Visualization of movieLen rating

Visualization Project Proposal

Zhao Chang u0931132@utah.edu u0931132
Junwei Shi u0943296@utah.edu u0943296
Tengda Shi tengda.shi@utah.edu u1015168

Project Repository: `https://github.com/VaultBOY/dataviscourse-pr-visualization`

1 Background and Motivation

The objective of our project is to perform an interactive visualization analysis on the relationship between different kinds of movies and people who give ratings on the movies. Our project deals with a really interesting problem. Note that thousands of movies are released over the world each year. We can easily find the corresponding information in some online websites (eg. IMDB[1]). However, in fact, we are only interested in a small part of them. We have no time to scan all the introductions to these movies. Thus, we hope we can perform an interactive visualization analysis and tell people what movies they may like to see based on people's attribute information. These attribute information regrading different people can also be added into existing recommendation systems and improve the accuracy of recommendation.

2 Project Objectives

The primary questions we are trying to answer with our visualization are:

- Tell people which movie they may like to see. If you are plan to watch a movie but you are not sure if you will like this movie, then you can see people's comments who are in your age through our visualization and decide whether to go to see the movie.
- Check people in different age what their favorite movie is and the comments of one certain movie. The basic function of our visualization is to search the information what you want to check about the movie.

[1] www.imdb.com/

- Help the movie makers to investigate who like their movies best. Our visualization will also benefit the movie makers and help them to make the movie that more people like to watch.

What we would like to learn is if any trends or patterns exist in our data set. In particular, the things we would like to accomplish include:

- Collect all the comments of each movie, including the information of people for each comment.
- Different comments of different movies for each person. People can search the information according to the audience's gender, oocupation, postal-zone etc..
- Movies can be searched by categor,theme,year and so on.

3 Data and Data Processing

We use MovieLens 1M Data Set[2]. This data set contains 1,000,209 ratings applied to 3,952 movies by 6,040 users of the online movie recommender service MovieLens. All ratings are contained in the file "ratings.dat" and are in the following format: UserID::MovieID::Rating::Timestamp. Ratings are made on a 5-star scale. Each user has at least 20 ratings. 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. By performing visualization analysis on the data set, we can answer some interesting questions, such as "what kinds of movies are liked by people with some specific attributes".

We pre-process the data set by clustering the data according to the different kinds of movies and different attributes of people. After generating such aggregation information, we use the transformed data set to perform interactive visualization analysis on it.

4 Visualization Design

4.1 Our Design

In our design, we divide the functions into four parts:

1) Present the rating info given a specific user.

2) Present the rating info given a specific movie

[2] http://datahub.io/dataset/movielens

3) Process the statistic given a specific user group using filter

4) Process the statistic given a specific movie group using filter

5) Find the dependencies between users and movies

4.2 Must-Have Features

The Movie must-have feature: name, year, genre

The user must-have feature: age, gender, occupation, location

4.3 Optional Features

All features in the data are clearly presented, hence they're all necessary features.

4.4 Part 1 Design

We can search the rating information given a specific user. Also we could filter out some type of movies we want, such as released year and genres.(figure 1) After that, we can present the information grouped by genre, or just a scatter plot along the year x-axis. (see figure 2 and 3) Also we can rank the movie as in figure 11.

4.5 Part 2 Design

We can search the rating information given a specific movie. We could present some interesting statistics, such as number of ratings or scores, categorized by ages. (figure 5 and 6), or categorized by some interesting features such as locations. Please note that the figure 6 is a distribution figure, and we use different colors to denote the distribution percentage.

4.6 Part 3 Design

First we need to filter out the users we want, by selecting some features as in figure 7. Then we show the statistics as figure 5 and 6.

4.7 Part 4 Design

We can filter out the movies by selecting features as figure 4. We can show the statistics as figure 9 and figure 10.

4.8 Part 5 Design

After some preprocessing of the data, such as clustering or collabrative filtering, We can show the dependencies of different groups of users and the movies. The groups are based on the feature we have selected.

5 Project Schedule

Week 1

1) Clean and pre-process the original data set

2) Check the feasibility of our project based on the operations above.

