Optimizing box office revenue through achieving a desired MPAA rating

WEI HAN KOH, LUKAS TAN, JEFFREY CHARLES DUNN, ZI FENG WEE

ABSTRACT

Our project explores the effect of the MPAA rating on a movie's box office earnings. We also analyze the content rating factors (`sex`, `violence`, and `language`) that influence the MPAA rating assigned to a movie. Finally, we investigate how a producer can achieve a desired MPAA rating by adjusting each content rating factor.

1. Our Project

a. Background information

The movie industry is a lucrative, multi-billion dollar industry that generates income from a few sources, such as box office revenues, DVD sales, licensing agreements, and streaming services. In 2019, box office revenues alone accounted for a record $42.3 billion (MPAA 2023), with international audiences driving the revenue growth by an estimated 4% (McClintock 2020) year-on-year.

Top performing movies can rake in as much as $2 billion in box office revenue, with "Avatar (2009)" topping the chart at a mind-blowing $2.8 billion (Childress and RT Staff 2022). With significant earnings at stake, movie production companies are highly motivated to maximise profitability.

One way movie production companies can improve profitability is by optimizing the MPAA ratings of its movies. With the "correct" MPAA rating, movie production companies can pitch their movies to the optimal target audience and thus increase their audience base. In the United States, the Motion Picture Association (MPAA) assigns ratings to movies on the basis of a variety of criteria. These ratings may advise age restrictions (e.g., "R") arising from the movie's content.

When developing movie content, production companies should carefully consider the MPAA ratings due to their potential to impact a movie's potential audience reach and revenue.

b. Problem framing

Our study has significant financial implications for movie production companies. Through enabling data-driven decision making on the optimal MPAA rating of movies, our study has two benefits:

Firstly, by identifying the factors that contribute to MPAA ratings and their impact on box office revenue, our research findings can help production companies finetune their movies' age-restricted content to stay within boundaries of their desired MPAA rating with the highest audience reach, which then translates to higher potential ticket sales.

Secondly, our study can provide valuable insights into the preferences and behaviours of moviegoers, thereby facilitating the development of more effective business strategies for movie production companies, gaining a competitive advantage in the highly profitable yet competitive industry.

c. Problem overview and general approach

For a movie production company, the MPAA rating of its movies may constrain the type and number of moviegoers paying to watch the movie in theaters. Furthermore, *ceteris paribus*, a movie's MPAA rating may influence the audience's perception and ratings of the movie.

Thus, obtaining a suitable age rating for its movies is crucial for a production company. Our project investigates the factors that contribute to such ratings with an aim to help a production company optimize a given movie's content by staying within certain rating thresholds to maximize its box office revenue potential.

We approach the problem by starting with descriptive analytics in Task I and move to predictive analytics in Task II.

In Task I, we fit linear and logistic regression models with respective numeric (dollar amount and critical ratings) and categorical (MPAA rating) features as our responses to investigate the effect of various predictors.

Next, we performed predictive analytics in Task II to investigate the effect of content ratings on MPAA ratings with logistic regression. We also investigated more advanced models such as XGBoost and Random Forests and compared their performance against the simpler models.

d. Initial hypotheses

From our initial understanding of the business problem and datasets, our initial hypotheses were:

1. The MPAA rating will have a significant effect on a movie's box office revenue and popularity (as measured by critical ratings).

2. release season will be correlated with a movie's box office performance.
3. The sex, violence, and language features will have predictive power of the MPAA rating, with language having the strongest coefficient.
4. The sex, violence, and language ratings will be correlated to a movie's overall box office performance.

2. The Data

a. Overview of data and sources

content_rating scraped dataset We started with the content rating dataset (Kids-In-Mind.com 2023) ('KIM') as our primary dataset. The website provides content rating and guidance to parents on the specific content of movies that their children would be exposed to.

Key variables in the dataset include MPAA rating (e.g., "G", "R"), and numeric ratings for each of the variables sex, violence, and language on a scale of 0 to 10 which were assigned by reviewers. Unlike MPAA, the content ratings provide details of a movie's content on a more granular level on each category but does not make age-specific recommendations.

budget_boxoffice scraped dataset We obtained data on movie production budget and box office revenue (The-Numbers.com 2023) which provides key financial data for our project. We modeled movie financials on their budget and revenue.

genre downloaded dataset The Internet Movie Database provides a number of rich datasets of movie information. We used movie genre data (IMDb.com 2023) to investigate whether genre has an effect on movie financials.

critical_review API data Movie reviews are made available on sites like Rotten Tomatoes and Metacritic. We obtained access of a site (imdb-api.com 2023) with movie review data, which we will use to proxy movie popularity. In our study, the terms "popularity" and "critical ratings" are used interchangeably.

inflation downloaded dataset Financial data obtained above spanned a number of years and was normalized for inflation for comparison in section 3.

b. Data cleaning process

Working with the two scraped datasets was difficult because they lacked a unique identifier field for joining with each other or the other datasets. The only common fields were movie title and year. Initially, we tried to join the datasets on these fields.

