推薦序

我在大專院校教資料分析及資料探勘多年，學生範圍包含資訊系及非資訊系的大學部及研究所。由於大環境的改變，傳統理論型的教科書已經無法適合大多數學生的學習。根據我的觀察，目前大多數的學生都希望學習不要過於理論，應該強調實用性。而且教材的內容要能連結大學、研究所及就業的需求，不要大學學一套，研究所學一套，就業後又必須學另外一套。同時，教材的廣度及深度必須足夠。非資訊系的學生能有豐富的工具，可以對特定的問題進行不同角度深入的探討；資訊系的學生可以利用撰寫程式或腳本來擴展應用。最重要的是，這些學習與研究必須可以輕鬆且有效的落實於未來的工作環境中！

何宗武教授這本書的內容，完全符合了現在資料分析及資料探勘學習的要求！這本書的核心是介紹 R 上面知名的資料探勘視窗套件 rattle。在說明的過程中，何教授利用範例，一步一步的帶領著大家了解 rattle 中各種資料分析及探勘工具的使用，並解釋結果的意思。由於理論性的論述較少，所以閱讀起來十分輕鬆。但遇到重要的部分，何教授會提出練習題目，希望大家思考看看。除了 rattle 之外，這本書也介紹 R 的基礎程式撰寫及其他重要套件，例如 R Commander, GGobi, 及 PMML。這些介紹提供了 rattle 擴展的能力。

何教授這本書的內容非常豐富，所以大學部與研究所的學生們都可以在這本書中找到資料分析及資料探勘學習的方向。當然，因為篇幅的限制，本書並不可能提供所有學習的主題，但大部分的主題幾乎都可以從此書起步，再慢慢往下挖掘找到。事實上，往下挖掘找到新知的過程更是一種重要的學習！對於資訊系或非資訊系學生在廣度及深度上的學習也是如此。

由於 R 本身是免費開放源碼的，本書除了豐富的內容，R 的學習成果是可以有效延續的。學生們可以在大學課堂中、研究所的研究計畫中、或者是未來工作的環境裡，直接使用本書學習的成果。

這本書是學習資料分析及資料探勘一個很好的教材，它完全符合了現在資料分析及資料探勘學習的要求！本人十分樂見此書的出版。

陳同孝

台中科技大學資訊工程系教授兼系主任

推薦序

R 語言是 1993 年開發的語言，主要用於統計分析、繪圖、資料探勘。由於 R 語言具有自由、開放源碼、應用廣泛以及大幅降低使用軟體的成本與提升產業競爭力等優點，近年來深受國外許多知名公司的使用，例如 Google, Facebook, New York Times, Bank of American 以及 Agoda 皆採用 R 語言進行資料分析。此外，Oracle 與 IBM 已將 R 語言融入其主要產品中，成為全世界兩大軟體公司的主要分析工具。

國內對於 R 語言有深入研究並兼具教學熱誠的財經學者首推何宗武教授。何宗武教授為了推展 R 語言以及讓學習者能很快的上手，不遺餘力地將其研究的心得寫成一系列的好書，讓莘莘學子能很快迎頭趕上世界的潮流，快速跨越 R 陡峭的學習曲線，功德實在無量，其精神讓人佩服不已。

高雄第一科技大學財務金融學院為了借重何宗武教授在 R 語言以及財金大數據的專長，從 2015 年 8 月起，禮聘何宗武教授擔任諮詢委員，提供財金大數據發展方向的建言，其遠見與觀點對於第一科大財務金融學院在財金大數據的發展具有極大的貢獻。面對金融科技的浪潮，天下雜誌在 2016 年 1 月出刊第 590 期的 FinTech 大衝擊的專刊中，報導高雄第一科大財務金融學院在數位金融教育的努力上，是台灣傳統商學院中轉變最快的學校。2016 年 2 月出刊的台灣銀行家雜誌報導高雄第一科技大學首創財金大數據中心，培育實作人才。高雄第一科技大學財務金融學院在迎接金融科技教育的挑戰上，能有如此快速發展與落實「金融 X 科技」跨領域的教育，何宗武教授所提出的建言著實功不可沒。

Data mining 是當今重要的大數據核心技術，提供對大數據的描述、探索以及模式的識別和預測。值此何宗武教授推出新書 Data mining with R Rattle 以饗讀者之際，獻上衷心的祝福。我堅信這本新書的問世，能造福更多 R 語言的學習者，並透過本書深入淺出的撰寫方式與內容，定能讓讀者輕鬆學習資料探勘的技術與應用，成為 R rattle 的高手！

林楚雄 教授
高雄第一科大財務金融學院院長
2016/9/2

機率原理

02
CHAPTER

樣本(sample)和母體(population)是統計分析的核心觀念。實際所收集的數據，不論量有多大，都稱為樣本，一個樣本內記錄的數據，稱為觀察值(observation)；例如，一間教室內有 50 個學生，教室是樣本，學生就是觀察值。這種關係，利用代數的名詞，樣本可以視為一個集合(set)，樣本內的觀察值則是集合內的元素(element)。一般我們會用下面的方式表示一個集合 X：

$$\{X|x1, x2, ...\}$$

樣本從哪裡來的？蒐集的。錯了，這不是一個腦筋急轉彎的問題，而是科學研究方法論的問題。所蒐集的數據不管多大，終究不是「所有」的數據。所以，樣本也意味著它只是部份資料。科學研究面對分析的對象，認為樣本是由一個預設的母體產生來的。母體就是種種理論上的機率分配，母體的性質，就是這些分配函數的性質。

然後透過對樣本的研究，推論母體的性質，這也就稱為抽樣(sampling)。舉一個例子，某大學校園所有的學生為母體，已知母體中男生和女生的比率為 6:4。如果我們不能知道母體性別比率，要如何推論這個比率？就是用抽樣。

隨機抽一次 100 個學生，記錄男女生比率為 4:6，這個數字和母體相差太大，兩者差距稱為偏誤(bias)。要降低偏誤，有兩個方法：

第 1。多抽幾次，例如，100 次；
第 2。抽多一點觀察值，例如，1000 人。

這樣，計算 100 次記錄的男女比率的平均數，這個平均數理論上會和母體的真實值很接近，因此，抽樣偏誤(sampling bias)就會大大降低。

在上例中，母體男女比率的已知是假設，統計實務上是不可能，所以，統計學的研究，提出了種種理論函數，透過假設來描述母體。機率學研究這些母體的性質，就是機率分配與隨機過程；統計學則是從分析抽樣，推論其母體，就是估計和檢定。

以「樣本-母體」為基礎的分析，類似「人類-上帝」，在學習統計數據分析時，需牢牢記住。此外，統計學的另一支發展，不從母體建立出發，稱為無母數統計(Nonparametric Statistics)，我們也會簡單介紹。本章簡單介紹機率，第 3 章再介紹統計。

2.1 數據分析第一步 -- 從樣本開始

資料分析的第一步就是要認識資料。所謂認識資料，就是要知道資料的性質為何。好比當你要分析一個人，是不是要先蒐集有關他特徵的資訊？身份證就是最基本有關這個人特徵的紀錄，再多一點，就是健康檢查報告。這些就是資料的性質(properties)，對這些性質的分析，就是資料分析。

以某股市上市公司平均報酬率來說 (檔名：AverageReturns.csv)，這筆資料也稱做樣本 (sample) 或樣本變數，檔內記錄 811 家公司的數值，稱為觀察值 (observation)。所謂的小樣本是指觀察值少，大樣本是觀察值夠多。一個標準的抽樣，一個樣本，至少要有 1 千個觀察值 (理論說是 1042 個)。

