R Base Graphics

Prepared by Dr. Jose Isagani B. Janairo

2023

This document is for the exclusive use of students enrolled in the course Data Science for Life Scientists at De La Salle University.

Creating Plots in R

One of the advantages of using the R programming language is the ability to easily create and customize beautiful plots. There are two main methods of creating plots in R, the base graphics system, and the grammar of graphics (ggplot) system. In this exercise, we will focus on the basics of the R base graphics.

We will be using the iris dataset for this exercise. First, load the data.

```
data(iris)
```

Give a quick look at the contents of the dataset by using the head() and tail() command.

```
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
```

```
tail(iris)
```

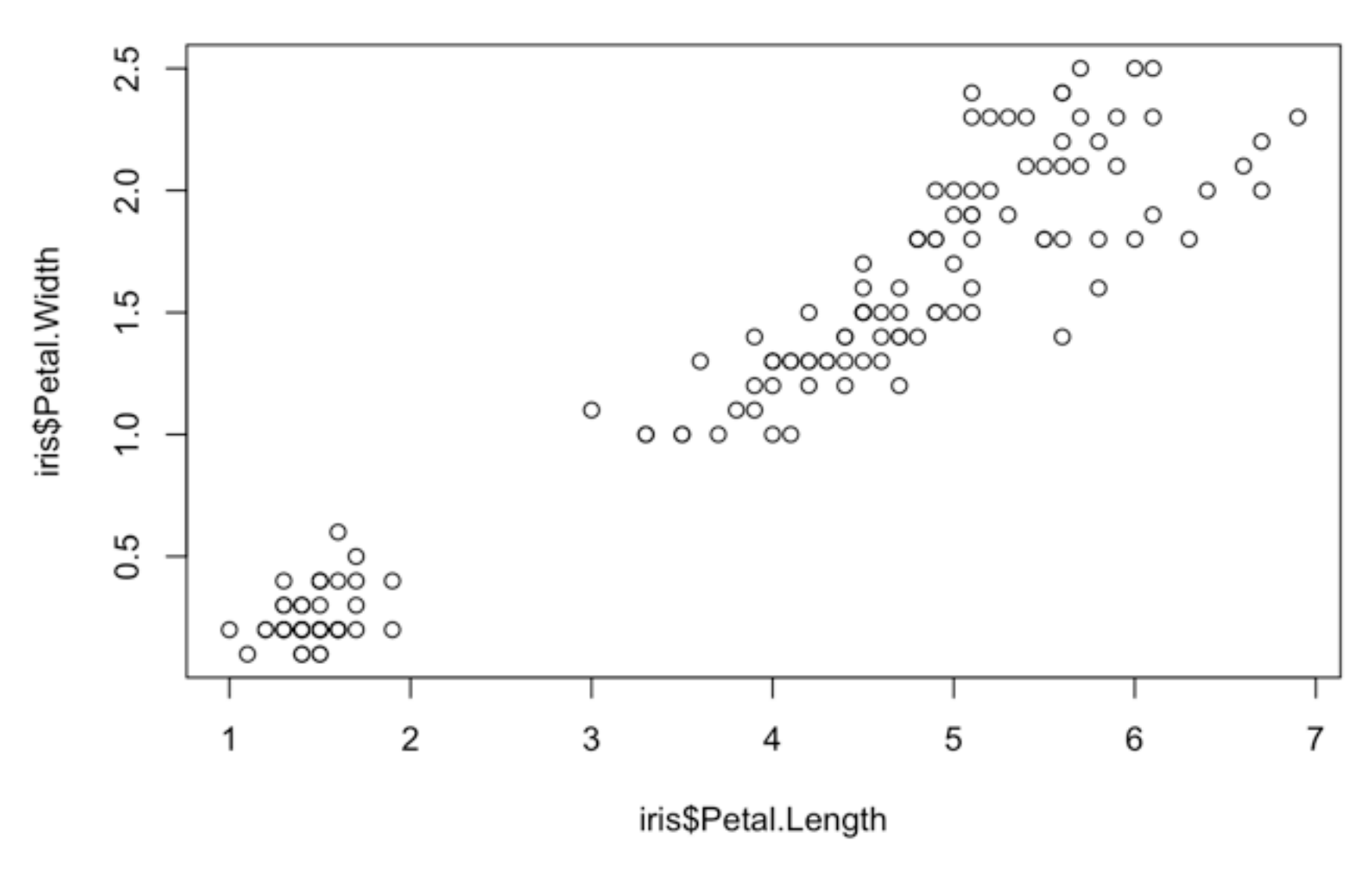
```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 145          6.7         3.3          5.7         2.5 virginica
## 146          6.7         3.0          5.2         2.3 virginica
## 147          6.3         2.5          5.0         1.9 virginica
## 148          6.5         3.0          5.2         2.0 virginica
## 149          6.2         3.4          5.4         2.3 virginica
## 150          5.9         3.0          5.1         1.8 virginica
```

As you can see, it has 4 continuous variables and a single categorical variable (species). Using the base R graphics, we will visualize these variables and explore possible relationships.

