Case Study 2 - Analyzing Data From MovieLens

Brendan Foley
Luis Castillo
Yuhui Gong
Mu Niu
Matt Weiss

October 27, 2016

1 Data Collection

The data collection process consisted of downloading the MovieLens 1 million ratings data set from the GroupLens Research website. How the data was collected by GroupLens is stated on their website (http://grouplens.org/datasets/movielens/):

> GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set.

More specific data about MovieLens (https://movielens.org/info/about):

> MovieLens is a research site run by GroupLens Research at the University of Minnesota. MovieLens uses "collaborative filtering" technology to make recommendations of movies that you might enjoy, and to help you avoid the ones that you won't. Based on your movie ratings, MovieLens generates personalized predictions for movies you haven't seen yet. MovieLens is a unique research vehicle for dozens of undergraduates and graduate students researching various aspects of personalization and filtering technologies.

2 Motivations

This topic interested us because it was the first case study to involve a large data set that was well structured. In the first case study we analyzed Twitter feeds and while the amount of data we collected could have easily exceeded 1 million Tweets, it was highly unstructured. With MovieLens we were presented with the possibility of analyzing movie ratings by gender, age, movie genre and occupation with clean and well structured data.

Furthermore, since the data was focused on one specific topic (movie ratings as opposed to the noise continuously flowing through Twitter) there was the possibility of looking at a specific topic in detail.

3 Data Analysis and Results

The data analysis techniques and questions in the Jupyter Notebooks are answered in each subsection below.

3.1 Problem 1

To answer this problem, we used pivot tables with different parameters and sorted tables so that we could get movies with an average rating higher than 4.5. For example, in finding highly rated movies based on gender, we set the parameter "columns" to "gender" in the pivot table. This resulted in splitting the table into male and female, sorting them separately and selecting movies with more than 4.5 average rating among males and females. Similar procedures were applied to generate the remaining results, which are listed below.

- There are in total 21 movies having an average rating higher than 4.5.
- There are in total 51 movies among women with an average rating higher than 4.5.
- There are in total 23 movies among men with an average rating higher than 4.5.
- There are in total 149 movies with a median rating over 4.5 among women over age 30.
- There are in total 86 movies with a median rating over 4.5 among men over age 30.

Our definition of popular considered the following three parameters:

- Highly Rated
- Most Ratings Received
- Gender Neutral

Here gender neutral refers to a movie that was popular to both men and women and is defined by:

$$gender\ ratio = \frac{|number\ of\ male\ voters - number\ of\ female\ voters|}{total\ number\ of\ voters} \tag{1}$$

where the larger $(1 - gender\ ratio)$ the more gender neutral a movie was.

Using these three criteria for popularity, we firstly chose the top 100 movies with the most ratings. Next we calculated a score that applies a 0.8 weight to average rating and a 0.2 weight to gender ratio since we considered the average rating as more important in categorizing popular.

The final scoring equation was:

$$score = (1 - gender\ ratio) \times 0.2 + average\ rating \times 0.8 \tag{2}$$

where the higher this score the more popular a movie. The results are in the table below.

	count		mean		all count	Avg_rating	ratio	Score
	F	M	F	M				
Shawshank Redemption, The (1994)	627	1600	4.53907496	4.560625	2227	4.554557701	0.436910642	0.841347
Schindler's List (1993)	615	1689	4.562601626	4.491415038	2304	4.510416667	0.466145833	0.828438
Casablanca (1942)	505	1164	4.300990099	4.461340206	1669	4.412822049	0.394847214	0.827082
Usual Suspects, The (1995)	413	1370	4.513317191	4.518248175	1783	4.517106001	0.536735838	0.81539
Sixth Sense, The (1999)	664	1795	4.477409639	4.37994429	2459	4.406262708	0.459943066	0.813013
Godfather, The (1972)	483	1740	4.314699793	4.583333333	2223	4.524966262	0.565452092	0.810904
Raiders of the Lost Ark (1981)	572	1942	4.332167832	4.520597322	2514	4.477724741	0.54494829	0.807446
Silence of the Lambs, The (1991)	706	1872	4.271954674	4.381944444	2578	4.351823119	0.452288596	0.805834
One Flew Over the Cuckoo's Nest (1975)	444	1281	4.310810811	4.418423107	1725	4.390724638	0.485217391	0.805472
American Beauty (1999)	946	2482	4.238900634	4.347300564	3428	4.317386231	0.448074679	0.801167

Figure 1: Top 100 Popular Movies

In addition, we calculated each group's average rating and standard deviation based on their gender, occupation and age to see how easy various groups are to please. Here "easier to please" means higher mean rating. Our results are as follows:

- Females tends to be easier to please, with more steady rating scores.
- People under 18 are the easiest group to please, people from 18 to 24 are the hardest group, but it gets easier as age grows.
- Homemakers and retired people are the easiest group to please. Writers and unemployed people are hardest to pleased.

