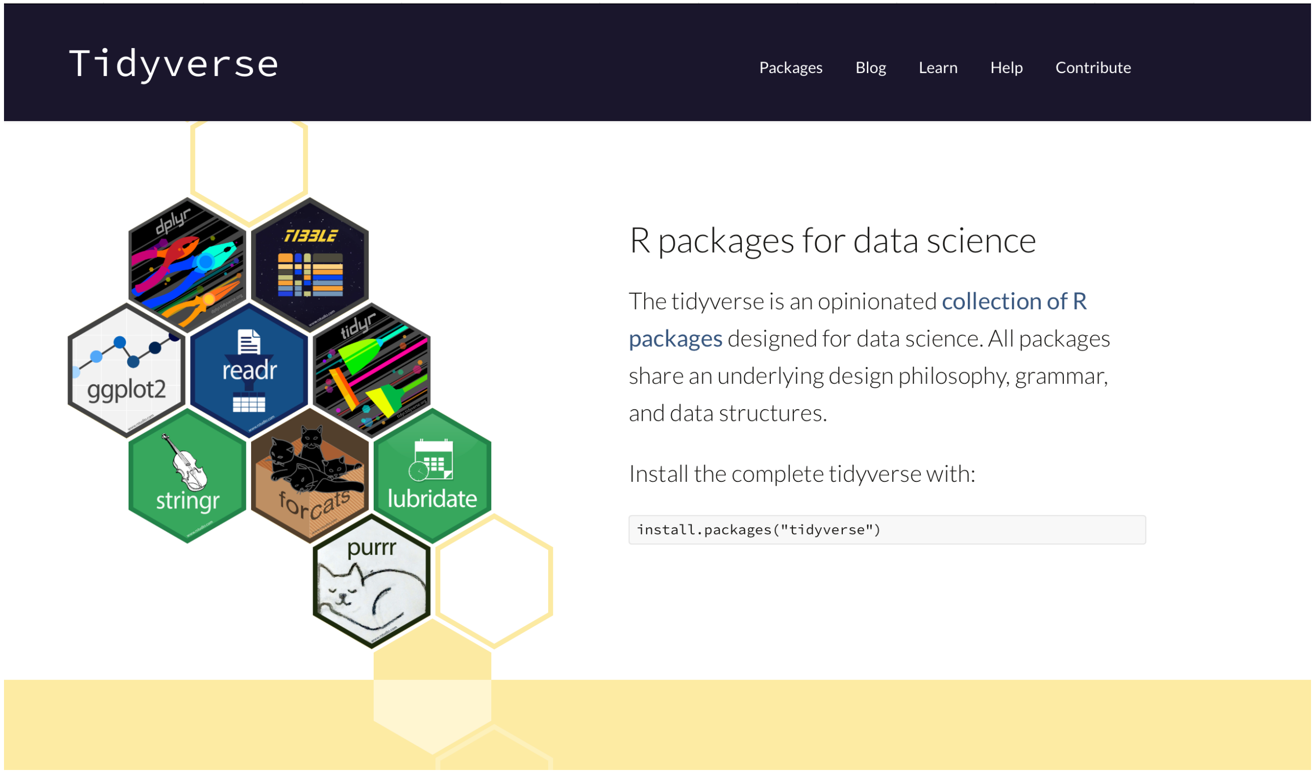
Demo Introduction to Data Wrangling with `dplyr`



Tidyverse

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

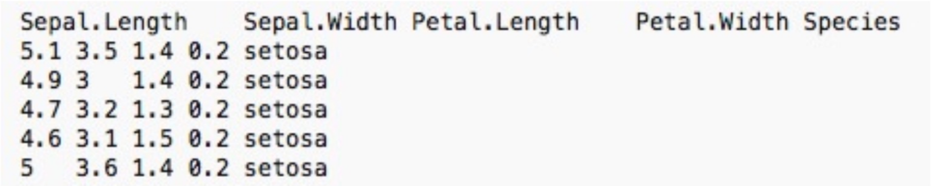
```
install.packages("tidyverse")
```

Reading Data with `readr` and Tidying Data with `tidyr`

Data File Formats

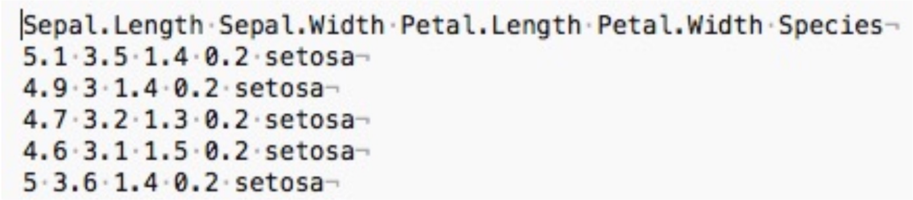
Data is stored in plain text files with a delimiter specifying the boundaries between data entries. The most common delimiters are tabs or commas.

tab separated values (TSV)



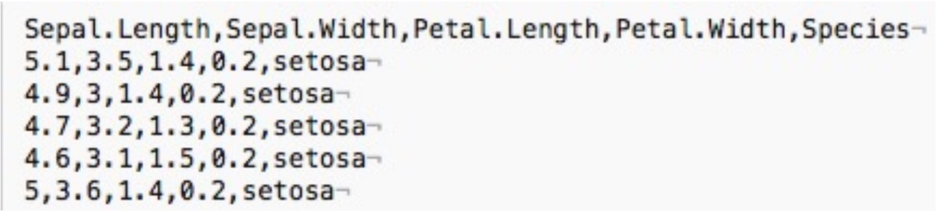
```
Sepal.Length    Sepal.Width Petal.Length    Petal.Width Species
5.1 3.5 1.4 0.2 setosa
4.9 3   1.4 0.2 setosa
4.7 3.2 1.3 0.2 setosa
4.6 3.1 1.5 0.2 setosa
5   3.6 1.4 0.2 setosa
```

spaces



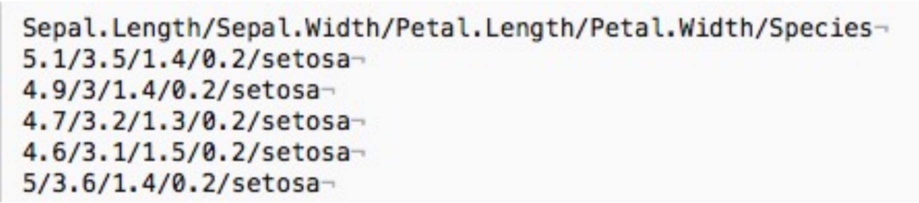
```
Sepal.Length·Sepal.Width·Petal.Length·Petal.Width·Species¬
5.1·3.5·1.4·0.2·setosa¬
4.9·3·1.4·0.2·setosa¬
4.7·3.2·1.3·0.2·setosa¬
4.6·3.1·1.5·0.2·setosa¬
5·3.6·1.4·0.2·setosa¬
```

comma separated values (CSV)



