Wrangling data: magrittr and tidyverse

Simon Farrell

9 July 2018

magrittr

magrittr

- Pipes (familiar to users of bash or hash)
- Passes output of one function as input to another
- fun_a %>% fun_b

magrittrExample.R

Le tidyverse

tidverse

Opinionated collection of R packages designed for data science

https://www.tidyverse.org

Packages

- **dplyr: running analyses on subsets**
- **tidyr: tidy up data**
- readr: helps reading in data consistently
- tibble: a tweak on data frames
- purr: haven't looked closely but it seems cool ¯_(ツ)_/¯
- ggplot2: pretty plotting using models as grammar

dplyr: apply for data frames

LM code

Exercise: see code

tidyr: data transformation and tidying

- See also the library reshape2, to which tidyr is an interface

Key functions

- Gather: Convert wide format to long format
- Spread: Convert long format to wide format
- Separate: Split a column into separate columns
- Unite: Join several columns into one

Look at code

dplyr also provides some useful functions for tidying data

- filter: return rows matching conditions
- select: exclude unwanted variables
- mutate: add a new variable
- join: join datasets
- arrange: sort data frame by one or more variables (see website at end of this section)

Together, tidyr and dplyr give you powerful tools to transform and summarise data

Useful for both data analysis and modelling (especially model fitting)

Further reading

https://www.tidyverse.org

http://rpubs.com/bradleyboehmke/data_wrangling

Advanced plotting with ggplot2

graphical grammar for plots

- Basic idea: data set is a structure, and different aspects of plot reflect different structures
- We build up (add to) a plot to add different features

summarise(simon)

- Rstudio projects are useful
- Rmarkdown is useful
- tidyverse is useful