CASE STUDY 2: ANALYZING DATA FROM MOVIELENS

TEAM 12

DS 501: INTRODUCTION TO DATA SCIENCE

TEAM MEMBERS

Chu Wang

Saranya Manoharan

Rishitha Kiran

Di You

Valerie Tuzel

1. MOTIVATION

GroupLens Research, which is a research lab in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities[1], has collected and made available rating datasets from the MovieLens web site, which was released on 2003. This dataset consists of three data files, users, movies, and ratings files. We wanted to explore this data and get some useful insights from it as much as we can by analyzing data. In what follows we dive into these analyses and explore the results we obtained.

2. DATA COLLECTION

As was mentioned above, the dataset we downloaded consists of three data files; users, movies, and, ratings files. These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. All ratings are contained in the file "ratings.dat" and are in the following format: UserID::MovieID::Rating::Timestamp. User information is in the file "users.dat" and is in the following format: UserID::Gender::Age::Occupation ::Zip-code. Movie information is in the file "movies.dat" and is in the following format: MovieID::Title::Genres.

2.1. Importing the MovieLens data set and merging it into a single Pandas DataFrame

Importing

We first downloaded the 1 million dataset zip file which contained the three data files.

Merging

Each individual file was read into Python as a Pandas Data Frame. The data was merged into a single Data Frame taking advantage of the relational nature of these files.

Storing Data

Finally, we stored this data frame as HDF5 file. An example of the first 5 rows of this data frame appears below.

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
0	1	1193	5	978300760	F	1	10	48067	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2	1193	5	978298413	M	56	16	70072	One Flew Over the Cuckoo's Nest (1975)	Drama
2	12	1193	4	978220179	M	25	12	32793	One Flew Over the Cuckoo's Nest (1975)	Drama
3	15	1193	4	978199279	M	25	7	22903	One Flew Over the Cuckoo's Nest (1975)	Drama
4	17	1193	5	978158471	M	50	1	95350	One Flew Over the Cuckoo's Nest (1975)	Drama

Table 1: First 5 rows of the data frame.

3. DATA ANALYSIS:

The following section will discuss in detail regarding the different analysis that were performed over the MovieLens dataset.

3.1. REPORT SOME BASIC DETAILS OF THE DATA YOU COLLECTED.

3.1.1. How many movies have an average rating over 4.5 overall?

We made use of pivot table for computing this. We calculated the average of all the ratings for each the aggregate function = np.mean and selected those whose average was greater than 4.5. The total number of movies that had an average rating over 4.5 were 21 and below is the table of these 21 movies along with their ratings.

title	mean rating
Apple, The (Sib) (1998)	4.666667
Baby, The (1973)	5.000000
Bittersweet Motel (2000)	5.000000
Close Shave, A (1995)	4.520548
Follow the Bitch (1998)	5.000000
Gate of Heavenly Peace, The (1995)	5.000000
Godfather, The (1972)	4.524966
I Am Cuba (Soy Cuba/Ya Kuba) (1964)	4.800000
Lamerica (1994)	4.750000
Lured (1947)	5.000000
One Little Indian (1973)	5.000000
Sanjuro (1962)	4.608696
Schindler's List (1993)	4.510417
Schlafes Bruder (Brother of Sleep) (1995)	5.000000
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	4.560510
Shawshank Redemption, The (1994)	4.554558
Smashing Time (1967)	5.000000
Song of Freedom (1936)	5.000000
Ulysses (Ulisse) (1954)	5.000000
Usual Suspects, The (1995)	4.517106
Wrong Trousers, The (1993)	4.507937

Table 2: Movies with average rating over 4.5.

3.1.2. How many movies have an average rating over 4.5 among men? How about women?

We used the pivot table for computing this as well. We found the average of all the ratings for each title based on gender and selected those titles which corresponded to the male gender that had average rating greater than 4.5. Among all the movies rated, there were only 23 movies that had average rating of over 4.5 among men. Table 3 shows the top ten of these movies.

gender title	F	M
Angela (1995)	3.000000	5.000000
Apple, The (Sib) (1998)	4.750000	4.600000
Baby, The (1973)	NaN	5.000000
Bells, The (1926)	4.000000	5.000000
Dangerous Game (1993)	4.000000	5.000000
Follow the Bitch (1998)	NaN	5.000000
For All Mankind (1989)	3.333333	4.583333
Gate of Heavenly Peace, The (1995)	5.000000	5.000000
Godfather, The (1972)	4.314700	4.583333
I Am Cuba (Soy Cuba/Ya Kuba) (1964)	5.000000	4.750000

Table 3: Top ten movies that have an average rating over 4.5 among men.

We then selected those titles which corresponded to the female gender that had average rating greater than 4.5. Among all the movies rated, there were 51 movies that had average rating over 4.5 among women. Table 4 below shows the top ten of these movies.

Top ten movies that have an average rating over 4.5 among women:

gender	F	M
title		
24 7: Twenty Four Seven (1997)	5.000000	3.750000
Among Giants (1998)	4.666667	3.333333
Aparajito (1956)	4.666667	3.857143
Apple, The (Sib) (1998)	4.750000	4.600000
Ayn Rand: A Sense of Life (1997)	5.000000	4.000000
Ballad of Narayama, The (Narayama Bushiko) (1958)	5.000000	3.428571
Battling Butler (1926)	5.000000	3.222222
Before the Rain (Pred dozhdot) (1994)	4.600000	4.173913
Belly (1998)	5.000000	3.000000
Big Combo, The (1955)	5.000000	3.600000

Table 4: Top ten movies that have an average rating over 4.5 among women.

Since we did not filter movies according to the number of times they were rated we thought that these results might not be useful as there could be movies that has very few ratings like 1 or 2 ratings. Because of this reason we wanted to make these calculations for the movies that had more than 250 ratings to start with. For this, we first found the number of the movies that had at least 250 ratings. 1,214 of the 6,040 movies have at least 250 ratings. Table 5 shows the top 20 most rated movies. Then, using the same pivot table results from before we selected the ones that had at least 250 ratings and sorted them in descending order to find the top 10 movies, that have at least 250 reviews, with the highest average ratings among men and women as shown in Tables 6 and 7, respectively. In Table 6 we see that among those movies rated at least 250 times only 5 of them had a rating over 4.5 among men. However, among women there were 11 movies that had a rating over 4.5.

title	
American Beauty (1999)	3428
Star Wars: Episode IV - A New Hope (1977)	2991
Star Wars: Episode V - The Empire Strikes Back (1980)	2990
Star Wars: Episode VI - Return of the Jedi (1983)	2883
Jurassic Park (1993)	2672
Saving Private Ryan (1998)	2653
Terminator 2: Judgment Day (1991)	2649
Matrix, The (1999)	2590
Back to the Future (1985)	2583
Silence of the Lambs, The (1991)	2578
Men in Black (1997)	2538
Raiders of the Lost Ark (1981)	2514
Fargo (1996)	2513
Sixth Sense, The (1999)	2459
Braveheart (1995)	2443
Shakespeare in Love (1998)	2369
Princess Bride, The (1987)	2318
Schindler's List (1993)	2304
L.A. Confidential (1997)	2288
Groundhog Day (1993)	2278

