Data wrangling

Susi Zajitschek

May 2025

Wranging data…

Data collection: raw /messy

→ Organise

→ Clean

→ Summarise

→ Report / communicate

Forth and back between R and Excel?

- Possible, but….

“finalDataset_Chapter2_filtered_version4b_THIS_one.xlxs”
(and about 100 similar files in various folders and subfolders)

- Using mainly R:

  Reproducible workflow (for others AND YOUSELF):

  Raw data – clean data – summarised – analysed

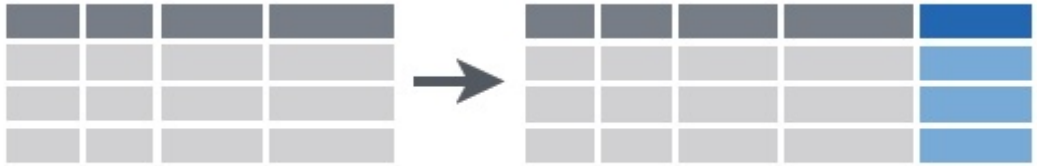
“Manipulating data”

- Manipulation of dataframes, not raw data
- Organisation
- Necessary for analysis & plotting

    - Making new variables
    - Subsetting data
    - Summarising data
    - Combining datasets
    - Reshaping the dataframe

“Manipulating data”

- Making new variables

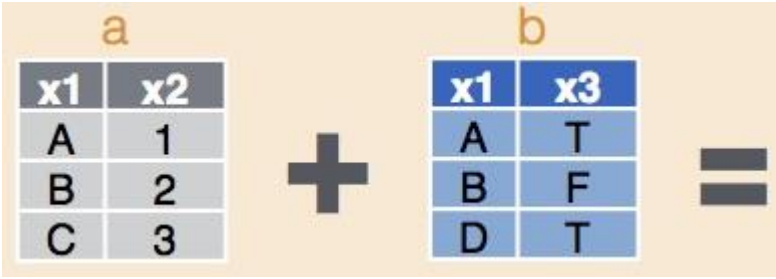


- Subsetting data

- Summarising data

“Manipulating data, continued”

- Combining datasets



- Reshaping the dataframe

“Long” format

country	year	metric
x	1960	10
x	1970	13
x	2010	15
y	1960	20
y	1970	23
y	2010	25
z	1960	30
z	1970	33
z	2010	35

“Wide” format

country	yr1960	yr1970	yr2010
x	10	13	15
y	20	23	25
z	30	33	35

Definition, “tidy data”

Each

- Colum: variable
- Row: observation
- Value: cell

In R

library(“tidyverse”) OR subsets: library(“tidyr”) / library(“dplyr”)

The “pipe”: %>%

- Operator “%>%”
- Chains multiple steps into a sequence of analysis
- “read it out loud” : a more logical sequence than the base code version offers

Example: data(iris)

Count how many "setosa" are wider than 0.2

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor

Base R

```
sum(nrow(subset(iris, Species == "setosa" & Petal.Width > 0.2)))
```

library(dplyr)

```
iris  %>%
    filter(Species == "setosa", Petal.Width > 0.2)  %>%
    summarise(sum_n = n())
```