WELCOME TO THE 2026 NDACAN SUMMER TRAINING SERIES!

The session will begin at 12pm EST.

This session is being recorded.

Please submit questions to the Q&A box!

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Duke University, Cornell University, UC San Francisco, & Mathematica

LAYING THE GROUNDWORK

FOUNDATIONAL SKILLS FOR USING ADMINISTRATIVE DATA IN CHILD WELFARE RESEARCH

NDACAN SUMMER TRAINING SERIES SCHEDULE

- **July 1st, 2026**
  - Overview of NDACAN administrative datasets
- **July 8th, 2026**
  - Data cleaning and management
- **July 15th, 2026**
  - Linking NCANDS and AFCARS
- **July 22rd, 2026**
  - Handling missing data
- **July 29th, 2026**
  - Data presentation and visualization

SESSION AGENDA

- Data presentation and visualization
- Demonstration in R
- Q & A

DATA PRESENTATION AND VISUALIZATION

DATA PRESENTATION

- Data presentation is foremost about communication.
  - *What is the message that you are trying to send?*
  - *How will your audience receive your intended message?*
- Data presentation is highly conventional and sometimes rule-bound, but it should be intentional and creative.
- With successful data presentation, tables and figures coordinate to communicate a coherent and convincing story.

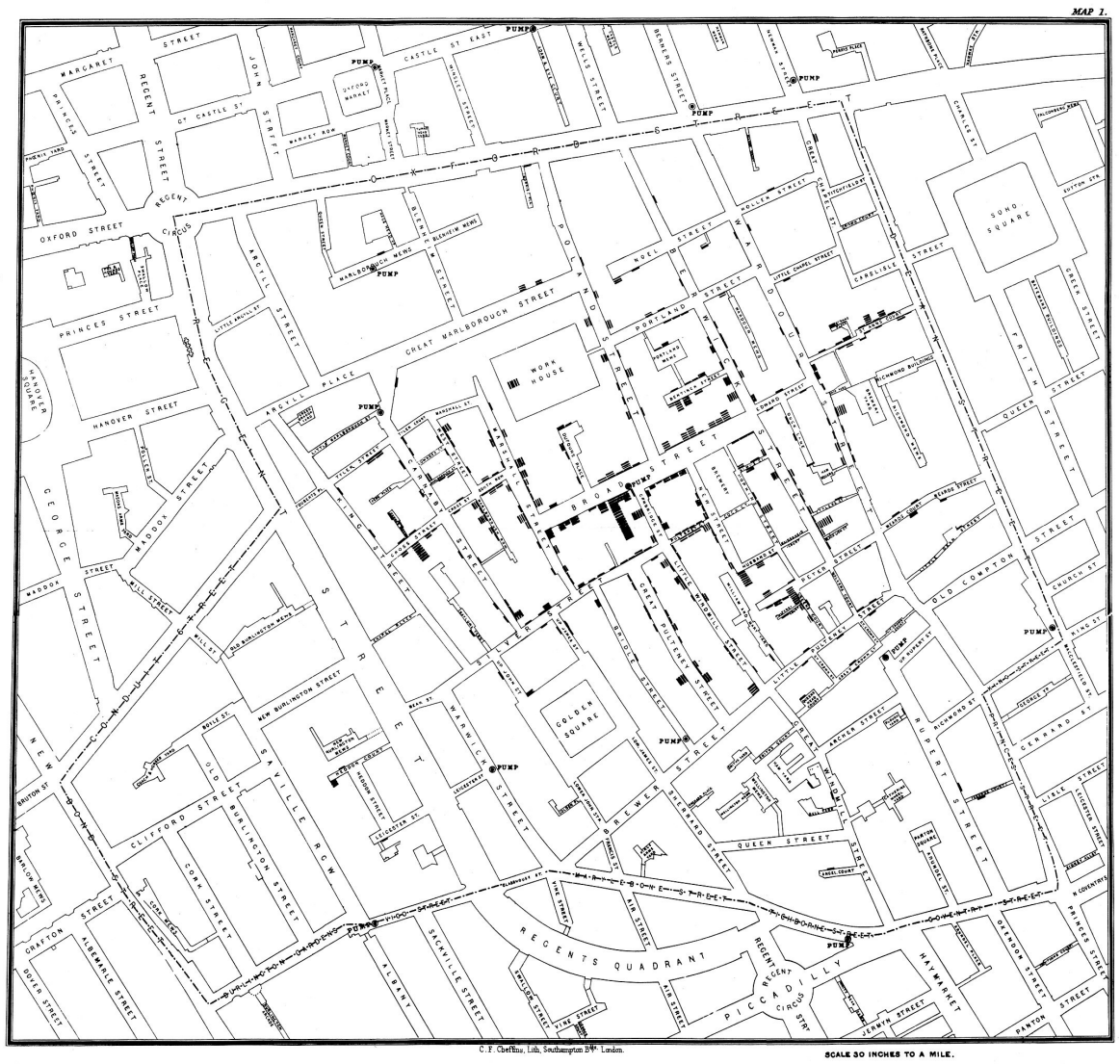
VISUALIZATION AS DATA PRESENTATION

- Data visualization aids communication by visually encoding data.
- Visualization leverages pre-attentive attributes like size, color, or position.
  - Compared to numerical information, pre-attentive attributes are processed more efficiently.
  - Pre-attentive attributes have natural interpretations, e.g. "darker" is "greater."
- Numerical information remains essential, but visualized quantities and other quantities of interest can be presented in appendix tables.

DATA VISUALIZATION AS EXPLORATION

Map by John Snow showing spatial distribution of cholera cases.