Table 5: Top 20 most rated movies along with the number of ratings they received.

gender	F	M
title		
Godfather, The (1972)	4.314700	4.583333
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	4.481132	4.576628
Shawshank Redemption, The (1994)	4.539075	4.560625
Raiders of the Lost Ark (1981)	4.332168	4.520597
Usual Suspects, The (1995)	4.513317	4.518248
Star Wars: Episode IV - A New Hope (1977)	4.302937	4.495307
Schindler's List (1993)	4.562602	4.491415
Wrong Trousers, The (1993)	4.588235	4.478261
Close Shave, A (1995)	4.644444	4.473795
Rear Window (1954)	4.484536	4.472991
Double Indemnity (1944)	4.282051	4.468354
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963)	4.376623	4.464789

Table 6: Top 12 movies, that have at least 250 reviews, with the highest average ratings among men.

gender	F	M
title		
Close Shave, A (1995)	4.644444	4.473795
Wrong Trousers, The (1993)	4.588235	4.478261
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	4.572650	4.464589
Wallace & Gromit: The Best of Aardman Animation (1996)	4.563107	4.385075
Schindler's List (1993)	4.562602	4.491415
Shawshank Redemption, The (1994)	4.539075	4.560625
Grand Day Out, A (1992)	4.537879	4.293255
To Kill a Mockingbird (1962)	4.536667	4.372611
Creature Comforts (1990)	4.513889	4.272277
Usual Suspects, The (1995)	4.513317	4.518248
It Happened One Night (1934)	4.500000	4.163934
Rear Window (1954)	4.484536	4.472991

Table 7: Top 12 movies, that have at least 250 reviews, with the highest average ratings among women.

3.1.3. How many movies have a median rating over 4.5 among men over age 30? How about women over age 30?

Going back to our original dataset, we first extracted the data where "age" was greater than 30. Then we used a pivot table to find the median of all the ratings for each title based on gender on this data. We then selected those titles which corresponded to the male gender with median rating over 4.5. According to our results, 86 movies have median rating over 4.5 among men over the age of 30. We then selected those titles which corresponded to the female gender with median rating over 4.5. Our results show that 149 movies have median rating over 4.5 among women over the age of 30. Tables 8 and 9 show the top ten movies that have median rating over 4.5 among men and women over the age of 30, respectively.

gender	F	M
title		
42 Up (1998)	4.0	5.0
Rear Window (1954)	5.0	5.0
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	5.0	5.0
Seven Chances (1925)	4.0	5.0
See the Sea (Regarde la mer) (1997)	NaN	5.0
Schlafes Bruder (Brother of Sleep) (1995)	NaN	5.0
Schindler's List (1993)	5.0	5.0
Saving Private Ryan (1998)	4.0	5.0
Sanjuro (1962)	5.0	5.0
Return with Honor (1998)	4.5	5.0

Table 8: Top ten movies that have median rating over 4.5 among men over the age of 30.

gender	F	M
title		
24 7: Twenty Four Seven (1997)	5.0	3.0
Philadelphia Story, The (1940)	5.0	4.0
Producers, The (1968)	5.0	4.0
Promise, The (La Promesse) (1996)	5.0	4.0
Psycho (1960)	5.0	5.0
Quiet Room, The (1996)	5.0	2.5
Raiders of the Lost Ark (1981)	5.0	5.0
Rain (1932)	5.0	4.0
Ran (1985)	5.0	4.0
Raw Deal (1948)	5.0	3.0

Table 9: Top ten movies that have median rating over 4.5 among women over the age of 30.

3.1.4 What are the ten most popular movies?

Our definition for most popular movies is as follows: First the movies should be rated at least 2,500 times and should have the average rating of 4 or more.

In order to get the ten most popular movies we first found the average rating per movie and then selected the ones that have at least 2,500 ratings. We could have gone even higher than this number to make sure we chose the popular movies among the most rated movies however once we went higher than 2,500, e.g. 2,700, this dataset came short in terms of the movies—there were only 4 movies that had at least 2,700 ratings. Hence, we chose this number to be 2,500. Once we selected the movies that had at least 2,500 ratings we then sorted them by their average ratings in descending order to get the top ten most popular movies as shown in Table 10. Once we sorted them we observed that all those movies for this dataset already had an average rating more than 4 hence we did not filter them for the ratings as well, however if

this was not the case, we should have filtered them by choosing the ones that had average rating greater than 4.

	rating	
	size	mean
title		
Raiders of the Lost Ark (1981)	2514	4.477725
Star Wars: Episode IV - A New Hope (1977)	2991	4.453694
Silence of the Lambs, The (1991)	2578	4.351823
Saving Private Ryan (1998)	2653	4.337354
American Beauty (1999)	3428	4.317386
Matrix, The (1999)	2590	4.315830
Star Wars: Episode V - The Empire Strikes Back (1980)	2990	4.292977
Fargo (1996)	2513	4.254676
Terminator 2: Judgment Day (1991)	2649	4.058513
Star Wars: Episode VI - Return of the Jedi (1983)	2883	4.022893

Table 10: Top ten most popular movies.

We did an extensive analysis based on age groups later in to see which age groups are easier to please. So please refer to later sections for our conjecture on that.

3.2. EXPAND OUR INVESTIGATION TO HISTOGRAMS

3.2.1. Plot a histogram of the ratings of all movies.

Here we plotted a histogram of the "rating" column of the merged data using matplotlib and seaborn packages, regardless of how many times the movies were rated. From Figure 1, it seems that the users tend to give ratings of 3 or higher more frequently than the ratings of 1 and 2.

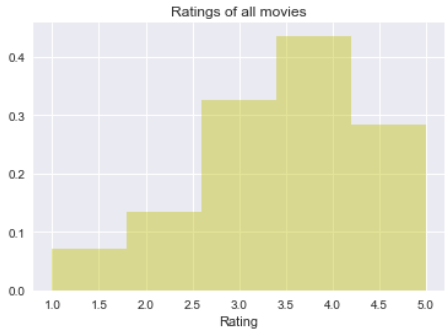

Fig. 1: Histogram plot of the ratings of all movies.

3.2.2. Plot a histogram of the number of ratings each movie received.

We then plotted a histogram of the number of ratings each movie received. Figure 2 shows both the histogram (on the left) and the semi-log distribution (on the right). From the histogram, we see that most

movies seem to get very few ratings. By taking the log of the y axis (plot on the right in Figure 2) we also see that this distribution exhibit nearly exponential decay with some deviations in the tail.

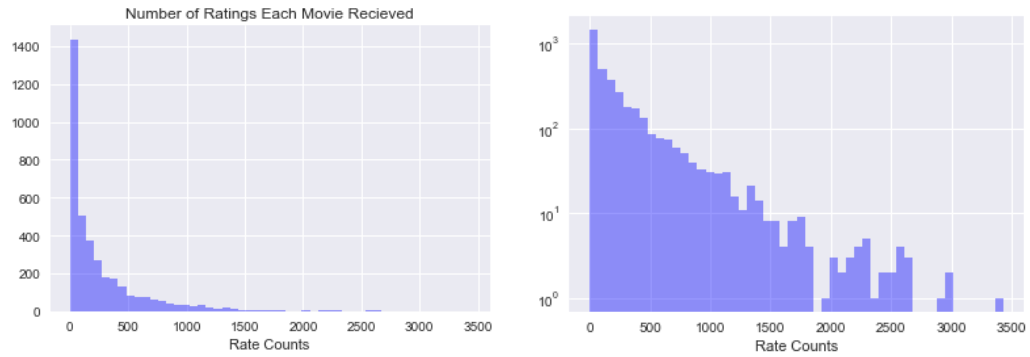

Fig. 2: Histogram of the number of ratings each movie received on the left,

semi-log distribution on the right.

