UNIVERSITY INSTITUTE of COMPUTING
Asia's Fastest Growing University

NAAC GRADE A+
Accredited University

PROJECT

Student Name: Aditya kumar UID: 24MCI10006

Branch: MCA-AIML Section/Group:1-A

Semester:1st Date of Performance:26-10-24

Subject Name: Statistical Technique Using R Subject Code: 24CAP-614

1.Aim:

To perform exploratory data analysis on the Iris dataset to understand the distribution of its variables, identify outliers, and visualize relationships between them.

2.Objective/problem definition:

- Understand the distribution of sepal and petal dimensions.
- Identify any outliers in the dataset.
- Visualize relationships between the features using various plots.

3.Programming language used:

R

4.Block diagram/design flow/flow chart

.Load Data

. Data Cleaning

.Exploratory Data Analysis

5.Algorithm or pseudo code

.Load necessary libraries (ggplot2, dplyr)

. Import the Iris dataset

. Explore the dataset structure and summary statistics

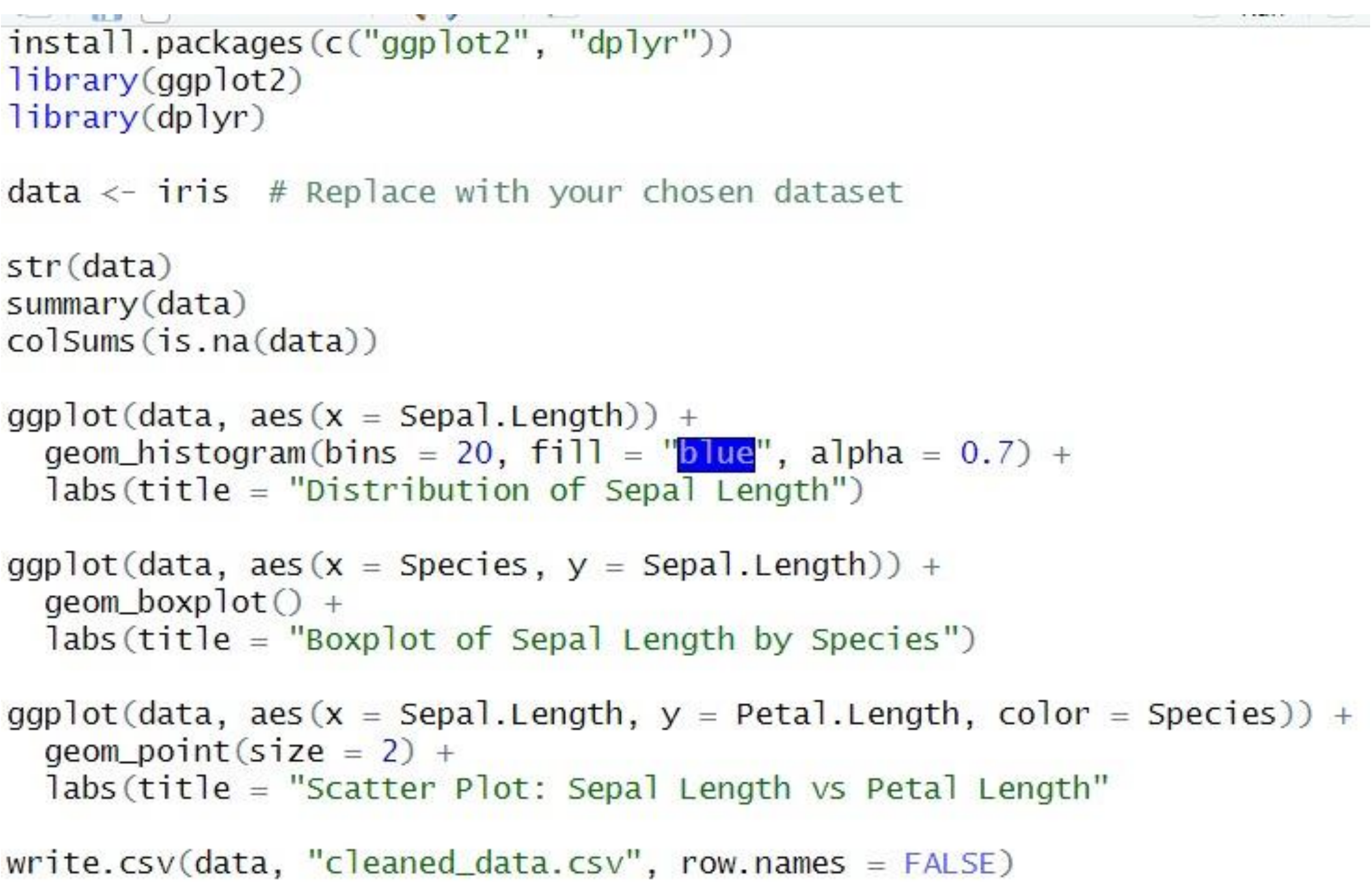
. Create histograms for each feature

. Generate boxplots to identify outliers

. Create scatter plots to visualize relationships between features

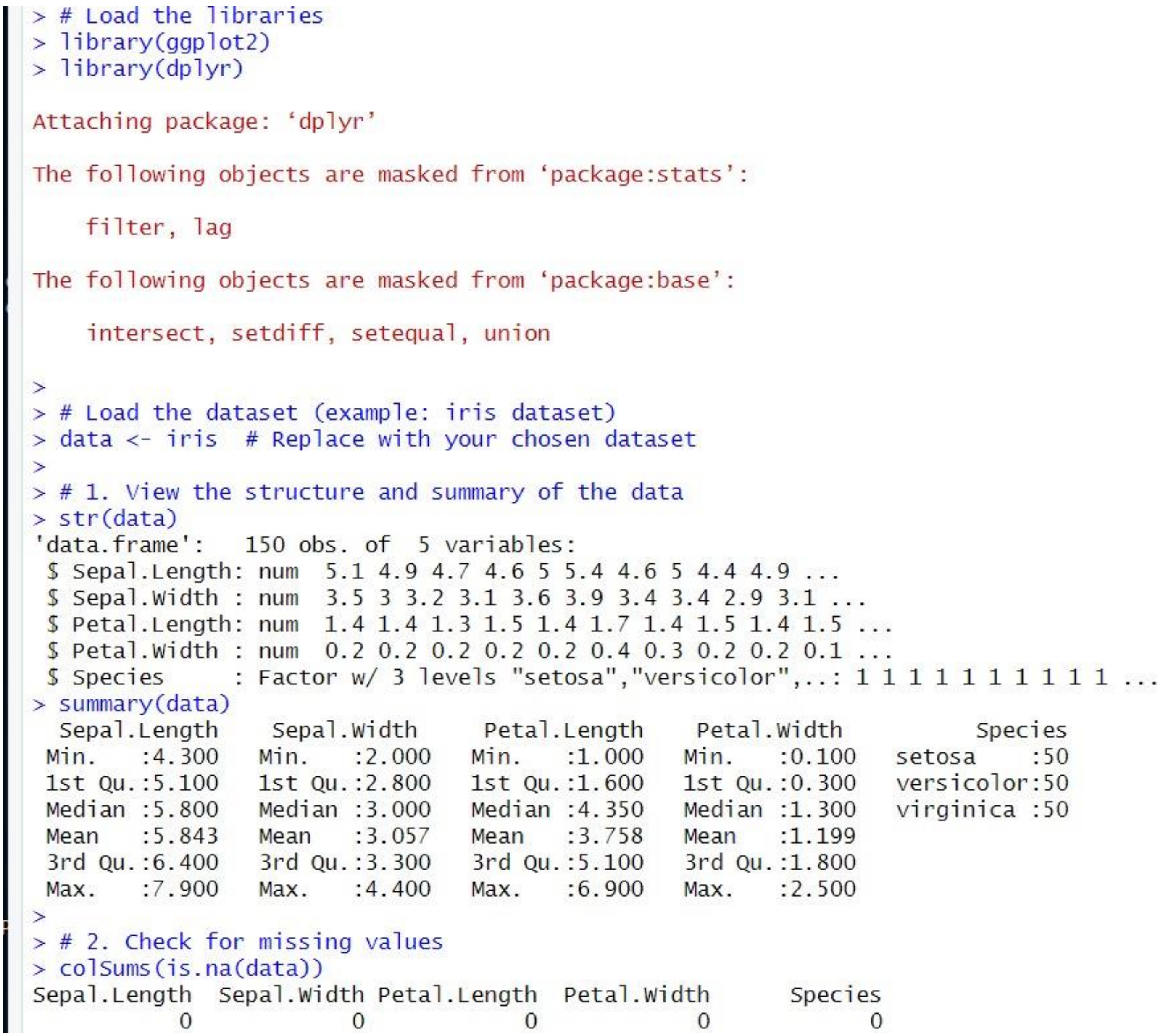
. Summarize findings

6.Implementation



```
install.packages(c("ggplot2", "dplyr"))
library(ggplot2)
library(dplyr)

data <- iris  # Replace with your chosen dataset

str(data)
summary(data)
colSums(is.na(data))

ggplot(data, aes(x = Sepal.Length)) +
  geom_histogram(bins = 20, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Sepal Length")

ggplot(data, aes(x = Species, y = Sepal.Length)) +
  geom_boxplot() +
  labs(title = "Boxplot of Sepal Length by Species")

ggplot(data, aes(x = Sepal.Length, y = Petal.Length, color = Species)) +
  geom_point(size = 2) +
  labs(title = "Scatter Plot: Sepal Length vs Petal Length"

write.csv(data, "cleaned_data.csv", row.names = FALSE)
```

