

資料科學 Data Science 系列

R 語言資料分析實務 (1)

從資料大海中，提煉出有價值的資料

姓名：羅左欣

日期：2016/10/3 (一)



本著作係採用創用 CC 姓名標示-非商業性-相同方式分享 3.0 台灣 授權條款授權。

部落格：<http://shouzo.github.io/>

Agenda

- (一) Prepare：預備工作**
- (二) Basic：基本介紹與操作**
- (三) 學習資源**

(一) Prepare : 預備工作

**(一) Prepare :
預備工作**

(一) Prepare：預備工作

安裝工作環境

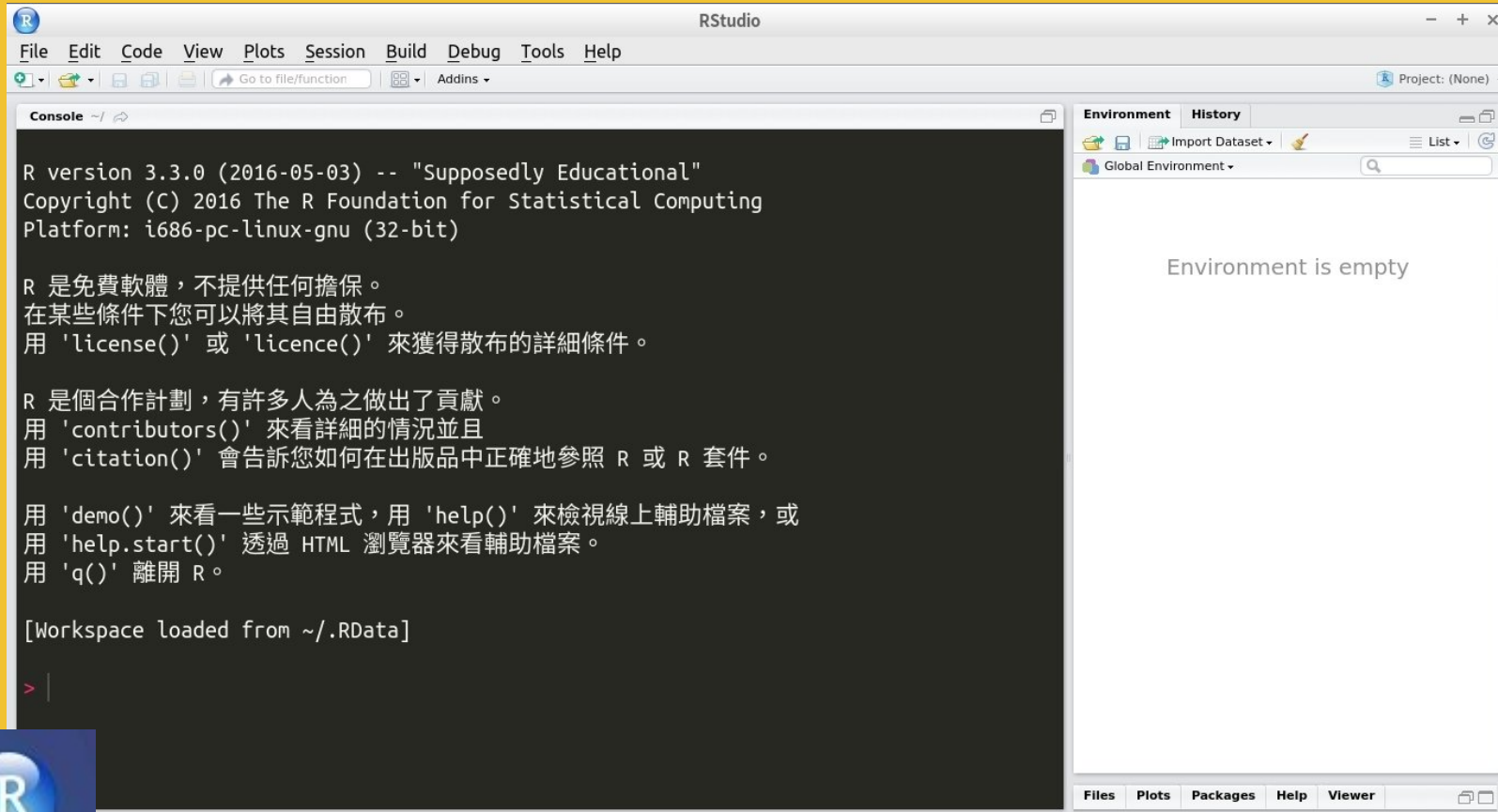


<https://youtu.be/fcd6zSk0yd8>

請參照上方影片 0:00 - 0:37 處

(一) Prepare：預備工作

啟動 RStudio



The screenshot shows the RStudio application window. The title bar reads "RStudio". The menu bar includes "File", "Edit", "Code", "View", "Plots", "Session", "Build", "Debug", "Tools", and "Help". The toolbar contains icons for file operations and a search bar. The main area is divided into two panes: "Console" on the left and "Environment" on the right. The console displays the R version information and a message in Chinese. The environment pane shows "Global Environment" and "Environment is empty".

```
R version 3.3.0 (2016-05-03) -- "Supposedly Educational"  
Copyright (C) 2016 The R Foundation for Statistical Computing  
Platform: i686-pc-linux-gnu (32-bit)  
  
R 是免費軟體，不提供任何擔保。  
在某些條件下您可以將其自由散布。  
用 'license()' 或 'licence()' 來獲得散布的詳細條件。  
  
R 是個合作計劃，有許多人為之做出了貢獻。  
用 'contributors()' 來看詳細的情況並且  
用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。  
  
用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或  
用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。  
用 'q()' 離開 R。  
  
[Workspace loaded from ~/.RData]  
> |
```