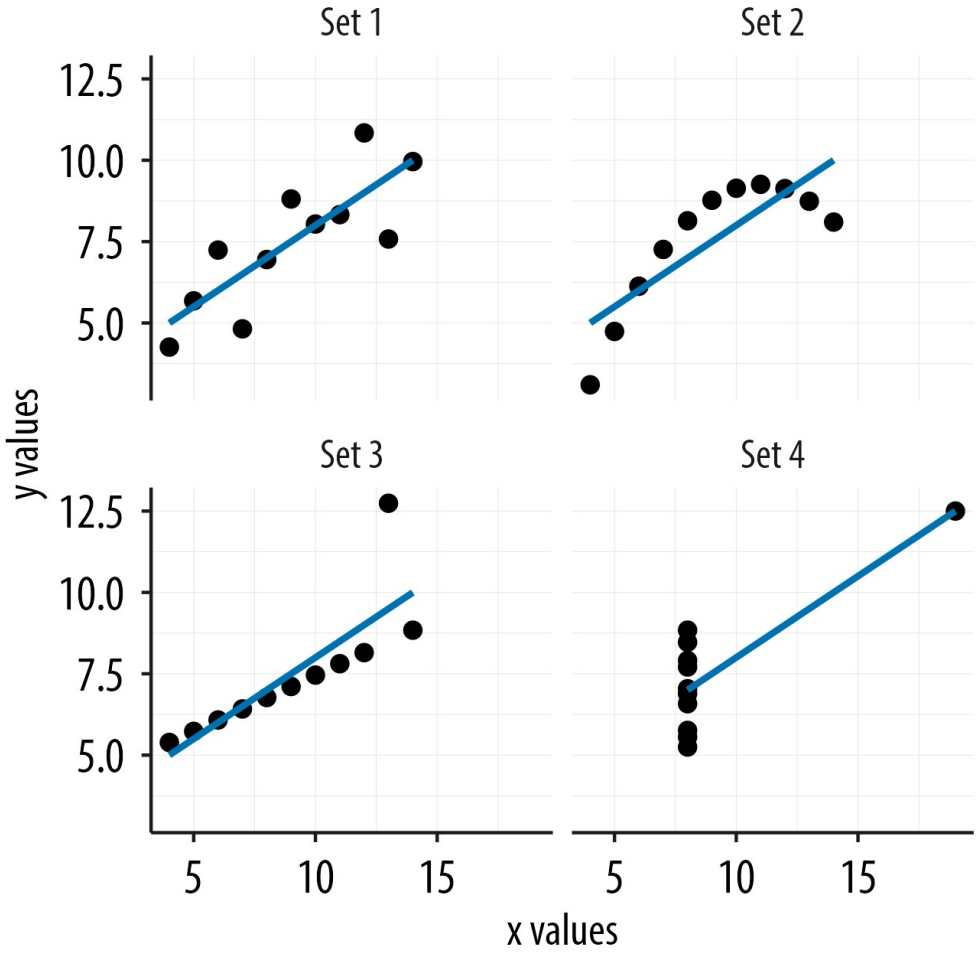
Source: Wikimedia Commons.



DATA VISUALIZATION AS DIAGNOSIS

Anscombe's quartet.

Source: Healy (2026)



GOALS OF VISUALIZATION AS DATA PRESENTATION

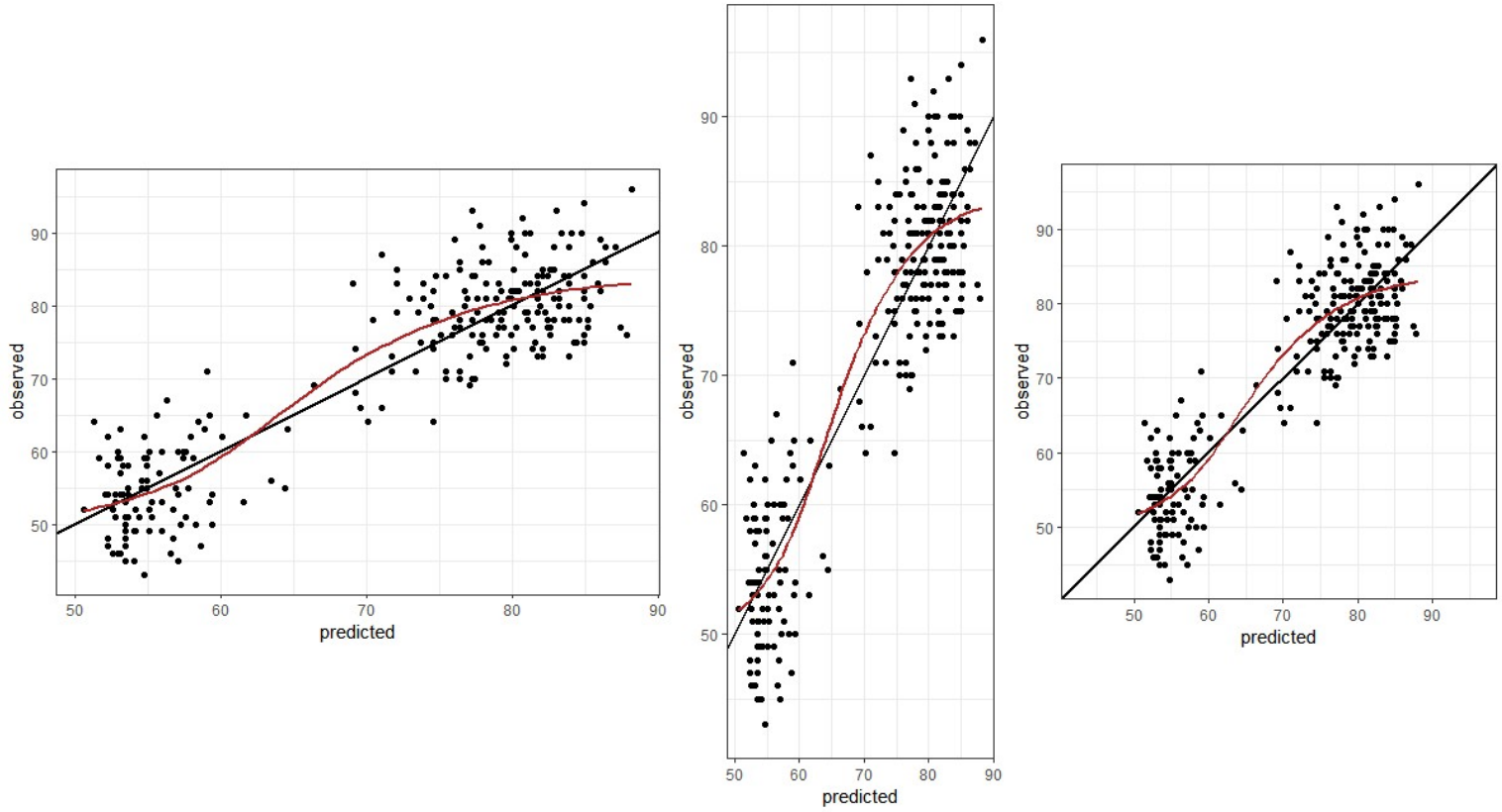
- Capture and focus attention
- Support intuitive, efficient, and accurate interpretation
- Increase transparency and bolster credibility

SELECTED BEST PRACTICES FOR DATA VISUALIZATION

- Keep it simple
- Label directly and comprehensively
- Don’t trust the defaults
- Avoid 3D
- Always report uncertainty
- Prioritize your message
- Remember your audience

