IMDB Movies Analysis - Factors That Contribute to Success

Andrew Fiegleman, Arisa Nguyen, Ethan Duncan (Group 2)

12/6/2021

Overview

For this analysis, we wanted to analyze the attributes of successful movies. We found a dataset that scrapes all movies in IMDB with at least 100 ratings as of 1/1/2020 and contains various information about each movie. Next, we defined successful movies as movies with high IMDB ratings and profitability. The specific factors analyzed in relation to movie success were budget, seasonality (when it was released), and genre. In addition to the IMDB dataset, a Consumer Price Index (CPI) dataset was used to calculate inflation and The Movie Database website was scraped to add additional context to the IMDB dataset.

- IMDB dataset[1]
- Consumer Price Index[2]
- The Movie Database (API)[3]

Research Questions

How does the budget, seasonality, and genre of the movie correlate with its IMDB rating and profitability? How has it changed over time?

General Data Cleansing

We first decided to limit our dataset to movies that were released in the USA, because we have a deeper understanding of American movie culture and history compared with other cultures. For this reason, we also analyzed its gross profit in the US only (rather than worldwide). These numeric aspects of the dataset were cleaned by stripping any characters that could not be converted to a numeric (such as the $ symbol). It was then converted from string to numeric so that the data could be analyzed.

[1] Stefano Leone, “IMDB Movies Extensive Dataset,” Kaggle, September 14, 2020, https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb%2Bratings.csv.
[2] “Consumer Price Index for All Urban Consumers: All Items in U.S. City Average.” FRED. Federal Reserve Bank of St. Louis, November 10, 2021. https://fred.stlouisfed.org/series/CPIAUCSL.
[3] “API Overview.” The Movie Database (TMDB). Accessed December 8, 2021. https://www.themoviedb.org/documentation/api.

Next, we made this decision to use the weighted average rating instead of the average rating. This is done in order to ensure the most accurate rating of each movie. According to the IMDb website[4], the IMDb accepts and considers all votes received by IMDb users, although not all votes have the same impact (or 'weight') on the final rating. This is mainly due to discourage unfair voting practices which, when detected, an alternate weighting calculation may be applied in order to preserve the reliability of the IMBd system.

To account for inflation, we adjusted the budget and profit to be represented in 2021 dollars. To do this, we joined a dataset that contains the Consumer Price Index (CPI) of every year to the IMDB movies dataset on the release year of the movie, then calculated the inflation rate (see figure A.1 in appendix).

We spot checked our adjusted-for-inflation budget and gross profit columns to ensure ours by using various other sources to adjust a movie's budget to 2021 dollars. Fortunately, we found results similar to ours, meaning that our formula and CPI dataset was working correctly.

Unfortunately, the IMDB movies dataset was not sufficient for us to complete our analysis, because its genre column contained multiple genres that the movie fell in. This would make our analysis difficult and complicated. Our solution was to scrape The Movie Database website (another database separate from IMDB), because The Movie Database noted a main genre that the movie fell into. The scraped dataset was joined to the IMDB dataset on IMDB ID.

Budget

The budget of the movie was hypothesized to be correlated with movie success, because a movie with a large budget has more ability to increase its production value and be an overall well-made movie. To perform analyses on the budget of the movie, the dataset was limited to movies released in the year 2000 and onward. This was done for two reasons: (1) because the IMDB website was fully in-use by the 2000's and (2) we have a better understanding of movie culture in the last 20 years than the last 50 or so years. We also limited the dataset to movies with budgets of $300,000 or more, because our literature research claimed that is the approximate lower limit for proper movies (that are to be seen by the public, rather than student films or projects).[5] Movies that were missing values for USA gross income, budget, IMDB rating, and vote count were also dropped. After these filters, there were 3937 movies left in the dataset.

Our preliminary findings show there is strong positive correlation between budget and profit, weak positive correlation between budget, and no correlation between budget and return on investment (ROI), per figure 1.

Average Values and Count of Movies by Budget Bracket

Budget (2021 dollars $)	Average Profit	Average Return on Investment	Average IMDB Rating	Count of Movies

[4] IMDb. IMDb.com. Accessed December 8, 2021. https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK?ref_=helpsect_pro_2_8#.

