112年度衛生福利部衛生福利資料科學中心

統計軟體推廣課程

健康資料管理與R軟體

《基礎篇》

劉品崧　統計諮詢暨分析師

花蓮慈濟醫院高齡健康中心

課程大綱

- 課程導覽
  - 課程目的、設計思維
- R軟體安裝
  - 名詞定義、下載及安裝、RStudio環境設定、R軟體的互動模式
- R軟體實作
  - 套件與函數、資料管理、統計分析

課程導覽

- 課程目的
- 設計思維

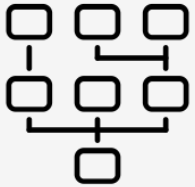
課程目的：健康資料管理與研究的基石

- 知識
  - 機率分布、資料特性、統計分析、報表解讀 ...
- 觀念
  - 研究設計、流行病學、研究設計、樣本估算 ...
- 技術
  - 撰寫指令、程式管理、除錯偵測、維護編修 ...
- 實務
  - 獨立研究、團隊合作、工作規劃、執行計畫 ...

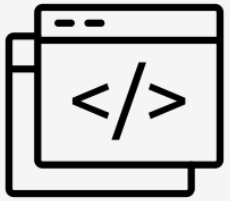
設計思維

- 思考核心理念

- 建構工作心流



- 應用技術實踐



R軟體安裝

- 名詞定義
- 下載及安裝
- RStudio環境設定
- R軟體的互動模式

關於R軟體

1. R是一個軟體（software），具資料管理、統計運算與視覺化等功能
2. R語言（R language）是泛指與R軟體溝通的代碼（code）
   又稱為指令（command）或函數（function）
3. 一支程式（program）由許多代碼及註解（annotation）所組成
4. 代碼之間依循語法（syntax）撰寫組成
5. 資料會以物件（object）的方式存於工作環境（environment）
6. 套件（package）包含一群相關功能的函數（function）
   讓使用者完成目的，使用者依據需求可以自己安裝（install）套件

關於Rtools軟體

1. 只有Windows使用者需要安裝Rtools軟體
2. 在R套件的安裝中擔任一個編譯（compile）的角色

https://cran.r-project.org/bin/windows/Rtools/rtools40.html

關於RStudio軟體

1. RStudio是整合開發環境

   （integrated development environment, IDE）

2. 核心仍然是R，所以要先安裝R軟體，之後RStudio才可以運作

3. 圖形化使用者介面（graphical user interface, GUI）

   讓初學者比較好上手

https://posit.co/download/rstudio-desktop/

各類電腦作業系統下載需求

- R系列皆為免費開源軟體，Google搜尋軟體名稱即可找到下載點
- Windows使用者名稱，不可以是中文
- 依據所使用的作業系統不同，你需要安裝以下軟體

作業系統 \ 安裝軟體	R	Rtools	Xcode	RStudio
Windows 務必以系統管理員身分執行安裝	V	V		V
Mac 請先確認iOS夠新可以安裝	V		V	V
Linux	V			V

https://www.r-project.org/

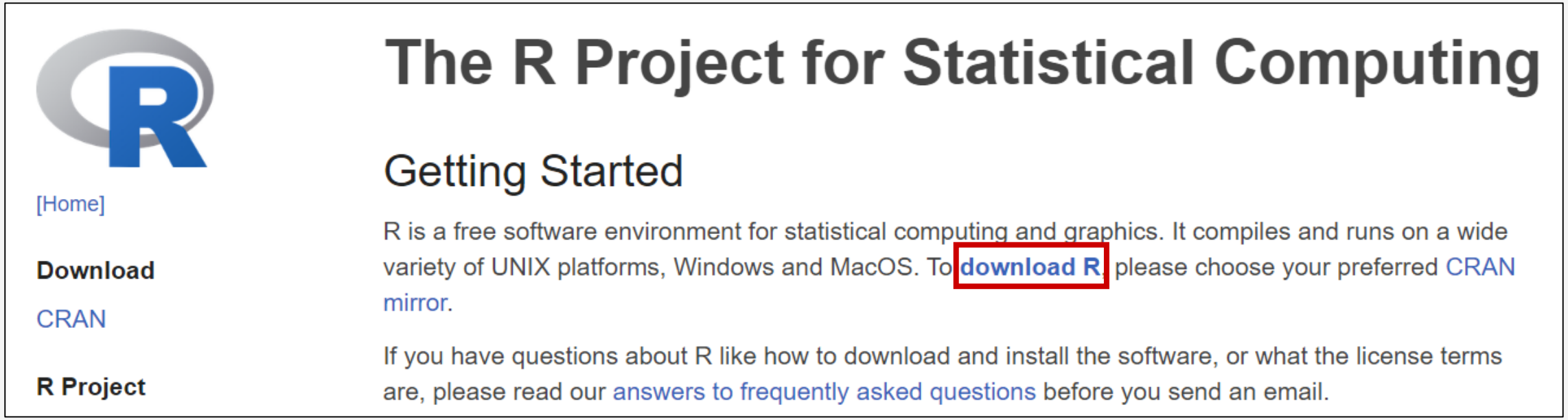
R軟體下載方式

- Google search「R」

- R官方首頁

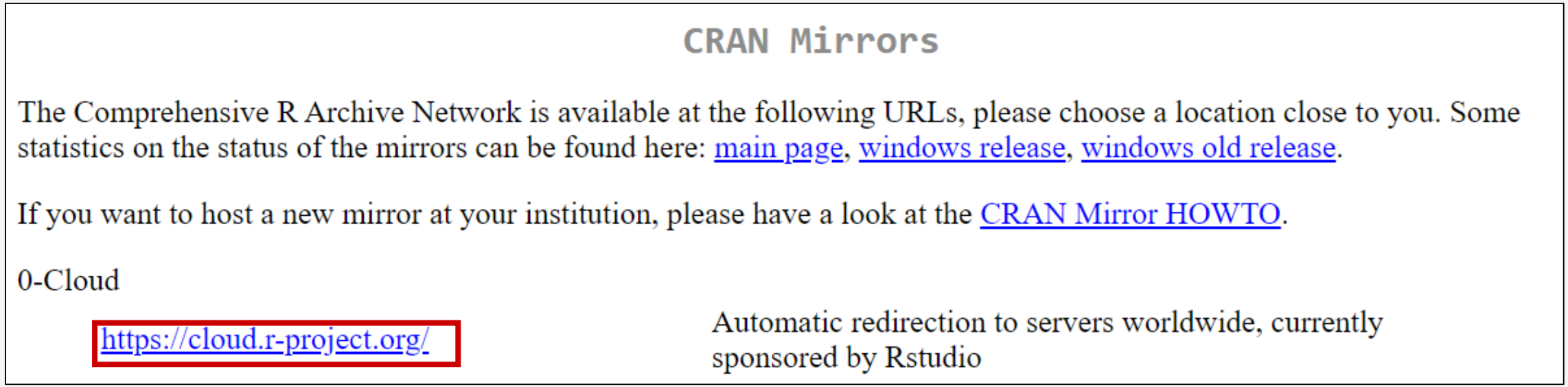
R軟體下載方式

- 從官方首頁點選「download R」



R軟體下載方式

- 點選其中一個下載點網址



CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

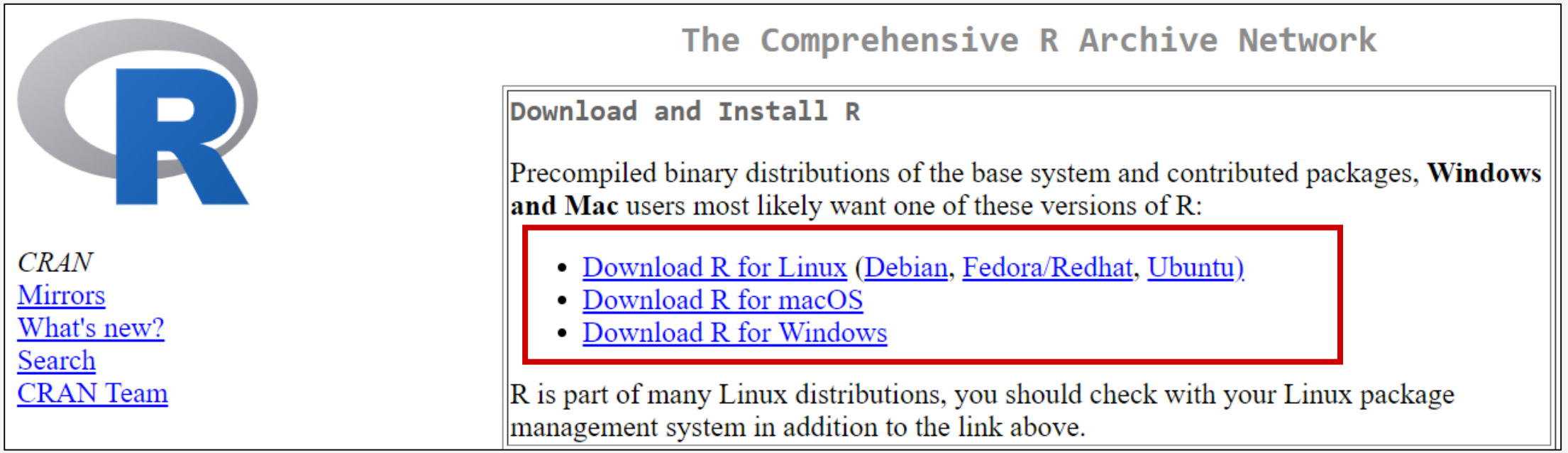
If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO.

0-Cloud

https://cloud.r-project.org/ — Automatic redirection to servers worldwide, currently sponsored by Rstudio

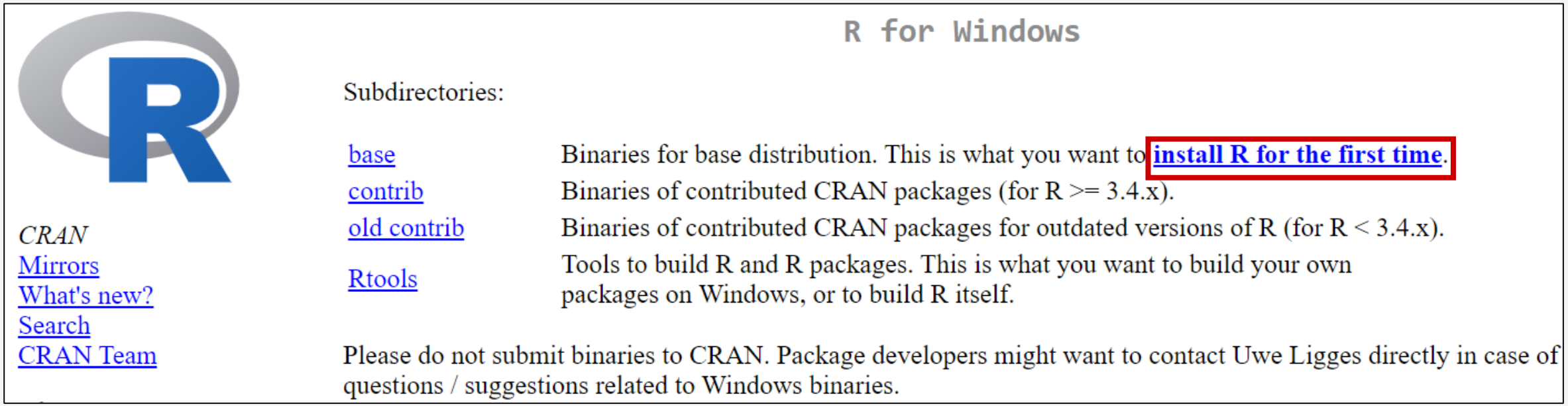
R軟體下載方式

- 從電腦作業系統對應的網址點選進去



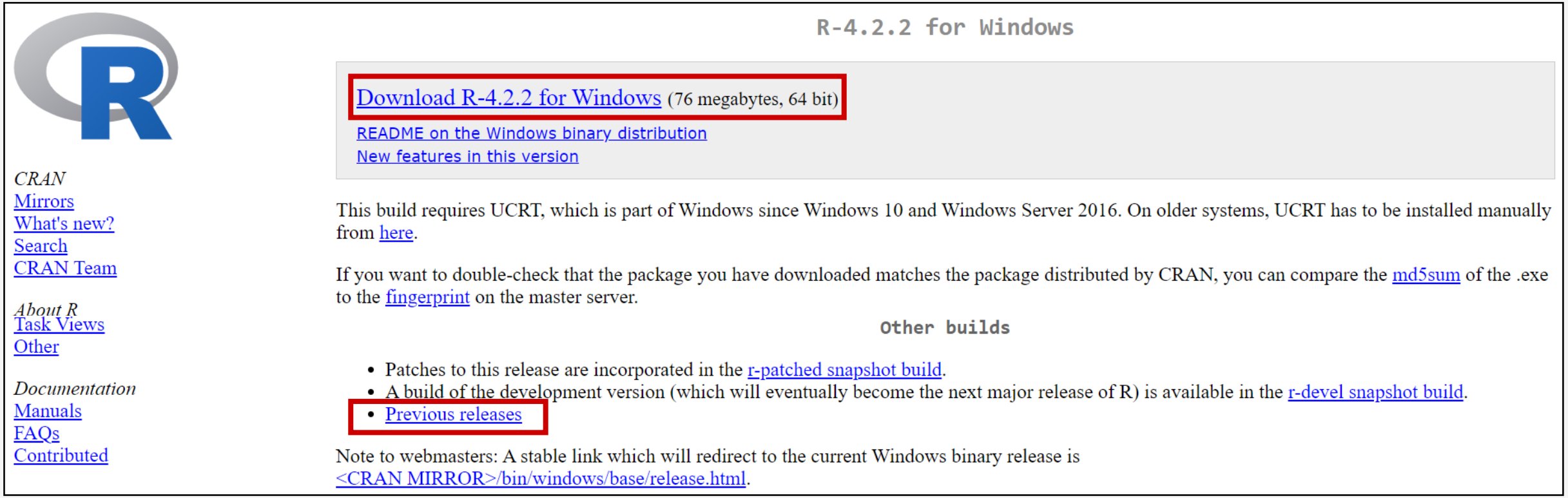
R軟體下載方式

- 點選「install R for the first time」下載最新版本



R軟體下載方式

- 點選「Download R-x.x.x for Windows」下載最新版本安裝檔
- 點選「Previous releases」下載較舊版本安裝檔



CRAN
Mirrors
What's new?
Search
CRAN Team

About R
Task Views
Other

Documentation
Manuals
FAQs
Contributed

R-4.2.2 for Windows

Download R-4.2.2 for Windows (76 megabytes, 64 bit)
README on the Windows binary distribution
New features in this version

This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT has to be installed manually from here.

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server.

