Recommender System: MovieLens 100K Dataset

Wei Zhong
University of Illinois at Urbana-Champaign

Matrix Factorization

The resulting dot product $q_i^T p_u$ captures the interaction between user u and item i, the users' overall interest in the item characteristics. Thus the user u's rating of item i can be denoted as r_{ui}, given by,

$$r_{ui} = q_i^T p_u$$

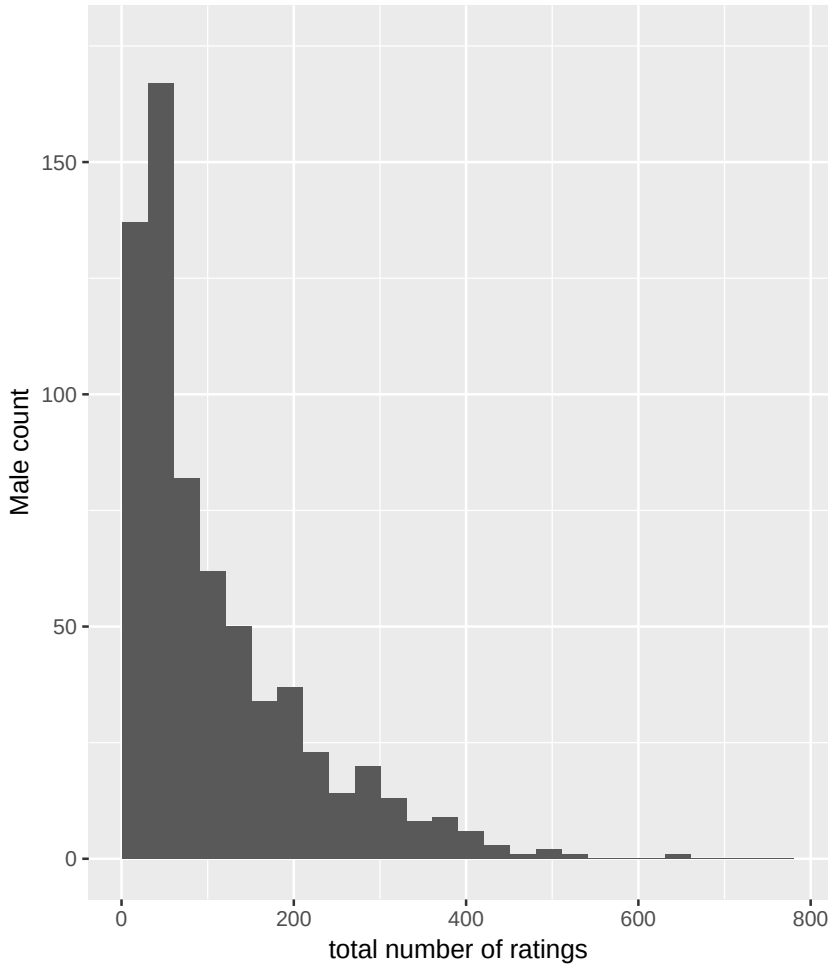
To learn the factor vectors, p_u, q_i, the system minimizes the regularized squared error on the set of known ratings:

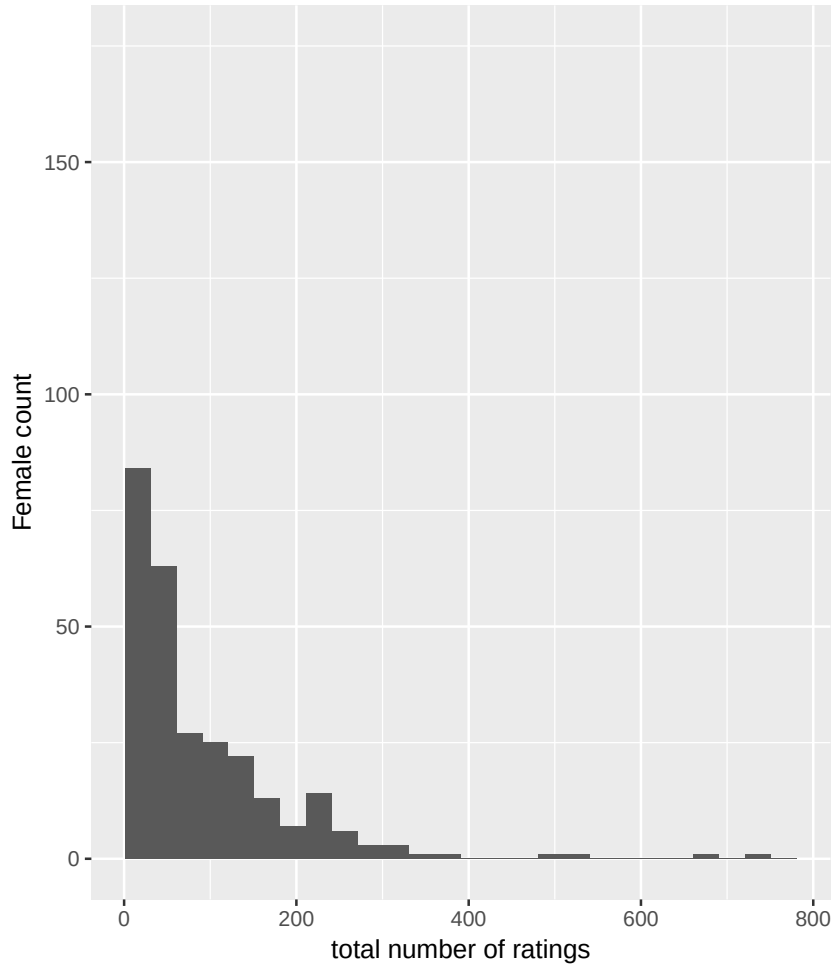
$$\min_{q^*, p^*} \sum_{(u,i) \in \Omega} (r_{ui} - q_i^T p_u)^2 + \lambda(||q_i||^2 + ||p_u||^2)$$

where Ω is the observed set of the (u, i) pairs.

