Analyzing Data from MovieLens: In General Sequels Stink

Case Study 2
October 26, 2017

Group 1
Claire Danaher
Jonny Friedman
Janvi Kothari
Renee Sweeney
Erin Teeple

INTRODUCTION

The goal of this case study was to analyze the MovieLens data set containing movie reviews for 1 million movies(1). Specific research goals included gaining insight into movie ratings generated by diverse audiences, to evaluate conjectures about movie ratings data using graphical and quantitative analysis, and to develop a story explaining our findings. We were also asked to propose a business case application for an analysis of MovieLens data that would be of interest to a movie company.

When formulating our business case application, we first carefully considered our initial analysis of movie ratings data and considered the relationship between movie reviews and movie profitability. We identified the costs of movie content generation: production budget, marketing, theater distribution, talent compensation, and financing costs (2-4). Revenue sources over the product life cycle of a movie were also outlined. For a given movie, revenue is generated not only from ticket sales during movie theater runs, but also from home entertainment sales (e.g. DVD and video), television agreements, video streaming (e.g. Amazon, Netflix), screenings in hotels and on airlines, through merchandising, and soundtrack sales (2-4).

The high costs of movie production can potentially be offset by income from multiple revenue sources over a movie product life cycle extending multiple years. Nonetheless, in one study of financial information from 279 Hollywood movies for which production budget and profit/loss data were available, only about 51% of these movies generated a profit, while 49% resulted in a loss (3). Furthermore, this study also reported that about half of all the money lost on movies that were not profitable was lost on the worst-performing 6% of films (3). When undertaking our business case analysis, we noted that it is possible for audience reviews of movies to be inconsistently linked to movie profitability. This is because some poorly rated movies may appeal to audiences for reasons other than artistic merit, for example sequels in popular movie series, adaptations of bestselling books, or features starring specific talent. Nonetheless, over the product lifecycle of a movie, we expect that interest in a movie will wane more quickly if the content itself is of lower quality. We assumed that movie ratings are one method for assessing such product quality.

Methods

Given the size of the data set and potential issues with data transfer via email, our group created cloud based database in MongoDB. Data included in the data sets were as follows: user_id, movie_id, rating, timestamp, title, genre, gender, age, occupation, and zipcode.

Pivot tables was used to calculate the number of movies with specific average and median ratings and to further examine other subgroup characteristics according to the available information fields. Feature-specific data frames were created for these analyses by selecting records of interest based on specific extraction criteria.

For our analyses, we constructed operational definitions for popular movies and the most poorly rated movies. Popular movies were conceptualized as those which had been widely viewed and highly rated, and we then operationalized these characteristics as movies with ratings frequencies above the median and as those with the highest ratings within this grouping. The most poorly rated movies were then conceptualized as those which had been widely viewed and poorly rated, and we operationalized these as movies with ratings frequencies above the median and as those with the lowest 5% of ratings within this grouping. For our business case analysis, Word Cloud was used to visualize the most frequently used words in popular and poorly rated movies by genre.

Results

Problem 1: Basic Details

The first question to be investigated asked about some basic details of the data set. The table below provides a summary of the movies with an average overall rating of 4.5 for various populations including: all people, men, women, men over 30 and women over 30.

Table 1. Basic details

Number of movies with an average rating over 4.5 overall	29
Number of movies with an average rating over 4.5 among men	29
Number of movies with an average rating over 4.5 among women	70
Number of movies with a median rating over 4.5 among men over age 30	105
Number of movies with a median rating over 4.5 among women over age 30	187

We defined “popular” movies as those that are both widely viewed and which receive high ratings. To identify the top ten most popular movies, movies were first sorted by rating frequency. Popular movies were then defined as the highest-rated movies among movies with a rating frequency greater than the median rating frequency.

