

# R 語言基礎教學

David Chiu  
2017/01/20

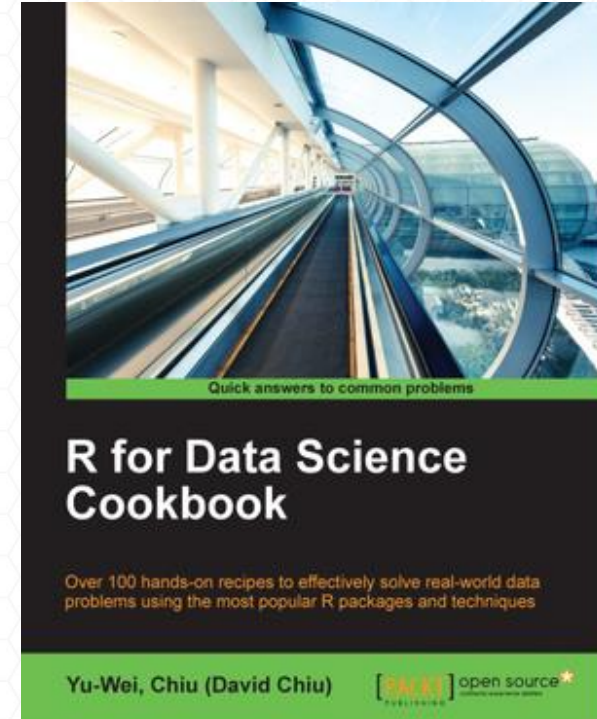
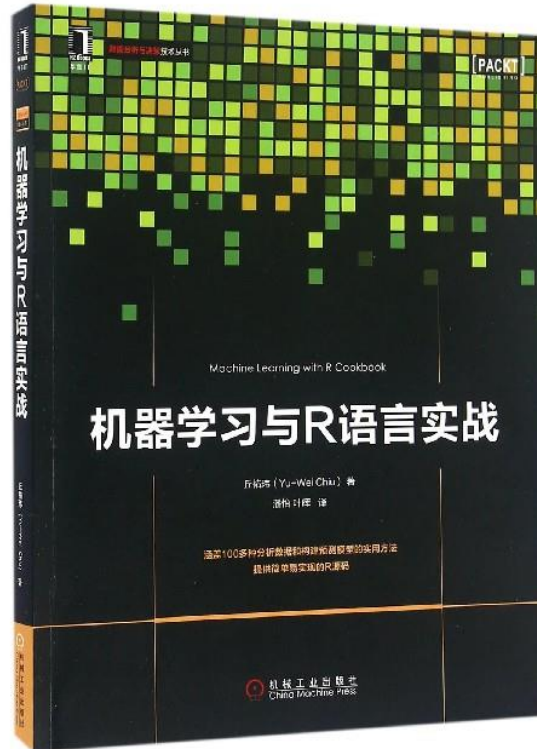
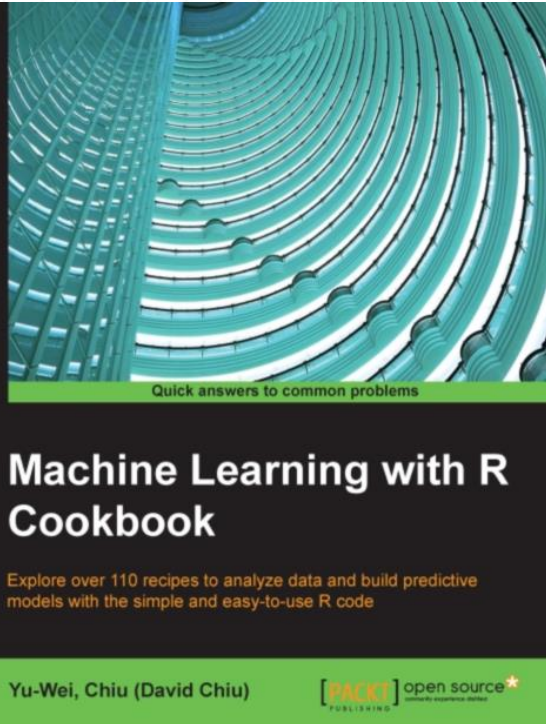
# 關於我



- 大數軟體有限公司創辦人
- 前趨勢科技工程師
- [ywchiu.com](http://ywchiu.com)
- 大數學堂  
<http://www.largitdata.com/>
- 粉絲頁  
<https://www.facebook.com/largitdata>
- R for Data Science Cookbook  
<https://www.packtpub.com/big-data-and-business-intelligence/r-data-science-cookbook>
- Machine Learning With R Cookbook  
<https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r-cookbook>



# Machine Learning With R Cookbook (机器学习与R语言实战) & R for Data Science Cookbook



Author: David (YU-WEI CHIU) Chiu

# 環境資訊頁面

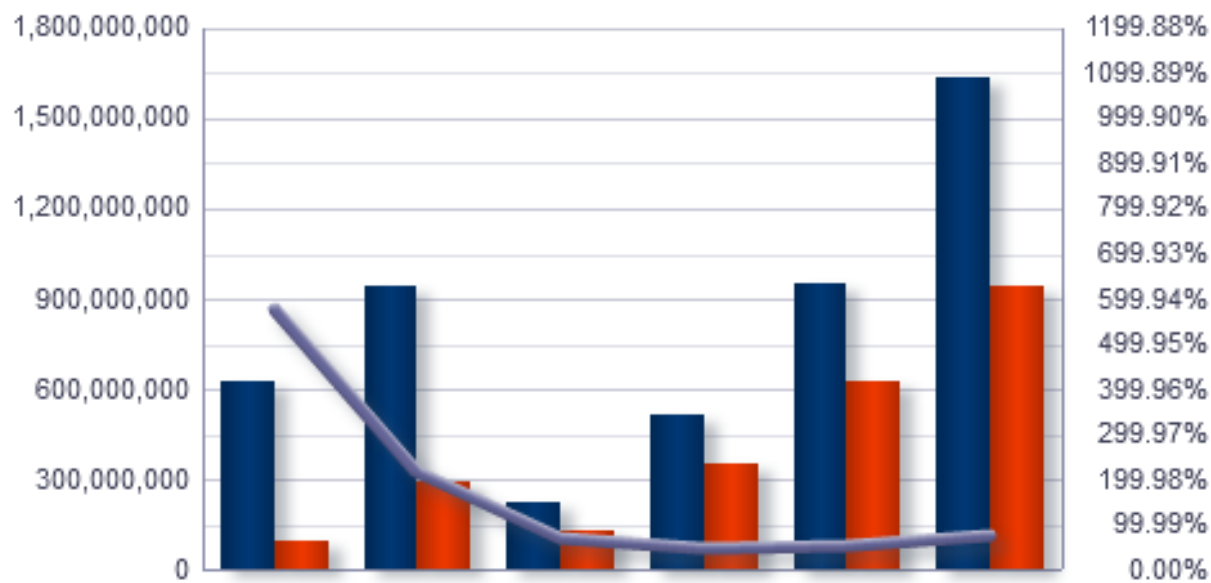
- 所有課程補充資料、投影片皆位於
  - <https://github.com/ywchiu/ipost>

# R語言與資料分析



# 資料分析實作 - 一個簡單的問題

- 試想如果今天老闆要你找出哪個年齡層的客户最多，並畫出資料分佈圖的話，該怎麼做？



# 不同的做法

## ■ 資料庫派的

- 先下個SQL 做個資料聚合
- 使用視覺化工具呈現到報表上
- 或許使用Excel 比較容易些



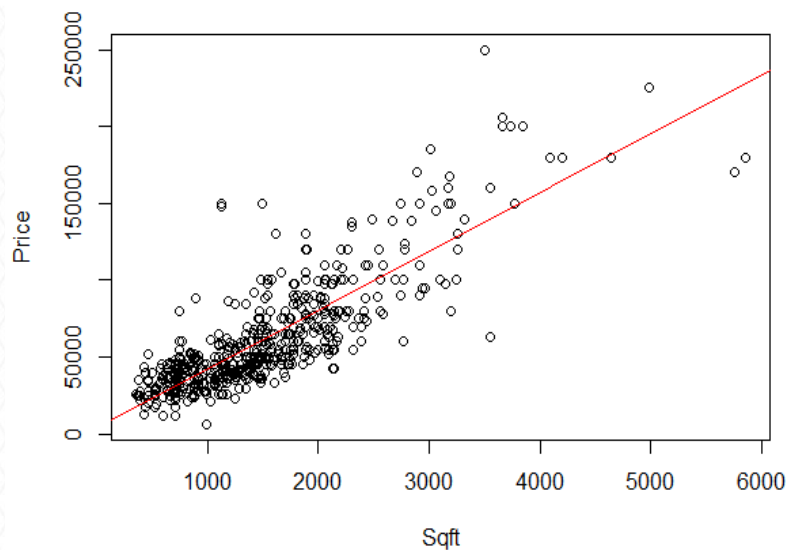
## ■ 軟體工程師派的

- 寫一個For迴圈掃過資料後，依條件規則進行聚合
- 使用圖表套件呈現圖表



# 相關性分析 - 更複雜的問題

## 統計房屋坪數與房價的關係



591 房屋交易 .com.tw 出租 目前所在縣市: 台北市

所有房屋 社區找房 地圖找房 找經紀人 手機找房 房東求租

縣市搜尋 | 商圈搜尋 | 學校搜尋 | 捷運搜尋 |

台北市 租屋 類型 請輸入社區、街道、商圈或房屋編號... 搜尋 地圖找房 找附近打工

目前共有 2,000 人在找房子, 出租中屋數 54,120 筆

我的搜尋條件 我的搜尋條件 (0) 我的收藏物件 (0)

