Iris Dataset

David C. King

2/17/2021

Builtin datasets

R has a package called “datasets” that provides many old but readily available datasets to test things out on.

In the console, do `?iris`

The help window will gives a description of the data, and the format of the `iris` and `iris3` data frames.

To see the full list of datasets in the package, click on ‘Index’ at the bottom of the help page.

`iris` dataset

Description

This famous (Fisher’s (FISHER 1936) or Anderson’s (Anderson 1936)) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris *setosa*, *versicolor*, and *virginica.*

```
data(iris) # copies the data frame into your workspace
head(iris) # see the first 10 rows. A data frame is basically a spreadsheet
```

```
##   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
```

```
iris$Petal.Length # access a whole column by name using the '$'
```

```
##   [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3 1.4
##  [19] 1.7 1.5 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2
##  [37] 1.3 1.4 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4 4.7 4.5 4.9 4.0
##  [55] 4.6 4.5 4.7 3.3 4.6 3.9 3.5 4.2 4.0 4.7 3.6 4.4 4.5 4.1 4.5 3.9 4.8 4.0
##  [73] 4.9 4.7 4.3 4.4 4.8 5.0 4.5 3.5 3.8 3.7 3.9 5.1 4.5 4.5 4.7 4.4 4.1 4.0
##  [91] 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1 6.0 5.1 5.9 5.6 5.8 6.6 4.5 6.3
## [109] 5.8 6.1 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9 5.0 5.7 4.9 6.7 4.9 5.7 6.0
## [127] 4.8 4.9 5.6 5.8 6.1 6.4 5.6 5.1 5.6 6.1 5.6 5.5 4.8 5.4 5.6 5.1 5.1 5.9
## [145] 5.7 5.2 5.0 5.2 5.4 5.1
```

```
iris[[3]] # or by its column number in [[ ]]
```

```
##   [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3 1.4
##  [19] 1.7 1.5 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2
##  [37] 1.3 1.4 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4 4.7 4.5 4.9 4.0
##  [55] 4.6 4.5 4.7 3.3 4.6 3.9 3.5 4.2 4.0 4.7 3.6 4.4 4.5 4.1 4.5 3.9 4.8 4.0
##  [73] 4.9 4.7 4.3 4.4 4.8 5.0 4.5 3.5 3.8 3.7 3.9 5.1 4.5 4.5 4.7 4.4 4.1 4.0
##  [91] 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1 6.0 5.1 5.9 5.6 5.8 6.6 4.5 6.3
## [109] 5.8 6.1 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9 5.0 5.7 4.9 6.7 4.9 5.7 6.0
## [127] 4.8 4.9 5.6 5.8 6.1 6.4 5.6 5.1 5.6 6.1 5.6 5.5 4.8 5.4 5.6 5.1 5.1 5.9
## [145] 5.7 5.2 5.0 5.2 5.4 5.1
```

Visualizing the iris data

How do we get a sense of Sepal and Petal dimensions among iris species? Given these measurements, let's look at some of their properties.

Boxplots

The distributions of each measurement reveal a greater variability in certain species.

To make a paneled version of the four boxplots, we make a new data structure. I.e., stick all the value columns together vertically, then add a descriptive column as to which measurement they are.

