Data Wrangling in R

Putting it all together

Steps in an EDA

Set up Github project

Create local project

Link projects

Get raw data

Figure out what it is

Read in data

Pre-process it

Look at dimensions

Look at values

Make tables

Hunt for messed up values

Hunt for NAs

Plot it

Don’t fool yourself

Set up project

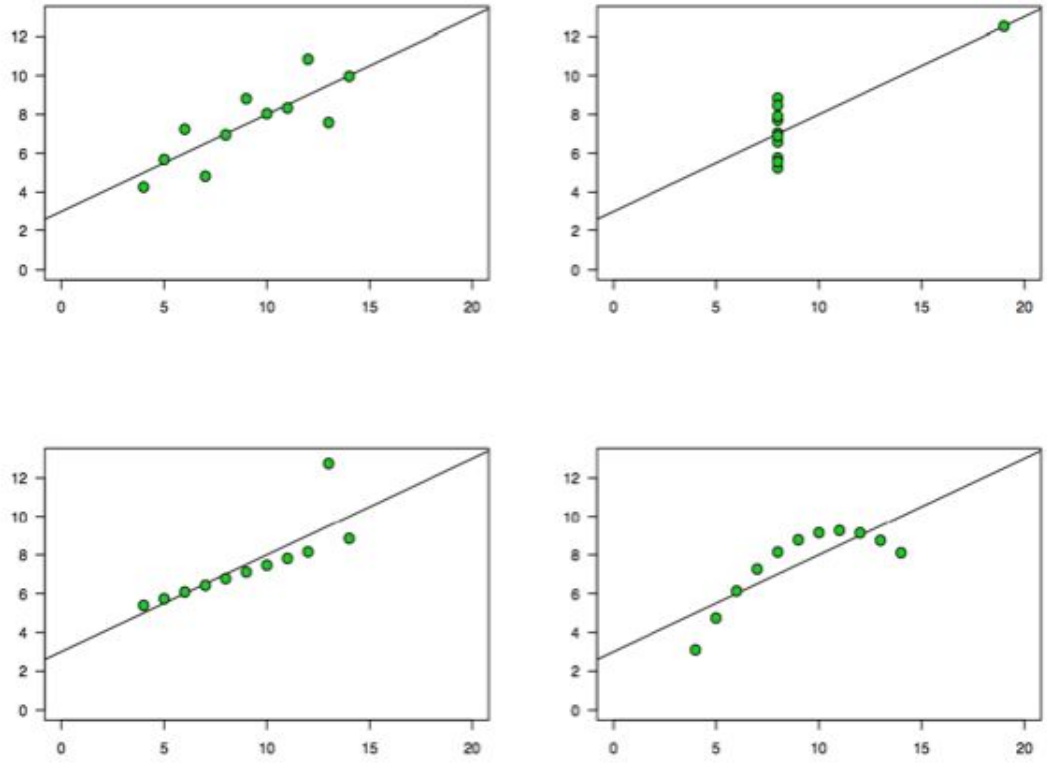
- Create repo simplystats_analysis on Github
- Clone project to rstudio.cloud
- Add data/ or API keys, tokens to .gitignore (if needed)
- Add, commit, and push
- Set up folder structure
- Add, commit, and push

Characteristics of exploratory plots

- They are made quickly
- A large number are made
- The goal is for personal understanding
- Axes/legends are generally cleaned up
- Color/size are primarily used for information

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

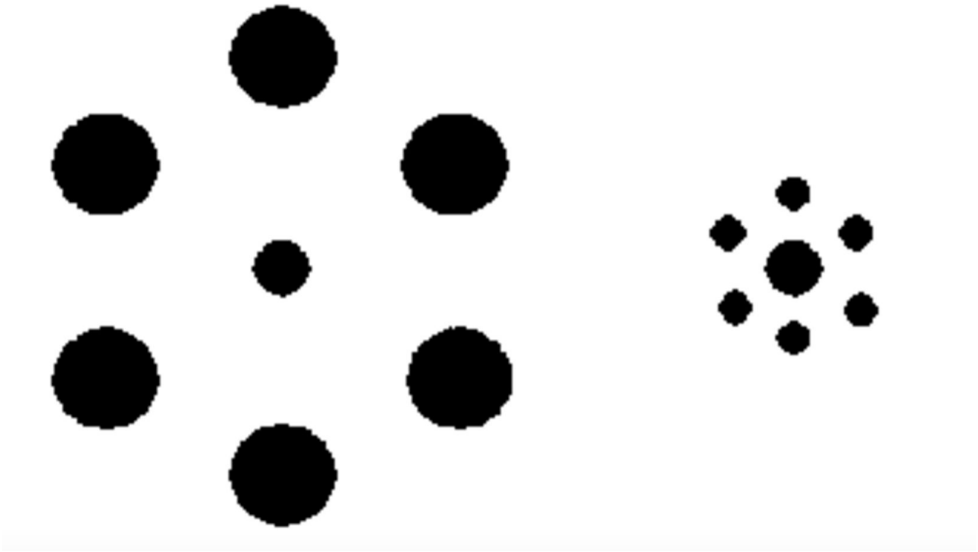
$\hat{\beta}_0 = 3.0, \quad \hat{\beta}_1 = 0.5, \quad$ p-value (slope) = 0.002, $\quad R^2 = 0.67.$

