第七章

偵測非常態分布資料集中的異常值

在第 6 章中，我介紹了三種當資料呈常態分布時，用來把異常值視覺化的方法。但是，在實際應用中，資料往往不一定符合常態分布，如果一律使用以假設資料呈常態分布為前提的分析方法，反而可能會導致錯誤的結論或誤導決策者。因此，第 4 章介紹的探測資料分布的技巧就顯得非常重要。

在這個章節裡，我會介紹三種方法，讓你在面對尚未標準化的資料時，仍然可以找出離群值並且以視覺化表示。這三種方法分別是：平均絕對離差、Tukey's Fences 方法，以及改良版的 z 分數檢定。

認識中位數絕對離差（MAD：Median Absolute Deviation）

「中位數絕對離差（MAD）」是一種用於衡量資料集的分散程度或變異性的統計量。其計算方式是：首先計算每個資料點與中位數之間的絕對離差（即資料點減去中位數後，再取絕對值），然後再計算這些「絕對離差」的中位數即為 MAD 值。數學上 MAD 的計算公式，可表示如下：

$$\text{MAD} = \text{Median}\left(\left|X_i - \text{Median}\right|\right)$$

的特性，因此，以 MAD 法檢測異常值非常有效。此外，MAD 法正是第 4 章曾介紹過的，所謂的無母數方法之一。要知道，無母數模型不限定何種資料分布，也就是無須假設特定的函數形式，因此，本身就具有較高的靈活性。

總而言之，MAD 法不但易於實作，還是一種功能強大且應用廣泛的檢測方法。

如何在 Tableau 中實作 MAD

在這個小節裡，我將帶你認識如何在 Tableau 中實作 MAD。首先請連接到「範例 – 超級市場」範例資料集，然後建立「銷售總額 vs. 月份」的折線圖。此外，我特別喜歡使用雙軸折線圖，也就是重複添加相同的資料欄位到「列」架上，並且將其中一組的標記外型設為圓形（詳參第 6 章使用「標準差」來偵測異常值章節），一系列調整視覺化的操作後，得到如圖 7-1 所示的雙軸折線圖。這樣的設定有助於後續的條件格式化操作，把異常值視覺化。