在這個系列的簡報中，主要以 "RStudio" 做為主要軟體。

(二) Basic : 基本介紹與操作

(二) Basic : **基本介紹與操作**

(二) Basic：什麼是資料探勘？

1. 什麼是資料探勘？

(二) Basic：什麼是資料探勘？

基本概念

資料探勘 = 資料庫之知識發掘

(**K**nowledge **D**iscovery in **D**atabases，**KDD**)

亦可稱作

- A. 數據挖掘
- B. 資料挖掘
- C. 資料採礦

它是"資料分析"技術裡的一個環節。

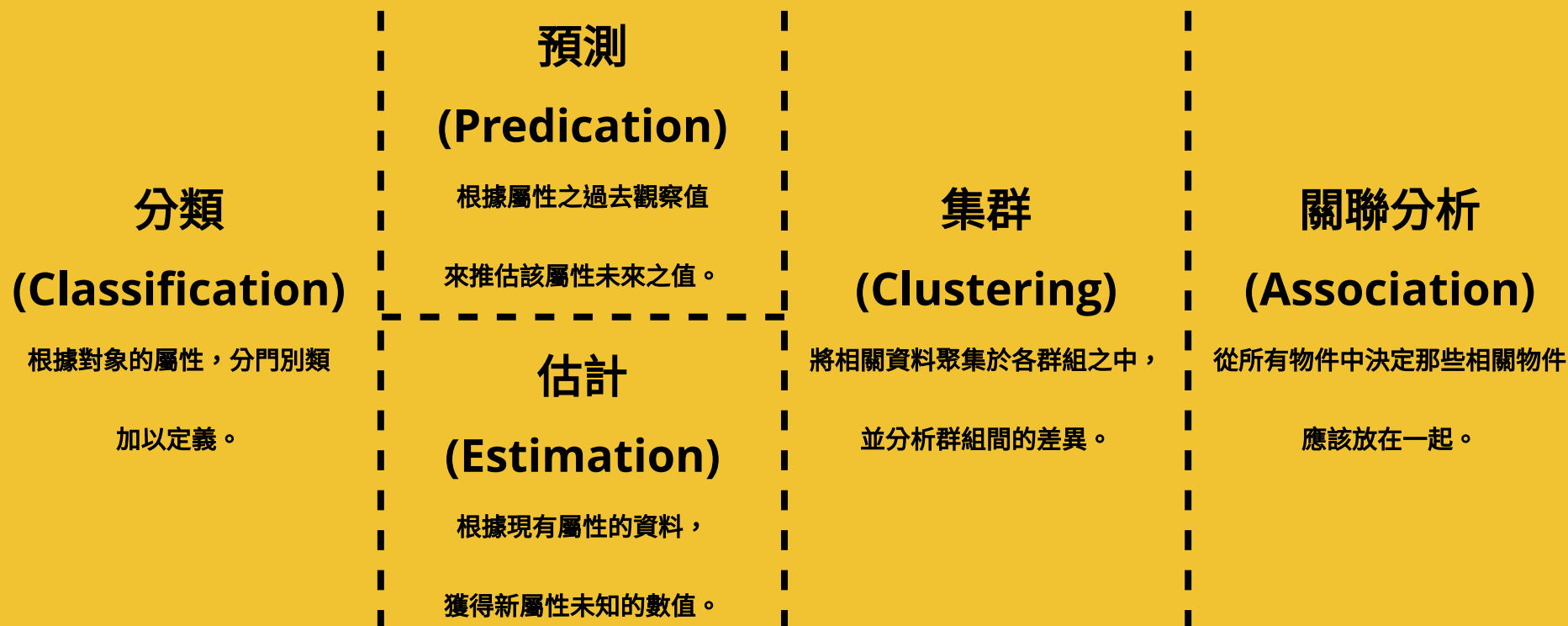
目的：從大量的資料中提取"有價值的資訊"

(二) Basic：什麼是資料探勘？

資料探勘技術

知識發掘(Knowledge Discovery)：由現有的資料中，得知一些我們不知道的事情。

假設檢定(Hypothesis Testing)：嘗試去證實或舉反證來驗證預設的想法。



資料引用自：http://myweb.fcu.edu.tw/~mhsung/Ecommerce/Data_Mining/DM_methods.htm

(二) Basic：什麼是資料探勘？

工作流程

STEP1：收集資料



訂立目標及問題之後，
進行資料蒐集及前置處理。

STEP2：處理資料



使用適當的工具，
進行資料分析。

STEP3：輸出結果



觀察者從分析後的結果得知資料代
表的意義，並思考出目標及問題的
處理方案。

(二) Basic : 軟體基礎

2. 軟體基礎

(二) Basic：軟體基礎

專有名詞

(1) 向量(Vectors)

(2) 因子(Factors)

(3) 陣列(Arrays)、矩陣(Matrices)

(4) 資料框(DataFrame)

(5) 列表(Lists)

以上名詞是在資料處理中，重要的概念。

(二) Basic：軟體基礎

基本算術運算

運算次序：括號 > 指數 > 乘法 > 除法 > 加減法

```
> 1 + 1
[1] 2
>
> 1 + 2 + 3
[1] 6
>
> 3 * 7 * 2
[1] 42
>
> 4 / 2
[1] 2
>
> 4 / 3
[1] 1.333333
>
```

```
> 4 * 6 + 5
[1] 29
>
> (4 * 6) + 5
[1] 29
>
> 4 * (6 + 5)
[1] 44
>
```

(二) Basic：軟體基礎

物件的操作

"指派"物件

```
> # 指派物件 (可以使用"<-"或"=")，建議使用"<-"
> x <- 2
>
> x          # 顯示"x"的值
[1] 2
>
> y = 5
> y          # 顯示"y"的值
[1] 5
>
>
> # 可以連續指派
> a <- b <- 7
> a
[1] 7
> b
[1] 7
>
```

