DATA SCIENCE

Statistics for Data Science

Correlation



UNC GREENSBORO

WHAT IS CORRELATION

Correlation: Understanding Relationships Between Variables

- What does it mean for two variables to be related?
- How can we visualize a relationship?
- How can we measure the strength and direction of a relationship?
- Does correlation imply causation?

Example: study hours and exam scores



Study hours

Explanatory variable X

Exam score

Response variable Y

Research questions give X and Y different roles

➢Explanatory Variable and Response Variable

- Correlation describes **linear** relationships between **quantitative** variables.
- Correlations tell us to the degree that two variables are associated with each other. It is a measure of association.
- Correlations are bivariate or multivariate in nature, i.e., relational between variables.
- Correlations measure association by measuring how when one variable changes another changes with it, which means it is measuring dynamic relationships (change).

Country	Per Capita GDP (*X*)	Life Expectancy (*Y*)
Austria	21.4	77.48
Belgium	23.2	77.53
Finland	20.0	77.32
France	22.7	78.63
Germany	20.8	77.17
Ireland	18.6	76.39
	21.5	78.51
	22.0	78.15
Switzerland	23.8	78.99
United Kingdom	21.2	77.37

Explanatory Variable (X): A variable used to explain, predict, or investigate differences in another variable.

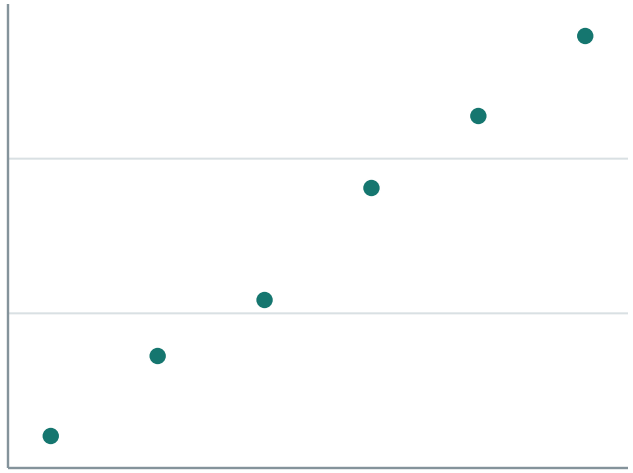
Response Variable (Y): A variable whose variation we want to understand or predict.

WHAT IS CORRELATION

Correlation describes direction and strength

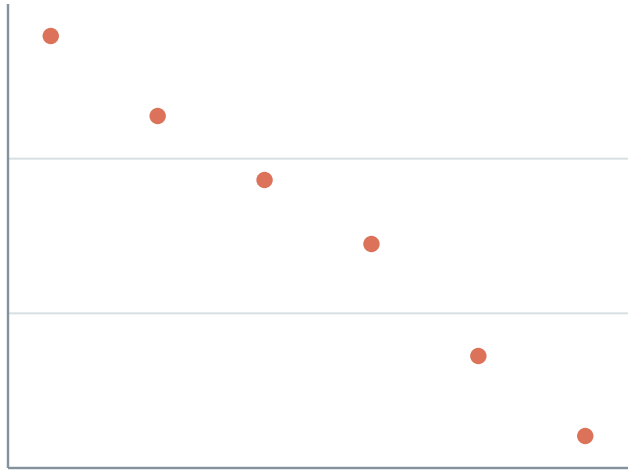
A correlation summarizes how two variables vary together.

Positive association



Positive: larger X tends to pair with larger Y.

Negative association



Negative: larger X tends to pair with smaller Y.

No clear linear pattern



No clear trend: the points show little linear structure.

Correlation Does Not Imply Causation

Ice cream sales and drowning incidents both rise in summer.

