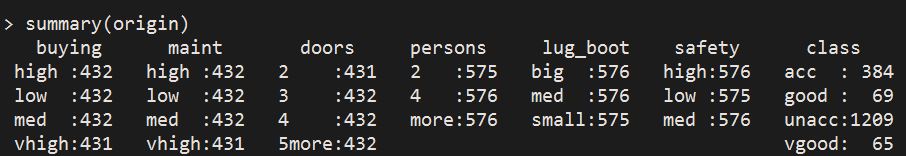
**購車偏好度：關聯分析與預測**

1. 資料型態



一共1727筆資料物件和7個屬性。

* 屬性資料
  + 費用

|  |  |
| --- | --- |
| Buying 價格 | vhigh、high、med、low |
| Maint 維修費用 | vhigh、high、med、low |

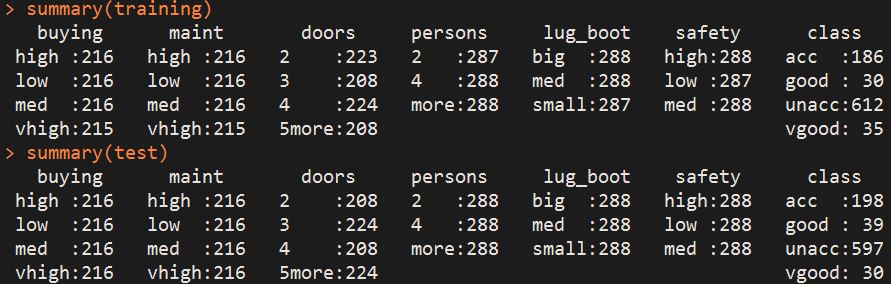
* 車子屬性

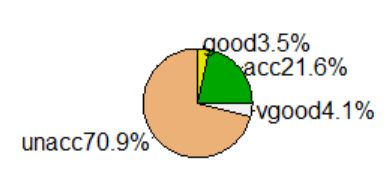
|  |  |
| --- | --- |
| doors車門數量 | 5以上、4、3、2 |
| persons承載人數 | 5以上、4、2 |
| lug\_boot行李箱尺寸 | big、med、small |
| safety安全性 | high、med、low |

* 分類

|  |  |
| --- | --- |
| class接受度 | vgood、good、acc、unacc |

由於資料龐大，足夠將資料分成training和test data，故將原data切成兩份存檔。並且兩份資料散布情況相同。



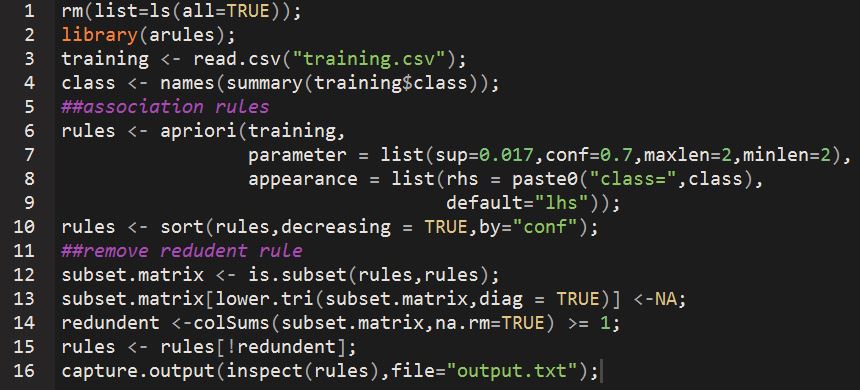


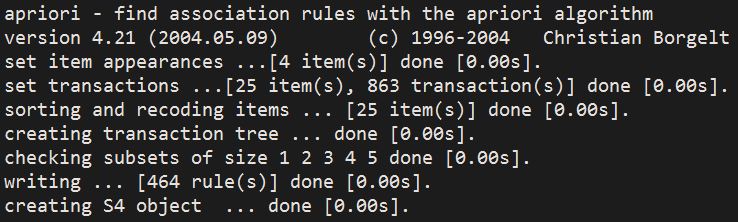
可以看出四個類別的比例不平均，unacceptable的比例遠遠超過其他三者。在探索關聯規則時關於unacceptable理論上會比較多。

