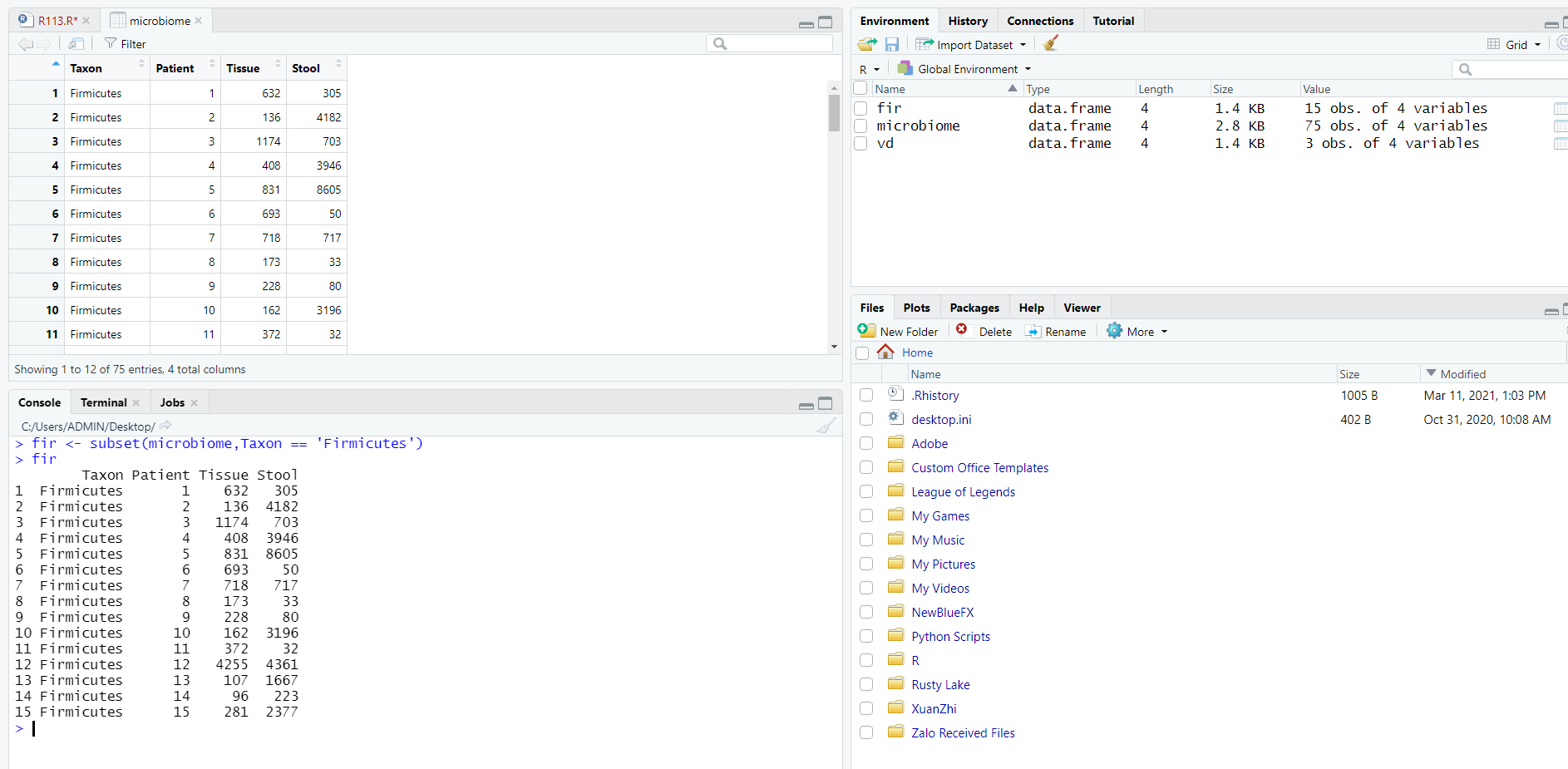
Bài tập: tóm tắt trang 20 -> 30 ( Intro\_to\_R\_Vietnamese)

