Sử dụng table1

Bước 1:

Thử đọc lại dữ liệu trực tiếp từ Internet và xem qua dữ liệu với hàm `glimpse(…)` trong thư viện `dplyr`:

```
library(dplyr)
df = read.csv('https://thachln.github.io/datasets/bank/bank-additional-full.csv', sep=';')


glimpse(df)
```

```
Rows: 41,188
Columns: 21
$ age            <int> 56, 57, 37, 40, 56, 45, 59, 41, 24, 25, 41, 25, 29,...
$ job            <chr> "housemaid", "services", "services", "admin.", "ser...
$ marital        <chr> "married", "married", "married", "married", "marrie...
$ education      <chr> "basic.4y", "high.school", "high.school", "basic.6y...
$ default        <chr> "no", "unknown", "no", "no", "no", "unknown", "no",...
$ housing        <chr> "no", "no", "yes", "no", "no", "no", "no", "no", "y...
$ loan           <chr> "no", "no", "no", "no", "yes", "no", "no", "no", "n...
$ contact        <chr> "telephone", "telephone", "telephone", "telephone",...
$ month          <chr> "may", "may", "may", "may", "may", "may", "may", "m...
$ day_of_week    <chr> "mon", "mon", "mon", "mon", "mon", "mon", "mon", "m...
$ duration       <int> 261, 149, 226, 151, 307, 198, 139, 217, 380, 50, 55...
$ campaign       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ pdays          <int> 999, 999, 999, 999, 999, 999, 999, 999, 999, 999, 9...
$ previous       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ poutcome       <chr> "nonexistent", "nonexistent", "nonexistent", "nonex...
$ emp.var.rate   <dbl> 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1...
$ cons.price.idx <dbl> 93.994, 93.994, 93.994, 93.994, 93.994, 93.994, 93....
$ cons.conf.idx  <dbl> -36.4, -36.4, -36.4, -36.4, -36.4, -36.4, -36.4, -3...
$ euribor3m      <dbl> 4.857, 4.857, 4.857, 4.857, 4.857, 4.857, 4.857, 4....
$ nr.employed    <dbl> 5191, 5191, 5191, 5191, 5191, 5191, 5191, 5191, 519...
$ y              <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no...
```

Bước 2:

Cài đặt thư viện `table1`:

```
install.packages('table1')
```

Thử phân tích vài số liệu bằng cách gọi hàm table1 với cú pháp:

- Tham số đầu tiên bắt đầu bằng dấu ~ (đọc là till). Tiếp theo là các biến cần phân tích được ghép với nhau bằng dấu +
- Tham số thứ hai `data=df` cho biết dữ liệu cần phân tích lấy từ biết `df`.

```
library(table1)
```

```
table1(~ age + job + education + contact + emp.var.rate +
cons.conf.idx + euribor3m + nr.employed, data=df)
```

Kết quả

	Overall (N=41188)
age	
Mean (SD)	40.0 (10.4)
Median [Min, Max]	38.0 [17.0, 98.0]
job	
admin.	10422 (25.3%)
blue-collar	9254 (22.5%)
entrepreneur	1456 (3.5%)
housemaid	1060 (2.6%)
management	2924 (7.1%)
retired	1720 (4.2%)
self-employed	1421 (3.5%)
services	3969 (9.6%)
student	875 (2.1%)
technician	6743 (16.4%)
unemployed	1014 (2.5%)
unknown	330 (0.8%)
education	
basic.4y	4176 (10.1%)
basic.6y	2292 (5.6%)
basic.9y	6045 (14.7%)
high.school	9515 (23.1%)
illiterate	18 (0.0%)
professional.course	5243 (12.7%)
university.degree	12168 (29.5%)
unknown	1731 (4.2%)
contact	
cellular	26144 (63.5%)
telephone	15044 (36.5%)
emp.var.rate	
Mean (SD)	0.0819 (1.57)
Median [Min, Max]	1.10 [-3.40, 1.40]
cons.conf.idx	
Mean (SD)	-40.5 (4.63)
Median [Min, Max]	-41.8 [-50.8, -26.9]

	Overall (N=41188)
euribor3m	
Mean (SD)	3.62 (1.73)
Median [Min, Max]	4.86 [0.634, 5.05]
nr.employed	
Mean (SD)	5170 (72.3)
Median [Min, Max]	5190 [4960, 5230]

Vài diễn giải:

- Tổng số quan sát (số records): N=41188
- Các biến liên tục như `age`, `emp.var.rate`, `cons.conf.idx`, `euribor3m`, `nr.employed` thì lệnh `table1` báo cáo số Trung bình (Mean), độ lệch chuẩn (SD), Trung vị (Median), Nhỏ nhất (Min), và Lớn nhất (Max).

