統計計算期中書面報告

S&P 500 分群分析

清大計財所

學號：108071503

姓名：張庭瑄

目錄

一、研究目的

一般量化基金經理人在挑選標的股票時的做法，是將一股票池中股票依各種財務指標做線性加權，再將綜合得出的結果根據分數高低做分群，其中權重的大小通常是依經理人主觀判斷來增減，因此希望透過非監督式的分群方法，S&P 500 成分股做股票的分類以及異常股票的挑選，以利之後投資組合的組建。

二、研究方法

本報告分成兩個部分，第一部分用切割式分群(Partitional clustering)來將成分股分群並以 Gap statistic method 來尋找最適分群數目; 第二部分用階層式分群(Hierarchical Clustering)方法來分群並以 Average Silhouette Method 來尋找最適分群數目。

研究議題：S&P 500 成份股分群分析

2.1 議題介紹

透過 Clustering 的方法來將 S&P 500 成分股依照市場、財務狀況做分群，並以分群的結果為參考，根據選股策略從 cluster 之中挑選相似的個股，以利組建投資組合，並利用分群測試成分股中是否存在著異常財務或市場狀態的股票，該股可能隱含超額報酬，可以考慮進一步將這些股票做回測，並加入投資組合之中。

2.2 資料說明與處理

2.2.1.變數說明：

資料樣本為 4 月 30 號當天，美國標準普爾 500 指數內成分股，共 505 間公司的最新一季季報公布的財務指標以及最新三個月內市場狀況指標。變數介紹如下：

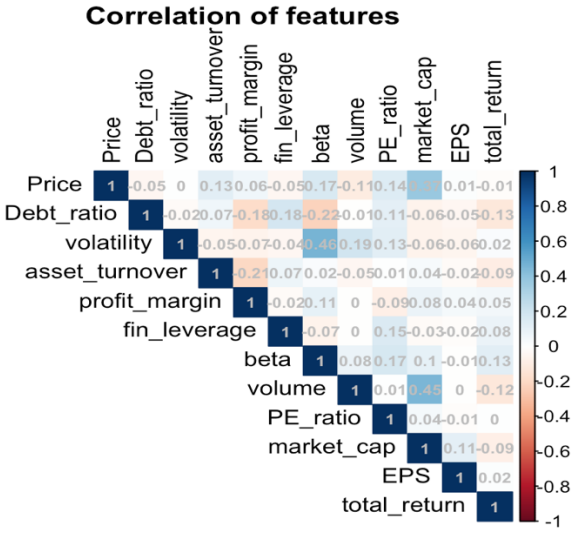
市場指標

Stock Price	近三個月內股價平均值
Volatility	近三個月股價波動率
Total Return	近三個月內股票報酬率
Volume	近三個月內交量加總
Market cap	近三個月公司市值(三個月股價平均＊流通在外股數)
Beta	近三個月市場風險

