Project 3

- Predictive Modeling -

(ShowTime Project)

1.Introduction

1.1.Problem Definition

Business Context:

Over-the-top (OTT) platforms, such as ShowTime, are reshaping how consumers access and enjoy media by offering on-demand video content through the internet. The global OTT market is rapidly expanding, driven by shifts in consumer habits and a growing preference for streaming services over traditional television. Consequently, these platforms face pressure to consistently deliver fresh and compelling content that captures viewer interest right from its release.

Competition within the OTT sector has intensified, particularly with the surge in content consumption observed during the COVID-19 pandemic. As the market becomes saturated with new offerings, ShowTime is keen on enhancing its first-day viewership, which serves as a vital metric for gauging content success.

Business Problem

ShowTime has detected inconsistencies in first-day content viewership, which is crucial for boosting user engagement and retaining subscribers. Several factors may be influencing these viewership figures, such as a drop in platform traffic, decreased advertising expenditure, competition from significant sports events, and the timing of content releases (e.g., weekends, holidays). Gaining insights into these factors is essential for ShowTime to formulate effective strategies to enhance viewership on its platform.

Objective

The main goal of this analysis is to uncover the key factors that drive first-day content viewership. By utilizing data from ShowTime's platform incorporating variables like visitor counts, ad impressions, trailer views, timing of content releases, and genre a predictive model will be constructed to identify which elements most significantly affect viewership.

1.2.Key Questions

This analysis seeks to address the following inquiries:

- What factors have a significant impact on first-day viewership on ShowTime's platform?
- How do variables such as ad impressions, visitor counts, and trailer views influence viewership levels?
- Does the timing of content releases (considering day of the week and season) affect viewer engagement?
- Are there external influences, such as major sports events, that contribute to decreased viewership?

Deliverable

The anticipated outcome of this project will be a linear regression model that identifies the main drivers of first-day content viewership. Based on the findings, actionable insights and recommendations will be presented to ShowTime, enabling the company to make strategic decisions to optimize its content strategy and enhance viewership performance.

1.3.Data Background and Contents

Dataset Overview

The dataset provided by ShowTime contains various factors that could potentially influence the first-day viewership of content on their platform. The goal is to analyze these factors to determine which ones drive viewership and how ShowTime can improve content performance.

Data Source

The dataset was collected from ShowTime's platform and includes metrics related to platform engagement, marketing efforts, content characteristics, and external factors. This data serves as the foundation for building a predictive model to understand first-day content viewership trends.

Target Variable

The target variable for this analysis is views content, which represents the number of first-day views (in millions) for a specific piece of content. This variable reflects how well the content performed on its first day of release and is the main metric ShowTime wants to optimize.

Independent Variables (Features)

The dataset contains the following independent variables, which will be analyzed to understand their impact on first-day viewership:

- **Platform Engagement**:
  - **visitors**: The average number of visitors (in millions) to ShowTime's platform in the past week. This variable captures the overall traffic on the platform leading up to the content's release.
- **Marketing and Exposure**:
  - **ad_impressions**: The number of ad impressions (in millions) generated by all ad campaigns for the content (both ongoing and completed). A higher number of ad impressions is expected to increase content visibility.
  - **views_trailer**: The number of views (in millions) the content's trailer received before the content's release. Trailer views can serve as an indicator of interest in the content.
- **Content and Release Timing**:
  - **genre**: The genre of the content, such as Drama, Comedy, Action, etc. Different genres may attract different levels of viewership based on audience preferences.
  - **dayofweek**: The day of the week on which the content was released (e.g., Monday, Friday). Viewership may vary depending on whether the content was released on a weekday or a weekend.
  - **season**: The season in which the content was released (e.g., Winter, Summer). Certain seasons may see higher viewership due to holidays or vacation periods.

- **External Factors**:
    - **major_sports_event**: A binary variable indicating whether a major sports event occurred on the day of the content's release (Yes/No). Major sports events could divert attention from ShowTime's platform, potentially reducing viewership.

Potential Impact of Variables

Each of these variables is expected to have a varying degree of influence on first-day content viewership:

- **visitors**, **ad_impressions**, and **views_trailer** are expected to have a positive impact, as greater platform traffic, ad visibility, and pre-release engagement could drive higher viewership.
- **genre**, **dayofweek**, and **season** may reveal patterns in content consumption based on audience behavior and preferences.
- **major_sports_event** is likely to negatively impact viewership, as large audiences may be drawn to sports broadcasts instead of streaming content on ShowTime.

1.4.Univariate Analysis :

1. visitors to the platform in the past week:

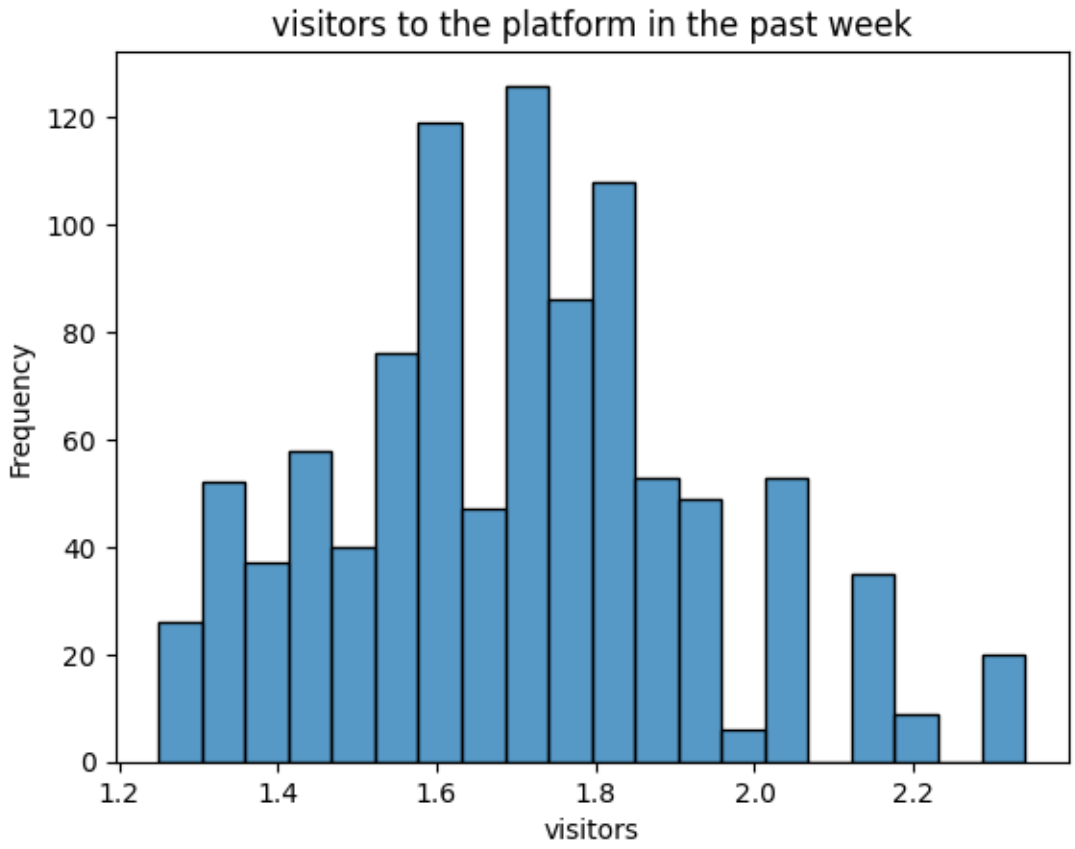

Figure 1. visitors to the platform in the past week

Interpretation:

The histogram of visitors to the ShowTime platform shows an average of 1.70 million visitors, indicating solid engagement. Most content attracts between 1.55 million and 1.83 million viewers, with some pieces significantly outperforming others. This distribution suggests variability in audience interest and highlights the differences in content performance on the platform.

2. Distribution of Average Weekly Visitors to ShowTime:

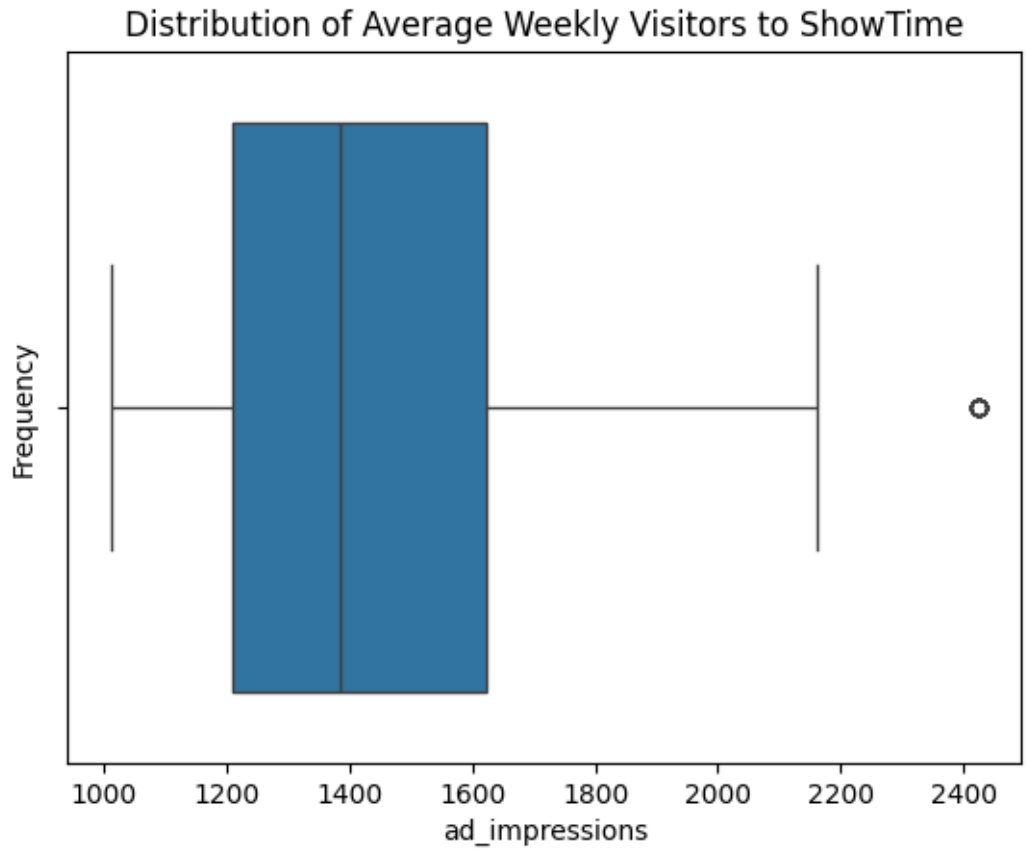

Figure 2. Distribution of Average Weekly Visitors to ShowTime

Interpretation:

The boxplot of ad impressions illustrates the central tendency and variability of advertising exposure on the ShowTime platform. The median ad impressions are approximately 1,434.71 million, while the interquartile range spans from about 1,210.33 million to 1,623.67 million. Outliers highlight instances of unusually high or low ad impressions, reflecting the diversity in advertising strategies across different content releases.

