Once Upon A Time In Hollywood

Ahmet Alperen Güngör [*1] **Arda Gürsul** [*2] **Hagen Carstensen** [*3]

Abstract

The COVID-19 pandemic introduced a severe structural shock on the global film ecosystem, accelerating the shift towards digital platformization. This project[1] aims to quantify the persistence and interaction of these changes using data extracted from TMDb and IMDb across three temporal regimes: Pre-COVID (2017-2019), COVID-Shock (2020-2021), and Post-COVID Adjustment (2022-2024). We trained an XGBoost classifier to test the predictive power of our dataset and conducted statistical tests to quantify this shift from both industrial and viewer experience perspectives. As a result we identified features such as runtime and budget as primary indicators as well as further emphasizing the divergence between theatrical and streaming releases.

1. Introduction

The COVID-19 pandemic functioned as an unprecedented disruption for the global film industry: cinemas were closed, physical productions were halted, and release schedules were delayed or diverted to streaming services while audiences remained confined to their homes. While this immediate effect was visible, the deeper question is whether this shock fundamentally changed the industry's DNA. Did the pandemic only affect the industry for a short period, or did it trigger a permanent structural evolution in how movies are made and consumed?

This problem matters significantly because the film industry serves as an example for how creative sectors adapt to global crises. Understanding these shifts has both economic value for studios deciding where to allocate budgets and for understanding viewers' behaviour. Recent literature has highlighted that the simultaneous closure of exhibition halls and the decrease of physical production caused severe material losses on the industry, forcing a rapid and involuntary adaptation to digital practices (Rahmouni, 2023). This disruption served as a critical accelerator for the platformization of this sector, consolidating the economic dominance of US-based streaming giants who capitalized on global lockdowns to radically expand their subscriber bases (Vlassis, 2021). Consequently, this shift appears to have entrenched new consumer norms; the aggressive shortening of theatrical exclusivity windows from the traditional 90 days to as few as 17 days has conditioned a significant portion of the audience to bypass cinemas entirely in favor of home viewing, signaling a potentially irreversible change of the demand (Dermott, 2024).

In this paper, we aim to move beyond financial metrics to analyze the structural characteristics of the films themselves. We analyze this structural shift using a dataset extracted from TMDb and IMDb, segmented into three distinct temporal regimes: Pre-COVID (2017-2019), COVID-Shock (2020-2021), and Post-COVID Adjustment (2022-2024). As mentioned previously, the pandemic accelerated the adoption of streaming platforms as primary distribution channels, raising questions about whether this was a temporary response to cinema closures or a permanent realignment. By employing an XGBoost classifier to identify the primary features distinguishing these eras, alongside non-parametric statistical inference, we examine the changes between theatrical and streaming release strategies.

2. Data and Methods

The initial dataset was extracted using the TMDB API and IMDb Non-Commercial Datasets. These datasets were merged using the movies' IMDb IDs, resulting in a unified dataset of 23,998 movies. The movies were categorized into three time intervals based on their release dates: "Before COVID" (released on or before March 10, 2020), "During COVID" (March 11, 2020, to December 31, 2021), and "After COVID" (released on or after January 1, 2022). We

[*]Equal contribution [1]Matrikelnummer 7264397, MSc Machine Learning [2]Matrikelnummer 7415777, MSc Machine Learning [3]Matrikelnummer 7276378, MSc Computer Science. Correspondence to: AAG <ahmet-alperen.guengoer@student.uni-tuebingen.de>, AG <arda.guersul@student.uni-tuebingen.de>, HC <hagen-paul.carstensen@student.uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025.