對 811 家公司的平均報酬，X-Y 軸的二維直方圖(histogram)是一個認識 811 筆平均報酬性質的好方法。圖 2.1-1 繪製了這筆資料的直方圖：

A. X 軸是報酬率的排序，可以知道極大值和極小值都是±5%延展。長方形有長(length)與寬(width)，寬是貼 X 軸的部分，長就是 Y 軸的高度。

B. 寬(width)就是 X 軸上，報酬率分群的一個群距刻度。好比 0 左邊長方形的寬，可以是(-0.15, 0)這樣的區間。

C. 長(length)在 Y 軸，則是落在 X 軸特定的群距，有多少家公司，就是次數。承上，約有 275 家。

D. 將所有的報酬率，在 X 軸分多個間隔區間，Y 軸就成為一個分佈，整體就稱為次數頻率(Frequency)。

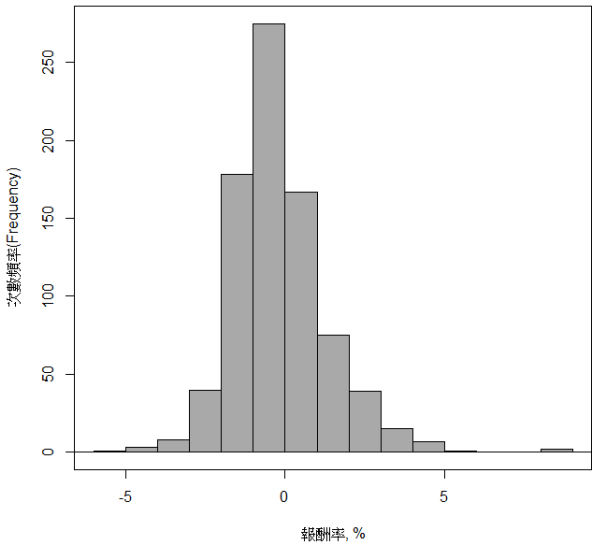


圖 2.1-1 上市公司平均報酬直方圖

直方圖告訴我們資料的基本性質，例如，我們可以發現幾點：

A. 極大值超過 5%，極小值也小於-5%。

B. 接近 0 的負報酬最多，接近 300 家公司。

C. 811 家公司的平均報酬應該在 0 附近。

如果我們把次數頻率(Frequency)的計數除以總公司數，就可以得到比率數據，如圖 2.1-2。

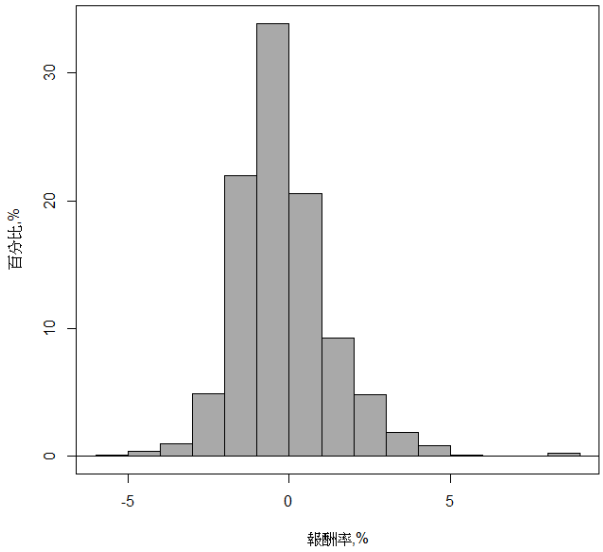


圖 2.1-2 以百分比表示的直方圖

圖 2.1-2 的直方圖是由樣本畫出來的。回到章首提到的「樣本-母體」關係，我們很自然要問：這個圖是由哪個母體產生的？或這個 811 筆觀察值的樣本，是由哪一個母體抽樣出來的？

樣本的統計學學名稱為隨機變數(random variable)。隨機這個字，代表了抽樣的非人為加工的特性。根據數字的特性，可以分成兩類隨機變數：離散型隨機變數(discrete random variable)和連續型隨機變數(continuous random variable)。離散型也稱間斷型，指數字是 0, 1, 2,…這樣的整數；連續型則是所有的實數，包含整數或小數都。

接下來，我們介紹幾種常用的理論機率分佈。

2.2 離散型理論機率分佈與 R Commander

離散隨機事件的機率，稱為機率質量函數 pmf (probability mass function)，機率的認識由稱為柏努利實驗(Bernoulli experiment)的隨機事件開始。

一個柏努利實驗是一個隨機實驗，實驗的結果可以被分類成兩種互斥且周延的事件，例如：成功或失敗、生或死。

令 X=1 為成功，X=0 為失敗，成功機率為 p，失敗為 q=1-p。

一個柏努利事件 X 的機率質量函數，可寫成：$f(x)=p^x(1-p)^{1-x}$。

也就是說：

成功 $f(x)=p^1(1-p)^0=p$

失敗 $f(x)=p^0(1-p)^1=1-p$

一串柏努利實驗稱為柏努利試行(Bernoulli trials)，指的是柏努利實驗被獨立地進行若干次之後的機率。這若干次的結果就是隨機變數或樣本。看以下的範例，會瞭解如何計算一個柏努利試行機率

某人投籃 5 次，成功進籃為 1，失敗為 0，投籃命中率為 0.8(p)。假設 5 次投籃為獨立，請問在這個投籃事件，只有第 1,3,5 次進籃的機率為何？

$$0.8\times(1\text{-}0.8)\times0.8\times(1\text{-}0.8)\times0.8=(0.8)^3\times(0.2)^2$$

認識了柏努利實驗的概念，我們就利用 R Commander 介紹 5 種離散機率分佈。R Commander 主選單上的【機率分佈】用在教學和抽樣的機率分佈函數，內有 5 種離散型機率分佈函數，每一種函數內皆提供 5 類性質。如圖 2.2-1 所示。

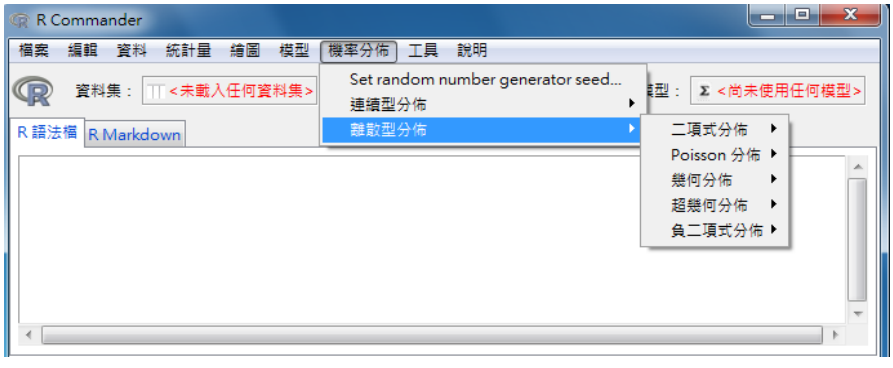


圖 2.2-1 R Commander 的離散型分佈函數

視覺化與探索性資料分析

06 CHAPTER

講到 R 的特色，不能不談到視覺化 (Visualization) 這一部份。然而視覺化相當專業，不是一個講酷比炫的電腦畫圖行為，而是透過視覺化解釋資料，一般也歸之於探索性資料分析 (Exploratory Data Analysis)。視覺化就資料階段而言，可以分成兩種層次：第 1 個層次是對原始資料的關係作分析；第 2 個層次是在資料探勘的模型對結果製作視覺化。

只要提到 R 套件中的專業繪圖套件，ggplot2 幾乎是無人不知。ggplot2 的重點在於有許多主流媒體的主題版型和號稱優雅的配色技術。這些套件如果要用程式語法才能畫圖的話，很多人都望之卻步。很幸運的，rattle 的視覺化除了內建 ggplot2 和 CairoDevice 之外，還有一個很專業的獨立套件 GGobi。因此本章第 2 節，專門介紹 GGobi。第 3 節則是筆者開發的 iClick 模組，將許多不容易的繪圖程式簡化為按鈕。這 3 個套件，對於資料探勘的視覺化資料解說相當有用。本章將一一介紹。

資料部份，我們基本上使用 rattle 內建的 weather.csv 氣象數據和 audit.csv 的報稅資料。氣象資料是一年的 22 個變數的氣象數據，記錄明天是否下雨 (RainTomorrow (No/Yes)) 與其他 21 個今天變數的關係，第 9 章的變數說明有詳細的介紹。載入後，依照內建的變數分類，不做更動。

6.1 基本探索分析圖

探索性資料分析 (EDA) 是資料探勘很重要的一環，依賴視覺化技術對資料結構和關係的處理。第 4 章簡單介紹過介面，這一章不再重複介紹資料載入和介面。因為視覺化必須仰賴色彩，故本章以彩圖逐節說明 rattle 視覺化執行 EDA 的重要功能。

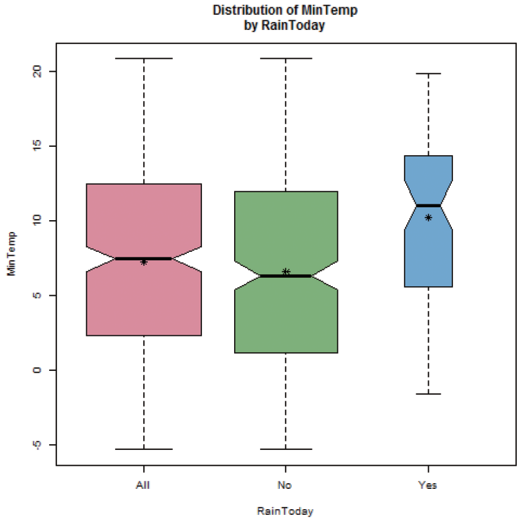
圖 6.1-1 是 Explore 選單的 Distribution 選項，以 4 個步驟完成繪製：

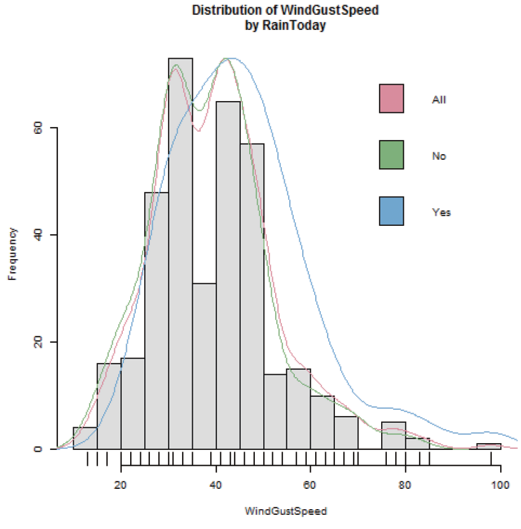
STEP1 選擇 RainToday 當作分組變數 (Group By)。

STEP2 選四個變數配圖。MinTemp 繪製 Box Plot(盒鬚圖)，WindGustSpeed 繪製 Histogram(直方圖)，Sunshine 繪製 Cumulative(累積機率密度)，WindGustDir 繪製 Mosaic(馬賽克圖)。