A survey of literature (Gomaa and Fahmy 2013) suggests three possible types of approaches: string-based, corpus-based, and knowledge-based. We attempted using string-based joins as the other two types dealt with larger corpora that exceeded the scope of this project.

Character-based (i.e., Damerau-Levenshtein) and term-based (Cosine similarity) approaches were attempted and were complicated by the following:

1. Inconsistent punctuation, e.g. "B*A*P*S" vs. "B.A.P.S."
2. Omission of parts of the title, e.g. "Star Wars: The Rise of Skywalker" vs "Star Wars: Episode IX - The Rise of Skywalker"

Furthermore, such approaches required comparing each movie against every other movie. This combinatorial explosion added to the challenge of performing the text-based join. For instance, with 5,000 movies in each of the three datasets, a total of 5000^3, i.e. 1.25e11 rows is produced. After joining, each row is compared against every other row to apply the string-based approaches.

We found these approaches to be unfeasible, with each run taking multiple hours as we searched for the optimal parameters. Moreover, we observed lower accuracy after a trial with a subset of data. Hence, we searched for an alternative.

c. Joining data

We obtained access to an application programming interface (API) with data from the Internet Movie Database.

This allowed us to query the best result with a search string of title and release_year to return only the best result with only a simple query, such as: https://imdb-api.com/en/API/SearchMovie/{api_key}/{title_year}. This API also had other endpoints which helped us generate the critical_review dataset by querying the movies from the scraped datasets.

Fig 1 provides an overview of the data schema across datasets after cleaning and joining.

d. Feature engineering

1) Release season

As movie box office (sales) exhibit strong seasonality through the calendar year (Einav 2007), we had to isolate the effects of seasonality from the other effects we intended to stud.y Thus, we created discrete

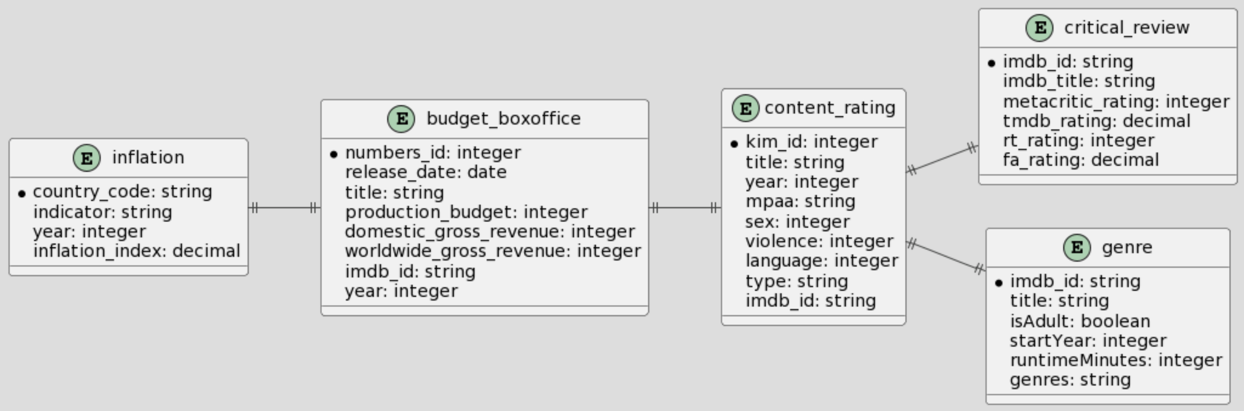

FIG. 1. Entity-relationship diagram of the individual datasets after cleaning. Movie data are related by imdb_id and inflation is related by year.

release seasons by grouping the `release date` feature into specific factors of:

1. `early_year`, calendar weeks 1-15 of the year
2. `summer`, calendar weeks 16-34 of the year
3. `fall`, calendar weeks 35-45 of the year
4. `holiday`, calendar weeks 46-53 of the year

The discretized feature was used successfully in (iii) to investigate how MPAA ratings affect box office earnings or popularity.

2) GENRE

`genre` was available as a list that included multiple genres per movie, with 24 total genres available in total. In order to effectively analyze our models, we used a multi-label binarization approach to one-hot encode the genre to make a feature for each individual genre.

The encoded feature was used successfully in (iii) to investigate the effect of genre on box office success and popularity.

3) INFLATION ADJUSTMENT

As previously mentioned, movie financial data (`production_budget` and `box_office`) were used to model the financial performance of the movies.

