Biên tập dữ liệu với R

Tuan V. Nguyen
Garvan Institute of Medical Research
University of New South Wales (UNSW Sydney), Australia
University of Technology, Sydney (UTS), Australia
Ton Duc Thang University, Vietnam

© Tuan V. Nguyen

Biên tập dữ liệu với R

- Làm quen với $
- Mã hóa (coding)
- Hoán chuyển (transformation)
- Chọn dữ liệu
- Biên tập dữ liệu với **`tidyverse`**

Biên tập dữ liệu
với hàm cơ bản trong R

Dấu "$"

- Rất quan trọng!
- $ nối kết dataset và biến số (dataframe và variable)

```
dat$var1
```

```
df$var1
```

- có nghĩa là biến "`var1`" thuộc dataset "`dat`", và `var1` thuộc dataset `df`

Dấu "$"

```
bw = read.csv("birthwt.csv")

head(bw, 3)
  id low age lwt race smoke ptl ht ui ftv  bwt
1 85   0  19 182    2     0   0  0  1   0 2523
2 86   0  33 155    3     0   0  0  0   3 2551
3 87   0  20 105    1     1   0  0  0   1 2557
```

```
> weight = lwt*0.453592
Error: object 'lwt' not found
> bw$weight = bw$lwt*0.453592

> head(bw, 3)
  id low age lwt race smoke ptl ht ui ftv  bwt   weight
1 85   0  19 182    2     0   0  0  1   0 2523 82.55374
2 86   0  33 155    3     0   0  0  0   3 2551 70.30676
3 87   0  20 105    1     1   0  0  0   1 2557 47.62716
```

Mã hoá (coding)

Chúng ta muốn tạo ra một biến mới "**lowbw**" mã hóa từ biến **low**. Nếu low=1 thì lowbw="Yes"; nếu low=0 thì lowbw="No"

```
bw$lowbw[low=1] <- "Yes"
bw$lowbw[low=0] <- "No"
```

```
> head(bw, 3)
  id low age lwt race smoke ptl ht ui ftv  bwt   weight lowbw
1 85   0  19 182    2     0   0  0  1   0 2523 82.55374   Yes
2 86   0  33 155    3     0   0  0  0   3 2551 70.30676   Yes
3 87   0  20 105    1     1   0  0  0   1 2557 47.62716   Yes
```

ifelse

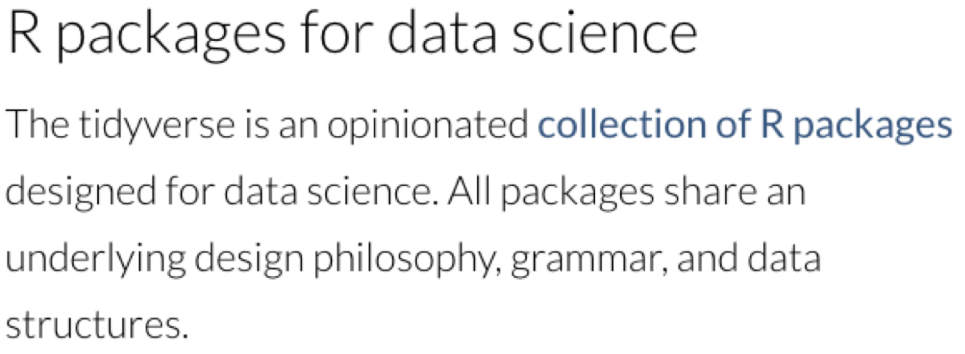
Chúng ta muốn tạo ra một biến mới "**smoker**". Nếu smoke=1 thì smoker=Yes, tất cả các giá trị khác thì smoker=No

```
bw$smoker = ifelse(bw$smoke==1, 1, 0)
```

Biên tập dữ liệu với "tidyverse"

Package `tidyverse`

- Hadley Wickham phát triển
- Tổng hợp từ `dplyr`, `ggplot2`
- Dùng cho khoa học dữ liệu
- Rất tiện cho quản lí dữ liệu phức tạp
- Chuẩn bị dữ liệu cho phân tích



https://www.tidyverse.org

Những hàm chính trong tidyverse/dplyr

5 'động từ' chính:

- `select()` chọn những cột/field liên quan
- `filter()` chọn những dòng quan tâm
- `mutate()` thêm cột/field
- `arrange()` thứ tự hóa dòng dữ liệu
- `summarize()` tóm tắt dữ liệu theo dòng

%>% (pipe) operator

"Văn phạm" chính của tidyverse

select(dataframe, conditions)

filter(dataframe, conditions)

mutate(dataframe, conditions)

arrange(dataframe, conditions)

hoặc

dataframe %>% **hàm**(conditions)

Dữ liệu thực hành: cân nặng trẻ sơ sinh

```
bw = read.csv("birthwt.csv")

head(bw)
```

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	85	0	19	182	2	0	0	0	1	0	2523
2	86	0	33	155	3	0	0	0	0	3	2551
3	87	0	20	105	1	1	0	0	0	1	2557
4	88	0	21	108	1	1	0	0	1	2	2594
5	89	0	18	107	1	1	0	0	1	0	2600
6	91	0	21	124	3	0	0	0	0	0	2622

select() – chọn những cột liên quan

full data (bw) 189 dòng, 11 cột)

