

R 語言基礎教學

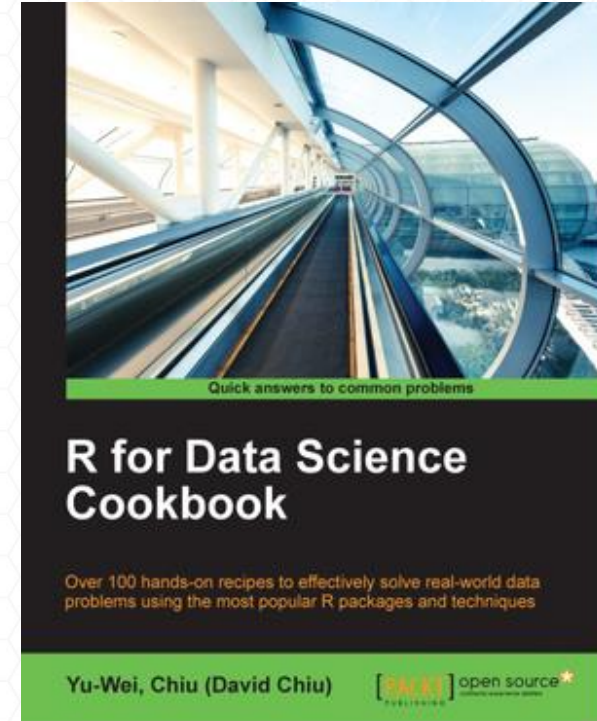
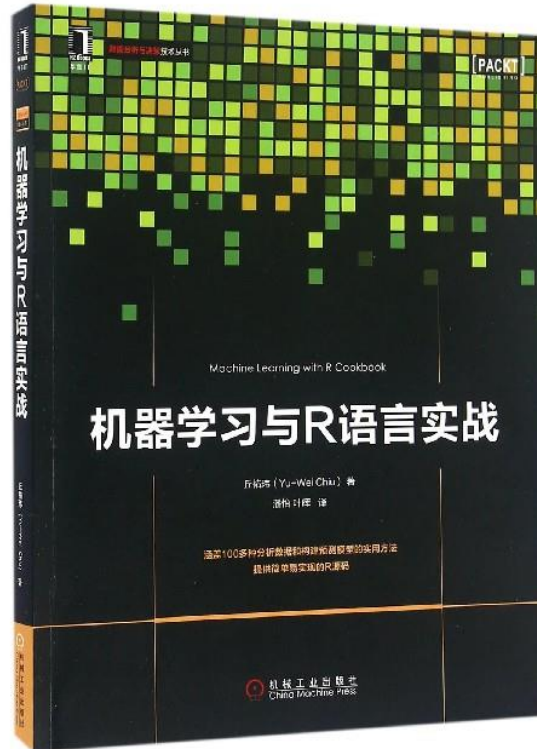
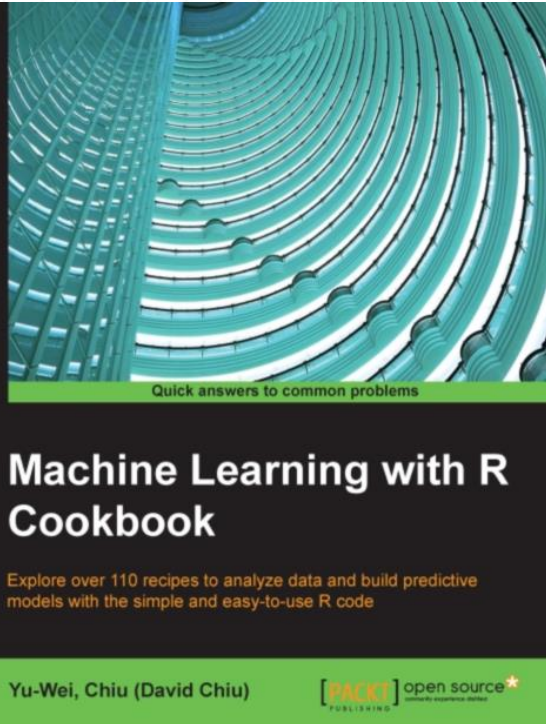
David Chiu
2017/01/16

關於我



- 大數軟體有限公司創辦人
- 前趨勢科技工程師
- ywchiu.com
- 大數學堂
<http://course.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

