Introduction to Data Science

DS501 Case Study 2 Team 5

Analyzing data from MovieLens

Submitted By
Harsh Nilesh Pathak,
Jidapa Thadajarassiri,
Krushika Tapedia,
Pitchaya Wiratchotisatian,
Prince Shiva Chaudhary.

Supervisor:
Dr. Randy Paffenroth

Overview and Motivation

In this study, we assume that our business is a movie company that telecast movies nationwide in the United States. To improve our business, various business decisions need to be wisely made. For example,

1. What movies/tv shows should we make and telecast and when should we do?
2. When should we launch upcoming movie trailers?
3. Which movies/tv shows should we telecast at which location?

These questions are not obvious to answer. In order to assure wisely decisions, we do data analysis using 1 million movie rating dataset from MovieLens to support our strategic making. Frequency analysis, correlation analysis and many visualizations such as histogram or scatter plots are applied on this dataset to help us understanding and making further decisions for our business.

Data Description

The GroupLens Research has collected and made available rating data sets from the MovieLens web site. We select the 1M data set consisting of 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 users who joined MovieLens in 2000[1]. Each of the three files contains the following features:

File name	**Features**
rating.dat	UserID::MovieID::Rating::Timestamp
users.dat	UserID::Gender::Age::Occupation::Zip-code
movies.dat	MovieID::Title::Genres

Data Preparation

Data Download: Download the 1 million movie ratings dataset for data files namely (Movies, Ratings and Users) from http://grouplens.org/datasets/movielens/.

Data Merging**:** Converted these data files into pandas data frames and merged together using below steps:

- Left join on rating and movies data frames so that we get all the movies which received ratings from users (If we do movie to rating left join, there can be movies which don't have any ratings associated with it leading to produce NaN values).
- Left join on resultant dataframe of [ratings and movies] with users data frame. This result will give us data where users have given rating and excluding those who never gave any rating to any movie. We can consider such userId dormant.

Data storage: After merging data frames, stored resultant data frame in HDF5 file on disk.

[1] http://files.grouplens.org/datasets/movielens/ml-1m-README.txt

	UserID	MovieID	Rating	Timestamp	Title	Genres	Gender	Age	Occupation	Zip-code
0	1	1193	5	978300760	One Flew Over the Cuckoo's Nest (1975)	Drama	F	1	10	48067
1	1	661	3	978302109	James and the Giant Peach (1996)	Animation|Children's|Musical	F	1	10	48067
2	1	914	3	978301968	My Fair Lady (1964)	Musical|Romance	F	1	10	48067
3	1	3408	4	978300275	Erin Brockovich (2000)	Drama	F	1	10	48067
4	1	2355	5	978824291	Bug's Life, A (1998)	Animation|Children's|Comedy	F	1	10	48067

Data Analysis and Results

Part 1: Some basic details of the data

Below are the results of basic details about the data:

- *How many movies have an average rating over 4.5 overall?*
  There are 21 movies having average rating over 4.5.
- *How many movies have an average rating over 4.5 among men? How about women?*
  There are 23 movies which received rating above 4.5 among men and there are 51 movies which received rating above 4.5 among women.
- *How many movies have a median rating over 4.5 among men over age 30? How about women over age 30?*
  There are 86 movies having a *median* rating over 4.5 among men over age 30 and 149 movies having a *median* rating over 4.5 among women over age 30.
- *What are the ten most popular movies?*
  Popular movies: A popular movie is one, which received highest number of ratings. However, ratings can be a subjective to individual to an popular movie, depending on individual's taste or liking of cinema.

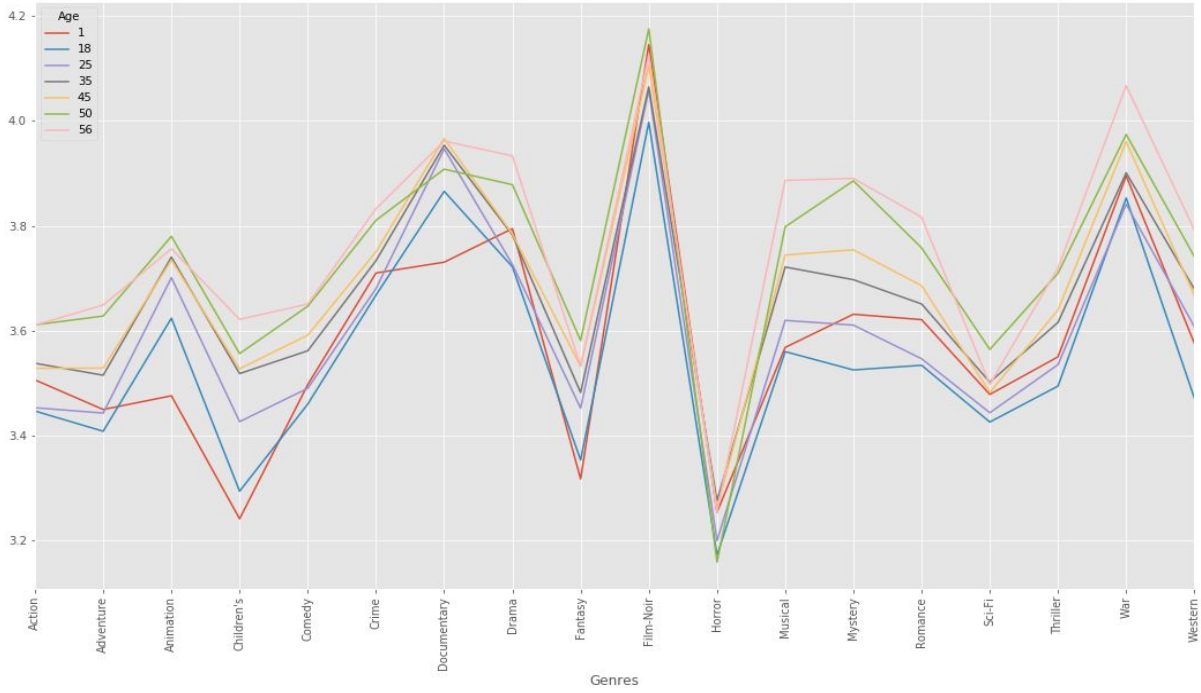
  Screenshot attached gives most popular movies:

Title	Rating
American Beauty (1999)	3428
Star Wars: Episode IV - A New Hope (1977)	2991
Star Wars: Episode V - The Empire Strikes Back...	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

Conjecture: People in different age groups have different rating behavior. Some specific age group might be easier to please.