5. Biên tập số liệu5.1 Tách rời dữ liệu: subset

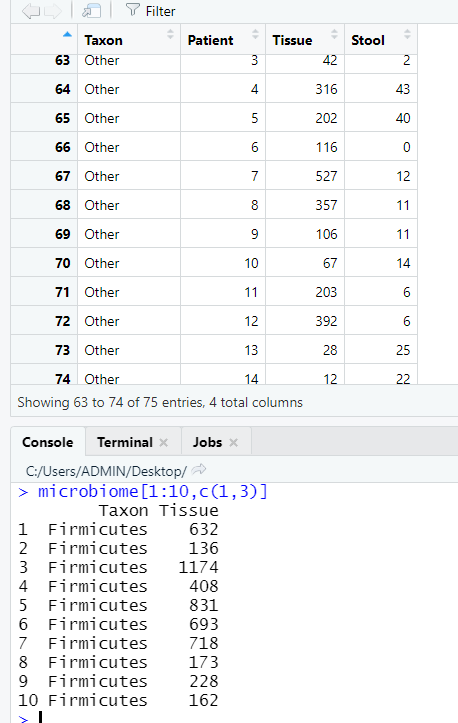
# Tách rời dữ liệu (data) thành dataframe mới đáp ứng điều kiện (cond)