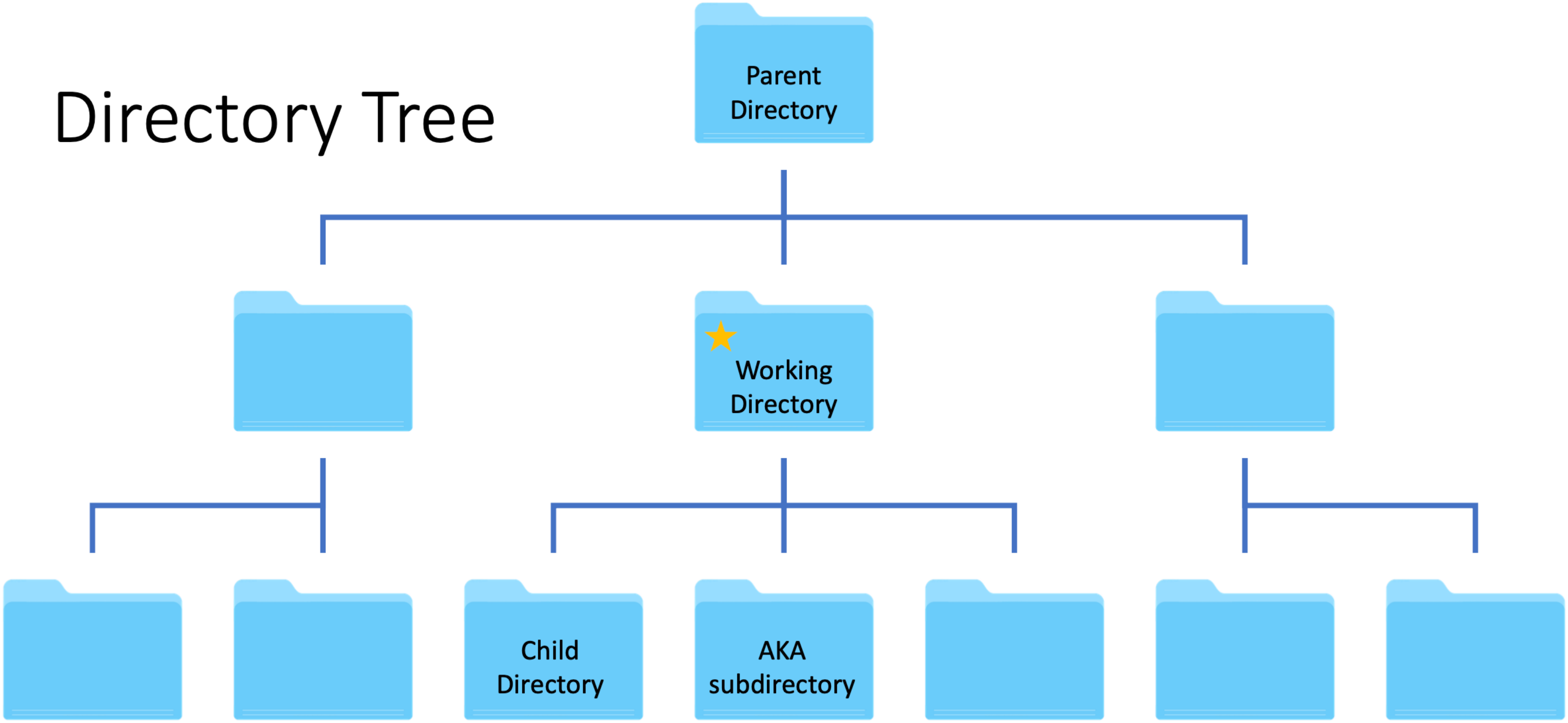
```
Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Species¬
5.1,3.5,1.4,0.2,setosa¬
4.9,3,1.4,0.2,setosa¬
4.7,3.2,1.3,0.2,setosa¬
4.6,3.1,1.5,0.2,setosa¬
5,3.6,1.4,0.2,setosa¬
```

Or any other character (BUT NEVER DO THIS)



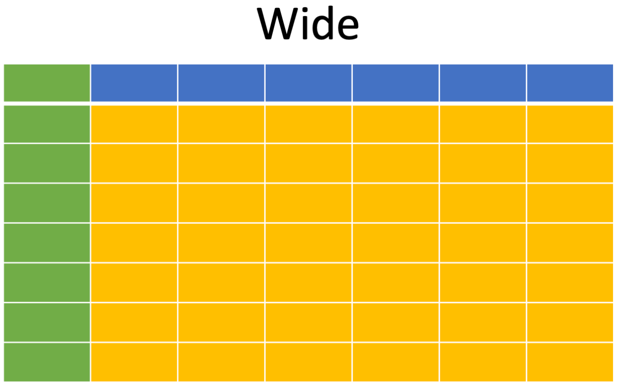
```
Sepal.Length/Sepal.Width/Petal.Length/Petal.Width/Species¬
5.1/3.5/1.4/0.2/setosa¬
4.9/3/1.4/0.2/setosa¬
4.7/3.2/1.3/0.2/setosa¬
4.6/3.1/1.5/0.2/setosa¬
5/3.6/1.4/0.2/setosa¬
```

Directory Tree

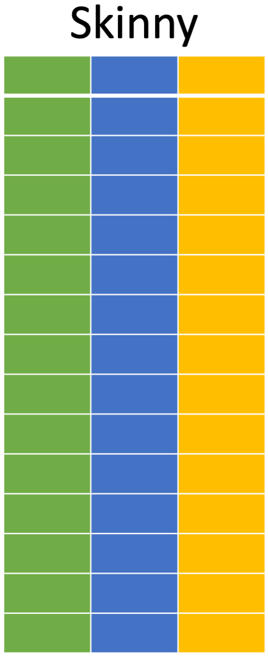


File Path: `working_directory/child_directory`

Wide vs Skinny Data



Wide



Skinny

Demo

Plotting with `ggplot2`

Quick Review: Types of Variables

Categorical

A **categorical** variable is a variable with a limited number of fixed descriptions; basically a label.

- unordered
  - No natural ordering
  - Ex: sample IDs, genotypes, phenotypes
- ordered
  - Natural way to order them
  - Ex: survery responses (poor, fine, ok, very good, good), chromomsomes (chr1, chr2, ch3, etc.)

Numeric

- discrete
  - Values are indivisible (or dividing them makes no sense); aka count data.
  - Ex: counts of people, read counts
- continuous
  - Values can be divided and expressing them as a divided value, even if the divisions aren’t necessary are present, is fine.
  - Ex: height, weight

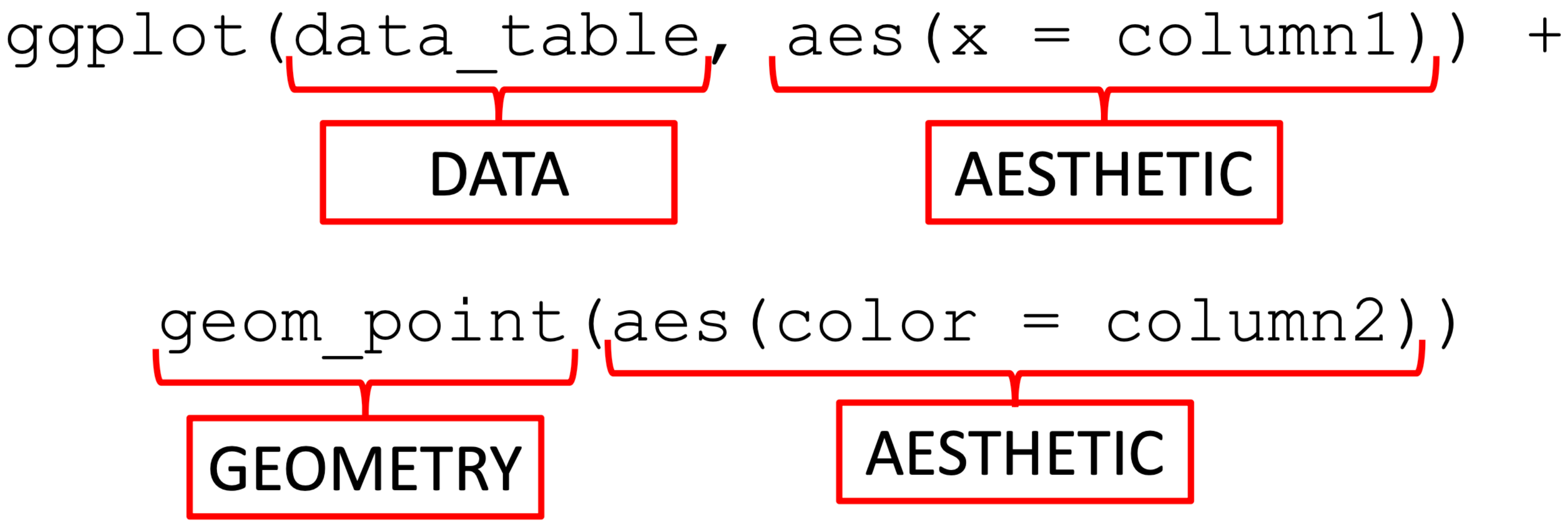
ggplot2

“ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.”

