The results in Table 1 don't seem to correspond to those in Figure 2

Mine Çetinkaya-Rundel
University of Edinburgh + Duke University + RStudio

🔗 bit.ly/tab1-fig2-iscb

@minebocek
mine-cetinkaya-rundel
cetinkaya.mine@gmail.com


The results in Table 1 don’t seem to correspond to those in Figure 2!

61 45 4

94 12 3

```r
# set.seed
set.seed(20190314)

# generate 8 random numbers between 0 and 99
runif(8, 0, 99) %>% round()
```

20 44

more than 70 percent

have tried and **failed** to reproduce
another scientist's experiments

Baker, Monya. "1,500 scientists lift the lid on reproducibility." Nature News 533.7604 (2016): 452.

more than **50** percent

have tried and **failed** to reproduce
their **own** experiments

Baker, Monya. "1,500 scientists lift the lid on reproducibility." Nature News 533.7604 (2016): 452.

Google Scholar yields

152

results containing the term **reproducibility crisis** just in **2019**

Google Scholar Search, March 13, 2019.

1992

earliest reference **reproducibility research***

Claerbout, Jon F., and Martin Karrenbach. "Electronic documents give reproducible research a new meaning."
SEG Technical Program Expanded Abstracts 1992. Society of Exploration Geophysicists, 1992. 601-604.

that I could find…

Electronic Documents Give Reproducible Research a New Meaning

RE1.3

Jon F. Claerbout and Martin Karrenbach, Stanford Univ.

SUMMARY

A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provides a concrete definition of reproducibility in computationally oriented research. Experience at the Stanford Exploration Project shows that preparing such electronic documents is little effort beyond our customary report writing; mainly, we need to file everything in a systematic way.

In 1990 we began experimenting with electronic documents that merge our scientific software with our word-processing software. A year later we manufactured a CD-ROM containing a new textbook, Joe Dellinger's doctoral dissertation, and two progress reports of the Stanford Exploration Project. We distributed these CD-ROMs[1] to sponsors and many friends at the 1991 SEG meeting.

In 1990, we set this sequence of goals:

- Learn how to merge a publication with its underlying computational analysis.
- Teach researchers how to prepare a document in a form where they themselves can reproduce their own research results a year or more later by "pressing a single button".
- Learn how to leave finished work in a condition where coworkers can reproduce the calculation including the final illustration by pressing a button in its caption.
- Prepare a complete copy of our local software environment so that graduating students can take their work away with them to other sites, press a button, and reproduce their Stanford work.
- Merge electronic documents written by multiple authors (SEP reports).
- Export electronic documents to numerous other sites (sponsors) so they can readily reproduce a substantial portion of our Stanford research.

We met all these goals and set new ones:

- produce all new documents in this form, including lab reports in formal classes and "lab notebooks" of research progress.
- make incremental improvements in electronic-document software
- seek partners for broadening standards (and making incremental improvements).

Our basic goal is reproducible research. The electronic document is our means to this end. In principle, reproducibility in research can be achieved without electronic documents and that is how we started. Our first nonelectronic reproducible document was a textbook in which the paper document contained the name of a program script in every figure caption. The program scripts were organized by book chapter and section so they could be correlated to an accompanying magnetic tape dump of the file system. The magnetic tape also contained all the necessary data to feed the program script.

Now that we have begun using CD-ROM publication, we can go much further. Every figure caption contains a pushbutton that jumps to the appropriate science directory (folder) and initiates a figure rebuild command and then displays the figure, possibly as a movie or interactive program. We normally display seismic images of the earth's interior, but to reach wider audiences, Figure 1 shows a satellite weather picture which the pushbutton will animate as seen on commerical television. We include all our plot software as well as freely available software from many sources, including compilers and the LaTeX word processing system. Naturally we cannot include licensed software, but with the exception of Fortran and C compilers and the UNIX system itself, our publication includes source code for everything needed. The CD-ROM, at 680 megabytes, is so large we have had room for many executable programs on popular brands of workstations. The presence of these executables gives our readers a fast start.

Nearly everyone would rather read a paper book than the bitmapped page images on a screen that you see with an electronic document. But the illustrations in the electronic book are mostly in color, many are movies, and some are interactive. So the electronic book gives the reader a better understanding of the results. We typically use an interactive movie program to compare seismic sections where successive frames include processing with various parameters. The movie medium is much more informative than comparing seismic sections side by side. 3-D volumes are much better exhibited by movies than static paper illustrations. We are delivering a volume of software that is accessed like a book.

[1]SEP-CD-1 is available from Stanford University Press, $15 plus shipping, tel 415-723-1593

In 1990, we set this sequence of goals:

- Learn how to merge a publication with its underlying computational analysis.
- Teach researchers how to prepare a document in a form where they themselves can reproduce their own research results a year or more later by "pressing a single button".
- Learn how to leave finished work in a condition where coworkers can reproduce the calculation including the final illustration by pressing a button in its caption.
- Prepare a complete copy of our local software environment so that graduating students can take their work away with them to other sites, press a button, and reproduce their Stanford work.
- Merge electronic documents written by multiple authors (SEP reports).
- Export electronic documents to numerous other sites (sponsors) so they can readily reproduce a substantial portion of our Stanford research.

Claerbout, Jon F., and Martin Karrenbach. "Electronic documents give reproducible research a new meaning." SEG Technical Program Expanded Abstracts 1992. Society of Exploration Geophysicists, 1992. 601-604.

Electronic Documents Give Reproducible Research a New Meaning **RE1.3**

Jon F. Claerbout and Martin Karrenbach, Stanford Univ.

SUMMARY

A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provides a concrete definition of reproducibility in computationally oriented research. Experience at the Stanford Exploration Project shows that preparing such electronic documents is little effort beyond our customary report writing; mainly, we need to file everything in a systematic way.

In 1990 we began experimenting with electronic documents that merge our scientific software with our word-processing software. A year later we manufactured a CD-ROM containing a new textbook, Joe Dellinger's doctoral dissertation, and two progress reports of the Stanford Exploration Project. We distributed these CD-ROMs[1] to sponsors and many friends at the 1991 SEG meeting.

In 1990, we set this sequence of goals:

- Learn how to merge a publication with its underlying computational analysis.
- Teach researchers how to prepare a document in a form where they themselves can reproduce their own research results a year or more later by "pressing a single button".
- Learn how to leave finished work in a condition where coworkers can reproduce the calculation including the final illustration by pressing a button in its caption.
- Prepare a complete copy of our local software environment so that graduating students can take their work away with them to other sites, press a button, and reproduce their Stanford work.
- Merge electronic documents written by multiple authors (SEP reports).
- Export electronic documents to numerous other sites (sponsors) so they can readily reproduce a substantial portion of our Stanford research.

We met all these goals and set new ones:

- produce all new documents in this form, including lab reports in formal classes and "lab notebooks" of research progress.
- make incremental improvements in electronic-document software
- seek partners for broadening standards (and making incremental improvements).

[1]SEP-CD-1 is available from Stanford University Press, $15 plus shipping, tel 415-723-1593

Our basic goal is reproducible research. The electronic document is our means to this end. In principle, reproducibility in research can be achieved without electronic documents and that is how we started. Our first nonelectronic reproducible document was a textbook in which the paper document contained the name of a program script in every figure caption. The program scripts were organized by book chapter and section so they could be correlated to an accompanying magnetic tape dump of the file system. The magnetic tape also contained all the necessary data to feed the program script.

Now that we have begun using CD-ROM publication, we can go much further. Every figure caption contains a pushbutton that jumps to the appropriate science directory (folder) and initiates a figure rebuild command and then displays the figure, possibly as a movie or interactive program. We normally display seismic images of the earth's interior, but to reach wider audiences, Figure 1 shows a satellite weather picture which the pushbutton will animate as seen on commerical television. We include all our plot software as well as freely available software from many sources, including compilers and the LaTeX word processing system. Naturally we cannot include licensed software, but with the exception of Fortran and C compilers and the UNIX system itself, our publication includes source code for everything needed. The CD-ROM, at 680 megabytes, is so large we have had room for many executable programs on popular brands of workstations. The presence of these executables gives our readers a fast start.

