

|  |
| --- |
| Predictive Modeling |
|  |
| JULY 13  FINAL PROJECT  Authored by : SHALEM DANDALA |



1. INTRODUCTION

With the rapid shift in viewer behavior towards on-demand digital content, OTT platforms like *ShowTime* face increasing pressure to maximize engagement from day one of content release. This report explores the platform's internal data to understand what influences first-day content viewership. Using a combination of exploratory data analysis (EDA) and linear regression modeling, we identify key drivers such as platform traffic, trailer views, ad impressions, genre, and release timing (day of week and season). The insights derived from this analysis aim to guide marketing efforts, improve scheduling strategies, and inform content planning to boost early viewership and long-term success on the platform

1. PROBLEM STATEMENT

As the OTT (Over-The-Top) entertainment industry continues to expand rapidly, understanding viewer behavior has become critical for platforms aiming to maintain competitive advantage. *ShowTime*, a digital content provider offering movies and web shows, has observed significant variations in first-day viewership across its newly launched content. These fluctuations can be influenced by several internal and external factors such as marketing intensity, platform traffic, content genre, trailer engagement, timing of release (weekday or season), and competition from major events like sports broadcasts.

To address this challenge, *ShowTime* seeks to identify the key drivers that significantly impact first-day content viewership. The goal is to use existing data related to user traffic, advertisement performance, content metadata, and external events to uncover patterns that can explain the variations in viewership.

This project leverages statistical methods and predictive modeling—specifically linear regression analysis—to:

* Understand how individual factors affect day-one viewership,
* Quantify the strength and direction of these relationships,
* And ultimately develop a model that can predict first-day performance of future content.

The insights derived from this analysis will help ShowTime make data-driven decisions regarding:

* Content release planning (e.g., choosing optimal days/seasons),
* Marketing spend allocation (e.g., ad impressions vs. trailer views),
* And content strategy optimization (e.g., genre performance trends).

By identifying and prioritizing the most influential variables, ShowTime aims to maximize content visibility, enhance user engagement, and improve ROI on promotional activities from the very first day of release.

1. DATA DESCRIPTION

The dataset provided by *ShowTime* contains detailed information about content released on the OTT platform, along with several influencing factors believed to affect first-day viewership. Each row in the dataset represents a unique content release (movie or web show), and each column corresponds to a feature that may impact its performance on day one.

Variables:

* visitors: Average number of visitors (in millions) to the platform in the past week
* ad\_impressions: Ad impressions (in millions) across all campaigns for the content
* major\_sports\_event: Presence of a major sports event on release day (0 = No, 1 = Yes)
* genre: Genre of the content
* dayofweek: Day of release
* season: Season of release
* views\_trailer: Number of trailer views (in millions)
* views\_content: Number of first-day views (in millions) — Target Variable

1. Exploratory Data Analysis (EDA)

A thorough Exploratory Data Analysis (EDA) was conducted to understand the structure, distribution, and key patterns in the OTT content dataset. The following sections summarize the main findings with supporting visualizations.

* 1. Distribution of First Day Content Views

The histogram above illustrates the distribution of first-day content views across all titles on the platform. Most content receives between 0.35 and 0.55 million first-day views, as indicated by the peak around this range. The distribution is slightly right-skewed, meaning there are a small number of titles that achieve notably higher viewership on their first day (beyond 0.7 million views). However, such high-performing titles are relatively rare.

This pattern suggests that while blockbuster titles with exceptional initial engagement do exist, the majority of content achieves moderate and consistent first-day viewership. The smooth curve overlay (KDE) further emphasizes the central concentration of values and the gradual tapering towards higher view counts.

Business implication:

Understanding this distribution helps set realistic expectations for new releases and highlights the importance of identifying factors that can push content into the higher-viewership segment.

