Transformasi Data

Muhammad Aswan Syahputra

4/9/2019

## Non-tidy menjadi Tidy dataset

Anda akan menggunakan fungsi spread dari paket tidyr untuk mengubah memperbaiki dataset ‘table2’ (juga dari paket tidyr). Aktifkanlah paket tidyr, lihat dataset ‘table2’. Apakah yang membuat dataset tersebut *non-tidy* data? Bukalah dokumentasi fungsi spread dengan menjalankan ?nama\_fungsi atau help(nama\_fungsi)!

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.5.3

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

Dataset ‘table2’ dapat diperbaiki dengan menjalankan kode berikut:

table2\_tidy <- spread(table2, key = "type", value = "count")  
t

## function (x)   
## UseMethod("t")  
## <bytecode: 0x0000000019196920>  
## <environment: namespace:base>

Selanjutnya Anda juga akan memperbaiki dataset ‘table4a’. Cetaklah dataset tersebut dan dapatkah Anda menyebutkan alasan mengapa dataset tersebut tidak *tidy* dan *tame*?

table4a # cetak dataset table4a

## # A tibble: 3 x 3  
## country `1999` `2000`  
## \* <chr> <int> <int>  
## 1 Afghanistan 745 2666  
## 2 Brazil 37737 80488  
## 3 China 212258 213766

Dataset ‘table4a’ dapat diperbaiki dengan menggunakan fungsi gather dari tidyr. Anda dapat mempelajari fungsi tersebut dengan menjalankan ?gather.

gather(table4a, key = "year", value = "case", 2:3)

## # A tibble: 6 x 3  
## country year case  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 745  
## 2 Brazil 1999 37737  
## 3 China 1999 212258  
## 4 Afghanistan 2000 2666  
## 5 Brazil 2000 80488  
## 6 China 2000 213766

table4a\_tidy <- table4a %>%   
 gather(key = "year", value = "cases", 2:3) # menggunakan tidyverse syntax, pipe %>%

Silakan lakukan hal serupa pada dataset ‘table4b’ namun dengan menggunakan “population” sebagai isian argumen value. Tuliskan juga dengan menggunakan *tidyverse syntax* dan simpan obyek tersebut dengan nama ‘table4b\_tidy’!

table4b

## # A tibble: 3 x 3  
## country `1999` `2000`  
## \* <chr> <int> <int>  
## 1 Afghanistan 19987071 20595360  
## 2 Brazil 172006362 174504898  
## 3 China 1272915272 1280428583

table4b\_tidy <- table4a %>%   
 gather(key="year", value="population", c(2:3))  
table4b\_tidy

## # A tibble: 6 x 3  
## country year population  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 745  
## 2 Brazil 1999 37737  
## 3 China 1999 212258  
## 4 Afghanistan 2000 2666  
## 5 Brazil 2000 80488  
## 6 China 2000 213766

Dataset ‘table4a\_tidy’ dan ‘table4b\_tidy’ tersebut dapat digabungkan menjadi satu dataset. Hal tersebut dapat dilakukan dengan menggunakan fungsi left\_join dari paket dplyr seperti contoh berikut:

library(dplyr) # mengaktifkan paket dplyr

## Warning: package 'dplyr' was built under R version 3.5.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

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

table4a\_tidy

## # A tibble: 6 x 3  
## country year cases  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 745  
## 2 Brazil 1999 37737  
## 3 China 1999 212258  
## 4 Afghanistan 2000 2666  
## 5 Brazil 2000 80488  
## 6 China 2000 213766

table4b\_tidy

## # A tibble: 6 x 3  
## country year population  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 745  
## 2 Brazil 1999 37737  
## 3 China 1999 212258  
## 4 Afghanistan 2000 2666  
## 5 Brazil 2000 80488  
## 6 China 2000 213766

mydata <- left\_join(table4a\_tidy, table4b\_tidy, by = c("country", "year"))  
mydata

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

mydata

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

## Data Wrangling

Dataset mydata tersebut merupakan subset dataset Tubercolusis yang diolah dari data ‘who’ dan ‘population’ (dari paket tidyr). Lihatlah ringkasan kedua tersebut dengan menggunakan glimpse!

