Can R Notebook help with the reproducibility?

By Thibiga and Ingrid

Introduction

Reproducibility and replicability has been an ongoing topic among scientists and academics. The importance of the matters has been widely discussed considering the ongoing "reproducibility crisis" (Serra-Garcia and Gneezy, 2021) The crisis refers to the failure to reproduce and replicate research and experiments that has been published (Serra-Garcia and Gneezy, 2021) Additionally, it has also been found in Serra-Garcia and Gneezy (2021) research that "nonreplicable publications are cited more."

Defining the terms

In order to further look at the findings, in a variation of studies, to discuss the title question, we need to define the relevant terms that will be used in this paper.

Reproducibility is defined by NSF (U.S national Science Foundation) as "(. . .) The ability of a researcher to duplicate the result of a prior study using the same materials as were used by the original investigator" (Goodman et al., 2016) . While reproducibility includes secondary researchers using the same material, *replicability* refers to collecting new data that leads to the same findings as the original study (Sciences et al., 2019). A source to perform and test both reproducibility and replicability, is R Notebooks - a markdown document which allows the user both "independent and interactive execution of the code chunks" (Boehmke, 2016). The user is also able to view the developed output immediately underneath the implied input (Grolemund and Wickham, n.d.) The notebook further allows users to interact with the software R while also create output which is both reproducible and of great publication quality (Grolemund and Wickham, n.d.).

The aim of this paper is to define reproducibility and replicability and further discuss if R notebook can help with reproducibility. We will look at the problem and further discuss the scope of the problem. Lastly, we will also look at potential solutions with R notebooks as a relevant tool.

Short literature review

This literature review will undertake a research of reproducibility and replicability in relevance to *R and R-studio*. It will further present relevant and credible sources. We will be using sources deducted from the shared Zotero library, which have been collected through detailed work by our professor.

As defined in the introduction are both reproducibility and replicability an important topic among scientists and in the academic world. Multiple articles have been written about the topic and the importance of it. McNutt (2014) states in her article from 2014 that reproducibility is important for researchers in order to support their conclusions and results. The article further states that there are various reasons of irreproducible work. Firstly, it may be hard for the researcher to control all the independent variables, some of the authors may also want to keep their methods private and not reveal the methods behind the concluded result (McNutt, 2014). Another factor is the production of false positives studies, where some of the results

indicates a "wrong" outcome. In an attempt to create reproducible studies must the above mentioned factors be avoided. McCullough et al. (2008) is also looking at the lack of reproducible published data. They found in their research of economics journal archives that authors tend to not contribute enough data in the published articles for them to be reproduced (McCullough et al., 2008). Through McNutt (2014) and McCullough et al. (2008) is it clear that there are multiple reasons of why articles and publications are irreproducible. However, given the reasons there are also a number of reasons that supports the creation of reproducible publications. Markowetz (2015) lists

1. Avoid disaster
   If the author is transparent with the data and the outcome, it will be easier to spot issues and mistakes, and it will sooner be edited and updated (Markowetz, 2015)
2. Easier
   It is not only easier for the reviewer, but it is also easier for the author when using transparent data. It is easier to track both your work and potential mistakes (Markowetz, 2015).
3. Reviewers gets a better understanding
   The transparent data makes the reader able to test their own ideas through the use of the authors data. This may help the reader to be on board with the authors thoughts, while also get a better understanding of the data (Markowetz, 2015).
4. Continuity of the work
   Producing reproducible and transparent work will further gain the author as readers can continue the research, which will save the original author both time and stress (Markowetz, 2015).
5. Building reputation
   Lastly, being transparent and open with the research will create a positive reputation for the author. The outcome will be a reputation of "an honest and careful researcher" (Markowetz, 2015).

After looking at the the sources above both the pros and the cons of reproducibility, it is clear that more transparency may lead to a more positive outcome. However, we are yet to look at a possible solutions of the reproducibility problem. Here, R Notebooks is a relevant factor.

The introduction stated that R notebooks helps the user to view output as the input is created, while the user can create transparent and reproducible papers (Grolemund and Wickham, n.d.) (Boehmke, 2016). In R notebooks, the code chunks helps both the author and the reviewer to view both the paper and the data sets in a more transparent way. As Markowetz (2015) states, it is easier to get an overview of the work, and it is also helpful for the reader to get a better understanding of the research. Additionally, using R Notebooks in combination with Git gives the researcher a great opportunity to "track and compare versions, retrace errors, explore new approaches in a structures manner, while maintaining a full audit trail" (Boehmke, 2016). Lastly, the creation of dynamic documents may also be a suggested solution of the problem. The fact that R notebooks has a transparent outcome and underlines the steps the author performs, makes the notebook a great possible solution for reproducibility. However, if it solves the problem is not yet found. This will further be discussed in the next part of the paper.

