Joining tables & Creating and publishing scholarly articles with Quarto and GitHub pages

Research Beyond the Lab: Open Science and Research Methods for a Global Engineer

Lars Schöbitz

2024-04-11

rbtl-fs24.github.io/website/

Learning Objectives (for this week)

1. Learners can apply functions from the dplyr R Package to join multiple data sets.
2. Learners can add literature references to Quarto files using the navigation menu of RStudio visual editor and using an exported collection in .bib format from Zotero Reference Management software
3. Learners can use the GitHub pages service to publish a repository as a standalone website.

Part 1: Homework module 6

Metadata: data about data

WHAT!?

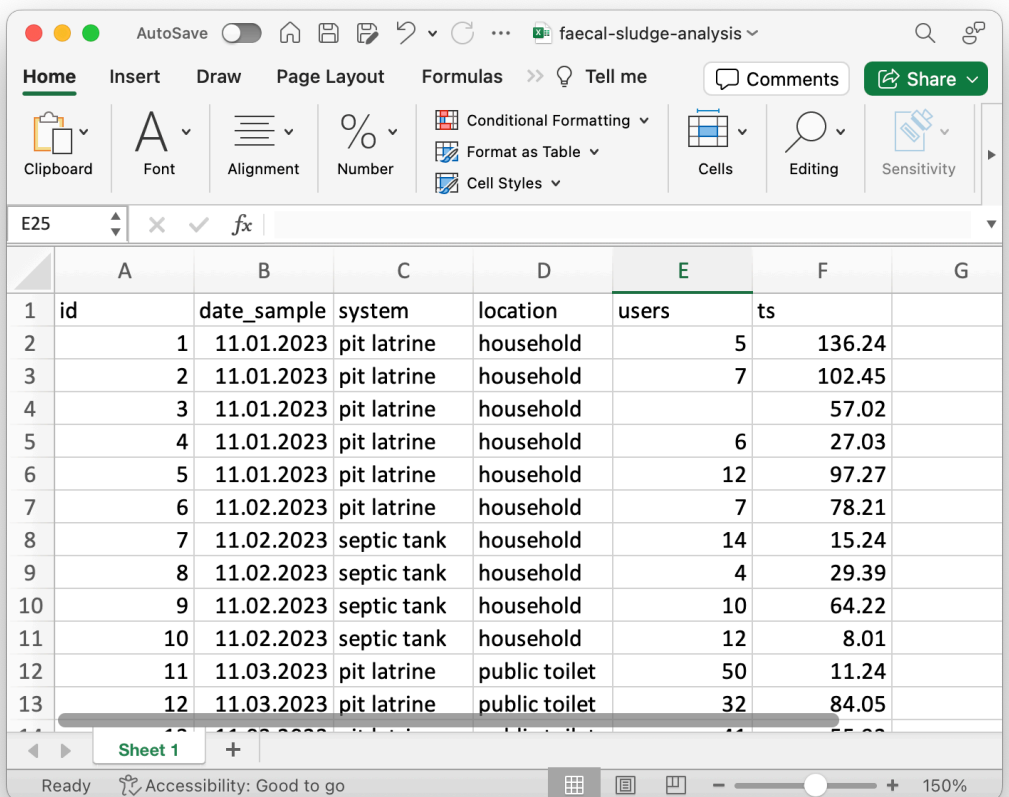
Faecal sludge samples

Imagine:

- you are new to WASH research and you have never heard of feacal sludge management.
- you are interested in learning more about the topic and you want to find some data to play with.
- you find a publication with a dataset on faecal sludge characteristics.

Faecal sludge samples

You download the XLSX file that contains the data and you open it in Excel. You see the following:

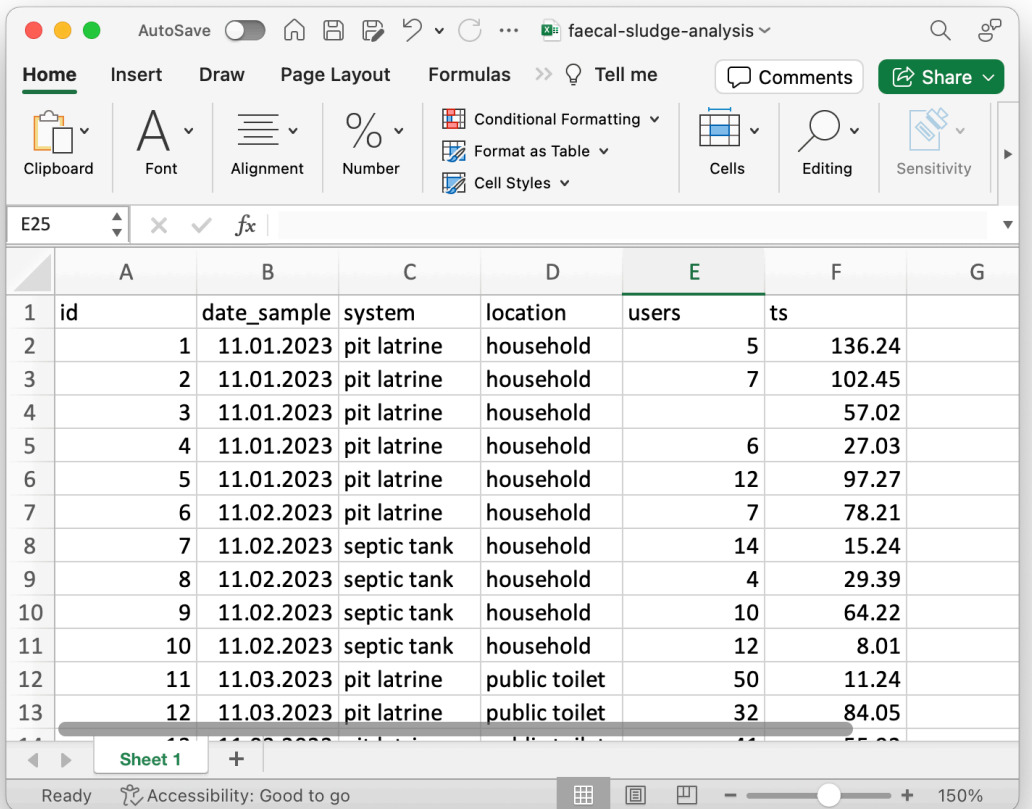


	A	B	C	D	E	F	G
1	id	date_sample	system	location	users	ts	
2	1	11.01.2023	pit latrine	household	5	136.24	
3	2	11.01.2023	pit latrine	household	7	102.45	
4	3	11.01.2023	pit latrine	household		57.02	
5	4	11.01.2023	pit latrine	household	6	27.03	
6	5	11.01.2023	pit latrine	household	12	97.27	
7	6	11.02.2023	pit latrine	household	7	78.21	
8	7	11.02.2023	septic tank	household	14	15.24	
9	8	11.02.2023	septic tank	household	4	29.39	
10	9	11.02.2023	septic tank	household	10	64.22	
11	10	11.02.2023	septic tank	household	12	8.01	
12	11	11.03.2023	pit latrine	public toilet	50	11.24	
13	12	11.03.2023	pit latrine	public toilet	32	84.05	

Faecal sludge samples

Open questions:

- What unit does `users` refer to?
- What does `ts` stand for?
- The `date` of what?
- Where was this data collected?
- Which method was used to collect the samples?



	A	B	C	D	E	F	G
1	id	date_sample	system	location	users	ts	
2	1	11.01.2023	pit latrine	household	5	136.24	
3	2	11.01.2023	pit latrine	household	7	102.45	
4	3	11.01.2023	pit latrine	household		57.02	
5	4	11.01.2023	pit latrine	household	6	27.03	
6	5	11.01.2023	pit latrine	household	12	97.27	
7	6	11.02.2023	pit latrine	household	7	78.21	
8	7	11.02.2023	septic tank	household	14	15.24	
9	8	11.02.2023	septic tank	household	4	29.39	
10	9	11.02.2023	septic tank	household	10	64.22	
11	10	11.02.2023	septic tank	household	12	8.01	
12	11	11.03.2023	pit latrine	public toilet	50	11.24	
13	12	11.03.2023	pit latrine	public toilet	32	84.05	

Questions that only the original author may have the answers to.

You as an author

have the chance to document your data properly once to make it easier for everyone else to know what it contains.

Documentation

Goes into a separate README file

- General information (authors, title, date, geographic location, etc.)
- Sharing / access information (license, links to publications, citation)
- Methodological information (sampling, analysis, etc.)

Data dictionary

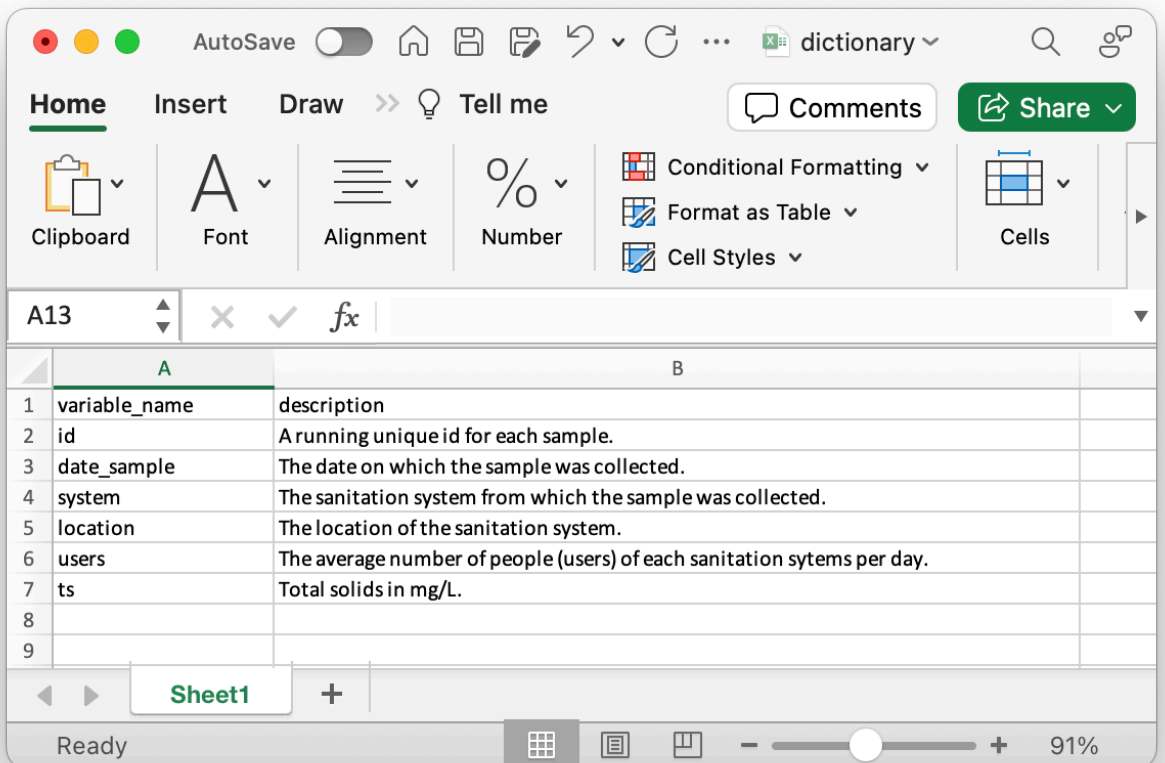
Goes into a separate file (`dictionary.csv`).

