職業表現和社會狀態：

2000 年代初 MovieLens 評論分析

一、研究方法與架構

此次研究資料為西元 2000 年代 MovieLens 的評論資料，其中包括了評分、評論者職業、評論電影種類等等相關資訊。

我想以不同職業為主要分析單位，主要以電影品味、評論次數及評論高低，去推斷當時每個職業的表現和社會的狀態．

二、資料探討與整理

先把所有表接起來，並把職業為“其他或未指定(Occupation = 0)”的列刪除，提升分析的精度。

新增欄位[Occupation]，設為職業的繁體中文翻譯

新增欄位[Datetime]，設為 Timstamp 的西元格式

資料最早時間: 2000-04-25 23:05:32

資料最晚時間: 2003-02-28 17:49:50

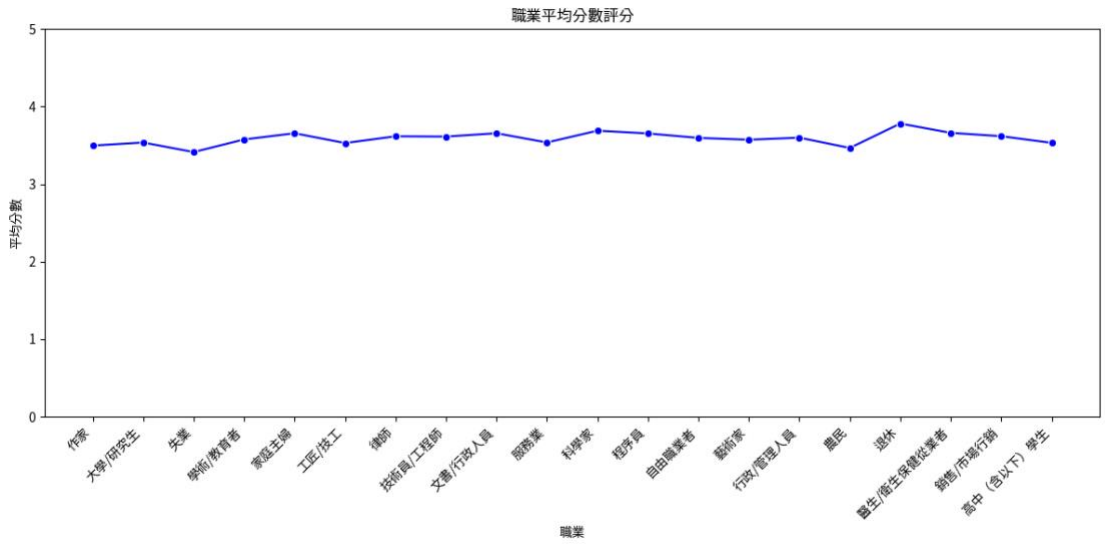
以上整理可以得知此次資料全為 2000 年初所產生。

```python
# 合併ratings, users
merged_data = pd.merge(ratings_df, users_df, on='UserID')
# 再合併movies
merged_data = pd.merge(merged_data, movies_df, on='MovieID')
# 把其他職業消去
merged_data = merged_data[merged_data['Occupation'] != 0]
# 翻譯表
unique_genres = merged_data['Genres'].unique()

merged_data.head()
# 多加一個欄位，職業轉成繁體中文
merged_data['Occupation_TW'] = merged_data['Occupation'].map(occupation_mapping)
# 多加一個欄位，Datetime翻成西元格式
merged_data['Datetime'] = pd.to_datetime(merged_data['Timestamp'], unit='s')

#找timstamp最早和最晚，轉格式
start_time = pd.to_datetime(merged_data['Timestamp'], unit='s').min()
end_time = pd.to_datetime(merged_data['Timestamp'], unit='s').max()

start_datetime = pd.to_datetime(merged_data['Datetime']).min()
end_datetime = pd.to_datetime(merged_data['Datetime']).max()

print(f"資料最早時間: {start_datetime}")
print(f"資料最晚時間: {end_datetime}")
merged_data
```

