數據分析師假日精修班 Lab1

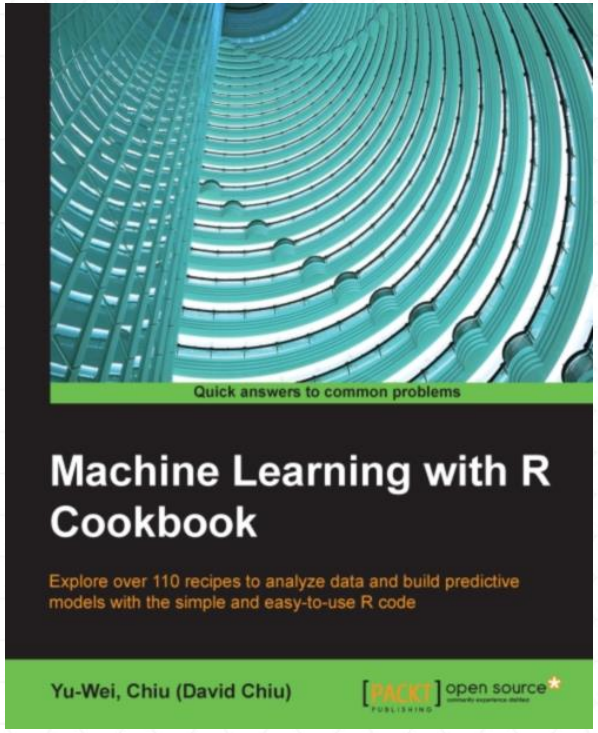
David Chiu

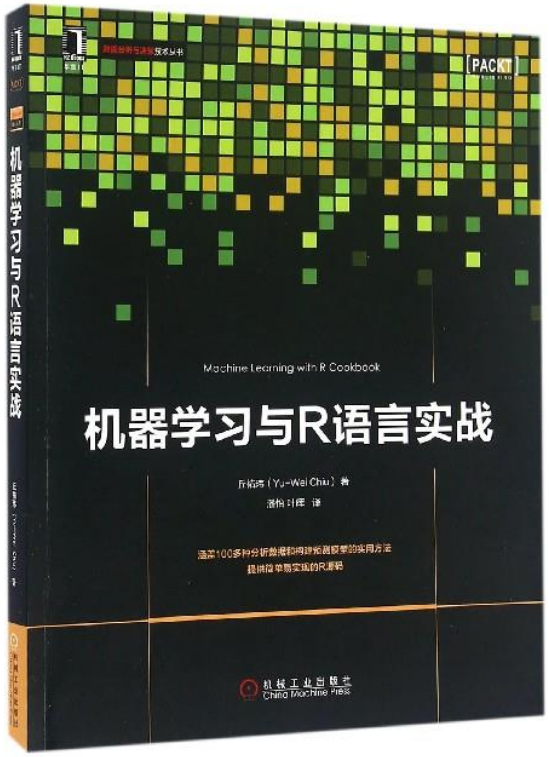
2016/09/01

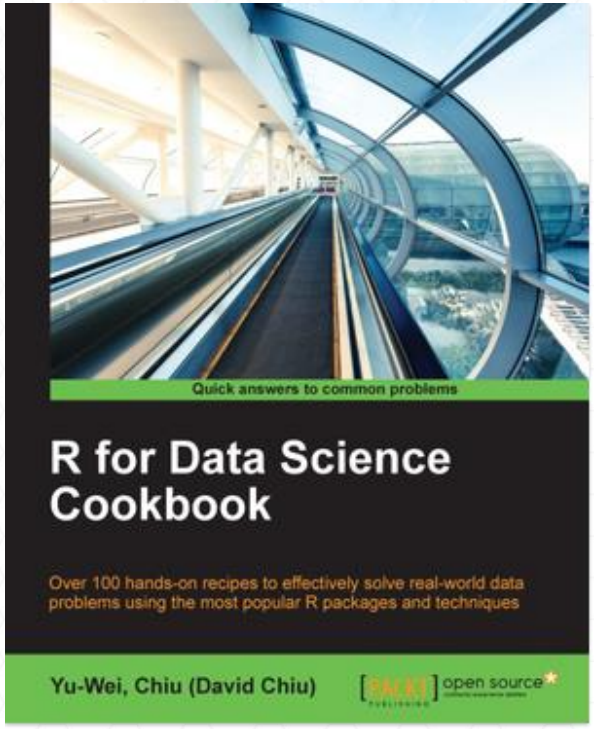
關於我

- 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