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

We then plotted a histogram of the average ratings for movies for all the movies in the dataset and then compared it to the histogram of the number of ratings of average rating for movies that were rated more than 100 times.

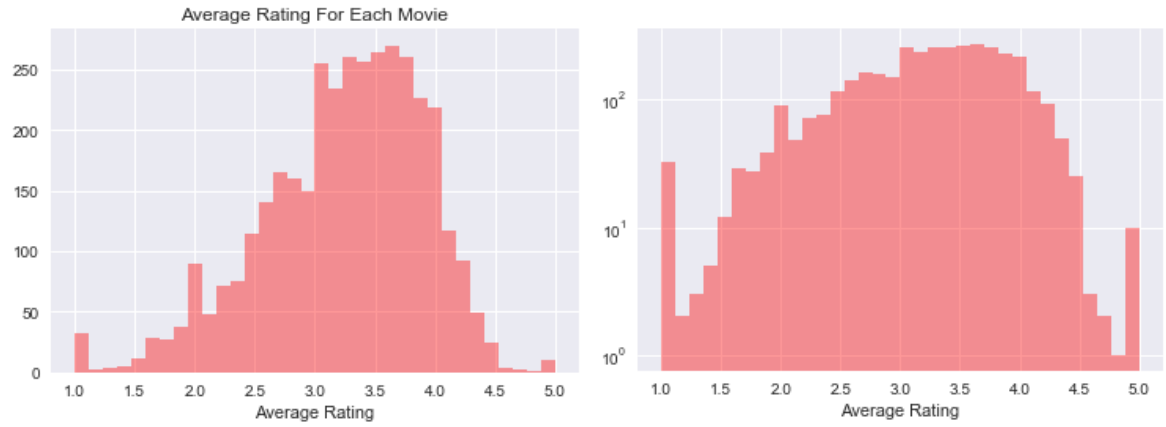

Fig. 3: Histogram of the average rating for movies for all the movies (on the left),

semi-log distribution (on the right).

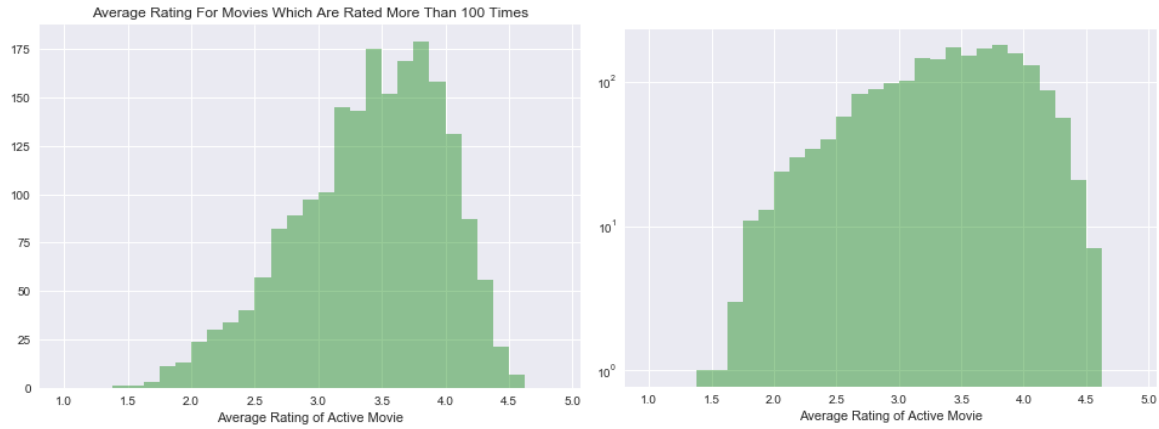

Fig. 4: Histogram of the average rating for movies for the movies which are rated more than 100 times (on the left), semi-log distribution (on the right).

In both plots shown in Figures 3 and 4 we see a similar general trend. In Fig. 3, we see that there are peaks at the ratings 1 and 5. By taking the log of the y axis we looked at the scaling of the distribution tail. The semi-log distribution shows that this distribution is negatively skewed or left skewed and, hence the mean is less than the median. For values less than the mean, the decay resembles a Gaussian distribution however for values larger than the mean, the decay is stronger than a Gaussian distribution. However, the presence of the peaks at the ratings 1 and 5 indicate a breakdown of this scaling. In comparison, in Fig. 4 where we show the histogram for the movies rated more than 100 times, these peaks disappear and tails follow the distribution nicely.

3.2.4. Make some conjectures about the distribution of ratings.

In this data set age groups are chosen as follows: 1: "Under 18", 18: "18-24", 25: "25-34", 35: "35-44", 45: "45-49", 50: "50-55", 56: "56+" Following this, we first looked at the distribution of the users' ages as shown in Figure 5. Most of the users in this dataset are between the ages of 25 to 35 followed by the age groups of 18 to 25 and 35 to 45.

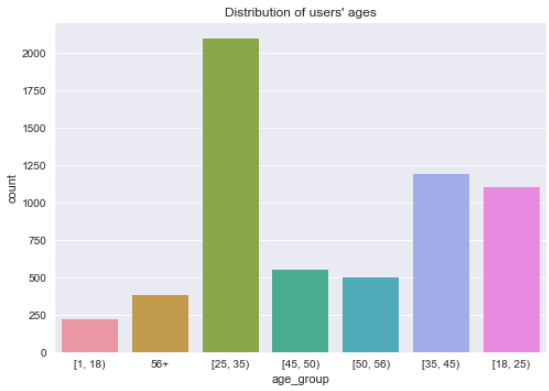

Fig. 5: Distribution of the users' ages.

We also looked at the rating size of these age groups and the average rating they gave. As seen in Table 11. Age group between 25 to 35 gave the most ratings followed by the age groups 18 to 25 and, 35 to 45, in

sync with the age groups who had the most number of users as we discussed above. Average rating was mostly around 3.5 to 3.6 for all the age groups though people who were 50 or older seemed to be the most enthusiastic raters of all the groups with an average around 3.7.

	rating	
	size	mean
age_group		
56+	38780	3.766632
[1, 18)	27211	3.549520
[18, 25)	183536	3.507573
[25, 35)	395556	3.545235
[35, 45)	199003	3.618162
[45, 50)	83633	3.638062
[50, 56)	72490	3.714512

Table 11: Rating size and mean per each age group.

We then explored the rating distribution among these age groups as shown in Figure 6.

