Concept of Tidy Data, Vectors & Pivoting

CVEN 5837 - Summer 2022

Lars Schöbitz

https://cven5837-ss22.github.io/website/

Mortenson Center in Global Engineering
UNIVERSITY OF COLORADO BOULDER

Learning Objectives (for this week)

1. Learners can apply functions from the dplyr R Package to transform their data from a wide to a long format and vice versa
2. Learners can list the four main atomic vector types in R
3. Learners can explain the three characteristics of tidy data

Part 1: Data types and vectors

Why care about data types?

via GIPHY

Example: survey data

id	job	price_glass
1	Student	0
2	Retired	0
3	Other	0
4	Employed	10
5	Employed	See comment
6	Student	05-Oct
7	Student	0
8	Retired	0
9	Student	10
10	Employed	0
11	Employed	20 (2chf per person with 10 people in the WG)
12	Student	10
13	Student	10

id	job	price_glass
14	Employed	0
15	Student	10
16	Student	0
17	Employed	5 to 10
18	Other	0
19	Student	0
20	Employed	10
21	Employed	0
22	Employed	5

Oh why won't you work?!

```
1 survey_data_small |>
2   summarise(mean_price_glass = mean(price_glass))
```

```
# A tibble: 1 × 1
  mean_price_glass
             <dbl>
1               NA
```

Oh why won't you still work??!!

```
1 survey_data_small |>
2   summarise(mean_price_glass = mean(price_glass, na.rm = TRUE))
```

```
# A tibble: 1 × 1
  mean_price_glass
             <dbl>
1               NA
```

Take a breath and look at your data

id	job	price_glass
1	Student	0
2	Retired	0
3	Other	0
4	Employed	10
5	Employed	See comment
6	Student	05-Oct
7	Student	0
8	Retired	0
9	Student	10
10	Employed	0
11	Employed	20 (2chf per person with 10 people in the WG)
12	Student	10
13	Student	10

id	job	price_glass
14	Employed	0
15	Student	10
16	Student	0
17	Employed	5 to 10
18	Other	0
19	Student	0
20	Employed	10
21	Employed	0
22	Employed	5

Very common data tidying step!

```
survey_data_small |>
  mutate(price_glass_new = case_when(
    price_glass == "5 to 10" ~ "7.5",
    price_glass == "05-Oct" ~ "7.5",
    str_detect(price_glass, pattern = "20") == TRUE ~ "20",
    str_detect(price_glass, pattern = "See comment") == TRUE ~ NA_charact
    TRUE ~ price_glass
  ))
```

Very common data tidying step!

id	job	price_glass_new	price_glass
1	Student	0	0
2	Retired	0	0
3	Other	0	0
4	Employed	10	10
5	Employed	NA	See comment
6	Student	7.5	05-Oct
7	Student	0	0
8	Retired	0	0
9	Student	10	10
10	Employed	0	0
11	Employed	20	20 (2chf per person with 10 people in the WG)
12	Student	10	10
13	Student	10	10

id	job	price_glass_new	price_glass
14	Employed	0	0
15	Student	10	10
16	Student	0	0
17	Employed	7.5	5 to 10
18	Other	0	0
19	Student	0	0
20	Employed	10	10
21	Employed	0	0
22	Employed	5	5

Sumamrise? Argh!!!!

```
survey_data_small |>
  mutate(price_glass_new = case_when(
    price_glass == "5 to 10" ~ "7.5",
    price_glass == "05-Oct" ~ "7.5",
    str_detect(price_glass, pattern = "20") == TRUE ~ "20",
    str_detect(price_glass, pattern = "See comment") == TRUE ~ NA_charact
    TRUE ~ price_glass
  )) |>
  summarise(mean_price_glass = mean(price_glass_new, na.rm = TRUE))
```

```
# A tibble: 1 × 1
  mean_price_glass
             <dbl>
1               NA
```

Always respect your data types!

```r
survey_data_small |>
  mutate(price_glass_new = case_when(
    price_glass == "5 to 10" ~ "7.5",
    price_glass == "05-Oct" ~ "7.5",
    str_detect(price_glass, pattern = "20") == TRUE ~ "20",
    str_detect(price_glass, pattern = "See comment") == TRUE ~ NA_charact
    TRUE ~ price_glass
  )) |>
  mutate(price_glass_new = as.numeric(price_glass_new)) |>
  summarise(mean_price_glass = mean(price_glass_new, na.rm = TRUE))
```

```
# A tibble: 1 × 1
  mean_price_glass
             <dbl>
1             4.76
```

Live Coding Exercise: Vectors

https://cven5837-ss22.github.io/website/

live-04a-vectors

1. Head over to rstudio.cloud
2. Open the workspace for the course (cven5837-ss22)
3. Open “Projects”
4. Open the “course-materials” project
5. Follow along with me