Discussion

The *R notebook* used in the software called *R (R Core Team, 2021)* , is similar to a *computable compendium* (Gentleman, 2005), which means you can write texts and codes. This notebook is supported by tools like *knitr* (Xie, 2015) (Xie, 2014), an integrated package in R, and *rmarkdown* (Allaire et al., 2021) (Xie et al., 2018) (Xie et al., 2020), which helps to integrate the codes, figures, and similar functions. By using R Markdown you can easily convert the notebook to formats like Microsoft World, PDF and others formats. Sharing the notebook on-line and giving access to other authors is considered as simple. Therefore R Notebook may be a solution for reproducibility.

In addition, there is a data code called *session info* in *R studio*. This function will tell about the author's working environment. That means it would give information about the author's version of *R*. By using this

data code, it could help other authors to reproduce the project or find out codes that deviates from what is normal (*R Session Info Info - Rostrum.blog*, n.d.).

```
sessionInfo(package = NULL)
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## 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
```

In our society the focus on the computational science has increased and therefore there have been some fast-moving developments in many scientific areas (Peng, 2011). One of these areas is the reproducibility in data science. In other words, other authors can reproduce the document by using the same data which have been collected and written by someone else (Gentleman, 2005). This kind of reproducibility is based on that every computational experiment has a detailed log of every action made in a program, like R Studio (Gentleman, 2005).

The documents and its computations are referred to as a compendium, a distributable object and which is a mix between text and data (Gentleman, 2005). These compendiums are "easily comprehend, modified and extended." Which means an external author/reader could be able to reproduce without any complications. The reproducibility is important and necessary, first of all it creates more opportunity for new insights. This is done by using the same compendium by adjusting the data with the aim of achieving the same results. By executing this you would see other possibilities that one has not had in their mind. The risk of doing error will reduce by reproducing. This will ensure that texts and codes are correct, which again would increase the reliability. The result of a research or a studies may be incorrect and by letting your document/compendium be reproducible, will allow readers to understand your work better ("The Significance of Reproducible Data," 2017).

The choice of reproducing the data or not will be based on your work. As written over, the reproducibility is based on that all computations done in research are recorded and is available to reuse. This could also be seen as a disadvantageous side of reproducibility if the data codes are no longer available to readers.

As Peng (2011) stated; **"*Interactive software systems often used for exploratory data analysis typically do not keep track of users' actions in any concrete form.*"**. Changing the behaviour of the software systems or use other software systems that support reproducibility may resolve the problem, but because of the human nature, it will not happen quickly. One of the main reasons is the hours someone have spent to learn a specific program.

Even though if author changes their software systems, they have to make the data codes available. If they choose not to share their data codes, the point of reproducibility would not be relevant. Another barrier to

reproducibility is the understanding of its culture. An ingrained culture that requires reproducibility for all scientific claims is necessary, and not everyone is familiar with this culture (Peng, 2011).

Conclusion

Reproducibility is about using one's author's file to *replicate* by others authors and this topic about reproducibility is heated among scientist in today's society. We can conclude that reproducibility in *data science* have more benefits than disadvantages, and therefore it would be seen as necessary tool for the current and future generations. The *R notebook* in *R studio* would be seen as a great platform to create a file, which can be replicated by others.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2021). *Rmarkdown: Dynamic documents for r* [Manual]. https://github.com/rstudio/rmarkdown

Boehmke, B. (2016). *Data wrangling with r*. https://doi.org/10.1007/978-3-319-45599-0

Gentleman, R. (2005). Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1034

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Grolemund, G., and Wickham, H. (n.d.). *R for Data Science*. Retrieved August 19, 2020, from https://r4ds.had.co.nz/

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology*, *16*(1), 274. https://doi.org/10.1186/s13059-015-0850-7

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

R Core Team. (2021). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing. https://www.R-project.org/

R session info info - rostrum.blog. (n.d.). Retrieved September 23, 2021, from https://www.rostrum.blog/2018/10/13/sessioninfo/

Sciences, N. A. of, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. https://doi.org/10.17226/25303

Serra-Garcia, M., and Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

The significance of reproducible data. (2017). In *Labfolder*. https://www.labfolder.com/the-significance-of-reproducible-data/

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman and Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595