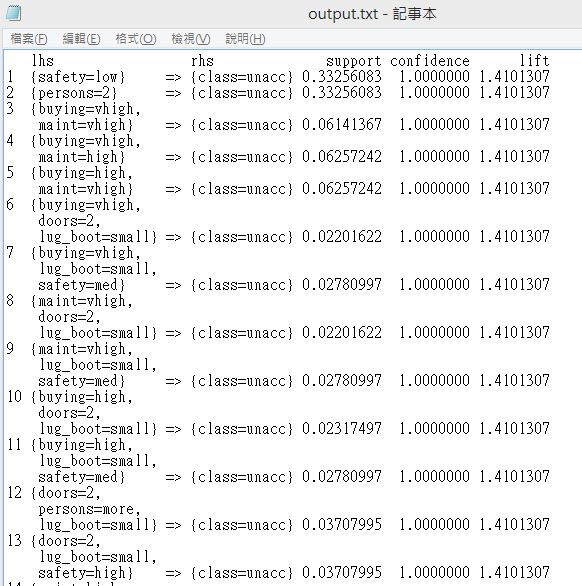
1. 關聯分析

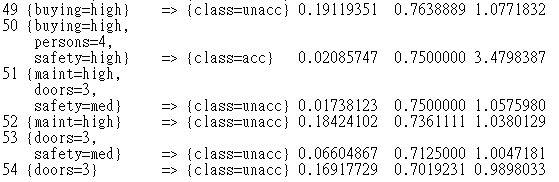
* 參數 support：0.017（約15人）、confidence：0.7

由於組合太多，先分析confidence較高的，而且與分類相關的規則，並在過程中就把冗於規則移除，最後依照confidence高低排序。









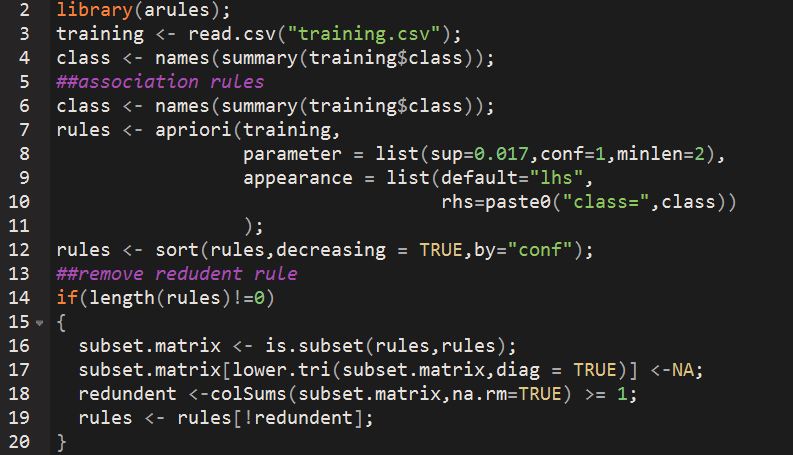
總共產生464條規則，刪去冗餘規則後剩餘54條規則，其中53條與unacceptable有關，只有1條是關於acceptable。

