# Case Study 2: Analyzing data from MovieLens

Team 9

Tingting Ma   Jinyan Lyu   Jiani Gao
Tianhao Guo   Mo Cheng

# Background & Motivation

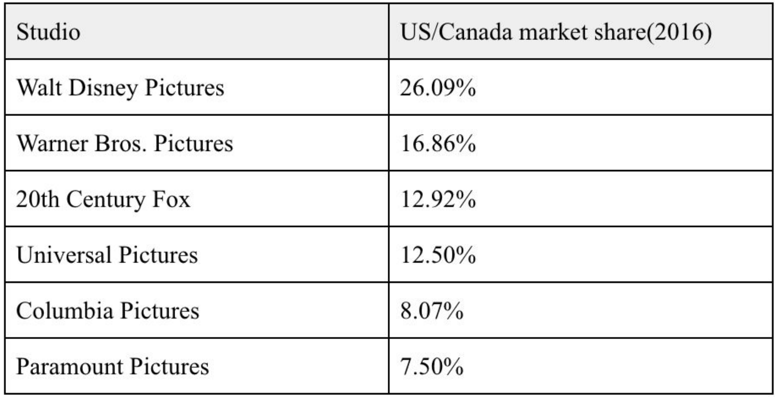
Business Intelligence

Movie Industry -- one of the most important and biggest industries in the world

BIG 6  -- 6 major studios

Production Companies – The Big 6

20th CENTURY FOX
A NEWS CORPORATION COMPANY
UNIVERSAL
WARNER BROS. PICTURES
Paramount
A VIACOM COMPANY
Walt Disney
PICTURES
COLUMBIA



| Studio | US/Canada market share(2016) |
|---|---|
| Walt Disney Pictures | 26.09% |
| Warner Bros. Pictures | 16.86% |
| 20th Century Fox | 12.92% |
| Universal Pictures | 12.50% |
| Columbia Pictures | 8.07% |
| Paramount Pictures | 7.50% |