STEP3 在 Setting 處，把 Verbose 和 Advanced Graphics 取消，改選 Use CairoDevice。

STEP4 執行 Execute。





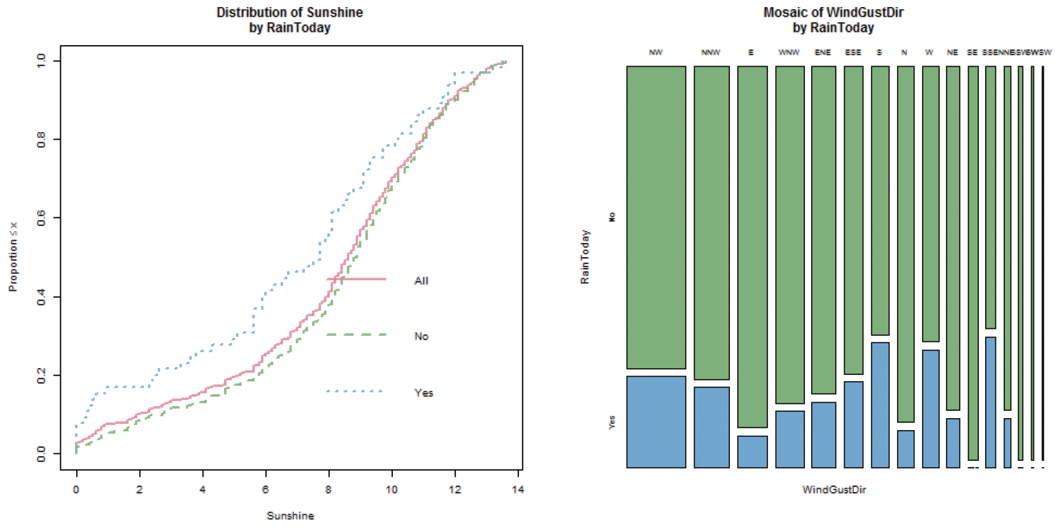


圖 6.1-1　四個變數的分佈圖

圖 6.1-1 的產製，是用 CairoDevice 繪製的，讀者可以試著用內定選項 Advanced Graphics 重畫一次。差異在於如果我們畫的四種圖有邊界問題，CairoDevice 能夠將之調整為放在同一頁，Advanced Graphics 則會變成兩頁。因此，什麼時候要用哪一種功能，要試過才會知道。而 ggplot2 產製的盒鬚圖，見圖 6.1-2。

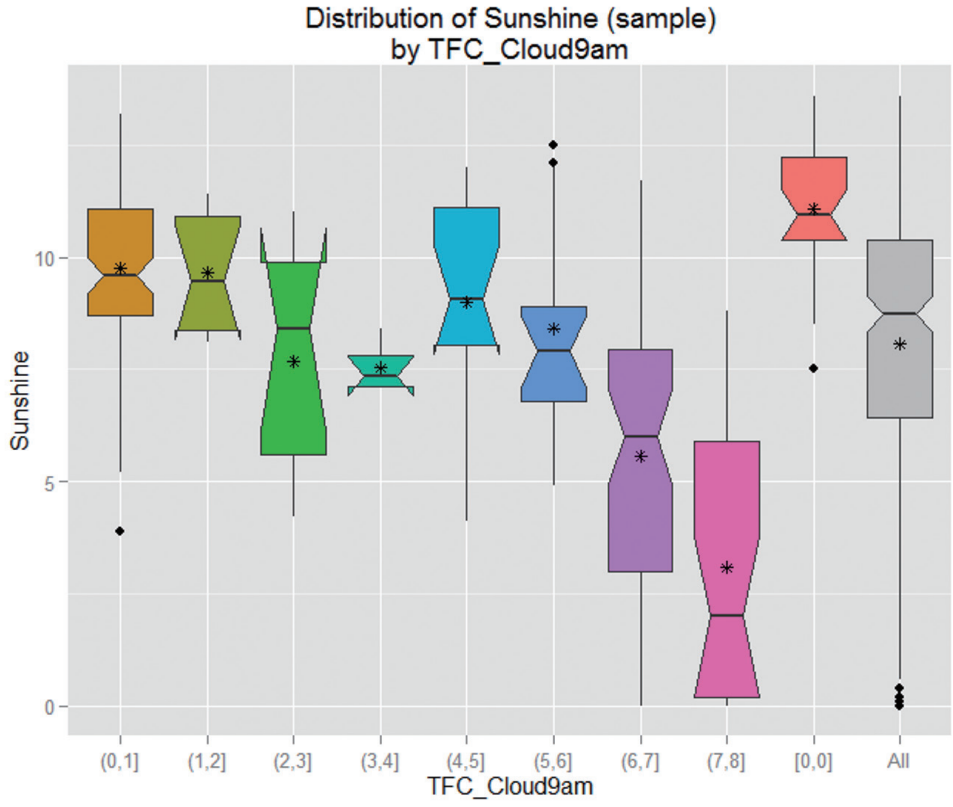


圖 6.1-2 Boxplot 的呈現，ggplot2

在 Distribution 選項的右邊有一個 Benford's law 選項，Benford's law 是用在判讀資料的奇異性，用在報稅資料可以找出可能有問題的報稅資料，這在會計資訊的大數據分析中是常用的工具。為了顯現這個工具的特色，我們用報稅資料 audit.csv 來畫圖 6.1-3。步驟如下：

STEP1 轉換變數為類別。因為 Group by 變數要選 TARGET_Adjusted。但是這個變數是數值，所以我們依照第 5 章 Transform 的介紹將之轉換為 As Category，產生新變數 TFE_TARGET_Adjusted。產生後，回到 Data 頁面，Execute 一下。

STEP2 進入「Explore」→「Distribution」。在 Benford 下方，勾選 Income 與之交會。然後在上面的 Group By 選 TFE_TARGET_Adjusted。

STEP3 Execute 後，產生圖 6.1-3。

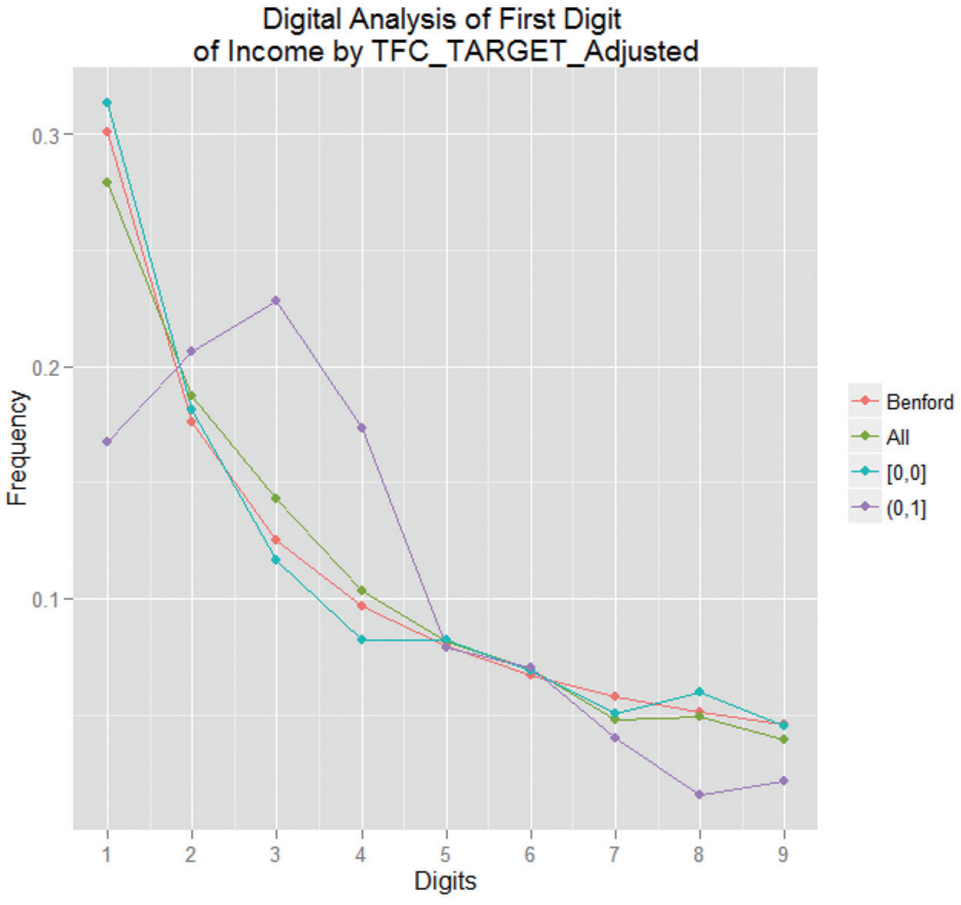


圖 6.1-3 Benford's Law 繪圖

接下來就是分析相關係數 (Correlation)，步驟為「Explore」→「Correlation」，再選「Ordered」。我們用 rattle 內建的氣象資料執行後，6.1-4(A) 是 Advanced Graphics，6.1-4(B) 是 CairoDevice。就視覺化的呈現，6.1-4(B) 的 CairoDevice 畫圖效果比較細緻。6.1-4(B) 的解讀，可以由主對角線類推：主對角線是自我相關係數，所以是 +1，圖形越細面積越小 (相關係數檢定之 P-value)，相關係數越顯著；負相關同理類推。

