國立台北大學

統計學系

在職專班

110學年度多變量分析

期末報告

社群平台客群 - 多媒體推薦方法研究

|  |  |
| --- | --- |
| 課程教授： | 鍾麗英 |
| 學生姓名： | 李玉柔711078911 |

中華民國一一一年六月十三日星期二

目錄

[一、 研究動機 6](#_Toc106396309)

[二、 資料介紹及研究內容 7](#_Toc106396310)

[(一)、 資料介紹 7](#_Toc106396311)

[(二)、 研究內容 7](#_Toc106396312)

[I. 研究範圍 7](#_Toc106396313)

[II. 研究流程圖 8](#_Toc106396314)

[三、 敘述統計 9](#_Toc106396315)

[(一)、 基本統計 9](#_Toc106396316)

[四、 相關係數 11](#_Toc106396317)

[五、 多變量分析 13](#_Toc106396318)

[(一)、 變異數分析 13](#_Toc106396319)

[I. 性別中音樂及影視喜好類程度有顯著差異 13](#_Toc106396320)

[II. 年齡層中音樂及影視喜好類程度有顯著差異 15](#_Toc106396321)

[III. 教育程度中音樂及影視喜好類程度有顯著差異 16](#_Toc106396322)

[(二)、 因素分析 18](#_Toc106396323)

[I. 音樂類型變數因素分析 18](#_Toc106396324)

[II. 影視類型變數因素分析 21](#_Toc106396325)

[III. 影音綜合變數因素分析 21](#_Toc106396326)

[(三)、 主成分分析 22](#_Toc106396327)

[I. 女性興趣愛好類型變數的主成分分析 23](#_Toc106396328)

[II. 男性興趣愛好類型變數的主成分分析 25](#_Toc106396329)

[(四)、 集群分析與多變量變異數分析 27](#_Toc106396330)

[I. 集群數判別 27](#_Toc106396331)

[II. 各集群顯著變數判別-多變量分析 28](#_Toc106396332)

[(五)、 集群後相關係數分析 29](#_Toc106396333)

[I. 女性興趣愛好語音樂喜好類型 29](#_Toc106396334)

[II. 男性興趣愛好語音樂喜好類型 31](#_Toc106396335)

[六、 結論 36](#_Toc106396336)

[附件、一 37](#_Toc106396337)

[附件、二 37](#_Toc106396338)

**表目錄**

[表、 一 人口輪廓數量比例表 9](#_Toc106392799)

[表、 二 性別與影音類別喜好程度交叉表 9](#_Toc106392800)

[表、 三 年齡層與影音類別喜好程度交叉表 10](#_Toc106392801)

[表、 四 教育程度與影音類別喜好程度交叉表 10](#_Toc106392802)

[表、 五 各性別對音樂及電影敘述統計表 13](#_Toc106392803)

[表、 六 性別共變數矩陣檢定表 14](#_Toc106392804)

[表、 七 性別多變量各統計量檢定表 14](#_Toc106392805)

[表、 八 性別與影音喜好程度參數估計表 14](#_Toc106392806)

[表、 九 年齡層對音樂及電影敘述統計表 15](#_Toc106392807)

[表、 十 年齡層多變量各統計量檢定表 15](#_Toc106392808)

[表、 十一 年齡層對音樂及電影敘述統計表 16](#_Toc106392809)

[表、 十二 年齡層對音樂及電影敘述統計表 16](#_Toc106392810)

[表、 十三 教育程度與影音喜好程度參數估計表 17](#_Toc106392811)

[表、 十四 音樂喜好程度問項表 18](#_Toc106392812)

[表、 十五 音樂類型變數因素分析表 19](#_Toc106392813)

[表、 十六 音樂類型變數特徵值相關矩陣表 19](#_Toc106392814)

[表、 十七 音樂類型變數因素分析表 20](#_Toc106392815)

[表、 十八 影視類型喜好程度表 21](#_Toc106392816)

[表、 十九 各興趣喜好程度問項表 22](#_Toc106392817)

[表、 二十 女性興趣愛好類型變數 23](#_Toc106392818)

[表、 二十一 女性興趣愛好類型變數 23](#_Toc106392819)

[表、 二十二 女性興趣愛好類型變數與主成分相關係數表 24](#_Toc106392820)

[表、 二十三 男性興趣愛好類型變數 25](#_Toc106392821)

[表、 二十四 男性興趣愛好類型變數 25](#_Toc106392822)

[表、 二十五 男性興趣愛好類型變數與主成分相關係數表 26](#_Toc106392823)

[表、 二十六 男性各群參數估計值與 28](#_Toc106392824)

[表、 二十七 女性各群參數估計值表 28](#_Toc106392825)

**圖目錄**

[圖、 一 興趣愛好和音樂變數間相關係數熱圖 11](#_Toc106392826)

[圖、 二 興趣愛好和影視變數間相關係數熱圖 12](#_Toc106392827)

[圖、 三 音樂類型變數特徵值陡坡圖 19](#_Toc106392828)

[圖、 四 女性興趣愛好類型變數陡坡圖 24](#_Toc106392829)

[圖、 五 男性興趣愛好類型變數陡坡圖 26](#_Toc106392830)

[圖、 六 女性集群分析陡波圖 27](#_Toc106392831)

[圖、 七 女性集群分析分群圖 27](#_Toc106392832)

[圖、 八 女性集群分析陡波圖 27](#_Toc106392833)

[圖、 九 男性集群分析分群圖 27](#_Toc106392834)

[圖、 十 女性第一群興趣與音樂類型變數相關係數熱圖 29](#_Toc106392835)

[圖、 十一 女性第二群興趣與音樂類型變數相關係數熱圖 30](#_Toc106392836)

[圖、 十二 女性第三群興趣與音樂類型變數相關係數熱圖 30](#_Toc106392837)

[圖、 十三 男性第一群與影視類型變數相關係數熱圖 31](#_Toc106392838)

[圖、 十四 男性第一群與音樂類型變數相關係數熱圖 31](#_Toc106392839)

[圖、 十五 男性第二群與影視類型變數相關係數熱圖 32](#_Toc106392840)

[圖、 十六 男性第二群與音樂類型變數相關係數熱圖 32](#_Toc106392841)

[圖、 十七 男性第三群與影視類型變數相關係數熱圖 33](#_Toc106392842)

[圖、 十八 男性第三群與音樂類型變數相關係數熱圖 33](#_Toc106392843)

[圖、 十九 男性第四群與影視類型變數相關係數熱圖 34](#_Toc106392844)

[圖、 二十 男性第四群與音樂類型變數相關係數熱圖 34](#_Toc106392845)

# 研究動機

在去年底知名串流流音樂平台，開出強大功能結合影視、社群媒體平台一起合作，藉由各平台客群資料互相推薦客戶可能喜歡產品，除了增加自身客戶黏著度，能夠保留自身原有客戶也能透過合作平台來吸引潛在客群，產生音樂、影視事及社群媒體三方互惠的強大商業模式。

但在今年2022年台灣4月開始爆發Omicron快速染的Covid-19變形種確診案例越來越多時，知名網路影音平台卻因發文失誤而引發退訂潮，但這不經讓人困惑，既然在前年底進行了這麼強大的商業模式，但為何僅憑一篇貼文，而造成這麼重大的損失。是因為商品不在吸引人又或者是目標推薦商品貼標不再精準，又或者是隨者訂閱客群年齡層上升，生活習慣愛好有所改變?這些我們無從得知。

但推薦系統總是在目前網路發達的時代時常聽到，甚至廣泛的被運用在網路行銷，無論是有形又或者是無形各種商品上進行網路廣告推播，目前以及未來多媒體影音仍然會是非常龐大的商機，而這次想透過在本份期末報告，應用在課堂中學習到的分析方式，來嘗試做出影音推薦方法。

# 資料介紹及研究內容

## 資料介紹

資料由Kaggle網站擷取而來，為英國FSEV University 為英國專門研究社會科學和經濟科學的學院，2013年統計課程中學生製作的調查問卷資料。問卷調查目標族群為年輕族群，調查該族群個人口輪廓、生活方式以及金錢觀…等資訊。

受調查年齡層介於15至30歲之間，問卷題型為各音樂類別喜好程度、各影視類別喜好程度、興趣及愛好、各事物恐懼程度、健康習慣、人生觀點、資金運用習慣及人口輪廓等8大類型問題。資料內容包含1,001筆資料，150個問項。

## 研究內容

本次研究為嘗試使用在課堂上學習的分析做出推薦多媒體影音方法，主要推薦產品為影音多媒體，主要會要專注在各音樂類別喜好程度、各影視類別喜好程度、興趣及愛好三大類喜好問項以及人口輪廓，共四大項。

在實務作業上，通常會以用戶瀏覽資料對客戶進行貼標分群動作，進而推薦同群人普遍喜愛的影音內容。又或者是以點及過的影音類別再推薦相同的類別的影音內容。

深為用戶也發現一種現象，若是透過目前推薦方式，用戶僅能吸收到同一種類型的資訊，對目前資訊爆炸時代下養成的年輕人可能不夠滿足，有可能使用戶一轉移至其他影音平台吸收更多元的資訊，容易使流量損失、廣告收益及品牌使用意願…等商業損失。

思考到要況展推薦影音類別，這邊這次研究的方法也會將有潛在機會的影音類別納入推薦表的行列。

### 研究範圍

本次研究將聚焦在音樂類別喜好程度共19題、各影視類別喜好程度共12題以及各興趣及愛好喜好程度共32題以上題型皆由1分不喜愛至5分非常喜愛評分，另外還有人口資料包含性別、年齡、教育程度…等，共73欄位資訊做接下來方法。

資料中若有任一欄位為遺漏值，將直接移除該筆資料，最終資料樣本數從1,001筆縮減至774筆。

### 研究流程圖

**資料收集**

**資料篩選**

**基本統計**

**變數類型決定**

**客群分析**

**變數分析**

**推薦方法**

# 敘述統計

## 基本統計

首先介紹這次資料中填寫人的年齡、性別、教育程度等分布情形。問卷受訪者年齡分佈表中看到受訪者為15~30歲，在後續將分成3個年齡層如下人口輪廓數量比例表(表、一)。分別為15至19歲(43%)、20至24歲(47%)以及25至30歲(9%)。性別中女性資料共460筆(59%)，男性資料共314筆(41%)。教育程度中就讀小學資料共4筆(0.5%)、小學資料共62筆(8%)、中學資料共482筆(62%)、學士資料共161筆(21%)、碩士資料共62筆(8%)、博士資料共3筆(0.4%)

