Advanced Data Science and Fallacies

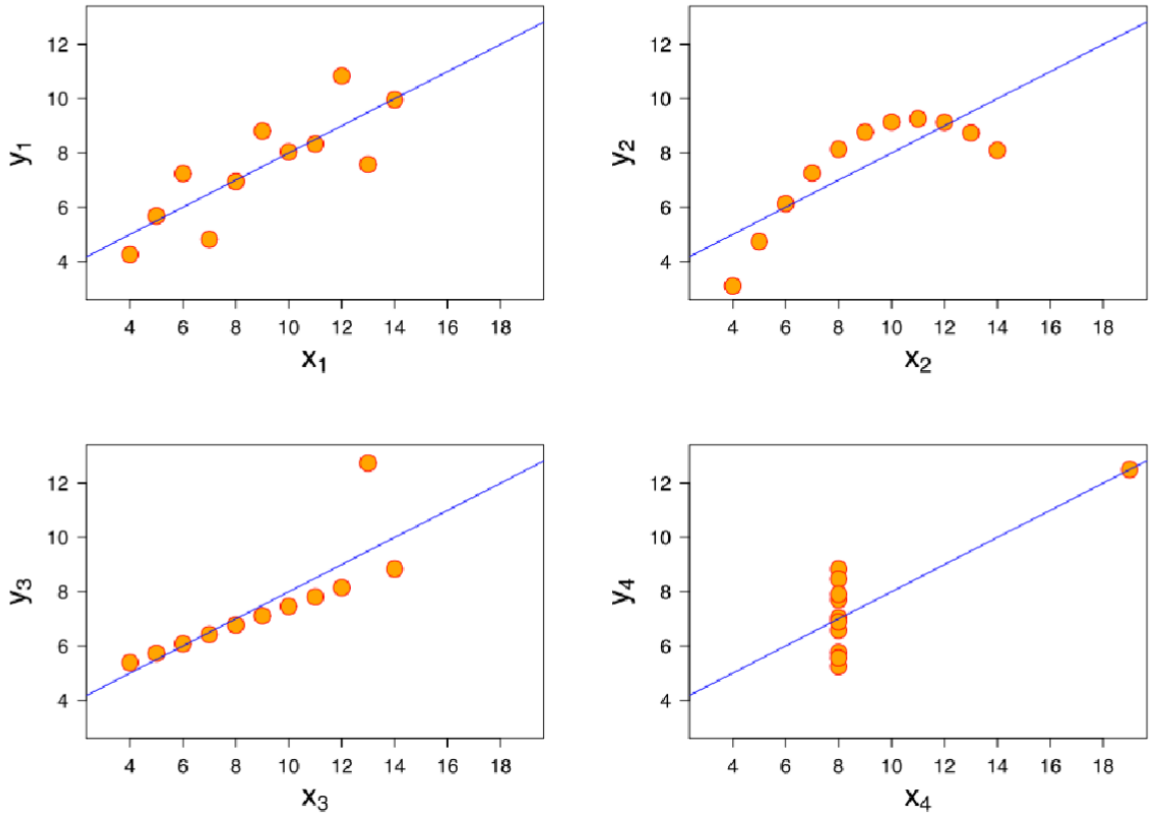
MIE223

Winter 2025

1 Potential Data Interpretation Fallacies

1.1 Anscombe Quartet: Visualize

When we look at mean of x, sample variance of x, mean of y, sample variance of y, correlation between x and y, and linear regression line, the accuracy differs.
If datasets have varying levels of accuracy, the mean, variance, and correlation will be the same. However, the linear regression line will differ.

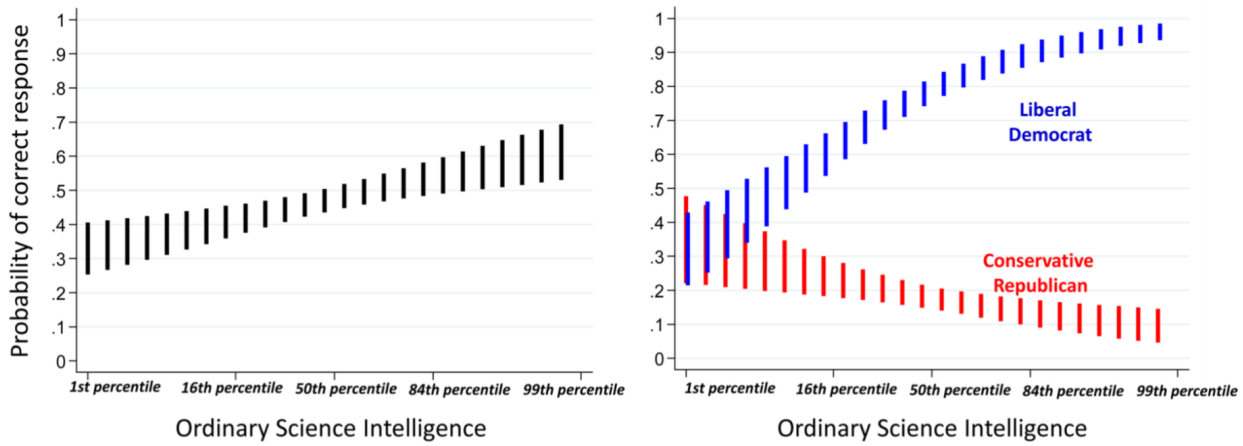


1.2 Spurious Correlation

Ratios induce spurious correlations. Solution: compositional data analysis.
x, y, and z are all statistically independent. The ratios induce high correlation between them.
Normalizing by a common number (such as in ratios) results in spurious correlation.