"移除"物件

```
> # 移除物件
> a          # 查看變數"a"
[1] 7
>
> rm(a)      # 移除變數"a"
>
> a          # 現在變數"a"被移除了，所以"a"不存在
錯誤：找不到物件 'a'
>
```

(二) Basic：軟體基礎

向量(Vectors)

(1) 向量(Vectors)

1. "同型別"元素的集合：不能涵蓋不同型別的元素
2. 向量(Vectors)沒有維度：沒有"行(colume)"跟"列(row)"
3. 最基礎的"物件(Objects)"

```
> # 把幾個元素結合在一起，形成 "向量(vectors) "  
> x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)  
>  
> # 顯示"x"的內容  
> x  
[1] 1 2 3 4 5 6 7 8 9 10  
>
```

常見的"型別"：邏輯(logical)、整數(integer)、數值(numeric)、
複數(complex)、字元(character)

(二) Basic : 軟體基礎

向量(Vectors)

(1) 向量(Vectors)

向量的運算(1)

```
> x <- c(1, 2, 3, 4, 5)      # 建立向量 "x"  
>  
> x * 3      # 把每一個元素乘以 3  
[1] 3 6 9 12 15  
>  
> x + 2      # 把每一個元素都加上 2  
[1] 3 4 5 6 7  
>  
> x - 4      # 把每一個元素都扣掉 4  
[1] -3 -2 -1 0 1  
>  
> x / 10     # 把每一個元素都除以 10  
[1] 0.1 0.2 0.3 0.4 0.5  
>  
> x ^ 2      # 把每一個元素都平方  
[1] 1 4 9 16 25  
>
```


(二) Basic : 軟體基礎

向量(Vectors)

(1) 向量(Vectors)

向量的運算(2)

```
> # 使用 ":" 產生連續的數字
> 1:10
[1] 1 2 3 4 5 6 7 8 9 10
> -5:4
[1] -5 -4 -3 -2 -1 0 1 2 3 4
>
> # 製造兩個相同長度的向量
> x <- 1:10
> y <- -5:4
> x + y      # 相加
[1] -4 -2 0 2 4 6 8 10 12 14
> x - y      # 相減
[1] 6 6 6 6 6 6 6 6 6 6
> x * y      # 相乘
[1] -5 -8 -9 -8 -5 0 7 16 27 40
> x / y      # 相除
[1] -0.2 -0.5 -1.0 -2.0 -5.0 Inf 7.0 4.0 3.0 2.5
> x ^ y      # 指數型態
[1] 1.000000e+00 6.250000e-02 3.703704e-02 6.250000e-02 2.000000e-01 1.000000e+00
[7] 7.000000e+00 6.400000e+01 7.290000e+02 1.000000e+04
```

(二) Basic：軟體基礎

向量(Vectors)

(1) 向量(Vectors)

向量的處理

1. 查看向量的型態和長度

```
> x <- 1:10      # 建立一個1 ~ 10的向量
>
> # 使用"mode()"來查看向量的型態
> mode(x)
[1] "numeric"
>
> # 使用"length()"來查看向量的長度
> length(x)
[1] 10
>
```

2. 用中括號查看向量裡的特定元素

```
> x <- 1:10      # 建立一個1 ~ 10的向量
>
> # 使用中括號"[]"查看特定元素
> x[1]
[1] 1
>
> x[1:2]
[1] 1 2
>
> x[c(1, 4)]     # 查看"第1個元素"和"第4個元素"的值
[1] 1 4
>
```

(二) Basic：軟體基礎

(2) 因子(Factors)

1. 儲存"類別型態"的資料
2. 具有"level"屬性
3. 分為"無順序"或"有順序"的

因子(Factors)

屬於「類別」資料的例子：

- 性別：「男、女」
- 地區：「台北、台中、台南、高雄」
- 血型：「A、B、AB、O」

註1：在 R 中，預設是以"字母順序"排列 levels；可使用 `ordered = TRUE`，讓 Factor 變成"有序")

註2：因子(Factors)像是"經過分級"後的向量(Vectors)

```
> x <- c(1, 2, 4, 3, 1, 2, 3, 4, 1)
>
> factor(x)
[1] 1 2 4 3 1 2 3 4 1
Levels: 1 2 3 4
>
>
> # 自訂 Level 的名稱。
> factor(x, labels = c("一", "二", "三", "四"))
[1] 一 二 四 三 一 二 三 四 一
Levels: 一 二 三 四
```

```
> factor(x, ordered = TRUE) # 進行排序
[1] 1 2 4 3 1 2 3 4 1
Levels: 1 < 2 < 3 < 4
>
```

(二) Basic：軟體基礎

陣列(Array)、矩陣(Matrices)

(3) 陣列(Array)、矩陣(Matrices)

陣列(Array)

1. "同型別"元素的集合
 2. 有"列(row)"、"行(colume)"
- 一種"多維度"的向量(vectors)

```
> # 建立一個 2x3x2 的陣列
> theArray <- array(1:12, dim = c(2, 3, 2))
> theArray
, , 1
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

, , 2
     [,1] [,2] [,3]
[1,]    7    9   11
[2,]    8   10   12
```

矩陣(Matrices)

1. "同型別"元素的集合
 2. 有"列(row)"、"行(colume)"
- 一種"只有二維"的向量(vectors)

```
> # 建立一個 3x4 的矩陣
> theMatrix <- matrix(1:12, 3, 4)
> theMatrix
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
>
```

(二) Basic：軟體基礎

資料框(DataFrame)

(4) 資料框(DataFrame)

◎ 有"列(row)"、"行(colume)"

1. "行"：代表"變數"