Break One

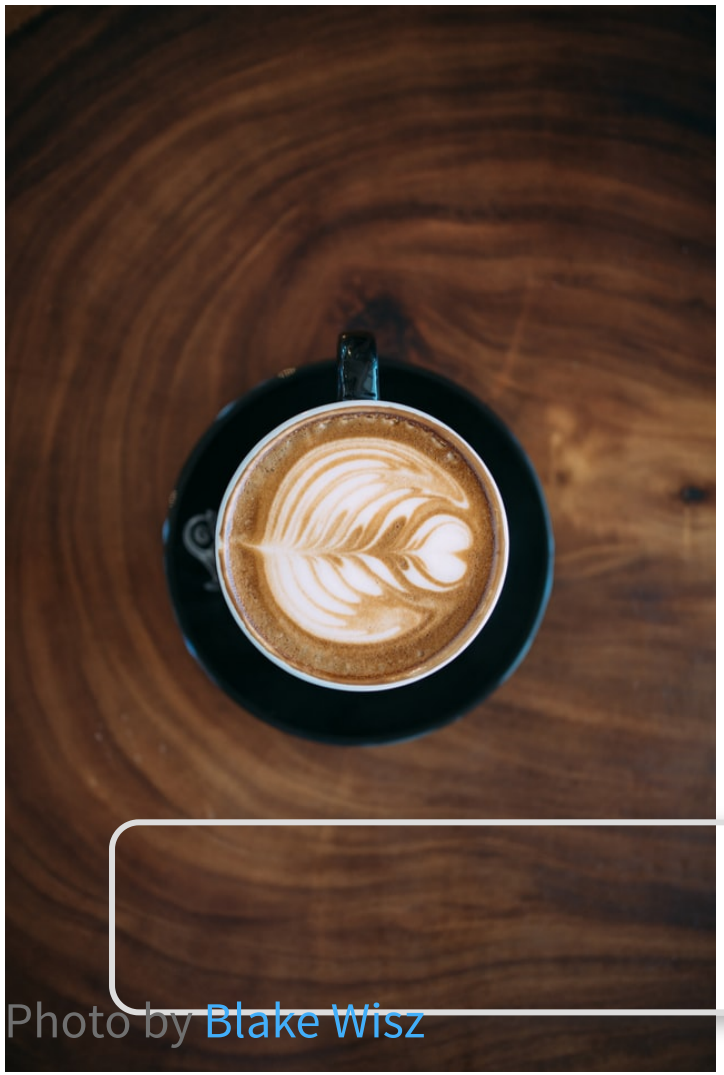


Photo by Blake Wisz

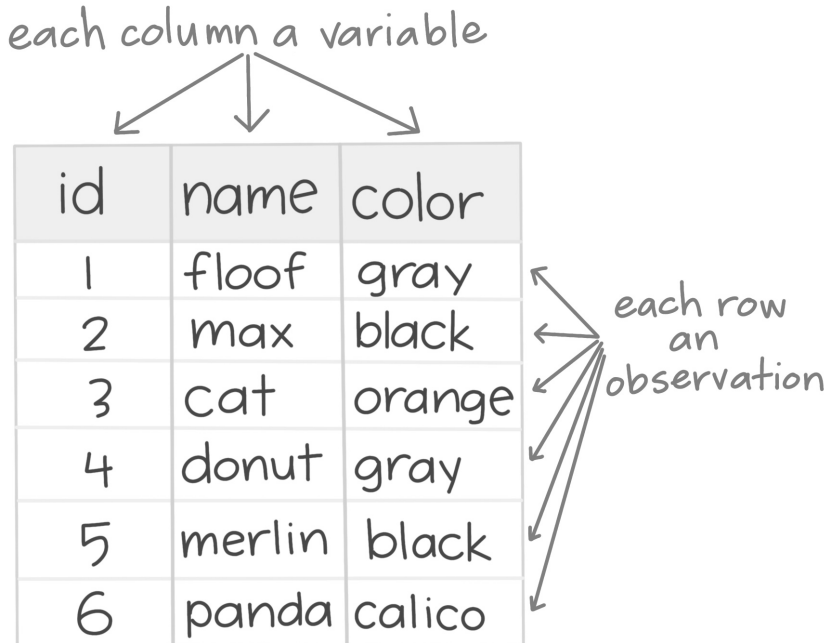
10:00

Part 2: tidyr - long and wide formats

> TIDY DATA is a standard way of mapping the meaning of a dataset to its structure.
>
> –HADLEY WICKHAM

In tidy data:

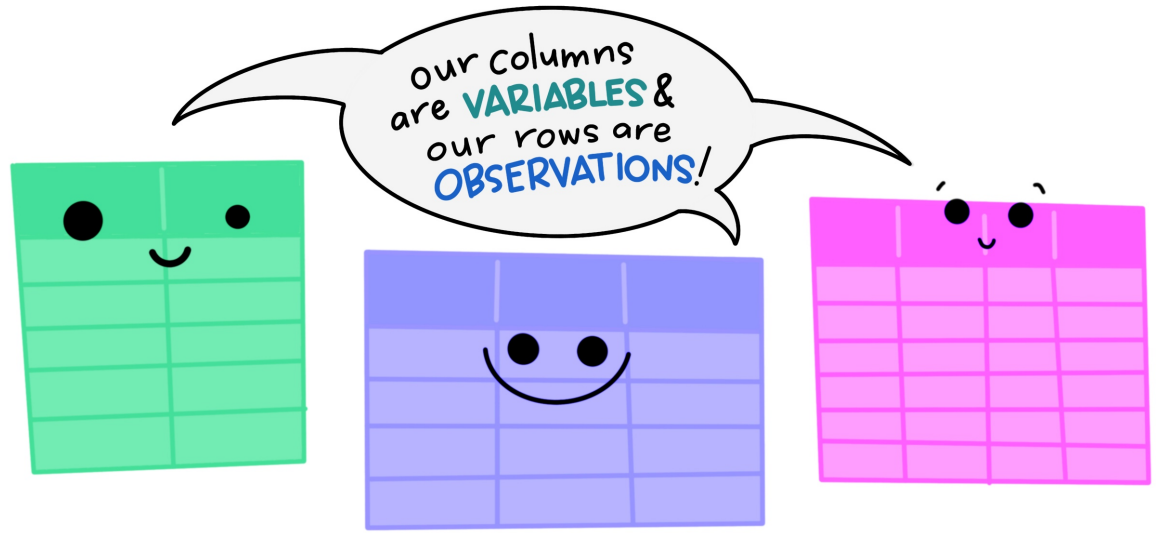
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



