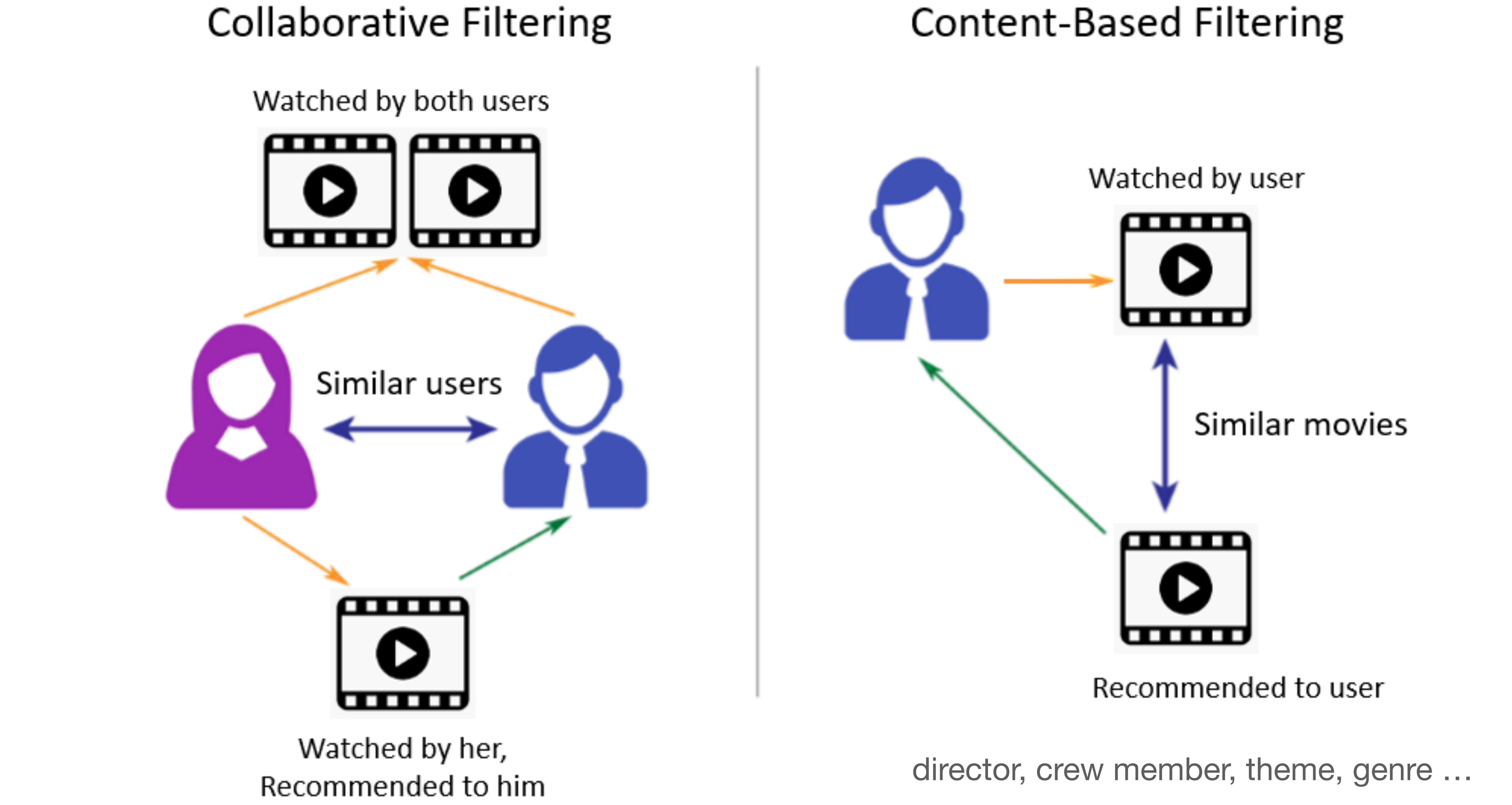
A Movie Recommender System

based on collaborative-based method

What to watch?



Collaborative Filtering

Content-Based Filtering

director, crew member, theme, genre …

graph from https://www.kdnuggets.com/2019/11/content-based-recommender-using-natural-language-processing-nlp.html

Data

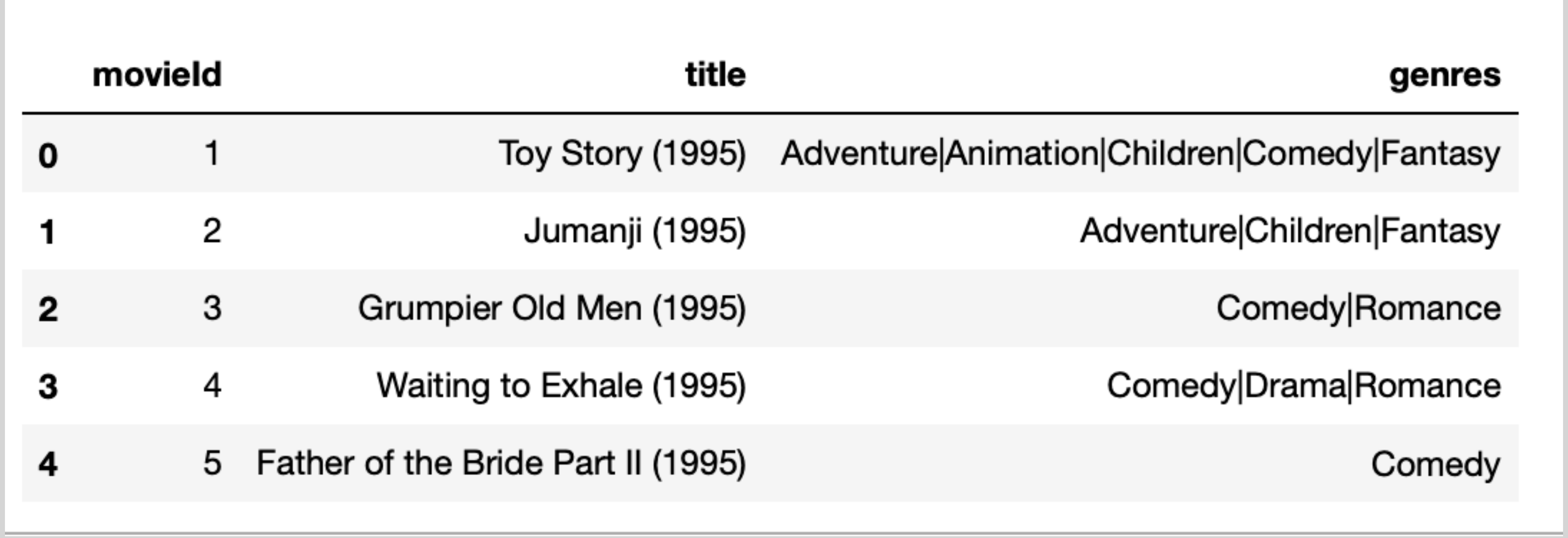
- 62,000 Movies
- 163,000 users
- 25 million 5-star rating records (0.5 lowest, 5.0 highest with step 0.5) between 1995 and 2019

