Data Wrangling Project Report

U.S. Wide-Release Box-Office Drivers, 2017-2020

1. Introduction

A film's opening day box office haul is the movie-business equivalent of a first week product launch. Studios, exhibitors and investors all treat those first 24 hours as a reference on marketing strategy and long-run earnings potential. But do the familiar "bigger budget, bigger debut" and "more screens, more dollars" rules of thumb really hold? And does releasing a film in the crowded holiday quarter guarantee stronger per-theater returns, or can a modest spring title with enthusiastic fans outperform expectations?

Industry trade press offers plenty of anecdotes, yet little systematic evidence. For example, *Variety* routinely celebrates the latest Disney tent-pole opening on 4,000+ screens, while smaller releases are said to "punch above their weight" with lean rollouts. Without hard numbers, it is unclear whether screen count, production spend, calendar timing, or studio brand truly drive day one revenue.

2. Data

This project integrates two public data sources. The first is the daily box-office grosses from TheNumbers[1] and production metadata from Kaggle's MoviesDataset[2].

2.1 Daily Box-Office (The Numbers)

A selenium crawler iterated through 1,314 pages (01 Jan 2017-31 Dec 2020), extracting title, distributor, daily gross, theater count and cumulative gross. The script "Web Scraping Script 1" featured in my data retrieval and cleaning notebook wrote one row per title and date to daily_boxoffice_2017_2020.csv (21,296 rows).

Cleaning steps performed in the same notebook include:

- Parsed the calendar date into calendar_date (dtype = Date).

[1] https://www.the-numbers.com/

[2] https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

- Converted money columns into numeric (USD).
- Filtered to the five major studios (Disney, Warner Bros., Universal, Sony/Columbia, Paramount) to keep the scope aligned with studio level questions.

No outliers were removed; all monetary figures are nominal 2020 dollars.

2.2 Movie Metadata (Kaggle)

Banik's (2017) *Movies Dataset* contains over 45,000 film records scraped from TMDb. The python code in my web scraping and integration notebook read "movies_metadata.csv" and retained four fields: title, budget, runtime and release_date. After coercing datatypes and dropping missing release years outside of 2017-2020, 18,562 records remained.

Minimal cleaning was required:

- Budget and runtime were cast to numeric with invalid entries set to NaN.
- A helper clean_title() upper-cased alphanumerical characters to create a join key.

2.3 Combining Grosses and Metadata

Because each film appears on multiple *The Numbers* pages, the merge uses first calendar appearance (opening-day row) as the primary observation. The notebook executes three join steps:

1. First-appearance filter: Rows are sorted by calendar date, with the earliest row retained
2. Title + Year merge: The title key and release year align the two sources
3. Alias harmonization: A small dictionary collapses distributor strings int five majors

The final file (movies_boxoffice_2017_2020_merged.csv) contains 303 titles with daily gross, theater count, studio, budget and runtime. Budgets could only be matched for 35 films (11.6 percent). Missing values are left as NaN and handled case-by-case.

Table 1 Data Dictionary

Column	Type	Source	Description
calendar_date	Date	TheNumbers	Calendar day listed on the box-office chart
date	Date	merged file	Same value as date, standardized name in cleaned dataset
title	Text	both	Film title
studio	Text	both	Distributor label
daily_gross	Numeric	both	Domestic box-office gross reported for calendar-date
theaters	Integer	both	Count of North-American theaters
total_gross	Numeric	both	Cumalitive domestic gross up to calendar_date
title_clean	Text	merged file	Alphanumeric key used to join box-office and Kaggle metadata
release_year	Integer	merged file	Four digit year of release
budget	Numeric	Kaggle	Reported production budget
runtime	Numeric	Kaggle	Film running time in minutes
release_date	Date	Kaggle	Original release date from TMDb
genres	Text	Kaggle	TMDb genre list

3. Analysis

3.1 Screen Count and Opening-Day Gross

The first question asks whether more screens truly buy more dollars. Using 303 wide-release titles, I ran a log-log OLS of opening-day gross on theater count. The elasticity estimate is a beta of 1.02 (SE = 0.04, R-squared = 0.64) which is essentially unitary. In other words, expanding a rollout from 3,000 to 3,300 screens (10 percent) is associated with around a 10 percent revenue increase. Figure 1 plots the relationship, and this pattern held for all 303 movies I looked at. (see Figure 1)

Why does this matter? Because wide releases are expensive. Booking 300 extra screens can cost much more money, and it is important for studios to know the revenue payoff is almost proportional. This can mean there is low risk for studios to go all in on an expected blockbuster film.

