**統計計算期中書面報告**

題目：英國大學生問卷資料分析

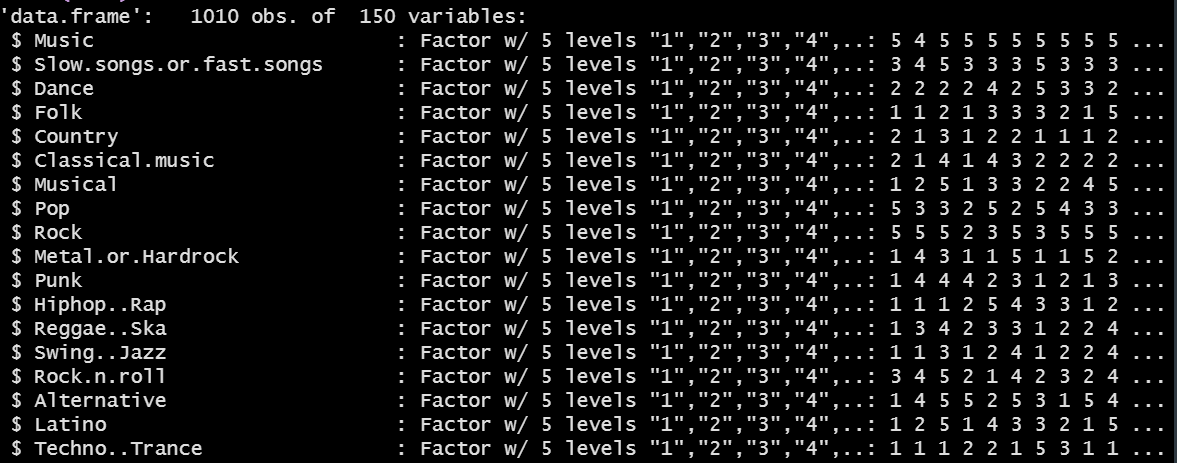
系級/學號/姓名：計財所 108071601 賴冠維

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

**目錄**

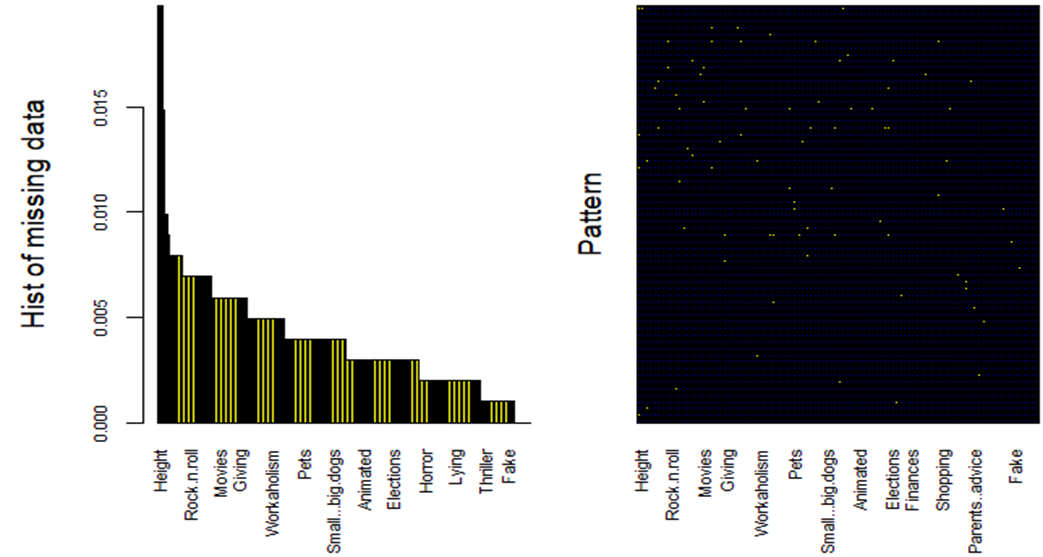
1. 資料介紹
2. 以CART對遺失值填補
3. 以Hierarchical Clustering 探索資料
4. 以Xgboost預測酒精上癮的學生並找出相關特徵
5. 針對Xgboost篩選出的變數，以因素分析找出變數間的相關
6. 將各個問題以Xgboost模型所得到的Importance Value進行加權，並以Gaussian Mixture Model(GMM)進行分類
7. 資料介紹

