Visualizing data with R and RStudio

ME 447/547

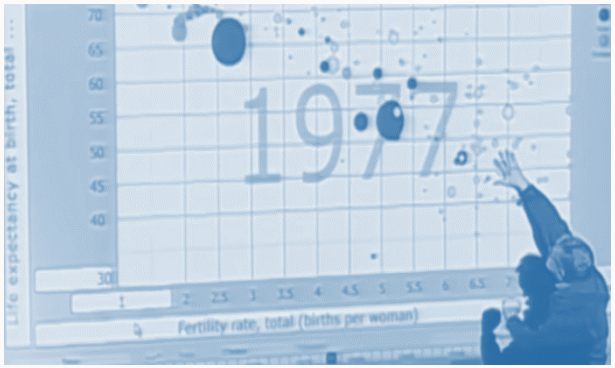
Richard Layton

March 2019

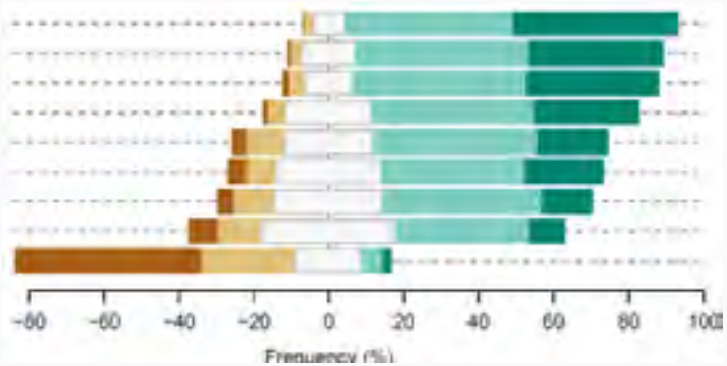
Rose-Hulman Institute of Technology

The course is designed to develop your skills in three areas

Rhetoric

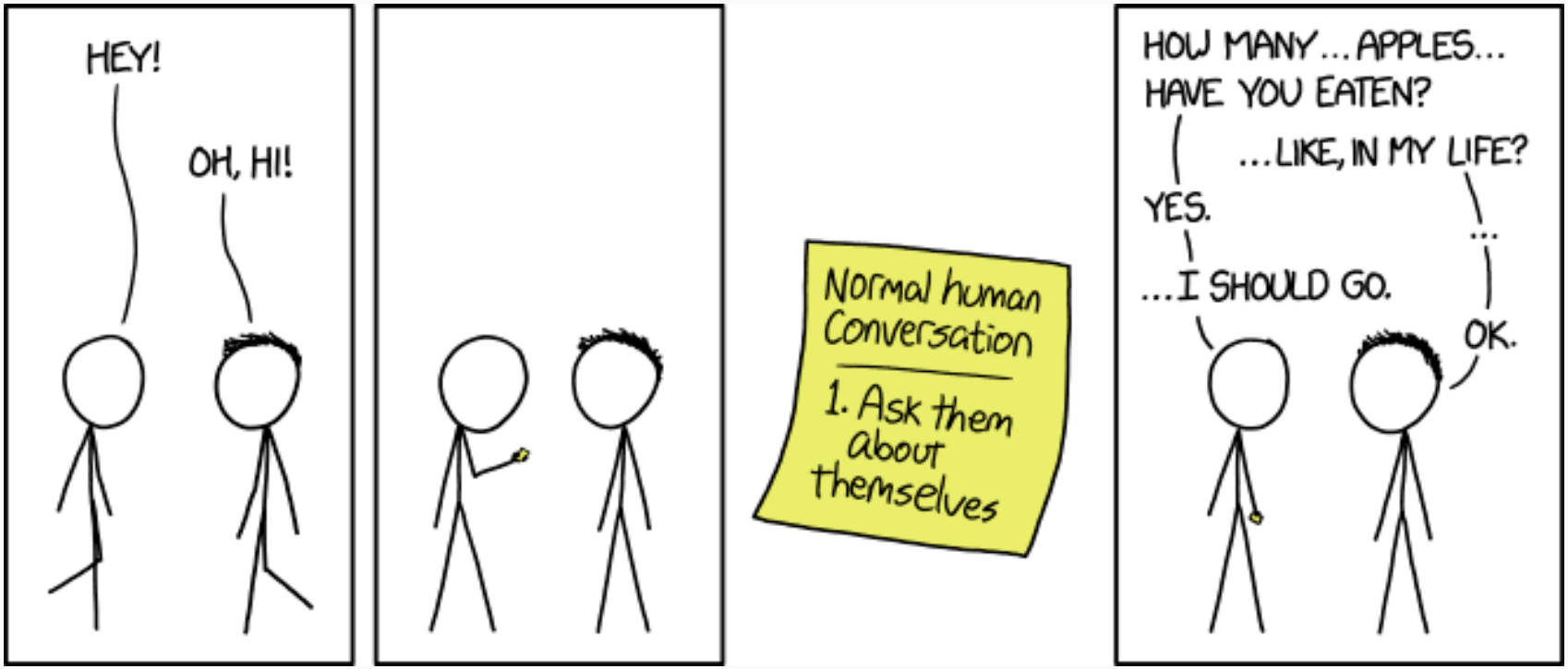


Repertoire



Means

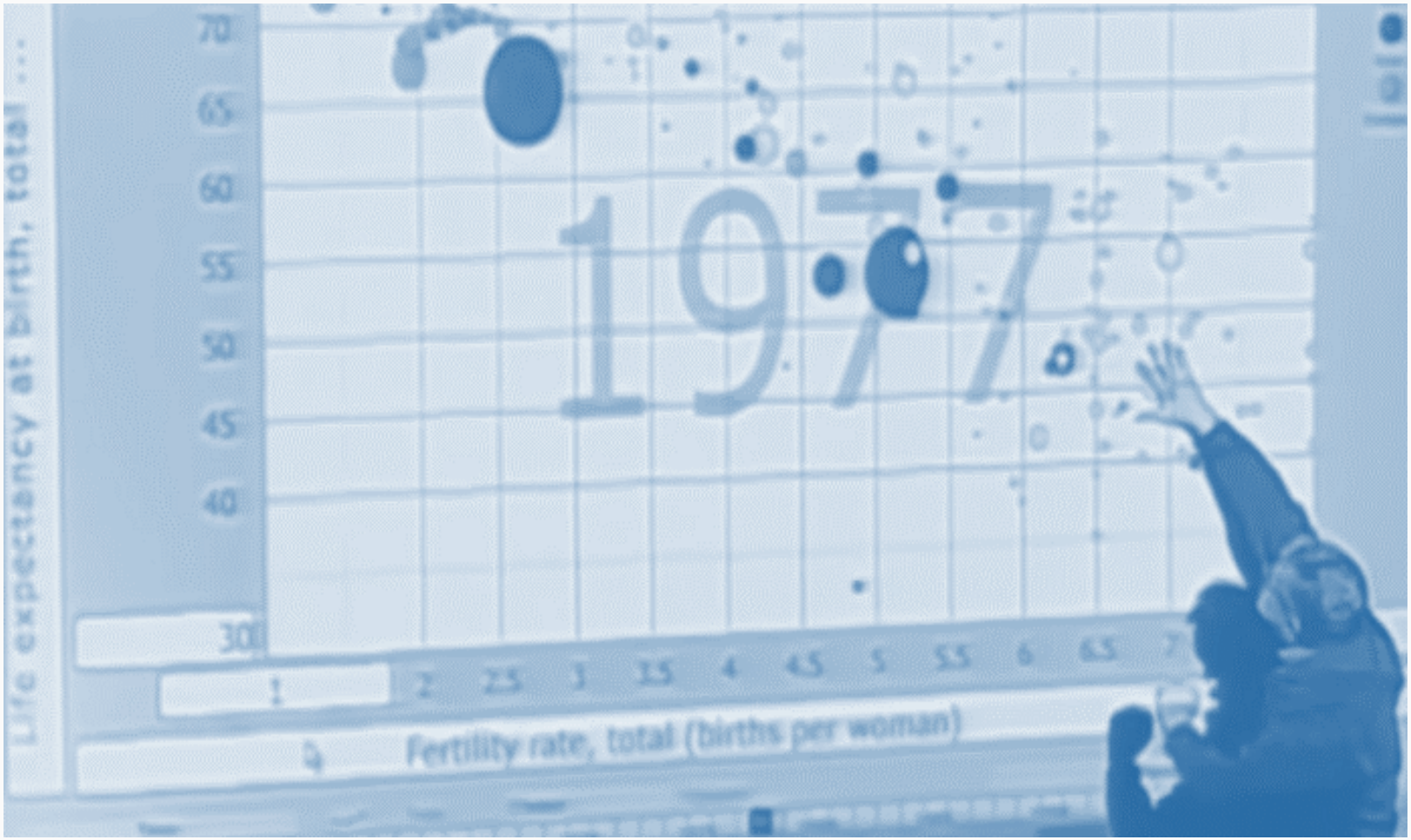
Please sit with someone you don’t know and introduce yourself



https://www.xkcd.com/1976/

Visual rhetoric

Designers shape information visually for rhetorical ends



Hans Rosling 2006 TED Talk

Consider the argument

How did Hans shape the information visually?

What were his rhetorical goals?

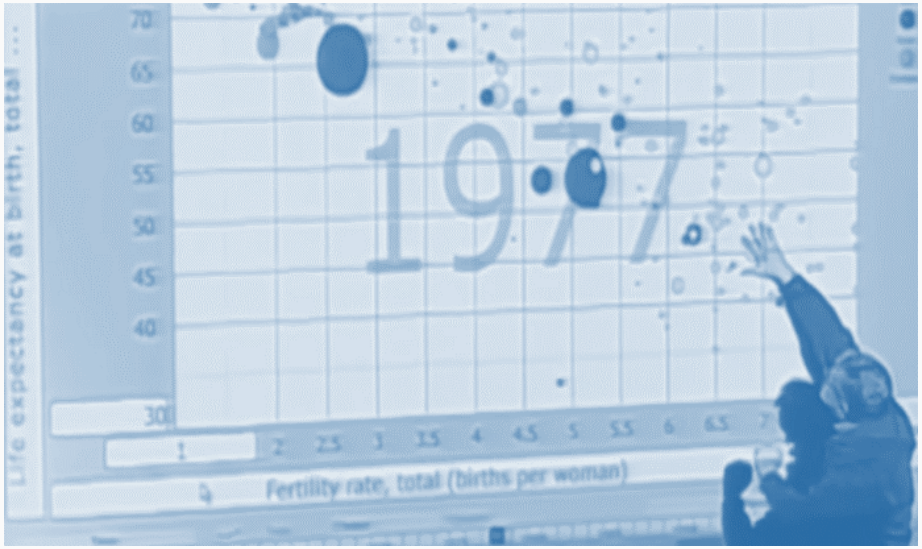


Image: TED2006

Consider a different, less credible, visual argument

True or False: $N_{\text{people on welfare}} > N_{\text{people with a full time job}}$

Image: Media Matters

Consider a different, less credible, visual argument

True or False: $N_{\text{people on welfare}} > N_{\text{people with a full time job}}$

