資科系 111016011 陳奕

機器學習HW3\_Clustering

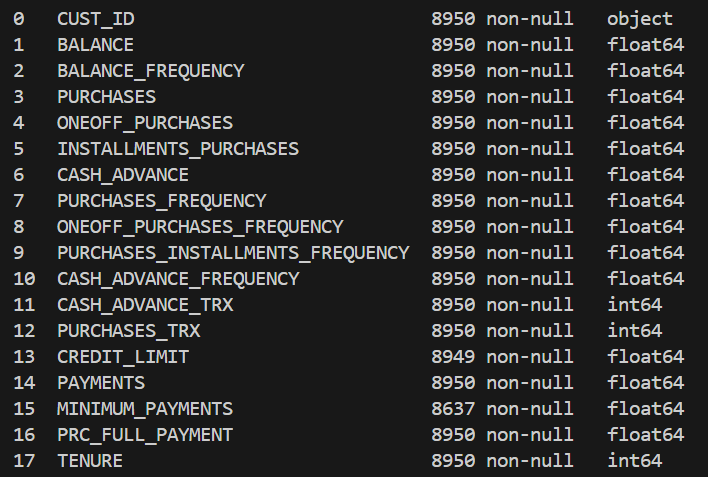
程式碼語言:

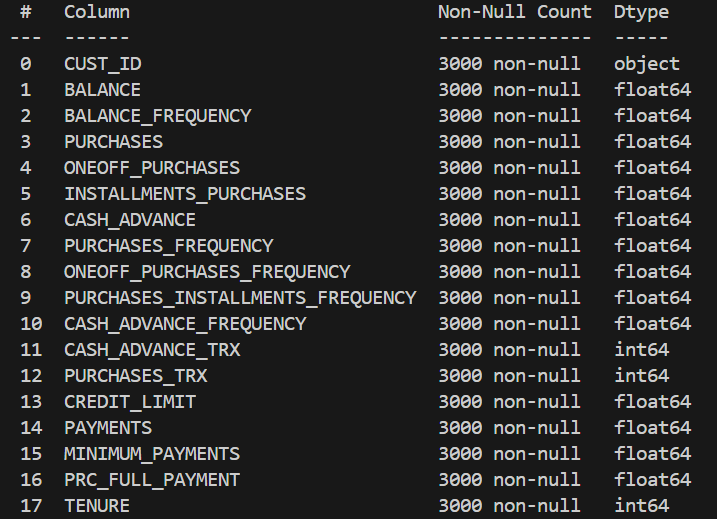
內容:

1. 資料清理與視覺化圖表 (參考程式碼Part 1)
2. 資料清理與選取

先使用function 1-1-1，發現在CREDIT\_LIMIT與PRC\_FULL\_PAYMENT有資料遺失(圖一)，

由於這次作業只需選取3000筆，所以直接將他們刪除，再使用function 1-1-2刪除代有遺失值的資料，最後用function 1-1-3取得隨機3000筆資料，將選取出來的資料放在"random\_3000\_credit\_card\_data.csv”，未來只需讀取該檔案即可，使用function 1-1-1確認沒有遺失值(圖二)，且為3000筆，同時發現除了CUST\_ID外的特徵皆為數值型。

(圖一)

(圖二)

1. 資料標準化

接下來使用function 1-2對除CUST\_ID外所有數值資料做標準化，並儲存在” standardized\_credit\_card\_data.csv”，由於未來的動作若保留CUST\_ID會不方便，所以在此先不保留，未來需要會再從” random\_3000\_credit\_card\_data.csv”取出CUST\_ID。

1. 敘述性統計分析

使用function 2-1 ，可以得出下列表格(圖三)，存於”descriptive\_statistics.csv”，值得注意的是，此處使用的資料為原始資料，並非標準化過後的資料，且此表格不包含CUST\_ID，而表格代表的含意由上到下代表：