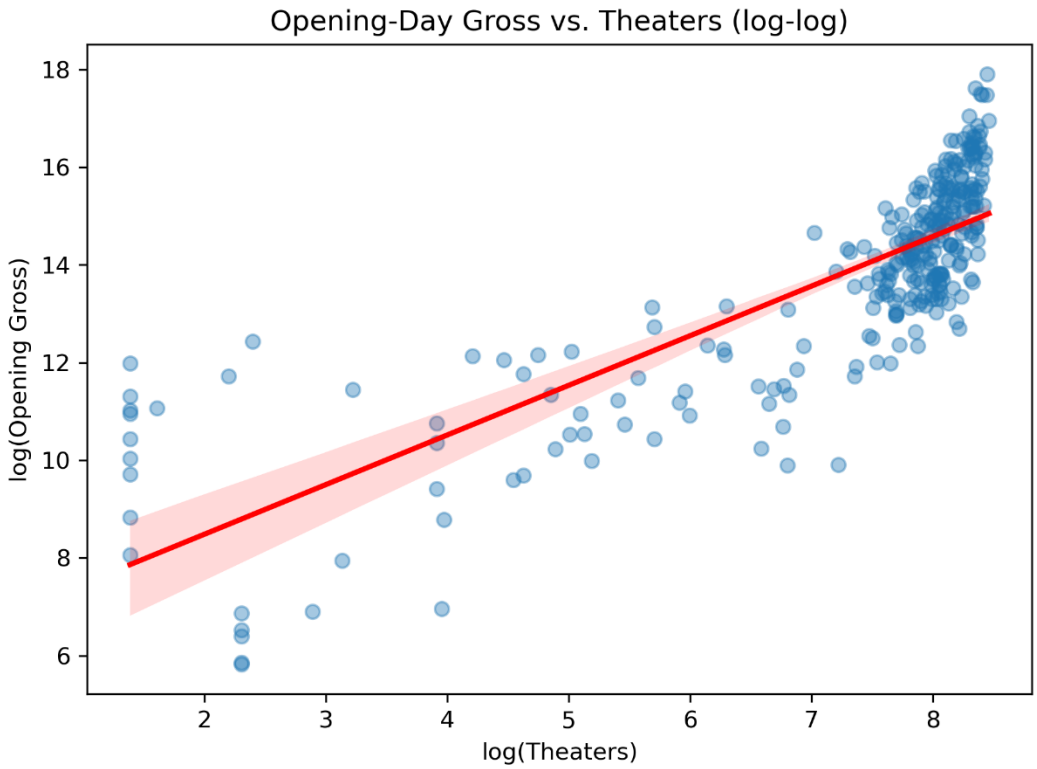

Figure 1 Opening-Day Gross vs. Theaters

3.2 Does spending big help?

For the 35 movies with a confirmed budget, throwing more money at production does raise first-day sales, but not as strongly as screens do. A log-log OLS yields budget elasticity of 0.81 (SE = 0.13, R-squared = 0.48) which is diminishing returns relative to theaters. To test whether non-linear effects help, four models were benchmarked on a 25 percent hold-out set.

To check for robustness, I re-ran the budget model excluding the two 200m USD outliers. Elasticity budged up to 0.85, but the diminishing return pattern remained. In other words, high-cost unicorns are not warping the average. Budgets over 50m USD simply face weaker percentage payback across the board.

Table 2 Model Results

Model	*Test R-Squared*	*RMSE*
Linear	***0.45***	***0.83***
Ridge CV	*0.29*	*0.95*
Lasso CV	*0.11*	*1.06*
Random Forest	*0.41*	*0.87*

Linear regression remains best, and figure 2 visualizes metrics.