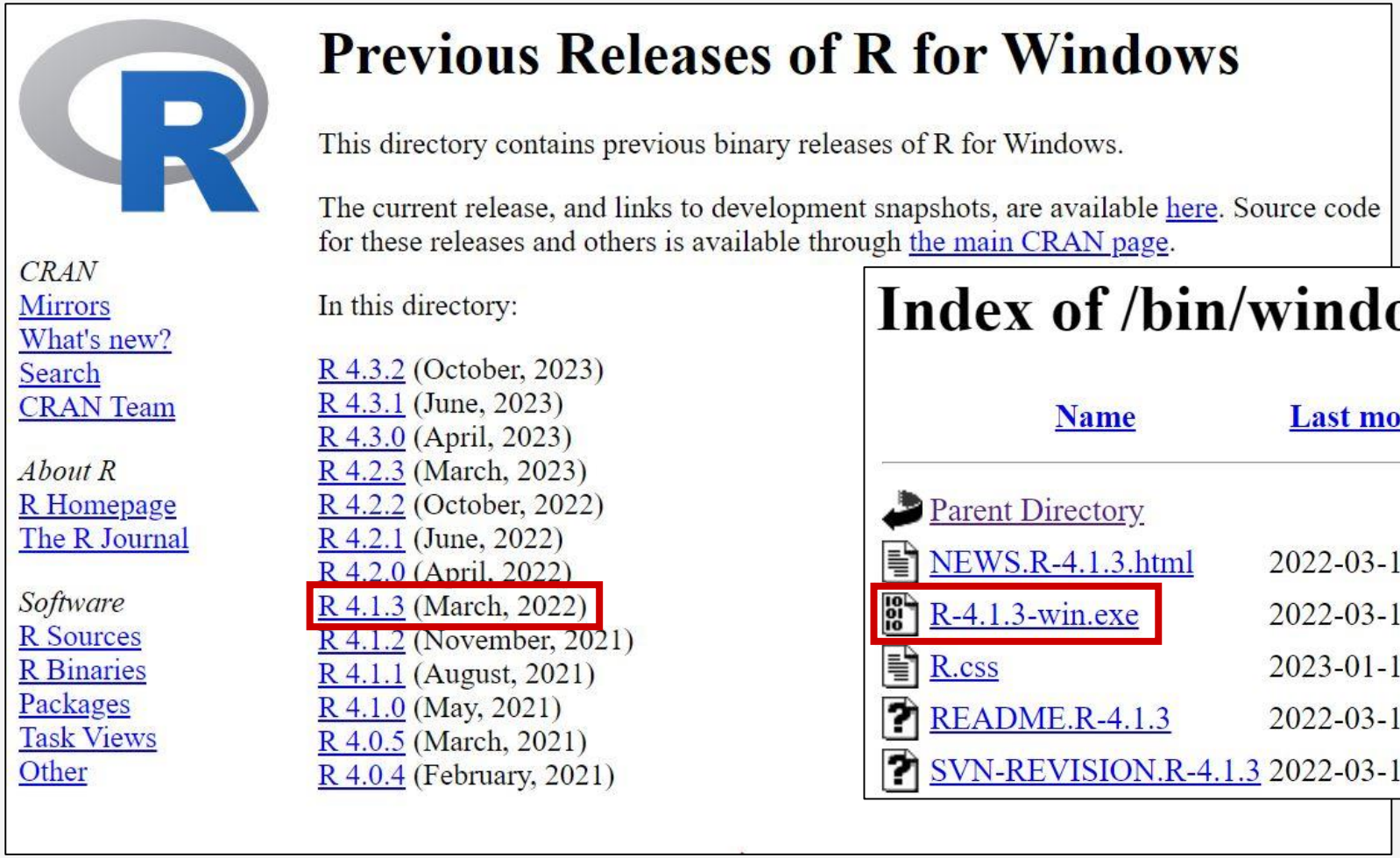
Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.html.

R軟體下載方式

- 下載較舊版本安裝檔方式，建議安裝4.1.3版



Previous Releases of R for Windows

This directory contains previous binary releases of R for Windows.

The current release, and links to development snapshots, are available here. Source code for these releases and others is available through the main CRAN page.

In this directory:

R 4.3.2 (October, 2023)
R 4.3.1 (June, 2023)
R 4.3.0 (April, 2023)
R 4.2.3 (March, 2023)
R 4.2.2 (October, 2022)
R 4.2.1 (June, 2022)
R 4.2.0 (April, 2022)
R 4.1.3 (March, 2022)
R 4.1.2 (November, 2021)
R 4.1.1 (August, 2021)
R 4.1.0 (May, 2021)
R 4.0.5 (March, 2021)
R 4.0.4 (February, 2021)

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Task Views
Other

Index of /bin/windows/base/old/4.1.3

Name	Last modified	Size	Description
Parent Directory		-	
NEWS.R-4.1.3.html	2022-03-10 09:16	109K	
R-4.1.3-win.exe	2022-03-10 10:32	86M	
R.css	2023-01-19 13:33	1.8K	
README.R-4.1.3	2022-03-10 09:16	8.5K	
SVN-REVISION.R-4.1.3	2022-03-10 09:16	46	

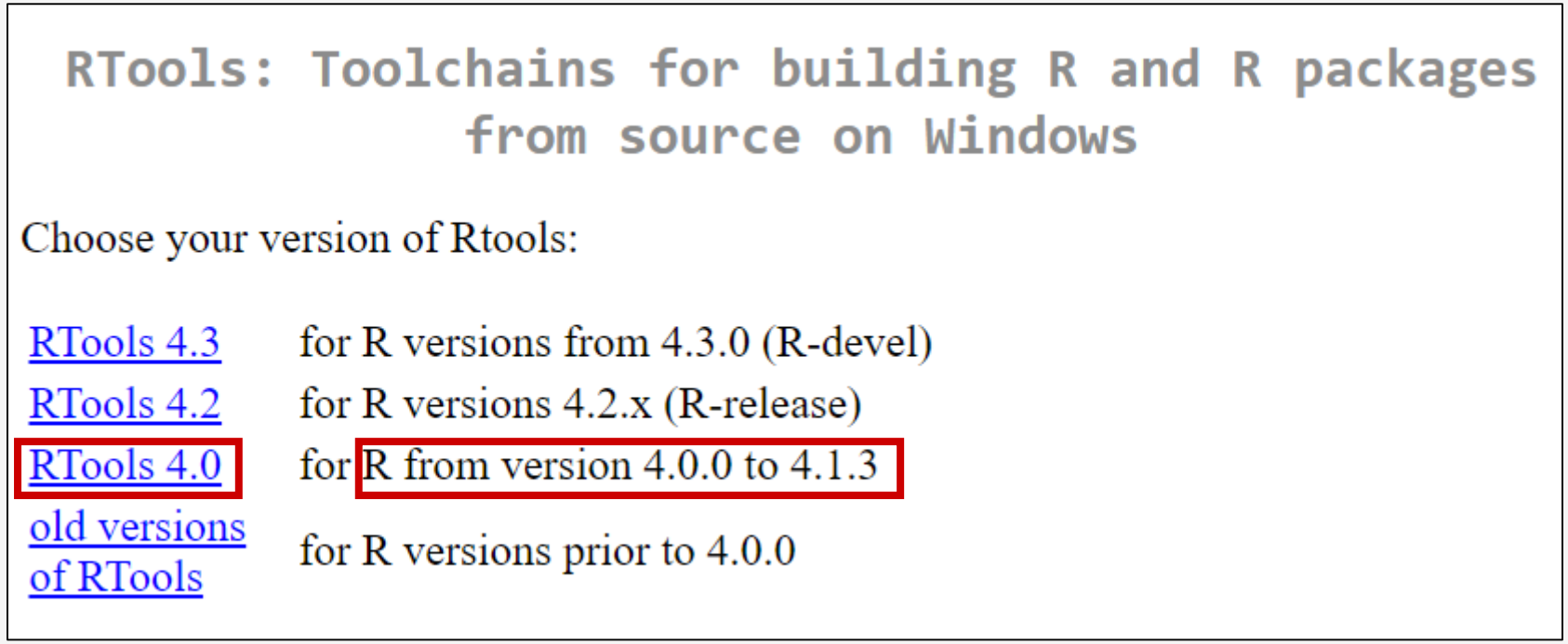
RTools軟體下載方式

- Google search「RTools」

- RTools官方首頁

RTools軟體下載方式

- 點選對應自己R軟體版本的Rtools下載



RTools: Toolchains for building R and R packages from source on Windows

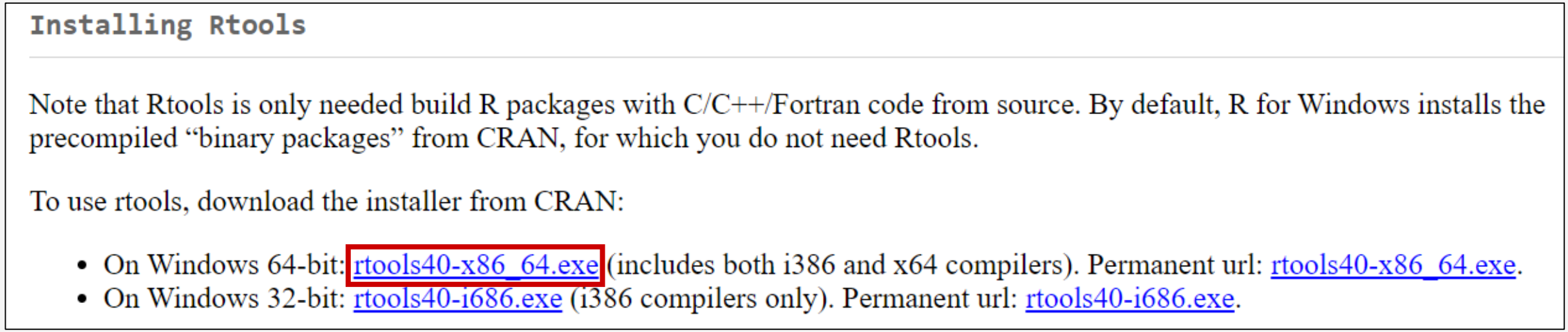
Choose your version of Rtools:

RTools 4.3	for R versions from 4.3.0 (R-devel)
RTools 4.2	for R versions 4.2.x (R-release)
RTools 4.0	for R from version 4.0.0 to 4.1.3
old versions of RTools	for R versions prior to 4.0.0

https://cran.r-project.org/bin/windows/Rtools/

RTools軟體下載方式

- 點選對應自己電腦位元數版本的Rtools下載
  - 電腦位元：從「本機」空白處，點「右鍵」選擇「內容」，檢視「系統類型」



Installing Rtools

Note that Rtools is only needed build R packages with C/C++/Fortran code from source. By default, R for Windows installs the precompiled "binary packages" from CRAN, for which you do not need Rtools.

To use rtools, download the installer from CRAN:

- On Windows 64-bit: rtools40-x86_64.exe (includes both i386 and x64 compilers). Permanent url: rtools40-x86_64.exe.
- On Windows 32-bit: rtools40-i686.exe (i386 compilers only). Permanent url: rtools40-i686.exe.

https://cran.r-project.org/bin/windows/Rtools/

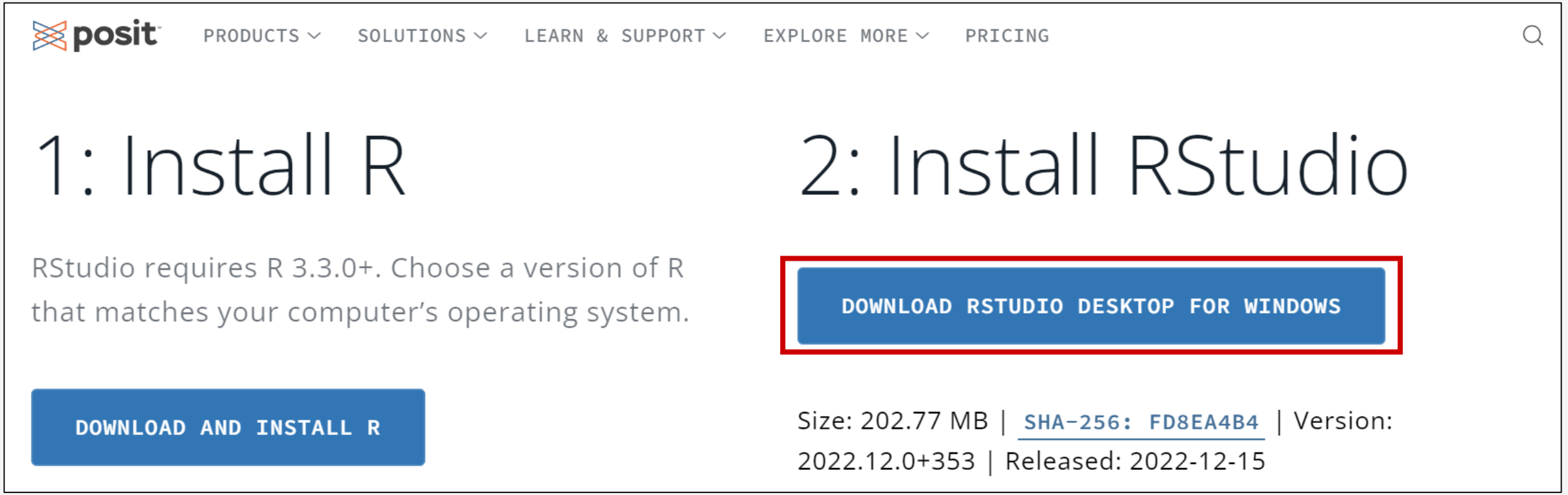
RStudio軟體下載方式

- Google search「RStudio」

- RStudio官方首頁

RStudio軟體下載方式

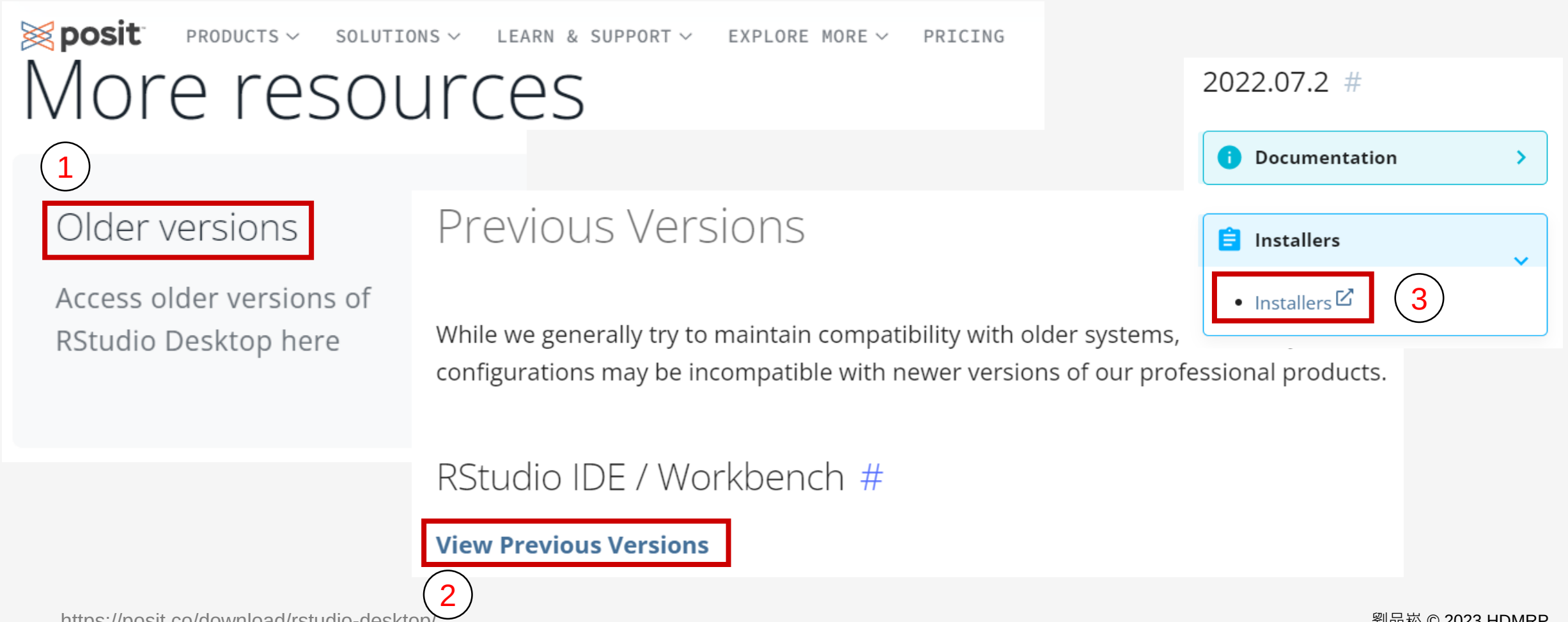
- 點選「Download RStudio desktop for Windows」下載安裝檔