count（計數）：非缺失值的數量。

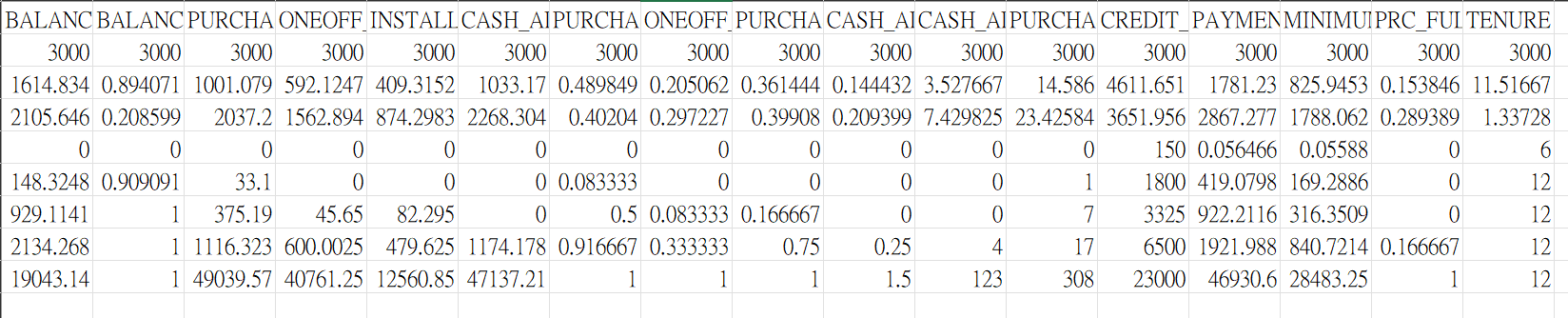
mean（平均值）：數值欄位的平均值。

std（標準差）：數值欄位的標準差。

min（最小值）：數值欄位的最小值。

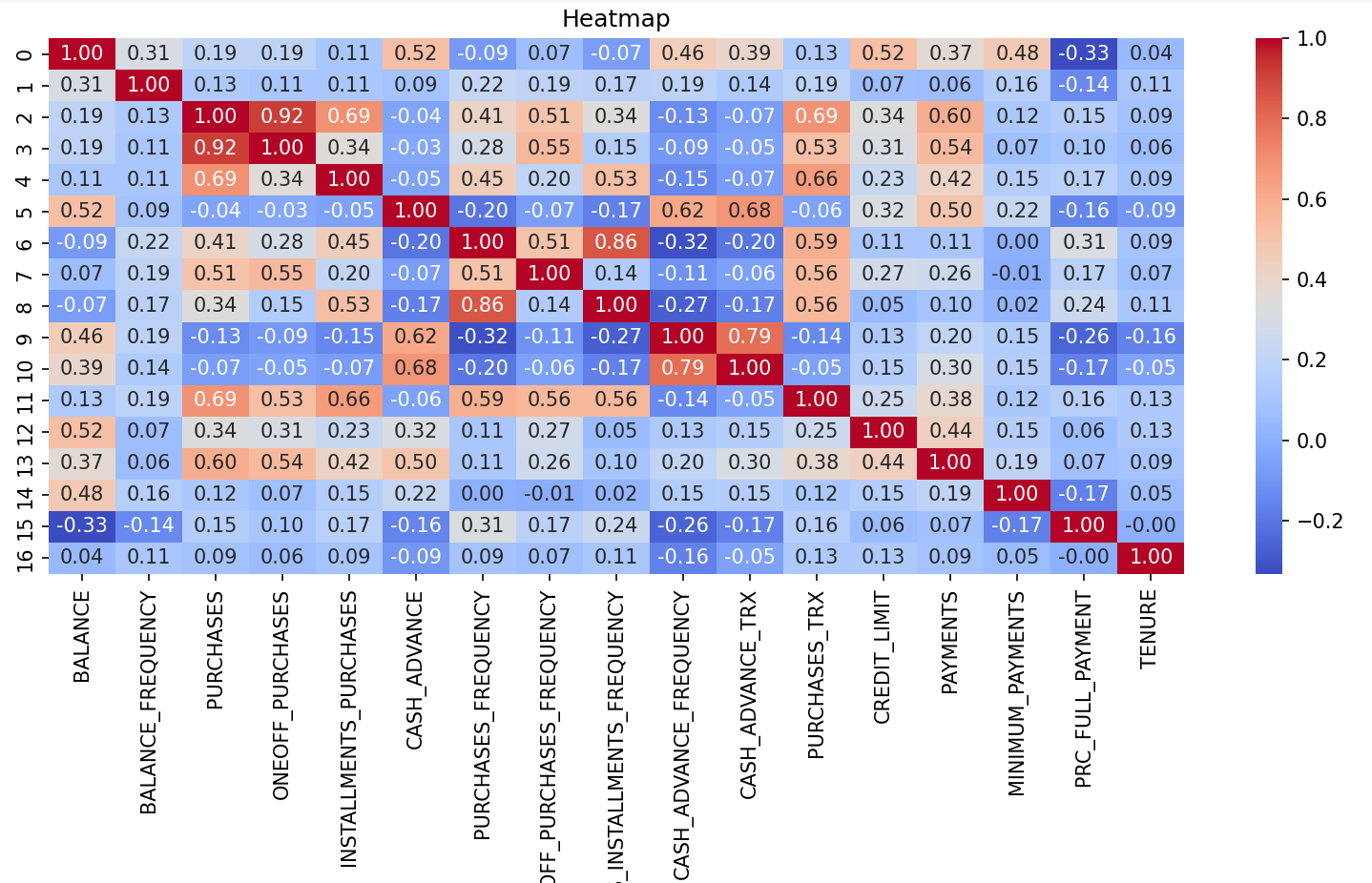
25%、50%、75%（四分位數）：四分之一、中位數和四分之三的百分位數。

max（最大值）：數值欄位的最大值。

(圖三)

1. 特徵相關性分析

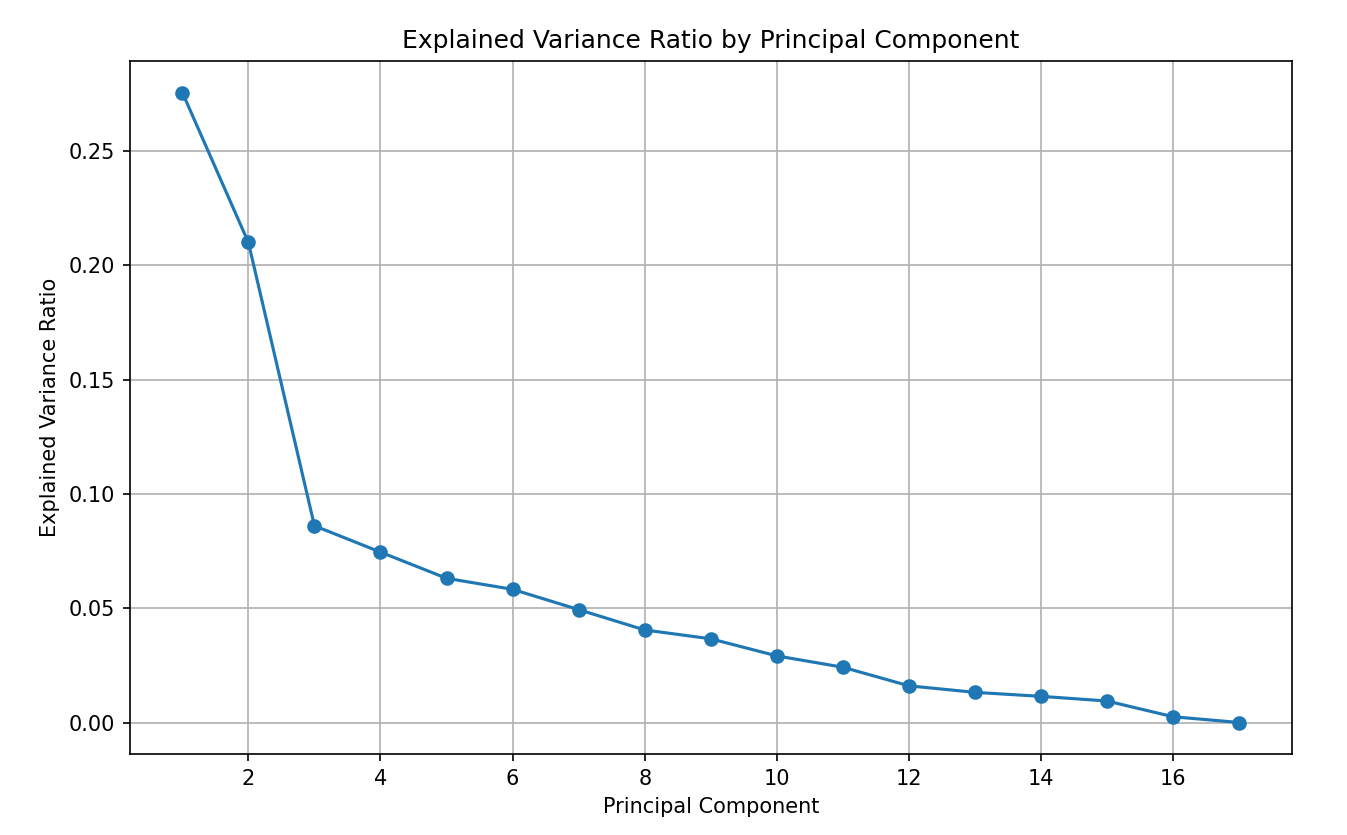
使用function 2-2-1 計算皮爾森相關係數，並用function 2-2-2畫成熱度圖(圖四)，值得注意的是，此處資料使用標準化後的資料。圖中可以發現ONEOFF\_PURCHASES與PURCHASES的相關係數極高，代表客戶單次最大購買金額與累積購買金額呈正相關，也許代表大部分客戶在單次交易中傾向於購買更高金額的物品，或者單次交易中購買金額較高的用戶可能對整體購買金額貢獻更多。

