«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 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 present our conclusion.

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 and Davey, 2021)

An *R Notebook* is a document in R Markdown format that contains chunks(Grolemund and Wickham, n.d.). Such an R Notebook is a document that has direct interaction with R, but it is also a document that are reproducible (Grolemund and Wickham, n.d.). When you are going to publish the document you can publish immediately or you can *knit* it into another format like HTML, PDF or Word.

Short literature review

Peng (2011) 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*" . 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).

@mccullough2008 argues that the data and the code in an article have to replicate. Several publishers og scientific journals now want a system that makes the authors include the data and code with their published articles, but most of them fail to achieve this. There are replication policies that they can follow, that includes some requirements, on how to do it, and then have them in archives. The reason so many of them fail is that it is all up to the authors to do this 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 mention that many economists don't see the reason why to replicate, but what else is the meaning of archives?

According to @mccullough2008 there are several authors who do not include data and codes when they publish article e.t.c.. Possible reasons why they do not publish is that they themselves have to sort out all the data and check that the codes are correct, and it will take a lot of time to do it each time. So in many cases the authors do not take the time to include data and code when publishing. Sometimes they will not let other authors do more research on the study they did, or use it as a base, which then means that they do not publish with code and data for the content (McCullough et al., 2008).

It is a solution on how to increase the possibilities of reproducing other works. Some articles show examples that archives can be mandatory, this means that the authors must include data and code when they publish so that the article e.t.c. will be in a system in the archive and can be used again later. Using the R Markdown / R Notebook will make it easier for the authors to have control over the code and data, because it will always be included in the program. There are different chunks that allow the data and coding to be included in the article without the author himself having to set them up and sort them afterwards.

Code chunks are a series of commands in different programming language, for 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 partially. The R notebook has the potential to be a great tool for any researcher, but it requires researchers to know how to use the program, and they must have exactly the same packages that were used during the study, otherwise they might not necessarily be able to reproduce the study. When you use R studio you can store your project in a repository on Github, where you can save it as a public or private project. After saving it in Github, you can drag it down to R studio whenever you need it. Other Github users can also use your repository/project if you make it public, and will then be able to use your data and code if they have the same packages that you used. There are also many researchers who want to protect their work and do not want to make the codes available, and then it becomes difficult for others to reproduce. For beginners in R it will be a lot to get acquainted with, there are many different things you need to know before you can use it. There are also many programs you need to install to do it optimally. We think there might be some problems because when we use the R studio/R Notebook in this way we always have to save it, commit it and push and pull when we use github as our backup.

Under is some cons and pro on how R Notebook can solve the problem with reproducibility.

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.
- You need to download a lot of programs to use the notebook properly.

