Data import & organization spreadshee

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

Lars Schöbitz

2025-03-13

Learning Objectives (for this m

1. Learners can import data from files in CSV
   sub-directories of the root directory.
2. Learners can explain the difference betwee
   character and the vector class factor.
3. Learners can discuss the difference betwee
   data, processed analysis-ready data, and da
   publication.
4. Learners can apply 12 principles for data or
   spreadsheets to the layout of a provided da
5. Learners can design a survey with five ques
   different types using Google Forms.

Homework modul

Task 3: Data for a country of yo

- for a country of your choice
- for the year 2000 and 2020
- for all variables that are not “safely managed sanitation service

Task 3: Data for a country of yo

- ☑ for the country you live or work in
- ☑ for the year 2000 and 2020
- ☑ for all variables that are not “safely managed sanitation services”

```
sanitation_u
  filter(iso
         yea
         var
```

Task 3: Data for a country of yo

- ☑ for the country you live or work in
- ☑ for the year 2000 and 2020
- ☑ for all variables that are not “safely managed sanitation services”

```
sanitation_u
  filter(iso
         yea
         var
```

Task 3: Data for a country of yo

- ☑ for the country you live or work in
- ☑ for the year 2000 and 2020
- ☑ for all variables that are not “safely managed sanitation services”

```
sanitation_u
  filter(iso
         yea
         var
```

```
sanitation_u
  count(iso3
```

iso3
UGA
UGA
UGA
UGA
UGA
UGA
UGA
UGA

Task 3: Data for a country of yo

- ☑ for the country you live or work in
- ☒ for the year 2000 and 2020
- ☑ for all variables that are not “safely managed sanitation services”

```
sanitation_u
  filter(iso
         yea
         var
```

```
sanitation_u
  count(iso3
```

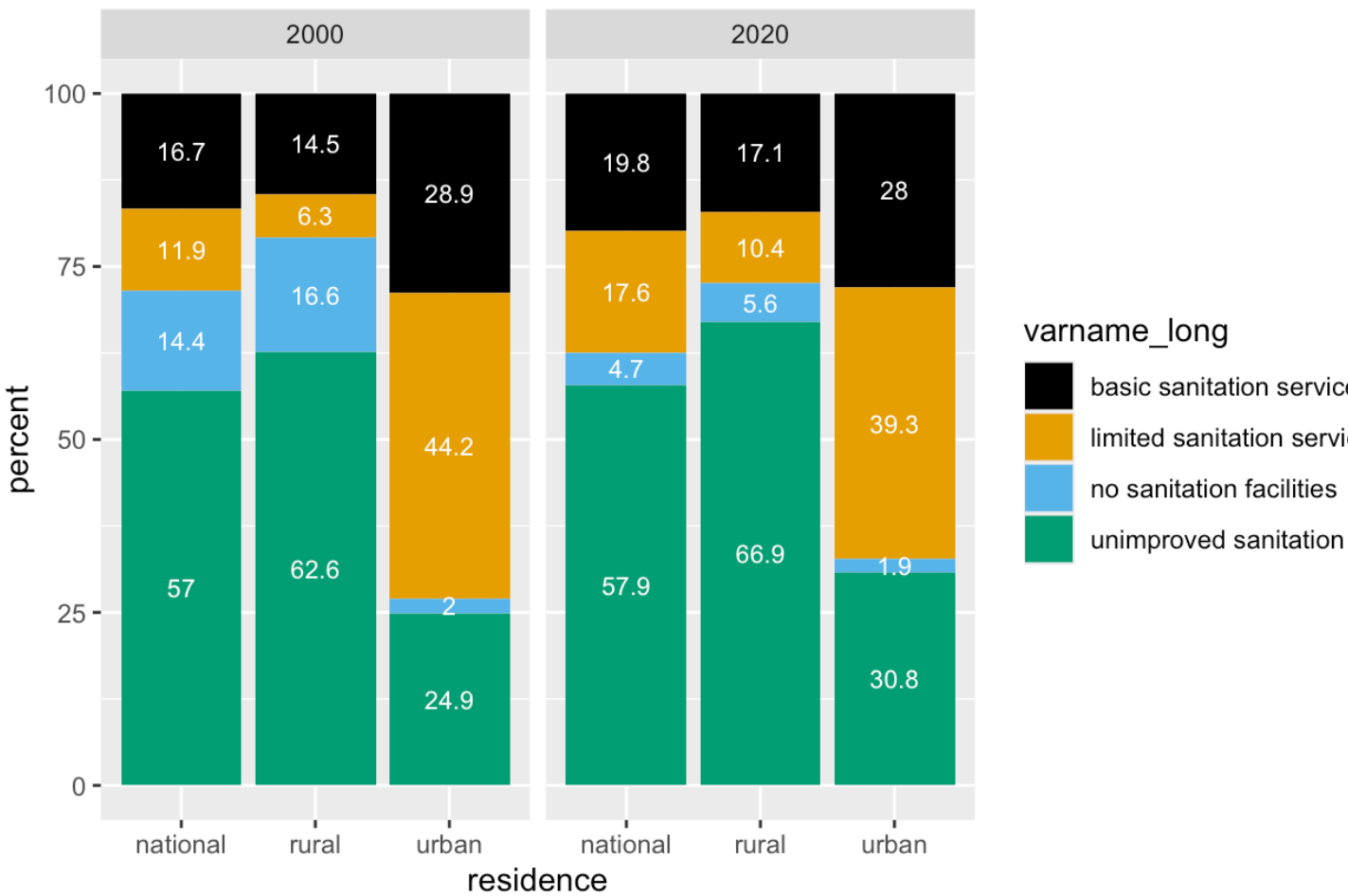
iso3

```
# A tibble: 0 × 4
# ℹ 4 variables:
varname_short <ch
```

