PM PROJECT BUSINESS REPORT

TITLE:

Analysing First-Day Viewership Drivers for OTT Content on ShowTime.

Problem Statement:

Objective:

ShowTime, an OTT service provider, aims to understand the key factors influencing the first-day viewership of the content on their platform. By identifying these driving variables, the company can implement targeted strategies to improve content consumption, optimize marketing spend, and enhance user engagement. Various factors, such as platform visits, marketing efforts, content release timing, and holiday effects, may impact first-day viewership. The goal is to develop a linear regression model to pinpoint these drivers and assist ShowTime in decision-making to boost content viewership List of content:

List of figures

1.Introduction

The primary objective of this report is to address concerns regarding the efficiency of marketing campaigns and other factors affecting first-day content viewership on ShowTime's platform. By analyzing the current data and understanding the key drivers of first-day viewership, the report aims to provide actionable insights and strategies that ShowTime can implement to enhance its viewer engagement and improve business performance

2. Methodology

2.1 Overall Approach

Exploratory Data Analysis (EDA):

- Conducting univariate and bivariate analysis to understand the distribution and relationships among the variables involved.
- Addressing specific questions related to content viewership, such as the impact of genre, day of release, season, and trailer views.

Data Preprocessing:

- Checking and managing missing values and duplicates within the dataset.
- Identifying and handling outliers to ensure robustness in the analysis.
- Feature engineering to create meaningful variables that may help in understanding first-day viewership patterns.

- Preparing the data for modelling by scaling or transforming variables as needed.

Model Building:

- Constructing a linear regression model to determine the driving factors for first-day content viewership.
- Interpreting the model coefficients to understand the significance and impact of each predictor variable.

Model Evaluation:

- Testing the assumptions of linear regression, such as linearity, homoscedasticity, normality of residuals, and multicollinearity.
- Evaluating the performance of the model using appropriate metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Insights and Recommendations:

- Deriving actionable insights from the results of the model.
- Providing recommendations to ShowTime based on the findings to help optimize content releases and improve first-day viewership.

2.2 Tools and Libraries

To execute this project efficiently, a range of tools and libraries will be utilized:

Data Manipulation:

- Pandas: Essential for data manipulation and analysis, including data cleaning, transformation, and aggregation.
- NumPy: Used for numerical operations and handling arrays.

Data Visualization:

- Matplotlib: A fundamental library for creating static, animated, and interactive visualizations.
- Seaborn: A visualization library based on Matplotlib, providing advanced features for creating attractive and informative statistical graphics.

Statistical Analysis and Machine Learning:

- SciPy: For statistical analysis and hypothesis testing.
- Stats models: For building and evaluating statistical models.
- scikit-learn: For implementing machine learning algorithms, including linear regression, and performing model evaluation.

3.Data Overview

3.1 importing the libraries :

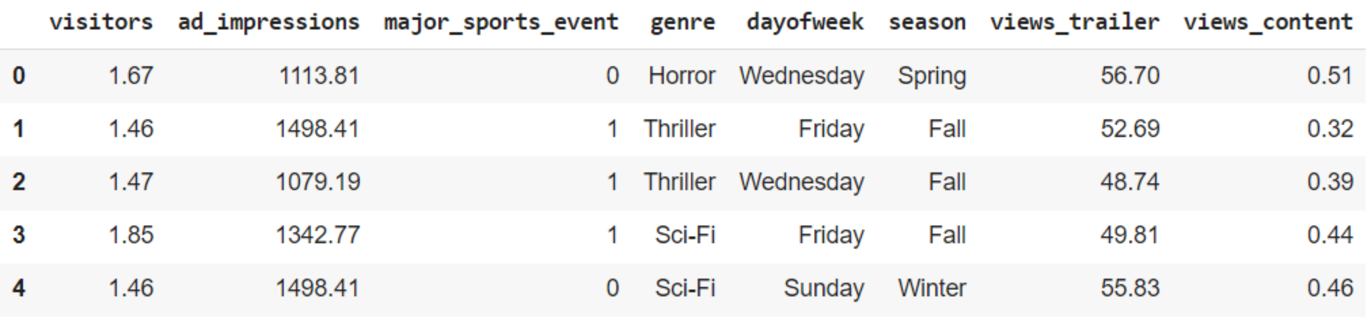
- To conduct the data analysis, we utilized several key Python libraries, including pandas, numpy, matplotlib, and seaborn. These libraries were chosen for their powerful capabilities in data manipulation, numerical computation, and visualization.

3.2 Data loading :

- The dataset was imported into google colab for analysis
- The data file was loaded using pandas functions, which allowed us to efficiently handle and manipulate the dataset

3.3 check the structure of data

- Used the function like head() and tail() in order to understand the data with the first five rows here's it.



	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content
0	1.67	1113.81	0	Horror	Wednesday	Spring	56.70	0.51
1	1.46	1498.41	1	Thriller	Friday	Fall	52.69	0.32
2	1.47	1079.19	1	Thriller	Wednesday	Fall	48.74	0.39
3	1.85	1342.77	1	Sci-Fi	Friday	Fall	49.81	0.44
4	1.46	1498.41	0	Sci-Fi	Sunday	Winter	55.83	0.46

Here is the data description(columns) :

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

There are 1000 rows and 8 columns

3.4 check datatype

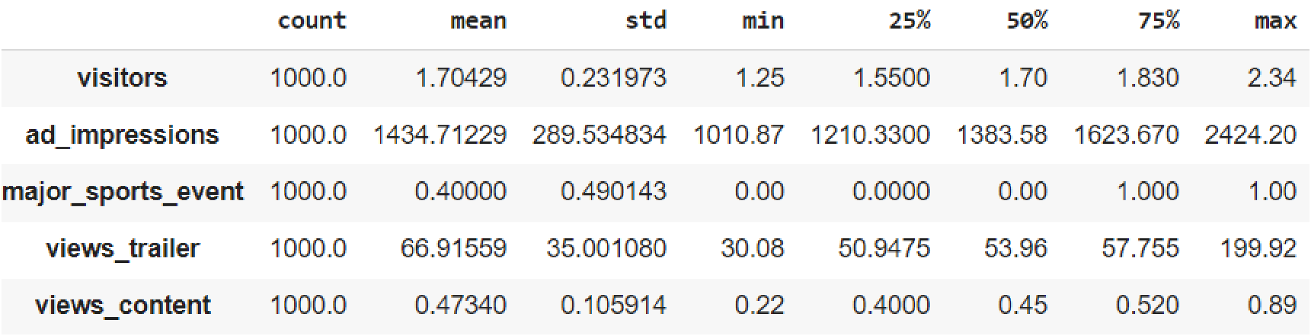
- **Numerical Variables**: visitors, ad_impressions, views_trailer, views_content
- **Categorical Variables**: major_sports_event, genre, dayofweek, season
- There are total 8 columns in which 3 are categorical that is object and the remaining 5 are numerical

- When proceeding with the analysis, these data types guide how we treat each column. For instance:

- Numerical columns may undergo statistical analysis and visualization through histograms or scatter plots.
- Categorical columns might be analysed using bar charts, and in some cases, converted to numerical format using encoding techniques for machine learning models.

3.5 Checking the statistics of data:

We get the statistical summary of the data by using the describe() , this provides a very insightful statistical summary it tells the average, count, maximum, minimum, 25th percentage , 50th percentage and 75th percentage of the data which makes it easy to draw the conclusions



	count	mean	std	min	25%	50%	75%	max
visitors	1000.0	1.70429	0.231973	1.25	1.5500	1.70	1.830	2.34
ad_impressions	1000.0	1434.71229	289.534834	1010.87	1210.3300	1383.58	1623.670	2424.20
major_sports_event	1000.0	0.40000	0.490143	0.00	0.0000	0.00	1.000	1.00
views_trailer	1000.0	66.91559	35.001080	30.08	50.9475	53.96	57.755	199.92
views_content	1000.0	0.47340	0.105914	0.22	0.4000	0.45	0.520	0.89

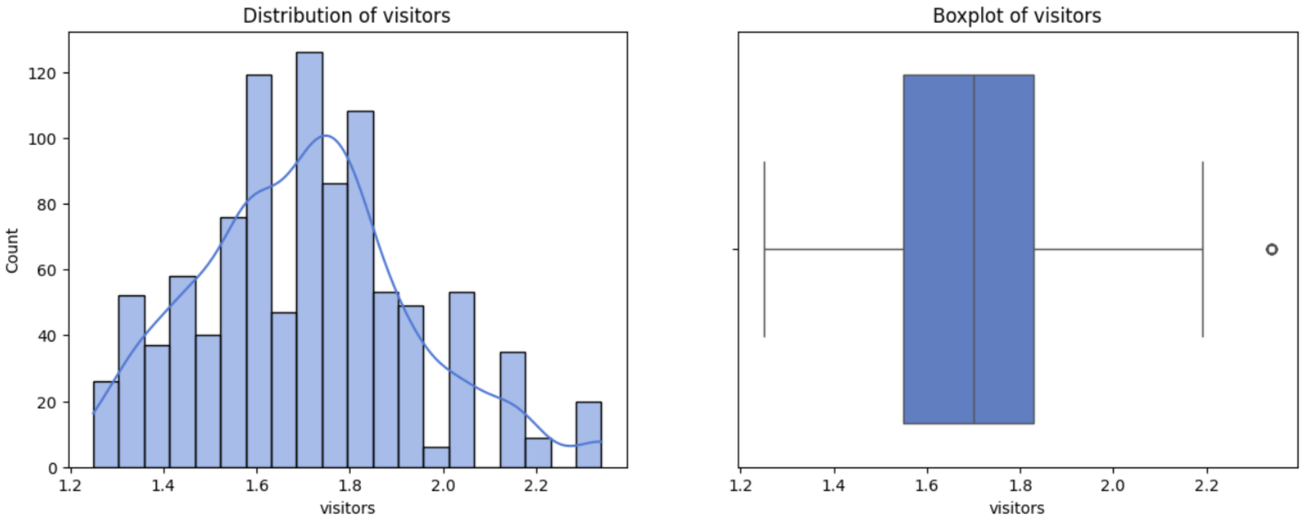
4. Exploratory Data Analysis (EDA)

We perform Analysis byusing Univariate, Bivariate and multivariate analysis

4.1 Univariate Analysis:

- **numerical columns** :

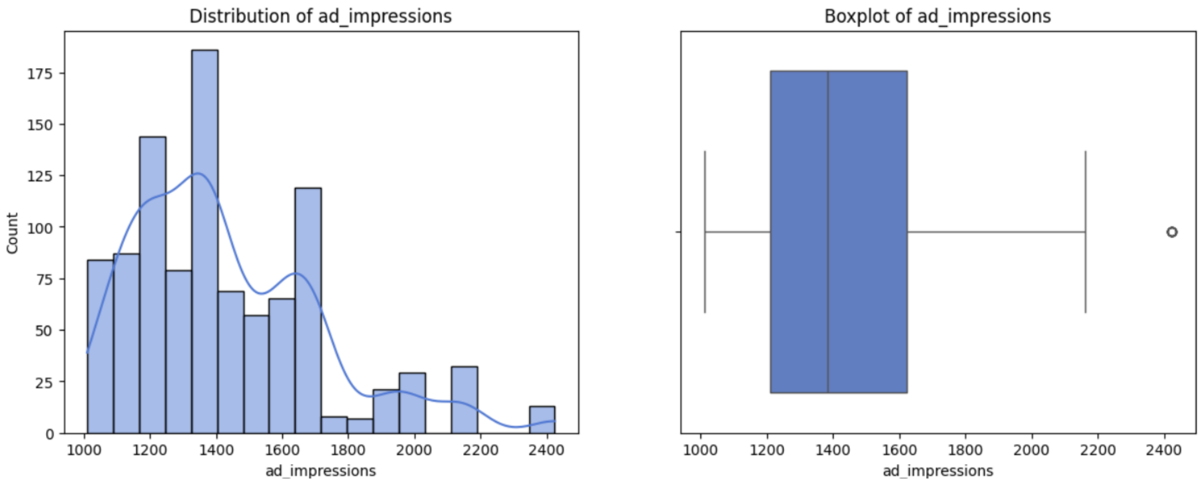
4.1.1.Observing the distribution of visitors columns :



- Distribution Shape: The histogram shows the distribution of visitors, and it looks approximately bell-shaped, indicating a normal distribution with some right skewness.

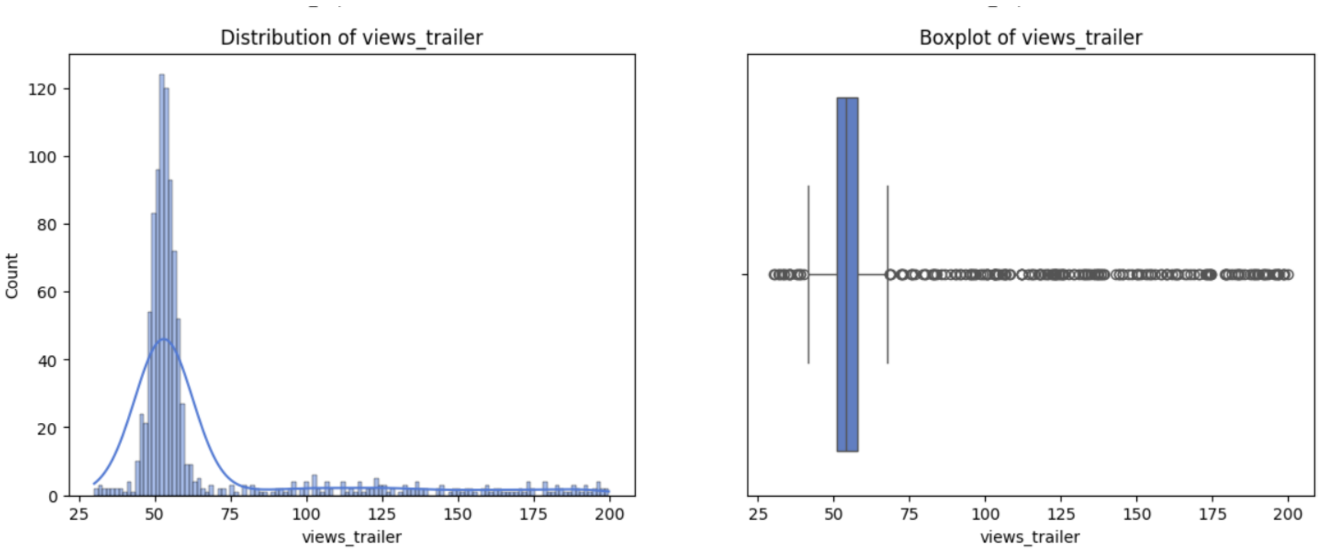
- Peak: The highest frequency occurs between 1.6 and 1.8 visitors, where the count is over 120. This suggests that most visitors fall in this range.
- Skewness: There is a slight right skew (longer tail on the right side), as seen from the tapering of the KDE curve toward the right (i.e., values above 2.0).
- Outliers: There seem to be some values that are quite low in frequency toward the higher end (around 2.2 visitors).