glimpse(who)

## Observations: 7,240  
## Variables: 60  
## $ country <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afg...  
## $ iso2 <chr> "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF", "...  
## $ iso3 <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", ...  
## $ year <int> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1...  
## $ new\_sp\_m014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_m65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sp\_f65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_m65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_sn\_f65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_m65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ new\_ep\_f65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_m65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f014 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ newrel\_f65 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...

glimpse(population)

## Observations: 4,060  
## Variables: 3  
## $ country <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afgha...  
## $ year <int> 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 200...  
## $ population <int> 17586073, 18415307, 19021226, 19496836, 19987071, 2...

summary(population)

## country year population   
## Length:4060 Min. :1995 Min. :1.129e+03   
## Class :character 1st Qu.:1999 1st Qu.:6.029e+05   
## Mode :character Median :2004 Median :5.319e+06   
## Mean :2004 Mean :3.003e+07   
## 3rd Qu.:2009 3rd Qu.:1.855e+07   
## Max. :2013 Max. :1.386e+09

population%>%  
 distinct(country)

## # A tibble: 219 x 1  
## country   
## <chr>   
## 1 Afghanistan   
## 2 Albania   
## 3 Algeria   
## 4 American Samoa   
## 5 Andorra   
## 6 Angola   
## 7 Anguilla   
## 8 Antigua and Barbuda  
## 9 Argentina   
## 10 Armenia   
## # ... with 209 more rows

Sekarang kita akan membuat versi utuh dari dataset ‘mydata’ dengan menggunakan data seluruh negara pada dataset ‘who’ sebagai berikut:

# Menjalankan fungsi satu per satu  
(tb1 <- gather(who, key = "key", value = "case", new\_sp\_m014:newrel\_f65))

## # A tibble: 405,440 x 6  
## country iso2 iso3 year key case  
## <chr> <chr> <chr> <int> <chr> <int>  
## 1 Afghanistan AF AFG 1980 new\_sp\_m014 NA  
## 2 Afghanistan AF AFG 1981 new\_sp\_m014 NA  
## 3 Afghanistan AF AFG 1982 new\_sp\_m014 NA  
## 4 Afghanistan AF AFG 1983 new\_sp\_m014 NA  
## 5 Afghanistan AF AFG 1984 new\_sp\_m014 NA  
## 6 Afghanistan AF AFG 1985 new\_sp\_m014 NA  
## 7 Afghanistan AF AFG 1986 new\_sp\_m014 NA  
## 8 Afghanistan AF AFG 1987 new\_sp\_m014 NA  
## 9 Afghanistan AF AFG 1988 new\_sp\_m014 NA  
## 10 Afghanistan AF AFG 1989 new\_sp\_m014 NA  
## # ... with 405,430 more rows

tb2 <- select(tb1, country, year, case)  
 View(tb2)  
(tb3 <- group\_by(tb2, country, year))

## # A tibble: 405,440 x 3  
## # Groups: country, year [7,240]  
## country year case  
## <chr> <int> <int>  
## 1 Afghanistan 1980 NA  
## 2 Afghanistan 1981 NA  
## 3 Afghanistan 1982 NA  
## 4 Afghanistan 1983 NA  
## 5 Afghanistan 1984 NA  
## 6 Afghanistan 1985 NA  
## 7 Afghanistan 1986 NA  
## 8 Afghanistan 1987 NA  
## 9 Afghanistan 1988 NA  
## 10 Afghanistan 1989 NA  
## # ... with 405,430 more rows

(tb4 <- summarise(tb3, cases = sum(case, na.rm = TRUE)))

## # A tibble: 7,240 x 3  
## # Groups: country [219]  
## country year cases  
## <chr> <int> <int>  
## 1 Afghanistan 1980 0  
## 2 Afghanistan 1981 0  
## 3 Afghanistan 1982 0  
## 4 Afghanistan 1983 0  
## 5 Afghanistan 1984 0  
## 6 Afghanistan 1985 0  
## 7 Afghanistan 1986 0  
## 8 Afghanistan 1987 0  
## 9 Afghanistan 1988 0  
## 10 Afghanistan 1989 0  
## # ... with 7,230 more rows