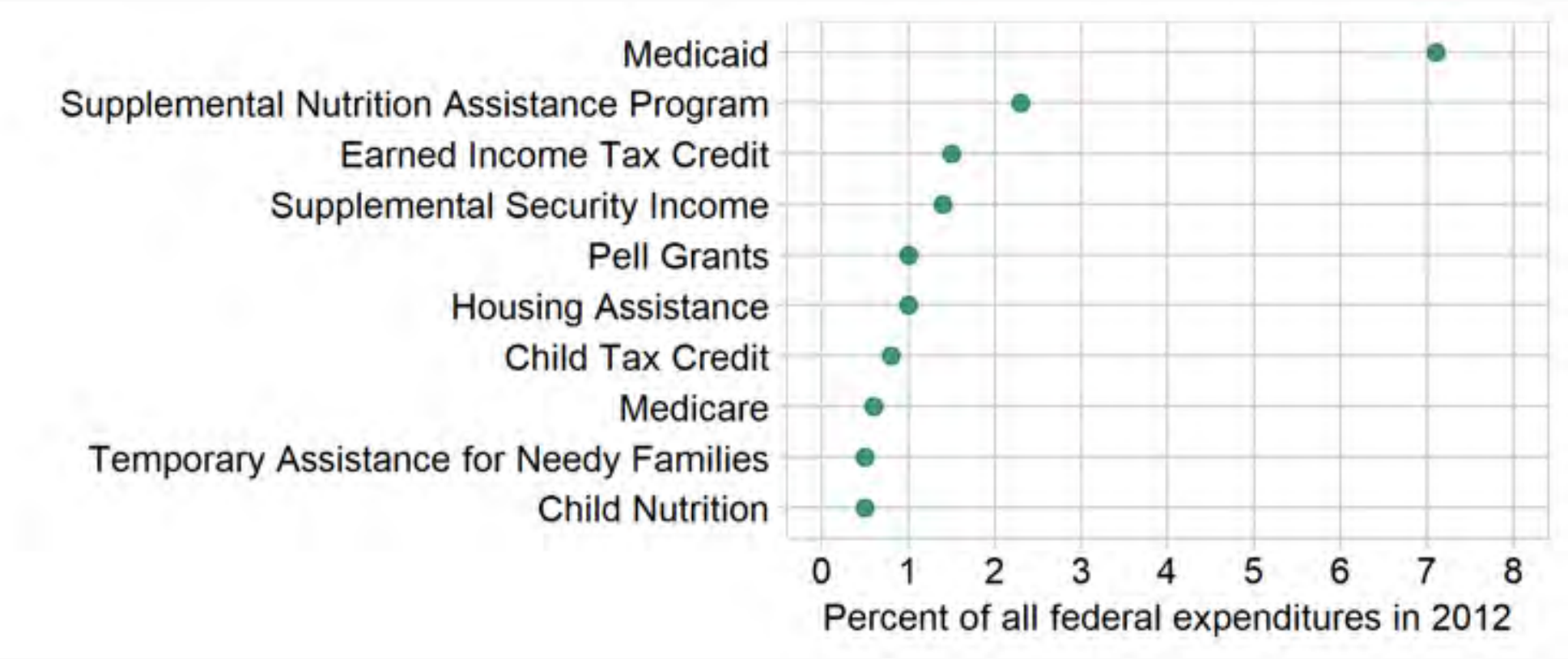
Image: Media Matters

False. One count is artificially high; the other is artificially low.

The counts use different definitions of “people”.

What does it mean to receive “welfare” benefits?

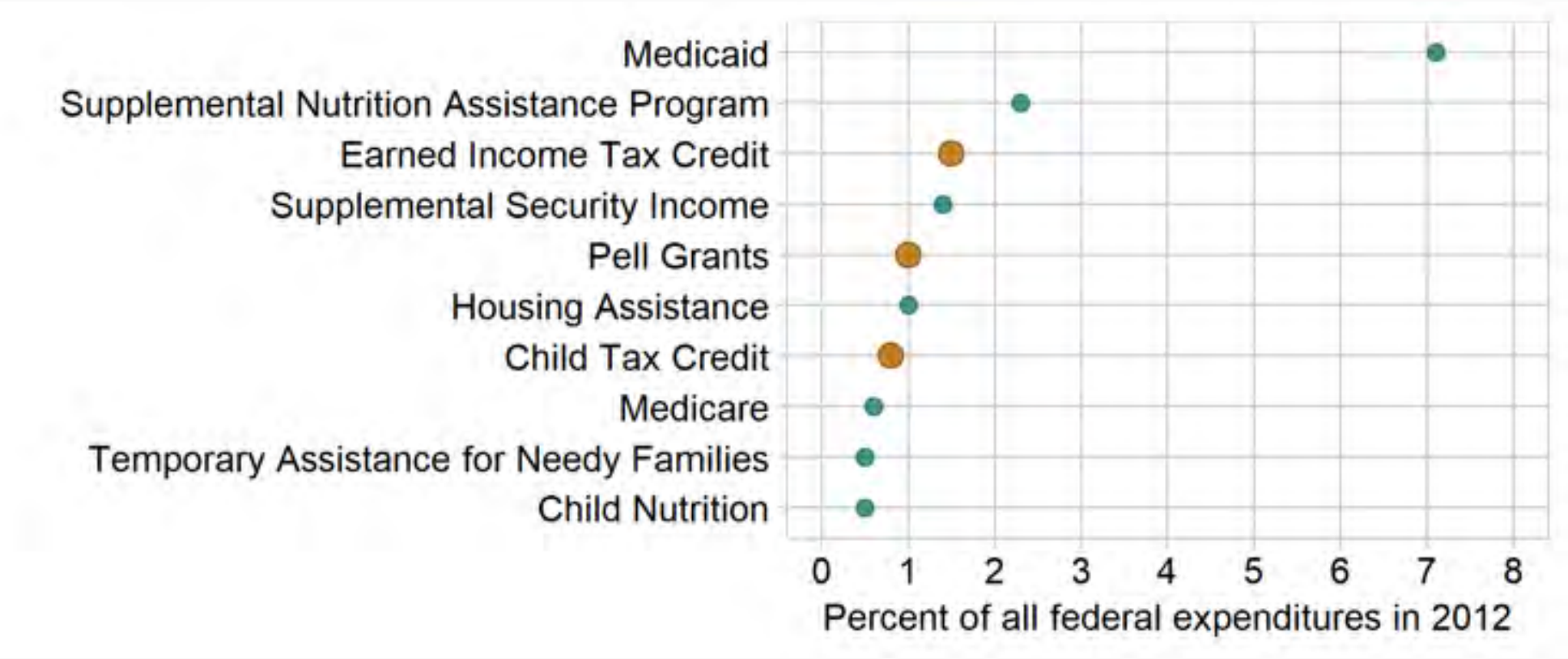
Federal means-tested programs and tax credits



In total, 17% of the 2012 US federal budget ($590 B / $3540 B).

What does it mean to receive “welfare” benefits?

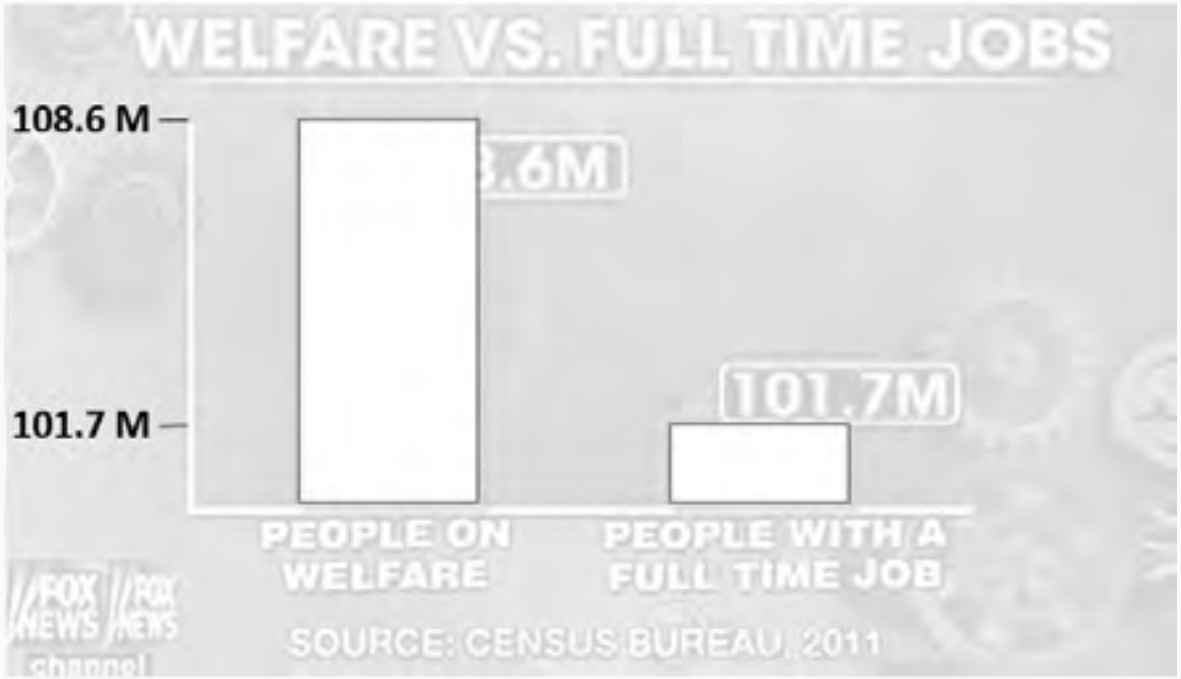
Federal means-tested programs and tax credits



In total, 17% of the 2012 US federal budget ($590 B / $3540 B).

Also, the visual argument belies the verbal argument

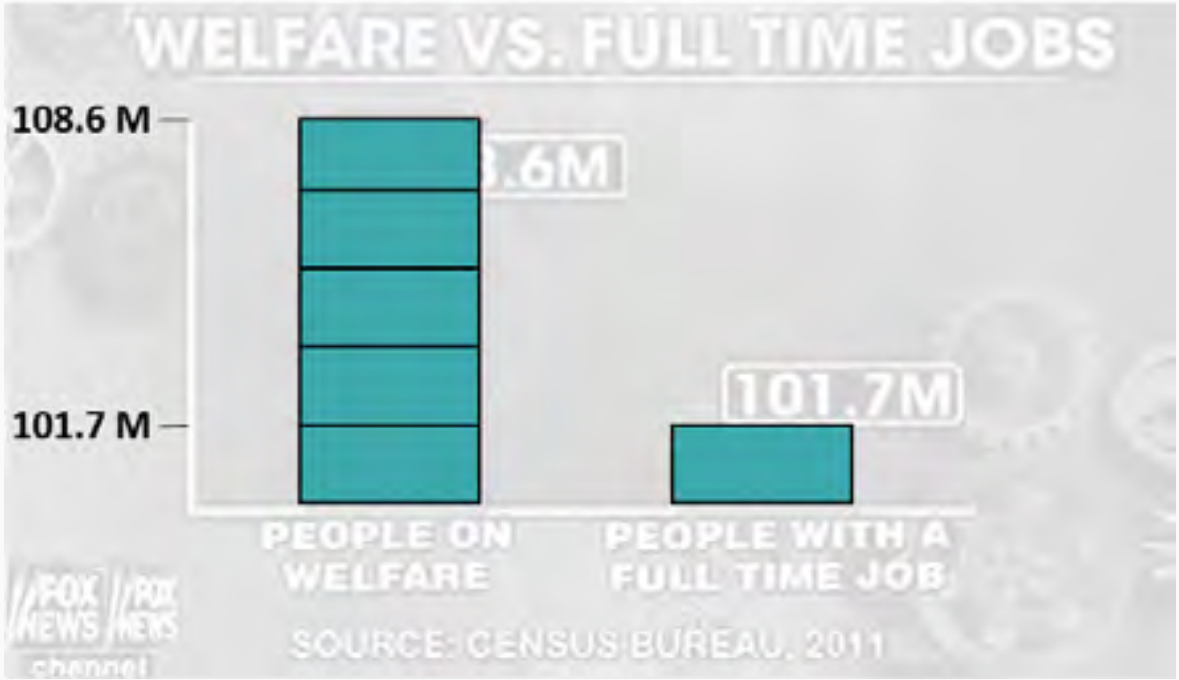
What is the visual lie?



