Using R for Data Processing

The What and Why of R

Glenn Williams

University of Sunderland

2020-09-14 (updated: 2020-09-22)

What is R?

- **Statistical programming language** used for processing, summarising, analysing, and graphing data.
- **Free and open source**, so anyone can see how it works and add to the development of the program.
- Packages and changes are **vetted by reviewers**, so we can be sure most things we do in R are appropriate.
- Often used with **RStudio** which makes working with **notebooks**, **projects**, and **version control** easier. More on these features later.

Why Should I Care?

- Free software means **anyone can use it at no cost**. This makes science more inclusive, and allows for easier confirmation of analyses.
- Programming languages force you to **document all decisions** made with the data. This makes your research more **transparent**.

You supporting Open Science

- With **Open Source** Software we can see how the program works and improve/extend it where needed. **Finding and fixing problems is easier and quicker**.
- Free, Open Source software is crucial for **Open Science**.

Sharing is Caring

- Is your reseach **replicable**? If I follow your methods with **new participants**, will I get **similar results**?
- Is your research **reproducible**? If I follow your methods with **your data**, will I get **the same results**?
- Checking these things is made possible if you share your materials, data, and steps for analysis.
- Sites like the Open Science Foundation (OSF) and GitHub allow us to host our research products online. They also track changes to your work.
- Opening up your research can be scary, but **it'll probably make you more careful**, allow people to build on your work, and allow science to be more reliable and able to self-correct.

Why Should I Care?

- Veldkamp et al. (2014): 63% of articles contained **at least one *p*-value that was incorrect**. In 20.5% of cases, these errors led to **erroneous decisions** about the statistical significance of an effect.

- How does this happen?

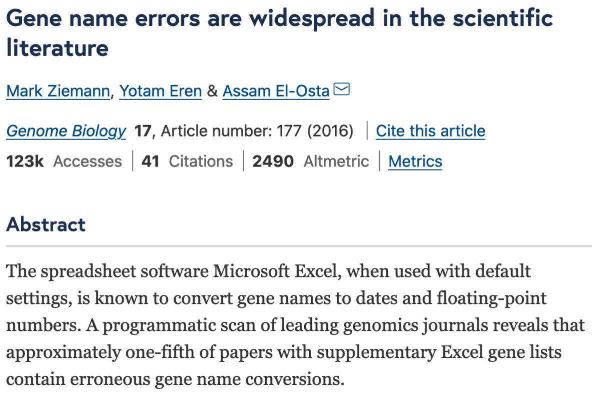
    - Not documenting your methods fully makes **detecting mistakes much more difficult**.
    - Transcribing results from visual interfaces like SPSS means you introduce **human error**.

- How can we fix this problem? **Notebooks**!

    - You never write the results, just the methods for making them.
    - Every step of your analysis is documented.
    - **If your data changes, your results are instantly updated**.

Still Not Convinced?

- Easy to learn is not Easy to use.
    - How do you **filter observations** in Excel? In R, use `filter()`
    - How do you do a **t-test** in SPSS? In R, use `t.test()`
    - Learning to code might take longer, but implementing it is often easier down the line.
- **You won't document everything you do in Excel**. That's a problem if you notice a mistake.
- **R scales up**. Imagine processing 1,000,000 rows of data in Excel. R can do that as easily as 10 rows.



Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

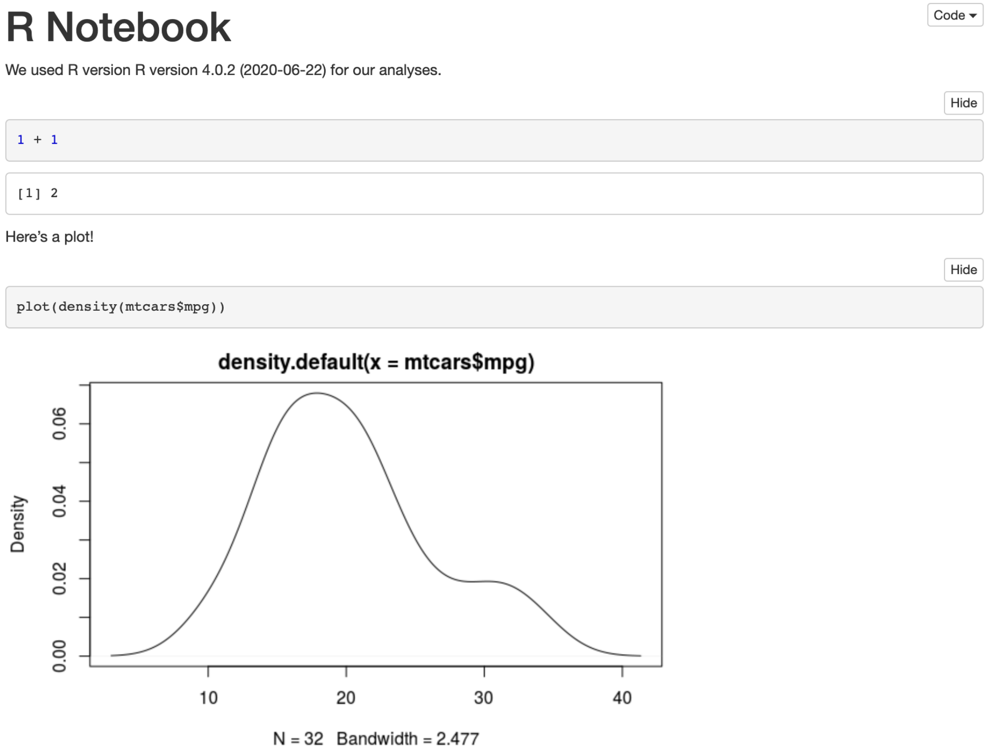
Genome Biology **17**, Article number: 177 (2016) | Cite this article

123k Accesses | **41** Citations | **2490** Altmetric | Metrics

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Example Notebook



R Notebook

We used R version R version 4.0.2 (2020-06-22) for our analyses.

```
1 + 1
```

```
[1] 2
```

Here's a plot!

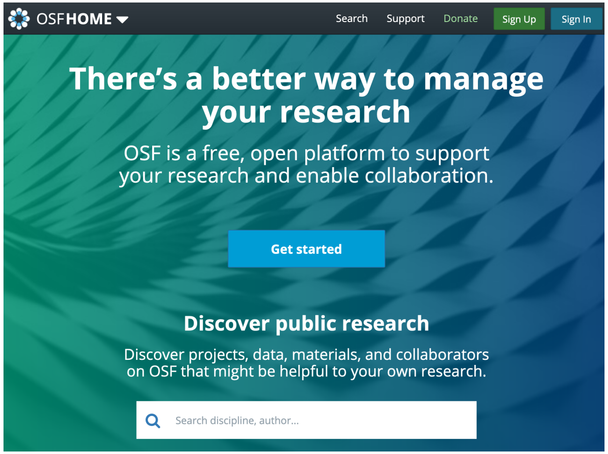
```
plot(density(mtcars$mpg))
```

Software versions are easy to report, and we can **see both the code to produce results as well as the results**.

How Do I Share My Work?

- Using R without sharing will help to solve many issues with reproducibility, but making it accessible to others is even better.

- The OSF is the most user-friendly method of sharing all your research.
    - Servers paid for 50 years, with servers in Europe.
    - **Drag and drop** to upload files.
    - Automatically **tracks updates** to files.
    - Can link **pre-registrations, licenses, and pre-prints** to your project.



There's a better way to manage your research

OSF is a free, open platform to support your research and enable collaboration.

Discover public research

Discover projects, data, materials, and collaborators on OSF that might be helpful to your own research.

The OSF

- GitHub is another option, which has more advanced **version control**, but can be trickier to use.

What We'll Do

After these lessons, you will have:

- used rstudio.cloud to learn the basics of how R works.
- created your first notebook.
- processed some raw data and cleaned it up.
- made summaries of our data.
- made graphs of data.
- shared all of this on The OSF.

It can be difficult at first, but it **will make things easier** when it comes to the assessment. Don't worry:

- You will struggle to remember the code, but **that's fine**.

- It's standard in most sciences, so if you have a question/issue, it's already been asked and answered somewhere. **Google is your friend**.

- You'll learn a valuable, **transferable skill**.