資科系 110916015 林翡

機器學習HW3\_Clustering

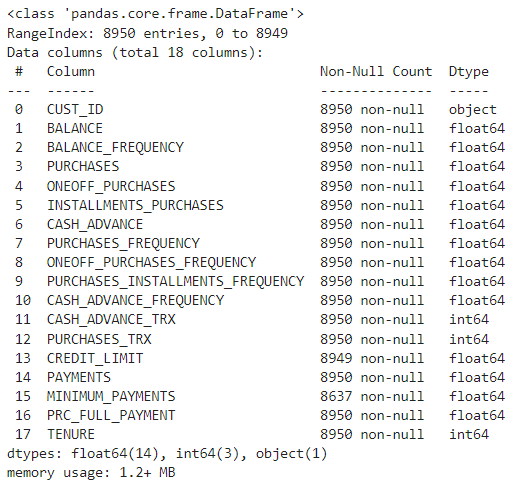
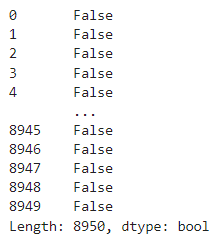
程式碼語言:Python(在jupyter lab上寫)

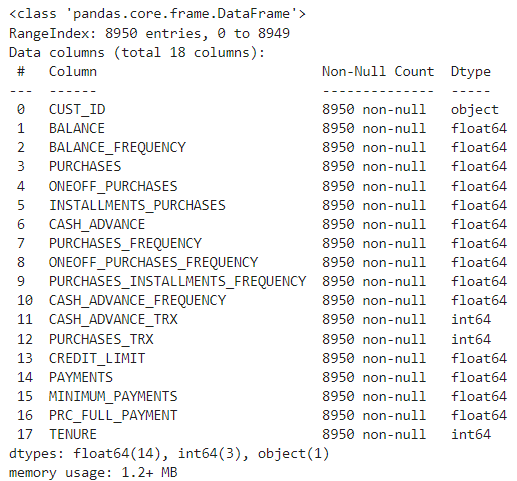
內容:

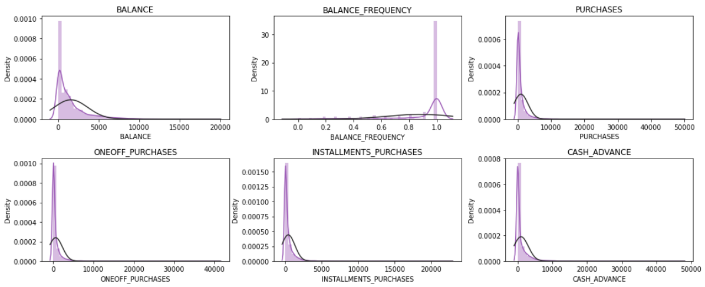
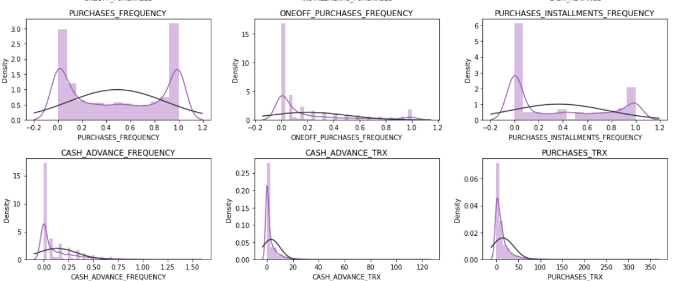
1. 資料清理與視覺化圖表 (程式碼[1]~程式碼[10])

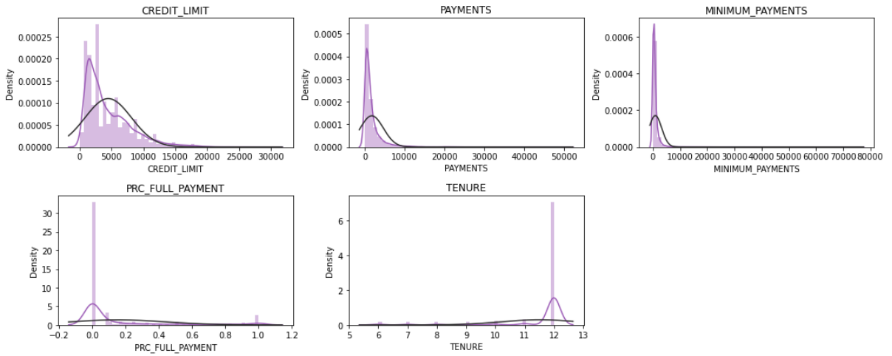
* 資料清理

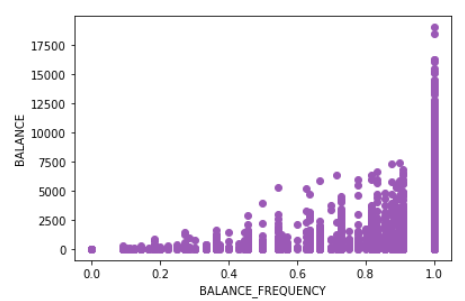
1. 檢查資料集有無遺漏值及重複值，由圖中可看出此資料集的CREDIT\_LIMIT和