A visual argument prevails—as the designer well knows

Verbal lie: 7% more people receiving benefits than not

Visual lie: 500% more people receiving benefits than not



What were the designer's rhetorical goals?

Ethical obligations are inherent in graph design

In data visualization, journalism meets engineering — Alberto Cairo

journalism increase knowledge among the public while minimizing harmful side effects

engineering give information a visual shape—model it, sculpt it—effectively and efficiently

(Cairo, 2014)

Repertoire

Graph design begins by understanding the data structure ...

Number of variables?

Continuous or discrete?

Number of variables?

Nominal or ordinal?

Number of levels each?

… and by knowing the prior art suited to that structure

62

strip plot

box and whisker plot

multiway

scatterplot

dot plot

line graph

conditioning plot

scatterplot matrix

63

parallel coordinate plot

cycle plot

mosaic plot

financial (OHLC) plot

diverging stacked bar

linked micromaps

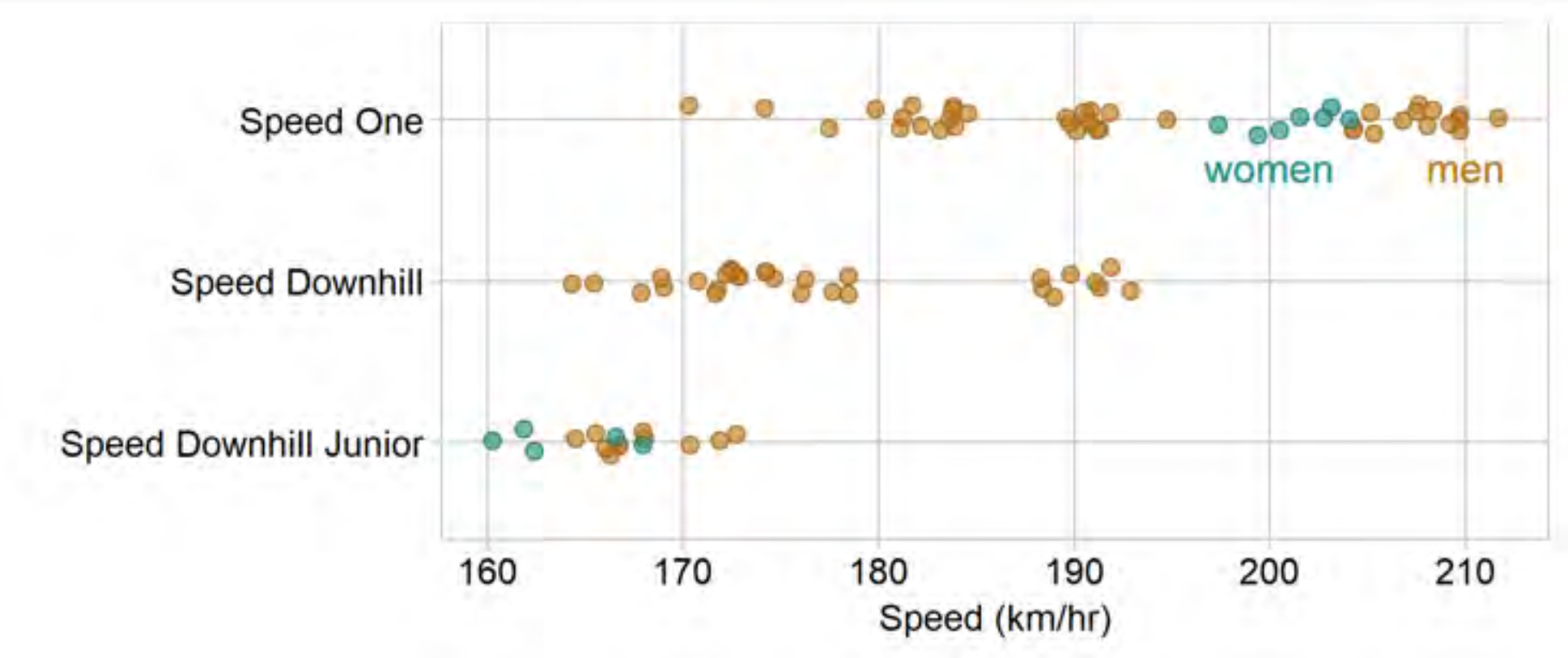
proportional symbol map

dot density map

Gallery — strip plot, jitter plot, or 1D scatterplot

Quantitative: speed (continuous), $N_{\text{obs}} = 91$

Categorical: event (nominal, 3 levels), sex (nominal, 2 levels)

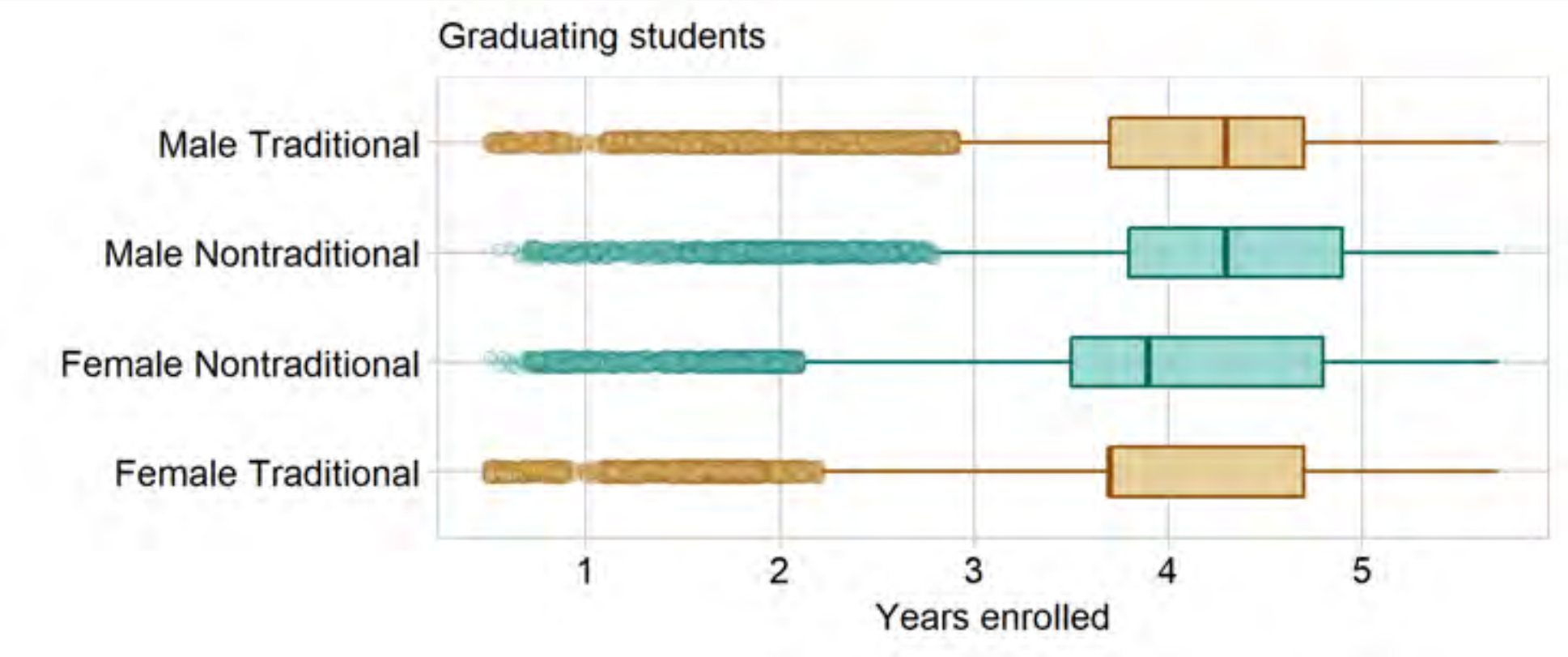


Data source (Unwin, 2015)

Gallery — box and whisker or box plot

Quantitative: Years enrolled (continuous), $N_{obs} = 269057$

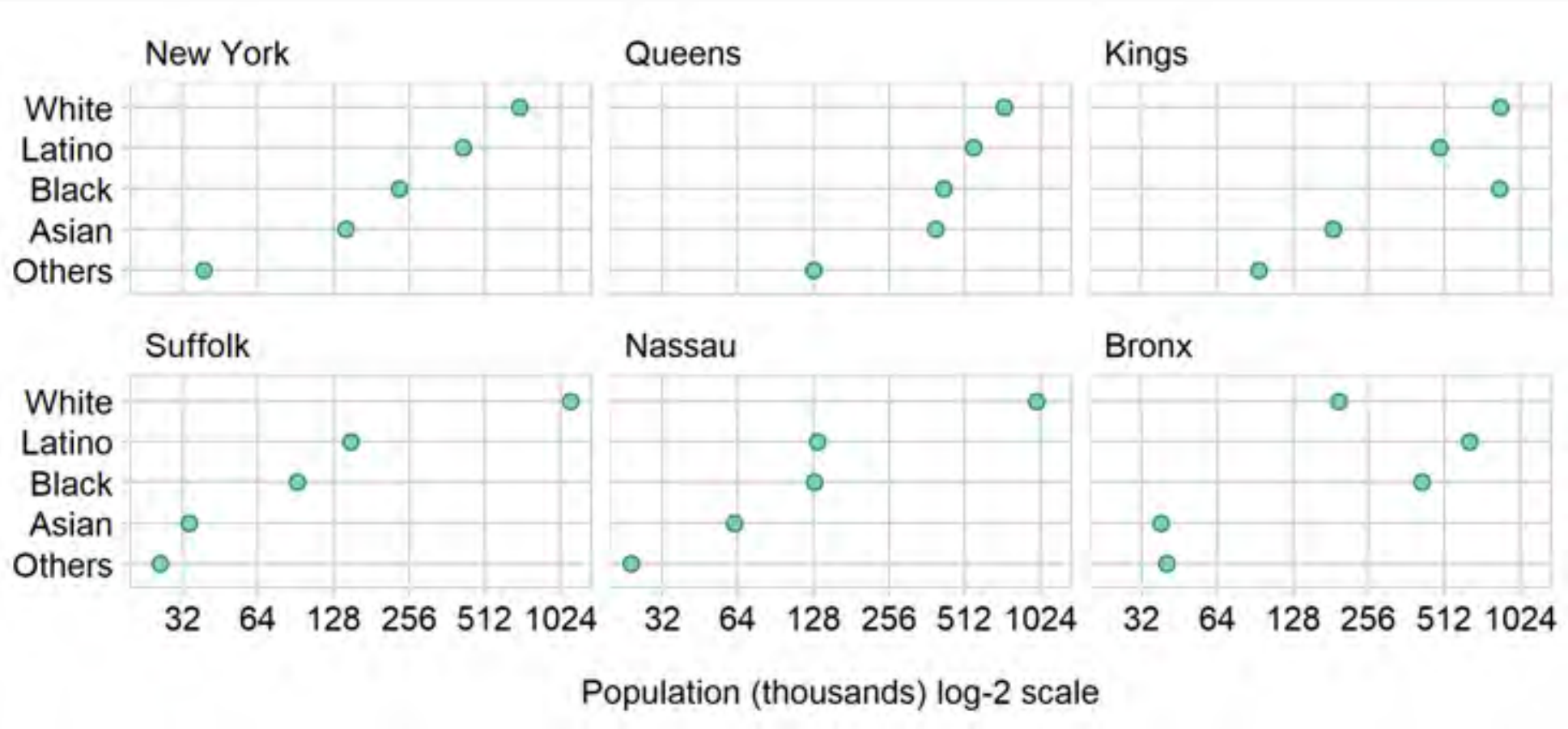
Categorical: Path + sex (nominal, 4 levels)



