

Giới thiệu ngôn ngữ 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



Nội dung

- Ngôn ngữ R
- Vận hành
- Văn phạm
- Đọc dữ liệu



Một chút lịch sử

- R là "*a language and environment for statistical computing and graphics.*"
- Xuất phát từ software “S” vào thập niên 1980s
- Ross Ihaka và Robert Gentleman (Đại học Auckland, New Zealand) tạo ra R vào 1990s
- Từ 1997: quốc tế hóa "R-core" gồm 15 người



Ross Ihaka



Robert C. Gentleman

Ngôn ngữ (phần mềm) R

- **Mã nguồn mở - hoàn toàn miễn phí !**
- Chạy trên Windows, Unix, MacOS.
- Do các chuyên gia thống kê phát triển
- Rất nhiều phương pháp phân tích, cơ bản đến nâng cao
- Biểu đồ chất lượng cao
- Các đại học và viện nghiên cứu rất chuộng R
- "Dân chủ hóa" phương pháp thống kê
- Cộng đồng R ở Việt Nam lớn mạnh



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2019-07-05, Action of the Toes) [R-3.6.1.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

[What are R and CRAN?](#)

Installation of R on Windows

- Select **Windows**
- Select “**base**”
- Run → OK → Next
- Then **Finish**
 - R icon on your desktop



Giao diện R (Macbook)

R File Edit Format Workspace Packages & Data Misc Quartz Window Help

R Console

```
> x = rnorm(1000)
> y = 0.5 + 1.2*x + rnorm(1000)
> plot(y ~ x, pch=16, col="blue")
> # Running a regression model
> m = lm(y ~ x)
> summary(m)

Call:
lm(formula = y ~ x)

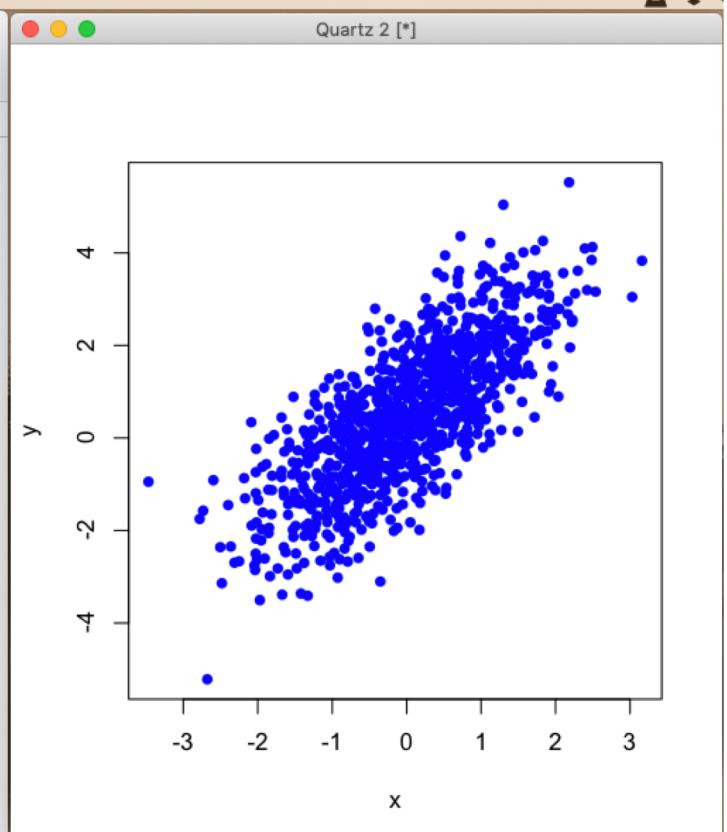
Residuals:
    Min      1Q  Median      3Q     Max 
-3.1998 -0.6768 -0.0346  0.6961  3.0297 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.50658   0.03150   16.08   <2e-16 ***
x            1.15967   0.03164   36.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.996 on 998 degrees of freedom
Multiple R-squared:  0.5737,
Adjusted R-squared:  0.5733
F-statistic: 1343 on 1 and 998 DF,  p-value: < 2.2e-16

>
```

Quartz 2 [*]