- One row canno
  2020) for the s
- One year canno
  time
- One year is eith

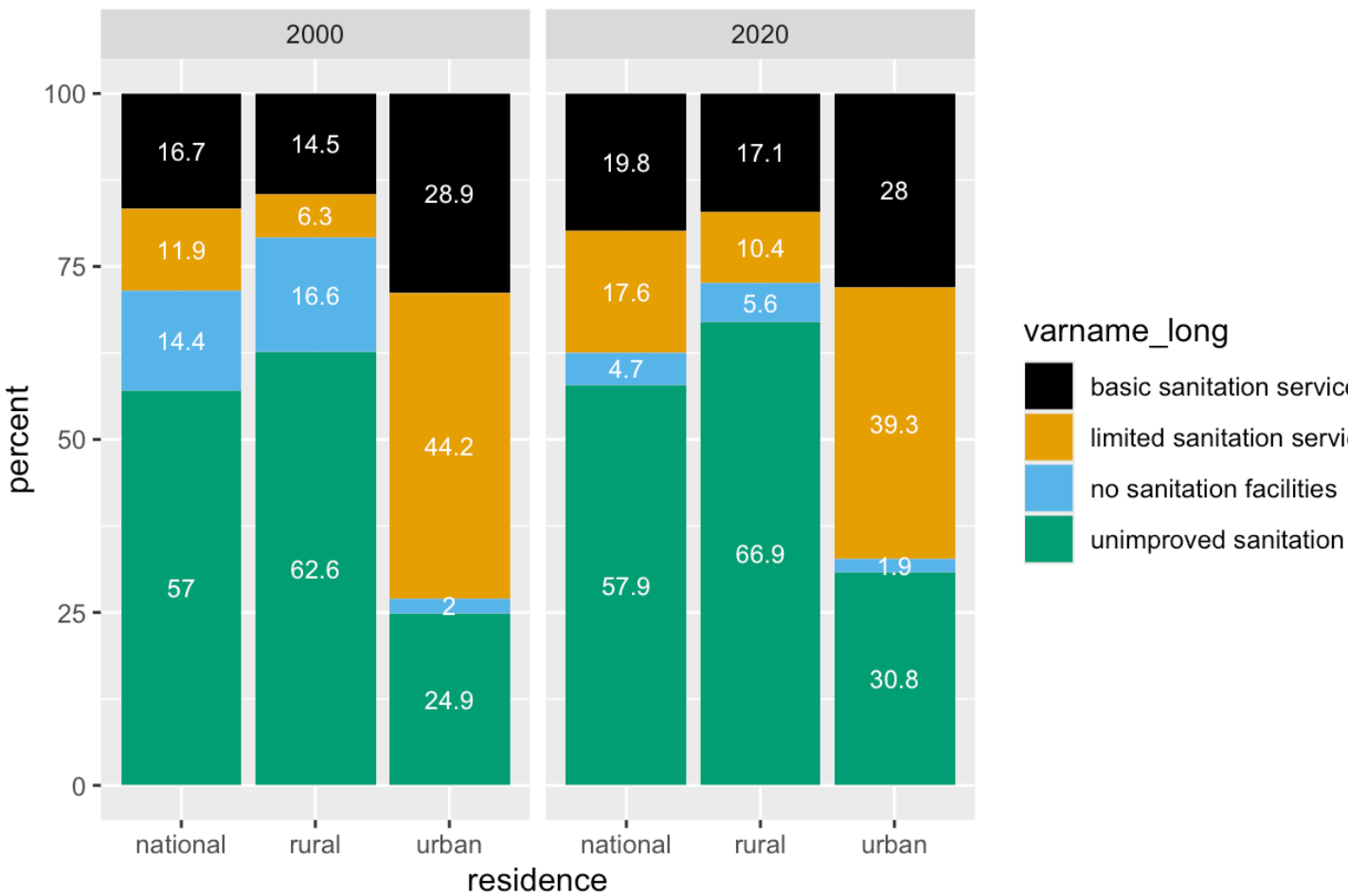
Task 5 & 6: Make a plot & inspe

1. Look at the plot that you created. What do you notice about the order of the bars / or
2. What would you want to change?
3. Why did we remove “safely managed sanitation services” from the data set in Task 3



