Assignment 2 – Empirical workflow
Due date: Wednesday, June 10th, 2020 by 5:00pm

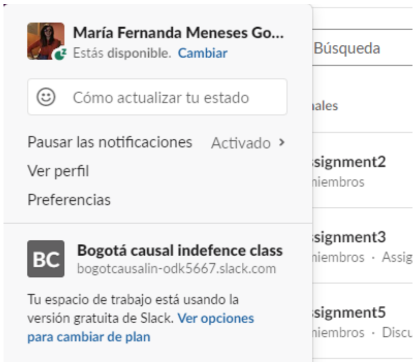
Name: María Fernanda Meneses González

1. Slack

Install slack on your phone and desktop. Create an account and log into our class slack channel. Note, this is for grade. You have two days to get enrolled, otherwise you won't get credit for this part. It is worth **10% of your grade** and is due by end of day **June 10th**. Please do this immediately as it's an easy 10%. Change your profile with a new picture of your choosing, your name and a short description. Say hello in one of the main channels once you're in! Be sure to tag me and Carolina. This is where we will talk as a class regularly. We can also DM this way.

Answer/

I already have enrolled in Slack, I downloaded the app on my phone, I got into the channel and I said hello to Carolina and the teacher:



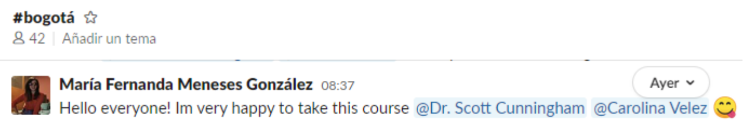
María Fernanda Meneses Go...
Estás disponible. Cambiar

Cómo actualizar tu estado

Pausar las notificaciones Activado ›
Ver perfil
Preferencias

BC Bogotá causal indefence class
bogotcausalin-odk5667.slack.com

Tu espacio de trabajo está usando la versión gratuita de Slack. Ver opciones para cambiar de plan



#bogotá ☆
42 | Añadir un tema

María Fernanda Meneses González 08:37 Ayer
Hello everyone! Im very happy to take this course @Dr. Scott Cunningham @Carolina Velez 😋

2. **Gentzkow and Shapiro**

Read Gentzkow and Shapiro "Code and Data for the Social Sciences" in the "Helpful stuff" Github directory and answer the following

a) Summarize briefly the point of chapters 2- 8 in less than one page.

In the research process it is necessary to determine and think about which steps are capable of being automated. The idea of automating processes has two main reasons: i) replicability and ii) efficiency. Because in real situations, there are countless steps to follow from raw data to final output, it is important to have the ability and speed to replicate an analysis correctly, to the extent that it prevents errors, encourages efficiency, reduces researcher frustration and stress.

For this, having a guide file on the steps to follow (script key steps) is necessary, to be able to leave the process easy to read for someone else if they are going to replicate it and for ourselves when we replicate the analysis. Equally, maintaining an order in the different files used is key, to avoid future confusion. For example, having clear file names and not having too many outputs that might cast doubt on which file was last used. The efficiency issue is key, for example when transferring Excel files to CSV it can be easy if there are only three files, but when the databases get bigger and more complex, the manual task becomes unsustainable and dangerous.

On the other hand, it is important to have tools that allow the control of the different versions of the files to avoid additional confusion, for example, naming the versions with the dates of modification or the initials of the authors' names in case of group work. In addition, it allows you to go back in the process in an orderly manner or it just lets you simply see your previous ideas. This also allows you to fearlessly edit the files, because if there is a mistake, you can roll back the changes.

Another important element to keep in mind is the directories. These should preferably include separation of files according to their functionality and type. For example, the segmentation of the directory in file between the construction and cleaning of data, and the analysis of the same. Additionally, the separation of the inputs and outputs allows a better visualization of the elements and variables.

Considering now the structure of the databases, it is advisable to try to use databases that are logical, understandable, and self-documenting without ambiguities. Additionally, it is recommended to save the original databases (without modifications) in an understandable and normalized way. Later build another file that includes the transformations of the variables, and finally make a merge between the bases.

Writing a good code also requires abstraction and well documented issues. In order to make you code more readable and easier to understand for others, at the time it eliminates redundancy. Besides, sometimes people could think that having a code with full of comments should be useful, but sometimes it becomes a bad idea if it is not put in a proper way. The comments and additional ideas of the codes must me written clearly and updated. Finally, the management of tasks related to this topic, must be believed as something that requires a management system. Wrongly a researcher could use an email, but its advisable to look for one good online tool.

b) Why do Genztkow and Shapiro think these elements of modern empirical work are so important? What problems does each element solve?

The author mentions that these elements are important because they help social scientists find better ways to work. Below each element and its contribution to solve problems.

