

Investigating Netflix Streaming Engagement Analytics

Djovan Velasco

July 13, 2026

Abstract

This project investigates Netflix platform analytics utilizing a semi-annual engagement database (July 2023 - June 2024). Through a combination of data extraction through SQL and statistical analysis in R, we evaluate content longevity, audience retention, catalog decay, and the impact of global licensing. The analysis reveals algorithmic advantages for shorter feature films, quantifies the significant premium of global distribution rights, and identifies TV series as the optimal format for international audience engagement.

1 Introduction

Utilizing a semi-annual engagement database spanning from late 2023 through mid-2024, this study applies a combination of SQL data querying and non-parametric statistical testing to decode the platform's viewership dynamics. Specifically, this analysis seeks to identify the optimal parameters for content acquisition, catalog longevity, and international distribution by investigating the following eight research questions:

- **RQ1: Relationship of Movie Release Date and Duration** Is there a statistically significant relationship between a movie's release date and its runtime?
- **RQ2: Audience Engagement and Series Retention Analysis** How does viewership volume distribute across the top-performing franchises, and what is the season-over-season retention rate?
- **RQ3: Maximum Chart Comparison** Do television series or feature films sustain longer cumulative durations on the Global Top 10 charts?

- **RQ4: Movie Runtime** Does the platform's engagement algorithm favor specific movie runtimes?
- **RQ5: Viewership S1 vs S2** What is the natural rate of audience view depreciation between the first and second seasons of a show?
- **RQ6: Viewership Decay** At what rate does viewership depreciate across consecutive reporting periods?
- **RQ7: Global vs Domestic Views** What is the quantifiable engagement advantage of securing worldwide distribution rights versus regional exclusivity?
- **RQ8: Global Preference** Among globally licensed content, which format achieves higher median viewership?

2 Methodology

Data was extracted from a local SQLite database(`netflixdb.sqlite`) containing tables for movies, TV shows, seasons, episodes, and semi-annual view summaries spanning late 2023 through mid-2024. Joins, CTEs, and window functions were utilized to extract data using the DBI and RSQLite packages. Results were subsequently saved as data frames in R which were used for analysis.

The viewership distributions were found to be dominantly right skewed. Since Netflix's catalog contains a massive volume of titles with low viewership, and a small cluster of viral outliers, non-parametric statistical methods were prioritized. Wilcoxon Rank-Sum tests and Kruskal-Wallis analyses were utilized to perform significance testing without the assumptions required by standard testing methods.

The individual methods used for each research question are listed below.

2.1 Relationship of Movie Release Date and Duration

To establish the baseline dataset for feature films, a query was executed on the movie table to extract the movie_id, title, release_date, and runtime. To ensure the data was valid for analysis in R, the query utilized a WHERE clause to filter out any records containing null values in either the release date or runtime fields.

Table 1: Movie Dataset Summary and Technical Metadata

movie_id	title	release_date	runtime
1	Damsel	1709856000000	110
2	Lift	1705017600000	107
3	Society of the Snow	1704326400000	146
4	Under Paris	1717545600000	104
6	Mother of the Bride	1715212800000	90

The table was saved as a data frame and processed to convert the data to a YYYY/MM/DD format. A linear model was then generated to investigate if there was a relationship between release_date and runtime.

2.2 Audience Engagement and Series Retention Analysis

First, the tv_show, season, and view_summary tables were joined to aggregate total platform views per series. To isolate the top-performing franchise, CTE was constructed utilizing the RANK() OVER window function, partitioned by total descending views. The outer query then joined the ranked data back to the season table, extracting the sequential season_number and corresponding total_season_views exclusively for the number-one ranked show, preparing the data for period-over-period retention calculations in R.

Table 2 details the overall total views accumulated by the platform’s top ten most-watched series.

