Unit 10, Module 4: Stratified Graphs

CSCI 5a: Programming in R

Let's clear the global computing environment:

```
rm( list = ls() )
```

Module Overview and Learning Outcomes

Hello! And welcome to Module 4: Stratified Graphs.

In this module, we will explore two fascinating graphing techniques.

- In Section 1, we'll learn how to construct boxplots across each level of a factor.
- In Section 2, we'll see how to construct stripcharts across each level of a factor.

When you've completed this module, you'll be able to:

- Construct a stratified boxplot.
- Construct a stratified stripchart.

There are no new built-in R functions in this module.

All right! Let's get started by learning how to construct a stratified boxplot.

Section 1: Stratified boxplots

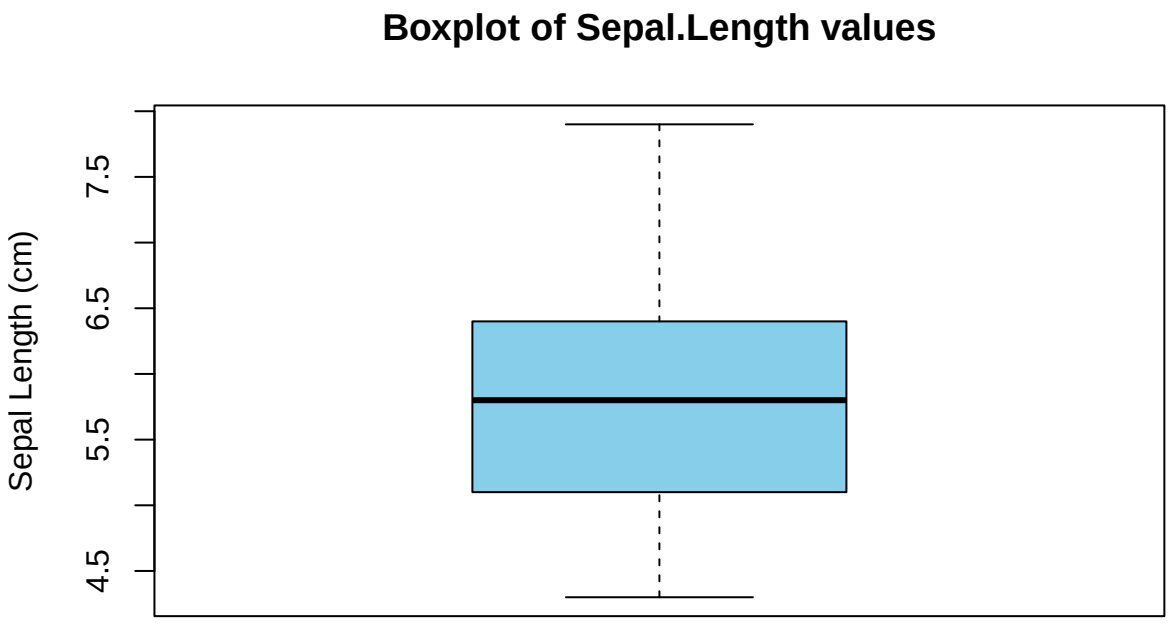
> **Main Idea:** *We can display a boxplot for a numeric vector across each level of a factor*

In this section, we'll learn how to construct boxplots across each level of a factor.

Let's consider the `iris` data frame, and in particular the `Sepal.Length` variable.

We can make a simple boxplot of the values in this column vector:

```
boxplot(
    iris$Sepal.Length,
    main = "Boxplot of Sepal.Length values",
    ylab = "Sepal Length (cm)",
    col = "skyblue"
)
```



That's nice, but remember that the `iris` dataset consists of measurements on three different species of iris.

It would be interesting to look at a boxplot for each of the species; perhaps there's some difference between the different species.