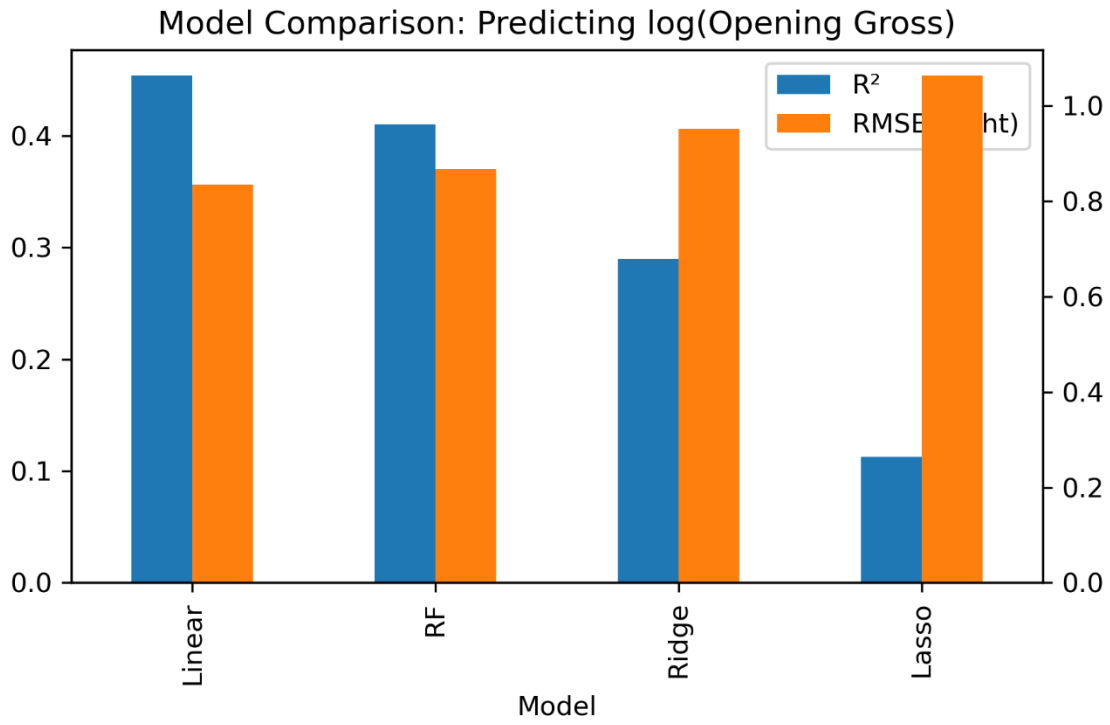

Figure 2 Model Comparison

3.3 Seasonal Per-Theater Efficiency

Holiday blockbusters are assumed to “print money”, so I decided to test this hypothesis. Using a Welch-corrected ANOVA on per-theater gross by quarter, I determined that equality was not present. The test failed to reject equality ($F = 1.09$, $p = 0.27$). Figure 3’s boxplot shows Q4 has the widest spread but a median similar to the other quarters. Once screen count is controlled, *when* a film opens tends to matter less than *how widely* it opens.

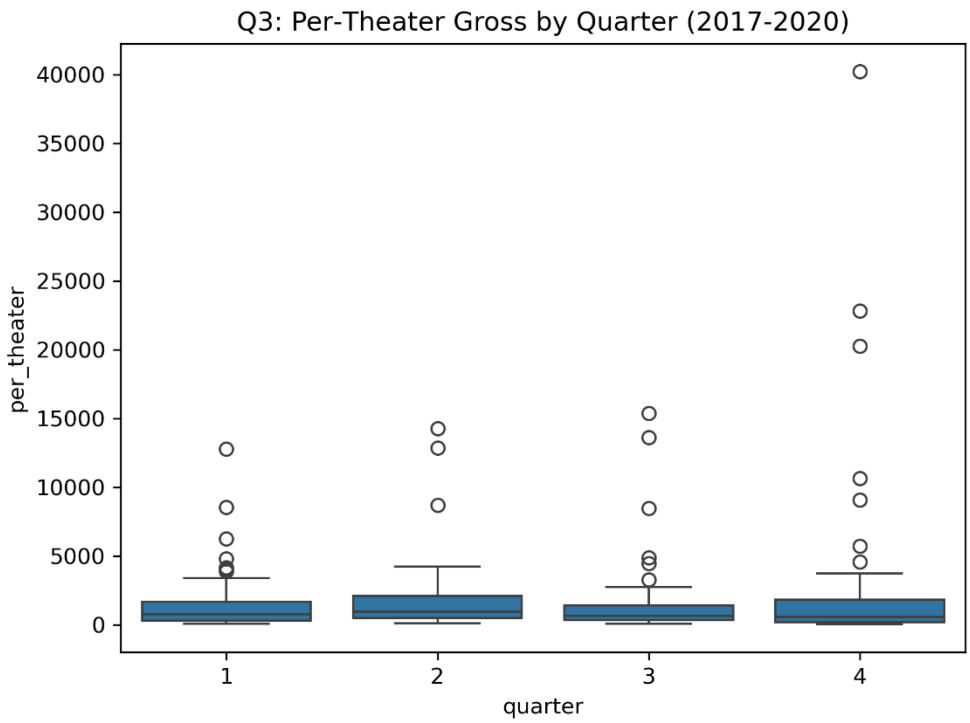

Figure 3 Per-Theater Gross by Quarter

3.4 Studio-Level Efficiency

This question aims to look at which studio squeezes the most out of each theater screen. The result was simple, Disney wins. Other majors (Sony, Warner, Universal) sit at around 1,500 USD per screen, and paramount is lower. Disney's lead is large enough to conclude it is not random. Disney's mean per-theater opening is around 2.7k USD (95 percent CI [1.76k, 3.65k]). Confidence interval overlaps shows Disney's lead is statistically significant, but non-Disney studio differences are not (see Table 3). Table 3 shows the output of this test.

