Demographic Analysis of Movie Ratings

INFO 370: Team Pringles Supreme

Raymond Duong Timothy Ha Alex Lee Gabe Youn

1. INTRODUCTION

Films play a major role in not only shaping views of modern culture, but also in reflecting on societal attitudes of a time period. It is one of the essential forms of art in the entertainment industry and it is reflected by its continually increasing revenue. According to the PwC, the film industry is projected to bring in $35.3 billion from the United States alone in 2017 which will account for over a third of the global film industry market (“Electronic home”, n.d.).

The relevance of movie ratings parallel the growing presence of the film industry. Ratings by respected critics and well-known websites like Rotten Tomatoes and IMDb, are a few ways that films are measured. Movie ratings now have a deserved spot in famous magazines like Rolling Stone and The New York Times. These ratings often influence the attitudes of the public, therefore, can sway box office sales and film revenues.

But what can be learned from the movie reviews of average citizens? In this paper, we explore the relationships between ratings given by users and their demographic information, including age, gender, occupations, and location. We want to find out if there are significant differences or trends between various demographic groups and their preferences in movies. Additionally, we want to find the most accurate reviewers based on these demographic factors. This is defined as the user profile described by one or more demographic traits, that rates movies most representatively to the overall population. Lastly, we want to replicate a movie recommendation system based on this user rating data. Variations of algorithms similar to those used for movie recommendations are used in various popular websites, softwares, and applications for many types of media, including Netflix for TV shows and movies, and Spotify for music. Insight into this data can provide valuable information about how different demographics respond to movies and personal movie selections.

2. DATA

2.1 MovieLens

In our research we used data that was gathered by the GroupLens Research Project from the University of Minnesota[1]. These data sets contain 1,000,209 anonymous ratings of 3,952 unique movies made by 6,040 users on MovieLens. MovieLens is a commercial-free, community based website run by GroupLens where users can signup to collectively discuss and review movies. GroupLens collects data from the website and makes it publically available. The data sets we used contain demographic[1] information including: age range, gender, location by zip code, and occupation of the users. Included with this data set are the users' ratings on a 1-5 scale for specific movies along with the genre of each movie. Movie information such as titles and genres were pulled from the IMDb API. The latest movie release date was the year 2000. The following figures were created to visually display the distribution of information provided in the data sets:

Figure 1. Distribution of User Gender

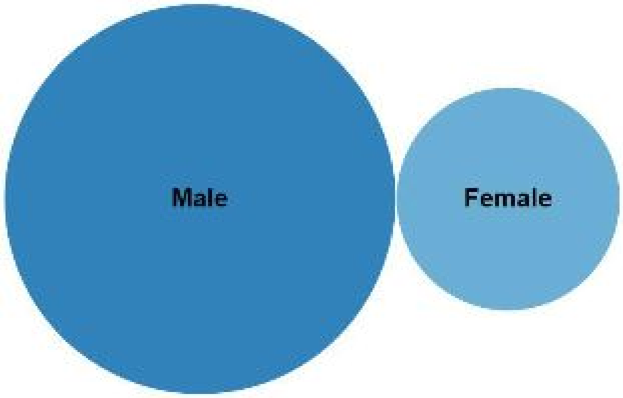
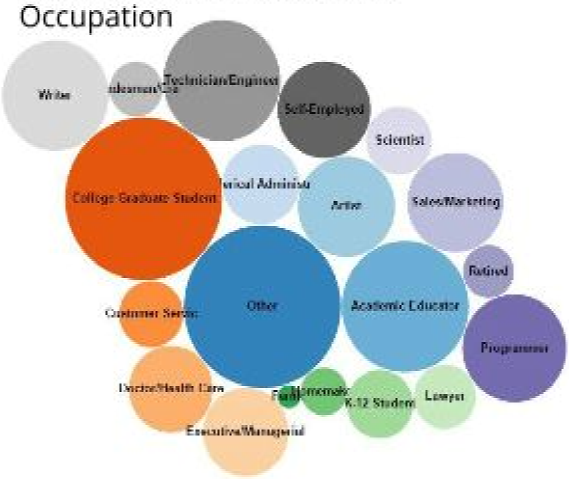


