FDA hw4

F74046030 李昕

1. 前情提要 :

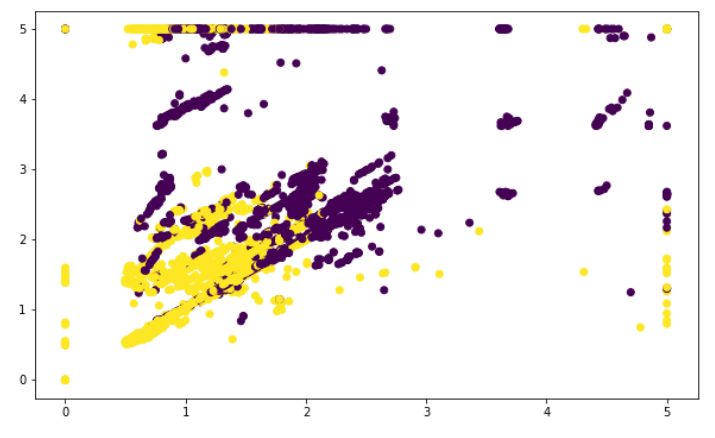
這次使用的資料是對於旅遊景點的評論分數資料。但因為給出分數的可能原因相當多，一個分數可以代表的意思也就相對多了起來。所以這次作業在對於分數討論的時候，會以高分代表有去過、低分代表沒去過這樣的前提和設定來進行討論。如果出現沒辦法以這種設定來解釋的狀況，才會利用其他的角度來嘗試說明。

另外報告使用的結果是某次執行完整理的資料。因為寫完報告之後有重跑過程式，最後上傳的程式碼和整理出來的資料的cluster順序會有不一樣。但除了人數會有浮動以及cluster的順序編號會不同之外，分群出來的結果是大致差不多的。只要是同一類高分項目就會差不多。所以在整理比較的時候都是和報告上已經整理好的cluster比較，如果造成困擾真的非常抱歉QAQQ

1. dataset :

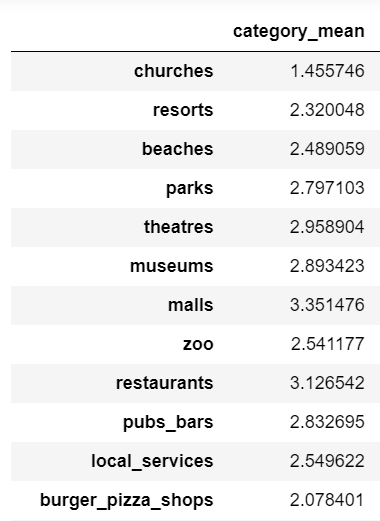
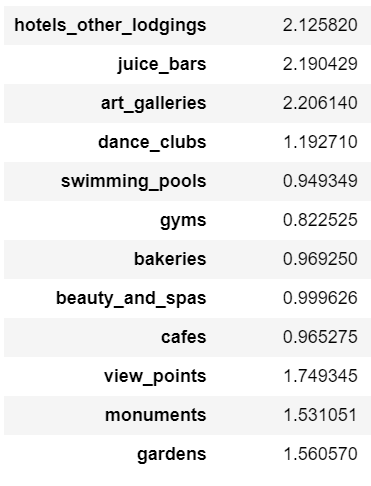
Travel Review Rating

1. analyze :
   * + 畫圖觀察一下資料在平面上的分布



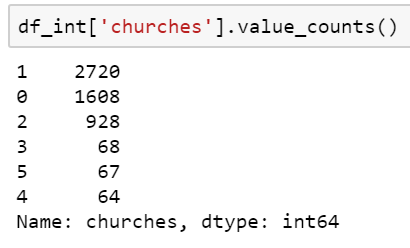
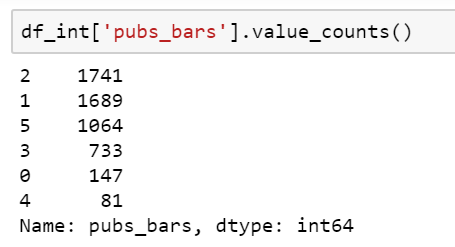
如果只分兩群的話，可以勉強看出圖上的分群區域。但分到12群就很難在圖上看出東西了。

* + - 找出每一個景點的平均分數

算出每一個景點的平均分數偷看一下。如果平均分數有2~3分的話，代表這個景點在這些使用者中有一定的支持度和評論分數，才有辦法拿到這個平均分數。如果平均分數只有0.幾的，可能就表示這個景點在這筆資料中的分數大多是低分，有需要的話可以優先考慮把這個景點拿掉。

* + - 找出每一個景點0~5分的個數

找每一個景點0~5各個分數的數量。先把整個dataframe轉成int type，再一個一個觀察。

可以看到在這些資料中，0分和1分的比例算是比較高的。我想在這種情況下，偏高分的345分的重要性就會提高。也就是當分完群之後，群中比較高分的景點可能就有足夠代表這個群體的重要程度與重要性。

1. problem :

利用各個景點的評論分數來嘗試分類客群?

