Sử dụng R Markdown

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

RMarkown, RStudio, R

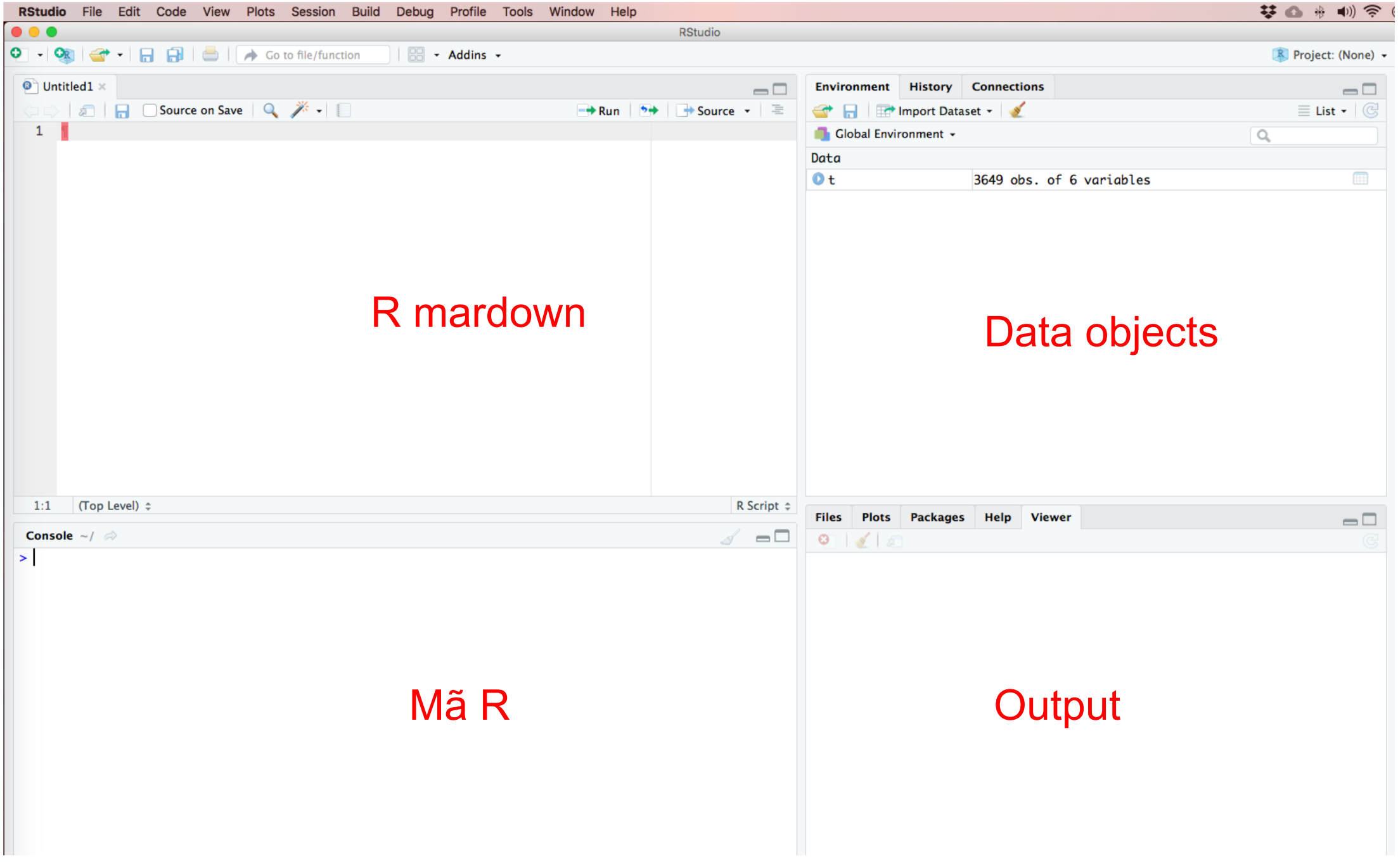
- **R** là nền tảng
- **R Studio** được xây dựng trên R
  - có 'menu' và giao diện biểu đồ
- **R Mardown** là một phần của RStudio
  - giúp 'document' mã R
  - tái lập (reproducibility)

R Studio

- Download R Studio

  https://www.rstudio.com/products/RStudio/#Desktop

- R Markdown có trong R Studio


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Connections
Import Dataset
List
Global Environment
Data
t
3649 obs. of 6 variables
Files
Plots
Packages
Help
Viewer
Console ~/
>
R mardown
Data objects
Mã R
Output

