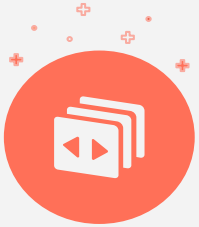


Iterating Efficiently with Parameterized Reporting in R Markdown

Jay Jeffries

Start now!

01.

Background

Non-exhaustive definition and summary of software & package utility

02.

Introduction

Setup, relevancy and intrigue into parameterized reporting

03.

Code Examples

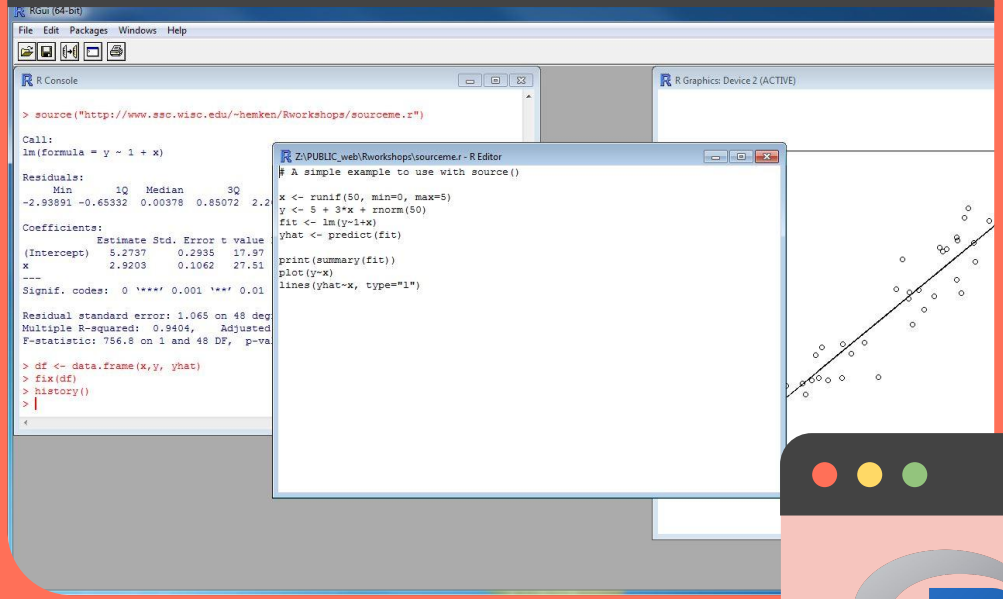
Social Capital Atlas
Grads on the Go
NCES

01.

Future Directions

How do I plan to implement this?
Where can you take this?

R Interface



R Language

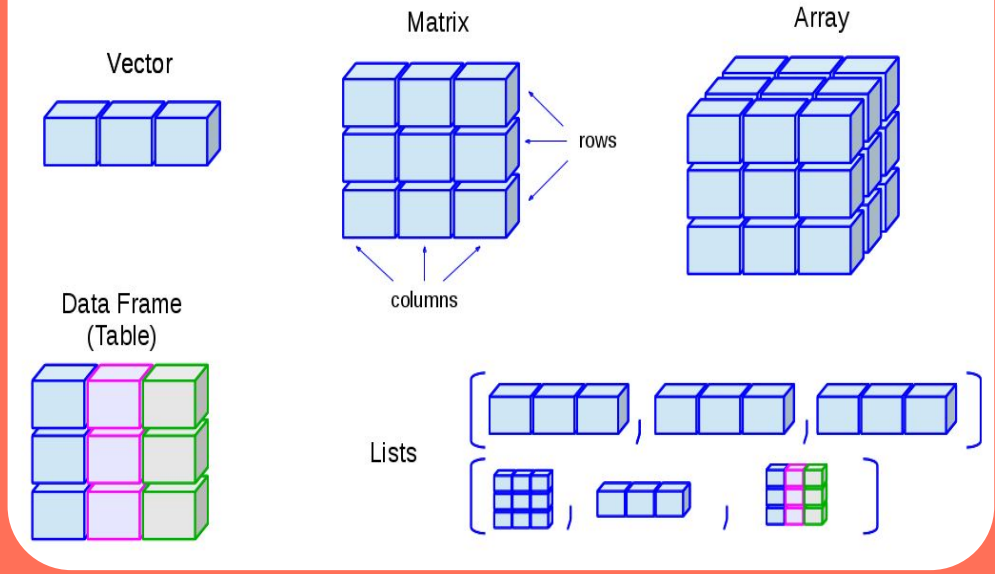
Object-oriented language

- 5 object "classes": character, numeric, integer, complex, and logical (T/F)

Can read and use a variety of languages and statistical software syntax

- Reads C, C++, C#, .NET, Java, and Fortran (if compiler available), Perl, Ruby, F#, and Julia
- {MplusAutomation} package
- {SASmarkdown} package and {SAS} engine in R Markdown
- {reticulate} package and {python} engine in R Markdown
- SQL with {sql} engine in R Markdown

R Interface



Data structures in R. Source: Ceballos and Cardiel 2013.