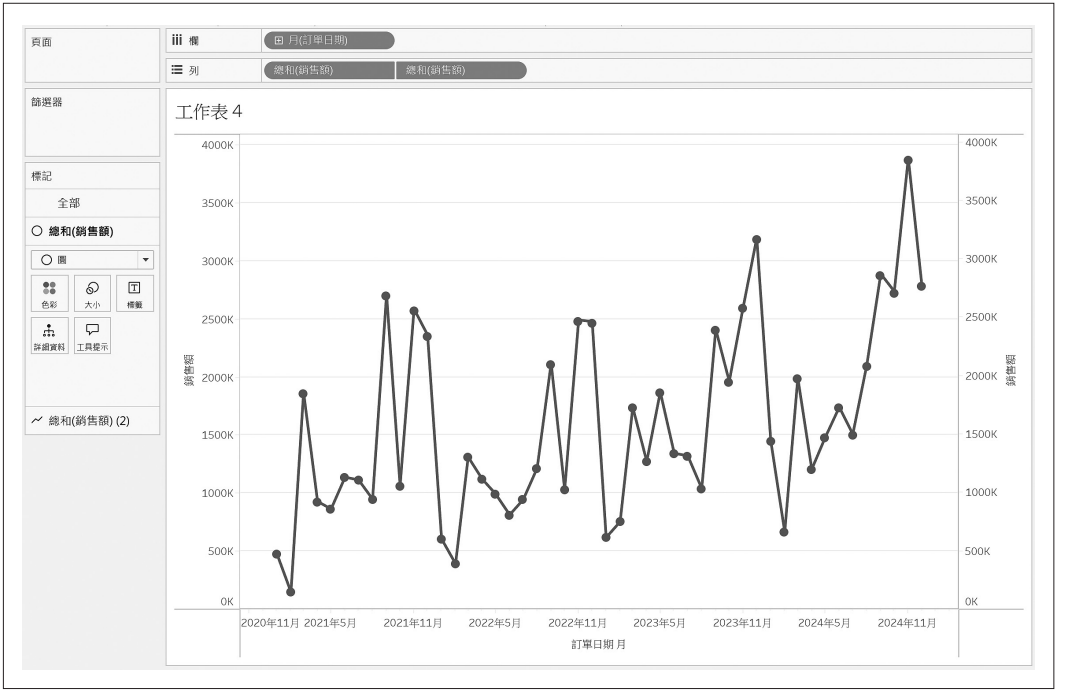

圖 7-1　銷售總額 vs. 月份雙軸折線圖

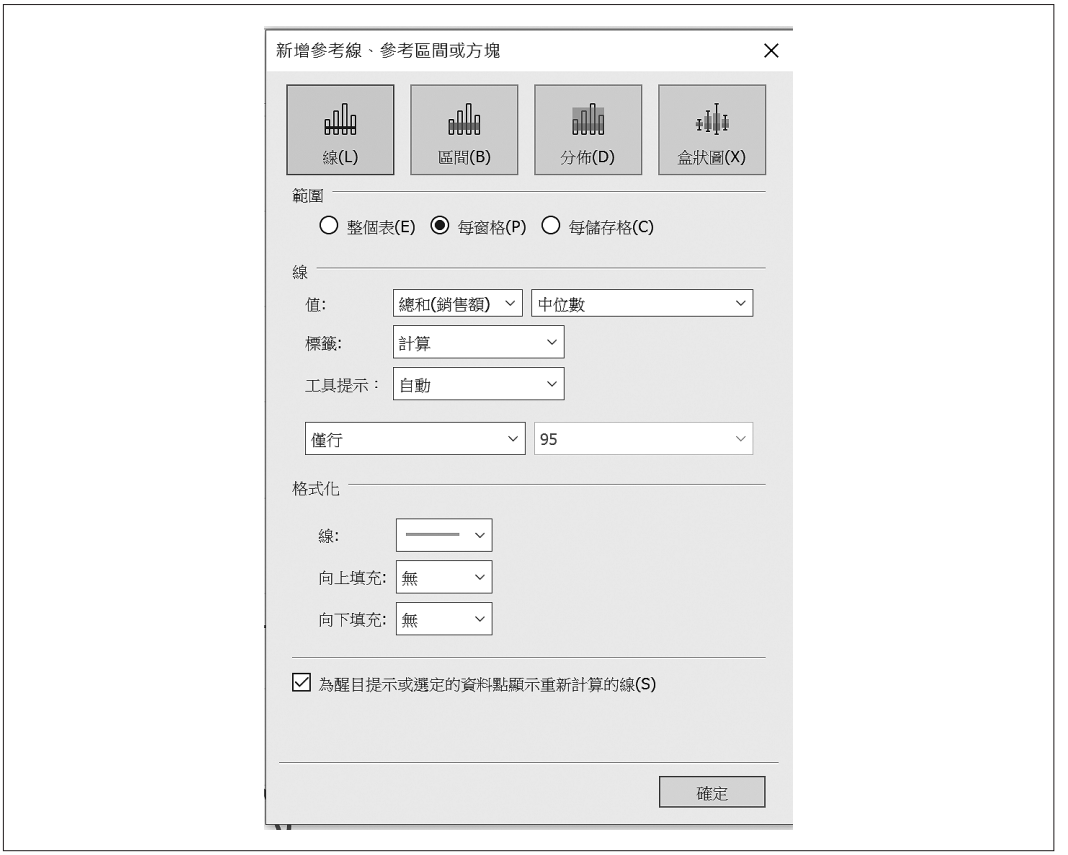

圖 7-3　「新增參考線、參考區間或方塊」的對話視窗

現在，你可以開始建立進行 MAD 法所需要的「計算欄位」了。回顧一下本章開頭的解說（請參見第 119 頁的「認識中位數絕對離差」），你會需要建立三個「計算欄位」：MAD 值、下界和上界。首先，我先建立 MAD 統計量的「計算欄位」（見圖 7-4）。

```
WINDOW_MEDIAN(ABS(SUM([銷售額])-WINDOW_MEDIAN(SUM([銷售額]))))
```