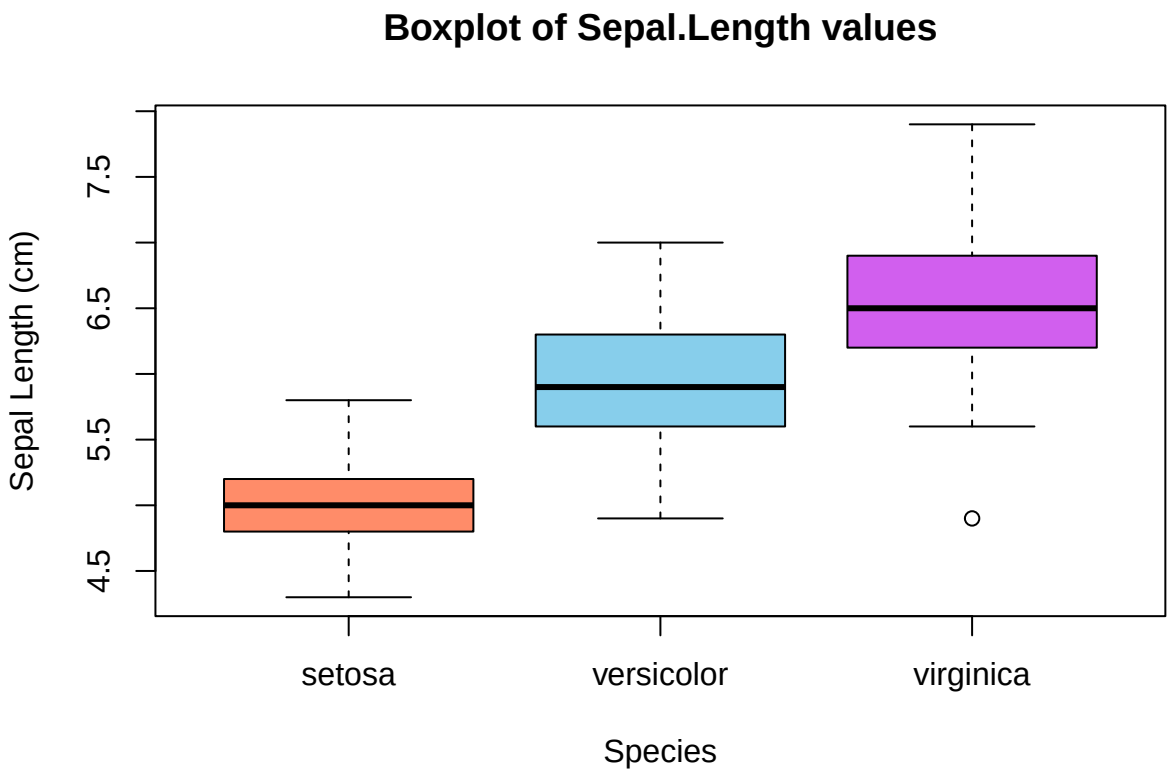
A *stratified* boxplot is a display which consists of a separate boxplot for each level of a factor.

In this case, the `Species` variable is the factor, with three levels.

Be aware that this term "stratified boxplot" is my own and if you use with other people they won't know what you're talking about.

We make a stratified boxplot in R by using a formula, very similar to how we specified a linear model in the previous lecture:

```
boxplot(
    iris$Sepal.Length ~ iris$Species,
    main = "Boxplot of Sepal.Length values",
    xlab = "Species",
    ylab = "Sepal Length (cm)",
    col = c("salmon1", "skyblue", "mediumorchid2")
)
```



