Correlation

Ling250/450: Data Science for Linguistics
C.M. Downey
Spring 2025

What is correlation?

What is correlation?

- Measures the **strength** and **direction** of a **relationship between variables**

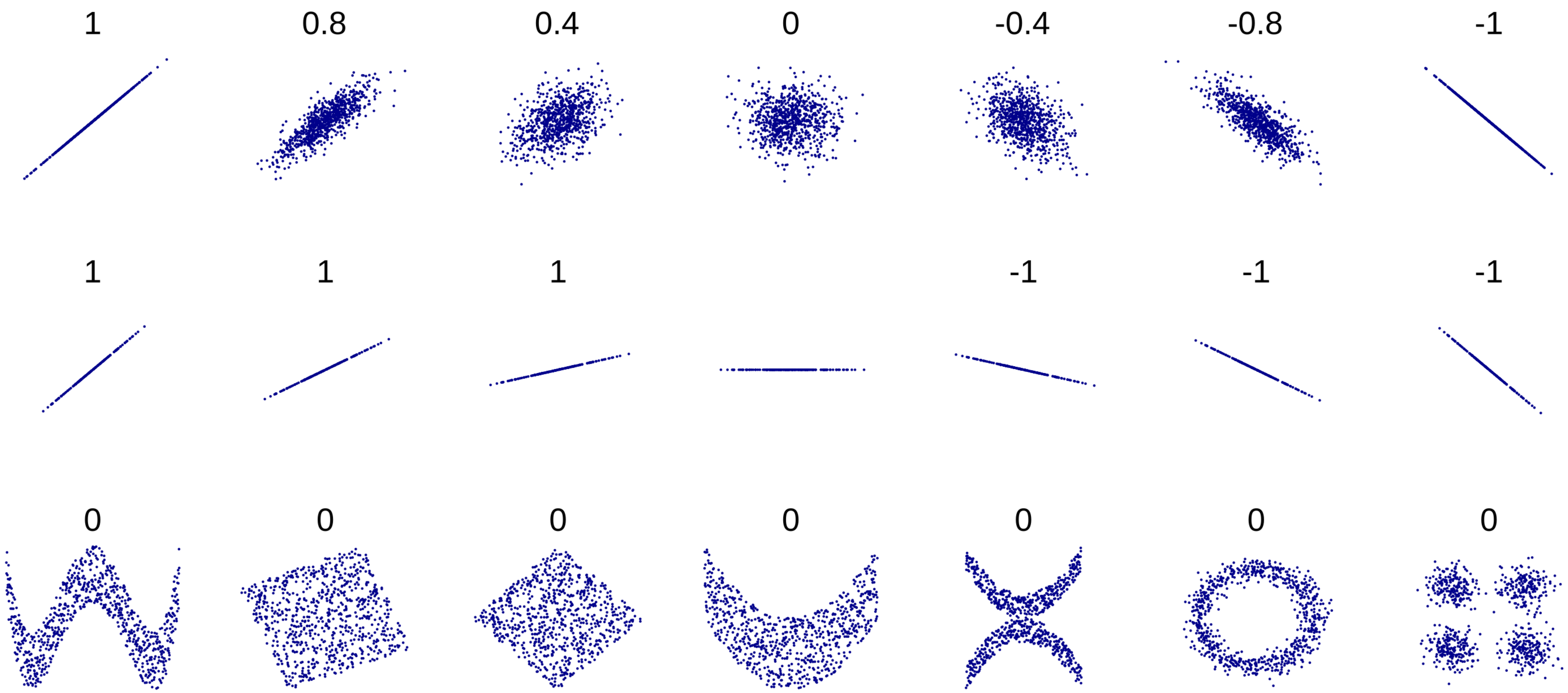
What is correlation?

- Measures the **strength** and **direction** of a **relationship between variables**
- General intuition
  - **Positive correlation:** as X increases, so does Y
  - **Negative correlation:** as X increases, Y decreases

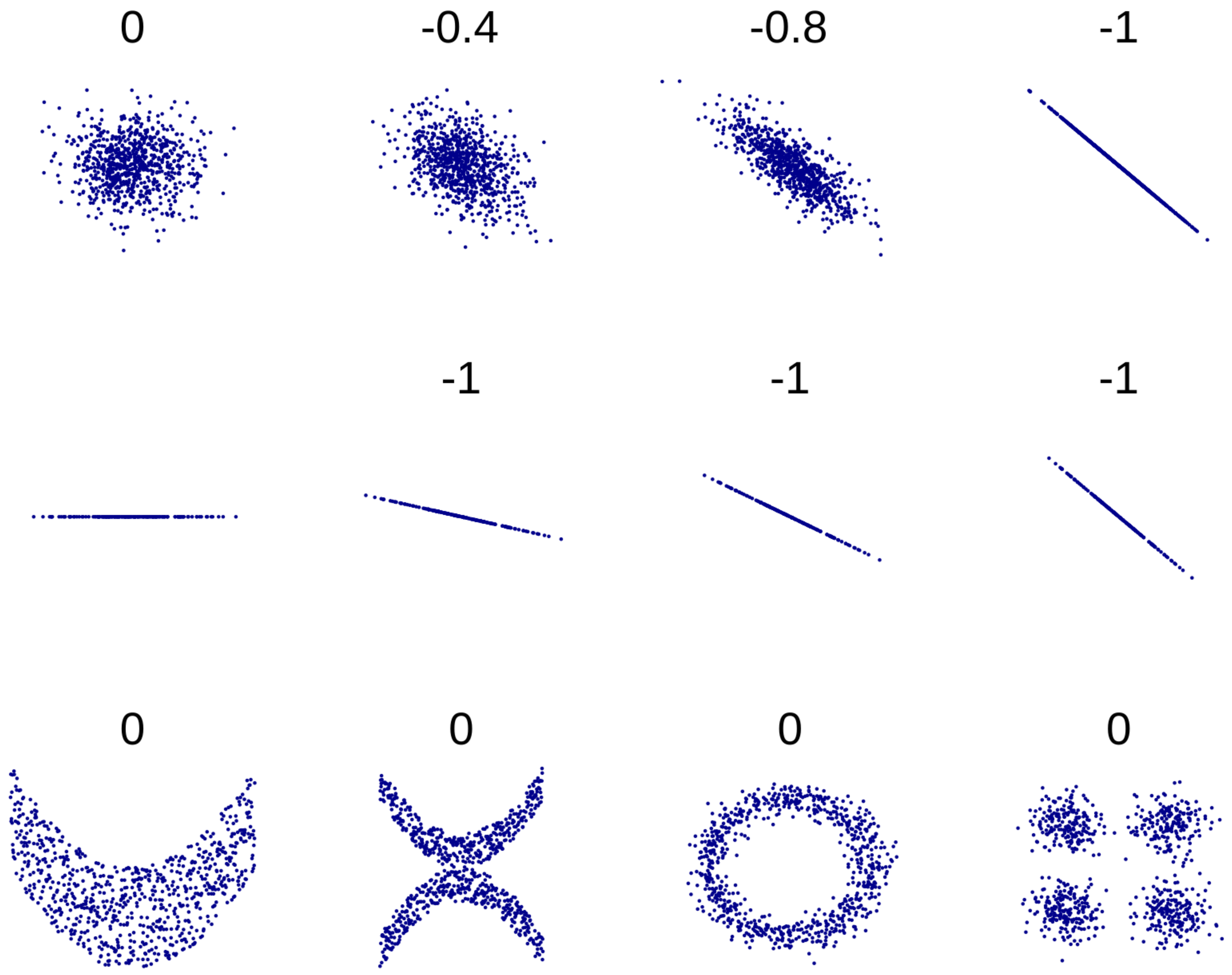
What is correlation?