grammar	description
data	The table you want to visualize
geometry	What shape you want to give that visualization, ex: scatter plot, boxplot, violin plot, bar plot, histogram, density plot
aesthetic	The appearance of the geometry, ex: size, shape, color

The philosophy of `ggplot`

Data, geometry, and aesthetics are **independent.**



```
ggplot(data_table, aes(x = column1)) +
      geom_point(aes(color = column2))
```

DATA

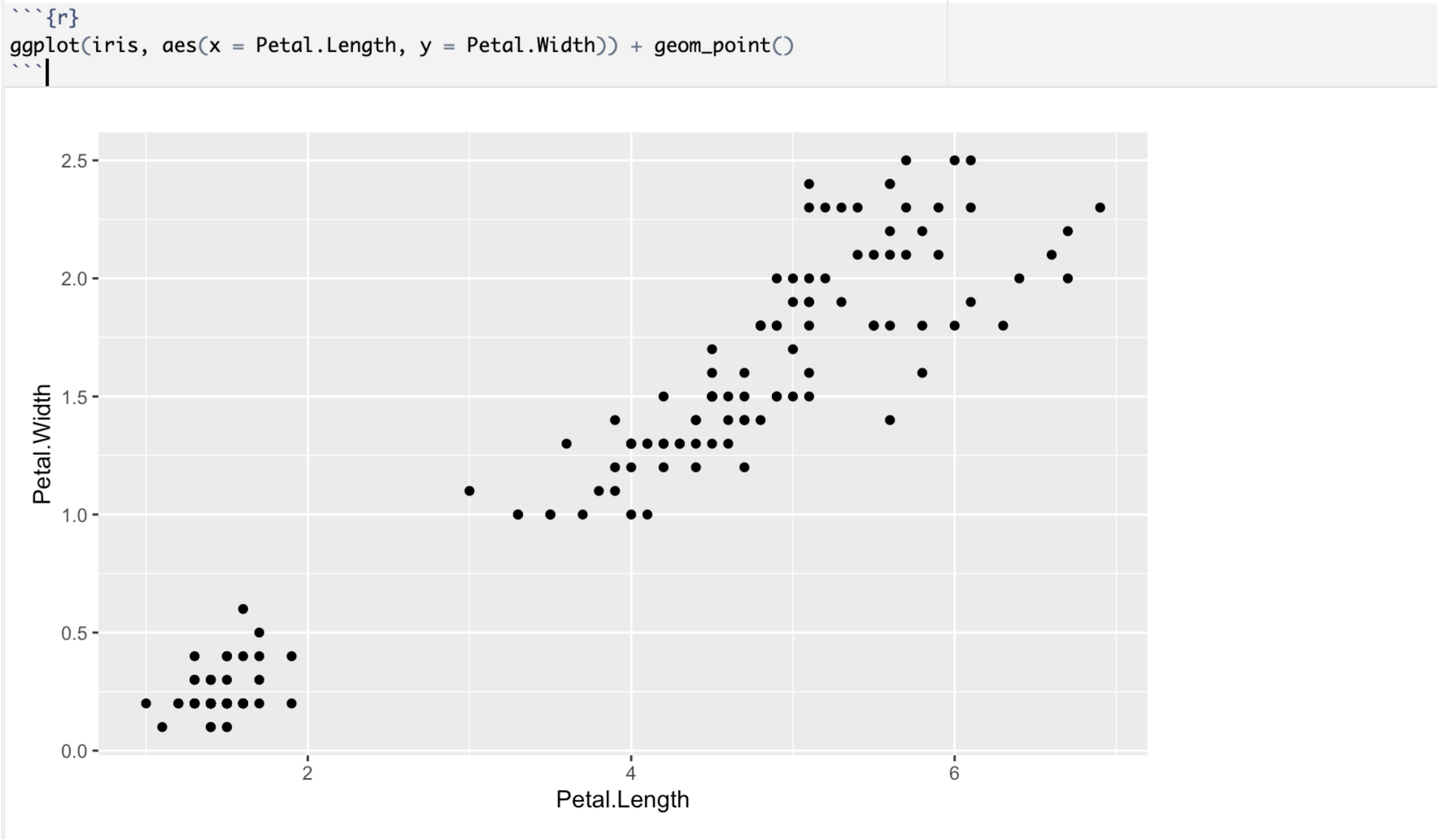
AESTHETIC

GEOMETRY

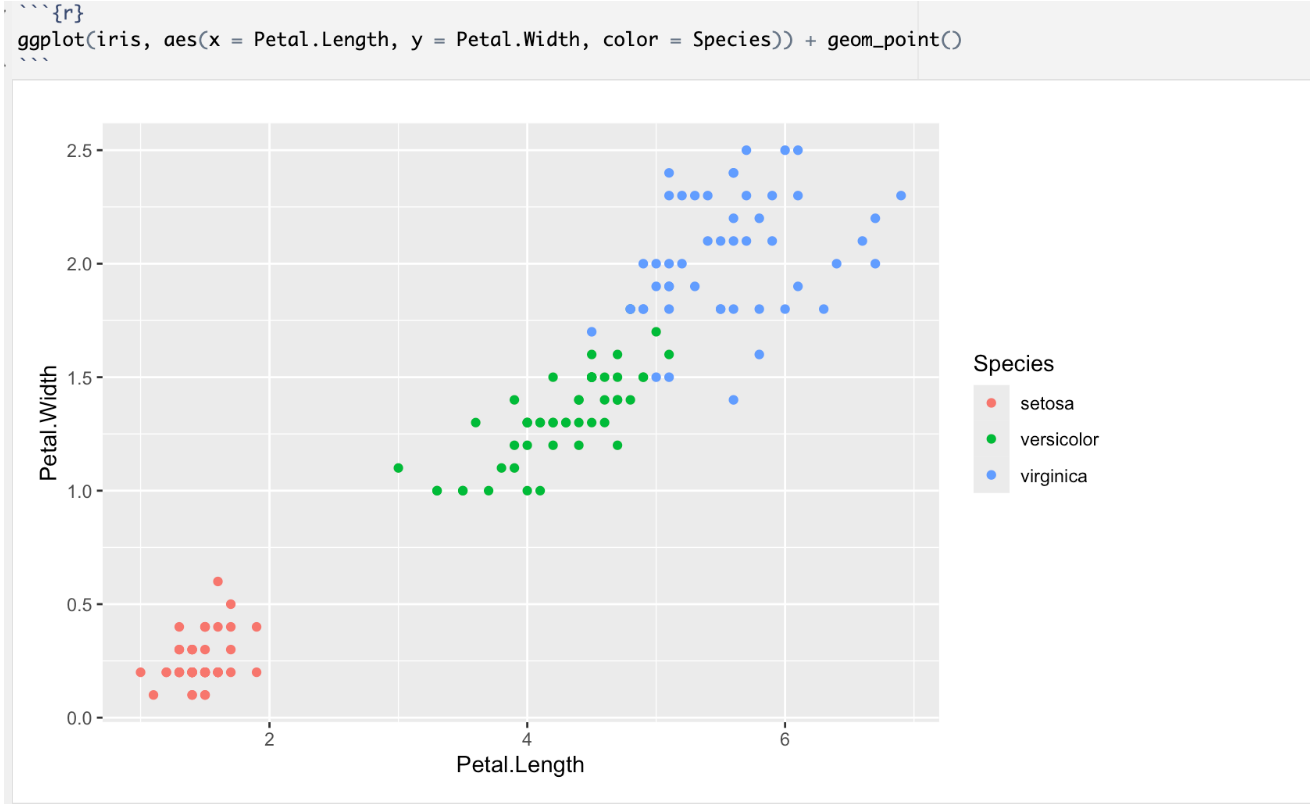
AESTHETIC

Demo

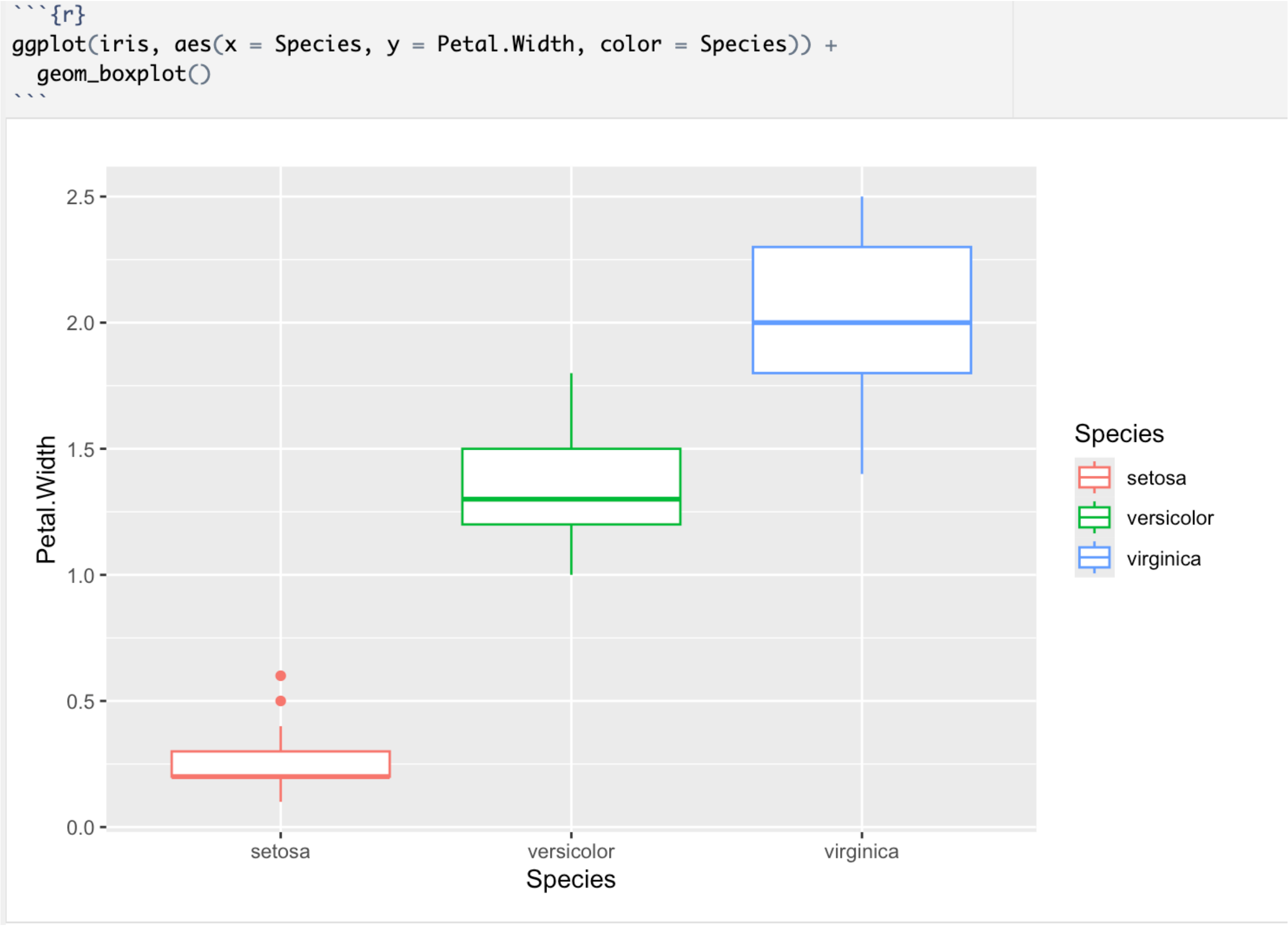
Scatter plot



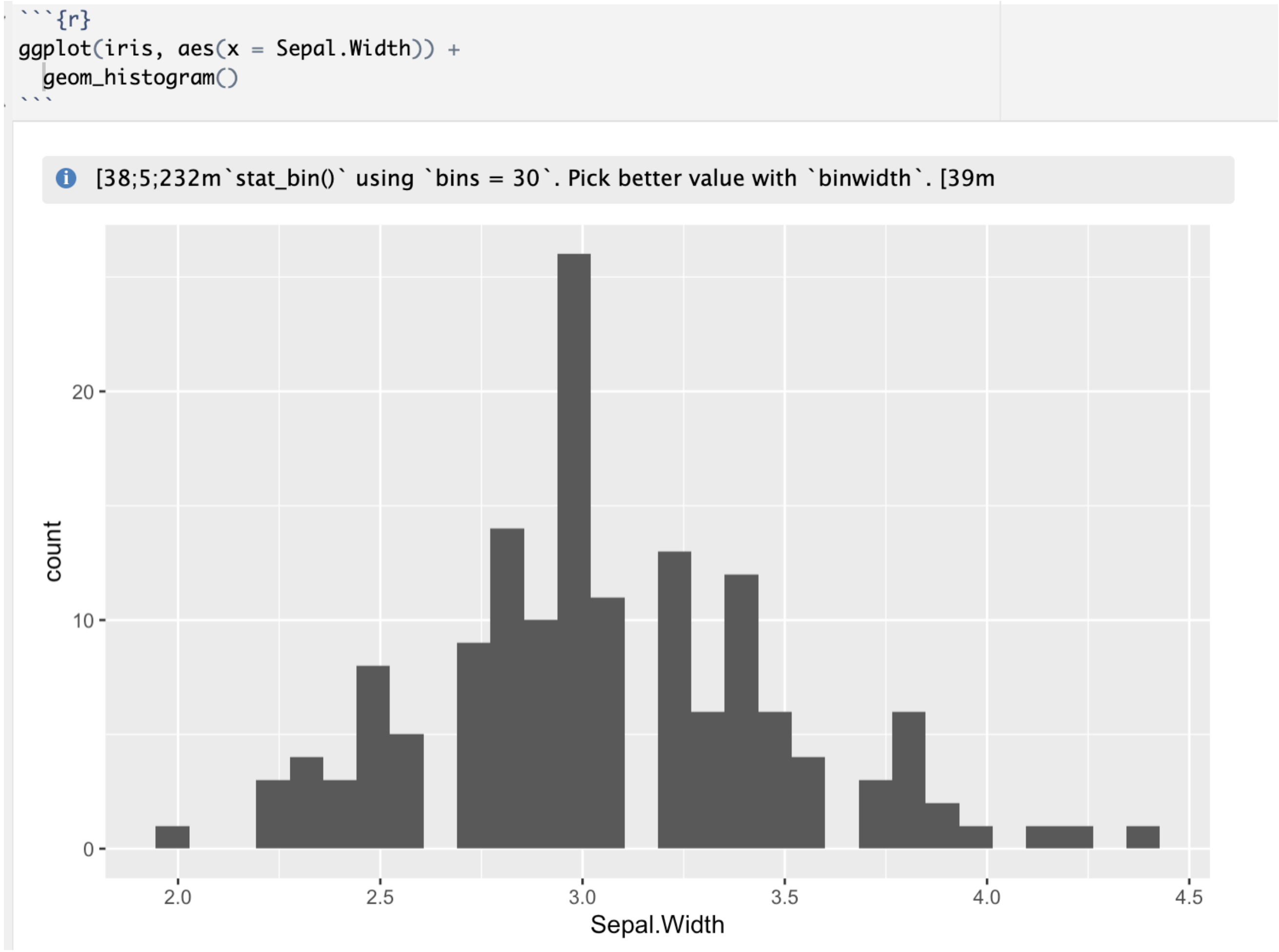
```{r}
ggplot(iris, aes(x = Petal.Length, y = Petal.Width)) + geom_point()
```



~~~
```{r}
ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species)) + geom_point()
```
~~~

