FROM CAREER NIGHTMARES TO CINEMATIC SCARES: CRAFTING A BLOCKBUSTER HORROR STRATEGY

- Statistics for Data Scientists -

JBP061-B-6

June 5, 2025

By

Angelo Filiol de Raimond - *2154269*

Jheronimus Acadamy of Data Science

Contents

1 Understanding the Horror Movie Landscape

1.1 Introduction

What makes a horror movie financially successful? This question drives the present analysis. Success in this industry depends on multiple factors — genre trends, audience reach, perceived quality, production choices — but the ability to identify and quantify these elements is rare.

To tackle this challenge, we use a publicly available dataset curated for the TidyTuesday project (October 29, 2024), based on IMDb horror and monster movies. The dataset includes metadata such as runtime, genre, ratings, and vote counts — all of which can inform creative and strategic decisions for future productions.

Our goal in this section is to clean and explore the data, define our performance metrics, and lay the foundation for a statistical analysis that will support a viable production strategy for a horror film.

1.2 Dataset Overview and Market Landscape

The original dataset contains 630 entries. We applied several filtering steps to isolate full-length horror movies with complete data. The cleaned dataset includes 169 films.

Filtering Step	Remaining Entries
Initial dataset	630
Filter: only `movie` type	528
Filter: genre contains `Horror`	223
Remove duplicates on title	193
Remove missing values (votes, rating, runtime)	169

Table 1: Table 1 — Dataset filtering steps

Beyond cleaning, the dataset offers a glimpse of the horror market's internal structure. Sub-genres such as *Comedy*, *Sci-Fi*, and *Action* are overrepresented, suggesting these niches are popular — but also likely to be more competitive.

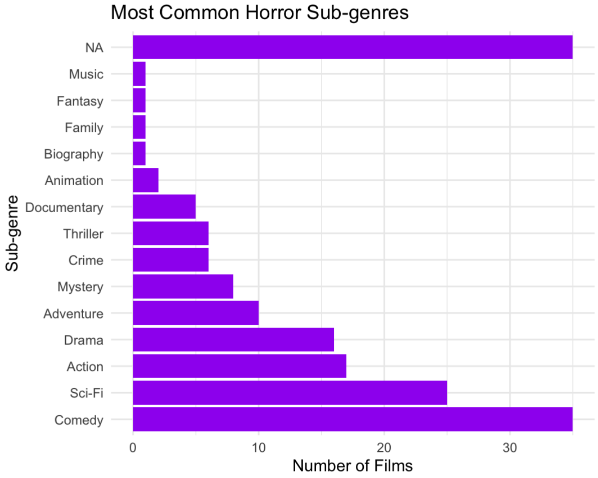

Figure 1: Distribution of horror sub-genres

1.3 Variable Engineering

To support meaningful comparisons and decisions, we derived two key variables: `sub_genre` and `duration_group`.

Sub-genre. Since all selected movies include "Horror" by design, we extracted the secondary genre listed (if applicable) to define a clearer thematic niche: e.g., *Horror-Comedy*, *Horror-Sci-Fi*, etc. This distinction allows us to analyze market positioning at a finer level.

Duration group. The `runtime_minutes` variable was grouped into four bins:

- `<60` minutes (short films)
- `60-89` minutes (short-format features)
- `90-119` minutes (standard format)
- `120+` minutes (extended format)

This grouping simplifies format-based comparisons. For instance, it helps assess whether longer horror films perform better — or if tighter runtimes are more appealing commercially.

These engineered variables enhance interpretability and directly support creative decisions about format and theme.

1.4 Performance Metrics and Revenue Potential

To measure success, we define two key performance indicators:

1. **IMDb vote count (`num_votes`)**, used as a proxy for popularity and reach. The number of votes likely reflects higher visibility and exposure, which in turn relates to revenue potential, especially if the film is distributed on monetized platforms such as streaming services or theatrical releases.

2. **IMDb average rating (`average_rating`)**, used as a proxy for perceived quality. Although quality does not guarantee commercial success, it may contribute to long-term profitability via reviews, word-of-mouth, and critical reputation.

The vote count distribution is highly skewed — a few movies receive thousands of votes, while most receive less than a hundred. To avoid distortion, we apply a log transformation:

$$\texttt{log\_votes} = \log(1 + \texttt{num\_votes})$$

(This transformation makes the data more stable and avoids letting blockbuster outliers dominate the analysis.)