Gallery — multiway

Quantitative: Population (continuous), $N_{obs} = 30$

Categorical: Race/ethnicity (nominal, 5L) & county (nominal, 6L)

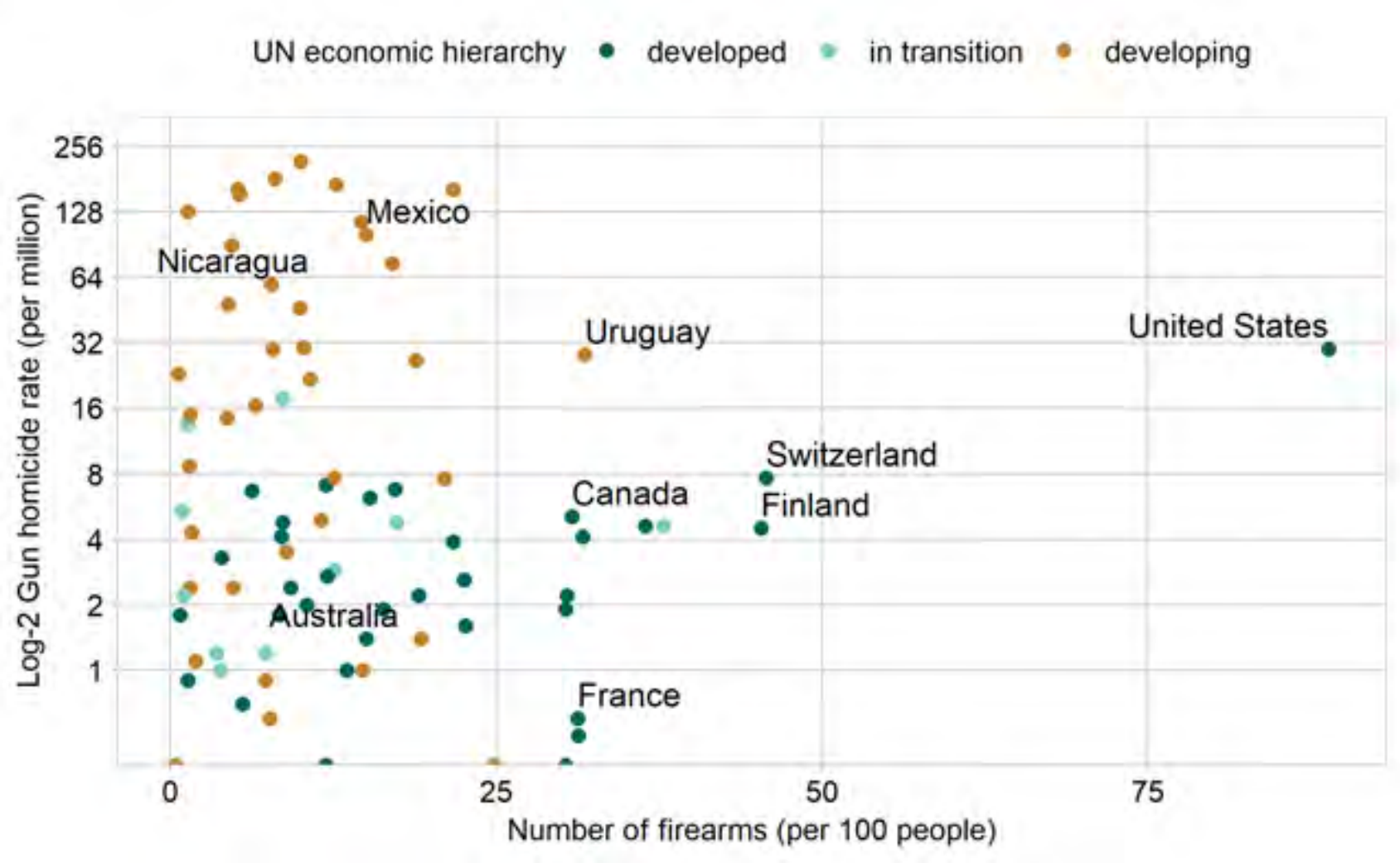


Adapted from (Robbins, 2013)

Gallery — scatterplot

Quantitative: Gun homicides & gun ownership (continuous), $N_{obs} = 90$

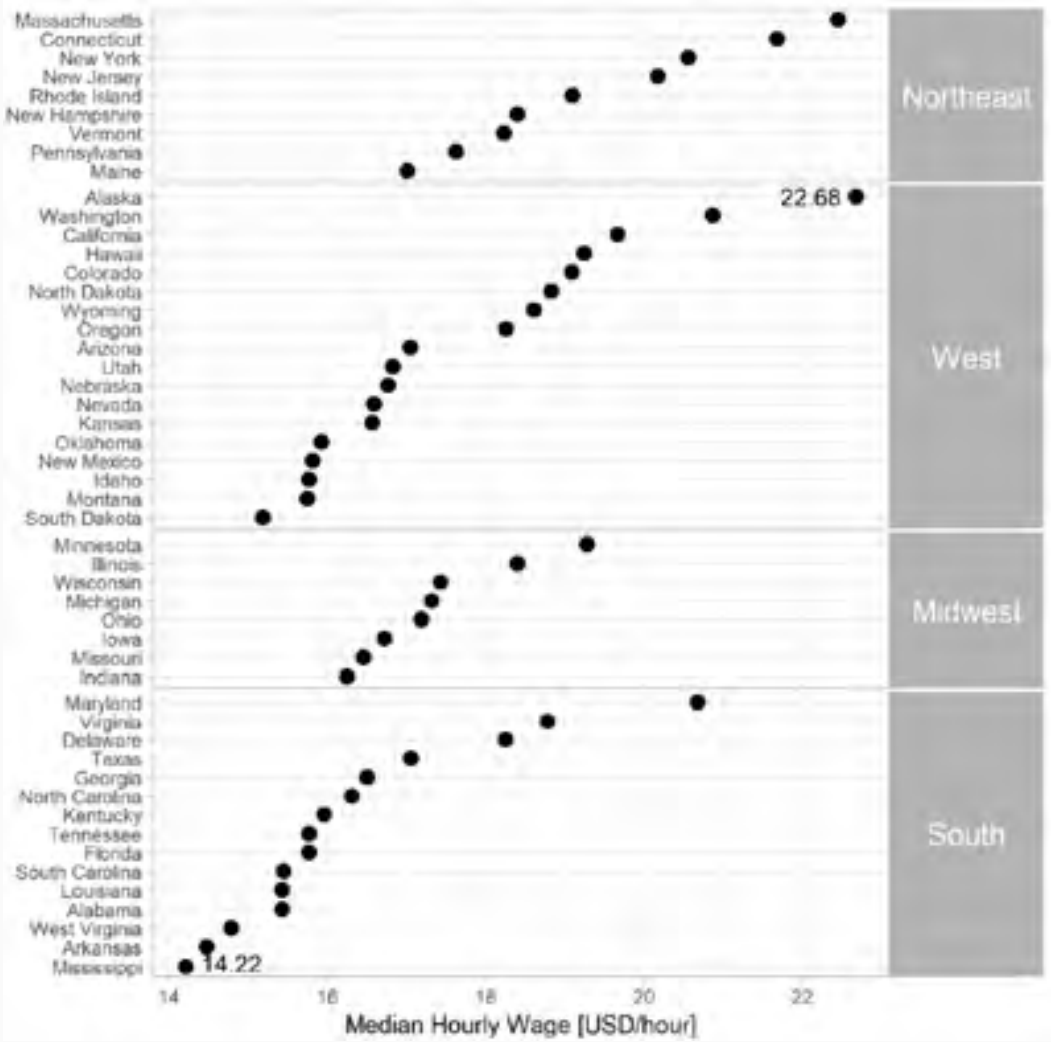
Categorical: Country (nominal, 90L) & economic hierarchy (nominal, 3L)



Gallery — Cleveland dot plot

Quantitative: 2016 median hourly wage (continuous), $N_{\text{obs}} = 50$

Categorical: State (nominal, 50 levels) & region (nominal, 4 levels)

