Correlation

Frank Edwards

10/26/2021

Correlation (math time): Z-scores

First, we need the variables to be comparable, so we transform them to be on a standard deviation scale.

A *z*-score scales a variable measures the number of standard deviations an observation is away from it's mean.

$$z \text{ score of } x_i = \frac{x_i - \bar{x}}{S_x}$$

Where $\bar{x}$ is the mean, and S_x is the standard deviation of variable x. Z scores have a mean zero, and a range defined by the range of the data on a standard deviation scale.

For a normally (Gaussian) distributed variable, this will typically range between $[-3, 3]$

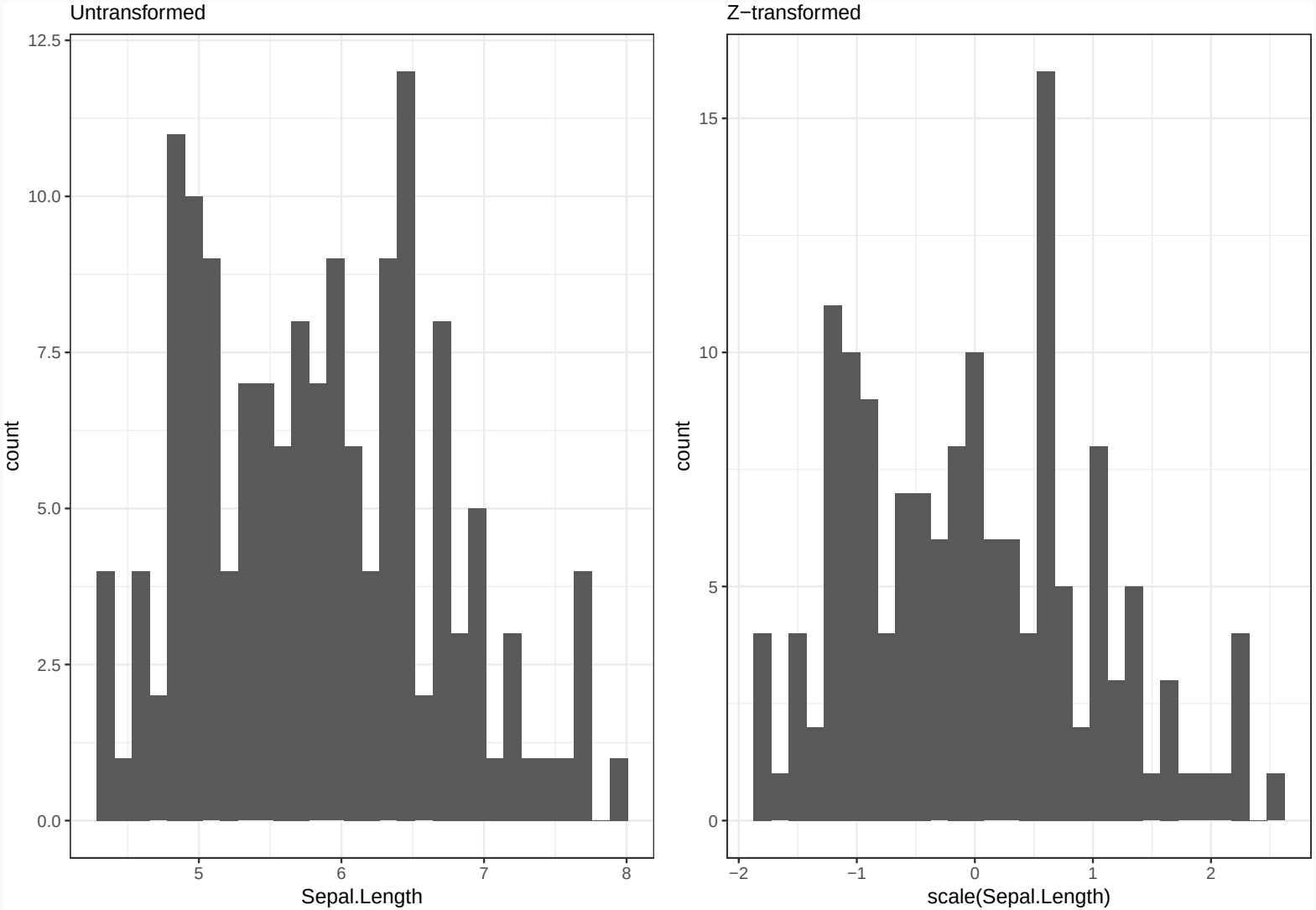
In R, we can transform a numeric into a *z*-score using `scale()`