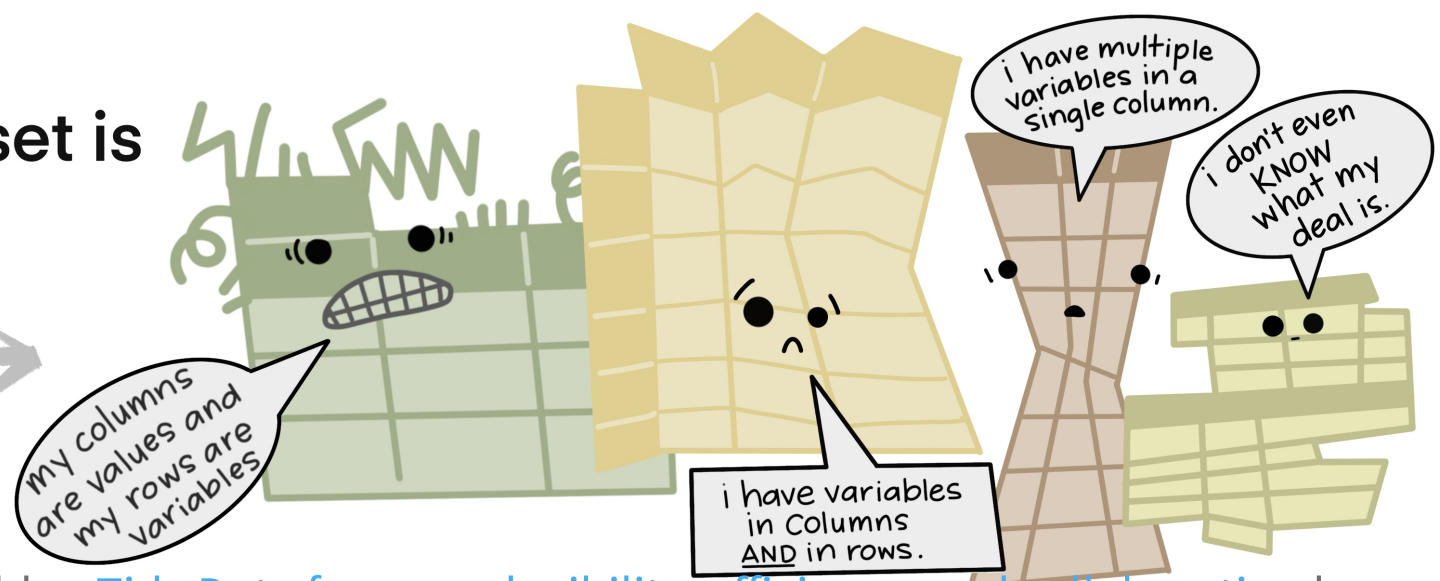
Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10). DOI: 10.18637/jss.v059.i10

Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

- The standard structure of tidy data means that "tidy datasets are all alike..."

