資料視覺化的奇幻之旅

彭其捷

資料視覺化的前置作業：資料準備與資料清洗

作者介紹

彭其捷鑽研網路服務多年，出版過 3 本 UX 使用者經驗專書，曾大量接觸設計、工程，使用者體驗等相關工作，近年因大數據 & 物聯網概念蓬勃發展，觀察到越來越多數據導向的服務興起。然而，艱澀的數據需要良好的設計輔助，才能創造良好的閱讀體驗，其中特別依賴資料視覺化的相關能力。因此，本專欄特別針對各類網路服務的資料呈現提出美學觀點，分享一些國內外資料視覺化的概念、工具與案例。

系列文章介紹

FINDIT 的目標是『發現趨勢，看見未來』，事實上，眾多的趨勢就隱藏在眾多的數據當中，等待著人們去發現、去解讀。透過『資料視覺化』的輔助，能夠把冷冰冰的數據圖像化，協助人們掌握趨勢，更能夠協助新創企業在創新創業過程中找出正確方向，或是幫助投資人找到潛在投資標的，是一項強大的武器。

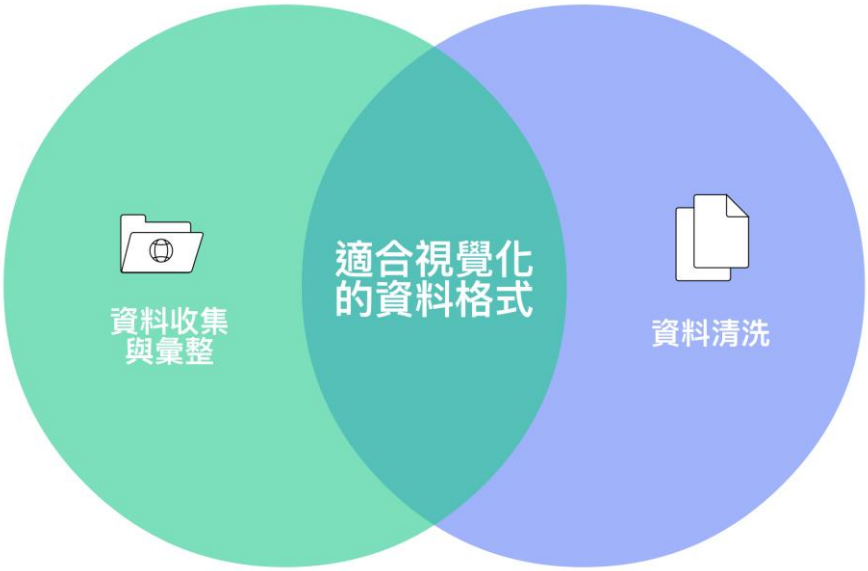
本系列文章將從資料視覺化的概念開始著手，之後會陸續分享視覺化的經典案例，相關工具介紹等等，希望讀者們能夠在大數據時代，透過視覺化的輔助，說出好的故事。

前兩個月我們介紹了許多 Tableau 的操作方式，但讀者可能會想說，通常接到的資料都是雜亂不堪，甚至有時還要自己去尋找合適的資料，現實世界很少有可以直接使用的資料等等。本期內容具聚焦在正式進行資料視覺化任務之前，如何資料收集與彙整，以及資料清洗，讓讀者可以事半功倍，順利進行資料視覺化之旅。

前兩個月我們介紹了許多 Tableau 的操作方式，但讀者可能會想

說，通常接到的資料都是雜亂不堪，甚至有時還要自己去尋找合適的資料，現實世界很少有可以直接使用的資料等等。也就是說，現實會遇到的多數情況是，正式進行資料視覺化任務之前，需要先完成兩大步驟：

- 資料收集與彙整：從網路上、其他組織或是次級資料檢索的方式尋找合適資料素材
- 資料清洗：處理雜亂資料，並不同的資料格式進行對齊，讓其產生對應關聯性



資料來源：本文作者。

圖 1 進行資料視覺化任務的兩個前置作業

Part 1：資料收集與彙整

一、視覺化資料三大類型

我們總期待能夠有一個結構良好、乾淨的資料表，但現實世界中其實是會遇到各式各樣的資料類型，可能是資料格式不一致，或是格式需要自己定義等等狀況，許多資料整理的挑戰可能超越想像。我們可以將視覺化資料區分為三大類，分別是『結構化資料』、『半結構化

資料』與『非結構化資料』，以下分別介紹：

（一）結構化資料 (structured data)

結構化資料表示資料擺放整齊，在當初收集的的時候就已經有完整的定義，且均是依照一致的收集邏輯進行，不論是欄位、格式、順序等等都是相同的，非常容易理解，不太會有意外。

結構化資料通常是透過程式所產生，多數是從傳統的關聯式資料庫取出，清楚明瞭，可直接提供相關軟體進行後續利用，也是視覺化作業的最優起跑點。

（二）非結構化資料 (unstructured data)

非結構化資料，從名稱就可以看出其麻煩之處，也就是『沒有特別結構』的資料屬性，相對較難以處理，有各式各樣的變異性等等，但還是有可能存在著一定的規則性，所以仍可歸在資料的範疇當中。

這世界存在著非常大量的非結構化資料，甚至包括著許多需要等待著被人類所定義的非結構化資料。舉例來說，網路上大量的部落格文章，就可以算是一種非結構化資料，每個人會有各自的寫作風格，每個人也可能使用不同工具進行文章的撰寫等等，然而，這些部落格依然可以根據某些特定資料屬性的定義，而被 Google 之類的檢索工具所檢索到，像是『文章中有特定關關字』、『常常被人們點閱』或是『部落格文章使用的語言』等等，也就是說，雖然乍看之下是非結構化資料，但只要經過人類的定義，就有可能是寶貴的資料資產。

（三）半結構化資料 (Semi-structured data)

有些資料格式屬於『半結構化資料』，從名稱可以看出介於『結構化資料』與『非結構化資料』之間，為什麼會有這樣的格式呢？因為這世界許多資料累積的狀況並沒有辦法像結構化資料如此完美，例如某些屬性只存在部分資料，或是某些欄位的資料可能存在多種格式等等。相對非結構化資料更有邏輯性，但還是需要整理過才能拿來使用。