Minimum required information

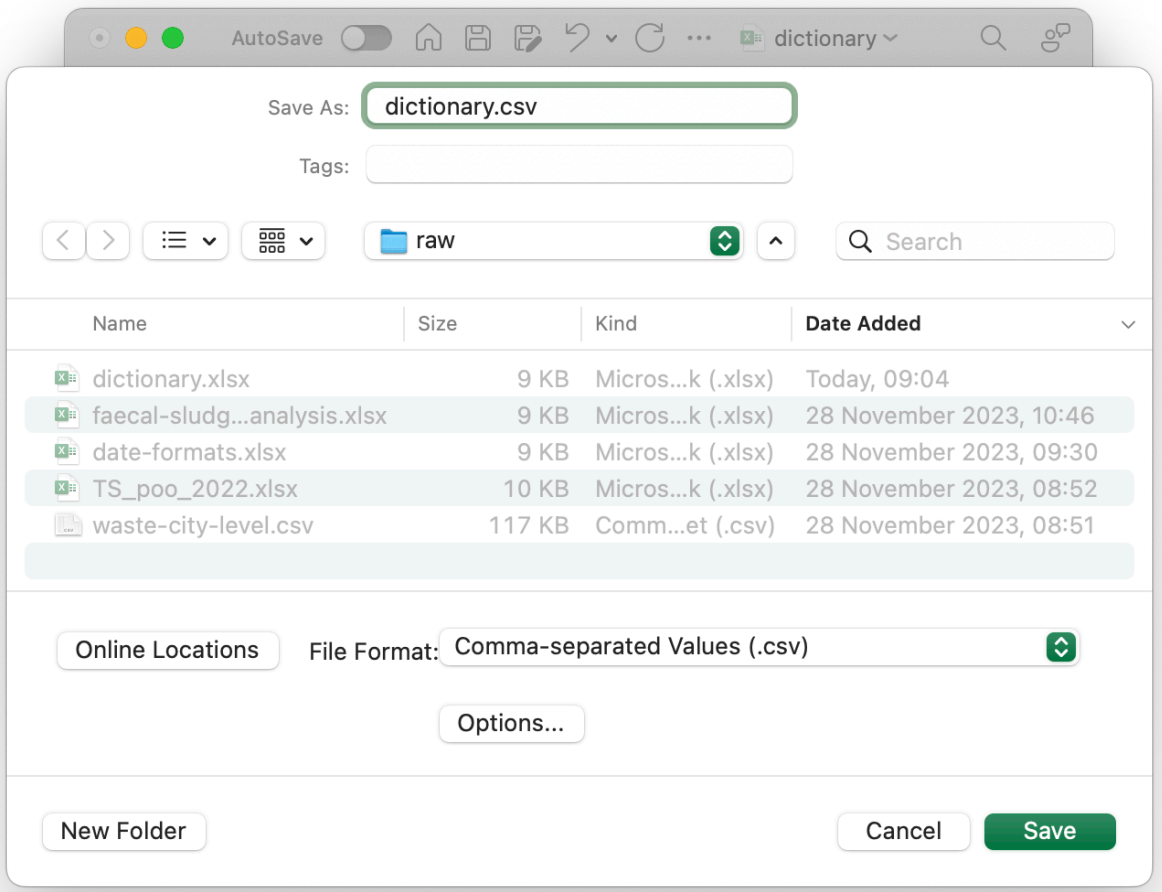
- Variable name
- Variable description

Data dictionary for faecal sludge samples

- Edit in spreadsheet software (e.g. MS Excel)