- Measures the **strength** and **direction** of a **relationship between variables**
- General intuition
  - **Positive correlation:** as X increases, so does Y
  - **Negative correlation:** as X increases, Y decreases
- Correlation is expressed as a **coefficient** (a number)
  - Most commonly used is **Pearson's Correlation Coefficient** (written r)
  - Based on another measure called **covariance**

Visualizing correlation

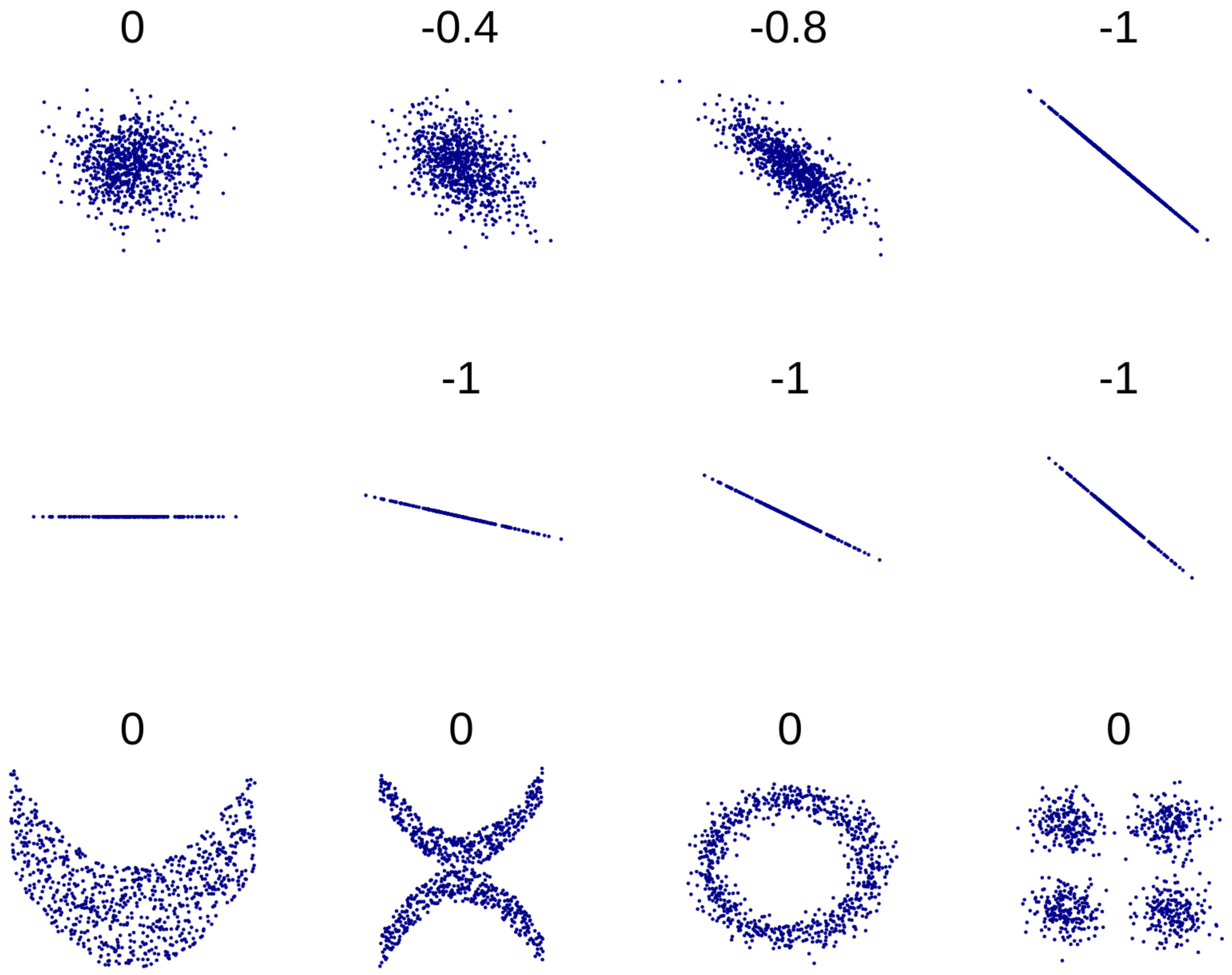


Notes on correlation



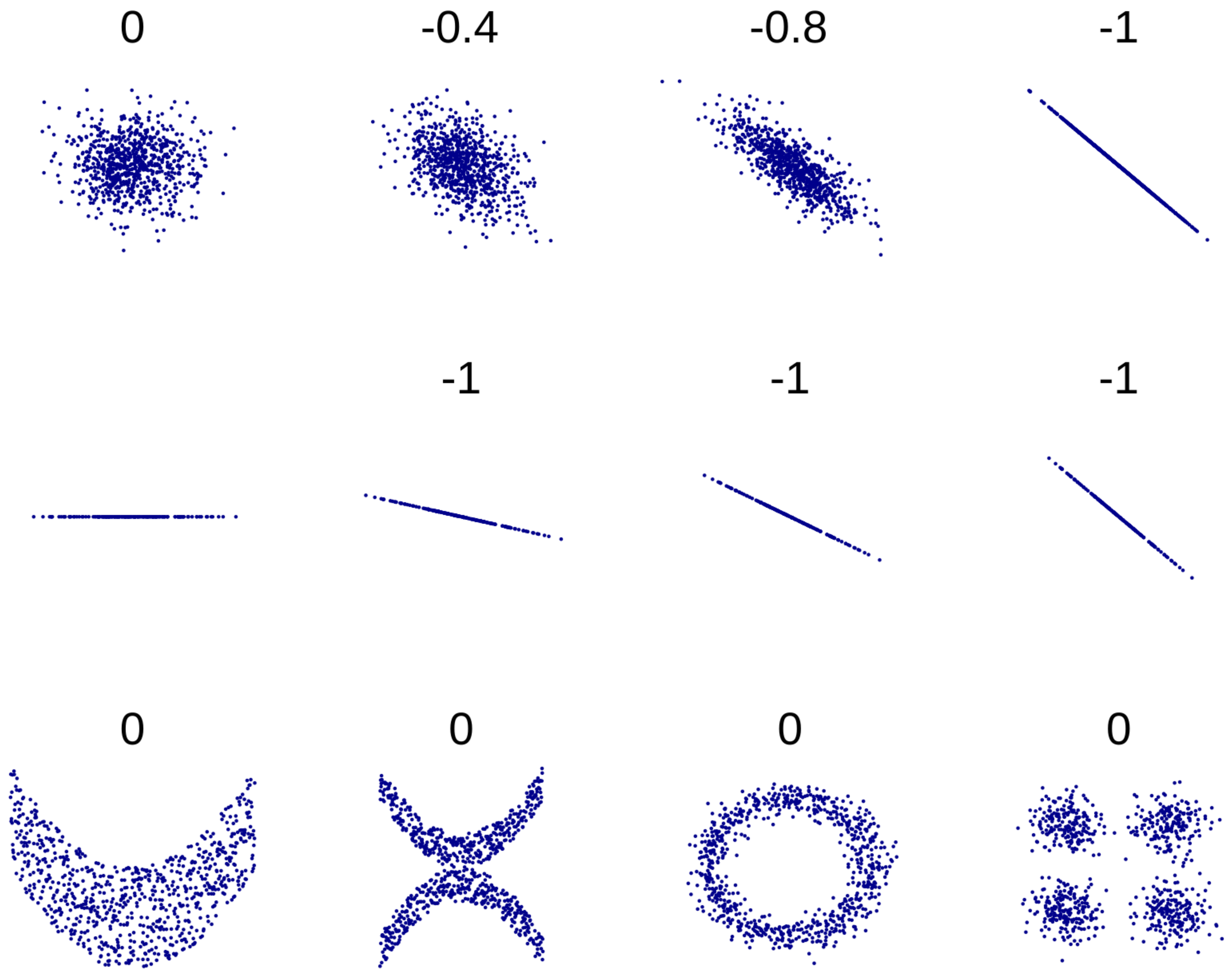
Notes on correlation

- The coefficient's range is **-1 to +1**



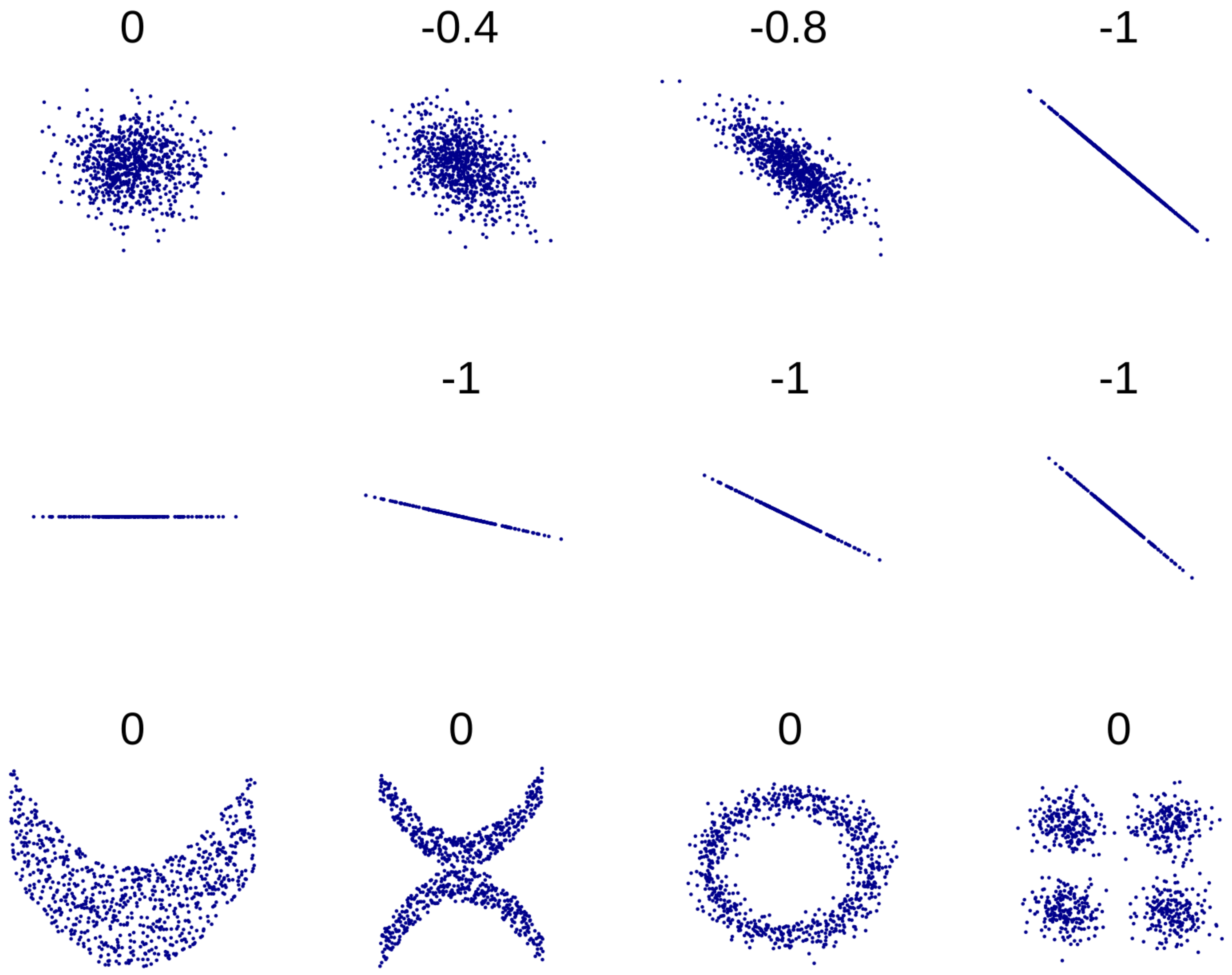
Notes on correlation

- The coefficient's range is **-1 to +1**
- 0 indicates **no correlation**



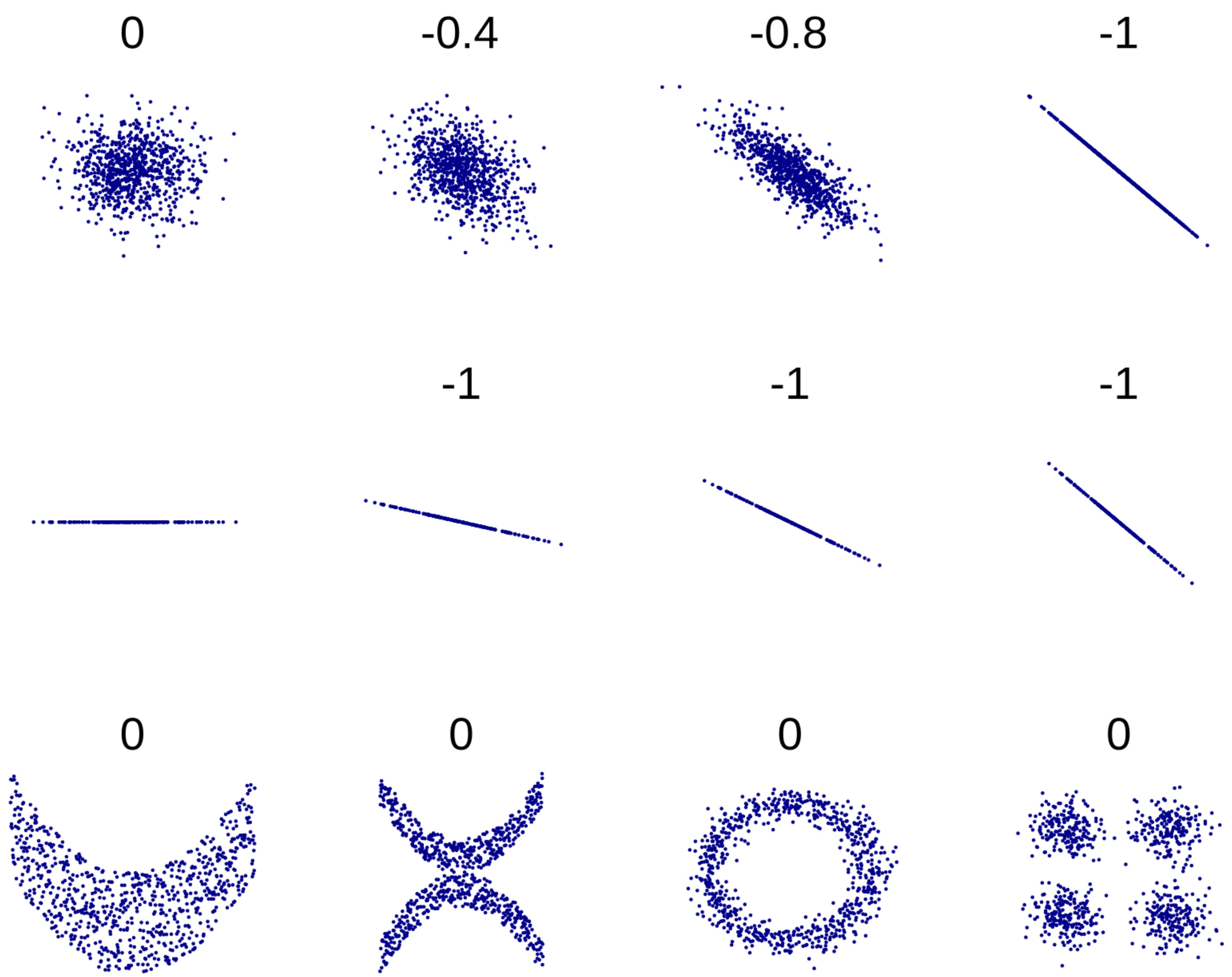
Notes on correlation

- The coefficient's range is **-1 to +1**
- 0 indicates **no correlation**
- The absolute value of the coefficient tells the **"exactness"**, but **not the "slope"** of the relationship



Notes on correlation

- The coefficient's range is **-1 to +1**
- 0 indicates **no correlation**
- The absolute value of the coefficient tells the **"exactness"**, but **not the "slope"** of the relationship
- Data can have **structure without correlation**



Covariance

Covariance

