資料科學導論 HW2

學號:B11015030
姓名:張睿麟
系級:四資工三乙

資料清理

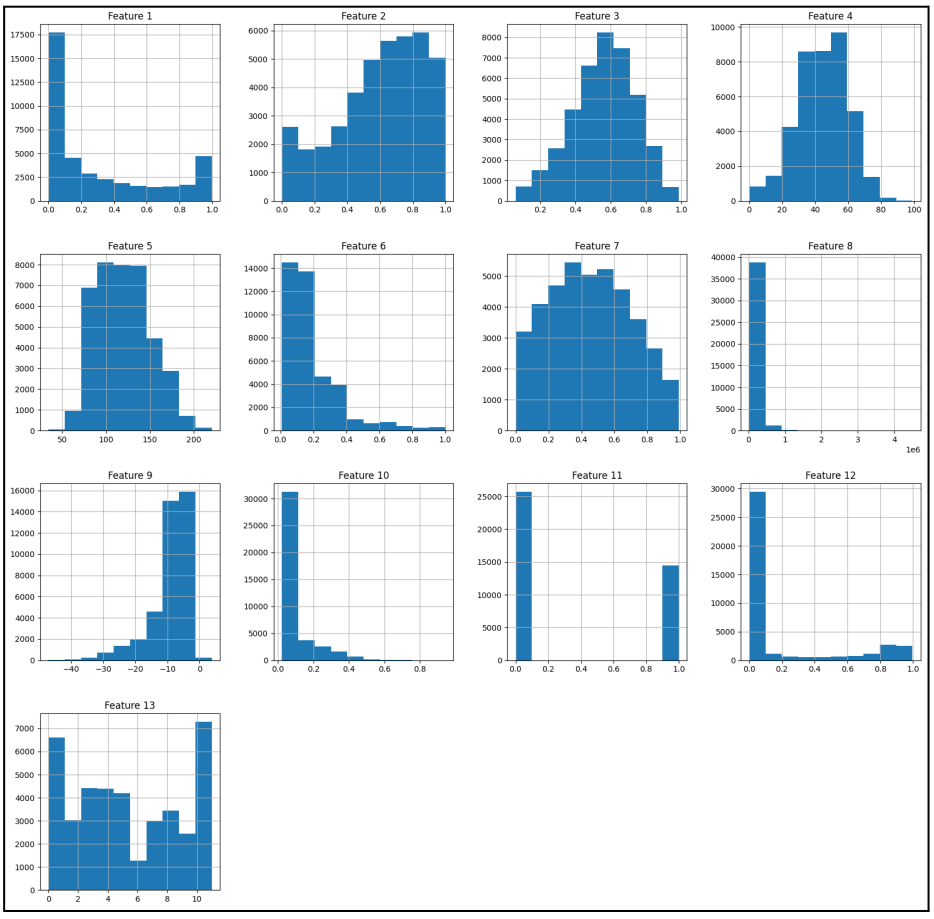
1. 本次資料無缺失值
2. 先直接對移除 ['song_id'] 欄位

資料分析

1. 敘述性統計描述

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11	Feature 12
count	40114.000000	40114.000000	40114.000000	40114.000000	40114.000000	40114.000000	40114.000000	4.011400e+04	40114.000000	40114.000000	40114.000000	40114.000000
mean	0.306247	0.599714	0.559131	44.464302	119.810783	0.193903	0.456827	2.456232e+05	-9.141892	0.094101	0.360149	0.180956
std	0.340959	0.264198	0.178867	15.477318	30.622201	0.161792	0.247000	1.103774e+05	6.156216	0.101841	0.480049	0.325070
min	0.000000	0.000792	0.059600	0.000000	34.347000	0.009670	0.000000	1.550900e+04	-47.046000	0.022300	0.000000	0.000000
25%	0.020200	0.433000	0.443000	34.000000	94.859500	0.096900	0.259000	1.904800e+05	-10.851000	0.036100	0.000000	0.000000
50%	0.145000	0.642000	0.570000	45.000000	119.693500	0.126000	0.449000	2.273730e+05	-7.290000	0.049000	0.000000	0.000159
75%	0.550000	0.815000	0.688000	56.000000	140.205000	0.243000	0.648000	2.756732e+05	-5.191000	0.099500	1.000000	0.150000
max	0.996000	0.999000	0.986000	99.000000	220.276000	1.000000	0.992000	4.497994e+06	3.744000	0.942000	1.000000	0.996000