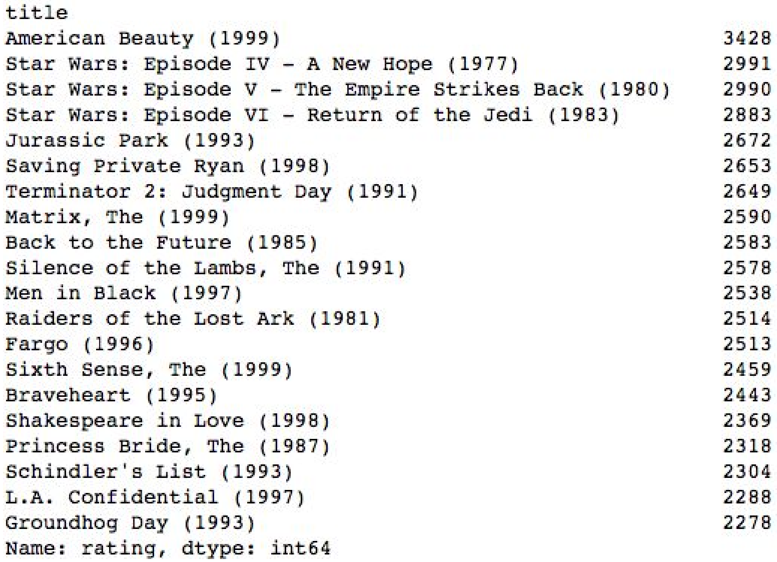
# Background & Motivation

MovieLens  1M Data Set

Python, Pandas, NumPy and Matplotlib

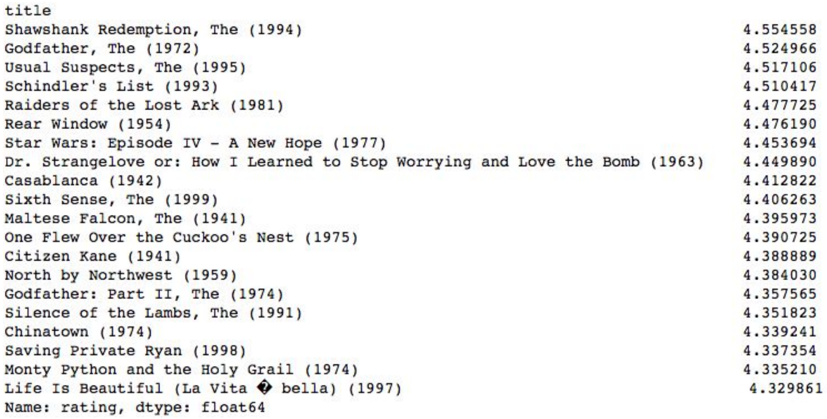
# Basic details of data

top 20 movies by rating times



```
title
American Beauty (1999)                                      3428
Star Wars: Episode IV - A New Hope (1977)                   2991
Star Wars: Episode V - The Empire Strikes Back (1980)       2990
Star Wars: Episode VI - Return of the Jedi (1983)           2883
Jurassic Park (1993)                                        2672
Saving Private Ryan (1998)                                  2653
Terminator 2: Judgment Day (1991)                           2649
Matrix, The (1999)                                          2590
Back to the Future (1985)                                   2583
Silence of the Lambs, The (1991)                            2578
Men in Black (1997)                                         2538
Raiders of the Lost Ark (1981)                              2514
Fargo (1996)                                                2513
Sixth Sense, The (1999)                                     2459
Braveheart (1995)                                           2443
Shakespeare in Love (1998)                                  2369
Princess Bride, The (1987)                                  2318
Schindler's List (1993)                                     2304
L.A. Confidential (1997)                                    2288
Groundhog Day (1993)                                        2278
Name: rating, dtype: int64
```

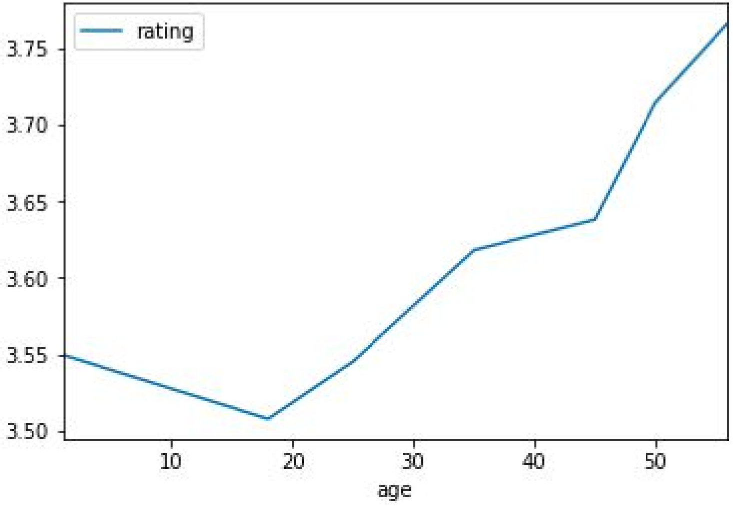
top 20 movies by mean ratings with rating times no less than 1000

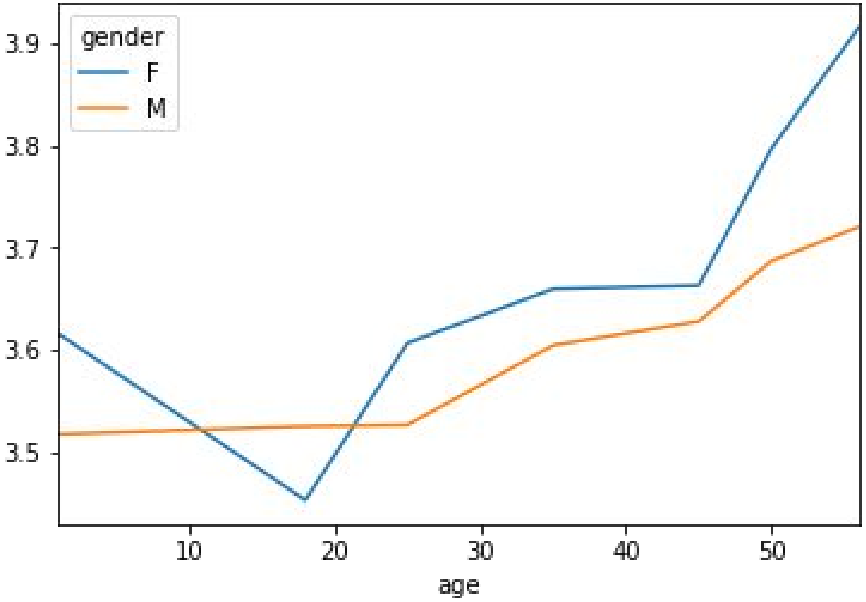


```
title
Shawshank Redemption, The (1994)                                             4.554558
Godfather, The (1972)                                                        4.524966
Usual Suspects, The (1995)                                                   4.517106
Schindler's List (1993)                                                      4.510417
Raiders of the Lost Ark (1981)                                               4.477725
Rear Window (1954)                                                           4.476190
Star Wars: Episode IV - A New Hope (1977)                                    4.453694
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963)  4.449890
Casablanca (1942)                                                            4.412822
Sixth Sense, The (1999)                                                      4.406263
Maltese Falcon, The (1941)                                                   4.395973
One Flew Over the Cuckoo's Nest (1975)                                       4.390725
Citizen Kane (1941)                                                          4.388889
North by Northwest (1959)                                                    4.384030
Godfather: Part II, The (1974)                                               4.357565
Silence of the Lambs, The (1991)                                             4.351823
Chinatown (1974)                                                             4.339241
Saving Private Ryan (1998)                                                   4.337354
Monty Python and the Holy Grail (1974)                                       4.335210
Life Is Beautiful (La Vita � bella) (1997)                                    4.329861
Name: rating, dtype: float64
```