3. Distribution of Trailer Views for ShowTime Content:

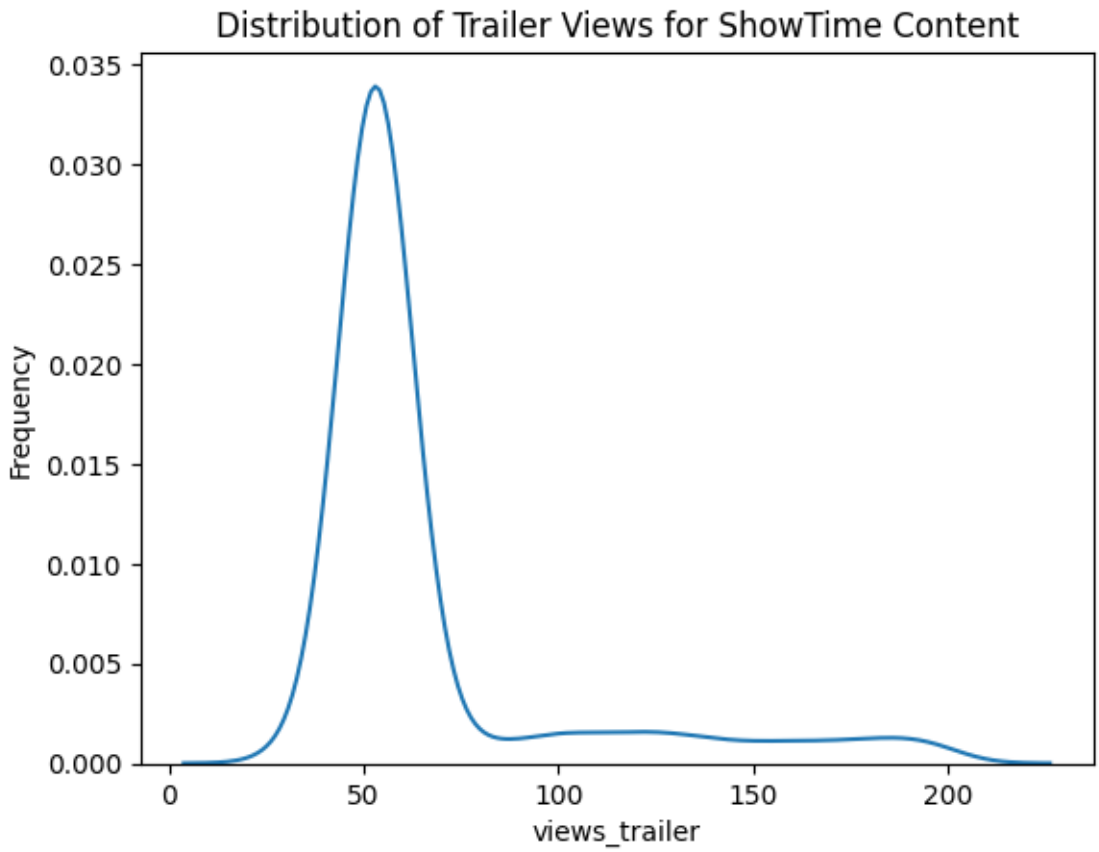

Figure 3. Distribution of Trailer Views for ShowTime Content

Interpretation:

The kernel density estimate (KDE) plot of trailer views reveals the distribution of audience engagement with ShowTime content. The peak of the distribution, approximately 66.92 million views, indicates strong interest in trailers. The spread suggests that while many trailers receive moderate views, a few gain significantly higher engagement, reflecting variability in audience attraction to different content offerings.

4. Distribution of First-Day Views for ShowTime Content:

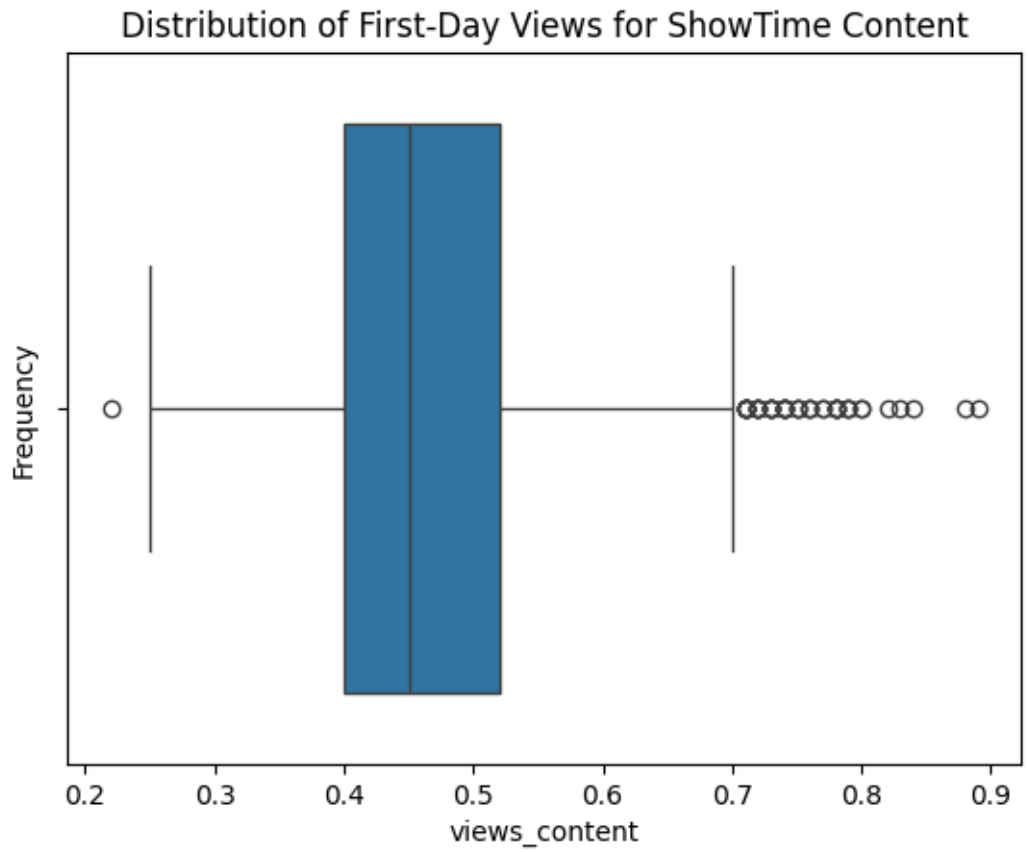

Figure 4.Distribution of First-Day Views for ShowTime Content

Interpretation:

The boxplot of first-day views for ShowTime content illustrates the distribution of audience engagement upon release. The median first-day views are approximately 0.47 million, indicating a typical level of initial interest. The interquartile range spans from about 0.40 million to 0.52 million, with outliers reflecting instances of unusually high or low engagement, showcasing variability in content performance.

5. Frequency of Major Sports Events on ShowTime Release Days:

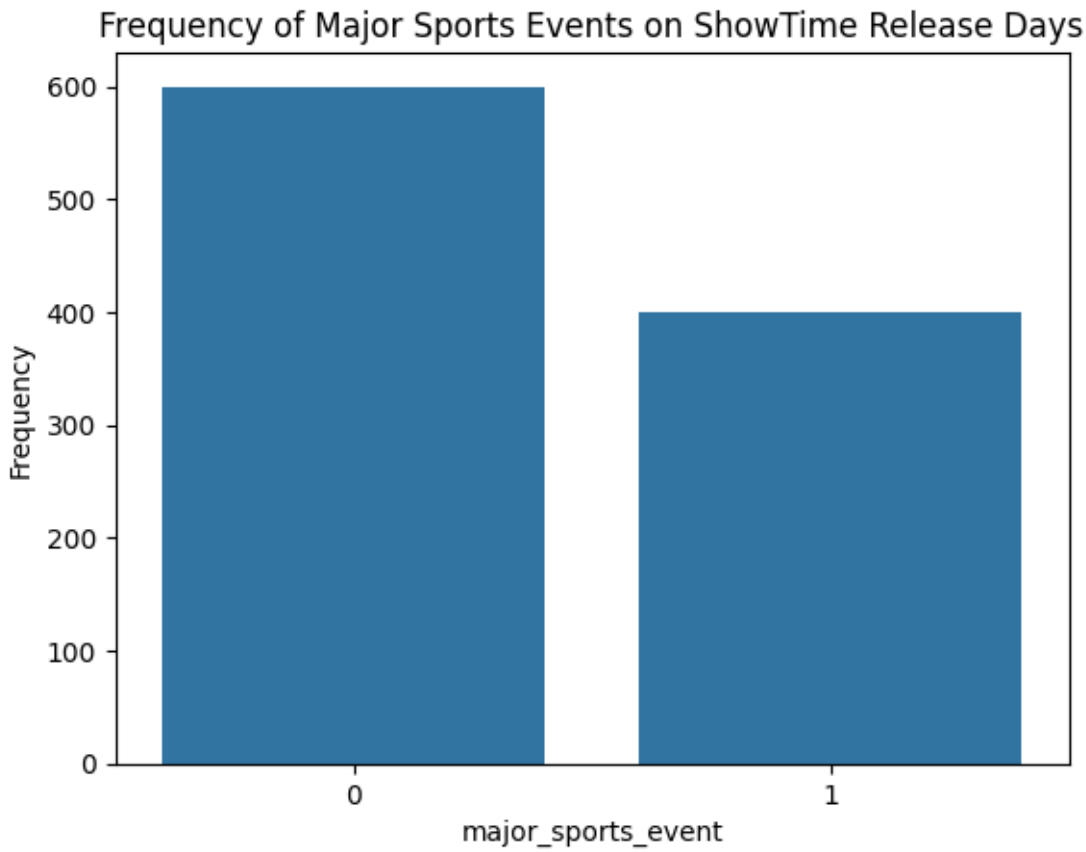

Figure 5.Frequency of Major Sports Events on ShowTime Release Days

Interpretation:

The count plot for major sports events shows a clear distinction between the two categories. There are 600 instances (60%) of content releases on days without major sports events and 400 instances (40%) on days with major sports events. This distribution highlights the frequency of content released in relation to the occurrence of major sports events, indicating potential competition for viewer attention.

6. Distribution of Content Genres on ShowTime:

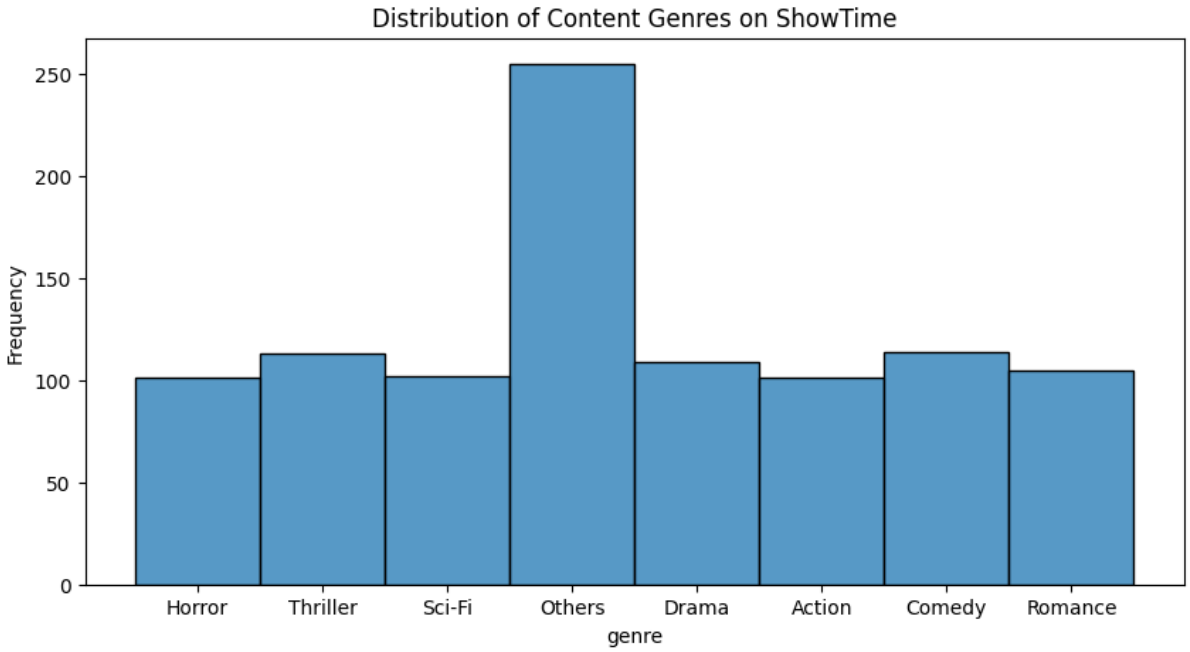

Figure 6: Distribution of Content Genres on ShowTime

Interpretation:

The histogram for content genres on ShowTime reveals the distribution of various genres among the releases. The genre "Others" dominates with 255 instances, followed by "Comedy" (114), "Thriller" (113), "Drama" (109), and "Romance" (105). The genres "Sci-Fi" and "Action" have slightly lower counts at 102 and 101, respectively, while "Horror" also has 101 instances. This distribution indicates a diverse range of genres, with "Others" being the most prevalent.

