Movie Ratings Over Time

Sam N.

Dataset

The dataset used is from the MovieLens website. It contains 20 million movie ratings applied to 27,000 movies by 138,000 users. The data was released in April 2015, and thus contains no ratings after this date.

Below is a download link for the data.

Data Source: MovieLens web site (filename: ml-20m.zip)
Location: https://grouplens.org/datasets/movielens/

Motivation

The reason for this investigation is that that some movies seem to have a far too high rating upon their release.

I was curious to find out if this is common across many movies and if the ratings tend to trend to a more reasonable rating over time.

This could be useful to know when looking for movies to watch. If the findings support this curiosity, recently released film ratings may need to be taken with a grain of salt when compared to older movie ratings.

Research Question

Are movie ratings more biased at time of release, and do they become less one sided, and thus potentially more reasonable, over time?

Findings

While the difference is small, it does appear that movie ratings made in year of release show an upward bias. Further investigation in to particular years or for particular movies could show interesting results.

Below are statistics and percentiles of the difference between Year of release ratings and Subsequent year ratings.

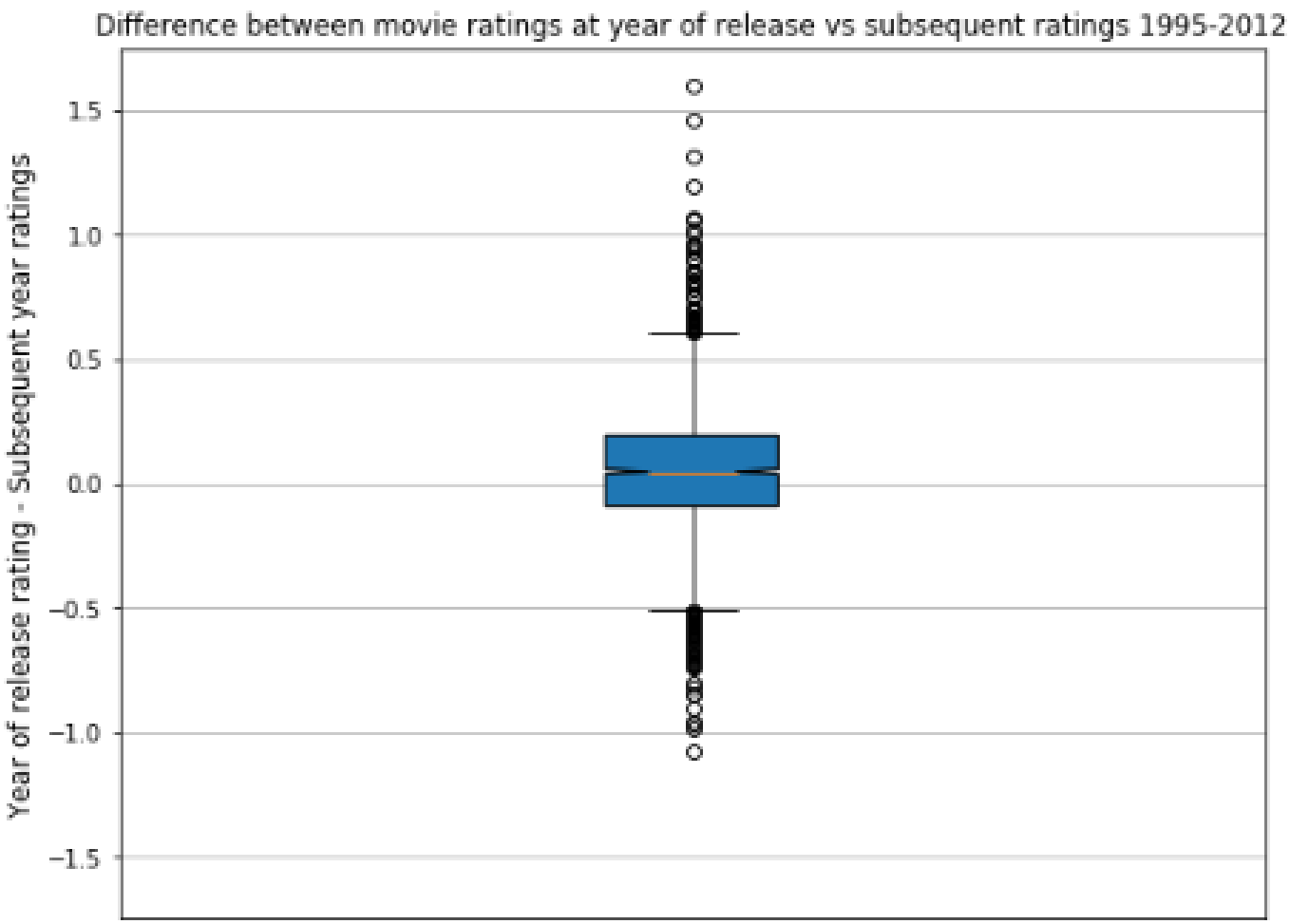
Mean	0.054308	25%	-0.089042
Min	-1.079475	50%	0.049242
Max	2.033898	75%	0.191230

Findings

On average, movies in our dataset have a 0.054 higher rating in the year of release compared to subsequent year ratings. The percentile ranges also show signs that year of release ratings tend to be higher.