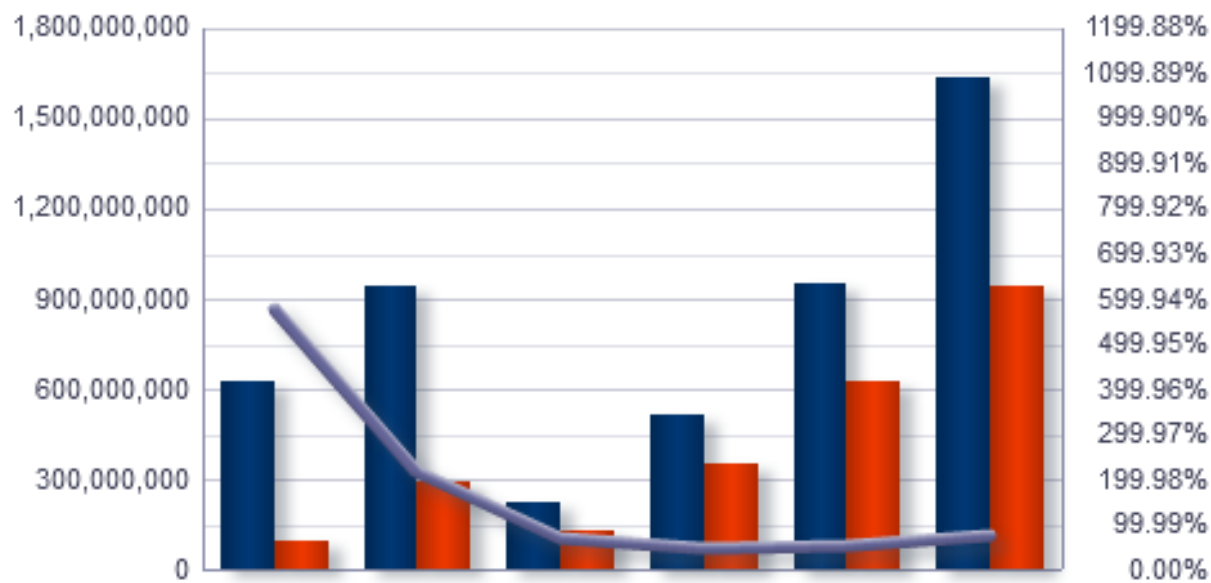
環境資訊頁面

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

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



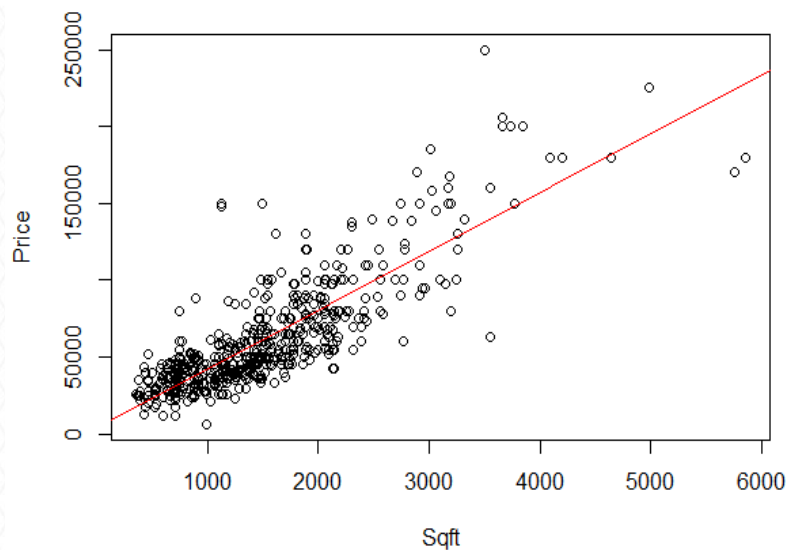
■ 軟體工程師派的

- 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

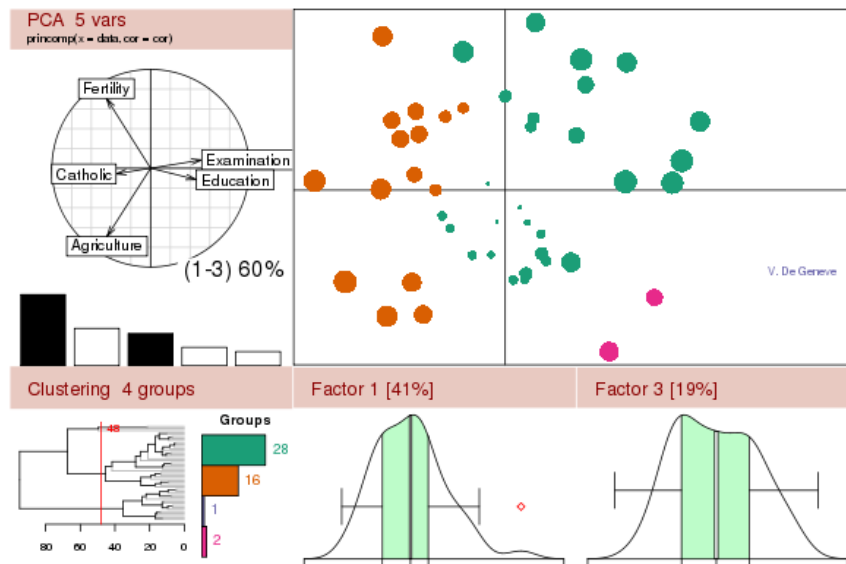
我也更出意見 體驗新版

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

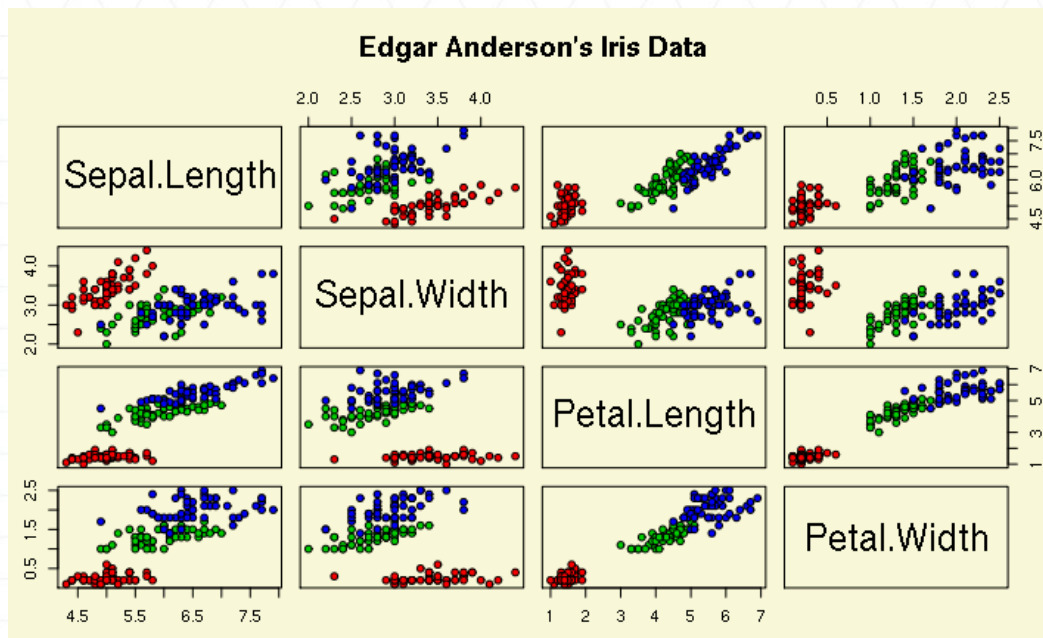
為什麼使用R

- 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

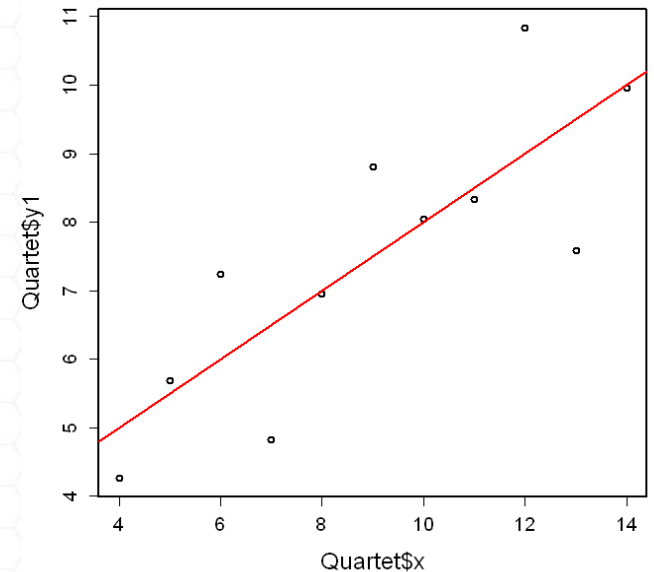


影像辨識

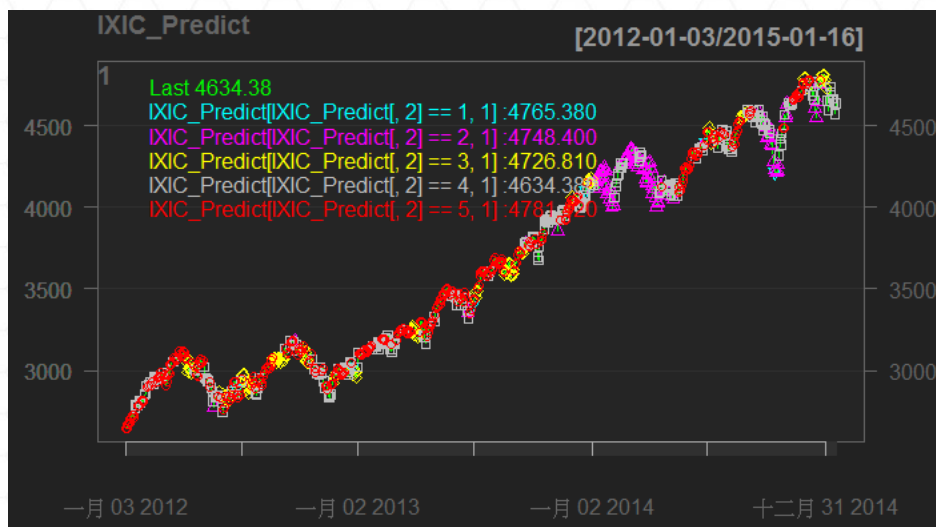


用R做簡單迴歸分析

```
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

