**Assignment 2 – Daniela Segura Valencia**

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

1. **Summarize briefly the point of chapters 2- 8 in less than one page.**

The main objective the authors try to explain, is how to optimize the time spent during a research project, through 8 advises or steps, they consider key.

In the first place, researchers should automate the processes they run at different programs. This implies the creation of a structure that allows to repeat or correct specific steps along the time, without the need of repeating or recreating the procedure from zero every time a revision is necessary. In the future, third parties or even the authors, will be thankful for the time spent in the automation, because this increases the efficiency, probable big mistakes and reduces great costs of repeat steps that do not give an added value to the project. The second advise, introduces a methodology that is used by different study branches and is the creation of a “version control”, that includes the monitoring of different kind of files, that goes from pdf’s to programming codes. This step can be seen as a more efficient version of a binnacle, because tracks all the editions, changes, corrections, mistakes, and so on, made by any of the authors, without creating concretely a bunch of different versions of the file, that can cause confusions and delays.

Further, it is advised to optimize the research through the use of directories in a way the authors and thirds can separate different kind of information and its respective uses. One useful rule is to separate the original data set, the code related with the project in question and the regression results shown in different kind of visualizations as tables or graphs, in different directories an subdirectories. These actions allow to identify easier and faster which are the sources of each kind of questions that can appear during the research. Furthermore, and in relation with the way the data should be managed, the fifth step, called “keys” implies three steps. The first one, is the importance to save the original data set, with all the different tables that contain each one an identification key, in a specific place; the second one, is the creation of a second set of files that includes the change in the variables and the creation of new variables or rows, that are going to be used in the analysis; by last, if it is required, researchers should do a merge of the data set, using the key of each table, to make the estimations.

The sixth advise, promote the development of an efficient programming process that is not redundant and goes from the most detail fact, to the most general way to represent a specific situation. As an advantage, the authors will save a lot of time trying to understand and unreadable code or unnecessary lines that does not give an added value to the analysis. Finally, they advise to optimize the comments included in the directory through an efficient variable naming process that makes more understandable the code reading without redundant or confusing clarifications; and the use of extra tools that allows to manage all the tasks and responsibilities all the team have been assigned.

1. **Why do Genztkow and Shapiro think these elements of modern empirical work are so important?**

All these tools used together, try to improve the research process, through the principal objective most of economics researchers wants to reach: minimize the errors and maximize the efficiency. In our particular situation, avoiding as most mistakes as possible, can save a lot of time that can be best invested, for example, discussing the principal results and its impact to the hypothesis.

**What problems does each element solve?**

* **Automation:** Increases the efficiency when some steps need to be changed or corrected.
* **Version control:** Eliminates or reduces the high cost of making mistakes or change the approach of a research during a period of time, by maintaining and unique version of the directory at each specific moment of the time, which saves all the actualizations, without creating a bunch of versions with confusing names, that not necessarily are well connected with the rest of the directories and sub directories.
* **Directories:** Makes easier the process of developing a group project, to recognize the origin of the data and each specific goal according to the directory
* **Keys:**  Do not lose the original data sets, so mistakes can be easily tracked.
* **Abstraction:** Saves research time, makes the code reading easier, reduce redundancy, improves clarity and reduces the probability of making mistakes.
* **Documentation:** Increase efficiency, avoid confusion, allow other researchers or interested people to analyze the results by themselves without witnessing errors in the code.
* **Management:** avoid ambiguity, keep in mind the track of all the important objectives, share with all the team the goals and tasks.

1. **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.**

An empirical project must have a routine that includes all of these principles, because they all try to improve the way all the information recollected during the process can be found or identified in any moment of the time, and by any person who needs it. This is why failing in the creation of a version control could end up in mistakes related with the unnecessary replication of an old file in which the researcher had already identified a specific error in the code and hence in the results. For example, if the researcher in the past fall into a procedure mistake with the use of missing values presented in the data set, and find an efficient way to solve it, then, in the present, if this mistake occurs again, he can have two possible ways to react: the first one, implies investing time thinking about how he did solve that problem before, which is not what he should want because of the time constraint; and the second one, that implies to move back some versions and replicate the correct code, without investing valuable time trying to solve some problem that has been already solved before.

1. **How do you plan to incorporate these solutions into your own work?**

I think I should implement all of them in some grade but in a progressively way, because changing a methodology I already have developed during time, that has worked good enough for me and for my work team is not efficient either. But in general terms, the most useful tool is the implementation of a version control, because projects in the financial sector are always having innovations and changes, which does not necessarily imply that the former versions were wrong, just means that decisions are always trying to improve the results. But, to prove this, is important to keep control of all the previous versions, to compare different results and take the best possible decisions.

**Git**

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.
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?
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?
4. Name the four main Git operations. What does each operation do and how are is each operation different from one another?
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.
6. Post a link to your repository
7. Please clone our course github repository on your desktop

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