Figure 2. Distribution of User Occupation



[1] http://grouplens.org/datasets/movielens/

Figure 3. Distribution of User Age

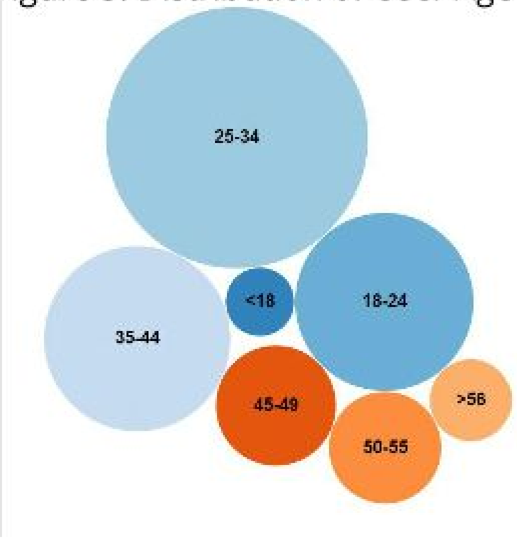


Figure 4. Distribution of User Location

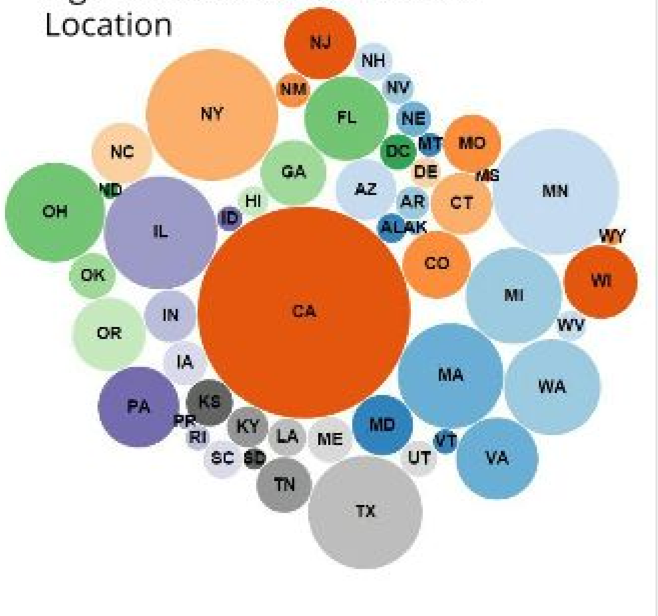
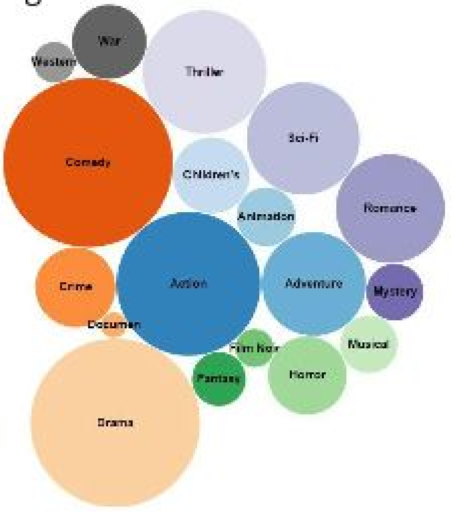


Figure 5: Distribution of Film Genre



As displayed by the Figure 1, males dominated the gender distribution representing 71.1% of all users (28.3% female). In addition, Figure 2 shows that most of the users were from a younger demographic which is also reflected in the occupation distribution in Figure 3, with college students making up a substantial fraction. Geographically, Figure 4 shows that there were users from a wide variety of states nation wide with California being the most prevalent. Figure 5 shows the variety of movie genres represented in the dataset, with comedy, drama, and action being the most frequent of the eighteen total genres.

2.2 Descriptive Statistics

Based on the user demographic data, males were far more represented than females. Not only was over 70% of the users male, but, on average, males gave more film ratings than females did and overall represented around 75% of the total ratings given. Females gave a higher mean rating than males did, and we will discuss if there was a statistically significant difference in our analysis.

In the appendix, Figure 10 shows the number of movie ratings by rating that each age group gave, and Figure 11 shows the mean rating given by males and females for each genre.