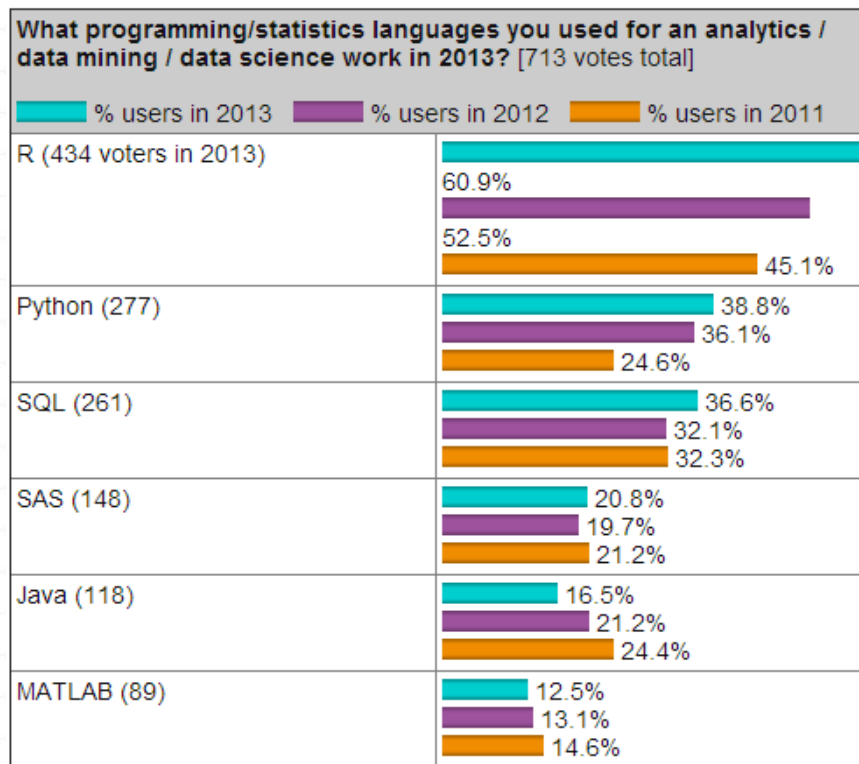
人臉辨識



最廣泛被用來做資料分析的語言

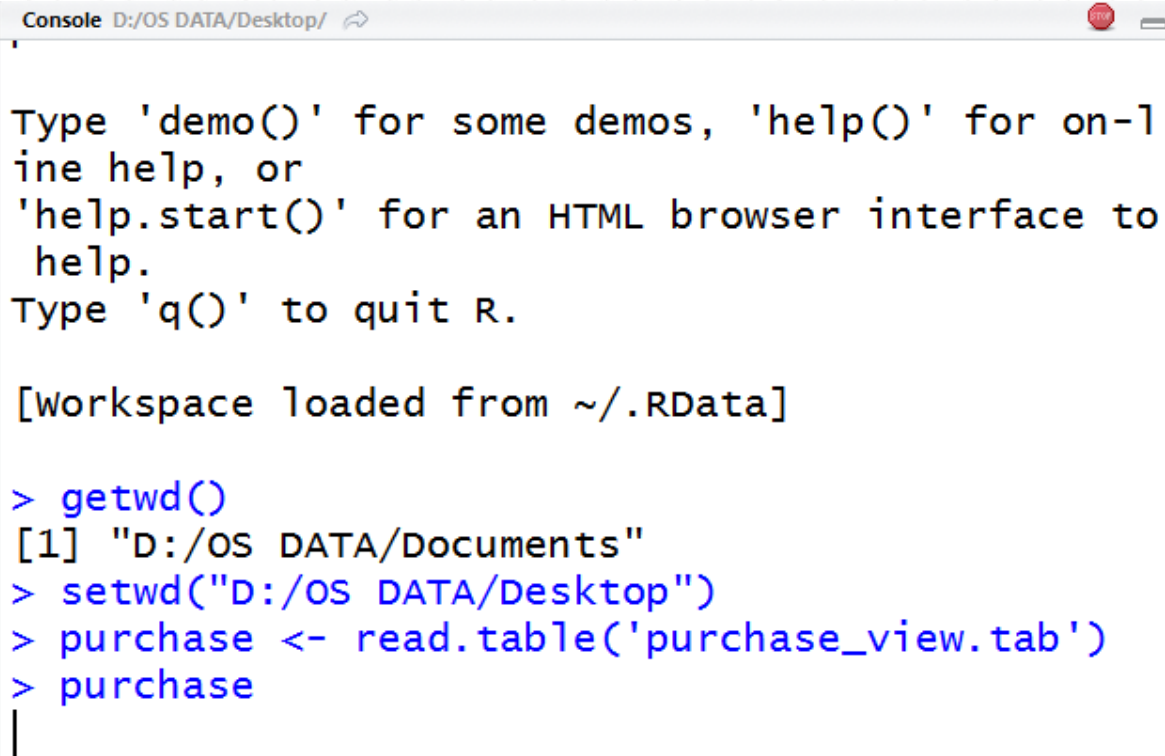
最受歡迎的語言持續為 *R*, *Python* (39%), 及 *SQL* (37%). *SAS* 大約在 20% 上下.

By Gregory Piatetsky, Aug 27, 2013.



當資料超過一定限度時？

- 雖然資料筆數只有幾百萬筆，但要花上很久時間才能讀取跟存儲資料



```
Console D:/OS DATA/Desktop/

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.

[workspace loaded from ~/.RData]

> getwd()
[1] "D:/OS DATA/Documents"
> setwd("D:/OS DATA/Desktop")
> purchase <- read.table('purchase_view.tab')
> purchase
|
```

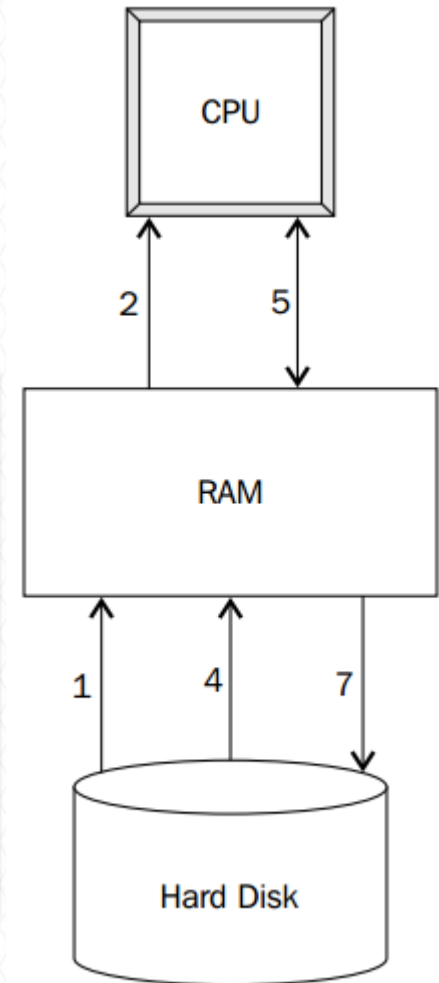
使用開源R 會碰到的問題

- CPU, RAM 與 Disk I/O 的限制
- R 是直譯式語言
- R 是單一線程 (Single-Threaded)
- R 需要將所有資料讀進記憶體

CPU, RAM 以及 Disk IO