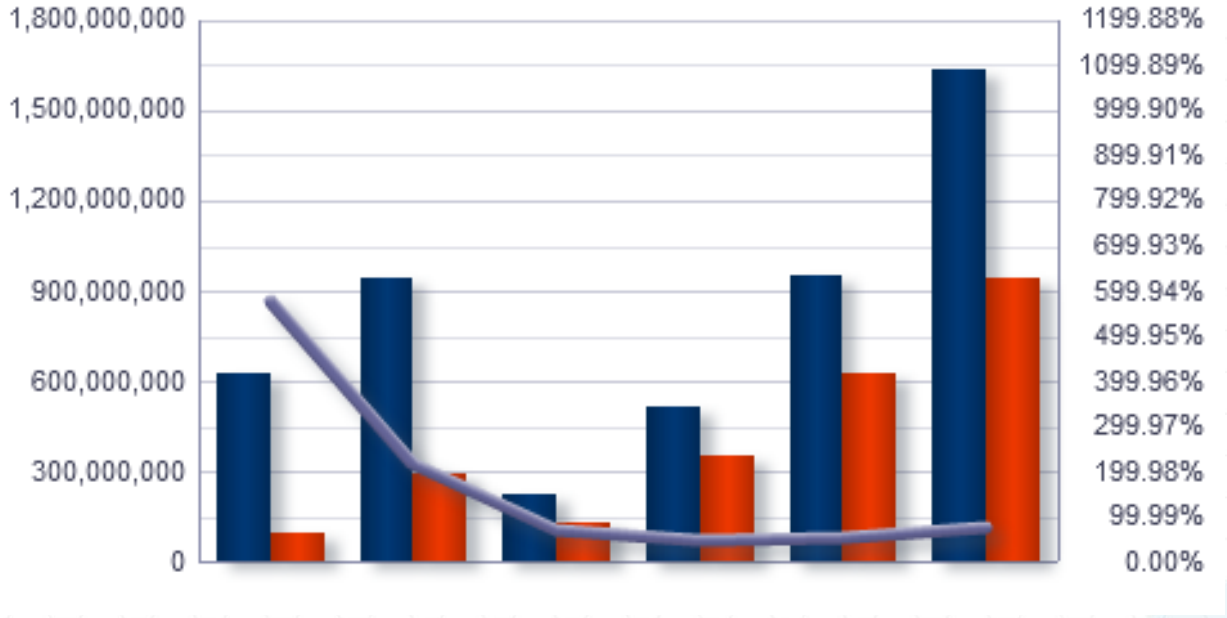
環境資訊頁面

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

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戶最多，並畫出資料分佈圖的話，該怎麼做?



