Case Study 2

Analyzing data from MovieLens

Team 9

Tingting Ma
Jiani Gao
Jinyan Lyu
Tianhao Guo
Mo Cheng

I. Motivation and Background

Business Intelligence is a user-oriented process of gathering, exploring, interpreting and analyzing of data, which leads to the streamlining and rationalization of the decision-making process. Those systems support managers in business decision-making in order to create economic value growth of an enterprise.

Movie industry is one of the most important and biggest industry in the world. The worldwide theatrical market had a box office of US$38.6 billion in 2016. In United States, there are six companies whose various film production and distribution subsidiaries collectively command approximately 80 to 85 percent of U.S. and Canadian box office revenue. Besides these six major studios, there are several major independent studios who are responsible for actual movie production while big six are primarily backers and distributors of movies.

As we see in the table below, it's highly competitive for the big six, and for every studio, business intelligence is quite important as it provides insights for them to make decisions on what movie to make and how to distribute the movie in order to get the most profit and market share.

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%

Table 1: major studios' market share

Movie review is a key factor that impacts the box office of a movie. Many people will look at the IMDb score and read other people's reviews when choosing what movie to see. So we choose the MovieLens dataset analyze the movie rating data to support the business intelligence process for a movie studio. The MovieLens 1M dataset provides us 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users. User information includes id, gender, age, occupation and zip-code. This detailed demographic information helps us get more useful results through the analysis.

We choose to use Python, Pandas, NumPy and Matplotlib to analyze and visualize the data. Pandas offers us the data structures and operations to read and manipulate the data. Using the statistical tools enables us to transform, merge and analyze the data. Matplotlib helps us translate the results to charts that can be read by people with different backgrounds.

II. Basic details of data

At the beginning of the analysis, we collected some basic information of the 1 million dataset. The result shows that only 29 movies have average rating over 4.5 overall. Moreover, 29 movies have average rating over 4.5 among men, and 70 movies have average rating over 4.5 among women. We also analyzed the median ratings of movies among men and women over age 30 respectively. Surprisingly, we found that the number of movies with median rating over 4.5 among men and women over age 30 are 105 and 187, which is far more than 29. Note: the dataset only provides us the age range of each user. Since there exist the range 25-34, we cannot count the people whose age is between 30-34.

Then, we summarized the most popular movies. To get an accurate result, we first collected the top 20 mean ratings among all the movies.

```
title
Gate of Heavenly Peace, The (1995)                                  5.000000
Lured (1947)                                                        5.000000
Ulysses (Ulisse) (1954)                                             5.000000
Smashing Time (1967)                                                5.000000
Follow the Bitch (1998)                                             5.000000
Song of Freedom (1936)                                              5.000000
Bittersweet Motel (2000)                                            5.000000
Baby, The (1973)                                                    5.000000
One Little Indian (1973)                                            5.000000
Schlafes Bruder (Brother of Sleep) (1995)                           5.000000
I Am Cuba (Soy Cuba/Ya Kuba) (1964)                                 4.800000
Lamerica (1994)                                                     4.750000
Apple, The (Sib) (1998)                                             4.666667
Sanjuro (1962)                                                      4.608696
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954) 4.560510
Shawshank Redemption, The (1994)                                    4.554558
Godfather, The (1972)                                               4.524966
Close Shave, A (1995)                                               4.520548
Usual Suspects, The (1995)                                          4.517106
Schindler's List (1993)                                             4.510417
Name: rating, dtype: float64
```

Then, we considered the rating times of each movie and plot the top 20 rating times of movies.

```
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
```

We found that some movies are with very high rates, but have very few ratings. Combining these two aspects, we finally decided to plot movies of top 20 mean ratings with rating times no less than 1000 as the most popular movies.

```
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
```

We made the conjecture that 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. To find our conjecture is true or false, we did the following analysis. First, we grouped people by age and gender. We can see that people of age 25-34 are most likely to rate movies, while people of age under 18 rated movies the least times. Figure 1 suggested that rating showed a significant upward trend with age, which means people with older age are easier to pleased. On the other hand, in figure 2, by comparing the mean ratings rated by males and females, we can draw the conclusion that females tend to rate higher than males do.

```
age
25    395556
35    199003
18    183536
45     83633
50     72490
56     38780
1      27211
Name: age, dtype: int64
```

