Analysis Of Data Distribution Of Iris Dataset

Student Name: Jaspreet Kaur	**UID: 24MCI10062**
Branch: MCA(AIML)	**Section/Group: 1-B**
Semester: 1st	**Date of Performance: 25-10-24**
Subject Name: R programming	**Subject Code: 24CAP-614**

Title : Exploratory Data Analysis of the Iris Dataset Using Boxplots to Identify Distribution and Outliers in Sepal and Petal Measurements.

Aim : The aim of this project is to perform exploratory data analysis (EDA) on the **Iris dataset** to visualize the distribution, spread, and presence of outliers in the sepal and petal measurements. This includes the use of boxplots and other visualizations to understand the relationships between the variables.

Task to be done:

- Load and explore the structure of the Iris dataset.
- Visualize the distribution of sepal and petal measurements using histograms.
- Create boxplots for each variable to detect outliers and assess data spread.
- Analyze relationships between variables using scatter plots.
- Label axes and provide descriptive titles for all plots and visualizations.
- Summarize key findings from the visualizations and statistics.

Question :

Choose a dataset from a repository like Kaggle or UCI Machine Learning Repository and perform exploratory data analysis using R. Explore the distribution of variables, identify outliers, and visualize relationships between variables using plots like histograms, scatter plots, and boxplots.

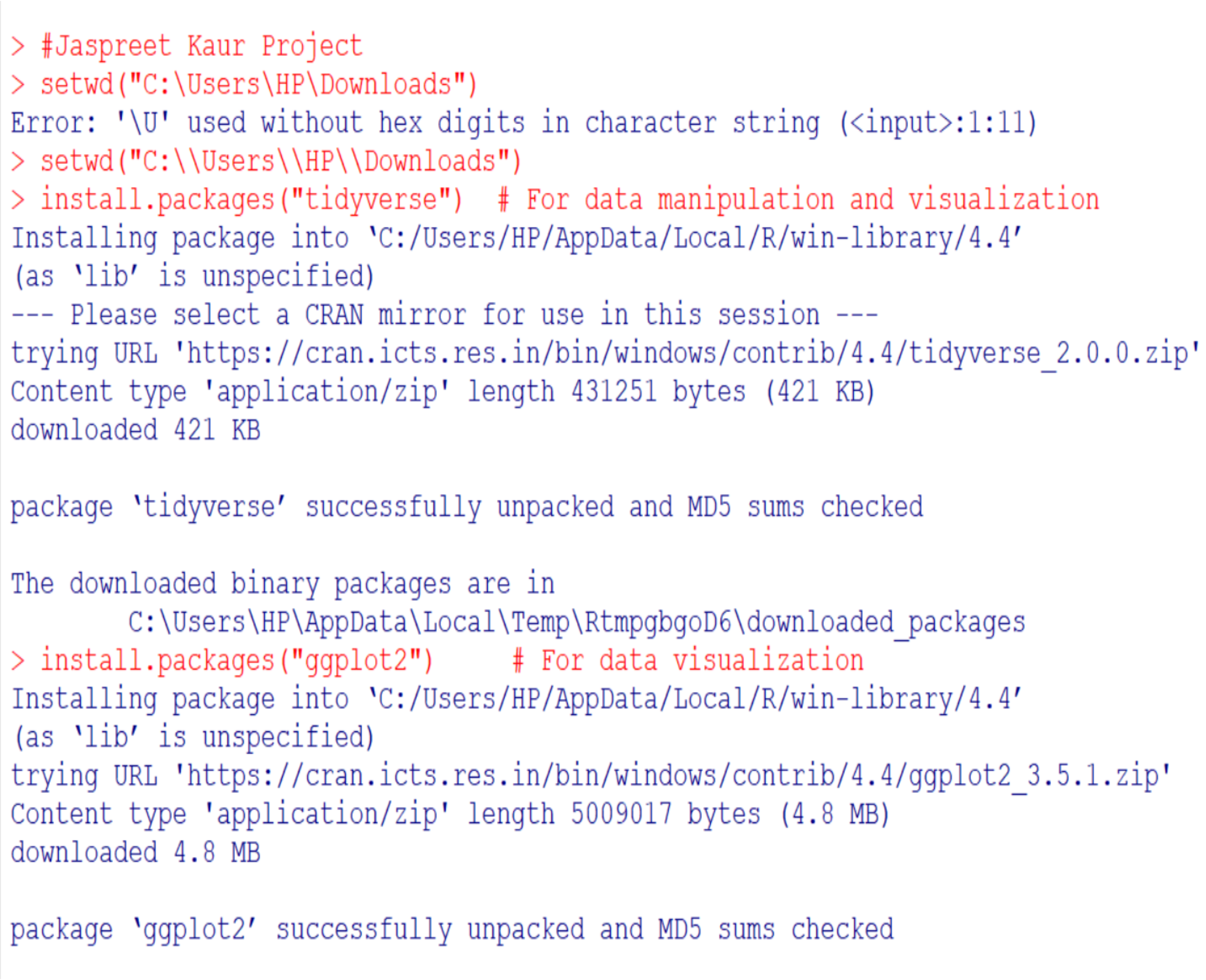
Code/Implementation :

Step 1: Choose a Dataset

- Go to **Kaggle** or the **UCI Machine Learning Repository**.
- Pick a dataset that interests you. For this guide, let’s use the **Iris dataset** (a popular dataset with four numerical variables: sepal length, sepal width, petal length, petal width, and one categorical variable: species).