25% of ratings are 0.191 higher in the year of release at the upper end, and at the bottom end, 25% of ratings are only 0.089 lower on year of release. Lets visualized this in a boxplot.

Mean	0.054308	25%	-0.089042
Min	-1.079475	50%	0.049242
Max	2.033898	75%	0.191230


Difference between movie ratings at year of release vs subsequent ratings 1995-2012
Year of release rating - Subsequent year ratings
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

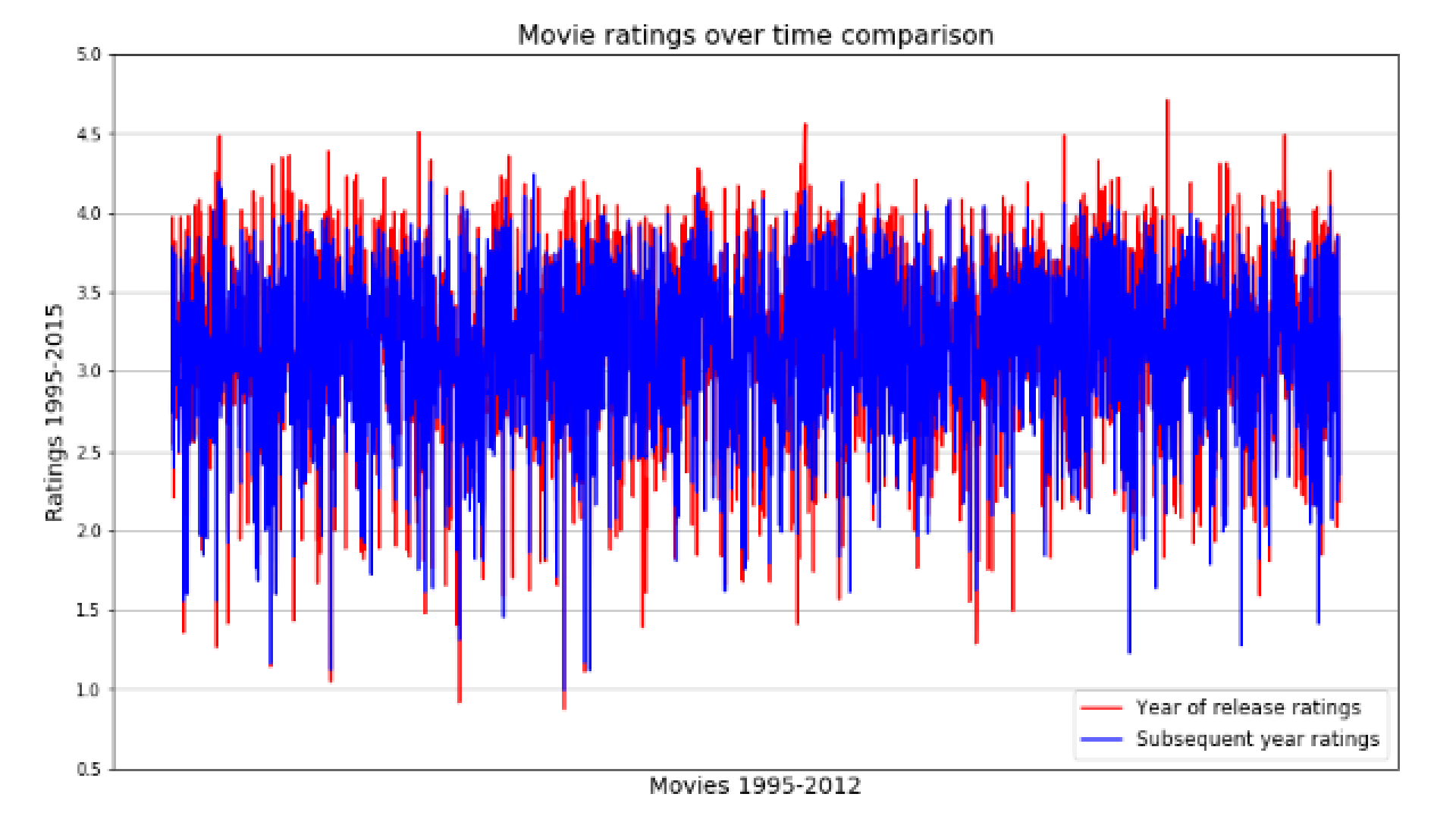
Findings

We can also visualized this in a line chart, comparing Year of release ratings and Subsequent year ratings.

Here we can see a clear dominance of year of release ratings compared to subsequent year ratings for higher rated movies.

For lower rated movies it is less clear, although year of release ratings do seem to be lower.

For medium rated movies it is hard to tell from this visualization.

