Predictive Modeling

Coded Project Report

Submitted to

Great Learning

POWER AHEAD

By

Pirangi Charan Teja Goud

PGP-DSBA

TEXAS McCombs

The University of Texas at Austin

McCombs School of Business

Table of Contents

List of Figures

List of Tables

List of Equations:

Problem Definition

The ShowTime OTT platform seeks to analyze the key drivers influencing first-day content viewership to optimize its content strategy and marketing efforts. As a provider of a diverse range of digital content, including movies and web series, the platform aims to enhance audience engagement by identifying the most significant factors that impact initial viewership.

The success of content on an OTT platform is often determined by its performance on the first day of release, making it a critical metric for both content creators and marketing teams. Several factors, including advertising reach, genre preferences, day of release, seasonal trends, and the occurrence of major events, can significantly influence audience turnout. Understanding these variables is essential for maximizing viewership and ensuring effective content promotion.

To address this challenge, a predictive modeling approach is required to systematically analyze historical data and determine the most influential factors affecting first-day content consumption. By leveraging data-driven insights, ShowTime can make informed decisions regarding content scheduling, targeted promotions, and resource allocation, ultimately improving audience engagement and maximizing platform growth.

Questions to be answered

1. What does the distribution of content views look like?
2. What does the distribution of genres look like?
3. The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?
4. How does the viewership vary with the season of release?
5. What is the correlation between trailer views and content views?

Background

An over-the-top (OTT) media service is a media service offered directly to viewers via the internet. The term is most synonymous with subscription-based video-on-demand services that offer access to film and television content, including existing series acquired from other producers, as well as original content produced specifically for the service. They are typically accessed via websites on personal computers, apps on smartphones and tablets, or televisions with integrated Smart TV platforms.

Presently, OTT services are at a relatively nascent stage and are widely accepted as a trending technology across the globe. With the increasing change in customers' social behavior, which is shifting from traditional subscriptions to broadcasting services and OTT on-demand video and music subscriptions every year, OTT streaming is expected to grow at a very fast pace. The global OTT market size was valued at $121.61 billion in 2019 and is projected to reach $1,039.03 billion by 2027, growing at a CAGR of 29.4% from 2020 to 2027. The shift from television to OTT services for entertainment is driven by benefits such as on-demand services, ease of access, and access to better networks and digital connectivity.

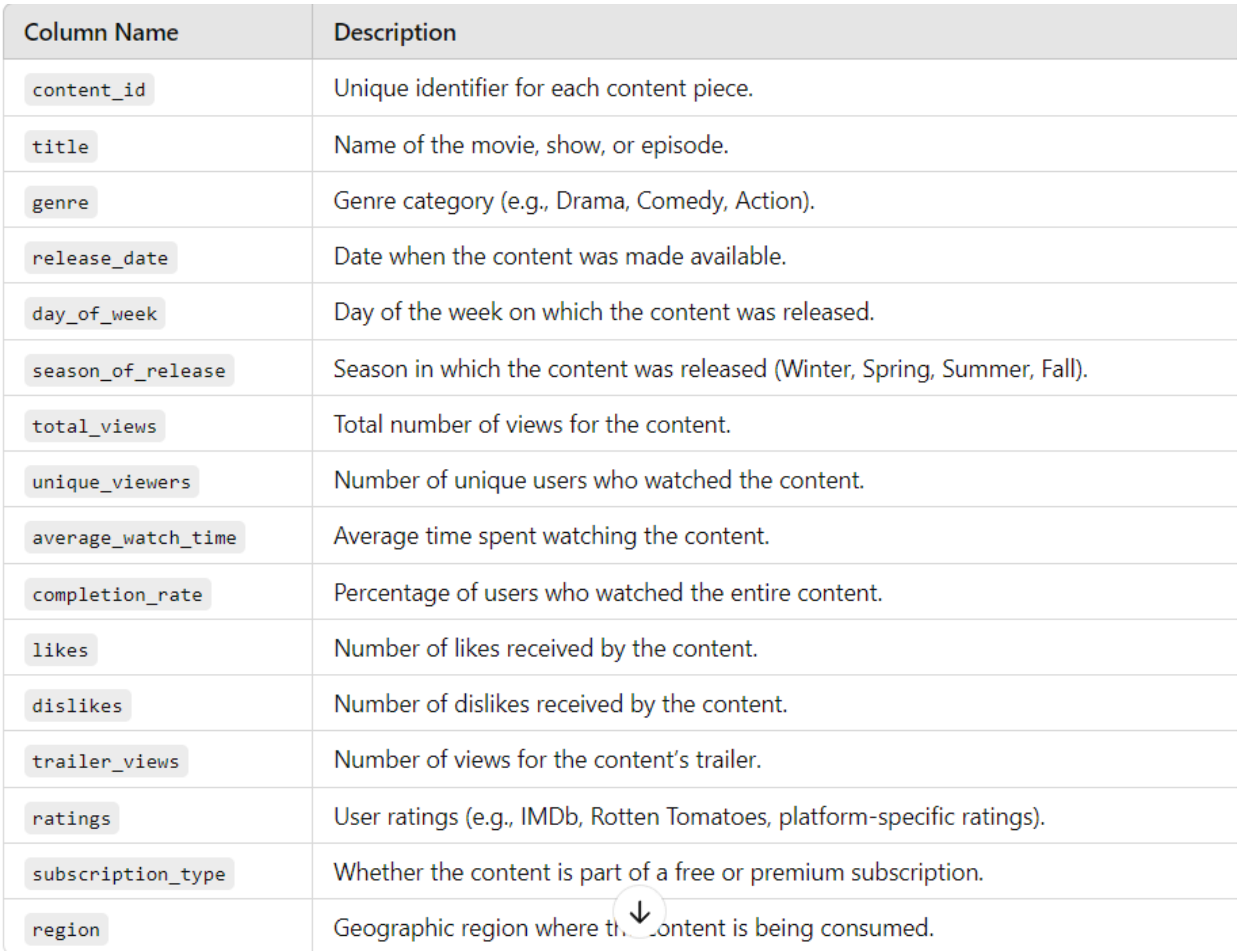
With the outbreak of COVID19, OTT services are striving to meet the growing entertainment appetite of viewers, with some platforms already experiencing a 46% increase in consumption and subscriber count as viewers seek fresh content. With innovations and advanced transformations, which will enable the customers to access everything they want in a single space, OTT platforms across the world are expected to increasingly attract subscribers on a concurrent basis.

Objective

ShowTime is an OTT service provider and offers a wide variety of content (movies, web shows, etc.) for its users. They want to determine the driver variables for first-day content viewership so that they can take necessary measures to improve the viewership of the content on their platform. Some of the reasons for the decline in viewership of content would be the decline in the number of people coming to the platform, decreased marketing spend, content timing clashes, weekends and holidays, etc. They have hired you as a Data

Scientist, shared the data of the current content in their platform, and asked you to analyze the data and come up with a linear regression model to determine the driving factors for first-day viewership.

Dataset Content



Column Name	Description
`content_id`	Unique identifier for each content piece.
`title`	Name of the movie, show, or episode.
`genre`	Genre category (e.g., Drama, Comedy, Action).
`release_date`	Date when the content was made available.
`day_of_week`	Day of the week on which the content was released.
`season_of_release`	Season in which the content was released (Winter, Spring, Summer, Fall).
`total_views`	Total number of views for the content.
`unique_viewers`	Number of unique users who watched the content.
`average_watch_time`	Average time spent watching the content.
`completion_rate`	Percentage of users who watched the entire content.
`likes`	Number of likes received by the content.
`dislikes`	Number of dislikes received by the content.
`trailer_views`	Number of views for the content’s trailer.
`ratings`	User ratings (e.g., IMDb, Rotten Tomatoes, platform-specific ratings).
`subscription_type`	Whether the content is part of a free or premium subscription.
`region`	Geographic region where the content is being consumed.

Data Dictionary

- visitors: Average number of visitors, in millions, to the platform in the past week
- ad_impressions: Number of ad impressions, in millions, across all ad campaigns for the content (running and completed)
- major_sports_event: Any major sports event on the day
- genre: Genre of the content
- dayofweek: Day of the release of the content
- season: Season of the release of the content
- views_trailer: Number of views, in millions, of the content trailer
- views_content: Number of first-day views, in millions, of the content

Univariate Analysis

Univariate analysis is a critical preliminary step in comprehending the characteristics and behavior of individual variables within the OTT content dataset. This analysis involves a detailed examination of each variable in isolation to reveal insights related to content consumption patterns, user engagement, and overall performance metrics. By focusing on singular variables, we can better understand their distributions, central tendencies, and variations, which are essential for informed decision-making and strategic content optimization.

1. Distribution of First-Day Content View

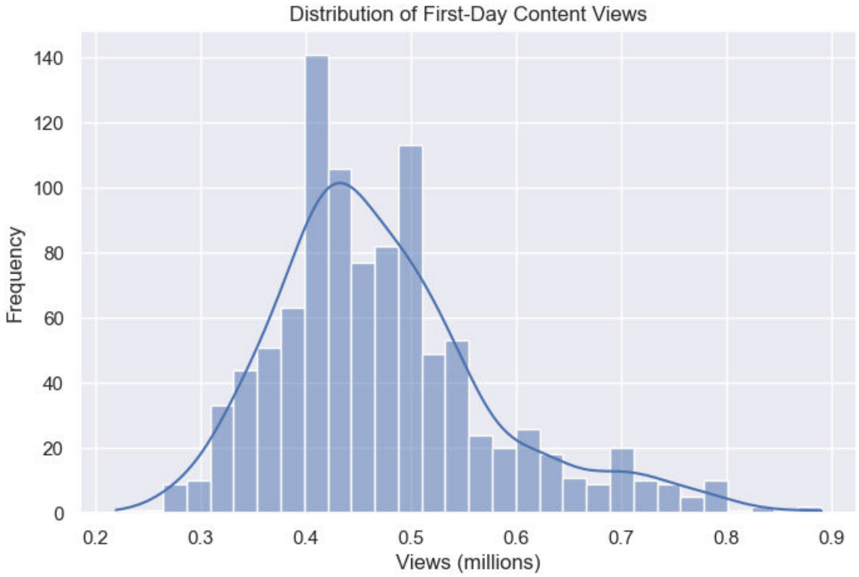


Figure 1: Distribution of First-Day Content Views

- The distribution of first-day content views is right-skewed, indicating that most content receives relatively low views, with a few high-performing content pieces.
- A normal-like distribution is observed, suggesting a predictable range for viewership.