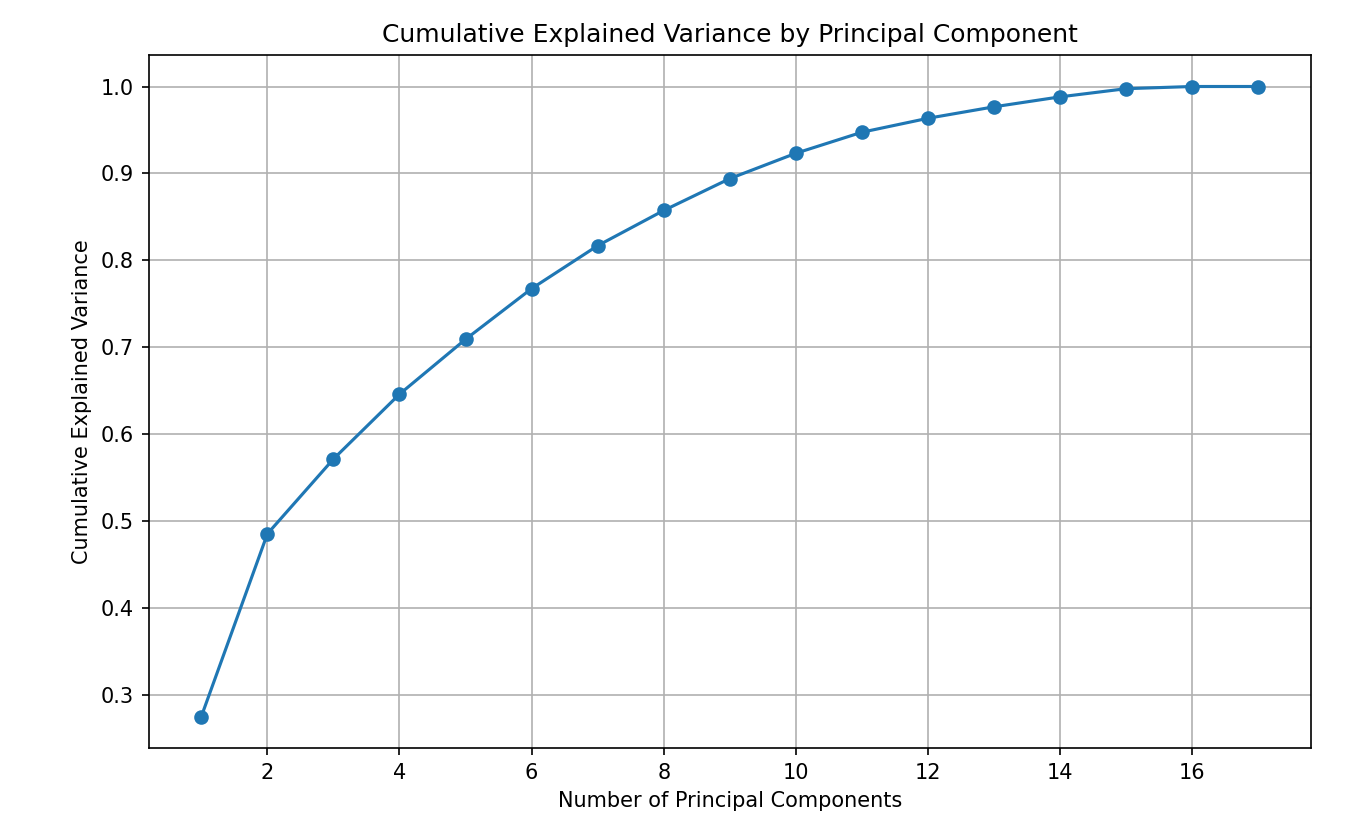


(圖四)

1. PCA降維處理與分析

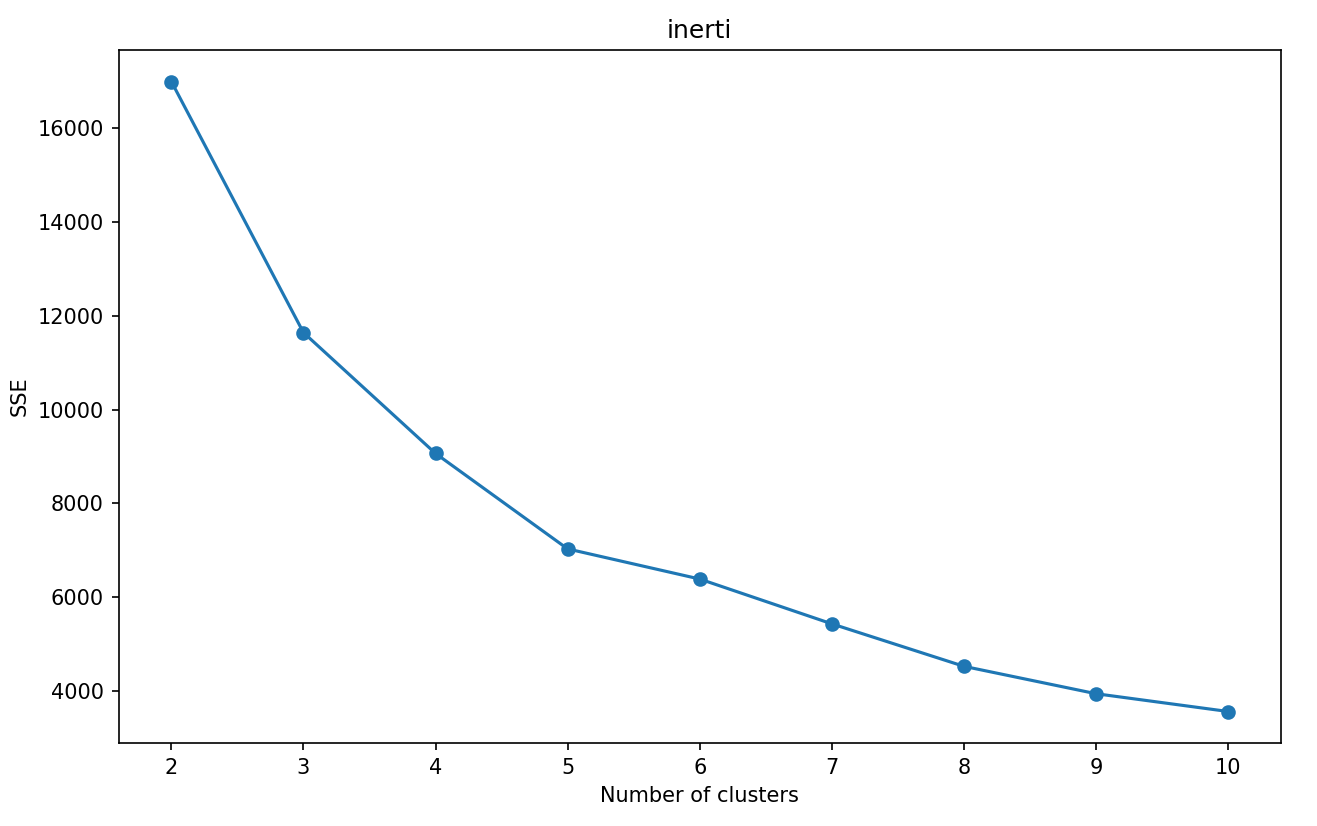
使用function 2-3-1 計算解釋變異數比率並畫出(圖五)，再用function 2-3-2 計算累進的變異數比率並畫出(圖六)，當將資料降至7維時，能解釋資料80%的變異性；若降至10維，則能達到90%，但為了提高計算效率以及可視化的需求，將資料降成2維並存於” pca\_data\_2PC.csv”