[1]https://github.com/Ampelman123/Onceuponatimeinhollywood

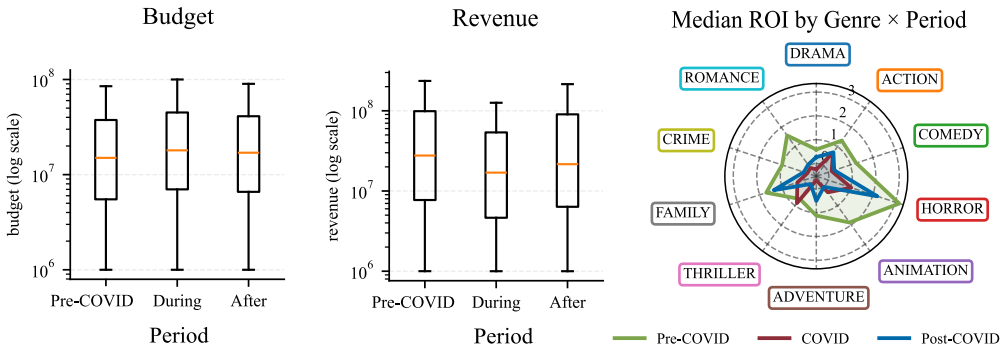


Figure 1. Distribution of budget and revenue across eras, illustrating the revenue contraction during the COVID-shock despite stable budgets.

defined and computed the Return on Investment (ROI) as:

$$\text{ROI} = \frac{\text{Revenue} - \text{Budget}}{\text{Budget}}$$

for a better understanding of the significance of the change in film industry economics by observing the effects of the changes in budget and revenue in a single variable. We simplified the dataset by parsing string metadata lists (e.g. production companies) into count features (0, 1, 2+) and collapsing high-cardinality categories (e.g., language of the movie) to top 10 + "Other". We used median imputation for data cleaning (e.g. NaN and inf values) strictly on the training set to prevent data leakage. We used XGBoost as an interpretable probe to test the dataset's predictive power for distinguishing intervals in structured tabular data. Hyperparameters (e.g., number of trees, depth, learning rate) were tuned with Optuna to maximize validation Gini score.

To quantify the shifts identified by the model, we adopt a dual-metric approach targeting both volumetric and distributional changes. First, we evaluate changes in genre market share using a two-sample Z-test for proportions:

$$Z = \frac{\hat{p}_{post} - \hat{p}_{pre}}{\sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_{pre}} + \frac{1}{n_{post}})}}$$

where n_{pre} and n_{post} represent total industry release counts. This metric identifies significant fluctuations in genre prevalence relative to total industry output. Second, we address the financial distribution. The revenue data exhibits extreme Pareto-like characteristics ($Gini = 0.86$), violating the normality assumptions required for parametric tests (e.g., t-tests). To detect shifts in the underlying revenue structure without assuming a specific distribution, we utilize the non-parametric Epps-Singleton (ES) test, which compares the empirical characteristic functions of the two distributions:

$$W = \left(\frac{1}{n_{pre}} + \frac{1}{n_{post}}\right)^{-1} (\mathbf{g}_{pre} - \mathbf{g}_{post})^T \hat{\mathbf{\Omega}}^{-1} (\mathbf{g}_{pre} - \mathbf{g}_{post})$$

Here $\mathbf{g}_{pre}$ and $\mathbf{g}_{post}$ denote the Empirical Characteristic Function vectors evaluated at standardized points, and $\hat{\mathbf{\Omega}}$ represents the estimated covariance matrix of the difference vector. The null hypothesis $H_0 : F_1(x) = F_2(x)$ posits that the revenue distributions are identical.

This approach is superior to mean-standardization for this dataset as it accounts for changes in the entire distribution shape.

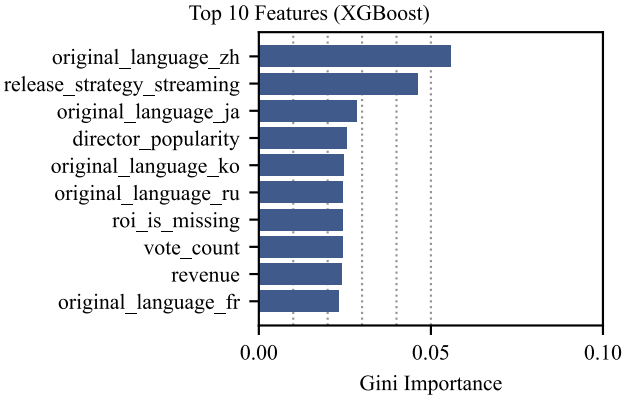
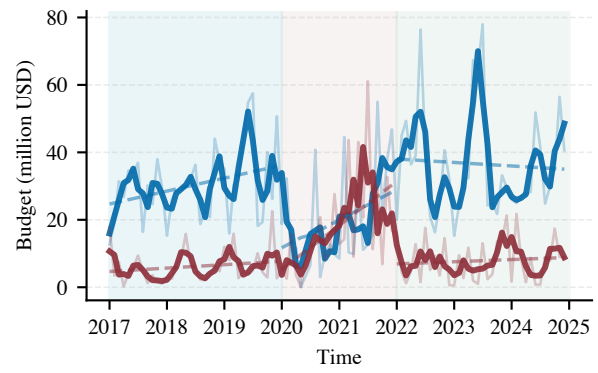


Figure 2. XGBoost feature importance for distinguishing Pre-, During-, and Post-COVID periods.

