Pivoting & Joining & Metadata

ds4owd - data science for openwashdata

Lars Schöbitz

2023-12-12

Learning Objectives (for this week)

1. Learners can apply functions from the tidyr R Package to transform their data from a wide to a long format and vice versa.
2. Learners can apply functions from the dplyr R Package to join multiple data sets.
3. Learners can understand the relevance a data dictionary to describe variable names in a data set.

Homework module 6

Your turn: Discuss the reading

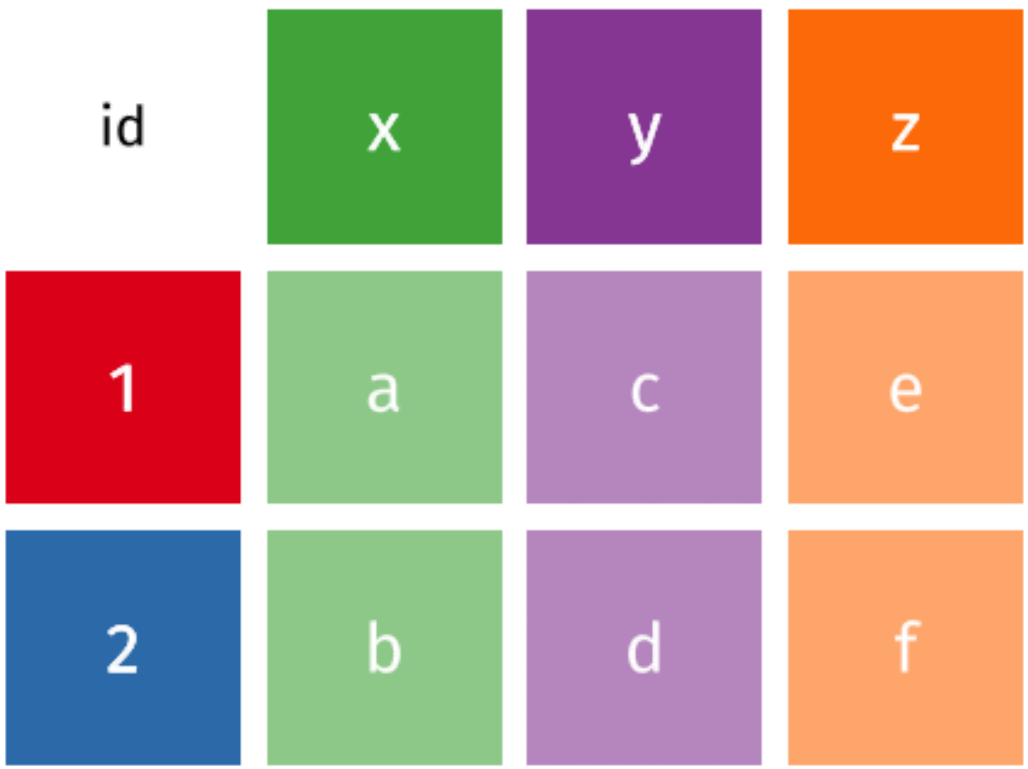
> In discussion groups of 3, share your examples and discuss how the recommendations would improve your workflows.

1. In your homework you have read Wilson et al. (2017): “Good enough practices in scientific computing”.
2. You chose one of the recommended practices (Data Management, Software, Collaboration, Project organization, Keeping track of changes, Manuscripts), then:
3. You came up with a real-world example or scenario where the recommended practice could be applied in your research or academic work.
4. You are prepared to explain how the recommended practices would improve your workflows. This will be in the class setting as part of small discussion group (max 3 people).