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance

- Conceptually similar to the **variance**, but for two variables

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance

- Conceptually similar to the **variance**, but for two variables
  - Like variance, the values are (approximately) **squared**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance

- Conceptually similar to the **variance**, but for two variables
  - Like variance, the values are (approximately) **squared**
  - Also like variance, covariance is **not intuitive** to think about

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance

- Conceptually similar to the **variance**, but for two variables
  - Like variance, the values are (approximately) **squared**
  - Also like variance, covariance is **not intuitive** to think about
  - Essentially: **average product of deviation**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance

- Conceptually similar to the **variance**, but for two variables
  - Like variance, the values are (approximately) **squared**
  - Also like variance, covariance is **not intuitive** to think about
  - Essentially: **average product of deviation**
- The covariance has **no minimum/ maximum value**

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

Pearson's Correlation Coefficient

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Pearson's Correlation Coefficient

- The correlation coefficient (r) is the covariance **normalized by the standard deviations**
  - "How much is the covariance **compared** to the standard deviations?"

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Pearson's Correlation Coefficient

- The correlation coefficient (r) is the covariance **normalized by the standard deviations**
  - "How much is the covariance **compared** to the standard deviations?"
- Normalization puts the value in the **-1 to +1 range**

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Pearson's Correlation Coefficient

- The correlation coefficient (r) is the covariance **normalized by the standard deviations**
  - "How much is the covariance **compared** to the standard deviations?"
- Normalization puts the value in the **-1 to +1 range**
- R command: `cor(x, y)`

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Interpreting correlation

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

Interpreting correlation

- Interpreting a coefficient **depends heavily on context**
  - In some fields/situations, $r < 0.95$ is considered "weak"
  - In others, $r > 0.3$ is considered "strong"

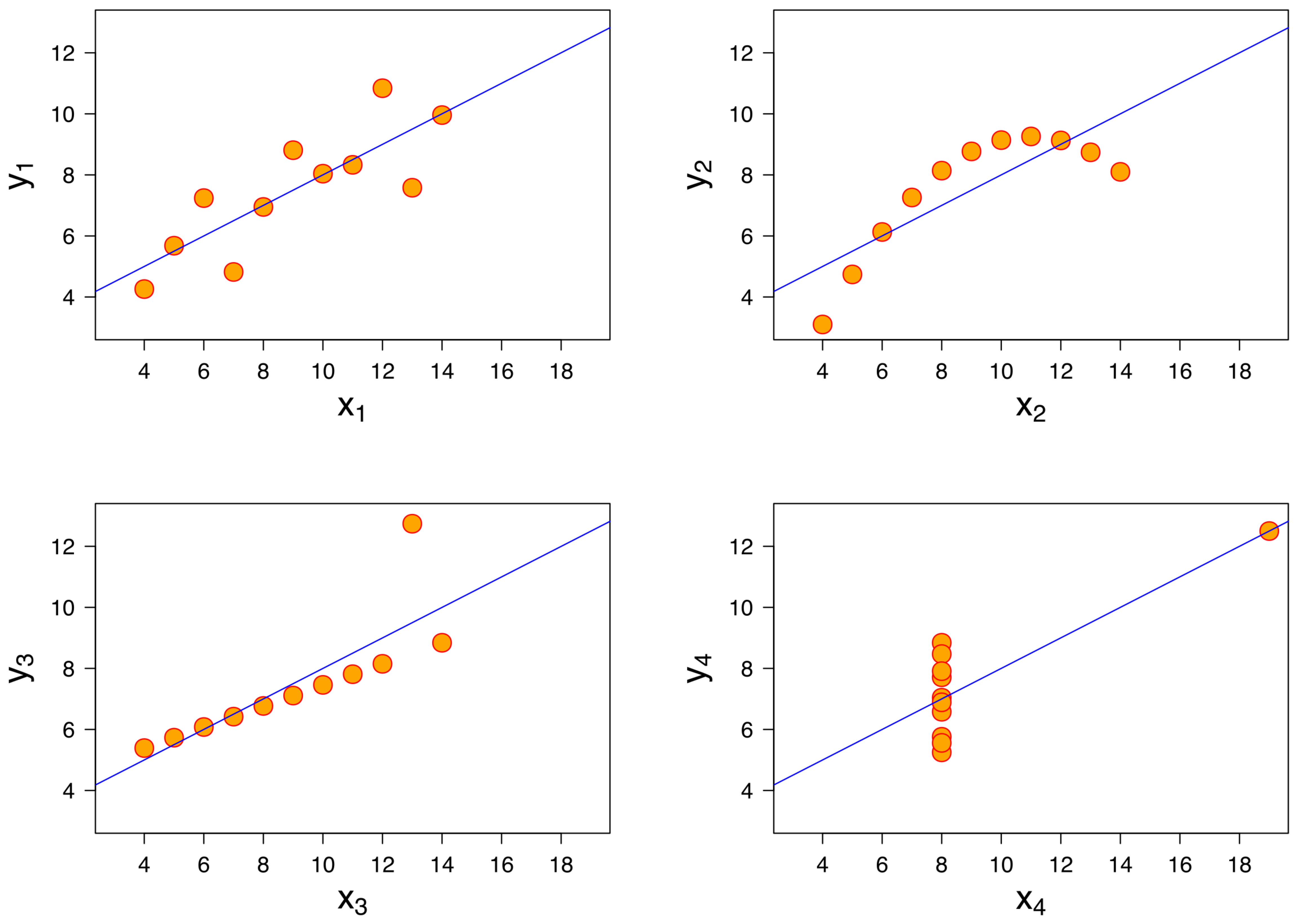
Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

Interpreting correlation

- Interpreting a coefficient **depends heavily on context**
  - In some fields/situations, $r < 0.95$ is considered "weak"
  - In others, $r > 0.3$ is considered "strong"
- Strong correlation does **not** mean "statistical significance". It's **just a measurement**

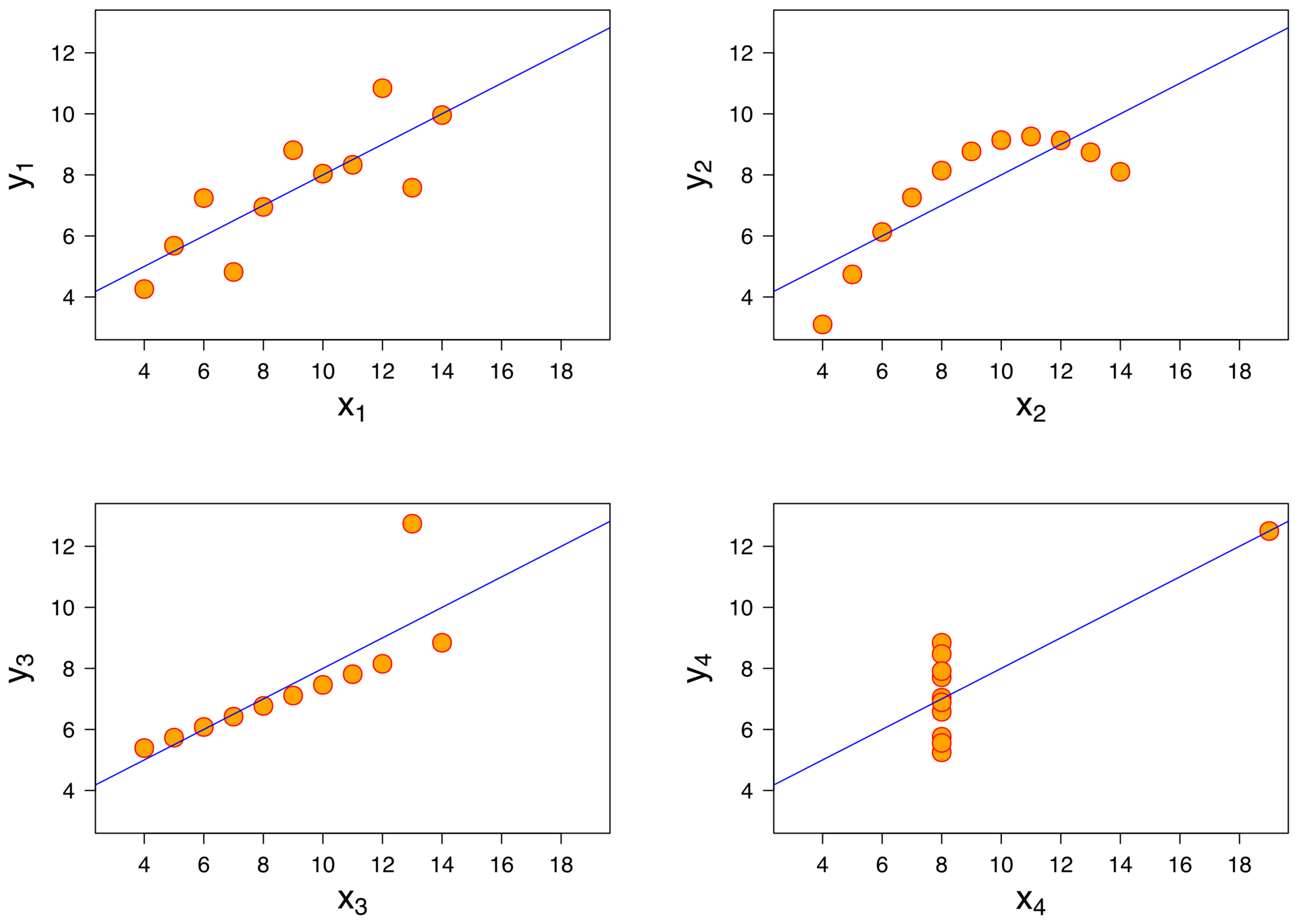
Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