Table 2: Top 10 Most Popular Movies

title	rating
Gate of Heavenly Peace, The (1995)	5.0
Schlafes Bruder (Brother of Sleep) (1995)	5.0
One Little Indian (1973)	5.0
Song of Freedom (1936)	5.0
Baby, The (1973)	5.0
Smashing Time (1967)	5.0
Follow the Bitch (1998)	5.0
Lured (1947)	5.0
Ulysses (Ulisse) (1954)	5.0
Bittersweet Motel (2000)	5.0

We then made some conjectures about how easy various groups are to please. We wondered if older viewers preferred older movies and/or if younger viewers preferred more recent movies. We observed that people age 50 and over did rate older movies slightly higher on average (Figure 1&Figure 2). Average ratings by viewers under age 18 were not found to demonstrate a similar trend or a trend suggesting preference for more recent movies.

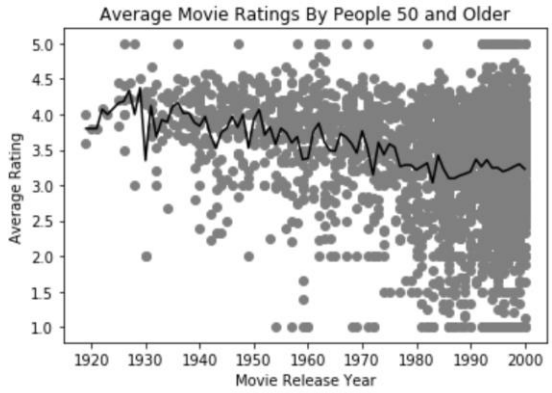

Figure 1

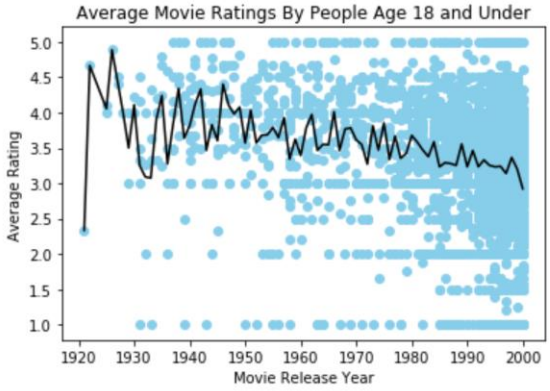

Figure 2

Problem 2: Histograms

We then expanded our investigation to histograms. One specific issue with inferences drawn from Problem 1 was that the number of times a movie was rated was not fully considered. The use of histograms can help investigate questions about frequency.

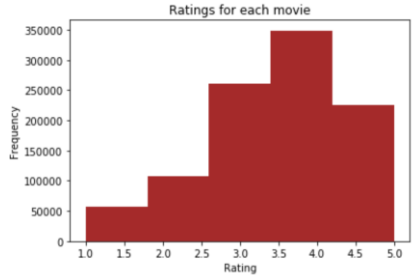

Figure 3:Histogram of the ratings of all movies

We observed that the most frequent ratings were between 3 and 4 and that rating frequencies were much higher for some movies compared with others (Figure 2). The observation of a non-linear relationship in which popular things are exponentially more popular than unpopular things can be described as following Zipf's Law, which was formulated to characterize natural language word frequencies but which also been found to apply in other ranking cases, as well (5).

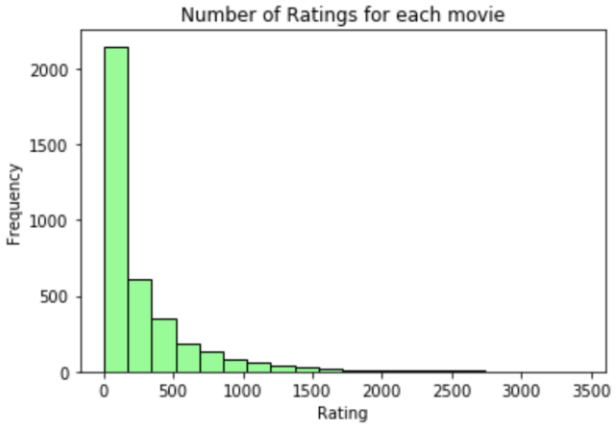

Figure 4:Histogram of the number of ratings each movie received

In Figure 5, we see that among rated movies, the average ratings for each movie appear to be approximately normally distributed.

