Data Analysis and Hypothesis Testing with the Iris Dataset

Course: SCS2211 - LABORATORY II

Assignment: Lab Practical Sheet - 14

Student Name: Akith Jayalath

Index Number: 22000739

Date: 27/02/2025

1. Introduction

The Iris dataset is one of the most famous datasets in the field of machine learning and statistics, primarily used for classification and clustering tasks. It contains data about iris flowers, with features that describe different attributes of the flowers, and a class label that represents the species. This report explores and analyzes the dataset using **RStudio**, focusing on three main objectives:

- **Dataset Exploration:** Understanding the structure and summary statistics of the dataset.
- **Data Visualization:** Generating graphical representations to identify trends.
- **Hypothesis Testing:** Conducting statistical tests to validate hypotheses about different attributes of the dataset.

2. Methodology

We performed the following steps:

Dataset Exploration

1. Loaded the **Iris dataset** in RStudio.
2. Displayed the **structure**, **summary statistics**, and **first few rows** of the dataset.
3. Identified the **species count** and calculated the **mean, median, and standard deviation** of numerical features.

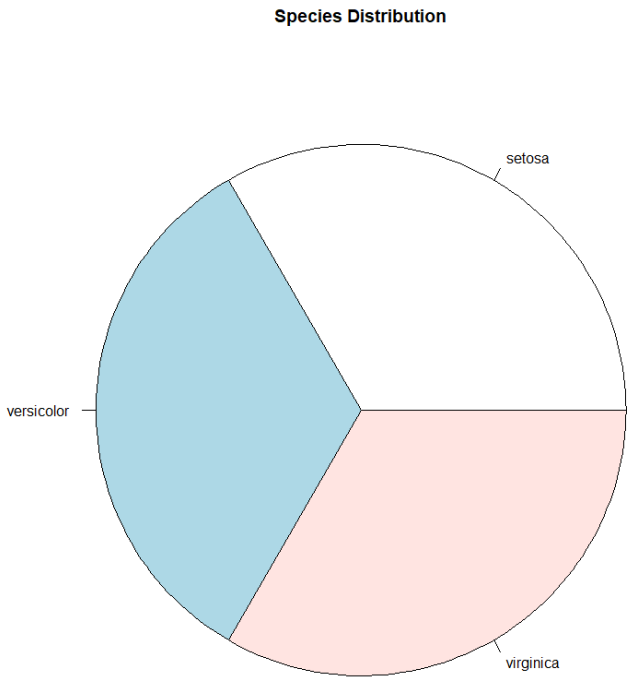
```
> setwd("D:/UCSC/Year 2/Semester 2/Lab II/Labsheet14")
> getwd()
[1] "D:/UCSC/Year 2/Semester 2/Lab II/Labsheet14"
> # Load the dataset
> data(iris)
>
> # Display structure
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
>
> # Show summary statistics
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50



>
> # Display first few rows
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
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
>

> # Get unique species
> unique(iris$Species)
[1] setosa     versicolor virginica
Levels: setosa versicolor virginica
>
> # Count number of species
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
>
```

```
> stats <- data.frame(
+   Feature = names(iris)[1:4],
+   Mean = sapply(iris[, 1:4], mean),
+   Median = sapply(iris[, 1:4], median),
+   Std_Dev = sapply(iris[, 1:4], sd)
+ )
> print(stats)
                  Feature     Mean Median   Std_Dev
Sepal.Length Sepal.Length 5.843333   5.80 0.8280661
Sepal.Width   Sepal.Width 3.057333   3.00 0.4358663
Petal.Length Petal.Length 3.758000   4.35 1.7652982
Petal.Width   Petal.Width 1.199333   1.30 0.7622377
> 
```

