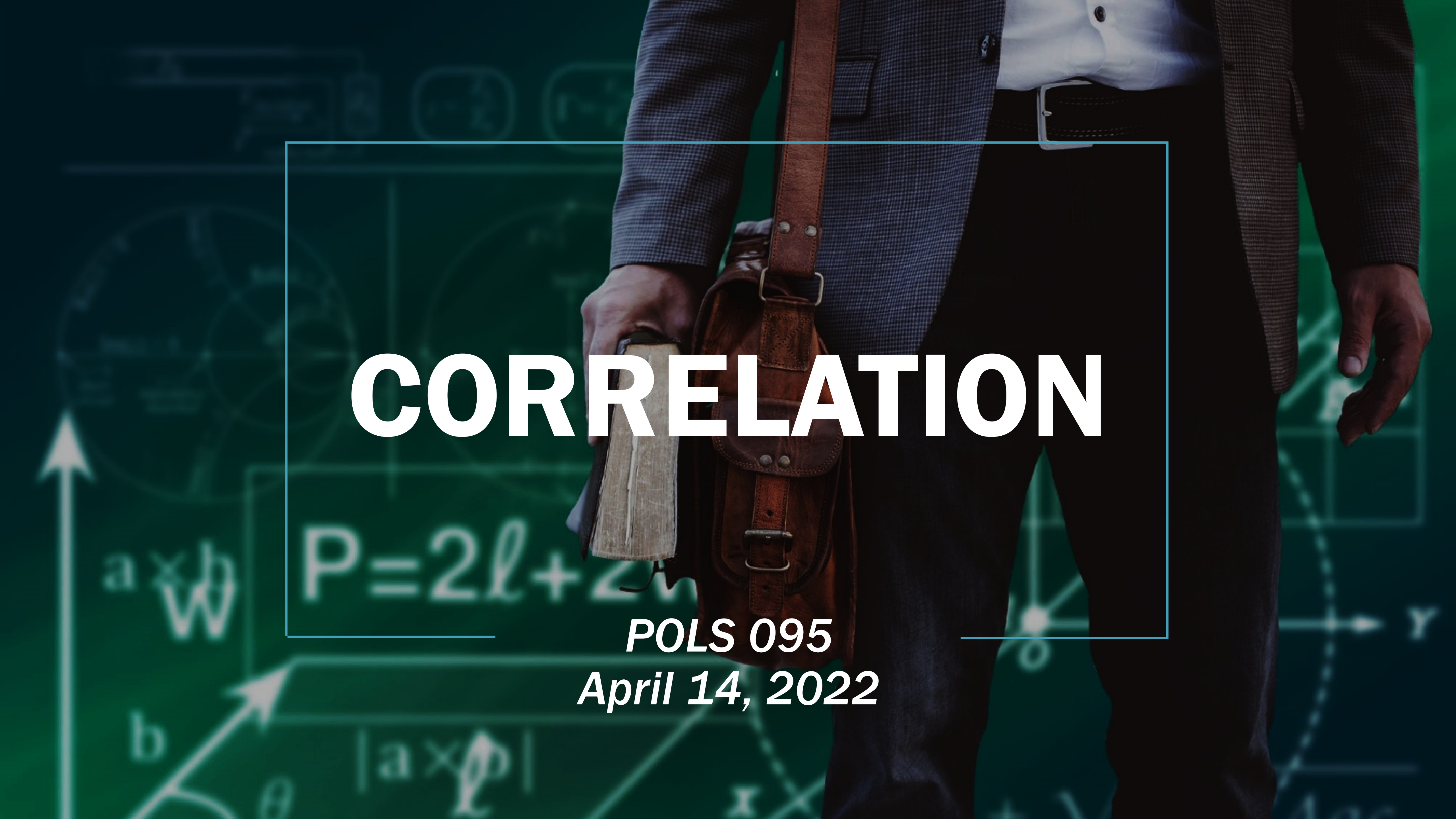
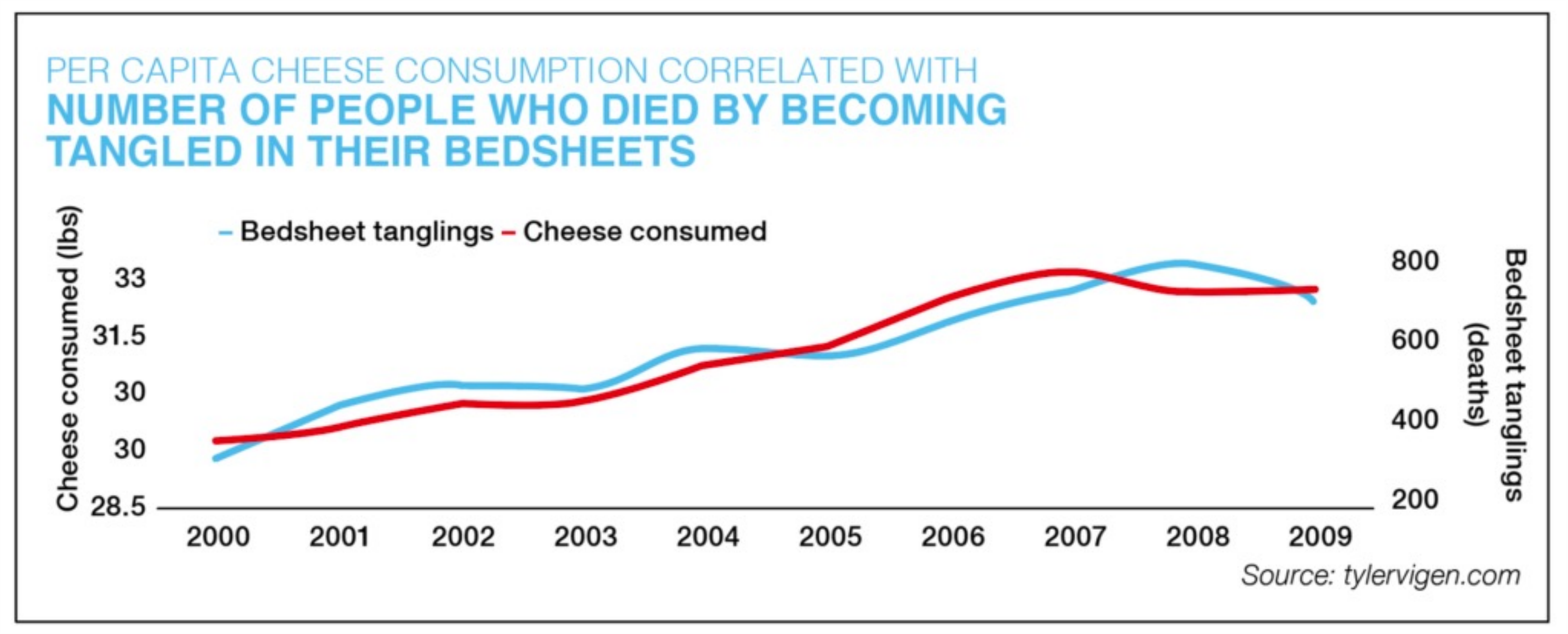


CORRELATION

POLS 095
April 14, 2022

WHAT IS IT?



