**MovieLens Datasets Study**

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

Summary

…details needed…

Part 1 Introduction

**) Background of the recommender systems**

Recommender systems apply statistical and knowledge discovery techniques to the problem of making product recommendations.[2] Good recommender systems can help customers easily find the products they want, and create a value-added relationship between buyers and sellers.

There are generally two types of recommender systems: Content-based filtering & Collaborative filtering (CF). Two types of systems share one “consistency” assumption which is those who agreed in the past tend to agree again in the further. Content-based filtering is based on the attributes of products. Consumers preference’s is generated by the attributes of products their have bought before. However, on the other hand, collaborative filtering is based on the performance of similar users. In other words, defining groups of similar users is the key process in the type of recommender systems.

We focus on the application of CF on movies recommendations. So much emphasis will be put on this particular area in the rest of introduction part. Based on CF we can predict the rating of an unrated movies or create a top-N list of unrated movies for a specific user. Memory-based CF focus on finding K nearest neighbors for the user in the user-item matrix. Then recommendation will be given based on the items liked by the K nearest neighbors. Model-based CF tries to build a model from the rating data (clustering, latent semantic structure, etc.) and then us this model to predict missing ratings.

**) Background of the MovieLens datasets**

MovieLens datasets was first released in 1998. People’s preferences for movies were collected in this dataset. Nowadays, these 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]

Part 2 General Process and Key Question:

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 sub 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 stage. Interesting stories are found via visualization and basic data summary. Along the way, we decided to focus on dataset “u.user” and building an “affinity” matrix to describe the similarity of users which is also the first step of CF recommender system. The closeness is defined by a distance mainly based four demographic variables in the sub dataset. In the fourth part, we use this “affinity” matrix to make rating prediction of users.

…details needed…

Part 3 Initial Exploration:

**Different behavior between movie lovers and enthusiasts**

For each user we counted the number of movies they rated and plot the result.

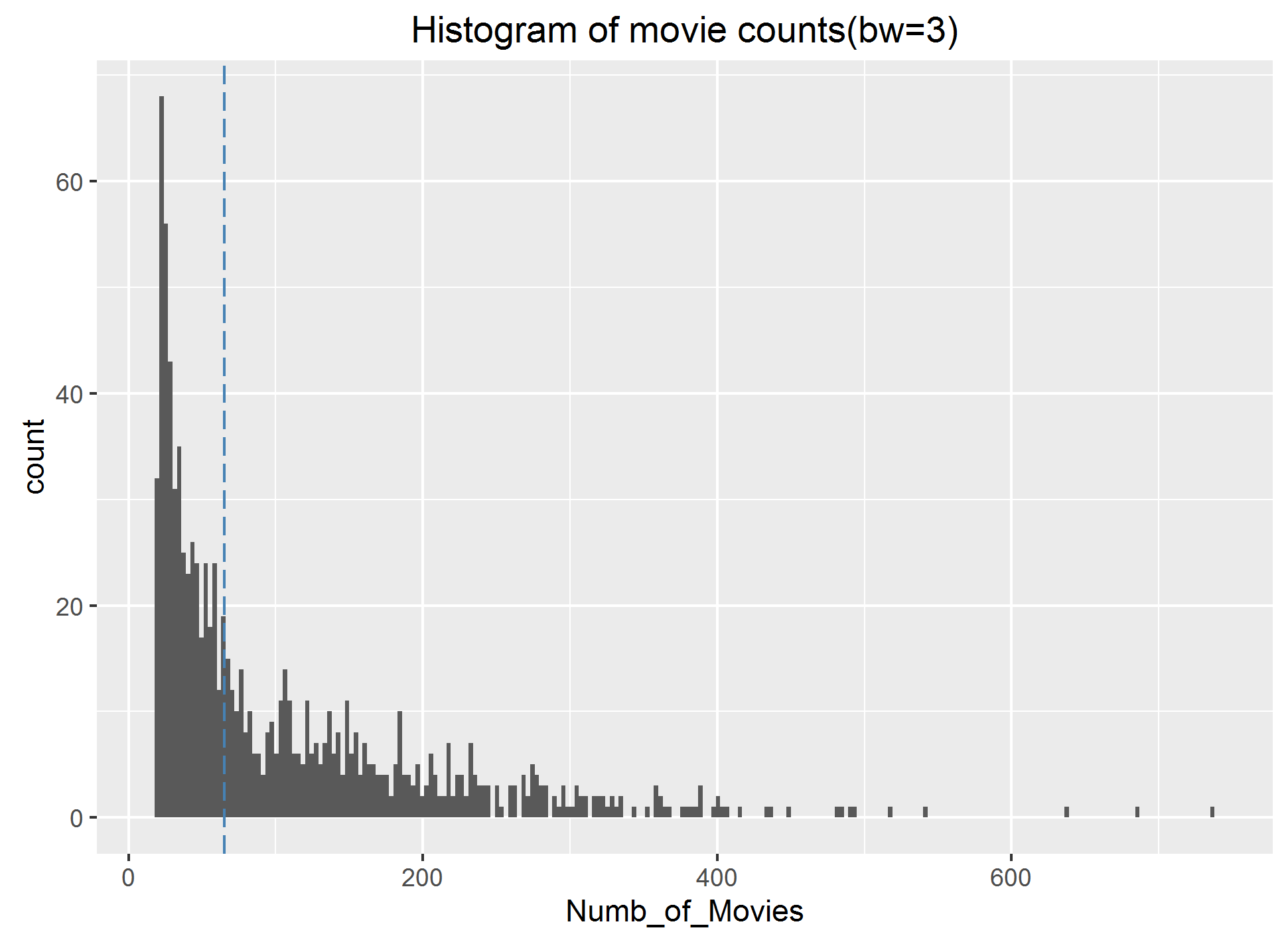


Figure1. Histogram of the number of movies users rated (Blue dashed line indicates median)