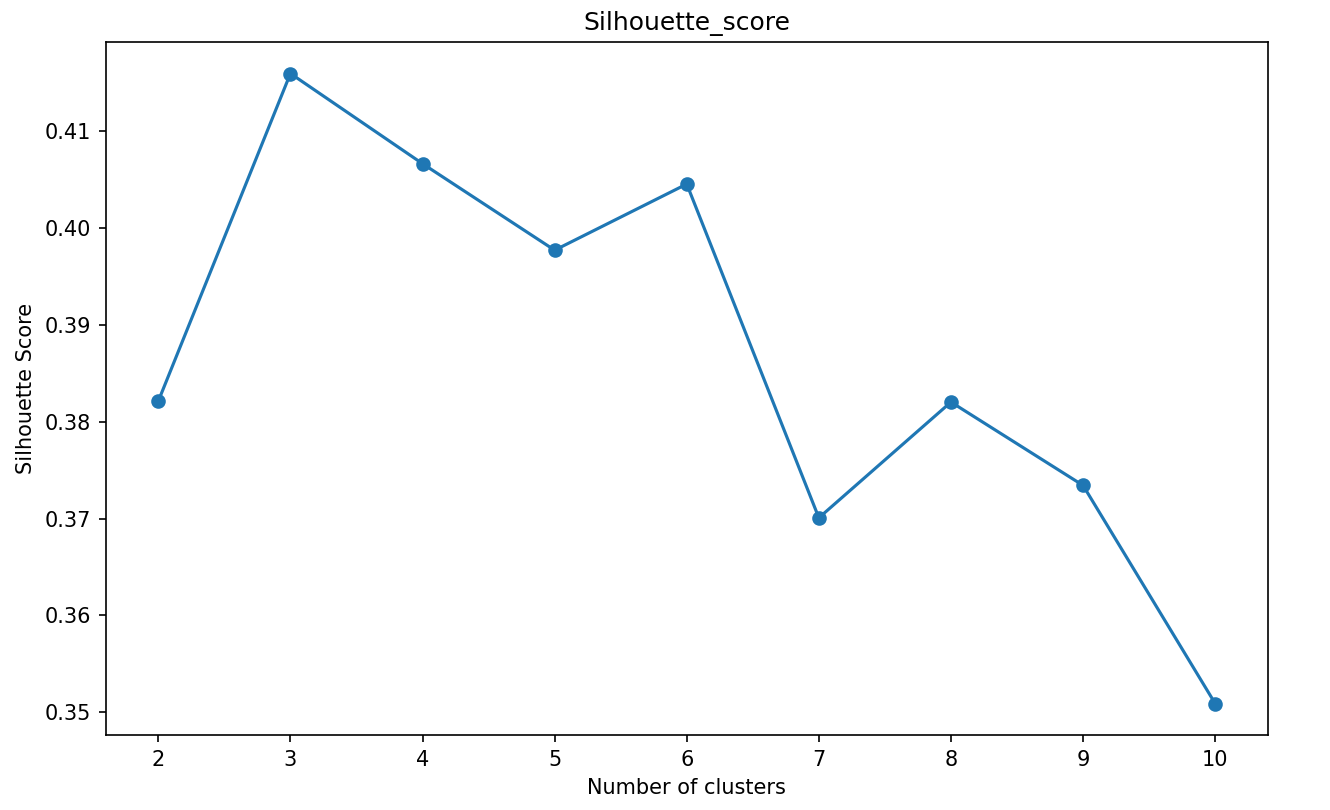
(圖五)

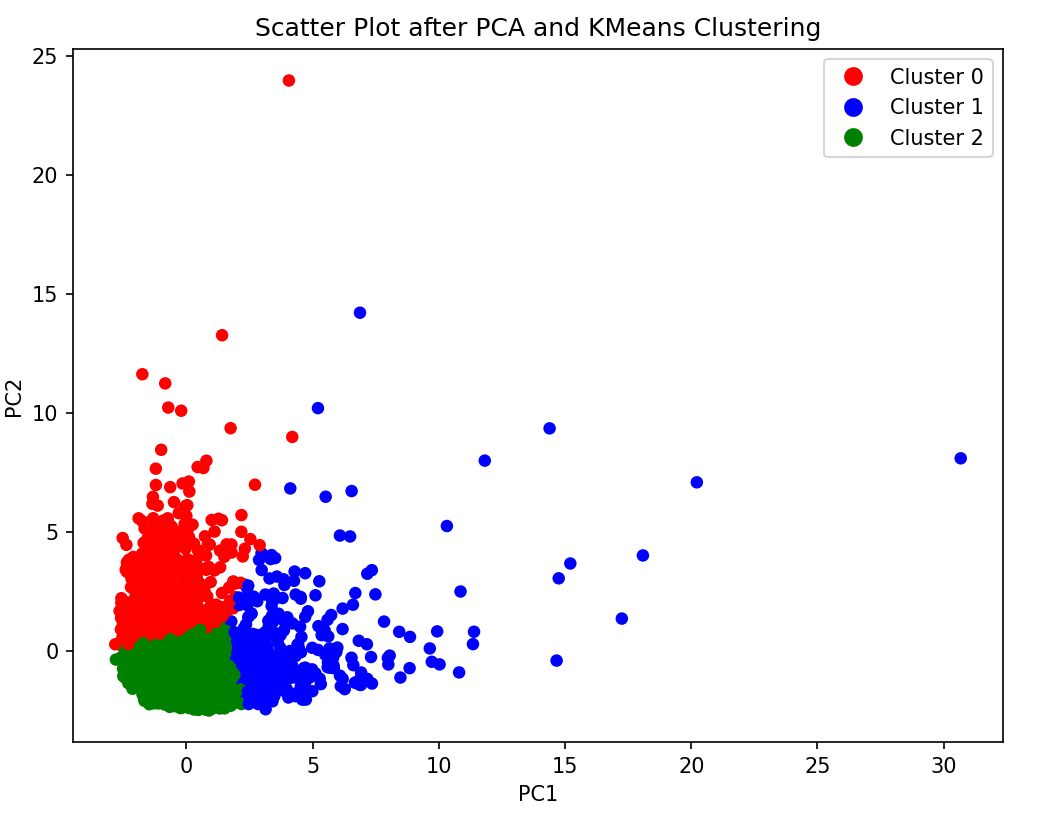
(圖六)