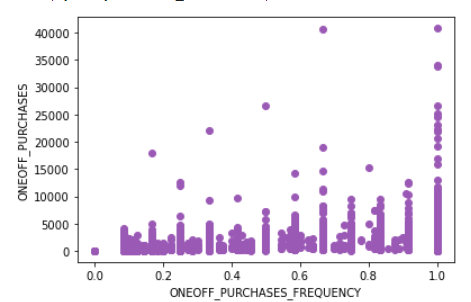
MINIMUM\_PAYMENTS兩個欄位有遺漏值，並用填入平均值的方式填補遺漏值，無重複值。 (程式碼[1]、程式碼[3]~程式碼[5])

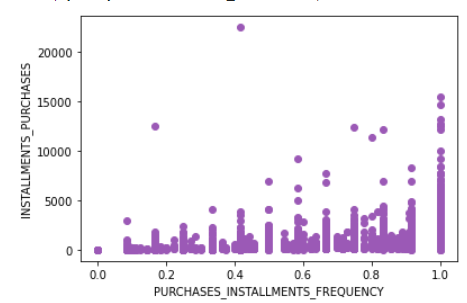
 處理遺漏值後

* 視覺化圖表 (程式碼[7]~程式碼[10])

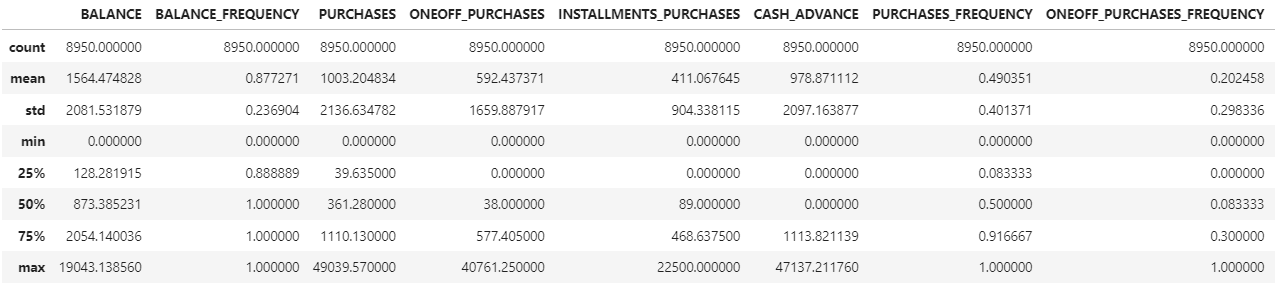
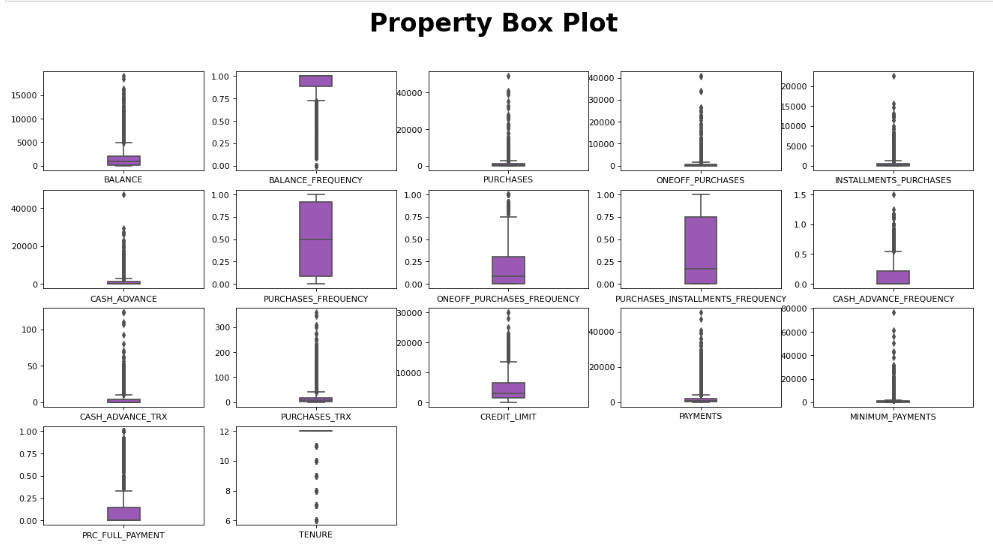
17個欄位的分佈圖