三、圖表分析

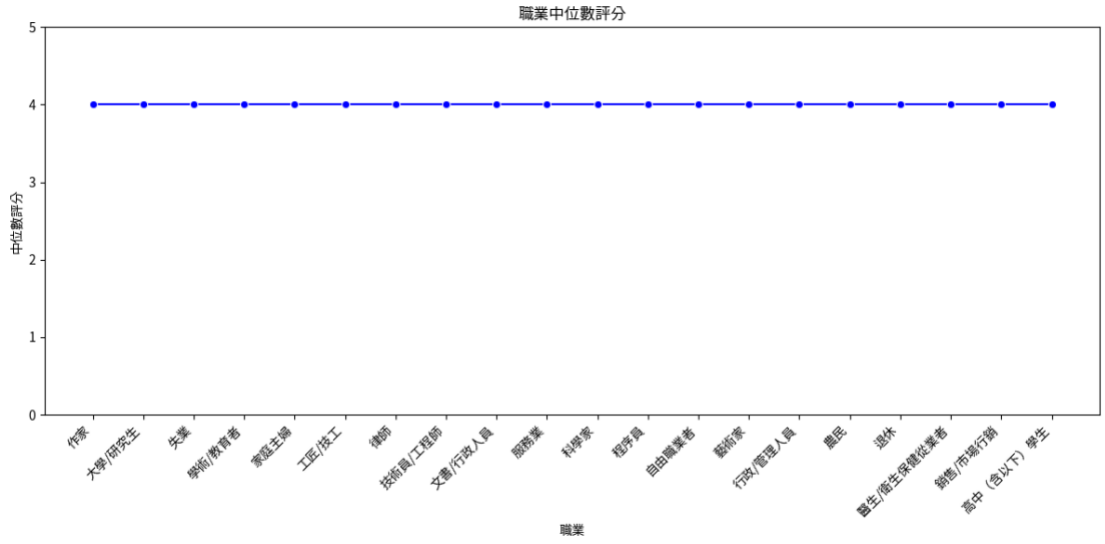
（一）不同職業的平均分數

每個職業的平均分數相差不遠，都在 3.5 分上下



（二）不同職業的評分中位數

所有職業的中位數皆為 4 分

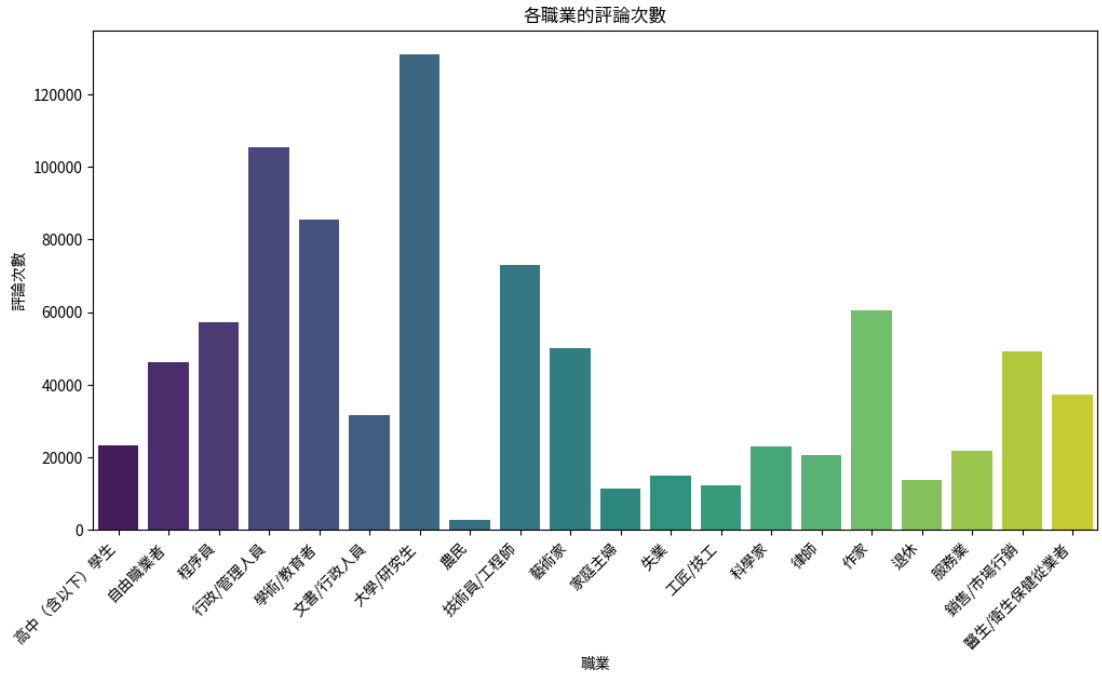


小結：

所有職業的群體平均分數皆為 3 分以上，可以得知大多數人都願意給不低的分數，照理來說一定會有難看的電影，但是平均分數都仍在及格之上，且中位數皆為 4 分代表這些職業有一半以上的人的評分都為 4 分。