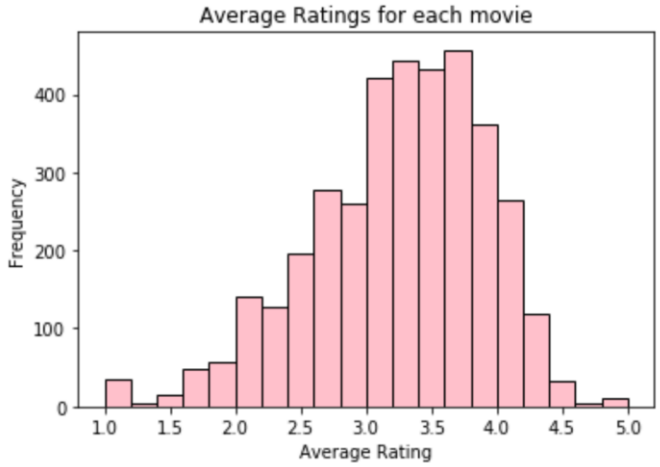

Figure 5: Histogram of the average rating for each movie

Breaking average ratings down to compare the average ratings for movies rated more than 100 times with those rated less than 100 times, we see that the distribution for movies rated more than 100 times appears approximately normal with a slight left skew. In comparison, the distribution for movies rated less than 100 times is less normal-appearing. This observation is not surprising, given the distributions are created from different numbers of ratings.

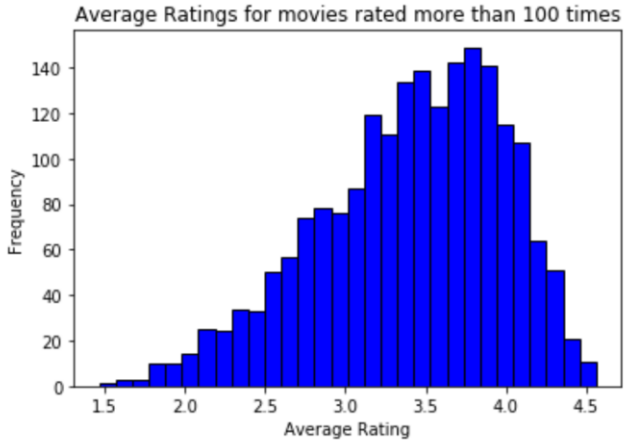

Figure 6:Plot a histogram of the average rating for movies which are rated more than 100 times.

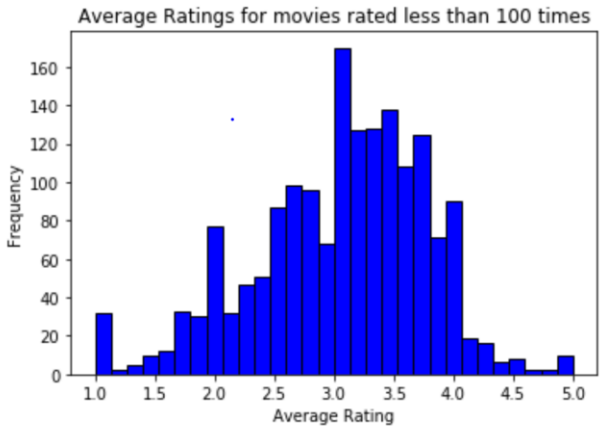

Figure 7:Average ratings for movies rated less than 100 times

Comparing the tails of the histogram composed of all the movies versus the one composed only of movies rated more than 100 times, we observe that the tails are smoother in the histogram including only movies rated more than 100 times. In answer to the question of which highly rated movies would you trust are actually good (those rated more than 100 times or those rated less than 100 times), movies rated more than 100 times would have a greater sample size for estimation of the population mean for perceived movie quality. We would therefore most trust the quality of the highly rated movies which had been rated the higher number of times.