Constraints

- Also recorded in IMDB dataset (drop 3.6%)
- Movies have at least 10 reviewers

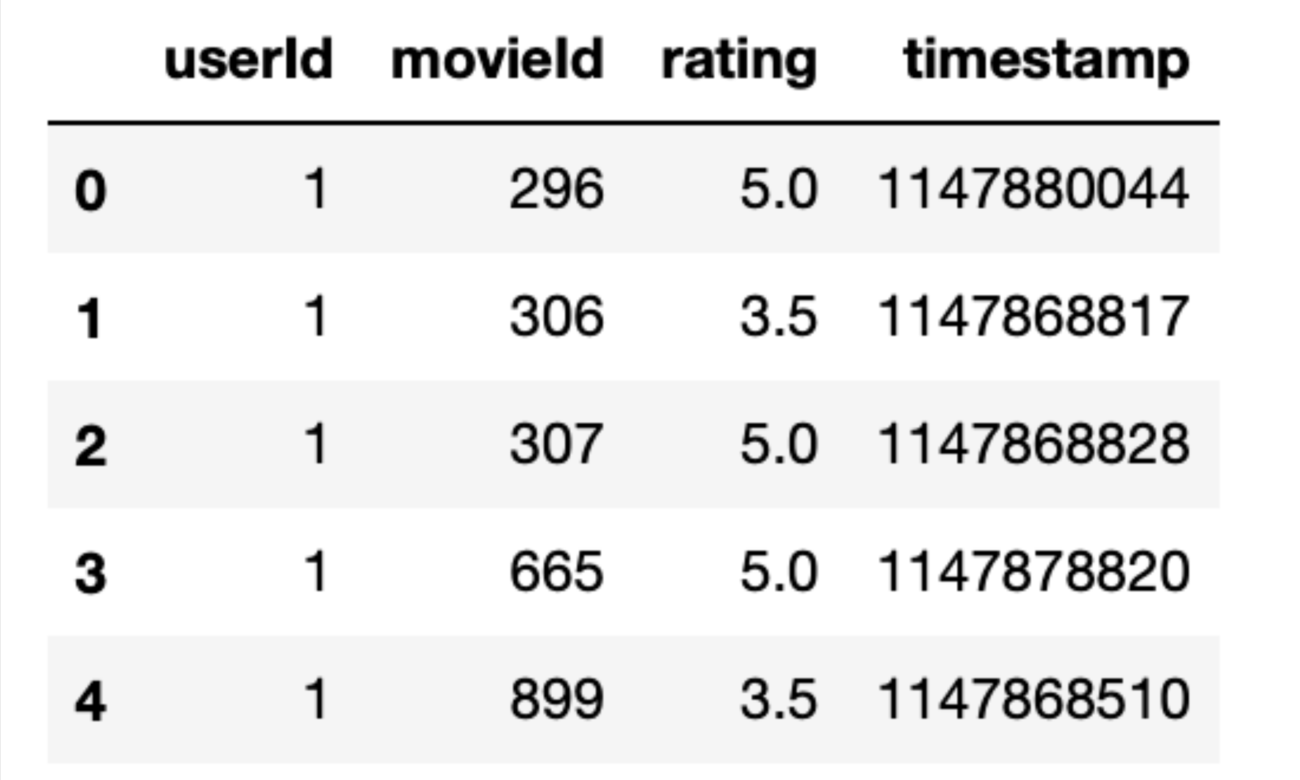
24,430 movies left in the dataset

Data Manipulation and Analysis



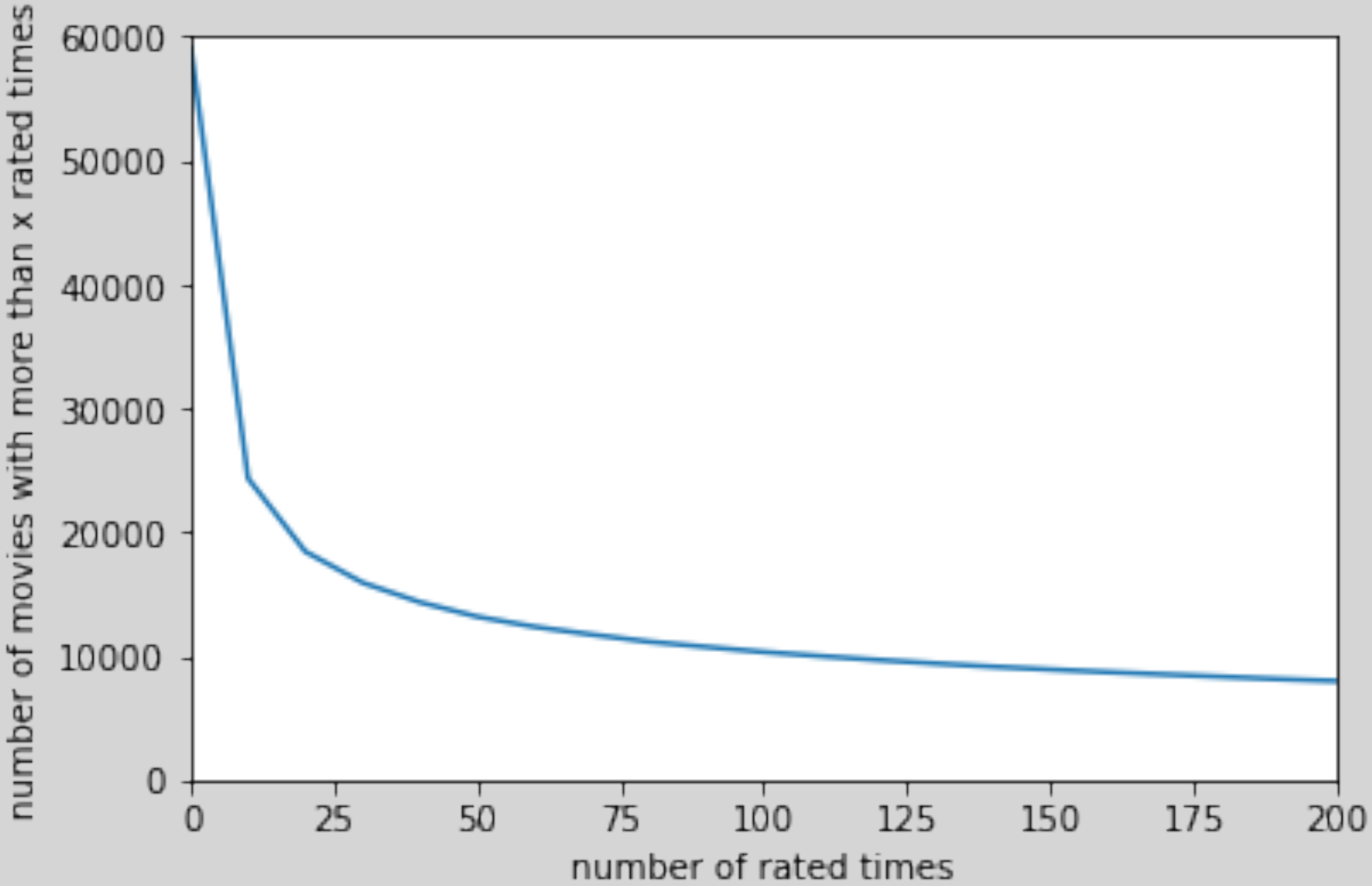
	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

Movies

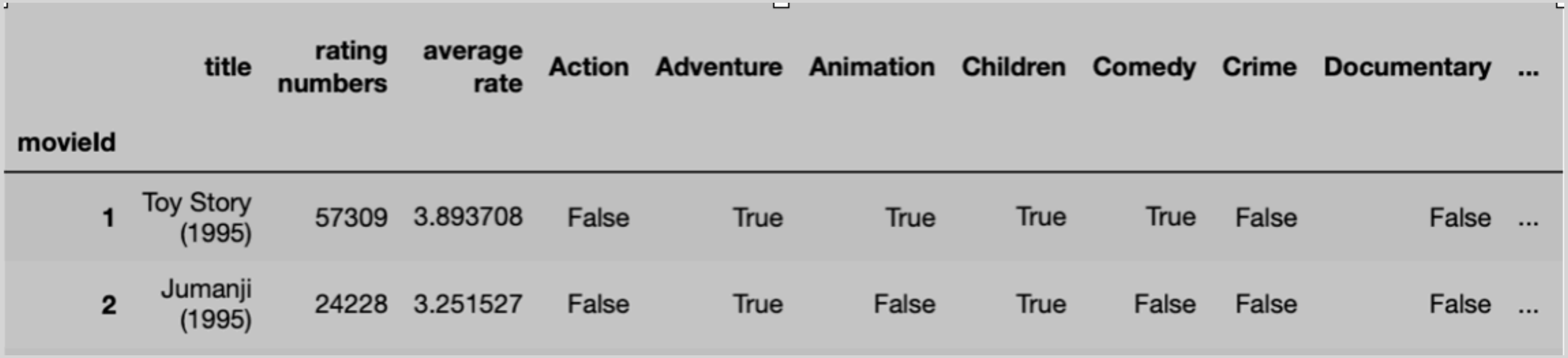


	userId	movieId	rating	timestamp
0	1	296	5.0	1147880044
1	1	306	3.5	1147868817
2	1	307	5.0	1147868828
3	1	665	5.0	1147878820
4	1	899	3.5	1147868510