我認為這時段的人在普遍友善，又或著說網路剛興起，大家都還懵懂無知，給予的評論都是積極正向的，還不像現在的網路評論環境一樣有話直說，甚至會出現平均分數 1 分的情況。

（三）不同職業的評論次數



小結：

此圖可以看到評論次數最高為大學生 / 研究生及行政管理人員等，可以看到商管（文）組、工程師等辦公室工作族群評論次數多次，我推斷是因為當時正在從傳統產業轉型，多了很多辦公室、資訊及文書等新型工作，這些工作也大多需要使用網路，也算是前幾批會使用網路的人，而此次資料為網站搜集，當時使用網路的應該也多為年輕人，相對的，在網路上評論的次數也會較多，

而此圖也直接地表示出一些一級產業（農民）、傳統產業，應多為年長者，很少在用網路，更不可能去網路上評論電影，導致評論次數較少。

資料調整：

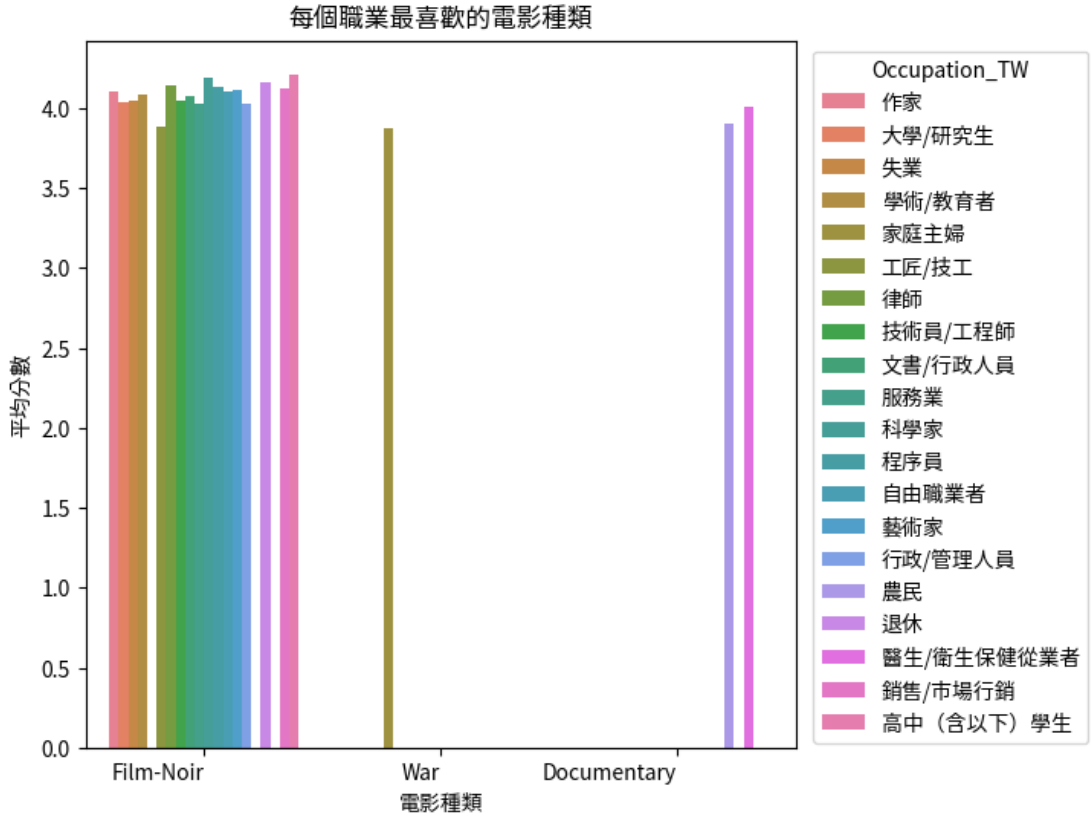
這次資料有許多 row 的[Genres]（電影種類）有一種以上，為了更精準的判斷種類，我把有多個種類的 row 拆成不同 row，同時 row 數也變成 1829937 rows，例：

1000207	5851	3607	5	957756608	F	18	20	55410	One Little Indian (1973)	Comedy|Drama|Western	作家