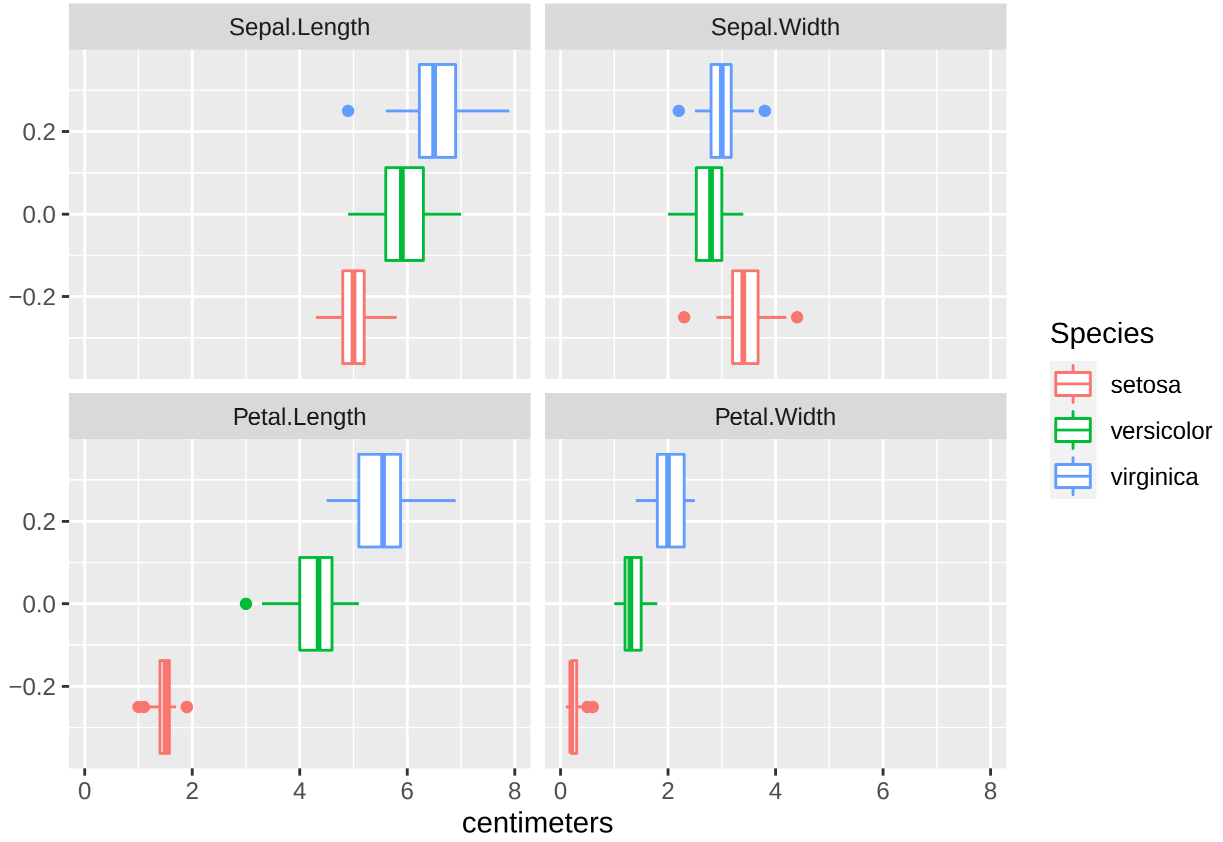
```r
library(reshape2) # for melt
iris_long = melt(iris,
               id = "Species",
               value.name = "centimeters",
               variable.name = "flower_part")

head(iris_long)
```

```
##   Species  flower_part centimeters
## 1  setosa Sepal.Length         5.1
## 2  setosa Sepal.Length         4.9
## 3  setosa Sepal.Length         4.7
## 4  setosa Sepal.Length         4.6
## 5  setosa Sepal.Length         5.0
## 6  setosa Sepal.Length         5.4
```

Split into different boxplots based on the flower part.

```r
ggplot(iris_long, aes(x = centimeters, col = Species)) +
  geom_boxplot() +
  facet_wrap( ~ flower_part)
```