Nearly everyone would rather read a paper book than the bitmapped page images on a screen that you see with an electronic document. But the illustrations in the electronic book are mostly in color, many are movies, and some are interactive. So the electronic book gives the reader a better understanding of the results. We typically use an interactive movie program to compare seismic sections where successive frames include processing with various parameters. The movie medium is much more informative than comparing seismic sections side by side. 3-D volumes are much better exhibited by movies than static paper illustrations. We are delivering a volume of software that is accessed like a book.

Now that we have begun using CD-ROM publication, we can go much further. Every figure caption contains a pushbutton that jumps to the appropriate science directory (folder) and initiates a figure rebuild command and then displays the figure, possibly as a movie or interactive program. We normally display seismic images of the earth's inte-

Claerbout, Jon F., and Martin Karrenbach. "Electronic documents give reproducible research a new meaning." SEG Technical Program Expanded Abstracts 1992. Society of Exploration Geophysicists, 1992. 601-604.

NEWS · 28 FEBRUARY 2019 ·

Pioneering 'live-code' article allows scientists to play with each other's results

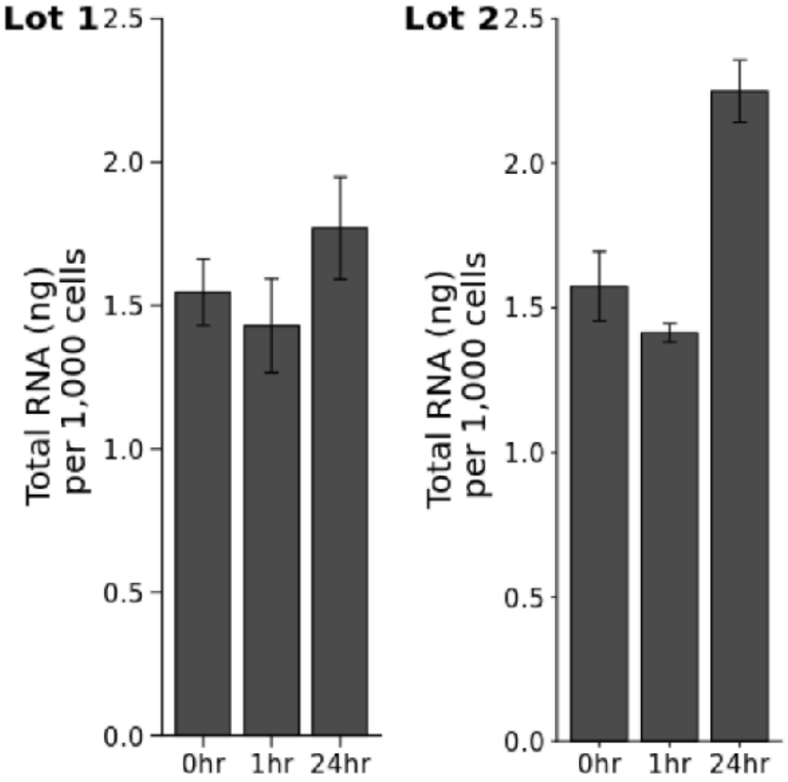
eLife's prototype lets scientists modify the software underlying figures to validate, build on, or better understand the work.

This is a Reproducible document. See the original article or source.

NOTE: Below is a reproducible version of Figure 1B. You can inspect the code, make changes and run the code by pressing SHIFT+ENTER. The data used can be downloaded here.

R Script

status: ready (run code with ⇧↵)



Total RNA levels following c-Myc overexpression

Lewis, L. Michelle, et al. "Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc." Elife 7 (2018): e30274.

setting the stage

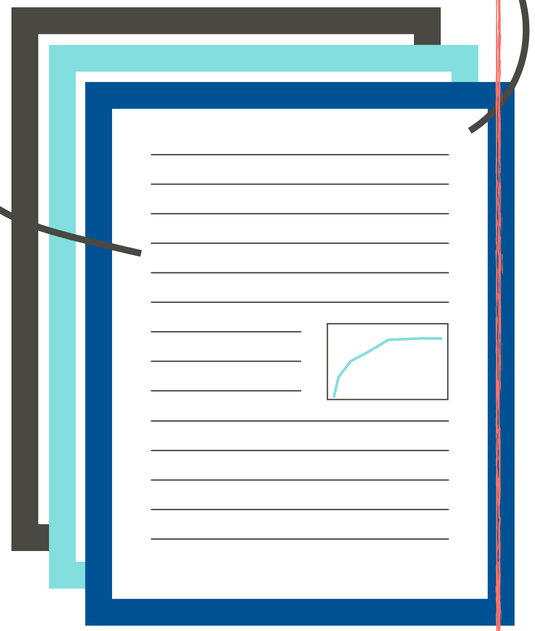
Photo by Alexander Dummer on Unsplash].

replicability

same research question

same results

new data



reproducibility

same research question

same results

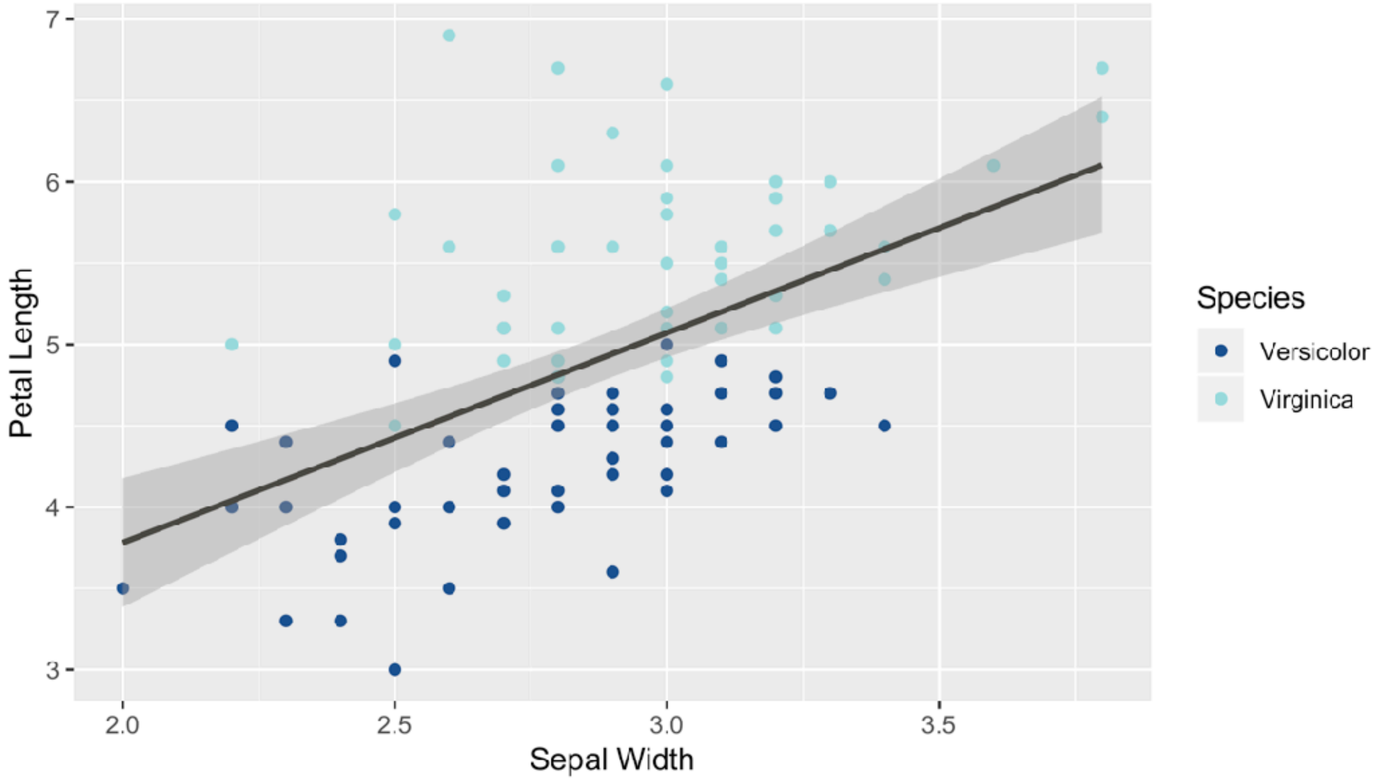
same data

e.g.