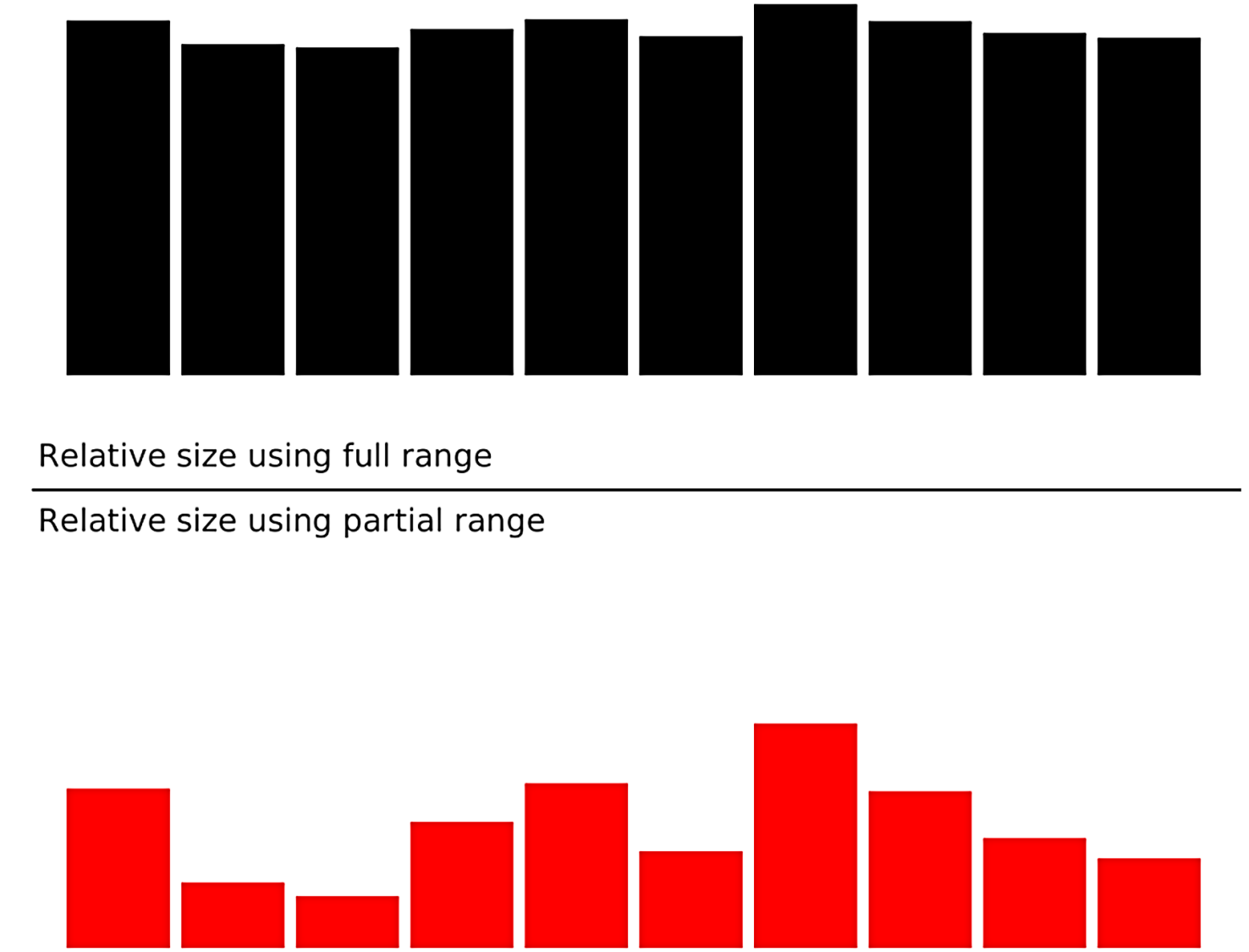
MISLEADING THROUGH ASPECT

Source: Krause, Rennie, and Tarran (2024)



MISLEADING THROUGH SCALE

Source: Rougier, Droettboom, and Bourne (2014)