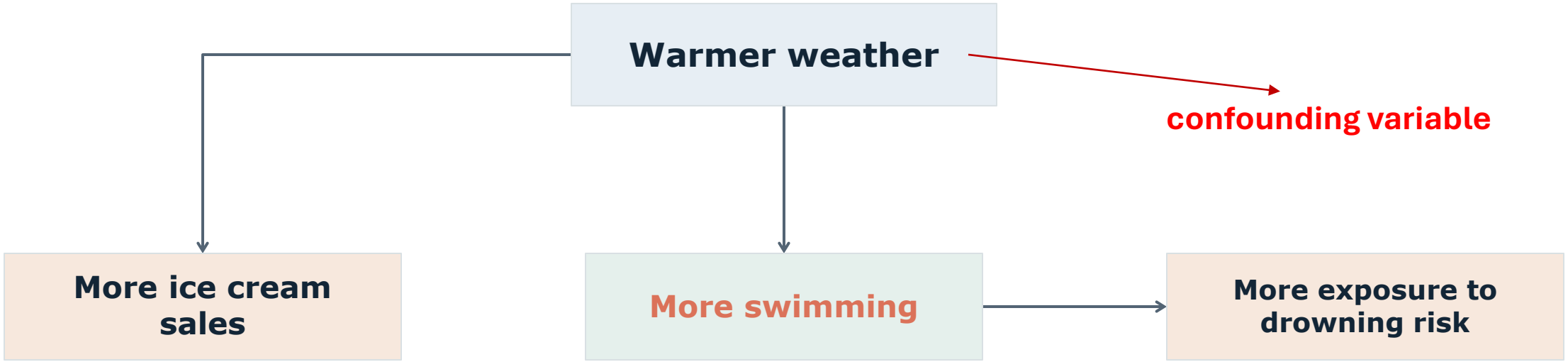




WHAT IS CORRELATION

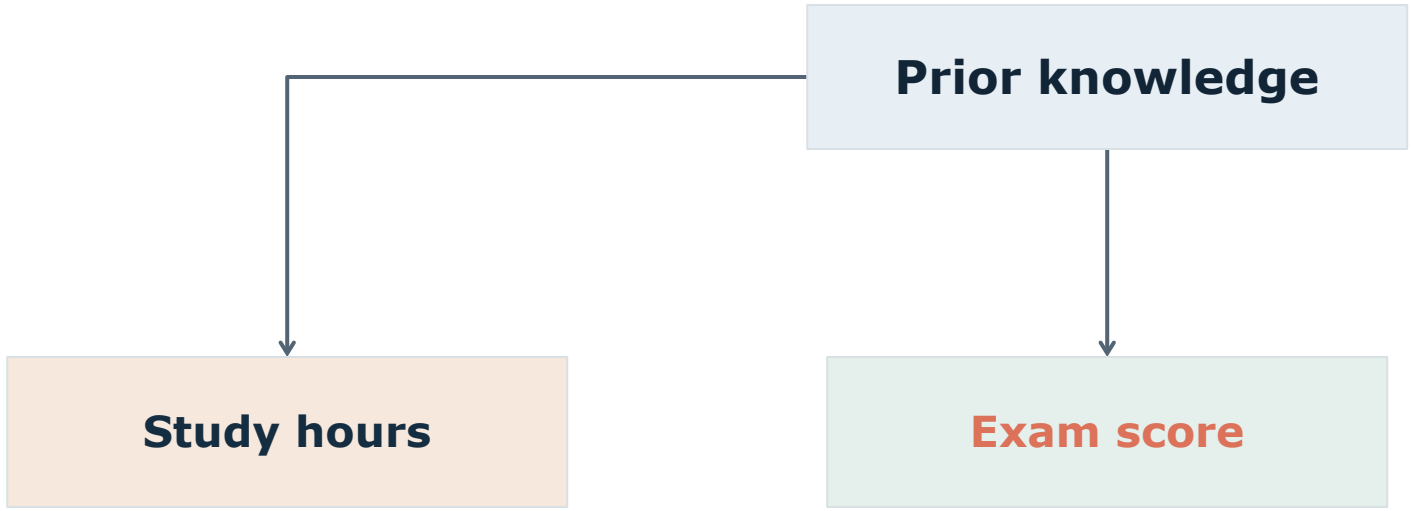
A third variable can explain a correlation

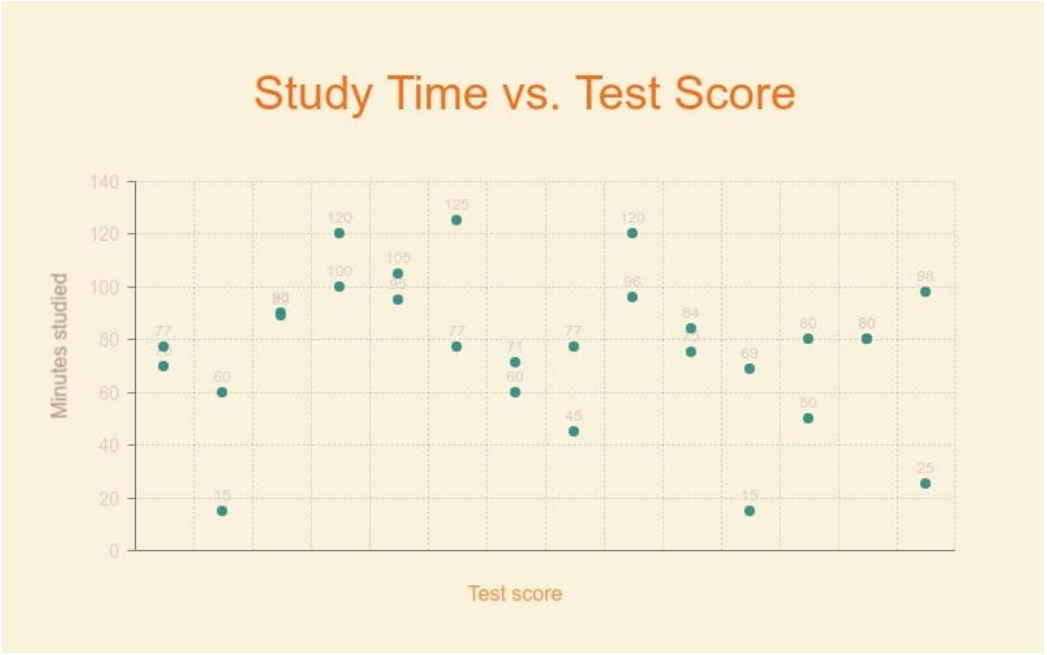
Ice cream sales and drowning incidents both rise in summer.