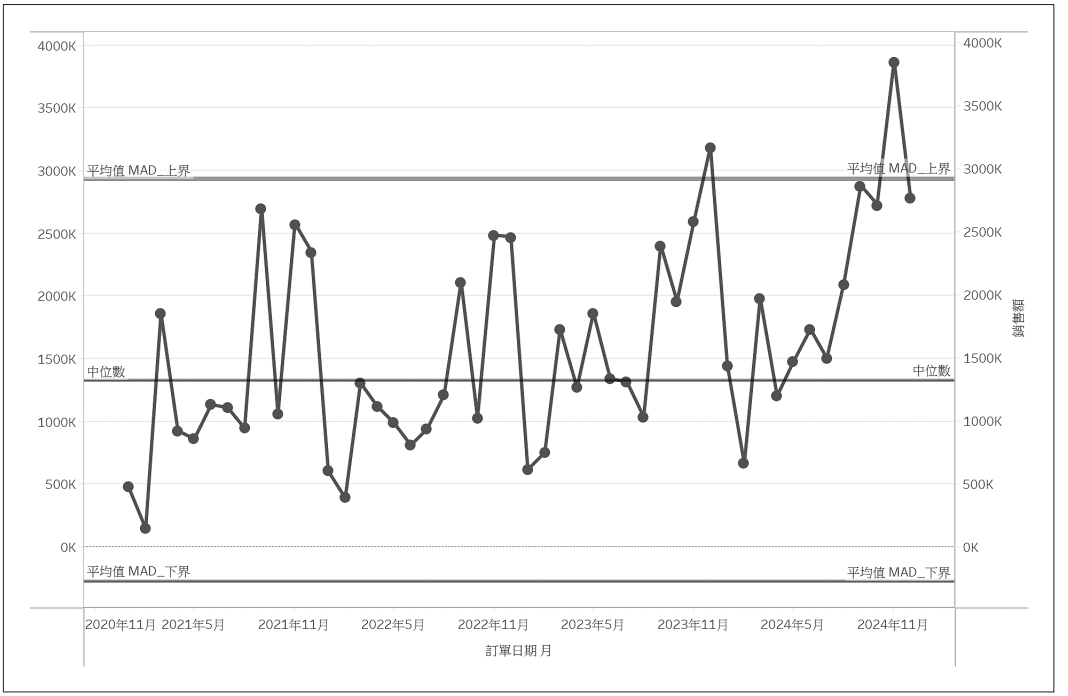

圖 7-9　二條參考線都已經套用在視圖上

你可以看到，資料中有兩個值超過了上界。對於這些資料來說，就可以視它們為離群值或是偏離常態的情況，並且進行更深入的分析，研究看看為什麼這幾個月份會有這樣的情形！

認識改良版的 z 分數方法

在第 6 章中，我介紹了一種使用「z 分數」來偵測離群值的方法。當時的公式是把「資料點的數值」減去「資料集平均數」，然後再除以「標準差」。從數學角度來看，這個公式可表示如下：