縮小搜尋範圍 重置

租金 不限 5000元以下 5000-10000元 10000-15000元

精選推薦

- 信義路稀有金店面, 近通... 大安區 - 店面 27.09坪 140,000元
- 附月租, 年租勿來, 雙連... 中山區 - 獨立套房 15坪 30,000元
- 設計感對外窗套房-近劍... 士林區 - 分租套房 6.2坪 9,500元
- 近台北車站-滿廷長安大... 大同區 - 辦公 64.59坪 85,000元

租屋列表頁 新版上線

我也更出意見 體驗新版

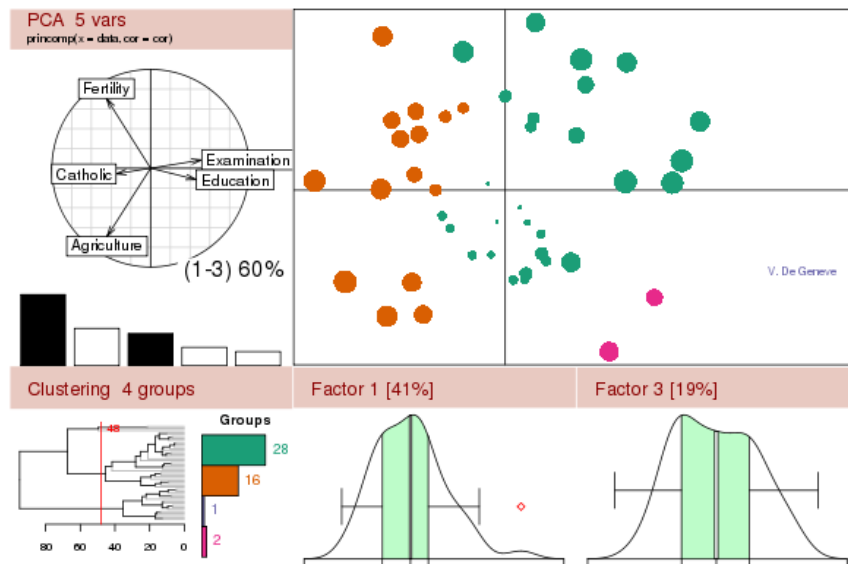
下載591 收藏頁面 意見反饋 回報錯誤





# 什麼是R

- AT&T貝爾實驗室暨S語言所發展出來的GNU 專案
- 提供統計分析與圖形視覺化功能的開源程式語言
- 使用C, Fortran 編程的函式語言



# S 語言

- 1976 年 John Chambers 在貝爾實驗室開發出 S，用來取代 SAS 與 SPSS
  - ▣ 1976 年使用 Fortran 實現的第一代 (S Version 1)
  - ▣ 1978 年支援 Linux 系統 (S Version 2)
  - ▣ 1983 ~ 1992 年引入萬物皆物件的概念 (S version 3)
  - ▣ 1993 年被 MathSoft 買斷，改版為 S-PLUS(當時三大統計軟體之一)
  - ▣ 1995 年更新後變為 (S Version 4)
  - ▣ 1998 年 S 獲得 ACM 的軟體系統獎
  - ▣ 2008 年 S-PLUS 被 TIBCO 收購



# R 語言

- S 語言的方言 (分支)
- 受到函數式編程語言Scheme 的啟發，因而想將該功能加入到 S 語言當中
- 1992年Ross Ihaka 與 Robert Gentleman 為了教授統計，因此開發出了 R語言
- 除了R 以外，還有S-Plus，但兩個分支走向不同，一個走向社群，一個走向商業

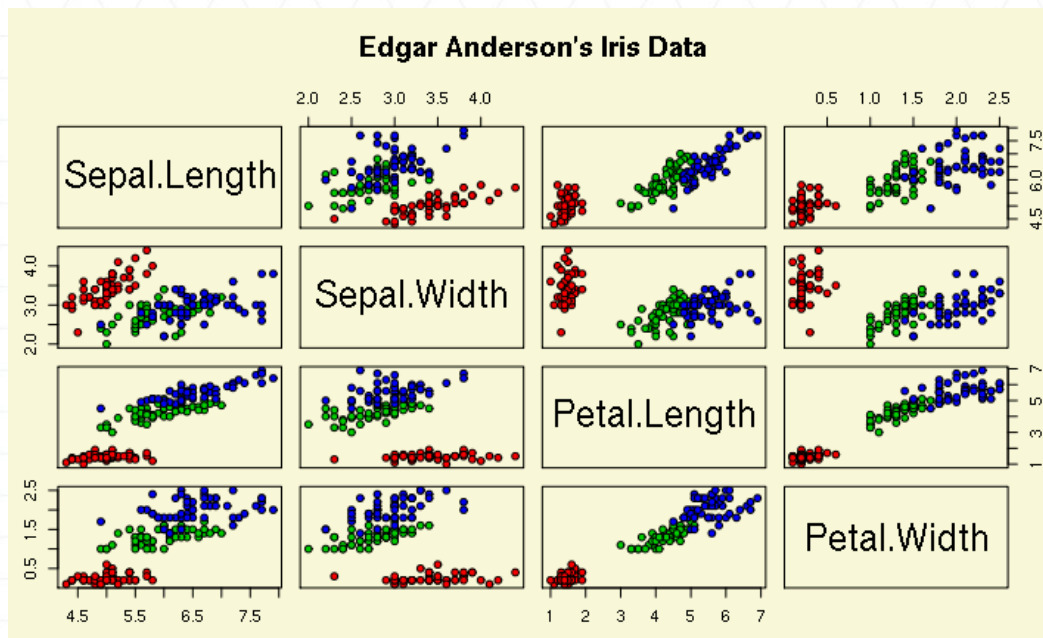
# 為什麼使用R

- 立即完成統計分析
  - 資料處理
  - 資料分析
  - 報表製作
- 內建許多數學函式及圖形套件(也可安裝第三方套件)
  - 可以結合其他語言：如Java, C++
- 免費且開源
  - <http://cran.r-project.org/src/base/>
  - 驚人的潛力和彈性
  - 容易擴充和客製化
  - 只要你願意且有能力，就可以貢獻並且改進



