Movie Complexity and Ratings

Ingo Nader

Dataset(s)

In the analysis, the IMDB movie dataset was used[(1)]. It describes 5-star rating and free-text tagging activity from [MovieLens](http://movielens.org), a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These data were created by 138493 users between January 09, 1995 and March 31, 2015.

Of this dataset, two tabular files were used for the analysis:

- `movies.csv`: movie IDs and movie genres (and title – not used in this analysis)
- `ratings.csv`: movie IDs and ratings per user (and userID, timestamp – not used)

---

(1) available from: http://files.grouplens.org/datasets/movielens/ml-20m.zip

Motivation

Personally, I find movies that are too simple not entertaining. I tend to like more complex movies better. Hence, in order to shed light on this in a more empirical way, I wanted to research the relationship between complexity and movie ratings, the hypothesis being that this relationship of more complex movies being rated more positively holds for a more general population (i.e., in this dataset).

Research Question(s)

I tried to answer the following research questions in my work:

- Is there a relationship between (average) movie rating and movie complexity?
- Does this relationship vary for different movie genres?

Movie complexity was measured directly (because it is not part of this dataset), but via a heuristic: It was assumed that more complex movies are categorized into more genres. Hence, the measure for complexity is the number of genres that the movie was associated with.

Findings

- Overall, the hypothesis of a (linear) relationship between complexity and rating is not supported by data (Pearson correlation r = -0.02).
- It seems that the variance of movie ratings varies with movie complexity:
  - Less complex moves have higher spread of ratings, ranging from 0.5 to 5.0.
  - More complex movies have lower spread of average ratings.

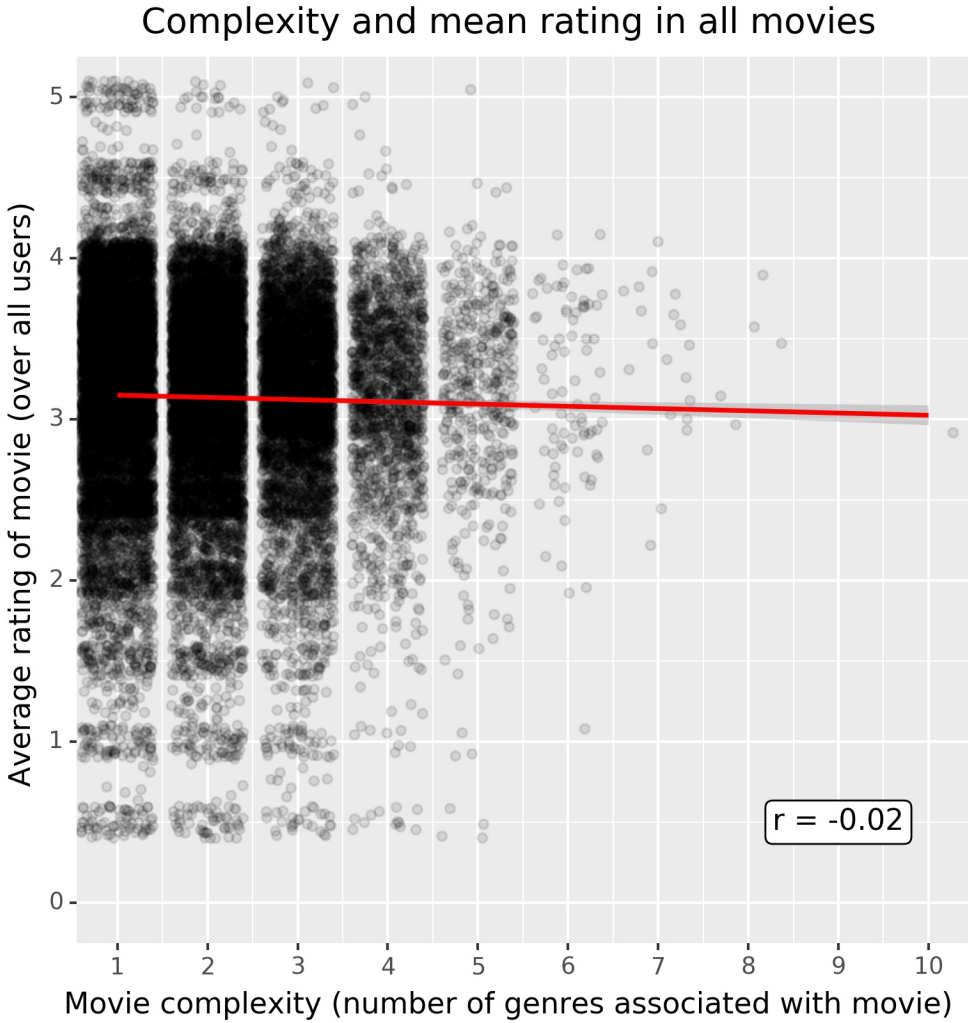
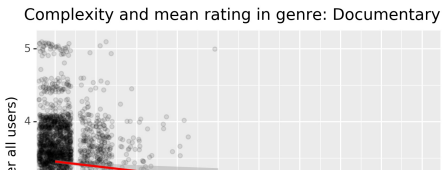
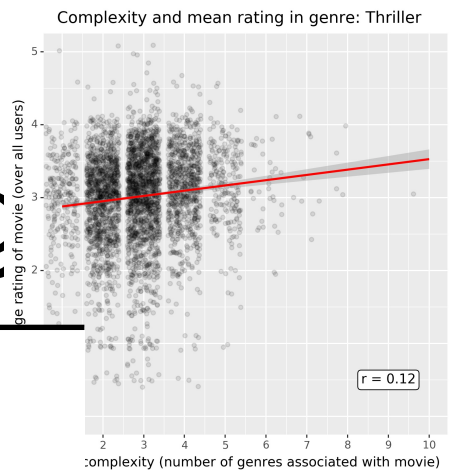