表、 一 人口輪廓數量比例表

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **性別** | **人數** | **%** | | **女性** | 460 | 59% | | **男性** | 314 | 41% | | **總和** | 774 | 100% | | |  |  |  | | --- | --- | --- | | **年齡層** | **人數** | **%** | | **01.15~19歲** | 335 | 43% | | **02.20~24歲** | 366 | 47% | | **03.25~30歲** | 73 | 9% | | **總和** | 774 | 100% | | |  |  |  | | --- | --- | --- | | **教育程度** | **人數** | **%** | | **01.小學就讀中** | 4 | 0.5% | | **02.小學** | 62 | 8% | | **03.中學** | 482 | 62% | | **04.學士** | 161 | 21% | | **05.碩士** | 62 | 8% | | **06.博士** | 3 | 0.4% | | **總和** | 774 | 100% | |

由於想透過人口輪廓有一步的判斷，觀察人口輪廓在各影音類型的平均數標準差做個簡單的事先了解。

從性別來看在音樂及影視平均分數前五名如下性別與影音類別喜好程度交叉表(表、二)。

觀察影視中在男女性別上排名差異，童話故事、動畫和浪漫愛情片女性較男性喜愛，動作、科幻及戰爭男性較女性喜愛，喜劇和紀錄片都在男女喜愛排行中的前五名。

觀察音樂中在男女性性別排名差異，拉丁音樂女性較男性喜愛，嘻哈說唱男性較女性喜愛，搖滾、流行、快節奏、搖滾都在男女喜愛排行中的前五名。

表、 二 性別與影音類別喜好程度交叉表

|  |  |
| --- | --- |
|  |  |

從年齡層來看在音樂及影視平均分數前五名如下年齡層與影音類別喜好程度交叉表(表、三)。觀察影視中在年齡層上排名差異，喜劇、童話及紀錄片在各年齡層喜愛排行中的前五名。動作和驚悚片不在15~19歲的前五名中。浪漫和驚悚不在20~24歲的前五名中。動畫和動作不在25~30歲的前五名中。觀察音樂中在年齡層上排名差異，拉丁音樂喜愛排行25~30歲較15~24歲高。舞曲喜愛排行15~24歲較25~30歲高。搖滾、流行、快節奏、搖滾都在各年齡層喜愛排行中的前五名。

表、 三 年齡層與影音類別喜好程度交叉表



從教育程度來看在音樂及影視平均分數前五名如下教育程度與影音類別喜好程度交叉表(表、四)。觀察影視中在年齡層上排名差異，喜劇、動畫及童話及紀錄片在各教育程度喜愛排行中的前五名，博士除外。在博士學歷上除喜劇外，排名前五的影視類型和其他教育水準的有明顯差別。觀察音樂中在年齡層上排名差異，搖滾、流行、快節奏、搖滾除博士及小學就讀中外在各教育水準喜愛排行中的前五名。古典類型在碩士及博士的排行中。

表、 四 教育程度與影音類別喜好程度交叉表



以上排行僅由喜好分數平均值來做個前導了解，無法直接說明有無顯著差異，實際上在各人口輪廓對影音媒體類型有無顯著差異仍須用MANOVA做進一步分析，才能有科學依據的說法。對接下來的推薦方法方能收斂到最有效率的方式。

# 相關係數

接下來利用相關係數來看音樂喜好、影視喜好和興趣愛好的變數間的相關性。

觀察音樂和興趣愛好變數間的相關係數熱圖如下，縱軸為音樂種類變數，橫軸為興趣愛好變數，從興趣愛好可看到：

喜歡歷史和音樂中的古典樂、歌劇以及爵士樂高度正相關。

喜歡閱讀和音樂的古典樂、舞台樂、爵士樂、歌劇樂以及地下樂團高度正相關。

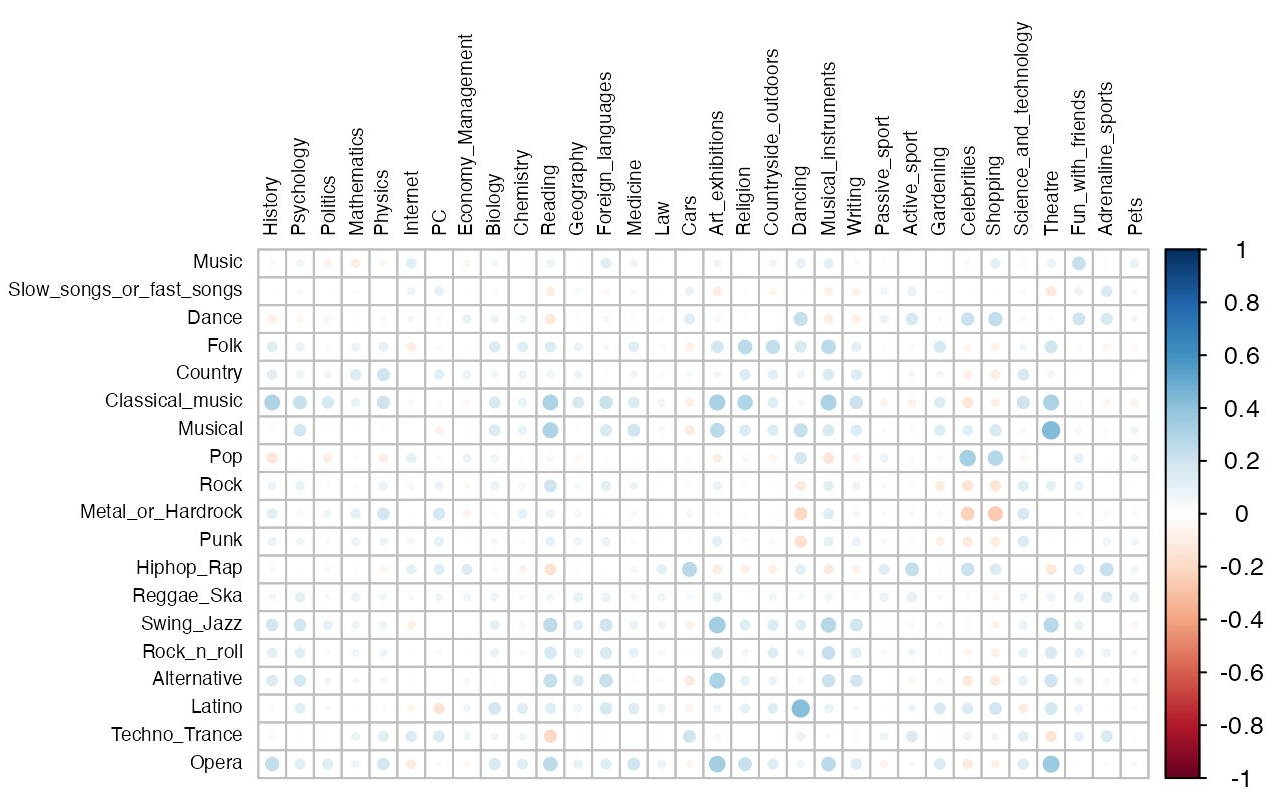
喜歡藝術展覽和喜歡閱讀對音樂的高度正相關變數相同。

喜歡跳舞和音樂的拉丁樂和舞曲高度正相關。

喜歡樂器演奏和喜歡閱讀對音樂的高度正相關變數相同。

喜歡藝人音樂的舞曲、流行樂以及嘻哈高度正相關，和重金屬搖滾負相關。

喜歡劇院的和音樂中古典樂、交響樂、爵士樂以及歌劇樂正相關。



圖、 一 興趣愛好和音樂變數間相關係數熱圖

目前看到各興趣愛好和對應影音類型相關性，若直覺推薦用用平時就喜愛的類型，就會產生前述情況發生，

影視和興趣愛好變數間的想關係數圖如下，縱軸為影視種類變數，橫軸為興趣愛好變數。從興趣愛好可看到：

歷史和影視中的戰爭以及紀錄片有高度正相關。

物理和影視中的科幻、西部牛仔正相關和浪漫片呈現負相關。

電腦和影視中的科幻、西部牛仔、動作、恐怖以及驚悚正相關和浪漫片呈現負相關。

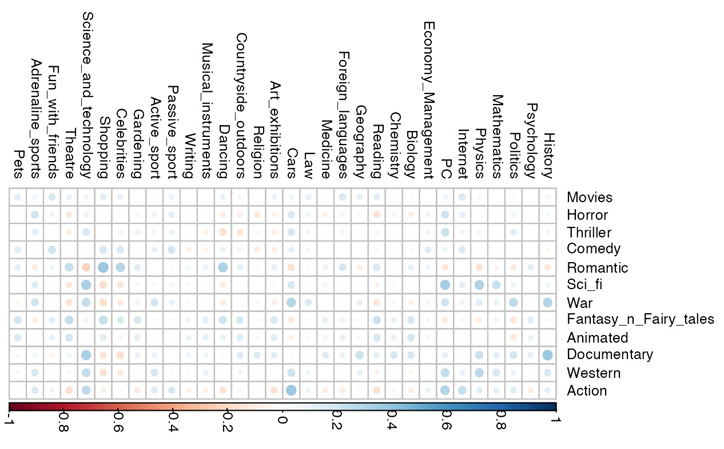
汽車和影視中的戰爭、動作、西部牛仔、恐怖、及驚悚正相關和浪漫片呈現負相關。

跳舞和影視中的和影視中以及童話夢幻正相關變數相關和驚悚、科幻

以及動作呈現負相關。

購物和影視中的和影視中、喜劇以及童話夢幻正相關，和科幻負、戰爭、紀錄片以及西部牛仔負相關。

科學技術和影視中的科幻、戰爭、紀錄片、西部牛仔以及動做正相關，和浪漫負相關。



圖、 二 興趣愛好和影視變數間相關係數熱圖

目前已了解各興趣愛好對應的影音類別，其實在這邊就可以直接對用戶依照興趣愛好來做影音類型推薦。直接依照此方法推薦會產生前章述說的情形，不斷推薦類似類別的影音，但僅侷限幾個類別，有機會使流量損失、廣告收益及品牌使用意願…等商業利益損失這是多媒體平台都不樂見的。