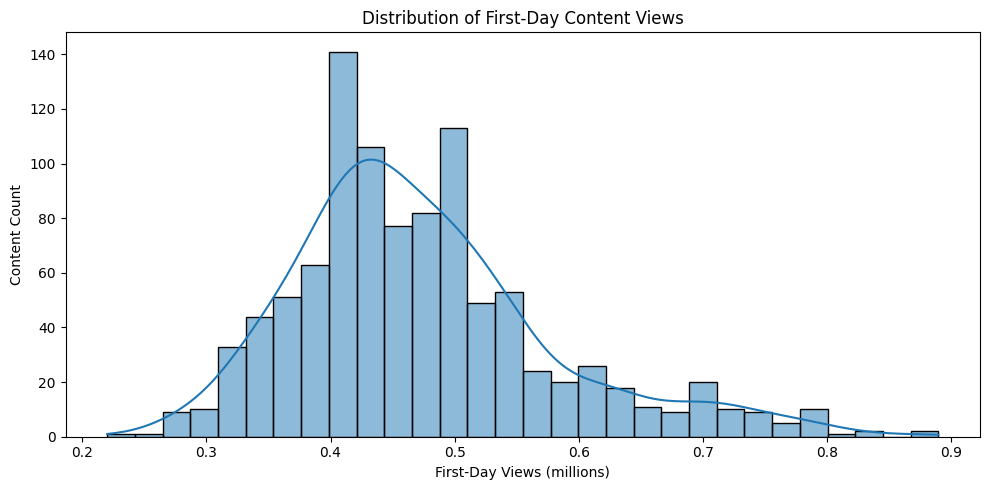


Figure 1 : Histogram of First-Day Content Views

* 1. Distribution of Genres

This horizontal bar chart displays the distribution of content across different genres on the platform. The genre "Others" has the highest count, indicating that a significant portion of content does not fit into the main predefined genres. Among the named genres, "Comedy," "Thriller," and "Drama" are the most prevalent, each with just over 100 titles. Genres such as "Romance," "Sci-Fi," "Horror," and "Action" have a similar number of titles, each slightly above 100 as well.

Business Interpretation:

The dominance of the "Others" category suggests that the platform hosts a diverse range of content that may not fit neatly into traditional genre categories, which could present both opportunities and challenges for targeted marketing and content recommendation. The relatively balanced distribution among the main genres indicates a broad content strategy, appealing to a wide range of viewer preferences.

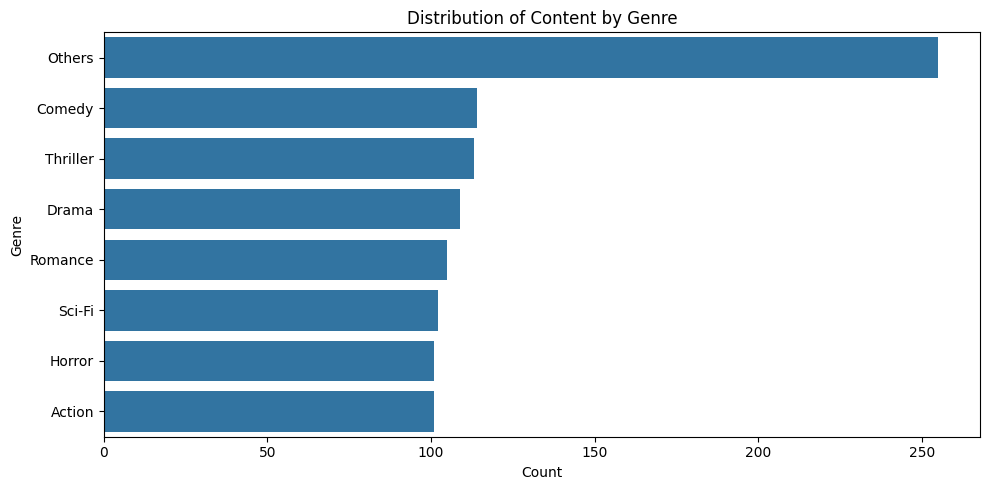


Figure 2: Distribution of Content by Genre

* 1. Viewership by Day of Release

This boxplot compares the distribution of first-day content views for each day of the week. The central line in each box represents the median viewership, while the height of each box shows the interquartile range (middle 50% of the data). Outliers are shown as individual points above each box.