View(tb4)  
(tb5 <- ungroup(tb4))

## # A tibble: 7,240 x 3  
## country year cases  
## <chr> <int> <int>  
## 1 Afghanistan 1980 0  
## 2 Afghanistan 1981 0  
## 3 Afghanistan 1982 0  
## 4 Afghanistan 1983 0  
## 5 Afghanistan 1984 0  
## 6 Afghanistan 1985 0  
## 7 Afghanistan 1986 0  
## 8 Afghanistan 1987 0  
## 9 Afghanistan 1988 0  
## 10 Afghanistan 1989 0  
## # ... with 7,230 more rows

(tb6 <- left\_join(tb5, population))

## Joining, by = c("country", "year")

## # A tibble: 7,240 x 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1980 0 NA  
## 2 Afghanistan 1981 0 NA  
## 3 Afghanistan 1982 0 NA  
## 4 Afghanistan 1983 0 NA  
## 5 Afghanistan 1984 0 NA  
## 6 Afghanistan 1985 0 NA  
## 7 Afghanistan 1986 0 NA  
## 8 Afghanistan 1987 0 NA  
## 9 Afghanistan 1988 0 NA  
## 10 Afghanistan 1989 0 NA  
## # ... with 7,230 more rows

View(tb6)  
(tb7 <- filter(tb6, !is.na(population)))

## # A tibble: 4,037 x 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1995 0 17586073  
## 2 Afghanistan 1996 0 18415307  
## 3 Afghanistan 1997 128 19021226  
## 4 Afghanistan 1998 1778 19496836  
## 5 Afghanistan 1999 745 19987071  
## 6 Afghanistan 2000 2666 20595360  
## 7 Afghanistan 2001 4639 21347782  
## 8 Afghanistan 2002 6509 22202806  
## 9 Afghanistan 2003 6528 23116142  
## 10 Afghanistan 2004 8245 24018682  
## # ... with 4,027 more rows

(tb8 <- mutate(tb7, proportion = 100\*cases/population))

## # A tibble: 4,037 x 5  
## country year cases population proportion  
## <chr> <int> <int> <int> <dbl>  
## 1 Afghanistan 1995 0 17586073 0   
## 2 Afghanistan 1996 0 18415307 0   
## 3 Afghanistan 1997 128 19021226 0.000673  
## 4 Afghanistan 1998 1778 19496836 0.00912   
## 5 Afghanistan 1999 745 19987071 0.00373   
## 6 Afghanistan 2000 2666 20595360 0.0129   
## 7 Afghanistan 2001 4639 21347782 0.0217   
## 8 Afghanistan 2002 6509 22202806 0.0293   
## 9 Afghanistan 2003 6528 23116142 0.0282   
## 10 Afghanistan 2004 8245 24018682 0.0343   
## # ... with 4,027 more rows

View(tb8)  
  
# Syntax menggunakan pipe %>%  
  
tb\_all <-   
 who %>%   
 gather(key = "key", value = "case", new\_sp\_m014:newrel\_f65) %>%   
 select(country, year, case) %>%   
 group\_by(country, year) %>%   
 summarise(cases = sum(case, na.rm = TRUE)) %>%   
 ungroup() %>%   
 left\_join(population) %>%   
 filter(!is.na(population)) %>%   
 mutate(proportion = 100\*cases/population)

## Joining, by = c("country", "year")

Dapatkah Anda membuat ringkasan apa saja hal apa saja yang dilakukan pada proses *data wrangling* diatas? (Petunjuk: ?nama\_fungsi)

1. …
2. …
3. …
4. …
5. …
6. …
7. …
8. …

Cek apakah dataset ‘tb8’ sama dengan dataset ‘tb\_all’! Menurut Anda, cara penulisan *syntax* manakah yang lebih mudah digunakan dan dipahami?

identical(tb8, tb\_all)

## [1] TRUE