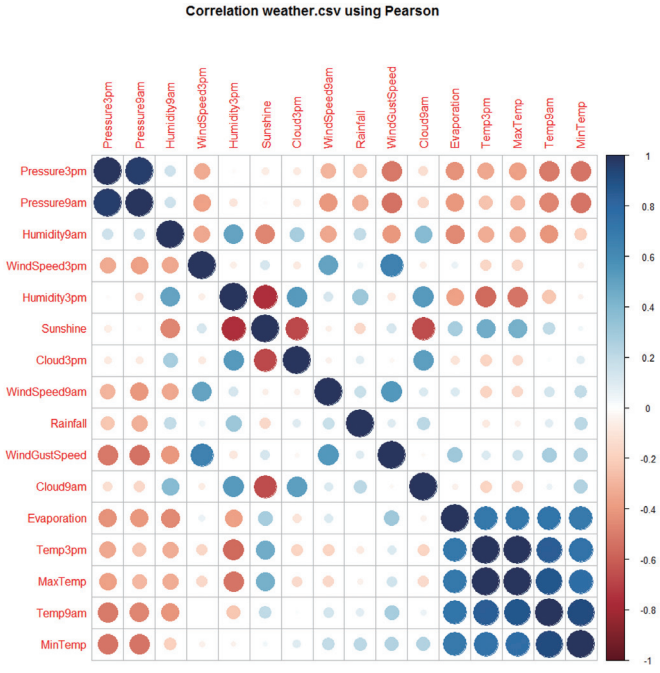


圖 6.1-4(A)　使用 Advanced Graphics

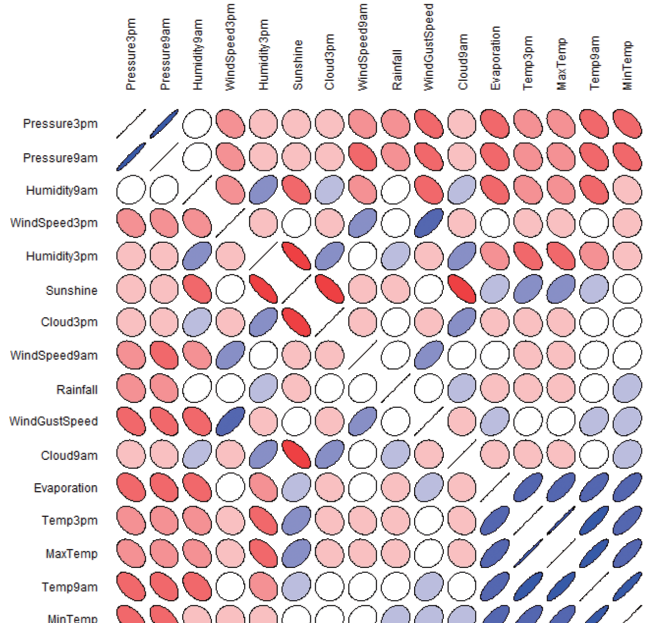


圖 6.1-4(B)　使用 CairoDevice

圖 6.1-5 是針對有缺值的變數，計算彼此間的相關係數。如果同時皆有缺值，就沒有相關係數。步驟為「Explore」→「Correlation」，再選「Explore Missing」。

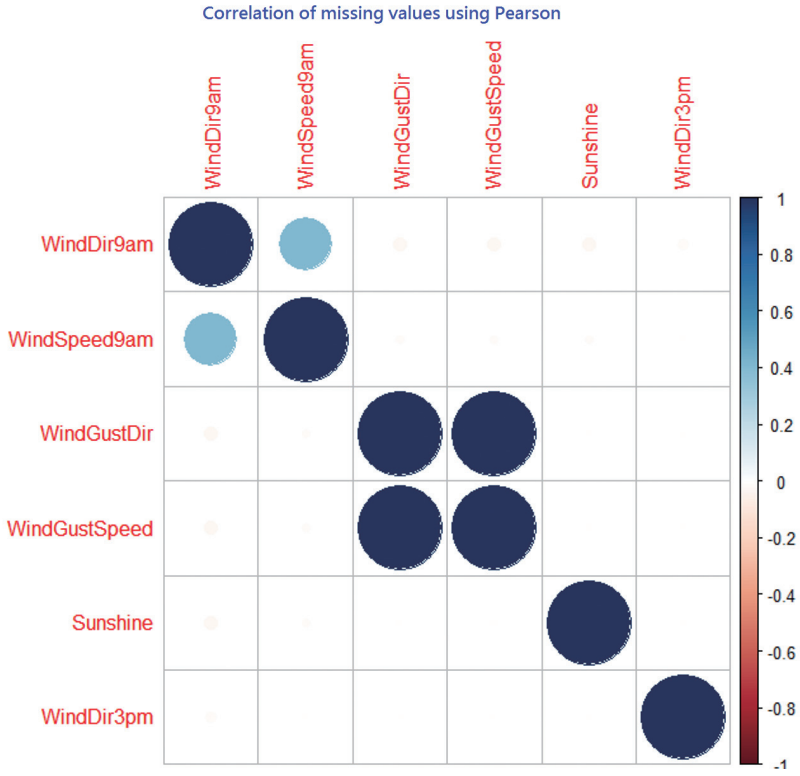


圖 6.1-5　**缺值的相關性** (Explore Missing)，Advanced Graphics

圖 6.1-6 是將兩兩相關的階層關係畫出來，步驟為「Explore」→「Correlation」，進入後選「Hierarchical Method」。這樣的階層圖，對資料探勘很有用。變數配對列在右邊，X 軸則是 0 ～ 3 的距離數字，代表了兩個變數的關聯強度：越短代表距離越近，相關性也越強。我們將階層結構以兩個方框框起來，有以下幾個觀察：

(1) 以兩兩為例，小框的兩個變數 {Temp3pm, MaxTemp} 距離最短，相關性最強。以 Pearson 係數看，它們的相關係數是 0.99。最下方的三個與風速 (Wind) 有關的變數，彼此關聯性成為最強的一組。上方的兩個，{Cloud3pm, Cloud9am }，相關性介於中間，Pearson 相關係數是 0.5。

(2) 大框框內的 5 個變數 {Temp9am, MinTemp, Evaporation, Temp3pm, MaxTemp} 看成一群，這群的關聯，高於它們和其他變數。

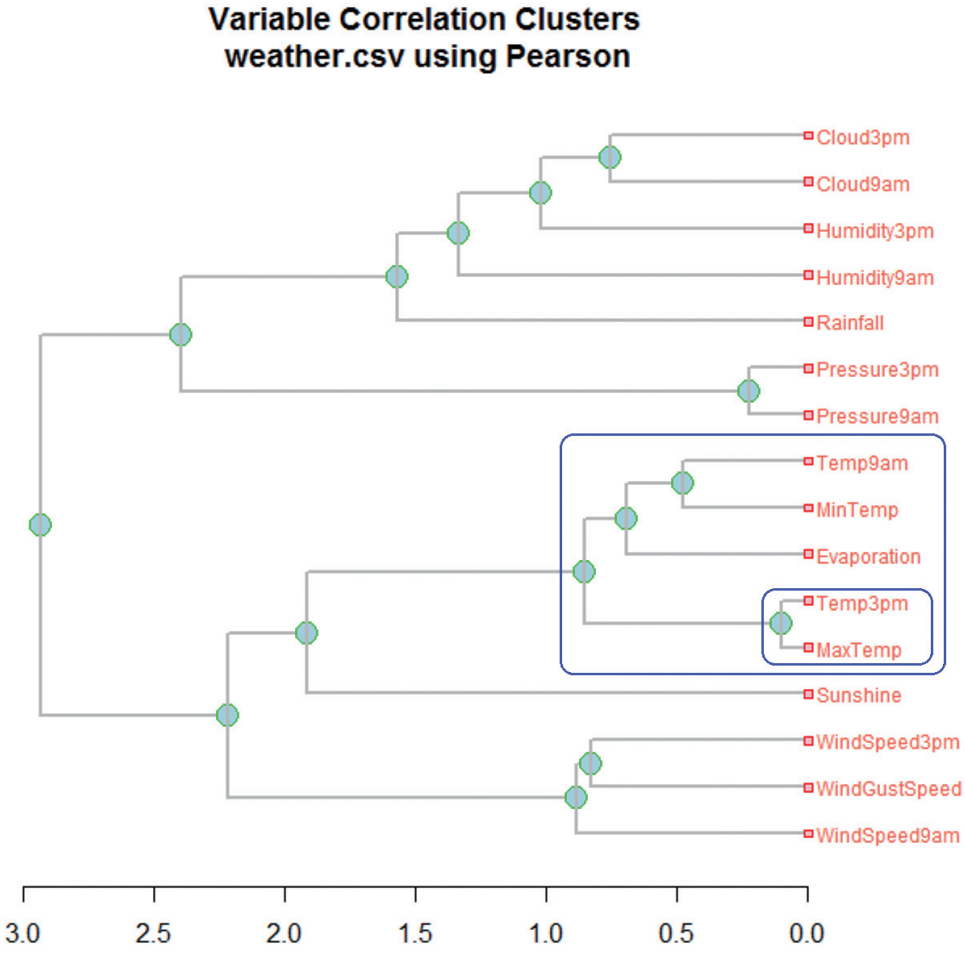


圖 6.1-6 **以階層法** (Hierarchical Method) **顯示相關**，Advanced Graphics

6.1-7 的散佈矩陣圖，則是依步驟：「Explore」→「Distributions」後，什麼項目都不選，直接按「Execute」鈕就會產生兩兩「散佈圖」和「相關係數檢定」的兩個半三角矩陣的合併結果；此處的散佈矩陣圖是只限數值變數。下三角是兩兩散佈配上 linear fit 和信任區間；上三角則是 Pearson 相關係數檢定的 P-value，同時，數值越大，數字顯示就越大。

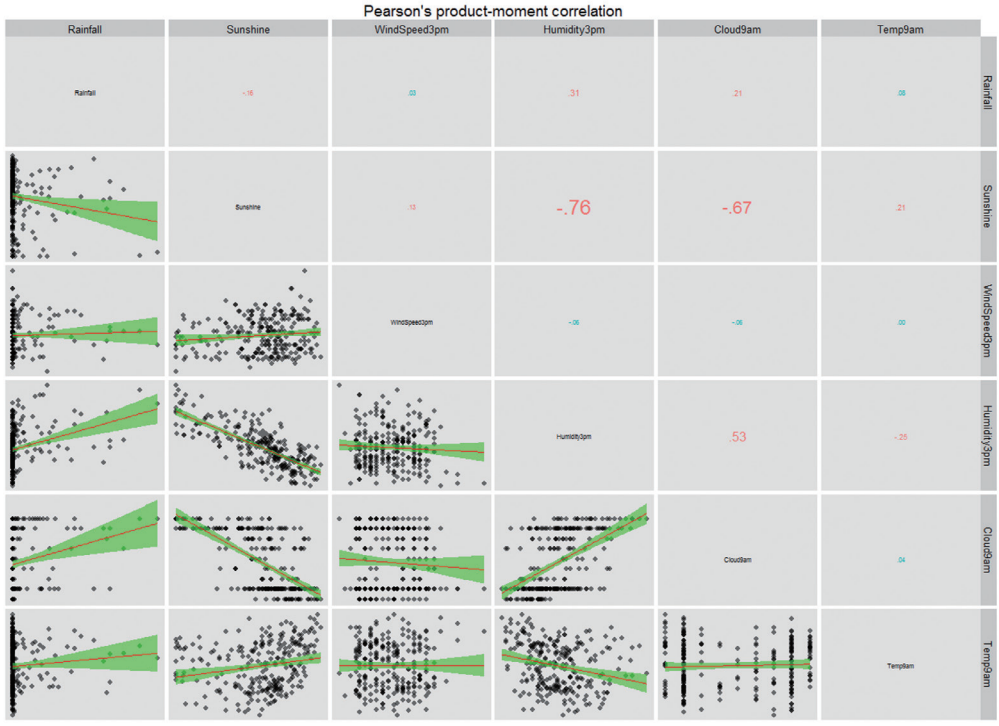


圖 6.1-7　散佈矩陣

6.2　GGobi 互動式繪圖

GGobi 基本上是一個獨立的開放繪圖軟體，爾後寫成 R 的套件可在 R 內部載入，但要單獨使用也可以。參考圖 6.2-1，由 rattle 主介面進入 GGobi 的互動視覺化區域步驟如下：

