(A bit of) Advanced R

Part 3 - a tour of the `tidyverse`

Julien Chiquet

https://github.com/jchiquet/CourseAdvancedR

Université Paris Dauphine, Juin 2018

Outline

1. Introduction to tidy data and tidyverse

2. magrittr

3. tidyr

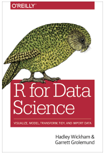
4. dplyr

5. tibble

References

Many ideas/examples inspired/stolen there:

R for data science (Wickham & Grolemund, 2016), http://r4ds.had.co.nz



Tidyverse website, https://www.tidyverse.org/

Prerequisites

Data Structures in base R

1. Atomic vector (integer, double, logical, character)
2. Recursive vector (list)
3. Factor
4. Matrix and array
5. Data Frame

R base programming

1. Control Statements
2. Functions
3. Functionals
4. Input/output
5. Rstudio API (application programming interface)

Outline

1. Introduction to tidy data and tidyverse

2. magrittr

3. tidyr

4. dplyr

5. tibble

Tidy data: motivation

Collected data are (never) under a proper canonical format

> *"Happy families are all alike; every unhappy family is unhappy in its own way." – Leo Tolstoy*

> *"Tidy datasets are all alike, but every messy dataset is messy in its own way." – Hadley Wickham*[1]

[1] Rstudio's chief scientific advisor

Tidy data: motivation

Collected data are (never) under a proper canonical format

> *"Happy families are all alike; every unhappy family is unhappy in its own way." – Leo Tolstoy*

> *"Tidy datasets are all alike, but every messy dataset is messy in its own way." – Hadley Wickham*[1]

[1] Rstudio's chief scientific advisor

Tidy data: what?

First, a subjective question

What is the *observation/statistical unit* in your data?

Definition

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

In tidy data,

1. each variable forms a column,
2. each observation forms a row,
3. each type of observational unit forms a table.

Tidy data: what?

First, a subjective question

What is the *observation/statistical unit* in your data?

Definition

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

In tidy data,

1. each variable forms a column,
2. each observation forms a row,
3. each type of observational unit forms a table.

Tidy data: what?

First, a subjective question

What is the *observation/statistical unit* in your data?

Definition

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

In tidy data,

1. each variable forms a column,
2. each observation forms a row,
3. each type of observational unit forms a table.

Tidy data: why?

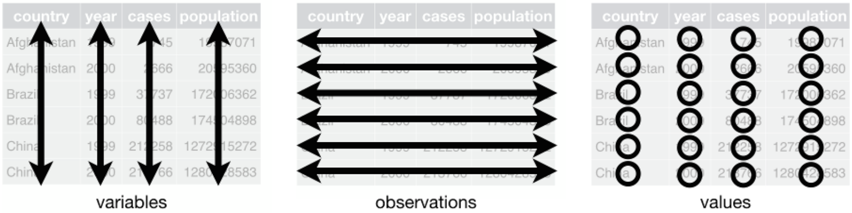


Figure 1: Tidy data

- make manipulation, visualization and modelling easier
- a common structure for all packages
- a philosophy for data representation (beyond the R framework)

Tidy or not ?

```r
tidyr::table3
```

```
## # A tibble: 6 x 3
##   country      year rate
## * <chr>       <int> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

Tidy or not ?

```r
tidyr::table2
```

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <int> <chr>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

Tidy or not ?

```
tidyr::table1
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

The process of data analysis

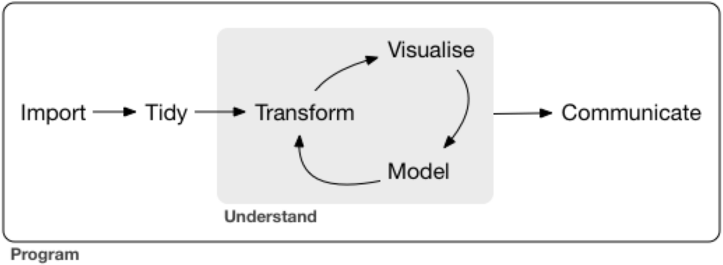


Figure 2: scheme for data analysis process

- **import:** read/load the data
- **tidy:** formating (individuals/variables data frame)
- **transform:** suppression/creation/filtering/selection
- **visualization:** representation and validation
- **model**: statistical fits
- **communication**: diffusion (web/talk/article)