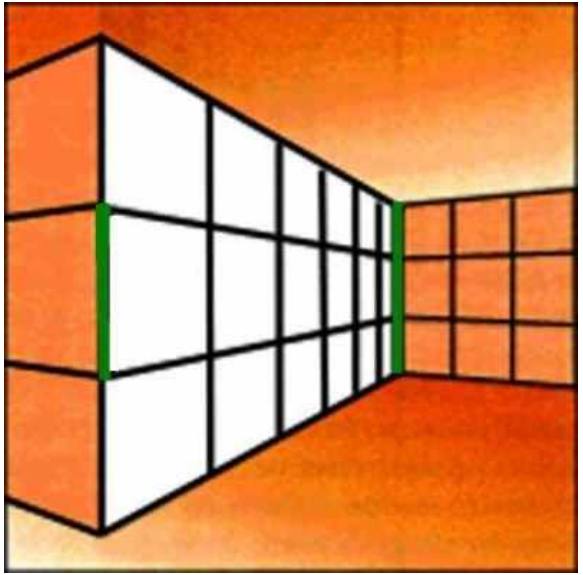


EDA

- EDA is part statistics, part psychology
- Unfortunately we (humans) are designed to find patterns even when there aren't any
- Visual perception is biased by your humanness.
- The key goal in exploratory EDA is to not trick yourself