https://posit.co/download/rstudio-desktop/

RStudio軟體下載方式

- 下載較舊版本安裝檔方式

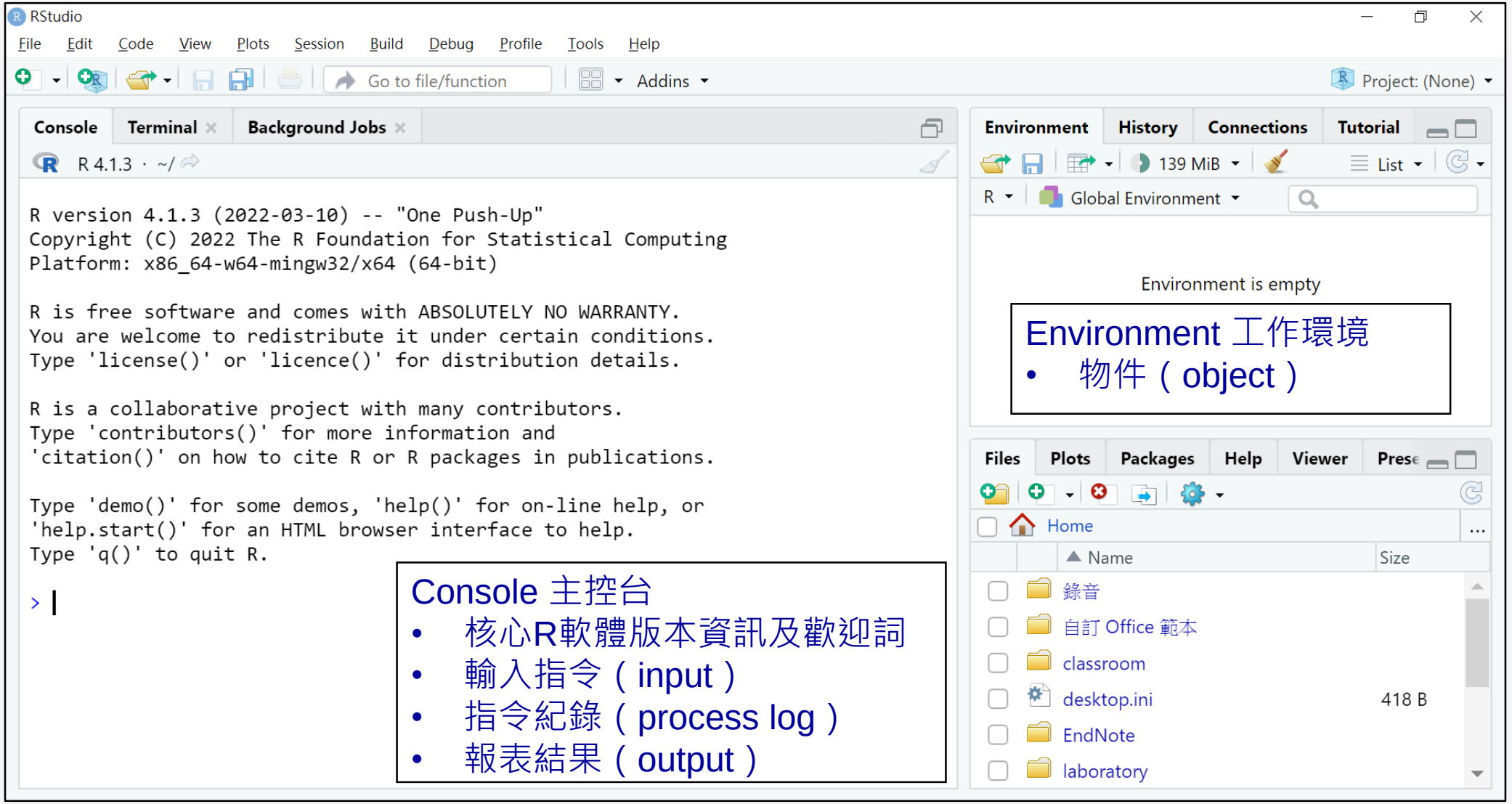


https://posit.co/download/rstudio-desktop/

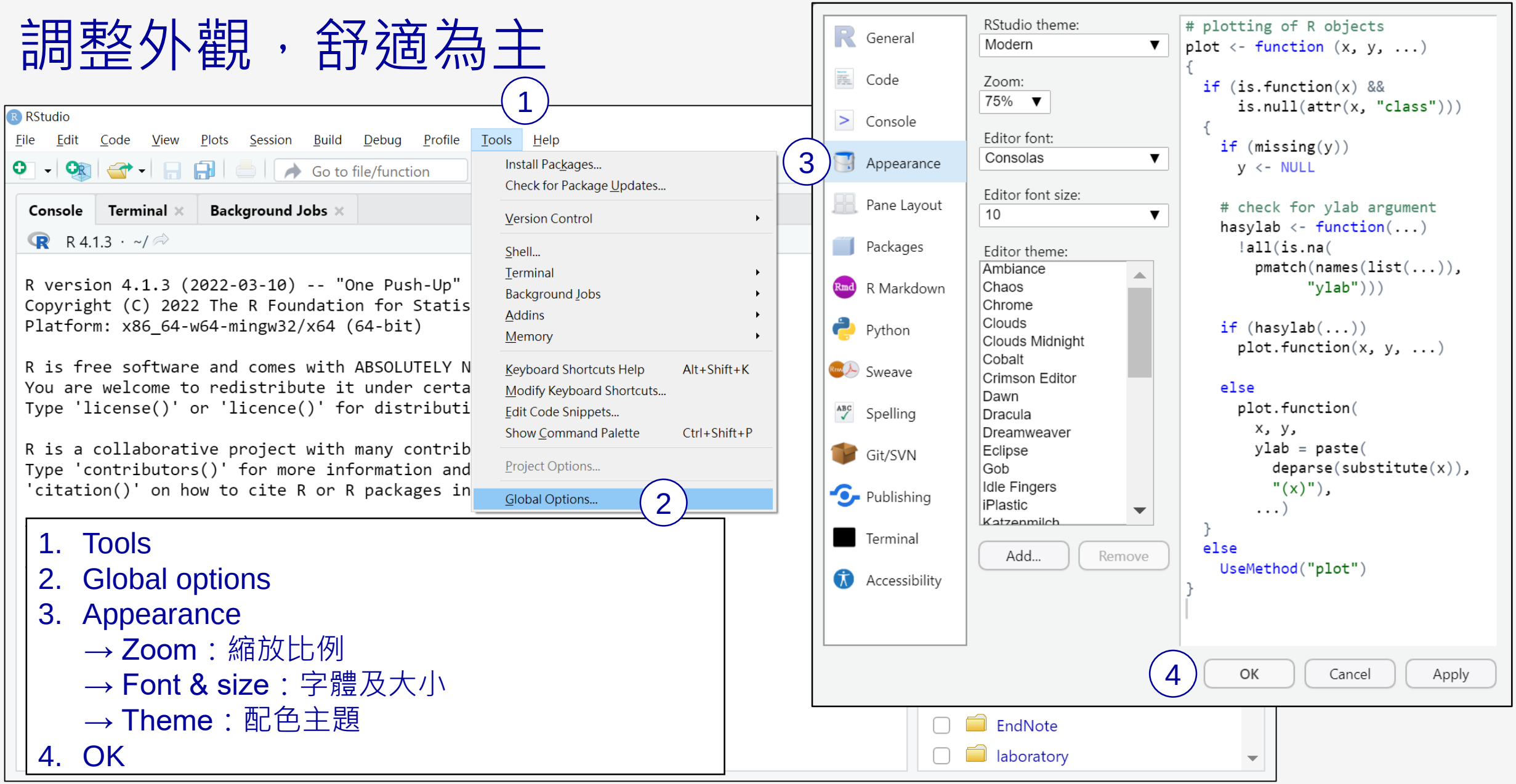
軟體安裝

- Windows使用者全部都要對安裝檔點右鍵「以系統管理員身分執行」
- R
  - 全部使用預設並點選「下一步」
- RTools
  - 全部使用預設並點選「Next」
- RStudio
  - 全部使用預設並點選「下一步」

開啟RStudio 跟著一起作！

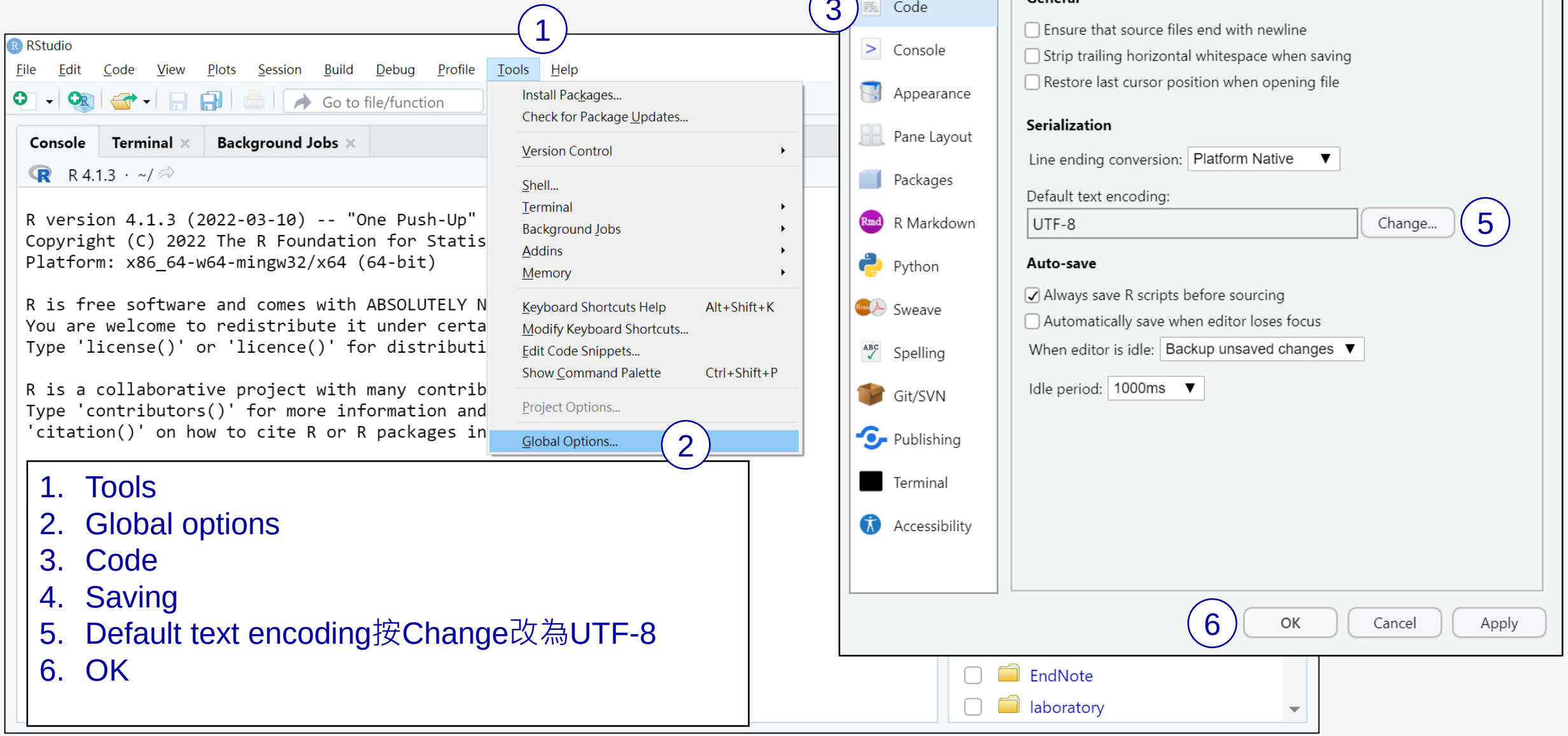


調整外觀，舒適為主



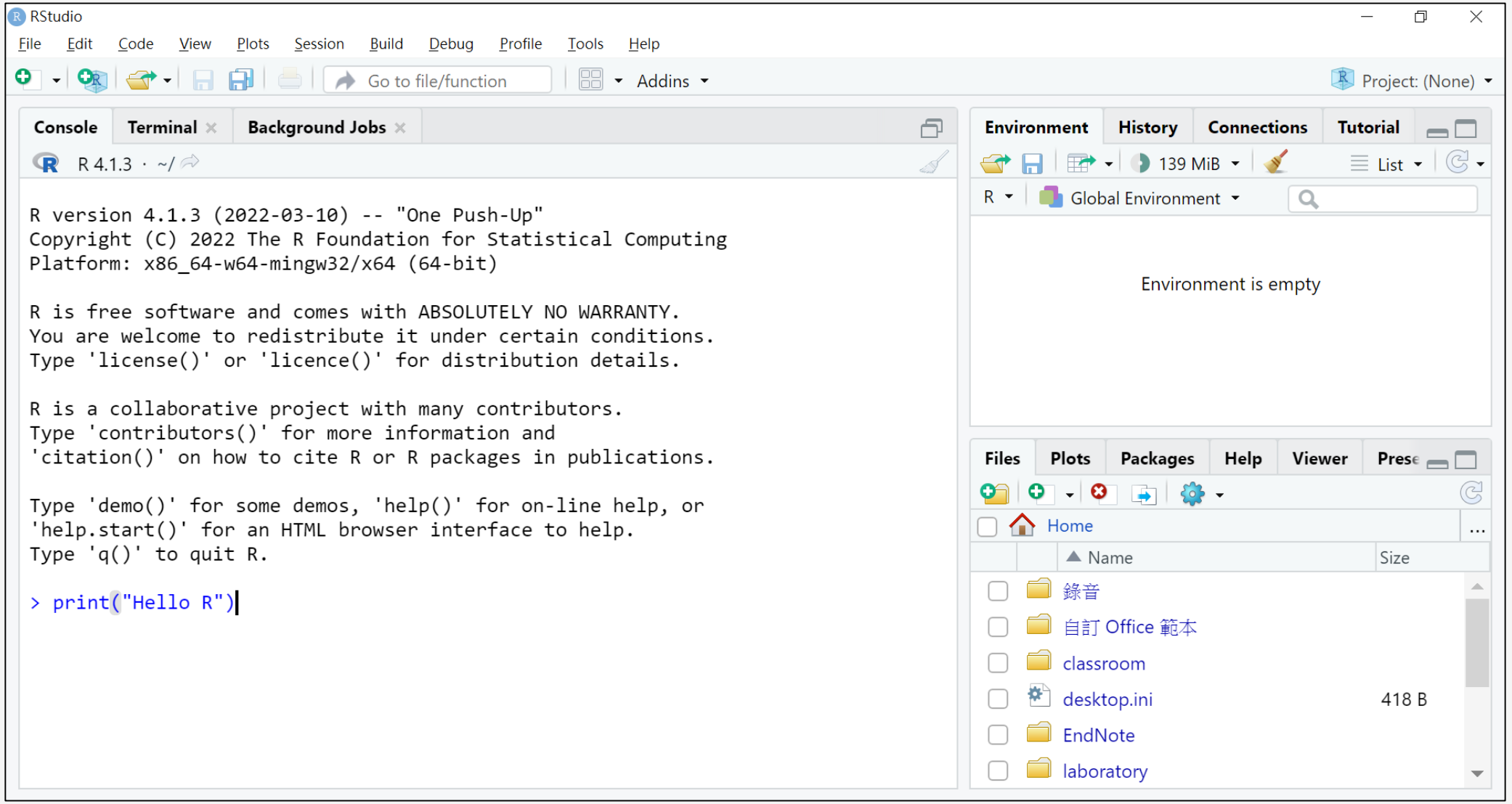
1. Tools
2. Global options
3. Appearance
   → Zoom：縮放比例
   → Font & size：字體及大小
   → Theme：配色主題
4. OK

