Can R Notebook help with reproducibility?

Introduction

This text will focus on how R Notebook can help with reproducibility. The text is based on reproducibility as a standard for general research being published, as well as reproducibility in the field of economical studies. R Notebook will be discussed as a solution to meet the criteria of reproducibility.

Literature review

"Reproducible Research in Computational Science," by Peng 2011.

Roger D. Peng presents reproducibility as the bare minimum standard when evaluating all published scientific findings. The reasoning being the change in technology leading to large public data sets/databases used by multiple scientist to produce independent research. This change making it harder to investigate the credibility of different studies/papers.

According to Peng, the studies/papers needs to have data and code public to make it possible for it to be reproducible. An issue following this requirement is that the computer code often is not available due to software systems and packages being private.

As the main goal presented in the article is to develop a culture that require reproducibility for computational science to be publish, two solutions are mentioned:

1. Authors make their data and code available for everyone by using the free code repositories. Such as Github and SourceForge.
    a. This would make it possible to provide a minimum code as online material which would be informative, and could be used to reveal any potential problems.
    b. The next step would be to be able to provide datasets in durable non-proprietary format, but this would require additional costs.

2. The scientific community creates a single place for authors in all fields to make their studies reproducible.
    a. In order to accomplish this the community would need to create a common DataMed central, CodeMed central and PubMed central where publishers/authors data, metadata, and code can be stored with their studies. This would require additional costs and would need to be coordinated and supported by the government.

The article makes it clear that even if a analysis is reproducible it does not guarantee the validity, quality and correctness even when it is published to others.

"Do economics journal archives promote replicable research?" by McCullough.

McCullough's article puts focus on reproducibility and replication in the field of economical studies. In this article the ability to replicate a study is put as the gold standard to evaluate the reliability of a scientific

claim, and should be aimed for in all archives. Replicability meaning being able to collect new data and get the same results/conclusions.

The article shines light on the fact that long-standing archives of economics do not accustom to reproducibility. And some of those who claim to be data-only archives fail in reality due to authors being in charge of submitting their code and data, but neglecting to do so. In these cases the papers get published regardless as a result of lack of controlling by the journal archives. Meaning there is improvement to be done to the promotion of replicable research.

Additionally, the article shows that editors will defend published work even if it is not replicable with the argument of the papers being published before the requirements for data availability. This only showcases the importance of stricter data+code requirements.

"Statistical Analyses and Reproducible Research," by Gentleman and Lang 2004

In the discussion on how R notebook can help meet the criteria of reproducibility Gentleman and Lang's definitions on compendium, code chunks and text chunks are used.

> *Compendium is both a container for the different elements that make up the document and its computations (i.e. text, code, data, . . .), and as a means for distributing, managing and updating the collection.*

> ***code chunks are sequences of commands in some programming language such as R or Perl. Code chunks are intended to be evaluated according to the language in which they are written. These perform the computations needed to produce the appropriate output within the paper, and also to produce intermediate results used across different code chunks.***

> *Text chunks describe the problem, the code, the results and often their interpretation. Text chunks are intended to be formatted for reading.*

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

```
## 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
```

Discussion

In the discussion we will focus on three main areas:

1. Can R Notebook solve the problems surrounding reproducibility?
2.
3.
4.

Can R Notebook solve the problems surrounding reproducibility?

To meet the criteria of reproducibility computable documents is a solution. R Notebook is one of the available

R Notebooks features:

1. Uses R markdown, collects all parts into one document
2. Codes goes through sequentially, but can also go through individually
3. Uses Latex - making it possible to write mathematics
4. Uses pandoc as a transformer - html. pdf. and word.

R Notebook is a tool to have the best foundation to be able to produce reproducible papers/studies. On the other hand, one can not guarantee that each user of R Notebook creates reproducible research, as the program does not guarantee for human mistakes to be made.

As

Conclusion

References