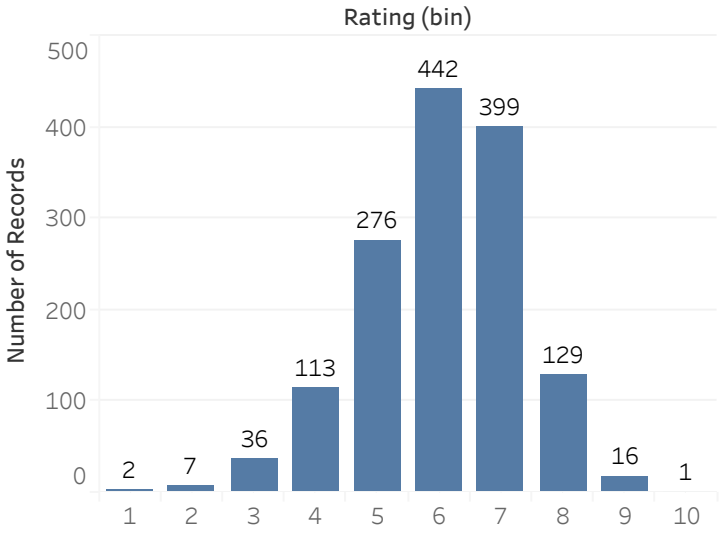
IMDB Exploratory Data Analysis

Interactive versions can be found at:
https://public.tableau.com/views/eda_imdb/D1?:embed=y&:display_count=yes
https://public.tableau.com/views/eda_imdb/D2?:embed=y&:display_count=yes
https://public.tableau.com/views/eda_imdb/D3?:embed=y&:display_count=yes
https://public.tableau.com/views/eda_imdb/D4?:embed=y&:display_count=yes

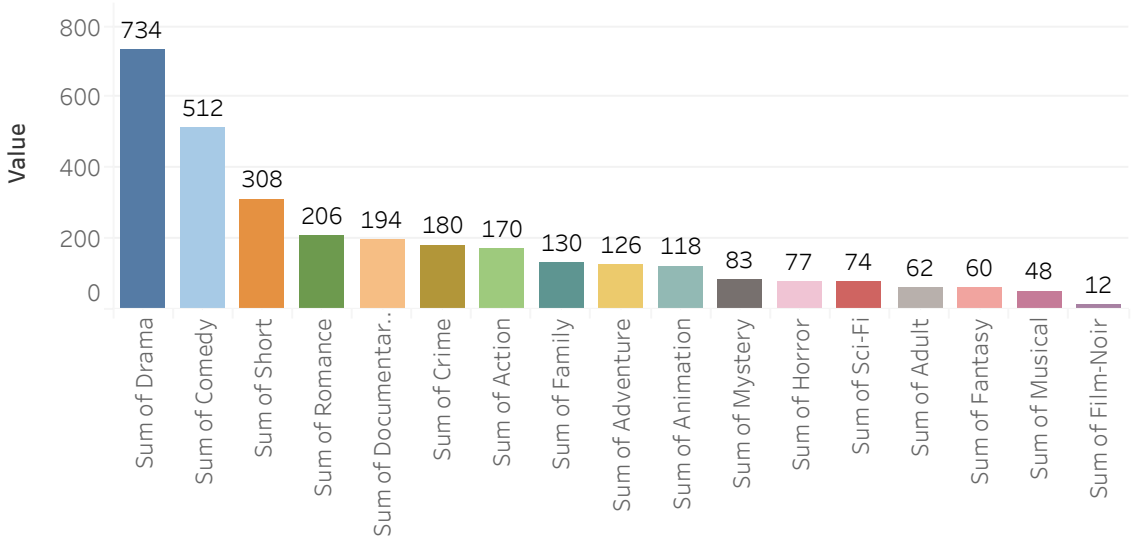
Drama and comedy are the most popular genres.

For the 2000-movie random sample we selected from IMDB, we decided to primarily look at genre by extracting 17 binary columns of genre categories that represent the (non-comprehensive) list of the 17 most commonly found genres as recommended by the IMDBpy documentation. The most popular genre is drama, followed by comedy. Each of these genres is found in over one-quarter of the movies sampled.