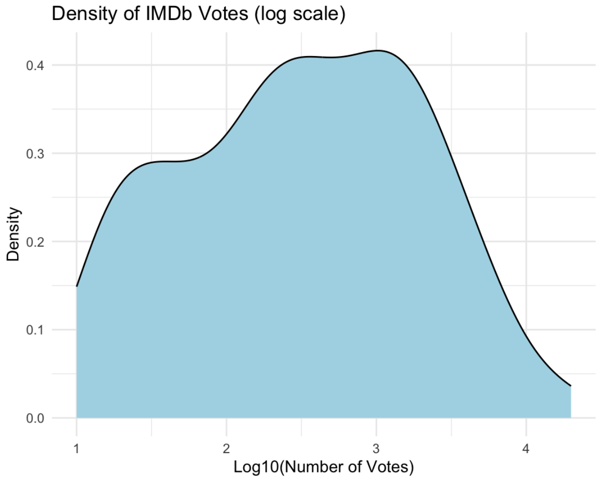

Figure 2: IMDb vote distribution (log scale)

Interestingly, the correlation between votes and ratings is almost null (≈ 0.03). This means that popularity and quality should be treated as distinct dimensions in any strategic analysis.

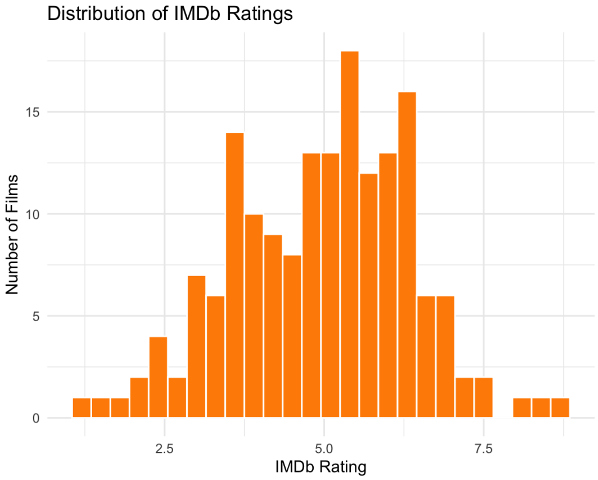

Figure 3: IMDb rating distribution

From a strategic perspective, our goal is not just to create a good horror film — but a successful one. If our goal is commercial success, maximizing vote count becomes crucial, as it reflects broader audience exposure on monetized platforms. A film with a high rating but low visibility may still fail financially. Popularity, not just quality, drives revenue.

Monetization potential. While our dataset does not include production costs or revenue figures, we assume that vote count reflects audience exposure on monetized platforms. For instance, a horror film with 10,000+ votes on IMDb likely received wide distribution via streaming or theatrical release. Assuming even modest monetization (e.g., \$1–2 per viewer), the difference between a film with 300 votes and one with 10,000 becomes financially significant.

Cost assumptions. Industry reports suggest that low-budget horror films are often produced for under \$5 million — sometimes even below \$1 million. Given this, even moderate exposure (e.g., 5,000–10,000 viewers) may be sufficient to break even, particularly on platforms with low distribution costs (e.g., Netflix, Amazon Prime Video). Budget-awareness should thus guide decisions on duration, cast, and sub-genre complexity.

This revenue-cost balance is why we must identify not only what makes a film good — but what makes it profitable. The next section explores precisely this question.

1.5 Strategic vs Control Variables

To clarify our analytical framework, we distinguish between variables that can be directly manipulated by the producer (strategic levers), and those that must be controlled for but cannot be changed at will.

Variable	Type	Role in Decision-making
sub_genre	Categorical	Strategic (narrative choice)
duration_group	Categorical	Strategic (editing/format)
average_rating	Continuous	Control (perceived quality)
runtime_minutes	Continuous	Control (fine adjustment)

Table 2: Table 2 — Strategic and control variables

The following variables will be used to inform production and marketing choices:

- **sub_genre** — narrative positioning: determines the thematic direction of the film.
- **duration_group** — format decisions: guides the expected length and structure.
- **num_votes** — popularity proxy: used to measure and aim for exposure.
- **average_rating** — perceived quality: a control variable for audience reception.

