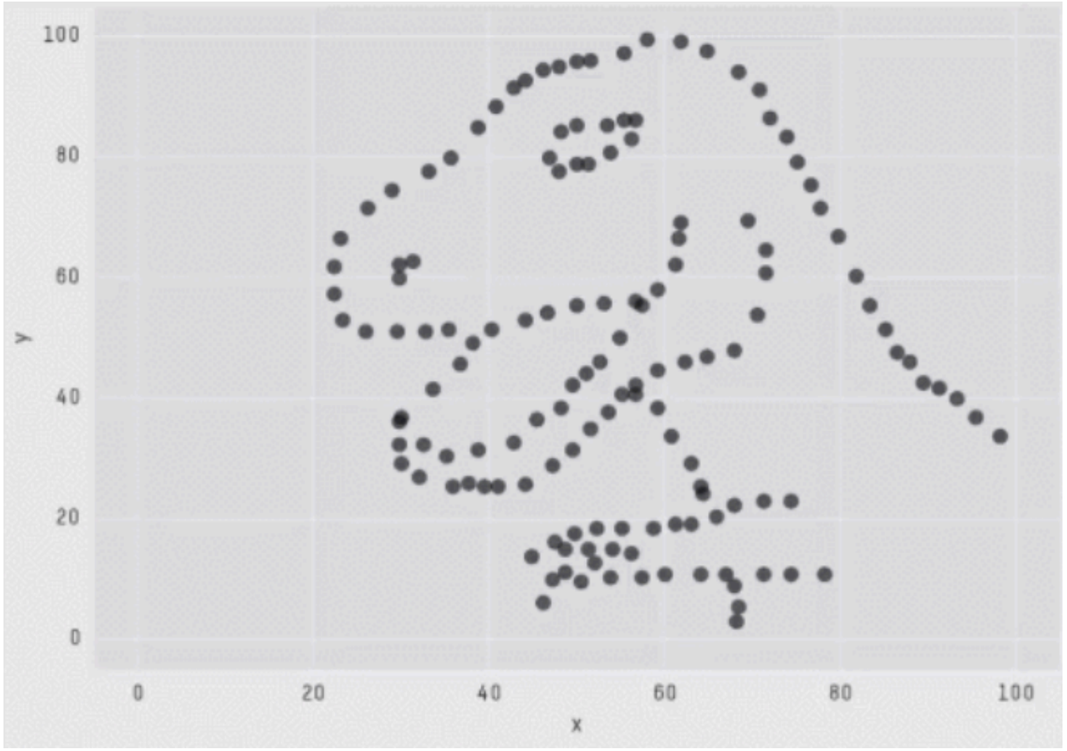
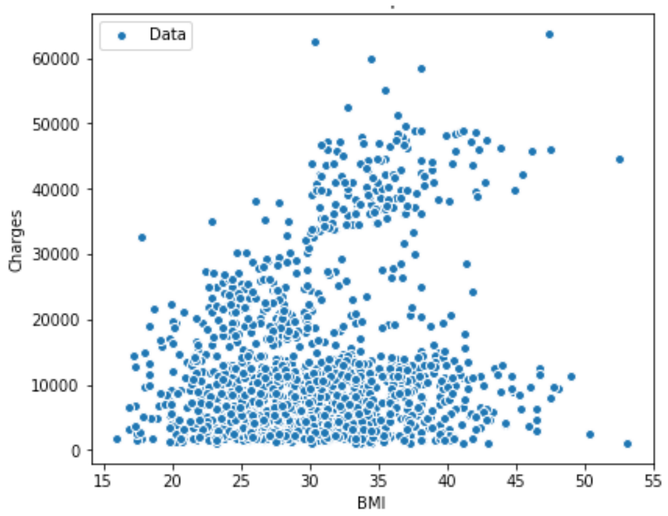
Lecture 14: Data relationships and causation

Instructor: Michael Szell
Oct 13, 2023

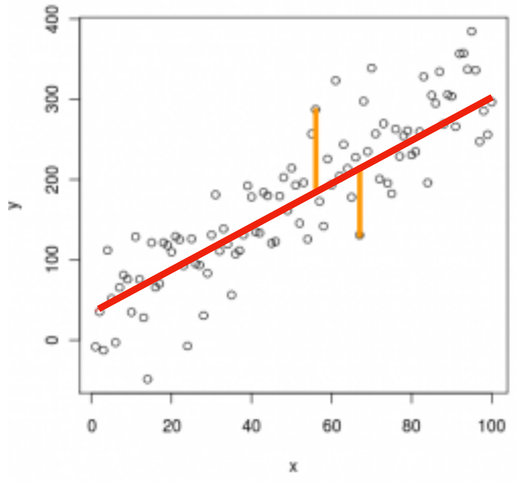


Today you will learn about relating different variables

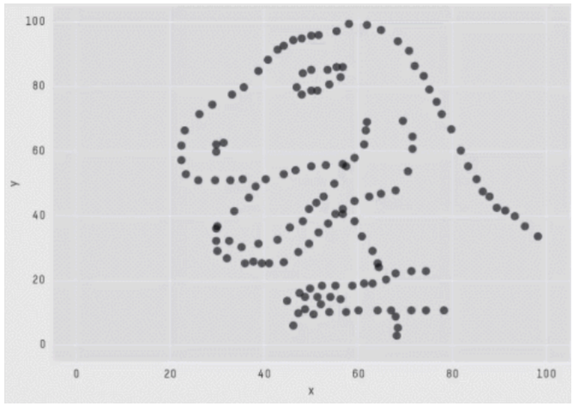
Scatterplots



Correlation and regression



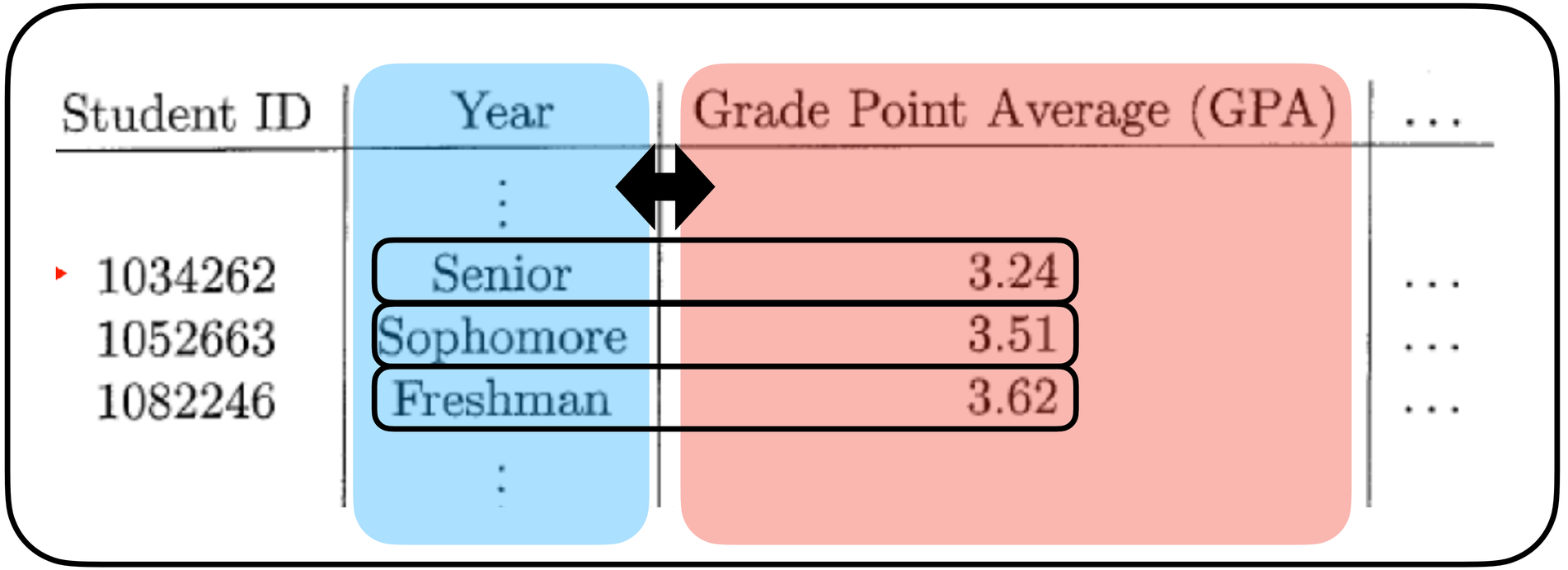
Associations and causation



Today you will learn about relating different variables

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

We are interested in the connection of variables



Bivariate data analysis

Two variables are associated if some values of one variable tend to occur more often with some values of the second variable

Smoking

Life expectancy

Statistical associations are tendencies which allow individual exceptions (anecdotal evidence)

100-year-old woman says drink and cigarettes keep her young

A woman who toasted her 100th birthday today with a cigarette and a tot of whisky said she would also be raising a glass to 70 years as a committed smoker and drinker.

https://www.telegraph.co.uk/news/newstopics/howaboutthat/7360761/100-year-old-woman-says-drink-and-cigarettes-keep-her-young.html

We can relate different types of variables

Categorical   Categorical

Quantitative   Quantitative

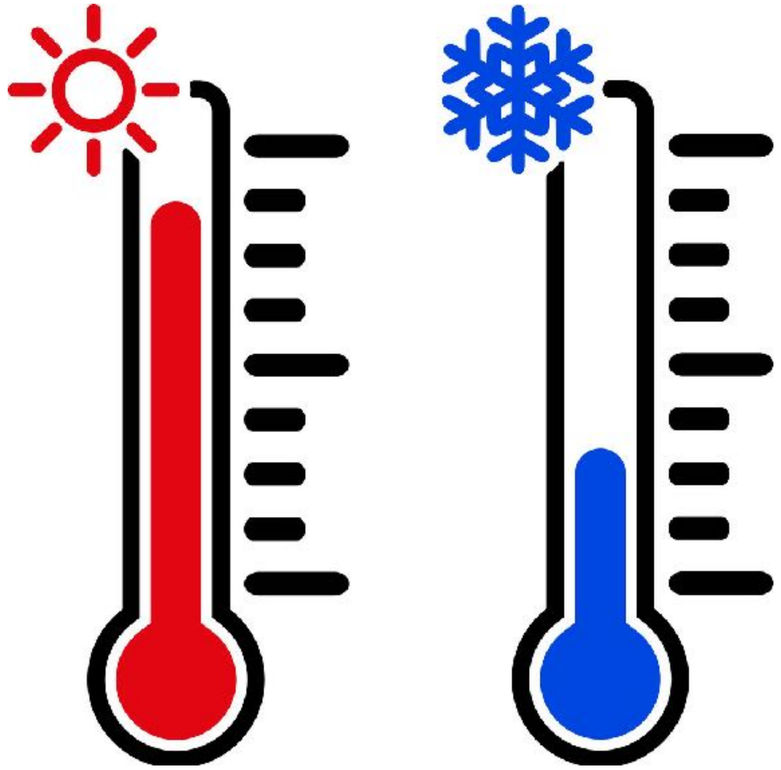
Categorical   Quantitative

There can be different reasons to relate variables

- Explore the relationship
- Show that one can explain/predict variation in the other

A response variable measures an outcome of a study.
An explanatory variable explains or causes its changes.

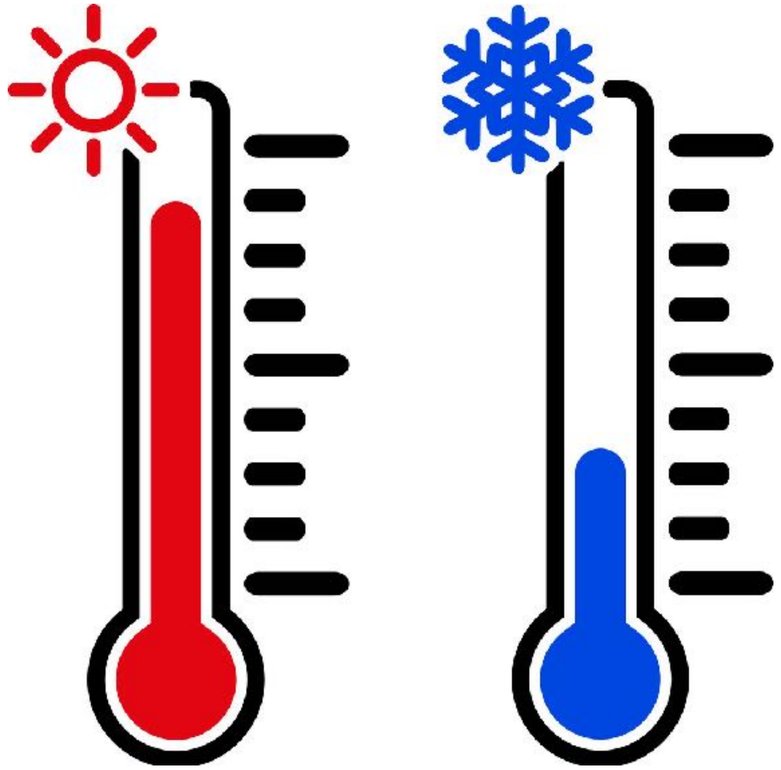
Alcohol
Explanatory



Drop in body temperature
Response

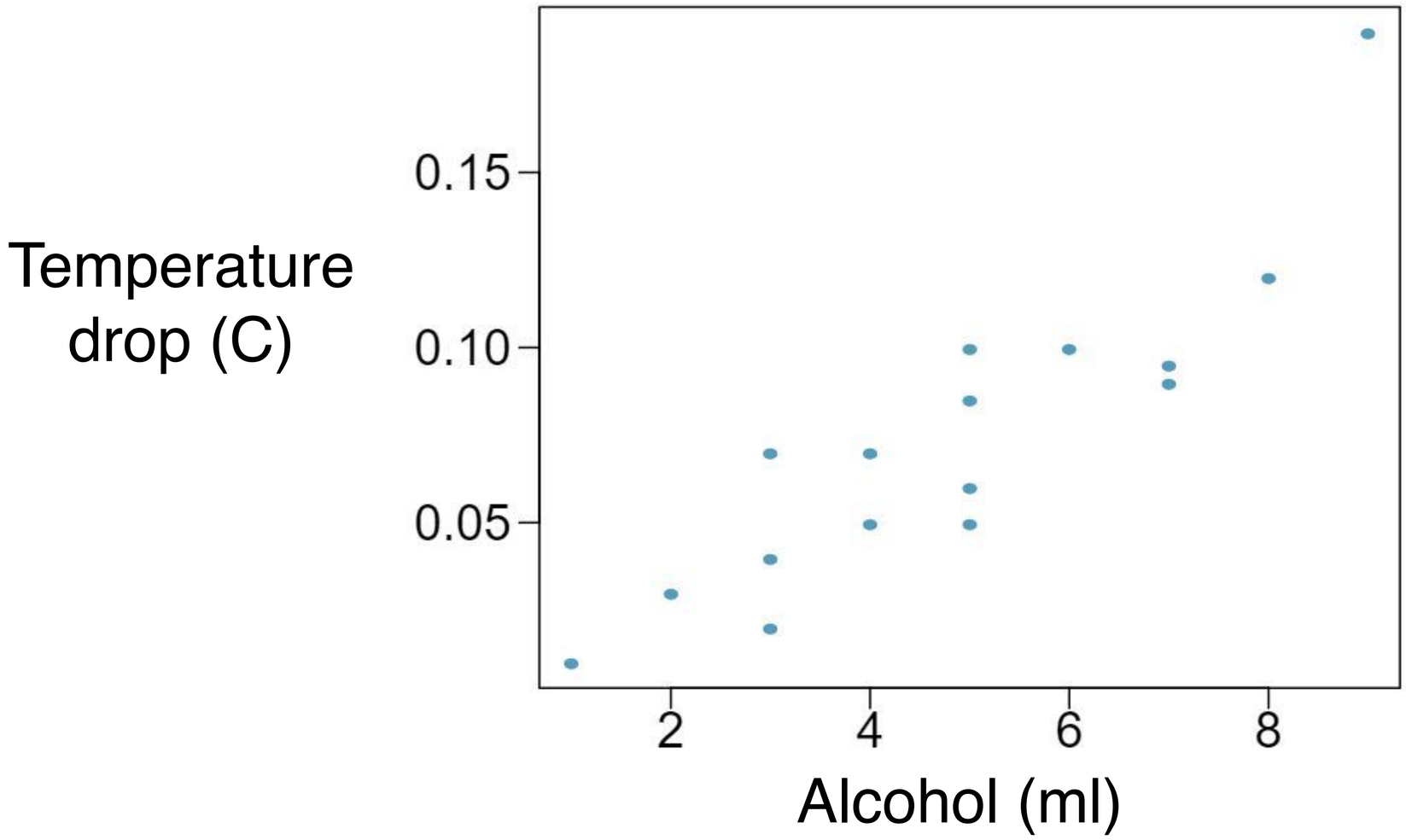
A response variable measures an outcome of a study.
An explanatory variable explains or causes its changes.