二、適合視覺化的資料格式

怎樣才算是一個對資料視覺化作業來說品質良好的資料呢？一般來說是『定義清楚的結構化資料』，即一個蘿蔔一個坑，且每個坑規格都相同，資料欄位格式都相同，對於視覺化作業就相當方便。一個適合資料視覺化的結構化資料，通常包括兩大部分：

- 資料表頭（第一列，說明該欄的用途）
- 資料本體（第二列以下，擁有相同的格式的資料）

熱點名稱	地址	緯度	經度
立法院大門會客室	100臺北市中正區中山南路1號	25.043965	121.519581
立法院議場	100臺北市中正區中山南路1號	25.043717	121.520635
國家圖書館2樓參考室	100臺北市中正區中山南路20號	25.03717	121.516567
國家圖書館5樓閱覽室	100臺北市中正區中山南路20號	25.03717	121.516567
國家圖書館6樓閱覽室	100臺北市中正區中山南路20號	25.03717	121.516567
國家圖書館B1樓閱覽室	100臺北市中正區中山南路20號	25.03717	121.516567
國家圖書館閱覽室	100臺北市中正區中山南路20號	25.03717	121.516567
國立中正文化中心國家戲劇院福華劇院軒入口處	100臺北市中正區中山南路21-1號	25.036675	121.519046
國立中正文化中心國家音樂廳售票口左側	100臺北市中正區中山南路21-1號	25.036675	121.519046
國立中正紀念堂管理處堂內大忠門及大孝門服務台	100臺北市中正區中山南路21號	25.0347299	121.521932
教育部一樓大廳	100臺北市中正區中山南路5號	25.042749	121.51915

資料來源：本文整理。

圖 2 適合視覺化的資料格式示意

三、從哪裡取得資料？

了解資料類型之後，我們關心的下一個議題是：『可以從哪裏取得資料呢？』，雖然我們號稱大數據時代已經來臨，但多數資料都是屬於未開放的『組織內部業務資料』，許多分析團隊遇到的第一個挑戰，就是沒有合適的資料可使用，也很難進行後續的業務整合，以下介紹一些常見的資料取得方式：

（一）內部業務資料或是跨單位合作資料

多數數據都是組織內部管理的， 如果是其他公司的業務資料當然就更難取得了，但即使是自己公司的資料，也並非一定可順利取得，其中可能會遇到政治問題（不一定有權限取得相關資料），也可能遇

到技術性的問題（內部儲存的格式不符合視覺化的期待），前者需要依賴跨單位的溝通來整合彼此對於業務上的共同目標，後者則需透過資料的整理技巧進行修正。

（二）透過第三方 API 取得資料

另一種資料取得方式，是透過第三方服務取得，多是透過 API（指提供資料窗口）的方式來進行，例如 Facebook 就開放讓開發者在取得使用者的同意之下，取得特定使用者的朋友名單、Email、興趣等等個人資訊，並存放於企業資料庫當中。

（三）手動製作資料

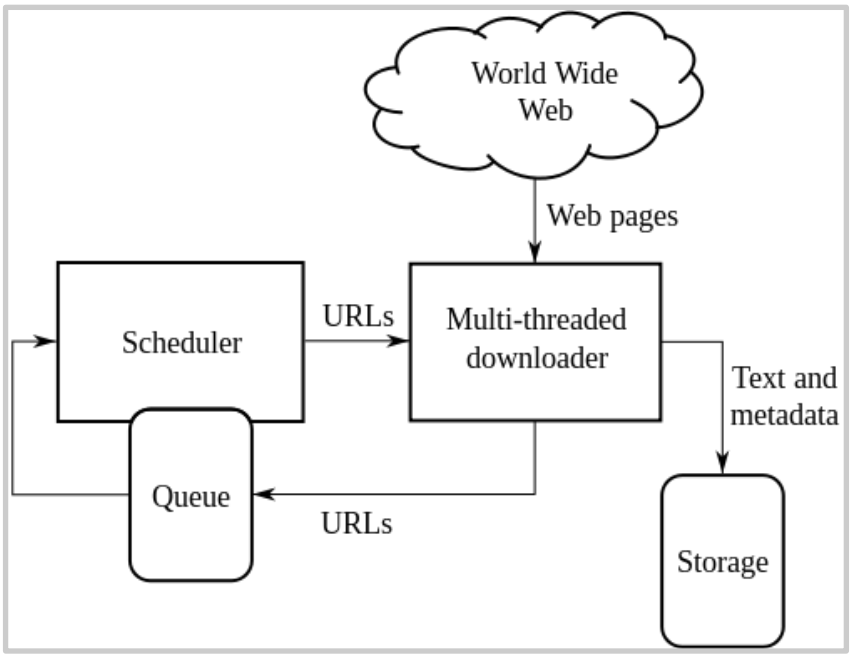
初期如果真的找不到適合的資料的話，由人工或是程式重新累積也是一種選擇，例如：

- 聘雇一些人參與實驗來累積統計數據
- 請人現場進行紀錄，人工紀錄收集數據
- 用程式的方式模擬可能產生的行為數據
- 人工編輯資料表來產生結構化資料

手動製作的方式最大好處在於流程透明，我們能夠很清楚的知道目前資料產出的邏輯，對於數據分析或是視覺洞察很有幫助。我們也可以考慮用外包的方式產生數據，像是委託第三方公司產生數據，或是取得社群的支援來累加數據等等，都可以製作出可供後續利用的資料。

（四）撰寫網路爬蟲程式取得資料

所謂的網路爬蟲，指的是透過程式方法全自動進行外部資料撈取，例如 Google 就透過無數的爬蟲無時無刻在更新資料庫的數據狀態，整理後便能夠建立龐大的檢索系統，讓使用者輸入關鍵字後，取得過濾後的有價值資訊。



資料來源：https://zh.wikipedia.org/wiki/網路爬蟲

圖 3　網路爬蟲的流程圖，透過排程機制，定期前往目標網頁下載資料回來

（五）透過開放資料（Open Data）取得全世界資料

最後一種取得資料的方式，是透過開放資料 (Open data) 的方式，這是近幾年全世界的一股趨勢，指的是經特定組織或是業務單位挑選過後的資料，願意提供給公眾進行使用，這些資料不受著作權、專利權等等的法律所限制，有些也不受相關商業的考量限制，可以公開進行使用。