1.6 Descriptive Statistics

We now summarize the distributions of the three most important continuous variables. These values already offer strategic insights:

- The median vote count is low (332), but the upper quartile exceeds 1,400 — suggesting that some films succeed exceptionally well in visibility.
- The rating distribution is centered around 5.0, indicating generally modest audience reception.
- Most runtimes fall between 75 and 95 minutes, reinforcing the dominance of compact formats.

Variable	Min	Q1	Median	Mean	Q3	Max
num_votes	10	69	332	1273	1413	19895
average_rating	1.2	3.9	5.1	4.94	5.9	8.8
runtime_minutes	50	75	84	85.8	93	295

Table 3: Table 3 — Descriptive statistics of key variables

These figures suggest that the horror market rewards films with tight runtimes and accessible narratives — but only a few reach significant popularity.

1.7 Conclusion

The cleaned dataset now gives us a solid foundation to evaluate which production choices — such as sub-genre or runtime — are linked to higher exposure and potential profitability. We have defined measurable indicators of success, identified strategic levers, and documented the competitive landscape.

Having mapped the key elements of the horror movie landscape, we now turn to the fundamental question: which production decisions actually drive audience exposure and critical reception — and ultimately, profitability?

2 Which Production Choices Drive Exposure?

2.1 Modeling Strategy

To go beyond descriptive statistics and explore which features genuinely influence audience exposure, we use statistical modeling to predict our main performance indicator: the logarithm of the IMDb vote count.

Why log? Raw vote counts are heavily skewed — a handful of horror films receive thousands of votes, while most receive fewer than a hundred. Taking the logarithm stabilizes this distribution and allows us to interpret the model coefficients more reliably (i.e., as marginal percentage changes). The formula we estimate is:

$$\log(1+\texttt{num\_votes}) = \beta_0 + \beta_1 \cdot \texttt{average\_rating} + \beta_2 \cdot \texttt{runtime\_minutes} + \text{dummies}$$

We test three types of models to understand and compare results:

- **Linear regression (OLS)** — a standard benchmark, interpretable and widely used.
- **Lasso regression** — introduces a penalty that shrinks less important coefficients to zero, helping variable selection.
- **Random Forest** — a flexible, non-linear model that captures interactions and complex relationships, but sacrifices interpretability.

By comparing these models, we aim to balance two goals: understanding which factors influence popularity, and predicting vote counts with reasonable accuracy.

2.2 Key Relationships in the Data

Before diving into model outputs, we explore simple bivariate relationships to guide expectations.

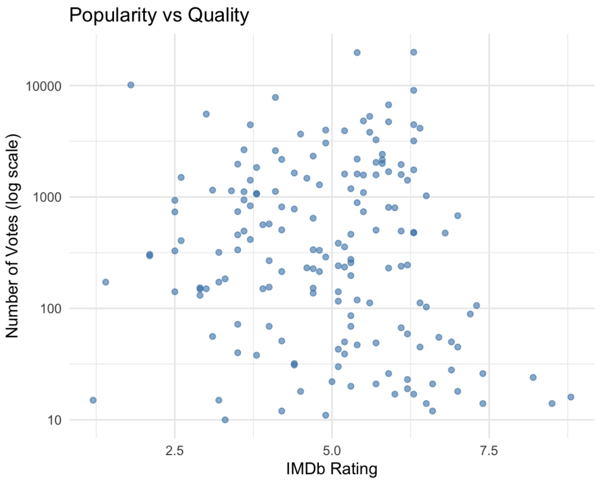

Figure 4: Popularity vs quality (IMDb rating vs log number of votes)

Figure 4 confirms a surprising but important insight: well-rated films are not always popular, and vice versa. The correlation between IMDb rating and vote count is near zero (≈ 0.03). This reinforces the idea that perceived quality (critical or fan reception) is only one piece of the puzzle. Popularity and reach are driven by other factors — possibly sub-genre appeal, platform visibility, or marketing efforts.

To explore those, we disaggregate vote counts across sub-genres and runtime formats. Boxplots in Figure 5 show that `Sci-Fi`, `Action`, and `Adventure` horror films tend to attract more viewers than, say, `Documentary` or `Family` variants. Likewise, excessively short or long films tend to underperform in vote counts.