Task 5 & 6: Make a plot & inspe

1. Look at the plot that you created. What do you notice about the order of the bars / or
2. What would you want to change? put in order of the “sanitation ladder”
3. Why did we remove “safely managed sanitation services” from the data set in Task 3

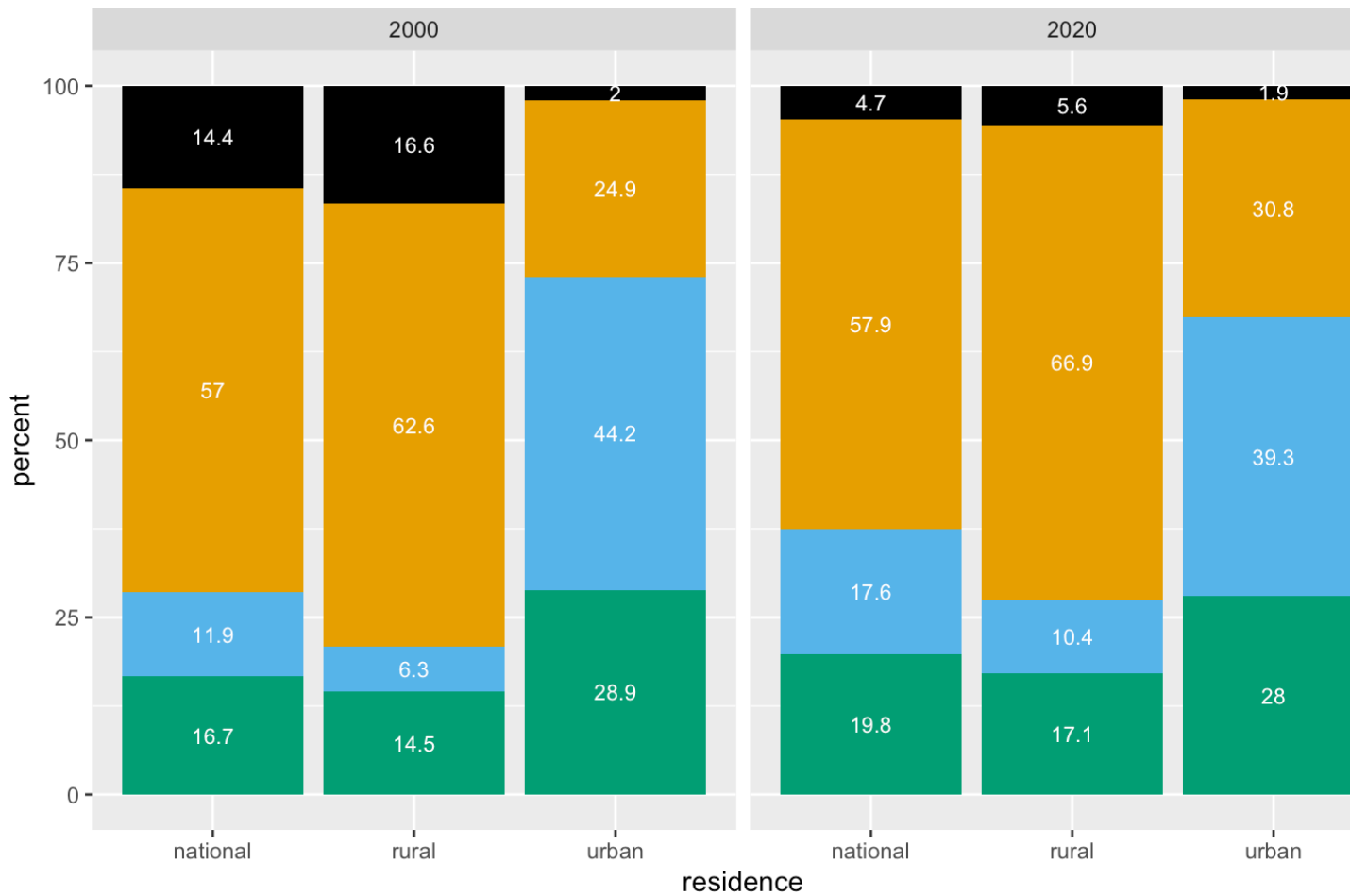


Sanitation ladder?