Scatter plot

The first type of plot that we will create is the scatter plot. The **plot()** function creates a scatter plot and uses the numeric variables as the arguments. As you can see from the plot that we just created, it is not visually appealing.

```
plot(iris$Petal.Length, iris$Petal.Width)
```

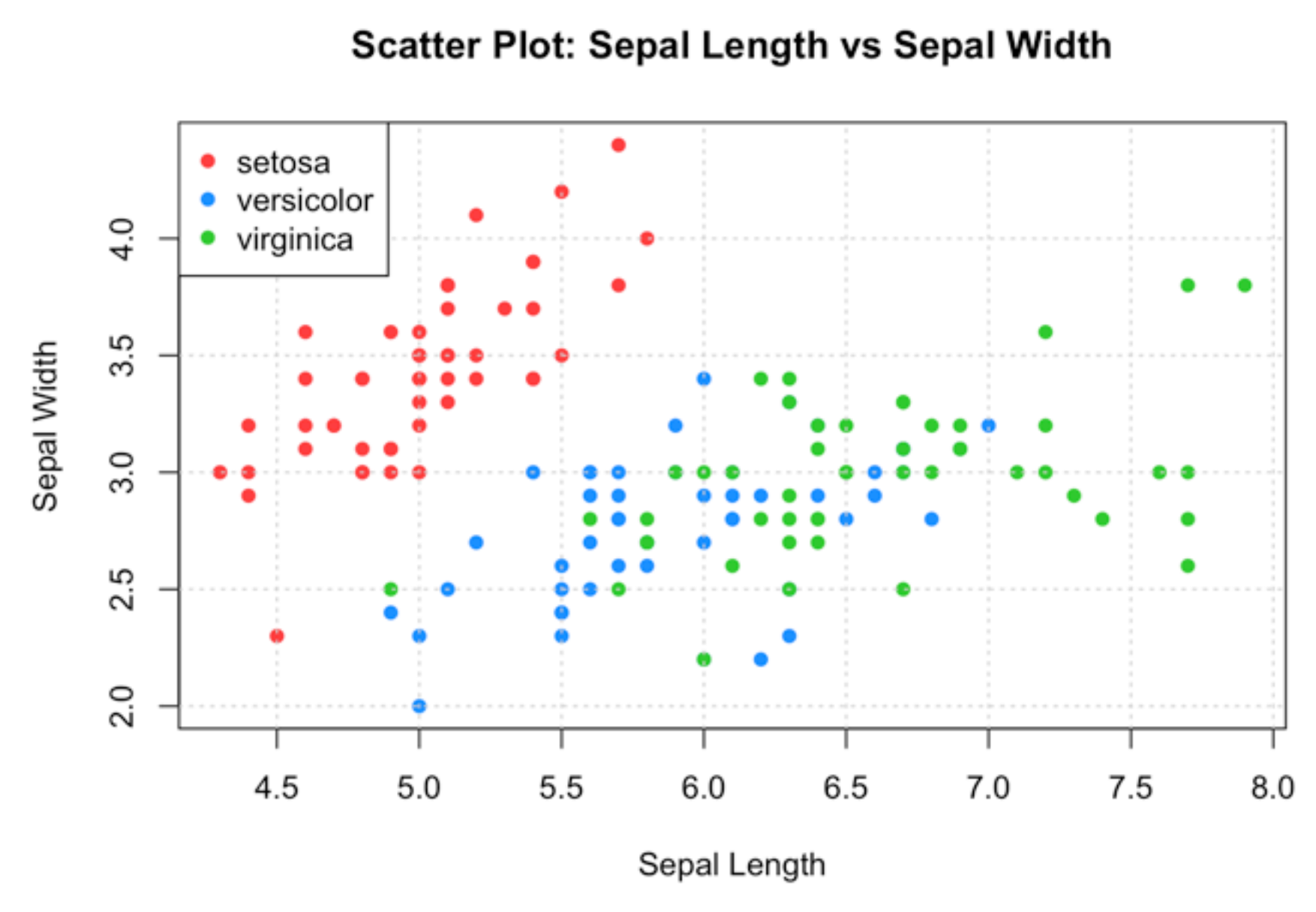


You can also create a scatter plot where the points are differentiated based on the categorical variable. By doing this, we are essentially creating a 3D scatter plot.

```
# Create a scatter plot
plot(iris$Sepal.Length, iris$Sepal.Width,
     xlab = "Sepal Length",
     ylab = "Sepal Width",
     main = "Scatter Plot: Sepal Length vs Sepal Width",
