Group Members: Branden Diniz, Xuanyu Li, Dennis Silva, Chuxin Wei

**DS 501: Case Study 2 Report**

**Section 0 – Introduction and Background**

Business Intelligence is a technology driven process that creates efficient and rational decision-making processes through the gathering, exploring, interpreting and analyzing of data. One takes information from the business itself, its users or consumers, in additional to other internal systems and external sources to generate economic growth. This case study about analyzing data from movie ratings provided by MovieLens, reflects this type of process. If we imagine a movie company is interested in bettering some area of their company, we can attempt to use this data to make rational and informative decisions in that area. For example, the data has potential to answer the question of what movie should be made to entice additional users to subscribe to their movie service. Python, Pandas, and the statistical tools offered within them are forms of technology that help one to interpret the data analytically. Once the data is analyzed, one still needs to translate the results in a way which will express both why and how the decisions should be made to individuals from diverse mathematical, data science, and business backgrounds. Business Intelligence is a core principle that our group must achieve to have success in both this case study as well as our much of our future projects outside of the classroom.

For this case study our group decided to use the MovieLens 1M dataset. We chose this dataset over the other available datasets for two major reasons. Firstly, the 1M dataset is the largest dataset available from MovieLens that contains user demographic information. Individual user demographic information is purposefully not included in larger datasets. A reason for this is not given, but it may be due to the fact that companies have failed to maintain anonymization of its users because of the publically available data they have collected on them. One of the biggest and most recent instance of this was the de-anonymization of the Netflix Prize data [1]. We did not want to limit ourselves on the type of data available to us, and wanted to keep user demographics as an option to analyze if desired. Our second major reason was that although this dataset only has 1M ratings, it has a very similar sparsity or greater than that of the other datasets. If we attempted to convert this data into an m-by-n ratings data matrix with m = 3952 users and n = 6040 items, we would only cover about 4.18% of all possible user-movie rating pairs. In other words, the percentage of non-empty elements would be 4.18%. If we compare this to say the MovieLens 20M dataset with m = 138,493 users and n = 27,278 movies, we would have a sparsity of roughly 0.53%. Although datasets with more data points overall could lead to more accurate insights about certain aspects of the movie rating system, we chose to use a dataset with a maximal number of percentage of known data points to better cover the overall structure of the system.

**Section 1 –Basic Statistics and Popularity: User Ratings**

In order to get familiar with the data and investigate interesting paths for analysis, we focused first on answering many of the example questions specified in the Case Study 2 notebook. Beginning with Problem 1, we looked at some summary statistics on ratings over men and women. Of the one-million ratings, only 21 movies had an average of over 4.5 stars over all users. Men and women had 23 and 51 movies respectively whose average rating was over 4.5 stars. We found that 86 movies have a median rating over 4.5 among men over the age of 30 and 149 movies have a median rating over 4.5 among women over the age of 30.

We then wanted to find the top ten most popular movies. To do this, we had to first define what it means to be “popular.” After much discussion, we decided that a movie’s popularity was directly proportional to its number of ratings given to it regardless of the actual ratings score it received. We made the assumption that the volume of ratings were indicative of people talking about and watching these movies. The users who have rated these movies had to first hear about the movie, seek it out, and watch it during their free time. They also had to recognize the movie as being so popular that they wanted to provide input on whether other users would like/dislike it or to find similar/dissimilar movies to it for themselves. We decided the popularity of a movie is unrelated to the ratings scores it received as a movie that several users have viewed may not always be highly rated. We felt including the score may lead to more of what makes a movie “good” rather than “popular.” Under this assumption, it would make sense that movies that were available to be watched and rated earlier may have more ratings than movies that were added to the system later. This dataset does not provide when the movies were added to the database, thus we make the assumption that all movies were available around roughly the same time period. That being said, we found the top ten most popular movies by the number of ratings provided.

|  |  |  |
| --- | --- | --- |
| MovieID | Movie Title | Number of Ratings |
| 3858 | American Beauty (1999) | 3428 |
| 260 | Star Wars: Episode IV – A New Hope (1977) | 2991 |
| 1196 | Star Wars: Episode V – The Empire Strikes Back (1980) | 2990 |
| 1210 | Star Wars: Episode VI – Return of the Jedi (1983) | 2883 |
| 480 | Jurassic Park (1993) | 2672 |
| 2028 | Saving Private Ryan (1998) | 2653 |
| 589 | Terminator 2: Judgment Day (1991) | 2649 |
| 2571 | The Matrix (1999) | 2590 |
| 1270 | Back to the Future (1985) | 2583 |
| 593 | The Silence of the Lambs (1991) | 2578 |

