數資系 110813037 余珮綺

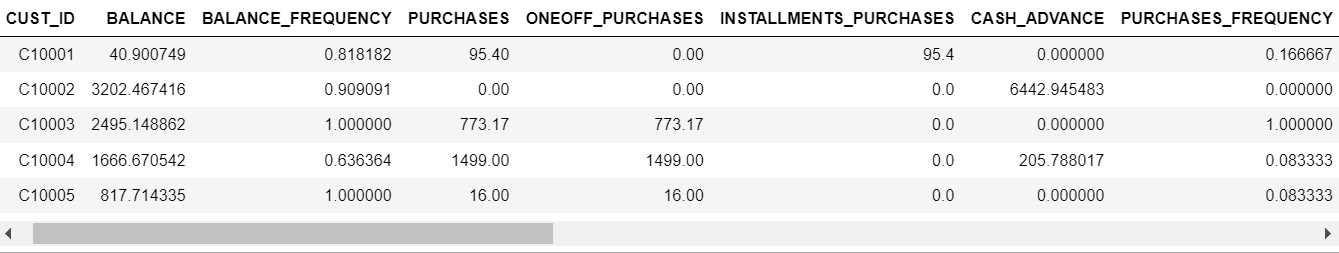
機器學習HW3\_Clustering

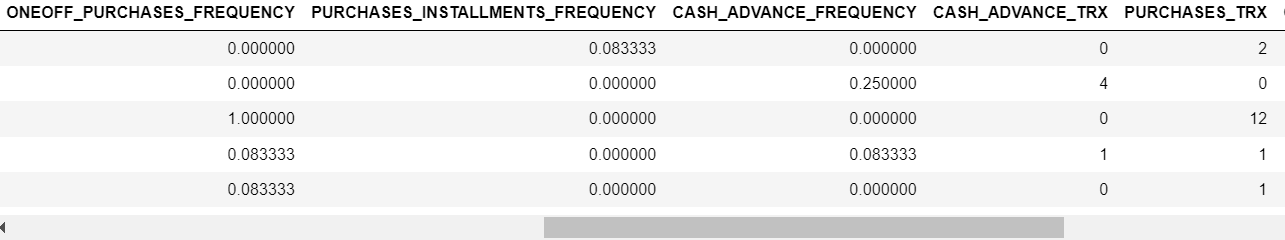
程式碼語言: python

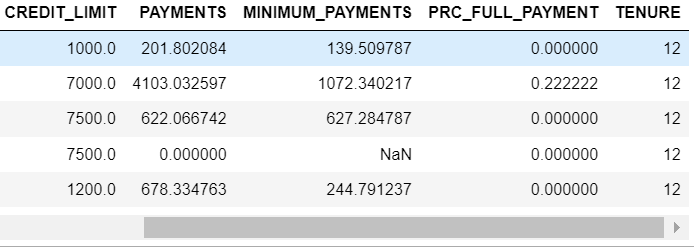
內容:

1. 資料清理與視覺化圖表

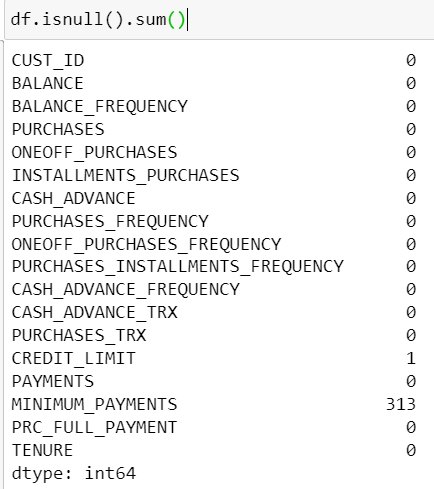
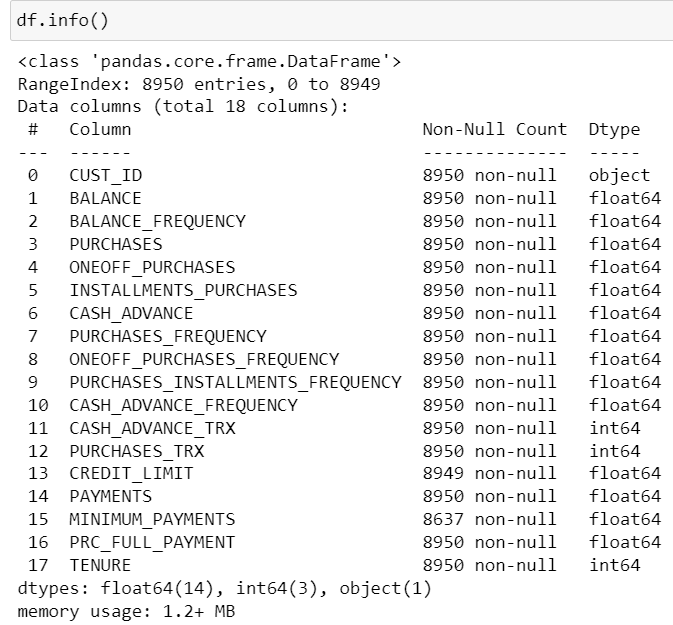
首先，先查看前幾筆的資料狀況