R Language

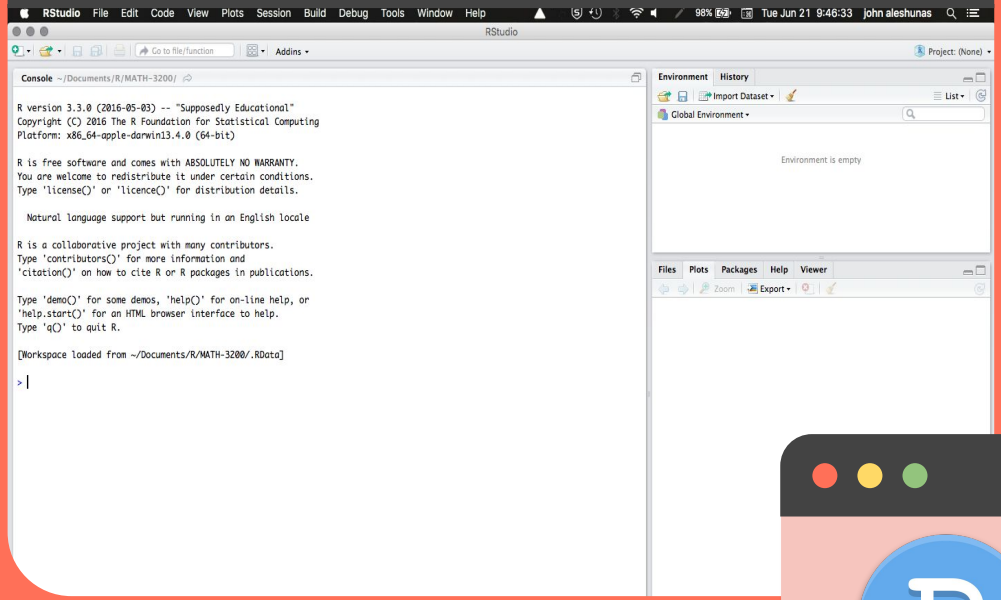
Object-oriented language

- 5 object "classes": character, numeric, integer, complex, and logical (T/F)

Can read and use a variety of languages and statistical software syntax

- Reads C, C++, C#, .NET, Java, and Fortran (if compiler available), Perl, Ruby, F#, and Julia
- {MplusAutomation} package
- {SASmarkdown} package and {SAS} engine in R Markdown
- {reticulate} package and {python} engine in R Markdown
- SQL with {sql} engine in R Markdown

R Studio Interface



R Studio

RStudio is an integrated development environment, or *IDE*

- Write & store your own functions
- Use others' functions via a library of "packages" from *CRAN*
  -
  -
- Create your own packages!

Many packages are crowd-sourced or created from voluntary contribution

Vibrant community of R package developers, collaborators, users!

1976

Statistical programming language, *S*, developed by John Chambers at Bell Labs as part of AT&T Corp.

1991

Ross Ihaka and Robert Gentleman develop *R* at the University of Auckland's Department of Statistics

1995

GNU General Public Licensing establishes R as free software in resistance to the commercialized *S*-Plus

Total Google Scholar citations: **429,466**

1997

R Core Team non-profit group formed for public interest to control source code, develop, and interact with community

2000

R Version 1.0.0 released!

2014

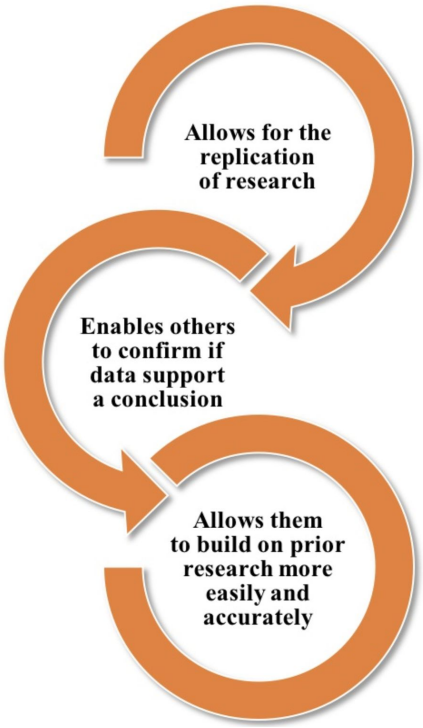
R Markdown pkg released by team: J.J. Allaire, Yihui Xie, Jonathan McPherson, Hadley Wickham, and others.

2022

R version 4.2.1 (Funny-Looking Kid) released on June 23rd

2022

R Markdown version 2.16 released



The CITI Program (2017)

Transparency & Code Sharing

jjeffries13

- *Sharing your code improves reproducibility*
- *Your code makes things transparent*
  - "Researcher degrees of freedom"
- *Online code supports collaboration*
  - Large costs of time and energy if collaborators wish to reuse/repurpose/extend code
- *More eyes on a project can catch more errors*
- *Creating shareable code from the beginning can actually save you time*
  - Can prevent irreparably breaking code
  - Can help you keep track of the changes you've made
  - Especially useful for "back burner" projects

