Doing Reproducible Data Analysis

Mark Andrews

28/06/2020

What you will learn

- In this course, you will learn:
  1. How to create reproducible data science reports with *RMarkdown*.
  2. How to use *Git* & *GitHub* for version control and collaboration.
  3. How to create R packages, particularly for distributing code, data, and scripts for data science.
  4. How to use *Docker* containers to create reproducible computing environments.
  5. How to use *Make* to automate your workflows.

What is reproducible data analysis

- The end product of any data analysis is usually a set of tables, figures, and seemingly countless quantities, which are usually communicated in reports, slides, etc.
- This is the end product of an often long an arduous process that began with raw data.
- The aim of reproducible data analysis is to allow others, including our future selves, to be able to reproduce any given figure, table, statistical quantity.
- This is often motivated by a commitment to doing *open science*.
- It can also be motivated as a general means to doing more high quality and robust analysis.