Histogram of Ratings



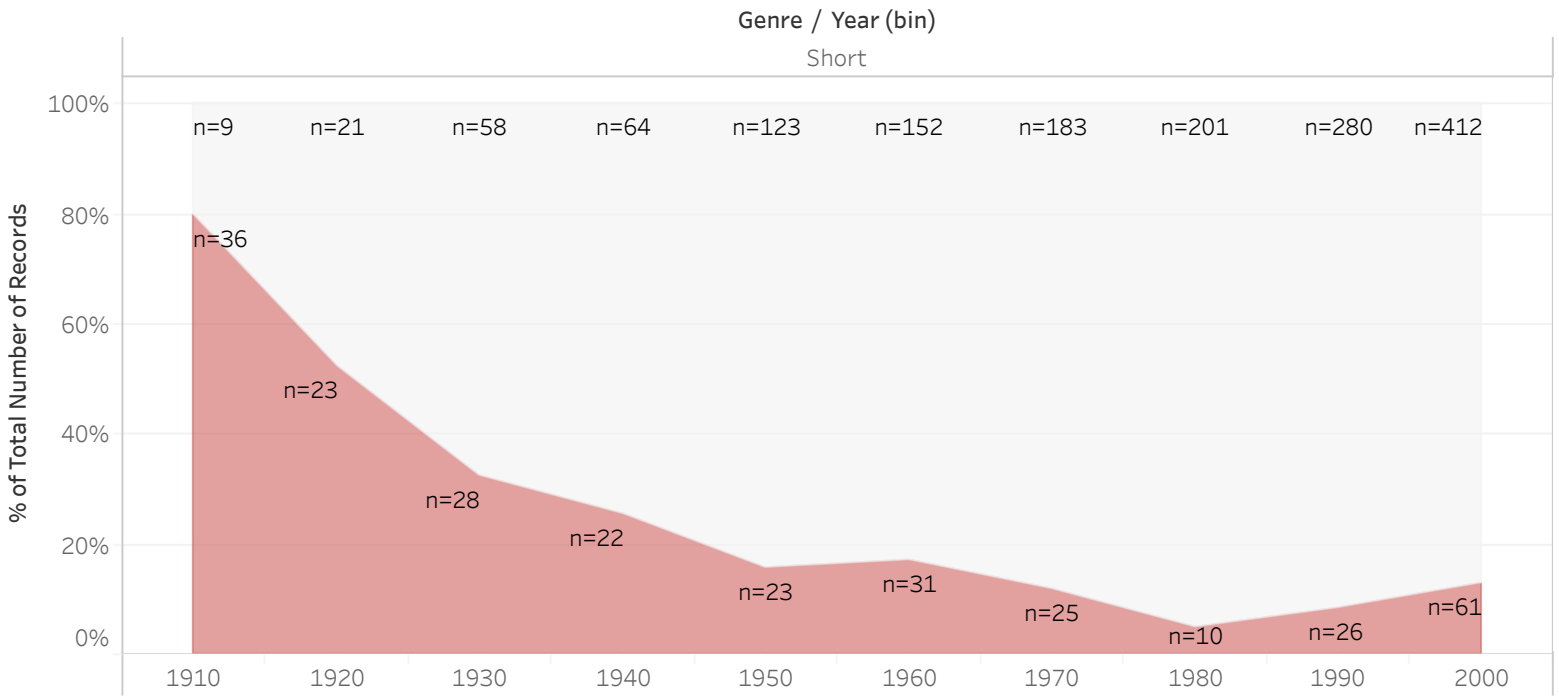
Most popular genres



Interactive: Explore the prevalence of each genre over time

The Short genre makes up a smaller percentage of our movie sample over time, whereas the thriller genre increases in prevalence. Many genres stay relatively constant over time.

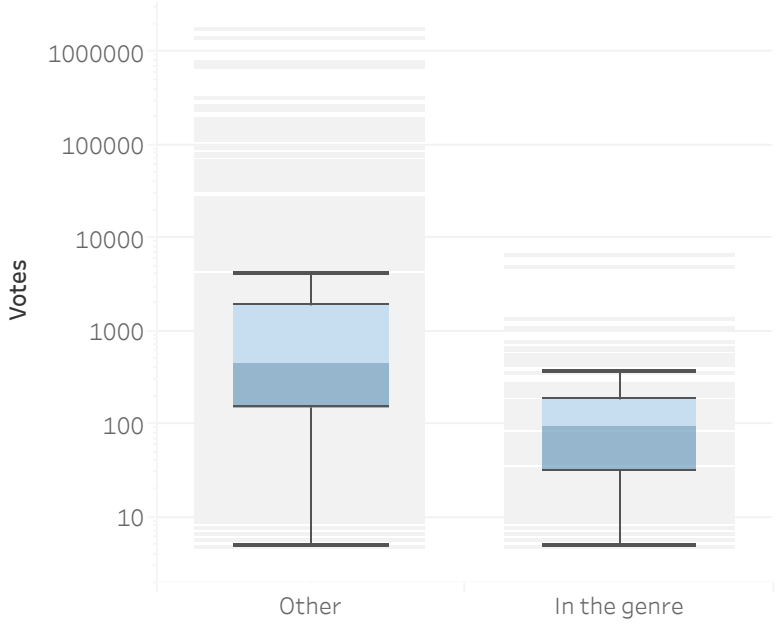
% of movies that fall into the Short genre by decade



