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[2], 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

[2] http://movielens.org

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, and find out whether my assumption holds for a more general population (i.e., in this dataset).

The hypotheses that I was trying to investigate was that more complex movies are rated more positively.

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 not 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.

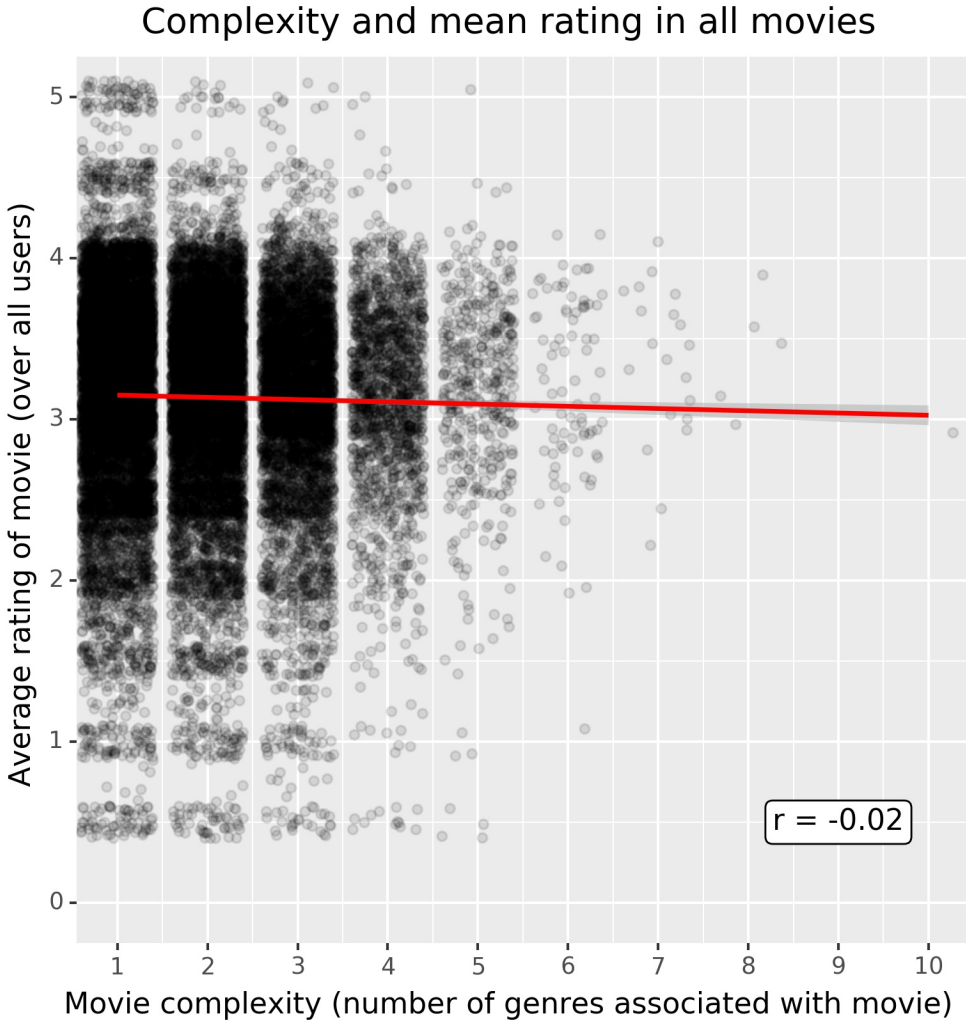


Figure: Movie ratings averaged over all users for each movie, complexity as measured by the number of genres indicated for each movie. Data is jittered to better show spread (random jitter added for ratings: ±0.1; for complexity: ±0.4). Red line depicts a linear regression model fitted to the data in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

Findings

- Some genres show slight correlations of complexity and movie rating (higher complexity associated with higher ratings)
- 6 Genres have correlations greater than 0.10 (see table and figure on the right)
- Highest correlation for the horror genre (r = 0.184)

Genre	n	r
Horror	2611	0.184
Action	3520	0.161
Sci-Fi	1743	0.150
Children	1139	0.128
Thriller	4178	0.118
Mystery	1514	0.108
Animation	1027	0.097
IMAX	196	0.096
Comedy	8374	0.092
Western	676	0.090
Musical	1036	0.059
Adventure	2329	0.057
Romance	4127	0.054
Crime	2939	0.042
Fantasy	1412	0.026
Film-Noir	330	-0.037
Drama	13344	-0.047
Documentary	2471	-0.058
War	1194	-0.063