4.1.2.Observing the distribution ads impression columns :



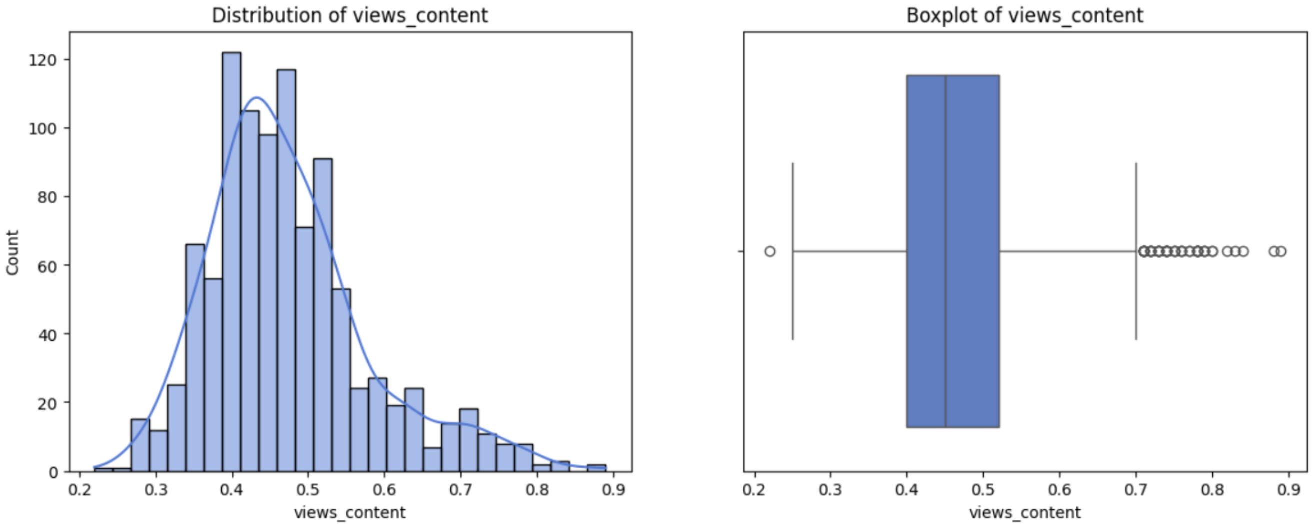
- Distribution Shape: The ad_impressions distribution is right-skewed, with a higher concentration between 1200 and 1400 impressions.
- Peak Count: Most ad_impressions fall around 1400, with the highest count close to 175.
- Outliers: There's a noticeable outlier around 2400 ad impressions, which is much higher than the rest of the data.
- Boxplot Summary: The interquartile range (IQR) spans from 1200 to 1600, with the median near 1400. The outlier beyond 2400 confirms the right skew.
- spread: Most data lies between 1000 and 1800 ad impressions.

4.1.3.Observing the distribution views trailer columns :



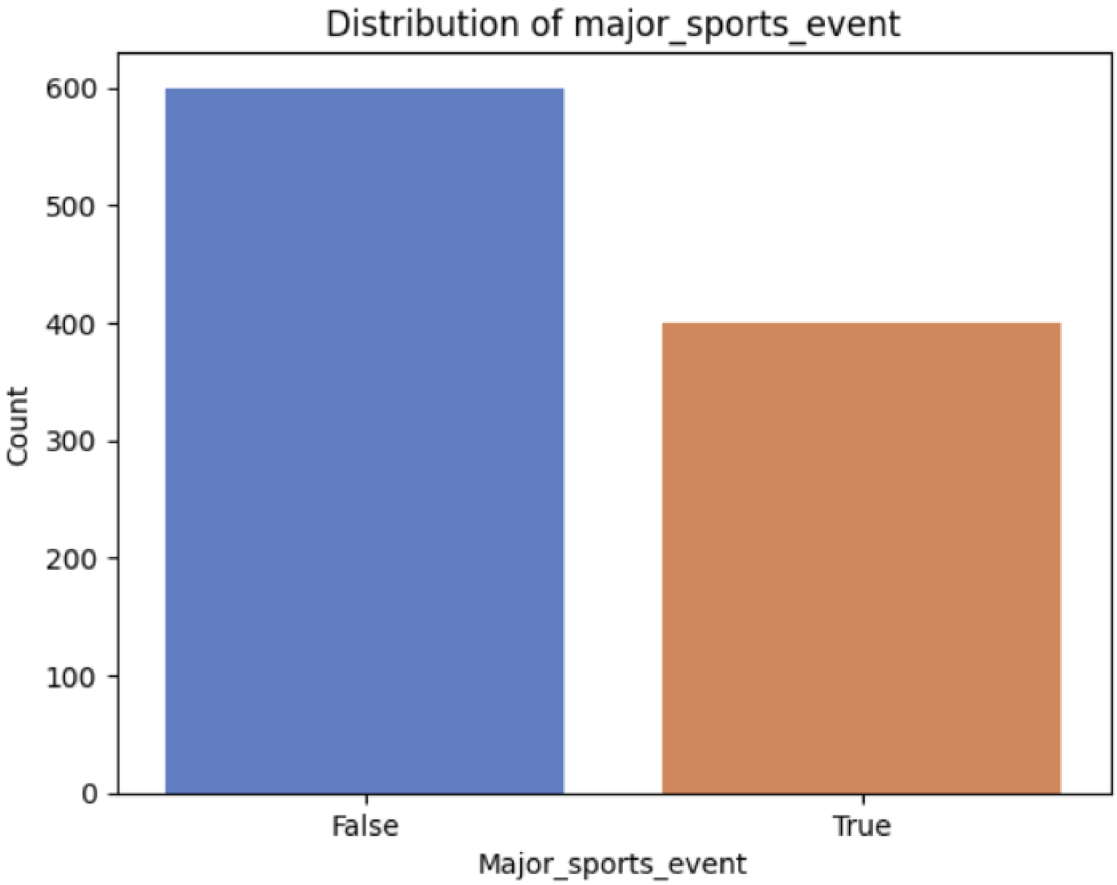
- Distribution Shape: The views_trailer distribution is highly right-skewed, with most data concentrated around 50 views.
- Peak Count: The majority of the trailer views are around 50, with a sharp peak in the histogram.
- Outliers: There are numerous outliers, especially after 75 views, as seen from the boxplot. The tail extends up to 200 views.
- Boxplot Summary: The boxplot shows a narrow range for the middle 50% of the data, with the IQR between 50 and 60 views, and a long tail of outliers beyond 75.
- Spread: Despite the high outlier count, the majority of views are clustered around 50 to 60 views, indicating a highly concentrated data distribution.

4.1.4.Observing the distribution views_content columns :



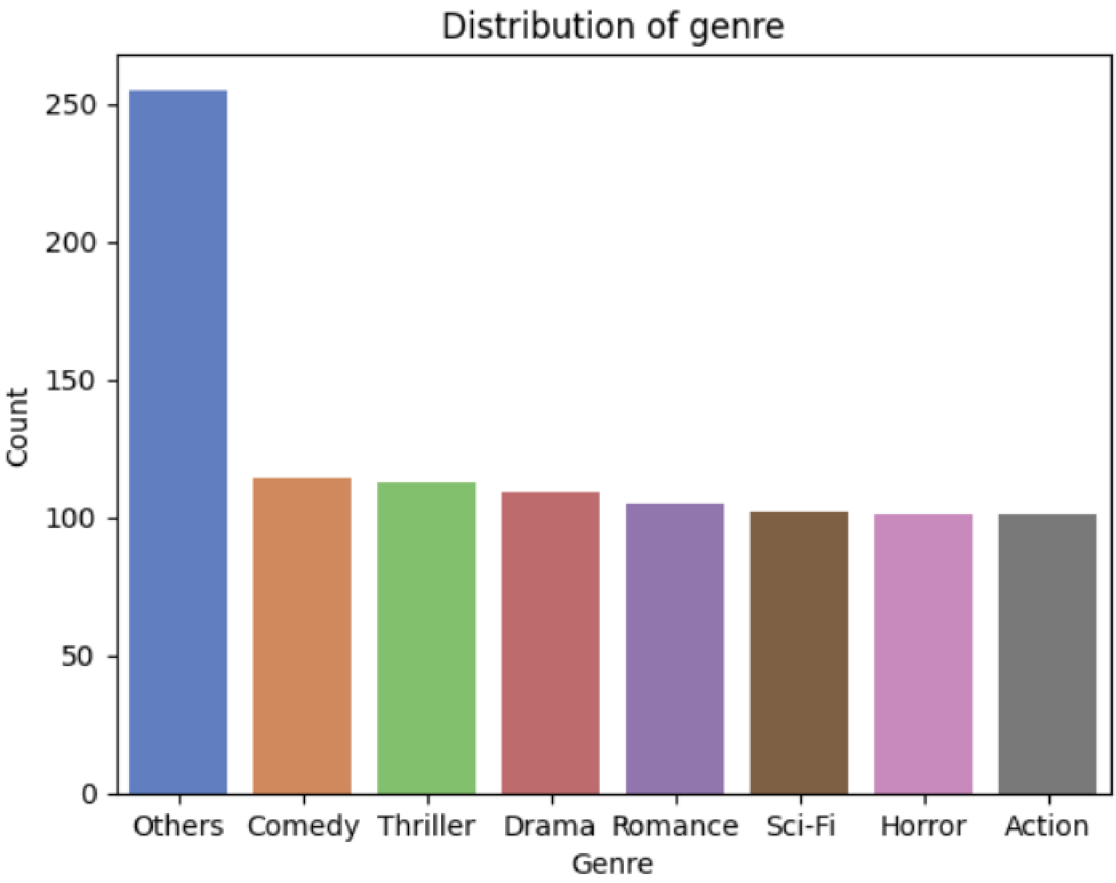
- **Shape:** The histogram shows a **right-skewed** distribution, indicating that there are a few observations with very high values (i.e., a large number of views_content) that pull the tail of the distribution to the right.
- **Central Tendency:** The majority of the data points cluster between 0.4 and 0.6, suggesting that this is the most common range for views_content.
- **Spread:** The distribution is relatively spread out, with a noticeable range between the minimum and maximum values.
- **Density Curve:** The superimposed density curve provides a smoother representation of the distribution, confirming the right-skewed shape and the central tendency
- **Outliers:** The individual points outside the whiskers (the lines extending from the box) are potential outliers, suggesting that they are significantly different from the majority of the data.

4.1.5 Distribution of major sports events



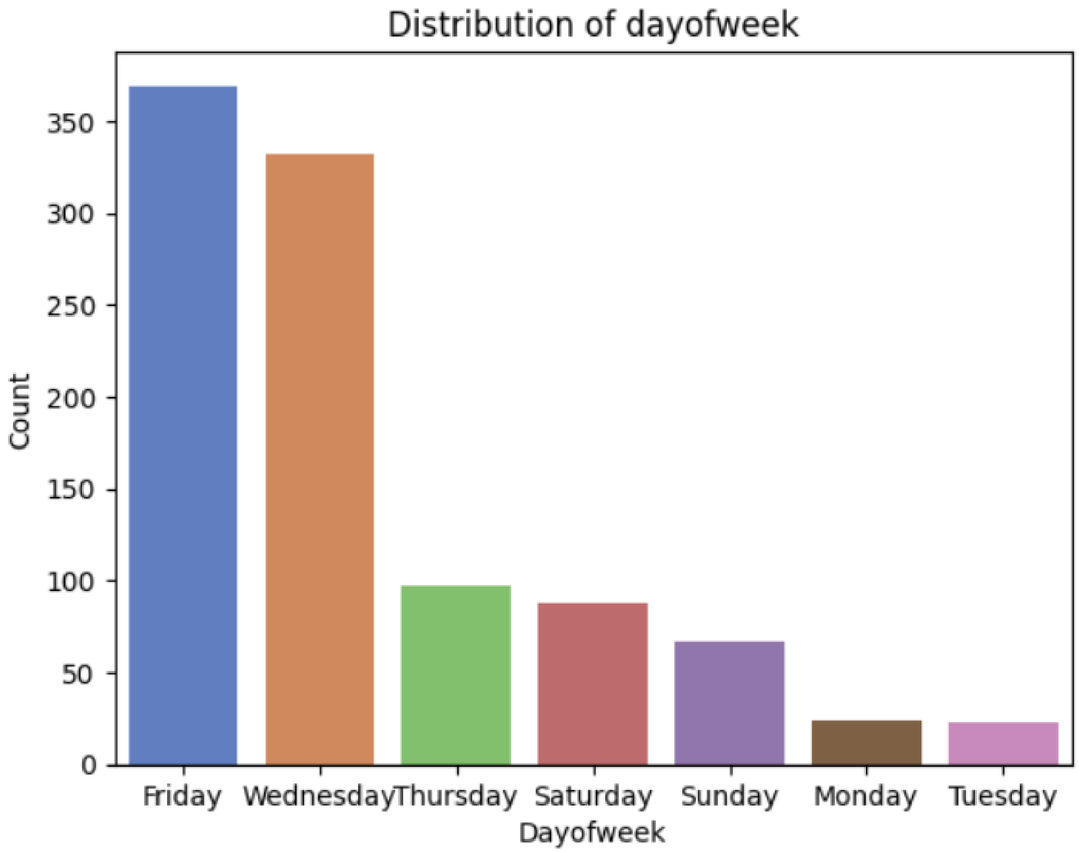
The graph illustrates the distribution of the major_sports_event variable. A majority of the observations (approximately 600) fall into the "False" category, indicating that most of the data points do not correspond to major sports events. Conversely, a smaller number of observations (around 400) belong to the "True" category, suggesting that a minority of the data points are associated with major sports events.

4.1.6 Distribution of Genre



- The dominance of the "Others" category in the distribution of genre.
- The relative frequencies of the other genres and any patterns or trends that may be observed.
- The potential reasons or implications of the dominance of the "Others" category.
- How the findings from this graph relate to other variables or aspects of your study.