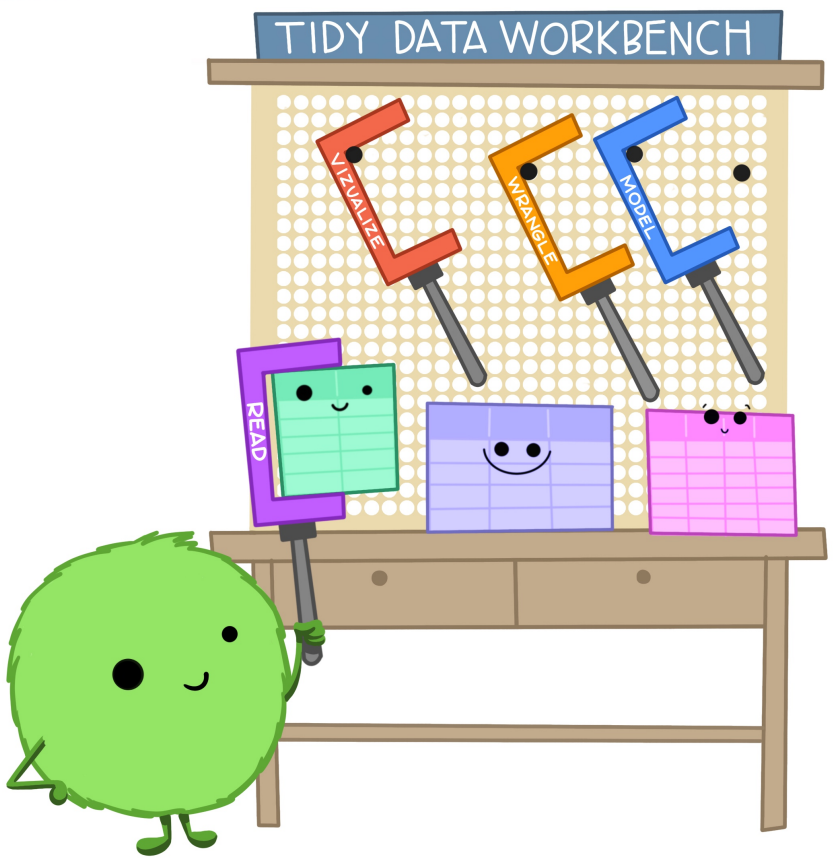


"...but every messy dataset is messy in its own way."
–HADLEY WICKHAM

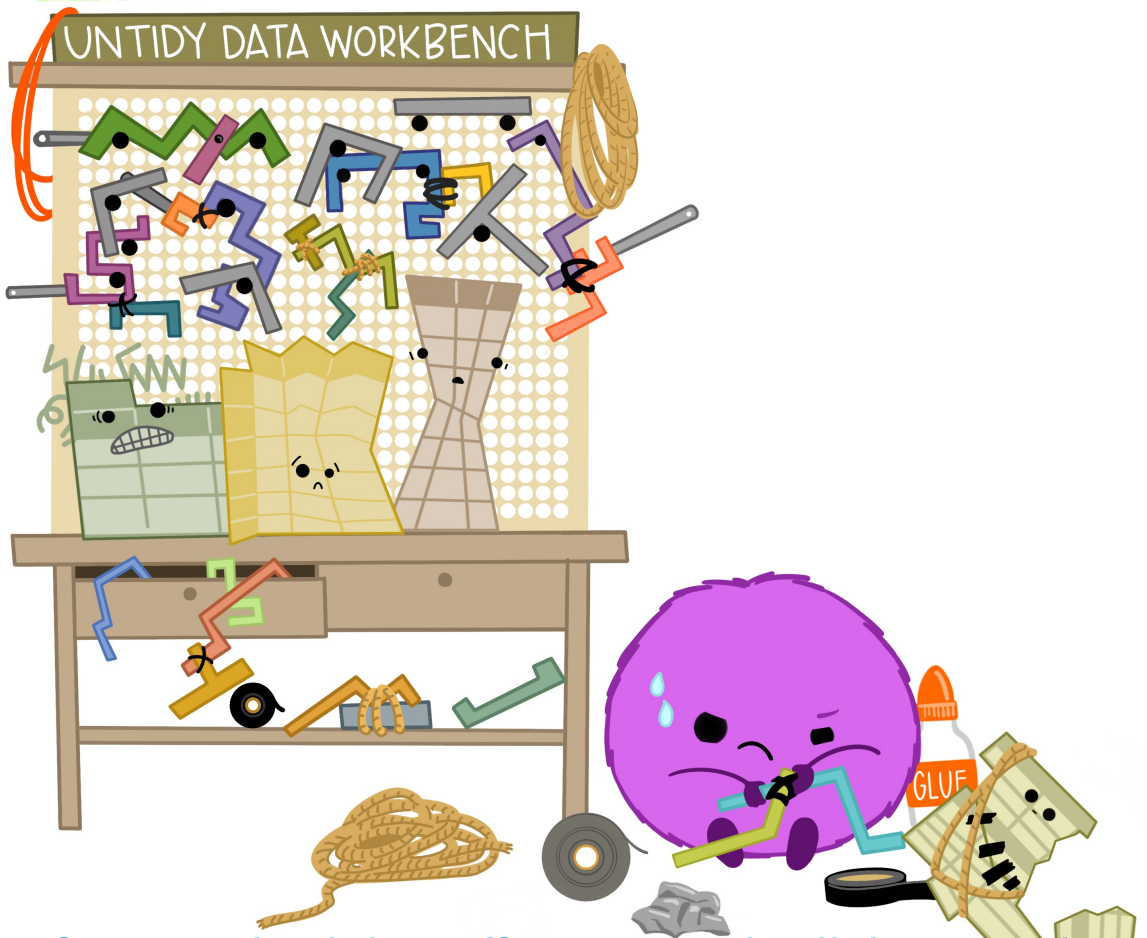


Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

When working with tidy data, we can use the same tools in similar ways for different datasets...

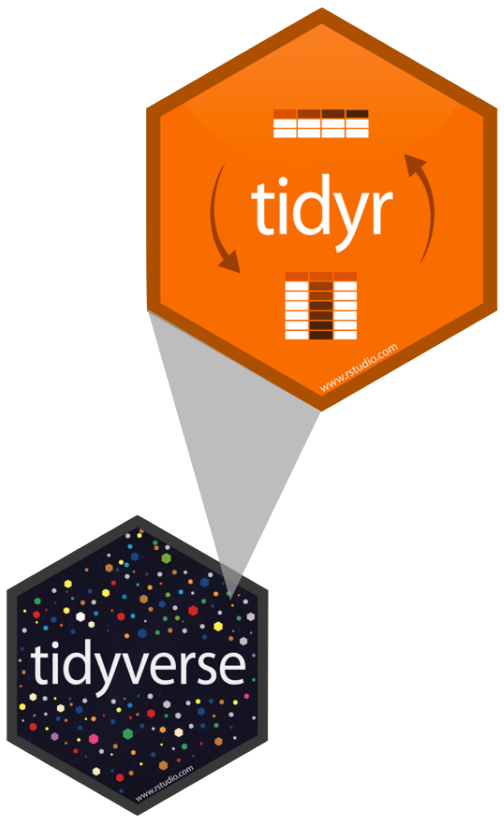


...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

A grammar of data tidying



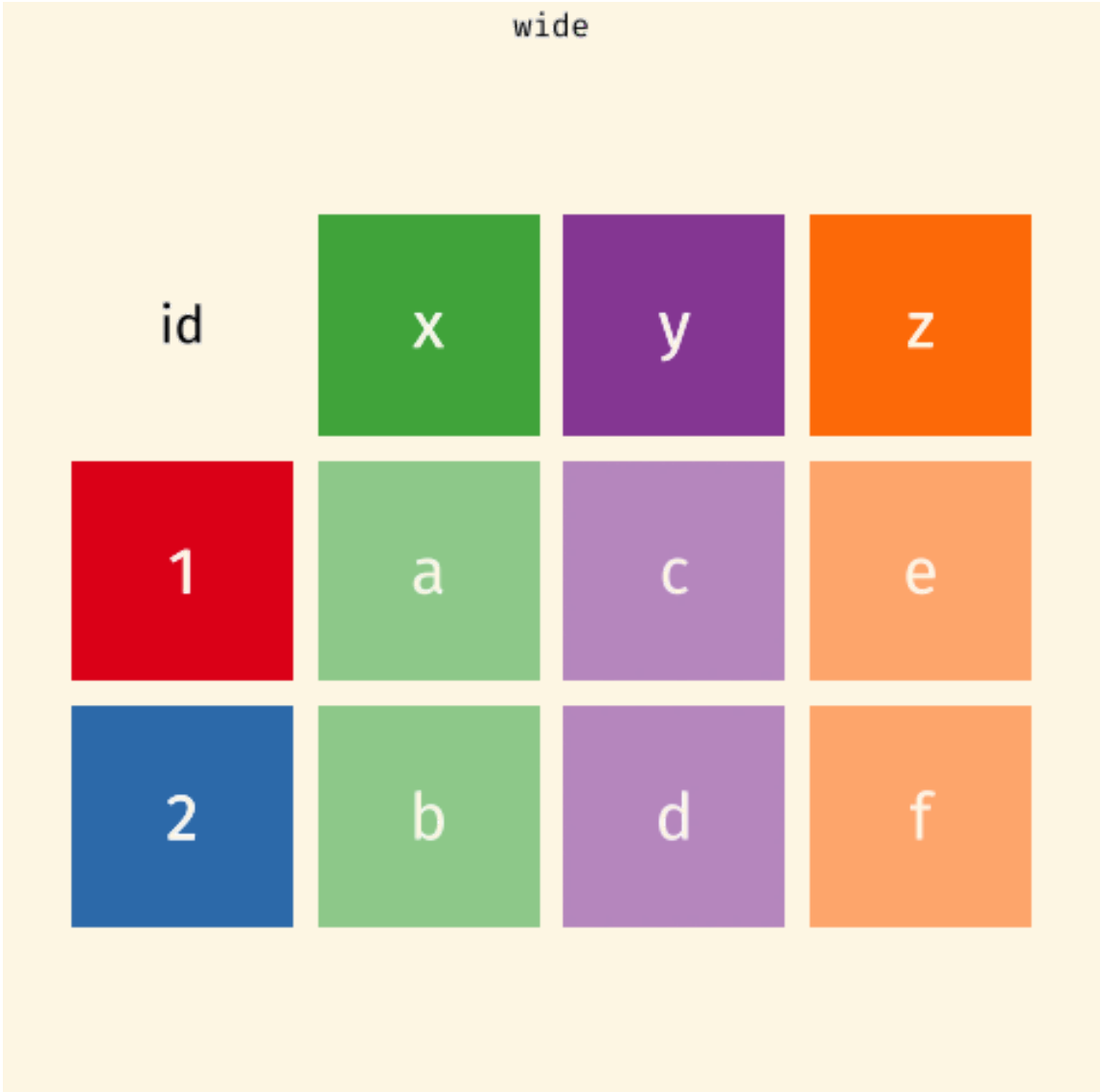
The goal of tidyr is to help you tidy your data via