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.1.1    magrittr_2.0.1    tools_4.1.1       htmltools_0.5.1.1
##  [5] yaml_2.2.1        stringi_1.7.3     rmarkdown_2.10    knitr_1.33
##  [9] stringr_1.4.0     xfun_0.25         digest_0.6.27     rlang_0.4.11
## [13] evaluate_0.14
```

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 found a lot of information about how researchers feel about reproducibility. They want to protect the work they do, and often they don't want to publish the 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 would 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 those who do not want to publish. 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 already have done, it also makes it easier to back up what they got.

In our conclusion we will say that R Notebook can be a solution to fix the problem of reproducibility, but only partially. We can use different tools to make a good article e.t.c., and it also contains the code and the data. But it is a little bit complicated to understand without knowledge.

References

Grolemund, G., and Wickham, H. (n.d.). *R for Data Science*.

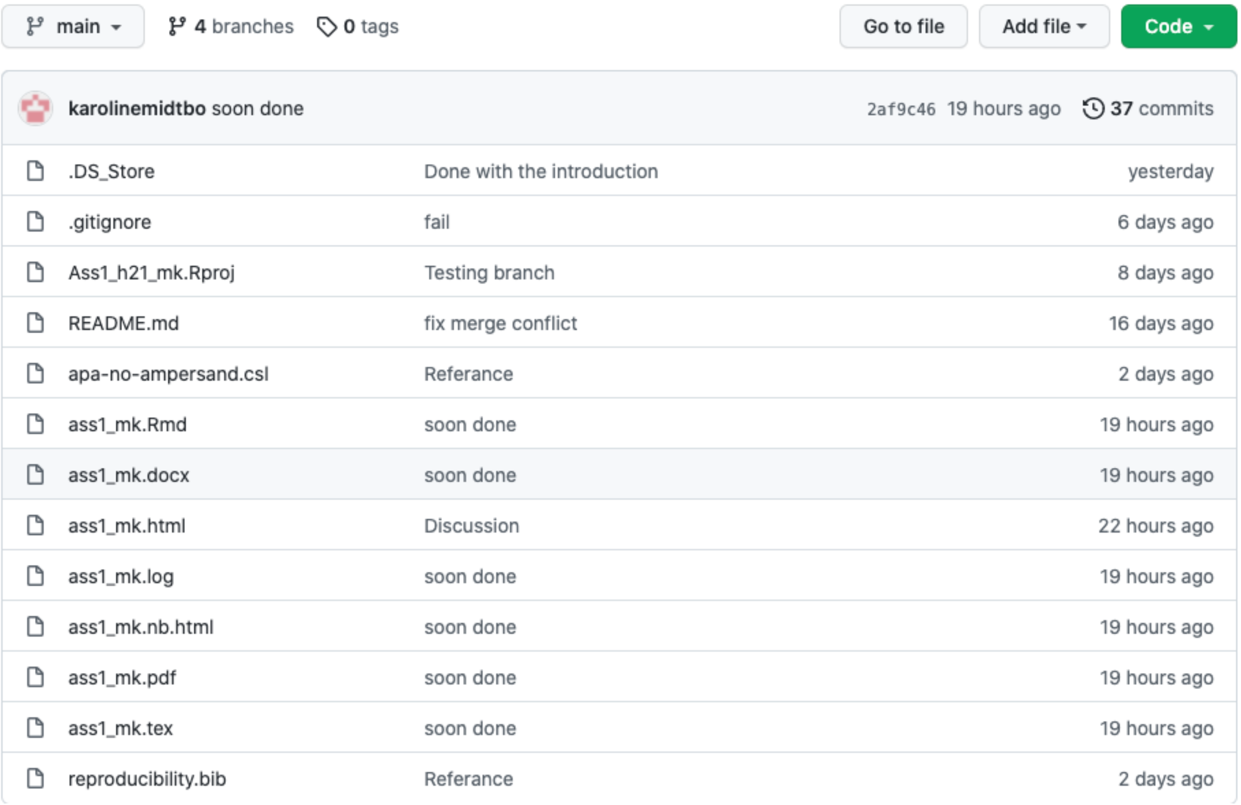
McCullough, B. D., McGeary, K. A., and Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'économique*, *41*(4), 1406–1420. https://doi.org/10.1111/j.1540-5982.2008.00509.x

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

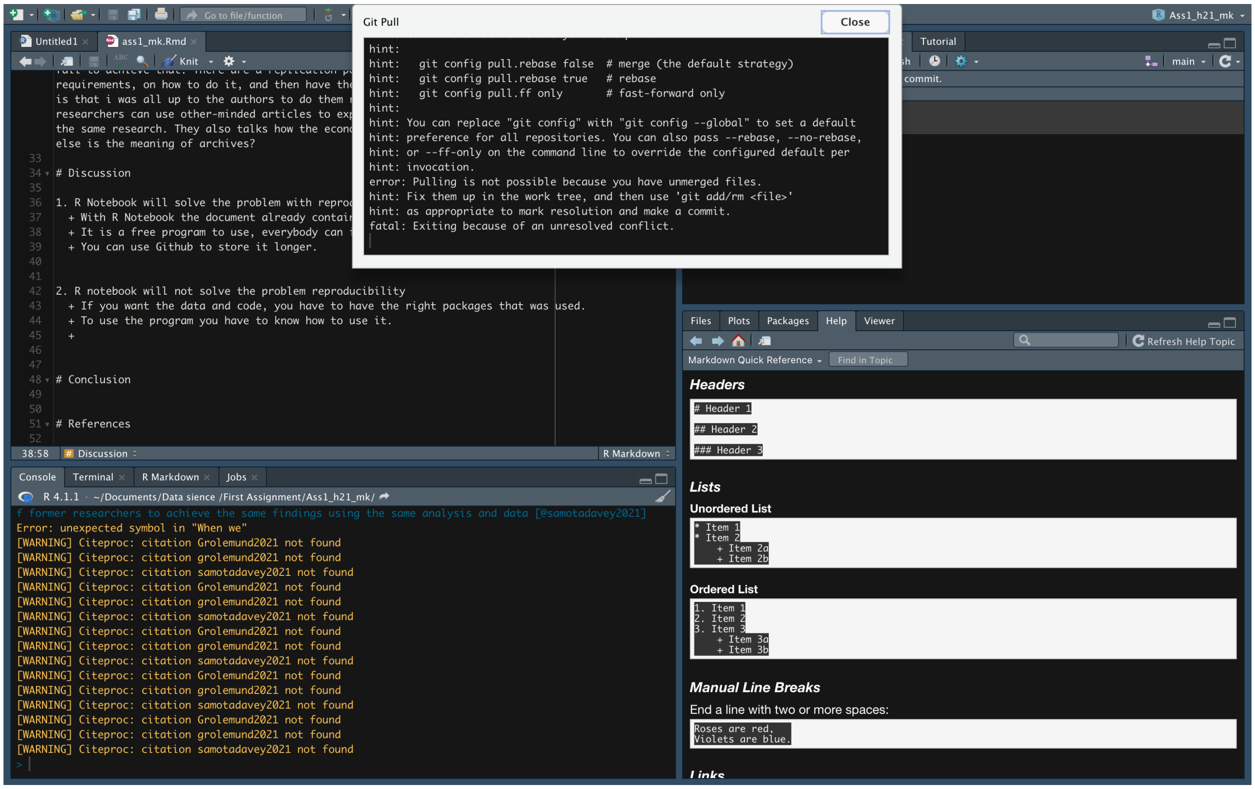
Samota, E. K., and Davey, R. P. (2021). Knowledge and attitudes among life scientists toward reproducibility within journal articles: A research survey. *Frontiers in Research Metrics and Analytics*, *6*, 678554. https://doi.org/10.3389/frma.2021.678554

Appendix



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