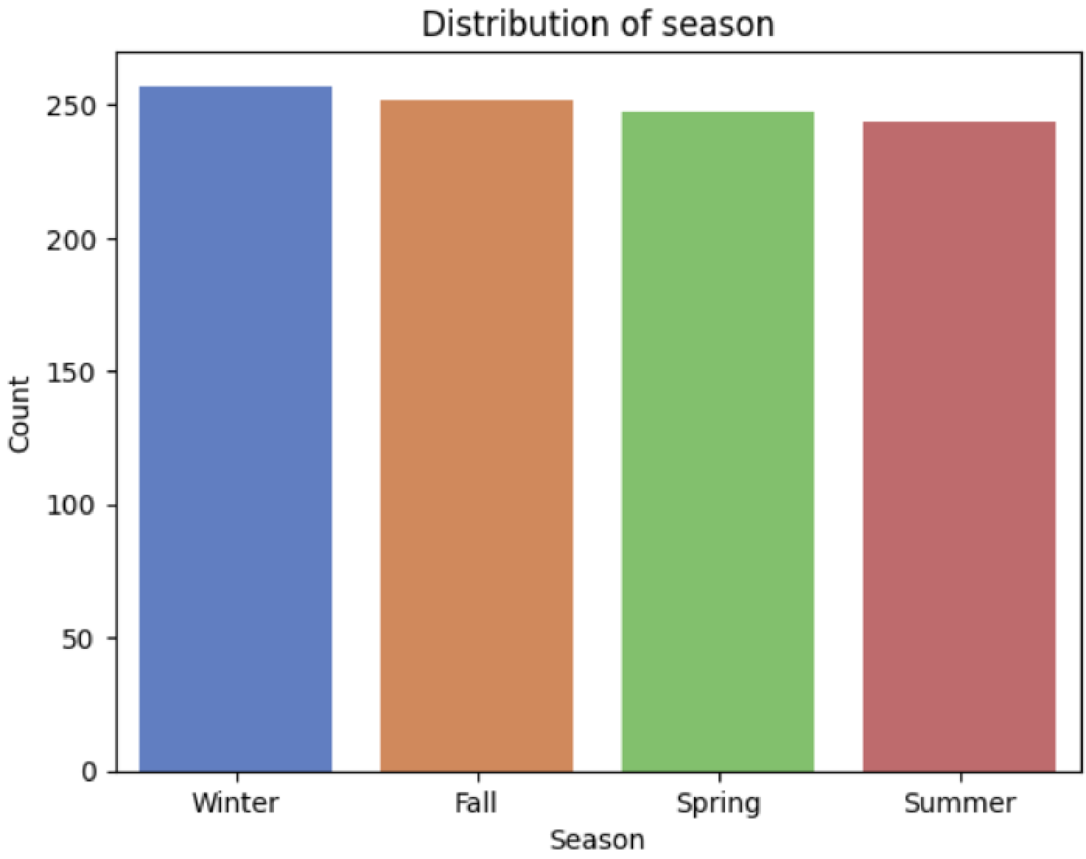
4.1.7Distribution of day of week



The graph illustrates the distribution of the dayofweek variable. The "Friday" category dominates the distribution, indicating that a substantial portion of the data points correspond to Fridays. The other days of the week have lower frequencies, with "Wednesday/Thursday" and "Saturday" having the next highest counts. The remaining days, "Sunday," "Monday," and "Tuesday," have relatively lower frequencies.

- The dominance of the "Friday" category in the distribution of dayofweek.
- The relative frequencies of the other days of the week and any patterns or trends that may be observed.
- The potential reasons or implications of the dominance of the "Friday" category.
- How the findings from this graph relate to other variables or aspects of your study.

4.1.8 Distribution of Seasons:

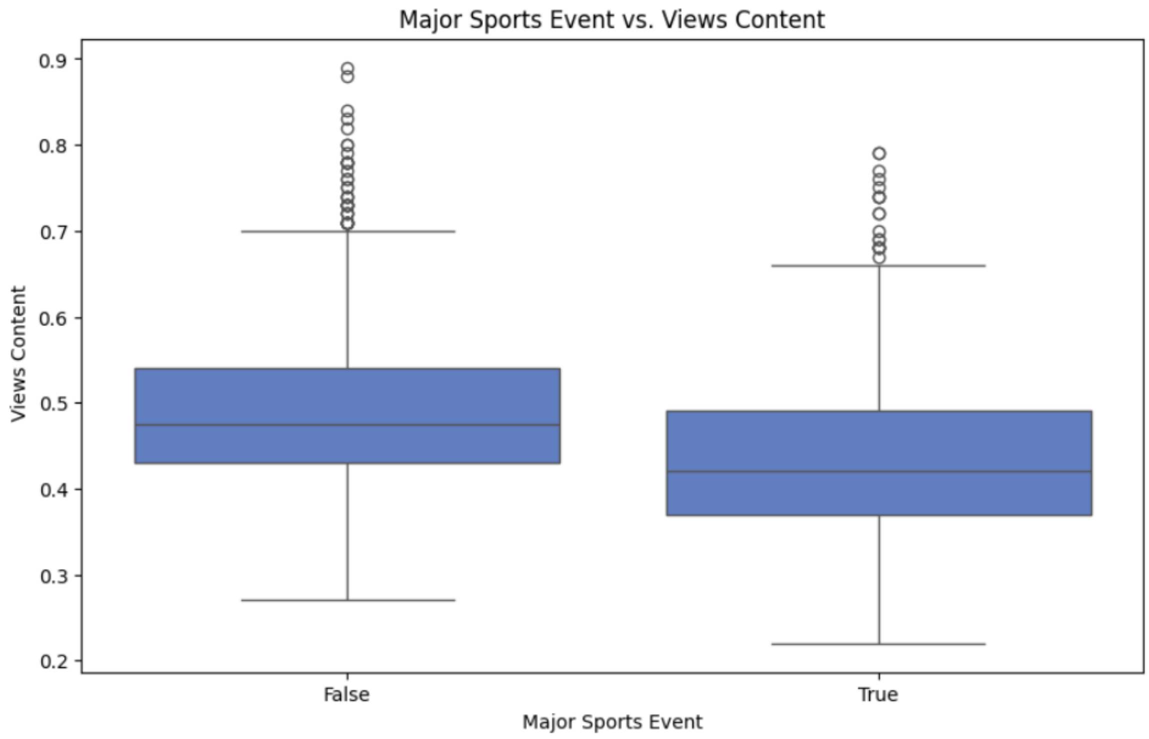


The graph illustrates the distribution of the season variable. The data appears to be fairly evenly distributed across the four seasons, indicating that there is no significant preference or bias towards any particular season in the dataset.

- The relatively even distribution of data across the four seasons.
- Any potential implications or insights that can be drawn from this observation.
- How the findings from this graph relate to other variables or aspects of your study.

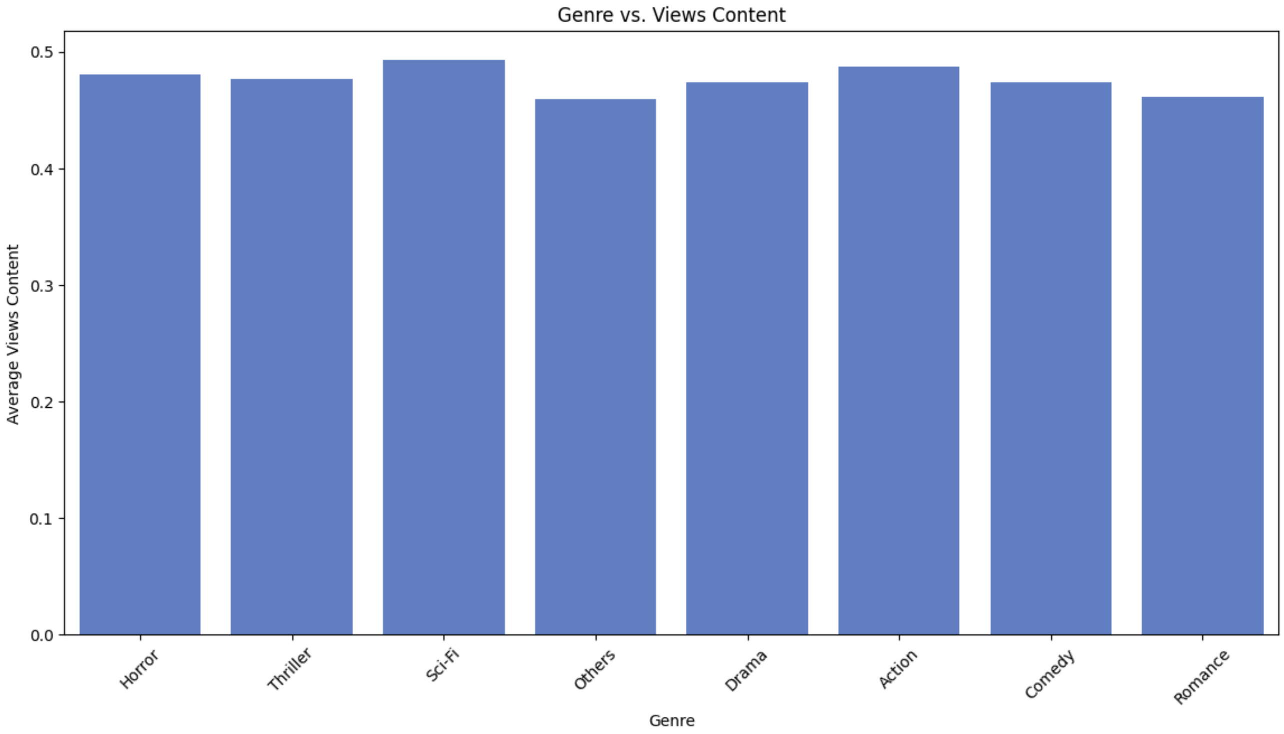
4.2 Bivariate Analysis:

4.2.1 Major sports events vs views content



Outliers: There are a few outliers present in both categories, indicated by the individual points outside the whiskers. These outliers suggest that there are some observations with unusually high or low Views Content values, even when considering the category of Major Sports Event.

4.2.2 Genre vs views content



Dominance of Certain Genres:

- Sci-Fi: This genre tops the list with an impressive 4.9 million views, indicating a significant viewer preference for science fiction content.
- Action: Alongside Sci-Fi, the Action genre also garners high viewership, suggesting that high adrenaline and exciting content are popular among viewers.

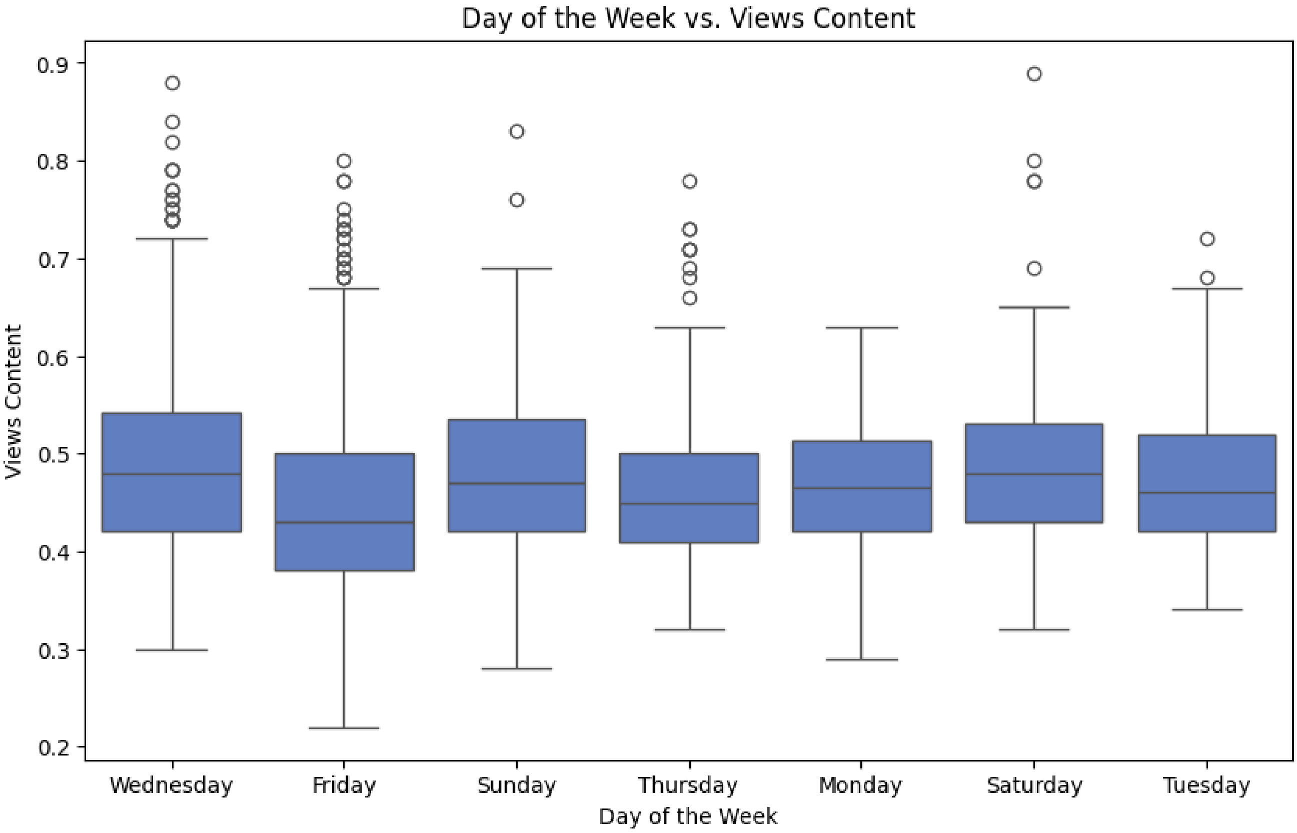
Moderately Popular Genres:

- Horror, Thriller, and Comedy: These genres follow Sci-Fi and Action in terms of viewership. This trend highlights a balanced audience interest in these categories. The moderate popularity suggests that while they are not as dominant as Sci-Fi and Action, they still attract a consistent and substantial number of viewers.

Least Viewed Genres:

- Others: This category has the least views among all genres. This could indicate that content classified under 'Others' might not be striking enough, or there is less content available in these genres, making them less appealing or less promoted.

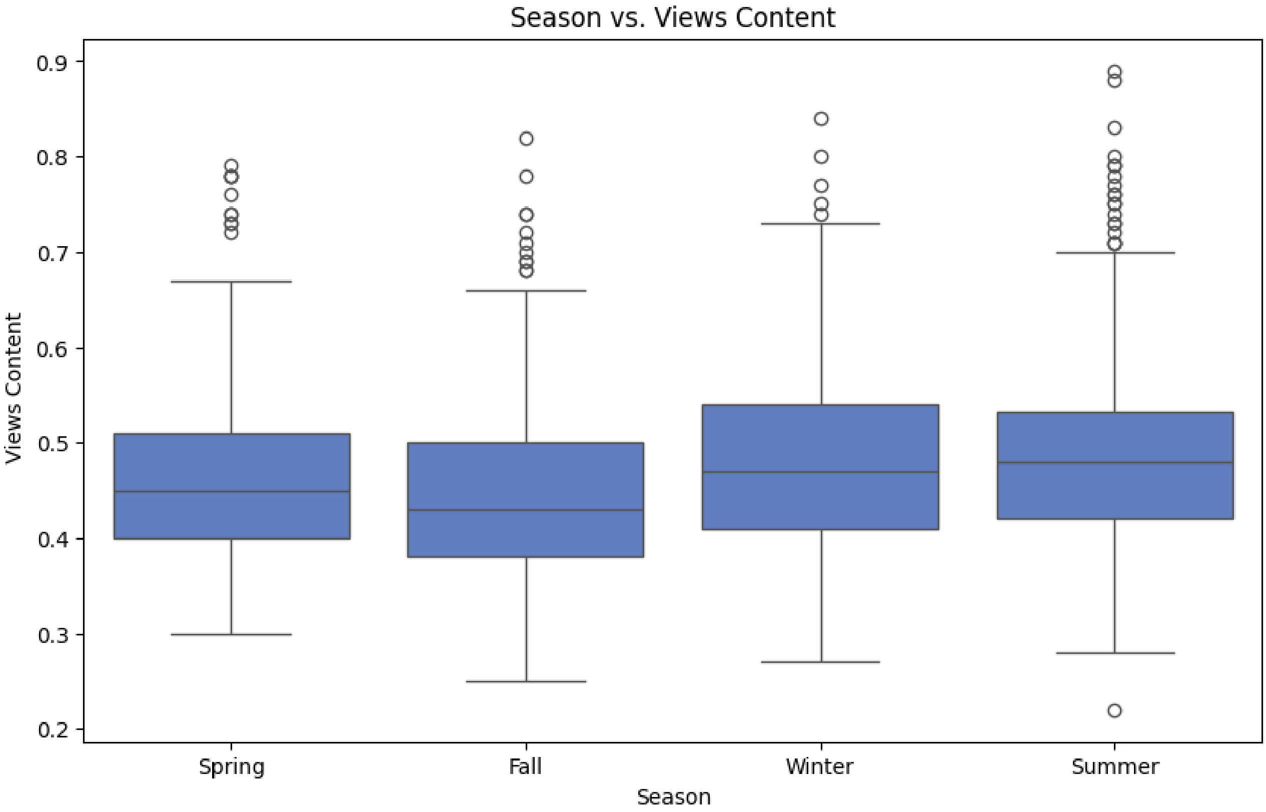
4.2.3 Day of week vs views content



- Saturday stands out as the day with the highest viewership. This can be attributed to more viewers being available to consume content during their weekend leisure time.
- Following Saturday, Wednesday, Tuesday, and Sunday also show higher viewership figures, which suggest that these mid-week and weekend days are popular for content consumption

- This may be due to a mix of mid-week breaks and end-of-week relaxation periods. On the other hand, Friday sees the lowest viewership. This could be because viewers might be occupied with social events or other activities as they kick off their weekends.
- In summary, while the viewership does not vary significantly across the different days of the week, Saturday emerges as the day with the highest engagement. This insight can help ShowTime better schedule their content releases to maximize viewership.