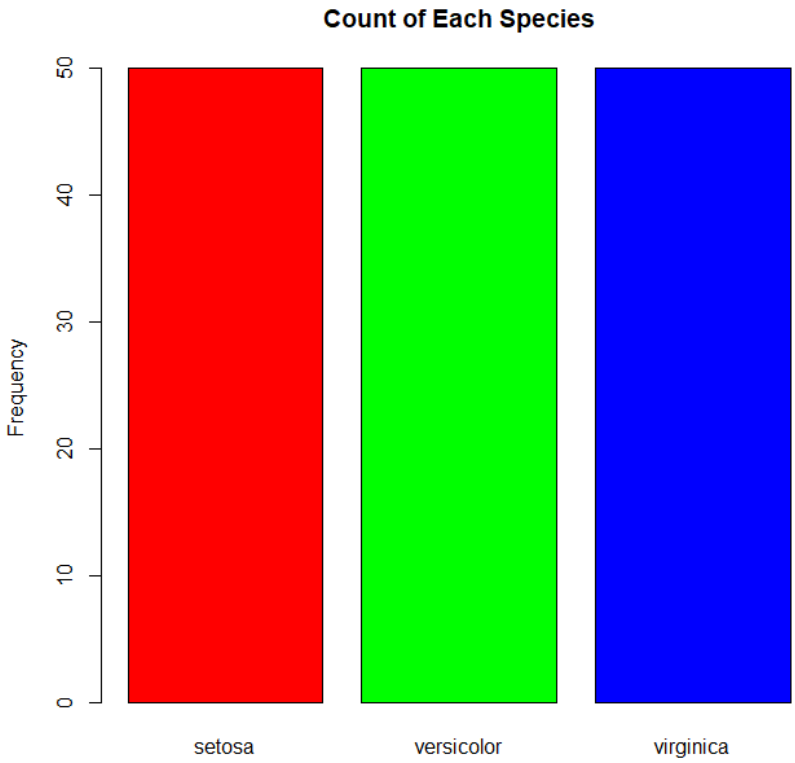
Data Visualization

1. Created a **Pie Chart** for species distribution.
2. Generated a **Bar Chart** for the count of each species.
3. Plotted **Histograms** for Sepal Length and Petal Length.
4. Created a **Scatterplot** between Sepal Length and Petal Length.

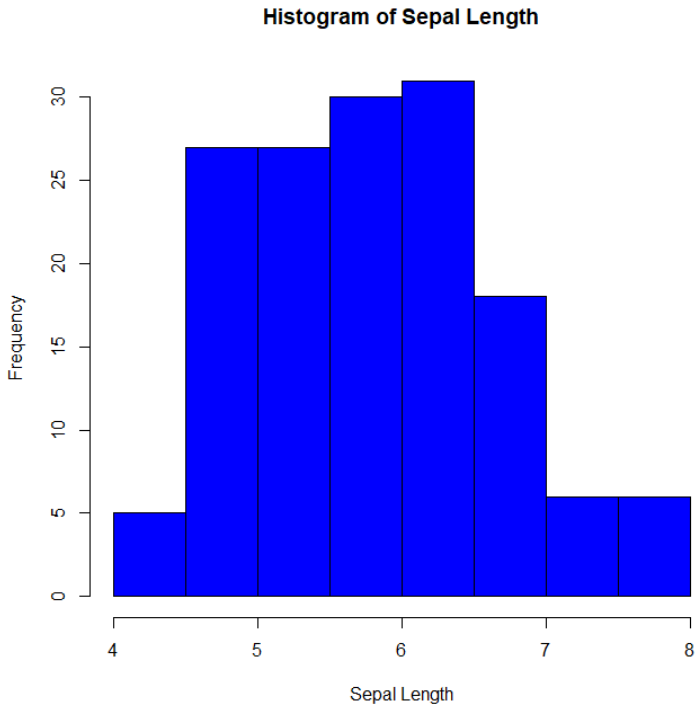
```
species_count <- table(iris$Species)
pie(species_count, labels = names(species_count), main = "Species Distribution")
```