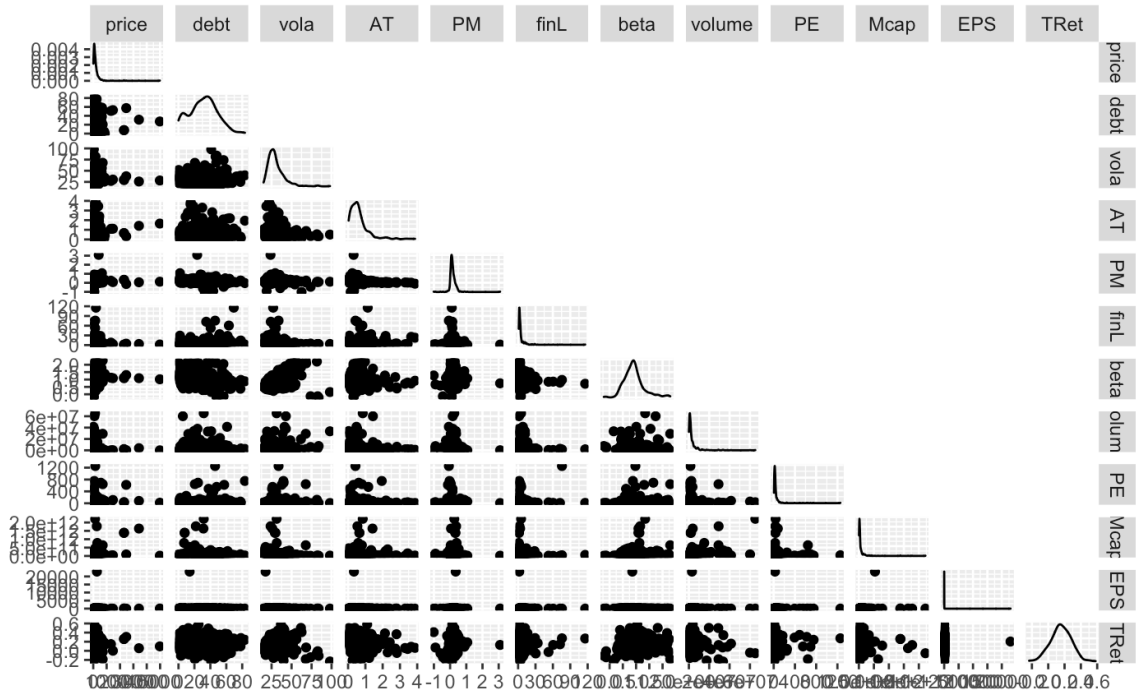
財務指標：

Asset Turnover	最新一季公司季報的資產週轉率，衡量公司營運狀況
Financial leverage	最新一季公司季報的財務槓桿，衡量財務風險
Current ratio	最新一季公司季報的流動比率，衡量公司短期償債能力
Debt ratio	最新一季公司季報的流動比率，衡量公司長期償債能力
Profit margin	最新一季公司季報的毛利率，衡量公司銷貨之獲利能力
EPS	最新一季公司季報的每股盈餘，衡量公司的獲利能力
PE ratio	最新一季公司季報的本益比，衡量公司股價的便宜與否

2.2.2.探索性資料分析：



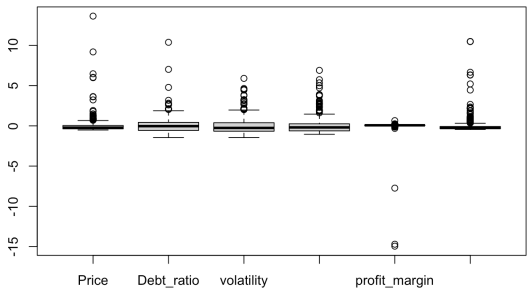
上圖為各變數之間的相關性關係圖，可以從中看出各個變數之間的相關性並不高，不存在共線性問題。

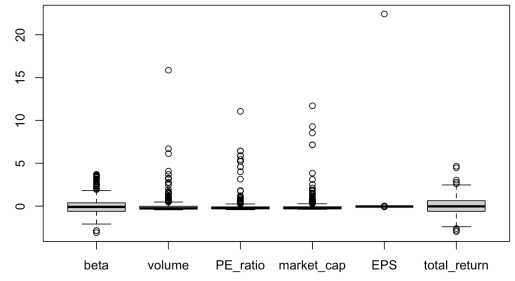


上圖為各變數之間的散佈圖以及變數分配圖，可以從中看出每個變數的分配皆非常態分配。

Standardized & Winsorized :

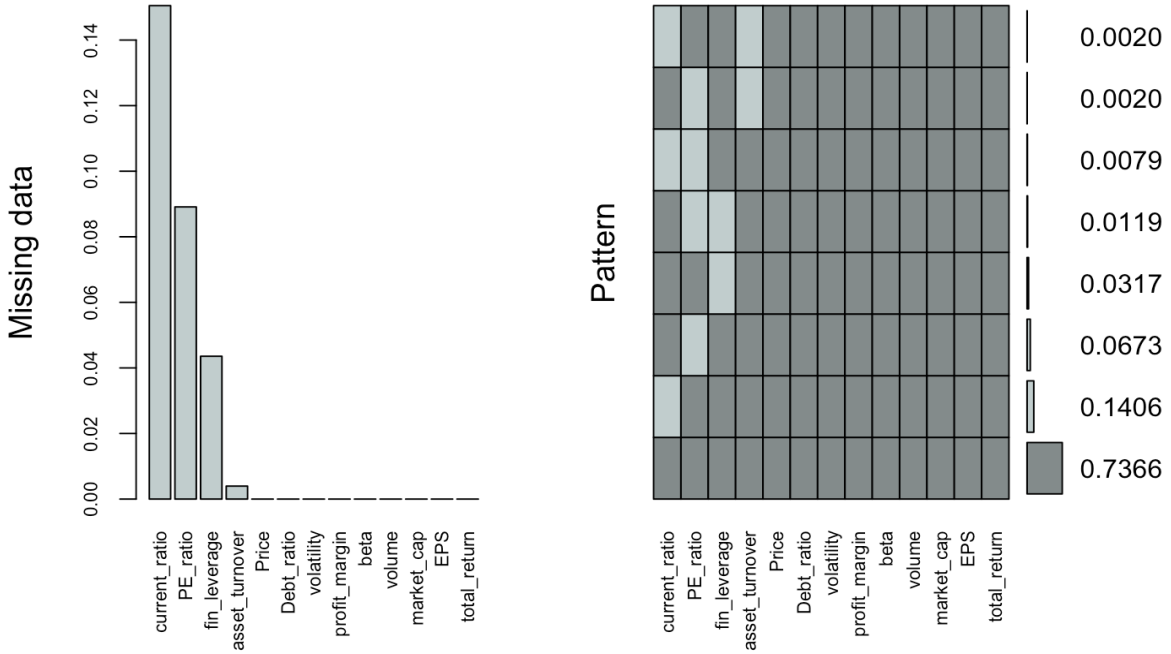
各變數之間 scale 不盡相同，並且在做 clustering 分析時要求資料事先標準化以利分群，因此將資料做標準化後以盒鬚圖觀察其分配





