

IMDb

TOP 1000 Movies



LET'S GET STARTED



組員：黃詩涵、陳玟儒、林柏辰、黃宥芯、劉
貞莉、溫展德



| CONTENTS

01

資料簡介

Codebook與IMDb

02

研究方向

建立電影推薦系統
設定目標客群

03

變數分析與處理

處理變數資料與分
析變數關係

04

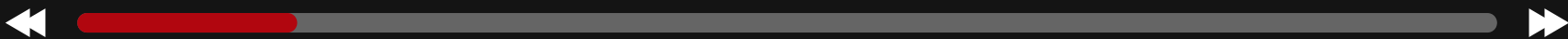
成果與展望

利用R shiny 製作
推薦系統





01 資料簡介



Codebook

變數名稱	變數類型	資料描述	代號表示
...1	Number	電影編號	評分排列順序
Poster_Link	Character	海報連結	圖片網址
Series_Title	Character	電影名稱	
Released_Year	Character	發行年份	西元年份(有一筆資料:“PG”)
Certificate	Character	電影分級	包含各國不同分級方式
Runtime	Character	電影時間長度	單位:min
Genre	Character	電影類型	
IMDb_Rating	Number	IMDb上的評分(大眾)	1-10顆星, 星數越多評價越高
Overview	Character	電影摘要	
Meta_score	Number	專業影評人評分	百分制, 數值越高評價越高
Director	Character	導演名稱	
Star1,Star2,Star3,Star4	Character	演員名稱	
No of votes	Number	IMDb上評分的用戶數量	整數(單位:人)
Gross	Number	電影獲利(毛利)	整數(單位:美金)

| 如何篩選前1000 IMDb Rating?

$$WR = \frac{Rv + Cm}{v + m}$$

- R = 該電影平均分數 = (Rating)
- v = 投票人數 = (votes)
- m = 要求最小人數 (currently 25000)
- C = 全部電影的平均分數 (currently 7.1)

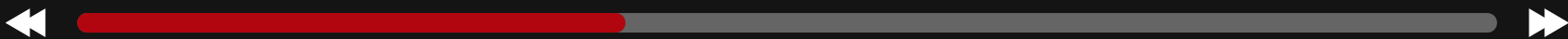


避免投票人數少導致分數過於極端的情況發生



02

研究方向



I 研究目標

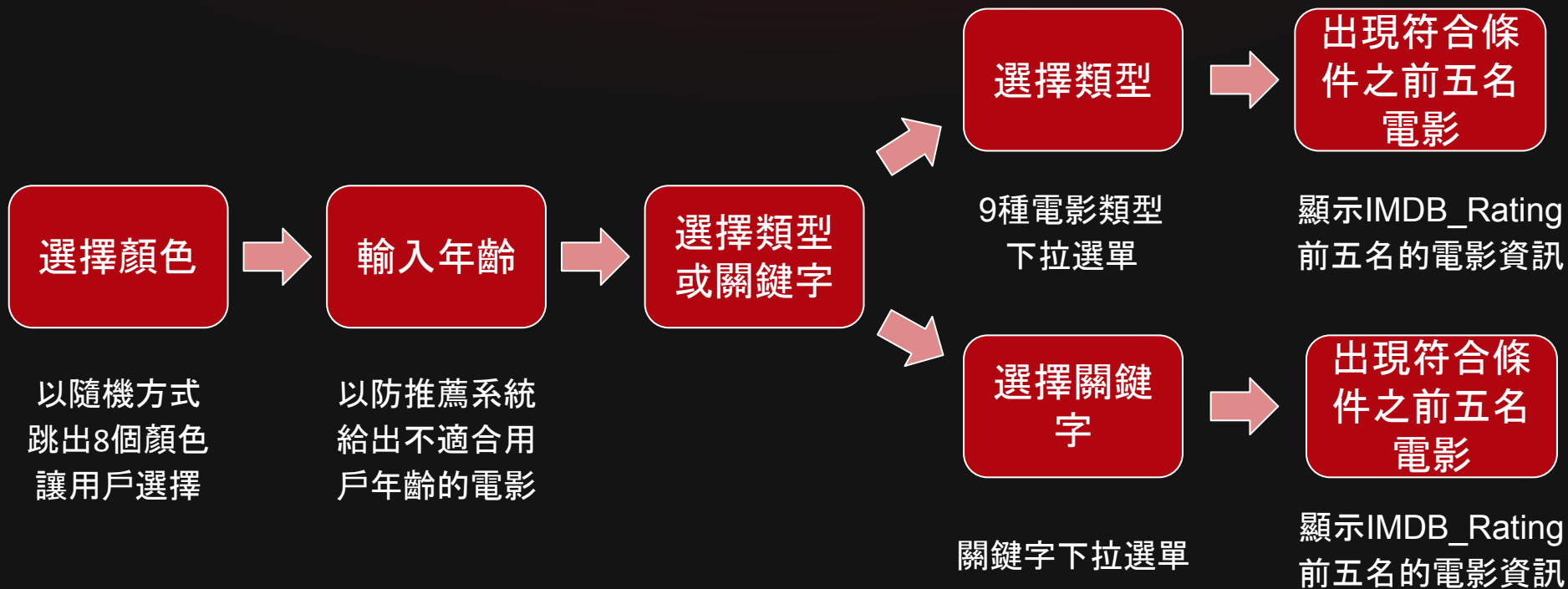
我們希望為IMDb建立一個推薦系統，透過優化這個推薦系統來提高舊有用戶的滿意度，同時吸引對推薦系統有興趣的新用戶。們