# 應用範圍

- 統計分析
- 迴歸分析
- 資料分群
- 資料分類
- 推薦系統
- 文字探勘



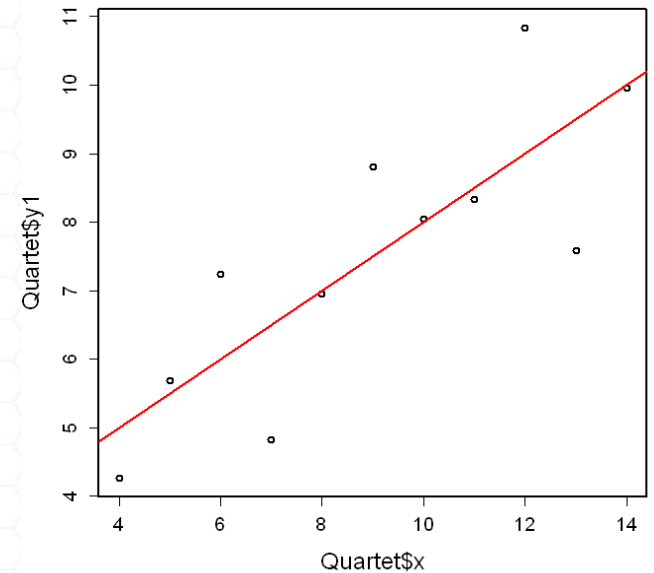


# 影像辨識

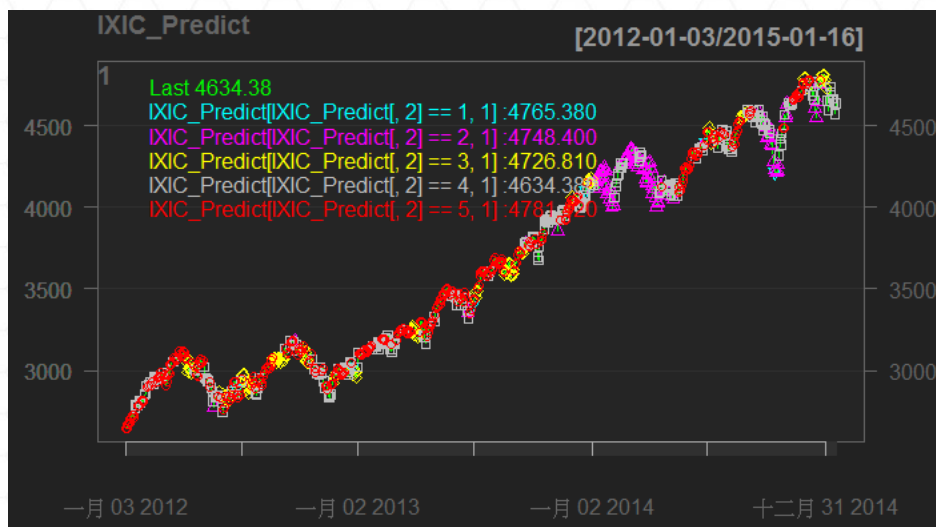


# 用R做簡單迴歸分析

```
data(anscombe)
plot(y1 ~ x1, data = anscombe)
lmfit <- lm(y1~x1, data=anscombe)
abline(lmfit, col="red")
```



# 更複雜的分析



預測股票

人臉辨識

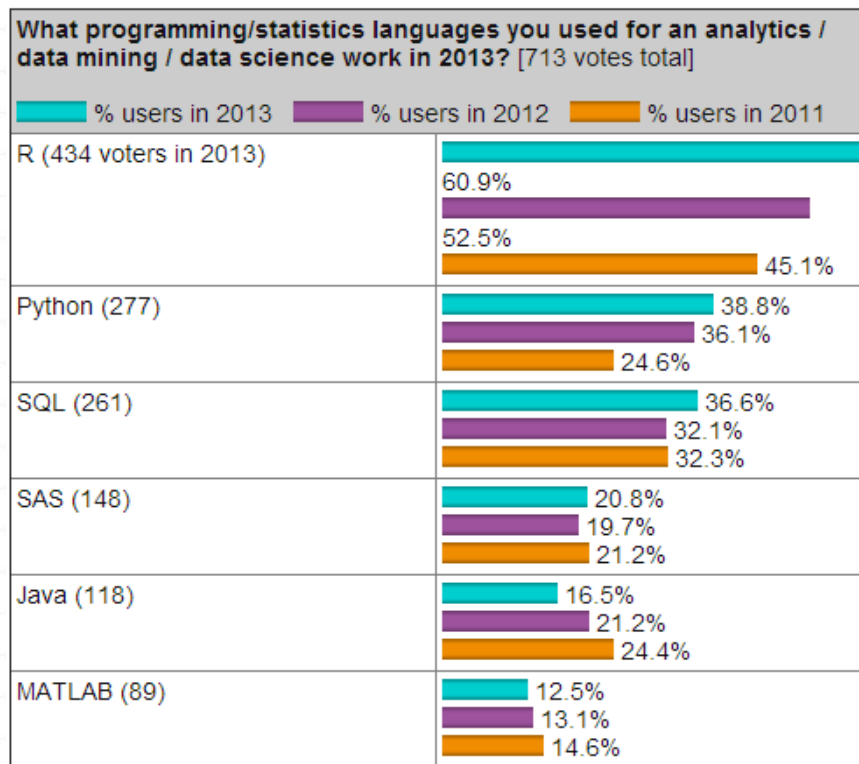




# 最廣泛被用來做資料分析的語言

最受歡迎的語言持續為 *R*, *Python* (39%), 及 *SQL* (37%). *SAS* 大約在 20% 上下.

By Gregory Piatetsky, Aug 27, 2013.



# R語言環境設定

# R Download

■ <https://cran.r-project.org/bin/windows/base/>

R-3.3.2 for Windows (32/64 bit)

[Download R 3.3.2 for Windows](#) (62 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

下載3.3.2 版本

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

## Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)




# Microsoft R Open

- Microsoft R Open, they **automatically use all available cores and processors to significantly reduce computation times**
- 使用數學核心函數庫 ( Math Kernel Libraries , MKL ) 優化多執行緒處理器 ( multi-threaded processor ) 來強化效能

MRAN About R Microsoft R Open Community Download

Find an R Package

Microsoft R Open: The Enhanced R Distribution



Microsoft R Open, formerly known as Revolution R Open (RRO), is **the enhanced distribution of R** from Microsoft Corporation. It is a complete open source platform for statistical analysis and data science.

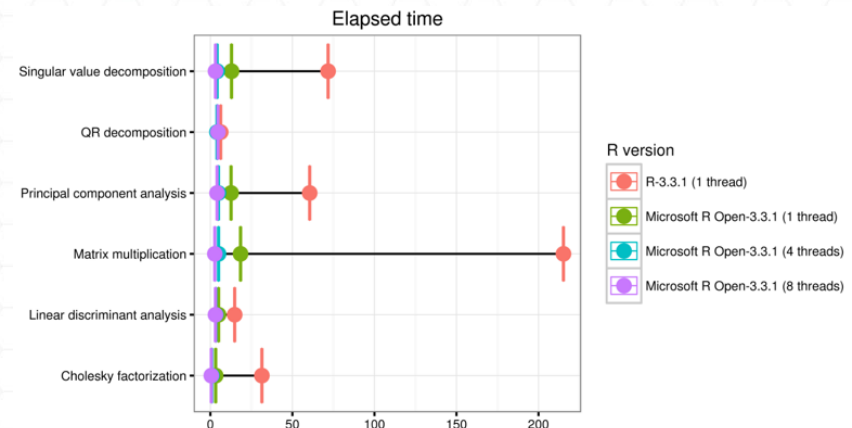
The current version, Microsoft R Open 3.2.5, is based on (and 100% compatible with) R-3.2.5, the most widely used statistics software in the world, and is therefore fully compatible with all packages, scripts and applications that work with that version of R. It includes additional capabilities for **improved performance, reproducibility**, as well as support for **Windows and Linux-based platforms**.

Like R, Microsoft R Open is open source and free to download, use, and share.

[Learn more...](#)

[DOWNLOAD](#)

[Release News](#)



<https://mran.revolutionanalytics.com/documents/rro/multithread/>


# 下載Microsoft R Open

■ <https://mran.microsoft.com/open/>

MRAN

About RMicrosoft R OpenCommunityDownload

Find an R Package




Microsoft R Open, formerly known as Revolution R Open (RRO), is **the enhanced distribution of R** from Microsoft Corporation. It is a complete open source platform for statistical analysis and data science.

The current version, Microsoft R Open 3.3.2, is based on (and 100% compatible with) R-3.3.2, the most widely used statistics software in the world, and is therefore fully compatible with all packages, scripts and applications that work with that version of R. It includes additional capabilities for **improved performance, reproducibility**, as well as support for **Windows and Linux-based platforms**.

Like R, Microsoft R Open is open source and free to download, use, and share.