     col=c("brown1","dodgerblue1","limegreen")[as.integer(iris$Species)],
     pch = 16)

# Add grid lines
grid()

# Add a legend
legend(x="topleft",
       legend=c("setosa","versicolor","virginica"),
       col=c("brown1","dodgerblue1","limegreen"),
       pch=16)
```



You can use different markers by altering the "pch" argument, wherein the arguments takes a number from 0 - 25. Take a look at the various markers available. Run the following command.

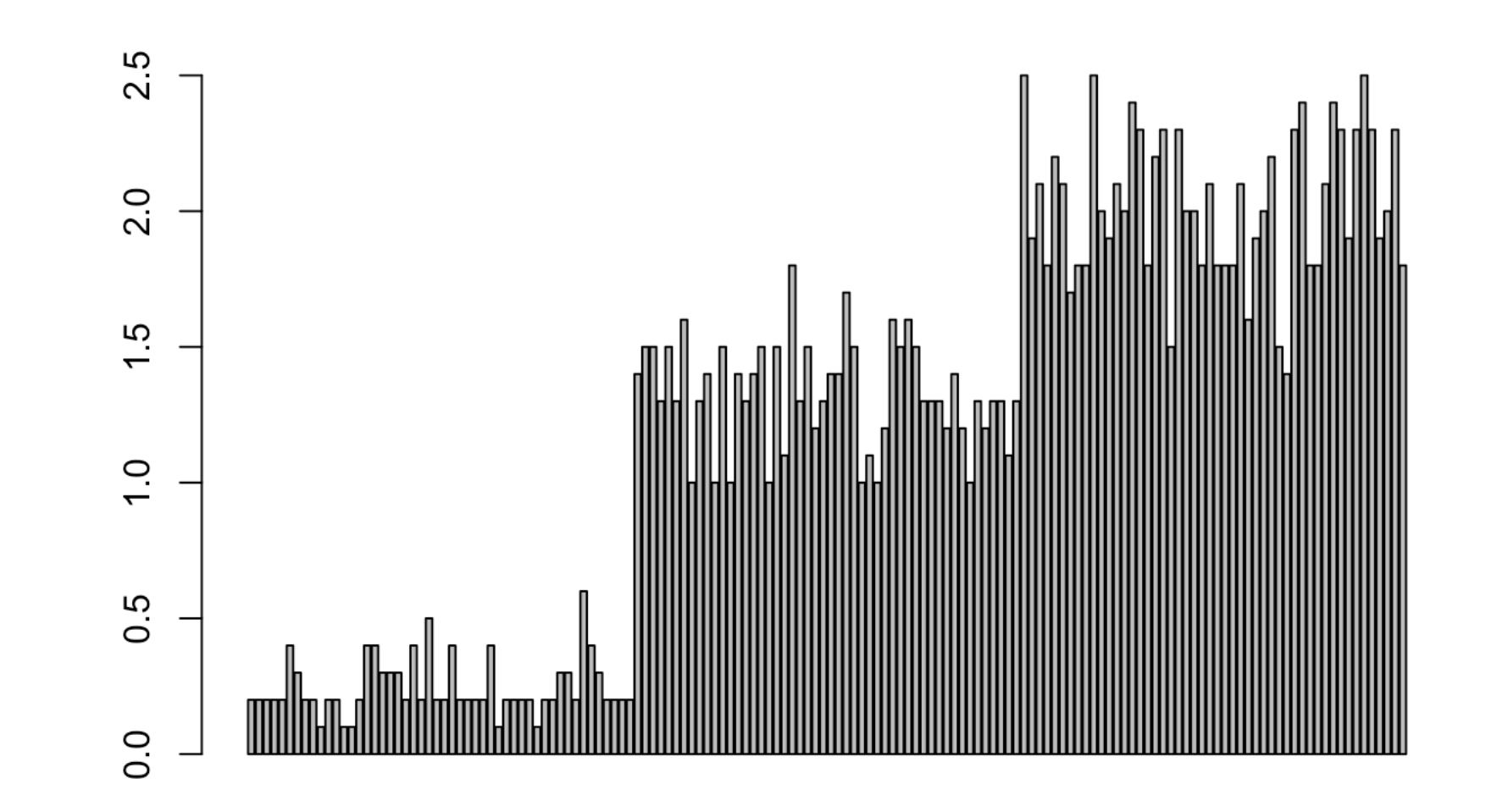
```
?pch
```

HANDS-ON OUTPUT 1 (HOO1): Create a single scatter plot using different markers and colors for each iris species. Save the image and upload it as an output for today's activity.