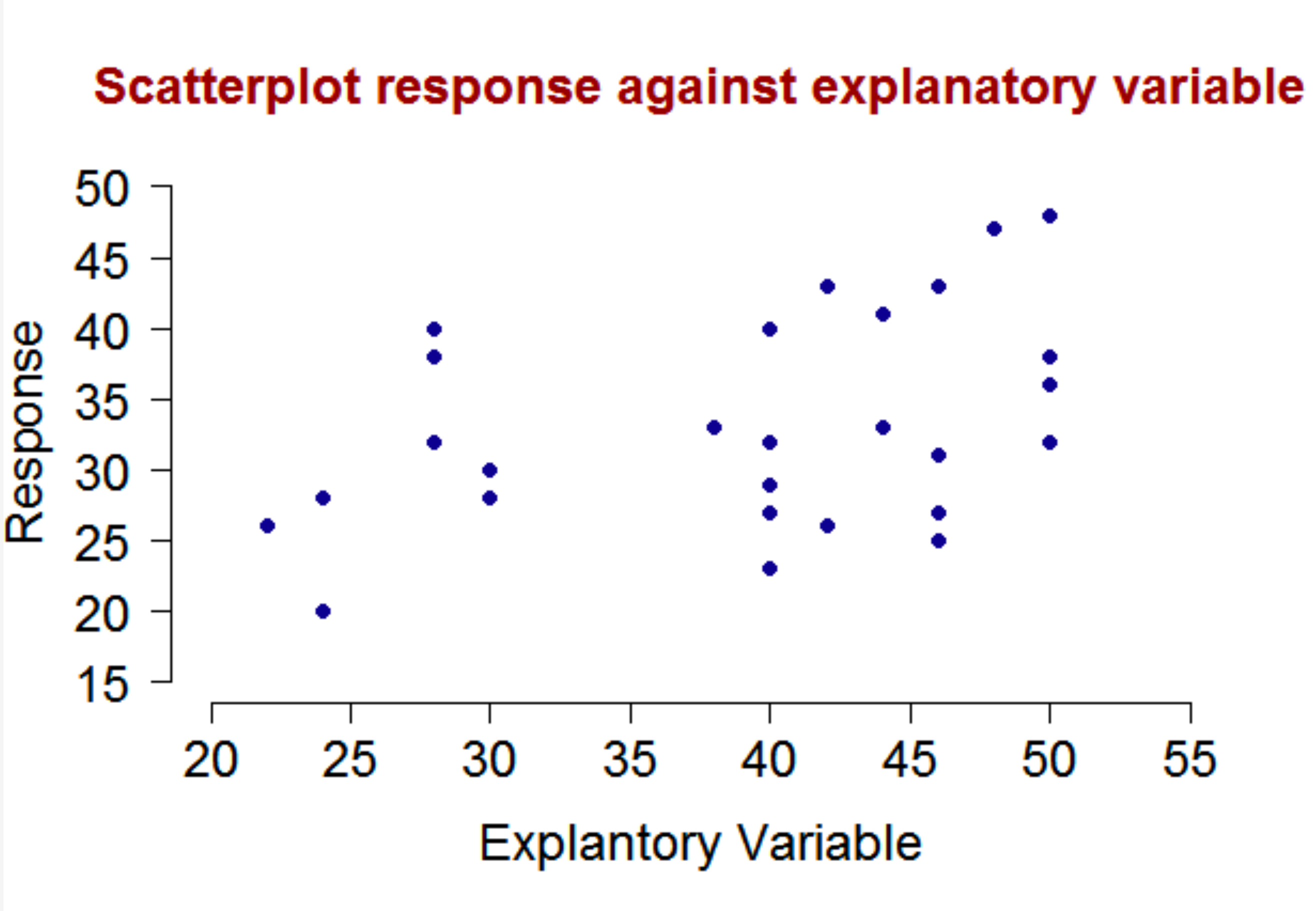
Source: tylervigen.com

- Co ("together) – Relation between two interval level variables

THE SCATTERPLOT

Dependent variable, Y, goes on the vertical/left axis

Independent variable, X, goes on the horizontal/bottom axis



THE SCATTERPLOT

Evaluate the relationship between crime and unemployment in cities

Hypothesis: In a comparison of cities, those with higher rates of unemployment will have higher rates of crime than those with lower rates of unemployment.

For each city, we have data on:

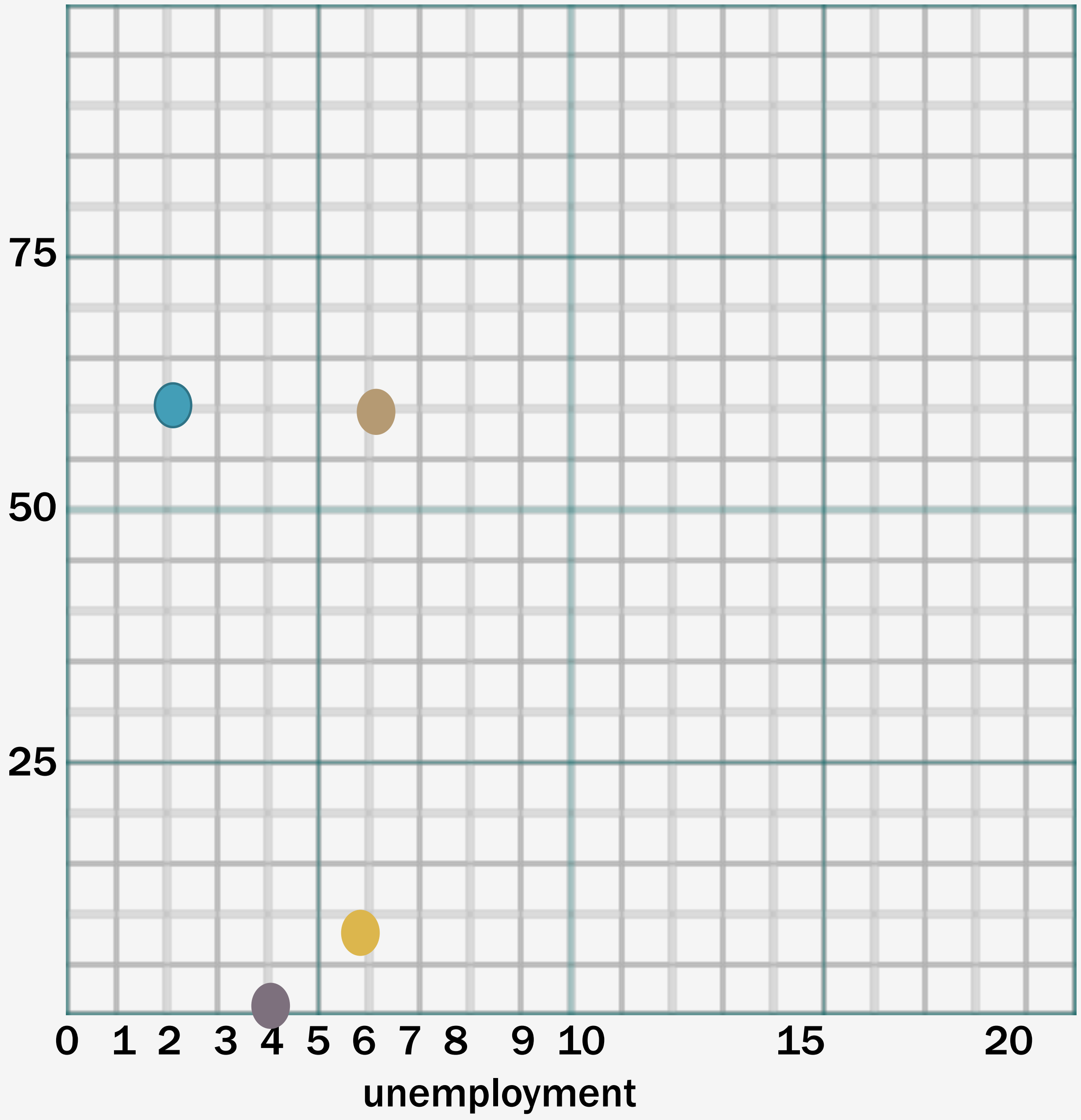
- the number of crimes per 1000 people
- and the rate of unemployment