7. Distribution of Content Releases by Day of the Week:

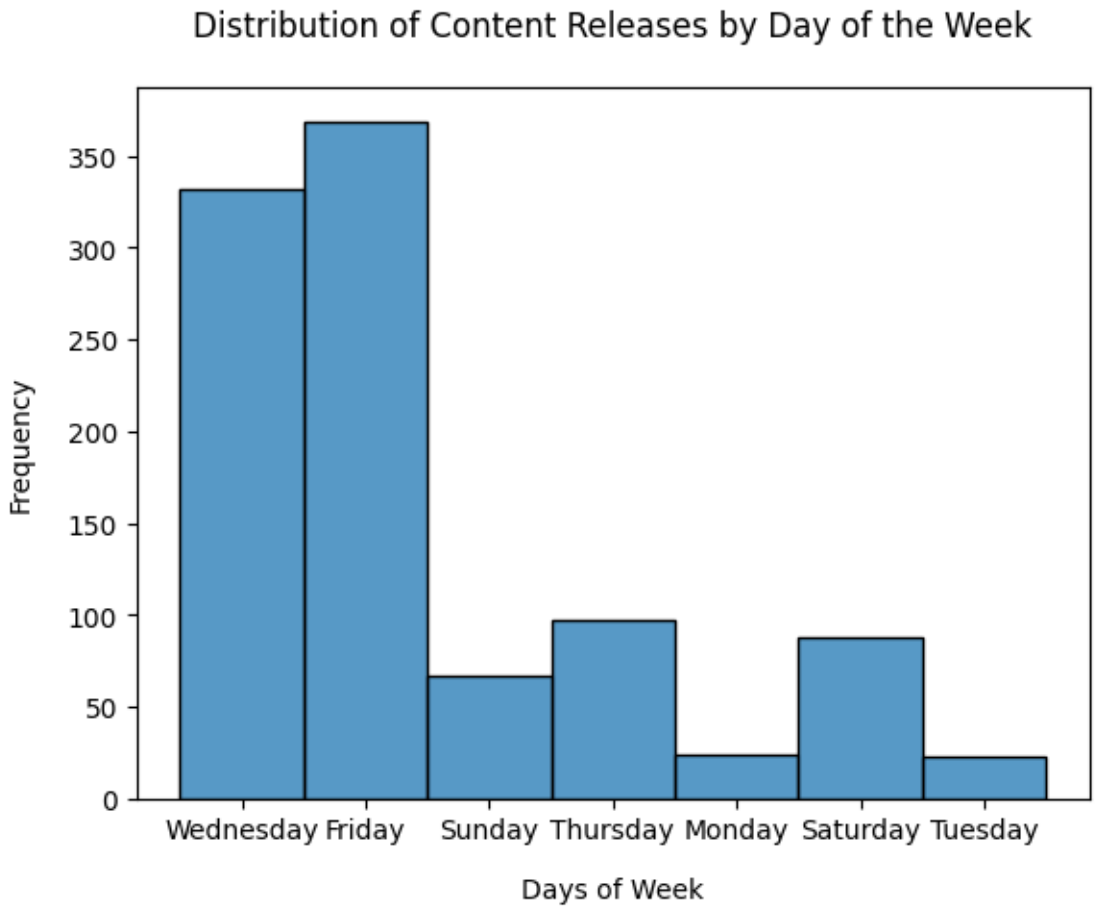

Figure 7. Distribution of Content Releases by Day of the Week

Interpretation:

The histogram for content releases by day of the week reveals the frequency of releases across different weekdays. The highest frequency is observed on Friday with 369 releases, followed by Wednesday with 332 releases. The remaining days show significantly fewer releases, with Monday (24) and Tuesday (23) having the least. This distribution indicates that ShowTime primarily schedules releases towards the end of the week, likely to capture higher viewer engagement.

8. Distribution of Content Releases by Season on ShowTime:

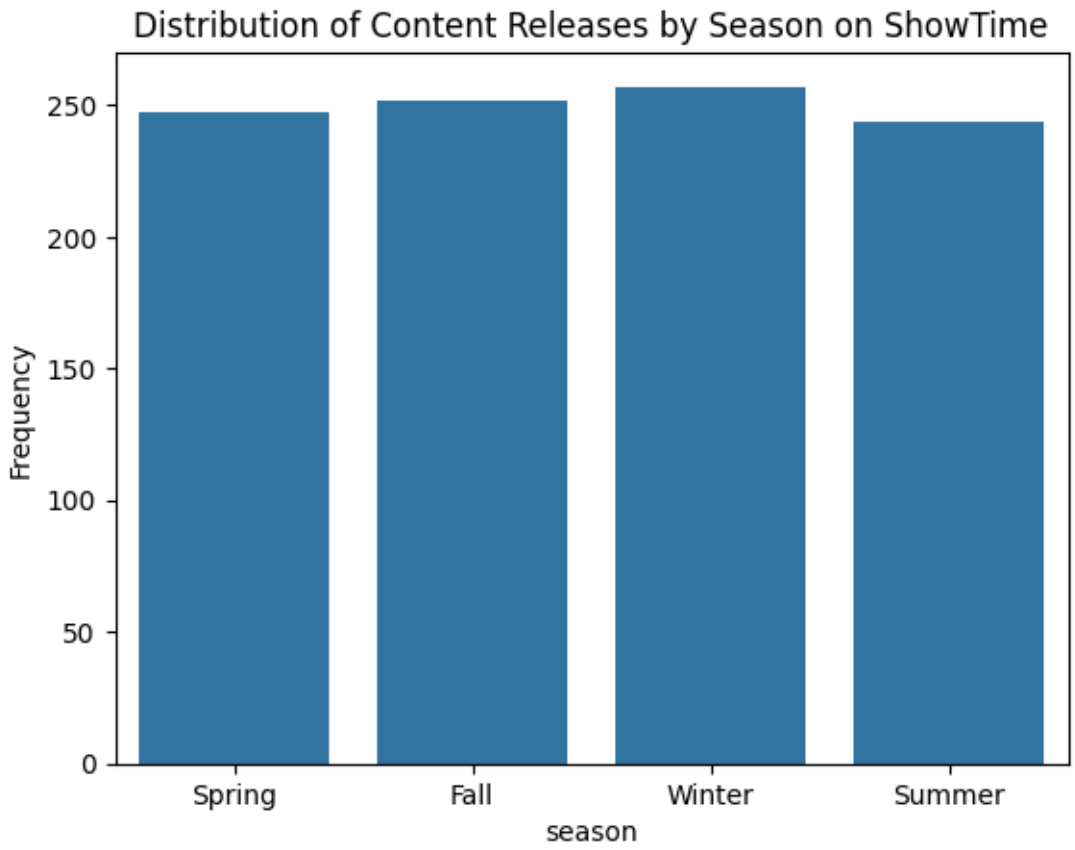

Figure 8.Distribution of Content Releases by Season on ShowTime

Interpretation:

The count plot for content releases by season illustrates the frequency of releases throughout the year. The highest number of releases occurs in Winter with 257 instances, followed closely by Fall with 252 releases. Spring and Summer show similar frequencies, with 247 and 244 releases, respectively. This distribution suggests a relatively balanced approach to content releases across seasons, with a slight preference for Winter and Fall.

1.5.Bivariate Analysis:

9. Relationship between all numeric variables:

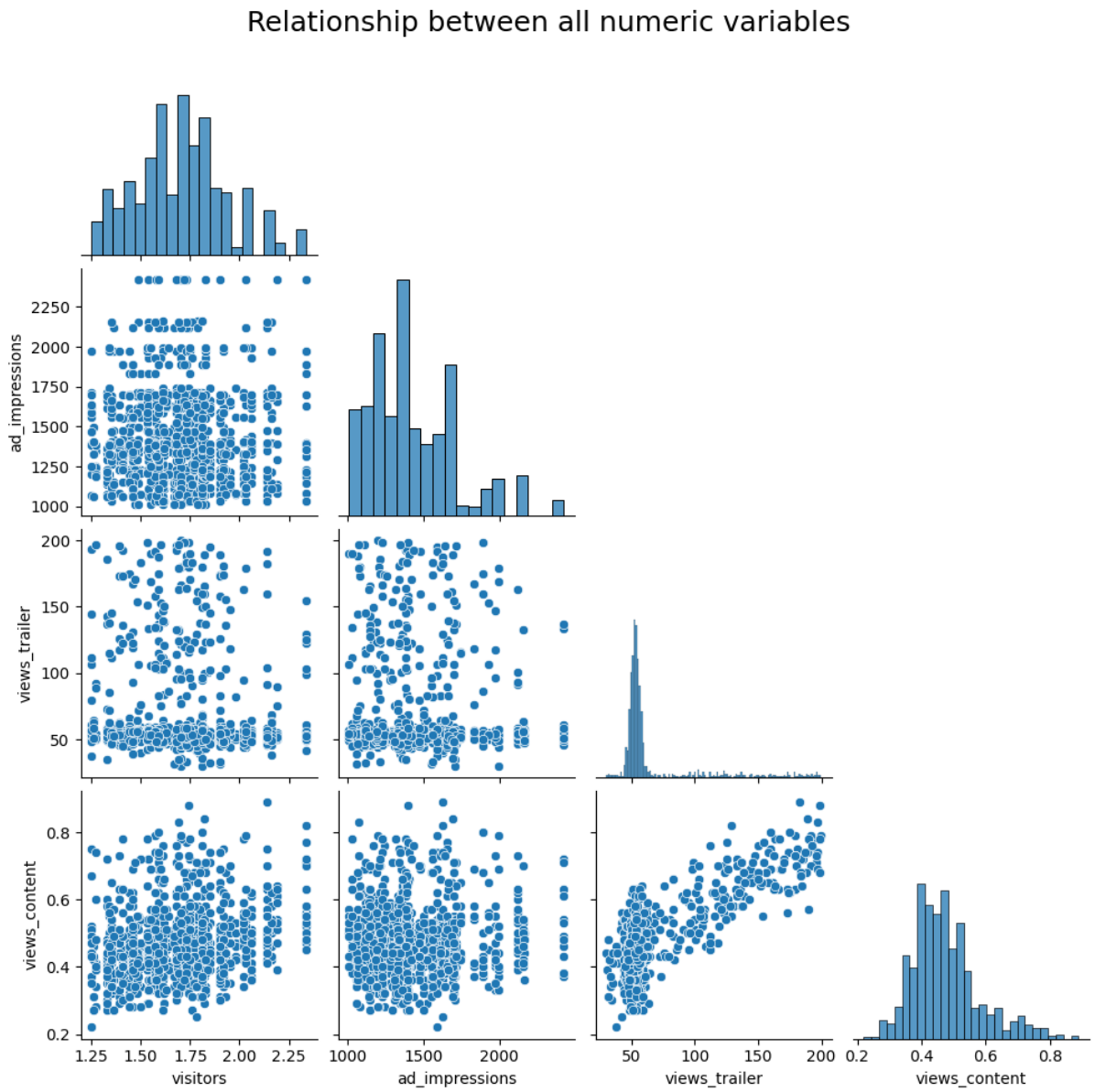

Figure 9.Relationship between all numeric variables

Interpretation:

The pairplot visualizes the relationships among numeric variables: visitors, ad impressions, views trailer, and views content. It reveals a positive correlation between visitors and ad impressions, indicating that increased advertising drives more traffic. Additionally, higher trailer views are linked to greater visitor counts, emphasizing effective marketing. A strong connection between views trailer and views content suggests that engaging trailers lead to increased first-day viewership. Furthermore, more ad impressions correlate with higher views content, highlighting the impact of advertising on viewership. These insights underscore the importance of marketing strategies in enhancing content performance and audience engagement for ShowTime.