Alcohol
Independent



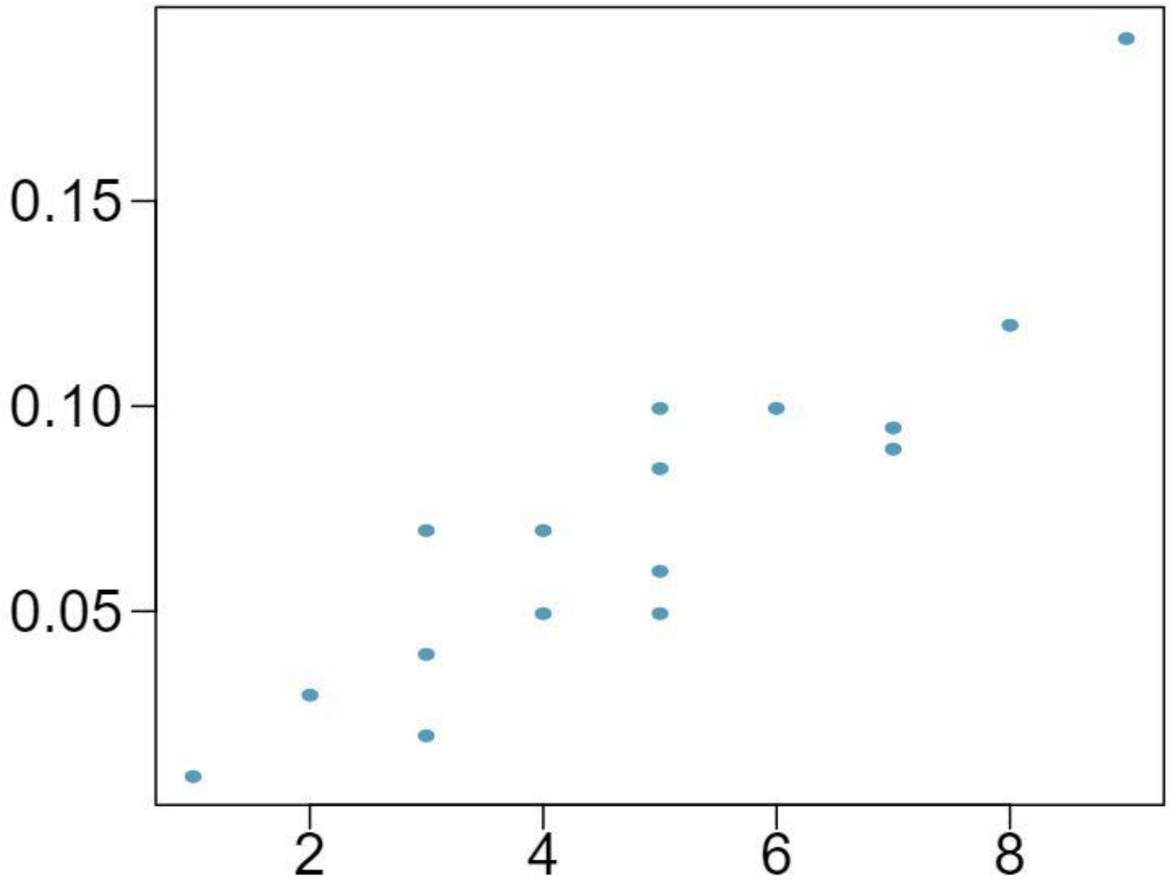
Drop in body temperature
Dependent

A scatterplot shows the relationships between two quantitative variables



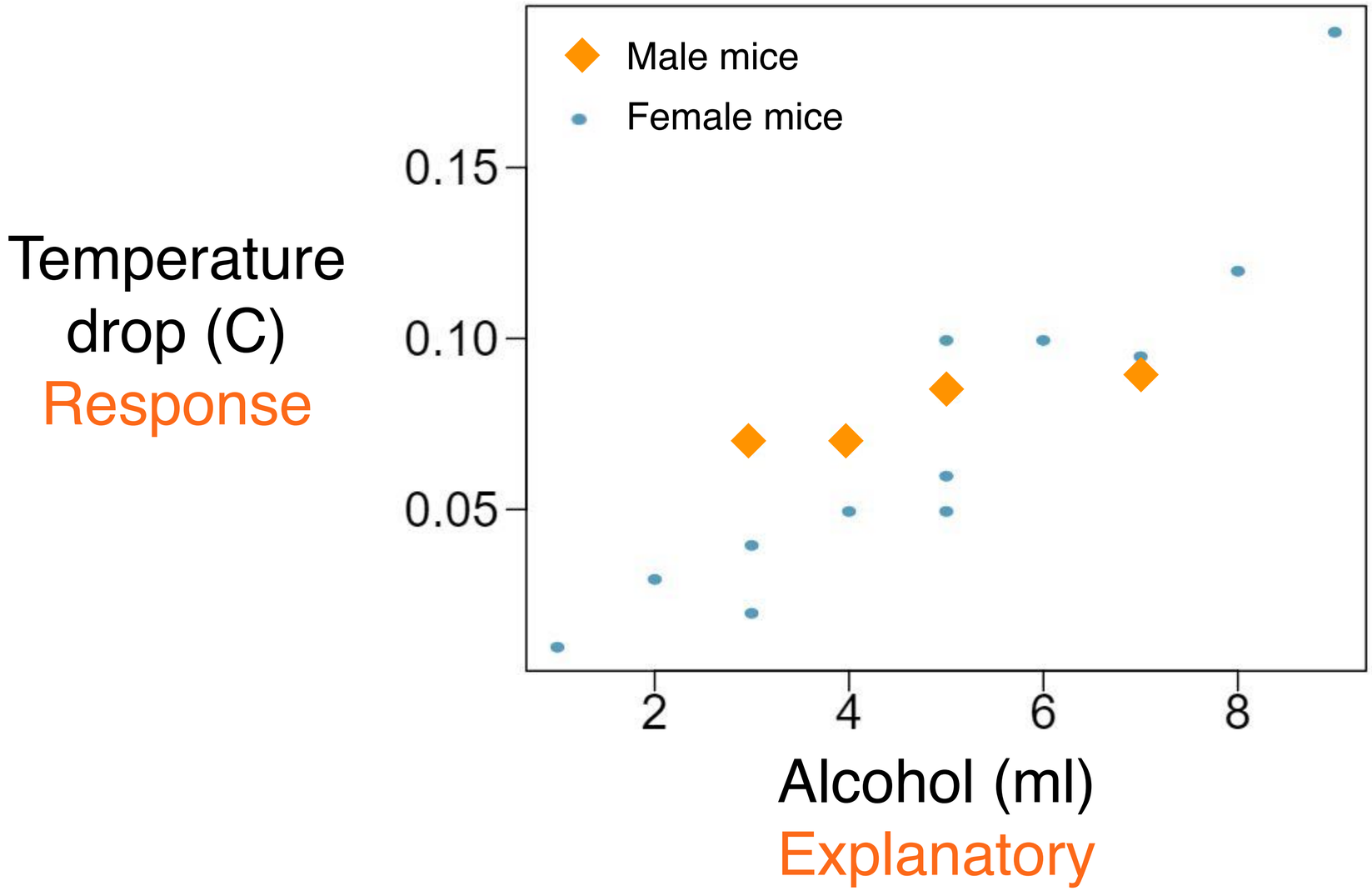
If there is an explanatory variable, we always plot it on the horizontal x-axis

Temperature drop (C)
Response



Alcohol (ml)
Explanatory

To add a categorical variable to a scatterplot, use a different plot color or symbol for each category



There can be different reasons to relate variables

- Explore the relationship
- Show that one can explain/predict variation in the other
- Show that one causes variation in the other

High school grades can predict university success, but do not cause it

The recipe for two-variable exploration is the same as for one

1) Make a scatterplot

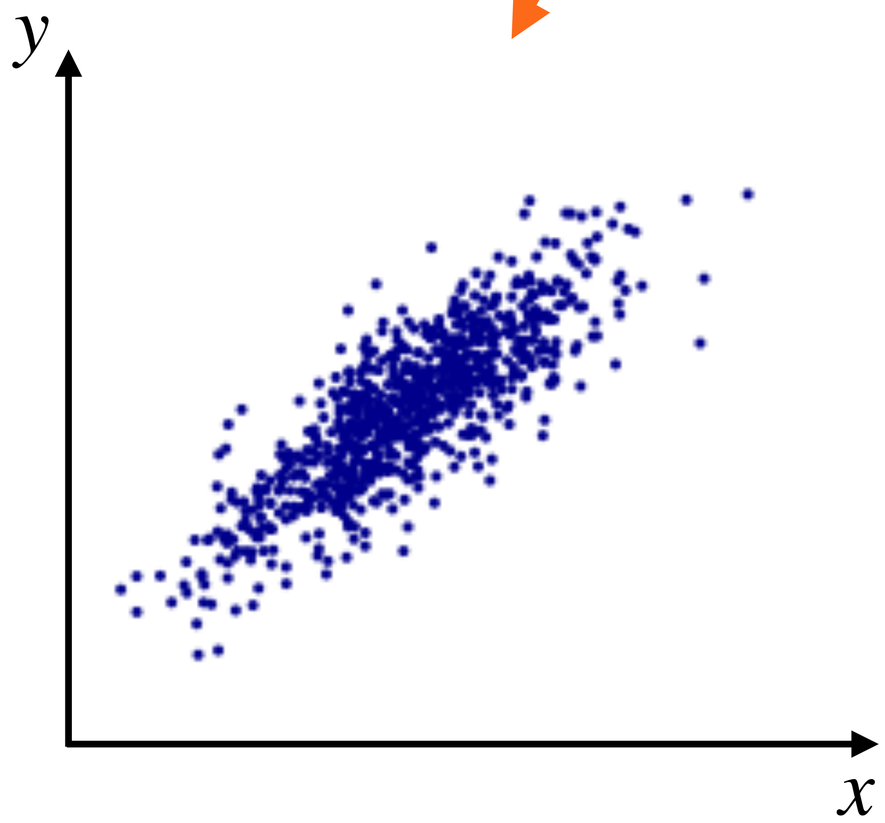
2) Look for the overall pattern and deviations, outliers

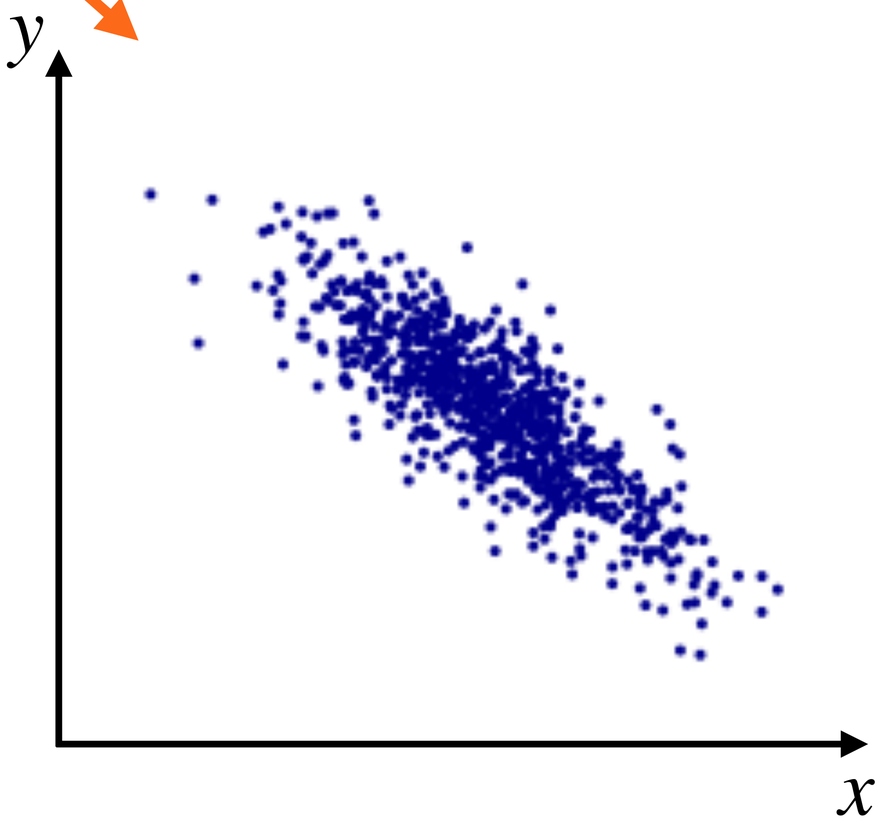
3) Calculate a numerical summary

4) When the relationship is regular we can describe it with a smooth curve (a mathematical model)

