Using MovieLens 1M Dataset to aid Movie Production Houses in data driven decision making

Submitted By:

Todd Hay
Mukund Khandelwal
Zeling Lei
Brandon Werner
Zhe Lyu

Submitted To:

Prof. Randy Paffenroth
Instructor: DS501
WPI
10/26/2017

Table of Contents

Using MovieLens 1M Dataset to aid Movie Production Houses in data driven decision making

Todd Hay, Mukund Khandelwal, Zeling Lei, Brandon Werner, Zhe Lyu

Abstract

Producing a movie can represent a significant investment of time and resources by a movie company so due diligence is required to ensure that a new movie will be popular with potential audiences. Our team (Team 14) proposes that useful information can be extracted from a data set of over one million movie reviews and used to guide the development of a new movie. We first performed exploratory data analysis (EDA) to better understand the primary characteristics of the movie review data set. Guided by the information collected through EDA, we made various conjectures on the interactions and relationships between movie reviewers and their assessments of movie quality to identify areas of risk and opportunity.

Next, we identified popular movies based on number of reviews and analyzed them in search of common attributes. We observed that the action genre is associated with 7 out of 10 popular movies. While the standard deviation for ratings wasn't significantly different between action and non-action genre movies, we did find that the correlation coefficient for gender improved when considering only movies with the Action genre indicator. So, our analysis indicated that action genre movies are similarly liked between Men and Women, the standard deviation of the ratings is like non-action movies and seven out of ten movies identified as popular are associated with the action genre.

Finally, we recommend that the movie company DreamWorks create a minimally viable movie, in the form of a movie trailer, for an action genre movie. Doing so will make the movie available to potential consumers as soon as possible and provide the opportunity to collect feedback as potential consumers watch, comment on and rate the trailer in various channels such as YouTube or IMDB.com.

Motivation

Our team believes there is actionable information hidden in the data set that can help a movie production company like DreamWorks experience success at the box office. It is our desire to unlock the value through the application of statistical rigor to identify the types of movies that are popular and experience fiscal success at the box office. Furthermore, we hope to share our findings with DreamWorks which is undoubtedly anxious to improve their performance at the box office and improve their ranking within the movie production industry.

Introduction

As a starting point our team will convey some basic details regarding the MovieLens 1 million movie reviews data set. This data set contains data about users, movies and ratings provided by thousands of

users for thousands of movies. Specifically, we analyzed three files which, when combined, collectively contained “1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000” [1.1].

User information comes from the users.dat file and contains demographic information regarding age range, gender, ZIP code and occupation categories.

Movie information comes from the movies.dat file and contains information such as Movie ID, title and genres. “Genre is the term for any category of literature or other forms of art or entertainment, e.g. music, whether written or spoken, audio or visual, based on some set of stylistic criteria. Genres are formed by conventions that change over time as new genres are invented and the use of old ones are discontinued. Often, works fit into multiple genres by way of borrowing and recombining these conventions” [1.2]. Genres are pipe-separated and are selected from the following genres: Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western [1.3].

Ratings information comes from the ratings.dat file and contains information for linking ratings to users and movies. The file of course also contains ratings (which are made on a 5-star scale and are whole-star ratings only) and a timestamp “represented in seconds since the epoch as returned by time (2)” [1.4]. Each user has at least 20 ratings in the ratings.dat file.

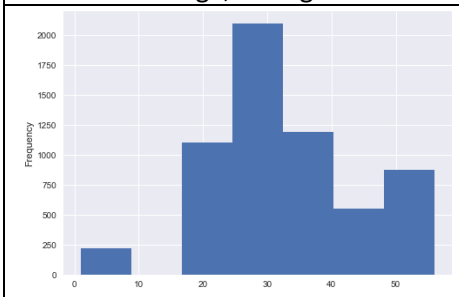
This data set contains significant insights into the relationships and interactions between user traits and movie traits and ultimately their potential effects on users’ ratings for movies. Team 14 will explore these areas and share our findings in this paper. As a first step, we will do some exploratory data analysis and present some basic details of the data our team collected and analyzed.

Problem 1: Importing the MovieLens 1M data set and performing Exploratory Data Analysis (EDA)

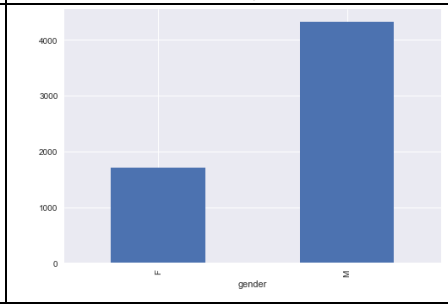
The perfunctory step was to import and merge the three files that make up the 1 million movie reviews data set. As mentioned earlier, three data files were imported then merged into a single Pandas dataframe. With the single dataframe established the results was then save to an external file to preserve this initial, combined data set. The file was saved in the HDF5 format which is “a data model, library, and file format for storing and managing data. [1.5]”

Some basic details for the MovieLens users are shown below with help from the Matplotlib library:

User Age, Histogram



User Gender, Bar Chart



Age and Gender Table

age	F	M
1	78	144
18	298	805
25	558	1,538
35	338	855
45	189	361
50	146	350
56	102	278
	1,709	4,331

Next, we looked at the ratings data. Using pandas data frames and pivot tables we examined how many movies had an average rating greater than 4.5 star (on a 5-star scale) for the entire user population, then again by gender. With this first analysis, we already see some variation in ratings depending on the gender of the user. The findings are below and details are available in Appendix 1-A:

All Users	**Male Users**	**Female Users**
Question: How many movies have an average rating over 4.5 overall? Answer: 21	Question: How many movies have an average rating over 4.5 among men? Answer: **23**	Question: How many movies have an average rating over 4.5 among women? Answer: **51**

We then considered gender and age when assessing movies with a median rating greater than 4.5. The first step was to restrict the ratings to Men and Women over the age of 30. Then we applied the aggregate function 'median' which effectively returned all movies with a rating of 5 stars due to the nature of the rating system where users could provide only whole star ratings.

It is important to note that the filter for age > 30 returns age ranges slightly out of scope. The age field is a code for an age range. So, when we restricted the data frame to "> 30" we are analyzing a collection of age ranges; "35-44", "45-49", "50-55" and "56+". So, the median ratings are for Men and Women over 34. The findings are below and details are available in Appendix 1-B:

Men over 30 with Median Movie Rating > 4.5	**Women over 30 with Median Movie Rating > 4.5**
Question: How many movies have a median rating over 4.5 among men over age 30? Answer: **86**	Question: How many movies have a median rating over 4.5 among women over age 30? Answer: **149**

A final basic analysis was to determine what are the Top 10 Popular Movies. The definition of "popular" can be subject to debate but our method was to simply identify the number of ratings per movie using pandas and pivot tables, sort the list in descending order by number of ratings and slice off the top 10 rows. The justification for this method being that "popular" movies are those which have the greatest number of people sharing their opinion about the movie. If a person shares their opinion about a movie we can safely assume that means the person saw the movie. So, we can then consider each movie review to indicate a movie was watched, and each viewing of a movie is a "vote" in favor of the movie being popular.

A movie can be good or bad by any number of movie critic standards but what likely matters to a movie company is revenue. In fact, some movies can be so bad by most movie critic standards that they develop a cult following. So, regardless of the ratings, what ultimately matters are if people were willing to pay for a ticket to see a given movie.