目標客群

- 平台舊用戶
- 對推薦系統感興趣之新用戶



| 推薦系統流程圖





03

變數處理與分析

電影海報、電影類型、關鍵字、電影分級



| 變數處理-海報資訊與色彩探索

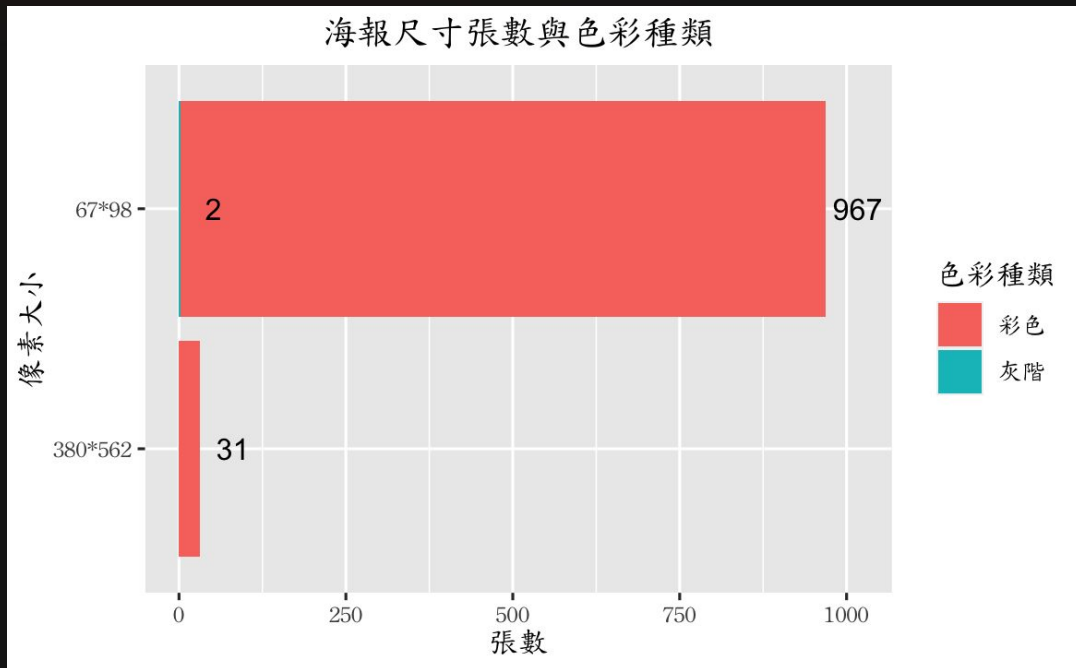
● 目標

- 透過海報資訊(如:連結是否有效等)了解資料集有否遺失值
- 取得每張海報中所使用的三種主要顏色
- 了解所有海報主要顏色之分佈
- 利用色彩差異數值辨別相近色彩

變數處理-海報資訊與色彩探索

● 海報資訊

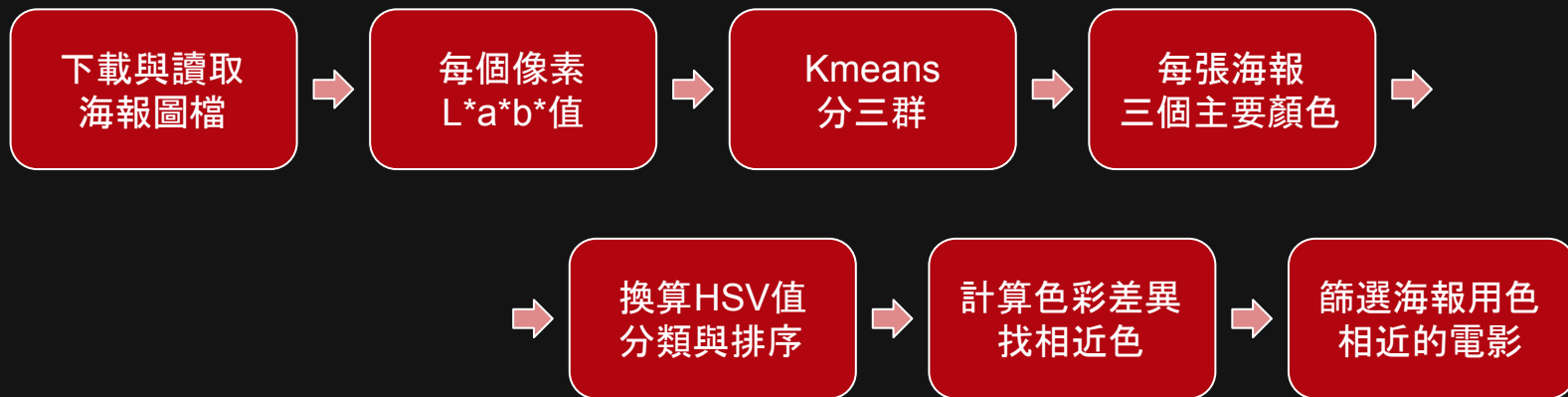
- 確認連結是否有效
- 進行補值
- 取得圖像大小
- 確認色彩種類



| 變數處理-海報資訊與色彩探索

● 色彩探索

- Python: OpenCV, scikit-learn
- R: ColorNameR



變數處理-海報資訊與色彩探索

● 色彩探索

- CIELAB色彩空間($L^*a^*b^*$)
 - $0(\text{黑}) \leq L^* \leq 100(\text{白})$:感知的亮度
 - $-127(\text{綠}) \leq a^* \leq 127(\text{紅})$
 - $-127(\text{藍}) \leq b^* \leq 127(\text{黃})$
- 感知上統一的空間
- 給定的數字變化對應於相似的感知顏色變化

A L 50
a 10
b 10

B L 50
a 10
b 20

C L 50
a 10
b 50

D L 10
a -50
b -50

下載與讀取海報圖檔



每個像素 $L^*a^*b^*$ 值



Kmeans分三群



每張海報三個主要顏色



換算HSV值分類與排序



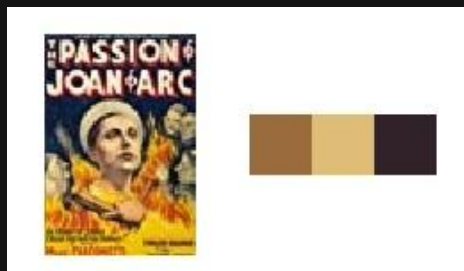
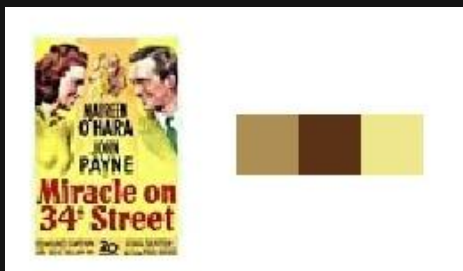
計算色彩差異找相近色



篩選海報用色相近的電影

變數處理-海報資訊與色彩探索

● 色彩探索



下載與讀取海報圖檔

每個像素L*a*b*值



Kmeans分三群



每張海報三個主要顏色



換算HSV值分類與排序



計算色彩差異找相近色

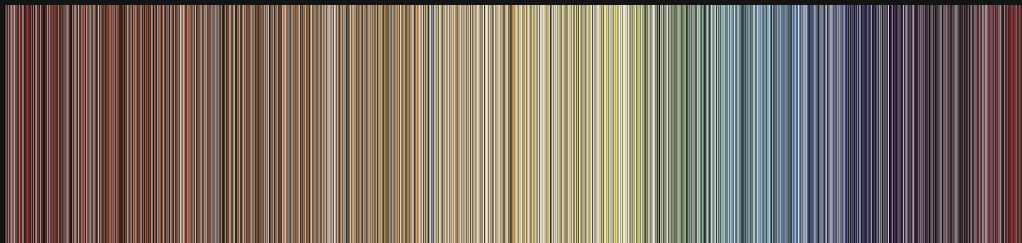


篩選海報用色相近的電影

| 變數處理-海報資訊與色彩探索

● 色彩探索

- HSV色彩空間(OpenCV)
 - $0 \leq H \leq 180$: 色相 / 顏色名稱
 - $0 \leq S \leq 255$: 飽和度 / 色彩純度
 - $0 \leq V \leq 255$: 明度 / 亮度
- 所有海報主要顏色依色相H遞增排序



下載與讀取海報圖檔

每個像素L*a*b*值

Kmeans分三群

每張海報三個主要顏色



換算HSV值分類與排序

計算色彩差異找相近色



篩選海報用色相近的電影

變數處理-海報資訊與色彩探索

● 色彩探索

- HSV基本顏色分量範圍(OpenCV)