***subset(data, cond)***



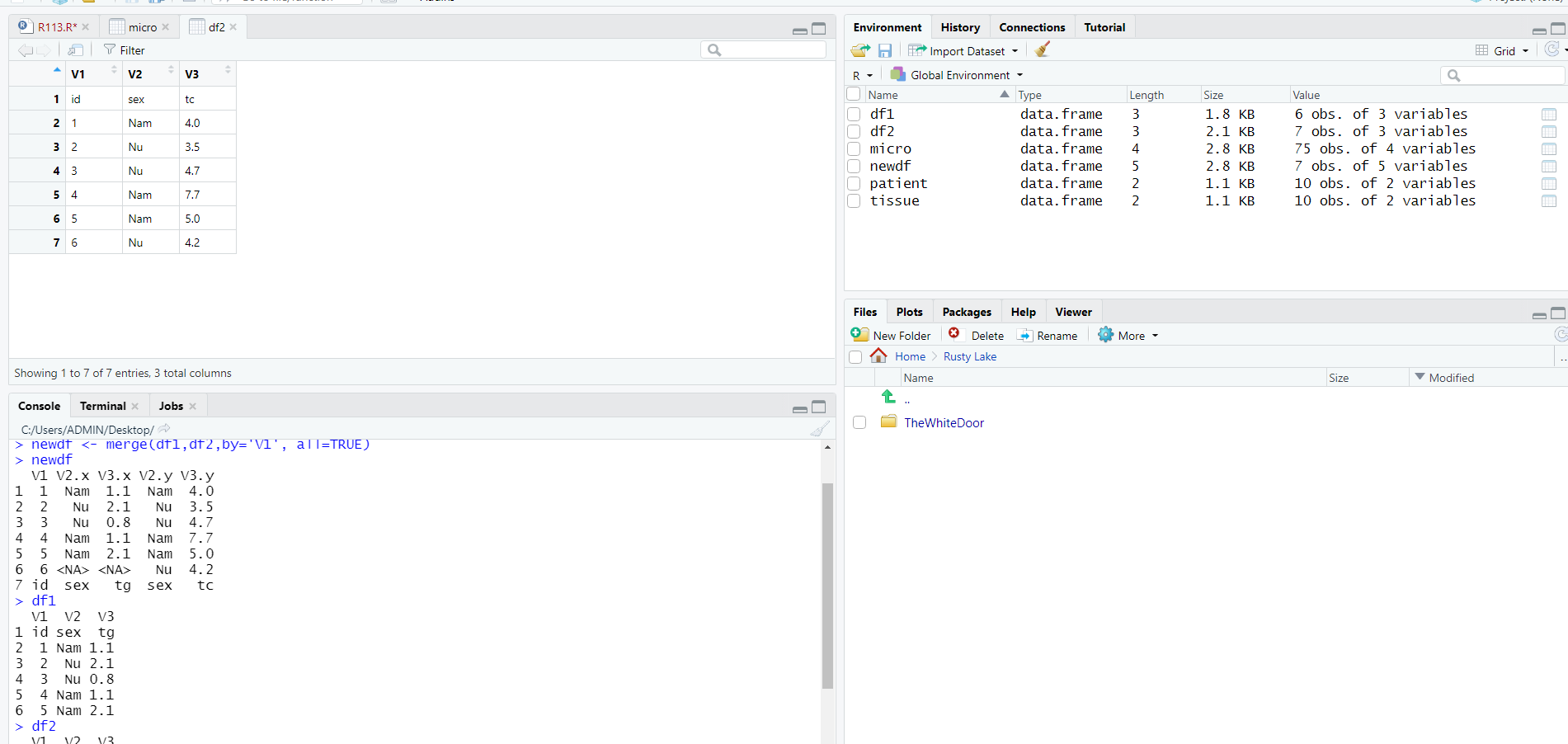
**5.2 Chiết số liệu từ một data.frame**

*Dataframe[row,col]*



**5.3 Nhập hai data.frame thành một:**

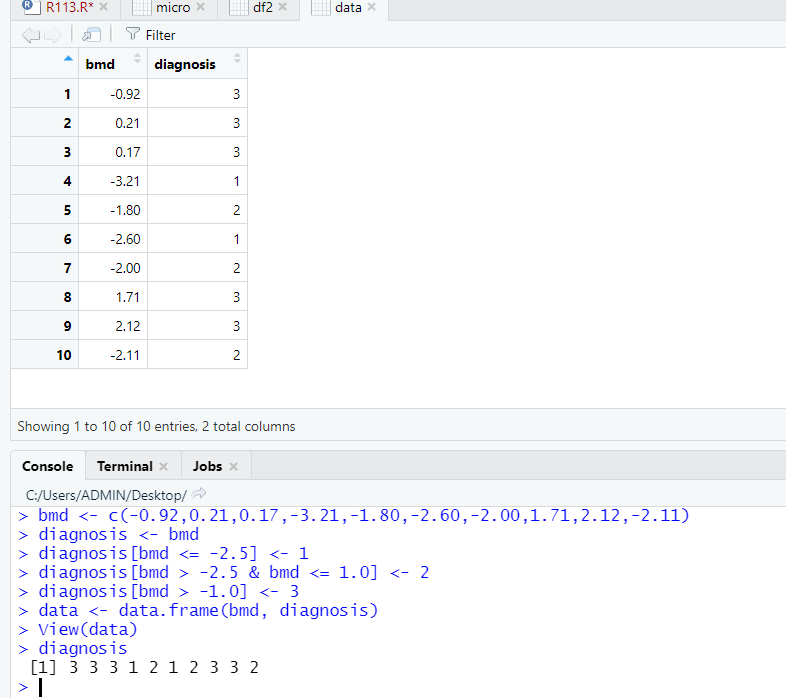
merge(**dataframe1**, **dataframe2**, by=’***colname***’, all=TRUE)



**5.4 Biến đổi số liệu (data coding)**

#Duplicate biến đó rồi gán các giá trị mới và gộp thành dataframe

**Data.frame( series1, series1fixed )**



**5.5 Biến đổi số liệu bằng cách dùng *replace***

**replace(data, from, to)**

Một cách biến đổi số liệu khác là dùng replace, dù cách này có vẻ rườm rà chút ít.

Tiếp tục ví dụ trên, chúng ta biến đổi từ bmd sang diagnosis như sau:

diagnosis <- replace(diagnosis, bmd <= -2.5, 1)

**5.6 Biến đổi thành yếu tố (*factor*)**

Trong phân tích thống kê, chúng ta phân biệt một biến số mang tính *yếu tố* (factor) và  
biến số liên tục bình thường. Biến số yếu tố không thể dùng để tính toán như cộng trừ  
nhân chia, nhưng biến số số học có thể sử dụng để tính toán.