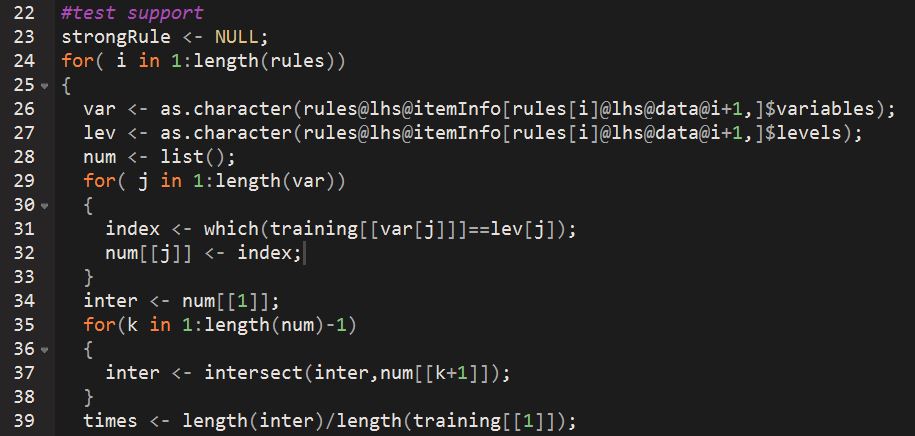
其中safety=low和person=2在資料中出現的比率和safety=low→class=unacc、person=2→class=unacc的support相等，可以明確表示兩者為強規則。

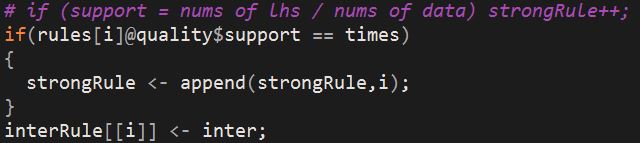
對good、vgood來說，support=0.017的門檻太高，完全無法產生相關規則；但對unacc來說，support=0.017太低，取出太多規則。