Table 2: Top 10 Series by Overall Platform Views

title	overall_total_views
CoComelon	315,000,000
Young Sheldon	194,100,000
Bridgerton	178,600,000
Gabby’s Dollhouse	172,600,000
Suits (2011)	161,000,000
Peppa Pig	139,700,000
Lottie Dottie Chicken	114,900,000
Brooklyn Nine-Nine	114,500,000
Little Angel	111,000,000
Lupin	109,600,000

Table 3 breaks down performance across all ten seasons of the top-performing title, *CoComelon*.

Table 3: *CoComelon* Viewership Distribution by Season

title	season_number	total_season_views
CoComelon	1	51,400,000
CoComelon	2	41,400,000
CoComelon	3	36,800,000
CoComelon	4	33,600,000
CoComelon	5	23,700,000
CoComelon	6	32,200,000
CoComelon	7	24,100,000
CoComelon	8	50,900,000
CoComelon	9	11,100,000
CoComelon	10	9,800,000

Finally, to map viewership decay and audience fatigue, season-over-season retention rates were derived. Table 4 measures the proportion of the audience carried over from the previous season.

Table 4: *CoComelon* Season-over-Season Retention Metrics

title	season	total_views	prev_views	rate	pct
CoComelon	2	41,400,000	51,400,000	0.8054	80.5%
CoComelon	3	36,800,000	41,400,000	0.8889	88.9%
CoComelon	4	33,600,000	36,800,000	0.9130	91.3%
CoComelon	5	23,700,000	33,600,000	0.7054	70.5%
CoComelon	6	32,200,000	23,700,000	1.3586	135.9%
CoComelon	7	24,100,000	32,200,000	0.7484	74.8%
CoComelon	8	50,900,000	24,100,000	2.1120	211.2%
CoComelon	9	11,100,000	50,900,000	0.2181	21.8%
CoComelon	10	9,800,000	11,100,000	0.8829	88.3%

2.3 Maximum Chart Comparison

To evaluate longevity, the maximum number of cumulative_weeks_in_top10 was extracted for each title. Because movies and TV shows are stored separate tables, two distinct CTEs were engineered to aggregate the maximum charted weeks for each format. A column (format_type) was injected into both CTEs. The datasets were then vertically stacked using a UNION ALL operator, resulting in a single, unified dataframe comparing the maximum weeks charted alongside a standardized format identifier.

Table 5 highlights the top five titles with the highest longevity, categorized by their format type.

Table 5: Top 5 Titles by Maximum Weeks Charted

title	format_type	max_weeks
KPop Demon Hunters	Movie	50
Squid Game	TV Show	32
Stranger Things	TV Show	29
Wednesday	TV Show	28
Ms. Rachel	TV Show	25

The subsequent data frame was then saved in R and used to perform a Wilcoxon Rank-Sum Test.

2.4 Movie Runtime

To test for engagement preferences regarding movie length, continuous runtime data was transformed into standardized categorical fields. A CTE first isolated the maximum charted weeks for all valid movies. The main query then utilized a conditional CASE WHEN statement to assign each title to one of four discrete categorical buckets based on runtime(Short, Standard, Long, or Very Long).

Table 6: Sample Classification of Movie Runtime Buckets

title	max_weeks	runtime_bucket
"Sr."	NA	1. Short (<90m)
#Alive	NA	2. Standard (90-119m)
#AtFirstSight	NA	2. Standard (90-119m)
#FriendButMarried	NA	2. Standard (90-119m)
#FriendButMarried 2	NA	2. Standard (90-119m)

After querying the data was used to perform a Kruskal-Wallis rank sum test.

2.5 Viewership S1 vs S2

Comparing the performance of a first and second season requires a paired statistical test, which required data to be formatted wide rather than long. To achieve this, a HAVING MAX(season_number) = 2 clause was used to strictly isolate series that ended after two seasons. The data was then pivoted directly within the SQL query using conditional MAX(CASE WHEN ...) aggregations, transforming the row-based season data into two distinct columns (season_1_views and season_2_views) tied to a single series title.