ratings

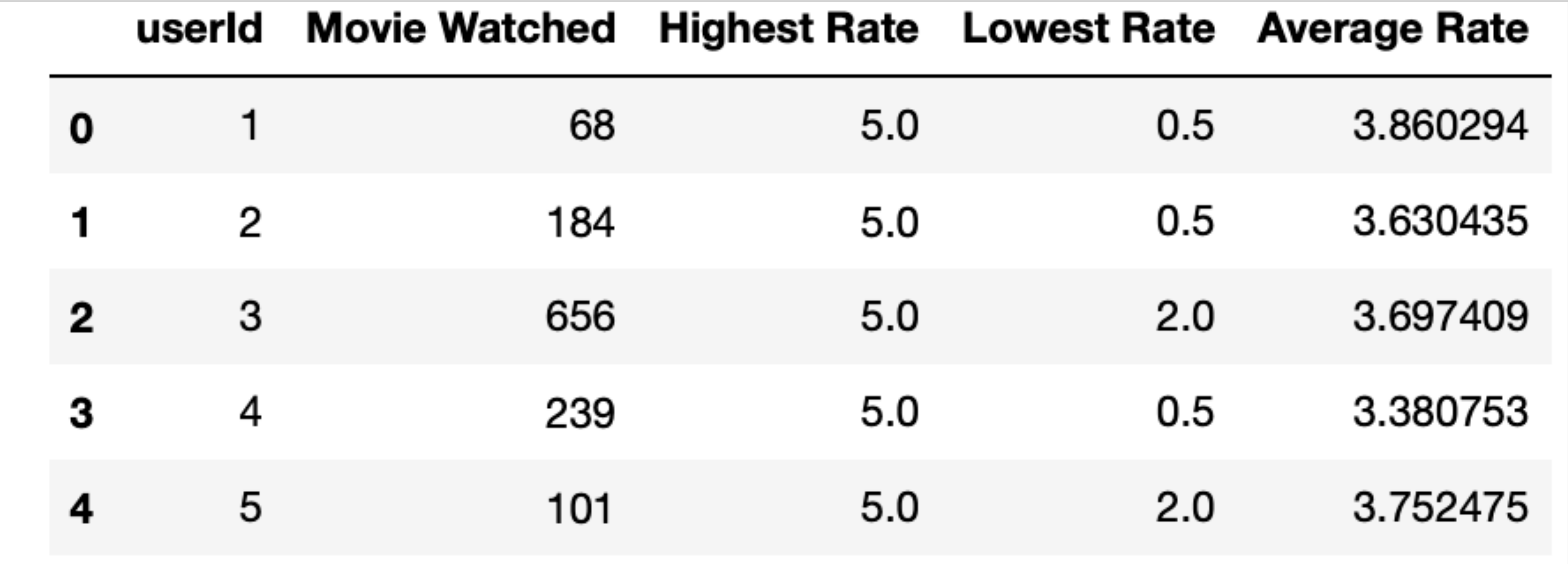


expand genres in boolean columns, add rating numbers, average rates for movies

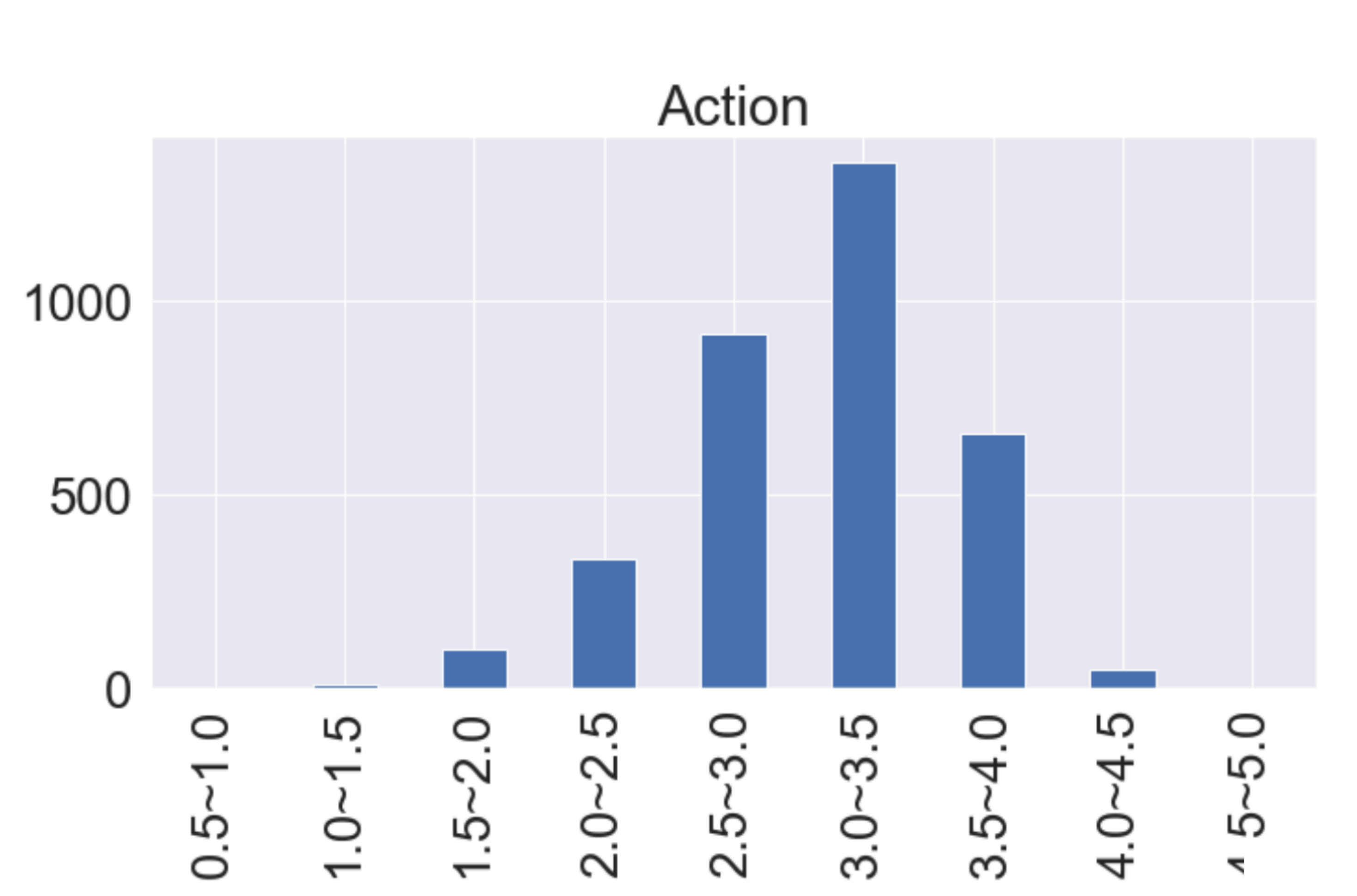


	title	rating numbers	average rate	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	...
movieId											
1	Toy Story (1995)	57309	3.893708	False	True	True	True	True	False	False	...
2	Jumanji (1995)	24228	3.251527	False	True	False	True	False	False	False	...