# Conjecture

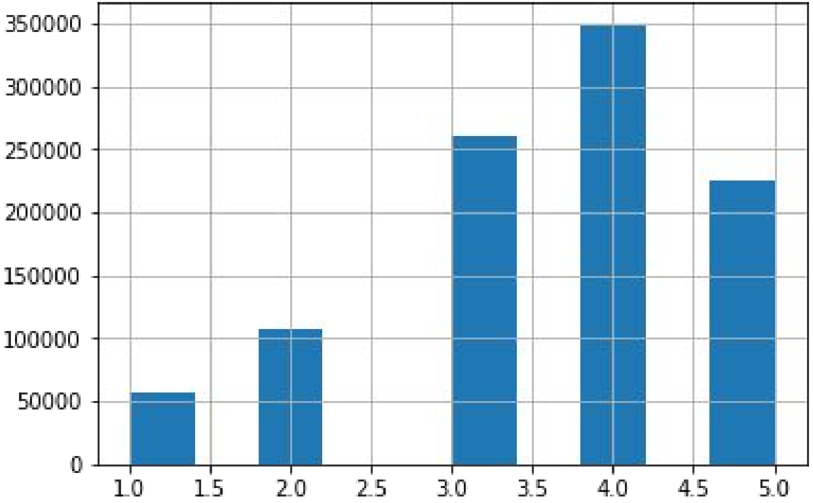
Elderly people are the easiest to please since they are more easygoing than young people, and female are easier to please than male since they are more emotional.