This dataset has information on movies spanning over 100 years, beginning from 1915. To allow for comparisons across this extended period, we adjusted `production_budget` and `box_office` by the annual headline Consumer Price Index from the World Bank (Ha et al. 2023).

The inflation-adjusted data was used throughout our paper, with some noted limitations c.

3. Modeling

We used the below models in our paper to answer our research questions. A summary of our process flow and the models used can be found in Fig 2.

We used the following models in our project:

1. Linear regression
2. Logistic regression
3. XGBoost
4. Random Forests

Our paper addresses two main tasks and aims to solve the business problem identified in section c.

Task I: Does MPAA rating affect a movie's box office performance and critical reviews?

We first investigate the effect of MPAA rating on box office performance and critical reviews.

Our source dataset includes four different ratings from:

1. `metacritic_rating`, from metacritic
2. `tmdb_rating`, from The Movie Database
3. `rt_rating`, from Rotten Tomatoes
4. `fa_rating`, from movie Affinity

Thus, we need to define the meaning of “critical reviews” among these four different sources.

First, we observed missing data in some movies, making it necessary to decide which rating to use for the analysis. Given the importance of each rating to the dataset, we computed a mean score (not considering nulls) to preserve as much information as possible.

Nevertheless, this approach posed a challenge since not all ratings were on the same scale. We plotted the distribution of each of the four ratings features (see Fig 3) and observed that three out of four

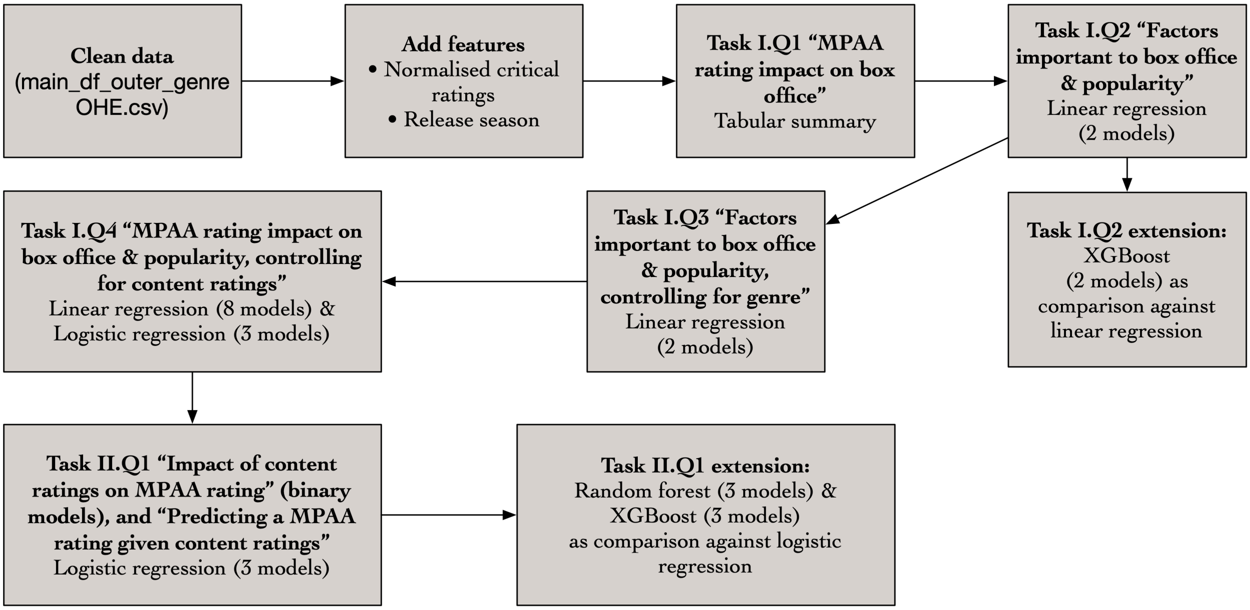
Clean data (main_df_outer_genre OHE.csv)
Add features
• Normalised critical ratings
• Release season
Task I.Q1 "MPAA rating impact on box office"
Tabular summary
Task I.Q2 "Factors important to box office & popularity"
Linear regression (2 models)
Task I.Q2 extension:
XGBoost (2 models) as comparison against linear regression
Task I.Q3 "Factors important to box office & popularity, controlling for genre"
Linear regression (2 models)
Task I.Q4 "MPAA rating impact on box office & popularity, controlling for content ratings"
Linear regression (8 models) & Logistic regression (3 models)
Task II.Q1 "Impact of content ratings on MPAA rating" (binary models), and "Predicting a MPAA rating given content ratings"
Logistic regression (3 models)
Task II.Q1 extension:
Random forest (3 models) & XGBoost (3 models) as comparison against logistic regression

FIG. 2. Overview of models in our paper in relation to our research questions.