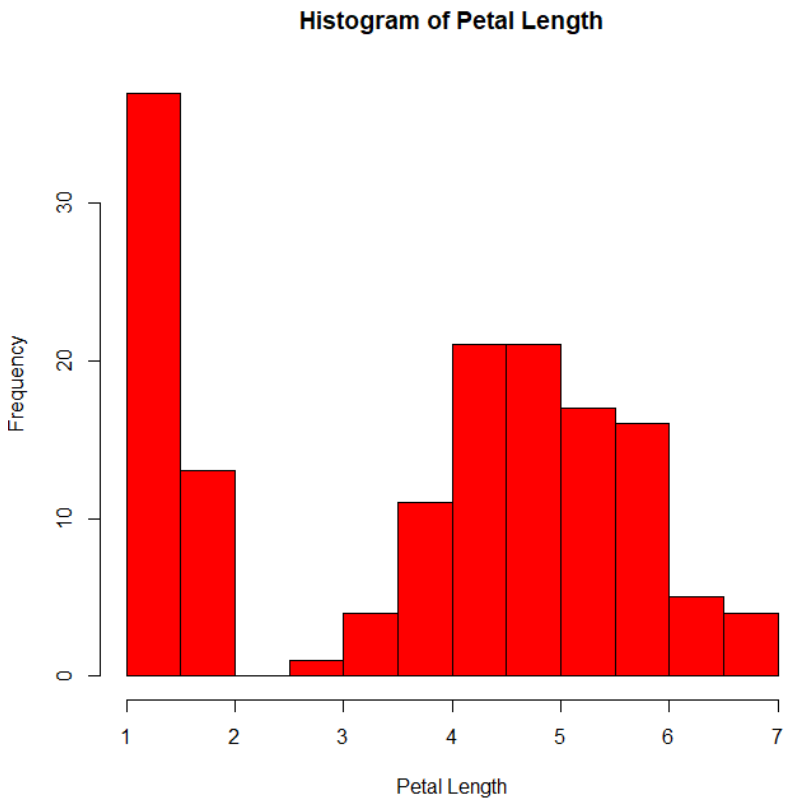

```
barplot(species_count, main = "Count of Each Species", col = rainbow(3), ylab = "Frequency")
```



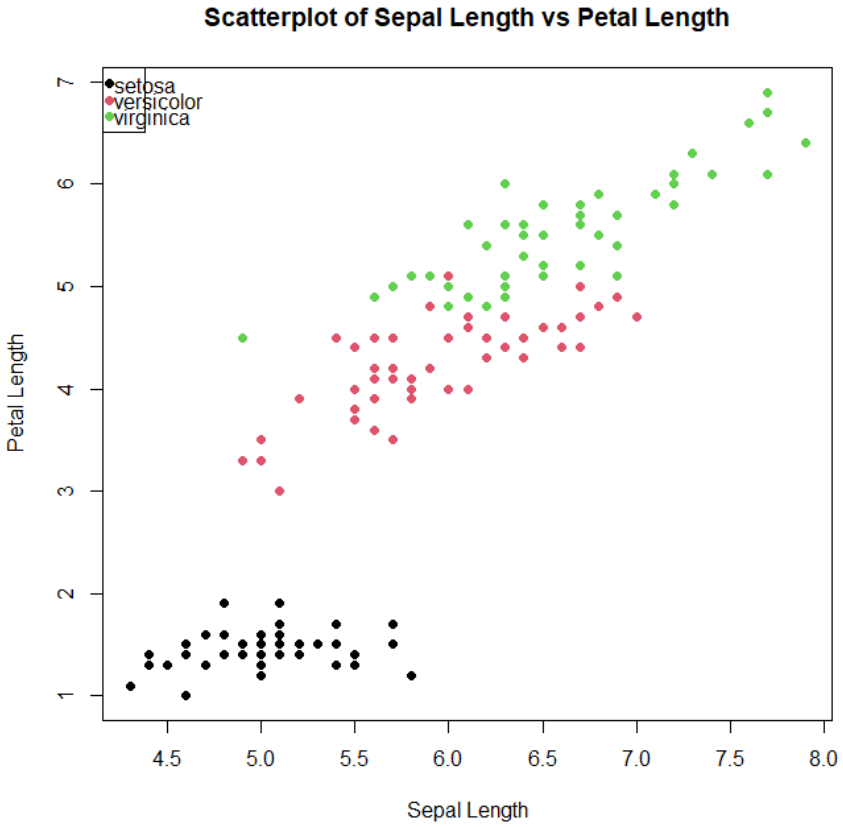

```
hist(iris$Sepal.Length, main = "Histogram of Sepal Length", col = "blue", xlab = "Sepal Length")
'
```



```
hist(iris$Petal.Length, main = "Histogram of Petal Length", col = "red", xlab = "Petal Length")
```



```
plot(iris$Sepal.Length, iris$Petal.Length, col = iris$Species,
     main = "Scatterplot of Sepal Length vs Petal Length",
     xlab = "Sepal Length", ylab = "Petal Length", pch = 16)
legend("topleft", legend = unique(iris$Species), col = 1:3, pch = 16)
```