```
data <- read.csv("mydata.csv")
totals <- colSums(data)
write.csv(totals, "totals.csv")
```

1. R 的程式碼會被讀進RAM.
2. R 直譯器將 R 程式碼變為機器碼，然後才將機器碼讀進CPU中
3. CPU 執行程式
4. 程式將資料從硬碟讀進 RAM (read.csv).
5. 資料會使用以Chunk 的形式讀到CPU 中執行
6. CPU 一次執行一個Chunk, 並與 RAM 交換Chunk 直到所有資料都被處理過 (colSums).
7. 最後, 資料會被寫回去磁碟之中 (write.csv).

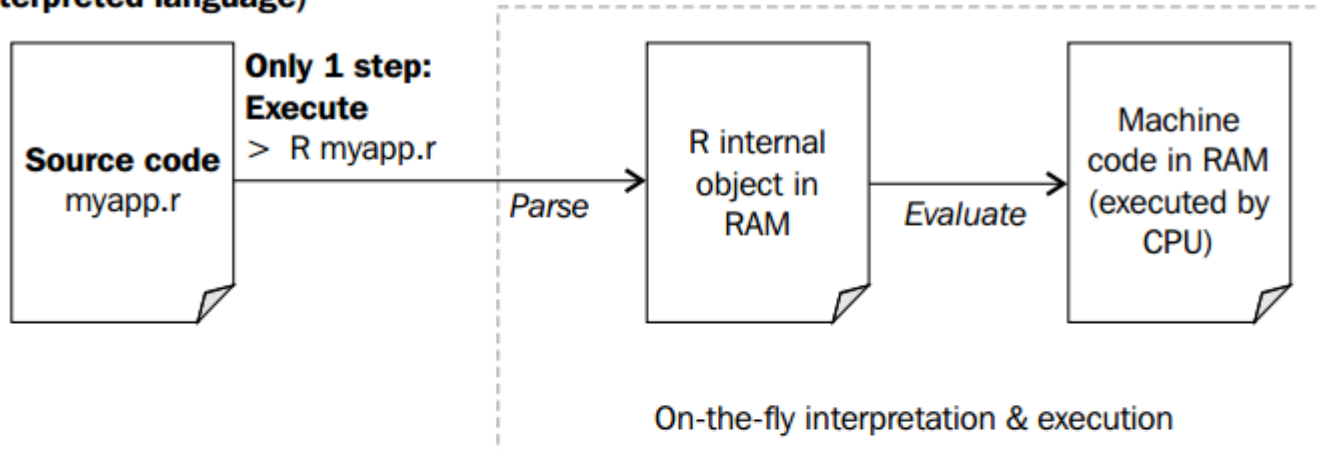


CPU, RAM 以及 Disk IO

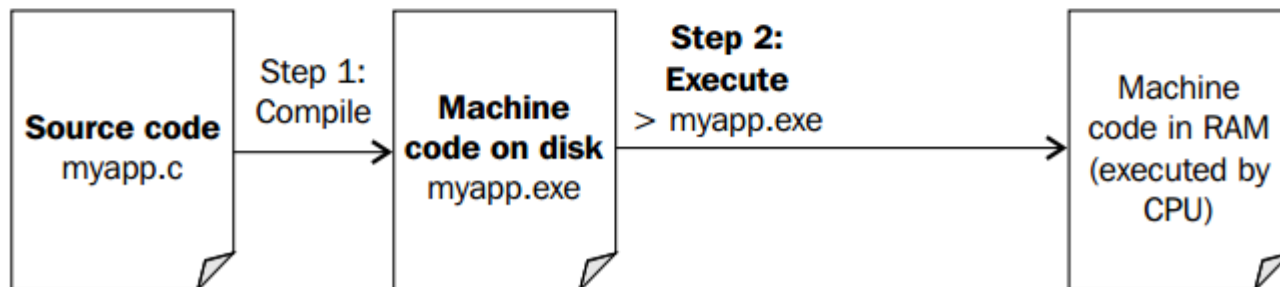
- CPU 的效能決定多快可以處理完資料，例如執行 `colSums` 時，包含了將 R 程式碼轉譯成機器碼所花的時間以及處理資料的時間
- RAM 的大小決定可以處理的資料量.假如使用 `read.csv` 所讀取到的資料量遠大於 RAM，執行將會失敗
- Disk IO 的速度(例如在做 `read.csv` 以及 `write.csv` 時) 決定多塊可以將資料讀進記憶體或寫回磁碟之中

R 是直譯式語言

R(interpreted language)



C(compiled language)



R 是單線程 (Single Threaded)

- 即使買的是64核心超強伺服器，R也只會使用一個核心做運算



解決 R 的效率問題

■ 養成良好的寫作風格

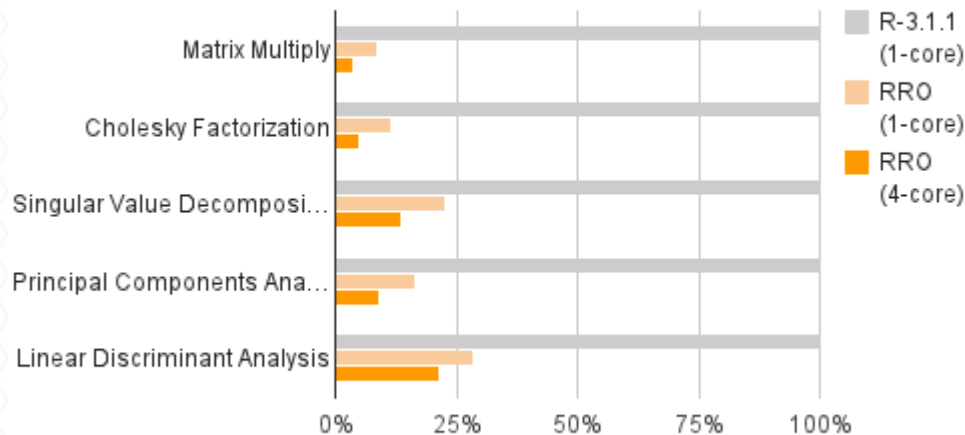
- 向量化計算Vectorization
- 使用內件函式
- 預先配置(Preallocating)好記憶體
- 使用簡單的資料結構
- 使用 hash tables
- 從 CRAN中找好的套件 (使用RCPP 或 RJAVA的套件)

■ 平行化運算

Revolution Analytics

- 2007 年成立，為統計軟體 R 語言提供商業服務方案
- 除了維護社群版本的 R 軟體，貢獻部分 R 語言程式碼，還提供企業版本的 R 軟體，販售給企業用戶

Performance comparison



REVOLUTION
ANALYTICS

微軟在2015收購了Revolution R

[Store](#) ▾[Products](#) ▾[Support](#)[Official Microsoft Blog](#)[The Fire Hose](#)[Microsoft On the Issues](#)[Next](#)[Transform](#)

Microsoft to acquire Revolution Analytics to help customers find big data value with advanced statistical analysis

Posted January 23, 2015 By **Joseph Sirosh** - *Corporate Vice President, Data Group, Microsoft*



Update: April 6, 2015: Microsoft has closed the acquisition of Revolution Analytics. For more details, please read the blog post by Joseph Sirosh [here](#).

I'm very pleased to announce that Microsoft has reached an agreement to acquire

Featured Posts

Microsoft to acquire LinkedIn



Microsoft and LinkedIn Corporation on Monday announced they have entered ... [Read more »](#)

Physical and virtual worlds intersect with Windows Holographic, now opening to partners for a new era of mixed reality