Table 1. Regression output for predicting petal length from sepal width.

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	9.06	0.929	9.76	1.13e-17
Sepal.Width	-1.74	0.301	-5.77	4.51e- 8

Figure 2. Relationship between petal length and sepal width

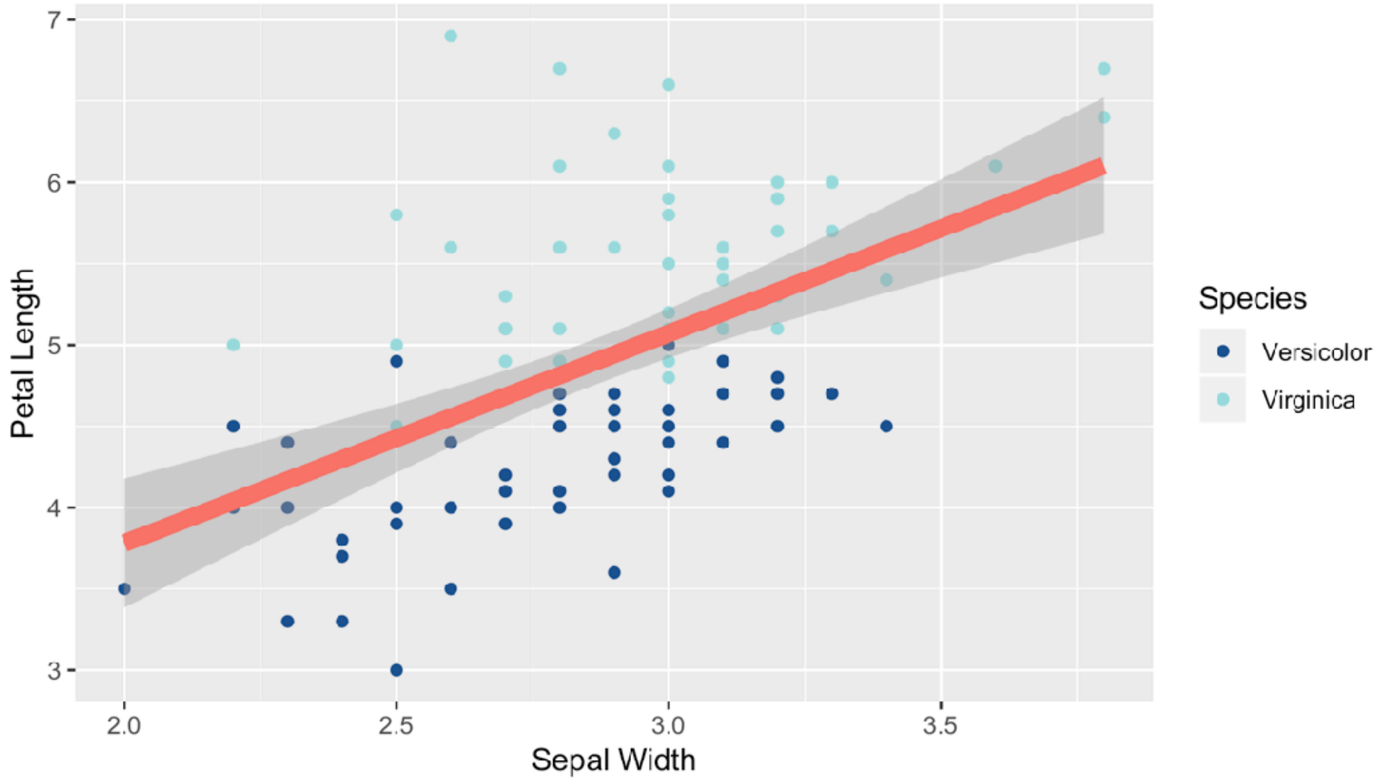


e.g.

Table 1. Regression output for predicting petal length from sepal width.

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	9.06	0.929	9.76	1.13e-17
Sepal.Width	-1.74	0.301	-5.77	4.51e- 8

Figure 2. Relationship between petal length and sepal width



analysis

```r
# fit model
model <- lm(Petal.Length ~ Sepal.Width, data = iris)

# print model summary
tidy(model)
```

report

Table 1. Regression output for predicting petal length from sepal width.

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	9.06	0.929	9.76	1.13e-17
Sepal.Width	-1.74	0.301	-5.77	4.51e- 8

analysis

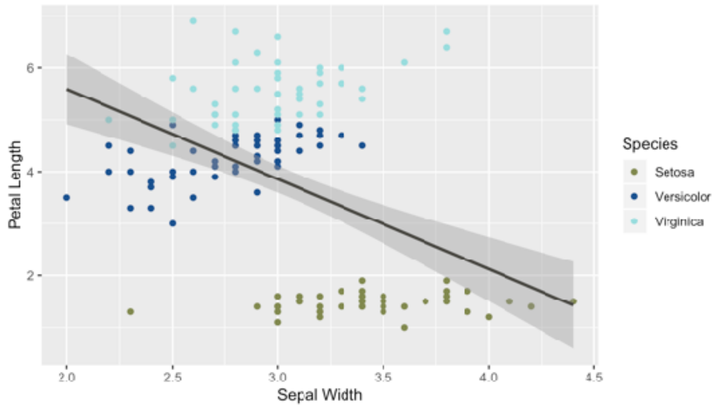
```
# visualize the relationship
ggplot(iris) +
  geom_point(
    aes(x = Sepal.Width, y = Petal.Length, color = Species)
  ) +
  geom_smooth(
    aes(x = Sepal.Width, y = Petal.Length),
    method = "lm"
  )
```

report

Table 1. Regression output for predicting petal length from sepal width.

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	9.06	0.929	9.76	1.13e-17
Sepal.Width	-1.74	0.301	-5.77	4.51e- 8

Figure 2. Relationship between petal length and sepal width



analysis

```r
# filter out Setosas
iris_nonsetosa <- iris %>%
  filter(Species != "setosa")

# visualize the relationship
ggplot(iris_nonsetosa) +
  geom_point(
    aes(x = Sepal.Width, y = Petal.Length, color = Species)
  ) +
  geom_smooth(
    aes(x = Sepal.Width, y = Petal.Length),
    method = "lm"
  )
```

report

Table 1. Regression output for predicting petal length from sepal width.

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	9.06	0.929	9.76	1.13e-17
Sepal.Width	-1.74	0.301	-5.77	4.51e- 8