故接下來想透過音樂、影視、興趣和人口輪廓變數各別進行分析，找出最合適的推薦方法。

# 多變量分析

本章節將嘗試使用更嚴謹的分析方法找出最適合的推薦方式。

將優先對人口輪廓進行對多媒體喜好程度的變異數分析(MANOVA)，找出對用戶貼標的第一指標。音樂、影視類型變數進行變數分群，找出變數間是否能夠有群組關係，進而將潛在影音類型推薦給客戶。

## 變異數分析

推薦特定音樂類型和影視類型前，須先了解人口輪廓對這兩類別下的變數是否有顯著差異性。

利用性別、年齡層和教育程度先個別觀察在音樂及影片中是否有明顯的偏好差異，以上三個人口類型輪廓，進而做影音多媒體平台推薦排序，接下來以變異數分析來了解。

### 性別中音樂及影視喜好類程度有顯著差異

資料中女性為460筆、男性為314筆，依照下列各性別對音樂及電影敘述統計表(表、五)，音樂平均值上女性與男性的平均值較男性略多且標準差小，但這無法直接判斷是否在音樂上有顯著差異。

而在影視喜好上平均女性與男性的平均值差異略小，標準差雖比男性小但也無法直接判斷是否有顯著差異。

表、 五 各性別對音樂及電影敘述統計表



為判斷性別對音樂及影視喜好程度差異的顯著性，由共變數矩陣檢定表(表、六)共變數矩陣檢定P-Value小於0.001以及性別多變量各統計量檢定表(表、七)結果可看出P-Value 在各統計量上皆小於0.05，性別上對音樂和影視有顯著喜好差異。

表、 六 性別共變數矩陣檢定表



表、 七 性別多變量各統計量檢定表



而男性和女性要優先推薦音樂又或者是影視，可從性別與影音喜好程度參數估計表(表、八)看出女性和音樂P-Value<0.05有顯著性，在音樂上男女有顯著的差異，從參數估計值中可看出女性在音樂上的係數較高，綜合上述，在客戶是女性的族群中會以音樂為優先推薦，而男性則以後續分群後對音樂和影視的相關性程度來推薦。

表、 八 性別與影音喜好程度參數估計表



### 年齡層中音樂及影視喜好類程度有顯著差異

資料中年齡集劇為15至19歲。年齡層分為3組，組一為15~19歲335人(43%)、組二為20~24歲366人(47%)以及組三24~30歲73人(9%)。想了解各年齡層對音樂與影視的喜好有無差異。

依照下列年齡層對音樂及電影敘述統計表(表、九)在音樂和影視喜好在平均值及標準差看不出太大差異，但這無法直接判斷無顯著差異，接下來做多變量變異數(MANOVA)分析，來看出年齡層是否有顯著差異。

表、 九 年齡層對音樂及電影敘述統計表



可從下年齡層多變量各統計量檢定表(表、十)看出各統計量皆P-Value皆大於0.05，在各年齡層對影音多媒體的喜好無顯著差異。

表、 十 年齡層多變量各統計量檢定表



依照上述年齡層無法作為推薦優先推薦影視或音樂的參考變數。

### 教育程度中音樂及影視喜好類程度有顯著差異

問卷填寫者中教育程度從小學、中學、大學、碩士、博士都有，但樣本數在各教育程度上的分布可從教育程度對音樂及電影敘述統計表看出並不均勻，甚至有些教育程度樣本僅有3筆數量不足，有可能會影響到最後是否要拿來做推薦的變數。

從年齡層對音樂及電影敘述統計表(表、十一)可看出各個教育程度間除了碩士、博士和其他教育程度平均數和標準差有明顯差異外，其他教育程度的平均數和標準差並無太大差異，接下來由多變量變異數(MANOVA)分析，來看出年齡層是否有顯著差異。

表、 十一 年齡層對音樂及電影敘述統計表



可從下性別多變量各統計量檢定表(表、十二)看出各統計量P-Value皆小於0.05，在各教育程度間對影音多媒體的喜好有顯著差異。

表、 十二 年齡層對音樂及電影敘述統計表



前述說明到教育程度間的分佈並不均勻，且由教育程度與影音喜好程度參數估計表(表、十三)可發現教育程度推薦音樂或者影視的變數中，僅有博士教育程度在音樂上顯著，但博士教育程度樣本僅有3筆，且在網路平台上資料蒐集來的資訊中很難判斷該名用戶教育程度，故最終決定不使用教育程度作為接下來的推薦主要參考變數。

表、 十三 教育程度與影音喜好程度參數估計表



整合上列3項人口輪廓，性別和教育程度有顯著的對音樂及影視的偏好，年齡層對音樂和影視沒有顯著的偏好。而性別中女性對音樂的喜好程度較男性顯著的高，教育程度則是因為目前蒐集的為卷資料分佈不夠均勻，以及在實務很難對用戶的教育程度資訊做蒐集，故僅使用性別做為第一步的推薦分類。

先判斷男性或女性，判斷為女性則優先推薦音樂平台，再由後續的分析來判斷用戶被貼標後的群組與各類音樂喜愛相關成對做推薦。而判斷出為男性的用戶，則不先判別推薦影視或音樂，則是以後續分析方式來判斷用戶被貼標後的群組與各類影音多媒體喜愛相關程度做最後的推薦。

## 因素分析

推薦多媒體不僅是指給用戶同一種影音多媒體，也想找出相似類型的影音類型給推薦給用戶，又或者可能是用戶潛在喜好影音多媒體，讓用戶有耳目一新有的新鮮選擇。

接下來規劃分別將音樂類型變數和影視類型變數以及綜合音樂影視變數，下去做因素分析，找出特定影音類型是否有共同因素做為後續推薦標準。

### 音樂類型變數因素分析

音樂類型變數在本次問卷中共有19題，皆以五分量表形式，由1分為不喜歡至5分非常喜歡給分，題目如下音樂喜好程度問項表(表、十四)：

表、 十四 音樂喜好程度問項表

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 音樂類型題目 | | 各分次數 | | | | | 各分次數比(%) | | | | |
| **1** | **2** | **3** | **4** | **5** | **1** | **2** | **3** | **4** | **5** |
| 1 | 對我來說聽音樂非常享受以及喜愛 | 7 | 8 | 25 | 102 | 632 | 1% | 1% | 3% | 13% | 82% |
| 2 | 我喜歡搖滾樂 | 36 | 87 | 162 | 226 | 263 | 5% | 11% | 21% | 29% | 34% |
| 3 | 我喜歡流行音樂 | 45 | 123 | 201 | 232 | 173 | 6% | 16% | 26% | 30% | 22% |
| 4 | 我喜歡搖滾樂 | 83 | 147 | 226 | 195 | 123 | 11% | 19% | 29% | 25% | 16% |
| 5 | 我喜歡拉丁樂 | 144 | 201 | 179 | 133 | 117 | 19% | 26% | 23% | 17% | 15% |
| 6 | 我喜歡古典月 | 102 | 188 | 222 | 146 | 116 | 13% | 24% | 29% | 19% | 15% |
| 7 | 我喜歡另類音樂 | 154 | 172 | 196 | 136 | 116 | 20% | 22% | 25% | 18% | 15% |
| 8 | 我喜歡嘻哈、說唱 | 165 | 160 | 159 | 181 | 109 | 21% | 21% | 21% | 23% | 14% |
| 9 | 我喜歡舞蹈、迪斯科、放克 | 76 | 173 | 236 | 187 | 102 | 10% | 22% | 30% | 24% | 13% |
| 10 | 我喜歡音樂劇 | 157 | 181 | 212 | 133 | 91 | 20% | 23% | 27% | 17% | 12% |
| 11 | 我喜歡重金屬搖滾樂 | 298 | 171 | 121 | 101 | 83 | 39% | 22% | 16% | 13% | 11% |
| 12 | 我喜歡搖擺樂、爵士樂 | 153 | 183 | 199 | 162 | 77 | 20% | 24% | 26% | 21% | 10% |
| 13 | 我比起慢節奏音樂我喜歡快節奏 | 17 | 51 | 454 | 178 | 74 | 2% | 7% | 59% | 23% | 10% |
| 14 | 我喜歡雷鬼、斯卡(Ska) | 134 | 201 | 224 | 146 | 69 | 17% | 26% | 29% | 19% | 9% |
| 15 | 我喜歡電音迷幻樂 | 304 | 155 | 150 | 103 | 62 | 39% | 20% | 19% | 13% | 8% |
| 16 | 我喜歡朋克(Punk) | 246 | 180 | 170 | 118 | 60 | 32% | 23% | 22% | 15% | 8% |
| 17 | 我喜歡歌劇 | 293 | 227 | 133 | 78 | 43 | 38% | 29% | 17% | 10% | 6% |
| 18 | 我喜歡民間音樂 | 223 | 256 | 184 | 68 | 43 | 29% | 33% | 24% | 9% | 6% |
| 19 | 我喜歡鄉村音樂 | 255 | 272 | 154 | 65 | 28 | 33% | 35% | 20% | 8% | 4% |

經過因素分析音樂類型內部變數後，會先以KMO值評估資料是否適合基本上KMO 0.8，資料即非常適合做因素分析。音樂類型變數因素分析表(表、十五)中可看出KMO=0.767小於0.8，故沒有非常適合做因素分析。

表、 十五 音樂類型變數因素分析表



但仍然希望能做到上述中提到的藉由相同類型的音樂來做推薦，嘗試著進行往下的分析。通常使用的是特徵值大於1、可解釋總變異的70%以上或者是陡坡圖轉肘點判斷要有幾種共同因素。由於變數有18個若要依照可解釋總變異的70%以上選因素，由音樂類型變數特徵值相關矩陣表(表、十六)可觀察出需要挑選9個因素，但如此一來共同因素裡面可能僅會有1~2個變數特徵，無法達到推薦預期，以特徵值大於1或音樂類型變數特徵值陡坡圖(圖、三)選擇則僅需選出4或5個共同因素特徵值。

|  |  |
| --- | --- |
| 表、 十六 音樂類型變數特徵值相關矩陣表 | 圖、 三 音樂類型變數特徵值陡坡圖 |

變數和共同因素間需要高相關才能解釋因素的意涵，課文中建議loadings最少要高於0.6稱為高度或至少0.4。而轉軸後音樂類型變數與特徵值表相關係數表(表、十七)中選出來的各特徵值因素僅對1個變數有高loading，可解釋成每種音樂都有自己非常獨特性風格。