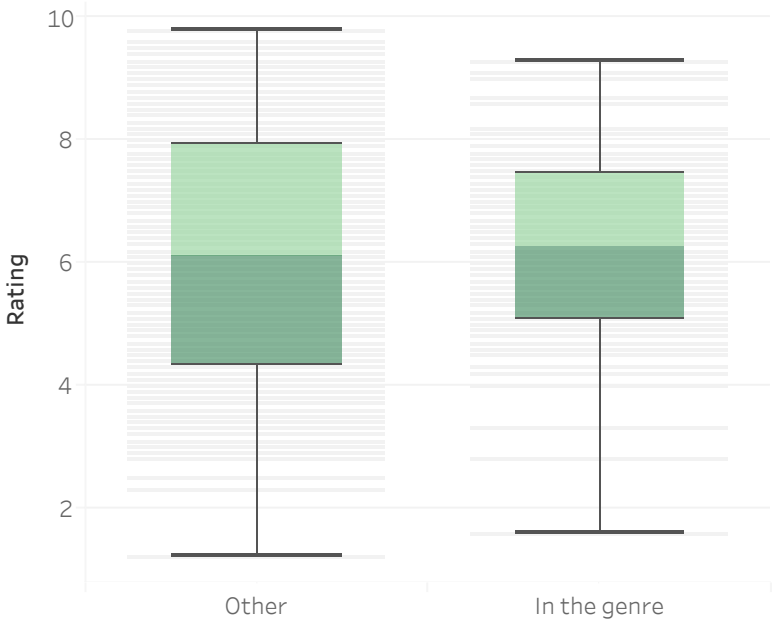
Interactive:
Explore the distribution of ratings and votes for each genre

Many of the negative ratings seem to come from the horror and comedy genres. Some genres (such as Film-noir and documentary) have higher than average ratings.

