Reducing the pain with good R tools

John Tillinghast
American University
For Data Wranglers DC

The dirty secret of data science

- Everyone like to pretend it's all about cool software and cutting-edge algorithms
- When you actually do it, you spend 80% of your time on data wrangling
- Much of this uses (almost) no statistical judgment and should be (almost) automatable
- Includes: reading it in; arranging it; fixing "obvious" errors; rescaling/transforming; linking items in different data sets; generating more appropriate variables; etc.

Improving your workflow

- 80% on data wrangling means you are spending only 20% of your time on brain work
- 40% would mean that you're spending 60% of your time on brain work
- You have just made yourself 3x more productive! ☺

Arranging data

- “Long” format vs. “wide” format
- Grouping data for separate analysis of subsets
- Two approaches
  - Have a general tool that can do the arranging for you
  - Add analysis tools to *avoid* the rearranging (do it behind the scenes)

Tools in this talk

- Lubridate: helps interpret and manipulate dates
- Plyr (the original, very popular):
  - Allows subgrouping *without* rearranging
- Dplyr: new (2014) version of plyr
  - Faster because of Rcpp
  - Some changes in syntax

Dplyr

- Update of very popular plyr package
- Dplyr has compiled code, so it's much faster
- Requires R 3.0.2
- "Basic verbs": filter, arrange, select, mutate, summarise
- Groupwise is split-apply-combine
- Join vs merge
- Has very consistent syntax (but not the same as plyr)

Split-apply-combine analysis

- Very often, you want to analyze different subgroups within a data set.
- In this example, we are looking at the growth over time for each patient. The patient is the subgroup.
- Subgroups can be much more complicated, e.g., broken down by sex, race, age category, and region.

The one example for this whole talk

- Multiple patient visits over a few years
- Kids get their heights measured!
- How fast is each kid growing over the years?

Wrangling Issues

- Multiple files must be linked
- Dates must be converted
- Analyze subgroups of one big dataset
- Long or wide format?
- Missing data
- Unequal visit numbers

Related: the gospel of reproducible analysis

- In science we pay lip service to “reproducible experiments”
- Someone else should be able to reproduce your experiment by following the same steps
- Frequently this does not happen because it’s expensive, time-consuming, and difficult

Reproducible analysis on computers

- For computer-based analysis THERE IS NO EXCUSE! If you do it right your steps can be reproduced at the push of a button.
- Most shops do not yet insist on reproducibility (but it's catching on)
- If you do this right, reproducibility is simple and will save YOU lots of time and trouble later

REPL to routine

- Pull in id and one data set
- Try each step for one data set: join, get age, group
- After working out the right commands, make a single script [example: correct options for read.csv, e.g. ‘as.is’] that can do it with one command
- Convert that to a single script for all the data sets

The KISS Rule

- As a child, I thought it was “Keep It Simple, Stupid”
- My brilliant boss Tom Tang always said “Keep It Simple AND Stupid”
- I thought it was a mistake, since English was not his first language
- BUT AS USUAL, TOM WAS RIGHT!