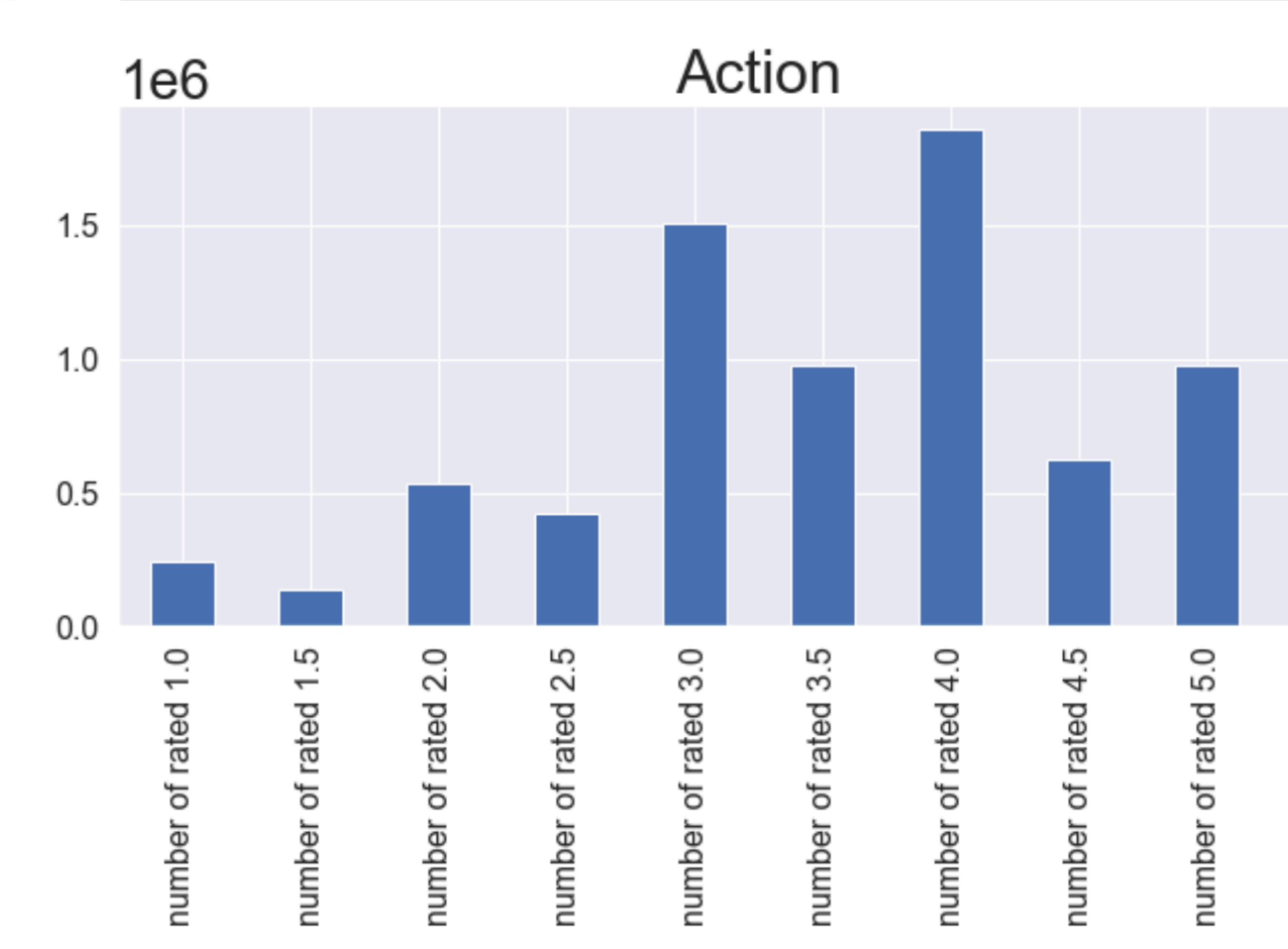
Add statistics for each user



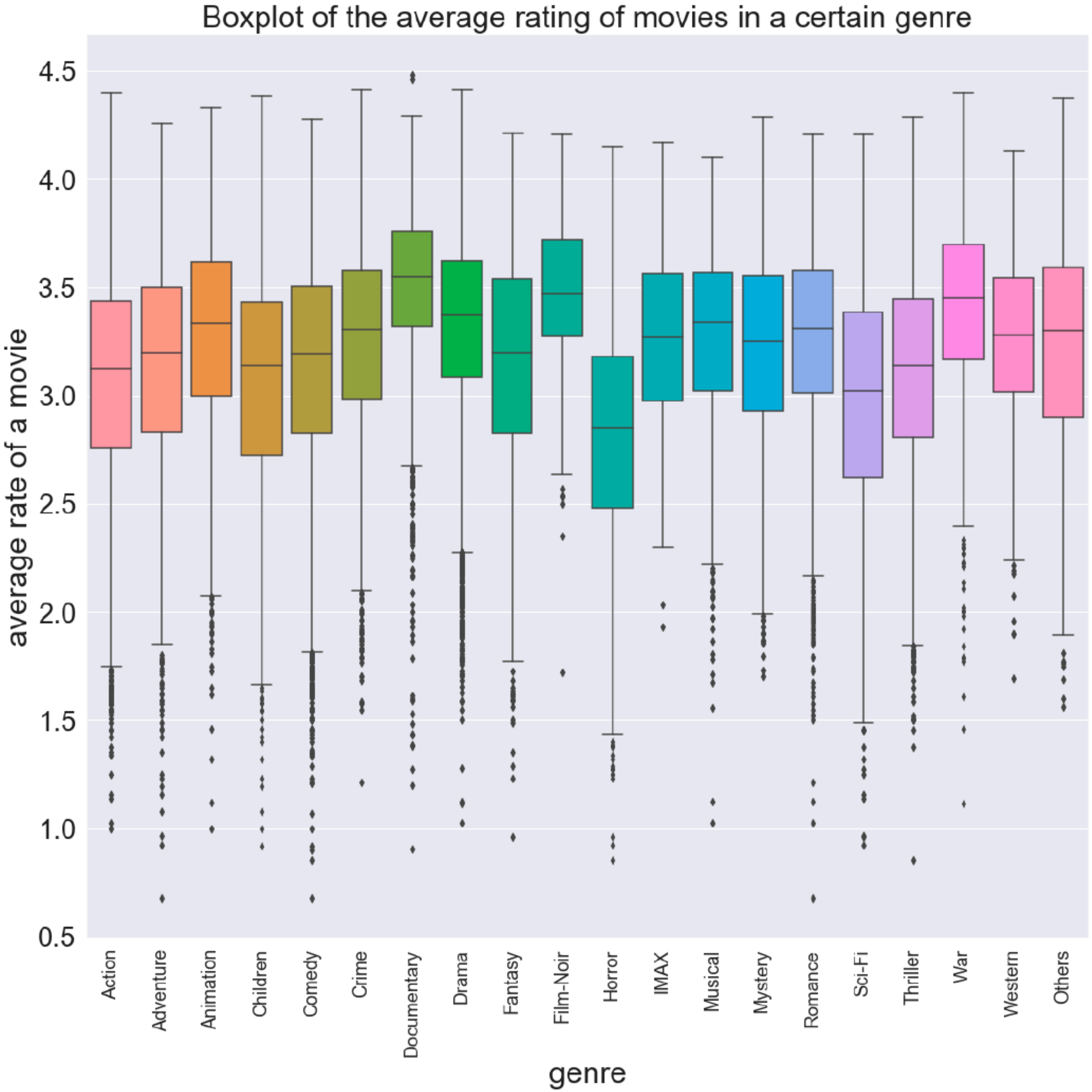
	userId	Movie Watched	Highest Rate	Lowest Rate	Average Rate
0	1	68	5.0	0.5	3.860294
1	2	184	5.0	0.5	3.630435
2	3	656	5.0	2.0	3.697409
3	4	239	5.0	0.5	3.380753
4	5	101	5.0	2.0	3.752475



distribution of the average rating of users



distribution of the average rating of movies


Boxplot of the average rating of movies in a certain genre
average rate of a movie
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Action
Adventure
Animation
Children
Comedy
Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
IMAX
Musical
Mystery
Romance
Sci-Fi
Thriller
War
Western
Others
genre

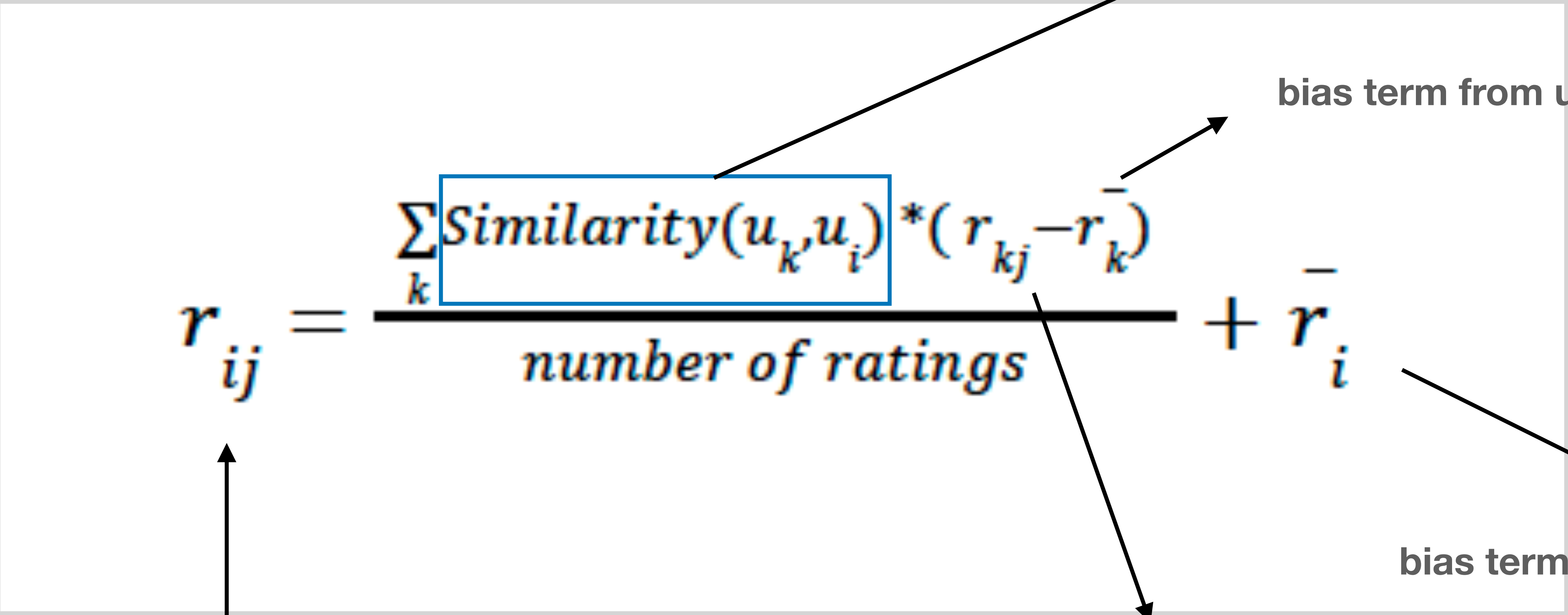
Model

collaborative-based method

(Euclidean distance, Pearson’s coefficient and cosine similarity)

similarity function between user k and i

bias term from user records



$$r_{ij} = \frac{\sum_{k} \boxed{Similarity(u_k,u_i)} *(r_{kj} - \bar{r}_k)}{number\ of\ ratings} + \bar{r}_i$$

