統計計算期中個人報告

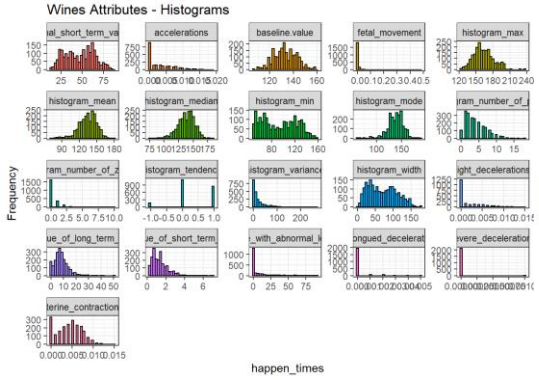
106070020 何羿樺

1. 介紹:

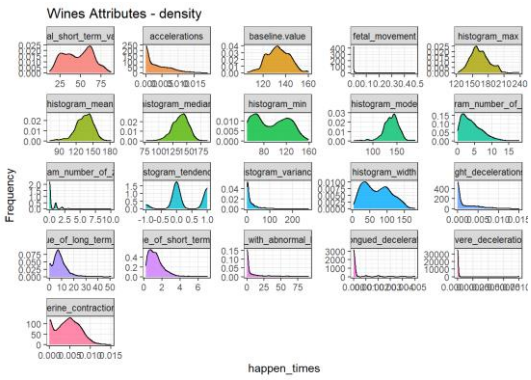
降低兒童死亡率體現在聯合國的幾個可持續發展目標中，是衡量人類進步的一個關鍵指標。聯合國期望到 2030 年，各國結束可預防的新生兒和 5 歲以下兒童的死亡，所有國家的目標是將 5 歲以下兒童的死亡率至少降低到每千名活產兒 25 例。與兒童死亡率概念平行的當然是孕產婦死亡率，在懷孕和分娩期間及之後的死亡人數為 295 000 人（截至 2017 年）。這些死亡中的絕大部分（94%）發生在低資源環境中，而且大多數是可以預防的。鑒於上述情況，心電圖（CTG）是評估胎兒健康的一個簡單且成本低廉的選擇，允許醫療保健專業人員採取行動，以防止兒童和孕產婦死亡。該設備本身的工作原理是發送超聲脈衝並讀取其反應，從而揭示胎兒心率（FHR）、胎兒運動、子宮收縮等情況。我所使用的數據資料集包含 2126 條從心電圖檢查中提取的特徵記錄，然後由三位產科專家將其分為三類，正常，疑似，病理性。因此，我想使用上課所教的分類方法 K-Means Clustering, Hierarchical Clustering, 和 Gaussian Mixture Model(GMM)，來分類 CTG 特徵。

2. 探索式資料分析(EDA)

首先，如圖一和圖二所示，我畫出了這份資料中每一欄的直方圖以及分布圖，可以看出資料的分布。從圖上可以看出某些參數在各個值的分布上極不均勻，但也有某些參數有均勻地分布。

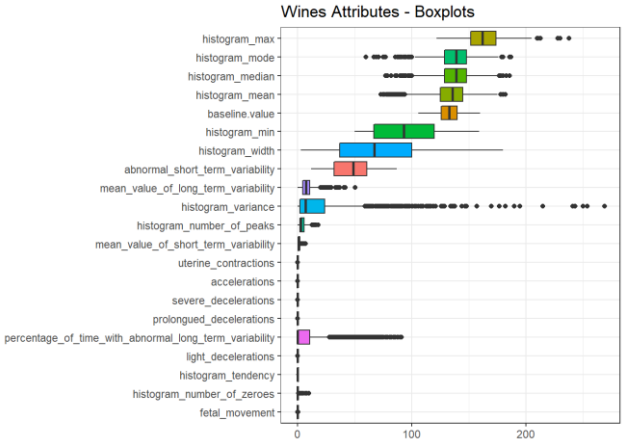


(圖一)



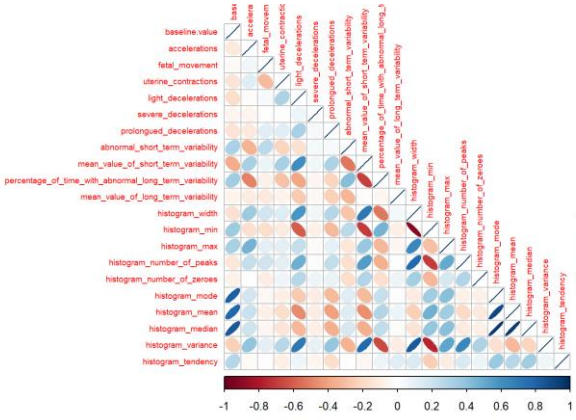
(圖二)

由圖三可以看出資料的盒狀圖，排序方式是按照各資料中位數的大小。

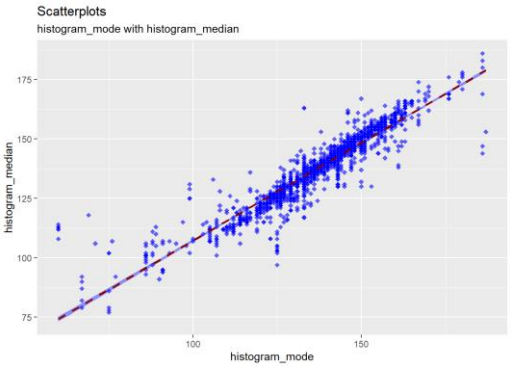


(圖三)

我使用 corrplot()函數來創建一個相關矩陣的圖形來顯示不同屬性之間的關係。由圖中顯示的結果可以看出 histogram_mode 跟 histogram_median 有強烈的正相關，因此我通過擬合一個線性模型來類比這兩個變數之間的關係(如圖五所示)。