  Ví dụ tuổi trong nghiên cứu này trung bình là 40, độ lệch chuẩn 10.4, trung vị là 38, người thấp tuổi nhất là 17 tuổi, người cao tuổi nhất là 98 tuổi.

age	
Mean (SD)	40.0 (10.4)
Median [Min, Max]	38.0 [17.0, 98.0]

- Các biến phân loại (hay định tính, categorical variables) như nghề nghiệp (job), trình độ học vấn (education), hình thức liên lạc (`contact`) thì `table1` liệt kê các giá trị ở cột bên trái và báo cáo số lượng từng loại, kèm tỉ lệ % so với tổng số N.

  Ví dụ nhìn vào báo cáo của `job` có thể hình dung sơ bộ tỉ lệ công việc: việc "`admin.`", chắc là công việc văn phòng nói chung gồm có 10422 người, chiếm 25.3% trong tổng số dữ liệu quan sát; "`unknown`" có nghĩa là không biết nghề nghiệp (có thể là khi thực hiện nghiên cứu quên hỏi, hoặc quên nhập liệu hoặc người ta không chịu cung cấp) là 330 người chiếm 0.8% - không đáng kể.

job	
admin.	10422 (25.3%)
blue-collar	9254 (22.5%)
entrepreneur	1456 (3.5%)
housemaid	1060 (2.6%)
management	2924 (7.1%)

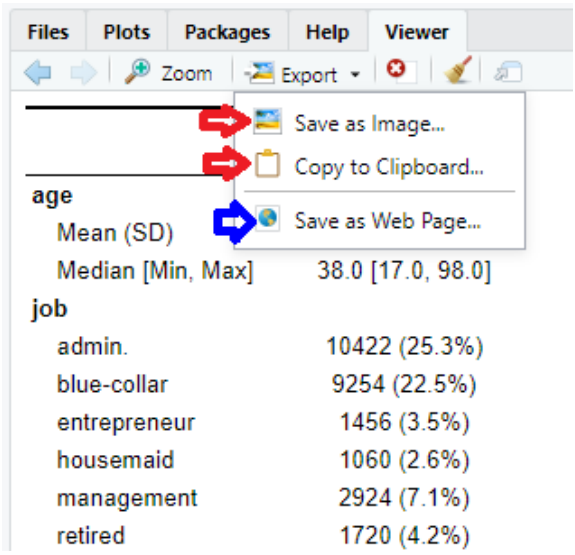
retired	1720 (4.2%)
self-employed	1421 (3.5%)
services	3969 (9.6%)
student	875 (2.1%)
technician	6743 (16.4%)
unemployed	1014 (2.5%)
unknown	330 (0.8%)

Gợi ý:

- Bạn thử phân tích thêm biến y rồi tự lý giải xem sao?

```
table1(~ age + job + education + contact + emp.var.rate +
cons.conf.idx + euribor3m + nr.employed + y, data=df)
```

- Chú ý trong RStudio để lấy được cái hình đưa vào tài liệu theo cách thông thường như dùng menu bên dưới (chỗ mũi tên màu đỏ) thì hình có thể bị cắt mất. Nên dùng chức năng Save as Web Page…để lưu thành trang web rồi copy nội dung vào tài liệu.