$$z = (x - \mu) \div \sigma$$

然而，由於標準的 z 分數源自「平均數」和「標準差」，所以待檢測的資料本身就必須已經符合常態分布，才能使用這個公式計算 z 分數。在這個小節，你可以嘗試調整這個

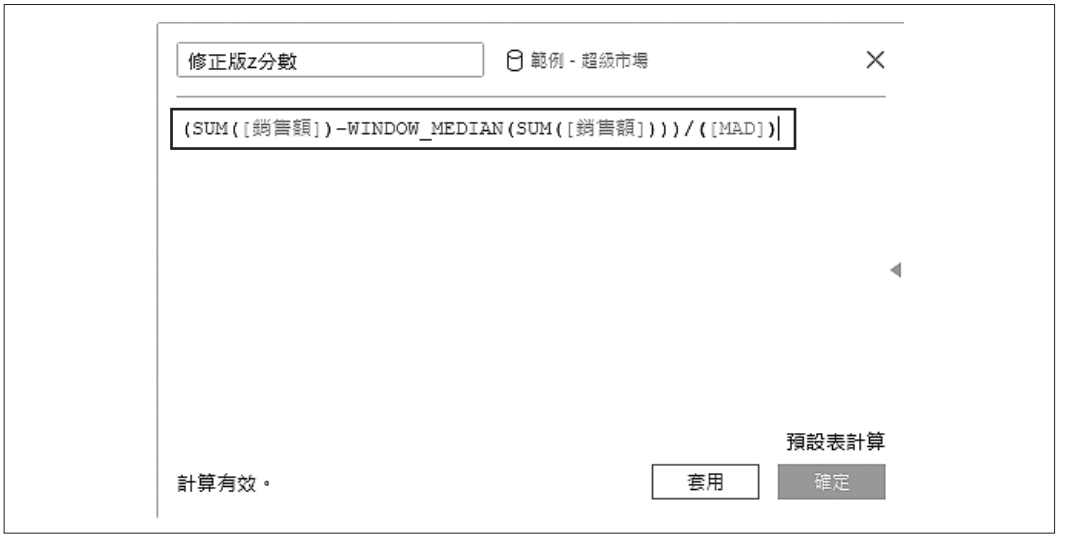

圖 7-12　計算改良版 z 分數

下一步，就是將「改良版 z 分數」的「計算欄位」內所含的計算公式套用到視圖上。操作很簡單，若要把這個新的度量值加入視圖中，只需把被我命名為「改良版 z 分數」的欄位資料，拖曳到「標記」架上的「標籤」屬性上，如圖 7-13 所示。