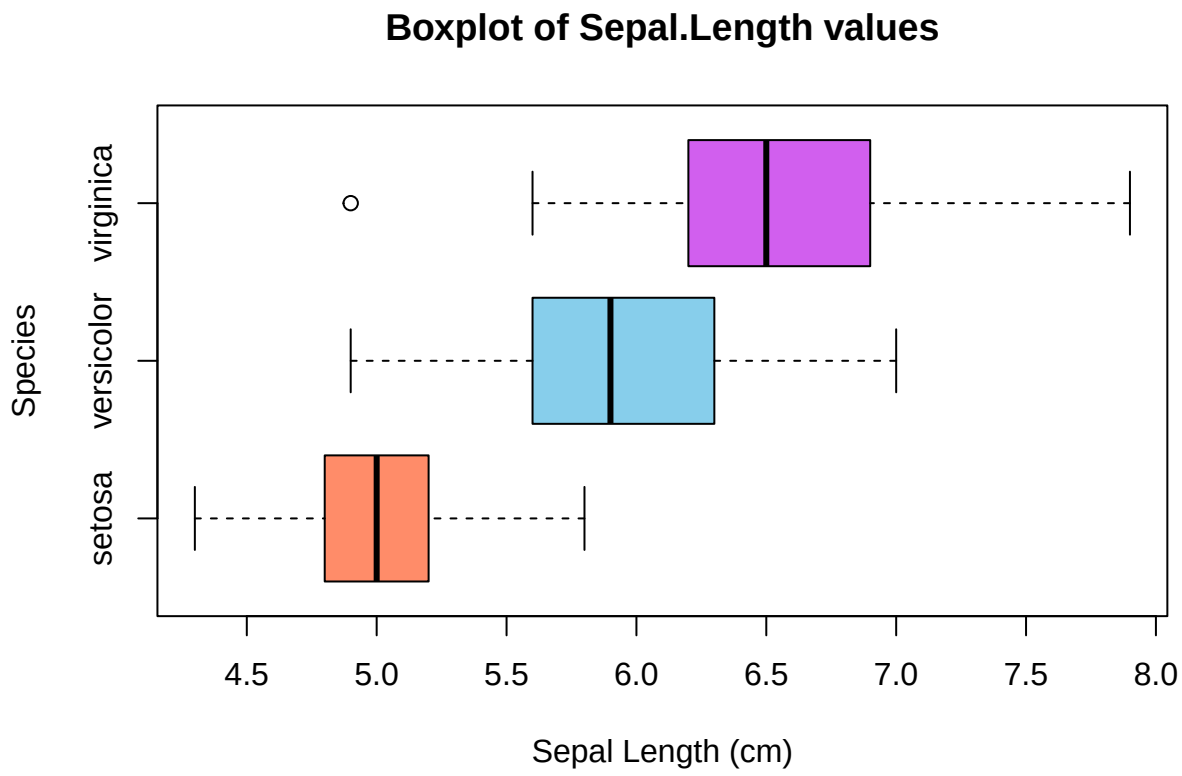
Wow!

There really is a big difference in the `Sepal.Length` values across the three species.

We couldn't see that with the basic boxplot, but when we stratified it became immediately apparent.

Incidentally, we can also display these boxplots horizontally:

```
boxplot(
    iris$Sepal.Length ~ iris$Species,
    main = "Boxplot of Sepal.Length values",
    xlab = "Sepal Length (cm)",
    ylab = "Species",
    col = c("salmon1", "skyblue", "mediumorchid2"),
    horizontal = TRUE
)
```



Personally, I think the more levels there are for the factor, the more I prefer the vertical display.

So that's how to construct a stratified boxplot display.

Now let's see how to construct a stratified stripchart display.

Exercise 4.1

Create a simple boxplot for the `Sepal.Width` variable. Then construct a stratified boxplot for the `Sepal.Width` variable across the three species. What do you conclude from your stratified display?

