Seen any Good Movies Lately?

Understanding Film Preferences for Recommendations Solely on Background Information

Qiuyi (Helen) Hong, Yanhong (Haley) Huang, Tom Meagher, and Tyler Reese

INTRODUCTION− It's a question we hear all the time during lulls in conversation− "Seen any good movies lately?" The subtext of this question, of course, is "Have you seen any movies that I would like to see?" Answering this question is not so easy. We know what movies that *we ourselves* enjoy, but how are we to know what movies another person might prefer? We don't know what movies they like, or even what movies they've *seen* for that matter. The question is more daunting than its face value. We certainly don't want to disappoint our friends, nor lead them into a two-hour wasted investment of their precious free time. Yet we must attempt to recommend a movie based solely on their personal background: age, gender, geographic location, and occupation for example.

This question extends far beyond maintaining personal social relationships. Movie streaming businesses, such as Netflix®, acquire new customers every day. Each time a new member logs-in for the first time, Netflix® must recommend a movie to that new user *based solely on login demographic information*. Ensuring a positive first experience is vital in maintaining and growing a satisfied subscriber-base, making this question all the more important.

How do we recommend movies to others based solely on their personal information? This question seems both incredibly difficult and complicated to answer, yet surprisingly relevant both socially for individuals, and economically for direct-to-consumer movie services. A question that is both difficult yet relevant is precisely one that is *interesting.* Moreover it's a question worth *asking* and, as we will do in this report, worth considering how to *answer* (or, at the very least, answer up to some reasonable level of certainty). We will begin to explore how to answer this question through the following four objectives:

Objective 1: Collect data and analyze basic details. We will acquire a data set which is appropriate for considering this question, and begin to explore and analyze some of its basic attributes.

Objective 2: Extending our investigations to histograms. Beyond analyzing standard statistics, we consider cumulative properties of the data. This leads to studying various relevant distributions, which can illustrate how certain groups or types of people tend to feel towards movies.

Objective 3: Men vs. Women. Likely the simplest differentiator between individuals, due to its binary nature, is gender. Understanding movie preferences between genders, particularly when the preferences of one gender can be predicted from those of the other gender, is an excellent first step in answering this question.

Objective 4: Business Intelligence. We direct our attention to the first-time user problem for movie streaming companies. We consider one specific facet of this larger problem: what is the best *genre* of move to recommend to a first-time user? We will explore this question by asking a very specific series of sub-questions.

.

OBJECTIVE 1: Collect data and analyze basic statistics

Data: This report uses the *MovieLens 1 Million Dataset*, made available by Grouplens: http://grouplens.org/datasets/movielens/. This data poses the advantage that it has already been cleaned and organized, and is ready for analysis. Grouplens offers four packages of this data set, containing 100 thousand or 1, 10, and 22 Million movie reviews. Based on various technical and functional constrains, we determined that the 1 Million data set contains sufficient quantities of data, but is most computationally feasible for the group to analyze and report. The data set is comprised of three data files, *users*, *movies*, and *ratings.* *Users* contains the profile information for 6,040 users who submitted ratings – including a user ID, as well as gender, age, occupation and zip code. *Movies* contains basic information of the 3,883 movies that were rated– including a movie ID, title and genres. Finally *ratings* contains data about each of the 1 Million movie ratings submitted– including the user and movie ID's, the rating and a timestamp.

1. Import, merge, and store data. Each of the three data files described above were downloaded directly from GroupLens. Each individual file was read into Python as a Pandas Data Frame. Thanks to the relational nature of these files, the data was merged into a single Data Frame containing pertinent information for every movie rating. This main Data Frame was then stored in an HDF5 file. An example of the first 5 rows of this data frame appears below.

Figure 1.1: Data Sample

	User ID	Movie ID	Rating	Time Stamp	Gender	Age	Occ.	Zip	Title	Genres
0	1	1193	5	978300760	F	1	10	48067	One Flew Over…(1975)	Drama
1	2	1193	5	978298413	M	56	16	70072	One Flew Over…(1975)	Drama
2	12	1193	4	978220179	M	25	12	32793	One Flew Over…(1975)	Drama
3	15	1193	4	978199279	M	25	7	22903	One Flew Over…(1975)	Drama
4	17	1193	5	978158471	M	50	1	95350	One Flew Over…(1975)	Drama

2. Basic statistics and analysis of data collected.

How many movies have an average rating over 4.5 overall?
First, we considered the number of movies with an average rating above 4.5 overall (out of a best-possible 5). Using a Pandas pivot table, we determined there were **29** movies with an average rating of at least 4.5. Five examples including movie name and overall rating are displayed below.

Table 1.1: Sample Movies with Average Rating Over 4.5

Title	Average Rating
Apple, The (Sib) (1998)	4.67
Baby, the (1973)	5.00
Bells, The (1926)	4.50
Bittersweet Motel (2000)	5.00
Callejon de los Milagros, El (1995)	4.50

How many movies have an average rating over 4.5 among men? Among women?

These rating were submitted by both men and women who, likely, have differing tastes in movies. Thus while only 29 movies had an average rating of 4.5 overall, many movies may receive high average ratings amongst a specific gender. We counted the number movies with an average rating over 4.5 among men and women (again with a pivot table). There were **29** movies satisfying a men's average rating above 4.5, however among women's ratings, there were **70** movies with an average rating over 4.5. Five examples of each, including the average rating amongst the other gender, are shown below.

Table 1.2: Movies with Average Rating Over 4.5 among Men

Title	Average Rating by Men	Average Rating by Women
Angela (1995)	5.00	3.00
Apple, The (1998)	4.60	4.75
Baby, The (1973)	5.00	NaN
Bells, The (1926)	5.00	4.00
Callejon … (1995)	4.50	NaN

Table 1.3: Movies with Average Rating Over 4.5 among Women

Title	Average Rating by Women	Average Rating by Men
24:7 (1997)	5.00	3.75
Among Giant (1998)	4.67	3.33
Aparajito (1956)	4.67	3.86
Apple, The (1998)	4.75	4.60
Arguing the…(1996)	4.50	3.78

These results do not indicate that women are "easy to please" where as men or more difficult. Among the users submitting ratings, 4,331 are males while only 1,709 are females. Accordingly, over 750 thousand ratings were submitted by men, compared to only 250 thousand by women. Thus while there are 70 movies with high average ratings by women (more than twice the number for men, at 29), these ratings are being masked in the overall average ratings by the dominant number of male ratings. This is why the number of movies with an overall average of 4.5, and an average of 4.5 among men, are so similar- the comparatively large number of male ratings is driving these averages.

How many movies have a median rating over 4.5 among men over age 30? Among women over age 30?

We used a pivot table to collect all movies (by title) with a median rating above 4.5 among both men and women whose age is over 30. There were **105** movies with a median above 4.5 among men over 30, and **187** qualified movies among women's rating. Five examples of each (including the median among the opposite gender) are shown below.

Table 1.4: Movies with Median Rating Above 4.5 among Men Over 30

Title	Median Rating by Men	Median Rating by Women
42 Up (1998)	5.00	4.00
All Quiet … (1930)	5.00	4.00
American Beauty (1999)	5.00	4.00
Among Giants (1998)	5.00	5.00
Angela (1995)	5.00	3.00

Table 1.5: Movies with Median Rating Above 4.5 among Women Over 30

Title	Median Rating by Women	Median Rating by Men
24: 7: (1997)	5.00	3.00
400 Blows… (1959)	5.00	4.00
Above the Rim (1994)	4.50	3.00
Across the Sea … (1995)	5.00	NaN
African Queen, The (1951)	5.00	4.00

Observe that many of these "popular" movies (in terms of median) have a median rating of 5.0. In fact, this means that these movies have *all* ratings of 5.0. This doesn't indicate overwhelming success of the movie, as these calculations do not factor in the *number* of times each film was rated. This will be explored further in the second objective.

What are the ten most popular movies in the data set?

