數位系 110919030 汪怡廷

機器學習 HW3_Clustering

程式碼語言:PYTHON

內容:

1. 資料清理與視覺化圖表 (參考程式碼 Part 1)

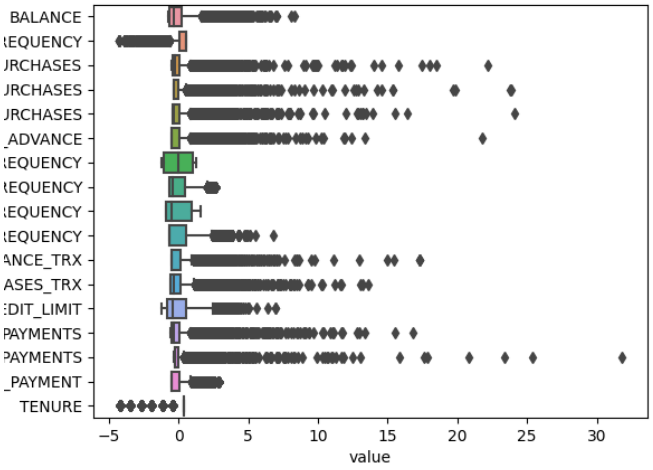
   。透過 info()跟 shape 得知 CREDIT_LIMIT 和 MINIMUM_PAYMENTS 特徵有一些缺失值。將缺失值整列去掉，從(8950, 18)變成(8636, 18)。

   。丟掉不必要的欄位

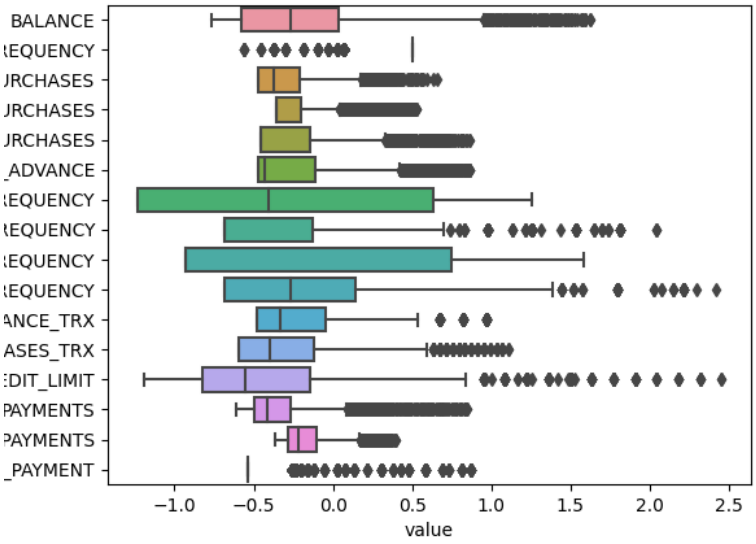
```
data.drop('CUST_ID', axis=1, inplace=True)
```

   。用盒鬚圖檢查離群值再刪除離群之後檢查盒鬚圖

   (下圖為未刪除離群值的盒鬚圖)



   (下圖為刪除離群值的盒鬚圖)



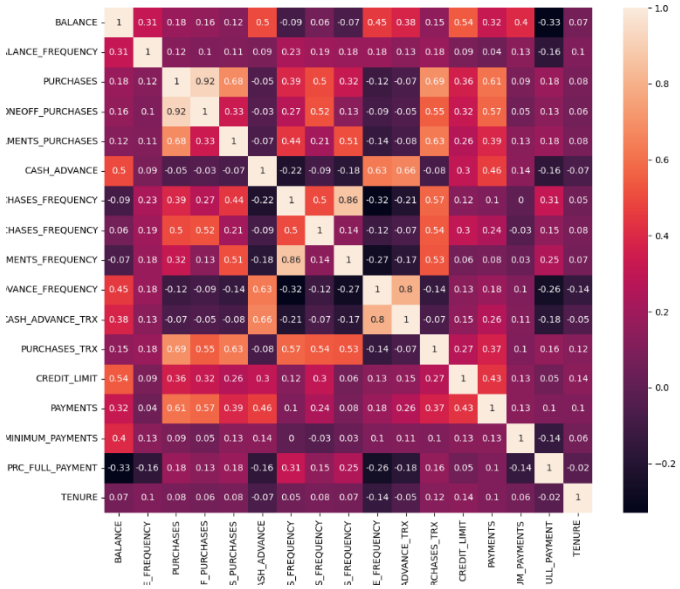
◦ 標準化數據

```
# 使用 StandardScaler 進行標準化
scaler = preprocessing.StandardScaler().fit(data)
data_scaled = scaler.transform(data)
```