10. Impact of Major Sports Events on First-Day Views of ShowTime Content:

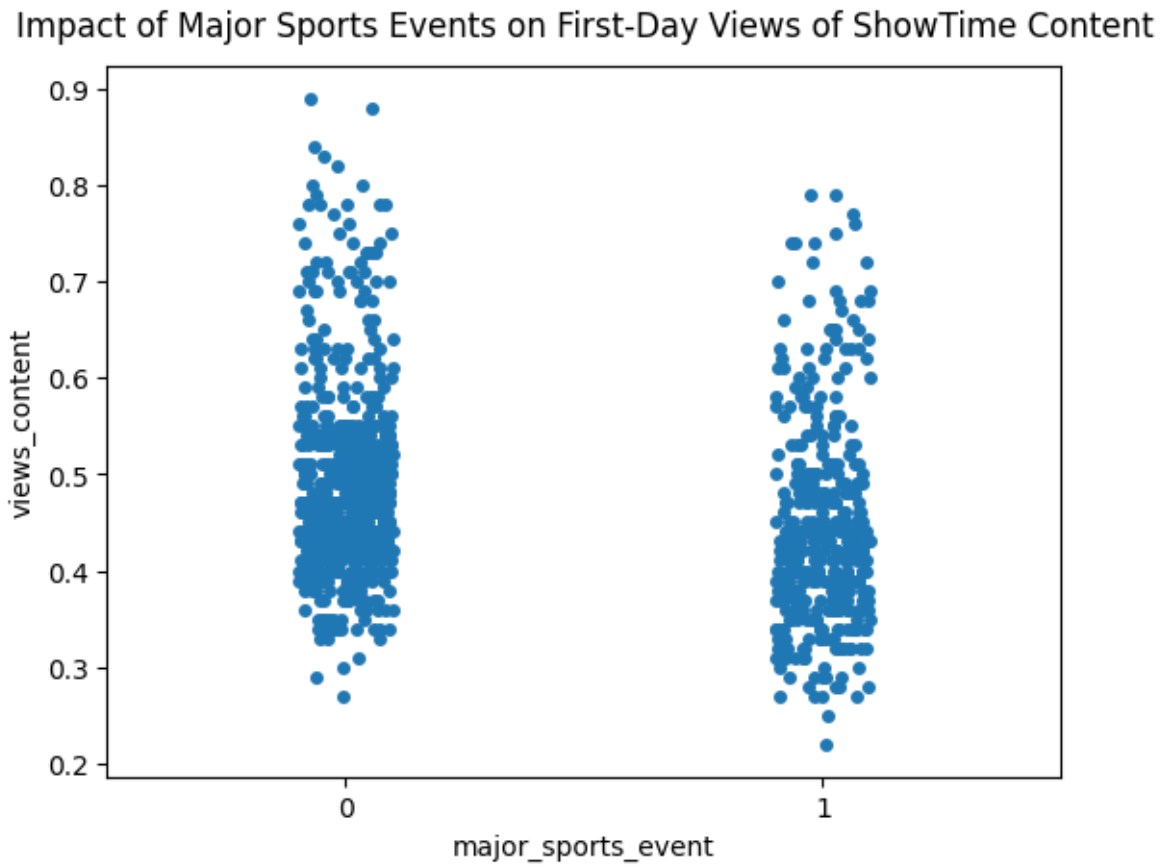

Figure 10.Impact of Major Sports Events on First-Day Views of ShowTime Content

Interpretation:

The strip plot illustrates the impact of major sports events on first-day views of ShowTime content. Content released without major sports events generally achieves higher viewership, falling with in the range of 0.35 million and 0.6 million views. Conversely, releases coinciding with sports events typically attract fewer viewers, often falling below 0.45 million views. This highlights how external factors can significantly affect audience engagement with streaming content.

11. Average First-Day Views by Genre for ShowTime Content:

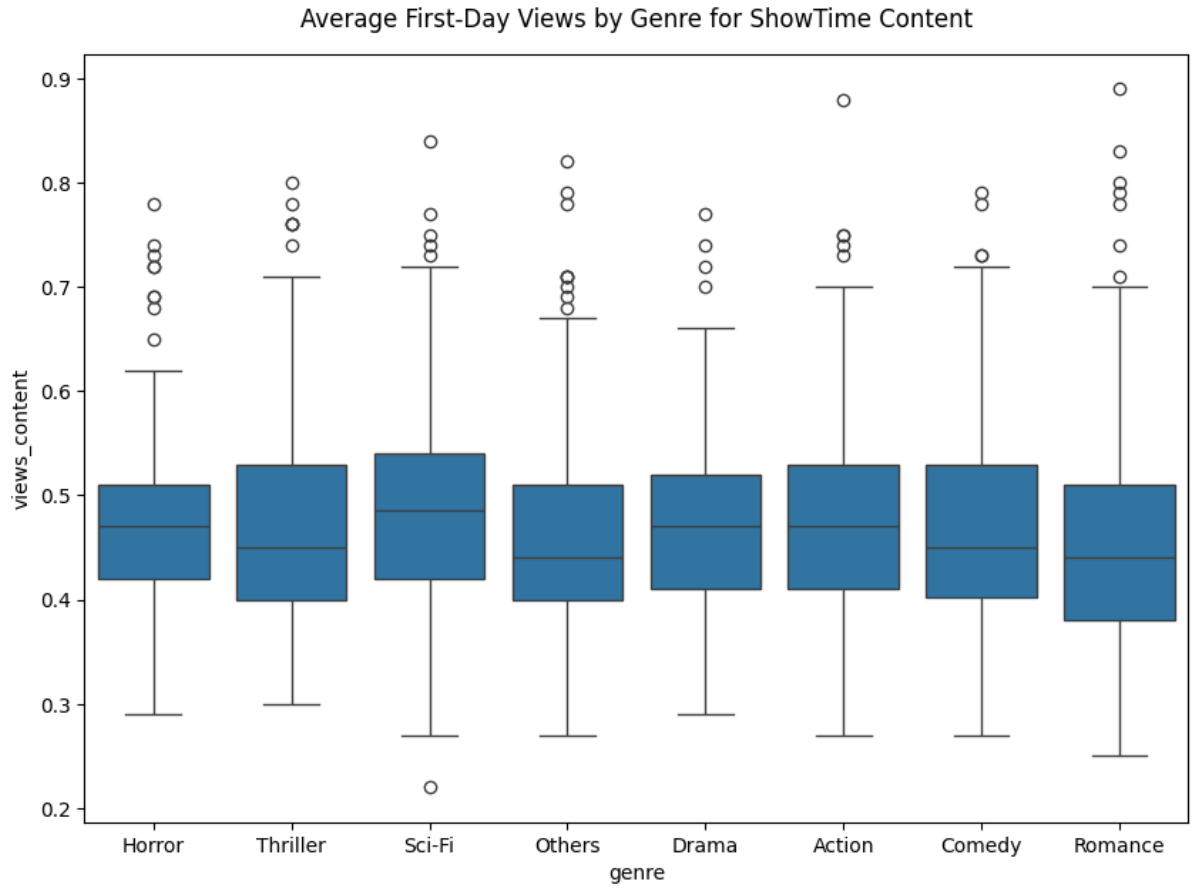

Figure 11. Average First-Day Views by Genre for ShowTime Content

Interpretation:

Approximately 50% of all genres fall within the content views range of 3.8 to 5.5 million. Each genre contains a significant number of outliers. Notably, the sci-fi genre has the highest content views, while the romance genre has comparatively lower content views. This distribution indicates variability in audience engagement across genres, highlighting the need to consider both average views and outliers when analyzing genre performance.

12. Average First-Day Views by Day of the Week for ShowTime Content

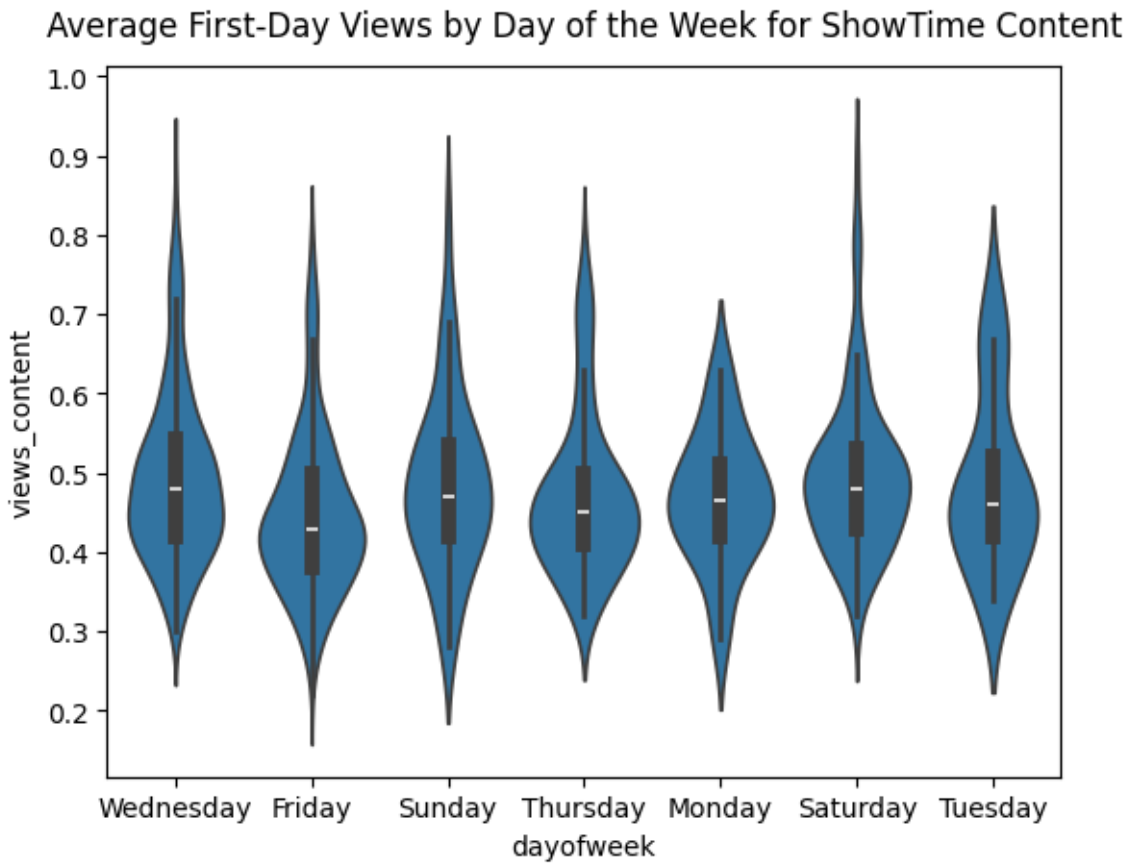

Figure 12. Average First-Day Views by Day of the Week for ShowTime Content

Interpretation:

Daily content views display considerable variability, with around 50% falling within the range of 3.5 to 5.5 million. Notably, Wednesday leads with content views, while Friday typically records lower views. This observation emphasizes the importance of considering both average views and outliers when evaluating trends in audience engagement throughout the week.