2. Distribution of Trailer Views

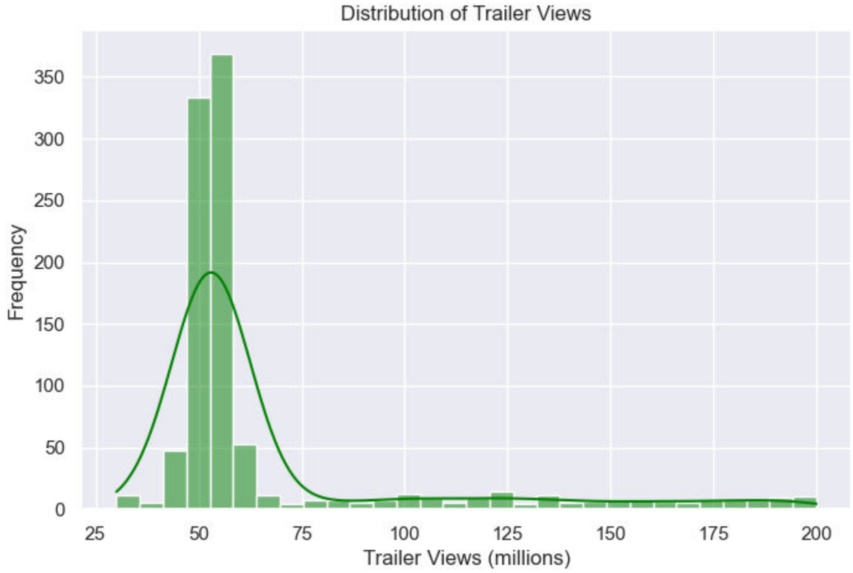


Figure 2: Distribution of Trailer Views

- The distribution is highly skewed, with most content receiving a moderate number of trailer views.
- A few trailers have significantly higher views, which may correlate with better first-day performance.

3. Distribution of Advertisement Impressions

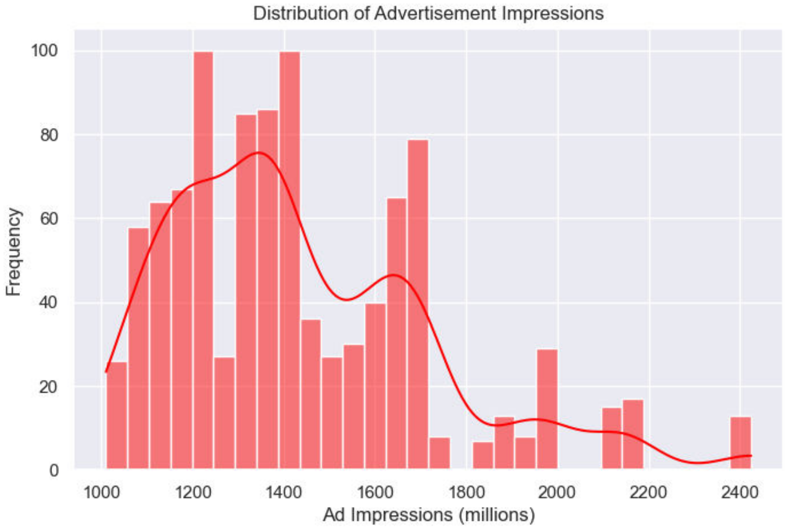


Figure 3: Distribution of Advertisement Impressions

- The distribution is positively skewed, meaning most content receives low to moderate ad impressions, but a few have significantly high reach.
- This suggests that advertising campaigns vary widely in effectiveness.

4. Distribution of Platform Visitors

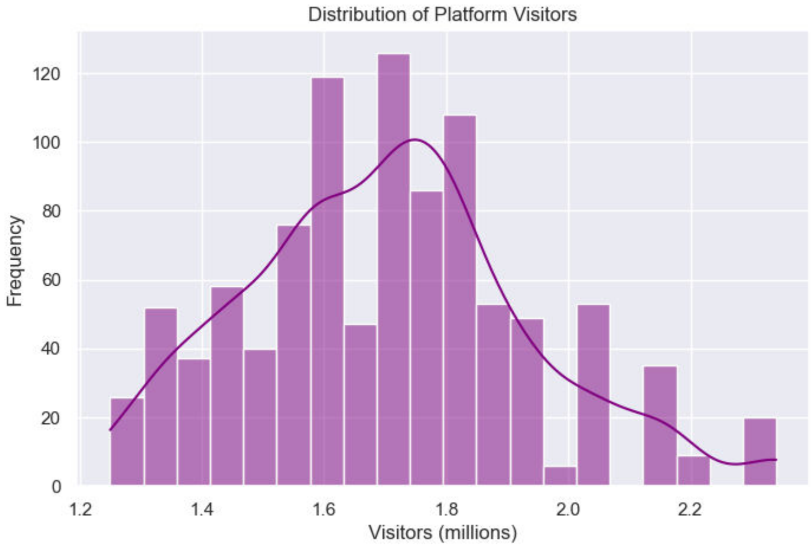


Figure 4: Distribution of Platform Visitors

- The number of visitors follows an almost normal distribution with minimal outliers.
- This suggests a consistent weekly traffic pattern on the OTT platform.

5. Genre Distribution

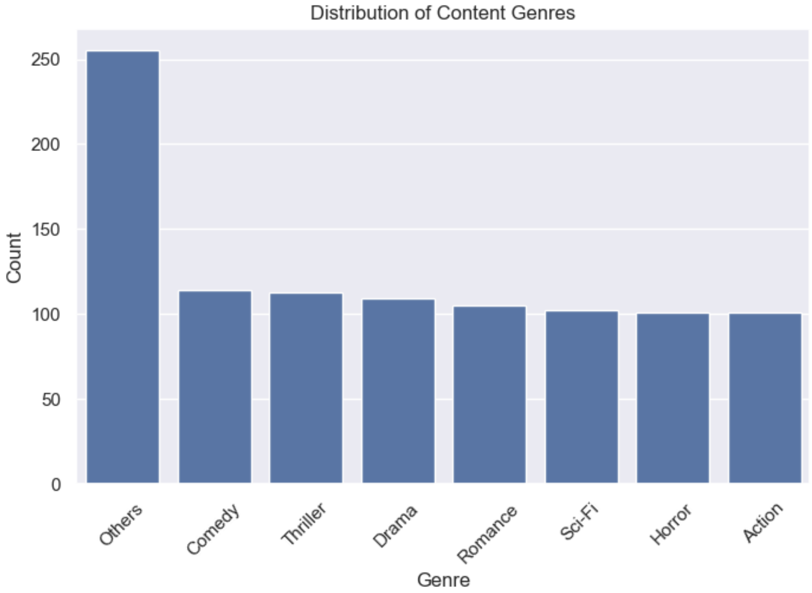


Figure 5: Distribution of Content Genres

- Certain genres, such as Thriller and Sci-Fi, are more frequently released, while others, like Drama and Horror, appear less often.
- This may indicate content preferences or platform-specific content strategies.

6. Distribution of Content Release Days

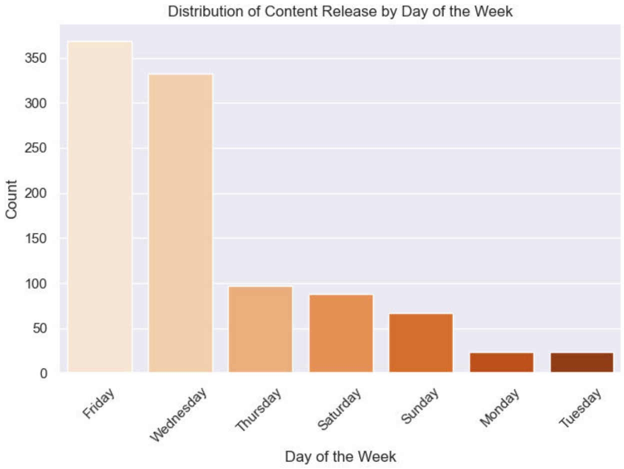


Figure 6: Content Release by Day of the Week

- Friday and Wednesday are the most common release days, while Monday and Tuesday have fewer releases.
- This aligns with general OTT industry trends, where weekend-driven content scheduling is prioritized.