varname_short	varname_long	simplified
san_sm	safely managed sanitation services	a decent toilet th moved & treated
san_bas	basic sanitation services (improved sanitation facilities which are not shared)	a decent toilet th
san_lim	limited sanitation services (improved sanitation facilities which are shared)	a decent toilet th
san_unimp	unimproved sanitation facilities	an inadequate toi
san_od	no sanitation facilities (open defecation)	no toilet

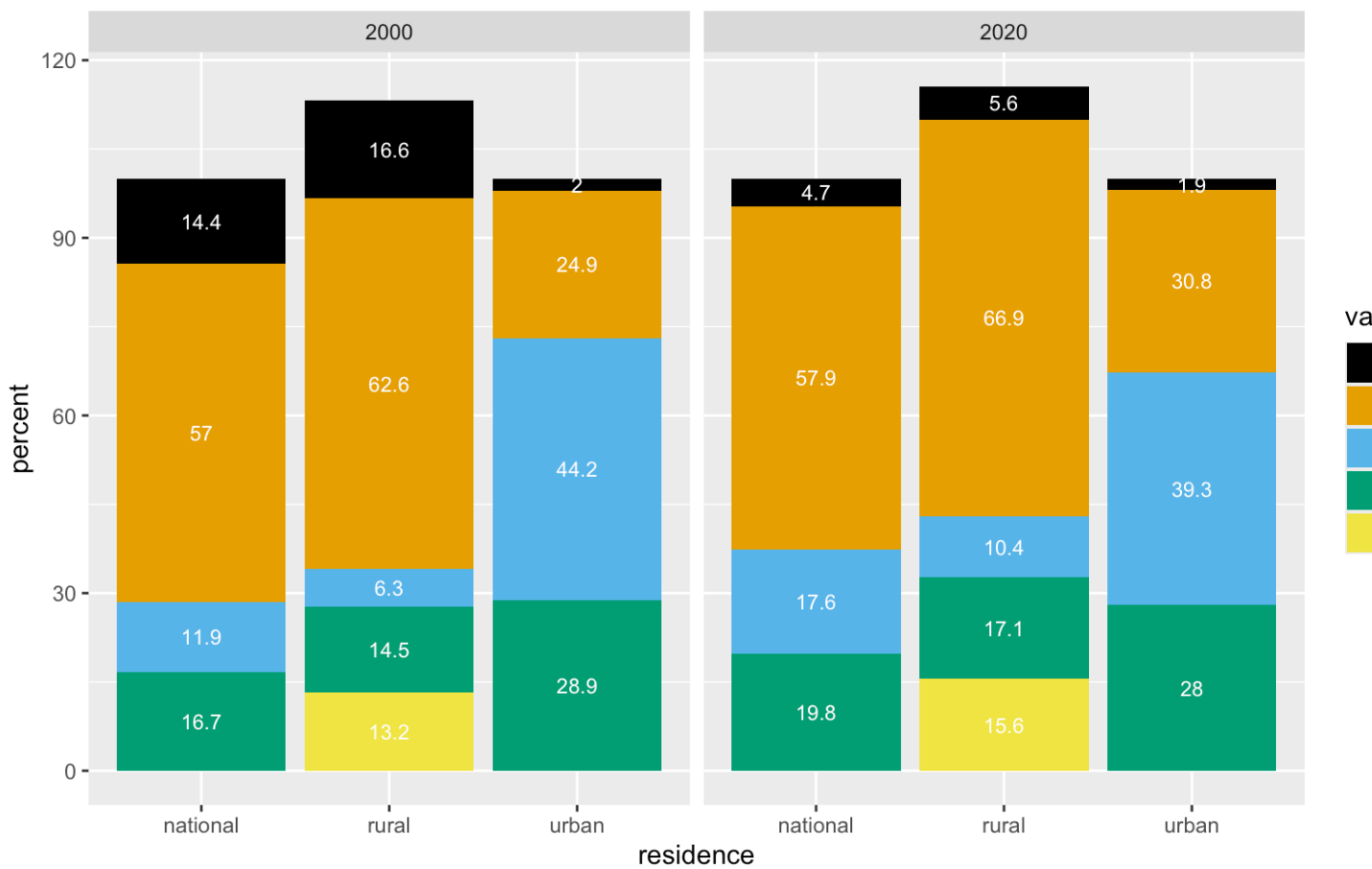
Task 5 & 6: Make a plot & inspe

1. Look at the plot that you created. What do you notice about the order of the bars / or
2. What would you want to change? put in order of the “sanitation ladder”
3. Why did we remove “safely managed sanitation services” from the data set in Task 3