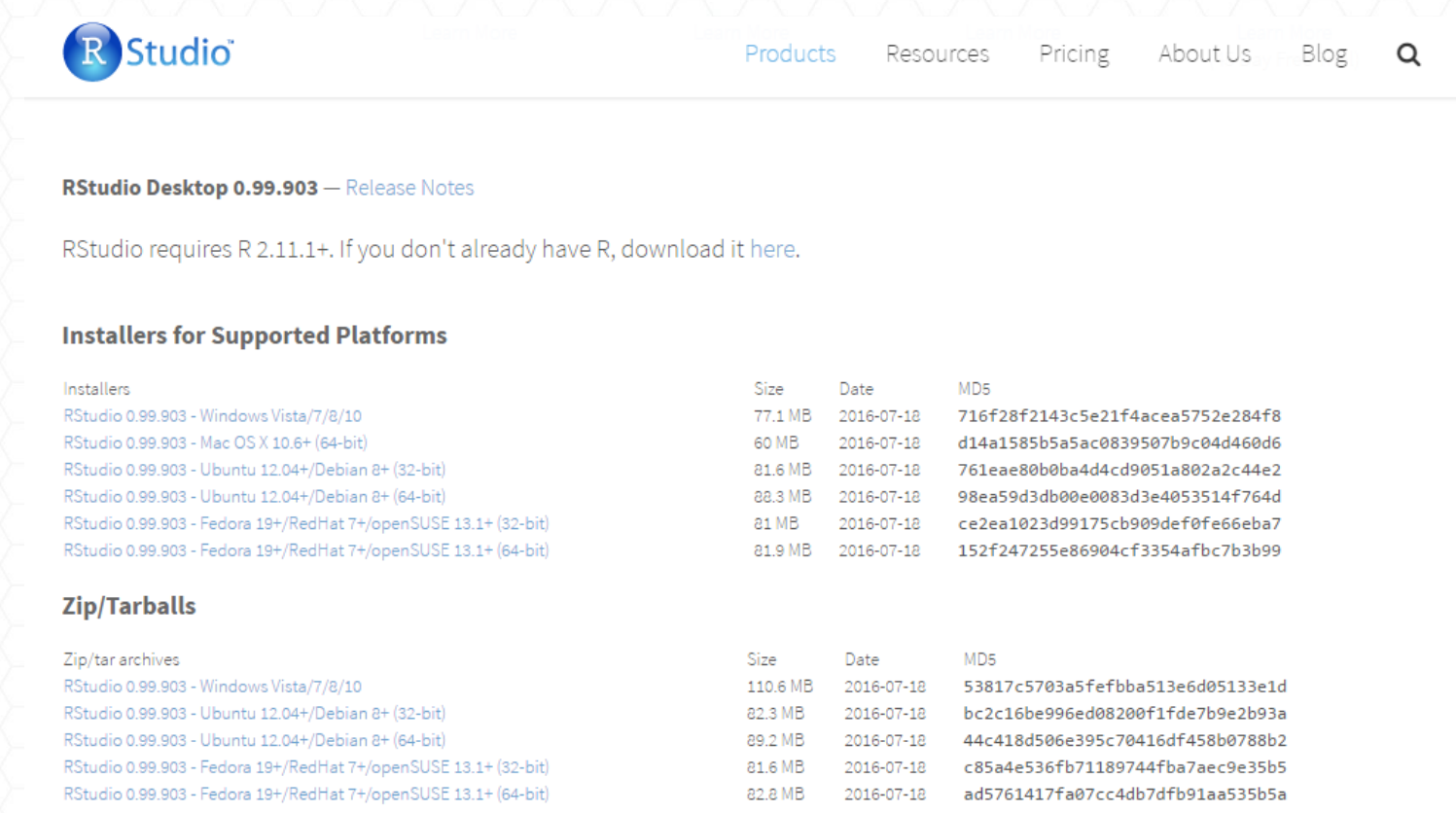
[Learn more...](#)

 **DOWNLOAD**

[Release News](#)

# 下載RStudio

■ <https://www.rstudio.com/products/rstudio/download3/>



The screenshot shows the RStudio website's download page. At the top is the RStudio logo and a navigation menu with links for Products, Resources, Pricing, About Us, and Blog. Below the navigation bar, the page title is "RStudio Desktop 0.99.903 — Release Notes". A paragraph states that RStudio requires R 2.11.1+ and provides a link to download R. The main content is divided into two sections: "Installers for Supported Platforms" and "Zip/Tarballs". Each section contains a table with columns for the installer name, size, date, and MD5 hash. The "Installers" section lists installers for Windows, Mac OS X, Ubuntu, and Fedora. The "Zip/Tarballs" section lists zip and tar archives for the same operating systems.

## RStudio Desktop 0.99.903 — Release Notes

RStudio requires R 2.11.1+. If you don't already have R, download it [here](#).

### Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	77.1 MB	2016-07-18	716f28f2143c5e21f4acea5752e284f8
RStudio 0.99.903 - Mac OS X 10.6+ (64-bit)	60 MB	2016-07-18	d14a1585b5a5ac0839507b9c04d460d6
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (32-bit)	81.6 MB	2016-07-18	761eae80b0ba4d4cd9051a802a2c44e2
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (64-bit)	88.3 MB	2016-07-18	98ea59d3db00e0083d3e4053514f764d
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	81 MB	2016-07-18	ce2ea1023d99175cb909def0fe66eba7
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	81.9 MB	2016-07-18	152f247255e86904cf3354afbc7b3b99

### Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	110.6 MB	2016-07-18	53817c5703a5fefbba513e6d05133e1d
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (32-bit)	82.3 MB	2016-07-18	bc2c16be996ed08200f1fde7b9e2b93a
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (64-bit)	89.2 MB	2016-07-18	44c418d506e395c70416df458b0788b2
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	81.6 MB	2016-07-18	c85a4e536fb71189744fba7aec9e35b5
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	82.8 MB	2016-07-18	ad5761417fa07cc4db7dfb91aa535b5a



# Rstudio

編輯區

歷史&環境

控制臺

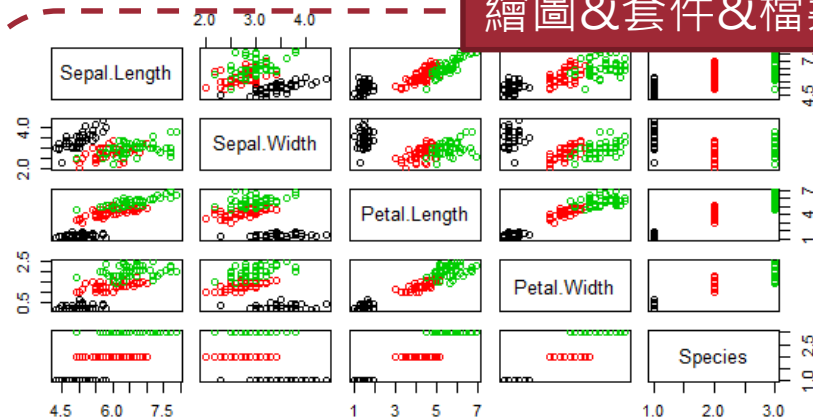
繪圖&套件&檔案

```
1 library(rvest)
2 appledaily <- html("http://www.berich.com.tw/DP/Cr
3 article <- appledaily %>% html_nodes("table") %>%
```

```
table(tw2330$tf)
hist(tw2330$Close)
pairs(iris)
pairs(iris, col="iris$Species")
pairs(iris, col=iris$Species)
```

[Workspace loaded from ~/.RData]

```
> pairs(iris)
> pairs(iris, col="iris$Species")
Error in plot.xy(xy, type, ...) : invalid color name 'iris$Species'
> pairs(iris, col=iris$Species)
> |
```



# R 語言基礎

# 數學運算

# 數字相加

3 + 8

# 數字相減

3 - 8

# 數字相乘

5 \* 5

# 數字相除

11 / 2

# 指數

2^10

# 取餘數

11%%2

可以將R 當成計算機使用





# 設定變數

# 指定變數

a <- 3

a

可以使用 = 或 <- 指定變數

# 變數相加

b <- 5

c <- a + b

c

# 基礎資料型態

# 數值型態

numer <- 17.8

# 字串型態

char <- "hello world"

# 布林邏輯

logic <- TRUE

# 使用class 檢查資料型態

class(logic)

# 不同型態資料做運算

```
card_length <- 3
```

```
card_width <- "5 inches"
```

```
card_length * card_width
```

```
Error in card_length * card_width :  
  non-numeric argument to binary operator
```

```
#重新將card_width 指到5
```

```
card_width <- 5
```

```
card_length * card_width
```



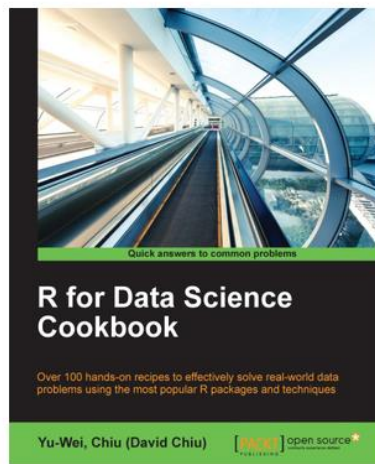
# 計算一本書的價錢

RRP <- 35.99

Exchange <- 31.74

NTD <- RRP \* Exchange

NTD



## R for Data Science Cookbook

Yu-Wei, Chiu (David Chiu)  
July 2016



★★★★★ feefo  
1 customer reviews

Over 100 hands-on recipes to effectively solve real-world data problems using the most popular R packages and techniques