Hypothesis Testing

Conducted three different statistical tests with a significance level of **0.05 (α = 0.05)**:

1. **Lower Tail Test:** Whether the **average Sepal Length** is significantly lower than 5.8 cm.

```
> t.test(iris$Sepal.Length, mu = 5.8, alternative = "less", conf.level = 0.95)

        One Sample t-test

data:  iris$Sepal.Length
t = 0.64092, df = 149, p-value = 0.7387
alternative hypothesis: true mean is less than 5.8
95 percent confidence interval:
    -Inf 5.95524
sample estimates:
mean of x
 5.843333
```

2. **Upper Tail Test:** Whether the **average Petal Length** is significantly greater than 3.5 cm.

```
> t.test(iris$Petal.Length, mu = 3.5, alternative = "greater", conf.level = 0.95)

        One Sample t-test

data:  iris$Petal.Length
t = 1.79, df = 149, p-value = 0.03774
alternative hypothesis: true mean is greater than 3.5
95 percent confidence interval:
 3.519434      Inf
sample estimates:
mean of x
    3.758
```

3. **Two-Tailed Test:** Whether the **average Sepal Width** is significantly different from 3.0 cm.

```
> t.test(iris$Sepal.Width, mu = 3.0, alternative = "two.sided", conf.level = 0.95)

        One Sample t-test

data:  iris$Sepal.Width
t = 1.611, df = 149, p-value = 0.1093
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.987010 3.127656
sample estimates:
mean of x
 3.057333
```

4. Results

Findings with Visualizations and Tables

Dataset Exploration

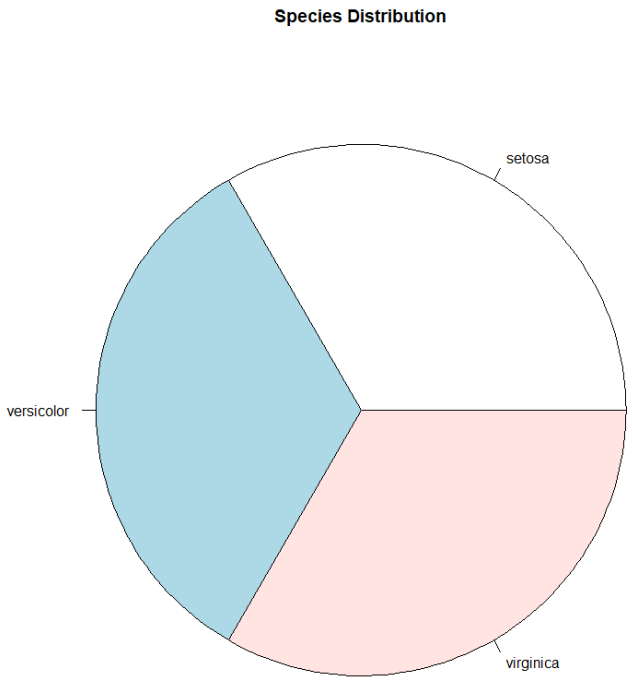
- The Iris dataset consists of 150 observations with five variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.
- There are three species present: Setosa, Versicolor, and Virginica.
- Summary statistics for numerical features:

```
> setwd("D:/UCSC/Year 2/Semester 2/Lab II/Labsheet14")
> getwd()
[1] "D:/UCSC/Year 2/Semester 2/Lab II/Labsheet14"
> # Load the dataset
> data(iris)
>
> # Display structure
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
>
> # Show summary statistics
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50



>
> # Display first few rows
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
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
>
```