	Avg_rating	Occupation	STD
9	3.4868	homemaker	0.872799
13	3.4611	retired	0.862036
3	3.4020	clerical/admin	0.934391
12	3.3918	programmer	0.935027
15	3.3816	scientist	0.868145
10	3.3637	K-12 student	1.026851
18	3.3617	tradesman/craftsman	0.825533
5	3.3574	customer service	0.903788
8	3.3545	farmer	0.790726
16	3.3522	self-employed	0.901704
14	3.3449	sales/marketing	0.910138
6	3.3257	doctor/health care	0.943223
17	3.3222	technician/engineer	0.903138
0	3.2916	other or not specified	0.972678
7	3.2638	executive/managerial	0.947056
2	3.2563	artist	0.97688
11	3.2332	lawyer	0.949641
1	3.2153	academic/educator	0.931278
4	3.1897	college/grad student	1.008008
20	3.1468	writer	0.965803
19	3.1431	unemployed	0.981837

Age	1	18	25	35	45	50	56
Avg_rating	3.34	3.16	3.23	3.28	3.29	3.34	3.37
STD	1.0261	1.0209	1.0032	0.9496	0.9207	0.9053	0.9089

Gender	F	M
Avg_rating	3.29	3.21
STD	1.0045	1.0054

Figure 2: "Ease To Please"

3.2 Problem 2

In problem 2 we analyzed the grouping of movie ratings using histograms. There were four histograms in total.

The first histogram (ratings of all movies) showed the distribution of the five possible movie rating values from 1-5. The x-axis represents the rating and the number of movies receiving a specific rating is represented on the y-axis. The second histogram (number of ratings each movie received) showed the distribution of how many ratings an individual movie received. Here the x-axis represents the number of ratings a movie received and the y-axis how many movies received the rating spanned by a given bin. The third histogram(average rating for each movie) looked at the average rating of each movie. The x-axis represents the average rating and the y-axis the number of movies receiving that rating. The last histogram shows the same data and the previous histogram, but only for movies that received more than 100 ratings. One noticeable difference between the last two histograms is in the tails. This will be discussed in more detail below in the results section. Below are the four histograms.