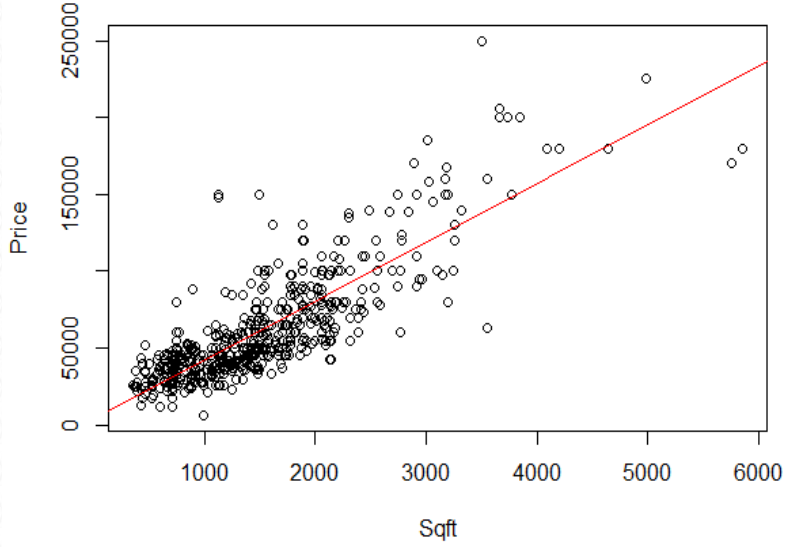
不同的做法

- 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

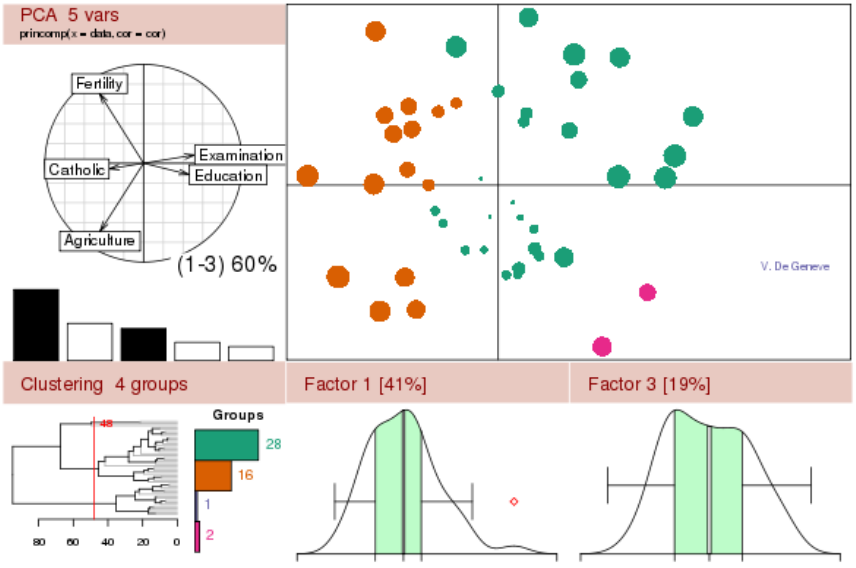
- 統計房屋坪數與房價的關係




WWW
iDeA
45% is SALE!
25% time
TEAM
internet
MARKETING
CLOUD

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

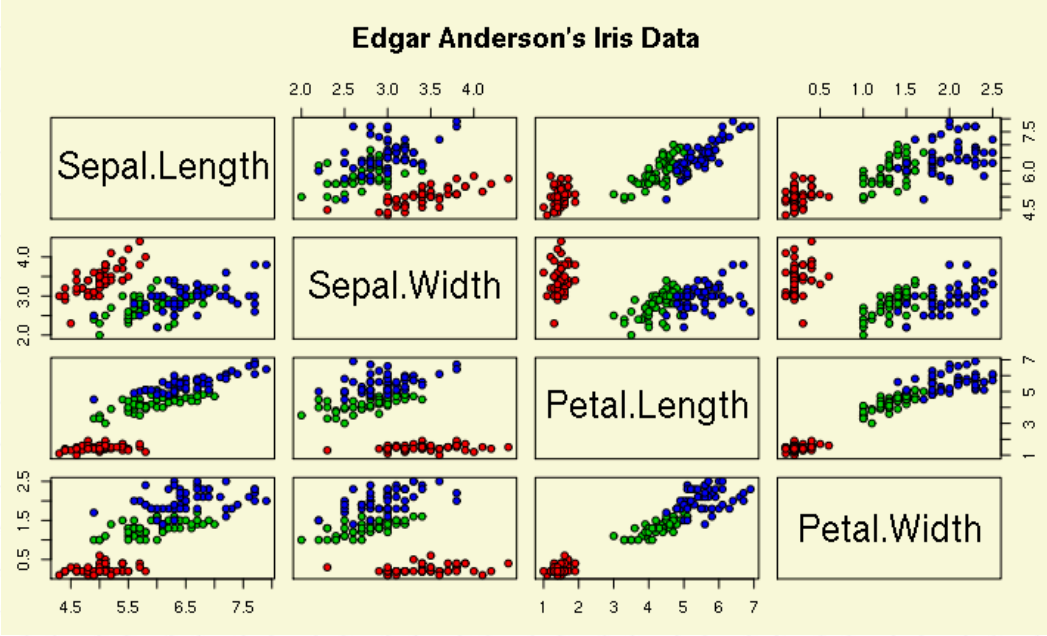
- S 語言的方言 (分支)
- 受到函數式編程語言Scheme 的啟發，因而想將該功能加入到 S 語言當中
- 1992年Ross Ihaka 與 Robert Gentleman 為了教授統計，因此開發出了 R語言
- 除了R 以外，還有S-Plus，但兩個分支走向不同，一個走向社群，一個走向商業