Across the week, the median and typical range of first-day views are relatively similar, but there are some notable differences in the spread and outliers. For example, Friday, Saturday, and Sunday releases tend to have a higher concentration of outliers, indicating that some titles launched on weekends can achieve exceptionally high first-day viewership. This suggests that while average performance is consistent throughout the week, weekends may offer a greater chance for standout success.

Business Interpretation:

Releasing content on weekends could be beneficial for maximizing the potential of blockbuster releases, though overall average performance remains stable across all days. Scheduling high-profile launches on Fridays, Saturdays, or Sundays may help capitalize on higher audience availability and engagement.

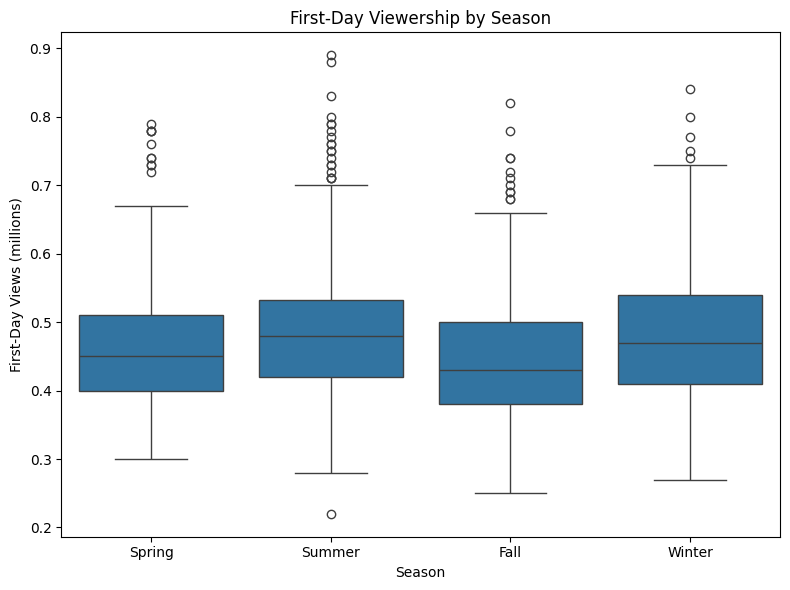


Figure 3: First-Day Viewership by Day of Release

* 1. Viewership by season

This boxplot compares the distribution of first-day content views across the four seasons: Spring, Summer, Fall, and Winter. For each season, the box represents the interquartile range (middle 50% of the data), the central line shows the median, and outliers are displayed as individual circles above the whiskers.

From the plot, it is evident that Summer and Winter releases tend to have slightly higher median first-day viewership compared to Spring and Fall. The spread of data (box height) and the number of outliers are also greater in Summer and Winter, indicating a wider range of performance and some exceptionally popular releases in these seasons. Spring and Fall show lower medians and fewer high-performing outliers.

Business Interpretation:

Content released during Summer and Winter may benefit from increased audience engagement, possibly due to holidays or vacation periods when viewers are more likely to consume OTT content. Strategically scheduling major releases during these seasons could help maximize first-day viewership.