Vận hành của R Studio

RStudio

Untitled1

```
1
```

```
> # Ví dụ vận hành
> set.seed(123)
> height = rnorm(n=1000, mean=156, sd=5.5)
> error = rnorm(1000)
> p.weight = -30 + 0.55*height + error
> # fit linear regression model
> m = lm(p.weight ~ height)
> summary(m)

Call:
lm(formula = p.weight ~ height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0279 -0.6914  0.0043  0.7087  3.2911

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.456296   0.911762  -35.60   <2e-16 ***
height        0.566009   0.005838   96.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.006 on 998 degrees of freedom
Multiple R-squared:  0.904,     Adjusted R-squared:  0.9039
F-statistic:  9401 on 1 and 998 DF,  p-value: < 2.2e-16

> hist(height, col="blue", border="white")
>
```

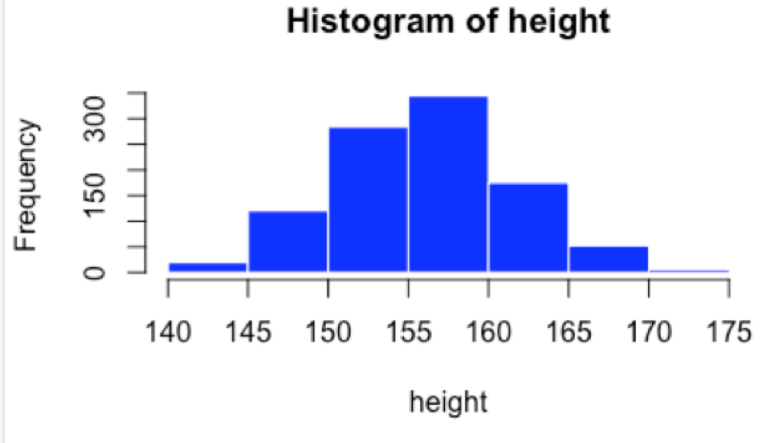
Có thể lưu mã R và output bằng cách dùng File/Save

Environment	History	Connections
Import Dataset	List	
Global Environment		

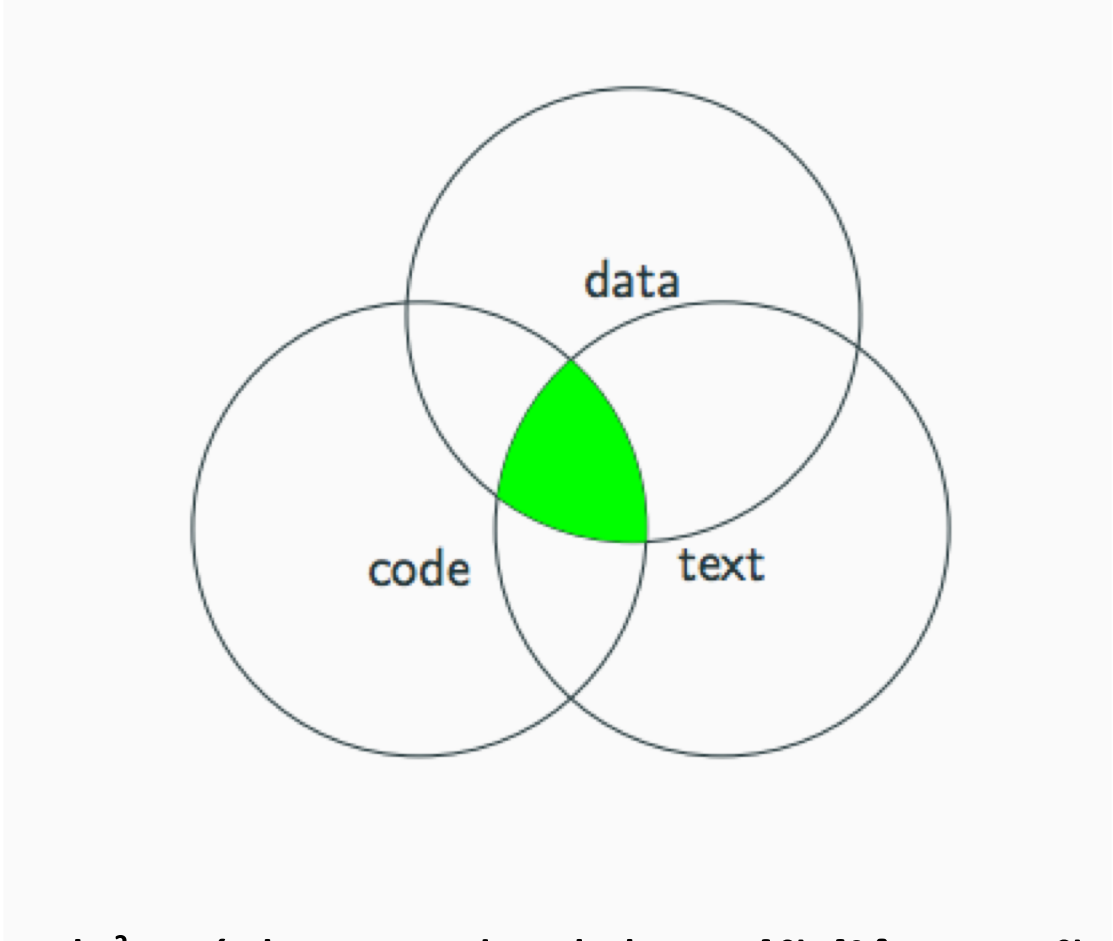
Data	
m	List of 12
t	3649 obs. of 6 variables
Values	
error	num [1:1000] -0.996 -1.04 -0.018 -0.132 -…
height	num [1:1000] 153 155 165 156 157 ...
p.weight	num [1:1000] 53.1 54.1 60.5 55.9 53.6 ...