On Wednesday at Computex in Taipei, Terry Myerson, executive vice president, ... [Read more »](#)

Windows veteran Dona Sarkar is new head of Windows Insider Program


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.

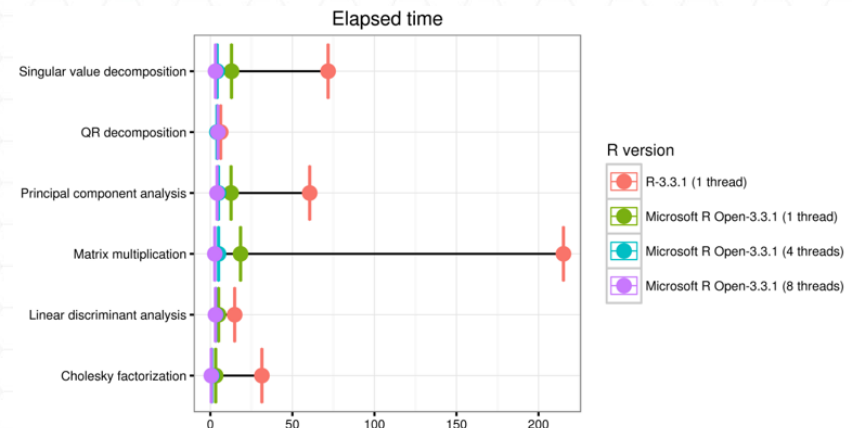
The current version, Microsoft R Open 3.2.5, is based on (and 100% compatible with) R-3.2.5, the most widely used statistics software in the world, and is therefore fully compatibility with all packages, scripts and applications that work with that version of R. It includes additional capabilities for **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](#)



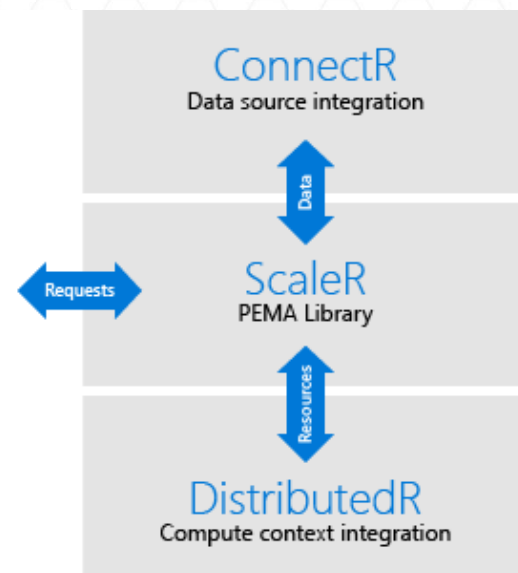
<https://mran.revolutionanalytics.com/documents/rro/multithread/>

Microsoft R Server

- 可以用ConnectR連結不同來源的資料源：支援Hadoop、Linux和Teradata平臺
- 在不用修改程式碼的情況下讓R程式碼能平行運算
- 可選擇在單機或分散式群集（ Distributed Cluster ）建立和執行R模型，程式碼不因執行平台不同而需要修改

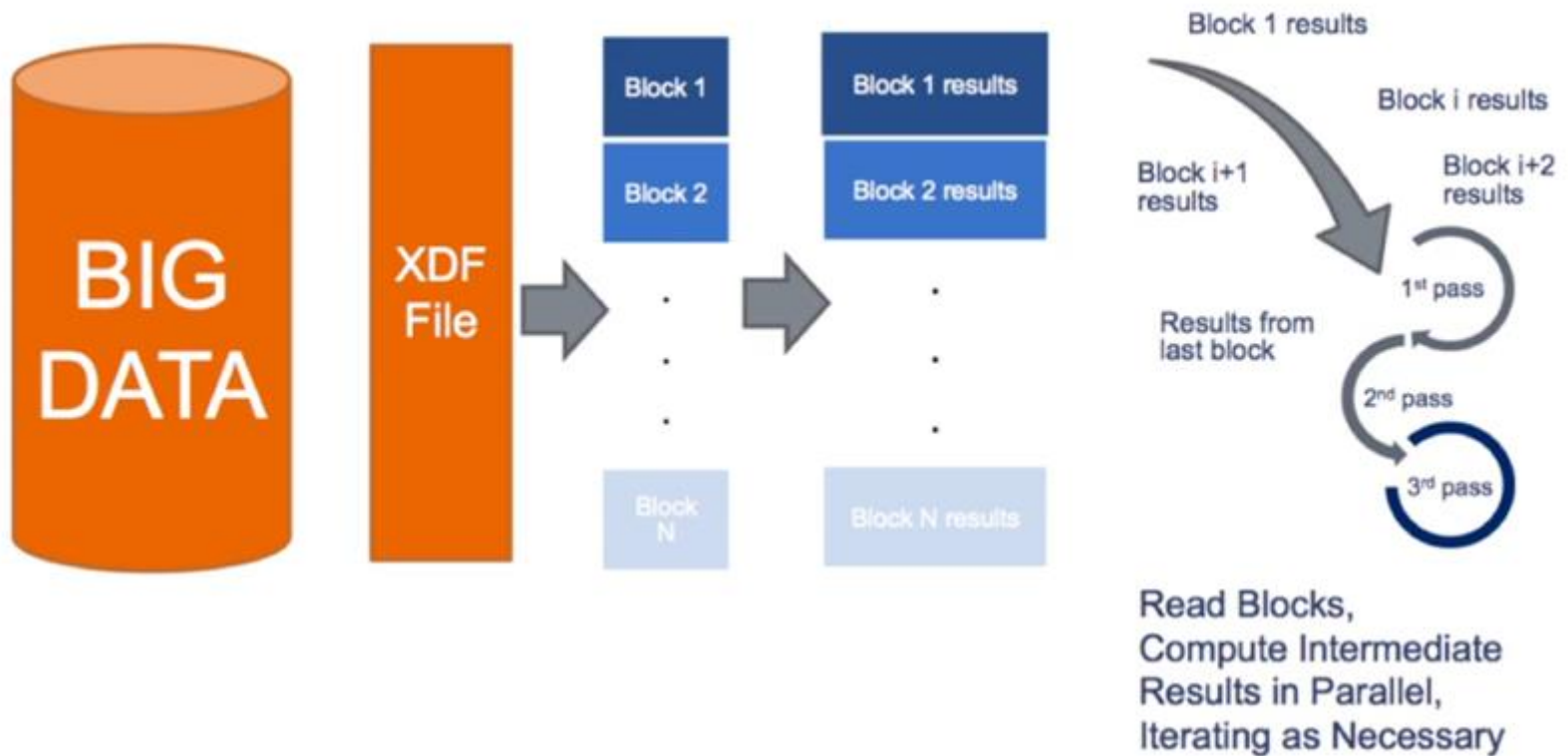
R+CRAN

Enhanced R Interpreter

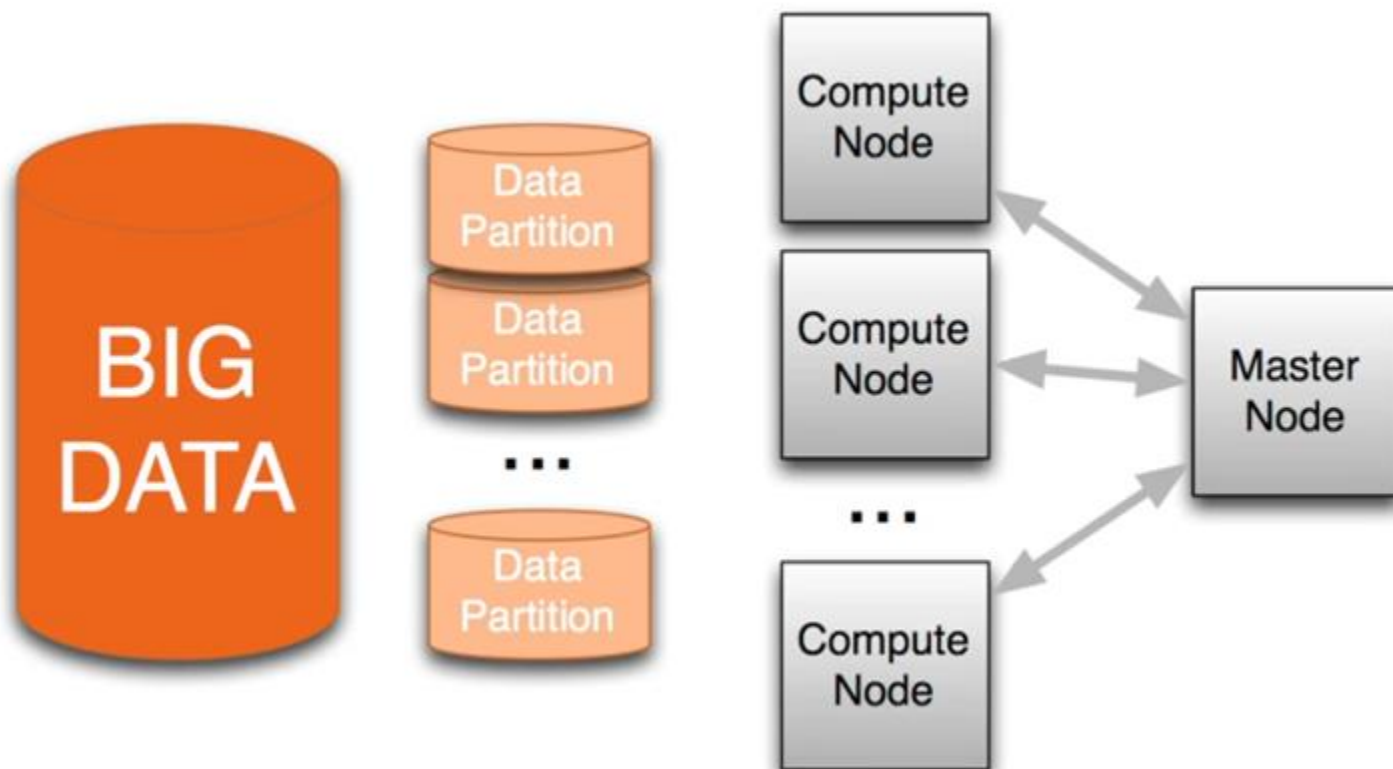


XDF 檔案運算模式

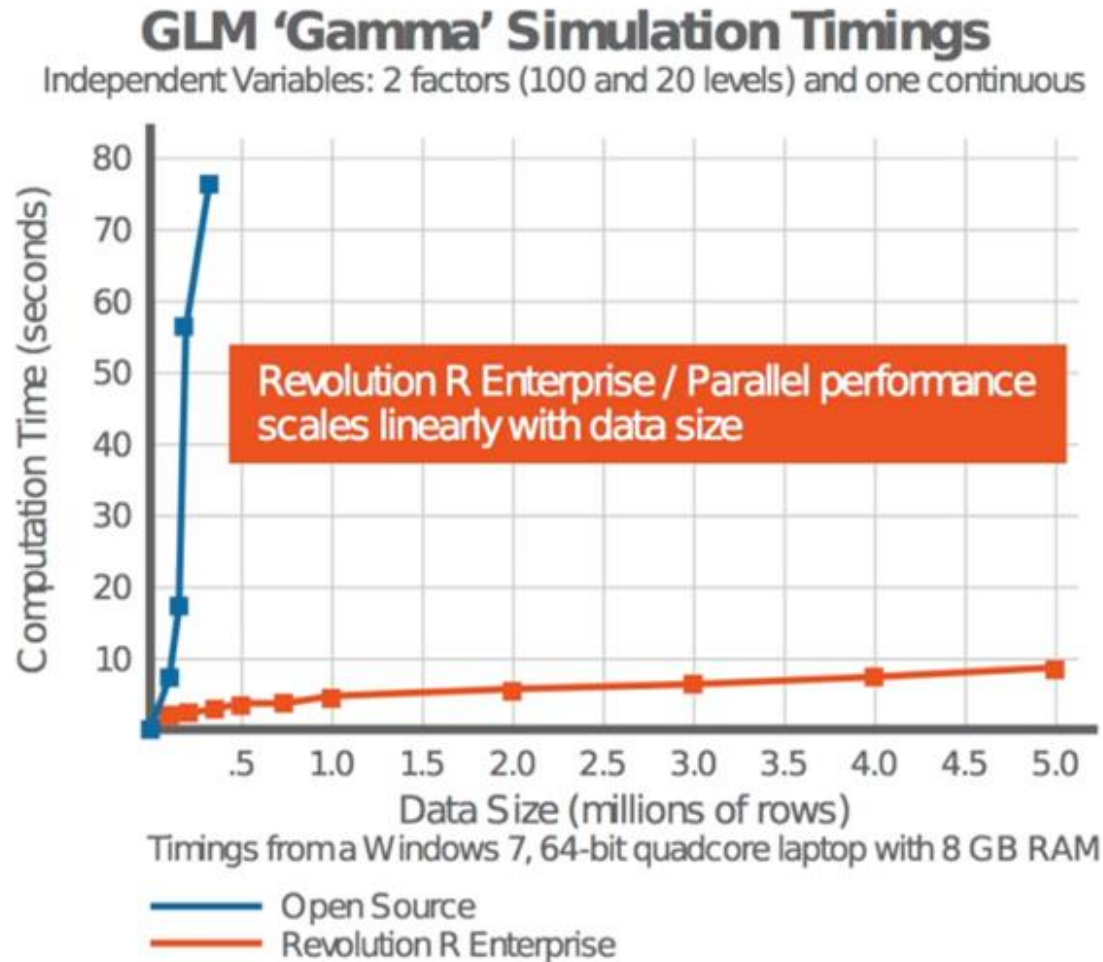
■ Parallel External Memory Algorithm



在分散式環境下的處理方式



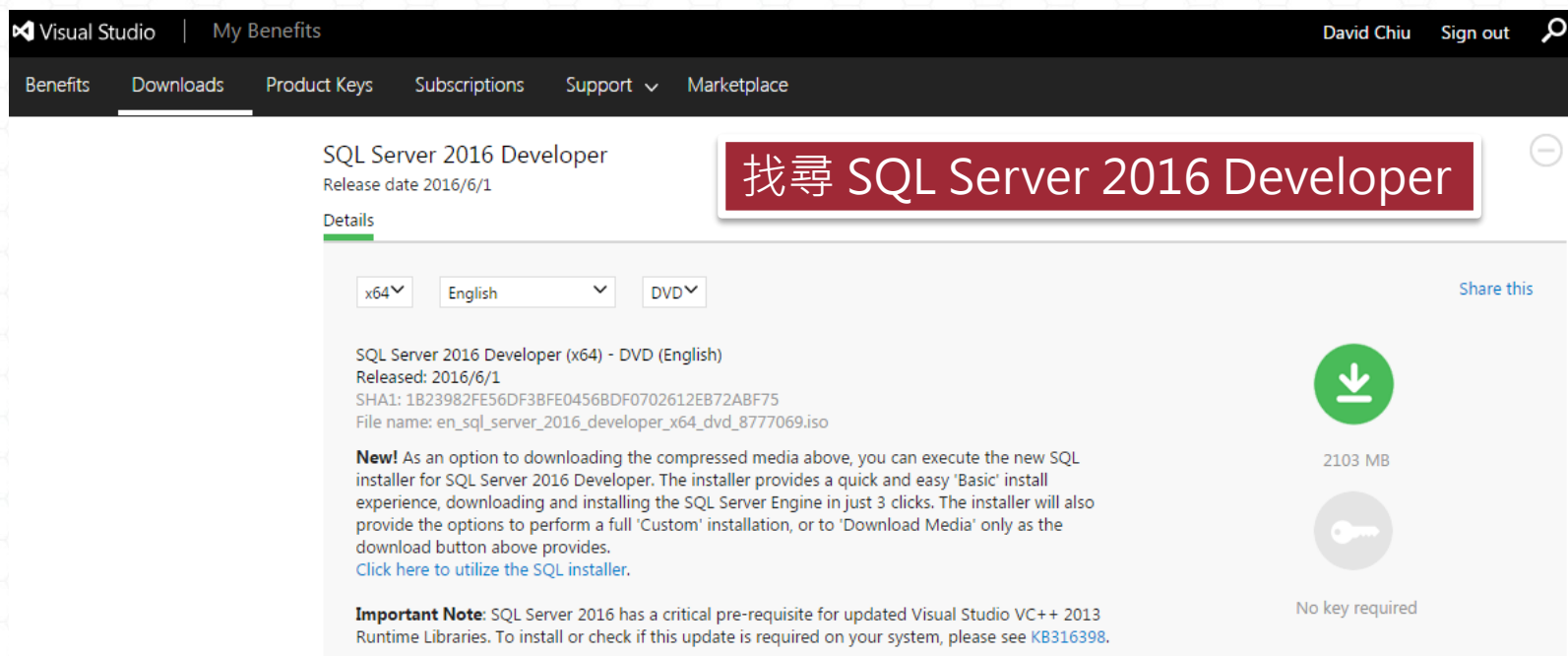
開源R 與 Microsoft R Enterprise效能評比



R語言環境設定

下載 SQL Server 2016 Developer

■ <https://my.visualstudio.com/downloads>



The screenshot shows the Visual Studio web interface. At the top, there's a navigation bar with 'Visual Studio' and 'My Benefits'. Below it, a menu bar includes 'Benefits', 'Downloads' (which is active), 'Product Keys', 'Subscriptions', 'Support', and 'Marketplace'. The main content area is titled 'SQL Server 2016 Developer' with a release date of '2016/6/1'. A red banner with white text says '找尋 SQL Server 2016 Developer'. Under the 'Details' tab, there are dropdown menus for 'x64', 'English', and 'DVD'. The download information includes: 'SQL Server 2016 Developer (x64) - DVD (English)', 'Released: 2016/6/1', 'SHA1: 1B23982FE56DF3BFE0456BDF0702612EB72ABF75', and 'File name: en_sql_server_2016_developer_x64_dvd_8777069.iso'. A 'New!' section describes a new SQL installer. On the right, there's a green download button with a white arrow, labeled '2103 MB', and a key icon labeled 'No key required'. A 'Share this' link is also present.