13. Average First-Day Views by Season for ShowTime Content:

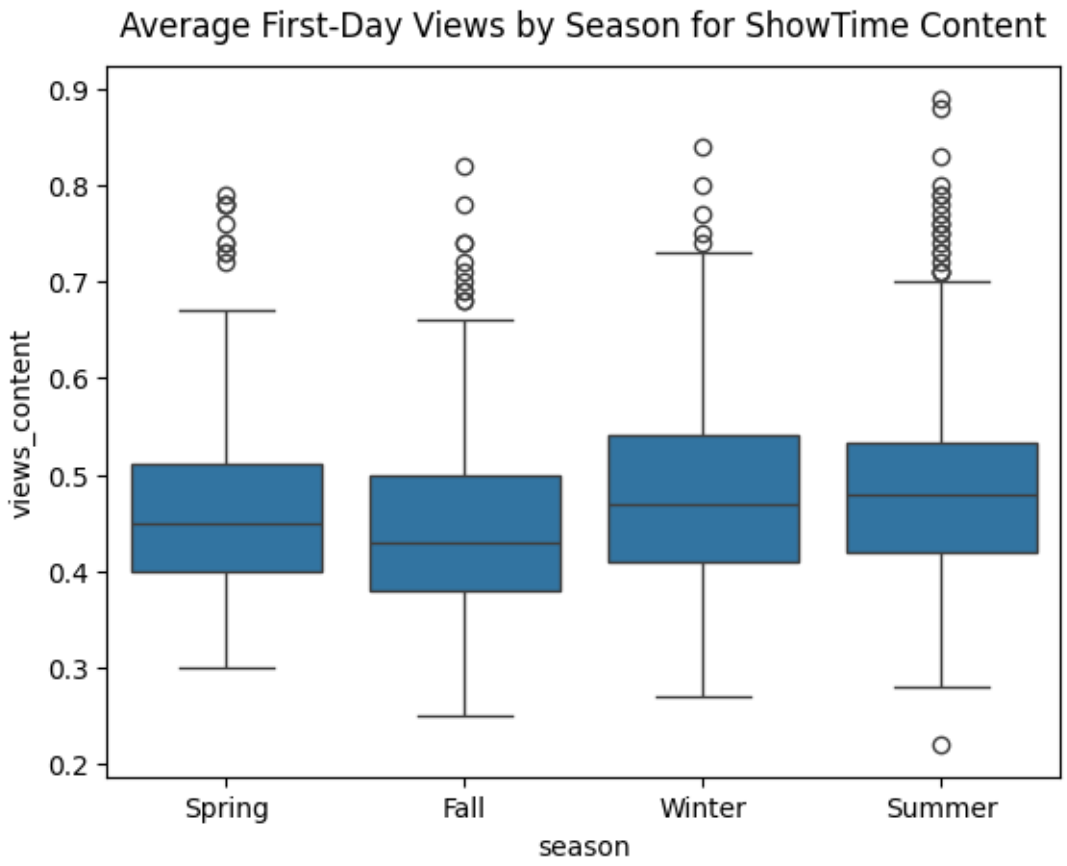

Figure 13. Average First-Day Views by Season for ShowTime Content

Interpretation:

Seasonal content views exhibit significant variability, with approximately 50% falling between 3.8 and 5.5 million. Notably, the fall season records the lowest views, while winter content tends to engage audiences more effectively than summer releases. This finding underscores the importance of accounting for both average views and outliers when analyzing trends in audience engagement over the seasons.

1.6.Answers to the key questions provided :

1. Distribution of content views:

Distribution of First-Day Views for ShowTime Content:

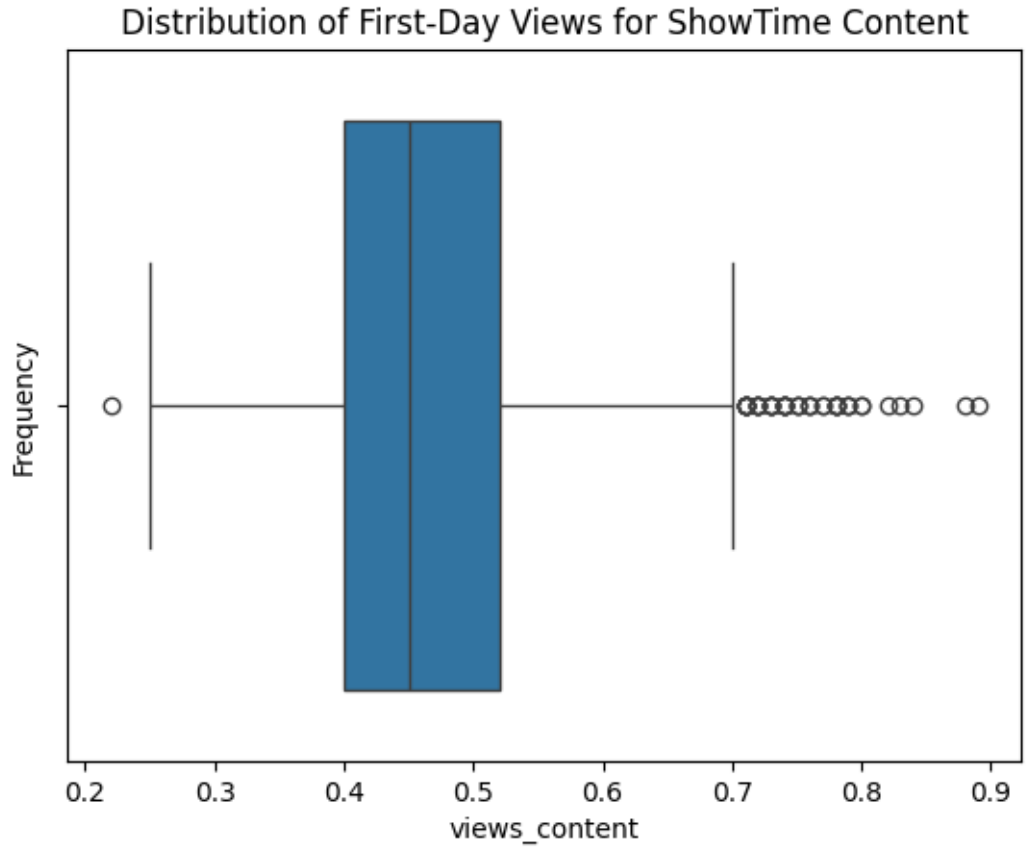

Figure 14. Distribution of First-Day Views for ShowTime Content

Answer 1:

The boxplot for first-day views of ShowTime content illustrates audience engagement upon release. The median views are about 0.47 million, indicating a typical level of interest. The interquartile range is approximately 0.40 million to 0.52 million, while outliers reveal cases of unusually high or low engagement. This highlights the variability in content performance and suggests that some releases significantly capture or miss audience attention compared to others.

2. Distribution of genres:

Distribution of Content Genres on ShowTime:

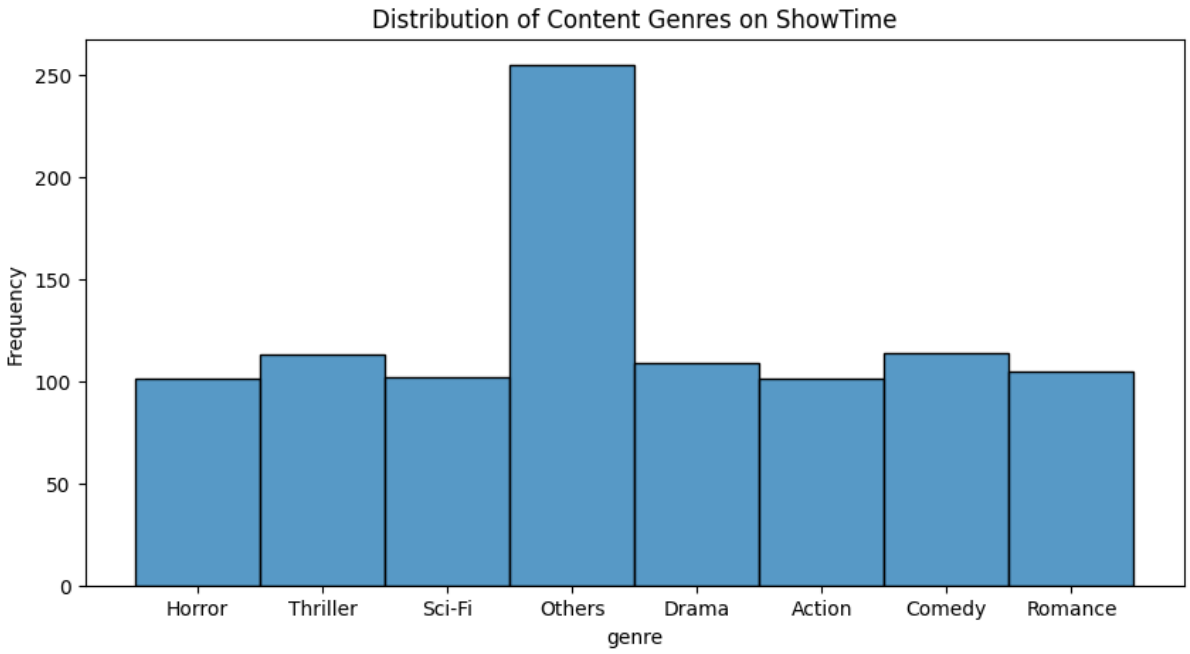

Figure 15. Distribution of Content Genres on ShowTime

Answer 2:

The histogram of content genres on ShowTime illustrates the distribution of various releases. The "Others" category is the most prevalent, with 255 instances, followed by "Comedy" (114), "Thriller" (113), "Drama" (109), and "Romance" (105). "Sci-Fi" and "Action" have slightly lower counts at 102 and 101, respectively, while "Horror" also stands at 101 instances. This distribution reflects a diverse array of genres, highlighting "Others" as the dominant category among the releases.