前述是希望能夠推薦相似類型的音樂，讓用戶能夠有更多選擇，增加用戶對平台的黏著度以及瀏覽器的停留時間，無法從因素分析中在同一特徵值內找到多個變數，近無法達成推薦目標，故不對音樂類型變數採用因素分析。

表、 十七 音樂類型變數因素分析表



### 影視類型變數因素分析

影視類型變數在本次問卷中共有10題，皆以五分量表形式，由1分為不喜歡至5分非常喜歡給分，題目如下影視類型喜好程度表(表、十八)：

表、 十八 影視類型喜好程度表

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 影視類型題目 | | 各分次數 | | | | | 各分次數比(%) | | | | | 4.5分 (%) |
| **1** | **2** | **3** | **4** | **5** | **1** | **2** | **3** | **4** | **5** |
| 1 | 我喜歡看電影 | 2 | 6 | 58 | 153 | 555 | 0% | 1% | 7% | 20% | 72% | 91% |
| 2 | 我喜歡看喜劇片 | 1 | 17 | 67 | 195 | 494 | 0% | 2% | 9% | 25% | 64% | 89% |
| 3 | 我喜歡看動畫片 | 39 | 92 | 153 | 191 | 299 | 5% | 12% | 20% | 25% | 39% | 63% |
| 4 | 我喜歡看奇幻/童話片 | 33 | 91 | 181 | 183 | 286 | 4% | 12% | 23% | 24% | 37% | 61% |
| 5 | 我喜歡看記錄片 | 28 | 100 | 201 | 241 | 204 | 4% | 13% | 26% | 31% | 26% | 57% |
| 6 | 我喜歡看動作片 | 50 | 126 | 171 | 217 | 210 | 6% | 16% | 22% | 28% | 27% | 55% |
| 7 | 我喜歡看浪漫的片 | 48 | 120 | 215 | 192 | 199 | 6% | 16% | 28% | 25% | 26% | 51% |
| 8 | 我喜歡看驚悚片 | 63 | 133 | 189 | 245 | 144 | 8% | 17% | 24% | 32% | 19% | 50% |
| 9 | 我喜歡看戰爭片 | 103 | 156 | 197 | 138 | 180 | 13% | 20% | 25% | 18% | 23% | 41% |
| 10 | 我喜歡看科幻片 | 105 | 162 | 194 | 163 | 150 | 14% | 21% | 25% | 21% | 19% | 40% |
| 11 | 我喜歡看恐怖片 | 197 | 159 | 167 | 135 | 116 | 25% | 21% | 22% | 17% | 15% | 32% |
| 12 | 我喜歡看西部牛仔片 | 285 | 253 | 145 | 58 | 33 | 37% | 33% | 19% | 7% | 4% | 12% |

經過因素分析音樂類型內部變數後，會先以KMO值評估資料是否適合基本上KMO 0.8，資料即非常適合做因素分析，而下表中可看出KMO=0.646小於0.8，甚至低於音樂類型變數KMO=0.767，可得知沒有不適合做因素分析故不再繼續往下做因素分析步驟。



### 影音綜合變數因素分析

由於音樂及影視變數皆不適合做因素分析，則選擇不在繼續對兩者類型結合之變數進行因素分析。

## 主成分分析

人口輪廓以多變數變異數分析方法可有效判別何種用戶在音樂及影音上有明顯差異，故保留此分析方式。而影音類型變數在前述運用的因素分析方法，均不適合，故不將因素分析對影音類型變數分析方式納入後續推薦方法中。

接下來要對興趣愛好類型的喜好程度找出是否能縮減變數再進一步對推薦方法有下一步的動作。

興趣愛好指標問卷中共有32題，皆以五分量表形式，由1分為不喜歡至5分非常喜歡給分，題目如各興趣喜好程度問項表(表、十九)：

表、 十九 各興趣喜好程度問項表

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 興趣愛好類型題目 | | 各分次數 | | | | | 各分次數比(%) | | | | | 4.5分 (%) |
| **1** | **2** | **3** | **4** | **5** | **1** | **2** | **3** | **4** | **5** |
| 1 | 社交 | 0 | 18 | 64 | 170 | 522 | 0% | 2% | 8% | 22% | 67% | 89% |
| 2 | 互聯網 | 7 | 28 | 148 | 235 | 356 | 1% | 4% | 19% | 30% | 46% | 76% |
| 3 | 外語 | 28 | 79 | 172 | 227 | 268 | 4% | 10% | 22% | 29% | 35% | 64% |
| 4 | 戶外運動 | 57 | 90 | 164 | 234 | 229 | 7% | 12% | 21% | 30% | 30% | 60% |
| 5 | 寵物 | 161 | 97 | 122 | 121 | 273 | 21% | 13% | 16% | 16% | 35% | 51% |
| 6 | 體育和休閒活動 | 106 | 113 | 164 | 146 | 245 | 14% | 15% | 21% | 19% | 32% | 51% |
| 7 | 競技水平的運動 | 154 | 96 | 144 | 142 | 238 | 20% | 12% | 19% | 18% | 31% | 49% |
| 8 | 購物 | 83 | 149 | 183 | 182 | 177 | 11% | 19% | 24% | 24% | 23% | 46% |
| 9 | 詩歌朗讀 | 156 | 120 | 147 | 131 | 220 | 20% | 16% | 19% | 17% | 28% | 45% |
| 10 | 科學和技術 | 85 | 135 | 212 | 179 | 163 | 11% | 17% | 27% | 23% | 21% | 44% |
| 11 | 歷史 | 79 | 153 | 218 | 163 | 161 | 10% | 20% | 28% | 21% | 21% | 42% |
| 12 | 電腦軟件、硬件 | 106 | 162 | 191 | 172 | 143 | 14% | 21% | 25% | 22% | 18% | 41% |
| 13 | 心理學 | 87 | 170 | 204 | 184 | 129 | 11% | 22% | 26% | 24% | 17% | 40% |
| 14 | 地理 | 98 | 157 | 220 | 158 | 141 | 13% | 20% | 28% | 20% | 18% | 39% |
| 15 | 腎上腺素運動 | 167 | 151 | 176 | 142 | 138 | 22% | 20% | 23% | 18% | 18% | 36% |
| 16 | 劇院 | 122 | 167 | 207 | 132 | 146 | 16% | 22% | 27% | 17% | 19% | 36% |
| 17 | 汽車 | 222 | 168 | 148 | 118 | 118 | 29% | 22% | 19% | 15% | 15% | 30% |
| 18 | 經濟、管理 | 205 | 193 | 159 | 117 | 100 | 26% | 25% | 21% | 15% | 13% | 28% |
| 19 | 生物學 | 198 | 207 | 154 | 95 | 120 | 26% | 27% | 20% | 12% | 16% | 28% |
| 20 | 政治 | 191 | 197 | 182 | 118 | 86 | 25% | 25% | 24% | 15% | 11% | 26% |
| 21 | 演奏樂器 | 368 | 114 | 98 | 76 | 118 | 48% | 15% | 13% | 10% | 15% | 25% |
| 22 | 藝術 | 203 | 193 | 184 | 103 | 91 | 26% | 25% | 24% | 13% | 12% | 25% |
| 23 | 跳舞 | 297 | 154 | 134 | 82 | 107 | 38% | 20% | 17% | 11% | 14% | 24% |
| 24 | 藥物 | 233 | 204 | 162 | 68 | 107 | 30% | 26% | 21% | 9% | 14% | 23% |
| 25 | 數學 | 289 | 149 | 167 | 90 | 79 | 37% | 19% | 22% | 12% | 10% | 22% |
| 26 | 名人生活方式 | 273 | 172 | 173 | 99 | 57 | 35% | 22% | 22% | 13% | 7% | 20% |
| 27 | 化學 | 354 | 187 | 81 | 66 | 86 | 46% | 24% | 10% | 9% | 11% | 20% |
| 28 | 宗教 | 315 | 169 | 143 | 78 | 69 | 41% | 22% | 18% | 10% | 9% | 19% |
| 29 | 法律 | 282 | 213 | 144 | 83 | 52 | 36% | 28% | 19% | 11% | 7% | 17% |
| 30 | 物理 | 348 | 184 | 113 | 82 | 47 | 45% | 24% | 15% | 11% | 6% | 17% |
| 31 | 詩歌寫作 | 464 | 112 | 79 | 65 | 54 | 60% | 14% | 10% | 8% | 7% | 15% |
| 32 | 園藝 | 405 | 175 | 105 | 50 | 39 | 52% | 23% | 14% | 6% | 5% | 11% |

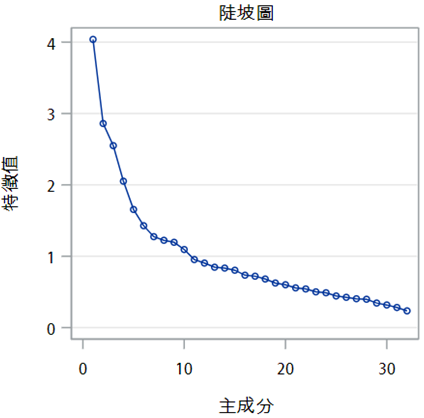
前述章節中性別能看出對影音有明顯差異，故接下來興趣愛好變數資料，將依照性別先拆分後再進行主成分分析。

### 女性興趣愛好類型變數的主成分分析

如下表(表、二十)所示，可看出沒有變數標準差佔比最大，則無須擔心個別變數的影響權重，則利用標準化資料作主成分分析。

通常使用的是特徵值大於1、可解釋總變異的70%以上(本次使用50%以上)或者是陡坡圖轉肘點判斷要有幾種特徵值。資料標準化後，將在32個變數中要用幾個主成分進行分類。如下表(表、二十一)能解釋總變異50%以上需要8個主成分，特徵值大於1需要10個主成分，而陡坡圖(圖、四)並不能明顯看出轉肘點。故最後以選取8個主成分來找出其相關性最大變數。

|  |  |
| --- | --- |
| 表、 二十 女性興趣愛好類型變數  基本統計敘述 | 表、 二十一 女性興趣愛好類型變數  相關矩陣特徵值表 |



圖、 四 女性興趣愛好類型變數陡坡圖

由女性興趣愛好類型變數與主成分相關係數表(表、二十二)可看出從32個變數收斂到8個主成分，但由於相關係數皆太低，並且僅應用了13個變數，其餘19個變數無法被歸類進來，且考慮到後續實際，主成分僅針對變數做收斂分群，但沒辦法對觀察值做後續動作，故不對女性各個主成分做命名動作與應用此分析方式。