調整程式檔案文字編碼

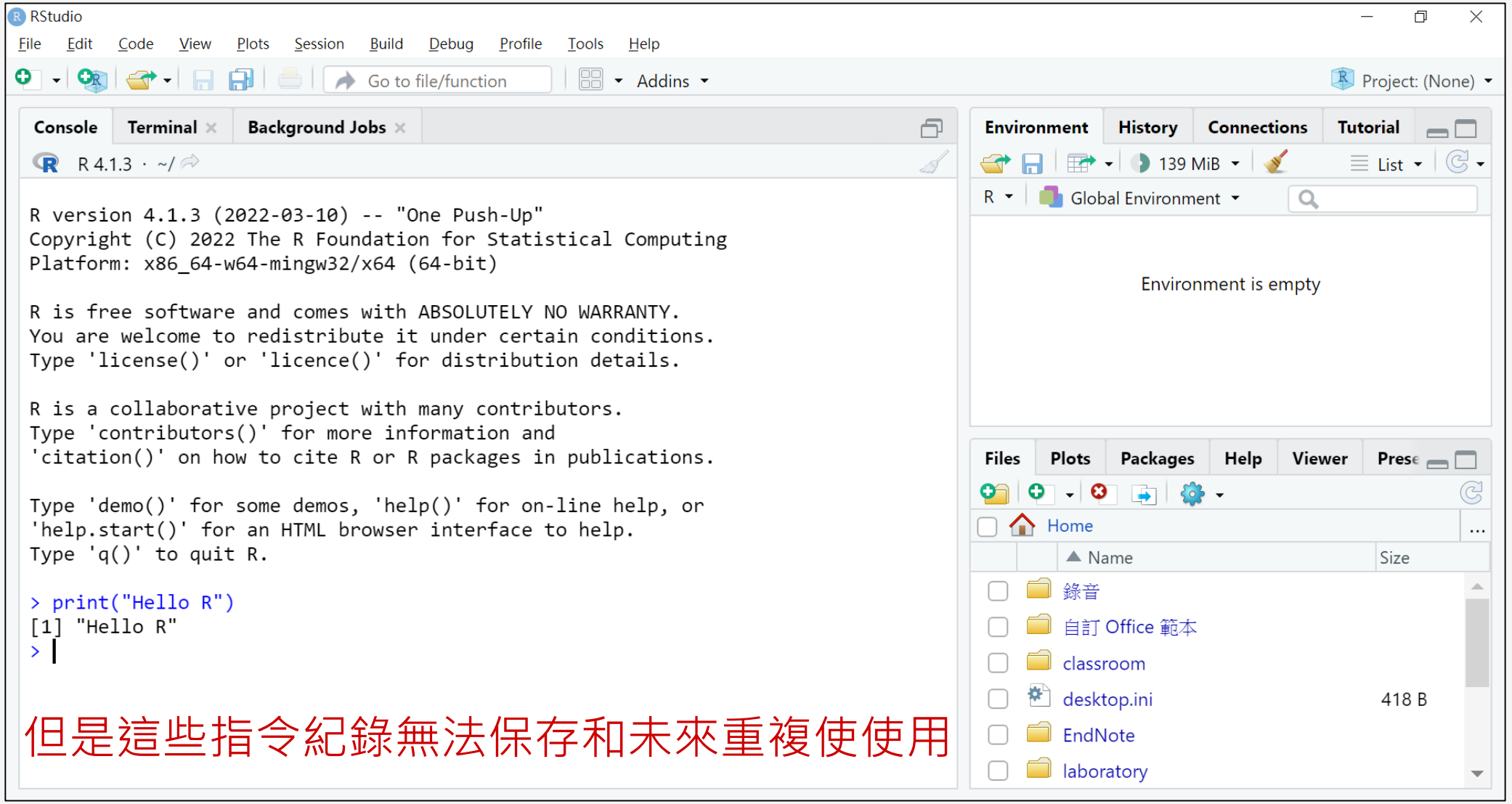


1. Tools
2. Global options
3. Code
4. Saving
5. Default text encoding按Change改為UTF-8
6. OK

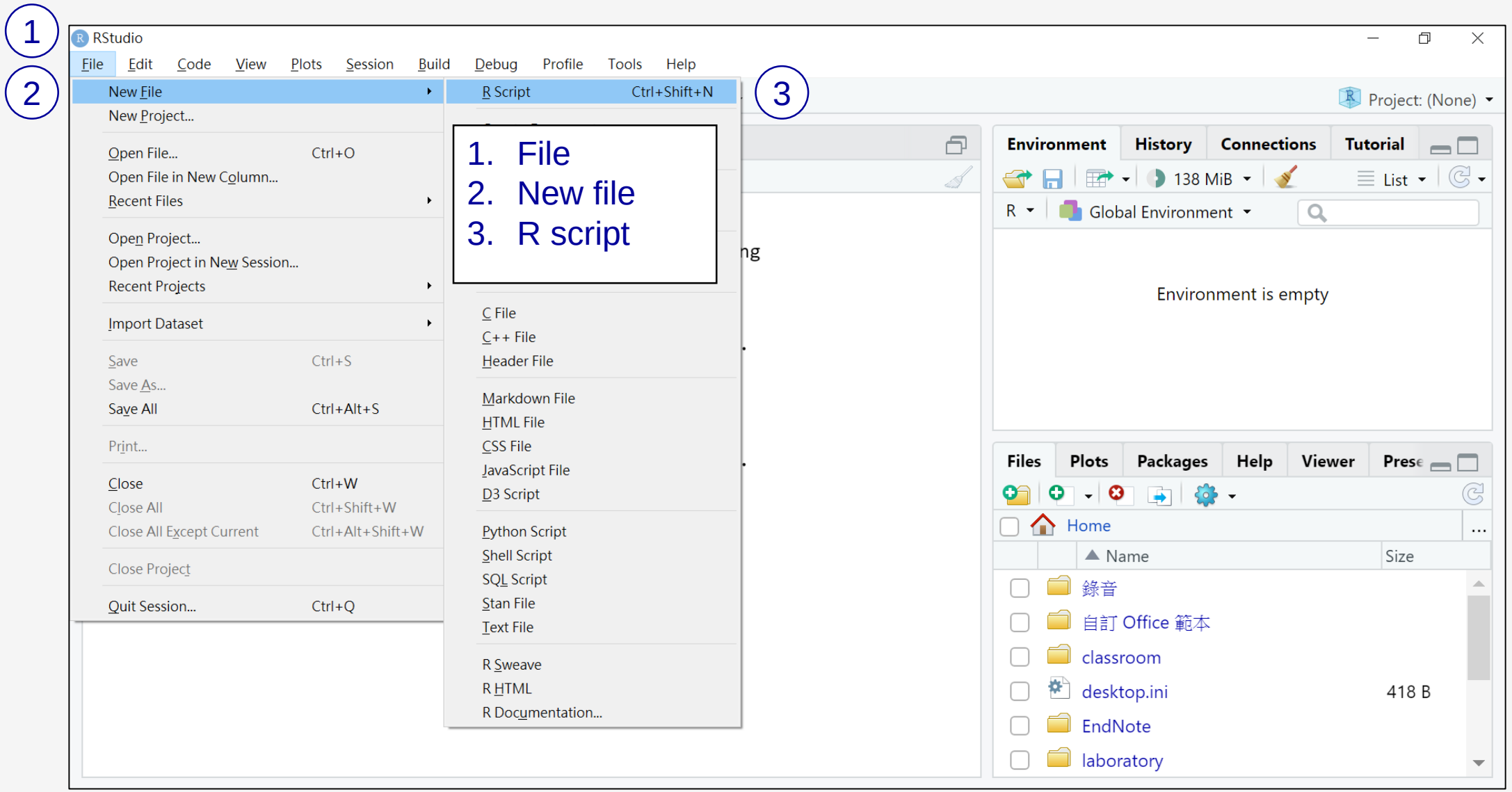
在Console區域輸入指令，按Enter



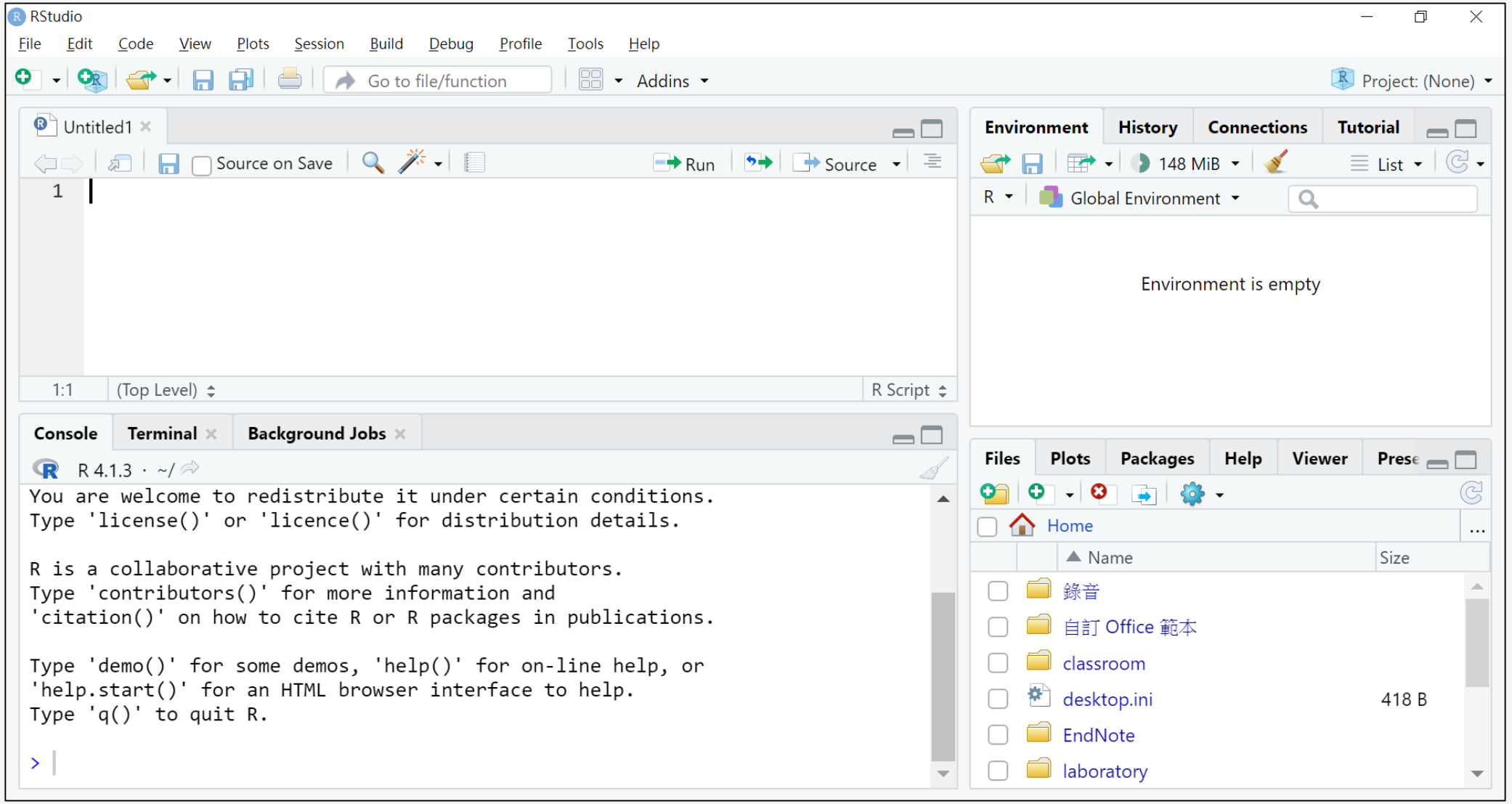
在Console區域得到輸入指令的執行 / 回傳結果



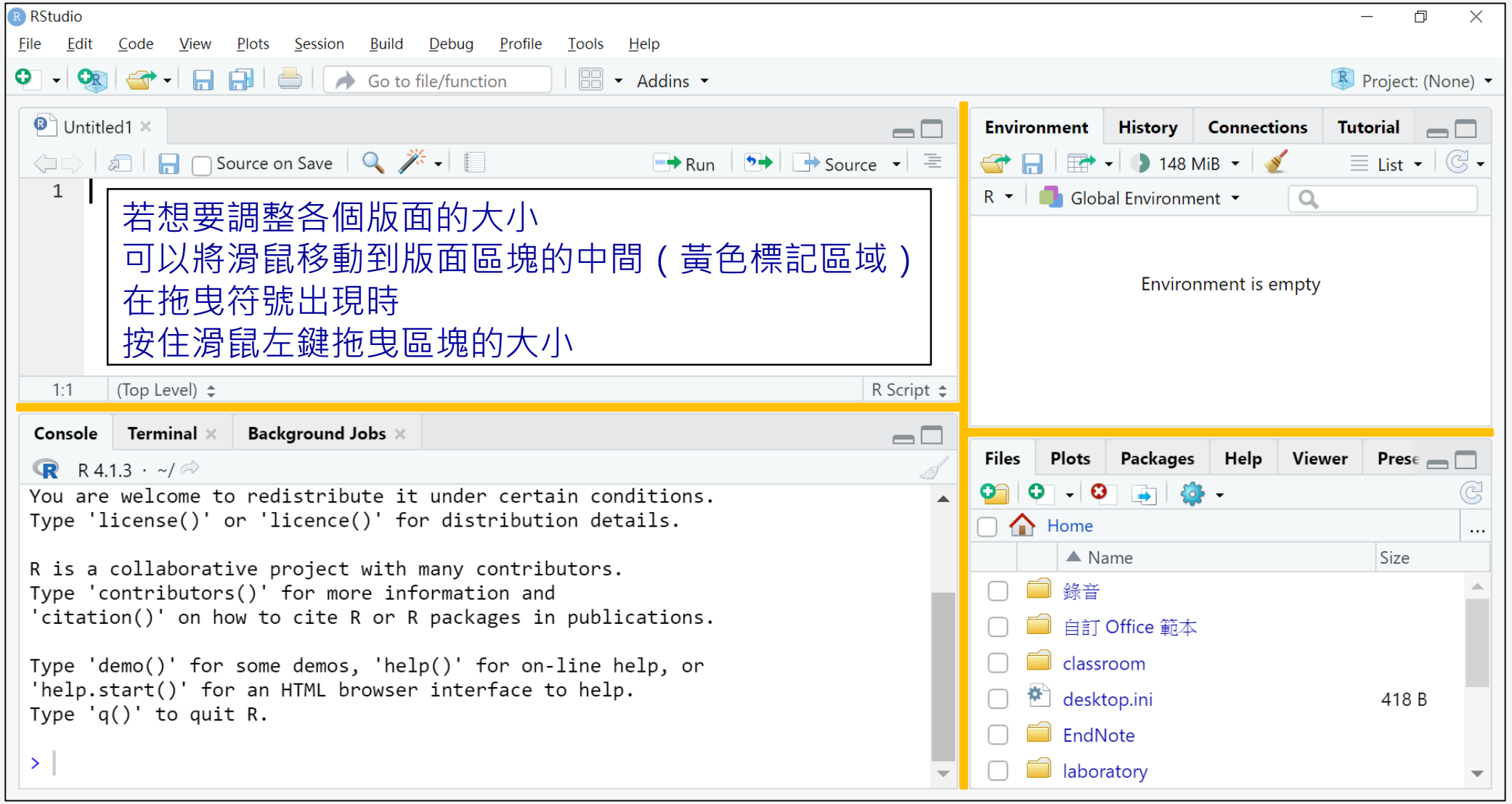
開啟指令稿



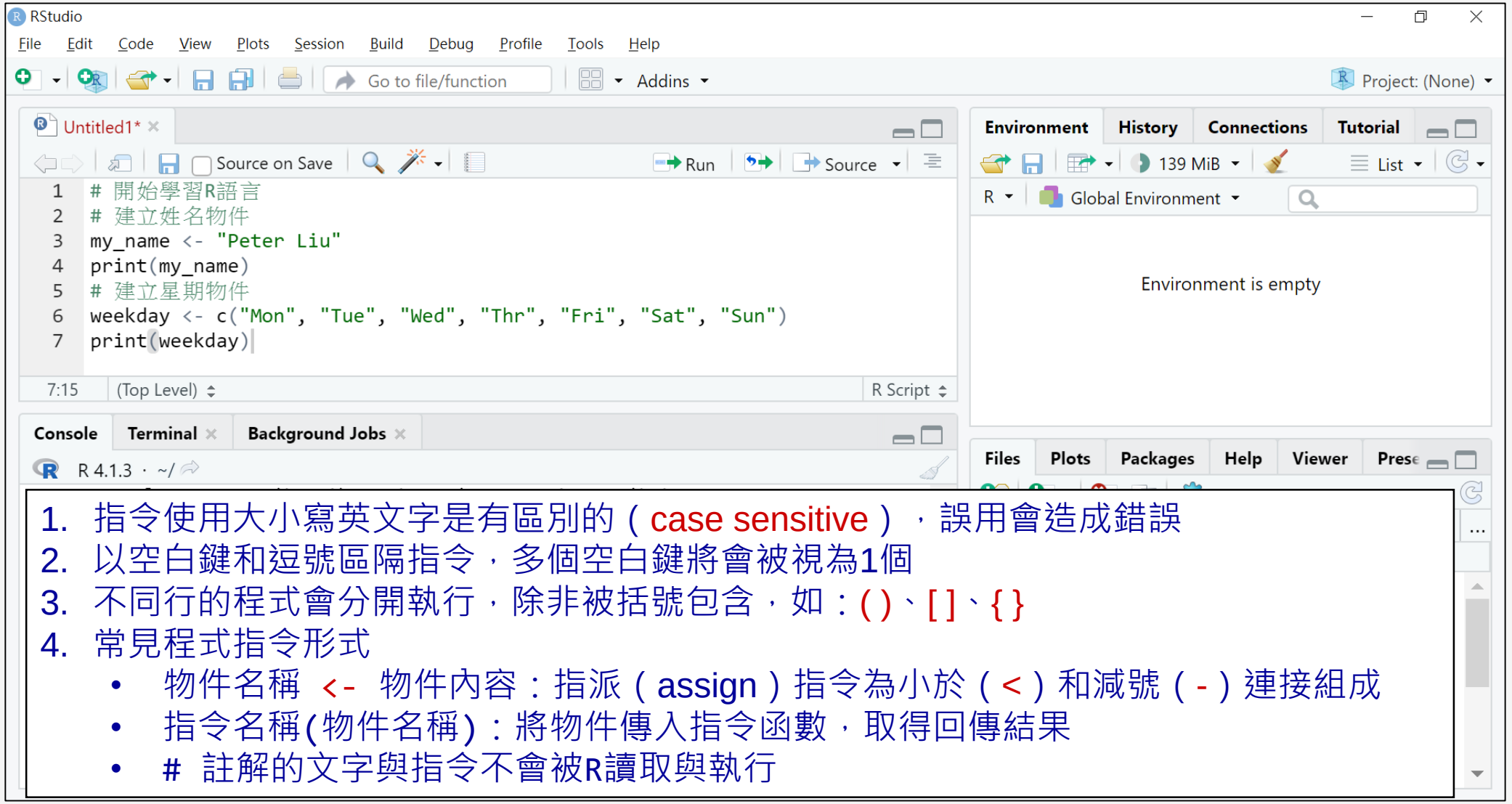