[5] Caroline Brophy, "8 Levels of Film Budgets and Rates," The Film Fund, May 29, 2020, https://www.thefilmfund.co/8-levels-of-film-budgets-and-rates/.

Up to $50,000,000	$8,257,181	71%	6.1	2736
Up to $100,000,000	$11,758,130	16%	6.2	644
Up to $150,000,000	$11,384,727	9%	6.3	280
Up to $200,000,000	$26,337,512	15%	6.7	170
Up to $250,000,000	$56,599,165	26%	6.8	73
Up to $300,000,000	$129,771,497	47%	7.1	25
Up to $350,000,000	$119,028,058	35%	7.2	6
Up to $400,000,000	$145,417,970	40%	7.2	3

Figure 1: Summary of average values and count of movies by budget brackets. Preliminary insight into how budget affects movie success.

Budget and IMDB Rating

The first set of attributes analyzed were the movie's budget and IMDB rating, weighted by its number of votes. Weighting the movies ensures that relatively unknown movies did not influence the data as much as more known movies, which more likely have more votes. For example, a movie with 10,000 votes would be given more weight than a movie with 100 votes.

From this analysis, we found a positive but weak correlation between the movie's budget and IMDB rating with a R-Value of 0.16 and a nearly horizontal linear regression trendline. The weak correlation indicates that budget might not affect the IMDB rating of the movie, and the movies of all types of IMDB ratings can be found across all budget sizes. A movie with a high budget may not yield a high IMDB rating.

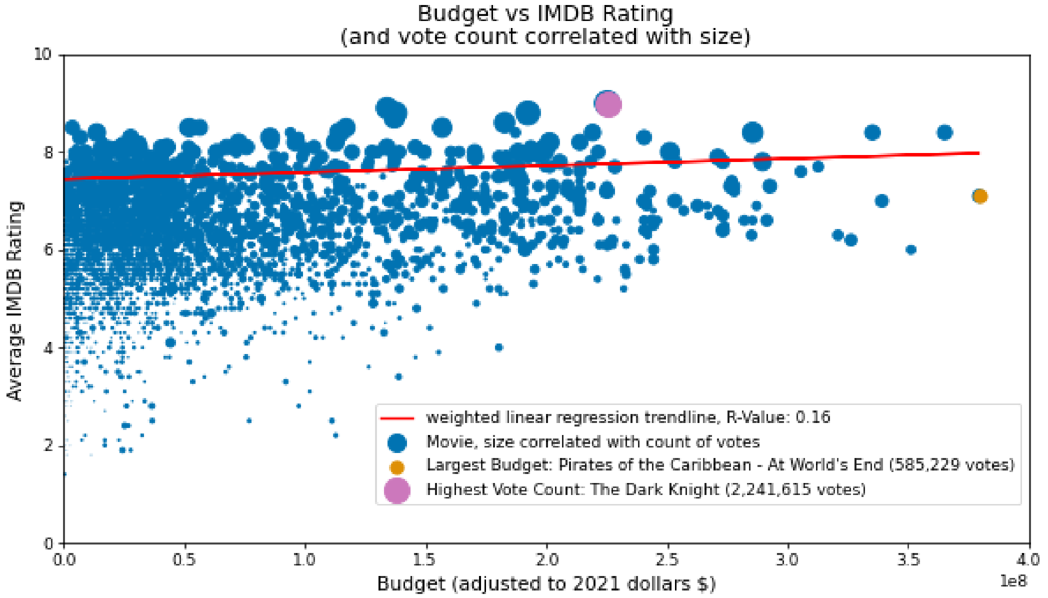


Figure 2: Shows movies plotted by their budget (represented in 2021 dollars) and IMDB rating. The size of each movie's marker correlates with the number of votes or ratings the movie received in IMDB (the movie with the highest vote count has the largest marker). The linear regression trendline was weighted by the number of votes the movie received. Notable movies are marked in unique colors.

Budget and Profit

By analyzing budget and profit similarly, it was shown that there was a strong positive correlation between the two attributes. For this analysis, profit was calculated by subtracting the budget from gross US income. The weighted linear regression trendline is clearly sloped positively and has an R-value of 0.32. A movie that has a higher budget is more likely to have a higher profit. This supports the initial hypothesis that budget affects a movie's success.