轉變成

1829933	5851	3607	5	957756608	F	18	20	55410	One Little Indian (1973)	Comedy	作家
1829934	5851	3607	5	957756608	F	18	20	55410	One Little Indian (1973)	Drama	作家
1829935	5851	3607	5	957756608	F	18	20	55410	One Little Indian (1973)	Western	作家

（四）每個職業的第一喜歡（平均評分高）電影種類



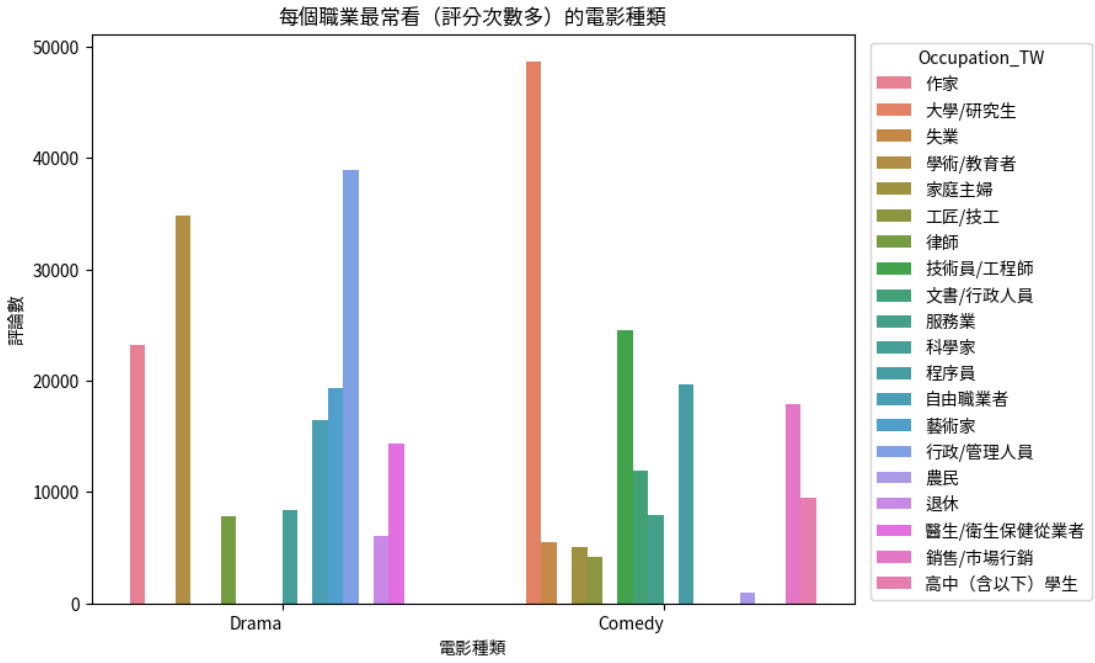
小結：

大多數職業都很喜歡看 Flim-Noir（黑色電影），其中大多風格晦暗、悲觀且憤世嫉俗，我認為在這個年代，開始興起故事架構複雜值得反思的電影，人們不再只挑選善惡分明的敘事結構、英雄打壞蛋的普通電影，反而開始踴躍地去欣賞位於底層社會的寫實電影，人們喜歡黑色電影也反映出 2000 年代初的社會環境，包括了金融風暴和道德意識抬頭，人們更能和黑色電影共情．