Solution

Section 2: Stratified Stripcharts

> **Main Idea:** *We can display a stripchart for a numeric vector across each level of a factor*

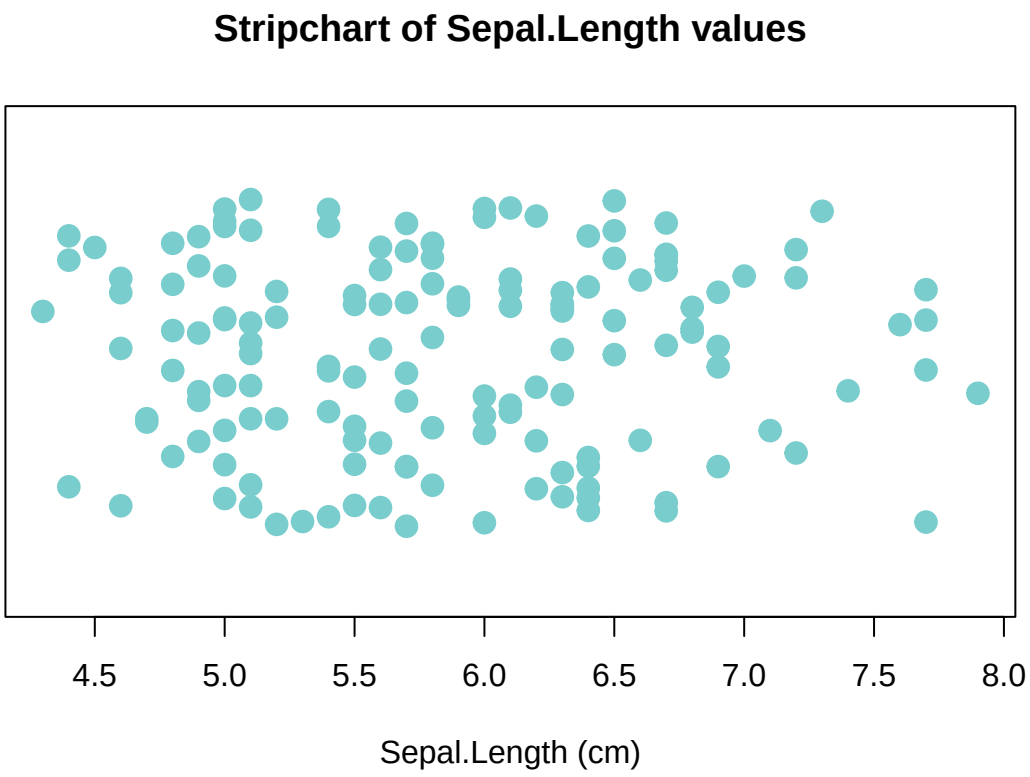
In this section, we'll learn how to construct stripcharts across each level of a factor.

Now we're going to do the same analysis, but with stripcharts.

First, let's make a simple stripchart of the `Sepal.Length` values:

```
stripchart(
    iris$Sepal.Length,
    ylim = c(0, 2),
```