With that justification in mind, we applied the method described above which yielded the following results for Top 10 Popular Movies. A pivot table was created to calculate the mean rating and count of ratings by movie title as well as show the genres associated with each movie. The results were sorted in

descending order by number of reviews and the top ten rows were sliced from the data frame which yielded the following:

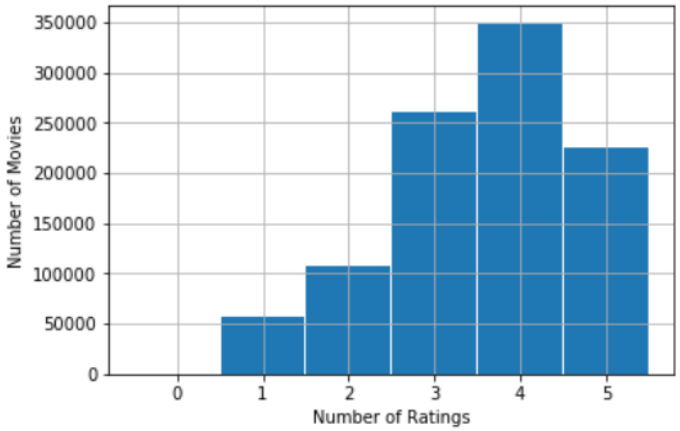
		rating	reviews
		mean	
title	genres		
American Beauty (1999)	Comedy|Drama	4.317386	3428
Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Fantasy|Sci-Fi	4.453694	2991
Star Wars: Episode V - The Empire Strikes Back (1980)	Action|Adventure|Drama|Sci-Fi|War	4.292977	2990
Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Romance|Sci-Fi|War	4.022893	2883
Jurassic Park (1993)	Action|Adventure|Sci-Fi	3.763847	2672
Saving Private Ryan (1998)	Action|Drama|War	4.337354	2653
Terminator 2: Judgment Day (1991)	Action|Sci-Fi|Thriller	4.058513	2649
Matrix, The (1999)	Action|Sci-Fi|Thriller	4.315830	2590
Back to the Future (1985)	Comedy|Sci-Fi	3.990321	2583
Silence of the Lambs, The (1991)	Drama|Thriller	4.351823	2578

The veracity of our list of popular movies is supported by the following data from Box Office Mojo (http: //www. boxofficemojo.com). This supplemental data from Box Office Mojo shows domestic, wide-opening weekend gross dollars earned for each movie title in our Ten Most Popular Movies list. Nine out of ten movie titles in our Ten Most Popular Movies list were the top grossing movies for their wide-opening weekend.

Movie Title	**Wide-Opening Weekend: Rank & Gross Dollars**
American Beauty:	No.3 rank, 706 theaters, $11,598 average
Star Wars: Episode IV - A New Hope:	No.1 rank, 757 theaters, $8,992 average
Star Wars: Episode V - The Empire Strikes Back 1980):	No.1 rank, 823 theaters, $13,171 average
Star Wars: Episode VI - Return of the Jedi (1983):	No. 1 rank, 1,002 theaters, $22,973
Jurassic Park (1993):	No. 1 rank, 2,404 theaters, $19,561 average
Saving Private Ryan (1998):	No. 1 rank, 2,463 theaters, $12,414 average
Terminator 2: Judgment Day (1991):	(Non-3D) No. 1 rank, 2,463 theaters, $12,414 average (3D) No. 25 rank, 371 theaters, $1,490 average
Matrix, The (1999):	No. 1 rank, 2,849 theaters, $9,753 average
Back to the Future (1985):	No. 1 rank, 1,420 theaters, $7,853 average
Silence of the Lambs, The (1991):	No. 1 rank, 1,497 theaters, $9,196 average

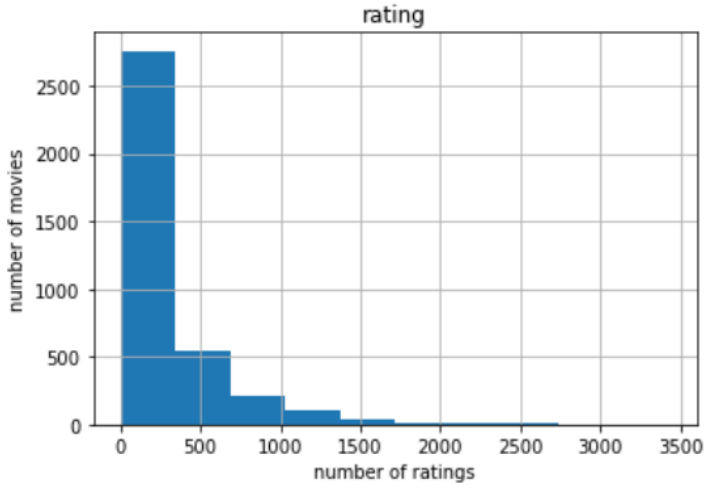
Problem 2: Expand our investigation to histograms

Below is the histogram that shows the ratings of all movies.

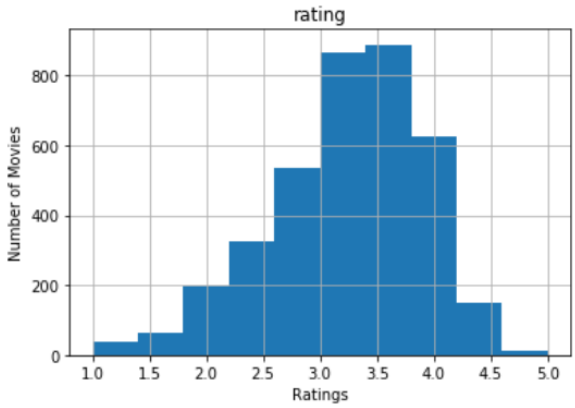


The y-label is the number of movies, the x-label is ratings. We have more than 50000 movies rated 1, more than 100000 movies rated 2, more than 250000 movies rated 3, almost 350000 movies rated 4 and approximately 255000 movies rated 5.

The histogram below shows the number of ratings each movie received.

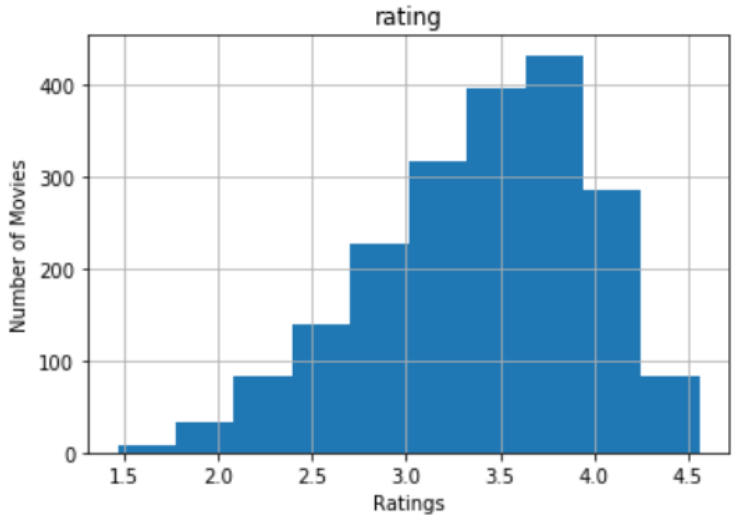


The x-label shows the number of ratings, the y-label shows the number of movies. So approximately 2750 movies have 350 ratings, 500 movies have 700 ratings, 200 movies have 1000 ratings, 100 movies have 1350 ratings, 50 movies have 1750 ratings, and less than 50 movies have more than 1750 ratings. Below is the histogram shows the average ratings for each movie. The total number of rated movies in this histogram is 3706.