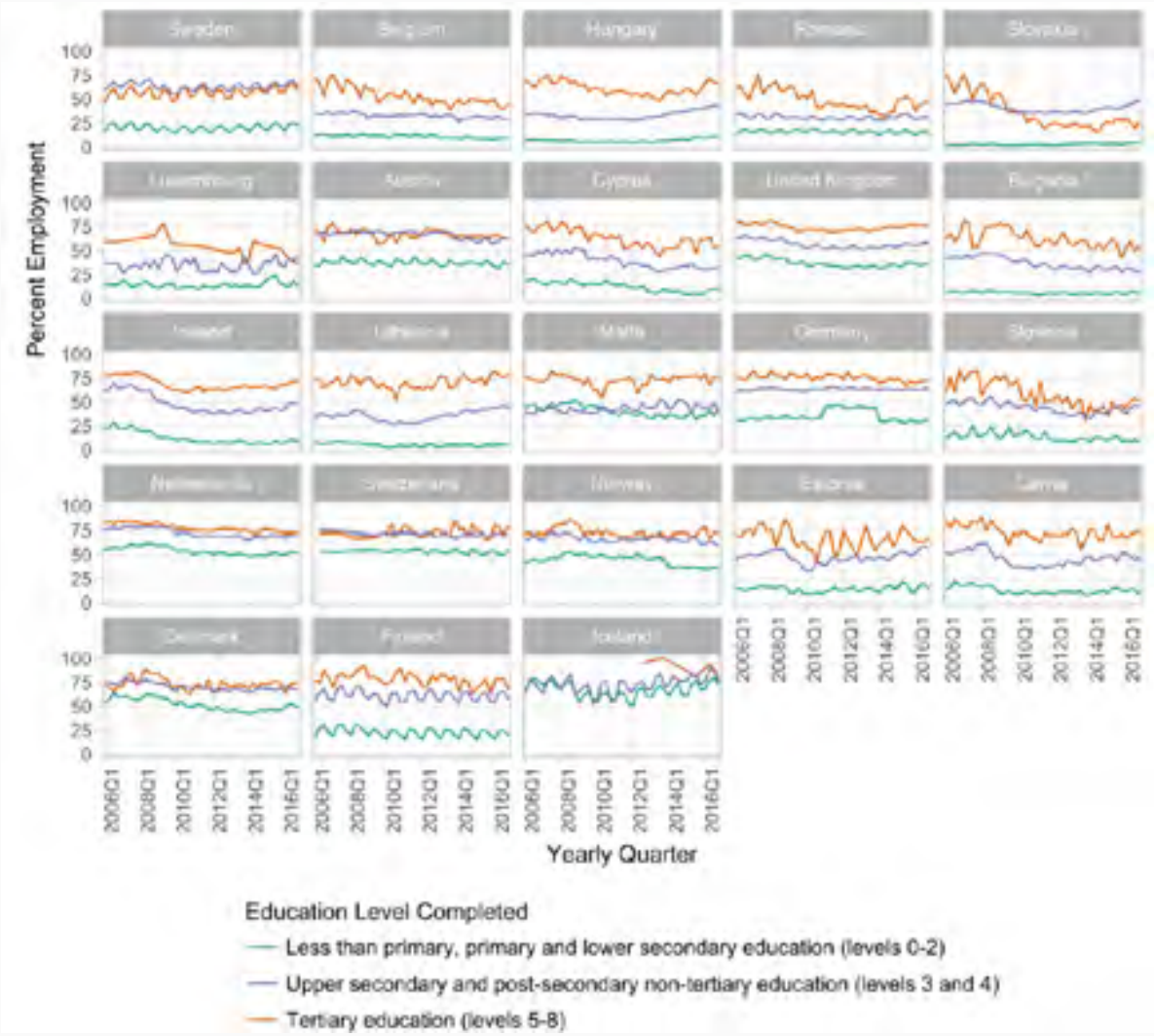


Joseph Hubach
2017 portfolio

Gallery — line graph

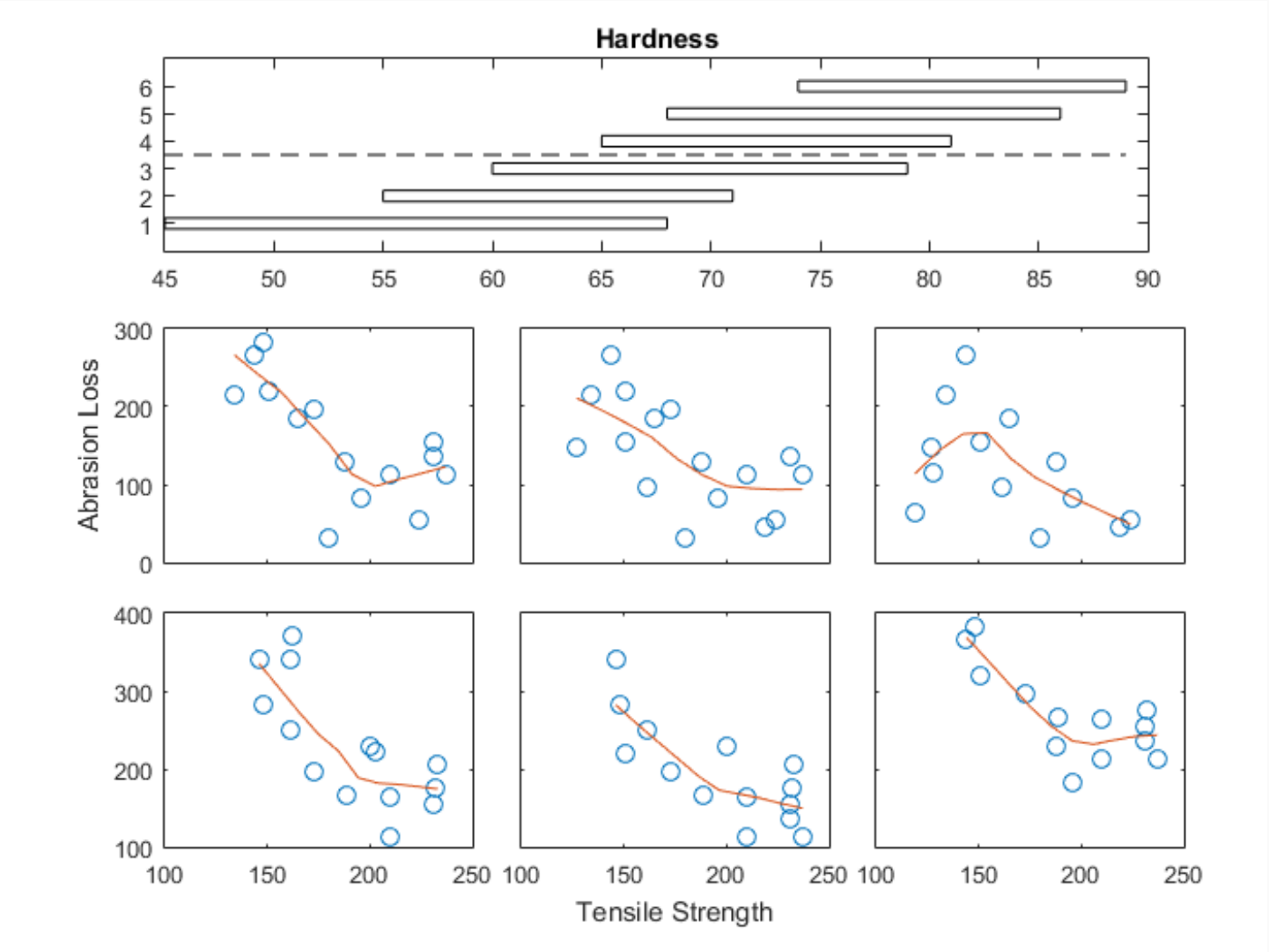
Quantitative: Percent employment (continuous), $N_{obs} = 1656$

Categorical: Country (nominal, 23L), education (ord, 3L), quarter (ord, 24L)



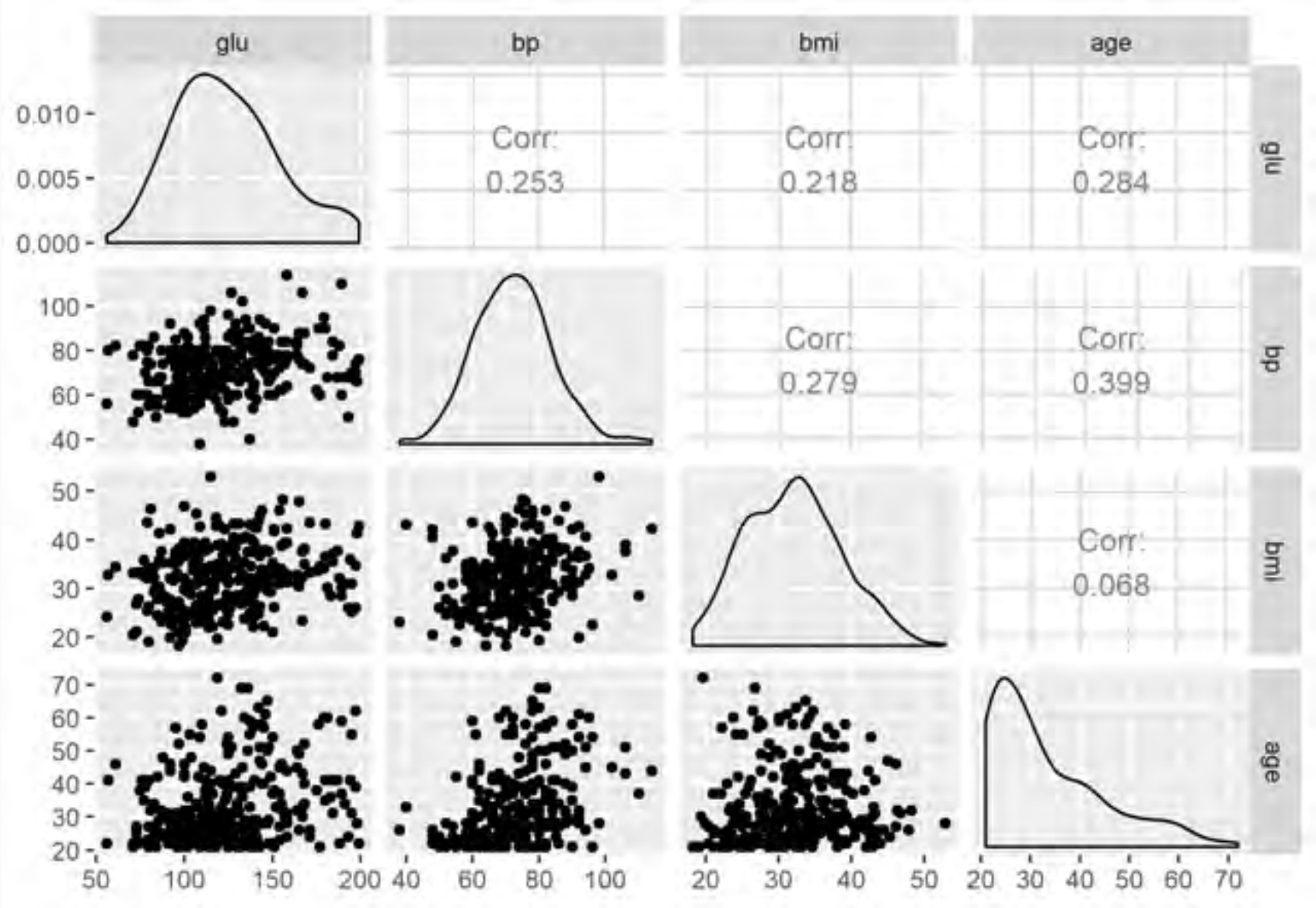
Gallery — conditioning plot

Quantitative: Rubber abrasion loss, tensile strength, & hardness (all continuous), $N_{\text{obs}} = 30$



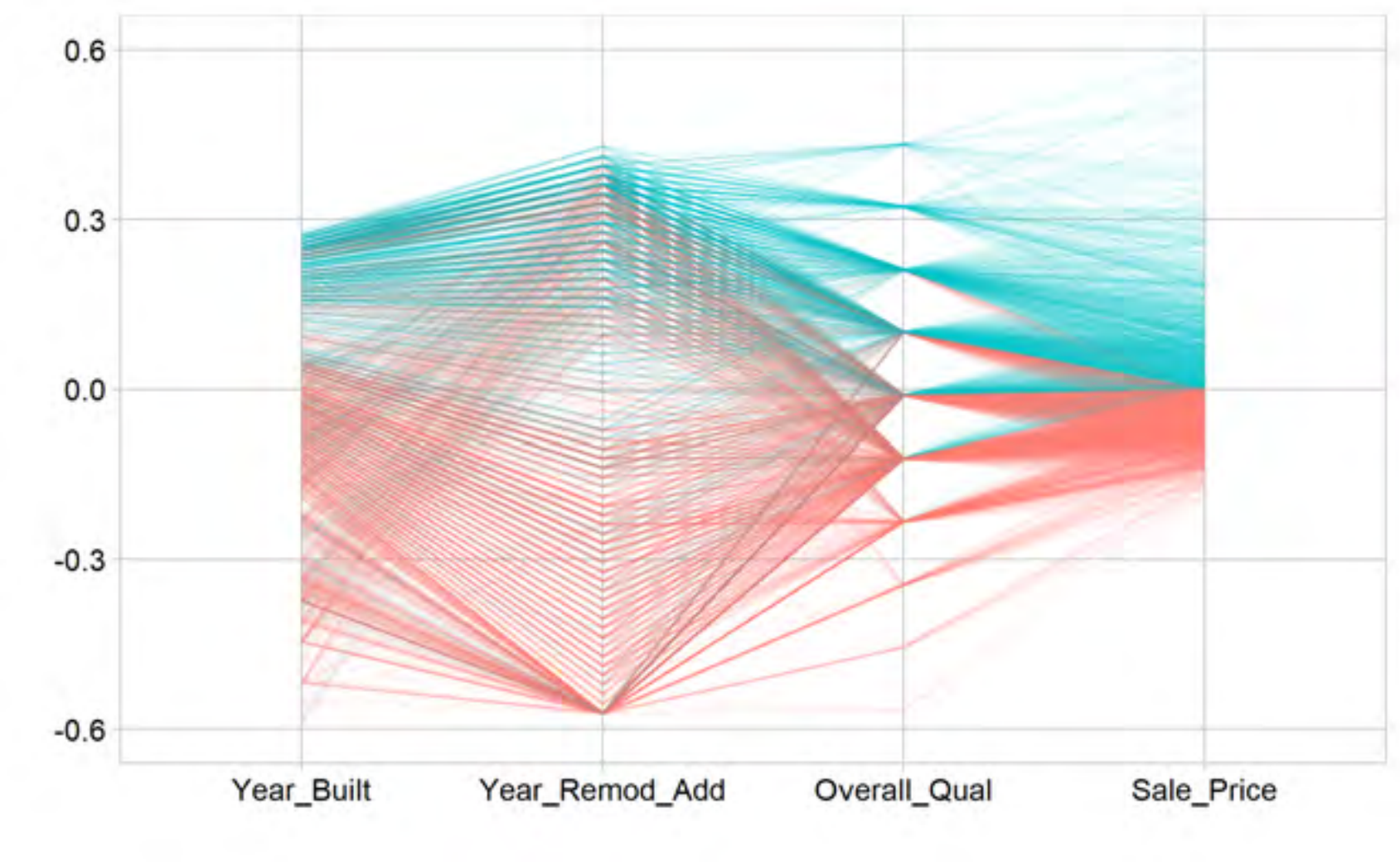
Gallery — scatterplot matrix

Quantitative: glucose, blood pressure, BMI, age (continuous), $N_{\text{obs}} = 300$