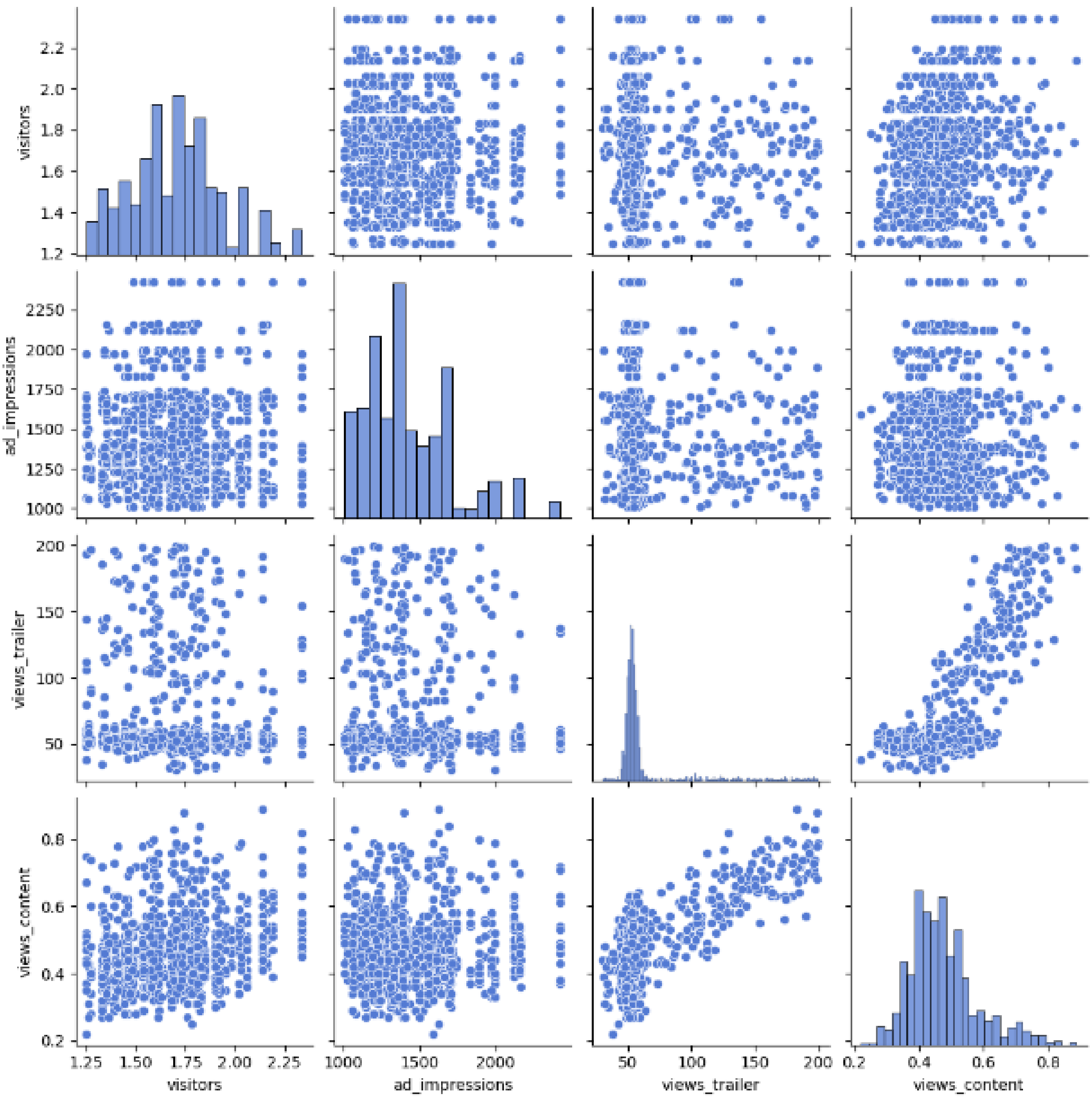
4.2.3 Seasons vs views content



- Summer emerges as the season with the highest viewership, averaging 4.8 million views. This spike in viewership can be attributed to factors such as increased leisure time during vacations, leading to higher content consumption. Winter follows with substantial viewership, indicating that audience engagement remains high as people likely spend more time indoors consuming content. Spring also shows considerable viewer activity, although slightly less than winter.
- The season with the least viewership is Fall, averaging 4.5 million views. This lower viewership might be influenced by various factors, including back-to-school schedules and the resumption of regular activities post-summer.
- In summary, seasonality plays a role in viewership patterns, with summer showing the highest engagement. ShowTime can leverage these insights to plan content releases and marketing strategies to align with peak viewership season

4.2.4 Visualizing Relationships Between Numerical Variables

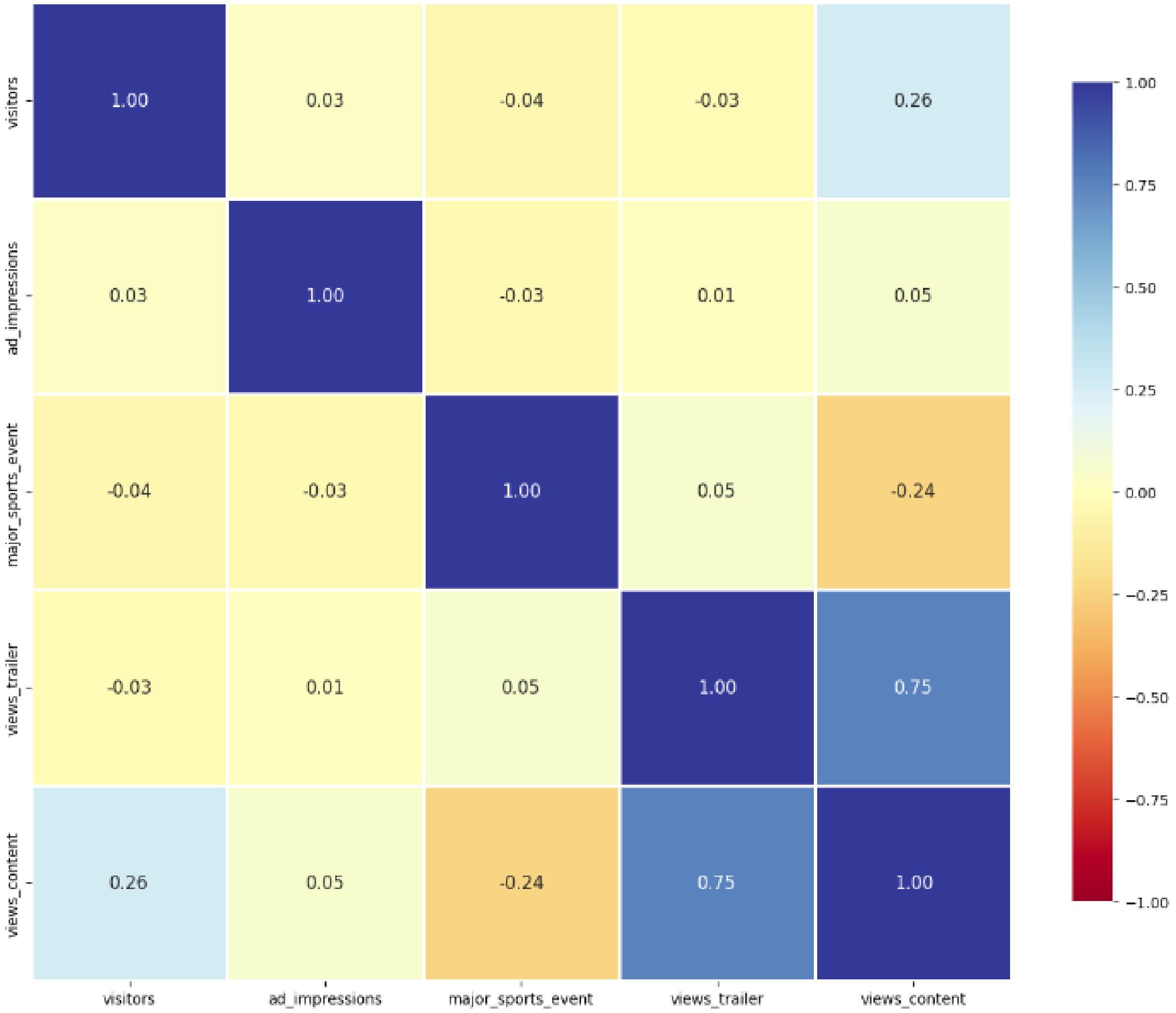
The pairplot is a useful visualization tool that displays the relationships between multiple variables in a dataset. Each subplot in the matrix represents the relationship between two variables, with the diagonal subplots showing the distribution of each variable



- **Visitors vs. Ad Impressions:** There seems to be a weak positive correlation between these two variables. As the number of visitors increases, the number of ad impressions tends to increase slightly, but the relationship is not very strong.
- **Visitors vs. Views Trailer:** There appears to be a moderate positive correlation between these variables. As the number of visitors increases, the number of views for trailers tends to increase.

- **Visitors vs. Views Content:** There seems to be a strong positive correlation between these variables. As the number of visitors increases, the number of views for content tends to increase significantly.
- **Ad Impressions vs. Views Trailer:** There appears to be a weak positive correlation between these variables. As the number of ad impressions increases, the number of views for trailers tends to increase slightly, but the relationship is not very strong.
- **Ad Impressions vs. Views Content:** There seems to be a moderate positive correlation between these variables. As the number of ad impressions increases, the number of views for content tends to increase.
- **Views Trailer vs. Views Content:** There appears to be a strong positive correlation between these variables. As the number of views for trailers increases, the number of views for content tends to increase significantly.

4.3 Multivariate analysis



Visitors and Views Content (0.26):

- There is a weak positive correlation (correlation coefficient: 0.26) between the number of visitors and views content.
- Implication: This suggests that as the number of visitors increases, views content also tends to increase. However, the relationship is not very strong, indicating that other factors may also be influencing views content.

Ad Impressions and Other Variables:

- Ad Impressions have a very weak correlation with all other variables:
- Views Trailer (0.01) and Views Content (0.05): Both show weak correlations, implying almost no linear relationship with ad impressions.
- Major Sports Event (-0.03): Also has a very weak negative correlation, suggesting that the number of ad impressions does not change significantly based on whether there is a major sports event.

Major Sports Event and Views Content (-0.24):

- There is a moderate negative correlation (correlation coefficient: -0.24) between major sports events and views content.
- Implication: This suggests that during major sports events, the number of views content tends to decrease. Major sports events might be diverting viewers' attention away from OTT content.

Views Trailer and Views Content (0.75):

- There is a strong positive correlation (correlation coefficient: 0.75) between views of the trailer and views of the content.
- Implication: This indicates that people who view the trailer are highly likely to view the content as well. This strong relationship highlights the importance of trailer views as a predictor for content views.

Visitors and Other Variables:

- Ad Impressions (0.03): Shows a very weak positive correlation, indicating almost no relationship.
- Views Trailer (-0.03): Also shows a very weak negative correlation, which suggests that the number of visitors does not strongly relate to trailer views.

Overall Conclusion of Exploratory Data Analysis (EDA)

The exploratory data analysis conducted offers a comprehensive view of various factors influencing content viewership on ShowTime OTT platform. The insights gathered from univariate, bivariate, and multivariate analyses shed light on patterns in user behaviour, content consumption, and external factors like sports events and seasons. Below is a summary of the key findings from the analysis:

1. **Univariate Analysis:**

  - **Visitors:** The distribution is approximately normal with some right skewness. The majority of visitors fall in the 1.6 to 1.8 range, with fewer extreme values, indicating a relatively stable user base.
  - **Ad Impressions:** The ad impressions show a right-skewed distribution, with most impressions ranging between 1200 and 1400. There are outliers indicating occasional high ad activity.
  - **Views (Trailer & Content):** Both metrics exhibit a highly skewed distribution. While most trailer views cluster around 50 views, content views tend to center between 0.4 and 0.6 million views. This suggests concentrated viewing activity around specific content.
  - **Major Sports Events:** A majority of the data points are not associated with major sports events, with fewer observations for those times when they do occur.
  - **Genres & Day of Week:** "Sci-Fi" and "Action" genres dominate viewership, while "Others" receive minimal engagement. Fridays and Saturdays tend to see the highest content consumption, pointing to potential audience preferences over the weekend.
  - **Seasons:** Viewership is fairly balanced across seasons, though summer shows a slight uptick in content consumption, potentially due to increased leisure time.

2. **Bivariate Analysis:**
    - **Major Sports Events vs Views:** During major sports events, there is a noticeable decline in content views, suggesting that sports events may divert user attention away from ShowTime's platform.
    - **Genre vs Views:** Genres such as Sci-Fi and Action see significantly higher viewership than others. Less popular genres (like Horror and Thriller) still attract a consistent number of viewers but to a lesser extent.
    - **Day of the Week vs Views:** Saturday emerges as the day with the highest viewership, likely due to increased weekend leisure time. Other mid-week days (Wednesday and Tuesday) also experience higher content consumption compared to the rest of the week.
    - **Season vs Views:** Summer shows the highest engagement, with significant viewership spikes, while Fall records the lowest viewership, likely due to seasonal factors like back-to-school schedules.
3. **Multivariate Analysis:**
    - **Visitors and Views Content (Correlation 0.26):** There is a weak positive correlation, indicating that while an increase in visitors leads to more content views, other variables may also significantly influence views.
    - **Ad Impressions and Other Variables:** Ad impressions show very weak correlations with views, suggesting that impressions alone do not drive user engagement with content.

  - **Major Sports Events and Views Content (-0.24):** There is a moderate negative correlation, meaning viewership decreases during major sports events. This indicates a diversion of attention towards sports-related content or activities.
  - **Views Trailer and Views Content (0.75):** There is a strong positive correlation, indicating that trailer views are a strong predictor of content consumption. This highlights the importance of trailers in attracting users to watch content.