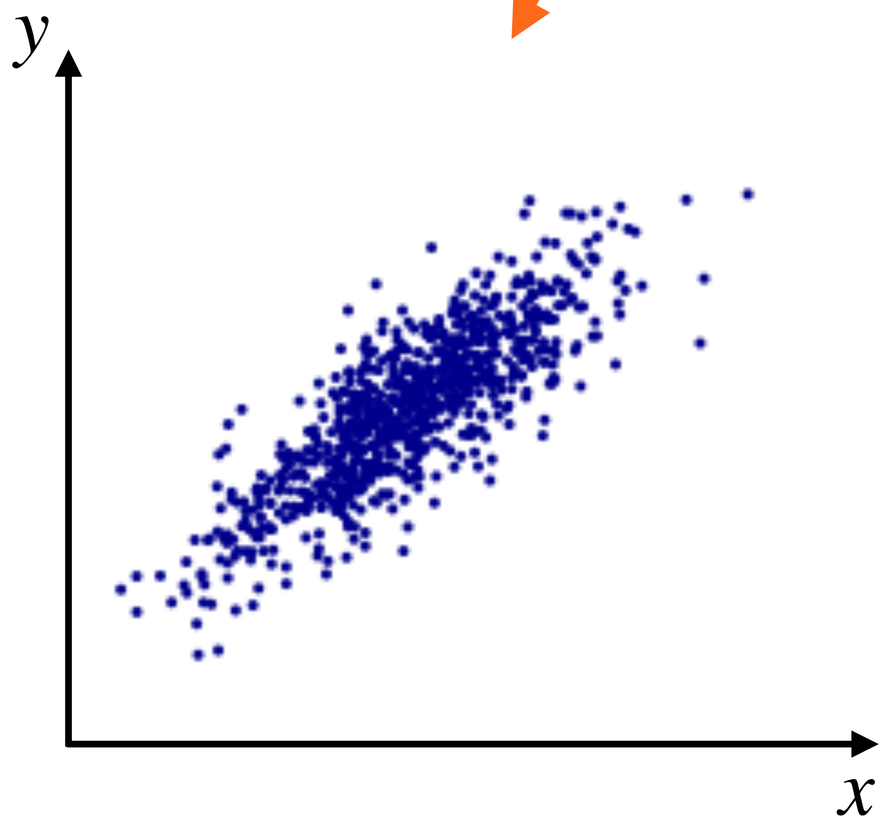
Correlation

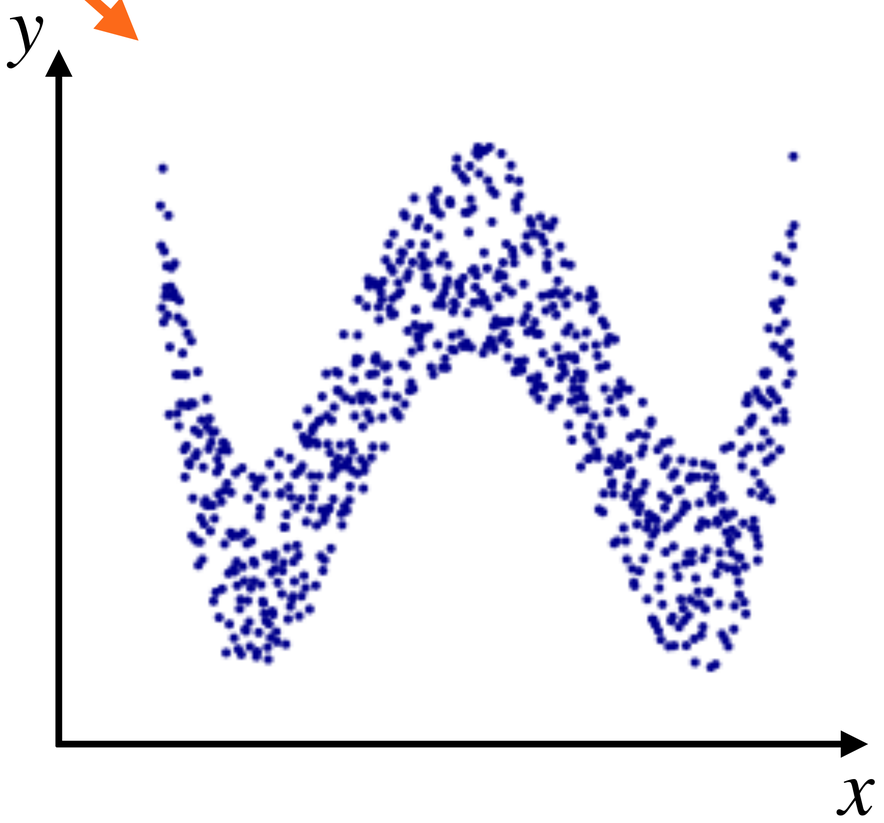
Association can be positive or negative



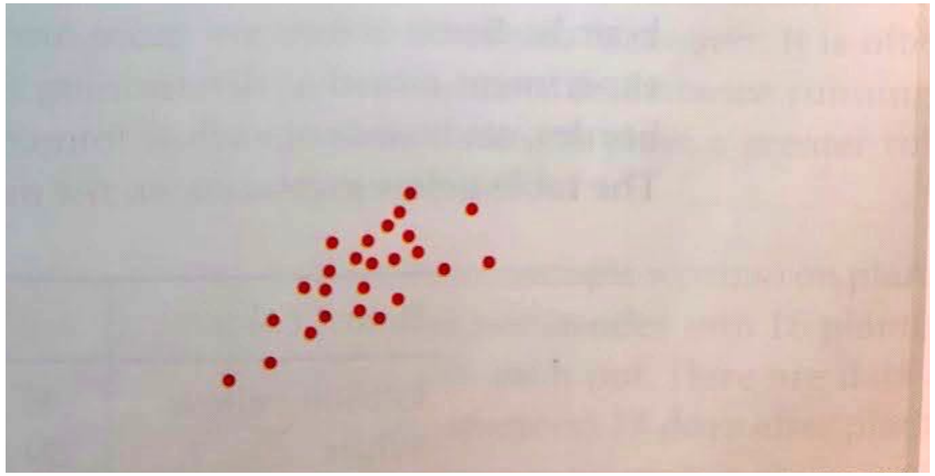


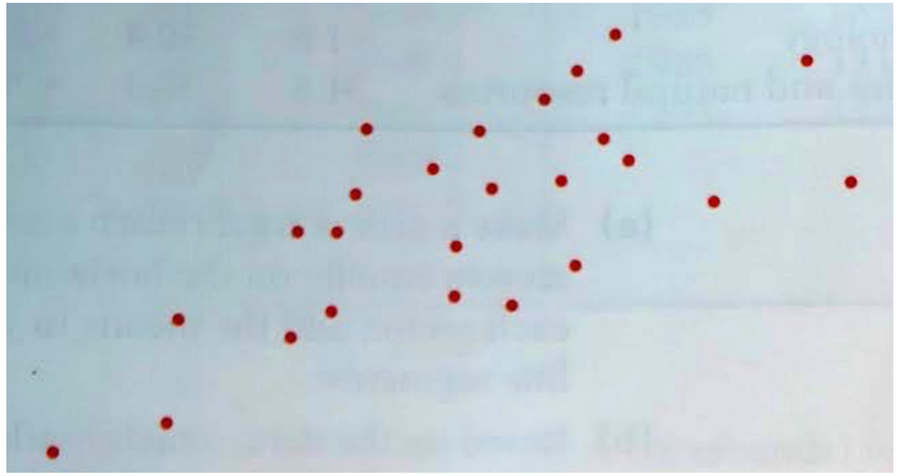
Association can be linear or nonlinear



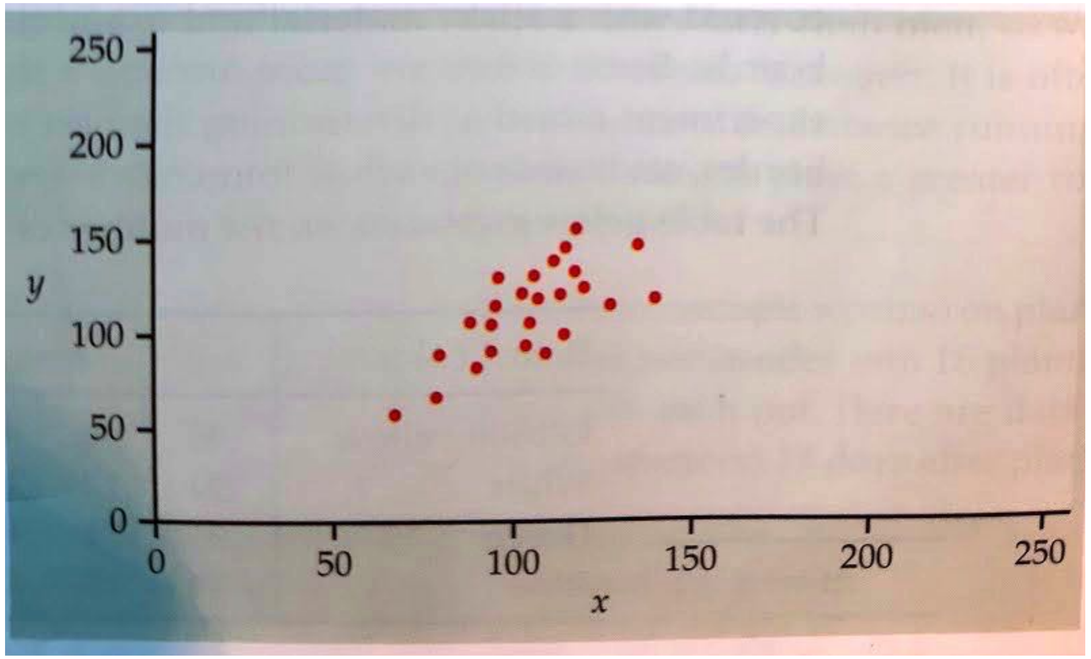


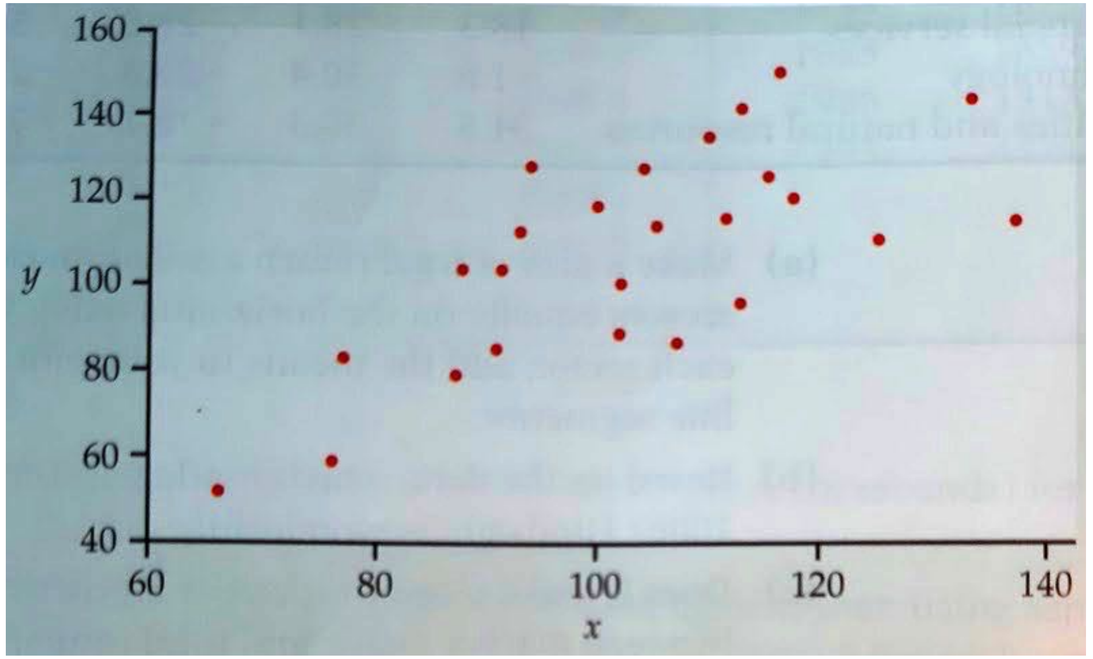
Which points have the stronger linear relationship?





Our eyes are not good for judging linear relationship





The mean $\bar{x}$ is the average value

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum x_i$$

It gives an idea of the "center" of the distribution

The standard deviation s measures spread

variance: $$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

$$= \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

The correlation coefficient r is defined as

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

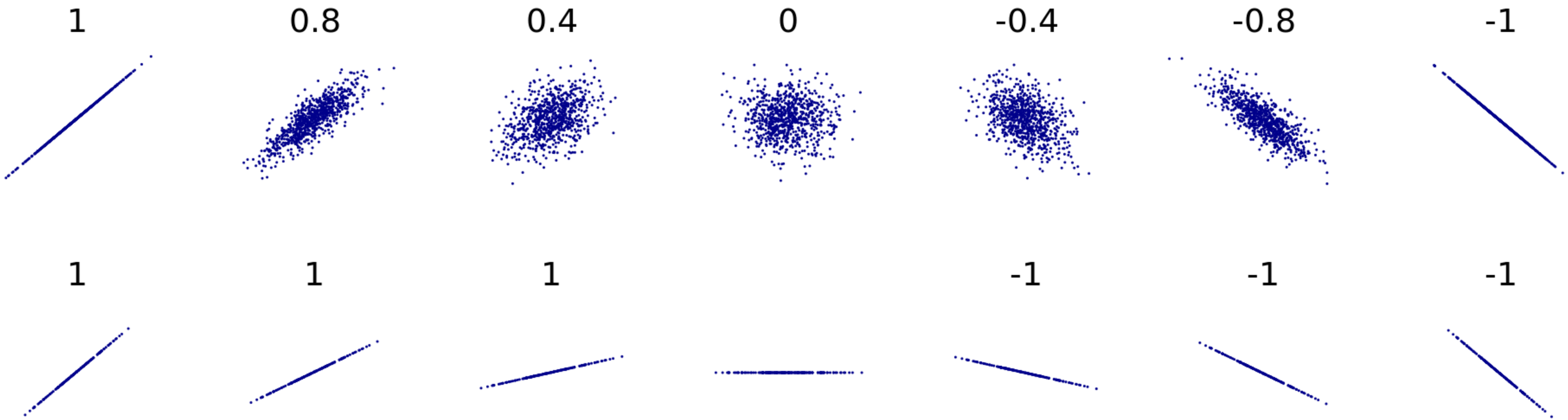
The correlation coefficient r is defined as

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

standardized x values standardized y values

Therefore r does not change with linear transformations

Correlation measures strength of linear relationship



I'M CHALLENGING YOU

GUESS THE CORRELATION

http://guessthecorrelation.com/

GUESS THE CORRELATION

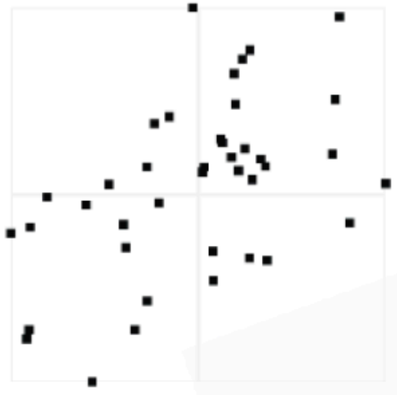
ABOUT THE GAME

THE AIM OF THE GAME IS SIMPLE. TRY TO GUESS HOW CORRELATED THE TWO VARIABLES IN A SCATTER PLOT ARE. THE CLOSER YOUR GUESS IS TO THE TRUE CORRELATION, THE BETTER.

YOUR GUESS SHOULD BE BETWEEN ZERO AND ONE, WHERE ZERO IS NO CORRELATION AND ONE IS PERFECT CORRELATION. NO NEGATIVE CORRELATIONS ARE USED IN THE GAME. HERE ARE SOME EXAMPLES:

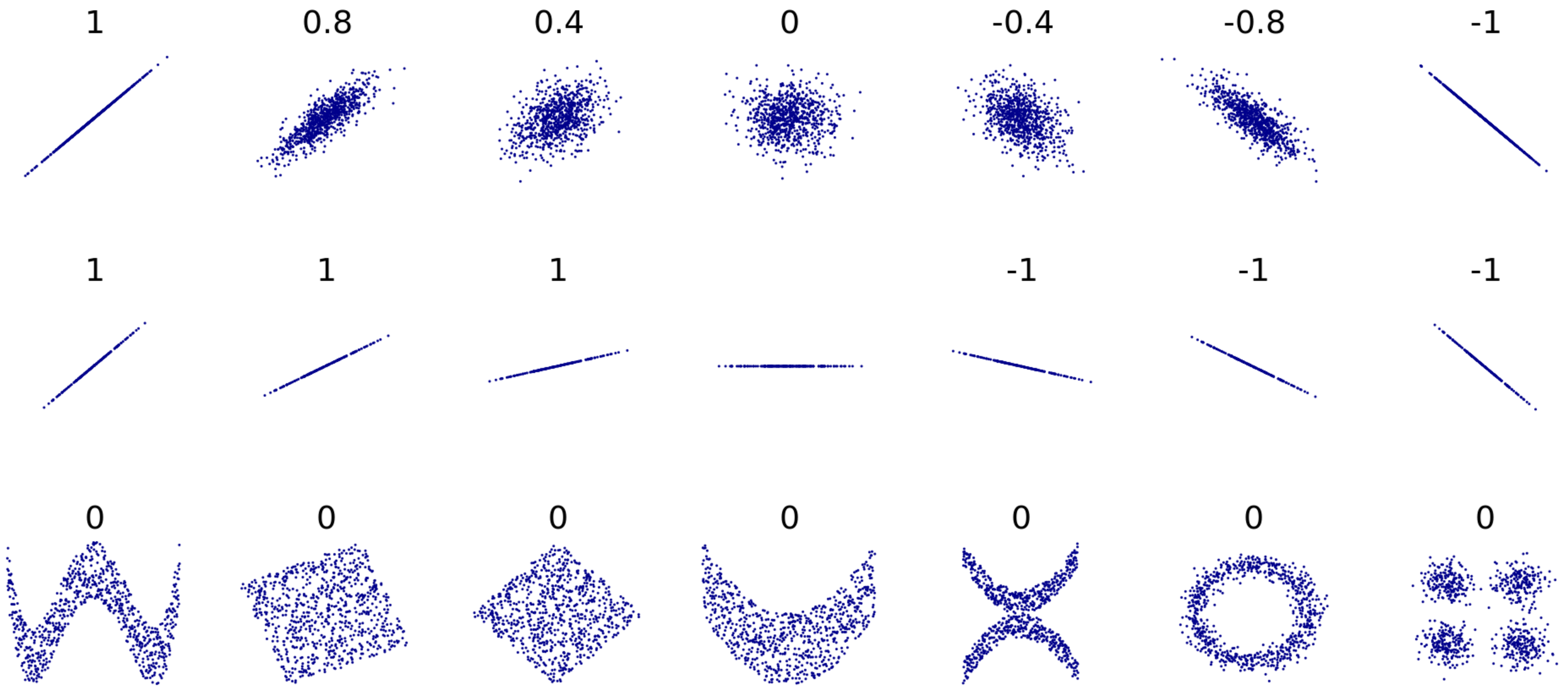
R=0.0

R=0.5



R=0.75

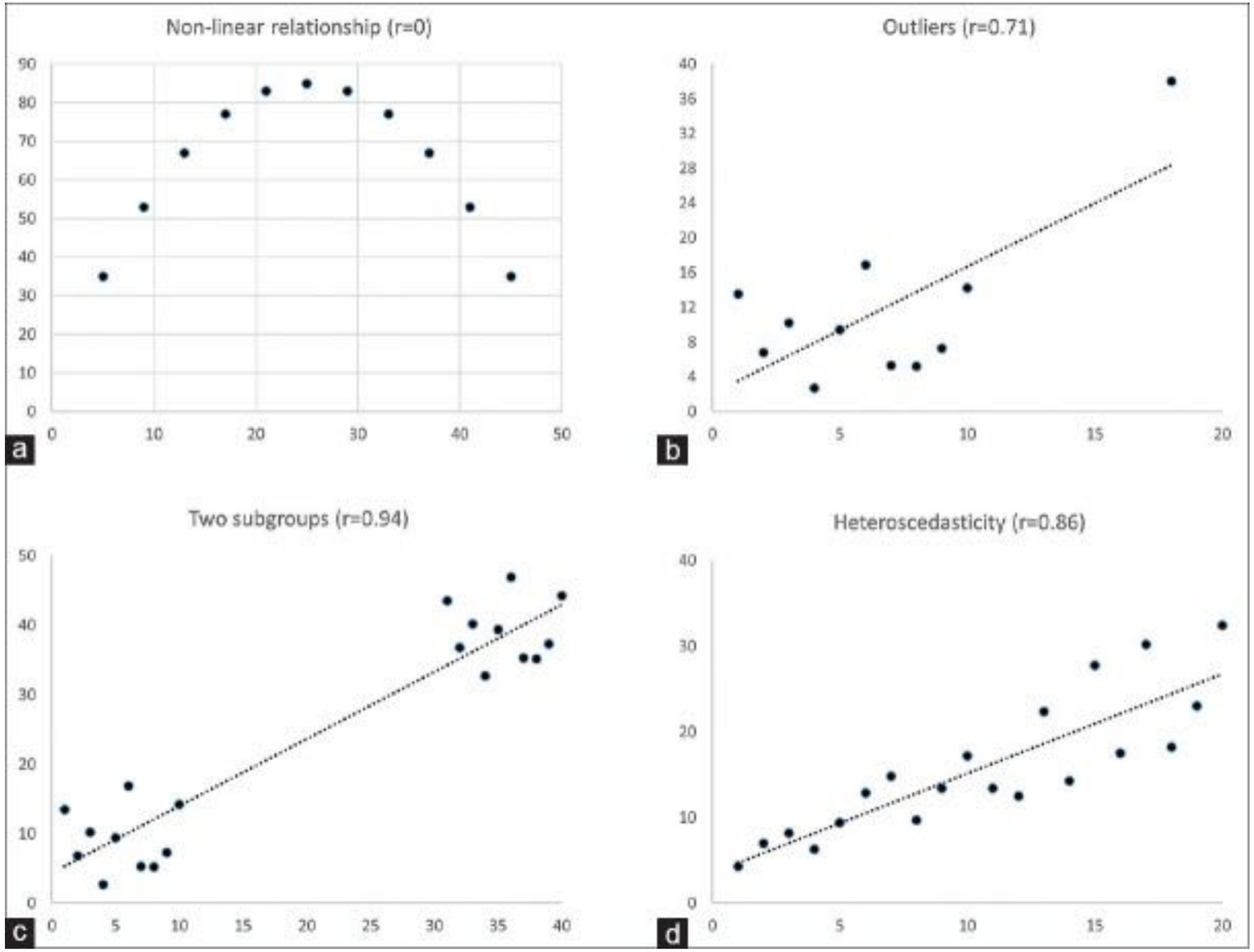
Correlation is a very incomplete description of two-variable data



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

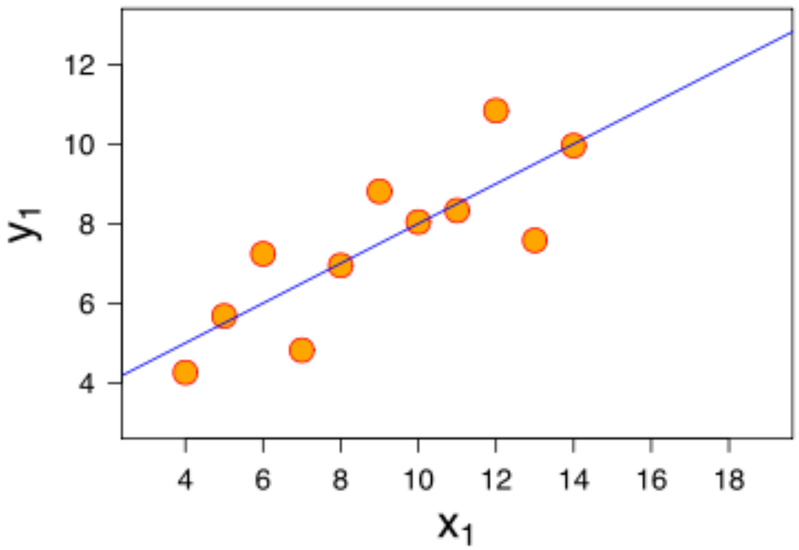
GUESS THE CORRELATION IS A **GAME WITH A PURPOSE**. THIS MEANS, WHILE IT AIMS TO BE ENTERTAINING, DATA ON THE GUESSES IS COLLECTED AND USED TO ANALYSE HOW WE PERCEIVE CORRELATIONS IN SCATTER PLOTS. SO THE MORE PEOPLE THAT PLAY, THE MORE DATA IS GENERATED!

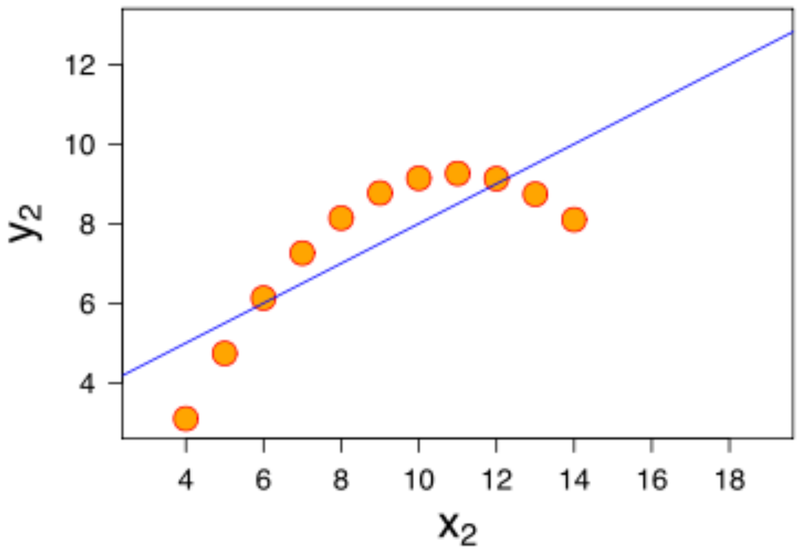
Correlation should NOT be used in many situations

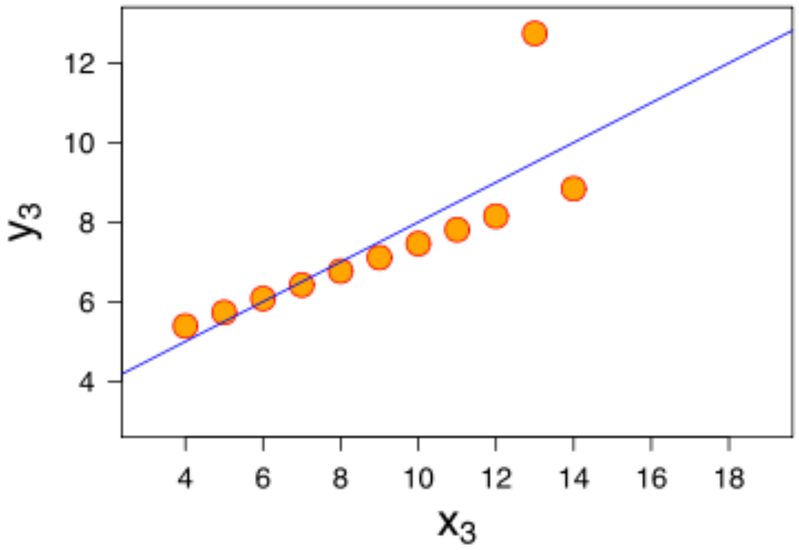


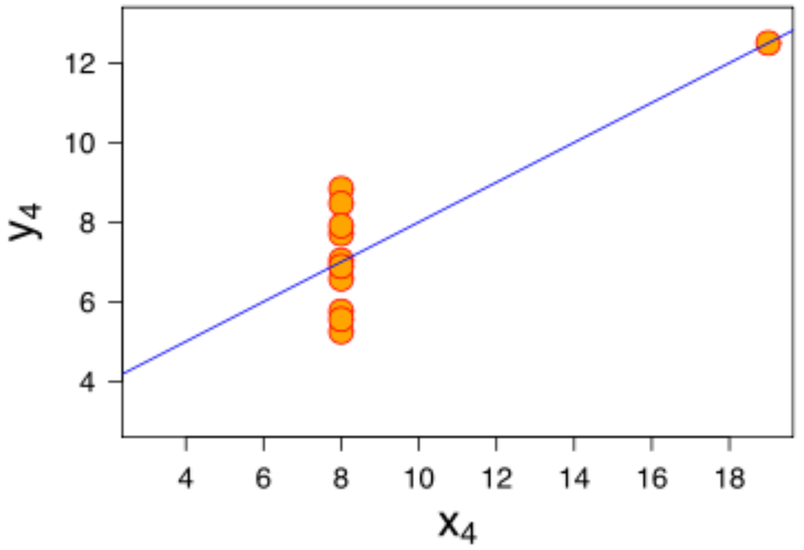
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079093/

