Journey to the Center of Movies: An Analysis of What Tracks with Positive Engagement on IMDb

Rebecca Lubera, Amee Yang, James Bryant II, Greg Spahlinger

Introduction

It’s rare to find a person who is unfamiliar with movies, who can’t quote a few memorable lines from a favorite movie, or articulate some cogent opinion of what separates the movies they can quote from the ones they find forgettable. Movies are culturally ubiquitous and commercially important enough that it might feel strange to reflect on the fact that they have only been around for about 100 years.

In their early years, the business of deciding what constituted a good or bad movie largely fell to entertainment writers, and scholarly commentators. At the present time any person with a web browser and an inclination can make their opinions known on a variety of websites. One of these, the Internet Movie Database (IMDb) houses a large collection of movies and allows members to vote on the quality of those films. They also house quantitative figures on the movies, and allow the submission of user reviews.

While people invariably have opinions on what constitutes a good movie, people can often be influenced by factors outside their awareness when forming their opinions. In prior years it might have been very hard to look for those patterns, but the existence of large movie databases like IMDb, which aggregate user engagement data, make this task much more tractable. In order to pursue this objective, we obtained the *IMDb Movies Extensive Dataset* (Leone, 2020) from Kaggle, an online data hosting site. This set contains entries on 85,855 movies from 1894 to 2020. In order to augment the set with financial data we obtained another set from Kaggle – *The Opening Weekend Box Office Performance 50MM+* (Tharmalingam, 2020) dataset. This set is smaller, containing data on 893 movies dating from 1983 to the present, with information on opening weekend revenue and budget. We were especially interested in whether the amount of money spent on making a movie would track with metrics of engagement, and we hoped to do the same thing for some measure of the revenue a movie makes. These questions have implications not just from a cultural interest standpoint, but because they shed light on whether the money spent on large film productions makes good business sense.

Findings – Historical Trends in Engagement and Perception of Quality

The first set of trends we decided to explore is the perception of quality and level of engagement that people on IMDb have with movies from various historical periods. We used the full IMDb database for this task, and the Pandas, Matplotlib, Math and Seaborn libraries in the Python programming language to manipulate our data.

Throughout this paper we will use the average score a movie gets from users – “average vote” or “average user score” as a metric to quantify the level of merit or enjoyment that the audience feels for a movie. While this metric is the best we can do, it is important to understand that such a measurement does not have a straightforward interpretation. The vote a person makes on IMDb is the product of an opinion, which could be influenced by factors including, but not limited to, their level of enjoyment of the movie, how artistically meritorious they think that movie is, their feelings about personnel associated with the movie, or their level of interest in subject matter presented in the movie.

In contrast, we will use the total number of votes as a metric for engagement. This is the number of people who have given a score to a particular movie, and can be understood as the number of people who cared enough about a movie to find it on IMDb and rate it.

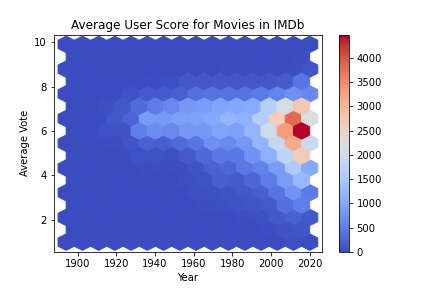
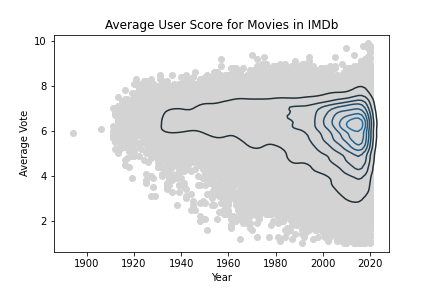


Figure 1 – The Average Score Distribution of Movies on IMDb Broadens Downward Over Time.

The first figure presented here (Fig. 1) shows the 85,855 movies in IMDb plotted by their user score and year of release. Because this many movies cannot be sensibly displayed on a scatter plot, we used a kernel density estimate contour plot from the Seaborn library to show a contoured distribution of points (left) and a hexbin plot from Pandas which shows the quantity of movies in a hexagonal area (right).

Several conclusions can be drawn from these plots. The first and most obvious is that throughout time the central tendency of the score data doesn’t change all that much. From the 1920s onward, most of the movies seen here are getting average scores somewhere between a 5 and a 7. The most obvious change that can be observed is the broadening of the scores into the lower reaches of the scale. In the earliest movies in the database (those released prior to 1930) there are practically no cases of movies with average scores lower than 4, but by 2000 these became fairly common, and they got more common as the 21st century progressed. There are a number of interpretations we could make for this trend, but all of them are speculative. It could be that as the movie industry progressed, more movies are being released that are considered “bad” (boring, tasteless, offensive, etc.) or it could be that the bad ones from the early days have been forgotten, and are not included in the database.