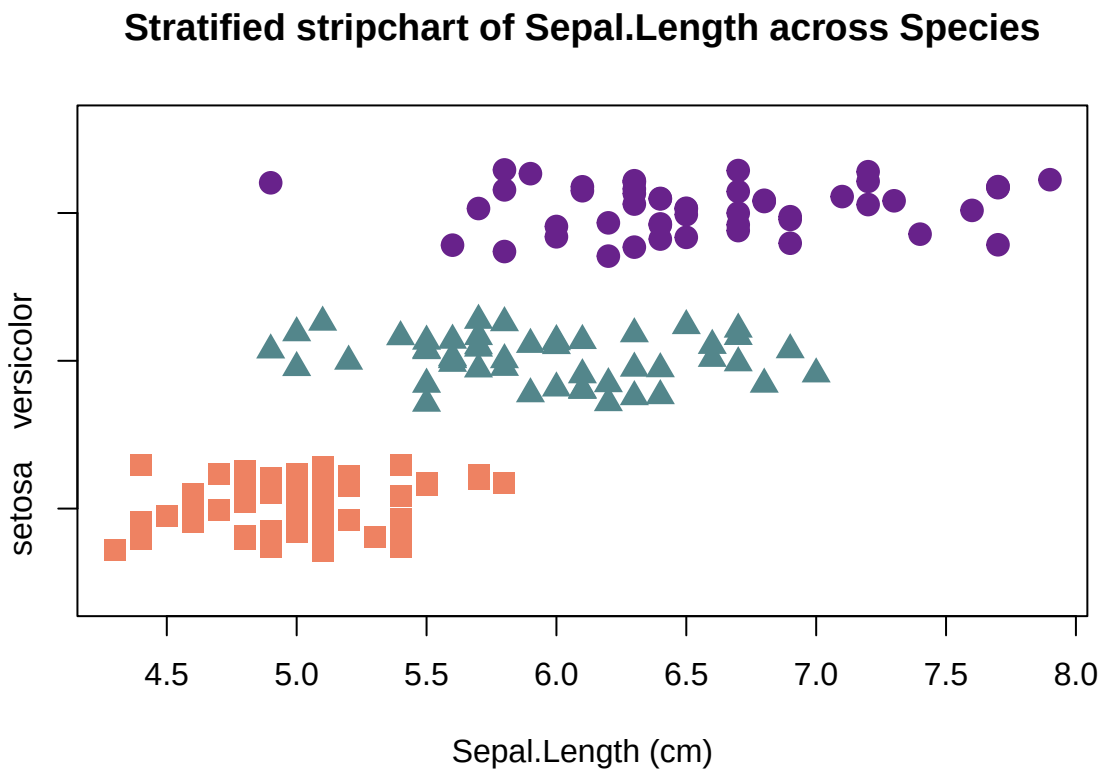
```
    main = "Stripchart of Sepal.Length values",
    xlab = "Sepal.Length (cm)",
    method = "jitter",
    jitter = 0.7,
    pch = 19,
    cex = 1.5,
    col = "darkslategray3"
)
```



Again, as with the simple boxplot, it's hard to see any sort of pattern here.

Now let's make a stratified stripchart, using the same method of specifying a formula:

```
stripchart(
    iris$Sepal.Length ~ iris$Species,
    main = "Stratified stripchart of Sepal.Length across Species",
    xlab = "Sepal.Length (cm)",
    method = "jitter",
    jitter = 0.3,
    pch = c(15, 17, 19),
    cex = 1.5,
    col = c( "salmon2", "cadetblue4", "darkorchid4" )
)
```



Once again, now we can really see the differences between the species.

Two important points:

- Do NOT specify the `ylim` values when you have stratified stripchart, because R will take care of that.
- Jitter should be no more than 0.5.

So that's how to construct a stratified stripchart.

Now let's review what we've learned in this module.

Exercise 4.2

Create a simple stripchart for the `Sepal.Width` variable. Then construct a stratified stripchart for the `Sepal.Width` variable across the three species. What do you conclude from your stratified display?

Solution

Module Review

In this module, we explored two fascinating graphing techniques.

- In Section 1, we learned how to construct boxplots across each level of a factor.
- In Section 2, we saw how to construct stripcharts across each level of a factor.

Now that you've completed this module, you should be able to:

- Construct a stratified boxplot.
- Construct a stratified stripchart.

There were no new built-in R functions in this this module.