1. approach :

利用k-means來把資料中的users分群

因為不知道k要設多少，所以嘗試用輪廓係數來找k的值。另外因為景點總共也才24個，如果太多群就會變成一個景點配一群使用者，感覺有點沒意義。所以決定在7~16之間利用輪廓係數找k值。

執行了幾次之後找出來k = 12的次數最多，所以決定把K設為12。

1. result :

12 clusters 特徵 :

Cluster 0 :

476人

4分以上 : 博物館、購物中心、

3~4分 : 劇院、餐廳、

2~3分 : 公園、動物園、

1~2分 : 教堂、渡假勝地、海灘、酒吧、當地服務、漢堡披薩店、飯店住宿、果汁吧、畫廊、紀念碑、

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點、花園

在這一群之中平均分數最高的景點是博物館跟購物中心。但是這一群的度假勝地和海灘這種比較需要花時間遊玩的景點分數就很低，可能是比較少人去。因此看起來大概比較喜歡四處逛逛的旅客。另外因為高分的有購物中心，所以我猜這些人可能是不住在那附近的本地人，也就是從其他地方過來玩的遊客。但是飯店住宿的分數也不高，並且餐廳有很高的分數，所以大概是一日遊的遊客，來逛一逛、吃個飯就回家。

Cluster 1 :

434人

4分以上 : 渡假勝地、購物中心、餐廳、

3~4分 : 博物館、動物園、酒吧、果汁吧、畫廊、

2~3分 : 海灘、公園、劇院、當地服務、漢堡披薩店、

1~2分 : 教堂、飯店住宿、

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點、紀念碑、花園

這一群高分的有度假勝地，也有購物中心，再加上酒吧餐廳，大概就是白天血拚晚上喝酒的大玩特玩行程。另外在這裡可以看到，這一群中飯店住宿的分數很低。我猜可能是google分類中的resorts就有包含住宿了，所以其他標記為飯店住宿的地標才會相對變的低分。因為已經住在resorts地方遊客們根本不需要再自己去找地方住。不過一般旅行社的套裝行程中常出現的景點和紀念碑這種，在這群中分數普遍偏低。所以我猜這群大概不是跟團的那種旅客，而是自己排行程，去自己想去的地方的自由行遊客。

Cluster 2 :

762人

4分以上 : 餐廳、酒吧、當地服務、

3~4分 : 購物中心、動物園、

2~3分 : 公園、劇院、博物館、漢堡披薩店、

1~2分 : 教堂、渡假勝地、海灘、飯店住宿、果汁吧、畫廊、跳舞俱樂部、景點、紀念碑、花園

1分以下 : 游泳池、健身房、麵包店、美容中心、咖啡店、

在這一群中，當地服務拿到了高分。我其實一直沒有很懂所謂的local services到底是指甚麼，但大概就是類似當地的交通之類的?如果是的話，我猜這一群大概就是為了特定目的而出門的當地歐洲人。例如想去逛街，或是和朋友約在某個餐廳之類的。所以一些比較會吸引在地人的景點拿到了高分，例如餐廳、酒吧、動物園。購物中心雖然比較像是去旅行的人逛的，但在地人應該也是可以逛，所以還算合理。而景點、紀念碑、度假勝地這種吸引觀光客的地方，對於在地人就比較沒有吸引力，得到的分數也較低。另外這一群的人數相對其他群有稍微多一點，我想可能是因為不論是當地人或是出來玩的遊客，都會需要利用到當地的一些服務，因此使得當地服務平均分數較高的這一群人數上升。

Cluster 3 :