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：



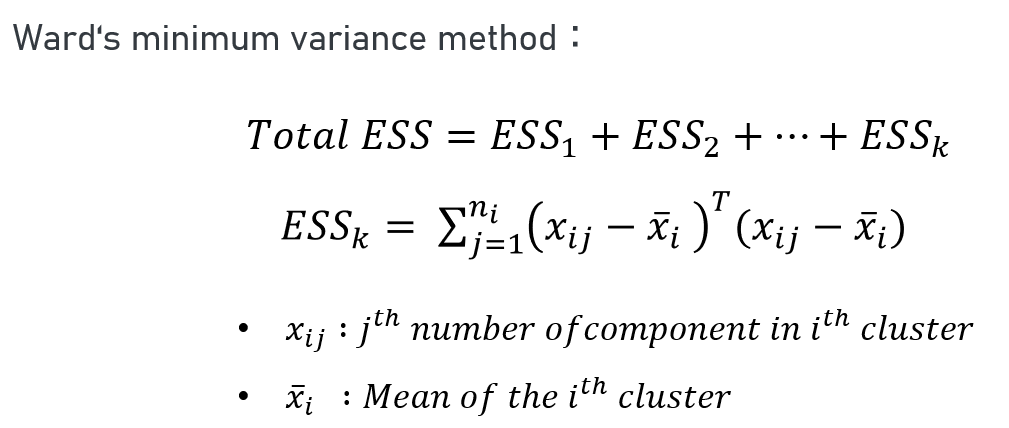
1. 遺失值填補

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。

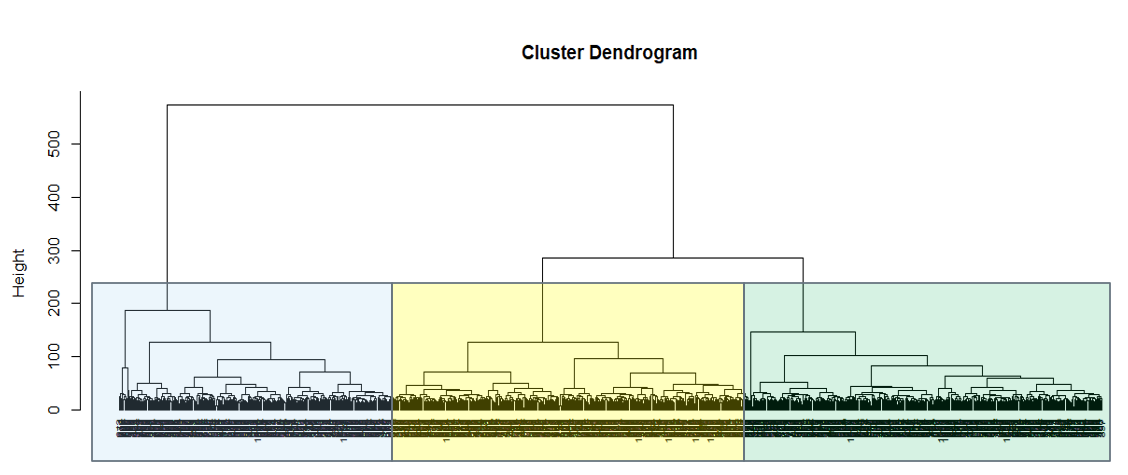


三、 Hierarchical Clustering

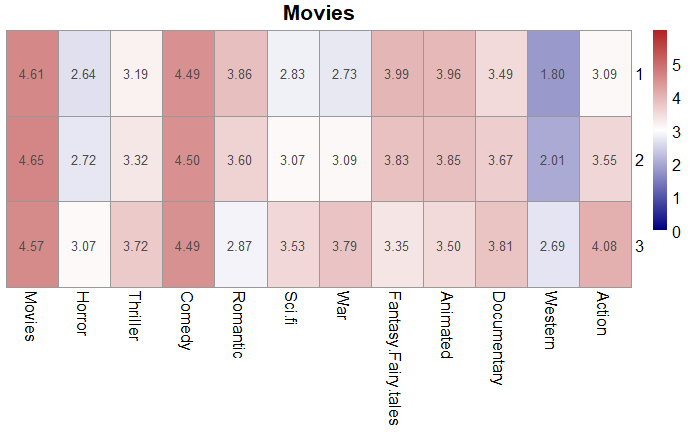
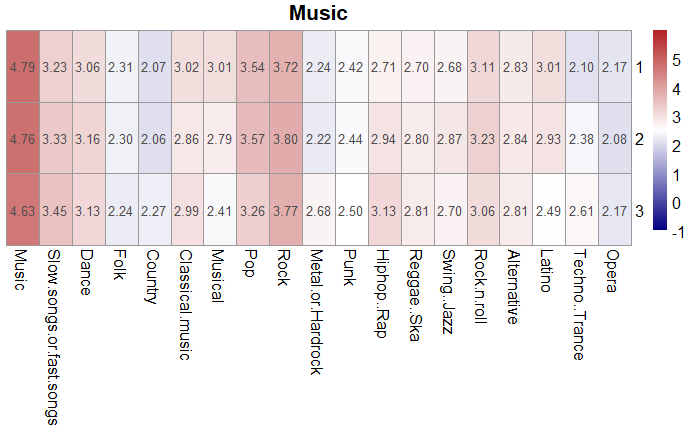
以Ward最小組內變異法作為Hierarchical的合併標準，Ward法主要想法，為反覆比較每對資料合併後的群內總變異數的增量，並找增量最小的組別優先合併。越早合併的子集表示其間的相似度越高。而使用華德最小變異法的前提為，初始各點資料距離必須是歐式距離的平方和(Squared Euclidean Distance)，其定義如下：。

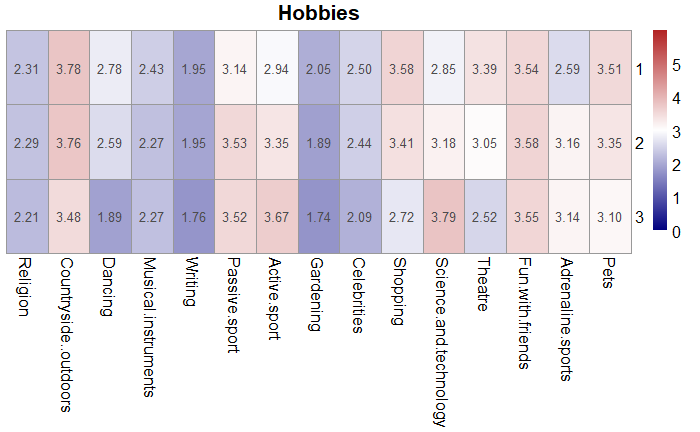
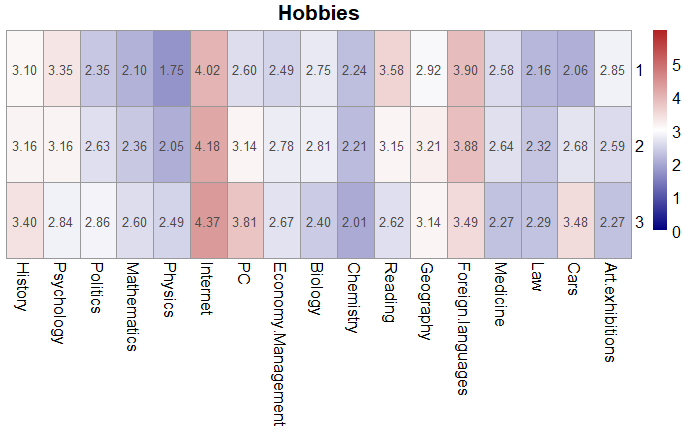