Bar Plots

Bar plots are commonly used in data visualization. The command **barplot()** is used to create bar plots in base R.

```
barplot(iris$Petal.Width)
```

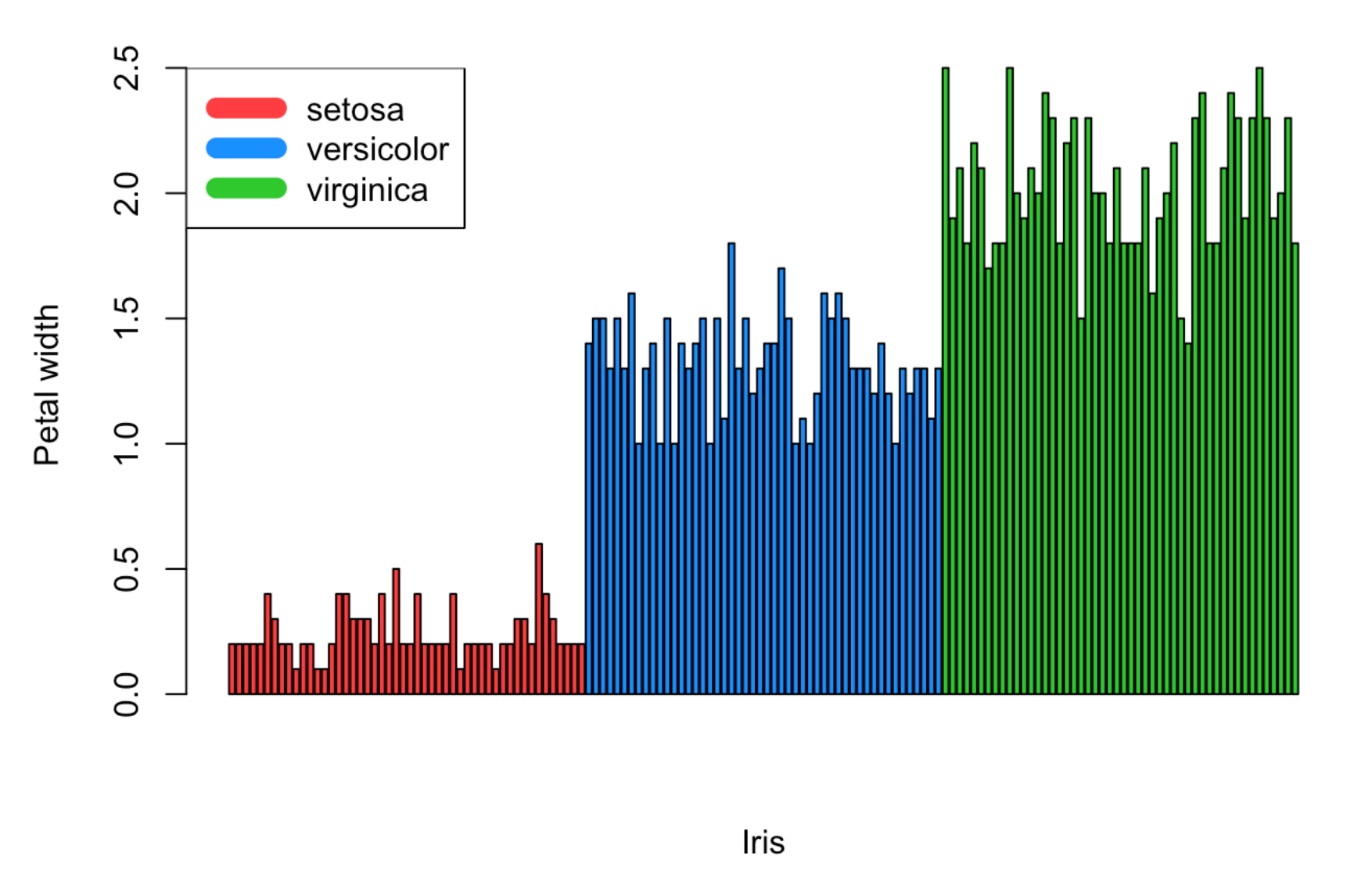


In the above chart that was created, each bar represents an iris flower and its corresponding petal width. As you can see, this is an incomplete graph since axes labels are not even supplied.