從上面兩張 boxplot 圖可以看出資料並非常態分配，且存在著許多離群值，為減少離群值對模型的影響，對資料進行縮尾處理，但因為變數中的 outliers 存在財務意涵，所以本報告只針對 3 個標準差之外的極值做縮尾處理。

2.2.3.遺失值處理：



本報告利用多重填補法(Mice)來對缺漏值比率低於 10%的樣本進行補值，其中變數 current_ratio 因為缺漏值比率超過 10%逕行刪除。 Mice 主要是透過 Gibbs Sample 來對多元變數產生多組填補值，概念為預測目標欄位(target column)時，會將其他非目標欄位作為預測變數(predictors)來進行建模，之後再用預測的結果來填補目標欄位的遺漏值，其中在建模預測時本報告採用 Random forest 的方法。

2.3 模型說明

本報告使用的分群模型可以分成切割式分群以及階層式分群兩類，並透過下列兩種常見的方法來決定最佳分群數目，分別為 Average Silhouette method（側影圖法）、Gap statistic（Gap 統計量）。

2.3.1 切割式分群 (Partitional Clustering)

2.3.1.1 K-Medoid

本報告採用切割式分群中的 K-medoid 法來將資料分群，因為一般的 K-mean 法容易受離群值影響，而此樣本存在眾多離群值，在分群上並不適合，而 K-medoid 在中心點選擇上，將選擇 cluster 內某個真實觀測值，而非群內平均值，較不易受離群值所影響。另外 K-Medoids 比 K-Means 更強之處在於他的目標是最小化相異度加總值，而非只是歐式距離平方和，因此處理類別變數資料。但在遇到樣本數較大時，複雜度較高，運算較複雜。

K-medoid 步驟如下：

1. 隨機選取 K 個質心的值（質心必須是某些樣本點的值，非任意值）

2. 計算各個點到質心的距離

3. 將點的類劃分為離他最近的質心，形成 K 個群集

4. 根據分類好的集群，在每個集群內重新計算質心：

1) 計算群集內所有樣本點到其中一個樣本點的曼哈頓距離和（絕對誤差）

2) 選出使群集絕對誤差最小的樣本點作為質心

5. 重複迭代步驟 2~4 直到滿足迭代次數或誤差小於指定的值

2.3.1.2 決定最適分群數目

Average silhouette method 是衡量各分群結果的品質的指標，其是透過計算各 cluster 中，每個觀察值的 silhouette width，並取群內 silhouette width 平均值而得。其中，觀察值的 silhouette with 衡量的是該觀察值是否被很好的歸類在合適的群聚。若 silhouette with 若為正數且值越大，則表示該觀測植被很好的分派到合適的群聚；相反的，若 silhouette with 值很小或甚至為負數，則表示該觀測值的分群結果不是很合適。而整個 cluster 的 average silhouette with 越大表示分群做得越好。接著

計算 k=1~n 群 cluster 與對應 average silhouette width。而其中極大化 average silhouette width 的 k 值即為最佳分群數目。