Gallery — parallel coordinate

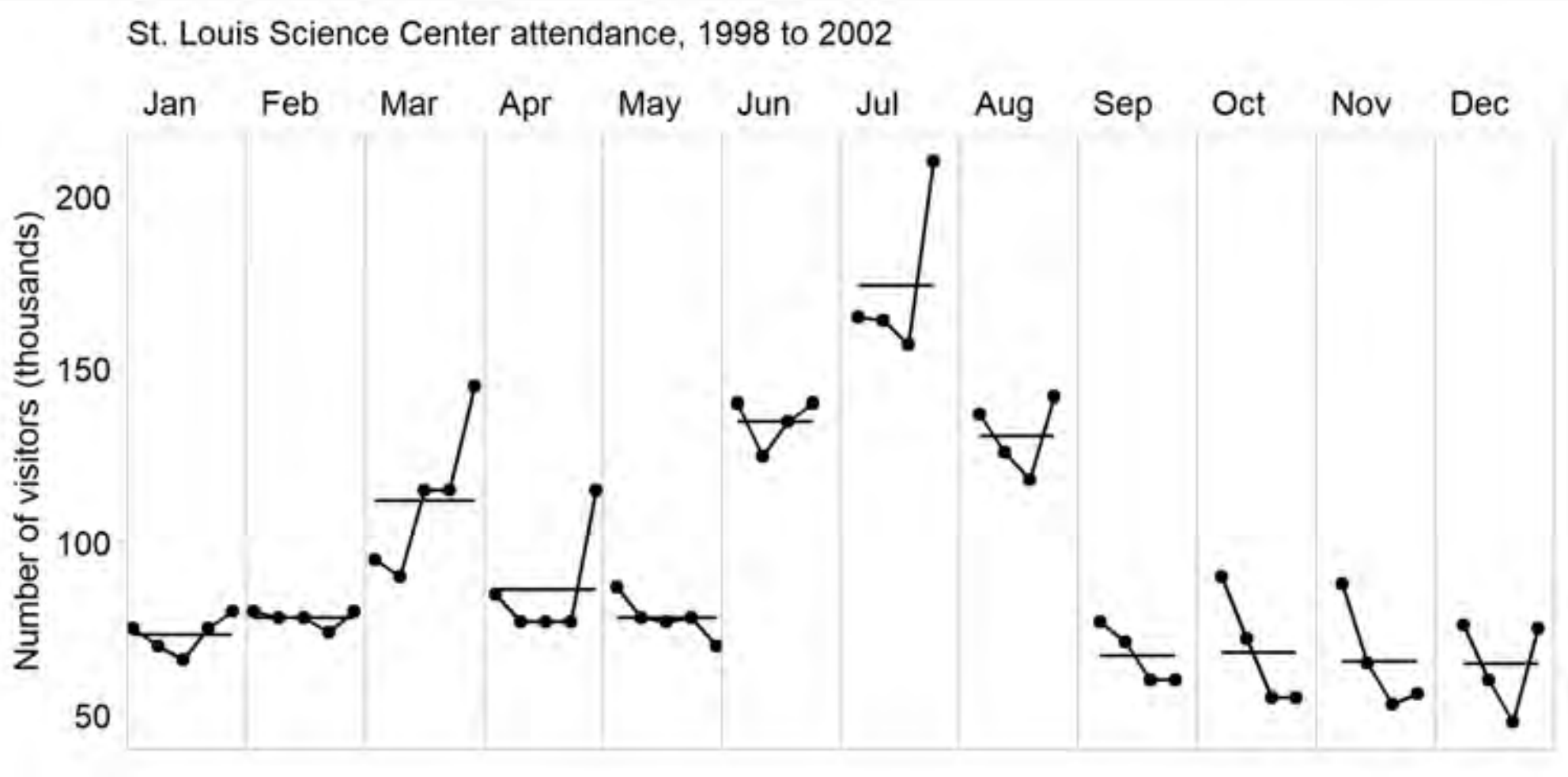
Quantitative: Year built, remodeled, & sale price (continuous),
quality (discrete) $N_{\text{obs}} = 2930$



Gallery — cycle plot

Quantitative: Number of visitors (continuous), $N_{obs} = 53$

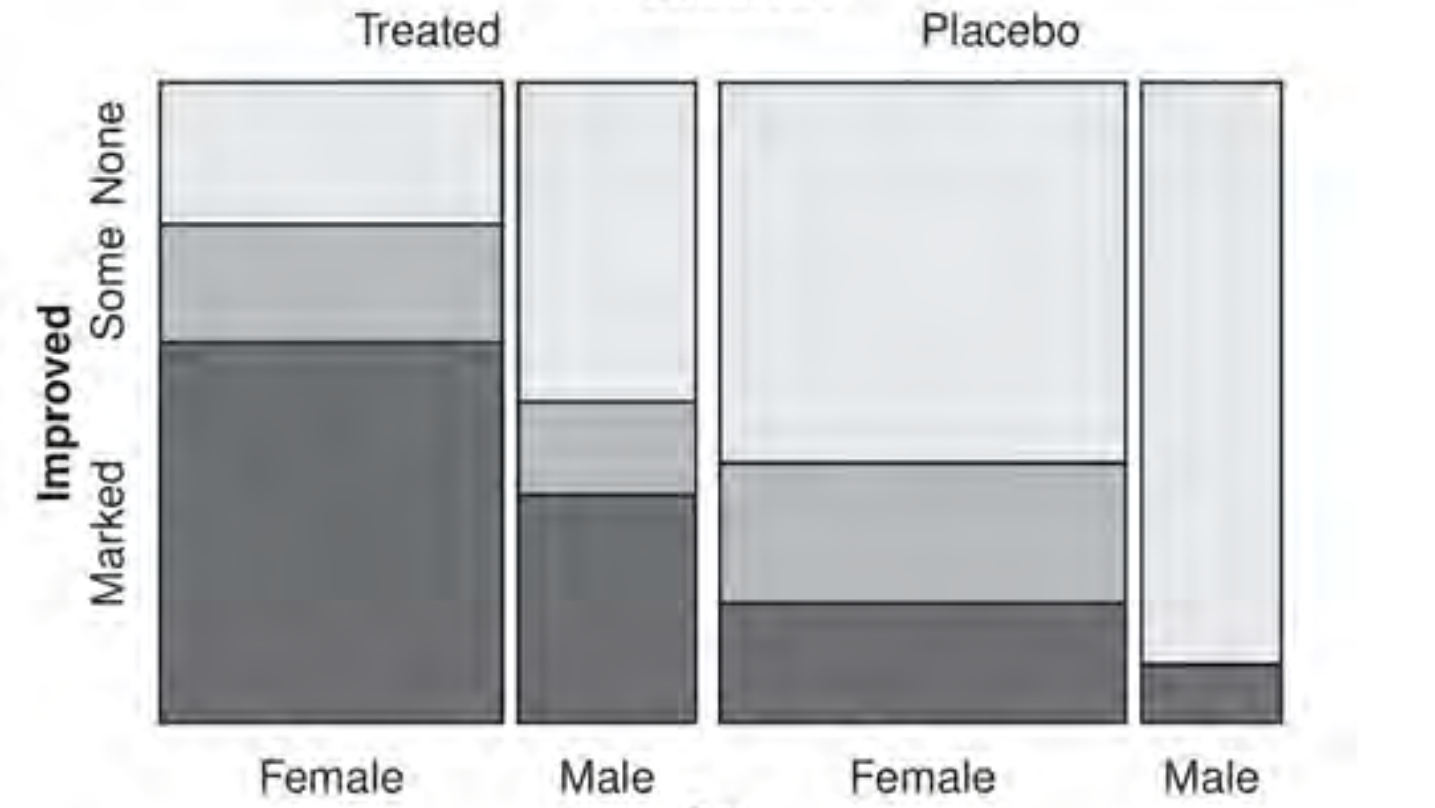
Categorical: Month (ordinal, 12 levels), year (ordinal, 5 levels)



Gallery — mosaic plot

Quantitative: Frequency (continuous), $N_{\text{obs}} = 84$

Categorical: Sex (nomi, 2L), treatment (nomi, 2L), outcome (ordi, 3L)

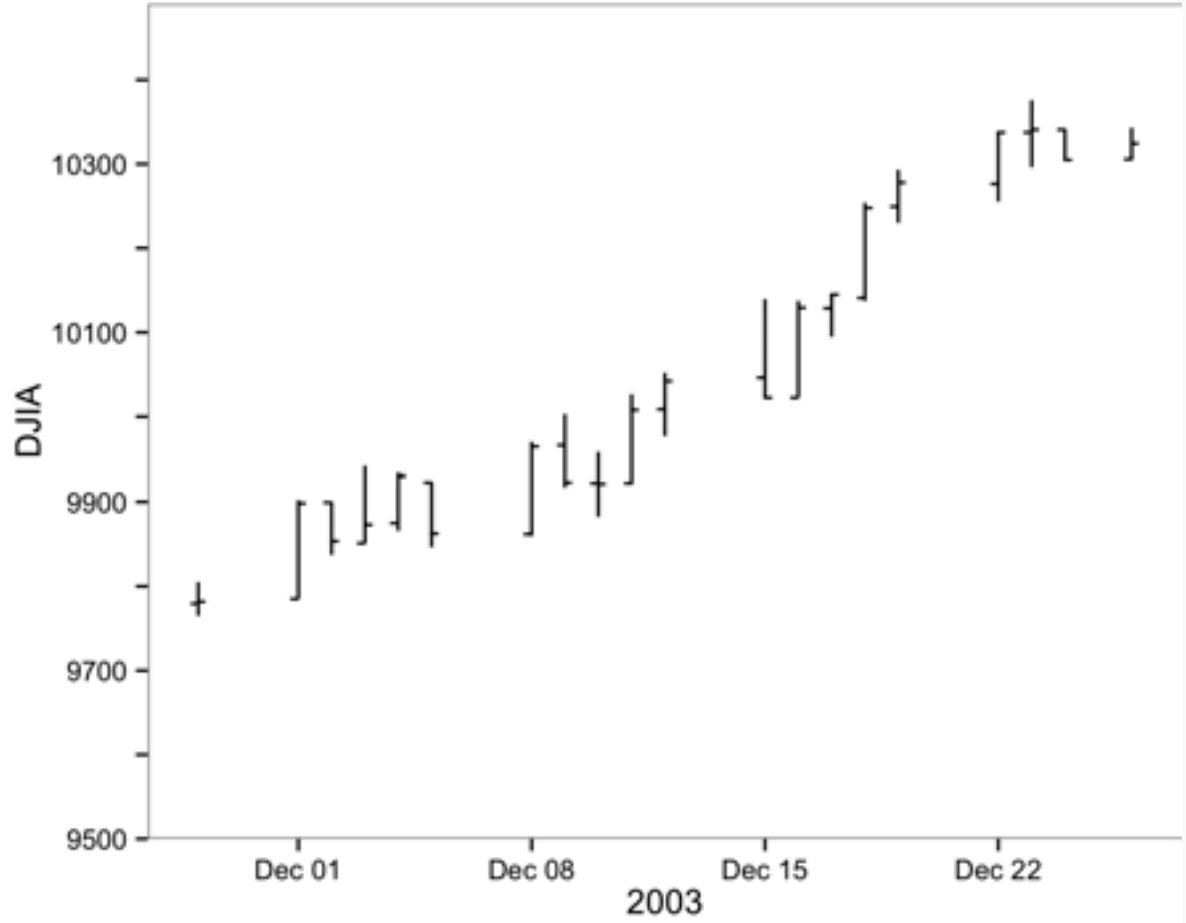


David Meyer, Achim Zeileis, and Kurt Hornik (2017) *vcd: Visualizing Categorical Data*, R package version 1.4-4, arthritis treatment data.