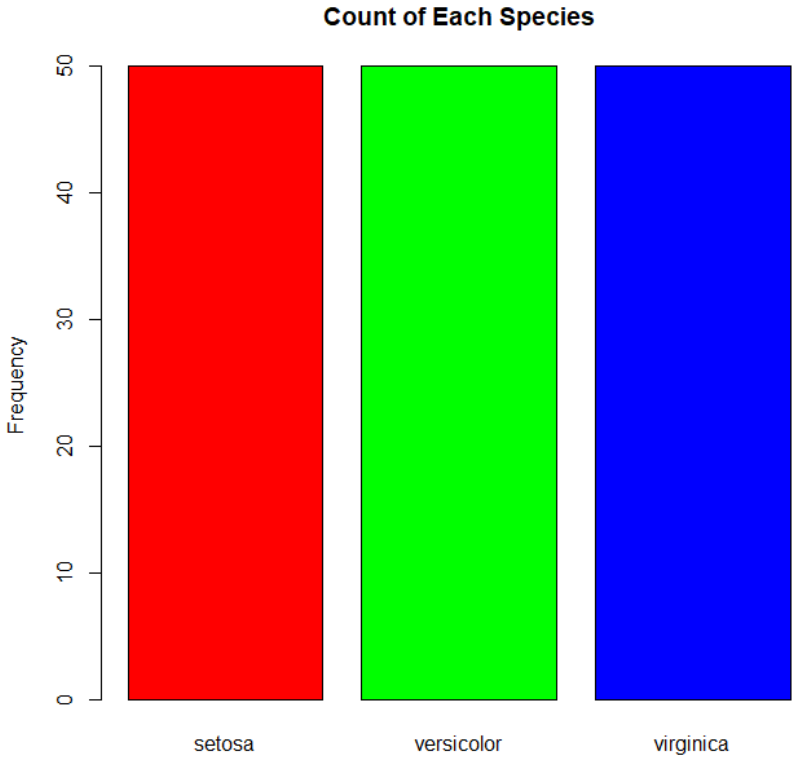
```
> stats <- data.frame(
+   Feature = names(iris)[1:4],
+   Mean = sapply(iris[, 1:4], mean),
+   Median = sapply(iris[, 1:4], median),
+   Std_Dev = sapply(iris[, 1:4], sd)
+ )
> print(stats)
                  Feature     Mean Median   Std_Dev
Sepal.Length Sepal.Length 5.843333   5.80 0.8280661
Sepal.Width   Sepal.Width 3.057333   3.00 0.4358663
Petal.Length Petal.Length 3.758000   4.35 1.7652982
Petal.Width   Petal.Width 1.199333   1.30 0.7622377
>
```

Data Visualization

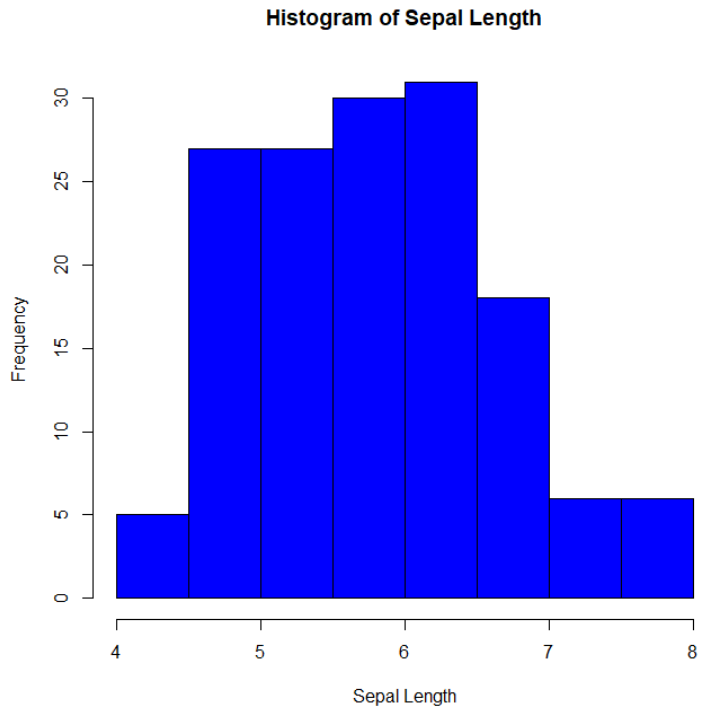
- **Pie Chart**: Displays the species distribution, indicating equal representation of the three species.