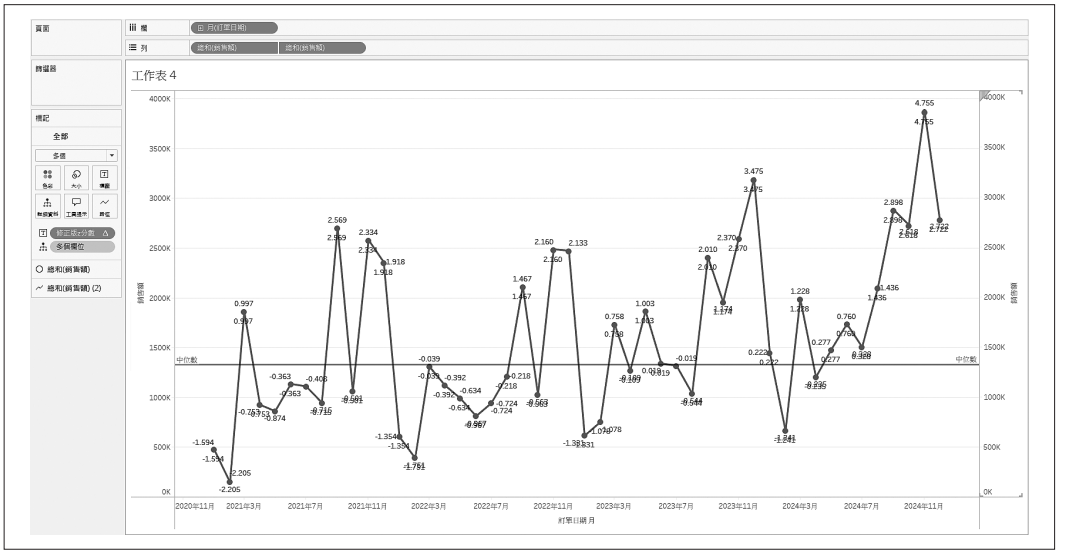

圖 7-13　拖曳「計算改良版 z 分數」欄位資料到「標記」架上的「標籤」屬性上

如何在 Tableau 中實作叢集分析

首先，請連接至「範例 – 超級市場」範例資料集，你必須先建立一個顯示顧客平均折扣的點狀圖。因此，在新的工作表上，請將「折扣」的資料欄位，拖曳到「欄」架上，然後在「欄」架上「總和（折扣）」的綠色膠囊按鈕上，按下滑鼠右鍵將「度量（總和）」改為「度量（平均值）」。接著，將「客戶名稱」的資料欄位拖曳到「標記」架上的「詳細資料」屬性上，並且自「標記」架上的下拉選單中把「自動」改為「圓」。此時，你應該會得到一個類似圖 11-7 的點狀圖。

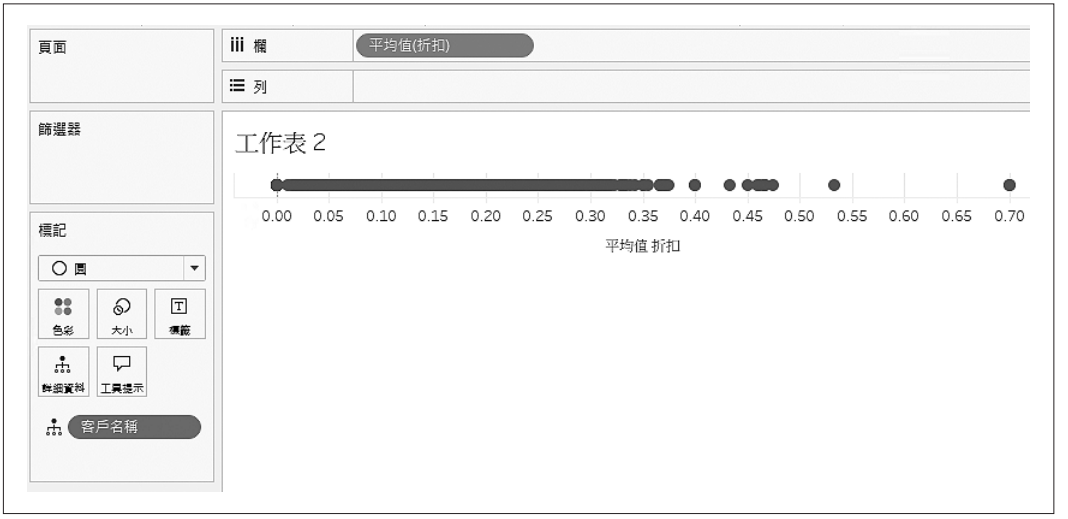

圖 11-7　平均折扣 vs. 客戶名稱的點狀圖

接下來，你需要對點狀圖進行「抖動」處理。想要實現「抖動」運算，可以在「列」架空白處點擊滑鼠右鍵，點選「新建計算」，然後在「列」架上空白膠囊中的輸入欄位直接輸入 random()，按 Enter。這一系列的操作，可以使視圖中的每個「標記」在 y 軸上隨機擺放產生如圖 11-8 的效果。這樣做的好處是，將游標懸停在任何一個「標記」上，就可以直接在視圖中看到它們的內容。此外，上述操作也可以透過先前使用過的「計算欄位」來實作。

