Tidy data reshaping & summaries

Athanasia Monika Mowinckel

Sept. 15th 2020

LCBC
LIFESPAN CHANGES
in brain and cognition

Part 2

Tidy data reshaping & summaries

Tidy data reshaping & summaries

- pivoting data with tidyr (~25 min)
- grouped summaries with dplyr (~25 min)
- working with nested data using purrr (~25 min)

tidyr

pivoting / altering data shape

tidyr

The goal of tidyr is to help you create tidy data.

Tidy data is data where:

- Every column is variable.
- Every row is an observation.
- Every cell is a single value.

Tidy data describes a standard way of storing data that is used wherever possible throughout the tidyverse. If you ensure that your data is tidy, you'll spend less time fighting with the tools and more time working on your analysis. Learn more about tidy data in `vignette("tidy-data")`.

Tall/long vs. wide data

- Tall (or long) data are considered "tidy", in that they adhere to the three tidy-data principles

- Wide data are not necessarily "messy", but have a shape less ideal for easy handling in the tidyverse

Example in longitudinal data design:

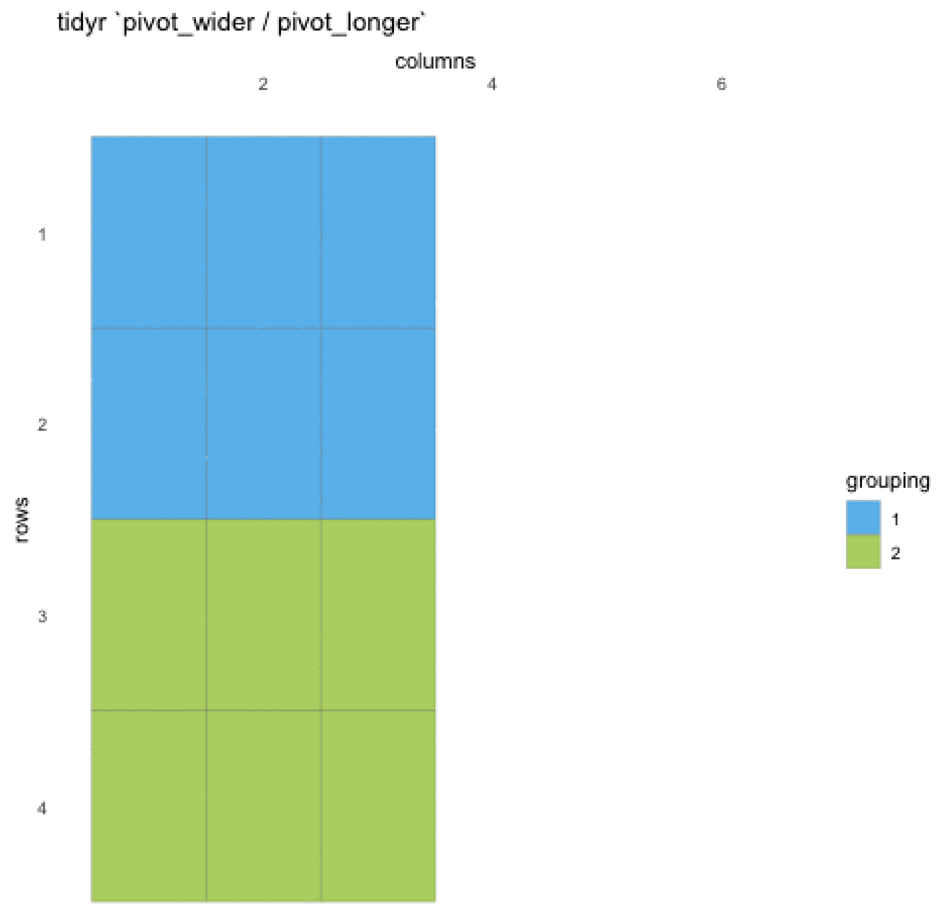
- wide data: each participant has a single row of data, with all longitudinal observations in separate columns
- tall data: a participant has as many rows as longitudinal time points, with measures in separate columns

tidyr

pivoting

`pivot_longer()` - wide to long
`pivot_wider()` - long to wide

Transforms data shape



Pivoting longer

takes tidy-select column arguments, so it is easy to grab all the columns you are after.

```
penguins %>%
  pivot_longer(contains("_"))
```