Solutions to the Exercises

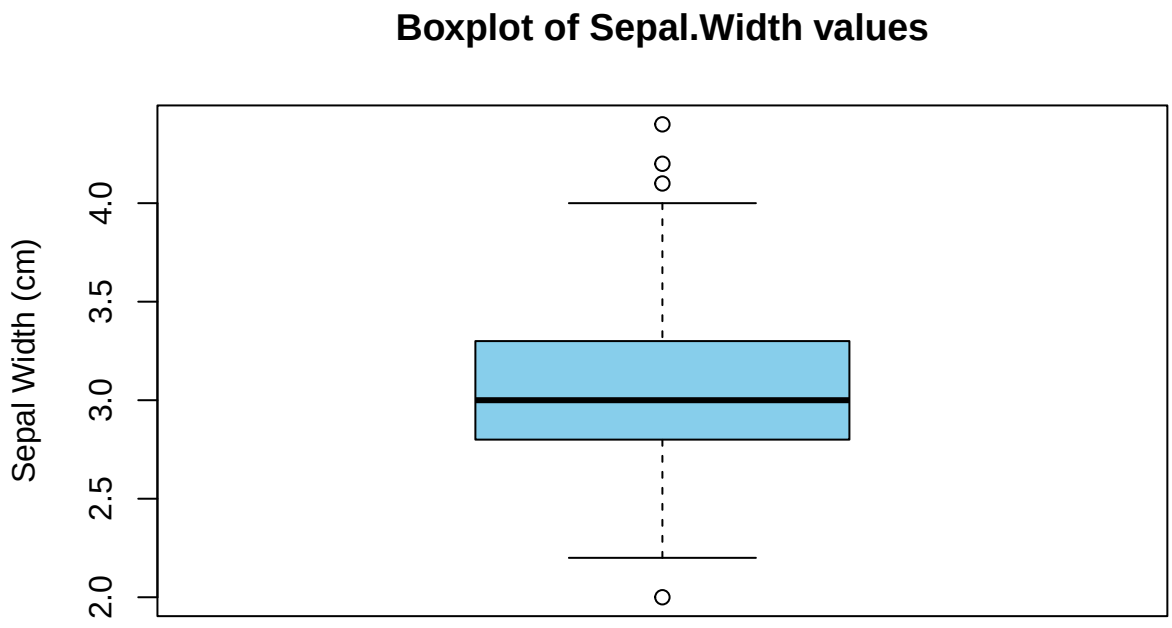
Exercise 4.1

Create a simple boxplot for the `Sepal.Width` variable. Then construct a stratified boxplot for the `Sepal.Width` variable across the three species. What do you conclude from your stratified display?