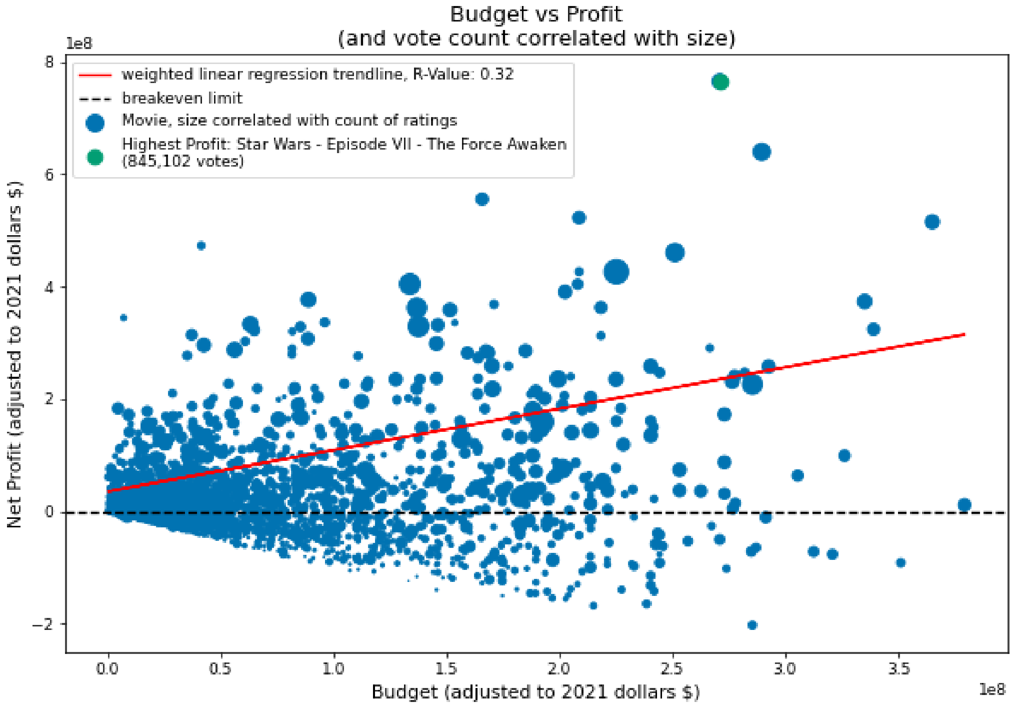


Figure 3: Shows movies plotted by their budget (represented in 2021 dollars) and net US profit (also represented in 2021 dollars). The size of each movie's marker correlates with the number of votes or ratings the movie received in IMDB (the movie with the highest vote count has the largest marker). The linear regression trendline was weighted by the number of votes the movie received. Notable movies are marked in unique colors.

Return on Investment/ Budget

Having high profit on a movie is beneficial for filmmakers, and the previous section shows that having a high budget is correlated with it. However, it is also important to have high Return on Investment (ROI) to be considered a profitable, successful movie. ROI was calculated using the formula in appendix figure A.2, with budget as the cost and net gross income as the return.

According to figure 4, ROI increased with IMDB rating. This indicates that despite profitability and budget being strongly correlated (figure 2), movies can be profitable relative to their budget if they are well received. This makes logical sense because a film that is profitable is also typically well liked by the audience. The outlier in the 1- 2 ratings range was "Saving Christmas" with a rating of 1.4, budget of $553,846 (2021 dollars), income of $3,083,784 (2021 dollars), and ROI of 457%.