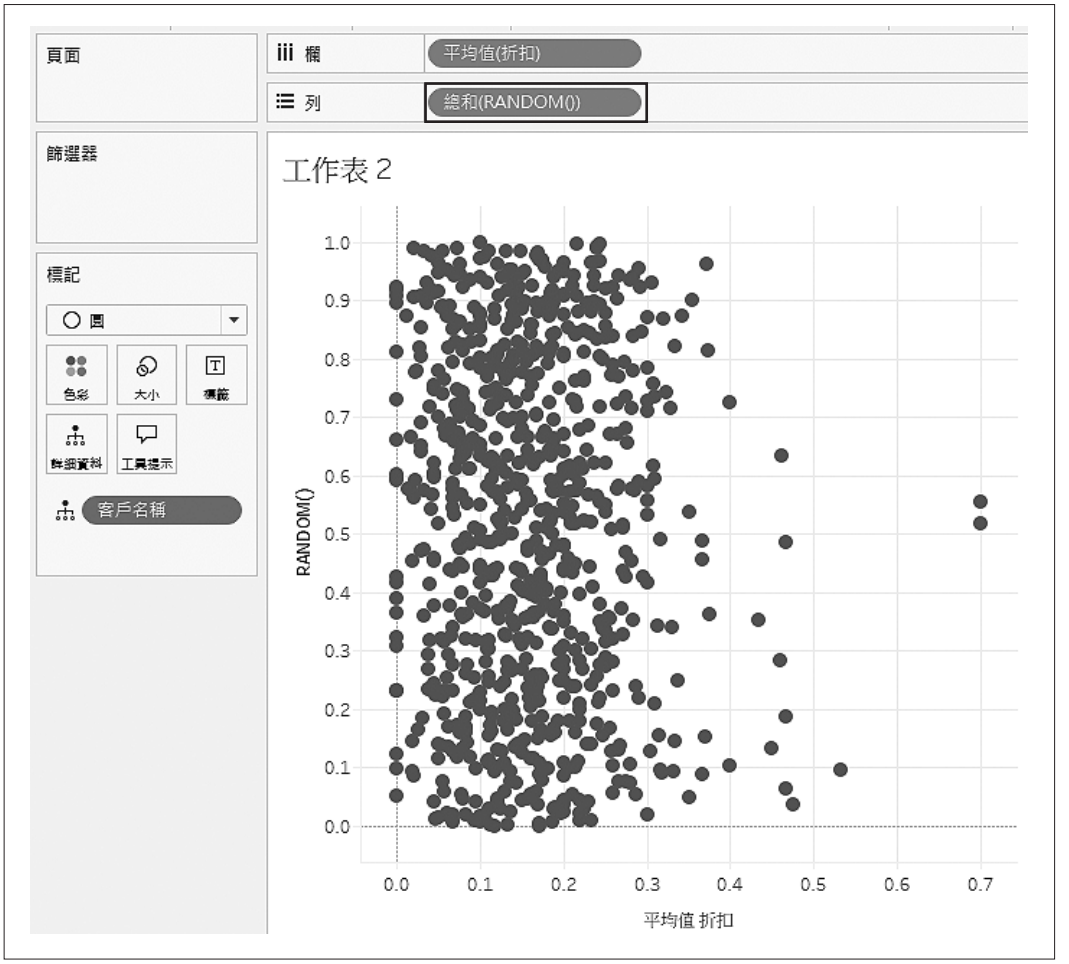

圖 11-8　將平均折扣 vs. 客戶名稱的點狀圖進行抖動處理

從這裡開始，實作叢集分析模型變得非常簡單。接著切換到「分析」窗格，再將「叢集」往視圖中拖曳，不要放開滑鼠，一路拖曳到彈出的小視窗中的「叢集」上，如圖 11-9 所示。

此時，工作表上會彈出「叢集」視窗，顯示當前模型中的變數，並且在「叢集數」對話框中，你還可以輸入期望的 k 值（預設值為 5），如圖 11-10 所示。除此之外，在「標記」架上的「色彩」屬性中將會出現一個名為「叢集（1）」的新欄位資料。