BALANCE\_FREQUENCY-BALANCE散點圖

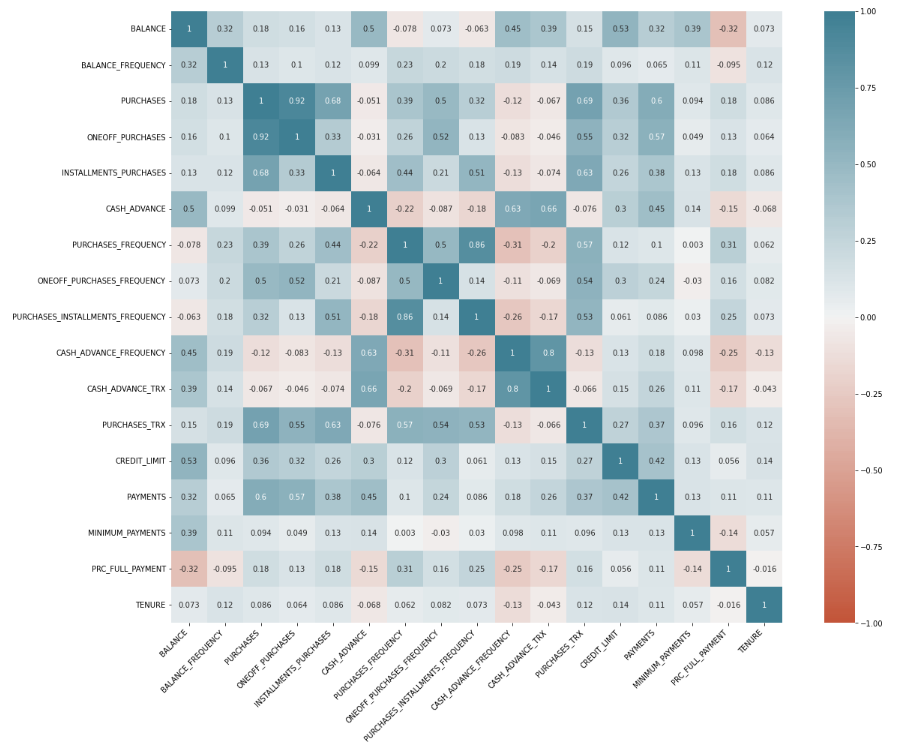
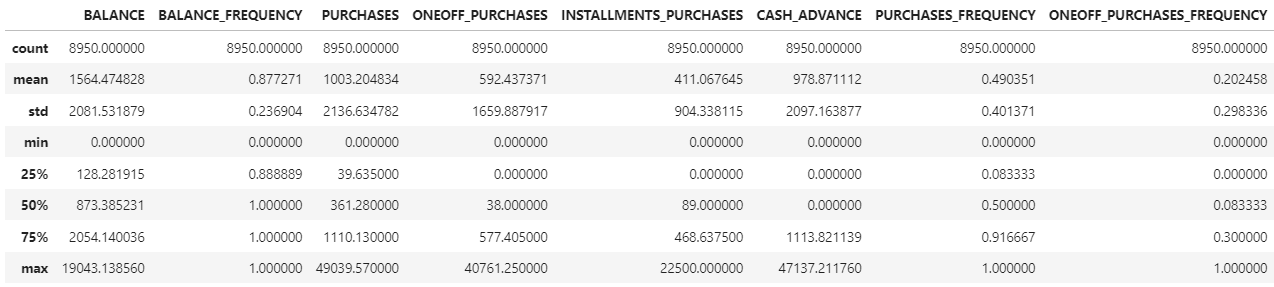
ONEOFF\_PURCHASES\_FREQUENCY-ONEOFF\_PURCHASES散點圖

 PURCHASES\_INSTALLMENTS\_FREQUENCY-INSTALLMENTS\_PURCHASES散點圖

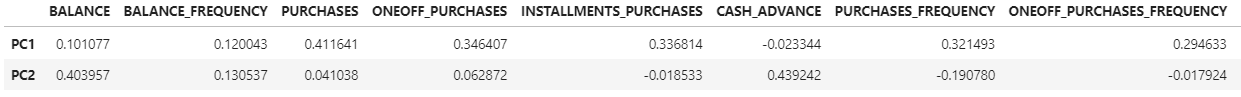
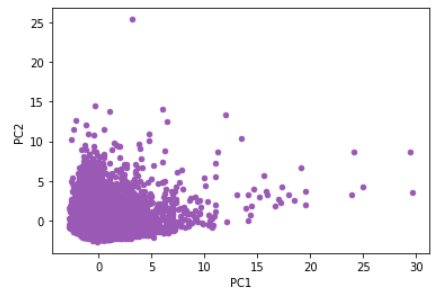
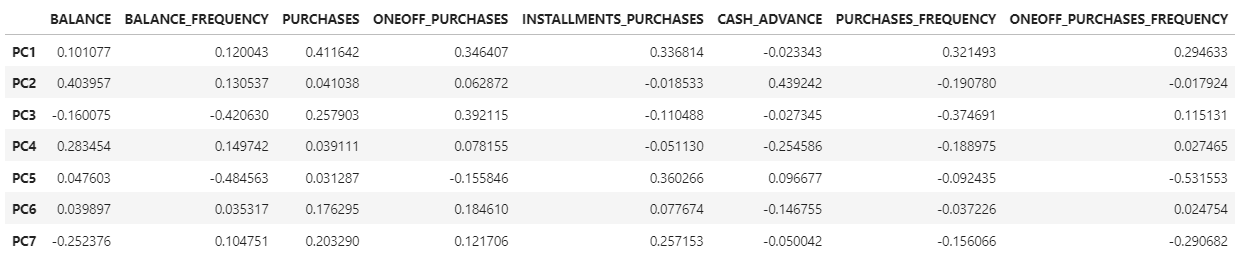
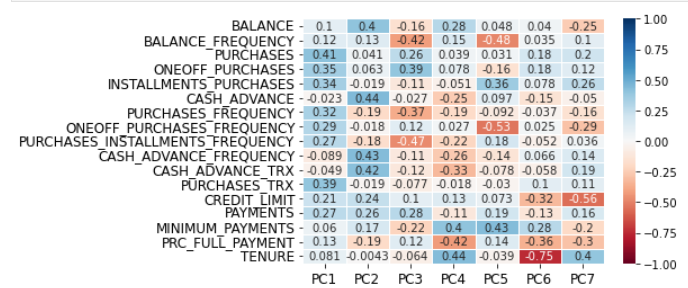
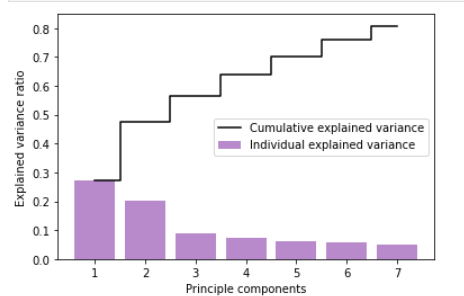
1. 敘述性統計分析 (程式碼[11]~程式碼[12])