Z-scores in R

```
iris %>%
  mutate(Sepal.Length.sc = scale(Sepal.Length)) %>%
  select(Sepal.Length, Sepal.Length.sc)
```

```
##     Sepal.Length Sepal.Length.sc
## 1            5.1     -0.89767388
## 2            4.9     -1.13920048
## 3            4.7     -1.38072709
## 4            4.6     -1.50149039
## 5            5.0     -1.01843718
## 6            5.4     -0.53538397
## 7            4.6     -1.50149039
## 8            5.0     -1.01843718
## 9            4.4     -1.74301699
## 10           4.9     -1.13920048
## 11           5.4     -0.53538397
## 12           4.8     -1.25996379
```

Z-score transformed distributions have the same shape as the original data



Correlation

Correlation measures the degree to which two variables are associated with each other. We often use the letter r to denote a correlation.

$$r(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_x} \times \frac{y_i - \bar{y}}{S_y}$$
$$= E[z(x) \times z(y)]$$

In R, you can use `cor()`

Evaluate correlations using cor()

- Compute the correlation between Sepal.Length and Sepal.Width. What does it mean?
- Compute the correlation between Petal.Length and Petal.Width. What does it mean?
- Compute the correlation between Petal.Length and Sepal.Width. What does it mean?

Bivariate visuals for continuous data: Scatterplots

Make a scatterplot

```
ggplot(iris,
       aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point()
```

Scatterplot ingredients

- a `data.frame` with two continuous variables
- `aes()` with an x and y parameter
- `geom_point()`

Practice

- Scatterplot petal length on the x and petal width on the y
- Flip the axes (move length to y, width to x)

Scatterploting with clusters

We often have bivariate measures that are *clustered*, or have some structure caused by a third (often categorical) variable.

Scatterploting with clusters

We often have bivariate measures that are *clustered*, or have some structure caused by a third (often categorical) variable.

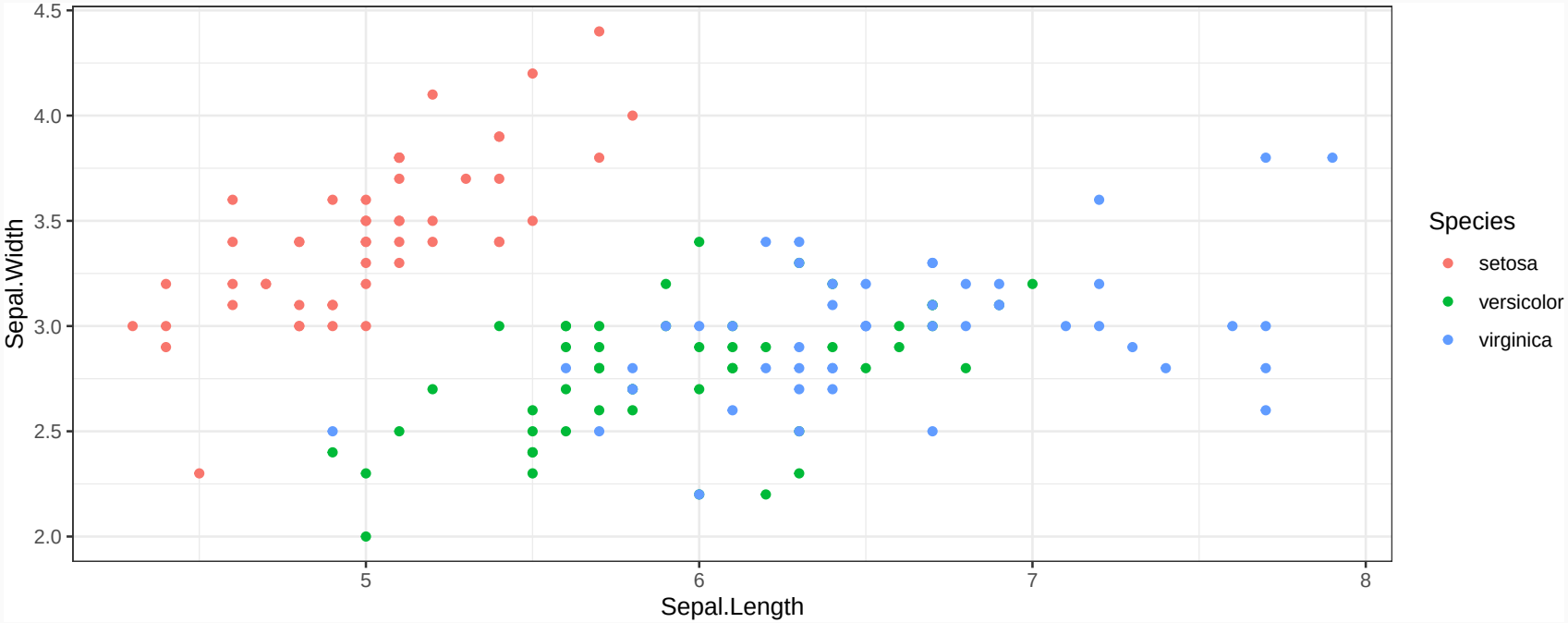
Do we see structure here? What could be causing it?

```
ggplot(iris,
       aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point()
```

