«R Notebooks» and reproducibility

Assignment 1 i kurset Data Science 2021 - Karoline Midtbø og Morten Knutsen

Introduction

In this paper we will look at reproducibility and how R notebook can be a solution to this problem. First we are going to look at literature review, where we present what other scientific authors writes about reproducing. Then we will have our on discussion on how the necessity of reproducibility in research and whether the use of "R - Notebooks" is a possible solution to the problem on lack of reproducibility. In the end we will represent a conclusion to the chapter discussion.

When we are talking about ***reproducibility*** it is about getting confidence in the conclusion to the scientists (McNutt, 2014). The definition of reproducibility is how other researchers can use the analysis of former researchers to achieve the same result using the same analysis and data (Samota og Davey, 2021)

R Notebook is a document from R Markdown that contains chunks(Grolemund og Wickham, u.å.). R Notebook is a document that has direct interaction with R, but it is also a document that are reproducible (Grolemund og Wickham, u.å.). When you are going to publish the document you can publish Immediately or you can knitted to another document like HTML,PDF or Word.

Short literature review

Roger D. Peng has written a paper that tells us more about reproducibility. He says that "*A critical barrier to reproducibility in many cases is that the computer code is no longer available*" Peng (2011). This is one of the problems to reproducibility. "*Researchers across a range of computational science disciplines have been calling for reproducibility, or reproducible research, as an attainable minimum standard for assessing the value of scientific claims*" Peng (2011). Even if reproducibility becomes a minimum standard, it does not guarantee the quality. The "R" kite-mark is to indicate the idea that a knowledgeable has reviewed the data and code and found it reproducible. To make researches reproducible it is recommended for everyone that use any computing in there research to publish there code. Even though the code isn't clean, they should publish it, it just need to be available Peng (2011).

There is on article about "*Do economics journal archives promote replicable research?*" written by McCullough et al.@mccullough2008. The article show how the data and the code to

an article is important to have to replicate. There is several companies that are publishing articles that want a system that makes the authors include the data and code when they publish, but most of them fail to achieve that. There are a replication policies that they can follow, that includes some requirements, on how to do it, and then have them in archives. The reason to many of them failed is that it is all up to the authors to do them right. The goal of replication is that authors or researchers can use other-minded articles to explore further, so they can avoid wasting time doing the same research. They also talks how the economics don't see the reason to replicate, but what else is the meaning of archives?

According to McCullough et al. there are several authors who do not include data and codes when they are publishing article etc. Possible reasons why they do not publish is that they themselves have to sort to see that all data and codes are in order, or else it is not usable to reproduce. Sometimes they don't want to let other authors do further research to the study they did, or to use it as a base. When it comes to publishing with code and data, the authors have to include it by them self in the publishing McCullough et al. (2008).

There is some solution on how to increase the opportunities to reproduce others work. Some articles show examples that archives can be mandatory, that will mean the authors have to include data and code when they publishing so the article etc. will be in a system in the archive and can be used again later on. With help of R Markdown/R Notebook it will be easier for the authors to have control on the code and data, and it will always include in the program. There is to different chunks that helps with data and coding.

Code chunks are a series of commands in different programming language, in example R. Code chunks preform calculations needed to produce the appropriate output. Also to create intermediate results used across different code chunks.

A text chunk, on the other hand, describes the results, codes, problems and the interpretation. Text chunks is formatted for the user to read it, not the computer.

Discussion

R notebook can be a solution to fix the problem of reproducibility, but only partly. R notebook has the opportunity to be a good tool for any researcher, but it requires that the researchers knows how to use the program and they need to have the exact same packages that was used under the study, or else they will not manage to reproduce the study. When you are using R studio you can use Github to store your repository, then you can store it as private or public. After you store it in Github you can pull it down to R studio whenever you need it. Other Github users can also use your repository if you make it public, and use your data and code, if they have the same packages you used. There is also a lot researchers that want to protect their work and will not make their codes available, and then it will be difficult to reproduce for others. For beginners in R there will be a lot to get into, there is many different things you need to know before you can use it, there is a lot of programs you need to install for making it optimal.

1. R Notebook will solve the problem with reproducibility

- With R Notebook the document already contain codes.
- It is a free program to use, everybody can install it.
- You can use Github to store it longer.

2. R notebook will not solve the problem reproducibility

- If you want the data and code, you have to have the right packages that was used.

- To use the program you have to know how to use it.

```
{r-første chunk} sessionInfo()
```