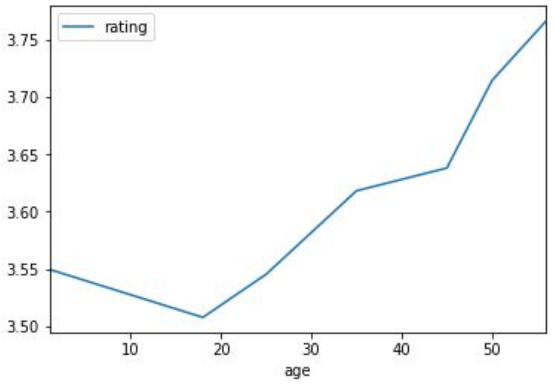

figure 1: mean rating of each age group

gender	F	M
age		
1	3.616291	3.517461
18	3.453145	3.525476
25	3.606700	3.526780
35	3.659653	3.604434
45	3.663044	3.627942
50	3.797110	3.687098
56	3.915534	3.720327

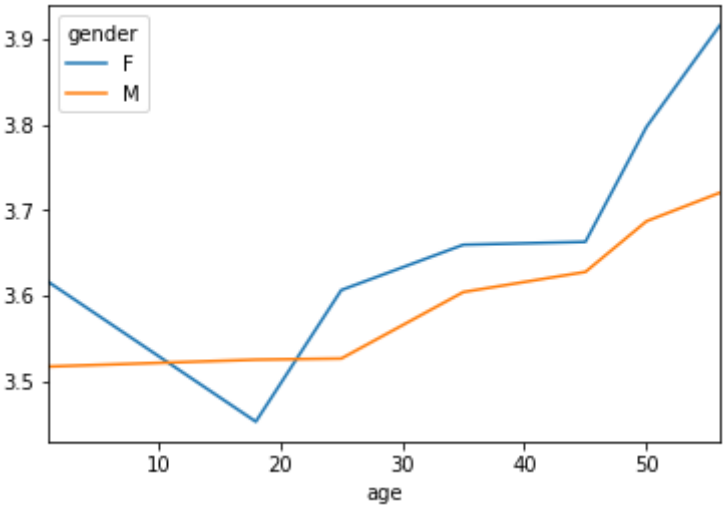

figure 2: mean rating of each age group and gender

III. Investigation to histograms

To get a deeper understanding of our dataset, we designed a number of histograms. First, we analyzed the distribution of the ratings of all the movies. In figure 3, we can see that number of ratings equal to 4.0 are biggest. So most people tend to rate the movie 4 star. In figure 4, we plot the histogram of number of ratings each movie received. Only about 200 movies among 3900 movies get more than 1000 ratings, while others get much fewer ratings. The plot is exponential.

Next, we analyzed the average rating of each movie. In figure 6 we removed the movies with less than 100 ratings and we get a histogram with fewer outlier data. In figure 5, we can see there exist some movies with average rating 5.0 in the tail, but they are removed in figure 6. We think only when a movie gets enough ratings, the average value is meaningful. So these movies with rating 5.0 cannot be judged as good movies since they didn't get enough ratings.