	黑		白	紅		橙	黃	綠	藍	紫
Hmin	0		0	0	156	10	25	34	77	124
Hmax	180		180	10	180	25	34	77	124	156
Smin	0	0	0	43						
Smax	255	43	43	255						
Vmin	0	46	220	46						
Vmax	46	220	255	255						

下載與讀取海報圖檔

每個像素L*a*b*值

Kmeans分三群

每張海報三個主要顏色



換算HSV值分類與排序

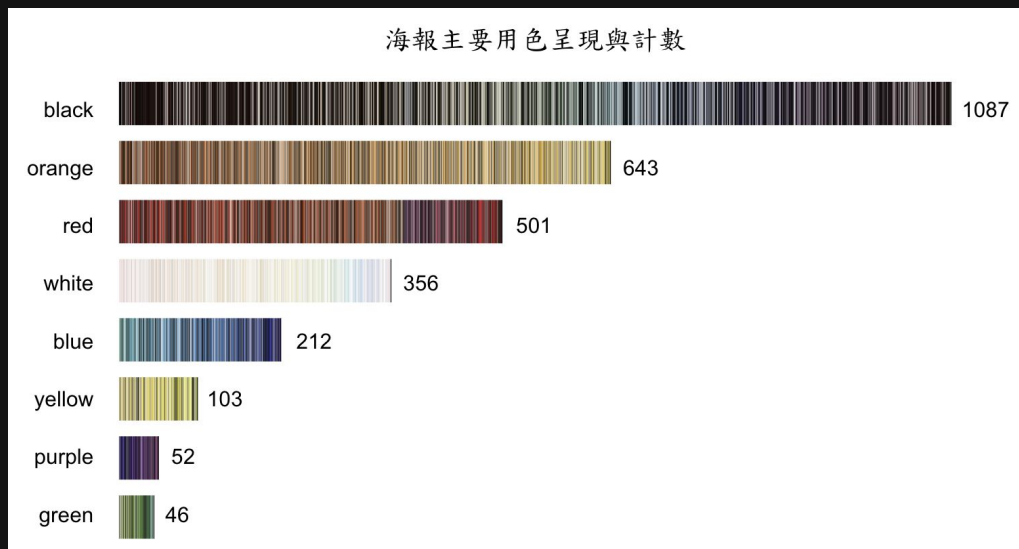
計算色彩差異找相近色

篩選海報用色相近的電影

變數處理-海報資訊與色彩探索

● 色彩探索

- HSV基本顏色分量範圍(OpenCV)



下載與讀取海報圖檔

每個像素L*a*b*值

Kmeans分三群

每張海報三個主要顏色



換算HSV值分類與排序



計算色彩差異找相近色

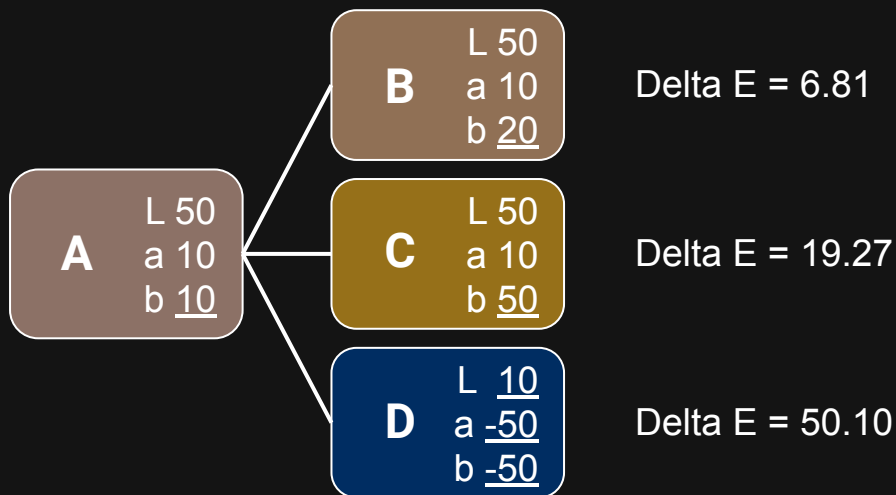


篩選海報用色相近的電影

變數處理-海報資訊與色彩探索

● 色彩探索

- L*a*b*色彩差異 Delta E
- CIEDE2000:複雜、考慮知覺非均勻特性



下載與讀取海報圖檔

每個像素L*a*b*值

Kmeans分三群

每張海報三個主要顏色

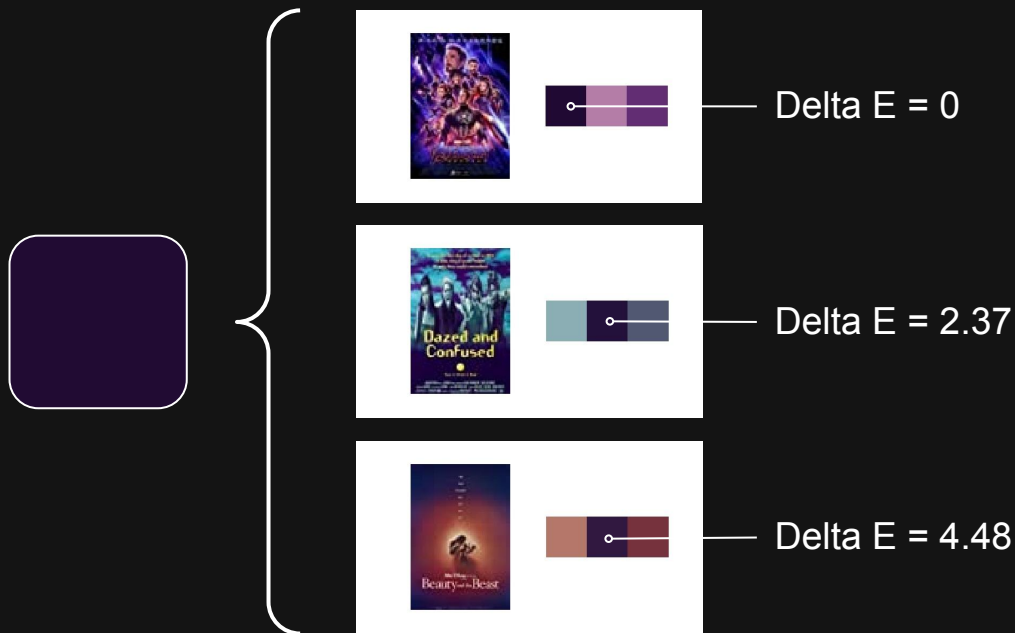
換算HSV值分類與排序

計算色彩差異找相近色

篩選海報用色相近的電影

變數處理-海報資訊與色彩探索

● 色彩探索



下載與讀取海報圖檔

每個像素L*a*b*值

Kmeans分三群

每張海報三個主要顏色

換算HSV值分類與排序

計算色彩差異找相近色

篩選海報用色相近的電影

| 變數處理-海報資訊與色彩探索

● 小結

- 1000張海報圖檔皆取得, 無遺失值
- 採用L*a*b*色彩並分群取得每張海報的三個主要顏色
- 採用HSV色彩定義基本顏色分量範圍, 將3000個顏色分成八類
- 觀察HSV色相排序, 發現海報色彩以黑色最多橘色次之
- 以CIEDE2000計算色彩差異, 於推薦系統中篩選海報用色相近的電影