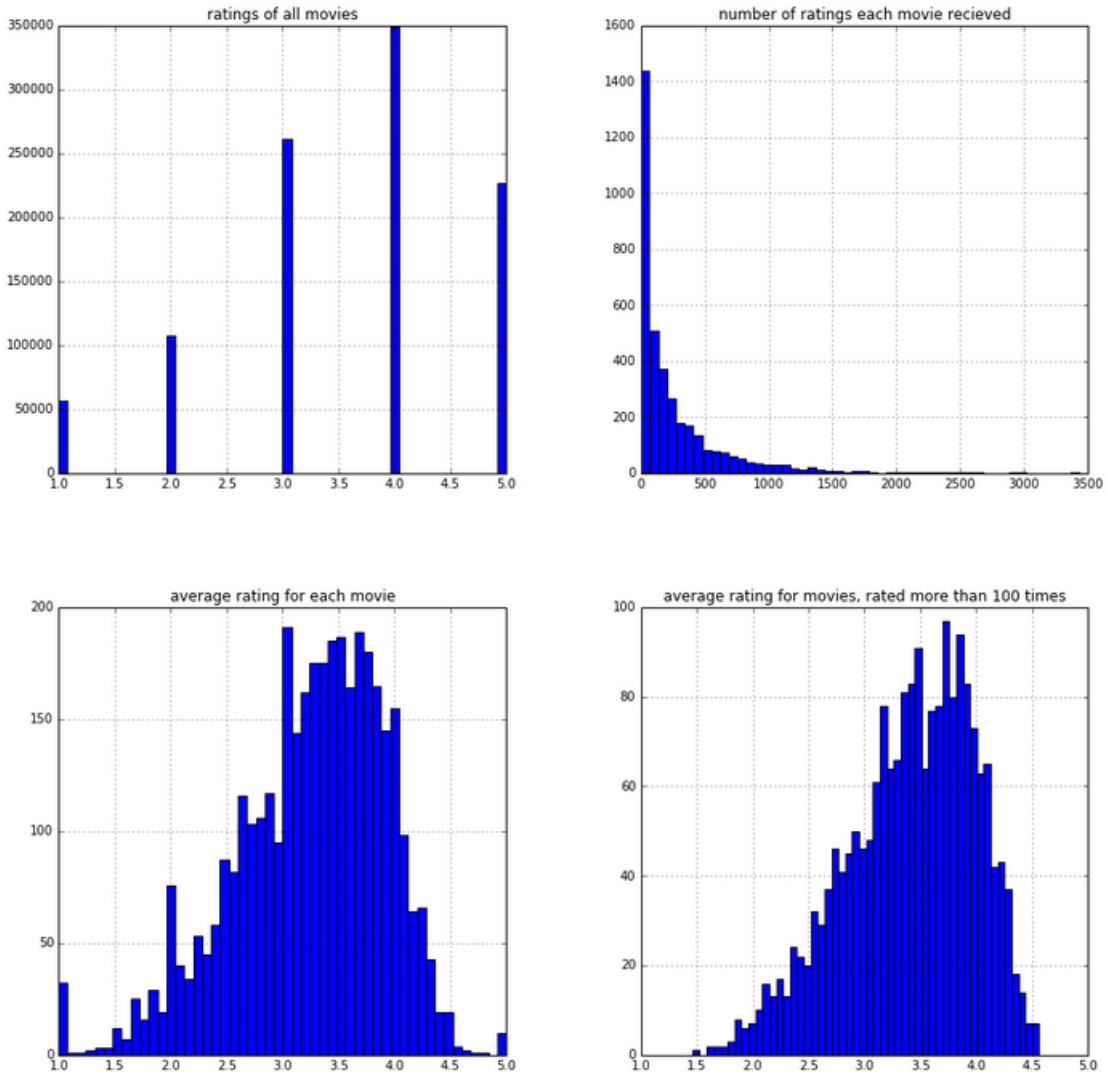

Figure 3: Problem 2 Histograms

Looking at the histogram showing the number of ratings each movie received (top right), the first bin is much larger than the next bin by approximately three times and contains movies that received much less than 100 ratings. This tells use there were a lot of movies that received very few ratings. In fact, this single bin accounts for approximately one-third of the 4000 movies in the MovieLens data set. Yet, this fact alone does not tell us all (or the majority) of these movies with just a few ratings received a disproportionate number of extreme scores (here extreme score means a rating of 1 or 5).

When comparing the bottom two histograms we notice two things. First if both histograms are scaled the one incorporating all movies is roughly twice as high as the other. This is consistent with the claim above that approximately one-third of all rated movies received less than 100 ratings. Second, the tails for the histogram with movies receiving more than 100 ratings drop off sooner at the tails. In fact, movies with ratings less than 1.5 and greater than 4.5 (approximately) are, for the most part, not present. This, combined with the evidence in the preceding paragraph, tells us that while the majority of the movies receiving less than 100 ratings may not have been extreme ratings, the majority of extreme ratings were for movies receiving less than 100 ratings.

Our conclusion is, only movies receiving ratings above a minimum threshold should be considered trustworthy. The reasoning is similar in baseball, where to win the batting title one must have a minimum number of at bats. Otherwise, it would be possible for someone with one at bat to have a hit and their batting average be perfect.

The next question is "can we reasonably speculate on who is making these ratings at the extreme ends of the rating spectrum?"

Perhaps these are obscure movies with a small fan base that is very loyal; something like a cult following. Until this point we have only claimed the majority of extreme ratings occur for movies with less than 100 ratings. Looking at the first histogram though we see there are many more 5 ratings compared with 1 ratings. This tells us the extreme ratings tend to be high. So who would be responsible for a movie with a small number of very high ratings? This depends and it is hard to make any solid inferences from this data. Perhaps it is young people who are less judicious in their choices or have less impulse control. Individuals who are accustomed to sending out whatever crosses their mind in 140 characters or less before giving it much thought. This is only half-serious speculation but it does open the door for further investigation to determine just who (or whom) is responsible for high ratings on movies with less than 100 ratings.

3.3 Problem 3

In problem 3 we wanted to know what social factors could influence whether or not male and female tastes were similar or not. The data was broken up based on various different factors. Then regression tests of Male vs. Female ratings of films within that subset of data were run.