Thay đổi cú pháp một chút bằng cách dùng dấu | (gọi là dấu more) để phân tích theo một biến phân nhóm. Ví dụ lệnh sau sẽ phân tích các biến `age`, `job`, `education`, `emp.var.rate`, `cons.conf.idx`, `euribor3m`, `nr.employed` theo các giá trị của hình thức liên lạc (`contact`): cellular là gọi số di động; telephone là gọi số để bàn.

```
table1(~ age + job + education + emp.var.rate + cons.conf.idx +
euribor3m + nr.employed | contact, data=df)
```

Kết quả như bên dưới. Cột bên trái là các biến cần phân tích. Nhưng số liệu được phân tích theo các hình thức gọi điện (các giá trị của biến `contact`)

	cellular (N=26144)	telephone (N=15044)	Overall (N=41188)
age			
Mean (SD)	40.0 (11.0)	40.1 (9.43)	40.0 (10.4)
Median [Min, Max]	38.0 [17.0, 98.0]	39.0 [18.0, 86.0]	38.0 [17.0, 98.0]
job			
admin.	7126 (27.3%)	3296 (21.9%)	10422 (25.3%)
blue-collar	5090 (19.5%)	4164 (27.7%)	9254 (22.5%)
entrepreneur	855 (3.3%)	601 (4.0%)	1456 (3.5%)
housemaid	640 (2.4%)	420 (2.8%)	1060 (2.6%)
management	1902 (7.3%)	1022 (6.8%)	2924 (7.1%)
retired	1231 (4.7%)	489 (3.3%)	1720 (4.2%)
self-employed	893 (3.4%)	528 (3.5%)	1421 (3.5%)
services	2311 (8.8%)	1658 (11.0%)	3969 (9.6%)
student	671 (2.6%)	204 (1.4%)	875 (2.1%)
technician	4637 (17.7%)	2106 (14.0%)	6743 (16.4%)
unemployed	620 (2.4%)	394 (2.6%)	1014 (2.5%)
unknown	168 (0.6%)	162 (1.1%)	330 (0.8%)
education			
basic.4y	2350 (9.0%)	1826 (12.1%)	4176 (10.1%)
basic.6y	1247 (4.8%)	1045 (6.9%)	2292 (5.6%)
basic.9y	3452 (13.2%)	2593 (17.2%)	6045 (14.7%)
high.school	5928 (22.7%)	3587 (23.8%)	9515 (23.1%)
illiterate	15 (0.1%)	3 (0.0%)	18 (0.0%)
professional.course	3478 (13.3%)	1765 (11.7%)	5243 (12.7%)
university.degree	8657 (33.1%)	3511 (23.3%)	12168 (29.5%)
unknown	1017 (3.9%)	714 (4.7%)	1731 (4.2%)
emp.var.rate			
Mean (SD)	-0.387 (1.66)	0.897 (0.971)	0.0819 (1.57)
Median [Min, Max]	-0.100 [-3.40, 1.40]	1.10 [-3.40, 1.40]	1.10 [-3.40, 1.40]
cons.conf.idx			
Mean (SD)	-41.4 (5.02)	-39.0 (3.35)	-40.5 (4.63)
Median [Min, Max]	-42.7 [-50.8, -26.9]	-36.4 [-50.8, -26.9]	-41.8 [-50.8, -26.9]
euribor3m			
Mean (SD)	3.10 (1.82)	4.54 (1.09)	3.62 (1.73)
Median [Min, Max]	4.08 [0.634, 4.97]	4.86 [0.634, 5.05]	4.86 [0.634, 5.05]
nr.employed			
Mean (SD)	5150 (79.2)	5190 (48.6)	5170 (72.3)
Median [Min, Max]	5200 [4960, 5230]	5190 [4960, 5230]	5190 [4960, 5230]

Bạn hãy thử phân tích các biến theo trình độ (`education`) rồi diễn giải xem thế nào?

```
table1(~ age + job + contact + emp.var.rate + cons.conf.idx +
euribor3m + nr.employed | education, data=df)
```

Mở rộng một chút, bạn thử phân nhóm theo 2 biến thì như thế nào?

```
table1(~ age + job + emp.var.rate + cons.conf.idx + euribor3m +
nr.employed | education + contact, data=df)
```

Kết quả là `table1` trình bày hàng ngang được phân nhóm thành 2 cấp, bên trên là `education`, trong mỗi giá trị của `education` thì gồm các giái trị của `contact` như sau (tôi cắt bớt các cột chỉ chừa lại một cột **university.degree** và vài dòng để minh họa:

	university.degree		Overall	
	cellular (N=8657)	**telephone (N=3511)**	**cellular (N=26144)**	**telephone (N=15044)**
age				
Mean (SD)	38.7 (9.75)	39.4 (9.29)	40.0 (11.0)	40.1 (9.43)
Median [Min, Max]	36.0 [20.0, 91.0]	37.0 [22.0, 83.0]	38.0 [17.0, 98.0]	39.0 [18.0, 86.0]
job				
admin.	4229 (48.9%)	1524 (43.4%)	7126 (27.3%)	3296 (21.9%)
blue-collar	64 (0.7%)	30 (0.9%)	5090 (19.5%)	4164 (27.7%)
entrepreneur	378 (4.4%)	232 (6.6%)	855 (3.3%)	601 (4.0%)
housemaid	87 (1.0%)	52 (1.5%)	640 (2.4%)	420 (2.8%)
management	1408 (16.3%)	655 (18.7%)	1902 (7.3%)	1022 (6.8%)
retired	217 (2.5%)	68 (1.9%)	1231 (4.7%)	489 (3.3%)
self-employed	510 (5.9%)	255 (7.3%)	893 (3.4%)	528 (3.5%)
services	124 (1.4%)	49 (1.4%)	2311 (8.8%)	1658 (11.0%)
student	111 (1.3%)	59 (1.7%)	671 (2.6%)	204 (1.4%)
technician	1319 (15.2%)	490 (14.0%)	4637 (17.7%)	2106 (14.0%)
unemployed	183 (2.1%)	79 (2.3%)	620 (2.4%)	394 (2.6%)
unknown	27 (0.3%)	18 (0.5%)	168 (0.6%)	162 (1.1%)
euribor3m				

	university.degree		Overall	
	cellular (N=8657)	telephone (N=3511)	cellular (N=26144)	telephone (N=15044)
Mean (SD)	3.19 (1.82)	4.37 (1.28)	3.10 (1.82)	4.54 (1.09)
Median [Min, Max]	4.12 [0.634, 4.97]	4.86 [0.634, 5.05]	4.08 [0.634, 4.97]	4.86 [0.634, 5.05]
nr.employed				
Mean (SD)	5150 (82.2)	5190 (56.9)	5150 (79.2)	5190 (48.6)
Median [Min, Max]	5200 [4960, 5230]	5190 [4960, 5230]	5200 [4960, 5230]	5190 [4960, 5230]

Sử dụng compareGroups

Cài đặt thư viện:

```
install.packages('compareGroups')
```

Thử gọi hàm compareGroups với tham số thứ nhất bắt đầu bằng biến phân nhóm (`contact`), sau đó là dấu ~, và danh sách các biến cần phân tích cách nhau bởi dấu +. Sau đó gọi tiếp hàm `createTable(…)` với tham số là kết quả của hàm `compareGroups`.

```
cg = compareGroups(contact ~ age + education + emp.var.rate +
cons.conf.idx + euribor3m + nr.employed, data=df)

createTable(cg)
```

Kết quả như sau:

```
--------Summary descriptives table by 'contact'---------

_____________________________________________________________
                          cellular      telephone   p.overall
                           N=26144       N=15044
_____________________________________________________________
age                      40.0 (11.0)   40.1 (9.43)    0.138
education:                                            <0.001
    basic.4y            2350 (8.99%)  1826 (12.1%)
    basic.6y            1247 (4.77%)  1045 (6.95%)
    basic.9y            3452 (13.2%)  2593 (17.2%)
    high.school         5928 (22.7%)  3587 (23.8%)
    illiterate           15 (0.06%)     3 (0.02%)
    professional.course 3478 (13.3%)  1765 (11.7%)
    university.degree   8657 (33.1%)  3511 (23.3%)
    unknown             1017 (3.89%)   714 (4.75%)
emp.var.rate            -0.39 (1.66)   0.90 (0.97)    0.000
cons.conf.idx           -41.39 (5.02) -38.97 (3.35)   0.000
euribor3m                3.10 (1.82)   4.54 (1.09)    0.000
nr.employed              5152 (79.2)   5193 (48.6)    0.000
_____________________________________________________________
```

Diễn giải một chút kết quả:

- Hàm compareGroups và createTable cũng tương tự như hàm table1 với cách phân tích phân nhóm trên. Tức là cột bên trái là các biến cần phân tích, giá trị phân tích theo các giá trị của biến phân nhóm ở cột bên phải.

- Khác với hàm table1, cú pháp phân tích theo nhóm ở đây thì để biến phân nhóm bên trái.

- Điểm khác biệt tiếp theo là có thêm trị số P (cột p.overall).