Part 1: Pivoting data

Pivoting data

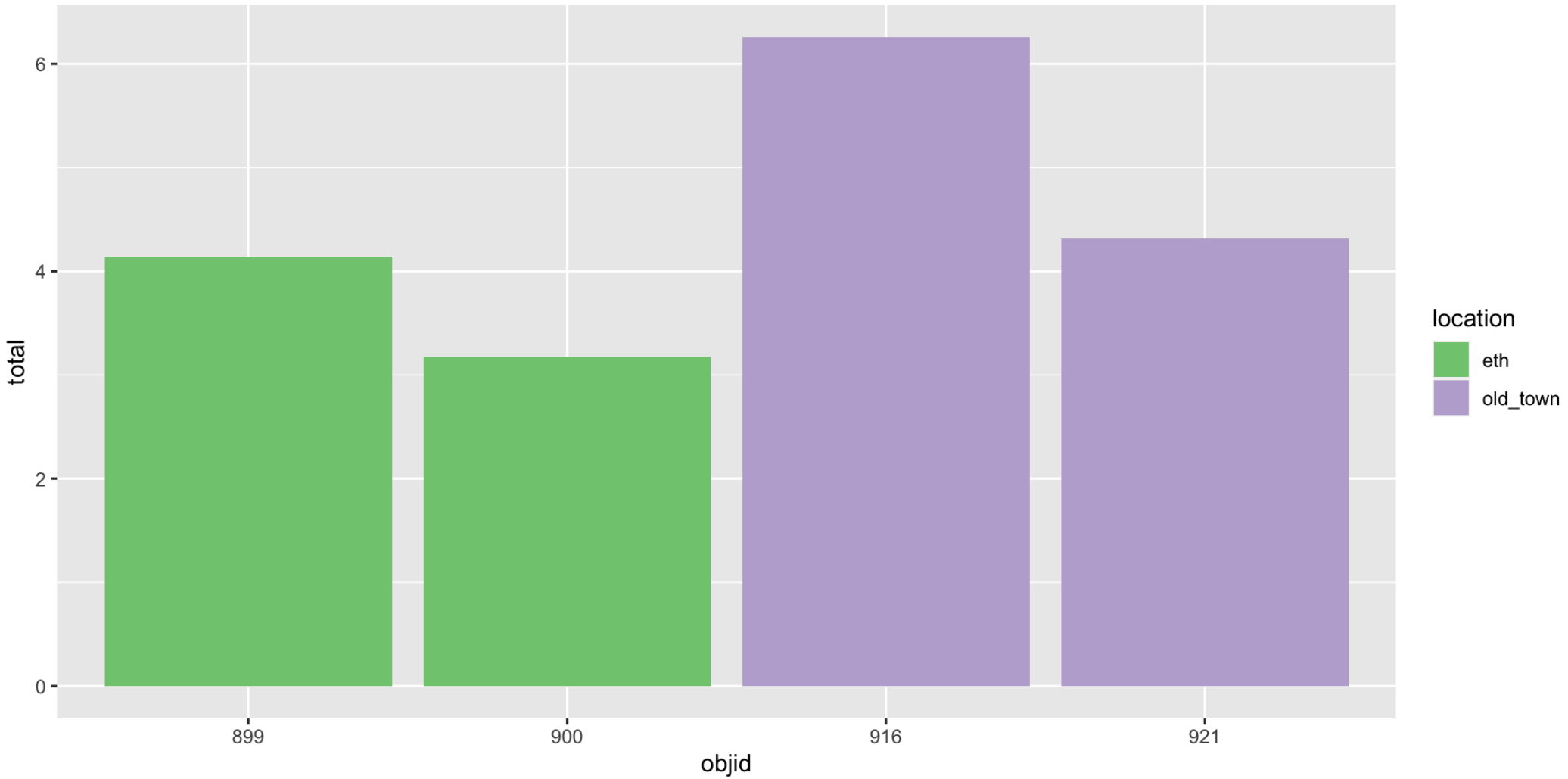
wide



Waste characterisation data

objid	location	pet	metal_alu	glass	paper	other	total
900	eth	0.06	0.06	0.58	0.21	1.14	2.05
899	eth	0.14	0.01	0.18	0.28	3.04	3.64
921	old_town	0.00	0.00	0.00	0.41	1.57	1.99
916	old_town	0.17	0.04	0.80	0.55	0.62	2.19
900	eth	0.10	0.04	0.00	0.40	0.58	1.12
899	eth	0.08	0.03	0.00	0.05	0.34	0.50
921	old_town	0.08	0.03	0.30	0.40	1.52	2.33
916	old_town	0.11	0.04	0.92	1.01	1.99	4.07

How would you plot this?