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

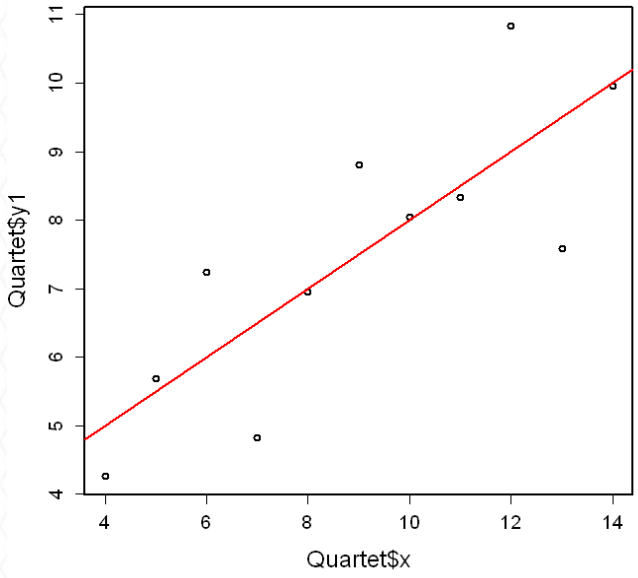
- 統計分析
- 迴歸分析
- 資料分群
- 資料分類
- 推薦系統
- 文字探勘