接著，再來確認是否有缺失值，發現有兩個欄位存在這樣的狀況



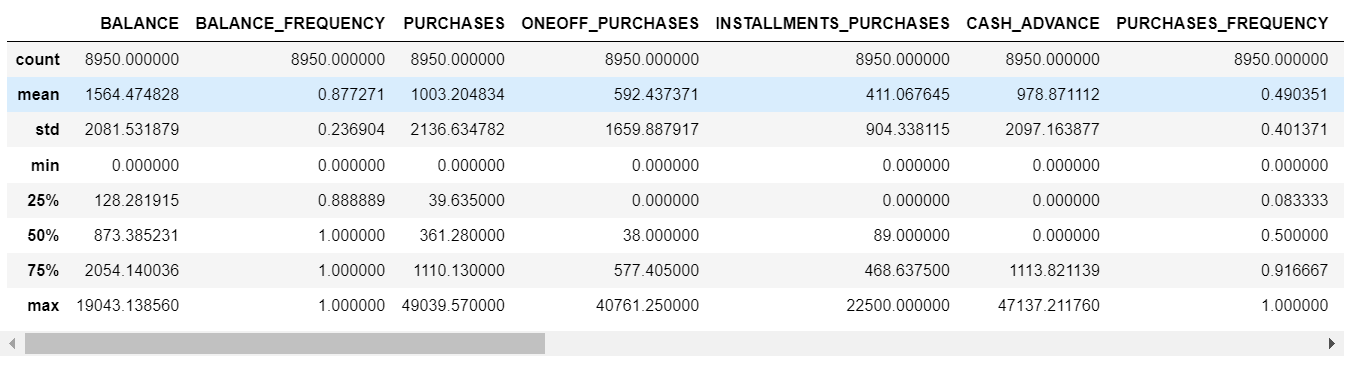
選擇以平均值的方式進行缺失值的填補

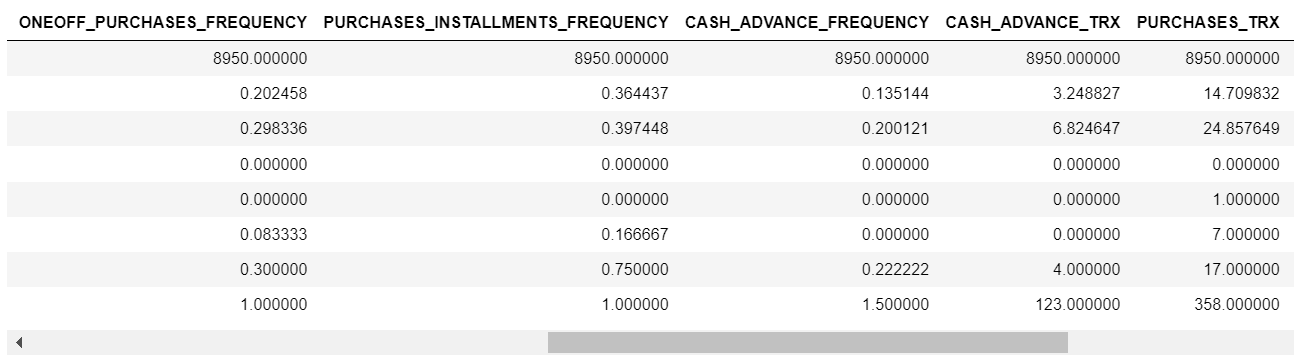


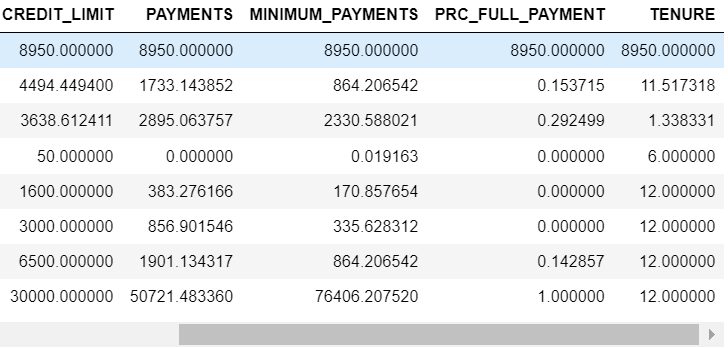


1. 敘述性統計分析

進行初步的資料整理後，查看整體資料的統計狀況，可發現某些欄位的最大最小值呈現極大的差異，待會將用圖表進行更值觀的察看

