Import data

Isaac Verminck

2018-04-06

Tidy data

At what age do people marry?

The first question we want to investigate is at what age people marry. This data is available in the *year_of_birth* tab of the *marriage* files.

We'd like to have a single dataframe indicating for each age how many people marry. However, when loading the data isn't coded in the way we want.

```r
select_worksheet(worksheet = "year_of_birth",
                 year = "2013",
                 type = "divorce")[1:20, ]
```

```
## # A tibble: 20 x 12
##    `Echtscheidingen~ X__1  X__2  X__3  X__4  X__5  X__6  X__7  X__8  X__9 
##    <chr>             <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 Leeftijd          Gebo~ Belg~ <NA>  Brus~ <NA>  Vlaa~ <NA>  Waal~ <NA> 
##  2 <NA>              <NA>  Eers~ Twee~ Eers~ Twee~ Eers~ Twee~ Eers~ Twee~
##  3 Totaal            Tota~ 24872 24872 5948  5948  11567 11567 7357  7357 
##  4 - dan 12 jaar     2001~ 0     0     0     0     0     0     0     0    
##  5 12 jaar           2001  0     0     0     0     0     0     0     0    
##  6 <NA>              2000  0     0     0     0     0     0     0     0    
##  7 13 jaar           2000  0     0     0     0     0     0     0     0    
##  8 <NA>              1999  0     0     0     0     0     0     0     0    
##  9 14 jaar           1999  0     0     0     0     0     0     0     0    
## 10 <NA>              1998  0     0     0     0     0     0     0     0    
## 11 15 jaar           1998  0     0     0     0     0     0     0     0    
## 12 <NA>              1997  0     0     0     0     0     0     0     0    
## 13 16 jaar           1997  0     0     0     0     0     0     0     0    
## 14 <NA>              1996  0     0     0     0     0     0     0     0    
## 15 17 jaar           1996  0     0     0     0     0     0     0     0    
## 16 <NA>              1995  0     0     0     0     0     0     0     0    
## 17 18 jaar           1995  0     0     0     0     0     0     0     0    
## 18 <NA>              1994  0     1     0     1     0     0     0     0    
## 19 19 jaar           1994  0     1     0     0     0     0     0     1    
## 20 <NA>              1993  0     6     0     2     0     4     0     0    
## # ... with 2 more variables: X__10 <chr>, X__11 <chr>
```

There are several issues:

- the name of the analysis is repeated as header
- there's a total row at the top
- at the bottom the source of the data is mentioned
- we have the data by year while we want a single dataframe
- regions are separate variables while we want a single variable region with the regions as values
- distinction is made between first and second spouses while we don't care about this distinction
- people with the same ages can be born in two different years

- the year in which the data were gathered isn't mentioned explicitly
- the type of data (divorce or marriage) isn't mentioned explicitly
- variables are coded as character instead of factors (e.g. region)

We create a function to solve each issue, apply these to each dataframe and then bind all dataframes for each year together. Small note: additional data on gender is provided from 2015 on. Making a modification to incorporate this hasn't been done yet so for the moment the focus is on the data 2013-2014.

```r
types <- c("divorce","marriage")
years <- paste0("201", 3:4)

rep_years <- rep(years, times = length(types))
rep_types <- rep(types, each = length(years))

worksheets <- map2(.x = rep_years,
                   .y = rep_types,
                   .f = ~select_worksheet(worksheet = "year_of_birth",
                                          year = .x,
                                          type = .y)
                   )

tidy_args <- list(worksheet = worksheets,
                  year = rep_years,
                  type = rep_types)
tidy_dfs <- pmap(tidy_args,
                 function(worksheet, year, type) tidy_year_of_birth(worksheet, year, type)
                 )

year_of_birth <- bind_together(tidy_dfs)

save(year_of_birth,
     file = paste(data_path, "year_of_birth.RData", sep = "/")
     ) # data_path is object in tidy-year_of_birth.R
```