Next we consider the most popular movies in this data set. According to Merriam-Webster[1], "Popular" is defined as:

> *1: liked or enjoyed by many people*
> *2: accepted, followed, used, or done by many people*
> *3: of, relating to, or coming from most of the people in a country, society, or group*

Therefore we define a "Popular" movie in a way consistent with these three conditions. A movie is "Popular" if:

[1] http://www.merriam-webster.com/dictionary/popular

1 – It has above-average ratings amongst both men and women (i.e. an average rating among men which is higher than the average of *all* ratings given by men)

2 – The total number of ratings is higher than the average number of ratings per movie

Our interest is the most popular movie *in this data set.* Therefore the "society" or "group" (Merriam-Webster condition 3) in consideration is specifically the individuals who submitted ratings. This is some (small) subset of the U.S. population. Due to sampling bias, these ratings are certainly not reflective of the world population, and likely not the U.S. population. Therefore these movies are "Popular" in the sense that they are the most popular of the 3,900 movies rated, *among the individuals who submitted ratings.* Once we determined the "Popular" movies given this definition, the 10 *most* popular were chosen based on highest average overall rating.

Table 1.6: Top 10 Popular Movies

Rank	Title	Average Rating	Total Number of Ratings
1	Seven Samurai (The Magnificent Seven) (1954)	4.56	628
2	Shawshank Redemption, The (1994)	4.55	2227
3	Godfather, The (1972)	4.52	2223
4	Close Shave, A (1995)	4.52	657
5	Usual Suspects, The (1995)	4.52	1783
6	Schindler's List (1993)	4.51	2304
7	Wrong Trousers, The (1993)	4.51	882
8	Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	4.49	470
9	Raiders of the Lost Ark (1981)	4.48	2514
10	Rear Window (1954)	4.48	1050

Conjectures about how easy various groups are to please.

Before we even recommend a movie, our success or failure could depend on the inherent *happiness* level in various groups of people: how easy they are to please. We propose 2 distinct ideas.

Conjecture 1: The older a person gets, the more difficult they are to please.

As a person grows in years and life experience, it becomes more and more difficult to impress them, or show them something "new." Therefore we postulate that as a person ages, they are more difficult to please. Within movie data, this will be reflected by lower ratings from higher age groups.

First, we computed the average among all ratings provided by each age group:

Table 1.7: Average Rating per Age Group

Age	1	18	25	35	45	50	56
Average Rating	3.550	3.508	3.545	3.618	3.638	3.715	3.767

As seen in Table 1.7, there is actually an *increase* in average rating in higher age groups. While this difference of 0.25 in the average is not enough to support or refute our hypothesis, it suggests that there may be a difference in overall ratings among age groups, and thus we investigate further.

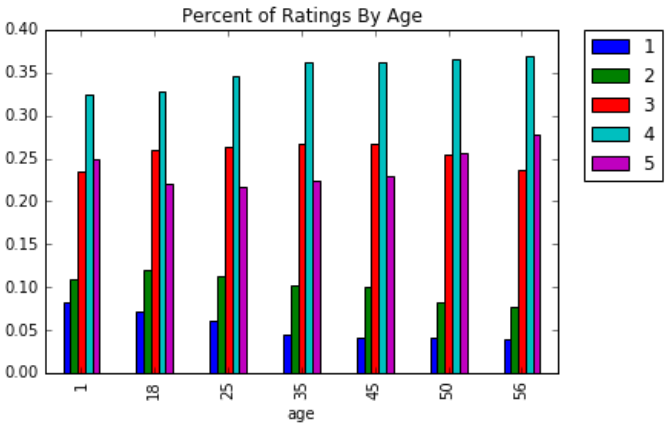
Beyond average ratings, a better measure of how "easy" or "difficult" it is to please a group of people is *percent* of high and low ratings given. That is, if a group is difficult to please, the percentage of *low* ratings out of *all* ratings submitted by that group will be higher than a group that is easier to

please. Similarly, a difficult group will submit a lower total fraction of *high* ratings. The following table and plot displays the percentage of each rating (1 – 5) submitted by each age group.

Table 1.8 Percentage of Age Group Rating

Age Group	1	2	3	4	5
1-17	8.2%	11.0%	23.4%	32.4%	25.0%
18-24	7.1%	12.0%	25.9%	32.8%	22.1%
25-34	6.0%	11.3%	26.4%	34.6%	21.7%
35-44	4.6%	10.2%	26.6%	36.2%	22.5%
45-49	4.1%	10.1%	26.7%	36.3%	22.9%
50-55	4.1%	8.3%	25.5%	36.5%	25.7%
56+	4.0%	7.7%	23.6%	36.9%	27.8%

Figure 1.2 Distribution of Age Group Ratings



These results confirm the general trend observed in the average ratings. Of those ratings submitted by individuals over the age of 56, 4% of those ratings were 1 whereas 27.8% of them were 5. On the other hand, among 18-24 year olds, 7.1% of ratings were 1 whereas only 22.1% were 5. That is, the oldest age groups submitted 5% more ratings of **5**, and 3% fewer ratings of **1** than did the 18-24 age group. The plot in Figure 1.2 demonstrates that this, in fact, appears to be a trend: in general as age increases, the percentage of both 4 and 5 ratings increases whereas the percentage of both 1 and 2 ratings decreases.

Conclusion: Given this movie-rating data, this conjecture appears to be false. In fact, our data supports a trend that is just the opposite: an older person is more likely to rate a movie as 5 and less likely to rate a movie as 1 than is a younger person.

Conjecture 2: Tired people are easier to please

As it gets later in the day (or early in the morning), it seems logical that individuals are less likely to be carefully and critically analyzing the world, but more likely simply relaxing. As a result, the more tired a person becomes, the easier they are to please.

Each movie rating in the data set includes a timestamp, which we converted to local time and then extracted the hour. As in the first conjecture, we first calculated average movie ratings submitted in each hour. These averages ranged between 3.52 and 3.62, showing no discernible pattern in relation to the hour. Therefore we again consider the *percentage* of ratings submitted at each hour. For analysis purposes, we considered the hours between 11PM and 5AM to be the time window when most people are tired. We then calculate the percentage of 1 and 5 ratings among all ratings submitted in this "tired" window. The table below compares these percentages to those of all ratings.

Table 1.9 Percentage of 1 and 5 ratings by time of day

Time	Rating of 1	Rating of 5
11PM-5AM	0.058	0.224
Total	0.056	0.226

Conclusion: Our data does not support this conjecture, not indicating that people watching movies in the middle of the night are "easier" to please. The percentage of high and low rating submitted are very similar to those submitted at all other times of the day. Note, however, that neither does it support the *opposite* conclusion (i.e. tired people are not more difficult to please).

OBJECTIVE 2: Expand our investigation to histograms

The inferences drawn in the previous objective were largely based on averages and percents. One shortcoming of such inferences is the ignorance of the *number* of times each movie was rated. Regardless of rating numbers, a movie with 1,000 ratings is surely considered more "popular" than a movie only viewed once. Therefore we expand our analysis to consider how many times various movies were rated, relative to various predictors in our data set. An excellent way to visualize such cumulative data is using histograms.

Plot a histogram of the ratings of all movies.

First, we tabulate the *total* number of each type of rating (1-5) submitted in the data set. This can be summarized using a histogram on all movie ratings. This histogram can be found in Figure 2.1 below.

