WJP - R Coding Handbook

Carlos A. Toruño P.

October, 2022

Table of contents

Preface

Figure 0.1: Zacatal by nicaraguan painter Raúl Marín

This coding handbook is a continuous work mantained by the Data Analytics Unit (DAU) from The World Justice Project (WJP) with the aim of unifying the different aspect of the carried by the unit. In this book, you will find not only general guidelines but also several issues that will help the reader to understand and contribute in our tasks.

The handbook cover aspects related to the general workflow, the coding process and the visualization guidelines used by the team. As mentioned earlier, this book is a continuous work in progress and, as such, the rules and guides are not written in stone but rather, subject to improvements that every member of the team is open to discuss and include in this handbook.

To learn more about the work of The World Justice Project, please visit the official website.

1 Workflow

In this chapter, we will cover the basic guidelines and issues related to the DAU workflow when programming with R and R Studio. Therefore, in order to assimilate the content in this chapter (and the entire handbook) we will require you to have the following:

- An intermediate knowledge of R language.
- R Studio is installed in your computer, however, you are free to use any other IDE of your choice.
- A basic knowledge of Git.
- A GitHub account.
- Access to the DAU's SharePoint
- Although is not required, we strongly advise you to install GitHub Desktop.

If you are new to R, I would recommend you to check the Hands-On Programming with R book written by Garrett Grolemund. If you are already comfortable with the R programming language but you need an introduction on how to analyze data using it, I would suggest you to read the R for Data Science book edited by the R4DS team.

1.1 SharePoint

All files used by the DAU are stored in the Data Analytics folder within the organization's SharePoint. And, as a rule, you are required to work and modify these files directly from this online directory. In order to achieve this, you will need to download the OneDrive app in your computer and sign in using your WJP account. If you are already using One Drive with your personal account, you can add an additional work account which will function parallel to your personal account in your computer. For more information, see the following website.

Given that the DAU SharePoint is a directory created and administered by the organization, you will have to sync this directory to your own OneDrive WJP Account. In order to do this, go to the Data Analytics directory using OneDrive Online and, once you are there, you will see the ***Add Shortcut to My Files*** option in the main panel headers. Follow the instructions and sync it to your WJP folder in your computer.

By syncing your work with the SharePoint, the whole team is going to be able to see and review changes as they go. However, if more than one person is working on the same file, it is likely that this working flow will produce several mistakes. Therefore, we need a version control tool that will allow us to modify and collaborate in team projects without these kind of issues. Given that most of our work is focused in coding, we use Git and the GitHub platform for this.

1.2 Git

Git is a free and open source software that allows users to set up a version control system designed to handle projects. Given the nature of its features, it is normally used for collaboratively developing code and data integrity. GitHub is a website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code. For a gentle introduction to Git and GitHub, see the following post published by Kinsta. For a more in depth introduction, please refer to the GitHub documentation.

Using the Git features allow us to simultaneously work on the same project and even in the same code without worrying about interfering with other members of the team. As a rule, every project carried by the DAU has a code administrator who is in charge of setting up the GitHub repository and add other members of the team as project collaborators. Additionally, the code administrator is in charge of setting the *main branch* and the initial structure of the code (see the (**data-management?**) section on this chapter). It is required that the GitHub repository have its *main branch* in its respective SharePoint folder.

We use convergence development[1] to collaboratively code in the same project. For this, it is highly recommended that each team member works on a separate *branch* and, once the data routine is done, the auxiliary branches can be merged into the main branch of the repository. Collaborators that are not the code administrator have the option to clone the GitHub repository in their computer in a local directory that it is not sync to the SharePoint and work in their respective branch from a local copy outside the SharePoint. In other words, it is only the functional final version contained inside the *main branch* the one that it is going to be sync in SharePoint.

Important: GitHub is used to keep track of the code we use in each project. Under no circumstance, we will include the data sources in the online repositories.

[1] For more information about convergence development and branches, we advise you to refer to this article from Pluralsight.

1.3 Projects

When you load a data set or source an R script, you will have to set up the working directory where these files are located in your computer. However, the path to this working directory is quite different for all the members of the team. R Studio allow us to enclose all of our analysis, code and auxiliary files into a project.

A project is a feature that allow us to work with the analysis we are carrying without having to worry about where does these files are stored or who is working on them. Say goodbye to `setwd("...")`. Besides managing relatives paths, R projects allow users to keep a history of actions performed and even keep the objects in your environment. Because of this, projects are the cornerstone of our work when performing analysis with R.

As a rule, every project has a file named `project-name.Rproj` in its root directory and open it should be the first action when working on a project. For more information on working with R projects, refer to the Workflow section from the R for Data Science book.

1.4 Data Management

The University of California San Diego (UC San Diego) has a Data Management Best Practices that reviews common guidelines for managing research data. In this handbook, I will focus on two topics mentioned by those guidelines: Data Organization and Data Documentation.

1.4.1 Data Organization

The data organization involves three important elements: filing system, naming conventions and data granularity.

A filing system is basically the organization structure (directories, folders, sub-folders, etc) in which files are stored. There are no standard rules about how this should be done. However, the chosen filing system needs to make sense not only to the person currently working on a given project but to anyone going through these files in the future. As a rule, each project would have the following sub-folders:

- Code: Depending on the complexity of the project, you could choose to create separate directories within the Code folder for Stata, R or Python files.
- Data: Depending on the complexity and nature of the project, you could choose to create separate directories within the Data folder for RAW, INTERMEDIATE or CLEAN data sets.

- Outcomes: The outcomes of the project might have several different formats. For example, images could be in PNG and/or SVG format, Reports might be created in PDF, some tables might be exported as TEX files, etc. The outcomes folder should have a separate directory for each one of these formats.

In some cases, the creation of a PDF report is key, for example, the Regional Country Reports. For these projects, we strongly advise to create a separate directory to store the code files used for the report. At the moment of writing this handbook, these reports are created using R Markdown, but a migration to Quarto is feasible in the future.

- Markdown/Quarto

In relation to the naming conventions, these are a set of rules designed to complement the filing system and help collaborators in understanding the data organization. Each project have the flexibility to use a specific set of naming rules to use in the filing system. However, there are a few general rules to note:

- Use descriptive file names that are meaningful to you and your colleagues while also keeping them short.
- Avoid using spaces and make use of hyphens, snake_case, and/or camelCase.
- Avoid special characters such as $, #, ^, & in the file names.
- Be consistent not only along the project but also across different projects. If all different data files and routines are named in the same way, it's easier for you to use those tools across projects and re-factor routines.