Methodology: A pivot function (df.pivot_table) is used to create a pivot table of the average rating of each movie genre type.

Result: After going through the result of our analysis we found that people in aged group (18-24 and 25-34) generally give 'low' ratings to almost each segment of movies when compared against old people of age group belonging to 56+ generally give 'high' ratings and they are easy to please.



Part 2: Expand our investigation to histograms

In this part, histograms are used to expand our study. The number of histograms are plotted for the investigation: a histogram of the ratings of all movies (figure 2.1), a histogram of the number of ratings each movie received (figure 2.2), a histogram of the average rating for each movie (figure 2.3) and a histogram of the average rating for movies which are rated more than 100 times (figure 2.4).

Figure 2.1: The ratings of all movies

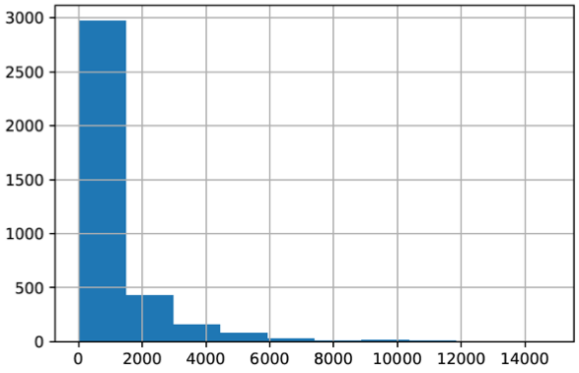


Figure 2.2: The number of ratings each movie received

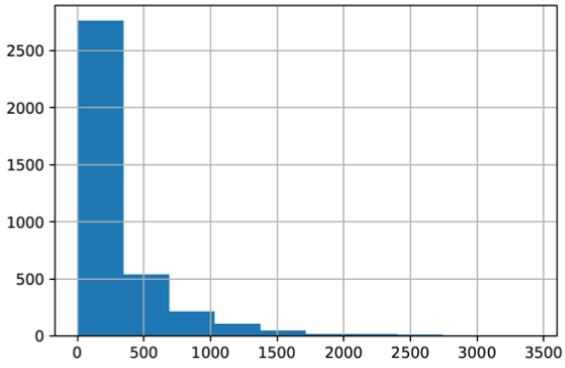


Figure 2.3: The average rating for each movie

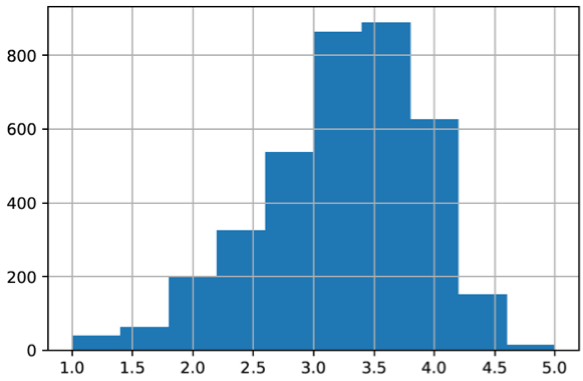
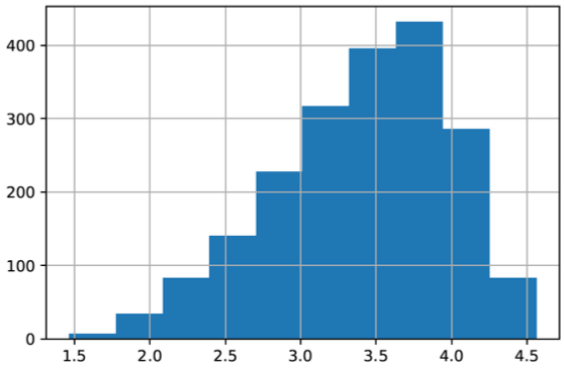


Figure 2.4: The average rating for movies which are rated more than 100 times



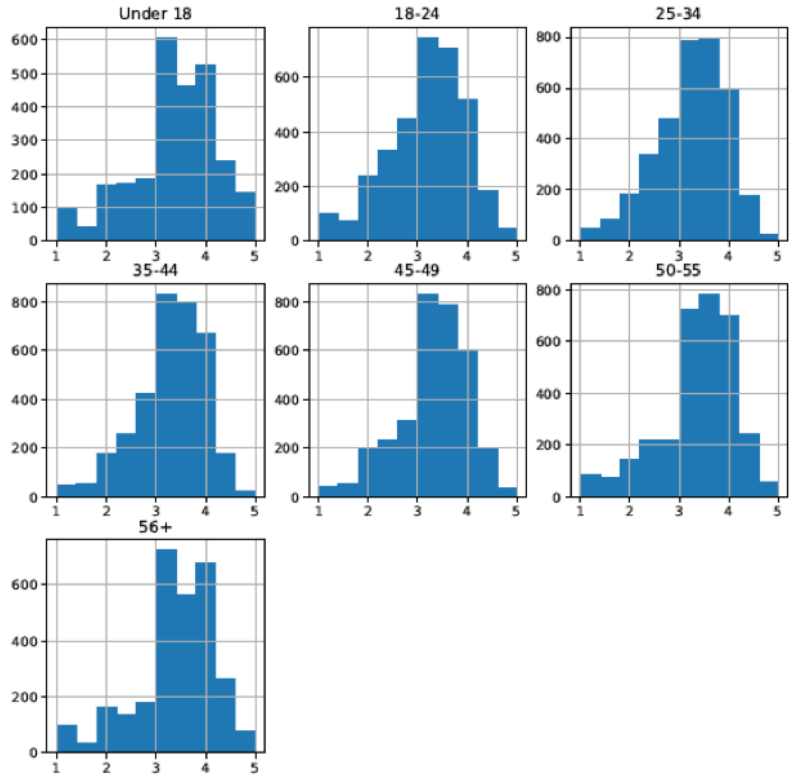
The wider 2-sided tails in figure 2.3 when compared to the tails in figure 2.4 shows that the average rating for highly rated movies (rated more than 100 times) has less spread, which implies less variance and more stable. Therefore, in order to consider the distribution of movie rating, we prefer the average rating of movies that are rated more than 100 times. The distribution of ratings is further investigated; some conjectures are as followings:

Conjecture 1: The distribution of movie rating may be different according to age group.

Methodology : A pivot function (df.pivot_table) is used to create a pivot table of the average rating of each movie against age group and a histogram is plotted using .hist() function Python.

Result : The distribution of average rating on each age group is illustrated in figure 2.5. The most fluctuated takes place in age group 1 (1-18 year-old); they are easier to give either 1 or 5 rating. It might be interpreted that when children under 18 like or dislike a movie, they are not tempting to give the highest or lowest score. For teenagers (age group 18: 18-25 year-old), giving movie rating is less fluctuated than children; they tend to consider more for giving 5-scored rating but still easily give 1-scored. The distribution from adults (age group 25, 35, 45 and 50: 25-55 year-old) is clearly more compacted with less spread in the tails; adults hardly give rating at either peak sides. However, movie rating by elderly people with age higher than 55 year-old seems to be more fluctuated; the distribution is likely as the distribution from children.

Figure 2.5: The average rating for each movie on each age group

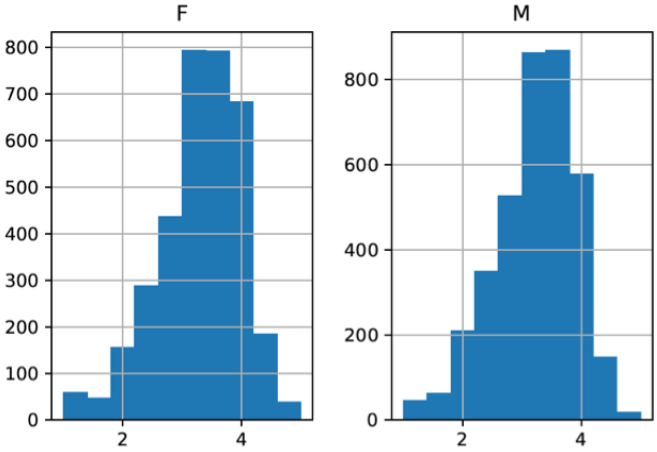


Conjecture 2: The distribution of movie rating may be different according to gender.

Methodology : A pivot function (df.pivot_table) is used to create a pivot table of the average rating of each movie against gender and a histogram is plotted using .hist() function Python.

Result : The difference of distribution of average rating between female and male is shown in figure 2.6. The average rating plot for female in the left hand side shows more left-skewed than the plot for male. It seems that women generally tend to give higher rating than men. However, women seem easier giving 1-score for the movies they do not like.