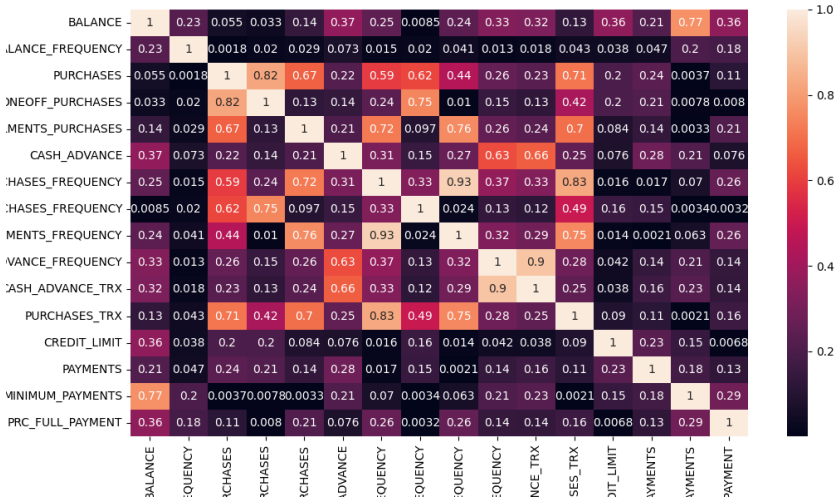
2. 敘述性統計分析(參考程式碼 Part 2)

```
3. mean = data.mean()
4. std = data.std()
5. normed_data = (data - mean) / std
```

3.特徵相關性分析(參考程式碼 Part 3)



上圖為沒有去掉離群前的 Heatmap



上

上圖為去掉離群後的 Heatmap

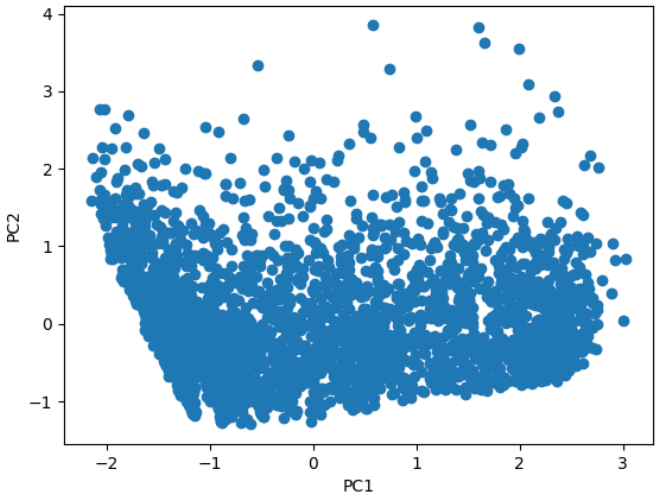
在此程式碼中有曾經為了要提高輪廓係數，試圖將部分相關係數較低的特徵資料 drop 掉，但結果卻導致輪廓係數進一步降低，因此將其註解掉。

```
# # 丟掉相關性相對不高的
# newdata=normed_data.drop(['BALANCE_FREQUENCY','PAYMENTS'], axis=1)
```

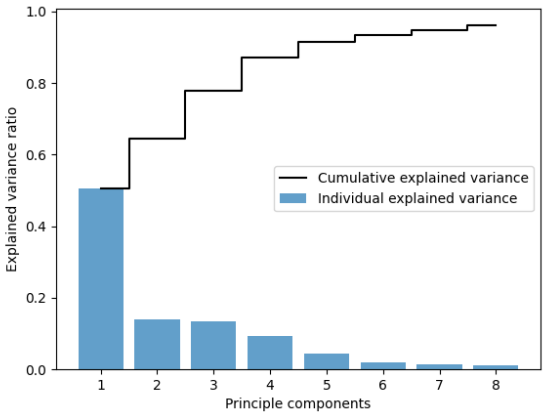
4.PCA 降維處理與分析(參考程式碼 Part 4)