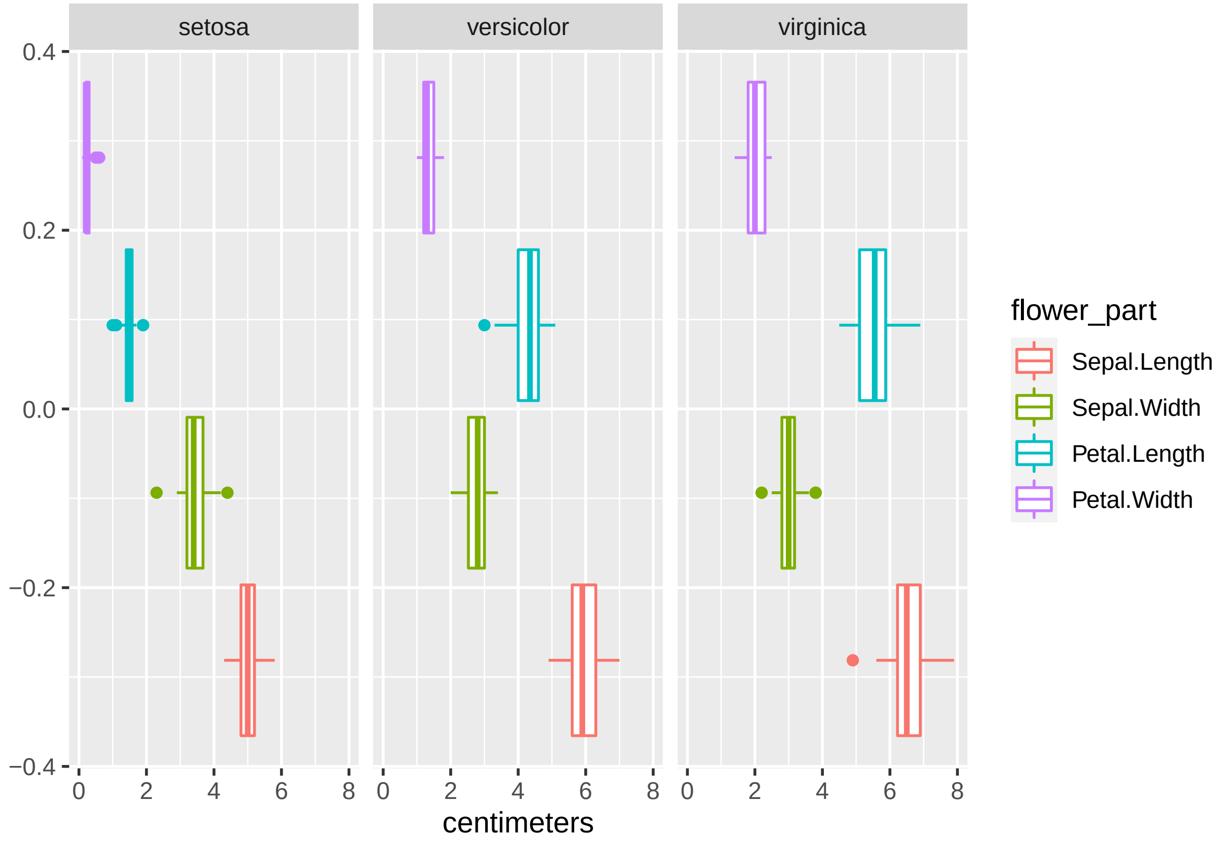


The most variable measurements appear to be the sepal and petal lengths of *versicolor* and *virginica*. The least amount of variation appears to be in the petal width and lengths of *setosa*.