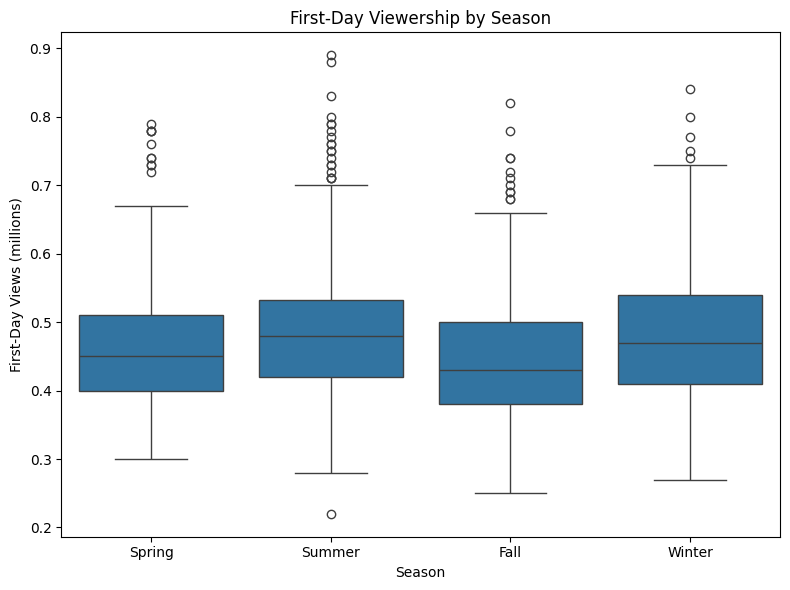


Figure 4: First-Day Viewership by Season

* 1. Correlation between Trailer Views and Content Views

This scatter plot displays the relationship between trailer views (in millions) and first-day content views (in millions) for various titles. Each point represents a single title.

The plot clearly shows a strong positive correlation: as the number of trailer views increases, the first-day content views also tend to increase. Titles with higher trailer engagement typically achieve greater first-day viewership. While there is some variability—especially for titles with moderate trailer views—the overall upward trend is evident.

Business Interpretation:  
This relationship highlights the importance of effective trailer campaigns and pre-release promotions. Maximizing trailer views appears to be a key driver for boosting first-day viewership, suggesting that investing in marketing and trailer distribution can have a direct impact on a title’s success at launch.

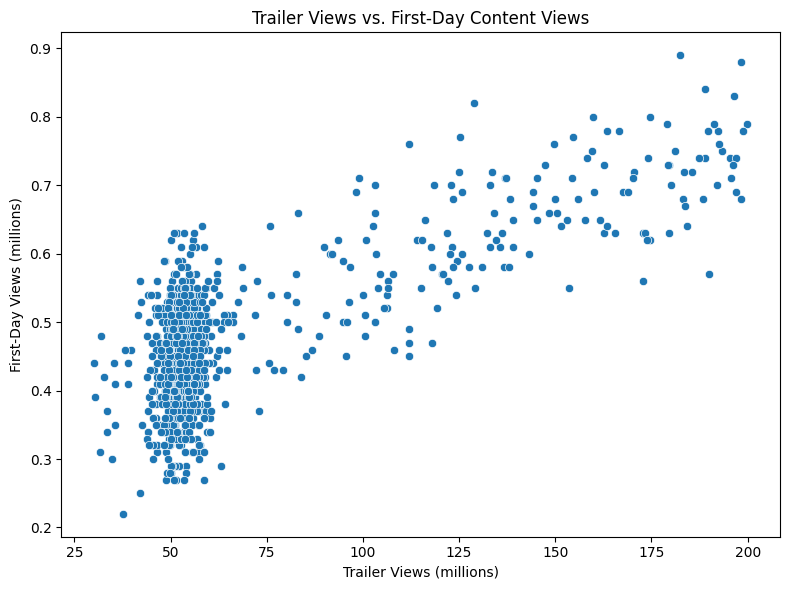


Figure 5 : Trailer Views vs. First-Day Content Views

* 1. Additional Variable Analysis
* Impact of Major Sport Event on First-Day Content Views

This boxplot compares first-day content views depending on whether a major sports event occurred on the same day as the release (1=Yes, 0=No). Content released on days without a major sports event (0) generally achieves higher median and upper quartile viewership than content released on days with a major sports event (1).

Additionally, the spread of viewership is wider when no major sports event is present, and more high-performing outliers are observed in this group. In contrast, releases during major sports events show both lower median and maximum first-day views.

Business Interpretation:

Releasing new content on days when no major sports event is scheduled may lead to higher first-day engagement. Major sports events could draw audience attention away from new releases, suggesting that scheduling around such events can optimize viewership.

