FDA_HW3-2 Report

F74066307 陳品修

1. 資料選擇與前處理:

選擇**Tarvel Review Ratings Data Set做為資料集**

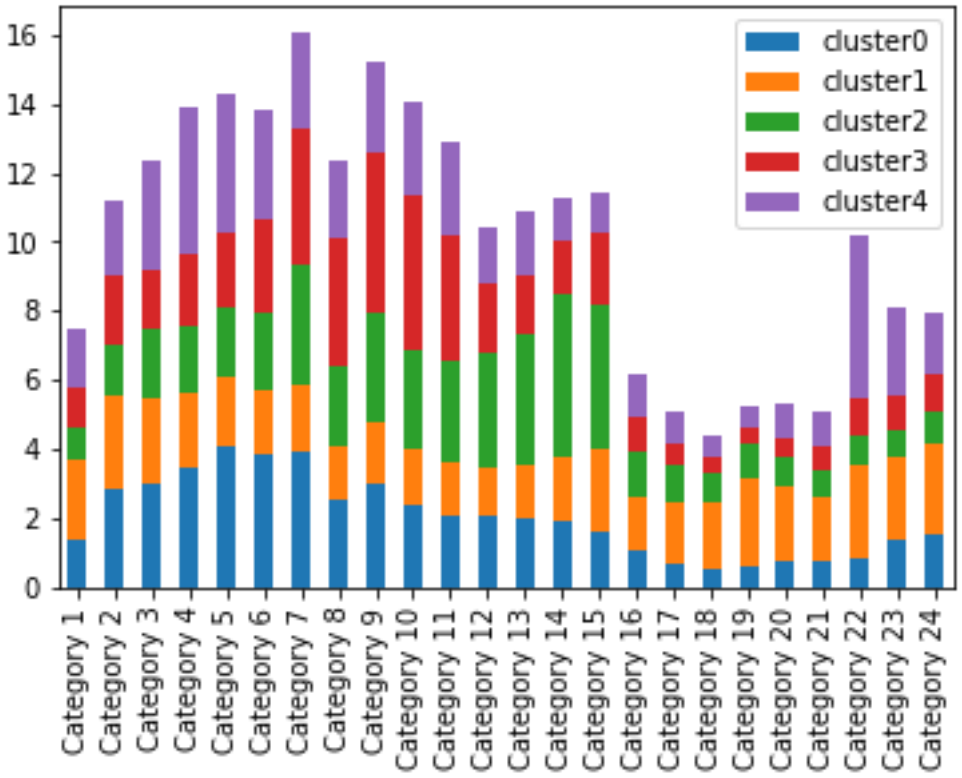
首先將用不到的UserId以及Unnamed的column 25 del掉

在嘗試直接拿資料去分群時，發現會有「2\t2.」與「NaN」出現，評分範圍應落在0到5之間，因此將2\t2.替換成2.2，NaN替換成0

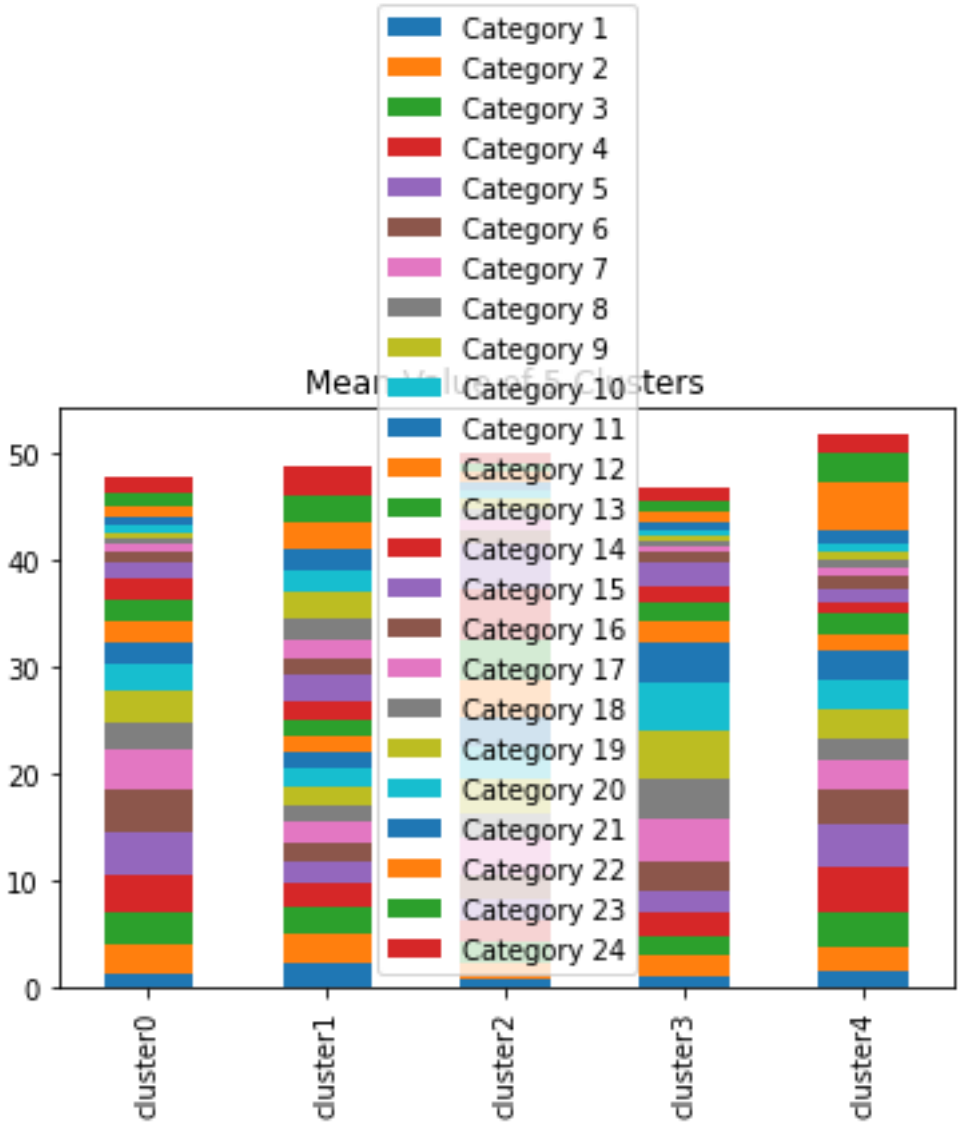
另外還有，雖然不知道原因，只有Category 11的type是Object而不是float，因此使用astype將Category 11改成float

2. 分群:

因為尚未明白資料的關係性，因此第一次先用KMeans，嘗試將資料分成5群，分完之後，對每個Category裡同一clusters id取平均值，畫成疊加型的Bar chart，看各Category中，在哪個cluster裡的平均值最高，再進行轉置，利用轉置後畫出的圖看各Category在哪個cluster占最多



轉置後:



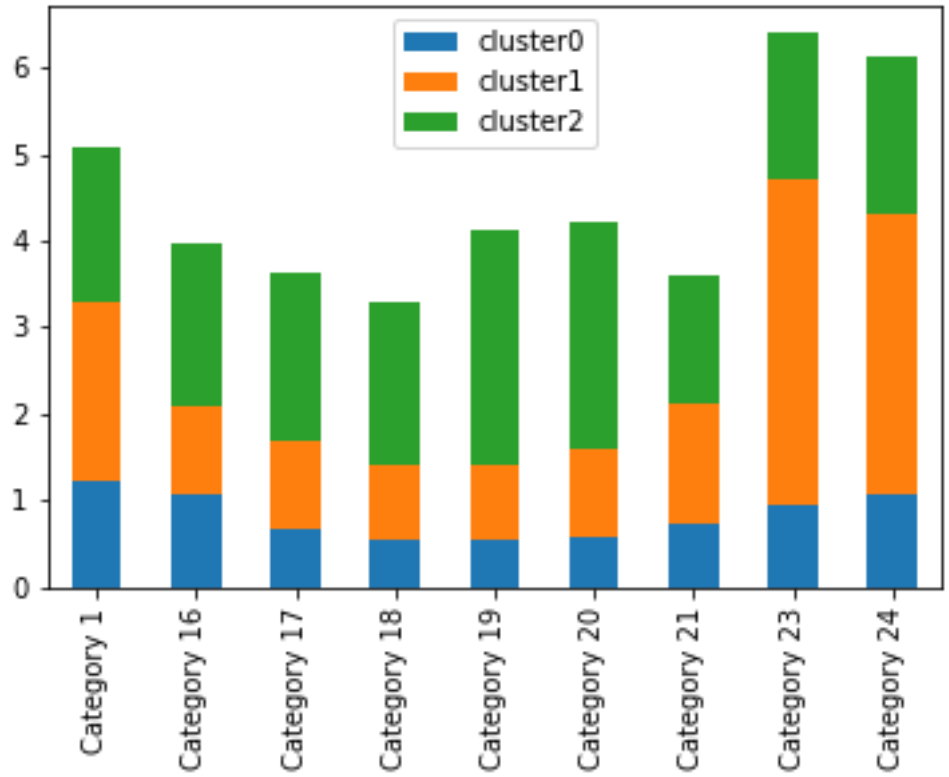
之所以看平均值，是認為因為這是評分的資料集，若是一群資料的平均值較另一群高，可能代表高得分的集中在一群中，能看出評分的走向，因此選用這個標準來分類

3. 再分群:

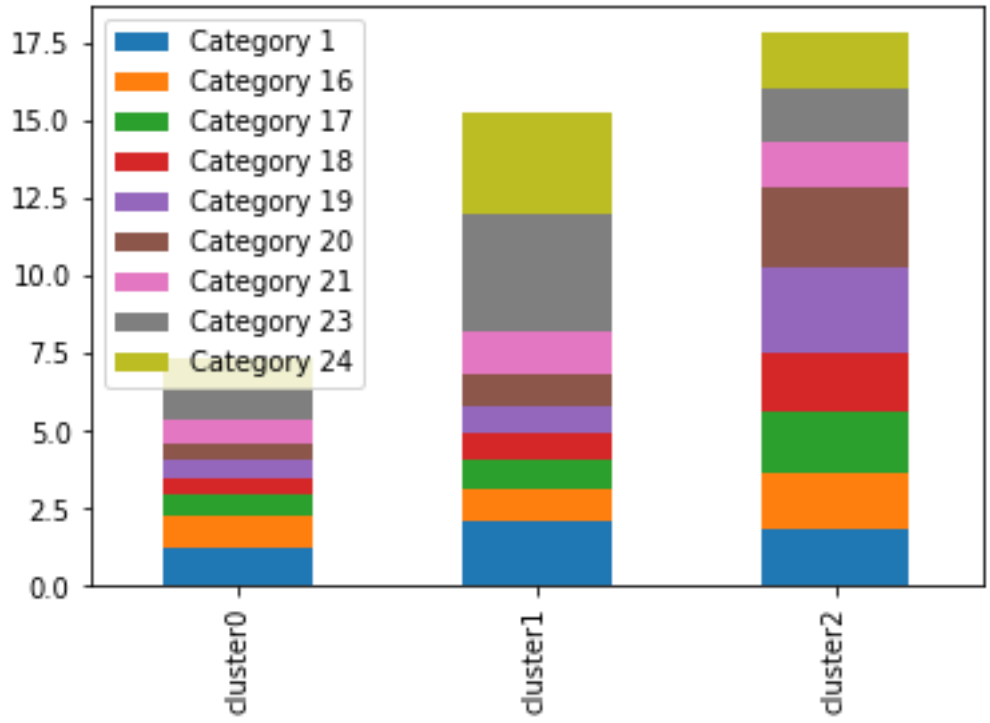
第一次分完的結果是:

```
#在 cluster0 bar 高:
c0=df[['Category 2','Category 5', 'Category 6','Category 7']]

#在 cluster1 bar 高:
c1=df[['Category 1', 'Category 16','Category 17', 'Category 18', 'Category 19'
, 'Category 20', 'Category 21', 'Category 23', 'Category 24']]

#在 cluster2 bar 高:
c2=df[['Category 12', 'Category 13', 'Category 14', 'Category 15']]

#在 cluster3 bar 高:
c3=df[['Category 8', 'Category 9', 'Category 10', 'Category 11']]

#在 cluster4 bar 高:
c4=df[['Category 3','Category 4','Category 22']]
```