RStudio

Một dạng "add-on" R

RStudio <http://rstudio.org>



Giao diện RStudio

RStudio

Untitled1 x Addins ▾

Source on Save Go to file/function Run Source ▾

Environment History Connections

Files Plots Packages Help Viewer

Zoom Export ▾ Publish ▾

1

1:1 (Top Level) R Script

Console Jobs

~/Dropbox/_Conferences and Workshops/Dai hoc Duoc 6-2019/Datasets/

```
> x = rnorm(1000)
> y = 0.5 + 1.2*x + rnorm(1000)
> plot(y ~ x, pch=16, col="blue")
> # Running a linear regression model
> m = lm(y ~ x)
> summary(m)
```

Call:
lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-2.7323	-0.7092	-0.0146	0.6981	3.2494

Coefficients:

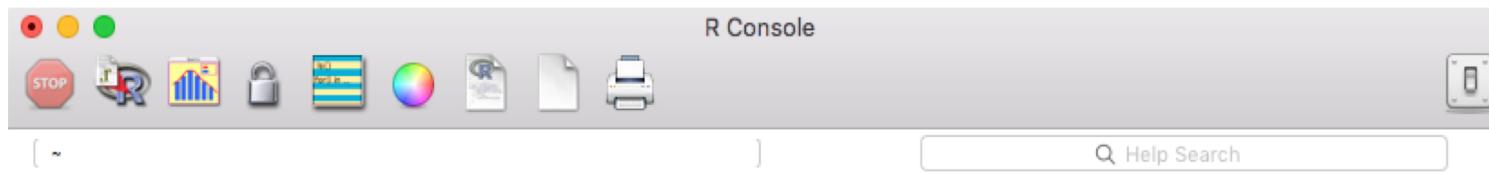
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.50240	0.03305	15.20	<2e-16 ***
x	1.16262	0.03308	35.15	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 998 degrees of freedom
Multiple R-squared: 0.5531, Adjusted R-squared: 0.5527
F-statistic: 1235 on 1 and 998 DF. p-value: < 2.2e-16

A scatter plot showing a positive linear relationship between x and y. The x-axis ranges from -2.5 to 3.5, and the y-axis ranges from -3 to 6. The data points are blue dots.

Vận hành R



```
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"  
Copyright (C) 2017 The R Foundation for Statistical Computing  
Platform: x86_64-apple-darwin15.6.0 (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
Natural language support but running in an English locale
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

```
[R.app GUI 1.70 (7338) x86_64-apple-darwin15.6.0]
```

```
[History restored from /Users/tuannguyen/.Rapp.history]
```

```
>
```

R là tập hợp nhiều "packages"

R = Base + Packages

- **Base** là phần mềm cơ bản bao gồm một số hàm dùng cho phân tích dữ liệu
- **Packages** là những modules dùng cho các phân tích chuyên dụng
- Có hơn 6000 packages trong R
- Có thể tải về và cài đặt packages trực tiếp từ mạng

Một số packages phổ biến

Vận hành và phân tích cơ bản

foreign: đọc dữ liệu từ Stata, SPSS, SAS

ggplot2: biểu đồ

table1: bảng số liệu

compareGroups: bảng số liệu

DescTools: Phân tích mô tả

dslabs: Data science labs (datasets)

AER: Applied econometrics with R

Phân tích và mô hình

rms: Các mô hình hồi qui

car: Companion to regression analysis

survival: Phân tích sống còn

cluster: Phân tích cụm

psych: Psychometrics

visreg: Regression visualization

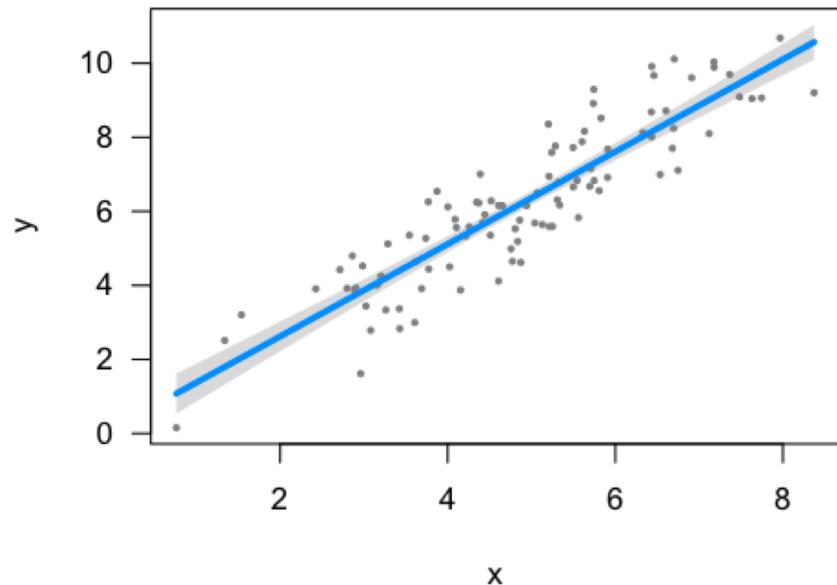
Package có thể cài đặt trực tiếp nếu **máy nối mạng**:

```
install.packages ("DescTools")
```

Sử dụng package

- Mỗi package có "hàm" (function) riêng
- Trước khi dùng hàm, cần phải "gọi" package
- Ví dụ: hiển thị mô hình hồi qui tuyến tính

```
x = rnorm(100, mean=5, sd=1.5)  
y = 0.5 + 1.2*x + rnorm(100)  
m = lm(y ~ x)  
# gọi package visreg  
library(visreg)  
visreg(m)
```



Văn phạm R

Văn phạm R

- Objects - Đối tượng
- Variables – biến số
- Functions – hàm

Object = đối tượng

- R vận hành theo đối tượng (object)
- Đối tượng là biến số, dataset, input, output, v.v.
- Đối tượng phải có *tên*
- Tên đối tượng phân biệt chữ hoa và chữ thường

tuan = 5 + 7

TUAN = lm(y ~ x)

Tuan = plot(y ~ x)

Tên biến số (variable)

- Dùng mẫu tự, số, kí hiệu (., -, _)
- Kí hiệu "assignment": <- hoặc =
- Phân biệt chữ thường và chữ hoa

```
Genotype = 5; genotype <- 7;
```

```
Geno.type = Genotype + genotype
```

R Session

```
> Genotype = 5  
> genotype <- 7  
> Geno.type = Genotype + genotype  
> Geno.type  
[1] 12
```

Hàm (function)

- Lệnh R = function
- Hàm phải có **arguments**
- Ví dụ: Phân tích mô hình $y = a + bx$