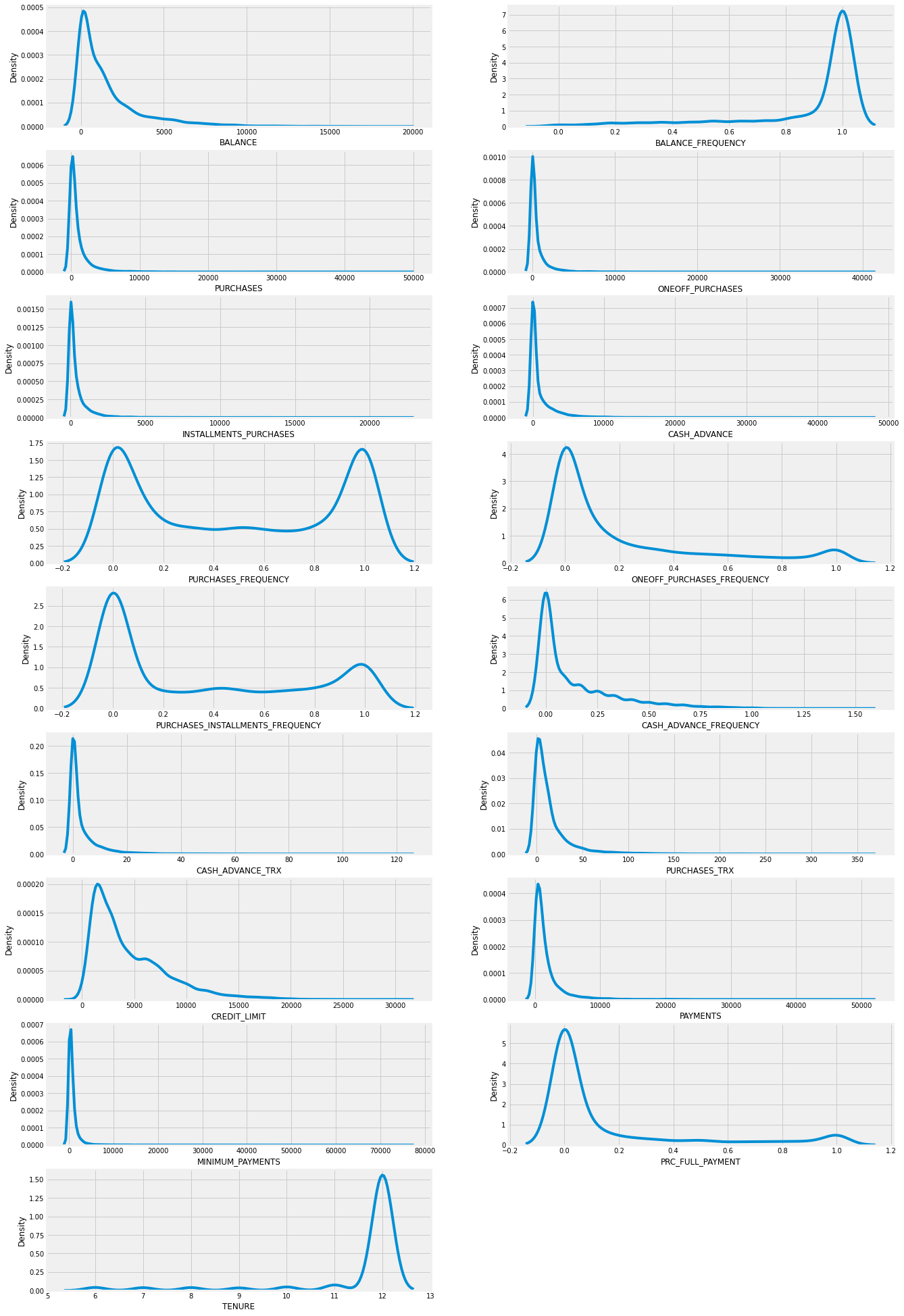




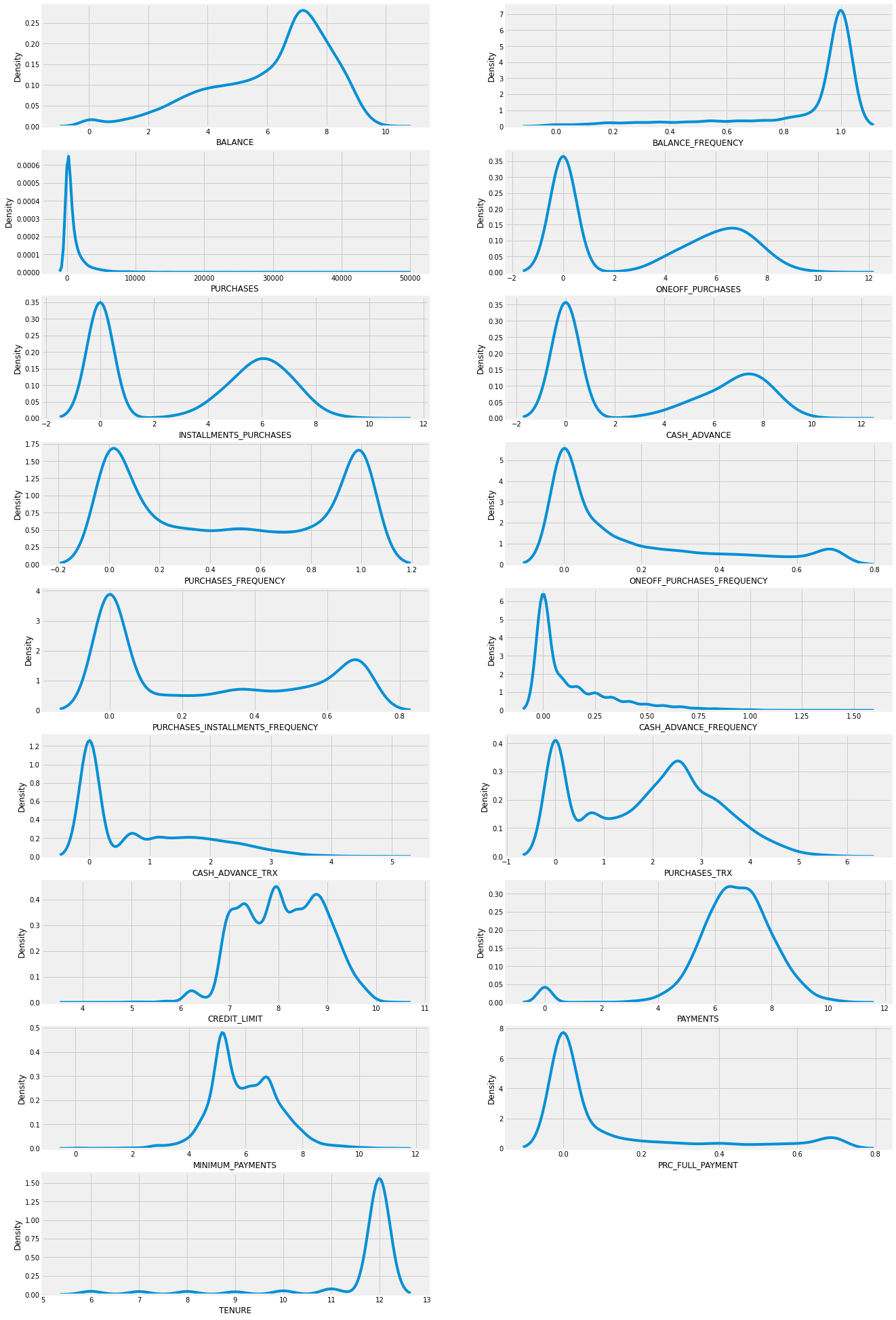


1. 特徵相關性分析

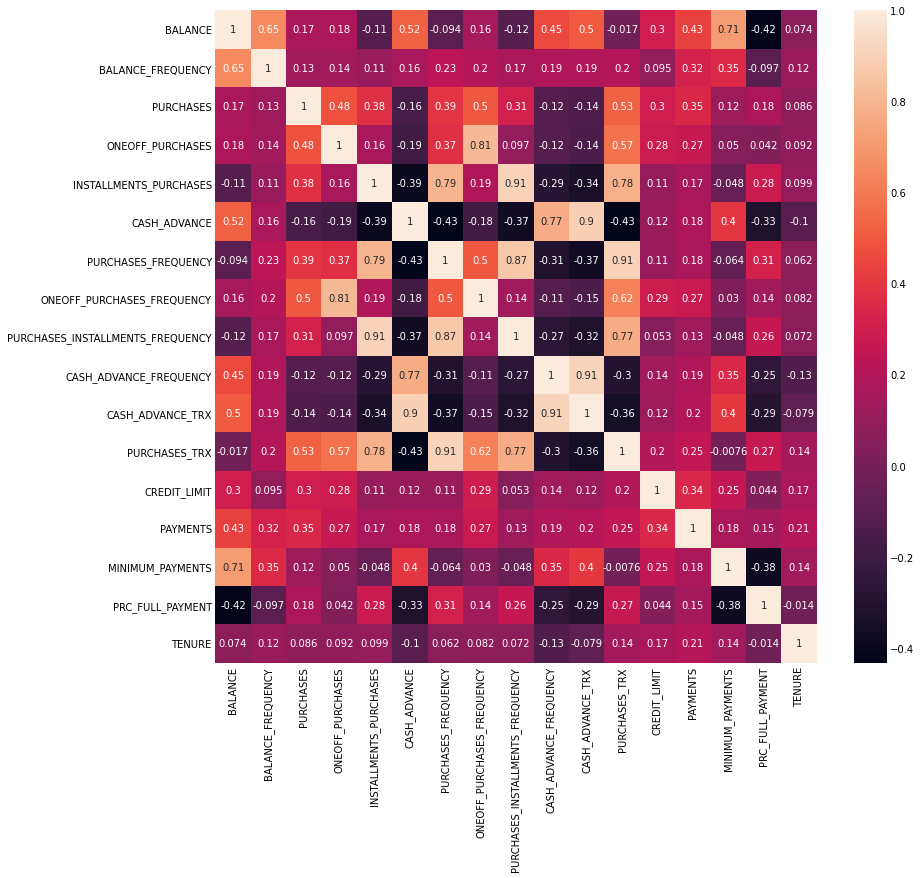
以圖表查看數值的偏移狀況，發現就如敘述性統計初步顯示的，偏移狀況明顯



 因此使用數學公式進行資料偏移的校正



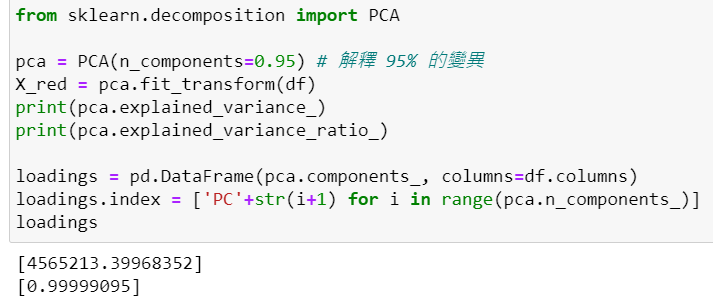