Correlation shows association. Causal claims need stronger evidence and a suitable study design.

A third variable can explain a correlation

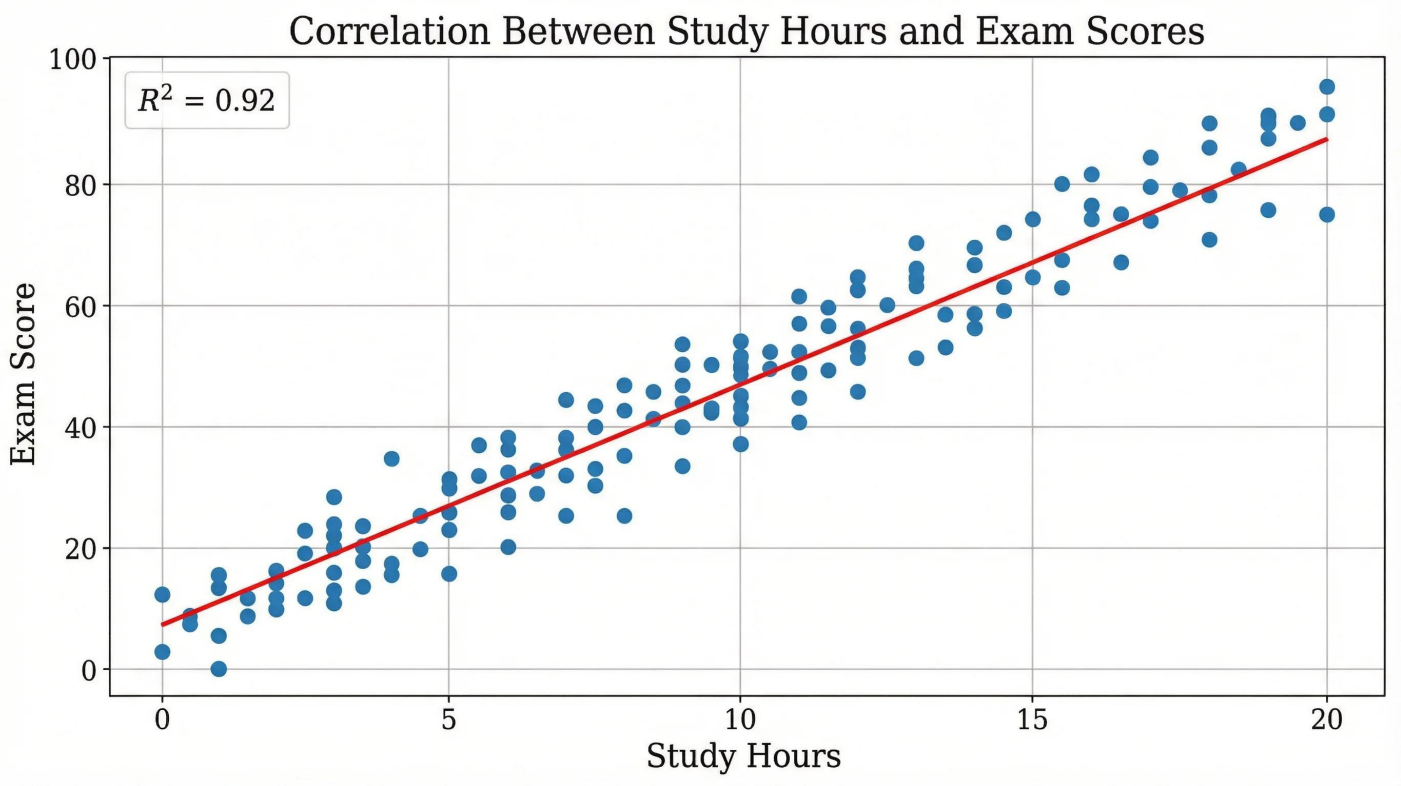




SCATTERPLOTS

A scatterplot turns each data row into a point

Example: study hours and exam scores



Scatterplot: A scatterplot visualizes the relationship between two numerical variables.

SCATTERPLOTS

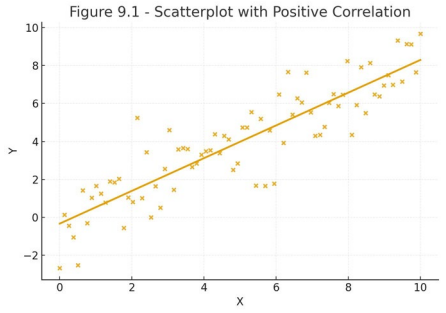
Read a scatterplot in four steps

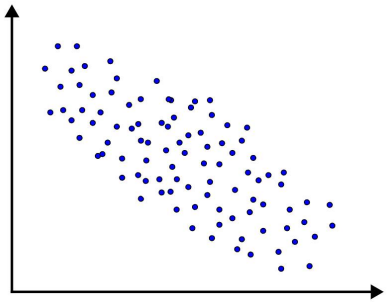
01 **Direction** — Positive, negative, or no clear association?