Strategic Insights for ShowTime:

- Leverage High-Engagement Days and Seasons: The data shows that Saturdays, as well as the summer season, experience the highest engagement. ShowTime should consider aligning content releases and marketing campaigns around these peak engagement times to maximize viewership.
- Focus on Popular Genres: Sci-Fi and Action dominate viewership, suggesting ShowTime could invest in producing more content within these genres. Simultaneously, identifying niche marketing strategies for underperforming genres like "Others" may help increase engagement in those categories.
- Address the Impact of Major Sports Events: During major sports events, content views decline. To combat this, ShowTime could explore cross-promotional opportunities with sports events or create content that complements the interests of sports viewers.
- Utilize Trailer Views as a Predictor of Success: The strong correlation between trailer views and content views highlights the importance of promoting trailers. Investing in engaging trailers and ensuring they reach the audience may lead to higher overall content consumption.

5.Key question Answers

5.1What does the distribution of content views look like?

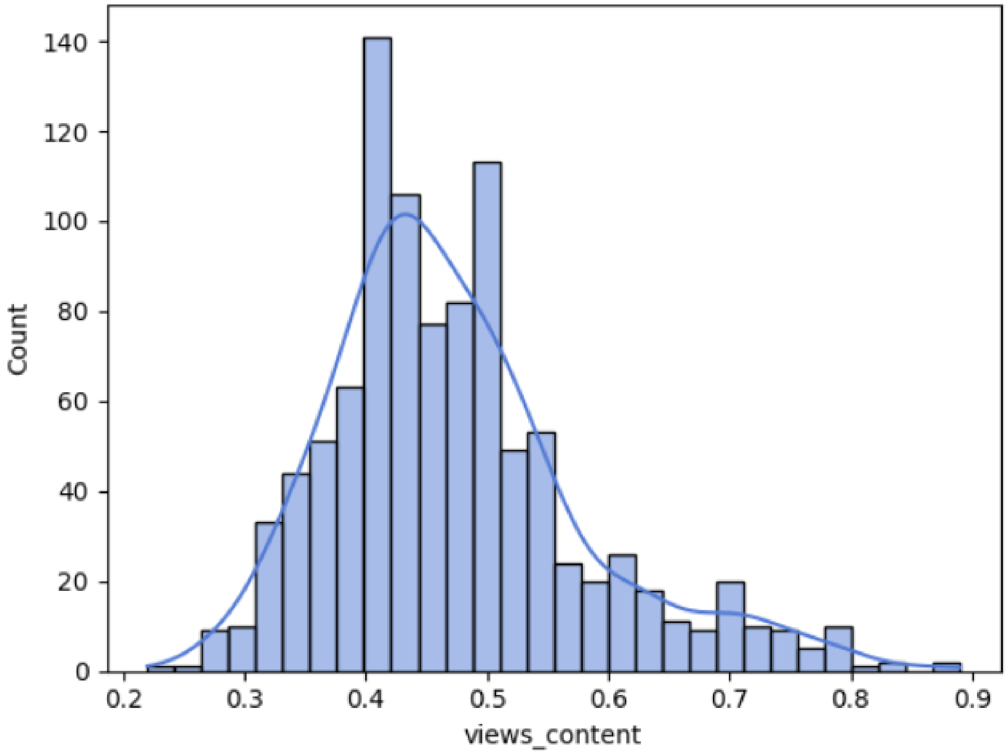
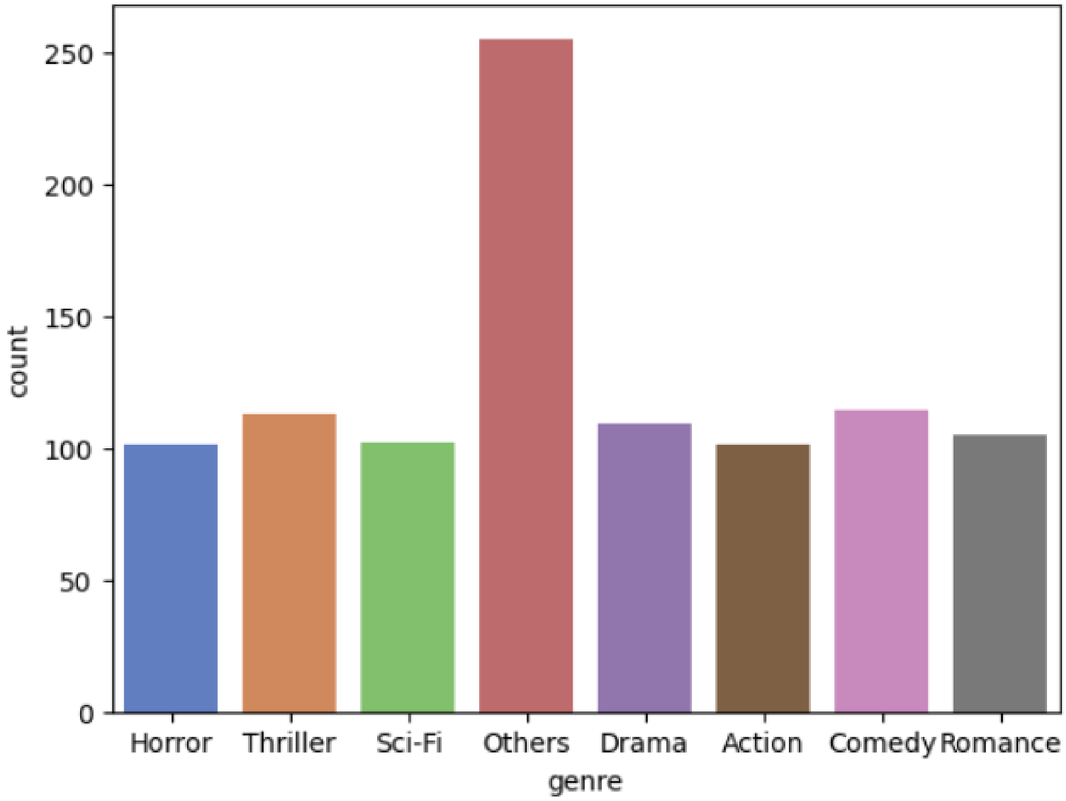


Fig 5.1 views content distribution

- It is normally distributed
- The views content data is right skewed Right-skewed with many outliers. This means that most content receives fewer views, but there are a few that get significantly higher views. This is common in streaming services where a few hit shows/movies drive most of the traffic
- Most content receives fewer views, clustered around the lower end of the viewership range. There are a few content pieces that receive significantly higher views, which extend the tail of the distribution to the right.
- the highest count is 130 which is of 0.4million views

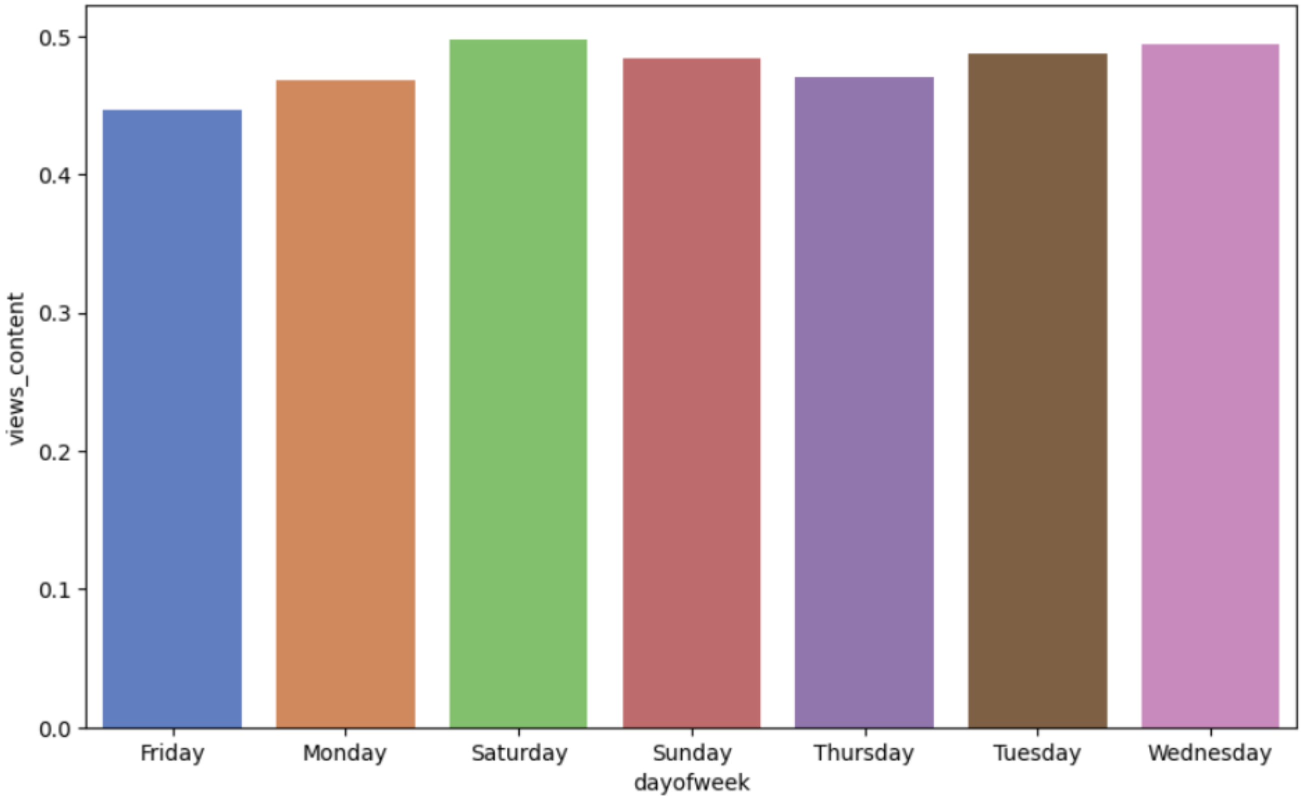
5.2 What does the distribution of genres look like?



5.2 Genre distribution

The analysis of the distribution of genres on the ShowTime platform revealed some interesting trends. The 'Others' category stands out as the most watched genre, potentially indicating a collection of niche or mixed-genre content that appeals to a broad audience. Beyond this, Comedy and Thriller genres are equally popular, suggesting that the audience enjoys a mix of laughter and suspense in their viewing experience. Based on these insights, content strategy can be tailored to focus on producing or acquiring more content in these popular genres, ensuring broader audience engagement and satisfaction

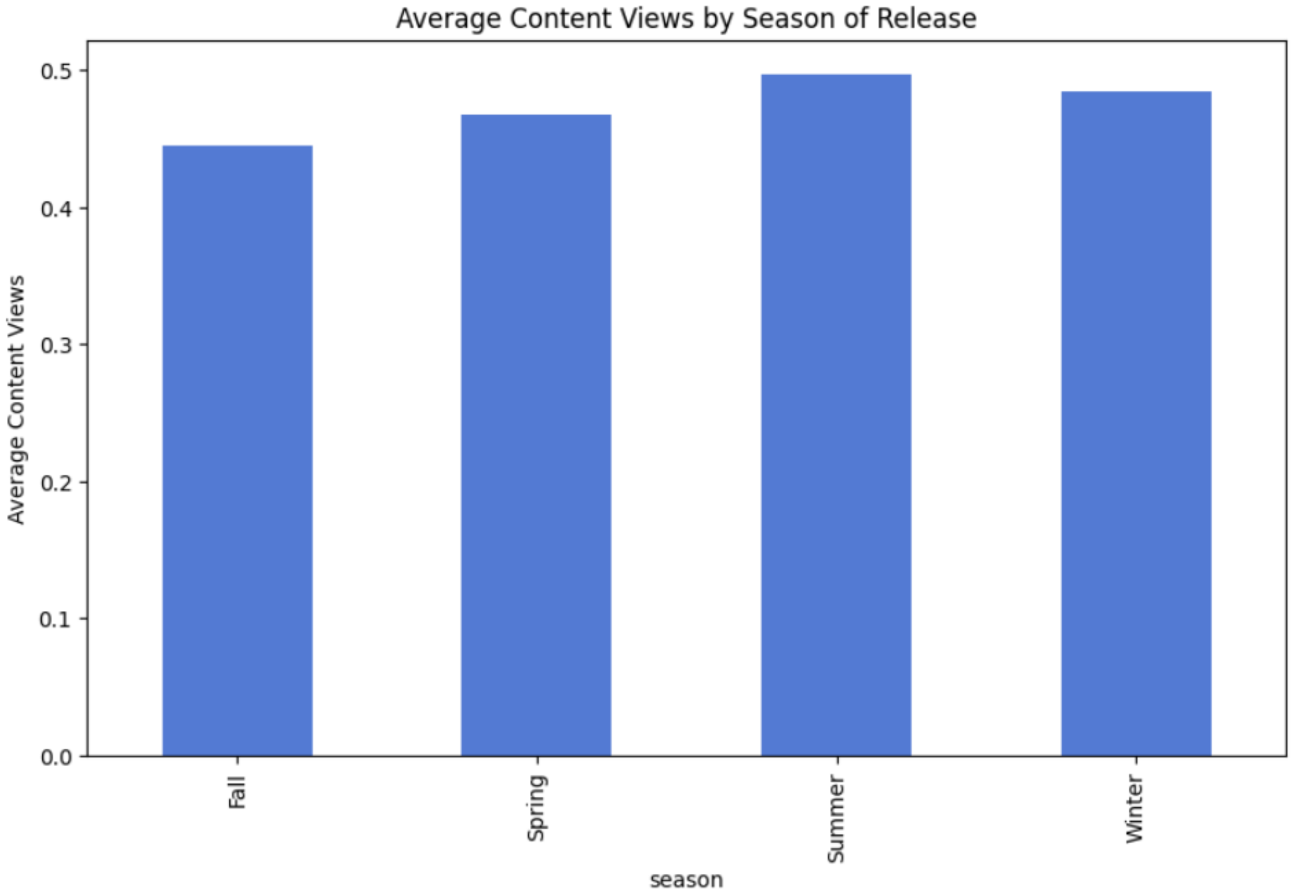
5.3The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?



5.3 Viewership vs day of week

- Saturday stands out as the day with the highest viewership. This can be attributed to more viewers being available to consume content during their weekend leisure time.
- Following Saturday, Wednesday, Tuesday, and Sunday also show higher viewership figures, which suggest that these mid-week and weekend days are popular for content consumption
- This may be due to a mix of mid-week breaks and end-of-week relaxation periods. On the other hand, Friday sees the lowest viewership. This could be because viewers might be occupied with social events or other activities as they kick off their weekends.
- In summary, while the viewership does not vary significantly across the different days of the week, Saturday emerges as the day with the highest engagement. This insight can help ShowTime better schedule their content releases to maximize viewership.