Week 2

1) Build the basic visual structure/layout using JavaScript.

2) Import the transformed data into MySQL.

Week 3

1) Finish the framework of our project based on the data format.

2) Finish the draft of our Process Book.

Week 4

1) Polish the style of modules in our project.

2) Begin discussing Project Screen-Cast.

Week 5

1) Continue polishing the style of modules in our project.

2) Finish the public website for our project.

Week 6

1) Finish our Project Screen-Cast.

2) Describe our project in detail and finish our Process Book.

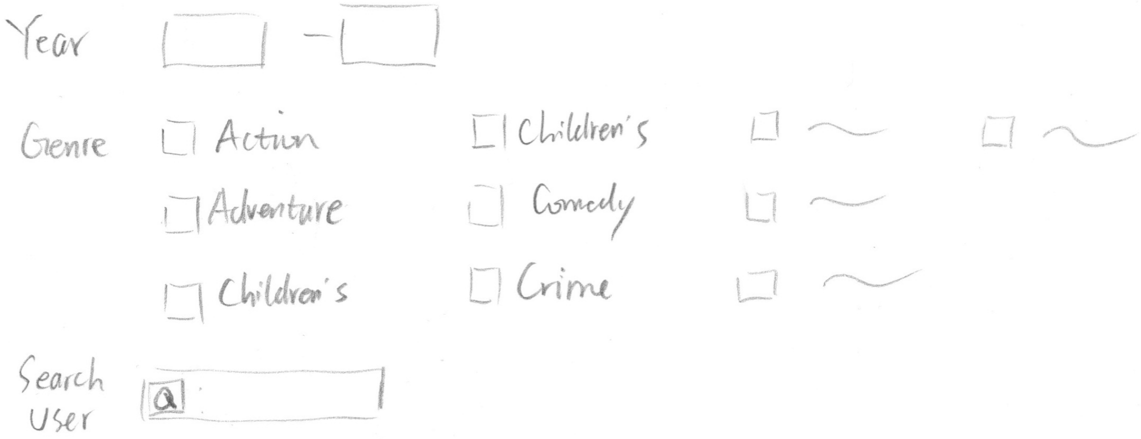


Figure 1: Search for user and filter the movie information

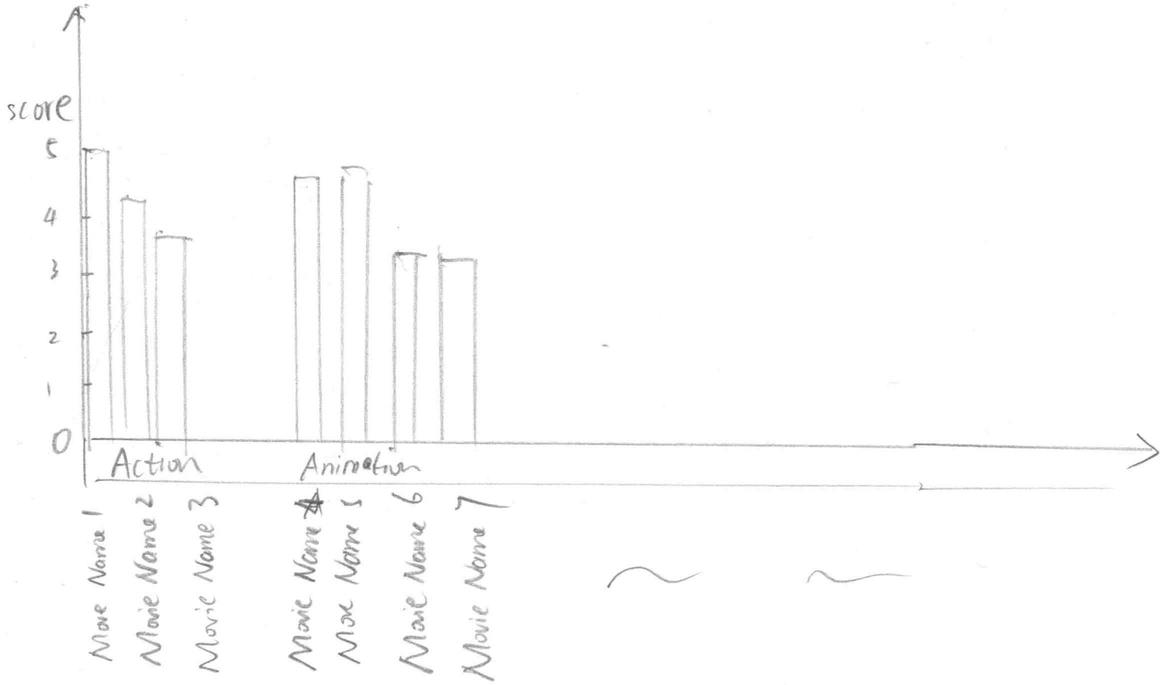


Figure 2: Histogram of scores grouped by genre

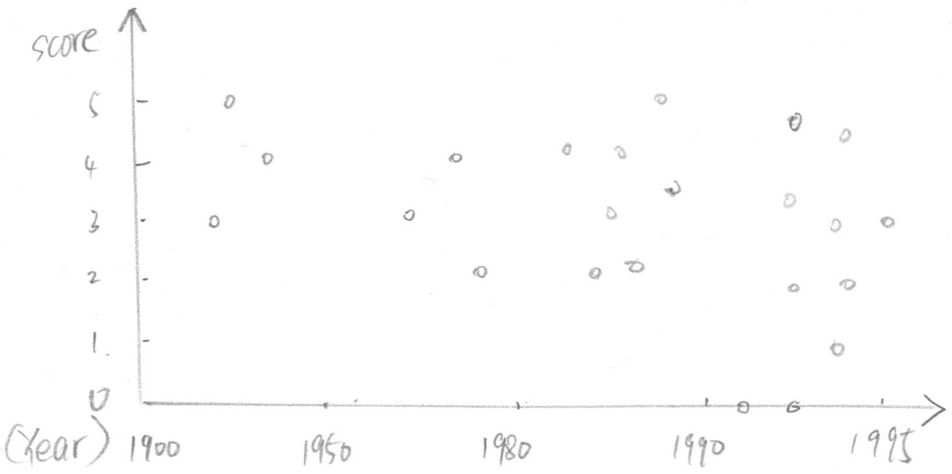
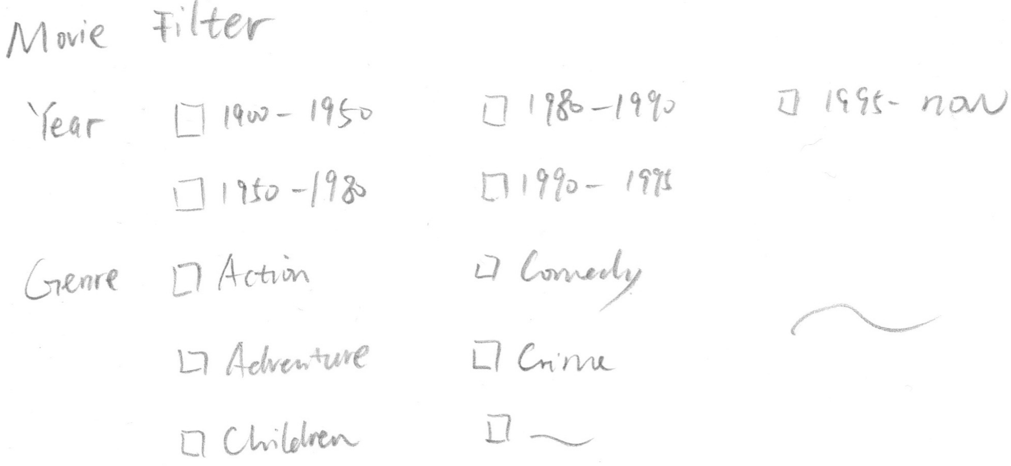