I 變數處理-Genre電影類型資料分割

由於電影類型變數是由1~3種不同的類型所組成，因此為了方便後續分析，我們對類型資料進行分割。

Genre
Drama
Crime, Drama
Action, Crime, Drama
Crime, Drama
Crime, Drama
Action, Adventure, Drama
Crime, Drama



genre1	genre2	genre3
Drama	NA	NA
Crime	Drama	NA
Action	Crime	Drama
Crime	Drama	NA
Crime	Drama	NA
Action	Adventure	Drama
Crime	Drama	NA

| 21電影類型

Genre	類型
Comedy	喜劇
Romance	浪漫
Action	動作
Adventure	冒險
Sport	體育
Drama	劇情
Family	家庭

Genre	類型
Music	音樂
Musical	音樂劇
Crime	犯罪
Film-Noir	悲劇
Horror	恐怖
Mystery	懸疑
Thriller	驚悚

Genre	類型
Biography	傳記
History	歷史
War	戰爭
Western	西方
Fantasy	奇幻
Sci-Fi	科幻
Animation	動畫

I 合併相近類型

為了方便我們的TA，也就是平台用戶做選擇，我們將類型濃縮成以下9類。

Genre	類型
Comedy, Romance, Family	歡樂
Action, Adventure, Sport	熱血
Drama	劇情
War, Western	西方&戰爭
Biography, History	傳記&歷史

Genre	類型
Crime, Film-Noir, Horror, Mystery, Thriller	黑暗
Music, Musical	音樂
Fantasy, Sci-Fi	奇幻
Animation	動畫

I 建立Dummy Variable

在Top1000的電影中可被歸類為這9種類型的電影分別有多少

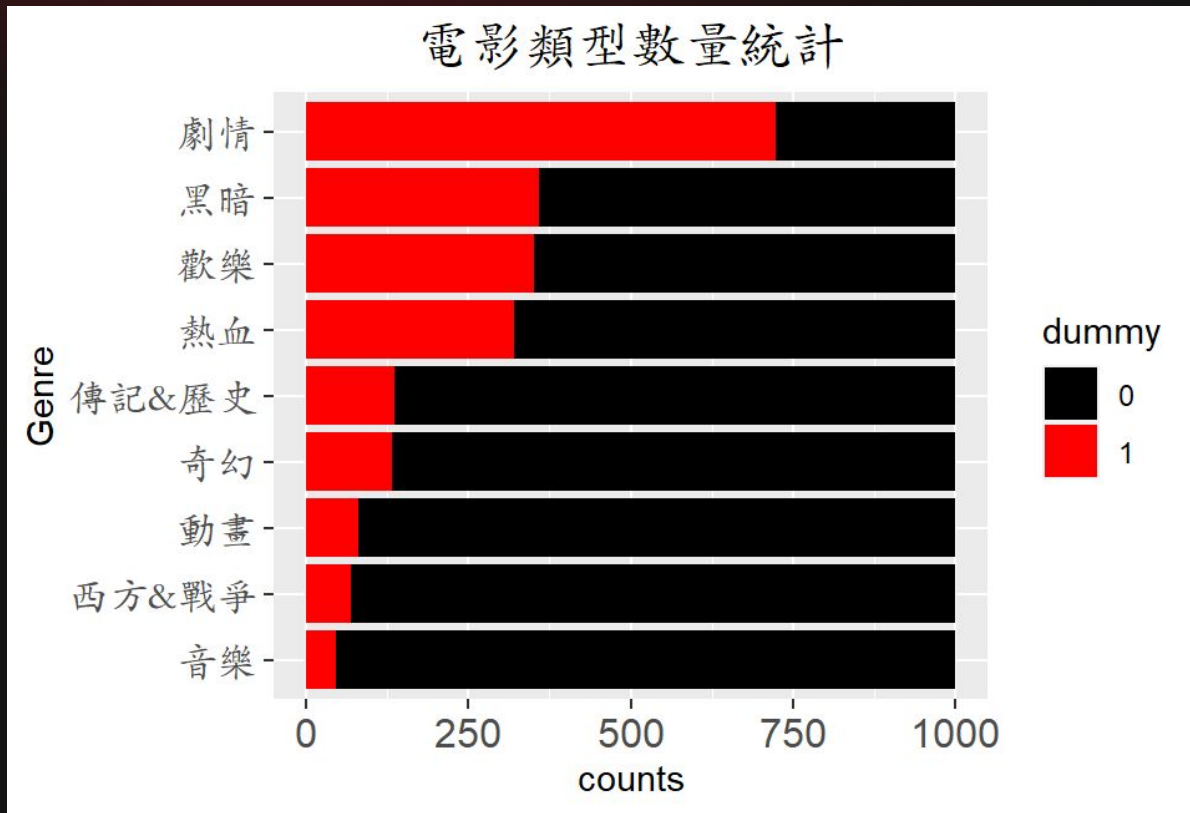
類型	個數	比例
歡樂	352	35.2%
熱血	320	32%
劇情	724	72.4%
西方&戰爭	70	7%
傳記&歷史	137	13.7%