Figure 2.6: The average rating for each movie on each gender

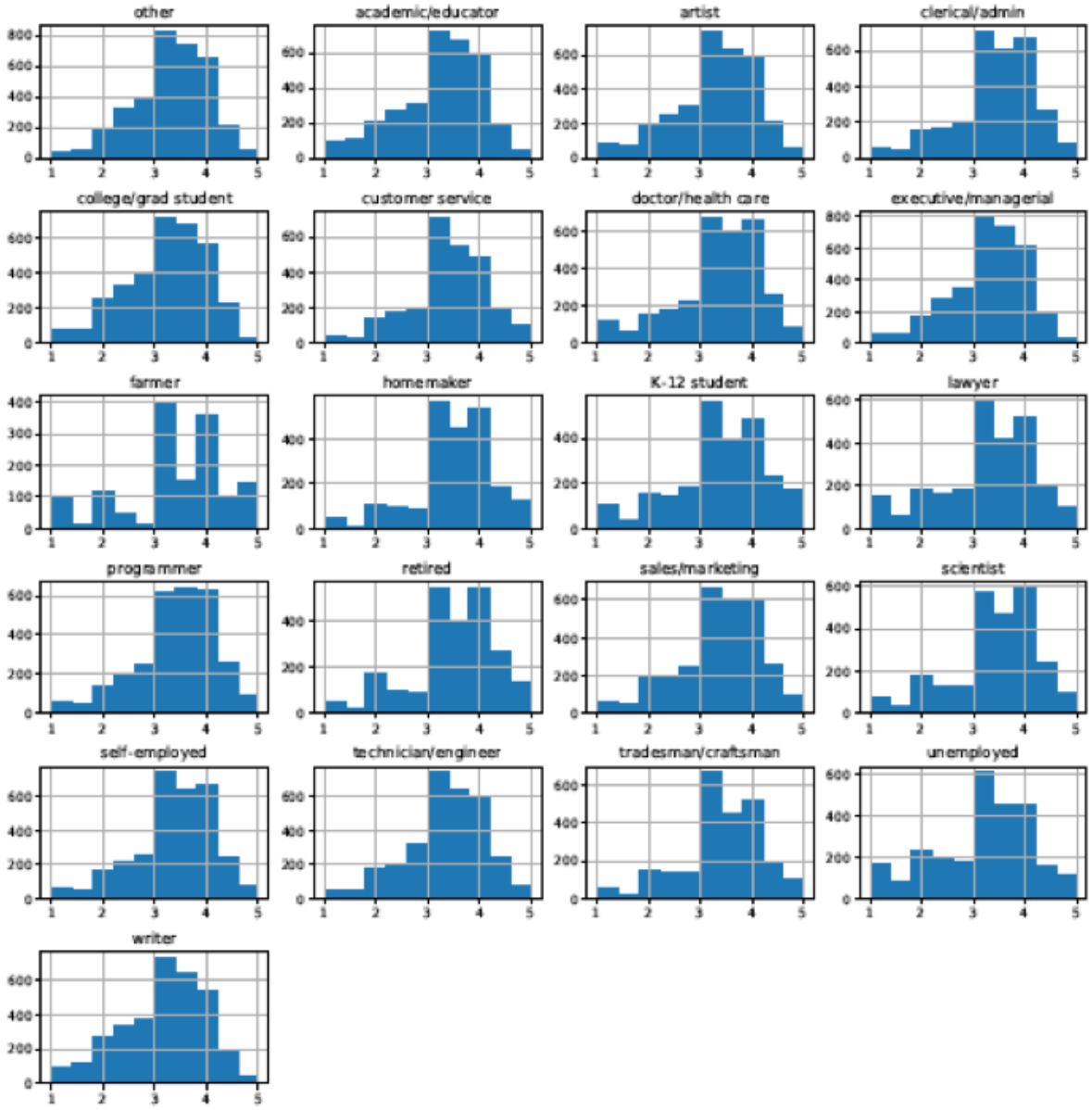


Conjecture 3: The distributions of movie rating may differ according to occupations.

Methodology : A pivot function (df.pivot_table) is used to create a pivot table of the average rating of each movie against age occupation and a histogram is plotted using .hist() function Python.

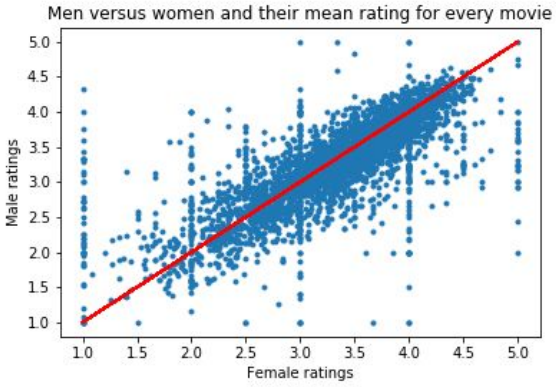
Result : The distributions of ratings by most occupations (figure 2.7) are in the same pattern with left-skewed and having the peak in between giving 3 to 4 of rating score. However, some occupations are easier to give extreme ratings such as farmer, K-12 student, lawyer, retired or unemployed.

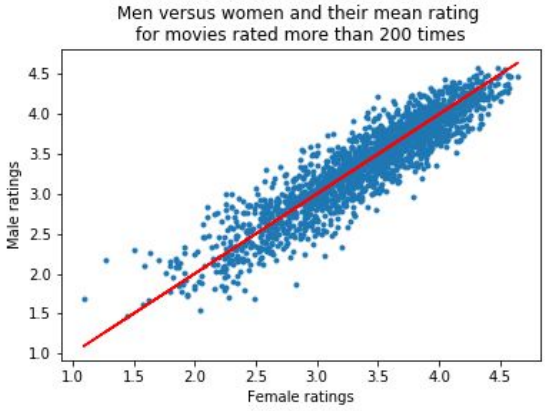
Figure 2.7: The average rating for each movie on each occupation



Part 3: Correlation - men versus women

In this part, we want to look more closely at the relationship between men and women and their mean rating for the movies by investigating scatter plots and correlations of the mean rating of men and women in different aspects.





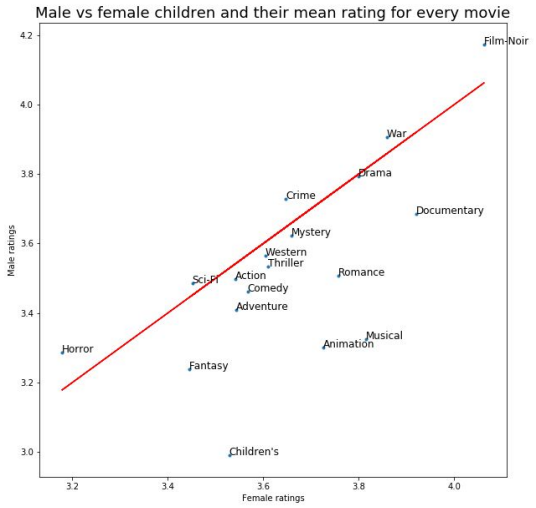
	Correlation
The mean rating of men and women for every movie	0.760231
The mean rating of men and women for movies rated more than 200 times	0.920391

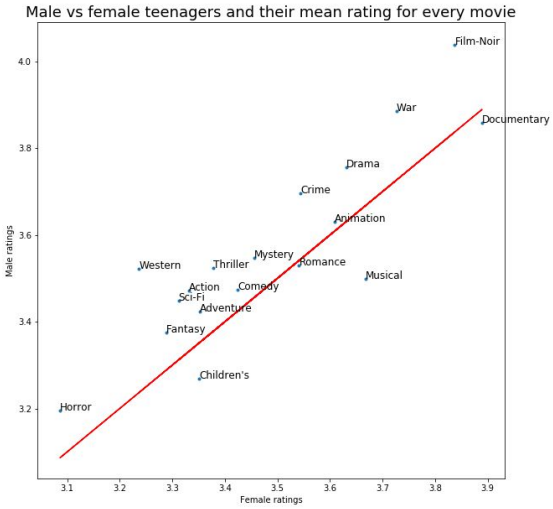
Including every movie, the correlation coefficient between the mean ratings of men and women are strongly positive, around 0.76. After selecting only the movies rated more than 200 times, the correlation coefficient between the mean ratings of men and women increases significantly to 0.92. Both relationships indicate the positive correlation between the mean ratings of men and women.