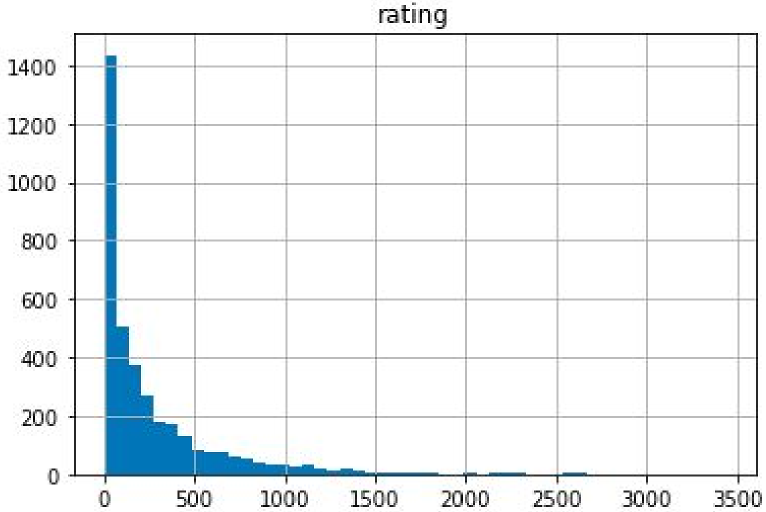


# Investigation to histograms

histogram of the ratings of all movies

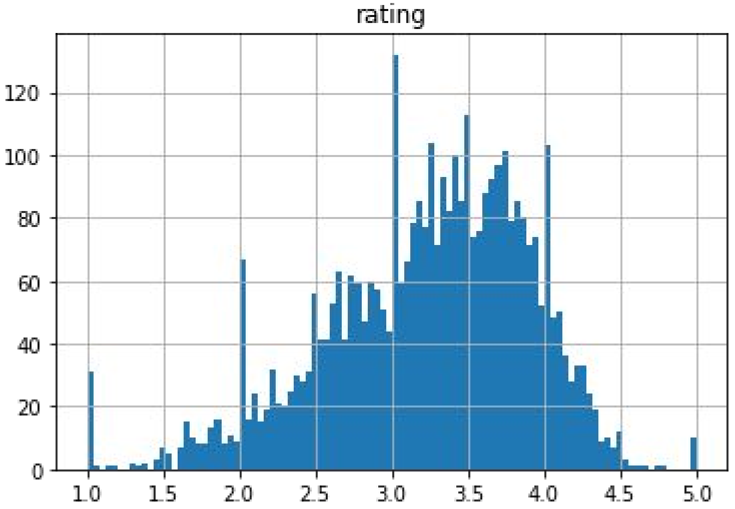


histogram of the number of ratings each movie received

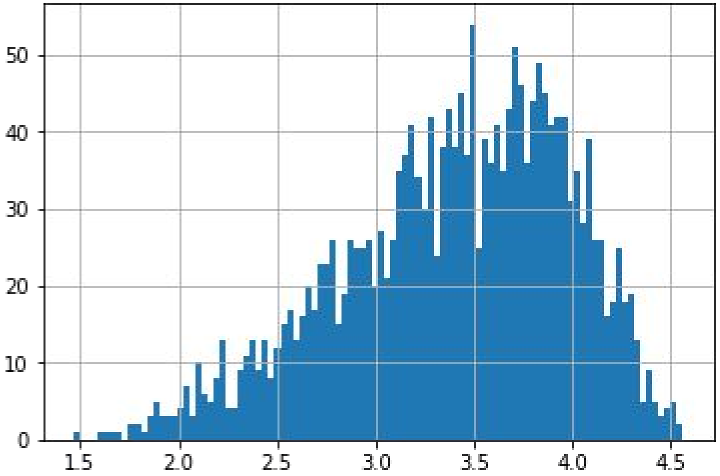


# Investigation to histograms

histogram of the average rating for each movie

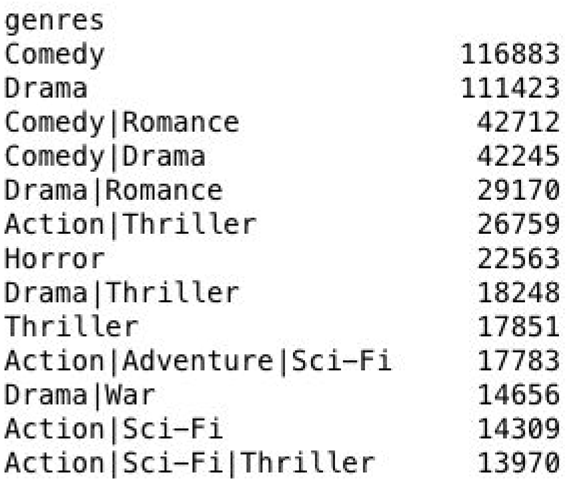


histogram of the average rating for movies which are rated more than 100 times



# Conjecture

Comedy movies have most rating numbers since people want to have fun when watching movies.