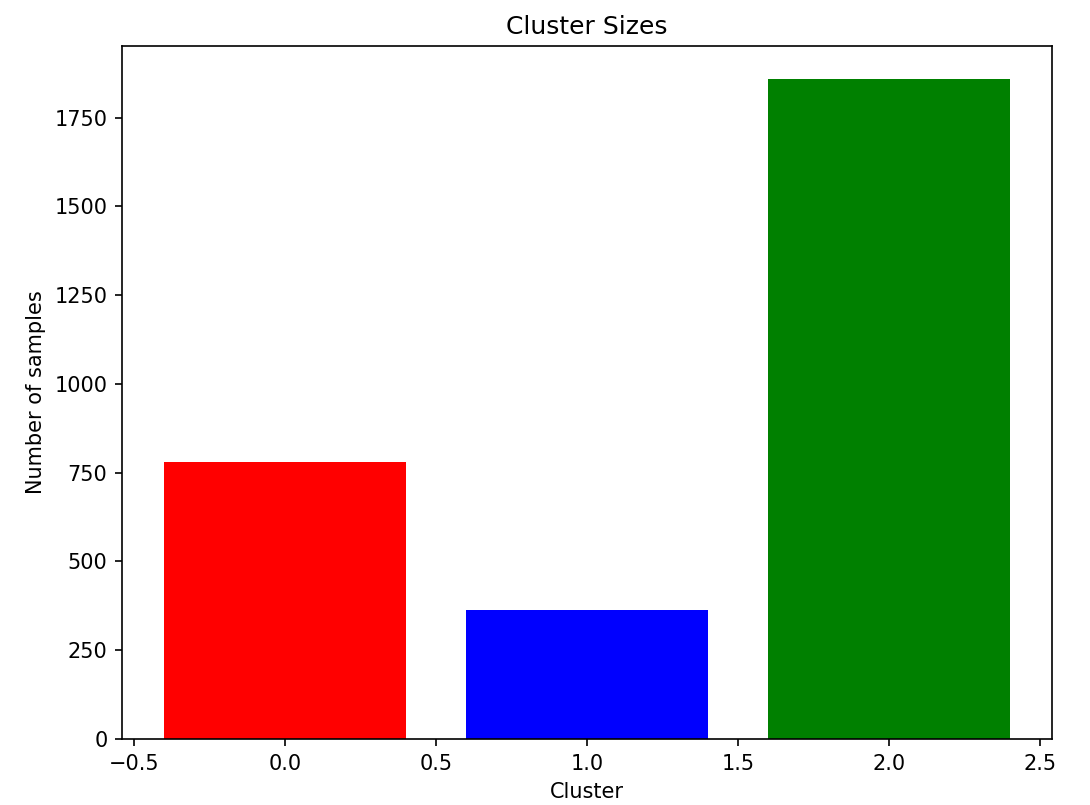
1. 資料分割與建置3個分群模型
2. K-means

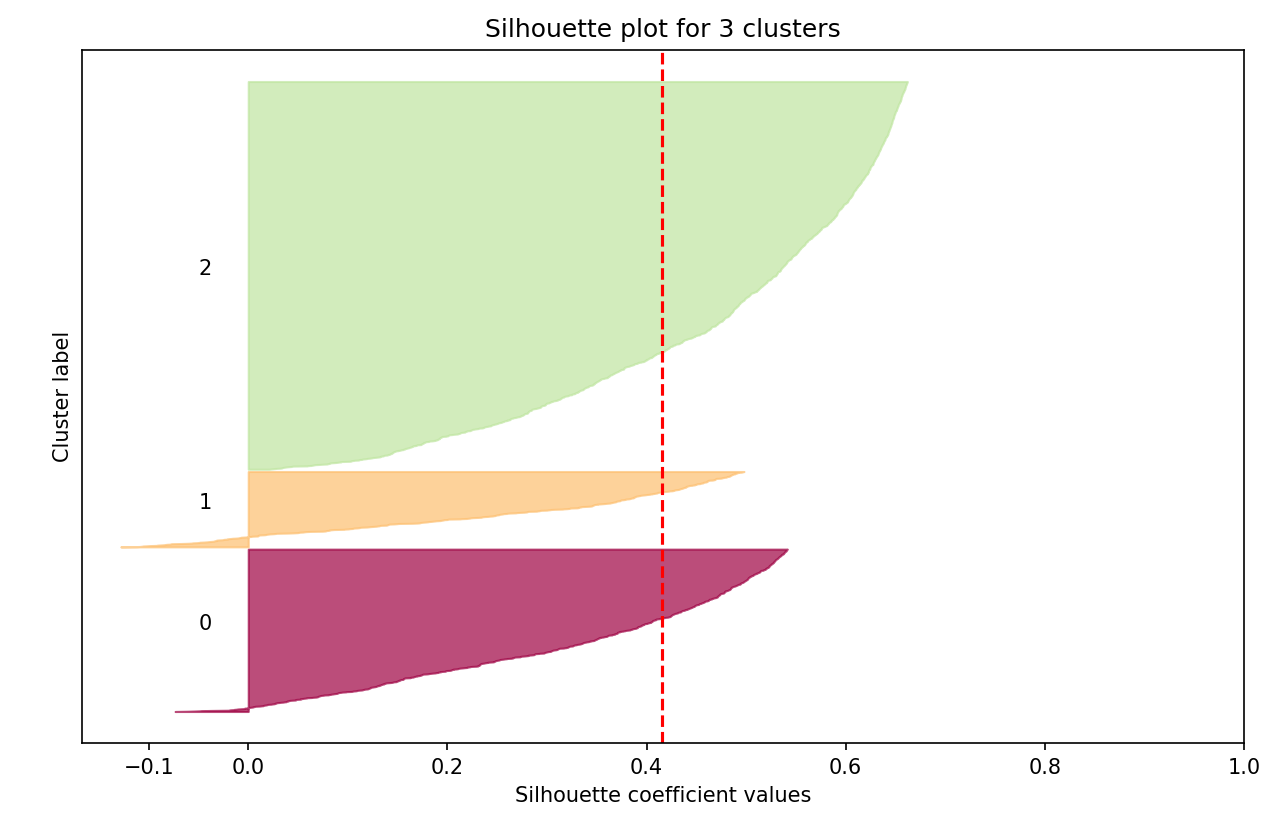
使用function 3-1-1畫出集群內的誤差平方和(圖七)以及function 3-1-2畫出的平均輪廓係數(圖八)並得出在k = 3時，會有最高值0.415。用function 3-1-3將其分成三群並畫成散佈圖(圖九)以及各群的數量(圖10)，最後用function 3-1-4畫出輪廓係數圖(圖十一)

 (圖七)

 (圖八)