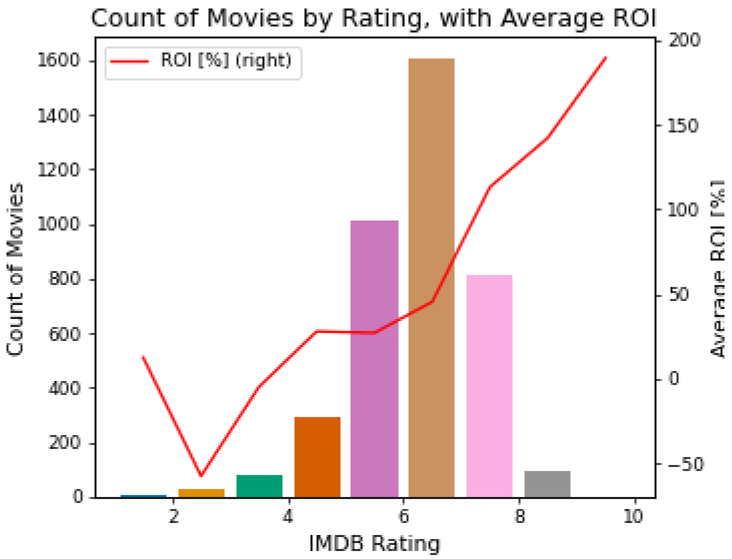


Figure 4: Plot of IMDB rating (x axis) against count of movies (left y axis) and average ROI (right y axis). There were no movies in the 9 - 10 ratings range after filtering the dataset.

Seasonality

For seasonality, we were curious about whether a titles' release date (i.e. release month) correlates with that movies' success. To do this, we first had to adjust our data to include data from The Movie Database (TMDB), as we found that our IMDB database had a film's publish date, but not its actual release date. We scraped this data via TMDB's api to collect their 700K plus titles, which we then joined back to our IMDB dataset via the imdb_id field which both dataset had. In this analysis we looked at how a film's release month (averaged over every year) correlates with that film's U.S. total box office (adjusted for inflation) and it's weighted imdb score. Furthermore, we analyzed how seasonality in the film industry has changed over time in terms of its relationship with box office.

Seasonality and Box Office

The first insight we pulled out of our Seasonality study is that movies in the summer months and around the holidays (Nov. / Dec.) appear to have the highest average U.S. box office, with June exhibiting the highest average. In fact, June and December alone make up more than 28% of all U.S. box office on an aggregate basis. Furthermore, January appears to be the worst month to release a film in relation to U.S. box office, historically. This confirms the industry's notion of 'Dump Months' in that January, August, and September are traditionally where film studios bury their film projects they expect to flop.[6] We will provide evidence on whether this trend is consistent year over year or only on an aggregate basis.

In the figure below, we also see the count of movies for each month. It's interesting to note that some of the months exhibiting the highest average box office actually had the lowest count of distinct movies being released in them. One cause of this could be that movies which are released with less competition will perform better financially. On the other hand, it may be that since studios have historically planned to release their major blockbuster films in the summer and holiday months, smaller films aim for off months (January, September, October) to release their films, as to not compete with these 'Tent-Pole' films.

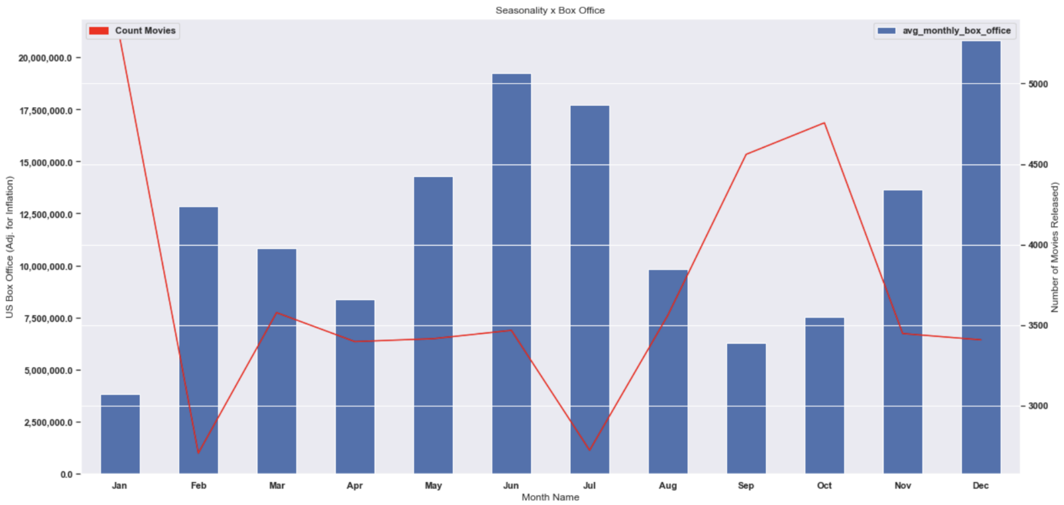


Figure 5: Plot of Month (x axis) against average box office (left y axis) and count of movies (right y axis). Averages all U.S. - released movies over each month from 1940 onwards.