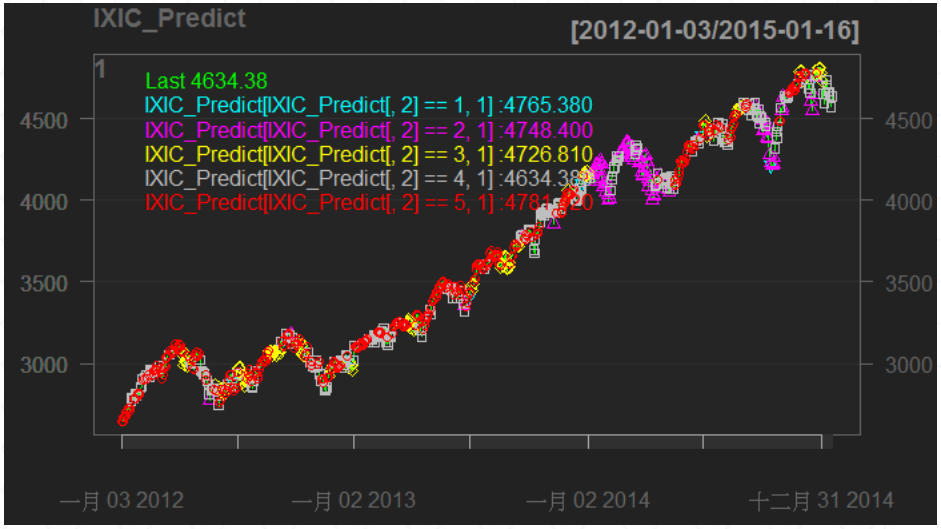
影像辨識

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

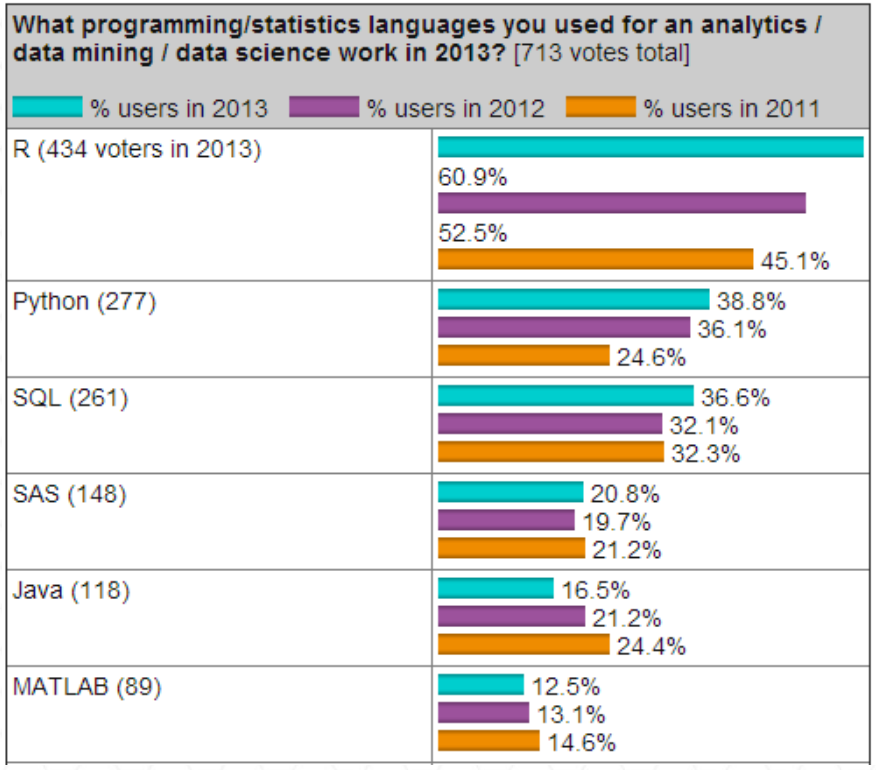


預測股票

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.

Revolution R

- 社群使用版本
  - http://www.revolutionanalytics.com/downloads/

	Base R 2.14.2 64	Revolution R (1-core)	Revolution R (4-core)	Speedup (4 core)
Matrix Calculation	17.4 sec	2.9 sec	2.0 sec	7.9x
Matrix Functions	10.3 sec	2.0 sec	1.2 sec	7.8x
Program Control	2.7 sec	2.7 sec	2.7 sec	Not Appreciable

  - http://www.revolutionanalytics.com/why-revolution-r/benchmarks.php

微軟在2015收購了Revolution R

Microsoft 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*

f 563 in 1918

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

- https://mran.microsoft.com/open/

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

R語言環境設定

下載R

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

R-3.3.1 for Windows (32/64 bit)

Download R 3.3.1 for Windows (70 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

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?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

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

下載RStudio

- https://www.rstudio.com/products/rstudio/download3/

RStudio

Products Resources Pricing About Us Blog

RStudio Desktop 0.99.903 — Release Notes

RStudio requires R 2.11.1+. If you don't already have R, download it here.