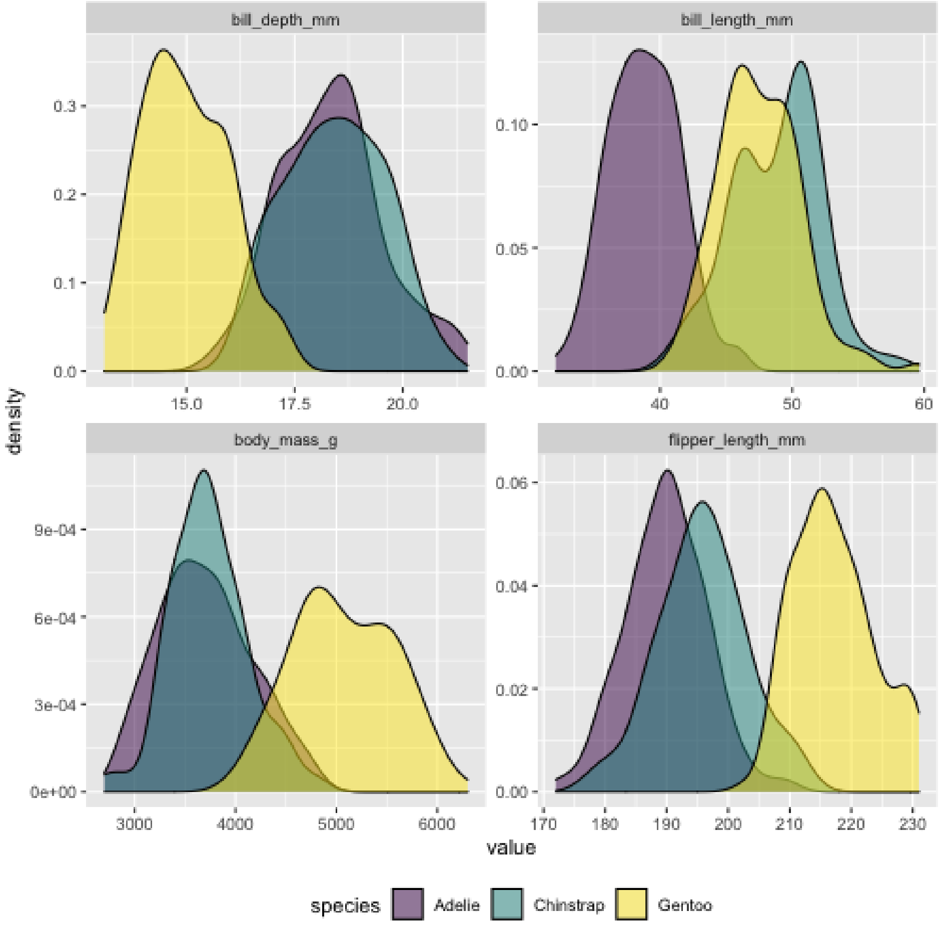
```
## # A tibble: 1,376 x 6
##    species island    sex     year name              value
##    <fct>   <fct>     <fct>  <int> <chr>             <dbl>
##  1 Adelie  Torgersen male    2007 bill_length_mm     39.1
##  2 Adelie  Torgersen male    2007 bill_depth_mm      18.7
##  3 Adelie  Torgersen male    2007 flipper_length_mm 181  
##  4 Adelie  Torgersen male    2007 body_mass_g      3750  
##  5 Adelie  Torgersen female  2007 bill_length_mm     39.5
##  6 Adelie  Torgersen female  2007 bill_depth_mm      17.4
##  7 Adelie  Torgersen female  2007 flipper_length_mm 186  
##  8 Adelie  Torgersen female  2007 body_mass_g      3800  
##  9 Adelie  Torgersen female  2007 bill_length_mm     40.3
## 10 Adelie  Torgersen female  2007 bill_depth_mm      18  
## # … with 1,366 more rows
```

Why pivot longer?

Can be convenient for easy sub-plots with ggplot

```
penguins %>%
  pivot_longer(contains("_")) %>%

  ggplot(aes(x = value, fill = species)) +
  geom_density() +
  facet_wrap(~ name, scales = "free") +
  scale_fill_viridis_d(alpha = .5) +
  theme(legend.position = "bottom")
```



pivoting wider

```
penguins_long <- penguins %>%
  mutate(id = row_number()) %>%
  pivot_longer(contains("_"),
               names_to = c("body_part",
                            "measure",
                            "unit"),
               names_sep = "_")

penguins_long %>%
  pivot_wider(names_from = c("body_part", "measure", "unit"), # pivot these columns
              values_from = "value", # take the values from here
              names_sep = "_") # separate names_from with this character
```

```
## # A tibble: 344 x 9
##    species island sex     year    id bill_length_mm bill_depth_mm
##    <fct>   <fct>  <fct>  <int> <int>          <dbl>         <dbl>
##  1 Adelie  Torge… male    2007     1           39.1          18.7
##  2 Adelie  Torge… fema…   2007     2           39.5          17.4
##  3 Adelie  Torge… fema…   2007     3           40.3          18  
##  4 Adelie  Torge… <NA>    2007     4           NA            NA  
##  5 Adelie  Torge… fema…   2007     5           36.7          19.3
##  6 Adelie  Torge… male    2007     6           39.3          20.6
##  7 Adelie  Torge… fema…   2007     7           38.9          17.8
##  8 Adelie  Torge… male    2007     8           39.2          19.6
```