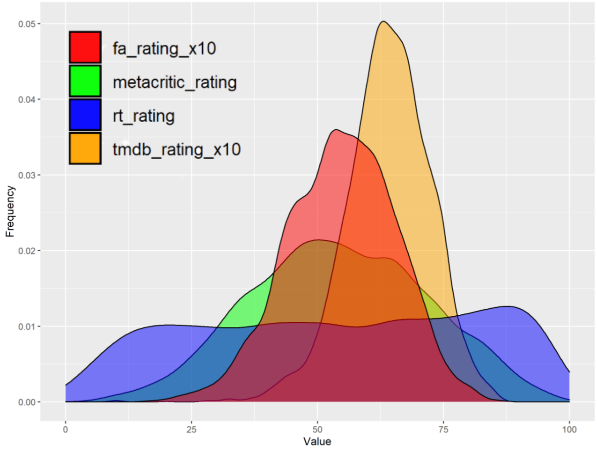

FIG. 3. Frequency distribution of normalized ratings of critical ratings

ratings appeared to be normally distributed, while the fourth rating feature (`rt_rating`) followed the uniform distribution. This is confirmed by the Q-Q plots (see Fig 4).

The ratings that fit a normal distribution were first standardized.

The resulting z-scores were then converted into percentiles to scale the data from the original range to a 0-100 range.

The uniformly distributed rating was also divided into percentiles, making all four columns comparable.

We engineered a new feature (`mean_Norm_score`) to capture a single critical rating for each movie.

Following this, we further divided *Task 1* into four questions.

QUESTION 1: WHICH MPAA RATINGS ARE ASSOCIATED WITH BOX OFFICE EARNINGS AND POPULARITY?

MPAA Rating	*Profit$M*	*Revenue$M*	*Popularity*
G	266	354	57.1
PG	182	267	47.2
PG-13	163	247	44.7
R	70	115	53.4

TABLE 1. Mean box office performance (adjusted for inflation) and popularity by MPAA rating

We analyzed the effect of the MPAA rating on box office revenue, profit, and critical reviews (Table 1). We show that G-rated movies are both the most profitable and highest earning, as well as most popular. We also observe PG-13-rated movies to be less popular than R-rated movies, but fare better at the box office with twice the amount of revenue and profit.

QUESTION 2: WHAT FACTORS ARE IMPORTANT TO A MOVIE'S BOX OFFICE SUCCESS AND POPULARITY?

We hypothesized that "time of year released" is important to a movie's box office performance.

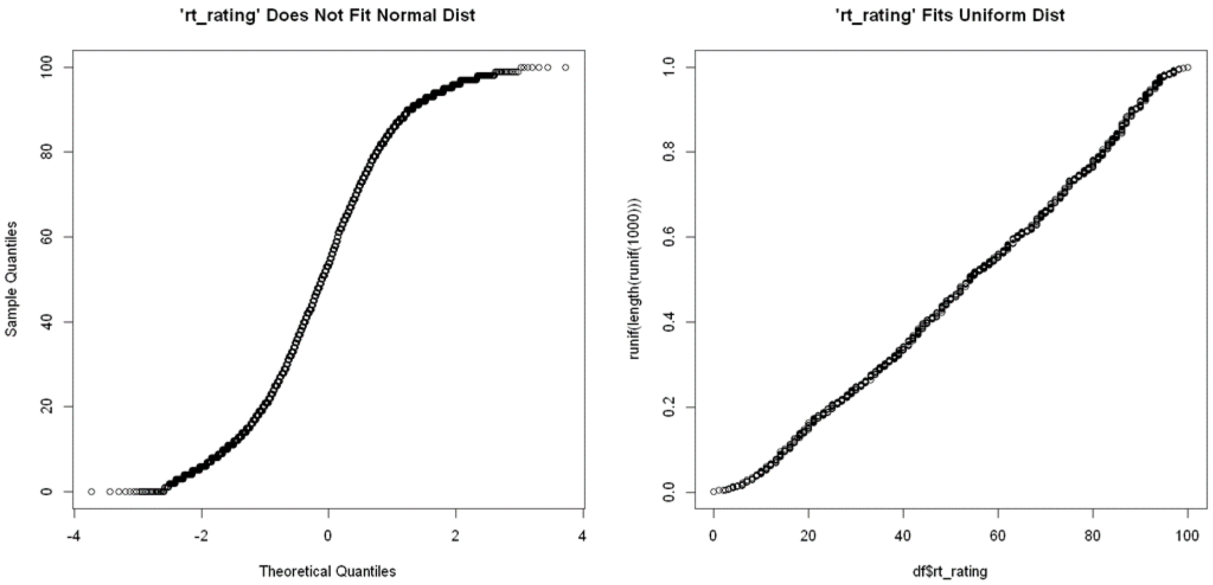

FIG. 4. Q-Q plots of rt_rating

Thus, we created a new feature by discretizing the `release_date` feature into bins, then categorized the release date into a few categories. The illustration below (Fig 5) shows the number of movies released in each week of the year.