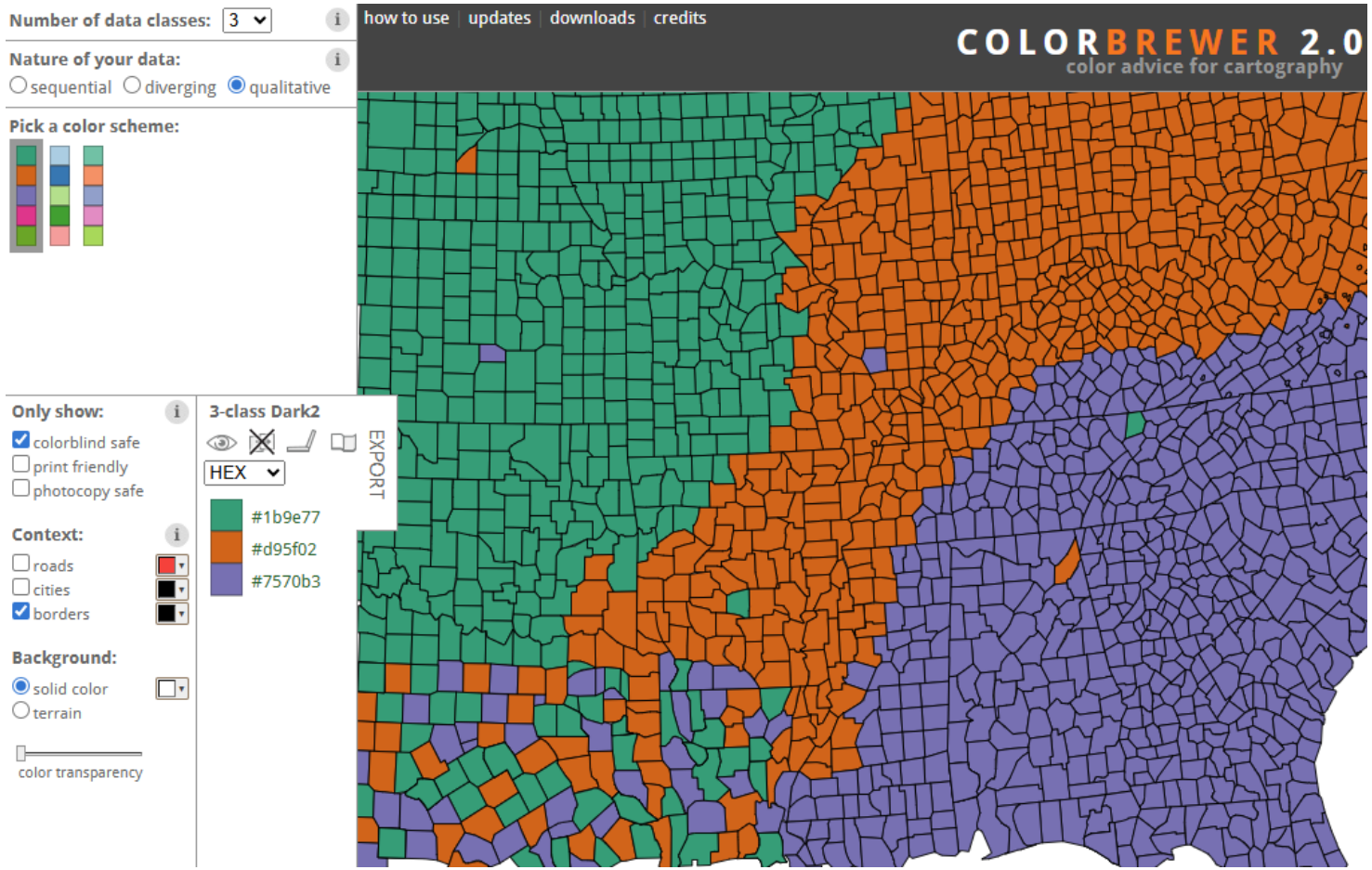


PRESENTING COMPARISONS EFFECTIVELY

- When plotting variables with different scales, avoid dual-axis charts and instead use multiple panels.
- When reporting regression coefficients for variables with different scales, use standardization or log-transformation.
- When visualizing ratio variables, scale axes logarithmically.
- When analyzing heterogeneity, plot predicted values or marginal effects rather than reporting interaction terms.

USING COLOR CAREFULLY

Source: ColorBrewer (https://colorbrewer2.org)



ENSURING ACCESSIBILITY

- Outline colorized objects
- Ensure contrast (4.5:1 for text, 3:1 for objects)
- Label directly
- Consider multiple encoding
- Use alt-text

RELATED CONSIDERATIONS

- Pre-registration
- Supplemental materials
- Replication packages
- Open access and pre-prints

DEMONSTRATION IN R

ADDITIONAL RESOURCES

- *Fundamentals of Data Visualization*, Clouse O. Wilke
  - https://clauswilke.com/dataviz/
- *Data Visualization: A Practical Introduction* (2e), Kieran Healy
  - https://socviz.co/
- *R Graphics Cookbook* (2e), Winston Chang
  - https://r-graphics.org/

QUESTIONS?