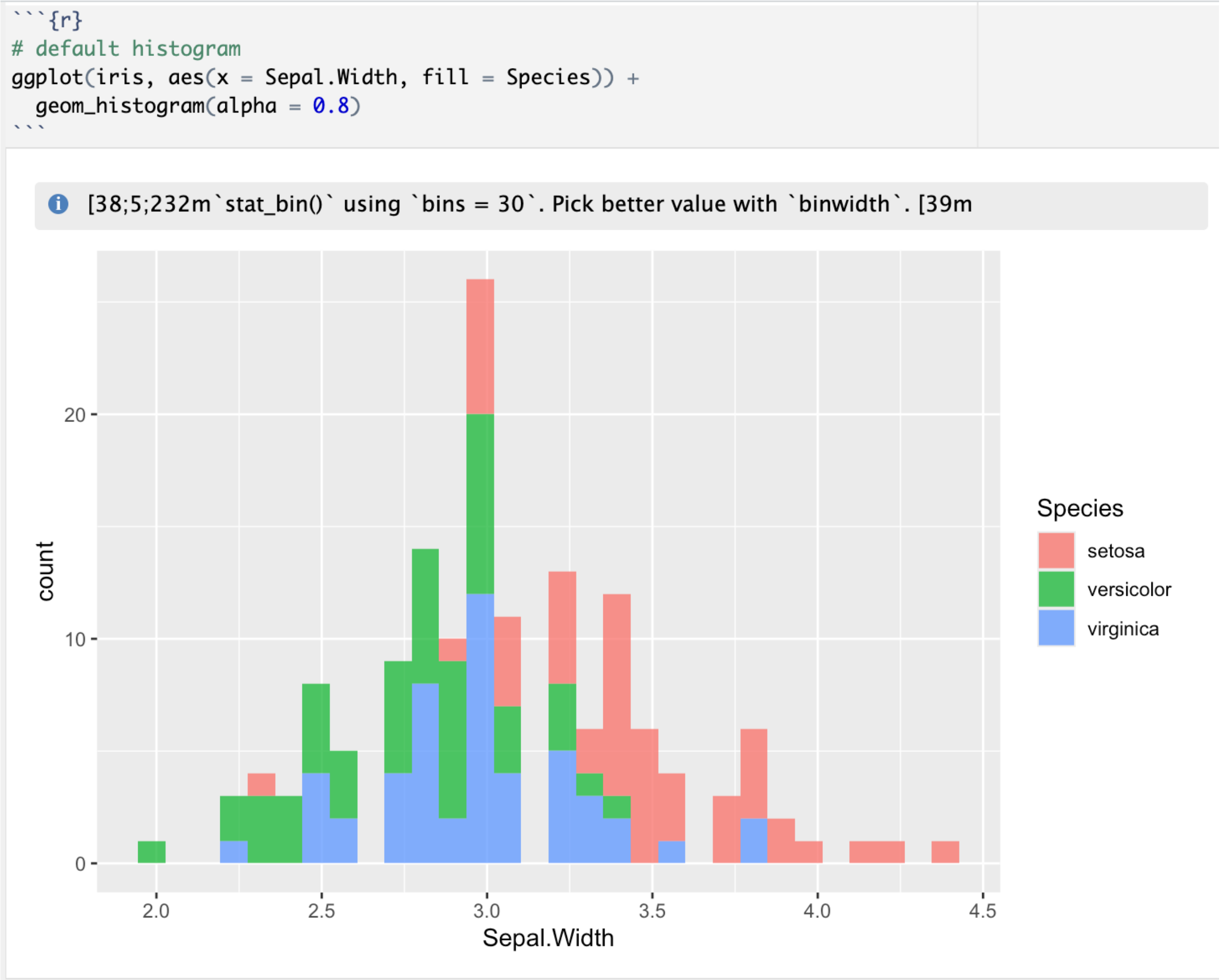


~~~
```{r}
ggplot(iris, aes(x = Species, y = Petal.Width, color = Species)) +
  geom_boxplot()
```
~~~



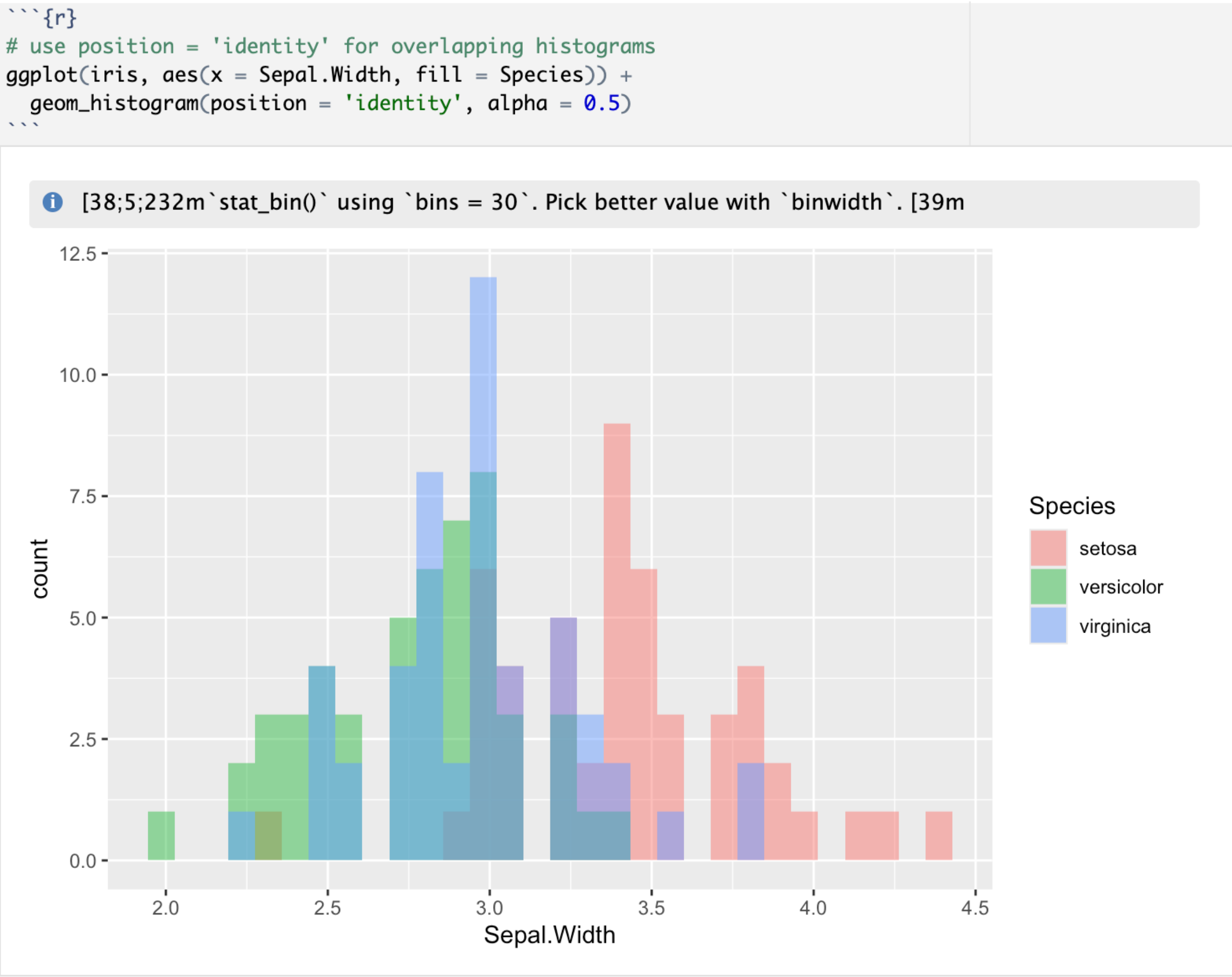
~~~~
```{r}
ggplot(iris, aes(x = Sepal.Width)) +
  geom_histogram()
```
~~~~

```
[38;5;232m`stat_bin()` using `bins = 30`. Pick better value with `binwidth`. [39m
```