We observed that the holiday season exhibited the highest number of releases, while January and late summer have comparatively fewer releases. As the distribution lacks obvious clusters, we proceeded with classifying releases into `early_year`, `summer`, `fall`, and `holiday`.

We fit a linear regression model (model 1.1) to identify factors associated with box office profits and found that production budget (adjusted for inflation) has a strong positive correlation with box office profits, with an increase of $1.00 in budget associated with a $2.03 increase in profits.

Movies released during the `summer` and `holiday` seasons tend to have higher box office profits, but this correlation does not necessarily mean that the seasons directly cause higher profits. The model only shows a correlation between the variables. It is likely that movies with lower budgets or expectations simply choose not to compete during these seasons to avoid intense competition.

Finally, our model shows that R-rated movies are likely to underperform G-rated movies in the box office, with an expected difference of $101.6 million.

We modeled (model 1.2) the factors predictive of the critical ratings of a movie and observed the following:

(i) Release season First, releases during the `summer`, `fall`, and `holiday` are associated with an increase in critical ratings. Of these, `holiday` is associated with the strongest effect. These results are statistically significant and indicate that release timing is significantly correlated with critical reception.

(ii) Production budget When adjusted for inflation, `production_budget` has a statistically significant coefficient of 4.327e-08. This indicates that a $100 million-dollar increase in a movie's budget is associated with an increase in the combined rating by 4.33 points. We again note that this does not imply a causal relationship between budget and critical ratings.

(iii) Content ratings Furthermore, the model shows that higher `sex` and `violence` content ratings are associated with lower critical ratings, while `language` does not have a significant effect. We observe that R-rated movies are associated with higher critical ratings, with a nearly 15-point advantage over G-rated movies. However, this advantage is somewhat diminished due to the fact that R-rated movies tend to have higher `sex` and `violence` content ratings, which have been shown to lower critical ratings as we saw above.

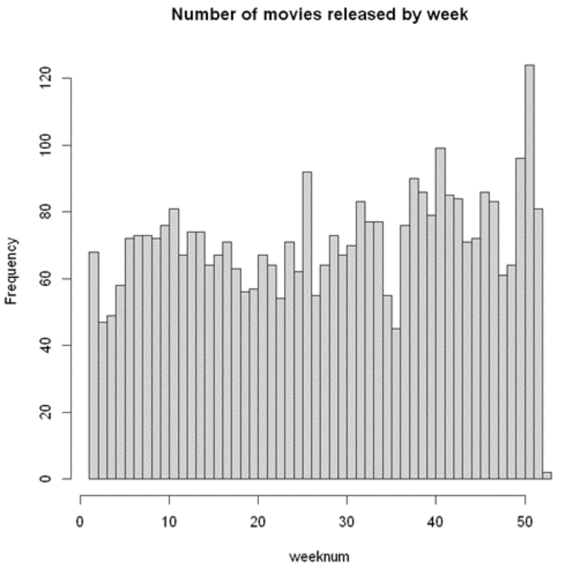

FIG. 5. Number of movies released by week

Question 3: How does the movie genre influence movie's box office success and popularity?

Our genre dataset contains up to 3 genre labels for each movie. Encoding each of the genres as a separate binary variable, we can incorporate these into the models from Question 2 to understand the impact of genre on the profits and popularity of a movie.

The genres of Action, Comedy, Drama, and Sport were all statistically significant (at α = 0.01) in regressing box office profit, each being associated with a decrease in profit.

The results were more varied when it comes to popularity:

- Action, Comedy, Horror, and Romance movies were associated with statistically significant decreases in popularity of 5.5, 6.7, 1.2, and 3.3 points.
- Animation, Biography, Documentary, Drama, and Musical movies were associated with increases in popularity of 1.6, 8.1, 2.0, 1.3, and 1.2, respectively.

Question 4: Holding constant movie content, does the MPAA rating affect box office earnings or popularity?

Within the context of a movie production company, an inquiry into whether the MPAA rating directly impacts a movie's popularity or profitability is of significant interest. However, a simple comparison between two MPAA ratings would not be enough to determine the effect of the MPAA rating, as movies with distinct ratings may inherently possess divergent content. An optimal approach would involve acquiring a substantial number of movies, randomly allocating either a "PG-13" or "R" MPAA rating prior to their release, and subsequently analyzing the final box office results and popularity ratings. Discrepancies in the findings would suggest the MPAA rating as being a potential causal factor.

A primary determinant of MPAA ratings includes the degree of sexual content, language, and violence portrayed in a movie. Hence, we can use the content ratings to determine "borderline" movies between two MPAA ratings to simulate a random assignment of MPAA ratings.

