Practices for reproducible data analysis

Ben Lambert[1]
ben.c.lambert@gmail.com

[1]University of Oxford

Wednesday 23rd June, 2021

Lecture content

- How to write unsmelly code
- When to test and when not to test in research
- How to structure your project for reproducibility

Outline

1. Ensuring code works

Defensive programming

"Assumes mistakes will happen and tries to find them as soon as they arise."

Data science work

Question: How do we programme defensively in:

1. Exploratory analyses?
2. Mature analyses?
3. Publications?

Exploratory analyses

Characteristics:

- Scoping analysis when first get hands on data
- Includes some basic analyses: potentially some data munging; potentially some plotting
- Eventual research path not decided
- May sit in a Jupyter / Markdown notebook

Question: should we programme defensively here?

Exploratory analyses

Steps I take:

- Graphical checks of sensibleness
- Add assertions into code chunks

Mature analysis

Characteristics:

- Scoping analysis when first get hands on data
- Includes some basic analyses: potentially some data munging; potentially some plotting
- Eventual research path not decided
- May sit in a Jupyter / Markdown notebook