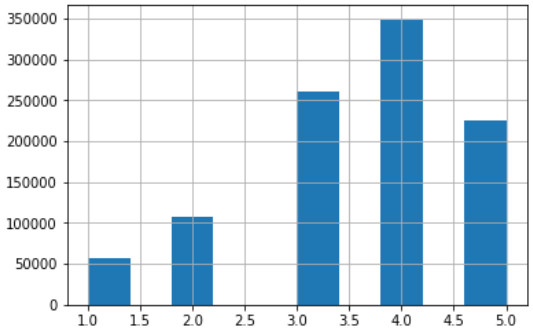

figure 3: Ratings of all movies

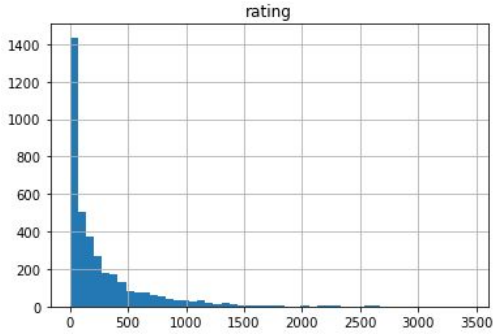

figure 4: The number of ratings each movie received

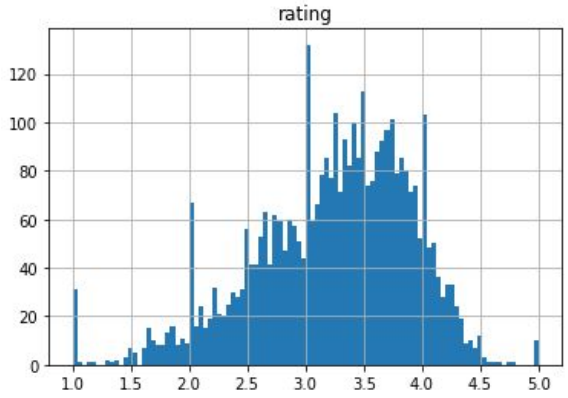

figure 5: The average rating for each movie

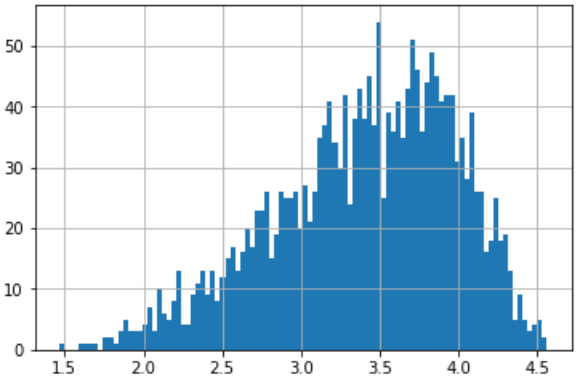

figure 6: The average rating for movies which are rated more than 100 times

To find more distribution data of ratings, we made the conjecture that comedy movies have most rating numbers since people want to have fun when watching movies. So we group the data by genre and count the rating numbers. Comedy movies have most rating numbers.

```
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
```

IV. Correlation: Men versus women

To investigate the correlation of movie preference between two genders, we made a scatter plot of men versus women and their mean rating for every movie and a scatter plot of men versus women and their mean rating for movies rated more than 200 times, and computed correlation coefficient respectively. As we can see in figure 7, men and women's rating values of most movies are highly correlated with correlation coefficient 0.76319. And when we only plot movies that are rated more than 200 times, we find this correlation becomes even stronger with correlation coefficient 0.918361. This indicates that men and women tend to have similar preference to movies in general.

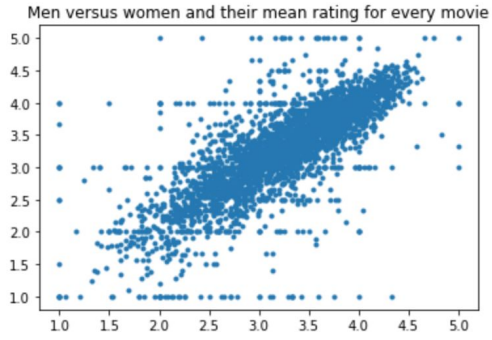

figure 7

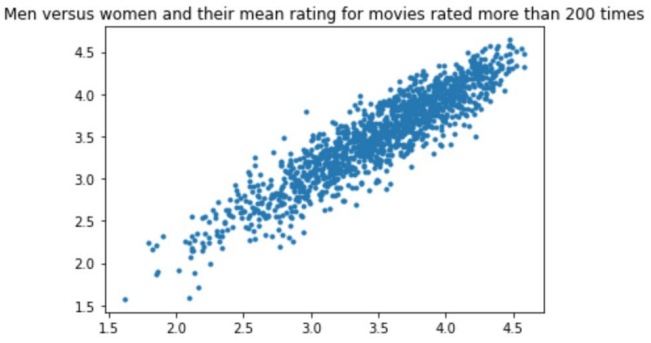

figure 8

We explored correlation of movie preference between two genders in different ages. We chose movies that rated more than 200 times as our data. Correlation coefficients are shown below. Men and women at 25 to 34 years old have the most correlated preference to movies, while those under 18 years old have a big different taste of movies.