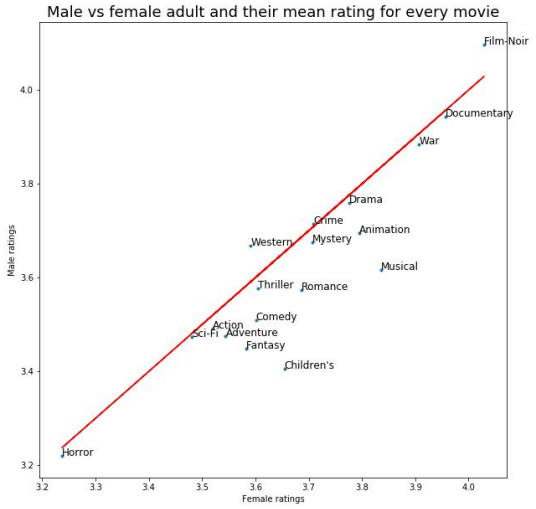
If we select only popular movies, the movies rated more than 200 times, the result of the scatter plot has more predictability. The scatter plot shows points scattering around the straight line y = x which indicating an even strongly positive correlation between the mean ratings of men and women; both genders tends to have similar opinions about popular movies.

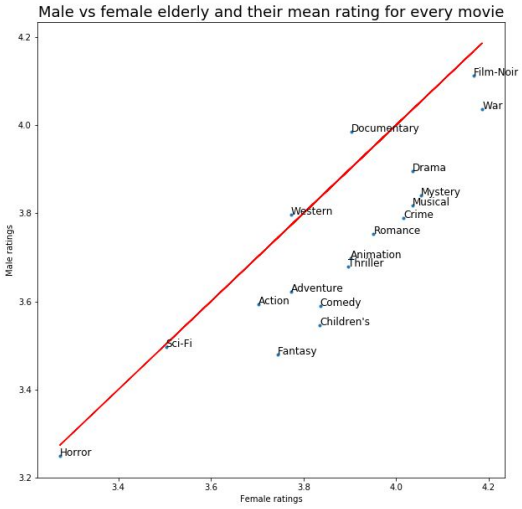
We make three conjectures under circumstances that the rating given by one gender can be used to predict the rating given by another gender.

Conjecture 1: Men and women are more similar for specific age groups, considered their rating over the same genres.









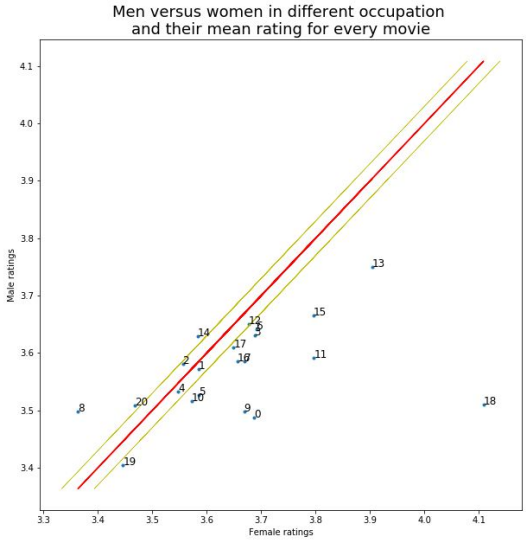
Correlation of male vs female over genres

Age group	
Children	0.6824
Teenager	0.8764
Adult	0.9135
Elderly	0.8821

We take genres into account. From the correlation of male vs. female over the same movie genres across different age groups, men and women more similar when they are older (larger positive correlation).

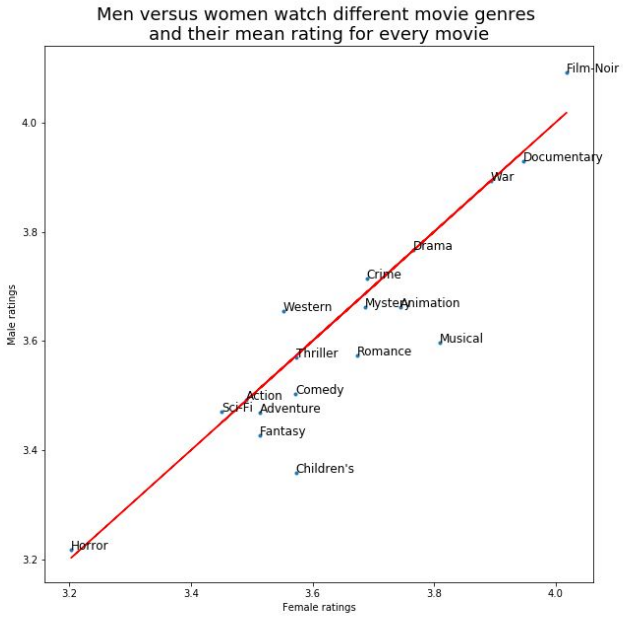
Conjecture 2: Men and women are more similar for some occupations.

The correlation coefficient of the mean rating of men and women in different occupation is around 0.44. This shows that not every occupation has similar ratings of men and women on movies. We wonder in which occupations that men and women are more similar. We make a scatter plot of the mean rating of men vs. women in different occupation below and set a threshold of residual to be 0.03.



From the scatter plot of mean ratings, men and women are more similar when they work as academic/educator (1), artist (2), college/grad student (4), and programmer (12).

Conjecture 3: Men and women are more similar for some genres.



From the scatter plot, we are interested in movie genres that lie on the straight line in which men and women have similar ratings. We find that men and women are more similar when they watch movies in the following genres: Action, Drama, War, and Triller.