註：行與行之間可儲存不同的資料型別

2. "列"：代表"觀測值"

```
> a <- data.frame(id = 1:10, scores = matrix(c(80:99), nrow = 10, ncol = 2 ) )
> a
  id scores.1 scores.2
1  1      80      90
2  2      81      91
3  3      82      92
4  4      83      93
5  5      84      94
6  6      85      95
7  7      86      96
8  8      87      97
9  9      88      98
10 10      89      99
>
```

(二) Basic : 軟體基礎

列表(Lists)

(5) 列表(Lists)

1. 能儲存任意物件

2. 可包含任何型別和長度的資料

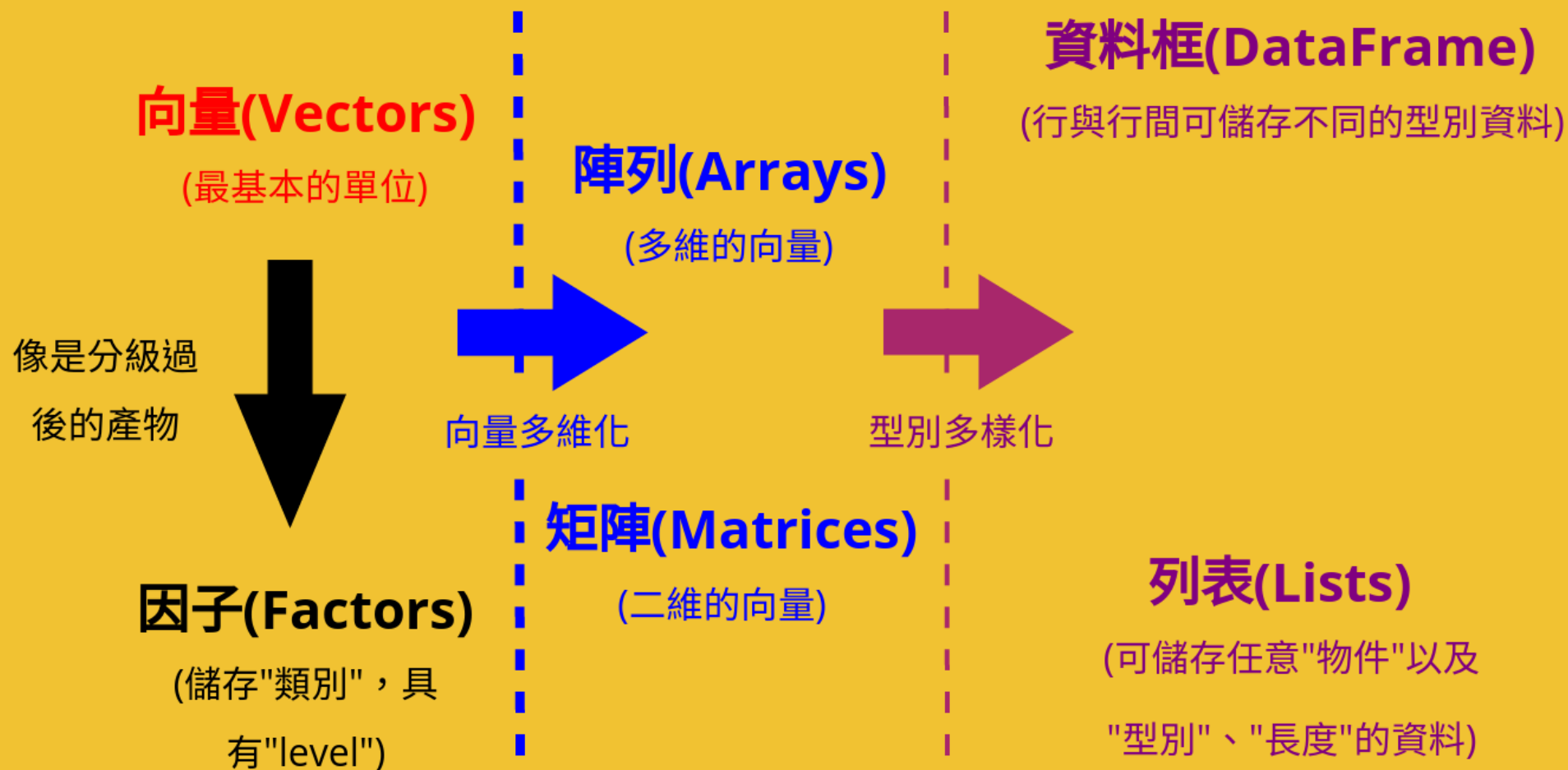
◎ 例如：numeric、character、DataFrame、list...

```
> x <- list(iris = iris, cars = cars, n = 2)
> x
$iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2   setosa
2           4.9           3.0           1.4           0.2   setosa
3           4.7           3.2           1.3           0.2   setosa
4           4.6           3.1           1.5           0.2   setosa
5           5.0           3.6           1.4           0.2   setosa
6           5.4           3.9           1.7           0.4   setosa
7           4.6           3.4           1.4           0.3   setosa
8           5.0           3.4           1.5           0.2   setosa
9           4.4           2.9           1.4           0.2   setosa
10          4.9           3.1           1.5           0.1   setosa
11          5.4           3.7           1.5           0.2   setosa
12          4.8           3.4           1.6           0.2   setosa
13          4.8           3.0           1.4           0.1   setosa
```

(二) Basic：軟體基礎

名詞之間的關係如下...

