

Correlation is not causation. D8A.ACADEMY/RESOURCES/CORRELATION-CAUSATION

WHAT R SAYS

$r = \text{corr}(x, y)$	-1 to +1, linear only
$r = 0.9$	moves together, no more
$r^2 = 0.81$	variance shared, not caused

THE CLASSICS

ice cream ↔ drowning	blame summer
shoe size ↔ reading	confounder: age
metric ↔ metric	← both grow with time.

TOWARD CAUSATION

randomized test	the gold standard
natural experiment	when you cannot A/B
control for z	compare within groups

THREE WAYS IT LIES

third variable	z drives both x and y
reverse direction	maybe y drives x
pure chance	100 pairs, 5 look real

CHECK YOURSELF

plot it first	r misses curves
$r = 0 \neq$ no link	a parabola scores 0
one outlier	← can fake a 0.8 alone.

GOTCHAS

"drives" in slides	say "moves with"
$p < 0.05$ + big r	significant ≠ causal
de-trend first	time links everything

WORTH MEMORIZING

01 The two lines that fool people

A strong correlation computed honestly, on two series that share nothing but a season.

```
df = beach.groupby("month")[["icecream_sales",
                             "drownings"]].sum()

df.corr().iloc[0, 1]
# 0.91

# same pairs, within one season only
df[df.season == "summer"].corr().iloc[0, 1]
# 0.08 <- the "link" was the sun all along
```

→ Controlling for the confounder (compare within summer only) is the fastest desk check: if r collapses, the third variable owned it.

02 Trap: what it looks like vs what it is

Three correlations your stakeholders will call causal this quarter.

LOOKS LIKE	IS OFTEN	SO
feature use ↔ retention	power users do both	run an experiment
ad spend ↔ revenue	both rise in Q4	de-trend, compare YoY
$r = 0.8, n = 12$	chance plus one outlier	plot it, drop the point

→ Interview phrasing: correlation is a screening tool. It tells you where to look, never what to conclude.