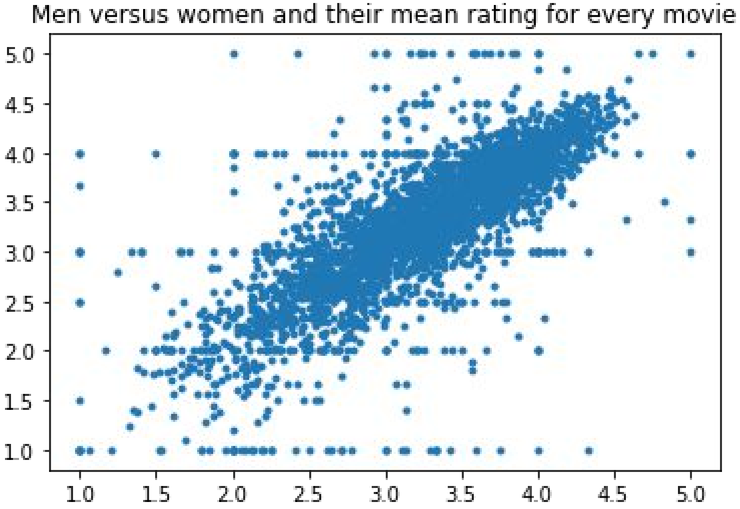


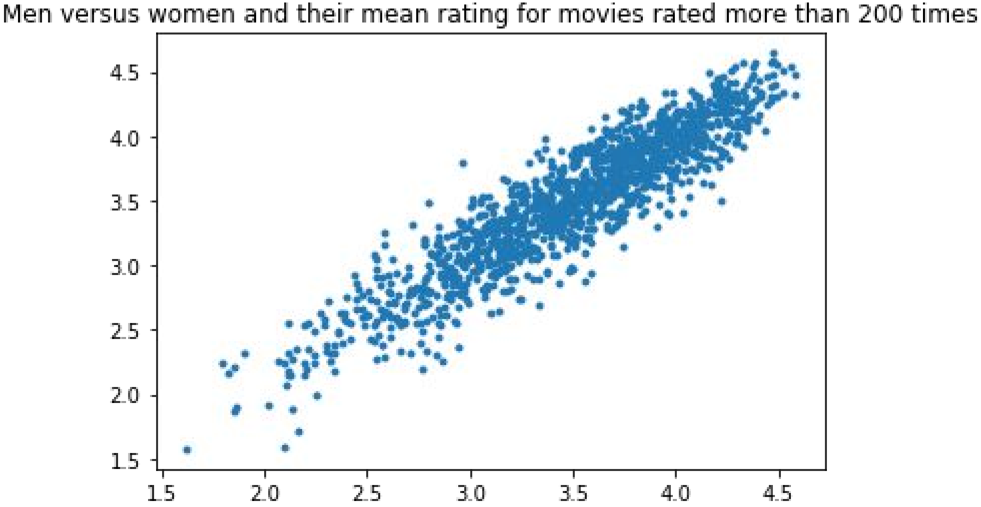
```
genres
Comedy                        116883
Drama                         111423
Comedy|Romance                 42712
Comedy|Drama                   42245
Drama|Romance                  29170
Action|Thriller                26759
Horror                         22563
Drama|Thriller                 18248
Thriller                       17851
Action|Adventure|Sci-Fi        17783
Drama|War                      14656
Action|Sci-Fi                  14309
Action|Sci-Fi|Thriller         13970
```

# Correlation: Men versus women

scatter plot of men versus women and their mean rating for every movie. corr = 0.76319

scatter plot of men versus women and their mean rating for movies rated more than 200 times. corr = 0.918361





# Conjecture

Men and women in the same occupation has the same preference to movies

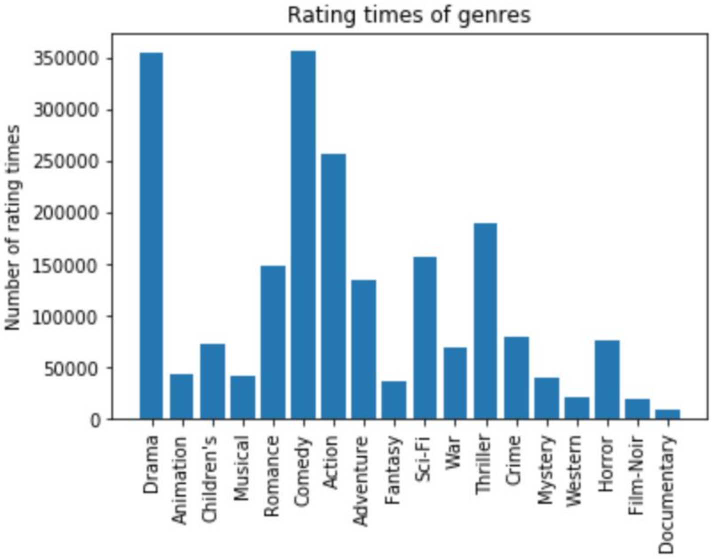
| Occupation | Correlation coefficient | Occupation | Correlation coefficient |
|---|---|---|---|
| academic/educator | 0.636357634705 | lawyer | 0.394055882261 |
| artist | 0.472413764133 | programmer | 0.450083759757 |
| clerical/admin | 0.438775296571 | retired | 0.294298338909 |
| college/grad student | 0.572648438461 | sales/marketing | 0.533524348122 |
| customer service | 0.329810126208 | scientist | 0.479621348720 |
| doctor/health care | 0.518478827401 | self-employed | 0.468766904706 |
| executive/managerial | 0.572695642366 | technician/engineer | 0.579449959376 |
| farmer | 0.275236368043 | tradesman/craftsman | 0.276750813049 |
| homemaker | 0.276577331069 | unemployed | 0.408121713176 |
| K-12 student | 0.330525786667 | writer | 0.606829865489 |