Files | Plots | Packages | Help | Viewer

Zoom | Export



R Markdown và mục tiêu tái lập

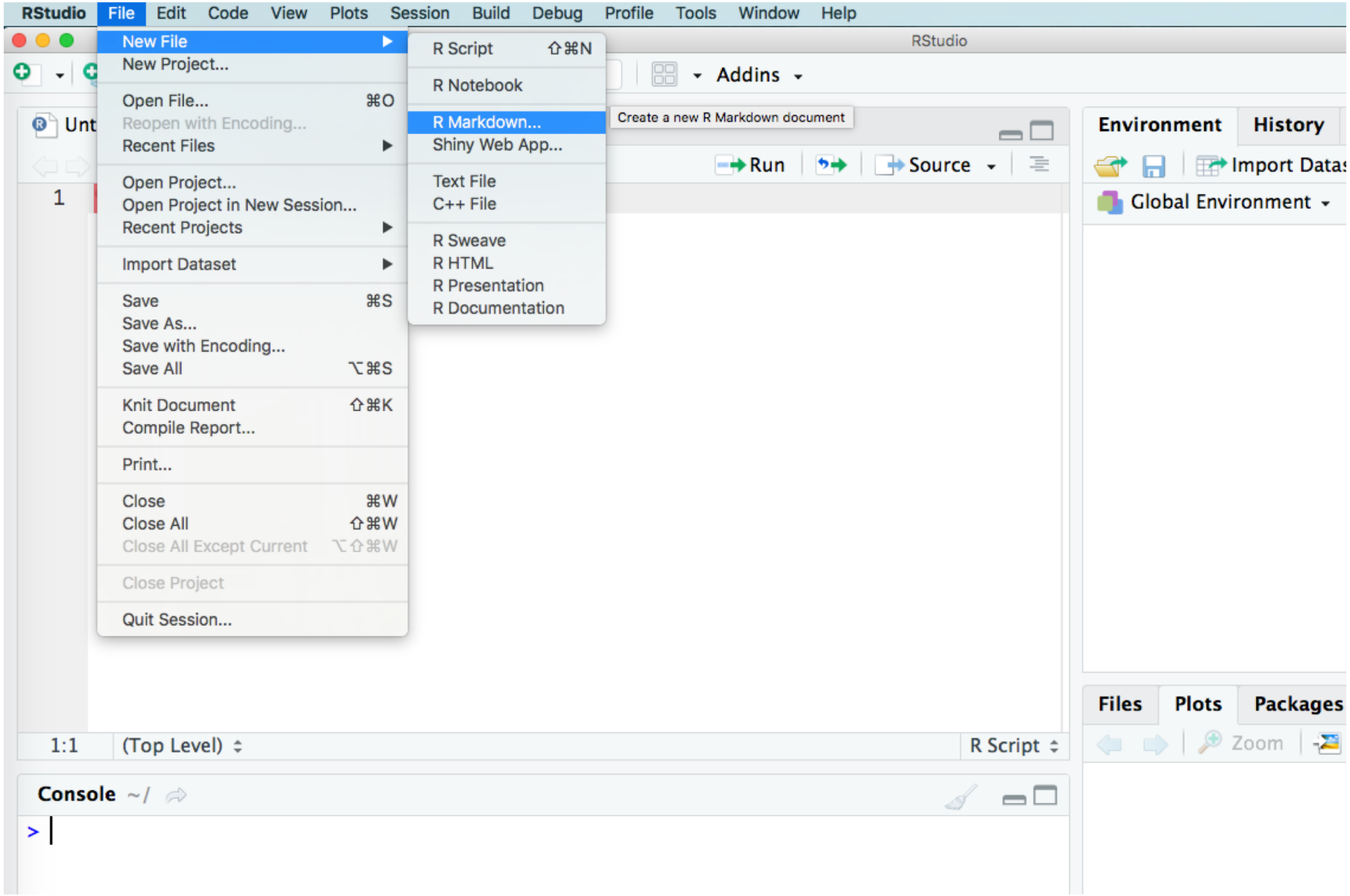


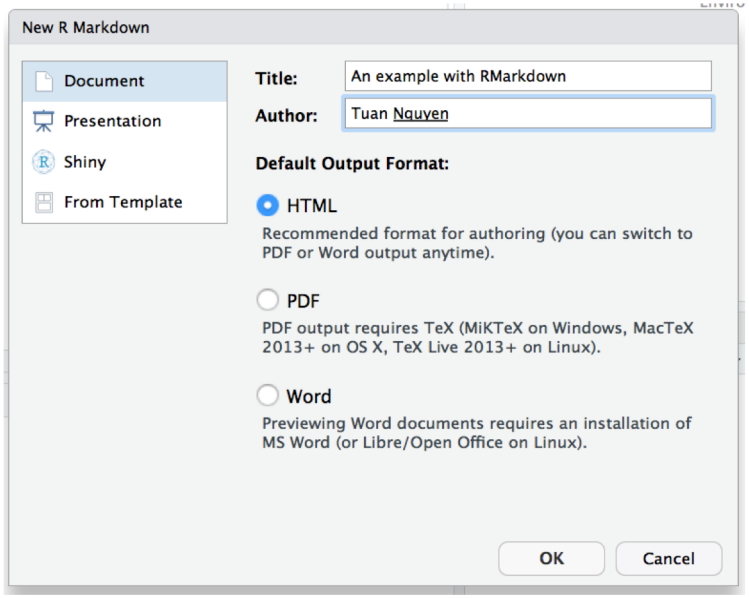
Đảm bảo tính reproducibility: **dữ liệu + mã R + văn bản**

R Markdown

- R + Markdown = RMarkdown
- Có thể xem là một "markup language"
- Lưu trữ mã R
- Export output sang **Word, pdf, html**

Một session với R Markdown





RStudio

Go to file/function

Addins

Untitled1

Untitled2

Knit

Insert

Run

```
1  ---
2  title: "An example with RMarkdown"
3  author: "Tuan Nguyen"
4  date: "4/22/2019"
5  output: html_document
6  ---
7
8  ```{r setup, include=FALSE}
9  knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for
   authoring HTML, PDF, and MS Word documents. For more details on using R
   Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the **Knit** button a document will be generated that
   includes both content as well as the output of any embedded R code chunks
   within the document. You can embed an R code chunk like this:
17
18 ```{r cars}
19 summary(cars)
20 ```
21
22 ## Including Plots
23
24 You can also embed plots, for example:
25
```