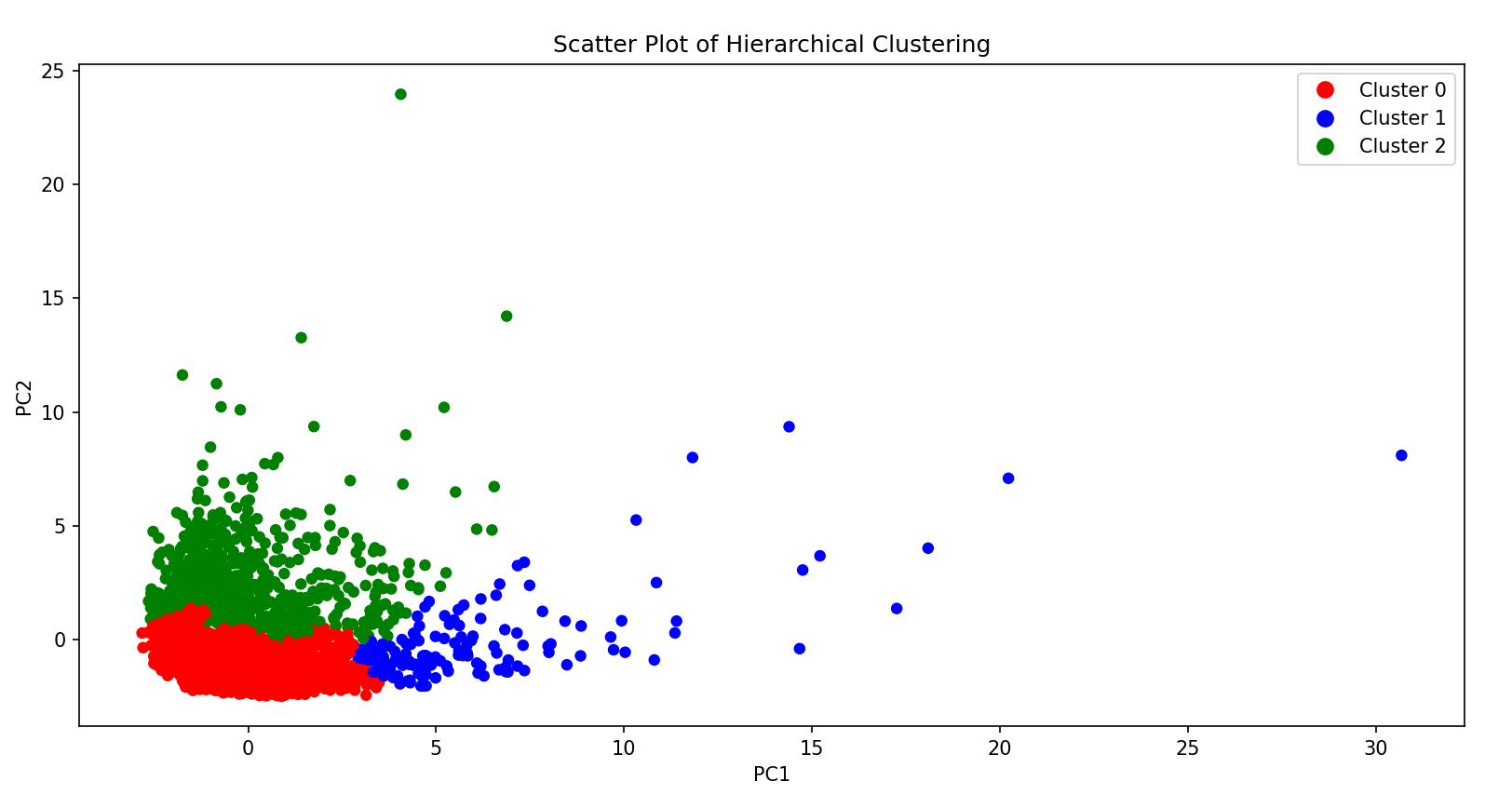
(圖九)

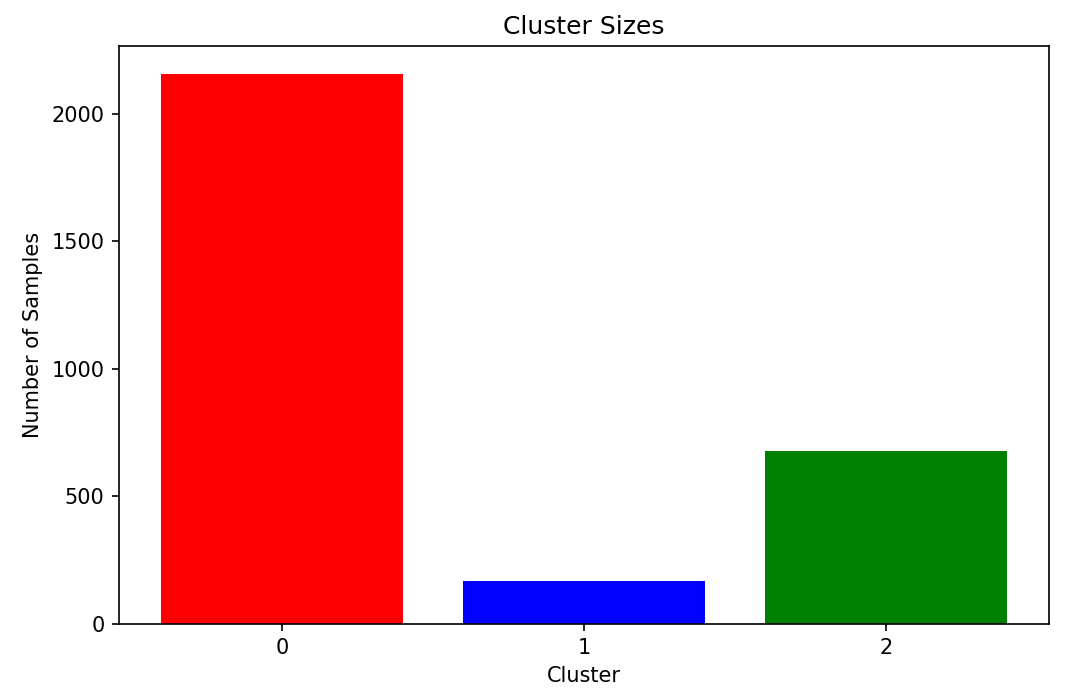
(圖十)