Visual Studio | My Benefits

David Chiu Sign out

Benefits Downloads Product Keys Subscriptions Support Marketplace

SQL Server 2016 Developer
Release date 2016/6/1

Details

x64 English DVD

SQL Server 2016 Developer (x64) - DVD (English)
Released: 2016/6/1
SHA1: 1B23982FE56DF3BFE0456BDF0702612EB72ABF75
File name: en_sql_server_2016_developer_x64_dvd_8777069.iso

New! As an option to downloading the compressed media above, you can execute the new SQL installer for SQL Server 2016 Developer. The installer provides a quick and easy 'Basic' install experience, downloading and installing the SQL Server Engine in just 3 clicks. The installer will also provide the options to perform a full 'Custom' installation, or to 'Download Media' only as the download button above provides.
[Click here to utilize the SQL installer.](#)

Important Note: SQL Server 2016 has a critical pre-requisite for updated Visual Studio VC++ 2013 Runtime Libraries. To install or check if this update is required on your system, please see [KB316398](#).

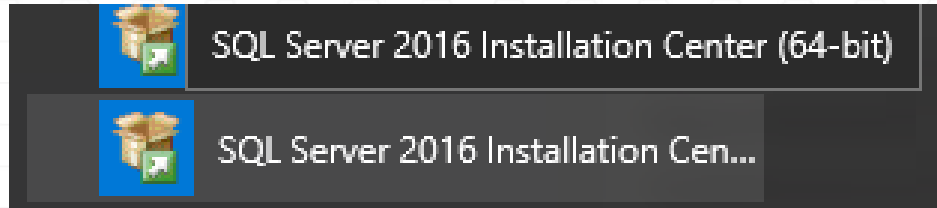
Share this

2103 MB

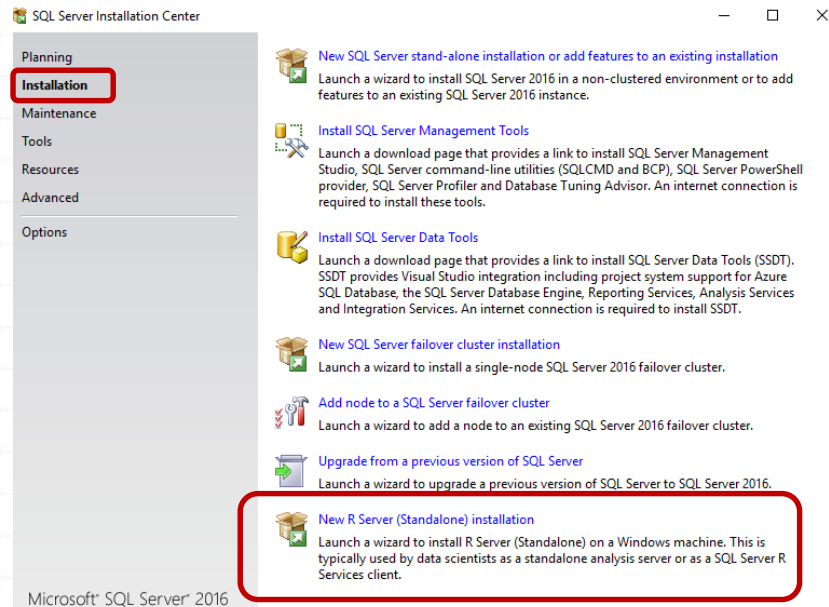
No key required

SQL Server 2016 Installation Center

■ 點選 SQL Server 2016 Installation Center (64-bit)

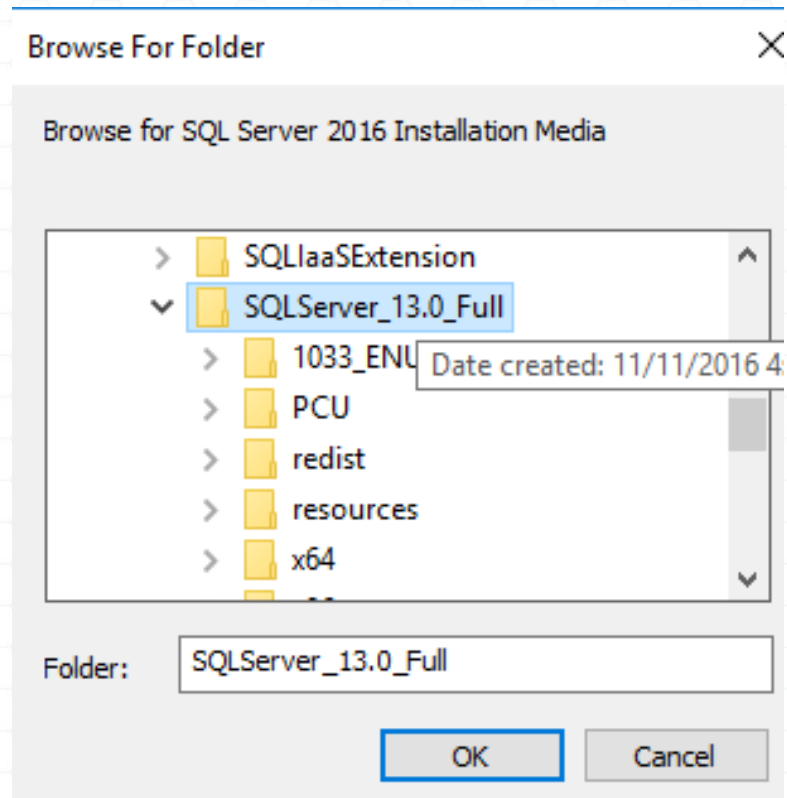


■ 安裝 New R Server(Standalone) Installation



選擇安裝目錄

■ 選擇安裝於SQLServer_13.0_Full



選擇增加新功能

■ 增加 R Server (Standalone)

SQL Server 2016 Setup

Feature Selection

Select the Enterprise Edition: Core-based Licensing features to install.

Global Rules	Features:	Feature description:
Product Updates	Instance Features	Includes Microsoft R Server for Windows that provides parallelized big data analysis, transformation, modeling and model deployment capabilities.