使用 PCA 進行降維，將原始數據集轉換為兩個主成分（PC1 和 PC2）。這兩個主成分可以視為新的特徵軸，將原始數據映射到二維空間中。我們可以通過可視化 PC1 和 PC2 在散點圖上的分佈來觀察數據的結構和聚集情況。

另外，我們還計算了每個主成分的特徵值（explained variance）和解釋變異比例（explained variance ratio）。特徵值反映了每個主成分對數據變異性的貢獻程度，解釋變異比例則表示每個主成分解釋了數據變異性的百分比。以下式得出得圖形:



將原本 17 維的特徵資料降維成 8 維，並取前兩維畫分群狀況的散佈圖，去了解前兩維特徵資料的分群效果



```
print(pca.explained_variance_) # 特徵值
print(pca.explained_variance_ratio_) # 解釋變異比例
print(pca.explained_variance_ratio_.sum())
```

```
[2.03824158 0.55937417 0.53738171 0.37177868 0.17560235 0.07749492
 0.05651153 0.05128696]
[0.50615999 0.13891034 0.13344891 0.09232443 0.04360763 0.01924445
0.0140336  0.01273618 0.01053509 0.00844015]
0.9794407587726157
```

如上圖所述，其 explained_variance 為降維後每個新特徵向量上所帶信息量大小(可解釋變異的大小)

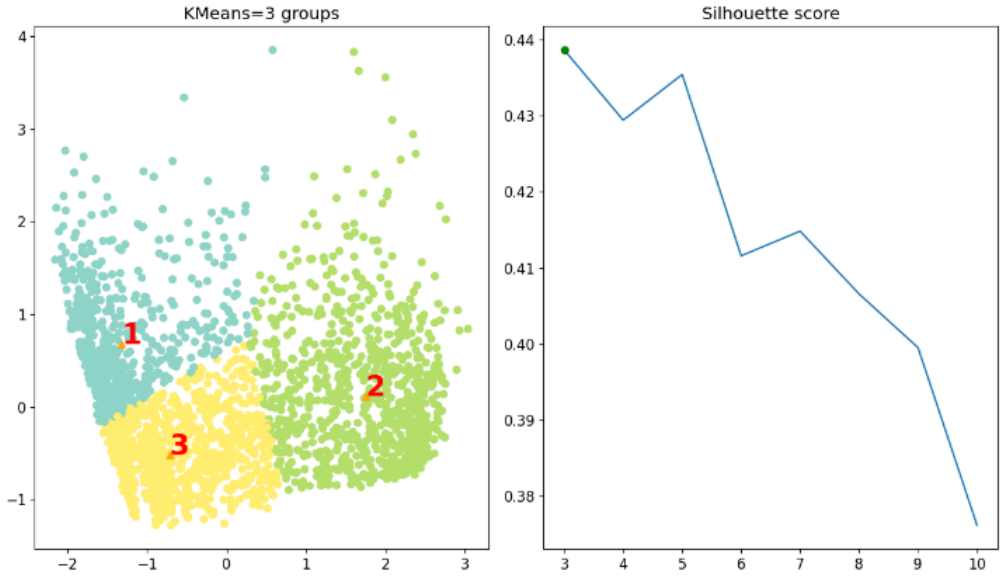
pca.explained_variance_ratio_可解釋變異比例。為了讓 explained_variance 跟 explained_variance_ratio_數值高一點，有試過調整不同的維度來讓 pca.explained_variance_ratio_.sum()更接近 1。

5.資料分割與建置 3 個分群模型(參考程式碼 Part 5)