2.3.2 Hierarchical Clustering

階層分群法並不需要預先設定分群數，並會產生的分群結果為一目瞭然的樹狀結構圖（又稱作 dendrogram）。其中群數(number of clusters)可由大變小(divisive hierarchical clustering)，或是由小變大(agglomerative hierarchical clustering)，透過群聚反覆的分裂和合併後，再選取最佳的分群數目 k。

2.3.2.1 決定最適分群數目

除上面所提及之 Average silhouette method 之外，也可考慮使用 Gap statistic，相關介紹如下：

Gap statistic（Gap 統計量，predicted-observed）

比較不同 k 水準值下，實際觀測值分群結果的群內總變異(W_k)和 Boostrap 抽樣法 B 次所產生樣本分群結果期望群內總變異($\overline{W_k}$) 的差異。 公式如下：

$$Gap(k) = \overline{W_k} - W_k = \frac{1}{B}\sum_{i=1}^{B} log\,(W_{ki}) - W_k$$

並選擇使$Gap(k)$最大化的最小 k 值，以符合：

$$Gap(k) = Gap(k+1) - s_{k+1}$$

也就表示 k 群的分群結構和虛無假設的 uniform 分配(沒有分群) 有很大的差距。

2.3.2.2 聚合式階層群聚法

透過計算每個觀察值之間的兩兩距離，將相近的兩點分成一 cluster，之後再次計算不同 cluster 之間的距離，相近者會再分成同一 cluster，反覆上面步驟，資料會由樹狀結構的底部開始開始逐次合併 (bottom-up)，最終會產生一樹狀結構圖(dendrogram)。

Dendrogram 中的葉節點（末梢節點）代表個別資料點，垂直座標軸的 Height 代表群聚間的不相似度(dissimilarity)， 群聚的高度(height)越高，代表觀測值間越不相似（組內變異越大）。特別注意的是要總結兩個觀測值的相似度，只能用兩觀測值何時被第一次合併的群聚高度來做判斷，而不能以水平軸的距離來評估。

在計算兩兩資料間的相異度矩陣時有常見的幾種演算法：

單一連結聚合演算法(single-linkage agglomerative algorithm): 定義 clusters 之間距離為兩 cluster 中最近的兩個點的距離。

完整連結聚合演算法(complete-linkage agglomerative algorithm): 定義 clusters 之間距離為兩 clusters 中最遠的兩個點的距離。

平均連結聚合演算法(average-linkage agglomerative algorithm): 定義 clusters 之間距離為兩 clusters 中各點與各點距離總和的平均。

中心連結聚合演算法(centroid-linkage agglomerative algorithm): 定義 clusters 之間距離為不同 clusters 中心點之間的距離。

華德最小變異法(Ward’s Minimum Variance): 最小化各群聚內變異加總(minimize the total within-cluster variance)。主要用來尋找緊湊球型的群聚。反覆比較每對資料合併後的群內總變異數的增量，並找增量最小的組別優先合併。越早合併的子集表示其間的相似度越高。而使用華德最小變異法的前提為，初始各點資料距離必須是歐式距離的平方和(Squared Euclidean Distance)。

2.3.2.3 分裂式階層群聚法

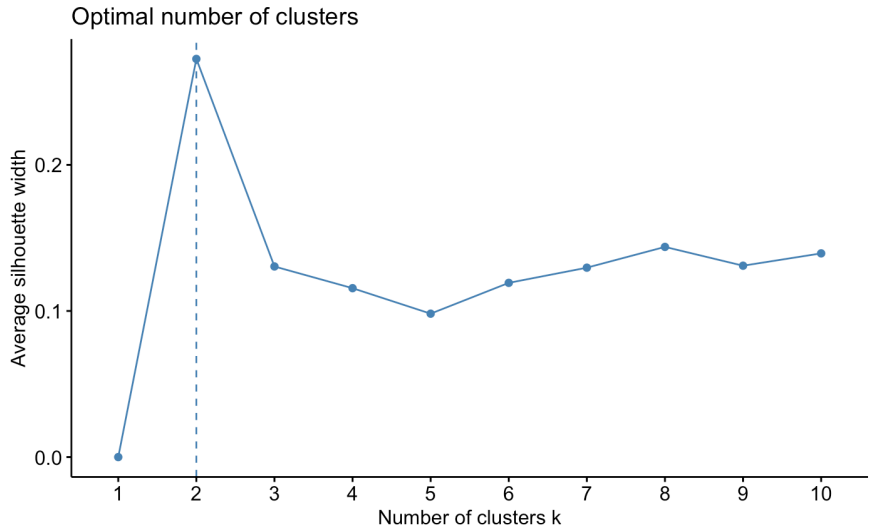
與聚合式階層群聚法的不同點為，資料會由樹狀結構的頂部開始逐次分裂(top-down)。