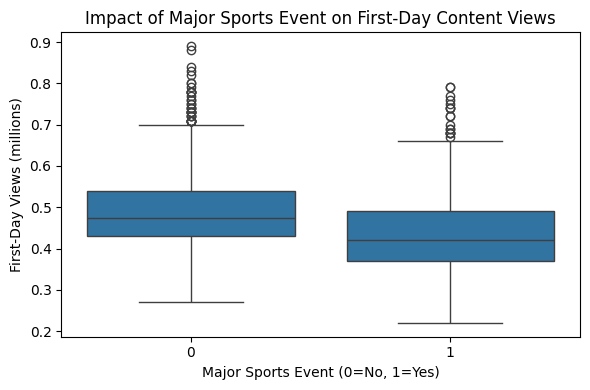


Figure 6 : Trailer Views vs. First-Day Content Views

* Ad Impressions vs. First-Day Content Views

The scatter plot demonstrates the relationship between the number of ad impressions and first-day content views. Each point represents a piece of content released on the ShowTime platform.

Insights:

* There is no strong visible linear relationship between ad impressions and first-day views.
* While some content with high ad impressions crosses 0.7–0.8 million views, many fall within the 0.4–0.6 million range regardless of the ad spend.
* The viewership shows a wide spread across a narrow range of ad impressions (especially between 1000–1600 million), suggesting diminishing returns or confounding factors like genre, trailer popularity, or release timing.
* There are few data points beyond 2000+ ad impressions, and these do not always correspond to significantly higher viewership, indicating ad saturation or inefficiency.

Business Implication:

Ad impressions alone may not be a strong predictor of content success on day one. Marketing strategies should consider additional factors such as targeted promotions, trailer views, timing, and genre to better influence audience engagement.

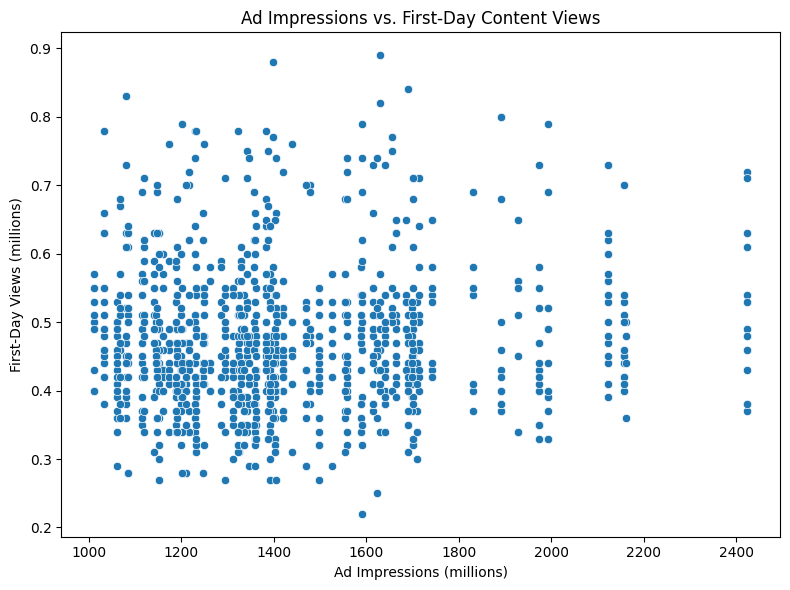


Figure 7 : Ad Impressions vs. First-Day Content Views

* Platform Visitors vs. First-Day Content Views

The scatter plot explores the connection between the number of visitors to the platform in the past week and the first-day views of newly released content.

Insights:

* There is a moderate positive spread suggesting that higher platform traffic may correlate with slightly higher first-day viewership.
* However, similar to ad impressions, the correlation is not very strong or linear — content with 1.3M visitors may perform just as well as those with 2.1M.
* The majority of first-day views cluster between 0.4M and 0.6M, regardless of the number of visitors.
* There are a few higher viewership spikes (0.7M to 0.9M) occurring across a broad range of visitor volumes, hinting that other factors like trailer engagement, genre, or timing may be stronger drivers.