We first generated two scatter plots:

- men versus women and their mean rating for every movie
- men versus women and their mean rating for movies rated more than 200 times

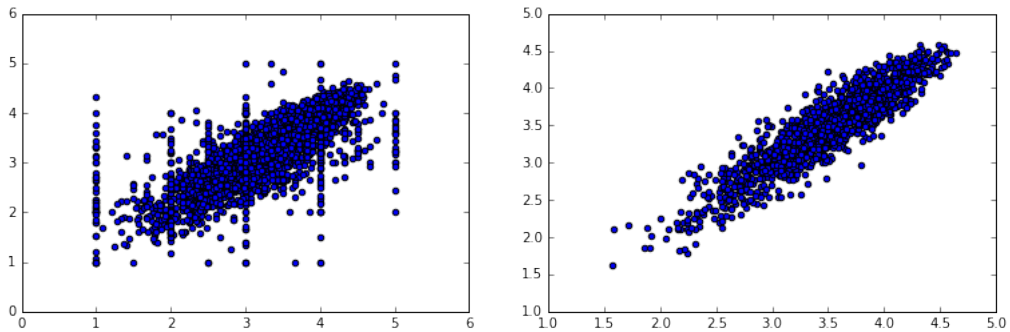

Figure 4: Left: men versus women and their mean rating for every movie. Right: men versus women and their mean rating for movies rated more than 200 times

The correlation coefficient between men and women is 0.763 with a p-value of 0. This means that the correlation is very significant. But is it meaningful? This question is more subjective,

but we say that a correlation coefficient of 0.76 shows that the ratings are similar. For young viewers the correlation is 0.56, while for old viewers it is 0.75. This seems to imply that movie tastes among genders are more similar for older people than younger people. To answer this we calculated the correlation coefficient for both young and old people below.

The correlation coefficient for young people is higher than that of old people at a significance of $p = 2.6 \times 10^{-43}$, which is very significant.

For active movies the correlation is 0.92, while for non-active movies the correlation is 0.68, this seems to imply that movie tastes among genders are more similar for movies that are widely viewed. The correlation coefficient for active movies is higher than that of non-active movies at a significance of p = 0, which is extremely significant. There are two possible interpretations of this.

1. Movies that are less mainstream can be targeted more to a certain gender. An example of this might be the Netflix show Jessica Jones which is rare for a Marvel movie on the big screen to have a female protagonist due to the mainstream nature of these films, and the return on investment they need to achieve. But shows on Netflix have the freedom to be more selective to the group they want to target.

2. There actually is no major divergence between genders among non-active films and all that we are observing is the noise of a small sample size. Put differently if a movie in the non-active movie list were to be seen by a larger audience, and rated by a larger audience, the male and female average rankings would begin to move closer to one another.

For dramas the correlation is 0.58, while for comedies the constellation is 0.68, this seems to imply that movie tastes among genders are more similar for comedies. The correlation coefficient for active movies is higher than that of non-active movies at a significance of $p = 4.5 \times 10^{-14}$, which is extremely significant. The correlations don't seem to change based on location as we got a p-value of 0.78 for ratings by people in the North and South. Here North/South was divided based on zip code.

In summary, the factors that seem to influence whether or not men and women will be more homogeneous with respect to film tastes are age, how mainstream a film is, and the genre of the film (in particular whether it is a drama or a comedy). Factors that don't seem to influence the homogeneity of the rankings are location (in particular if you live in the south, or the northeast), and male-popularity (in particular if men really like the film, or really hate the film).

3.4 Problem 4

Based on the Top 10 list in problem 2, nine out of the ten movies are within the genre of drama. Only "Raiders of the Lost Ark" is classified as an adventure movie. This tells us that Drama is the preferred genre among male and female raters in the database. All ten movies have a considerable amount of ratings as they all surpass 1500 ratings. In addition, the overall average rating for these movies is greater than 4, which is consistent with the histogram data above.

If a studio would want to make a decision on a movie, then the story should be targeted to women, in middle age. This is based on the fact (supported in our data) female raters and homemakers are the easiest groups to please. The story would have to be a drama, related preferably to a criminal story; since it is one theme that is consistent in the top 10 movies. By

incorporating a criminal element the film would appeal to men as well.

The business question that this data set can answer would be: "which story should the studio invest into a movie?" Is there any story that would be worth making a sequel or prequel in order to extend the movie franchise? One proposal that could be pitched to the producers in Warner Brothers is making a prequel to "The Shawshank Redemption". Showing Ellis Boyd 'Red' Redding's life (Morgan Freeman's character) before meeting Andy Dufresne (played by Tim Robbins). The expansion of the movie franchise would add to the $58.5 million the original movie made at the box-office made in 1994.