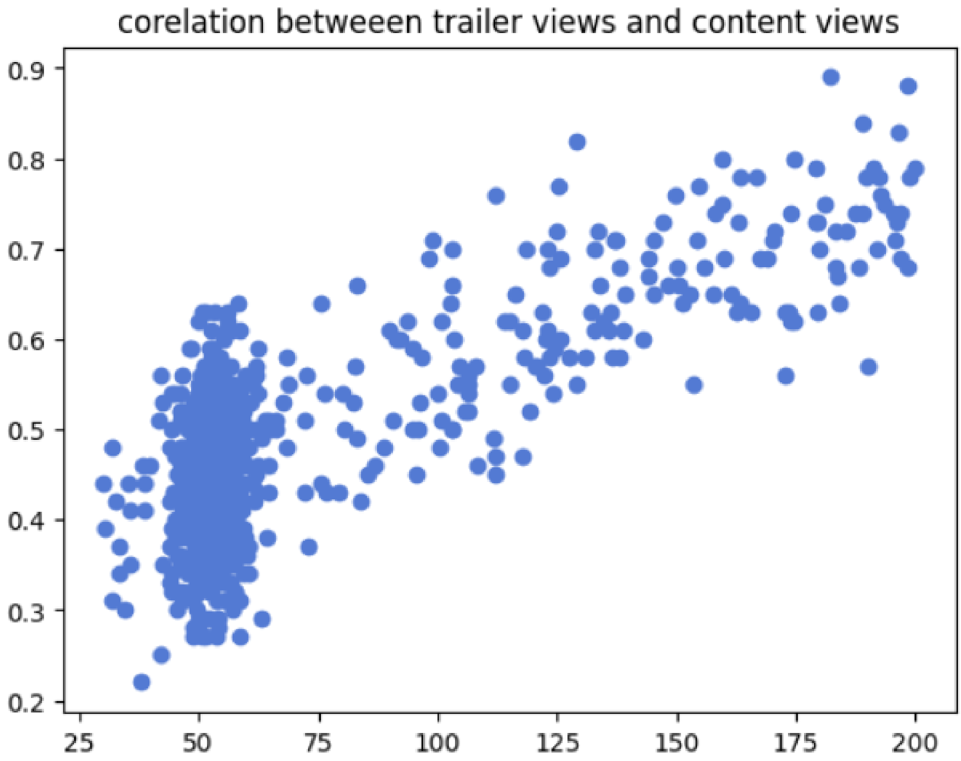
5.4How does the viewership vary with the season of release?



5.4 Season vs content views

- Summer emerges as the season with the highest viewership, averaging 4.8 million views. This spike in viewership can be attributed to factors such as increased leisure time during vacations, leading to higher content consumption. Winter follows with substantial viewership, indicating that audience engagement remains high as people likely spend more time indoors consuming content. Spring also shows considerable viewer activity, although slightly less than winter.
- The season with the least viewership is Fall, averaging 4.5 million views. This lower viewership might be influenced by various factors, including back-to-school schedules and the resumption of regular activities post-summer.
- In summary, seasonality plays a role in viewership patterns, with summer showing the highest engagement. ShowTime can leverage these insights to plan content releases and marketing strategies to align with peak viewership season

5.5 What is the correlation between trailer views and content views?



5.5 Corelation between trailer view and content view

- The correlation coefficient between trailer views and content views is 0.7. This indicates a strong positive correlation
- A correlation of 0.7 suggests that there is a significant positive relationship between the number of trailer views and the number of first-day content views. This means that content that receives a higher number of trailer views is likely to also receive a higher number of first-day views.
- A scatter plot of trailer views against content views visually supports this finding, showing an upward trend where higher trailer views generally correspond to higher content views

6. Data processing

Data preprocessing is a crucial step in the data analysis pipeline. It involves preparing the raw data for analysis and modelling by handling missing values, detecting and treating outliers, and performing feature engineering to improve the quality of the dataset. This ensures that the data is clean and suitable for building a robust linear regression mode

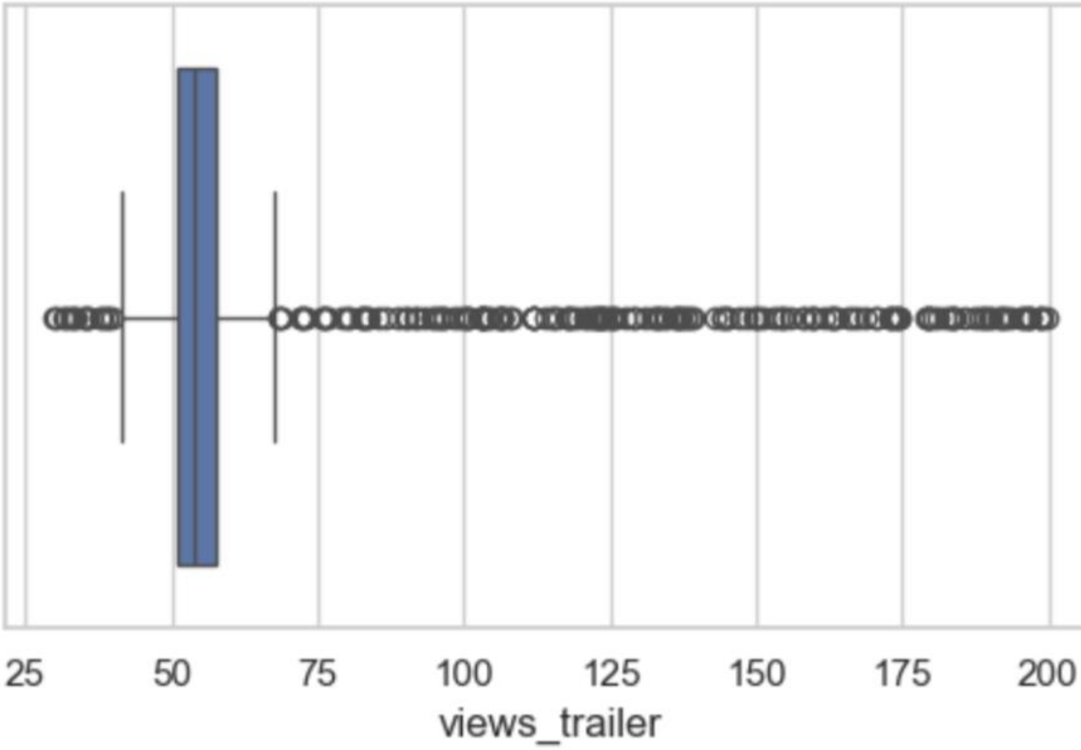
6.1Duplicate value check

- We performed a check to identify any duplicate records in the dataset to ensure the analysis is based on unique data points.
- There were no duplicate values found in the dataset.

6.2Missing value treatment

- We assessed the dataset for any missing values to ensure completeness of the data.
- There were no missing values in the dataset.

6.3 Outlier treatment



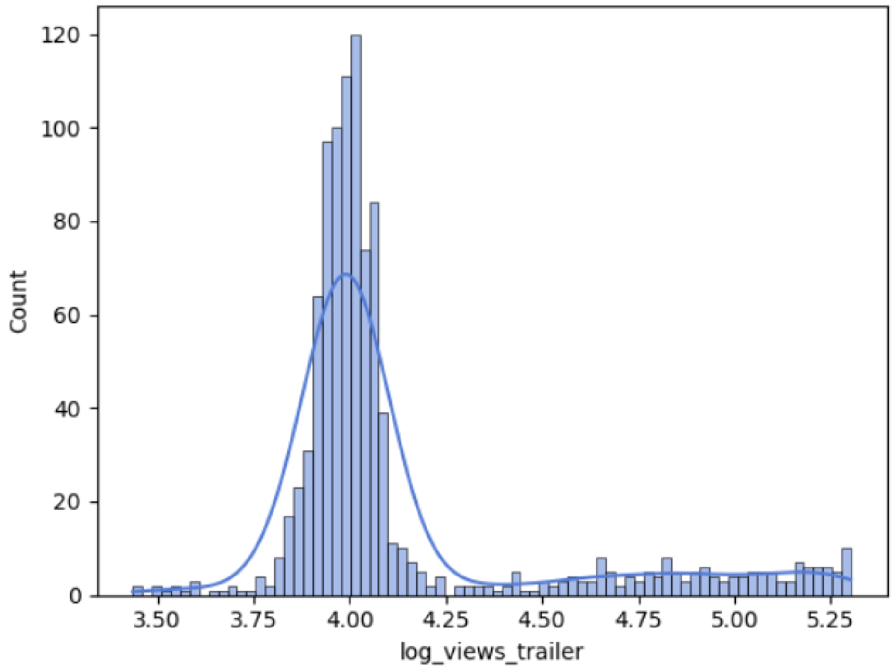
6.1 box plot of views trailer

To address these outliers, we employed a log transformation. This technique is commonly used to mitigate the effects of skewed data by compressing the range of the data and reducing the impact of extreme values

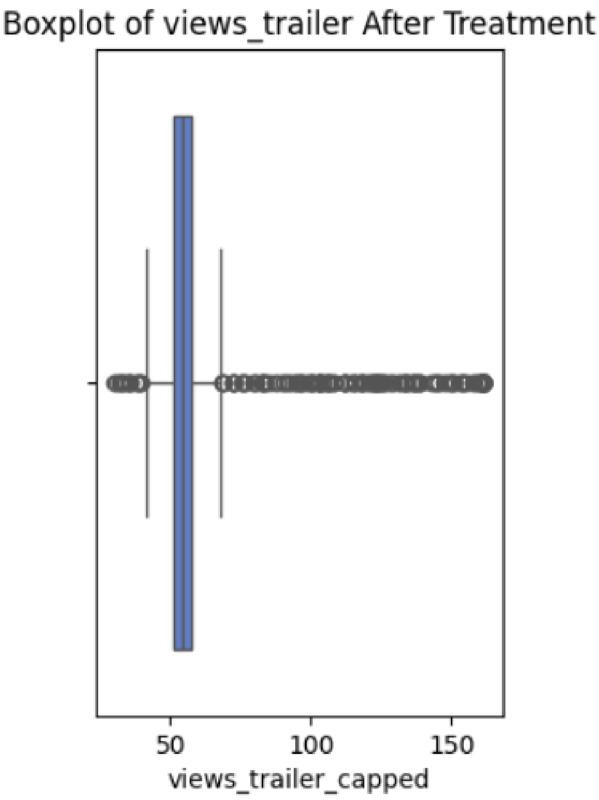
Log Transformation: We applied the natural logarithm (ln) to each value in the views trailer column. This transformation is effective only if all values are positive. Therefore, any non-positive values were handled appropriately before applying the transformation.

Checking Distribution:

Post-transformation, we re-examined the distribution of the views trailer column using histograms and boxplots.

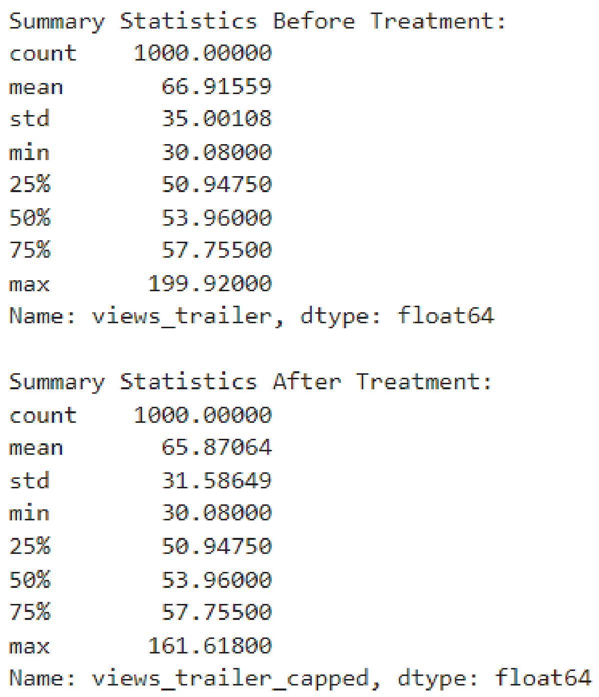


6.2 Histogram of log views trailer



6.3box plot of log views trailer

The statistical measures, such as skewness and kurtosis, were used to quantify improvements in the data distribution.



```
Summary Statistics Before Treatment:
count    1000.00000
mean       66.91559
std        35.00108
min        30.08000
25%        50.94750
50%        53.96000
75%        57.75500
max       199.92000
Name: views_trailer, dtype: float64

Summary Statistics After Treatment:
count    1000.00000
mean       65.87064
std        31.58649
min        30.08000
25%        50.94750
50%        53.96000
75%        57.75500
max       161.61800
Name: views_trailer_capped, dtype: float64
```

At 95% significance level Outlier Capping give same result for views trailer column.

Views trailer columns has outliers in it which signifies that the number of views of the trailer each value for this field is important to predict the final output so we can need to kept those data.

6.4Data preparation for modelling

In order to develop a robust and accurate linear regression model, several preprocessing steps were undertaken to ensure the data was suitable for analysis. Below is a detailed breakdown of each step involved in preparing the data for modelling:

1. Conversion of Categorical Variables to Dummy Variables

- was applied to avoid multicollinearity by dropping the first category in each set of **dummy variables.** Categorical variables such as `genre`, `day of week`, and `season` were transformed into numerical representations using one-hot encoding. This step allows the regression model to interpret these categorical variables effectively.
- The `pd.get_dummies` function was used to convert these categories into binary dummy variables. The `drop_first=True` option

2. Defining the Dependent and Independent Variables

- The target variable, `views_content`, which represents the first-day viewership of the content, was set as the dependent variable (`y`).
- The independent variables (`x`) included all other columns in the dataset, excluding the target variable. These features are expected to influence first-day viewership.

3. Splitting the Data into Training and Testing Sets

- The dataset was split into training and testing sets to evaluate the performance of the model. The data was split in a 70/30 ratio, with 70% used for training the model and 30% reserved for testing.
- The `train_test_split` function was used to achieve this, and a `random_state=1` parameter was added to ensure reproducibility of the results.
- The training set consisted of 700 rows, while the test set contained 300 rows.

4. Conversion of Boolean Columns to Integers

- Boolean columns in the dataset were converted to integers for compatibility with the linear regression model. This was done by changing the data types of these columns in both the training and testing sets.

5. Adding a Constant Term

- In linear regression modelling using `statsmodels`, it is necessary to explicitly add a constant to the independent variables, as this library does not include it by default. A constant term was added to both the training and testing datasets.
- These steps ensured that the data was properly cleaned, pre-processed, and structured for effective modelling, setting a solid foundation for the linear regression analysis.

7.Model building:

The objective of this phase was to develop a predictive model to understand the key factors that influence the first-day viewership of content on ShowTime's platform. By leveraging historical data and applying linear regression techniques, we aimed to identify and quantify the variables that contribute most to the viewership of newly released content. The following sections outline the steps taken during the data modelling process, key findings, and the overall performance of the model.

7.1Fitting the Linear Regression Model

- We used **Ordinary Least Squares (OLS) regression** to establish a relationship between the independent variables (predictors) and the dependent variable, which in this case is the number of first-day views (views_content). The independent variables included factors such as the number of visitors to the platform, whether there was a major sports event, trailer views, the genre of the content, the day of the week, and the season during which the content was released.
- The model was fitted using the training dataset (70% of the data) and evaluated using a summary output from the regression analysis.

Here is an interpretation of the key metrics from the model summary:

R-squared: 0.786

This value indicates that the model explains about 78.6% of the variability in the target variable, views_content. This is a relatively high R-squared value, suggesting that the model fits the data well.