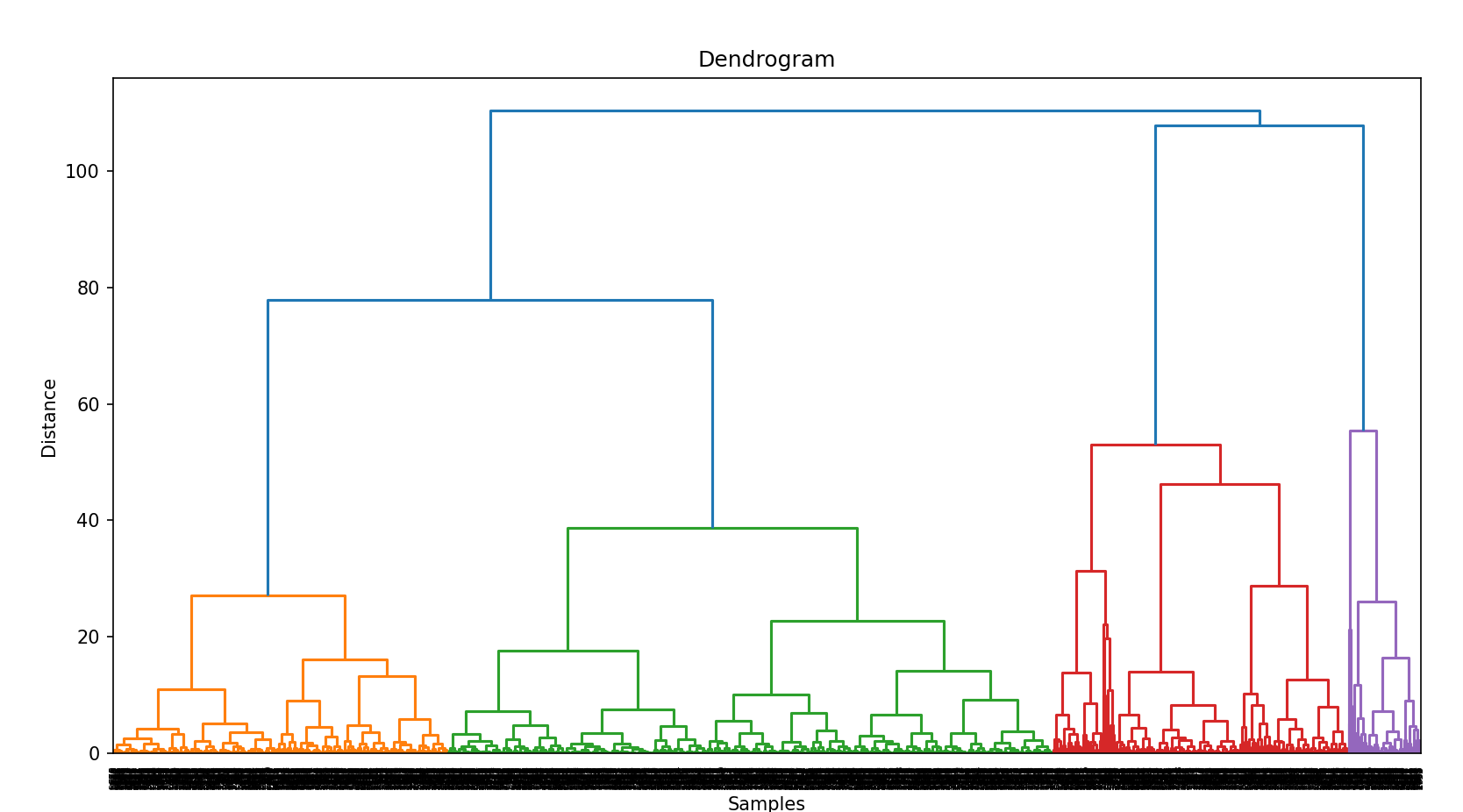
(圖十一)

1. Hierarchical Clustering (Agglomerative)

使用function 3-2畫出散佈圖(圖十二)、集群數量長條圖(圖十三)、樹狀圖(圖十四)與計算輪廓係數(0.42)，

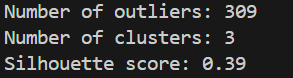
(圖十二)

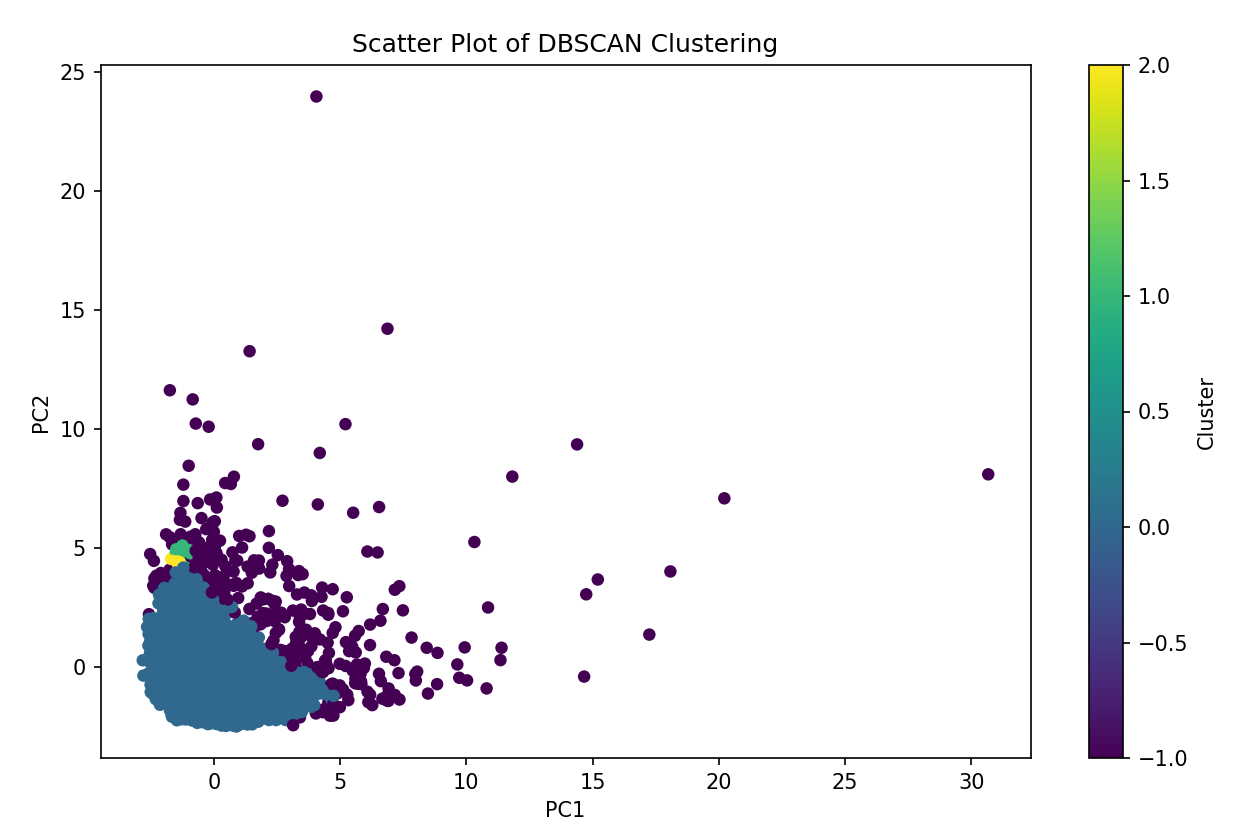
(圖十三)