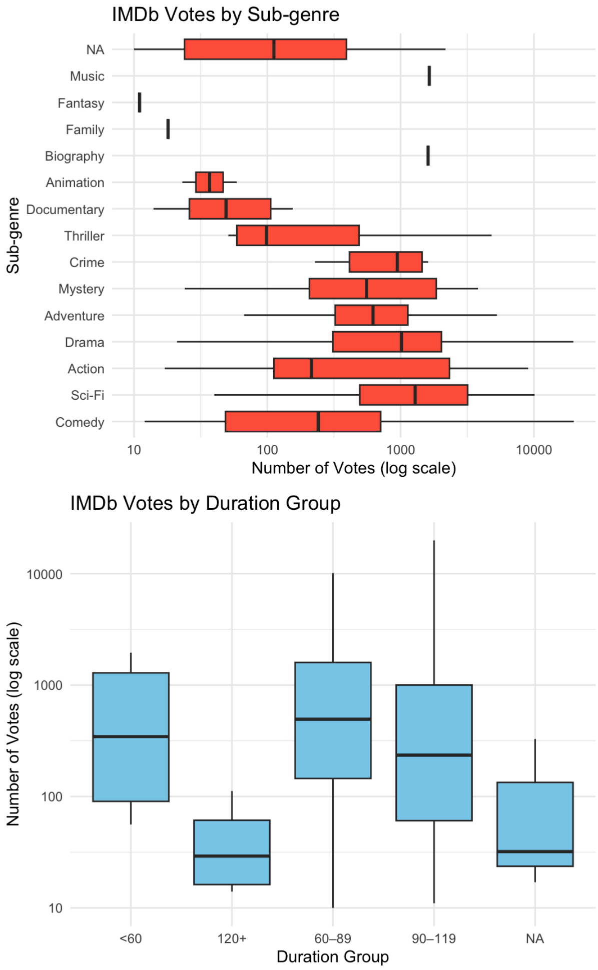

Figure 5: IMDb votes by sub-genre (top) and duration group (bottom)

This confirms what industry professionals often suspect: certain formats and genres are simply more bankable.

Additional figures provide further nuance. Figure 6 displays the distribution of duration groups among horror movies, revealing a strong concentration in the 60–89 minute range. This validates that most horror films follow a relatively compact format.

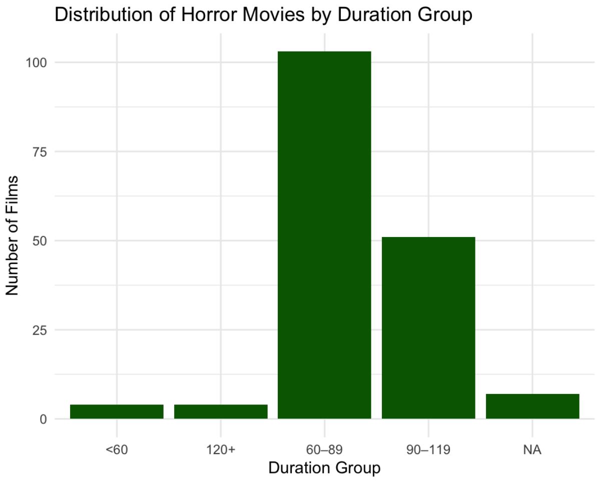

Figure 6: Distribution of horror movies by duration group

Moreover, average IMDb ratings vary significantly across sub-genres and duration groups. Figure 7 shows that longer films (120+ min) receive slightly better ratings, while short films (<60 min) show high variance. Ratings also differ by sub-genre — `Documentary` and `Animation` score higher, while `Crime` and `Thriller` show more mixed results.