Xie, Y. (2015). *Dynamic documents with r and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.org/knitr/

Xie, Y., Allaire, J. J., and Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Dervieux, C., and Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook

Appendix

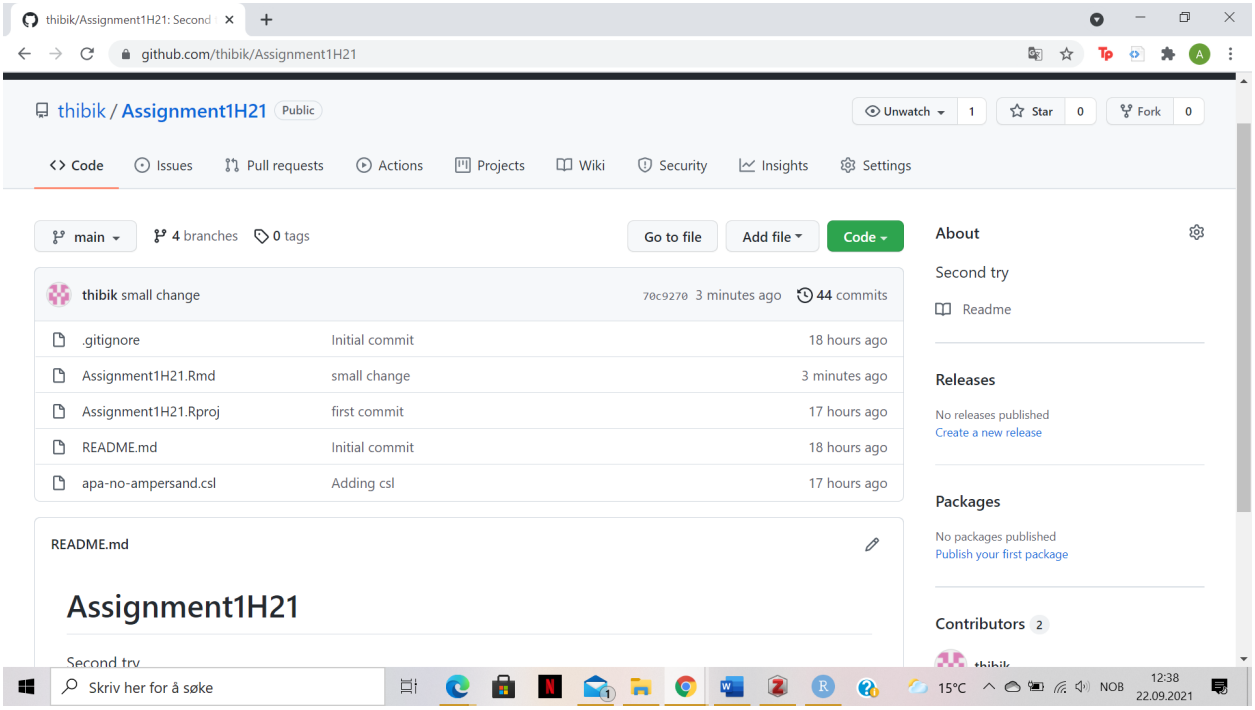

Figure 1: Overview

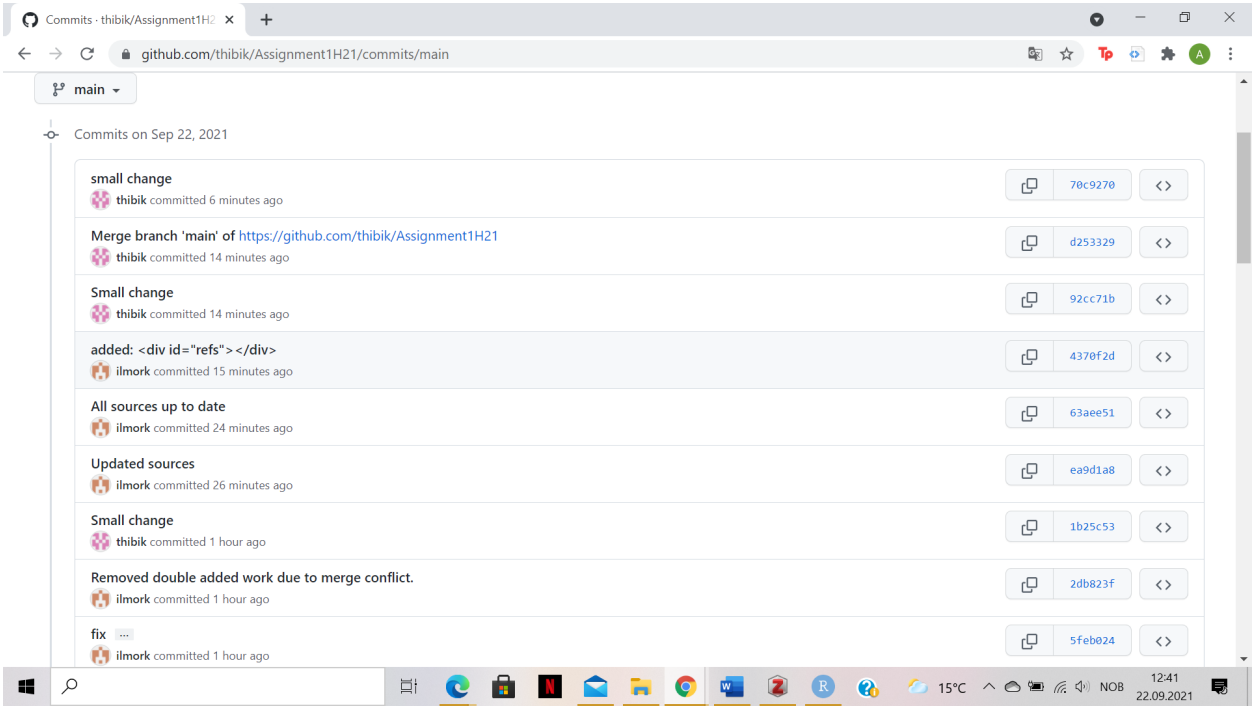

Figure 2: Commit history

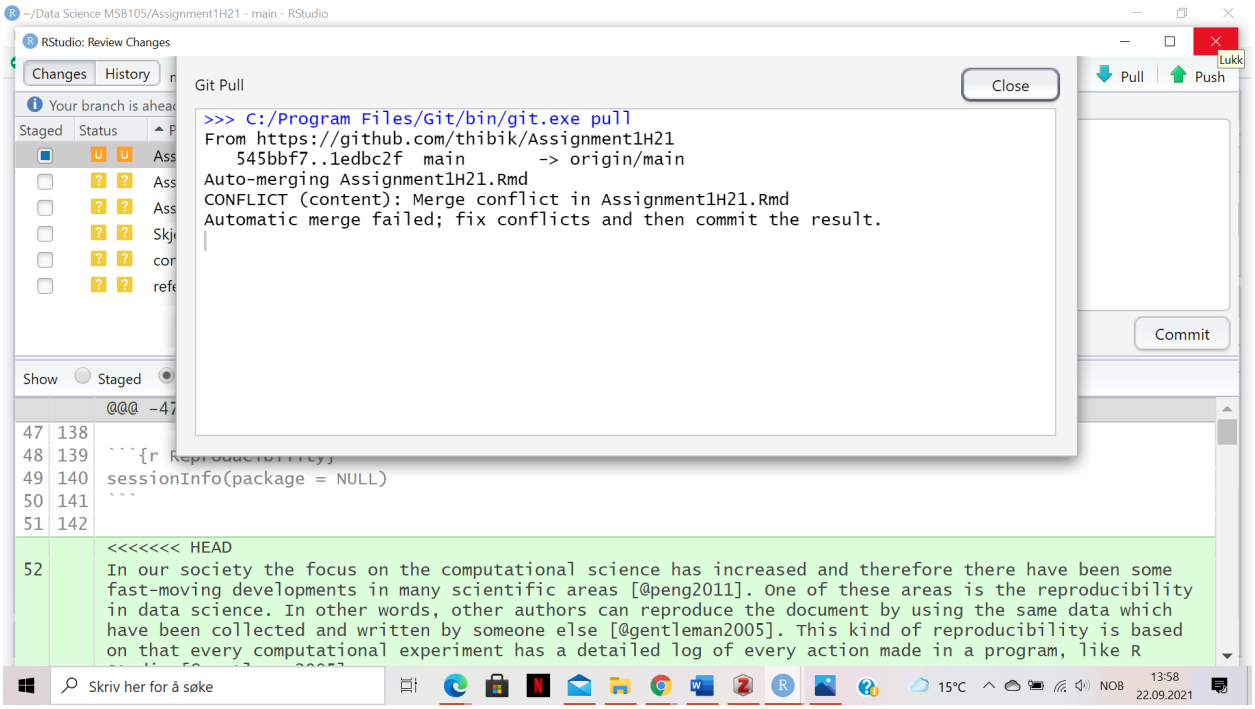

Figure 3: Merge conflict