IV = unemployment

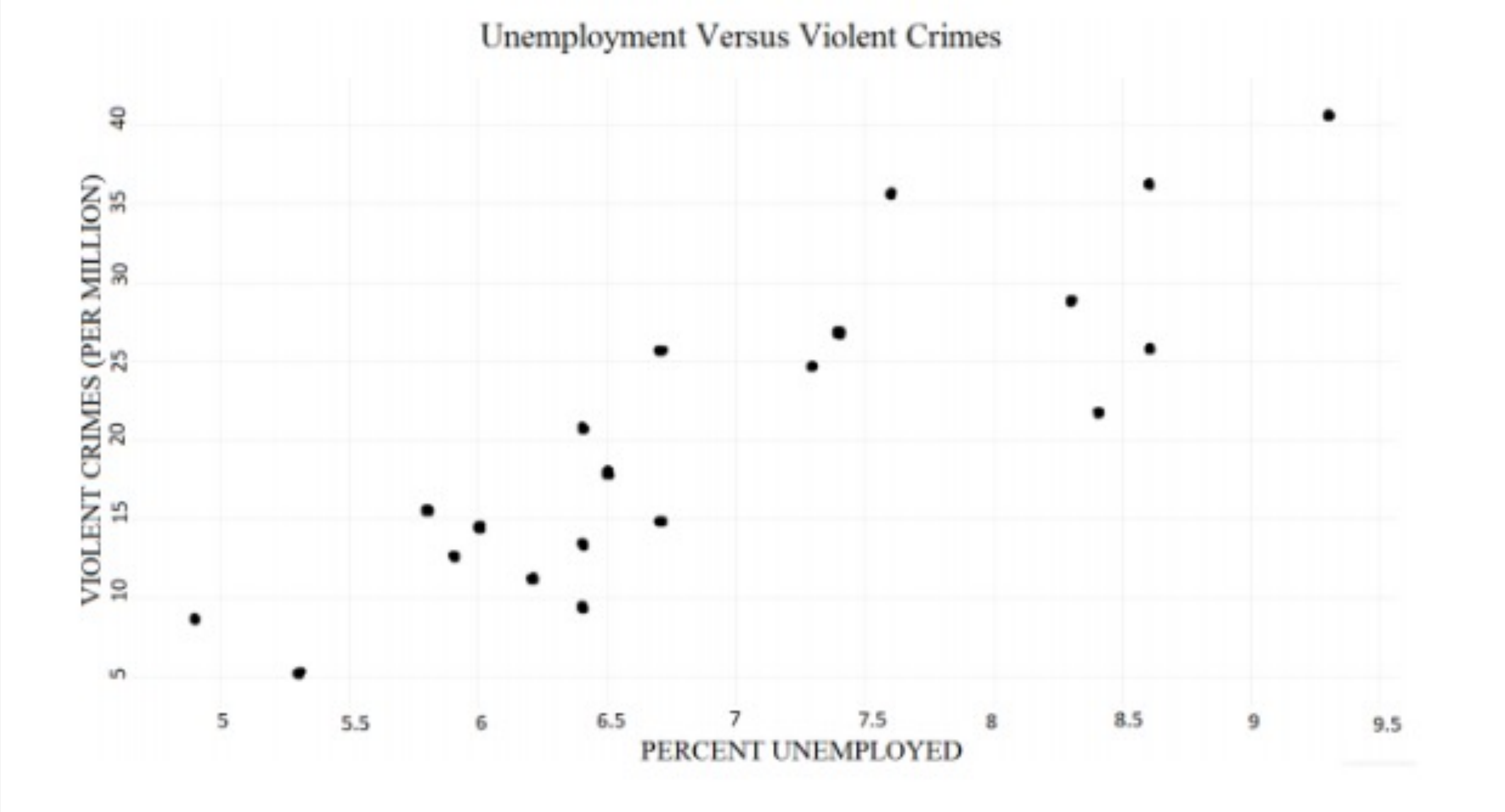
DV = crime

PLOTTING

Seattle unemployment = 3% crime = 60
Tucson unemployment = 6% crime = 8
Des Moines unemployment = 4% crime = .13
New Orleans unemployment = 6% crime = 59



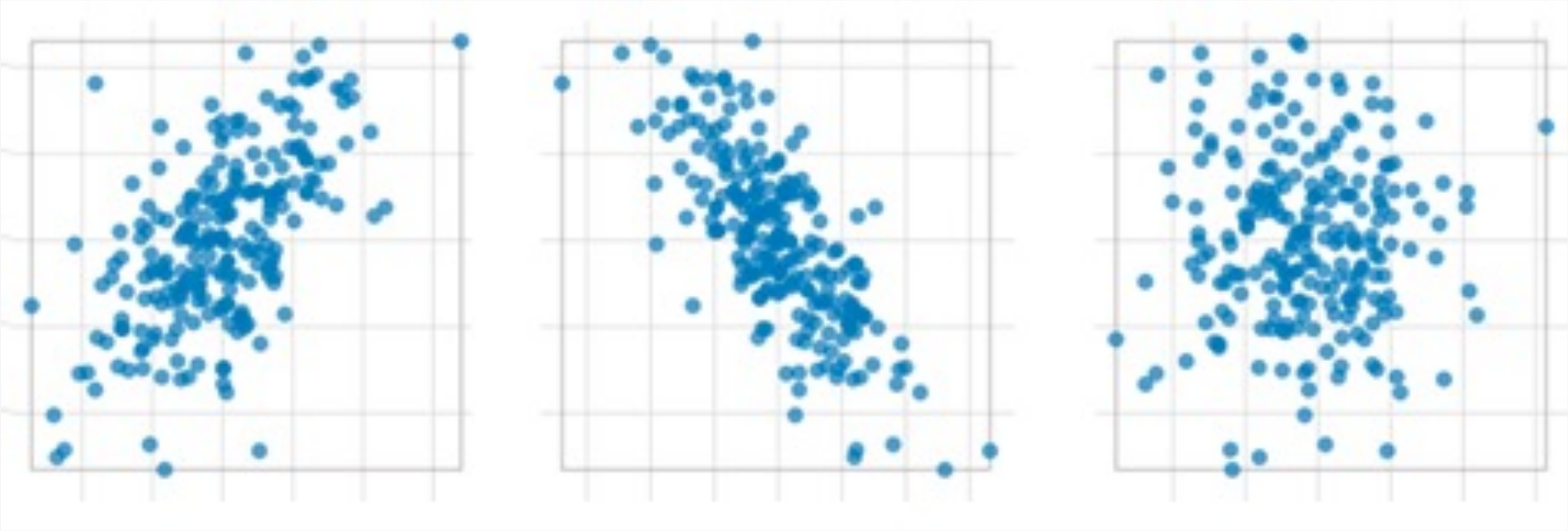
WE CAN LOOK AT THE RELATIONSHIP



What do we see?