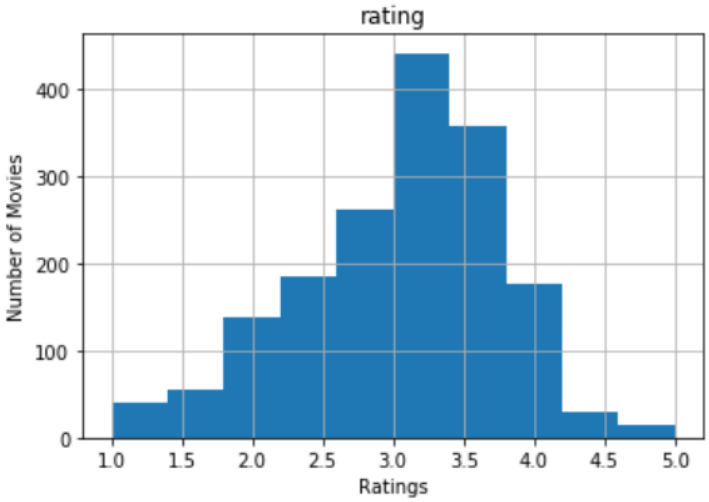
The x-label is the ratings, the y-label is the number of movies. Less than 50 movies receive 1.0~1.4 score, 50~100 movies receive 1.5 ~ 1.8 score. Approximately 200 movies receive 1.8~2.2 score. More than 300 movies receive 2.2~2.6 score. More than 500 movies receive 2.6 ~ 3.0 score. Almost 850 movies get 3.0~3.4 score. More than 850 movies get 3.4~3.75 score. More than 600 movies receive 3.75~4.2 score. Approximately 160 movies get 4.2~4.6 score. Less than 20 movies get 4.6~5.0 score.

The histogram below shows the average rating for movies rated more than 100 times. Total number of movies is 2006.



Based on the above histogram, the x-label is rating, and y-label is the number of movies. So less than 20 movies have 1.4~1.7 score. more than 30 but less than 50 movies have rating 1.7~2.1. Almost 80 movies have score in the range of 2.2~2.4. 140 movies have score 2.4~2.7. Approximately 225 movies have score between 2.7 to 3.0. 320 movies have rating between 3.0 to 3.3. Almost 390 movies have ratings between 3.4 to 3.6. Approximately 425 movies have rating 3.6~3.9. Around 280 movies have score 3.9 to 4.3. Around 70 movies have score 4.3 to 4.6.

The histogram below shows the average ratings for movies rated less than or equal 100 times. The total number of movies in the histogram is 1700.



The x-label is rating, and the y-label is the number of movies. Almost 40 movies have score 1.0~1.4. Around 50 movies have score between 1.4~1.8. Around 140 movies receive 1.8~2.2 score. Around 180 movies get 2.2~2.6. Around 260 movies get score 2.6~3.0. Approximately 440 movies receive 3.0~3.4 score. Around 350 movies get 3.4~3.8. More than 170 movies get 3.8~4.2. Around 25 movies get score 4.2~4.6. The rest receives score 4.6~5.0.

By observation about the tails of the histogram where we use all the movies versus the one where we use movies rated more than 100 times, I find histogram of movies with more than 100 reviews has fewer data points in the left and right tails; fewer outliers.

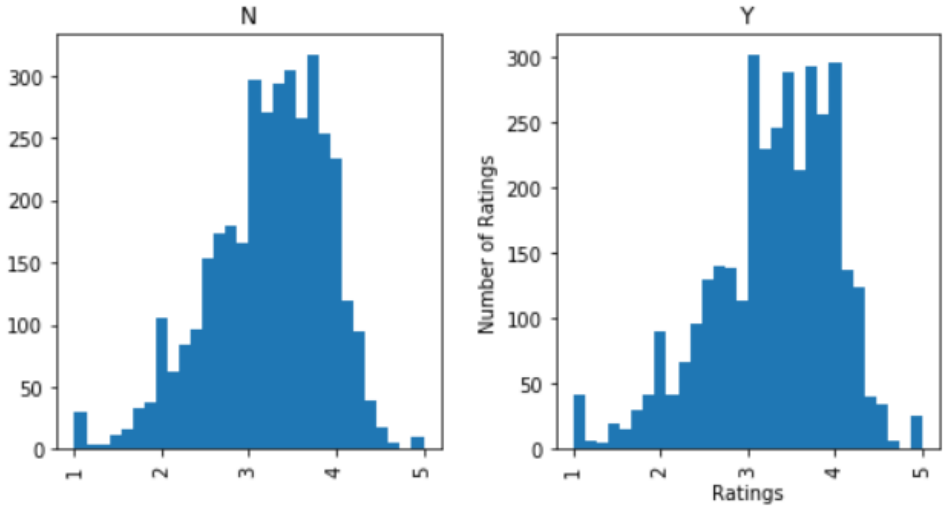
After observing the histograms of all the movies, movies rated more than 100 times, and movies rated less than 100 times, we would trust the movie ratings in the data frame where there are more than 100 reviews.

Movies with fewer than 100 reviews appear to have more outlier ratings and the larger the sample size, in general, the more accurate the results.

Conjectures

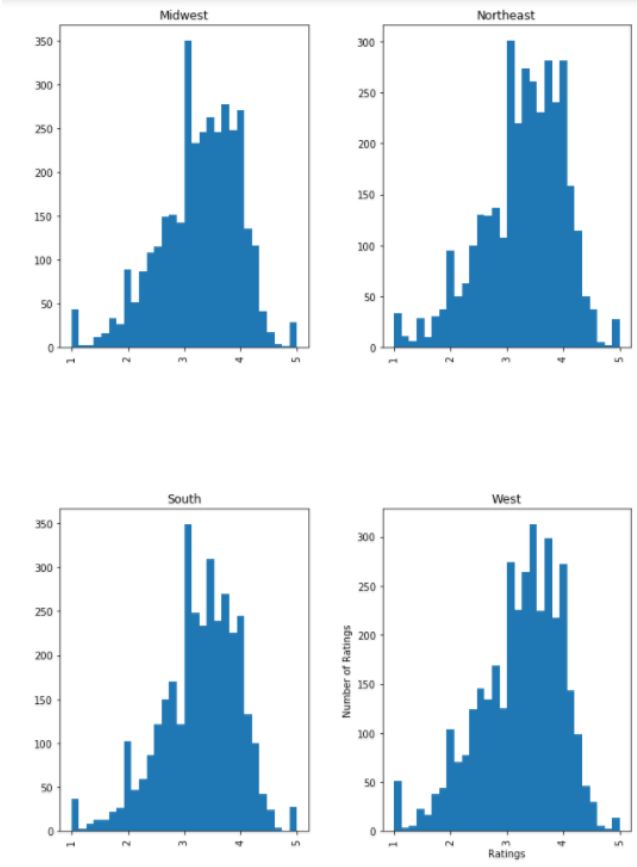
Based on the distribution of ratings, we figured out the following two conjectures.

Conjecture 1: Reviewers in Science, technology, engineering, education & math (STEEM) occupations have more extreme rating than non-STEEM occupations.
Below we plot two histograms that shows the difference between ratings from non-STEEM occupations (left), STEEM occupations(right), and which occupation is more likely to have extreme ratings 1 or 5?



Based on the figure above, our conjecture is right, movies rated by non-STEEM occupations have around 25 ratings for 1 score and 10 ratings for 5 score, while movies rated by STEEM occupations have approximately 45 ratings for 1 score, and 20 ratings for 5 score.