Gallery — financial (OHLC) plot

Quantitative: Opening, high, low, closing price (continuous), $N_{\text{obs}} = 20$

Categorical: Date (ordinal, 20 levels)

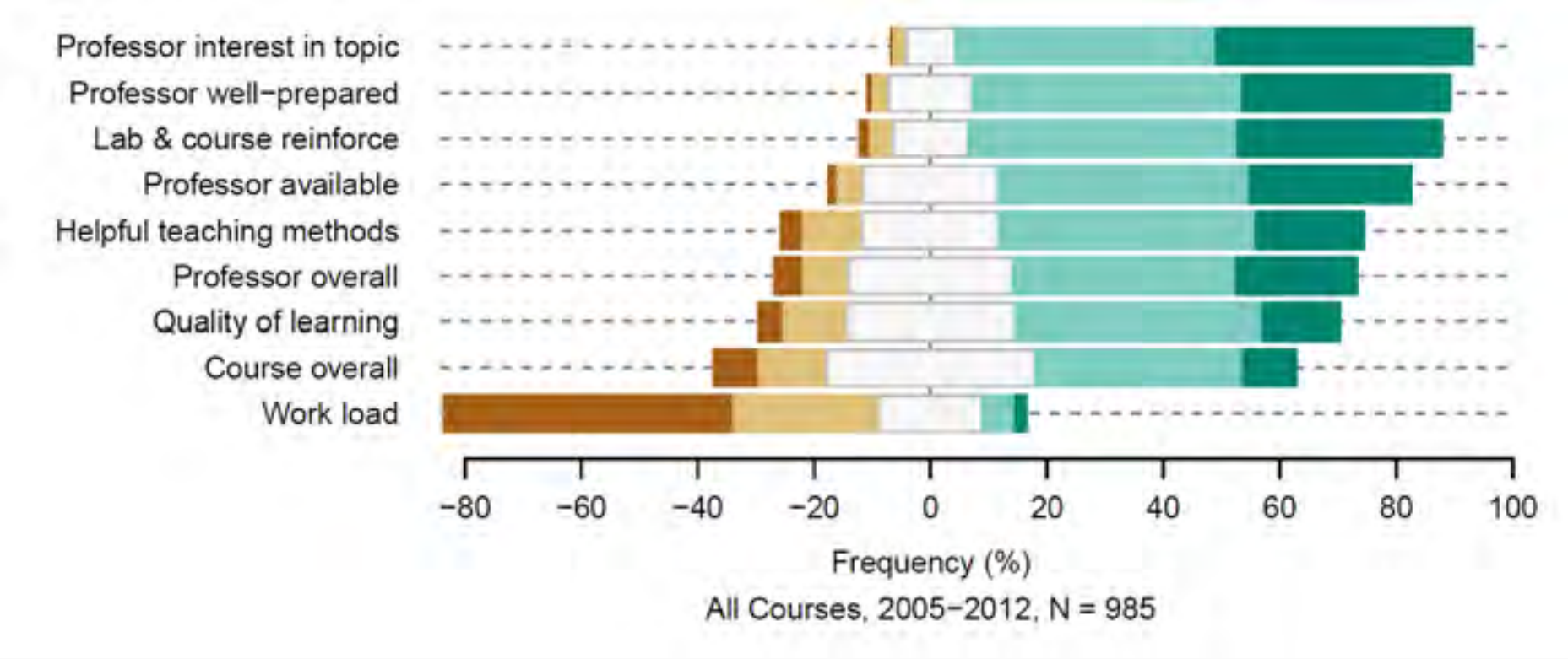


Adapted from (Robbins, 2013)

Gallery — diverging stacked bar

Quantitative: Frequency (continuous), $N_{obs} = 985$

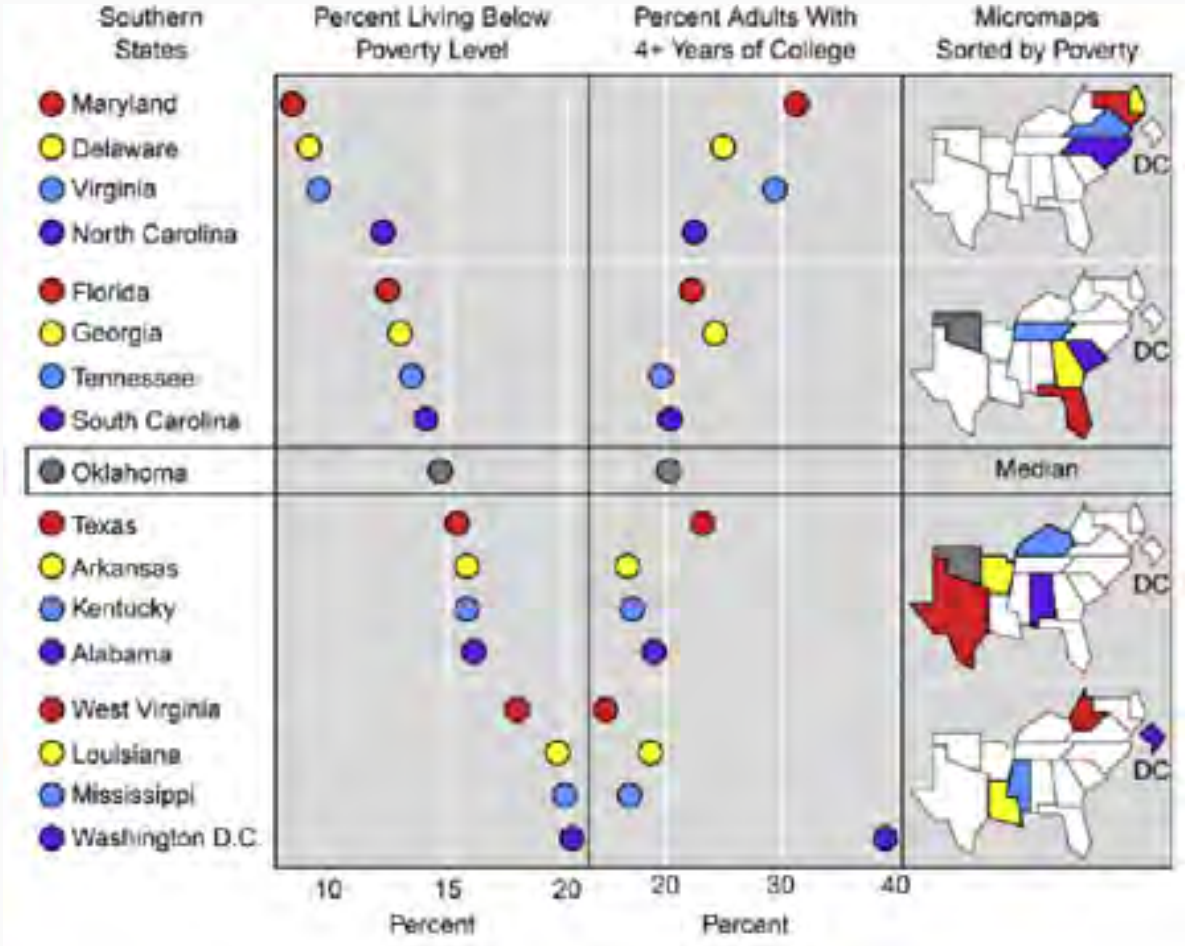
Categorical: Survey questions (nominal, 7L), responses (ordinal, 5L)



Gallery — linked micromaps

Quantitative: Percent poverty, percent college (continuous), $N_{\text{obs}} = 17$

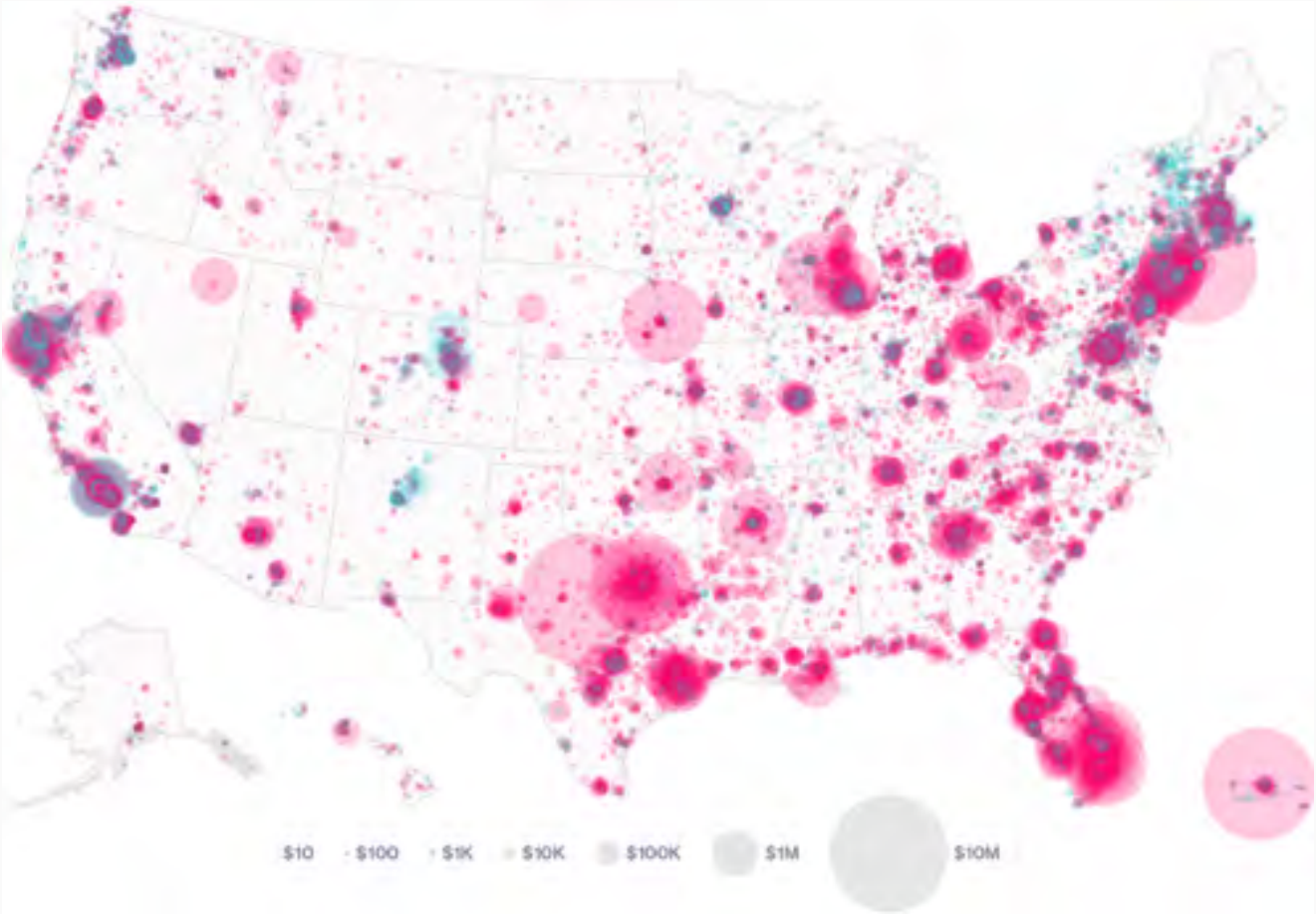
Categorical: State and geographic location (nominal, 17L)