Data dictionary for faecal sludge samples



- Save as CSV file

Directory tree of a project

Capstone project of Rainbow Train: https://github.com/rbtl-fs24/project-rainbow-train

```
.
├── R
│   └── 01-data-preparation.R
├── data
│   ├── processed
│   │   ├── README.md
│   │   ├── dictionary.csv
│   │   └── faecal-sludge-analysis.csv
│   └── raw
│       └── Faecal sludge Analysis_05112023.xlsx
├── docs
│   ├── index.html
│   ├── index.qmd
│   └── index_files
│       └── libs
└── project.Rproj
```

Directory tree of a project

- `R` folder: R scripts for data cleaning
- `data` folder: raw and processed data
- `docs` folder: the actual report that imports the processed data
- `project.Rproj`: RStudio project file

Inside the `data` folder

- `raw`: data as it was downloaded / as you received it (e.g. Excel file)
- `processed`: data that is ready to be used in the report

Inside the `processed` folder

- `README.md`: general information about the data
- `dictionary.csv`: data dictionary
- `faecal-sludge-analysis.csv`: cleaned data for which `dictionary.csv` applies

Part 2: Joining data

We...

...have multiple data frames

...want to bring them together

```
professions <- read_csv(here::here("data/scientists/professions.csv"))
dates <- read_csv(here::here("data/scientists/dates.csv"))
works <- read_csv(here::here("scientists/works.csv"))
```

Data: Women in science

Information on 10 women in science who changed the world

name
Ada Lovelace
Marie Curie
Janaki Ammal
Chien-Shiung Wu
Katherine Johnson
Rosalind Franklin
Vera Rubin
Gladys West
Flossie Wong-Staal
Jennifer Doudna

Inputs

professions | dates | works

name	profession
Ada Lovelace	Mathematician
Marie Curie	Physicist and Chemist
Janaki Ammal	Botanist
Chien-Shiung Wu	Physicist
Katherine Johnson	Mathematician
Rosalind Franklin	Chemist
Vera Rubin	Astronomer
Gladys West	Mathematician
Flossie Wong-Staal	Virologist and Molecular Biologist
Jennifer Doudna	Biochemist

Desired output

name	profession	birth_year	death_year	known_for
Ada Lovelace	Mathematician	NA	NA	first computer algorithm
Marie Curie	Physicist and Chemist	NA	NA	theory of radioactivity, discovery of elements polonium and radium, first woman to win a Nobel Prize
Janaki Ammal	Botanist	1897	1984	hybrid species, biodiversity protection
Chien-Shiung Wu	Physicist	1912	1997	confim and refine theory of radioactive beta decy, Wu experiment overturning theory of parity
Katherine Johnson	Mathematician	1918	2020	calculations of orbital mechanics critical to sending the first Americans into space
Rosalind Franklin	Chemist	1920	1958	NA
Vera Rubin	Astronomer	1928	2016	existence of dark matter
Gladys West	Mathematician	1930	NA	mathematical modeling of the shape of the Earth which served as the foundation of GPS technology
Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA	first scientist to clone HIV and create a map of its genes which led to a test for the virus
Jennifer Doudna	Biochemist	1964	NA	one of the primary developers of CRISPR, a ground-breaking technology for editing genomes

Inputs, reminder

```
1 names(professions)
```

```
[1] "name"       "profession"
```

```
1 names(dates)
```

```
[1] "name"       "birth_year" "death_year"
```

```
1 names(works)
```

```
[1] "name"      "known_for"
```

```
1 nrow(professions)
```

```
[1] 10
```

```
1 nrow(dates)
```

```
[1] 8
```

```
1 nrow(works)
```

```
[1] 9
```

Joining data frames

```
1 something_join(x, y)
```

- `left_join()`: all rows from x
- `right_join()`: all rows from y
- `full_join()`: all rows from both x and y
- ...

Setup

For the next few slides...

```r
x <- tibble(
  id = c(1, 2, 3),
  value_x = c("x1", "x2", "x3")
  )
```

```r
x
```

