CSC 490 Seminar in Computer Science

Yussef Saidi & Kailie Yuan

Data Mining

Software Used: RGui

Real Dataset Used: Official IMDb Datasets

As we searched for interesting data sets to explore, we could not find many free available ones. However, we ended up coming across IMDb's data sets; which consists of all kinds of information about movies and television shows. With that information, we asked ourselves how this data could help us learn something useful. What makes movie more or less successful? How can a producer create a more successful movie? In this context, the word "successful" means success with users and reviewers in the form of user ratings (on scale from 1-10).

Our goal of analyzing IMDb's data sets is to look for correlations between user ratings to other characteristics of a movie. We believe that those factors allow for better choices to be made when trying to maximize your movie's ratings on IMDb.

The IMDB dataset (datasets.imdbws.com) available is broken into 7 different tsv (Tab Separated Values) files:

1. title.akas.tsv : has qualitative information about the film
2. title.basics.tsv : has quantitative information about the film title
3. title.crew.tsv : directors and writers information
4. title.episode.tsv : tv-shows episode information
5. title.principals.tsv : principal cast and crew information
6. title.ratings.tsv : IMDb rating and Number of user ratings
7. name.basics.tsv : has information about all individuals involved in the scene

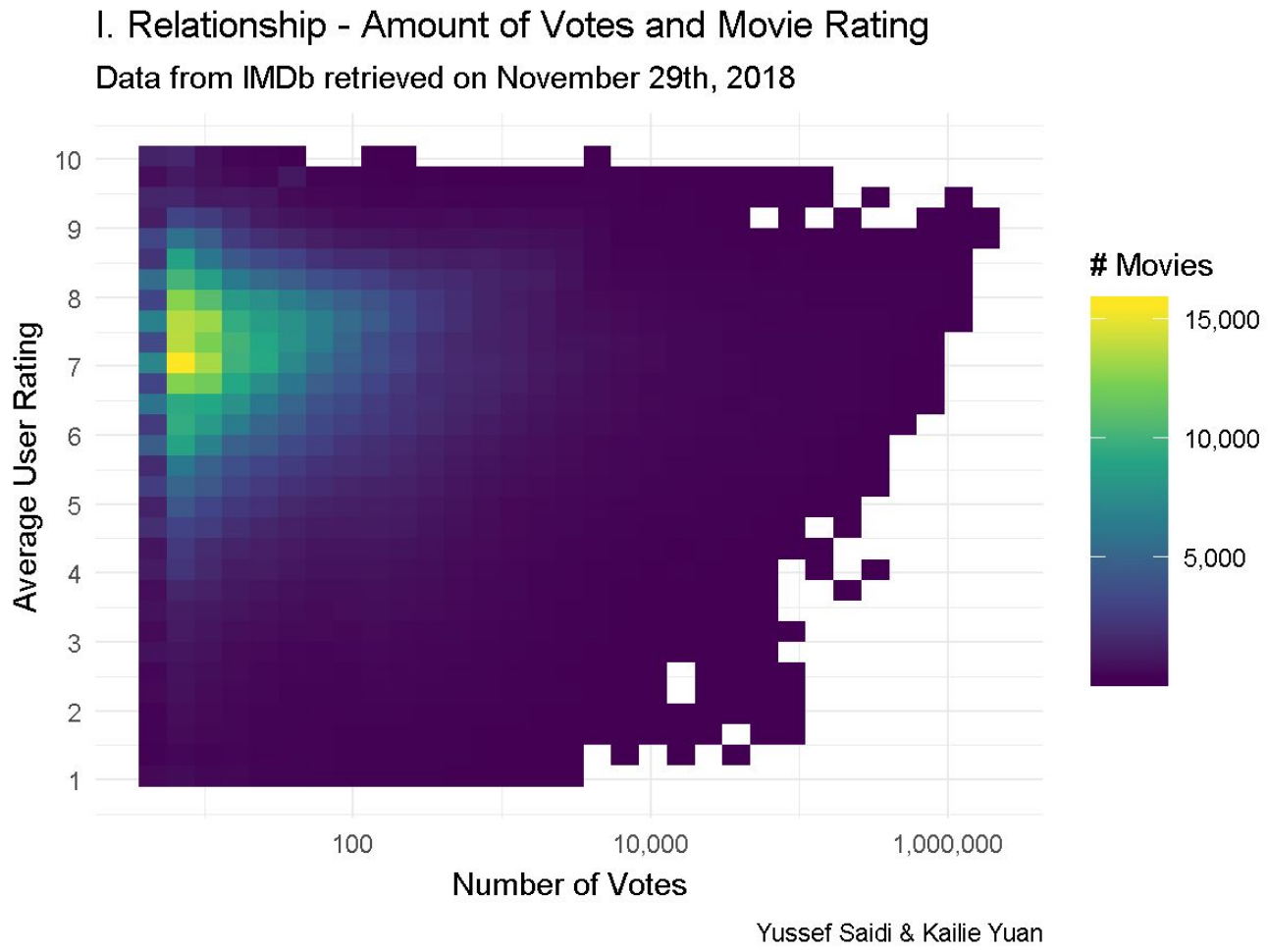
To implement our approach in solving this problem, we will be using RGui and ggplot. R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS (www.r-project.org/). ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. All we need to provide it are the data, how to map variables to aesthetics, and what graphical primitives to use.