Step 2:

- Set the working directory where your dataset is saved.
- Install and Load necessary packages:
  3

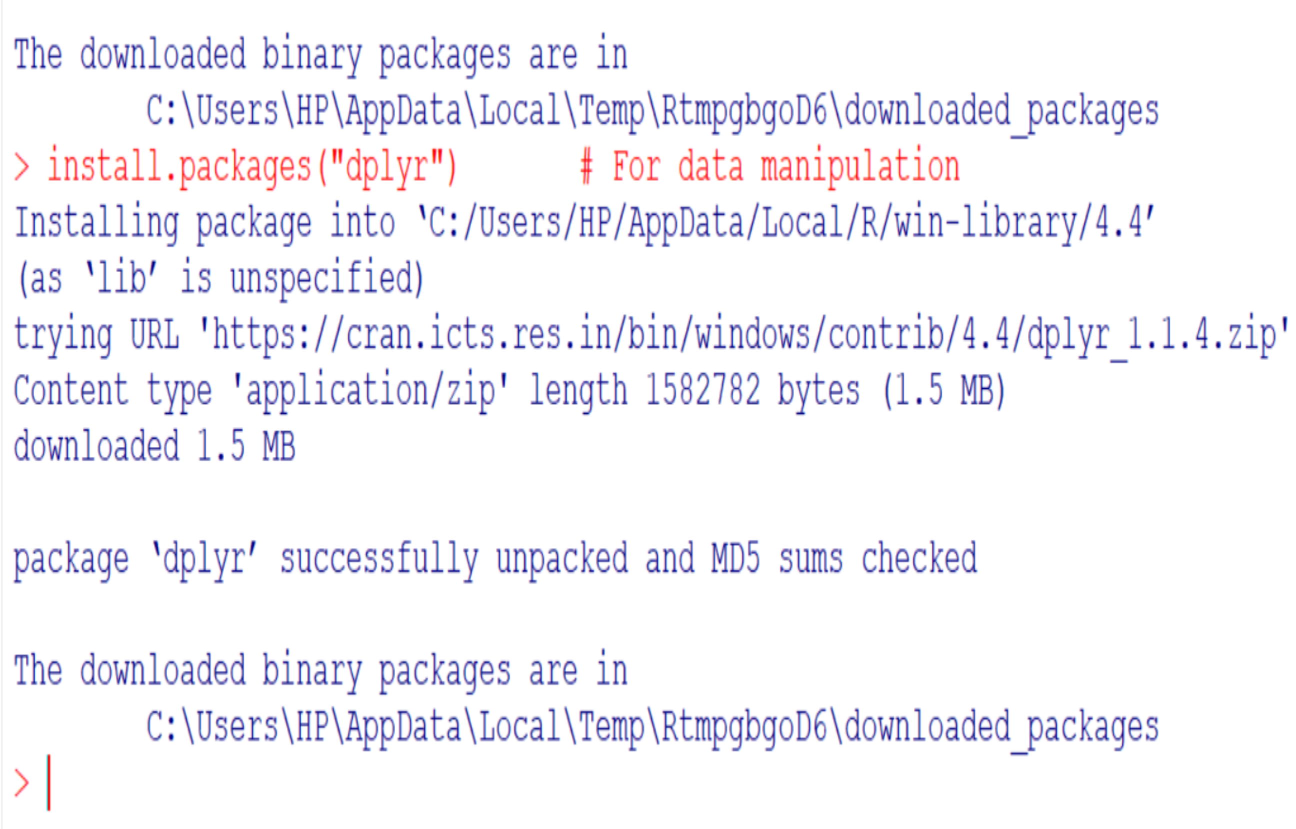


```
> #Jaspreet Kaur Project
> setwd("C:\Users\HP\Downloads")
Error: '\U' used without hex digits in character string (<input>:1:11)
> setwd("C:\\Users\\HP\\Downloads")
> install.packages("tidyverse")  # For data manipulation and visualization
Installing package into ‘C:/Users/HP/AppData/Local/R/win-library/4.4’
(as ‘lib’ is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cran.icts.res.in/bin/windows/contrib/4.4/tidyverse_2.0.0.zip'
Content type 'application/zip' length 431251 bytes (421 KB)
downloaded 421 KB

package ‘tidyverse’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\HP\AppData\Local\Temp\RtmpgbgoD6\downloaded_packages
> install.packages("ggplot2")     # For data visualization
Installing package into ‘C:/Users/HP/AppData/Local/R/win-library/4.4’
(as ‘lib’ is unspecified)
trying URL 'https://cran.icts.res.in/bin/windows/contrib/4.4/ggplot2_3.5.1.zip'
Content type 'application/zip' length 5009017 bytes (4.8 MB)
downloaded 4.8 MB

package ‘ggplot2’ successfully unpacked and MD5 sums checked
```