USER SUPPORT
NDACANSUPPORT@CORNELL.EDU

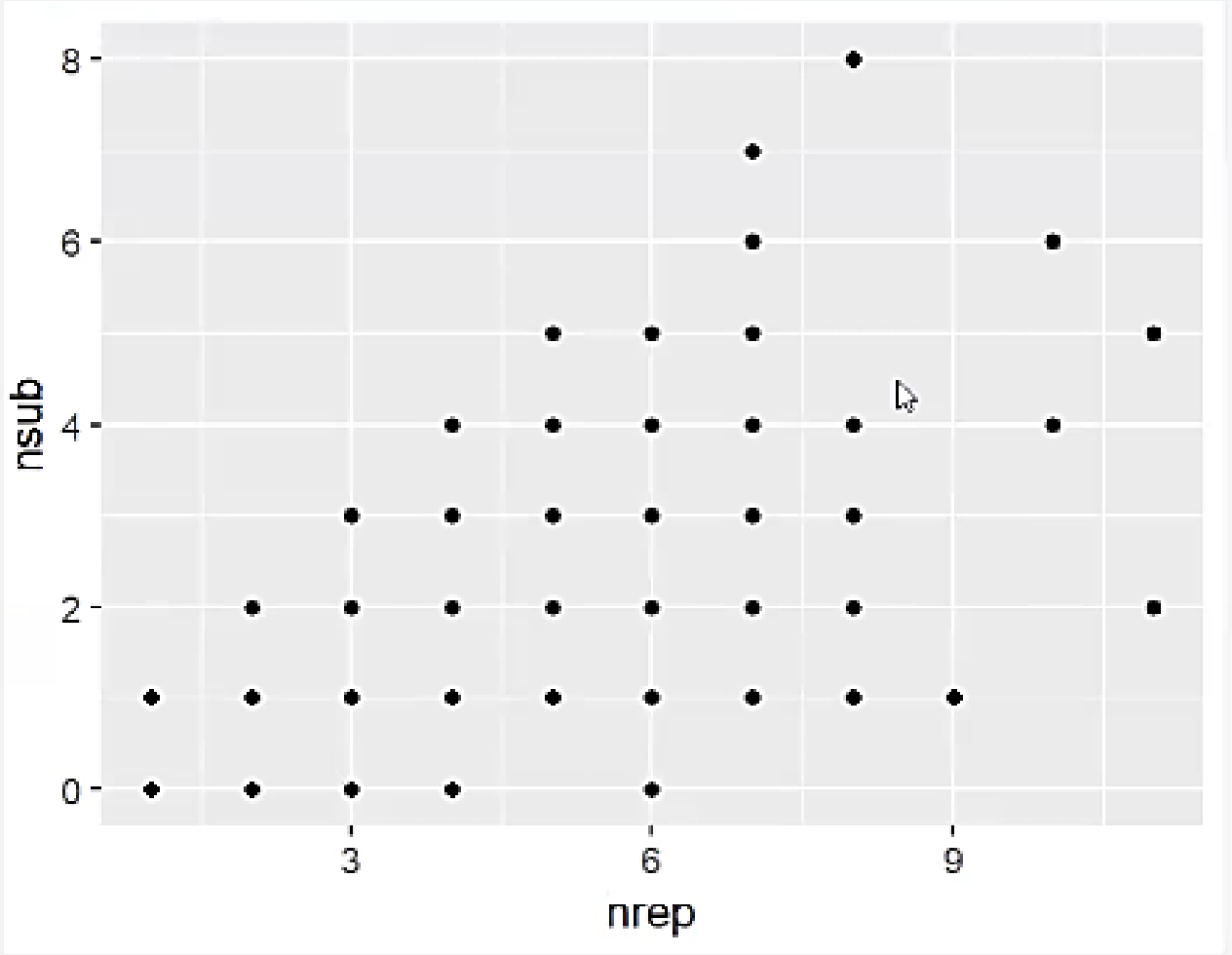
ALEX ROEHRKASSE
AROEHRKASSE@BUTLER.EDU

ALYSSA LINDSEY
ALYSSA.LINDSEY@UCSF.EDU

COMMENTS IN THE ZOOM CHAT

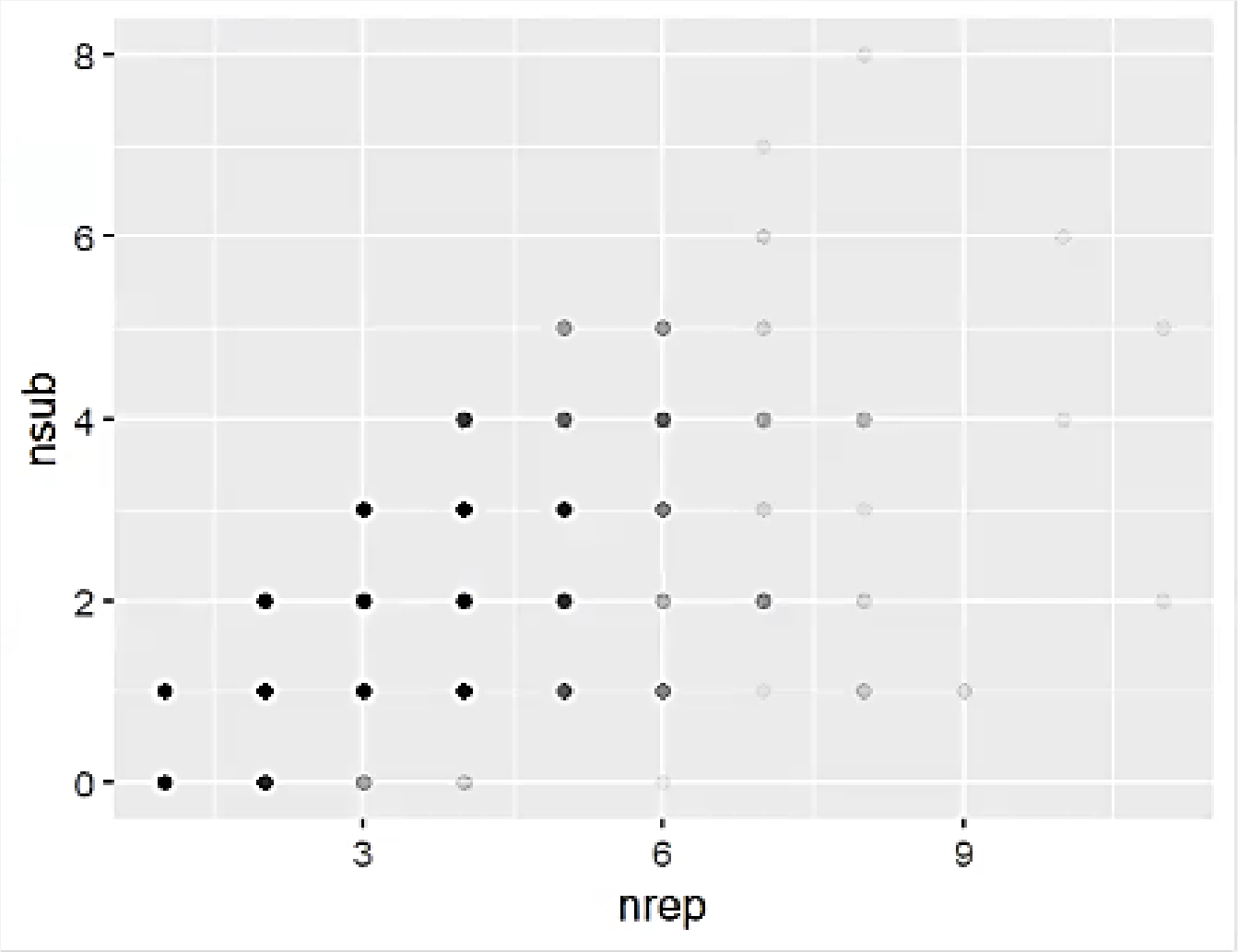
- Email the NDACAN archive at NDACAN@cornell.edu with general questions, and email NDACANsupport@cornell.edu with technical support questions.
- The NDACAN archive website https://www.ndacan.acf.hhs.gov contains the following:
  - The R code and output of this presentation.
  - Information and instructions for joining the Child-Maltreatment-Research-L (CMRL) list serve.
  - The FAQ Frequently Asked Questions about using the NDACAN archive datas including AFCARS, NCANDS, and NYTD.
  - The list of available datasets and ordering instructions.

R CODE IMAGE 1



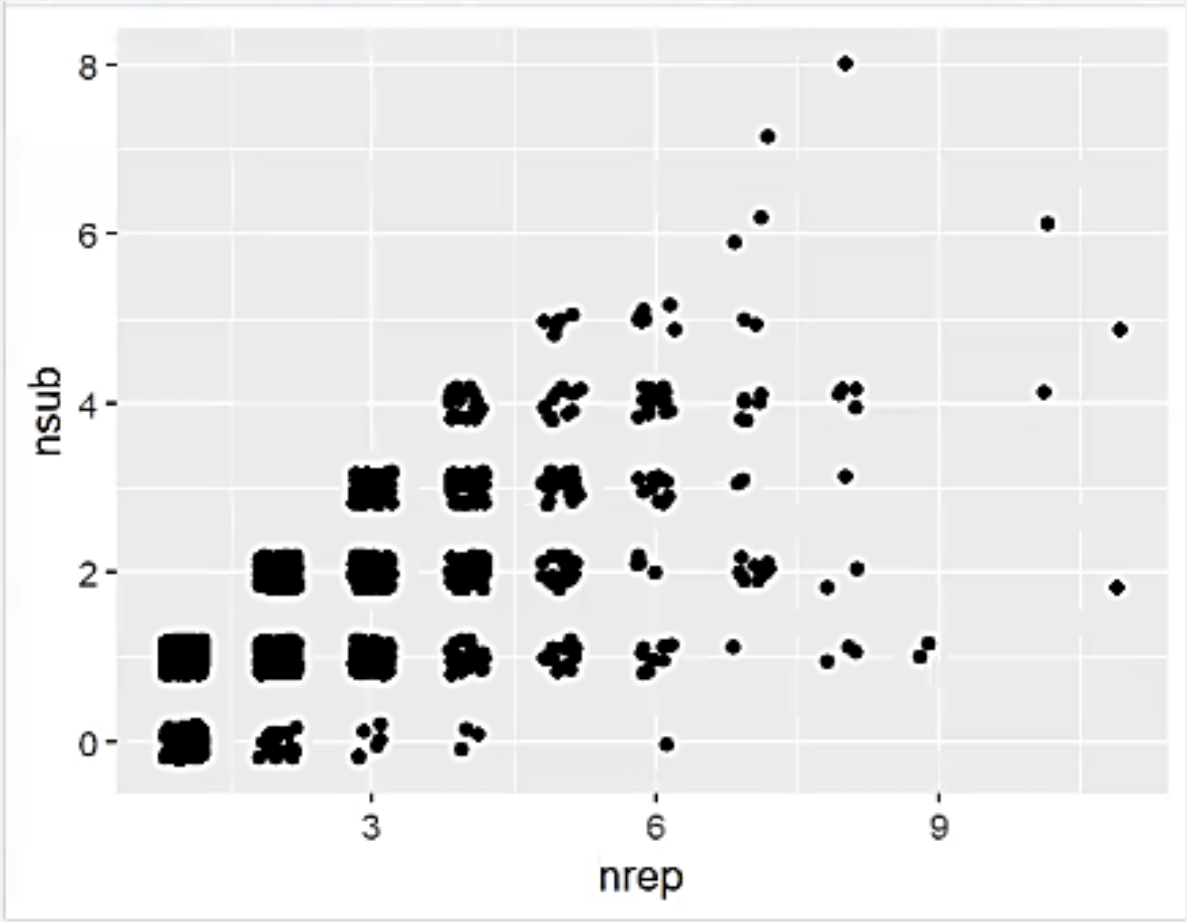
A plain dot plot with x-axis nrep (number of maltreatment reports for a child) and y-axis nsub (substantiated reports). Points show integer data, but each point is a large number of observations, and therefore the visualization does not illustrate the data distribution well.