Since PG-13 and R rated movies comprise most of the dataset, we first filtered the dataset to movies with these ratings. The subsequent analysis entailed examining the overlaid distributions of each `sex`, `violence`, and `language` category for PG-13 and R movies. For `sex`, movies with a rating between 3 and 5 were retained, while for `violence`, only those with a rating between 3 and 6 were considered. In contrast, the `language` variable exhibited a more bipolar nature, with a rating of 5 representing the majority of overlap between PG-13 and R movies.

Moreover, the interaction between content rating variables and overlapping distributions necessitated further consideration. Consequently, the dataset was refined to include only movies with a combined content rating score between 13 and 16.

The final filtered dataset contained 375 "borderline" movies indistinguishable in `sex`, `violence`, and `language` content. To assess the randomness of the data, we regressed `mpaa_R` on `release_season`, `sex`, and `violence` with a logistic regression model. No coefficients were significant at the α = 0.05 level, which suggests that the MPAA ratings allocated to the movies within this "borderline" dataset were effectively random.

With this "borderline" dataset, we found no effect of an R rating on a movie's inflation-adjusted profits, however a substantial effect (17.8 points) on its popularity score was observed. Nevertheless, investigating the genre balance between the two classes showed an imbalance across numerous genres. For instance, R-rated movies were predominantly `drama`, which correlated with an increase in popularity ratings, while PG-13 movies predominantly featured `action` and `comedy`, both of which were associated with lower popularity ratings. Consequently, incorporating significant genres as features removed the effect of MPAA rating on popularity ratings.

Our final attempt to control for the genre effect involved selecting a single genre, further homogenizing the "borderline" dataset. We filtered for `drama` movies since this genre encompassed the greatest number of movies. Regressing the MPAA rating against the `release_season`, `sex`, and `violence` again yielded no significant coefficients, indicating the MPAA rating was still "randomly" distributed.

Feature	*Estimate*	*t value*	*Sig.*
(Intercept)	67.88	3.021	**
production_budget_infAdj	4.949e-08	0.905	
release_seasonfall	7.087	1.091	
release_seasonholiday	14.24	1.924	.
release_seasonsummer	10.04	1.522	
sex	-3.089	-0.959	
violence	-2.839	-1.029	
mpaaR	17.44	2.819	**

TABLE 2. Results of linear regression model 1.4.16 (adj. R-squared = 0.07), which controls for the effects of genre and language ratings. We demonstrate that MPAA R rating increases popularity score by 17.44 points.

Despite the lack of impact on a movie's profitability, an MPAA rating of R was associated with a higher popularity score, exhibiting an effect of 17.44 points (refer to Table 2). As both the movie's content and genre were held constant, this finding supports the notion that an R rating contributes to elevated critical ratings. However, our analysis may still be constrained by the limited size of the dataset (168 observations) and other potential confounding variables, such as production company, director, and actors which were not explored in our analysis.

This part of our hypothesis was correct in that there was an effect of MPAA rating on a movie's popularity, but incorrect in that there was no effect on its box office performance.

Task II: How can a movie producer adjust a movie's content to achieve a desired MPAA rating?

The Kids-in-Mind.com content ratings, which are based on actual movie scenes, could be used to predict the MPAA rating assigned to a movie. While we did not analyze how individual scenes contribute to a content rating, the consistency of content ratings from this data source allows production companies to reasonably estimate these ratings for their movies in production based on the troves of historical data available.

Model		*Estimate*	*t value*	*Sig.*
PG vs. G	(Intercept)	0.513	15.837	***
PG vs. G	sex	0.047	4.464	***
PG vs. G	violence	0.034	3.634	***
PG vs. G	language	0.099	9.813	***
PG-13 vs. PG	(Intercept)	-0.288	-14.80	***
PG-13 vs. PG	sex	0.067	16.52	***
PG-13 vs. PG	violence	0.073	19.97	***
PG-13 vs. PG	language	0.144	32.25	***
R vs. PG-13	(Intercept)	-0.507	-25.68	***
R vs. PG-13	sex	0.043	15.82	***
R vs. PG-13	violence	0.048	18.07	***
R vs. PG-13	language	0.109	42.24	***

TABLE 3. Results of 3 binary-class logistic regression models (AIC = 525, 1,466 and 3,516, respectively)

Instead, our paper mainly focuses on adjusting a movie's content by way of its content rating scores, as this approach is sufficient for achieving the project's objectives.

Nonetheless, studying how individual movie scenes contribute to Kids-in-Mind.com ratings could be a fruitful area for future research.

Question 1: In what way do content ratings influence the MPAA rating? We will be using three distinct models to address the problem: G vs. PG, PG vs. PG-13, and PG-13 vs. R.

First, binary class models are more interpretable than a single multi-class model, aiding understanding in the factors influencing an MPAA rating.

