# Milestone report: Movie recommendation system

1. An introduction to the problem:

Analyzing the movielens dataset and building a recommendation system based on the rating given to the movies by the users in the dataset.

1. A deeper dive into the data set:

The dataset used for this exercise is the movielens dataset, which has around 6000 movies and they are rated by users belonging to different age groups and occupations.

The most important fields that will be primarily used in the recommender system are:

* + 1. Userids: Every user in the system is given a unique identifier.
    2. Movieids: Every movie in the system also has a unique identifier
    3. Ratings: All the ratings given by users to movies are recorded on a scale of 1 to 5 . Each user has at least 20 ratings.

The data is in the form of .dat files with 3 files for users, movies and ratings.

The .dat files had to be loaded in R and then transformed from positional strings to tab separated form and then further loaded into 3 data frames. These data frames were merged into one big data frame, on which all the preliminary exploratory analysis was conducted to get a better feel of the data.

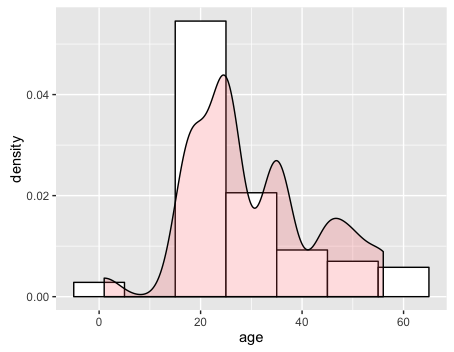
All the metadata fields in the 3 files are represented by ids for e.g. user age group, their occupation is represented as IDs which don’t make much sense on its own. So, reference data had to be extracted from the README file accompanying the dataset and transformed into csv which was then loaded in R and merged with the master data frame to get all the information in one place. This was a minor hurdle, since the exploratory graphs were meaningless without this exercise.

The data has limitations in terms of the genres assigned to the movies. Each movie falls under multiple genres which makes it difficult to clearly classify movies under a single or a common few genres. So, same movie can fall under 3 different genres, but it may not be like the movie that has one of its multiple genres in common to the first movie. This might introduce noise in the recommendations done based on just genre similarities.

This limitation and some exploratory analysis helped shape the approach to run recommendations based purely on ratings provided by the users.

1. Preliminary exploratory analysis of the dataset done in R:
   * 1. Plotting the trends for age groups of the users in the data sample:

Here majority of the users that have rated most movies fall in the 18-25 group category which has people from ages roughly 18 to 34 per the age group boundary definitions provided in the reference documents accompanying the data set. There is also a spike for age group around 35 and a smaller spike for age groups 50.



Reference information on age:

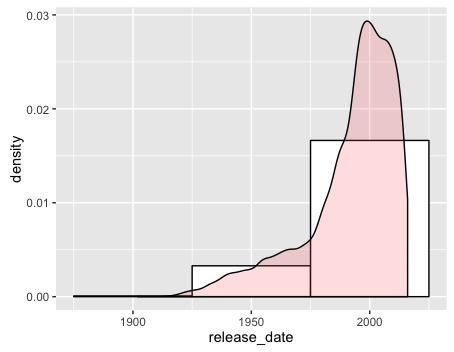
|  |  |
| --- | --- |
| Age\_ID | Age\_group\_desc |
| 1 | Under18 |
| 18 | 18-24 |
| 25 | 25-34 |
| 35 | 35-44 |
| 45 | 45-49 |
| 50 | 50-55 |
| 56 | 56+ |

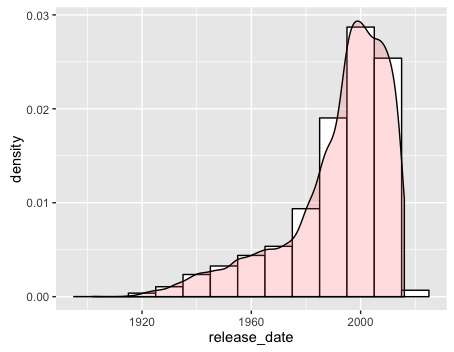
* + 1. Plotting the trends for movies based on their release dates:

It is clear from the high-density spike that majority of the movies are from 1990s.

Although the initial low density tail also suggests that the dataset does have a small number of movies from the past starting early 1900s.