Figure 3: Scatter plot of scores



Movie Filter

Year □ 1900 - 1950 □ 1980 - 1990 □ 1995 - now

□ 1950 - 1980 □ 1990 - 1995

Genre □ Action □ Comedy ~

□ Adventure □ Crime

□ Children □ ~

Figure 4: Filter of movie info

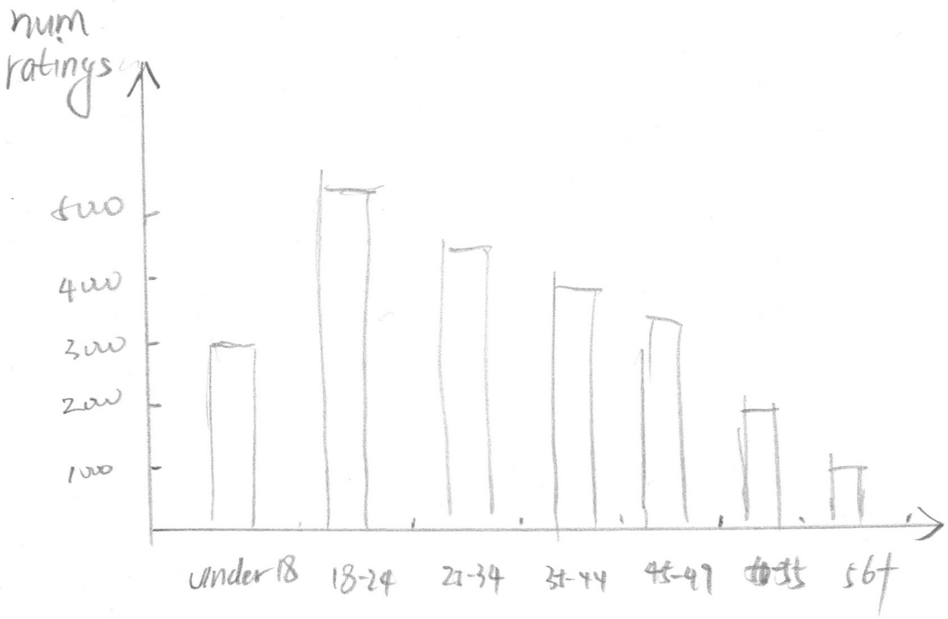


Figure 5: Num of ratings grouped by ages

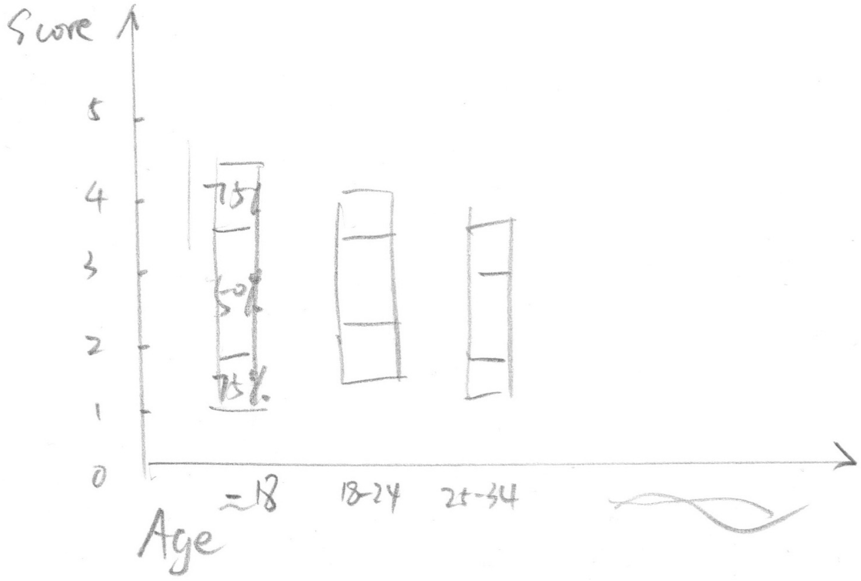


Figure 6: Distributions of scores grouped by ages

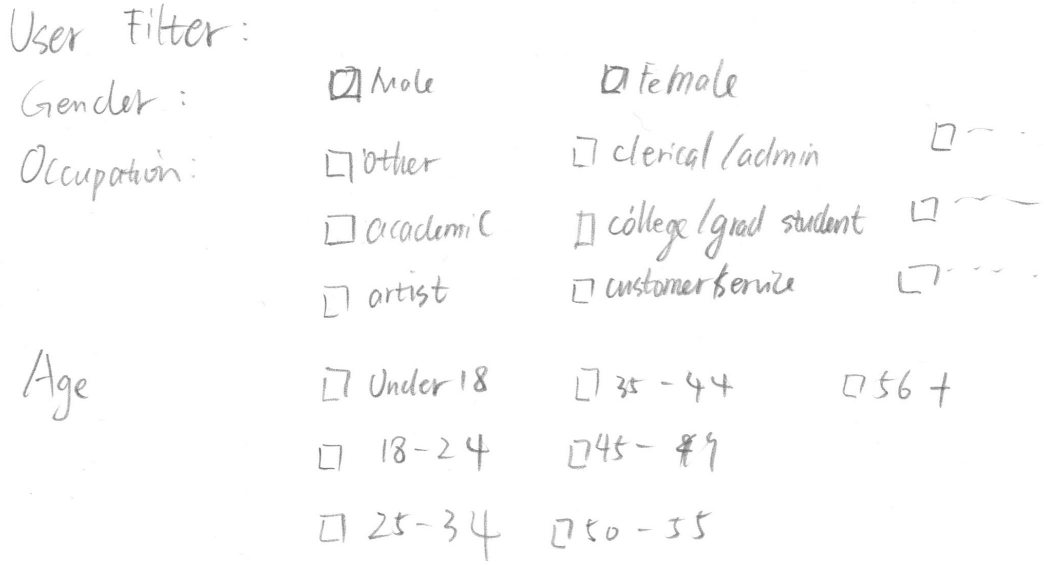


Figure 7: Filter of user info