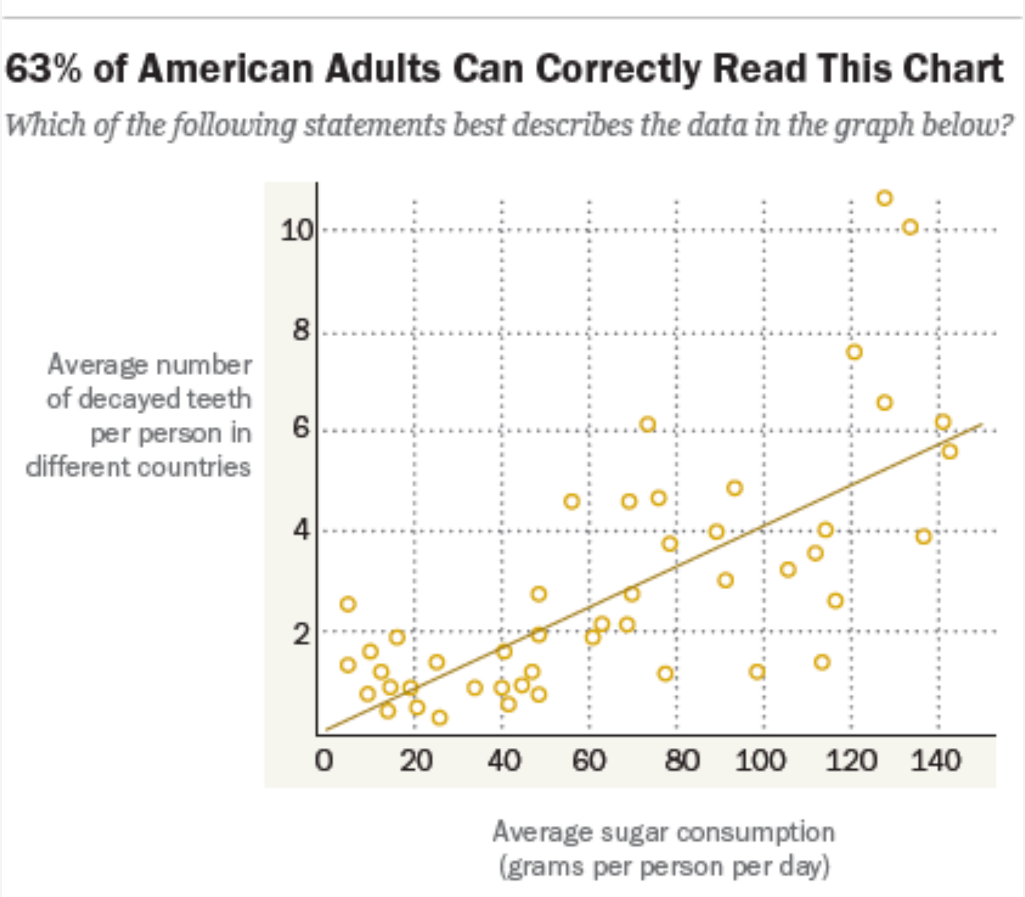
- Lots of variation on both IV and DV
- General upward trend
- Very few points in any clear line

PATTERNS



- General patterns?
  - Positive or negative relationship?
- Positive relationship:
  - High values of the IV are associated with higher values of the DV
  - Lower values with lower values
- Negative relationship:
  - If higher values of IV with lower values of the DV
- Or it may be unclear

READING A PATTERN



63% of American Adults Can Correctly Read This Chart

Which of the following statements best describes the data in the graph below?

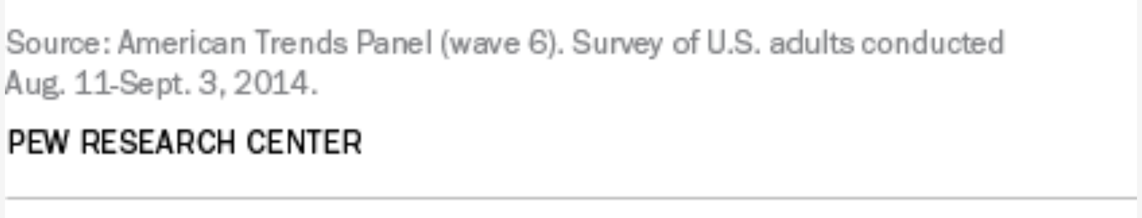
This chart indicates that which of the following:

(a) In recent years, the rate of cavities has increased in many countries.

(b) In some countries, people brush their teeth more frequently than in other countries.

(c) The more sugar people eat, the more likely they are to get cavities.

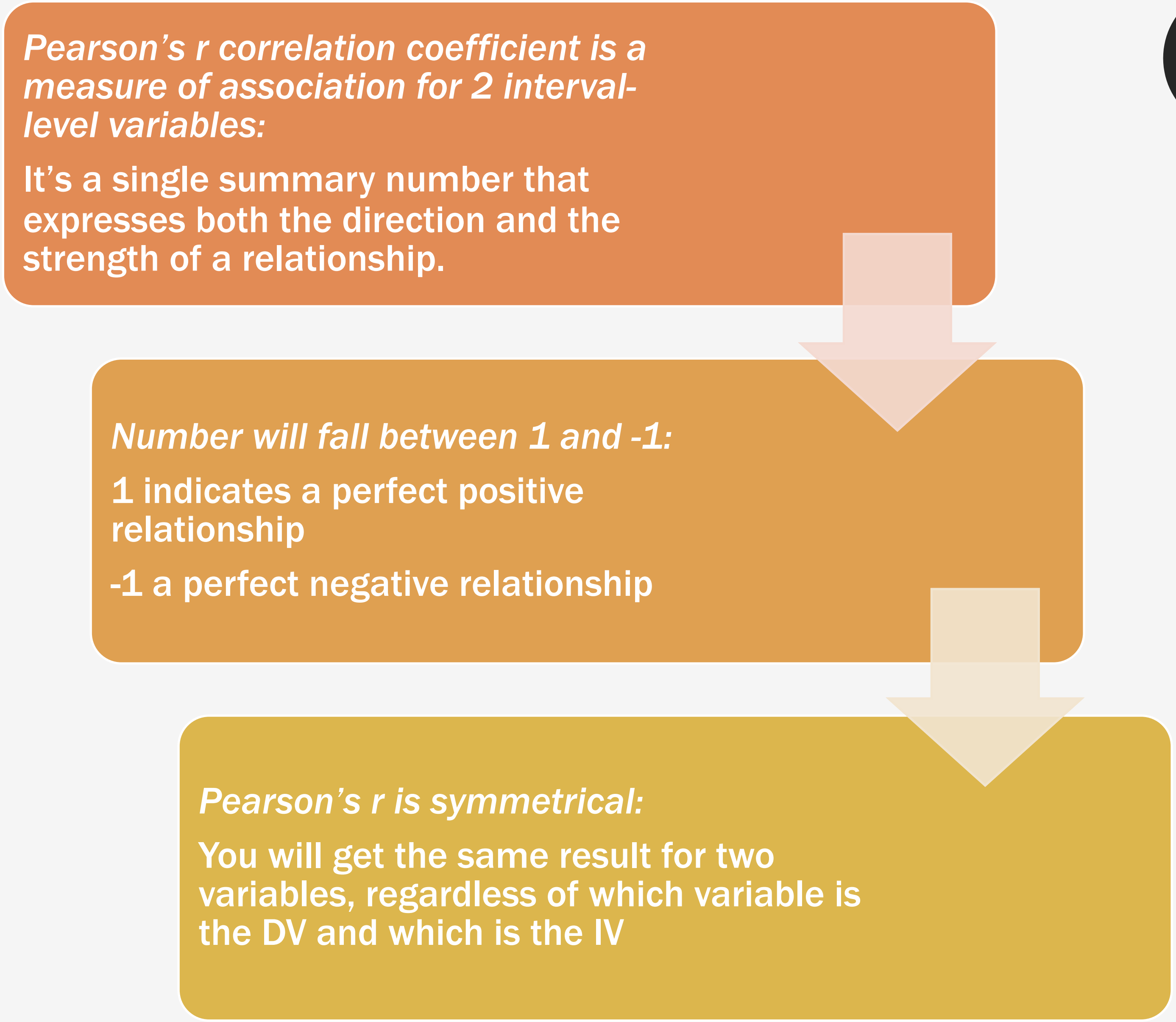
(d) In recent years, the consumption of sugar has increased in many countries.