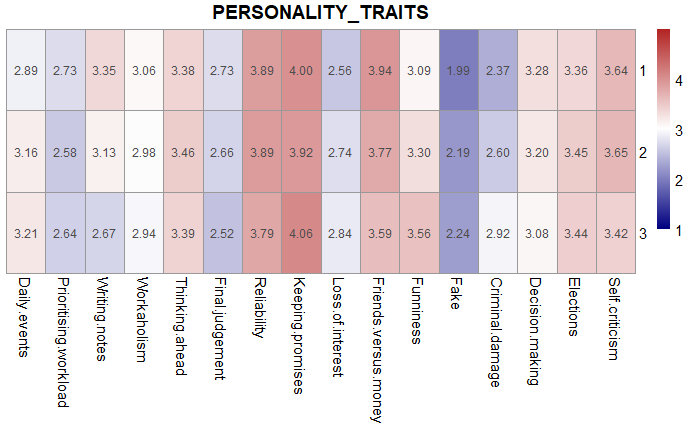
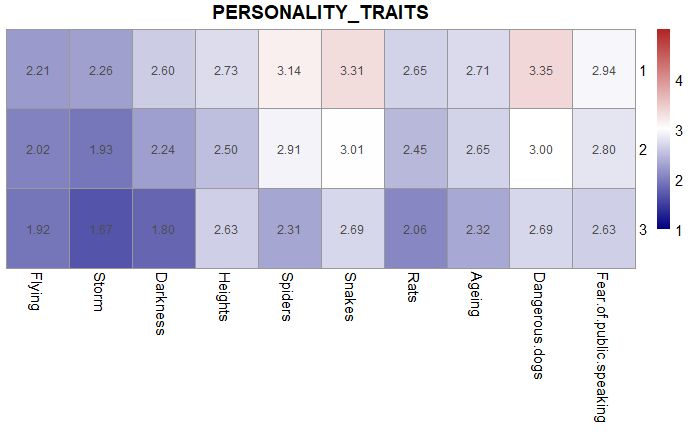
做出樹狀圖後，本篇決定分為三群，結果如下：

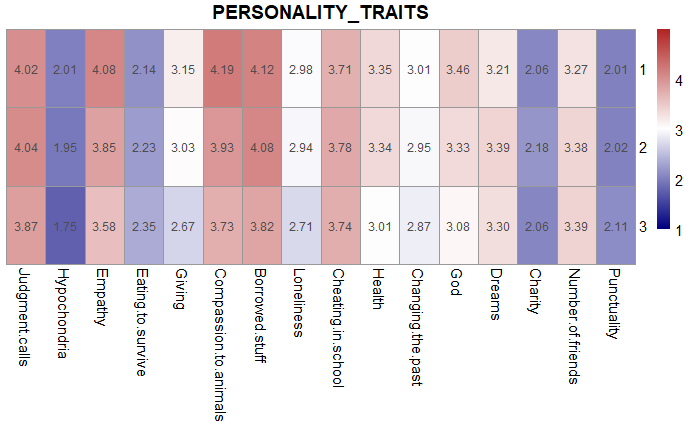
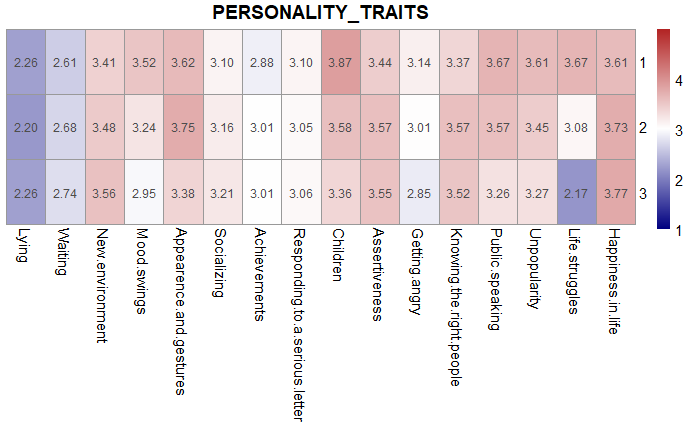


以這三群資料進行探索性資料分析，將每組問卷的填答資料計算取得該組平均，並將其數值以熱力圖表示，如下圖：





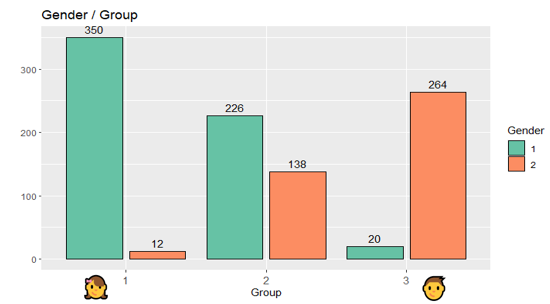


總結上述各圖呈現的結果，可得以下結論：

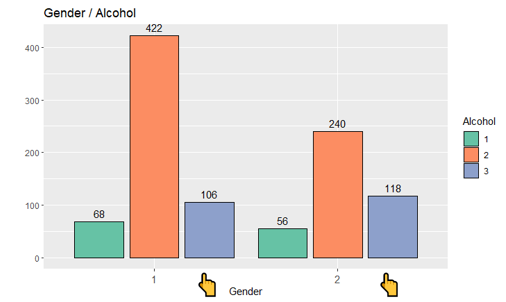
* 1. 第一組討厭物理，喜歡閱讀、外文，喜歡購物、討厭科學、科技，討厭極限運動並且比較感性。
  2. 第三組異常喜歡網路、電腦、車子，不喜歡藝術、討厭跳舞、喜歡運動、極限運動、科學，討厭戲劇、明顯不願意做筆記記事情，比較理性。