類型	個數	比例
黑暗	360	36%
音樂	47	4.7%
奇幻	133	13.3%
動畫	82	8.2%

| 電影類型統計



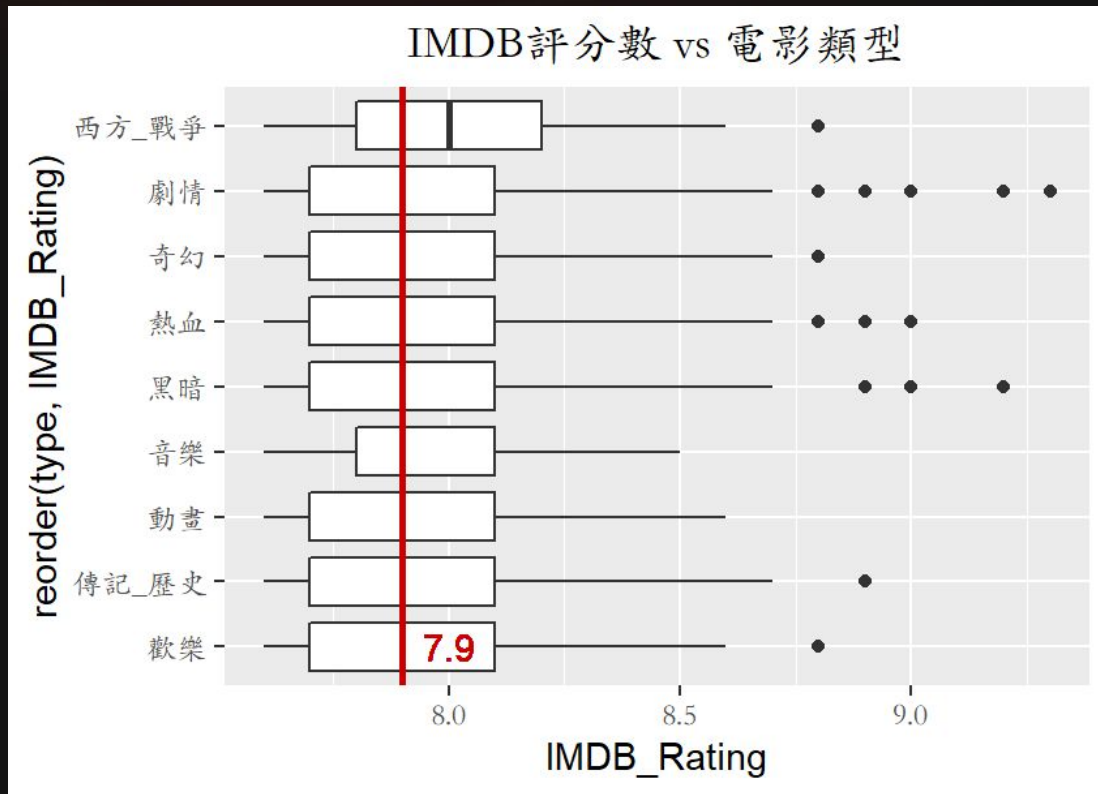
Top1000的電影中，類型為劇情的最多，且明顯高於其他類型。



| 電影類型 vs IMDb_Rating



除了西方&戰爭類型的電影IMDb評分數之中位數較高，為8分左右之外，其餘8種類型評分數的中位數皆為7.9分左右，可知不同類型電影的評分數並沒有明顯的差異。

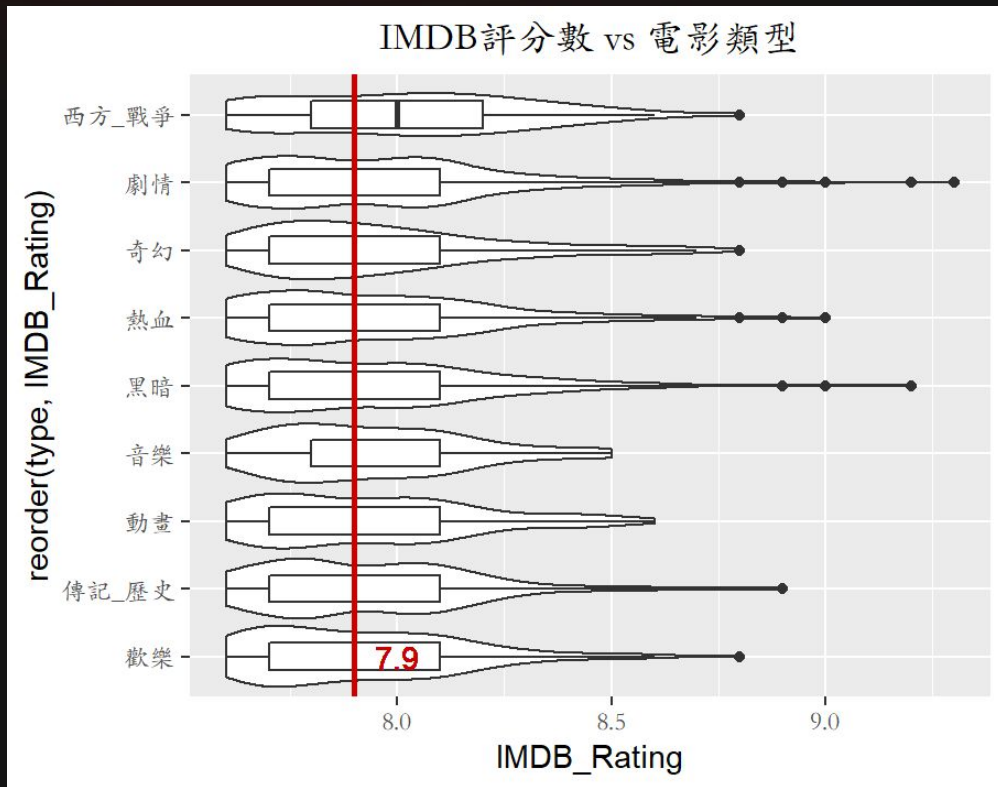




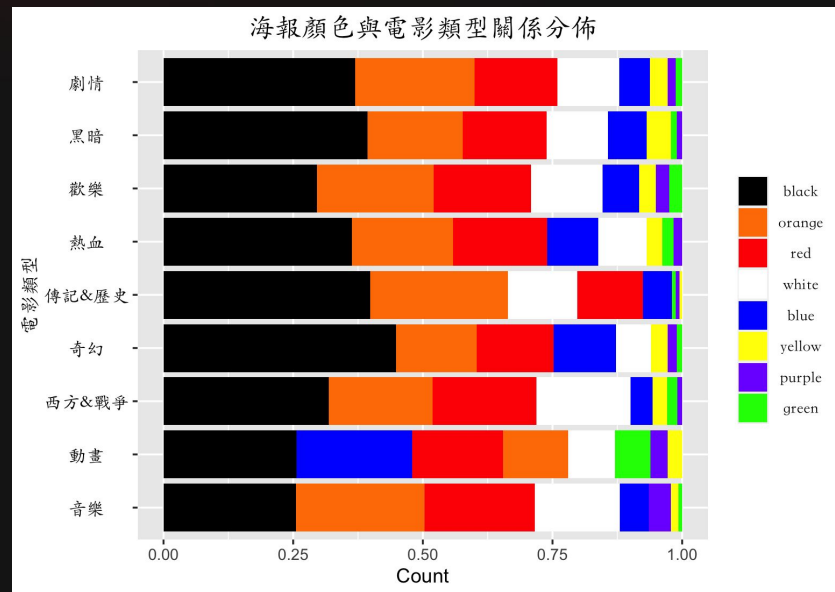
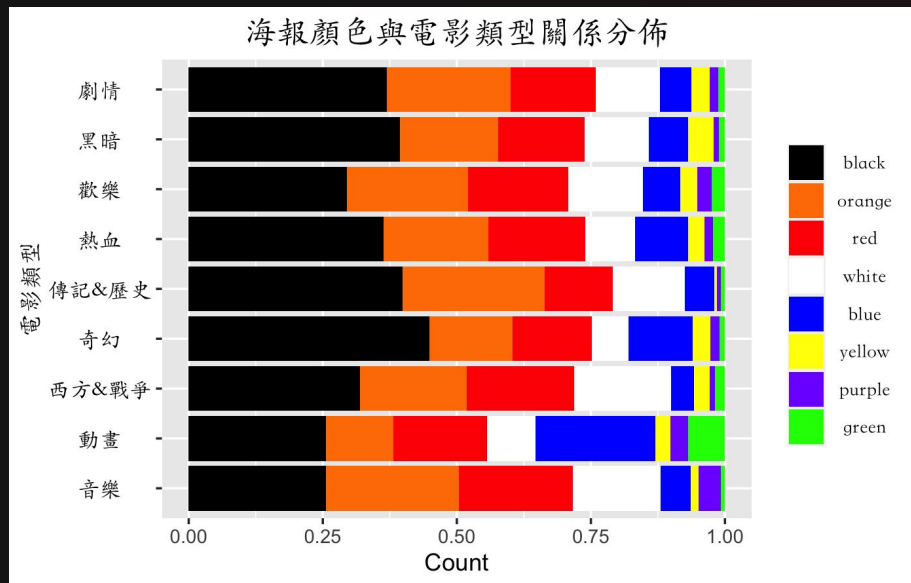
電影類型 vs IMDb_Rating