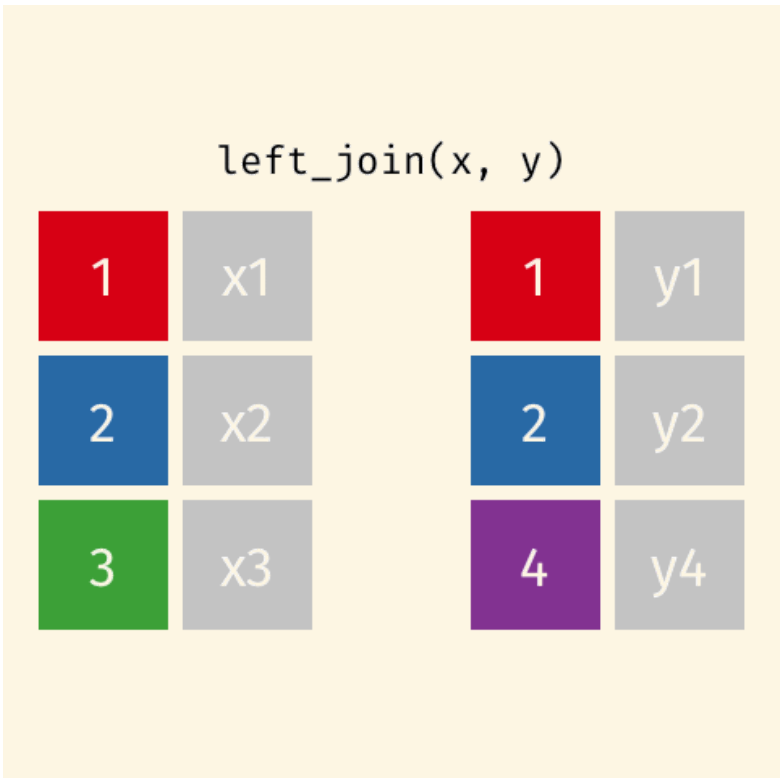
```
# A tibble: 3 × 2
     id value_x
  <dbl> <chr>
1     1 x1
2     2 x2
3     3 x3
```

```r
y <- tibble(
  id = c(1, 2, 4),
  value_y = c("y1", "y2", "y4")
  )
```

```r
y
```

```
# A tibble: 3 × 2
     id value_y
  <dbl> <chr>
1     1 y1
2     2 y2
3     4 y4
```

left_join()




```
1  left_join(x, y)
```

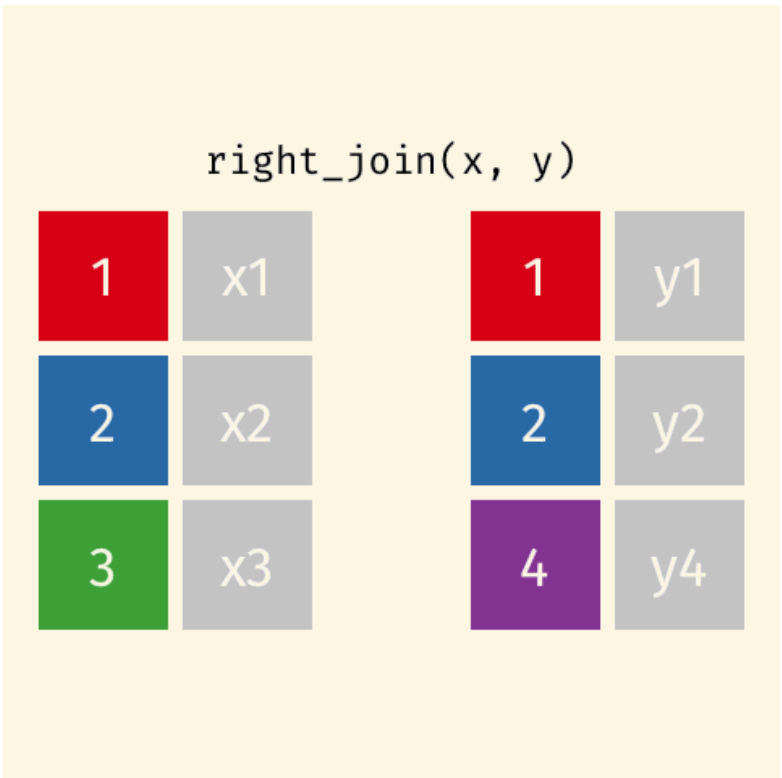
```
# A tibble: 3 × 3
     id value_x value_y
  <dbl> <chr>   <chr>
1     1 x1      y1
2     2 x2      y2
3     3 x3      <NA>
```

left_join()

```
professions %>%
  left_join(dates)
```

name	profession	birth_year	death_year
Ada Lovelace	Mathematician	NA	NA
Marie Curie	Physicist and Chemist	NA	NA
Janaki Ammal	Botanist	1897	1984
Chien-Shiung Wu	Physicist	1912	1997
Katherine Johnson	Mathematician	1918	2020
Rosalind Franklin	Chemist	1920	1958
Vera Rubin	Astronomer	1928	2016
Gladys West	Mathematician	1930	NA
Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA
Jennifer Doudna	Biochemist	1964	NA