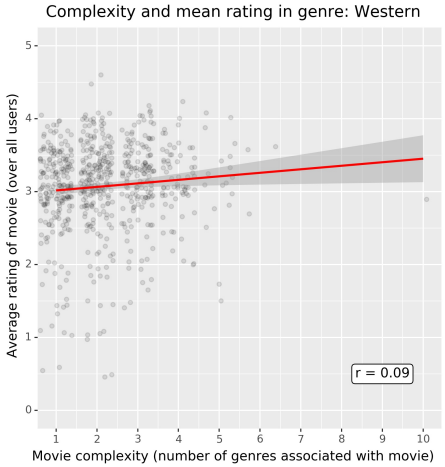
Figure: Movie ratings averaged over all users for each movie, complexity as measured by the number of genres indicated for each movie. Data is jittered to better show spread (random jitter added for ratings: ±0.1; for complexity: ±0.4). Red line depicts a linear regression model fitted to the data in the plot.

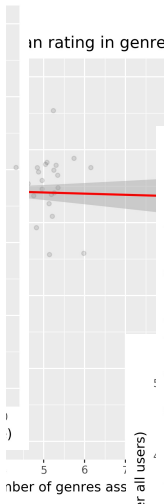
Finding

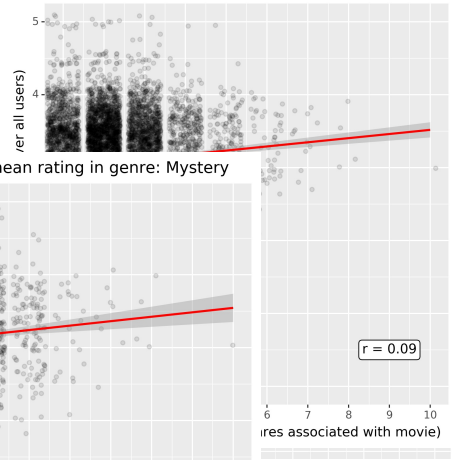
variable	n	cor
is_horror	2279	0.197753
is_children	986	0.182484
is_action	3073	0.157140
is_scifi	1561	0.147570
is_thriller	3681	0.132066
is_comedy	7173	0.116801
is_mystery	1295	0.098975
is_animation	895	0.096643
is_imax	193	0.090634
is_musical	918	0.089673
is_adventure	2019	0.069309
is_western	547	0.046833
is_romance	3561	0.044360
is_fantasy	1255	0.040684
is_crime	2531	0.033876
is_filmnoir	297	-0.000481
is_documentary	1827	-0.037644
is_drama	11397	-0.042163
is_war	1040	-0.073160
is_nogenreslisted	0	NaN