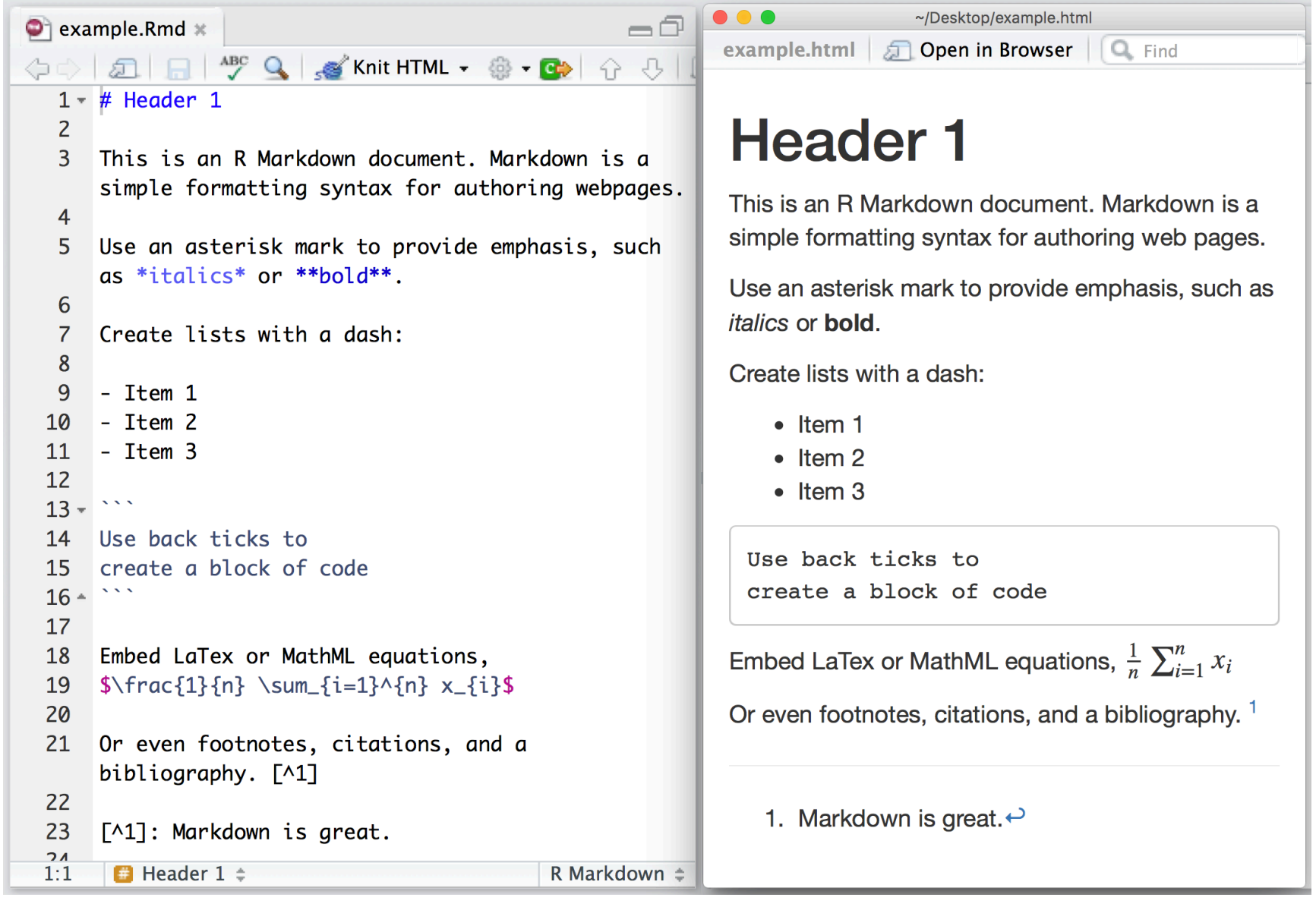
2:1 An example with RMarkdown R Markdown