Table 7: Viewership Comparison: Season 1 vs. Season 2

title	season_1_views	season_2_views
100 Days to Indy	300,000	100,000
3Below: Tales of Arcadia	2,600,000	1,800,000
7SEEDS	500,000	300,000
A Clean Sweep	100,000	100,000
A Killer's Mistake	200,000	300,000

2.6 Viewership Decay

To track viewership decay across distinct reporting periods, views were separated based on their start_date. Conditional SUM(CASE WHEN ...) aggregations were utilized to evaluate timestamps, splitting the view counts

into two columns: h2_2023_views, timestamps before 1704067200000 (Jan 1 2024) and h1_2024_views (timestamps after). After using UNION ALL to merge movies and TV shows, a strict WHERE filter was applied to ensure only titles that generated greater than zero views in both periods were retained.

Table 8: Sample Viewership Across Consecutive Reporting Periods

title	format_type	h2_2023_views	h1_2024_views
"Sr."	Movie	200,000	100,000
#Alive	Movie	4,600,000	4,300,000
#FriendButMarried	Movie	100,000	100,000
#FriendButMarried 2	Movie	100,000	100,000
#Manhole	Movie	300,000	200,000

The data was saved to plot in R and visualize the distribution of audience viewership decay rate over the two periods.

2.7 Global vs Domestic Views

Titles were evaluated based on their boolean available_globally flag. After stacking all titles and calculating their total historical views via CTEs, a CASE WHEN statement was used to standardize the boolean flag into a readable licensing_model category ('Global' or 'US_Exclusive'). The final query performed the mathematical aggregations, utilizing COUNT(*) and AVG(total_views) grouped by the licensing model to output the final comparative statistics.

Table 9: Average Viewership by Licensing Model

licensing_model	total_titles_charted	average_views
Global	7,409	8,292,496
US_Exclusive	1,902	3,407,505

2.8 Global Preference

To determine which format performs best internationally, the dataset was filtered to include only titles with worldwide rights (available_globally = 1). Total views were aggregated for both movies and TV shows (summing across all distinct seasons for TV). The global titles were then joined to their respective view counts, and a standardized format_type column was included to allow for a direct comparison between global preference for movies or tv.

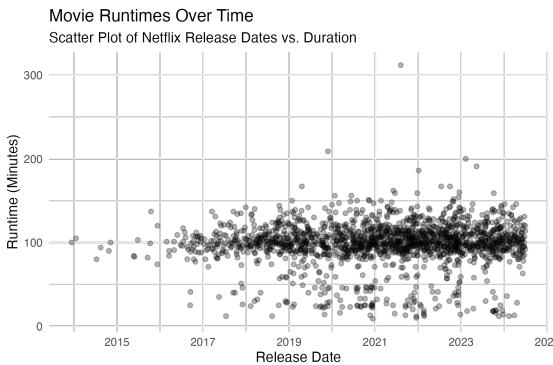
Table 10: Top Globally Licensed Titles by Format and Viewership

title	format_type	total_views
Lift	Movie	529,000,000
Bridgerton	TV Show	529,000,000
Squid Game	TV Show	465,600,000
The Beautiful Game	Movie	393,700,000
Stranger Things	TV Show	393,700,000

3 Results

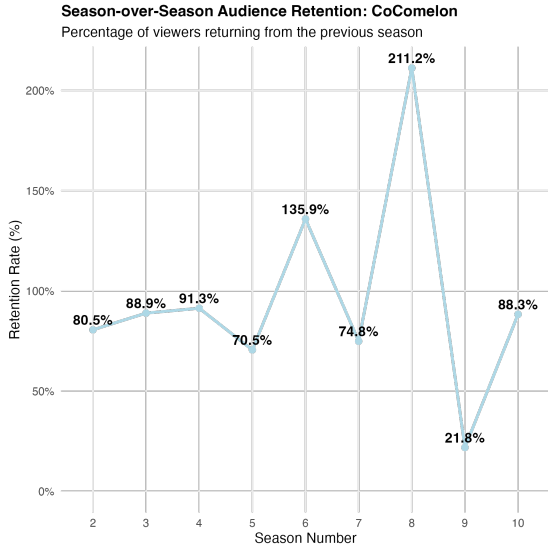
3.1 Relationship of Movie Release Date and Duration

