Transformasi Data

Muhammad Aswan Syahputra

4/9/2019

Contents

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")
```

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 <- table4b %>%
  gather(key = "year", value = "cases", 2:3) # menggunakan tidyverse syntax, pipe %>%
table4b_tidy
```

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

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
```

```
mydata <- left_join(table4a_tidy, table4b_tidy)
```

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

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...
```

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

```r
# Menjalankan fungsi satu per satu
tb1 <- gather(who, key = "key", value = "case", new_sp_m014:newrel_f65)
tb2 <- select(tb1, country, year, case)
tb3 <- group_by(tb2, country, year)
tb4 <- summarise(tb3, cases = sum(case, na.rm = TRUE))
tb5 <- ungroup(tb4)
tb6 <- left_join(tb5, population)
```

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

```r
tb7 <- filter(tb6, !is.na(population))
tb8 <- mutate(tb7, proportion = 100*cases/population)

# 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?

```r
tb8
```

```
## # 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
```

```
tb_all
```

```
## # 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
```