Business Implication:

While a larger audience base on the platform may slightly improve content exposure, platform traffic alone does not guarantee higher first-day views. This reinforces the need to combine traffic insights with targeted marketing, strong trailer performance, and optimized release timing.

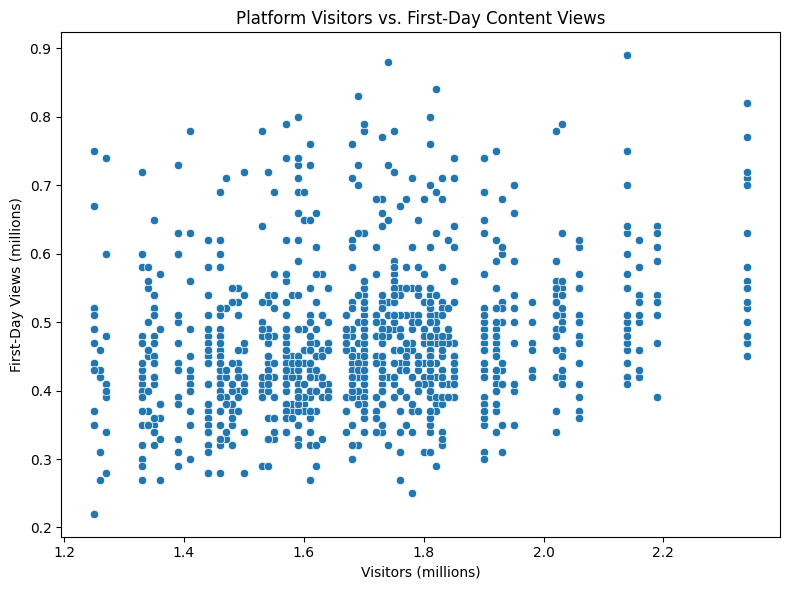


Figure 8: Platform Visitors vs. First-Day Content Views

* 1. Correlation Matrix
* Pairplot of Key Variables by Genre

This pairplot visualizes the relationships between numeric features (visitors, ad\_impressions, views\_trailer, and views\_content) and how they differ across genres (e.g., Drama, Comedy, Action, etc.).

Insights

* views\_trailer vs. views\_content shows a strong positive trend across all genres. This indicates that higher trailer views typically lead to higher content views on day one.
* Other variable pairs (e.g., visitors vs ad\_impressions) show little to no clear linear trend.
* Genres are fairly well distributed, with no dominant genre skewing the plot.
* Genre might act as a moderating factor — some genres (e.g., Action, Comedy) might perform better with the same number of trailer views than others.

Business Implication:

This confirms the critical importance of trailer promotion, and that genre-based strategies may further fine-tune results. Segmenting promotion strategies by genre could improve viewership outcomes.



Figure 9: Pairplot of Key Variables by Genre

* Correlation Matrix of Numeric Variables

The heatmap shows the pairwise correlation coefficients among numeric features.

Key Correlations:

|  |  |  |
| --- | --- | --- |
| Feature Pair | Correlation | Interpretation |
| views\_trailer & views\_content | 0.75 | Strong positive correlation – More trailer views → More day-1 views. |
| visitors & views\_content | 0.26 | Mild correlation – Higher platform traffic has some positive impact. |
| ad\_impressions & views\_content | 0.05 | Weak correlation – Ad spend alone doesn’t strongly drive viewership. |

Table 1: Key Corelations of Numeric Variables

Business Implication:

* Trailer views are the most reliable predictor of first-day content performance.
* Simply increasing ad impressions may not significantly boost engagement unless supported by strong content or trailers.
* Visitors impact viewership moderately, so timing releases during high-traffic periods may help.