Seasonality over Time

Looking at the U.S. box office (adjusted for inflation) over time, we have noticed a few changing trends. As mentioned before, the summer months and the holiday months appear to be the highest months in terms of U.S. box office, with June being the highest. However it wasn't always this way. In fact, in the 1950s, February was the most lucrative month, followed by October.

Additionally, one that is familiar with the history of the film industry would know that the 'Summer Blockbuster' was invented in the 1970s with films like Jaws (1975), Star Wars (1977), and Alien (1979).[7] We can see that in the visualization below as the summer months in the 1970s strongly outperform in relation to their historic performance. This trend continues into the 1980s and beyond and the rest of the industry begins to embrace the philosophy of the 'Summer Tent-Pole'. In fact, in the

[6] Jonathan Bernstein, "Why January Is a Good Month to Bury Bad Movies," the Guardian (The Guardian, January 9, 2007), https://web.archive.org/web/20140929051530/http://www.theguardian.com/film/2007/jan/08/awardsandprizes.features.

[7] Belton DeLaine-Facey, "From Jaws to Star Wars, the Defining Summer Blockbusters of the 1970s," CBR, June 30, 2020, https://www.cbr.com/defining-summer-blockbusters-of-the-1970s/.

1980s, May, June and July made up more than 43% of that year's box office off of the back of summer films like E.T. The Extra-Terrestrial, Raiders of the Lost Ark, Ghostbusers, and Back to the Future.

Over the same period, the month of December, which was the strongest month in terms of box office throughout the 1960s and 1970s, declined in relation to the summer months. Over the past 10 years, we observe a general normalization across all months of the year where we see a less pronounced difference between these two time periods.

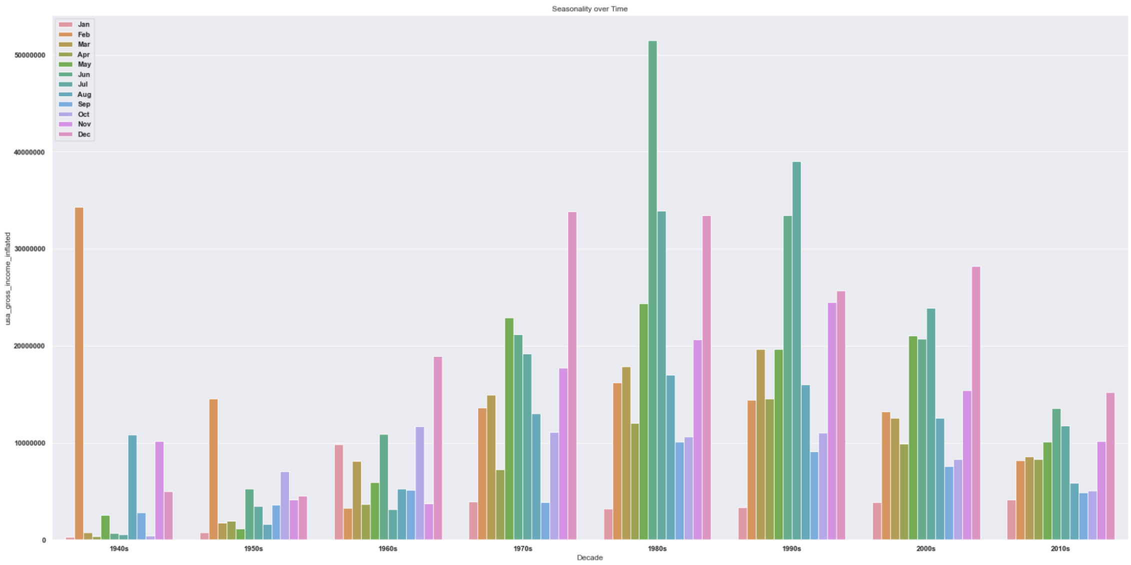


Figure 5: Plot of Decade (x axis) against average box office (left y axis). For each decade, we show a bar representing the box office average for each month over the span of the given 10-year period.

