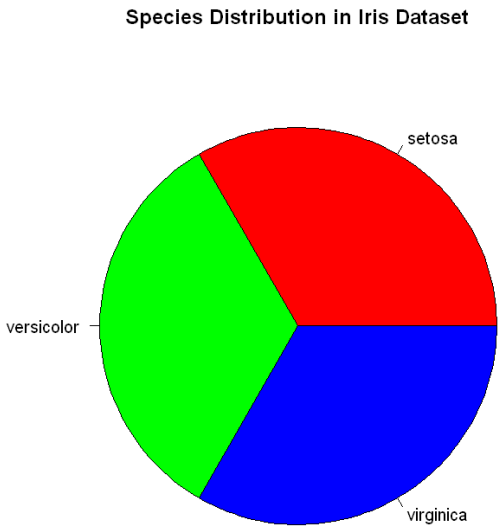

Figure 1: Pie Chart of Species Distribution in the Iris Dataset

The pie chart above visually represents the distribution of species in the Iris dataset. It clearly shows the proportion of each species (Setosa, Versicolor, and Virginica) within the dataset, with Setosa being the most prevalent species.

```
par(bg = "white")
#bacuse i used dark theme
ggplot(
    iris,
    aes(x = Species, fill = Species)) +
    geom_bar() +
    theme_minimal() +
    labs(title = "Count of Each Species", x = "
       Species", y = "Count"
)
```

The output of the code is the pie chart shown below:

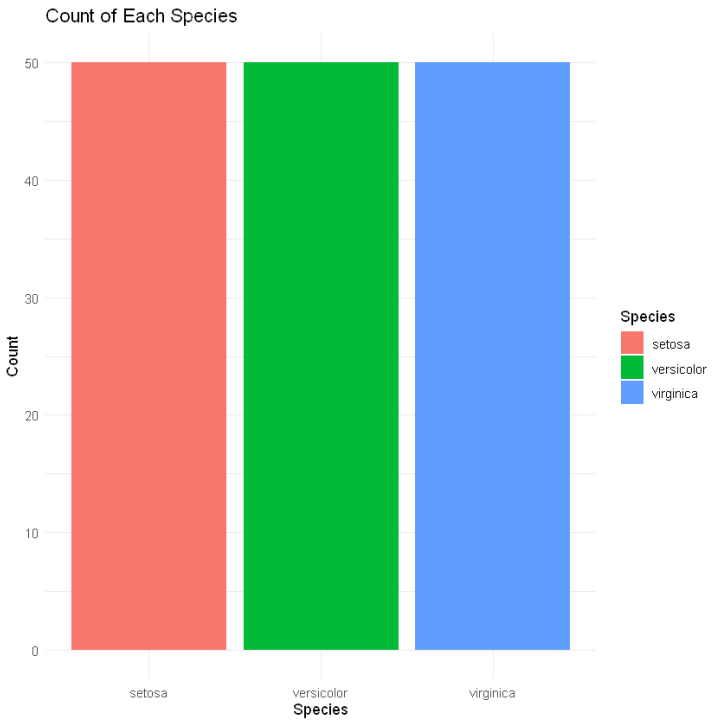

Figure 2: Pie Chart of Species Distribution in the Iris Dataset

The bar plot above illustrates the count of each species in the Iris dataset. As seen, the number of observations for each species is evenly distributed, with three species present in approximately equal quantities.

```r
par(bg = "white")
#bacuse i use dark theme
# Histogram for Sepal Length
ggplot(
    iris,
    aes(x = Sepal.Length)) +
    geom_histogram(
        binwidth = 0.5,
        fill = "blue",
        color = "black",
        alpha = 0.7) +
    theme_minimal() +
labs(title = "Histogram of Sepal Length", x = "
   Sepal Length", y = "Frequency")
```

The output of the code is the pie chart shown below:

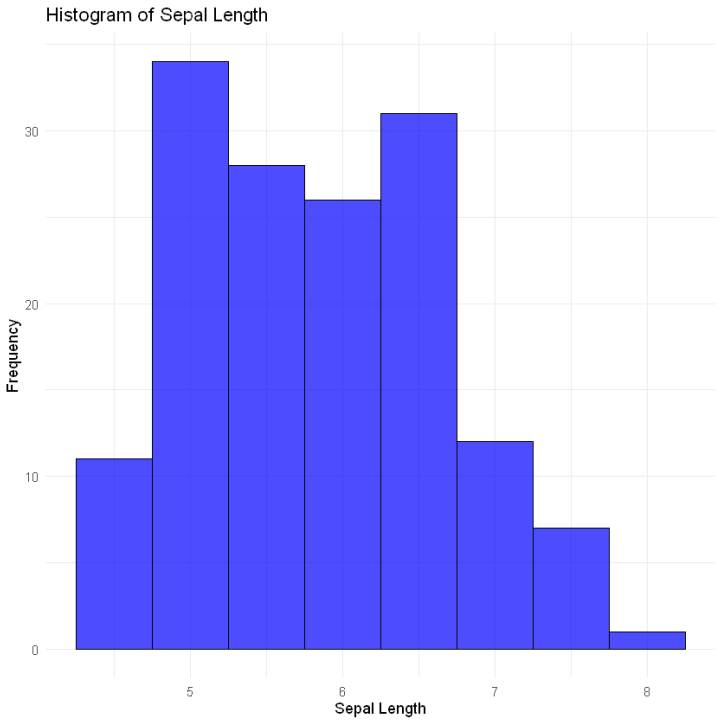

Figure 3: Pie Chart of Species Distribution in the Iris Dataset

This histogram displays the distribution of Sepal Length in the Iris dataset. The distribution appears roughly normal, with most Sepal Length values clustered around the mean value of 5.843.

```r
ggplot(iris, aes(x = Petal.Length)) +
geom_histogram(
    binwidth = 0.5,
    fill = "red",
    color = "black",
    alpha = 0.7) +
theme_minimal() +
labs(
    title = "Histogram of Petal Length",
    x = "Petal Length",
    y = "Frequency"
)
```

The output of the code is the pie chart shown below:

Figure 4: Pie Chart of Species Distribution in the Iris Dataset

The histogram above represents the distribution of Petal Length in the Iris dataset. It shows a right-skewed distribution, with Petal Length values mainly concentrated between 1 and 5.

```
        par(bg = "white")
        ggplot(iris, aes(x = Sepal.Length, y = Petal.
           Length, color = Species)) +
            geom_point() +
            theme_minimal() +
            labs(
                title = "Scatterplot of Sepal Length vs
                   Petal Length",
                x = "Sepal Length",
                y = "Petal Length"
            )
```

The output of the code is the pie chart shown below:

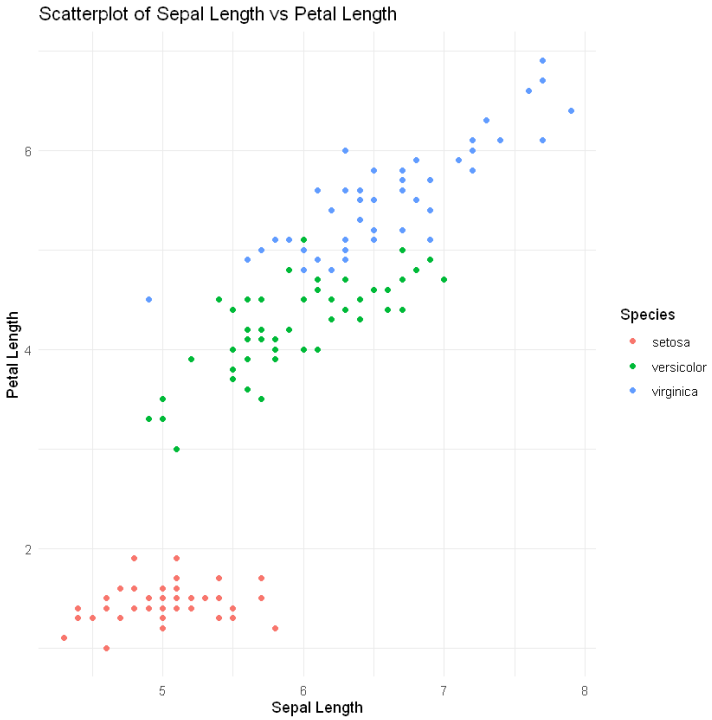

Figure 5: Pie Chart of Species Distribution in the Iris Dataset

The scatterplot above shows the relationship between Sepal Length and Petal Length. From this plot, we can observe a positive correlation between these two variables, where larger Sepal Length values tend to correspond to larger Petal Length values.

4 Discussion

The visualizations and statistical analysis provided valuable insights into the Iris dataset:

- The dataset consists of three species: Setosa, Versicolor, and Virginica, with equal distribution.
- Sepal Length and Petal Length exhibit a strong positive correlation.
- Hypothesis tests provided statistical confirmation regarding Sepal and Petal dimensions.

5 Conclusion

In conclusion, the analysis successfully explored the dataset through descriptive statistics, visualization, and hypothesis testing. Future work could involve implementing machine learning models for species classification.