Console ~/

>

Một document của R Markdown

- Header
- R codes
- Output



Một document của R Markdown

- Header
- R codes
- Output

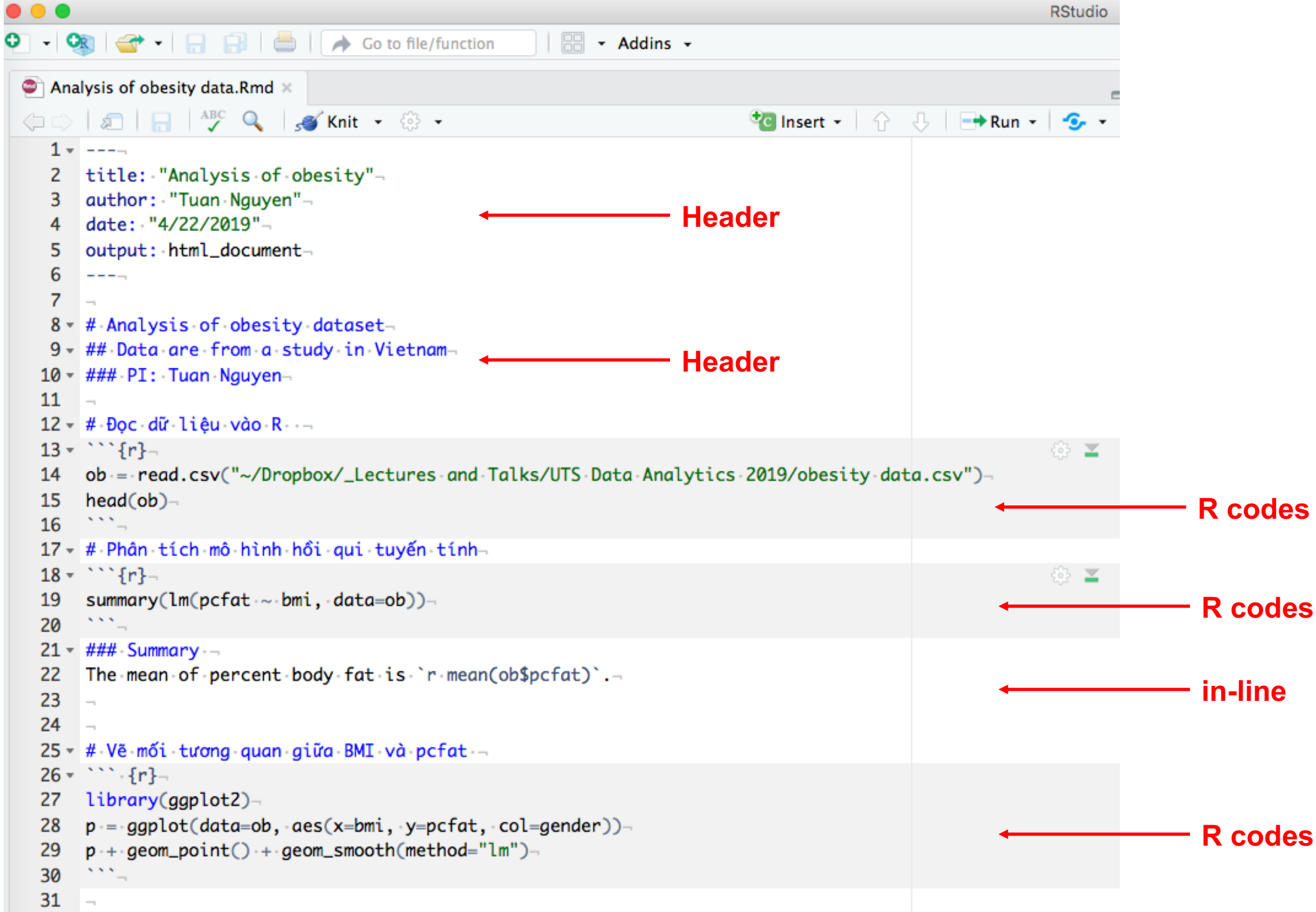
Giữa ``` và ``` là R codes

```
# Phân tích tương quan giữa BMI và pcfat
## Dữ liệu obesity data, Việt Nam
### PI: Tuan Nguyen

# Đọc dữ liệu vào R
``` {r}
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
summary(lm(pcfat ~ bmi, data=ob))

# Vẽ mối tương quan giữa BMI và pcfat
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, ol=gender))
p + geom_point() + geom_smooth(method="lm")
```

<text giữa hai dấu này không có trong phần output>
```


RStudio
Go to file/function
Addins
Analysis of obesity data.Rmd
Knit
Insert
Run
---
title: "Analysis of obesity"
author: "Tuan Nguyen"
date: "4/22/2019"
output: html_document

Analysis of obesity dataset
Data are from a study in Vietnam
PI: Tuan Nguyen
Đọc dữ liệu vào R
```{r}
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```
Phân tích mô hình hồi qui tuyến tính
```{r}
summary(lm(pcfat ~ bmi, data=ob))
```
Summary
The mean of percent body fat is `r mean(ob$pcfat)`.
Vẽ mối tương quan giữa BMI và pcfat
```{r}
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```
Header
Header
R codes
R codes
in-line
R codes

Output →

Analysis of obesity

Tuan Nguyen

4/22/2019

