HW 3-2 Report

嚴立言 F74056043

dataset selection : **Tarvel Review Ratings**

* Dataset Description

Data columns (total 26 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 User 5456 non-null object

1 Category 1 5456 non-null float64

2 Category 2 5456 non-null float64

3 Category 3 5456 non-null float64

4 Category 4 5456 non-null float64

5 Category 5 5456 non-null float64

6 Category 6 5456 non-null float64

7 Category 7 5456 non-null float64

8 Category 8 5456 non-null float64

9 Category 9 5456 non-null float64

10 Category 10 5456 non-null float64

11 Category 11 5456 non-null object

12 Category 12 5455 non-null float64

13 Category 13 5456 non-null float64

14 Category 14 5456 non-null float64

15 Category 15 5456 non-null float64

16 Category 16 5456 non-null float64

17 Category 17 5456 non-null float64

18 Category 18 5456 non-null float64

19 Category 19 5456 non-null float64

20 Category 20 5456 non-null float64

21 Category 21 5456 non-null float64

22 Category 22 5456 non-null float64

23 Category 23 5456 non-null float64

24 Category 24 5455 non-null float64

25 Unnamed: 25 2 non-null float64

---------------------------------------------------------------------------------------

- user 欄位為沒有意義的流水號

- category11 的 type 不知為何是 object，將其轉為 float

- category25 幾乎沒有資料，因此將其移除

- category12, 24 有缺值，將其以中位數補上

(後來有發現到其實是有兩欄位移了一格，不過考慮到資料集本身的就有些問題(後面會說明)，認為這兩欄不會嚴重影響到結果，因此沒有多作處理)

發現的問題：

**" Google user rating ranges from 1 to 5 and average user rating per category is calculated."**

根據敘述，可以理解為每個數值是「該使用者到過的所有該類別地點的評分平均」

。且根據敘述，應該不能評一分以下的數字，因此即使將零分視為該使用者沒有做過任何評分，也無法解釋 0 ~ 1 分是如何出現的。

解決方法：

- 無視 0 ~ 1 的值

- 將 0 視為缺失值，以 knn 補上

* Problem Decision

我在想分析目標時發現了幾個問題

- 各個類別間沒有因果關係

- 各個類別間沒有重要性差異（沒有預測單個類別的必要）

- 每個 user 對於不同類型的喜好差異極大

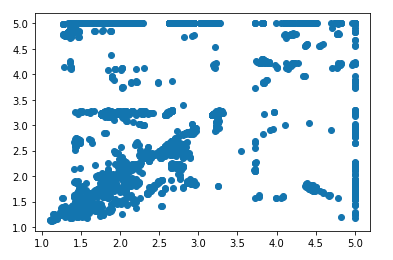
- 種類過多導致分群困難

因此我打算先算出類別間彼此的相關性

發現的問題：

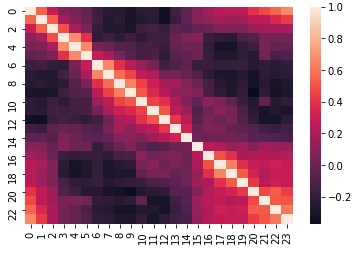
試著將不同 category 繪成 scatter plot 後，發現有非常多值在 x = 5 or y = 5 的線上。

覺得過於極端，因此將其全部視為 outlier ，以 knn 補上。



結果：

繪製出相關係數後感覺各分類之間沒有明顯的關聯，因此打算更改目標。



* Problem Decision version 2

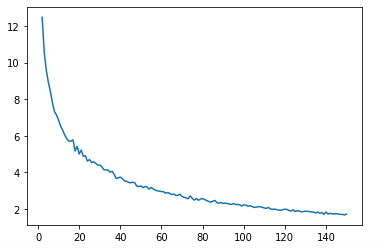
新目標嘗試將使用者分群，藉此找出喜好相近的人群

評估方法：

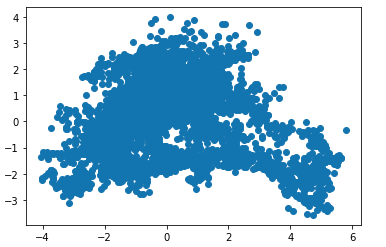
1. 利用 k means 將母體分成 n 個子群
2. 計算每個子群的變異數
3. 重複測試不同的 n，找出變異數最小的 n

觀察：

從折線圖看來，隨著 n 的增加變異數大概會持續減少直到 0 為止。不過分的太細就失去了分群的意義，所以應該要嘗試找出一個折衷的數字。



* PCA 主成份分析



用 pca 將資料降到 2 維看看有沒有明顯的分群。

不過在跟同學討論過後發現兩人降維後的資料長得不一樣，討論後認為是 preprocessing 的方式不一樣，在思考哪種方法比較適合的時候覺得原始資料有很多問題。

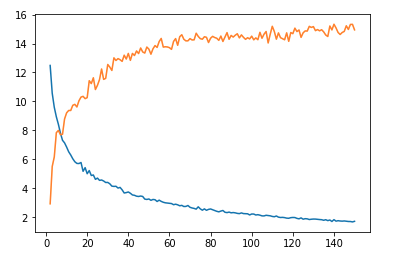
1. 一開始就有發現的 0 ~ 1 分的評分
2. 多數使用者的評分很極端，常常是某些給到 5 分，有些卻只給 1 分
3. 將分佈圖畫出來後發現多數分數集中在 3 分以下，然而 google map 上多數店家的評分都在 4 分以上

結論： 因為資料集感覺本身就有一些問題，因此不打算糾結在 preprocessing 的對錯上

* 最佳分群數目

根據 Elbow Method 的原理，「使群內的總變異最小；使群間的總變異最大」，接下來用類似的方式求群間變異數。

結果：群間變異數沒有變小的趨勢，因此嘗試別的方法。這次嘗試用 silhouette method。



結果：silhouette 和 Elbow 的結果有些類似，差別在於到了 6,70 有開始趨緩的跡象。

可能的解釋是此資料集的使用者無法做一個較好的分群