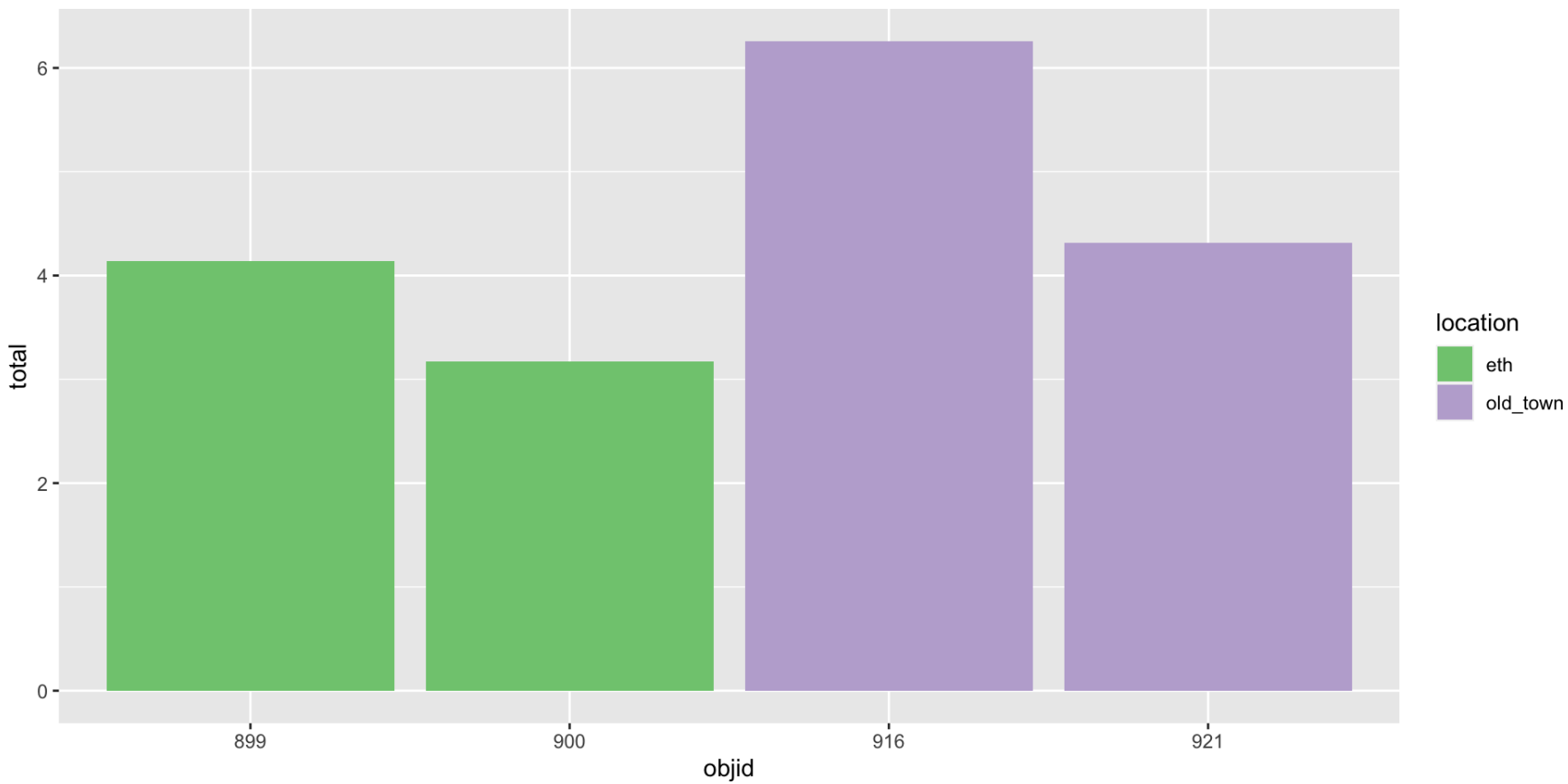


Three variables

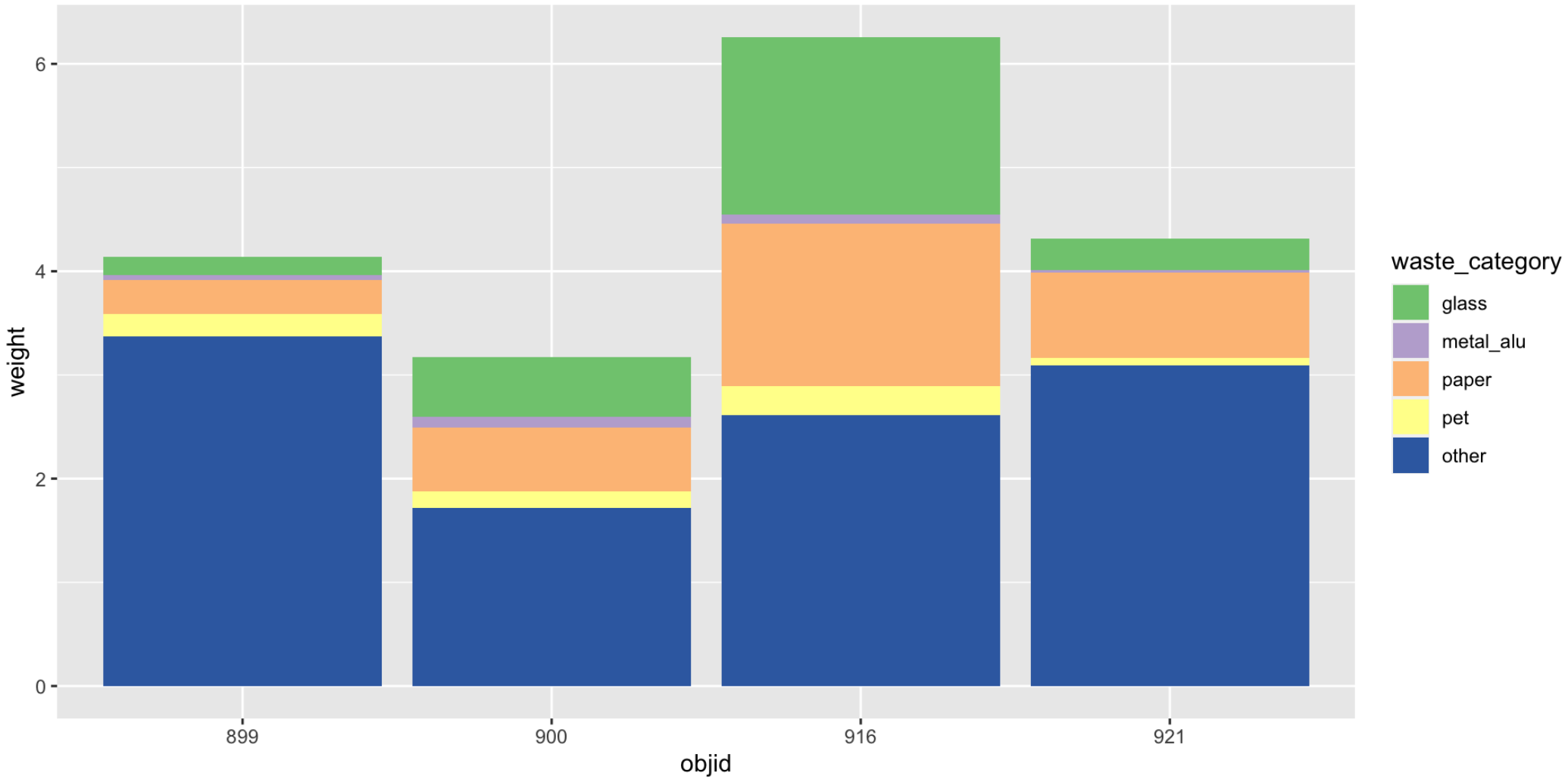
objid	location	total
900	eth	2.05
899	eth	3.64
921	old_town	1.99
916	old_town	2.19
900	eth	1.12
899	eth	0.50
921	old_town	2.33
916	old_town	4.07

Three variables -> three aesthetics

```
ggplot(data = waste_data_untidy,
       mapping = aes(x = objid,
                     y = total,
                     fill = location)) +
  geom_col() +
  scale_fill_brewer(type = "qual")
```



And how to plot this?



ds4owd-001.github.io/website/

Reminder: Data (in wide format)

objid	location	pet	metal_alu	glass	paper	other
900	eth	0.06	0.06	0.58	0.21	1.14
899	eth	0.14	0.01	0.18	0.28	3.04
921	old_town	0.00	0.00	0.00	0.41	1.57
916	old_town	0.17	0.04	0.80	0.55	0.62
900	eth	0.10	0.04	0.00	0.40	0.58
899	eth	0.08	0.03	0.00	0.05	0.34
921	old_town	0.08	0.03	0.30	0.40	1.52
916	old_town	0.11	0.04	0.92	1.01	1.99