```
m1 = lm(y ~ x, data = test)
```

Object name
m1

Function (hàm)
lm = linear model

Arguments (đối số):
variables: y, x
dataset name

Đọc dữ liệu vào R

Các dữ liệu R có thể đọc

- Đọc trực tiếp
- ASCII và text files
- Excel / **csv**

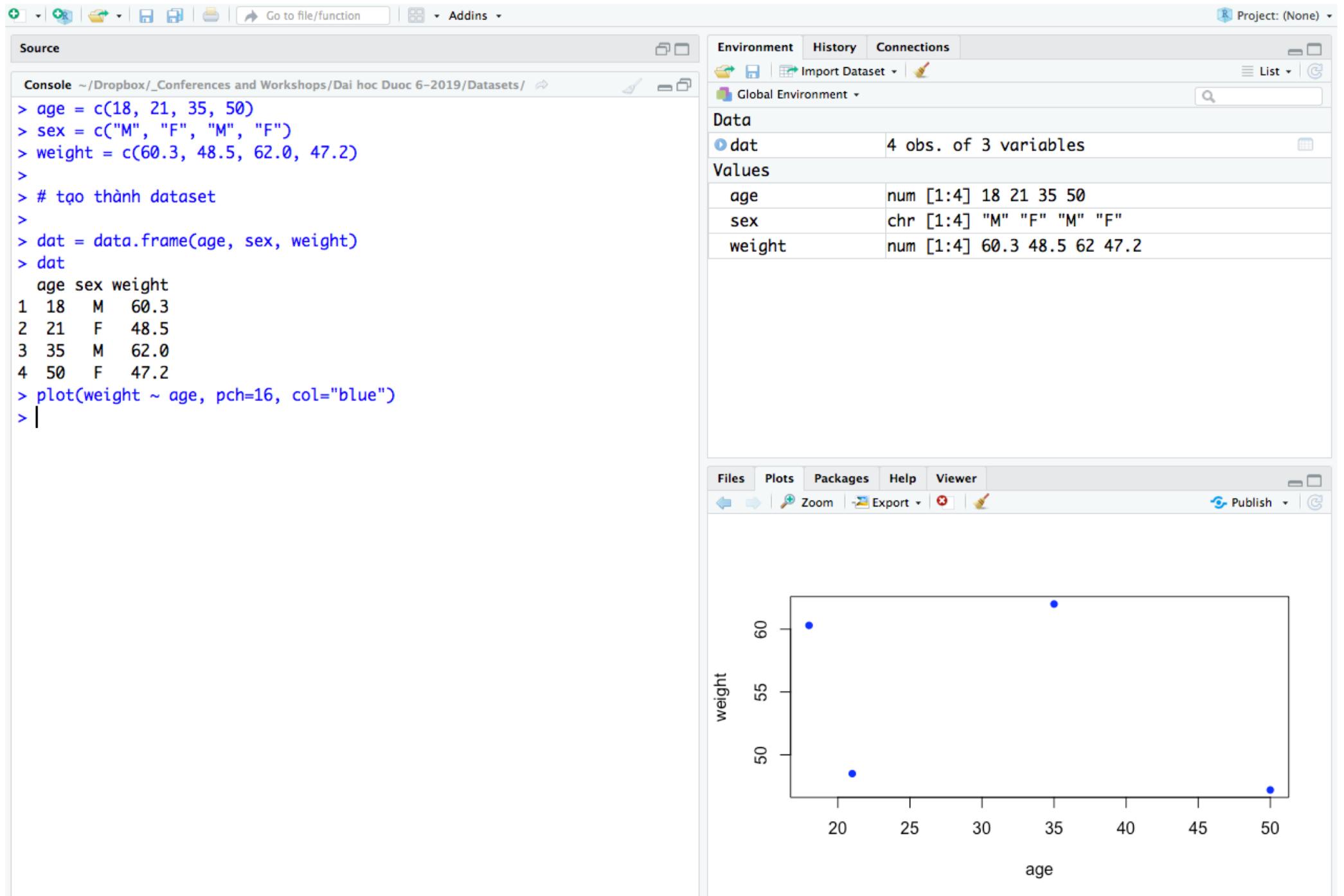
Đọc dữ liệu trực tiếp: c()

age	sex	weight
18	M	60.3
21	F	48.5
35	M	62.0
50	F	47.2

```
age = c(18, 21, 35, 50)  
sex = c("M", "F", "M", "F")  
weight = c(60.3, 48.5, 62.0, 47.2)
```

tạo thành dataset

```
dat = data.frame(age, sex, weight)  
dat
```



Dùng file.choose() để tìm file

Dùng `file.choose()`

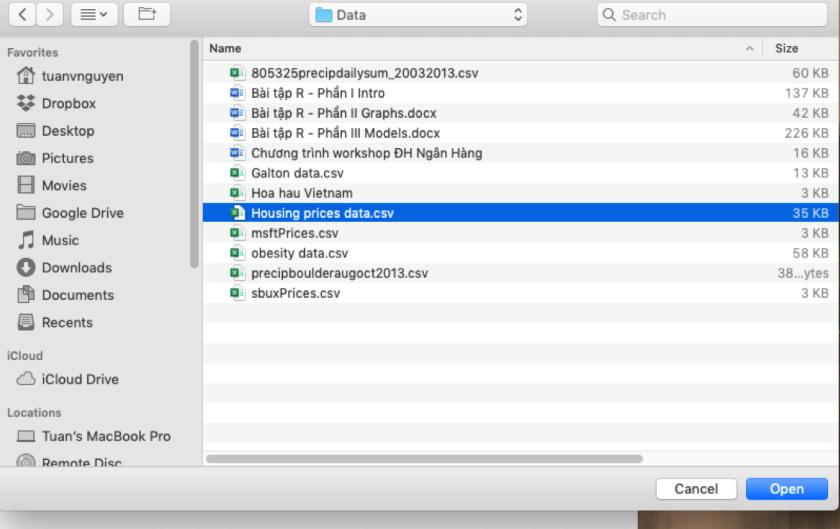
```
t = file.choose()
```