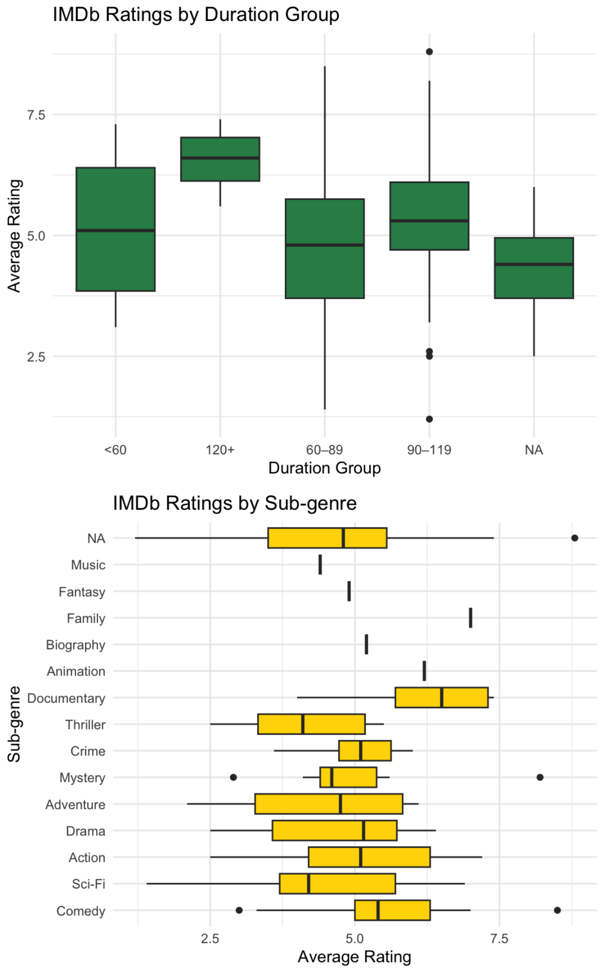

Figure 7: IMDb ratings by duration group (top) and sub-genre (bottom)

These figures enrich our understanding: while some sub-genres or formats yield higher viewer ratings, they may not necessarily drive visibility — a gap our predictive models aim to explain.

2.3 Linear Model and Coefficients

We first fit a multiple linear regression (OLS), using the formula above. The results show an adjusted R^2 of 0.19 — which means about 19% of the variation in popularity can be explained by the included features. In a domain as chaotic and creative as film production, this is a meaningful signal.

Here are selected findings from the model:

- **Sub-genres** like `Documentary`, `Animation`, and `Family` show negative associations with exposure — suggesting limited appeal in a horror context.
- **Sci-Fi** horror performs slightly better than the baseline, with a marginally positive coefficient.
- **Duration group 120+ minutes** has a significantly negative coefficient ($p = 0.03$), suggesting long horror movies may lose viewer engagement or suffer from distribution hurdles.
- Surprisingly, `average_rating` is not a significant predictor of exposure ($p \approx 0.96$).

For readers unfamiliar with R^2, here's an intuitive explanation: imagine we could "guess" a movie's popularity just from its genre, duration, and rating. Our guesses would be about 19% closer to the truth than simply taking the overall average vote count. Not bad — but not enough to rely on blindly.

2.4 Model Performance Comparison

We train Ridge, Lasso, and Random Forest models on a test set (20% of the data). Table 4 summarizes out-of-sample performance via Mean Squared Error (MSE):

Model	MSE (Test Set)
Ridge Regression	3.41
Lasso Regression	3.15
Random Forest	3.68

Table 4: Predictive error comparison across models

Although Lasso achieved the lowest raw MSE, its adjusted R^2 was highly negative (≈ -2.56), a sign of overfitting or unstable selection. Ridge regression provides a good compromise between simplicity and accuracy. Random Forest was expected to shine due to its flexibility, but it performed only marginally better in capturing extreme cases — perhaps limited by the small sample size (n = 169).

2.5 Residual Analysis and Model Diagnostics

Residuals — the difference between predicted and actual outcomes — are crucial to evaluating model quality. Ideally, residuals should be randomly distributed around zero, with no clear patterns or extreme deviations.

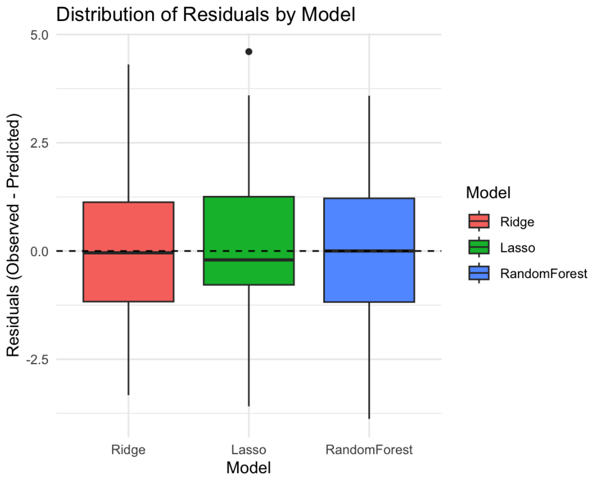

Figure 8: Distribution of residuals across models (Ridge, Lasso, RF)

As shown in Figure 8, all models suffer from considerable residual variance. This reflects the inherent unpredictability of success in the film industry. Some films will overperform (due to luck, memes, marketing virality), and some will flop — regardless of technical choices.