Just like in the scatter plot example, we can customize the bar plot to present more information.

```
#Create the bar plot
barplot(iris$Petal.Width, col = c("brown1","dodgerblue1","limegreen")[iris$Species], ylab = "Petal width", xlab = "Iris")

#Add the legend
legend("topleft", c("setosa", "versicolor", "virginica"), col=c("brown1","dodgerblue1","limegreen"), lwd=10)
```

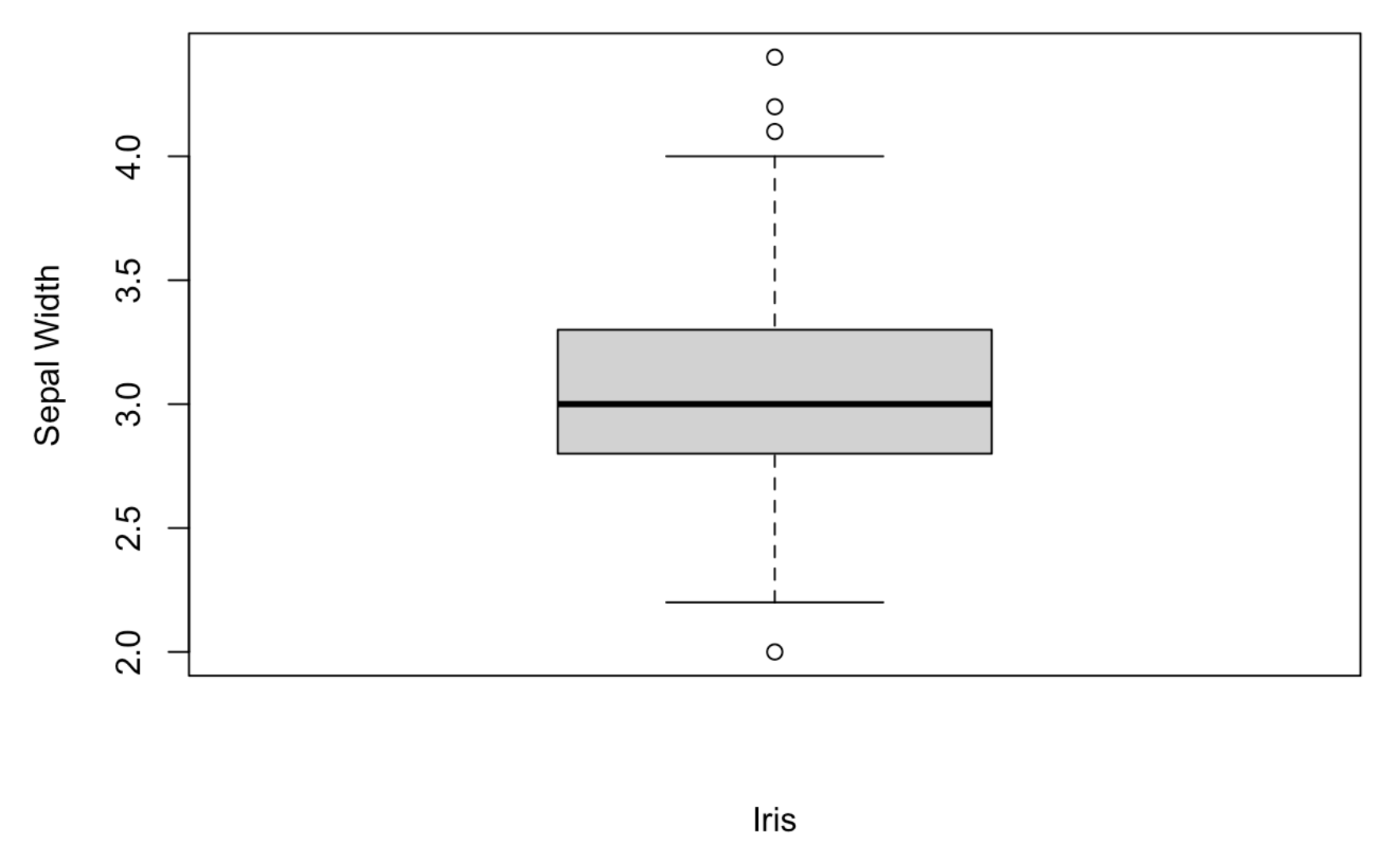


HOO2: Create horizontal bar plots for the remaining continuous variables in the iris dataset. Use the command ?barplot to see which argument to include in the command.