```
t = read.csv(file.choose())
```

- một cửa sổ sẽ hiện ra
- tìm file liên quan

R Console

```
> t = file.choose()
```



R Console

```
> t = file.choose()
> house = read.csv(t)
> head(house)
  crime zone industry river nox rooms age distance radial tax ptratio black
1 0.00632   18      2.31    0 0.538 6.575 65.2  4.0900    1 296 15.3 396.90
2 0.02731     0      7.07    0 0.469 6.421 78.9  4.9671    2 242 17.8 396.90
3 0.02729     0      7.07    0 0.469 7.185 61.1  4.9671    2 242 17.8 392.83
4 0.03237     0      2.18    0 0.458 6.998 45.8  6.0622    3 222 18.7 394.63
5 0.06905     0      2.18    0 0.458 7.147 54.2  6.0622    3 222 18.7 396.90
6 0.02985     0      2.18    0 0.458 6.430 58.7  6.0622    3 222 18.7 394.12
  lstat MEDV
1 4.98 24.0
2 9.14 21.6
3 4.03 34.7
4 2.94 33.4
5 5.33 36.2
6 5.21 28.7
>
```

Đọc dữ liệu từ excel

- Phức tạp (do cấu trúc excel thay đổi theo phiên bản)
- Cách tốt nhất:
 - "Xuất khẩu" sang dạng csv
 - Dùng hàm **read.csv()**

```
hh = read.csv("~/Dropbox/_Conferences and Workshops/Dai hoc  
Duoc 6-2019/Datasets/Hoa hau Vietnam.csv")
```

```
head(hh)
```

```
> head(hh)
```

	Group	Name	Name.Viet	City	Region	Crown.Year
1	Hoa Hau	Bich Phuong	B\x9di B\xcdch Ph_ng	Hanoi	North	1988
2	Hoa Hau	Dieu Hoa	Nguy_n Di_u Hoa	Hanoi	North	1990
3	Hoa Hau	Kieu Anh	H\x9a Ki_u Anh	Hanoi	North	1992
4	Hoa Hau	Thuy Thuy	Nguy_n Thu Th_y	Hanoi	North	1995
5	Hoa Hau	Thien Nga	Nguy_n Thi\x90n Nga	Saigon	South	1996
6	Hoa Hau	Ngoc Khanh	Nguy_n Th_ Ng_c Kh\x88nh	Saigon	South	1998

	Year	DOB	Age	Height	Bust	Waist	Hip	Weight
1	15/8/88	21/6/71	17.2	158	86	60	88	50
2	15/8/90	18/6/69	21.2	158	81	61	84	NA
3	15/8/92	7/7/76	16.1	169	85	62	87	NA
4	15/8/95	20/6/76	19.2	169	78	58	88	NA
5	15/8/96	25/6/75	21.2	170	87	64	92	NA
6	15/8/98	22/6/76	22.2	171	87	64	92	54

Tóm lược

- R là một trong những phát triển quan trọng của khoa học thống kê
- Hoàn toàn miễn phí
- Sử dụng rộng rãi trong các đại học trên thế giới
- R vận hành theo **packages**
- **RStudio** là một “add-on” nhưng vận hành gần như độc lập với R

Chú ý khi cài đặt package

- Có một số máy tính và trường hợp, việc cài đặt các package có thể gặp trắc ngai. Ví dụ:

"installation failed -non-zero exit status"

"package xxxx is not available (for R version yyy)"

- **Giải pháp 1**

- tìm địa chỉ trên Github (có trong <https://cran.r-project.org>)
- dùng hàm "install_github" trong package "devtools"
- ví dụ:

```
library(devtools)
```

```
install_github("ewenharrison/finalfit")
```

Chú ý khi cài đặt package

- **Giải pháp 2**
 - xác định đường link của package trong <https://cran.r-project.org>
ví dụ: https://cran.r-project.org/src/contrib/sas7bdat_0.5.tar.gz
 - dùng hàm "install_url" trong package "devtools"
 - ví dụ:

```
library(devtools)  
install_url("https://cran.r-project.org/src/contrib/sas7bdat_0.5.tar.gz")
```