Anscombe's quartet: Same mean, variation, correlation









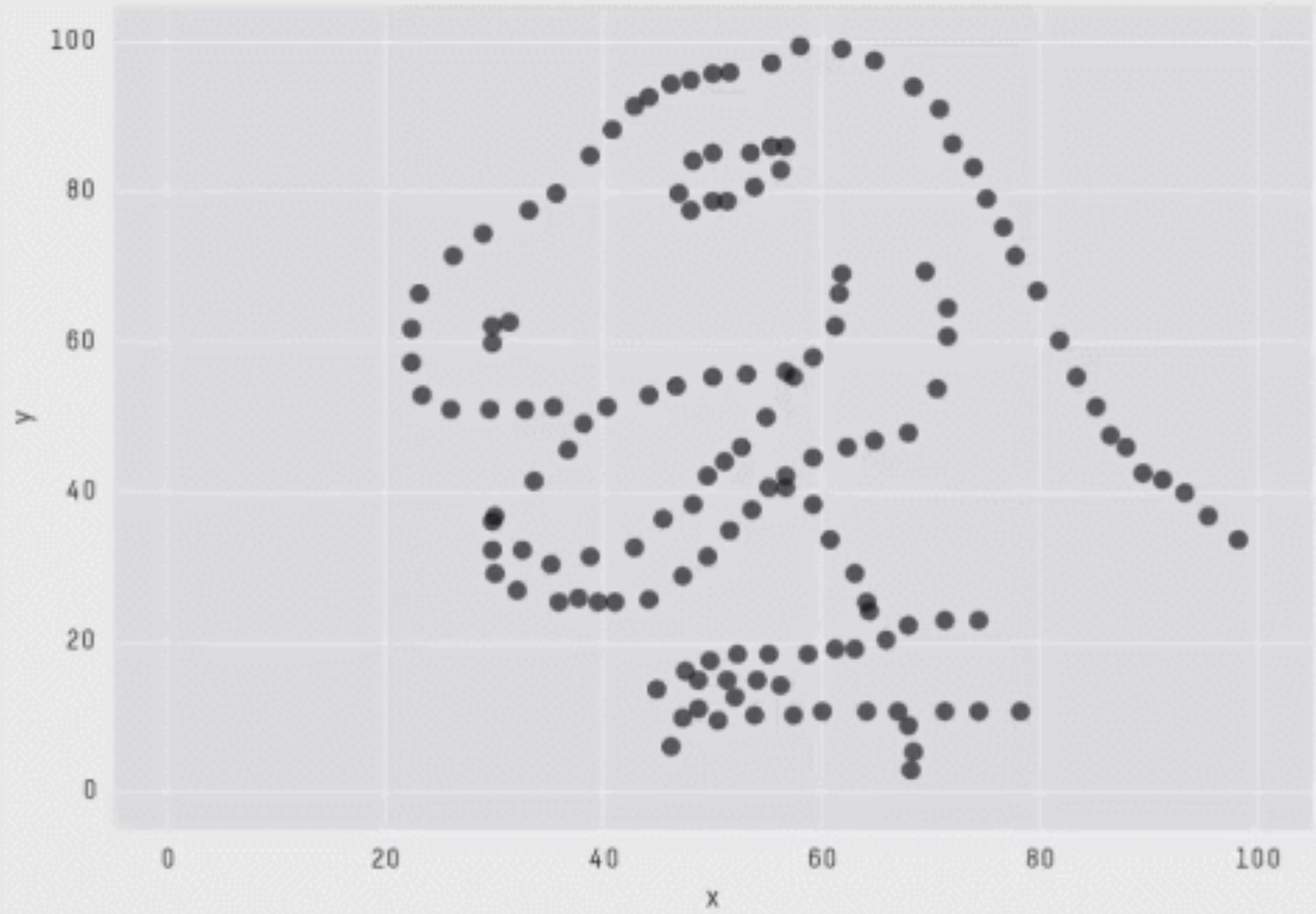
Anscombe, 1973

Anscombe's quartet: Same mean, variation, correlation

ALWAYS PLOT YOUR DATA

Anscombe, 1973

Datasaurus dozen



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

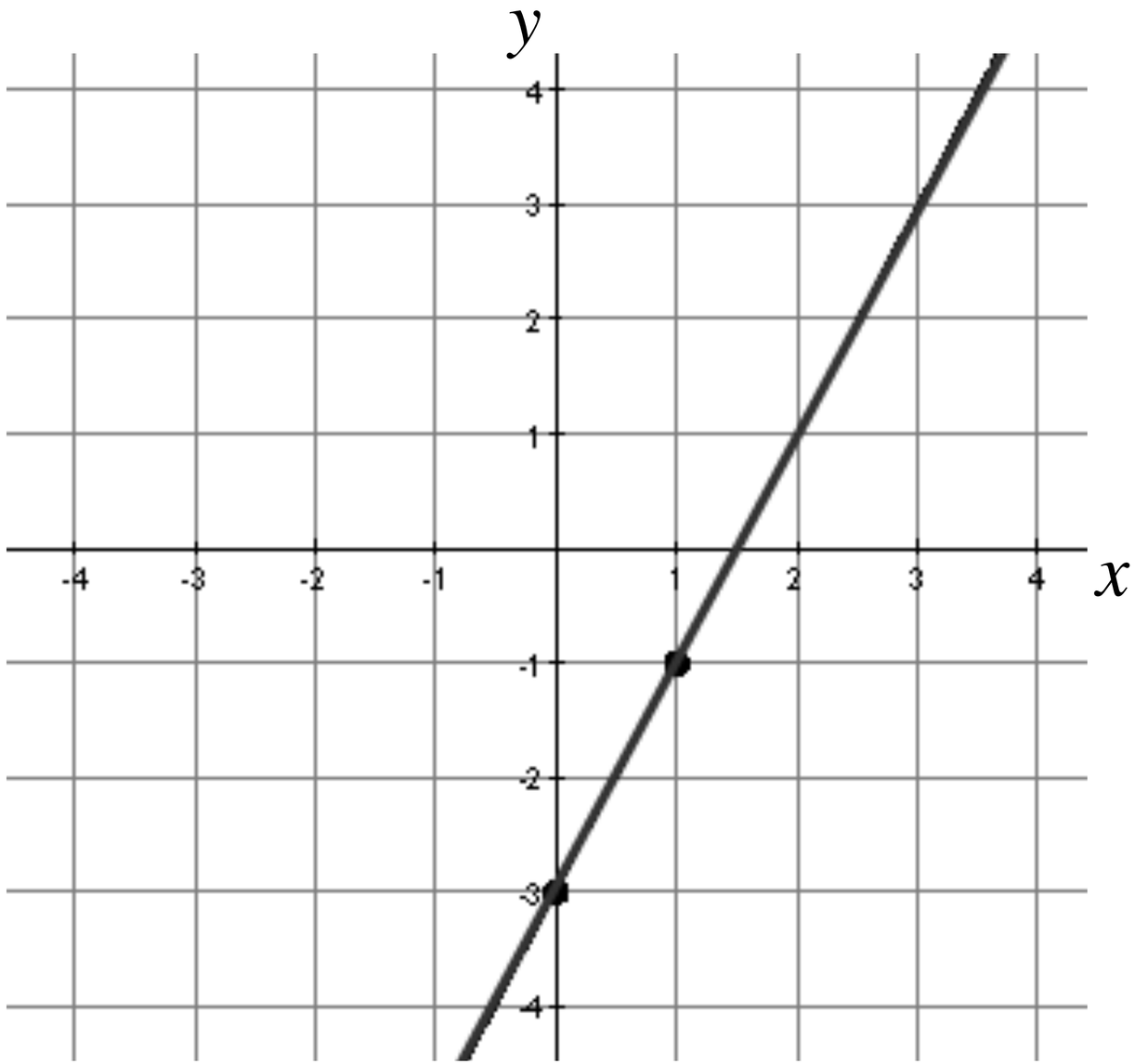
https://www.autodeskresearch.com/publications/samestats

Datasaurus dozen

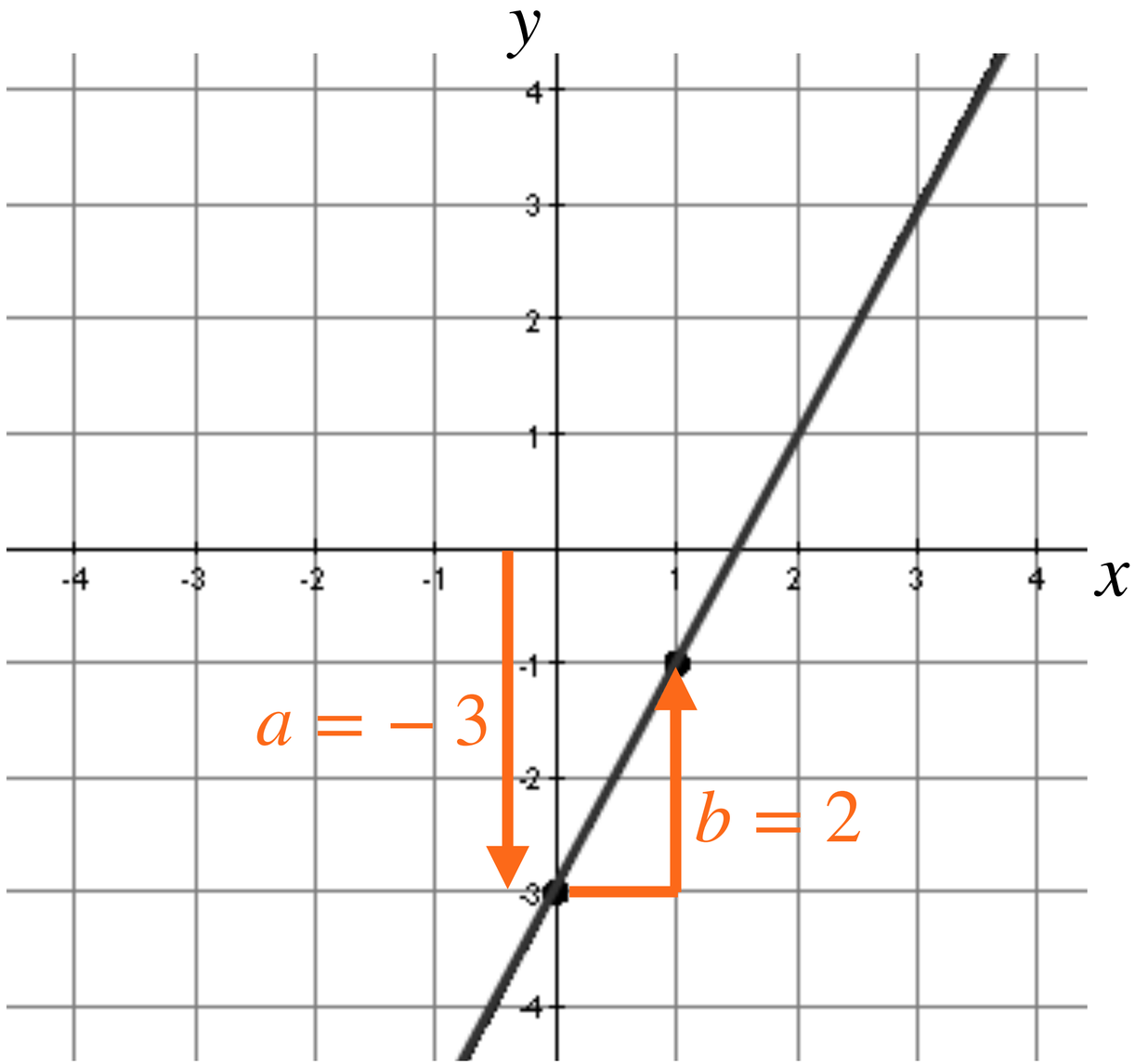
https://www.autodeskresearch.com/publications/samestats

Regression

A straight line $y = a + bx$ is determined by slope b and intercept a

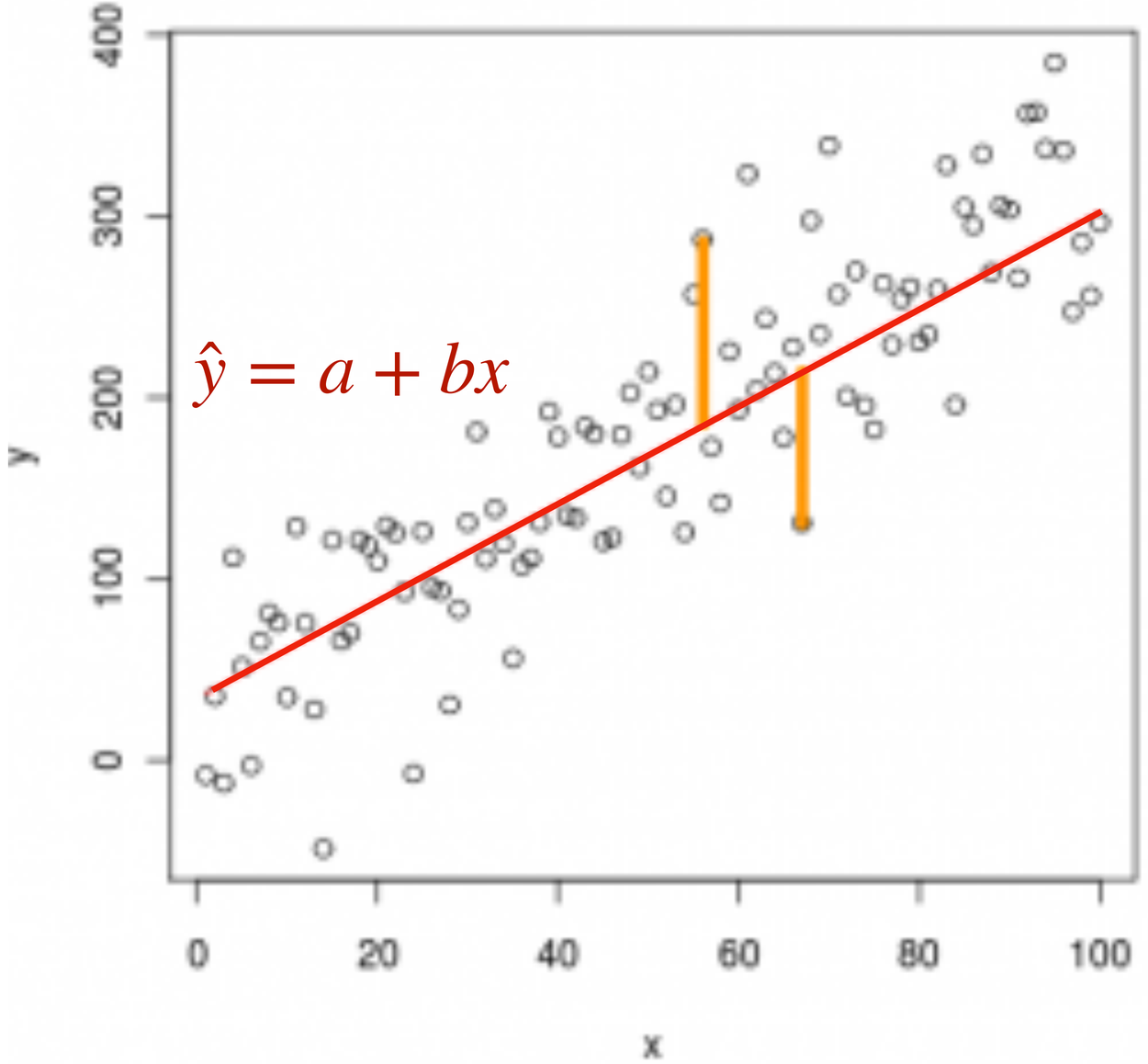


A straight line $y = a + bx$ is determined by slope b and intercept a



$$y = -3 + 2x$$

Least-squares regression fits a line $\hat{y}$ minimizing the distances along the dependent axis



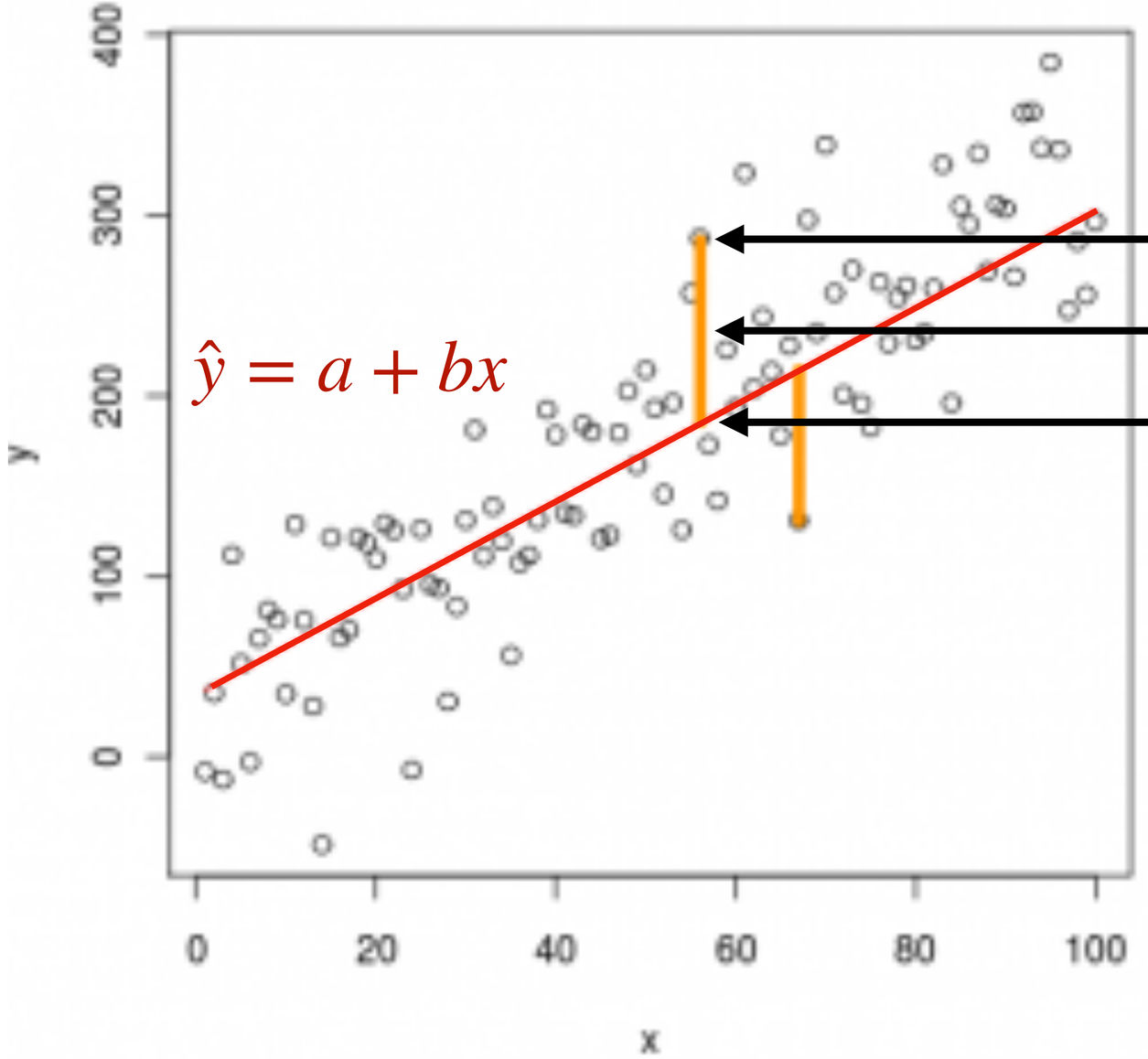
because we want to:

1) summarize the pattern

or

2) predict y from x

Least-squares regression makes the sum of squares of distances as small as possible