2. histogram 圖表 (Feature 13 經過 label encoding)
   經圖表觀察，有發現 Feature 8 的資料有嚴重偏斜，所以採取以下四種方式處理:
   1. 不處理
   2. log (對數) 轉換
   3. 平方根 轉換
   4. box-cox 轉換



特徵縮放 (Feature Scaling)

- 在分群問題中，對資料做特徵縮放有助於模型訓練的效率和準確度。原因如下：

    1. 可以使每個特徵的尺度在同一個標準底下，避免某些特徵主導整體分群的結果。
    2. 特徵縮放可以幫助這些分群演算法更快地收斂。
    3. 有些演算法，如 KMeans，是基於距離或相似度進行模型訓練和預測。特徵縮放可以確保這些算法對於各特徵的重要性更均衡，從而提高模型的準確性。

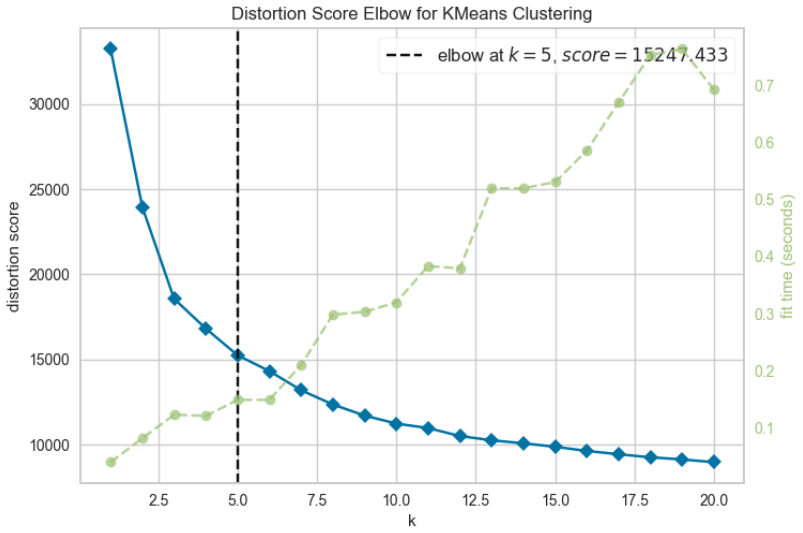
- 特徵縮方的方法：

    1. Standardization：平均值為 0、標準差為 1 的常態分佈。
    2. Normalization：將資料特徵縮放到一個特定的範圍，通常是 0 到 1 或 -1 到 1 之間。

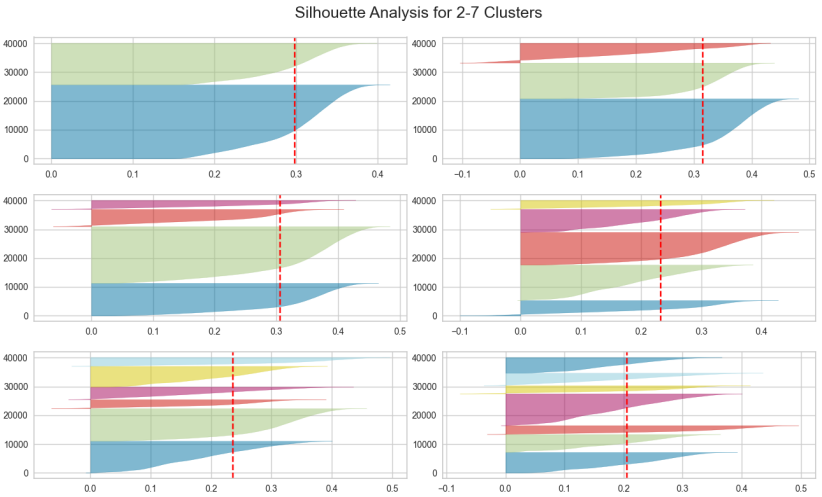
模型訓練

採用的分群演算法有以下選擇：

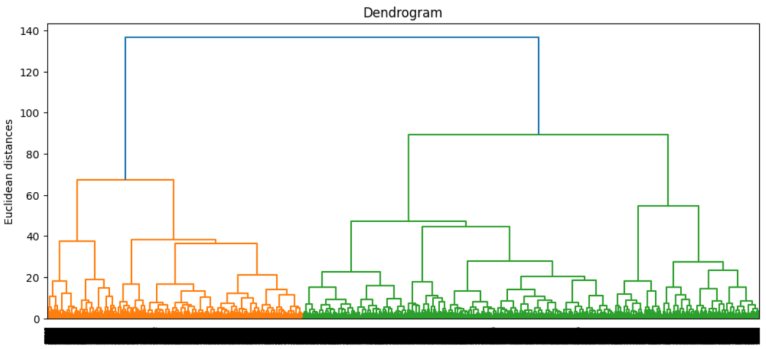
1. KMeans Clustering：
    a. 利用 inertia 做模型評估，決定要分群的 k 為多少 ⇒ Elbow 的位置大約在 4、5



    b. 使用 silhouette_scores，計算 k = 2~7 的輪廓分數，k = 3 的分群效果最好，但是 k = 4 時分數並不會低太多，反而 k = 5 時，分數剩下 0.23。



2. Agglomerative Clustering：
    a. 先做 Dendrogram 樹狀分析，觀察認為選擇 k = 3 進行分群。



3. DBSCAN Clustering：
    訓練時間長，且表現非常不好，參數難以調整，故不使用此演算法

模型測試

因為屬於分群問題，所以直接提交至 Kaggle 評估訓練效果，以下為比較的內容：

1. 資料偏斜處理：

   在只修改資料偏斜問題的選項中，得出的結果如下：

選項	public 評分結果