You need: A long format

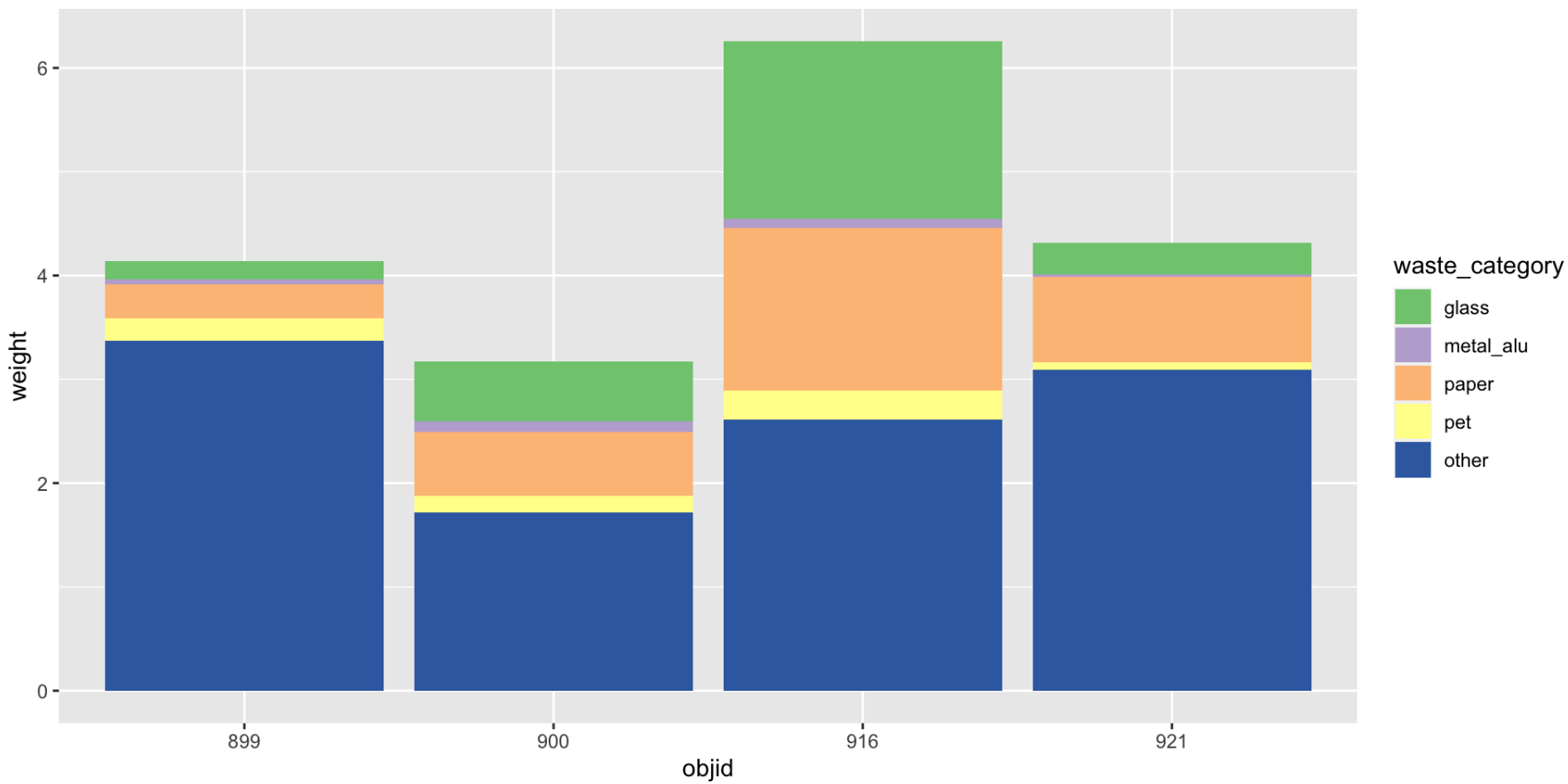
objid	location	waste_category	weight
900	eth	pet	0.06
900	eth	metal_alu	0.06
900	eth	glass	0.58
900	eth	paper	0.21
900	eth	other	1.14
899	eth	pet	0.14
899	eth	metal_alu	0.01
899	eth	glass	0.18
899	eth	paper	0.28
899	eth	other	3.04
921	old_town	pet	0.00
921	old_town	metal_alu	0.00
921	old_town	glass	0.00
921	old_town	paper	0.41

objid	location	waste_category	weight
921	old_town	other	1.57
916	old_town	pet	0.17
916	old_town	metal_alu	0.04
916	old_town	glass	0.80
916	old_town	paper	0.55
916	old_town	other	0.62
900	eth	pet	0.10
900	eth	metal_alu	0.04
900	eth	glass	0.00
900	eth	paper	0.40
900	eth	other	0.58
899	eth	pet	0.08
899	eth	metal_alu	0.03
899	eth	glass	0.00
899	eth	paper	0.05
899	eth	other	0.34

objid	location	waste_category	weight
921	old_town	pet	0.08
921	old_town	metal_alu	0.03
921	old_town	glass	0.30
921	old_town	paper	0.40
921	old_town	other	1.52
916	old_town	pet	0.11
916	old_town	metal_alu	0.04
916	old_town	glass	0.92
916	old_town	paper	1.01
916	old_town	other	1.99

Three variables -> three aesthetics

```
ggplot(data = waste_data_tidy,
       mapping = aes(x = objid,
                     y = weight,
                     fill = waste_category)) +
  geom_col() +
  scale_fill_brewer(type = "qual")
```



ds4owd-001.github.io/website/

How to

```
1 waste_data_untidy
```

objid	location	pet	metal_alu	glass	paper	other
900	eth	0.06	0.06	0.58	0.21	1.14
899	eth	0.14	0.01	0.18	0.28	3.04
921	old_town	0.00	0.00	0.00	0.41	1.57
916	old_town	0.17	0.04	0.80	0.55	0.62
900	eth	0.10	0.04	0.00	0.40	0.58
899	eth	0.08	0.03	0.00	0.05	0.34
921	old_town	0.08	0.03	0.30	0.40	1.52
916	old_town	0.11	0.04	0.92	1.01	1.99

How to

```
waste_data_untidy |>
  pivot_longer(cols = pet:other,
               names_to = "waste_category",
               values_to = "weight")
```