Part 4: Business Questions

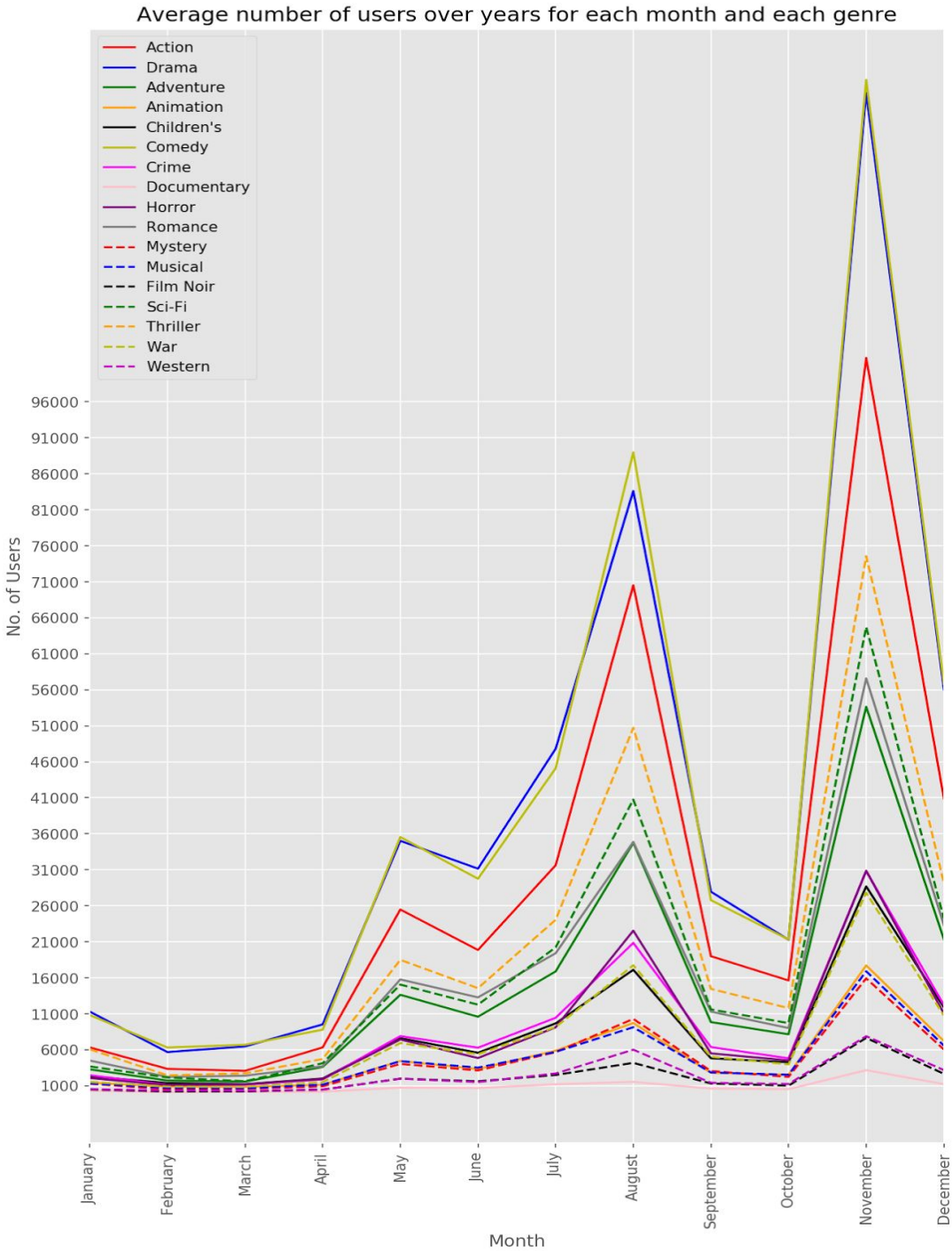
Defining Business Problem: We assume that we are in the business of making and telecasting Movie/TV shows company. We have coverage over world/US/local TV and theaters. So, by making use of the data available, business questions are raised. The detailed analysis with explanation about the questions are shown in this part.

Business Question 1: When and which type of movie most users watch for our movie company to make or telecast the most viewed movie?

To do this, we first create a pivot table by indexing the Genres and Month (extracted from the time stamp), and then counting the number of users who have rated in that particular month for a particular genre.

		UserID
Genres	Month	
Action	1	6341.333333
	2	3360.000000
	3	3080.000000
	4	6340.000000
	5	25485.333333
	6	19868.000000
	7	31630.666667
	8	70521.333333
	9	19002.666667
	10	15634.666667
	11	102094.666667
	12	40944.000000
Adventure	1	3288.000000
	2	1741.333333
	3	1582.000000
	4	3622.666667
	5	13644.000000
	6	10596.000000
	7	16885.333333
	8	34749.333333
	9	9852.000000
	10	8161.333333
	11	53642.666667
	12	21366.666667
Animation	1	1437.333333
	2	896.000000

This is how the pivoted dataframe in pandas looks like after indexing over Genre and Months, and then calculating the number of users who have rated for a particular genre in that particular month.