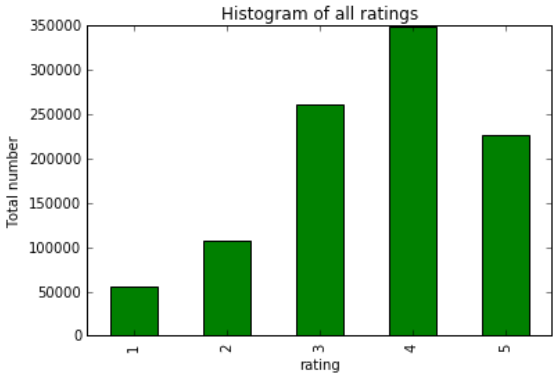


Figure 2.1: All Movie Ratings

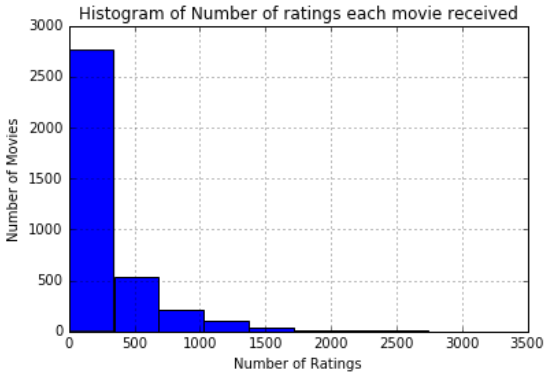


Figure 2.2: Number of Ratings Each Movie Received

Plot a histogram of the ratings each movie received.

On the other hand, we consider the per-movie *quantity* of ratings: how many ratings each movie received in the data set. This is akin to understanding how many times each movie has been *viewed* within the data set. A histogram of this data can be found in Figure 2.2, above. In this case, the horizontal axis represents the number of ratings received by a movie, and the vertical axis represents the total number of movies with that number of ratings.

Plot a Histogram of the average rating for each movie.

In addition to knowing the *quantity* of ratings for each movie, we are also interested in the *quality* of those ratings. Therefore we calculate the average overall rating for each movie, displayed as a histogram in Figure 2.3 below. The horizontal axis now represents the average rating, while the vertical axis represents the number of movies with that average rating.

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

Next, we combine the two previous measures, and consider the *quality* of ratings in terms of *quantity*. In particular, we apply a fixed quantification limit− in this case 100 ratings−eliminating movies with few numbers of ratings from analysis, and then once again consider the average rating of each movie. A histogram summarizing this data can be found in Figure 2.4, below. The reasoning for this quantity cut-off is explained by the following two questions.

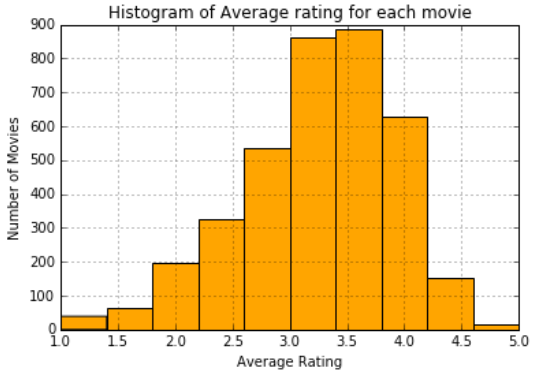


Figure 2.3: Average Movie Ratings

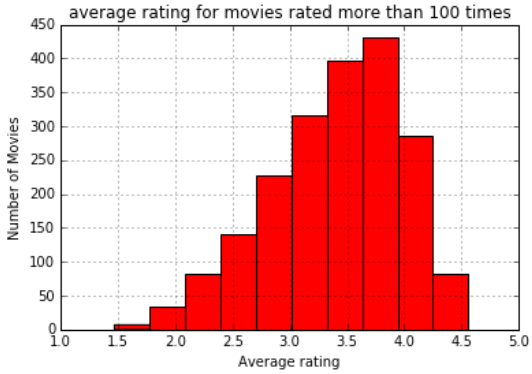


Figure 2.4: Average Movie Ratings of Movies with over 100 Ratings

What do you observe about the tails of the histograms?

Figure 2.3 displays a histogram of average ratings across all movies, whereas Figure 2.4 restricts to movies with at least 100 total ratings. The tails of the histograms are noticeably different. When considering all movies, there were a number of titles with average ratings between 1 and 1.5, as well as between 4.5 and 5. When we restrict to movies with at least 100 ratings, all such extreme averages disappear−all averages range between 1.5 and 4.5. This makes sense: if a movie is only rated once, and given a 1, then it has an average rating of 1. On the other hand, when a movie has been viewed many times, in order to achieve an average rating of 1 it needs to receive a very large percent of 1 ratings. For this reason, when we restrict to movies viewed more than 100 times, extreme average ratings disappear.

Which highly rated movies do you trust are actually good?

We would ultimately like to use the *quality* of ratings to determine the *quality* of a movie. If a movie has been viewed a single time, with a rating of 5, this movie now has an average rating of 5. This is an extremely high "quality" average rating. However the lack of *quantity* in ratings makes this average deceiving: it really only indicates that exactly one person enjoyed this movie. On the other hand, if a movie has over 100 views and an average rating of 4.5, in order to achieve a high average *many* people rated this movie highly. Therefore it is much more likely that this is a good movie. Thus a high quality rating, in the presence of a high *quantity* of ratings, is more indicative of a high quality *movie*. That is, among highly rated movies, we trust that those rated more than 100 times are actually good movies. Herein lies the motivation for restricting our data to at least 100 ratings.

Conjectures about the distribution of ratings.

Conjecture 1: Movies with few total ratings accumulate more low-end ratings, whereas highly-viewed movies accumulate more high-end ratings.

Movies which are "bad" likely receive a fewer total number of views simply because people do not want to watch them. On the other hand, "good" movies receive many total ratings thanks to eager viewers. This conjecture is testing the opposite: movies with fewer total views also receive (comparatively) more negative ratings, whereas highly-viewed movies receive more positive ratings. This question can be analyzed in terms of the distributions of total ratings (1-5) for each class of movie.

Data: We isolated from the data the movies with a low number of total ratings−less than half the average number of ratings−and a very high number of total ratings−more than twice the average number. We then tabulated the total number of ratings of each type (1-5) for each class of movie. This is summarized by the following two histograms.

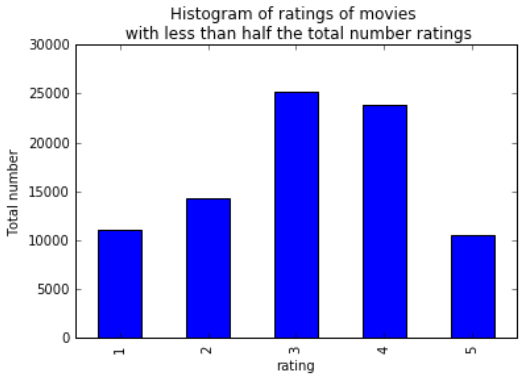


Figure 2.5: Total ratings, Poorly-viewed movies

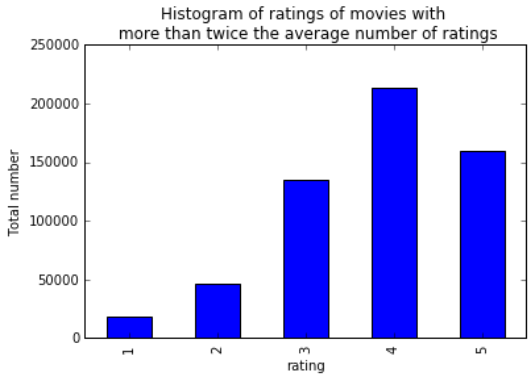


Figure 2.6: Total ratings, Highly-viewed movies

We phrase this question in terms of distributions, as discussions of quantity are biased- there is a difference in scale built-in to the way we separated our data (we explicitly selected high and low total views). However by considering the distribution of ratings shown above, we see a clear difference. First we consider the extreme ratings. Among highly-viewed movies, there was about a 10:1 ratio of ratings of 5 to those of 1. Among poorly-viewed movies, however, this ratio is essentially 1:1. Even beyond extreme ratings, it is clear that the distributions themselves are different. The distribution of highly-viewed movies accumulates towards the high-end of the rating scale. Among poorly viewed movies, the distribution accumulates towards the middle of the scale, with a much large representation of low-end ratings.

Conclusion: This conjecture is supported by our data, as seen in the distributions displayed above.