- The automation process eliminates operational errors from manual work, as well as inefficient processes.
- Version Control allows to have several versions of the files in a clear and orderly way, which allows tracking when there is an error.
- Directories well organized can contribute with order in the handling of the files and clarity of each one of them. It also makes it easy to run the data.
- Store cleaned data solves the problem of complex bases that turn out to be unintelligible to the researcher.
- Having the data in a normalized form until the last step, solves the problem of not having all the elements and files during each step, for example it is useful to trace an error.
- Abstraction eliminates redundancy and confusions
- Well documented notes and comments could reduce the lack of understanding of the processes, but if it is done wisely.
- Management reduces errors of ambiguity and lack of communication between co-authors.

c) Give an example of the sort of problem that could arise in the course of an empirical project if someone were to fail to adopt these principles.

I consider if these principles are not applied in the research work, there is a high probability of errors, inaccuracies, lack of clarity and efficiency, leading to possible erroneous conclusions.
For example, if there is a lack of documentation for the codes and files, it is very likely that when there is a job with a co-author, that person could use an older version of the code, overwrite some result, or simply will not find the latest file. This generates efficiency problems and a high probability of errors.

d) How do you plan to incorporate these solutions into your own work?

I think I could start with the basics of having a well-organized directory. Many times, I used to have many files in a single folder but separating each one into input or output already seems like a great tool. Another element that I can adapt is to reduce unnecessary and out-of-date code comments, because I usually write a lot in them unnecessarily. Also maintaining versions of the files seems key to me, because it allows having backups in case any unexpected error occurs.
Finally, in a longer term, betting on automating most of the process, so it avoids operational errors that can be very costly.

3. Git section

These next questions concern the software “git” and “github”.

As you saw in the previous section, Gentzkow and Shapiro believe version control to be one of the pillars of contemporary empirical research. One of the most popular methods today of version control is Git. But Git is a bit complicated the first time one learns about it. I encourage you to read Gentzkow and Shapiro closely, as well as google and Youtube, to learn enough to answer the following questions. One example is this deck of slides by Grant McDermott at the University of Oregon:

https://raw.githack.com/uo-ec607/lectures/master/02-git/02-Git.html#1

I have also included a deck of slides by Frank Pinter in "Helpful stuff". The completion of this section will satisfy the Github requirement of the course, not counting any additional assignments that use Github. You must have it done by **Wednesday June 10th** to receive the 10% credit.

1. Create a new section in the document you used to answer questions 1-4. Briefly explain what git and github are used for, how they are similar and how they are different.

Git is a high-quality version control system that was first developed in 2005. It is installed and maintained on your local system and gives you a record of your ongoing programming versions. Git is a free system; you only have to download it. In a few words Git lets you manage all the code history of your programs.

Github is an online hosting platform that provides an array of services built on top of the Git system. It is a hosting service that lets you manage Gt repositories. They are not the same, as McDermott said "…we don't need *Rstudio* to run R code, we don't need GitHub to use Git... But it will make our lives so much easier." While Git is installed in your local system, Git works exclusively cloud based. Through GitHub, is possible to share codes with others, giving them the power to make revisions or edits on various Git branches. It allows coordinated teamwork in real time.

2. Name a benefit of using git to organize your empirical research. What types of common problems can occur if you don't use git?

I have read Git is a very useful tool (in fact in some websites it is called best performer) for version control systems, because it is very fast and accurate, and also branching and merging are optimized for a better performance than other systems. You are not going to sleep worried about an elimination of something important because Git does not eliminate data, it offers security and reliability. It also lets you work offline. Using Git makes the workflow flexible and also branching and merging are easy processes.

3. What about using git is challenging for you for right now? What steps can you take to minimize those challenges such that you can adopt git for this class?

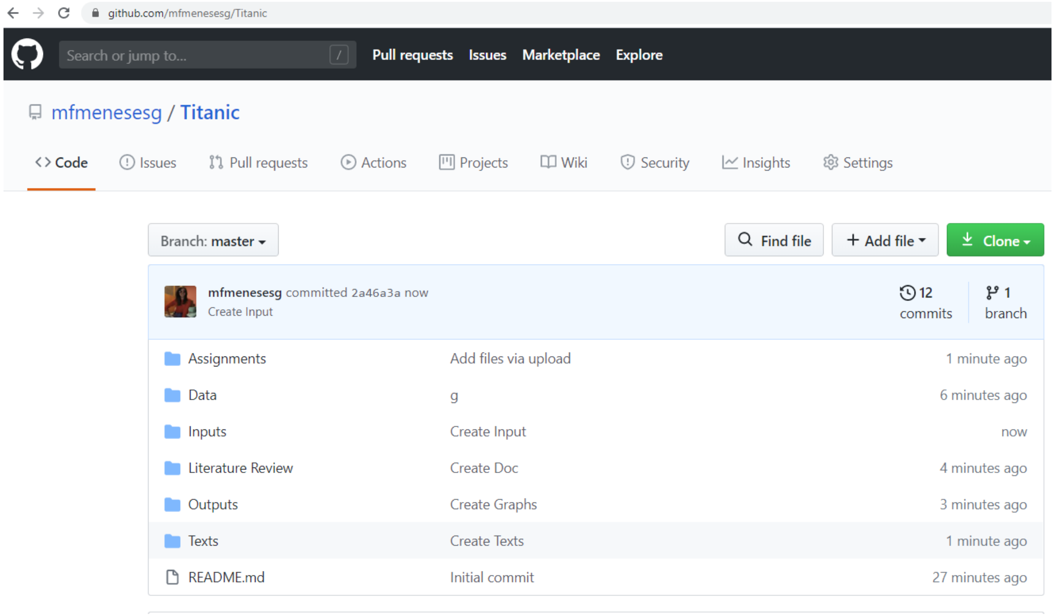
I think given the importance of maintaining version control of all files and processes in an investigation process, using Git is quite necessary. I think the main barrier is learning to use the system, with all its functionalities. I believe it could be solved with practice and research work on the functions and advantages of the system.