right_join()




```
1  right_join(x, y)
```

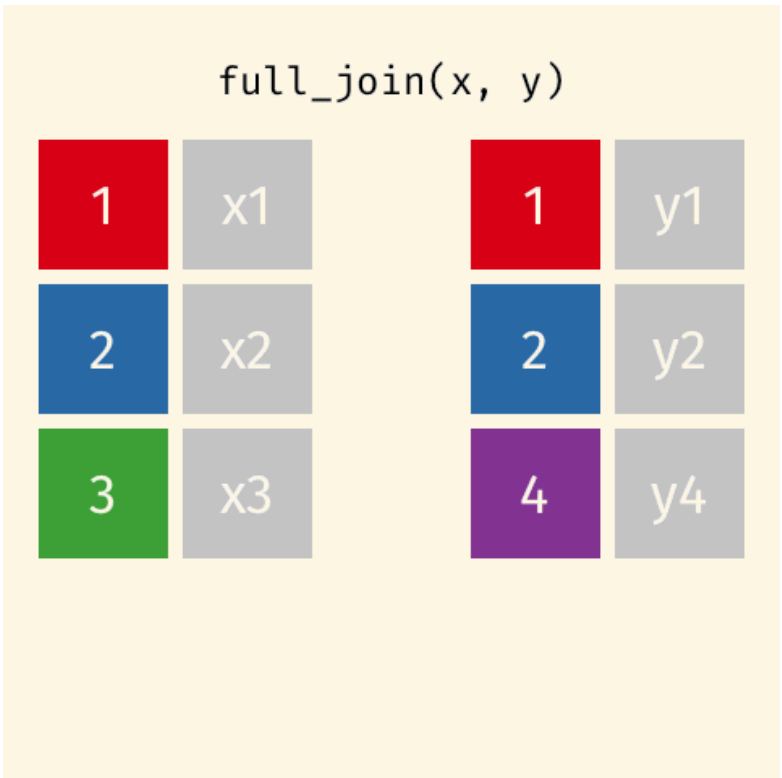
```
# A tibble: 3 × 3
     id value_x value_y
  <dbl> <chr>   <chr>
1     1 x1      y1
2     2 x2      y2
3     4 <NA>    y4
```

right_join()

```
professions %>%
  right_join(dates)
```

name	profession	birth_year	death_year
Janaki Ammal	Botanist	1897	1984
Chien-Shiung Wu	Physicist	1912	1997
Katherine Johnson	Mathematician	1918	2020
Rosalind Franklin	Chemist	1920	1958
Vera Rubin	Astronomer	1928	2016
Gladys West	Mathematician	1930	NA
Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA
Jennifer Doudna	Biochemist	1964	NA

full_join()




```
1  full_join(x, y)
```

```
# A tibble: 4 × 3
     id value_x value_y
  <dbl> <chr>   <chr>
1     1 x1      y1
2     2 x2      y2
3     3 x3      <NA>
4     4 <NA>    y4
```

full_join()

```
dates %>%
  full_join(works)
```

name	birth_year	death_year	known_for
Janaki Ammal	1897	1984	hybrid species, biodiversity protection
Chien-Shiung Wu	1912	1997	confim and refine theory of radioactive beta decy, Wu experiment overturning theory of parity
Katherine Johnson	1918	2020	calculations of orbital mechanics critical to sending the first Americans into space
Rosalind Franklin	1920	1958	NA
Vera Rubin	1928	2016	existence of dark matter
Gladys West	1930	NA	mathematical modeling of the shape of the Earth which served as the foundation of GPS technology
Flossie Wong-Staal	1947	NA	first scientist to clone HIV and create a map of its genes which led to a test for the virus
Jennifer Doudna	1964	NA	one of the primary developers of CRISPR, a ground-breaking technology for editing genomes
Ada Lovelace	NA	NA	first computer algorithm
Marie Curie	NA	NA	theory of radioactivity, discovery of elements polonium and radium, first woman to win a Nobel Prize

Putting it altogether

```
professions %>%
  left_join(dates) %>%
  left_join(works)
```

name	profession	birth_year	death_year	known_for
Ada Lovelace	Mathematician	NA	NA	first computer algorithm
Marie Curie	Physicist and Chemist	NA	NA	theory of radioactivity, discovery of elements polonium and radium, first woman to win a Nobel Prize
Janaki Ammal	Botanist	1897	1984	hybrid species, biodiversity protection
Chien-Shiung Wu	Physicist	1912	1997	confim and refine theory of radioactive beta decy, Wu experiment overturning theory of parity
Katherine Johnson	Mathematician	1918	2020	calculations of orbital mechanics critical to sending the first Americans into space
Rosalind Franklin	Chemist	1920	1958	NA
Vera Rubin	Astronomer	1928	2016	existence of dark matter
Gladys West	Mathematician	1930	NA	mathematical modeling of the shape of the Earth which served as the foundation of GPS technology
Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA	first scientist to clone HIV and create a map of its genes which led to a test for the virus
Jennifer Doudna	Biochemist	1964	NA	one of the primary developers of CRISPR, a ground-breaking technology for editing genomes

Take a break

Please get up and move! Let your emails rest in peace.

rbtl-fs24.github.io/website/

10:00

Part 2: Reference Management

Four terms

- Citation
- Reference
- Bibliography
- Citation Style Language (CSL)

What's a Citation?

- Inequality underpins waste management systems, structuring who can or cannot access services (Kalina et al., 2023).
- Many visitors still expect a personal pick-up, despite the availability of taxi services (Tilley & Kalina, 2021).
- In Tilley & Kalina (2021), the authors describe how visitors still expect a personal pick-up, despite the availability of taxi services.

What's a Citation?

- Inequality underpins waste management systems, structuring who can or cannot access services (Kalina et al., 2023).
- Many visitors still expect a personal pick-up, despite the availability of taxi services (Tilley & Kalina, 2021).
- In Tilley & Kalina (2021), the authors describe how visitors still expect a personal pick-up, despite the availability of taxi services.

> **Important: The period is after the citation.**

What's a Reference?