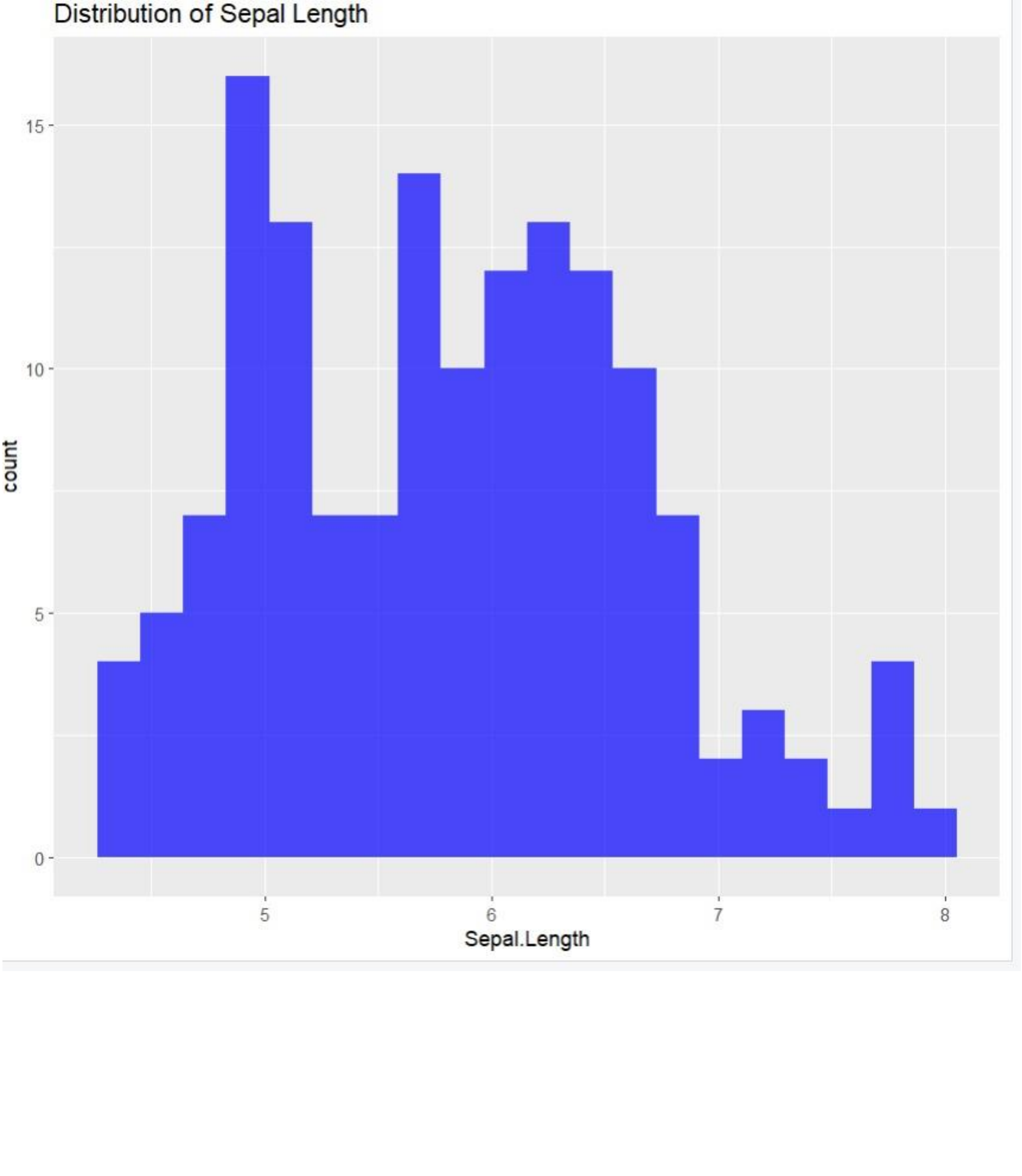
7..Output

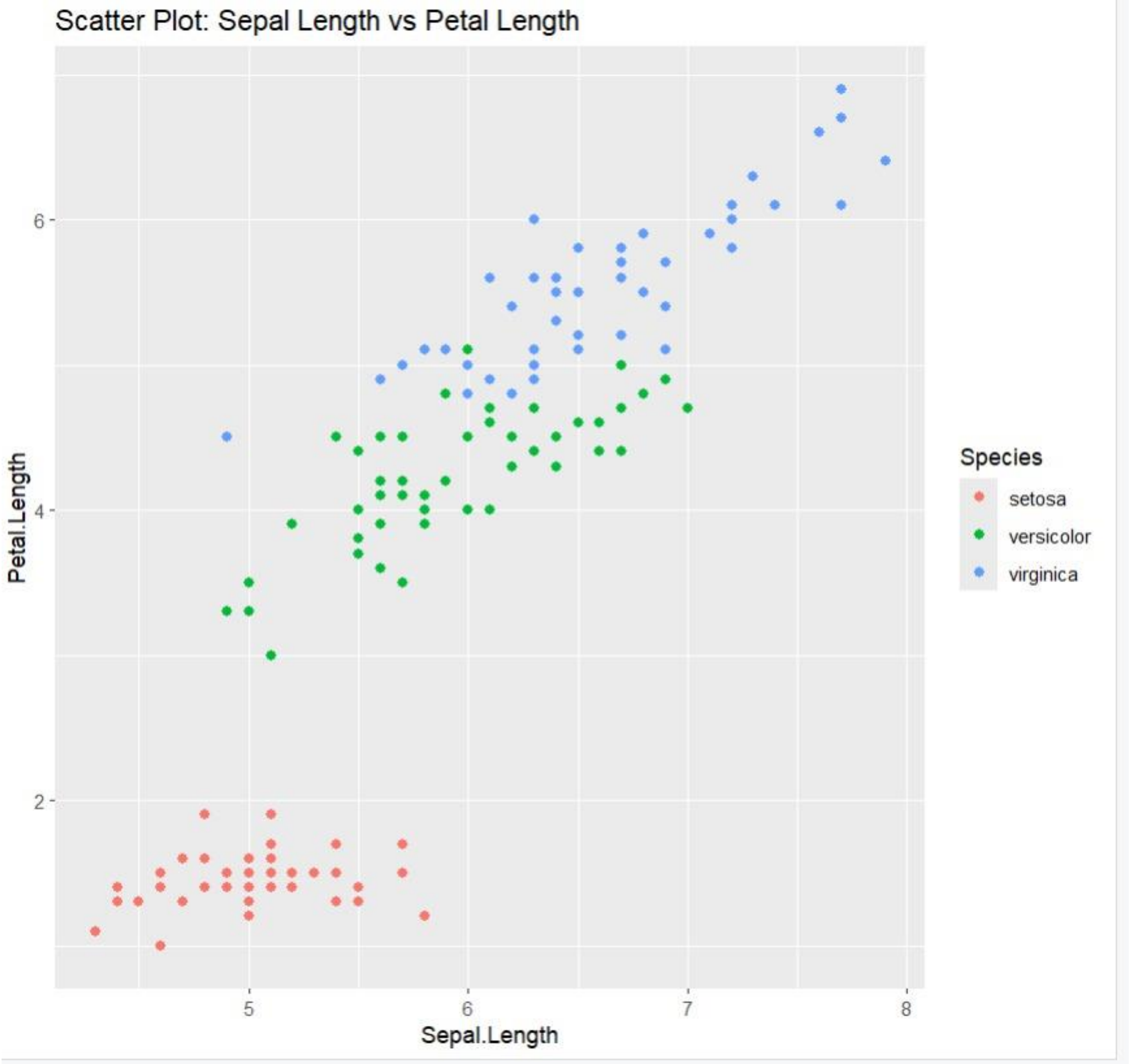


```
> # Load the libraries
> library(ggplot2)
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> 
> # Load the dataset (example: iris dataset)
> data <- iris  # Replace with your chosen dataset
> 
> # 1. View the structure and summary of the data
> str(data)
'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 ...
> summary(data)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 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                  
> 
> # 2. Check for missing values
> colSums(is.na(data))
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species 
           0            0            0            0            0 
```