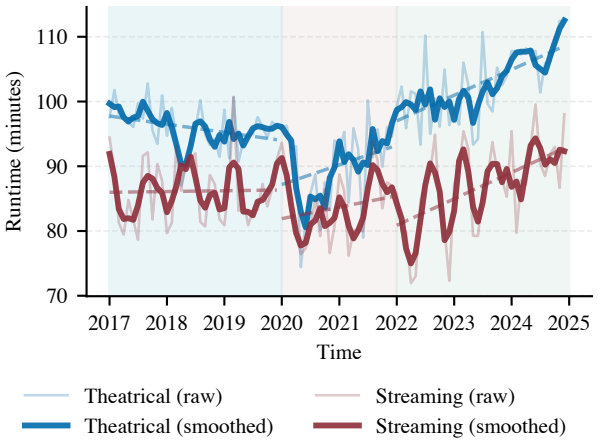
3. Results

The pandemic shock induced a visible systemic contraction in aggregate financial metrics. As illustrated in Figure 1, median theatrical revenue compressed significantly during the COVID-Shock period, declining by 42.8% (from $\$2.8 \times 10^7$ to $\$1.6 \times 10^7$), while production budgets remained statistically stable. Consequently, the ROI exhibited a uniform collapse across all genres, with no specific genre shielded from the downturn highlighting the immediate impact of theater closures and delayed releases. Post-pandemic metrics suggest a cautious recovery toward pre-2020 financial baselines, however, the interquartile range of revenue remains depressed compared to the 2017–2019 period.

The model achieved a Gini score of 0.5196 (AUC ≈ 0.76), indicating robust separation of the three periods from meta-



(a) Budget Evolution



(b) Runtime Evolution

Figure 3. Monthly evolution of (a) budget and (b) runtime for theatrical vs. streaming releases. Note the divergence during the shaded COVID period.

data. Figure 2 summarizes the XGBoost feature importances. Structural and market-based factors dominated: `original_language_zh` was most predictive, reflecting China's faster box-office recovery vs. the US, followed by `release_strategy_streaming`, capturing the shift to digital distribution.

Beyond aggregate metrics, theatrical and streaming releases responded to the shock with diverging structural adaptations, visualized in the time-series analysis in Figure 3. This figure displays monthly aggregated means for budget and runtime, with raw values plotted at reduced opacity and a 3-month moving average applied to reduce noise and clarify underlying trends (bold lines). Dashed linear trend lines fitted separately to each shaded period (pre-COVID, COVID, post-COVID) quantify the direction and magnitude of change within each era. All reported p-values are Bonferroni-corrected (p_{bonf}) to control the error rate across multiple comparisons. For each metric and release strategy, because we conducted three temporal comparisons (pre vs. COVID, COVID vs. post, and pre vs. post), correction factor is $m = 3$. The adjusted p-value is computed as $p_{\text{bonf}} = \min(1.0, p_{raw} \times m)$, where p_{raw} is the permutation test p-value for the mean difference. Theatrical runtimes followed a distinct "V-shaped" disruption-recovery pattern; the smoothed trendline dips sharply into the shaded COVID region ($95.8 \rightarrow 90.8$ min, $p_{\text{bonf}} = 0.0003$) before rebounding to pre-COVID baseline followed by a historic high of 102.5 min in 2024. In contrast, the streaming trendline remains flat throughout the shock ($p_{\text{bonf}} = 1.00$), visually confirming that the structural volatility was confined to the theatrical sector. A temporal inversion is further visible in budget dynamics: streaming budgets exhibited a distinct peak within the 2020–2021 window ($\$6.2M \rightarrow \$19.5M$, $p_{\text{bonf}} = 0.0003$), correlating with the period of cinema closures - indicating a temporary redirection of financial resources toward streaming production. This is consistent with broader observations of intensified platform competition and content investment (Lobato & Lotz, 2021), while theatrical budgets stagnated before a steep ascent in the 2022–2024 period ($\$35.0M$, $p_{\text{bonf}} = 0.0054$).