Source: American Trends Panel (wave 6). Survey of U.S. adults conducted Aug. 11-Sept. 3, 2014.

PEW RESEARCH CENTER

PEARSON’S R



Pearson’s r correlation coefficient is a measure of association for 2 interval-level variables:

It’s a single summary number that expresses both the direction and the strength of a relationship.

Number will fall between 1 and -1:

1 indicates a perfect positive relationship

-1 a perfect negative relationship

Pearson’s r is symmetrical:

You will get the same result for two variables, regardless of which variable is the DV and which is the IV

NO, YOU DO NOT NEED TO CALCULATE…

$$r = \frac{\sum z_x \cdot z_y}{n-1} = \frac{\sum \left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1}$$

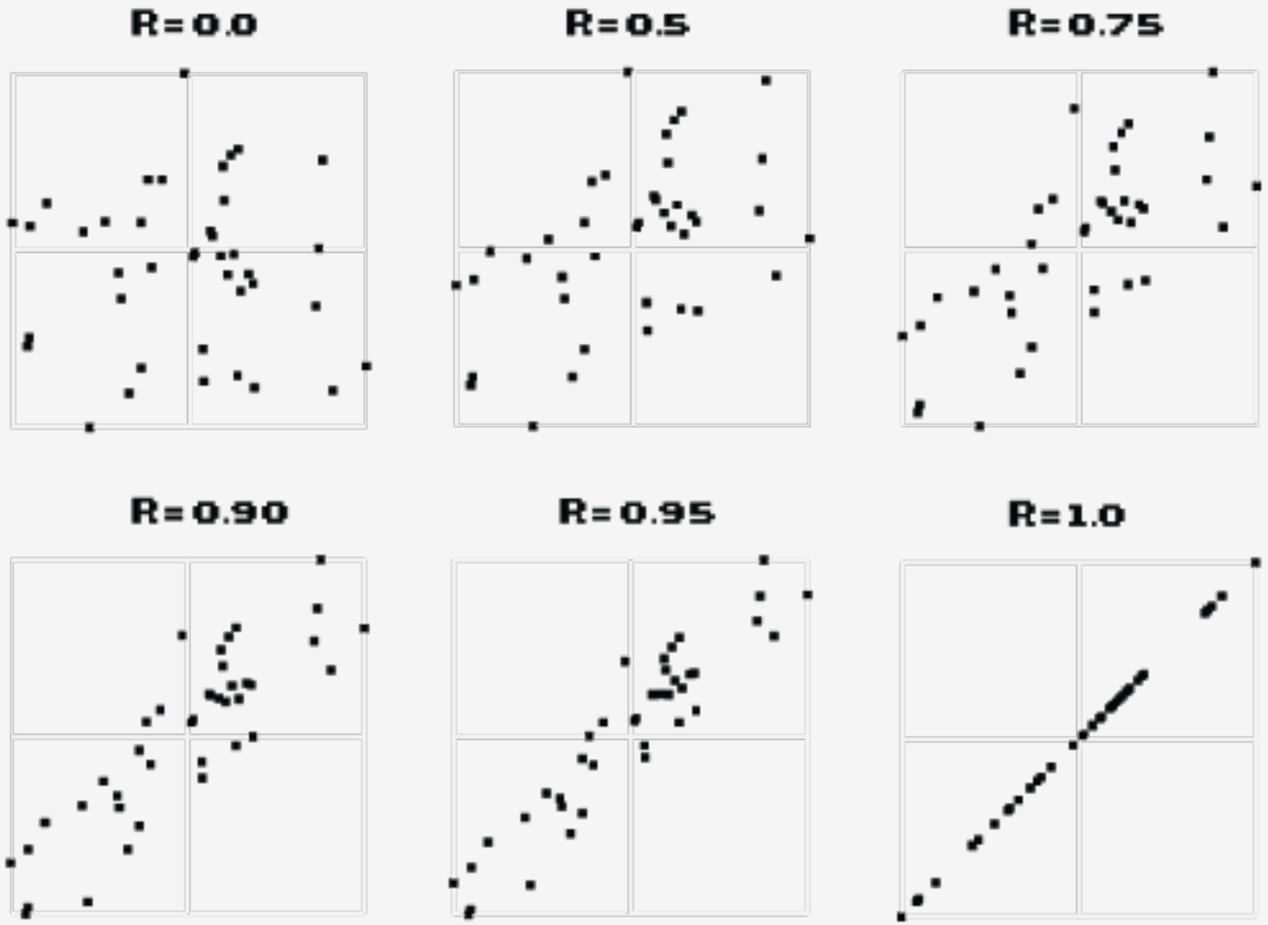
The logic of Pearson’s *r*:
For each observation, it examines how far x is from the mean of x and how far y is from the mean of y, standardized by the standard deviation of x or the standard deviation of y

If an observation is below the mean on both x and y, it will have a positive effect because it is a positive relationship.

INTERPRETING PEARSON'S R

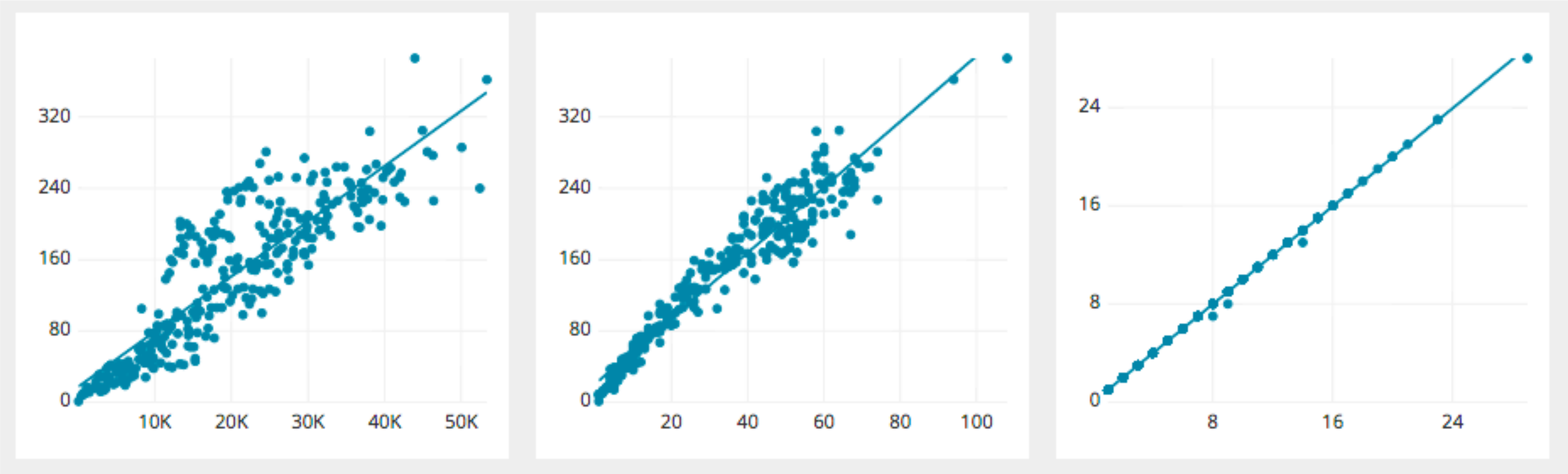
The deviation from the mean tells us how tightly or loosely packed together the data are.

