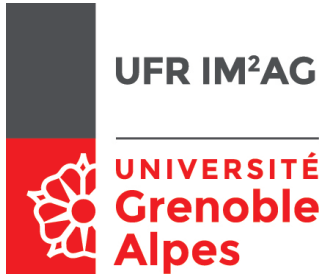
Data Manipulation

The dplyr way

Lucas Mello Schnorr, Jean-Marc Vincent

LIG/Inria – POLARIS

February 2017



Motivation

Institut national de la statistique et des études économiques

- First names given to newborns along years (*par départements français*)
- [https: //www.insee.fr/fr/statistiques/fichier/2540004/dpt2015_txt.zip][Link to =dpt2015_txt.zip=]] (12.24Mb, zipped – 85Mb pure text)
  - It has 3405311 rows (and one header line), 5 variables

Some questions that may arise

1. First name frequency evolves along time?
2. What can we say about " *Your name here* " (for each state, FR)?
3. Is there some sort of geographical correlation with the data?
4. Which state has a larger variety of names along time?

What would be your approach to tackle this?

- Need to manipulate data in a reproducible manner
- Leading to well elaborated plots for data interpretation

The `dplyr` R package (part of `tidyverse`)

Set of functions (called verbs) to perform common data manipulation

- Requirements: tidy data (columns are variables, rows are observations)
- With `magrittr` (the pipe operator `%>%`), it becomes a true workflow
  - Pipelining data manipulation

These are the basic verbs

- `select()`: select columns
- `filter()`: filter rows
- `arrange()`: reorder rows
- `mutate()`: create new columns
- `summarize()`: summarize values
- `group_by()`: group operations using *split-apply-combine*

Let's see them in action now → `TD5.Rmd`

References

Books/articles

- R for Data Science, by Garrett Grolemund and Hadley Wickham
  - Chapter 5 on Data transformation
- Tidy Data, by Hadley Wickham
  - See Section 2, or check directly the Table 3
- The Split-Apply-Combine Strategy for Data Analysis, by H Wickham
  - See Figures 4 and 7 (note that the paper uses an old version of dplyr)

Tutorials

- Introduction to dplyr 2016-06-23

Tools/packages

- magrittr
- dplyr