資料分析與學習基石 HW 3-2

F74062206 王聖中

> Q: Analyze the data (statistics. correlation…), and you should explain the details of analysis. (data preprocessing. improvements...)

A:

本次使用之資料集為 Tarvel Review Ratings Data Set。

data processing：

(1) 首先將檔案讀入後，呼叫 df.info() 觀察，發現幾件事：第一，Category 12 和 Category 24 有缺值。第二，多了一個 column，名為 Unnamed: 25，且有兩個 row 有值。第三，Category 11 的 Dtype 為 object，和其他 Category 不一樣。

sol: 於是先去看 Unnamed: 25 究竟是哪裡有值，發現是在 User 1348 和 User 2713 兩處，再觀察那兩個 row，發現資料似乎移位了，於是將其調正（詳細可以看 code 註解），同時處理兩格看起來較有問題的值，分別是 User 1348 在 Category 23 的值 (.26)，以及 User 2713 在 Category 11 的值 (2\t2. ，同時也是造成 Category 11 的 Dtype 變為 object 的一格。)，透過表格分布，校正值則取其前一位 User 在該 column 的值，看起來也較為合理。最後就可以將 Unnamed: 25 這個 column 丟棄，並把 Category 11 的 Dtype 設為 float64。

data analysis：

(1) 計算各 Category 的 0 分評論有幾個。

想法：在 Data Set Information 有提到 “Google user rating ranges from 1 to 5 and average user rating per category is calculated.”，因此若某 User 在某 Category 的值為 0 分則代表該 User 並沒有在此 Category 的 attractions 進行任何評分。

實作：在 code 將值算出後以 matplotlib 畫出長條圖，完整實作請見 .ipynb 檔。最後的長條圖為了容易觀看與分析，y 值由低到高顯示。

結論：

19 bakeries

18 gyms

20 beauty & spas

完整的內容可以去 code 看註解，在報告取前三個來討論分析。

如果思考在去歐洲的旅遊團上的話，或許代表著安排的行程通常不太會包含這些類型的景點，我猜這部分大多數的評分是由自助旅行的觀光客給的。

當然也可能代表這些景點本身就不太容易被注意到，也就不容易排進行程，如果是想開店來坑旅遊觀光客的話(?)，可能這三種店就不太適合開。

不過換個角度想，至少這些種類的景點應該是不太會到雷人的程度，因為我自己覺得有時候沒有去進行 Google 評分，代表沒有爛到想找地方批判，但也沒有讓人驚艷到想大推就是了 XD。（不過這個推斷要參考一下下面的指標，其實這三個類別綜合 (1)(2) 兩個指標來看，會發現並非如此 :P，但我認為這樣的敘述在一般情況下是沒有到太大的問題的。）

(2) 計算各 Category 的平均評分高低。

想法：若要評斷 Category 的好壞，當然少不了去看平均評分的數值為多少。

實作：基本上和上一點一樣，把平均值算出來後以長條圖呈現，不過這邊有做一個改良是：算平均的時候，把該 Category 裡的 0 當成 NaN，不計算至 mean 裡面。因為這部分應考量的是實際有在此 Category 評分的用戶意見，若為 0 分則如上所說為未在此 Category 評分，不應計算進去。而最後 y 值一樣由低到高排列。

結論：

先看前五高－－

malls
restaurants
theatres
museums
pubs/bars

再看倒數五名——

bakeries
beauty & spas
cafes
swimming pools
gyms

同樣的，完整的列表可至 code 註解看，這邊抓前後五名來討論分析。

其實不難想像，前五高的大多都屬於旅遊團或自助旅行時會想安排的行程，或許在商機蓬勃的情況下，這些類型的服務不斷精進，因此獲得了使用者平均較高的評分。

而倒數五名則代表給旅行者較不佳的體驗，配合 (1) 的指標來看，則可得出幾個類別，如 bakeries, gyms，或許會在旅遊界被當成遊歐的一大地雷。雖然在 (1) 有提到可能不適合開 bakeries, gyms 這類的店，但在這項指標下，換個角度想，其實要成為這些類別的 best 可能是相對容易的，因此商人們若能有好的

策略，就也能在歐洲觀光界殺出一條生路。

> Q: Define a reasonable problem (classification. regression. clustering...) and predict the results

A:　我選擇的是 clustering 的問題，目的是將這些 Users 分群，從某種意義上來說算是分出較適合一起去旅遊的的群體吧 XD。

有點不太確定這邊的 predict the result 要寫什麼，不過基本上我的問題是將 data set 中所有 users 進行分群，因此並沒有所謂 test data 要去預測結果。

> Q: Explain how you improved your results step-by-step (Original result -> Reasons -> Your approaches -> Improvement)

A:

使用 scikit-learn 裡的 KMeans 進行分群。

Original result：

一開始使用的是 Category 1 到 Category 24，也就是利用所有類別去進行 KMeans 的分群，然後群數設定為 16（其實我的想法很單純，有 5456 個 user，那就取一個不大也不小的因數好了，避免時間複雜度太高。5456 = 16 * 341，好，16 就決定是你了！）。

最後將 User ID 跟所在的群利用 matplotlib 畫成散佈圖，並且印出兩項資料，一、每群所占人數，由小到大排序；二、每群前 15 位 User 的 ID，透過這幾個來看出大致上分群的分佈。

Reasons：

還記得問題是「分出較適合一起去旅遊的群體」嗎，旅遊時通常要找的是興趣類似的旅伴。最初的結果的確是有符合這樣的結果，但是再仔細看一下景點類型，會發現其實這份資料將他分得很細，但實際上的考量若也這麼細會有點奇怪。舉例來說：如果兩個人都給 juice bars 這個 Category 很高分，可是其他都沒什麼評分，或是評分都不高，那在最初的情況，基本上他們會分到同一群，但是那樣的分類似乎變成是，好像他們的旅遊就是瘋狂去 juice bars，聽起來超怪。可是若能找到另一個人，或許 juice bars 這個類型沒有到 5 分，但也有 4 分左右，而他在 burger/pizza shops 這個 Category 評分很高，這樣或許一起出去時就能推薦好吃的漢堡或披薩店，而另一位就能提供好喝的果汁吧，似乎這樣出去旅遊才會比較快樂。也就是說，將「興趣類似」的定義變得比較廣了，可能都很愛吃，或是都很愛去娛樂場所（例如：malls 或 theatres 這種類型）……之類的，這樣的分群方式對問題本身來說才是較佳的解。

Your approaches：

於是根據上述，將類型靠近的放在一起，最後結果如下：

food：restaurants、burger/pizza shops、juice bars、bakeries、cafes

accommodation：resorts、hotels/other lodgings

humanities：churches、museums、art galleries、monuments

nightclub：pubs/bars、dance clubs

relax：theatres、malls、local services、beauty & spas

nature：beaches、parks、zoo、view points、gardens

sport：swimming pools、gyms

從 24 個類型改成 7 大類型。

Improvement：

剩下的方式與流程跟原本基本上一樣，最後可以看出各分群裡的人數全距減少了。這樣的分群方式，較不會有的人找不太到旅伴（原本的分群，人數最少的群為 108 位。而現在的分群，人數最少的群至少還有 203 位）。