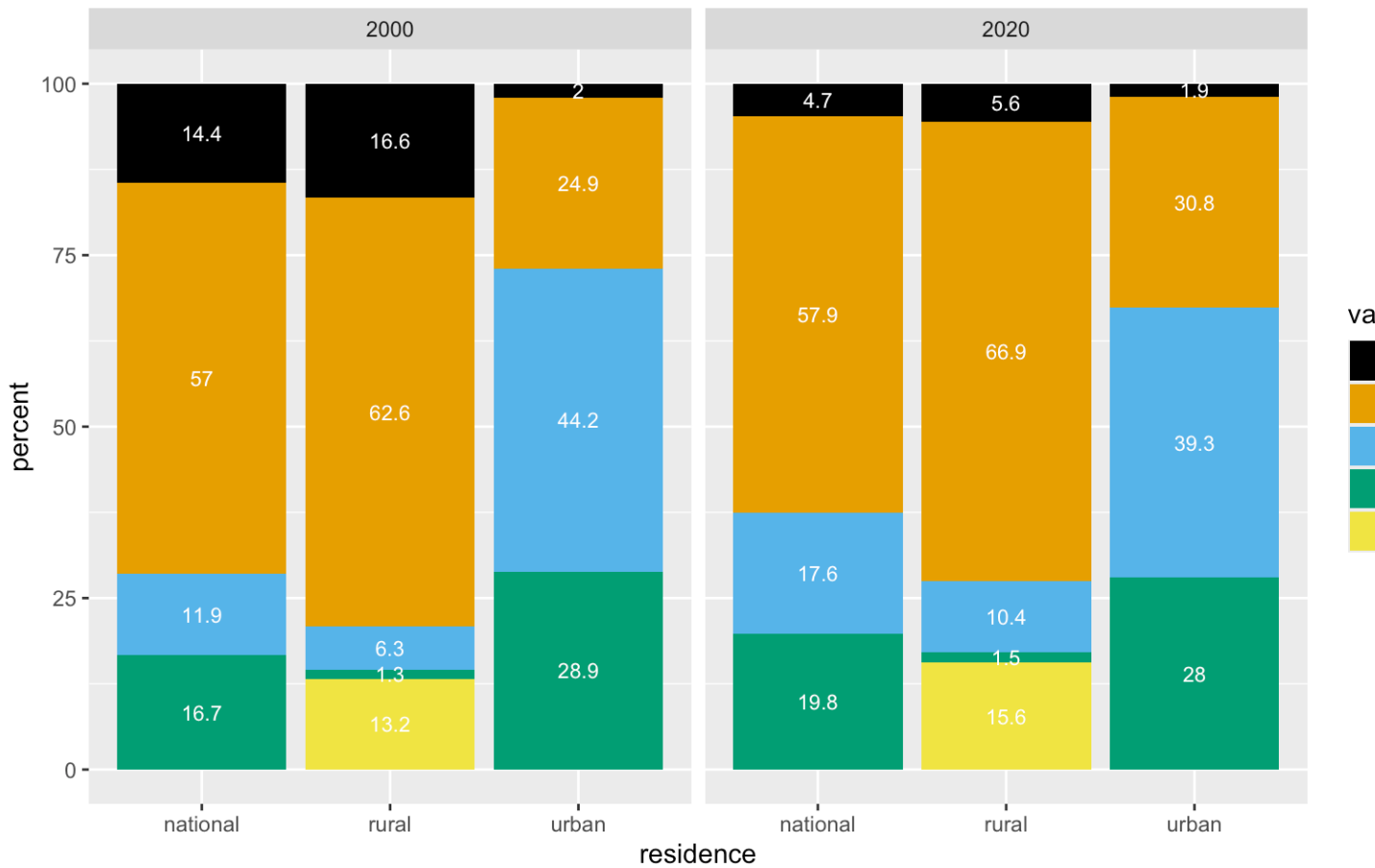
Task 5 & 6: Make a plot & inspe

1. Look at the plot that you created. What do you notice about the order of the bars / or
2. What would you want to change? put in order of the “sanitation ladder”
3. Why did we remove “safely managed sanitation services” from the data set in Task 3
   100%, a fraction of people with basic services have safely managed services



Task 5 & 6: Make a plot & inspe

1. Look at the plot that you created. What do you notice about the order of the bars / or
2. What would you want to change? put in order of the “sanitation ladder”
3. Why did we remove “safely managed sanitation services” from the data set in Task 3
   100%, a fraction of people with basic services have safely managed services



Types of variables - Remember

numerical

discrete variables

- non-negative
- whole numbers
- e.g. number of students, roll of a dice

continuous variables

- infinite number of values
- also dates and times
- e.g. length, weight, size

non-num

categorica

- finite num
- distinct gr countries,
- **ordinal** if ordering (school gra

Factors in R

My turn: Factors in R

Sit back and e

Take a break

Please get up and move!