Conjecture 2: The distribution of all ratings of older movies is less normally-distributed than that of newer movies.

Newer movies are likely being watched by audiences looking to be entertained. They watch these films with a critical eye, analyzing their level of enjoyment. This should lead to a distribution that appears normal. Audiences watching much older movies (i.e. movies made before 1960) likely have some nostalgia tied up in these older films. This will lead to unbalanced distributions.

Data: Although the release year of each movie was not a predictor in the MovieLens data set, it was included within the title for each film. We extracted the year from each title, and classified the movies by year of release. The average release year among these films was 1960. Therefore we separated the movies into "older" and "newer" films, based on pre- or post- 1960 release. We then calculated the total number of ratings of each type (1-5) for these two sets. This is summarized in the following two histograms.

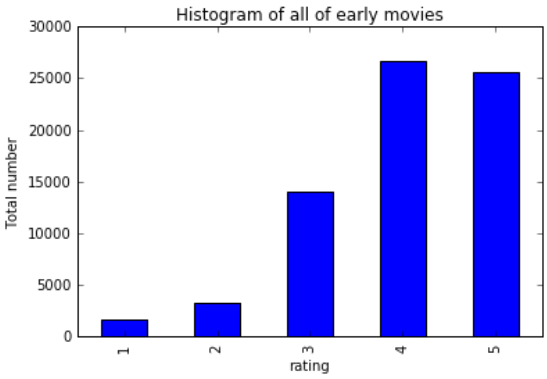


Figure 2.5: Total ratings, Older movies

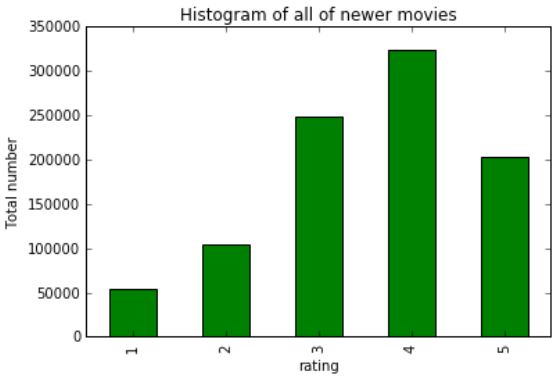


Figure 2.6: Total ratings, Newer movies

Once again, we frame this question in terms of *distributions* as there is an inherent gap in quantity of ratings: newer movies have been rated much more frequently. Nevertheless, we can see there is a distinct difference between these two distributions. Among older movies, the total number of 4 and 5 ratings is nearly equal. This total dominates the low (1 and 2) ratings, and is approximately twice the number of 3 ratings. This leads to an unbalanced distribution accumulating towards the high-end. The distribution of total ratings among newer movies, however, is much closer to a normal distribution, as seen in Figure 2.6. There is a non-trivial fraction of each rating type, with an expected value near 4.

Conclusion: This conjecture is supported by our data, as seen in the histograms above.

OBJECTIVE 3: Men Versus Women

Having explored many facets of the data concerning movies and their ratings, we now begin to analyze trends among those *submitting* the ratings, the users. We are interested in determining when the ratings among one group of users may be indicative of those from another group. In particular, we will compare the ratings of men versus the ratings of women.

Make a scatter plot of men versus women and their mean rating for every movie.

First, we compute (via pivot table) the average rating among men, and among women, for each movie title in the data set. This can be summarized in a scatter plot, shown in Figure 3.1.

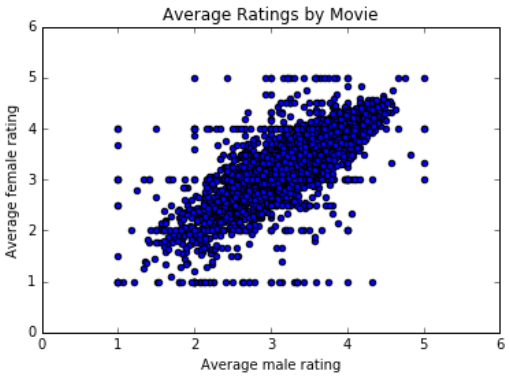


Figure 3.1: Average men's vs. women's ratings

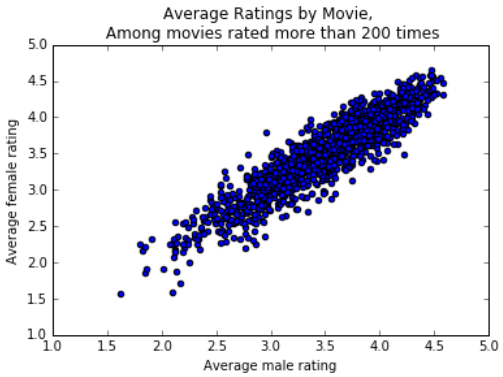


Figure 3.2: Average men's vs. women's ratings, Movies with at least 200 ratings.

Make a scatter plot of men versus women and their mean ratings for movies rated more than 200 times.

As we've seen previously, mean ratings of movies with few total ratings can be extremely misleading, and are not an accurate measure of the perceived quality of a movie. Therefore we restrict our scatter plot to only those movies viewed at least 200 times, shown above in Figure 3.2.

Compute the correlation coefficient between the ratings of men and women.

As seen in Figure 3.1 and 3.2 above, there seems to be an overall linear dependence between the average ratings of men, and the average ratings of women, particularly in those movies with more than 200 views. Given this perceived linear relationship, we therefore calculate the correlation coefficient between the mean ratings of men and the mean ratings of women.

Table 3.1: Correlation between average men's and average women's ratings.

	Correlation Coefficient
All Movies	0.76
Movies with at least 200 Ratings	0.92

What do you observe?

These correlation coefficients are both relatively high. A correlation coefficient closer to one indicates a stronger linear relationship between the data being tested. Therefore, there is a reasonable linear relationship between the average male rating and average female rating among all movies, and a rather strong linear relationship among movies with at least 200 ratings. This agrees with the visually-perceptible linear trends that seem to appear in Figures 3.1 and 3.2.

Are the ratings similar or not?

This data presented above is somewhat misleading. First, the correlation coefficient is a measure of the linear relationship between 2 sets of data. Thus a high (for example 0.92 as above) correlation indicates a linear relationship: that is, the average men's rating is predictable *in a linear fashion* from the average women's rating. This alone does not, however, mean that the average ratings need be *equal* (i.e. "similar"). A high correlation indicates linear predictability, but it is the nature of the trend in Figure 3.2 that indicates that this linear relationship is likely equality. Moreover, based on the high correlation values, one may predict that *ratings* between men and women are similar. However, this correlation is for the *mean* rating per title between men and women. That is, *on average* men and women rate movies in a manner that is linearly related. This doesn't indicate that the ratings *themselves* are similar. For example, there could be a movie for which both men and women have an average rating of 3, where women always rate as 1 or 5 and all men rate it as 3. We explore the data further to understand if the ratings between men and women are actually related.

As we have done in previous problems, we consider the percentage of extreme ratings made by each gender, now on a per-movie basis. That is, for each film we determine the percentage of 1 or 2 ratings−out of all ratings *for that film*−made by men, and made by women. We also calculate the percentage of 5 ratings for each film, again by gender. We then calculate the correlation between genders for these sets of by-movie statistics, summarized below.

Table 3.2: Correlation between average men's and average women's ratings.

		Correlation Coefficient
Percent 1 or 2 Ratings	**All Movies**	0.81
	At least 200 Ratings	0.90
Percent 5 Ratings	**All Movies**	0.73
	At least 200 Ratings	0.89

These high correlation coefficients once again indicate a linear dependency among the data, particularly among movies rated more than 200 times. That is for any movie, given the percentage of 5 ratings given by females, the percentage of 5 ratings given by men is related in a linear fashion. Similarly, the percentage of low-ratings is also linearly related across the genders. Once again, we consider trends in scatter plots, shown below in Figures 3.3 and 3.4, to postulate that this linear relationship is, in fact, equality (i.e. that these rating statistics are "similar").