1. 進入主選單的 Explore
2. 點選最右邊的 Interactive
3. 自動內選 GGobi

如圖 6.2-1 所示，Execute 啟動 GGobi 兩個小視窗，左邊的 Control Panel 是控制面板，控制面板主選單，可以啟動許多視覺化功能；右邊是圖形顯示區，是顯示圖形的獨立視窗。說明如下：

1. 啟動軟體後，會自動的取最上面兩筆數據，繪製散佈圖 (Scatterplot)。如左圖顯示，X 軸是最低溫度 (MinTemp)，Y 軸是最高溫度 (MaxTemp)。

2. 右圖繪製兩兩散佈圖。內建顯示是兩軸沒有刻度，要顯示軸刻度，可以點選「Options」→「Show Axes」。

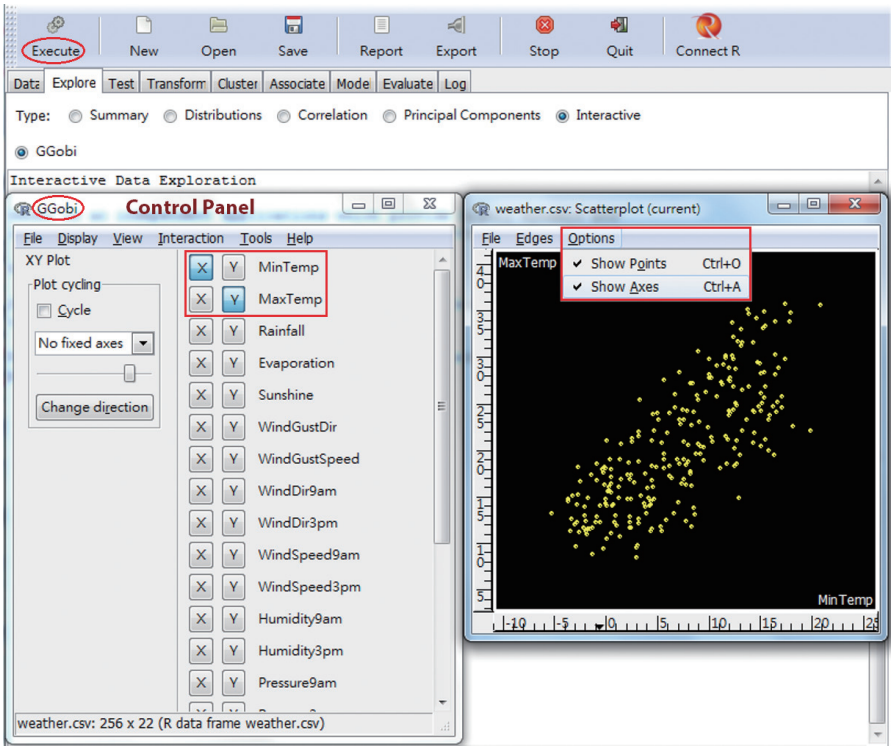


圖 6.2-1　啟動 GGobi 的 Control Panel

如果讀者對視覺化的要求不高，可以在 rattle 內看一看資料效果，但如果有進階處理的需求（例如，輸出高品質圖檔），那就需要使用獨立的 GGobi 軟體了，兩者顯示介面雖然完全一樣，但是在 rattle 內使用 GGobi 的效果，不如使用完整獨立的軟體要好。因此，建議讀者至 http://www.ggobi.org/ 下載適合作業系統的版本的執行檔 .exe，啟動後，從選單 File 處載入資料 weather.csv。GGobi 可以載入 .csv 和 .xml 兩類數據格式檔案。底下的解說便是使用獨立軟體，而不再從 rattle 內部使用。

圖 6.2-1 的散佈圖內，要標出明天是否下雨時，該如何標示？ GGobi 提供的「刷一刷 (Brushing) 」上色工具，就是為了解決這樣的視覺化需求。

6.2-1 刷一刷 (Automatic Brushing) 上色

如圖 6.2-2 所示，點選「Tools」→「Automatic Brushing」就會跳出右上方刷色的對話視窗。這個視窗分成上下兩部份：

上面是選擇要刷一刷上色的變數：選擇 RainTomorrow。

下半部是色彩配置：因為 RainTomorrow 是二元變數 (No/Yes)，依字母順序，從左至右：No 是黃色，Yes 是紫色。因為是二元變數，所以上方的 1.11~1.89 就沒有用，只有下方的 215 和 41 兩個數字，這兩個數字是 R 色版的顏色編號。1.11~1.89 是連續變數的分距，也就是說，如果要刷一刷上色的變數是連續變數，GGobi 會將之分成幾個群距，稍候會展示作法。

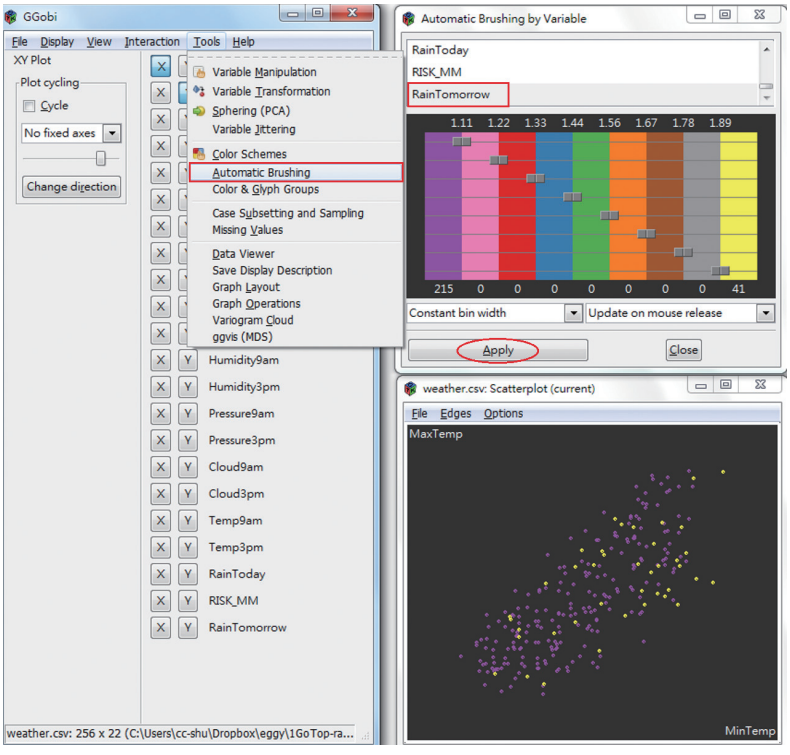


圖 6.2-2　刷 RainTomorrow

從圖 6.2-2 看來，整體的正相關，間錯著明天不下雨 (No) 的紫色。因為是隨機散佈，所以沒有顯示出具體的類型 (Pattern) 或集中度；因此，看起來是無關的。

接下來，我們刷 Rainfall (降雨量) 這筆變數。如圖 6.2-3 所示，降雨量高是黃色，低是紫色。顯示的圖形中，低降雨量似乎是主要的特徵。

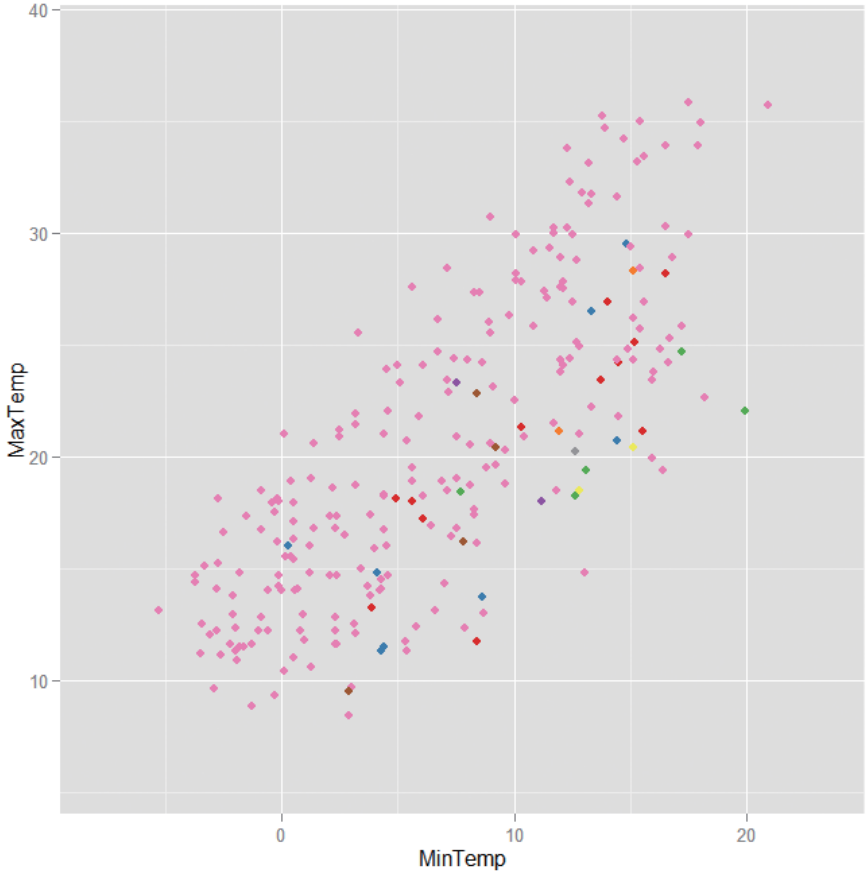


圖 6.2-3　以 Rainfall 刷色

圖 6.2-3 和圖 6.2-2 的黑底圖差異甚大。因為 GGobi 不能直接存取圖形，必須透過兩行語法，才能在 R 內產生高品質和解析度的圖形。在 GGobi 產生圖形後，輸出的步驟如下：