資料來源：https://www.mcc.gov/

圖 4 透過開放資料（Open Data）是近幾年流行的資料取得方法

根據國際開放（Open Definition）的定義，所謂的『開放性』指的是：推動任何人都可以在最大互通性下進行參與的共享領域，讓知識提供自由存取、使用、修改，以及分享，下表整理了國際上對於開放的定義。

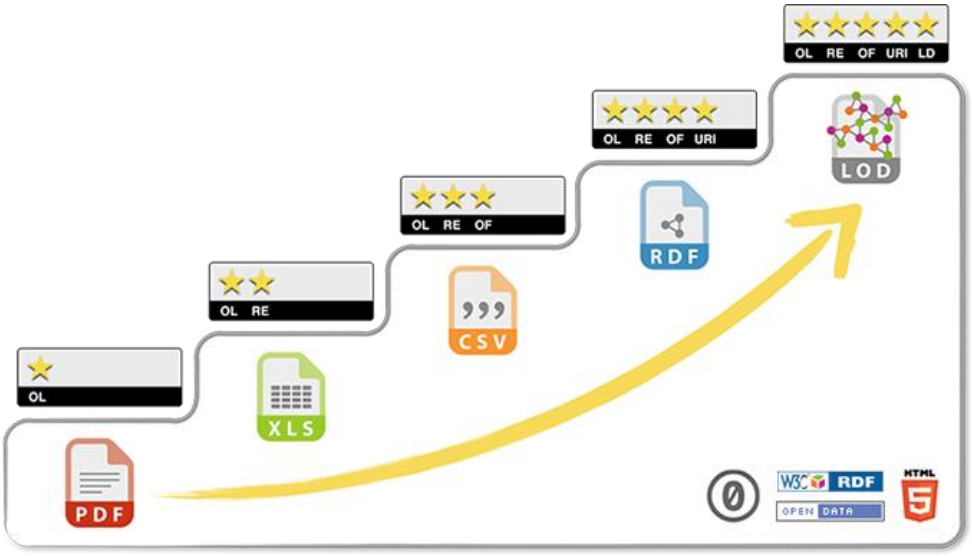
表 1 國際上對於『開放』的定義（Open Definition）

項目	說明
開放授權	該作品必須以開放的授權條款進行提供。 任何附加於該作品的額外條款（例如使用條款，或授權人所持有的專利權），皆不得 與前述授權條款產生衝突。
方便近用	該作品應以其完整狀態，且僅收取一次性合理重製工本費用的方式來提供，如果是透過網際網路，那較佳模式就是以免費的方式來提供下載。任何授權遵循所需要的額外資訊（例如因註引出處而被要求的貢獻者姓名表）也 必須 伴隨作品一併提供。
開放格式	該作品必須以合宜且可修改的格式來提供，亦即無不必要的技術限制置於授權權利的行使上。明確的說，資料應該是以機器可讀、批次不零散，且開放格式（例如，該格式之規格可被自由公開披露，且無收費或其他限制於其使用上）的方式來提供；或最低限度，可被至少一款的自由開源軟體工具進行處理。

資料來源：http://opendefinition.org/od/2.0/zh-tw/）

四、開放資料等級

開放資料除了代表歡迎提供給其他人利用之外，之中也存在品質的層級議題。全球資訊網（World Wide Web）發明者和鏈結資料的創始者：提姆・柏納-李 (Tim Berners-Lee) 設計了一個開放資料五顆星等級分類架構，所下所示：



資料來源：http://5stardata.info/zh-TW/

圖 5 開放資料的五星級架構，越高層級代表揭露越多的資料之間串接邏輯

以下說明五種等級分別代表的意義：

- 一顆星：採用開放授權，讓手上的資料（任何資料格式）可以在網路上取得
- 二顆星：讓這份資料能以結構化的方式取得（例如用 Excel 取代掃描的表格）
- 三顆星：使用開放格式取代專屬格式(例如用 CSV 取代 Excel）
- 四顆星：使用『固定網址』來表示資料，使其它人可以連結到資料在網路上的位置
- 五顆星：鏈結你的資料到其它資料，可提供資料之間的脈絡關係，例如兩份資料間的相等關係（owl:sameAs）

由於網路大抵上是一個開放的世界，所以以上分類主要是一個概念，而非強制性的規範，所以我們並不是一定能在開放資料頁面看到相關資訊的揭露，但此圖依然明確提供了最主流的開放資料品質標準。

五、台灣政府資料開放平台

- 網址：http://data.gov.tw/
- 特色：有許多很棒的台灣區資料，可免費下載

政府資料開放平臺，是中華民國政府根據《政府資訊公開法》規定，所建立的開放資料計畫，上面有許多台灣的資料集，也採用許多標準開放資料格式像是 CSV、XML、JSON、OLAP、TXT 等等，也歡迎各組織引用或進一步應用。

資料來源：http://data.gov.tw/

圖 6 台灣政府開放資料首頁

Part 2：資料清洗

即使已經取得資料，對於企業來說，最花人力的部分可能是資料整理的工作，在資料視覺化的任務中，許多專案的清洗時間甚至會佔到一半以上，但我們須要投入時間來避免『Garbage-in Garbage-out』的問題，再強大的視覺分析能力，也拯救不了品質不好或定義模糊的

資料。

許多資料產生的初衷並非為了執行視覺化作業，絕大多數的資料取得後，並無法直接進行分析，資料產出的過程當中會有各種原因導致資料結構不完整，或是俗稱的：『很亂的資料』，需執行像是：修正錯字、修正格式、彌補空缺等等任務，此部分的時間成本容易被低估，在視覺化流程當中是相對容易被忽略的部分，花費比預期之外更多的時間。

一、雜亂的資料

資料很容易亂，到底是為什麼呢？我們先來看看雜亂資料有哪些類型，通常我們都會先將資料做一些定義，像是資料儲存的結構，欄位的設計等等，但通常只要有透過人為操作，或是時間拉的比較長的話，甚至是透過機器產生的數據，都會產生各種資料不完整。舉例來說：