A linear regression model was fitted to evaluate the relationship between a movie’s original release date and its runtime. The resulting model yielded a negligible R^2 value and a high p -value ($p > 0.05$), indicating no statistically significant correlation. The data confirms that feature length remains highly variable and independent of the release date.



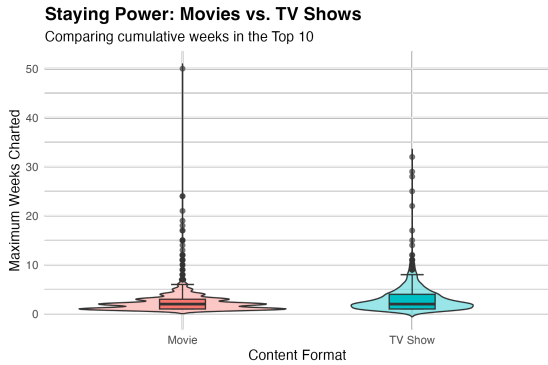
3.2 Audience Engagement and Series Retention Analysis

Season-over-season retention analysis reveals significant volatility. While baseline retention between standard seasons hovers between 70% and 90%, specific installments (such as Season 8) experienced massive viewership spikes exceeding a 200% retention rate. This suggests that engagement is highly susceptible to external algorithmic promotion, catalog placement, or viral resurgence.



3.3 Maximum Chart Comparison

A Wilcoxon Rank-Sum test was conducted to compare the cumulative weeks spent in the Global Top 10 between movies and television shows. The analysis confirms a statistically significant difference in staying power, with television series exhibiting a higher median (3.2 weeks) charting duration than feature films (2.6 weeks). While individual movie outliers exist, TV shows tend to yield higher engagement.



3.4 Movie Runtime

A Kruskal-Wallis rank-sum test evaluated median charting longevity across four distinct runtime categories. The test yielded statistically significant variance among the groups. Subsequent pairwise Wilcoxon testing confirmed the hypothesis that “Short” feature films (< 90 minutes) possess a distinct engagement advantage over all other film lengths.

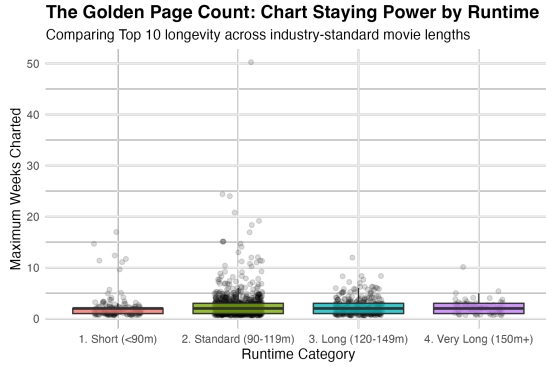


Table 11: Pairwise Wilcoxon Test p -values by Movie Runtime

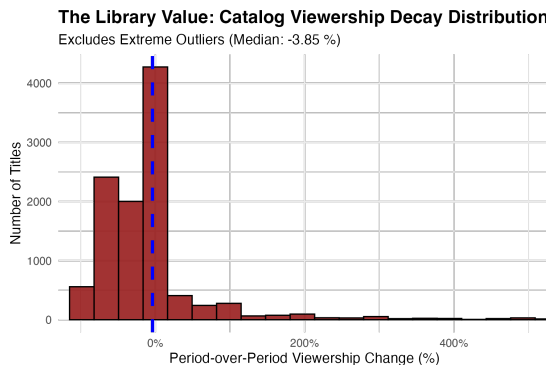
	1. Short (<90m)	2. Standard (90-119m)	3. Long (120-149m)
2. Standard (90-119m)	0.00019	-	-
3. Long (120-149m)	0.01324	0.21240	-
4. Very Long (150m+)	0.00465	0.95981	0.50390