Box Plots

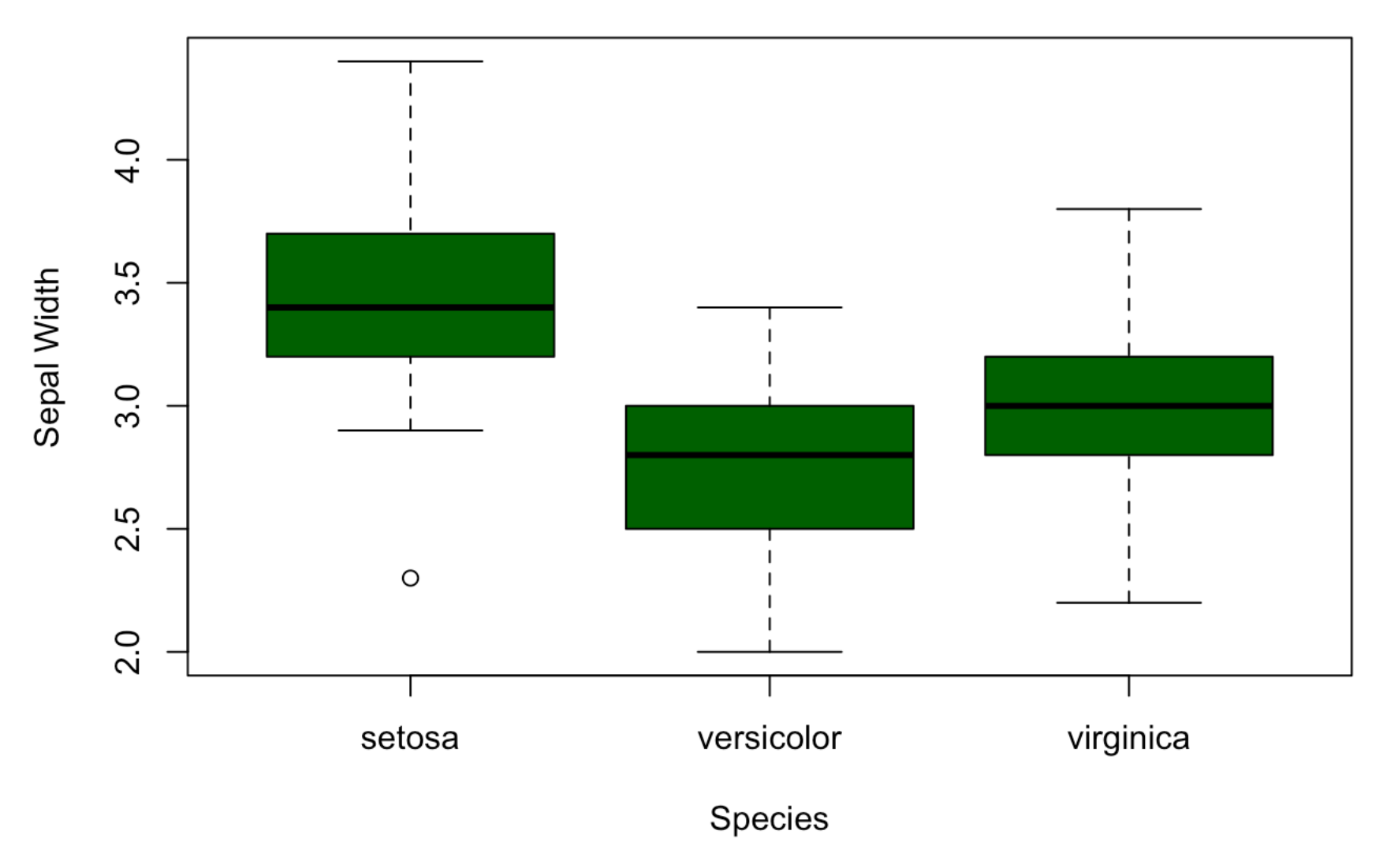
Box plots are useful graphs since they convey the distribution of the data into quartiles. The **boxplot()** command is called to create this type of graph.

```
boxplot(iris$Sepal.Width, xlab = "Iris", ylab = "Sepal Width")
```



You can create box plots based on the factor variable. You need to use the tilde *(~)*. You can also change the color of the plots.

```
boxplot(Sepal.Width~Species, data = iris, xlab = "Species", ylab = "Sepal Width", col = "darkgreen")
```



You can also perform subsetting on your dataset to focus on a particular species. This can also be performed with other types of graph.