Age groups	Correlation coefficient
Under 18	0.347655941238
18-24	0.738763533061
25-34	0.875939510758
35-44	0.802941598954
44-49	0.72521653777
50-55	0.666829431418
56+	0.49148544214

table 2: correlation of each age group

In order to find under what circumstances the rating given by one gender can be used to predict the rating given by the other gender, we made the following conjectures: men and women in the same occupation has the same preference to movies. We explored the correlation of movie ratings by occupation, and correlation coefficient are shown below. From table 3 we can see that men and women whose career are in academic and education have the most correlated preference to movies, and on the other hand, that whose career are farmer have the largest difference in the preference of 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

table 3: correlation of each occupation

Except the correlation of movie preference between men and women in different age scales and different occupations, we want to pay attention to the correlation between the two genders in different genres. using the similar function, we can get the result of correlation coefficient below. Surprisingly, we can find that the animation movie has the largest correlation coefficient between men and women and it is 1.0, which says that in animation movie, there is no preference difference between men and women. On the contrary, we can find that in film-noir movie, the correlation coefficient between men and women is quite small (below 0.05), which means that in this genre, there are very huge difference in the preference between men and women. From this table below, we can notice that there are quite different correlation coefficient in different genres.

Genre	Correlation coefficient	Genre	Correlation coefficient
Action	0.711877491492	Horror	0.782836582261
Adventure	0.951933144961	Musical	0.806155288294
Animation	1.000000000000	Mystery	0.808735268515
Children’s	0.903243944225	Romance	0.342194346703
Comedy	0.810762045468	Sci-Fi	0.632506301801
Crime	0.218831790897	Thriller	0.813697321219
Documentary	0.267361512501	War	0.308714702195
Drama	0.58870390968	Western	0.372076844225
Film-Noir	0.0270375334741		

table 4: correlation of each genre

V. Business Intelligence

If we plan to make a new movie, how we can make it a successful one?

A successful movie needs to be well accepted by audience and has a high box office earnings. So we try to answer this question in two ways: What content is loved by audience? To appeal as many audience as possible, how do we advertise this movie?

Choice of content

To investigate which genre is most popular, we analyzed mean ratings of each genre, and rating times of each genre. From the figure 10, we surprisingly found film-noir has the highest rating points with 4.08, followed by war and documentary, with score 3.89 and 3.93. These three genres are not main stream movies in the market, which has few rating times in figure 9. Drama, animation, musical has relatively higher mean ratings scores: 3.77, 3.68, 3.67 respectively. Horror has the lowest rating scores 3.21. In terms of rating times, drama and comedy has top two rating times, which in some degree reflects people are more willing to watch this kind of movies. Thus, from these plots, drama is a good choice, with both high rating scores and rating times. If we only focus on making a movie with high rating scores, film-noir, war, and documentary are also good choices.