Analysis of obesity dataset

Data are from a study in Vietnam

PI: Tuan Nguyen

Đọc dữ liệu vào R

```
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```

```
##   id gender height weight  bmi age WBBMC wbbmd   fat  lean pcfat
## 1  1      F    150     49 21.8  53  1312  0.88 17802 28600  37.3
## 2  2      M    165     52 19.1  65  1309  0.84  8381 40229  16.8
## 3  3      F    157     57 23.1  64  1230  0.84 19221 36057  34.0
## 4  4      F    156     53 21.8  56  1171  0.80 17472 33094  33.8
## 5  5      M    160     51 19.9  54  1681  0.98  7336 40621  14.8
## 6  6      F    153     47 20.1  52  1358  0.91 14904 30068  32.2
```

Phân tích mô hình hồi qui tuyến tính

```
summary(lm(pcfat ~ bmi, data=ob))
```

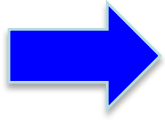
```
##
## Call:
## lm(formula = pcfat ~ bmi, data = ob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.612  -4.181   1.392   4.690  18.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.39889    1.36777   6.141 1.11e-09 ***
## bmi          1.03619    0.06051  17.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.45 on 1215 degrees of freedom
## Multiple R-squared:  0.1944, Adjusted R-squared:  0.1937
## F-statistic: 293.2 on 1 and 1215 DF,  p-value: < 2.2e-16
```

Summary

The mean of percent body fat is 31.6047859.

R Markdown codes and output