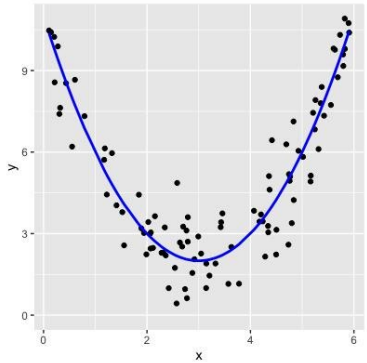
02 **Strength** — How closely do points follow a recognizable pattern?

03 **Form** — Is the pattern linear, curved, or another shape?

04 **Unusual observations** — Look for outliers, clusters, gaps, and restricted ranges.







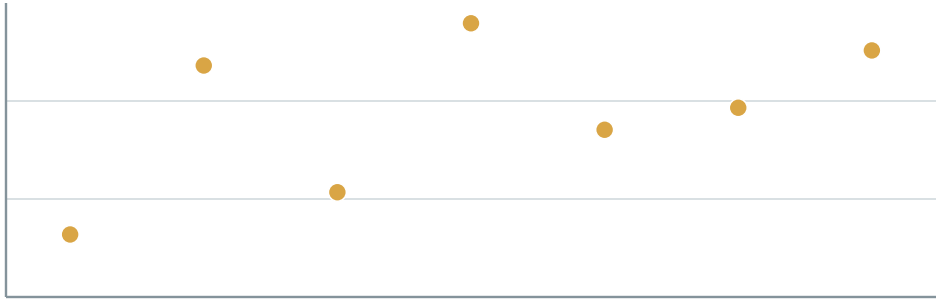
SCATTERPLOTS

Strength describes scatter around a pattern

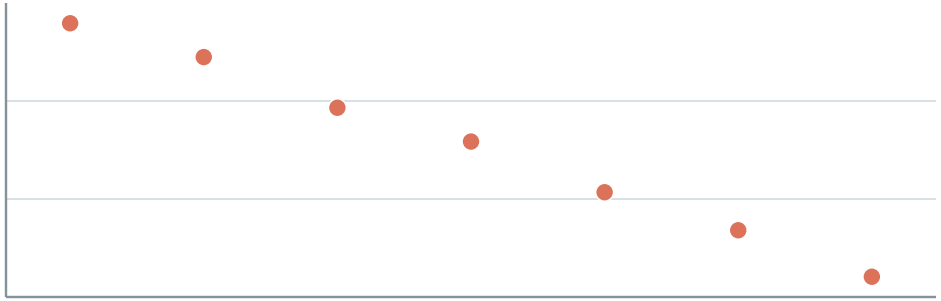
Strong positive



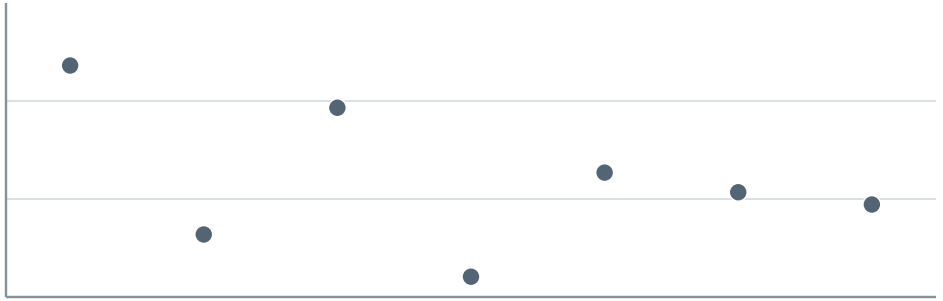