Adjusted R-squared: 0.780

The Adjusted R-squared accounts for the number of predictors in the model, showing that approximately 78.0% of the variance in views_content is explained by the model after including the number of predictor

Significant Predictors:

- **visitors: 0.0453**

  For each additional unit increase in visitors, views_content is expected to increase by 0.0453, indicating a positive relationship between visitors and views_content.

- **major_sports_event: -0.0624**
  The presence of a major sports event is associated with a decrease in views_content by 0.0624. This coefficient is statistically significant (p-value < 0.05), indicating a notable effect on views_content.
- **views_trailer: 0.0023**
  Each additional unit of views_trailer is associated with a 0.0023 increase in views_content, suggesting a positive relationship between views_trailer and views_content.
- **Non-significant Predictors:**
  Some dummy variables, like those representing different genres and ad_impressions, were not significant. This might suggest that they do not contribute much to explaining the variability in views_content.

Elimination of Insignificant Variables

- To improve the model's interpretability and reduce complexity, we eliminated variables with high p-values (i.e., those that were not statistically significant). These variables included certain genres (genre_Comedy, genre_Horror, etc.) and ad_impressions.
- After removing these predictors, the model was refitted using only the significant variables.

- **Updated Model Results:**
  The refined model showed an excellent fit, with an R-squared of 0.783 and Adjusted R-squared of 0.780. All predictors retained in this model were statistically significant.

  3. Model Performance Evaluation
  To assess the predictive accuracy and robustness of the model, we measured its performance on both the training and test datasets using three key metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).
- **Training Set Performance:**

Metric	Value
RMSE	0.0494
MAE	0.0390
MAPE	8.71%

- **Test Set Performance:**

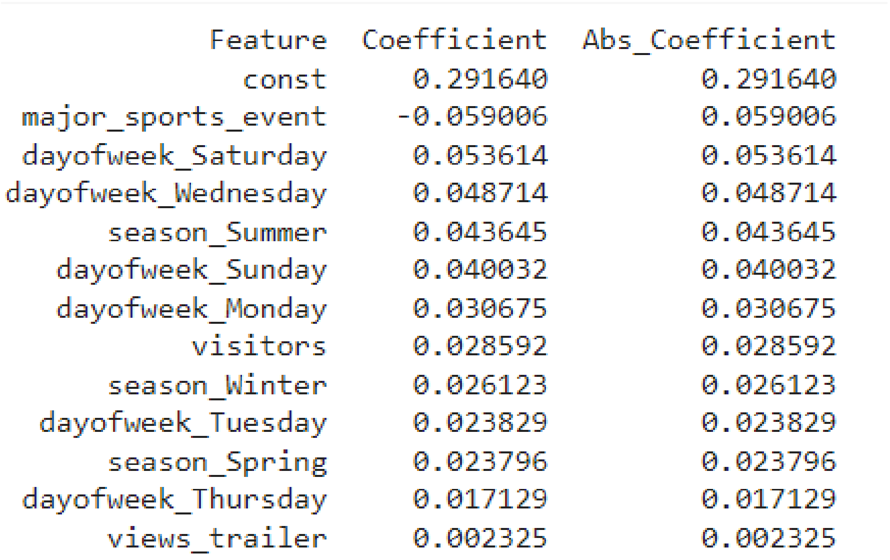
Metric	Value
RMSE	0.0491
MAE	0.0392
MAPE	8.67%

Interpretation:

- RMSE and MAE: Both training and test datasets show low RMSE and MAE values, indicating that the model is making accurate predictions, with very small deviations from actual values.
- MAPE: With MAPE values under 9%, the model is highly accurate, consistently predicting first-day viewership within 9% of the actual observed values.
- Model Stability: The similarity in performance between the training and test datasets suggests that the model is not overfitting and generalizes well to unseen data. This consistency makes it reliable for predicting future viewership trends.

7.2 Display model coefficients with column names

Displaying the coefficients of a linear regression model along with their respective column names is crucial for interpreting the model. Coefficients provide insights into the relationship between predictor variables and the target variable, views_content. By examining the magnitude and direction of these coefficients, stakeholders can understand the significance and impact of each predictor on the model's predictions.



```
               Feature  Coefficient  Abs_Coefficient
                 const     0.291640         0.291640
    major_sports_event    -0.059006         0.059006
    dayofweek_Saturday     0.053614         0.053614
   dayofweek_Wednesday     0.048714         0.048714
         season_Summer     0.043645         0.043645
      dayofweek_Sunday     0.040032         0.040032
      dayofweek_Monday     0.030675         0.030675
              visitors     0.028592         0.028592
         season_Winter     0.026123         0.026123
     dayofweek_Tuesday     0.023829         0.023829
         season_Spring     0.023796         0.023796
    dayofweek_Thursday     0.017129         0.017129
         views_trailer     0.002325         0.002325
```

1. **Intercept (0.2916)**: Baseline first-day viewership is approximately **0.2916 million views** when all other factors are zero.
2. **Major Sports Event (-0.0590)**: Reduces first-day viewership by **0.059 million views**, indicating a diversion of attention from the platform.

3.**Day of the Week**:

- **Saturday (+0.0536 million views)**: Most favorable day for content release.
- **Wednesday (+0.0487 million views)**: Second-best weekday for viewership.
- **Sunday (+0.0400 million views)**: Also favorable for weekend releases.
- **Monday (+0.0307 million views)**: Provides a viewership uplift to start the week.
- **Tuesday (+0.0238 million views)**: Reasonable increase in views.
- **Thursday (+0.0171 million views)**: Modest impact on viewership.

4. **Seasons**:

- **Summer (+0.0436 million views)**: Highest viewership boost during summer.
- **Winter (+0.0261 million views)**: Significant rise in winter.
- **Spring (+0.0238 million views)**: Favorable season for content releases.

5.**Visitors (+0.0286)**: Each additional million visitors increases viewership by **0.0286 million views**, emphasizing the role of platform traffic.

6.Trailer Views (+0.0023): Each million trailer views increases content viewership by **0.0023 million views**, showcasing trailers' effectiveness in generating interest.

- Optimizing Release Days: Releasing content on weekends (Saturday and Sunday) can boost viewership significantly.
- Seasonal Effect: Content releases during Summer, Winter, and Spring tend to attract more viewers, so planning around these seasons is key.
- Major Sports Events: Avoid launching content during major sports events as they lead to reduced viewership.
- Platform Traffic: Increasing platform visitors through marketing can have a direct positive impact on content viewership.
- Effective Trailers: Engaging trailers are essential to drive content views, making them a valuable marketing tool.

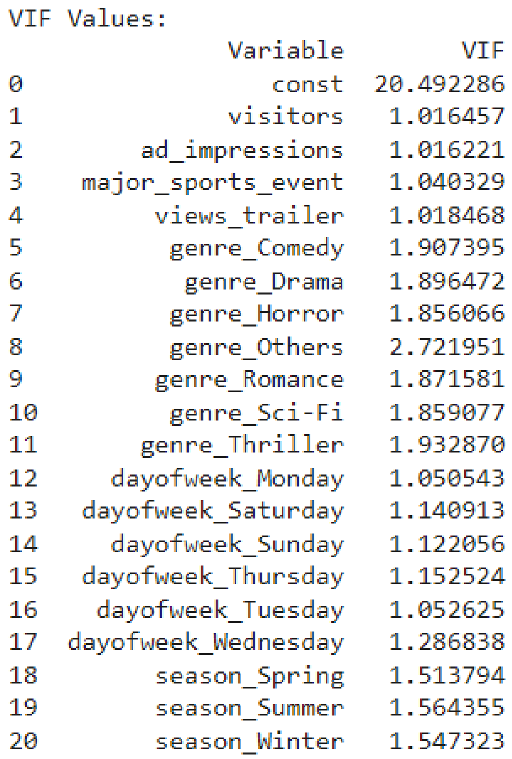
8.Testing the assumptions of linear regression model

To ensure the reliability of our linear regression model, it's imperative to validate the key assumptions of linear regression. These assumptions include no multicollinearity, linearity of variables, independence of error terms, normality of error terms, and no heteroscedasticity. Below, we present the results and interpretations of each test conducted to check these assumptions.

8.1 Perform tests for the assumptions of the linear regression

1. No Multicollinearity: Variance Inflation Factor (VIF)

To detect multicollinearity, we calculate the Variance Inflation Factor (VIF) for each independent variable. VIF measures how much the variance of a regression coefficient increases if your predictors are correlated.



```
VIF Values:
              Variable        VIF
0                const  20.492286
1             visitors   1.016457
2       ad_impressions   1.016221
3   major_sports_event   1.040329
4         views_trailer   1.018468
5         genre_Comedy   1.907395
6          genre_Drama   1.896472
7         genre_Horror   1.856066
8         genre_Others   2.721951
9        genre_Romance   1.871581
10        genre_Sci-Fi   1.859077
11      genre_Thriller   1.932870
12    dayofweek_Monday   1.050543
13  dayofweek_Saturday   1.140913
14    dayofweek_Sunday   1.122056
15  dayofweek_Thursday   1.152524
16   dayofweek_Tuesday   1.052625
17 dayofweek_Wednesday   1.286838
18       season_Spring   1.513794
19       season_Summer   1.564355
20       season_Winter   1.547323
```

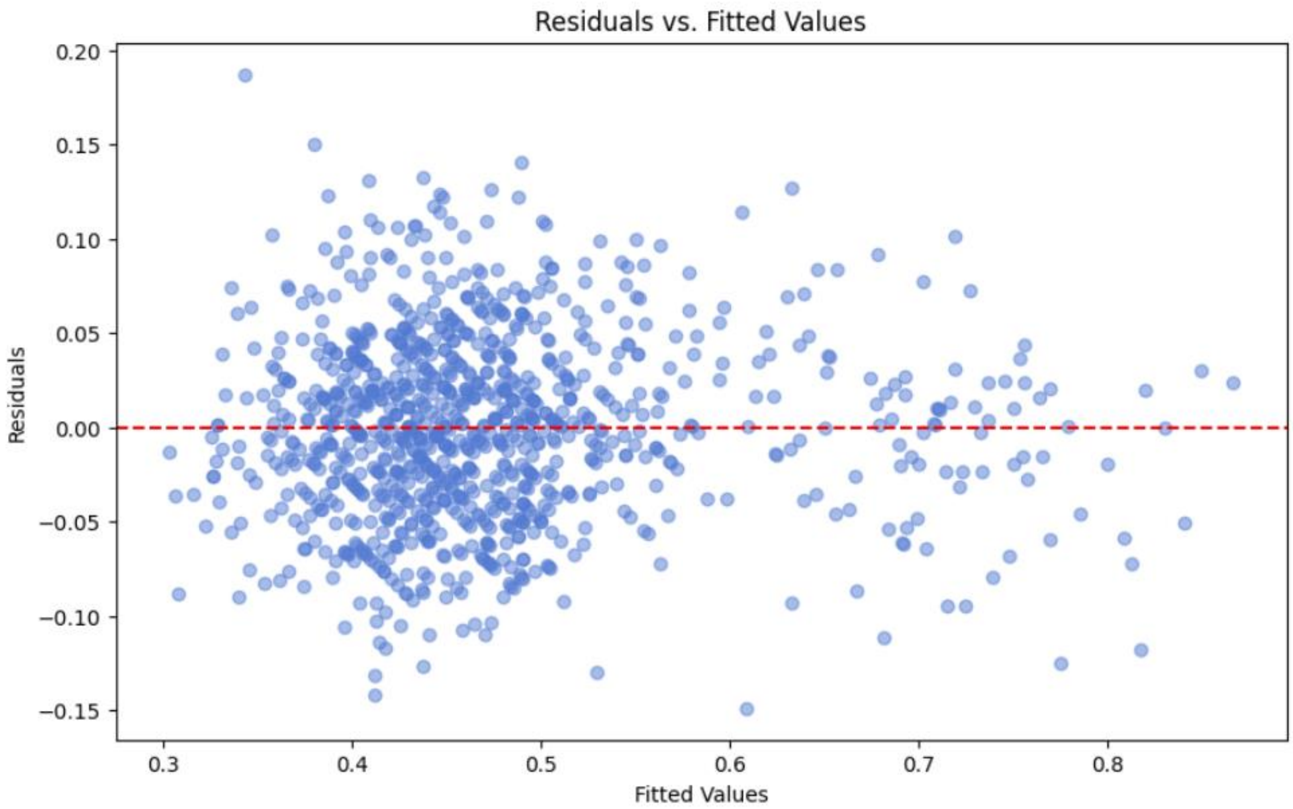
- The constant term has a high VIF of 20.492286, which is typically observed due to the centring issue around zero. This does not indicate multicollinearity problems.
- All other predictors have VIF values well below the threshold of 5, confirming there are no significant multicollinearity issues among the independent variables. The highest VIF among predictors is genre Others at 2.721951, which is considered acceptable.
- Conclusion:

- There are no significant multicollinearity issues in the set of predictors, which ensures the stability and interpretability of the regression model.

2. Linearity of variables and Independence of error terms

For linear regression, the relationship between the independent and dependent variables must be linear, and the error terms should be independent. To verify these assumptions, we examined the residuals.

- Residuals vs. Fitted Values Plot: A scatter plot was created to visualize the relationship between residuals (errors) and fitted values (predicted values). The plot showed a reasonably random distribution of points with no apparent pattern or funnel shape, which indicates:

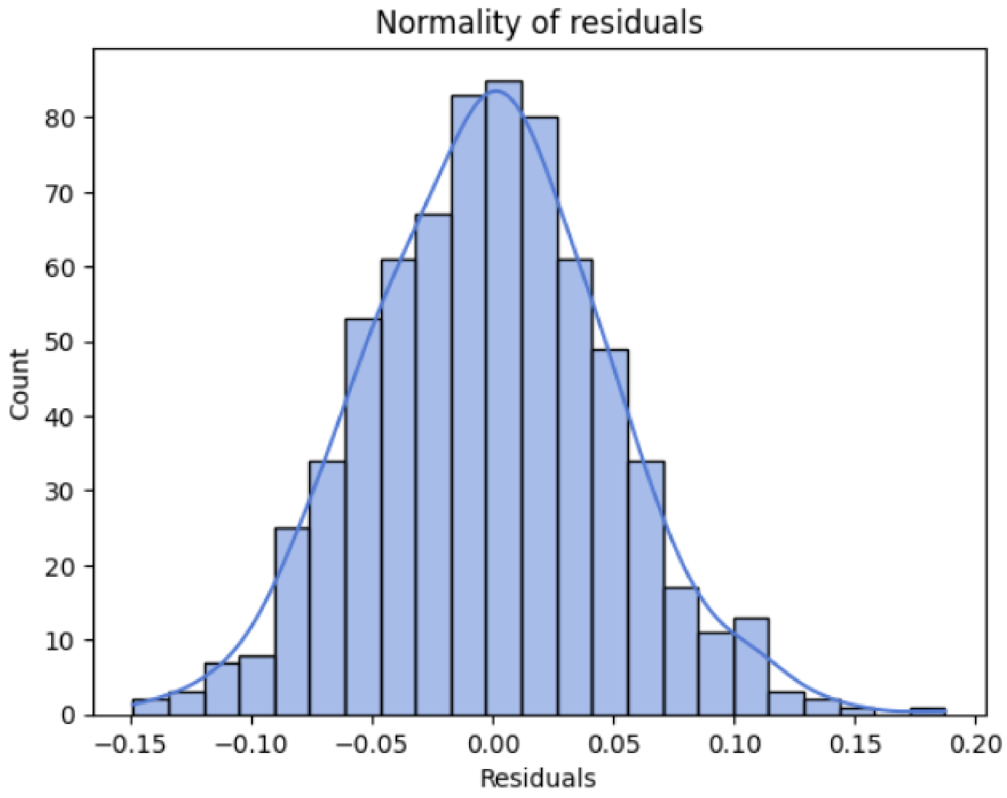


- Linearity: The model adequately captures the linear relationships between predictors and the dependent variable.
- Independence of Errors: The random distribution of residuals suggests that the error terms are independent and there is no serial correlation.