* 敘述性統計分析表 (程式碼[11]):計算各個欄位的平均值、標準差、最小值、四分位數、中位數、最大值。
* 敘述性統計分析圖 (程式碼[18])

1. 特徵相關性分析

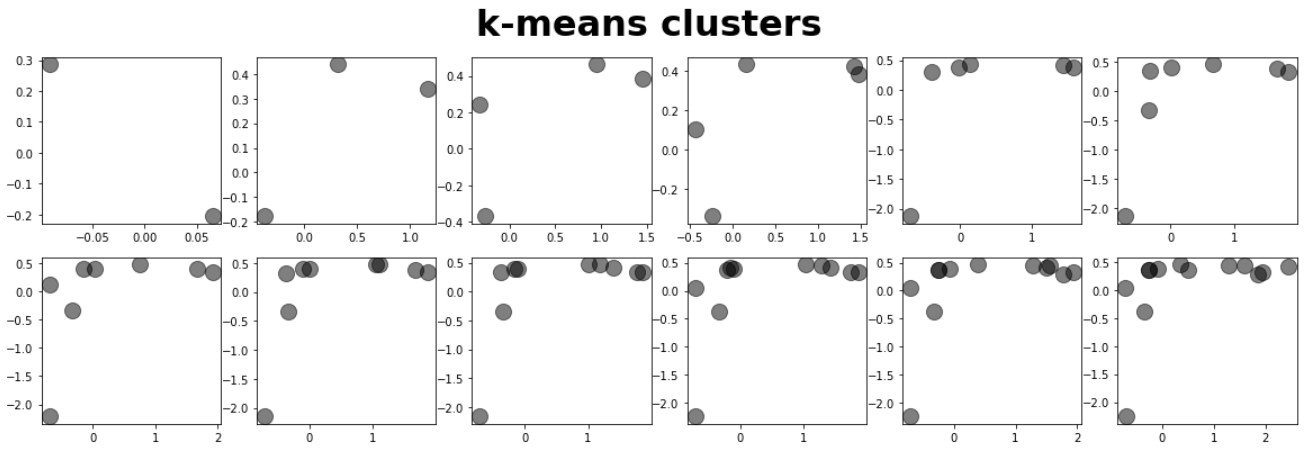
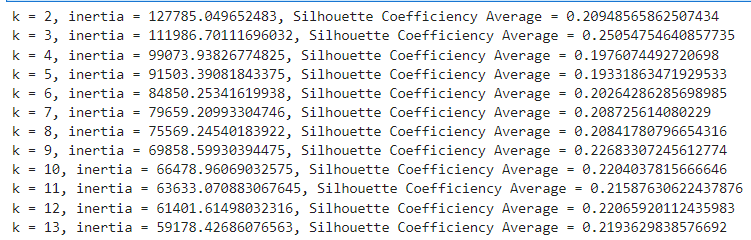
* 相關係數熱度圖 (程式碼[13])
* 特徵標準化(程式碼[14]):欄位選擇除CUST\_ID外的所有欄位做標準化。

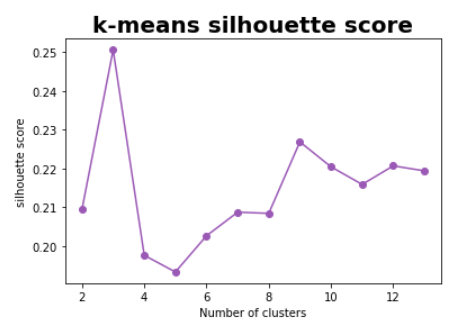
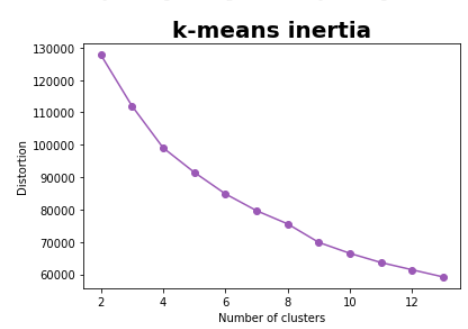
1. PCA降維處理與分析 (程式碼[15]~程式碼[20])