\$35.99

RRP \$35.99

☒ eBook

☐ Print + eBook



Add to Cart

# 向量 (Vector)

# 使用向量存放多個變數的資料

# 不同型態的向量

```
height_vec <- c(180,169,173)
```

```
name_vec <- c("Brian", "Toby", "Sherry")
```





# 向量的運算

# 兩個向量進行數學運算

$x \leftarrow c(1, 2, 3, 7)$

$y \leftarrow c(2, 3, 5, 1)$

$x + y$

$x * y$

$x - y$

$x / y$

# 自動產生向量

## ■ 產生1到20

```
x <- 1:20
```

```
x
```

```
y <- seq(1,20)
```

```
y
```

## ■ 使用? 或help 去觀看seq 的用法

```
?seq
```

```
help(seq)
```

# 將向量作加總

# 透過sum 將向量資料作加總

```
x <- c(1,2,3,5,7)
```

```
sum(x)
```

# 查詢該如何使用sum函式

```
?sum
```

```
help(sum)
```



# 指定名稱

- 可以使用names 指定向量名稱

```
height_vec <- c(180,169,173)
```

```
height_vec
```

```
names(height_vec) <- c("Brian", "Toby", "Sherry")
```

```
height_vec
```

```
name_vec <- c("Brian", "Toby", "Sherry")
```

```
names(height_vec) <- name_vec
```

# 判斷向量內容是否符合條件

`height_vec > 175`

`height_vec < 175`

`height_vec >= 175`

`height_vec <= 175`

`height_vec == 180`

`height_vec != 180`

■ 可以篩選符合條件的資料

`height_vec[height_vec > 175]`

# 使用向量計算BMI

- Brian的身高為180, 體重是73公斤;Toby身高是169公分, 體重是87公斤; Sherry身高為173公分,體重是 43公斤。請用Vector找出誰的BMI是異常的?
- BMI值計算公式:  $BMI = \text{體重(公斤)} / \text{身高}^2(\text{公尺}^2)$

	身體質量指數(BMI) (kg/m <sup>2</sup> )
體重過輕	$BMI < 18.5$
正常範圍	$18.5 \leq BMI < 24$
異常範圍	過重: $24 \leq BMI < 27$ 輕度肥胖: $27 \leq BMI < 30$ 中度肥胖: $30 \leq BMI < 35$ 重度肥胖: $BMI \geq 35$



# 陣列 (Matrix)

# 建立陣列

## ■ 學生兩次考試的成績

```
kevin <- c(85,73)
```

```
marry <- c(72,64)
```

```
jerry <- c(59,66)
```

```
mat <- matrix(c(kevin, marry, jerry), nrow=3,  
byrow= TRUE)
```

# 新增欄位與列的名稱

```
colnames(mat) <- c('first', 'second')  
rownames(mat) <- c('kevin', 'marry', 'jerry')
```

OR

```
mat2 <- matrix(c(kevin, marry, jerry), nrow=3, byrow=TRUE,  
dimnames=list(c('kevin', 'marry', 'jerry'), c('first', 'second')))
```



# 取矩陣維度、列與欄數

- 取維度

`dim(mat2)`

- 取列數

`nrow(mat2)`

- 取行數

`ncol(mat2)`

# 依欄或列取矩陣資料

- 取第一列

`mat2[1,]`

- 取第一行

`mat2[:,1]`

- 取第二、三列

`mat2[2:3,]`

- 取第二列第一行的元素

`mat2[2,1]`

# 新增列與行

## ■ 新增學生資料

```
mat3 <- rbind(mat2, c(78,63))  
rownames(mat3)[nrow(mat3)] <- 'sam'  
mat3
```

## ■ 新增考試分數

```
mat4 <- cbind(mat2, c(82,77,70))  
colnames(mat4)[ncol(mat4)] <- 'third'  
mat4
```



# 使用rowSums 及colSums

- 使用rowSums 及 colSums 針對列及欄加總

rowSums(mat2)

colSums(mat2)

# 矩陣運算

## ■ 矩陣宣告

```
m1 <- matrix(1:4, byrow=TRUE, nrow=2)
```

```
m2 <- matrix(5:8, byrow=TRUE, nrow=2)
```

## ■ 矩陣運算

```
m1 + m2
```

```
m1 - m2
```

```
m1 * m2
```

```
m1 / m2
```

# 使用矩陣計算考試成績

## ■ 學生兩次考試的成績

```
kevin <- c(85,73)
```

```
marry <- c(72,64)
```

```
jerry <- c(59,66)
```

```
mat <- matrix(c(kevin, marry, jerry), nrow=3,  
byrow= TRUE)
```

- 如果老師希望給每個人最後總成績，以加權為第一次考試佔40%，第二次佔60%；請問該怎麼用矩陣運算達成？



# 階層 (Factor)

# 將資料轉換為類別資料(Factor)

```
Weather <- c("sunny", "rainy", "cloudy", "rainy",  
"cloudy")
```

```
weather_category <- factor(weather)
```

```
weather_category
```

```
levels(weather_category)
```

character 跟 Factor 屬於不同東西  
請善用class 檢查資料型態

# 清單(Lists)



# 使用list 包裝不同類型資料

## ■ 使用list 包裝類型不同的資料

```
person <- list(name='James', height=180, Employ=TRUE)  
person
```

## ■ 使用lapply 套用函式到list 裡面的元素

```
li = list( c(98,82,66,54), c(83,72,77))  
lapply(li, sum)
```

# Data Frame

# 如何建立Data Frame

# 建立 Vector

```
days <- c('mon','tue','wed','thu','fri')
```

```
temp <- c(22.2,21,23,24.3,25)
```

```
rain <- c(TRUE, TRUE, FALSE, FALSE, TRUE)
```

# 使用 Vector 建立Data Frame

```
df <- data.frame(days,temp,rain)
```

df



# 檢視 Data Frame

# 檢視資料形態

`class(df)`

# 檢視架構

`str(df)`

# 檢視資料摘要

`summary(df)`

# 使用R 內建的資料集

- 表列資料集

`data()`

- 使用資料集

`data(iris)`

- 觀察讀取到的資料集型態

`class(iris)`

# Iris 資料集

■ [http://en.wikipedia.org/wiki/Iris\\_flower\\_data\\_set](http://en.wikipedia.org/wiki/Iris_flower_data_set)



*Iris setosa*



*Iris versicolor*



*Iris virginica*



# 觀看資料集的前幾筆資料與後幾筆資料

## ■ 觀看前幾筆資料

`head(iris)`

`head(iris, 10)`

## ■ 觀看後幾筆資料

`tail(iris)`

`tail(iris, 10)`

請善用?檢視  
函式說明

# 取得指定列與行的部分資料集

- 取前三列資料

```
iris[1:3,]
```

- 取前三列第一行的資料

```
iris[1:3,1]
```

- 也可以用欄位名稱取值

```
iris[1:3,"Sepal.Length"]
```

- 取前兩行資料

```
iris[,1:2]
```

取特定欄位向量值

```
iris$"Sepal.Length"
```

df[列, 欄]

# 資料篩選

- 取前五筆包含length 及 width 的資料

```
five.Sepal.iris <- iris[1:5, c("Sepal.Length",  
"Sepal.Width")]
```

- 可以用條件做篩選

```
setosa.data <- iris[iris$Species=="setosa",1:5]
```

- 使用which 做資料篩選

```
which(iris$Species=="setosa")
```



# 資料排序

- 用Sort 作資料排序

```
sort(iris$Sepal.Length, decreasing = TRUE)
```

- 用order做資料排序

```
iris[order(iris$Sepal.Length, decreasing = TRUE),]
```

# 實價登錄資料分析案例

# 不動產交易實價資料

■ <http://plvr.land.moi.gov.tw/DownloadOpenData>

內政部  
不動產交易實價查詢服務網

歡迎蒞臨不動產交易實價查詢服務網  
累積人數：0072528321  
線上人數：000416



NEWS ▶ 內政部105年幸福列車單身聯誼開始報名，內政部補助活動費用2成，歡迎單身未婚男女踴躍參加。

系統訊息 系統維護訊息

- 9月1日提供7月1-15日成交案件查詢及下載，歡迎查詢利用。 105.09.01
- 內政部105年幸福列車單身聯誼開始報名，內政部補助活動費用2成，第一場苗栗花露農場 105.06.29  
新竹南寮漁港，第2場高雄農場高雄外港，6月26日截止報名；第3場新店碧潭，7月3日截止報名，歡迎參加。...[活動連結](#)
- 105年3月15日起opendata及付費提供批次資料原提供「交易年月」增提供「交易日期」。 105.03.15
- 本部實施「不動產實價登錄」制度，推動各項地政便民服務措施，及財政部實行網路報稅 104.12.02  
等措施，降低申請不動產移轉登記所需稅捐成本。因此世界銀行於經商環境(Doing