Interpreting correlation



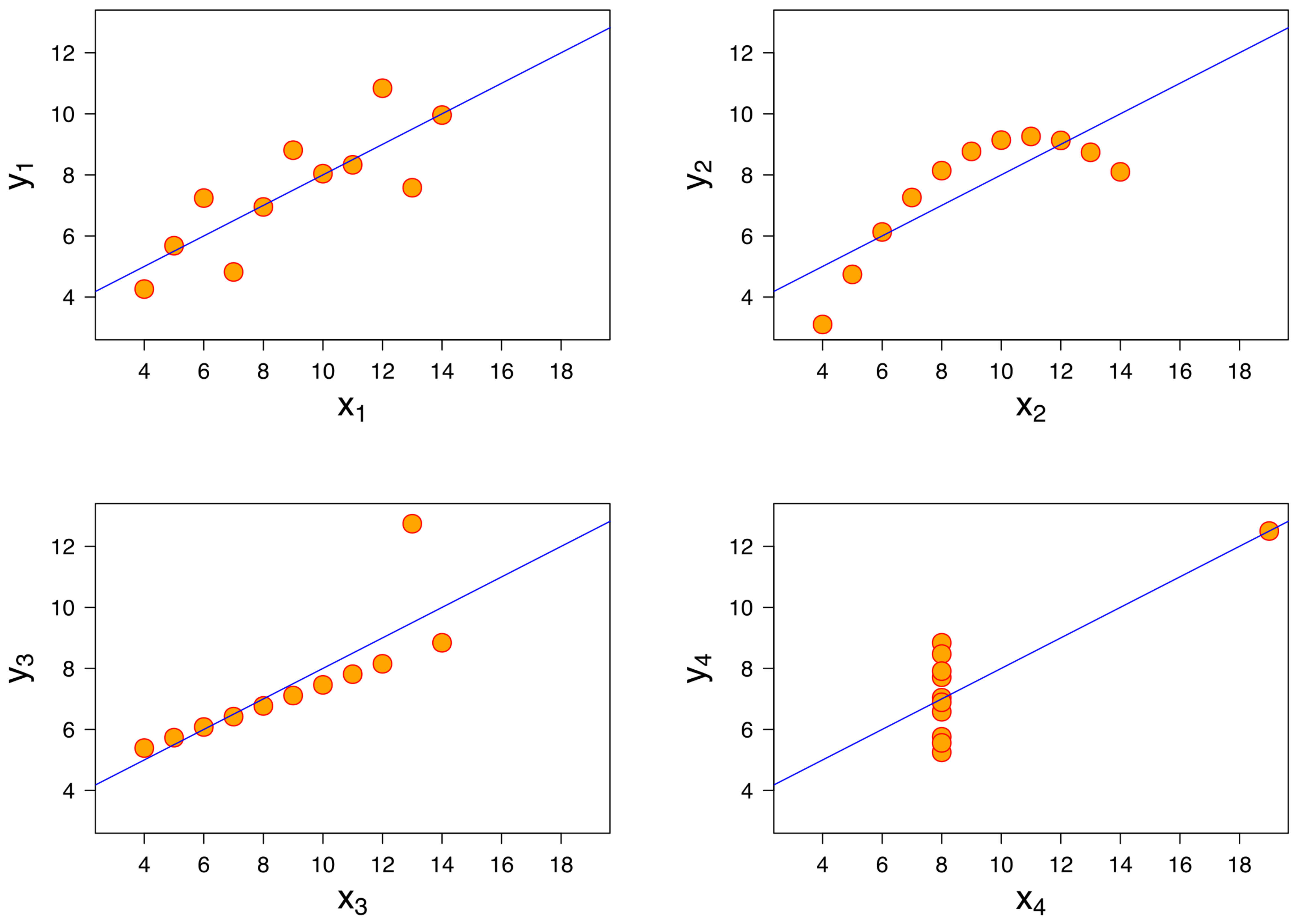
Interpreting correlation

- "Anscombe's quartet" are four datasets with the **exact same correlation**
  - Shows that correlation **does not** give all important aspects of the data

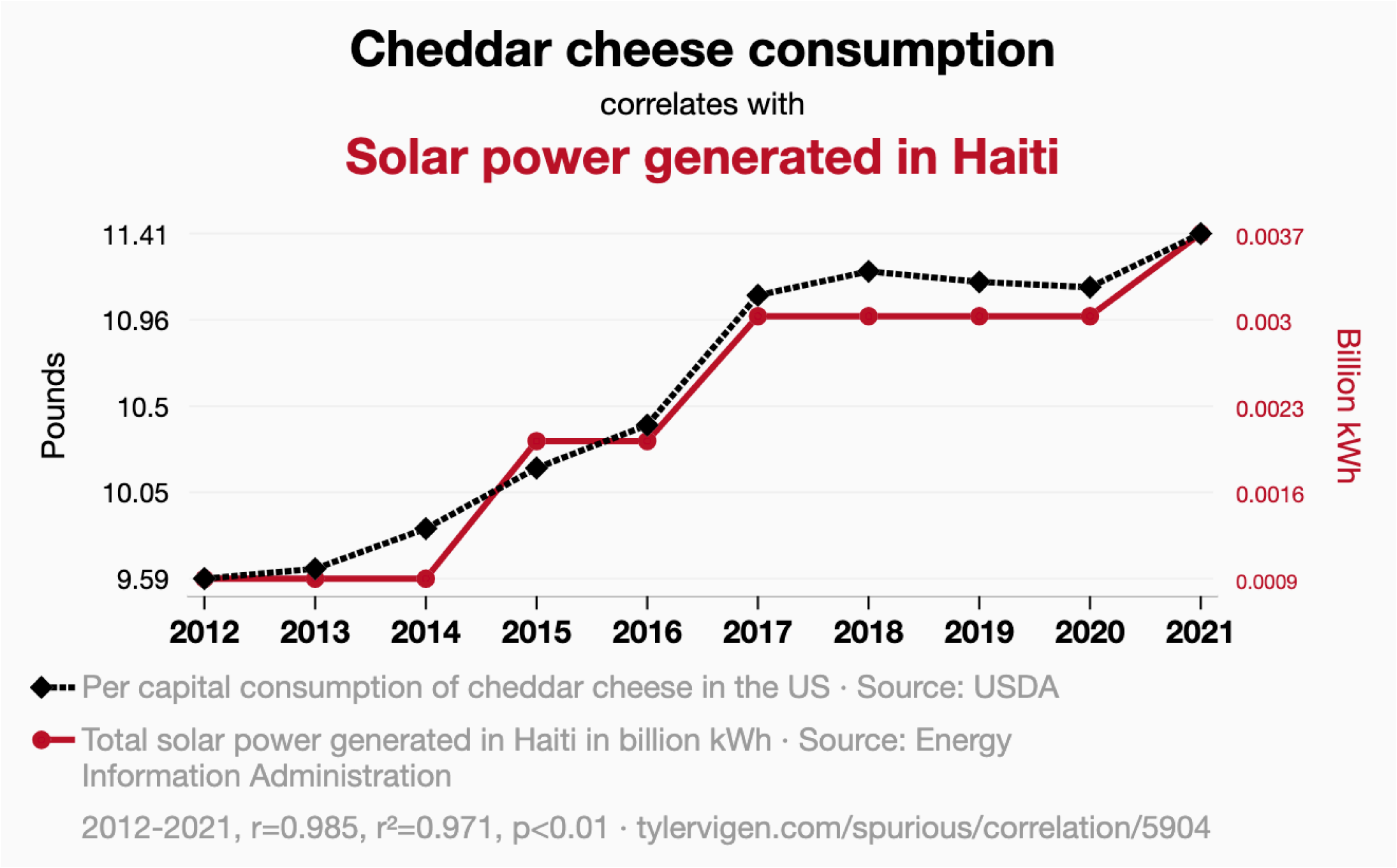


Interpreting correlation

- "Anscombe's quartet" are four datasets with the **exact same correlation**
  - Shows that correlation **does not** give all important aspects of the data
- Make sure to **always use visualizations** in combination with descriptive statistics