setosa is easiest to distinguish from the other species.

Split the boxplots by species instead of flower part.

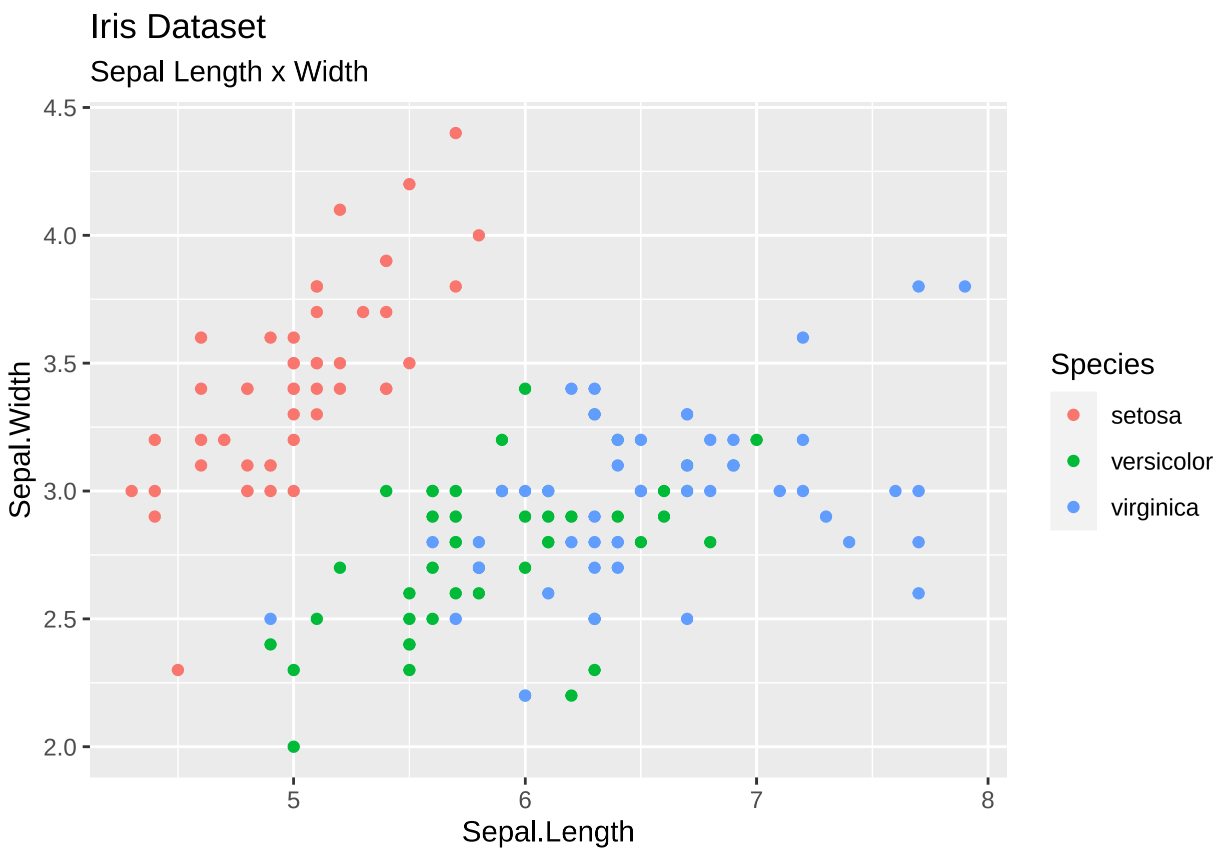
```
ggplot(iris_long, aes(x=centimeters, col=flower_part)) +
  geom_boxplot() +
  facet_wrap(~Species)
```