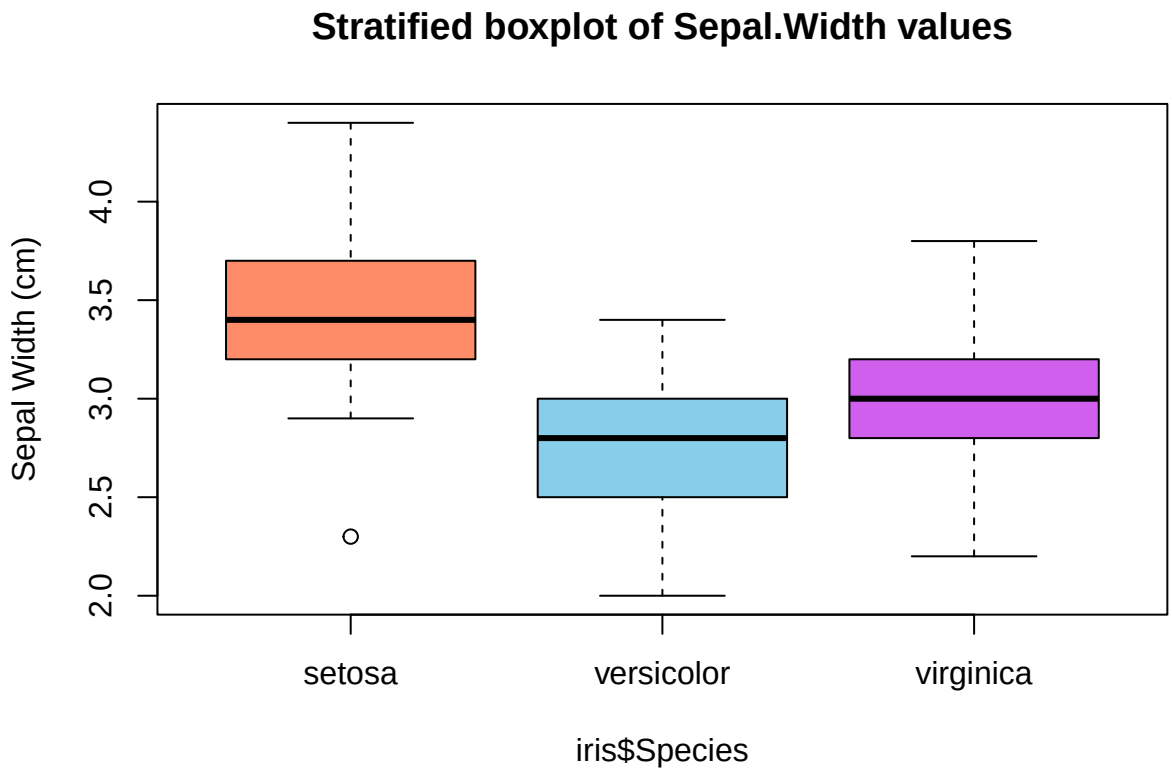
Solution

Here's the simple boxplot:

```
boxplot(
    iris$Sepal.Width,
    main = "Boxplot of Sepal.Width values",
    ylab = "Sepal Width (cm)",
    col = "skyblue"
)
```



```
boxplot(
    iris$Sepal.Width ~ iris$Species,
    main = "Stratified boxplot of Sepal.Width values",
    ylab = "Sepal Width (cm)",
    col = c("salmon1", "skyblue", "mediumorchid2")
)
```



Although the `setosa` species had the shortest sepal lengths, it seems from this graph that the species has the largest sepal widths.

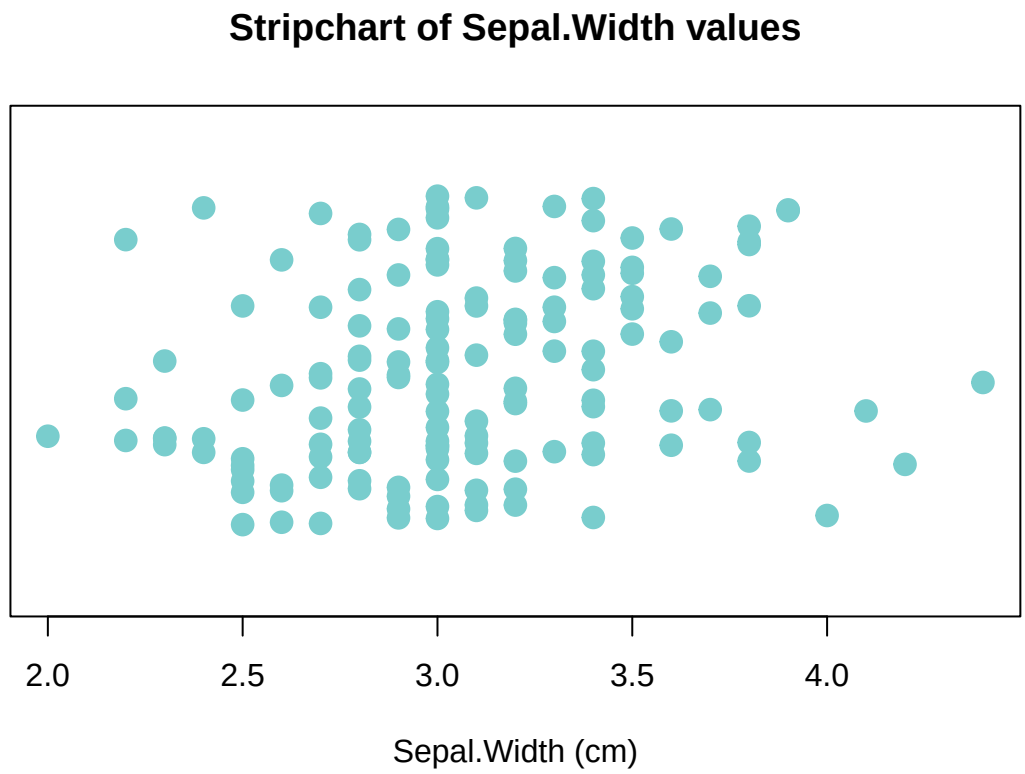
Exercise 4.2

Create a simple stripchart for the `Sepal.Width` variable. Then construct a stratified stripchart for the `Sepal.Width` variable across the three species. What do you conclude from your stratified display?