## 重要公告

提醒民眾及房仲業者：不動產仲介經紀業服務報酬之計收，主管機關並未規定固定收費比率，消費者可與房仲業者自行議訂，其向買賣或租賃之一方或雙方收取報酬之總額，合計不得超過該不動產實際成交價金6%或1.5個月租金。



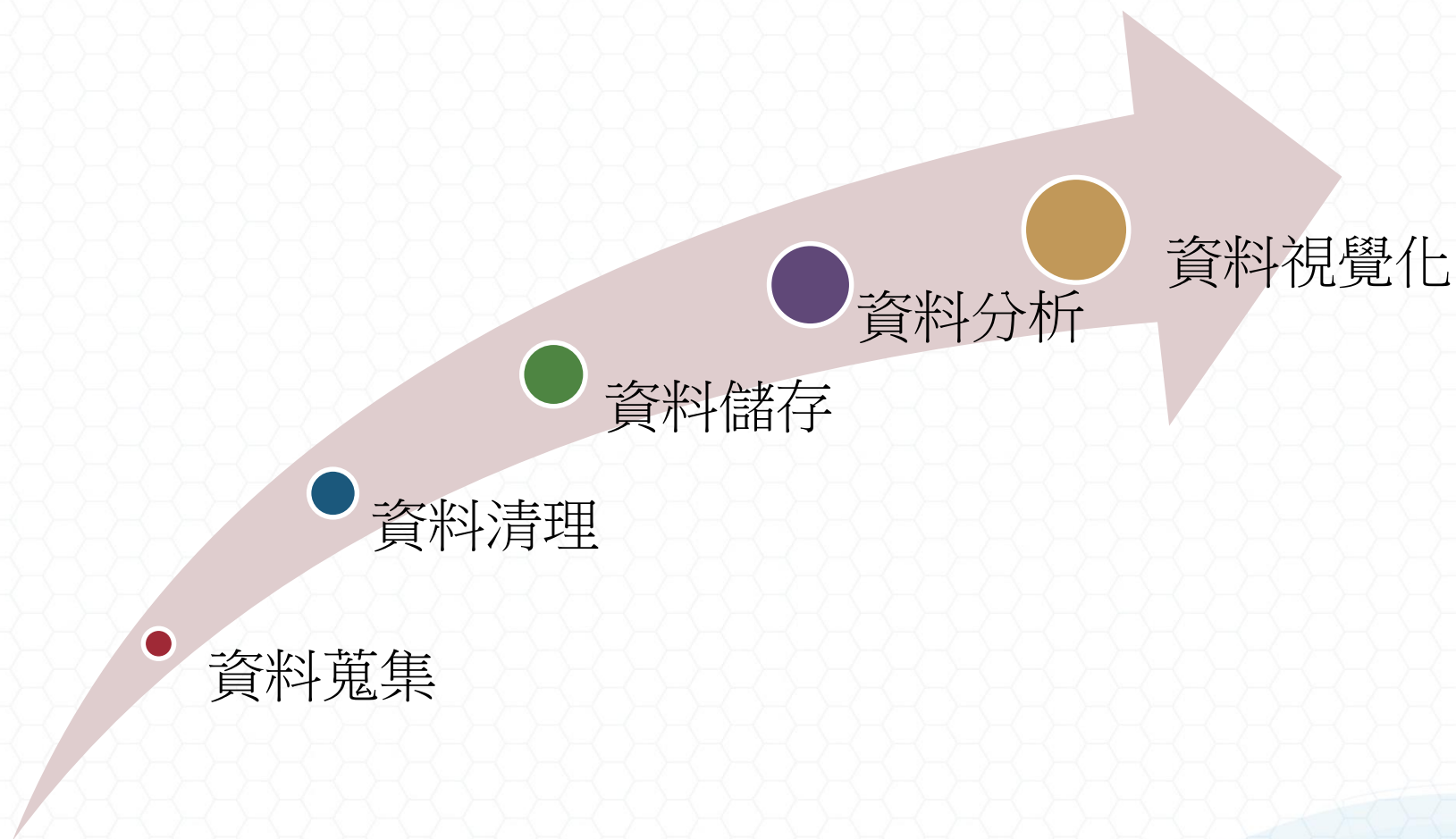
App及文件下載



查詢須知



# 數據科學步驟



# 資料蒐集

# 透過R操作一般文字檔案

- 檔案是最基本的資料保存模式
- 透過檔案，我們可以保存與讀取資料

1, David, M  
2, Mary, F  
3, John, M





# 使用R 讀取csv檔案

## ■ 檢視目錄所在

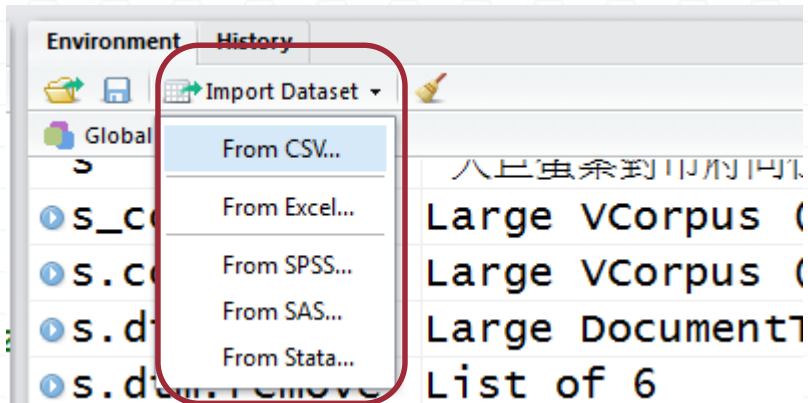
`getwd()`

可以使用`setwd` 修改目錄位置

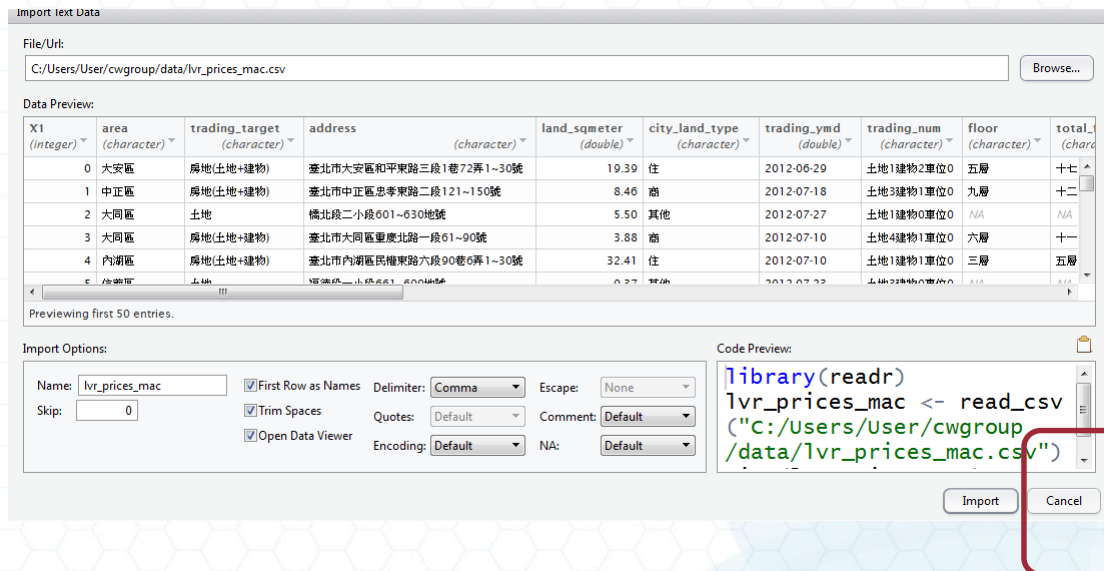
## ■ 下載檔案資料

`download.file('https://github.com/ywchiu/ipost/raw/master/data/lvr_prices_mac.csv', 'lvr_price.csv')`

# 使用Rstudio 讀取檔案



也可以用read.csv 讀取檔案資料



# 資料探索



# 資料探索

- 計算大安區所有成交物件的的總價(total\_price)總合

```
daan <- lvr_price[lvr_price$area == '大安區',]  
sum(as.numeric(daan$total_price), na.rm = TRUE)
```

- 計算總價(total\_price)平均

```
mean(as.numeric(daan$total_price), na.rm = TRUE)
```

## 資料探索(II)

- 列舉中山區所有成交物件總價(total\_price)前三名的地址(address)以及(total\_price)

```
zhongshan <- lvr_price[lvr_price$area == '中山區',c('address', 'total_price')]
```

```
idx <- order(zhongshan$total_price, decreasing = TRUE)
```

```
res <- zhongshan[idx,]
```

```
res[1:3,]
```