開啟指令稿完成



版面區塊調整

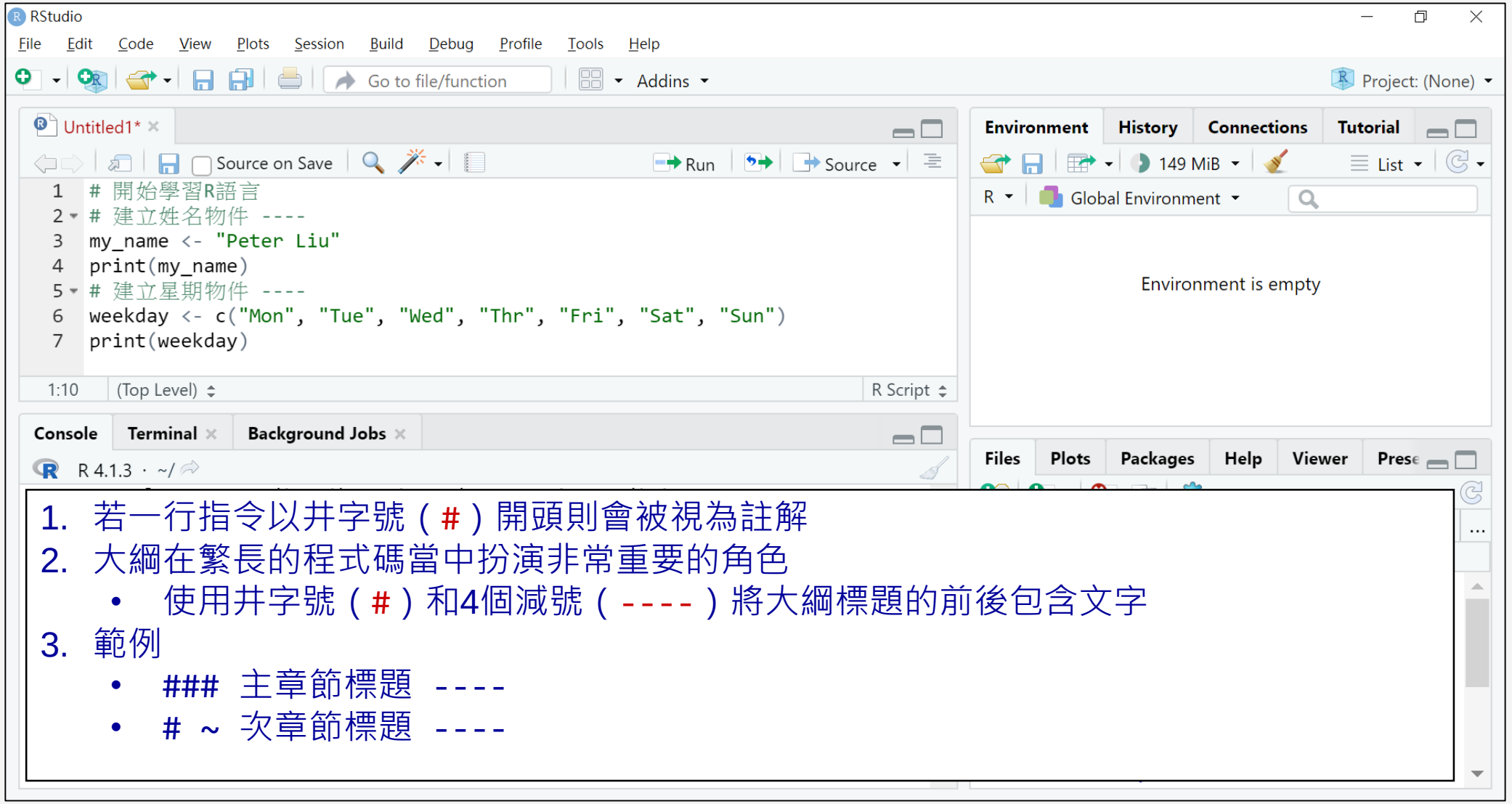


編輯指令稿內容

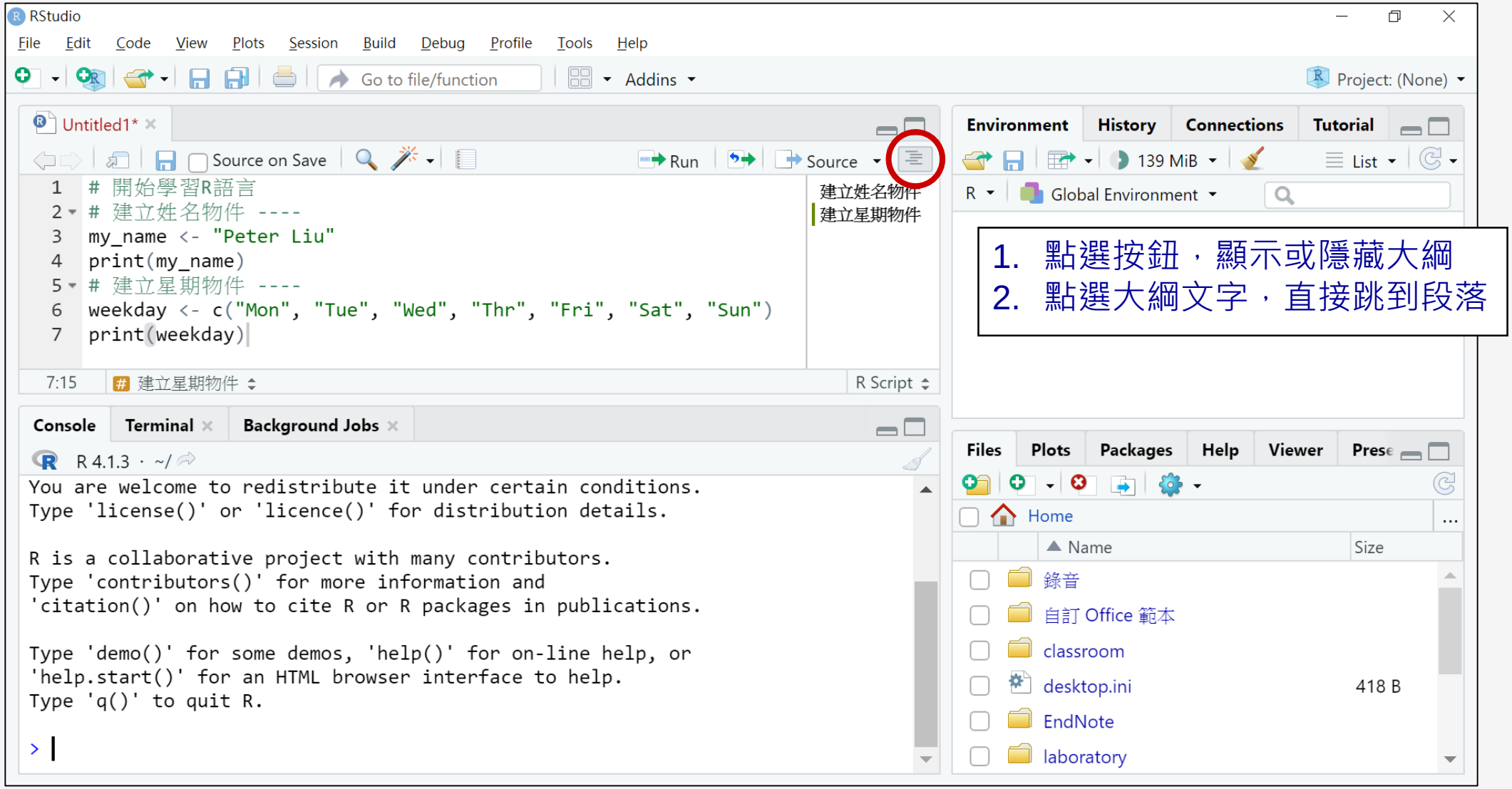


1. 指令使用大小寫英文字是有區別的（case sensitive），誤用會造成錯誤
2. 以空白鍵和逗號區隔指令，多個空白鍵將會被視為1個
3. 不同行的程式會分開執行，除非被括號包含，如：( )、[ ]、{ }
4. 常見程式指令形式
   - 物件名稱 <- 物件內容：指派（assign）指令為小於（<）和減號（-）連接組成
   - 指令名稱(物件名稱)：將物件傳入指令函數，取得回傳結果
   - # 註解的文字與指令不會被R讀取與執行