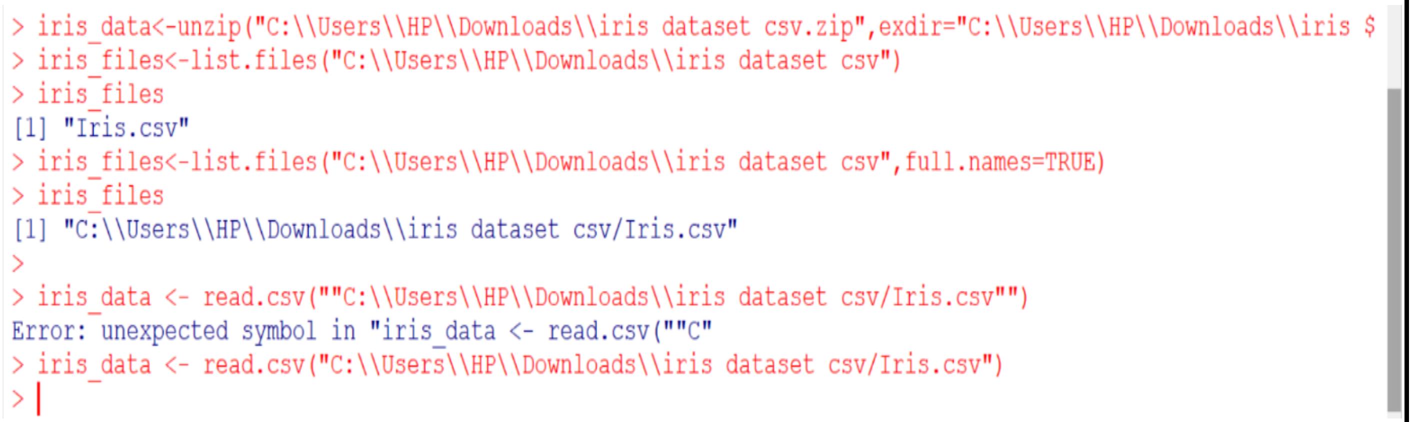
```
The downloaded binary packages are in
        C:\Users\HP\AppData\Local\Temp\RtmpgbgoD6\downloaded_packages
> install.packages("dplyr")       # For data manipulation
Installing package into ‘C:/Users/HP/AppData/Local/R/win-library/4.4’
(as ‘lib’ is unspecified)
trying URL 'https://cran.icts.res.in/bin/windows/contrib/4.4/dplyr_1.1.4.zip'
Content type 'application/zip' length 1582782 bytes (1.5 MB)
downloaded 1.5 MB

package ‘dplyr’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\HP\AppData\Local\Temp\RtmpgbgoD6\downloaded_packages
> 
```