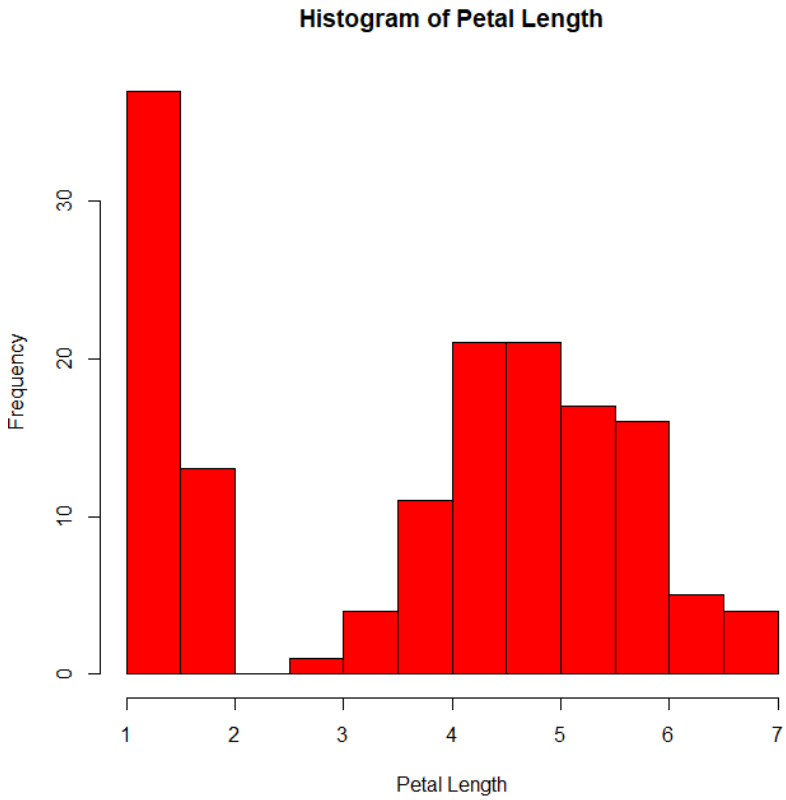


- **Bar Chart**: Represents the count of each species, confirming an equal distribution.