Install Setup Files	<input checked="" type="checkbox"/> Database Engine Services	Prerequisites for selected features:
Install Rules	<input checked="" type="checkbox"/> SQL Server Replication	Already installed:
Installation Type	<input checked="" type="checkbox"/> R Services (In-Database)	... Microsoft Visual Studio 2010 Redistributables
Feature Selection	<input checked="" type="checkbox"/> Full-Text and Semantic Extractions for Sea	... Microsoft MPI v7
Feature Rules	<input checked="" type="checkbox"/> Data Quality Services	Disk Space Requirements
Consent to install Microsoft R ...	<input type="checkbox"/> PolyBase Query Service for External Data	Drive C: 482 MB required, 102527 MB available
Feature Configuration Rules	<input checked="" type="checkbox"/> Analysis Services	
Installation Progress	<input checked="" type="checkbox"/> Reporting Services - Native	
Complete	Shared Features	
	<input checked="" type="checkbox"/> R Server (Standalone)	
	<input type="checkbox"/> Reporting Services - SharePoint	
	<input type="checkbox"/> Reporting Services Add-in for SharePoint Proc	

Select All Unselect All

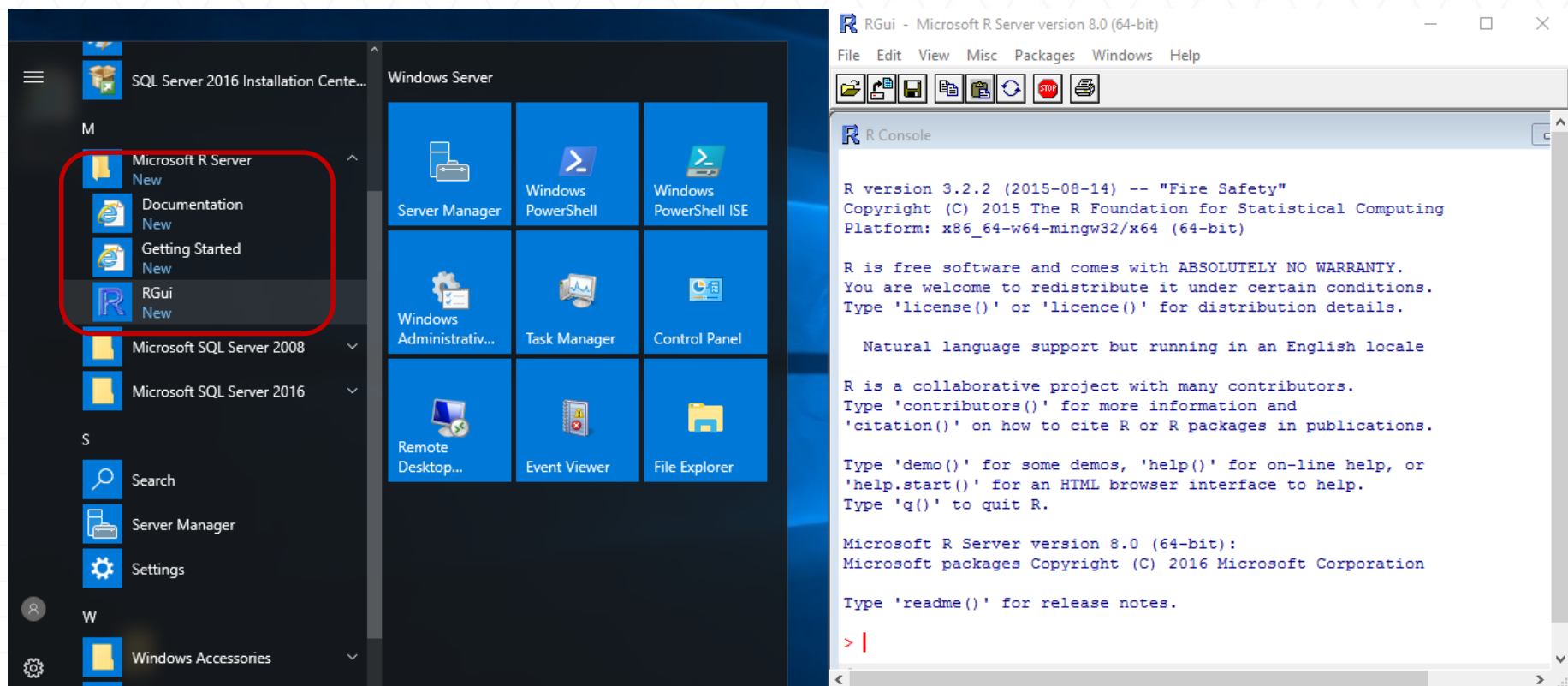
Instance root directory: C:\Program Files\Microsoft SQL Server\

Shared feature directory: C:\Program Files\Microsoft SQL Server\

Shared feature directory (x86): C:\Program Files (x86)\Microsoft SQL Server\

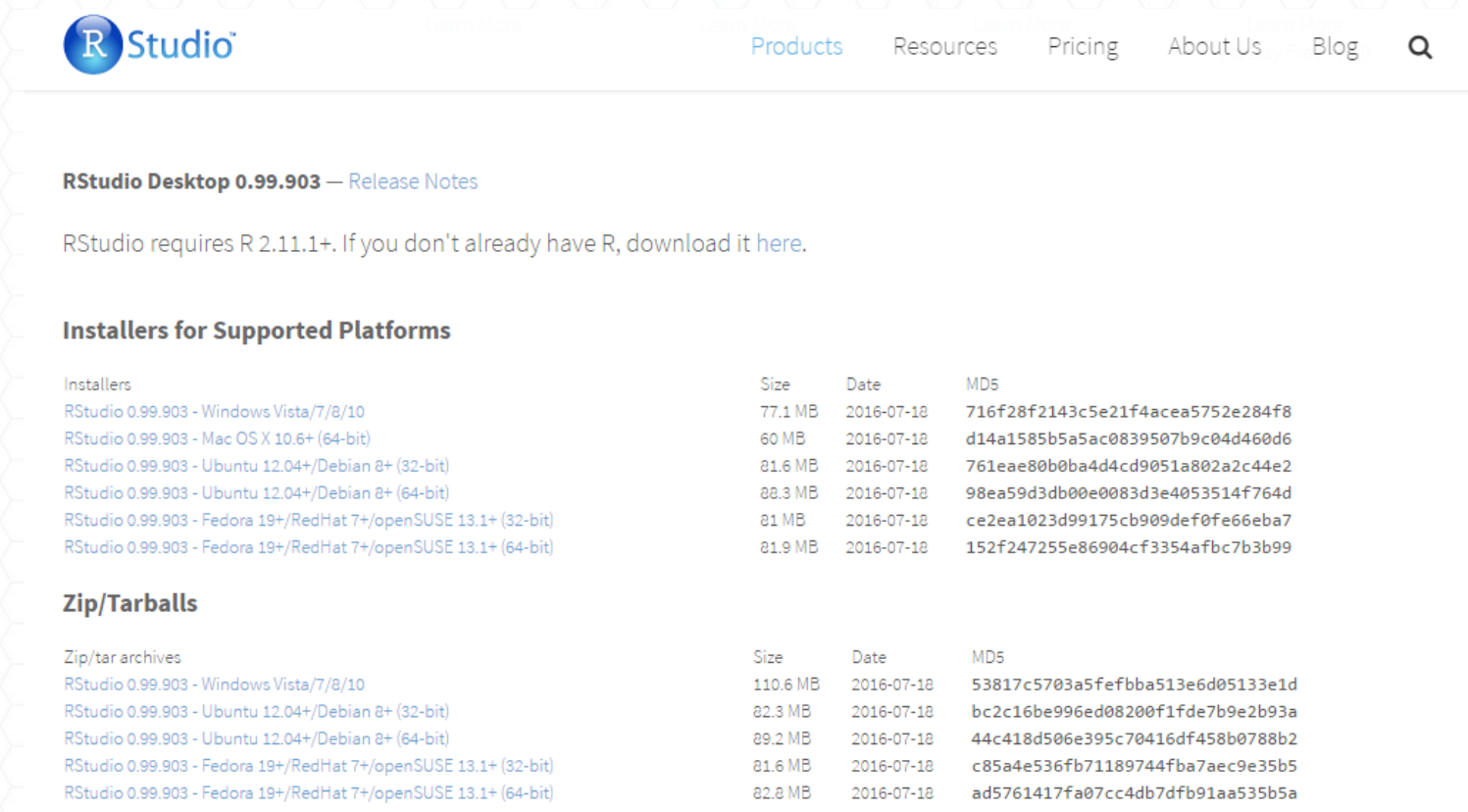
< Back Next > Cancel

開啟Microsoft R Server Version 8.0



下載RStudio

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



The screenshot shows the RStudio website's download page for RStudio Desktop 0.99.903. The page has a navigation bar with links for Products, Resources, Pricing, About Us, and Blog. The main content area is titled "RStudio Desktop 0.99.903 — Release Notes". It states that RStudio requires R 2.11.1+ and provides a link to download R. The page 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.

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

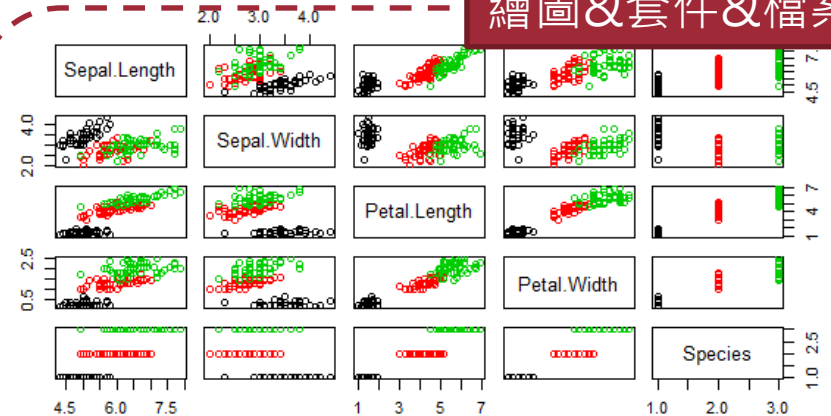
繪圖&套件&檔案