由 PCA 分析可得知前兩為佔比較多，因此在切割測試集以及訓練集方面，只以前兩維的特徵資料為代表

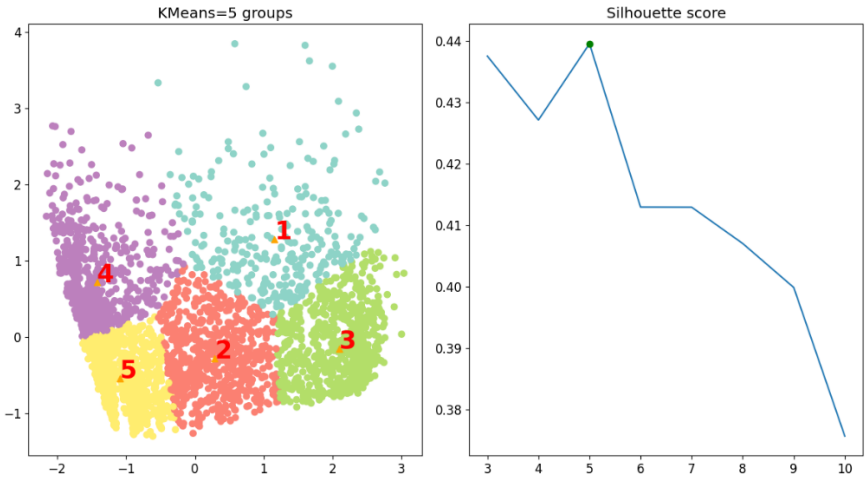
```
train_x,test_x = train_test_split(pca_data[:,:2], test_size=0.2,
random_state=42)
```

1. K-means，以轉折圖決定的集群數量進行分群，統計每個集群的大小，繪製散佈圖，並計算輪廓係數、調整蘭德指數。(參考程式碼 Part 5 中的 K-means)



圖上的左圖為使用前兩維的特徵資料進行畫分佈圖，右圖則為不同群在輪廓

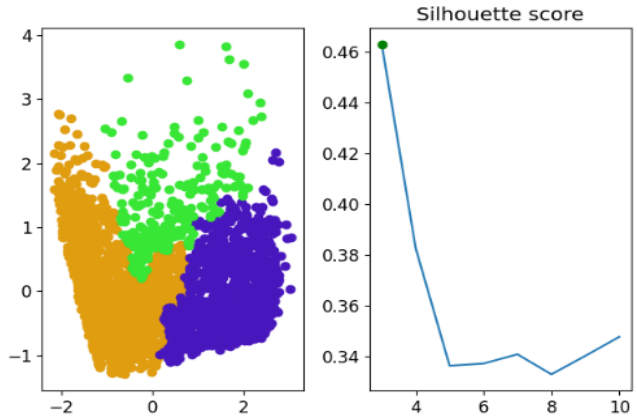
係數的表現

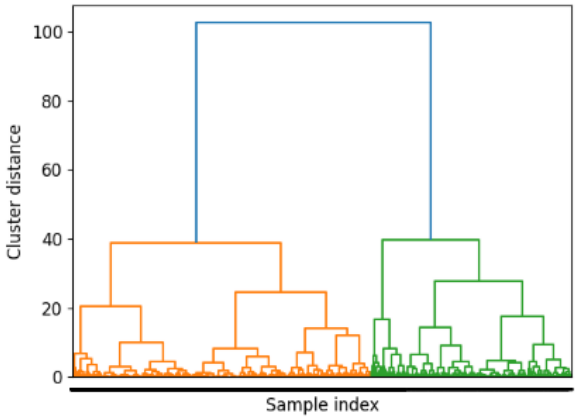


由圖上可以得知:最佳分群結果為 3 群，我們取 pca 降維後特徵資料中解釋變異比例最高的前兩維來進行 kmeans 分群，另外我們的 train_x 跟 test_x 的輪廓係數分別為 0.44016622135753336，0.42677290866026835

不過經過比較，我發現沒有使用資料分割進行分群的話會有 5 群，可能是因為去掉離群值加上資料分割後造成資料量太少。
由於沒有分割資料集的輪廓係數或是分類情況都表現得較好，因此後面的資料也是以不分割資料的結果為主