**Top Ten Most Popular Movies**

Next we wanted to make some conjectures about how easy various groups are to please. Of the seven age groups provided by MovieLens (<18, 18-24, 25-34, 35-44, 45-49, 50-56, and >56 years old), we made the conjecture that the extrema of these groups (<18 and >56) were the most easiest to please. We believed the tastes of these groups to be very similar and therefore there rating habits too. To answer this question, we first looked at the variance of average movie ratings over each age group.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age Group | <18 | 18-24 | 25-34 | 35-44 | 45-49 | 50-56 | >56 |
| Rating Variance | 0.7920 | 0.6275 | 0.5139 | 0.4877 | 0.5123 | 0.6219 | 0.6725 |

**Rating Variance by Age Group**

One can notice that the <18 and the >56 groups actually have the greatest amount of variance in their average ratings of different movies with 0.7920 and 0.6725 respectively. The group with the least amount of variance is the 35-44 age group with a variance of 0.4877. Thus we find that our conjecture was wrong and the 35-44 age group was the most easy to please. Looking at this we were interested in the correlation between the average ratings of movies between different age groups. The table of correlation between each pair of movies as well as the overall correlation with all movie pairs is given below.

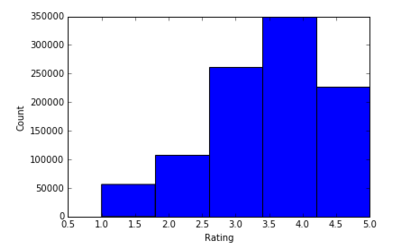
**Rating Correlation by Age Group**

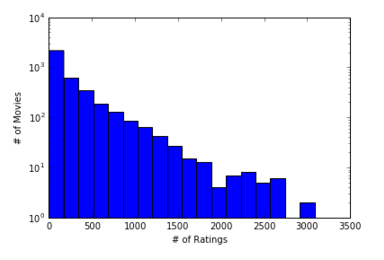
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age Group | <18 | 18-24 | 25-34 | 35-44 | 45-49 | 50-56 | >56 | Average Correlation |
| <18 | 1.000 | 0.5328 | 0.5654 | 0.5074 | 0.4936 | 0.4631 | 0.4089 | 0.5673 |
| 18-24 | 0.5328 | 1.000 | 0.6914 | 0.6027 | 0.5667 | 0.5364 | 0.4513 | 0.6259 |
| 25-34 | 0.5654 | 0.6914 | 1.000 | 0.7633 | 0.6886 | 0.6789 | 0.5845 | 0.7103 |
| 35-44 | 0.5074 | 0.6027 | 0.7633 | 1.000 | 0.6878 | 0.6903 | 0.6188 | 0.6957 |
| 45-49 | 0.4936 | 0.5667 | 0.6886 | 0.6878 | 1.000 | 0.6633 | 0.5926 | 0.6704 |
| 50-56 | 0.4631 | 0.5364 | 0.6789 | 0.6903 | 0.6633 | 1.000 | 0.6123 | 0.6635 |
| >56 | 0.4089 | 0.4513 | 0.5845 | 0.6188 | 0.5926 | 0.6123 | 1.000 | 0.6098 |
| Average Correlation | 0.5673 | 0.6259 | 0.7103 | 0.6957 | 0.6704 | 0.6635 | 0.6098 |  |

We find that the 35-44 year olds have high correlations with each of the age groups, most notably the 25-34 year olds. 35-44 year olds make up roughly 18% of all the known ratings, yet they have similar ratings habits to other groups. Thus, if we please this age group we will likely please most of the other age groups in the process. It should be noted that the 25-34 year olds actually have the overall highest average correlation with all the other age groups. However, this is only a slight change, and may be due to the fact that this age group accounts for roughly 40% of the entire 1M ratings.