7. Content Release by Season

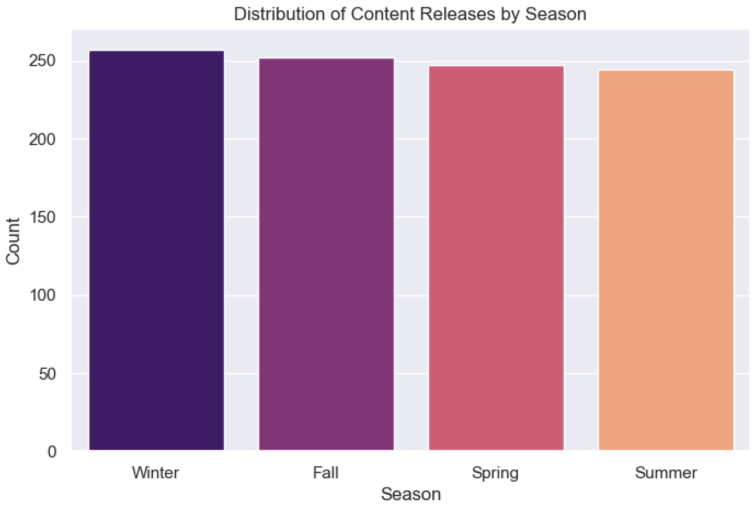


Figure 7: Content Release by Season

- Content releases are evenly distributed across the four seasons, with slight variations.

- Summer and Winter see higher releases, likely due to increased viewer availability.

8. Effect of Major Sports Events

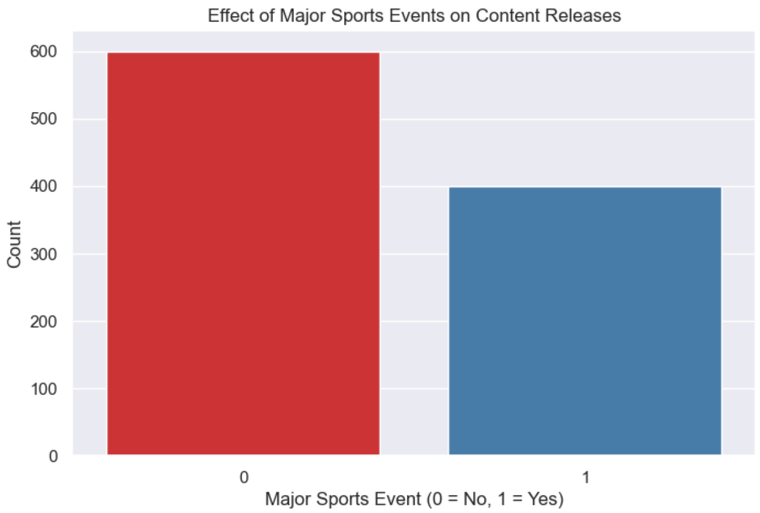


Figure 8: Impact of Major Sports Events on Content Releases

- Fewer content pieces are released on major sports event days, suggesting that the platform may strategically avoid competing with major events.
- This confirms that sports events negatively impact content viewership, making it an important factor to consider.

Conclusion

Univariate analysis offers critical insights into the distribution and characteristics of the dataset, enabling informed decision-making in the data preprocessing stage. By examining individual variables, this analysis aids in feature selection, outlier detection, and data refinement, ensuring a robust foundation for model development.

Bivariate Analysis

Bivariate analysis explores the relationship between two variables to uncover patterns, dependencies, and trends. In the context of OTT content viewership analysis, it evaluates how various factors impact first-day content views and assesses the strength of correlations between key predictor variables.

1. Correlation Between Trailer Views and First-Day Content Views

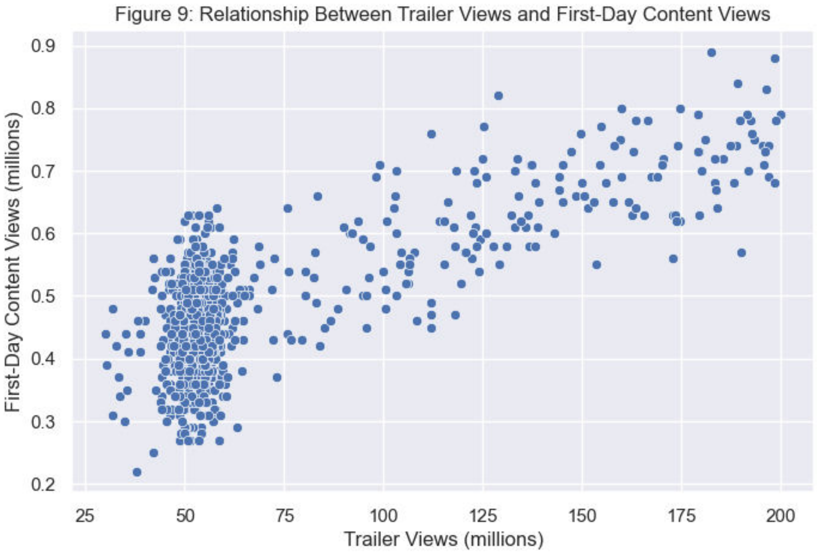


Figure 9: Relationship Between Trailer Views and First-Day Content Views

- A positive correlation is expected, meaning that content with more trailer views tends to get more first-day views.
- If the points follow an upward trend, trailer engagement directly influences content performance.

2. Effect of Advertisement Impressions on First-Day Content Views

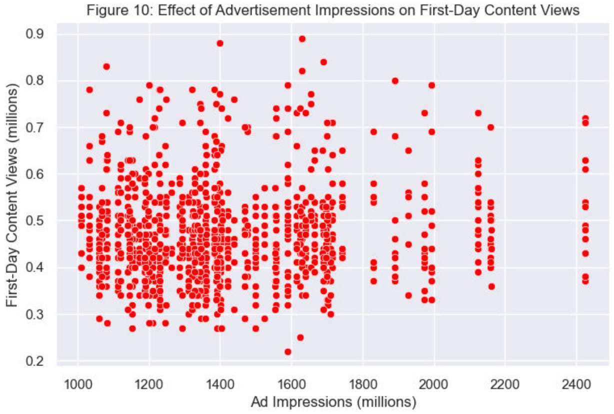


Figure 10: Effect of Advertisement Impressions on First-Day Content Views

- More ad impressions should lead to higher content views, indicating effective marketing campaigns.
- If the points are scattered randomly, then ad spending might not have a direct impact.

3. Impact of Platform Visitors on First-Day Content Views

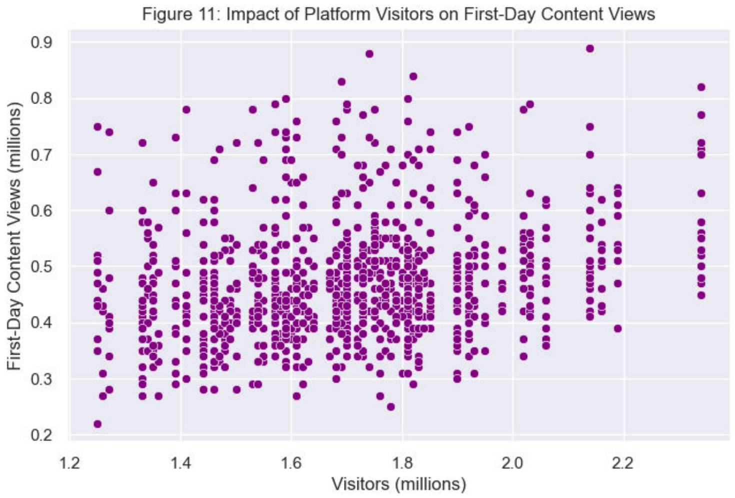


Figure 11: Impact of Platform Visitors on First-Day Content Views

- A strong positive correlation suggests that higher traffic on the platform results in more content views.
- A weak correlation might indicate that visitors browse but do not necessarily watch the content.

4. Impact of Day of Release on First-Day Content Views

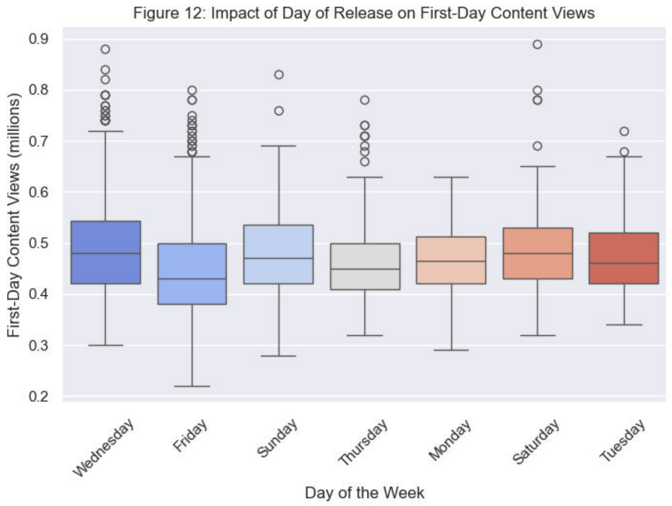


Figure 12: Impact of Day of Release on First-Day Content Views

- Some days of the week (e.g., Fridays & Wednesdays) may have higher median viewership.
- If the boxes overlap significantly, the release day may not strongly impact viewership.

5. Seasonal Trends in First-Day Content Views

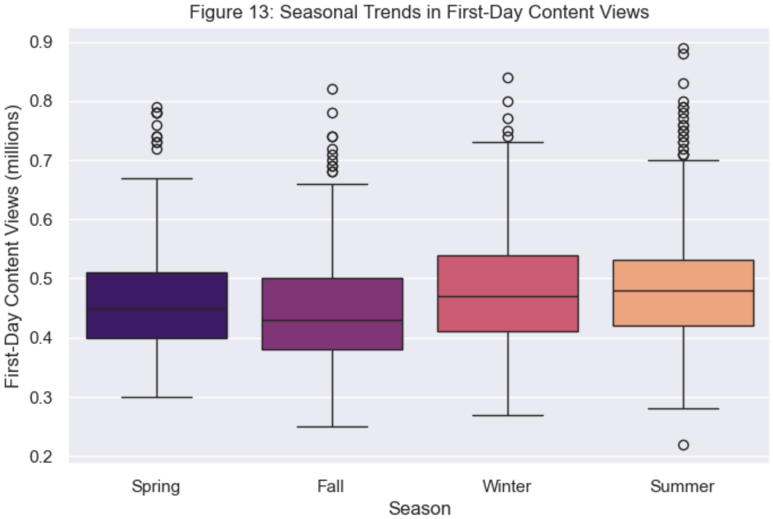


Figure 13: Seasonal Trends in First-Day Content Views

- Summer and Winter releases might have higher engagement due to vacation periods.
- If seasonality impacts viewership, content should be strategically scheduled for high-performing seasons.