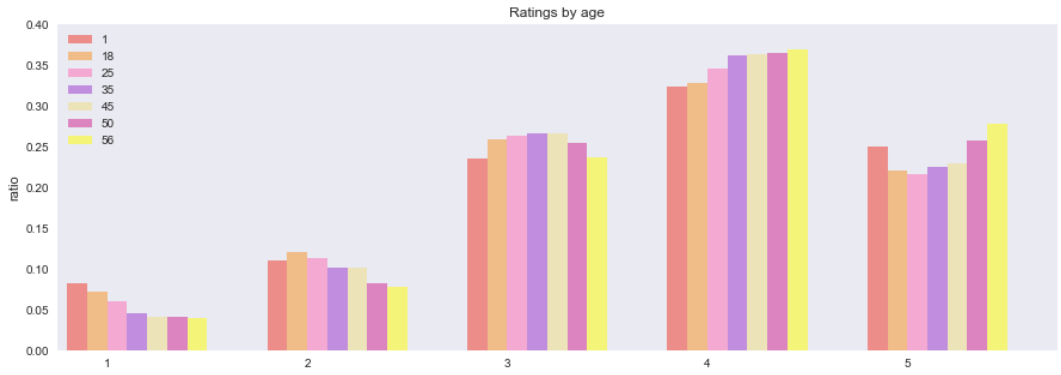

Fig. 6: Distribution of the ratings among users' age groups.

For this distribution, we first found the number of ratings done per rating *i*—where *i* is from 1 to 5, for each age group, and then divided that number over the total number ratings to get the ratio. Fig 6 shows that people younger than 25 seem to be more critical of the movies compared to other age groups, giving mostly the ratings of 1 and 2. On the other hand, they seem to be nearly as enthusiastic as people over 50, giving ratings of 5 more than the middle-aged group, ranging from 35 to 50. People between the ages of 35 and 50 seems to give the highest number of rating 4 compared to the other age groups. It seems that middle aged users tend to ratings greater than or equal to 3 more than ratings of 1 or 2. This may suggest that middle age users have mostly positive opinions it comes to rating a movie or maybe this suggests that they just give ratings mostly to the movies they like. In comparison, the data also suggests that ratings 4 and 5 are predominantly given by users in the age groups above 50 suggesting that older people tend to give higher ratings more frequently compared to the younger generations. Hence our conjecture is that, young users seem to have either very negative or very positive opinions, on the other hand, middle aged users

often give ratings between 3 to 4 and finally the users who are older than 50 seem to be overly happy and positive about the movies they watch or they tend to rate the movies they like the most more.

3.3 CORRELATION: MEN VERSUS WOMEN

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

We first calculated the average ratings for each movie based on gender using pivot_table. Then plotted average ratings of men versus women for every movie in a scatter plot as shown in Fig. 7. There seems to be a good correlation between both genders average rating tendencies.

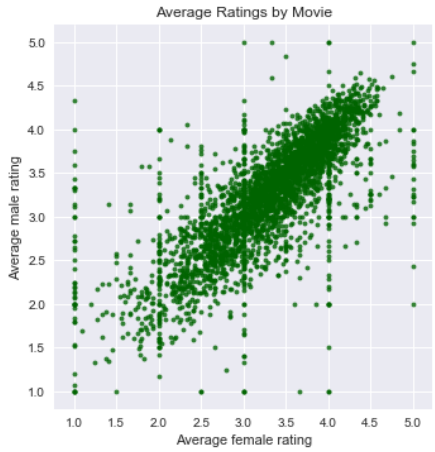

Fig. 7: Scatter plot average ratings of men vs. average ratings of women.

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

We grouped the movies based on the title and calculated the rating size for each movie. Then we selected those that had more than 200 ratings. We then calculated the average ratings of male and female users by locating the movie titles in the data created in the previous step. Finally, we plotted average ratings of men versus women for every movie that had more than 200 ratings in a scatter plot as shown in Fig. 8. There seems to be a stronger correlation than the one we observed in Fig. 7 between both genders average rating tendencies for the movies that are rated more than 200 times. Which suggests that the more the number of ratings we see that there is actually a good correlation between both genders' average rating tendencies.

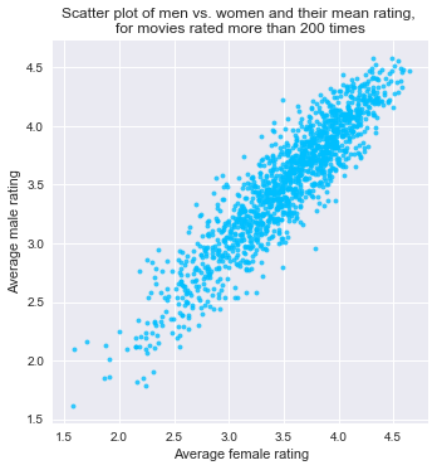

Fig. 8: Scatter plot average ratings of men vs. average ratings of women,

for movies rated more than 200 times.

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

We calculated the correlation between the average ratings of men and women for all and also for the movies that were rated more than 200 times, the using the corr() function, to actually see the value of the correlation that we observed in Figures 7 and 8, shown below. The correlation between genders' average ratings for all the movies in the dataset is 0.763 which suggests a good positive linear relationship between the average rating tendencies of men versus women. This correlation increases from 0.763 to 0.918 for the movies that were rated more than 200 times which suggests that men and women tend to rate on average the same way regardless of the genre of the movie.

```
Correlation:
gender         F         M
gender
F        1.00000   0.76319
M        0.76319   1.00000

Correlation with movie over 200 ratings:
gender          F          M
gender
F        1.000000   0.918361
M        0.918361   1.000000
```

3.3.4 Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender. For example, are men and women more similar when they are younger or older?

To know under what circumstances the rating given by one gender can be used to predict the rating given by the other gender we segregated the data into above defined age groups by querying technique. The groups are as follows: "Under 18", "18-24", "25-34", "35-44", "45-49", "50-55" and "56+". Then for each group we calculated the mean and plotted in a scatter plot as shown in Fig. 9.