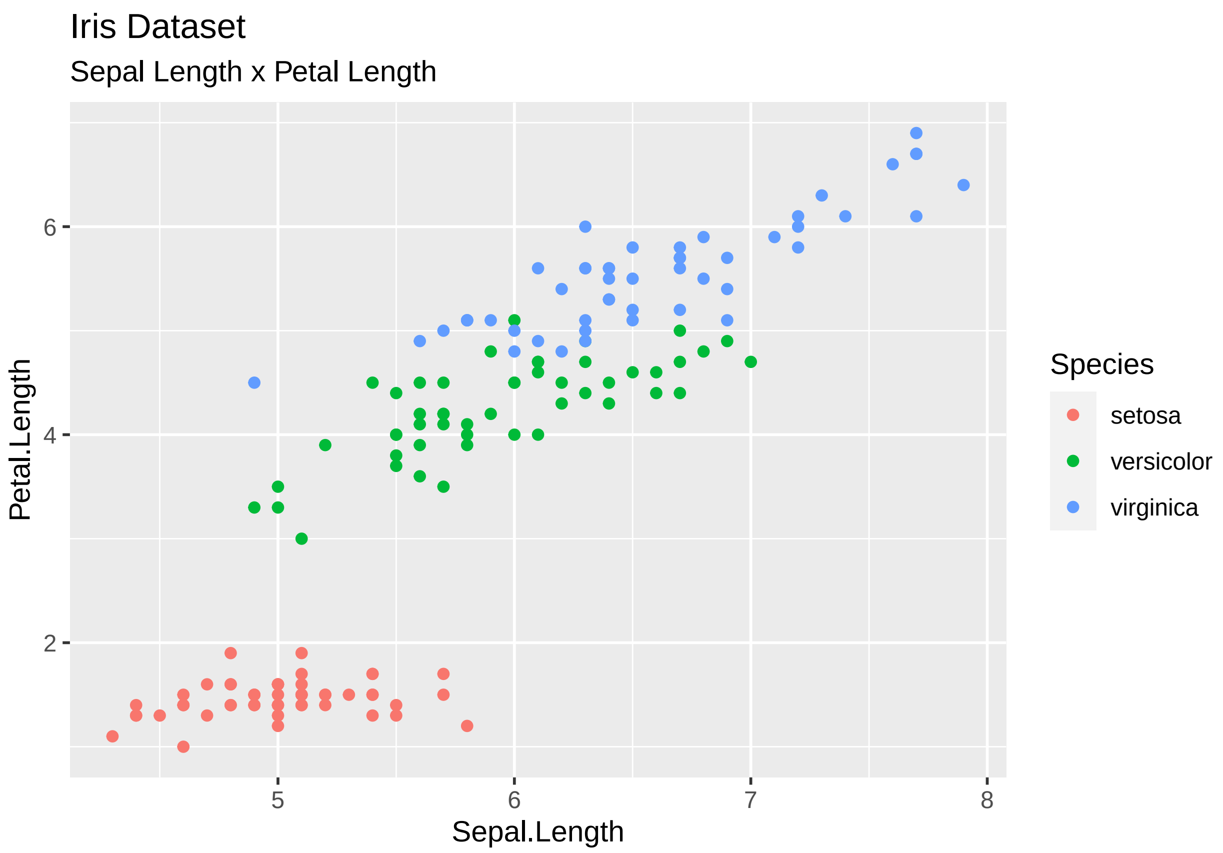


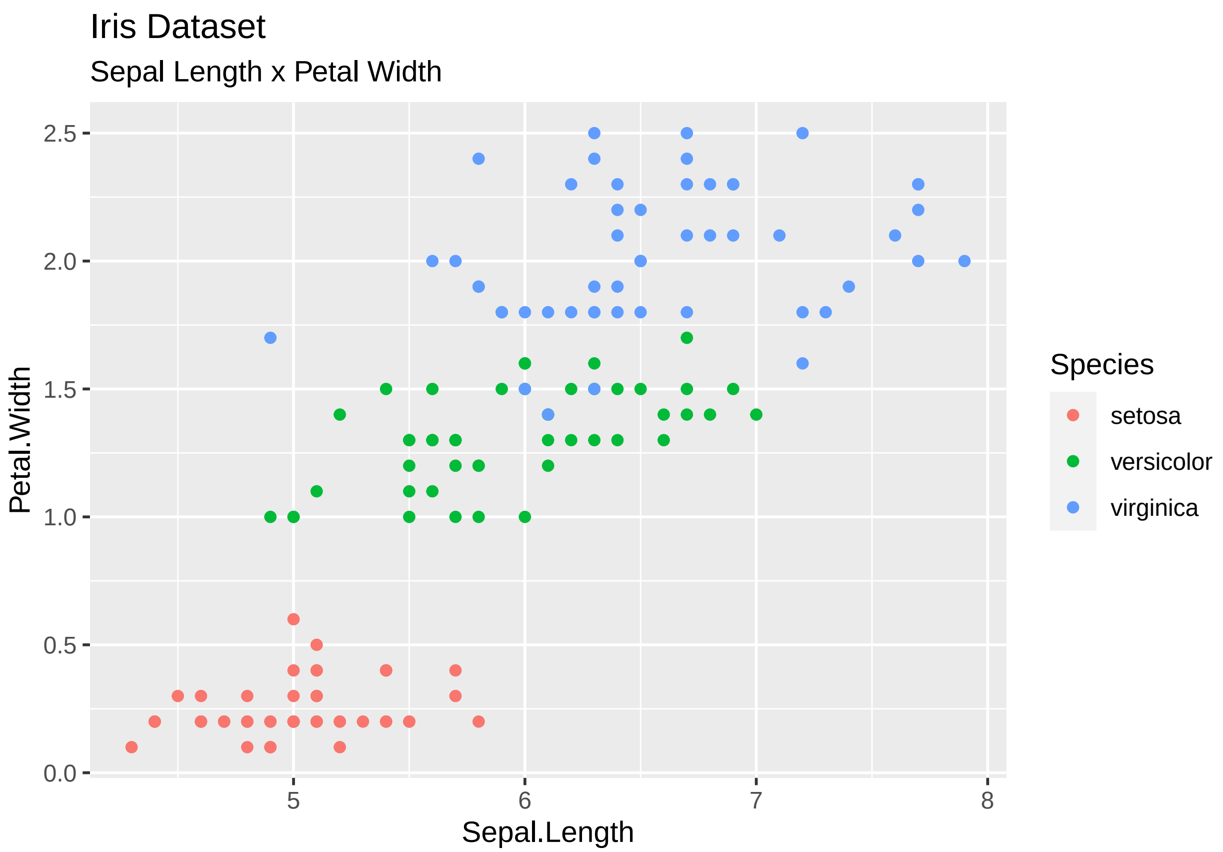
Again, *setosa* has the most characteristic pattern, whereas *versicolor* and *virginica* are more similar to each other.

Scatter plots

Comparing measurements is a way to get a sense of the data. Plot the columns in the data frame against each other.