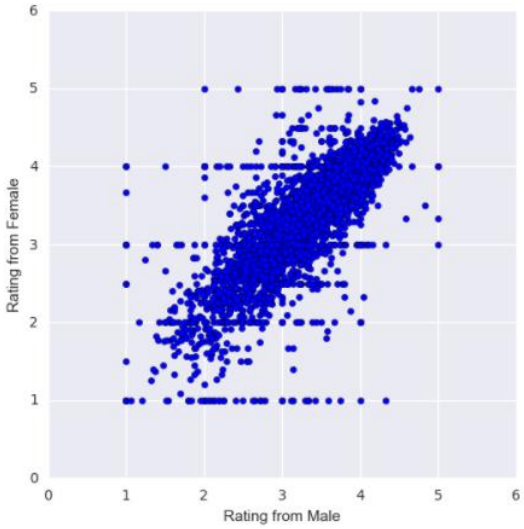
Conjecture 2: Reviewers from the West region are more likely to have extreme ratings than all other regions.
Below are four histograms of ratings from people from Midwest, Northeast, South and West.



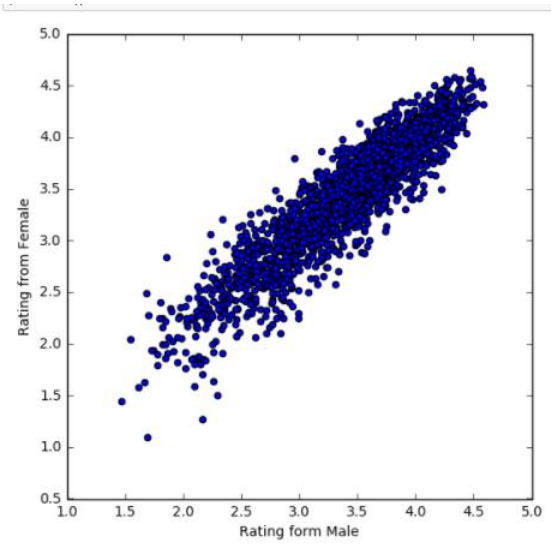
Based on these four figures, it is obvious that people from west and Midwest gives more 1 score (west 50 + 45 Midwest = 95) than people from other regions. But for 5 score it is not obvious. So, we can conclude people from West are more likely to give rating of 1, and we need more data and observations to decide if 5 is the same case, for the data we have now, it's not.

Problem 3: Correlation: Men versus women

For problem 3, we made a scatter plot of men versus women and their mean rating for every movie. The scatter plot shows a fuzzy line between the mean rating of men versus women, which shows there is a strong correlation between the mean rating of men versus women. The correlation coefficient between the average ratings of men and women is .76, which shows a strong positive correlation. The covariance between the average ratings of men and women is .51, which shows a positive medium correlation. We also show the number of men versus women and their mean rating for movies rated more than 200 times. This scatter plot is still fuzzy, but less fuzzy than the previous line, which shows a high correlation coefficient. The correlation coefficient between the average ratings of men and women where movies are rated more than 200 times is .920, which shows a very strong positive correlation. The covariance between the average ratings of men and women where movies are rated more than 200 times is .315, which shows a positive medium correlation.



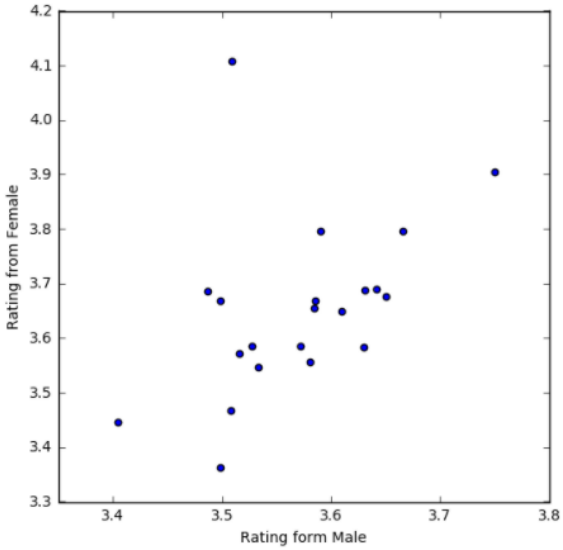
Scatter plot of men versus women and their mean rating for every movie



Scatter plot of men versus women and their mean rating for movies rated more than 200 times

Conjecture 1

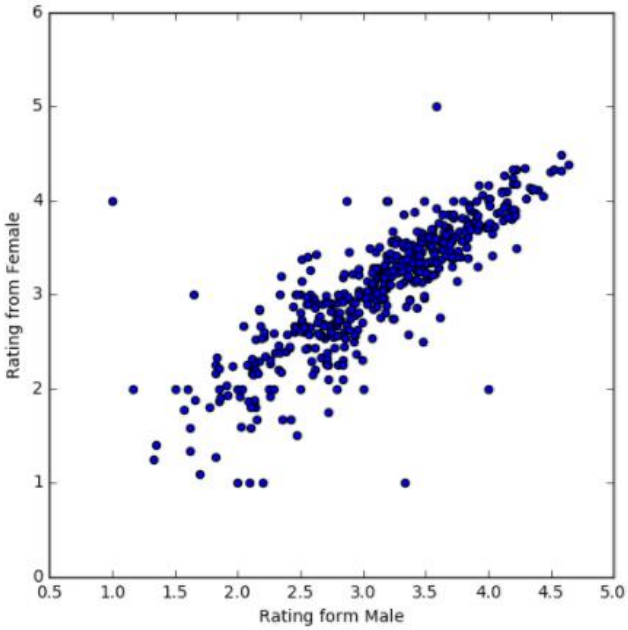
We plotted a scatter plot showing the movie ratings for men and women for different occupations as shown by the scatter plot. The scatter plot shows less similarity between the data points. The correlation coefficient is .437, which shows a medium positive relationship between movie for rating men and women for different occupations. We also plotted a scatter plot showing men versus women and their mean rating by Action movie genre. This shows a strong positive correlation coefficient, and the scatter plot shows a fuzzy straight line. The correlation coefficient in this case is .84.



Scatter plot of men versus women and their mean rating grouped by occupation description

Conjecture 2

For conjecture 2, we found that movie ratings for men and women are more similar for movies with the action genre. We plotted the movie ratings of men versus women and the data formed a fuzzy line. The correlation coefficient for the ratings between men and women is .84, which shows a strong positive correlation. While men and women do not always provide the same ratings for action movies, they are highly and positively correlated. Using the correlation coefficient, we calculated, you can find the average ratings for women for an action movie if you know what men rate the same action movie, and vice-versa.



Scatter plot of men versus women and their mean rating by Action movie genre

Problem 4: Business Intelligence

How can Movie Lens dataset be used to guide DreamWorks Production Studio devise a business strategy to climb up the ranking from 10th spot in Top 10 Production Houses in the United States?

https://images.google.com

DreamWorks Animation SKG, Inc. (more commonly known as DreamWorks Animation, or simply DreamWorks) is an American animation studio that is a subsidiary of Universal Studios, a division of NBCUniversal, itself a division of Comcast. It is based in Glendale, California and produces animated feature films, television programs and online virtual games. [4.1]

According to [4.2] and [4.3], DreamWorks stands last in the ratings for Top 10 Movie production houses in the Hollywood, with Time Warner and Sony grabbing the top 2 positions in the list. DreamWorks has currently released a total of 35 feature films, including the franchises Shrek, Madagascar, Kung Fu Panda, How to Train Your Dragon and Trolls [4.1]. However, to move up the list, the production house has to analyse and come up with investment avenues in movie genres that offer significant business growth and high revenue stream, and not just remain confined to Animation genre.