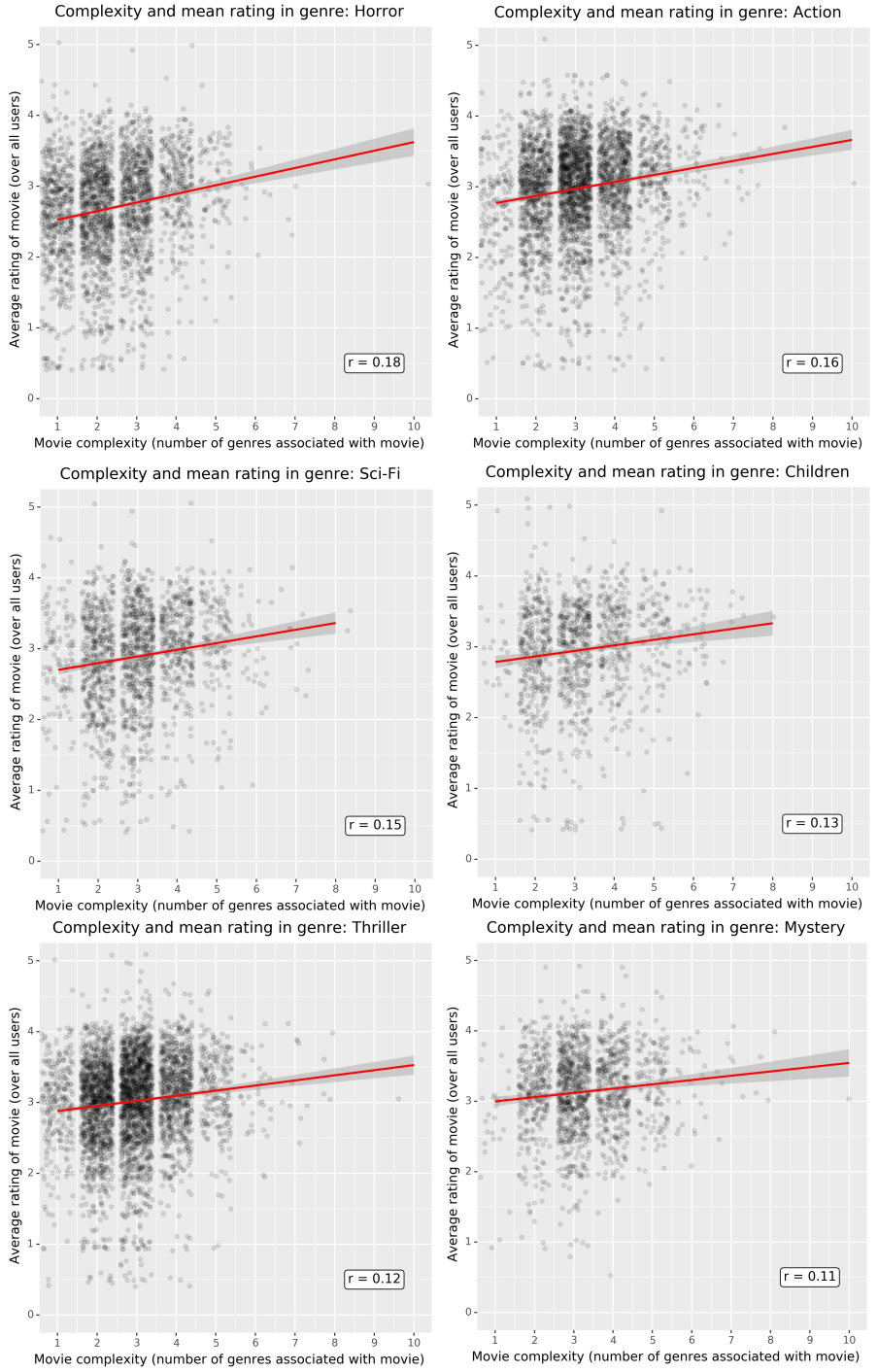
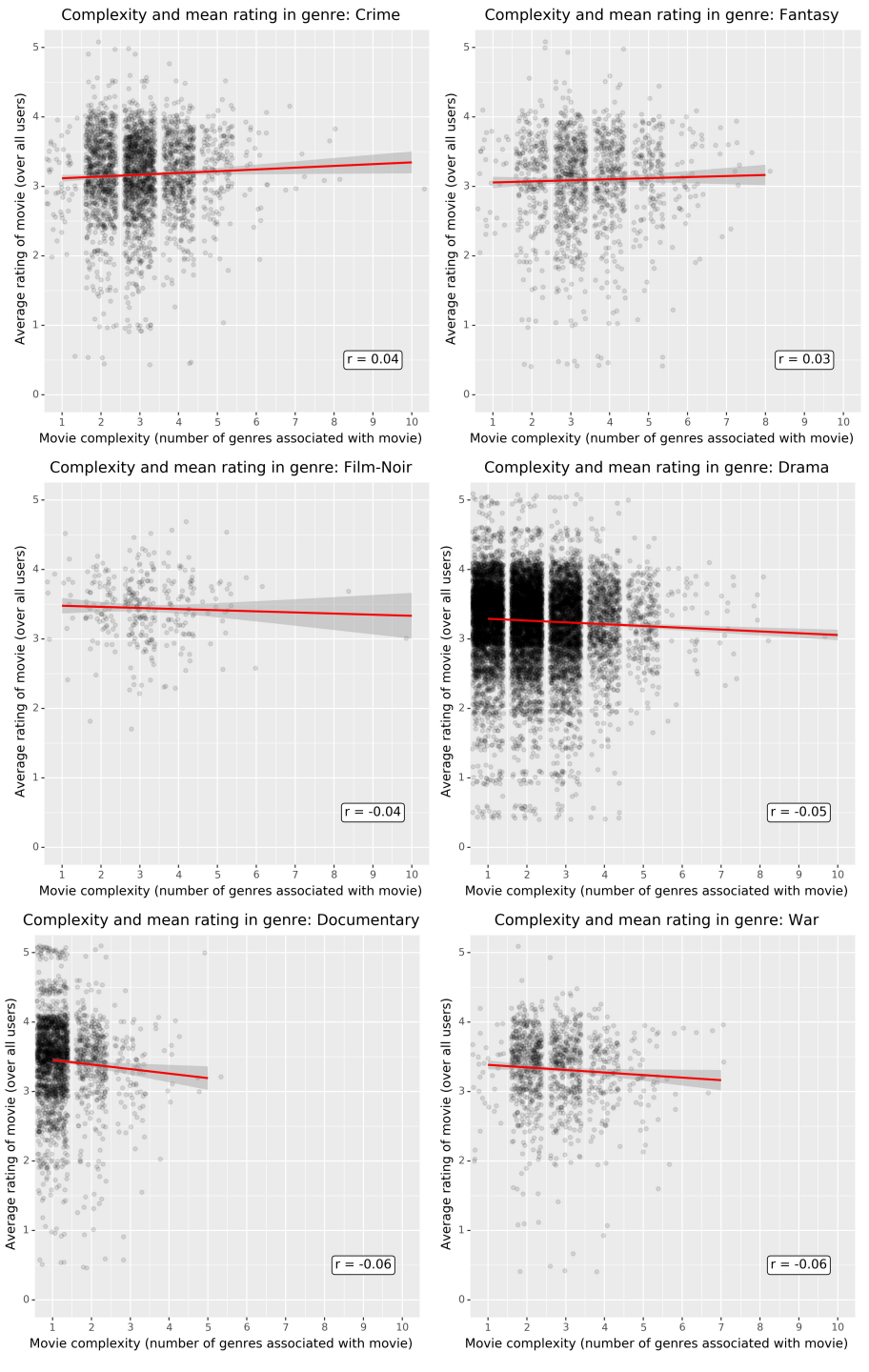