3. Variance in viewership with the day of release:

Average First-Day Views by Day of the Week for ShowTime Content:

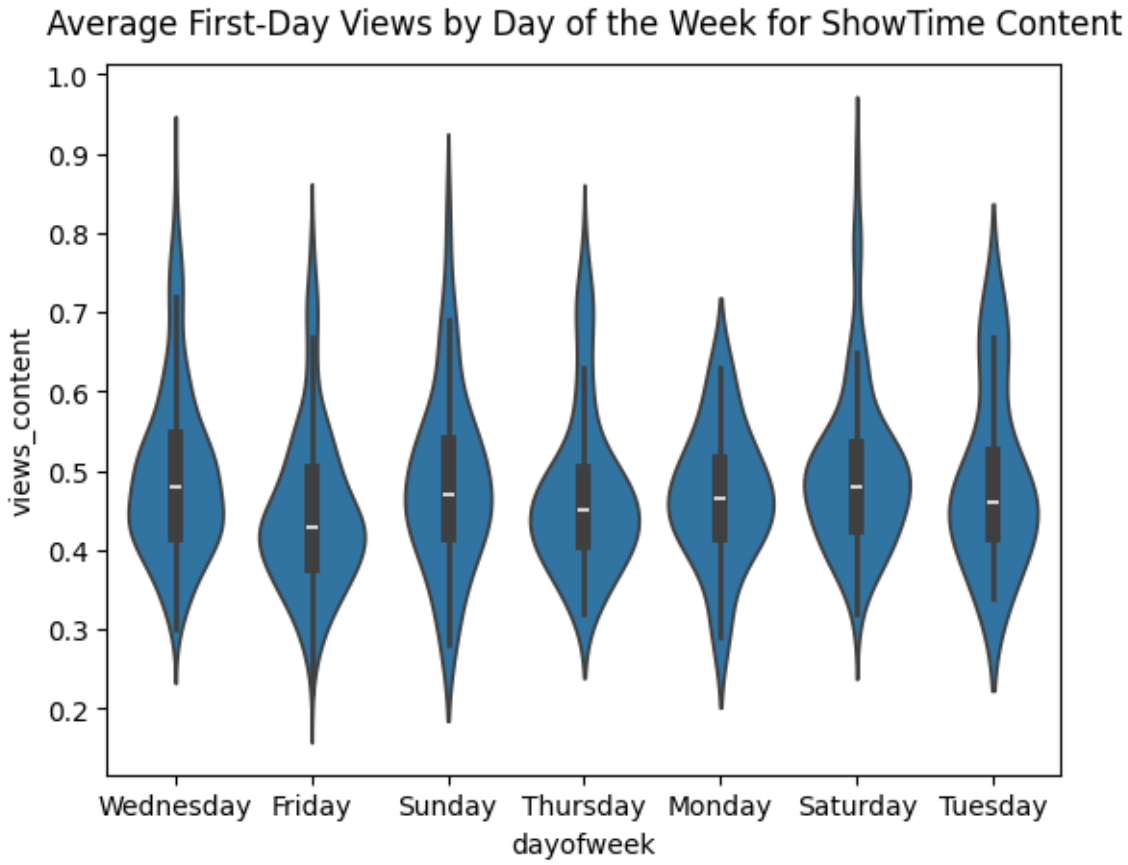

Figure 16. Average First-Day Views by Day of the Week for ShowTime Content

Answer 3:

Viewership varies significantly with the day of release, highlighting the impact of timing on audience engagement. Approximately 50% of daily content views fall between 3.5 and 5.5 million. Notably, Wednesday emerges as the day with the highest content views, while Friday typically sees lower engagement. This pattern underscores the importance of analyzing both average views and outliers to understand trends in viewership throughout the week, illustrating how the day of release can influence overall audience interest.

4. Variance in viewership by season of release:

Average First-Day Views by Season for ShowTime Content:

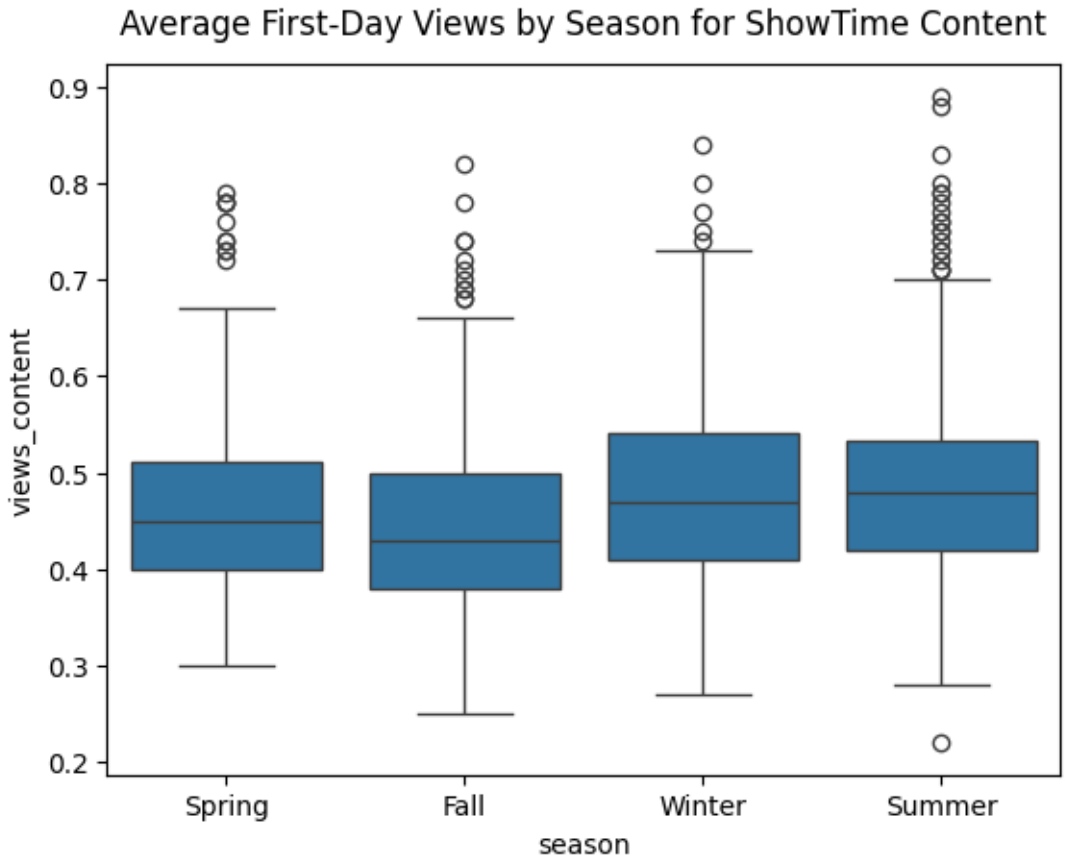

Figure 17. Average First-Day Views by Season for ShowTime Content

Answer 4:

Viewership varies considerably with the season of release, indicating seasonal trends in audience engagement. Approximately 50% of seasonal content views range from 3.8 to 5.5 million. Notably, fall recordings show the lowest viewership, while winter content tends to engage audiences more effectively than summer releases. This analysis highlights the significance of considering both average views and outliers when evaluating trends in audience engagement across different seasons, emphasizing how seasonality can influence overall viewership patterns.

5. correlation between trailer views and content views:

Answer 5:

The correlation between trailer views and content views is approximately 0.75, indicating a strong positive relationship. This suggests that as trailer views increase, content views tend to rise as well. A high correlation implies that effective trailer marketing likely enhances audience interest and engagement with the actual content. This relationship underscores the importance of trailers as a promotional tool, as they can significantly influence initial viewership, making them a crucial aspect of content release strategies.

1.7.Insights Based on EDA

ShowTime, a leading OTT platform, seeks to identify key factors influencing first-day content viewership. In an increasingly competitive market, maximizing audience engagement from the moment content is released is critical. To achieve this, the data from ShowTime's platform was analyzed through Exploratory Data Analysis (EDA) to uncover patterns and insights related to content viewership. This analysis examines factors such as platform traffic, marketing efforts, content characteristics, and external influences like major sports events to understand their impact on first-day views. These insights will help guide ShowTime's strategy to enhance content performance and boost audience engagement on release days.

1. Data Background and Contents:

The dataset includes 1,000 rows and 8 columns representing various factors that could influence viewership. These factors are:

- **visitors**: Average visitors to the platform
- **ad_impressions**: Number of ad impressions
- **major_sports_event**: Indicator of major sports events on release day
- **genre**: Genre of the content
- **dayofweek**: Day of content release
- **season**: Season of release
- **views_trailer**: Trailer views before release
- **views_content**: First-day views (target variable)

2. Univariate Analysis:

- **Visitors**: Distributed around 1.7 million, with moderate variance.
- **Ad Impressions**: Mostly concentrated around 1,434 million but with substantial variability.
- **Major Sports Event**: 40% of content released on days with major sports events.
- **Genre**: 'Others' is the largest category, followed by Comedy, Thriller, and Drama.
- **Day of the Week**: Most content is released on Fridays and Wednesdays.
- **Season**: Distribution is almost even across seasons, with Winter and Fall being slightly more popular for releases.
- **Trailer Views**: Trailer views range widely, from 30 million to nearly 200 million.
- **First-Day Views (Target Variable)**: Most content garners between 0.40 to 0.52 million views on the first day.

3. Bivariate Analysis:

- **Visitors vs. Views**: A positive relationship exists between platform traffic and first-day content views. More visitors equate to more views.
- **Ad Impressions vs. Views**: Higher ad impressions lead to more first-day views, underscoring the importance of marketing efforts.
- **Major Sports Event vs. Views**: Content released on days with major sports events tends to receive fewer first-day views, possibly due to audience attention being diverted.
- **Genre vs. Views**: Action, Sci-Fi, and Comedy genres tend to have higher first-day views, while Romance and Horror perform relatively lower.
- **Day of the Week vs. Views**: Friday releases attract the highest first-day views, while Monday and Tuesday releases see the lowest.
- **Season vs. Views**: Winter and Fall releases perform better than Summer, likely due to seasonal behaviors like holidays.
- **Trailer Views vs. Views**: A strong positive correlation is seen between trailer views and first-day content views, indicating the value of early engagement.

4.Conclusive Insights:

The analysis reveals that platform traffic (visitors)**,** ad impressions, and trailer views are key drivers of first-day content viewership. Genre, day of the week**,** and season also play a significant role in viewership variations. The presence of major sports events negatively impacts views. ShowTime should focus on maximizing ad impressions, trailer engagement, and timing content releases (e.g., Fridays and during non-sport days) to boost first-day.

2.Data preprocessing:

2.1.Duplicate value check:

A thorough check for duplicate entries in the dataset was performed to ensure data integrity. Duplicate values can distort the analysis and impact the accuracy of the model. After conducting the analysis, it was confirmed that there are no duplicate rows in the dataset. Therefore, no further action related to duplicate data is necessary.

2.2.Missing Value Treatment

In the analysis of the dataset, a thorough examination for missing values was conducted. The results indicated that there are no missing values present in any of the columns. Each column, including visitors, ad_impressions, major_sports_event, genre, dayofweek, season, views_trailer, and views_content, contained complete data with 1000 non-null entries. This completeness ensures that subsequent analysis and modeling processes can be performed without the need for imputation or removal of any rows, thereby maintaining the integrity and reliability of the dataset.