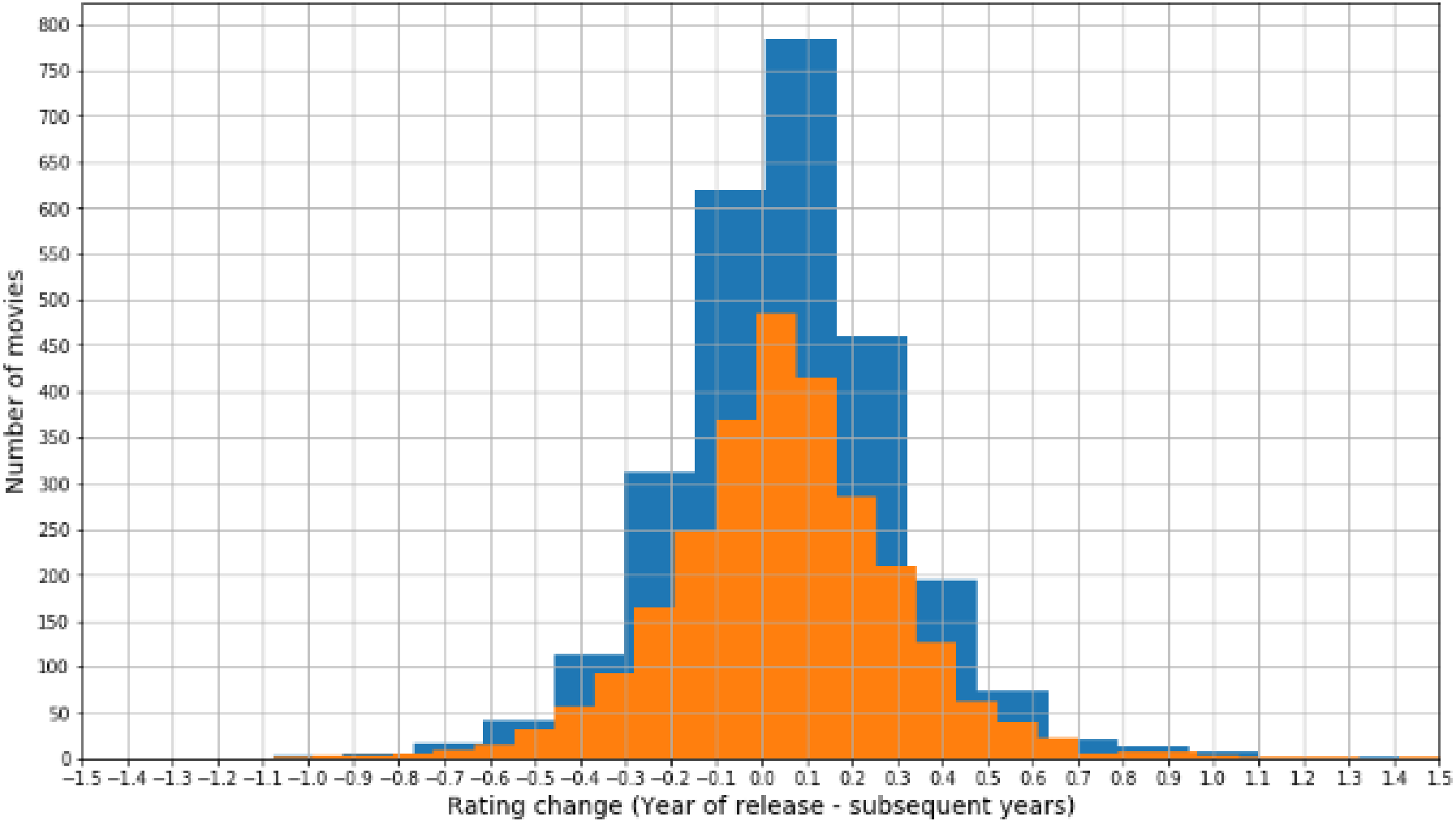

Movie ratings over time comparison
Ratings 1995-2015
Movies 1995-2012
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Year of release ratings
Subsequent year ratings

Findings

We must ask ourselves why this is. Perhaps movie ratings change over time due to movies being too contemporary and dating quickly. Or could it be due to hype of movies upon release causing year of release ratings to be higher, while being lower in subsequent years once the hype is gone?

Further investigation is required to answer these questions.

Finally, lets plot the change in ratings in a frequency distribution. The slight bias towards higher ratings in year of release is easily noticeable here with a larger portion of the rating differences on the right (positive) side of the distribution.


Number of movies
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
−1.5 −1.4 −1.3 −1.2 −1.1 −1.0 −0.9 −0.8 −0.7 −0.6 −0.5 −0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5
Rating change (Year of release - subsequent years)

Acknowledgements

I discussed the project with my lady – thank you lovely.

References

I did the work on my own, with help from the UCSD Jupyter Notebooks completed in the weeks leading up to this project.

These notebooks used the same movie ratings dataset in Week 4 – Pandas.

Jupyter Notebook

I was unable to attach the Jupyter Notebook to the upload for this project. Please find a link to my GitHub below containing the notebook created.

It includes further bits of information including movies whose ratings changed the most after year of release. You are welcome to explore further yourself.

https://github.com/stnorling/Python-DS-UCSD/blob/master/Week-6-Mini-Project/movie_rating_bias.ipynb