MovieLens Datasets Study

-Exploration of Recommender Systems and Data Visualization

Zhuangye Chen(Michael), Alexander Trent Dewey, Xingxing Tong(David)

MovieLens: MovieLens datasets was first released in 1998. People's preferences for movies were collected in this dataset. Nowadays, this datasets are widely used in education, research, and industry. They are downloaded hundreds of thousands of times each year. This popularity is, to a certain degree, a reflation of the incredible rate of growth of personalization and recommendation research, in with datasets such as these have substantial value in exploring and validating ideas. [1]

100K datasets: Our project focus on the MovieLens 100K dataset which contains 100,000 ratings (1-5) from 943 users on 1682 movies. Three crucial subsets we focused are "u.data", "u.genre" and "u.user". These three datasets contains information of movie rating scores, user's demographic information and movie's genre. Emphasis is put on digging the raw datasets in the first week. We hope find some interesting stories (Details will be introduced in the next section) about the data via visualization. Along the way, we get familiar with movies data and the recommendation system which is quite helpful for the further exploration in this area couple of days later.

Different behavior between movie lovers and enthusiasts: For each user we counted the number of movies they had rated. It turns out that median of user rated movie number is 65.

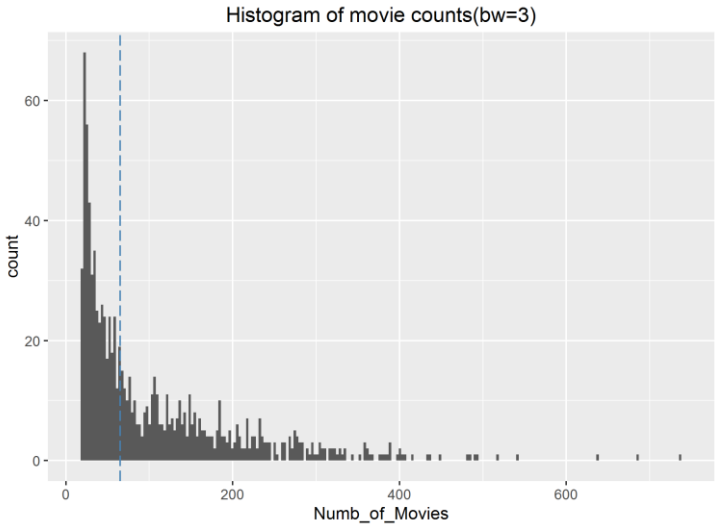

Figure1. Histogram of the number of movies users rated (Blue dashed line above is the median)

Based on the median we divided our users into two different groups. 1st group (movie lovers) rate number less or equal to 65 & 2nd group (movie enthusiasts) rate larger than 65. Then we combined the demographic information from dataset "u.user" and explore the different behaviors in light of the demographic information. Two graphs below are examples.

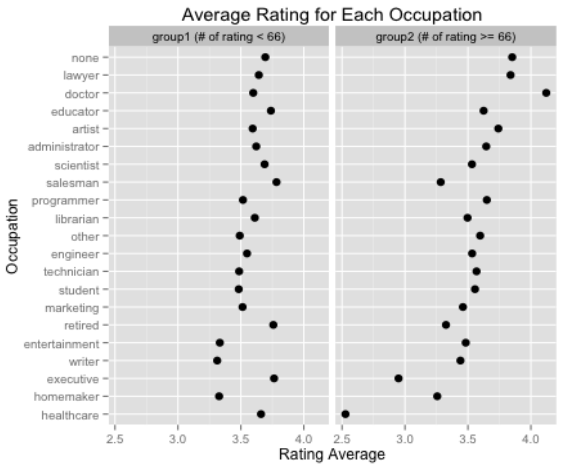

Figure2. Distribution of average rating scores for two groups given each occupation

Figure3. Distribution of occupation and gender for two groups

Genre information: This time we focus on each unique movie instead of user.

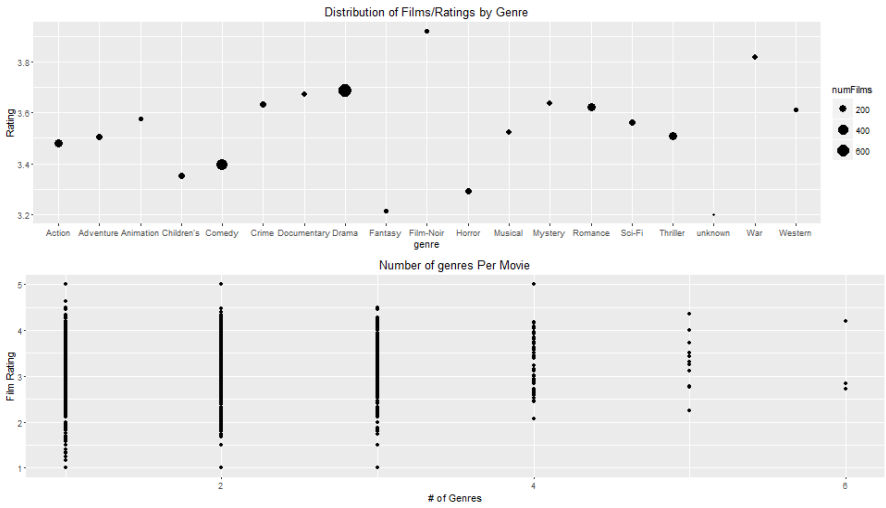

Figure4. Genre graphical summary

Key question: Obviously it's impossible to know the "true" rating of any movie, but without knowing the actual difference in quality between two movies, can we explain the primary force which drives this rating disparity between silly and serious? Are heavy users rating serious movies more heavily and at a higher rating? Is there a broader base of users watching comedies? Can this be explained by the little demographic information we have (age, gender, and occupation)? Finally, the large difference in ratings between comedy and drama exists in part because a small proportion of films (89/1682) are in both categories while more (1052/1682) are in one or the other. How much do genres overlap, and to what extent do they overlap (and to what extent does correlation help to influence ratings)?

GitHub: Our repository name is “P2_Recom_Sys_MovieLens”. Different “.md” files are created to record project schedule and off-line group discussion. Mark Holder once said “Your closest collaborator is you six months ago, but you don’t reply to emails.” So we try to record our group discussions and project schedules as clear as possible for future.

Reference: [1] F. Maxwell Harper and Joseph A. Konstan. 2015. TheMovieLens datasets: History and context. ACM Trans. Interact. Intell. Syst. 5, 4, Article 19 (December 2015), 19 pages.