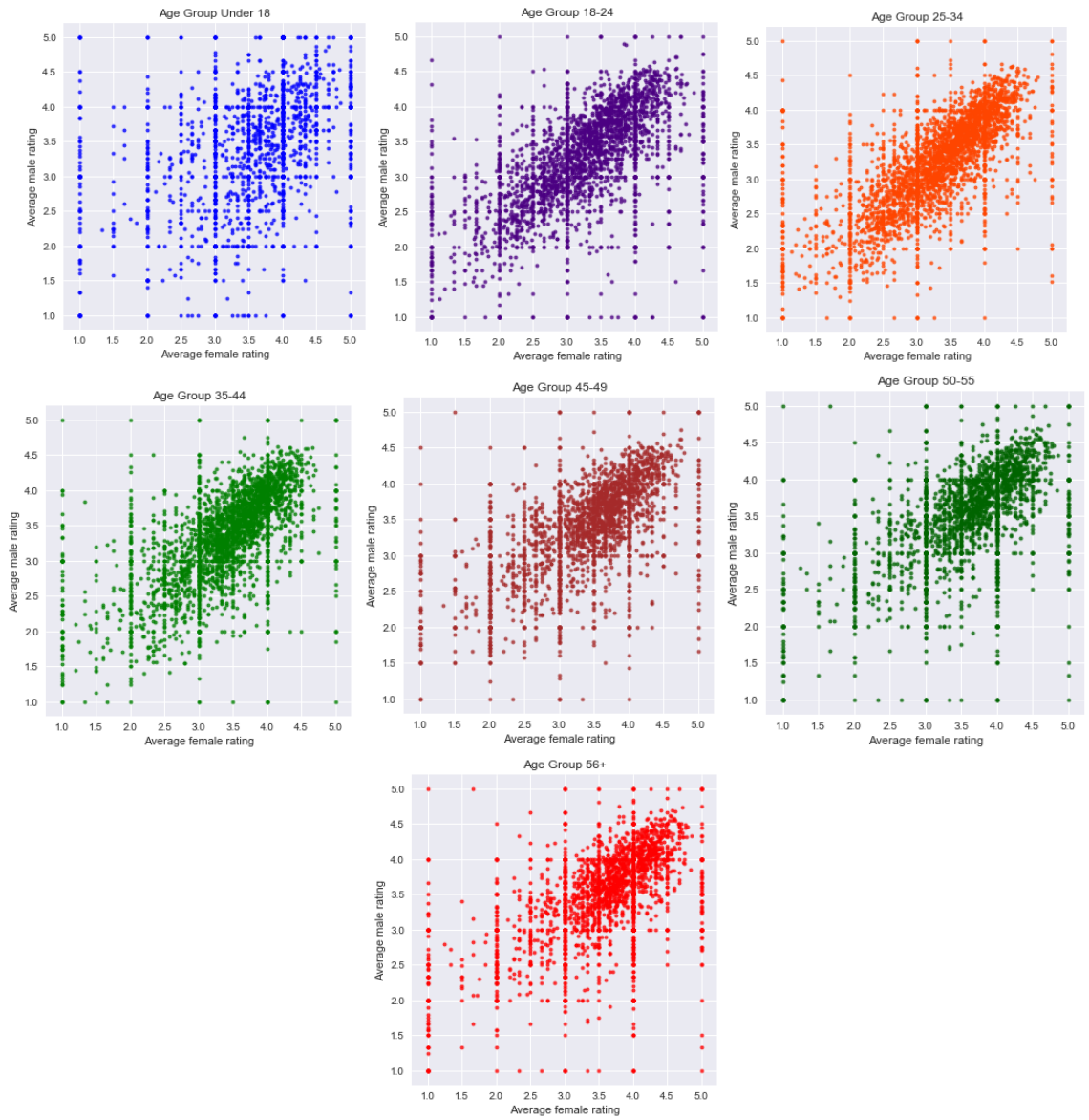

Fig. 9: Scatter plot average ratings of men vs. average ratings women based on age groups.

We also calculated the correlation for each we got the following values: "Under 18" : 0.348, "18-24": 0.576, "25-34": 0.686, "35-44": 0.599, "45-49": 0.569, "50-55": 0.537, "56+": 0.492.

From these results we observe that the correlations among men and women rating tendencies are somewhat similar between the ages of 18 to 55. We looked at different combinations of age groups to see if we could get a better understanding of the correlation between men and women's average rating tendencies. We first looked at the relation for the ages older than 25 as shown on the left in Fig. 10. The corellation between male and female average rating for users older than 25 is 0.705. Then we combined the ages of 18 to 55, as shown Figure 10, and the correlation was stronger as it rose to 0.759 for the ages between 18 to 55.

Fig. 10: Scatter plot average ratings of men vs. average ratings women for users older than 25 (on the left), and for users between the ages 18 and 55 (on the right).

Hence in order to be able to use the rating given by one gender to predict the rating given by the other gender we first need our data to have large number of ratings for each movie preferably more than 200 and we also need our users to be between the ages of 18 and 55 as there seems to be stronger correlation between those age groups even if we look at the whole movie dataset without any restrictions on the rating number.

3.4 BUSINESS INTELLIGENCE

We decided to do some more exploration on this dataset to gain more insights based on gender, genre, location and occupation and wanted to answer the following question: "How a movie company could promote the movie according to the genre, location, age, gender and occupation related insights from the data". Below we share are our findings and discuss our business question.

3.4.1 Exploring average rating differences based on gender among movies rated more than 1,500 times.

After finding the correlations between male and female average ratings we also wanted to see the differences between the average ratings based on gender for the movies that were selected based on having ratings more than 1,500. Choosing this number ensured that we were looking among the movies that were more on the popular side since highly rated movies in this dataset seemed to have higher average ratings both by men and women. Results are shown in Fig. 11. Positive difference in the figure means that the movie was favored more by men than women. One interesting result was that the movie "Independence Day" was favored more by women than men even though it is an "Action" movie. Considering it is an "Action" movie one might think it would have been favored more by men. So perhaps, for action movies a movie company could promote their movie not only with men in mind but also women as they seemed to favor it as well—according to our results in later sections as well.

3.4.2 Finding unique genres and calculating the average rating for each genre based on gender.

We retrieved unique genres from the genres in movies dataset and for each movie we assigned values according to the genres they belonged. For example, if a movie belong to "Animation and Comedy" then for that movie we assigned the value 1 both to Animation and to Comedy. Based on this data we calculated the number of movies in each genre. Figure 12 shows the bar chart we obtained from on this calculation.

From this figure we see that the major number of movies in this dataset belong to “Drama” and “Comedy” genres.