Conjectures

One conjecture we made was that age might be differentially related to mean rating scores. We observed that mean rating increased with increasing age. A possible explanation for this finding is that older individuals have likely watched more movies than young people, resulting in differences in assessment of movie production value, in contrast to younger audiences. Perhaps older audiences have watched more poor

quality movies, resulting in the observed increase in average movie ratings with increasing age over 20 years.

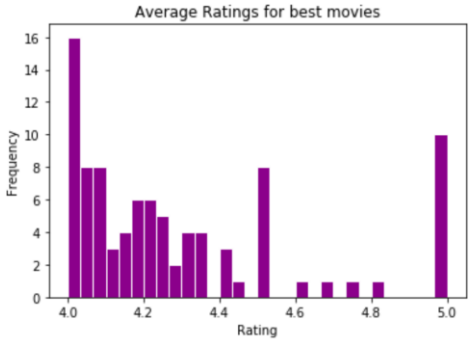

Figure 9:Average ratings for the best movies

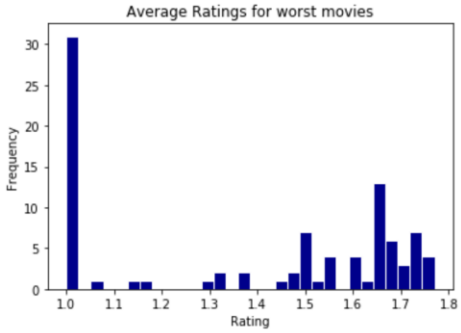

Figure 8: Average ratings for the worst movies

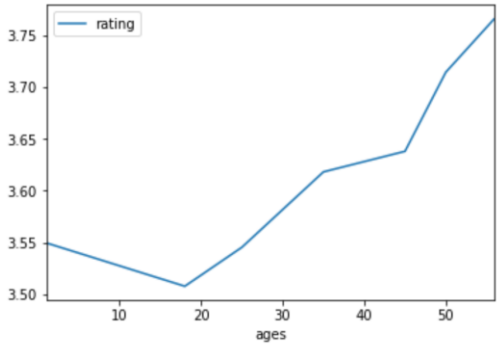

Figure 10:. Average ratings by age

We then looked at differences in the average ratings for the best and worst movies (Figure 8&9). We observed that the best movies were more frequently rated closer to 4 than 5. Movies receiving the worst possible average rating were more frequent in our data set than movies receiving the best possible average rating. These observations could occur because movies of the worsts quality are made more frequently than movies of the best quality. Another possibility is that raters assign ratings for best and worst quality asymmetryically - ie perceived worst quality is not the same cognitive distance from neutral as perceived best quality.

Problem 3: Correlation: Men versus women

In this section we looked more closely at relationships between the pieces of data we have.We see in Figures 7 and 8 that average ratings for men and women are relatively similar, both for mean ratings for every movie and for movies rated more than 200 times. We observe that among movies rated more than 200 times there is even greater agreement in men's and women's mean ratings. The *correlation coefficient* between the

ratings of men and women was found to be 0.91836139. This suggests that the ratings are quite similar, and that men's ratings can be used generally to predict women's rating

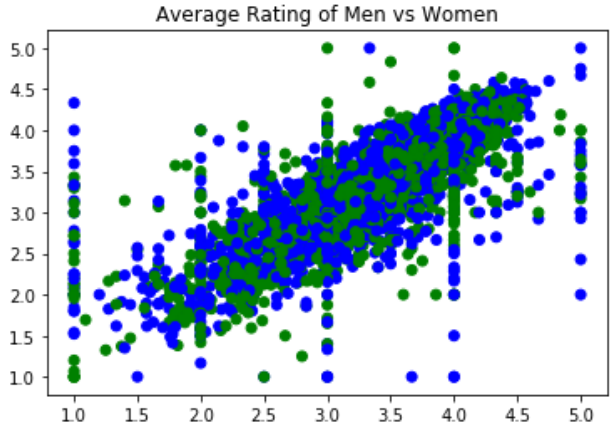

Figure 11:Scatter plot of men versus women and their mean rating for every movie; blue = female, green = male.