- pivoting for going between wide and long data
- splitting and combining character columns
- nesting and unnesting columns
- clarifying how `NA`s should be treated

https://cven5837-ss22.github.io/website/

Slide taken from Data Science in a Box

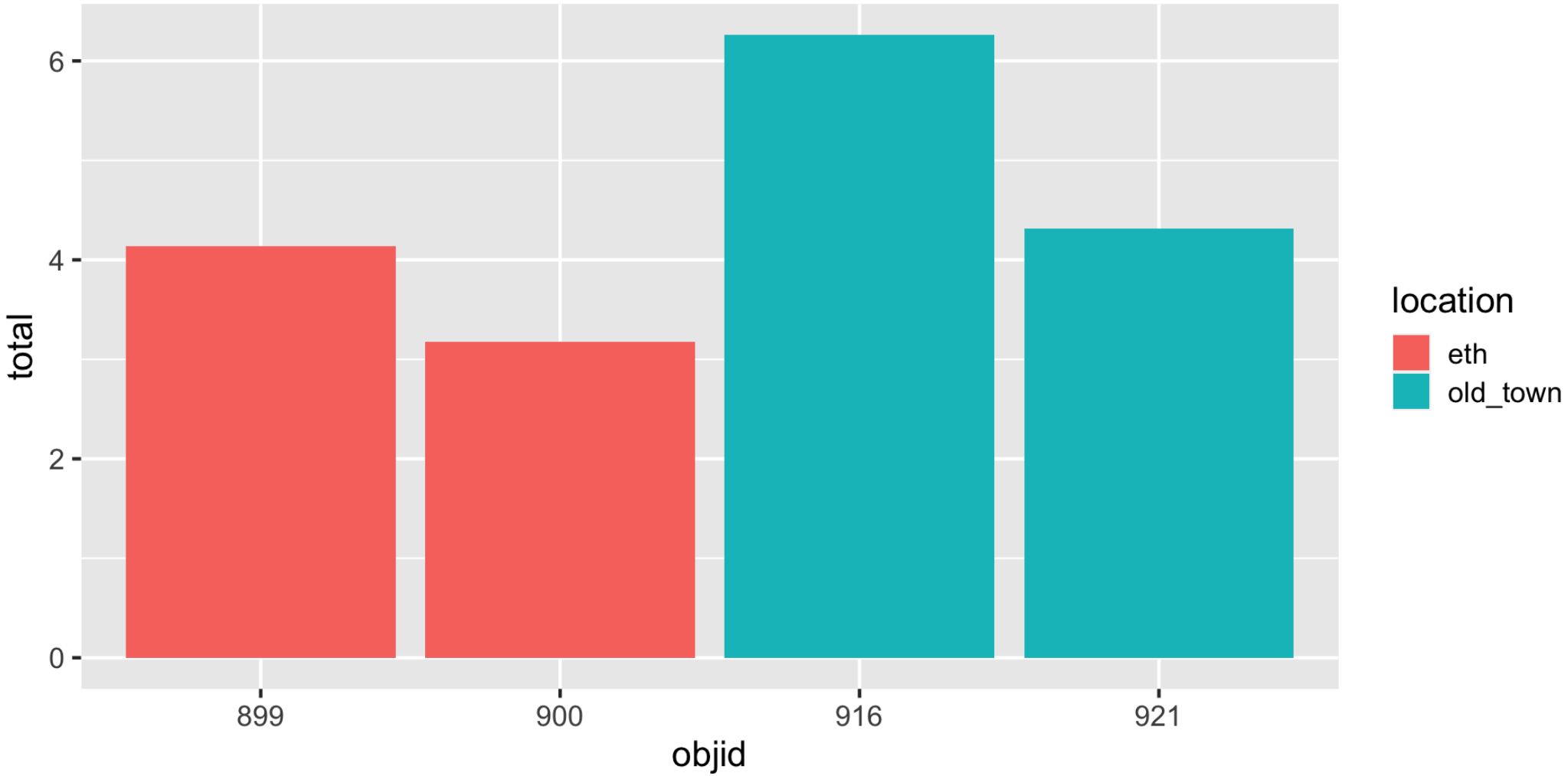
Pivoting data



Waste characterisation data

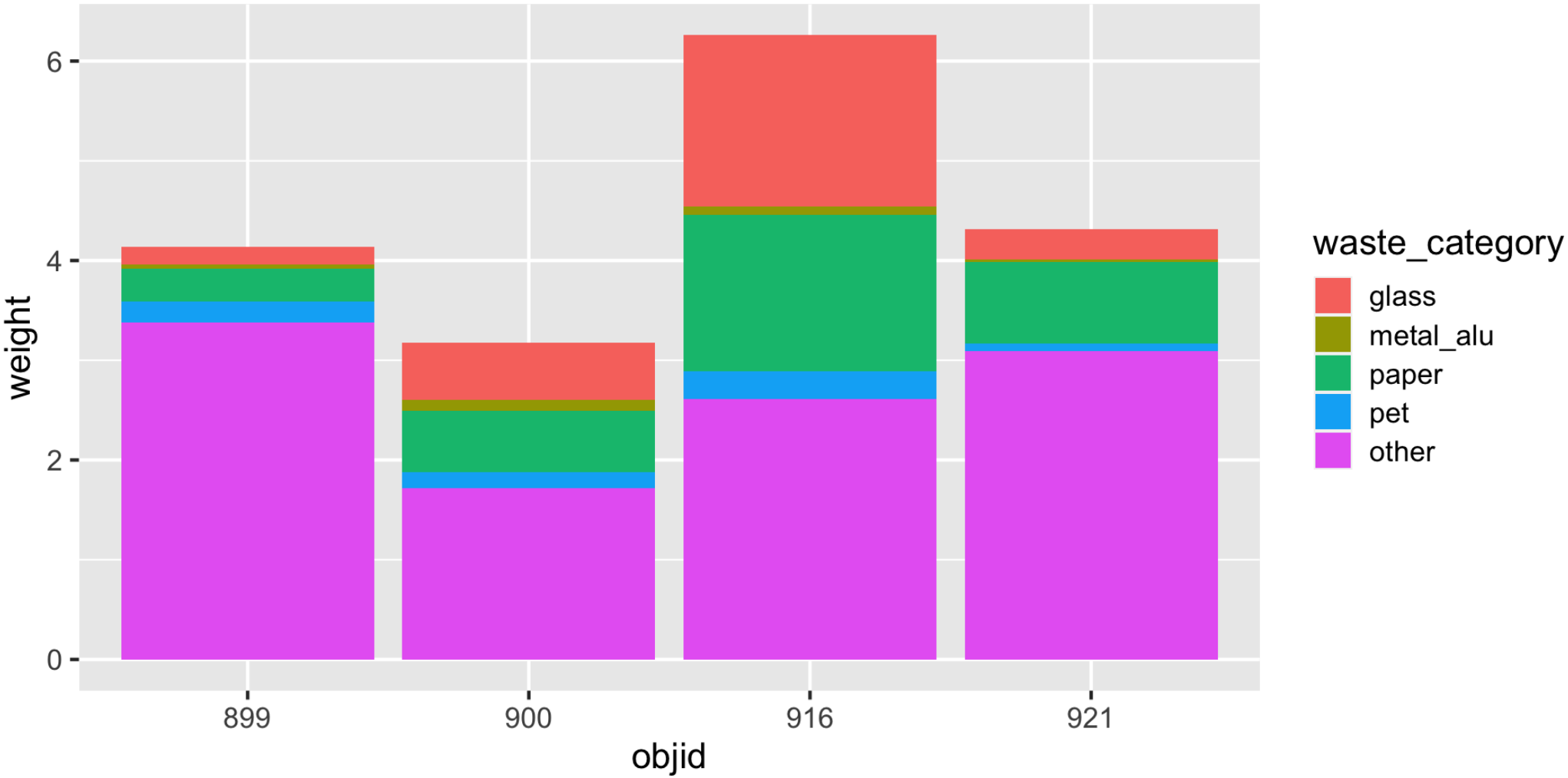
objid	location	pet	metal_alu	glass	paper	recyclable	non_recyclable	total
900	eth	0.06	0.06	0.58	0.21	0.92	1.14	2.05
899	eth	0.14	0.01	0.18	0.28	0.61	3.04	3.64
921	old_town	0.00	0.00	0.00	0.41	0.41	1.57	1.99
916	old_town	0.17	0.04	0.80	0.55	1.56	0.62	2.19
900	eth	0.10	0.04	0.00	0.40	0.54	0.58	1.12
899	eth	0.08	0.03	0.00	0.05	0.16	0.34	0.50