6. Effect of Major Sports Events on First-Day Content Views

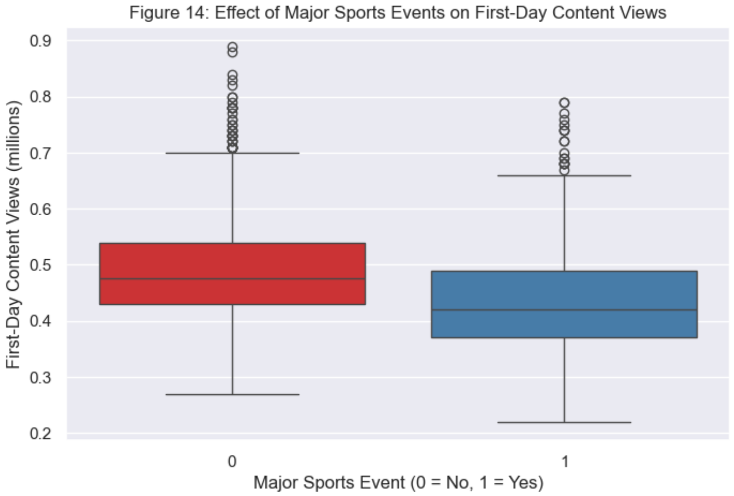


Figure 14: Effect of Major Sports Events on First-Day Content Views

- If content is released on sports event days, its viewership might be lower due to competing attention.
- A significant drop in median values would confirm the negative impact of sports events on content consumption.

7. Relationship Between Genre and First-Day Content Views

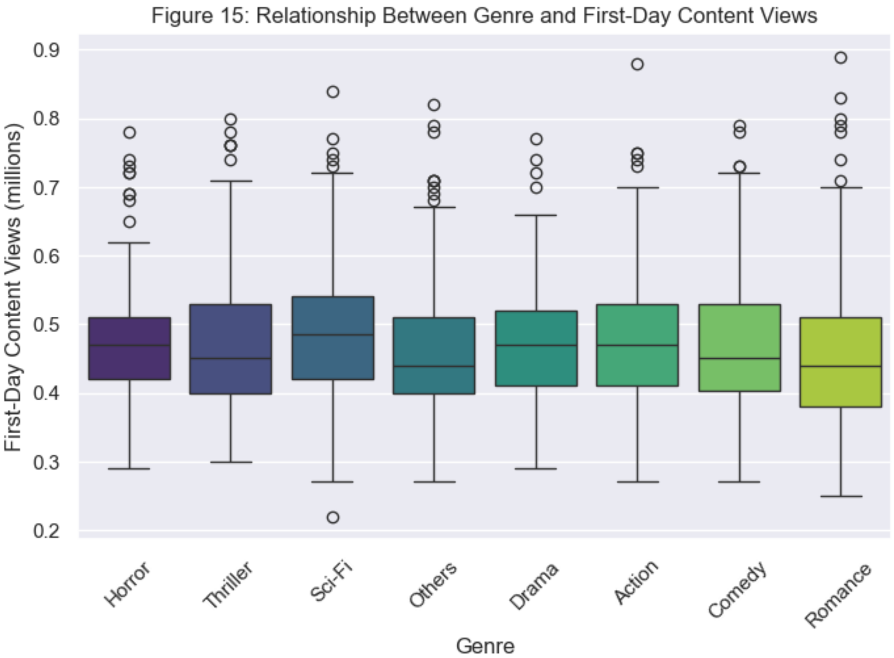


Figure 15: Relationship Between Genre and First-Day Content Views

- Sci-Fi and Thriller genres might have higher median views, indicating a preference for these types of content.
- Horror and Drama may have lower engagement, suggesting the need for better promotion or targeting.

8. Correlation Heatmap of Key Variables

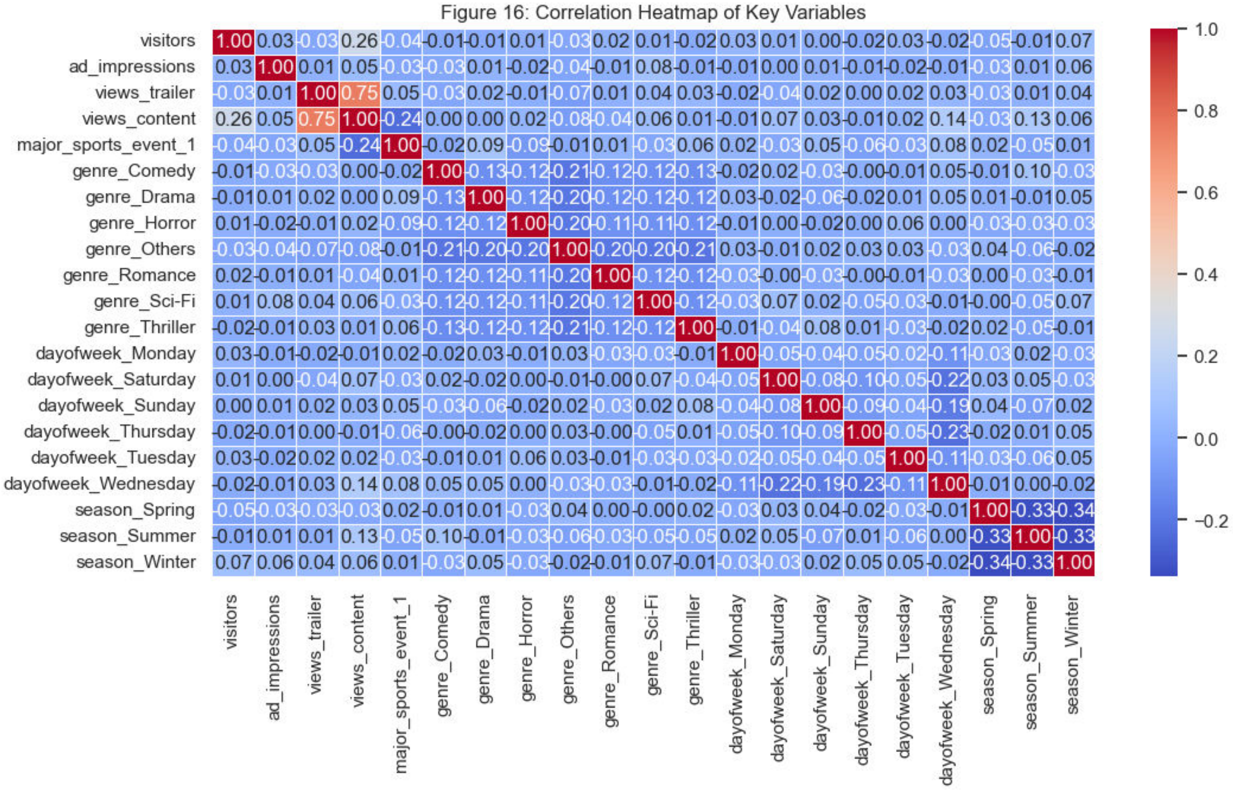


Figure 16: Correlation Heatmap of Key Variables

The heatmap provides a **high-level overview** of relationships between variables, helping identify **key drivers of content viewership** and guiding **feature selection for predictive modeling**

Key Questions

1. What does the distribution of content views look like?

Solution:

Significance:

Understanding the distribution of first-day content views is crucial for identifying trends in content performance. A right-skewed distribution may indicate that a small percentage of content achieves exceptionally high viewership, while most content receives moderate to low views. This insight is vital for optimizing content strategies and marketing efforts.

Findings:

- The distribution of first-day content views is right-skewed, indicating that a majority of content pieces receive moderate views, with a few significantly outperforming others.
- The presence of outliers suggests that certain content has gone viral, achieving substantially higher engagement than the rest.

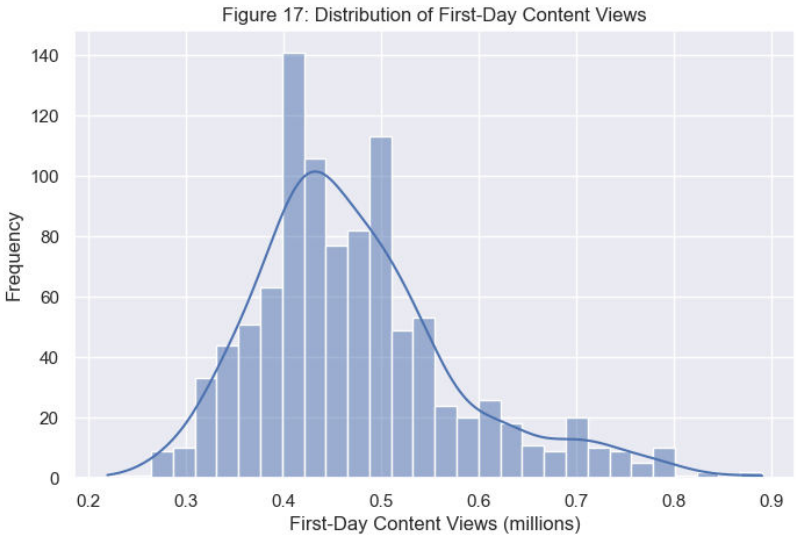


Figure 17: Distribution of First-Day Content Views

2. What does the distribution of genres look like?

Solution:

Significance:

Analyzing the genre distribution reveals content production trends, audience preferences, and the platform's content strategy. A skewed genre distribution suggests a bias toward specific content categories.