1.3 Confounding Variables

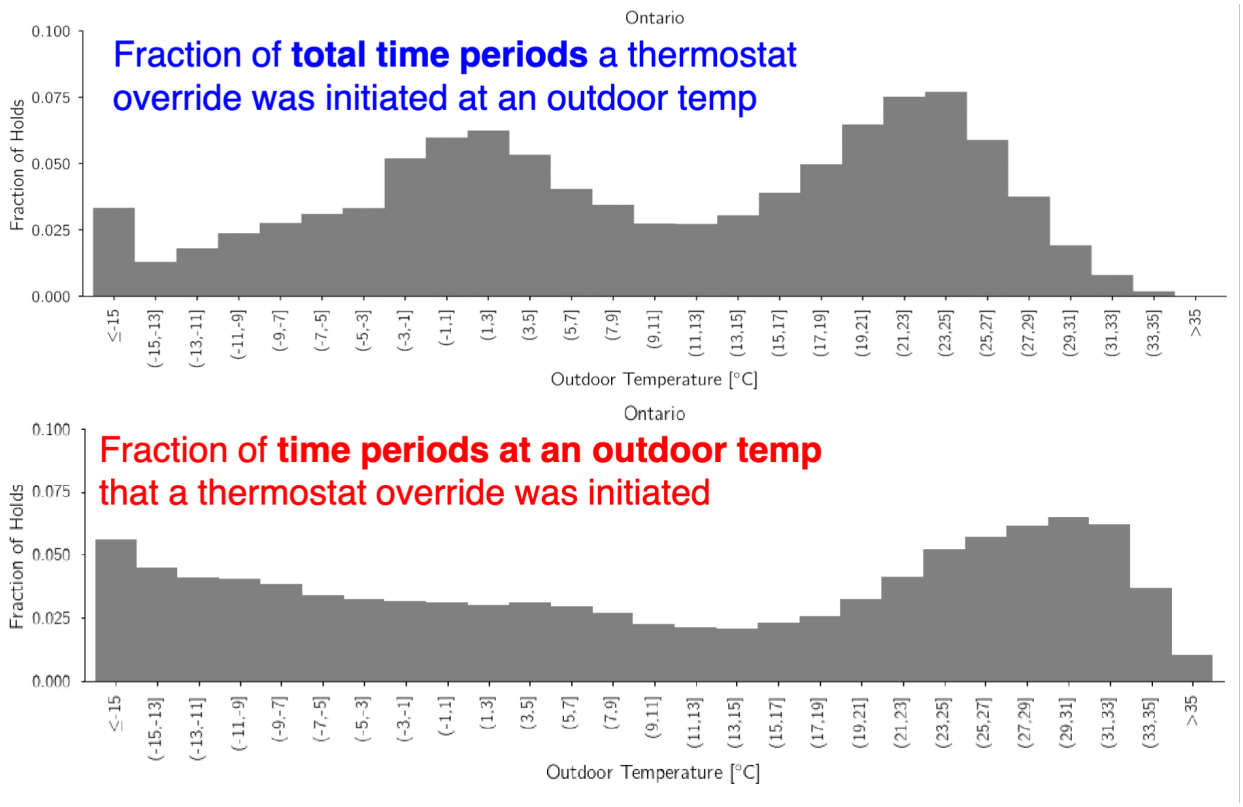
- Beware of (latent) confounding variables
  - Averaging over them can hide important trends



– How to find confounders? Search.
– What if latent? Learn mixture models (EM), deep learning, etc.

Political affiliation is a confounder. Belief in a scientific statement is conditioned on political affiliation.

1.4 Correctly Normalizing



blue

- peaks between 0 and 1 and 23 and 25

- Not at certain temperatures with equivalent probability
- Looking at a joint measurement not a conditional measurement