Two solutions to plotting clustered data: aesthetics

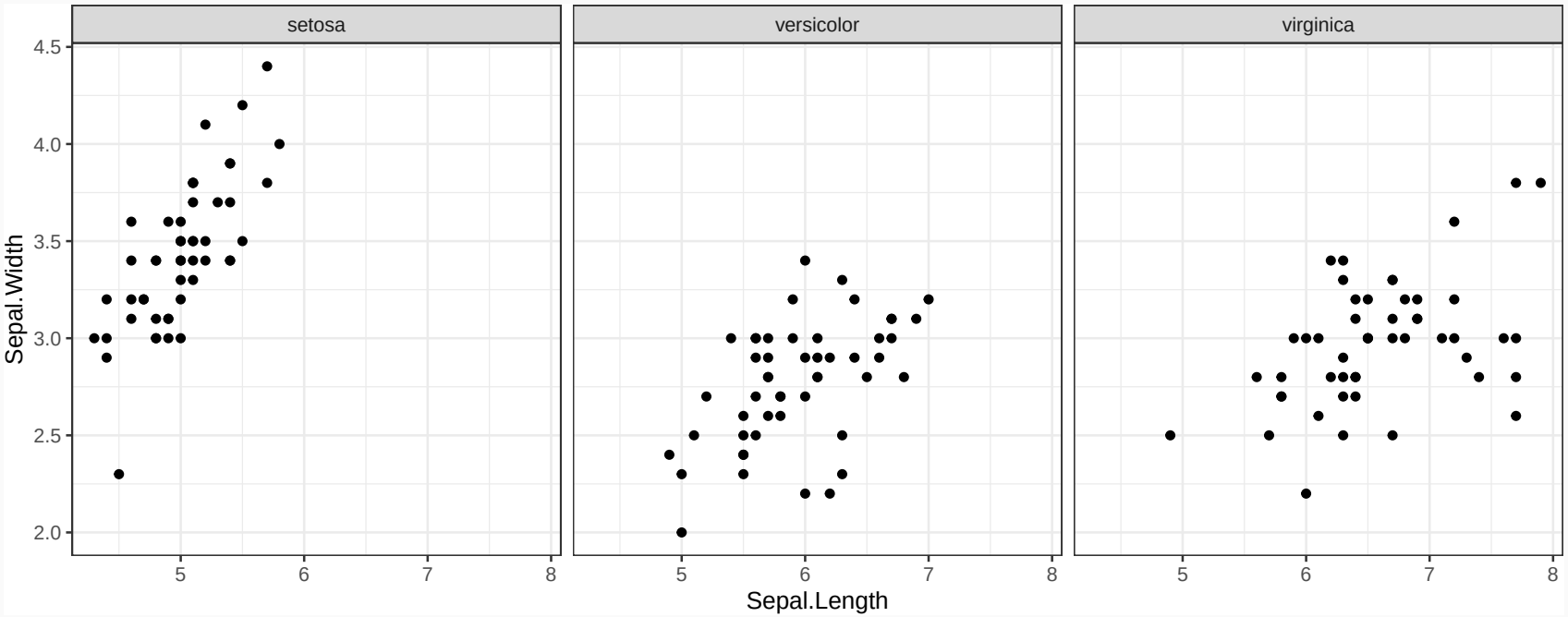
Use the clustering variable to add another aesthetic element to our plot, like color

```
ggplot(iris,
       aes(x = Sepal.Length, y = Sepal.Width,
           color = Species)) +
  geom_point()
```