Figure 2. Relationship between petal length and sepal width

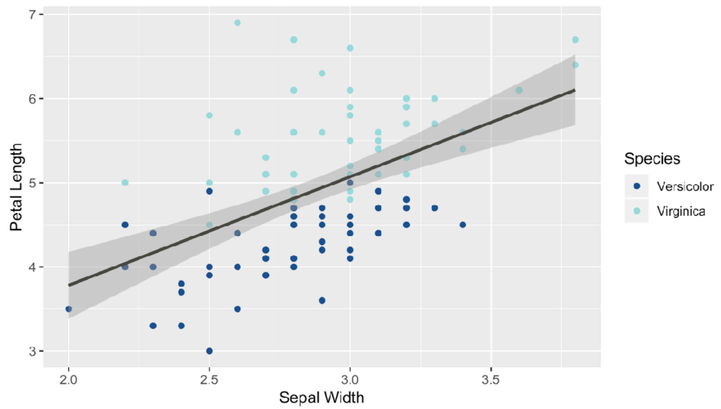
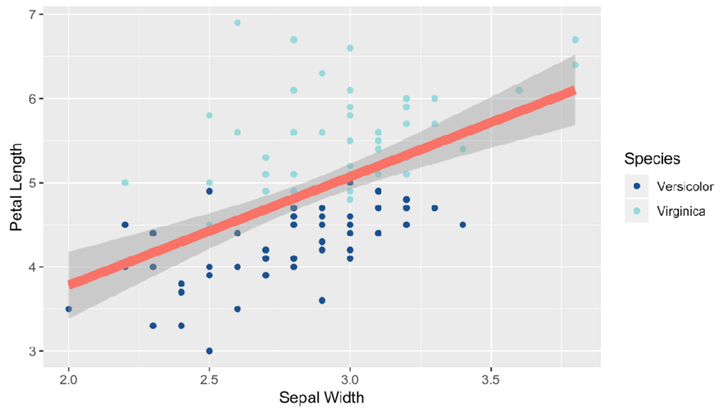
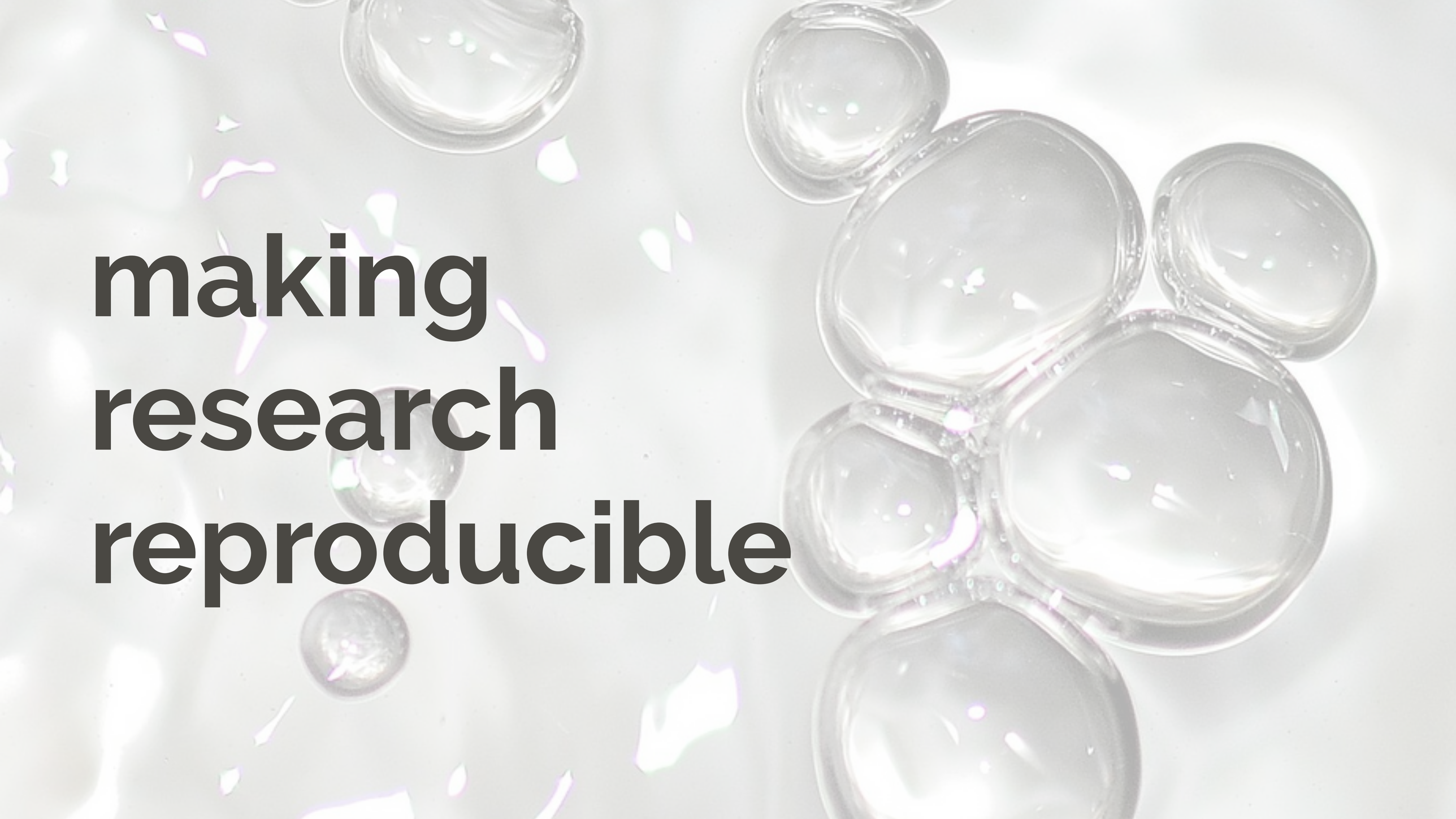


Table 1. Regression output for predicting petal length from sepal width.

Term	Estimate	Std. Error	Statistic	p-value
(Intercept)	9.06	0.929	9.76	1.13e-17
Sepal.Width	-1.74	0.301	-5.77	4.51e- 8

Figure 2. Relationship between petal length and sepal width




making
research
reproducible

“An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”

– David Donoho, paraphrasing Jon Claerbout

Buckheit, Jonathan B., and David L. Donoho. "Wavelab and reproducible research." Wavelets and statistics. Springer, New York, NY, 1995. 55-81.

make

raw data

code & documentation to reproduce the analysis

specifications of your computational environment

available and accessible

Peng, Roger. "The reproducibility crisis in science: A statistical counterattack." Significance 12.3 (2015): 30-32.
Gentleman, Robert, and Duncan Temple Lang. "Statistical analyses and reproducible research." Journal of Computational and Graphical Statistics 16.1 (2007): 1-23.

“The most important tool is the **mindset**, when starting, that the end product will be reproducible.”

– Keith Baggerly

nobody,
not even yourself,
can recreate any part
of your analysis

push button
reproducibility
in published work

“There’s **no one-size-fits-all solution** for computational reproducibility.”

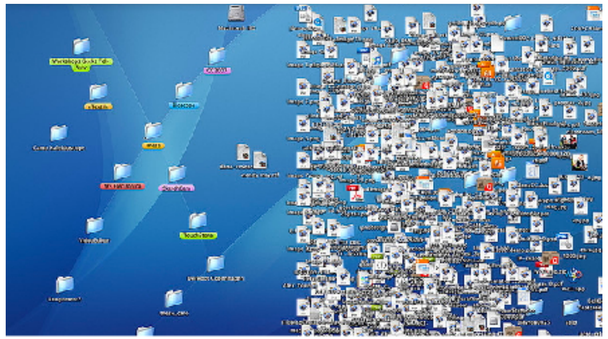
Perkel, Jeffrey M. "A toolkit for data transparency takes shape." Nature 560 (2018): 513-515.

8 principles

but the following might help…

1 organize your project

level of organization



simpler analysis

- raw-data
- processed-data
- manuscript
  - manuscript.Rmd

more complex analysis

- raw-data
- processed-data
- scripts
- figures
- manuscript
  - manuscript.Rmd

stick with the conventions of your peers

write READMEs liberally

raw-data

├ README.md

├ airlines.csv

├ airports.csv

├ flights.csv

├ planes.csv

├ weather.csv

processed-data

scripts

figures

manuscript

README

This folder contains the raw data for the project.
All datasets were downloaded from openflights.org/data.html on 2019-04-01.