Computational Reproducibility

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

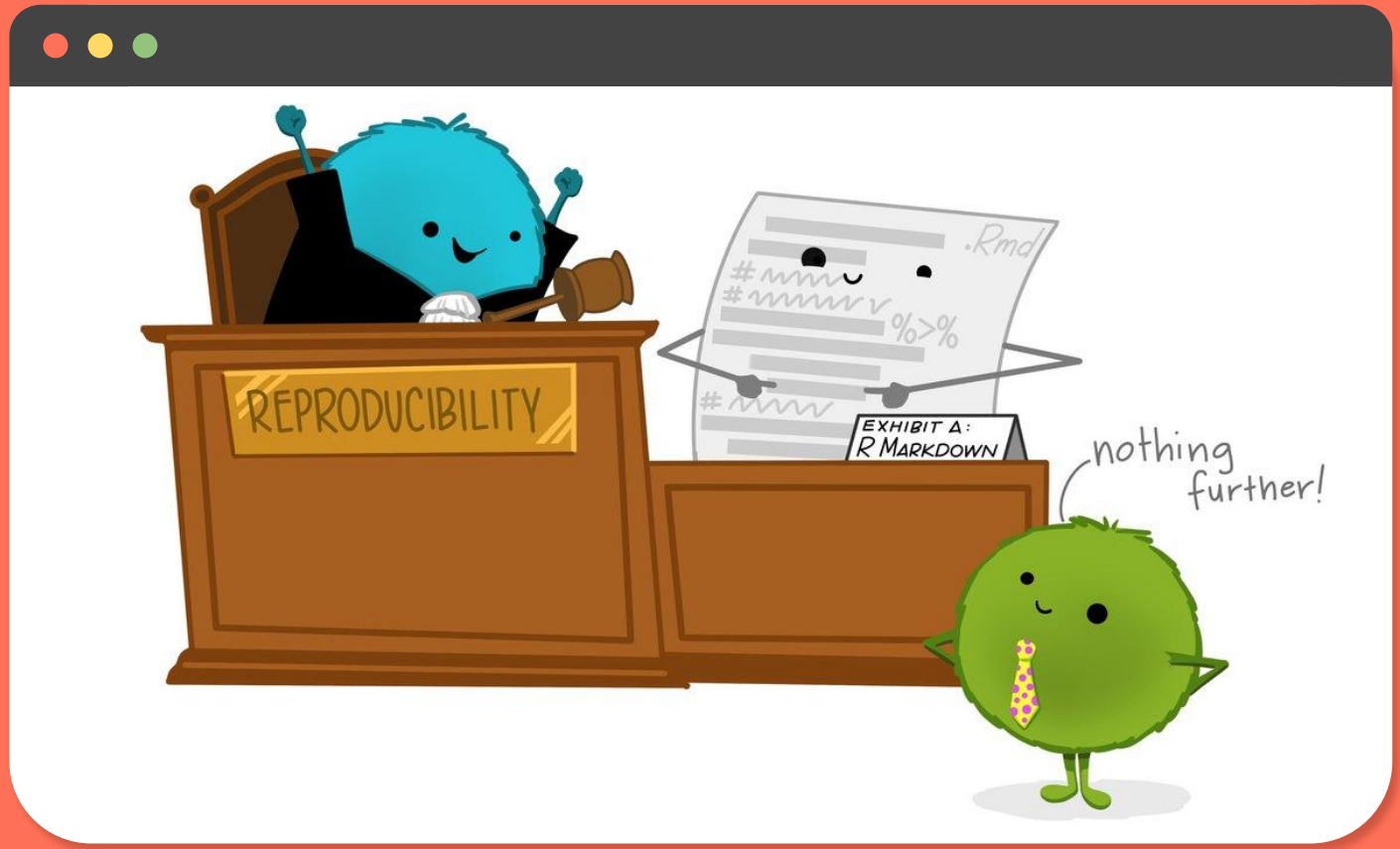
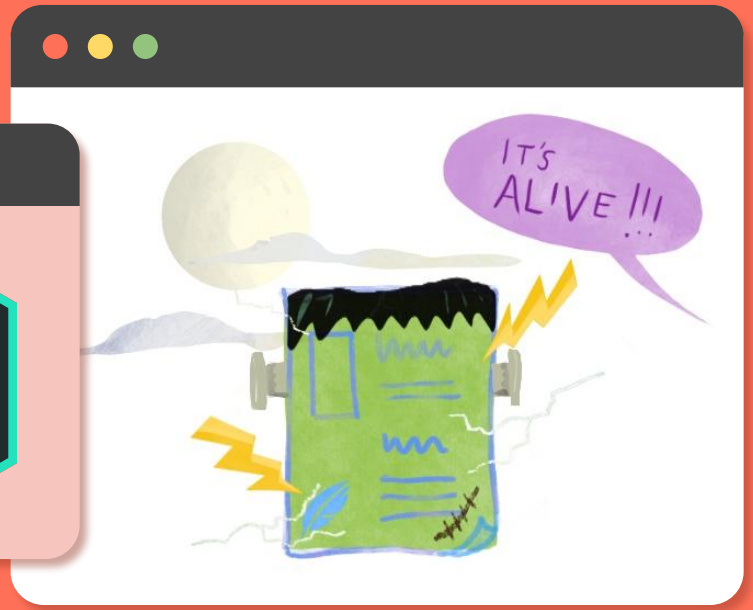