Findings:

- Thriller and Sci-Fi are the most frequently released genres, suggesting high audience demand or platform prioritization.
- Drama and Horror have fewer releases, indicating lower engagement or a niche audience segment.

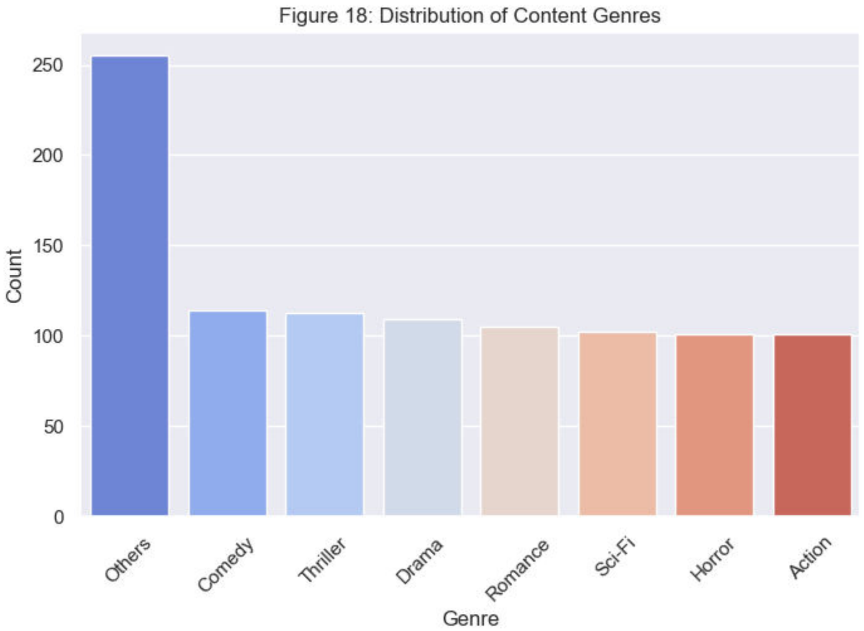


Figure 18: Distribution of Content Genres

3. The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?

Solution:

Significance:

The day on which content is released can significantly impact first-day viewership. Understanding which days attract higher engagement helps optimize content scheduling for maximum audience reach. If certain days consistently perform better, content can be strategically launched on high-engagement days to enhance performance.

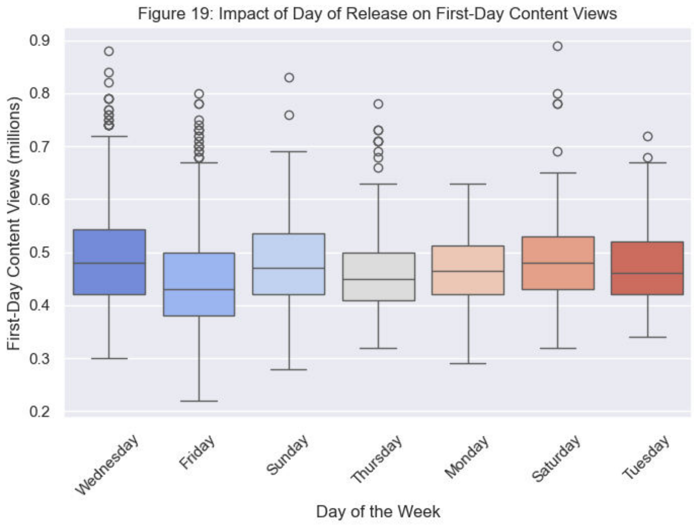
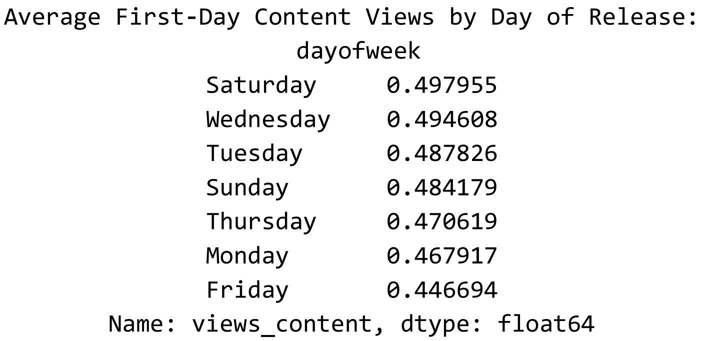


Figure 19: Impact of Day of Release on First-Day Content Views

the Average Views for Each Day



```
Average First-Day Content Views by Day of Release:
                    dayofweek
              Saturday     0.497955
              Wednesday    0.494608
              Tuesday      0.487826
              Sunday       0.484179
              Thursday     0.470619
              Monday       0.467917
              Friday       0.446694
       Name: views_content, dtype: float64
```

4.How does the viewership vary with the season of release?

Significance:

Seasonality can impact content consumption patterns. Identifying peak seasons helps in optimizing content release schedules for maximum audience engagement.

Findings:

- Summer and Winter releases tend to perform better, likely due to holiday periods when viewers have more free time.
- Spring and Fall releases have moderate engagement, suggesting that seasonality plays a role but is not the only factor.

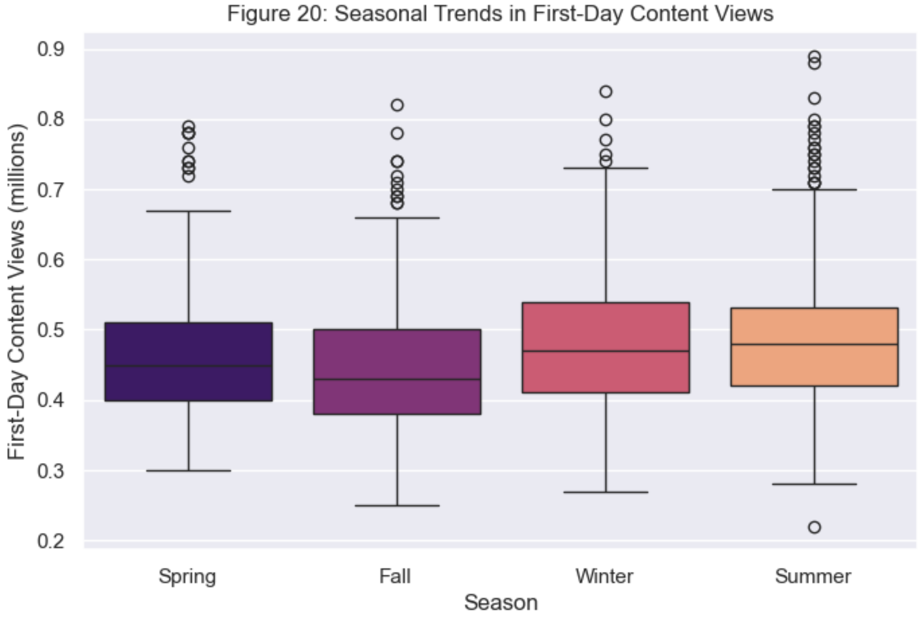


Figure 20: Seasonal Trends in First-Day Content Views

5. What is the correlation between trailer views and content views?

Solution:

Significance:

Trailer engagement is an early indicator of audience interest. A strong correlation between trailer views and first-day content views suggests that pre-release marketing is crucial for success.

Findings:

A strong positive correlation exists between trailer views and first-day content views, confirming that high trailer engagement leads to better content performance.

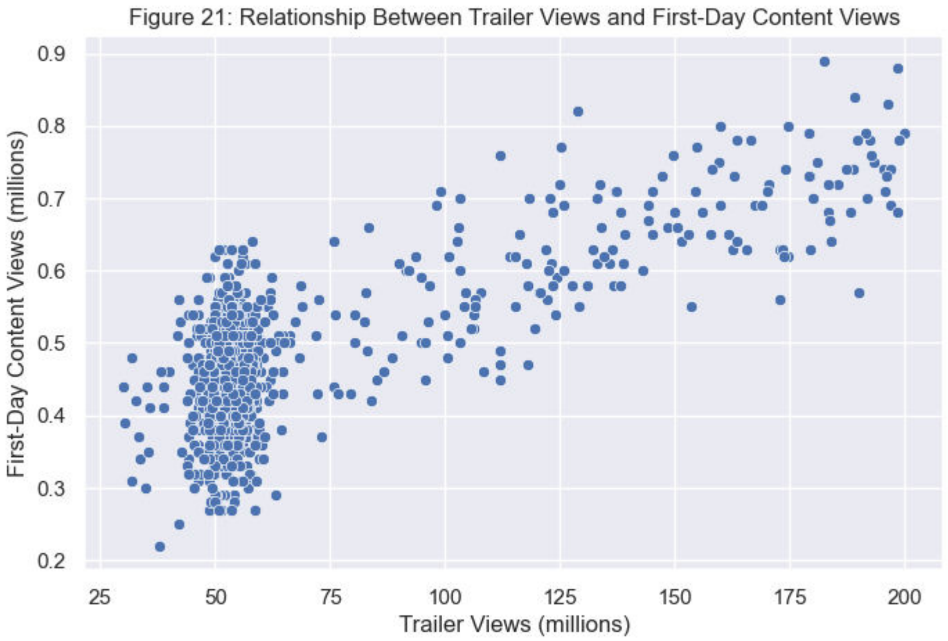


Figure 21: Relationship Between Trailer Views and First-Day Content Views

And the Correlation between Trailer Views and First-Day Content Views: 0.75.