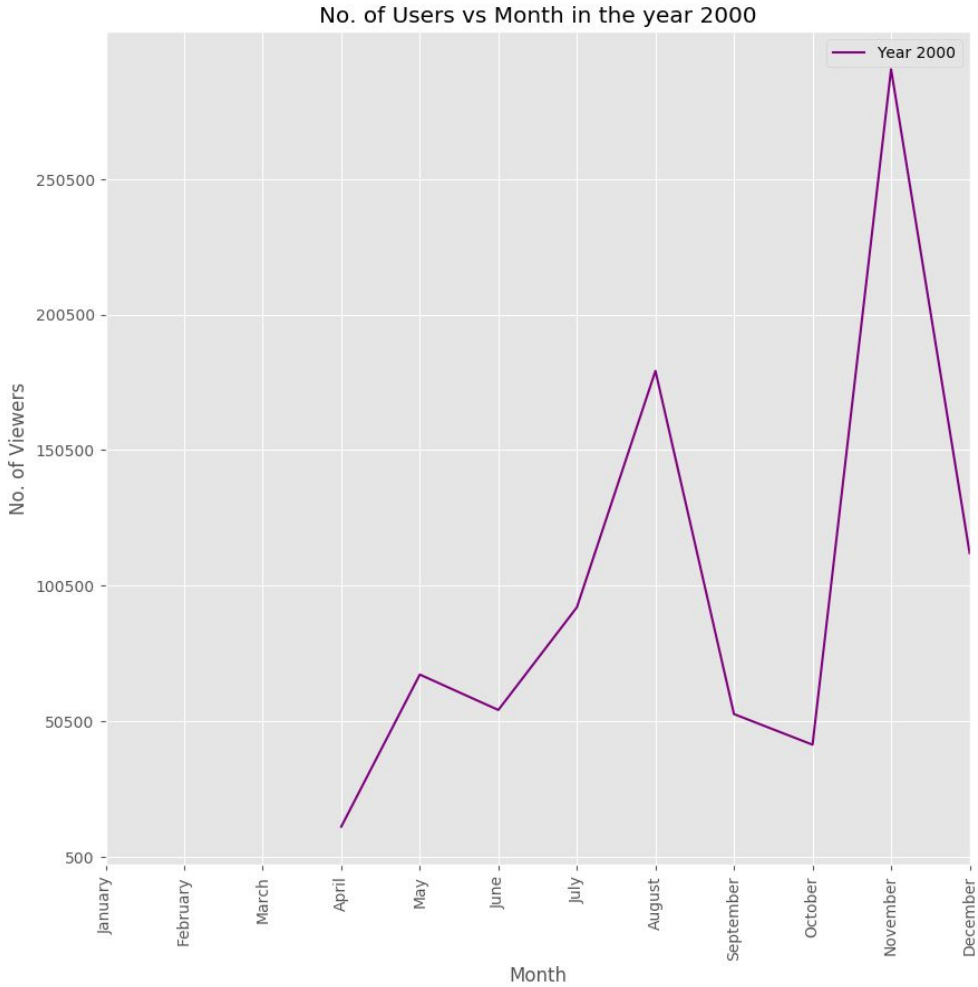
From the above plot we know in which months we have the highest average number of users watching which type of movie. This information is useful in deciding which movie to make/telecast and what timeline could be set on the making and the releasing of the movie. This information could be very significant in improving the sales of the movie company because of the knowledge and trend we have about the target audience's movie watching and rating habits.

One of the observations is that the most watched movie is of Comedy, Drama, and Action, respectively, in November. Another peak time is in August. Since we found that

Comedy and Drama are all time favorite types of movies and people love watching it at anytime of the year, we should invest in making our own comedy and drama movies for cost reduction; therefore, our profit will increase in the long run (valued investment) because these movies can be telecasted any time of a year and once we own them our expense on telecasting will significantly decrease.

Moreover, since people watch a lot of comedy and drama in August and November, in these months we should increase the variety of comedy and drama for telecasting in order to persuade our audiences to stick to our channel. On the other hand, in some months that people watch less different genres such as October or April, our telecasting plan should change according to audience behaviors by increasing the variety of type of movies in order to targeting as variable as audience preferences

Business Question 2: When is our target audience active, and when should we launch upcoming movie trailers?



Since the year 2000 has the maximum information, we choose to plot the rating behavior of the users among 12 months of the year 2000. So, we analyze the data in this year to understand the pattern of preferred time in a year people highly tends to watch movie in order to implement strategic launching for our new movie trailers and any other promotional events.

We can observe from the plot that in the months of November and August most number of users watch movies and rate them, hence advertising or releasing any movie trailers during or before that period to grab the target audience's attention would be profitable.

Business Question 3: How the occupation influence the movie companies? As a movie company it is to have state wise information about the genres but having an added information about the occupation helps the company to explore people with what occupation tend to rate more. Since feedbacks are important for any movie or TV show (local or country-wide) we would have knowledge beforehand, that how much feedback we are going to get from whom and where exactly.

- **Approach :** To answer this we started with grouping by multiindex ordering [location, Occupation ,Genres] and got a custom column we named it user rate frequency (URF). Formally, **URF** can be defined as the number of users of a particular occupation in a particular genre and in a particular location.

Zip-code	Occupation	Genres	UsersRateFreq
00231	0	Romance	74
		Action	51
00606	8	Comedy	78
		Drama	53
00681	10	Sci-Fi	55
00918	3	Drama	112
00961	4	Drama	104
		Romance	56
		Action	172

Zip-code	Occupation	Genres	UsersRateFreq
Alabama	0	Action	36
		Comedy	86
		Drama	76
		Romance	24
		Thriller	51
	1	Action	69
		Adventure	48
		Comedy	143
		Documentary	20