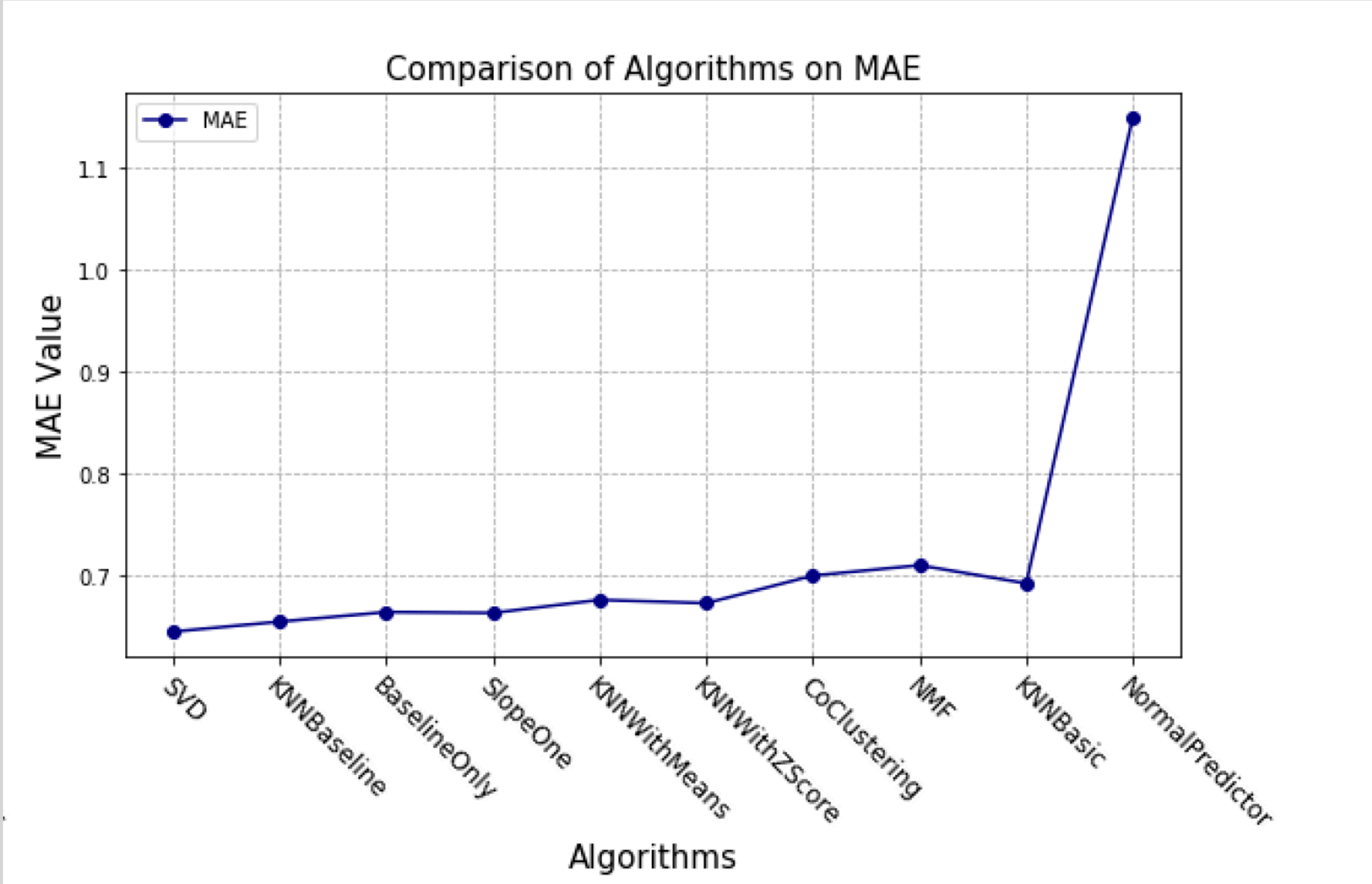
bias term of target user

rate record of user k giving to movie j

target: predicted rate of user i to movie j

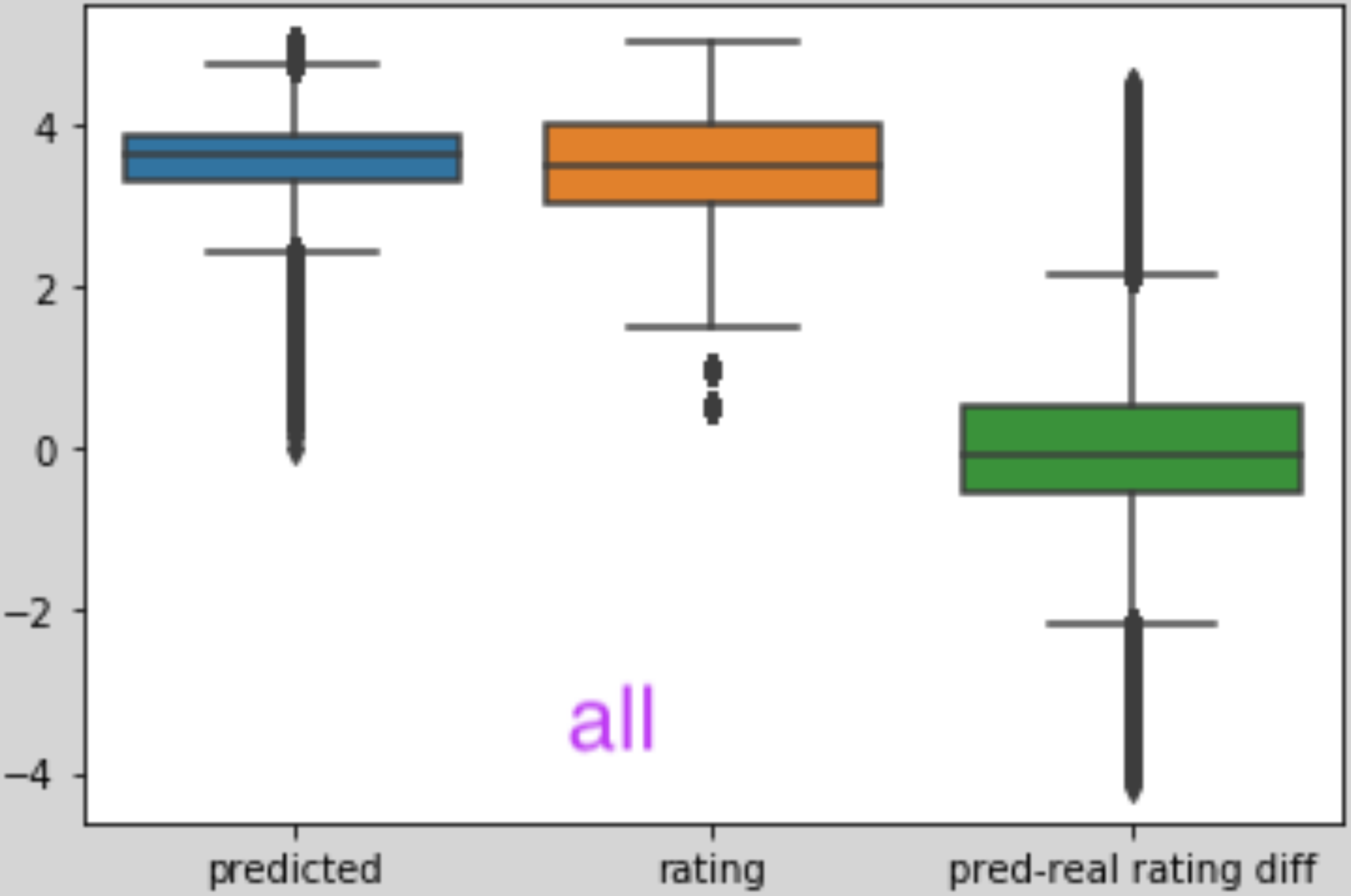
the problem of scarcity and sparsity : matrix factorization

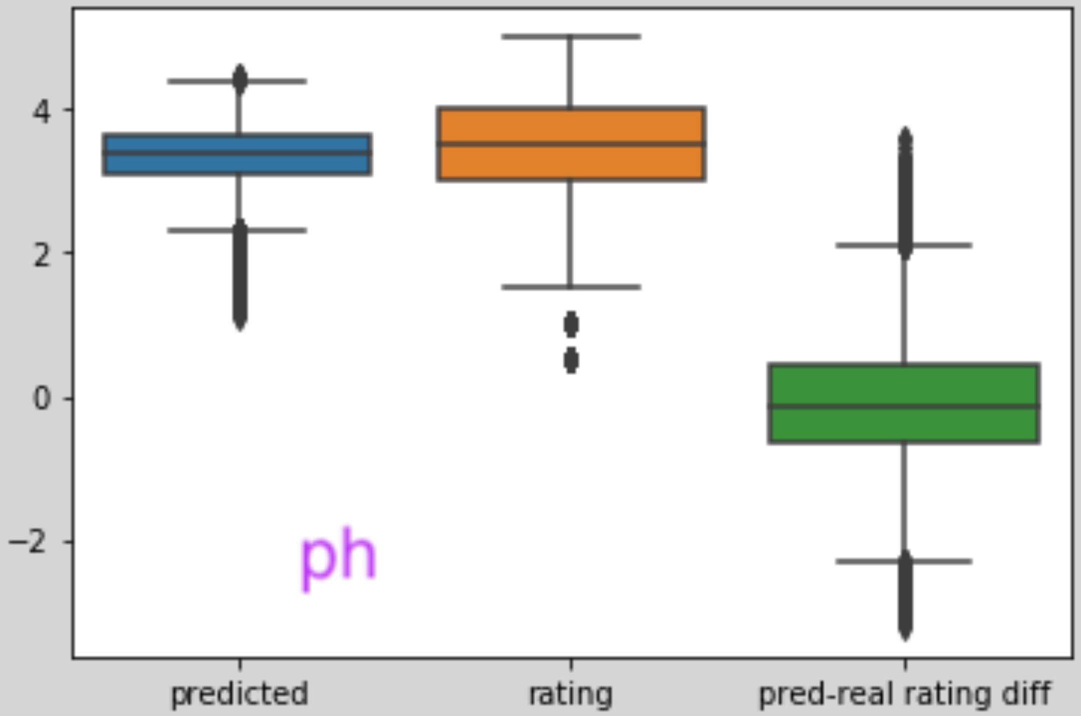
Algorithm test: Singular Value Decomposition (SVD) wins! MAE = 0.68