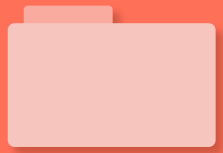
Image credit: Allison Horst

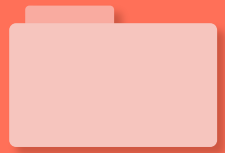
Introducing R Markdown

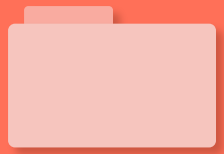
- An R package that enhances collaboration, communication, and reproducibility
- A file format (.Rmd) for making dynamic documents through R
- A tool that integrates R code, output, and text
- Works in conjunction with {knitr}

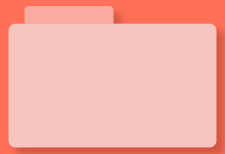


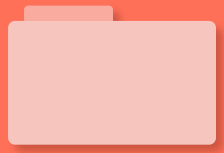
Artwork by Desireé De Leon

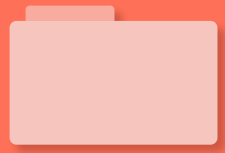

Raw Data

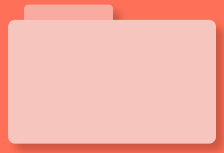

Processed Data

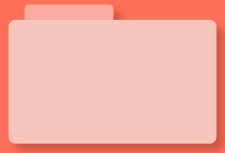

Descriptives


Models


Visuals


Write-Up & Reports


Presentation


Codebook

(Likely) Typical Workflow

- Separate files for raw and clean data
  a.
- Data codebook and users guide Word file(s)
- Software output
  a.
  b.
  c.
- Word files for write-up, separate reports
- PowerPoint presentation file

R Markdown script

YAML sets title, date, and output type

Code chunk loads packages and data

```
---
title: "Outbreak Situation Report"
date: "4/24/2021"
output: word_document
---

```{r setup, echo=FALSE}
pacman::p_load(rio, here, tidyverse, janitor, incidence2, flextable)
linelist <- rio::import(here::here("data", "case_linelists", "linelist_cleaned.rds"))
```

This report is for the Incident Command team of the fictional outbreak of Ebola cases.
**As of `r format(max(linelist$date_hospitalisation, na.rm=T), "%d %B")` there have
been `r nrow(linelist)` cases reported as hospitalized.**

Summary table of cases by hospital

```{r, echo=F, out.height="75%"}
linelist %>%
  filter(!is.na(hospital)) %>%
  group_by(hospital) %>%
  summarise(cases = n(),
    deaths = sum(outcome == "Death", na.rm=T),
    recovered = sum(outcome == "Recover", na.rm=T)) %>%
  adorn_totals() %>%
  qflextable()
```

Epidemic curve by age

```{r, echo=F, warning=F, message=F, out.height = "75%", out.width="100%"}
# create epicurve
age_outbreak <- incidence(
  linelist,
  date_index = date_onset,    # date of onset for x-axis
  interval = "week",          # weekly aggregation of cases
  groups = age_cat)

# plot
plot(age_outbreak, n_breaks = 3, fill = age_cat, col_pal = muted, title = "Epidemic
curve by age group")
```
```

Text and in-line code

Code chunk makes table

Headings

Code chunk makes plot

Output (e.g. Word document)

outbreak_report - Read-Only - Compatibility Mode

Search

ayout References Mailings Review View Zotero Help Picture Format

Outbreak Situation Report

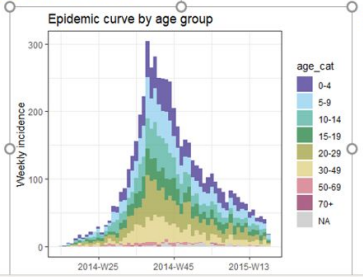
4/24/2021

This report is for the Incident Command team of the fictional outbreak of Ebola cases. **As of 30 April there have been 5888 cases reported as hospitalized.**

Summary table of cases by hospital

hospital	cases	deaths	recovered
Central Hospital	454	193	165
Military Hospital	896	399	309
Missing	1,469	611	514
Other	885	395	290
Port Hospital	1,762	785	579
St. Mark's Maternity Hospital (SMMH)	422	199	126
Total	5,888	2,582	1,983

Epidemic curve by age



Display Settings Focus

https://epirhandbook.com/en/reports-with-r-markdown.html

YAML header

Text prose

Code

Input

~/Documents/r4ds/r4ds/rmark

Addins

diamond-sizes.Rmd

Knit · Insert · Run

```
1  ---
2  title: "Diamond sizes"
3  date: 2016-08-25
4  output: html_document
5  ---
6
7  ```{r setup, include = FALSE}
8  library(ggplot2)
9  library(dplyr)
10
11 smaller <- diamonds %>%
12   filter(carat <= 2.5)
13 ```
14
15 We have data about `r nrow(diamonds)` diamonds. Only
16 `r nrow(diamonds) - nrow(smaller)` are larger than
17 2.5 carats. The distribution of the reminder is shown
18 below:
19
20 ```{r, echo = FALSE}
21 smaller %>%
22   ggplot(aes(carat)) +
23   geom_freqpoly(binwidth = 0.01)
24 ```
```

8:17 Chunk 1: setup — R Markdown

Console | R Markdown

~/Documents/r4ds/r4ds/rmarkdown/

```
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
```