除了奇幻類型的電影之IMDb評分數為右偏分布之外，其餘8種類型電影的IMDb評分數皆有雙峰情形，皆在7.9分左右的位置有些微凹陷。

9種電影類型之IMDb評分數皆集中在7.6分到8.1分之間。



| 電影類型 vs 海報色彩



| Genre小結

- Top1000的電影中，類型為劇情的最多，且明顯高於其他類型。
- 不同類型電影的IMDb評分數並沒有明顯的差異。
- 9種電影類型之IMDb評分數皆集中在7.6分到8.1分之間。
- 並沒有特定電影類型偏好使用特定色彩來製作海報的情形發生。



綜上所述，由於不同電影類型的海報色彩與IMDb評分數皆無明顯差異，所以推薦系統在推薦電影時，9種類型的電影都有機會被推薦到

| 變數處理-電影摘要斷詞

●目標

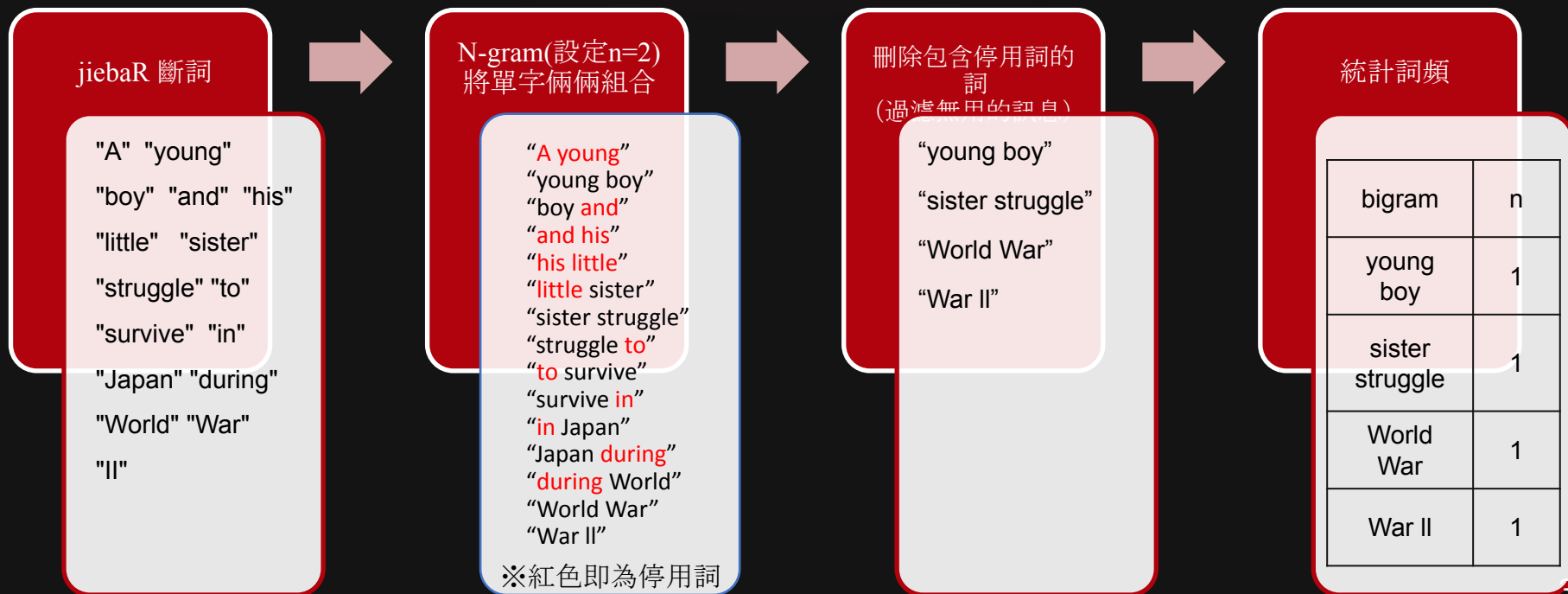
- 透過斷詞、計算詞頻，分析電影裡常出現的題材
- 整理出能包含9成以上電影的關鍵字，後續在推薦系統上供使用者勾選有興趣的內容

●斷詞處理方式

1. One-gram
2. Bi-gram
3. TF-IDF

變數處理-電影摘要斷詞

- Bi-gram處理流程 Ex. A young boy and his little sister struggle to survive in Japan during World War II.



| 變數處理-電影摘要斷詞

one-gram及bi-gram斷詞結果

※n為關鍵字在所有電影摘要中出現的次數

One-gram			Bi-gram		
1	onegram	n	1	bigram	n
2	life	111	2	world war	31
3	world	85	3	war ii	23
4	family	66	4	los angeles	11
5	war	66	5	york city	11
6	woman	65	6	true story	8
7	story	63	7	african american	6
8	love	61	8	boarding school	5
9	boy	46	9	civil war	5
10	father	45	10	police detective	5

- 從one-gram結果可看出有許多電影內的角色關係與家庭有關, 愛情相關的電影也是較常出現的題材
- 從bi-gram斷詞結果可以看出IMDb top1000的電影裡最常出現的戰爭相關的題材, 由真實故事改編的電影也佔有一定的比例, 另外電影中最常出現的地點可能為洛杉磯及紐約

| 變數處理-電影摘要斷詞

TF-IDF

- $tf = \frac{\text{單詞在一份文本中出現的次數}}{\text{該文本的總字數}}$
- $idf = \ln\left(\frac{\text{文本總數}}{\text{出現過該單詞的文本數量}}\right)$
- $tf-idf = tf * idf$
tf-idf分數越高, 代表該單詞越重要

【TF-IDF分數前10名關鍵字】

※n = 關鍵字在該電影摘要中出現的次數
total = 該電影摘要的總字數

movie	word	n	total	tf	idf	tf-idf
The Last Emperor	emperor	1	8	0.125	6.214	0.777
Double Indemnity	insurance	3	25	0.12	6.214	0.746
Witness for the Prosecution	surprise	2	17	0.118	6.214	0.731
The Last Emperor	china	1	8	0.125	5.808	0.726
Short Cuts	day	2	11	0.182	3.816	0.694
8½	harried	1	10	0.1	6.907	0.691
8½	retreats	1	10	0.1	6.907	0.691
Das Cabinet des Dr. Caligari	caligari	1	10	0.1	6.907	0.691
Das Cabinet des Dr. Caligari	cesare	1	10	0.1	6.907	0.691
Das Cabinet des Dr. Caligari	hypnotist	1	10	0.1	6.907	0.691