儘管某些欄位仍偏移嚴重，但已比未調整前佳



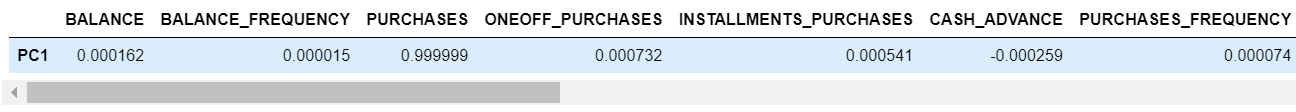
將調整過後的數值用熱圖查看欄位間的相關性，發現欄位名稱有相同單字的通常都呈現較高的相關性

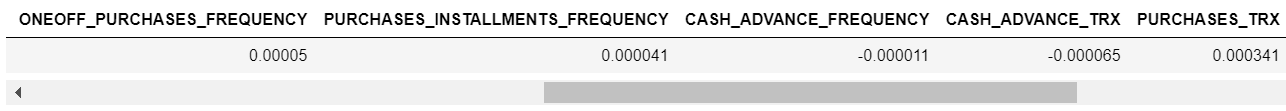
1. PCA降維處理與分析

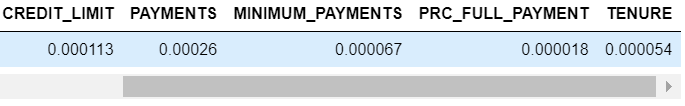
由於欄位眾多，因此採用PCA降維以利後續分析，在此採用接式95%變異的方式