也可以換成head(res,3)

那如果想要換區統計呢？是否可以將動作包裝成函式？

# 函式 (Function)

- 回傳值為最後被執行的語句

```
f = function(<arguments>) {  
    #任何腳本  
}
```

- 可帶預設參數

```
f = function(a, b = 2, c = NULL) {  
}
```



# 將資料探索的程式包裝成函式

```
getTopThree <- function(area){  
  zhongshan <- lvr_price[lvr_price$area == area,]  
  idx <- order(zhongshan$total_price, decreasing = TRUE)  
  res <- zhongshan[idx,c('area', 'address', 'total_price')]  
  return(res[1:3,])  
}
```

```
getTopThree('大安區')
```

那如果不想一一打入台北市12 區的話  
如何該列舉各區的統計數據？

# tapply

計算屬性

分組條件

```
■ tapply(lvr_price$total_price, lvr_price$area,  
function(e)mean(e,na.rm=TRUE))
```

匿名函式

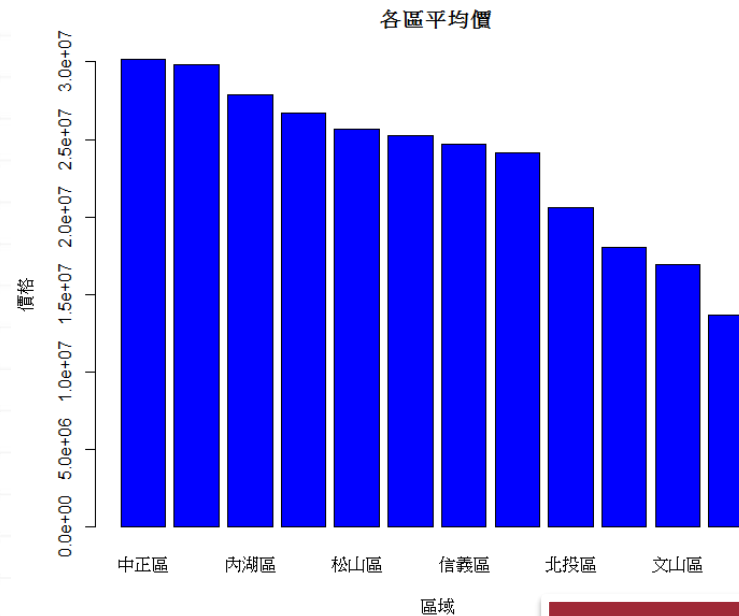
士林區	大同區	大安區	中山區	中正區	內湖區	文山區
24139903	18063872	29798170	26708805	30154011	27905514	16953869
北投區	松山區	信義區	南港區	萬華區		
20626410	25652125	24725051	25235793	13642289		

該如何將這些資料視覺化?

# 使用長條圖比較平均房價高低

```
price_per_sec <- tapply(lvr_price$total_price, lvr_price$area,  
function(e)mean(e,na.rm=TRUE))
```

```
barplot(sort(price_per_sec, decreasing = TRUE), main= "各區平均價",  
xlab = "區域", ylab = "價格", col="blue")
```

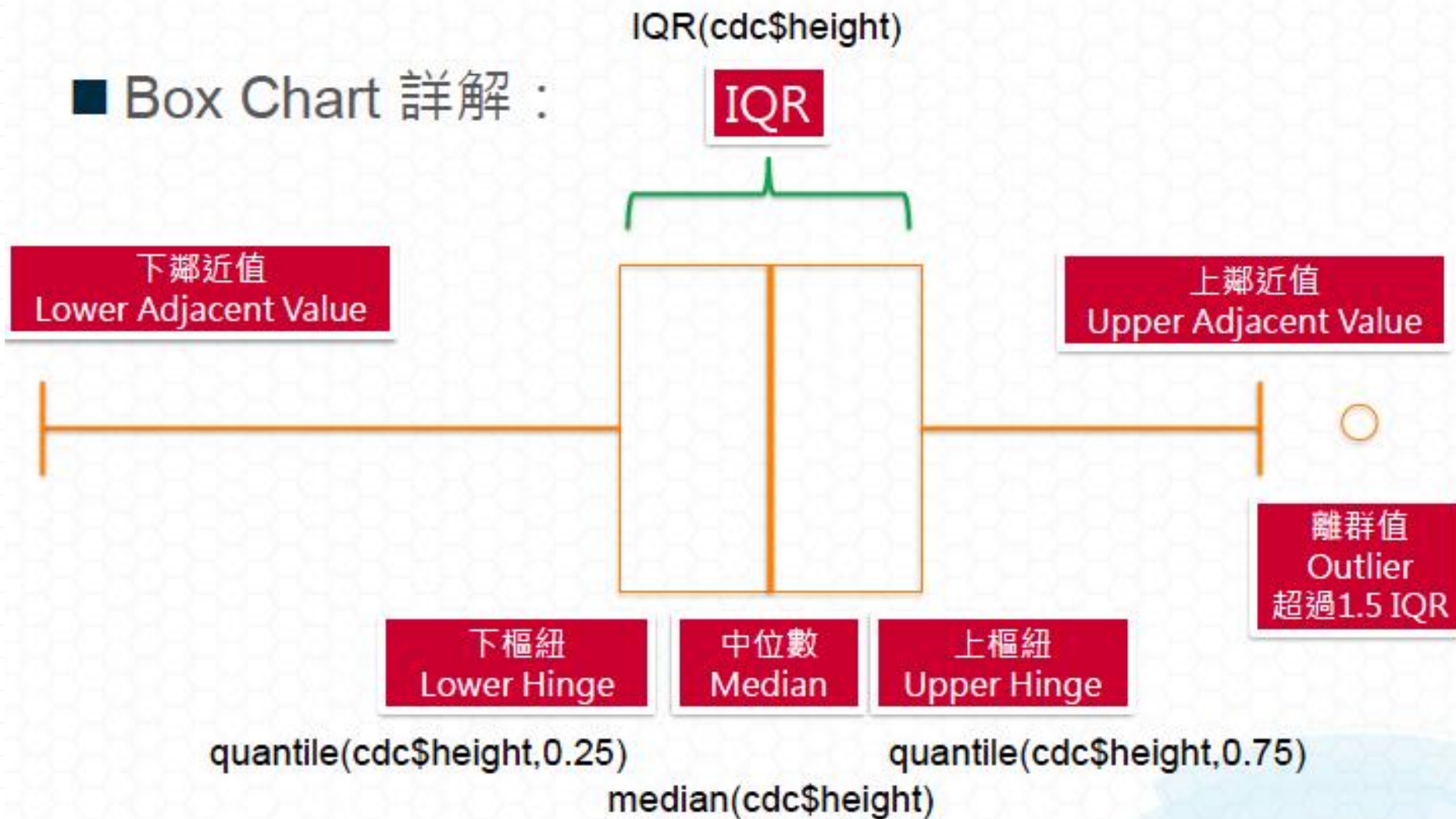


但房屋買賣價格很大，要考慮離群值



# Box Chart

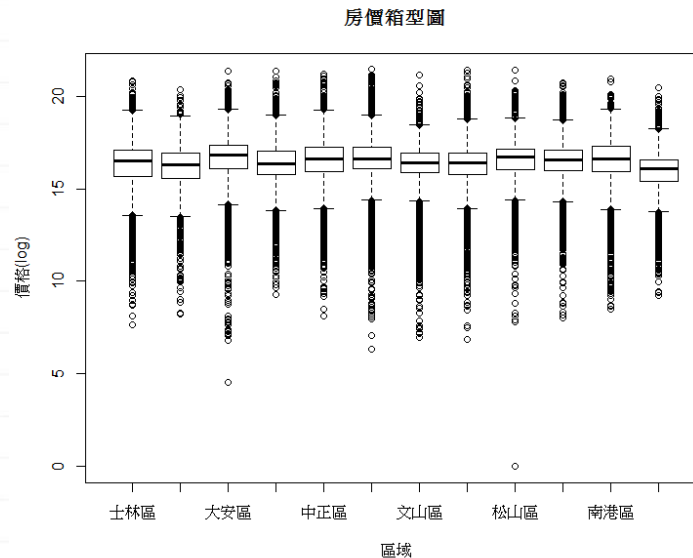
## ■ Box Chart 詳解：



# 繪製箱形圖

- 繪製台北市各區域(根據Area 區分)房價的箱型圖  
(Y軸為總價 total\_price)