Solution

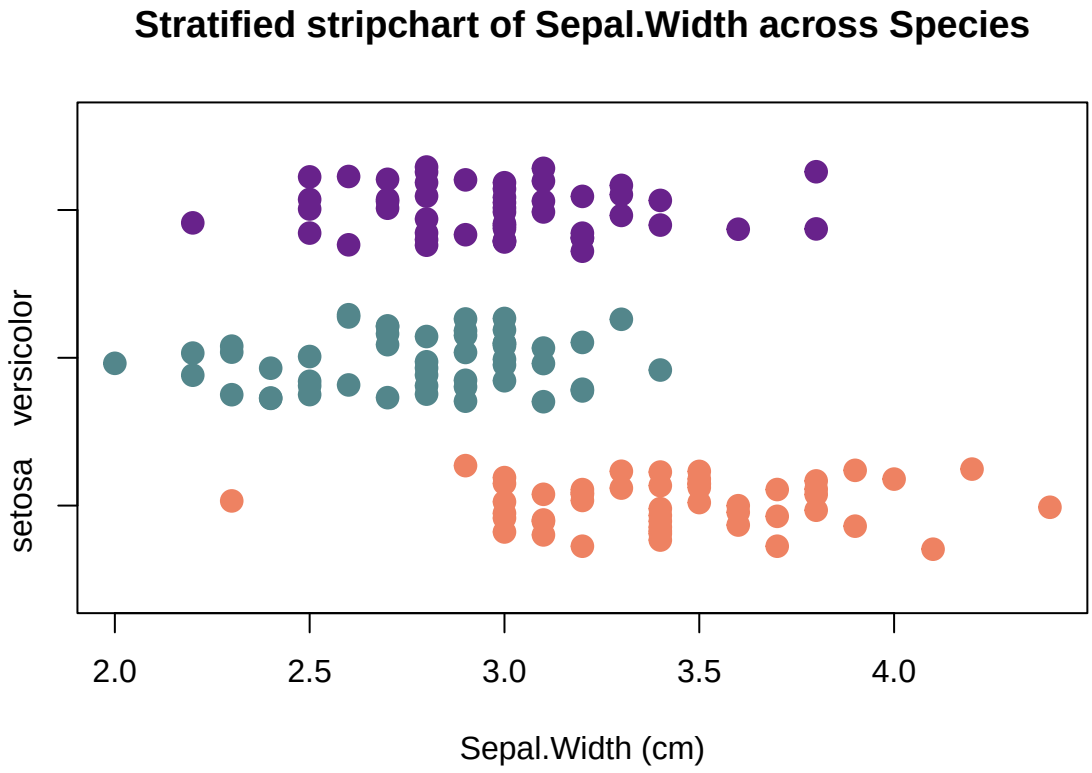
```
stripchart(
    iris$Sepal.Width,
    ylim = c(0, 2),
    main = "Stripchart of Sepal.Width values",
    xlab = "Sepal.Width (cm)",
    method = "jitter",
    jitter = 0.7,
    pch = 19,
```

```
    cex = 1.5,
    col = "darkslategray3"
)
```



Now here's the stratified display:

```
stripchart(
    iris$Sepal.Width ~ iris$Species,
    main = "Stratified stripchart of Sepal.Width across Species",
    xlab = "Sepal.Width (cm)",
    method = "jitter",
    jitter = 0.3,
    pch = 19,
    cex = 1.5,
    col = c( "salmon2", "cadetblue4", "darkorchid4" )
)
```


Stratified stripchart of Sepal.Width across Species
setosa
versicolor
2.0
2.5
3.0
3.5
4.0
Sepal.Width (cm)