Notes on Workflow for Mental Health and the Justice System in Durham County

Maria Tackett

2023-06-05

Table of contents

> **Note**
>
> The source code for this document is available at github.com/matackett/mhjs-workflow. Please open an issue or submit a pull request if you have any edits or additions.

Reproducibility

In *Telling Stories with Data* (Alexander 2023), Rohan Alexander describes reproducibility as the following (emphasis mine):

> *"Alexander (2019) defines reproducible research as that which can be exactly redone, given* all *the materials used. This underscores the importance of providing the* ***code, data, and environment****. The minimum expectation is that another person is independently able to use your code, data, and environment to get your results, including figures and tables."*

Using this definition as a guide, we can think about how to make our research reproducible by establishing good practices in each of the following components:

1. Data
2. Code (and analysis)
3. Organization
4. Environment

We'll address each of these components separately to help organize our thinking, but in reality all these components intertwine, so we'll often need to discuss one while discussing the other. For example, we can't establish best practices for coding without thinking about the data.

- Why is it important to do the work on this project in a reproducible way?
- What are practical considerations to keep in mind as we develop reproducible practices for this project?

Data

In the 2021 paper Datasets for Datasheets (Gebru et al. 2021), Gebru et al. name two stakeholders for a data set: *data creators* and *data consumers*. For many data sets the creator and consumer are one in the same. In either case, clear and detailed data documentation is critical for both parties to understand what's in a data set, how the data can be analyzed, and the context and limitations of any results gleaned from the data.

At a minimum, every data set should have a *codebook* (aka data dictionary) that documents what each row represents, and every every column in the data set

- Column name
- Data type
- Precise definition
- Possible values / categories
- Units (if applicable)

Another important piece of documentation is the datasheet that provides more information regarding the motivation for the data set, data collection, contents fo the data, and ethical considerations. Gebru et al. (2021) gives a list of questions we can use to create "a datasheet

that documents its motivation, composition, collection process, recommended uses, and so on." Datasheets serve a specific purpose for data creators and data consumers.

- **Data creators:** "to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use"
- **Data consumers:** "to ensure they have the information they need to make informed decisions about using a dataset"

- What information from a datasheet might be useful and important for our project?
- Use the questions from Gebru et al. (2021) to build a template datasheet that could be used as documenation for data sets in our project. The datasheet doesn't need to include all the questions from the paper, but be sure it includes at least one question from each section: Motivation, Composition, Collection Process, Preprocessing/cleaning/labeling, Uses, Distribution, Maintenance.

Code

R Scripts

- R script for infrequent tasks, i.e. creating data sets or writing functions
  - Reference the R for Data Science Chapter
  - Use comments throughout to explain what is happening

Quarto documents

- Literate programming for analysis
  - Use quarto
  - Reference the R for Data Science Chapter on Quarto
  - Use narrative to describe analysis steps and results. You want it detailed enough that someone else could understand your analysis process and results (or you can understand them if you read it a year later)
  - Don't need to write fancy prose as you're working on the analysis - can use bullet points.

  - Nice thing about Quarto is that you can easily make reports and presentations from the same document

General tips

- General tips
  - Naming conventions - R for Data Science Chapter
  - Code Style - Tidyverse style guide

> Tip
>
> You can use the visual editor to make report writing in RStudio more similar to Google Docs / MS Word. Learn more about Quarto and visual editor.

Version Control

- What is version control
- Why it's important
- Telling Stories with data?
- How it works for our project without Github

Organization

- Telling stories with data file organization.
- File names mean something - R for Data Science
- Use a project for each new analysis

Environment

Putting it all together

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2021. "Datasheets for Datasets." https://arxiv.org/abs/1803.09010.