Reproducibility in data analysis

Max Masnick, PhD

https://maxmasnick.com

What is reproducibility for data analysis?

- Same inputs, same outputs

Motivating examples

- Business context: You summarize revenue for 2017-2019, and your boss asks for the 2016 numbers. When you go to do this, you get different numbers for 2017-2019.
- Research context: You get a “revise and resubmit” for a paper, which asks you to change how a figure is formatted. But you can’t reproduce the original figure with your data.

What do you need for reproducibility?

Consistency in:

- Code
- Data
- Environment
- Process

Consistency with code

- Clear environment, run from start to finish
- Version control

Consistency with data

- **DO NOT** modify your raw source data
- Version control

Consistency with environment

- Use a package manager
  - `renv` for R
  - `pipenv` for Python
- Note the version of R/Python/etc. you use in your project's documentation

Consistency with process

- Use a well-considered folder structure
- Have a single “point of entry” to produce your analysis
- Write documentation

Consistency with process: documentation

- In-line code comments
- Descriptions for each file
- README

Version control

- https://github.com
- https://maxmasnick.com/kb/learn-git/

Mindset