Seasonality and Critical Reception

As shown in appendix figure A.3, we see how release month correlates with a movie's critical success, or their ‘Weighted Average Vote’ (range 0-10) from our IMDB dataset. On an aggregate basis, looking at all movies in our dataset, movies tend to average around a 5.30 to a 5.90 score, with most months having an average of around 5.70. However, some months appear to over or under-index as compared to the others. For example, the worst performing month in terms of critical reception, January, has an average score 8.05% lower than that of the total average score across all months. This reaffirms our understanding of ‘Dump Months’ in that January is typically where studios like to send their films that they expect won’t be well received. On the other hand, the best month, September, has an average score of about 3.70% greater than that of all movies. Overall, however, we can deduce that seasonality has a generally low correlation on a movie's critical success.

Genre

In our quest to understand what factors are indicative of a successful movie, we compared genre to our metrics of success for a movie (IMDb rating and gross profit/income). Similar to the other datasets, this dataset was limited to USA movies only. Also we limited the dataset to gross profit in the US only (rather than worldwide, which is a parameter in the dataset), in order to have a better understanding of the results of our data since our entire group was born in the US. This dataset was not limited to the year 2000 to the 2020 like in previous sections of this report, rather it was the entire IMDb dataset combined with “The Movie Database” which spans from the year 1894 to 2020 for a total of 48,747 movies. Further subsections will relate the results of our exploratory analysis.

Genre vs. IMDb Rating

The first metric that we looked at in the genre analysis was genre vs. IMDb rating. The IMDb rating system is on a scale of 1-10. The genre list[8] that was used from “The Movie Database” (TMDB) which contains 19 genres: Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Music, Mystery, Romance, Science Fiction, Thriller, TV Movie, War, and Western. According to the combination of IMDb and TMDB datasets the top 5 highest rated genres were: 1. Documentary, 2. History, 3. War, 4. Music and 5. Drama whereas the 5 lowest rated genres starting from the bottom were: 1. TV Movie, 2. Thriller, 3. Action, 4. Fantasy and 5. Family. These findings should be noted that all of the genres fall within range of less than two rating points. Documentary, the highest rating, received an IMBd rating of 6.40/10 and Horror, the lowest rating, received an IMBd rating of 4.59/10.

These results are interesting for a number of reasons. First off, a disclaimer should be made that any potential outlining data was not filtered for in any way; ideally making this data the most realistic view of genre success rate since 1894 (the publishing date for the oldest movie in our dataset). One insight that can be gleaned from this analysis is that all over an approximately 125 year timescale, the 19 genres included in analysis are within two rating points. Since this is a mean of all the movie ratings, it could be said that there is a fair degree of uniformity among all the genres of movies in the sense that there is a small standard deviation between all 19 genres. The results can be seen in the figure below.

[8] Genres - movie Bible - the movie database (TMDB). (n.d.). Retrieved December 7, 2021, from https://www.themoviedb.org/bible/movie/59f3b16d9251414f20000006.

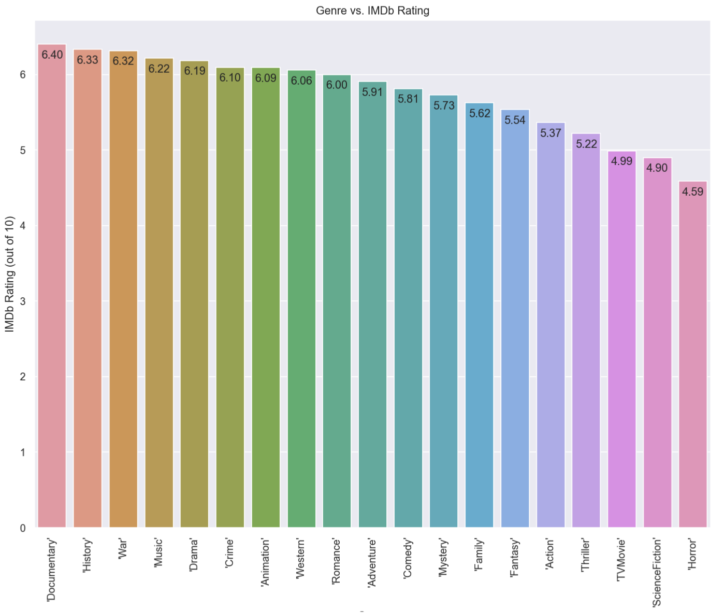


Figure 7: Plot of genre of movies (x axis) against IMDb rating (y axis). There were no movies higher than a 7/10 IMDb rating even when including all the entire datasets.