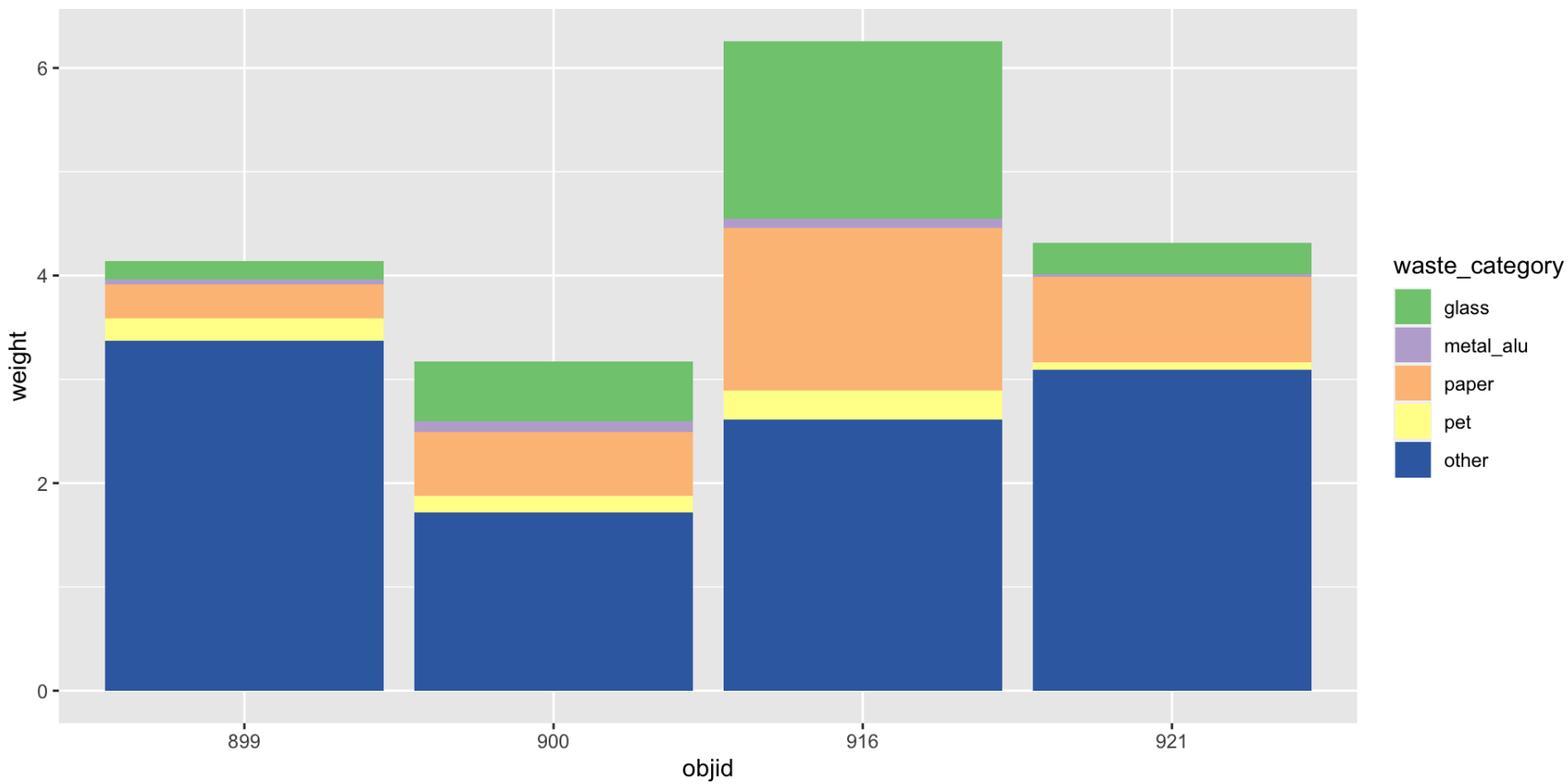
objid	location	waste_category	weight
900	eth	pet	0.06
900	eth	metal_alu	0.06
900	eth	glass	0.58
900	eth	paper	0.21
900	eth	other	1.14
899	eth	pet	0.14
899	eth	metal_alu	0.01
899	eth	glass	0.18
899	eth	paper	0.28
899	eth	other	3.04
921	old_town	pet	0.00

objid	location	waste_category	weight
921	old_town	metal_alu	0.00
921	old_town	glass	0.00
921	old_town	paper	0.41
921	old_town	other	1.57
916	old_town	pet	0.17
916	old_town	metal_alu	0.04
916	old_town	glass	0.80
916	old_town	paper	0.55
916	old_town	other	0.62
900	eth	pet	0.10
900	eth	metal_alu	0.04
900	eth	glass	0.00
900	eth	paper	0.40
900	eth	other	0.58
899	eth	pet	0.08
899	eth	metal_alu	0.03

objid	location	waste_category	weight
899	eth	glass	0.00
899	eth	paper	0.05
899	eth	other	0.34
921	old_town	pet	0.08
921	old_town	metal_alu	0.03
921	old_town	glass	0.30
921	old_town	paper	0.40
921	old_town	other	1.52
916	old_town	pet	0.11
916	old_town	metal_alu	0.04
916	old_town	glass	0.92
916	old_town	paper	1.01
916	old_town	other	1.99

Three variables -> three aesthetics

```
ggplot(data = waste_data_tidy,
       mapping = aes(x = objid,
                     y = weight,
                     fill = waste_category)) +
  geom_col() +
  scale_fill_brewer(type = "qual")
```



ds4owd-001.github.io/website/

Your turn: md-07-exercises - pivoting data

1. Open posit.cloud in your browser (use your bookmark).
2. Open the ds4owd workspace for the course.
3. In the File Manager in the bottom right window, locate the `md-07a-pirvoting.qmd` file and click on it to open it in the top left window.
4. Follow instructions in the file

30:00

Take a break

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

ds4owd-001.github.io/website/

10:00

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

ds4owd-001.github.io/website/

Source: Discover Magazine

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
```

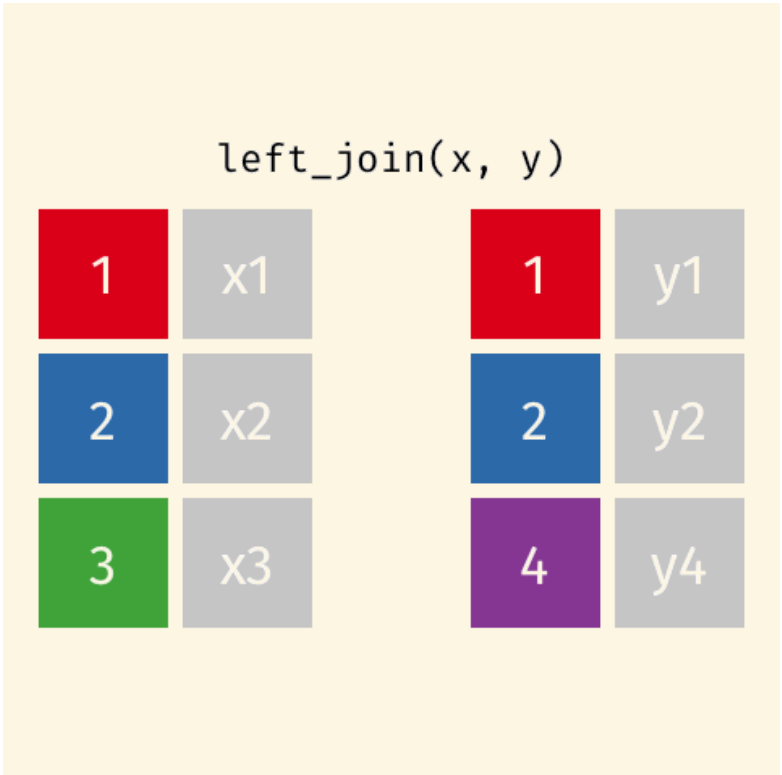
```
# A tibble: 3 × 2
     id value_x
  <dbl> <chr>
1     1 x1
2     2 x2
3     3 x3
```