Our team believes that there is valuable information associated with the MovieLens data that can help DreamWorks achieve its objective. However, producing a movie isn't a small deal. According to an article on Business Insider [4.4], the total cost of making "Pirates of the Caribbean: At World's End" (2007) was $341 million. That is more than the **GDP of Federated States of Micronesia $322 million in 2016**. This comparison must have brought an idea about the sheer magnitude of risk and investment involved in producing a movie. Therefore, what the production house need is a small-scale understanding of a market based on which they can magnify their business objective on a large scale. In technical terms, they need a **Minimum Viable Product.**

According to Wikipedia [4.5], A Minimum Viable Product (MVP) is a product with just enough features to satisfy early customers, and to provide feedback for future product development. In the film industry, **a small feature film or trailer** can be considered an MVP as it can allow a production house to gather data related to customer reviews and ratings, and subsequently model their movie sequences to satisfy audience demands.

According to Wikipedia, there are **24 Literary genres** [4.6]. And the big question is, which genre should the production house target. To answer this question, we performed Exploratory Data Analysis on the MovieLens 1M Data Set and derived interesting conjectures and observations.

In Problem 1 of the Python Notebook, we answered a question, *"What are the ten most popular movies?",* and obtained the following results:

		rating	reviews
		mean	
title	genres		
American Beauty (1999)	Comedy|Drama	4.317386	3428
Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Fantasy|Sci-Fi	4.453694	2991
Star Wars: Episode V - The Empire Strikes Back (1980)	Action|Adventure|Drama|Sci-Fi|War	4.292977	2990
Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Romance|Sci-Fi|War	4.022893	2883
Jurassic Park (1993)	Action|Adventure|Sci-Fi	3.763847	2672
Saving Private Ryan (1998)	Action|Drama|War	4.337354	2653
Terminator 2: Judgment Day (1991)	Action|Sci-Fi|Thriller	4.058513	2649
Matrix, The (1999)	Action|Sci-Fi|Thriller	4.315830	2590
Back to the Future (1985)	Comedy|Sci-Fi	3.990321	2583
Silence of the Lambs, The (1991)	Drama|Thriller	4.351823	2578

Based on these results, we see that out of 10 movies, **7 movies are action genre movies.**

Our definition of "popular" movies was those which have the greatest number of people sharing their opinion about the movie. If a person shares their opinion about a movie we can assume that the person saw the movie. We consider each review to represent a viewing, and each viewing of a movie is a "vote" in favor of the movie being "popular".

To further strengthen our conjecture, we cross-validated our results with the data from **Box Office Mojo (http://www.boxofficemojo.com)** which shows domestic, wide-opening weekend gross dollars earned for each title. We observed that, **9 out of 10 titles on our "Ten Most Popular Movies" were the top grossing movies for their opening weekend.**

The results from Box Office Mojo enabled us to support our conclusion that **action movies** are the most popular movies since a high grossing movie represents that the audience paid money to see the action movie.

Additionally, in Problem 3, we derived and supported our conjecture that *"Movie Ratings for Men and Women are more similar for movies with the Action genre"* by calculating the correlation coefficient between them.

gender	F	M
gender		
F	1.000000	0.840376
M	0.840376	1.000000

A correlation coefficient of *0.84* for action genre as opposed to *0.76* across all the movies indicated that Men and Women, across all the age groups, have relatively similar opinion for action movies.

Moreover, in Problem 1, we derived and supported our conjecture that *"Movies associated with the 'Action' genre are more universally liked (smallest standard deviation) than any other movie genre".*

Based on the above conjectures, we concluded that Action Movies are the most popular genre.

However, when investment is in millions of dollars, it becomes imperative to support our decision based on analysis from other sources. Therefore, we made use of the **IMDB and YouTube Statistics** to map, compare, and validate our conclusion and conjectures.

There are some key assumptions that frame our analysis and they are articulated here:

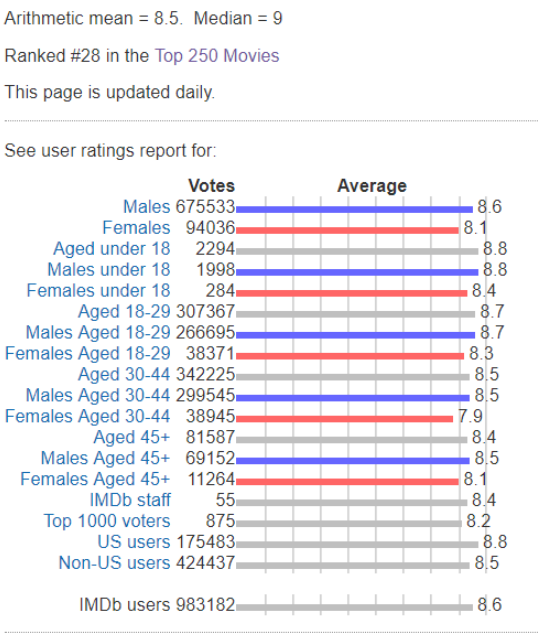
1) IMDB ratings are based on a *weighted average formula* which the website does not disclose to avoid tampering with ratings. Therefore, we used the basic weighted average formula to calculate the ratings.

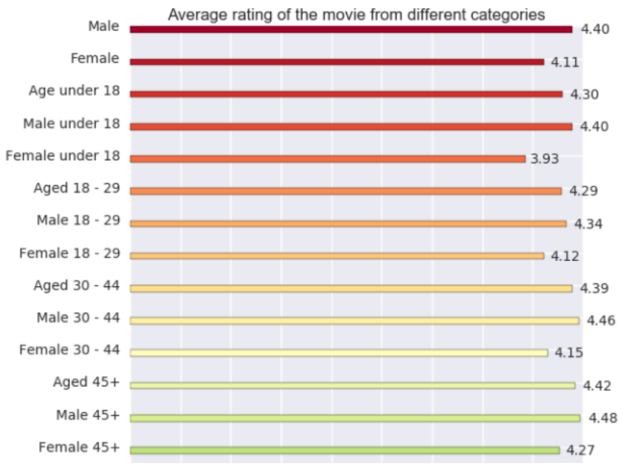
2) IMDB ratings are on a scale of 10, whereas the ratings we had were on a scale of 5. Therefore, our primary aim is to see if we are relatively accurate with respect to IMDB.

3) Since the list for Top 10 popular movies consist of movies made before YouTube was founded, we compared the YouTube Statistics for movie trailers uploaded in a relatively same time frame.

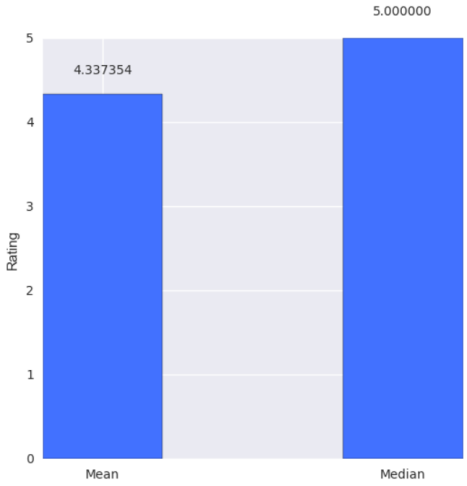
Our team selected reviews for ***Saving Private Ryan (1998)*** from the MovieLens dataset to test our conclusions. We performed analysis for Gender and Age based categories to exactly map and compare our results from those of IMDB.