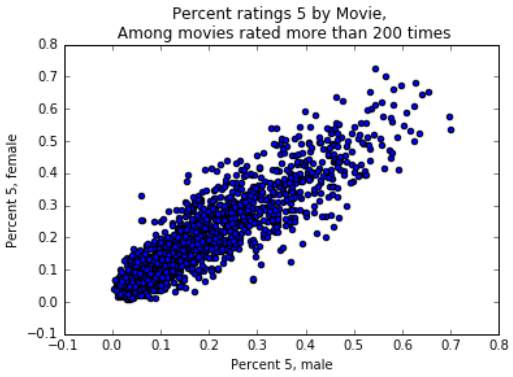


Figure 3.3: Percent 5 ratings men v. women, Movies with at least 200 ratings.

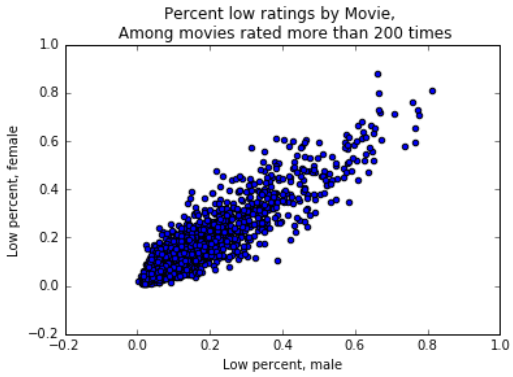


Figure 3.4: Percent low ratings men v. women, Movies with at least 200 ratings.

Based on this data, we conclude that males and females rate movies similarly, both on average and in distribution. This similarity is especially strong when considering movies rated more than 200 times. This does not, however, indicate we can predict a single male rating given all female ratings. The overall rating behavior between genders is similar, in the average and distributional sense, but single instances of ratings need not be similar.

Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.

Through the previous problem, we have developed a working definition for the rating of one gender to be predictable by the other. In particular, we have observed that it is likely unreasonable to ask for a prediction of a single-instance rating, but instead we analyze if the overall ratings between genders is similar. Therefore, we will say that a *circumstance* allows the rating of one gender to be predicted by the other gender if:

1. The average ratings per-movie is predictable between genders
2. The percentage 5-ratings (by movie) is predictable between genders
3. The percentage 1 or 2 ratings (by movie) is predictable between genders

As discussed previously, a high correlation coefficient indicates a linear relationship in the data−linear relationships are a standard model for making predictions. Therefore if the correlation between genders for the three statistics above are high it indicates that the two genders, both on average and in distribution, rate movies in a way which is linearly predictable.

Conjecture 1: Genders agree on what is "popular": Ratings between genders can be predicted among highly-watched movies made most recently.

Data: We restrict our total data set two-fold. First we isolate those movies with at least 200 total ratings, and then extract those movies with a release date after 1990.

Next we compute the three correlation coefficients described in (1) – (3) above. These are summarized in Table 3.3.

Table 3.3: Correlation between Men's and Women's ratings, Movies with at least 200 ratings, made after 1990

	Correlation
Average Ratings by Movie	0.924
Percent ratings 5 by Movie	0.902
Percent ratings 1 or 2 by Movie	0.911

These very high correlation values indicate a linear relationship between two genders' average ratings, percent 5 ratings, and percent low ratings *for each movie*. That is, for each highly-watched movie made after 1990, the average rating made by men relates linearly to the average women's rating. We may use this linear dependence to develop a linear (regression) model, and thereby estimate how *predictable* the average ratings are between genders. All reported error is *Mean Squared Error*.

We split our data (the double-restricted data as described above) into a training and testing subset. As the high correlation indicates a strong linear relationship, we train a linear regression model on 20% of our data, and reserve 80% for testing. The resulting (linear model) is:

[Average Women's Rating] = 0.2162 + 0.9420 [Average Men's Rating]

A plot of this model, along with the training data (red) and testing data (green) is shown in Figure 3.5. What is more, this linear model had a testing error of 0.04. Therefore we conclude that, for highly-watched movies made recently, the average women's rating is well-predictable from the average men's rating.

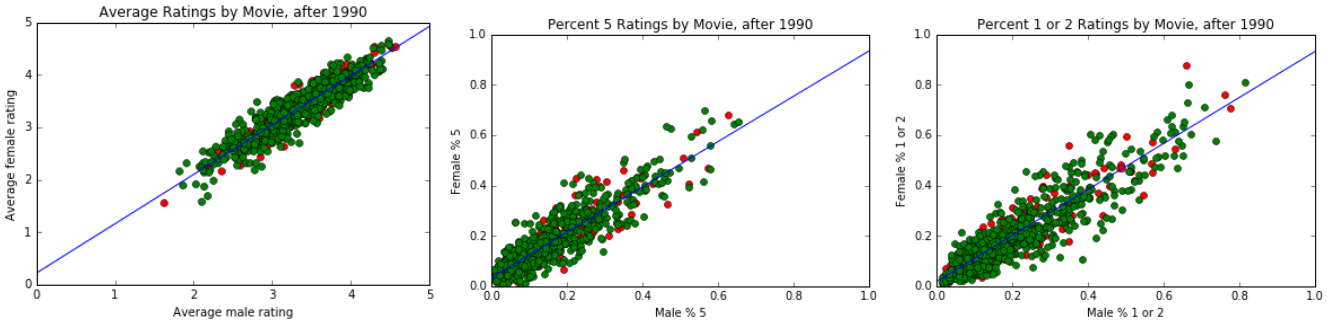


Figure 3.5: Linear model predicting: (1) average women's rating from average men's rating, (2) women's percent 5 ratings from men's percent 5 ratings, (3) women's percent 1 or 2 ratings from men's percent 1 or 2
Training Data: red, Testing Data: Green, Linear Model: Blue Line

Similarly, we trained a linear model to predict the percentage (per movie) of 5 ratings between genders. Once again using only 20% of data for training, we found the following linear regression model (where, again, these percents are the fraction of ratings of 5 out of all ratings for *each specific movie*).

[Women's % Ratings 5] =0.023 + 0.945 [Men's % Ratings 5]

This model had a mean squared testing error of 0.003. This model is shown in the second figure of Figure 3.6. Finally, we trained a linear model on the percentage (per movie) of 1 or 2 ratings between genders. Once again using only 20% of data for training, we found the following linear regression model:

[Women's % Ratings 1 or 2] =0.019 + 0.913 [Men's % Ratings 1 or 2]

This model had a mean squared testing error of 0.004, and can be found in the third figure of Figure 3.6.

Conclusion: Given testing errors which are small compared to the magnitude of the data, the average rating per movie, percent 5 rating per movie, and percent 1 or 2 rating per movie are all reliably predictable between genders. Therefore this data therefore supports the conjecture that ratings of one gender are predictable from those of the other given highly-watched movies produced recently.

Conjecture 2: Genders agree on what is "funny"! That is, the ratings between genders can be predicted among comedies.

Data: We restrict our total data set two-fold. First we isolate those movies with "Comedy" among its genres, and then extract those movies with at least 100 total ratings to avoid misleading extreme-value averages as shown in part 1.

We compute the three correlation coefficients described in (1) – (3) above. These are summarized in Table 3.4.

Table 3.4: Correlation between Men's and Women's ratings, Comedies with at least 100 total views

	Correlation
Average Ratings by Movie	0.914
Percent ratings 5 by Movie	0.890
Percent ratings 1 or 2 by Movie	0.863

These high correlation values once again indicate a linear relationship between two genders' average ratings, percent 5 ratings, and percent low ratings *for each movie*. Therefore, for each comedy (with at least 100 ratings), the average rating made by men depends linearly on the average women's rating. We may use this linear dependence to develop a linear (regression) model, and thereby estimate how *predictable* the average ratings are between genders.