of votes for Short movies vs others



Avg rating for Short movies vs others



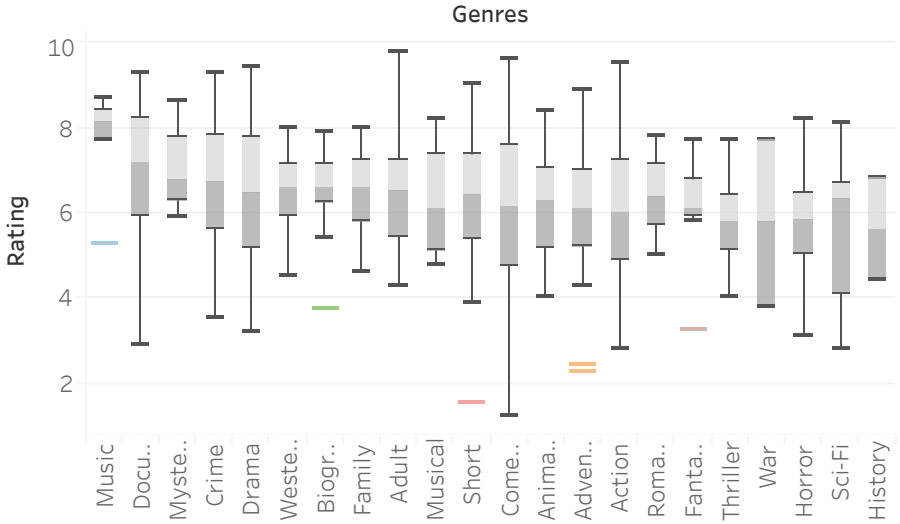
Genre
Short

Exploring the genre that is returned first for each movie

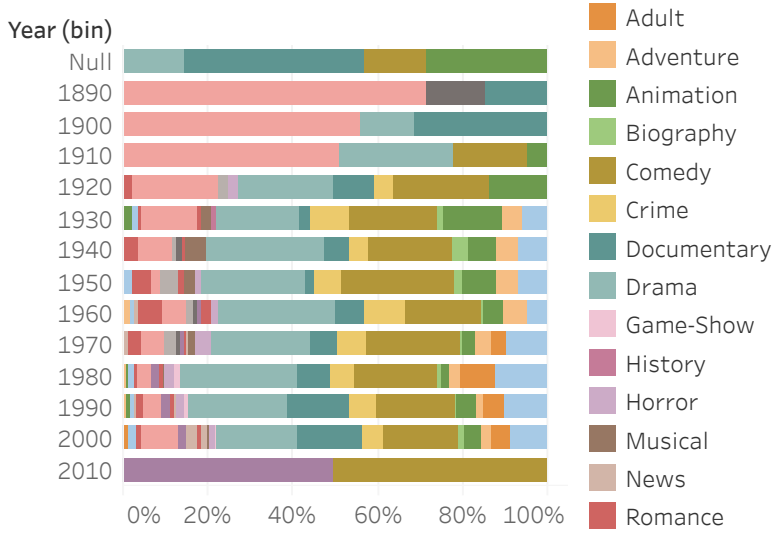
One (very imperfect) way of converting multi-label lists of genres to multi-class mutually exclusive genre categories is by taking the first genre returned by IMDB. I am not totally confident this represents any significant, but it gives us a sense of the data.

We can see that there are some genres here which were not included in our list from IMDBPy. Namely, our list does not include: Talk-Show, News, Game-Show, Reality-TV, History, or Sport. We may want to add some of these such as Sport and History back in for the larger dataset in Milestone 2.