```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()




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

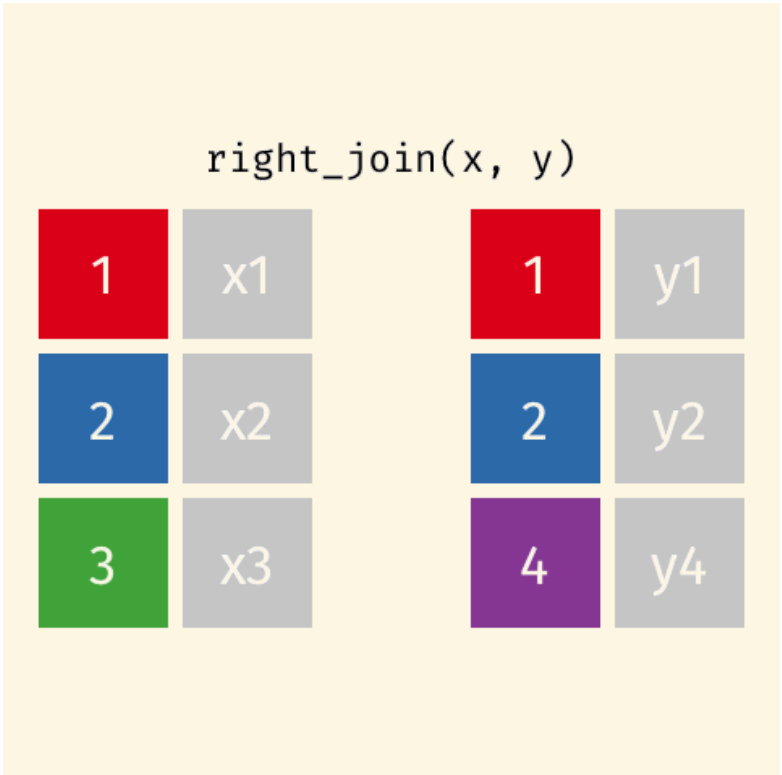
```
# A tibble: 3 × 3
     id value_x value_y
  <dbl> <chr>   <chr>
1     1 x1      y1
2     2 x2      y2
3     3 x3      <NA>
```

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()




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

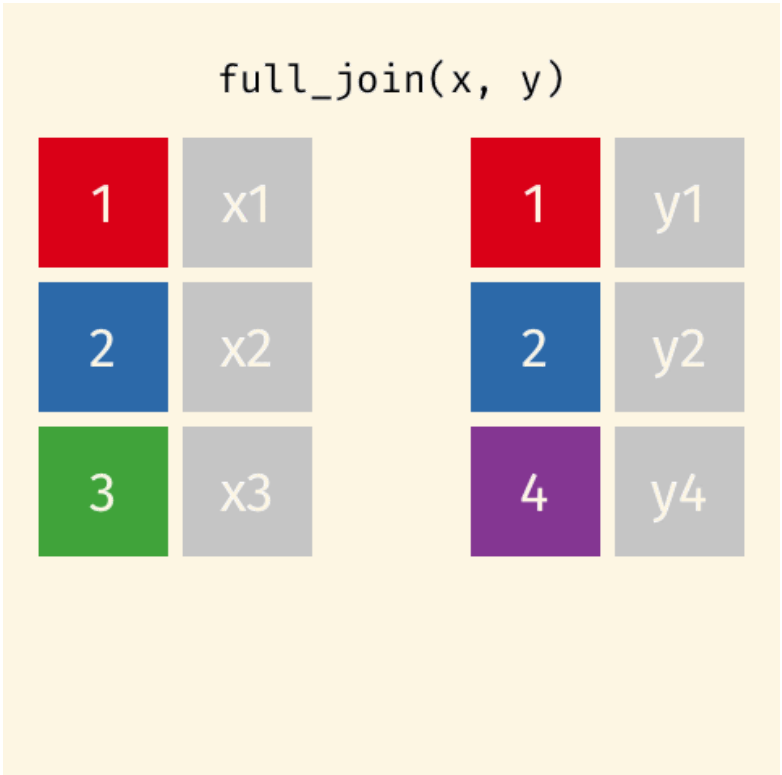
```
# A tibble: 3 × 3
     id value_x value_y
  <dbl> <chr>   <chr>
1     1 x1      y1
2     2 x2      y2
3     4 <NA>    y4
```

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

Part 3: Metadata

Metadata: data about data

WHAT!?

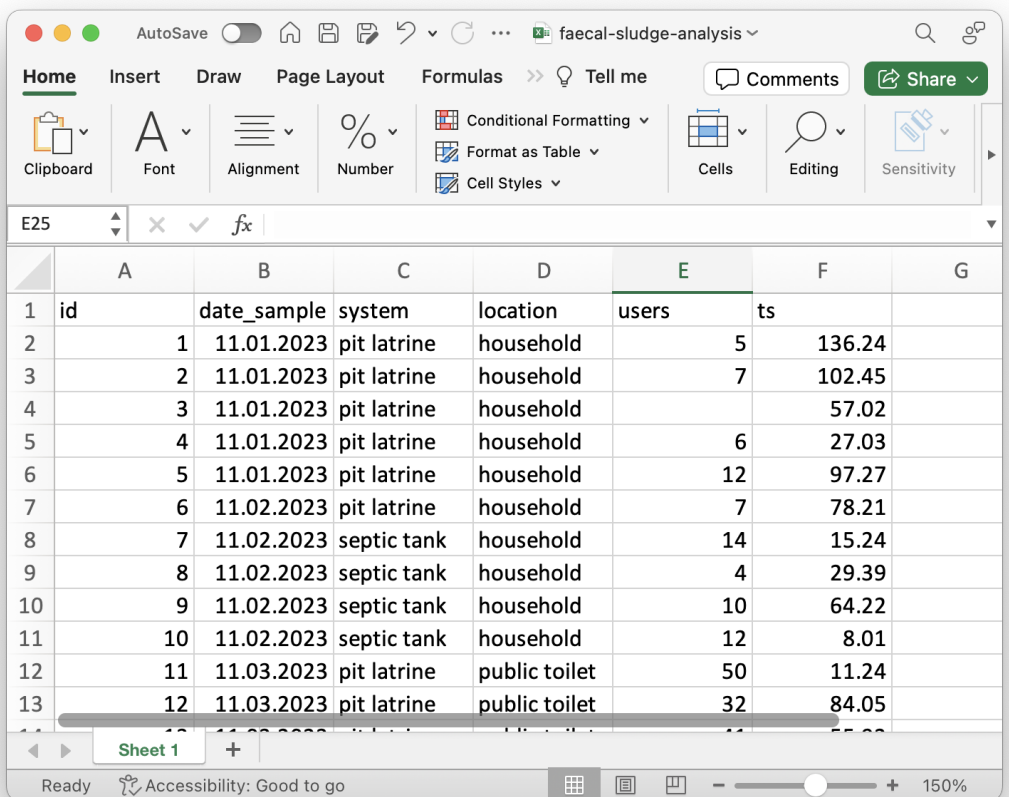
Faecal sludge samples

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:

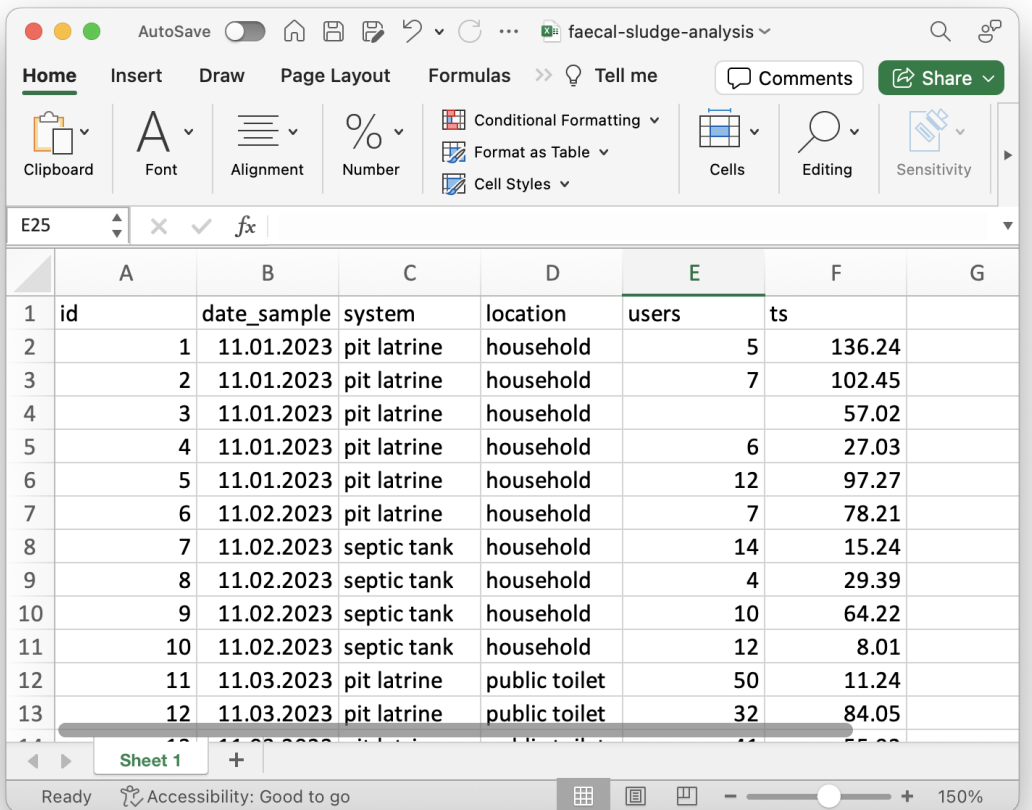


	A	B	C	D	E	F	G
1	id	date_sample	system	location	users	ts	
2	1	11.01.2023	pit latrine	household	5	136.24	
3	2	11.01.2023	pit latrine	household	7	102.45	
4	3	11.01.2023	pit latrine	household		57.02	
5	4	11.01.2023	pit latrine	household	6	27.03	
6	5	11.01.2023	pit latrine	household	12	97.27	
7	6	11.02.2023	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

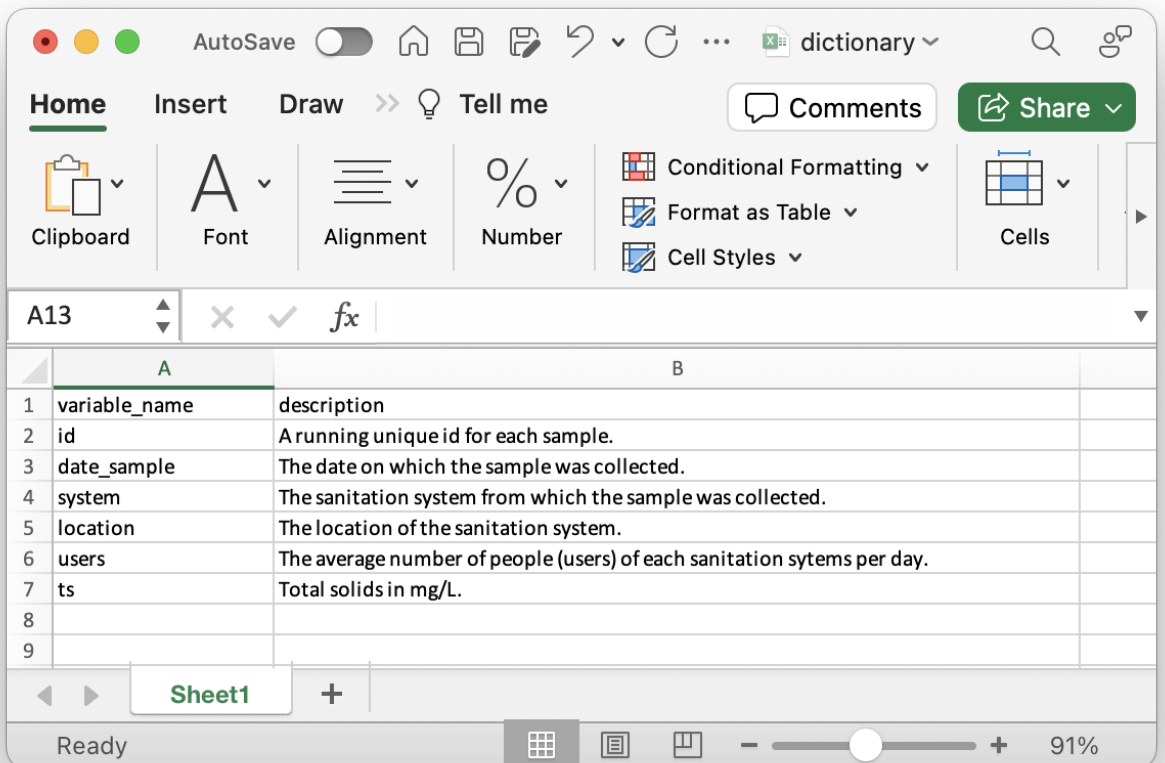
Goes into a separate file (`dictionary.csv`).