Figure 6: Gender Descriptive Statistics

Gender	Number of Users	Number of Ratings	Number of Ratings per User	Mean Rating
Male	4,331	753,769	174	3.57
Female	1,709	246,440	144	3.62
Total Population	6,040	1,000,209	165	3.58

Figure 7: Age Group Descriptive Statistics

Age Group	Number of Ratings	Mean Rating
Under 18	27,211	3.55
18 - 24	183,536	3.51
25 - 34	395,556	3.55
35 - 44	199,003	3.62
45 - 49	83,633	3.64
50 - 55	72,490	3.71
56+	38,780	3.77

3. ANALYSIS AND METHODS

3.1 Statistical Significance

One question our team proposed was whether there was statistically significant differences between ratings of different demographics. To analyze our data, we used a combination of t-test and ANOVA with a significance level of alpha = 0.05 to compare means within gender, age groups, and the different occupations.

To test for a significant difference between male (3.57) and female (3.62) mean ratings, we performed a t-test and found that the p-value was less than alpha, showing statistical difference between the two groups. Running ANOVA tests for both age groups and occupations, we found that there was statistical difference within both groups. This tells us that a statistical significance exists among the age groups but fails to tell us where within the age groups the significance is seen.

3.2 Determining Each State's Accuracy

We examined zip code information for each user in the database to see if geography had an influence on the mean ratings of each state. To do this, we compiled the ratings for all users for every movie they had rated. We calculated an Overall National Mean Rating (ONMR) of 3.5815644 out of 5. Then for each user, utilizing an online zip-code database[2], we determined the state each user live in. For example, if a user had a zip code of 98012, they would be clustered with other users from Washington state. With a simple filter method in Python, we pulled the average mean for all users in an individual state. Washington had an average mean of 0.04629 below the ONMR. By taking the absolute difference between each individual state mean rating and the ONMR, we were able to analyze how accurate each state was at rating movies in comparison to the national mean rating (Figure 8). A larger absolute difference demonstrated that the state was less accurate in rating movies when using the ONMR as a benchmark. A smaller absolute difference demonstrated that the state was more accurate in rating movies when using the ONMR as a benchmark. Some states that stood out were California, New York, and Oregon. These states are the most accurate states, with the lowest absolute mean difference and led us to believe that this visualization could demonstrate another analytical concept. Understanding that a greater number of users in our dataset originated from states such as California, New York, and Oregon, it is safe to

[2] federalgovernmentzipcodes.us

assume that not only does this visualization show the accuracy of each state's ratings, but also how well each state is represented in the national mean ratings.