變數處理-電影摘要斷詞

電影推薦系統關鍵字挑選



- 依照顏色分類目的: 為避免使用推薦系統時, 用戶選擇顏色及關鍵字後沒有符合的電影跳出
- 各顏色關鍵字示意圖

black	white	red	orange	yellow	green	blue	purple
agent	agent	agent	agent	agent	boy	accident	agent
american	american	american	american	american	child	agent	aging
army	assassin	army	army	assassin	city	american	american
assassin	battle	assassin	assassin	battle	classroom	army	boy
battle	boy	battle	band	boy	community	assassin	child
boy	child	boy	battle	capture	conflict	battle	city
child	city	child	black	christmas	crime	boy	day
city	crime	city	boy	city	daughter	child	dead
crime	daughter	couple	british	crime	dead	city	death
daughter	day	crime	child	day	destruction	crime	delivery

| 變數處理-電影分級Certificate

- 目的

不同國家地區的电影分級制度各不相同，但實質上的含義差異並不大。

因此我們決定將不同的分級制度統一轉換成台灣的分級制度，以減少變數值的數量，讓之後的資料分析能更直接地呈現。

變數處理-電影分級Certificate

- 先觀察資料原有的變數值，共17種

電影分級	含義	出現次數	電影分級	含義	出現次數
"Approved"	允許發行。	11	"R"	美國限制級。	146
"Passed"	允許發行。	34	"TV-PG"	電視台訂立的保護級。	3
"U"	通用級(Universal)	234	"TV-14"	電視台訂立的輔導級。	1
"U/A"	1983年訂立的保護級。	1	"TV-MA"	電視台訂立的成人級。	1
"A"	1983年訂立的成人級(Adult)	197	"Unrated"	未分級。	1
"UA"	印度保護級。	175	"GP"	保護級。	2
"G"	美國大眾級。	12	"16"	日本輔導級。	1
"PG"	美國保護級。	37	NA	遺失值。	101
"PG-13"	美國特別輔導級。	43			

| 變數處理-電影分級Certificate

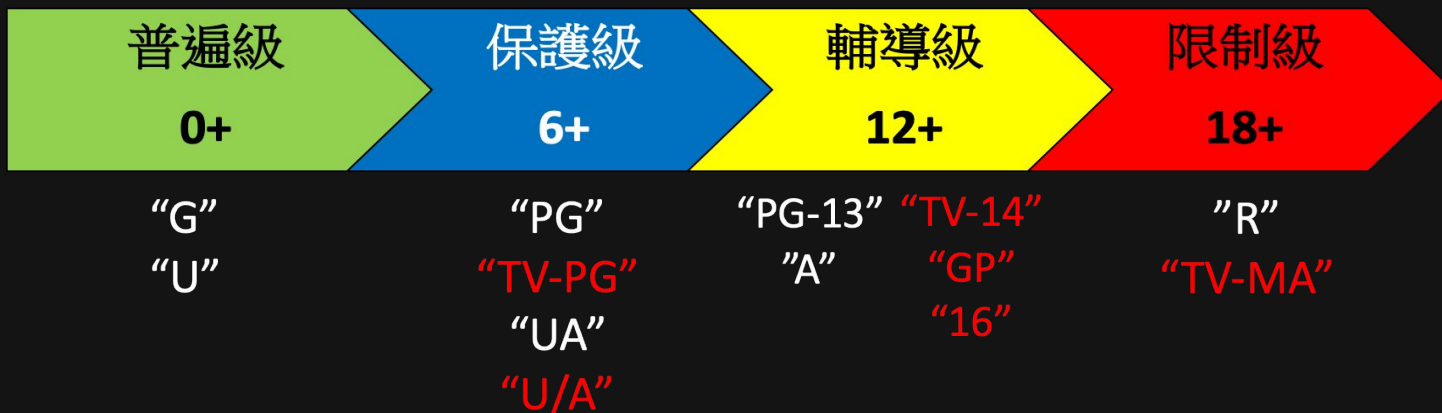
- 處理NA值,"Unrated" ,"Approved","Passed"

電影分級	含義	出現次數
"Approved"	1968年前發行的電影尚未分級，僅用能夠發行與否進行分類。	11
"Passed"	1968年前發行的電影尚未分級，僅用能夠發行與否進行分類。	34
NA	遺失值。	101
"Unrated"	未分級。	1

電影分級中共有147筆尚未分級的資料，但因為電影的分級資料都是既有的資料，所以我們根據網路上的相關資料填補遺失值。

變數處理-電影分級Certificate

- 將不同的電影分級分類



其中紅字標注的屬於樣本數小的冷門分類(不超過五件), 所以我們將屬於這些分類的電影依照台灣代理商的分級自行重新歸類。

| 變數處理-電影分級Certificate

- 冷門分類特別備註

“TV-PG”:[350]海灘的那一天(普通級)、[461]瘦子(輔導級)、
[877]隱形人(輔導級)

“TV-14”:[92]7號房的禮物(保護級)

“TV-MA”:[199]佈局(輔導級)

“GP”:[228]午後七點零七分(輔導級)、[992]戰略大作戰(輔導級)

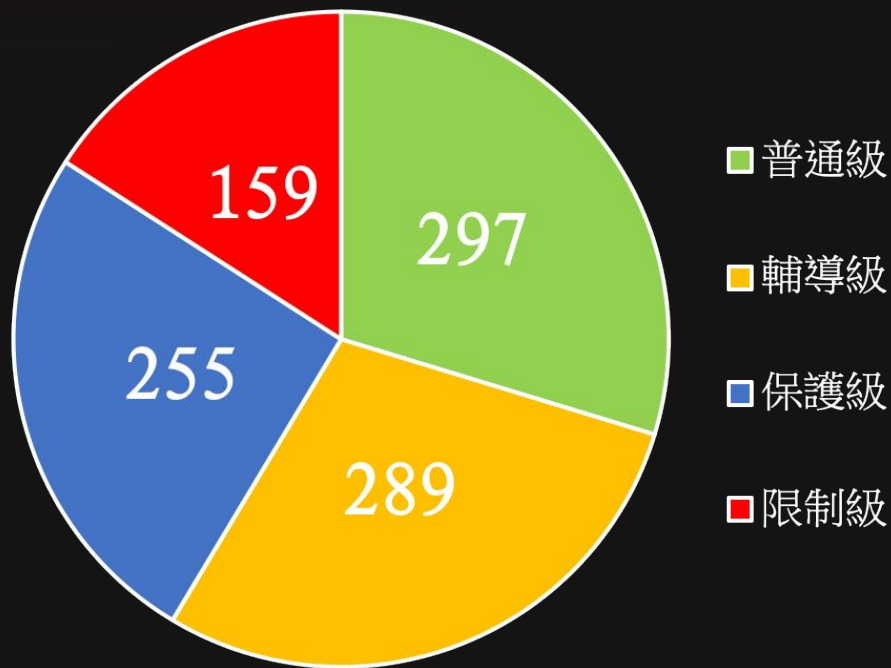
“U/A”:[879]人肉搜索(輔導級)

“16”:[198]聲之形(普遍級)