- detailed description of the source of information
- author's name, title, year of publication, publisher, DOI, etc.

Tilley, E., & Kalina, M. (2021). "My flight arrives at 5 am, can you pick me up?": The gatekeeping burden of the african academic. *Journal of African Cultural Studies, 33*(4), 538–548. https://doi.org/10.3929/ethz-b-000493677

What's a Bibliography?

- list of references in a research paper or project
- includes all sources used, whether they were directly quoted or not
- listed alphabetically by the author's last name in the reference list

References

Kalina, M., Makwetu, N., & Tilley, E. (2023). "The rich will always be able to dispose of their waste": A view from the frontlines of municipal failure in Makhanda, South Africa. *Environment, Development and Sustainability*. https://doi.org/10.1007/s10668-023-03363-1

Tilley, E., & Kalina, M. (2021). "My flight arrives at 5 am, can you pick me up?": The gatekeeping burden of the african academic. *Journal of African Cultural Studies*, *33*(4), 538–548. https://doi.org/10.3929/ethz-b-000493677

What's the Citation Style Language (CSL)?

- It's what your citation and generated bibliography look like
- APA (American Psychological Association) Style, Chicago Style, IEEE Style, Vancouver Style, etc. (over 10,000 styles in Zotero Style Repository)

What's the Citation Style Language (CSL)?

author-date: Many visitors still expect a personal pick-up, despite the availability of taxi services (Tilley & Kalina, 2021).

numeric Many visitors still expect a personal pick-up, despite the availability of taxi services [1].

Why use a reference management tool?

Managing references manually:

- is a lot of work
- is prone to mistakes
- makes you lose track

Why use Zotero?

- free
- open source: developed in public
- transparent about access to your own data
- cross-platform (Windows, Mac, Linux)
- collaboration in groups
- integration with word processors

rbtl-fs24.github.io/website/

Scholarly Articles in Quarto

Quarto supports

- a standardized schema for authors and affiliations that can be expressed once int the source document,
- the use of Citation Style Language (CSL) to automate the formatting of citations and bibliographies, and
- outputting to `pdf`, `html`, and `docx` with custom formatting,

according to the styles required for various journals,

and creating the LaTeX required for submission to multiple journals.

Front matter

Quarto provides a rich set of YAML metadata keys to describe the details required in the front matter of scholarly articles.

- title
- author
- affiliation
- abstract
- keywords
- citation
- licensing
- etc.


```yaml
---
title: "Give me a title"
date: 2023-10-17
author:
  - name: First Last
    orcid: 0000-0003-2196-5015
    email: me@web.org
    affiliation:
      - name: ETH Zurich
        url: https://ethz.ch/de.html
abstract: >
  The abstract will be placed here. Breaks can be added by hitting return
  on the keyboard.
```

Our turn: md-07-exercises

1. Open posit.cloud in your browser (use your bookmark).
2. Open the `rbtl-fs24 workspace` for the course.
3. Open `md-07-exercises`.
4. In the File Manager in the bottom right window, locate the `scholarly-writing.qmd` file and click on it to open it in the top left window.
5. Follow along on the screen using the instructions in the document.