Snapshots shown below compare the results **(Left: IMDB Stats | Right: MovieLens Stats)**





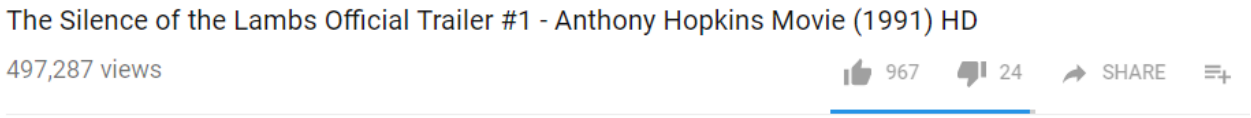
Snapshot below is a bar chart showing the mean and median ratings for ***Saving Private Ryan (1998)*** from MovieLens dataset.



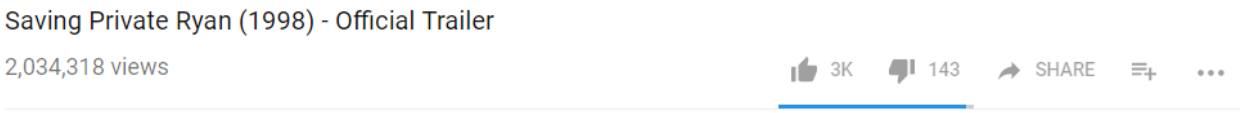
Since IMDB ratings are considered as a benchmark by the industry as well as the viewers, our rating analysis of action movie ***Saving Private Ryan (1998)*** correlated largely with the IMDB ratings.

From YouTube Statistics, our primary goal was to see if there is a relation between *more trailer views* and *Like/Dislike ratio* to a movie success *(Popular*) as Trailers are the **Minimum Viable Product** to be tested.

To compare, we chose ***The Silence of the Lambs (1991),*** a non-action movie, also a part of the Top 10 popular movies.



https://www.youtube.com/watch?v=W6Mm8Sbe__o&t=3s



https://www.youtube.com/watch?v=zwhP5b4tD6g&t=2s

We observed that the number of views for the action movie was approximately ***4.09 times*** more than the non-action movie. Even though the Likes /Dislikes ratio was lower for action movie as compared to non-action movie, it could probably be due to a significant difference between the views.

These results from YouTube helped as conclude that Trailers with more views correlated to a more popular movie. **(Saving Private Ryan is ranked 6th | The Silence of the Lambs ranked 9th)**

*Based on all the analysis we performed, we can recommend DreamWorks to produce an **Action Genre** movie and release a trailer that captures the attention of the audience. More captivating the trailer is, more are the views. More the views, more the probability of success of the movie.*

Scope for Improvement

1) Had there been more information related to **Cost incurred in movie production** and **Revenue generated,** we could have performed exploratory data analysis to see if there is any **trend or correlation** between movie ratings/reviews and **Return on Investement.**

2) Had there been more information related to the **number of distributors** for a movie, we could have performed exploratory data analysis to see if there is any trend or correlation between **number of distributors** and **popularity of a movie (number of reviews)**

3) Analysis of **Socio-Economic Status** such as Median Household Income could have helped us mine more interesting insights on how statistics like these can influence movie ratings or popolarity.

4) Analysis of **Geographical Distribution** for the reviews could have helped us mine any interesting correlation between ratings and area from where the ratings were given.

Conclusion

For case study 2, we downloaded 1 million reviews from MovieLens for analysis. We then merged all the data into a single Pandas Dataframe and stored the data into an HDF5 file. In problem 1, we reported the number of movies that have an average rating over 4.5 overall. We also reported how many movies have an average rating over 4.5 among men. Also, we reported the ten most popular movies. These are just a few of the statistics we reported for problem 1.

In our report, we showed the 9 out of 10 titles our “Ten Most Popular Movies” based on gross dollars for opening weekend. We gathered this data from www.boxofficemojo.com. We took initiative to gather this information on our own to support the reasoning we talked about in the business question. 7 out of the 10 movies in the top 10 were action movies, which further supports why we decided to recommend in our business question that a movie studio produce an action movie trailer.

For our business question, we are proposing that the movie studio create a trailer as a Minimum Viable Product to test their idea for a new action movie. The movie studio will use the trailer to gather consumer sentiment on whether to continue and make the capital investment to produce the movie. This method of using the trailer as a MVP reduces risk for the movie studio as if the trailer does not enough positive consumer sentiment, they can cancel the project or change movie sequences.

Citations

1.1) http://files.grouplens.org/datasets/movielens/ml-1m-README.txt

1.2) https://en.wikipedia.org/wiki/List_of_genres#Film_and_television_formats_and_genres

1.3) http://files.grouplens.org/datasets/movielens/ml-1m-README.txt

1.4) http://files.grouplens.org/datasets/movielens/ml-1m-README.txt

1.5) https://support.hdfgroup.org/HDF5/

4.1) https://en.wikipedia.org/wiki/DreamWorks_Animation

4.2) http://www.therichest.com/rich-list/the-biggest/the-10-biggest-hollywood-studios/

4.3) https://reelrundown.com/film-industry/Top-10-Movie-Production-Companies

4.4) http://www.businessinsider.com/most-expensive-movies-ever-2014-6/#2-cleopatra-1963-340-million-29

4.5) https://en.wikipedia.org/wiki/Minimum_viable_product

4.6) https://en.wikipedia.org/wiki/List_of_genres

Appendices

Appendix 1-A: Average rating greater than 4.5 star

All users		Male Users			Female Users		
title	rating	title	gender	rating	title	gender	rating
One Little Indian (1973)	5.000000	Angela (1995)	M	5.000000	24 7: Twenty Four Seven (1997)	F	5.000000
Schlafes Bruder (Brother of Sleep) (1995)	5.000000	Gate of Heavenly Peace, The (1995)	M	5.000000	Dancemaker (1998)	F	5.000000
Bittersweet Motel (2000)	5.000000	Ulysses (Ulisse) (1954)	M	5.000000	Raw Deal (1948)	F	5.000000
Ulysses (Ulisse) (1954)	5.000000	Smashing Time (1967)	M	5.000000	Prisoner of the Mountains (Kavkazsky Plennik) (1996)	F	5.000000
Follow the Bitch (1998)	5.000000	Small Wonders (1996)	M	5.000000	Song of Freedom (1936)	F	5.000000
Gate of Heavenly Peace, The (1995)	5.000000	Schlafes Bruder (Brother of Sleep) (1995)	M	5.000000	Other Side of Sunday, The (Søndagsengler) (1996)	F	5.000000
Song of Freedom (1936)	5.000000	Lured (1947)	M	5.000000	One Little Indian (1973)	F	5.000000
Lured (1947)	5.000000	Follow the Bitch (1998)	M	5.000000	Message to Love: The Isle of Wight Festival (1996)	F	5.000000
Baby, The (1973)	5.000000	Dangerous Game (1993)	M	5.000000	Lamerica (1994)	F	5.000000
Smashing Time (1967)	5.000000	Bells, The (1926)	M	5.000000	I Am Cuba (Soy Cuba/Ya Kuba) (1964)	F	5.000000
I Am Cuba (Soy Cuba/Ya Kuba) (1964)	4.800000	Baby, The (1973)	M	5.000000	Twice Upon a Yesterday (1998)	F	5.000000
Lamerica (1994)	4.750000	Time of the Gypsies (Dom za vesanje) (1989)	M	4.833333	Woman of Paris, A (1923)	F	5.000000
Apple, The (Sib) (1998)	4.666667	I Am Cuba (Soy Cuba/Ya Kuba) (1964)	M	4.750000	Gate of Heavenly Peace, The (1995)	F	5.000000
Sanjuro (1962)	4.608696	Lamerica (1994)	M	4.666667	Gambler, The (A Játékos) (1997)	F	5.000000
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	4.560510	Window to Paris (1994)	M	4.666667	For the Moment (1994)	F	5.000000
Shawshank Redemption, The (1994)	4.554558	Sanjuro (1962)	M	4.639344	Skipped Parts (2000)	F	5.000000
		Apple, The (Sib) (1998)	M	4.600000	Saltmen of Tibet, The (1997)	F	5.000000
		For All Mankind (1989)	M	4.583333	Belly (1998)	F	5.000000