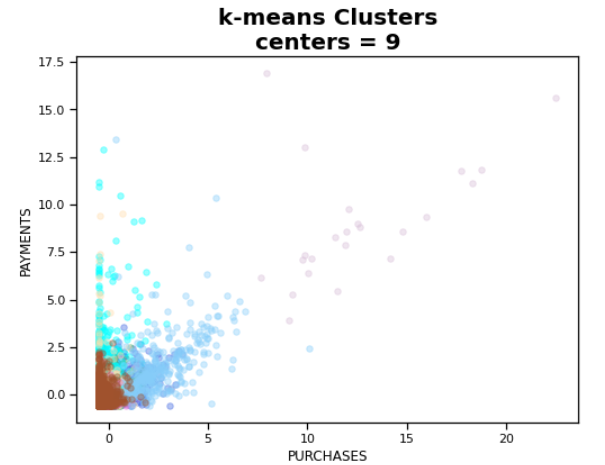
* 用PCA將數值特徵降到二維 (程式碼[15]~程式碼[16])
* 分析:從PCA結果得到的兩個主成分中可看出PURCHASES在PC1的值較大，可能有較大的正向貢獻，但我覺得其他特徵的值好像都沒有差太多，有負號的也沒有到負很多；而PC2的PURCHASES\_FREQUENCY和PRC\_FULL\_PAYMENT算是當中比較低的值。
* 解釋80%的變異性 (程式碼[18]~程式碼[20]):將數值特徵降到七維。

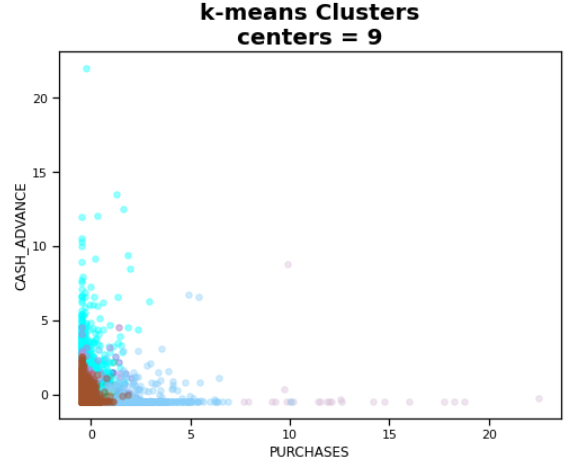
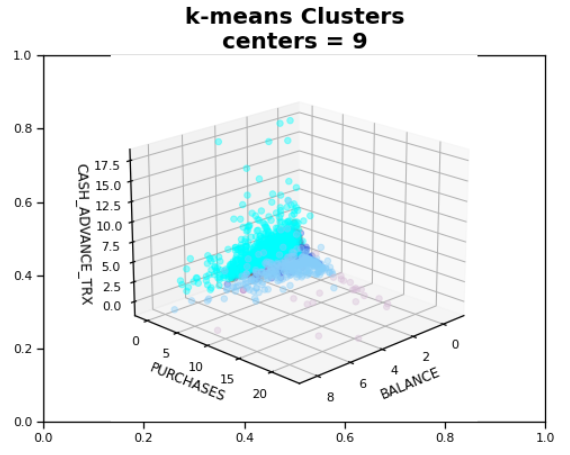
1. 資料分割與建置3個分群模型(程式碼[23]~程式碼[44])

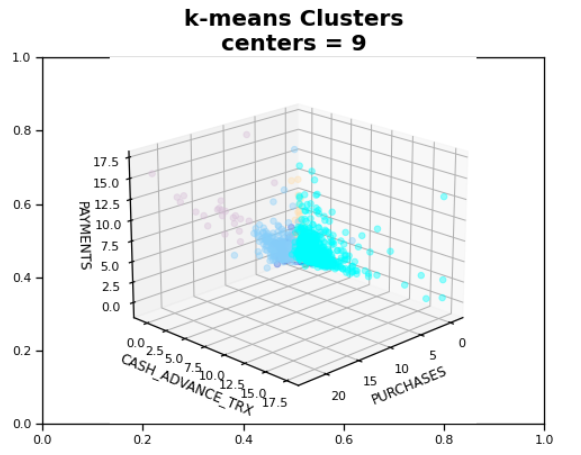
1.K-means:將標準化後的資料集(8950筆)以K-means分群，k值設為2到13並把分群結果視覺化，同時依照k值計算每個點到其最近的叢集中心的平方距離的和(inertia)及平均輪廓係數，並繪製inertia及平均輪廓係數的轉折圖。選擇k=9作為分群的群集數，再把分群結果存到clusterData[‘pred’]。特徵選擇BALANCE、PURCHASES、CASH\_ADVANCE、CASH\_ADVANCE\_TRX、PAYMENTS 這五個特徵，並用迴圈從五個特徵挑選X、Y軸繪製2D及3D散佈圖，最後繪製統計各群集大小的長條圖，繪圖部分有參考學姊的程式碼。(程式碼[24]~程式碼[30])