R CODE IMAGE 2



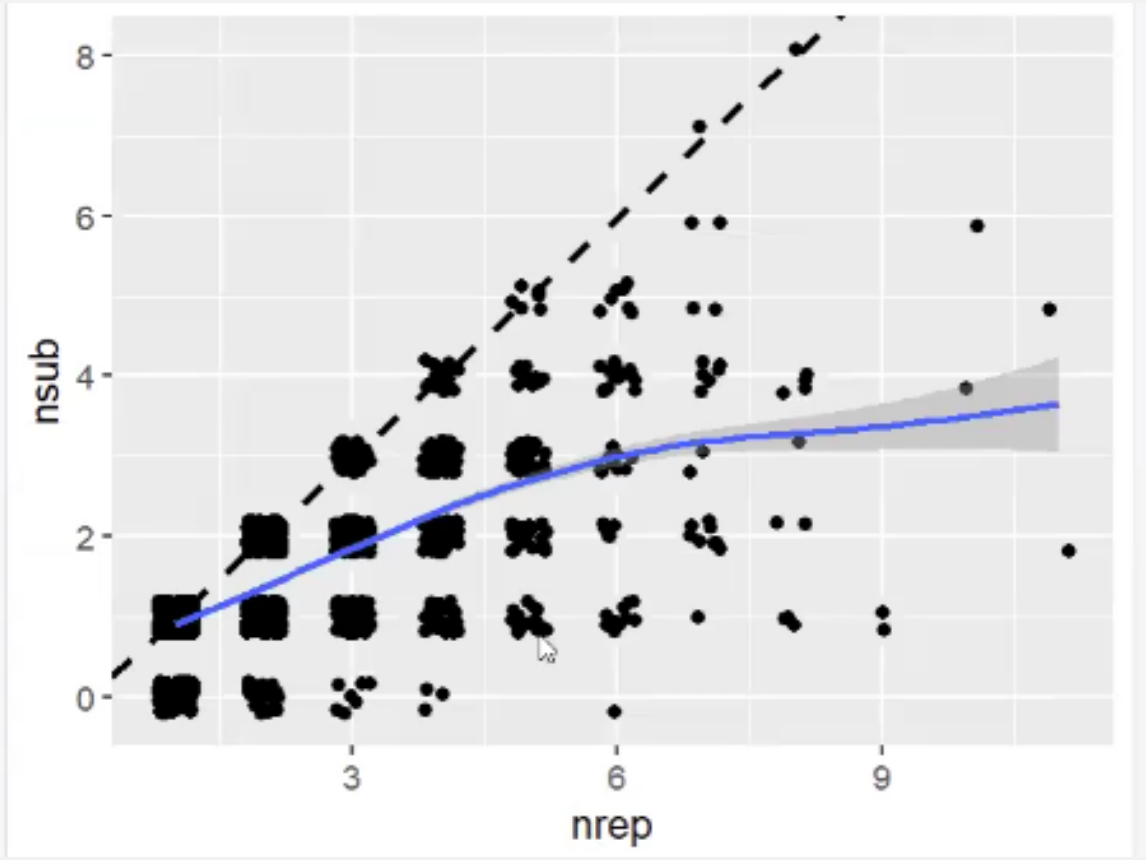
A plot with x-axis nrep (number of maltreatment reports for a child) and y-axis nsub (substantiated reports). Points are faded to visually indicate lower counts.