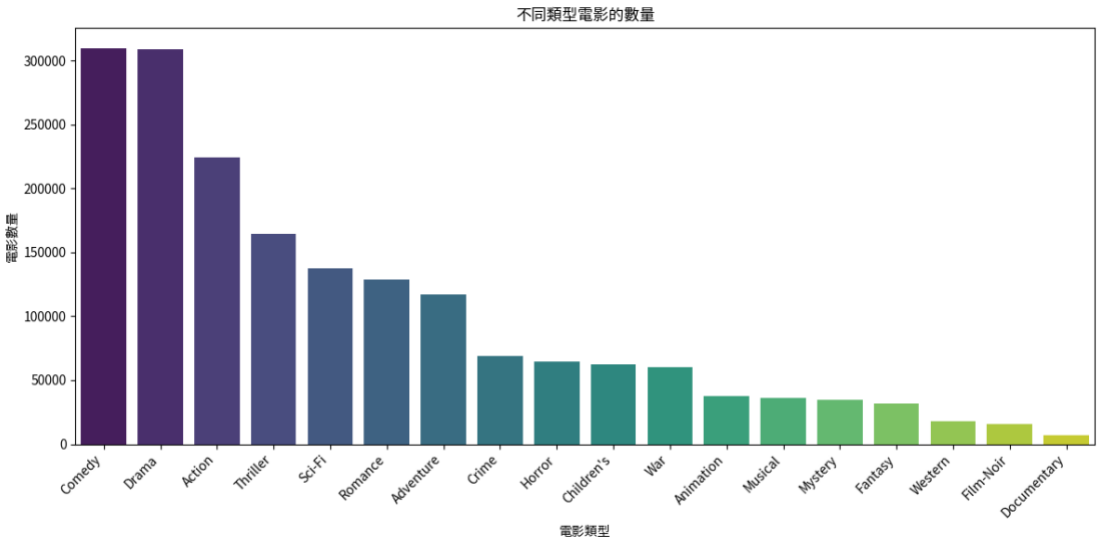
這張表讓我最意外的是家庭主婦竟然最喜歡看 War（戰爭片），我認為應該是因為戰爭片大多感人、悲傷及不捨，特別觸動家庭主婦（絕大部分是女生），很容易與死亡、戰爭等悲慘情節共情．

而農民和醫生最喜歡看紀錄片，我認為還算合理，紀錄片就像論文，利用拍攝影像來向觀眾輸出觀點，讓觀眾反思其中的知識，農民和醫生的專業知識是能夠一直進步的，並且農民的工作是可以自行改良的，他們時常觀看紀錄片來學習．

（五）不同職業最常看（加總評論 row 數最多）電影種類



（六）評論最多的電影類型



小結：

普遍所有職業都集中在 Drama（劇情片）和 Comedy（喜劇），我認

為這兩種是最長青、最老少咸宜的電影類型，大部分的電影也都為這兩個類型，且這次資料被評論的電影種類也是 Drama 及 Comedy 類型為前二多，綜述以上原因，Drama（劇情片）和 Comedy（喜劇）之所以最多人看，主要是歸咎於這兩種類型的電影數量基數大，發行數量多，導致大多數人常看．