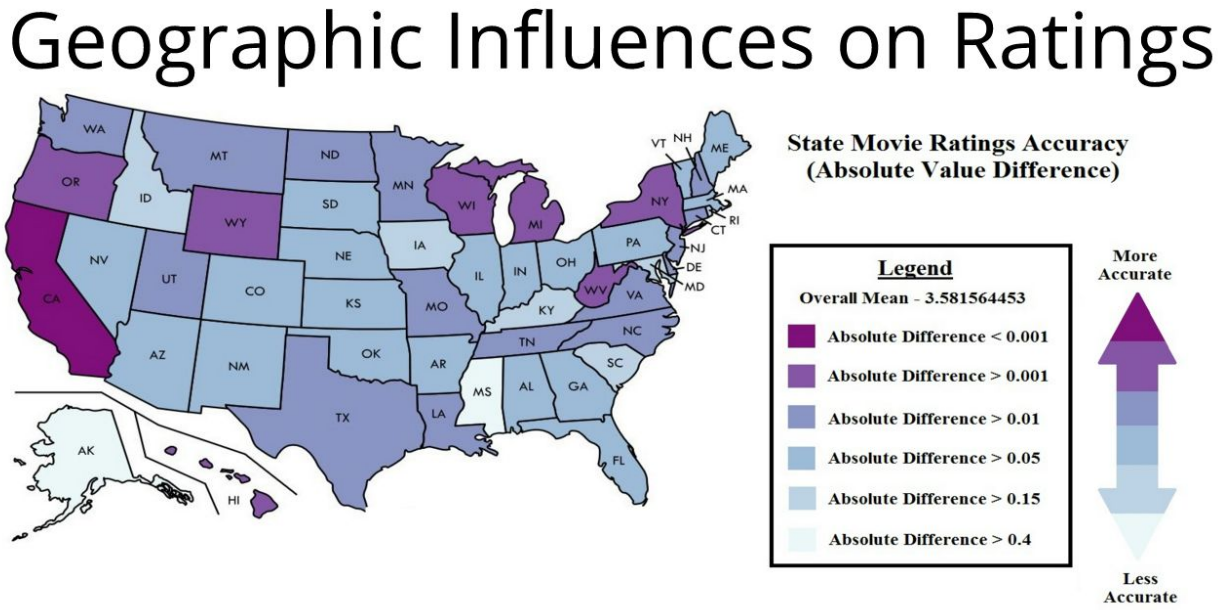

Geographic Influences on Ratings
State Movie Ratings Accuracy
(Absolute Value Difference)
Legend
Overall Mean - 3.581564453
Absolute Difference < 0.001
Absolute Difference > 0.001
Absolute Difference > 0.01
Absolute Difference > 0.05
Absolute Difference > 0.15
Absolute Difference > 0.4
More Accurate
Less Accurate
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
OH
KY
TN
MS
AL
GA
FL
SC
NC
VA
WV
PA
NY
VT
NH
ME
MA
RI
CT
NJ
DE
MD
AK
HI

Figure 8: Accuracy of State Ratings

3.3 Most Accurate Reviewer

We wanted to find out who the most accurate reviewers were but due to time limitations we decided to narrow our scope to the most accurate occupations. We define accuracy as being closest to the mean and we determined the accuracy by finding the minimum percent error.

As a case scenario, we set out to find the most accurate occupation for the film Toy Story (1995). To give some background, Toy Story had 2077 total ratings, 80% of which were given by users between the ages of 18 – 44. Comparing the overall mean rating of Toy Story to the mean rating given to Toy Story by each occupation would not give us the most accurate occupation. This is because the overall mean of rating for Toy Story can be heavily influenced by one group of users. For example, if 2000 of the 2077 ratings came from college students, the most "accurate" reviewers would no doubt be college students because they skew the overall mean towards their own due to the sheer number of ratings coming from college students. To account for this, we recalculated the overall mean of Toy Story as a weighted mean, calculated by summing the means of each occupation and then dividing that number by the number of occupations, 21. This allows us to determine the most accurate reviewer by giving each occupation equal weight in calculating the overall mean.

After adjusting our algorithms above, we found the weighted overall mean for Toy Story to be 4.15 out of 5. After comparing each occupation's mean rating for Toy Story to the weighted mean, we found the occupation that had the closest mean, 4.14, and thus, the smallest percent error. The occupation that was most accurate in rating Toy Story, much to our surprise, were lawyers.

In addition to finding the most accurate occupation for Toy Story, we calculated the most accurate occupations for all eighteen movie genres (Figure 9).

Figure 9: Most Accurate Occupation Per Genre

Movie Genre	Occupation with Most Accurate Ratings
Action	K-12 Student
Adventure	Customer Service
Animation	Doctor/Health Care
Children's	Programmer
Comedy	Academic/Educator
Crime	Technician/Engineer
Documentary	Farmer
Drama	Executive/Managerial
Fantasy	Academic/Educator
Film-Noir	Other
Horror	K-12 Student
Musical	Other
Mystery	Executive/Managerial
Romance	Executive/Managerial
Sci-Fi	Doctor/Health Care
Thriller	Self-Employed
War	Academic/Educator
Western	Executive/Managerial

3.4 Predictive Model of Movie Recommendations

GroupLens, the research organization that compiled the MovieLens dataset, utilized their compiled data to create a movie recommendation system as one of their foundational goals. Our group wished to emulate it and discover how it can be further optimized. We went with a k-clustering approach, the k's being the various demographics a particular user may fall under. In our dataset, we have three main demographics to cluster with: gender, age-range, and occupation. We would first cluster all of one gender and found what the top rated movies were for that gender (this was done by calculating the mean ratings for each movie within the specific gender). After, within this gender-divided subset, we would further cluster by age-range to find the movies highly rated by a user who falls under a specific gender and age-range combination. Lastly, we clustered once more for occupations and generated a list of top rated movies enjoyed by that demographic.