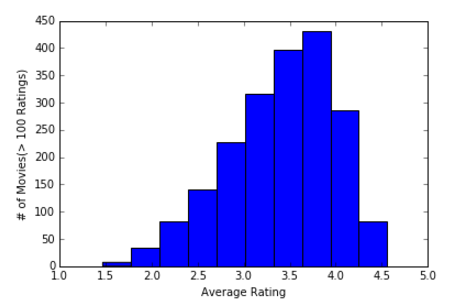
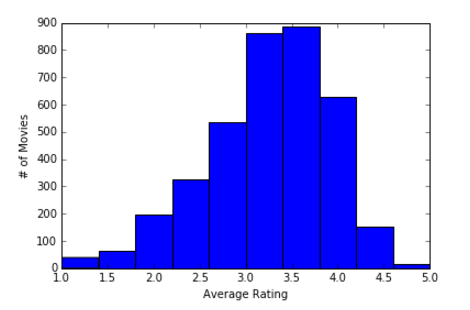
**Section 2 – Histograms and Distribution of Ratings**

Next we moved on to Problem 2 where we began to bin the data and create histograms. We plotted a histogram of the ratings of all movies as well as the number of ratings each movie received. Most people seem to rate a movie 4 stars, and overall rate movies on the higher end of the scale. In the number of ratings for each movie histogram, we decided to make the y-axis a log scale to see the differences in ratings easier. The plot looks relatively linear, therefore the data is exponential. This indicates there are only a few movies which have substantial user feedback on them.

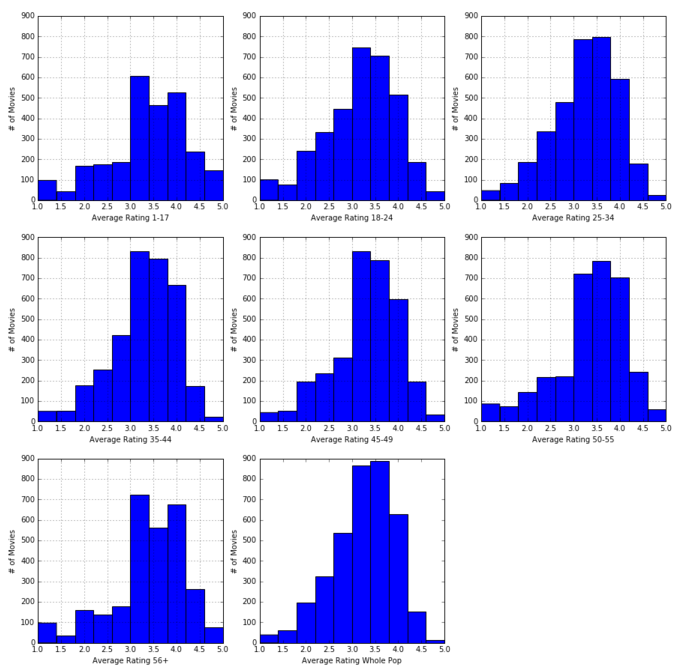
 **Ratings of All Movies Number of Ratings Each Movie Received**



Next we plotted the histograms of the average rating for each movie. In the first histogram we considered all movies independent of the number of ratings each movie received. In the second histogram we only included a movie if it had more than 100 ratings. In both cases, the majority of users tend to rate movies between 3 and 4 stars. One may notice that the tails disappear when we only use movies rated more than 100 times. For a movie to have an extremely high or extremely low rating, there must be a strong consensus of all users who rated that movie. If only a few users rate a certain movie, assuming all movies had relatively equal availability, then that movie may be described more as a niche than an overall popular movie. Individuals who are watching a specialized movie may be more likely to have a strong consensus in ratings of that movie. By removing movies that have less than 100 ratings, we remove most of these rating behaviors. To compare the two histograms directly we calculated the kurtosis for each. Kurtosis is a measure that helps describe the shape of a probability distribution. It reflects the ‘tailedness’ of a probability distribution of a real-valued random variable. It can range from -2 to +2. A positive kurtosis means the distribution has fat tails, while a negative kurtosis has thinner tails. The original histogram has a kurtosis of 0.3246, but after removing all movies with less than 100 ratings the kurtosis becomes -0.2478. Thus this provides a measure indicating the exact change in tails of the histogram, showing how they change from “fat” to “thin.”

 Average Movie Ratings (All Movies) Average Movie Ratings  
 (Movies with >100 Ratings)

After viewing the variance data collected from problem one and the histograms of average rating data from problem two, we made the conjecture that the 25-34 and the 35-44 age group’s average rating patterns followed the rating patterns of the entire population of users. To view this we first created histograms of the number of average ratings over each age group.