- airlines: Airline names
- airports: Airports metadata
- flights: Flight data
- planes: Plane metadata
- weather: Hourly weather data

keep data tidy & machine readable

Student	Exam Grade		
Name	1	2	Major
Barney Donaldson	89	76	Data Science, Public Policy
Clay Whelan	67	83	Public Policy
Simran Bass	82	90	Statistics
Chante Munro	45	72	Political Science, Statistics
Gabrielle Cherry	32	79	.
Kush Piper	98	sick	Statistics
Faizan Ratliff	82	75	Data Science
Torin Ruiz	70	80	Sociology, Statistics
Reiss Richardson	missed exam	34	Neuroscience
Ajwa Cochran	50	65	Data Science

record code + document non-code steps + write tests

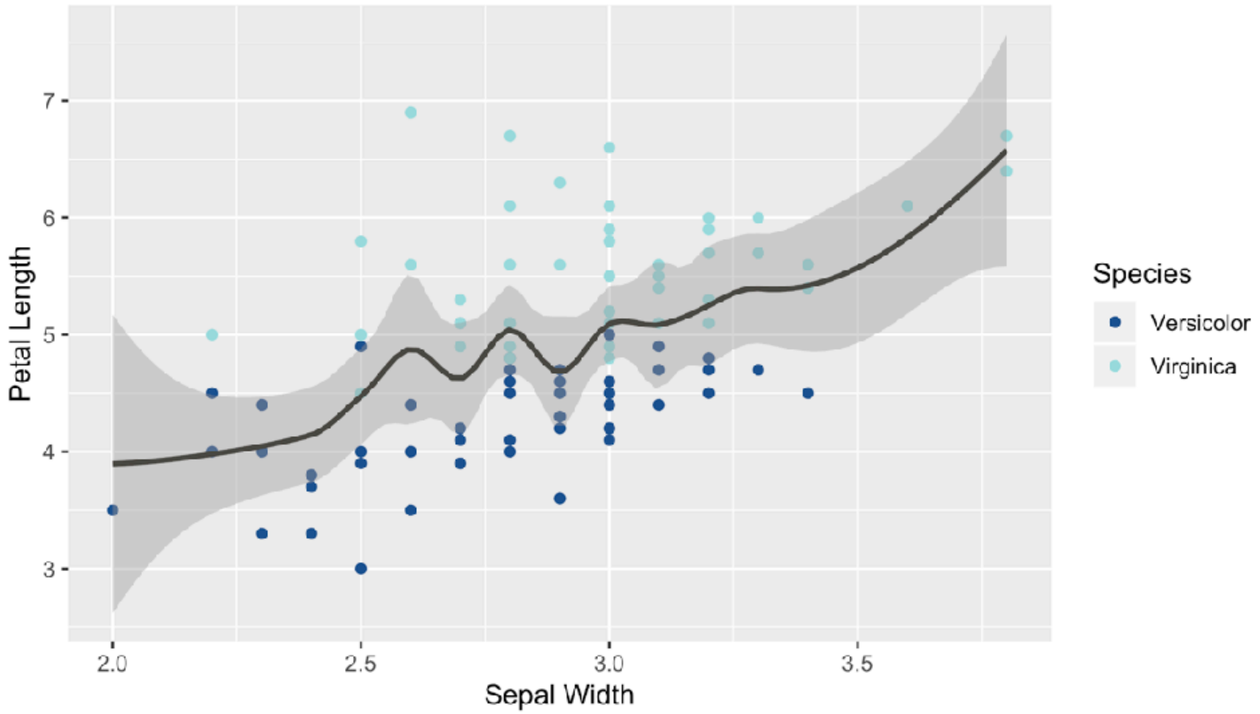
name	exam_1	exam_2	first_major	second_major	participation
Barney Donaldson	89	76	Data Science	Public Policy	ok
Clay Whelan	67	83	Public Policy	NA	ok
Simran Bass	82	90	Statistics	NA	ok
Chante Munro	45	72	Political Science	Statistics	Low
Gabrielle Cherry	32	79	NA	NA	ok
Kush Piper	98	NA	Statistics	NA	ok
Faizan Ratliff	82	75	Data Science	NA	ok
Torin Ruiz	70	80	Sociology	Statistics	ok
Reiss Richardson	NA	34	Neuroscience	NA	low
Ajwa Cochran	50	65	Data Science	NA	low

Low participation

Broman, Karl W., and Kara H. Woo. "Data organization in spreadsheets." The American Statistician 72.1 (2018): 2-10.

comment your code

```r
# use loess smoothing
ggplot(iris_nonsetosa) +
  geom_point(
    aes(x = Sepal.Width, y = Petal.Length, color = Species)
  ) +
  geom_smooth(
    aes(x = Sepal.Width, y = Petal.Length),
    method = "loess", span = 0.375
  )
```



use literate programming

~/Desktop/Talks/DataTech/datatech-2019 - master - RStudio Source Editor

tab1-fig2.Rmd

Knit | Insert | Run

```
---
title: "Table 1 matches Figure 2!"
author: "Mine Çetinkaya-Rundel"
date: "`r Sys.Date()`"
output:
  html_document:
    fig_caption: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, message = FALSE)
```

```{r}
library(tidyverse)
conflict_prefer("filter", "dplyr")
library(broom)
library(knitr)
```

In this report we evaluate the relationship between relationship between petal length and sepal width for irises.

```{r}
iris_nonsetosa <- iris %>%
  filter(Species != "setosa")
```

The original dataset has `r iris$Species %>% unique() %>% length()` species, but we will only work with `r iris_nonsetosa$Species %>% unique() %>% length()` of them.

Model

The model results are below.
```

2:34 | Table 1 matches Figure 2! | R Markdown

tab1-fig2.Rmd

Knit | Insert | Run

```
---
title: "Table 1 matches Figure 2!"
author: "Mine Çetinkaya-Rundel"
date: "`r Sys.Date()`"
output:
  html_document:
    fig_caption: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, message = FALSE)
```

```{r}
library(tidyverse)
conflict_prefer("filter", "dplyr")
library(broom)
library(knitr)
```

In this report we evaluate the relationship between relationship between petal length and sepal width for irises.

```{r}
iris_nonsetosa <- iris %>%
  filter(Species != "setosa")
```

The original dataset has `r iris$Species %>% unique() %>% length()` species, but we will only work with `r iris_nonsetosa$Species %>% unique() %>% length()` of them.
```

~/Desktop/Talks/DataTech/datatech-2019/scripts/tab1-fig2/tab1-fig2.html

tab1-fig2.html | Open in Browser | Find | Publish

Table 1 matches Figure 2!

Mine Çetinkaya-Rundel

2019-03-14

```
library(tidyverse)
conflict_prefer("filter", "dplyr")
library(broom)
library(knitr)
```

In this report we evaluate the relationship between relationship between petal length and sepal width for irises.

```
iris_nonsetosa <- iris %>%
  filter(Species != "setosa")
```

The original dataset has 3 species, but we will only work with 2 of them.

Model

The model results are below.

```
m_pl_sw <- lm(Petal.Length ~ Sepal.Width, data = iris_nonsetosa)
tidy_m_pl_sw <- tidy(m_pl_sw)
kable(tidy_m_pl_sw,
      caption = "Table 1. Regression output for predicting petal length
      from sepal width.",
      digits = 2)
```

Table 1. Regression output for predicting petal length from sepal width.

term	estimate	std.error	statistic	p.value
(Intercept)	1.20	0.62	1.94	0.06
Sepal.Width	1.29	0.21	6.02	0.00

tab1-fig2.Rmd

Knit | Insert | Run

```
---
title: "Table 1 matches Figure 2!"
author: "Mine Çetinkaya-Rundel"
date: "`r Sys.Date()`"
output:
  html_document:
    fig_caption: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
```

```{r}
library(tidyverse)
conflict_prefer("filter", "dplyr")
library(broom)
library(knitr)
```

In this report we evaluate the relationship between relationship between petal
length and sepal width for irises.

```{r}
iris_nonsetosa <- iris %>%
  filter(Species != "setosa")
```

