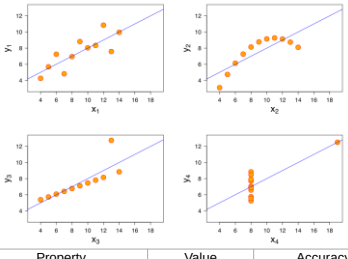
Advanced Data Science

Scott Sanner

Potential Data Interpretation Fallacies

Anscombe Quartet: Visualize!

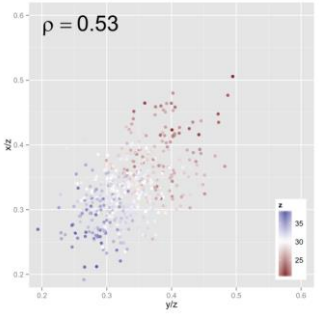


Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	plus/minus 0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively

Spurious Correlation

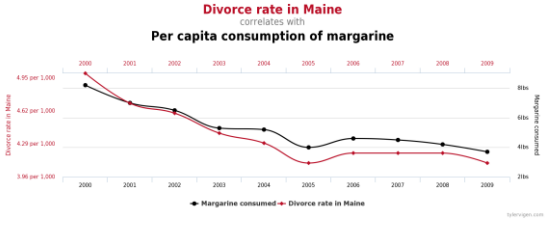
- x,y ~ N(10,1)
- z ~ N(30,9)
- Sample correlation ρ of 0.53 (high)
- Even though x,y,z, are all statistically independent!
  - Problem here: ratios induce correlations
- Solution: compositional data analysis

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



Beware: Spurious Correlations

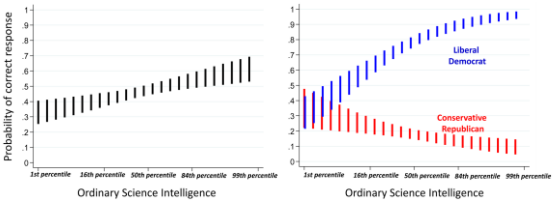
Beware: Spurious Correlations



Confounding Variables

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

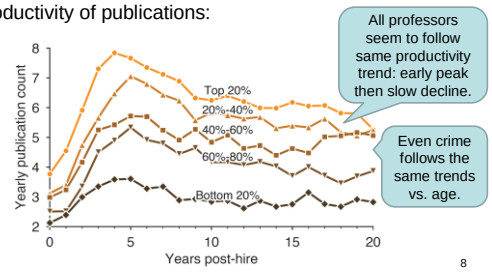
Responses to "there is evidence of human-caused climate change"



  - How to find confounders? Search.
  - What if latent? Learn mixture models, deep learning, etc.

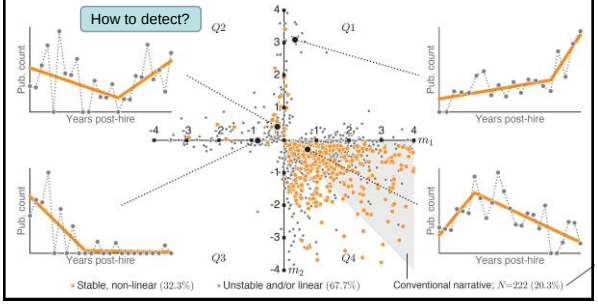
Myth of the Average

- High rate of plane crashes in the 40s
  - Seats built for "average" pilot were non-adjustable!
- Productivity of publications:



Myth of the Average (Cont.)

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



Simpson's Paradox

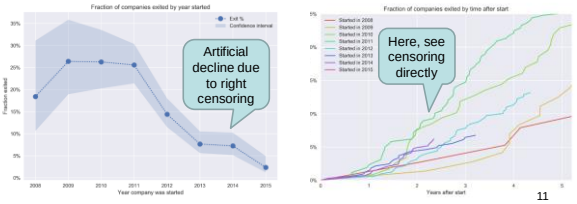
- Possibly the most problematic issues in data analysis – unobserved variables and/or sample bias can completely change your interpretation **and decision**!

	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

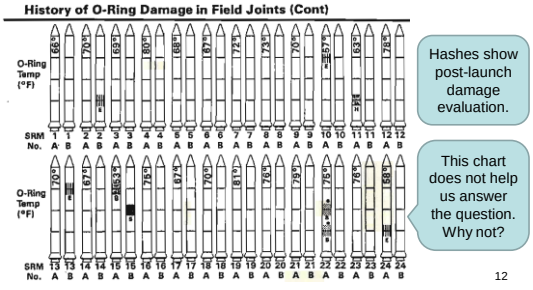
Plotting "Right Censored" Time Series Data

- Be aware of horizon effects with time
  - "right censored": event may not have happened in all samples
    - Or data may be incomplete (people left study, lost track, etc.)
  - Especially prominent in "survival/failure analysis"
    - Many ways to handle: https://en.wikipedia.org/wiki/Survival_analysis
  - For plotting: need to make incompleteness in data clear…



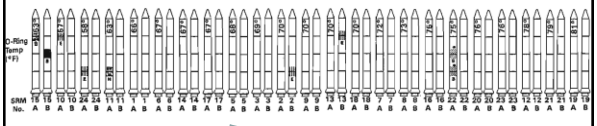
Tufte's Challenger Example

- Should the Space Shuttle Challenger have been launched when the launch temperature was -1C?



Tufte's Challenger Example (Cont)

- If temperature is what concerns us, let's reorder by temperature:



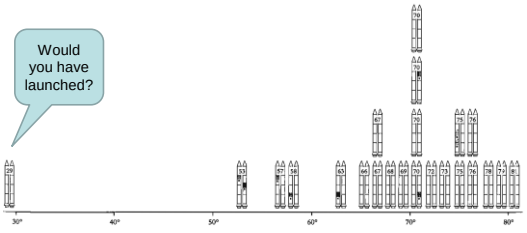
OK, some concentrated damage at lower temperatures, but not highly conclusive.

13

Tufte's Challenger Example (Cont)

- Let us accurately represent the temperature scale and try one more time:

Would you have launched?



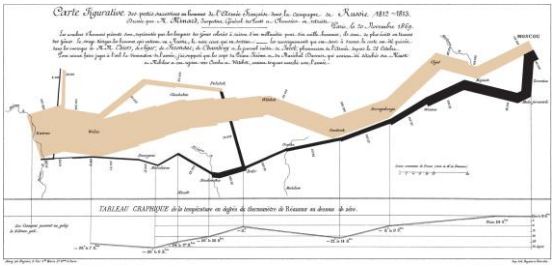
14

Custom Visualization

15

Minard's Map of the 1812 Russian Campaign

- Sometimes you have to construct your own visualization…



Modern redrawing of Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of **six types of data**: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates. Source: https://en.wikipedia.org/wiki/Charles_Joseph_Minard

16

Debugging Data-oriented Code

For machine learning or complex data science analysis

17

Synthetic Data for Debugging

- Your code will have bugs!
- **How to debug data analysis / machine learning?**
  - **Make a synthetic dataset**
    - E.g., (Features **x**: Age, Gender; Label y: {Snapchat, Facebook})
      - Everyone under 20 uses Snapchat
      - Everyone 20 and over uses Facebook
      - Random selection for gender

Or use label as a feature!

  - **Does your analysis uncover the expected trends?**
  - **What happens as you add noise?**
    - 90%, 80%, 70% of people under 20 use Snapchat?

Inability to uncover expected trends indicates a bug in your code or the need to revisit your analysis approach or hypothesis space.

18