以下為高低評分比例：

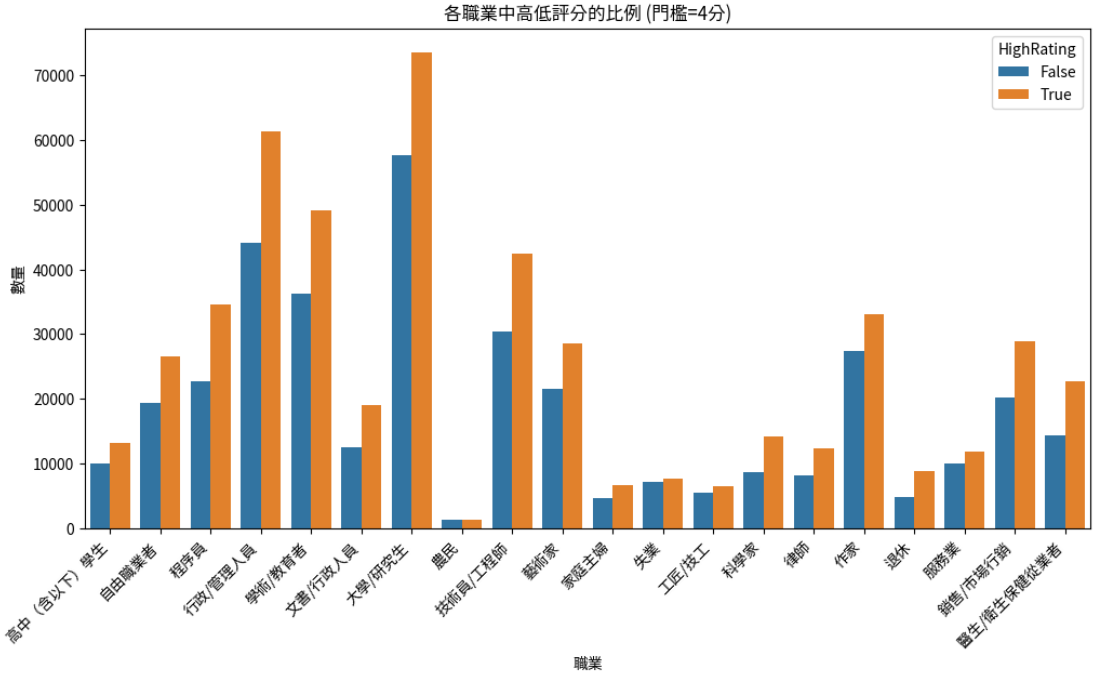
以評分 4 分作為門檻，

(Rating >= 4 分) = 高評分

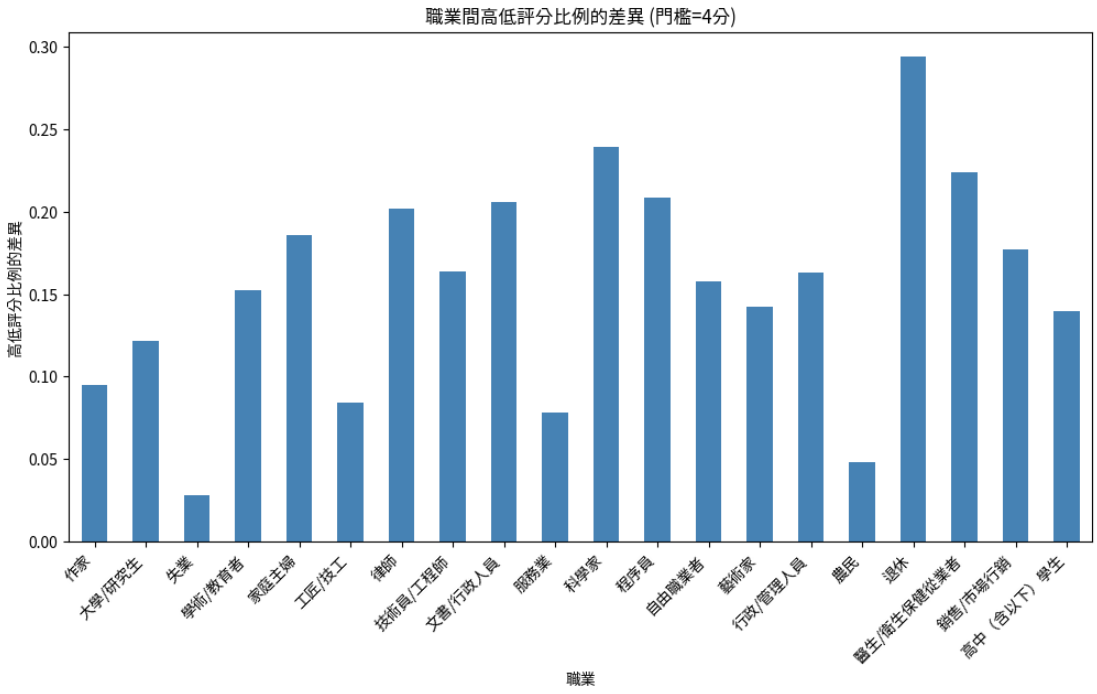
(Rating < 4 分) = 低評分

這張圖表示了不同職業高評分與低評分的比例

（七）各職業中高低評分的比例



（八）不同職業高比例與低比例數量差異圖

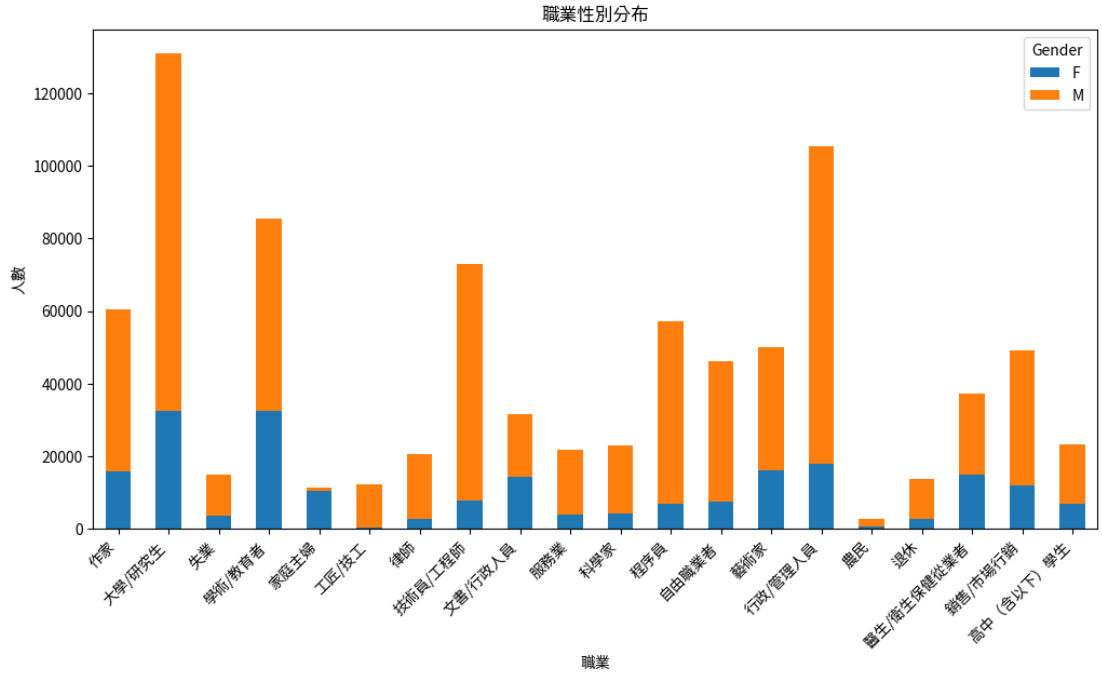


小結：

可以看出比例相差最大的為退休人士，我認為是年齡稍大的長者在批評事物的時候，不太在乎其他人的想法，不像年輕人在批評時，會思考同儕的想法是否和自己一樣，使得容易給出中規中矩正常的分數 (表中可以看出學生的差異量偏中間，不太大)，而年長者毫無顧慮，加上自身經驗較多，已經有自己的審美和道德觀，理解自己喜歡什麼不喜歡什麼，更容易給出主觀的評分，而其他偏高差異量的職業大多為科學家、醫生、工程師等等高專業職業，我認為他們的邏輯思維較為縝密，比較以理組思維的去理解所有事情，也清楚知道自己喜不喜歡一部電影，做出忠於自己的評分．

綜合以上論述，我認為在理不理解自己喜歡些什麼類型的電影和自身的人生經歷，對於做出主觀的評論是非常重要的．

（九）不同職業的男女佔比



小結：

這張圖表示了不同職業的男女佔比，可以明顯的看出來大多數的職業仍以男性為主，甚至壓倒性地佔據總比例，家庭主婦也是壓倒性地以女性為主，可以看出當時的社會狀態仍是傳統的男主外、女主內，上班的大多數還是男生，如果是現在的資料做分析，女性的佔比應該會明顯的增長許多．

四、研究總結