Output

rkdown - master - RStudio

rmarkdown

Environment | History | Git

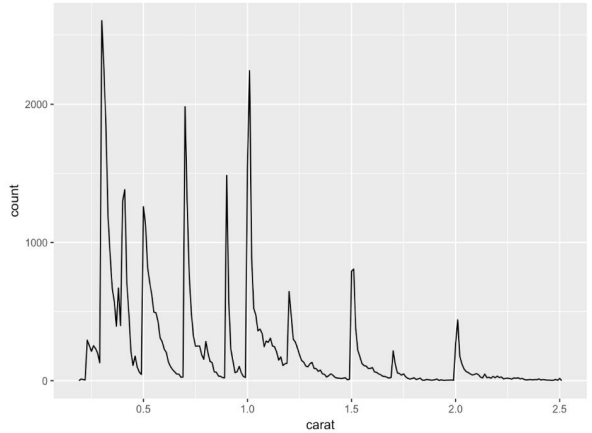
Files | Plots | Packages | Help | Viewer

Publish

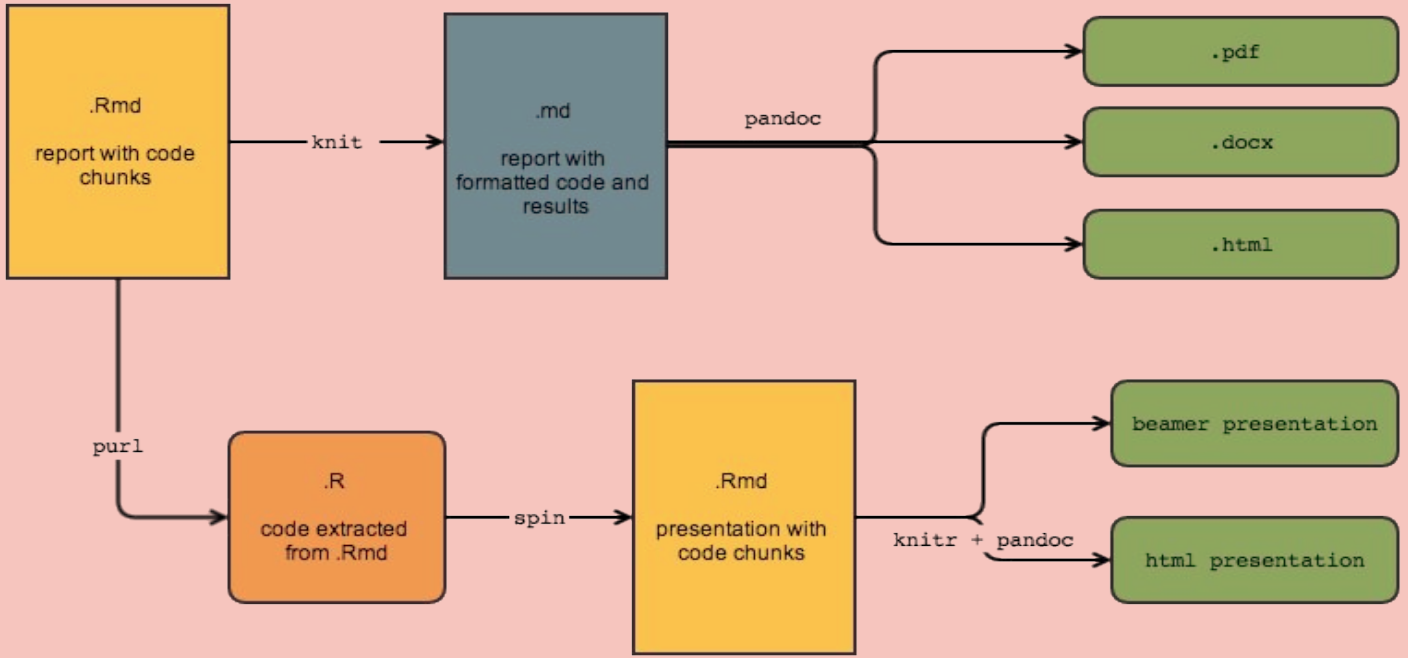
Diamond sizes

2016-08-25

We have data about 53940 diamonds. Only 126 are larger than 2.5 carats. The distribution of the reminder is shown below:



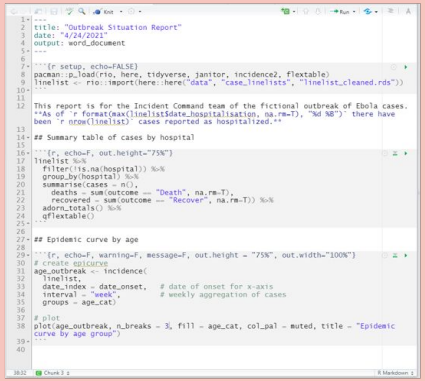
R Markdown Workflow



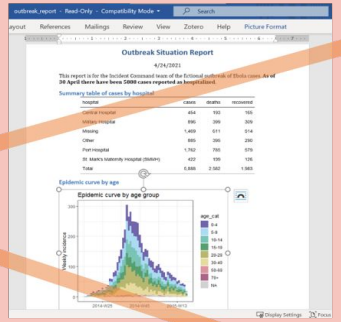
https://sachsmc.github.io/knit-git-markr-guide/knitr/knit.html