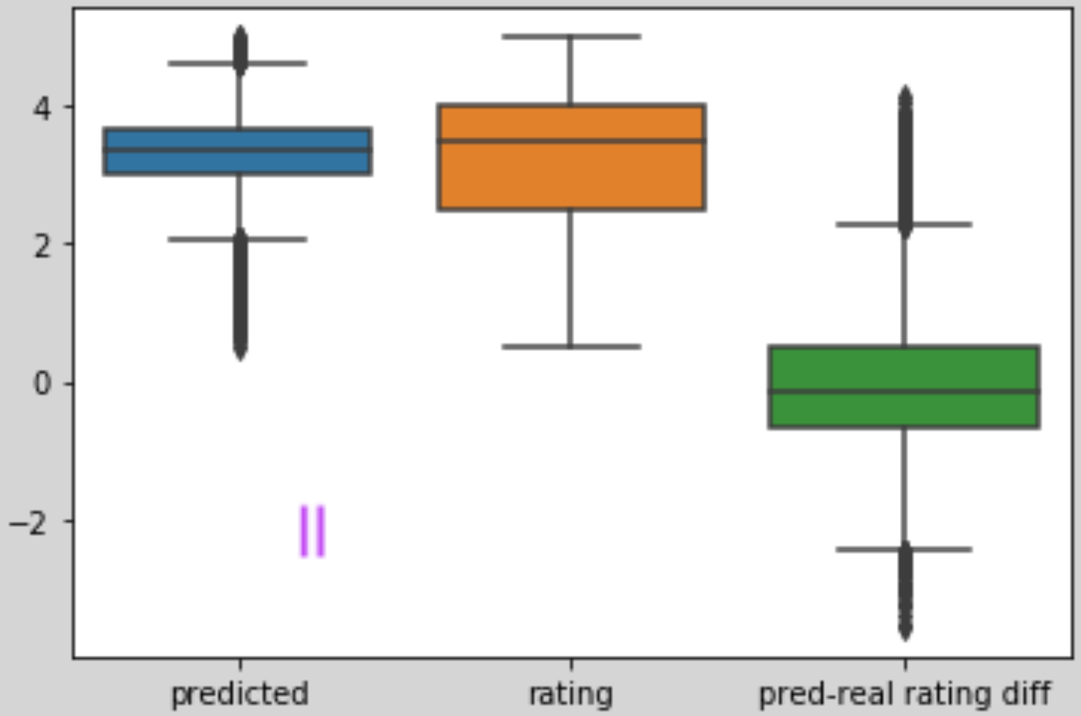


Performance Analysis

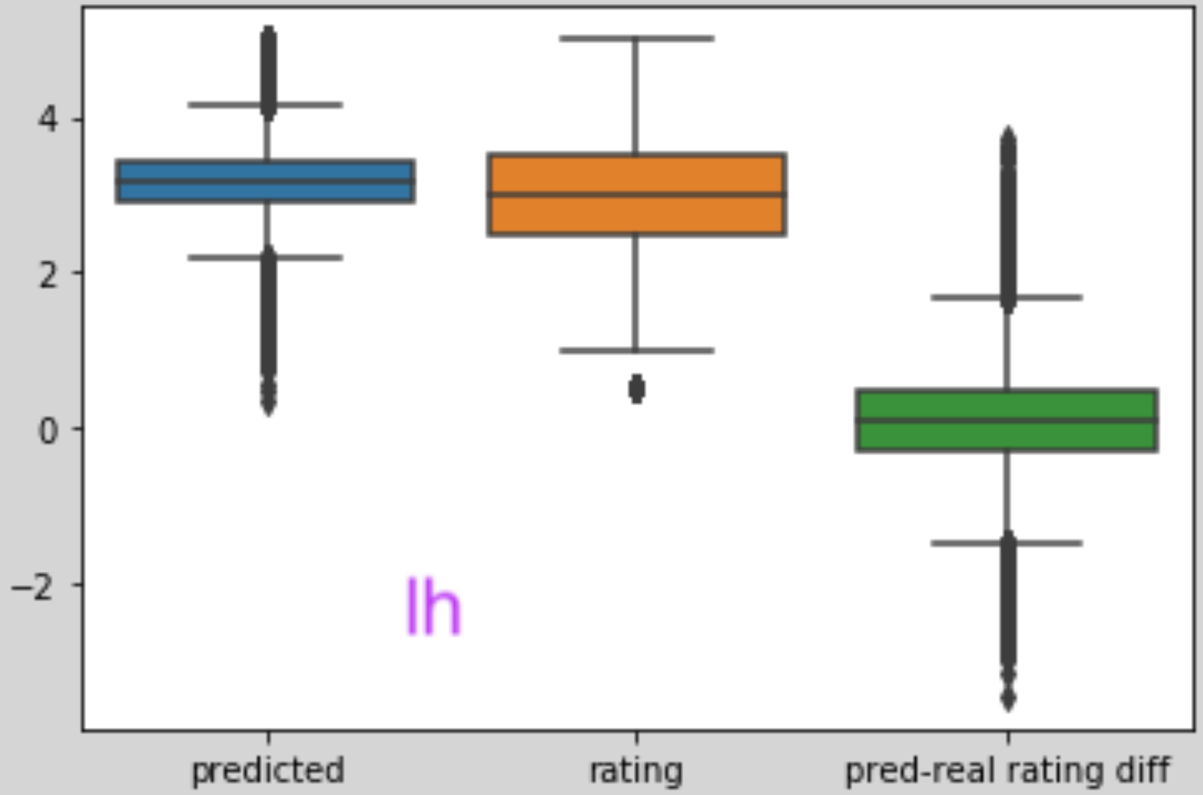
overall box diagrams of the predicted ratings comparing with the actual ratings