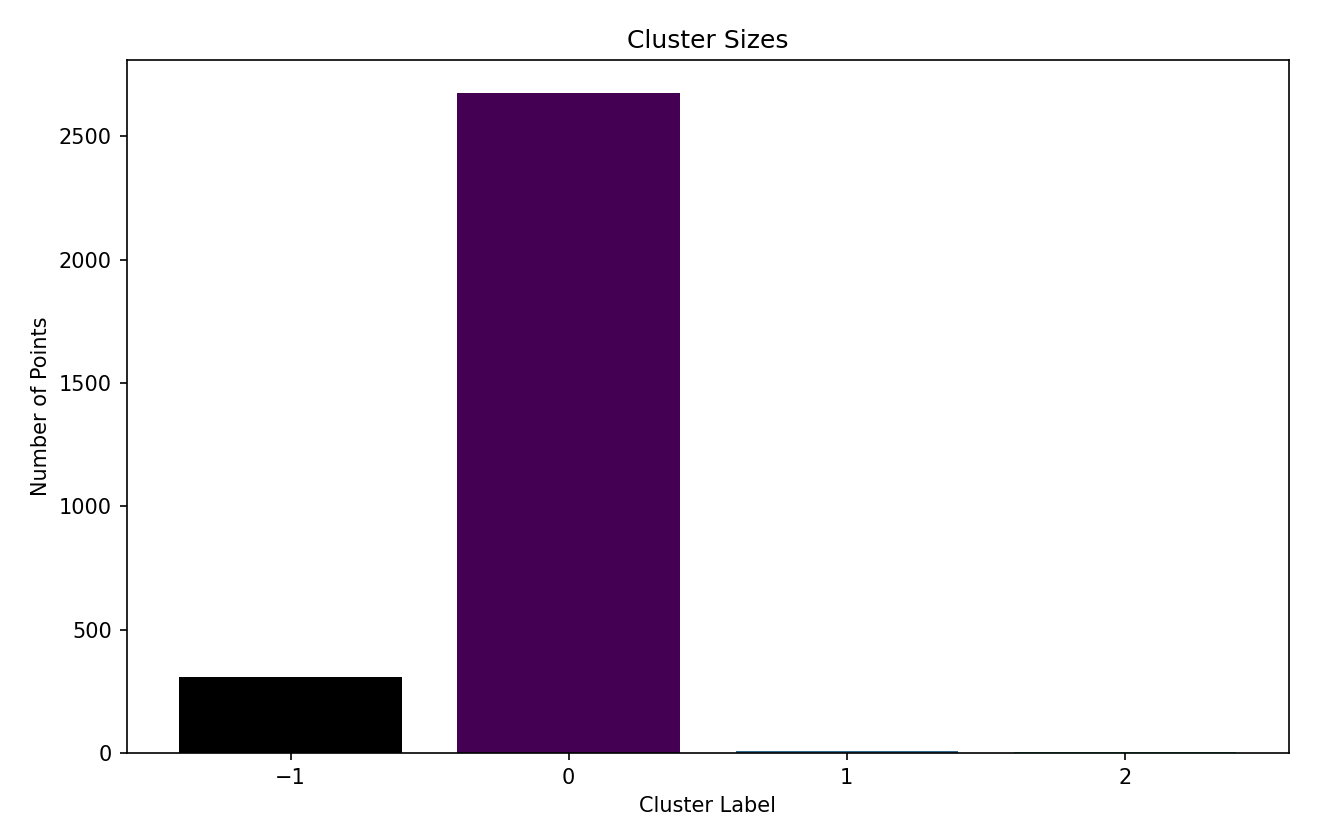
(圖十四)

1. DBSCAN

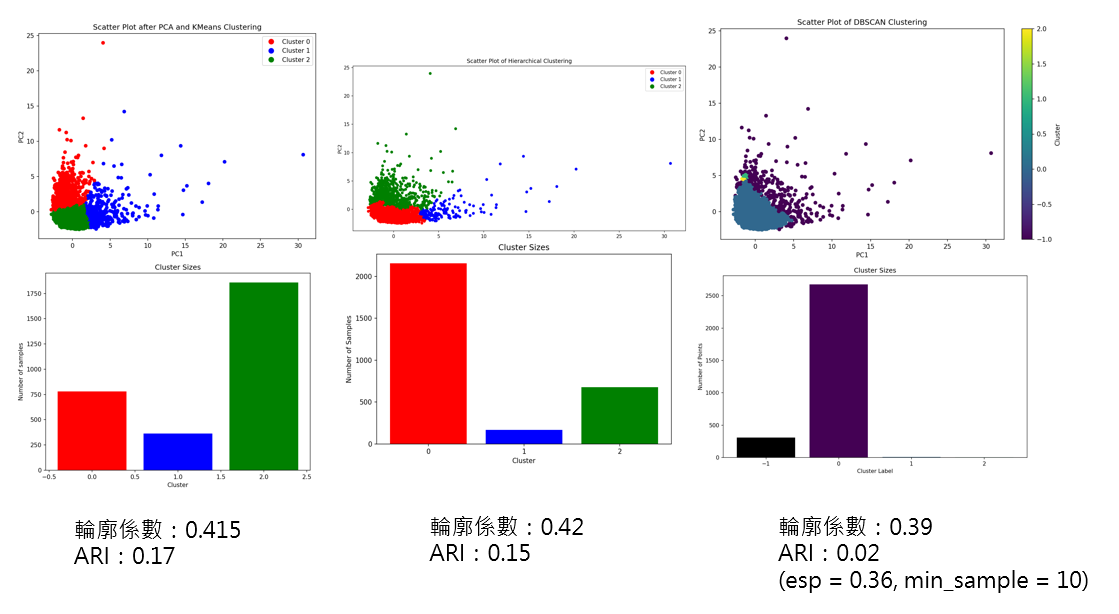
用function 3-3做DBSCAN，經過多次嘗試後得出將esp(最短距離)設為0.36，min\_samples 設為10可得以下數據(圖十五)，畫出散佈圖(圖十六)以及各集群大小(圖十七)

(圖十五)

(圖十六)

(圖十七)

1. 綜合比較3個模型的分群結果與分析討論

 (圖十八)

圖十八為三個模型的總表，從剛剛kmean的部分可以知道當集群數在3的時候表現會是最好的，所以Hierarchical Clustering的模型也將參數設為3群，而DBSCAN在更改esp與min\_sample時只需要些微的改動就可能讓結果完全不同，最後是經過一些嘗試才找到分群數量與前兩者相同，且輪廓係數表現較好的參數，不過實際分群出來的結果差強人意，Kmean的結果是三者中最平均的，其次是Hierarchical Clustering，最後才是DBSCAN。由於我三個都忘記做ARI所以在最後才補上，先用function 3-4-1對BALANCE做分類，把帳戶餘額分成三等份，當作真實的標籤，再把前面的function 3-1-3、function 3-2、function 3-3最後都加上function 3-4-2用以計算ARI，得出結果如圖十八，雖然三者表現都不是特別明顯，不過可以看到kmean是三者之中最能反映BALANCE的。