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。

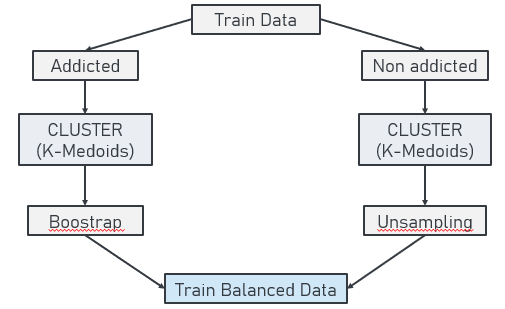


四、 以Xgboost預測酒精上癮的學生

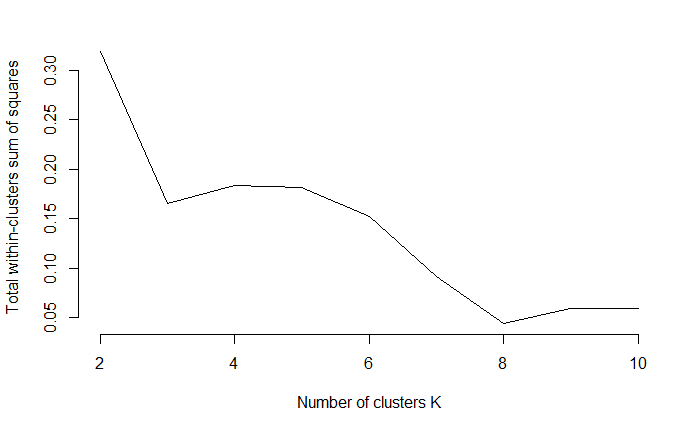
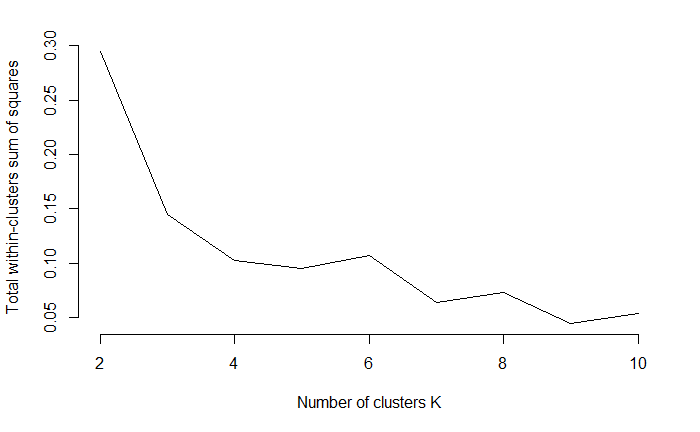
如下圖所示，填答為3代表學生勾選飲酒過量的選項，而可發現男性酒精上癮的比例明顯比女性要高。



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：

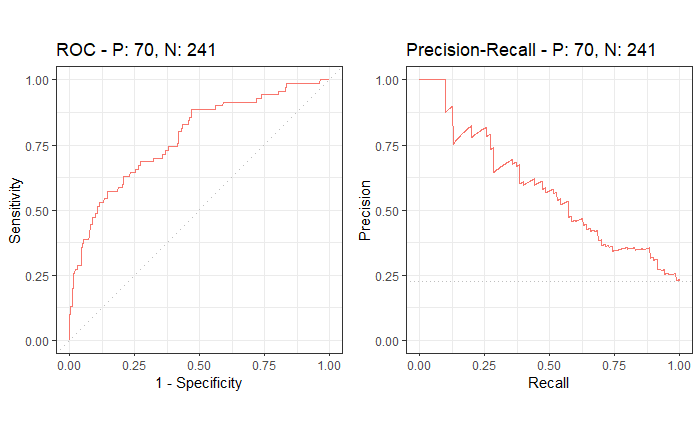
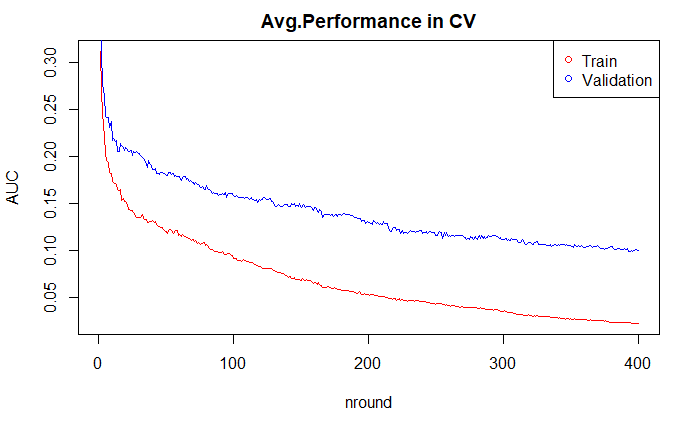


我們將訓練集資料以有酒精上癮、沒酒精上癮分為兩組，並對上癮組進行Boostrap重抽樣至兩組數量平衡，而兩組抽樣方法為先將資料以K-Medoids以總組內變異最小分群，再對各群抽樣，結果如下：