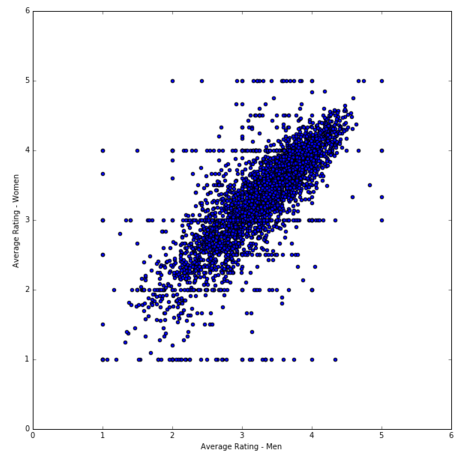
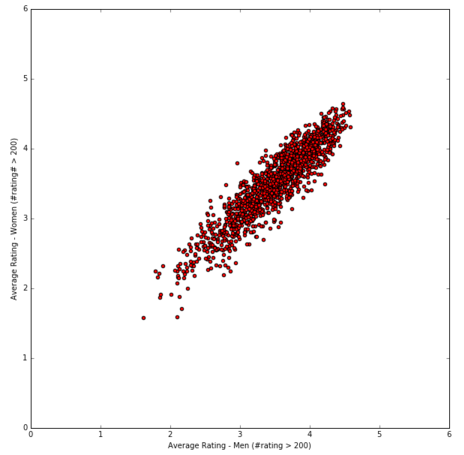
**Average Rating by Age Group**

Visually we could compare the shape of each histogram to one another, but this would overall depend on the comparer. We thus decided to conduct a one-way chi-squared test for goodness of fit. This test is applied when one has a categorical variable (each bin) from a single population (age group), and is used to determine whether the sample data is consistent with some hypothesized distribution (whole population). We first converted the data in each histogram into percentages of total movies rated by each individual population. Once done we were able to conduct the chi-squared test. We found that 25-34 year olds had a test statistic-value pair of (1.062, 0.9993) and the 35-44 year olds had (1.661, 0.9958). Since both pvalues are very high and the test statistics are relatively low, we accept the hypothesis that these two age group’s average rating percentages are consistent with the overall population’s rating percentages. In other words, if we know how either of these groups behave in terms of rating patterns we can extrapolate to the entire population fairly well. For comparisons sake, the test statistic-pvalue pair for the >56 year olds, one of the more variable groups, was (34.72, 6.677e-05). This indicates we cannot extrapolate to the entire population remotely well from this age group.

**Section 3 – Scatter Plots and Correlations: Men versus Women**

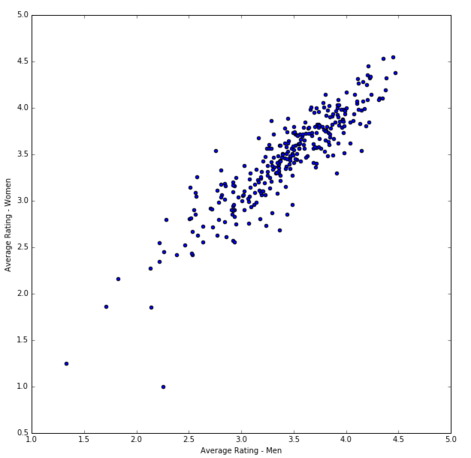
This brings us to the correlation between men and women. First we created a scatter plot of men versus women over their mean rating for every movie. We then created the same scatter plot, but this time only included movies that were rated more than 200 times.

**Men vs. Women Average Ratings Men vs. Women Average Ratings  
 (All movies) (Movies with >200 ratings)**



A few movies were only rated by a few people, majorly men or women. By only including movies rated more than 200 times, we find a better average of ratings for each movie for both the men and the women. When we calculate the correlation coefficient using all available data, we find it to be 0.7632. This indicates that there is a moderate correlation between the rating patterns of men and women on the same movies. However, when we remove the outliers from the data through filtering movies rated over 200 times, this correlation greatly increases to 0.9184. This better estimates the true correlation as each individual movie has a larger sample of ratings.

After seeing this, we made a conjecture that men and women must also rate similarly under other circumstances. We believed that men and women would also rate movie by genre similarly. A scatter plot was created similar to the ones above, but now each dot represents a different movie genre. We decided to differentiate movie genre by all genres a movie had. For example an Action | Comedy movie is different from just an Action movie. The correlation coefficient here is about 0.9015, meaning that men and women are also strongly correlated when rating different types of movie genre rather than solely individual movies.