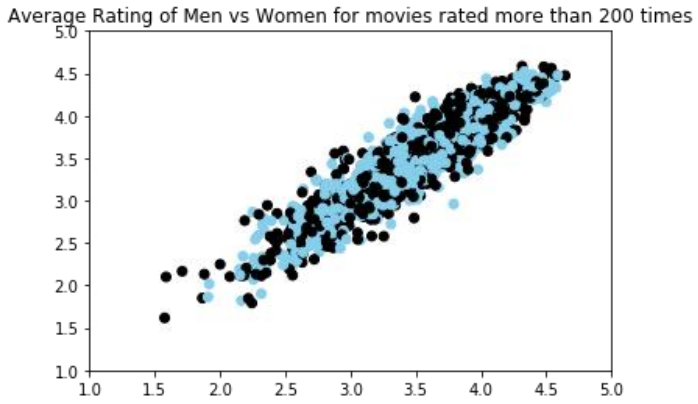

Figure 12:Scatter plot of men versus women and their mean ratings for movies rated more than 200 times (blue = female, black = male)

Using combined information on occupation and gender provides one circumstance in which ratings given by one gender can be used to predict the rating given by the other gender (Figure 13/Table 3). Occupation 1 (Academic/Educator) has a high male-female correlation versus Occupation 10 (K-12 Student), which has a low male-female correlation.

Table 3.

```
Correlation Coefficient between Men and Women's Average Movie Ratings based on Occupation
Occupation 0: 0.578786108205
Occupation 1: 0.636357634705
Occupation 2: 0.472413764133
Occupation 3: 0.438775296571
Occupation 4: 0.572648438461
Occupation 5: 0.329810126208
Occupation 6: 0.518478827401
Occupation 7: 0.572695642366
Occupation 8: 0.275236368043
Occupation 9: 0.276577331069
Occupation 10: 0.330525786667
Occupation 11: 0.394055882261
Occupation 12: 0.450083759757
Occupation 13: 0.294298338909
Occupation 14: 0.533524348122
Occupation 15: 0.47962134872
Occupation 16: 0.468766904706
Occupation 17: 0.579449959376
Occupation 18: 0.276750813049
Occupation 19: 0.408121713176
Occupation 20: 0.606829865489
```