Godfather, The (1972)	4.524966
Close Shave, A (1995)	4.520548
Usual Suspects, The (1995)	4.517106
Schindler's List (1993)	4.510417
Wrong Trousers, The (1993)	4.507937

Godfather, The (1972)	M	4.583333
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	M	4.576628
Shawshank Redemption, The (1994)	M	4.560625
Raiders of the Lost Ark (1981)	M	4.520597
Usual Suspects, The (1995)	M	4.518248

Big Combo, The (1955)	F	5.000000
Country Life (1994)	F	5.000000
Coldblooded (1995)	F	5.000000
Ayn Rand: A Sense of Life (1997)	F	5.000000
Clean Slate (Coup de Torchon) (1981)	F	5.000000
Brother, Can You Spare a Dime? (1975)	F	5.000000
Bittersweet Motel (2000)	F	5.000000
Ballad of Narayama, The (Narayama Bushiko) (1958)	F	5.000000
Battling Butler (1926)	F	5.000000
World of Apu, The (Apur Sansar) (1959)	F	4.842105
Hearts and Minds (1996)	F	4.833333
Apple, The (Sib) (1998)	F	4.750000
Rain (1932)	F	4.750000
Marcello Mastroianni: I Remember Yes, I Remember (1997)	F	4.666667
Among Giants (1998)	F	4.666667
Panther (1995)	F	4.666667
Aparajito (1956)	F	4.666667
Close Shave, A (1995)	F	4.644444
Firelight (1997)	F	4.600000
Before the Rain (Pred dozhdot) (1994)	F	4.600000
Flower of My Secret, The (La Flor de Mi Secreto) (1995)	F	4.600000
Wrong Trousers, The (1993)	F	4.588235
General, The (1927)	F	4.575758

		Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	F	4.572650
		Pather Panchali (1955)	F	4.571429
		Wallace & Gromit: The Best of Aardman Animation (1996)	F	4.563107
		Schindler's List (1993)	F	4.562602
		Grand Illusion (Grande illusion, La) (1937)	F	4.560976
		Shawshank Redemption, The (1994)	F	4.539075
		Grand Day Out, A (1992)	F	4.537879
		To Kill a Mockingbird (1962)	F	4.536667
		Creature Comforts (1990)	F	4.513889
		Usual Suspects, The (1995)	F	4.513317

Appendix 1-B: Women and Men over 30 with Median Movie Rating > 4.5

Movies with a median rating over 4.5 among Men over age 30		
42 Up (1998)	M	5.0
All Quiet on the Western Front (1930)	M	5.0
American Beauty (1999)	M	5.0
Among Giants (1998)	M	5.0
Angela (1995)	M	5.0
Ayn Rand: A Sense of Life (1997)	M	5.0
Bells, The (1926)	M	5.0
Bicycle Thief, The (Ladri di biciclette) (1948)	M	5.0
Boat, The (Das Boot) (1981)	M	5.0
Bridge on the River Kwai, The (1957)	M	5.0
Carmen (1984)	M	5.0
Casablanca (1942)	M	5.0
Chinatown (1974)	M	5.0
Chushingura (1962)	M	5.0
Citizen Kane (1941)	M	5.0
City Lights (1931)	M	5.0
Close Shave, A (1995)	M	5.0
Conformist, The (Il Conformista) (1970)	M	5.0
Creature Comforts (1990)	M	5.0
Double Indemnity (1944)	M	5.0
Dr. Strangelove or: How I Learned to Stop Worry...	M	5.0
Ed's Next Move (1996)	M	5.0
Fargo (1996)	M	5.0
Firelight (1997)	M	5.0
Follow the Bitch (1998)	M	5.0
For All Mankind (1989)	M	5.0
Gambler, The (A Játékos) (1997)	M	5.0
General, The (1927)	M	5.0
Godfather, The (1972)	M	5.0
Godfather: Part II, The (1974)	M	5.0
Grand Illusion (Grande illusion, La) (1937)	M	5.0
Grapes of Wrath, The (1940)	M	5.0
Grateful Dead (1995)	M	5.0
Hearts and Minds (1996)	M	5.0
I Am Cuba (Soy Cuba/Ya Kuba) (1964)	M	5.0
Inherit the Wind (1960)	M	5.0
It's a Wonderful Life (1946)	M	5.0
Jean de Florette (1986)	M	5.0
Jupiter's Wife (1994)	M	5.0
Lamerica (1994)	M	5.0
Lawrence of Arabia (1962)	M	5.0
Lured (1947)	M	5.0
Maltese Falcon, The (1941)	M	5.0
Manon of the Spring (Manon des sources) (1986)	M	5.0
Modern Times (1936)	M	5.0
North by Northwest (1959)	M	5.0
One Flew Over the Cuckoo's Nest (1975)	M	5.0
Palm Beach Story, The (1942)	M	5.0
Paralyzing Fear: The Story of Polio in America,...	M	5.0
Pather Panchali (1955)	M	5.0
Paths of Glory (1957)	M	5.0
Patton (1970)	M	5.0
Psycho (1960)	M	5.0
Raiders of the Lost Ark (1981)	M	5.0
Rear Window (1954)	M	5.0
Return with Honor (1998)	M	5.0
Sanjuro (1962)	M	5.0
Saving Private Ryan (1998)	M	5.0
Schindler's List (1993)	M	5.0
Schlafes Bruder (Brother of Sleep) (1995)	M	5.0
See the Sea (Regarde la mer) (1997)	M	5.0
Seven Chances (1925)	M	5.0
Seven Samurai (The Magnificent Seven) (Shichini...	M	5.0
Seventh Seal, The (Sjunde inseglet, Det) (1957)	M	5.0
Shawshank Redemption, The (1994)	M	5.0
Silence of the Lambs, The (1991)	M	5.0
Sixth Sense, The (1999)	M	5.0

Small Wonders (1996)	M	5.0
Smashing Time (1967)	M	5.0
Star Wars: Episode IV - A New Hope (1977)	M	5.0
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	M	5.0
Third Man, The (1949)	M	5.0
Tigrero: A Film That Was Never Made (1994)	M	5.0
Time of the Gypsies (Dom za vesanje) (1989)	M	5.0
To Kill a Mockingbird (1962)	M	5.0
To Live (Huozhe) (1994)	M	5.0
Treasure of the Sierra Madre, The (1948)	M	5.0
Two or Three Things I Know About Her (1966)	M	5.0
Usual Suspects, The (1995)	M	5.0
Vampyros Lesbos (Las Vampiras) (1970)	M	5.0
Wallace & Gromit: The Best of Aardman Animation...	M	5.0
Window to Paris (1994)	M	5.0
Wizard of Oz, The (1939)	M	5.0
Wrong Trousers, The (1993)	M	5.0
Yojimbo (1961)	M	5.0
Young Frankenstein (1974)	M	5.0