Second, these models are more relevant to the business problem than a single model would be.

A movie production company rarely has to decide between two extremely different ratings (e.g., R vs G) for their movie. Therefore, a logistic regression model that compares one rating against all others may not be helpful in understanding which factors influence a movie's rating to fall on one side of a two-rating spectrum or the other.

We first modeled G versus PG movies and found that `sex`, `violence`, and `language` were all significant, with `language` having the strongest effect (Table 3).

For PG versus PG-13 and PG-13 versus R, all three content rating variables were also significant, with `language` again showing the strongest effect.

These models can be used to predict a given movie's MPAA rating based on estimates of the content ratings. For example, if the estimated content rating factors of `sex`, `violence`, and `language` are 4, 7, and 5 respectively, our model (`model.PG13_R`) predicts this movie has a 55% probability of receiving an R rating as opposed to a PG-13 rating.

At this point, we split our datasets into test and train datasets in order to compare performance on classification via a confusion matrix, through which we determined our accuracy metric.

The performance of our three logistic regression models was documented in Table 4.

Model	Accuracy	Sensitivity	Specificity
PG vs. G	0.889	0.292	0.971
PG-13 vs. PG	0.907	0.809	0.951
R vs. PG-13	0.867	0.914	0.828

TABLE 4. Accuracy of 3 binary-class logistic regression models after train-test split

Comparing it to our alternative models of XGBoost and Random Forests, logistic regression performed well enough in terms of accuracy (refer to Table 5).

Model	Logistic Regression	Random Forest	XGBoost
PG vs. G	0.889	0.894	0.885
PG-13 vs. PG	0.907	0.913	0.920
R vs. PG-13	0.867	0.894	0.913

TABLE 5. Accuracy of 3 binary-class logistic regression models compared to XGBoost and Random Forests

In terms of choosing hyperparameters, we did manual trial-and-error with a few common values while searching for the right hyperparameter value. For instance, for Random Forests, we experimented with the default value of 500 for `ntree` but also went around searching for the best model in the range 100 through 900 in increments of 200.

In this case, there was no significant difference between accuracy metrics nor in the training time required for the different hyperparameter values. Hence, we went with the value that gave the most accuracy and which required the least amount of time to train. As a result of this process, we decided on `ntree` = 300.

4. Challenges and future research

a. Size of dataset

In order to answer our research questions, we had to isolate and control for the effects of certain variables, such as `genre` and `language`. We implemented this by filtering for movies in the "drama" genre with similar `language` ratings. As a result, certain models in the later parts of Task I were run on relatively smaller-sized datasets with *n* below 300 data points.

From our research, the common rule of thumb is 10 data points per independent variable, especially when using linear regression, which our linear regression models did satisfy.

Upon further investigation in extant literature, a study (Austin and Steyerberg 2015) goes even further and suggests that for linear regression only two data points are required per independent variable.

Nonetheless, it would be interesting to see if there could be additional nuances from the data if we had access to more data points.

b. Low adjusted R-squared

In certain models, particularly the models where genre and content ratings were controlled for (for an example, see Table 2), we observed low adjusted R-squared.

We believe that this might also be related to small dataset size. In controlling tightly for genre and content ratings, we might have also reduced the predictive power of the model.

It is also possible that low adjusted R-squared is due to other unidentified factors beyond what we could cover in our paper.

c. Inflation dataset

The World Bank dataset includes headline CPI data since 1970. Although most of the movies in our content rating dataset had release dates after 1970, there were four movies that were released before 1970 which also had content ratings.

For these four movies, we used the unadjusted dollar amounts since there were so few of them, and acknowledge that this may cause the effects of these movies to appear smaller than other movies.

d. Areas for future research

In the course of our analysis, we identified the following areas of future study:

(i) Analysis of actual scene content In place of analysing the content ratings as a proxy for a film's content, analyzing the actual movie script is a potential avenue of future research. Directly analysing content has the benefit of eliminating a separate source of potential bias and may provide direct feedback on the insertion or removal of specific scenes of content to optimize for critical reviews or a desired MPAA rating.

(ii) Impact of 'star' actors, directors, and movie production companies An obvious potential confounding factor we did not get a chance to investigate

is the contribution of 'star' power from actors, directors, and even movie production companies. As demonstrated, e.g. (Elberse 2007), group dynamics are such that cinema-goers may be inclined to watch a movie in the cinema simply because it stars their favorite actor. Hence, a future study of movie content ratings may be benefit from investigating this 'star' effect.

(iii) Point in time approach Our analysis is conducted at a point in time surveying movies released by January 2023. Given the changing habits and tastes of movie-goers, it may be helpful to reproduce the analysis periodically to observe future trends and seasonality.