We also examined production infrastructure through the number of production companies per film and international co-production rates. The proportion of films involving multiple countries declined during COVID as cross-border collaboration became logistically challenging. This structural contraction in production networks provides additional evidence that the pandemic disrupted not only distribution channels but also the organization of film production itself.

The Epps-Singleton (ES) test reveals profound structural changes in genre economics, corroborating the trends in Figure 4. The most acute erosion of theatrical market share occurred in the high-budget Action and Adventure sectors, characterized by significant reductions in theatrical release volume ($Z = -3.014$ and $Z = -2.078$, respectively). In the Adventure genre, visual inspection of Figure 4 reveals that the *Non-Franchise Gap* —the revenue disparity between total and franchise releases—has nearly vanished post-2022, indicating that non-IP releases no longer contribute meaningfully to the genre's gross revenue. Action franchises exhibit a decoupling: while franchise release volume increased ($Z = 1.593$), aggregate revenue trended downward. The Epps-Singleton test confirms a fundamental shift in the Action revenue distribution ($W = 9.894, p = 0.042$), validating that increased tentpole volume has not yielded commensurate financial returns compared to pre-pandemic baselines.

In contrast, the Family and Animation sectors demonstrated comparative resilience, with revenue shares successfully regressing to pre-pandemic means following the shock period. For Animation, this stability was structurally driven by franchise performance ($W = 11.000, p = 0.026$). Figure 4 illustrates that while the broader genre distribution remained stable, the recovery tracks closely with the franchise contribution line, suggesting that the sector's post-pandemic

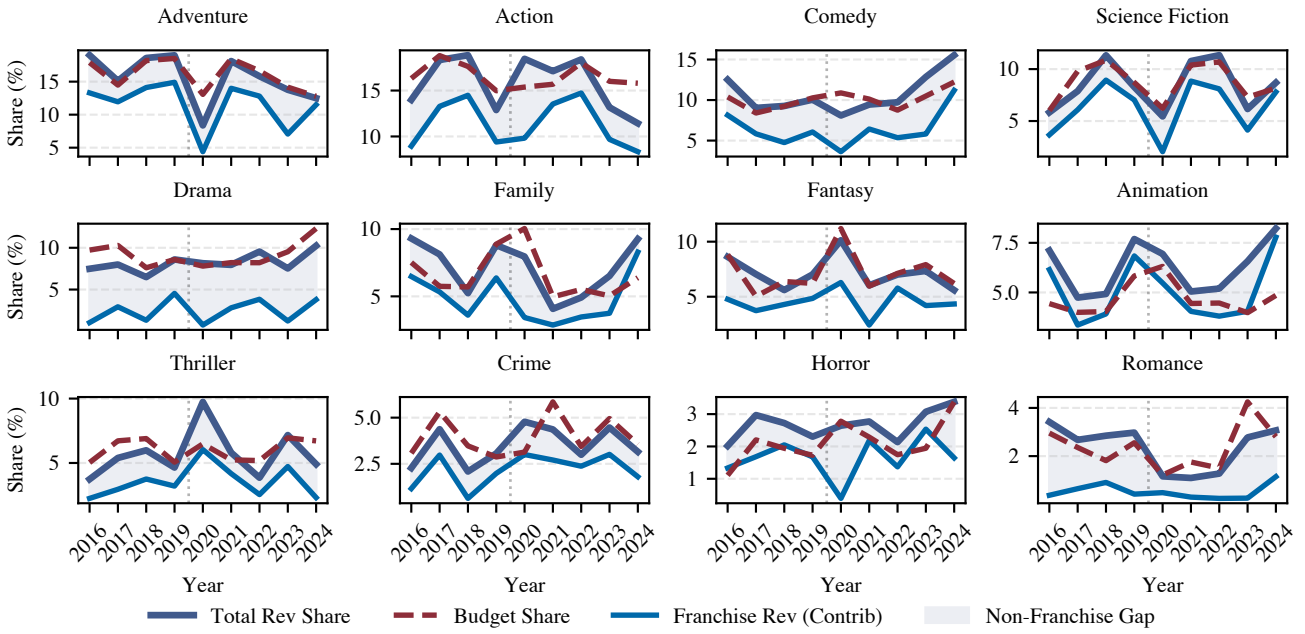


Figure 4. Longitudinal trends in revenue share vs. budget share by genre.

viability is anchored almost exclusively in utilising recognized Intellectual Property (IP) for movies.