名詞之間的關係



(二) Basic：簡易視覺化

小型成果實作

簡易視覺化

(二) Basic：簡易視覺化

空間資料視覺化

繪製地圖與資料分佈圖

參考資源：[R 的 ggmap 套件：繪製地圖與資料分佈圖，空間資料視覺化](#)

(二) Basic：簡易視覺化

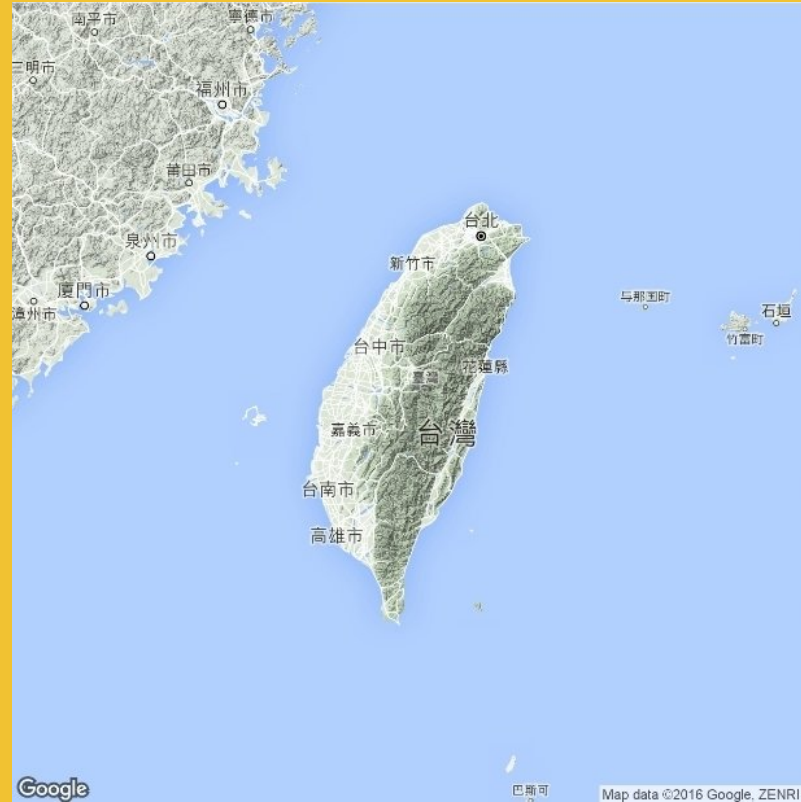
本範例應用之素材如下

參考資源：[R 的 ggmap 套件](#)：繪製地圖與資料分佈圖，空間資料視覺化

(二) Basic：簡易視覺化

使用素材(1)

台灣地圖



參考資源：R 的 [ggmap](#) 套件：繪製地圖與資料分佈圖，空間資料視覺化

(二) Basic：簡易視覺化

使用素材(2) ➡ 政府資料開放平臺 - 紫外線即時監測資料

<http://data.gov.tw/node/6076>

The screenshot shows the website interface for the 'Ultraviolet Radiation Real-time Monitoring Data' dataset. The page includes a navigation menu with options like '資料集下載', '互動專區', and '活化應用'. The main content area displays the dataset title, a rating of 3.4/5, and a list of metadata including '資料集描述', '主要欄位說明', '資料資源', '資料集類型', '資料集提供機關名稱', and '更新頻率'. The '資料資源' section lists available formats: XML, JSON, and CSV, each with a '檢視資料' link.

參考資源：R 的 gmap 套件：繪製地圖與資料分佈圖，空間資料視覺化

(二) Basic：簡易視覺化

地圖搭配開放資料，進行視覺化

參考資源：[R 的 ggmap 套件：繪製地圖與資料分佈圖，空間資料視覺化](#)

(二) Basic：簡易視覺化

STEP1：繪製基本地圖(1)

- 必備套件：ggmap、

```
# 安裝套件 (如果沒有才需要安裝)
> install.packages("ggmap")      # 安裝"ggmap"套件
> install.packages("mapproj")    # 安裝"mapproj"套件

# 載入套件
> library(ggmap)                 # 載入"ggmap"套件
> library(mapproj)              # 載入"mapproj"套件

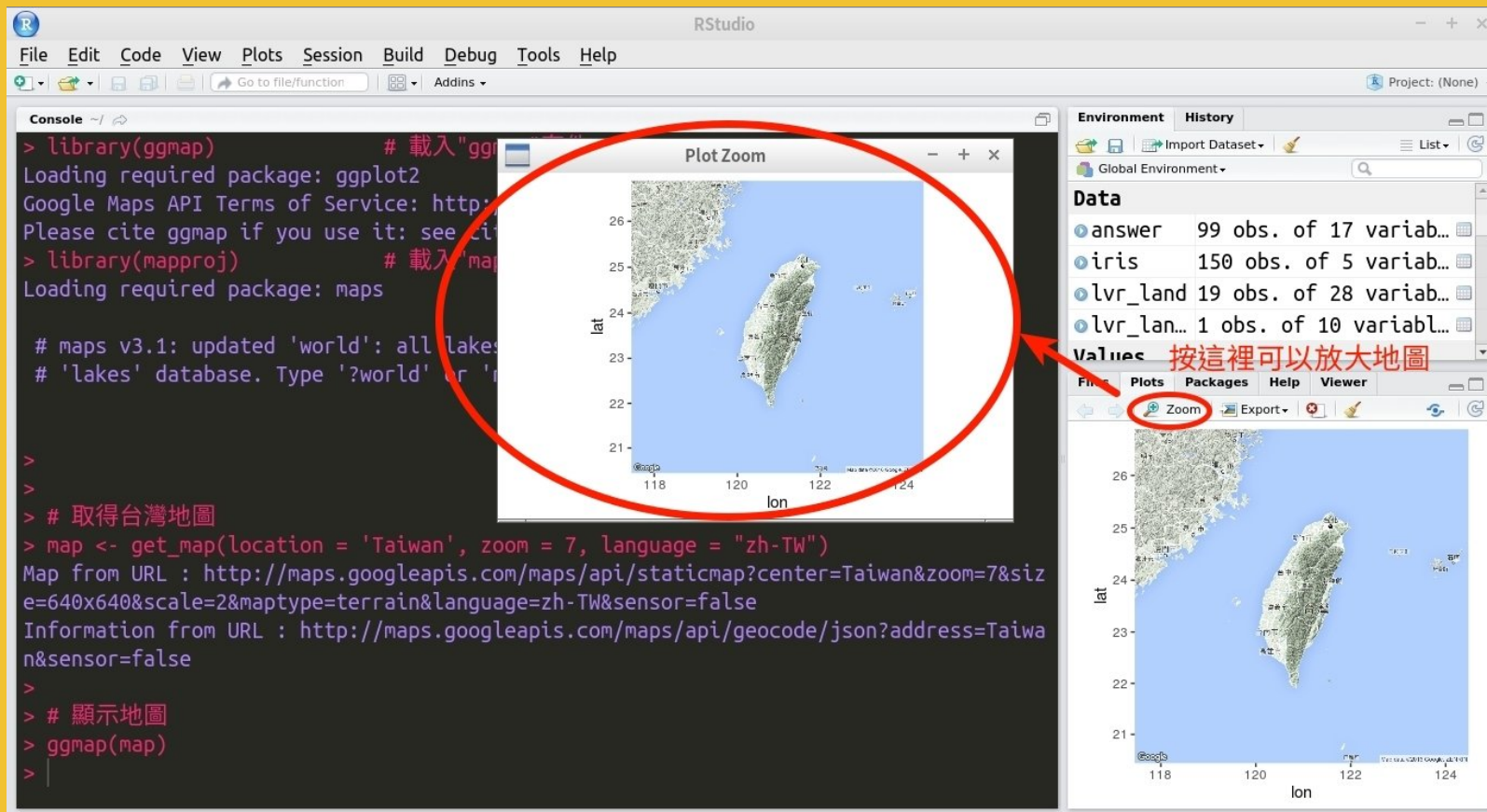
# 取得台灣地圖
> map <- get_map(location = 'Taiwan', zoom = 7, language = "zh-TW")

# 顯示地圖
> ggmap(map)
```