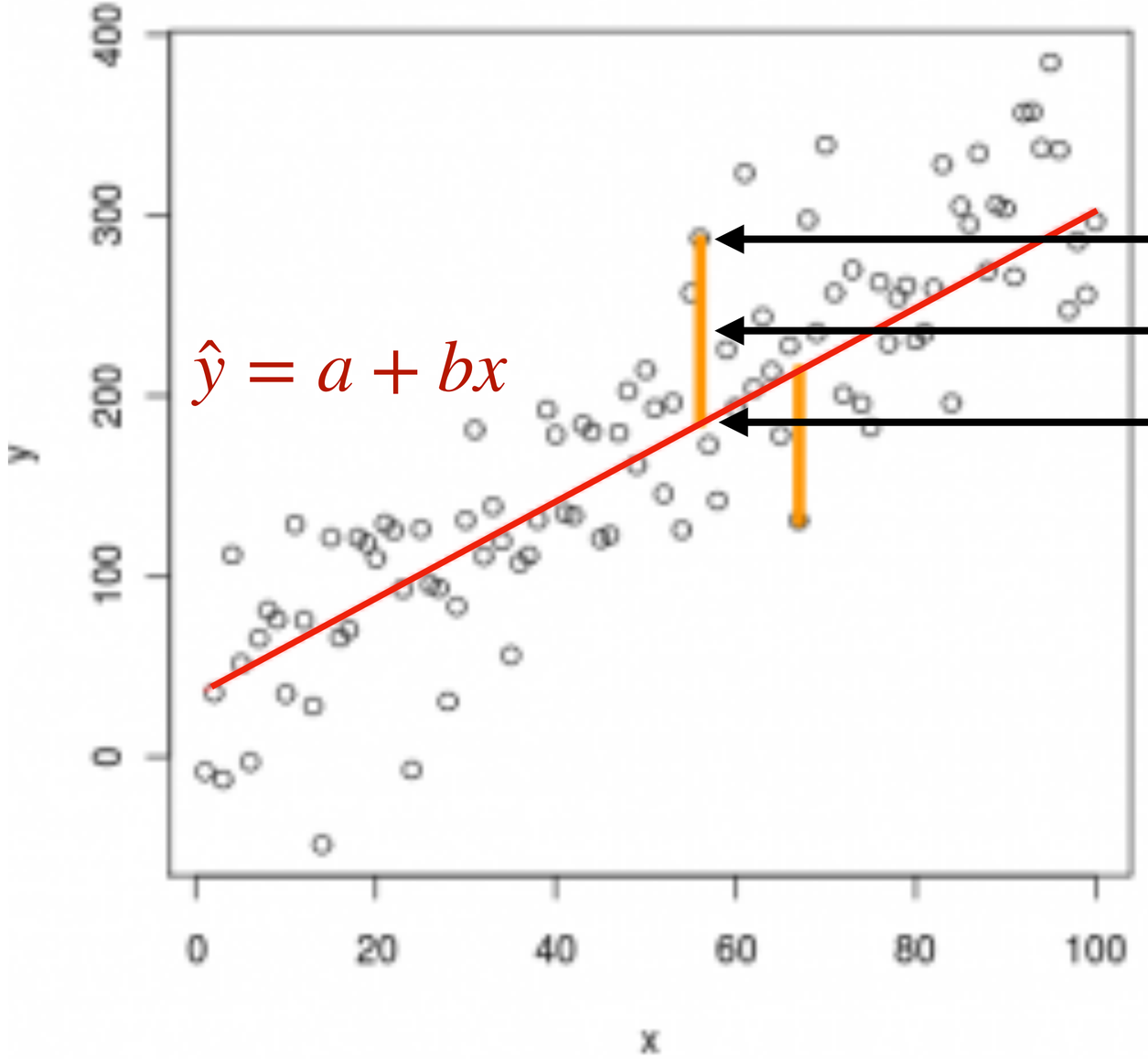


observed $y = 170$

error $= y - \hat{y}$

predicted $\hat{y} = a + b \cdot 58$

Least-squares regression makes the sum of squares of distances as small as possible



observed $y = 170$

error $= y - \hat{y}$

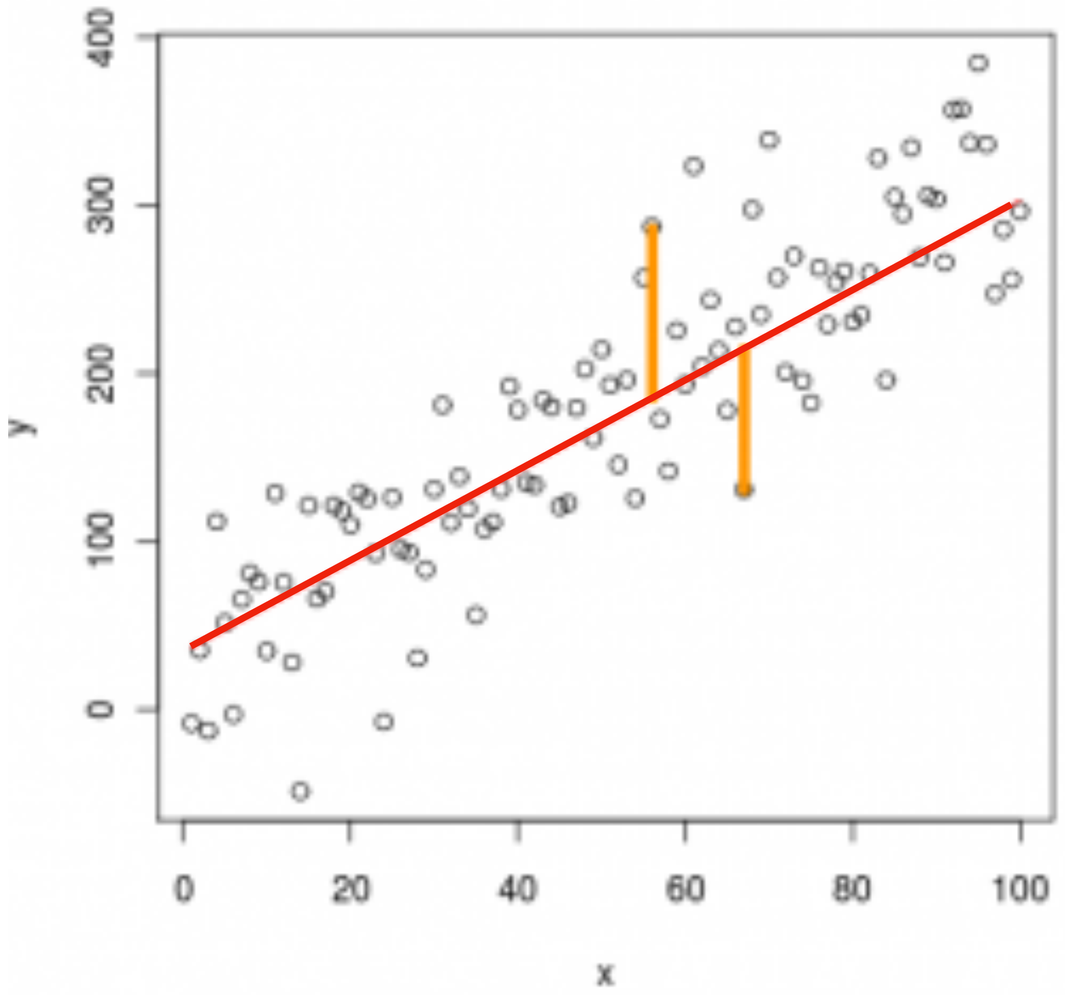
predicted $\hat{y} = a + b \cdot 58$

We need to find a and b such that

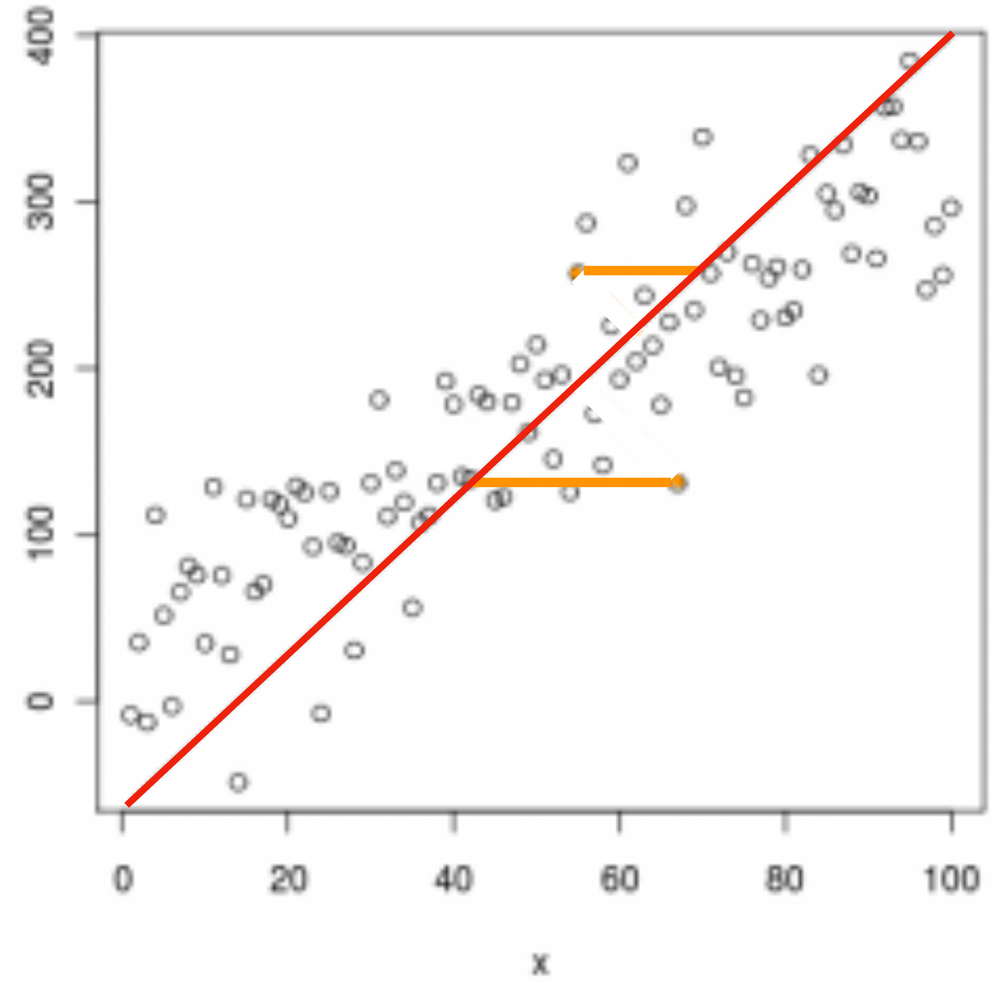
$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

is minimal.

Least-squares regression is not symmetric



≠



but correlation is symmetric

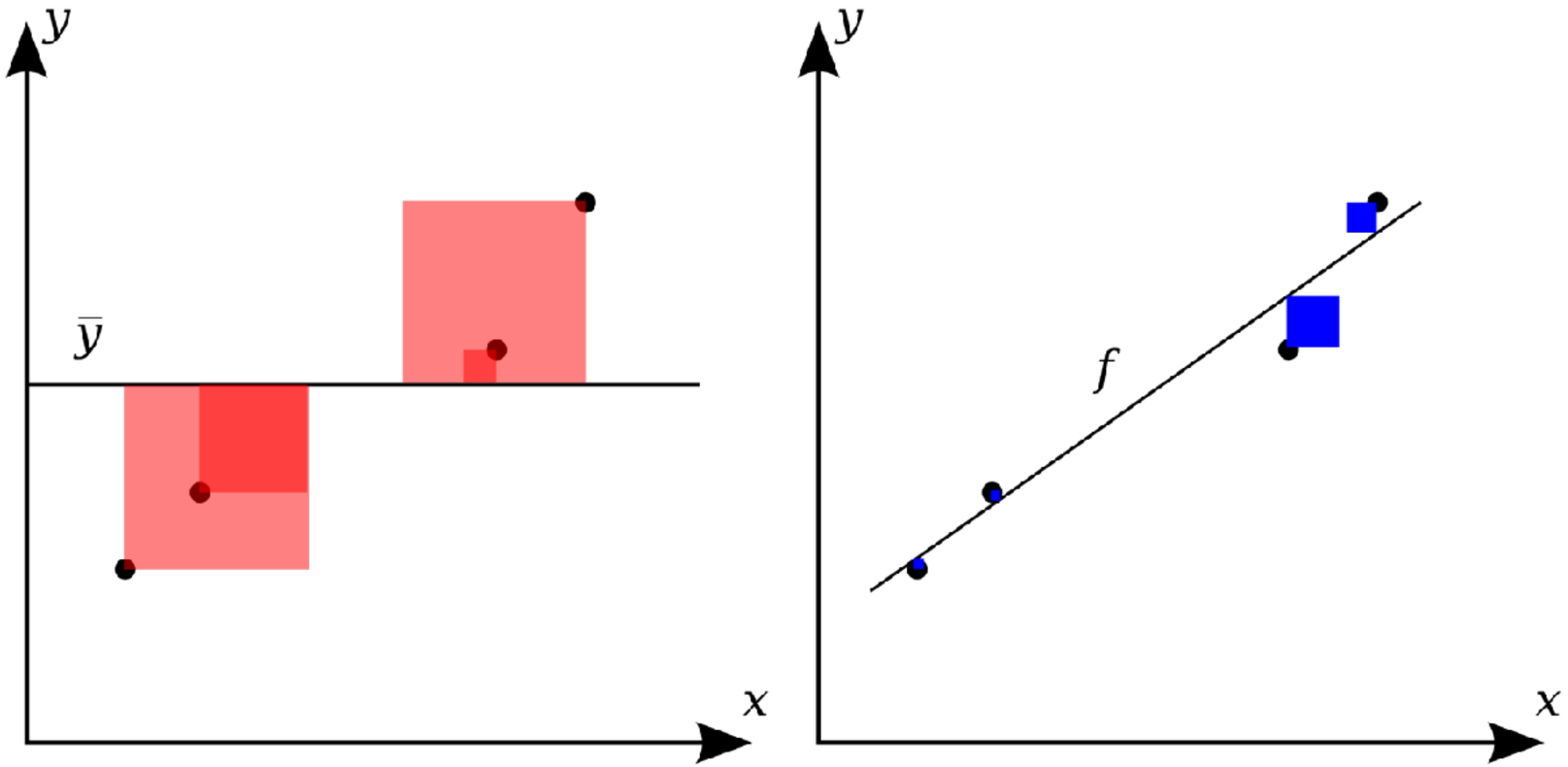
The connection between correlation and regression is r^2

The square of correlation r^2 is the fraction of the variation in the values of y that is explained by least-squares regression of y on x.

The connection between correlation and regression is r^2

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2 \qquad SS_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2$$

$$r^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$



https://en.wikipedia.org/wiki/Coefficient_of_determination

The connection between correlation and regression is r^2

When doing linear regression, it is common to report r^2 to give some information about the goodness of fit.

r^2 closer to 1 means better fit

Causation


TIME
https://www.youtube.com/watch?v=B85VBIeUCbU

A lurking variable is a variable additional to explanatory and response that may influence the interpretation of their relationship

Playing violent computer games

Spree killer

A lurking variable is a variable additional to explanatory and response that may influence the interpretation of their relationship

Male, 20-30

Playing violent computer games

?

spurious correlation

Spree killer

A lurking variable is a variable additional to explanatory and response that may influence the interpretation of their relationship

Male, 20-30

?

?

spurious correlation

Playing violent computer games

Spree killer

Reverse causation means that between two associated variables X and Y, the causal direction is opposite of what was expected

Playing violent computer games

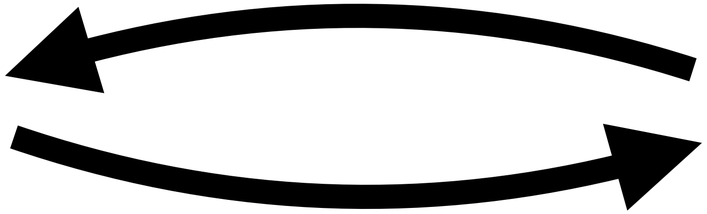
?

reverse causation

Spree killer

Reality is often more complicated:
When both X causes Y and Y causes X, we have Simultaneity

Depression



?