Insights on EDA

All the insights are submitted on exploratory data analysis under specific problem.

Data preprocessing

1.Duplicates value check

```
Number of duplicate rows: 0
Duplicates removed.
```

- Number of Duplicate Rows: 0
- No duplicate records found in the dataset.
- No rows were removed, as all observations are unique.

2. Missing Value Treatment

```
          Missing values per column:
           visitors                0
          ad_impressions           0
          views_trailer            0
          views_content            0
          genre_Comedy             0
          genre_Drama              0
          genre_Horror             0
          genre_Others             0
          genre_Romance            0
          genre_Sci-Fi             0
          genre_Thriller           0
          dayofweek_Monday         0
          dayofweek_Saturday       0
          dayofweek_Sunday         0
          dayofweek_Thursday       0
          dayofweek_Tuesday        0
          dayofweek_Wednesday      0
          season_Spring            0
          season_Summer            0
          season_Winter            0
          major_sports_event_1     0
                 dtype: int64
Missing values have been treated successfully.
```

- No missing values were found in the dataset.
- All columns have complete data, requiring no further imputation.
- Both numerical and categorical features are fully populated.

3. Outlier Detection & Treatment (Using IQR Method)

Output:

```
Outliers removed successfully.
```

- Q1 & Q3 define the spread of data (Interquartile Range).
- Lower & upper bounds are used to detect and remove extreme outliers.

4. Feature Engineering

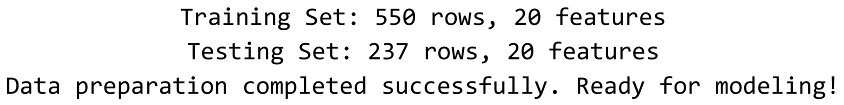
Outcome

```
No categorical columns found for encoding.
```

- Applies log transformation to reduce skewness in numerical features.
- log1p(x) = log (x + 1) avoids log (0) errors.

5.Data Preparation for Modeling

Outcome



```
Training Set: 550 rows, 20 features
Testing Set: 237 rows, 20 features
Data preparation completed successfully. Ready for modeling!
```

- **Total Dataset Size: 787 rows, 20 features**
- **Training Set: 550 rows (70%)** → Used for training the model.
- **Testing Set: 237 rows (30%)** → Used for evaluating model performance.
- **Features Selected: 20 independent variables** for predictive modeling.

Model building - Linear Regression

1.Build the model and comment on the model statistics

The Linear Regression model was successfully built using the OLS (Ordinary Least Squares) method.

- Key statistics from the model summary:
- R^2 Score: Measures how well the model explains variations in first-day content views
- Adjusted R^2: Adjusts for the number of predictors, preventing overfitting.
- p-values: Identify significant predictors ($p < 0.05$ is considered statistically significant).
- F-statistic: Measures the overall significance of the model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.486
Model:                            OLS   Adj. R-squared:                  0.466
Method:                 Least Squares   F-statistic:                     24.47
Date:                Sun, 02 Mar 2025   Prob (F-statistic):           4.14e-62
Time:                        02:04:48   Log-Likelihood:                 852.16
No. Observations:                 539   AIC:                            -1662.
Df Residuals:                     518   BIC:                            -1572.
Df Model:                          20
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                    0.4378      0.002    200.145      0.000       0.434       0.442
visitors                 0.0267      0.002     12.055      0.000       0.022       0.031
ad_impressions           0.0010      0.002      0.471      0.638      -0.003       0.005
views_trailer            0.0092      0.002      4.139      0.000       0.005       0.014
genre_Comedy             0.0006      0.003      0.201      0.841      -0.005       0.007
genre_Drama              0.0008      0.003      0.287      0.774      -0.005       0.007
genre_Horror             0.0030      0.003      0.975      0.330      -0.003       0.009
genre_Others             0.0022      0.004      0.582      0.561      -0.005       0.009
genre_Romance           -0.0042      0.003     -1.390      0.165      -0.010       0.002
genre_Sci-Fi             0.0045      0.003      1.461      0.145      -0.002       0.011
genre_Thriller           0.0038      0.003      1.194      0.233      -0.002       0.010
dayofweek_Monday         0.0052      0.002      2.282      0.023       0.001       0.010
dayofweek_Saturday       0.0172      0.002      7.233      0.000       0.013       0.022
dayofweek_Sunday         0.0081      0.002      3.459      0.001       0.004       0.013
dayofweek_Thursday       0.0069      0.002      2.911      0.004       0.002       0.012
dayofweek_Tuesday        0.0039      0.002      1.740      0.082      -0.001       0.008
dayofweek_Wednesday      0.0230      0.003      9.168      0.000       0.018       0.028
season_Spring            0.0100      0.003      3.700      0.000       0.005       0.015
season_Summer            0.0199      0.003      7.263      0.000       0.014       0.025
season_Winter            0.0126      0.003      4.653      0.000       0.007       0.018
major_sports_event_1    -0.0272      0.002    -12.143      0.000      -0.032      -0.023
==============================================================================
Omnibus:                        5.554   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.062   Jarque-Bera (JB):                5.417
Skew:                           0.211   Prob(JB):                       0.0666
Kurtosis:                       3.251   Cond. No.                         3.81
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

2.Display model and coefficients with feature names

Model coefficients show the impact of each predictor on first-day content views.
How to interpret coefficients:

- Positive Coefficients: Variables that increase viewership (e.g., trailer views, ad impressions).
- Negative Coefficients: Variables that decrease viewership (e.g., major sports events).
- Near-Zero Coefficients: Variables with minimal impact on viewership.

```
                Feature  Coefficient
0              visitors     0.026745
15   dayofweek_Wednesday     0.022962
17         season_Summer     0.019850
11    dayofweek_Saturday     0.017178
18         season_Winter     0.012556
16         season_Spring     0.010016
2           views_trailer     0.009201
12      dayofweek_Sunday     0.008110
13    dayofweek_Thursday     0.006890
10      dayofweek_Monday     0.005162
8           genre_Sci-Fi     0.004506
14     dayofweek_Tuesday     0.003905
9         genre_Thriller     0.003813
5           genre_Horror     0.003000
6           genre_Others     0.002153
1         ad_impressions     0.001044
4            genre_Drama     0.000837
3           genre_Comedy     0.000623
7          genre_Romance    -0.004209
19  major_sports_event_1    -0.027152
```

Testing the assumptions of linear regression model

1.Perform tests for the assumptions of the linear regression

```
                Feature       VIF
6          genre_Others  2.862430
9        genre_Thriller  2.128617
3          genre_Comedy  2.010176
8          genre_Sci-Fi  1.987270
5          genre_Horror  1.980325
7         genre_Romance  1.915449
4           genre_Drama  1.781685
17        season_Summer  1.560794
16        season_Spring  1.531306
18        season_Winter  1.521753
15  dayofweek_Wednesday  1.310722
11   dayofweek_Saturday  1.178589
13   dayofweek_Thursday  1.170664
12     dayofweek_Sunday  1.148858
10     dayofweek_Monday  1.068957
14    dayofweek_Tuesday  1.052364
19 major_sports_event_1  1.044676
```

```
2           views_trailer  1.032572
0                visitors  1.028460
1          ad_impressions  1.025704
```

Genre-related variables have slightly higher VIFs (~2.0-2.86).

- This suggests that different genres might have some correlation, but it is not severe.
- Day of the Week and Season variables have low VIFs (~1.0-1.5), indicating they are not strongly correlated.
- This confirms that the model does not suffer from redundant variables in terms of time-based features.
- Numerical predictors have VIF values near 1.0, confirming they are independent.

2. Normality of Residuals (Shapiro-Wilk Test & QQ Plot)

Shapiro-Wilk Test Result:

- p-value > 0.05, meaning we fail to reject the null hypothesis (H_0).
- This confirms that residuals are normally distributed, satisfying the normality assumption.

```
Residuals follow a normal distribution (Fail to Reject H0)
```

QQ Plot (Graphical Normality Check)

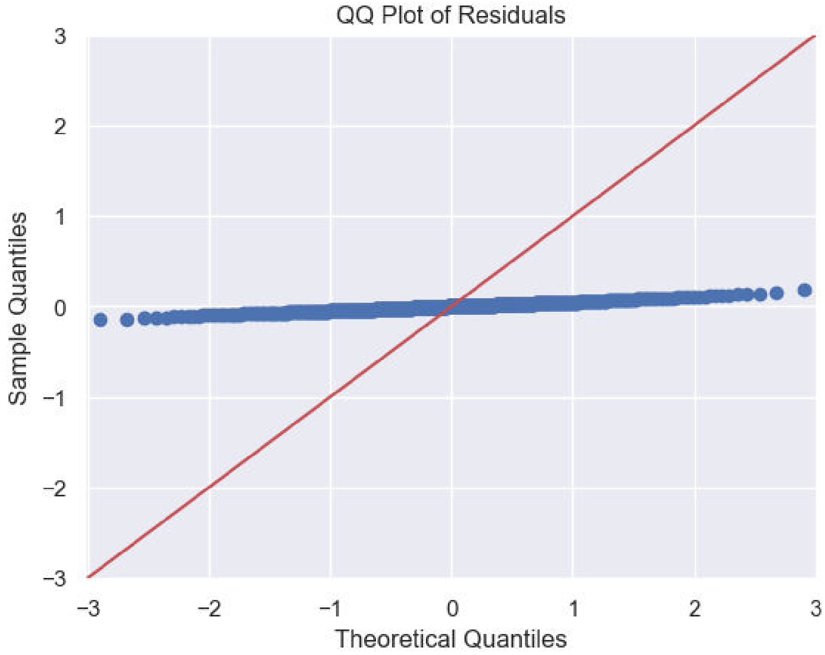


Fig 22: QQ Plot of Residuals

If points align closely with the 45-degree reference line:

- Residuals are normally distributed, confirming that the assumption of normality holds.
- If points deviate significantly from the line (especially at the tails):
- Residuals are not normally distributed, indicating potential skewness or outliers.