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	85	0	19	182	2	0	0	0	1	0	2523
2	86	0	33	155	3	0	0	0	0	3	2551
3	87	0	20	105	1	1	0	0	0	1	2557
4	88	0	21	108	1	1	0	0	1	2	2594
5	89	0	18	107	1	1	0	0	1	0	2600
6	91	0	21	124	3	0	0	0	0	0	2622

```
temp = select(bw, c("low", "bwt", "lwt", "age"))
temp = bw %>% select(c("low", "bwt", "lwt", "age"))
head(temp)
```

```
> head(temp)
  low  bwt lwt age
1   0 2523 182  19
2   0 2551 155  33
3   0 2557 105  20
4   0 2594 108  21
5   0 2600 107  18
6   0 2622 124  21
```

filter() – chọn những dòng liên quan

```
temp = bw %>% filter(race==1, bwt<2500)
temp = filter(bw, race==1, bwt<2500)
head(temp)
```

```
> temp
    id low age lwt race smoke ptl ht ui ftv  bwt
1   10   1  29 130    1     0   0  0  1   2 1021
2   20   1  21 165    1     1   0  1  0   1 1790
3   22   1  32 105    1     1   0  0  0   0 1818
4   23   1  19  91    1     1   2  0  1   0 1885
5   26   1  25  92    1     1   0  0  0   0 1928
6   27   1  20 150    1     1   0  0  0   2 1928
7   29   1  24 155    1     1   1  0  0   0 1936
8   33   1  19 102    1     0   0  0  0   2 2082
9   34   1  19 112    1     1   0  0  1   0 2084
10  35   1  26 117    1     1   1  0  0   0 2084
11  36   1  24 138    1     0   0  0  0   0 2100
12  42   1  22 130    1     1   1  0  1   1 2187
13  45   1  17 110    1     1   0  0  0   0 2225
14  51   1  20 121    1     1   1  0  1   0 2296
15  56   1  31 102    1     1   1  0  0   1 2353
16  57   1  15 110    1     0   0  0  0   0 2353
17  65   1  30 142    1     1   1  0  0   0 2410
18  67   1  22 130    1     1   0  0  0   1 2410
19  68   1  17 120    1     1   0  0  0   3 2414
20  69   1  23 110    1     1   1  0  0   0 2424
21  77   1  26 190    1     1   0  0  0   0 2466
22  79   1  28  95    1     1   0  0  0   2 2466
23  84   1  21 130    1     1   0  1  0   3 2495
```

mutate - tạo ra biến số mới

```
temp = bw %>% mutate(mother.wt=lwt*0.453592, weight=bwt/1000)
temp = mutate(bw, mother.wt=lwt*0.453592, weight=bwt/1000)
head(temp)
```

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt	mother.wt	weight
1	85	0	19	182	2	0	0	0	1	0	2523	82.55374	2.523
2	86	0	33	155	3	0	0	0	0	3	2551	70.30676	2.551
3	87	0	20	105	1	1	0	0	0	1	2557	47.62716	2.557
4	88	0	21	108	1	1	0	0	1	2	2594	48.98794	2.594
5	89	0	18	107	1	1	0	0	1	0	2600	48.53434	2.600
6	91	0	21	124	3	0	0	0	0	0	2622	56.24541	2.622

arrange – sort dữ liệu theo dòng

```
arrange(temp, mother.wt, weight)
```

```
    id low age lwt race smoke ptl ht ui ftv  bwt mother.wt weight
1   44   1  20  80    3     1   0  0  1   0 2211  36.28736  2.211
2   15   1  25  85    3     0   0  0  1   0 1474  38.55532  1.474
3  137   0  22  85    3     1   0  0  0   0 3090  38.55532  3.090
4   32   1  25  89    3     0   2  0  0   1 2055  40.36969  2.055
5  118   0  24  90    1     1   1  0  0   1 2948  40.82328  2.948
6  132   0  18  90    1     1   0  0  1   0 3062  40.82328  3.062
....
```

Hàm "sample" có thể dùng để chọn mẫu ngẫu nhiên

```
d5 = sample_n(bw, 10)   # Chọn ngẫu nhiên 10 đối tượng từ bw
d5
```

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
57	144	0	21	110	3	1	0	0	1	0	3203
159	43	1	27	130	2	0	0	0	1	0	2187
124	220	0	22	129	1	0	0	0	0	0	4111
93	187	0	19	235	1	1	0	1	0	0	3629
48	135	0	19	132	3	0	0	0	0	0	3090
73	164	0	23	115	3	1	0	0	0	1	3331
17	102	0	15	98	2	0	0	0	0	0	2778
154	35	1	26	117	1	1	1	0	0	0	2084
30	116	0	17	113	2	0	0	0	0	1	2920
79	172	0	20	121	2	1	0	0	0	0	3444

Hàm "sample_frac()" có thể dùng để chọn mẫu ngẫu nhiên

```
d6 = sample_frac(bw, 0.05) # Chọn ngẫu nhiên 5% đối tượng từ bw
d6
```

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
127	223	0	35	170	1	0	1	0	0	1	4174
125	221	0	25	130	1	0	0	0	0	2	4153
141	22	1	32	105	1	1	0	0	0	0	1818
146	27	1	20	150	1	1	0	0	0	2	1928
31	117	0	17	113	2	0	0	0	0	1	2920
48	135	0	19	132	3	0	0	0	0	0	3090
128	224	0	19	120	1	1	0	0	0	0	4238
116	212	0	28	134	3	0	0	0	0	1	3941
97	191	0	29	154	1	0	0	0	0	1	3651

Tóm tắt tidyverse

- `select()` chọn những cột/field liên quan
- `filter()` chọn những dòng quan tâm
- `mutate()` thêm cột/field
- `arrange()` thứ tự hóa dòng dữ liệu
- `sample()` lấy mẫu ngẫu nhiên