Simultaneity

Smoking

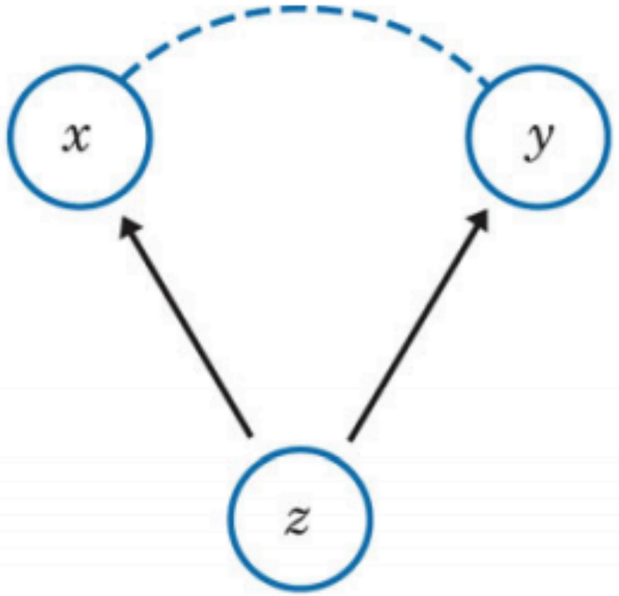
Association can come from causation

Causation

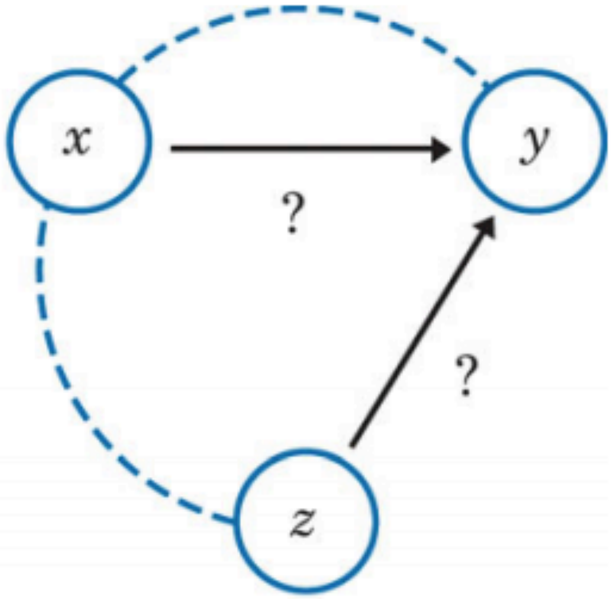
Association can come from causation but also from lurking variables

Causation

Common response



Confounding

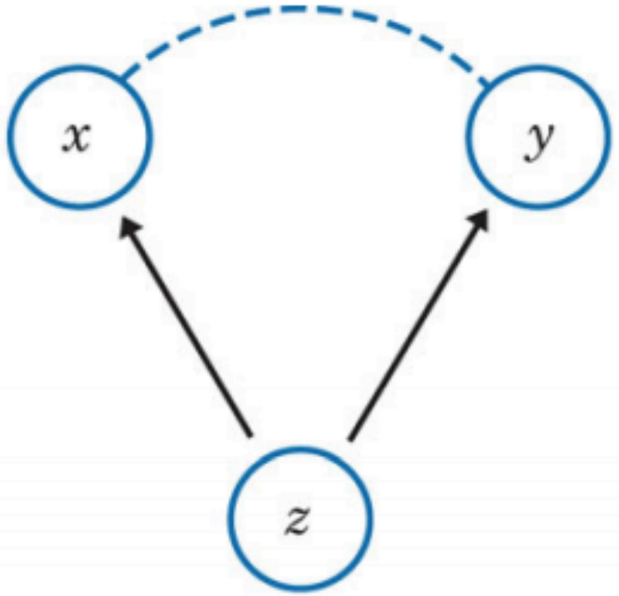


Association can come from causation but also from lurking variables

Causation

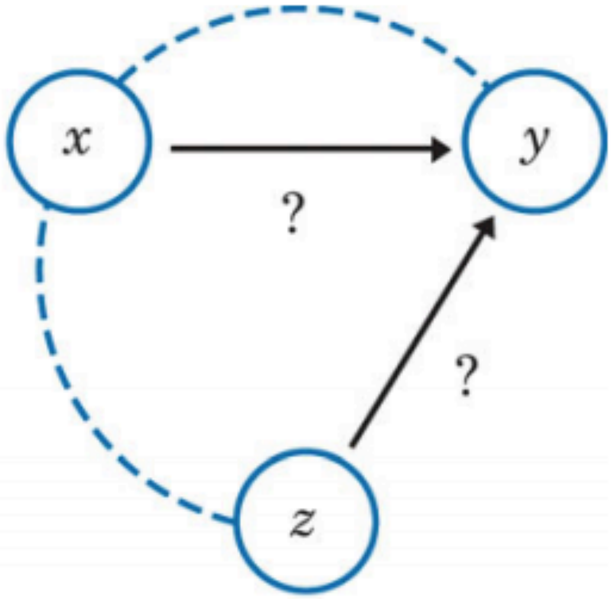
Change in x
causes change in y

Common response



Both x and y respond to
changes in some other variable

Confounding



The effect of x on y cannot
be distinguished from the
effect of other variables on y

High school grades can predict university success, but do not cause it

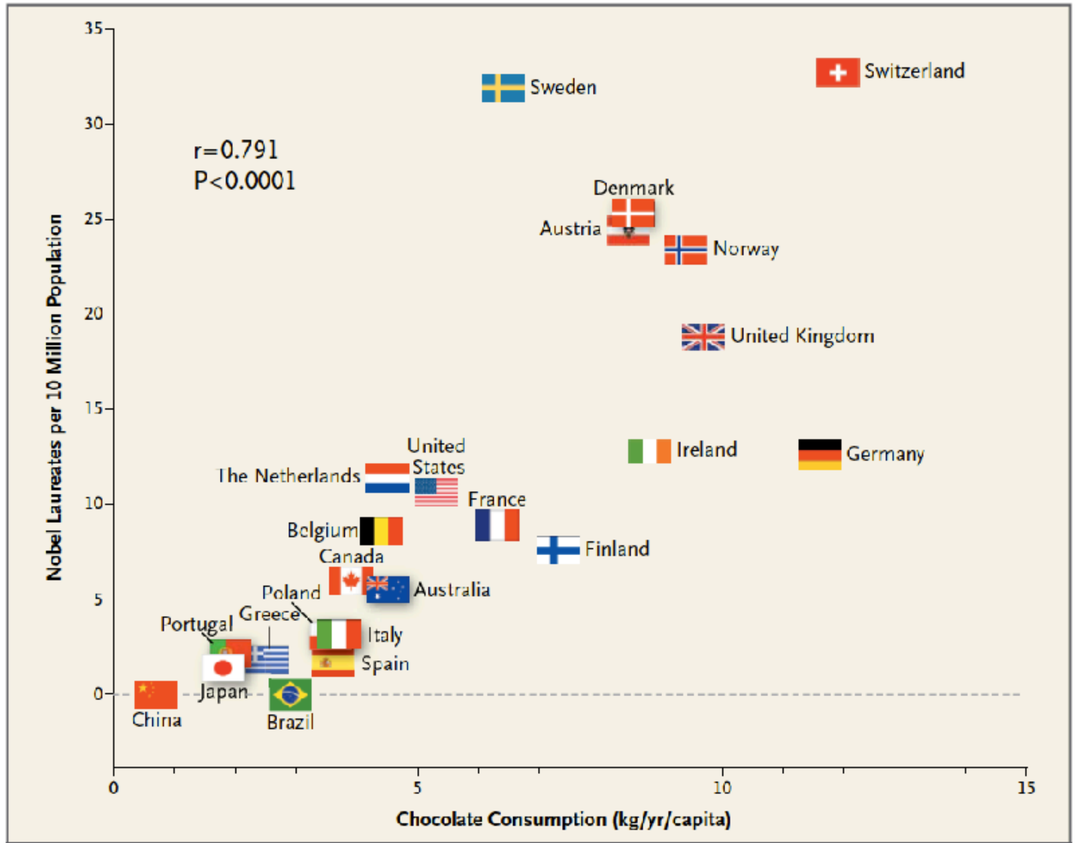


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

TIME

Secret to Winning a Nobel Prize? Eat More Chocolate

By Olivia B. Waxman @OBWax | Oct. 12, 2012

BBC | Sign in | News | Sport | Weather | Shop | Ear

NEWS

Home | Video | World | UK | Business | Tech | Science | Magazine | En

Magazine

Does chocolate make you clever?

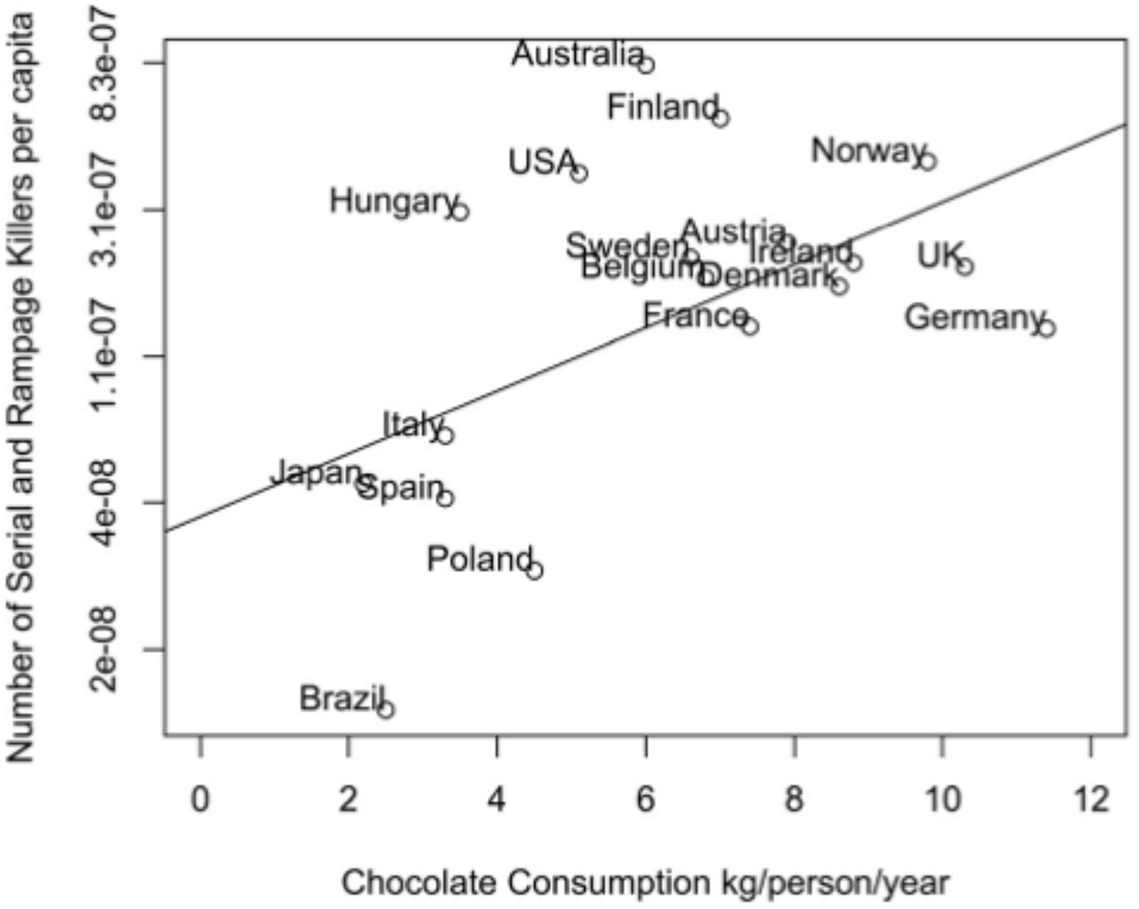
By Charlotte Pritchard

BUSINESS INSIDER

SCIENCE

There's A Shocking Connection Between Eating More Chocolate And Winning The Nobel Prize

Joe Weisenthal

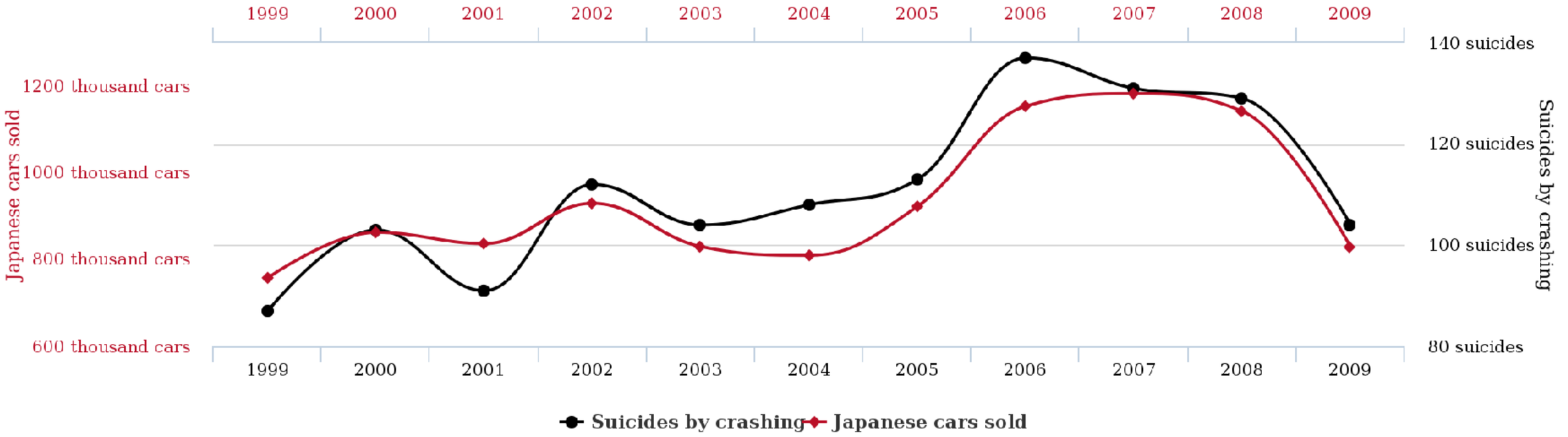

Number of Serial and Rampage Killers per capita
2e-08
4e-08
1.1e-07
3.1e-07
8.3e-07
Australia
Finland
Norway
USA
Hungary
Austria
Sweden
Ireland
UK
Belgium
Denmark
France
Germany
Italy
Japan
Spain
Poland
Brazil
0
2
4
6
8
10
12
Chocolate Consumption kg/person/year