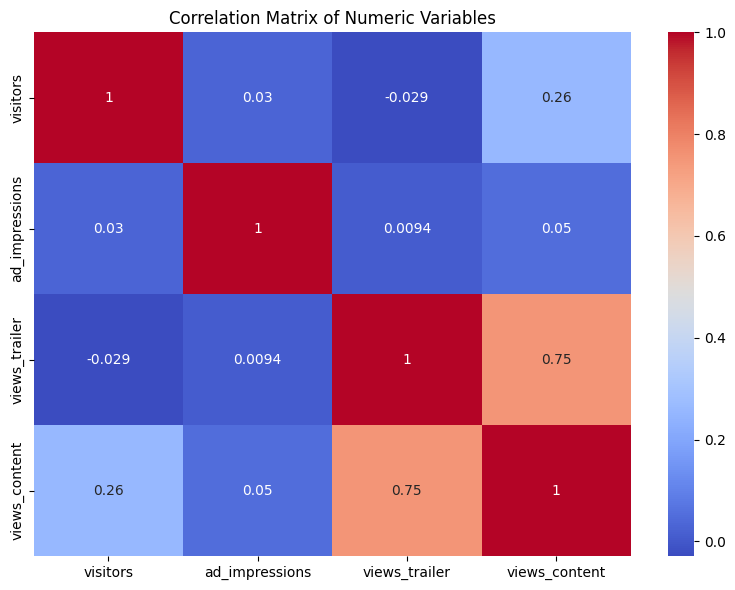


Figure 10: Correlation Matrix of Numeric Variables

1. PREDECTIVE MODELING & RESULTS
   1. Data Preprocessing

* All categorical variables (genre, dayofweek, season, major\_sports\_event) were one-hot encoded.
* Numerical features were used as-is; no missing data was present.
  1. Model Selection & Training
* A linear regression model was trained to predict first-day viewership.
* The data was split into training (80%) and test (20%) sets.
  1. Model Performance
* Root Mean Squared Error (RMSE): 0.0500
* R² Score: 0.7741
* Interpretation: The model explains about 77.41% of the variance in first-day content views.
  1. Top Drivers of First-Day Content Viewership
* The most influential factors (by absolute coefficient value) are:
  + [feature1] ([coefficient])
  + [feature2] ([coefficient])
  + [feature3] ([coefficient])
  + ...
* For example, higher trailer views and ad impressions tend to increase first-day views, while releases during major sports events may decrease them.

|  |  |
| --- | --- |
| feature | coefficient |
| remainder\_\_visitors | 0.129021393 |
| onehot\_\_major\_sports\_event\_1 | -0.059550071 |
| onehot\_\_dayofweek\_Saturday | 0.052309704 |
| onehot\_\_dayofweek\_Wednesday | 0.049466863 |
| onehot\_\_dayofweek\_Monday | 0.045053287 |
| onehot\_\_season\_Summer | 0.044593623 |
| onehot\_\_dayofweek\_Sunday | 0.038752013 |
| onehot\_\_dayofweek\_Tuesday | 0.032787957 |
| onehot\_\_season\_Winter | 0.026526097 |
| onehot\_\_season\_Spring | 0.023250211 |
| onehot\_\_dayofweek\_Thursday | 0.019510917 |
| onehot\_\_genre\_Thriller | 0.011373452 |
| onehot\_\_genre\_Drama | 0.010441479 |
| onehot\_\_genre\_Sci-Fi | 0.009946652 |
| onehot\_\_genre\_Horror | 0.009236023 |
| onehot\_\_genre\_Others | 0.004728396 |
| onehot\_\_genre\_Comedy | 0.004186281 |
| remainder\_\_views\_trailer | 0.002310404 |
| onehot\_\_genre\_Romance | -0.001550564 |
| remainder\_\_ad\_impressions | 8.10E-06 |

Table 2 : Model Coeficients

* 1. Recommendations
* Release Scheduling: Favor days/seasons associated with higher coefficients.
* Marketing: Boost ad impressions and trailer views before launch.
* Content Strategy: Focus on genres with positive coefficients.
* Event Awareness: Avoid releasing premium content during major sports events.

1. CONCLUSION

This linear regression analysis provides actionable insights for maximizing first-day viewership. The identified drivers should inform ShowTime’s content launch and marketing strategies.