921	old_town	0.08	0.03	0.30	0.40	0.81	1.52	2.33
916	old_town	0.11	0.04	0.92	1.01	2.08	1.99	4.07

How would you plot this?



https://cven5837-ss22.github.io/website/

And this?



https://cven5837-ss22.github.io/website/

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

https://cven5837-ss22.github.io/website/

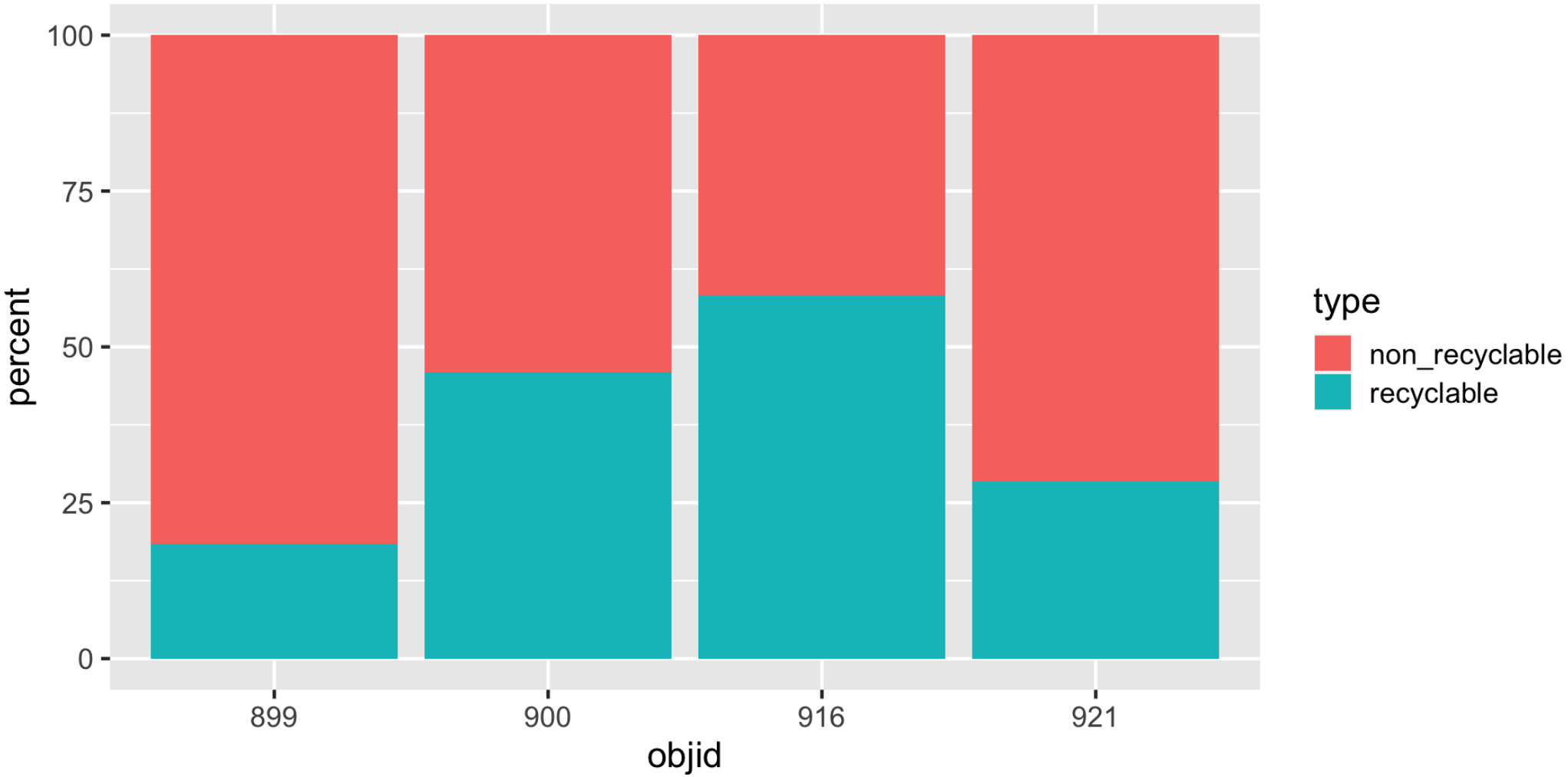
objid	location	waste_category	weight
921	old_town	paper	0.41
921	old_town	other	1.57
916	old_town	pet	0.17
916	old_town	metal_alu	0.04
916	old_town	glass	0.80
916	old_town	paper	0.55
916	old_town	other	0.62
900	eth	pet	0.10
900	eth	metal_alu	0.04
900	eth	glass	0.00
900	eth	paper	0.40
900	eth	other	0.58
899	eth	pet	0.08
899	eth	metal_alu	0.03
899	eth	glass	0.00
899	eth	paper	0.05