Your turn: md-04a-exercises - f

1. Open posit.cloud in your browser (use your bookmark)
2. Open the rbtl-fs25 workspace for the course.
3. In the File Manager in the bottom right window, locat
   `factors-your-turn.qmd` file and click on it to op
   window.
4. Follow instructions in the file

Data import

Reading rectangular data into R

CSV & XLSX

readr

- `read_csv()` - comma delimited files
- `read_csv2()` - semicolon separated files (common in countries where , is used as the decimal place)
- `read_tsv()` - tab delimited files
- `read_delim()` - reads in files with any delimiter
- ...

readxl

- `read_excel(`
- ...

Reading data from CSV files

- import unprocessed raw data

```r
waste <- read_csv("data/raw/waste-city-level.csv")

waste
```

```
# A tibble: 367 × 113
   iso3c region_id country_name       income_id city_name additional
   <chr> <chr>     <chr>              <chr>     <chr>     <chr>
 1 AFG   SAS       Afghanistan        LIC       Jalalabad <NA>
 2 AFG   SAS       Afghanistan        LIC       Kandahar  <NA>
 3 AFG   SAS       Afghanistan        LIC       Mazar-E-… <NA>
 4 AFG   SAS       Afghanistan        LIC       Kabul     <NA>
 5 AFG   SAS       Afghanistan        LIC       HiratÂ    <NA>
 6 AGO   SSF       Angola             LMC       Luanda    <NA>
 7 ALB   ECS       Albania            UMC       Korca     <NA>
 8 ALB   ECS       Albania            UMC       Vlora     <NA>
 9 ARE   MEA       United Arab Emira… HIC       Abu Dhabi <NA>
10 ARE   MEA       United Arab Emira… HIC       Dubai     <NA>
# ℹ 357 more rows
# ℹ abbreviated name: ¹additional_data_annual_budget_for_waste_manag
# ℹ 107 more variables: additional_data_annual_solid_waste_budget_ye
#   additional_data_annual_swm_budget_2017_year <dbl>,
#   additional_data_annual_swm_budget_year <dbl>,
#   additional_data_annual_waste_budget_year <dbl>,
```

Writing data as CSV files

- transform data
- export processed analysis-ready data

```r
# data transformation
waste_sml <- waste |>
  select(country_name, city_name, iso3c, income_id,
         total_msw_total_msw_generated_tons_year,
         population_population_number_of_people) |>
  rename(country = country_name,
         city = city_name,
         generation_tons_year = total_msw_total_msw_generated_t
         population = population_population_number_of_people)

# export processed analysis-ready data
write_csv(waste_sml, "data/processed/waste-city-level-sml.csv")
```

Reading data from XLSX files

- import unprocessed raw data

```r
sludge <- read_excel("data/raw/tbl-01-faecal-sludge-analysis.xl
                     sheet = 1)
sludge
```

```
# A tibble: 20 × 6
      id date_sample         system      location      users     ts
   <dbl> <dttm>              <chr>       <chr>         <dbl>  <dbl>
 1     1 2023-11-01 00:00:00 pit latrine household         5 136.
 2     2 2023-11-01 00:00:00 pit latrine household         7 102.
 3     3 2023-11-01 00:00:00 pit latrine household        NA  57.0
 4     4 2023-11-01 00:00:00 pit latrine household         6  27.0
 5     5 2023-11-01 00:00:00 pit latrine household        12  97.3
 6     6 2023-11-02 00:00:00 septic tank household         7  78.2
 7     7 2023-11-02 00:00:00 septic tank household        14  15.2
 8     8 2023-11-02 00:00:00 septic tank household         4  29.4
 9     9 2023-11-02 00:00:00 septic tank household        10  64.2
10    10 2023-11-02 00:00:00 septic tank household        12   8.01
11    11 2023-11-03 00:00:00 pit latrine public toilet    50  11.2
12    12 2023-11-03 00:00:00 pit latrine public toilet    32  84.0
13    13 2023-11-03 00:00:00 pit latrine public toilet    41  55.9
14    14 2023-11-03 00:00:00 pit latrine public toilet   160  15.3
15    15 2023-11-03 00:00:00 pit latrine public toilet    20  22.6
16    16 2023-11-04 00:00:00 septic tank public toilet    26   8.72
```

Writing data as CSV files

- transform data
- export data underlying a publication

```r
# data transformation
tbl_sludge_summary <- sludge |>
  filter(!is.na(users)) |>
  group_by(system, location) |>
  summarise(
    count = n(),
    mean_ts = mean(ts),
    sd_ts = sd(ts),
    median_ts = median(ts)
  )

# export data underlying a publication
write_csv(tbl_sludge_summary, "data/final/tbl-01-faecal-sludge-
```

system	location	count	mean_ts	sd_ts	media
pit latrine	household	4	90.7	45.9	
pit latrine	public toilet	5	37.8	31.3	
septic tank	household	5	39.0	30.8	
septic tank	public toilet	5	20.4	14.3	