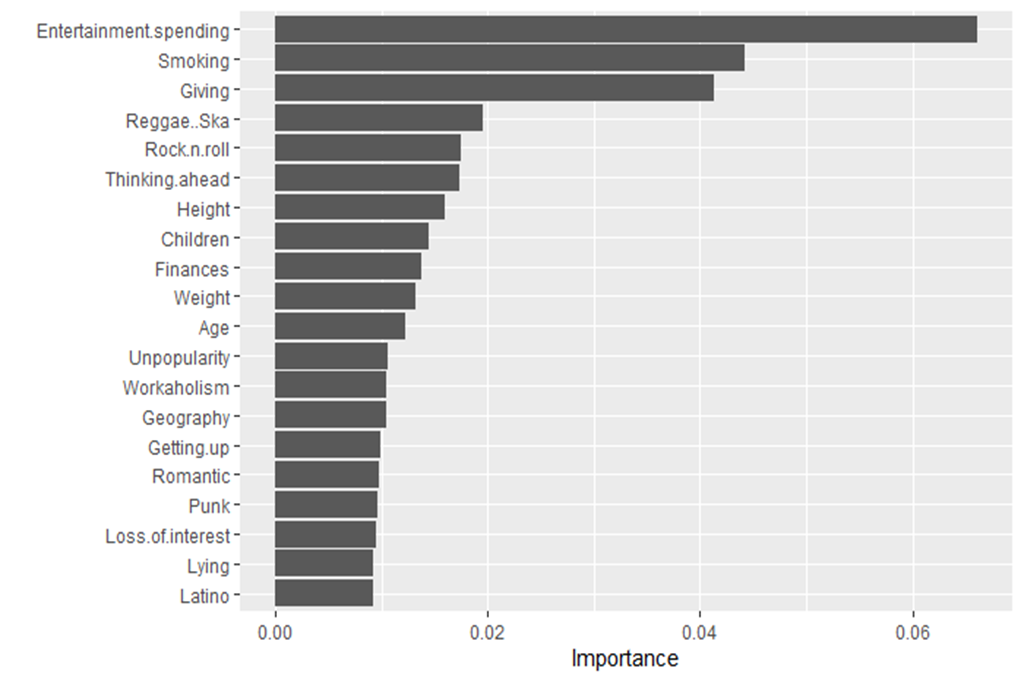
分析一至十群間分幾群有最低的總組內變異，可看到上癮組分八群，而沒上癮組分九群有最低的組內變異，將兩組數量平衡後，形成新的平衡後訓練集。