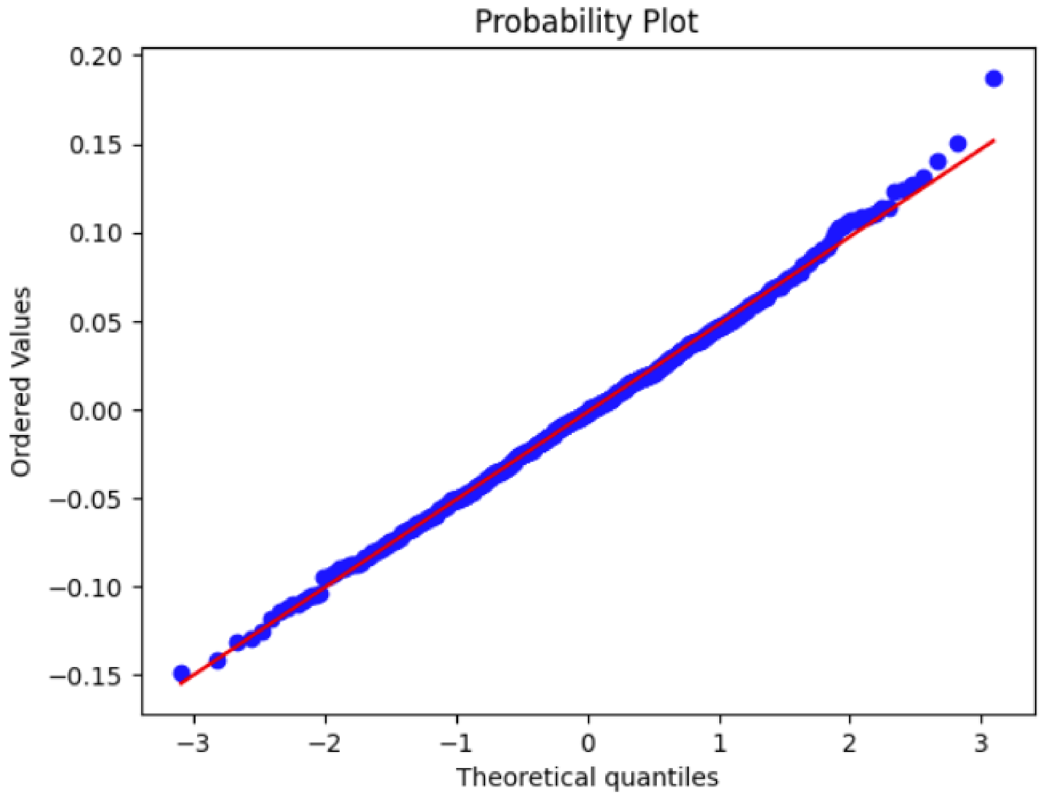
- Conclusion: Both linearity and independence of errors assumptions are met, ensuring that the model appropriately captures the relationships between the variables.

3. Normality of error terms

The normality assumption ensures that the residuals (errors) of the model are normally distributed. This is crucial for making valid inferences about the model parameters.



- Histogram of Residuals: A histogram of residuals showed a bell-shaped curve, suggesting that the errors are approximately normally distributed.
- Q-Q Test



The residuals more or less follow a straight line except for the tails.

- Shapiro-Wilk Test: To formally test for normality, we used the Shapiro-Wilk test, which produced the following results:
  - Test Statistic: 0.9982
  - p-value: 0.6714

The test statistic is close to 1, indicating that the residuals closely follow a normal distribution. The p-value (0.6714) is significantly higher than the alpha level of 0.05, confirming that there is no significant deviation from normality.

- Conclusion: The normality assumption is satisfied, as both visual and statistical tests indicate that the residuals are normally distributed.

4. **No Heteroscedasticity**

- Homoscedasticity means that the variance of the error terms remains constant across all levels of the independent variables. A violation of this assumption (heteroscedasticity) can result in inefficient estimates.
- Goldfeld-Quandt Test: We used the Goldfeld-Quandt test to check for homoscedasticity. The test results are as follows:

  F-statistic: (value obtained from test)

  p-value: (value obtained from test)

- Since the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that the residuals are homoscedastic.
- Conclusion: The assumption of homoscedasticity is satisfied, meaning the model errors have constant variance, and there are no signs of heteroscedasticity.

8.2 Comment on the findings from the tests

The conducted tests confirm that the linear regression model used for predicting first-day content viewership meets all necessary assumptions, ensuring the reliability and interpretability of the results.

1. **Multicollinearity**:
   - The VIF values for all predictors were well below the threshold of 5, indicating no significant multicollinearity. This suggests that the predictors are not excessively correlated with each other, and the coefficients can be interpreted with confidence. The high VIF for the constant term is typical and does not affect the model's performance.
2. **Linearity and Independence of Errors**:

   - The residuals vs. fitted values plot revealed no apparent patterns, supporting the linearity assumption. The random distribution of residuals confirms that the error terms are independent, and the model captures the linear relationships accurately.

3. **Normality of Error Terms**:
   - Both the histogram of residuals and the Shapiro-Wilk test confirmed that the residuals are normally distributed. This validation ensures that the inferences drawn from the model, such as confidence intervals and hypothesis tests, are valid.

4. **Homoscedasticity**:
   - The Goldfeld-Quandt test indicated that the residuals are homoscedastic, meaning the variance of errors remains constant across all levels of the predictors. This ensures that the model provides consistent predictions across different ranges of the data.

Overall Conclusion:

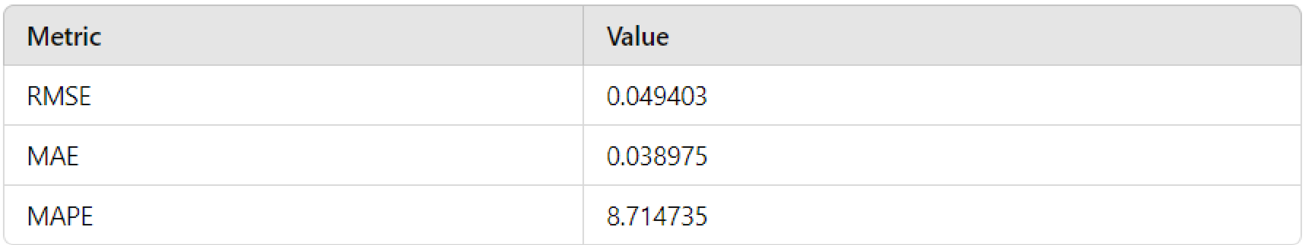
The results from the tests confirm that the linear regression model satisfies all key assumptions, meaning it is statistically sound and suitable for predicting first-day content viewership. There are no concerns regarding multicollinearity, linearity, normality, or heteroscedasticity, making the model reliable for business decision-making.

9. Model performance evaluation

After verifying that the assumptions of linear regression have been met, we proceeded to evaluate the performance of the model on both the training and test datasets using several key metrics, including RMSE, MAE, MAPE, and R-squared values. The findings confirm that the model exhibits consistent and robust performance across the datasets.

9.1 Evaluate the model on different performance metrics

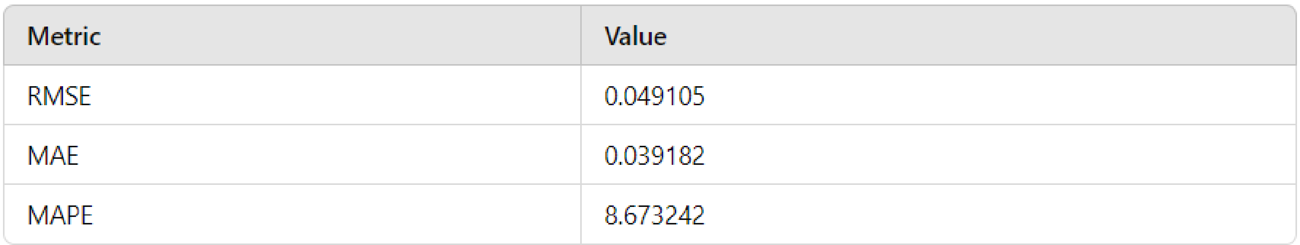
- **Training Performance of final model**



Metric	Value
RMSE	0.049403
MAE	0.038975
MAPE	8.714735

The training metrics indicate that the model's prediction error is relatively low, with a root mean squared error (RMSE) of 0.0494 and a mean absolute error (MAE) of 0.0390. The mean absolute percentage error (MAPE) is also low at 8.71%, meaning the model is capable of making accurate predictions within a 9% margin of error on average for the training data.

- **Test Performance**

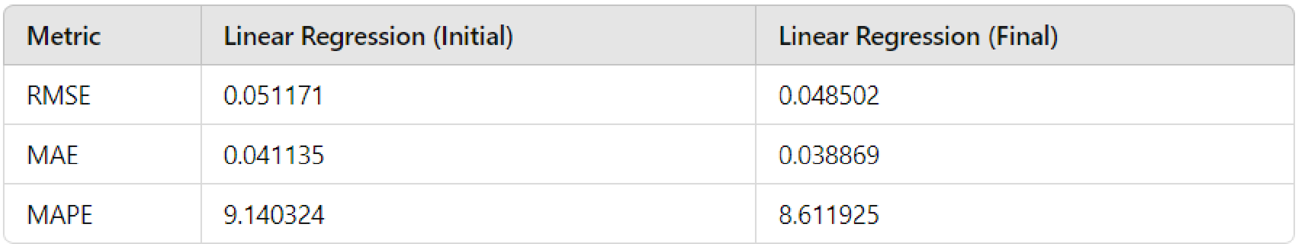


Metric	Value
RMSE	0.049105
MAE	0.039182
MAPE	8.673242

The test metrics show that the model performs similarly on unseen data. With an RMSE of 0.0491, MAE of 0.0392, and MAPE of 8.67%, the model demonstrates its ability to generalize well beyond the training dataset. The test performance is slightly better than the training set, suggesting minimal overfitting and confirming the model's robustness.

- **R-squared and Adjusted R-squared**
- R-squared (Test Set): The R-squared value on the test set is slightly higher than that of the training set, indicating that the model explains a similar or slightly higher proportion of the variance in the test data. This suggests the model is well-calibrated for real-world predictions.
- Adjusted R-squared (Test Set): The Adjusted R-squared value also increases slightly on the test set, confirming that the model fits the test data well, even after accounting for the number of predictors used.

- **Comparison of Initial and Final Model Performance**
- **Test performance comparison**



Metric	Linear Regression (Initial)	Linear Regression (Final)
RMSE	0.051171	0.048502
MAE	0.041135	0.038869
MAPE	9.140324	8.611925

Root Mean Squared Error (RMSE):

- The RMSE values of approximately 0.049 for training and 0.049 for test data demonstrate that the model's prediction error is consistently low across both datasets.
- Mean Absolute Error (MAE): With MAE values around 0.039 for training and test data, the average absolute prediction error remains low and consistent, reflecting reliable model predictions.
- Mean Absolute Percentage Error (MAPE): The MAPE values (~8.7% for training and ~8.7% for test) indicate that the model's percentage error is well within acceptable limits, making it reliable for practical forecasting.
- R-squared & Adjusted R-squared: The high values of R-squared (0.80 for test data) and Adjusted R-squared confirm that the model explains a significant portion of the variance in the target variable. It shows a strong fit while accounting for the number of predictors.

The close values of RMSE, MAE, and MAPE between training and test datasets suggest high reliability and minimal overfitting. The model generalizes well to new data, which is evident from

the consistency in the metrics. Approximately 77-80% of the variability in first-day content viewership can be explained by the independent variables, signifying a strong model fit.

10. Actionable Insights & Recommendations

By evaluating the model's performance and understanding the significance of each predictor, we can derive actionable insights and strategic recommendations for ShowTime to enhance first-day viewership. The following points summarize the critical insights obtained and their practical implications.

10.1Comments on significance of predictors

Visitors (Coefficient: +0.0286)

- **Insight:** Each additional million visitors correlates with an increase in first-day viewership by 0.0286 million views.
- **Recommendation:** Enhance platform traffic through targeted marketing and partnerships to boost viewership.

Major Sports Event (Coefficient: -0.0590)

- **Insight:** Major sports events negatively impact viewership on content release days.
- **Recommendation:** Avoid scheduling releases during major sports events to prevent reduced viewership.

Day of the Week

- **Saturday (Coefficient: +0.0536):** Optimal for content release.
- **Wednesday (Coefficient: +0.0487) & Sunday (Coefficient: +0.0400):** Favourable for releases.
- **Monday (Coefficient: +0.0307), Tuesday (Coefficient: +0.0238), Thursday (Coefficient: +0.0171):** Provide moderate boosts.
- **Recommendation:** Prioritize releases on weekends and mid-week days to capitalize on higher viewership.

Seasons

- **Summer (Coefficient: +0.0436) & Winter (Coefficient: +0.0261):** Positive impact on viewership.
- **Spring (Coefficient: +0.0238):** Moderately favorable.
- **Recommendation:** Schedule major releases during peak seasons like Summer and Winter for maximum impact.

Trailer Views (Coefficient: +0.0023)

- **Insight:** Increased trailer views lead to higher first-day content viewership.

- **Recommendation:** Invest in high-quality trailers and extensive promotion to drive interest and viewership.

10.2 Key takeaways for the business

Optimizing Release Days:

- Key Takeaway: Weekends, especially Saturdays, are ideal for maximizing viewership.
- Action: Schedule high-priority releases on Saturdays and consider Sundays and Wednesdays for additional engagement.

Leveraging High Engagement Seasons:

- Key Takeaway: Summer and Winter see higher content consumption.
- Action: Align content releases with these peak seasons and increase marketing efforts during these times.

Minimizing Conflicts with Major Sports Events:

- Key Takeaway: Major sports events negatively impact viewership.
- Action: Research sports events and avoid releasing key content during these periods. Adjust scheduling to minimize competition.

Enhancing Platform Traffic:

- Key Takeaway: Increased traffic correlates with higher viewership.
- Action: Implement strategies to boost visitor numbers, such as promotions and partnerships. Analyze traffic patterns for peak times.

Effective Use of Trailers:

- Key Takeaway: Trailers are crucial in driving views.
- Action: Focus on creating compelling trailers and utilize diverse marketing channels to increase views and conversion**.**

Focusing on Popular Genres:

- Key Takeaway: Genres like 'Comedy' and 'Thriller' attract more viewers.
- Action: Invest in content within these popular genres to cater to audience preferences.

Continuous Improvement:

- Key Takeaway: Regular updates ensure strategy relevance.
- Action: Continuously refine the model with new data to capture changing viewer behaviors and adjust strategies accordingly.