tidyr

LearningSpoonsR

2018-05-20

beamer@RMarkdown

- RMarkdown
- pdf 조판을 위한 texlive 엔진
- slide형태의 pdf를 만드는 beamer 패키지 (R 패키지가 아니라 tex 패키지)
- 한글 및 twocolumn layout을 위한 `latex-topmatter.tex` (베포해드리는 `rmd-beamer.Rmd` 템플릿의 하위 폴더에 있습니다.)

0. Let's start!

```
source("infile-tidyr.R")
library(tidyverse) # Wickham's
library(sqldf)
```

- Part I. Join
- Part II. Tidy data

Part I. Join

```
df1
```

```
##   CustomerId Product
## 1          1 Toaster
## 2          2 Toaster
## 3          3 Toaster
## 4          4   Radio
## 5          5   Radio
```

```
df2
```

```
##   CustomerId State
## 1          2 Seoul
## 2          4 Seoul
## 3          6 Busan
```

- 4 types of join

I-1. Inner Join

```
inner_join(df1, df2)
merge(x = df1, y = df2, by = "CustomerId")
sqldf("SELECT CustomerId, Product, State
       FROM df1 JOIN df2 USING(CustomerID)")
```

```
## Joining, by = "CustomerId"
##   CustomerId Product State
## 1          2 Toaster Seoul
## 2          4   Radio Seoul
```

I-2. Left Join

```
left_join(df1, df2)
merge(x = df1, y = df2, by = "CustomerId", all.x = TRUE)
sqldf("SELECT CustomerId, Product, State
       FROM df1 LEFT JOIN df2 USING(CustomerID)")
```

```
## Joining, by = "CustomerId"
##   CustomerId Product State
## 1          1 Toaster  <NA>
## 2          2 Toaster Seoul
## 3          3 Toaster  <NA>
## 4          4   Radio Seoul
## 5          5   Radio  <NA>
```

I-3. Outer Join (full)

```r
full_join(df1, df2)
merge(x = df1, y = df2, by = "CustomerId", all = TRUE)
```

```
## Joining, by = "CustomerId"
##   CustomerId Product State
## 1          1 Toaster  <NA>
## 2          2 Toaster Seoul
## 3          3 Toaster  <NA>
## 4          4   Radio Seoul
## 5          5   Radio  <NA>
## 6          6    <NA> Busan
```

I-4. Right Join

```
right_join(df1, df2)
merge(x = df1, y = df2, by = "CustomerId", all.y = TRUE)
```

```
## Joining, by = "CustomerId"
##   CustomerId Product State
## 1          2 Toaster Seoul
## 2          4   Radio Seoul
## 3          6    <NA> Busan
```

Summary

- Summary

```
inner_join(df1, df2)
left_join(df1, df2)
full_join(df1, df2)
right_join(df1, df2)
```

- Variations

```
inner_join(df1, df2)
inner_join(x=df1, y=df2)
inner_join(x=df1, y=df2, by = "CustomerId")
inner_join(x=df1, y=df2, by = c("CustomerId"))
inner_join(x=df1, y=df2, by = c("CustomerId"="CustomerId"))
```

blank

Part II. Tidy data

```
table1
```

```
##   ISO3 year  cases      popul
## 1  AFG 1999    745   19987071
## 2  AFG 2000   2666  201595360
## 3  BRA 1999  37737  172006362
## 4  BRA 2000  80488  174504898
## 5  CHN 1999 212258 1272915272
## 6  CHN 2000 213766 1280428583
```

II-0. Short Review (mutate)

```
table1
```

```
##   ISO3 year  cases      popul
## 1  AFG 1999    745   19987071
## 2  AFG 2000   2666  201595360
## 3  BRA 1999  37737  172006362
## 4  BRA 2000  80488  174504898
## 5  CHN 1999 212258 1272915272
## 6  CHN 2000 213766 1280428583
```

```
table1 %>% mutate(rate = cases / popul * 100)
```

```
##   ISO3 year  cases      popul        rate
## 1  AFG 1999    745   19987071 0.003727410
## 2  AFG 2000   2666  201595360 0.001322451
## 3  BRA 1999  37737  172006362 0.021939305
## 4  BRA 2000  80488  174504898 0.046123634
## 5  CHN 1999 212258 1272915272 0.016674951
## 6  CHN 2000 213766 1280428583 0.016694879
```