表、 二十二 女性興趣愛好類型變數與主成分相關係數表

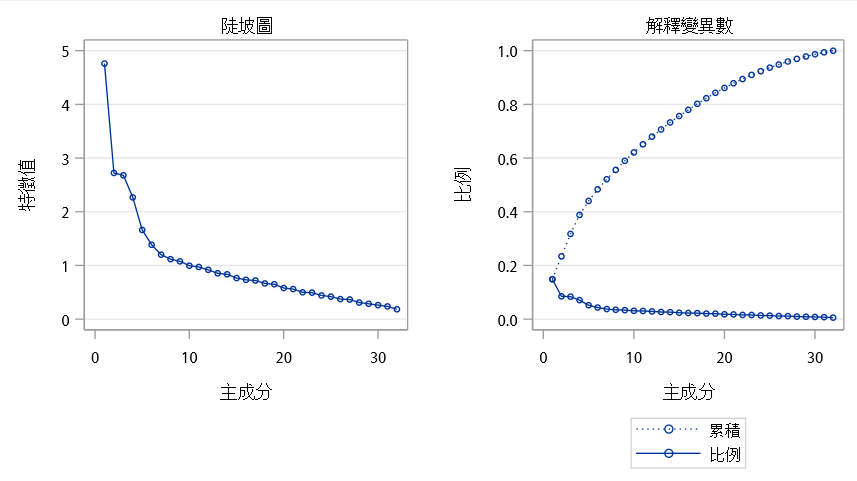


### 男性興趣愛好類型變數的主成分分析

如下表(表、二十三)所示，可看出沒有變數標準差佔比最大，則無須擔心個別變數的影響權重，則利用標準化資料作主成分分析。

通常使用的是特徵值大於1、可解釋總變異的70%以上(本次使用50%以上)或者是陡坡圖轉肘點判斷要有幾種特徵值。資料標準化後，將在32個變數中要用幾個主成分進行分類。如下表(表、二十四)能解釋總變異50%以上需要7個主成分，特徵值大於1需要9個主成分，而陡坡圖(圖、五)並難以判斷出轉肘點。故最後以總變異50%以上選取7個主成分來找出其相關性最大變數。

|  |  |
| --- | --- |
| 表、 二十三 男性興趣愛好類型變數  基本統計敘述 | 表、 二十四 男性興趣愛好類型變數  相關矩陣特徵值表 |



圖、 五 男性興趣愛好類型變數陡坡圖

由男性興趣愛好類型變數與主成分相關係數表(表、二十五)可看出從32個變數收斂到7個主成分，但由於相關係數皆太低，並且僅應用了16個變數，其餘16個變數無法被歸類進來，且考慮到後續實際，主成分僅針對變數做收斂分群，但沒辦法對觀察值做後續動作，故也不對男性各個主成分做命名動作與應用此分析方式。

表、 二十五 男性興趣愛好類型變數與主成分相關係數表



## 集群分析與多變量變異數分析

在進行音樂及電影推薦之前，我們希望可以依照用戶群體的面相來區別出在不同興趣的群體推薦其感興趣的音樂及電影類型，因此我們使用了階層式集群分析來區分出不同興趣的族群，目的是希望可以達到精準族群推薦。

### 集群數判別

本次集群分析判斷分群數量是依造陡坡圖轉肘點進行判斷。

#### 女性集群數判別

下左圖為女性群及分析陡坡圖，觀察出在點4有略為明顯轉肘點，故將女性分為4群，右下圖為以四群做為切分整份調查問卷女性的分群圖。

|  |  |
| --- | --- |
| 圖、 六 女性集群分析陡波圖 | 圖、 七 女性集群分析分群圖 |

#### 男性集群數判別

男性判別後集群數與女性相同，下左圖為男性群及分析陡坡圖，觀察出在點4有略為明顯轉肘點，故也將男性分為4群。右下圖為以四群做為切分整份調查問卷男性的群集分群圖。

|  |  |
| --- | --- |
| 圖、 八 女性集群分析陡波圖 | 圖、 九 男性集群分析分群圖 |

### 各集群顯著變數判別-多變量分析

在這也使用多變量變異數分析(MANOVA)進行，各別對女性、男性4群人觀察各變數是否有顯著差異，將有顯著差異變數挑出再以參數估計值來判斷最終能代表各群的變數為何。

在同一變數有兩群(含)以上有顯著差異，則會以參數估計值大的分給該群組，若一變數僅有一群有顯著差異仍然會觀察參數估計值為正才認列為該群組代表變異數。若特定族群在任一遍數均無顯著性，則會因應不同情形作不同方式的挑選。

以上述挑選方式以女性第四群並無變異數對其有顯著差異，並觀察期各變數之平均數及標準差與第三群十分相似，故將女性第四群與第三群合併為一群，最終女性分為三群。

在女性分為三群與男性維持四群後，代表女性各群變異數如下右表，男性各群變異數如下左表。

在代表變數選出來後，女性各群代表興趣愛好變數會與音樂類型變數直接進行相關性分析找出與相關性高的音樂類型進而推薦給被貼標到各群用戶。

而男性則是代表該群的興趣喜好變數與音樂及電影變數進行相關性分析，找出該群用戶最適合推薦音樂或電影。

|  |  |
| --- | --- |
| 表、 二十六 男性各群參數估計值與  變數平均值表 | 表、 二十七 女性各群參數估計值表 |

## 集群後相關係數分析

由集群分析先將性別女性、男性分開，在進行各別分群得到透過興趣愛好確定用戶分群貼標，以相關係數來對各群人要做什麼樣的影音媒體做推薦。

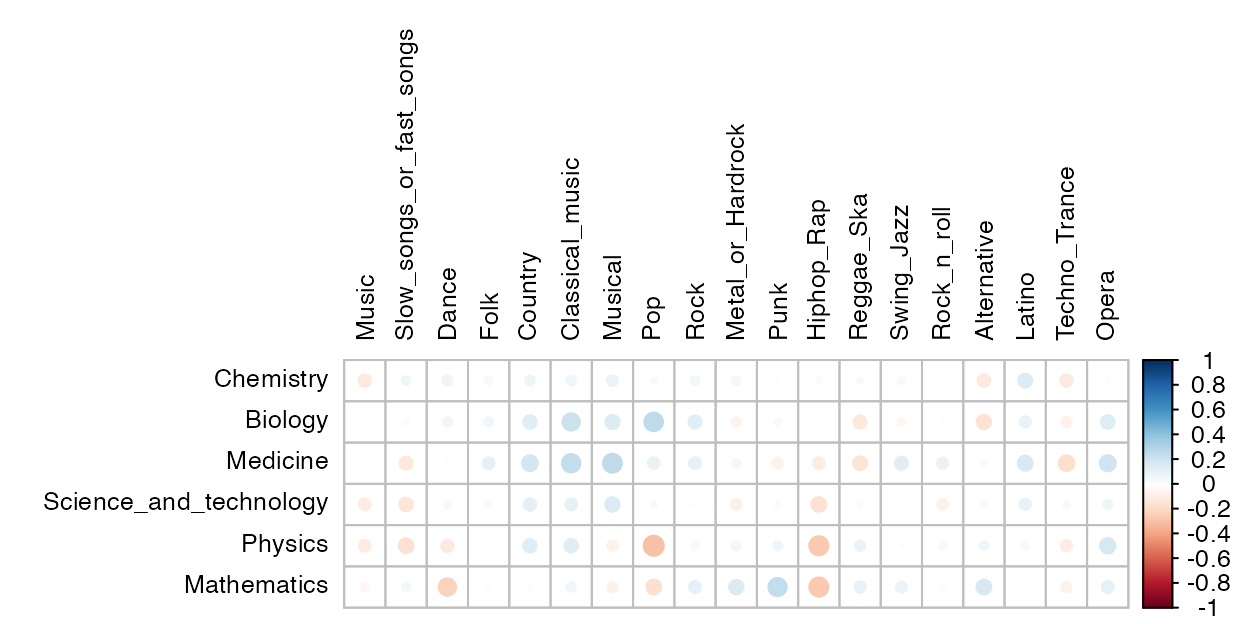
在多變數變異數分析中得知，女性應該先推薦音樂，而要推薦何種類型的音樂，由前一章節集群分析我們將分成3群，依照各群興趣愛好對應到的音樂類型高相關係數值喜好類型進行推薦。

而男性變異數分析我們得知，參數估計值在音樂和電影都為零，故先進行分群動作後，由前一章節集群分析我們將男性分成4群，依照各群興趣愛好對應到的音樂或電影以高相關係數進行推薦。