R CODE IMAGE 3



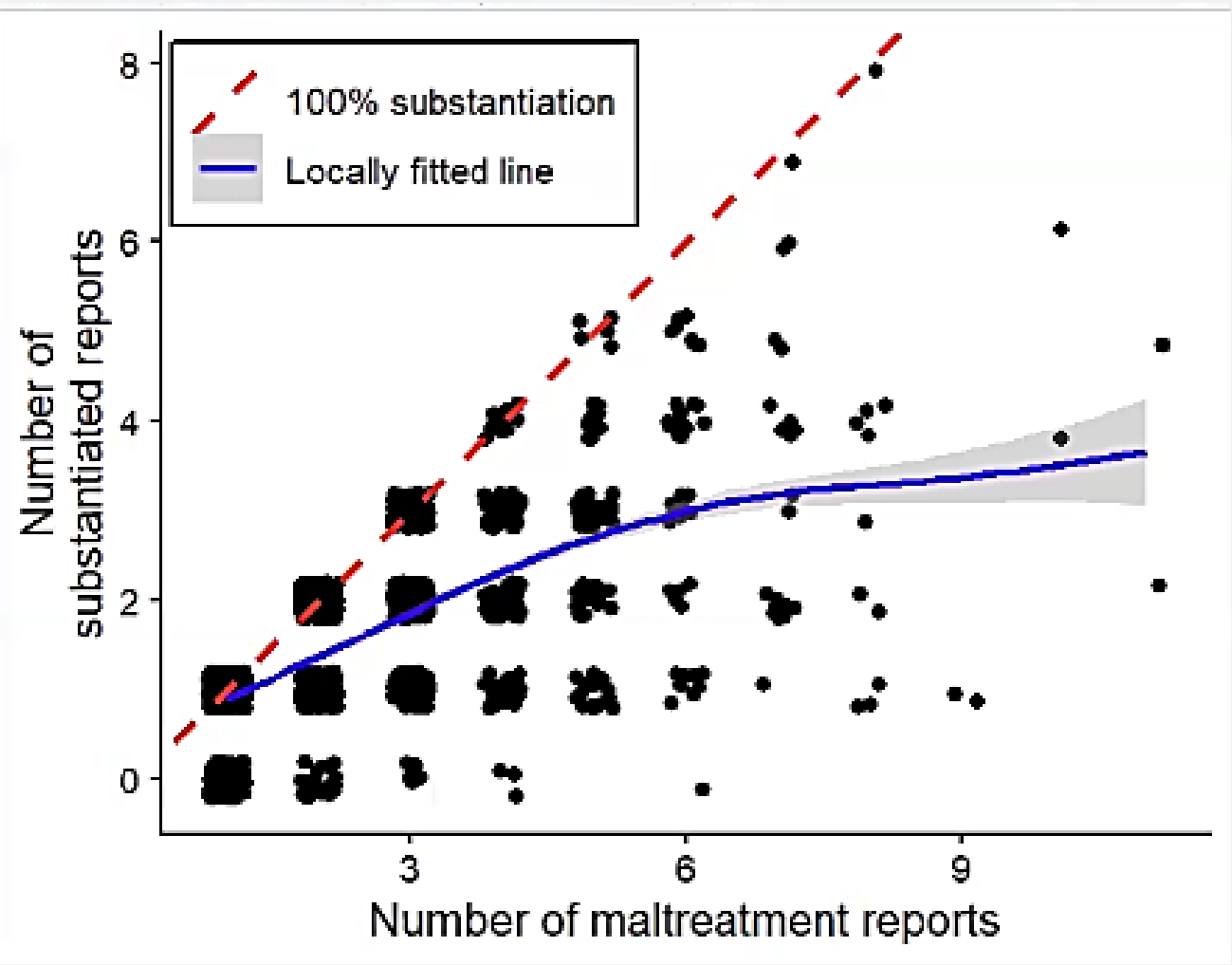
A jitter style plot with x-axis nrep (number of maltreatment reports for a child) and y-axis nsub (substantiated reports). Points are arranged in clumps to visually indicate counts/frequency.

R CODE IMAGE 4



A graphic with three geoms 1) a jitter style plot with x-axis nrep (number of maltreatment reports for a child) and y-axis nsub (substantiated reports), 2) a dashed line with slope of one overlaid to indicate whether the number of reports is equal to the number of substantiated reports, and 3) a locally estimated line of best fit.

R CODE IMAGE 5



A graphic with a legend ("100% substantiation", "Locally fitted line"), labeled axes (x-axis is "Number of maltreatment reports", y-axis is "Number of substantiated reports"), and colored lines.

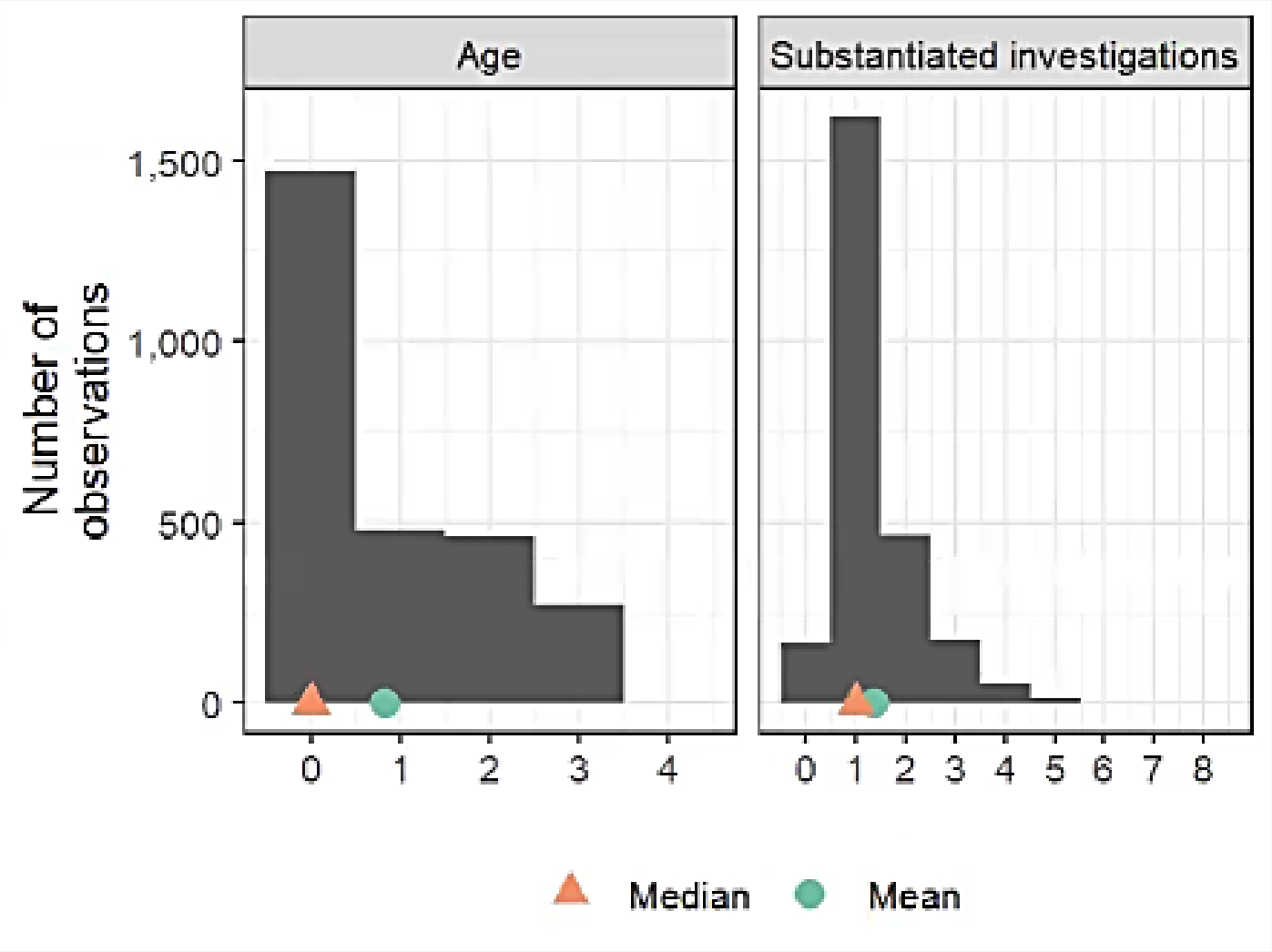
R CODE IMAGE 6

Characteristic	N = 2,689[1]
nsub	
0	164 (6.6%)
1	1,620 (65%)
2	463 (19%)
3	175 (7.0%)
4	57 (2.3%)
5	16 (0.6%)
6	3 (0.1%)
7	1 (<0.1%)
8	1 (<0.1%)
Unknown	189
ageatlatrem	
0	1,475 (55%)
1	481 (18%)
2	459 (17%)
3	272 (10%)
4	2 (<0.1%)

[1] n (%)

A basic descriptive table generated by gtsummary that shows the frequency and proportional frequency of each value of each variable (nsub is number of substantiated report, ageatlatrem is age at last removal).