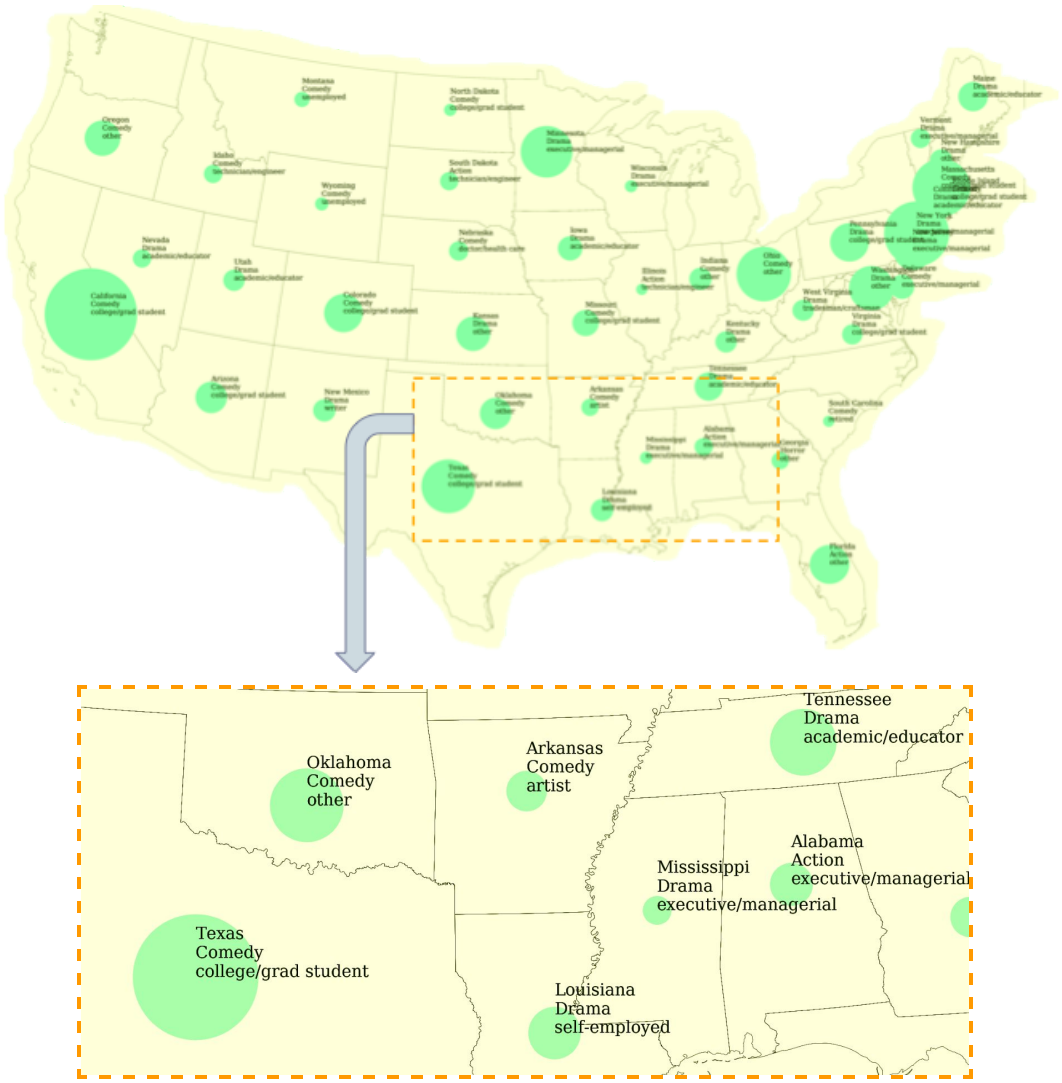
- **Transformations :** Here we applied two transformation to the tables.
  - Exploded the genres. [DRILL DOWN on genres]
  - Contracted the zip-codes to states. [DRILL UP on Zip-codes]
- Getting Top **URFs** (defined above) for various states (best one per state).

Zip-code	Occupation	Genres	
California	4	Comedy	8354
New York	7	Drama	4221
Massachusetts	4	Comedy	3352
Ohio	0	Comedy	2980
Texas	4	Comedy	2816
Minnesota	7	Drama	2650
Washington	0	Drama	1949
Florida	0	Action	1532
New Jersey	7	Drama	1495
Colorado	4	Comedy	1476
Pennsylvania	4	Drama	1445
Oregon	0	Comedy	1257
Kansas	0	Drama	1198

Plotting above table to give a proper visualization was important so that every aspect of the table is covered. So we choose to go with USA map and plot the circles of radius URF and then add other information to it. Below you can see USA map with circles of different sizes. In the zoomed view we can see that the popular genre in among Texas people with

occupation college/gred student is comedy and being a movie company we want feedbacks for our movies and local TV-shows.

The sense behind taking into account occupation in this plot is who are giving the reviews to movies of different genres. Without occupation you may get the most rated genres with respect to location but in that sense feedback interpretation would have been missed.



Minimum Viable Product

So using all of the above analysis on the MovieLens data we can design a MVP for our business problem. In the case that we want to launch a local state movie as our MVP that is to be launched in a particular state only. So in our case

1. In the location California.
2. Genre we would most likely to use is Comedy | Action.
3. Launch period for the movie should be around November.
4. And we are most likely to get our feedbacks from College/Grad student.

Summary

For making decisions on our movie business, we analyzed 1 million movie rating data from MovieLens. The data was stored in HDF5 and manipulated using Pandas library. First, some basic analysis were made. We found that 21 movies got average rating greater than 4.5; more

information was gained by taking gender into consideration. Top 10 popular movies were shown by ranking movies with the number of people rated on each movies and selecting the top 10 highest. Doing basic analysis also showed that people in age group between 18-24 and 25-34 are the most difficult to please (give low ratings) while elderly people with age 56+ are the easiest to please (give high ratings).

Histograms, then, were used for further analysis. The histogram of average ratings illustrated that the distribution is left-skewed. By filtering only movies rated more than 100 times, we obtained a more preferable histogram with less variance. The distribution of ratings under different conditions helped us to more understand our customers as followings:

- Adults with age 25-55 have the least fluctuated in movie rating following by teenagers (age 18-24), elderly (age 56+) and children (age under 18) respectively
- Women are generally give higher rating score (more left-skewed distribution) than men but easier to rate 1-score.
- Most occupations have left-skewed distribution with peak between 3 to 4 score except farmer, K-12 student, lawyer, retired and unemployed who are easier to give extreme rating

Further analysis were made by considering correlation of average rating of men versus women. We found that both genders tend to have similar opinions towards popular movies (rated more than 200 times). Additionally, men and women have similar opinion when: they watch movies in Action, Drama, War, and Triller; when they are older, consider their rating over the same genres; and when they work as academic/educator, artist, college/grad student, and programmer.

For the business perspective, we tried to answer when, how, where to launch the movie/TV show and from which occupation we are most likely to get feedbacks. Accordingly we also designed our MVP for our business which gives us a good direction to start.