### 女性興趣愛好語音樂喜好類型

女性被分為4群，每群興趣愛好特項上節已說明，本節僅會說明，各群興趣愛好變數對應到音樂相關係數值高的進行優先推薦建議

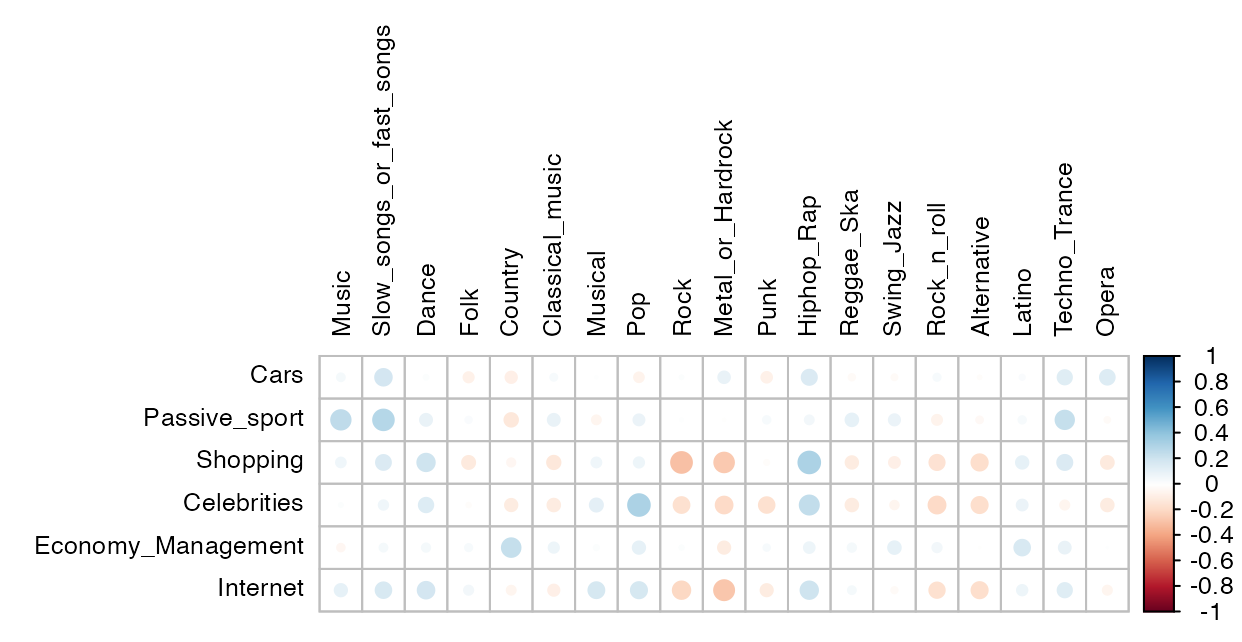
#### 女性第一群

下圖(圖、十)為興趣對音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的音樂類型；例如：推薦「鄉村音樂」、「古典樂」、「音樂劇」...等等。****

圖、 十 女性第一群興趣與音樂類型變數相關係數熱圖

#### 女性第二群

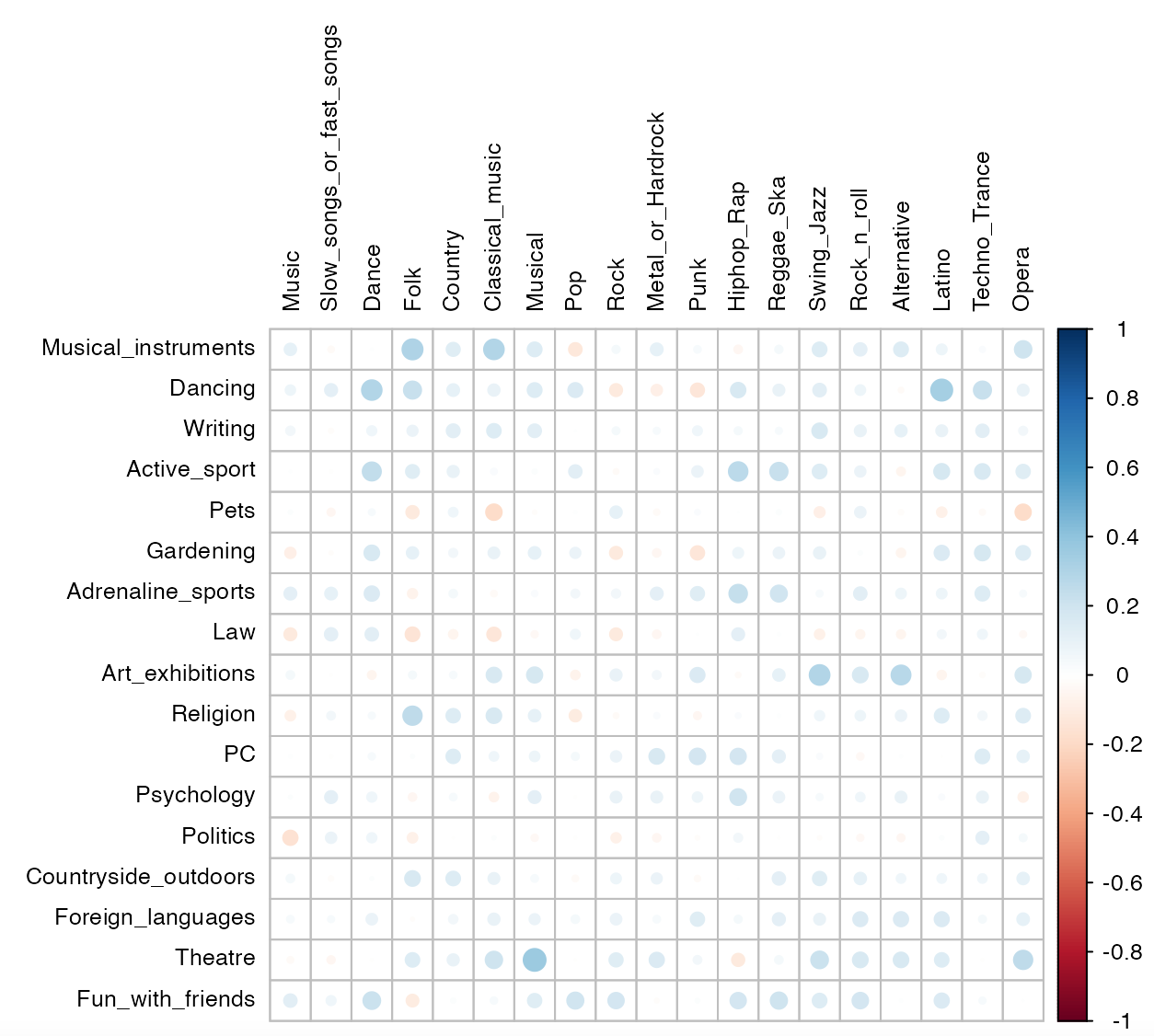
下圖(圖、十一)為興趣對音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的音樂類型；例如：推薦「快歌」、「舞曲」、「流行樂」、「嘻哈」、「電子快節奏音樂」...等等。

****

圖、 十一 女性第二群興趣與音樂類型變數相關係數熱圖

#### 女性第三群

下圖為(圖、十二)興趣對音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的音樂類型；例如：推薦「舞曲」、「民歌」、「音樂劇」、「古典樂」、「爵士樂」、「拉丁樂」...等等。