2.6 Limitations and Causal Inference

Causality. We must emphasize: these models reveal associations, not causes. A strong correlation between sub-genre and votes does not mean that simply labeling your movie "Sci-Fi Horror" will boost your audience. External variables (budget, platform, distribution) may drive both genre and exposure.

Unobserved Confounders. Many essential variables are missing from our dataset: marketing spend, star power, production studio, release date, and especially platform availability (Netflix? Shudder? Cinema?). These missing variables could distort the observed relationships.

Sample Size and Generalizability. With only 169 movies, our results are sensitive to outliers. A single popular movie can skew patterns. Thus, conclusions should be interpreted with humility.

Model Robustness. Ridge regression strikes a good balance: it performs better than naive OLS, avoids Lasso's instability, and is easier to interpret than Random Forests. Still, no model here exceeds 22–23% explained variance. This is a useful starting point, not a magic formula.

2.7 Conclusion

Our analysis confirms that strategic choices — such as sub-genre and duration — can moderately influence visibility. However, success remains partially unpredictable and influenced by factors beyond creative control. Data can guide us, but it cannot replace creative intuition or smart distribution strategy.

3 Strategic Recommendation: Producing a Successful Horror Film

3.1 Executive Summary

Drawing on our exploration of the horror movie landscape (Section I) and the statistical modeling of key performance factors (Section II), we now provide a strategy recommendation to maximize the potential for commercial success. Our approach is grounded in empirical evidence, guided by patterns in viewer behavior, and constrained by realistic production and marketing conditions.

In short: The data suggest that to maximize exposure — and, by extension, revenue — a horror film should aim for a tightly edited format (∼80 minutes), embrace a hybrid genre like `Horror-Comedy` or `Horror-Sci-Fi`, and avoid extremes in runtime or niche genres that show low engagement.

We break down this recommendation across four pillars: content, format, audience targeting, and distribution. We also highlight limitations and areas for further data collection.

3.2 Content Strategy: Which Story to Tell?

Our analysis shows that some sub-genres systematically attract more votes than others — a clear proxy for popularity and monetization potential. For example:

- **Horror-Comedy** and **Horror-Sci-Fi** dominate the dataset and receive higher median votes.
- Sub-genres like `Family`, `Documentary`, or `Animation` underperform significantly.

We thus recommend that **Grading to the Grave** be positioned as a **dark satire or workplace horror-comedy**, possibly with dystopian or surreal overtones to echo *Sci-Fi* tropes. This approach aligns thematically with the blurb and statistically with audience preferences.

Beyond statistical patterns, this genre choice also aligns with historical trends: hybrid horror films like *Shaun of the Dead* (Horror-Comedy) or *The Mist* (Horror-Sci-Fi) have proven cult appeal and strong word-of-mouth success despite modest budgets. These cases show that blending genres does not confuse audiences but instead invites broader demographics and repeat viewings.

Moreover, a satirical setting such as an academic or corporate environment leverages a universal cultural reference point — bureaucracy, burnout, and absurdity — which resonate across international audiences and lend themselves to dark humor. This aligns perfectly with the conceptual DNA of *Grading to the Grave*.

3.3 Format Strategy: Runtime and Structure

Our boxplots and regression models consistently show that excessively short (<60 minutes) and overly long (>120 minutes) horror films attract significantly fewer votes.

- The sweet spot lies between **60 and 89 minutes** — short-format features that are easier to distribute on streaming platforms and less costly to produce.
- Viewer engagement tends to decline with runtime: long horror movies may suffer from pacing issues or limited rewatchability.

We recommend a **runtime of 75 to 85 minutes** — compact, commercially viable, and consistent with current streaming trends. From a production perspective, shorter formats also allow for tighter shooting schedules, smaller crews, and greater agility in post-production — all of which reduce risk in low-budget filmmaking.