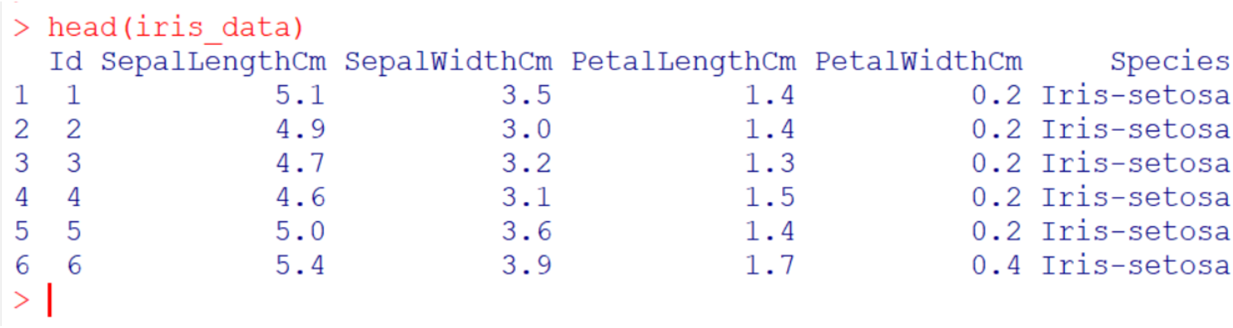
Step 3: Load the Dataset

Load the dataset into R using the read.csv() function.



```
> iris_data<-unzip("C:\\Users\\HP\\Downloads\\iris dataset csv.zip",exdir="C:\\Users\\HP\\Downloads\\iris $
> iris_files<-list.files("C:\\Users\\HP\\Downloads\\iris dataset csv")
> iris_files
[1] "Iris.csv"
> iris_files<-list.files("C:\\Users\\HP\\Downloads\\iris dataset csv",full.names=TRUE)
> iris_files
[1] "C:\\Users\\HP\\Downloads\\iris dataset csv/Iris.csv"
>
> iris_data <- read.csv(""C:\\Users\\HP\\Downloads\\iris dataset csv/Iris.csv"")
Error: unexpected symbol in "iris_data <- read.csv(""C"
> iris_data <- read.csv("C:\\Users\\HP\\Downloads\\iris dataset csv/Iris.csv")
> 
```

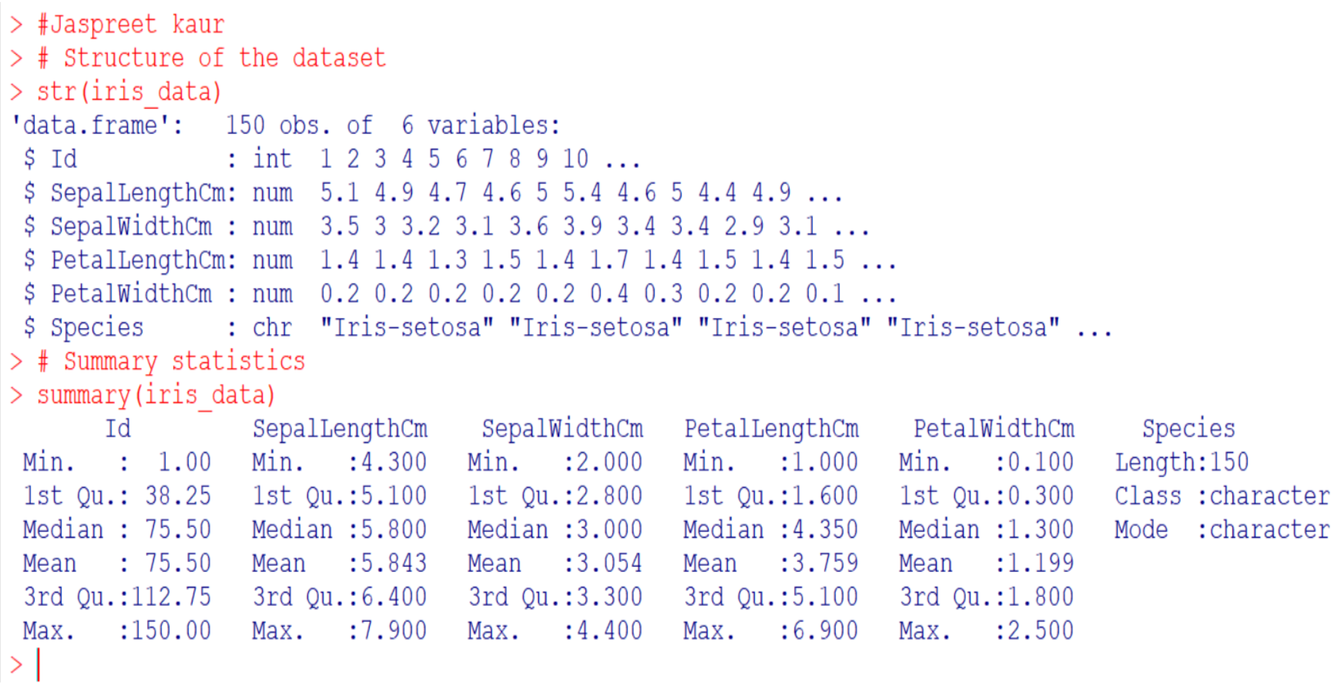
Check the first few rows to ensure the data has been loaded properly:



```
> head(iris_data)
  Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
1  1           5.1          3.5           1.4          0.2 Iris-setosa
2  2           4.9          3.0           1.4          0.2 Iris-setosa
3  3           4.7          3.2           1.3          0.2 Iris-setosa
4  4           4.6          3.1           1.5          0.2 Iris-setosa
5  5           5.0          3.6           1.4          0.2 Iris-setosa
6  6           5.4          3.9           1.7          0.4 Iris-setosa
> 
```

Step 4: Exploratory Data Analysis

□ str() gives you information about data types and structure.
□ summary() provides descriptive statistics (mean, median, min, max, etc.).