****

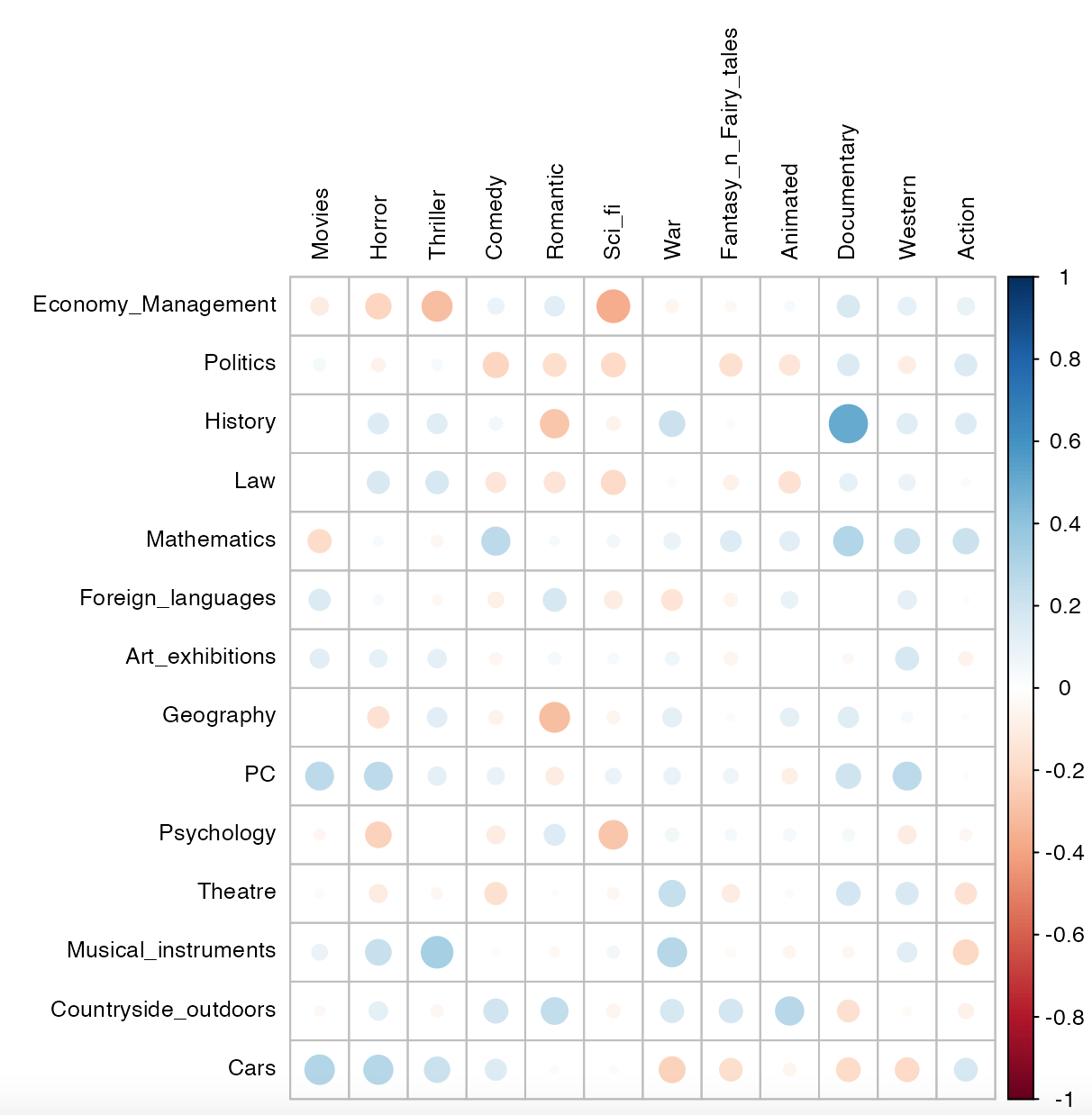
圖、 十二 女性第三群興趣與音樂類型變數相關係數熱圖

### 男性興趣愛好語音樂喜好類型

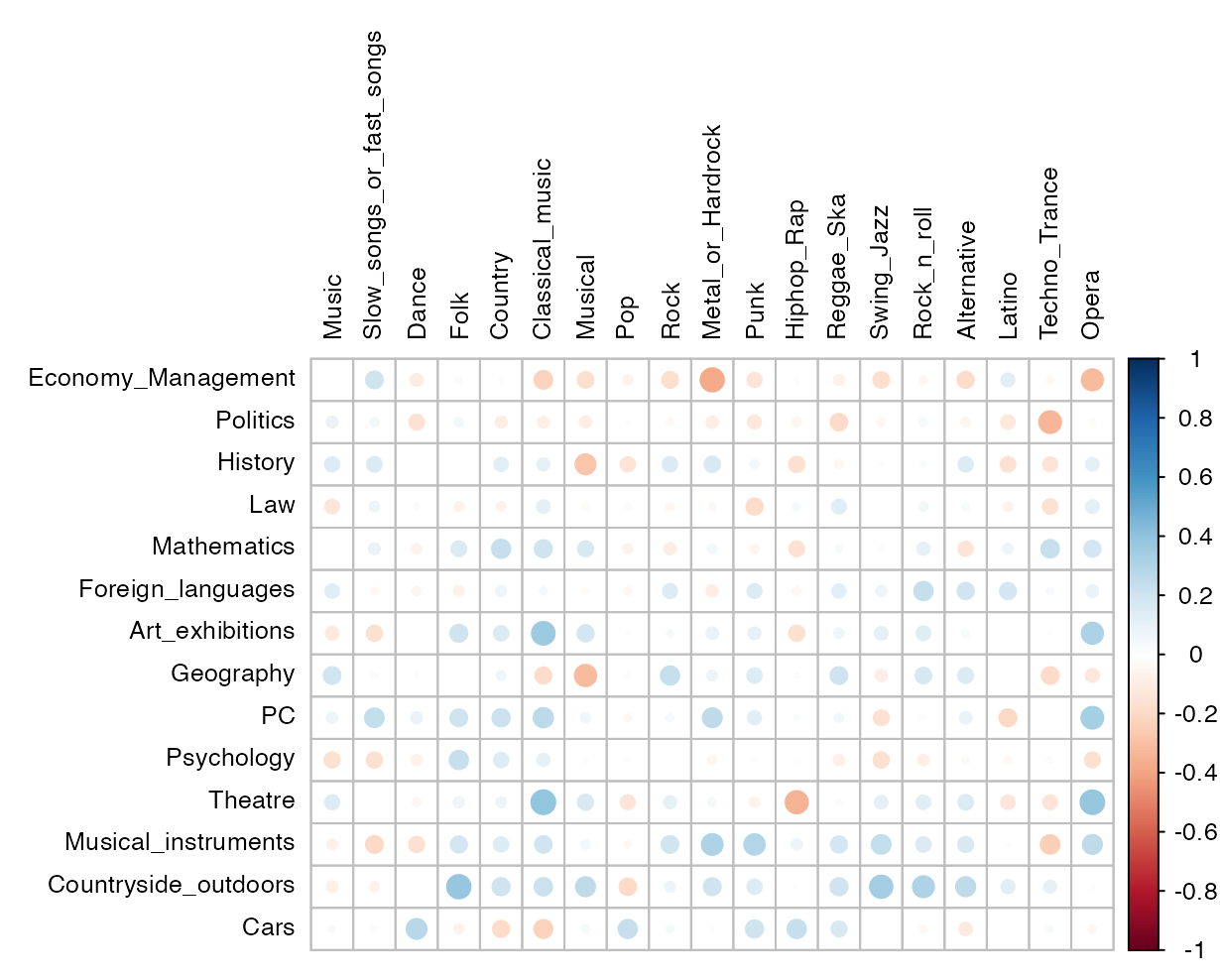
男性被分為4群，每群興趣愛好特項上節已說明，本節僅會說明，各群興趣愛好變數對應到音樂或電影相關係數值高的優先進行推薦建議

#### 男性第一群

下兩張圖(圖、十三 圖、十四)為興趣對電影類型及音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的電影及音樂類型；例如：推薦「恐怖片」、「驚悚片」、「戰爭片」、「紀錄片」...等等。音樂則會推薦「古典樂」、「歌劇」、「民謠」、「鄉村音樂」類型的音樂。



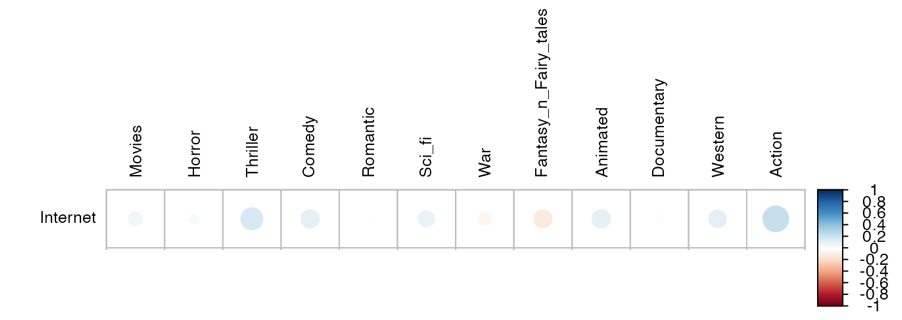
圖、 十三 男性第一群與影視類型變數相關係數熱圖

****

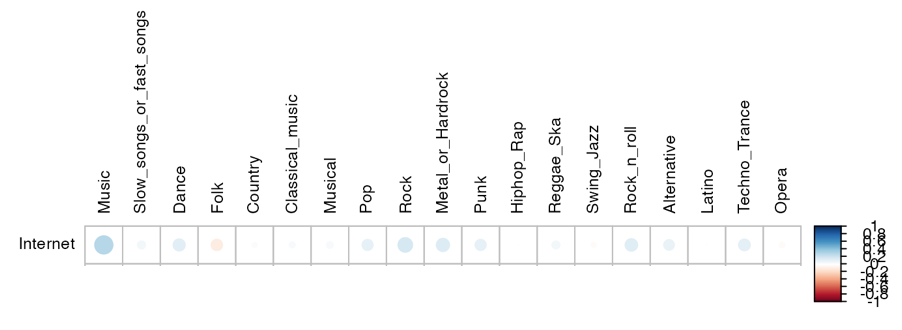
圖、 十四 男性第一群與音樂類型變數相關係數熱圖

#### 男性第二群

下兩張圖(圖、十五 圖、十六)為興趣對電影類型及音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的電影及音樂類型；例如：推薦「驚悚片」、「喜劇」、「動作片」類型電影。音樂則會推薦「搖滾樂」、「金屬樂」、「龐克音樂」...等等。

****

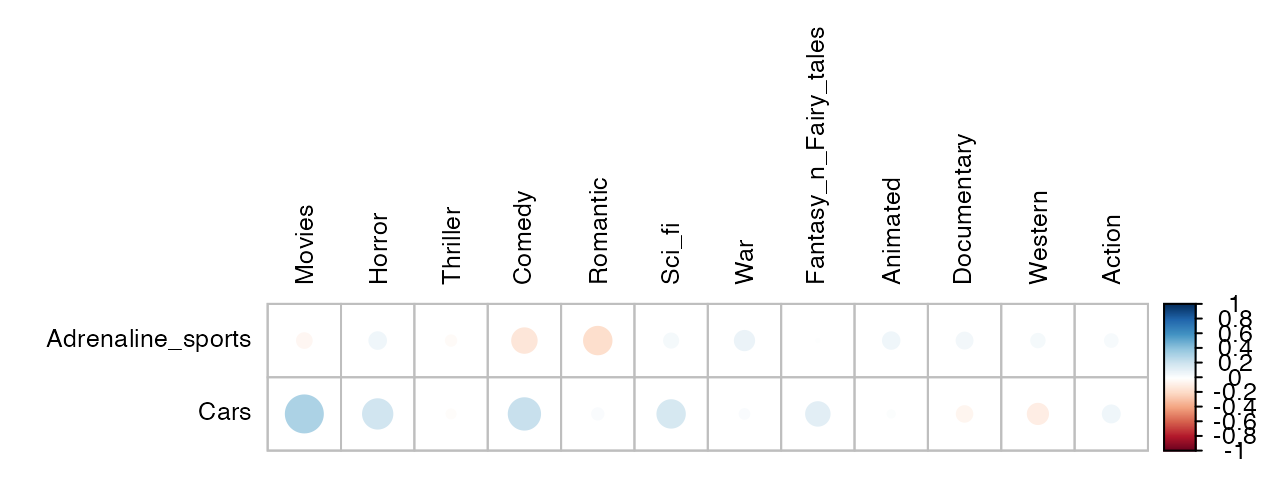
圖、 十五 男性第二群與影視類型變數相關係數熱圖

****

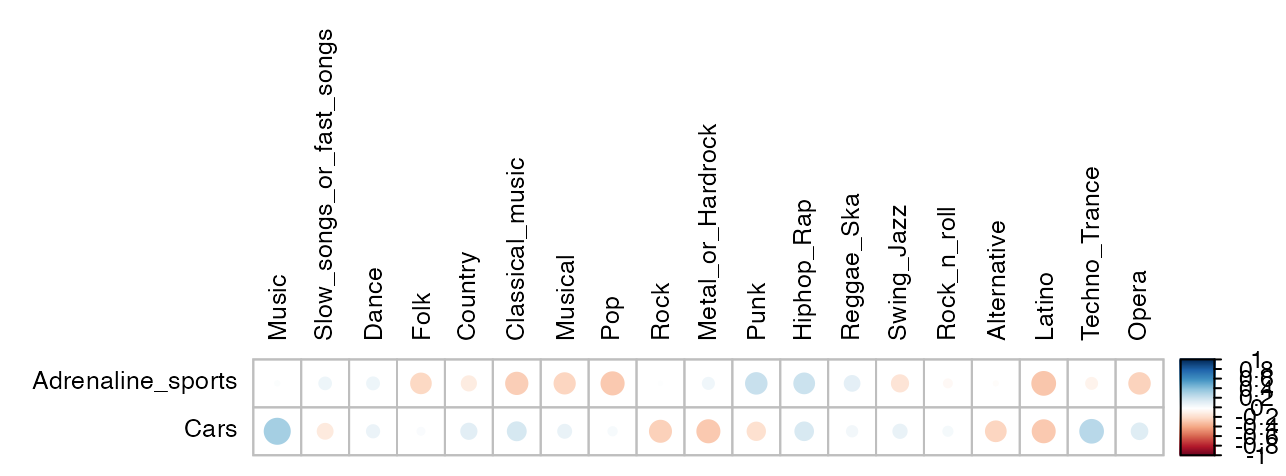
圖、 十六 男性第二群與音樂類型變數相關係數熱圖

#### 男性第三群

下兩張圖(圖、十七 圖、十八)為興趣對電影類型及音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的電影及音樂類型；例如：推薦「恐怖片」、「科幻片」、「幻想與童話」...等等。音樂則會推薦「龐克音樂」、「嘻哈饒舌音樂」、「電子快節奏」。