`boxplot(log(total_price) ~ area, data = lvr_price, main= "房價箱型圖", xlab = "區域", ylab = "價格(log)")`



# 如果想要繪製各區不同時間點的房價走勢？

- 但 `tapply` 只能針對一個變數(區域或時間)做聚合，是否有其他解決方案？
- 使用 `dplyr`
  - 提供操作資料的基本語法
    - `filter`, `select`, `arrange`, `mutate`, `summarise`, `group_by`
  - 提供資料合併功能(`JOIN`)
    - `Inner join`, `left join`
  - 可以操作資料表(`data table`) 或資料庫 (`Database`)的資料



# 安裝與使用dplyr

## ■ 安裝dplyr

- `install.packages("dplyr")`

## ■ 使用dplyr

- `library(dplyr)`

## ■ 觀看說明頁

- `help(package='dplyr')`

# 資料篩選

## ■ dplyr 的過濾功能

```
filter(lvr_price, area == '中山區')
```

## ■ dplyr 的欄位選取

```
select(lvr_price, total_price)
```

# 但如果我想同時選擇欄位又過濾資料呢？

## ■ 鏈接(Chaining)

- %>% (Then)

- 來自 magrittr

## ■ 使用Then (%>%)

```
lvr_price %>%
```

```
  select(area, total_price) %>%
```

```
  filter(area == '中山區')
```



# 資料做排序

- 使用Arrange 可以將資料做排序

```
lvr_price %>%  
  select(area, total_price) %>%  
  filter(area == '中山區') %>%  
  arrange(total_price)
```

- 由大到小排序 (desc)

```
lvr_price %>%  
  select(area, total_price) %>%  
  filter(area == '中山區') %>%  
  arrange(desc(total_price))
```

如同

```
SELECT total_price, area  
FROM lvr_price  
WHERE area = "中山區"  
ORDER BY total_price
```

# 分組計算 (group\_by, summarise)

## ■ 分組計算函式

- group\_by: 分組依據

- summarise: 依組別計算結果

## ■ 統計各區域各年月(2012年1月1日後)的價格總和

```
lvr_price$trading_ym <- as.Date(format(lvr_price$trading_ymd, '%Y-%m-01'))
```

```
lvr_stat <- lvr_price %>%
```

```
  select(trading_ym, area, total_price) %>%
```

```
  filter(trading_ym >= '2012-01-01') %>%
```

```
  group_by(trading_ym, area) %>%
```

```
  summarise(overall_price = sum(as.numeric(total_price), na.rm=TRUE))
```

# 繪製房價變化

- 使用折線圖繪製台北市各區域(根據Area 區分)從2012 年至今每月的房價。(X軸為交易月，Y軸為總價total\_price)

產生 3 X 4 的圖表

使用for 迴圈繪製各區域圖

```
par(mfrow=c(3,4))  
for (a in levels(lvr_stat$area)){  
  plot(overall_price ~ trading_ym  
    ,lvr_stat[lvr_stat$area == a,]  
    , type='l', main = a)  
}
```

type	description
p	點。
l	直線。
o	點+直線。(兩者重疊)
b	點+直線。(不重疊)
c	點+直線。(點為空白)
S/s	階梯狀。
h	Histogram狀。
n	空樣式。



# 產生pivot table

## ■ 可以使用tidyr 套件產生pivot table

```
library(tidyr)
```

```
price_pivot <- spread(lvr_stat, trading_ym, overall_price, fill=0)
```

```
View(price_pivot)
```

	area	2012-01-01	2012-02-01	2012-03-01	2012-04-01	2012-05-01	2012-06-01	2012-07-01	2012-08-01	2012-09-01	2012-10-01
1	士林區	661140000	231680000	359891504	205481036	2539010528	514470692	1612810630	2627628833	4072187784	5481187784
2	大同區	0	180150000	525210000	87110000	74806000	173360800	302473961	2056051662	1168712672	1168712672
3	大安區	139136600	123870000	64470991	127736080	49052000	668836500	2953852056	4914767352	5015582531	5015582531
4	中山區	17601653	140250000	176022439	3156689238	2374390095	925510000	3036981800	4692015079	7291463744	7291463744
5	中正區	258200000	112690000	660420000	935400000	550190000	364040000	1463004513	4498343873	3952174637	3952174637
6	內湖區	349930000	216810000	299907515	944724354	1444681765	615189000	2529917498	5319499128	8621316230	8621316230
7	文山區	166887497	147810000	553478681	739757581	192890000	297841800	1245614572	2878918358	2197557340	2197557340
8	北投區	43850000	68000	82490000	494942526	899718610	732233963	2308572984	4394034989	3995621228	3995621228
9	松山區	0	0	405003695	554400	405400000	270720000	1216371195	2230302414	3043203633	3043203633
10	信義區	40800000	177020000	1269564574	2517094802	1241890000	647360000	2809067000	2790734498	4376895822	4376895822

# 將結果存回檔案中

## ■ 使用write.csv

```
write.csv(price_pivot, 'taipei_house_price.csv')
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		area	2012/1/1	2012/2/1	2012/3/1	2012/4/1	2012/5/1	2012/6/1	2012/7/1	2012/8/1	2012/9/1	2012/10/1	2012/11/1	2012/12/1	2013/1/1	2013/2/1	2013/3/1
2	1	士林區	6.61E+08	2.32E+08	3.6E+08	2.05E+08	2.54E+09	5.14E+08	1.61E+09	2.63E+09	4.07E+09	4.35E+09	5.26E+09	9.21E+09	4.07E+09	3.01E+09	6.5E+09
3	2	大同區	0	1.8E+08	5.25E+08	87110000	74806000	1.73E+08	3.02E+08	2.06E+09	1.17E+09	3.74E+09	2.36E+09	2.2E+09	1.51E+09	8.8E+08	2.08E+09
4	3	大安區	1.39E+08	1.24E+08	64470991	1.28E+08	49052000	6.69E+08	2.95E+09	4.91E+09	5.02E+09	7.11E+09	5.24E+09	7.4E+09	5.08E+09	4.66E+09	7.33E+09
5	4	中山區	17601653	1.4E+08	1.76E+08	3.16E+09	2.37E+09	9.26E+08	3.04E+09	4.69E+09	7.29E+09	8.88E+09	8.14E+09	1.12E+10	8.48E+09	5.1E+09	1.22E+10
6	5	中正區	2.58E+08	1.13E+08	6.6E+08	9.35E+08	5.5E+08	3.64E+08	1.46E+09	4.5E+09	3.95E+09	7.33E+09	4.2E+09	5.67E+09	4.18E+09	2.47E+09	3.7E+09
7	6	內湖區	3.5E+08	2.17E+08	3E+08	9.45E+08	1.44E+09	6.15E+08	2.53E+09	5.32E+09	8.62E+09	7.08E+09	1.32E+10	1.11E+10	5.84E+09	6.36E+09	8.71E+09
8	7	文山區	1.67E+08	1.48E+08	5.53E+08	7.4E+08	1.93E+08	2.98E+08	1.25E+09	2.88E+09	2.2E+09	4.29E+09	3.56E+09	5.15E+09	2.64E+09	3.15E+09	4.16E+09
9	8	北投區	43850000	68000	82490000	4.95E+08	9E+08	7.32E+08	2.31E+09	4.39E+09	4E+09	3.21E+09	4.49E+09	6.23E+09	5.79E+09	3.97E+09	5.89E+09
10	9	松山區	0	0	4.05E+08	554400	4.05E+08	2.71E+08	1.22E+09	2.23E+09	3.04E+09	5.14E+09	4.11E+09	6.29E+09	2.36E+09	1.76E+09	4.58E+09
11	10	信義區	40800000	1.77E+08	1.27E+09	2.52E+09	1.24E+09	6.47E+08	2.81E+09	2.79E+09	4.38E+09	4.01E+09	6.86E+09	6.35E+09	3.42E+09	2.88E+09	4.89E+09
12	11	南港區	53560259	1.45E+08	4.3E+08	4.53E+08	7.35E+08	2.12E+08	1.43E+09	2.33E+09	4.29E+09	5.97E+09	2.85E+09	5.07E+09	1.95E+09	2.31E+09	3.81E+09
13	12	萬華區	17430000	14800000	7800000	5.14E+08	3.12E+08	88610000	7.59E+08	1.81E+09	1.32E+09	1.77E+09	1.79E+09	3.16E+09	1.19E+09	1.01E+09	2.09E+09

The background features a light blue hexagonal grid pattern. Overlaid on this is a large, faint, concentric circular design that resembles a stylized spiral or a series of overlapping rings. The text "THANK YOU" is centered in a bold, dark blue font.

**THANK YOU**