Data Manipulation (CMP595 PPGC/INF/UFRGS)

Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS

Porto Alegre, Brazil – October 2018

Motivation

Institut national de la statistique et des études économiques

- First names given to newborns (*par départements français par année*)
- Link to `dpt2015_txt.zip` (12.24Mb, zipped – 85Mb pure text)
  - It has 3405311 rows (and one header line), 5 variables

```
library(readr);
df <- read_tsv (file = "/tmp/dpt2015.txt",
                locale = locale(encoding = "ISO-8859-1"));
```

	sexe	preusuel	annais	dpt	nombre
1	2	MATHILDA	2009	33	5.00
2	2	ROSE-MARIE	1964	41	3.00
3	1	EDOUARD	1919	97	38.00
4	1	DIMITRI	1981	02	13.00
5	2	LINOA	2013	59	4.00
6	1	SÉBASTIEN	1953	97	16.00

Motivation $\rightarrow$ How to handle this amount of data?

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?

Motivation $\rightarrow$ How to handle this amount of data?

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

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

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

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

- Condition: 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 with `1_TD.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 DA, by Hadley 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