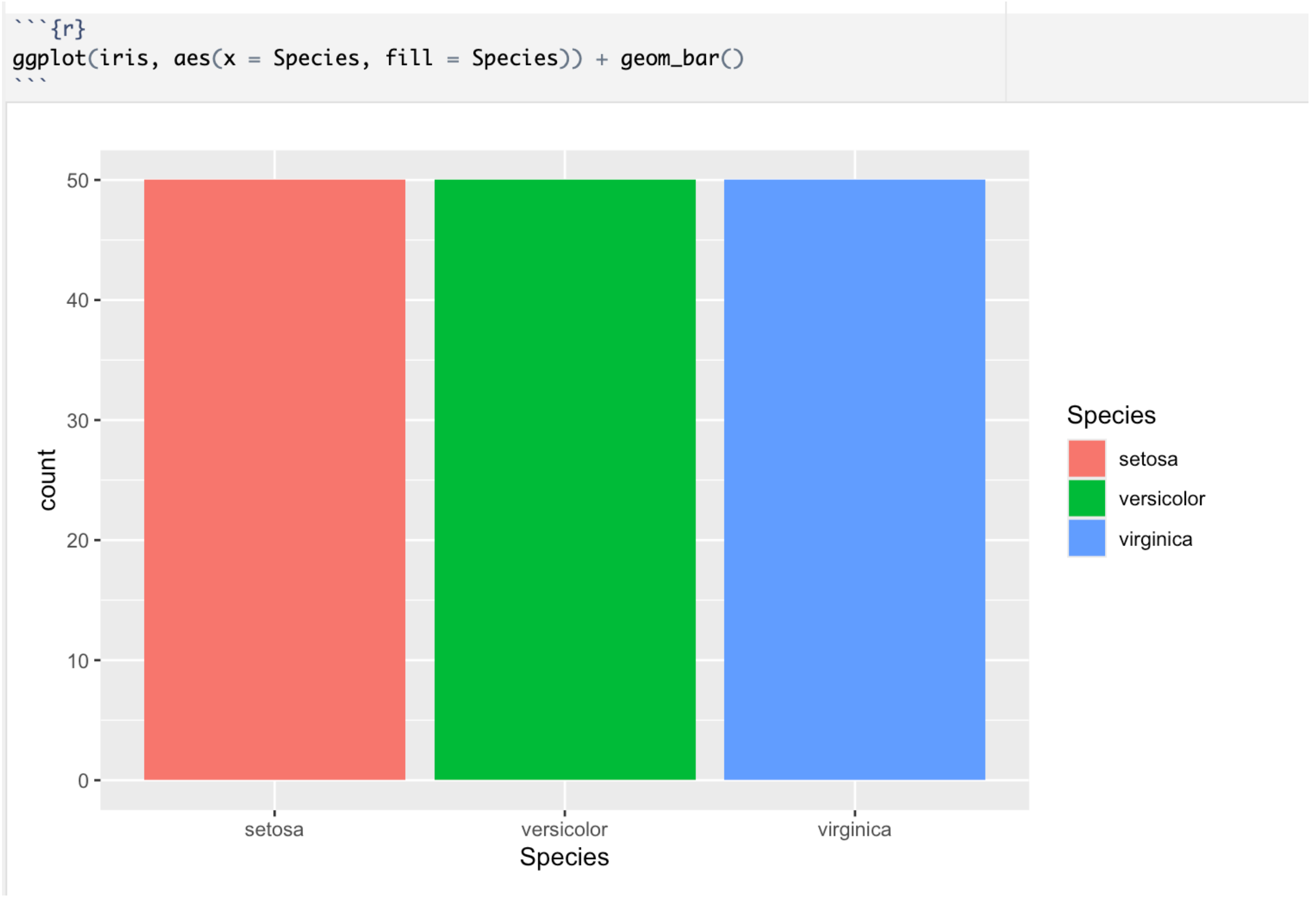
~~~
```{r}
# default histogram
ggplot(iris, aes(x = Sepal.Width, fill = Species)) +
  geom_histogram(alpha = 0.8)
```
~~~

```
[38;5;232m`stat_bin()` using `bins = 30`. Pick better value with `binwidth`. [39m
```

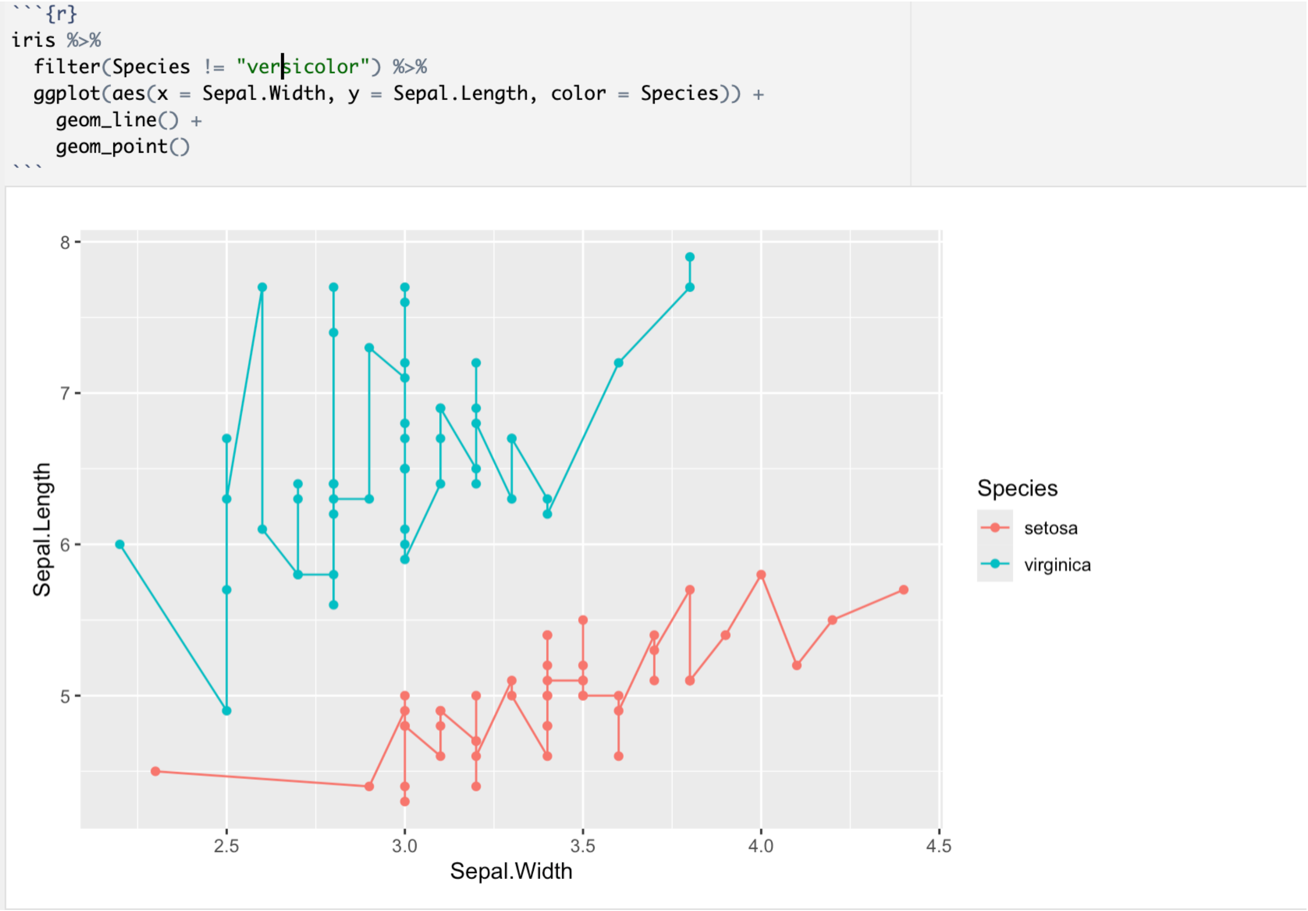


~~~
```{r}
# use position = 'identity' for overlapping histograms
ggplot(iris, aes(x = Sepal.Width, fill = Species)) +
  geom_histogram(position = 'identity', alpha = 0.5)
```
~~~