519人

4分以上 : 公園、劇院、景點、

3~4分 : 海灘、博物館、

2~3分 : 渡假勝地、購物中心、餐廳、酒吧、當地服務、紀念碑、

1~2分 : 教堂、動物園、漢堡披薩店、飯店住宿、果汁吧、畫廊、跳舞俱樂部、美容中心、咖啡店、花園

1分以下 : 游泳池、健身房、麵包店、

這一群的高分景點中有其他群比較少出現的公園和劇院。如果公園包含遊樂園那種類型的話，我猜這一群很有可能是全家一起出遊的遊客類型。也就是帶小孩出去玩的爸爸媽媽們，或是陪年長的父母出遊。劇院、博物館和看風景的地方都很適合闔家出遊，海灘也可以讓小孩玩開心。當然長輩可能比較不會去海灘，但如果是陪可愛的孫女的話就有可能，也就是一樣是闔家出遊的類型。而飯店住宿的分數較低，可能是趁著工作空檔或假期帶全家一起出去玩的一日遊。這樣的遊玩方式我想應該很多人會選擇自己開車，因此利用到當地服務的機會減少，分數也就降低。不過在這之中，我覺得比較有趣的是酒吧的平均分數將近一半，有2.4分之多。帶小孩出門一般不會喝酒吧?所以這一項數據就有可能把我前面的假設全部推翻。但也有可能是在小孩睡了之後偷偷和另一半一起出去喝一杯，或是帶爸爸媽媽一起去喝。但是我突然想到，前面也有假設自己開車去的機會很大。也就是如果還要開車的話，負責開車的人就只能點果汁了。可憐。

Cluster 4 :

485人

4分以上 : 漢堡披薩店、飯店住宿、果汁吧、

3~4分 : 當地服務、畫廊、

2~3分 : 購物中心、餐廳、酒吧、

1~2分 : 海灘、公園、劇院、博物館、動物園、麵包店、美容中心、

1分以下 : 教堂、渡假勝地、跳舞俱樂部、游泳池、健身房、咖啡店、景點、紀念碑、花園

這一群就比較讓我困惑了。因為飯店住宿的高分，所以我想這一群應該不會是歐洲的本地人。但在這樣的前提之下，比較偏向速食類型的漢堡披薩店卻拿到了高分。也就是這是一群去歐洲玩了卻沒有甚麼錢吃飯的傢伙，為了省預算所以只好吃吃漢堡?而且漢堡還比一般的餐廳高分。可能是因為相對便宜吧。便宜就是美味。大概是拿爸爸媽媽辛苦賺來的錢出國玩，所以不敢太奢侈的窮學生。

Cluster 5 :

518人

4分以上 : 無

3~4分 : 麵包店、

2~3分 : 教堂、渡假勝地、海灘、公園、劇院、博物館、購物中心、餐廳、果汁吧、畫廊、健身房、美容中心、花園

1~2分 : 動物園、酒吧、當地服務、漢堡披薩店、飯店住宿、跳舞俱樂部、游泳池、咖啡店、景點、紀念碑、

1分以下 : 無

這一群跟其他群比起來，最大的不同就是既沒有4分以上的景點，也沒有1分以下的。也就是一群愛恨不明顯的人，又或是一群不喜歡給5分和0分的人。但實際去看了資料之後，發現這些人還是會給5分的。大家也會有給特別低分的景點。但或許是因為這些人給分的特性沒有辦法被分到其他群中，所以最後被通通集中來這裡。也就是一群太有特色的。

不過也有可能就是一群愛吃麵包的人啦。麵包店拿最高分的就是這群了。

Cluster 6 :

689人

4分以上 : 海灘、公園、劇院、

3~4分 : 渡假勝地、博物館、購物中心

2~3分 : 動物園、餐廳、酒吧、當地服務、

1~2分 : 教堂、漢堡披薩店、飯店住宿、果汁吧、畫廊、跳舞俱樂部、紀念碑、花園

1分以下 : 游泳池、健身房、麵包店、美容中心、咖啡店、景點

這一群的高分景點和cluster 0的很像。比較不一樣的大概就是度假勝地和海灘在這裡拿到了高分。因為有海灘，所以我猜這些人去的是比較靠近南方的南歐地區，到海邊享受暖陽的行程。所以才會有海灘和度假勝地的高分。只是如果要去海邊的話，感覺不會去劇院和博物館那些的?所以我猜大概會是兩天一夜或三天兩夜這種行程，一天去海邊，剩下就去市區逛逛。而這種行程可能很多人會為了省錢而選擇便宜的青年旅館和簡單的食物，才導致了這裡的餐廳和飯店住宿分數偏低。

Cluster 7 :

