Movie Data Exploration and Recommendation

# -----DS501 Case Study 2: Analyzing data from MovieLens

### **Team 11**

### Phaterpekar, Ankit Manohar

### Yi, Lingyun

### Su, Yupeng

### Guram, Jui Saba

### Cui, Xiaoting

## Background

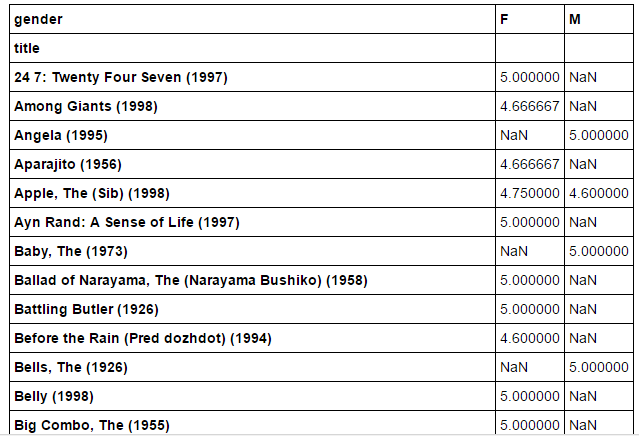
With the big data collected online or from internet of things, companies will be able to gain deeper understanding of their customers’ taste by exploring the data. In addition, they are trying to personalize their service for each individual customer by predicting their preference. Recommendation system plays a critical role in providing personalized service. Movie service providers, such as Netflix, personalize their movie recommendations for subscribers in order to better maintain customer loyalty. Electronic commerce companies, like Amazon, personalize the product recommendation for customers to sell more products. Social Media companies, like Instagram, personalize user recommendation to make customers spend more time on the website.

MovieLens, a movie recommendation website, predicts the rating on movies by each user using a variety of recommendation algorithm and they recommend the high rating movies to the user. At the same time, MovieLens also shares movie rating datasets for research purpose. Our team will do data exploration and build up movie recommendation engine with the 1M dataset, which contains 1 million ratings from 6000 users on 4000 movies.

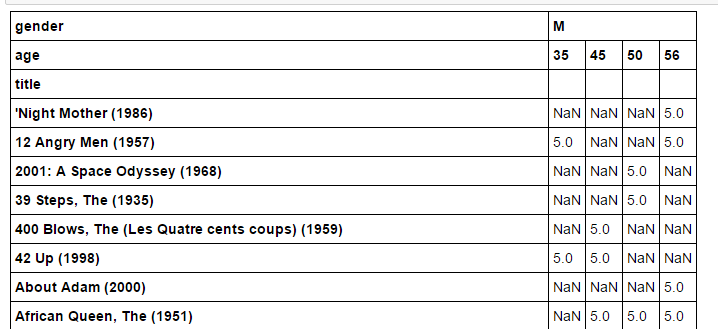
# Data Exploration and Analysis

## Problem1- Basic Data Details

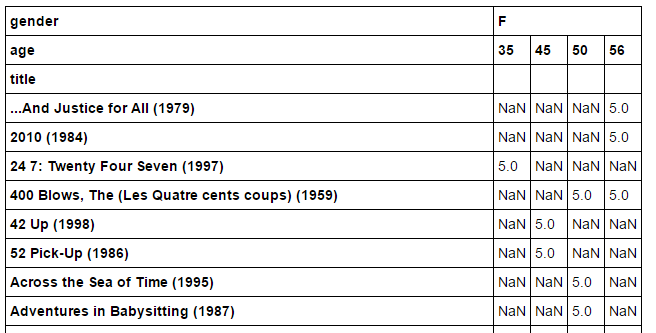
We did basic analysis for the movies that have average rating over 4.5 by gender. There are 68 movies have average rating over 4.5. Within these movies, 23 movies rated over 4.5 among men while 51 movies rated over 4.5 among women. The table below shows the part of table for the movies rated over 4.5.



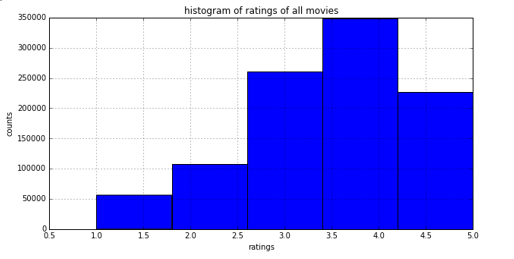
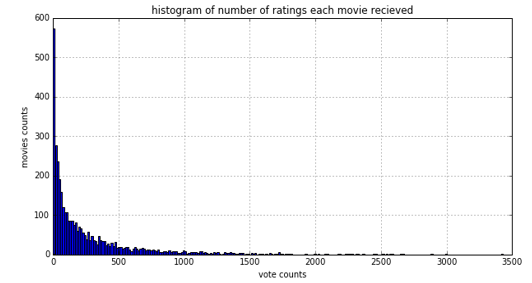
For the movies that have median rating over 4.5, 381 movies among men over 30 while 622 movies among women over 30. Here is a part of table for the movies have median rating over 4.5 among men:



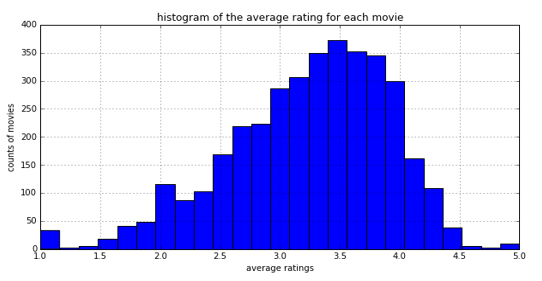
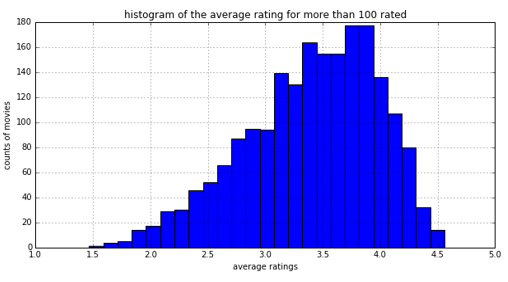
Here is a part of table for the movies have median rating over 4.5 among women:



## Problem2 – Analysis on Histograms

**

*Figure 1: Histogram of ratings of all movies Figure 2: Histogram of number of ratings each movie recieved*

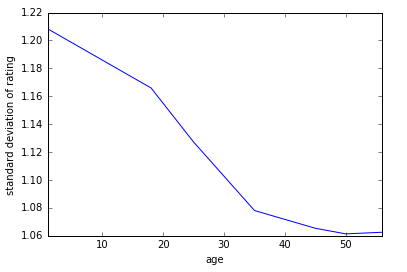
**

*Figure of Average Rating for all movies Figure of Average Rating for movies more than 100 rated*

From the histograms above, it’s easy to see that once the movie which is rated less than 100 times are eliminated from the data set, the distribution of histogram is basically belong to normal distribution, which is more reasonable than the original one.

So according to the analysis above, we confirm that those rated more than 100 times own the high ratings are actually good.

**Conjectures: the younger people tend to have more extreme ratings, and super – natural type movie is the one get more extreme ratings among the young group.**

Since standard deviation can measure the dispersion of the data, so all the following analysis is based on the standard deviation.

According to the figure above, it is easy to see that the age of under 18 tends to have largest standard deviation of rating. In the other words, they tend to give the extreme high or low ratings to the movies among other groups of people. On the other side, while people getting older, their ratings are closer to the mean. This supports well with the conjectures.

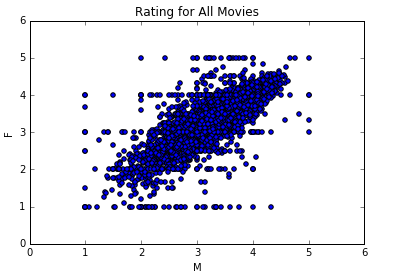
The following table shows the standard deviation on ratings by the Genres types for more than 100 rated movies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Genres (age>=24)** | **ST Dev** | **Genres (age < 24)** | **ST Dev** |
| Action|Animation|Children's|Sci-Fi|Thriller|War | 1.38 | Comedy|Crime|Horror | 1.59 |
| Adventure|Animation|Children's | 1.32 | Action|Animation|Children's|Sci-Fi|Thriller|War | 1.56 |
| Action|Horror | 1.28 | Adventure|Fantasy|Romance | 1.53 |
| Horror | 1.27 | Horror|Mystery|Thriller | 1.48 |
| Action|Comedy|Musical|Sci-Fi | 1.26 | Romance|Thriller | 1.47 |
| Comedy|Horror|Musical|Sci-Fi | 1.24 | Action|Romance|War | 1.44 |
| Adventure|Children's|Sci-Fi | 1.24 | Adventure|Sci-Fi | 1.40 |
| Horror|Thriller | 1.24 | Horror | 1.39 |
| Action|Adventure|Drama | 1.22 | Action|Adventure|Sci-Fi|War | 1.39 |
| Sci-Fi | 1.22 | Comedy|Horror|Thriller | 1.37 |

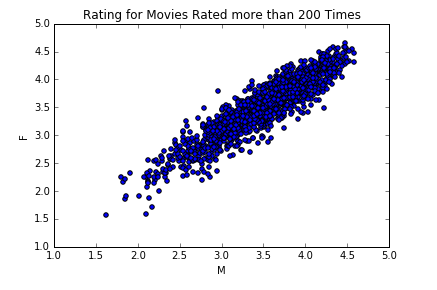
Based on this table, for the age less than 24 users, the Comedy/Crime/Horror movies have the highest difference within the ratings than other type movies. However, for the age larger than 24 users, the Action/Animation/Children's/Sci-Fi/Thriller/War movies have the highest difference within the ratings than other type movies. This observation supports the conjectures we made on this problem.