STEP1 在 GGobi 控制面板視窗點選「Tools」→「Save Display Description」後會出現對話視窗，詢問檔名和儲存位置。檔名可自行輸入，此處以 myFig.R 為名。要注意的是儲存位置，儲存位置的資料夾必須有載入的資料。

STEP2 底下程式檔 I-6_MakeFigure.R，載入套件 DescribeDisplay，執行語法就可以產生繪圖物件 tmp。

```
library(DescribeDisplay)
tmp<-dd_load("myFig.R")
plot(tmp)
print(ggplot2::ggplot(tmp))
```

繪製法有二，如最後兩行。plot() 是呼叫 R 內的高階繪圖，ggplot() 是呼叫套件 ggplot2 來畫。兩個不是任一即可，它們各有千秋，誰畫的好，要看資料形式，畫出來才知道。接下來的圖，都是這樣產生的。

ggplot2::ggplot() 的語法是說，我們不載入套件 ggplot2，但是使用它的繪圖函數，所以用 :: 宣告這個功能。使用量少的套件，不需要 library() 載入，可以節省電腦資源。雖然載入了套件 DescribeDisplay，讀者也可以將最前面兩行改成：

```
DescribeDisplay::dd_load()
```

另外，GGobi 選單上的 Display 內有一些圖形，在分析資料上效果相當好。

6.2-2 新型散佈矩陣圖 (New Scatterplot Matrix)

於控制台選單，點選「Display 」→「New Scatterplot Matrix」啟動散佈矩陣圖的控制面板。左邊會切換到只有選 X 沒有 Y 的選擇清單。GGobi 內建取前 4 個，我們稍作修改。右邊則和前面的介紹一樣，刷 Rainfall 變數。這樣的目的是讓被刷的變數是連續變數，方便切區塊。

按「Apply」鈕後，出現如圖 6.2-4 的散佈矩陣。這種散佈矩陣和 R 的散佈矩陣一樣，主對角線都是變數自己的密度分佈圖。其餘兩兩就是散佈。顏色上，就是依照刷色條件產生，相當清楚。

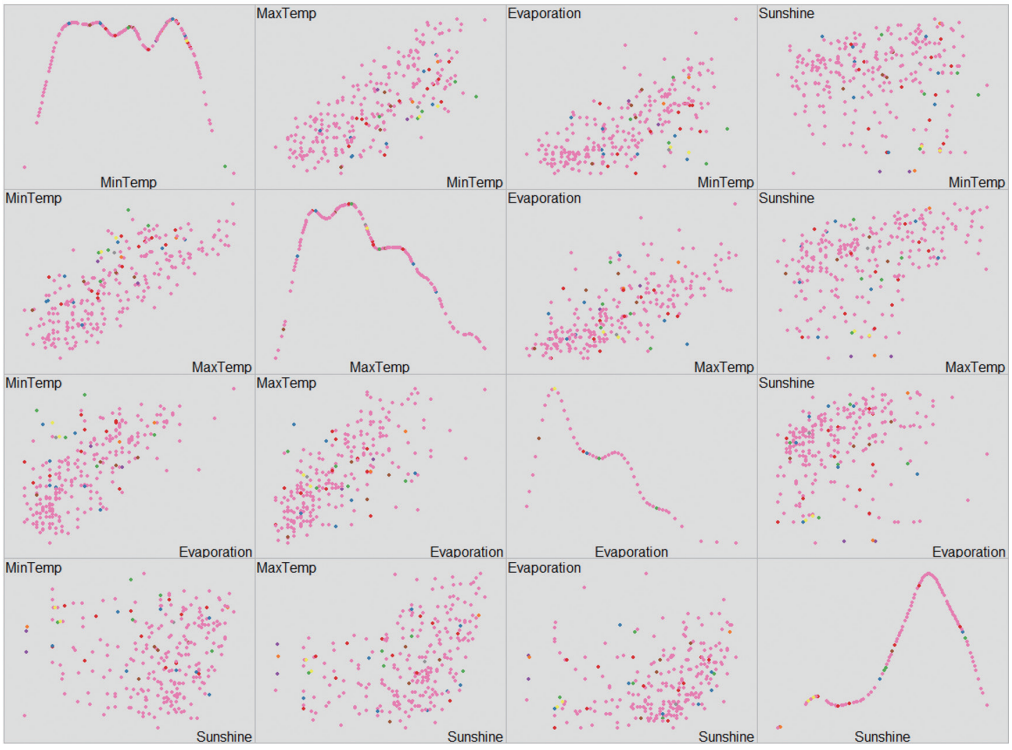


圖 6.2-4　刷色的散佈矩陣圖

6.2-3 New Parallel Coordinates

再來是處理多變量資料視覺化常用的圖形 Parallel Coordinates。由以下路徑啟動圖 6.2-5：點選「Display」→「New Parallel Coordinates Display」。

Parallel Coordinates 是一種多變量視覺化的技術，用於辨認分類集中度，我們可以看看 X 軸的變數是否具有區分類型的能力，這在資料探勘的領域中很常被使用。以圖 6.2-5 為例，這張圖以 RainTomorrow 刷色：No 是紫色，Yes 是黃色。選擇變數的先後，圖形 X 軸的排序，就是由左至右 (Sunshine, Evaporation, MinTemp, MaxTemp, Rainfall)。

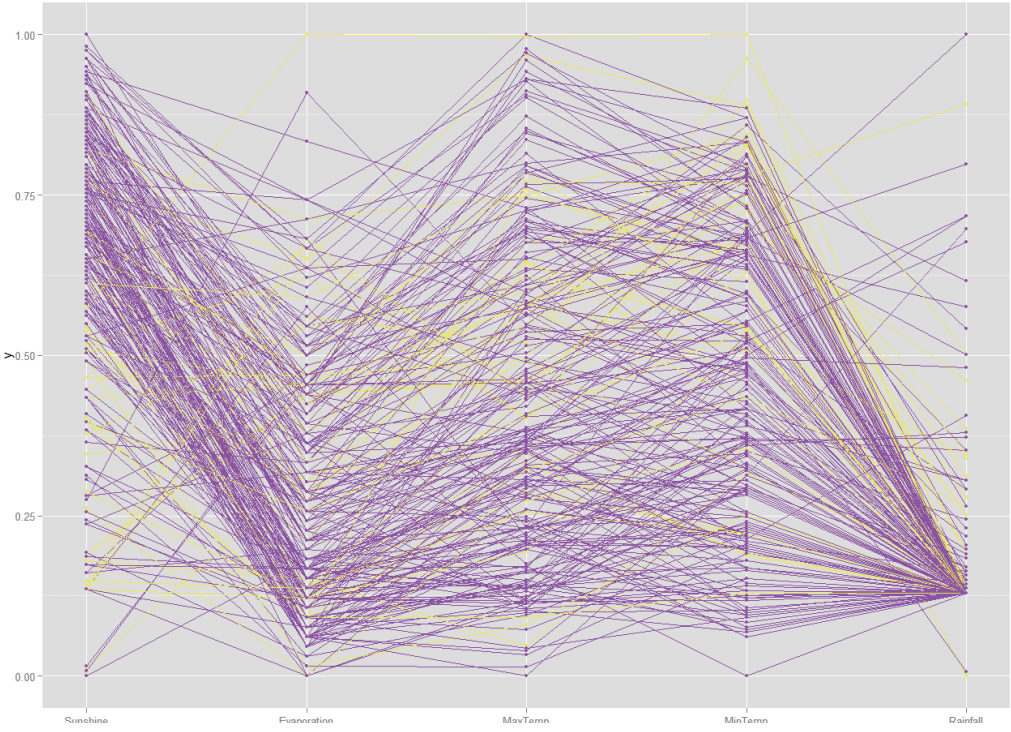


圖 6.2-5　以 RainTomorrow 刷色

我們可以看出最左邊 Sunshine 區塊中，高日照的紫色有群聚，也就是說，日照程度越高，明天降雨機率越低。最右邊的平均降雨量越低，明天降雨機率也越低。

圖 6.2-5 畫出來的是以 X 軸以外的變數 RainTomorrow 來刷色。圖 6.2-6 則是以 X 軸內的某一變數為刷色變數，我們取的是最左邊的日照，所以，在自己的區塊，色塊區分就很清楚。