(Research) Data Management

Examples of terms used when managing

term	folder	explanation
unprocessed raw data	raw	data that is not processed and r its original form and file
processed analysis-ready data	processed	data that is processed to prepar analysis and is exported in its ne a new file
final data underlying a publication	final	data that is the result of an analy descriptive statistics or data visu and shown in a report, but then a exported in its new form as a nev

Take a break

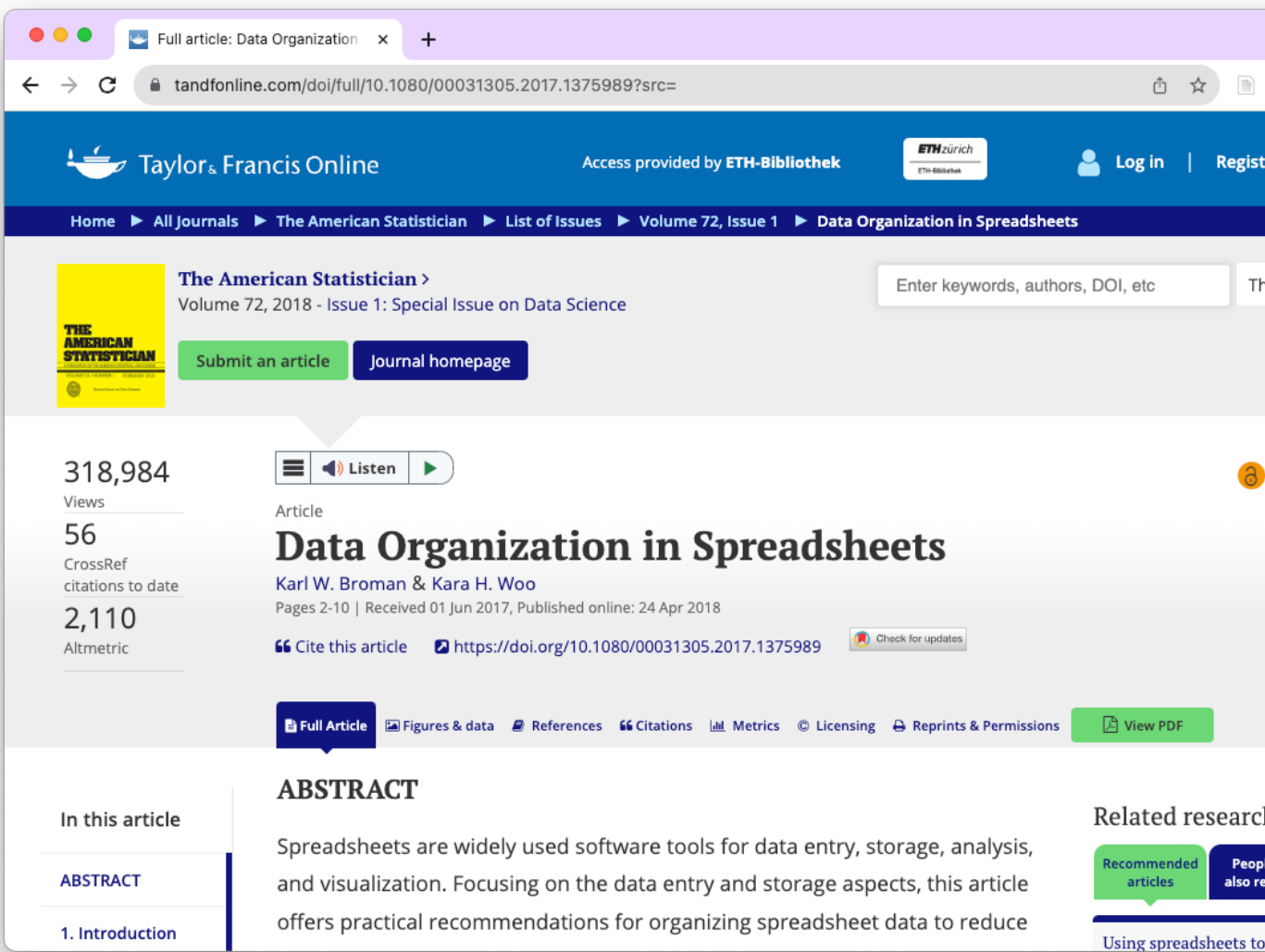
Please get up and move!