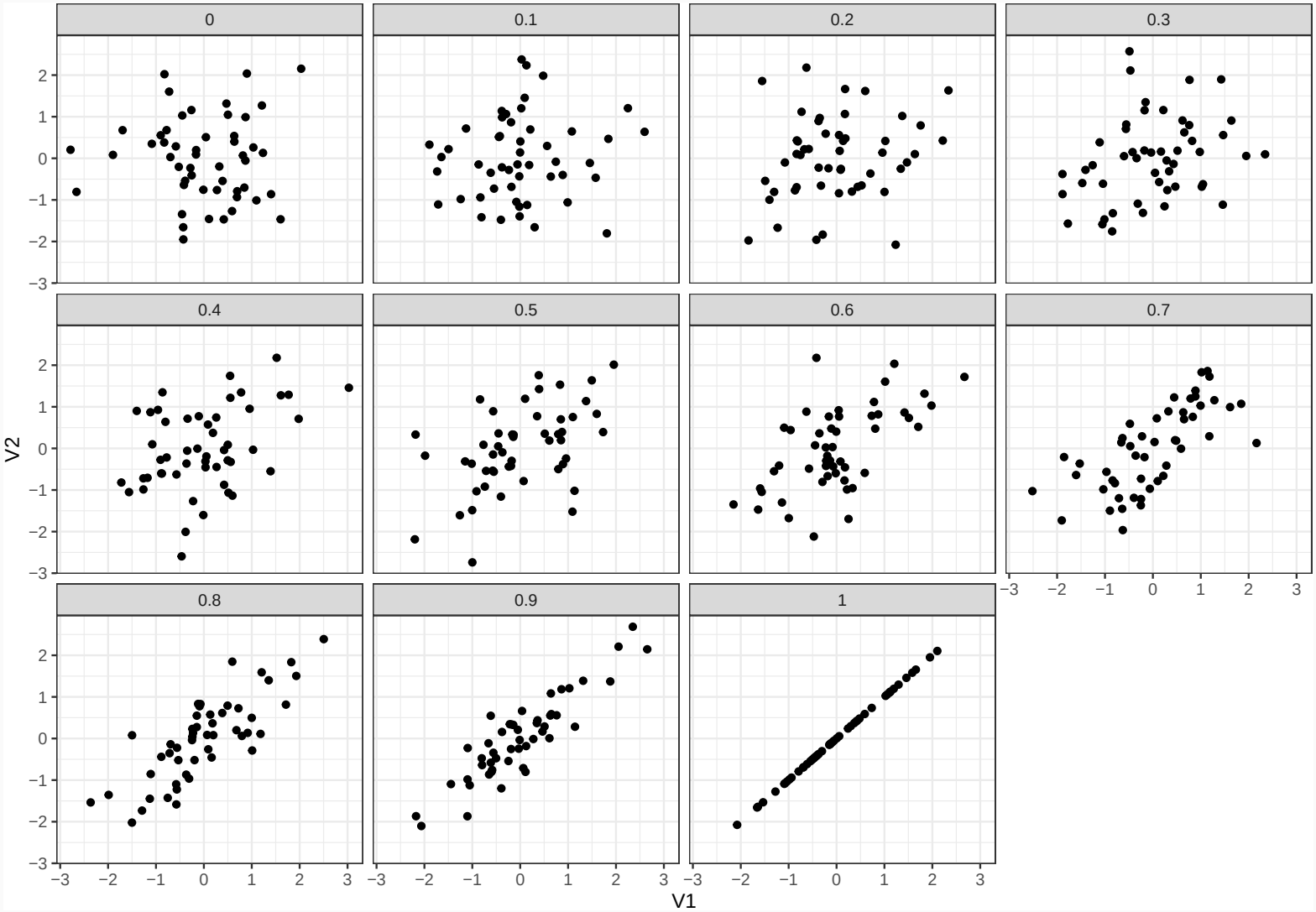


Two solutions to plotting clustered data: facets

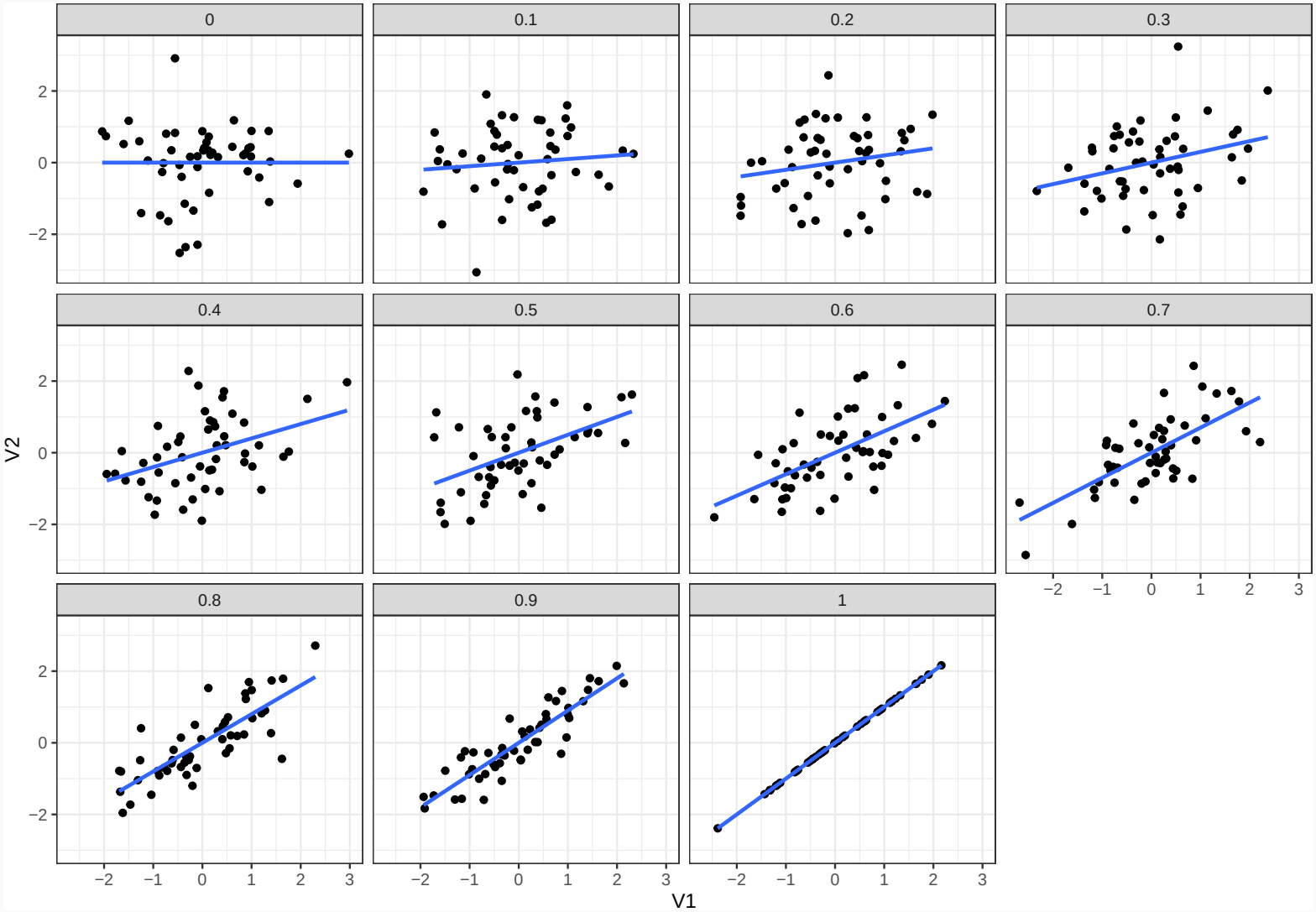
```
ggplot(iris,
       aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point() +
  facet_wrap(~Species)
```