Average Rating by Genre



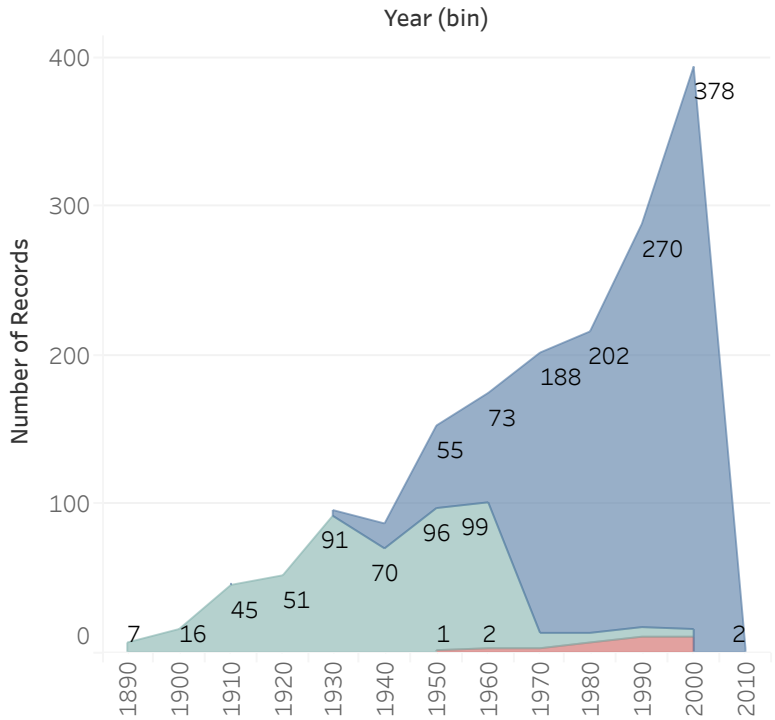
Mix of genres by Year



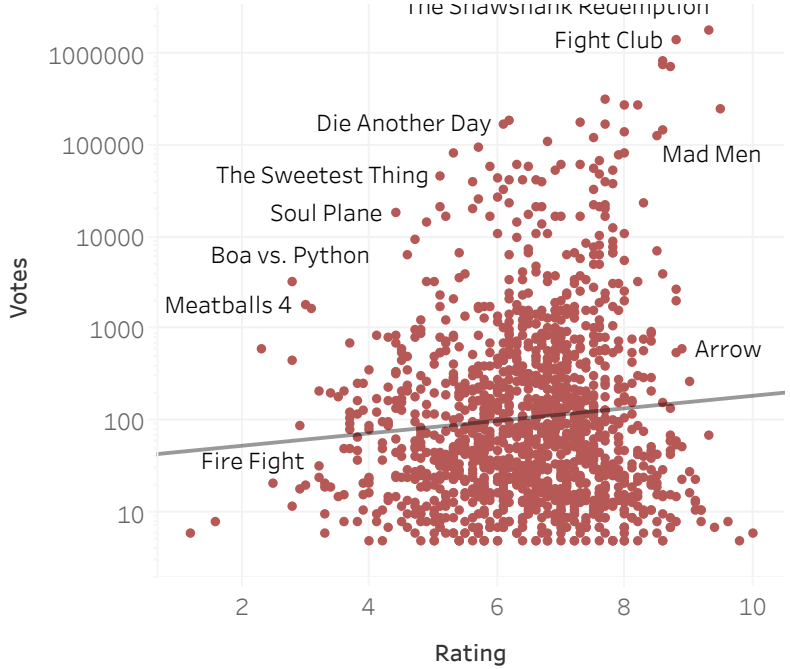
Other correlations

Color Info (group)
- Color
- Black and White
- Both color and b&w