三、結果與討論

3.1 切割式分群 (Partitional Clustering)

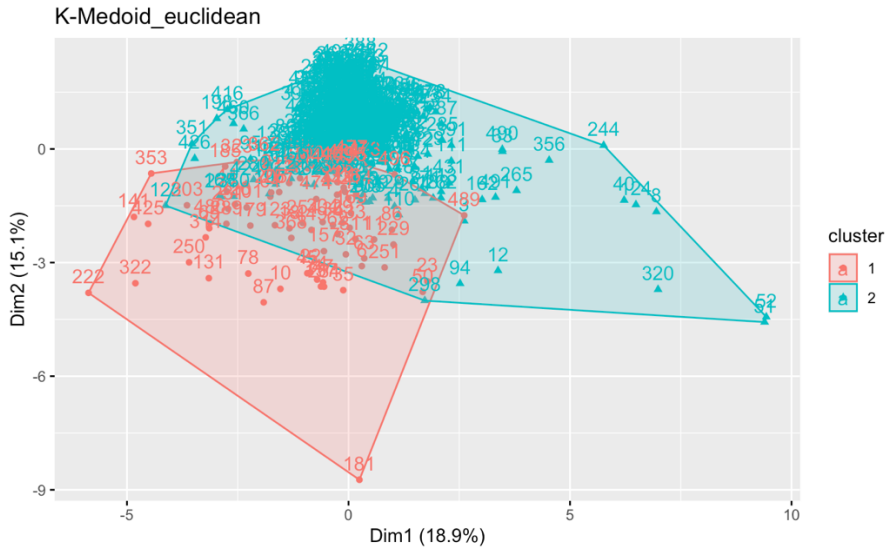
3.1.1 決定最適分群數目

分群數目 k 與對應 average silhouette width 如下圖



本報告選擇 k=2 作為切割式分群中最適分群數目

利用 K-mediod (歐式距離)所分群出的結果如下圖



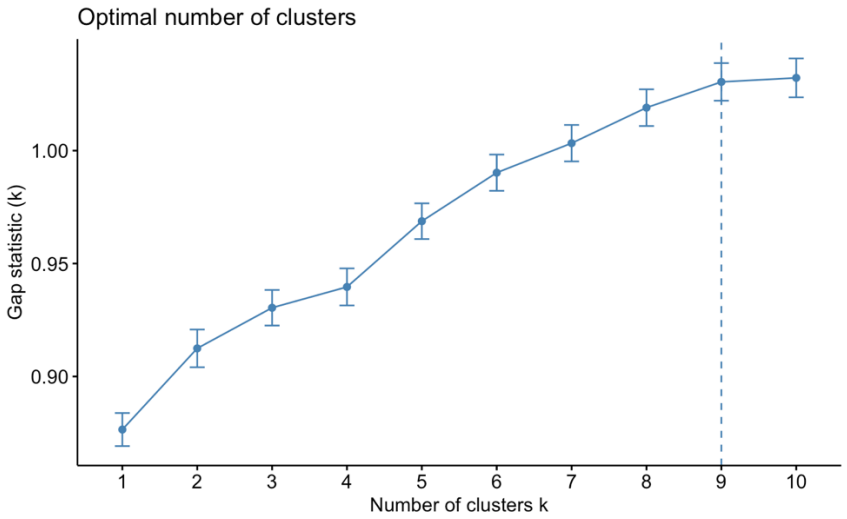
上圖的 X 軸與 Y 軸是分別是主成分分析的 PC1 與 PC2。

從圖中可以看到 K medoid 囊括 outliers 進 clusters 中，但兩群之間的資料點很相近，大部分資料再經過主成分分析降維之後，資料的分佈十分密集，因此分類出兩 clusters 之間的差異可能沒有很大。

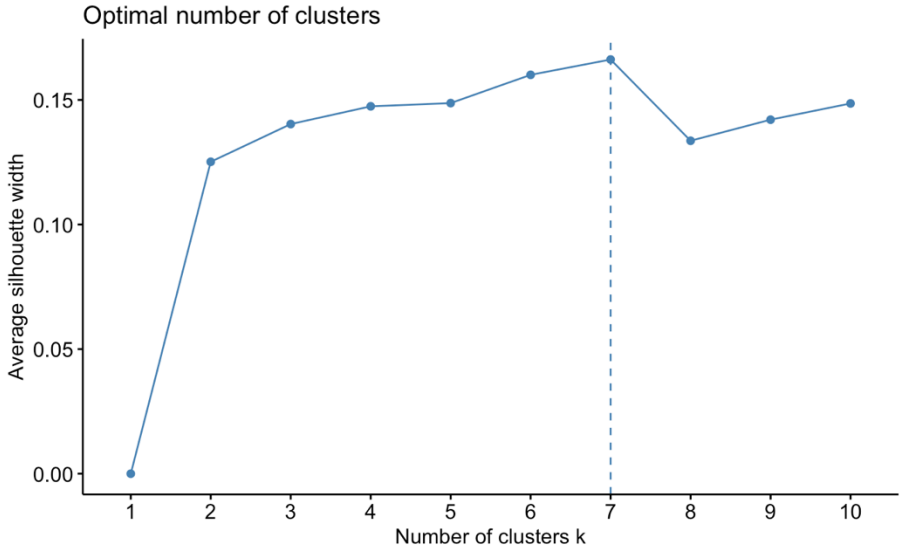
3.2 Hierarchical Clustering

3.2.1 決定最適分群數目

下圖是 Gap statistic 統計量所挑選出最適的分群數目



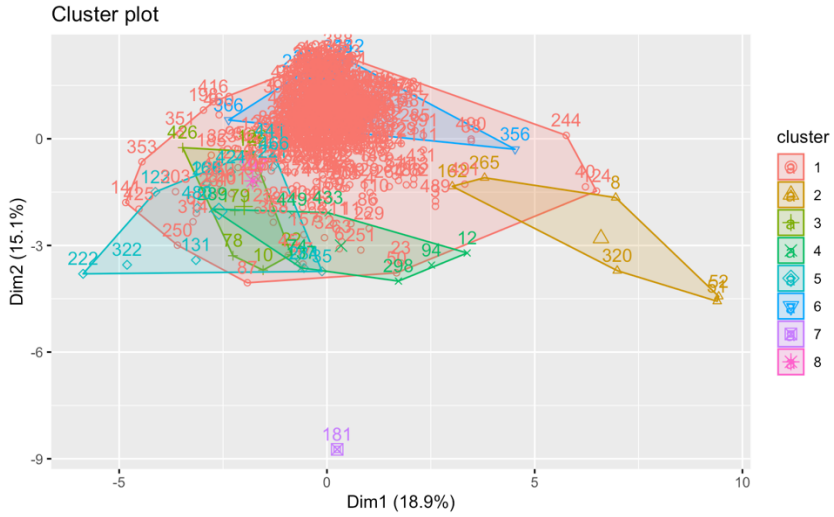
同時考慮 Average silhouette method 所挑選的最適分群數目，如下圖



綜合兩者結果，挑選 k=8 作為最適分群數目。

3.2.2 聚合式階層群聚法

下圖是利用聚合式分類法(曼哈頓距離-average 演算法)所得到之分群結果



3.2.3 分裂式階層群聚法

下圖是利用分裂式分類法所得到之分群結果

比較分裂式以及聚合式兩張圖的結果，可以明顯發現 observation_181 皆被獨自分成一類，代表該公司與其他公司之間的財務、市場狀況可能存在特別不同之處，極有可能該股票存在超額報酬，可以考慮將該公司進一步做回測，並加入投組之中。

3.3 結論

1. 股票分類上提供另外一種方法，供主動式基金經理人挑選標的。
2. 前 500 大成分股的變數之間資料十分相似，分類出的每個 cluster 之間差距可能不大，可以考慮擴展股票池的深度、增加股票數目，以及考慮市場預期的變數，可能可以分出更多 clusters。

3. 可以藉由 clustering 來找出與眾不同的個股，該個股極有可能隱含超額報酬，可以特別挑出來組建投資組合，回測過去績效，以賺取超額報酬。

四、參考資料

1.上課講義

2. An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

3.網站

https://www.jamleecute.com/hierarchical-clustering

https://www.jamleecute.com/missing-value-treatment

https://www.jamleecute.com/partitional-clustering-kmeans-kmedoid/

https://www.mdeditor.tw/pl/258g/zh-tw

https://zhuanlan.zhihu.com/p/30616837

https://zhuanlan.zhihu.com/p/55163617