As before, we split our data (the double-restricted data as described above) into a training and testing subset. We train a linear regression model on 20% of our data, and reserve 80% for testing. The resulting (linear model) is:

[Average Women's Rating] = 0.429 + 0.881 [Average Men's Rating]

A plot of this model, along with the training data (red) and testing data (green) is shown in Figure 3.6. Moreover, this linear model had a testing error of 0.054. Therefore we conclude that among comedies, the average women's rating is well-predictable from the average men's rating.

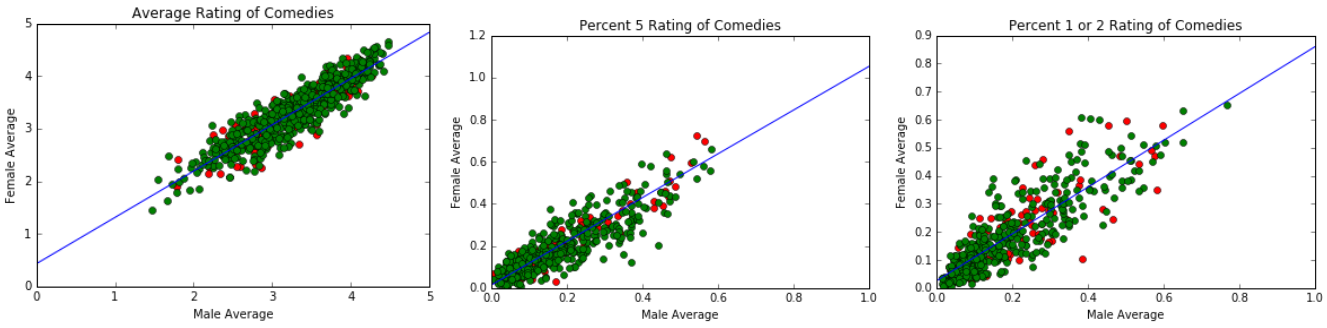


Figure 3.6: Linear model predicting: (1) average women's rating from average men's rating, (2) women's percent 5 ratings from men's percent 5 ratings, (3) women's percent 1 or 2 ratings from men's percent 1 or 2
Training Data: red, Testing Data: Green, Linear Model: Blue Line

Similarly, we trained a linear model to predict the percentage (per movie) of 5 ratings between genders. Once again using only 20% of data for training, we found the following linear regression model:

[Women's % Ratings 5] =0.023 + 1.048 [Men's % Ratings 5]

This model had a mean squared testing error of 0.0046. This model is shown in the second figure of Figure 3.6. Finally, we trained a linear model on the percentage (per movie) of 1 or 2 ratings between genders. Again using only 20% of data for training, we found the following linear regression model:

[Women's % Ratings 1 or 2] =0.026 + 0.834 [Men's % Ratings 1 or 2]

This model had a mean squared testing error of 0.005, and can be found in the third figure of Figure 3.6.

Conclusion: Given testing errors which are small compared to the magnitude of the data, the average rating per movie, percent 5 rating per movie, and percent 1 or 2 rating per movie are all reliably (and linearly) predictable between genders. This data therefore supports the conjecture that ratings of one gender are predictable from those of the other among comedies with at least 100 ratings. That is, genders tend to agree on what is funny!

OBJECTIVE 4: Business Intelligence

Online movie services such as Netflix® acquire new customers daily. Without any previous movie ratings to analyze, Netflix must recommend movies to new customers based solely upon their registration information. This initial recommendation is vital- Netflix wants its new customers to have a positive first experience. While there are many facets to this question, we ask the following:

What *genre* of movie should Netflix recommend to a first time user?

Before we begin, we acknowledge that it is unreasonable to expect *guaranteed* success in practical tests of new-user genre recommendations. An individual could certainly subscribe to Netflix for the sole purpose of watching documentaries about pre-World War II agriculture in Europe. Recognizing this *a priori* based on general trends in demographic information is implausible at best. Therefore we must formalize what we mean by "should" in the question above.

A genre Netflix *should* recommend to a first time user is one in which, with reasonably high probability, that user will find interest in watching a movie

There is subtlety here that deserves mention. This is a question of *interest*, not a question of *enjoyment*. We are *not* asking which genre of movie a new user is mostly likely to *enjoy*. At the end of the day, people watch movies that they dislike all the time−hence the many low ratings in our dataset. This displeasure, however, will likely be linked to the *movie itself* –disappointment in the actors, or directors, or even the script. The bulk of this unhappiness will not be held against the movie subscription service. What *will* create impatience towards Netflix, however, is if after logging in for the first time, a new user must spend a long time scrolling and searching through various pages to find a movie which seems interesting to them. People can't possibly choose a new movie based on how much they enjoyed it, without yet seeing it! Rather, they choose movies which they *think* they will enjoy−movies that look *interesting*. Therefore the underlying goal of this question is to understand *interest* among users based on their demographic information.

Metrics: User interest is at the heart of this question, so we will step away from relying upon movie ratings as a measure of *approval*. Instead the appropriate metrics to determine *interest* among a demographic group will be a combination of the *number* of ratings each type of movie receives, and the *percentage*−out of all ratings by a given group−a genre receives.

Data: There are 18 genres included in this data set: Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller,

War, and Western. Each title in the data set has attached to it some combination of these genres, for example "Animation/Children's/Comedy." Rather than considering the (many) such combinations, we consider the data within the 18 (overlapping) categories. So for example, any movie with "drama" listed *among* its genres will be considered a drama.

Preliminary Analysis

To begin, we determine which, among the numerous genres listed above, generated the most interest in users. By determining those "active genres" which are particularly well-watched, we can ask very specific demographic-based questions to analyze interest in these most popular genres. Within the scope of this report, we choose to restrict our focus to *three* such active genres. Figures 4.1 and 4.2 below illustrate the total number of movies within each genre, as well as the total number of ratings for films within each genre.

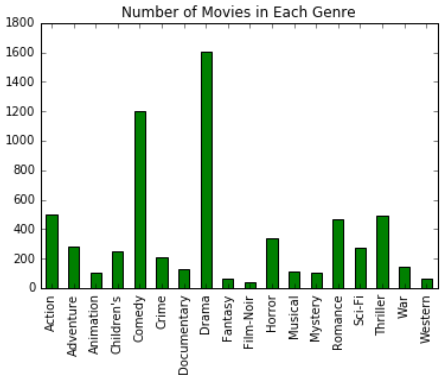


Figure 4.1: Number of movies classified in each genre

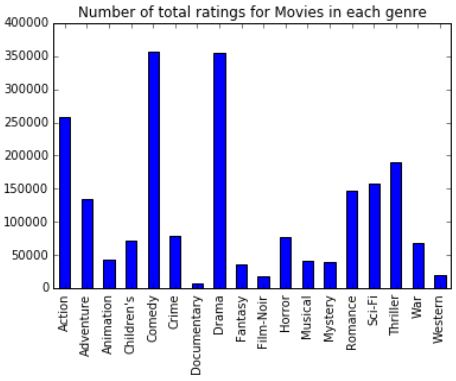


Figure 4.2: Number of ratings of movies in each genre

A few things are made clear by the graphs above. There are both many films *produced* classified as Comedy and Drama, and many films *watched* classified as Comedy and Drama. This is not overly surprising, as movie studios *produce* films with the goal of garnering audience interest. For these two reasons: high production rates, and high viewership rates, we choose Comedy and Drama as two of our active genres. In choosing our third active genre, the Adventure genre presents itself as a standout. It is ranked third in both production totals and ratings totals. What distinguishes Adventure films is the relationship between the two. There are far less than half as many Adventure movies as there are Dramas within the dataset, but these Adventures received much more than half as many ratings as those dramas. That is, the number of total ratings *per movie* of adventure films is larger than that for dramas. Thus for both high production and ratings per movie rates, we choose Adventure as the third "active genre." Thus we have re-focused our question:

Among Adventure, Comedy and Drama, which genre should Netflix recommend to a first time user?

Having narrowed the scope of our analysis, we now ask three very specific questions regarding these three genres.