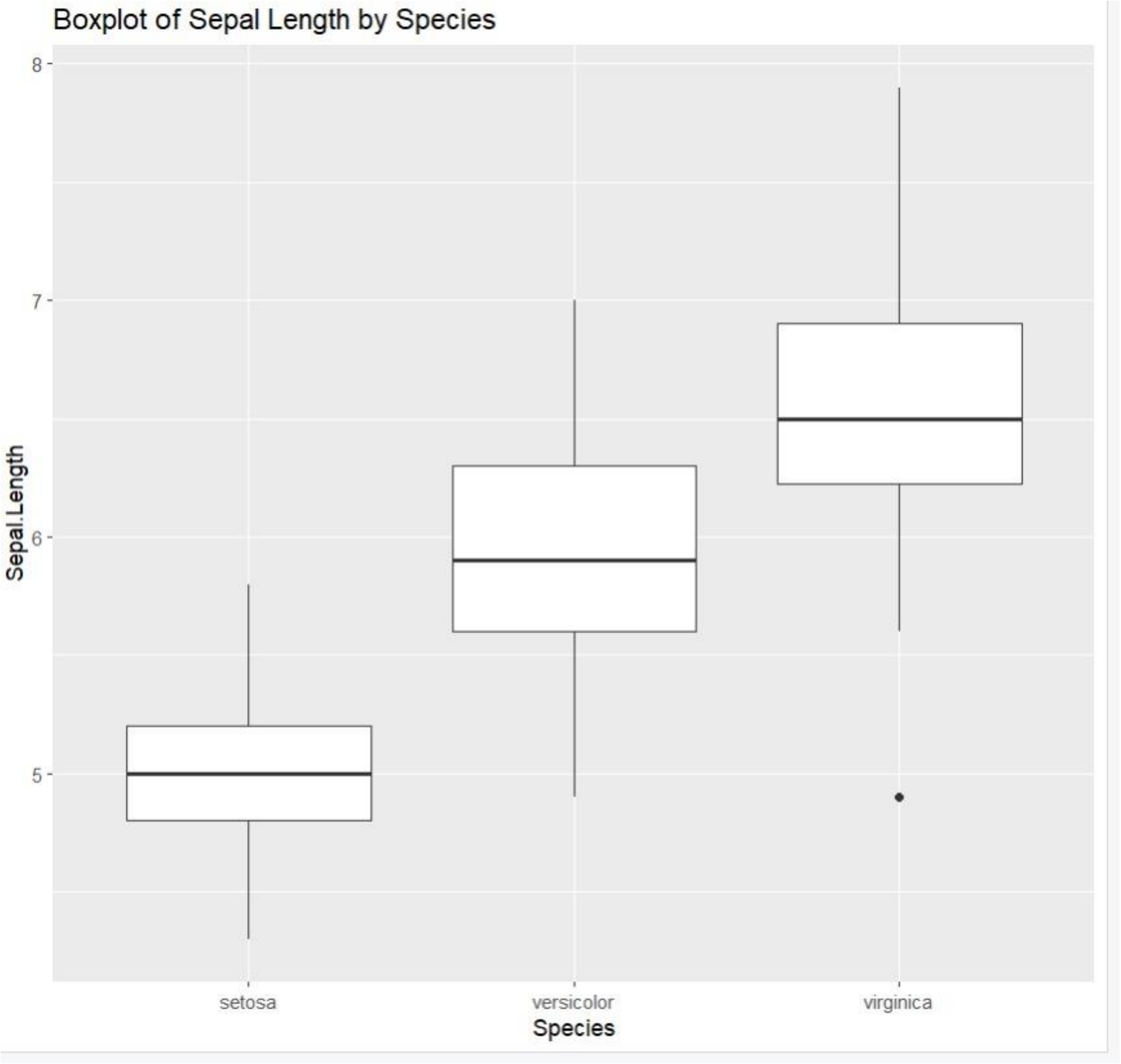
Data frames: In R, a data frame is a table-like structure used to store data. It is a list of vectors of equal length, allowing you to work with different types of data (numeric, character, factor, etc.) in a single object. Here's a quick overview of how to create and manipulate data frames in R:

GGplot2 : ggplot2 is an R package that provides a powerful and flexible framework for creating data visualizations based on the "grammar of graphics." Developed by Hadley Wickham, it allows users to build complex plots using a coherent system of layers and components.





Scatter Plot: Sepal Length vs Petal Length



8.**Conclusion:**

- Sepal and petal dimensions follow a normal distribution with some outliers present.
- Clear distinctions exist between the three species based on petal measurements.

9.Future Framework:

. Applying machine learning algorithms for classification based on the features.

. Expanding the dataset by including additional features or related datasets.

. Exploring temporal changes in species distribution if data is available.

10.Learning Outcomes

- Gained insights into data visualization techniques.
- Understood how to identify outliers and their implications.
- Developed skills in using R for data analysis and visualization

Sr. No.	Parameters	Marks Obtained	Maximum Marks
1.	Worksheet		8 Marks
2.	Viva		10 Marks
3.	Simulation		12 Marks
	Total		30 Marks