3.4 Platform and Distribution Targeting

While the dataset lacks detailed platform data, it is reasonable to assume that many highly rated and voted horror films found success on **streaming services**. These platforms favor content that is:

- easy to categorize and promote (e.g., "dark comedy" or "sci-fi horror"),
- suitable for casual viewing (<90 minutes), and
- capable of generating buzz with minimal marketing spend.

Given this, we recommend aiming for **direct-to-streaming distribution**, particularly on platforms like *Netflix*, or *Amazon Prime Video*. These outlets are more receptive to genre-blending and experimental horror formats.

In addition, these platforms often rely on algorithmic recommendation systems that benefit shorter, niche, and stylized content. Aligning the film with well-tagged tropes (e.g., "parody," "office horror," "satirical sci-fi") could enhance algorithmic discoverability and organic reach.

3.5 Audience Engagement and Marketing

Although we did not have access to marketing budget data, our strategy aims to **maximize exposure relative to cost**:

- A known actor or cameo could increase visibility disproportionately, especially if their screen presence can be leveraged in trailers or memes.
- A viral marketing campaign (e.g., campus horror screenings, faux faculty emails, or academic memes) could echo the movie's satirical theme and generate grassroots buzz.
- The script should be designed for **word-of-mouth virality** and meme potential. Short quotable lines, absurdist scenes, and referential humor are likely to resonate with online audiences.

This type of campaign favors organic traction over expensive ads and is well suited to student and young adult demographics — especially on platforms like TikTok, Reddit, and Letterboxd. **Summary Table — Recommended Strategic Choices:**

Production Dimension	**Recommended Strategy**
Sub-genre	Horror-Comedy with Sci-Fi or surreal elements
Runtime	75–85 minutes
Tone	Satirical, dark, workplace-based
Platform	Streaming-first (Prime, Netflix)
Budgeting	Micro-budget with marketing creativity
Star Power	Minor cameo or cult actor if feasible

Table 5: Table 5 — Strategic synthesis for film production

3.6 Limitations of the Recommendation

Despite the data-driven nature of our recommendation, we must acknowledge its limitations:

- **Absence of revenue and budget data:** Without actual financials, we rely on votes as a proxy for success — a reasonable but imperfect substitute.
- **Small sample size:** With only 169 observations, results are sensitive to outliers.
- **Missing causal variables:** Factors like platform placement, release timing, and marketing intensity remain unobserved but likely drive outcomes.
- **Potential survivorship bias:** The dataset may underrepresent failed or unreleased films, skewing results toward relatively successful entries.

Thus, we advise interpreting these findings as **strategic directions**, not deterministic rules. They should guide creative and business decisions without replacing qualitative judgment.

3.7 Further Data and Analysis Opportunities

To strengthen future recommendations, we would collect or integrate:

- **Budget and revenue data** (from Box Office Mojo, The Numbers, or production studios).
- **Streaming platform distribution**, release dates, and user demographics.
- **Social media virality metrics**, especially for recent horror hits.

Additional analysis could include:

- **Natural Language Processing (NLP)** of reviews and taglines to identify keywords linked to success.
- **Sentiment analysis** on ratings vs. written feedback.
- **Cluster analysis** to identify unobserved film archetypes or audience niches.

Such extensions would improve our understanding of how film content, structure, and external variables jointly influence exposure and performance.

3.8 Conclusion

In conclusion, the most effective production strategy for *Grading to the Grave* — and for maximizing commercial viability more broadly — lies in embracing audience-friendly formats (∼80 minutes), popular sub-genres (comedy, sci-fi), and a distribution-first mindset oriented toward streaming.

If paired with creative marketing and tight budget control, this strategy offers a realistic path to visibility and profitability in the fiercely competitive horror landscape. Ultimately, this data-informed strategy does not replace artistic vision; it simply provides a navigational compass in an unpredictable terrain. Creativity remains the core engine of any cinematic success — but when guided by insight, it is more likely to reach the audience it deserves.