5. Conclusion and key takeways

We discussed methodology and evaluation of results in detail above in 3. Below, we provide our overall conclusion of our project, including its key novelties and insights.

For our hypothesis #3, we indeed saw that the `sex`, `violence`, and `language` features had strong predictive power of the MPAA rating of a film, with `language` having the strongest. However, we were incorrect in hypothesis #4 - neither of those three variables were significantly correlated to a film's box office performance. Rather, our modeling showed that the time of year released and MPAA ratings were predictive of a film's box office performance (which was our second hypothesis).

Finally, as noted in the actual analysis, our hypothesis #1 was partially correct in that the MPAA rating was shown to have an effect on a film's popularity but not its box office performance.

Overall, our findings above demonstrate the practical utility of content ratings in facilitating informed decision-making by a movie production company seeking to optimize for critical ratings, MPAA ratings, or profitability. In particular, for movies in production that are straddling between two potential MPAA ratings, we showed three different models that are highly accurate in predicting which MPAA rating would eventually be assigned to the movie.

Our results also reveal that certain genres, specifically Documentary, Drama, and Musical, are associated with more favorable critical reviews. This insight will benefit movie production companies seeking to produce movies that are more critically reviewed, for instance, at the start of the production when selecting between pitches of different genres.

6. Literature survey

On top of the technical literature on methods and techniques cited elsewhere, we reviewed the following papers related to our topic.

The effect of MPAA ratings on movie financials was studied (Sundaram 2006). G-rated movies were associated with highest return on investment (ROI). Conversely, R-rated movies were found to have the lowest ROI.

A study on movie content and box office revenue found that violence in movies persisted over time and associated this with market demand for such content (Barranco et al. 2017).

The trend of MPAA ratings over time was investigated; the MPAA was found to have become more lenient over time towards violent content (Leone and Barowski 2011).

`genre` was identified as an important factor on movie profitability while the importance of MPAA rating was found to be lower than suggested in previous literature (Karniouchina et al. 2010).

Works cited

Austin, P. C., and E. W. Steyerberg, 2015: The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, **68**, 627–636, doi:https://doi.org/10.1016/j.jclinepi.2014.12.014. https://www.sciencedirect.com/science/article/pii/S0895435615000141.

Barranco, R. E., N. E. Rader, and A. Smith, 2017: Violence at the box office: Considering ratings, ticket sales, and content of movies. *Communication Research*, **44**, 77–95, doi:10.1177/0093650215614363.

Childress, E., and RT Staff, 2022: Rotten tomatoes: The 50 highest-grossing movies of all time. https://editorial.rottentomatoes.com/article/highest-grossing-movies-all-time/.

Einav, L., 2007: Seasonality in the u.s. Motion picture industry. *The RAND Journal of Economics*, **38**, 127–145, doi:https://doi.org/10.1111/j.1756-2171.2007.tb00048.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-2171.2007.tb00048.x.

Elberse, A., 2007: The power of stars: Do star actors drive the success of movies? *Journal of Marketing*, **71**, 102–120, doi:10.1509/jmkg.71.4.102.

Gomaa, W. H., and A. A. Fahmy, 2013: A survey of text similarity approaches. *International Journal of Computer Applications*, **68**, 13–18, doi:10.5120/11638-7118.

Ha, J., A. Kose, and F. Ohnsorge, 2023: One-stop source: A global database of inflation. *Policy Research Working Paper*, **9737**. https://www.worldbank.org/en/research/brief/inflation-database.

imdb-api.com, 2023: IMDb API documentation. https://imdb-api.com/api#Ratings-header.

IMDb.com, 2023: IMDb datasets. https://imdb.com/interfaces/.

Karniouchina, E., S. Carson, and W. L. Moore, 2010: A note on revenue versus profitability as indicators of motion picture performance. *SSRN*, doi:10.2139/ssrn.1712088.

Kids-In-Mind.com, 2023: Kids-in-mind.com | parents' movie guide, ratings and review. https://kids-in-mind.com/a.htm.

Leone, R., and L. Barowski, 2011: MPAA ratings creep. *Journal of Children and Media*, **5**, 53–68, doi:10.1080/17482798.2011.533488.

McClintock, P., 2020: The hollywood reporter: 2019 global box office revenue hit record $42.5B despite 4 percent dip in u.s. https://www.hollywoodreporter.com/news/general-news/2019-global-box-office-hit-record-425b-4-perc

MPAA, 2023: Statista: Global box office revenue 2005-2021. https://www.statista.com/statistics/271856/global-box-office-revenue/.

Sundaram, S., 2006: Profitability study of MPAA rated movies. *Seidman Business Review*, **12**. https://scholarworks.gvsu.edu/sbr/vol12/iss1/6.

The-Numbers.com, 2023: The numbers - movie budgets. https://www.the-numbers.com/movie/budgets/all.