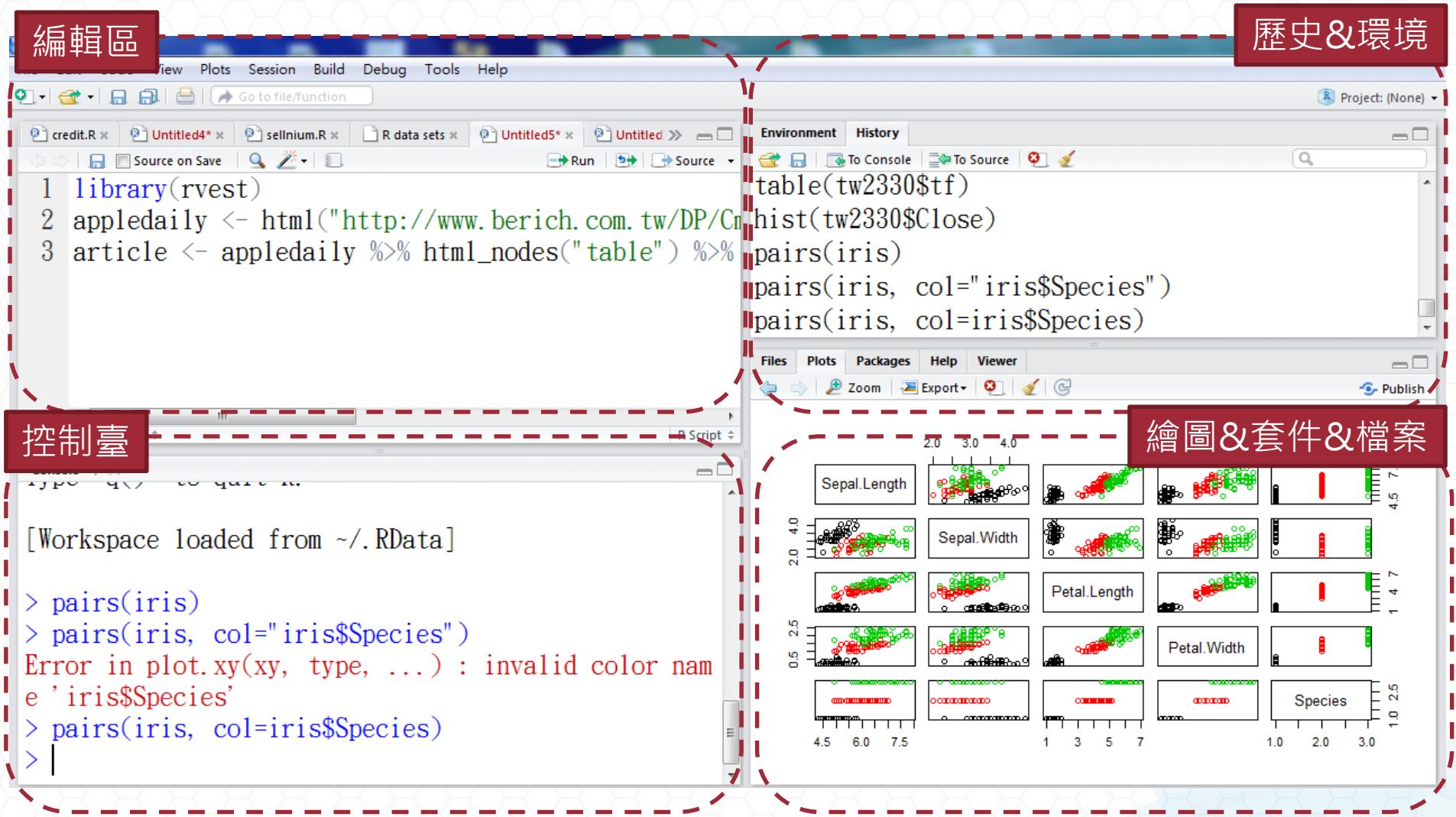
Installers for Supported Platforms

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
Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
credit.R
Untitled4*
selnium.R
R data sets
Untitled5*
Untitled
Source on Save
Run
Source
library(rvest)
appledaily <- html("http://www.berich.com.tw/DP/Cn
article <- appledaily %>% html_nodes("table") %>%
Environment History
To Console
To Source
table(tw2330$tf)
hist(tw2330$Close)
pairs(iris)
pairs(iris, col="iris$Species")
pairs(iris, col=iris$Species)
Files Plots Packages Help Viewer
Zoom
Export
Publish
Script
[Workspace loaded from ~/.RData]
> pairs(iris)
> pairs(iris, col="iris$Species")
Error in plot.xy(xy, type, ...) : invalid color nam
e 'iris$Species'
> pairs(iris, col=iris$Species)
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
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