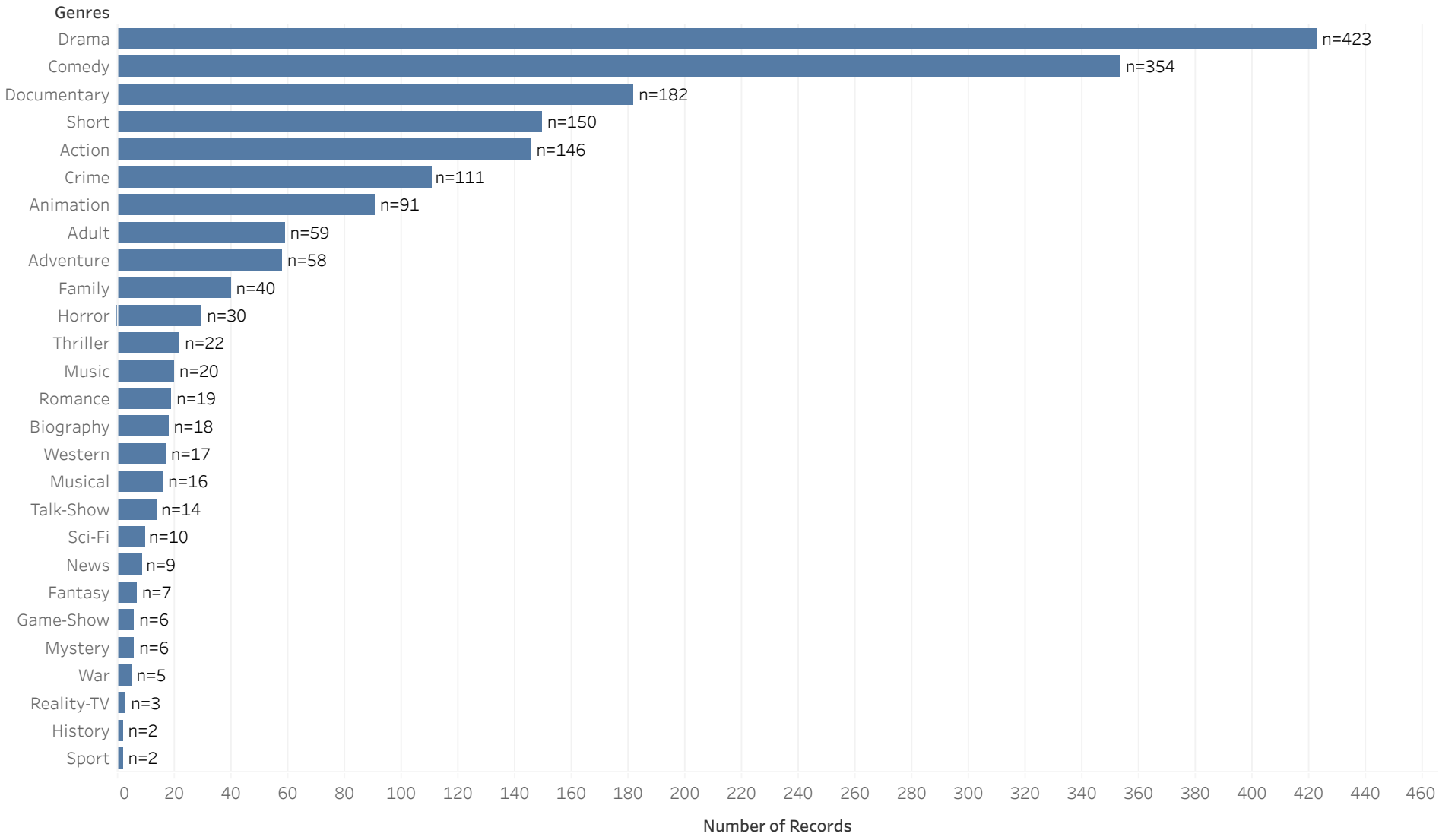
The rise of color movies over time



Only slight correlation between rating and votes

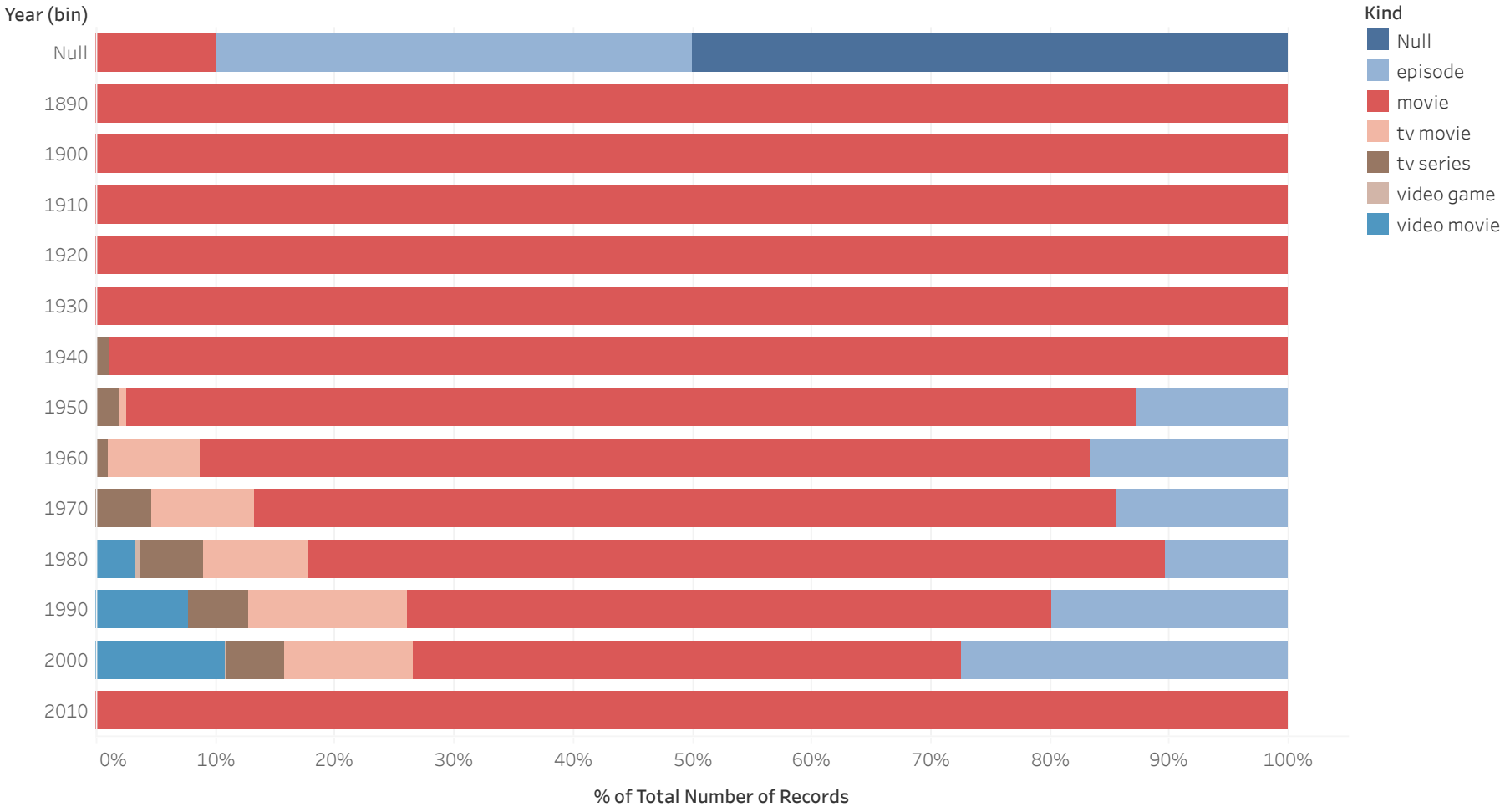


Most popular genres (first listed)