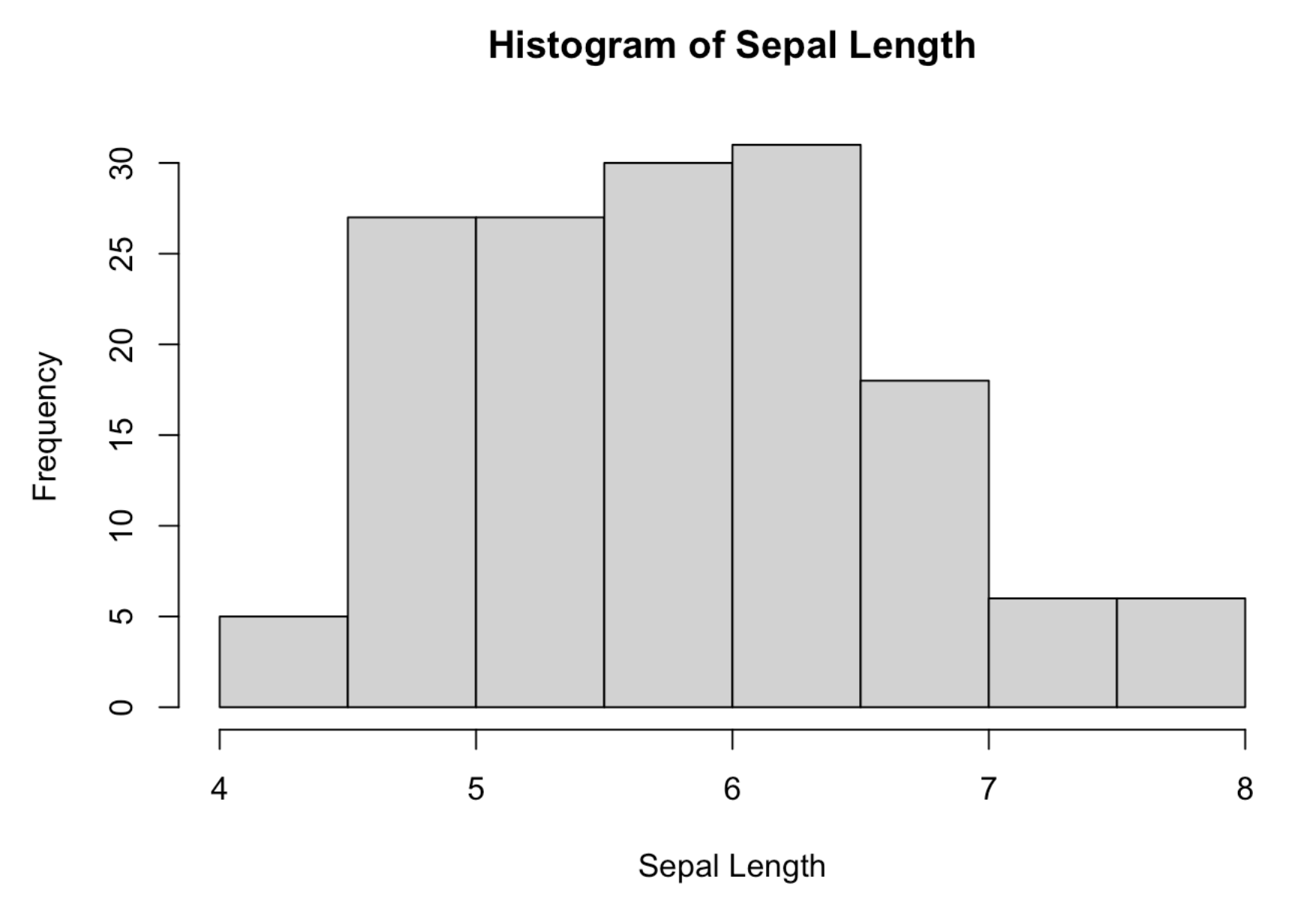
```
#Isolating the setosa species, and creating a new dataset
setosa_data <- subset(iris, Species == "setosa")

#Using the new dataset to create a box plot.
boxplot(setosa_data$Sepal.Width, xlab = "Setosa", ylab = "Sepal Width")
```