以下可見得每個欄位對於降維後的占比，發現PURCHASES相當重要

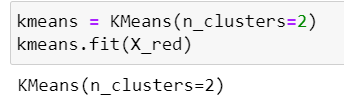
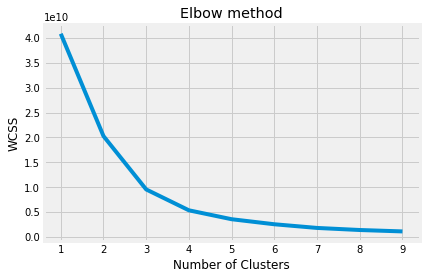




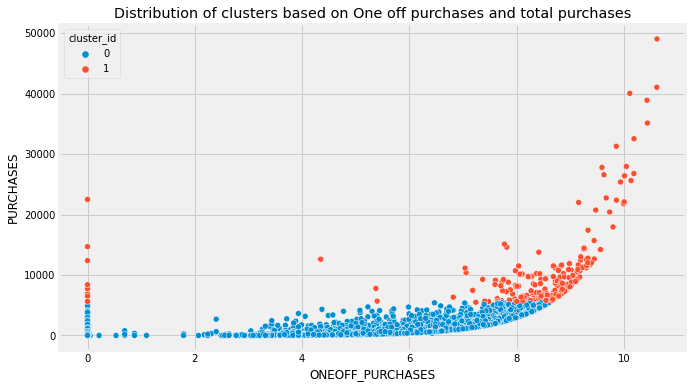


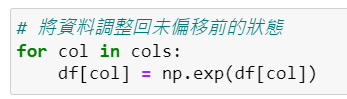
1. 資料分割與建置3個分群模型
2. K-means:

首先，先查看轉折圖，確認在分群數為多少時效果最好，發現n=2時表現應該是最佳

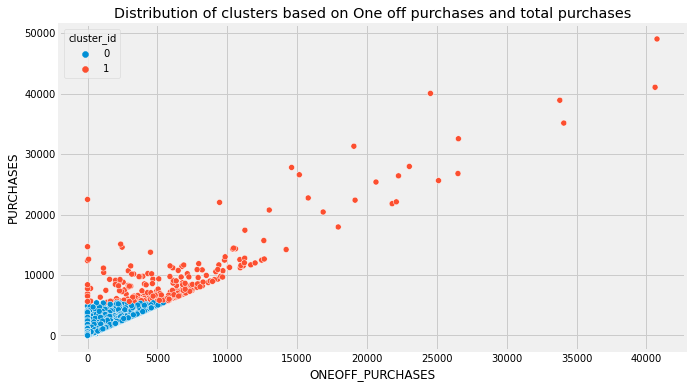


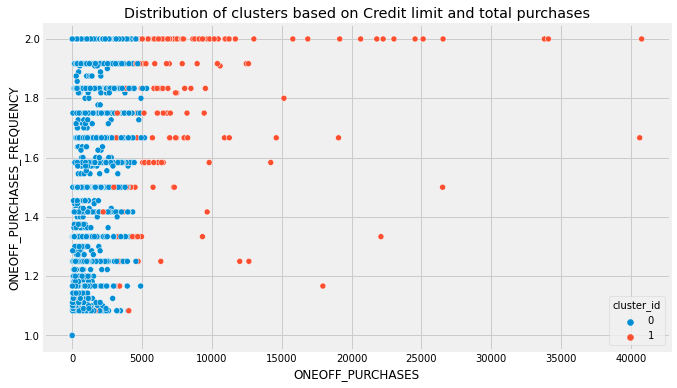
所以先查看了n=2的分群結果

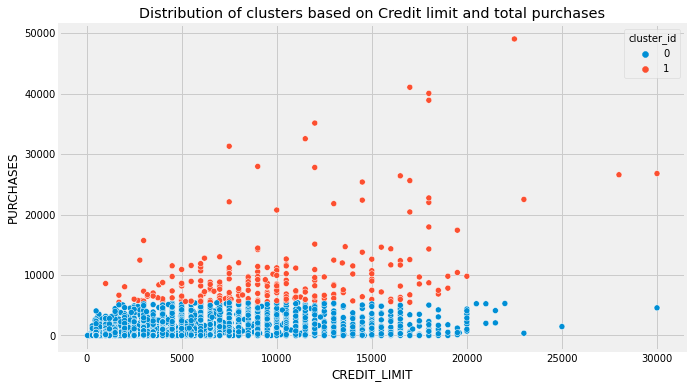




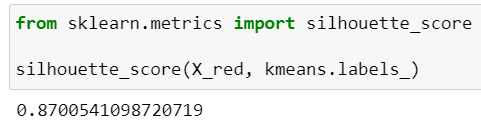
然而上圖為調整數值過後的分群結果將數值還原後再次進行查看



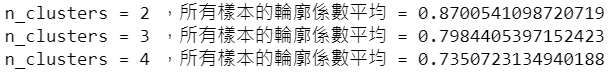


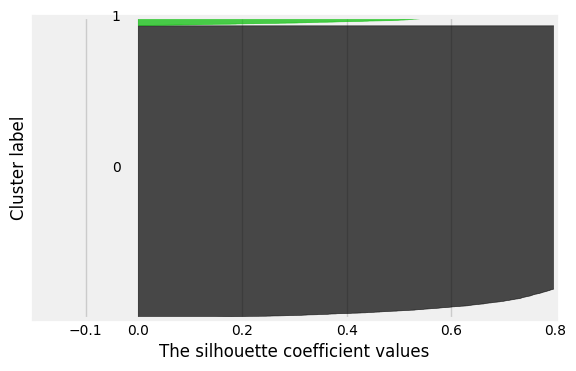


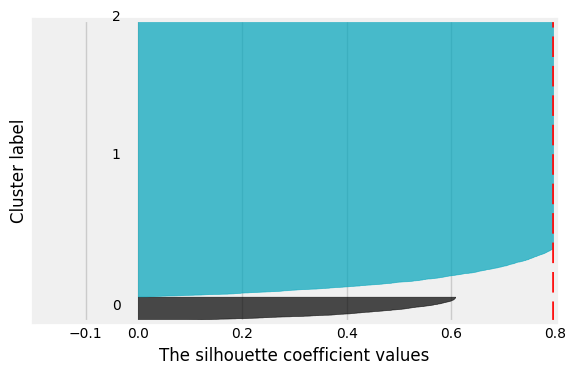
並針對不同的欄位相關程度進行分群結果查看。

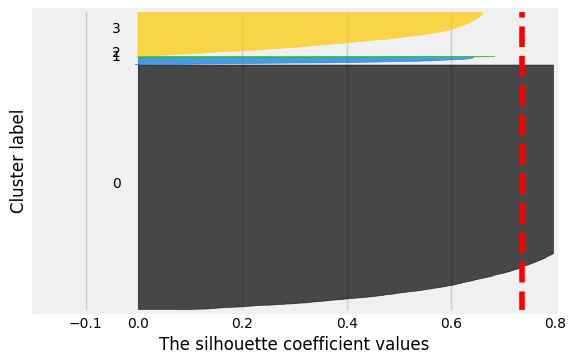


接著並進行分不同群數所產生的分數評比，發現確實分為2群的效果最好。



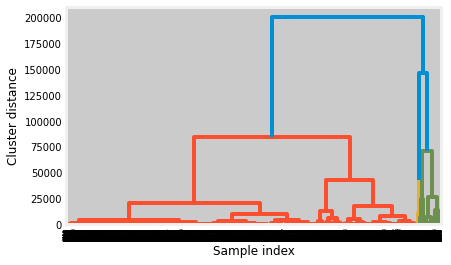