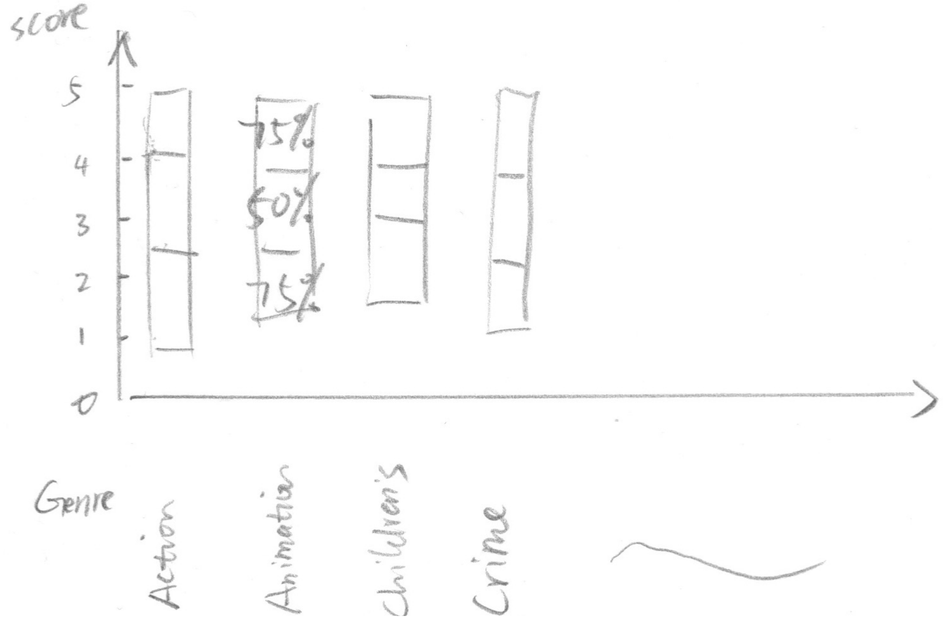


Figure 8: Distributions of scores grouped by genres

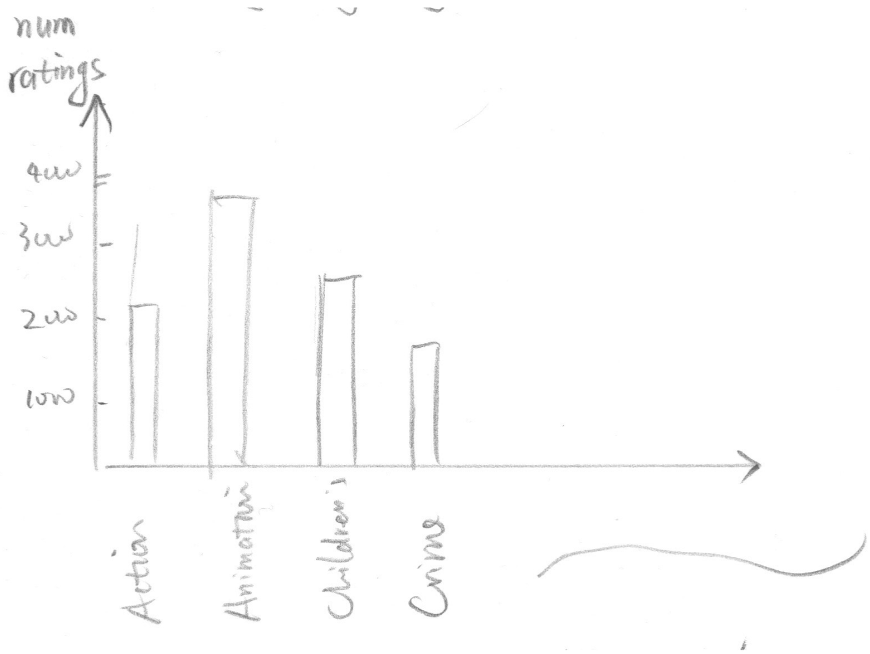


Figure 9: Num of ratings grouped by genres

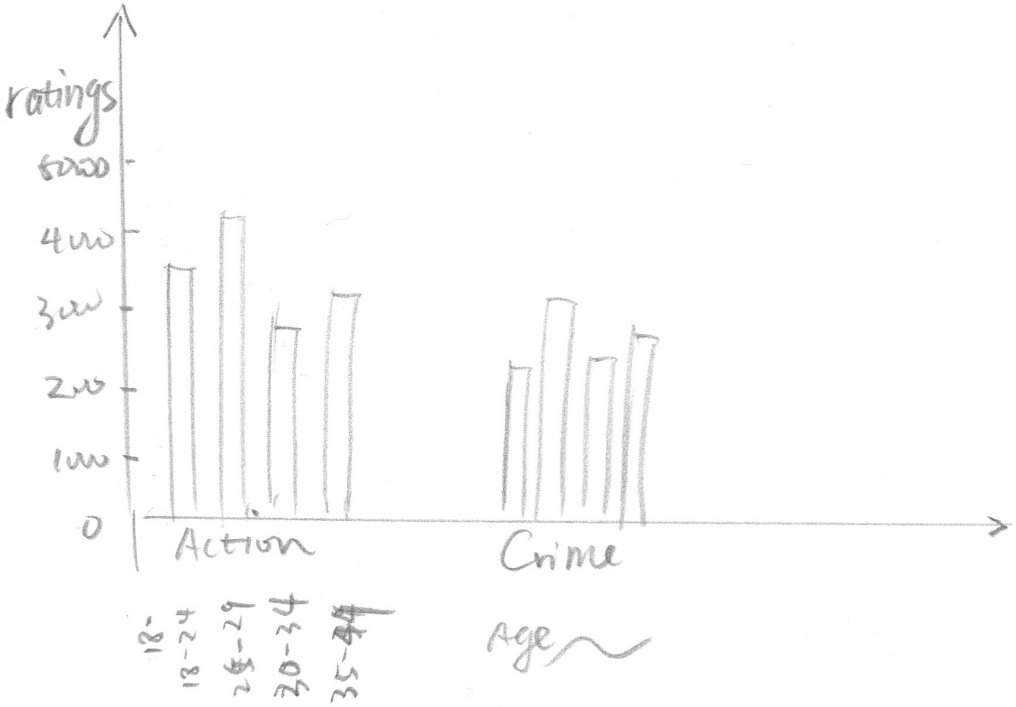