R Markdown Workflow

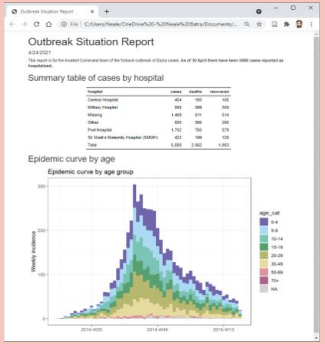
R Markdown script



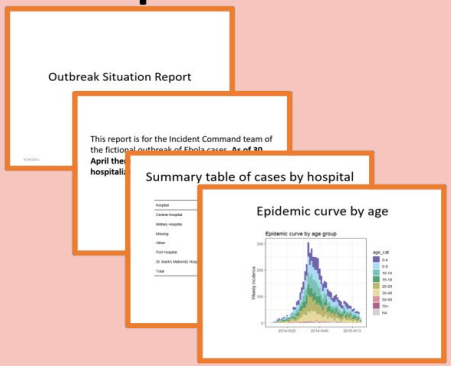
Word



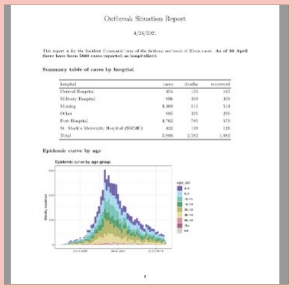
HTML



Powerpoint



PDF



https://epirhandbook.com/en/reports-with-r-markdown.html

Image credit: Allison Horst

Products

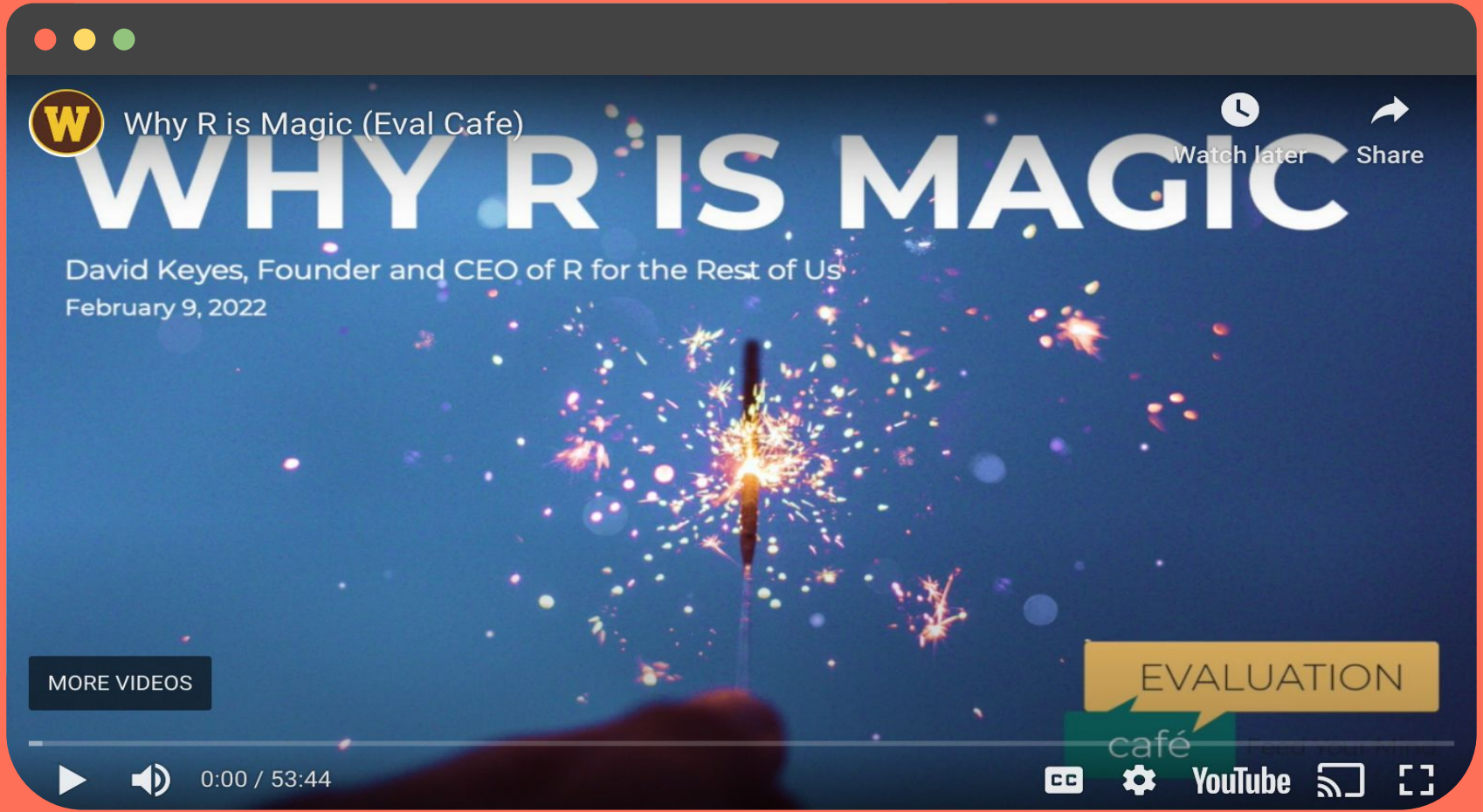
- Reports
- Slide shows
- Blogs, websites
- Theses, dissertation
- Manuscripts, papers
- Books
- Shiny apps, dashboards
- Generative art
- Maps and 3D models

Scale up the work you need.