參考資源：[R 的 ggmap 套件：繪製地圖與資料分佈圖，空間資料視覺化](#)

(二) Basic：簡易視覺化

STEP1：繪製基本地圖 (執行結果)

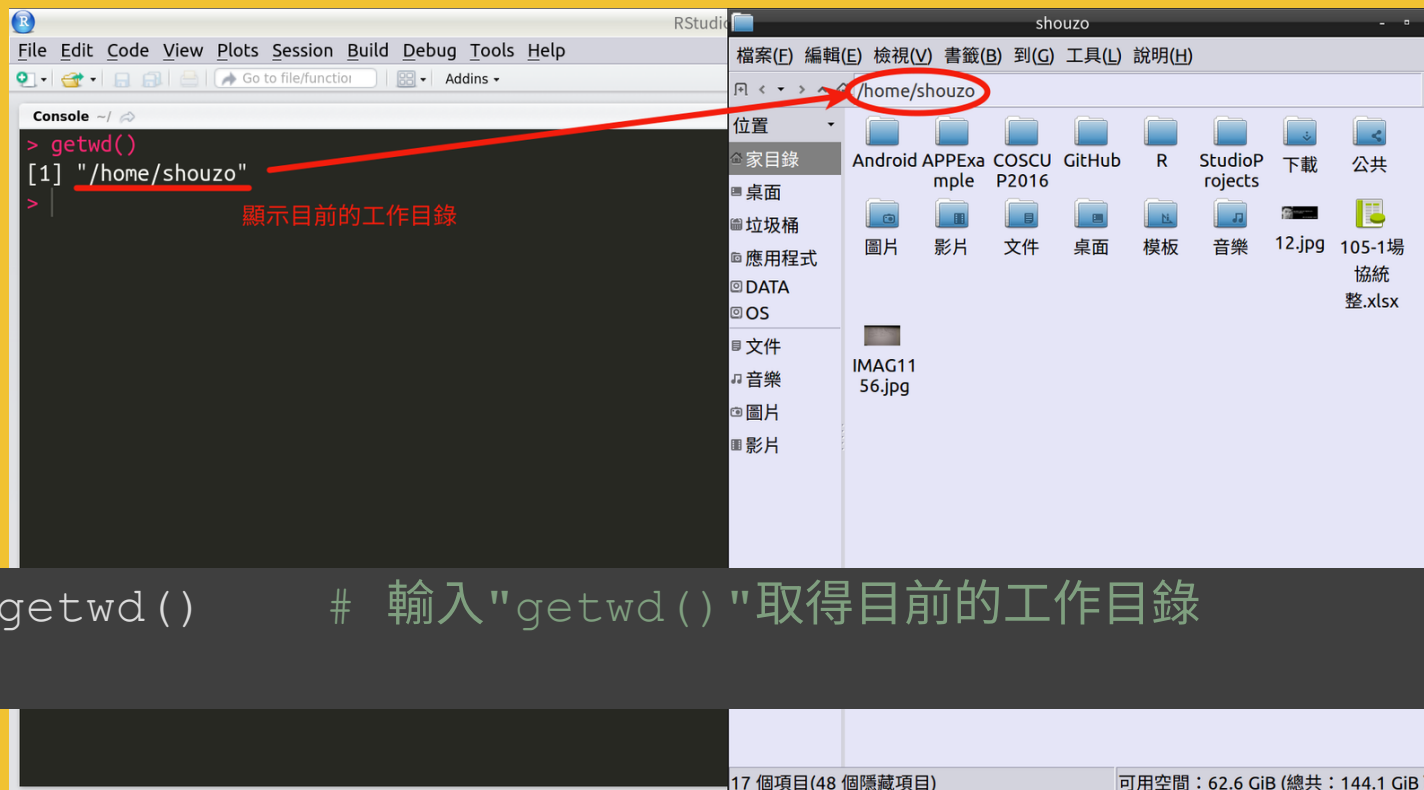


參考資源：R 的 `ggmap` 套件：繪製地圖與資料分佈圖，空間資料視覺化

(二) Basic：簡易視覺化

STEP2：取得資料(1)

- 先取得目前的工作目



The screenshot shows two windows side-by-side. On the left is the RStudio console window, and on the right is a Windows File Explorer window. The RStudio console shows the command `getwd()` being executed, resulting in the output `"/home/shouzo"`. A red arrow points from the output in the console to the address bar of the File Explorer window, which also displays `/home/shouzo`. Below the screenshot, a dark grey box contains the R command `> getwd()` followed by a comment in Chinese: `# 輸入 "getwd()" 取得目前的工作目錄`.