表 2 各類導致雜亂資料的可能原因

雜亂資料類型	情境
系統設計變更	某公司的系統在 xxxx 年某一天更改了幣別格式，從加拿大幣變成美金計價
資料定義變更	幣別名稱差異，像是新台幣就有可能有『新台幣』『台幣』『TWD』『$』等等描述方法，如果用符號表示，也會有跟美金混淆的問題
資料欄位格式差異	數值標準改變，時間、地點的描述名稱都有可能在長時間之下調整，例如：原本的高雄縣在 2010 年更名為高雄市，資料也需對應整理
系統或硬體不穩定	某筆資料只會在網路通暢時寫入資料，但如果網路不夠穩定則會中斷寫入，所以會產生空值問題
業務整合與認知差異	針對相同事物，在文字或是描述上的口語差異
檔案差異	最好用的當屬 excel 或是 csv 等標準格式，但有時候會收到 pdf 格式的檔案，甚至是紙本資料

資料來源：本文整理。

試算表的格式彈性，就像是一個開放表格，通常不會嚴格定義每

一格裡面到底能夠放怎樣的資訊，所以只要各種人為造成、流程改變、或是技術上的限制等等，就會造成資料格式異常，許多狀況都會導致亂資料的情形。

二、資料清洗常用八招

當資料亂掉時，需要進行『資料清洗』的流程。顧名思義，數據清洗就是將數據整理成方便後續利用的整個流程，其中包括像是：異常值的處理、缺失數據的處理、重複數據的處理、降噪程序等等，最終目的通常是產出結構化資料，供後續資料分析或是資料視覺化使用。

圖 7 針對雜亂資料進行『數據整理』幾乎是每個數據專案的必經過程

資料清洗階段，有許多技巧可使用，整理如下表：

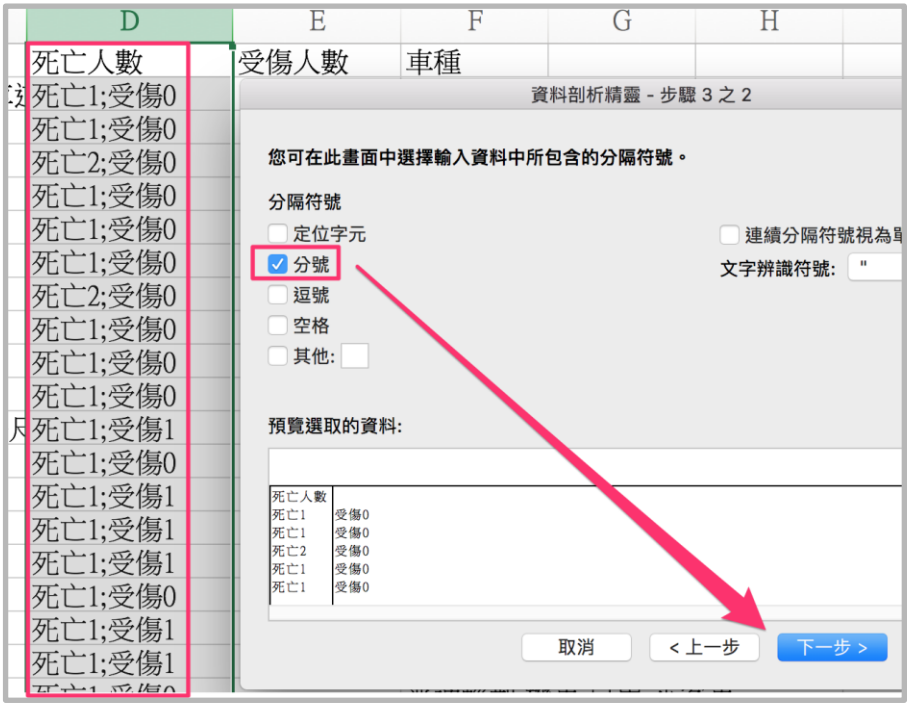
表 3 資料清洗常用八招

編號	項目	說明與情境
（一）	切割（Split）	指的是將單一欄位的內容分散到不同欄位，例如：單一欄位如果紀錄『國家與城市』像是『Taiwan, Taipei』，我們可以透過切割方式將其分成兩個欄位。
（二）	修剪（Trim）	修剪掉無意義的空白資料。
（三）	排序（Sort）與篩選（Filter）	有時取得資料後，發現實際上只需要其中部分來後續利用，這種情境可以用排序與過濾萃取其中重要資料。
（四）	合併（Merge）	跟切割相反，合併指的是取多個欄位的資料，合併為單一欄位，保留其字串完整性。
（五）	格式轉換（Format）	欄位格式的轉換，常見於『日期』與『貨幣』格式的轉換。
（六）	取代（Replace）	字串取代應該是許多人常用的技巧，像是『將男性從 Male 轉換成 1』這種常見情境，來解決數據定義對應問題。
（七）	移除重複列（Remove）	許多資料會出現同樣的資料列，我們需要將其進行整合，並消滅重複的資料列，避免數據重複計算。
（八）	轉置（pivoting ）	許多數據並不是以合適的『欄列』格式呈現，有時需要將其進行轉置作業，將原本的欄換成列，或是將列換成欄。

資料來源：本文整理。

（一）資料切割（Split）

切割、修剪是我們很常使用的處理技巧，主要用來將某個整合在一起的字串切開，因為當初業務上需求所收集的數據，很可能與最終使用方法不同，文字可能會包含前置字元、 結尾字元，或多個空格字元等等，但我們並不需要。



資料來源：本文整理。

圖 8 有時想要的資料會放在同一個欄裡面，就需要進行資料切割（Split）的處理

死亡人數	
死亡1	受傷0
死亡1	受傷0
死亡2	受傷0
死亡1	受傷0
死亡1	受傷0
死亡1	受傷0
死亡2	受傷0
死亡1	受傷0
死亡1	受傷0
死亡1	受傷0

資料來源：本文整理。

圖 9 成功切開成為兩欄，可獨立查或計算死亡與受傷資訊

（二）修剪（TRIM）

除了切割之外，資料有時會包括無意義的空白資料，有可能是人工登打失誤導致，透過修剪可移除文字的多餘空格，僅保留獨立一個空格。但如果要完全移除空格的話，則可以改用 SUBSTITUTE 函數，將空格完整取代掉。