在預設情況下，Tableau 叢集分析的視覺化效果會根據顏色來區分。不過，你也可以將其改為以形狀來區分。如果想要達成那樣的效果，請從「標記」架上的下拉選單中自「圓」改選為「圖形」，然後點擊「標記」架上「叢集（1）」欄位資料前的色彩符號，再點選「圖形」，產生的效果如圖 11-11 所示。

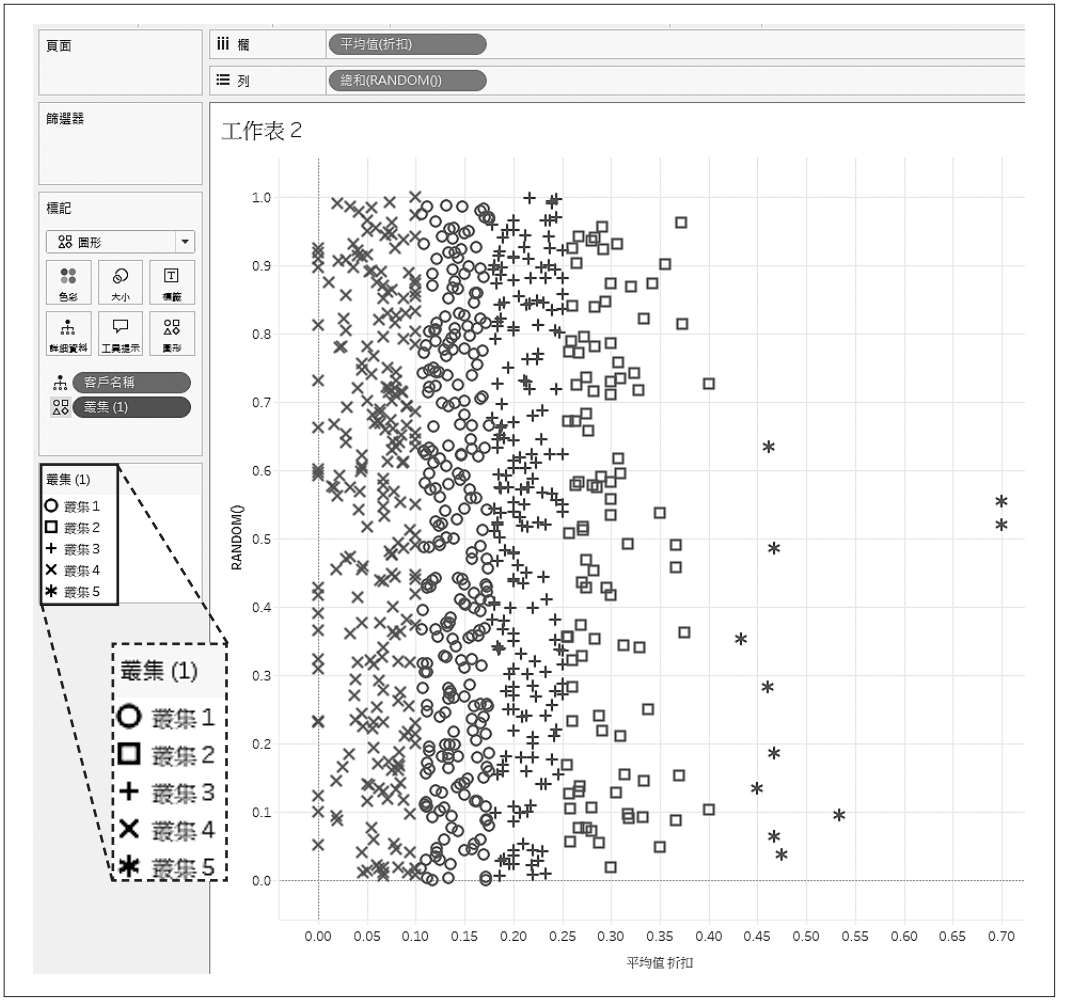

圖 11-11　以「形狀」來表現叢集分析可是化的效果