Skip the work you don't.

See evidence of reproducibility.

Parameterized Reporting



https://rfortherestofus.com/2022/03/why-r-is-magic/

Parameters

- Data parameters (i.e., filters) are a list of specific values of a variable of interest
- Often clusters or strata
  -
  -
- Useful when re-rendering the same report with distinct values for key inputs.
- Declared via **params:** field in the YAML

"Can you run that analysis again for a different time period?"

"Can you run it for group B instead of group A?"

"Can you do this for every department?"

- meters (i.e., f st of specific values of a var st
- Often clusters or st
  -

Which variable would you use to filter for the provided examples?

- Useful when re rendering the same report with distinct values for key inputs.
- Declared via field in the YAML

Examples

Student exam scores by class

Sales reports by company branch

Demographics across voting districts

Quality of life characteristics for those with scores greater than 40 on the BDI across countries

Pollution and emission statistics for private American companies with more than 1,500 employees who received $500 million+ in allocated tax credits in 2022.

Parameters

- Data parameters (i.e., filters) are a list of specific values of a variable of interest
- Often clusters or strata
  -
  -
- Useful when re-rendering the same report with distinct values for key inputs.
- Declared via **params:** field in the YAML

Examples

Student exam scores by class
(class ID)

Sales reports by company branch
(branch ID)

Demographics across voting districts
(district ID)

Quality of life characteristics for those with scores greater than 40 on the BDI across countries
(score + country ID)

Emission statistics for private U.S. companies with more than 7,500 employees who received $500 million or more in allocated tax credits during the 2022 fiscal year.
(public/private + country + employee + tax credits + year)

Define params

```
classID <- tibble(
class =
"EDPS-459-001",
"EDPS-459-002",
"EDPS-459-003",
"EDPS-459-004",
"EDPS-459-005",
"EDPS-459-006"
)
```

Saved as an object!

YAML

```
---
Title: "Final Exam Report"
Date: 10-12-2022
output: html_document
params:
  classID: "EDPS-459-001"
---
```

Examples

Student exam scores by class
(class ID)

Sales reports by company branch
(branch ID)

Demographics across voting districts
(district ID)

Quality of life characteristics for those with scores greater than 40 on the BDI across countries
(score + country ID)

Emission statistics for private U.S. companies with more than 7,500 employees who received $500 million or more in allocated tax credits during the 2022 fiscal year.
(public/private + country + employee + tax credits + year)

Define params

```
company_list <- tibble(
public = 1, #dummy coded
country = US,
employee_count => 7500,
taxcred_alloc => 500000000,
year = 2022
)
```

YAML

```
---
Title: "Emissions Report"
Date: 10-12-2022
output: html_document
params:
  params: "General Motors"
---
```

Examples

Student exam scores by class
(class ID)

Sales reports by company branch
(branch ID)

Demographics across voting districts
(district ID)

Quality of life characteristics for those with scores greater than 40 on the BDI across countries
(score + country ID)

Emission statistics for private U.S. companies with more than 7,500 employees who received $500 million or more in allocated tax credits during the 2022 fiscal year.
(public/private + country + employee + tax credits + year)

Parameterized Report Workflow

- Single .Rmd file "template" for
  a.
  b.
  c.
  d.
  e.
- An .R file that loops through params to repeatedly render input .Rmd file
  a. Includes a defined list or dataframe of parameters
- Place to save output reports
  a. $N_{\text{reports}} = N_{\text{groups of interest}}$


Code Examples Time :)

Study

The Influence of Systematic Recurrent Feedback to Teachers in Higher Education on Student Perception of Classroom Experience

Om Joshi, Justin Andersson, Jay Jeffries

- Student survey distributed four times
  - Survey adapted from Teaching Analysis By Students (TABS)
- Individual results rendered into reports for each participating GTAs
- Report will include:
  - Descriptive statistics from Likert-scale responses
  - Free-response text

Future Directions

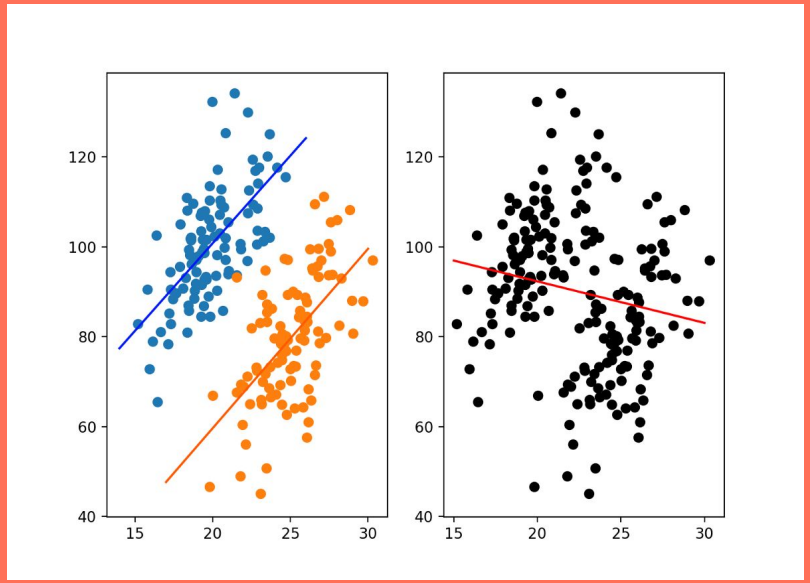
How can you see yourself implementing this in your work?