objid	location	waste_category	weight
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

Reminder: The wide format

objid	location	pet	metal_alu	glass	paper	recyclable	non_recyclable	total
900	eth	0.06	0.06	0.58	0.21	0.92	1.14	2.05
899	eth	0.14	0.01	0.18	0.28	0.61	3.04	3.64
921	old_town	0.00	0.00	0.00	0.41	0.41	1.57	1.99
916	old_town	0.17	0.04	0.80	0.55	1.56	0.62	2.19
900	eth	0.10	0.04	0.00	0.40	0.54	0.58	1.12
899	eth	0.08	0.03	0.00	0.05	0.16	0.34	0.50
921	old_town	0.08	0.03	0.30	0.40	0.81	1.52	2.33
916	old_town	0.11	0.04	0.92	1.01	2.08	1.99	4.07

Or this?


100
75
50
25
0
percent
899
900
916
921
objid
type
non_recyclable
recyclable

Calculate percentages

objid	location	waste_category	type	weight	percent
900	eth	pet	recyclable	0.06	2.02
900	eth	metal_alu	recyclable	0.06	1.95
900	eth	glass	recyclable	0.58	18.14
900	eth	paper	recyclable	0.21	6.74
900	eth	other	non_recyclable	1.14	35.78
899	eth	pet	recyclable	0.14	3.33
899	eth	metal_alu	recyclable	0.01	0.31
899	eth	glass	recyclable	0.18	4.30
899	eth	paper	recyclable	0.28	6.69
899	eth	other	non_recyclable	3.04	73.36
921	old_town	pet	recyclable	0.00	0.00
921	old_town	metal_alu	recyclable	0.00	0.00
921	old_town	glass	recyclable	0.00	0.00

objid	location	waste_category	type	weight	percent
921	old_town	paper	recyclable	0.41	9.60
921	old_town	other	non_recyclable	1.57	36.46
916	old_town	pet	recyclable	0.17	2.76
916	old_town	metal_alu	recyclable	0.04	0.69
916	old_town	glass	recyclable	0.80	12.73
916	old_town	paper	recyclable	0.55	8.82
916	old_town	other	non_recyclable	0.62	9.99
900	eth	pet	recyclable	0.10	3.09
900	eth	metal_alu	recyclable	0.04	1.35
900	eth	glass	recyclable	0.00	0.00
900	eth	paper	recyclable	0.40	12.60
900	eth	other	non_recyclable	0.58	18.33
899	eth	pet	recyclable	0.08	1.86
899	eth	metal_alu	recyclable	0.03	0.72
899	eth	glass	recyclable	0.00	0.00
899	eth	paper	recyclable	0.05	1.26

objid	location	waste_category	type	weight	percent
899	eth	other	non_recyclable	0.34	8.16
921	old_town	pet	recyclable	0.08	1.81
921	old_town	metal_alu	recyclable	0.03	0.70
921	old_town	glass	recyclable	0.30	6.89
921	old_town	paper	recyclable	0.40	9.32
921	old_town	other	non_recyclable	1.52	35.21
916	old_town	pet	recyclable	0.11	1.74
916	old_town	metal_alu	recyclable	0.04	0.70
916	old_town	glass	recyclable	0.92	14.63
916	old_town	paper	recyclable	1.01	16.20
916	old_town	other	non_recyclable	1.99	31.73

How to

```
1 waste_data_untidy
```

objid	location	pet	metal_alu	glass	paper	recyclable	non_recyclable	total
900	eth	0.06	0.06	0.58	0.21	0.92	1.14	2.05
899	eth	0.14	0.01	0.18	0.28	0.61	3.04	3.64
921	old_town	0.00	0.00	0.00	0.41	0.41	1.57	1.99
916	old_town	0.17	0.04	0.80	0.55	1.56	0.62	2.19
900	eth	0.10	0.04	0.00	0.40	0.54	0.58	1.12
899	eth	0.08	0.03	0.00	0.05	0.16	0.34	0.50
921	old_town	0.08	0.03	0.30	0.40	0.81	1.52	2.33
916	old_town	0.11	0.04	0.92	1.01	2.08	1.99	4.07

How to

```
waste_data_untidy |>
  select(objid:paper, non_recyclable)
```

objid	location	pet	metal_alu	glass	paper	non_recyclable
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 |>
  select(objid:paper, non_recyclable) |>
  rename(other = non_recyclable)
```

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_category_levels <- c("glass", "metal_alu", "paper", "pet", "other'

waste_data_untidy |>
  select(objid:paper, non_recyclable) |>
  rename(other = non_recyclable) |>
  pivot_longer(cols = pet:other,
               names_to = "waste_category",
               values_to = "weight") |>
  mutate(waste_category = factor(waste_category,
                                 levels = waste_category_levels))
```