Iris Dataset
Sepal Length x Width
Sepal.Width
4.5
4.0
3.5
3.0
2.5
2.0
5
6
7
8
Sepal.Length
Species
setosa
versicolor
virginica


Iris Dataset
Sepal Length x Petal Length
Petal.Length
6
4
2
5
6
7
8
Sepal.Length
Species
setosa
versicolor
virginica

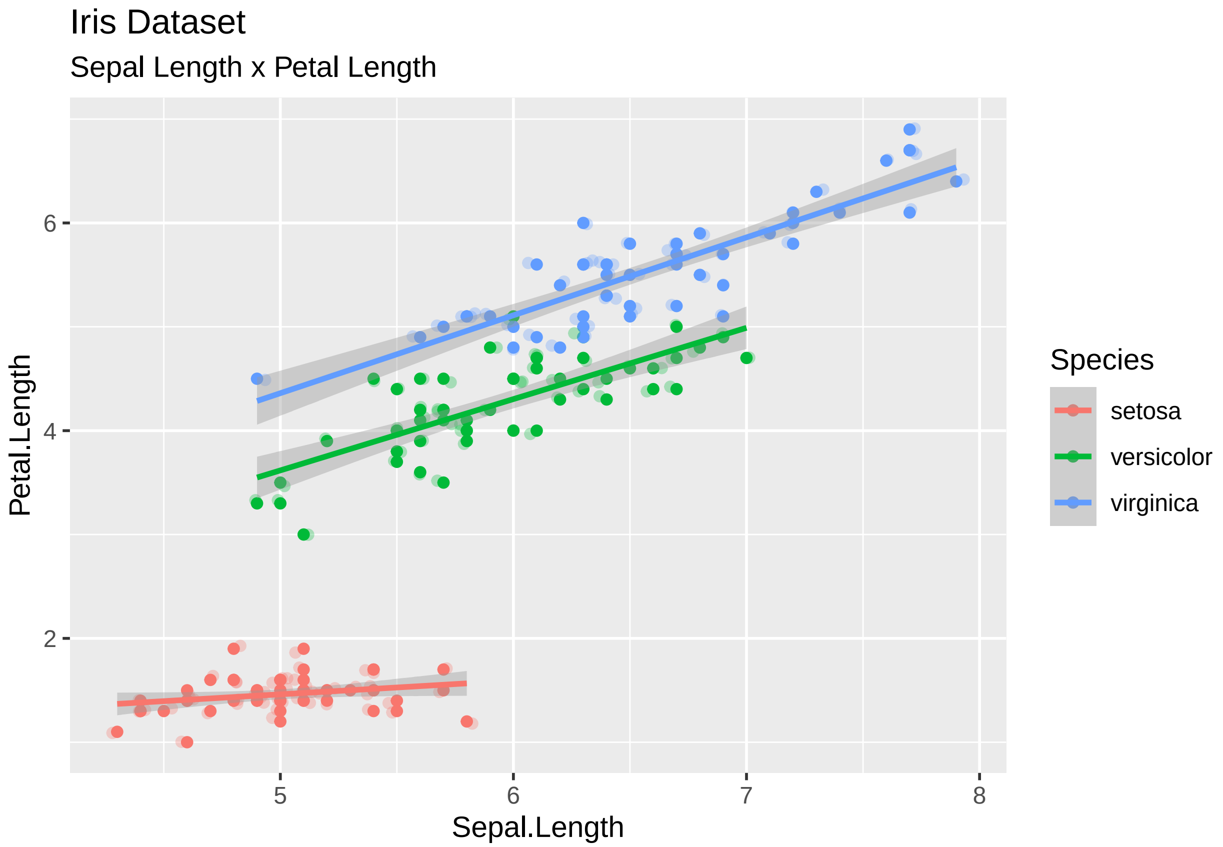


Different sepal to petal comparisons seem to separate better than others.

Adding a trendline with `stat_smooth()`

```
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length,col=Species)) +
  geom_point() +
  ggtitle("Iris Dataset", subtitle = "Sepal Length x Petal Length") +
  geom_point(alpha = 0.3,  position = position_jitter()) +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



It seems that petal length is tightly distributed when compared to sepal length. Perhaps these relationships could be combined to help distinguish these species?

Conclusion

A combination of sepal and petal measurements may elicit the best way of distinguishing *versicolor* from *virginica*, whereas *setosa* is quite distinct in these properties.

Bibliography

Anderson, Edgar. 1936. "The Species Problem in Iris." *Annals of the Missouri Botanical Garden* 23 (3): 457–509. http://www.jstor.org/stable/2394164.

FISHER, R. A. 1936. "THE Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88. https://doi.org/https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.