2.3.Outlier Treatment:

Outliers of numerical columns:

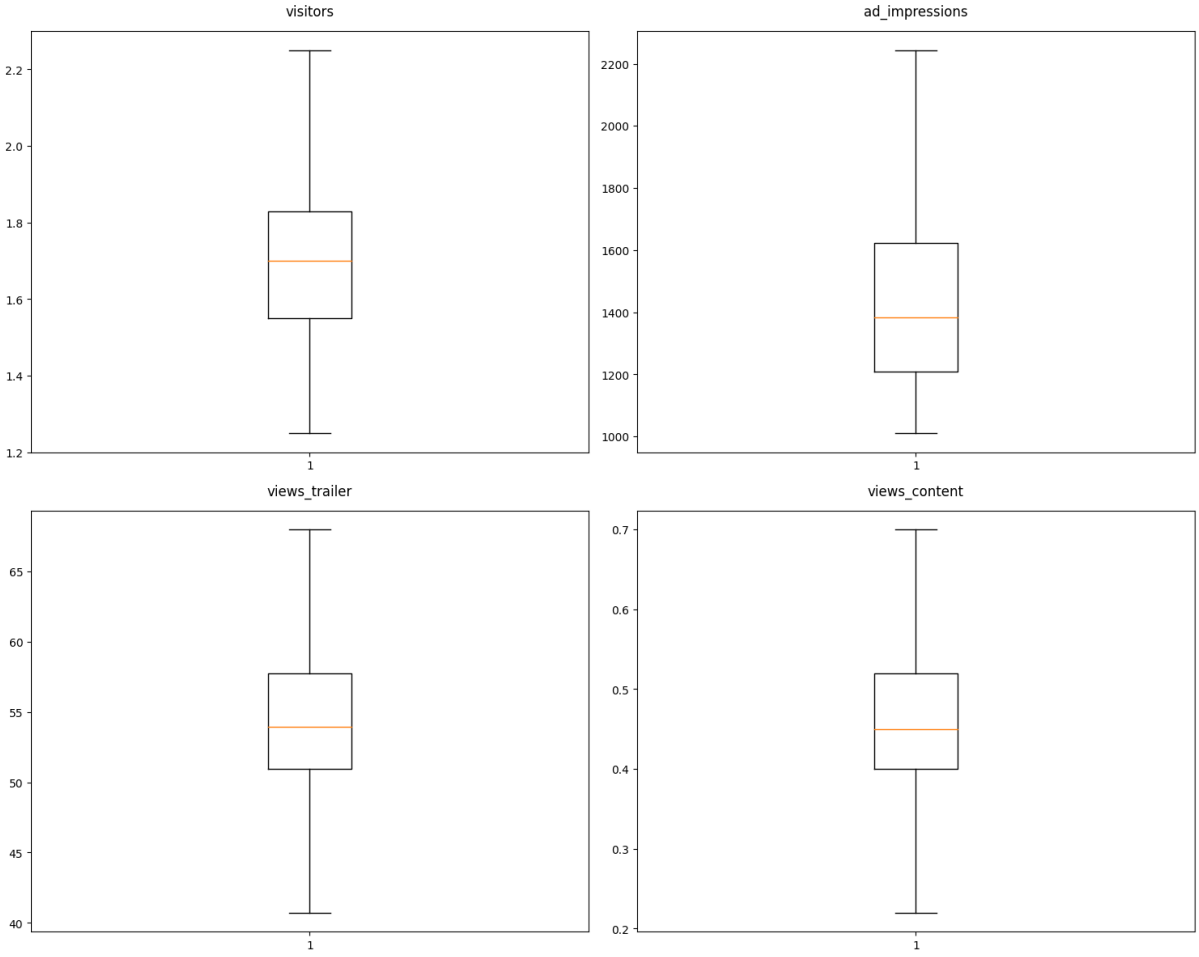

Figure 18. Outliers of numerical columns

In the dataset, outliers were initially identified using the interquartile range (IQR) method, which helps detect values that lie significantly outside the typical range of data. Outliers were found in several key columns:

- **Visitors**: 20 values (2%)
- **Ad Impressions**: 13 values (1.3%)
- **Trailer Views**: 189 values (18.9%)
- **First-Day Views**: 47 values (4.7%)

These outliers were either much smaller or larger than expected, falling outside the range of 1.5 times the IQR from the first and third quartiles. Because outliers can distort statistical analysis and influence the results of a linear regression model, it was essential to address them appropriately.

The outliers were treated using an IQR-based capping method, where values exceeding the upper and lower whiskers were adjusted to the respective limits. This approach ensures that extreme values are brought within a reasonable range without entirely removing them, preserving the overall structure of the data.

After performing this outlier treatment, the dataset was reevaluated, and no outliers remained in any of the columns. This thorough approach ensures the data is now clean, allowing for more reliable analysis and predictive modeling without skewing results due to extreme values.

2.4.Feature Engineering

In Feature Engineering, various feature transformations were applied to make the dataset suitable for modeling. First, one-hot encoding was employed to convert categorical variables into numeric format. This was applied to columns such as major_sports_event, genre, dayofweek, and season. Each category within these columns was transformed into binary columns representing whether a certain category was present (1) or not (0). For example, the genre column was expanded into features like genre_Comedy, genre_Drama, genre_Horror, etc. This step ensures that the model can handle categorical data without introducing multicollinearity, by dropping the first category for each feature.

Additionally, Boolean variables were converted into integers (0 and 1) to maintain a consistent numeric format across all features. This was applied to any remaining Boolean columns generated through the one-hot encoding process, ensuring that all columns in the dataset are numeric and ready for analysis. As a result of these transformations, the dataset now contains 21 columns.

2.5.Data Preparation for Modeling

In preparation for modeling, the dataset underwent a critical transformation to establish the features (independent variables) and target (dependent variable) for analysis. The target variable selected for regression analysis was views content, representing first-day viewership of the content. To facilitate the modeling process, features were extracted by dropping the views content column from the dataset.

The remaining variables, including visitors**,** ad impressions**,** views trailer, and the newly constructed categorical variables, were retained for modeling. To ensure effective training and validation, the dataset was split into training and testing subsets, with 70% of the data for training and 30% for testing. This split allows for robust model evaluation and helps prevent overfitting, ensuring the model generalizes well to new, unseen data.

3.Model building - Linear Regression

3.1.Build the Model and Comment on the Model Statistics:

An Ordinary Least Squares (OLS) regression model is developed to predict content viewership based on a set of independent variables. The final model was optimized by iteratively refining the variables based on their statistical significance and multicollinearity checks.

R-squared: The final model explains 63.9% of the variation in content viewership. This suggests a relatively strong fit between the predictors and the target variable.

Adjusted R-squared: The adjusted R-squared value of 63.2% accounts for the number of predictors in the model, confirming that the model's fit remains stable when adjusted for the number of variables.

Significant Variables:

- **Visitors**: The number of visitors to the platform is positively and significantly correlated with content viewership. This reaffirms that more visitors translate to higher content views.
- **Trailer Views**: Trailer views are highly significant and show a strong positive relationship with content viewership. Exposure to trailers continues to be a major driver of first-day engagement.
- **Major Sports Events**: The presence of major sports events negatively impacts content viewership, as expected, pulling viewers away from the platform.
- **Genre (Sci-Fi)**: Among the genres, Sci-Fi has a positive and statistically significant effect on content views. This indicates a clear preference among viewers for Sci-Fi content on first-day releases.

Day of the Week:

- **Saturday and Sunday**: These days have a significant positive impact on content views, indicating that the weekend is a prime time for content consumption.
- **Tuesday, Thursday, and Wednesday**: These weekdays also show positive effects on viewership, but to a lesser degree compared to the weekend.

Seasons: The seasonal effect is significant, particularly for Summer and Winter, with both showing a notable positive impact on viewership. Spring also has a significant but slightly smaller positive effect.

Multicollinearity: Variance Inflation Factor (VIF) analysis was performed to check for multicollinearity. All VIF values were below the commonly accepted threshold, indicating no serious multicollinearity issues among the remaining variables. Notably, the variable genre Others (VIF: 2.567207) was dropped from the model due to a high VIF, which indicated potential multicollinearity.

Non-Significant Variables (Dropped): Variables like ad impressions, several genres (Comedy, Drama, Horror, Romance, Thriller), and some weekdays (e.g., Monday) were found to be statistically insignificant in explaining the variance in content views and were dropped from the model.

Conclusion: The model reveals key drivers of first-day content viewership, including visitors, trailer views, specific genres, and certain days of the week. These insights can guide platform strategies to optimize content release schedules, marketing efforts, and genre preferences to maximize viewer engagement.

3.2.Display model coefficients with column names :

Column names	Model coefficients
Const (views content)	0.2672
visitors	0.1169
views trailer	0.0093
Major sports event l	0.0622
genre - Sci-Fi	0.0188
Day of week-Saturday	0.0515
Day of week-Sunday	0.0377
Day of week-Thursday	0.0173
Day of week-Tuesday	0.0452
Day of week-Wednesday	0.0397
Season-Spring	0.0265
Season-Summer	0.0435
Season-Winter	0.0295

Table 1. model coefficients with column names

4.Testing the assumptions of linear regression model

4.1.Perform Tests for the Assumptions of Linear Regression

various tests have been conducted to verify the key assumptions of the linear regression model: linearity, independence, normality of residuals, and homoscedasticity. Below are the results of these tests:

1.Linearity and Independence of Errors:

19. Fitted vs Residual plot

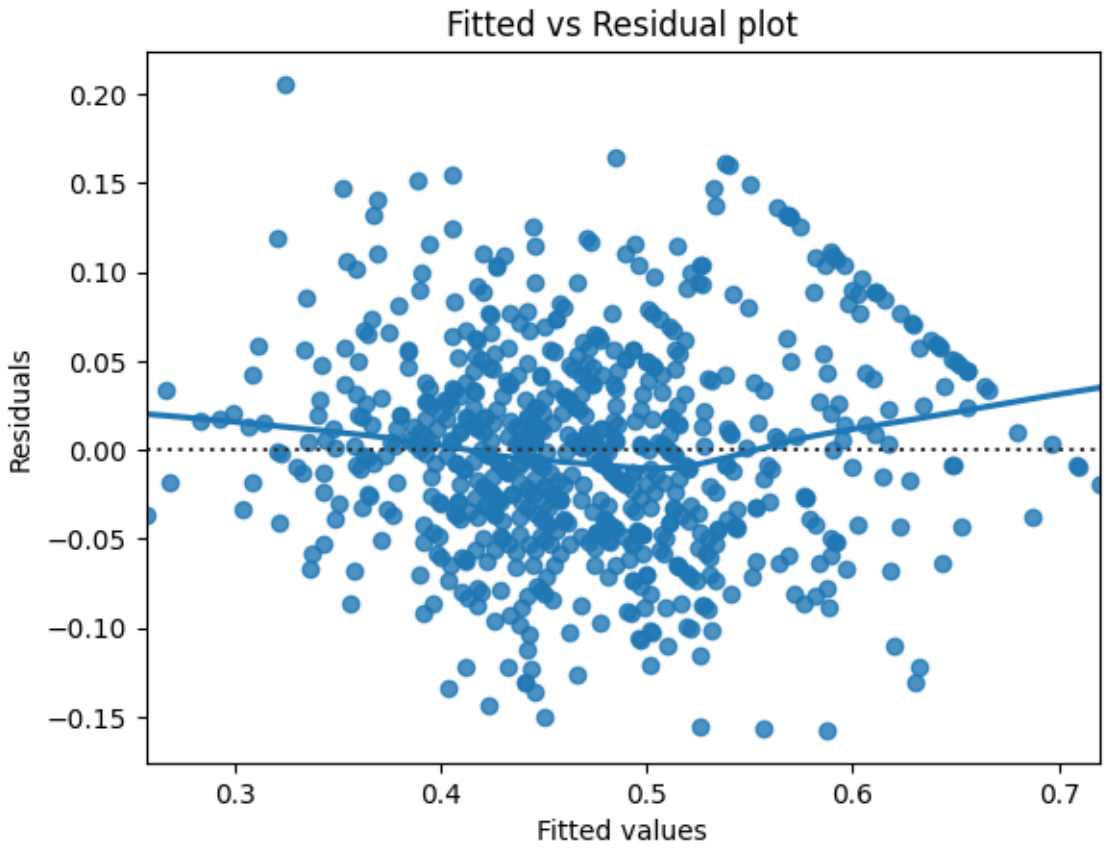

Figure 19. Fitted vs Residual plot

- A **Fitted vs. Residuals plot** was used to check the assumption of linearity and independence. The plot showed a random, dispersed distribution of residuals, which suggests that the model does not violate the assumption of independence. No clear patterns in the residuals confirmed that the relationship between the predictors and the response is linear.