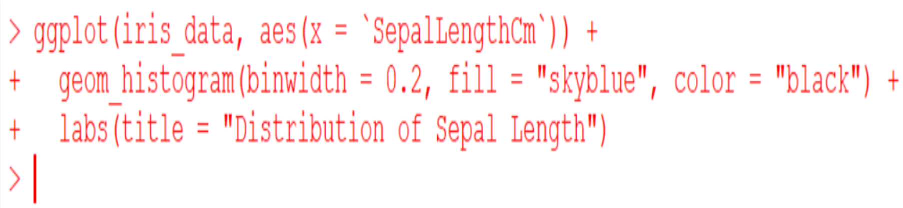


```
> #Jaspreet kaur
> # Structure of the dataset
> str(iris_data)
'data.frame':	150 obs. of  6 variables:
 $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SepalLengthCm: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ SepalWidthCm : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ PetalLengthCm: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ PetalWidthCm : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
> # Summary statistics
> summary(iris_data)
       Id         SepalLengthCm    SepalWidthCm   PetalLengthCm    PetalWidthCm     Species         
 Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Length:150        
 1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Class :character  
 Median : 75.50   Median :5.800   Median :3.000   Median :4.350   Median :1.300   Mode  :character  
 Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199                     
 3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                     
 Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                     
> |
```

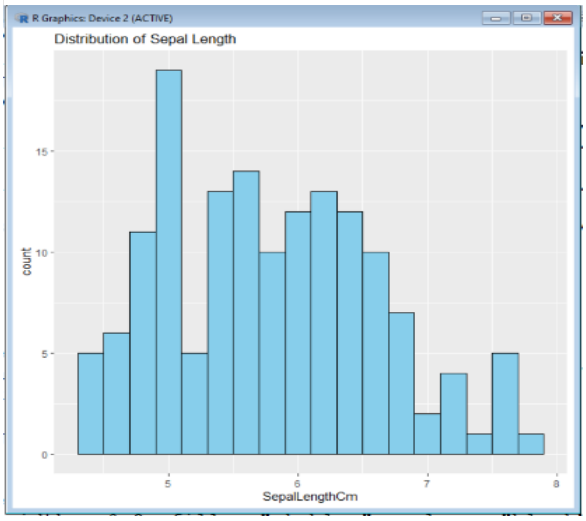
Step 5 : Distribution of Variables :

Start with histograms to visualize the distribution of numerical variables:

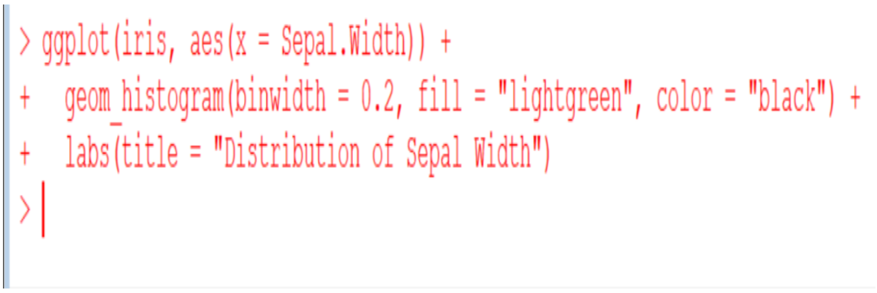
i. Histogram for Sepal Length:



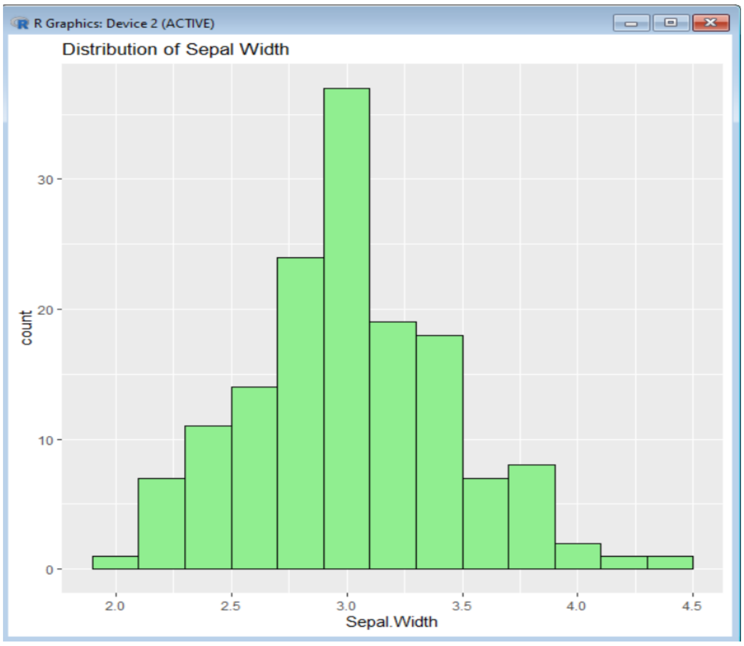
```
> ggplot(iris_data, aes(x = `SepalLengthCm`)) +
+   geom_histogram(binwidth = 0.2, fill = "skyblue", color = "black") +
+   labs(title = "Distribution of Sepal Length")
> |
```



ii. Histogram of Sepal Width :