What optical illusions teach us about plotting



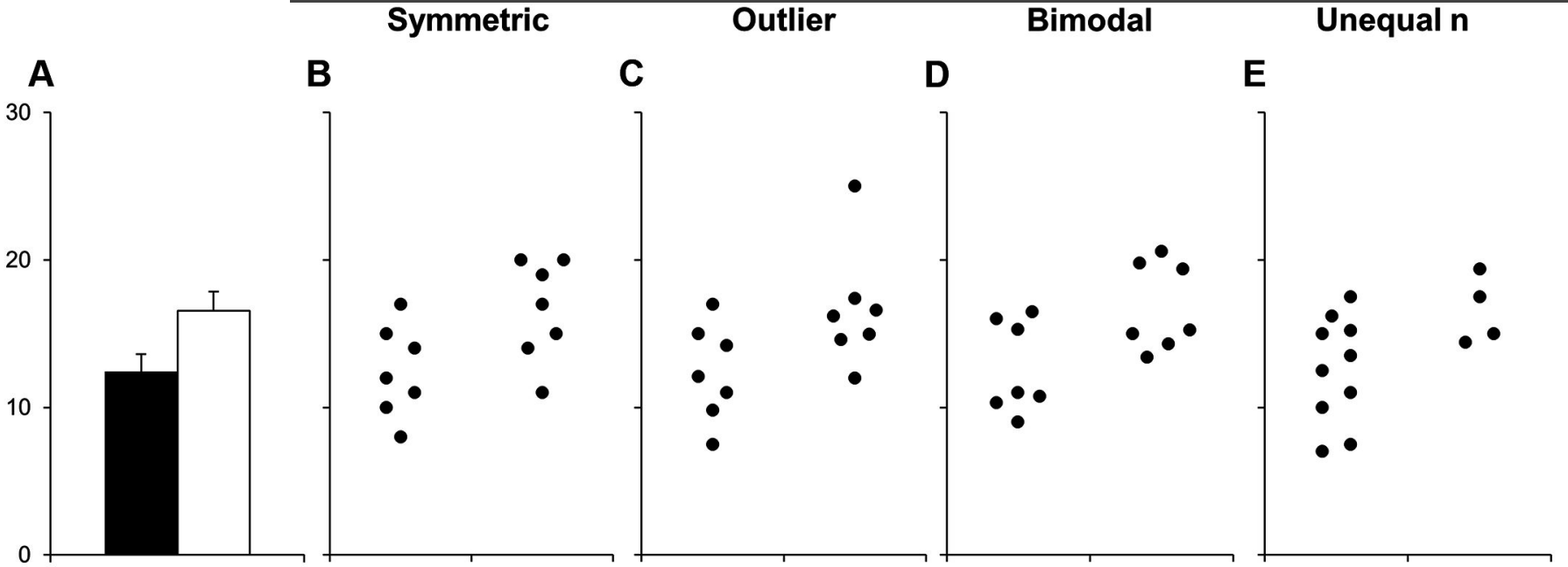


http://brainden.com/visual-illusions.htm

Basic principles

Show the data

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

