Smart ways to perform common tasks in R

April 2021

Although the tidyverse provides lot of packages that may be useful to perform common data preparation and univariate or multivariate statistical summaries, other solutions do exist. I am pretty confident that base-only R packages have been developed since I last used R for intensive data analysis (I can think of strengejacke's packages for data visualization and data manipulation, for instance), but the Hmisc packages existed long before plyr, then dplyr, were published on CRAN, and it already worked pretty well. It's still the best piece of software ever written for R after MASS, in my view. Together with the rms and ggplot2 packages, you will get the data munging triumvirate almost for free. Almost because you will have to learn a lot and to choose wisely among the numerous options. In addition, remember that "lightweight is the right weight". If you can perform 80% of your tasks with three packages and builtin stuff, then you're on the right side of the Pareto law. Here's your starter kit for Hmisc.

Disclaimer: I have nothing against the tidyverse way of doing things (I just barely understand why we really need another "rlang", and I regret the problems of reverse dependencies that it may have caused in the past). If that suits your needs, that's all fine. If on the contrary you get stuck on basic data manipulation stuff, or built-in one-liner R functions don't play well with "tibble", then you're probably going in the wrong direction. Start with Phil Spector's textbook [1], then eventually learn a bit of plyr [2]. The 80% of time spent in data manipulation, as we use to say, is not a joke: You will likely spend most of your time switching from long to wide format, aggregating data all over again and again, and not just before even you start building your super nice statistical model.

References

[1] Phil Spector. *Data Manipulation with R*. New York, NY: Springer, 2008.

[2] H Wickham. "The Split-Apply-Combine Strategy for Data Analysis". In: *Journal of Statistical Software* 40.1 (2011).