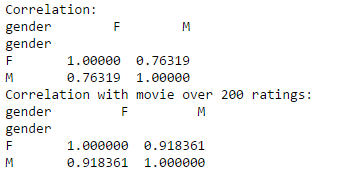
## Problem 3 – Analysis on Correlation

If the dataset gets broken down by gender, what is the relationship between two datasets and how well one dataset relates to the other dataset? For the first observation, we created a scatter plot for the mean rating from men versus women on each movie.

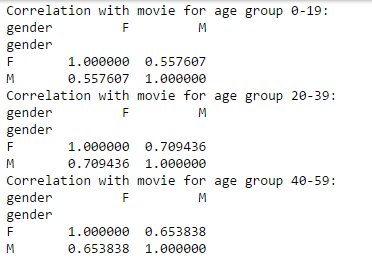


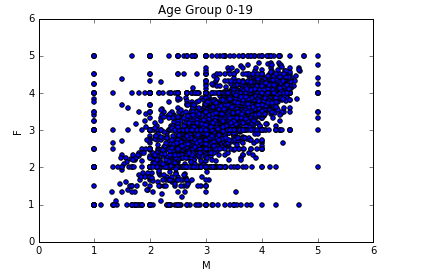
Based on the plot, there is positive linear correlation between men’s ratings and women’s ratings. Movies that received high rating from men were also rated hign by women on average. The dataset for this result is including all the movies, even the ones with few ratings. Too few ratings could results in skewed bias. So we created a new men-versus-women scatter plot only on the movies rated more than 200 times. Compared to the scatter plot for all movies, the linear trend is more obvious within the 2nd plot that covers only the movies rated more than 200 times. Also, the correlation coefficient of men versus women for all movies is 0.76 while 0.92 for the movies rated more than 200 times.

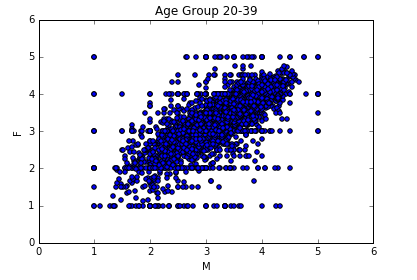


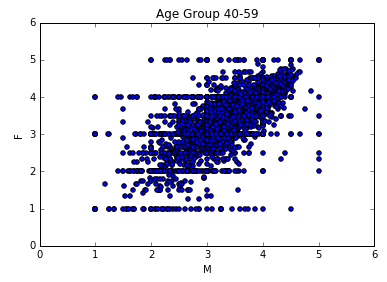


For different age group, would the degree of correlation between men’s rating and women’s rating be different? We separate the ages into 3 groups: Young (0-19), Middle (20-39) and old (40-59). For each age group, we created scatter plot for movie ratings by gender and calculate correlation coefficient. Based on the results, for different age group, there are different degree of correlation between ratings by one gender and the ratings by the other gender. For age group 20-39, the correlation is the highest.









### Conclusions for Problem 3

The movie ratings between genders have positive linear correlation. The higher the movie gets rated by one gender, the higher it gets rated by the other gender. If the movies with few ratings get removed from the dataset, the correlation will be stronger between men’s ratings and women’s rating. Among different age groups, the correlation coefficients between movie ratings on genders are quite different, which leads us to believe that the user age is a critical factor underlying the degree of correlation.

# Movie Recommendation Model

## Purpose:

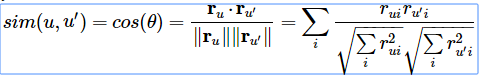
The goal of this recommendation model is to present a ranked list of movies given an input movie. The idea is that if one has enough user-to-movie data (ratings, age, occupation, etc...), then no other information is necessary to make decent recommendations, in contrast to regression and classification problems where one must explore various features in order to boost a model's predictive powers.

## Implementation:

We have the movie ratings by users (on a 1-5 scale). The main data file consists of user-id, movie-id, rating and timestamp which will be used as an input to our model. First, we created a user-by-movie matrix which maps user/movie ID’s to user/movie indices by building an “interaction” number in each cell. Every user has rated at least 20 movies which leads to a sparsity of 6.3%. This means that 6.3% of the user-movie ratings have a value. The missing ratings are attached a zero value. Next, we split our data into training and test sets by removing 10 ratings per user from the training set and placing them in the test set.