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。

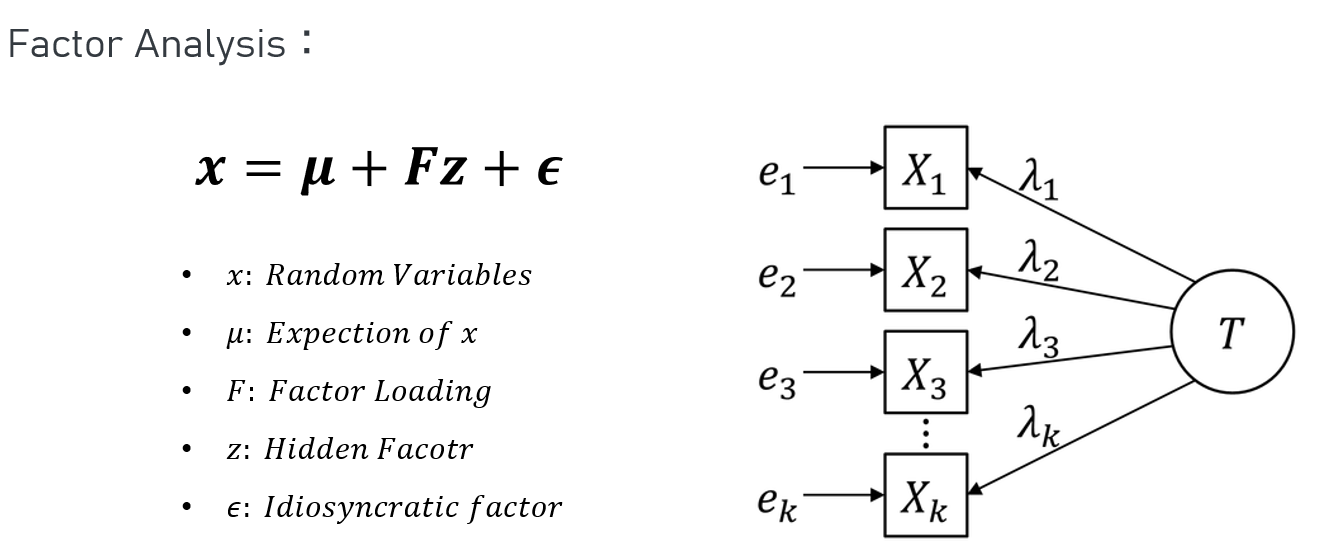


五、針對Xgboost篩選出的變數，以因素分析找出變數間的相關

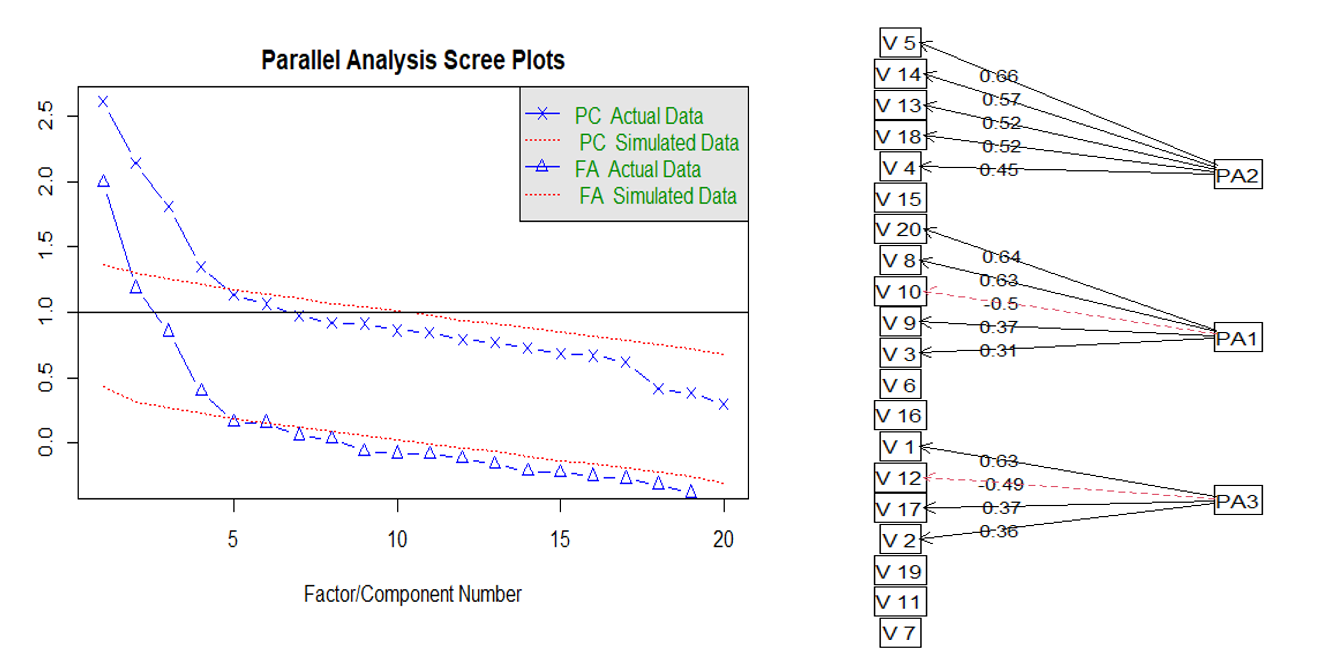
以Importance Value 篩選出前20有影響力的變數，如下圖：