- R = 0 means there is no relationship.
- Higher values indicate a stronger relationship.
- The sign (+ or -) indicates the direction of the relationship



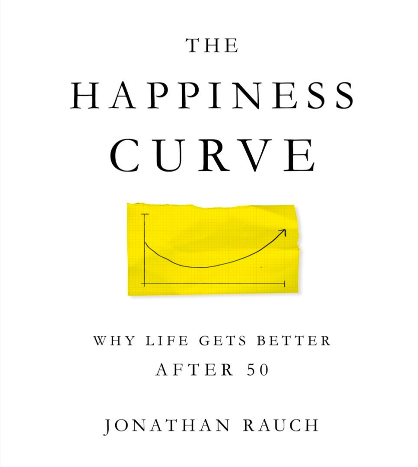
USING THE LAW OF LARGE NUMBERS

- Importantly, Pearson’s r assumes a linear relationship.
- One good reason to look at your scatterplot first is to ensure that this is a good assumption.



WHAT IF YOUR DATA LOOKS LIKE THIS?

This is the well-known Happiness U-curve
If we assume linearity, we may miss something important.... But we assume linearity a lot!



THE
HAPPINESS
CURVE

WHY LIFE GETS BETTER
AFTER 50

JONATHAN RAUCH

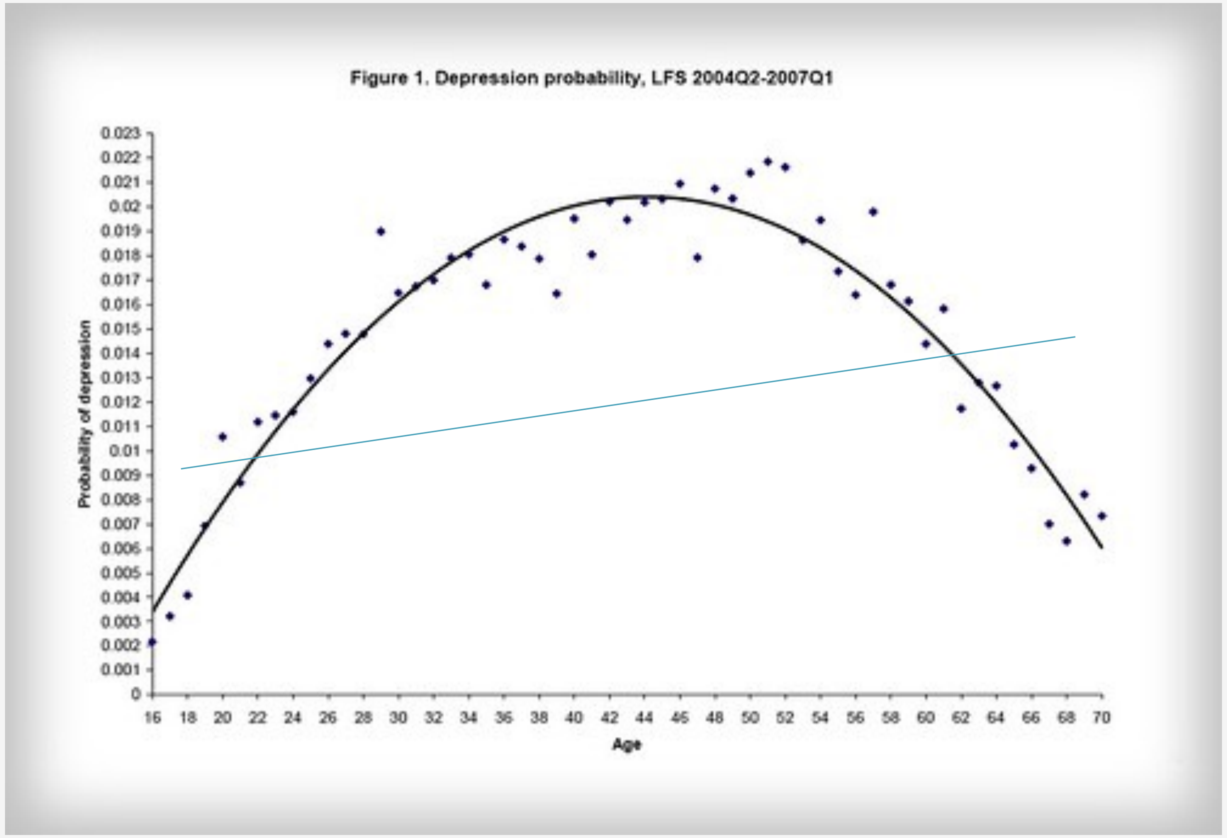


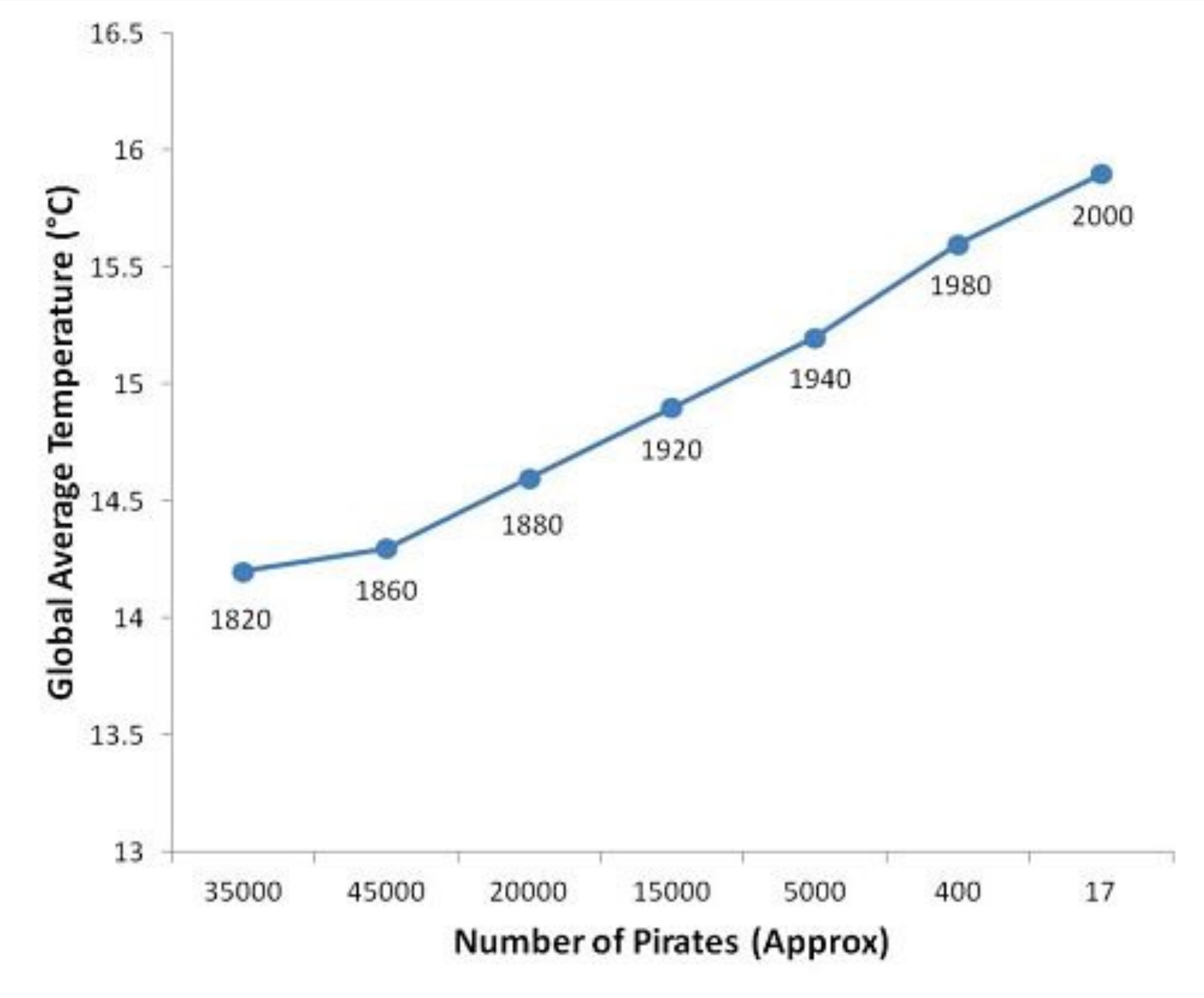
Figure 1. Depression probability, LFS 2004Q2-2007Q1

