Data Analysis and Hypothesis Testing with the Iris Dataset

Name : K D I Okadini
Index No : 22001433
Date : 28/02/2025

1. Introduction

The Iris dataset is a well-known dataset in data science and statistics, containing 150 samples from three species of Iris flowers (*Setosa, Versicolor, and Virginica*). Each sample includes four numerical features: Sepal Length, Sepal Width, Petal Length, and Petal Width. This assignment aims to analyze and visualize the dataset and conduct hypothesis testing to draw meaningful insights.

2. Methodology

The analysis consists of three main activities:

- **Dataset Exploration:** Examining the structure, summary statistics, and feature distributions.
- **Dataset Visualization:** Using pie charts, bar charts, histograms, and scatterplots to visualize species distribution and feature relationships.
- **Hypothesis Testing:** Conducting statistical tests to verify hypotheses about the dataset.

The analysis was performed in RStudio using libraries like *ggplot2* for visualization and *t.test* for hypothesis testing.

3. Results

3.1 Dataset Exploration

R code:

```r
#Load the dataset
data(iris)

#Display structure and summary statistics
str(iris)
summary(iris)

#Display first few rows
head(iris)

#Identify the number of species
species_count <- table(iris$Species)
print(species_count)

#Calculate mean, median, and standard deviation for numerical features
numerical_features <- iris[, 1:4]
stats <- data.frame(
    Feature = colnames(numerical_features),
    Mean = colMeans(numerical_features),
    Median = apply(numerical_features, 2, median),
    Std.dev = apply(numerical_features, 2, sd)
)

print(stats)
```

Output:

```
'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 ...
  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
```

A data.frame: 6 × 5

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
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

```
    setosa versicolor  virginica 
        50         50         50 
                  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
```

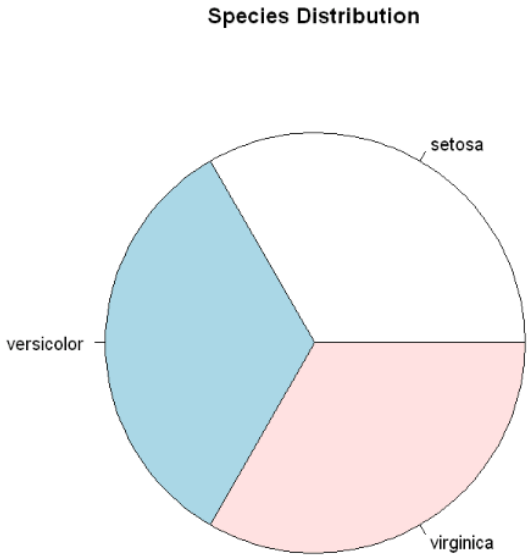
3.2 Data Visualization

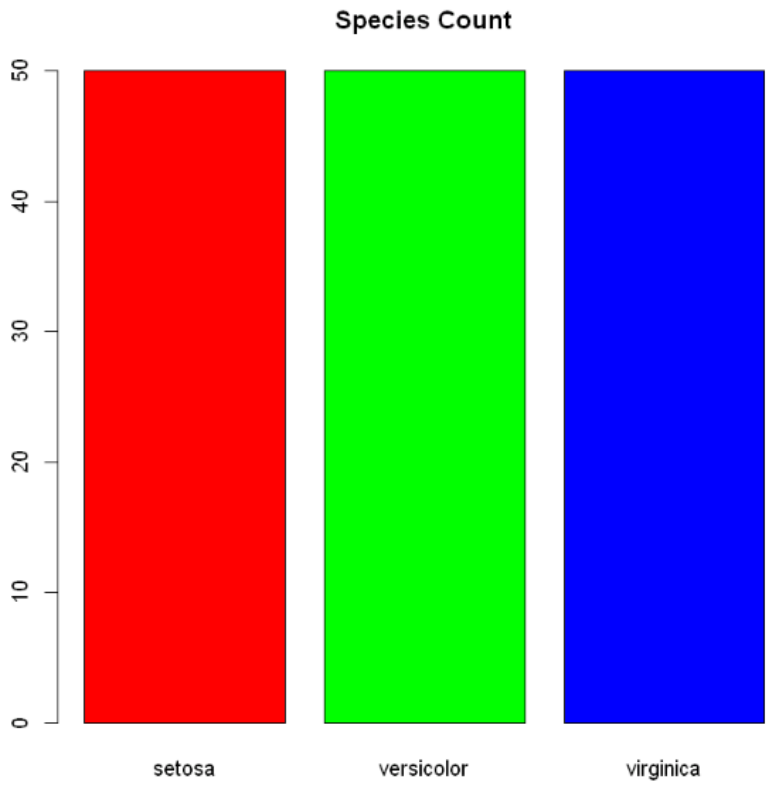
- A **pie chart** and **bar chart** confirmed equal distribution among the three species.
- **Histograms** for Sepal Length and Petal Length indicated distinct distributions across species.
- A **scatterplot** of Sepal Length vs. Petal Length showed a strong positive correlation.