In hopes of finding an answer to our problem, we will be analyzing:

1. Average Ratings VS Amount of votes
2. Average Ratings VS Movie Runtime
3. Average Ratings VS Release Year
4. Average Ratings VS Genre

Results and Analysis

The first thing we want to look at is how a movies' ratings are affected by the number of people that vote. By using the numVotes, and the averageRating columns from *title.ratings.tsv*, we were able to create the following graph:

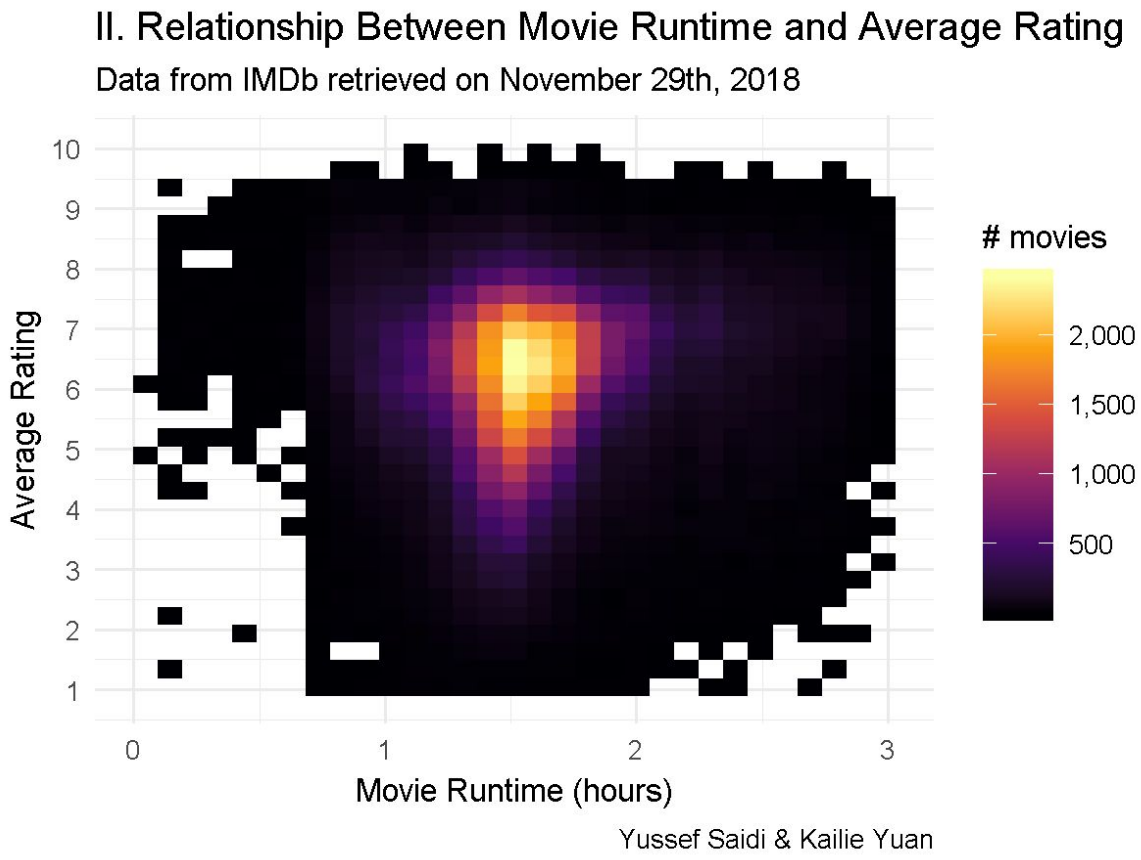


We can observe the following:

- Very few of the movies have a wider range of ratings (0 - 10)
- Majority of the movies have ratings in the range of (5 - 9)
- Majority of the movies movies don't have many votes (about <100)
- Ratings follow the theory of Four Point Scale: ratings tend to be between 6-10 even though the scale offers more options.

We can conclude that the Theory of Four Point Scale can be seen as negative or neutral. It can be seen as negative because majority of the movies usually are made by known studios who have a higher budget and a higher quality. They end up with a slightly higher average rating than the rest. But, it can be neutral because ratings naturally tend to stay a little above average (6/10), even for very average content. It seems that people are more inclined to write a review if they have a positive point of view of a certain movie, which creates more advertisement.

The second experiment is to see if there was any correlation between the movie runtime and its rating. By joining the two files *title.ratings.tsv*, and *title.basics.tsv*, we have movie ratings and their runtime in the same table, we are able to create the following graph:

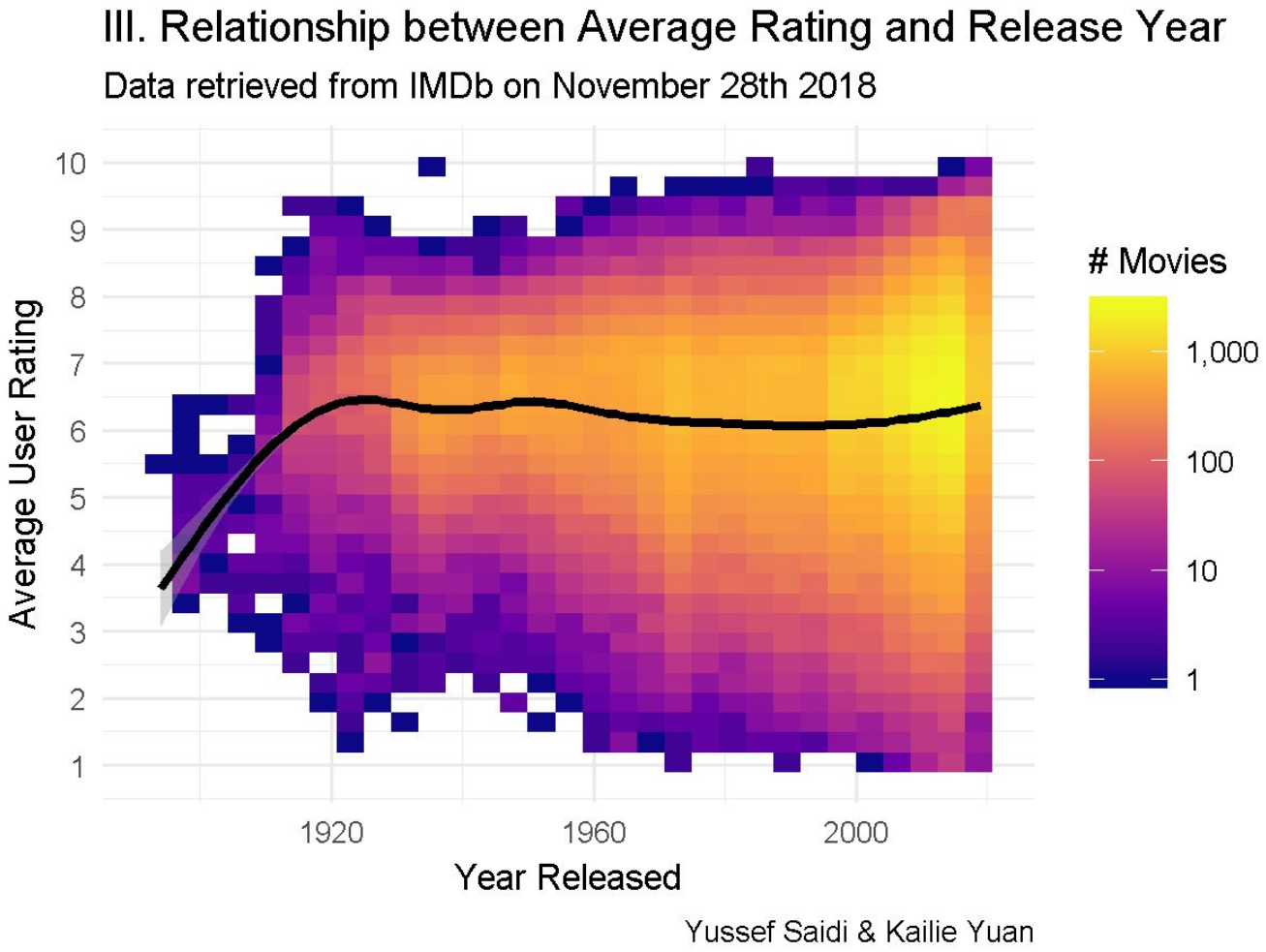


We observe the following:

- There are both long and short movies that are rated well
- Most average rated movies are in the 1.5 hour category

Because there are both movies long and short that are rated well, we cannot conclude that there a real correlation between a movie's runtime and a movie's rating on IMDb.

The third experiment is connected to the second. We want to see proof of how the four point scale came to be. We are expecting that over time, most movie ratings are shifted towards the 6 to 10 range. By joining the two files *title.ratings.tsv*, and *title.basics.tsv*, we have movie ratings and the year they were released in the same table. We are able to create this graph:

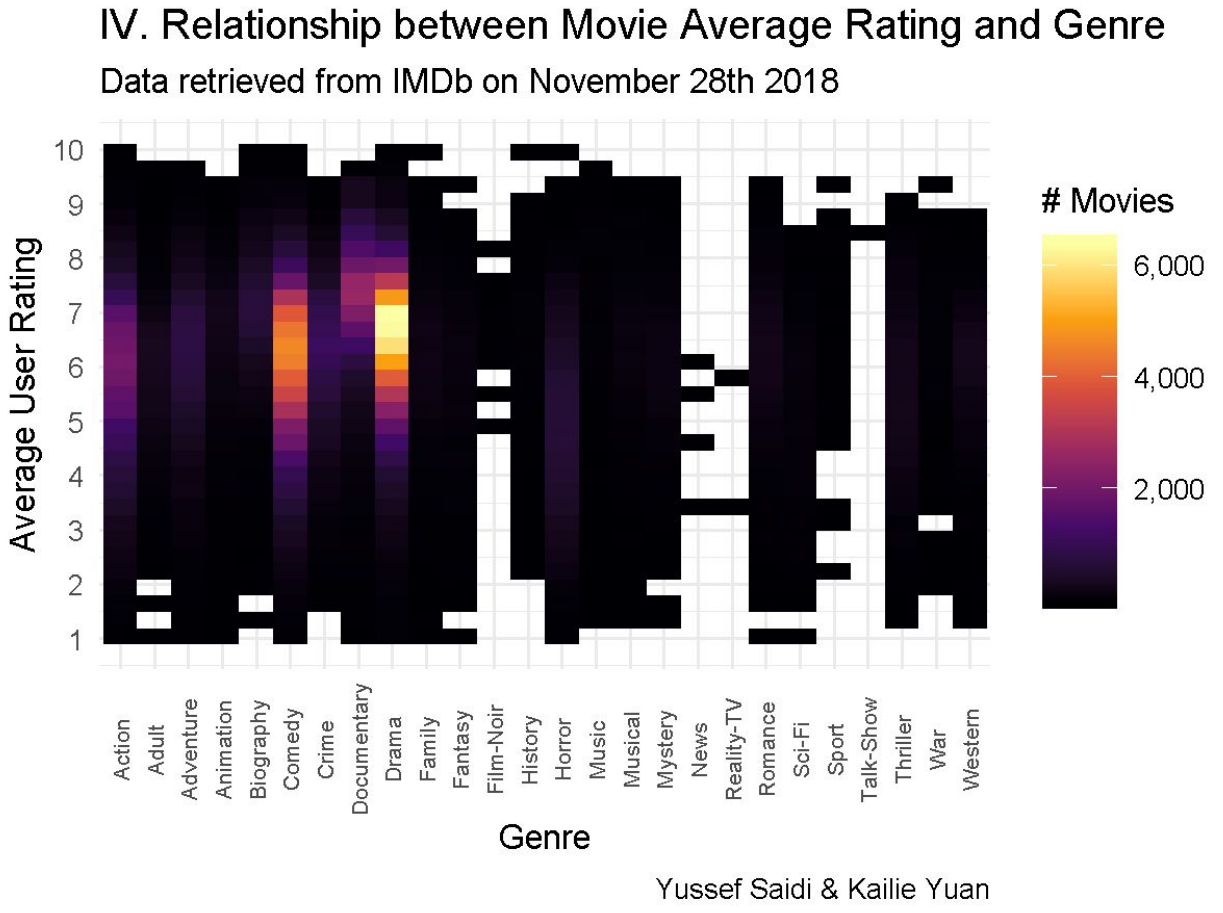


We observe the following:

- Ratings outside the Four Point Scale have increased
- Average ratings, represented by the line, have increased slightly since the 1980s

More movies can have an unusually low, or high rating (1 or 10) which depends on various factors. The average ratings of movies have increased over time which may be because reviewers are becoming more critical, and the quality of movies improving.

The final experiment is to see how the rating on average varies based on the genre of the movie. We have to join *title.ratings.tsv*, and *title.basics.tsv* once more, select only the averageRating, numVotes, and the Genres columns to display. Before we can graph this data, we also have to get the rid of the movies for which we do not have a genre. With that we are able to create this graph:



We observe the following:

- Comedy & Drama are the two genres with many movies rated, followed by action, adventure and horror
- The ratings for these genres of movies confirm the Four Point Rule we mentioned earlier

- Majority of Biographies and Documentaries have higher ratings on average than other genres.
- Majority of the horror movies have the worst average ratings

From these results, we can hypothesize that horror movies don't stand out because they have a very specific recipe, and tropes that most producers stick with. In other words, horror movies can be successful without being good movies because of their genre. This graph tells us that to make a movie with the best rating possible, we should pick a biography or a documentary. It makes sense because they provide a good amount/quality of information to the viewer, which will ultimately satisfy them.