In our example, we found the top movies a female 3rd grader might enjoy from our movie list. When first clustered just by gender, the top rated movies for females were: "The World of Apu", "A Close Shave", "The Wrong Trousers", "The General", and "Sunset Blvd.". Adding in an age-range of under eighteen, the top movies were: "Schindler's List", "The Sound of Music", "Mulan", "The Wizard of Oz", and "The Matrix". Finally, adding in the occupation of a K-12 student, we arrived with a suggested list: "Marry Poppins", "The Matrix", "Erin Brockovich", "Willy Wonka and the Chocolate Factory", and "The Princess Bride". As the 'k's' increased and more specific demographics were added to the clustered, the suggestion list became more and more catered towards the requested individual.

4. LIMITATIONS

The most significant limitation of our data comes from the data sets themselves. It was mentioned by the original research group in their README file that, "All demographic information is voluntarily provided by the users and is not checked for accuracy," and, "Movies are mostly entered by hand, so errors and inconsistencies may exist." The data provided may help to give us insight on the film industry but may not be entirely accurate. Additionally, the data sets were all collected in 2000 and could be argued to be too outdated to give us any relevant information for society today.

In regards to our statistical analysis, we found statistical significance differences between various demographic groups and their moving ratings by using two-tailed t-tests and analysis of variance testing. However, this does not necessarily mean that the results we found are practically significant. Therefore, as these results may help to give us insights about various demographic groups in a population, they do not give us substantial evidence to make accurate assumptions about these groups.

5. CONCLUSION AND FUTURE DIRECTIONS

5.1. Concluding Results

The results show us that there are statistical differences in movie ratings between various demographic groups. However, our research and analysis should be treated as introductory for the significance of these results can be interpreted in a variety of ways. To gain more confidence and substance, additional work must be done on several more samples and demographics to show clear patterns throughout analyses. In creating a movie recommendation system, our work can serve as one of the building blocks.

5.2. Future Work

In furthering this work, firstly, we would focus on collecting or finding data sets that are more current. With more current data, we would once again analyze if various demographic traits significantly affected movie ratings of users. With more than one sample, we could look for parallels in significant differences between demographic groups in the current sample compared with previous ones in order to make inferences about these groups. In addition to ANOVA comparisons, which only tell us whether a significant difference exists between groups within a factor and not which specific groups have significant differences, we can conduct post-hoc tests such as Tukey's HSD. Additionally, while analyzing multiple demographic factors, it would be interesting to test for interaction effects between these factors.

In gathering more data and results, we look to improve our movie recommendation system. The first step in this direction would be to improve the accuracy in our results for top movie ratings in each demographic. Additionally, we could perform analytical work on more demographics than the ones we have analyzed. In expanding our work in another direction, we look to apply the recommendation system for not only movies but also other types of media like television shows and music.

6. REFERENCES

Electronic home. (n.d.). Retrieved from pwc.com/us/en/industry/entertainment-media/publications/outlook.html