Figure 10: Num of ratings grouped by genres and subdivied by ages

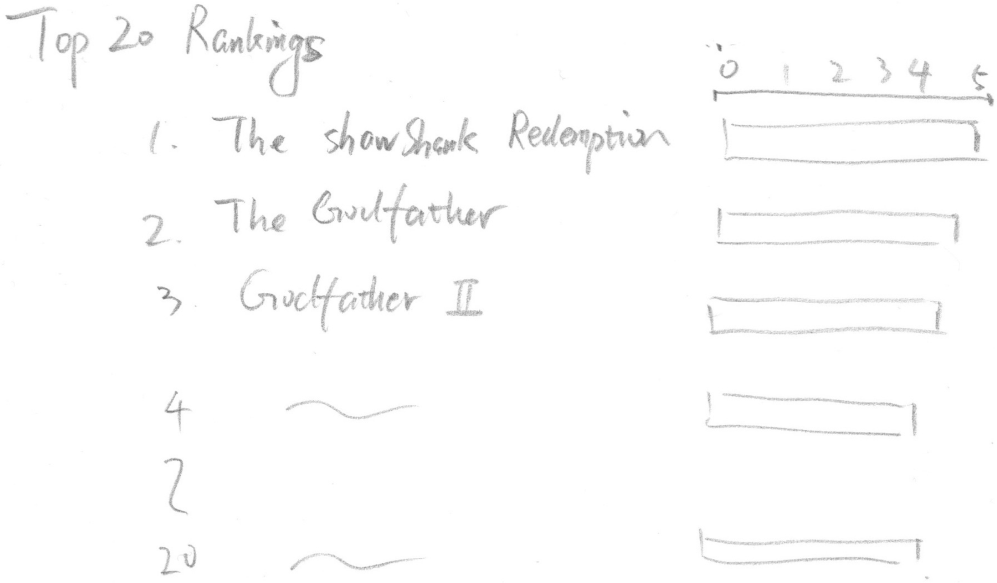


Figure 11: Rankings of movie scores

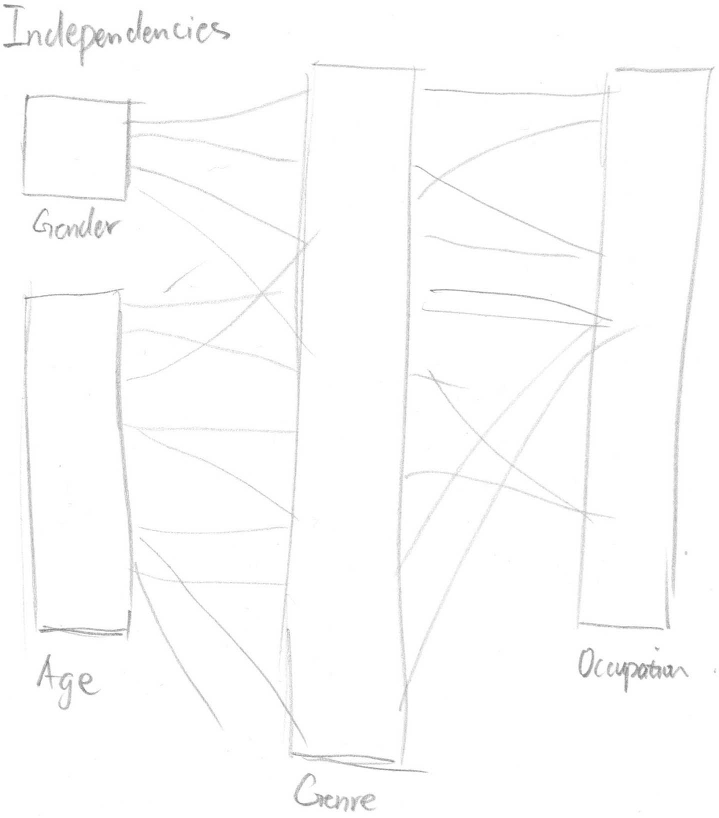


Figure 12: Dependencies of users and movies