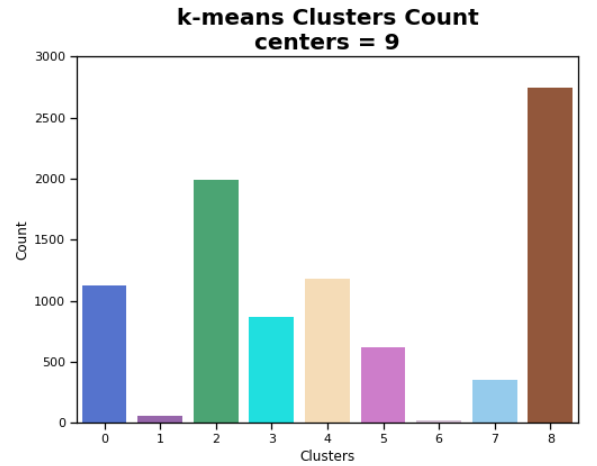
inertia及平均輪廓係數

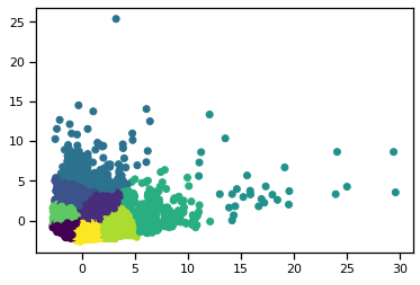


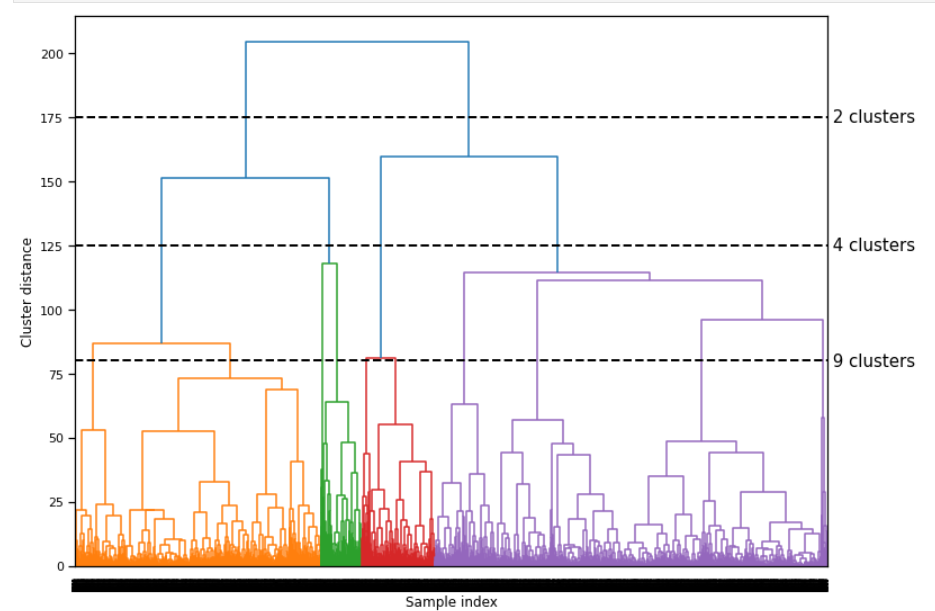
選了分群比較明顯的圖放上來，但好像都大同小異。

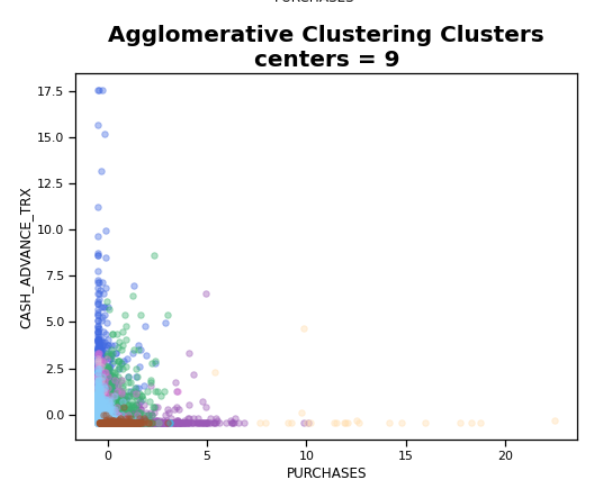
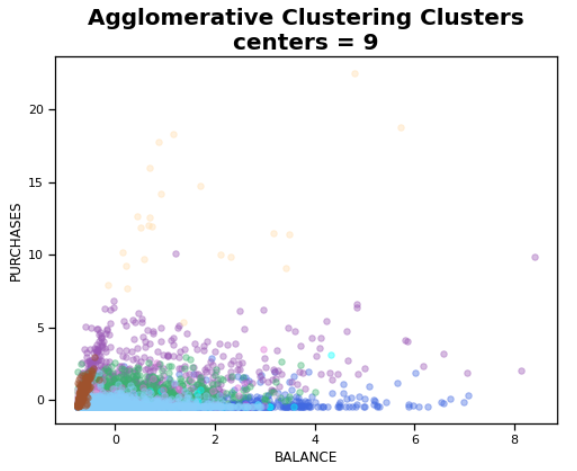