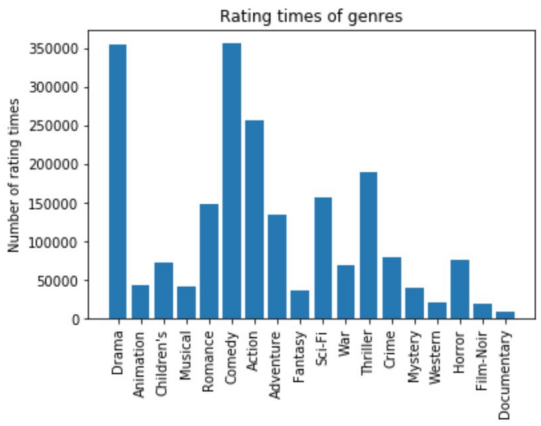

figure 9

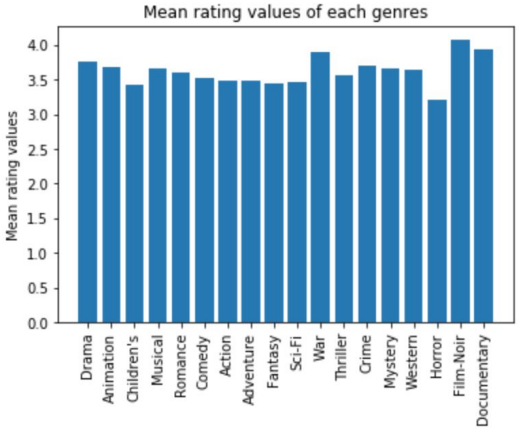

figure 10

And we were curious about if gender factor will affect our decision. We plotted mean rating values and rating times by gender. In figure 12, we calculated the rating percentage instead of rating times by each gender, to have a more clear idea of each gender's preference. As we can see from the figure 11, mean rating values by gender is consistent with the whole audience. Women like musical and children's more than men do, while men like western and film-noir more. In terms of rating times, both men and women rate drama and comedy most. The top four men rate are comedy, drama, action and thriller, while drama, comedy, romance and thriller are the top four for women.

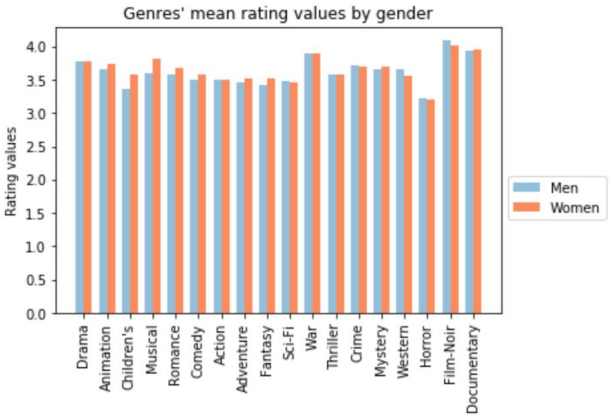

figure 11

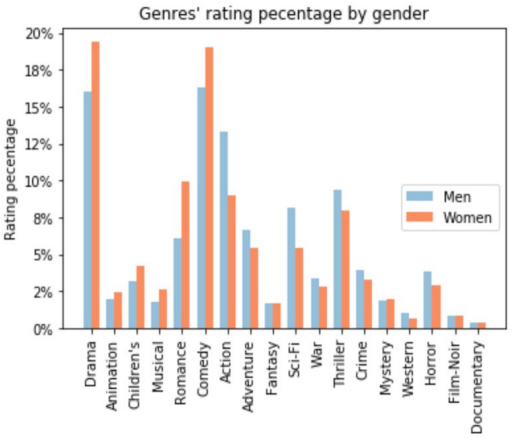

figure 12

We also want to find if there is correlation between the number of ratings and the average ratings. For each genre, we calculate each movie's rating number and mean rating value. As we see in figure 13, action and sci-fi movies have the highest correlation which means these two kinds of movies tend to have both good box office and reputation.

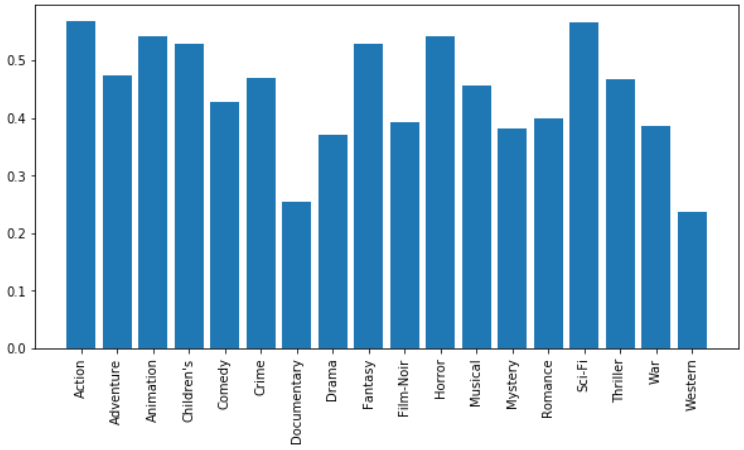

figure 13: correlation between the number of ratings and the average ratings

Advertising strategy

Suppose we have made a good quality movie, we want as many people as possible to know about our movie. We need to try to find those audience who likes to rate and rate high, so that more people will go to see the movie.