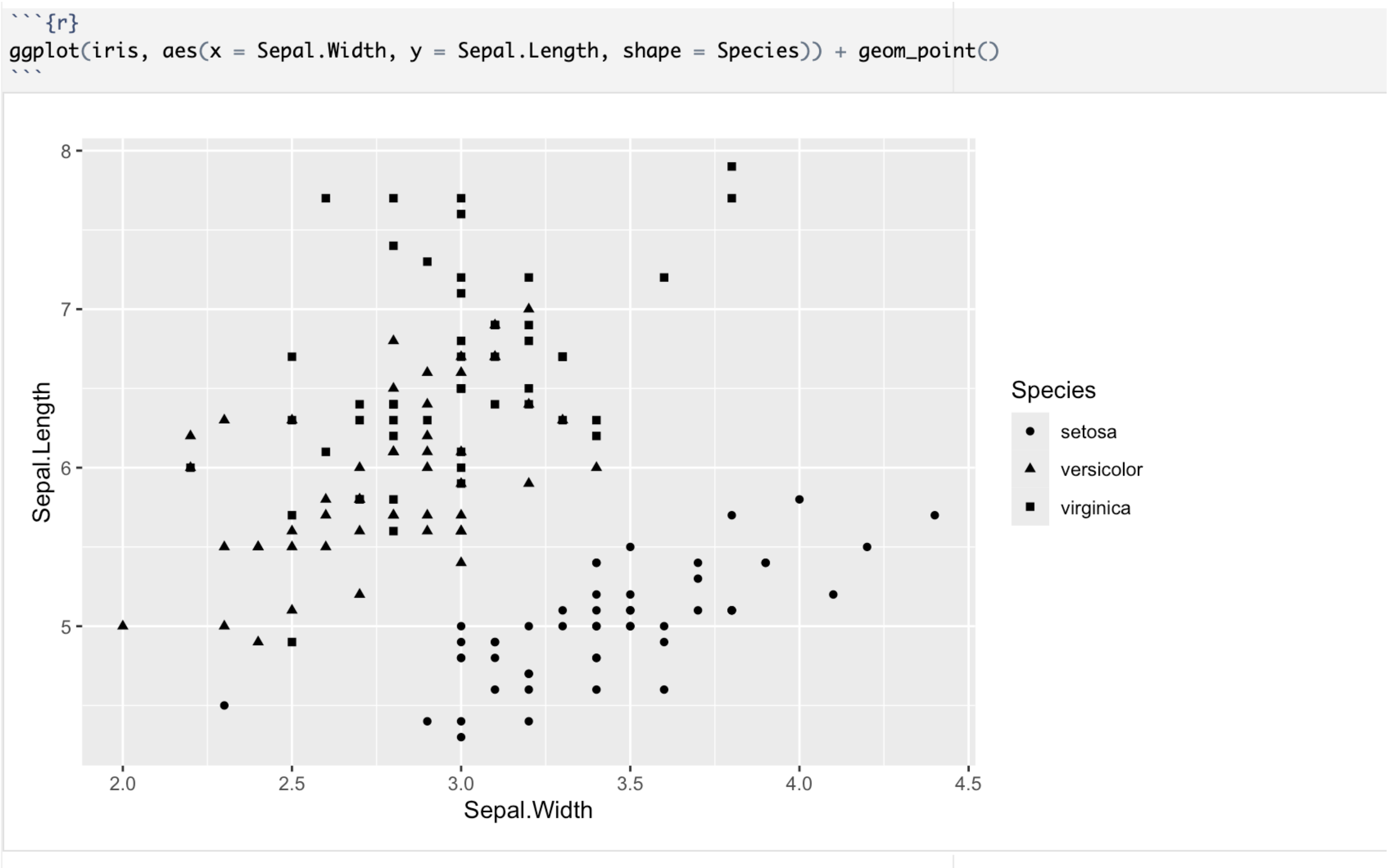
```
ℹ [38;5;232m`stat_bin()` using `bins = 30`. Pick better value with `binwidth`. [39m
```



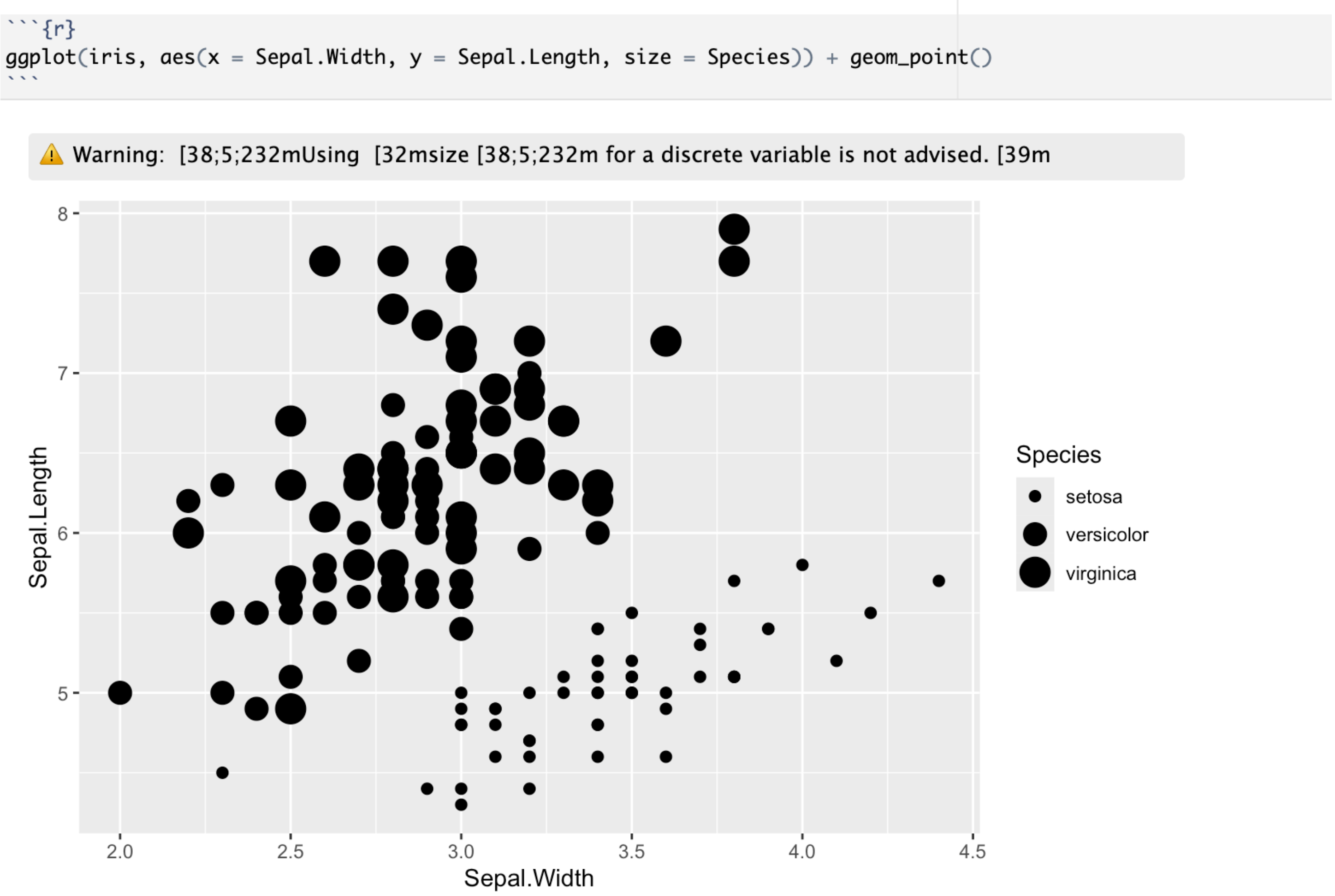
~~~~
```{r}
ggplot(iris, aes(x = Species, fill = Species)) + geom_bar()
```
~~~~



~~~
```{r}
iris %>%
  filter(Species != "versicolor") %>%
  ggplot(aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
    geom_line() +
    geom_point()
```
~~~