lindt

Japanese passenger cars sold in the US

correlates with

Suicides by crashing of motor vehicle

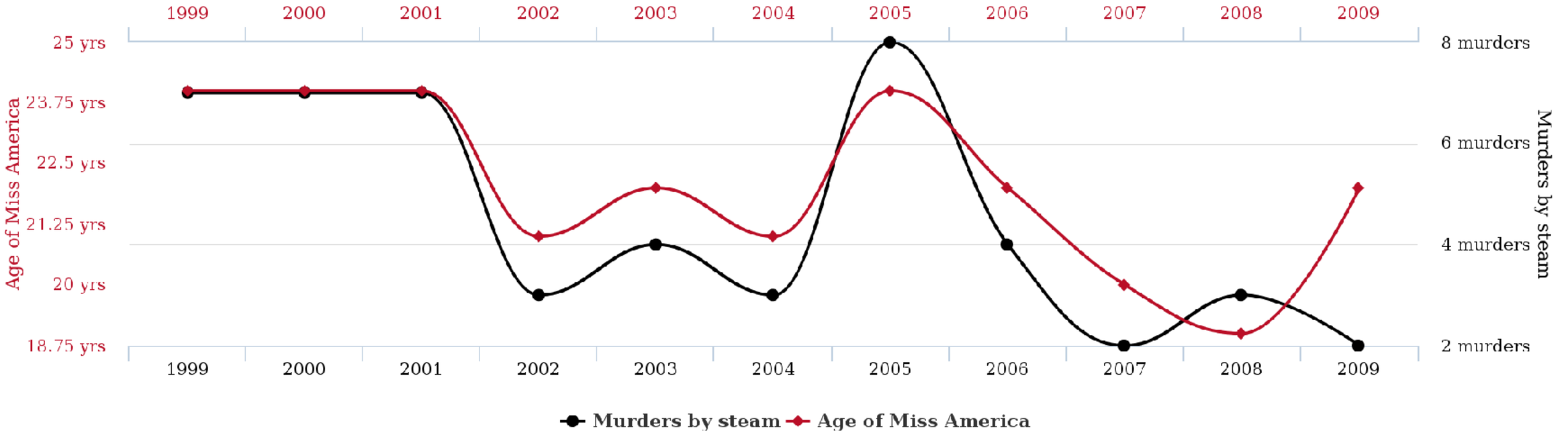


tylervigen.com

Age of Miss America

correlates with

Murders by steam, hot vapours and hot objects



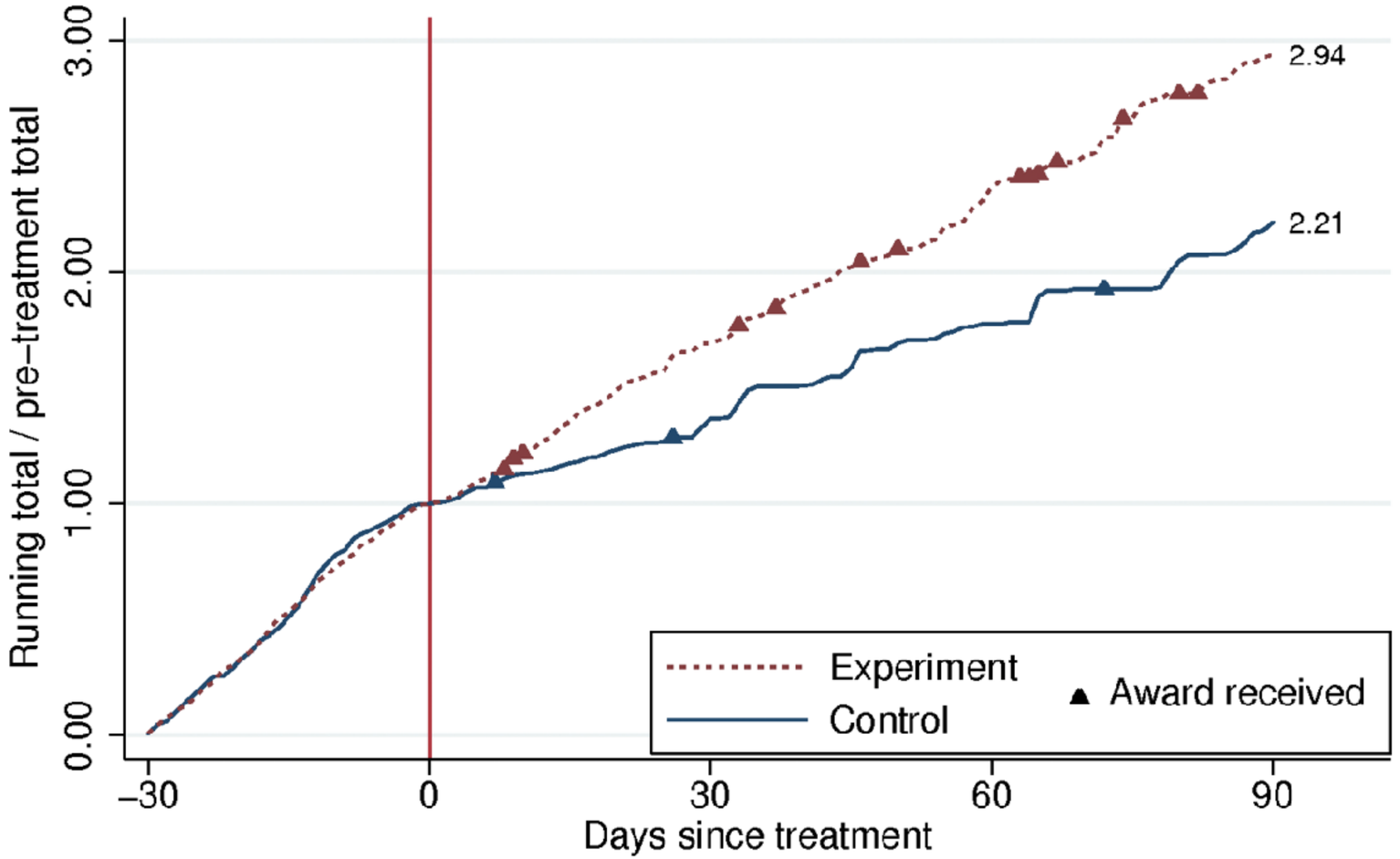
tylervigen.com

Correlation does not (necessarily) imply causation

Correlation can be a good starting point.

In science we want to understand the mechanisms that lead to the correlation

Experiments establish best whether an association is a causation because other influences can be controlled



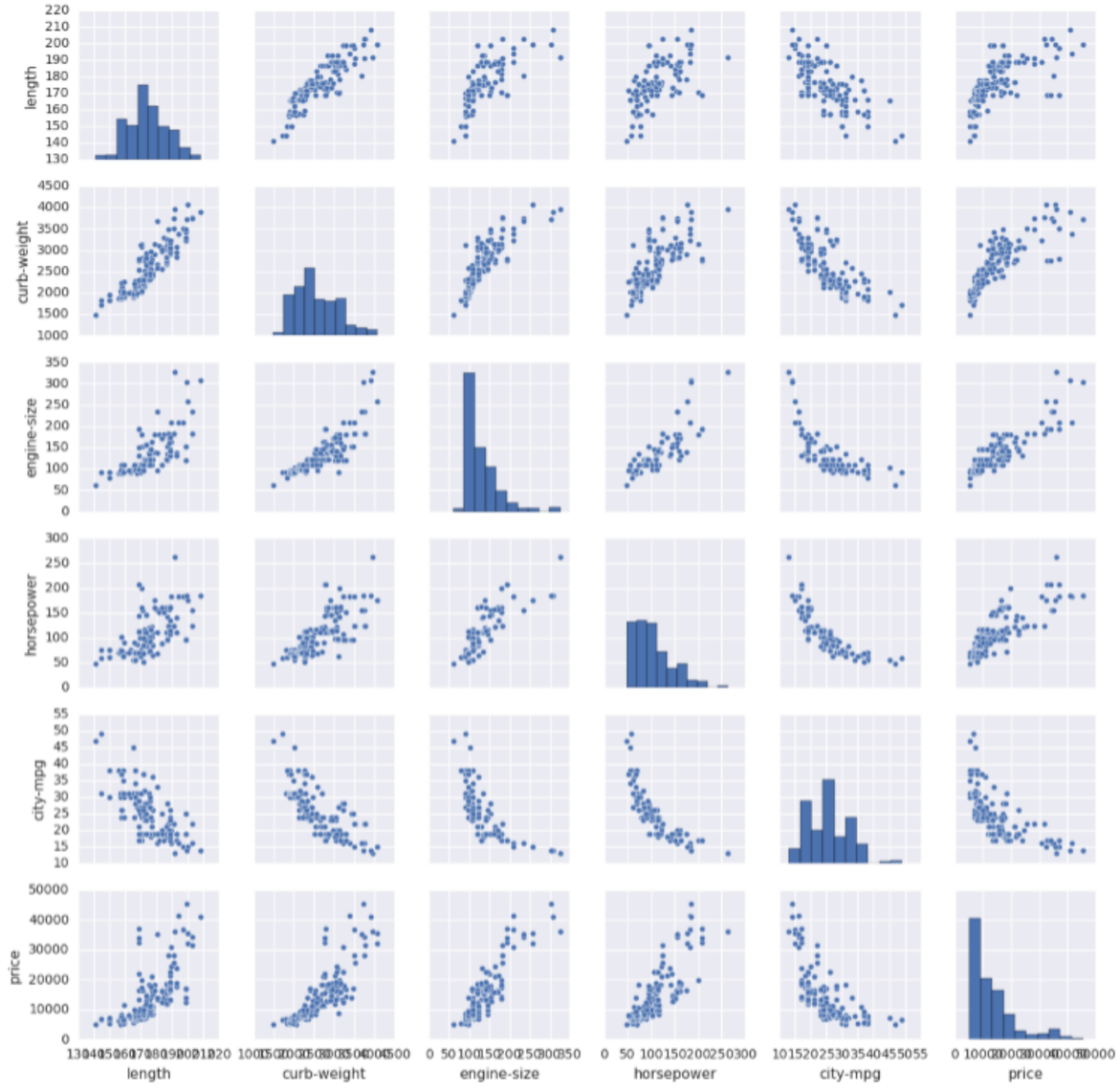
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0034358

If we cannot run experiments, we cannot be 100% sure we see causation, but there are good criteria

If we cannot run experiments, we cannot be 100% sure we see causation, but there are good criteria

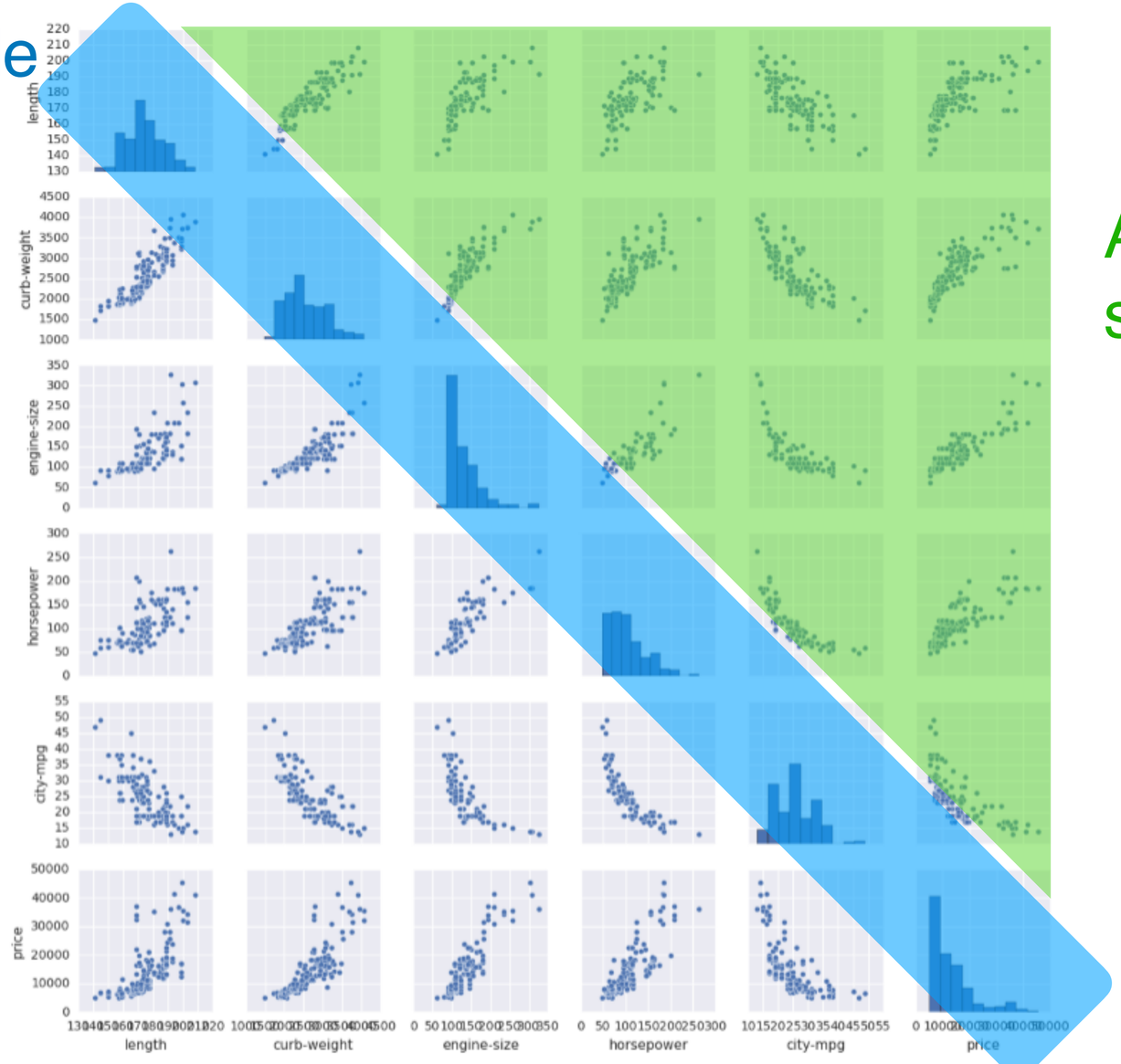
- Strong association
- Consistent association
- Association between dose and response
- Time consistency
- Plausibility

The scatterplot matrix gives you a summary plot of your whole data set



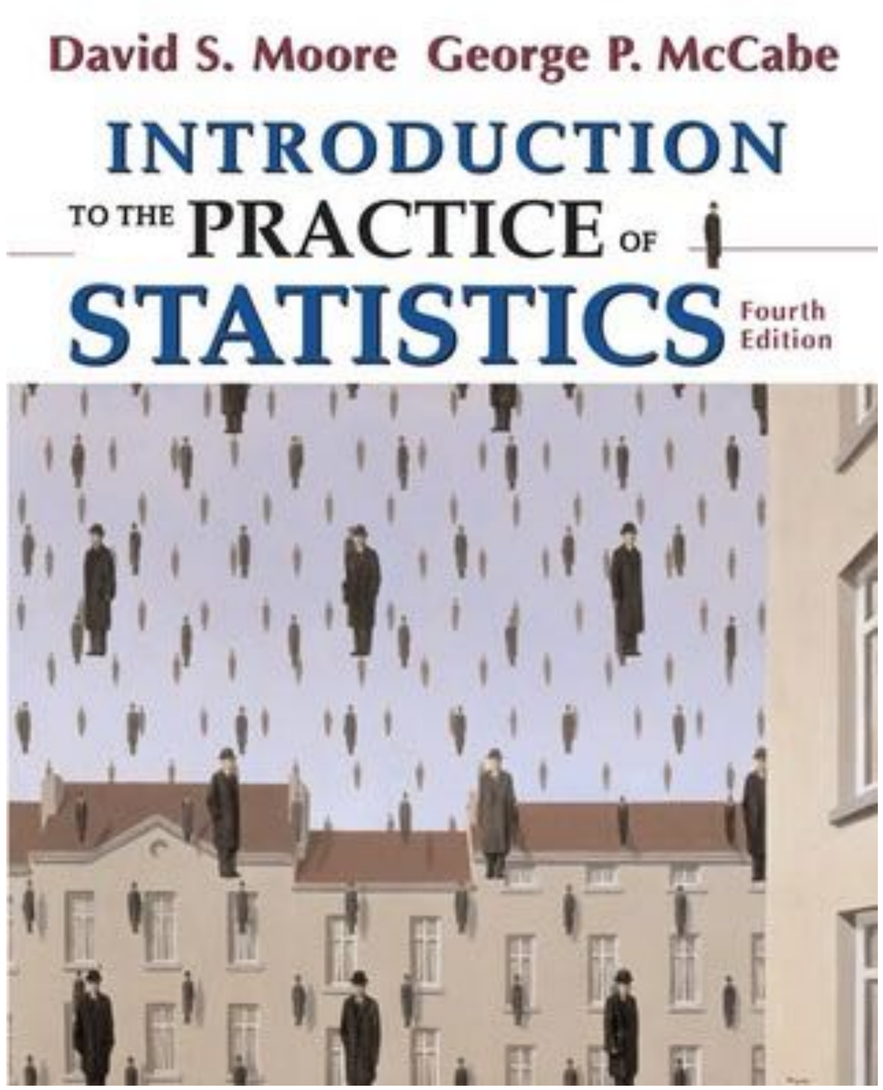
The scatterplot matrix gives you a summary plot of your whole data set

All single-variable distributions



All bivariate scatter plots

Sources and further materials for today's class



Chapter 2

Jupyter