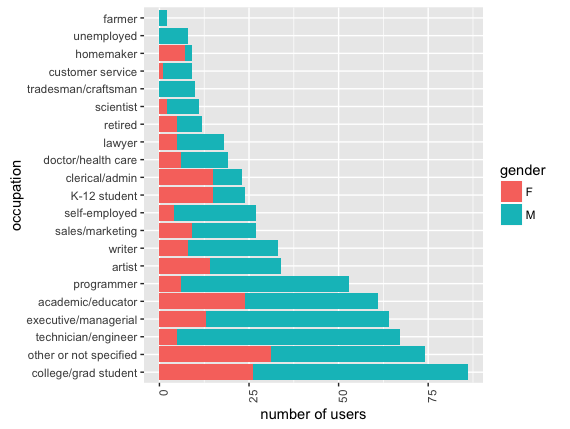
The second date plot is basically like the first date plot, but the binwidth is adjusted to get a closer look at the release date trend in our sample.





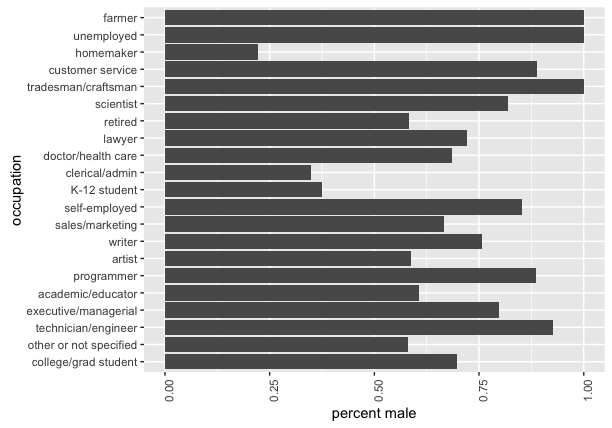
* + 1. Plotting the distribution of users per occupation and genders:

This is a very basic plot that delves into the distribution of the occupations and genders amongst those occupations of the users. Males in general are far more as compared to females in all occupations except homemaker, clerical/admin and K-12 students, unsurprisingly. This gender bias is plotted in the next graph.



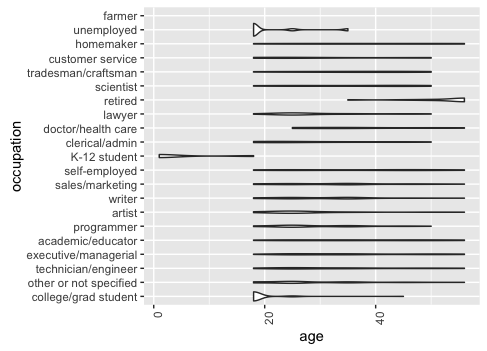
* + 1. Plotting the gender bias in each profession:

In this plot males are shown a percentage of the total users in each occupation to get a clearer picture of the gender distribution for each occupation in the sample.



* + 1. Plotting the age of users with respect to their occupation:

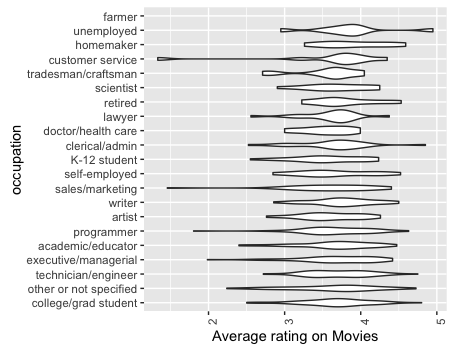
The following plot shows the exactly how old the make-up of an occupation is in our sample data. A few insights are clear, students have a very low average age as compared to the retired, same goes for the unemployed category. Almost all professions have a lower average age as compared to doctor/healthcare occupational category.



* + 1. Plotting the ratings of users with respect to their occupation:

The following plot explains how users in different occupations tend to rank different movies.

Users from different occupations do not tend to rank movies evenly. The customer service and sales marketing professionals rank movies extremely low as compared to the ratings of other professions. The unemployed seem to be doing the exact opposite, ranking movies higher. Healthcare professionals seem to have the least variance in their average ratings. It stays steady between 3 and 4.



1. Final approach:

As mentioned before, based on the data present for users and the ratings given by them a recommender system based on user ratings will be the best approach to move forward. Since the dataset does not have enough metadata around movies, content based filtering approach cannot be used with this dataset, since the inputs to the recommendation system are not enough.

But the dataset has enough data around users and their ratings as evident from the exploratory analysis.

The final approach will be to use a certain portion of the dataset as a training data set and then using the recommender lab package available in R the recommendation code will be tested on a test data.

For the final submission cleaned up versions of the same movielens data will be used, that are available as source files.