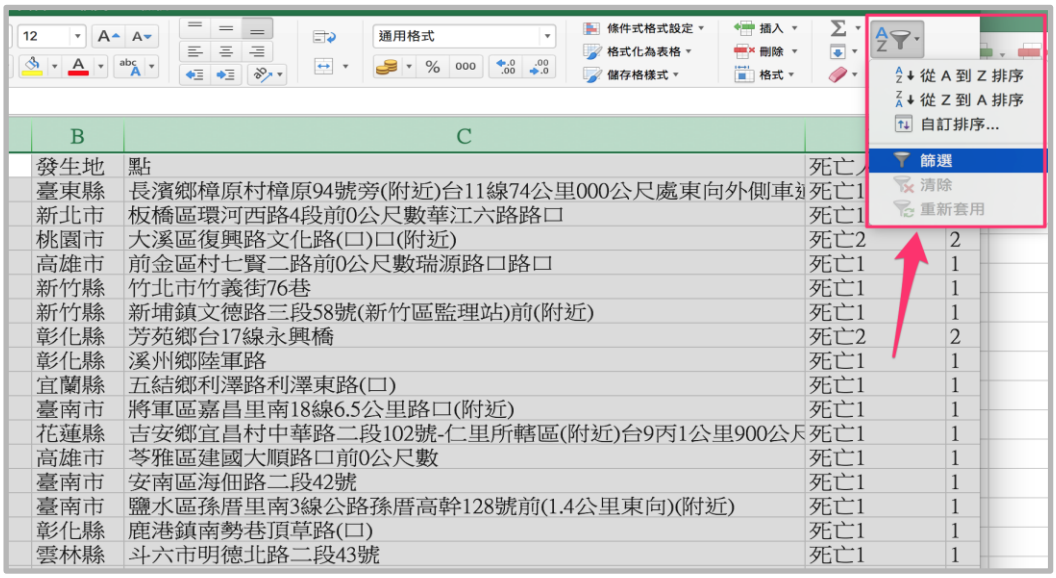
A	B	C
臺東縣　長濱鄉	臺東縣 長濱鄉	臺東縣長濱鄉
新北市　板橋區	新北市 板橋區	新北市板橋區
桃園市 大溪區	桃園市 大溪區	桃園市大溪區
高雄市　　前金區	高雄市 前金區	高雄市前金區

資料來源：本文整理。

圖 10 我們常常需要去除掉多餘的資料空白，處理成為此圖 C 欄的乾淨格式

（三）排序（Sort）與篩選（Filter）

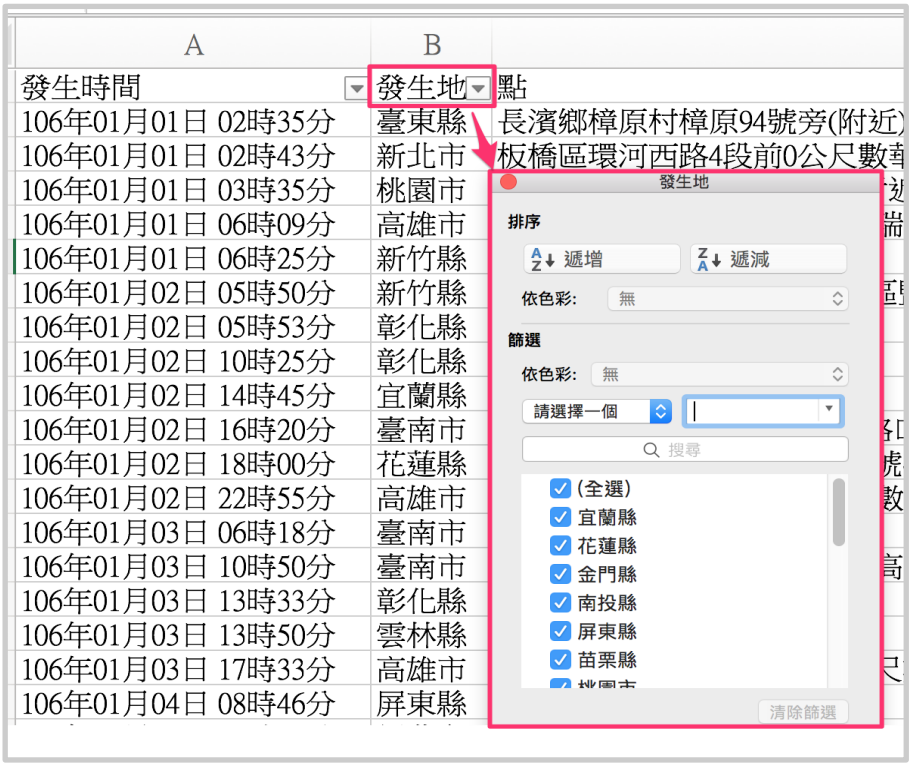
排序與過濾是資料取值常用的技巧，例如：我們只想要留下特定數字區間、時間區間，或是只選定特定類別資料等，都很適合使用排序與過濾技巧。



B	C		
發生地	點	死亡人	
臺東縣	長濱鄉樟原村樟原94號旁(附近)台11線74公里000公尺處東向外側車道	死亡1	
新北市	板橋區環河西路4段前0公尺數華江六路路口	死亡1	
桃園市	大溪區復興路文化路(口)口(附近)	死亡2	2
高雄市	前金區村七賢二路前0公尺數瑞源路口路口	死亡1	1
新竹縣	竹北市竹義街76巷	死亡1	1
新竹縣	新埔鎮文德路三段58號(新竹區監理站)前(附近)	死亡1	1
彰化縣	芳苑鄉台17線永興橋	死亡2	2
彰化縣	溪州鄉陸軍路	死亡1	1
宜蘭縣	五結鄉利澤路利澤東路(口)	死亡1	1
臺南市	將軍區嘉昌里南18線6.5公里路口(附近)	死亡1	1
花蓮縣	吉安鄉宜昌村中華路二段102號-仁里所轄區(附近)台9丙1公里900公尺	死亡1	1
高雄市	苓雅區建國大順路口前0公尺數	死亡1	1
臺南市	安南區海佃路二段42號	死亡1	1
臺南市	鹽水區孫厝里南3線公路孫厝高幹128號前(1.4公里東向)(附近)	死亡1	1
彰化縣	鹿港鎮南勢巷頂草路(口)	死亡1	1
雲林縣	斗六市明德北路二段43號	死亡1	1