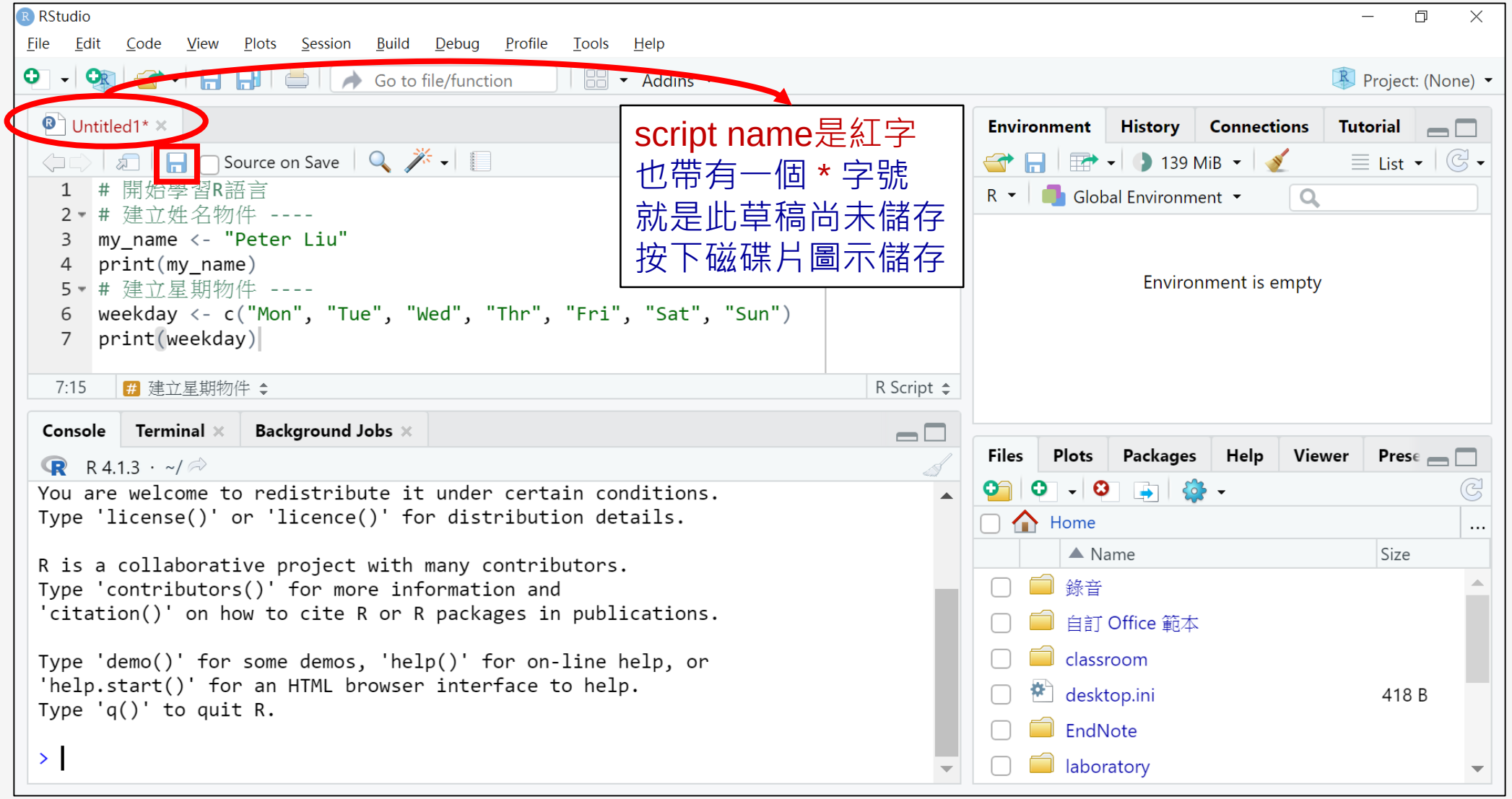
編輯指令稿註解與大綱



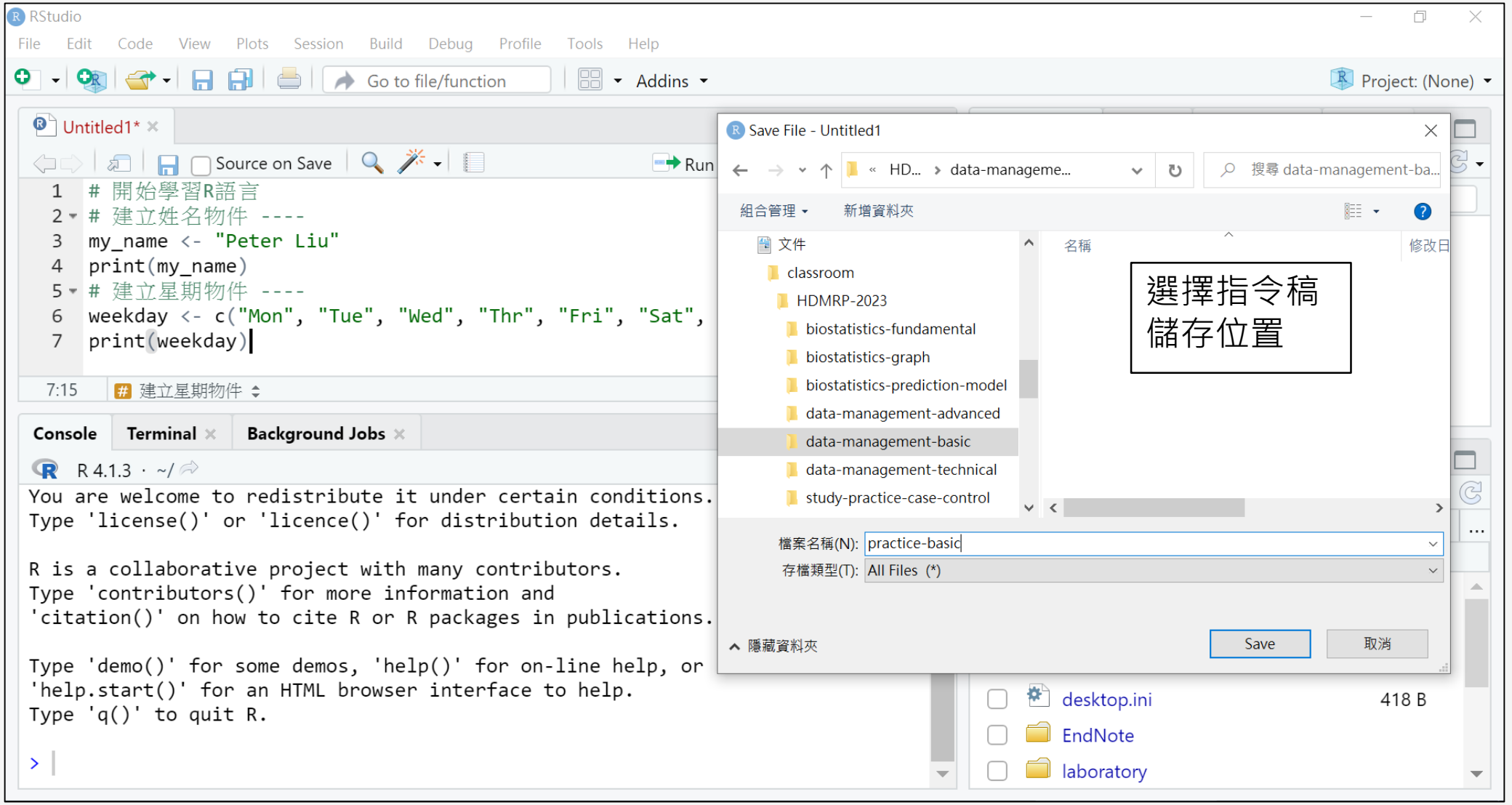
編輯指令稿註解與大綱



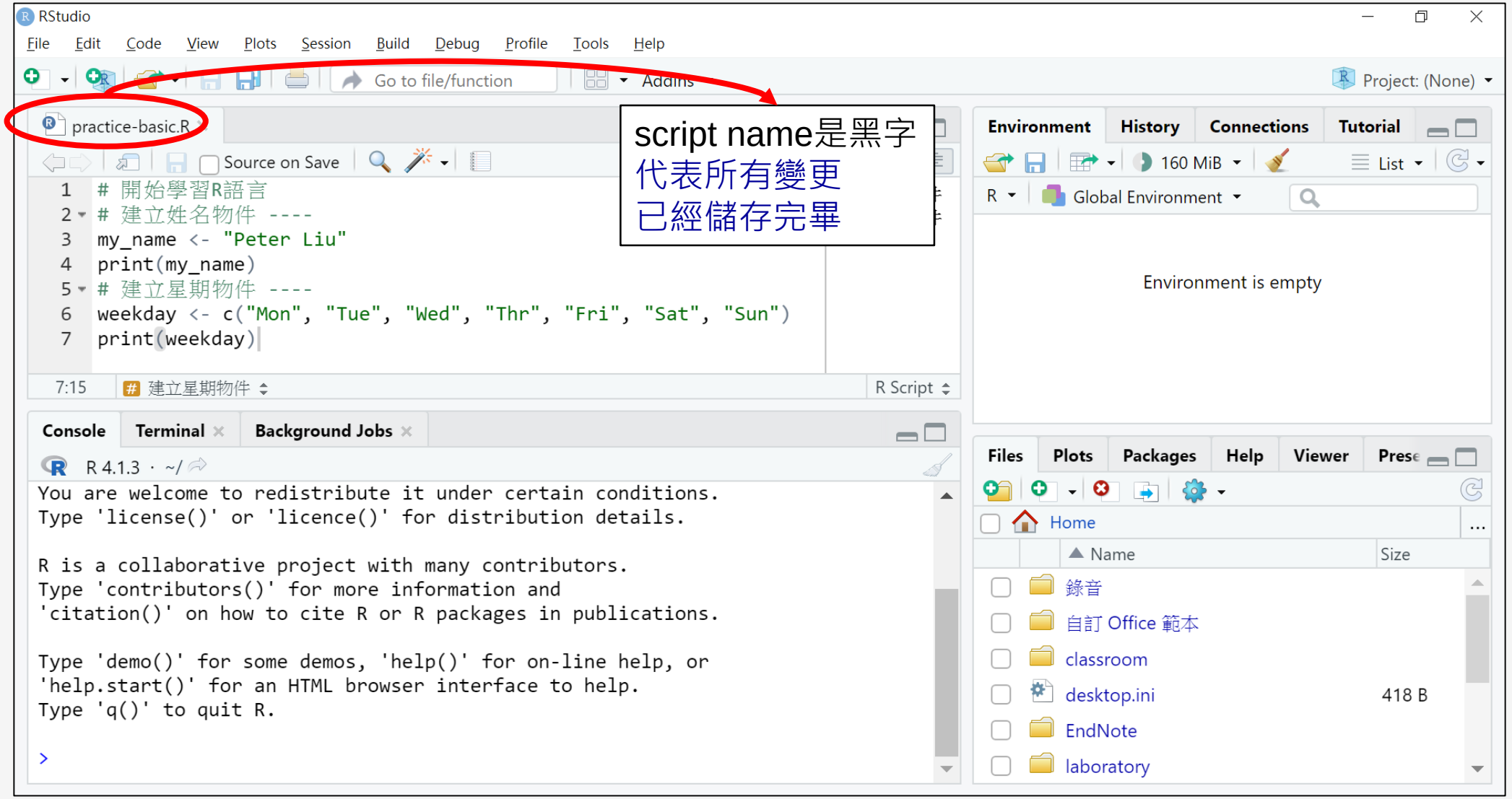
儲存指令稿



儲存指令稿

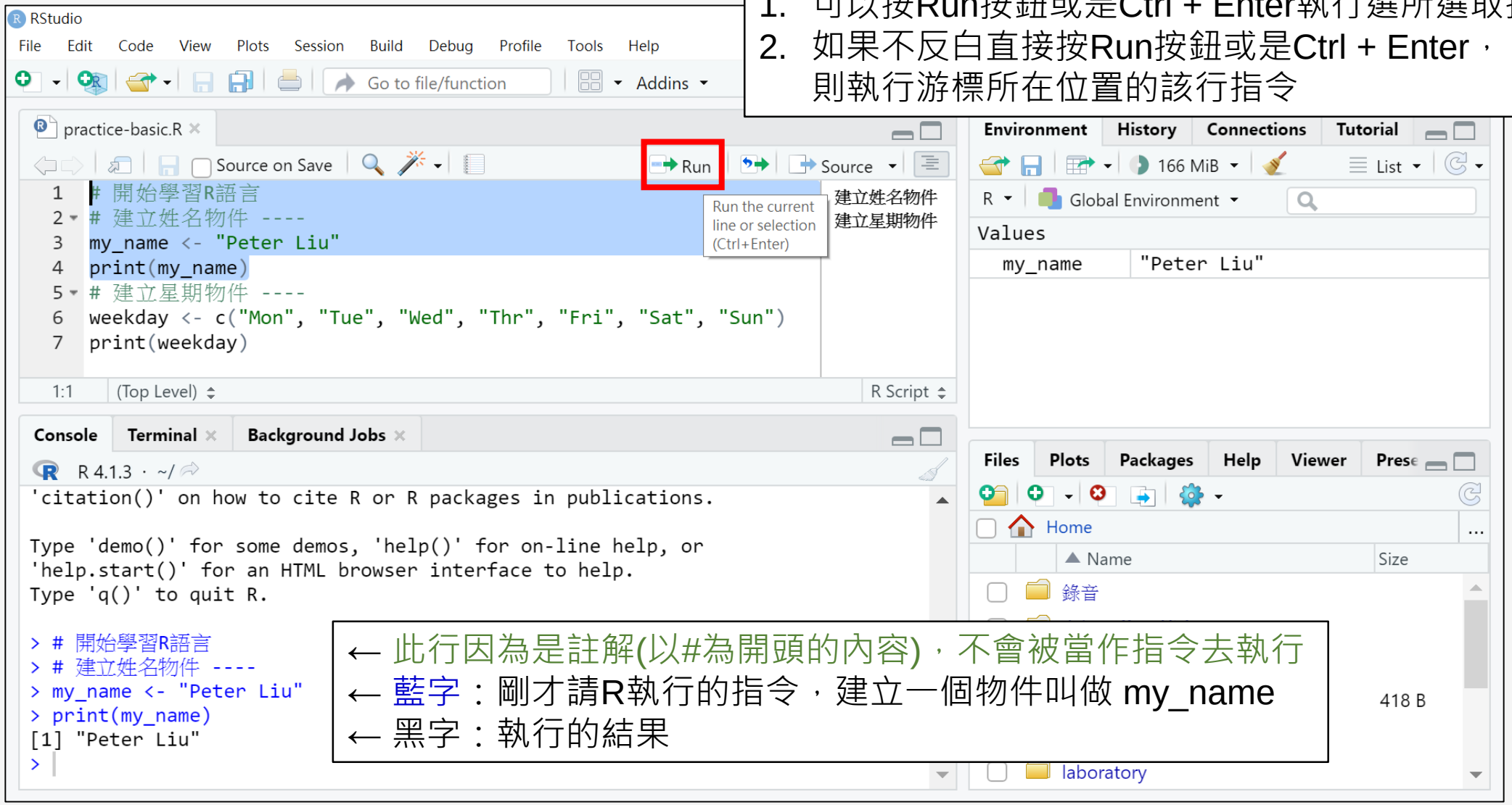


儲存指令稿完成



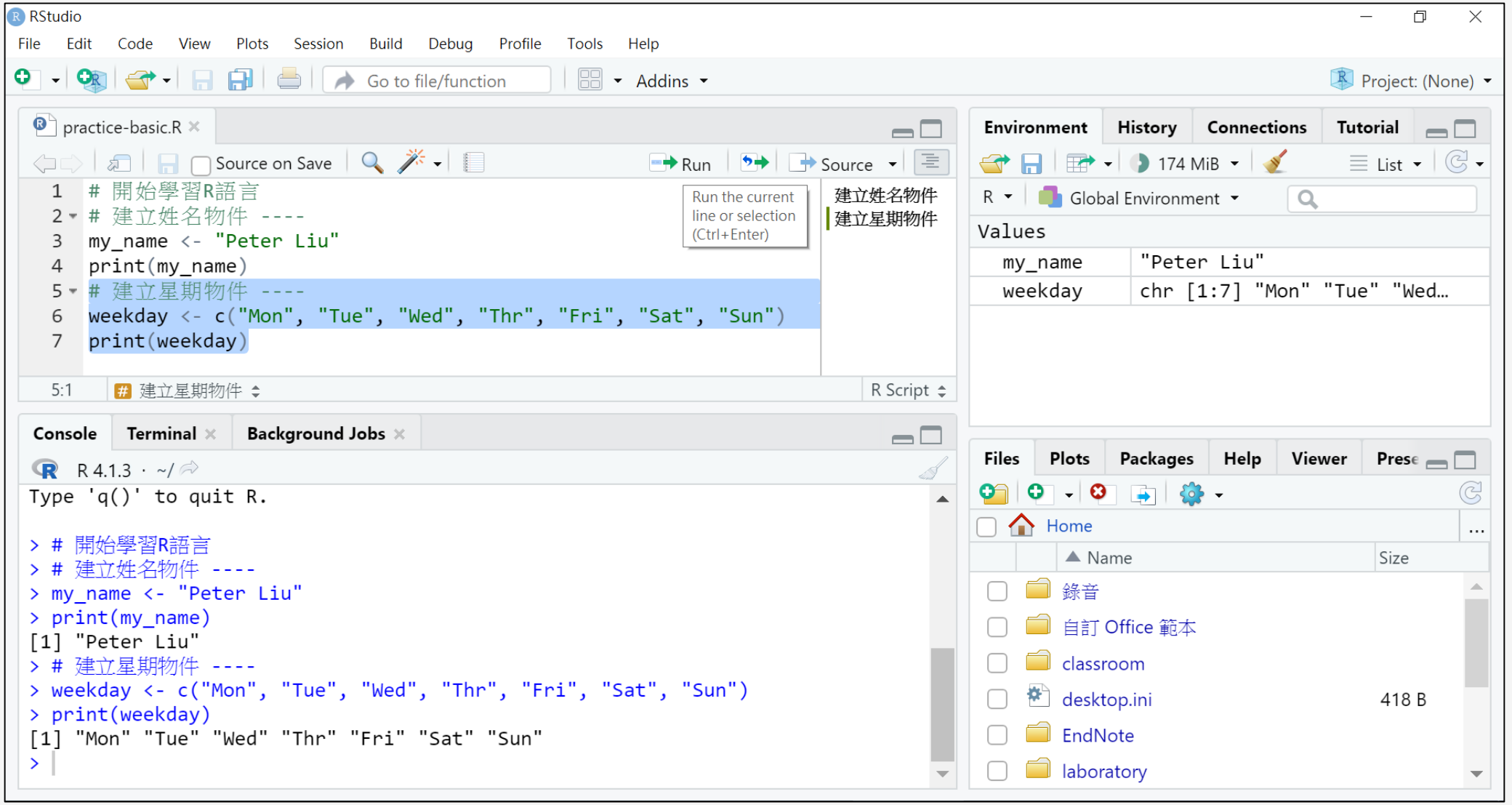
選取（反白）部分指令稿執行

1. 可以按Run按鈕或是Ctrl + Enter執行選所選取指令
2. 如果不反白直接按Run按鈕或是Ctrl + Enter，則執行游標所在位置的該行指令



← 此行因為是註解(以#為開頭的內容)，不會被當作指令去執行
← 藍字：剛才請R執行的指令，建立一個物件叫做 my_name
← 黑字：執行的結果

選取（反白）部分指令稿執行



R軟體實作

- 套件與函數
- 資料管理
- 實作練習

套件與函數的關係

- 套件（package）：包含一個或多個函數，通常具有類似意義或用途
  - `Survival`：存活分析套件

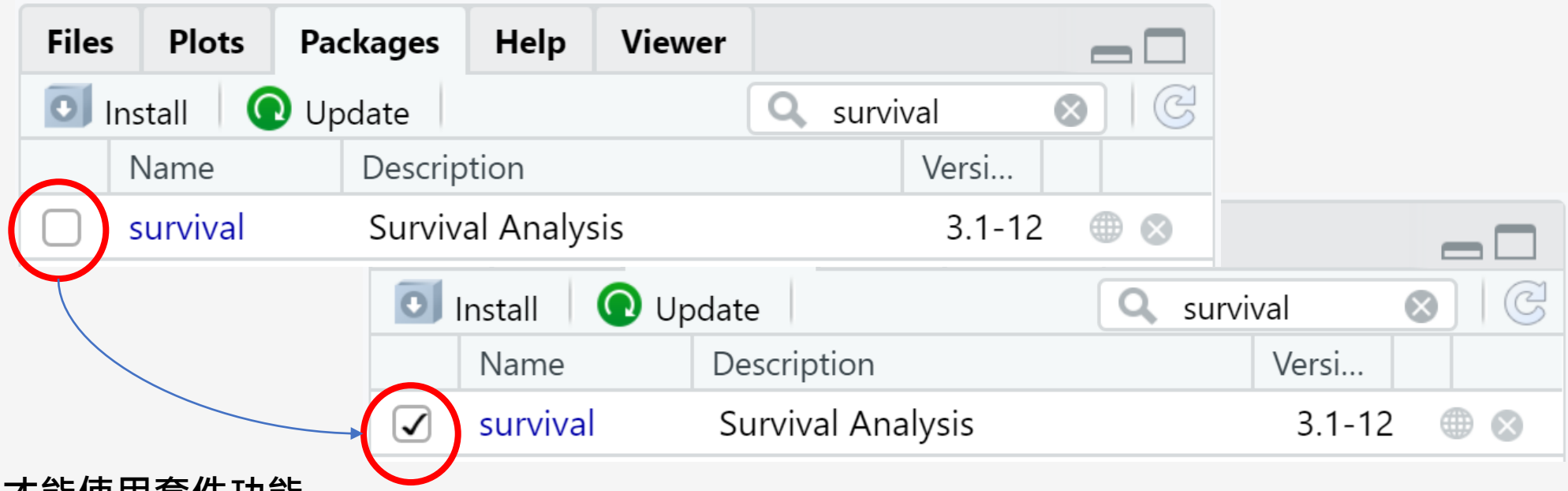
- 函數（function）：專一用途的指令代碼（code），注意要加`()`
  - `Surv()`：定義存活變數