Weak positive



Strong negative

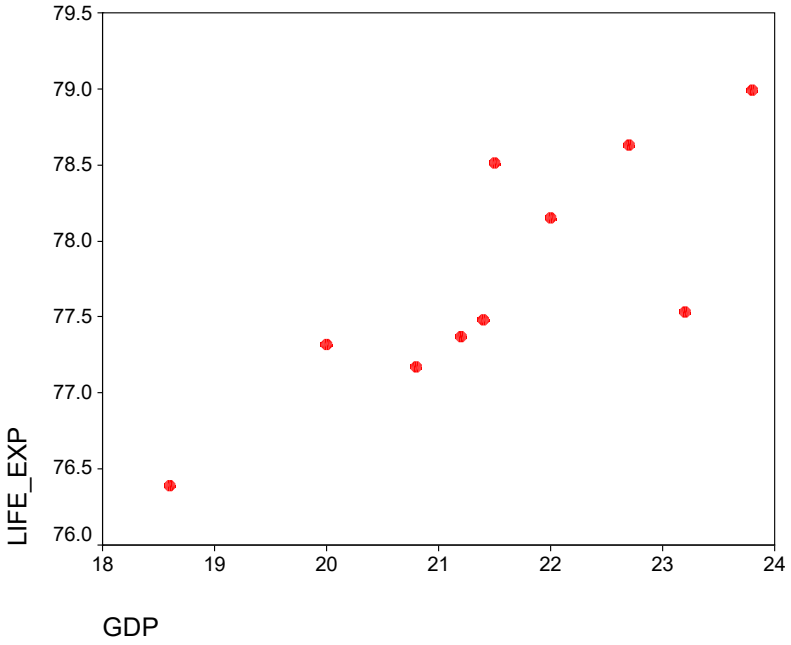


Weak negative



A steep relationship is not necessarily more strongly correlated than a shallow relationship.

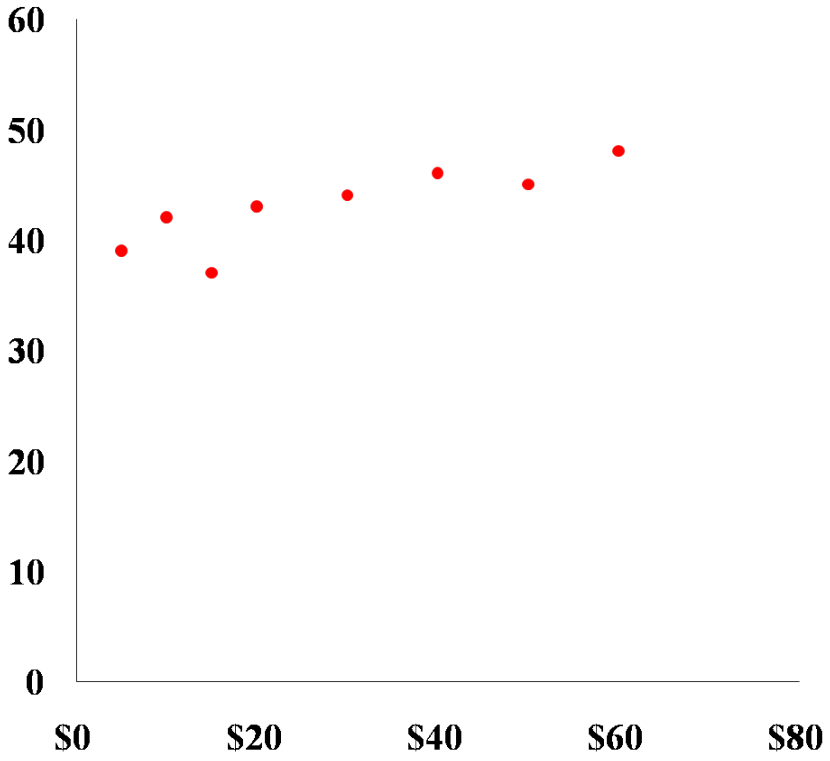
Interpreting Scatterplots



Interpretation:

- Form: linear (straight)
- Outliers: none
- Direction: positive
- Strength: difficult to judge by eye

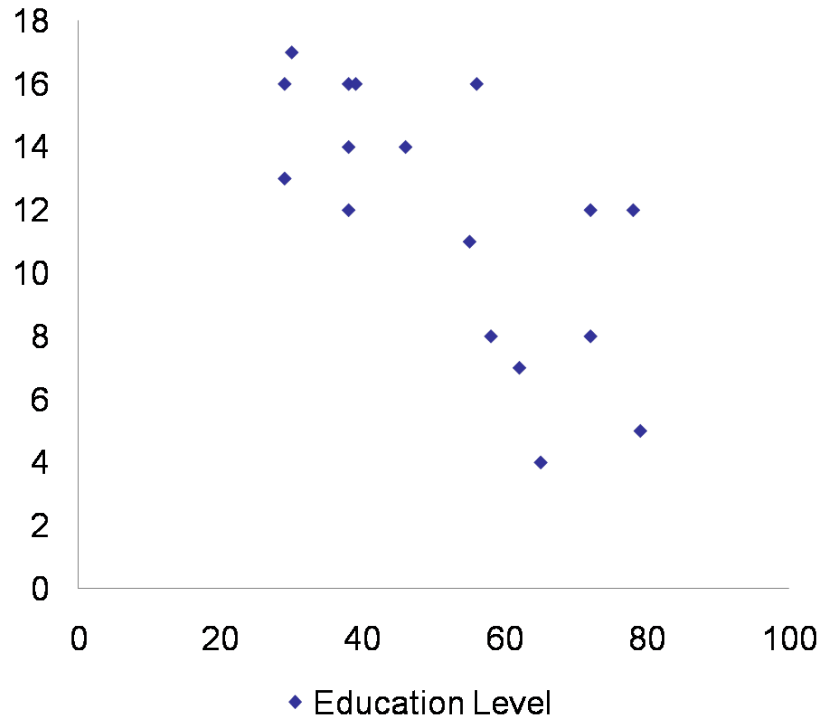
Interpreting Scatterplots



Interpretation:

- Form: linear
- Outliers: none
- Direction: positive
- Strength: difficult to judge by eye (looks strong)

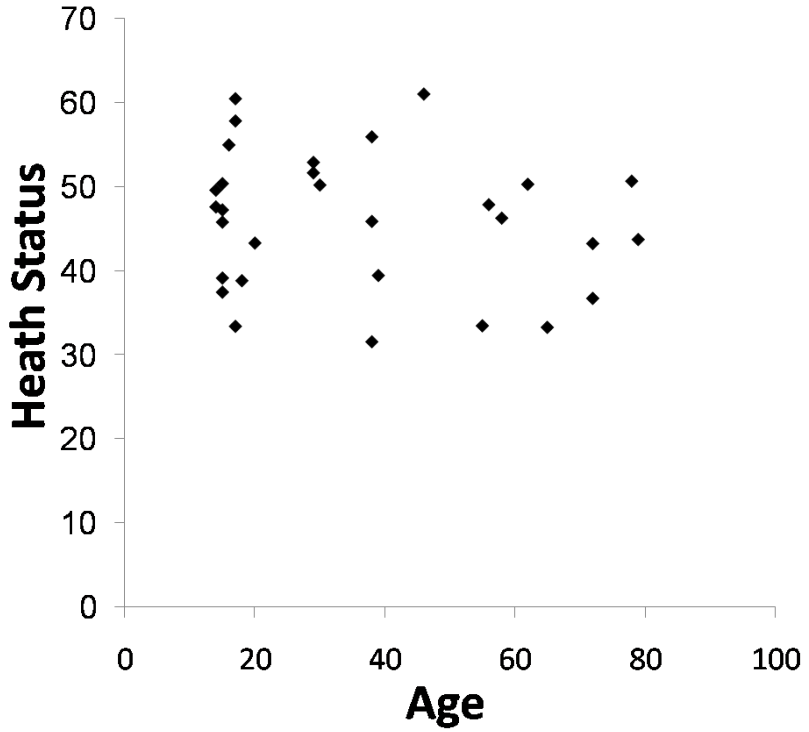
Interpreting Scatterplots



Interpretation:

- Form: linear
- Outliers: none
- Direction: negative
- Strength: difficult to judge by eye (looks moderate)

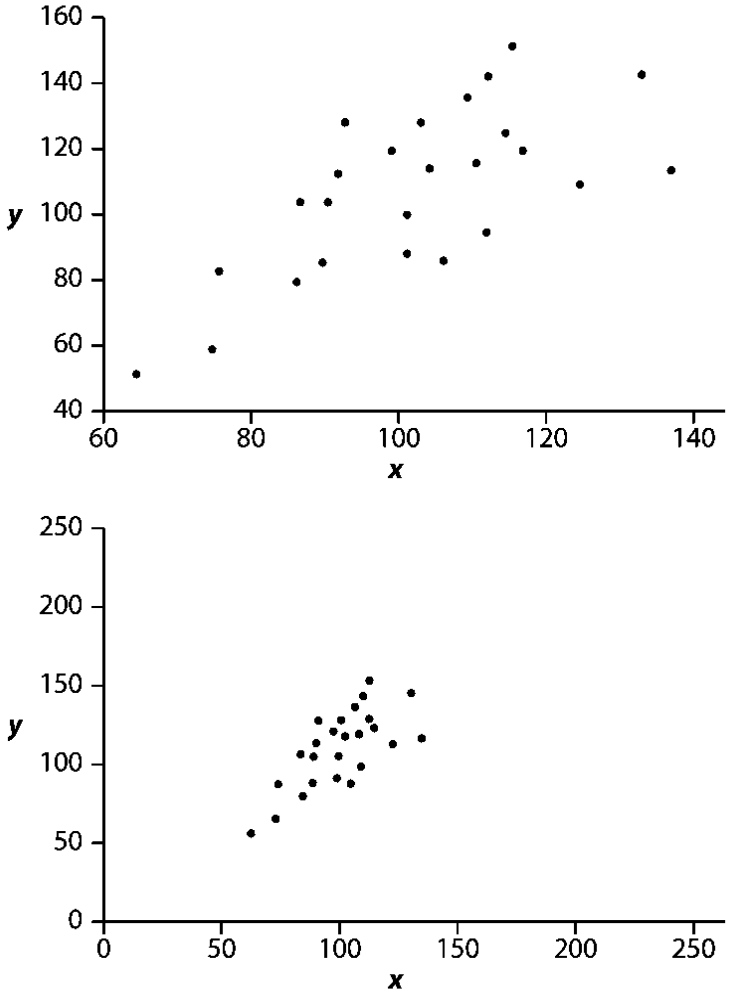
Interpreting Scatterplots



Interpretation:

- Form: linear(?)
- Outliers: none
- Direction: negative
- Strength: difficult to judge by eye (looks weak)

When Scatterplots Can Be Misleading



- It is difficult to judge correlational strength by eye alone;
- Here are identical data plotted on different axes;
- First relationship seems weaker than second;
- This is an artifact of the axis scaling;

We use a statistical called the correlation coefficient to judge strength objectively

Pearson r Correlation

Pearson r Correlation is the most commonly used statistic to measure the degree of the relationship between linearly related variables.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad = \quad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y

n = number of observations

x_i = value of x (for ith observation)

y_i = value of y (for ith observation)

Pearson r Correlation

r is always between −1 and +1

- r = 1: perfect positive
- r = -1: perfect negative
- r = 0: no linear relationship

0.0 < |r| < 0.3	weak correlation
0.3 < |r| < 0.7	moderate correlation
0.7 < |r| < 1.0	strong correlation
|r| = 1.0	perfect correlation

The closer r is to +1 or –1, the stronger the correlation.

Husband's age / Wife's age

r = .94 (strong positive correlation)

Husband's height / Wife's height

r = .36 (weak positive correlation)

Distance of golf putt / percent success

r = -.94 (strong negative correlation)

CORRELATION COEFFICIENTS

Pearson r Correlation



Correlation $r = 0$



Correlation $r = -0.3$



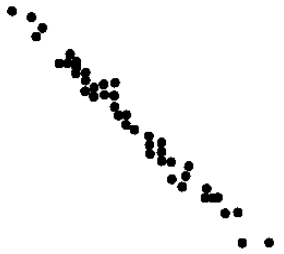
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

CORRELATION COEFFICIENTS

Pearson r Correlation

X=[1,2,3], Y=[2,4,6]

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson r Correlation

X=[1,2,3], Y=[2,4,6]

$\bar{X} = 2, \bar{Y} = 4$

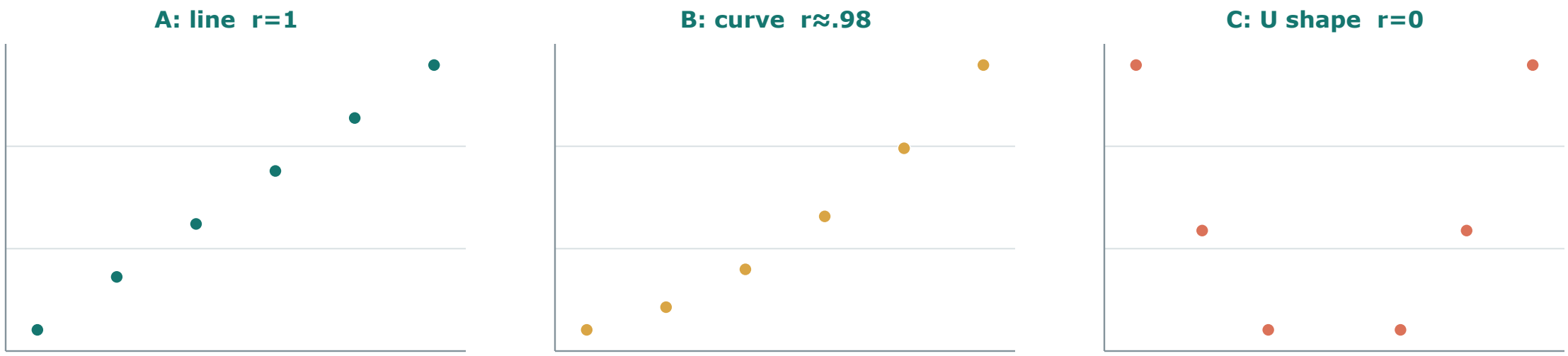
(X)	(Y)	$(\mathrm{X} - \bar{X})$	$(\mathrm{Y} - \bar{Y})$	Product
1	2	-1	-2	2
2	4	0	0	0
3	6	1	2	2

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r = \frac{4}{\sqrt{2} \times \sqrt{8}} = 1$$