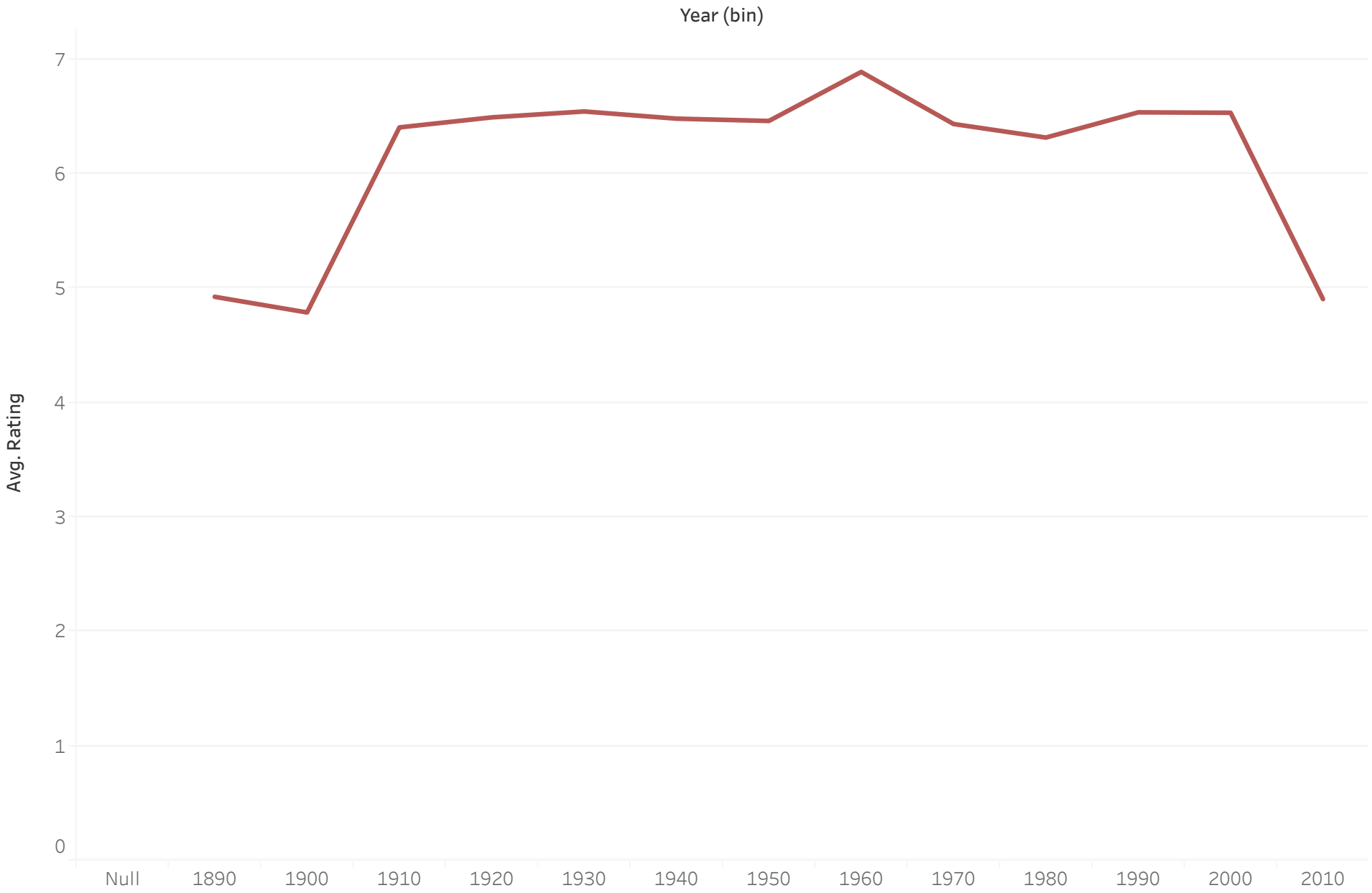


Sum of Number of Records for each Genres. The marks are labeled by sum of Number of Records. The view is filtered on Genres, which keeps 27 of 28 members.

Kind by Decade



Avg rating by Decade



MPAA Rating by Year