log (對數) 轉換	0.66666
平方根 轉換	0.67
box-cox 轉換	0.66916
不處理 ⇒ 最佳	0.67125

   ➢ 結論：不處理資料偏斜問題。

2. 特徵縮放：

   在只修改 Standardization 或 Normalization 的情況，得出的結果如下：

選項	public 評分結果
Normalization	0.56916
Standardization ⇒ 最佳	0.67125

   ➢ 結論：使用 **Normalization** 會使結果非常糟糕，所以使用 **Standardization** 對資料進行 **feature scaling**。

3. PCA 做特徵轉換：

   嘗試以 PCA 做特徵轉換，選用的特徵數量為 6，分群數量為 3。public 評分結果為 0.63458 ⇒ 並沒有使用原始特徵資料好。

   ➢ 結論：在本專案下，使用 **PCA** 無法幫助模型做特徵選擇。

4. 選擇特徵：

經過多次選擇不同欄位作為訓練的特徵，選擇級果最好的特徵組合，下表為不同特徵組合的評分結果：

欄位組合	public 評分結果
[‘Feature 1’, ‘Feature 2’, ‘Feature 3’, ‘Feature 4’, ‘Feature 5’, ‘Feature 7’, ‘Feature 9’, ‘Feature 10’, ‘Feature 11’, ‘Feature 12’,‘Feature 13’]	0.59666
[‘Feature 1’, ‘Feature 4’, ‘Feature 8’, ‘Feature 9’] ⇒ 次佳	0.67125
[‘Feature 4’, ‘Feature 8’, ‘Feature 9’] ⇒ 最佳	0.67291

➢ 結論：下兩組的表現接近，認為可以透過其他資料處理方式使結果更好。

5. Embedded：

嘗試使用多個模型進行分群，想法是把每個模型的結果綜合，取聯集。下表為 Embedded 與沒有 Embedded 的結果：

模型組合	public 評分結果
KMeans (k = 3) + agglomerative (k = 3)	0.655
KMeans ⇒ 最佳	0.67291

➢ 結論：使用 **Embedded** 的方法可能會導致誤判了更多的狀況，若使用兩個模型的交集，會失去許多原本正確的狀況，單純使用 **KMeans** 的結果最佳