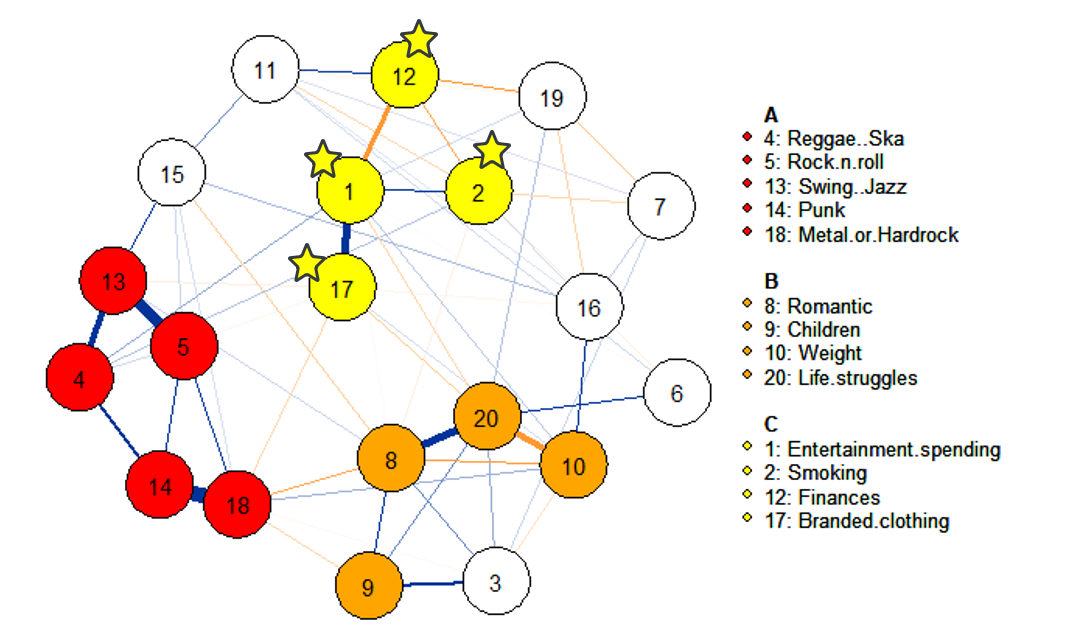
接著使用因素分析，找出篩選變數彼此間，是否存在關聯，下圖簡介因素分析：

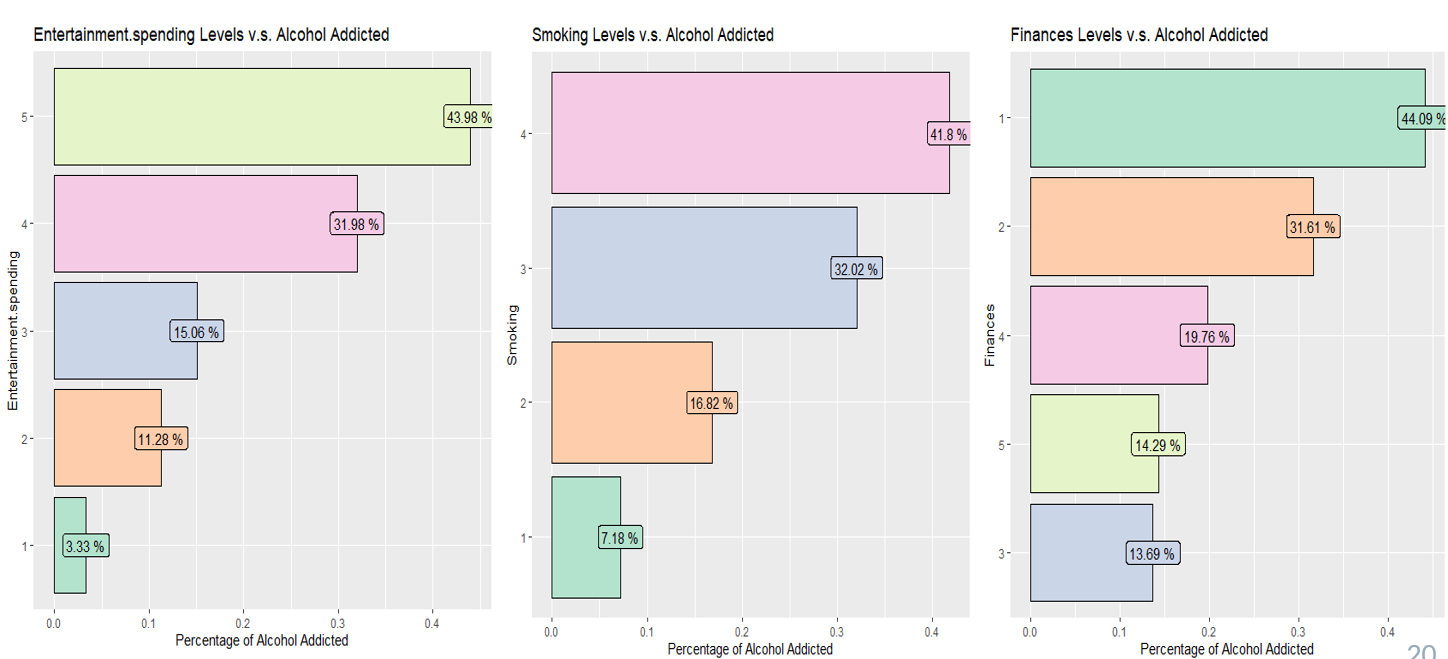


如下圖，我們以Scree Plots看到因素分析在第三因子時Eigen value便低於1，因此我們最多選到三因子，而右圖為三因子間，Eigen vector各變數間的關聯。