因此我們先嘗試會有哪些規則必然使class＝unacc，將這些規則分類出來的data從training data中移除。

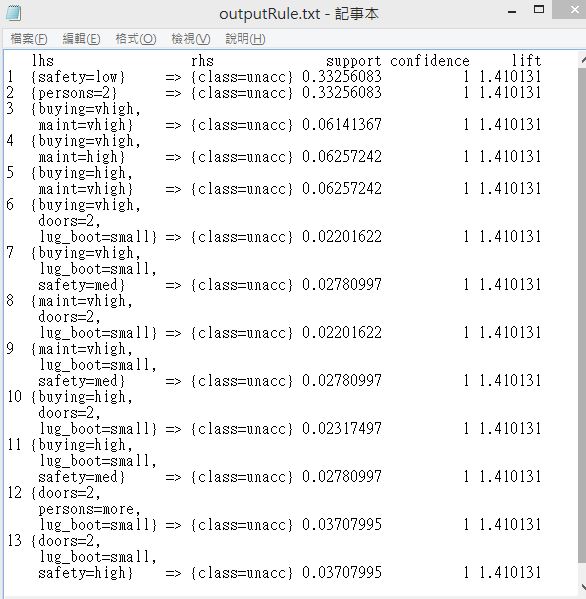
* 參數 support：0.017、confidence：1



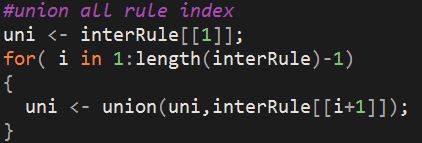


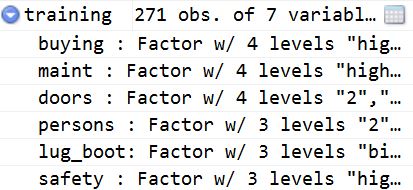


產生13條規則

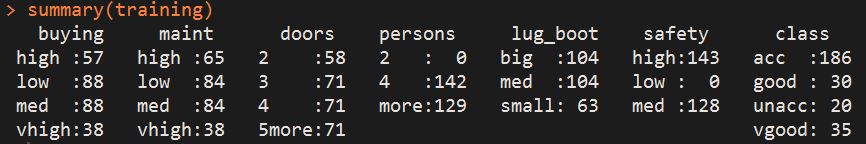


移除了593條資料，剩餘271條資料

 D:\alice\Desktop\擷取.JPG



但仍有class=unacc

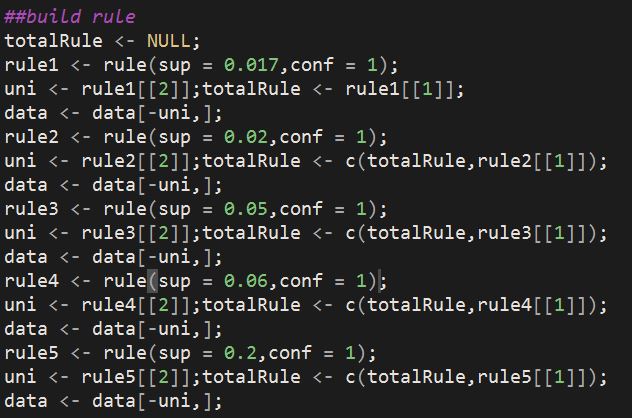


* 演算法：

support = （lhs && rhs）/資料總數

若support ＝lhs / 資料總數，則表示 lhs必然使得class=rhs。

1. 利用appriori找出關聯規則
2. 移除重複規則
3. 檢查是否lhs的出現比率是否相等於sopport ，若是，則保留規則
4. 移除與規則相符合的資料項目，接著重複步驟1



最後剩餘一筆資料無法分類。



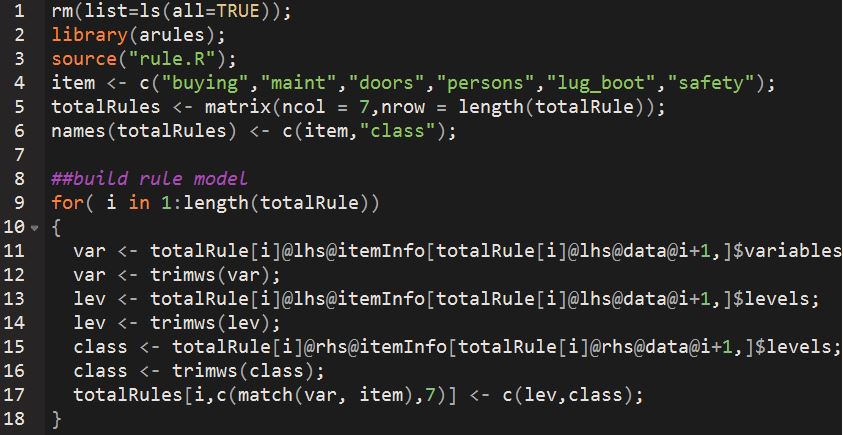
由於每次參數的設定都需要變化，因此無法使用迴圈自動進行。我覺得這部份非常可惜，希望可以自動偵測剩餘資料，找出最佳參數。