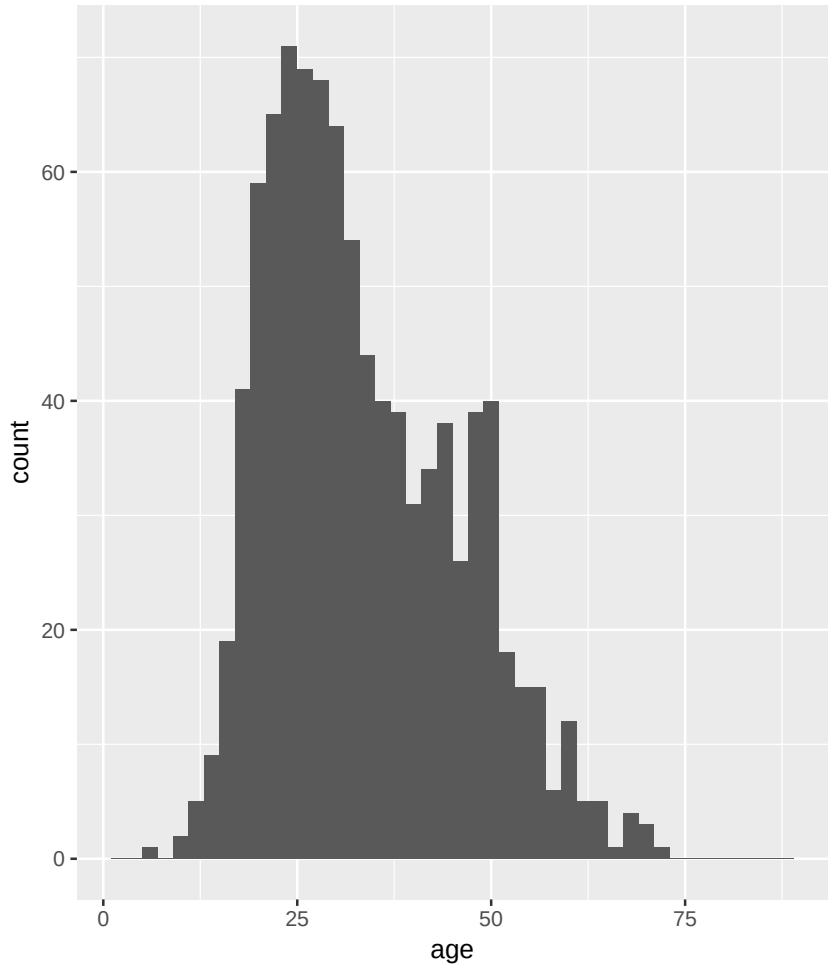
Missing Pattern of the Dataset

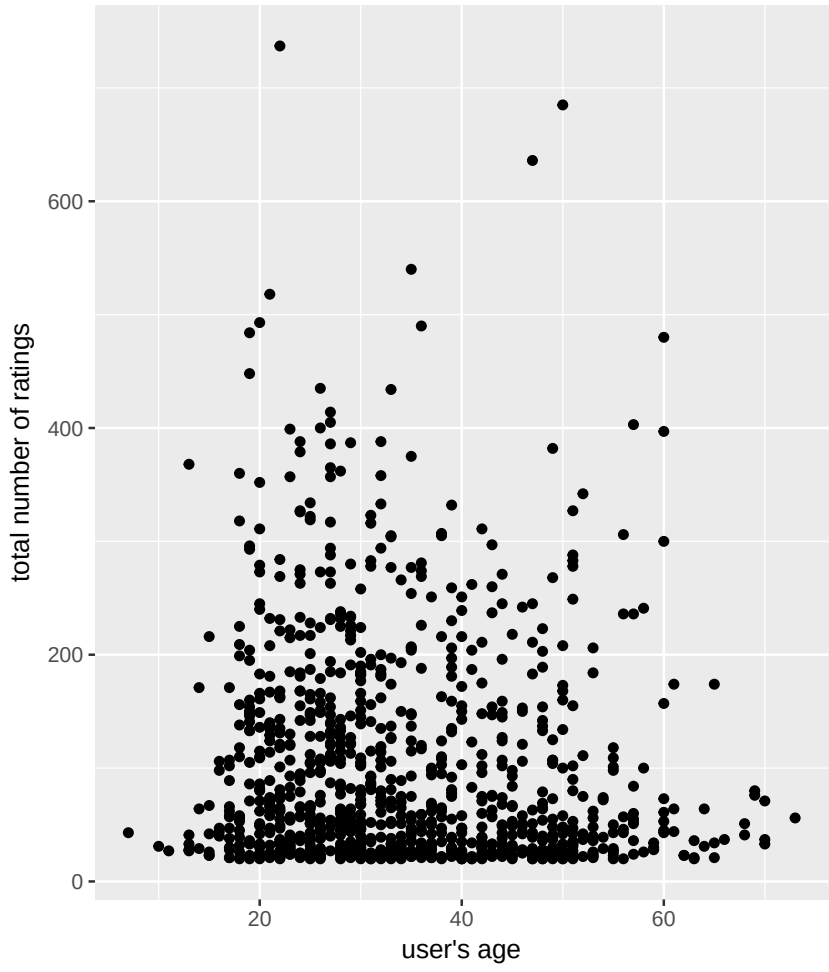
Gender Pattern



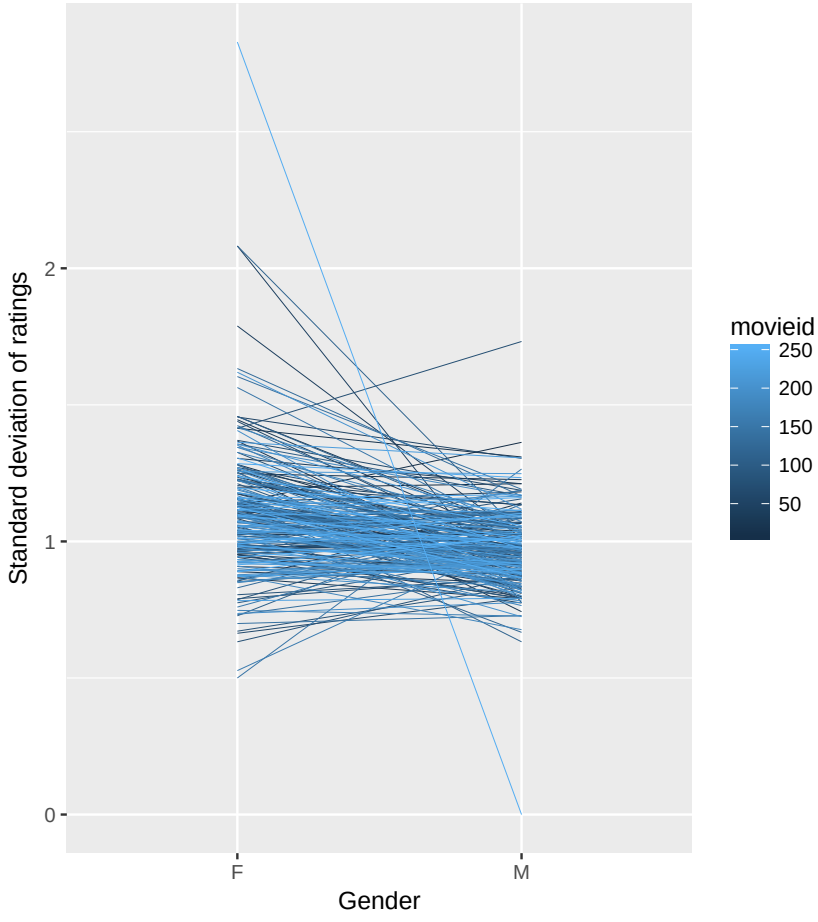


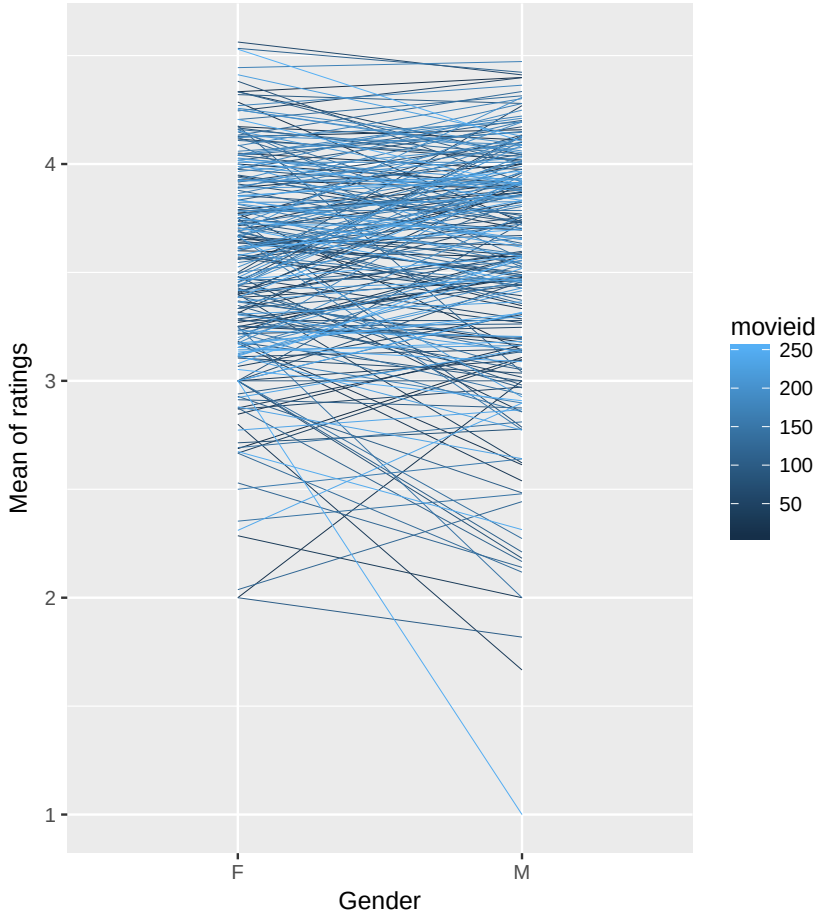
Age Pattern


count
60
40
20
0
0
25
50
75
age

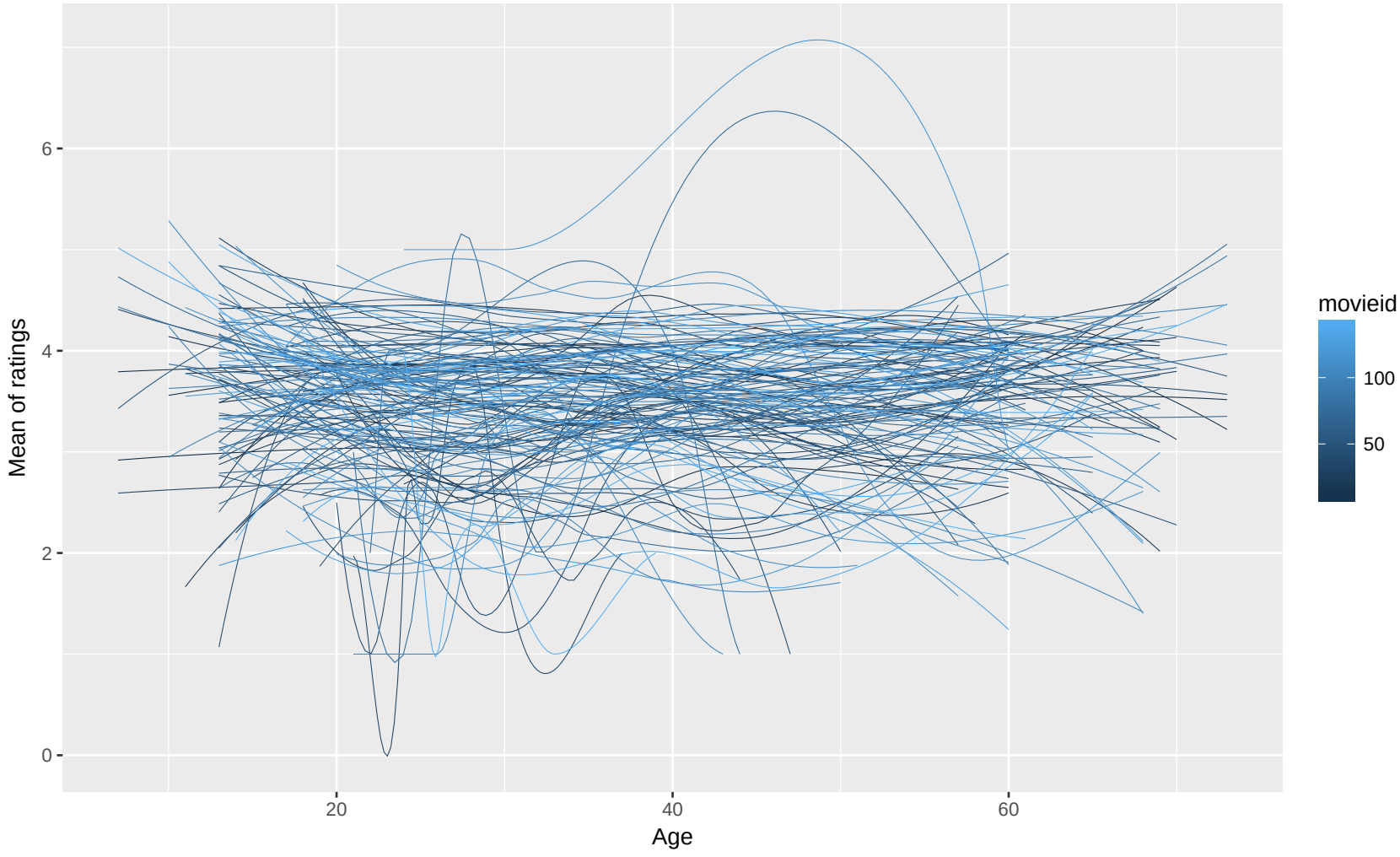