Minimum required information

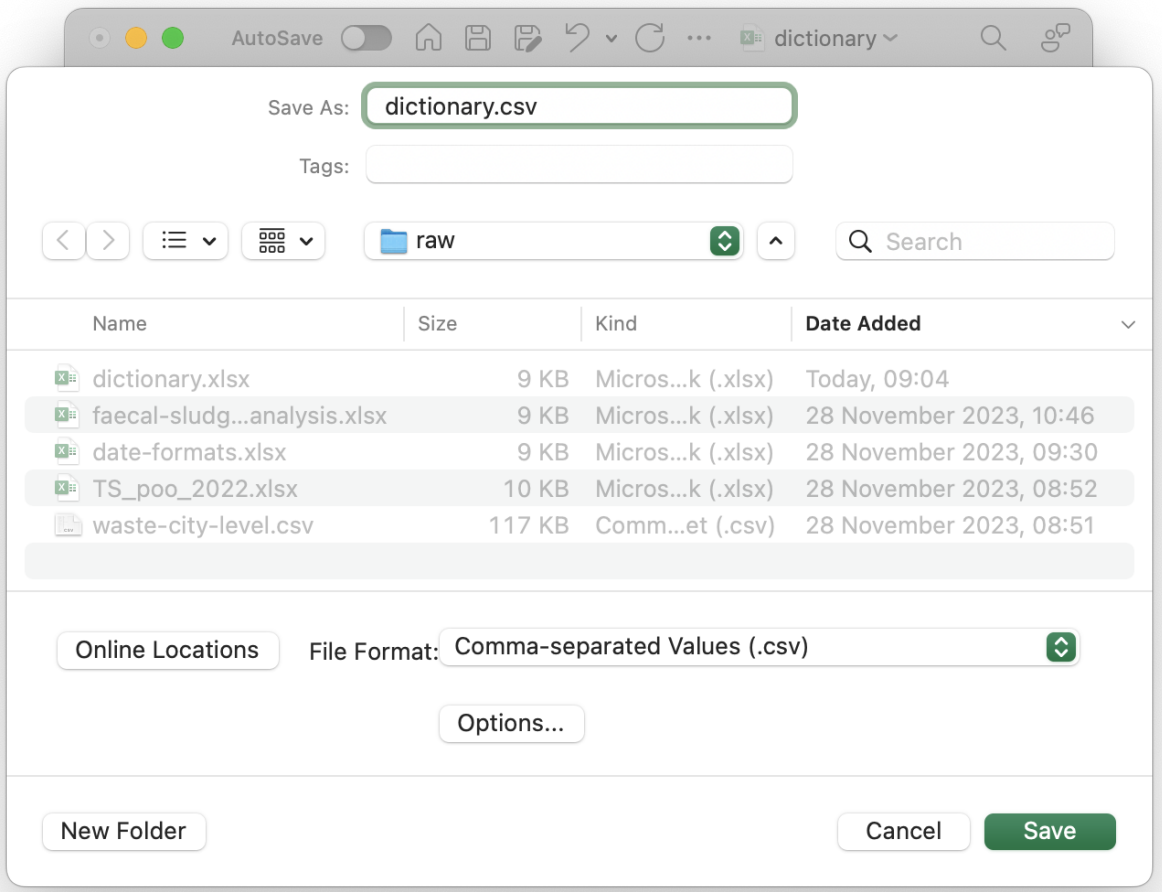
- Variable name
- Variable description

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/ds4owd-001/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

My turn: A tour of Rainbow Train's project

Sit back and enjoy!

ds4owd-001.github.io/website/

10:00

Homework assignments module 7

Module 7 documentation

ds4owd-001.github.io/website/modules/md-07.html

Module 7

Joining data & wrting functions

Learning Objectives

1. Learners can apply functions from the dplyr R Package to join multiple data sets.
2. Learners can write functions that reduce repition in dplyr and ggplot2 code sequences.

Slides

- In preparation

Readings

1. Read R for Data Science - Section 20 - Joins

Homework due date

- Homework assignment due: Monday, December 18th
- Feedback until: Thursday, December 21st

Wrap-up

ds4owd-001.github.io/website/

Final student hours of 2023

- Thursday, December 14th at 2 pm CET

First student hours of 2024

- Thursday, January 11th at 2 pm CET

Christmas break

- Lars will be on vacation from December 22nd until January 15th
- Mian and Sophia will be on vacation from December 22nd until January 8th

First lecture of 2024

Tuesday, January 16th at 2 pm CET

ds4owd-001.github.io/website/

Course calendar

date	week	topic	module
31 October 2023	1	Welcome & get ready for the course	module 1
07 November 2023	2	Data science lifecycle & Exploratory data analysis using visualization	module 2
14 November 2023	3	Data transformation with dplyr	module 3
21 November 2023	4	Data import & Data organization in spreadsheets	module 4
28 November 2023	5	Conditions & Dates & Tables	module 5
05 December 2023	6	Data types & Vectors & For Loops	module 6
12 December 2023	7	Pivoting & joining data	module 7
19 December 2023	8	Break	NA
26 December 2023	9	Break	NA
02 January 2024	10	Break	NA
09 January 2024	11	Work on Capstone project	NA
16 January 2024	12	Writing functions & Creating a website with Quarto and GitHub pages	module 8
23 January 2024	13	openwashdata webinar: a data sharing workflow that may please the publishers	NA
30 January 2024	14	Using AI for software development in R	module 9
06 February 2024	15	Work on Capstone project	NA
13 February 2024	16	Final submission date of Capstone project	NA
20 February 2024	17	Graduation party of openwashdata academy	module 10

Thanks! 🌻

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

References

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. “Good Enough Practices in Scientific Computing.” *PLOS Computational Biology* 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.