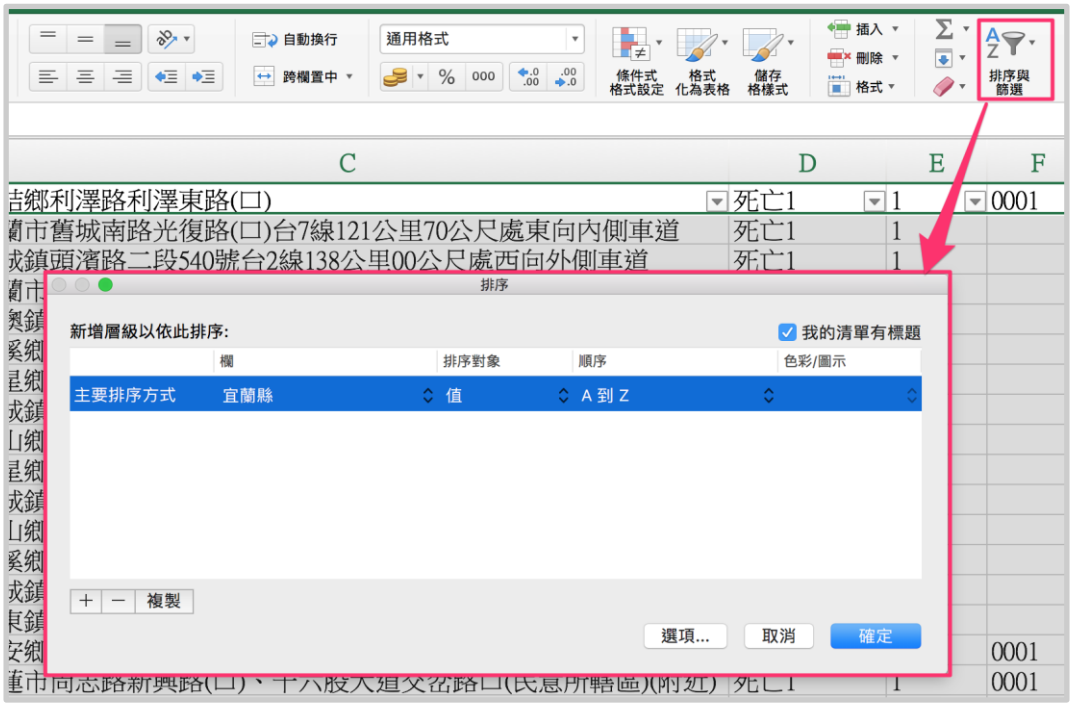
資料來源：本文整理。

圖 11 Excel 畫面右上角有排序與篩選的功能



資料來源：本文整理。

圖 12 開啟篩選後，許多欄位都可勾選瀏覽特定內容資料

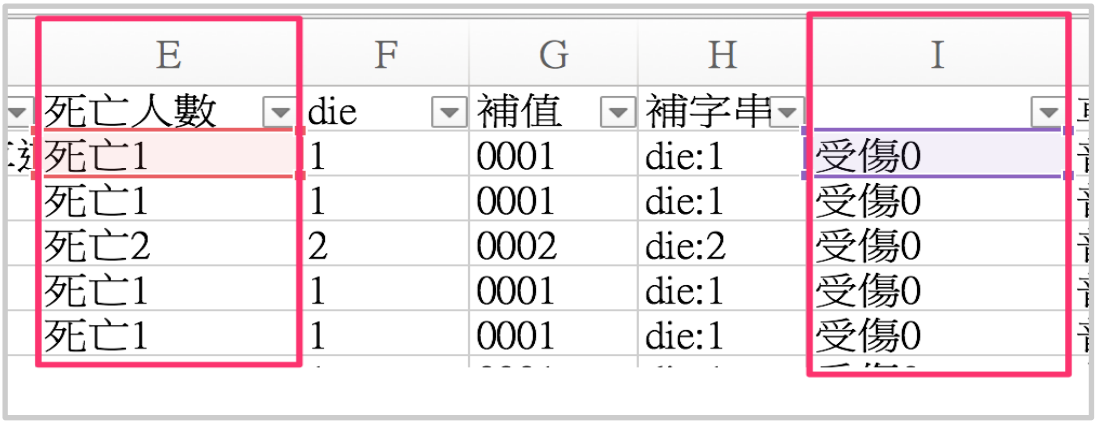


資料來源：本文整理。

圖 13 也可以開啟排序設定窗格，來設定排序方式，可設定多個排序並按照階層順序

（四）合併（Merge）

我們可能希望取得多個欄位的資料，合併為單一欄位，來保留字串完整性，例如：我們可能會想要重新組合姓名、地區的顯示資訊等等，讓該資訊能夠有更佳的應用。



	E	F	G	H	I
	死亡人數	die	補值	補字串	
	死亡1	1	0001	die:1	受傷0
	死亡1	1	0001	die:1	受傷0
	死亡2	2	0002	die:2	受傷0
	死亡1	1	0001	die:1	受傷0
	死亡1	1	0001	die:1	受傷0

資料來源：本文整理。

圖 14 本練習會用到的欄位資訊（E&I 欄資訊）

在組合的部分，我們可以使用 CONCATENATE 函數，概念很簡單，透過欄位的組合，並在欄位之間塞入想要組合的中間連接字串。

	A	B
	發生時間	縣市摘要
	106年01月01日 02時43分	臺東縣(死亡1,受傷0)
	106年01月01日 02時43分	新北市(死亡1,受傷0)
	106年01月01日 03時35分	桃園市(死亡2,受傷0)
	106年01月01日 06時09分	高雄市(死亡1,受傷0)
	106年01月01日 06時25分	新竹縣(死亡1,受傷0)
	106年01月02日 05時50分	新竹縣(死亡1,受傷0)
	106年01月02日 05時53分	彰化縣(死亡2,受傷0)
	106年01月02日 10時25分	彰化縣(死亡1,受傷0)
	106年01月02日 14時45分	宜蘭縣(死亡1,受傷0)
	106年01月02日 16時20分	臺南市(死亡1,受傷0)
	106年01月02日 18時00分	花蓮縣(死亡1,受傷1)
	106年01月02日 22時55分	高雄市(死亡1,受傷0)
	106年01月03日 06時18分	臺南市(死亡1,受傷1)