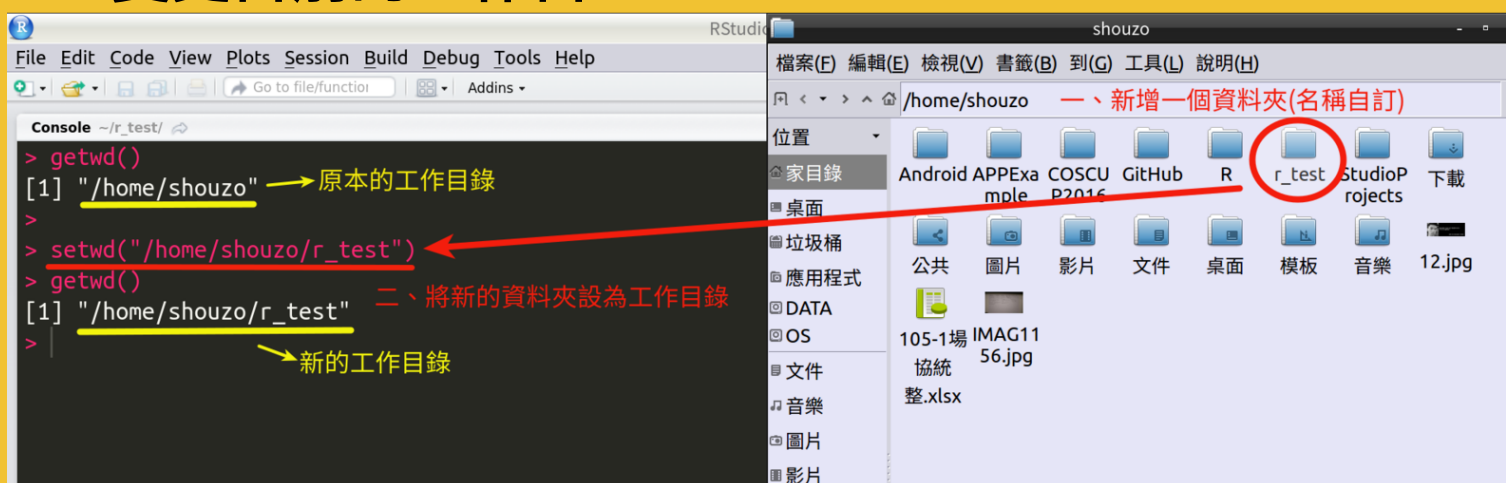
```
> getwd() # 輸入 "getwd()" 取得目前的工作目錄
```

參考資源：[R 的 ggmap 套件：繪製地圖與資料分佈圖，空間資料視覺化](#)

(二) Basic：簡易視覺化

STEP2：取得資料(2)

- 變更目前的工作目錄



```
> setwd() # 輸入"setwd()"變更目前的工作目錄

# 範例：將工作目錄變更到"r_test"這個資料夾底下
> setwd("../r_test")
```

參考資源：[R 的 ggmap 套件：繪製地圖與資料分佈圖，空間資料視覺化](#)

(二) Basic：簡易視覺化

紫外線即時監測資料：<http://data.gov.tw/node/6076>

STEP2：取得資料(3)

- 將資料存進工作目錄

The screenshot shows two windows. The left window is a web browser displaying the 'data.gov.tw' website. The right window is a file manager showing the directory structure of a file named 'UV_20160906115527.csv'.

Browser Window (Left):

- Page Title: 紫外線即時監測資料 | 政府資料開放平台
- URL: data.gov.tw/node/6076
- Page Content: 紫外線即時監測資料
- 資料集評分: 平均: 3.6 (8 投票)
- 資料集描述: 本資料集之舊有檔案下載連結 http://... 提供之檔案下載連結。環保署和中央氣象局資料已整合成1個檔。
- 主要欄位說明: 測站名稱(SiteName)、紫外線指數(UV指數)、緯度(WGS84)(TWD97Lat)、發布時間
- 資料資源: XML, JSON, CSV (circled in red), 檢視資料
- 資料集類型: 原始資料
- 資料集提供機關名稱: 行政院環境保護署
- 更新頻率: 每小時更新
- 授權方式: 政府資料開放授權條款-第1版

File Manager Window (Right):

- Window Title: r_test
- Location: /home/shouzo/r_test
- File List: UV_20160906115527.csv (circled in red)
- File Type: 文件
- File Size: 15527 bytes
- File Extension: .csv

A red arrow points from the 'CSV' option in the browser window to the 'UV_20160906115527.csv' file in the file manager window.