318人

4分以上 : 花園

3~4分 : 公園、劇院、博物館、購物中心

2~3分 : 教堂、渡假勝地、海灘、動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、紀念碑

1~2分 : 果汁吧

1分以下 : 畫廊、跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點

這一群中最值得一提的大概就是最高分的花園了。逛花園到底是個甚麼概念?我在做這個作業的時候一直很不解。是要買花的那種?還是單純坐著看花喝咖啡?總之我猜會喜歡花園的人，應該是對自己的眼光和品味有要求的人。所以在這樣的前提之下，高分的還有劇院、博物館、購物中心等等是合理的。可能是有錢貴婦優雅的一日遊。去花園喝個下午茶，再去去劇院提升一下自己的品味或是去博物館吹個冷氣，最後再到購物中心好好血拚一番，結束完美的一天。

Cluster 8 :

301人

4分以上 : 當地服務、飯店住宿

3~4分 : 公園、劇院、購物中心、餐廳、酒吧、景點、紀念碑

2~3分 : 渡假勝地、海灘、博物館、動物園、漢堡披薩店、

1~2分 : 教堂、果汁吧、畫廊、花園

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店

這也是很有特色的一群。在這一群中，拿到最高分的分別是當地服務和飯店住宿。最高分的居然不是遊玩的項目，而是偏向服務性質的類型。我想這會是客氣且有禮貌，願意表達感謝的一群人。對於旅行中的人們提供的服務給予高分，說明他們對於默默幫助自己的旅途更加順利的人們都看在眼裡，並經由打下的高分表示他們的謝意，希望透過評論分數讓他們的努力可以被更多人看見。

Cluster 9 :

208人

4分以上 : 畫廊、跳舞俱樂部

3~4分 : 購物中心、果汁吧、游泳池、

2~3分 : 公園、劇院、博物館、動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、

1~2分 : 教堂、渡假勝地、海灘、健身房、咖啡店、景點、紀念碑、花園

1分以下 : 麵包店、美容中心

這也是讓我困惑的一群。稍微上網查了一下，國外的club大概就是可以喝酒又可以跳舞的地方，類似台灣的夜店。所以我關於dance club的認知應該沒有太大的差錯。也就是這是一群白天喜歡逛逛畫廊游游泳，晚上就去夜店嗨的人?看來是喜歡熱鬧的美術系朋友呢。人數不多，大概是小眾。

不過也有可能單純只是分類器把這兩個靠近的欄位都拿到高分的使用者分在一起而已。看來要試試看把順序打亂的結果。

Cluster 10 :

428人

4分以上 : 購物中心、餐廳、畫廊

3~4分 : 博物館、動物園、酒吧、果汁吧

2~3分 : 海灘、公園、劇院、當地服務、飯店住宿

1~2分 : 教堂、渡假勝地、漢堡披薩店、花園

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點、紀念碑

這一群看起來就比cluster 9 正常多了。畫廊、博物館都拿到高分，大概是對藝術或文化有興趣的遊客。除了畫廊博物館這兩項之外，也有動物園和酒吧這些比較不同性質的景點拿到了高分，但飯店住宿卻在2~3分的區間。我猜可能跟cluster 6的情形有點像，只玩兩天或三天，所以住宿沒有太高分，但高分項目有比較大的差異。

Cluster 11 :

328人

4分以上 : 景點

3~4分 : 咖啡店、紀念碑、花園

2~3分 : 教堂、渡假勝地、海灘、畫廊、美容中心、

1~2分 : 公園、劇院、博物館、購物中心、動物園、餐廳、酒吧、當地服務、漢堡披薩店、果汁吧、跳舞俱樂部、游泳池、健身房、麵包店、

1分以下 : 飯店住宿

這一群的特色就很明顯。在其他群中都沒有太好表現的view point在這裡拿到了最高分。雖然不太確定，但我把view point理解為看風景的地方。會喜歡這類大自然風景的人應該是比較不會喜歡市區購物的行程，因此購物中心或俱樂部這類景點的分數都偏低。並且花園、紀念碑的分數較高，所以我想我的假設是合理的。

另外值得一提的是這裡的飯店住宿分數拿到了最低。可能是因為風景這類的景點都不是會需要過夜的，大概一個下午去看一看就可以走了，所以不需要用到飯店，自然也就沒甚麼評論分數。

也或許是因為不需要飯店住宿，去咖啡店的人也被分到了這群。應該沒有人會特地到外面住飯店只為了去咖啡店吧?所以去喝咖啡的大概都是當天來回，喝完就回家了。當然就沒有飯店的評論分數。所以這一群中的咖啡店平均分數也拿到了高分。

小結 :