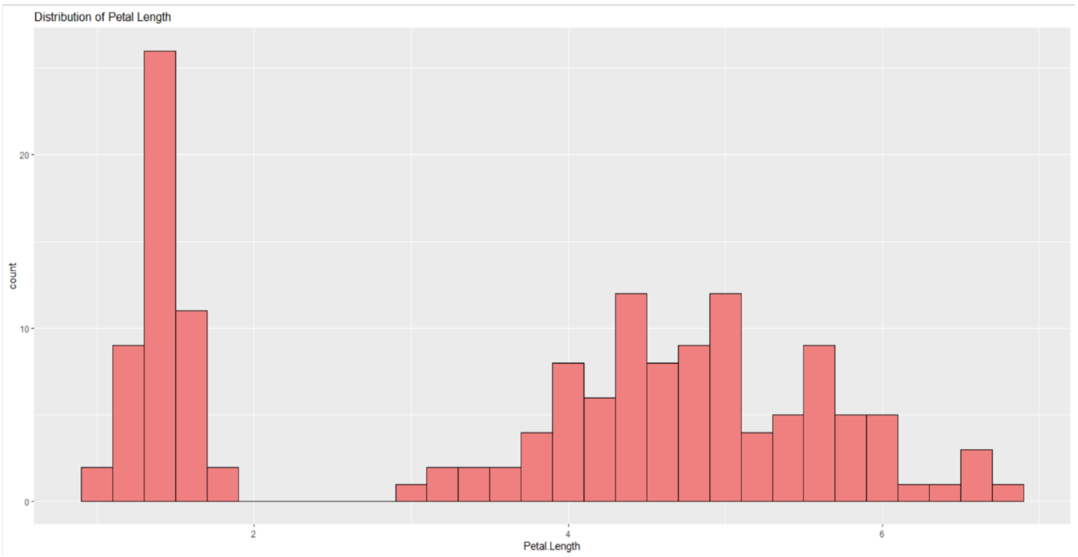


```
> ggplot(iris, aes(x = Sepal.Width)) +
+   geom_histogram(binwidth = 0.2, fill = "lightgreen", color = "black") +
+   labs(title = "Distribution of Sepal Width")
> 
```



iii. Histogram of Petal Length :

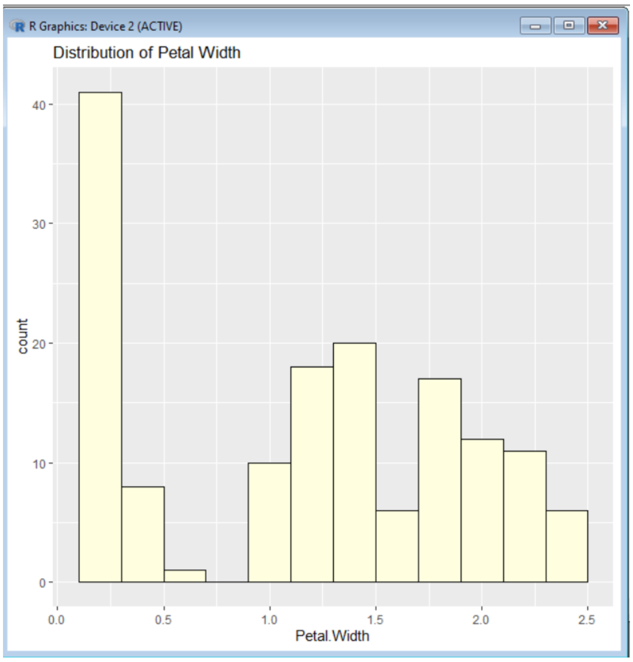
```
ggplot(iris, aes(x = Petal.Length)) +
+   geom_histogram(binwidth = 0.2, fill = "lightcoral", color = "black") +
+   labs(title = "Distribution of Petal Length")
```



iv.

Histogram of Petal Width :

```
ggplot(iris, aes(x = Petal.Width)) +
+   geom_histogram(binwidth = 0.2, fill = "lightyellow", color = "black") +
+   labs(title = "Distribution of Petal Width")
```

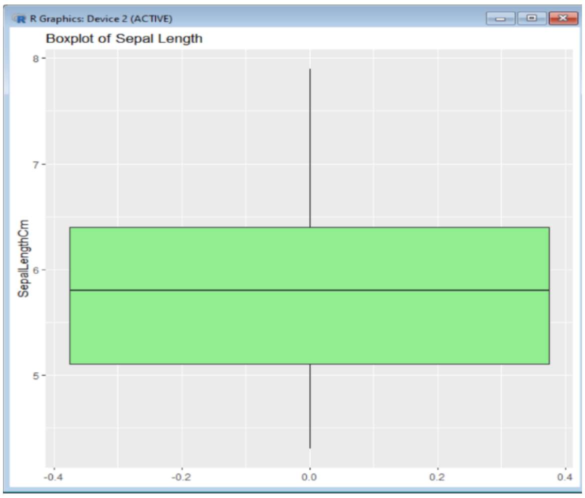


Step 6: Boxplot to Identify Outliers:

Boxplots are useful to identify outliers in the dataset.