2. Normality of Residuals

- A **histogram of the residuals** was plotted to check for normal distribution. The histogram displayed an approximate bell-shaped curve, supporting the assumption of normality.

20. Normality of residuals

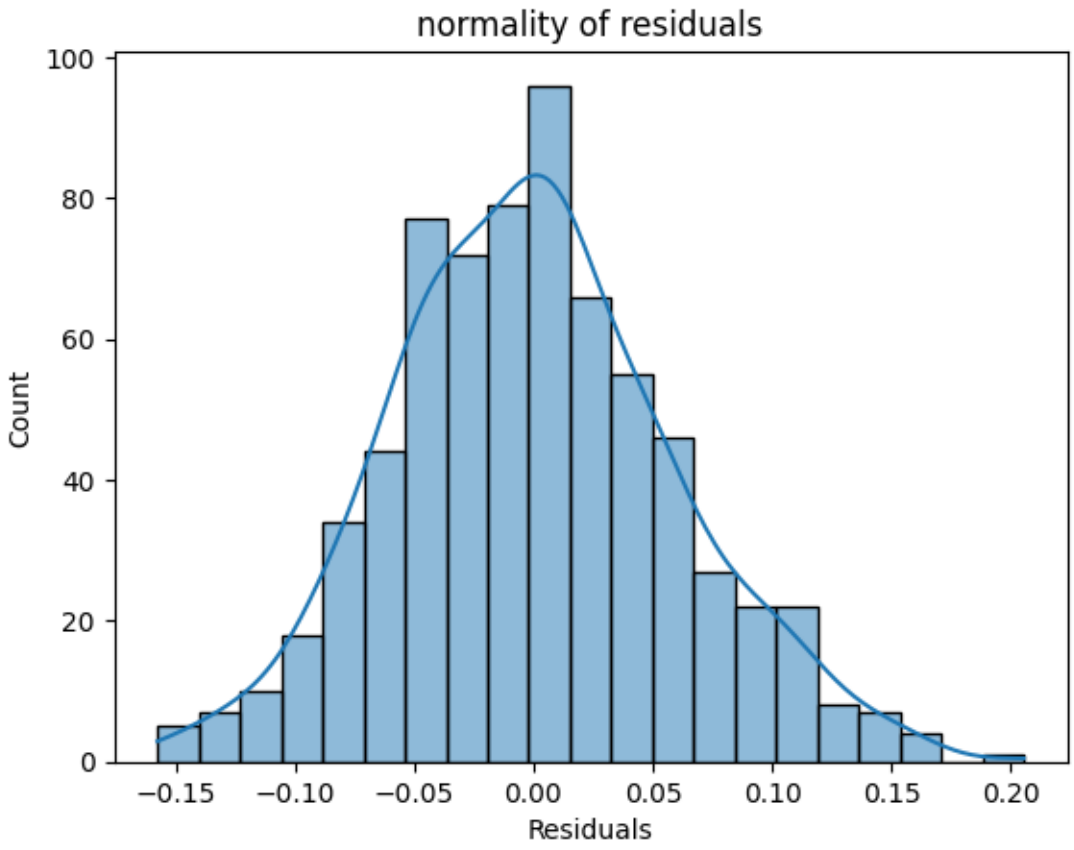

Figure 20. Normality of residuals

- Additionally, the Q-Q plot was used, which showed the residuals approximately aligning along a 45-degree line. This further reinforced the assumption that the residuals are normally distributed.

21. Probability Plot

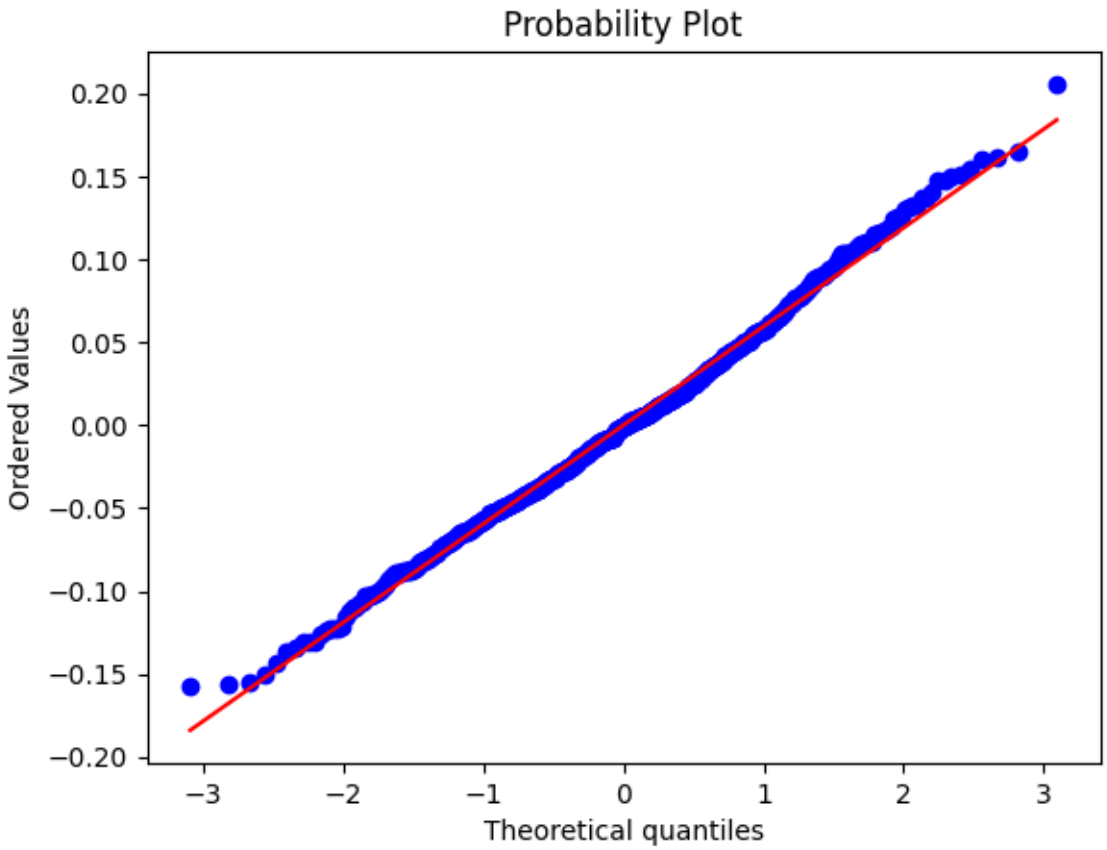

Figure 21. Probability Plot

- However, the Shapiro-Wilk test returned a p-value of 0.0419, slightly below 0.05, suggesting a minor deviation from normality. Despite this, the visual inspection of the residual plots indicated sufficient normality for the model to proceed.

3. Homoscedasticity

- To test for homoscedasticity (constant variance of errors), the Goldfeld-Quandt test was performed. The results indicated a p-value of 0.085, which is above the 0.05 threshold. This suggests that the errors exhibit constant variance, confirming that the homoscedasticity assumption holds.

4.2.Comment on the Findings from the Tests

The tests conducted to verify the assumptions of the linear regression model revealed several important insights:

1. **Linearity and Independence of Errors**:
   The residual plot displayed no clear patterns, confirming that the relationship between the predictors and the target variable is linear. Additionally, the dispersed nature of the residuals indicates that the independence assumption is satisfied, meaning the residuals are not correlated with each other.
2. **Normality of Residuals**:
   While the Shapiro-Wilk test resulted in a **p-value slightly below 0.05**, indicating a minor deviation from normality, the histogram and Q-Q plot visually suggested that the residuals approximate a normal distribution. In practical terms, the normality assumption is adequately met, and this minor deviation is not expected to significantly impact model performance.
3. **Homoscedasticity**:
   The Goldfeld-Quandt test confirmed that the residuals have constant variance, as evidenced by the **p-value of 0.085**. This means the assumption of homoscedasticity holds, ensuring that the variability in the residuals remains constant across all levels of the independent variables.

Conclusion:

The overall results from these tests suggest that the model meets the necessary assumptions to be considered reliable and robust. While there is a slight deviation from perfect normality in the residuals, the visual tests and the lack of any serious violations support the adequacy of the linear regression model for predicting content viewership.

5.Model performance evaluation

5.1.Evaluate the Model on Different Performance Metrics

In this phase, we evaluated the performance of the linear regression model using various metrics to assess its predictive accuracy on the test dataset. The key metrics used are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and their values for both training and test datasets are as follows:

1. **Root Mean Squared Error (RMSE)**:
    - **Training RMSE**: 0.0594
    - **Test RMSE**: 0.0656
2. **Mean Absolute Error (MAE)**:
    - **Training MAE**: 0.0468
    - **Test MAE**: 0.0531

These metrics indicate that the model performs relatively well, with RMSE values close in both training and test datasets, suggesting no significant overfitting. The MAE also shows a slight increase in the test dataset, indicating a marginal decrease in prediction accuracy, which is expected as the model encounters new data.

Overall, the performance metrics reflect a reliable model for predicting content viewership on the platform.

6.Actionable Insights & Recommendations

6.1.Comments on Significance of Predictors

The significance of the predictors in the regression model reveals important insights into what drives content viewership on the platform. Key findings include:

- **Visitors**: This variable had a positive coefficient of 0.1169, indicating that an increase in the number of visitors directly correlates with higher content views, reinforcing the importance of attracting more users to the platform.
- **Trailer Views**: With a coefficient of 0.0093, trailer views are a strong driver of engagement. This suggests that effective marketing through trailers significantly enhances viewer interest and should be prioritized in promotional strategies.
- **Major Sports Events**: This predictor has a negative coefficient of -0.0622, indicating that the occurrence of major sports events tends to divert viewers away from content on the platform, highlighting the need for strategic content scheduling during such events.
- **Genre (Sci-Fi)**: The positive coefficient of 0.0188 for Sci-Fi shows a clear viewer preference for this genre, which should be taken into account when curating content.
- **Days of the Week**: Weekend days (Saturday and Sunday) showed significant positive effects on viewership, suggesting that content releases should be strategically timed for these days to maximize engagement.

6.2.Key Takeaways for the Business

1. **Enhance Marketing Strategies**: With trailer views being a significant predictor, investing in targeted marketing campaigns and promotional content can effectively increase viewer engagement and content consumption.
2. **Optimize Content Scheduling**: The negative impact of major sports events on viewership highlights the need for careful planning around content release dates. The business should avoid launching significant content during major sporting events to prevent viewer diversion.
3. **Expand Popular Genres**: The clear viewer preference for Sci-Fi suggests that the platform should consider increasing the availability of Sci-Fi content. This could lead to higher viewer satisfaction and retention.
4. **Focus on User Acquisition**: Given that the number of visitors is positively correlated with content views, the business should focus on user acquisition strategies to attract more visitors. This could involve leveraging social media, partnerships, and targeted advertisements.
5. **Strategic Weekend Releases**: The analysis shows a marked increase in viewership during weekends. Thus, releasing new content on Fridays or Saturdays could maximize exposure and viewership, aligning with consumer viewing habits.

These insights can guide strategic decisions for optimizing content release schedules, enhancing marketing efforts, and tailoring content offerings to meet viewer preferences, ultimately driving higher engagement and satisfaction on the platform.