3.5 Viewership S1 vs S2

For series concluding after two seasons, a paired Wilcoxon Signed-Rank test was utilized to measure audience attrition. A significant p -value revealed that the median views for season 2 were distinctly different and lower than season 1 views, with season 2 median at 700,000 and season 1 at 900,000.

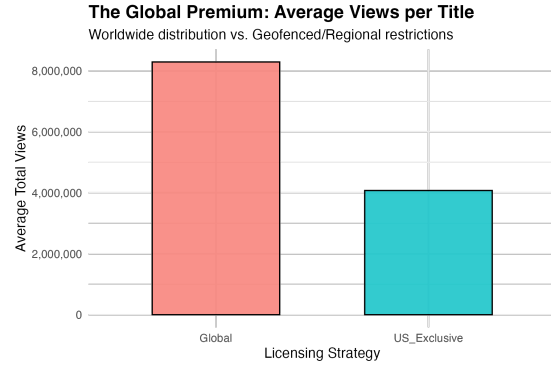
3.6 Viewership Decay

Analysis of titles active across consecutive reporting periods (H2 2023 and H1 2024) reveals the natural depreciation of viewership. After excluding extreme viral outliers (titles exhibiting period-over-period growth $> 500\%$), the platform's core catalog exhibited a standardized median decay rate. The heavily right-skewed distribution highlights that while the majority of older content steadily loses viewership, a select fraction of the library reliably maintains or grows its audience over time.



3.7 Global vs Domestic Views

The evaluation of licensing reveals a massive premium associated with worldwide distribution. Globally licensed titles generated an average of approximately 8.29 million views per title. In stark contrast, content restricted exclusively to the United States averaged only 4 million views. The data proves that regional geofencing statistically limits a title's average viewership.



3.8 Global Preference

Focusing exclusively on the 7,409 globally licensed titles, a Wilcoxon Rank-Sum test evaluated which format achieves deeper cross-cultural penetration. The statistical output confirms that television series generate significantly higher median viewership than feature films on an international scale.

Table 12: Summary Statistics: Global Views by Format

format_type	median_views	mean_views
Movie	1,600,000	7,879,781
TV Show	2,500,000	8,591,228

4 Discussion

The findings of this project provide insights into the mechanics of streaming platform engagement, revealing how content format, runtime, and licensing structures dictate platform success.

First, the data unequivocally demonstrates the financial and engagement necessity of global distribution rights. The analysis of licensing revealed that regional geofencing severely caps a title's potential yield, costing an average of over 4 million views per title compared to globally available counterparts. Furthermore, when evaluating how to allocate these global licensing budgets, the data strongly favors serialized television over feature films.

Second, our Kruskal-Wallis analysis of feature films proved that “Short” movies (under 90 minutes) possess a statistically significant engagement advantage. Because shorter films can be produced faster, and chart longer, it is favorable these films to be produced more for long term success.

Finally, the analysis of catalog decay and audience retention highlights the volatility of streaming libraries. While the season 2 slumps and standardized catalog decay confirm the natural aging of content, the extreme positive outliers—such as the massive 200% period-over-period viewership spike in *CoComelon* prove that streaming libraries do not experience strict linear decay. Instead, engagement is highly susceptible to external factors such as algorithmic resurrection and viral trends.

Acknowledgements

I would like to thank Dr. Michael Tsiang for his guidance and support throughout the course, teaching me the skills necessary to put this project together.