```
requirements, on how to do it, and then have the
is that i was all up to the authors to do them r
researchers can use other-minded articles to exp
the same research. They also talks how the econo
else is the meaning of archives?

# Discussion

1. R Notebook will solve the problem with reproc
  + With R Notebook the document already contain
  + It is a free program to use, everybody can t
  + 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.
  +


# Conclusion


# References

```

Git Pull

Close

```
hint:
hint:   git config pull.rebase false  # merge (the default strategy)
hint:   git config pull.rebase true   # rebase
hint:   git config pull.ff only       # fast-forward only
hint:
hint: You can replace "git config" with "git config --global" to set a default
hint: preference for all repositories. You can also pass --rebase, --no-rebase,
hint: or --ff-only on the command line to override the configured default per
hint: invocation.
error: Pulling is not possible because you have unmerged files.
hint: Fix them up in the work tree, and then use 'git add/rm <file>'
hint: as appropriate to mark resolution and make a commit.
fatal: Exiting because of an unresolved conflict.
```

```
R 4.1.1 · ~/Documents/Data sience /First Assignment/Ass1_h21_mk/
f former researchers to achieve the same findings using the same analysis and data [@samotadavey2021]
Error: unexpected symbol in "When we"
[WARNING] Citeproc: citation Grolemund2021 not found
[WARNING] Citeproc: citation grolemund2021 not found
[WARNING] Citeproc: citation samotadavey2021 not found
[WARNING] Citeproc: citation Grolemund2021 not found
[WARNING] Citeproc: citation grolemund2021 not found
[WARNING] Citeproc: citation samotadavey2021 not found
[WARNING] Citeproc: citation Grolemund2021 not found
[WARNING] Citeproc: citation grolemund2021 not found
[WARNING] Citeproc: citation samotadavey2021 not found
[WARNING] Citeproc: citation Grolemund2021 not found
[WARNING] Citeproc: citation grolemund2021 not found
[WARNING] Citeproc: citation samotadavey2021 not found
[WARNING] Citeproc: citation Grolemund2021 not found
[WARNING] Citeproc: citation grolemund2021 not found
[WARNING] Citeproc: citation samotadavey2021 not found
>
```

Headers

```
# Header 1

## Header 2

### Header 3
```

Lists

Unordered List

```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

Ordered List

```
1. Item 1
2. Item 2
3. Item 3
    + Item 3a
    + Item 3b
```

Manual Line Breaks

End a line with two or more spaces:

```
Roses are red,
Violets are blue.
```

Links