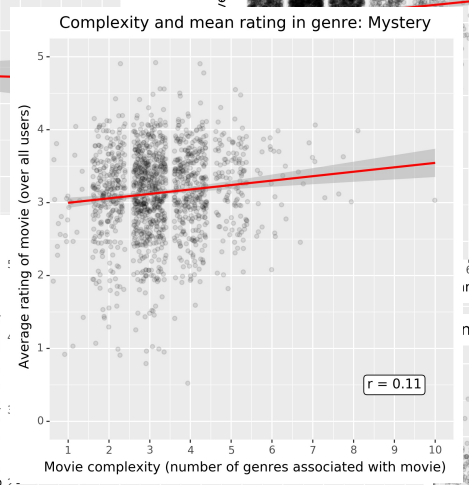


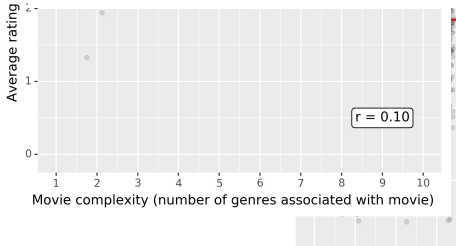


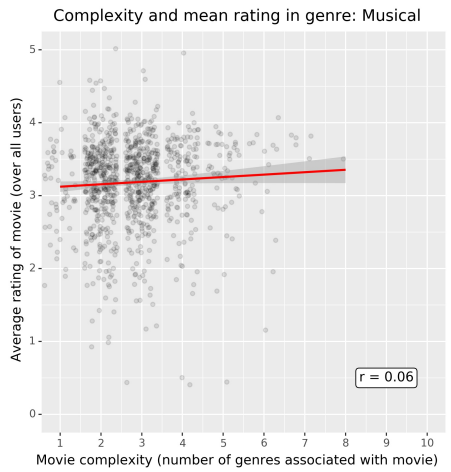


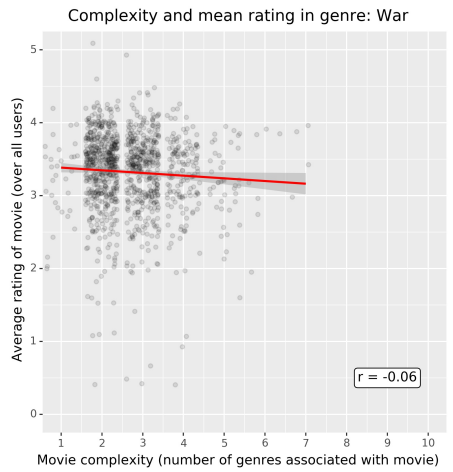


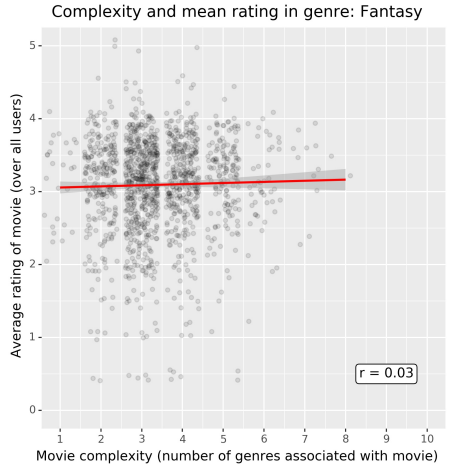


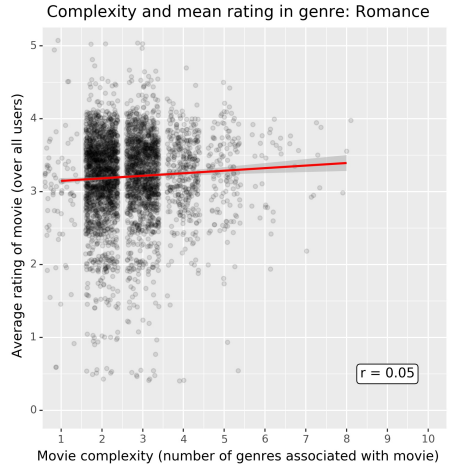


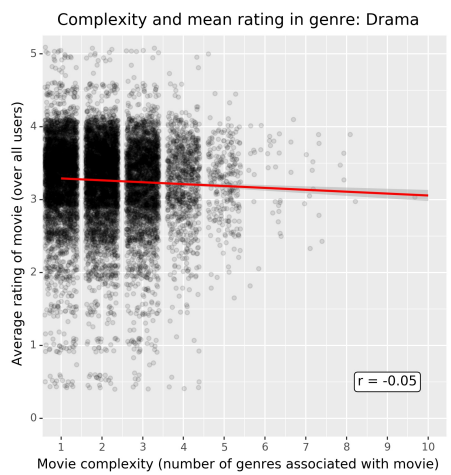


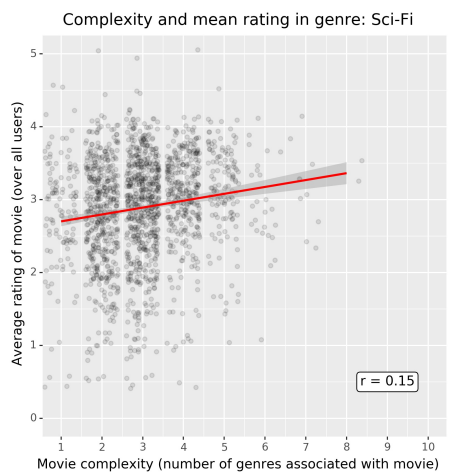












Findings

Acknowledgements

- Did you use other informal analysis to inform your work? Did you get feedback on your work by friends or colleagues? Etc.
- If you had no one give you feedback, it’s okay to say that.

References

- If applicable, report any references you used in your work. For example, you may have used a research paper from X to help guide your analysis. You should cite that work here. If you did all the work on your own, please state this.