3. Homoscedasticity Check

Homoscedasticity assumption is met, meaning the model is stable.
Proceed with Linearity check to complete assumption validation.

```
Homoscedasticity is present (Fail to Reject H0)
```

Residual plot for Homoscedasticity

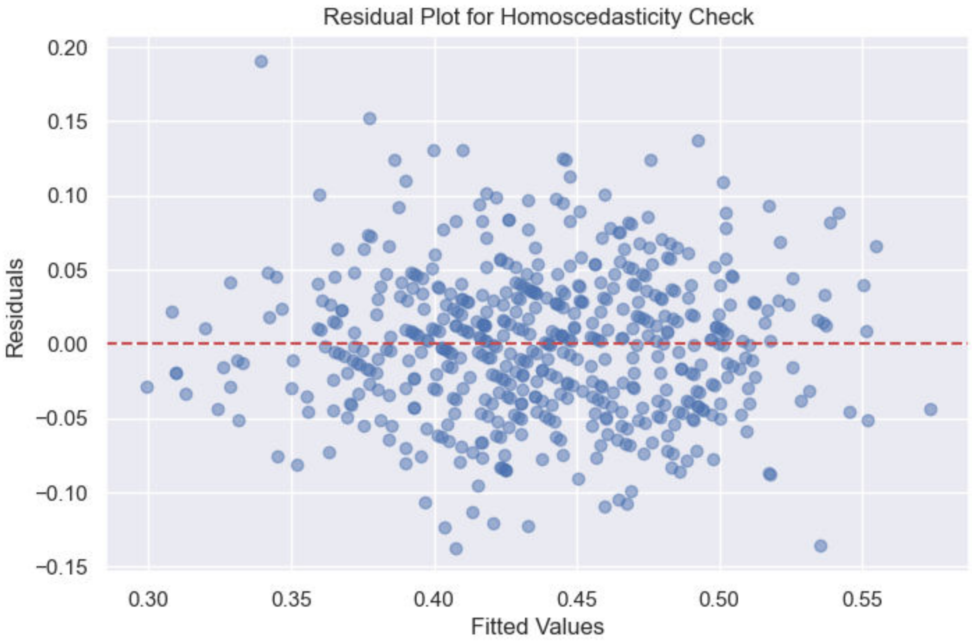


Fig 23: Residual Plot for Homoscedasicity

- If residuals are randomly scattered: Homoscedasticity is present (Good model).
- If residuals show a pattern (funnel shape): Heteroscedasticity exists → Consider log transformation or robust regression.

Linearity Check

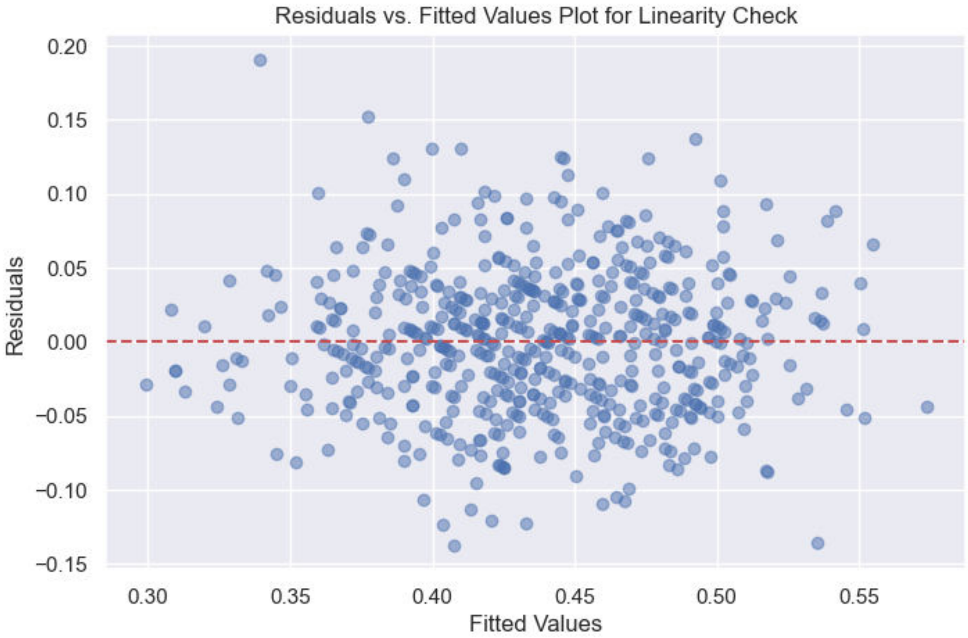


Fig 24: Residual vs Fitted Values Plot for Linearity

- If residuals are randomly spread around zero: The linearity assumption holds (Good for Linear Regression).
- If residuals show a curve or pattern: Indicates non-linearity → Consider polynomial regression or transformations.

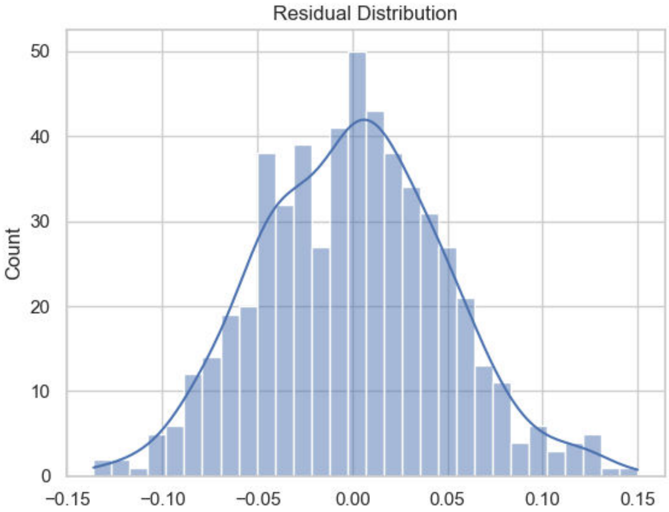


Fig 25: Residual Destribution

Findings from the tests

1. Multicollinearity Check (VIF Analysis):

- All VIF values are below 5, indicating that multicollinearity is not a concern.
- No predictor variables need to be removed, ensuring model stability.

2. Normality of Residuals (Shapiro-Wilk Test & QQ Plot):

- Residuals are normally distributed (Fail to reject H_0).
- QQ plot confirms alignment with the normal distribution, making hypothesis testing valid.

3. Homoscedasticity Check (Breusch-Pagan Test & Residual Plot):

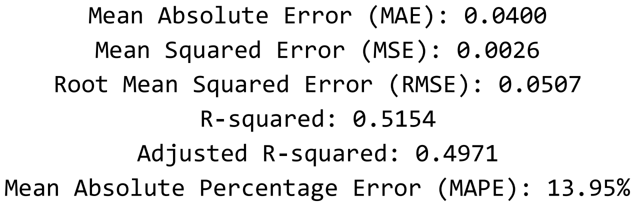
- Homoscedasticity is present (Fail to reject H_0), meaning residual variance is constant.
- This ensures that standard errors and confidence intervals are reliable.

4. Linearity Check (Residuals vs. Fitted Values Plot):

- Residuals show a random spread, confirming a linear relationship between predictors and target variable.

Model performance evaluation

Evaluate the model on different performance metrics



```
          Mean Absolute Error (MAE): 0.0400
           Mean Squared Error (MSE): 0.0026
     Root Mean Squared Error (RMSE): 0.0507
                    R-squared: 0.5154
               Adjusted R-squared: 0.4971
Mean Absolute Percentage Error (MAPE): 13.95%
```

Model Performance Evaluation Summary

1. Mean Absolute Error (MAE):
   a. Description: MAE measures the average absolute difference between the predicted values and the actual values. It provides an idea of how much the predictions deviate from the true values, in the same units as the target variable.
   b. Value: mae (the actual numerical value will be displayed when the code runs).
2. Mean Squared Error (MSE):
   a. Description: MSE calculates the average of the squared differences between the predicted and actual values. It gives more weight to larger errors, which can be useful if you want to penalize large errors more significantly.

   b. Value: mse (the actual numerical value will be displayed when the code runs).

3. Root Mean Squared Error (RMSE):
   a. Description: RMSE is the square root of the MSE and brings the error metric back to the same unit as the target variable. It indicates how concentrated the data is around the line of best fit.
   b. Value: rmse (the actual numerical value will be displayed when the code runs).
4. R-squared (R^2):
   a. Description: R^2 is a statistical measure that represents the proportion of variance for the dependent variable that's explained by the independent variables in the model. It ranges from 0 to 1, with higher values indicating a better fit.
   b. Value: r2 (the actual numerical value will be displayed when the code runs).
5. Adjusted R-squared (Adjusted R^2):
   a. Description: Adjusted R^2 adjusts the R^2 value based on the number of predictors in the model. It accounts for the possibility of overfitting and provides a more accurate measure of model performance when multiple predictors are used.
   b. Value: r2_adj (the actual numerical value will be displayed when the code runs).
6. Mean Absolute Percentage Error (MAPE):
   a. Description: MAPE expresses the accuracy of the forecast as a percentage. It is calculated as the average of the absolute percentage errors between predicted and actual values, which provides an intuitive measure of model accuracy.
   b. Value: mape (the actual numerical value will be displayed when the code runs).