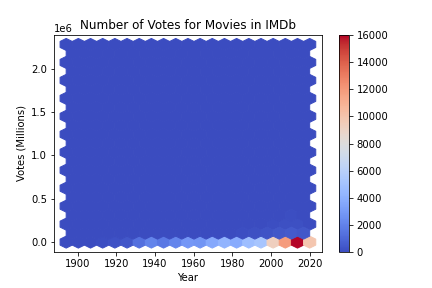
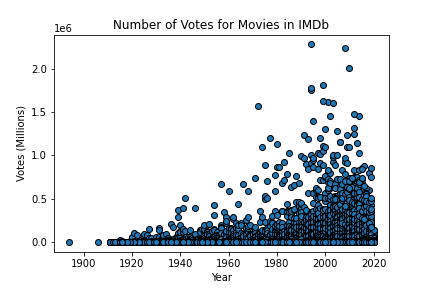


Figure 2 – The Number of Votes for an Average Movie is Consistently Low - With Significant Outliers.

In contrast to the data on average score over the years, the data on the number of votes movies are getting is quite a bit clearer. The hexbin plot in figure 2 has a y axis formatted in the millions, but it’s clear from the plot that almost all the data lined up on the bottom of the plot. This oddity is clarified by the scatter plot to the left of it. For the entire 20th and 21st centuries, the median number of votes for a movie in IMDb has stayed between 100 and 700 votes, with some oscillation (data not shown). It’s clear, though, that as time has passed, a significant number of outliers have appeared, which dwarf the level of engagement of the average movie by orders of magnitude. Sorting on votes will immediately clarify what sort of move we are talking about – the highest vote total of any movie in IMDb is The Shawshank Redemption (whose title is confusingly translated into Italian in the set) and others in the top 10 include titles like Pulp Fiction, and Forrest Gump. These are the films that you most likely think of when you hear the term “movies” – and it is important to stress that they are outliers with respect to the dataset as a whole.

Findings – Genre Appears to Correlate with Average User Score

We tried several grouping methods and scatter plots to look for correlations with average user score, and convincing relationships were somewhat hard to come by. Genre seemed to be something of an exception to this.

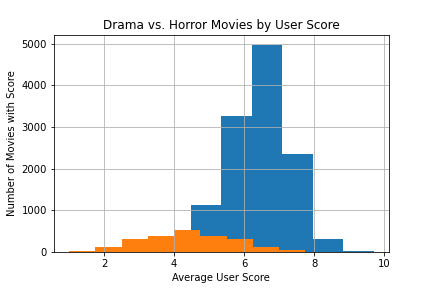
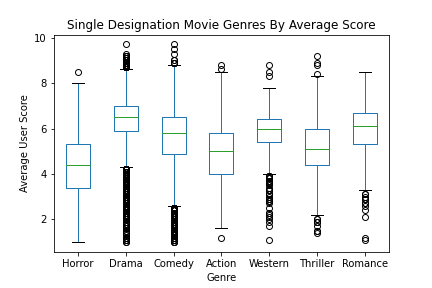


Figure 3 – Audiences are More Likely to Give High Marks to Drama, Less Likely to Praise Horror.

When we originally sorted on genre, the results were somewhat confusing, as movies in IMDb can have multiple genre designations. This led to our finding 1257 genre categories the first time we sorted. Because the genre designations in IMDb are combinations of single word identifiers, it was decided that the best way to look at genre was to single out movies that were given only a single genre designation, and compare those. Of these singly designated movies, the drama films were most numerous (12543 films) and had the highest median score (6.5). The least numerous single category we looked at was romance, which contained 439 films. The lowest rated genre was horror, with a median score of 4.4. Figure 3 shows a comparison of these categories using a boxplot, which shows some notable differences. The histograms on the right show the difference in the median, and abundance of drama films (blue) and horror films (orange).

In order to establish the statistical significance of the differences between the groups seen in figure 3 we implemented several hypothesis tests, to establish that the perceived differences are not the result of chance. A one-way ANOVA test was performed between the genres seen in figure 3, which returned a p-value of 0.0. This seemed strange and potentially untrustworthy so we followed up by running Welch’s t-tests between drama and horror, and drama and comedy. The two tests also returned p-values of 0.0. In order to further confirm the merit of this score a Welch’s t-test was implemented from scratch using functions from the Numpy and Math libraries, comparing the comedy and drama values. This test replicated the t statistic returned by the Scipy function of approximately -46.00, to four significant figures. The degrees of freedom for the set were further calculated to be approximately 13501. A web-based calculator at Social Science Statistics (n.d.) estimated the p-value to be less than 1.0 x 10-5.  We conclude that the cause of the low p-value is the large sample sizes we are working with. It seems safe to conclude that all of the differences shown in figure 3 are in fact statistically significant.

Findings – Correlation Between Total Votes or Average User Score and Money is Ambiguous in Big Budget Films

To look at how the revenue and budget correlate with average user score and engagement we found a dataset that contained opening weekend revenue, and budget for 454 movies. Because this set did not contain an IMDb id value, we were forced to merge these sets on the title, which led to some problems. The first thing that tipped us off to an issue was an entry that clearly showed that data from the 1983 film Ghostbusters had been merged with its 2016 reboot of the same name. Since both datasets contained a year column, we were able to filter out problematic instances by setting a filter to remove rows where the year values were not equal. An additional row was dropped for having a null value in the budget column. This left us with 304 movies in our dataset.