The number of movies users rated ranges from 20 to 737. The median is 65. So we divided our user group into two groups based on this median. Users who rated less or equal to 65 movies are named movie lovers and those who rated more than 65 movies are named movie enthusiasts. Then we combined the demographic information in “u.user” dataset with the division which gives us two graphs in the following.

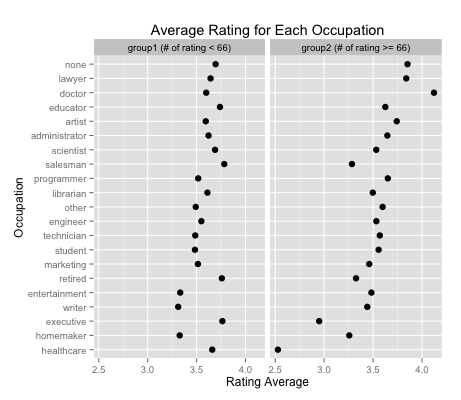


Figure2. Distribution of average rating scores for two groups given each occupation

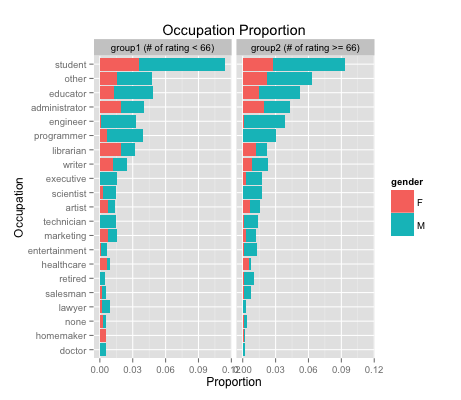


Figure3. Distribution of occupation and gender for two groups

…details needed…

**Genre**

A similar method is applied to “u.genre”. Instead of user, here we focus on the movie. The figure below shows the average rating for each genre of film. The figure below shows the number of genres associated with each film film. The exact distribution is (833, 569, 215, 51, 11, 3) for i = 1, 2,..., 6 labels. There is about a 50% chance that a given film will have multiple genres.

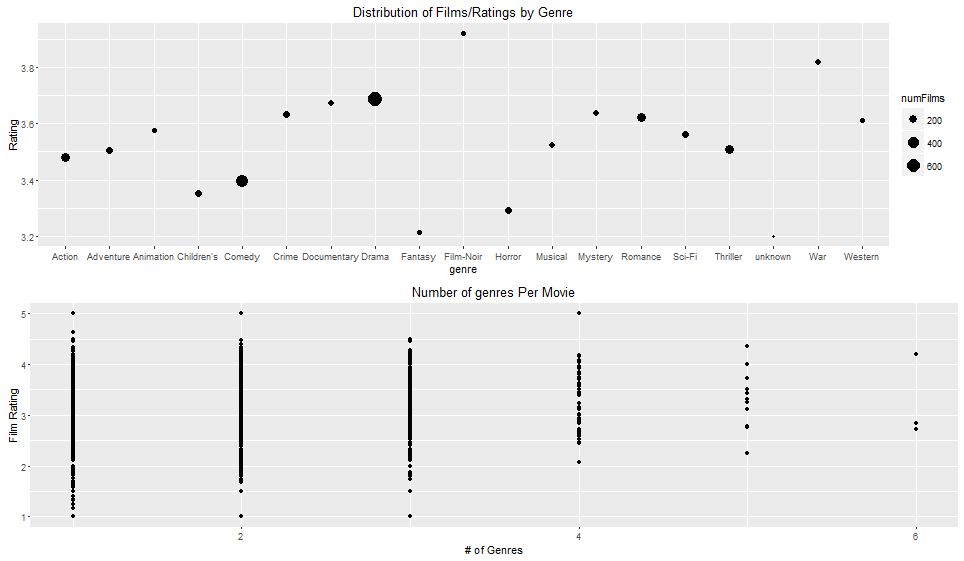


Figure4. Graphical summary for genre

In the figure above, note that comedies and dramas dominate the genre (reflected by the size of the points). Comedies are notably rated around 0.3 stars lower than dramas. Given that most of the variation in ratings by genre is primarily a result of having very few films, this sticks out as the most obvious trend--"fun" genres like action, children's, horror, and family are very low, while "serious" genres like Drama, Documentary, Crime, Mystery, Romance, and War are very high rated.

Part 4 Affinity Matrix

…details needed…

Part 5 Rating Prediction

…details needed…

**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.

[2] (Sarwar et al., 2000).