4. Name the four main Git operations. What does each operation do and how are is each operation different from one another?

- git config: Configure the author and email address to be used with commits.
- git add: It puts current working files into the stage (aka index or cache).
- git clone: It clones an existing repository into a new directory.
- git push: It sends changes to the master branch of your remote repository.

These operations are very different considering their different functionalities. On the one hand, one configures the author's name and the other clones information from a repository. Git operations differ but complement each other in the process.

5. The first step in your new empirical workflow is the creation of a Github repository ("repo"). You can either do this independently or do this through R functionality. You need to create a github account, then create your first repository called "Titanic". Initialize with a Readme and create the separate folders that we discussed in class on Monday.



mfmenesesg / Titanic

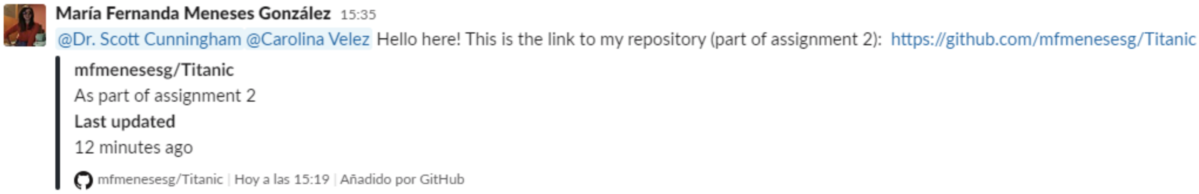
mfmenesesg committed 2a46a3a now
Create Input

12 commits

1 branch

Assignments	Add files via upload	1 minute ago
Data	g	6 minutes ago
Inputs	Create Input	now
Literature Review	Create Doc	4 minutes ago
Outputs	Create Graphs	3 minutes ago
Texts	Create Texts	1 minute ago
README.md	Initial commit	27 minutes ago

6. Post a link to your repository

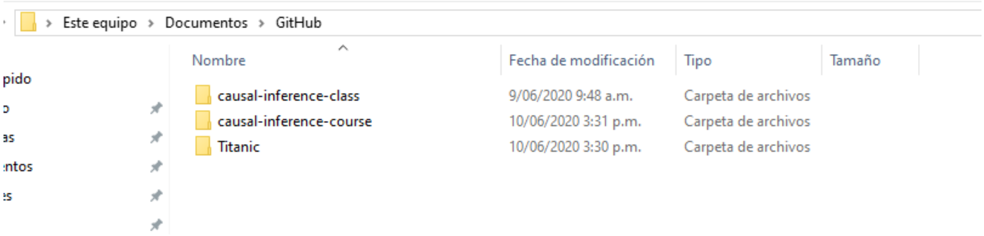


María Fernanda Meneses González 15:35
@Dr. Scott Cunningham @Carolina Velez Hello here! This is the link to my repository (part of assignment 2): https://github.com/mfmenesesg/Titanic

> **mfmenesesg/Titanic**
> As part of assignment 2
> **Last updated**
> 12 minutes ago
> mfmenesesg/Titanic | Hoy a las 15:19 | Añadido por GitHub

7. Please clone our course github repository on your desktop

https://github.com/scunning1975/causal-inference-class



Nombre	Fecha de modificación	Tipo	Tamaño
causal-inference-class	9/06/2020 9:48 a.m.	Carpeta de archivos	
causal-inference-course	10/06/2020 3:31 p.m.	Carpeta de archivos	
Titanic	10/06/2020 3:30 p.m.	Carpeta de archivos	

Bibliography Git section

1. Git vs. GitHub: What's the Difference?
https://blog.devmountain.com/git-vs-github-whats-the-difference/#:~:text=what's%20the%20difference%3F,help%20you%20better%20manage%20them.
2. 8 Reasons for Switching to Git
https://www.git-tower.com/blog/8-reasons-for-switching-to-git/
3. What is GIT?
https://www.toolsqa.com/git/what-is-git/