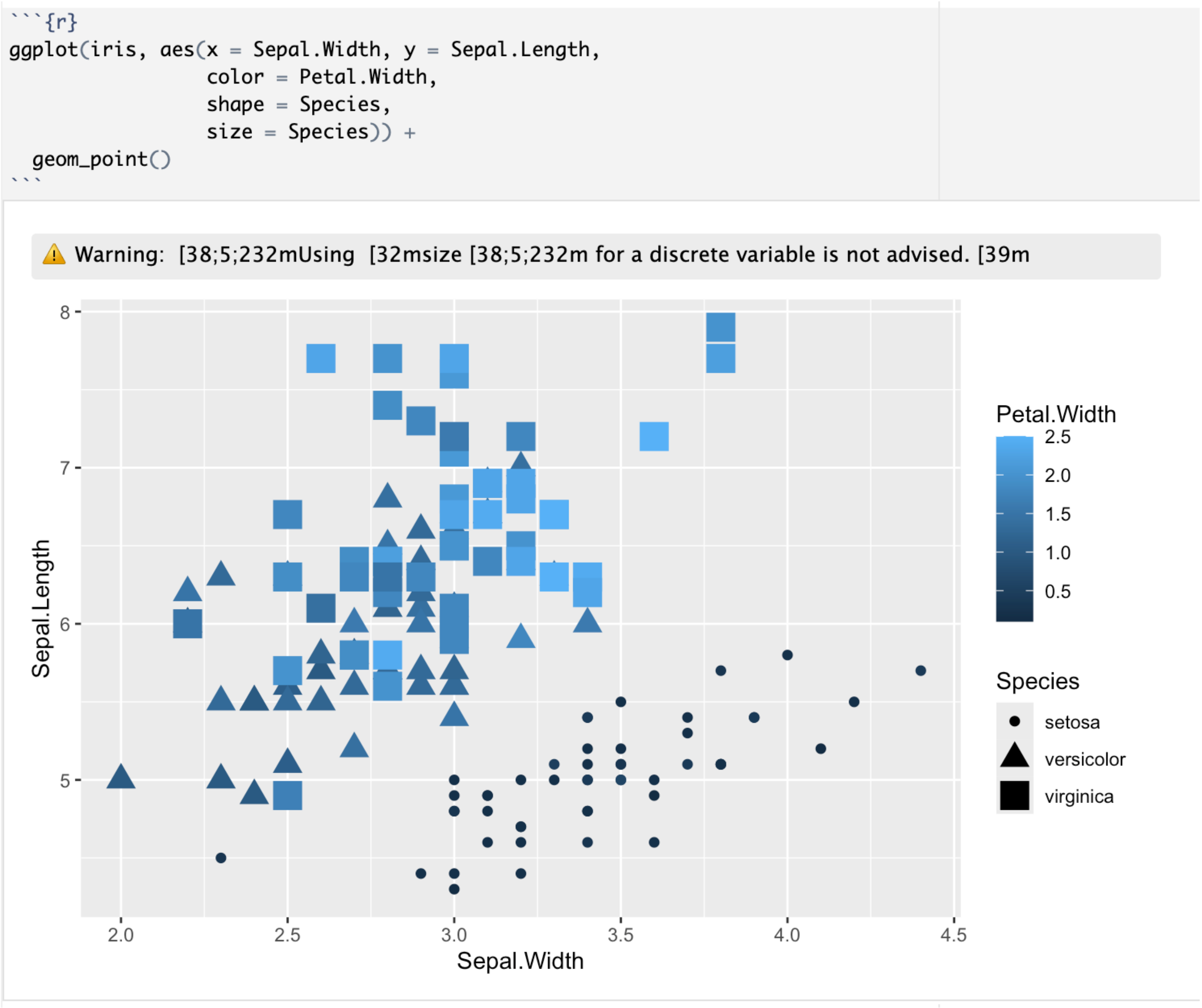


```{r}
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, shape = Species)) + geom_point()
```



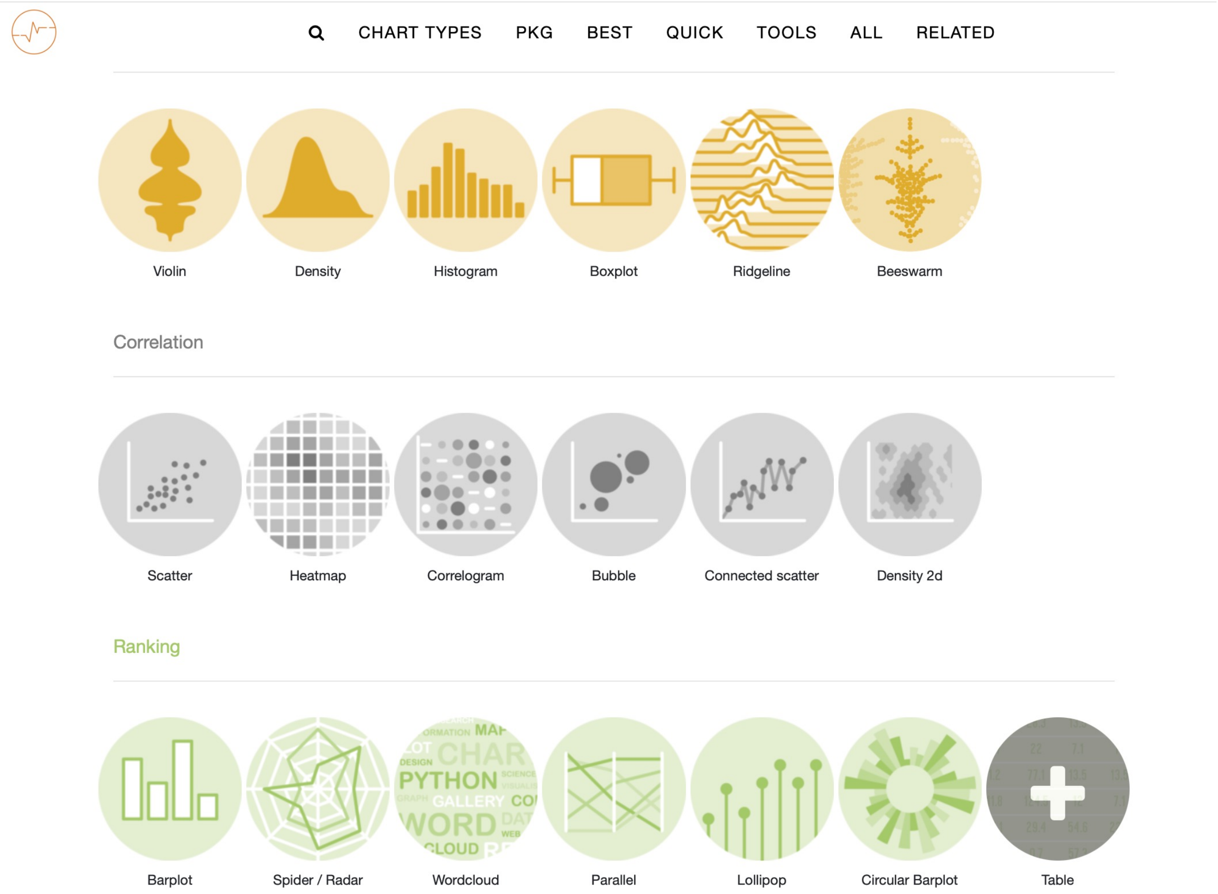
~~~
```{r}
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, size = Species)) + geom_point()
```
~~~

```
Warning:  [38;5;232mUsing  [32msize [38;5;232m for a discrete variable is not advised. [39m
```



```{r}
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length,
                 color = Petal.Width,
                 shape = Species,
                 size = Species)) +
  geom_point()
```

⚠ Warning: [38;5;232mUsing [32msize [38;5;232m for a discrete variable is not advised. [39m



https://r-graph-gallery.com