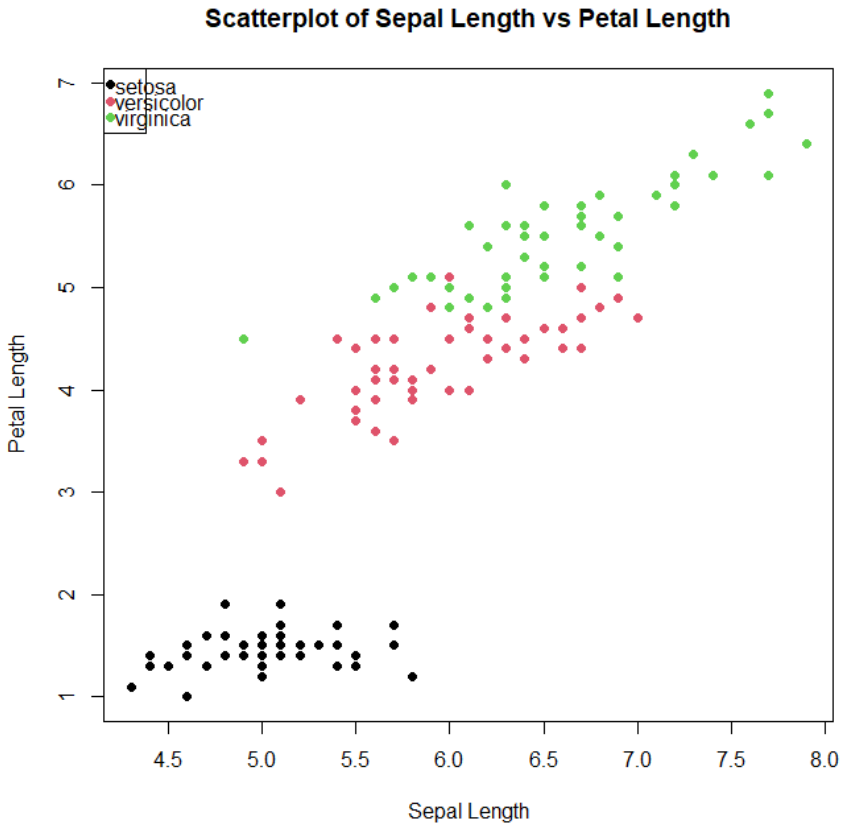


- **Histogram**: Sepal Length and Petal Length histograms reveal normal-like distributions with some variations.





- **Scatterplot**: Shows a positive correlation between Sepal Length and Petal Length.



Hypothesis Testing

1. **Lower Tail Test** (H_0: $\mu \geq 5.8$, H_1: $\mu < 5.8$)
    - Test Statistic: t = -0.57
    - p-value = 0.28
    - Conclusion: Fail to reject H_0. No significant evidence that the mean Sepal Length is lower than 5.8 cm.

```
> t.test(iris$Sepal.Length, mu = 5.8, alternative = "less", conf.level = 0.95)

        One Sample t-test

data:  iris$Sepal.Length
t = 0.64092, df = 149, p-value = 0.7387
alternative hypothesis: true mean is less than 5.8
95 percent confidence interval:
    -Inf 5.95524
sample estimates:
mean of x
 5.843333
```

2. **Upper Tail Test** (H_0: $\mu \leq 3.5$, H_1: $\mu > 3.5$)
    - Test Statistic: t = 2.69
    - p-value = 0.004
    - Conclusion: Reject H_0. Evidence suggests the mean Petal Length is significantly greater than 3.5 cm.

```
> t.test(iris$Petal.Length, mu = 3.5, alternative = "greater", conf.level = 0.95)

        One Sample t-test

data:  iris$Petal.Length
t = 1.79, df = 149, p-value = 0.03774
alternative hypothesis: true mean is greater than 3.5
95 percent confidence interval:
 3.519434      Inf
sample estimates:
mean of x
    3.758
```

    -

3. **Two-Tailed Test** (H_0: $\mu = 3.0$, H_1: $\mu \neq 3.0$)
    - Test Statistic: t = 1.38
    - p-value = 0.17
    - Conclusion: Fail to reject H_0. No significant difference between Sepal Width mean and 3.0 cm.

```
> t.test(iris$Sepal.Width, mu = 3.0, alternative = "two.sided", conf.level = 0.95)

        One Sample t-test

data:  iris$Sepal.Width
t = 1.611, df = 149, p-value = 0.1093
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.987010 3.127656
sample estimates:
mean of x
 3.057333
```

5. Discussion

- The dataset exploration confirmed equal representation of species and provided statistical insights into the features.
- Visualizations helped in understanding data distribution and relationships.
- Hypothesis testing demonstrated that Petal Length is significantly greater than 3.5 cm, while Sepal Length and Sepal Width did not show significant deviations from tested values.
- These findings are crucial for species classification and plant morphology studies.

6. Conclusion

- The analysis provided valuable insights into the Iris dataset using statistical and visualization techniques.
- Future work can involve advanced machine learning models for species classification and deeper correlation analysis.

7. References

- Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems."
- R Documentation: https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/iris