Boxplot for Sepal Length :

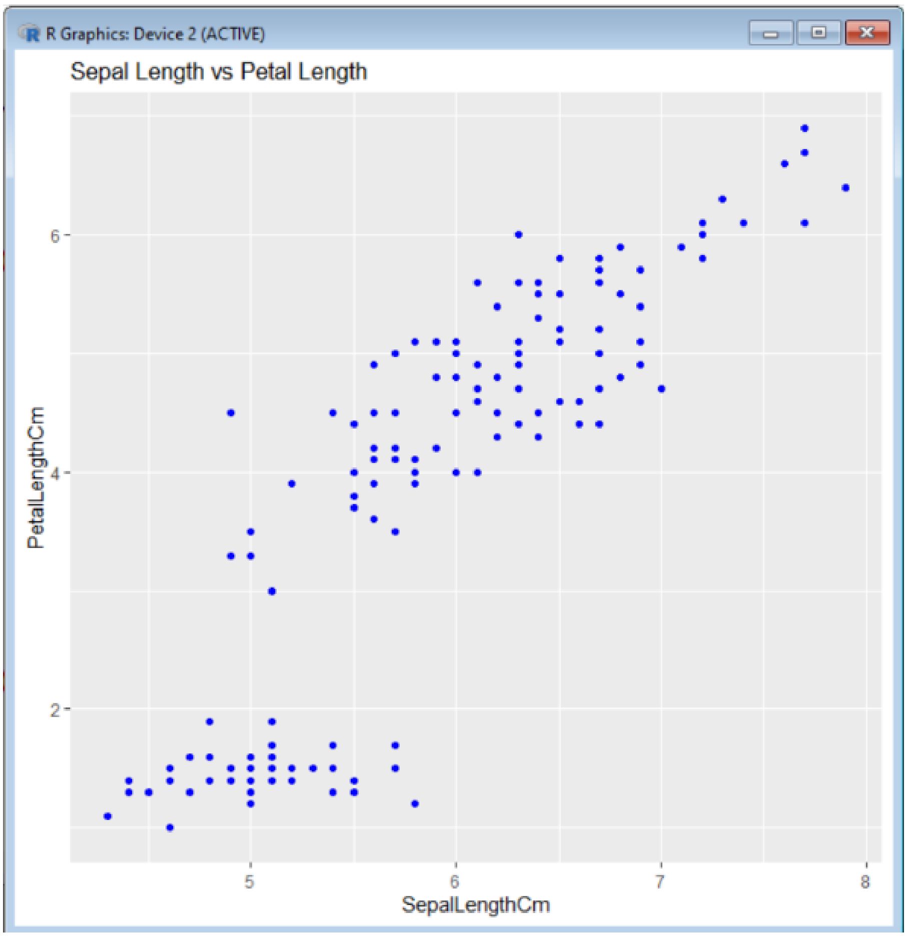
```
ggplot(iris_data, aes(y = SepalLengthCm)) +
+   geom_boxplot(fill = "lightgreen") +
+   labs(title = "Boxplot of Sepal Length")
```



Step 7: Relationship Between Variables:

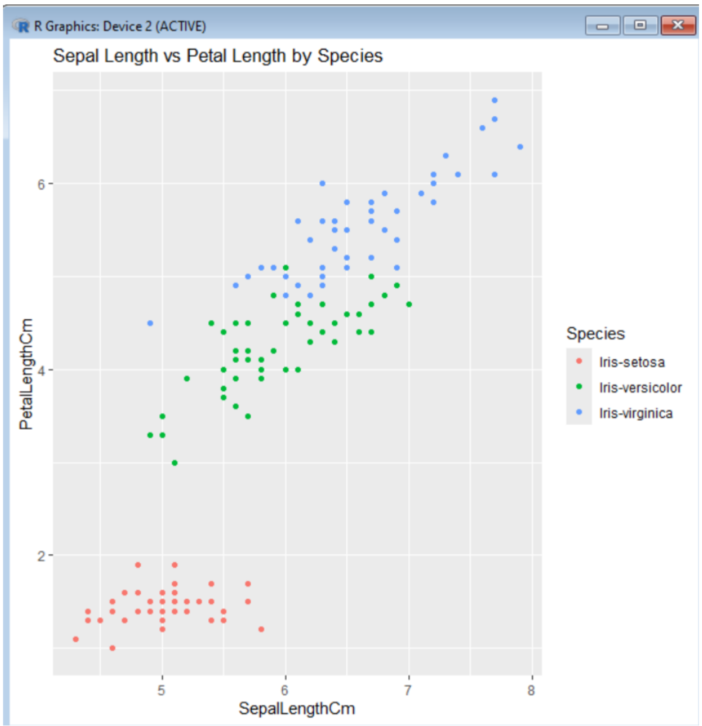
Scatter plots help visualize relationships between variables.

```
ggplot(iris_data, aes(x = SepalLengthCm, y = PetalLengthCm)) +
+   geom_point(color = "blue") +
+   labs(title = "Sepal Length vs Petal Length")
```