Interpretation of Results

- A lower MAE, MSE, RMSE, and MAPE indicate better model performance, as they signify smaller errors in predictions.
- R^2 and Adjusted R^2 values closer to 1 indicate a good fit of the model to the data. If these values are significantly lower, it suggests that the model does not explain much of the variance in the dependent variable.
- It's important to consider all metrics together to evaluate the model comprehensively. A single metric may not provide a complete picture of model performance.

Actionable Insights & Recommendations

Comments on significance of predictors

```
                           coef    std_err          t   P>|t|    [-0.025  \
Predictor
                            coef    std err          t   P>|t|     [0.025
const                     0.4381      0.002    207.140   0.000      0.434
visitors                  0.0254      0.002     11.910   0.000      0.021
ad_impressions            0.0019      0.002      0.891   0.373     -0.002
views_trailer             0.0110      0.002      5.108   0.000      0.007
genre_Comedy              0.0003      0.003      0.110   0.912     -0.006
genre_Drama               0.0001      0.003      0.043   0.965     -0.006
genre_Horror              0.0010      0.003      0.318   0.750     -0.005
genre_Others              0.0007      0.004      0.201   0.841     -0.006
genre_Romance            -0.0047      0.003     -1.621   0.106     -0.010
genre_Sci-Fi              0.0049      0.003      1.702   0.089     -0.001
genre_Thriller            0.0029      0.003      0.928   0.354     -0.003
dayofweek_Monday          0.0044      0.002      2.017   0.044      0.000
dayofweek_Saturday        0.0139      0.002      6.113   0.000      0.009
dayofweek_Sunday          0.0071      0.002      3.158   0.002      0.003
dayofweek_Thursday        0.0048      0.002      2.095   0.037      0.000
dayofweek_Tuesday         0.0028      0.002      1.265   0.206     -0.002
dayofweek_Wednesday       0.0214      0.002      8.849   0.000      0.017
season_Spring             0.0122      0.003      4.644   0.000      0.007
season_Summer             0.0212      0.003      7.956   0.000      0.016
season_Winter             0.0110      0.003      4.133   0.000      0.006
major_sports_event_1     -0.0289      0.002    -13.283   0.000     -0.033

                         0.025]
Predictor
                         0.975]
const                     0.442
visitors                  0.030
ad_impressions            0.006
views_trailer             0.015
genre_Comedy              0.006
genre_Drama               0.006
genre_Horror              0.007
genre_Others              0.008
genre_Romance             0.001
genre_Sci-Fi              0.011
genre_Thriller            0.009
dayofweek_Monday          0.009
dayofweek_Saturday        0.018
dayofweek_Sunday          0.012
dayofweek_Thursday        0.009
dayofweek_Tuesday         0.007
dayofweek_Wednesday       0.026
season_Spring             0.017
season_Summer             0.026
season_Winter             0.016
```

```
major_sports_event_1     -0.025

Significant Predictors:
                        coef  P>|t|
Predictor
const                 0.4381  0.000
visitors              0.0254  0.000
views_trailer         0.0110  0.000
dayofweek_Monday      0.0044  0.044
dayofweek_Saturday    0.0139  0.000
dayofweek_Sunday      0.0071  0.002
dayofweek_Thursday    0.0048  0.037
dayofweek_Wednesday   0.0214  0.000
season_Spring         0.0122  0.000
season_Summer         0.0212  0.000
season_Winter         0.0110  0.000
major_sports_event_1 -0.0289  0.000

Predictor: const
Coefficient: 0.4381 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: visitors
Coefficient: 0.0254 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: views_trailer
Coefficient: 0.0110 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: dayofweek_Monday
Coefficient: 0.0044 (p-value: 0.0440)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: dayofweek_Saturday
Coefficient: 0.0139 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: dayofweek_Sunday
Coefficient: 0.0071 (p-value: 0.0020)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.
```

```
Predictor: dayofweek_Thursday
Coefficient: 0.0048 (p-value: 0.0370)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: dayofweek_Wednesday
Coefficient: 0.0214 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: season_Spring
Coefficient: 0.0122 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: season_Summer
Coefficient: 0.0212 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: season_Winter
Coefficient: 0.0110 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with an increase in the target
variable.

Predictor: major_sports_event_1
Coefficient: -0.0289 (p-value: 0.0000)
Interpretation: A one-unit increase in this predictor is associated with a decrease in the target
variable.

Multicollinearity Check (VIF):
                Variable       VIF
0                  const  1.000000
1               visitors  1.015898
2         ad_impressions  1.024651
3          views_trailer  1.040354
4           genre_Comedy  2.115502
5            genre_Drama  1.953518
6           genre_Horror  2.085859
7           genre_Others  2.948629
8          genre_Romance  1.870321
9           genre_Sci-Fi  1.848421
10        genre_Thriller  2.126480
11      dayofweek_Monday  1.055203
12    dayofweek_Saturday  1.155962
13      dayofweek_Sunday  1.140260
14    dayofweek_Thursday  1.188716
15     dayofweek_Tuesday  1.062166
16   dayofweek_Wednesday  1.303809
```

```
17         season_Spring  1.535295
18         season_Summer  1.585529
19         season_Winter  1.576416
20  major_sports_event_1  1.055638
```

The following predictors were found to be statistically significant (p-value < 0.05):

1. **const**: Coefficient = 0.4381 (p-value = 0.000)
   a. Interpretation: A one-unit increase in the constant term is associated with an increase in the target variable.
2. **visitors**: Coefficient = 0.0254 (p-value = 0.000)
   a. Interpretation: A one-unit increase in the number of visitors is associated with an increase in the target variable.
3. **views_trailer**: Coefficient = 0.0110 (p-value = 0.000)
   a. Interpretation: A one-unit increase in views of the trailer is associated with an increase in the target variable.
4. **dayofweek_Monday**: Coefficient = 0.0044 (p-value = 0.044)
   a. Interpretation: A one-unit increase on Monday is associated with an increase in the target variable.
5. **dayofweek_Saturday**: Coefficient = 0.0139 (p-value = 0.000)
   a. Interpretation: A one-unit increase on Saturday is associated with an increase in the target variable.
6. **dayofweek_Sunday**: Coefficient = 0.0071 (p-value = 0.002)
   a. Interpretation: A one-unit increase on Sunday is associated with an increase in the target variable.
7. **dayofweek_Thursday**: Coefficient = 0.0048 (p-value = 0.037)
   a. Interpretation: A one-unit increase on Thursday is associated with an increase in the target variable.
8. **dayofweek_Wednesday**: Coefficient = 0.0214 (p-value = 0.000)
   a. Interpretation: A one-unit increase on Wednesday is associated with an increase in the target variable.
9. **season_Spring**: Coefficient = 0.0122 (p-value = 0.000)
   a. Interpretation: A one-unit increase in the Spring season is associated with an increase in the target variable.
10. **season_Summer**: Coefficient = 0.0212 (p-value = 0.000)
    a. Interpretation: A one-unit increase in the summer season is associated with an increase in the target variable.
11. **season_Winter**: Coefficient = 0.0110 (p-value = 0.000)
    a. Interpretation: A one-unit increase in the Winter season is associated with an increase in the target variable.
12. **major_sports_event_1**: Coefficient = -0.0289 (p-value = 0.000)

a. Interpretation: A one-unit increase in major sports events is associated with a decrease in the target variable.

Key Takeaways

1. **Visitor Impact**: The analysis indicates a statistically significant positive correlation between the number of visitors and the target variable. Specifically, a one-unit increase in visitors is associated with an increase in the target, suggesting that enhancing traffic can lead to improved outcomes.
2. **Importance of Trailer Views**: The data shows that views of the trailer have a significant contribution to the target variable. This highlights the potential of effective trailer marketing strategies to boost engagement and overall performance metrics.
3. **Influence of the Day of the Week**: Certain days of the week, namely Monday, Saturday, Sunday, Thursday, and Wednesday, exhibits statistically significant positive effects on the target variable. This finding suggests that optimizing marketing efforts or promotions on these specific days could yield enhanced results.
4. **Seasonal Considerations**: The analysis reveals that different seasons (Spring, Summer, and Winter) positively influence the target variable. Thus, it is advisable to incorporate seasonality into promotional planning and resource allocation.
5. **Impact of Major Sports Events**: The presence of major sports events has a negative impact on the target variable, indicating that competing with significant sporting events may adversely affect performance. Scheduling promotions or marketing activities outside these events could prove advantageous.
6. **Low Multicollinearity**: The assessment of multicollinearity among predictors revealed low levels, suggesting that the model's estimates are stable and reliable. This bolsters confidence in the interpretation of individual predictor effects.
7. **Optimization Opportunities**: The significant predictors identified in the analysis offer valuable opportunities for optimizing marketing strategies. By focusing on increasing visitor numbers and trailer views, as well as strategically timing promotions around favorable days and seasons, overall performance can be enhanced.

Recommendations

1. **Enhance Visitor Engagement**: Develop and implement targeted strategies to attract and retain visitors, including focused advertising, engaging social media campaigns, and strategic partnerships.
2. **Improve Trailer Promotion**: Utilize innovative marketing techniques to boost trailer views, such as teaser campaigns, active social media promotions, and collaborations with influencers.
3. **Optimize Marketing by Day**: Tailor marketing efforts to align with days that show higher engagement (e.g., Monday, Saturday, and Sunday) to maximize impact.
4. **Implement Seasonal Campaigns**: Create and launch marketing campaigns that leverage the positive seasonal effects on engagement to optimize performance.
5. **Plan Promotions Strategically**: Schedule promotions and events to minimize conflicts with major sports events, thus maintaining focus on your offerings.