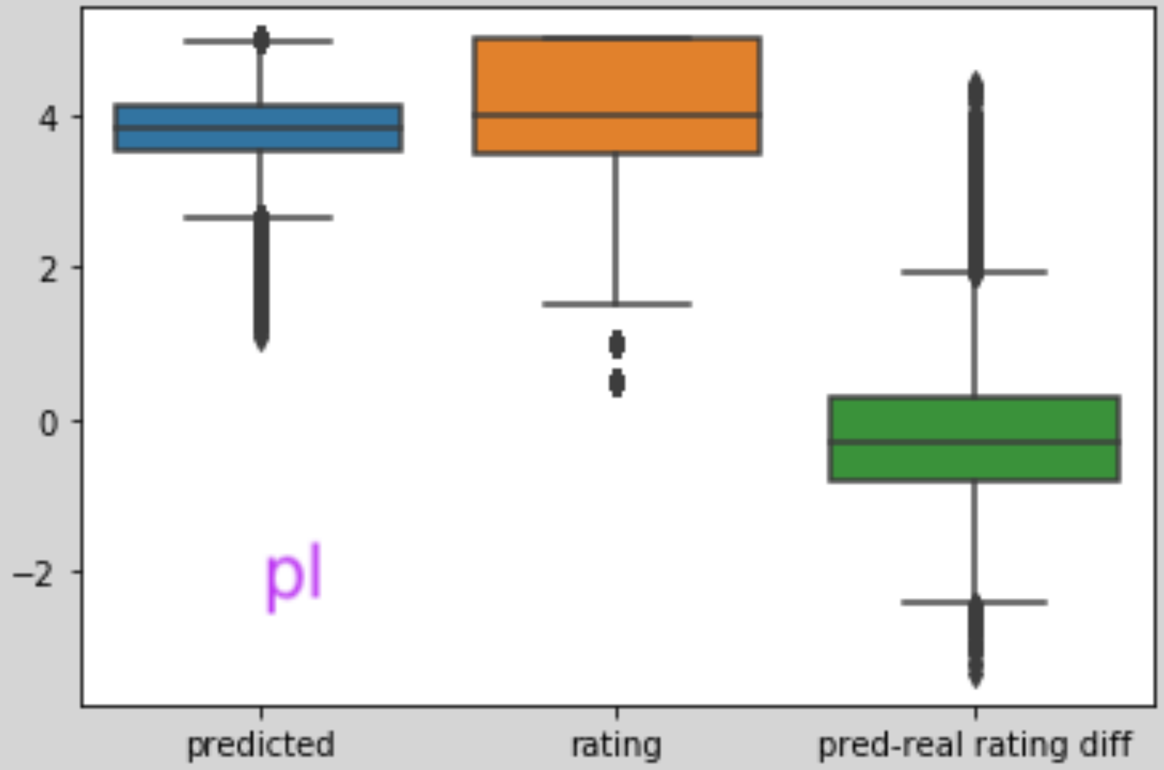


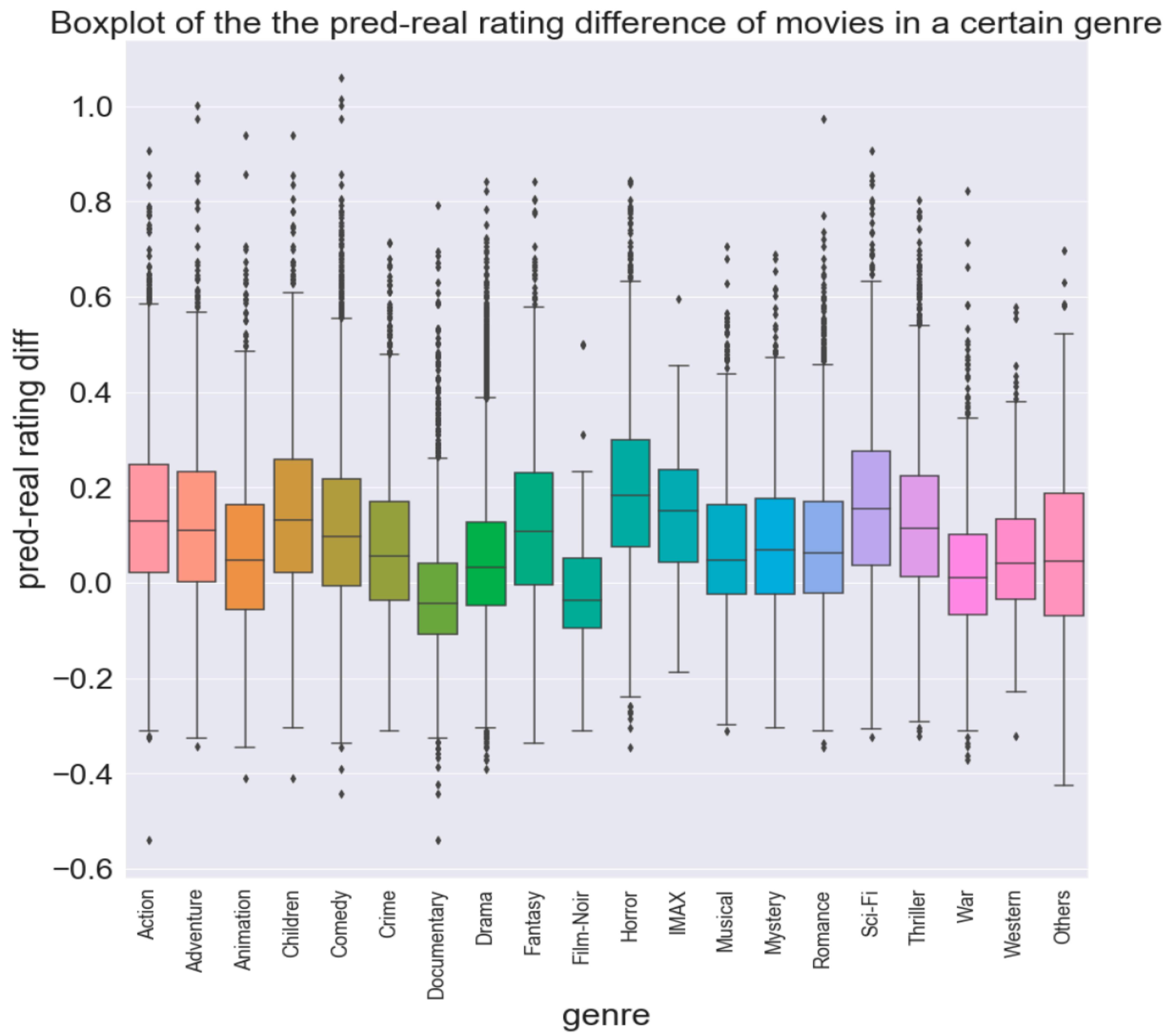




different cases: p(l)h(l) means popular (less-known) movies and heavy (light) users