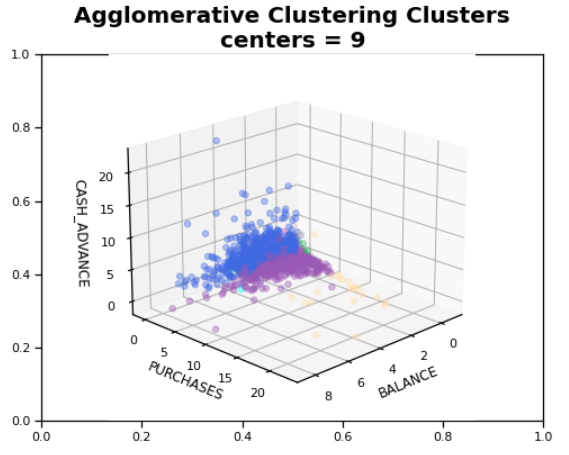
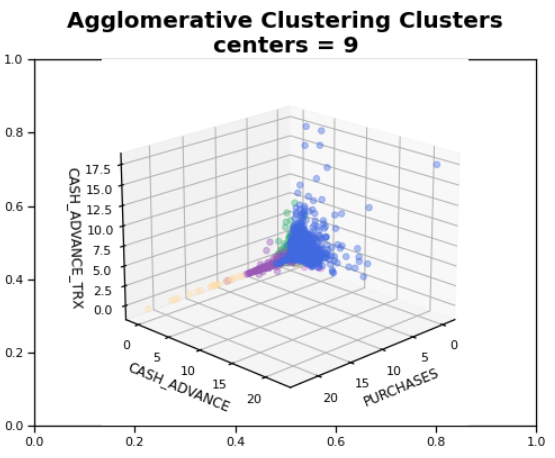
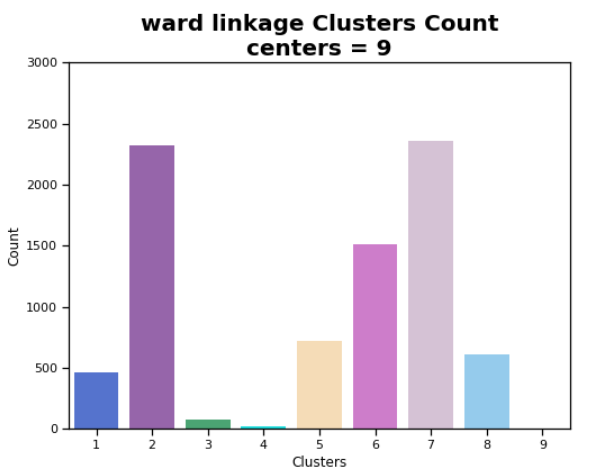


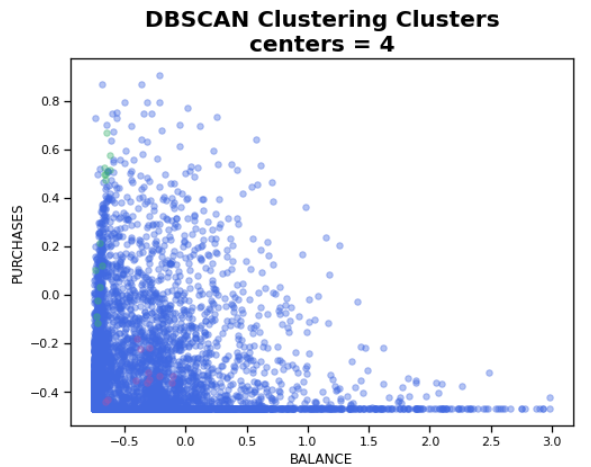
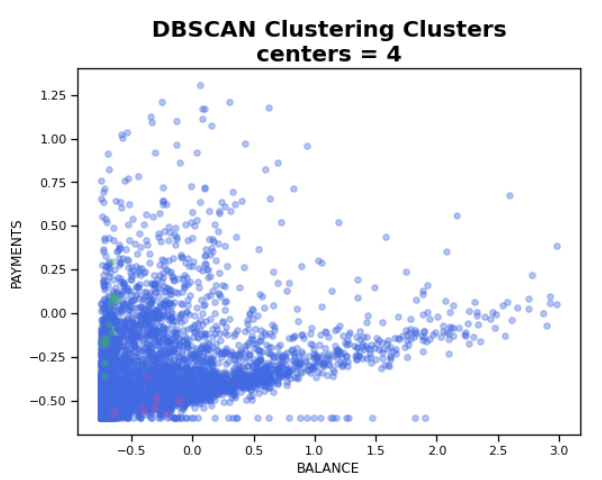
統計各群集大小的長條圖