# 變數相加
b <- 5
c <- a + b
c
```

可以使用 = 或 <- 指定變數

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)
```

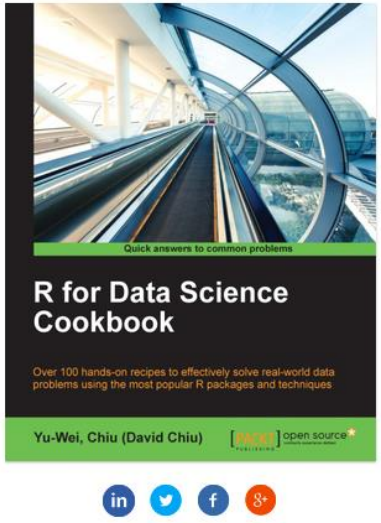
不同型態資料做運算

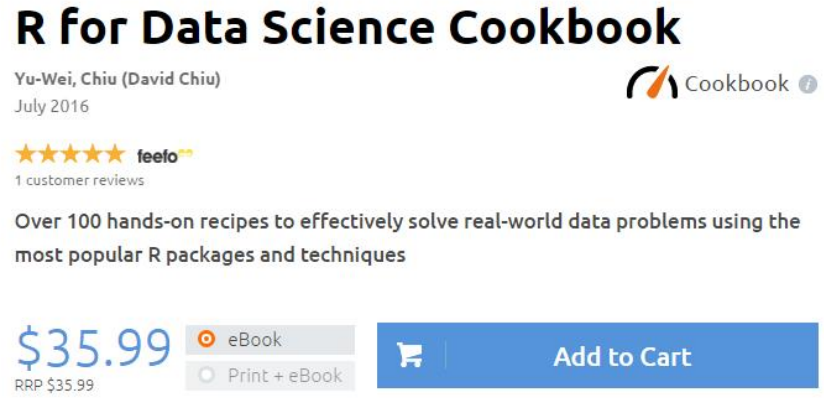
```
card_length <- 3
card_width <- "5 inches"
card_length * card_width
Error in card_length * card_width :
  non-numeric argument to binary operator

#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
```





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 <- c(1,2,3,7)
y <- 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
```

seq()

- 使用? 或help 去觀看seq 的用法

```
?seq
help(seq)
```

- 使用seq 產生不同類型向量

```
seq(1,20,2)
seq(1,3.5, by =0.5)
seq(1,10,length=2)
```

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 = 體重(公斤) / 身高^2(公尺^2)

	身體質量指數(BMI) (kg/m2)
體重過輕	BMI ＜ 18.5
正常範圍	18.5≦BMI＜24
異常範圍	過重：24≦BMI ＜27 輕度肥胖：27≦BMI ＜30 中度肥胖：30≦BMI ＜35 重度肥胖：BMI≧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
```

矩陣乘積

- m1 X m2

m1 %*% m2

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1*5+2*7 & 1*6+2*8 \\ 3*5+4*7 & 3*6+4*8 \end{bmatrix}$$

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

有順序的階層

- 產生可比較的類別資訊

```
temperature <- c("Low", "High", "High", "Medium", "Low", "Medium")
temperature_category <- factor(temperature, order = TRUE, levels = c("Low", "Medium", "High"))
temperature_category
temperature_category[3] > temperature_category[1]
temperature_category[4] > temperature_category[3]
```

- 檢查類別

```
levels(temperature_category)
```

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

清單(Lists)

清單(Lists)