Linda Pickle & Danial Carr (2010) Visualizing health data with micromaps, *Spatial and spatio-temporal epidemiology*, Vol. 1, pp. 143–50. https://bit.ly/2H967PH

Gallery — proportional symbol map

Categorical: Contribution (ordinal, 7 levels), party (nominal, 2 levels), ZIP code location (nominal, 42k levels), $N_{\text{obs}} = 42k$

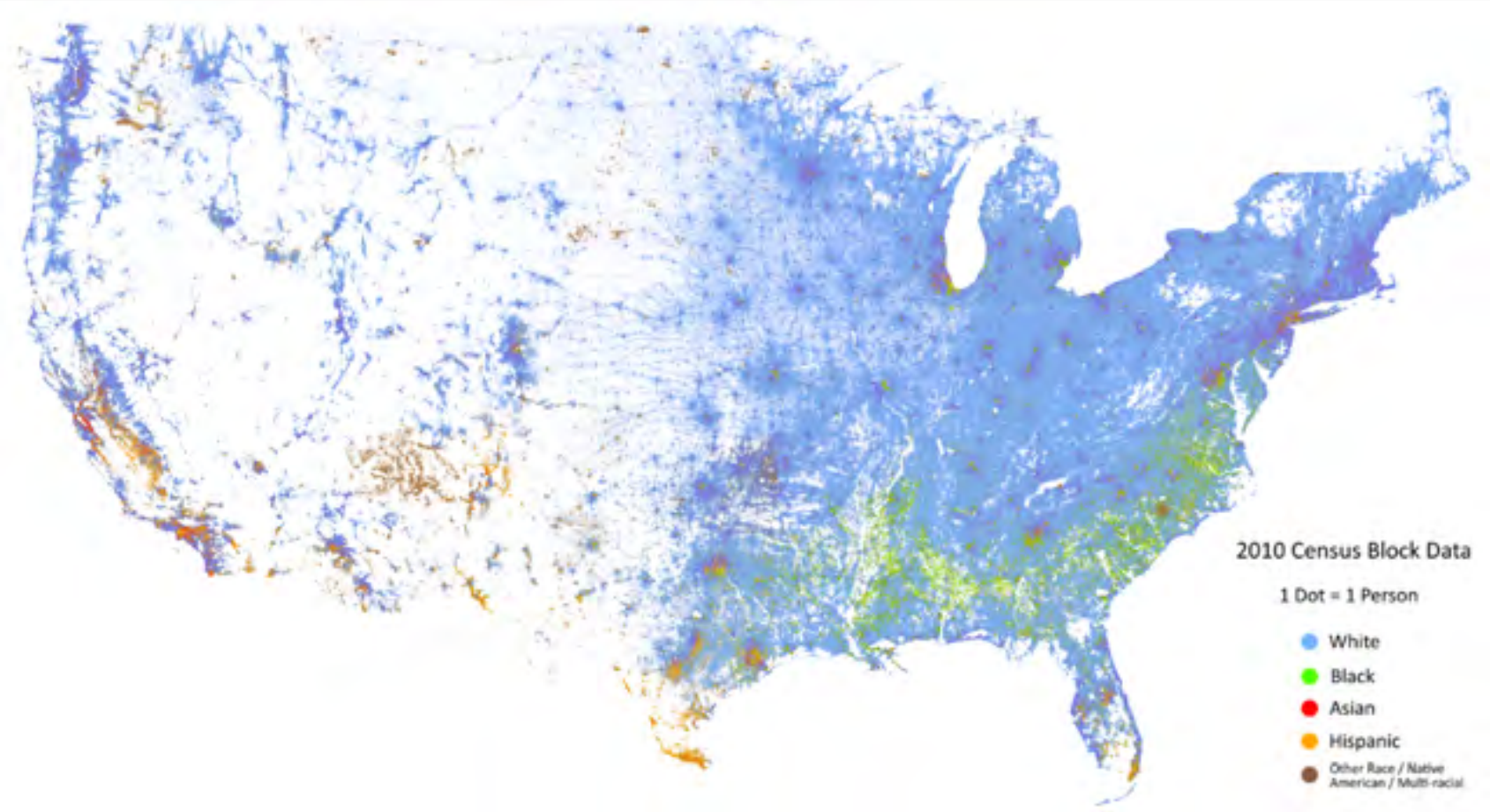


Zach Mider, Christopher Cannon, and Adam Pearce (Sep 15, 2015) Here's exactly where the candidates' cash came from, https://www.bloomberg.com/politics/graphics/2015-presidential-money-map/

Gallery — dot density map

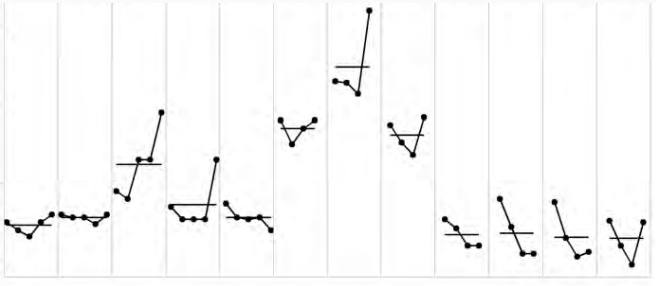
Quantitative: One dot per person, $N_{\text{obs}} = 308M$

Categorical: Race/ethnicity (nominal, 5L), geospatial location (nominal)

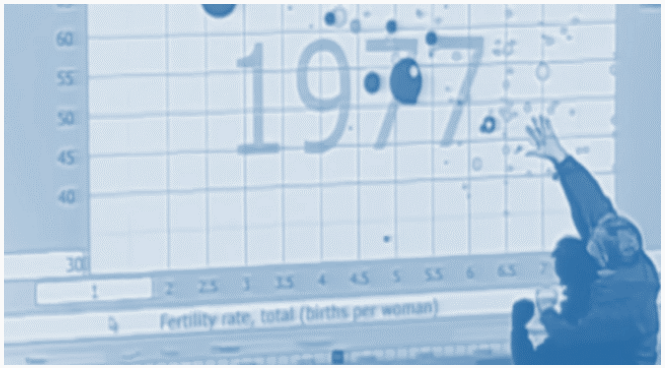


Implications for the designer

Grasp the data structure first



Explore data-suitable designs



Align the visual and verbal arguments

Means

Use the right tool for the job

RStudio — primary interface, integrates all our software

R — tidying data and creating graphs

R markdown — writing the portfolio, interleaving prose with code

Git — local version control

GitHub — collaborating and publishing the portfolio

The main topical threads weave through the calendar

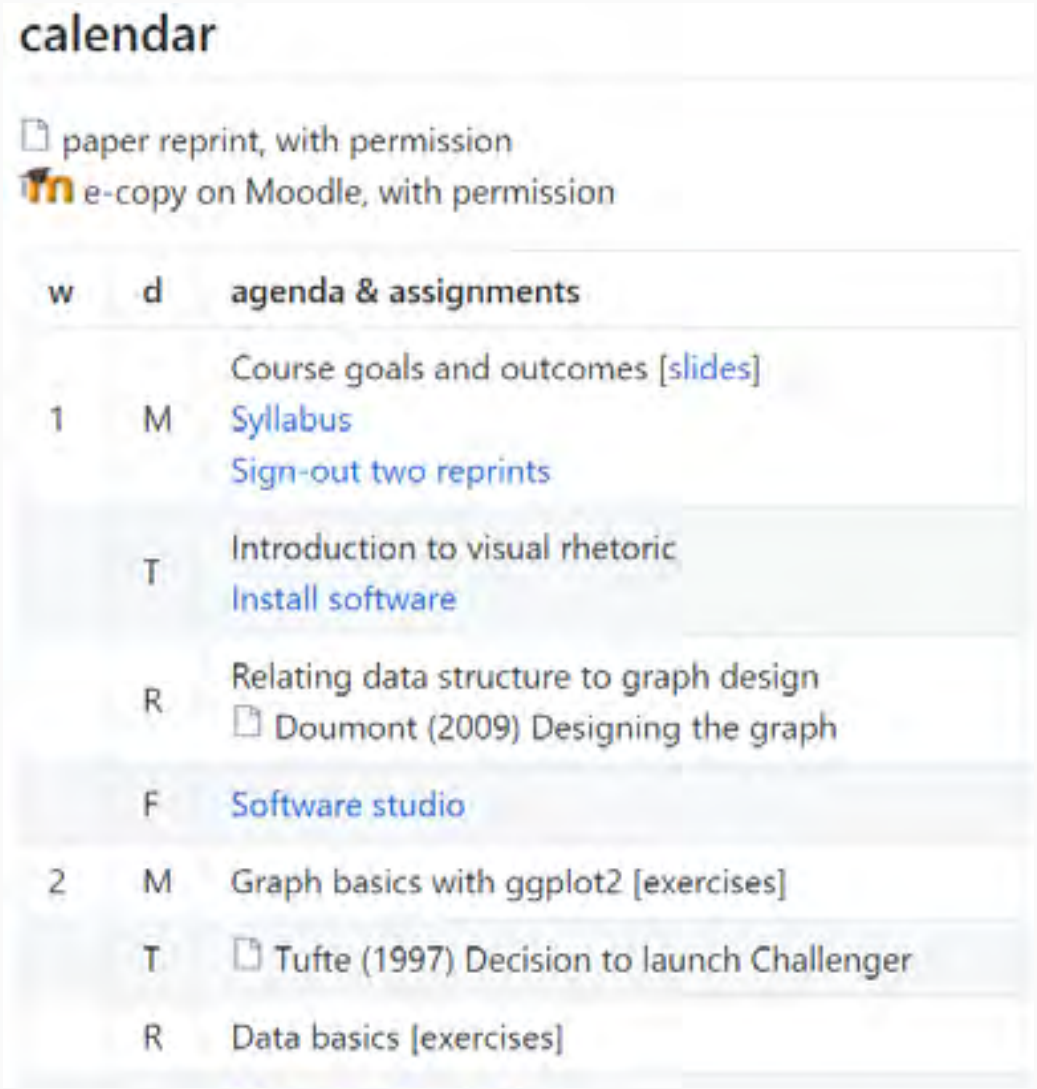
data

software

visual rhetoric

repertoire of graphs

portfolio



calendar

paper reprint, with permission

e-copy on Moodle, with permission

w	d	agenda & assignments
1	M	Course goals and outcomes [slides] Syllabus Sign-out two reprints
	T	Introduction to visual rhetoric Install software
	R	Relating data structure to graph design Doumont (2009) Designing the graph
	F	Software studio
2	M	Graph basics with ggplot2 [exercises]
	T	Tufte (1997) Decision to launch Challenger
	R	Data basics [exercises]

https://github.com/DSR-RHIT/me447-visualizing-data

References

Cairo A (2014) *Ethical infographics.* The Investigative Reporters and Editors Journal, Spring 2014
https://www.dropbox.com/s/pqgmg02yz0pgju4/EthicalInfographics.pdf

Robbins N (2013) *Creating More Effective Graphs.* Chart House, Wayne, NJ

Unwin A (2015) *GDAdata: Datasets for the book Graphical Data Analysis with R.* R package version 0.93
https://CRAN.R-project.org/package=GDAdata