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

objid	location	waste_category	weight
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
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

objid	location	waste_category	weight
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
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

How to

```
waste_category_levels <- c("glass", "metal_alu", "paper", "pet", "other"

waste_data_untidy |>
  select(objid:paper, non_recyclable) |>
  rename(other = non_recyclable) |>
  pivot_longer(cols = pet:other,
               names_to = "waste_category",
               values_to = "weight") |>
  mutate(waste_category = factor(waste_category,
                                 levels = waste_category_levels)) |>
  mutate(type = case_when(
    waste_category == "other" ~ "non_recyclable",
    TRUE ~ "recyclable")) |>
  relocate(type, .before = weight)
```

objid	location	waste_category	type	weight
900	eth	pet	recyclable	0.06
900	eth	metal_alu	recyclable	0.06
900	eth	glass	recyclable	0.58
900	eth	paper	recyclable	0.21

objid	location	waste_category	type	weight
900	eth	other	non_recyclable	1.14
899	eth	pet	recyclable	0.14
899	eth	metal_alu	recyclable	0.01
899	eth	glass	recyclable	0.18
899	eth	paper	recyclable	0.28
899	eth	other	non_recyclable	3.04
921	old_town	pet	recyclable	0.00
921	old_town	metal_alu	recyclable	0.00
921	old_town	glass	recyclable	0.00
921	old_town	paper	recyclable	0.41
921	old_town	other	non_recyclable	1.57
916	old_town	pet	recyclable	0.17
916	old_town	metal_alu	recyclable	0.04
916	old_town	glass	recyclable	0.80
916	old_town	paper	recyclable	0.55
916	old_town	other	non_recyclable	0.62

objid	location	waste_category	type	weight
900	eth	pet	recyclable	0.10
900	eth	metal_alu	recyclable	0.04
900	eth	glass	recyclable	0.00
900	eth	paper	recyclable	0.40
900	eth	other	non_recyclable	0.58
899	eth	pet	recyclable	0.08
899	eth	metal_alu	recyclable	0.03
899	eth	glass	recyclable	0.00
899	eth	paper	recyclable	0.05
899	eth	other	non_recyclable	0.34
921	old_town	pet	recyclable	0.08
921	old_town	metal_alu	recyclable	0.03
921	old_town	glass	recyclable	0.30
921	old_town	paper	recyclable	0.40
921	old_town	other	non_recyclable	1.52
916	old_town	pet	recyclable	0.11

How to

```r
waste_category_levels <- c("glass", "metal_alu", "paper", "pet", "other")

waste_data_untidy |>
  select(objid:paper, non_recyclable) |>
  rename(other = non_recyclable) |>
  pivot_longer(cols = pet:other,
               names_to = "waste_category",
               values_to = "weight") |>
  mutate(waste_category = factor(waste_category,
                                 levels = waste_category_levels)) |>
  mutate(type = case_when(
    waste_category == "other" ~ "non_recyclable",
    TRUE ~ "recyclable")) |>
  relocate(type, .before = weight) |>
  group_by(objid) |>
  mutate(percent = weight / sum(weight) * 100)
```

objid	location	waste_category	type	weight	percent
900	eth	pet	recyclable	0.06	2.02
900	eth	metal_alu	recyclable	0.06	1.95
900	eth	glass	recyclable	0.58	18.14

objid	location	waste_category	type	weight	percent
900	eth	paper	recyclable	0.21	6.74
900	eth	other	non_recyclable	1.14	35.78
899	eth	pet	recyclable	0.14	3.33
899	eth	metal_alu	recyclable	0.01	0.31
899	eth	glass	recyclable	0.18	4.30
899	eth	paper	recyclable	0.28	6.69
899	eth	other	non_recyclable	3.04	73.36
921	old_town	pet	recyclable	0.00	0.00
921	old_town	metal_alu	recyclable	0.00	0.00
921	old_town	glass	recyclable	0.00	0.00
921	old_town	paper	recyclable	0.41	9.60
921	old_town	other	non_recyclable	1.57	36.46
916	old_town	pet	recyclable	0.17	2.76
916	old_town	metal_alu	recyclable	0.04	0.69
916	old_town	glass	recyclable	0.80	12.73
916	old_town	paper	recyclable	0.55	8.82

objid	location	waste_category	type	weight	percent
916	old_town	other	non_recyclable	0.62	9.99
900	eth	pet	recyclable	0.10	3.09
900	eth	metal_alu	recyclable	0.04	1.35
900	eth	glass	recyclable	0.00	0.00
900	eth	paper	recyclable	0.40	12.60
900	eth	other	non_recyclable	0.58	18.33
899	eth	pet	recyclable	0.08	1.86
899	eth	metal_alu	recyclable	0.03	0.72
899	eth	glass	recyclable	0.00	0.00
899	eth	paper	recyclable	0.05	1.26
899	eth	other	non_recyclable	0.34	8.16
921	old_town	pet	recyclable	0.08	1.81
921	old_town	metal_alu	recyclable	0.03	0.70
921	old_town	glass	recyclable	0.30	6.89
921	old_town	paper	recyclable	0.40	9.32
921	old_town	other	non_recyclable	1.52	35.21

objid	location	waste_category	type	weight	percent
916	old_town	pet	recyclable	0.11	1.74
916	old_town	metal_alu	recyclable	0.04	0.70
916	old_town	glass	recyclable	0.92	14.63
916	old_town	paper	recyclable	1.01	16.20
916	old_town	other	non_recyclable	1.99	31.73

Live Coding Exercise: Pivoting

https://cven5837-ss22.github.io/website/

live-04a-tidyr-pivoting

1. Head over to rstudio.cloud
2. Open the workspace for the course (cven5837-ss22)
3. Open “Projects”
4. Open the “course-materials” project
5. Follow along with me

Homework week 4

https://cven5837-ss22.github.io/website/

Homework due dates

- All material on course website
- Homework assignment due: Friday, 29th July
- Learning reflection due: Monday, 1st August

https://cven5837-ss22.github.io/website/

Thanks! 

https://cven5837-ss22.github.io/website/

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