Movies with a median rating over 4.5 among Women over age 30		
24 7: Twenty Four Seven (1997)	F	5.0
400 Blows, The (Les Quatre cents coups) (1959)	F	5.0
Across the Sea of Time (1995)	F	5.0
African Queen, The (1951)	F	5.0
After Life (1998)	F	5.0
Amadeus (1984)	F	5.0
Among Giants (1998)	F	5.0
Aparajito (1956)	F	5.0
Apple, The (Sib) (1998)	F	5.0
Ballad of Narayama, The (Narayama Bushiko) (1958)	F	5.0
Before the Rain (Pred dozhdot) (1994)	F	5.0
Belly (1998)	F	5.0
Best Man, The (Il Testimone dello sposo) (1997)	F	5.0
Bicycle Thief, The (Ladri di biciclette) (1948)	F	5.0
Big Combo, The (1955)	F	5.0
Blue in the Face (1995)	F	5.0
Bogus (1996)	F	5.0
Bridge on the River Kwai, The (1957)	F	5.0
Brother, Can You Spare a Dime? (1975)	F	5.0
Buena Vista Social Club (1999)	F	5.0
Casablanca (1942)	F	5.0
Christmas Story, A (1983)	F	5.0
Cinema Paradiso (1988)	F	5.0
Citizen Kane (1941)	F	5.0
City Lights (1931)	F	5.0
Clean Slate (Coup de Torchon) (1981)	F	5.0
Cleo From 5 to 7 (Cléo de 5 à 7) (1962)	F	5.0
Close Shave, A (1995)	F	5.0
Cold Fever (Á köldum klaka) (1994)	F	5.0
Conformist, The (Il Conformista) (1970)	F	5.0
Country Life (1994)	F	5.0
Creature Comforts (1990)	F	5.0
Croupier (1998)	F	5.0
Dancemaker (1998)	F	5.0
Dark Command (1940)	F	5.0
Dr. Strangelove or: How I Learned to Stop Worry...	F	5.0
Dreaming of Joseph Lees (1998)	F	5.0
Eighth Day, The (Le Huitième jour) (1996)	F	5.0
Enchanted April (1991)	F	5.0
Everest (1998)	F	5.0
Eyes Without a Face (1959)	F	5.0
Faraway, So Close (In Weiter Ferne, So Nah!) (1...	F	5.0
Fargo (1996)	F	5.0
Firelight (1997)	F	5.0
Flower of My Secret, The (La Flor de Mi Secreto...	F	5.0
Force of Evil (1948)	F	5.0
Funeral, The (1996)	F	5.0
Gambler, The (A Játékos) (1997)	F	5.0

Garden of Finzi-Contini, The (Giardino dei Finz...	F	5.0
Gaslight (1944)	F	5.0
General, The (1927)	F	5.0
Godfather, The (1972)	F	5.0
Gone with the Wind (1939)	F	5.0
Grand Day Out, A (1992)	F	5.0
Grand Illusion (Grande illusion, La) (1937)	F	5.0
Grandfather, The (El Abuelo) (1998)	F	5.0
Green Mile, The (1999)	F	5.0
Haunted World of Edward D. Wood Jr., The (1995)	F	5.0
Hear My Song (1991)	F	5.0
Hearts and Minds (1996)	F	5.0
Idiots, The (Idioterne) (1998)	F	5.0
Innocents, The (1961)	F	5.0
It Happened One Night (1934)	F	5.0
It Takes Two (1995)	F	5.0
It's a Wonderful Life (1946)	F	5.0
Jupiter's Wife (1994)	F	5.0
Kagemusha (1980)	F	5.0
Killing Fields, The (1984)	F	5.0
King of Masks, The (Bian Lian) (1996)	F	5.0
Lamerica (1994)	F	5.0
Lawn Dogs (1997)	F	5.0
Lawrence of Arabia (1962)	F	5.0
Legend of 1900, The (Leggenda del pianista sull...	F	5.0
Local Hero (1983)	F	5.0
Love Serenade (1996)	F	5.0
Lucie Aubrac (1997)	F	5.0
Maltese Falcon, The (1941)	F	5.0
Man Facing Southeast (Hombre Mirando al Sudeste...	F	5.0
Manon of the Spring (Manon des sources) (1986)	F	5.0
Marcello Mastroianni: I Remember Yes, I Remembe...	F	5.0
Microcosmos (Microcosmos: Le peuple de l'herbe)...	F	5.0
My Fair Lady (1964)	F	5.0
Nights of Cabiria (Le Notti di Cabiria) (1957)	F	5.0
North by Northwest (1959)	F	5.0
Notorious (1946)	F	5.0
Official Story, The (La Historia Oficial) (1985)	F	5.0
Only Angels Have Wings (1939)	F	5.0
Other Side of Sunday, The (Søndagsengler) (1996)	F	5.0
Palookaville (1996)	F	5.0
Paradine Case, The (1947)	F	5.0
Paradise Lost: The Child Murders at Robin Hood ...	F	5.0
Paris, France (1993)	F	5.0
Paths of Glory (1957)	F	5.0
Perils of Pauline, The (1947)	F	5.0
Philadelphia Story, The (1940)	F	5.0
Prisoner of the Mountains (Kavkazsky Plennik) (...	F	5.0
Producers, The (1968)	F	5.0
Promise, The (La Promesse) (1996)	F	5.0
Psycho (1960)	F	5.0
Quiet Room, The (1996)	F	5.0
Raiders of the Lost Ark (1981)	F	5.0
Rain (1932)	F	5.0
Ran (1985)	F	5.0
Raw Deal (1948)	F	5.0
Rear Window (1954)	F	5.0
Rebecca (1940)	F	5.0
Red Sorghum (Hong Gao Liang) (1987)	F	5.0
Saltmen of Tibet, The (1997)	F	5.0
Sanjuro (1962)	F	5.0
Schindler's List (1993)	F	5.0
Seven Samurai (The Magnificent Seven) (Shichini...	F	5.0
Shadow of a Doubt (1943)	F	5.0
Shawshank Redemption, The (1994)	F	5.0
Singin' in the Rain (1952)	F	5.0
Sixth Sense, The (1999)	F	5.0
Skipped Parts (2000)	F	5.0
Sling Blade (1996)	F	5.0
Some Folks Call It a Sling Blade (1993)	F	5.0
Some Like It Hot (1959)	F	5.0

Song of Freedom (1936)	F	5.0
Sound of Music, The (1965)	F	5.0
Source, The (1999)	F	5.0
Star Wars: Episode IV - A New Hope (1977)	F	5.0
Still Breathing (1997)	F	5.0
Substance of Fire, The (1996)	F	5.0
Sum of Us, The (1994)	F	5.0
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	F	5.0
Sunset Strip (2000)	F	5.0
Taste of Cherry (1997)	F	5.0
Terrorist, The (Malli) (1998)	F	5.0
Third Man, The (1949)	F	5.0
Three Days of the Condor (1975)	F	5.0
To Kill a Mockingbird (1962)	F	5.0
To Live (Huozhe) (1994)	F	5.0
Trial, The (Le Procès) (1963)	F	5.0
Trouble in Paradise (1932)	F	5.0
Two Women (La Ciociara) (1961)	F	5.0
Two or Three Things I Know About Her (1966)	F	5.0
Usual Suspects, The (1995)	F	5.0
Vertigo (1958)	F	5.0
Wallace & Gromit: The Best of Aardman Animation...	F	5.0
What Happened Was... (1994)	F	5.0
Window to Paris (1994)	F	5.0
Wizard of Oz, The (1939)	F	5.0
Woman of Paris, A (1923)	F	5.0
World of Apu, The (Apur Sansar) (1959)	F	5.0
Wrong Trousers, The (1993)	F	5.0
Yojimbo (1961)	F	5.0
Young Frankenstein (1974)	F	5.0