The most extreme distributional shifts were observed in the *mid-budget* genres of Comedy ($W = 18.272, p = 0.001$) and Drama ($W = 12.295, p = 0.015$). Figure 4 contextualizes these shifts by decomposing revenue sources relative to budget allocation. Drama exhibits a persistent efficiency gap in the post-pandemic era: despite increased budget allocation, the revenue share shows only a marginal upward trend, failing to match expenditure. Comedy, conversely, displays a robust recovery, with revenue share outpacing budget allocation in the most recent period. Crucially, in both genres, the total revenue share is observed to track closely with franchise revenue contribution, indicating that the recovery in these traditionally original-led segments is correlated with an increased utilization of established IP.

4. Discussion & Conclusion

Our results suggest that the pandemic did not merely pause the industry but accelerated a bifurcation in content strategy. The *V-shaped* recovery of theatrical runtime and budget—overshooting pre-pandemic means—indicates a strategic pivot toward *Event Cinema*. Studios appear to be differentiating theatrical offerings from streaming content by investing in longer, more expensive spectacles. The null result for streaming runtime stability ($p = 1.00$) supports this, suggesting that digital platforms have maintained a consistent utility-maximizing format while theaters have been forced to upmarket their value proposition.

The divergence in genre economics indicates that the pandemic accelerated a structural bifurcation in theatrical strategy. The disappearance of the *Non-Franchise Gap* in Adventure and the franchise-driven stability in Animation reflect a defensive consolidation, where studios prioritize established Intellectual Property (IP) to mitigate risk. However, this strategy shows signs of saturation in the Action genre, where increasing franchise volume has yielded diminishing marginal returns, suggesting that scaling tentpole releases is no longer a guaranteed driver of growth.

In the *mid-budget* sector, the data reveals a significant displacement of original content. The tight correlation between franchise and total revenue in Comedy and Drama implies that theatrical viability in these genres is now conditional on IP attachment. For Drama, the persistent deficit between revenue share and budget allocation points to a model where theatrical releases may operate as loss leaders for downstream platforms rather than standalone profit centers. Consequently, the theatrical market appears to be consolidating around high-budget franchises, effectively ceding the traditional mid-budget segment to digital distribution.

We acknowledge that the reporting for financial data is often flawed, as movies tend to keep losses somewhat concealed. True streaming movies often lack financial data entirely and cinema movies now have a second financial livecycle in streaming, which is not reported. Though box office success seems to be a good proxy for streaming success aswell. We conclude the COVID-19 pandemic acted as a catalyst for a permanent structural realignment of the film industry. We conclude that the *Post-COVID* era is defined not by a return to 2019 norms, but by a new equilibrium: a theatrical sector dependent on high-budget, long-runtime franchise IP, and a streaming sector that has absorbed the mid-budget ecosystem.

Contribution Statement

Ahmet Alperen Güngör performed data sanity checks, data analysis and produced visualizations for release strategies and feature impacts. Arda Gürsul created the dataset, performed preprocessing and model creation, and produced visualizations of financial data. Hagen Carstensen analysed changes to genre composition and created the accompanying plot. All authors jointly wrote the text of the report.

References

Dermott, S. M. Analysis of film industry after the covid 19 pandemic. Masters thesis, National College of Ireland, 2024. URL https://norma.ncirl.ie/8483/.

Lobato, R. and Lotz, A. D. Beyond streaming wars: Rethinking competition in video services. *Media Industries Journal*, 8(1), 2021.

Rahmouni, L. The impact of covid-19 on the cinema industry. *ELWAHAT Journal for Research and Studies*, 16(1): 1084–1099, 2023. doi: 10.1177/0163443721994537.

Vlassis, A. Global online platforms, covid-19, and culture: The global pandemic, an accelerator towards which direction? *Media, Culture & Society*, 43(5):957–969, 2021. doi: 10.1177/0163443721994537.