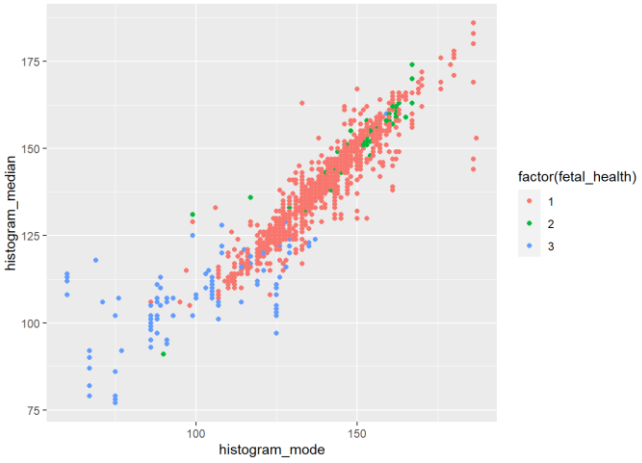


(圖四)



(圖五)

由圖六可以看出 histogram_mode 跟 histogram_median 之間以及跟各個 factors 之間的散狀圖分布



(圖六)

3. 資料分析(Data Analysis)

在結束 EDA 之後，我將利用上課教的三種方式去做分類，由於真實世界中的資料不會有 label，因此在這個環節，資料中的 label 會被忽略掉，我會利用三種分群方法進行分群，並看結果為何。

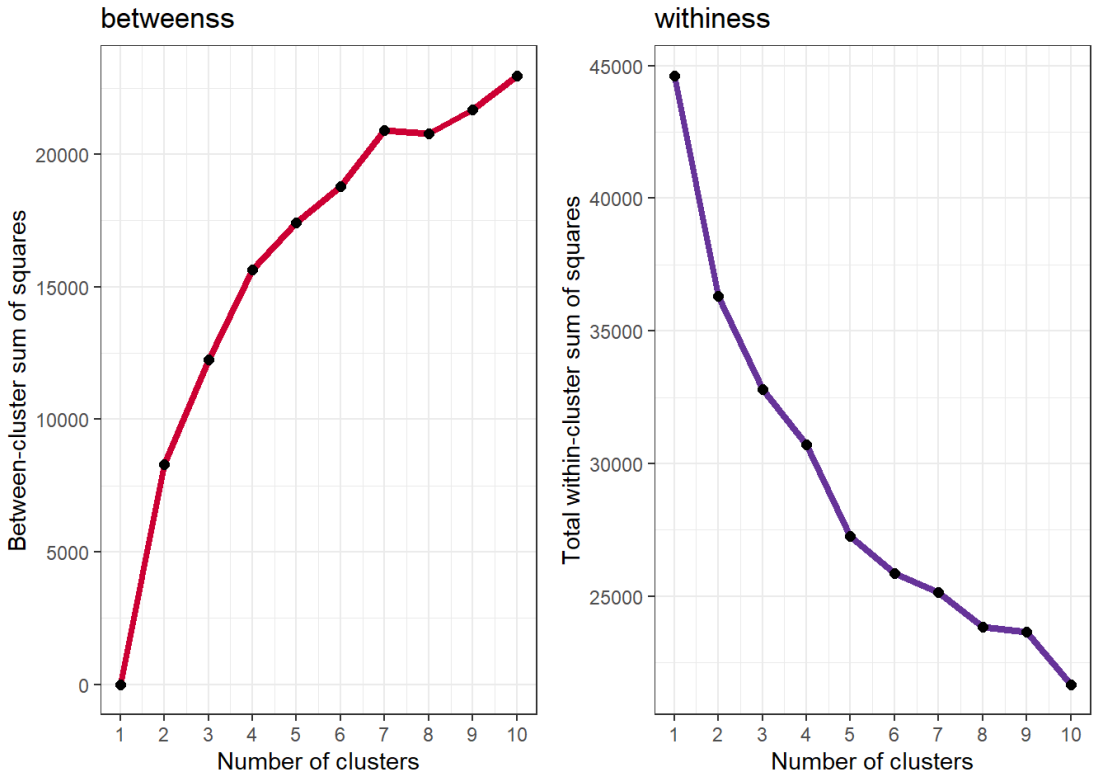
A. K-means

假設在全部資料中，可將資料分為 n 群。並反覆的閱讀資料點更新對應的群中心，可視作質心點 $\mu_1, \ldots, \mu_K$ ，使得找到最代表每個群的對應質心，換句話說，每個資料點會被分配到與他最相近的群裡。

若要使用 K-means，必須要先決定適當的分群數，以下我會使用三種方式來決定合適的分群數。

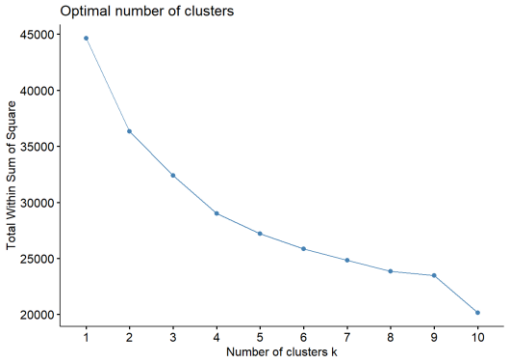
第一種:

希望使群內的總變異最小；使群間的總變異最大，我畫出 betweeness 和 withiness，我們的目標是讓 betweeness 愈大而 withness 愈小，如圖七所示，我顯示出了分群數從 1 到 10 betweeness 和 withiness 的變化。

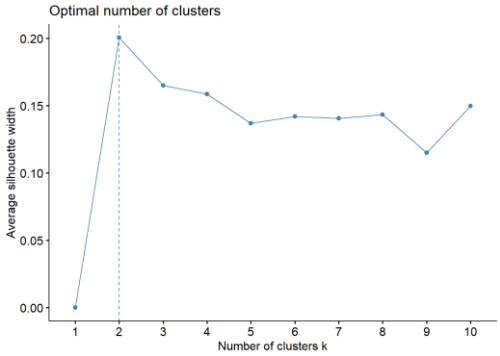


(圖七)

第二種方式就是找出一個數字 n，使得資料被分成 n 群時，群內的總變異(SSE)會最小，那麼 n = 最佳的分群數目，而這就是 Elbow Method！因為在 factoextra 的套件裡，已經有函式 fviz_nbclust()，我們可以直接實踐 Elbow Method(如圖八)。



(圖八)



(圖九)

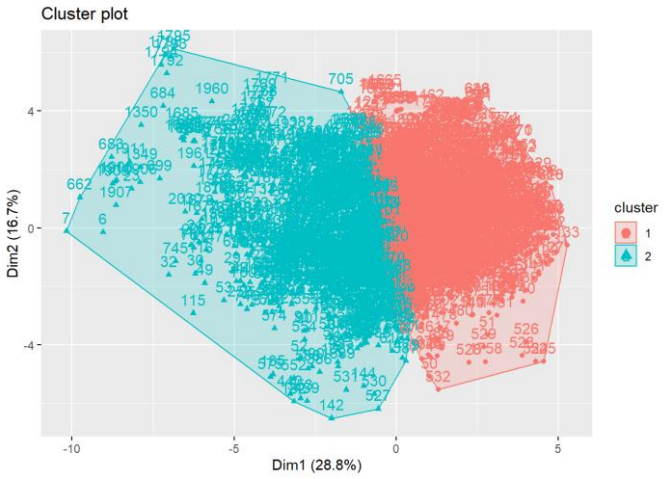
第三種方式為平均側影法(Average silhouette Method)。所謂側影系數(Silhouette Coefficient)會根據每個資料點(i)的內聚力和分散力，衡量分群的效果(quality)。其公式為:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- a(i)為資料點(i)，為它與群內其他資料點的平均距離
- b(i)為資料點(i)，為它與其他群內資料點的最小平均距離值
- s(i) = 側影係數，可以視為該資料點(i)在它所屬的群內是否適當的指標

取完每一個資料點的側影平均值，就可以將它當作當作衡量最佳分群數目的標準。

綜觀上述三種方式所計算出來的結果，我決定以 n=2 當作我的 K-means 分群值。

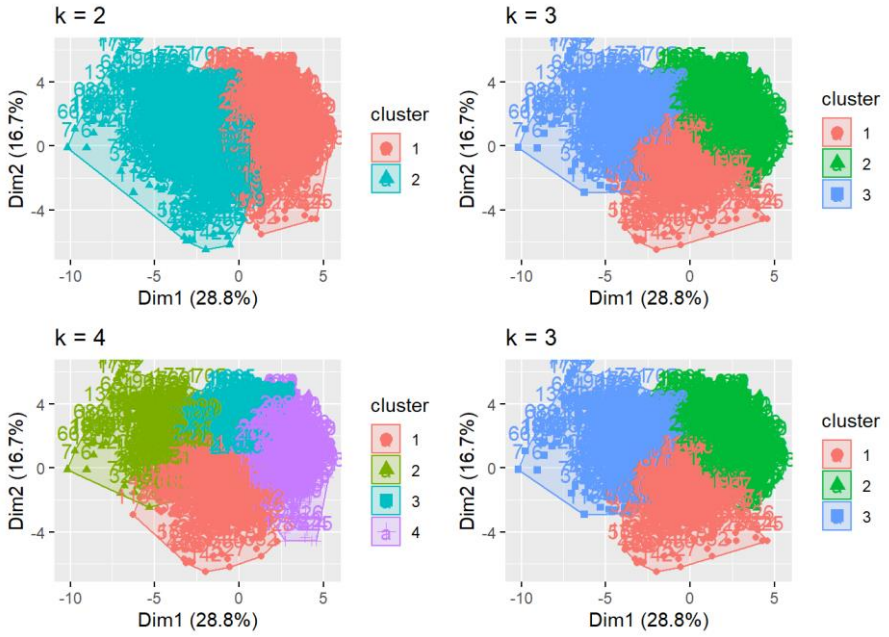


(圖 10)

由圖 10 的結果看來，當分群值為二時，他並無法完全的分開兩個群體，在

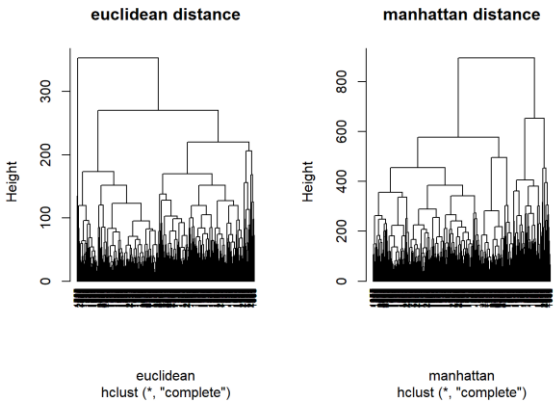
交接處仍舊重疊了許多值，而我認為這個結果是合理的，因為這是心電圖的資料，所以並無法單看心電圖就完美的分出患者的情況是否正常，或是直接判斷是否為病理性，會有許多模稜兩可的結果，舉例來說，有時候數字顯示可能是正常，但該患者有可能是病理性的患者。

除了 n=2，我也試了 n=3, n=4 的分群結果，如圖 11 所示，而重疊的部分也變得更多了，尤其是當 n=4 時，圖形的重疊部分變得十分多。
若特別看 k=3 的情況(由於原本的 label 分為正常、疑似和病理性)，我們可以看到中間的部分重疊了，也就是說，在許多情況下，直接利用 k-means 結果分群並無法準確的區分出這三種情況，舉例來說，可能一位正常的病患會被歸類於疑似或是病理的情況。



(圖 11)

B. Hierarchical Clustering



(圖 12)

我利用樹狀圖來或出資料的距離以及結構，圖 12 中左邊的那張圖為歐式距

離，右邊的為曼哈頓距離。歐氏距離也可以說是歐幾裡得距離，是一個通常採用的距離定義。它是在 m 維空間中兩個點之間的真實距離。在二維和三維空間中的歐氏距離的就是兩點之間的距離。使用這個距離，歐氏空間成為度量空間。曼哈頓距離是一種使用在幾何度量空間的幾何學用語，用以標明兩個點在標準坐標繫上的絕對軸距總和。

當算出了距離矩陣之後，我利用了五種方式把資料結合起來，而不同的方法會產生不同的效果。方法為以下五種:

(A)最近法(單一聯結法 Single Linkage)：

$$d_{A,B} = \underset{\substack{i \in A \\ j \in B}}{Min}\, d_{ij}$$

(B)最遠法(完全聯結法 Complete Linkage)：

$$d_{A,B} = \underset{\substack{i \in A \\ j \in B}}{Max}\, d_{ij}$$

(C)平均法(Average Linkage)：

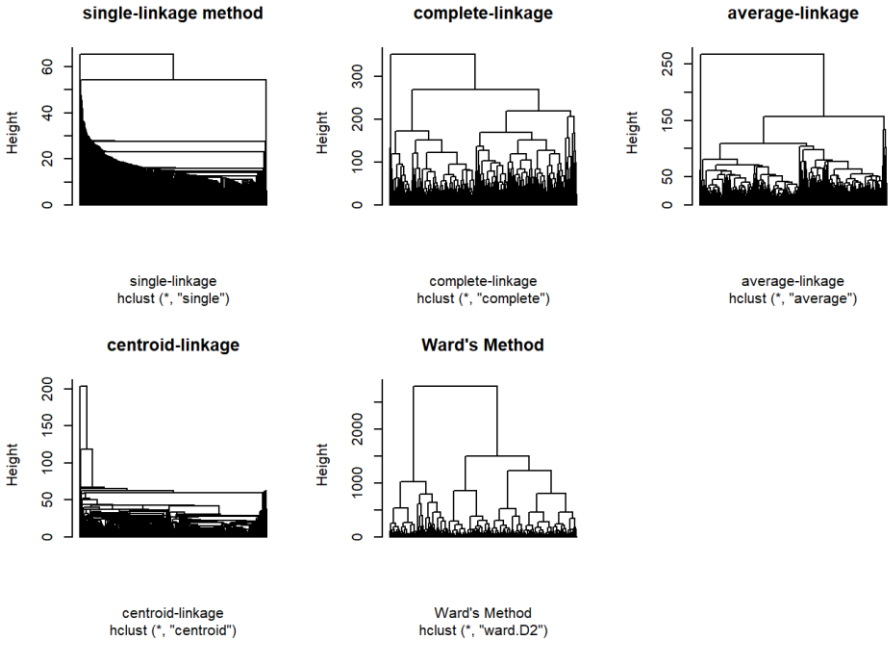
$$d_{A,B} = \Sigma\Sigma\, d_{ij} / n$$ ，n 為全部距離的個數

(D)中心法(Centroid Method)：

$$d_{A,B} = d(\bar{\bar{x}}_A, \bar{\bar{x}}_B) = \| \bar{x}_A - \bar{x}_B \|^2$$

(E)華德法(Wards Method 華德最小變異法)：

$$d_{A,B} = n_A \| \bar{x}_A - \bar{\bar{x}} \|^2 + n_B \| \bar{x}_B - \bar{\bar{x}} \|^2$$



(圖 13)

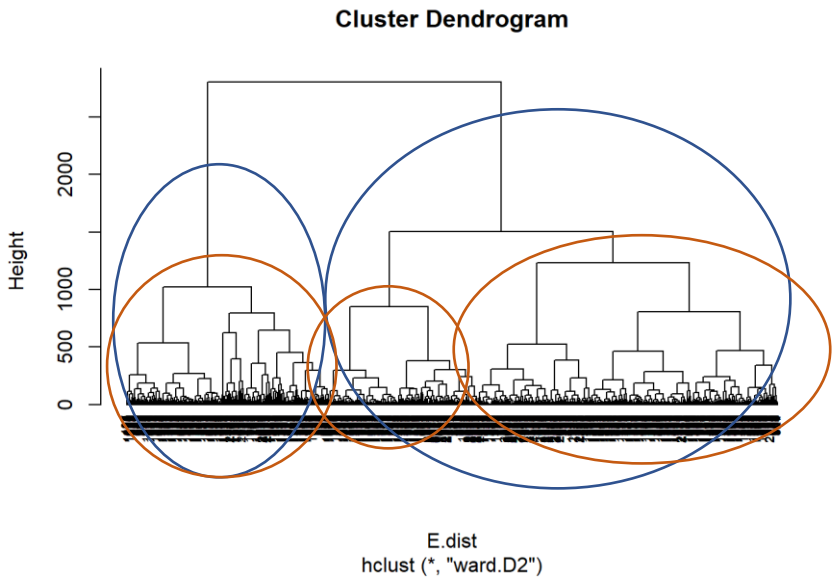
我採用歐式距離搭配不同聚合演算法，並算出聚合係數(agglomerative coefficient)衡量群聚結構被辨識的程度，聚合係數越接近 1 代表有堅固的群

聚結構(strong clustering structure)。

average	single	complete	ward
0.9635618	0.8641109	0.9715241	0.9963991

(表一)

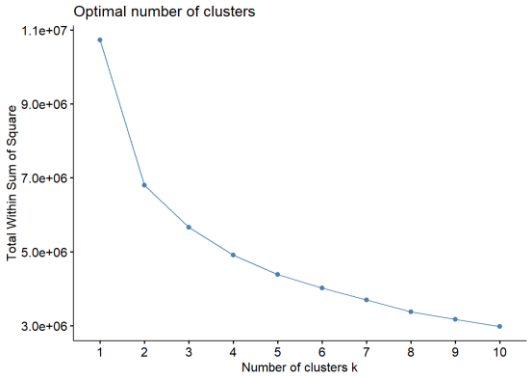
表一為以歐式距離搭配不同聚合演算法算出的聚合係數，在這個個資料中使用歐式距離搭配華德連結演算法的群聚係數有高達 99%的表現。



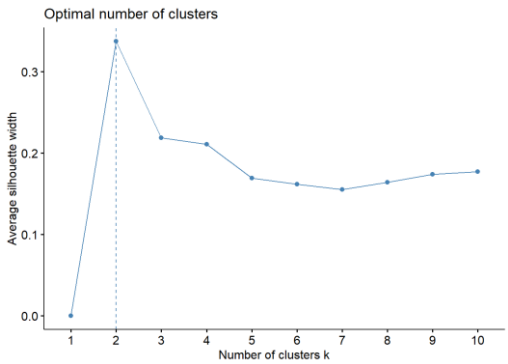
(圖 14)

圖 14 為歐式距離搭配華德連結演算法的結果，從樹狀圖看來，這份資料可以被分成兩群也可以被分成三群。

因此我們要決定合適的分群數量，我使用了剛剛說過的 elbow method 以及 Average silhouette Method，並將方法從 kmeans 改成 hcut，所得到的結果仍舊是 k=2 比較適當。



(圖 15)



(圖 16)

C. Gaussian Mixture Model (GMM)

在 GMM Clustering 中會實現聚類最大化(expectation-maximization ,EM)，一值持續實作 E-M 步驟，重複直到收斂：
E 步驟：對於每個點，找到每個聚類中成員的權重。
M 步驟：對於每個群集「權重」，根據所有數據點更新其位置。

我先對資料做標準化，並利用 GMM 的套件，找出最適當的 GMM 分群，由於上面兩個方法所得到的分群數都是二，我在這裡同樣也設定了二是最適當的分群數，最後得到了分群結果，如表二所示。

Cluster 0	Cluster 1
1884	242

(表二)

Cluster 1(正常)	Cluster 2(疑似)	Cluster 3(病理)
1655	295	176

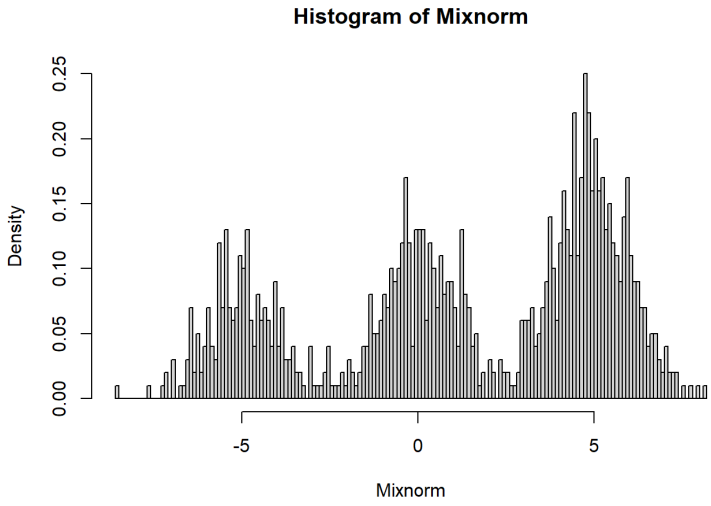
(表三)

表三為原本的的 label 資料，將兩者對比可以發現，正常的數量上升了，而剩下的患者被分在了同一群，也就是說有部分患者的結果並無法被正確分類，可能原本正常的患者被分到疑似或病理，或是疑似為患者的人被分到正常群體。

4. 資料生成與分群 Data Simulation

A. Mixed Normal distribution

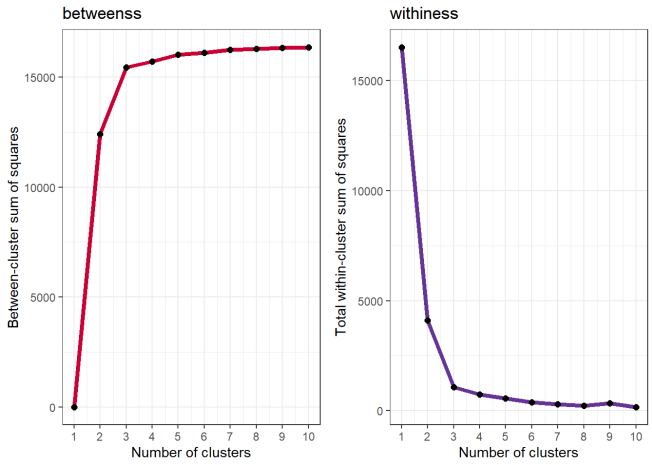
我首先生成了三個不同平均和變異數的常態分佈，並利用上述的三個方法 K-Means Clustering, Hierarchical Clustering, 和 Gaussian Mixture Model(GMM)，對這個 simulate 出的資料進行分群。
我設定的 distribution 是有 1000 個值，分別從 N(0,1), N(-5,1), 和 N(5,1)之中抽出來，而 mixing probability 是 0.3, 0.2, 0.5，也就是說各自從 N(0,1), N(-5,1), 和 N(5,1)抽取出的樣本分別會有 300, 200, 500 個。



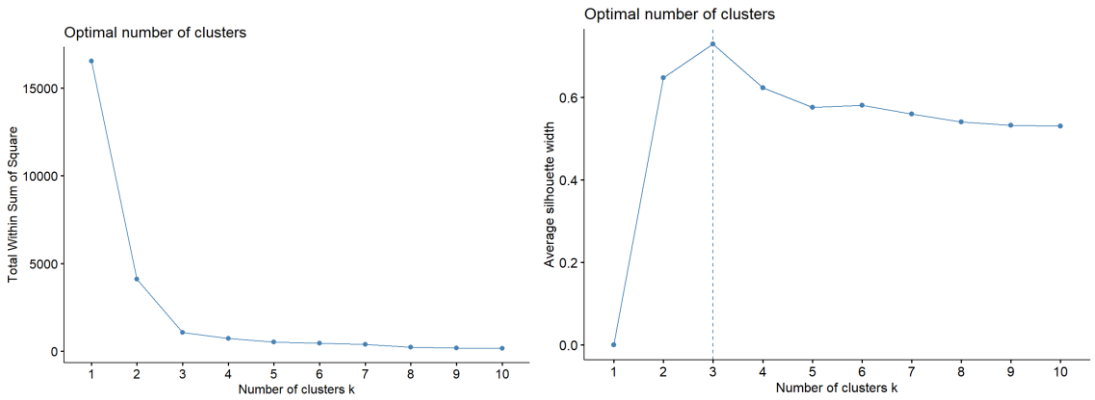
(圖 17)

圖 17 為這三個常態分佈的直方圖，從圖中很明顯可以看出來他們是屬於三個不一樣的族群，因此我期待分群結果可以很清楚地呈現 n=3 的這個結果。

圖 18 為這個 simulation data 的群間和群內資料。

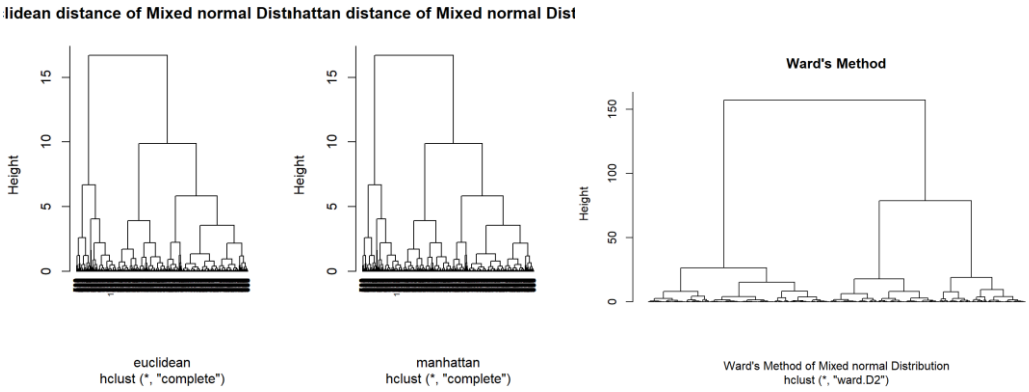


(圖 18)



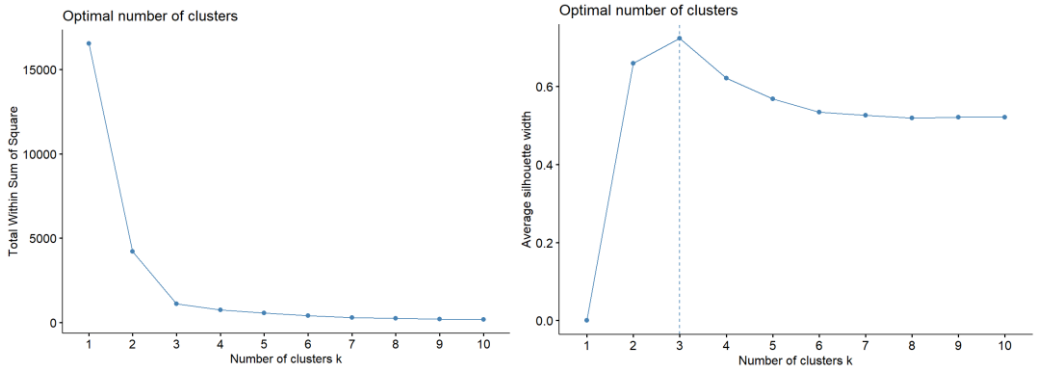
(圖 19)　　(圖 20)

圖 19、20 為這份資料以 kmeans 用 elbow method 以及 Average silhouette Method 所得到的分群結果，結果為 n=3，而這和我所期待的結果符合。



(圖 21)　　(圖 22)

我利用樹狀圖來或出資料的距離以及結構，圖 21 中左邊的那張圖為歐式距離，右邊的為曼哈頓距離。圖 22 為歐式距離搭配華德連結演算法的結果



(圖 23)　　(圖 24)

圖 19、20 為這份資料以 hcut 用 elbow method 以及 Average silhouette Method 所得到的分群結果，結果仍然為 n=3，而這也和我所期待的結果符合。

	Cluster1	Cluster2	Cluster3
data	300	200	500
K-means	306	223	471
GMM	307	218	475

(表四)

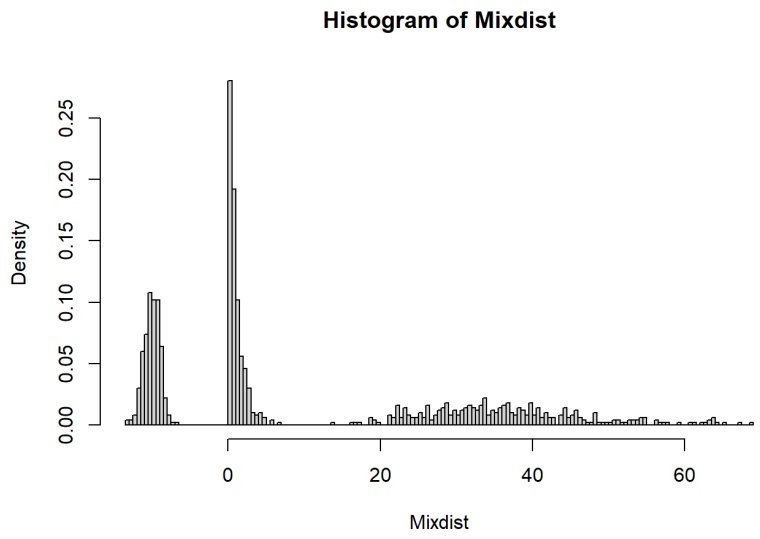
從表四可以看出 GMM 的結果比 K-means 更好一點點，而這是因為這個混和分布的組成元素都是常態(高斯)分佈，因此 GMM 的分配效果會更好。

因為這個問題原本就是完全由 Mixture Gaussian Distribution 生成的資料，GMM （如果能求得全域最優解的話）顯然是可以對這個問題做到的最好的建模。

B. Mixed Normal distribution

我首先生成了三個不同的分佈，並利用上述的三個方法 K-Means Clustering, Hierarchical Clustering, 和 Gaussian Mixture Model(GMM)，對這個 simulate 出的資料進行分群。

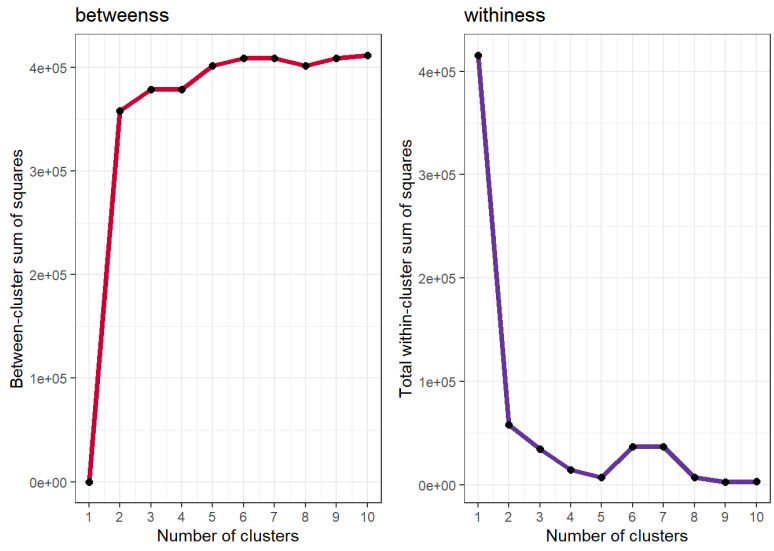
我設定的 distribution 是有 1000 個值，分別從 N(0,1), gamma(12,3), 和 exp(1) 之中抽出來，而 mixing probability 是 0.3, 0.35, 0.35，也就是說各自從 N(0,1), gamma(12,3), 和 exp(1)抽取出的樣本分別會有 300, 350, 350 個。



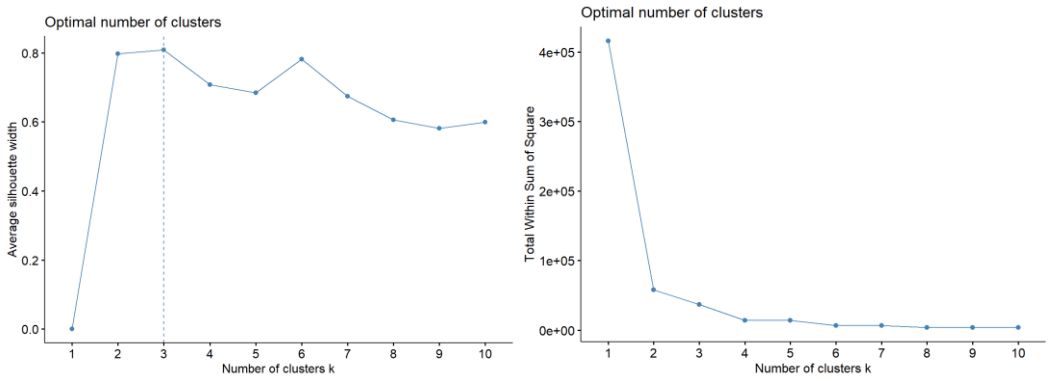
(圖 25)

圖 25 為這三個常態分佈的直方圖，從圖中很明顯可以看出來他們是屬於三個不一樣的族群，因此我期待分群結果可以很清楚地呈現 n=3 的這個結果。

圖 26 為這個 simulation data 的群間和群內資料。



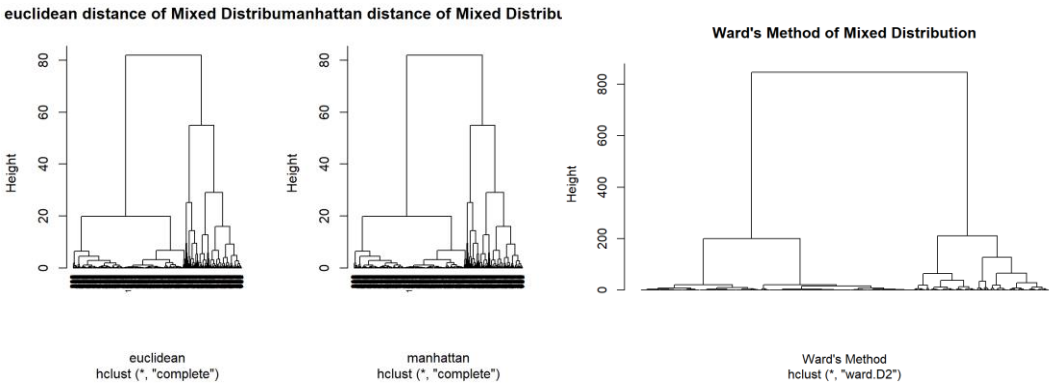
(圖 26)



(圖 27)

(圖 28)

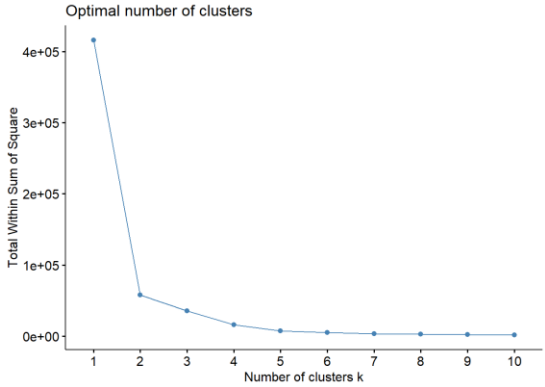
圖 27、28 為這份資料以 kmeans 用 elbow method 以及 Average silhouette Method 所得到的分群結果，結果為 n=3，而這和我所期待的結果符合。



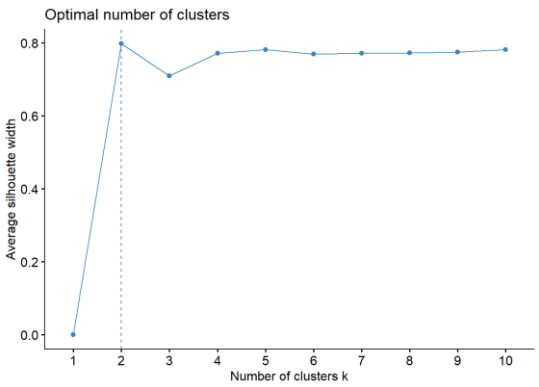
(圖 29)

(圖 30)

我利用樹狀圖來或出資料的距離以及結構，圖 29 中左邊的那張圖為歐式距離，右邊的為曼哈頓距離。圖 30 為歐式距離搭配華德連結演算法的結果



(圖 31)



(圖 32)

圖 19、20 為這份資料以 hcut 用 elbow method 以及 Average silhouette Method 所得到的分群結果，結果卻是 n=2，然而這並不符合我所期待的結果。我認為可能的原因是因為就距離上來看這三個分布的距離並沒有到很遠，所以樹狀圖才會顯示出分成兩群。

	Cluster1	Cluster2	Cluster3
data	300	350	350
K-means	295	380	325
GMM	295	360	345

(表 5)

從表五可以看出 GMM 的結果比 K-means 更好一些。

其實 GMM 和 K-means 的反覆運算求解法其實非常相似，都可以追溯到 EM 演算法，因此 GMM 並不能保證總是能取到最佳解，如果取到不好的初始值，就有可能得到很差的結果。對於 K-means 的情況，我們通常是重複一定次數然後取最好的結果，不過 GMM 每一次反覆運算的計算量比 K-means 要大許多，GMM 所得的結果（Px）不僅僅是資料點的 label ，而包含了資料點標記為每個 label 的概率，很多時候這實際上是非常有用的資訊，所以在分群的效果上會更好一些。

5. 討論(Discussion)

- 上述三個資料分析方法都顯示出了同一個結果，就是部分患者的族群並無法被正確的歸類到他原本的 label 之中。我認為分群的數量並沒有辦法和資料中所提供的 label 一樣的原因是由於資料中的 label 是專家集中他們的智慧、經驗、醫學知識以及臨床症狀所提供的判斷，若是想以資料中所提供的數據並分群達到相同的結果，我覺得會很困難，因為每一個患者的情況不同，即便擁有相同的數據並不代表就會有相同的結果，

這也是為甚麼在 k-means 的分群中會有許多的重疊，Hierarchical Clustering 可以被分成三群，也可以被分成兩群，GMM 所分群出來的結果也跟 label 相差甚遠，三者的準確度都很低。

- 機器對於那些很容易分辨的情況（患病或者不患病的機率很高）可以自動區分，但對於那種很難分辨的情況，比如，像我所使用的 data 的情況，疑似和病理性的機率其實非常相近，如果僅僅簡單地使用某一個閾值診斷患者的話，風險是非常大的，因此，在機器對自己的結果把握很小的情況下，會“拒絕發表評論”，而把這個任務留給有經驗的醫生去解決。
- 雖然無法單以分群來得出病患的正確群體，但是可以透過機器學習的方式建立一個適當的模型，並對之後的資料進行預測，這樣的話，若是之後拿到新的資料便可以對它進行預測，並先掌握患者可能的群體，並對其進行檢查，提早治療。
- 因為是自己模擬出來的資料，所以可以確定分群數的數量，而利用分群分群方法分出來的結果也會跟所預料的結果非常相近，準確度也很高，比起真實世界的資料，模擬的資料更好分群，因為沒有人知道真實世界的資料實際上的分群數，所以很難直接找出一個分群的數量，將所有資料去做分類，並且獲得很高的準確率。我上面做的那個心電圖的資料分析就是一個很好的例子，即便知道 label，卻仍舊無法準確分群。