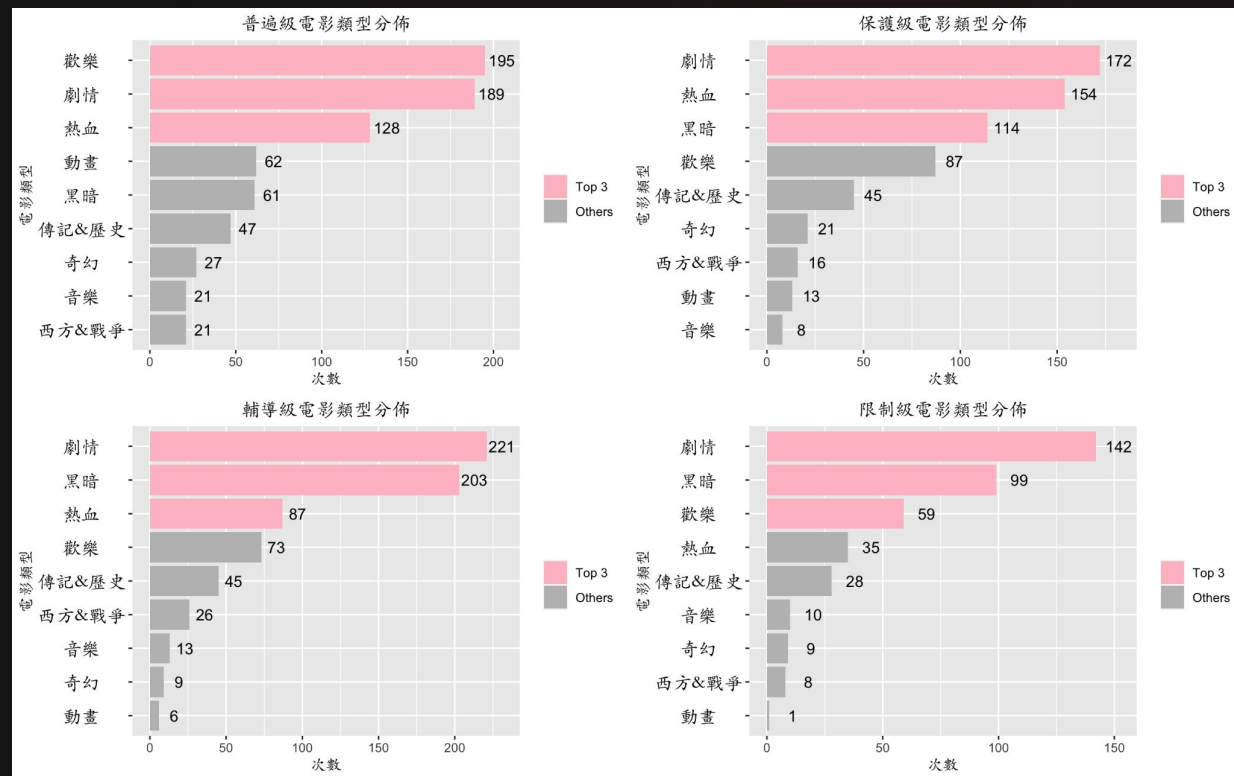
| 變數處理-電影分級Certificate

● 結果

電影分級	出現次數
普通級	297
保護級	255
輔導級	289
限制級	159



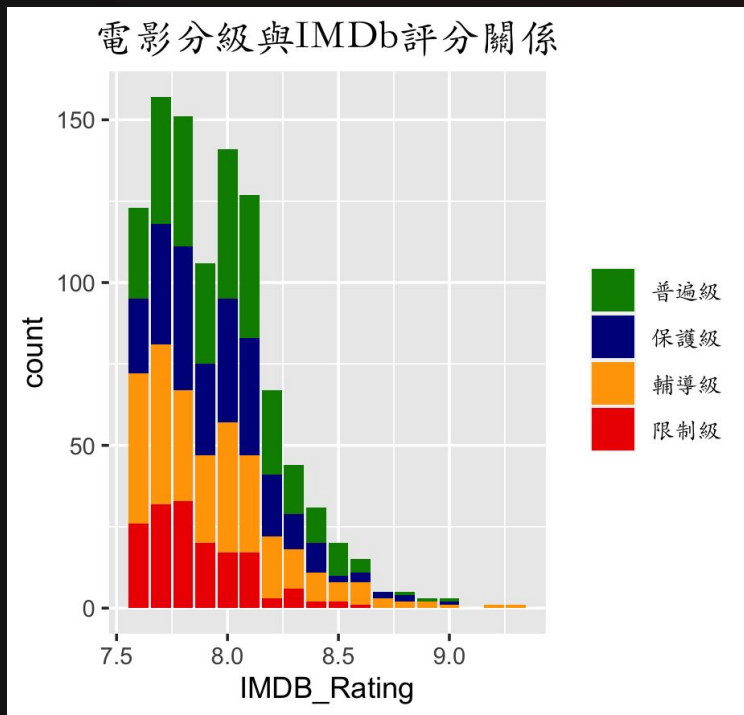
I 電影分級與電影類型的關係分佈



1. 劇情類型電影因本身樣本數多，所以在各分級佔的比例也高。普遍級中的歡樂片是唯一超越劇情片數量的類型。

2. 其中動畫片的數量隨著分級的提升而減少，因為動畫片多是为適合幼童觀看而製作。反之，黑暗類型的電影則多集中在保護級以上。

I 電影分級與IMDb評分的關係分佈



限制級電影明顯只分佈在8.6分以下，使電影被歸類在限制級的元素或許並不合觀影者的胃口。

反之，高分區段占比最高的是元素控制得宜的輔導級電影。



04

成果展示



系統介面

Movie Recommender

操作說明

推薦系統

顏色

其他色彩

年齡

18歲以上

篩選依據

電影類型

電影關鍵字

電影類型

熱血

開始推薦

資料探索

IMDb搜尋





開發人員

Top 1000 Movies by IMDb Rating

推薦電影清單

Show 10 entries

Search:

	海報	電影名稱	年份	分級	種類	簡介	IMDb Rating
1		Interstellar	2014	保護級	熱血, 劇情, 奇幻	A team of explorers travel through a wormhole in space in an attempt to ensure humanity's survival.	8.6
2		Song of the Sea	2014	保護級	熱血, 劇情, 動畫	Ben, a young Irish boy, and his little sister Saoirse, a girl who can turn into a seal, go on an adventure to free the fairies and save the spirit world.	8.1
3		Stand by Me	1986	普通級	熱血, 劇情	After the death of one of his friends, a writer recounts a childhood journey with his friends to find the body of a missing boy.	8.1
4		The Bourne Ultimatum	2007	保護級	熱血, 黑暗	Jason Bourne dodges a ruthless C.I.A. official and his Agents from a new assassination program while searching for the origins of his life as a trained killer.	8

SCAN ME



| 未來展望

1. 可加入更多變數來優化推薦體驗，如提醒電影時長過長、備註導演或演員為奧斯卡得主等
2. 針對顏色挑選欄位進行修正，避免選項中有過於相似的顏色同時出現
3. 提升操作介面流暢度，解決目前切換頁籤無法從標題處顯示分頁內容的問題
4. 新增手機使用介面，讓使用者能更方便的操作推薦系統，解決目前只能使用電腦操作的問題



THANKS!

Q&A



專案分配

單變數分析: 展德、宥芯

色彩探索: 貞莉

電影類型: 玫儒

摘要斷詞: 詩涵

電影分級: 柏辰

推薦系統: 宥芯、貞莉、玫儒、柏辰

報告: 展德、詩涵