red

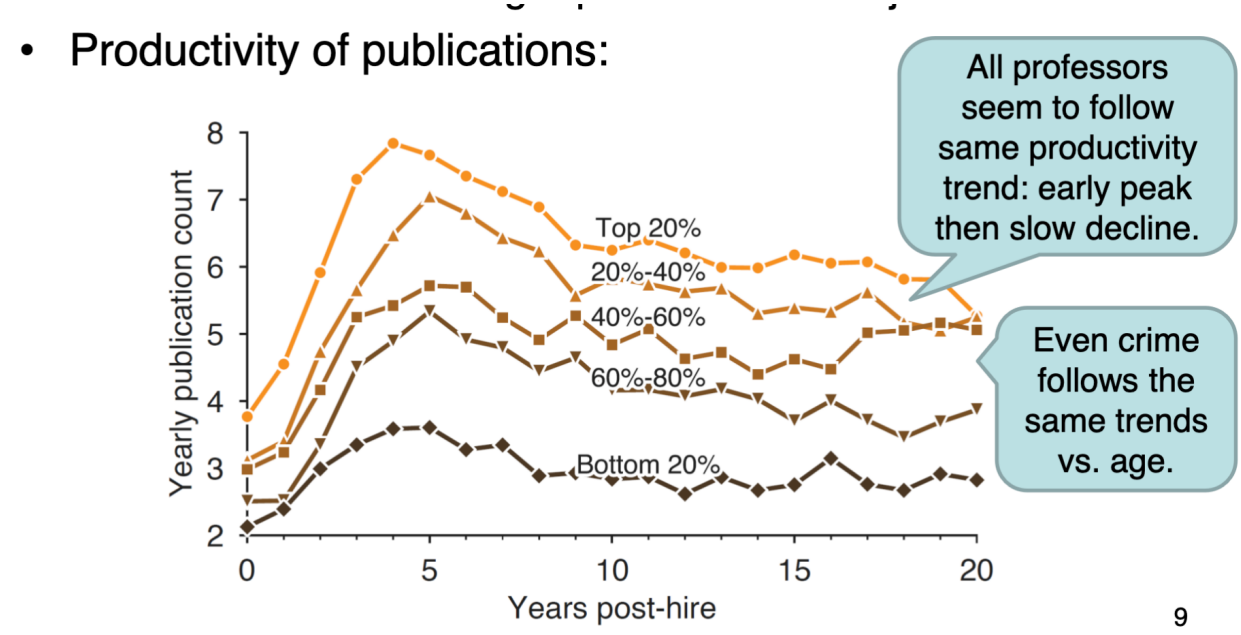
- Normalized by the number of days in each temperature range

Fallacies

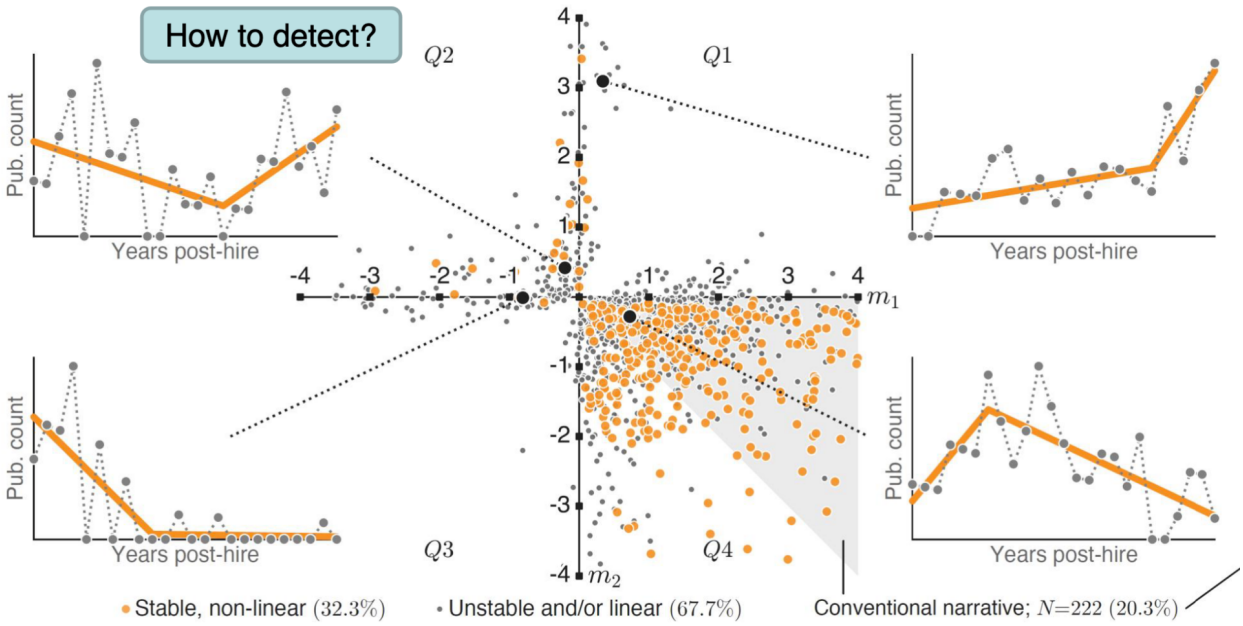
- Don't look at a visualization before you make a hypothesis

1.5 Myth of the Average

- High rate of plane crashes in the 40s
  - Seats built for "average" pilot were non-adjustable!
- Productivity of publications:
- We've averaged out the interesting behaviour of the data



If plot early and later career productivity, a more complex picture of productivity types emerges...



x axis m1, y axis m2. m1 is positive, m2 is negative.
We didn't see this many different behaviours because of averaging the data.
There are smaller subsets of trends that get shadowed by the majority trend

1.6 Simpson's Paradox

Possibly the most problematic issues in data analysis – unobserved variables and/or sample bias can completely change your interpretation and decision!

	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

How do resolve this? – Which answer is correct depends on more knowledge