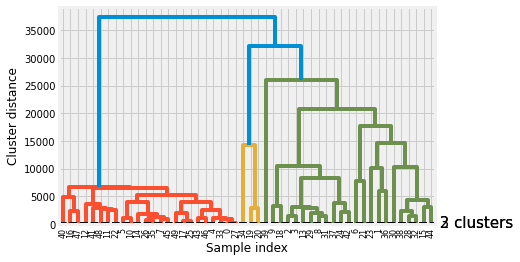


1. Hierarchical Clustering (Agglomerative) 採用ward linkage策略

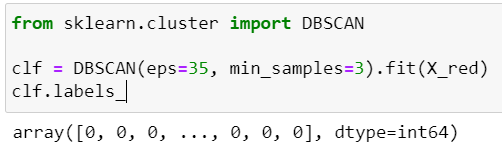
首先，先將所有資料都丟入模型進行分析，然而發現最終類別過堆，導致底下糊成一片



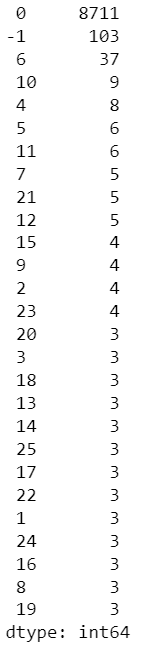
因此後來挑選出前50筆資料進行分析，可見得分析的結果更為清晰



1. DBSCAN



使用該方法進行分群，因為無須預設分群數，因此出來的結果更為細緻，雖然有些族群的結果過小，但特別的是還有分出-1的 noise。



1. 綜合比較3個模型的分群結果與分析討論

綜合以上，發現kmeans的分群可以很明確知道需要分出幾群，適合用在已知要達成的目標為合的狀況，而階層式的分析則適合用在資料筆數較少時，才不會花費過多時間還得到一個沒那麼清晰的結果，至於DBSCAN我認為他適合用在未知分群數與是否存在noise的情況。簡單來說，我認為各種分群方法皆有優缺點，只是在於追求的目標是什麼，就會有合適的應對方式。