> diag <- factor(diagnosis)  
> diag  
[1] 3 3 3 1 2 1 2 3 3 2  
Levels: 1 2 3

**Chú ý** R bây giờ thông báo cho chúng ta biết diag có 3 bậc: 1, 2 và 3. Nếu chúng ta yêu  
cầu R tính số trung bình của diag, R sẽ không làm theo yêu cầu này, vì đó không phải là  
một biến số số học:

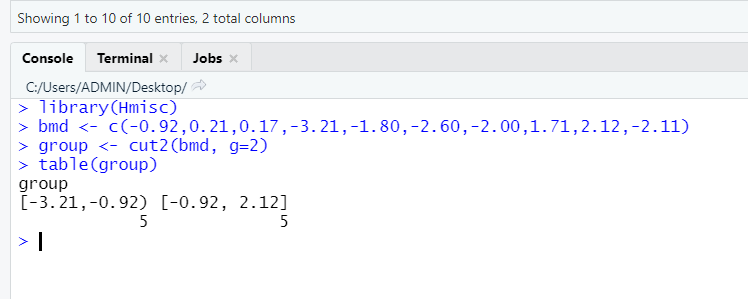
> mean(diag)  
[1] NA  
Warning message:  
argument is not numeric or logical: returning NA in: mean.default(diag)

**5.7 Phân nhóm số liệu bằng *cut2* (Hmisc)**

**Cài đặt thư viện Hmisc:**



Trong phân tích thống kê, có khi chúng ta cần phải phân **chia một biến số liên tục thành  
nhiều nhóm dựa vào phân phối của biến số**. Chẳng hạn như đối với biến số bmd chúng ta  
có thể **“cắt” dãy số thành các nhóm tương đương nhau** bằng cách dùng function cut2  
(trong thư viện Hmisc) như sau:



**6. Sử dụng R cho tính toán đơn giản**

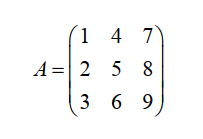
R có thể sử dụng cho các phép tính ma trận và lập chương.

**6.1 Tính toán đơn giản**

|  |  |
| --- | --- |
| Cộng hai số hay nhiều số với nhau: > 15+2997 [1] 3012 | Cộng và trừ: > 15+2997-9768 [1] -6756 |
| Nhân và chia > -27\*12/21 [1] -15.42857 | Số lũy thừa: (25 – 5)3 > (25 - 5)^3 [1] 8000 |
| Căn số bậc hai: 10 > sqrt(10) [1] 3.162278 | Số pi (π) > pi [1] 3.141593 > 2+3\*pi [1] 11.42478 |
| Logarit: loge > log(10) [1] 2.302585 | Logarit: log10 > log10(100) [1] 2 |
| Số mũ: *e*2.7689 > exp(2.7689) [1] 15.94109 > log10(2+3\*pi) [1] 1.057848 | Hàm số lượng giác > cos(pi) [1] -1 |
| Vector > x <- c(2,3,1,5,4,6,7,6,8) > x [1] 2 3 1 5 4 6 7 6 8 > sum(x) [1] 42 > x\*2 [1] 4 6 2 10 8 12 14 12 16 | > exp(x/10) [1] 1.221403 1.349859 1.105171 1.648 1.491825 1.822119 2.013753 1.822119 [9] 2.225541 > exp(cos(x/10)) [1] 2.664634 2.599545 2.704736 2.405 2.511954 2.282647 2.148655 2.282647 [9] 2.007132 |
| Tính tổng bình phương (sum of squares): 12 + 22 + 32 + 42 + 52 = ? > x <- c(1,2,3,4,5) > sum(x^2) [1] 55 | Tính tổng bình phương điều chỉnh (adjusted sum of squares): ( )2 1 *n i i x x* = ∑ - = ? > x <- c(1,2,3,4,5) > sum((x-mean(x))^2) [1] 10 Trong công thức trên mean(x) là số trung bình của vector x. |
| Tính sai số bình phương (mean square): | Tính phương sai (variance) và độ lệch chuẩn (standard deviation): |
| > x <- c(1,2,3,4,5) > sum((x-mean(x))^2)/length(x) [1] 2 Trong công thức trên, length(x) có nghĩa là tổng số phần tử (elements) trong vector x. | > x <- c(1,2,3,4,5) > var(x) [1] 2.5 Độ lệch chuẩn: *s*2 : > sd(x) [1] 1.581139 |