Genre vs. USA Gross Income

The second metric that we looked at in the genre analysis was genre vs. USA gross profit. The IMDb rating system is on a scale of 1-10 and when using the genre list[9] from "The Movie Database" there were 18 genres: Action, Adventure, Animation, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Music, Mystery, Romance, Science Fiction, Thriller, TV Movie, War, and Western. According to the dataset the top 5 highest grossing profit movies were: 1. Animation, 2. Adventure, 3. Fantasy, 4. Family and 5. Action whereas the 5 lowest rated genres starting from the bottom were: 1. Documentary, 2. Thriller, 3. Western, 4. Mystery and 5. History. Animation, the highest grossing film genre, received an average gross profit of $49,935,514 and Documentary, the lowest grossing film genre, received an average gross profit of $25,627. The results can be seen in the figure below.

[9] "Genres," Genres - movie Bible - the movie database (TMDB), accessed December 8, 2021, https://www.themoviedb.org/bible/movie/59f3b16d9251414f20000006.

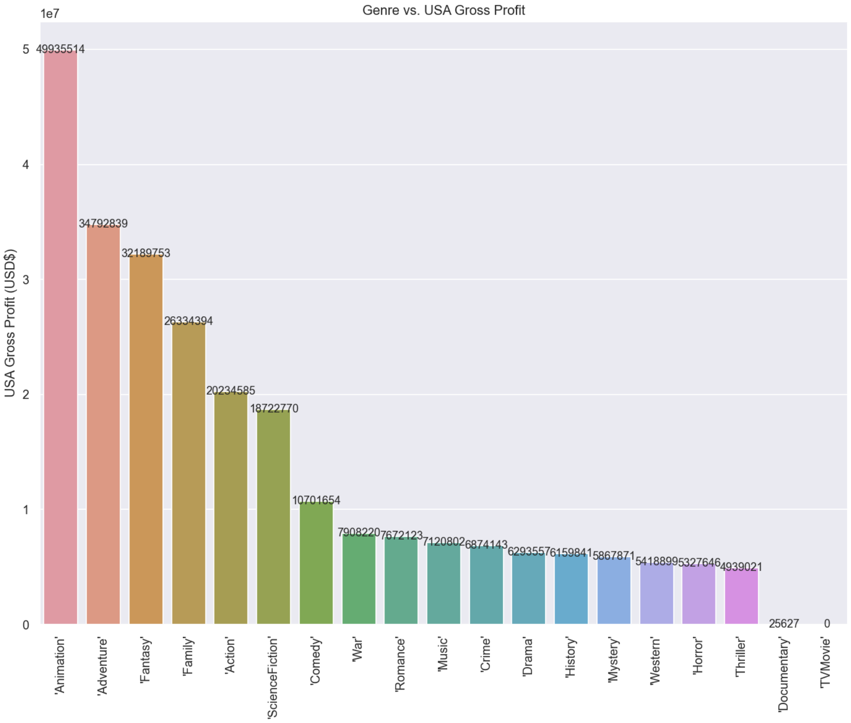


Figure 7: Plot of genre of movies (x axis) against USA Gross Profit (y axis). A wide spread of average gross profits from 1894-2020. Data missing for TV movie genre.

Conclusion

Regarding our first variable, there seems to be a strong correlation between budget and profits, and weak correlation between budget and IMDb ratings. Movies with large budgets tend to be more profitable, but not necessarily more well-received by the audience. However, it is important to note that even small-budget movies can still be profitable relative to their budget size, if it becomes well received by the audience.

