VIII LACSC
VIII Latin American Conference on Statistical Computing
San José, Costa Rica July 15-19, 2024

TUTORIAL

Reproducible Data analysis

Marcela Alfaro, University of California, Santa Cruz

Reproducible data analysis is essential to ensure the transparency, collaboration, and validity of the results of any research project. Additionally, it is a strategy to avoid repeating internal processes in research or a private company and to document the various analyses that are periodically conducted in an organization. In this tutorial, we will describe the fundamental techniques to create and share reproducible data analyses using the R programming language, the version control package Git, and the collaboration platform GitHub, as well as how to use AI tools to document the code and make it more readable.

Session 1. Introduction to Reproducible Data Analysis (30 minutes)

- What is reproducible data analysis and why is it important?
- Advantages and challenges of reproducibility in data analysis.
- Key tools and concepts: Quarto and version control.

Session 2. Fundamentals of Quarto (1 hour)

- Integration of R code and narrative text.
- Structure and syntax of Quarto.
- Generation of dynamic and visually attractive reports.
- Incorporation of interactive graphs and tables.

Session 3. Version Control with Git and GitHub (1 hour)

- Introduction to Git: tracking changes and collaboration.
- Creating and cloning repositories on GitHub.
- Integration of Quarto and Git for version tracking.
- Collaboration on data analysis projects.

Session 4. Practice and Use of CoPilot Tools (30 minutes)

- Development of a short data analysis project from scratch.
- Using coPilot tools to document code.

Requirements:

Before starting this workshop, please follow the following steps

- Download or update Quarto, R, RStudio (v2023.06 or higher), and git on your computer.
- Create an account on Github, if you already have one, have your username and password ready for the course.
- Download and install Copilot agent from your R Studio.

Marcela Alfaro is assistant Teaching Professor at the University of California, Santa Cruz. She is interested in developing novel statistical methods to address scientific questions related to the environment, and turning those experiences of interdisciplinary collaboration into teaching methods for statistics and data science. Her areas of application include climate models, biophysics, spectrophotometry, among others.