R CODE IMAGE 7



The number of observations for the two variables "substantiated reports" and the "age at last removal" are shown in separate histograms with central tendencies median and mean indicated by shape and color.

R CODE IMAGE 8

	(1)	(2)
(Intercept)	-0.418	-0.889
	(0.058)	(0.022)
chpriorYes	-0.161	0.463
	(0.094)	(0.033)
fcmoneyYes	0.429	0.005
	(0.177)	(0.120)
Num.Obs.	2066	17347
Num.Imp.		5

	(0.094)	(0.033)
fcmoneyYes	0.429	0.005
	(0.177)	(0.120)
Num.Obs.	2066	17347
Num.Imp.		5

AIC	2760.4	
BIC	2777.3	
Log.Lik.	-1377.205	
RMSE	0.49	

	(1)	(2)
(Intercept)	-0.418	-0.889
	(0.058)	(0.022)
chpriorYes	-0.161	0.463
	(0.094)	(0.033)
fcmoneyYes	0.429	0.005
	(0.177)	(0.120)
Num.Obs.	2066	17347
Num.Imp.		5

A basic descriptive table generated by the modelsummary command. It summarizes two logistic regression models from Session 3 in which we predict the likelihood of substantiation or indication of a maltreatment report as a function of a child's prior frequency of reports and whether they experienced financial distress. The table contains coefficients, standard errors, a number of different statistics reporting, and goodness of fit.

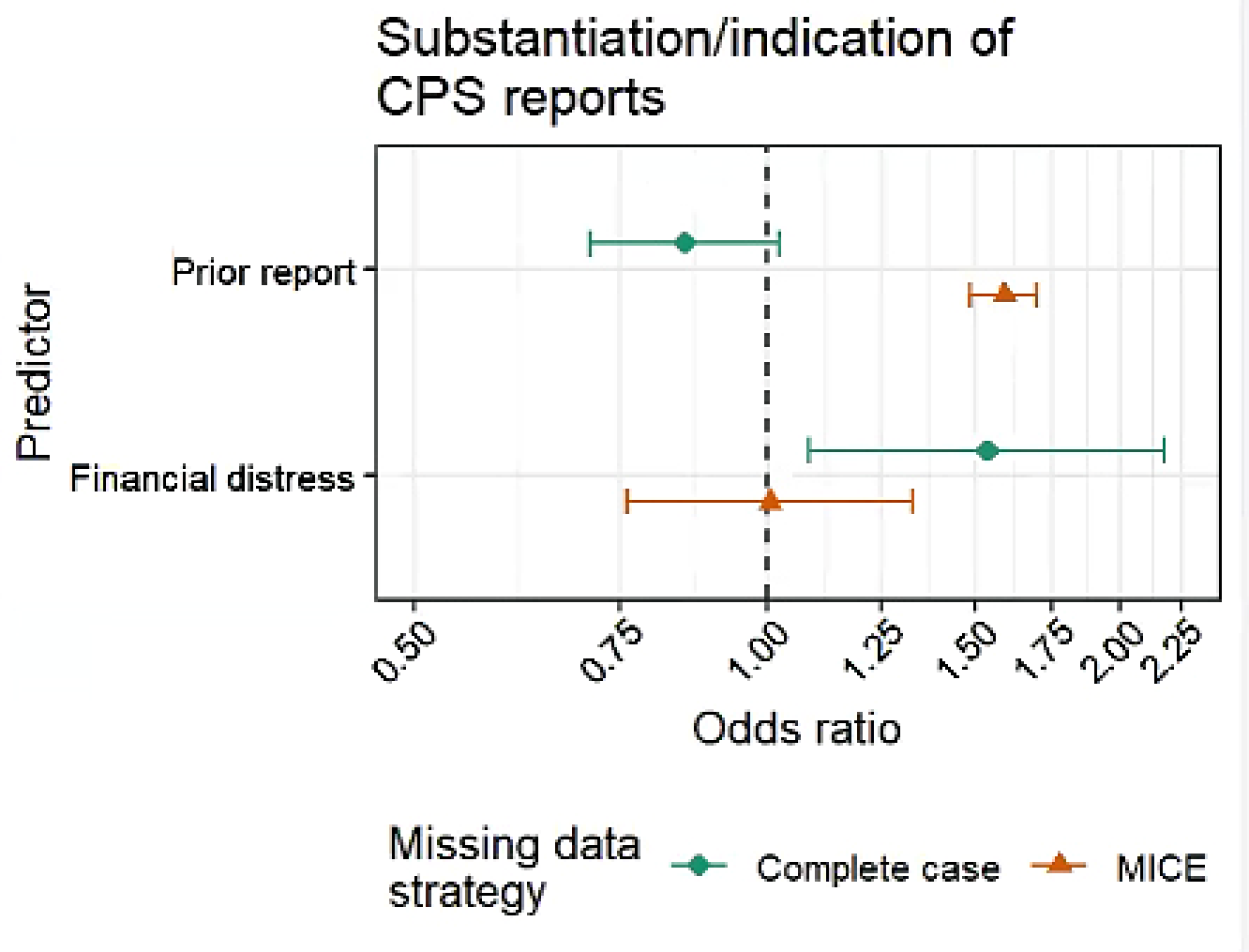
R CODE IMAGE 9

	Complete case	MICE
Prior report	-0.161+	0.463***
	[-0.346, 0.023]	[0.398, 0.528]
Financial distress	0.429*	0.005
	[0.082, 0.776]	[-0.275, 0.285]

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

A regression-results table comparing Complete case and MICE analyses for "Prior report" and "Financial distress".

R CODE IMAGE 10



A ggplot2 graphic titled "Substantiation/indication of CPS reports" which incorporates numerous visual indicators (e.g. highlight the null hypothesis as a visual anchor with a vertical line, plot point estimates as points, plot confidence intervals as error bars, log-transform axis of ratio estimand for comparison above/below 1, a colorblind-safe color palette, clearly label all plot features)