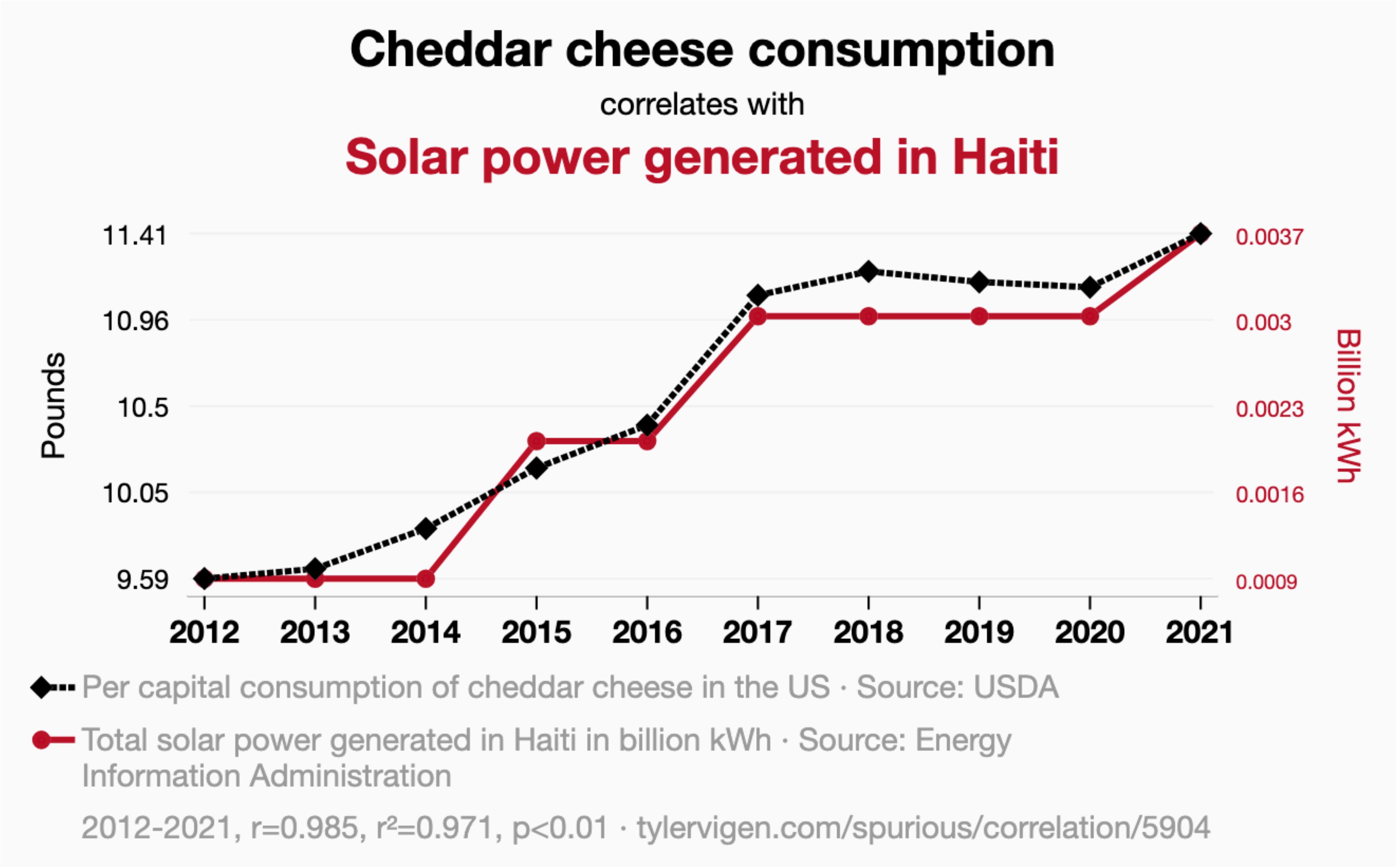


Interpreting correlation



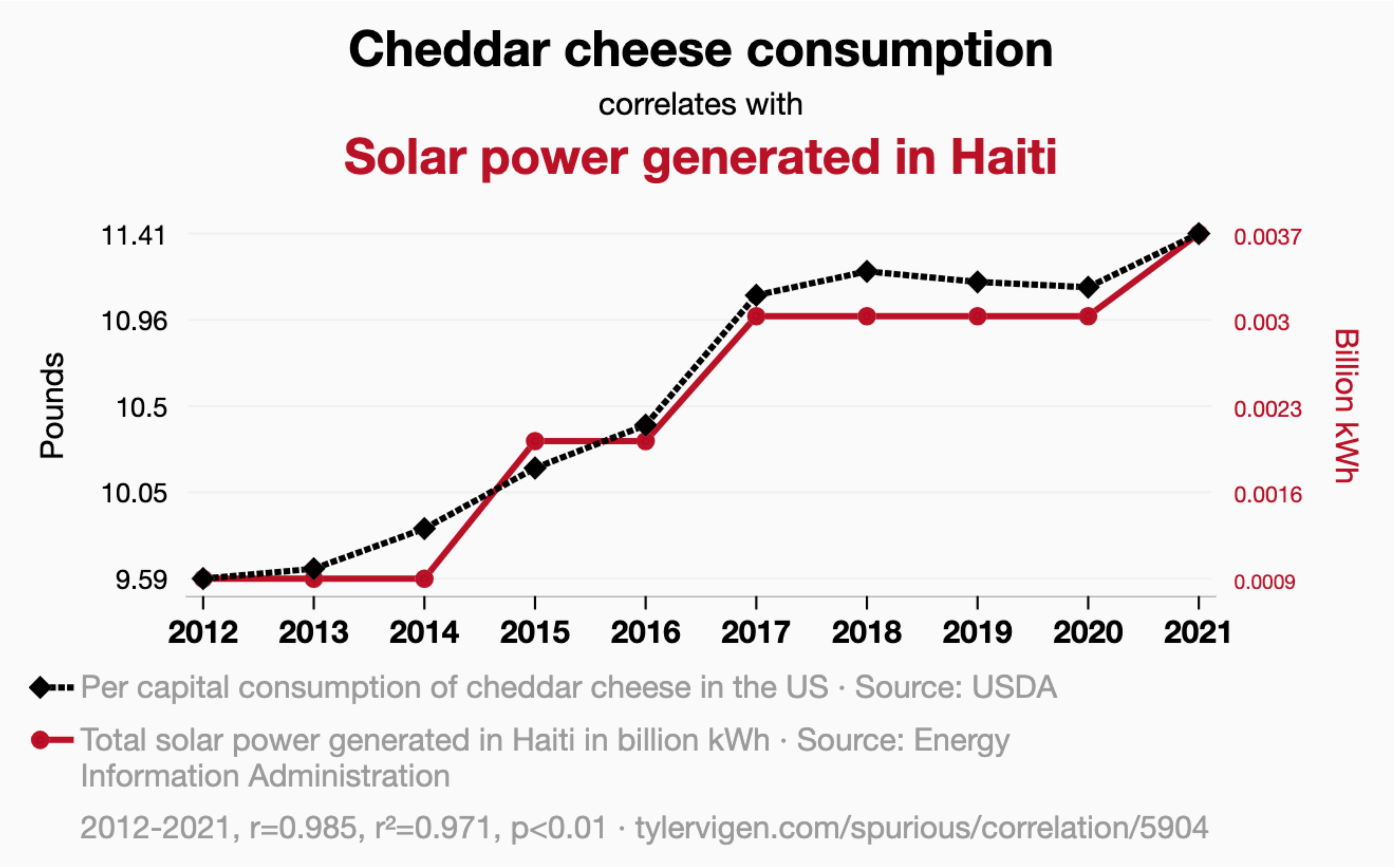
Interpreting correlation

- Important: **correlation does not imply causation!!!**
  - X might cause Y; Y might cause X
  - or **neither**!
  - Correlation **doesn't tell us**



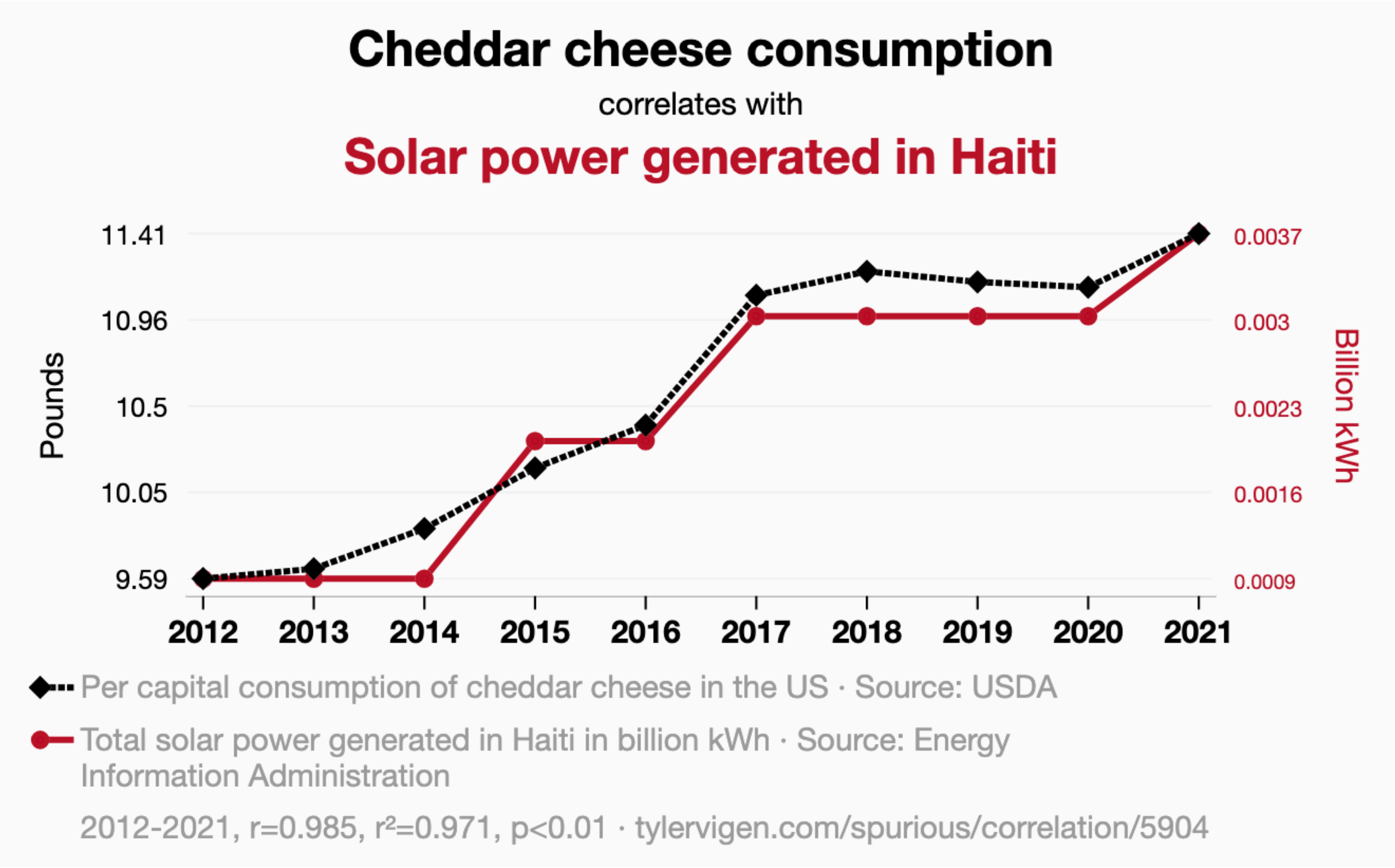
Interpreting correlation

- Important: **correlation does not imply causation!!!**
  - X might cause Y; Y might cause X
  - or **neither**!
  - Correlation **doesn't tell us**
- Lots of examples of funny **spurious correlations** (variables that inexplicably correlate)



Interpreting correlation

- Important: **correlation does not imply causation!!!**
  - X might cause Y; Y might cause X
  - or **neither**!
  - Correlation **doesn't tell us**
- Lots of examples of funny **spurious correlations** (variables that inexplicably correlate)
- Browse some at [this website]



Correlation matrix

- If you call `cor()` on a data frame, it gives the **correlation matrix**
  - i.e. the correlation between **all pairs of variables**
  - Note that every variables is **perfectly correlated with itself**

```
> just_numerical_columns = data.frame(vowels$F1, v
owels$F2, vowels$HEIGHT)
> head(just_numerical_columns)
  vowels.F1 vowels.F2 vowels.HEIGHT
1   848.070   1450.96           173
2   648.318   1126.22           173
3   259.000   1834.00           173
4   578.985   1715.22           173
5   405.000   1899.00           173
6   656.600   1414.40           173
> cor(just_numerical_columns)
               vowels.F1  vowels.F2 vowels.HEIGHT
vowels.F1      1.0000000 -0.1488059    -0.1996378
vowels.F2     -0.1488059  1.0000000    -0.1426429
vowels.HEIGHT -0.1996378 -0.1426429     1.0000000
```