從第一次的分類中可以看到，雖然會有一些小誤差，但大致上還是可以分出一些類別，每一類都會有一些特徵來表明他們是同一個族群。

但有一些資料在這幾群中都沒有特別的表現，也就是幾乎每一群對於這些景點的評論分數都不高也不低，例如果汁吧和游泳池等。因此決定試試看把一些比較不重要或是對於分類客群的幫助較小的景點資料拿掉，看看能不能分類出和第一次分類類似的結果。

另外因為在觀察第一次分類完的資料之後覺得好像有把靠近的column分數相近的user分在同一類，所以也想試試看打亂景點順序之後會不會影響分群結果。

1. Improve 1 : 拿掉不重要的資料，看看能不能分出類似的結果

* 果汁吧、游泳池、教堂 : 幾乎都在1、2分，就算上到3分以上，也有其他更有特徵的景點來表達那個群體，所以拿掉。
* 美容中心、健身房 : 客群明顯，但在第一次的分類中看起來對於這五千多筆資料的分群也沒有太大幫助，可能是因為在這些人裡面特別喜歡美容中心和健身房的人比較少，所以決定拿掉。

Result :

Cluster 0 :

677人

4分以上 : 劇院、博物館、購物中心、

3~4分 : 公園、酒吧、

2~3分 : 海灘、動物園、餐廳、當地服務、

1~2分 : 渡假勝地、漢堡披薩店、飯店住宿、畫廊、紀念碑、花園

1分以下 : 跳舞俱樂部、麵包店、咖啡店、景點、

這一群有點看不太出來對應到第一次分類的哪一類。但高分的景點有博物館、購物中心、劇院，應該可以勉強對應到第一次的cluster 0。就是一日遊的遊客。

Cluster 1 :

342人

4分以上 : 景點、

3~4分 : 咖啡店、紀念碑、花園

2~3分 : 渡假勝地、海灘、畫廊、

1~2分 : 公園、劇院、博物館、購物中心、動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、跳舞俱樂部、麵包店、

1分以下 : 無

類似第一次分類的cluster 11。景點、紀念碑、花園拿到了高分，也就是當天逛逛看風景的那一群。

Cluster 2 :

491人

4分以上 : 漢堡披薩店、飯店住宿、

3~4分 : 當地服務、畫廊、

2~3分 : 購物中心、餐廳、酒吧、

1~2分 :海灘、公園、劇院、博物館、動物園、麵包店、

1分以下 : 渡假勝地、跳舞俱樂部、咖啡店、景點、紀念碑、花園

類似第一次分類的cluster 4。飯店和漢堡拿到了高分，可以看出來就是去玩又沒錢吃飯，窮學生那群。

Cluster 3 :

271人

4分以上 : 當地服務、景點、

3~4分 : 公園、劇院、購物中心、餐廳、酒吧、飯店住宿、紀念碑

2~3分 : 渡假勝地、海灘、博物館、動物園、

1~2分 : 漢堡披薩店、畫廊、花園

1分以下 : 跳舞俱樂部、麵包店、咖啡店

這一群我猜是對應到第一次分類的cluster 8。當地服務拿到了最高分。雖然飯店住宿沒有到4分，但打開資料看會發現平均評論分數有3.98分，非常接近4分。可能是因為拿掉了一些資料造成的浮動，但還是可以相信這一群對應到的就是對飯店的人很好的那一群。

Cluster 4 :

727人

4分以上 : 餐廳、酒吧、當地服務、

3~4分 : 購物中心、動物園、

2~3分 : 公園、劇院、博物館、

1~2分 : 渡假勝地、海灘、漢堡披薩店、飯店住宿、畫廊、跳舞俱樂部、景點、紀念碑、花園

1分以下 : 麵包店、咖啡店、

這一群類似原本分類的cluster 2。就是出門晃晃的歐洲本地人，因為會用到大眾運輸工具所以有當地服務的高分。

Cluster 5 :

530人

4分以上 : 無

3~4分 : 麵包店、

2~3分 : 渡假勝地、海灘、公園、劇院、博物館、購物中心、果汁吧、畫廊、花園

1~2分 : 動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、跳舞俱樂部、咖啡店、景點、紀念碑、

1分以下 : 無

對應到原本分類的cluster 5。就是無欲無求，沒有高分也沒有低分的那群。或是愛吃麵包那群。

Cluster 6 :

440人

4分以上 : 渡假勝地、購物中心、餐廳、

3~4分 : 博物館、動物園、酒吧、畫廊、

2~3分 : 海灘、公園、劇院、當地服務、漢堡披薩店、

1~2分 : 飯店住宿、

1分以下 : 跳舞俱樂部、麵包店、咖啡店、景點、紀念碑、花園

對應到第一次分類的cluster 1。高分的景點類別中就只差在拿掉了果汁吧，其他都一樣。就是自己安排行程去度假的自由行散客。

Cluster 7 :