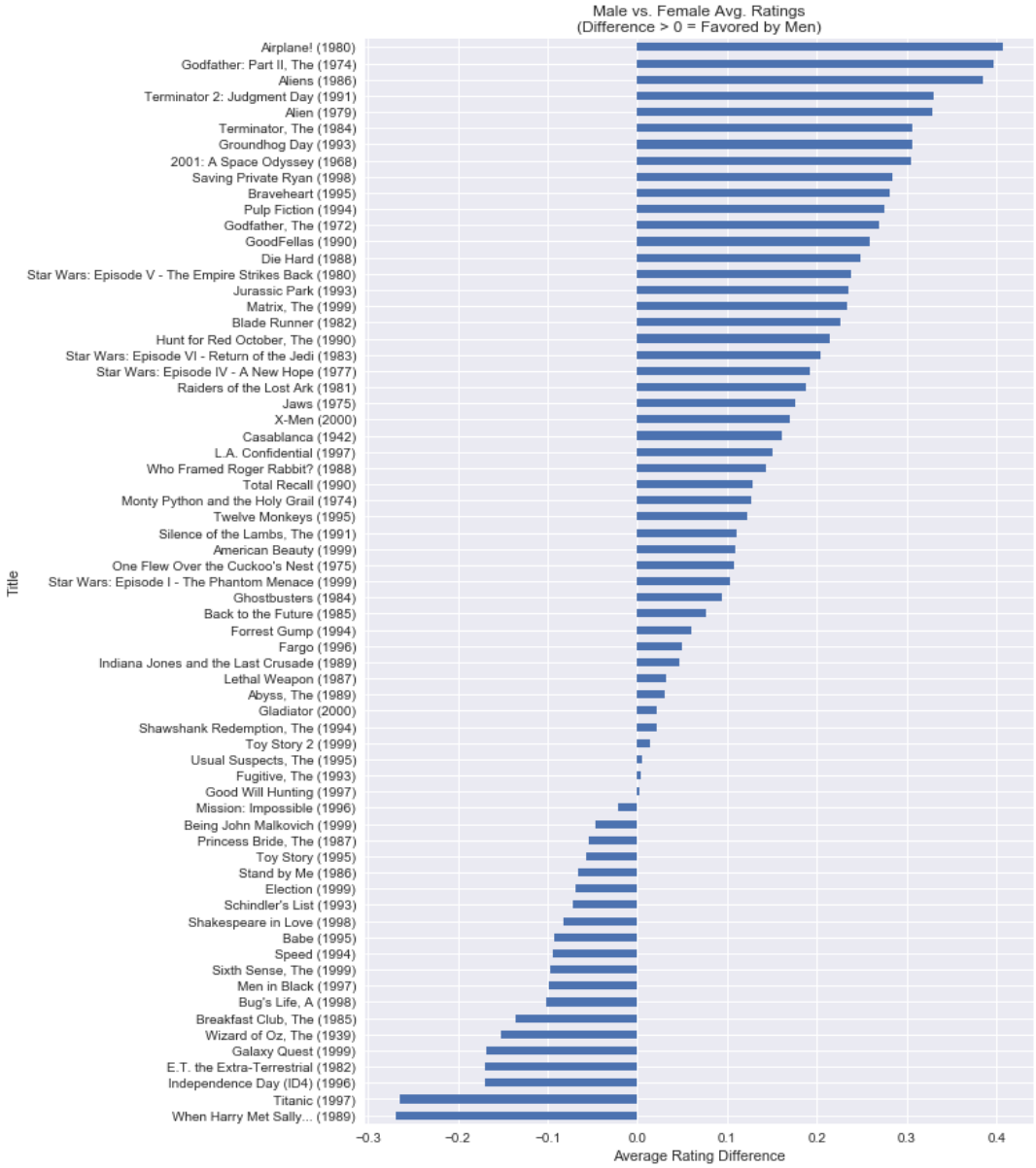

Fig. 11: Average rating difference between men vs women average ratings for movies rated more than 1500 times.

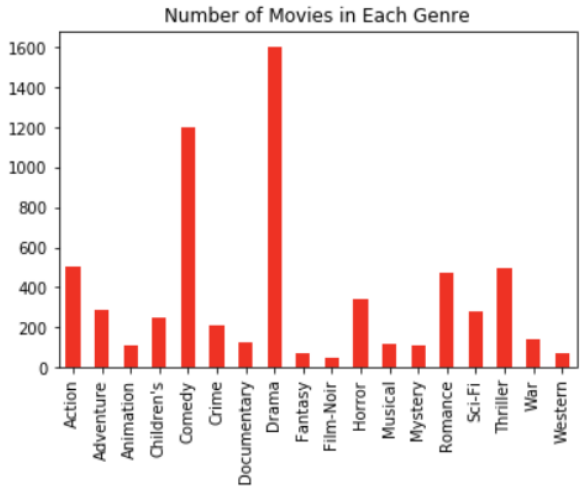

Fig. 12: Number of movies in each genre.

We then calculated the total rating given by users for each movie in these genres as shown in Figure 13. From this figure we see that users tend to rate "Comedy", Drama" and "Action" movies more frequently compared to other genres. Since some of the genre sizes—like "Documentrary" and, "Film-Noir"— are so small the results we obtained on these genres from our analysis might not be quite as the representative of that genre. We will need to have more sample points in our data to have a better understanding.

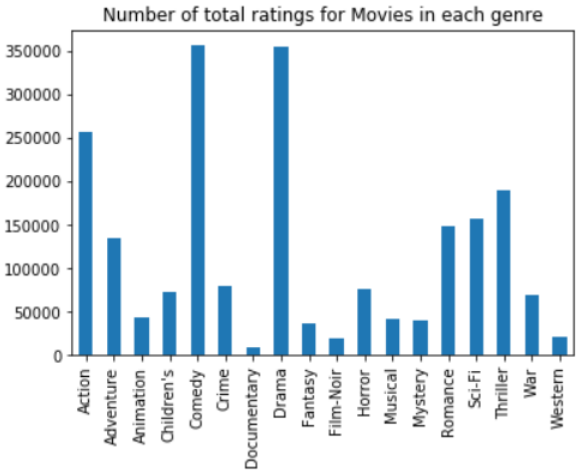

Fig. 13: Number of total ratings for movies in each genre.

We then looked at the movies that had more than 250 ratings. For each genres associated to the movie, they are splited and the new entry is combines into the dataset. Then we calculated the average rating for each genres based on gender. Then sorted the data in descending order based on the female and male average ratings. Table 11 shows both female and male preferences ranked.

genres	female	male
Documentary	4.135135	4.135769
Film-Noir	4.043739	4.124527
War	3.921579	3.953934
Musical	3.862081	3.657053
Drama	3.855978	3.855053
Animation	3.837322	3.749148
Romance	3.749225	3.630318
Crime	3.733396	3.772815
Children's	3.718653	3.533618
Mystery	3.716185	3.701563
Comedy	3.669676	3.621830
Thriller	3.632778	3.639406
Fantasy	3.566215	3.495121
Adventure	3.561616	3.526135
Western	3.556539	3.665983
Action	3.534964	3.553601
Sci-Fi	3.491003	3.517634
Horror	3.369593	3.435138

genres	female	male
Documentary	4.135135	4.135769
Film-Noir	4.043739	4.124527
War	3.921579	3.953934
Drama	3.855978	3.855053
Crime	3.733396	3.772815
Animation	3.837322	3.749148
Mystery	3.716185	3.701563
Western	3.556539	3.665983
Musical	3.862081	3.657053
Thriller	3.632778	3.639406
Romance	3.749225	3.630318
Comedy	3.669676	3.621830
Action	3.534964	3.553601
Children's	3.718653	3.533618
Adventure	3.561616	3.526135
Sci-Fi	3.491003	3.517634
Fantasy	3.566215	3.495121
Horror	3.369593	3.435138

Table 11: The average rating for movies in each genre based on gender. Ranked for female preference on the left and male on the right.

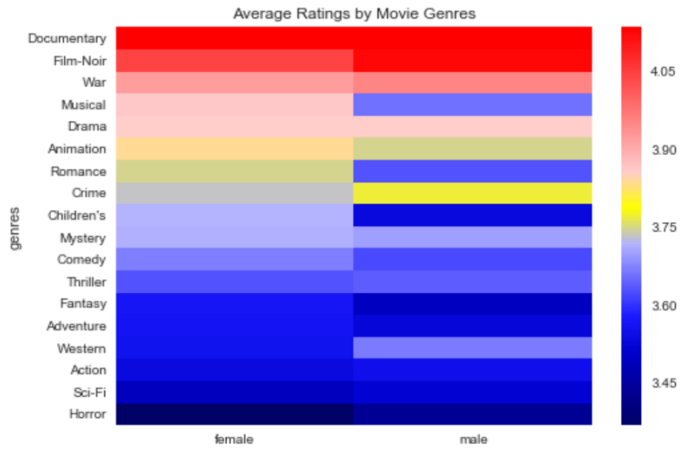

Fig. 14: Heat map of the Average ratings for movies in each genre based on gender.