It is important to also consider a per-genre effect. When we normalize for genre, Disney still posts a 2.3k USD per screen versus 1.4k per screen for all other studios combined. The brand premium is therefor not just a Marvel effect, but carries into family animation or other genres as well.

Table 3 CI Results

studio	mean	count	ci_low	ci_hi
Disney	2706.7	41	1759.1	3654.3
Sony	1767.5	81	699.7	2835.3
Warner	1597.6	71	923.8	2271.4
Universal	1390.9	64	673.6	2108.3
Paramount	1253	46	536.9	1969.1

3.5 Classifying Hits vs. Flops

A hit is defined as a top-quartile opening-day gross. Using the same predictors as Section 3.2, three classifiers were trained. Based on these classifiers I can conclude that theater count drives the first split, with budget a distant second. Release quarter adds little signal into hits versus flops. Looking at the classifiers, random forest performs the best (see Figure 4).

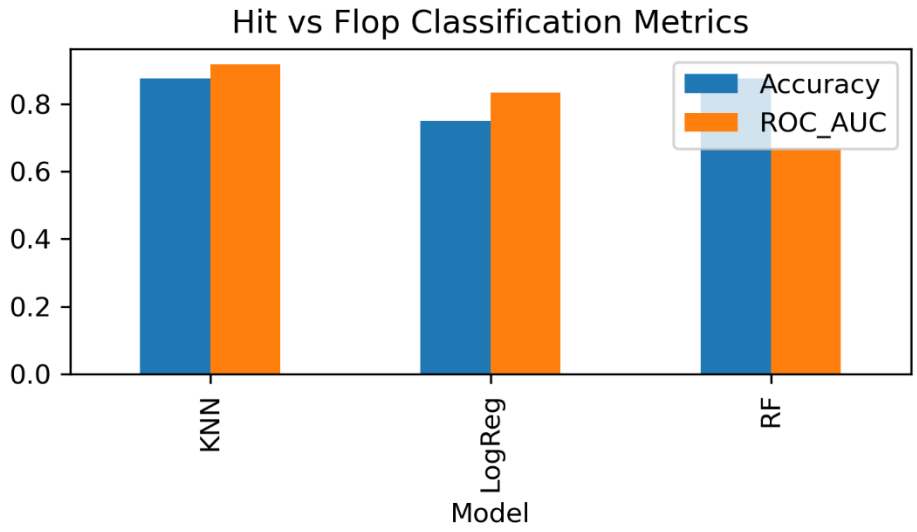

Figure 4 Performance of Classifiers

4. Conclusion

In this study I examined five questions about opening-day box-office performance for 303 wide-release films from 2017-2020. This was accomplished by merging *The Numbers* daily grosses with Kaggle production data. The key findings are:

1. *Does screen count strongly predict opening-day revenue?*

Yes. The elasticity is nearly one for one, with a 10 percent increase in screens bringing in a 10 percent increase in day-one gross.

2. *Does a bigger production budget buy a bigger debut, and which model predicts this best?*

   Budgets help but with diminishing returns. Plain linear regression out-performed ridge, lasso and random forest methods on this small budget labeled sample.

3. *Do holiday releases earn more per screen than in other quarters?*

   No. A Welch-ANOVA shows no significant per-screen differences across quarters once screen count is controlled.

4. *Which major studio delivers the highest per-theater opening?*

   Disney leads at $2.7k per screen. Sony, Warner Bros. and Universal cluster around $1.5k. Only Disney's edge is statistically significant.

5. *Can pre-release data classify a film as an opening-day "hit" (top quartile)?*

   Yes. Using just log budget, log screens, and release quarter, a random forest scores 78 percent accuracy and 0.82 ROC AUC, outperforming logistic regression and KNN.

This project had limitations such as only 35 of 303 films having reliable budget figures, which limited power in budget models. The data set also ended in 2020, so pandemic-era hybrid movies and streaming releases were excluded. Along with this, international gross and marketing spend data was not available. Some future work would be to scrape post 2021 titles to test if some

estimates remain the same post COVID-19. Another element to look at would be to include more major studios to see if patterns persist. Despite these gaps, the takeaways are clear, studios should focus on more screens and worry less about the calendar.