EXERCISE DEBRIEF

The same coefficient can hide different shapes



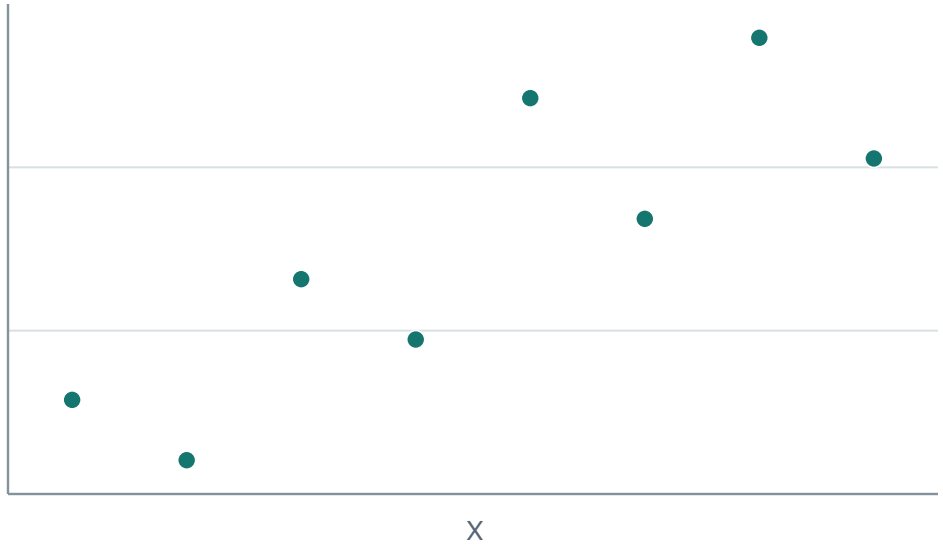
Dataset C has no linear association, yet X and Y are clearly related.

A scatterplot shows structure that a single number cannot.

OPTIONAL CHALLENGE

Bonus: Can two datasets share r but differ in shape?

Dataset A: roughly linear



Dataset B: two clusters



Both have Pearson r ≈ 0.8. Do they show the same relationship?

Plot both datasets, compute three coefficients, then explain what the numbers leave out.

Spearman ρ Correlation

Spearman ρ Correlation: Apply Pearson to ranks instead of raw values.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

ρ= Spearman rank correlation

di= the difference between the ranks of corresponding variables

n= number of observations

The assumptions of the Spearman correlation are that data must be at least ordinal and the scores on one variable must be monotonically related to the other variable.

Spearman ρ Correlation

Spearman ρ Correlation: Example

In a study of diagnostic processes, entering clinical graduate students are shown a 20-minute videotape of children's behavior and asked to rank-order 10 behavioral events on the tape in the order of the importance each has for a behavioral assessment. (1 = most important.) The data are then averaged to produce an average rank ordering for the entire class. The same thing was then done using experienced clinicians. The data follow:

Events	1	2	3	4	5	6	7	8	9	10
Experienced Clinicians:	1	3	2	7	5	4	8	6	9	10
New Students	2	4	1	6	5	3	10	8	7	9

Use Spearman correlation to measure the agreement between experienced and novice clinicians.

Spearman ρ Correlation

Spearman ρ Correlation: Example

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Events	1	2	3	4	5	6	7	8	9	10
Experienced Clinicians:	1	3	2	7	5	4	8	6	9	10
New Students	2	4	1	6	5	3	10	8	7	9

d_i: -1 -1 1 1 0 1 -2 -2 2 1

$$\sum d_i^2 = 18$$

$$\rho = 1 - \frac{6 \cdot 18}{10(100-1)} = 1 - \frac{108}{990} = 1 - 0.1091 = 0.891$$

Kendall rank τ correlation

Kendall rank τ correlation: Measure how often pairs agree in order.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Total pairs

Nc= number of concordant
Nd= Number of discordant

Key Terms

Concordant: Ordered in the same way.

Discordant: Ordered differently.

X=[1,2,3], Y=[3,2,1]

Pair	X order	Y order	Result
(1,2)	↑	↓	discordant
(1,3)	↑	↓	discordant
(2,3)	↑	↓	discordant

Spearman ρ Correlation

Kendall rank τ correlation : Example

In a study of diagnostic processes, entering clinical graduate students are shown a 20-minute videotape of children's behavior and asked to rank-order 10 behavioral events on the tape in the order of the importance each has for a behavioral assessment. (1 = most important.) The data are then averaged to produce an average rank ordering for the entire class. The same thing was then done using experienced clinicians. The data follow:

Events	1	2	3	4	5	6	7	8	9	10
Experienced Clinicians:	1	3	2	7	5	4	8	6	9	10
New Students	2	4	1	6	5	3	10	8	7	9

Use Spearman correlation to measure the agreement between experienced and novice clinicians.

Spearman ρ Correlation

Kendall rank τ correlation : Example

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Events	1	2	3	4	5	6	7	8	9	10
Experienced Clinicians:	1	3	2	7	5	4	8	6	9	10
New Students	2	4	1	6	5	3	10	8	7	9

Total Pairs: $C_n^2 = C_{10}^2 = 45$

Concordant pairs = 40

Discordant pairs = 5

$$\tau = \frac{C - D}{\text{total pairs}} = \frac{40 - 5}{45} = \frac{35}{45} \approx 0.778$$

Pandas calculates pairwise correlations

Use the method argument to compare Pearson, Spearman, and Kendall.

```
import pandas as pd

df = pd.DataFrame({
    "study_hours": [1, 2, 3, 4, 5, 6],
    "exam_score": [52, 58, 65, 72, 80, 88],
    "sleep_hours": [8, 7, 8, 6, 7, 6],
})

print(df.corr())  # Pearson is the default
print(df.corr(method="spearman"))
print(df.corr(method="kendall"))
```

The result is a correlation matrix for the numerical columns.