493人

4分以上 : 渡假勝地、海灘、劇院、

3~4分 : 公園、博物館、

2~3分 : 購物中心、動物園、餐廳、酒吧、

1~2分 : 當地服務、漢堡披薩店、飯店住宿、畫廊、跳舞俱樂部、紀念碑、花園

1分以下 : 麵包店、咖啡店、景點

這一群就有點混亂。單看高分項目，可以認為是對應到原本分類的cluster 6。只是前一次分類的cluster 6 的人數有689人，這一群只有493，差了快兩百。但因為這兩群的高分項目和低分項目都比較類似，所以認定這一群對應到的是上一次的cluster 6。可能是有一些特徵跟其他群也很像的user被分去其他群，所以人數才會下降。總之就是兩三天玩海邊。

Cluster 8 :

525人

4分以上 : 公園、劇院、景點、

3~4分 : 海灘、博物館、

2~3分 : 渡假勝地、購物中心、餐廳、酒吧、當地服務、紀念碑、

1~2分 : 動物園、漢堡披薩店、飯店住宿、畫廊、跳舞俱樂部、咖啡店、花園

1分以下 : 麵包店、

這一群類似第一次分類的cluster 3。全家一起去海邊玩的那群。真是羨慕。我也想結婚。

Cluster 9 :

228人(9

4分以上 : 畫廊、跳舞俱樂部

3~4分 : 購物中心、

2~3分 : 海灘、公園、劇院、博物館、動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、

1~2分 : 渡假勝地、咖啡店、景點、紀念碑、花園

1分以下 : 麵包店

類似第一次分類的cluster 9。夜店美術系的那群。高分的項目就差在被拿掉的果汁吧跟游泳池而已。

Cluster 10 :

291人

4分以上 : 劇院、

3~4分 : 公園、博物館、購物中心、漢堡披薩店、花園

2~3分 : 渡假勝地、海灘、動物園、餐廳、酒吧、當地服務、飯店住宿、紀念碑

1~2分 : 畫廊、

1分以下 : 跳舞俱樂部、麵包店、咖啡店、景點

這應該是原本cluster 7 的那一群。但最有特色的花園掉到了3~4分，反而是劇院上到4分。打開資料偷看一下，發現花園的平均分數還是有3.95分，可以算是接近4分了。所以認定這一群應該就是第一次分類中分出來的貴婦群。

Cluster 11 :

453人

4分以上 : 購物中心、餐廳、畫廊

3~4分 : 博物館、動物園、酒吧、

2~3分 : 海灘、公園、劇院、當地服務、飯店住宿

1~2分 : 渡假勝地、漢堡披薩店、花園

1分以下 : 跳舞俱樂部、麵包店、咖啡店、景點、紀念碑

畫廊跟博物館都拿到高分。大概就是第一次分類的cluster 10。出來玩兩三天的氣質文化派遊客。

小結 :

試試看用表格來比對第一次分群和拿掉不重要的資料後分群的情形 :

|  |  |  |
| --- | --- | --- |
| Improve後分群(拿掉不重要資料) | 第一次分群 | 族群特徵 |
| Cluster 0 / 677人 | Cluster 0 / 476人 | 一日遊的遊客 |
| Cluster 1 / 342人 | Cluster 11 / 328人 | 當天逛逛看風景 |
| Cluster 2 / 491人 | Cluster 4 / 485人 | 吃漢堡窮學生群 |
| Cluster 3 / 271人 | Cluster 8 / 301人 | 客氣人好 飯店高分群 |
| Cluster 4 / 727人 | Cluster 2 / 762人 | 出門晃晃的歐洲人 |
| Cluster 5 / 530人 | Cluster 5 / 518人 | 吃麵包那群 |
| Cluster 6 / 440人 | Cluster 1 / 434人 | 自己安排的自由行散客 |
| Cluster 7 / 493人 | Cluster 6 / 689人 | 兩三天玩海邊 |
| Cluster 8 / 525人 | Cluster 3 / 519人 | 全家一起去海邊 |
| Cluster 9 / 228人 | Cluster 9 / 208人 | 夜店美術系 |
| Cluster 10 / 291人 | Cluster 7 / 318人 | 去花園看花的貴婦團 |
| Cluster 11 / 453人 | Cluster 10 / 328人 | 氣質路線文化遊客 |