total number of ratings
600
400
200
0
20
40
60
user's age

Comparison of mean and variance of ratings among genders

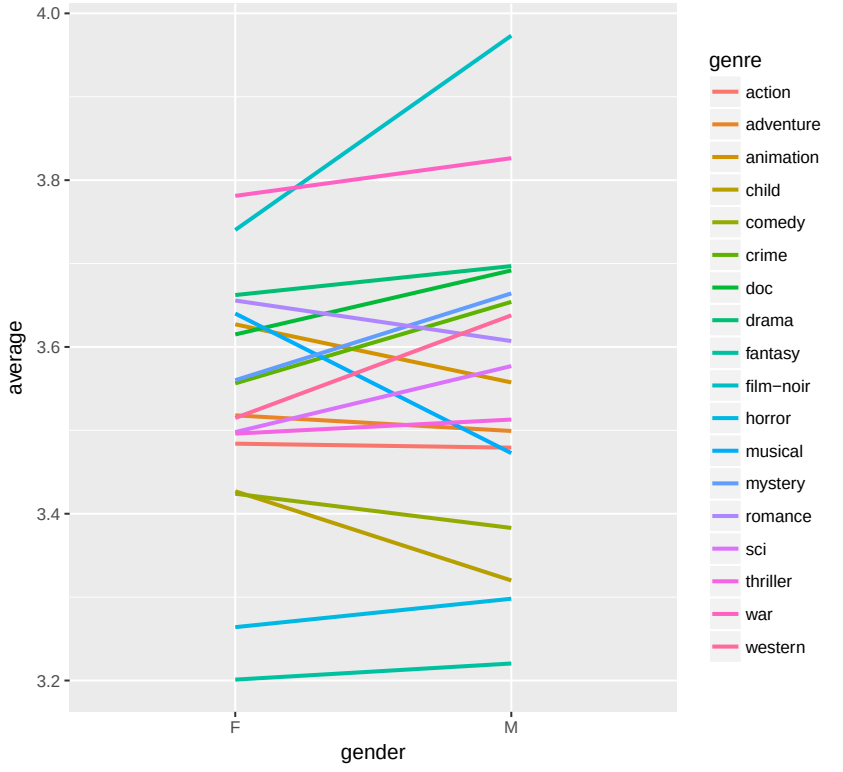


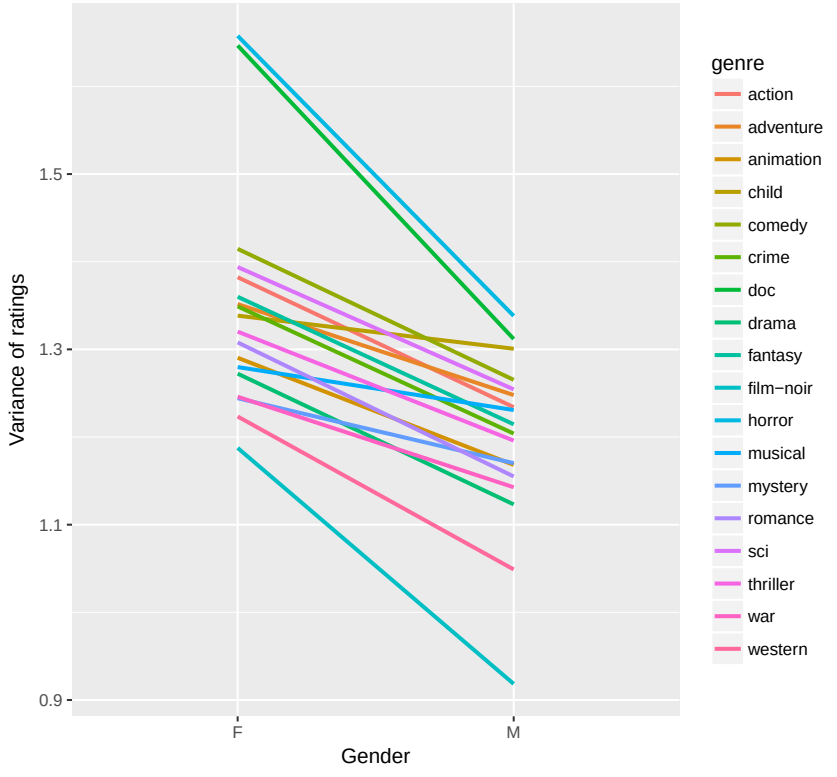


Comparison of mean of ratings among different ages



Comparison of mean and variance of ratings among genders for different movie genres





Content-Based Regression

From the previous analysis, we can see that for an individual movie, users with different age and gender would rate the movie differently. Since the variable age and gender might have effects, for each movie i, we find its all corresponding observed ratings. Then we regress against the corresponding gender and age of a user from each rating score, and calculate β_i. If a movie i does not have enough ratings (≤ 3), then we use the average of β_i from its five most similar movies, where similarily is measured based on movie's genres.