****

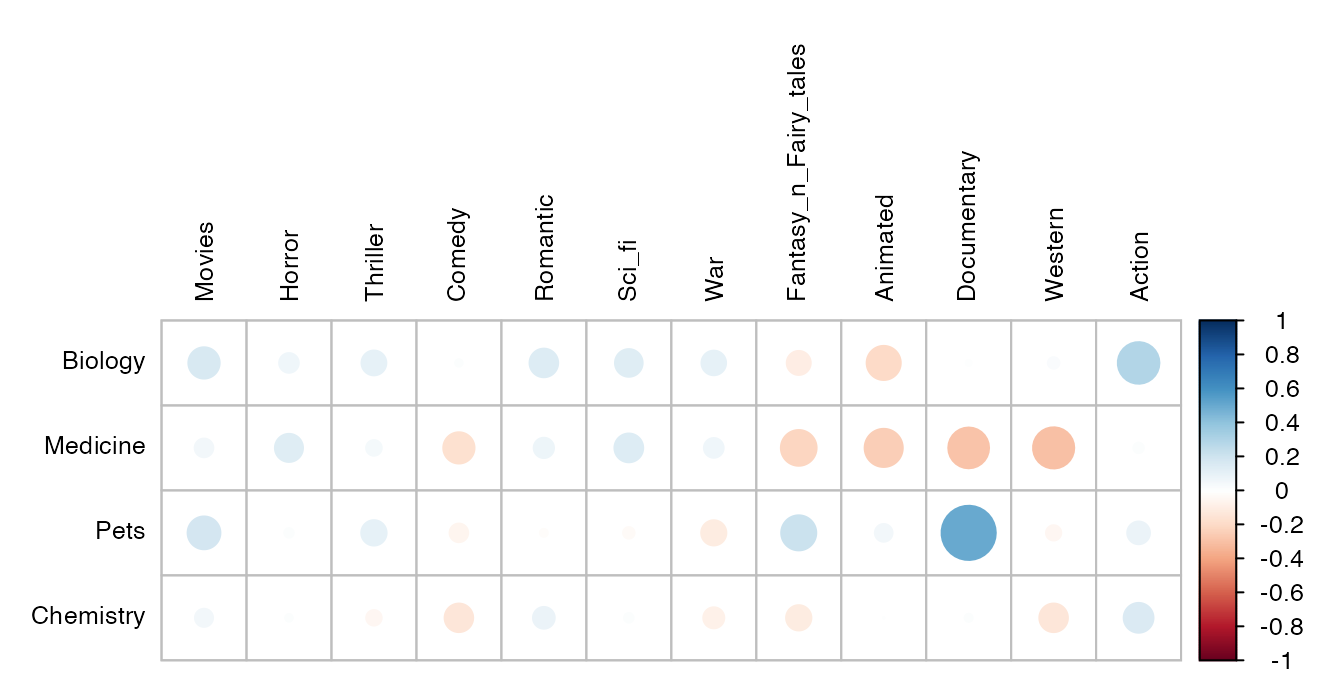
圖、 十七 男性第三群與影視類型變數相關係數熱圖

****

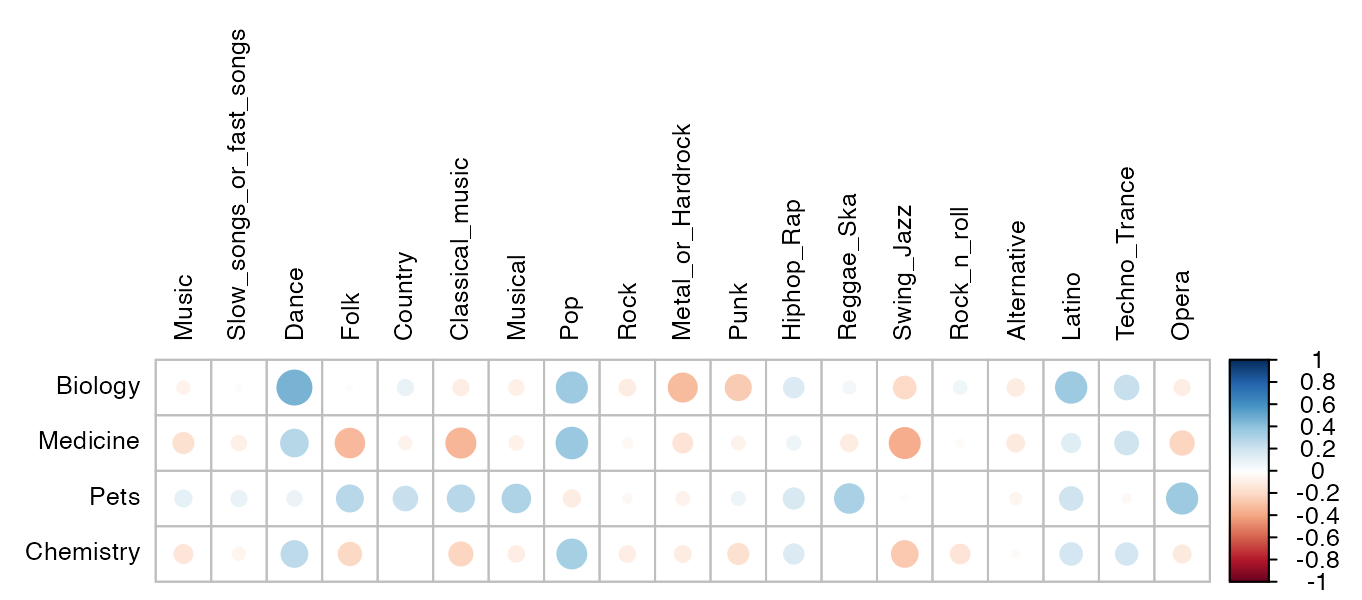
圖、 十八 男性第三群與音樂類型變數相關係數熱圖

#### 男性第四群

下兩張圖(圖、十九 圖、二十)為興趣對電影類型及音樂類型的相關係數熱圖。經由相關係數熱圖，我們會推薦與興趣呈現正相關較高的電影及音樂類型；例如：推薦「紀錄片」、「動作片」...等等。音樂則會推薦「流行音樂」、「舞曲」、「拉丁樂」...等等。

****

圖、 十九 男性第四群與影視類型變數相關係數熱圖

****

圖、 二十 男性第四群與音樂類型變數相關係數熱圖

總結本章興趣與相關係數推薦的方式，優先判斷該名用戶在該群體中興趣變數中再挑選高於客戶不排斥的分數高於3分做為該客戶的代表興趣。

再利用這些代表興趣變數找到相關係數為正的影音類別做相關係數排序，排序方法如下，以興趣分數大的相關係數由大至小做排序，再以興趣分數次大的相關係數由大至小做排序以此類推，若遇到數個興趣變數分數相同時，則由程式自動隨機對興趣變數先做排序後，再對其相對應的相關係數由大至小做排序。

推薦用戶兩種版位，第一種為依照該用戶喜好推薦影音類別，第二種為潛在價值影音類別。前述排序後前五名(含相關係數同分)的選入第一種喜好推薦影音類別清單，排序後五名的影音類別，由系統隨機挑選五種影音類別放入第二種為潛在價值影音類別清單。

# 結論

本次期末報告主要研究影音多媒體類別推薦方法，使用英國社會科學與經濟科學學院資料，透過在多變量分析課堂中學習的方法中挑出幾個可能使用的分析方法對變數以及資料群體進行分析，試圖嘗試找出最適合本次資料的推薦方法。

預先由人口輪廓類型變數，性別、年齡層、教育程度共三種變對音樂及電影喜好程度兩變數，進行多變量變異數分析，僅有性別對音樂及電影喜好程度兩變數有顯著差異，其中女性對音樂有顯著偏好，故第一推薦方向為，資料若為女性則優先推薦音樂，若為男性待後續分析結果再判斷優先推薦影音類別。

由音樂、影視及興趣等三大類型變數分別以因素分析及主成分分析試圖找出共同因素或變數組成分，但此資料結果顯示均不適合由變數面相出發。

進而由觀測值面相出發，採取群集分析方式嘗試。前述性別有顯著性差異，故分別對女性及男性進行群及分析，找出各別適合分群數，女性應分為四群，男性也應分為四群。

接著對各性別下之各族群與興趣喜好各變數進行多變數變異數分析，試圖找出各群代表興趣喜好變數，以便對各族群貼標及後續推薦依據，在此時發現男性及女性原在群及分析顯示應分為四群，但在多變數變異數分析下發現第四族群對各個變數均無參數估計值，且皆顯示為冗餘參數，故採取下列作法，對各族群下個興趣愛好變數之分數做出平均值後對應前三群，發現女性第四群興趣愛好變數平均值與第三群興趣愛好變數平均值相似，顧將第四群女性族群併入第三群中；而男性發現在第四群興趣愛好變數平均值有明顯與前三群不同，仍分為四群。故男性第四群代表變數則由平均數得出，其餘群體皆由參數估計值得出。

最後對女性群體代表變數僅對音樂類型變數進行相關性分析，找出適合推薦音樂類型以及潛在音樂類型。男性各群體代表變數則對音樂及影視類型變數均進行相關信分析，找出適合推薦音樂或影片類型以及潛在音樂或影片類型。

# 附件、一

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

# 附件、二

研究動機背景一：<https://meet.bnext.com.tw/blog/view/72612?>

研究動機背景二：<https://www.inside.com.tw/article/25719-spotify-netflix-hub-tiktok-discover>