- 可以混雜不同的資料型態

```
item <- list(thing="hat", size="8.25")
item
```

- 使用$取得內容物

```
test <- list(name="Toby", score = c(87,57,72))
test$score
test$score[2]
```

清單(Lists) (續)

- 沒有名字的清單

```
li <- list(c(3,5,12), c(2,4,5,8,10))
li
```

- 使用lapply將函式套用到list 上

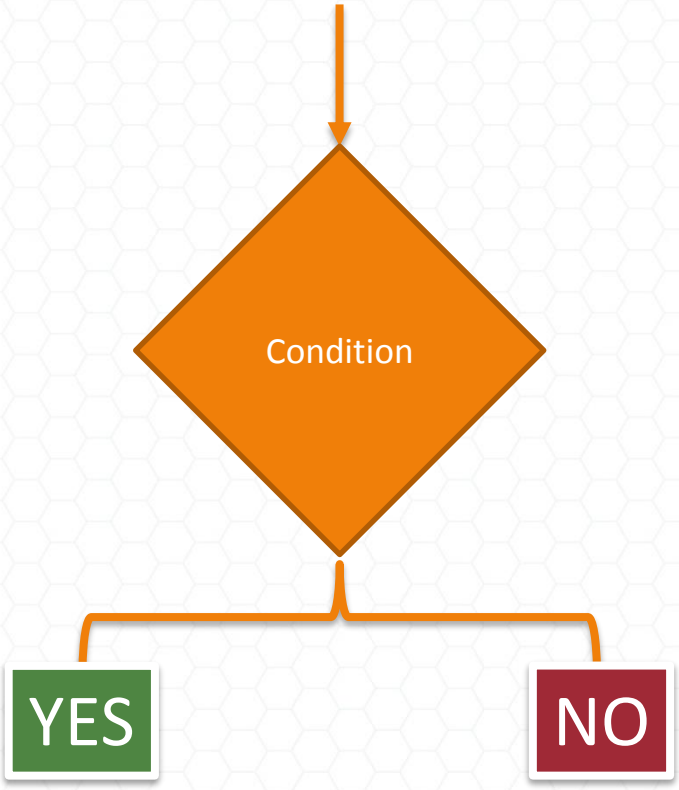
```
lapply(li, sum)
```

流程控制(Flow Control)

IF…ELSE…

- If 及 else 的判斷

```
x = 5;
if(x > 3){
    print("x > 3");
}else{
    print("x <= 3");
}
```



IF…ELSE IF…ELSE

- 使用else if

```
x = 5;
if(x > 3){
    print("x > 3");
} else if(x ==3){
    print("x == 3");
}else{
    print("x < 3");
}
```

FOR 迴圈

- For 迴圈

```
for(i in 1:10){
    print (i);
}
```

- 1~100的總和

```
s = 0
for(i in 1:100){
    s = s + i;
}
s
```

三種FOR 迴圈

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

for(i in 1:length(x)) {
 print(x[i])
}

for(i in seq_along(x)) {
 print(x[i])
}

for(letter in x) {
 print(letter)
}
```

使用while 迴圈

- 當不滿足while中定義的條件時，才會跳出迴圈

```
s = 0;
cnt = 0;
while(cnt <= 100){
    s = s + cnt;
    cnt = cnt + 1;
}
s
```

範例: 產生多筆頁面連結

```
url <- 'http://www.appledaily.com.tw/realtimenews/section/new/‘

for (i in seq(1,10)){
  print(paste0(url, i))
}
```

1 **2** 3 4 5 6 7 8 9 10 下10頁

函式 (Function)

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) {
}
```

Lazy Function

```
f = function(a, b) {
   a * 2
}
f(3)

f = function(a, b) {
     print(a+ b)
}
f(3)
```

範例: 撰寫函式計算文章詞頻

- 計算文章詞頻

```
wordcount <- function(article){
  article.split <- strsplit(article, ' ')
  table(article.split)
}

wordcount(a)
```

THANK YOU