The original dataset has `r iris$Species %>% unique() %>% length()` species, but
we will only work with `r iris_nonsetosa$Species %>% unique() %>% length()` of
them.
```

~/Desktop/Talks/DataTech/datatech-2019/scripts/tab1-fig2/tab1-fig2.html

tab1-fig2.html | Open in Browser | Find | Publish

Table 1 matches Figure 2!

Mine Çetinkaya-Rundel

2019-03-14

In this report we evaluate the relationship between relationship between petal length and sepal width for irises.

The original dataset has 3 species, but we will only work with 2 of them.

Model

The model results are below.

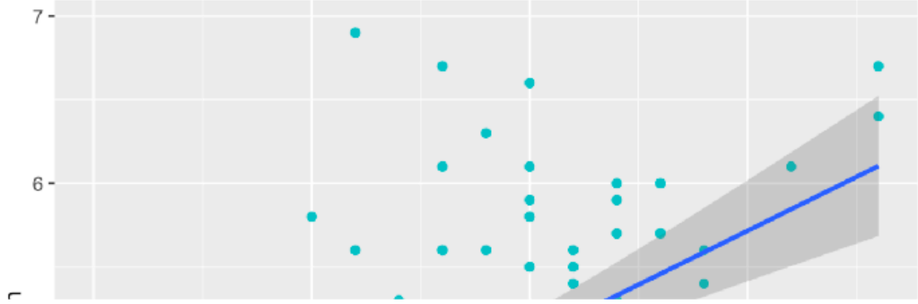
Table 1. Regression output for predicting petal length from sepal width.

term	estimate	std.error	statistic	p.value
(Intercept)	1.20	0.62	1.94	0.06
Sepal.Width	1.29	0.21	6.02	0.00

The slope of the regression model is 1.29.

Visualize

The figure below shows the relationship between these variables, and we observe a positive slope in this visualization as well.



use version control

tab1-fig2.Rmd

Knit | Insert | Run

```
---
title: "Table 1 matches Figure 2!"
author: "Mine Çetinkaya-Rundel"
date: "`r Sys.Date()`"
output:
  html_document:
    fig_caption: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
```

```{r}
library(tidyverse)
conflict_prefer("filter", "dplyr")
library(broom)
library(knitr)
```

In this report we evaluate the relationship between relationship between petal length and sepal width for irises.

```{r}
iris_nonsetosa <- iris %>%
  filter(Species != "setosa")
```

The original dataset has `r iris$Species %>% unique() %>% length()` species, but we will only work with `r iris_nonsetosa$Species %>% unique() %>% length()` of them.
```

changes tracked by  hosted on 

Happy Git and GitHub for the useR

Jenny Bryan, the STAT 545 TAs, Jim Hester

Let's Git started

Bryan, Jenny et. al. "Happy Git with R", happygitwithr.com.

7

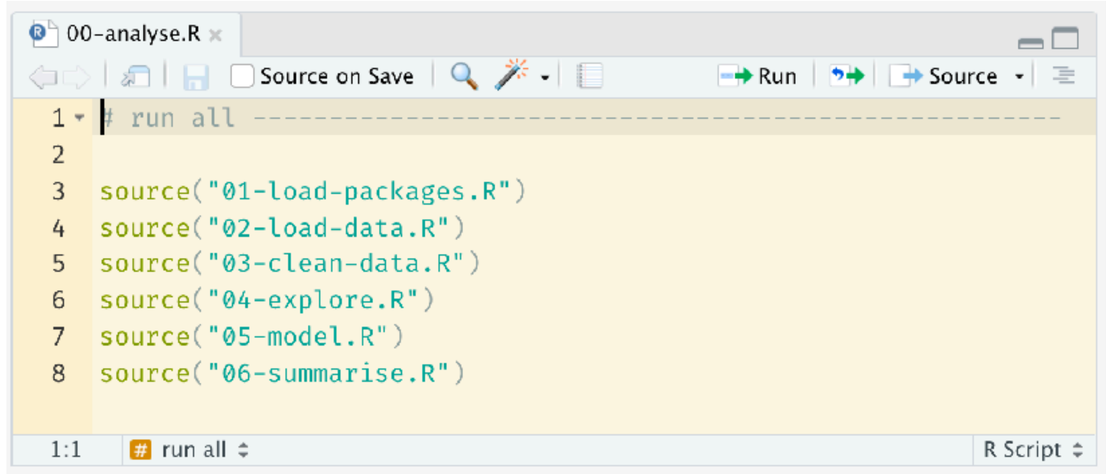
automate your process

raw-data

processed-data

scripts

- 00-analyse.R
- 01-load-packages.R
- 02-load-data.R
- 03-clean-data.R
- 04-explore.R
- 05-model.R
- 06-summarise.R



figures

manuscript

minimal make A minimal tutorial on make

I would argue that the most important tool for reproducible research is not Sweave or knitr but *GNU make*.

Consider, for example, all of the files associated with a manuscript. In the simplest case, I would have an R script for each figure plus a LaTeX file for the main text. And then a BibTeX file for the references.

Compiling the final PDF is a bit of work:

- Run each R script through R to produce the relevant figure.
- Run latex and then bibtex and then latex a couple of more times.

And the R scripts need to be run before latex is, and only if they've changed.

A simple example

GNU make makes this easy. In your directory for the manuscript, you create a text file called `Makefile` that looks something like the following (here using pdflatex).

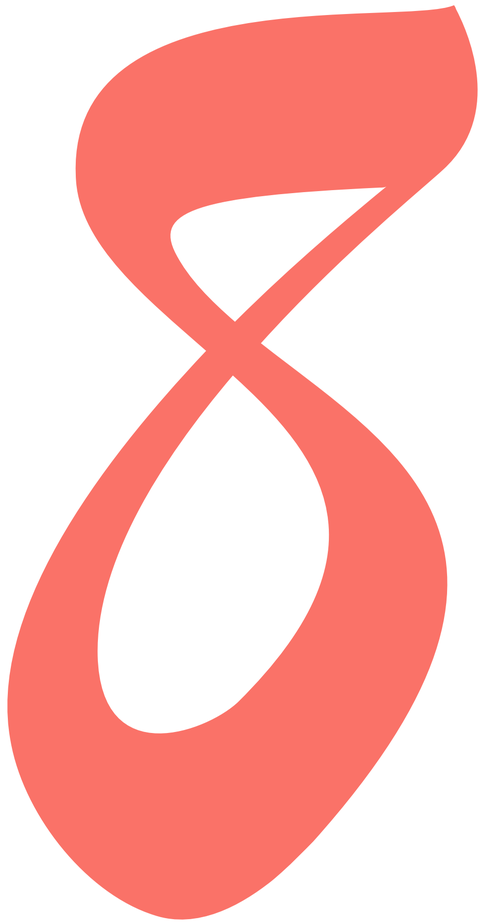
```
mypaper.pdf: mypaper.bib mypaper.tex Figs/fig1.pdf Figs/fig2.pdf
    pdflatex mypaper
    bibtex mypaper
    pdflatex mypaper
    pdflatex mypaper

Figs/fig1.pdf: R/fig1.R
    cd R;R CMD BATCH fig1.R

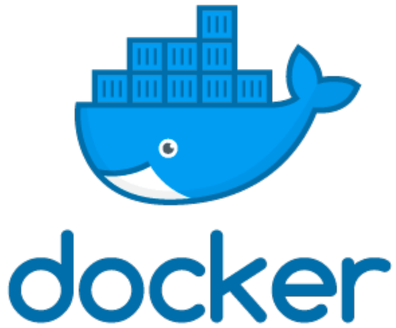
Figs/fig2.pdf: R/fig2.R
    cd R;R CMD BATCH fig2.R
```

Each batch of lines indicates a file to be created (the *target*), the files it depends on (the *prerequisites*), and then a set of commands needed to construct the target from the dependent files. Note that the lines with the commands *must* start with a **tab** character (**not spaces**).

Broman, Karl "Minimal Make", kbroman.org/minimal_make.



share computing environment


docker


binder


R Studio

1 **organize** your project

2 write **READMEs** liberally

3 keep data **tidy & machine readable**

4 **comment** your code

5 use **literate programming**

6 use **version control**

7 **automate** your process

8 share computing **environment**

Greg Wilson, Jennifer Bryan, Karen Cranston,
Justin Kitzes, Lex Nederbragt, Tracy K. Teal
“Good enough practices in scientific computing."
PLoS computational biology 13.6 (2017): e1005510.

PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson[1☯*], Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

1 Software Carpentry Foundation, Austin, Texas, United States of America, 2 RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, 3 Department of Biology, Duke University, Durham, North Carolina, United States of America, 4 Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, 5 Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, 6 Data Carpentry, Davis, California, United States of America

☯ These authors contributed equally to this work.
* gvwilson@software-carpentry.org

Author summary

Computers are now essential in all branches of science, but most researchers are never taught the equivalent of basic lab skills for research computing. As a result, data can get lost, analyses can take much longer than necessary, and researchers are limited in how effectively they can work with software and data. Computing workflows need to follow the same practices as lab projects and notebooks, with organized data, documented steps, and the project structured for reproducibility, but researchers new to computing often don't know where to start. This paper presents a set of good computing practices that every researcher can adopt, regardless of their current level of computational skill. These practices, which encompass data management, programming, collaborating with colleagues, organizing projects, tracking work, and writing manuscripts, are drawn from a wide variety of published sources from our daily lives and from our work with volunteer organizations that have delivered workshops to over 11,000 people since 2010.

Citation: Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

Editor: Francis Ouellette, Ontario Institute for Cancer Research, CANADA

Published: June 22, 2017

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Overview

We present a set of computing tools and techniques that every researcher can and should consider adopting. These recommendations synthesize inspiration from our own work, from the experiences of the thousands of people who have taken part in Software Carpentry and Data Carpentry workshops over the past 6 years, and from a variety of other guides. Our recommendations are aimed specifically at people who are new to research computing.

Introduction

Three years ago, a group of researchers involved in Software Carpentry and Data Carpentry wrote a paper called "Best Practices for Scientific Computing" [1]. That paper provided recommendations for people who were already doing significant amounts of computation in their research. However, as computing has become an essential part of science for all researchers, there is a larger group of people new to scientific computing, and the question then becomes, "where to start?"

looking into the future

Photo by Sweet Ice Cream Photography on Unsplash.

“The most important tool is the **mindset**, when starting, that the end product will be reproducible.”

– Keith Baggerly

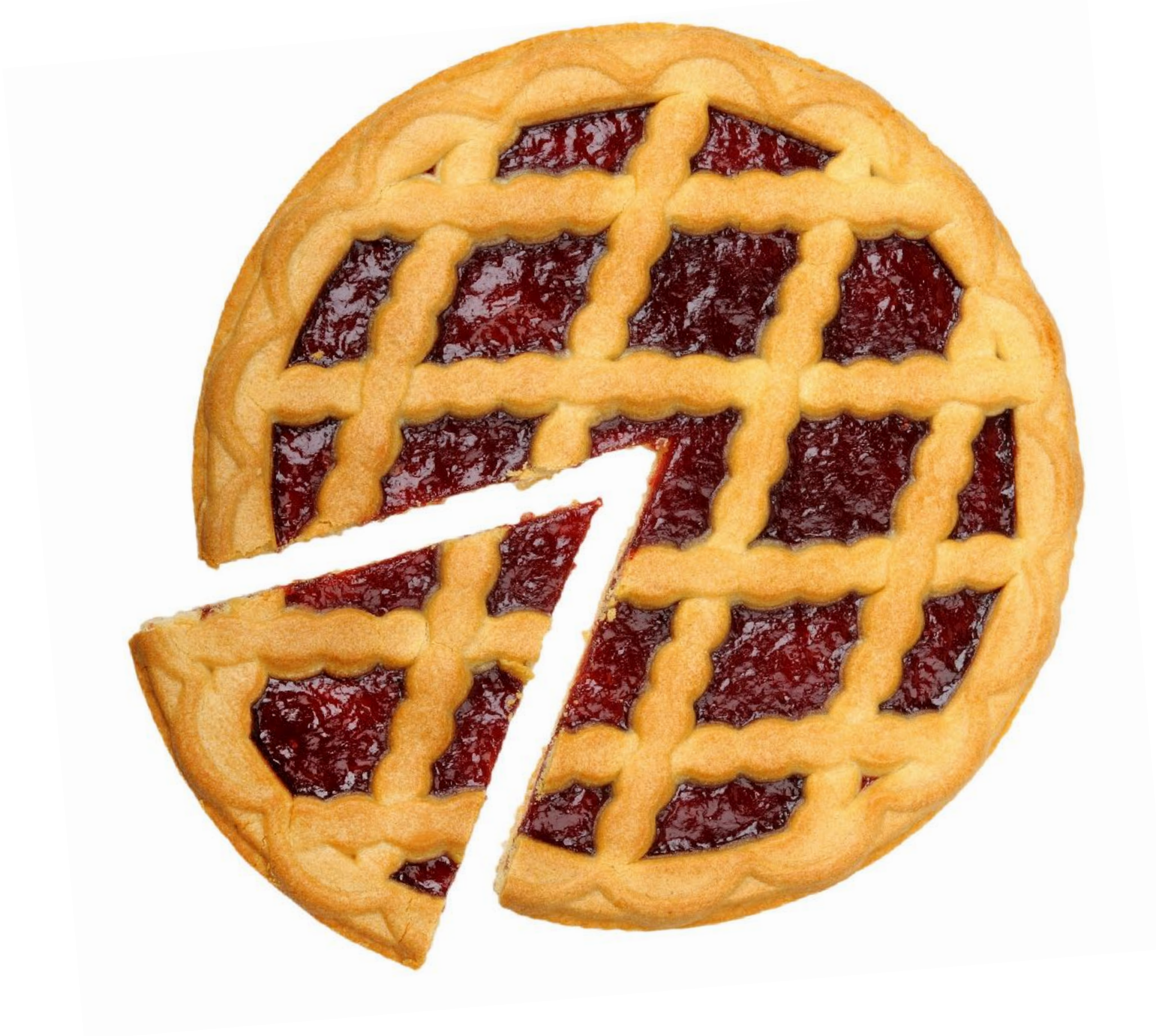
“The second most important tool is **training**.”

– Karl Broman

#1:
Convince data scientists to adopt a reproducible data analysis workflow

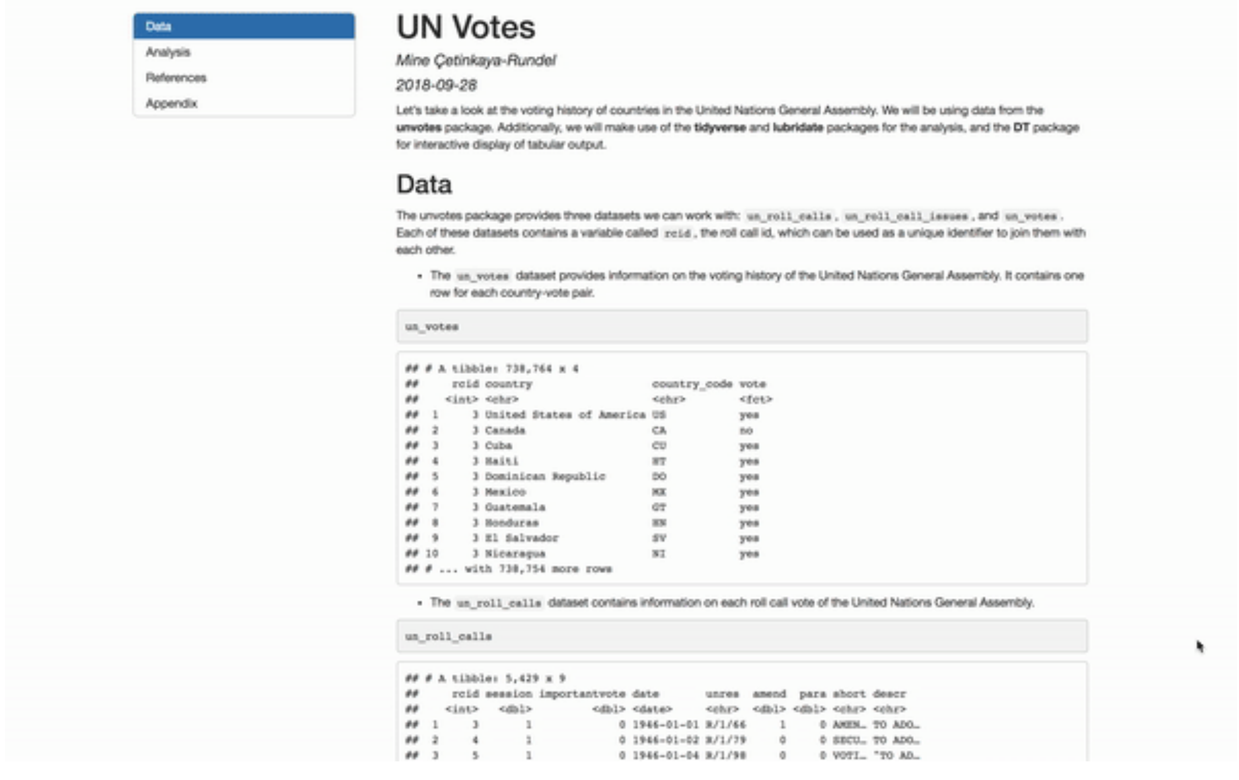
#2:
Train new data scientists who don't have any other workflow