Conclusion:

By using R and ggplot, we were able to manipulate IMDb's official data sets to study several factors that may affects a movie's average rating. It is clear that some have a significant effect on the rating of a movie while others do not.

First, we notice that the number of votes seem to simply restrict the ratings to the Four Point Scale. Albeit pointless, if someone wants to achieve a rating of 10, having a low amount of votes is the easiest way. Then, we see that over the last couple of decades, ratings have shifted outside the Four Point Scale, and the average user rating has been increasing slightly . Last, we observe that the movie genre and duration can affect a movie's rating. When it comes to movie genre, it can affect the difficulty of obtaining very high ratings. That seems to be because viewers have different expectations, depending on the genre of the movie. For movie length, most average rated movies are around 1.5 hours. We believe that it's harder to keep people interested with extra long movies and harder to develop their interest with extra short movies. It is riskier to go for longer, or shorter movies, if you are looking for the highest rating possible for your movie.

If we were producers, we would create a biography or documentary while keeping it around 1.5 hours.

References

https://datasets.imdbws.com/

https://ggplot2.tidyverse.org/index.html

https://www.r-project.org/

https://storyality.wordpress.com/2012/12/17/storyality-20-what-makes-a-film-successful/

https://tvtropes.org/pmwiki/pmwiki.php/Main/FourPointScale