Your turn: md-04a-exercises - i

1. Open posit.cloud in your browser (use your bookmark)
2. Open the rbtl-fs25 workspace for the course.
3. In the File Manager in the bottom right window, locat
   `your-turn.qmd` file and click on it to open it in the
4. Follow instructions in the file

Data Organization
Spreadsheets

Data Organization in Spreadsh



Data Organization in Spreadsh

Read the paper (it's part of your homework), b

- Go through the annotated slides: https://kbr
  Talk_DataOrg/dataorg_notes.pdf
- Watch Karl Broman give the talk (02:36 to 4
  youtu.be/t74E0a90gkA?t=156
- Read the content on a website: https://kbro

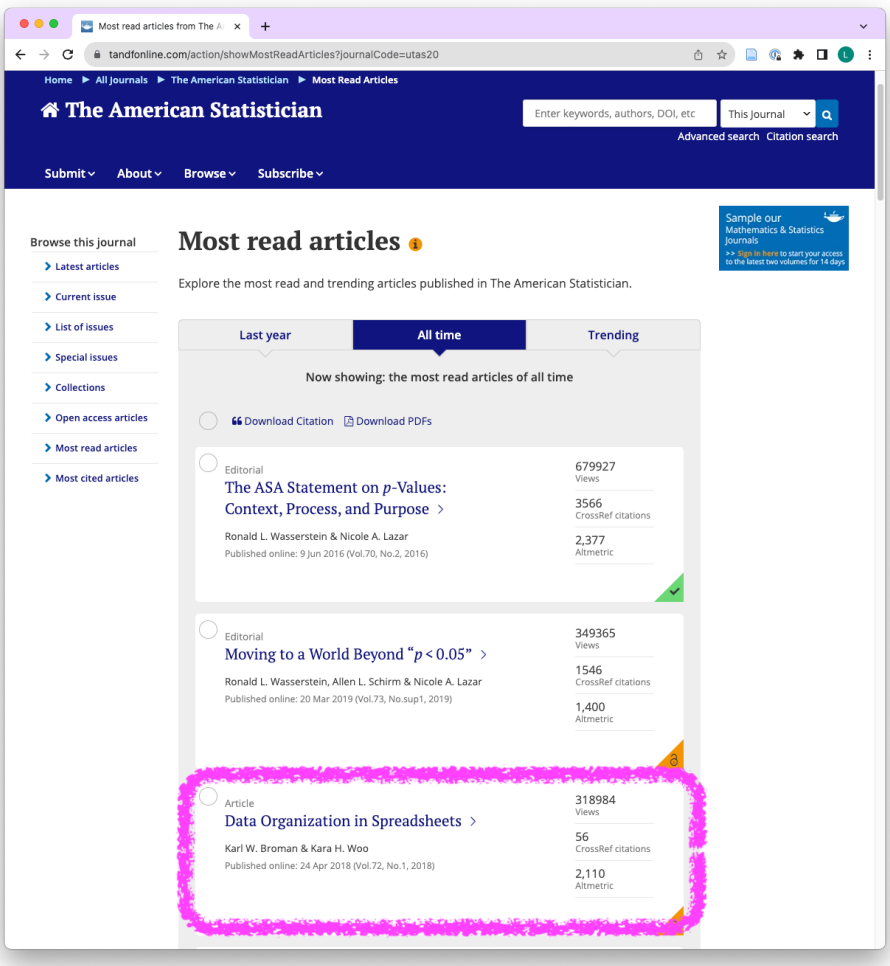
Data Organization in Spreadsh

But, especially apply it to your data


DATA

Data Organization in Spreadsh

Why? Because following a set of rules for orga
everyone's live a little better.



- 3rd most
American
- 310'000+
- widely acc
standards

rbtl-fs25.github.io/website/

Data Organization in Spreadsh

License? CC0 (!)

README.md

Data organization in spreadsheets

Slides for a talk for the OSGA Webinar Series, on 24 Sept 2021, based on my paper of the
with Kara Woo. Also see the related website.

PDF of slides: https://kbroman.org/Talk_DataOrg/dataorg.pdf

PDF of slides with notes: https://kbroman.org/Talk_DataOrg/dataorg_notes.pdf

Video of presentation: https://youtu.be/t74E0a90gkA

License

To the extent possible under law, Karl Broman has waived all copyright and related or neig
rights to "Data organization in spreadsheets". This work is published from the United State

PUBLIC DOMAIN

Homework assign
module 4

Module 4 documentation

rbtl-fs25.github.io/website/modu

Module 4

Data import & Data organization in spreadsheets

Learning Objectives

1. Learners can import data from files in CSV format located in s directories of the root directory.
2. Learners can explain the difference between the vector class character and the vector class factor.
3. Learners can discuss the difference between unprocessed raw processed analysis-ready data, and data underlying a publicati
4. Learners can apply 12 principles for data organisation in spreadsheets to the layout of a provided dataset.
5. Learners can design a survey with five questions of three diffe types using Google Forms.

Slides

Homework due date

- Homework assignment due: Wednesday, Ma
- Correction & feedback phase up to: Tuesda

Wrap-up

Thanks! 🌻

Slides created via revealjs and Quarto: https:/
presentations/revealjs/

Access slides as PDF on GitHub

All material is licensed under Creative Comm
Alike 4.0 International.