其中c1的欄位稍為過多，因此再對c1做一次KMeans，這次是分3群，結果:



轉置後:

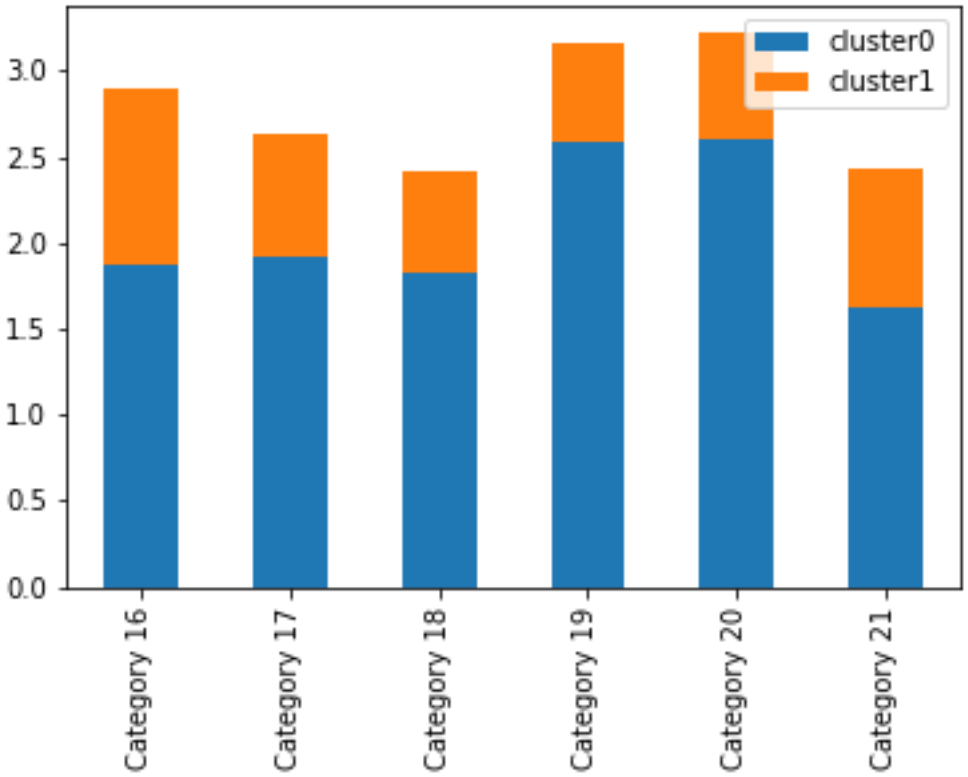


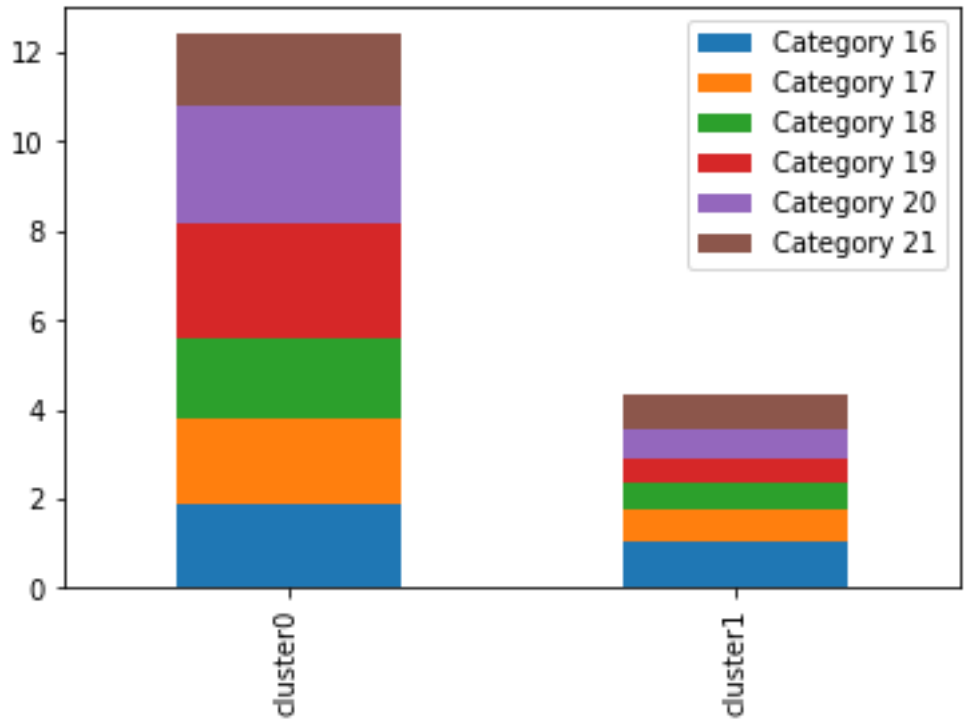
可以看見Category 1，23，24的分群走向與其他不同，在cluster 1最高，因此將其分出，變成:

```
c1_1=df[['Category 16','Category 17', 'Category 18', 'Category 19', 'Category 20', 'Category 21']]
c1_2=df[['Category 1', 'Category 23','Category 24']]
```

4. 嘗試分更多群:

在c1_1中仍有6個欄位，因此嘗試再利用KMeans分成2群，而結果為:





而對c0，c2，c3這三組有4個欄位的，進行再分群也無法分出走向不同的項，因此就做到這裡

5. 分群導出的結論:

最後結果:

```
g1=df[['Category 16','Category 17', 'Category 18', 'Category 19', 'Category 20
', 'Category 21']]
g2=df[['Category 1', 'Category 23','Category 24']]
g3=df[['Category 2','Category 5', 'Category 6','Category 7']]
g4=df[['Category 12', 'Category 13', 'Category 14', 'Category 15']]
g5=df[['Category 8', 'Category 9', 'Category 10', 'Category 11']]
g6=df[['Category 3','Category 4','Category 22']]
```

第一組:Category 16~21，以Attribute來看是17~22，也就是dance clubs，swimming pools，gyms，bakeries，beauty & spas跟cafes六項，這六項的打分走向類似，或許像是在豪華旅館內有這些設施，在旅館入住，享受館內設施時一齊打分，因此分數類似吧

第二組:Category 1，23，24，為churches，monuments跟garden，

第三組:Category 2，5，6，7，為resorts，theatres，museums，malls

第四組: Category 12~15，為burger/pizza shops，hotels/other lodgings，juice bars，art galleries

第五組: Category 8~11，為zoo，restaurants，pubs/bars，local services

第六組: Category 3，4，22，為beaches，parks，view points

分類在同一組的代表分數的走向類似，可能同高或是同低