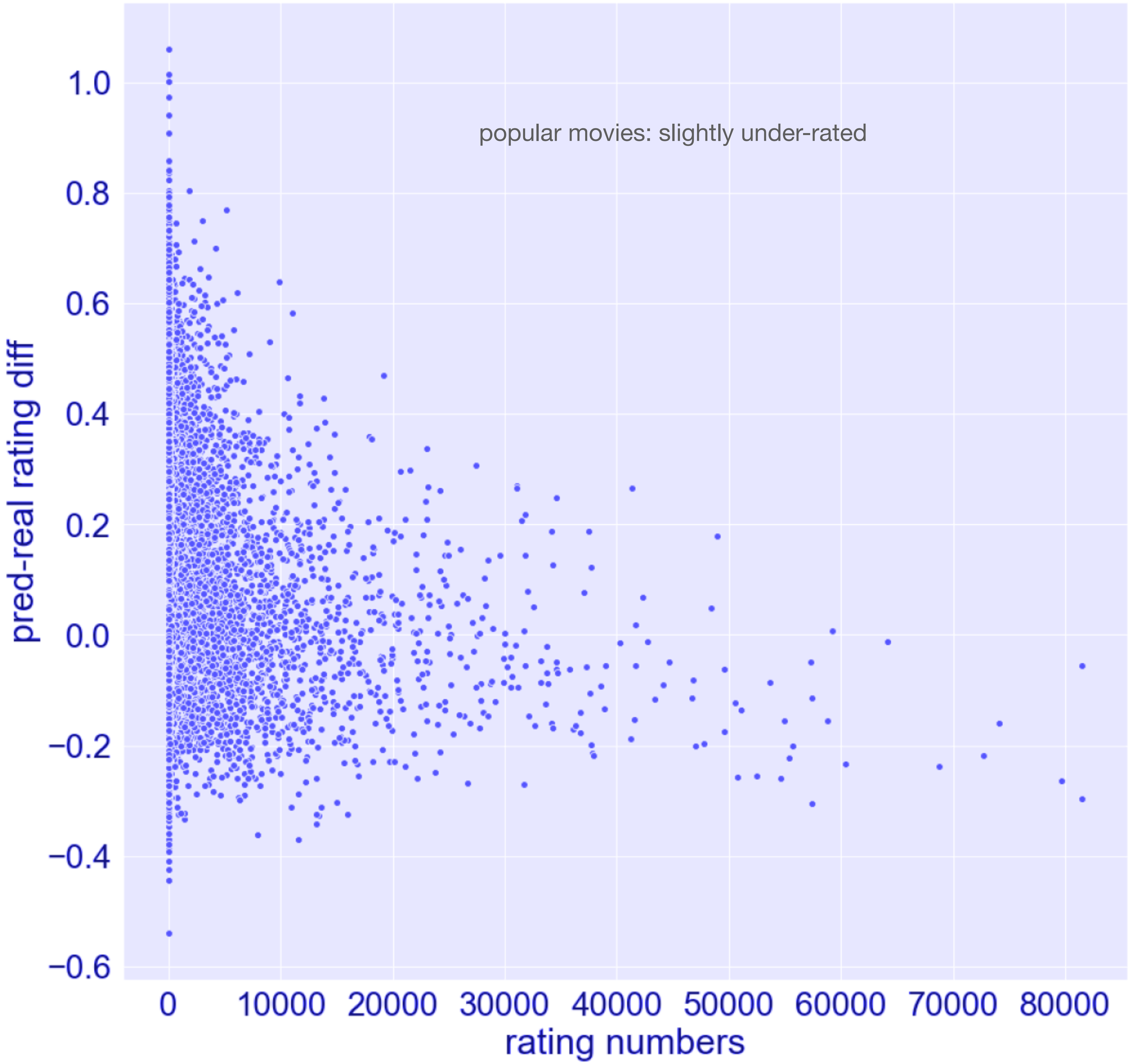




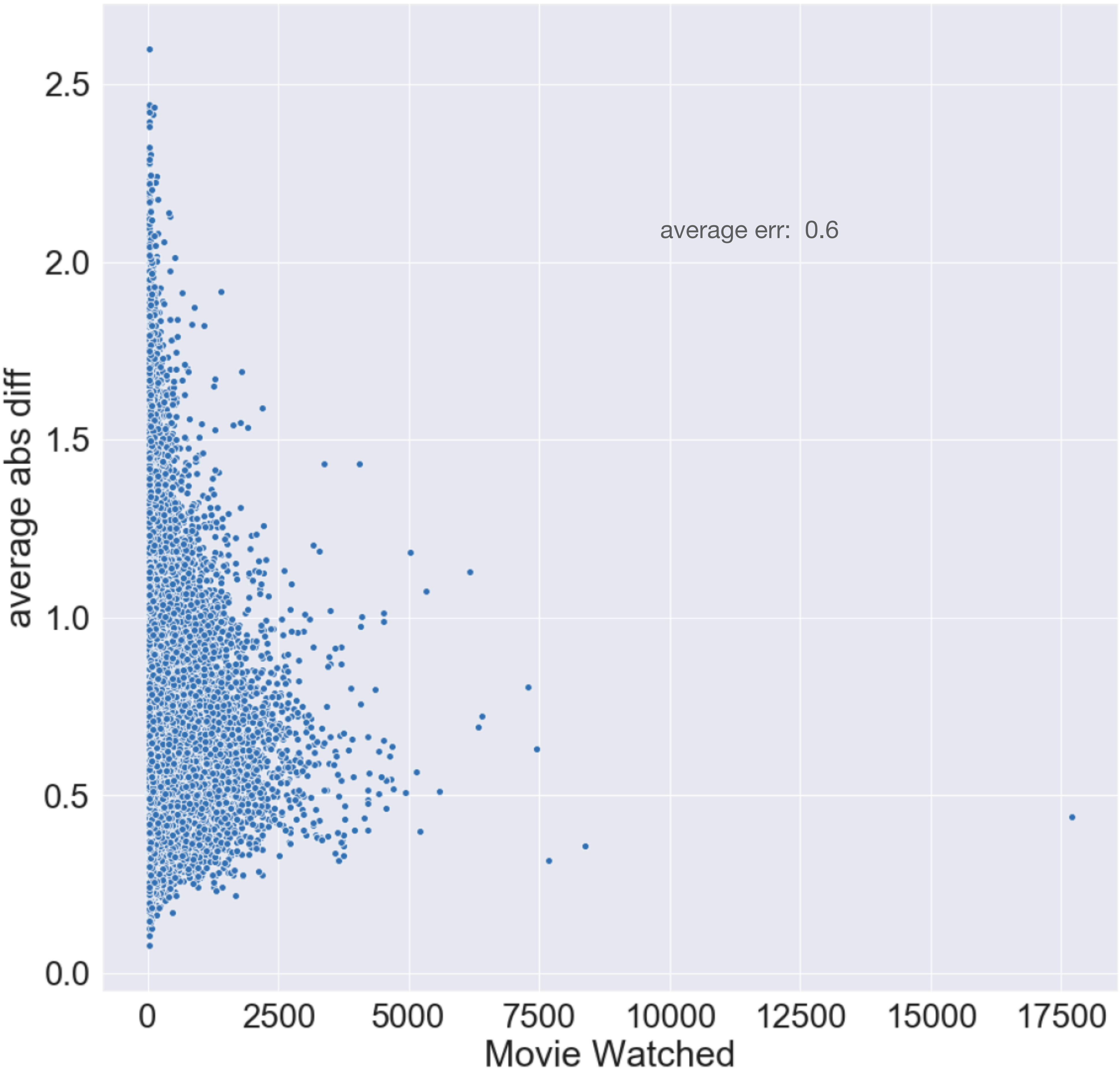


Boxplot of the the pred-real rating difference of movies in a certain genre

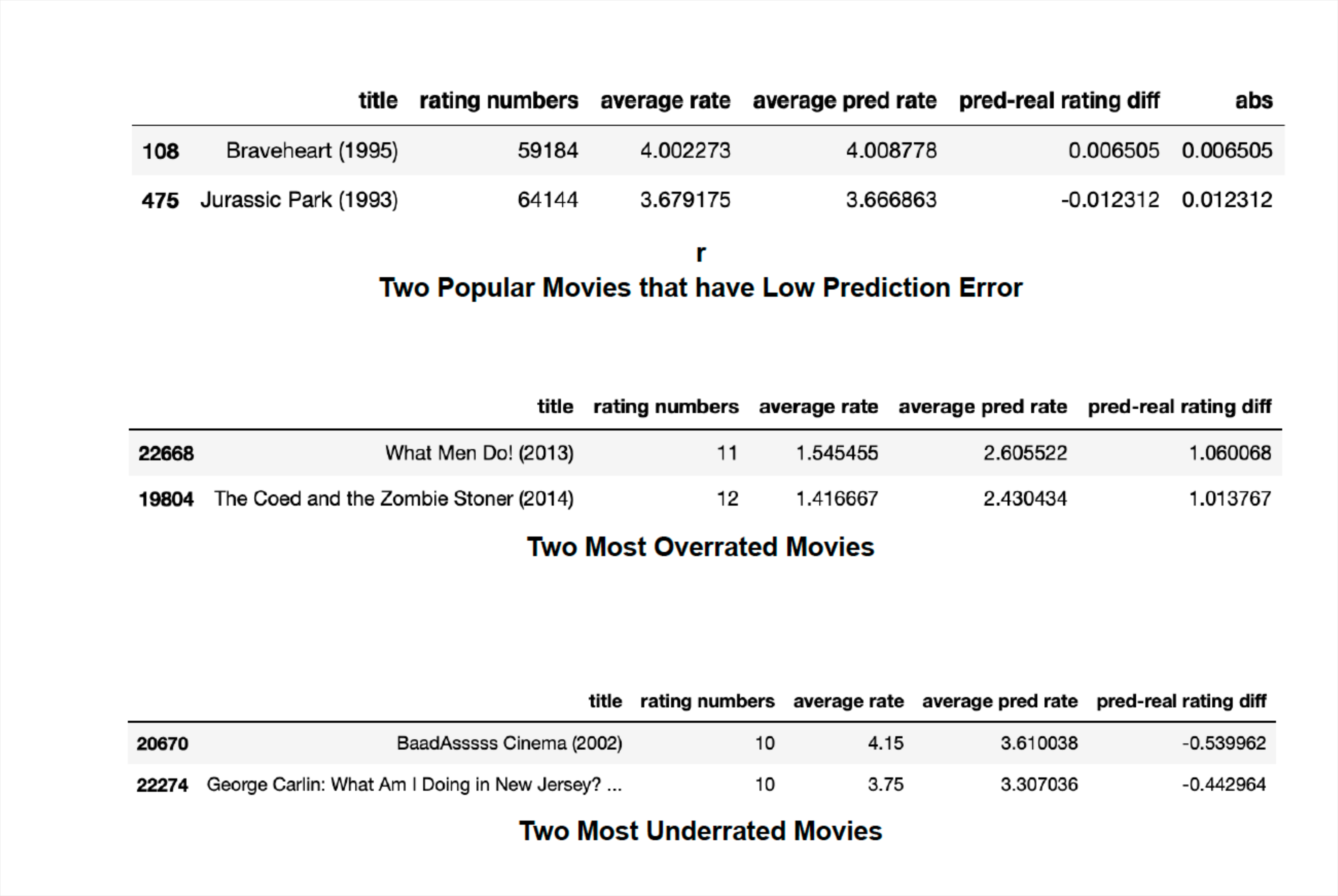
Scattered plot of pred-real rating difference vs. number of reviews of movies


popular movies: slightly under-rated

Scattered plot of pred-real rating difference vs. number of movies watched by users



Examples of concrete predictions



	title	rating numbers	average rate	average pred rate	pred-real rating diff	abs
108	Braveheart (1995)	59184	4.002273	4.008778	0.006505	0.006505
475	Jurassic Park (1993)	64144	3.679175	3.666863	-0.012312	0.012312

r

Two Popular Movies that have Low Prediction Error

	title	rating numbers	average rate	average pred rate	pred-real rating diff
22668	What Men Do! (2013)	11	1.545455	2.605522	1.060068
19804	The Coed and the Zombie Stoner (2014)	12	1.416667	2.430434	1.013767

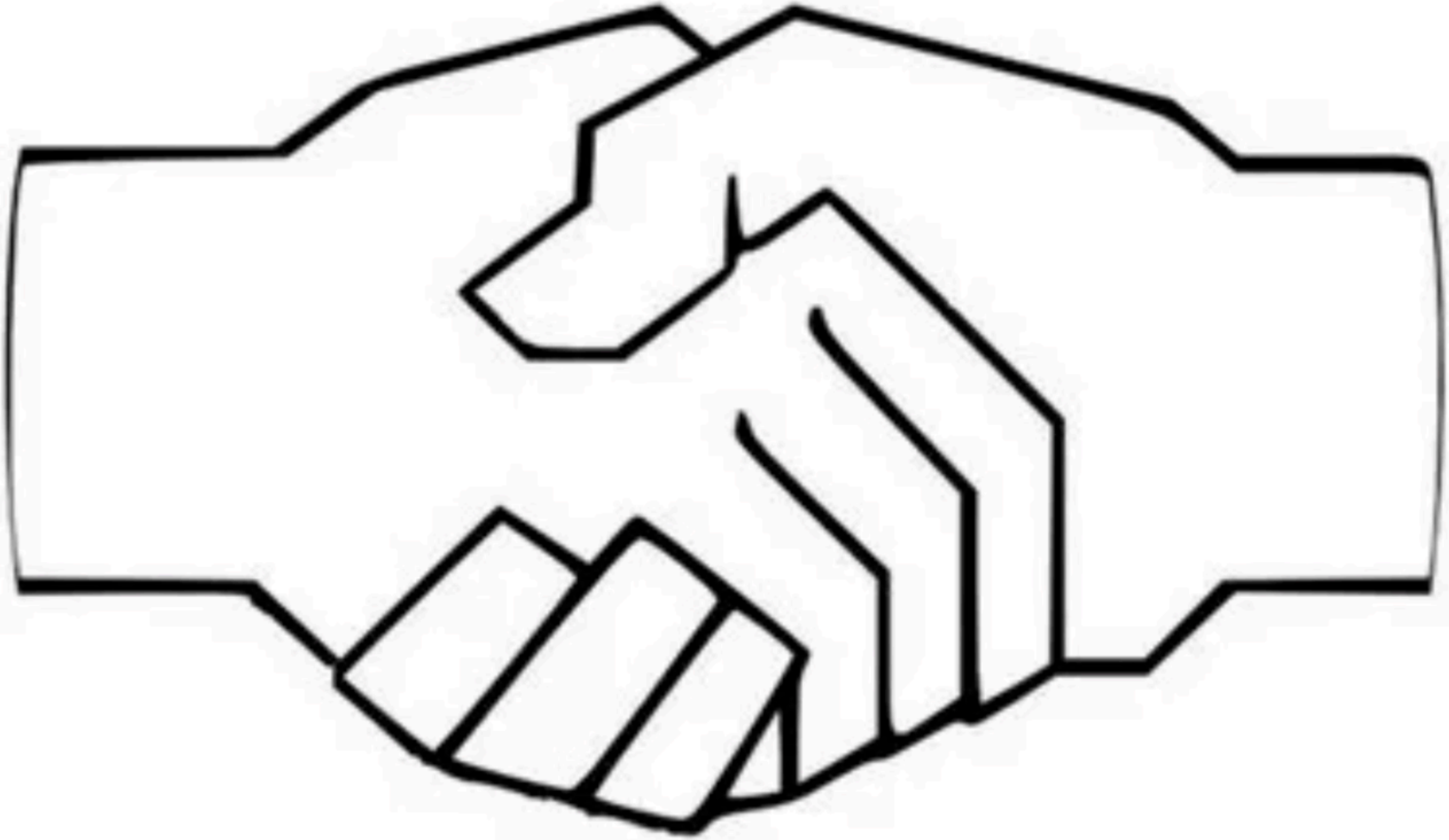
Two Most Overrated Movies

	title	rating numbers	average rate	average pred rate	pred-real rating diff
20670	BaadAsssss Cinema (2002)	10	4.15	3.610038	-0.539962
22274	George Carlin: What Am I Doing in New Jersey? ...	10	3.75	3.307036	-0.442964

Two Most Underrated Movies

Summary

- A collaborative-based method movie recommendation system is built by SVD algorithm, with MAE 0.68
- less-known movies tend to be over-predicted, while popular movies tend to be slightly under-predicted
- for almost every genre of movies, 75% of the median prediction error is within the range of -0.1 to 0.3, within the smallest rating step.
- for a user who has a large number of review records, typically our recommendation system has the average error around 0.6.



THANK YOU