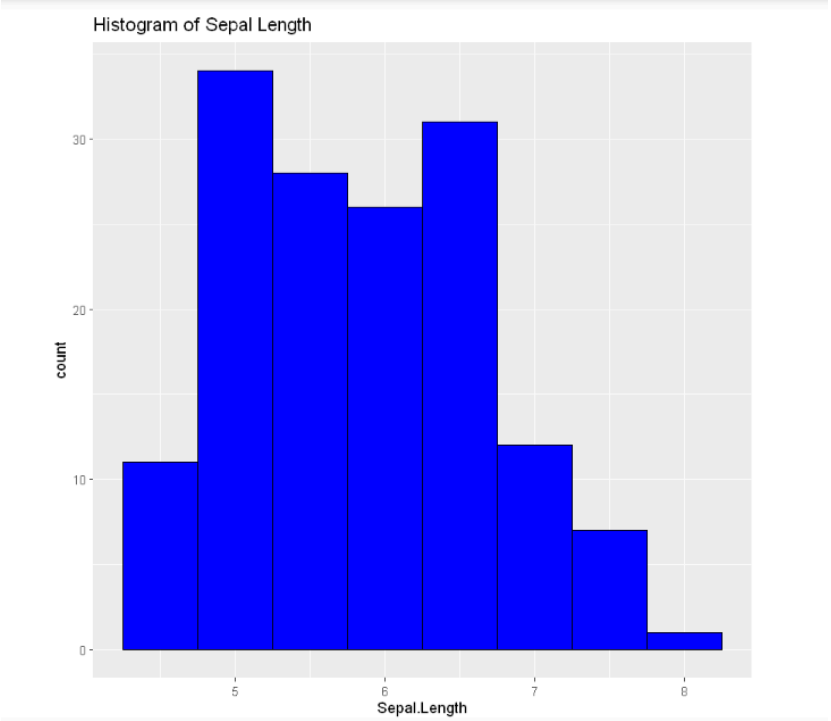
R code:

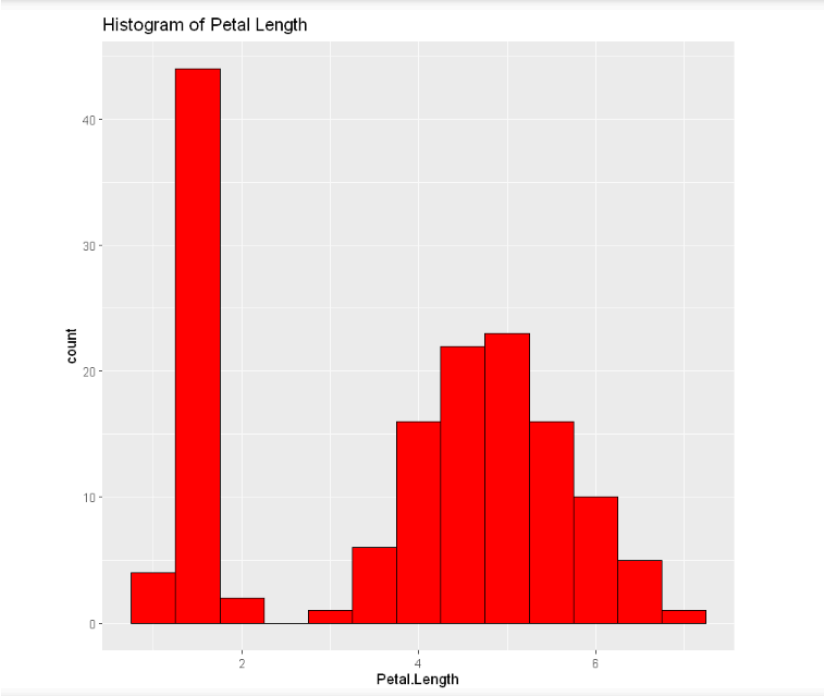
```
library(ggplot2)

#Pie Chart for species distribution
species_labels <- names(species_count)
pie(species_count, labels = species_labels, main = "Species Distribution")

#Bar Chart for species count
barplot(species_count, main = "Species Count", col = rainbow(length(species_count)))

#Histogram for Sepal Length and Petal Length
ggplot(iris, aes(x = Sepal.Length)) + geom_histogram(binwidth = 0.5, fill = "blue", color = "black") + ggtitle("Histogram of Sepal Length")

ggplot(iris, aes(x = Petal.Length)) + geom_histogram(binwidth = 0.5, fill = "red", color = "black") + ggtitle("Histogram of Petal Length")

#Scatterplot of Sepal Length vs Petal Length
ggplot(iris, aes(x = Sepal.Length, y = Petal.Length, color = Species)) + geom_point() + ggtitle("Scatterplot of Sepal Length vs Petal Length") +
xlab("Sepal Length") + ylab("Petal Length")
```