參考資源：R 的 [ggmap](#) 套件：繪製地圖與資料分佈圖，空間資料視覺化

(二) Basic：簡易視覺化

STEP2：取得資料(4)

- 載入並檢視資料

```
> # 載入要檢視的資料檔
> uv <- read.csv("UV_20160906115527.csv")
>
> head(uv)      # 檢視資料檔的前半段資料
  SiteName UVI PublishAgency County      WGS84Lon      WGS84Lat      PublishTime
1   屏東    0   環境保護署 屏東縣 120,29,16.92 22,40,23.09 2016-09-06 11:00
2   橋頭    1   環境保護署 高雄市 120,18,20.48 22,45,27.02 2016-09-06 11:00
3   新營    6   環境保護署 臺南市 120,19,2.10 23,18,20.28 2016-09-06 11:00
4   朴子    7   環境保護署 嘉義縣 120,14,50.46 23,27,55.11 2016-09-06 11:00
5  塔塔加    6   環境保護署 嘉義縣 120,52,50.06 23,28,14.19 2016-09-06 11:00
6  阿里山    2   環境保護署 嘉義縣 120,48,05.02 23,30,30.82 2016-09-06 11:00
>
```

參考資源：R 的 [ggmap](#) 套件：繪製地圖與資料分佈圖，空間資料視覺化

(二) Basic：簡易視覺化

STEP2：處理資料

A. 進行經緯度的轉換

原始的經緯度資料是以度分秒表示，在使用前要轉換為度數表示。

```
> lon.deg <- sapply(strsplit(as.character(uv$WGS84Lon), ","), as.numeric)
> uv$lon <- lon.deg[1, ] + lon.deg[2, ]/60 + lon.deg[3, ]/3600
> lat.deg <- sapply(strsplit(as.character(uv$WGS84Lat), ","), as.numeric)
> uv$lat <- lat.deg[1, ] + lat.deg[2, ]/60 + lat.deg[3, ]/3600
>
```

B. 把資料加入地圖中

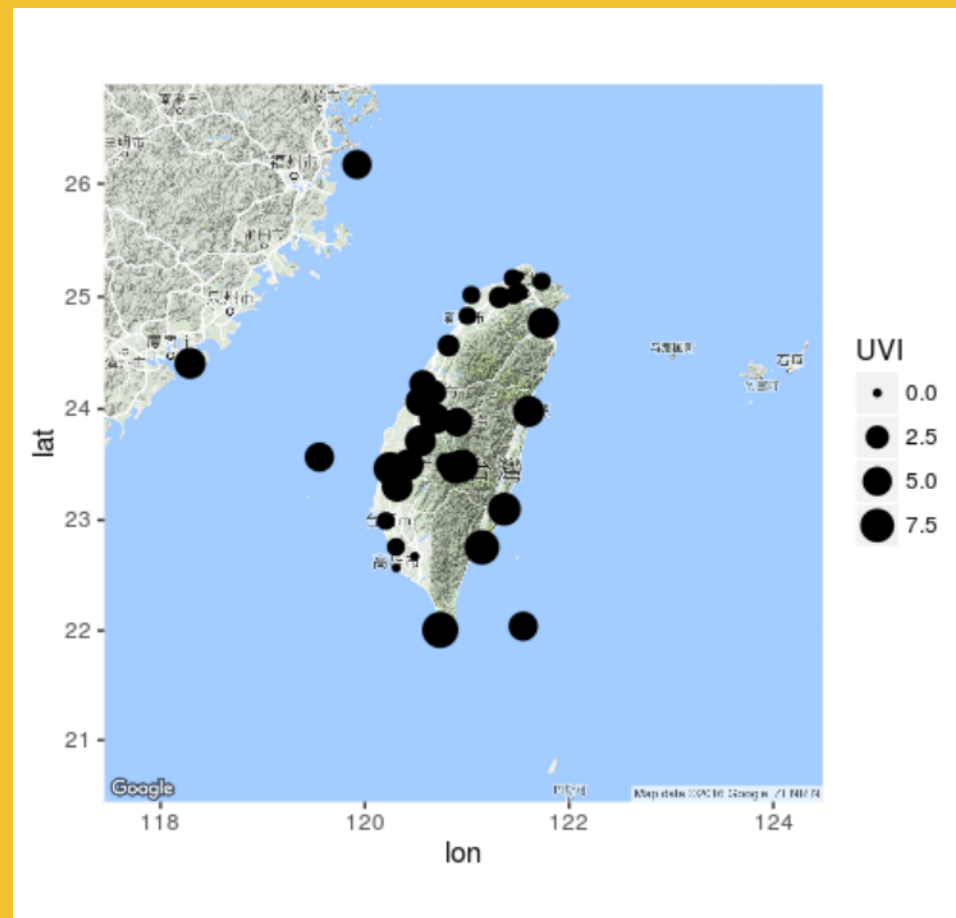
```
> ggmap(map) + geom_point(aes(x = lon, y = lat, size = UVI), data = uv)
```

參考資源：[R 的 ggmap 套件：繪製地圖與資料分佈圖，空間資料視覺化](#)

(二) Basic：簡易視覺化

STEP3：輸出結果

從結果圖中，我們可以得知全台灣
在取得資料當下的紫外線分佈情況。
紫外線強度越大，點就會越大。



參考資源：[R 的 ggmap 套件](#)：繪製地圖與資料分佈圖，空間資料視覺化

(三) Reference : 學習資源

**(三) Reference :
學習資源**

(三) Reference：學習資源

線上教材

一、中文教材

1. **R 語言翻轉教室** - Wush Wu、Chih Cheng Liang、Johnson Hsieh

<http://datascienceandr.org/>

2. **手把手教你 R 語言資料分析實務** - 張毓倫&陳柏亨

<http://goo.gl/18mwug>

3. **R 軟體與資料探勘之開發與應用** - 陳志華

<https://goo.gl/NPdzzP>



二、英文教材

1. **DataCamp**

<https://www.datacamp.com/>

2. **R for Data Science**

<http://r4ds.had.co.nz/>

(三) Reference：學習資源

推薦書籍



R 軟體資料分析基礎與應用

作者：Jared P. Lander

譯者：鍾振蔚

出版社：旗標

(三) Reference：學習資源

相關社群

台灣資料科學年會

<https://www.facebook.com/twdsconf/>



Taiwan R User Group

<https://www.facebook.com/Tw.R.User/>

資料視覺化 / Data Visualization

<https://www.facebook.com/data.visualize/>



Q & A