CORRELATION IS NOT CAUSATION

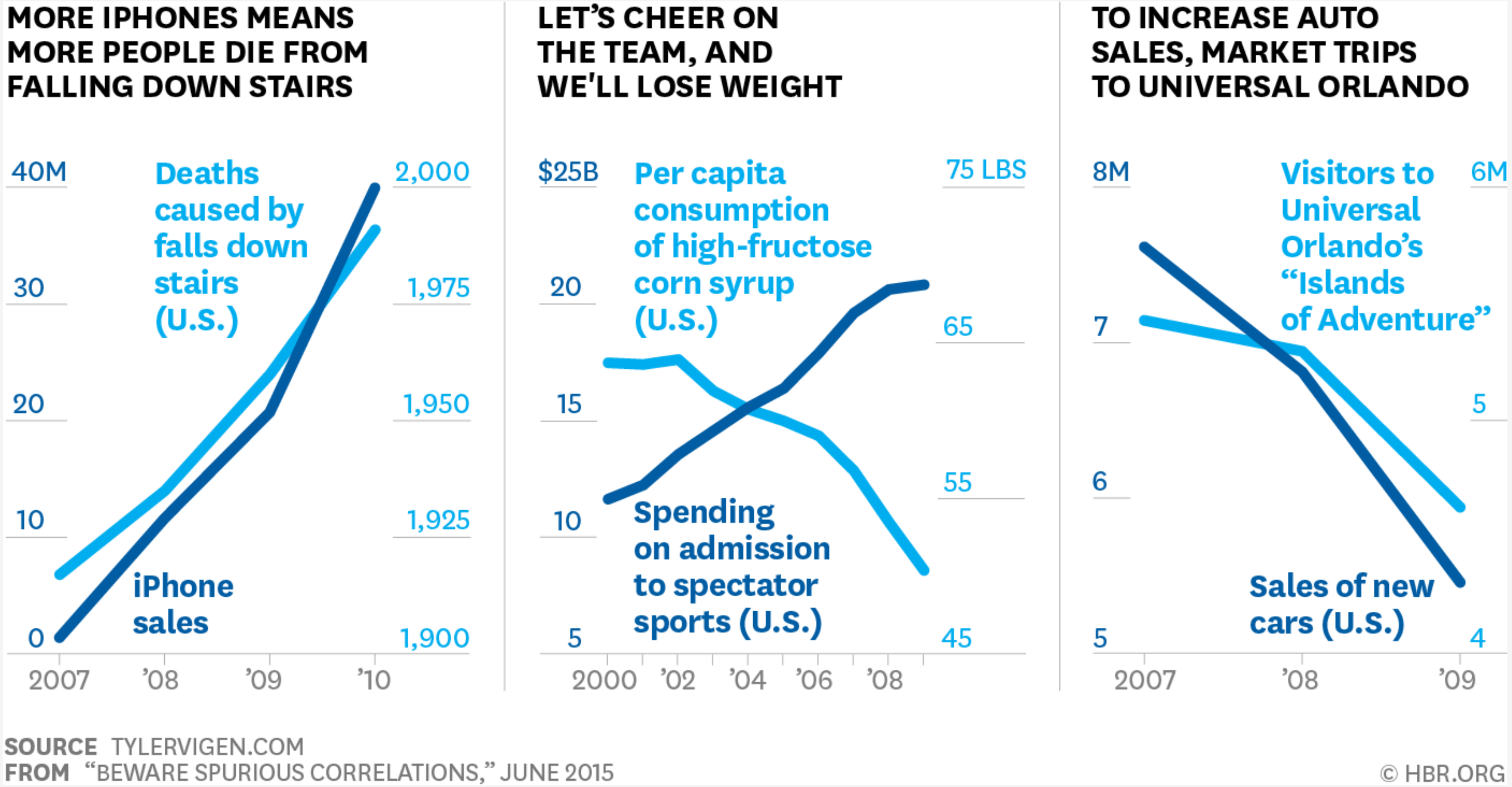
It’s a cliché for a reason!

How do we assess causality?

We will add controls!