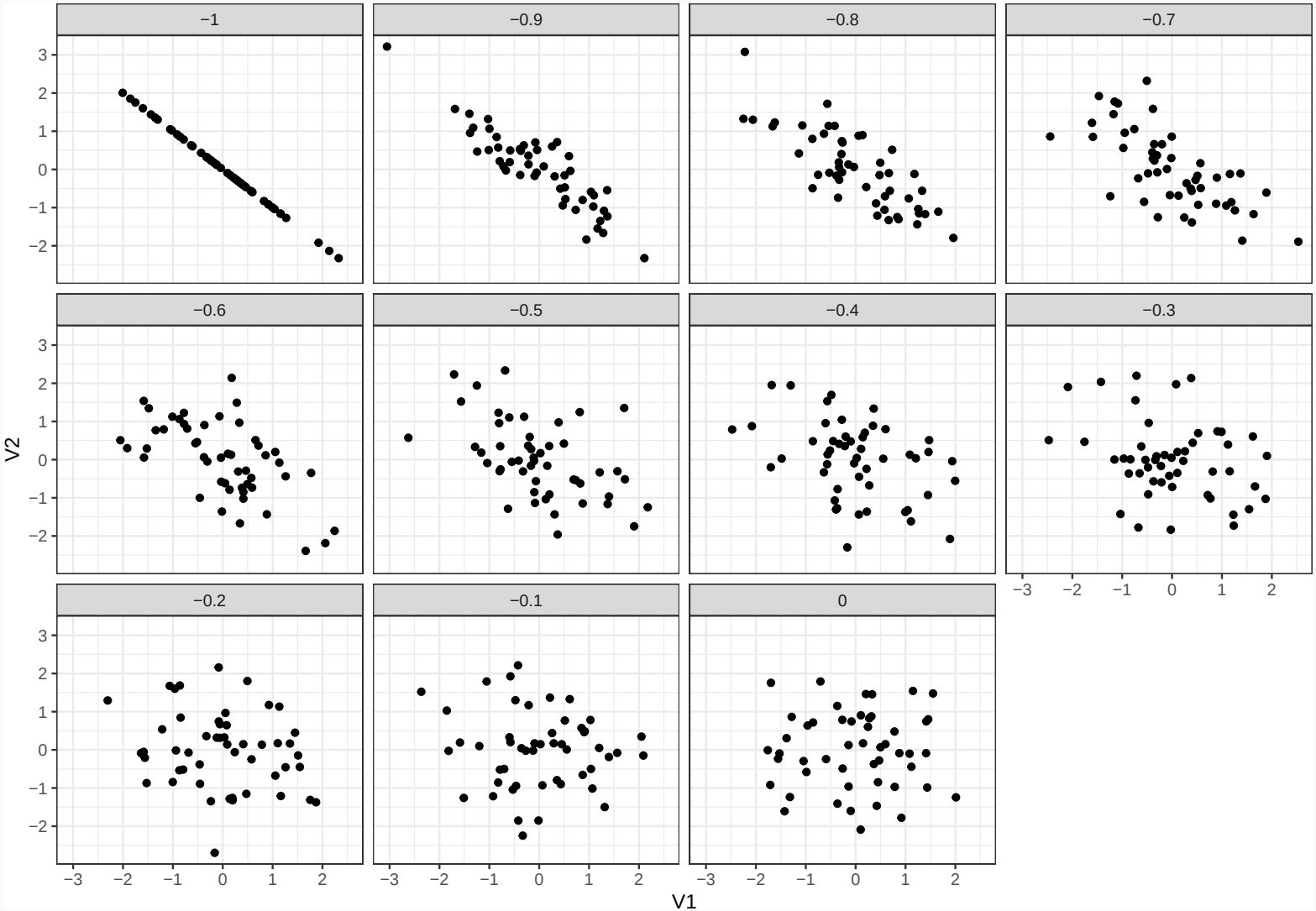
Correlation and scatterplots



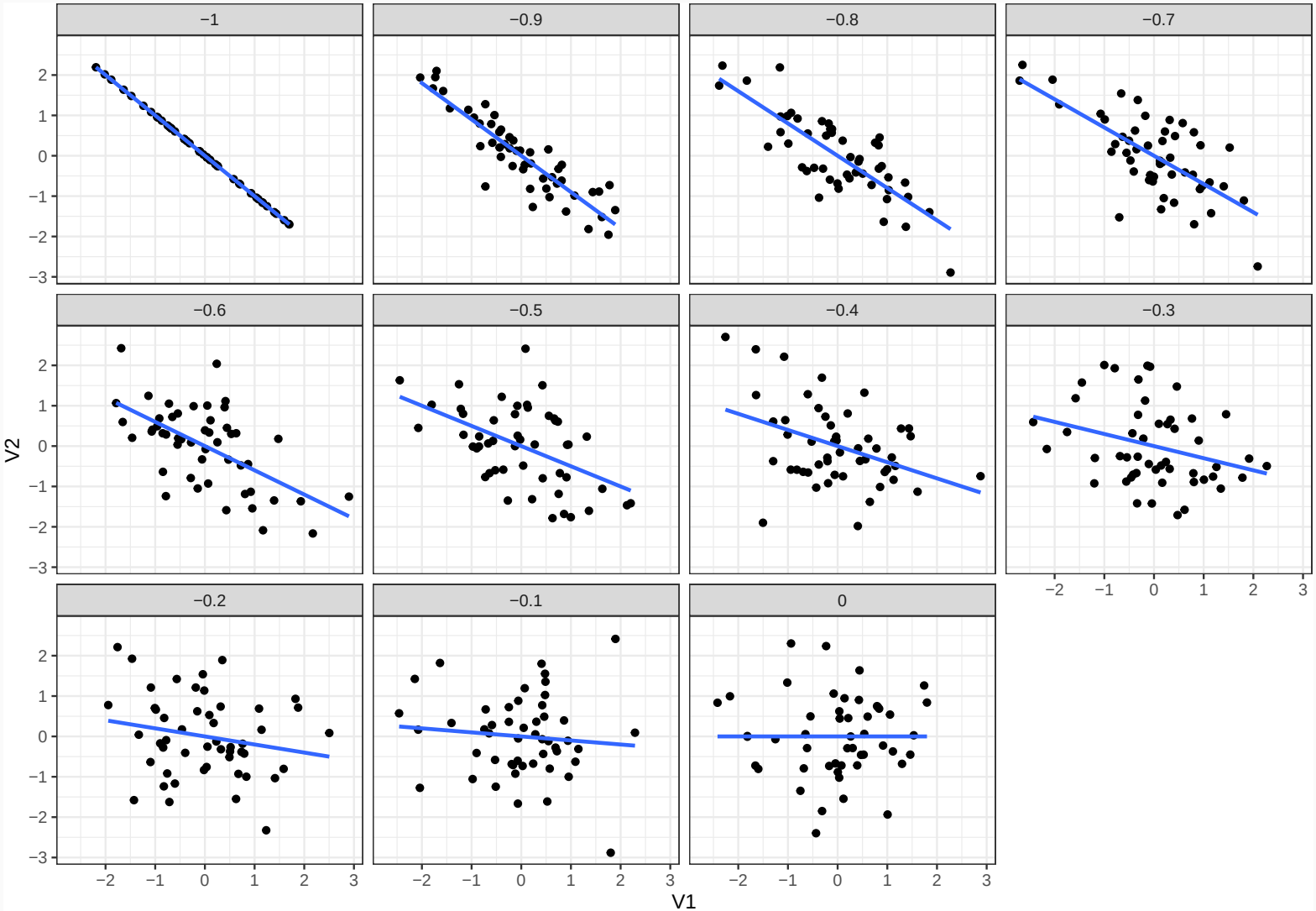
Correlation and scatterplots



Correlation and scatterplots



Correlation and Scatterplots



Return to iris

- Scatterplot Sepal.Width and Petal.Width
- How would you describe the relationship between the variables?
- Estimate the correlation
- Describe what you find using both the estimated correlation and your interpretation of the scatterplot