Tools for Reproducibility in Data Intensive Social Science Projects

Prof. José Manuel Magallanes, PhD.

August 2022

Tuesdays and Thursdays/ Terças e Quintas
August 9 - 25/9 – 25 de agosto de 2022
13:00 - 16:00

Class Room: TBC
Office: **1417**
E-mail:**jose.magallanes@fgv.br**

Course Description

This course presents a set of different tools needed to carry out data intensive social science projects with a reproducible approach. The contents are organized around four major components:

- Setting up cloud repositories.
- Understanding Version control.
- Authoring with Latex and Markdown.

- Managing references and authorship.
- Use of R and Python in Data Science Projects.

There is no pre-requisite for the course.

Educational Objectives

1. Have the participants become effective users of Data Science tools.
2. Have participants understand how to integrate R with other tools to produce reproducible social research.

Research Objectives

1. Guide the participants towards carefully planing a data science reproducible research project.
2. Choose a case for reproducible data science project.

Applications required

Students have to install the following software in their computers:

- R (choose according to your Operating System):
  https://cran.r-project.org/
- RStudio (choose according to your Operating System):
  https://rstudio.com/products/rstudio/download/
- Anaconda (choose according to your Operating System):
  https://www.anaconda.com/products/distribution
- GITHUB:
  Get an account at https://github.com/, and then download the desktop app from https://desktop.github.com/.
- ZOTERO:
  Get an account at https://www.zotero.org/, and then download the desktop app from https://www.zotero.org/download/.
- ZENODO:
  Get an account at https://zenodo.org/.

Community Conversation Norms

- Listening carefully and respectfully
- Sharing and teaching each other generously
- Clarifying the intent and impact of our comments
- Giving and receiving feedback in a "relationship –building" manner
- Working together to expand our knowledge by using high standards for evidence and analysis

Changes to the Syllabus

The professor reserves the right to make changes to the syllabus during the workshop. The professor will notify students immediately by email and in class if any changes are made.

Grading

Grades consider are based on exercises; which will be given along the workshop.

Working in groups

This course allows and encourages working in groups. You should have decided who will be in your before the end of the **third** classes.

Course Schedule

Week 01, 08/08 - 08/12: Setups and Intro to Coding.

1. Creating accounts and Installation
    - GitHub (web and desktop).
    - Google Drive and Dropbox and GitHub.
    - Zotero.
    - Zenodo.
    - R and Python. RStudio and Anaconda. R and Python without installation.
2. Collecting, storing and accessing data.
    - Collecting data tables from the web in R and Python.
    - Storing data files into the web in R and Python.
    - Accessing data files into the web in R and Python.

Week 02, 08/15 - 08/19: Data Pre-Processing and Exploration

1. Using Python:
   - Cleaning values.
   - Formatting tables.
   - Integration of tables.
2. Using R:
   - Exploring categorical data.
   - Exploring numerical data.

Week 03, 08/22 - 08/26: Integrating and Publishing

1. Publishing
2. Preparing bibliography file.
3. Publishing as webpage using markdown and R.
4. Publishing as a printout paper using Latex and R.
5. Publishing as a dahsboard using R.
6. Creating DOI for Project repository.

Recommended Reading

- Books:
  - Kitzes, Justin. Practice of Reproducible Research - Case Studies and Lessons from the Data-. University Of California Press, 2017.
  - Magallanes, José Manuel (2017). Introduction to Data Science for Social and Policy Research. Cambridge University Press.
- Links
  - GITHUB: https://docs.github.com/pt
  - RSweave: https://support.rstudio.com/hc/en-us/articles/200552056
  - ZOTERO: https://guides.library.sc.edu/zotero
  - RStudio: https://bookdown.org/gboccardo/manual-ED-UCH/

Acknowledgements

The material from this course was also developed from the support from:

- BITSS - Berkeley initiative for transparency in the social sciences.
- PUCP VRI - Annual Research Contest (CAP) 2019 - Project, managed by CISEPA (PI0512).