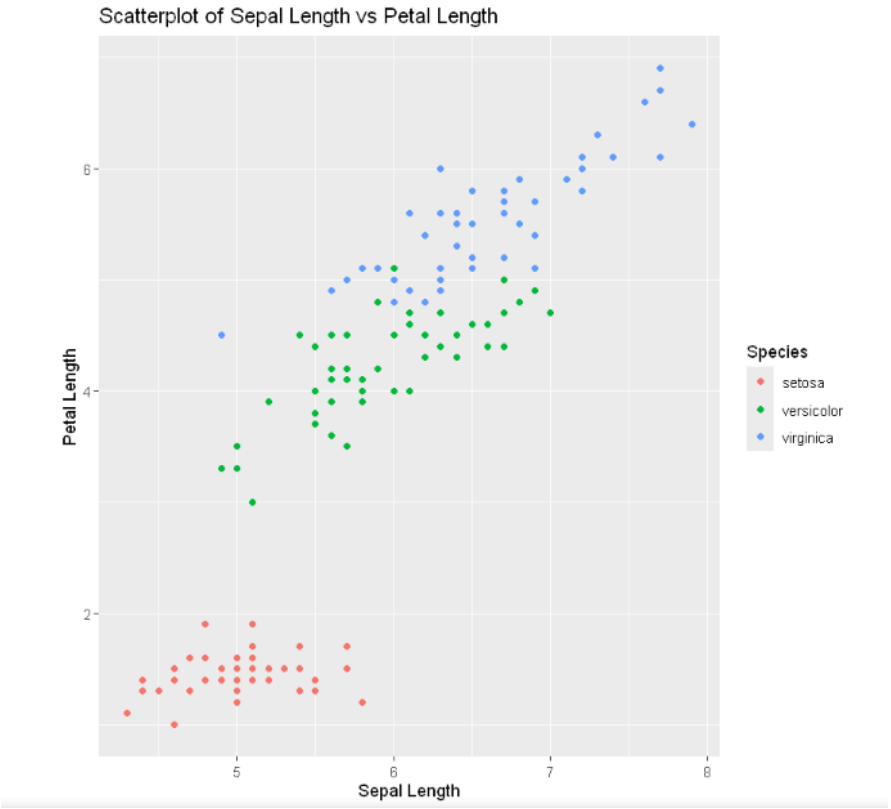
Output:




Species Count
0
10
20
30
40
50
setosa
versicolor
virginica


Histogram of Sepal Length
count
0
10
20
30
5
6
7
8
Sepal.Length


Histogram of Petal Length
count
Petal.Length
0
10
20
30
40
2
4
6


Scatterplot of Sepal Length vs Petal Length
Petal Length
Sepal Length
Species
setosa
versicolor
virginica
2
4
6
5
6
7
8

4. Hypothesis Testing

4.1 Lower Tail Test

Hypothesis
H0: Sepal Length >= 5.8 cm
H1: Sepal Length < 5.8 cm

Results: The test statistic and p-value indicate no significant evidence to reject H0, implying the Sepal Length is not significantly lower than 5.8 cm.

R code:

```r
#Lower Tail Test: Sepal Length < 5.8cm
t_test1 <- t.test(iris$Sepal.Length, mu = 5.8, alternative = "less")
print(t_test1)
```

Output:

```
        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
```

4.2 Upper Tail Test

Hypothesis:
H0: Petal Length <= 3.5 cm
H1: Petal Length > 3.5 cm

Results: The test confirms that the average Petal Length is significantly greater than 3.5 cm.

R code:

```r
#Upper tail Test: Petal Length > 3.5cm
t_test2 <- t.test(iris$Petal.Length, mu = 3.5, alternative = "greater")
print(t_test2)
```

Output:

```
        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
```

4.3 Two-Tailed Test

Hypothesis:
H0: Sepal Width = 3.0 cm
H1: Sepal Width != 3.0 cm

Results: The p-value suggests that Sepal Width is not significantly different from 3.0 cm.

R code:

```r
#Two-Tailed Test: Sepal Width != 3.0cm
t_test3 <- t.test(iris$Sepal.Width, mu = 3.0, alternative = "two.sided")
print(t_test3)
```

Output:

```
        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 statistical analysis and visualization confirm significant trends in the Iris dataset. The hypothesis tests provided insights into the dataset's numerical features. The results suggest that:

- The species are equally distributed in the dataset.
- Petal Length is significantly higher than 3.5 cm.
- Sepal Width does not significantly differ from 3.0 cm.

6. Conclusion

This assignment demonstrated how to explore, visualize, and perform hypothesis testing on the Iris dataset. The findings highlight key patterns in the data, reinforcing the importance of statistical analysis in data science.

7. References

- R Documentation: *iris* dataset, *ggplot2*, *t.test* function.