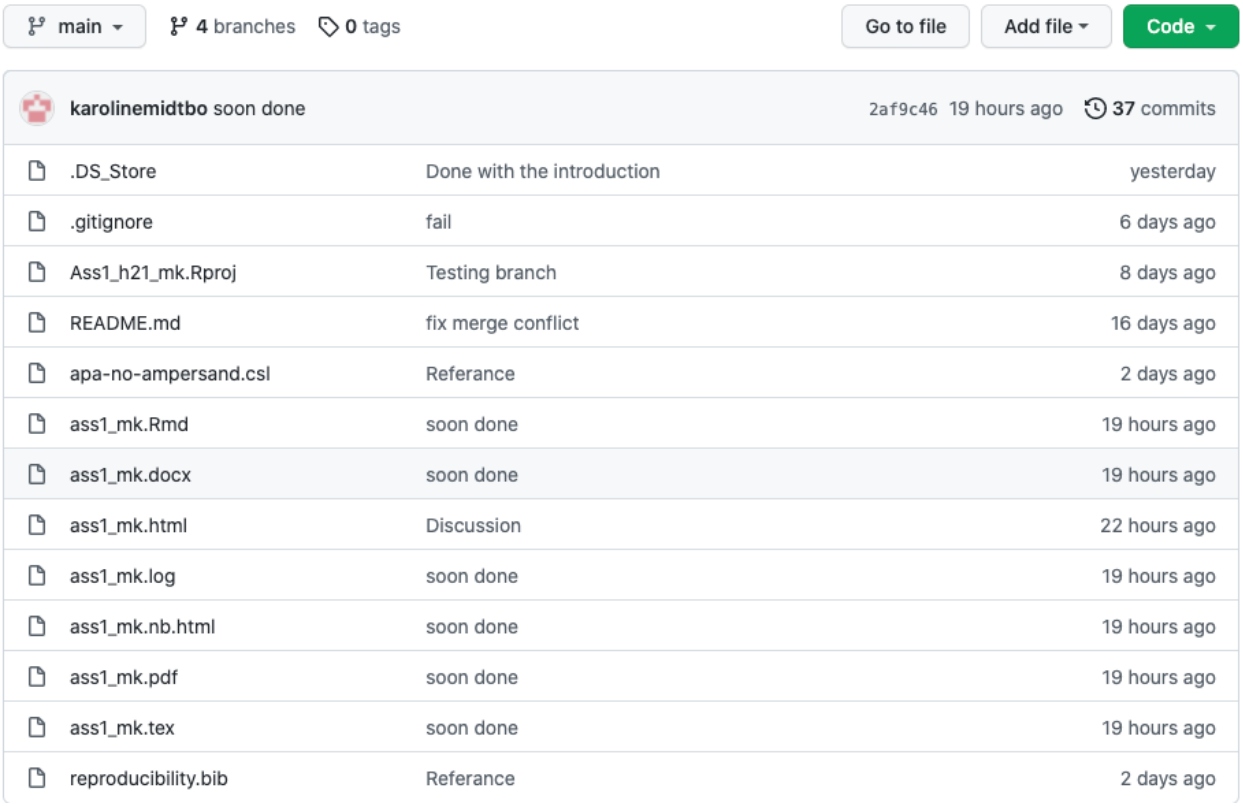
This function can help us to reproduce the research because it gives us information about which R version and packages we used. It can take time to install all the packages if you don't have them, but with chunks it will be easier to find which packages we used in our work.

After the literature work we did, we find a lot of information about how researchers feel about reproducibility. They want to protect the work they did, and often they don't want to publish their code and data they used to find their answers. Without the code and data other researchers can't get the exact same answer, and they will have to take the same exact test and use a lot of time to get the information. If the author publish the code and data it will be easier for the next researcher to just use the test that is already done and they will have the exact same answer and can use it to develop it. When authors or researchers publish with data and code, there will be a lot more of opportunities of reproducibility. But it is all on the authors, when they publish they have to check if the code and data are included.

Conclusion

After we have read different literature and had an discussion, we can see it is a split between the authors that want to publish their data and code and they who do not want to publish them. We conclude that the R Notebook will help the authors or researchers to have more control on the data and the code, but it's a program that need knowledge and require different packages to get the exact same answer. Predictability will help a lot when it comes to further resource, to don't waste time on what others have answer on, it also makes it easier to back up what they got.

References



main 4 branches 0 tags Go to file Add file Code

karolinemidtbo soon done 2af9c46 19 hours ago 37 commits

.DS_Store	Done with the introduction	yesterday
.gitignore	fail	6 days ago
Ass1_h21_mk.Rproj	Testing branch	8 days ago
README.md	fix merge conflict	16 days ago
apa-no-ampersand.csl	Referance	2 days ago
ass1_mk.Rmd	soon done	19 hours ago
ass1_mk.docx	soon done	19 hours ago
ass1_mk.html	Discussion	22 hours ago
ass1_mk.log	soon done	19 hours ago
ass1_mk.nb.html	soon done	19 hours ago
ass1_mk.pdf	soon done	19 hours ago
ass1_mk.tex	soon done	19 hours ago
reproducibility.bib	Referance	2 days ago

Grolemund, G., og Wickham, H. (u.å.). *R for Data Science*.

McCullough, B. D., McGeary, K. A., og Harrison, T. D. (2008). Do Economics Journal Archives Promote Replicable Research? *Canadian Journal of Economics/Revue canadienne d'économique*, *41*(4), 1406–1420. https://doi.org/10.1111/j.1540-5982.2008.00509.x

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229–229. https://doi.org/10.1126/science.1250475

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Samota, E. K., og Davey, R. P. (2021). Knowledge and Attitudes Among Life Scientists Toward Reproducibility Within Journal Articles: A Research Survey. *Frontiers in Research Metrics and Analytics*, *6*, 678554. https://doi.org/10.3389/frma.2021.678554