2. Hierarchical Clustering (Agglomerative) 採用 ward linkage 策略，統計每個集群的大小，繪製散佈圖與樹狀圖，並計算輪廓係數、調整蘭德指數。
(參考程式碼 Part 5 中的 Hierarchical Clustering)
這段程式碼使用 DBSCAN 演算法對 train_x 進行分群，並選擇 eps 和 min_samples 參數來調整分群的結果。然後，它去除了噪音點，並評估非噪音點的分群品質。最後，它將非噪音點的分群結果以散點圖的形式繪製

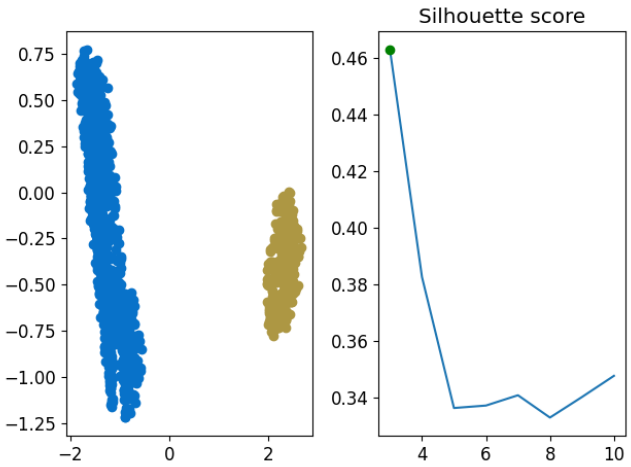




輪廓係數:0.4405208518739707，可分成 3 群

3. DBSCAN，統計每個集群的大小，繪製散佈圖，並計算輪廓係數、調整蘭德指數。(參考程式碼 Part 5 中的 DBSCAN)

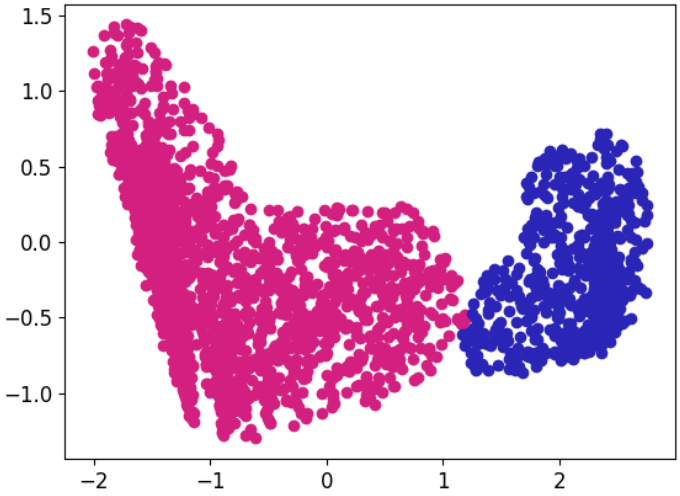
這個方法需要自行測試 eps,min_sample，但因為會產生噪點的關係，因此需要將噪點刪除。但如果刪掉太多噪點，分佈圖會呈現較好的分類情況，且輪廓係數達到 0.85，但印出 shape 會發現資料量從原本的 3000 多筆直接刪到 1000 多。下圖為刪掉太多噪點的結果:



再調整參數之後我們得到: clf = DBSCAN(eps=0.3, min_samples=75).fit(train_x)

利用這組參數，只會刪掉 484 筆噪點，雖然輪廓係數降低到

0.6319289742405935 但我認為這個情形更接近真實的分類結果。



6. 綜合比較 3 個模型的分群結果與分析討論

   就已輪廓係數方面:

   kmeans(0.43946541752680546)
   Hierarchical Clustering (0.46285157308379377)
   DBSCAN(0.6319289742405935)

DBSCAN> Hierarchical Clustering>kmeans
從圖形分析來看:
DBSCAN> kmeans>= Hierarchical Clustering
綜合比較方面，對於具有離群值或不明顯分群結構的資料集，DBSCAN 可能更適合。另外，K-means 和層次聚類模型通常對資料的線性可分性和標準化要求較高，因此可能需要進行更多的資料預處理，而 DBSCAN 則對資料的分佈情況要求較低，對密度不連續的區域有較好的適應性。因此我認為在這個 dataset 中 DBSCAN 的分群表現較好。