```
---
title: "Analysis of obesity"
author: "Tuan Nguyen"
date: "4/22/2019"
output: html_document
---
# Analysis of obesity dataset
## Data are from a study in Vietnam
### PI: Tuan Nguyen
```



Analysis of obesity

Tuan Nguyen

4/22/2019

Analysis of obesity dataset

Data are from a study in Vietnam

PI: Tuan Nguyen

R Markdown codes and output

````
# Đọc dữ liệu vào R

```{r}
ob = read.csv("~/Dropbox/_Lectures
and Talks/UTS Data Analytics
2019/obesity data.csv")
head(ob)
```
````

Đọc dữ liệu vào R

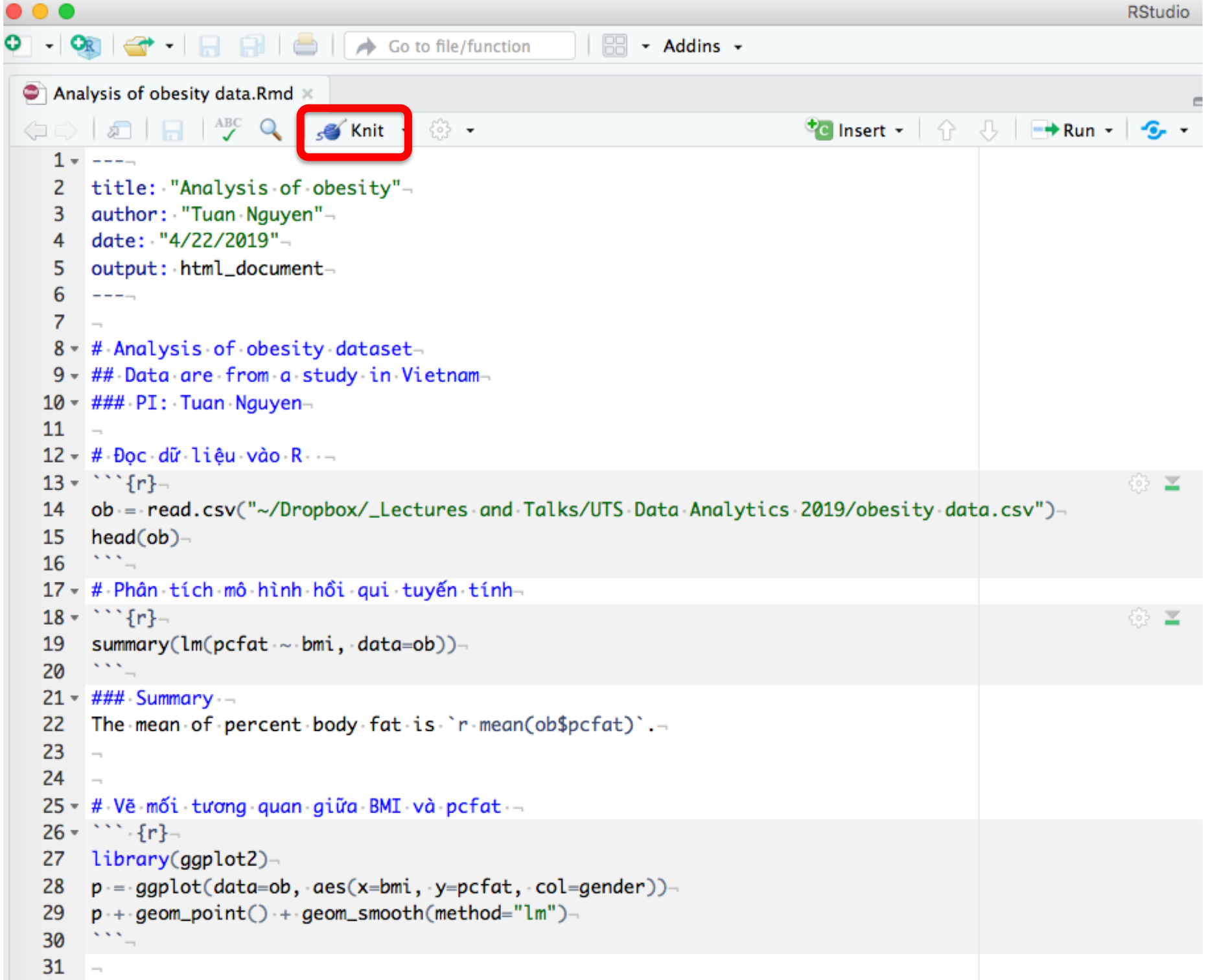
```
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```

```
##   id gender height weight  bmi age WBBMC wbbmd   fat  lean pcfat
## 1  1      F    150     49 21.8  53  1312  0.88 17802 28600  37.3
## 2  2      M    165     52 19.1  65  1309  0.84  8381 40229  16.8
## 3  3      F    157     57 23.1  64  1230  0.84 19221 36057  34.0
## 4  4      F    156     53 21.8  56  1171  0.80 17472 33094  33.8
## 5  5      M    160     51 19.9  54  1681  0.98  7336 40621  14.8
## 6  6      F    153     47 20.1  52  1358  0.91 14904 30068  32.2
```

```
# Phân tích mô hình hồi qui tuyến tính
```{r}
summary(lm(pcfat ~ bmi, data=ob))
```
Summary
The mean of percent body fat is `r mean(ob$pcfat)`.
```