Histogram

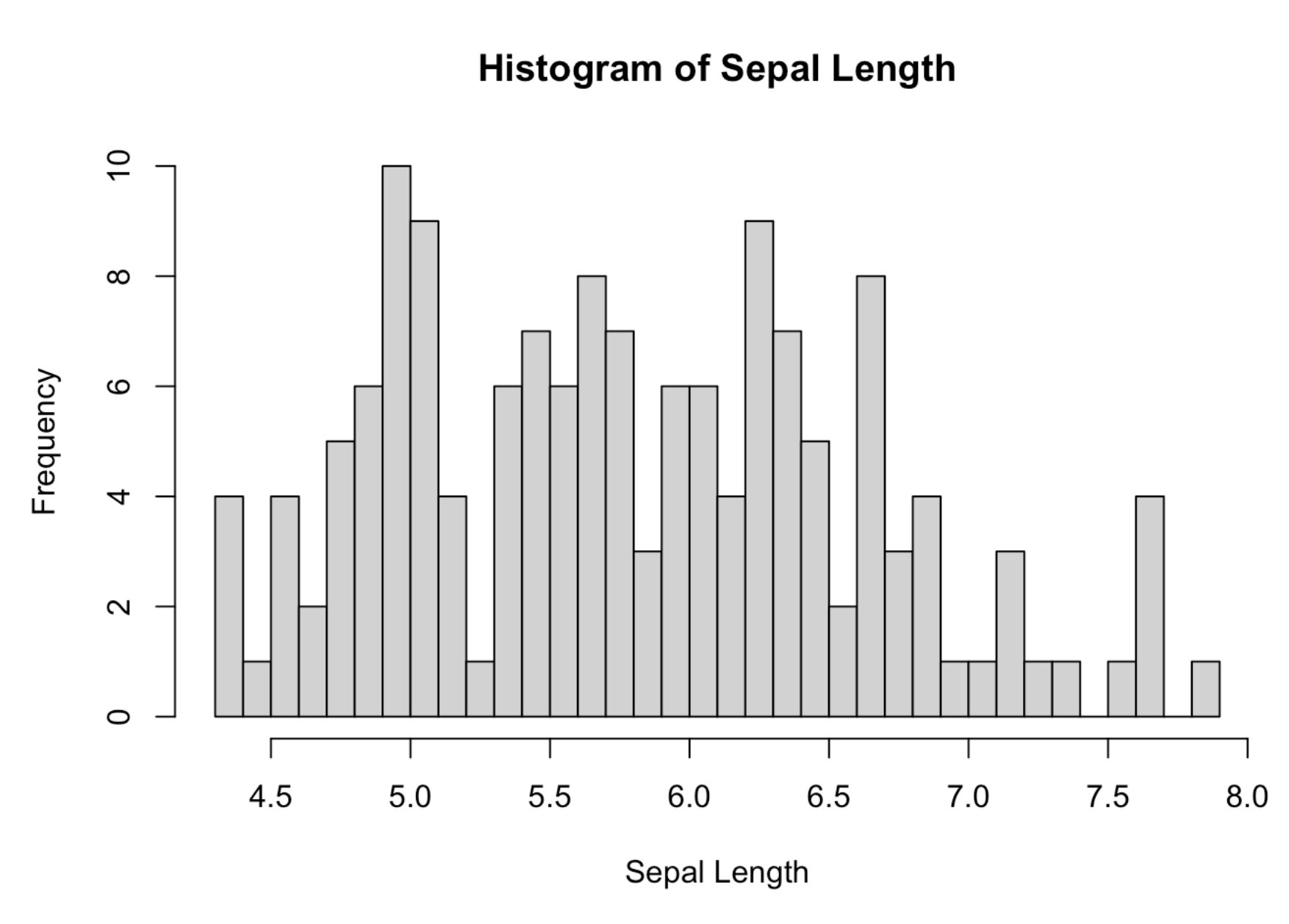
A histogram is composed of multiple bar plots that usually depict the frequency of occurrence of a particular variable. The command **hist()** is used to create histograms in the base R system.

```
hist(iris$Sepal.Length, xlab = "Sepal Length", ylab = "Frequency", main = "Histogram of Sepal Length")
```



The argument *breaks* in the hist() command is used to specify the number or locations of breaks (intervals) in the histogram. It determines the width of each histogram bin and affects the granularity of the histogram.

```
hist(iris$Sepal.Length, xlab = "Sepal Length", ylab = "Frequency", main = "Histogram of Sepal Length", breaks = 30)
```



HOO3: Create histograms using the sepal length for the versicolor species using the following breaks: 5, 10, 20, 30.

Exercises

Use the *mpg* dataset for the following exercises. You need to load the **ggplot2** package first. Exert effort to customize and improve the aesthetic quality of your plots. **Take a screenshot of your console showing the code, and the chart output.**

1. Create a scatterplot where the x axis is "cty" and the y axis is "hwy". Color the data points by class category.
2. Create the appropriate plot to determine which fuel type yields the best city driving mileage. Make sure that all axes are properly labeled. Why did you select your chosen plot?
3. Create a vertical barplot that shows the vehicle displacement (displ) and is colored based on the number of gears.

END