The tidyverse

Definition

- contraction of ‘tidy’ (“well arranged) and ‘universe’.
- an *opinionated collection* of R packages designed for data science.
- all packages share an underlying *design philosophy*, *grammar*, and *data structures*

Phylosophy

> *allows the user to focus on the important statistical questions rather than focusing on the technical aspects of data analysis*

Let’s have a look

The core tidyverse loads ggplot2, tibble, tidyr, readr, purrr, stringr, forecats, dplyr and others in a fancy and unconflicted way.

```
library(tidyverse)
tidyverse:::tidyverse_conflicts()
```

```
## -- Conflicts ------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
tidyverse:::tidyverse_deps()
```

```
## # A tibble: 25 x 4
##    package cran  local behind
##    <chr>   <chr> <chr> <lgl>
##  1 broom   0.4.4 0.4.3 TRUE
##  2 cli     1.0.0 1.0.0 FALSE
##  3 crayon  1.3.4 1.3.4 FALSE
##  4 dbplyr  1.2.1 1.2.1 FALSE
##  5 dplyr   0.7.5 0.7.4 TRUE
##  6 forcats 0.3.0 0.3.0 FALSE
##  7 ggplot2 2.2.1 2.2.1 FALSE
##  8 haven   1.1.1 1.1.1 FALSE
##  9 hms     0.4.2 0.4.2 FALSE
## 10 httr    1.3.1 1.3.1 FALSE
## # ... with 15 more rows
```

Packages roles and overview I

tibble

a modern re-imagining of the data frame

tidyr

a set of functions that help you get to tidy data

dplyr

a consistent set of verbs that solve the most common data manipulation challenges

Packages roles and overview II

readr

a fast and friendly way to read rectangular data (like csv, tsv, and fwf)

stringr

a cohesive set of functions designed to make working with strings as easy as possible

forcats

a suite of useful tools that solve common problems with factors

Packages roles and overview III

ggplot2

a system for declaratively creating graphics, based on The Grammar of Graphics

purrr

enhances R's functional programming (FP) toolkit

magrittr

offers a set of operators which make your code more readable

Data analysis with the tidyverse

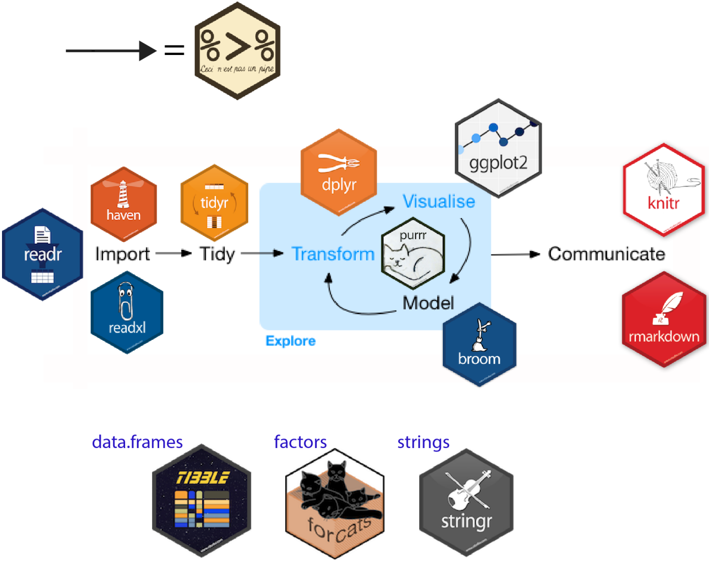


Figure 3: Updated scheme for data analysis process

Outline

1. Introduction to tidy data and tidyverse
2. magrittr
3. tidyr
4. dplyr
5. tibble

Outline

1. Introduction to tidy data and tidyverse

2. magrittr

3. tidyr

4. dplyr

5. tibble

Outline

1. Introduction to tidy data and tidyverse
2. magrittr
3. tidyr
4. dplyr
5. tibble

Outline

1. Introduction to tidy data and tidyverse

2. magrittr

3. tidyr

4. dplyr

5. tibble

References

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2014). *Advanced r*. CRC Press. Retrieved from http://adv-r.had.co.nz/

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc." Retrieved from http://r4ds.had.co.nz