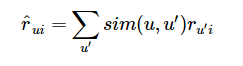
For our purposes we used Collaborative Filtering which can be split into two classes: **user-based** and **item-based collaborative filtering**. We built a similarity matrix for each for both user-based and item-based collaborative filtering.

For user-based collaborative filtering, the user-similarity matrix will consist of some distance metric that measures the similarity between any two pairs of users. Likewise, the item-similarity matrix will measure the similarity between any two pairs of items. A common distance metric is cosine similarity. The metric can be thought of geometrically if one treats a given user's (item's) row (column) of the ratings matrix as a vector. For user-based collaborative filtering, two users' similarity is measured as the cosine of the angle between the two users' vectors.  For users u and u′, the cosine similarity is,

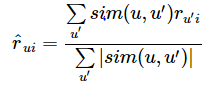


The results from this cosine similarity will range from 0 to 1 as there are no negative ratings. Using this similarity matrix we can now predict the ratings that were not included with the data. We will compare the results of our predictions with the test data to validate the quality of our recommender model.

For user-based collaborative filtering, we predict that a user's u's rating for item *i* is given by the weighted sum of all other users' ratings for item i where the weighting is the cosine similarity between the each user and the input user u.



Next we will normalize by the number of ru′i ratings:



**Since we are dealing with a domain where we have intuition about movies, we can look at our item similarity matrix and check if similar items make sense as part of or prediction model.**

## Results\Validation

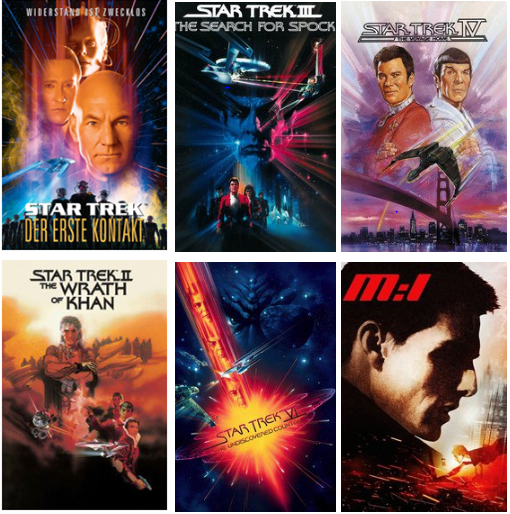
We used the website themoviedb.org to pull information about a movie such as Movie ID, Movie posters and Movie URL’s to tie it with the results of our model.

We built a dictionary for mapping the movie-indices from our similarity matrix to the urls of the movies. Next, we created a helper function to return the top-k most similar movies given some input movie. With this function, the first movie returned will be the input movie (as it is the most similar to itself).

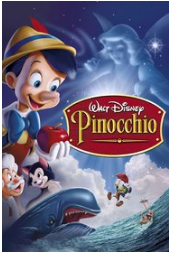
**Case 1**: **Model Input** : Star Trek: First Contact (1996) *Genre*: Action|Adventure|Sci-Fi



**Model Output**: Based on the input movie **Star Trek: First Contact** we received the following recommendations from our model that are most similar to our input. In this case we can see that it recommended other movies Sci-fi movies that are most similar to our Input, notably other Star Trek and Star Wars movies which are some of the well like movies within the same genre and rated highly by users.

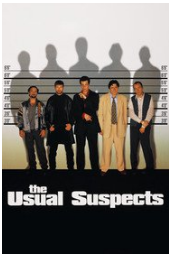
**Case 2: Model Input**: Pinocchio (1940) *Genre*: Animation|Family|Fantasy



**Model Output** : Based on the input movie **Pinocchio** we received the following recommendations from our model that are most similar to our input. In this case we can see that it recommended movies other Animation, Fantasy movies that are most similar to our input. Notably movies like Cinderella, Snow White, Alice in Wonderland & Mary Poppins which some of the other well liked movies in the same genre.



**Case 3 : Model Input**: The Usual Suspects (1995) *Genre*: Crime, Drama, Mystery



**Model Output:** Based on the input movie **The Usual Suspects** we received the following recommendations from our model that are most similar to our input. In this case we can see that it recommended other movies in the Crime, Drama, Mystery genre that are most similar to our input, notably movies like Seven , Pulp Fiction , Reservoir Dogs.

# Business Question

Within this case study, we did data exploration and analysis on the MovieLens dataset and created movie recommendation engine to predict users’ preference. All these results will be shown on dashboard for the decision support for the decision maker. For example, the project manager could use our analysis results to understand more on the taste of the customers in order to make better decision on which kind of movie should invest in. Also, the movie recommendation engine could help company personalize movie recommendations for customers to maintain their loyalty.