Annexes

Supplementary Figures for Section I

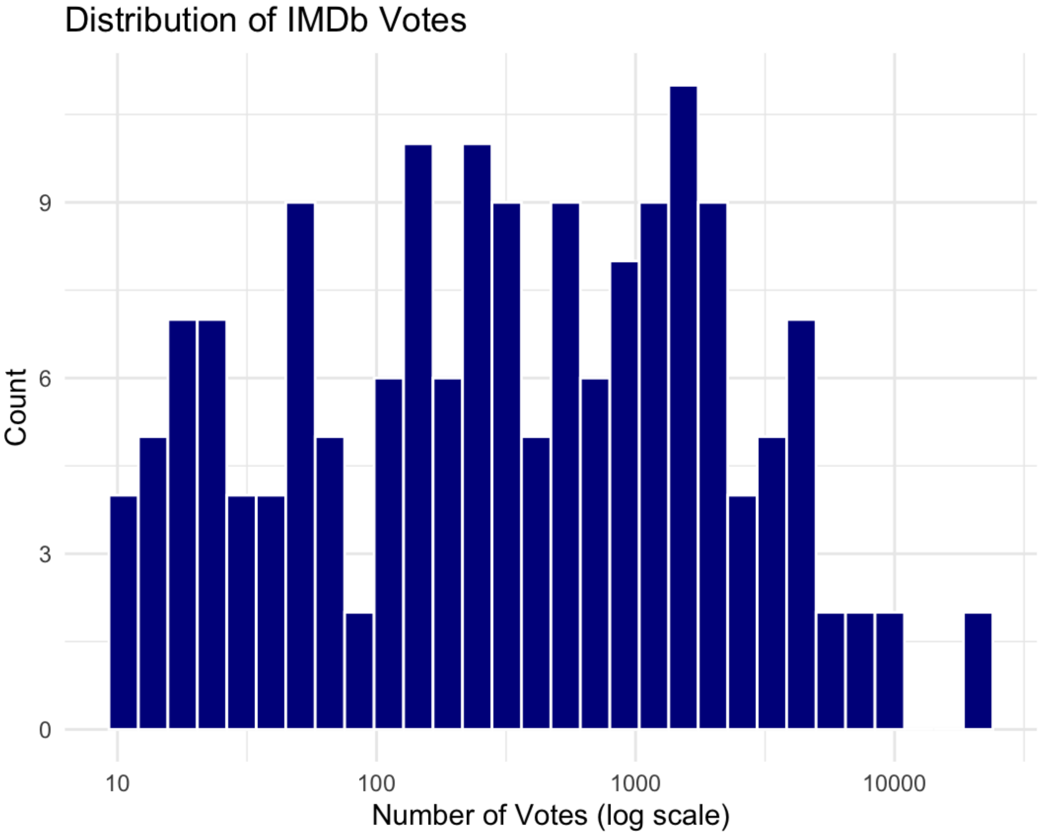

Figure A.1: Supplementary — IMDb votes by number of votes

References

[1] Jon Harmon. *Monster Movies Dataset – TidyTuesday (2024-10-29)*. Available at: https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-10-29/readme.md. Accessed May 1, 2025. Curated for the TidyTuesday project. 2024. URL: https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-10-29/readme.md.

[2] IMDb.com, Inc. *IMDb Dataset: name.basics.tsv.gz*. Available at: https://datasets.imdbws.com/name.basics.tsv.gz. Accessed May 1, 2025. Used with permission from IMDb. 2024. URL: https://datasets.imdbws.com/name.basics.tsv.gz.

[3] IMDb.com, Inc. *IMDb Dataset: title.basics.tsv.gz*. Available at: https://datasets.imdbws.com/title.basics.tsv.gz. Accessed May 1, 2025. Used with permission from IMDb. 2024. URL: https://datasets.imdbws.com/title.basics.tsv.gz.

[4] IMDb.com, Inc. *IMDb Dataset: title.crew.tsv.gz*. Available at: https://datasets.imdbws.com/title.crew.tsv.gz. Accessed May 1, 2025. Used with permission from IMDb. 2024. URL: https://datasets.imdbws.com/title.crew.tsv.gz.

[5] IMDb.com, Inc. *IMDb Dataset: title.principals.tsv.gz*. Available at: https://datasets.imdbws.com/title.principals.tsv.gz. Accessed May 1, 2025. Used with permission from IMDb. 2024. URL: https://datasets.imdbws.com/title.principals.tsv.gz.

[6] IMDb.com, Inc. *IMDb Non-Commercial Datasets*. Available at: https://developer.imdb.com/non-commercial-datasets/. Accessed May 1, 2025. Used with permission from IMDb. 2024. URL: https://developer.imdb.com/non-commercial-datasets/.