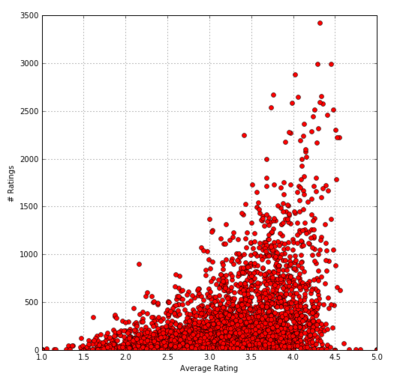


**Section 4 – Business Intelligence: Business Questions**

After much of the data analysis in this case study, we can review our findings in the light of business intelligence. For each section we proposed a conjecture and supported it with data. Our first conjecture, that the very young and old are the easiest to please, turned out to be false. However, in trying to support our conjecture we found those age groups who are the easiest to please in terms of overall ratings. The 25-34 year olds and the 35-44 year olds had the least variance among them, and also made up a significant portion of the data. Our second conjecture, that the 25-34 and 35-44 year old groups represented the rating behavior of the entire population, had data to support it too. Through histogram analysis and a chi-squared test, we have reason to believe that the understanding of how these two age groups behave can be extrapolated to the entire population of users who rate. Thus, a movie company may want to focus on creating content for these major demographics. They might also want to find ways to encourage the other age groups to provide more ratings in order to understand their specific tastes. Our third conjecture, that men and woman have similar rating tastes in movie genres, was also supported through looking at the correlation coefficient. A movie company may want to know this in order to spend less resources catering to any individual gender, and spend more resources in overall movie genre and content.

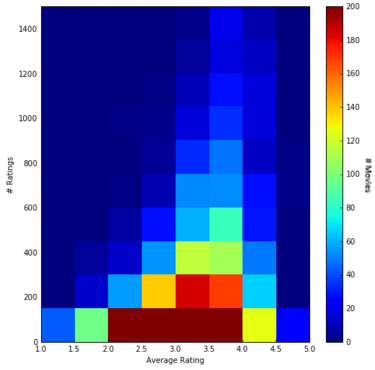
Two common business questions are to indicate what movies a company may want to remove from their system to make room for new content, as well as which movies to recommend to other users more frequently. To answer these questions, we decided to investigate the relationship between the number of ratings a movie has and the overall rating it receives. We first made a scatter plot of the data of the average rating of a movie versus the number of ratings it has.

**Average Rating vs. Number of Ratings**



From this we see that most of the data is concentrated less than 500 ratings per movie and between the ratings of 2 and 4, as would be expected. There seems to be a trend in the data that the more ratings a movie has, the higher likelihood it is rated moderately high. For example, if a movie has over 1500 ratings it seems to be likely that it will have an average rating greater than 3.5. Looking at the outliers, it seems like the most popular movies (those with the greatest number of ratings) also have an above average rating. Although this plot makes it easy to identify these popular movies, it makes it difficult to estimate the density of the clusters of movies. To visualize this density we decided to create a colored 2D image representing a two dimensional histogram of the data. In order to better indicate the changes in the density of movies we zoomed in the image to view only movies who had 1500 or less ratings, i.e. where most of the movies were located. We also reduced the maximum color density to 200 movies per bin to contrast the movies who had a great deal of ratings.

Average Rating vs. Number of Ratings   
(With added binned density)



With bins of 150 number of ratings high and 0.5 average star ratings across, we now see the movie density. The image visually strengthens the statements made about the scatter plot above. Most movies will only be rated less than 150 times, and those movies will receive average ratings. Movies that have a fair number of ratings (>600) yet have a relatively low score (<2.5) may want to be removed from the system to make room for additional content and save on costs to maintain those movies. Many individuals seem to not like these movies, and their overall contribution to system success seems to be minimal. On the other hand, movies that have a fair number of ratings (>600) and a relatively high score (>4.0) may want to be recommended to more people as the true rating average is more likely to be high when compared to movies with a high score but have fewer ratings. The roughly 140 movies with <150 ratings but are rated 4.0+ stars are on the cusp. With a few additional ratings, one could see if a movie falls into the average category (~3.0) or rises above and continues to receive high marks. Thus, through data analysis we were able to find supporting reasons to remove some specific movies while encouraging others.

**Citation**

[1] Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008.