# Business Intelligence

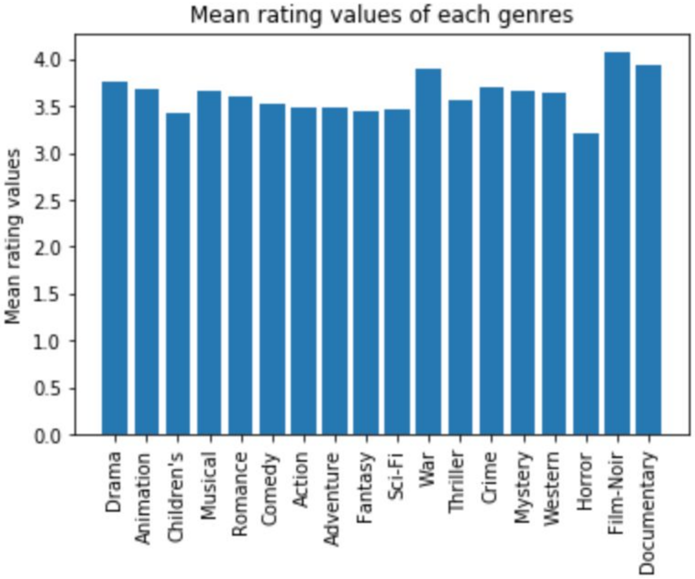
What content is loved by audience?

To appeal as many audience as possible, how do we advertise this movie?