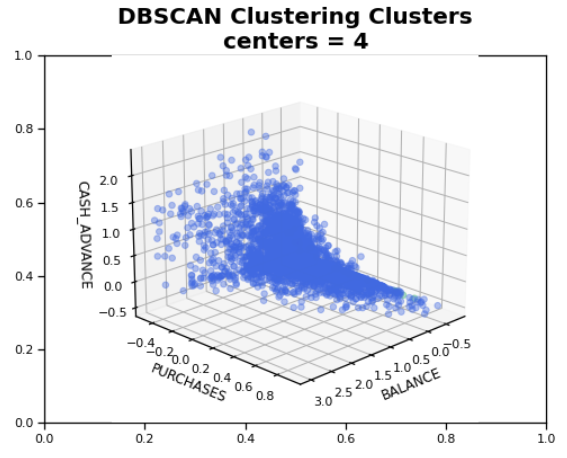
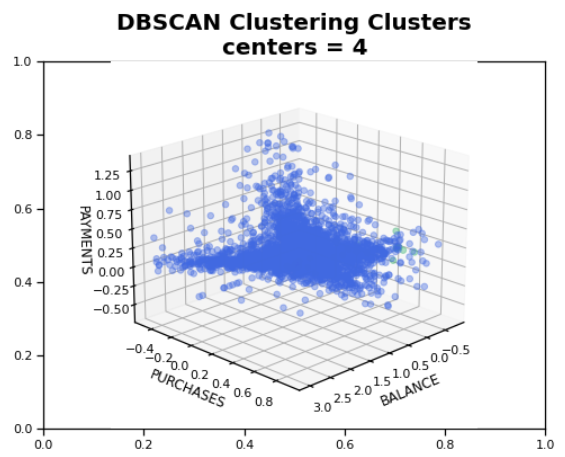
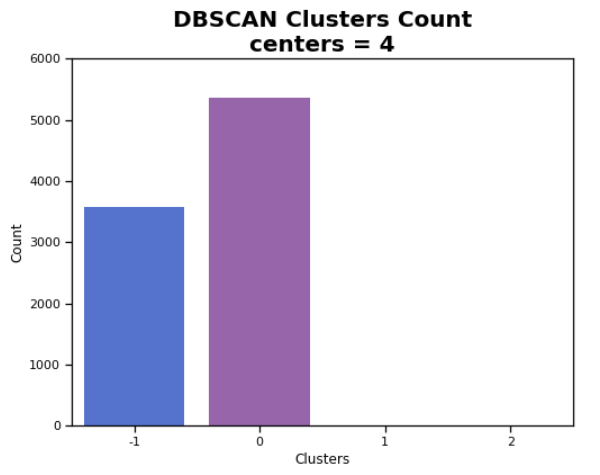
這邊多一個前面用PCA降到二維(PC1、PC2)且K-means分成9群的散點圖。

1. Hierarchical Clustering (Agglomerative): 將標準化後的資料集(8950筆)以ward linkage的凝聚分群法分群，將分群結果繪製成樹狀圖，群集數跟k-means一樣都選9群，並繪製2D、3D散佈圖及統計各群集大小的長條圖。(程式碼[31]~程式碼[38])





1. DBSCAN:一樣是將標準化後的資料集(8950筆)以DBSCAN進行分群，eps參數設為1、min\_samples設為9，繪製2D、3D散佈圖及統計各群集大小的長條圖。(程式碼[39]~程式碼[44])



1. 綜合比較3個模型的分群結果與分析討論

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | K-means | Hierarchical Clustering | DBSCAN |
| 訓練資料筆數 | 8950 | 8950 | 8950 |
| 集群數量 | 9 | 9 |
| 集群大小 |  |  |  |
| 集群分佈 |  |  |  |
| 調整蘭德指數  (CASH\_ADVANCE\_TRX當真實標籤) | 0.05801106539336735 | 0.04937301847184534 | 0.0431605786723538 |
| 輪廓係數 | 0.22683307245612774 | 0.15371895842035113 | 0.05242366252064325 |

* 分析討論:如果以調整蘭德指數來看的話，K-means的分群效果略好，但老實說我不太知道真實標籤要選哪個特徵，所以就看散佈圖哪個特徵可以分得比較明顯就選那個特徵當真實標籤，k-means的輪廓係數我拿k=9的來比較，k-means的輪廓係數也比其他兩個分群結果高。DBSCAN我如果照課本的eps設成35會8950筆資料都分到同一群，所以就設成1了，剛好分成4群，但看散佈圖都是藍色那一群，也不曉得這樣正不正常。而階層式分群的分群效果應該就趨於k-means和DBSCAN之間吧，只是在跑樹狀圖時跑得略久一些，因為我都用8950筆資料去做訓練，沒有特別隨機選幾筆出來訓練。