Have you seen this done in other contexts, fields or work, software?

A phenomenon in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Simpson's Paradox

- "The interpretation of interaction in contingency tables", Karl Pearson (1899)
- Early data showed case fatality rates for Covid-19 to be higher in Italy than in China overall, yet within every age group the fatality rate was higher in China than in Italy.
  -
- Parameterized reports may help disaggregate data to identify instances of Simpson's paradox



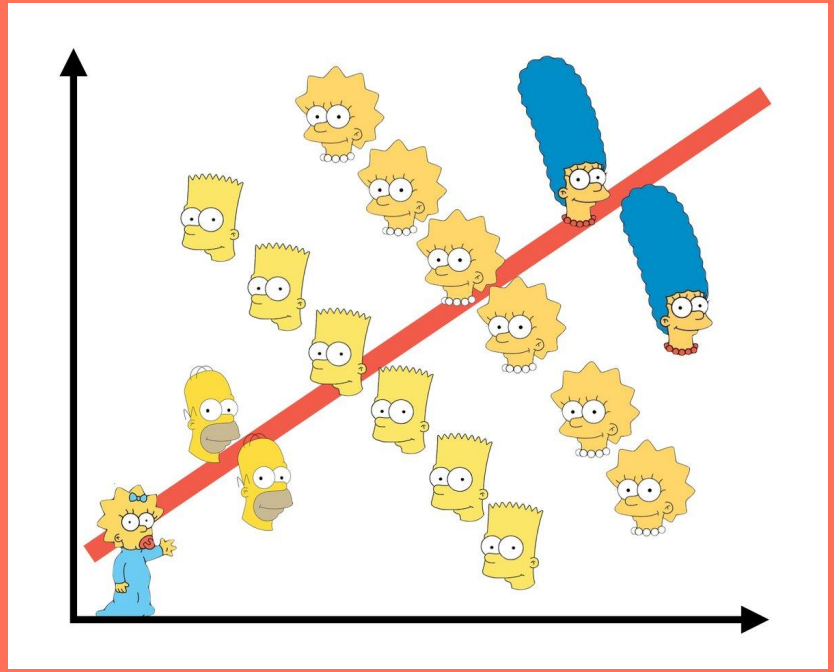
The curious case of Simpson's Paradox

Simpson's Paradox and Interpreting Data

A phenomenon in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Simpson's Paradox

- "The interpretation of interaction in contingency tables", Karl Pearson (1899)
- Early data showed case fatality rates for Covid-19 to be higher in Italy than in China overall, yet within every age group the fatality rate was higher in China than in Italy.
  -
- Parameterized reports may help disaggregate data to identify instances of Simpson's paradox



The curious case of Simpson's Paradox

References

Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2022). *rmarkdown: Dynamic Documents for R*. R package version 2.17, https://github.com/rstudio/rmarkdown.

Batra, N., Spina, A. Blomquist, P., Campbell, F., Laurenson-Schafer, H., Florence, I., Fischer, N., Ndiaye, A., Coyer, L., Polonsky, J., Izawa, Y., Bailey, C., Molling, D., Berry, I., Buajitti, E., Mousset, M., Hollis, S., Lin, W. (2021). The Epidemiologist R Handbook.

Grolemund, G., & Wickham, H. (2017). *R for Data Science*. O'Reilly Media.

Keyes, D. (2022, March 8). *Why R is Magic*. R for the Rest of Us. Retrieved October 13, 2022, from https://rfortherestofus.com/2022/03/why-r-is-magic/

Mohit, K. (2020). R Overview and History. *Medium*. https://medium.com/@ArtisOne/r-overview-and-history-75ecb036d0df

The CITI Program. 2017. "Data Management (RCR-Basic)". Accessed September 26. https://www.citiprogram.org/index.cfm?pageID=260.

Peng Roger. (2022). *R Programming for Data Science*. Leanpub. https://bookdown.org/rdpeng/rprogdatascience/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL .

Sprenger, J., & Weinberger, N. (2021, March 24). *Simpson's paradox*. Stanford Encyclopedia of Philosophy. Retrieved October 13, 2022, from https://plato.stanford.edu/entries/paradox-simpson/#:~:text=First%20published%20Wed%20Mar%2024,population%20is%20divided%20into%20subpopulations.

Xie Y, Dervieux C, Riederer E (2020). *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9780367563837, https://bookdown.org/yihui/rmarkdown-cookbook.

R and RStudio

Sisyphus punishment

R and RStudio

Atlas carrying the heavens