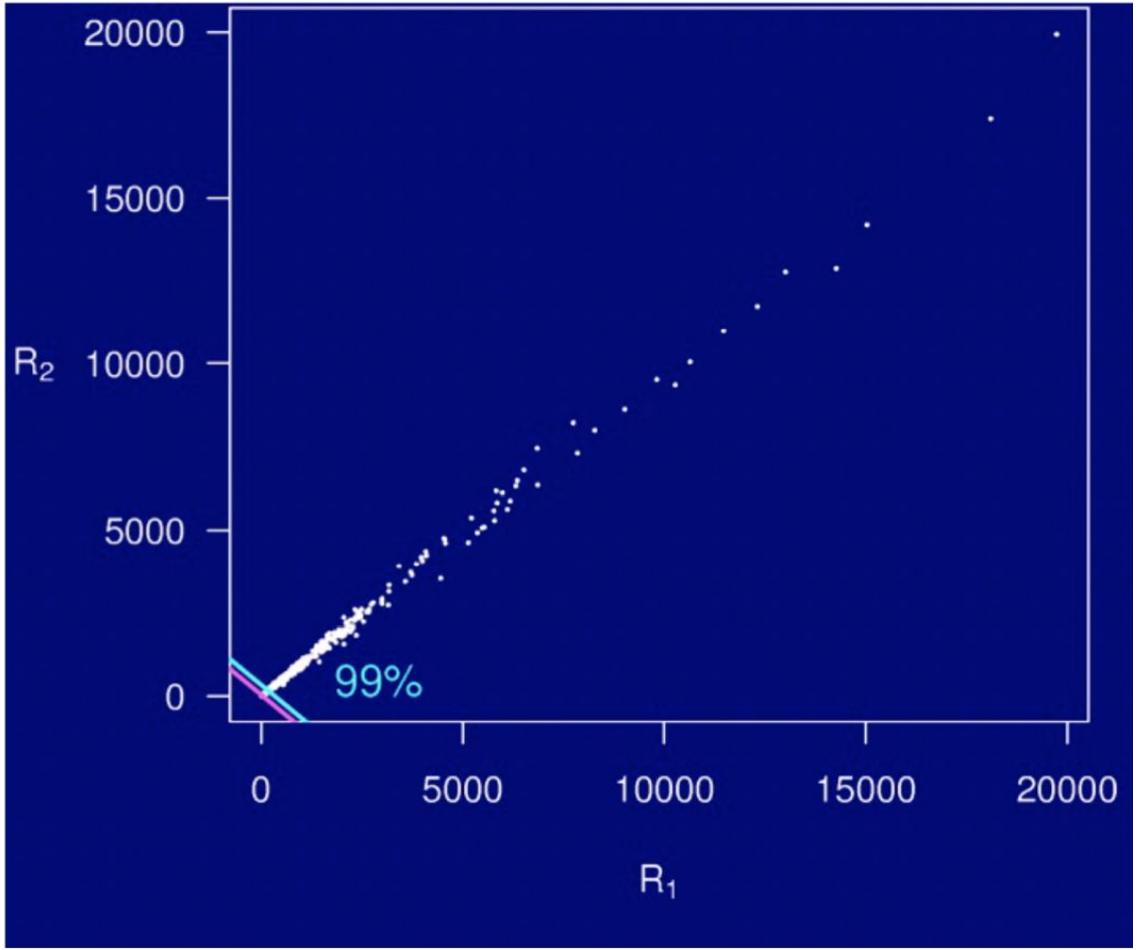


Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Basic principles

Be careful with scale

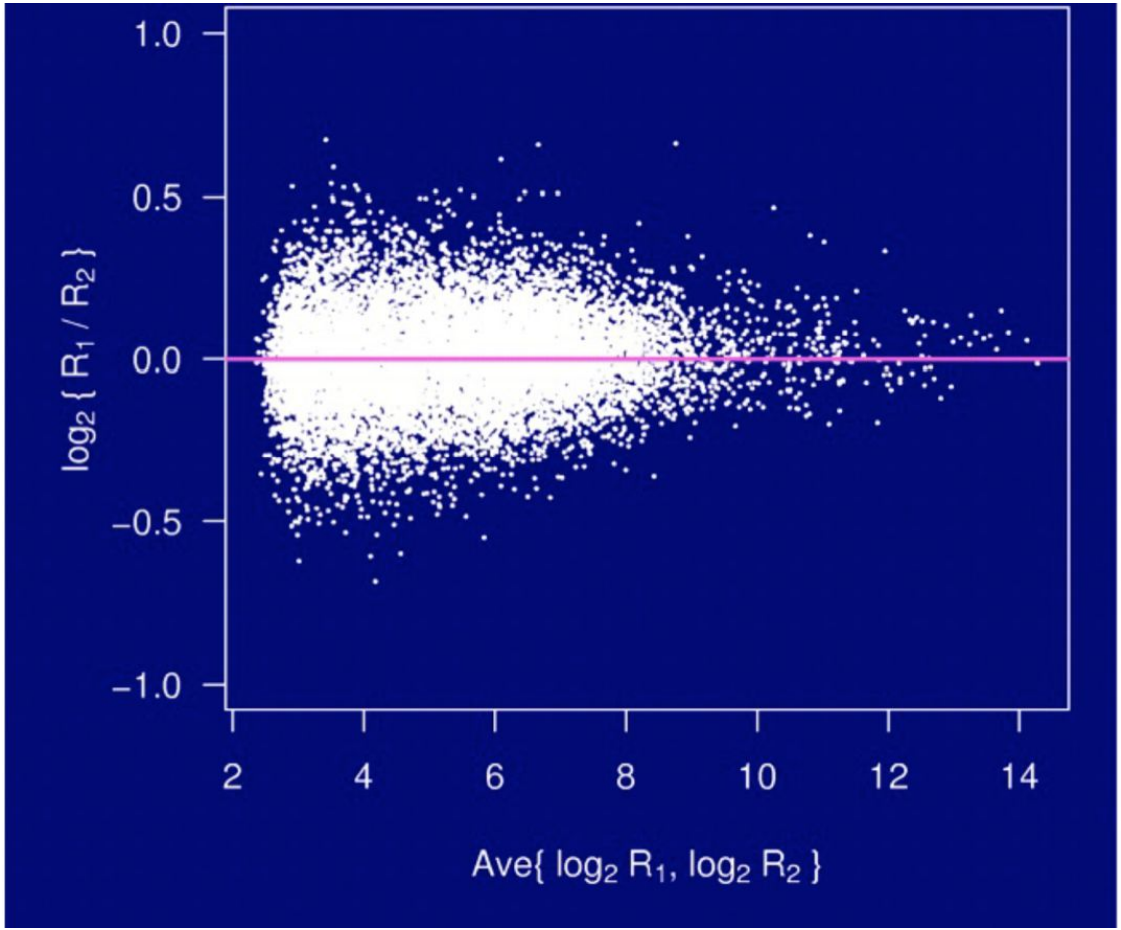
https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf



Basic principles

Compare things directly

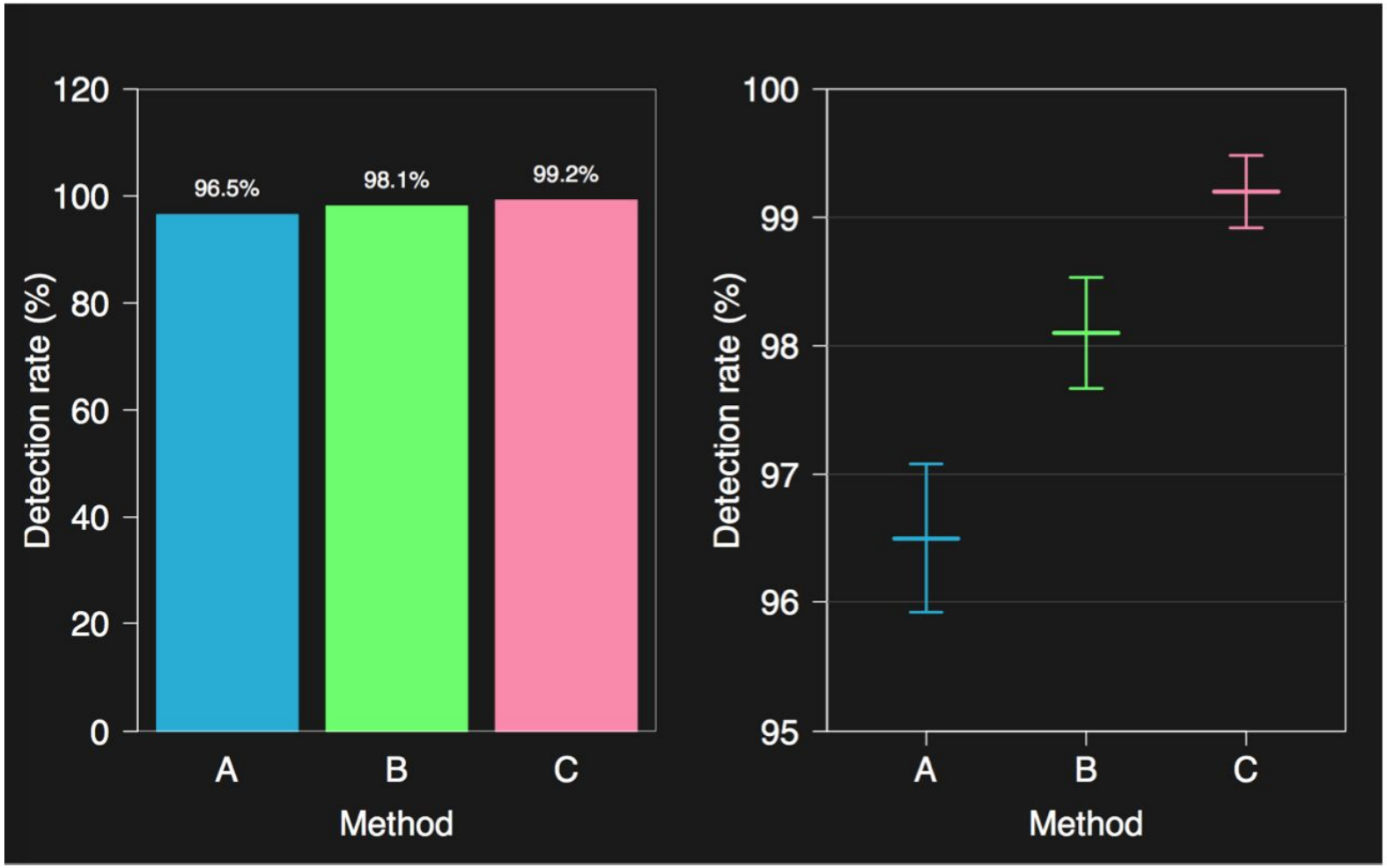
https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf



Basic principles

- Use common scales
- Start at zero

https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf



Round up

Further resources

https://r4ds.had.co.nz/index.html - section on “wrangling”

https://adv-r.hadley.nz/ - Advanced R: more on functional & object-oriented programming

https://rafalab.github.io/dsbook/introduction-to-data-wrangling.html - Great foundations for R, with more examples like the ones in this class

Open case studies: https://www.opencasestudies.org

- https://www.opencasestudies.org/ocs-bp-youth-disconnection/
- https://www.opencasestudies.org/ocs-bp-co2-emissions/

Further resources - git

https://happygitwithr.com/

https://lab.github.com/

https://www.katacoda.com/courses/git

https://git-school.github.io/visualizing-git/

https://rogerdudler.github.io/git-guide/

https://medium.com/quick-code/top-10-git-gui-clients-for-developers-b56d702579a6 - GUIs are a great way to get comfortable without memorizing commands

“It’s okay to google it”

http://stats.stackexchange.com/

http://stackoverflow.com/

https://support.bioconductor.org/

www.google.com

Three things: #1 Beware your humanness

aps
ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

General Article

Psychological Science
XX(X) 1–8

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

http://pss.sagepub.com/content/22/11/1359.abstract

Three things: #2 Be reproducible!

> “Your closest collaborator is you in six months, but you don’t respond to email.”

http://kbroman.org/Tools4RR/assets/lectures/06_org_eda_withnotes.pdf

Three things: #3 Just try it