**6.2 Sử dụng R cho các phép tính ma trận**

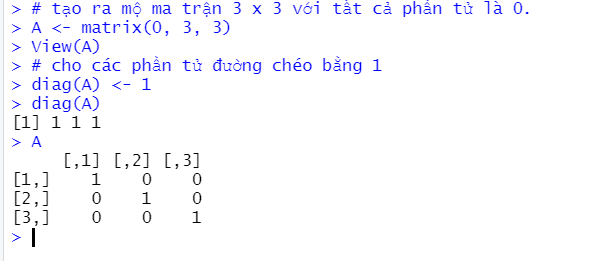
Khi viết **A**[*m*, *n*], chúng ta hiểu rằng ma trận **A** có *m* dòng và *n* cột. Trong R,  
chúng ta cũng có thể thể hiện như thế. Ví dụ: chúng ta muốn tạo một ma trận vuông A  
gồm 3 dòng và 3 cột, với các phần tử (element) 1, 2, 3, 4, 5, 6, 7, 8, 9, chúng ta viết:



Và với R:  
> y <- c(1,2,3,4,5,6,7,8,9)  
> **A <- matrix(y, nrow=3)**  
> A   
 [,1] [,2] [,3]  
[1,] 1 4 7  
[2,] 2 5 8  
[3,] 3 6 9  
Nhưng nếu chúng ta lệnh:  
> **A <- matrix(y, nrow=3, byrow=TRUE)**   
thì kết quả sẽ là:  
 [,1] [,2] [,3]  
[1,] 1 2 3  
[2,] 4 5 6  
[3,] 7 8 9

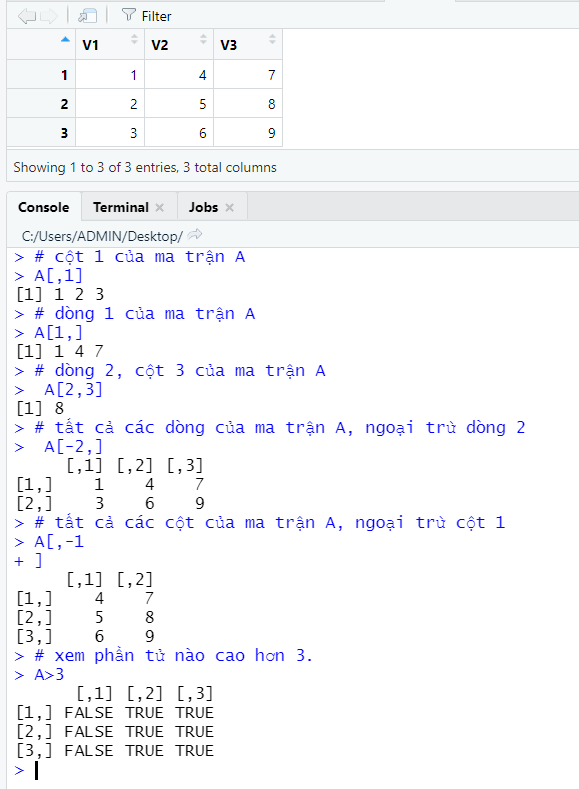
**\*Ma trận chuyển vị (transposed matrix) : t(matrix)**

**\*Ma trận vô hướng (scalar matrix)** là một ma trận vuông (tức số dòng bằng số cột), và  
tất cả các phần tử ngoài đường chéo (off-diagonal elements) là 0, và phần tử đường chéo  
là 1



**6.2.1 Chiết phần tử từ ma trận**

Tương tự chiết xuất giá trị của dataframe



**6.2.2 Tính toán với ma trận**

Cho 2 ma trận A và B

|  |  |
| --- | --- |
|  | **Cú pháp** |
| **Cộng và trừ hai ma trận**. | A + B |
| **Nhân hai ma trận** | A%\*%B |
| **Định thức (determinant)** Ma trận mà định thức bằng 0 là **ma trận suy biến** (**singular matrix**) và không thể đảo nghịch | det(A) |
| **Trị số eigen** | eigen(A) |

**\*Nghịch đảo ma trận và giải hệ phương trình**.

Ví dụ chúng ta có hệ phương trình sau đây:

