統計計算期中書面報告

題目：英國大學生問卷資料分析

系級/學號/姓名：計財所 108071601 賴冠維

資料來源：https://www.kaggle.com/miroslavsabo/young-people-survey

目錄

一、 資料介紹

二、 以 CART 對遺失值填補

三、 以 Hierarchical Clustering 探索資料

四、 以 Xgboost 預測酒精上癮的學生並找出相關特徵

五、 針對 Xgboost 篩選出的變數，以因素分析找出變數間的相關

六、 將各個問題以 Xgboost 模型所得到的 Importance Value 進行加權，並以 Gaussian Mixture Model(GMM)進行分類

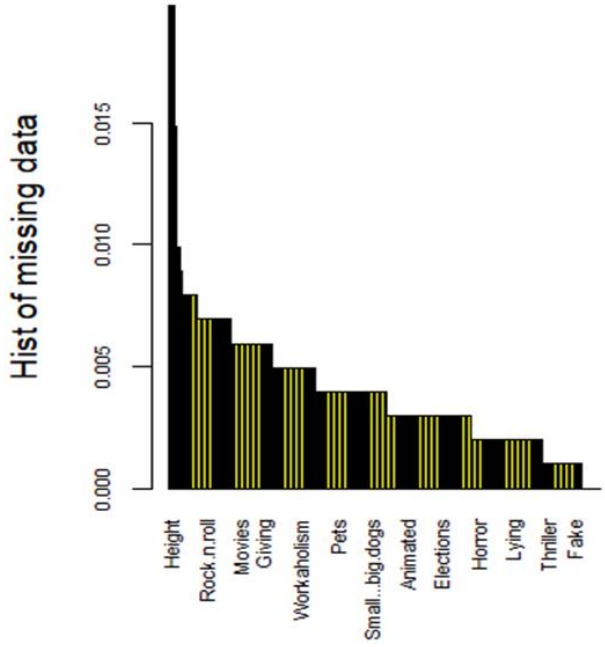
一、　資料介紹

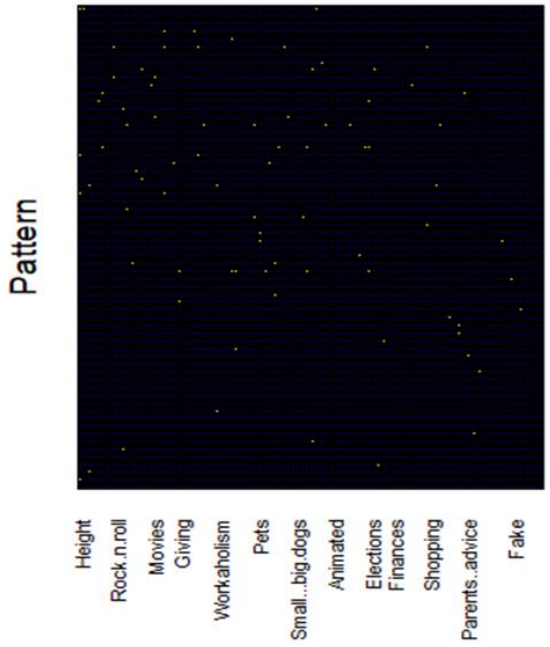
本筆資料為英國大學 Faculty of Social and Economic Sciences 對其學生進行問卷調查，所得到的填答資料，主要有八大類的問題，為音樂偏好、電影偏好、對什麼恐懼、嗜好興趣、人格特質、消費習慣、健康習慣以及基本的人口統計，共蒐集 1010 筆資料，取資料一小部分如下：

```
'data.frame':   1010 obs. of  150 variables:
 $ Music                      : Factor w/ 5 levels "1","2","3","4",..: 5 4 5 5 5 5 5 5 5 5 ..
 $ Slow.songs.or.fast.songs   : Factor w/ 5 levels "1","2","3","4",..: 3 4 5 3 3 3 5 3 3 3 ..
 $ Dance                      : Factor w/ 5 levels "1","2","3","4",..: 2 2 2 2 4 2 5 3 3 2 ..
 $ Folk                       : Factor w/ 5 levels "1","2","3","4",..: 1 1 2 1 3 3 3 2 1 5 ..
 $ Country                    : Factor w/ 5 levels "1","2","3","4",..: 2 1 3 1 2 2 1 1 1 2 ..
 $ Classical.music            : Factor w/ 5 levels "1","2","3","4",..: 2 1 4 1 4 3 2 2 2 2 ..
 $ Musical                    : Factor w/ 5 levels "1","2","3","4",..: 1 2 5 1 3 3 2 2 4 5 ..
 $ Pop                        : Factor w/ 5 levels "1","2","3","4",..: 5 3 3 2 5 2 5 4 3 3 ..
 $ Rock                       : Factor w/ 5 levels "1","2","3","4",..: 5 5 5 2 3 5 3 5 5 5 ..
 $ Metal.or.Hardrock          : Factor w/ 5 levels "1","2","3","4",..: 1 4 3 1 1 5 1 1 5 2 ..
 $ Punk                       : Factor w/ 5 levels "1","2","3","4",..: 1 4 4 4 2 3 1 2 1 3 ..
 $ Hiphop..Rap                : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 2 5 4 3 3 1 2 ..
 $ Reggae..Ska                : Factor w/ 5 levels "1","2","3","4",..: 1 3 4 2 3 3 1 2 2 4 ..
 $ Swing..Jazz                : Factor w/ 5 levels "1","2","3","4",..: 1 1 3 1 2 4 1 2 2 4 ..
 $ Rock.n.roll                : Factor w/ 5 levels "1","2","3","4",..: 3 4 5 2 1 4 2 3 2 4 ..
 $ Alternative                : Factor w/ 5 levels "1","2","3","4",..: 1 4 5 5 2 5 3 1 5 4 ..
 $ Latino                     : Factor w/ 5 levels "1","2","3","4",..: 1 2 5 1 4 3 3 2 1 5 ..
 $ Techno..Trance             : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 2 2 1 5 3 1 1 ..
```

二、　遺失值填補

將為各變數的遺失值標示出來，可發現整體資料缺失值不太多，本篇以 CART 演算法，針對每個變數間的缺失值進行估計，並填補。





三、 Hierarchical Clustering

以 Ward 最小組內變異法作為 Hierarchical 的合併標準，Ward 法主要想法，為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)，其定義如下：。

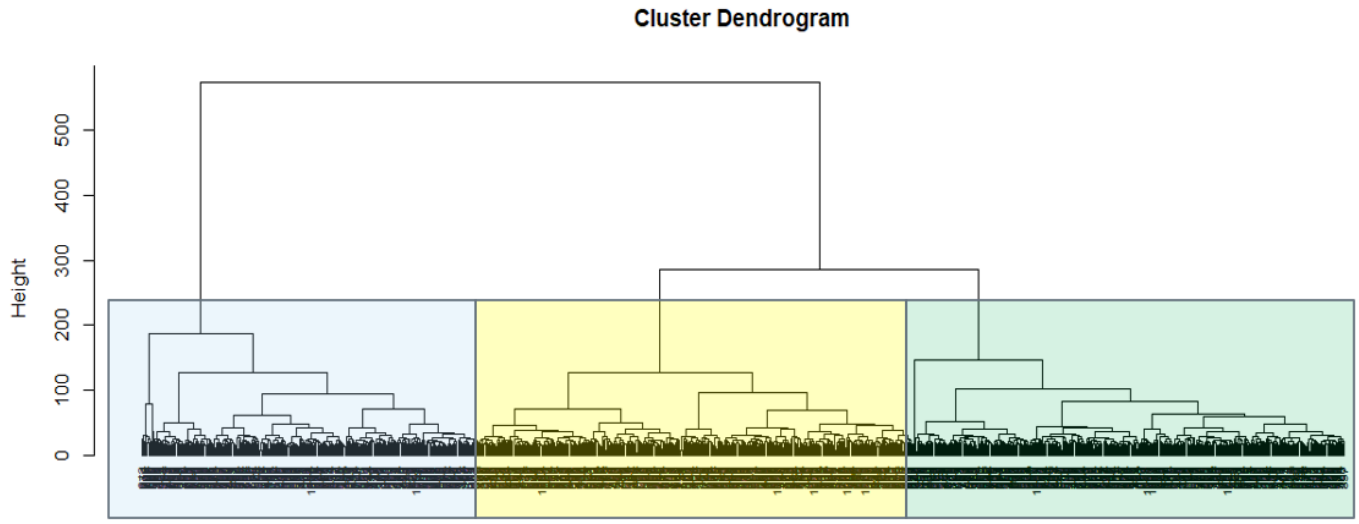
Ward's minimum variance method：

$$Total\ ESS = ESS_1 + ESS_2 + \cdots + ESS_k$$

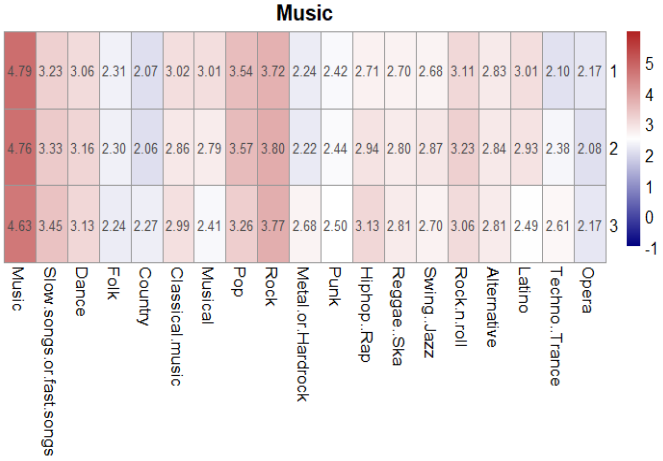
$$ESS_k = \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^T \left(x_{ij} - \bar{x}_i\right)$$

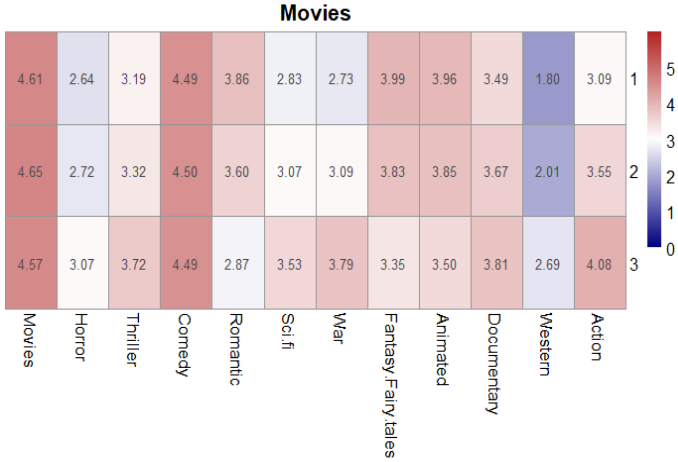
- x_{ij} : $j^{th}\ number\ of component\ in\ i^{th}\ cluster$
- $\bar{x}_i$: $Mean\ of\ the\ i^{th}\ cluster$

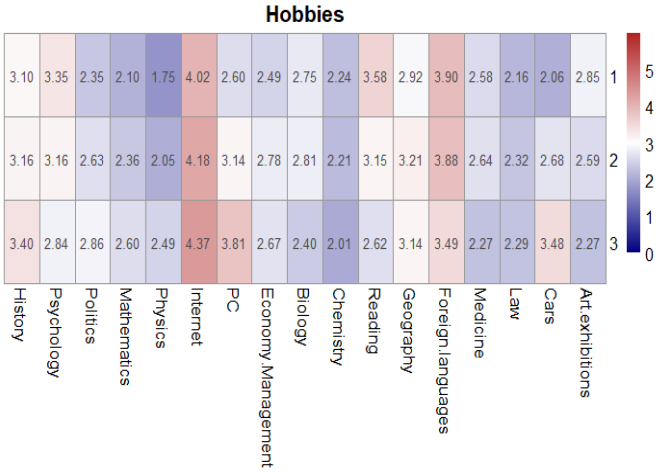
做出樹狀圖後，本篇決定分為三群，結果如下：

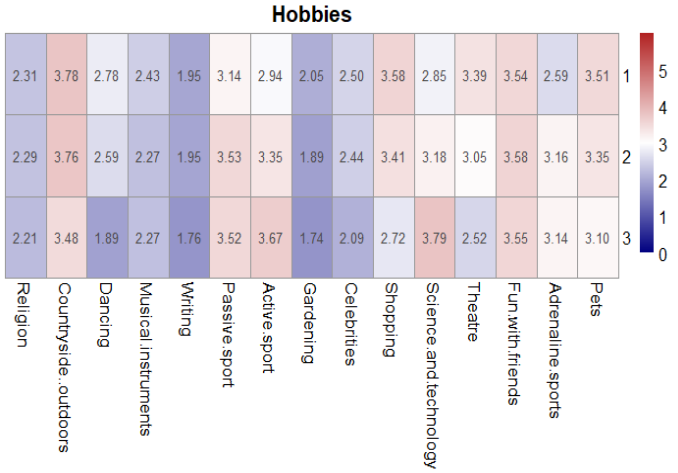


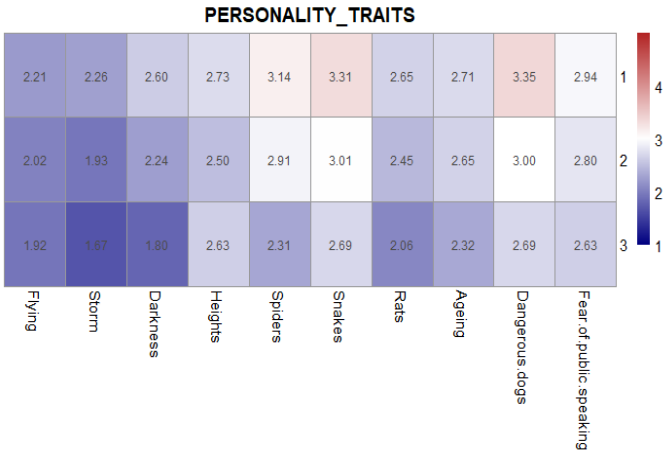
以這三群資料進行探索性資料分析，將每組問卷的填答資料計算取得該組平均，並將其數值以熱力圖表示，如下圖：

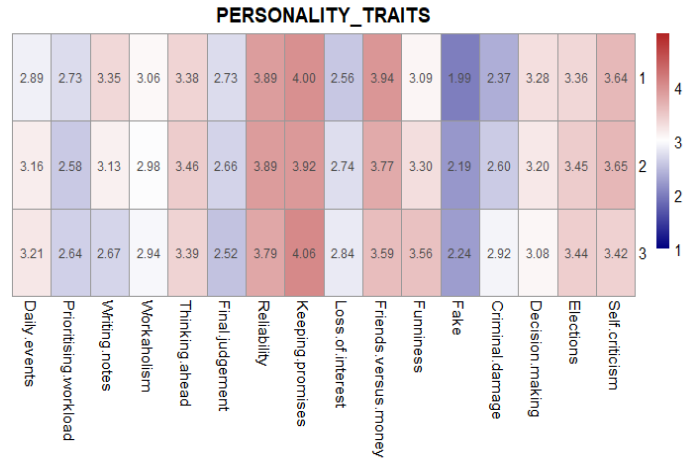

Music
Music Slow.songs.or.fast.songs Dance Folk Country Classical.music Musical Pop Rock Metal.or.Hardrock Punk Hiphop..Rap Reggae..Ska Swing..Jazz Rock.n.roll Alternative Latino Techno..Trance Opera
4.79 3.23 3.06 2.31 2.07 3.02 3.01 3.54 3.72 2.24 2.42 2.71 2.70 2.68 3.11 2.83 3.01 2.10 2.17 1
4.76 3.33 3.16 2.30 2.06 2.86 2.79 3.57 3.80 2.22 2.44 2.94 2.80 2.87 3.23 2.84 2.93 2.38 2.08 2
4.63 3.45 3.13 2.24 2.27 2.99 2.41 3.26 3.77 2.68 2.50 3.13 2.81 2.70 3.06 2.81 2.49 2.61 2.17 3
5 4 3 2 1 0 -1

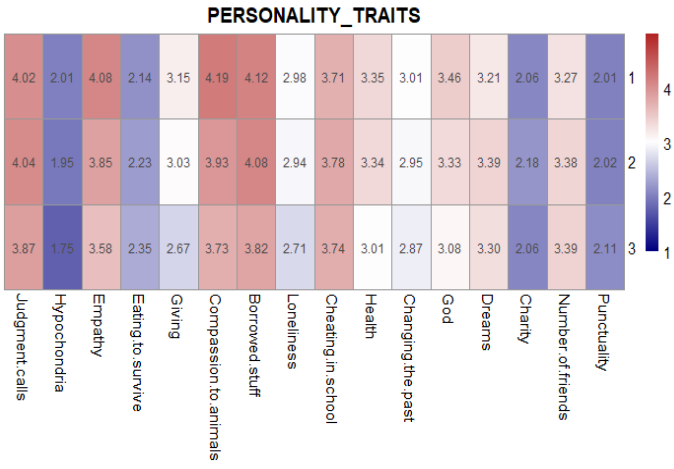

Movies
Movies Horror Thriller Comedy Romantic Sci.fi War Fantasy.Fairy.tales Animated Documentary Western Action
4.61 2.64 3.19 4.49 3.86 2.83 2.73 3.99 3.96 3.49 1.80 3.09 1
4.65 2.72 3.32 4.50 3.60 3.07 3.09 3.83 3.85 3.67 2.01 3.55 2
4.57 3.07 3.72 4.49 2.87 3.53 3.79 3.35 3.50 3.81 2.69 4.08 3
5 4 3 2 1 0

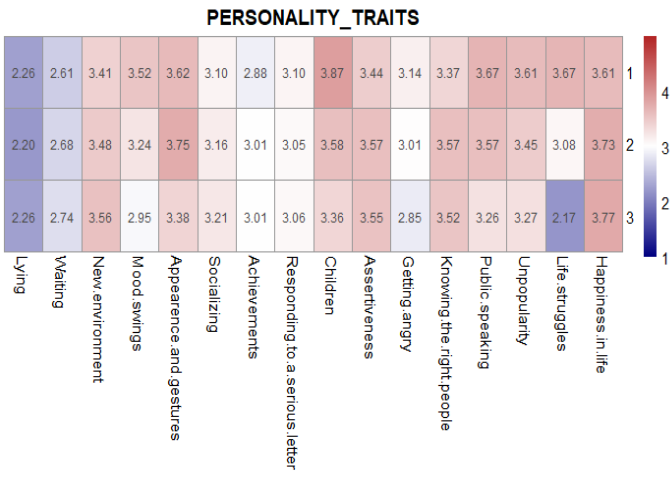

Hobbies
History Psychology Politics Mathematics Physics Internet PC Economy.Management Biology Chemistry Reading Geography Foreign.languages Medicine Law Cars Art.exhibitions
3.10 3.35 2.35 2.10 1.75 4.02 2.60 2.49 2.75 2.24 3.58 2.92 3.90 2.58 2.16 2.06 2.85 1
3.16 3.16 2.63 2.36 2.05 4.18 3.14 2.78 2.81 2.21 3.15 3.21 3.88 2.64 2.32 2.68 2.59 2
3.40 2.84 2.86 2.60 2.49 4.37 3.81 2.67 2.40 2.01 2.62 3.14 3.49 2.27 2.29 3.48 2.27 3
5 4 3 2 1 0


Hobbies
Religion Countryside..outdoors Dancing Musical.instruments Writing Passive.sport Active.sport Gardening Celebrities Shopping Science.and.technology Theatre Fun.with.friends Adrenaline.sports Pets
2.31 3.78 2.78 2.43 1.95 3.14 2.94 2.05 2.50 3.58 2.85 3.39 3.54 2.59 3.51 1
2.29 3.76 2.59 2.27 1.95 3.53 3.35 1.89 2.44 3.41 3.18 3.05 3.58 3.16 3.35 2
2.21 3.48 1.89 2.27 1.76 3.52 3.67 1.74 2.09 2.72 3.79 2.52 3.55 3.14 3.10 3
5 4 3 2 1 0


PERSONALITY_TRAITS
Flying Storm Darkness Heights Spiders Snakes Rats Ageing Dangerous.dogs Fear.of.public.speaking
2.21 2.26 2.60 2.73 3.14 3.31 2.65 2.71 3.35 2.94 1
2.02 1.93 2.24 2.50 2.91 3.01 2.45 2.65 3.00 2.80 2
1.92 1.67 1.80 2.63 2.31 2.69 2.06 2.32 2.69 2.63 3
4 3 2 1


PERSONALITY_TRAITS
Daily.events Prioritising.workload Writing.notes Workaholism Thinking.ahead Final.judgement Reliability Keeping.promises Loss.of.interest Friends.versus.money Funniness Fake Criminal.damage Decision.making Elections Self.criticism
2.89 2.73 3.35 3.06 3.38 2.73 3.89 4.00 2.56 3.94 3.09 1.99 2.37 3.28 3.36 3.64 1
3.16 2.58 3.13 2.98 3.46 2.66 3.89 3.92 2.74 3.77 3.30 2.19 2.60 3.20 3.45 3.65 2
3.21 2.64 2.67 2.94 3.39 2.52 3.79 4.06 2.84 3.59 3.56 2.24 2.92 3.08 3.44 3.42 3
4 3 2 1

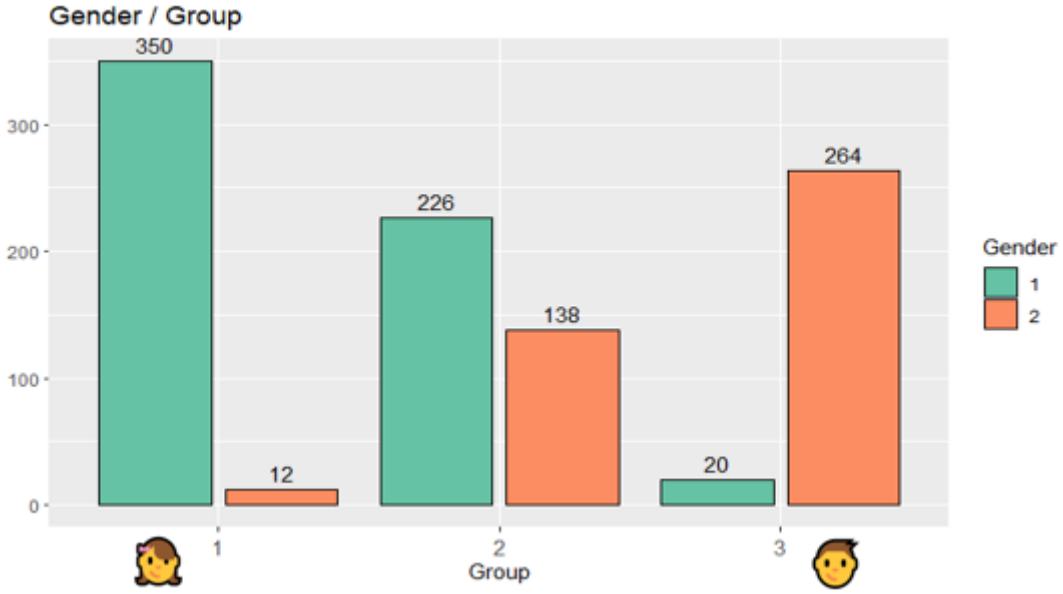

PERSONALITY_TRAITS
Judgment.calls Hypochondria Empathy Eating.to.survive Giving Compassion.to.animals Borrowed.stuff Loneliness Cheating.in.school Health Changing.the.past God Dreams Charity Number.of.friends Punctuality
4.02 2.01 4.08 2.14 3.15 4.19 4.12 2.98 3.71 3.35 3.01 3.46 3.21 2.06 3.27 2.01 1
4.04 1.95 3.85 2.23 3.03 3.93 4.08 2.94 3.78 3.34 2.95 3.33 3.39 2.18 3.38 2.02 2
3.87 1.75 3.58 2.35 2.67 3.73 3.82 2.71 3.74 3.01 2.87 3.08 3.30 2.06 3.39 2.11 3
4 3 2 1


PERSONALITY_TRAITS
Lying Waiting New.environment Mood.swings Appearance.and.gestures Socializing Achievements Responding.to.a.serious.letter Children Assertiveness Getting.angry Knowing.the.right.people Public.speaking Unpopularity Life.struggles Happiness.in.life
2.26 2.61 3.41 3.52 3.62 3.10 2.88 3.10 3.87 3.44 3.14 3.37 3.67 3.61 3.67 3.61 1
2.20 2.68 3.48 3.24 3.75 3.16 3.01 3.05 3.58 3.57 3.01 3.57 3.57 3.45 3.08 3.73 2
2.26 2.74 3.56 2.95 3.38 3.21 3.01 3.06 3.36 3.55 2.85 3.52 3.26 3.27 2.17 3.77 3
4 3 2 1

總結上述各圖呈現的結果，可得以下結論：

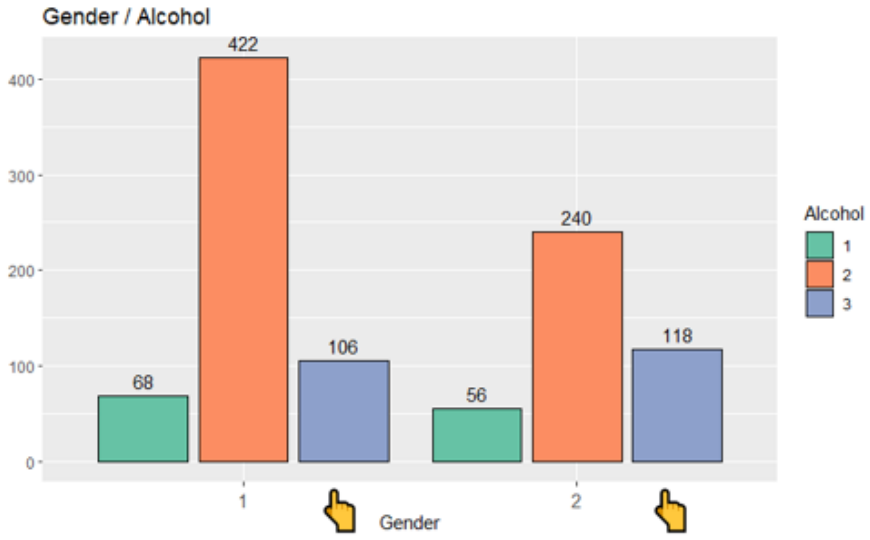
1. 第一組討厭物理，喜歡閱讀、外文，喜歡購物、討厭科學、科技，討厭極限運動並且比較感性。
2. 第三組異常喜歡網路、電腦、車子，不喜歡藝術、討厭跳舞、喜歡運動、極限運動、科學，討厭戲劇、明顯不願意做筆記記事情，比較理性。

最後可以看到第一組明顯由女生組成，第三組主要由男生組成。

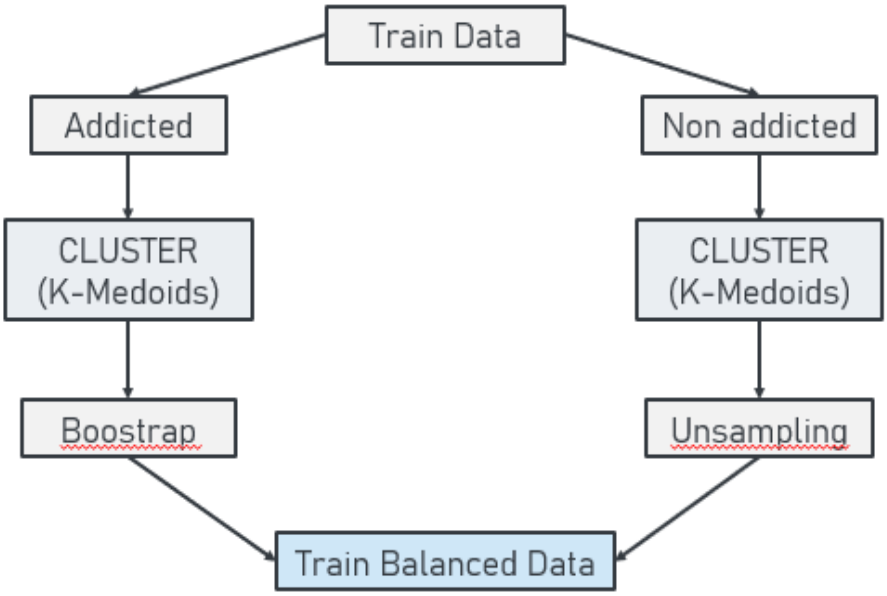


四、 以 Xgboost 預測酒精上癮的學生

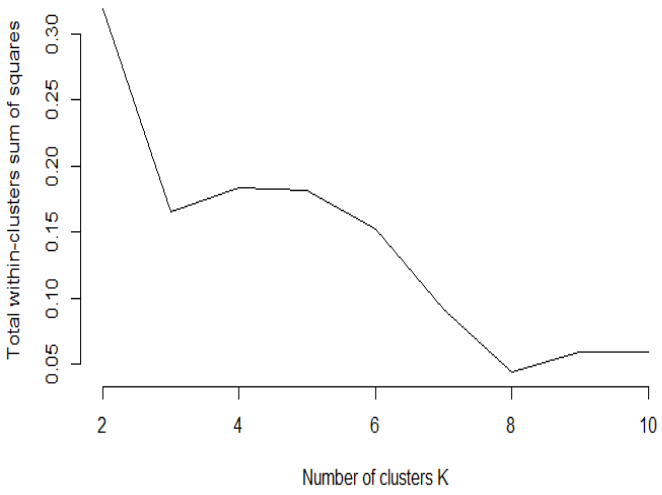
如下圖所示，填答為 3 代表學生勾選飲酒過量的選項，而可發現男性酒精上癮的比例明顯比女性要高。

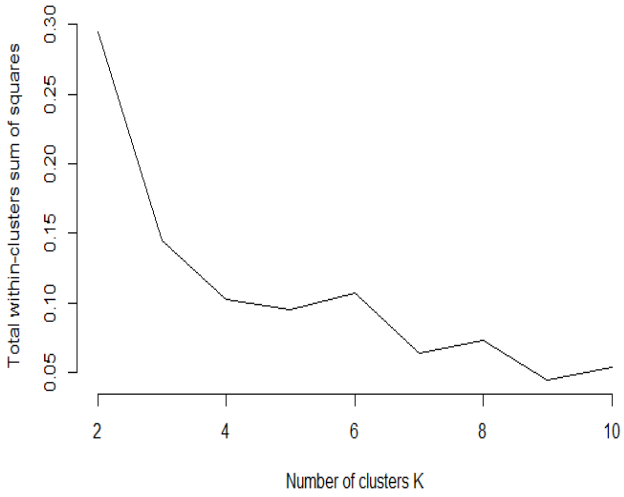


我們針對飲酒過量的學生以 Xgboost 進行預測並且找出有相關的其他特徵，首先我們先將資料以(70%，30%)分為訓練集與測試集，並對訓練集以目標變數結果進行平衡，以保證模型對各種結果都具有預測能力，流程如下：



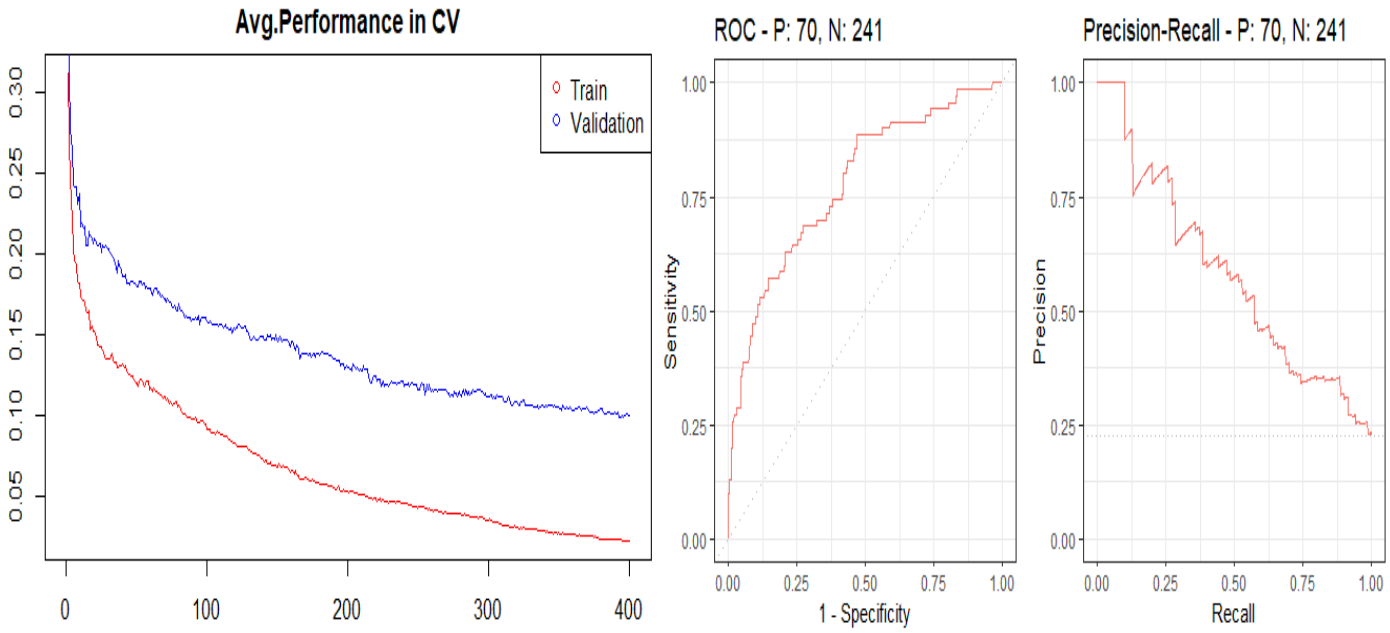
我們將訓練集資料以有酒精上癮、沒酒精上癮分為兩組，並對上癮組進行 Boostrap 重抽樣至兩組數量平衡，而兩組抽樣方法為先將資料以 K-Medoids 以總組內變異最小分群，再對各群抽樣，結果如下：





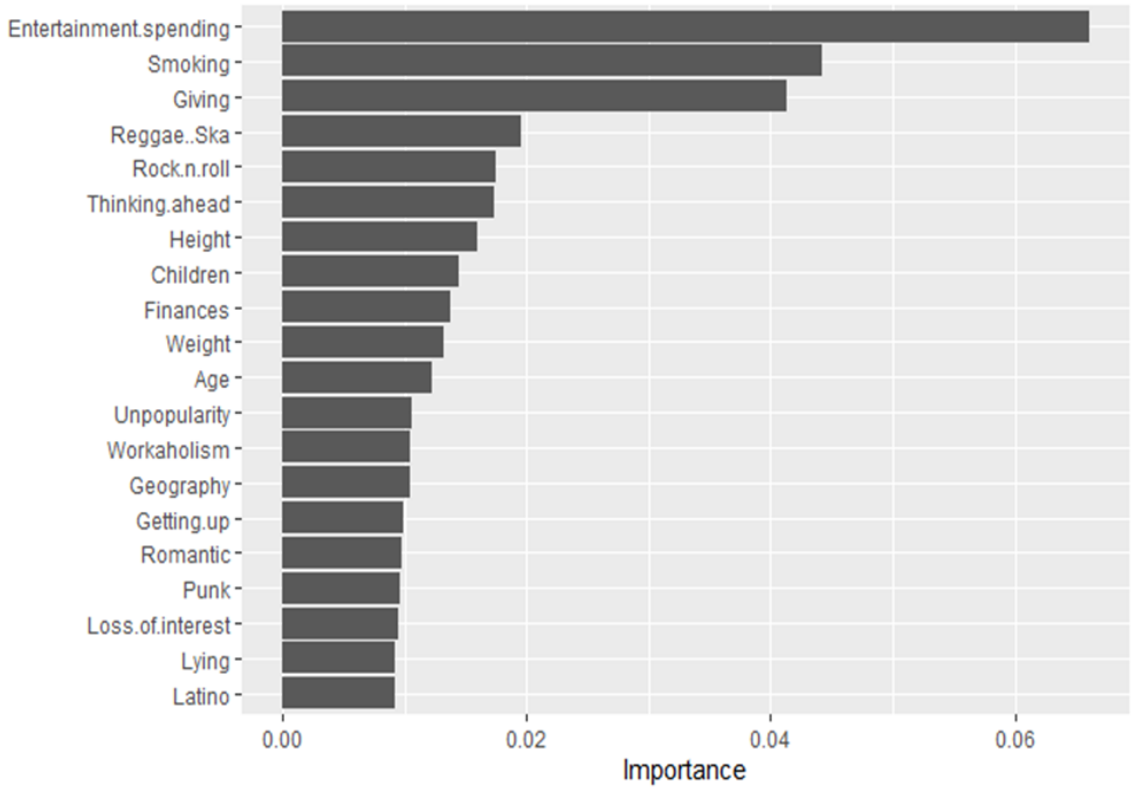
分析一至十群間分幾群有最低的總組內變異，可看到上癮組分八群，而沒上癮組分九群有最低的組內變異，將兩組數量平衡後，形成新的平衡後訓練集。

左圖為平衡訓練集 Train-Error/Validation-Error 在迭代時的表現，可發現大約在 400 左右收斂，右圖為模型在測試集的表現，可看到 ROC 雖然表現還可，但 Recall 表現並不佳，代表模型預測上癮的能力並不算太好。



五、針對 Xgboost 篩選出的變數，以因素分析找出變數間的相關

以 Importance Value 篩選出前 20 有影響力的變數，如下圖：

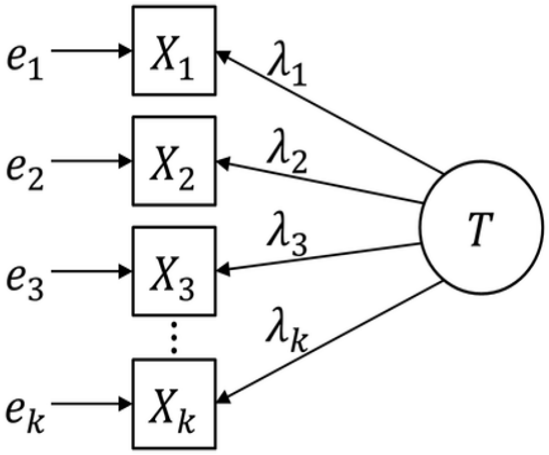


接著使用因素分析，找出篩選變數彼此間，是否存在關聯，下圖簡介因素分析：

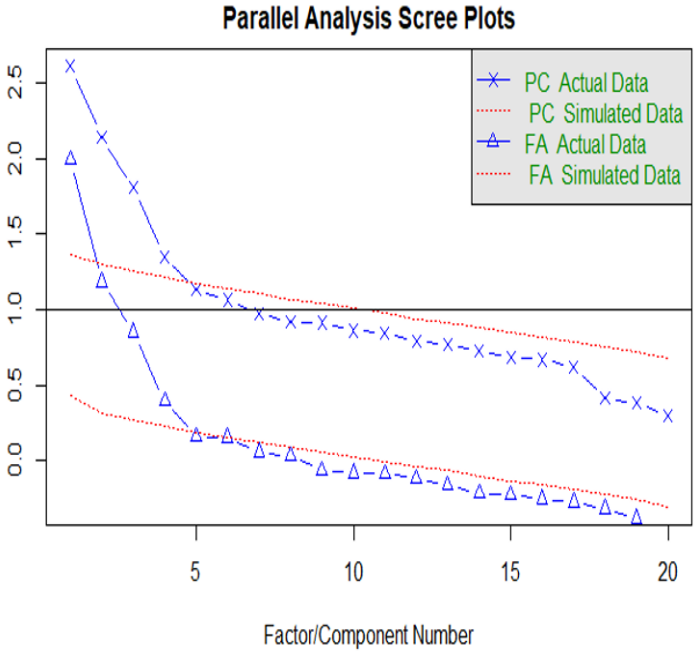
Factor Analysis：

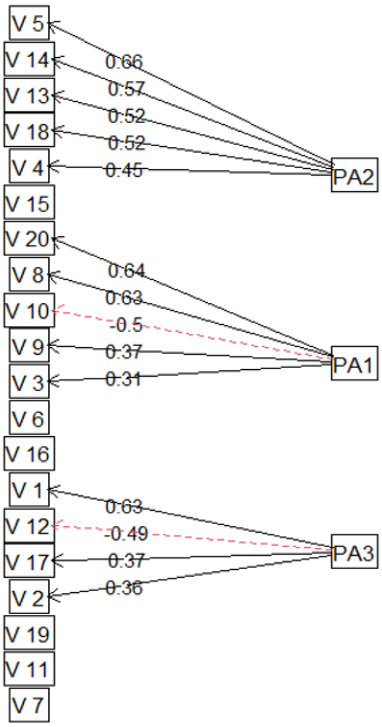
$$\boldsymbol{x = \mu + Fz + \epsilon}$$

- x: *Random Variables*
- μ: *Expection of x*
- F: *Factor Loading*
- z: *Hidden Facotr*
- ϵ: *Idiosyncratic factor*

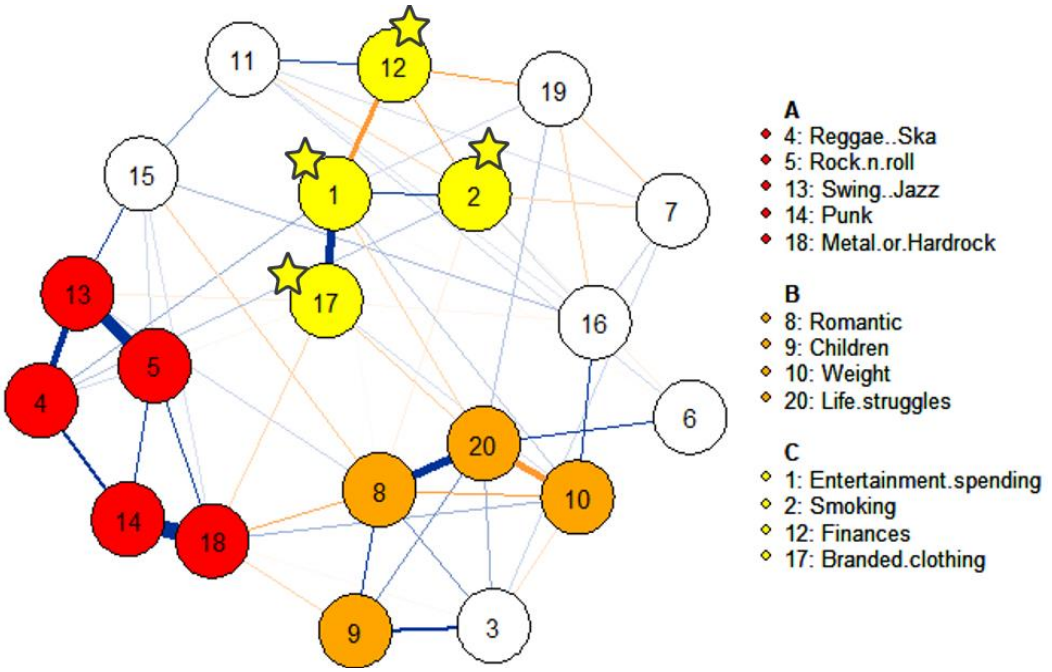


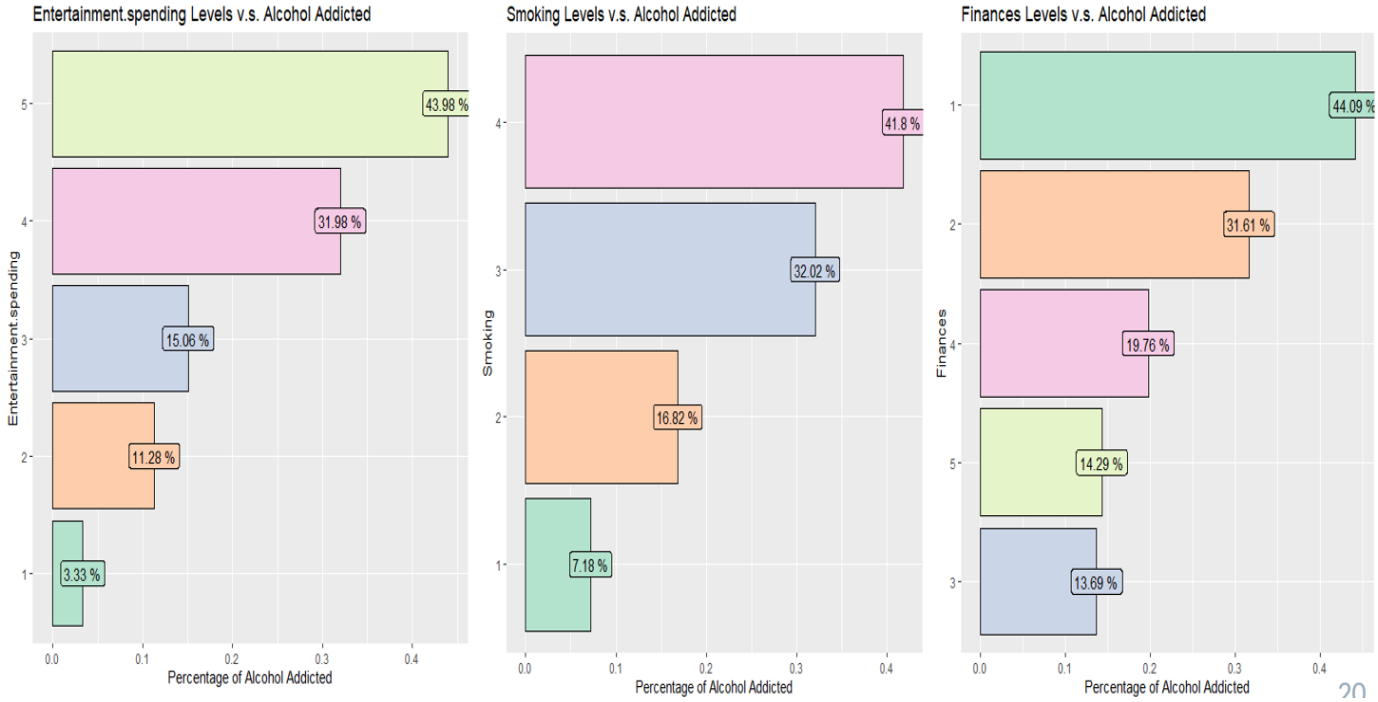
如下圖，我們以 Scree Plots 看到因素分析在第三因子時 Eigen value 便低於 1，因此我們最多選到三因子，而右圖為三因子間，Eigen vector 各變數間的關聯。





將所選的各組變數呈現如下圖，以第 C 組因子為例，可發現該因子將娛樂消費、抽菸、金融觀念、品牌服飾等問題歸類成一組，我們對該因子做探索性資料分析：

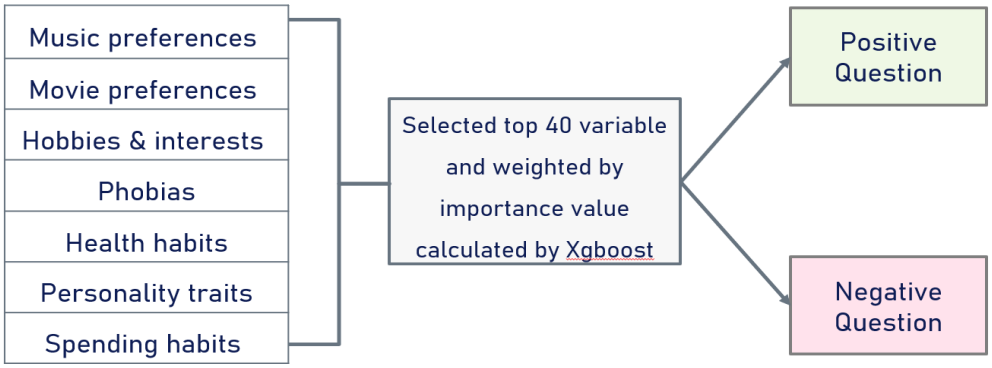


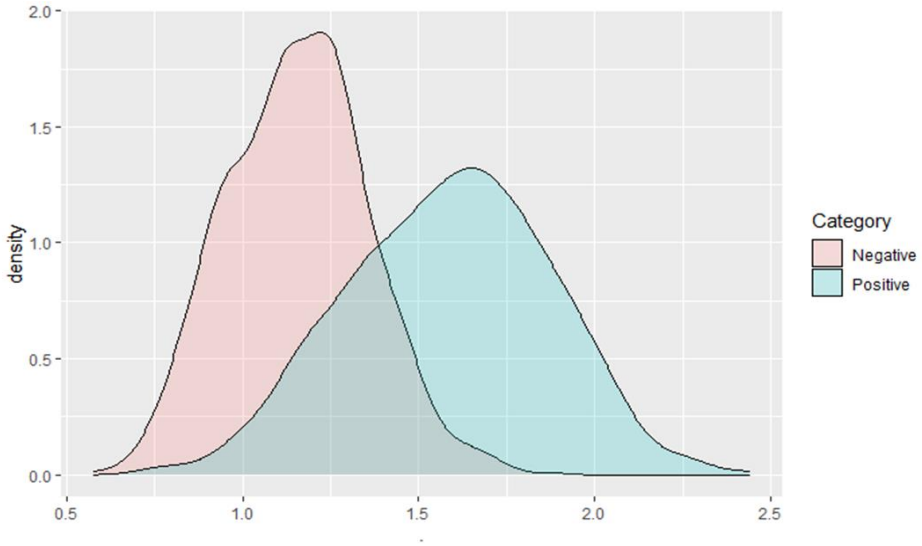


可發現娛樂消費、抽菸的頻率越高的族群裡，酒精上癮的比例也越高，而儲蓄觀念越差的族群酒精上癮的比例也越高。

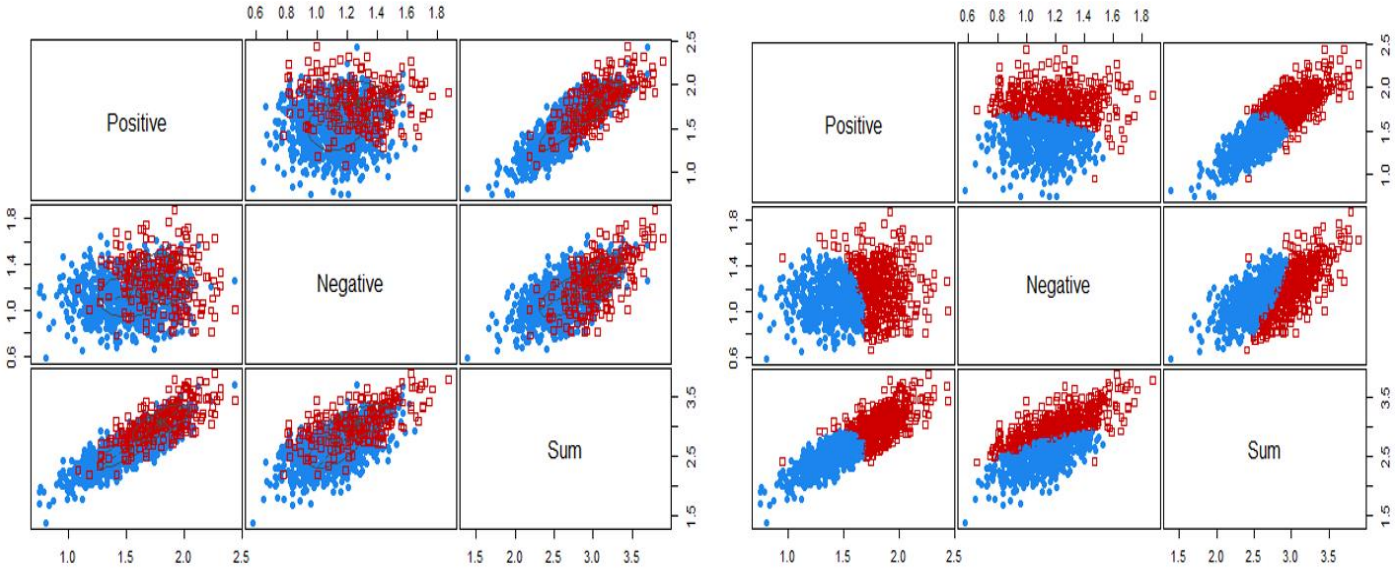
六、 將問卷填答資料加權，以 Gaussian Mixture Model 模型分類

第五章我們分析以 Xgboost 模型篩選出的變數，其變數間的關聯性，而本章我們將所篩選的變數以 Important Value 進行加權，並以問卷題目與上癮呈現正相關或負相關分為兩組，可得新的兩指標，其流程與機率分配圖如下：

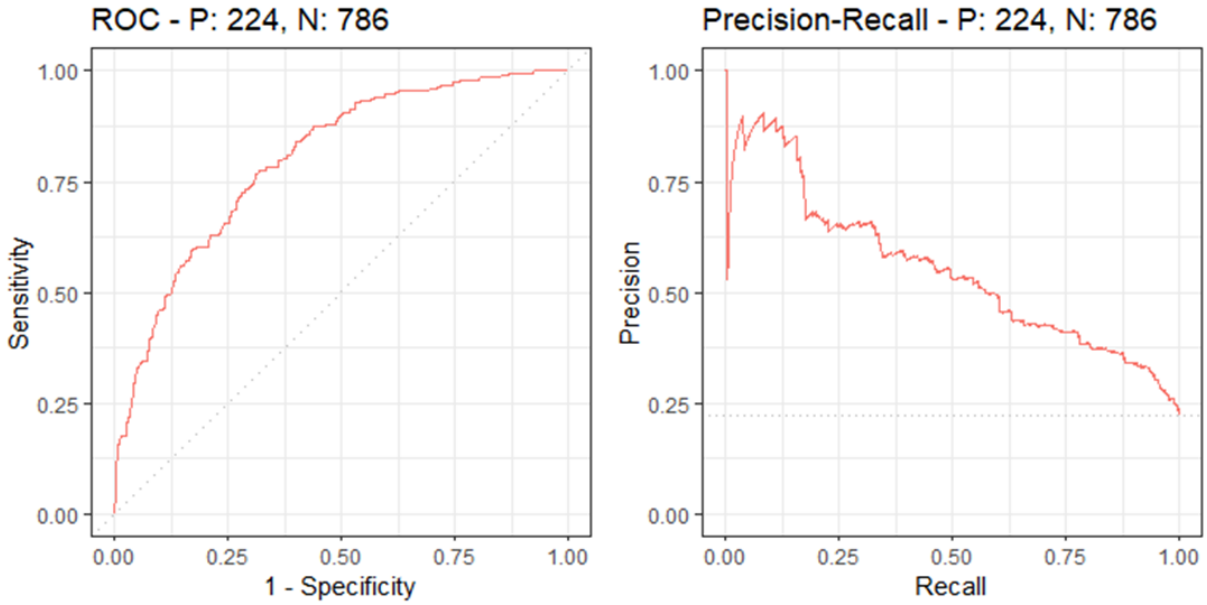




最後我們以正相關問題、負相關問題以及兩者總和進行 GMM 模型建立，可得結果如下：



上圖左邊為真實值，紅色為上癮，藍色為無上癮，上圖右邊為分類結果，可見當分數越高時，上癮的比例也會越高，但可看得出來資料在這兩組變數間重疊的部分很大，可見這種做法並不是個非常有解釋力的變數，我們觀察其模型表現：



可以發現表現比起 Xgboost 來說，並不會表現較差，甚至在 ROC 的表現上，要明顯比 Xgboost 來得更好。