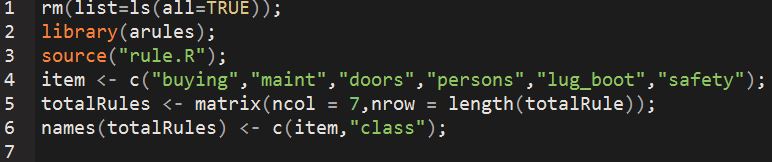
1. 分類

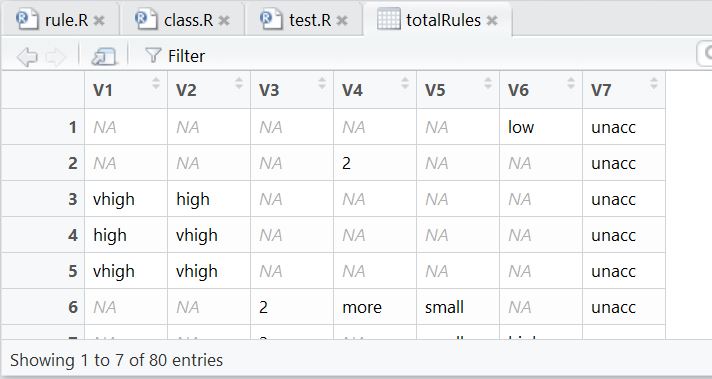
* 建立分類規則

接下來將上述rule1到rule5總共80筆規則依照順序，整理成容易讀取的資料格式，最後輸出到檔案裡。

檔案格式我選擇與data相同的格式，將規則當成一筆資料，規則中有提到的項目填入，未提到的資料則保持缺失。分類時讀入一筆資料物件，比對兩者之間是否相同（無視缺失值），若完全相同則符合規則。





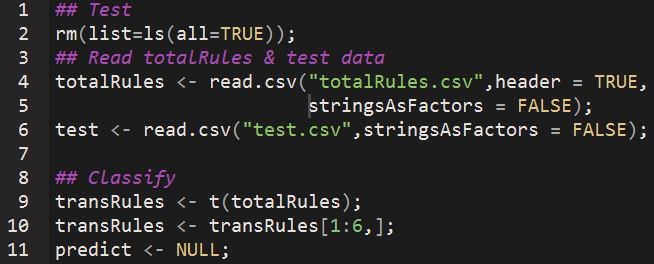


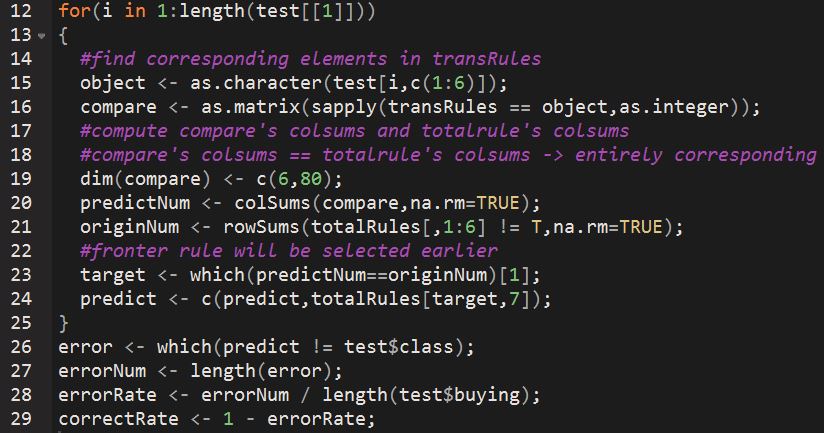
* Test檢測

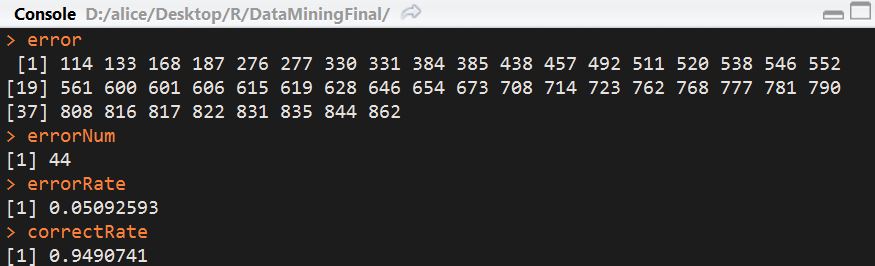
R語言中，當兩個向量的長度不相同， R會採用循環規則：較短的向量已經處理完所有的元素，而較長的向量還有未處理的元素時，R返回較短向量的開始位置，“循環”應用它的元素，繼續來與較長向量的元素進行相應操作，直到這個操作完成為止。

在這裡，資料物件為較短的向量，而rule則為較長的向量，可以讓資料物件循環比對所有rule而不需要跑迴圈。（增加效能）

但是向量比對的是向量元素，因此是否符合規則還需要看同一列中是否除了NA之外的所有元素皆為TRUE。若是的話，即符合規則。







將規則與test比對的結果，正確率約達0.95。