Publishing

rbtl-fs24.github.io/website/

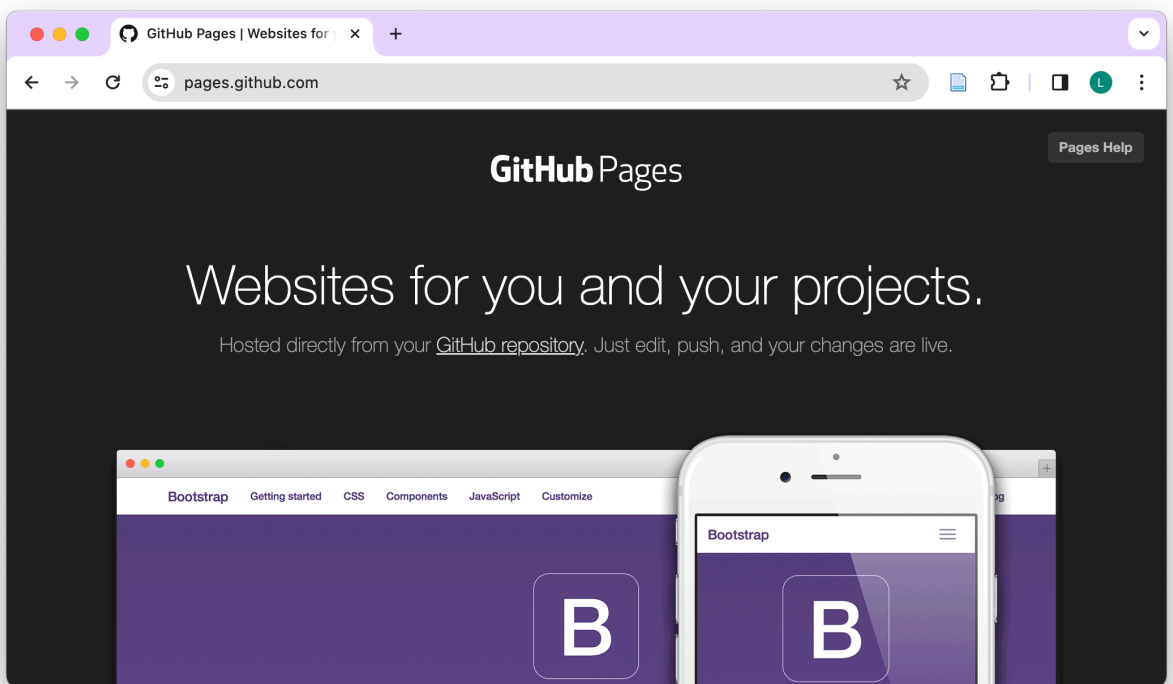
Our turn: md-07-publish-USERNAME

Clone the repository from GitHub & publish using GitHub Pages.

1. Open github.com in your browser and navigate to the GitHub organisation for the course: https://github.com/rbtl-fs24/.
2. Find the repository md-07-publish-USERNAME that ends with your GitHub username, and open it.
3. Click on the green “Code” button.
4. Copy the HTTPS URL to your clipboard.
5. Open the rbtl-fs24 workspace on posit.cloud
6. Click New Project > New Project from Git Repository
7. Paste the HTTPS URL from GitHub into the “URL of your Git Repository” field.
8. Wait until the project is deployed.
9. From the Files Manager in the bottom right window, open `docs` folder, then click on `index.qmd`.
10. Indicate when you are ready.

GitHub Pages

- GitHub Pages is a free service for hosting static websites. It is ideal for blogs, course or project websites, books, presentations, and personal hobby sites.



rbtl-fs24.github.io/website/

Minimal Example - Requirements

- Landing site needs to be stored as `index.qmd`
- The `index.qmd` needs to be stored in `docs` folder
- Example works well for a report/article as a stand-alone page
- Quarto provides a framework and examples for more complex websites: https://quarto.org/docs/websites/

Course Guide

- Steps for publishing the capstone project report are described on the capstone project page
- https://rbtl-fs24.github.io/website/project

Take a break

Please get up and move! Let your emails rest in peace.

rbtl-fs24.github.io/website/

10:00

Capstone project

rbtl-fs24.github.io/website/

Submission

- The submission due date is: Tuesday, 06th June.
- A complete report receives 30 points.
- We will use the GitHub issue tracker to communicate and ask questions about the capstone project.
- A list of required items for submission is covered on the course website: https://rbtl-fs24.github.io/website/project/

Your turn: Capstone project - Read and take notes

1. Open: https://rbtl-fs24.github.io/website/project/.
2. Read through the page.
3. For the list in “Required items” note down the numbers of those that are unclear to you and why.
4. After the time is up, we will discuss unclear items in class.

10:00

Our turn: Capstone project - Discuss unclear items

1. Share which unclear items you noted down.

10:00

Your turn: Capstone project - Share remaining unclear items on GitHub

1. Open your Capstone project repository (`project-USERNAME`) https://github.com/rbtl-fs24.
2. Add your questions for unclear items to the issue tracker.

05:00

Module 6 documentation

rbtl-fs24.github.io/website/modules/md-07.html

Module 7

Joining tables & Creating and publishing scholarly articles with Quarto and GitHub pages

Learning Objectives

1. Learners can apply functions from the dplyr R Package to join multiple data sets.
2. Learners can add literature references to Quarto files using the navigation menu of RStudio visual editor and using an exported collection in .bib format from Zotero Reference Management software
3. Learners can use the GitHub pages service to publish a repository as a standalone website.

Slides

View slides in full screen | Download slides as PDF

Homework due date

- Homework assignment due: Wednesday, April 17th

Next few weeks

- Lars will be off work until 30th May
- Elizabeth and Colin will support you during this time

module	date	topic
1	22 February 2024	Welcome & get ready for the course
2	29 February 2024	Data science lifecycle & Exploratory data analysis using visualization
3	07 March 2024	Data transformation with dplyr
4	14 March 2024	Data import & Data organization in spreadsheets
5	21 March 2024	Conditions & Dates & Tables
6	28 March 2024	Data types & Vectors & Pivoting
NA	04 April 2024	Easter Break
7	11 April 2024	Joining tables & Creating and publishing scholarly articles with Quarto and GitHub pages
8	18 April 2024	Waste Research
9	25 April 2024	Research Design
10	02 May 2024	Survey Design
NA	09 May 2024	Auffahrt Break
11	16 May 2024	Pre-test and logistics
NA	23 May 2024	Data collection
12	30 May 2024	Data analysis & report writing
NA	06 June 2024	Project Submission Deadline
NA	13 June 2024	Exam

Attribution

Content was re-used from a workshop hosted by Mine Çetinkaya-Rundel at the 2023 Symposium on Data Science and Statistics and stored at https://github.com/mine-cetinkaya-rundel/quarto-sdss. The original content is licensed under a Creative Commons Attribution 4.0 International License.

Thanks! 🌻

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/ Access slides as PDF on GitHub