We analyzed mean rating scores and rating times by occupation. As shown in the figure 14, retired audience tend to rate highest scores while unemployed rate the lowest. Doctors, scientist and homemakers also rate relative higher than other occupations. As for rating times, college and grad students and other occupation rate movies more than other occupations. Farmers has the lowest rating times.

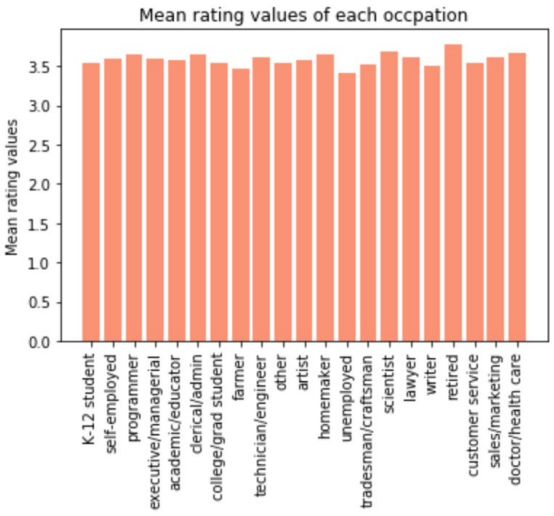

figure 14

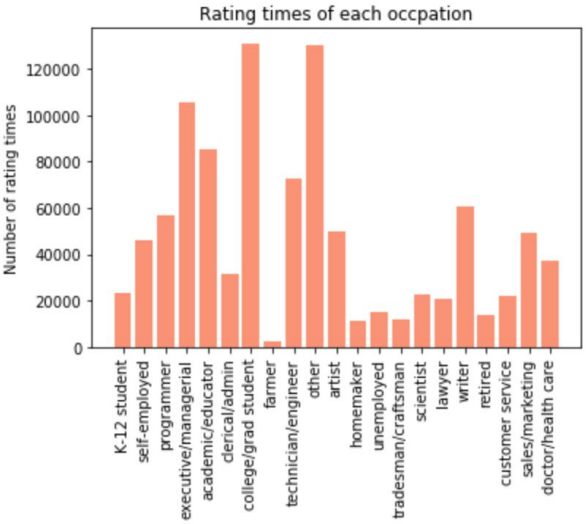

figure 15

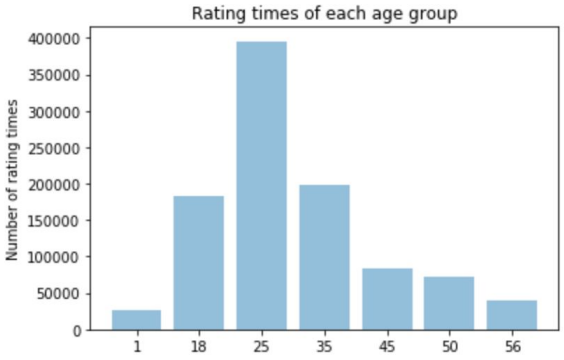

figure 16

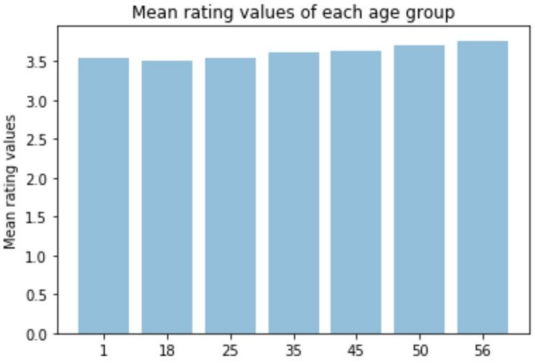

figure 17

We also analyzed mean rating scores and rating times by age. As shown in the figure 17, people older than 56 years old rate the highest scores, which is consistent with the result that retired audience rate the highest. In terms of rating times, adult at age 25-34 rate the most.

Conclusion

Based on these analysis, choosing to make drama, comedy and action movies are more likely to be successful. People rate higher and rate more are our valued audience. To have a good advising result, we need to advertise in these groups more. For example, we could send movie poster to college, graduate schools. Also we could give discount to senior and retired audience,since they rate high.