statistics and data science educators who teach data analysis should be instilling best practices in students before they set out to do research



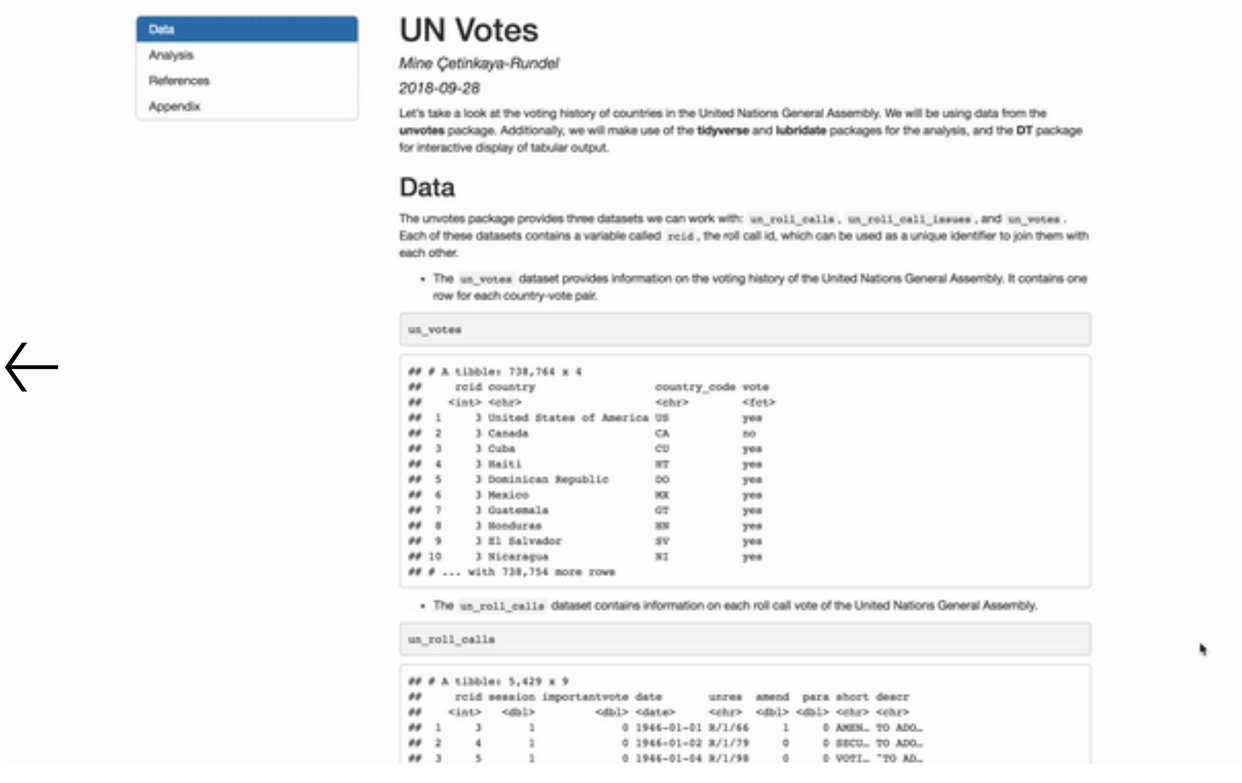

R Studio


rmarkdown


Data
Analysis
References
Appendix
UN Votes
Mine Çetinkaya-Rundel
2018-09-28
Let's take a look at the voting history of countries in the United Nations General Assembly. We will be using data from the unvotes package. Additionally, we will make use of the tidyverse and lubridate packages for the analysis, and the DT package for interactive display of tabular output.
Data
The unvotes package provides three datasets we can work with: un_roll_calls, un_roll_call_issues, and un_votes. Each of these datasets contains a variable called rcid, the roll call id, which can be used as a unique identifier to join them with each other.
The un_votes dataset provides information on the voting history of the United Nations General Assembly. It contains one row for each country-vote pair.
un_votes
A tibble: 738,764 x 4
rcid country country_code vote
<int> <chr> <chr> <fct>
1 3 United States of America US yes
2 3 Canada CA no
3 3 Cuba CU yes
4 3 Haiti HT yes
5 3 Dominican Republic DO yes
6 3 Mexico MX yes
7 3 Guatemala GT yes
8 3 Honduras HN yes
9 3 El Salvador SV yes
10 3 Nicaragua NI yes
... with 738,754 more rows
The un_roll_calls dataset contains information on each roll call vote of the United Nations General Assembly.
un_roll_calls
A tibble: 5,429 x 9
rcid session importantvote date unres amend para short descr
<int> <dbl> <dbl> <date> <chr> <dbl> <dbl> <chr> <chr>
1 3 1 0 1946-01-01 R/1/66 1 0 AMEN… TO ADO…
2 4 1 0 1946-01-02 R/1/79 0 0 SECU… TO ADO…
3 5 1 0 1946-01-04 R/1/98 0 0 VOTI… "TO AD…


R Studio


rmarkdown


Data
Analysis
References
Appendix
UN Votes
Mine Çetinkaya-Rundel
2018-09-28
Let's take a look at the voting history of countries in the United Nations General Assembly. We will be using data from the unvotes package. Additionally, we will make use of the tidyverse and lubridate packages for the analysis, and the DT package for interactive display of tabular output.
Data
The unvotes package provides three datasets we can work with: un_roll_calls, un_roll_call_issues, and un_votes. Each of these datasets contains a variable called rcid, the roll call id, which can be used as a unique identifier to join them with each other.
The un_votes dataset provides information on the voting history of the United Nations General Assembly. It contains one row for each country-vote pair.
un_votes
A tibble: 738,764 x 4
rcid country country_code vote
<int> <chr> <chr> <fct>
1 3 United States of America US yes
2 3 Canada CA no
3 3 Cuba CU yes
4 3 Haiti HT yes
5 3 Dominican Republic DO yes
6 3 Mexico MX yes
7 3 Guatemala GT yes
8 3 Honduras HN yes
9 3 El Salvador SV yes
10 3 Nicaragua NI yes
... with 738,754 more rows
The un_roll_calls dataset contains information on each roll call vote of the United Nations General Assembly.
un_roll_calls
A tibble: 5,429 x 9
rcid session importantvote date unres amend para short descr
<int> <dbl> <dbl> <date> <chr> <dbl> <dbl> <chr> <chr>
1 3 1 0 1946-01-01 R/1/66 1 0 AMEN… TO ADO…
2 4 1 0 1946-01-02 R/1/79 0 0 SECU… TO ADO…
3 5 1 0 1946-01-04 R/1/98 0 0 VOTI… "TO AD…

The results in Table 1 don’t seem to correspond to those in Figure 2

🔗 bit.ly/tab1-fig2-iscb

📂 bit.ly/tab1-fig2

@minebocek
mine-cetinkaya-rundel
cetinkaya.mine@gmail.com