We also displayed these average rating as a heat map as shown in Figure 14. From these results we see that, both genders seem to have similar taste for most genres. Furthermore, both genders seem to like the "Documentary", "Film-Noir" and "War" genres the most as they correspond to with the highest average ratings in the list. Though, considering the rating sizes of these genres one could also say that "Drama" is

most liked by females and “Crime” movies are favored more by men. Interestingly “Horror” genre was the least liked—with an average rating more than 3.3 for women and 3.4 for men—genre among both genders. Though this again could be due to the rating size of this genre in the dataset. From these findings, a movie company should choose to promote the movies in any genre by keeping both genders in mind except “Romance”, “Musical”, and “Children’s”. For the latter group movie company should promote the movie more towards females rather than males, or make a movie—say in a “Romance” genre—that might speak to both genders equally.

3.4.3 Calculating the average rating for each genre based on age and location.

We first filtered the dataset to select the movies that had more than 250 ratings. We then used the age groups we defined earlier and also extracted the first digits of zip code attribute as “area” using the assignment of zip code zones as shown in Figure 14 [2].

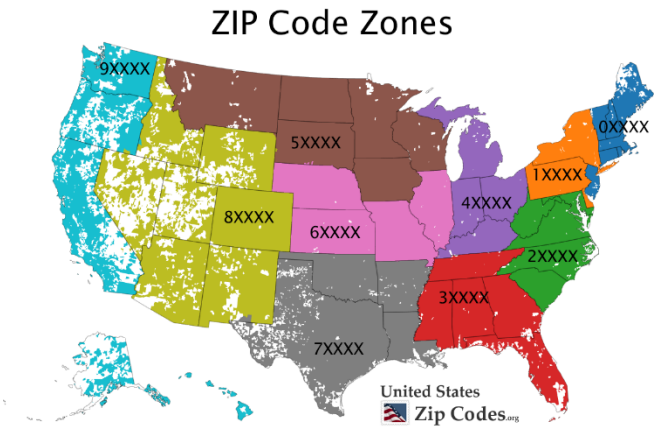

Fig. 14: Zip code zones.

We then plotted the heat maps of average ratings of users for each age group for each genre and average ratings of users in each area for each genre, as shown in Figures 15 and 16, respectively. From Figure 15, we see that “Documentary”, “War” and “Film-Noir” genres receive highest average rating among all age groups and areas, meanwhile, “Horror”, “Sci-Fi” and “Fantasy” genres tend to receive a relatively low rating in general. However, these findings may be affected by the number of ratings each genre received. Apart from these patterns, there are subtle differences between taste of different age groups. For example, younger people under the age 25 seems to think poorly of Western movies and Children’s movies, which is very interesting. It was also interesting to see that users under the age of 25 tend to like “Film Noir” genre. The preference of people over 50 years old seems similar though as we have found out from our analysis before users aged above 56 seem to have a tendency to be more generous on ratings, which is also reflected here. User over 50, seem to like “Film Noir” genre the best. As seen in Fig. 16, there seems to be no significant difference of taste among each genres in terms of areas except from the users from the area 8; Nevada, Idaho, Wyoming, Utah, Colorado, Arizona and New Mexico. Users from this area seem to be critical in their ratings, especially for documentary movies. From these we see that, the younger generations are more critical, so while promoting a movie, movie company might want to focus on the younger generations even more to increase the profitability of their movie.

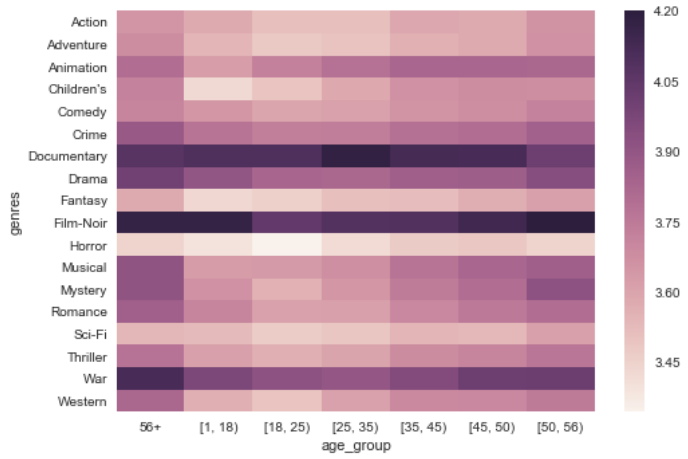

Fig. 15: Heat map of average ratings of users for movies in each genre based on each age group.

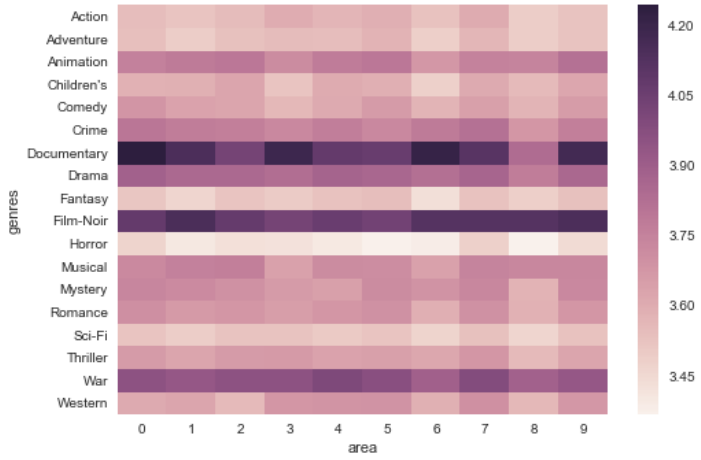

Fig. 16: Heat map of average ratings of users for movies in each genre based on each area.

3.4.5 Calculating the average rating for each genre based on occupation.

We first looked at the distribution of the occupation among users using the occupations in the original dataset. For this we fist made a column with occupation names using the occupation codes in the data; 0: "other" or not specified, 1: "academic/educator", 2: "artist", 3: "clerical/admin", 4: "college/grad student", 5: "customer service", 6: "doctor/health care", 7: "executive/managerial", 8: "farmer", 9: "homemaker", 10: "K-12 student", 11: "lawyer", 12: "programmer", 13: "retired", 14: "sales/marketing", 15: "scientist", 16: "self-employed", 17: "technician/engineer", 18: "tradesman/craftsman", 19: "unemployed", 20: "writer". Then calculated the number of users in each category as shown in Fig. 17.

occupation_names	user_id
K-12 student	195
academic/educator	528
clerical/admin	173
college/grad student	759
customer service	112
doctor/health care	236
executive/managerial	679
farmer	17
homemaker	92
lawyer	129
other	711
programmer	388
retired	142
sales/marketing	302
scientist	144
self-employed	241
technician/engineer	502
tradesman/craftsman	70
unemployed	72
writer	281