    `Surv(觀察時間, 事件發生)`
  - `coxph()`：擬和存活函數

    `coxph(Surv(觀察時間, 事件發生) ~ 自變數, 使用資料)`

套件載入

```
# 若需要使用的套件並非預設開啟後即載入的套件
# 使用 library 指令載入記憶體 每次重新打開R都要執行
library(survival)
```



成功載入之後才能使用套件功能
Packages面板會呈現打勾的狀態

函數使用說明：?Surv

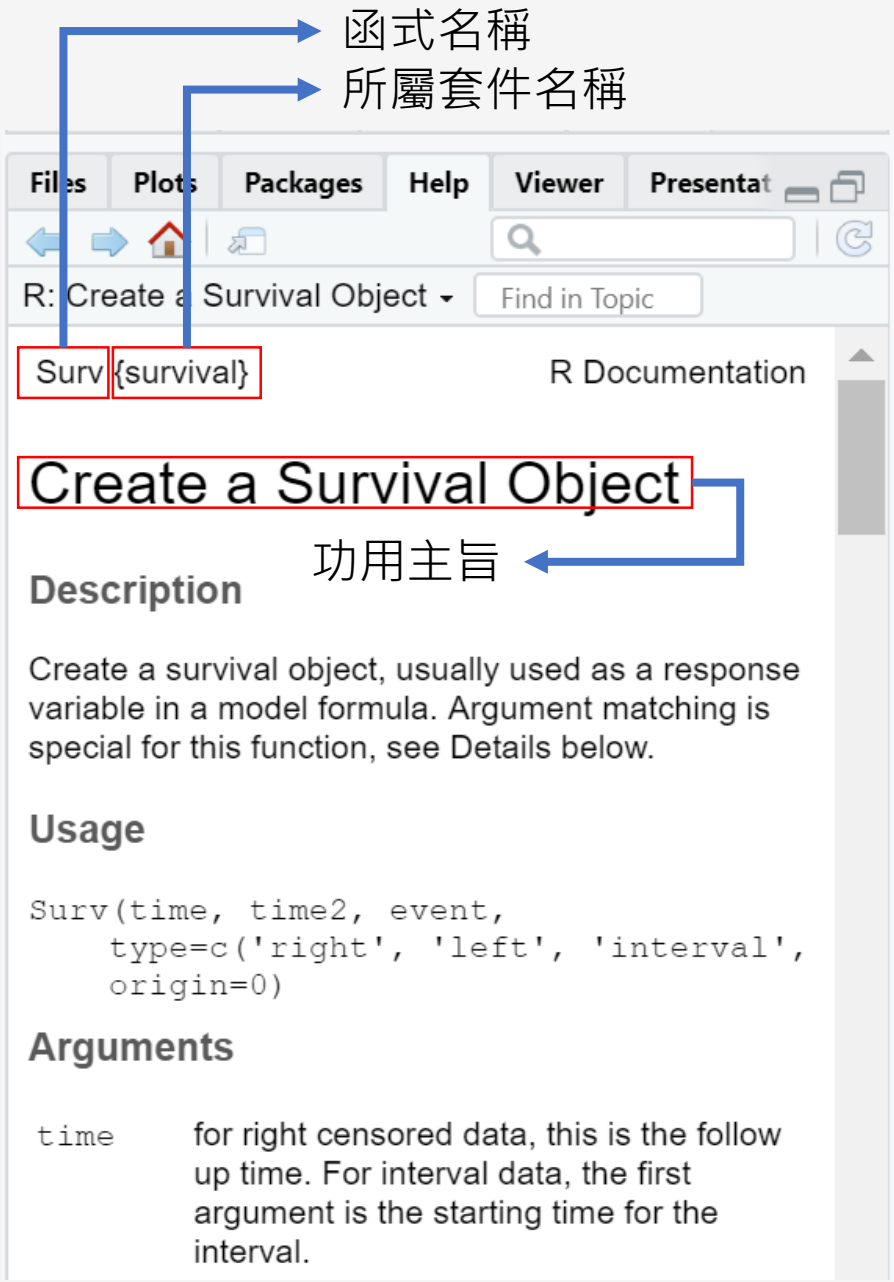
- 將正確的參數內容寫在正確的參數名稱裡面
- 參數 argument：執行的條件
  - `time`

    觀察時間，沒有預設值，必須填入否則無法執行
  - `event`

    事件發生，沒有預設值，必須填入否則無法執行
  - `type = c('right', 'left', 'interval')`

    指定截斷（censoring）的方式，必須是選項之一
  - `origin = 0`

    有預設值，亦可以修改



函數使用範例：Surv() + coxph()的撰寫

```
# 使用資料 dt
# 觀察時間 day
# 事件發生 death

# 完整寫出參數名稱與參數內容
# 使用dt資料集的day和event為存活狀態變數，treated為自變數
coxph(
  formula = Surv(time = day, event = death) ~ treated,
  data = dt)

# 簡化
coxph(Surv(day, death) ~ treated, dt)
```

函數使用結合：Surv() + coxph()的撰寫

```
# 設定一個存活函數
# 將此存活函數與變數擬和
# 模型資訊儲存為物件 mf
mf <- coxph(Surv(day, death) ~ treated, data = dt)

# 檢視模型報表
summary(mf)

# 繪製存活分析圖
ggsurvplot(mf)
```

套件下載，先確認有連上網路

```
# 下載一個套件
install.packages("data.table")

# 下載多個套件
install.packages(c("ggplot2", "lubridate"))

# 下載多個套件
need_packages <- c("ggplot2", "lubridate")

install.packages(need_packages)
```

常見問題：無法下載套件

```
# 沒有連結網路
> install.packages("data.table")

  trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.x/data.table_x.zip'

  Warning in install.packages :
    InternetOpenUrl failed: '無法解析伺服器名稱或位址'

  Error in download.file(url, destfile, method, mode = "wb", ...) :
    cannot open URL
    'https://cran.rstudio.com/bin/windows/contrib/4.x/data.table_x.zip'

  Warning in install.packages :
    download of package 'data.table' failed

# 趕快讓電腦連上網路吧
```

常見問題：無法下載套件

```
# 拼字錯誤，正確應為data.table，但是少打了中間的 .
> install.packages("datatable")

  將程式套件安載入
   'C:/Users/liu/Documents/R/win-library/4.1' (因為'lib'沒有被指定)

  Warning in install.packages :
    package ‘datatable’ is not available
    for this version of R

# 記得R指令的大小寫和符號必須一字不差
> install.packages("data.table")
```

常見問題：無法下載套件

```
# Windows使用者名稱為中文
> install.packages("data.table")

  "C:/Users/??"：檔案名稱、目錄名稱或磁碟區標籤語法錯誤

# 請關閉RStudio以及所有應用程式
# 把電腦使用者名稱修改為英文數字之後重新開機
# 再次打開R軟體
```

常見問題：無法下載套件

```
# Windows使用者的OneDrive資料夾"文件"是中文
> install.packages("data.table")
  "C:/Users/peter/OneDrive/??"：檔案名稱、目錄名稱或磁碟區標籤語法錯誤

# 在OneDrive以外之處新增一個資料夾名為 user_packages
# 複製這個資料夾的完整路徑，例如："C:\Users\peter\user_packages"
# Windows原本的路徑分隔斜線為"\"，要改為"/"
# 輸入指令修改下載資料夾為你剛才新創的資料夾
> .libPaths("C:/Users/peter/user_packages")

# 輸入指令確認套件資料夾應該要有兩個
> .libPaths()
  [1] "C:/Users/peter/user_packages"        # 下載套件
  [2] "C:/Program Files/R/R-4.2.3/library" # 內建套件
```

常見問題：無法載入套件

```
# 拼字錯誤
> library(DATA.TABLE)

  Error in library(gee) : 不存在叫 'DATA.TABLE' 這個名稱的套件



# 正確應為小寫data.table，請修正
> library(data.table)

  data.table 1.14.8 using 2 threads (see ?getDTthreads).
  Latest news: r-datatable.com
```

常見問題：無法載入套件

```
# 尚未下載
> library(ggplot2)
Error in library(ggplot2) : 不存在叫 'ggplot2' 這個名稱的套件

# 請執行install.packages(“ggplot2”)指令下載套件
> install.packages("ggplot2")

  package ‘ggplot2’ successfully unpacked and MD5 sums checked

# 再次載入套件
> library(ggplot2)
```

資料的產生：真實世界

- 2月18日深夜

- 一名8歲男性兒童由父母帶入急診

- CRIES分數為6分

- 經診斷為急性闌尾炎（acute appendicitis）

CRIES, Children Revised Impact of Event Scale

資料的儲存：樣態與編碼

- 結構化資料表（data table）
- 譯碼簿（codebook）

欄 / column / 變項 / variable

列 / row / 觀察值 / observation

id	date	male	age	pain	diagnosis
S1911	02-18	1	8	6	K35
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

變項名稱	中文意義	資料類型	編碼方式
id	身分證號	文字	S+四位數字
date	就醫日期	日期	mm-dd
male	男性	數值	1 = 男性；0 = 女性
age	年齡	數值	單位：歲
pain	疼痛指數	數值	CRIES量表分數
diagnosis	主診斷	文字	ICD-10-CM編碼

資料讀取：開啟目標資料（放入電腦記憶體）

```
# 路徑
path_data <- "C:/Users/liu/data-management-basic"

# 資料讀取後才可以進行分析與操作
setwd(path_data)
dt_1 <- fread("dt_1.csv")
```

	patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
1	A1	2014-02-04	Huang	120	88	0	NA
2	A1	2014-03-04	Huang	121	94	1	433
3	A1	2014-04-01	Huang	116	92	0	250
4	A1	2014-04-29	Lee	120	94	1	NA
5	A1	2014-05-27	Hsu	121	90	1	250
6	A1	2014-06-24	Lee	118	92	0	NA

資料讀寫：記憶體中的資料在RStudio關閉會消失

- 常見資料格式與R軟體讀取 / 寫出資料指令

資料格式	讀取資料		寫出資料	
	套件	函數	套件	函數
Excel	`xlsx`	`read.xlsx`	`xlsx`	`write.xlsx`
CSV	`data.table`	`fread`	`data.table`	`fwrite`
SAS	`haven`	`read_sas`	`haven`	`write_sas`
Stata	`haven`	`read_dta`	`haven`	`write_dta`
SPSS	`haven`	`read_sav`	`haven`	`write_sav`

資料檢視：理解資料的輪廓

```
# 資料表最前6列
head(dt_1)
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
A1	2014-02-04	Huang	120	88	0	NA
A1	2014-03-04	Huang	121	94	1	433
A1	2014-04-01	Huang	116	92	0	250
A1	2014-04-29	Lee	120	94	1	NA

```
# 變項摘要統計
summary(dt_1)
```

```
 patient_id          visit_date           physician              sbp             dbp          dose_modify    comorbidity
Length:60          Min.   :2014-02-03   Length:60          Min.   :112.0   Min.   :80.00   Min.   :0.0   Min.   :250.0
Class :character   1st Qu.:2014-04-03   Class :character   1st Qu.:118.0   1st Qu.:88.00   1st Qu.:0.0   1st Qu.:250.0
Mode  :character   Median :2014-06-11   Mode  :character   Median :120.0   Median :90.00   Median :0.0   Median :433.0
                   Mean   :2014-06-11                      Mean   :120.0   Mean   :90.07   Mean   :0.3   Mean   :351.7
                   3rd Qu.:2014-08-21                      3rd Qu.:122.2   3rd Qu.:92.00   3rd Qu.:1.0   3rd Qu.:433.0
                   Max.   :2014-10-17                      Max.   :127.0   Max.   :97.00   Max.   :1.0   Max.   :480.0
                                                                                                         NA's   :28
```

data.table概論

- data.table是非常好用的資料管理套件
  - R使用者必須要學習
- 請用class(dt)確認dt物件具有data.table屬性
  - 讀取、檢視、篩選、排序、修改、歸戶、合併
- dt[i, j, by]
  - i：選擇哪一些觀察值子集？
  - j：操作哪一些欄位(提取/建立/更新/整合計算)？
  - by：前述的處理動作要依照什麼欄位進行分組處理？

https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html

資料篩選：取得想要的觀察值《指定條件》

```
# physician變項的內容為 Lee
ot <- dt[physician == "Lee"]
```

patient_id	visit_date	physician	sbp	dbp
A1	2014-02-04	Huang	120	88
A1	2014-03-04	Huang	121	94
A1	2014-04-01	Huang	116	92
A1	2014-04-29	Lee	120	94
A1	2014-05-27	Hsu	121	90
A1	2014-06-24	Lee	118	92

patient_id	visit_date	physician	sbp	dbp
A1	2014-04-29	Lee	120	94
A1	2014-06-24	Lee	118	92
A1	2014-07-22	Lee	115	91
A1	2014-10-14	Lee	126	92
A3	2014-02-07	Lee	121	88
A3	2014-04-04	Lee	123	90

資料篩選：取得想要的觀察值《指定條件》

```
# physician變項的內容"不為" Lee
ot <- dt[!(physician == "Lee")]

# physician變項的內容"不為" Lee
ot <- dt[physician != "Lee"]
```

資料篩選：取得想要的觀察值《比大小》

```
# sbp變項的內容為大於等於120
ot <- dt [120 <= sbp]
```

patient_id	visit_date	physician	sbp	dbp
A1	2014-02-04	Huang	120	88
A1	2014-03-04	Huang	121	94
A1	2014-04-01	Huang	116	92
A1	2014-04-29	Lee	120	94
A1	2014-05-27	Hsu	121	90
A1	2014-06-24	Lee	118	92

patient_id	visit_date	physician	sbp	dbp
A1	2014-02-04	Huang	120	88
A1	2014-03-04	Huang	121	94
A1	2014-04-29	Lee	120	94
A1	2014-05-27	Hsu	121	90
A1	2014-08-19	Huang	124	89
A1	2014-09-16	Huang	124	90

資料篩選：取得想要的觀察值《數值模式》