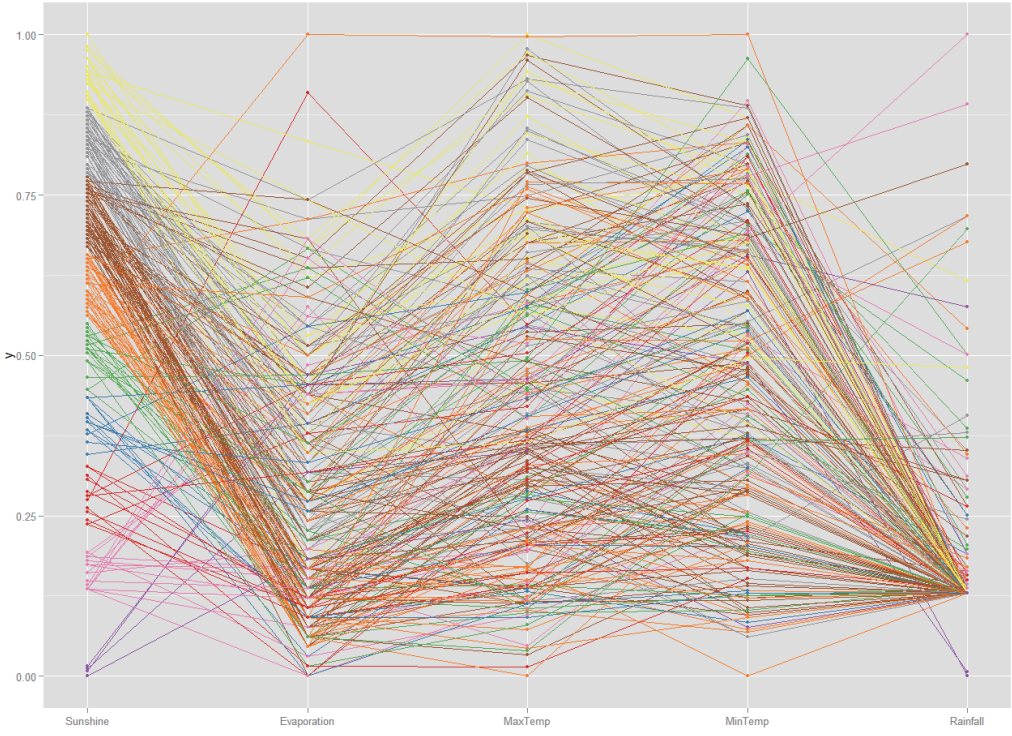


圖 6.2-6　依照 Sunshine 刷色

練習

請載入 oil.csv 的資料，對 region 刷色，繪出圖 6.2-7，解釋哪些變數區分產油的區域性？ Region 有三區：南方是紫色，黃色是北方，刷在最上方，綠色是 Sardinia 區。

圖 6.2-7

6.2-4 新堆疊長條圖 (Stacked Bar Chart)

在控制台選單，點選「Display」→「New Barchart」，選擇後會出現變數選單，一次只能選一個，圖 6.2-8 為選擇 Pressure3pm 然後用 Sunshine 刷色。這張圖的 X 軸是 COUNT 也就是計數「幾天」，這樣可以看出下午三點氣壓 (Pressure3pm) 極高或極低時，日照度偏低且一年之中沒幾天是這樣狀態。氣壓屬一般狀況時 (1015 上下)，各種日照度都出現且是最多的。

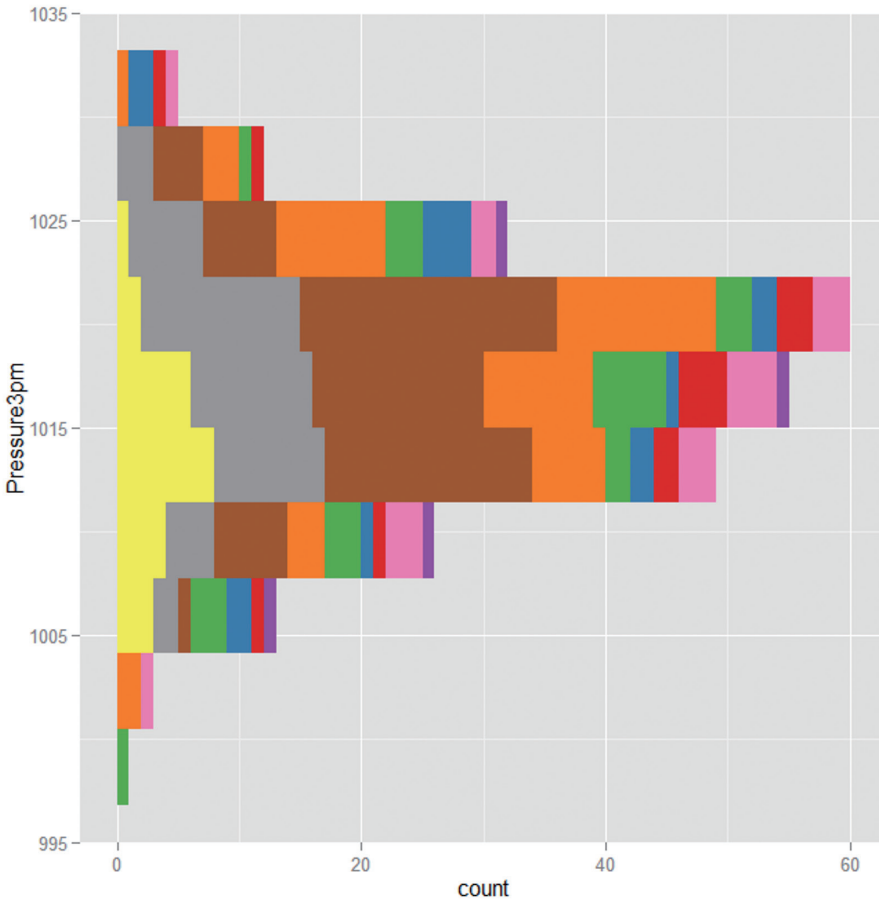


圖 6.2-8 New Barchart

6.3 iClick 的高階繪圖

iClick 是筆者開發的 R 套件，完整版本於 2015 年 12 月發佈在 R 的官網。iClick 是一個視窗介面，因為簡易，所以命名為「一指搞定」。iClick 設計的思維是希望單純只是要使用 R 的人，能夠用最簡易的 3 步驟，就可以完成複雜的演算程式。

iClick 的家族很多，本書提供兩個處理視覺化元件：VisMVCor 和 dailyTS。VisMVCor 是為了處理多變量數據的相關係數，dailyTS 是為了時間序列日資料設計的視覺化套件。

執行 iClick 所需要的 R 套件，都在 iClickPackages.R 這個檔案內，讀者只要在 R 主控台載入這個批次檔，執行 (Run all) 即可。

6.3-1 高階相關係數矩陣繪製：iClick.VisMVCor

依照以下 3 步驟，即可輕易啟動 iClick.VisMVCor。

STEP1 載入檔案：load("iClick.VisMVCor.RData")。

STEP2 載入資料：dat=read.csv("Cars93.csv")。

STEP3 啟動 iClick：iClick.VisMVCor(dat)。

如圖 6.3-1 所示，就可以啟動 iClick 介面。

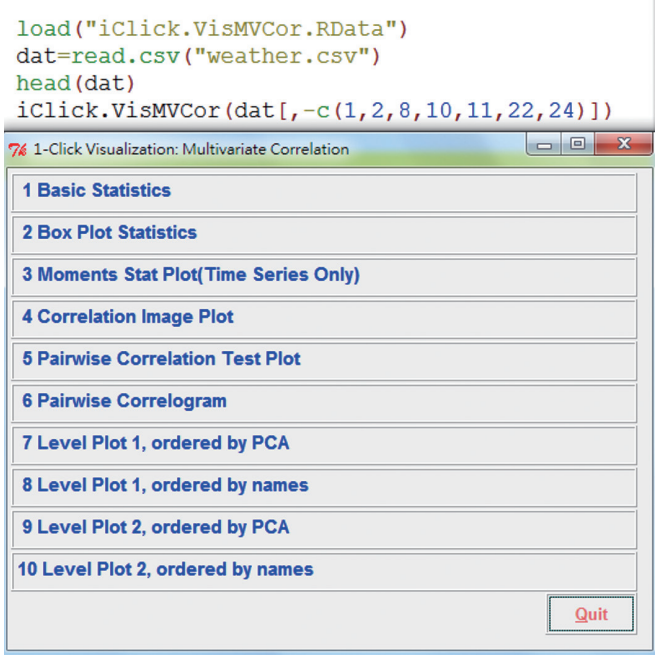


圖 6.3-1 iClick.VisMVCor 介面

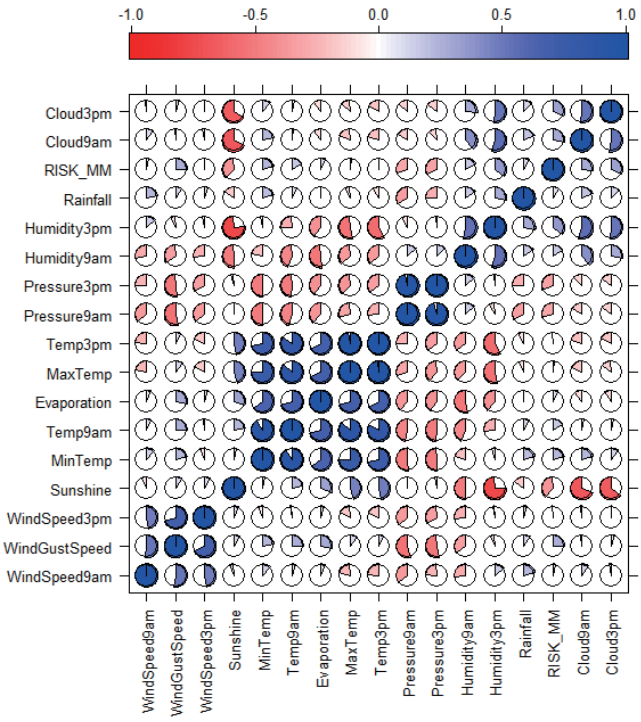


圖 6.3-4　編號 9 的相關係數之 levelplot

6.3-2 多個資產報酬率的比較

因為 Cars 資料的特性，iClick.VisMVCor 前兩個比較基礎敘述統計量的圖形不夠清晰。我們載入 24 個國家股市指數報酬率數據，如圖 6.3-5 上方輸入指令部分所示。

```
load("iClick.VisMVCor.RData")
dat=read.csv("returnsDaily24.csv")
iClick.VisMVCor(dat)
```