Thử thay thêm tham số `show.p.trend=T` trong hàm `createTable(…)`:

```
createTable(cg, show.p.trend = T)
```

Kết quả có thêm cột "`p.trend`": trị số P theo trend (tạm thời cách tính như thế nào thì bỏ quả nhé)

```
--------Summary descriptives table by 'contact'---------

_____________________________________________________________________
                          cellular      telephone   p.overall p.trend
                           N=26144       N=15044
_____________________________________________________________________
age                     40.0 (11.0)   40.1 (9.43)    0.138    0.138
education:                                           <0.001   <0.001
    basic.4y            2350 (8.99%)  1826 (12.1%)
    basic.6y            1247 (4.77%)  1045 (6.95%)
    basic.9y            3452 (13.2%)  2593 (17.2%)
    high.school         5928 (22.7%)  3587 (23.8%)
    illiterate           15 (0.06%)     3 (0.02%)
    professional.course 3478 (13.3%)  1765 (11.7%)
    university.degree   8657 (33.1%)  3511 (23.3%)
    unknown             1017 (3.89%)   714 (4.75%)
emp.var.rate            -0.39 (1.66)   0.90 (0.97)    0.000    0.000
cons.conf.idx           -41.39 (5.02) -38.97 (3.35)   0.000    0.000
euribor3m                3.10 (1.82)   4.54 (1.09)    0.000    0.000
nr.employed              5152 (79.2)   5193 (48.6)    0.000    0.000
_____________________________________________________________________
```

Quan sát biến emp.var.rate thì thấy độ lệch chuẩn (SD) lớn hơn trị trung bình (mean). Đây là dấu hiệu cho thấy biến này không tuân theo luật phân phối chuẩn. Như vậy dùng số trung bình và độ lệch chuẩn thì không phù hợp. Lúc này bạn thêm tham số `method=c(2)` cho hàm `compareGroups` như sau để báo cáo số trung vị (median), bách phân vị 25% (Lower quartile) và bách phân vị 75% (Upper quartile):

```
cg = compareGroups(contact ~ age + education + emp.var.rate +
cons.conf.idx + euribor3m + nr.employed, method=c(2), data=df)

createTable(cg)
```

```
--------Summary descriptives table by 'contact'---------

_____________________________________________________________________
                          cellular              telephone      p.overall
                           N=26144               N=15044
_____________________________________________________________________
age                     38.0 [32.0;47.0]      39.0 [33.0;47.0]  <0.001
education:                                                      <0.001
    basic.4y            2350 (8.99%)          1826 (12.1%)
```

```
    basic.6y                  1247 (4.77%)           1045 (6.95%)
    basic.9y                  3452 (13.2%)           2593 (17.2%)
    high.school               5928 (22.7%)           3587 (23.8%)
    illiterate                 15 (0.06%)             3 (0.02%)
    professional.course       3478 (13.3%)           1765 (11.7%)
    university.degree         8657 (33.1%)           3511 (23.3%)
    unknown                   1017 (3.89%)           714 (4.75%)
emp.var.rate                -0.10 [-1.80;1.40]      1.10 [1.10;1.40]      0.000
cons.conf.idx            -42.70 [-46.20;-36.10] -36.40 [-41.80;-36.40]   0.000
euribor3m                   4.08 [1.28;4.96]        4.86 [4.86;4.96]     <0.001
nr.employed                 5196 [5099;5228]        5191 [5191;5228]     <0.001
```

Kết quả cho thấy 3 biến age, emp.var.rate, và cons.conf.idx được báo cáo số trung vị và [bách phân vị 25%, bách phân vị 75%]

Gợi ý:

Dùng lệnh ? để đọc thêm tài liệu, ý nghĩa tham số các lệnh này:

```
?compareGroups
?createTable
```

Hàm `compareGroups` có giới hạn là số giá trị của biến phân nhóm lớn hơn 5 thì sẽ báo lỗi như sau (Do biết `education` có hơn 5 giá trị):

```
cg = compareGroups(education ~ age + job + contact + emp.var.rate +
cons.conf.idx + euribor3m + nr.employed, data=df)
```

```
Error in compareGroups.fit(X = X, y = y, include.label = include.label, :
number of groups must be less or equal to 5
```

Tham khảo https://youtu.be/hDQ0T6-i1rk