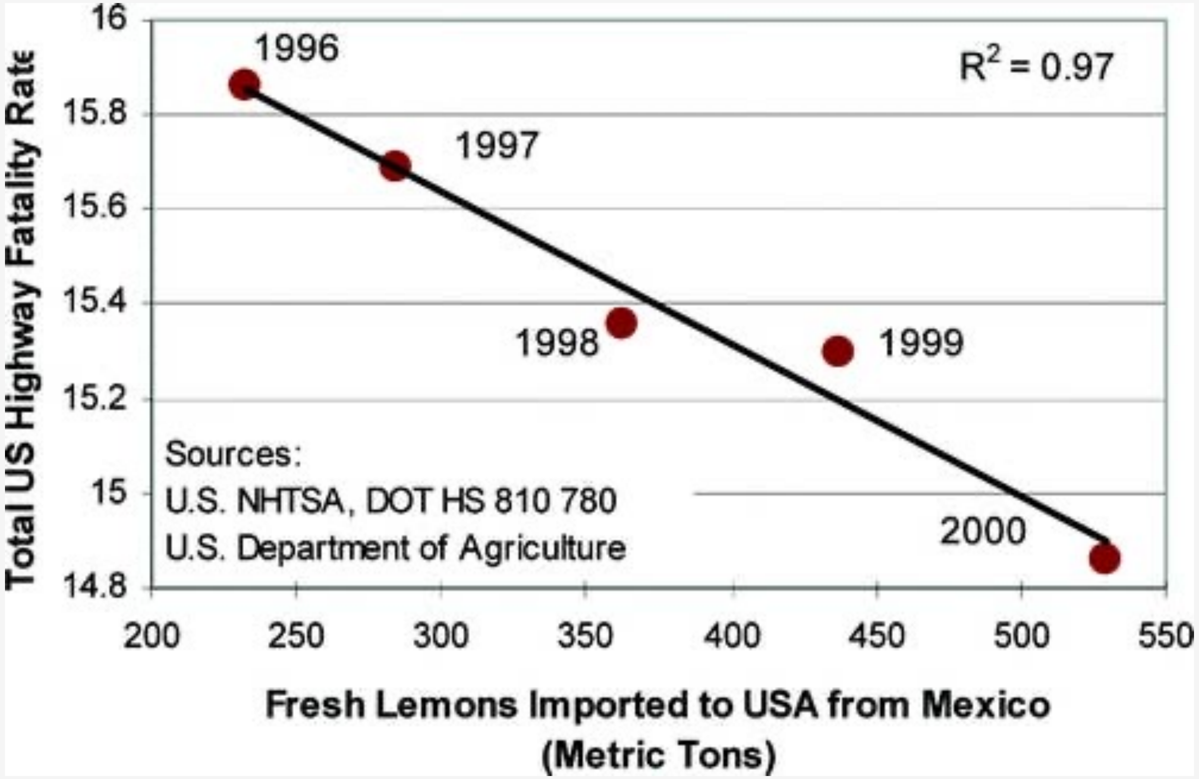
CORRELATION ≠ CAUSATION



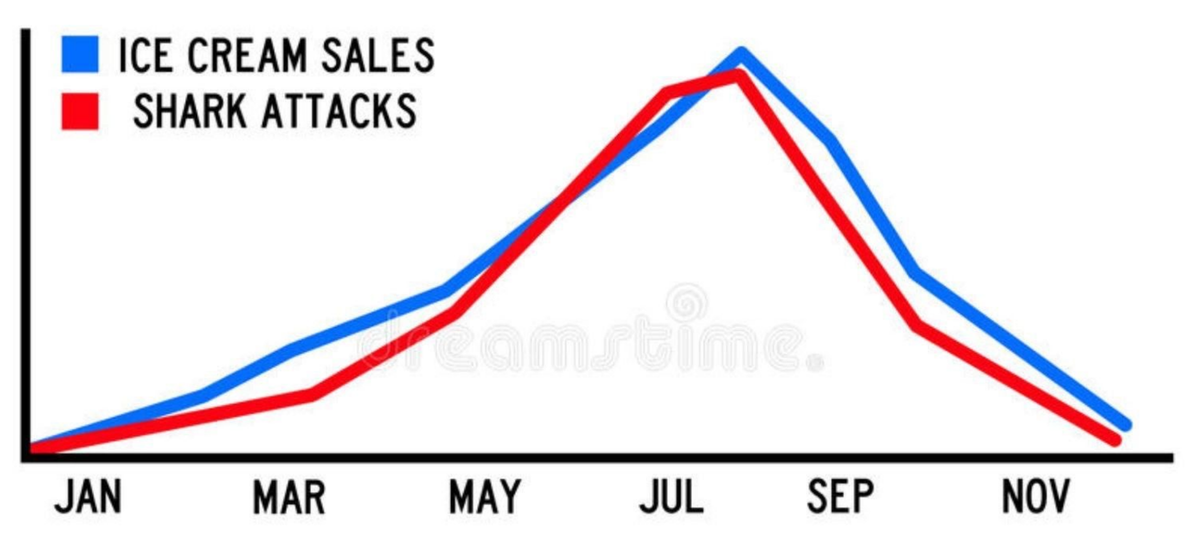
SOURCE TYLERVIGEN.COM
FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

© HBR.ORG

CORRELATION ≠ CAUSATION

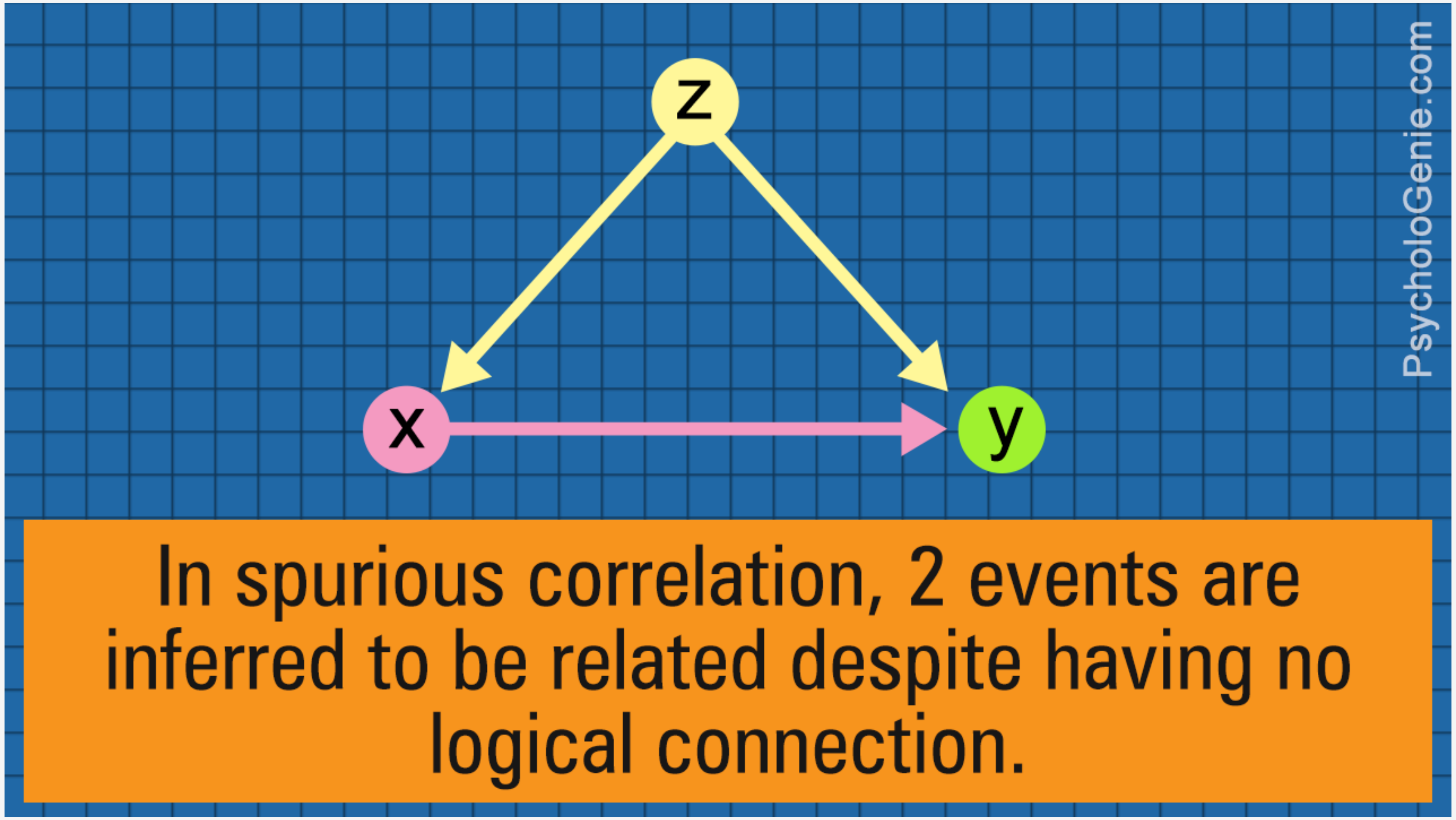


Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

SPURIOUS CORRELATION



In spurious correlation, 2 events are inferred to be related despite having no logical connection.

- The data we're working with won't be so wildly unconnected, but we still need to be concerned about spurious correlation
- That's why we need controls!
- https://www.tylervigen.com/spurious-correlations