Modern workflows in data science

Assignment 1

Alexandru Cernat

In the first unit we covered different types of workflows and new practices in reproducible research. For the first assignment we will concentrate on Git and GitHub as these are essential tools for reproducible and collaborative workflows with data.

Tasks

1. Set-up of Git and GitHub **(20%)**

- install Git on your computer
- make GitHub account *(get the free pro version with an educational account)*
- set-up Git in Rstudio *(if you get stuck you can try using GitHub Desktop)*

2. Do online GitHub course **(20%)**

https://github.com/skills/introduction-to-github

3. Do your first Git project **(total of 60%)**

- create a new project and set-up the project structure (folder, README, etc.) **(5%)**
- download two dataset about COVID-19 from this GitHub repository: "UID_ISO_FIPS_LookUp_Table" and "time_series_covid19_confirmed_global". Save both data locally **(5%)**
- merge the two datasets and create a long version of the data. Save both long and wide data locally **(20%)**
- create three graphs using `ggplot` and save them **(10%)**:
  i) overall change in time of log number of cases
  ii) change in time of log number of cases by country
  iii) change in time by country of rate of infection per 100,000 cases
- use the README file to write a mini-report **(10%)**. It should include:
  a. description of the project *(what you aim to do, where you got the data from, etc.)*
  b. explain the organization of the repo *(folder structure, where they can find the data and scripts, what were the steps in creating the report)*
  c. main findings where you include the three graphs and a sentence or two on their interpretation
  d. session info to help with reproducibility
- presentation **(10%)**. Clear structure of the folders, at least three commits with appropriate comments, well formatted README and graphs

Keep you repositories (for tasks 2 and 3) private. For your grading you will invite the teacher (`@alex-cernat`) to be collaborators to your repository. In that way we can mark your work.

Top tips

- for issues with setting up Git and Rstudio maybe check this link or try GitHub Desktop
- for an idea of folder organization you can check the slides
- you can check this link for info on the README file
- to download the data you can use the `read_csv()` from `tidyverse` with the url inside. You can add "?raw=true" at the end of the url to make sure it downloads the data.
- I would recommend not to include the data in the GitHub repo. If you want to not upload some files or folders you can add them to the `.gitignore` file. You can find an explanation here. *I find that the easiest way to do this is to open the `.gitignore` file with a text editor and add the files or folders I want to ignore. The file might be hidden so you need to be able to see hidden files on your computer.*
- for reshaping the data you can check the 'pivot' commands from `tidyverse`
- there are a lot of countries so for graphs 2 and 3 you can be creative. I expect time on the x axis. You could select just some countries to plot or you could plot all of them and then say in words what are the striking patterns *(you can always explore the data using tables to understand it better)*.
- for formatting the README you can look at this webpage. In the next unit we will discuss more about this. For this assignment you need to know two main things:
  - `#` creates a heading
  - `![](path)` inserts image
- learn how to add collaborators to your repo