Table: Pearson correlations (r) and sample sizes (n) within each genre subgroup. Each subgroup contains data of any movie that was associated with the respective genre (i.e., multiple plots might contain the same movie, if associated with multiple genres)

Figure: Movie ratings and complexity per genre. Each subplot contains data of any movie that was associated with the respective genre. As in main plot, data is jittered and the red line depicts a linear regression model fitted to the data in the plot. All subplots use the same axis limits for better comparability.

Findings

- Most genres do not display any relevant correlation.
- Of 19 genres:
  - 11 genres have an absolute correlation of 0.10 or lower
    (|r| ≤ 0.10), explaining less than 1% of variance
  - 4 genres have an absolute correlation of 0.05 or lower (|r| ≤ 0.05)
- Some genres even show a negative relationship, even though with a very low correlation. e.g., war movies or documentaries (both r = -0.06, but both with a relatively small spread of complexity)



Findings and Conclusion

- The data does not support the hypothesis for a general relationship of movie complexity and average rating (i.e., over all genres).
- There is an indication for a relationship between variance of ratings and movie complexity. Movies with low complexity have a high spread of ratings (full range), while more complex movies have a lower spread (more towards the scale midpoint or slightly above).
- Only some genres showed a (low) correlation between movie complexity and average ratings (e.g., horror, action, and sci-fi movies).
- In summary, the hypothesis was confirmed only partially.

Appendix

- Full analysis can be found on github: https://github.com/ingonader
  - Notebook containing this analysis (Jupyter notebook): https://github.com/ingonader/python-for-data-science-edx/blob/master/week-06-mini-project/imdb-movie-dataset-analysis-part-notebook.ipynb
  - Python file for full analysis (Jupyter lab): https://github.com/ingonader/python-for-data-science-edx/blob/master/week-06-mini-project/imdb-movie-dataset-analysis.py