```
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
```

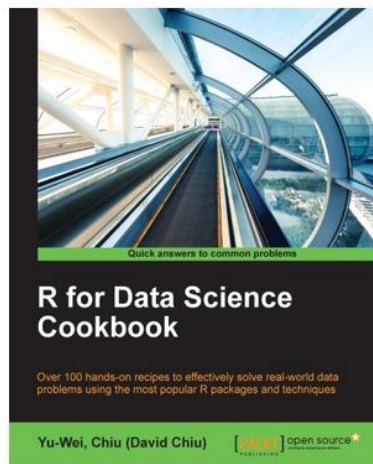
計算一本書的價錢

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 ²)
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)

產生陣列

■ 產生陣列

```
matrix(1:9, byrow=TRUE, nrow=3)
```

```
matrix(1:9, nrow=3)
```


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

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



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),]
```

實際範例

- 找出股票資料(stock_data)中日期大於2014年三月到八月間台積電最高收盤價(close)

- <http://finance.yahoo.com/quote/2330.TW?ltr=1>



The background features a light blue hexagonal grid pattern. Overlaid on this is a large, faint, light blue circular graphic composed of concentric rings and radial lines, resembling a stylized spiral or a target. A solid dark blue horizontal bar runs across the top of the image, and a similar but slightly textured dark blue bar runs across the bottom.

THANK YOU