Question 1: What is the best time of day to recommend a Drama to each gender?

That is, at what hour of the day is each gender most interested in watching a Drama? In order to judge interest, we will determine the percent of movies watched by each gender, per hour, that are classified as Drams. Figures 4.3 and 4.4 below illustrate the total number of dramas and the total number of movies rated per hour for each gender.

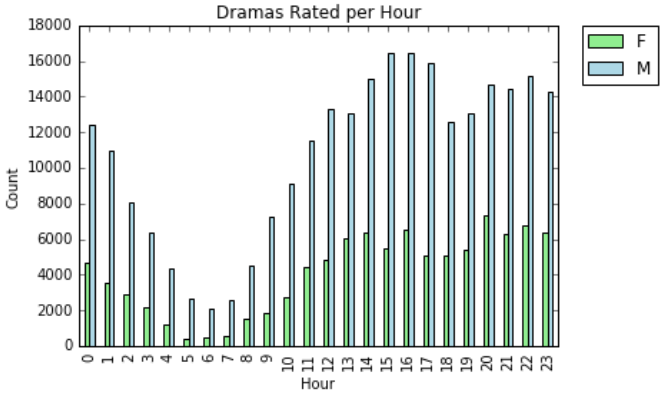


Figure 4.1: Number of Dramas rated per hour

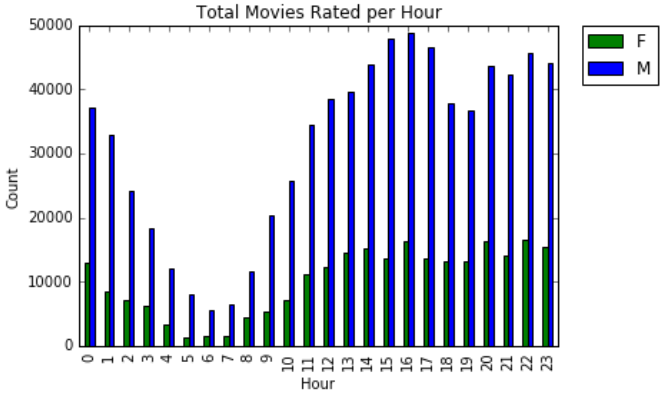


Figure 4.2: Number of Movies rated per hour

First, note that the disconnect of *volumes* in ratings between men and women is a function of our dataset, which has a disproportionate number of male participants. It will be the *percentage* data that follows which uncovers trends. In terms of *volume* of ratings, observe that the overall trends in this data are exactly what we expect to observe: the total number of movies (and dramas) being rated takes a severe dip during morning hours, when more people are sleeping or preparing for the day, and tends to peak in the later hours of the day. Notice there is also a dip in overall viewership between 6 and 7 PM – dinner time. Figure 4.5, however, illustrates the *percentage* of <u>all</u> movies watched per hour that are classified as drams. It is clear that there are distinctive trends in Drama viewership between genders. In particular of the movies watched, women watch nearly 15% *more* dramas in the evening than they do in the morning. On the other hand, men watch about 10% *fewer* drams in the evening than they do in the morning. These trends are precisely opposite. Women are most interested in watching dramas later in the day (particularly late in the evening and early in the morning) whereas men are most interested in watching dramas early in the day. It deserves mention that there is a consistent basis of interest in Dramas: at least 25% of movies watched every hour by both genders are dramas.

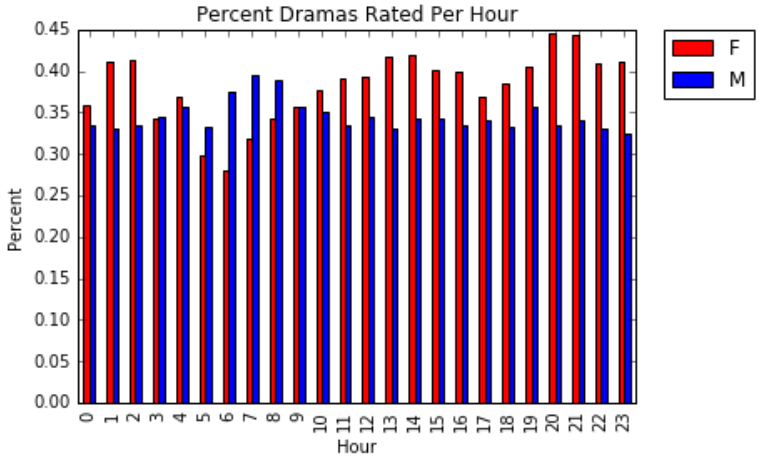


Figure 4.5: Percent of Movies that are Dramas

Conclusion: Women are significantly more interested in watching dramas later in the day, especially into the evening, than they are early in the morning. Men have a consistent interest in Dramas, and are more interested in viewing dramas in the morning than women.

Further Analysis: In order to make these conclusions more precise, more data is needed. In particular, movies are rated *after* they are viewed. Thus a movie which was rated at 10PM was likely *watched* starting at 8PM. Knowing the run-time for each movie would allow us to understand when people *start* to watch movies, allowing our conclusions to be more accurate.

Question 2: Which occupation is most likely interested in a comedy?

As before, we consider two metrics to analyze interest in the comedy genre: total number of comedies watched, and percent of *all* movies watched that are comedies. We consider these metrics for each of the 20 genders available in this data set: educator, artist, administration, college student,

customer service, doctor/health care, executive, farmer, homemaker, K-12 student, lawyer, programmer, retired, sales, scientist, self-employed, engineer, tradesman, unemployed, and writer. Figures 4.6 and 4.7 below display these results in chart form.

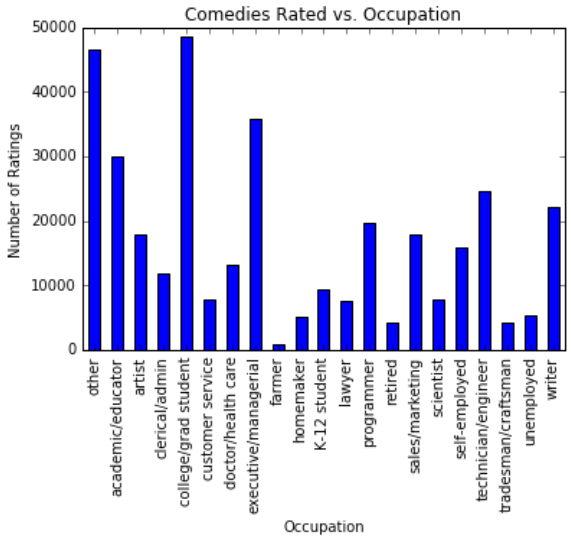


Figure 4.6: Number of Comedies rated

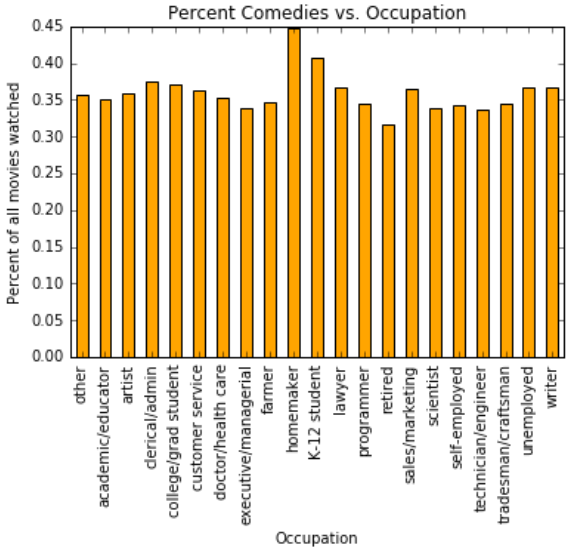


Figure 4.7: Percent of movies rated that are Comedies

We see that two occupations stand-out as showing high interest in comedies when considering each of these metrics. Though an occupation of "other" does not assist in our analysis, in considering sheer number of comedies rated one thing is clear: college students watch a lot of comedies. This is certainly not a surprising result to anyone who has lived in a dormitory. On the other hand, people of most occupations show a general interest in comedies, watching between 30 and 35% comedies among all other movies. Homemakers and K-12 students, however, watch a statistically significant higher percent of comedies overall, with 45 and 40% respectively. Once again, this result agrees with intuition.