在 2000 年代初的 MovieLens 評論資料中，透過對不同職業的分析，探討了當時的社會狀態和職業表現。以下為結

論：

1. 整體評分趨勢：
    - 平均分數：每個職業的平均分數相差不遠，都在 3.5 分上下。
    - 中位數：所有職業的中位數皆為 4 分。結果顯示，大多數人都願意給予不低的分數，且有一半以上的人的評分都在 4 分以上。
2. 評論次數分析：
    - 大學生 / 研究生及行政管理人員評論次數最高，顯示網路使用者以辦公室工作族群為主，這可能與當時網路使用者以年輕人居多有關。
3. 最常看的電影種類：
    - 大多數職業都喜歡看 Drama（劇情片）和 Comedy（喜劇），這兩種類型的電影數量基數大，發行數量多，因此受眾廣泛。
4. Flim-Noir（黑色電影）的受歡迎：
    - 大多數職業都喜歡看 Flim-Noir，這可能反映出 2000 年代初社會環境的一些黑暗、悲觀和反思特質。
5. 高低評分比例分析：

- 退休人士給出的高低評分比例相差最大，顯示年長者在評分時較為主觀，不太受同儕影響。
- 高專業職業（科學家、醫生、工程師等）給出的高評分比例較高，可能因其邏輯思維較為縝密。

6. 性別分佈：
    - 大多數職業仍以男性為主，反映出當時社會仍以男性為主導的特徵。

總的來說，2000 年代初的社會環境較為保守，但透過評論資料分析，我們也看到人們對於不同電影類型的喜好，以及職業在評分時的主觀性。

當下年代也算是網路剛興起的年代，MovieLens 算是線上影評的元老，這份資料提供了大量數據去探索 20 年前的電影品味，在做這次研究的時候一直想要獲取影評的文字內容，或許能夠使用現代的大型語言模型去解析他們的想法，讓這份研究更完整．