# Choice of content

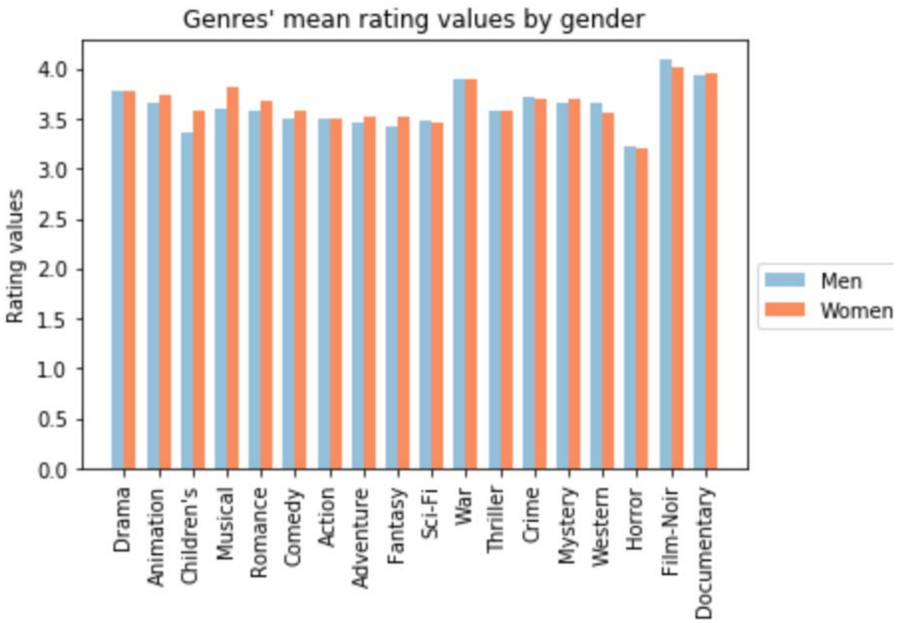


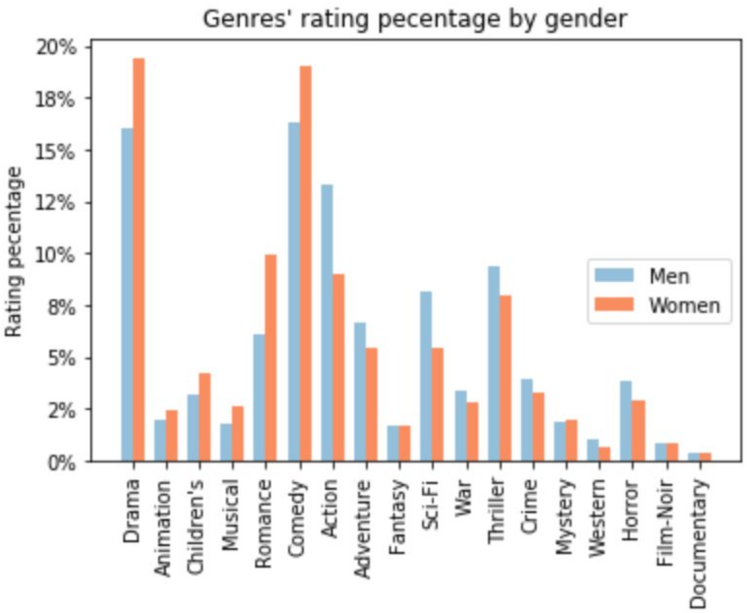
Rating times of genres



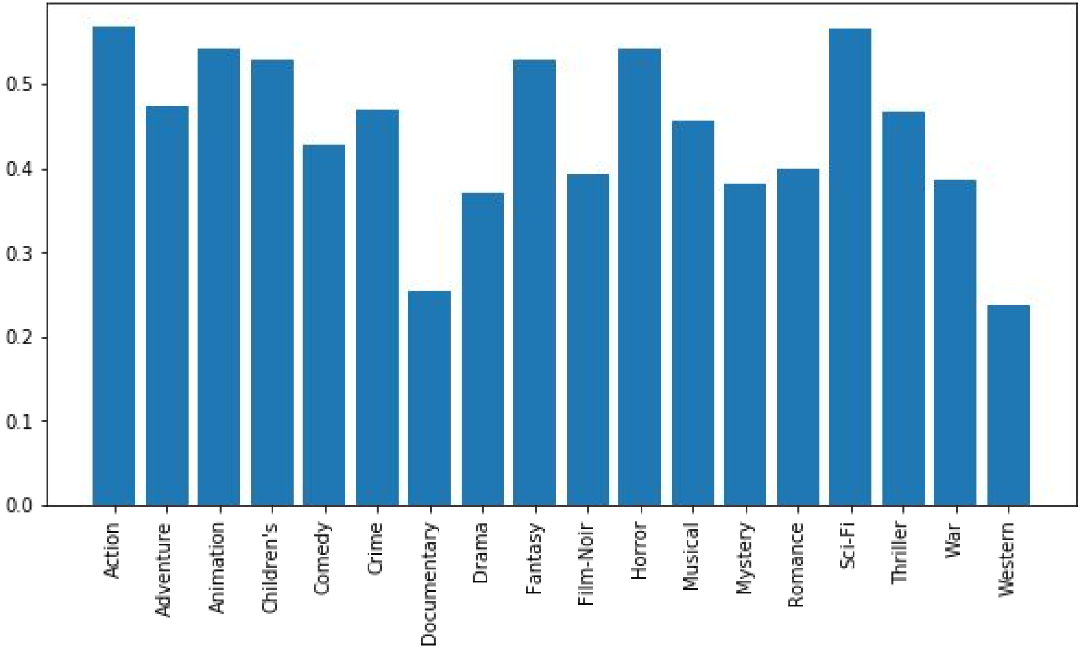
Mean rating values of each genres

# Choice of content





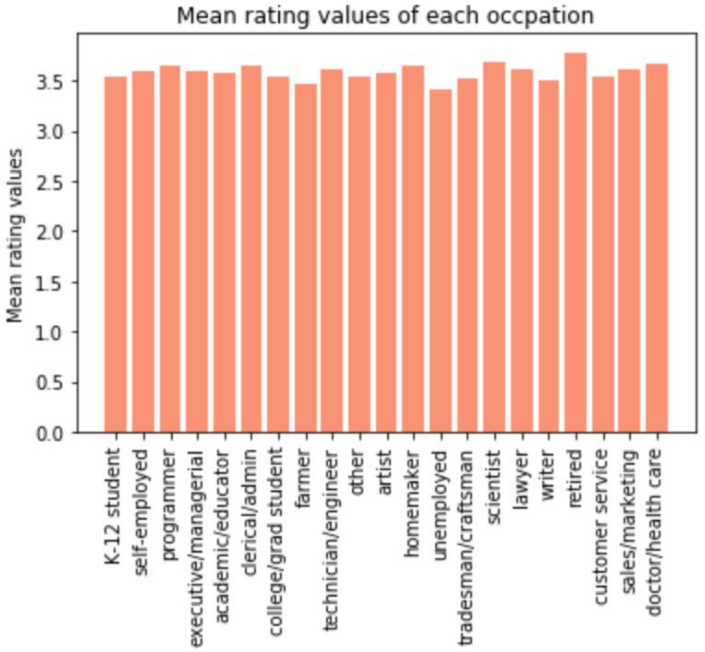
# Choice of content

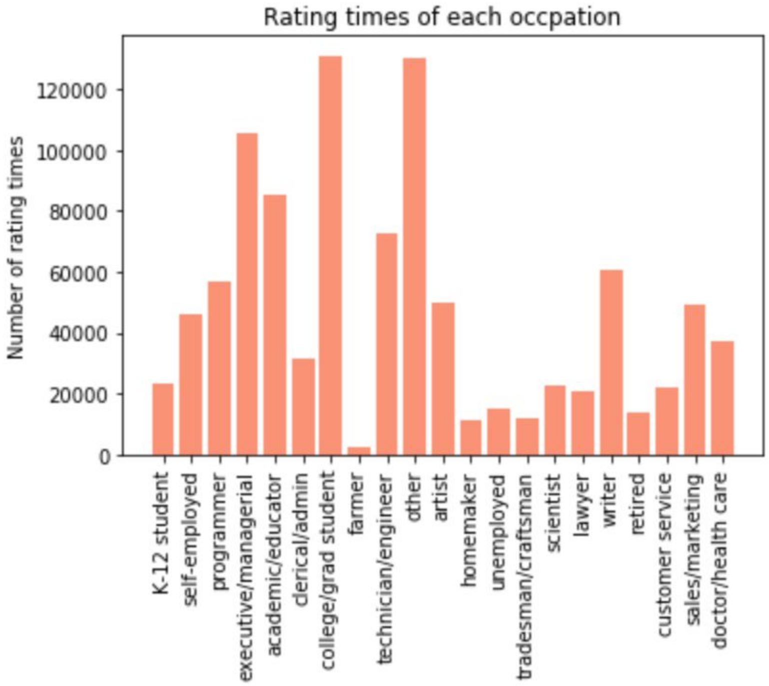


# Choice of content

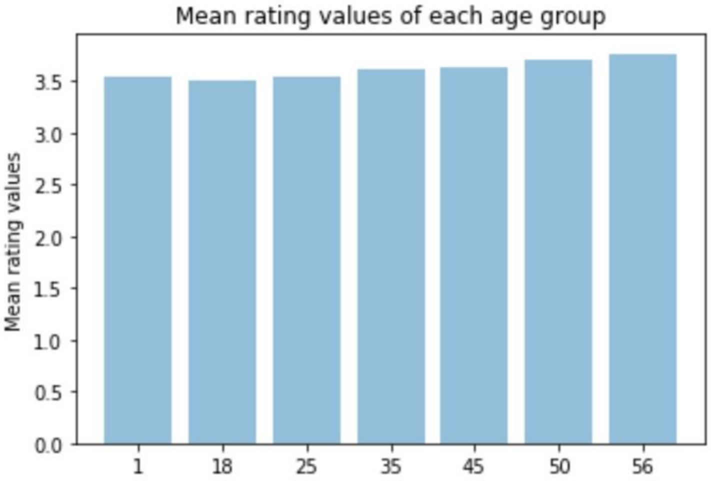
