Module 3: R

Submodule 5: Wrangling

Expected length: 0.5 day

Guiding question: How can we use data that isn’t already in our RStudio, isn’t arranged correctly, needs to be combined with other data, or is oversize?

Concepts: data import, pivoting, joining, tidy data, data.table

Description: This lesson introduces students to working with real data: importing, changing the layout entirely, and combining data as required. It also gives an overview of working with larger data using data.table.

Instructor Preparation: Run all code to ensure it generates the same output as described in the lesson. Select dataset for tidy data live coding.

| Materials and resources | Learning objectives |
| --- | --- |
| 05-wrangling\_deck.html | 1. Ability to import diverse data types  2. Ability to assess if data is long or wide, and pivot as needed  3. Ability to join data and select appropriate joins  4. Basic familiarity with data.table and how it looks different from tidyverse |

| Length | Lesson content | Guidelines, tips, and tricks |
| --- | --- | --- |
| 15min | Importing data (Wickham and Grolemund, 2017, Chapter 11; Timbers et al. 2021, Chapter 2) & Saving data |  |
| 30min | Pivot (Wickham and Grolemund, 2017, Chapter 12; Timbers et al. 2021, Chapter 3.4) |  |
| 30min | Joining data (Wickham and Grolemund, 2017, Chapter 13) |  |
| 60min | Live coding: Tidying data | Ask students in advance for datasets of interest or use data that you are familiar with. Walk through loading, pivoting, and joining live – ideally with little preparation beforehand, so students can see the process. Ask for contributions to make decisions during the process. |
| 30min | data.table (Wiley and Wiley, 2020, Chapter 7; https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html) | Participants should be following along in their own RStudio environments. |
| 60min | Formative exercises  1. Tidy the data below:  ```{r}  data <- tibble(  group = c("treat", "control"),  survival = c(17, 11),  deceased = c(3, 9)  )  ```  2. Join the dataset flights to the dataset airlines. What should the key(s) be? What do the different types of joins look like?  ```{r}  library(nycflights13)  data("flights")  data("airlines")  ```  3. Using flights and data.table, group based on cases that have dep\_delay < 0 and those that have arr\_delay > 0 and count the number in each group. How many groups/rows are there? How many in each group? | Allow 30 minutes of independent work time.  Take up for 30 minutes, with RStudio open to demonstrate.  Ask students for errors they encountered. |
|  | Summative assessment: Discuss datasets | At this point, students have the skills required to being their project. Give them the opportunity to ask questions about what they have learning so far, specifically in relation to their independent summative work. |