In order to understand which movies we are looking at for this part of the analysis, a couple of descriptive statistics are instructive. The median number of votes received on IMDb for movies in our combined dataset was 240,059.5 and the median budget is 75 million dollars. These figures alone, make it clear that these are the same outliers with respect to audience engagement that we found in the historical analysis. Moreover, the fact that the average budget of these films is so high provides a plausible explanation for why they are outliers. While we do not have stats on budget for the full IMDb set, it seems highly unlikely that the median budget would be on the order of millions or tens of millions of dollars. It’s important to keep this in mind, as we share these findings: they are specific to a small set of high-profile outliers, and almost certainly not applicable to movies as a whole.

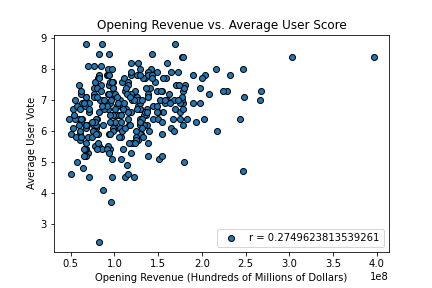
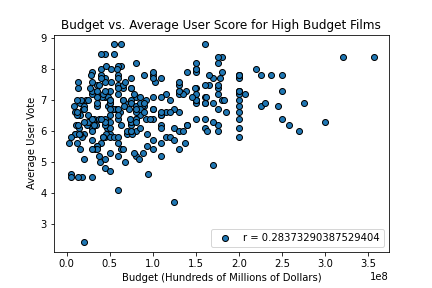


Figure 4 – Budget and Opening Revenue Show Weak Correlation with Average User Score in Big Budget Films

Budget and opening weekend revenue were scatter plotted along with average user vote (figure 4) and total votes (figure 5). For each scatter plot a Pearson r value was calculated to measure correlation between the variables. These appear in the legends of the figures. In figure 4 we can see that budget and opening revenue are weakly correlated to average user vote (roughly 0.28 and 0.27 respectively). Reading a little more into the data, we can see that these plots are vaguely wedge shaped. It appears that the more money is spent on a movie, or the more that it makes on opening weekend, the less likely it is to be panned by its audience. The opposite does not appear to be true though, as there are a large number of points that show lower budget films, and lower earning films getting high average scores. Budget and opening weekend revenue appear correlated with the total number of votes to a slightly higher degree (0.39 and 0.38 respectively) but still not all that strongly.

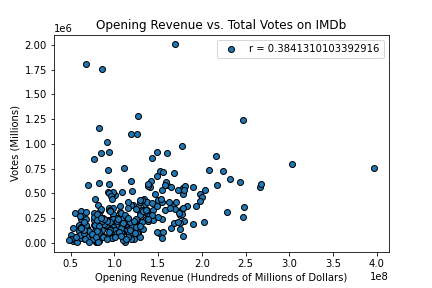
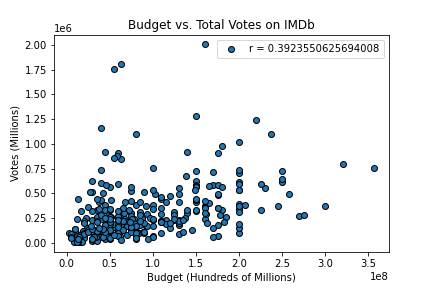


Figure 5 – Budget and Opening Revenue Show a Slightly Stronger Correlation with Total Votes in Big Budget Films

Conclusions

In conclusion, several interesting things have come out of our analysis. We were able to see from our large set scatter plots, that the central tendency of movie scores has not changed that much over time, but that IMDb users are more likely to give bad ratings to newer films. We also showed that the vast majority of films on IMDb get a relatively small amount of attention from IMDb users, with big budget Hollywood films appearing to constitute a substantial outlier population. While most metrics we looked at did not seem to relate to the average IMDb user score, genre did show a correlation, with users being more likely to rate drama films and pan horror flicks. Finally, it appears that neither user score nor votes correlate very strongly with either the budget of the film or the amount of money it makes on opening weekend. This is stated with the substantial caveat that having a large budget seems likely to determine which films will become outliers with respect to vote count, and we believe that our analysis only shows that beyond a certain point the budget ceases to be a major driving force for engagement.

References

Leone, S (2020) *IMDb Movies Extensive Dataset.* Kaggle,

<https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>

Social Sciences Statistics (n.d.) *P Value from T Score Calculator.*

<https://www.socscistatistics.com/pvalues/tdistribution.aspx>

Tharmalingam, N (2020) *The Opening Weekend Box Office Performance 50MM+*. Kaggle,

<https://www.kaggle.com/narmelan/opening-weekend-box-office-performance-50mm>