Phân tích mô hình hồi qui tuyến tính

```
summary(lm(pcfat ~ bmi, data=ob))
```

```
##
## Call:
## lm(formula = pcfat ~ bmi, data = ob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.612  -4.181   1.392   4.690  18.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.39889    1.36777   6.141 1.11e-09 ***
## bmi          1.03619    0.06051  17.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.45 on 1215 degrees of freedom
## Multiple R-squared:  0.1944, Adjusted R-squared:  0.1937
## F-statistic: 293.2 on 1 and 1215 DF,  p-value: < 2.2e-16
```

Summary

The mean of percent body fat is 31.6047859.

```
# Vẽ mối tương quan giữa BMI và pcfat
``` {r}
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```
```

Vẽ mối tương quan giữa BMI và pcfat

```
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```



Sau khi viết xong, có thể "Knit" để cho ra html output trên **rpubs.com**

Analysis of obesity

Tuan Nguyen

4/22/2019

Analysis of obesity dataset

Data are from a study in Vietnam

PI: Tuan Nguyen

Đọc dữ liệu vào R

```
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```

```
##   id gender height weight  bmi age WBBMC wbbmd   fat  lean pcfat
## 1  1      F    150     49 21.8  53  1312  0.88 17802 28600  37.3
## 2  2      M    165     52 19.1  65  1309  0.84  8381 40229  16.8
## 3  3      F    157     57 23.1  64  1230  0.84 19221 36057  34.0
## 4  4      F    156     53 21.8  56  1171  0.80 17472 33094  33.8
## 5  5      M    160     51 19.9  54  1681  0.98  7336 40621  14.8
## 6  6      F    153     47 20.1  52  1358  0.91 14904 30068  32.2
```

Phân tích mô hình hồi qui tuyến tính

```
summary(lm(pcfat ~ bmi, data=ob))
```

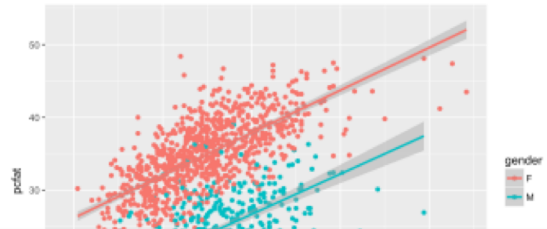
```
##
## Call:
## lm(formula = pcfat ~ bmi, data = ob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.612  -4.181   1.392   4.690  18.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.39889    1.36777   6.141 1.11e-09 ***
## bmi          1.03619    0.06051  17.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.45 on 1215 degrees of freedom
## Multiple R-squared:  0.1944, Adjusted R-squared:  0.1937
## F-statistic: 293.2 on 1 and 1215 DF,  p-value: < 2.2e-16
```

Summary

The mean of percent body fat is 31.6047859.

Vẽ mối tương quan giữa BMI và pcfat

```
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```



RStudio và RMarkdown

- Hai 'additions' rất quan trọng cho phân tích dữ liệu với R
- **RStudio** cung cấp giao diện 'thân thiện' hơn R
- **RMarkdown** là một 'ngôn ngữ bị chú' cho một công trình phân tích dữ liệu
  - một phương tiện rất có ích cho lưu trữ mã R
  - đảm bảo tính tái lập trong phân tích dữ liệu
  - có thể xuất sang Word, html và pdf


NGUYỄN VĂN TUẤN
PHÂN TÍCH DỮ LIỆU VỚI
R
Hỏi và Đáp
NHÀ XUẤT BẢN TỔNG HỢP THÀNH PHỐ HỒ CHÍ MINH
NGUYỄN VĂN TUẤN | PHÂN TÍCH DỮ LIỆU VỚI R