```
ot <- dt[inrange(x = dbp, lower = 80, upper = 90)]
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  80.00   87.75   88.50   88.11   90.00   90.00
```

```
ot <- dt[(80 <= dbp) & (dbp <= 90)]
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  80.00   87.75   88.50   88.11   90.00   90.00
```

```
ot <- dt[(80 < dbp) & (dbp < 90)]
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  85.00   86.00   88.00   87.58   89.00   89.00
```

資料篩選：取得想要的觀察值《字串模式》

```
# 找出physician字串裡面含有H的觀察值
ot <- dt[grepl(pattern = "H", x = physician)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
A1	2014-02-04	Huang	120	88	0	NA
A1	2014-03-04	Huang	121	94	1	433
A1	2014-04-01	Huang	116	92	0	250
A1	2014-05-27	Hsu	121	90	1	250
A1	2014-08-19	Huang	124	89	0	NA
A1	2014-09-16	Huang	124	90	0	433

資料篩選：取得想要的觀察值《字串模式》

```
# grepl + regexp 基本型態 ^ 代表開頭
ot <- dt[grepl("^H", physician)]
```

patient_id	visit_date	physician
A1	2014-02-04	Huang
A1	2014-03-04	Huang
A1	2014-04-01	Huang
A1	2014-05-27	Hsu

```
# grepl + regexp 基本型態 $ 代表結尾
Ot <- dt[grepl("g$", physician)]
```

patient_id	visit_date	physician
A1	2014-02-04	Huang
A1	2014-03-04	Huang
A1	2014-04-01	Huang
A1	2014-08-19	Huang

資料篩選：取得想要的觀察值《字串模式》

```
# 找出comorbidity字串裡面為433開頭或是434開頭的觀察值
ot <- dt [grepl("^433|^434", comorbidity)]

# 找出comorbidity字串裡面為43開頭，後面接著3或4的觀察值
ot <- dt [grepl("^43[34]",   comorbidity)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
A1	2014-03-04	Huang	121	94	1	433
A1	2014-09-16	Huang	124	90	0	433
A1	2014-10-14	Lee	126	92	0	433
A2	2014-03-07	Huang	127	90	0	434
A3	2014-08-22	Huang	120	89	0	434
A5	2014-08-18	Hsu	118	90	0	434

資料篩選：取得想要的變項

```
# 在j索引處寫一個句點，一對小括號，裡面寫入想要留下的變項名稱
ot <- dt[, .(patient_id, visit_date, comorbidity)]
```

patient_id	visit_date	comorbidity
A1	2014-02-04	NA
A1	2014-03-04	433
A1	2014-04-01	250
A1	2014-04-29	NA
A1	2014-05-27	250
A1	2014-06-24	NA

資料排序：讓資料成為想要的順序

```
# 依照就醫日期排序，預設為遞增（ascending）排序
ot <- ot[order(visit_date)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
A5	2014-02-03	Huang	120	88	1	NA
A1	2014-02-04	Huang	120	88	0	NA
A6	2014-02-06	Huang	120	90	1	480
A2	2014-02-07	Huang	119	94	0	433
A3	2014-02-07	Lee	121	88	1	433
A4	2014-02-07	Hsu	116	89	1	NA

資料排序：讓資料成為想要的順序

```
# 依照多個排序欄位之間以逗點隔開
ot <- ot[order(physician, patient_id)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
A1	2014-05-27	Hsu	121	90	1	250
A3	2014-03-07	Hsu	124	92	0	NA
A3	2014-06-27	Hsu	124	90	0	250
A3	2014-07-25	Hsu	119	89	0	433
A4	2014-02-07	Hsu	116	89	1	NA
A4	2014-03-07	Hsu	124	91	0	NA

資料排序：讓資料成為想要的順序

```
# 若要改為遞減（descending）排序，在變項前面加上負號（-）
ot <- ot[order(physician, -sbp)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity
A6	2014-05-01	Hsu	126	96	1	NA
A3	2014-03-07	Hsu	124	92	0	NA
A3	2014-06-27	Hsu	124	90	0	250
A4	2014-03-07	Hsu	124	91	0	NA
A4	2014-06-27	Hsu	124	92	0	NA
A6	2014-04-03	Hsu	124	93	0	433

資料排序：依據排序將觀察值選出

```
# 先排序
ot <- dt[order(patient_id, visit_date)]

# 第1筆
ot <- dt[, .SD[1], by = .(patient_id)]

# 第k筆，例如說K = 3
ot <- dt[, .SD[3], by = .(patient_id)]

# 第N筆(最後一筆)
ot <- dt[, .SD[.N], by = .(patient_id)]
```

資料修改：產生變數

```
# 新增一個變項名為vip，內容為數值9
ot <- dt[, `:=`(vip = 9)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity	vip
A1	2014-02-04	Huang	120	88	0	NA	9
A1	2014-03-04	Huang	121	94	1	433	9
A1	2014-04-01	Huang	116	92	0	250	9
A1	2014-04-29	Lee	120	94	1	NA	9
A1	2014-05-27	Hsu	121	90	1	250	9
A1	2014-06-24	Lee	118	92	0	NA	9

資料修改：更新變數

```
# 把既有的變項內容修改（更新）為0
ot <- ot[, `:=`(vip = 0)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity	vip
A1	2014-02-04	Huang	120	88	0	NA	0
A1	2014-03-04	Huang	121	94	1	433	0
A1	2014-04-01	Huang	116	92	0	250	0
A1	2014-04-29	Lee	120	94	1	NA	0
A1	2014-05-27	Hsu	121	90	1	250	0
A1	2014-06-24	Lee	118	92	0	NA	0

資料修改：有條件更新變數

```
# 如果是黃醫師看診的病人，都叫做VIP
ot <- ot[physician == "Huang", `:=`(vip = 1)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity	vip
A1	2014-02-04	Huang	120	88	0	NA	1
A1	2014-03-04	Huang	121	94	1	433	1
A1	2014-04-01	Huang	116	92	0	250	1
A1	2014-04-29	Lee	120	94	1	NA	0
A1	2014-05-27	Hsu	121	90	1	250	0
A1	2014-06-24	Lee	118	92	0	NA	0

資料修改：將長表轉置（transpose）成為寬表

```
# 依照patient_id相同的觀察值
# 把physician的內容轉換為新的欄位名稱
# 新的欄位內容由原有的dose_modify填入
ot <- dcast.data.table(
  data = dt,
  formula = patient_id ~ physician,
  value.var = "dose_modify")
```

長表（long）

patient_id	physician	visit_date	dose_modify
A1	Hsu	2014-05-27	1
A1	Huang	2014-02-04	0
A1	Lee	2014-04-29	1

寬表（wide）

patient_id	Hsu	Huang	Lee
A1	1	0	1

資料修改：將寬表轉置（transpose）成為長表

```
# 依照patient_id相同的觀察值
# 把Hsu, Huang, Lee三個欄位轉置成為變項標籤，名為physician
# 原始變項內容則成為新的變數dose_modify
ot <- melt.data.table(
  data = dt,
  id.vars = "patient_id",
  measure.vars = c("Hsu", "Huang", "Lee"),
  variable.name = "physician",
  value.name = "dose_modify")
```

寬表（wide）

patient_id	Hsu	Huang	Lee
A1	1	0	1

長表（long）

patient_id	physician	visit_date	dose_modify
A1	Hsu	2014-05-27	1
A1	Huang	2014-02-04	0
A1	Lee	2014-04-29	1

資料修改：去除重複

```
# 依照data.table內所有的變數進行比對，刪除重複的觀察值
ot_1 <- unique(dt[, .(patient_id)])
ot_2 <- unique(dt[, .(patient_id, physician)])
```

patient_id
A1
A2
A3
A4
A5
A6

patient_id	physician
A1	Huang
A1	Lee
A1	Hsu
A2	Huang
A3	Lee
A3	Hsu
A3	Huang

資料歸戶：依據變項內容相同者歸納資訊

```
#將分群統計資料存為新的變項，觀察值總數變少(與by的內容相同)
ot <- dt[, .(sbp_m = mean(sbp)), by = .(patient_id)]
```

	patient_id	sbp_m
1:	A1	120.5
2:	A2	119.0
3:	A3	120.3
4:	A4	120.0
5:	A5	120.2
6:	A6	120.2

資料歸戶：依據變項內容相同者統計資訊

```
# 將分群統計資料存為新的變項，觀察值總數不變
ot <- dt[, `:=`(sbp_m = mean(sbp)), by = .(patient_id)]
```

patient_id	visit_date	physician	sbp	dbp	dose_modify	comorbidity	vip	sbp_m
A1	2014-02-04	Huang	120	88	0	NA	1	120.5
A1	2014-03-04	Huang	121	94	1	433	1	120.5
A1	2014-04-01	Huang	116	92	0	250	1	120.5
A1	2014-04-29	Lee	120	94	1	NA	0	120.5
A1	2014-05-27	Hsu	121	90	1	250	0	120.5
A1	2014-06-24	Lee	118	92	0	NA	0	120.5
A1	2014-07-22	Lee	115	91	1	250	0	120.5
A1	2014-08-19	Huang	124	89	0	NA	1	120.5
A1	2014-09-16	Huang	124	90	0	433	1	120.5
A1	2014-10-14	Lee	126	92	0	433	0	120.5

資料合併：依據變項內容相同者串聯不同資料集

```
ot <- merge(dt_1, dt_2, by = c("patient_id", "visit_date"))
```

dt_1

patient_id	visit_date	sbp	dbp
A1	2014-02-04	120	88
A1	2014-03-04	121	94
A1	2014-04-01	116	92
A1	2014-04-29	120	94
A1	2014-05-27	121	90
A1	2014-06-24	118	92

dt_2

patient_id	visit_date	bmi
A1	2014-02-04	22.1
A1	2014-03-04	22.0
A1	2014-04-01	22.1
A1	2014-04-29	21.9
A1	2014-05-27	22.0
A1	2014-06-24	22.0

ot

patient_id	visit_date	sbp	dbp	bmi
A1	2014-02-04	120	88	22.1
A1	2014-03-04	121	94	22.0
A1	2014-04-01	116	92	22.1
A1	2014-04-29	120	94	21.9
A1	2014-05-27	121	90	22.0
A1	2014-06-24	118	92	22.0

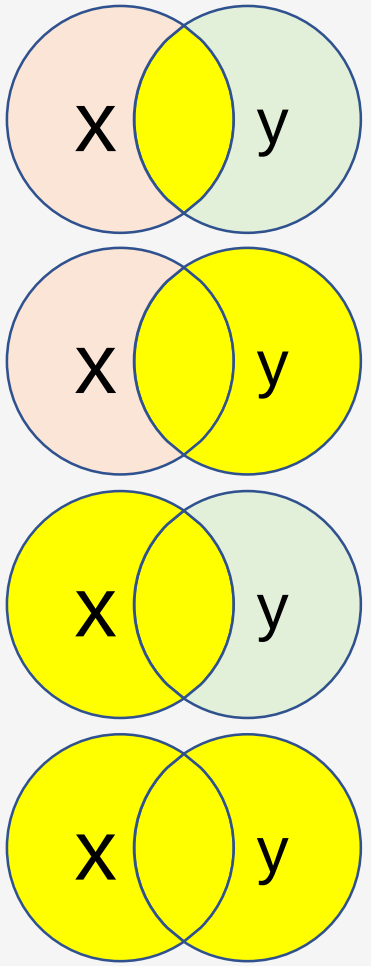
資料合併：不同水平合併的比較

```r
# 交集 inner join
ot <- merge(x = dt_1, y = dt_2, by = c("id"))

# 右聯結 right join
ot <- merge(x = dt_1, y = dt_2, by = c("id"), all.y = T)

# 左聯結 left join
ot <- merge(x = dt_1, y = dt_2, by = c("id"), all.x = T)

# 聯集 full join
ot <- merge(x = dt_1, y = dt_2, by = c("id"), all = T)
```



資料合併：依據變項名稱相同者堆疊不同資料集

```
ot <- rbind(dt_1, dt_2)
```

dt_1

patient_id	visit_date	physician	sbp	dbp
A1	2014-02-04	Huang	120	88
A1	2014-03-04	Huang	121	94
A1	2014-04-01	Huang	116	92
A1	2014-04-29	Lee	120	94
A1	2014-05-27	Hsu	121	90
A1	2014-06-24	Lee	118	92

dt_2

patient_id	visit_date	physician	sbp	dbp
A7	2014-02-04	Chang	118	90
A7	2014-03-04	Chang	119	90
A7	2014-04-01	Chang	120	87
A7	2014-04-29	Chang	113	88
A7	2014-05-27	Chang	120	91
A7	2014-06-24	Chang	124	92

ot

patient_id	visit_date	physician	sbp	dbp
A1	2014-02-04	Huang	120	88
A1	2014-03-04	Huang	121	94
A1	2014-04-01	Huang	116	92
A1	2014-04-29	Lee	120	94
A1	2014-05-27	Hsu	121	90
A1	2014-06-24	Lee	118	92
A7	2014-02-04	Chang	118	90
A7	2014-03-04	Chang	119	90
A7	2014-04-01	Chang	120	87
A7	2014-04-29	Chang	113	88
A7	2014-05-27	Chang	120	91
A7	2014-06-24	Chang	124	92

工作環境整理

```
# 刪除指定名稱的物件
rm(ot_merge)

# 用ls()函數找出環境中名稱符合指定模式的物件，刪除
rm(list = ls(pattern = "^ot_bind"))

# 用ls()函數找出環境中全部的物件名稱，刪除
rm(list = ls())

# 記憶體空間釋放
gc()
```

先決定資料管理目的，在選擇資料處理方式 實作看看！

操作處理

- 資料讀取
- 資料檢視
- 資料篩選
- 資料排序
- 資料修改
- 資料歸戶
- 資料合併

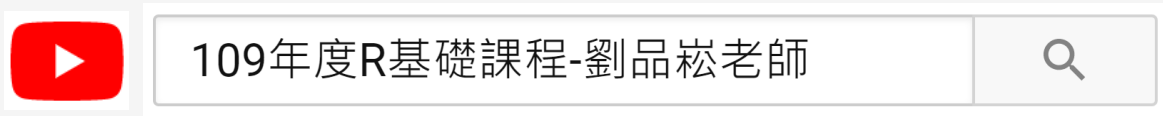
管理目的

- 開啟目標資料（放入電腦記憶體）
- 理解資料的輪廓
- 取得想要的資料
- 讓資料成為想要的順序
- 產生需要使用 / 更有意義的樣子
- 統計或集結資訊
- 串聯或堆疊不同資料集

Summary

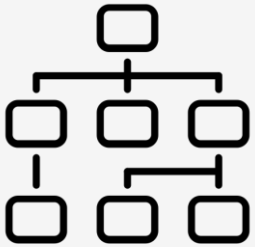
- 認識R軟體
  - 下載安裝
  - 指令互動
- 資料管理
  - 基本觀念
  - 實際操作
- 開放提問時間

- 劉品崧
- Peter Pin-Sung Liu
- psliu520@gmail.com
- https://github.com/PSLiu/

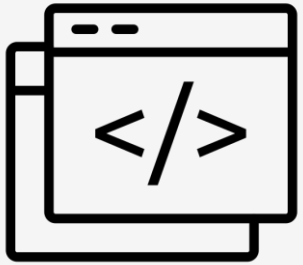


使用創用CC圖片宣告

Created by John Chapman
from the Noun Project



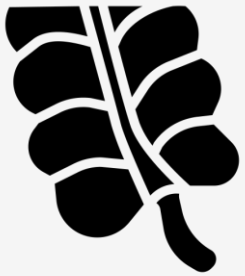
Created by QualityIcons
from the Noun Project



Created by SBTS
from the Noun Project

Created by Sunardi
from the Noun Project

Created by David Khai
from the Noun Project



Created by dDara