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：

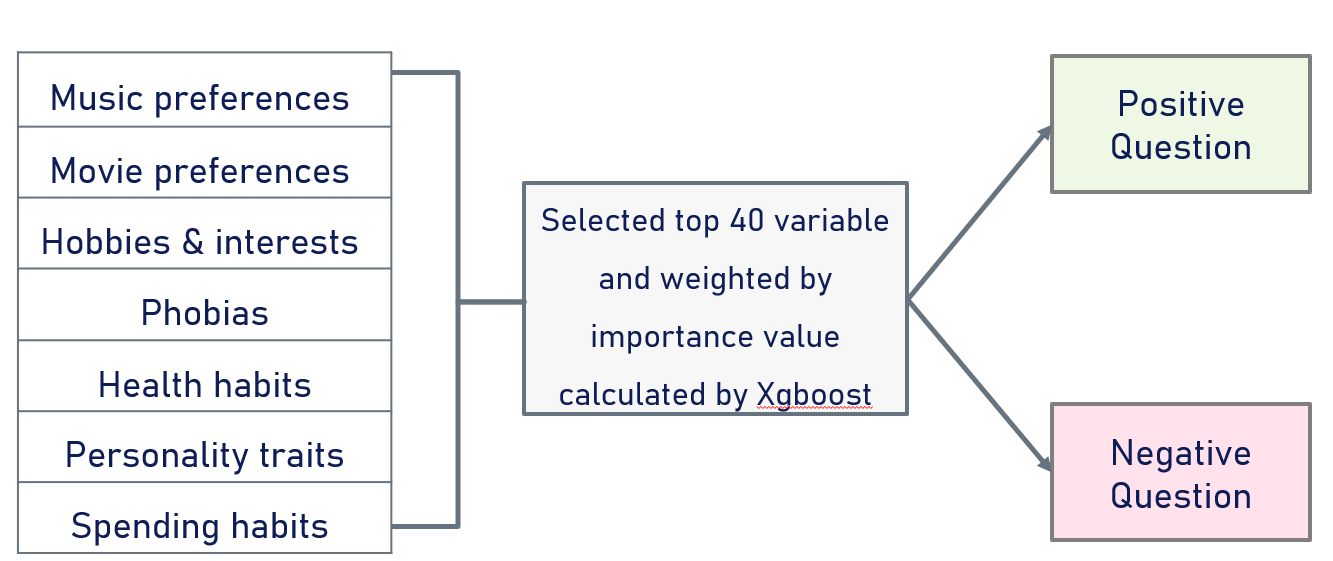


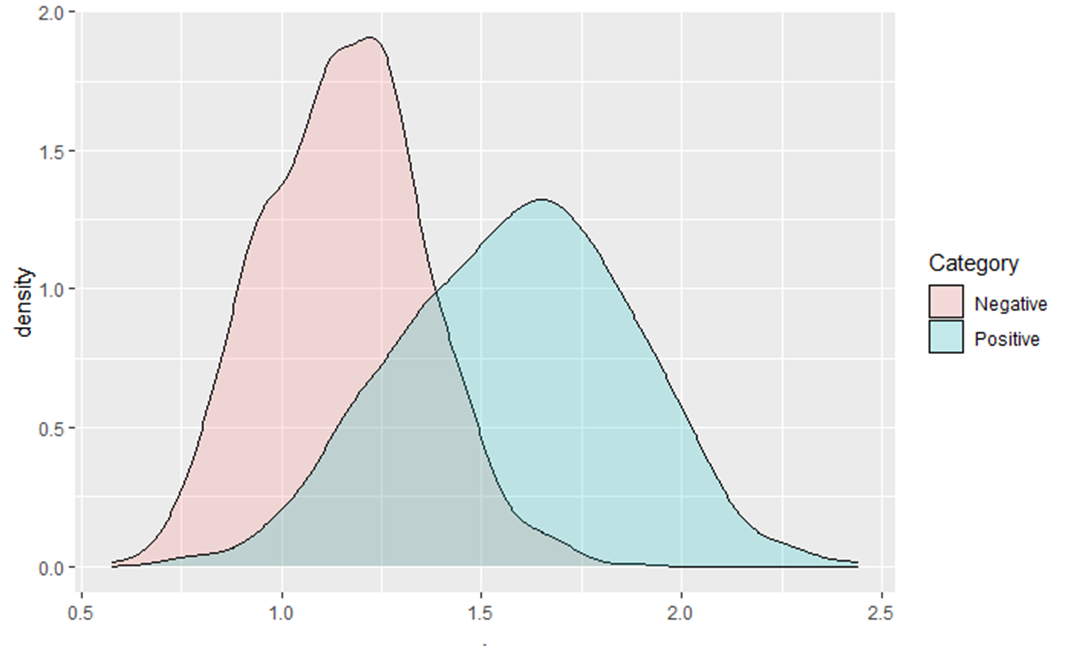


可發現娛樂消費、抽菸的頻率越高的族群裡，酒精上癮的比例也越高，而儲蓄觀念越差的族群酒精上癮的比例也越高。

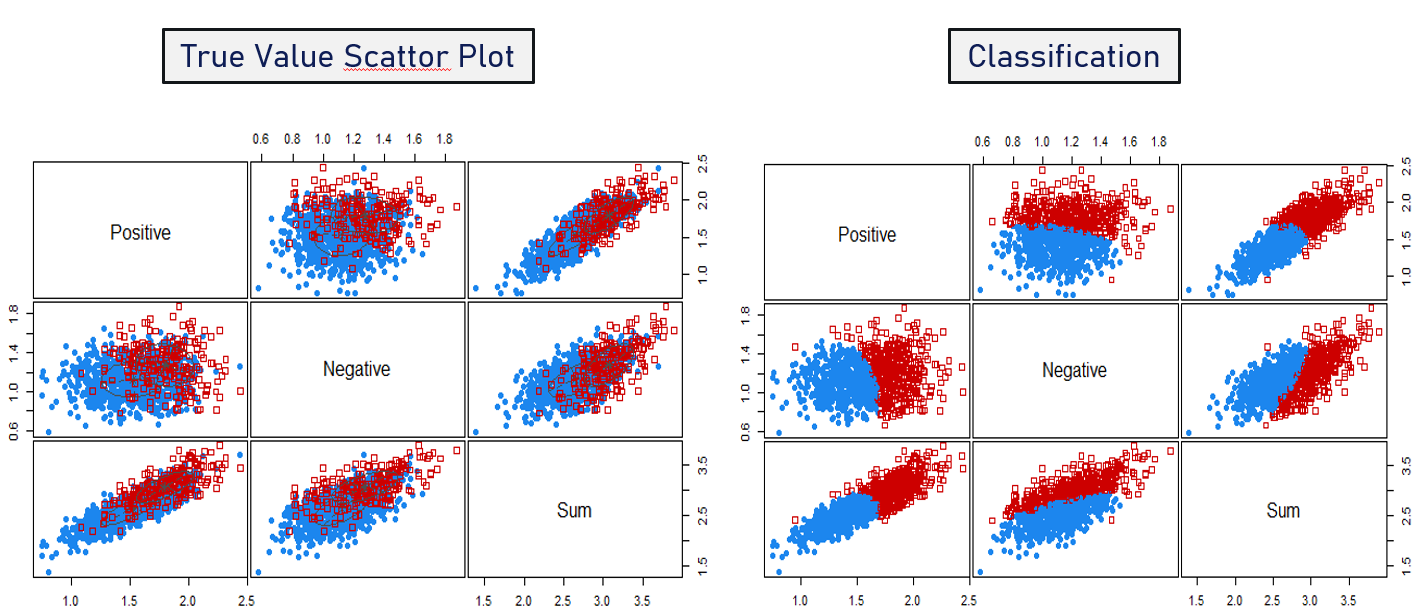
六、 將問卷填答資料加權，以Gaussian Mixture Model模型分類

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：

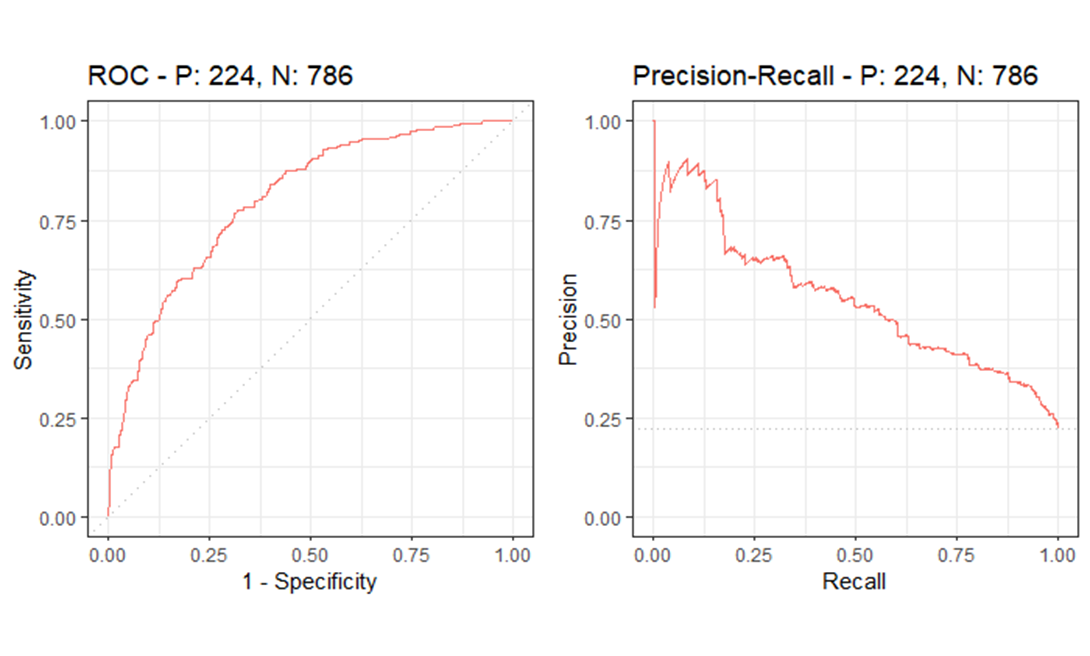




最後我們以正相關問題、負相關問題以及兩者總和進行GMM模型建立，可得結果如下：



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。