資料來源：本文整理。

圖 15 可透過多個欄位資訊合併出整合性摘要欄位

（五）格式轉換（Format）

欄位資料有時需要進行轉換，常見的像是『文字轉日期』、『民國轉西元』等等，不同情境之下我們可能會需要轉換資料顯示的內容，例如過去的縣市名稱跟現在的可能已經不同，或是不同國家對於幣別、時間的解釋格式不同時，需要用到格式轉換技巧。

	A	B	
1	發生時間	縣市摘要	發
2	106年01月01日 02時43分	臺東縣(死亡1,受傷0)	臺
3	106年01月01日 02時43分	新北市(死亡1,受傷0)	新
4	106年01月01日 03時35分	桃園市(死亡2,受傷0)	桃
5	106年01月01日 06時09分	高雄市(死亡1,受傷0)	高
6	106年01月01日 06時25分	新竹縣(死亡1,受傷0)	新
7	106年01月02日 05時50分	新竹縣(死亡1,受傷0)	新
8	106年01月02日 05時53分	彰化縣(死亡2,受傷0)	彰
9	106年01月02日 10時25分	彰化縣(死亡1,受傷0)	彰
10	106年01月02日 14時45分	宜蘭縣(死亡1,受傷0)	宜
11	106年01月02日 16時20分	臺南市(死亡1,受傷0)	臺
12	106年01月02日 18時00分	花蓮縣(死亡1,受傷1)	花
13	106年01月02日 22時55分	高雄市(死亡1,受傷0)	高
14	106年01月03日 06時18分	臺南市(死亡1,受傷1)	臺
15	106年01月03日 10時50分	臺南市(死亡1,受傷1)	臺

資料來源：本文整理。

圖 16 原始資料常常出現民國的時間區段，不好使用

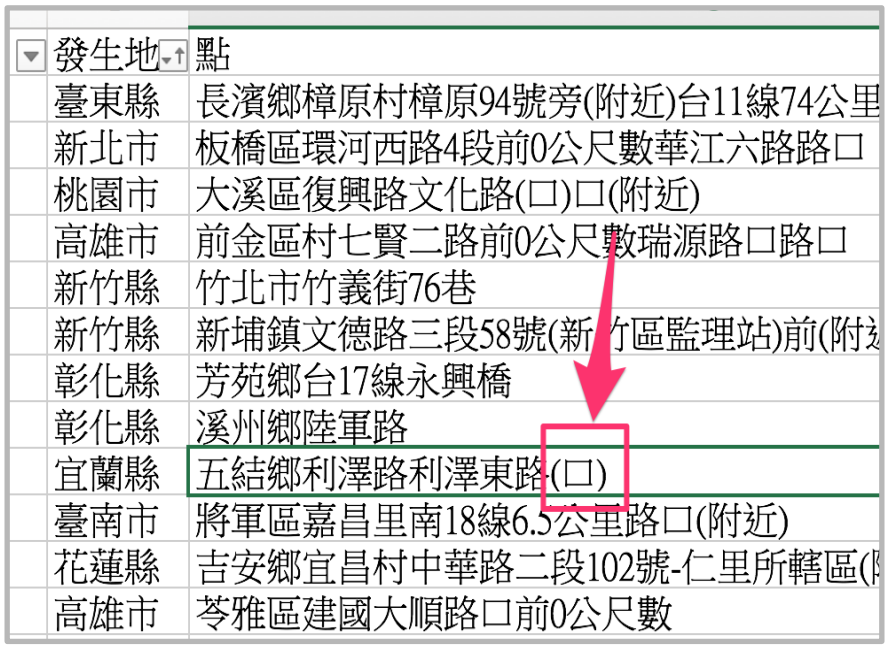
發生日期	發生時間(轉換)	發生時間	組合時間
106年01月01日	2017/01/01	2:43	2017/01/01 02:43:00
106年01月01日	2017/01/01	3:43	2017/01/01 03:43:00
106年01月01日	2017/01/01	4:43	2017/01/01 04:43:00
106年01月01日	2017/01/01	5:43	2017/01/01 05:43:00
106年01月01日	2017/01/01	6:43	2017/01/01 06:43:00
106年01月02日	2017/01/02	7:43	2017/01/02 07:43:00
106年01月02日	2017/01/02	8:43	2017/01/02 08:43:00
106年01月02日	2017/01/02	9:43	2017/01/02 09:43:00
106年01月02日	2017/01/02	10:43	2017/01/02 10:43:00
106年01月02日	2017/01/02	11:43	2017/01/02 11:43:00
106年01月02日	2017/01/02	12:43	2017/01/02 12:43:00
106年01月02日	2017/01/02	13:43	2017/01/02 13:43:00
106年01月03日	2017/01/03	14:43	2017/01/03 14:43:00

資料來源：本文整理。

圖 17 透過時間或是地理資料格式的轉換，較能夠通用於更多軟體

（六）取代（Replace）

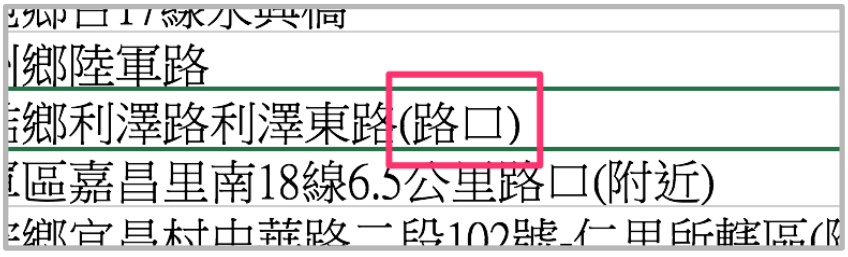
文字的定義與描述方式，有時會因為人的不同或是時間的不同而改變，有時我們會需要將資料進行一致化的處理，像是定義同義字，或是將名稱進行統一，方便供後續正確計算，這時就可能會需要使用到取代的技巧。



發生地	點
臺東縣	長濱鄉樟原村樟原94號旁(附近)台11線74公里
新北市	板橋區環河西路4段前0公尺數華江六路路口
桃園市	大溪區復興路文化路(口)口(附近)
高雄市	前金區村七賢二路前0公尺數瑞源路口路口
新竹縣	竹北市竹義街76巷
新竹縣	新埔鎮文德路三段58號(新竹區監理站)前(附近
彰化縣	芳苑鄉台17線永興橋
彰化縣	溪州鄉陸軍路
宜蘭縣	五結鄉利澤路利澤東路(口)
臺南市	將軍區嘉昌里南18線6.5公里路口(附近)
花蓮縣	吉安鄉宜昌村中華路二段102號-仁里所轄區(附
高雄市	苓雅區建國大順路口前0公尺數