In terms of seasonality, movies are seen to be over-performing both in the summer months (May, Jun., Jul.) and the Holiday months (Nov., Dec.). However, this tendency can only be traced back to around the 1970s. Before that, the seasonality of movies was much more sporadic. Furthermore, as it relates to critical reception, movies appear to be strongly in line with 'Dump Months' and the Academy awards; in that January tends to have the worst performing films (both financially and critically) while December represents the inverse.

In conclusion, we found there to be a correlation between genre of movie and USA gross profits and a weak-to-no correlation between the genre of movie and IMDb ratings. When looking at both IMDb rating and profits in tandem, there appears to be no correlation between metrics (i.e. Documentary is the highest IMDb rated film but it has the lowest gross profits). Future exploration of the dataset on issues relating to potential fluctuations in genre success rate globally versus just in the US could yield interesting results. It would also be interesting to compare these results with different databases or look at what is the demographics of the audience that participates in IMDb movie ratings.

Appendix

Figure A.1:

$$\text{Inflation Rate} = \frac{\text{CPI}_2 - \text{CPI}_1}{\text{CPI}_1} * 100$$

where:

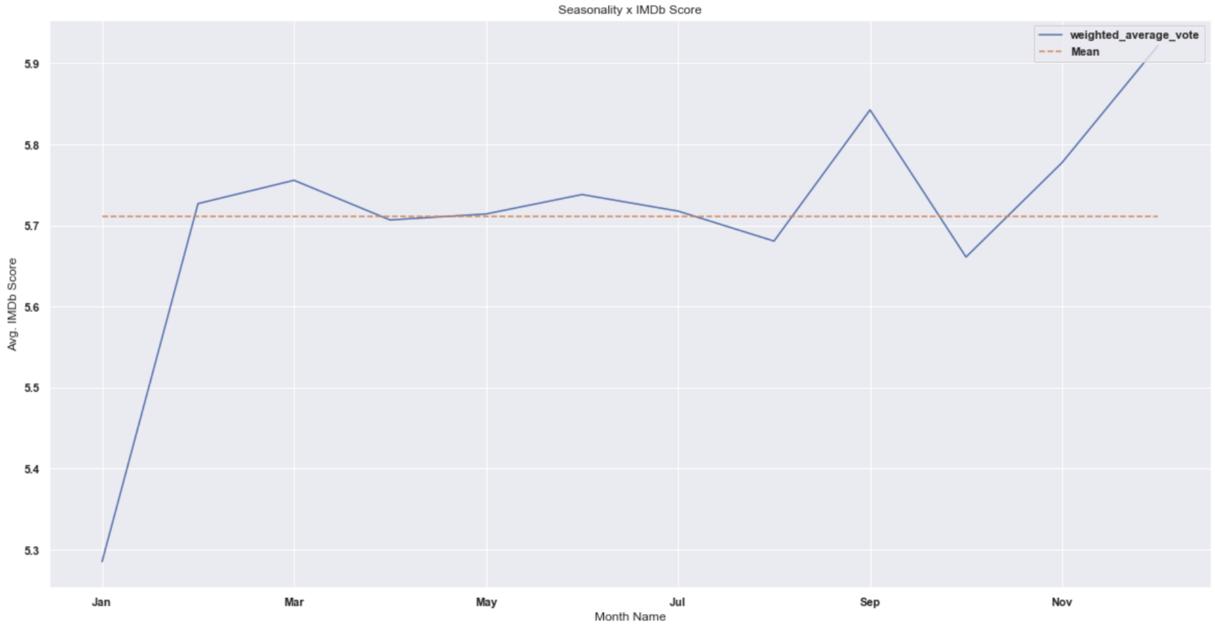
CPI_2 – is the CPI in the second period
CPI_1 – is the CPI in the previous period [10]

Figure A.2:

$$R = \frac{V_f - V_i}{V_i}$$

R = return
V_f = final value, including dividends and interest
V_i = initial value [11]

Figure A.3: Seasonality x IMDB Score



[10] Accessed December 8, 2021. https://cpiinflationcalculator.com/wp-content/uploads/2014/12/inflationrateformula.jpg .

[11] Accessed December 8, 2021. https://www.gstatic.com/education/formulas2/397133473/en/rate_of_return.svg.