拿掉了幾個景點資料以後，還是可以分出跟第一次分類差不多的族群。只是順序有點不太一樣，但跟分群內容沒有太大的關係，結果也大致符合第一次的分群狀況。只是還是有一兩群比較對不起來，可能就是拿掉資料造成的影響。不過也或許是原本就存在的誤差。

所以我猜拿掉這些資料是有意義的。可以減少資料量，而且成效也沒有太差。

1. Improve 2 : 打亂景點的順序

因為拿掉不重要的資料後分出來的群和第一次分群還是有一點點不同，所以決定打亂順序時就不拿掉資料，來控制並減少不確定的因素。

Result :

Cluster 0 :

319人

4分以上 : 花園

3~4分 : 公園、劇院、博物館、購物中心

2~3分 : 教堂、渡假勝地、海灘、動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、紀念碑

1~2分 : 果汁吧

1分以下 : 畫廊、跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點

Cluster 1 :

300人

4分以上 : 當地服務、飯店住宿

3~4分 : 公園、劇院、購物中心、餐廳、酒吧、景點、紀念碑

2~3分 : 渡假勝地、海灘、博物館、動物園、漢堡披薩店、

1~2分 : 教堂、果汁吧、畫廊、花園

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店

Cluster 2 :

757人

4分以上 : 餐廳、酒吧、當地服務、

3~4分 : 購物中心、動物園、

2~3分 : 公園、劇院、博物館、漢堡披薩店、

1~2分 : 教堂、渡假勝地、海灘、飯店住宿、果汁吧、畫廊、跳舞俱樂部、景點、紀念碑、花園

1分以下 : 游泳池、健身房、麵包店、美容中心、咖啡店、

Cluster 3 :

500人

4分以上 : 無

3~4分 : 麵包店、

2~3分 : 教堂、渡假勝地、海灘、公園、劇院、博物館、購物中心、餐廳、果汁吧、畫廊、健身房、美容中心、花園

1~2分 : 動物園、酒吧、當地服務、漢堡披薩店、飯店住宿、跳舞俱樂部、游泳池、咖啡店、景點、紀念碑、

1分以下 : 無

Cluster 4 :

437人

4分以上 : 購物中心、餐廳、畫廊

3~4分 : 博物館、動物園、酒吧、果汁吧

2~3分 : 海灘、公園、劇院、當地服務、飯店住宿

1~2分 : 教堂、渡假勝地、漢堡披薩店、花園

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點、紀念碑

Cluster 5 :

436人

4分以上 : 渡假勝地、購物中心、餐廳、

3~4分 : 博物館、動物園、酒吧、果汁吧、畫廊、

2~3分 : 海灘、公園、劇院、當地服務、漢堡披薩店、

1~2分 : 教堂、飯店住宿、

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、美容中心、咖啡店、景點、紀念碑、花園

Cluster 6 :

318人

4分以上 : 景點

3~4分 : 咖啡店、紀念碑、花園

2~3分 : 教堂、渡假勝地、海灘、畫廊、美容中心、

1~2分 : 公園、劇院、博物館、購物中心、動物園、餐廳、酒吧、當地服務、漢堡披薩店、果汁吧、跳舞俱樂部、游泳池、健身房、麵包店、

1分以下 : 飯店住宿

Cluster 7 :

515人

4分以上 : 渡假勝地、劇院、

3~4分 : 公園、海灘、博物館、

2~3分 : 購物中心、動物園、餐廳、酒吧、

1~2分 : 教堂、當地服務、漢堡披薩店、飯店住宿、果汁吧、畫廊、跳舞俱樂部、紀念碑、花園

1分以下 : 麵包店、咖啡店、景點、健身房、游泳池、美容中心

這一群像improve 1 拿掉不重要資料後的cluster 7 !!

只是海灘平均3.96掉到3~4分，但還是接近4分。其他的景點分數幾乎都一樣。

Cluster 8 :

485人

4分以上 : 漢堡披薩店、飯店住宿、果汁吧、

3~4分 : 當地服務、畫廊、

2~3分 : 購物中心、餐廳、酒吧、

1~2分 : 海灘、公園、劇院、博物館、動物園、麵包店、美容中心、

1分以下 : 教堂、渡假勝地、跳舞俱樂部、游泳池、健身房、咖啡店、景點、紀念碑、花園

Cluster 9 :

524人

4分以上 : 公園、劇院、景點、

3~4分 : 海灘、博物館、

2~3分 : 渡假勝地、購物中心、餐廳、酒吧、當地服務、紀念碑、

1~2分 : 教堂、動物園、漢堡披薩店、飯店住宿、果汁吧、畫廊、跳舞俱樂部、美容中心、咖啡店、花園

1分以下 : 游泳池、健身房、麵包店、

Cluster 10 :