Color points by species to see patterns across different classes:

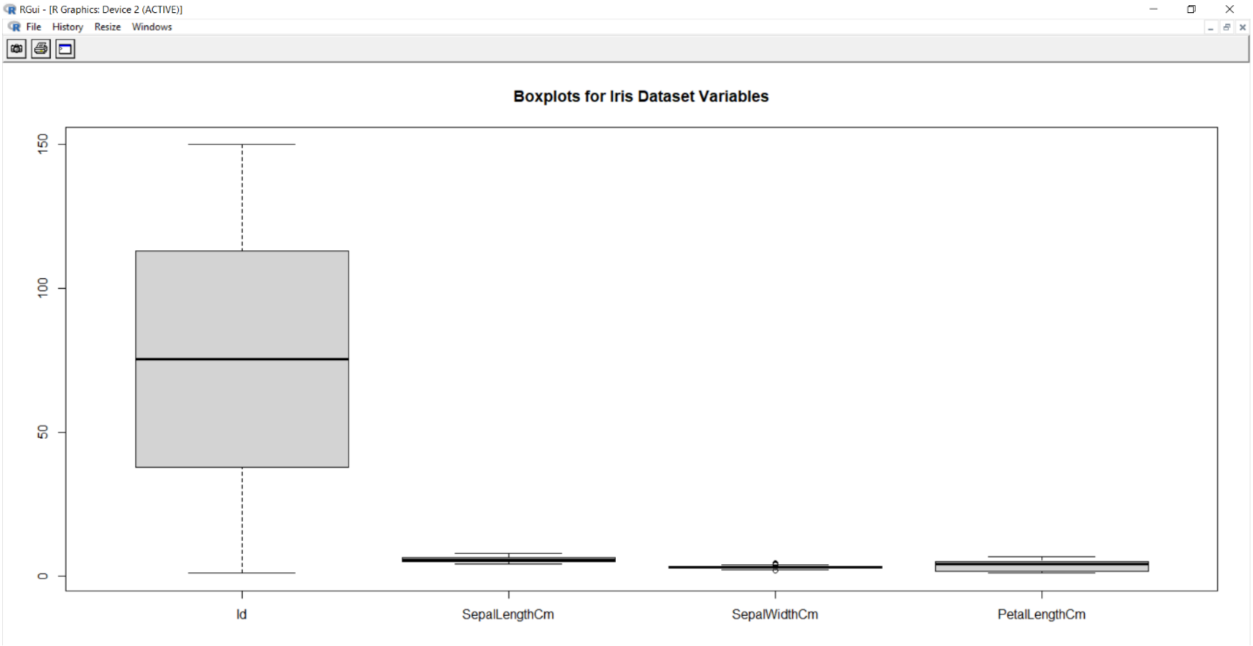
```
# Scatter plot colored by species
ggplot(iris_data, aes(x = Sepal.Length, y = Petal.Length, color = Species)) +
geom_point()
+ labs(title = "Sepal Length vs Petal Length by Species")
```



Step 8: Detecting Outliers:

Use boxplots for multiple variables to identify outliers. Look for points that are significantly above or below the whiskers:

boxplot(iris_data[, 1:4], main = "Boxplots for Iris Dataset Variables")


RGui - [R Graphics: Device 2 (ACTIVE)]
File
History
Resize
Windows
Boxplots for Iris Dataset Variables
0
50
100
150
Id
SepalLengthCm
SepalWidthCm
PetalLengthCm

Learning Outcomes :

1. Understand how to perform exploratory data analysis (EDA) on a dataset.
2. Learn how to visualize data distributions using histograms and boxplots.
3. Gain skills in detecting outliers and analyzing data spread.
4. Develop proficiency in using scatter plots to identify relationships between variables.
5. Gain experience in handling datasets and creating clear visual representations.
6. Improve the ability to draw insights from data visualizations and summary statistics.
7. Learn to use libraries such as **ggplot2** for advanced data visualization.