Go to RStudio

live demo

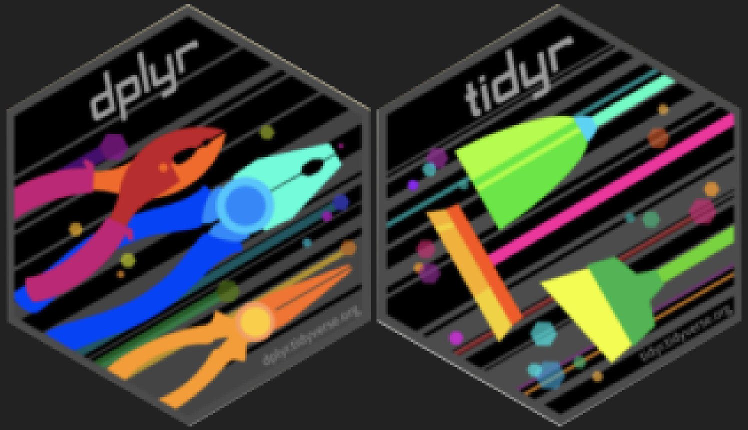
Go to subsetting exercises

```
learnr::run_tutorial("005-pivoting",
"tidyquintro")
```

08:00

dplyr + tidyr

data summaries



dplyr - comparison to base-R

tidy

```
penguins %>%
  summarise(mean(bill_length_mm, na.rm = TRUE))
```

base

```
mean(penguins$bill_length_mm, na.rm = TRUE)
```

https://dplyr.tidyverse.org/articles/base.html

Go to RStudio

live demo

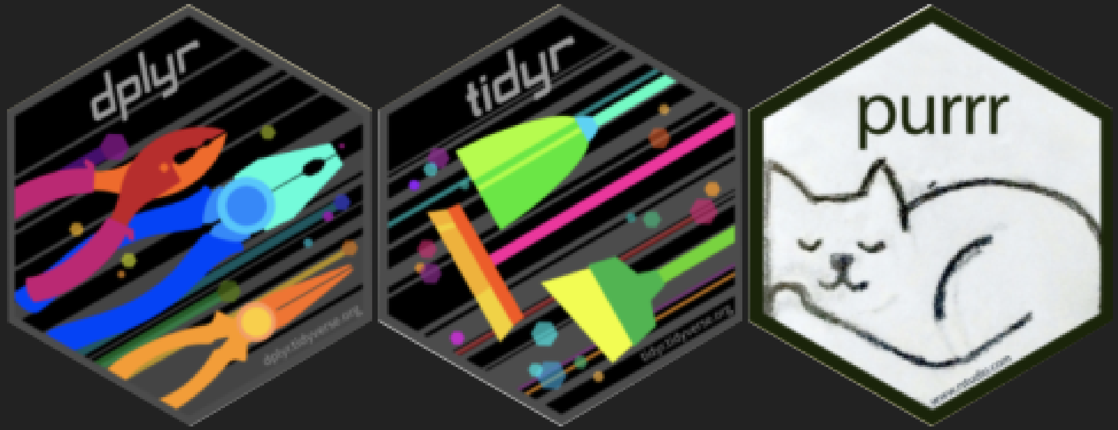
Go to subsetting exercises

```
learnr::run_tutorial("006-summarising",
"tidyquintro")
```

08:00

dplyr + tidyr + purrr

Working with nested data - avoiding loops



comparison to base-R

tidy

```r
penguins %>%
    nest_by(species, island) %>%
    mutate(lm_model = list(
      lm(bill_length_mm ~ bill_depth_mm, data = data)
      ))
```

base

```r
penguins$groups <- interaction(penguins$species, penguins$island)

models <- list()
for(i in 1:length(unique(penguins$groups))){
  tmp <- penguins[penguins$groups == groups[i],]
  models[[i]] <- lm(bill_length_mm ~ bill_depth_mm, data = data)
}

# or
lapply(unique(penguins$groups), function(x) lm(bill_length_mm ~ bill_depth_mm,
                                               data = penguins[penguins$groups == x,]))
```

Go to RStudio

live demo

Go to subsetting exercises

```
learnr::run_tutorial("006-nesting",
"tidyquintro")
```

08:00

End of part 2

LCBC LIFESPAN CHANGES in brain and cognition