II-0. Short Review (group_by & summarise)

```
table1
```

```
##   ISO3 year  cases      popul
## 1  AFG 1999    745   19987071
## 2  AFG 2000   2666  201595360
## 3  BRA 1999  37737  172006362
## 4  BRA 2000  80488  174504898
## 5  CHN 1999 212258 1272915272
## 6  CHN 2000 213766 1280428583
```

```
table1 %>% group_by(year) %>% summarise(n = sum(cases))
table1 %>% count(year, wt = cases) # equivalent to above
```

```
## # A tibble: 2 x 2
##    year      n
##   <dbl>  <dbl>
## 1  1999 250740
## 2  2000 296920
```

blank

II-1. gather from table4a & table4b

```
table4a
```

```
##   ISO3   1999   2000
## 1  AFG    745   2666
## 2  BRA  37737  80488
## 3  CHN 212258 213766
```

```
tidy4a <- table4a %>%
  gather(colnames(table4a)[-1],
         key = "year",
         value = "cases")
tidy4a
```

```
##   ISO3 year  cases
## 1  AFG 1999    745
## 2  BRA 1999  37737
## 3  CHN 1999 212258
## 4  AFG 2000   2666
## 5  BRA 2000  80488
## 6  CHN 2000 213766
```

```
table4b
```

```
##   ISO3       1999       2000
## 1  AFG   19987071  201595360
## 2  BRA  172006362  174504898
## 3  CHN 1272915272 1280428583
```

```
tidy4b <- table4b %>%
  gather(colnames(table4b)[-1],
         key = "year",
         value = "popul")
tidy4b
```

```
##   ISO3 year      popul
## 1  AFG 1999   19987071
## 2  BRA 1999  172006362
## 3  CHN 1999 1272915272
## 4  AFG 2000  201595360
## 5  BRA 2000  174504898
## 6  CHN 2000 1280428583
```

II-1. gather from table4a & table4b

```
left_join(tidy4a, tidy4b)
left_join(tidy4a, tidy4b, by = c("ISO3", "year"))
left_join(tidy4a, tidy4b, by = c("ISO3"="ISO3", "year"="year"))
```

```
## Joining, by = c("ISO3", "year")
##   ISO3 year  cases      popul
## 1  AFG 1999    745   19987071
## 2  BRA 1999  37737  172006362
## 3  CHN 1999 212258 1272915272
## 4  AFG 2000   2666  201595360
## 5  BRA 2000  80488  174504898
## 6  CHN 2000 213766 1280428583
```

II-2. spread from table2

```
table2
```

```
##   ISO3 year  type      count
## 1  AFG 1999 cases       745
## 2  AFG 1999 popul  19987071
## 3  AFG 2000 cases      2666
## 4  AFG 2000 popul 201595360
## 5  BRA 1999 cases     37737
## 6  BRA 1999 popul 172006362
```

```
table2 %>% spread(key = "type", value = "count")
```

```
##   ISO3 year cases     popul
## 1  AFG 1999   745  19987071
## 2  AFG 2000  2666 201595360
## 3  BRA 1999 37737 172006362
```

II-3. separate from table3

```
table3
```

```
##   ISO3 year              rate
## 1  AFG 1999       745/19987071
## 2  AFG 2000     2666/201595360
## 3  BRA 1999    37737/172006362
## 4  BRA 2000    80488/174504898
## 5  CHN 1999 212258/1272915272
## 6  CHN 2000 213766/1280428583
```

```
table3 %>% separate(rate, into = c("cases", "popul"), sep = "/")
```

```
##   ISO3 year  cases      popul
## 1  AFG 1999    745   19987071
## 2  AFG 2000   2666  201595360
## 3  BRA 1999  37737  172006362
## 4  BRA 2000  80488  174504898
## 5  CHN 1999 212258 1272915272
## 6  CHN 2000 213766 1280428583
```

Classic method

```
table3$cases <-
  sapply(strsplit(table3$rate, split = "/"),
         function(x) x[1])
table3$popul <-
  sapply(strsplit(table3$rate, split = "/"),
         function(x) x[2])
```

blank

blank