Choosing to make drama, comedy and action movies are more likely to be successful.

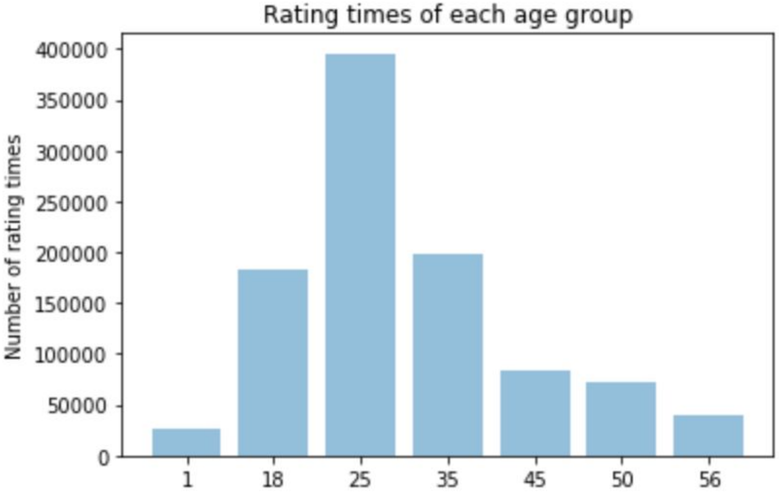
# Advertising strategy





# Advertising strategy





# Advertising strategy

We could send movie poster to college, graduate schools. Also we could give discount to senior and retired audience, since they rate high.

# Questions?