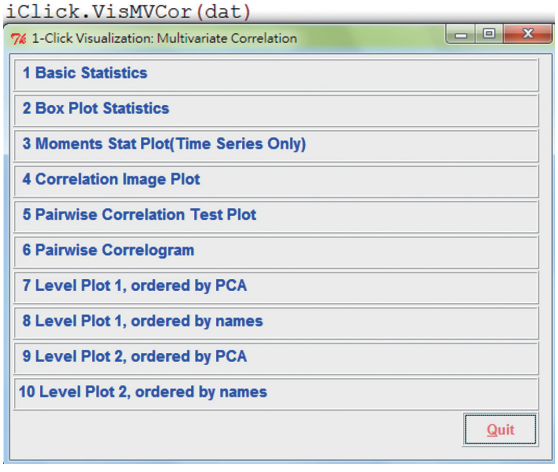


圖 6.3-5　在 iClick.VisMVCor 載入時間序列

使用這樣的功能，只要資料的第 1 欄是時間欄位即可。前 3 個是以視覺化的方式比較資料的敘述統計量，這功能和之前只是呈現個別資料相當不同。可以透過視覺化知道哪一個國家的平均數最大，或標準差最小。見圖 6.3-6。

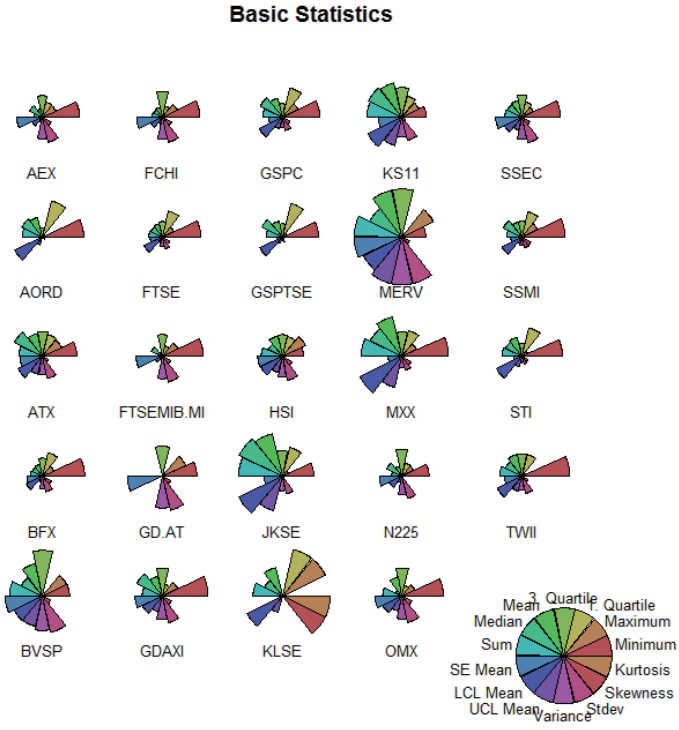


圖 6.3-6　編號 1 的比較基礎敘述統計 Basic Statistics

圖 6.3-6 比較了各國 14 個基本的資料性質。例如，極大值 (Maximum) 和峰態 (Kurtosis)，KLSE 馬來西亞這個股市，扇葉都是滿葉，也就是 24 個國家中，這兩個數據它都是最大。然後，極小值 (Minimum) 這個數據，多國幾乎都差不多，台灣是各國最大。圖 6.3-7 繪製了盒鬚圖比較，圖 6.3-8 則是將圖 6.3-6 簡化為 4 個敘述統計量，也就是 4 階動差。比較方式都相同，本書就不在此重述，讀者可以自行檢視。

6.3-3 時間序列日資料繪製：iClick.dailyTS

依照以下 4 步驟，即可輕易啟動 iClick.dailyTS，並以聯發科股價當數據解說。

STEP1 載入檔案：load("iClick.dailyTS.RData")。

STEP2 載入聯發科股價資料：

```
assetPrice=read.csv("mediaTek.csv")
```

STEP3 定義 2 欄資料，第 1 欄時間，第 2 欄欲繪製之資料，這裡是用 c(1, 5) 宣告原始資料的 2 個位置：

```
DAT=assetPrice[,c(1,5)]
```

STEP4 啟動 iClick：

```
iClick.dailyTS(DAT,color4=  r2b  ,color5=  jet  )
```

```
load("iClick.dailyTS.RData")
assetPrice=read.csv("mediaTek.csv")
head(assetPrice)
DAT=assetPrice[,c(1,5)]

iClick.dailyTS(DAT,color4="r2b",color5="jet")
```

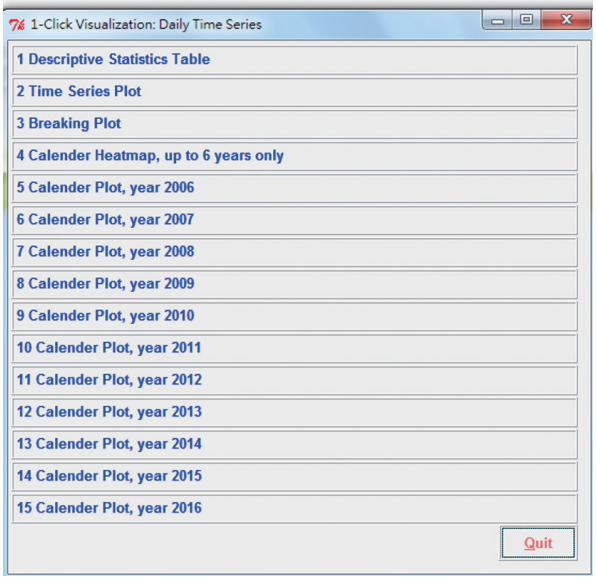


圖 6.3-9 iClick.dailyTS 介面

函數內的 "r2b" 是色彩，color4 這組有 3 種選項："r2b"、"r2g"、"w2b"；color5 這組也有 4 種選項："default"、"heat"、"jet"、"increment"。color5 是繪製編號 4 的 Calendar Heatmap；color4 是繪製編號 4 以降的 Calendar Plot。見以下兩圖。

編號 4 的 Heatmap 只限最近 6 年，這是因為視覺化要求使然；如果年份太多，就會跨頁或擁擠。但是編號 4 以降的 Calendar Plot 有多少年，就會依照資料自動產生選單，但如果資料太多年，會導致選單長度太長而容易超過螢幕，讀者在使用時請特別留意這點。

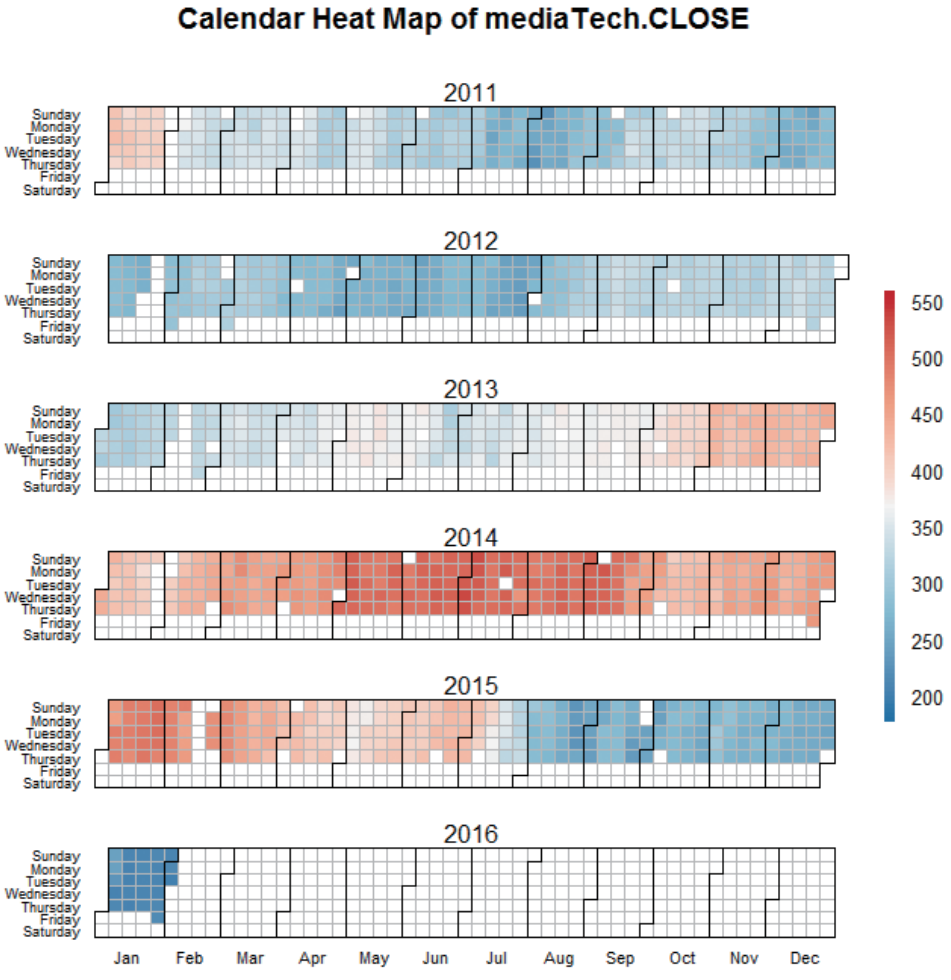


圖 6.3-10　編號 4 以日曆熱力圖顯示聯發科近 6 年的收盤價