208人

4分以上 : 畫廊、跳舞俱樂部

3~4分 : 購物中心、果汁吧、游泳池、

2~3分 : 公園、劇院、博物館、動物園、餐廳、酒吧、當地服務、漢堡披薩店、飯店住宿、

1~2分 : 教堂、渡假勝地、海灘、健身房、咖啡店、景點、紀念碑、花園

1分以下 : 麵包店、美容中心

Cluster 11 :

667人

4分以上 : 劇院、博物館、購物中心、

3~4分 : 公園、

2~3分 : 酒吧、海灘、動物園、餐廳、

1~2分 : 教堂、渡假勝地、漢堡披薩店、飯店住宿、當地服務、果汁吧、畫廊、紀念碑、

1分以下 : 跳舞俱樂部、游泳池、健身房、麵包店、咖啡店、景點、美容中心、花園

這群比較類似improve 1 拿掉不重要資料後的cluster 0 ?! 高分的景點項目幾乎一樣，人數也相近。驚恐。

小結 :

直接上圖。

|  |  |  |
| --- | --- | --- |
| 打亂景點順序後分群 | 第一次分群 | 族群特徵 |
| Cluster 0 / 319人 | Cluster 7 / 318人 | 去花園看花的貴婦團  一日遊的遊客 |
| Cluster 1 / 300人 | Cluster 8 / 301人 | 客氣人好 飯店高分群 |
| Cluster 2 / 757人 | Cluster 2 / 762人 | 出門晃晃的歐洲人 |
| Cluster 3 / 500人 | Cluster 5 / 518人 | 吃麵包那群 |
| Cluster 4 / 437人 | Cluster 10 / 428人 | 氣質路線文化遊客 |
| Cluster 5 / 436人 | Cluster 1 / 434人 | 自己安排的自由行散客 |
| Cluster 6 / 318人 | Cluster 11 / 328人 | 當天逛逛看風景 |
| Cluster 7 / 515人 | Improve 1 Cluster 7 / 493人 | 跟improve 1 的cluster 7很像!! |
| Cluster 8 / 485人 | Cluster 4 / 485人 | 吃漢堡窮學生群 |
| Cluster 9 / 524人 | Cluster 3 / 519人 | 全家一起去海邊 |
| Cluster 10 / 208人 | Cluster 9 / 208人 | 夜店美術系 |
| Cluster 11 / 667人 | Improve 1 Cluster 0 / 677人 | 跟improve 1 的cluster 0很像!! |

打亂順序之後的結果跟沒有打亂的結果幾乎一樣。連人數都很接近。看來是我太小看kmeans了。

但可怕的是不只是improve 1 的時候還是improve 2 的時候，原本分出來的cluster 0 和 cluster 6 都找不到對應的族群!!反而是兩個improve分出來沒辦法跟cluster 0 和 cluster 6 對的群有辦法互相對應。也就是說一開始分出來的cluster 0 和 cluster 6很有可能是錯的。尷尬。

1. 結論與討論 :

k-means分群確實有一定的準確度和穩定性，有辦法利用單純的評論分數資料來分類這些不知名的使用者。就算在嘗試improve的部分改變了資料的數量，拿掉幾筆資料之後還是有辦法做出差不多的分群結果。改變景點的資料順序後結果也很穩定。不過後來想想，改變資料順序這件事實在是沒甚麼意義呀。景點之間並沒有明顯的相關，使用者之間也沒有。就算可能有些景點在附近所以有地域關係、或是使用者之間是親友關係，但資料上沒有，所以都可以忽略。因為根本不知道有沒有這回事。

因此我猜，要在使用者分群這個問題中取得更好的解答，應該還需要更多資料。例如使用者的背景或是評分的時機、評分的地點等等。又或是景點的所在位置或各自之間的相對位置。有了這些依據，才有辦法更有效率，甚至更準確地看出我們的分群結果究竟準不準確，而不是像這次憑空想像亂猜。

不過資料一多就能做的更好這件事，好像對於其他問題來說也是一樣就是了。

另外在這次作業中，我討論時幾乎都是利用每一個群體的景點各自的平均分數來討論。平均分數雖然可以算是一個重要而且明確的指標，但全部只看平均好像還是有些偏頗。如果下次可以加入其他的資料屬性一起觀察討論的話，說不定可以看出更多有趣的現象。這或許也可以算是一種improve的方式?