資料來源：本文整理。

圖 18 由於『口』不好理解，想將其轉換為『路口』



資料來源：本文整理。

圖 19 簡單又好用的取代任務，能夠大大提升資料的理解性

（七）移除重複列（Remove）

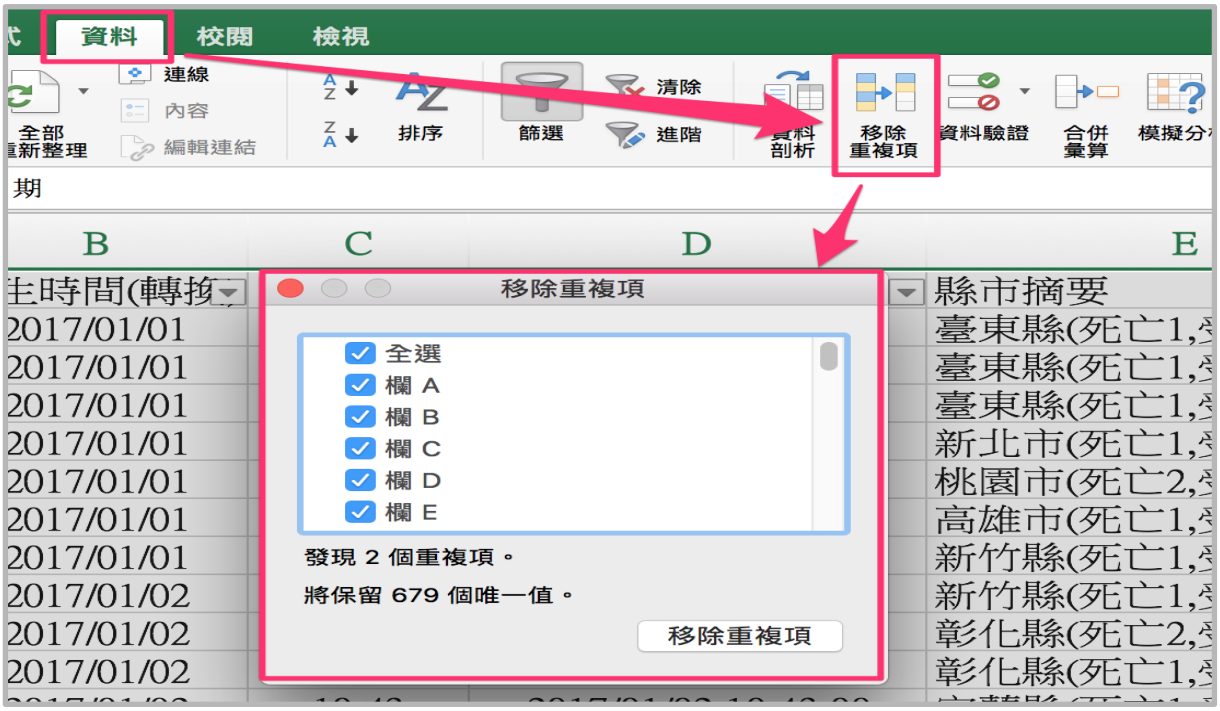
有的資料會出現重複列的狀況，有可能是重複寫入，或是人工不小心重複 key-in 了，但這類型資料有可能會導致後續分析的錯誤，我們需要自動找到這些重複的列，並將其移除。

發生日期	發生時間(轉換	發生時間	組合時間	縣市摘要
106年01月01日	2017/01/01	2:43	2017/01/01 02:43:00	臺東縣(死亡1,受傷0)
106年01月01日	2017/01/01	2:43	2017/01/01 02:43:00	臺東縣(死亡1,受傷0)
106年01月01日	2017/01/01	2:43	2017/01/01 02:43:00	臺東縣(死亡1,受傷0)
106年01月01日	2017/01/01	3:43	2017/01/01 03:43:00	新北市(死亡1,受傷0)
106年01月01日	2017/01/01	4:43	2017/01/01 04:43:00	桃園市(死亡2,受傷0)
106年01月01日	2017/01/01	5:43	2017/01/01 05:43:00	高雄市(死亡1,受傷0)
106年01月01日	2017/01/01	6:43	2017/01/01 06:43:00	新竹縣(死亡1,受傷0)

資料來源：本文整理。

圖 20 資料可能會出現重複的資料

這時候我們可以開啟 Excel『資料』頁籤，選擇『移除重複列』的功能，送出後，即可發現重複列已經被清除了。



資料來源：本文整理。

圖 21 Excel 貼心的『移除重複項』功能

（八）轉置（Transposing/pivoting ）

我們透過開放資料（Open Data）取到的資料五花八門，有時候對方在儲存資料的時候，不一定會考慮後續應用的用途，而是用人們很習慣的閱讀方式進行儲存，例如以下的格式就很常見，日期是往右延伸的，雖然人類眼睛容易閱讀，但是不利於後續資料視覺化所用。

A	B	C	D	E	F	G
死亡人數統計	2017/1/1	2017/1/2	2017/1/3	2017/1/4	2017/1/5	2017/1/6
臺東縣	1	1	2	0	0	1
新竹縣	2	0	0	1	1	1
彰化縣	0	1	1	1	1	2

資料來源：本文整理。

圖 22 常見的資料儲存格式

這時候我們可以透過資料轉置（Transposing/pivoting ）功能，轉換成方便視覺化之格式，將『城市』變更為欄位名稱，而『日期』則依照序列的方式往下延伸。

死亡人數統計	2017/1/1	2017/1/2	2017/1/3	2017/1/4	2017/1/5	2017/1/6
臺東縣	1	1	2	0	0	1
新竹縣	2	0	0	1	1	1
彰化縣	0	1	1	1	1	2

死亡人數統計	臺東縣	新竹縣	彰化縣
2017/1/1	1	2	0
2017/1/2	1	0	1
2017/1/3	2	0	1
2017/1/4	0	1	1
2017/1/5	0	1	1
2017/1/6	1	1	2

資料來源：本文整理。

圖 23 資料轉置可將資料調整為更適合視覺化軟體使用的格式