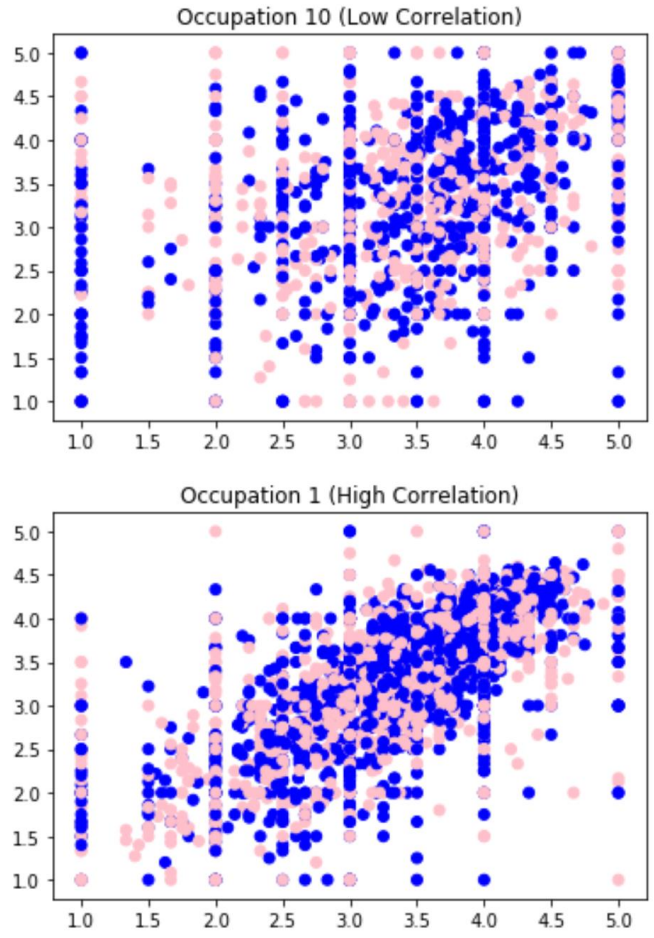

Figure 13: Occupational Plots

Problem 4: Business Intelligence

Movie ratings reflect received audience-perceived quality of a movie. A poorly rated movie may be profitable if an audience exists for it, but movie ratings do provide some

interesting insights into content that appeals to viewers. Given that mean ratings were observed to vary by age, care should be taken in generalizing findings across age cohorts.

Movie titles commonly make a first impression with an audience and are much like the hook of an article. When reading through a list of movies, the title can either grab the viewer's attention and incite further investigation or result in quickly continuing on to consideration of other movies. Furthermore, while some movie titles are abstract, consideration of nouns and verbs which attract viewership provide indication of themes that attract viewership within a genre.

Wordclouds are a commonly used approach for understanding the lexicon of a topic. Wordcloud produces a visual representation of word frequency whereby the size of the word corresponds to the frequency with which it is used. Outlined below is the results of an analysis which analyzed the word frequency by genre of both popular and unpopular movies. For the purposes of this analysis, we expanded the definition of popular to include all movies with reviews above the median number of reviews and accounting for the top/bottom 30%. While the information provided below does not provide hard and fast rules, it provides interesting context when considering what movies both movie titles and themes. .

One of the most pronounced findings of the word clouds was the consistent prominence of sequels notation(II and III) in word clouds for worst movies across genres. Provided below are the word clouds for thriller, action and comedy genres.

Thriller Word Cloud for Worst Movies

Comedy Word Cloud for Worst Movies

Horror Word Cloud for Worst Movies

While the movies were poorly rated, the frequency of the word is also a function of the number of titles containing this world. This would suggest that despite that movies are commonly reviewed poorly they also attract an audience.

The same logic can follow for the term “Movie”. As outlined below, they attract an audience but are generally are rated poorly. Further analysis is needed to explore relationships between ratings and profitability by genre.

Animation Word Cloud for Worst Movies

Adventure Word Cloud for Worst Movies

Children's Word Cloud for Worst Movies

CONCLUSIONS

In this case study we analyzed features of viewer-produced movie ratings, explored conjectures about movie ratings data, and developed a business case application for MovieLens data analysis. From a business perspective, the most prominent finding of our analysis was that certain movies, despite commonly being rated poorly still manage to attract an audience. If a movie studio was concerned with the rating-assessment and quality of a movie, they would be wise to steer away from sequels and movies lacking sufficient imagination that the word “movie” is used in the title. However, the ability to attract an audience regardless of quality presents a lower liability approach for studios. Costs of movie content generation includes production budget, marketing, theater

distribution, talent compensation, and financing costs. Given the popularity of these movies regardless of quality, a lower marketing budget would likely be required.

Subsequent analyses are needed to further explore relationships between ratings and profitability by genre.

REFERENCES

1. https://movielens.org/info/about
2. http://www.investopedia.com/articles/investing/093015/how-exactly-do-movies-make-money.asp
3. https://stephenfollows.com/hollywood-movies-make-a-profit/
4. http://www.nytimes.com/2012/07/01/magazine/how-does-the-film-industry-actually-make-money.html
5. https://en.wikipedia.org/wiki/Zipf%27s_law