7. APPENDIX

Figure 10

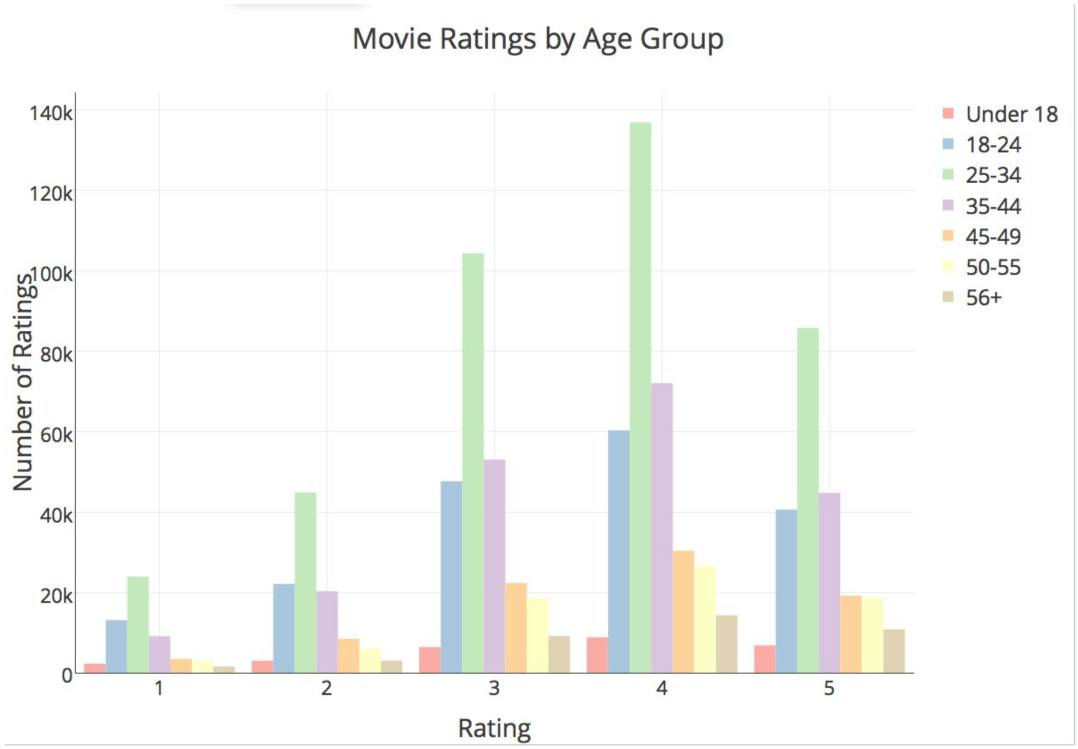
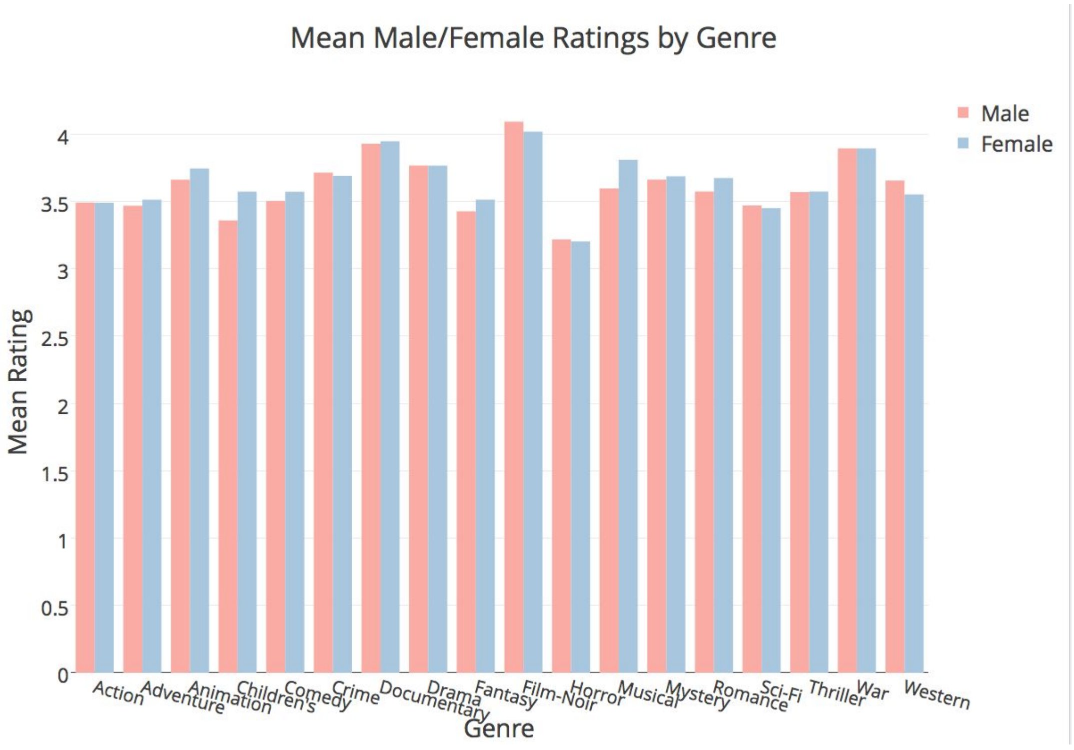


Figure 11


Mean Male/Female Ratings by Genre
Mean Rating
Action
Adventure
Animation
Children's
Comedy
Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
Musical
Mystery
Romance
Sci-Fi
Thriller
War
Western
Genre
Male
Female