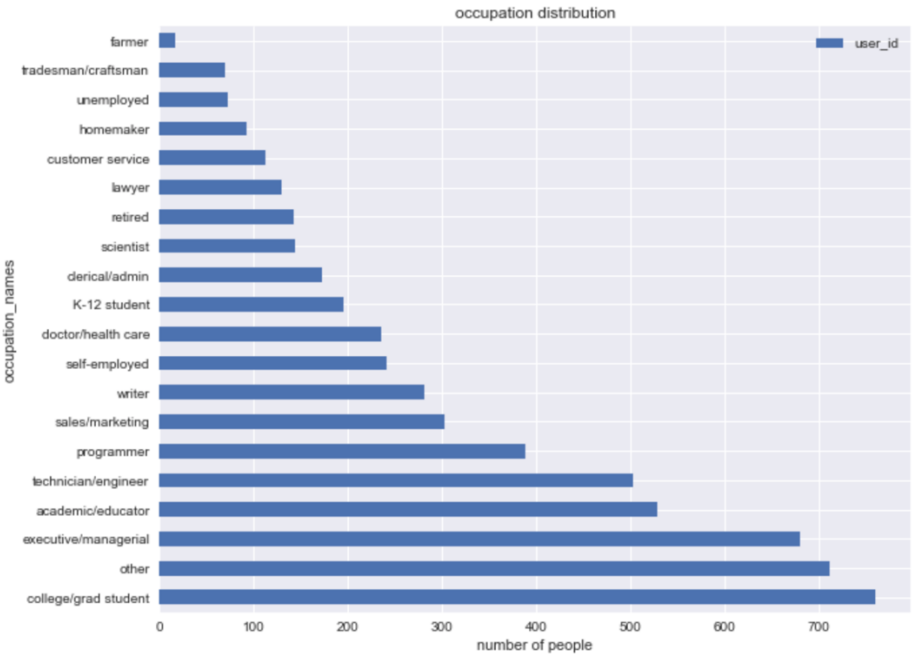

Fig. 17: Distribution of the occupation of the users.

From Figure 17, we see that the size of each group varies drastically, the smallest being 17 and the largest being 711. "College/grad student", "other", "executive/managerial" dominated the users' occupations. We then calculated the average ratings of movies based on occupation and the results are shown in Figure 18. Since some of the group sizes are so small the result we obtained for these occupations from our analysis might not be quite as the representative of that occupation. We will need to have more sample points in our data to have more reliable picture. From the Figure 18, we see that the average ratings seems to be around the same for most occupations. Retired users seem to be more enthusiastic in their rating in agreement with our findings before regarding age groups. Furthermore scientist, and, doctors seems to be most generous in their ratings as opposed to unemployed people, farmers, and, writers who were more critical in their ratings. Though again we have very little number of users that fall into "farmer" and "unemployed" hence these results might not be a representative for those occupations. We also do not know if there is any intersection between unemployed and college/grad student or K-12 student since some of the students might have chosen unemployed as their occupation.

We then plotted a heat map for the average rating for movies in each genres based on occupation shown in Figure 19. From this heat map , we see that "Documentary" and "Film-Noir" are the highest rated movies among all occupations agreeing with our gender based results though, homemakers and tradesman/craftsmans seem to favor "Documentary" and "Film-noir" a little less compared to the other occupations. "Horror" genre seems to be the least liked among all occupations again in agreement with our earlier results. However we need to emphasize again the number of movies that fall into these categories were so low hence we might be seeing this result because of that. Furthermore, no matter what the occupation, users seem to be more critical when it comes to "Fantasy", "Action", and "Adventure" movies. If we take into account that we are seeing discrepancies in the "Documentary", "War" and "Film-noir"

ratings most probably due to not having enough sample size on these genres one could say that over all for each genres, occupation does not seem to have a huge impact on peoples ratings. At least this is what we see from this dataset. Further analysis could be done on a larger dataset preferrably with equal number of occupation distribution to get a better understanding.

	rating	
	mean	std
occupation_names		
retired	3.781736	1.037604
scientist	3.689774	1.066078
doctor/health care	3.661578	1.109623
homemaker	3.656589	1.033470
clerical/admin	3.656516	1.097272
programmer	3.654001	1.083001
sales/marketing	3.618481	1.092451
lawyer	3.617371	1.148651
technician/engineer	3.613574	1.075178
executive/managerial	3.599772	1.083684
self-employed	3.596575	1.074638
academic/educator	3.576642	1.107499
other	3.537544	1.126175
customer service	3.537529	1.096425
college/grad student	3.536793	1.165478
K-12 student	3.532675	1.224677
tradesman/craftsman	3.530117	1.065832
writer	3.497392	1.154047
farmer	3.466741	1.103097
unemployed	3.414050	1.219017

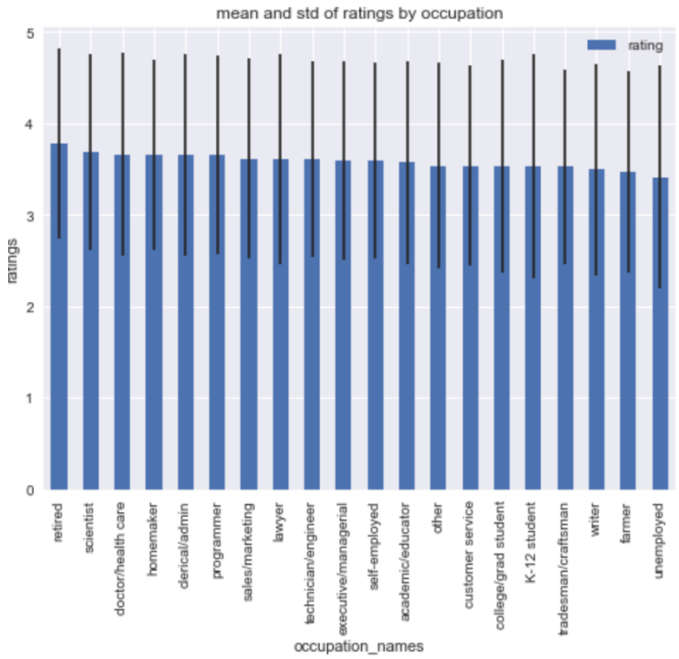

Fig. 18: Average ratings for based on occupation.

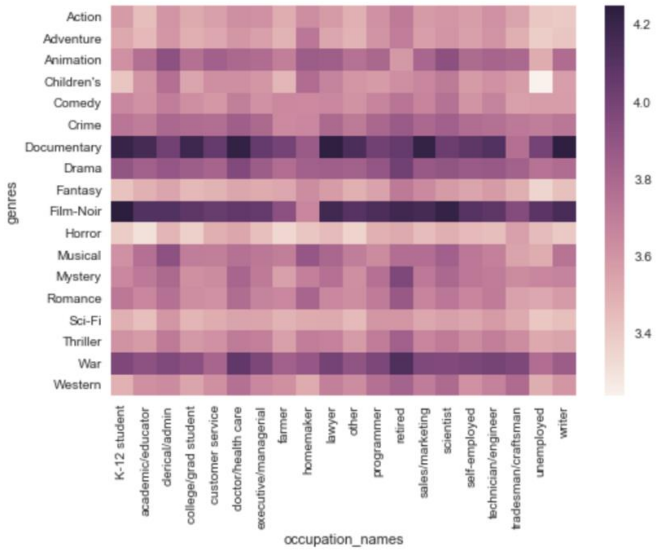

Fig. 19: Heat map of average ratings of users for movies in each genre based on occupation.

REFERENCES

[1] Grouplens. https://grouplens.org/datasets/movielens/

[2] United States Zip Codes. https://www.unitedstateszipcodes.org/