Conclusion: Compared to other occupational groups, college students, children (K-12 students) and homemakers show the most interest in the Comedy genre.
Further Analysis: More data would assist in making these conclusions more accurate. For example, a greater stratification within the "other" occupational group would be informative. It is clear in Figure 4.6 that farmers, for example, watch comparatively few comedies. There may be whole occupational groups masked within the "other" category which is more statistically significant than farmers for studying interest in comedies.

Question 3: Which age group (per gender) watches the most adventure movies?

One last time, we consider two metrics to analyze interest in the adventure genre: total number of adventures watched, and percent of *all* movies watched that are adventures. We consider these metrics for each gender, and each of the 7 available age groups: Under 18, 18-24, 25-34, 35-44, 45-49, 50-55, and over 56. Figures 4.8 and 4.9 below display these results in chart form.

These Figures display quite interesting trends. In terms of *total* number of Adventure movies watched, young men dominate the data. The top-three Adventure watchers in terms of *volume* are 25-34, 35-44, and 18-24 year old men. Though this seems to match with ones intuition that young men are interested in adventure movies, we cannot draw this conclusion based solely on *volume* data- recall that our dataset has a disproportionate number of male participants. The *percentage* statistics, found in Figure 4.9, reveal a trend that is actually quite surprising.

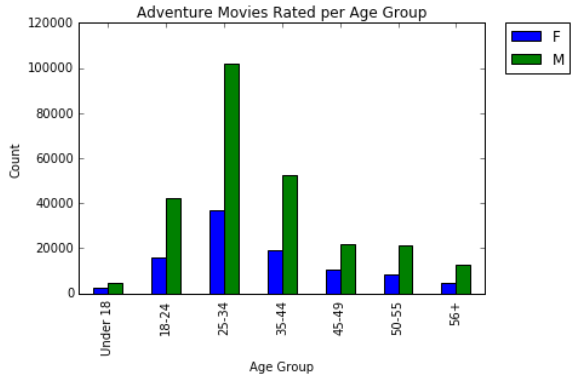


Figure 4.8: Number of Adventures rated

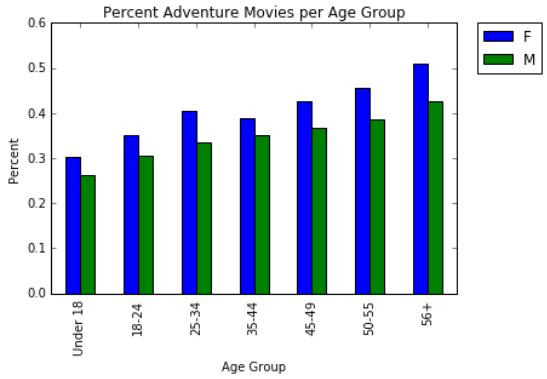


Figure 4.9: Percent of movies rated that are Adventures

Figure 4.9 displays the percentage of all movies watched, separated by gender and age group, that are classified as adventure movies. There are two clear very clear trends which emerge. Regardless of gender, the older a person gets they watch a higher percentage of adventure movies. On the other hand, regardless of age, *women* watch a larger percentage of adventure movies than *men*. Combine these trends together, and the data tells quite the opposite story than the misleading volume data. Amongst all movies watched, men under 18 watch approximately 25% Adventure movies, compared to **50%** amongst women over 56.

Conclusion: Women of every age are more interested in adventure movies than are men, and older people are more interested in adventure movies than younger individuals. In particular, nearly half of the movies watched by older women (in this data set) were adventure movies.
Further Analysis: More data, especially a greater representation of the women's population, would increase our confidence in our results. A finer segmentation of age-groups, for instance into 5-year increments, would also allow for more detailed analysis.

Overall Conclusions: Within this objective, we sought to understand which genre of movie would appeal, in terms of *interest*, to an individual based on their demographic information. Answering this question is vital to any online movie subscription service for making recommendations to first time users. By identifying "active genres" of movies, we were able to focus our analysis on genres which are highly-viewed. We further refined the scope of this broad problem by asking three specific questions. The trends we uncovered were somewhat surprising. Women show a significantly higher interest in Dramas in the evening than they do in the morning, whereas men are rather more interested in morning Drama-viewing compared to women. Recommending a comedy to a college student, a homemaker, or a child is a reasonable strategy, as all three show high interest in comedies compared to other occupations. Lastly women, and particularly aging women, show a particularly strong interest in Adventure movies.

We propose these conclusions with cautious confidence. Our sample users were a total of 6,000 people, likely not representative of *any* population as a whole. Never the less, these conjectures provide an excellent starting place for similar analysis on much larger (and proprietary) data sets.

Our Recommendation: At the end of the day, it is not necessary to determine a *singular* genre to recommend to each user based on their demographics. With the wide variety of people in the world, it is certainly unwise to "bet-the-farm" recommending a single genre, hoping that an individual matches overall demographic trends. A much more reasonable business practice is to determine a *few* genres (perhaps 4 or 5) that each individual is likely to show interest, and offer a small genre *selection*. This will not greatly detract from the user experience, yet certainly increase the success rate of first-time recommendations. The conclusions drawn within the specific questions posed here demonstrate good choices to add to such first time genre *selections*, based on a user's background information.

DATA LIMITATIONS

While the *MovieLens 1M Dataset* was sufficient to begin answering the questions presented in this report, we recognize that this data was certainly not without its limitations. To begin with, all movies rated in this data set were released *before* the year 2,000. Since that time, hundreds of movies generating billions in revenue have been produced. While a more up-to-date data set would be advantageous, however, it certainly was not necessary for this paper. The *MovieLens 1M* data set provided a wide breadth of films (ranging from 1920-2000) and, moreover, included a complete set of demographic information for each user. More importantly, the data set was readily available−the movie rating data for large streaming companies such as Netflix® is proprietary, making the *MovieLens 1M* set sufficient for our analysis. Although the quantity and breadth of the data collected was appropriate for this study, we recognize that the rating-responses carry significant bias. Most of this bias arises due to under-representation, as users resided within the United States, and only about 6,000 users submitted ratings. Additional sampling bias arises from unbalanced gender-sampling, with a 3:1 male to female ratio among the users.

CONCLUSIONS

In this paper, we began to explore the difficult-yet-relevant problem of making individual movie recommendations based solely on background information. First we retrieved the *MovieLens 1M Ratings* Data set, and performed some basic analysis. After examining "popular" movies based on submitted ratings, we found that our data suggested that older people are easier to please than younger folks: the oldest sample users gave the *highest* percentage of 5 ratings and *lowest* percentage of 1 ratings among all ratings they submitted. Next we extended our analysis to histograms to consider cumulative data. After generating summary-based histograms, we found that rating quantity and rating quality are related. That is, movies with few total ratings also had considerably more low-end ratings. On the other hand, *older* movies receive disproportionately more high-end ratings than do newer movies, which we postulate is due to a nostalgia factor. Next we considered the ratings of men versus those of women, and particularly when the two were predictable. Using a combination of correlation and linear regression, we determined that ratings between genders could be predicted among highly-viewed movies made recently, and comedies. That is, genders tend to agree on both what is "popular" and what is "funny"! Lastly, we extended our demographic-based recommendation problem to one of business intelligence. In particular, among "active genres," which genre of movie should Netflix recommend to a first-time user? By asking specific questions about these active genres (Drama, Comedy, and Adventure), we uncovered a number of trends between *interest* in a certain genre and various demographic groups. While none of these trends suggest a *singular* recommendation, each of the conclusions drawn proposes an excellent choice of genre to be included in a genre *selection* which is presented to a first-time user.