Predictive Modeling Project:

Linear Regression Analysis for ShowTime

Submitted by:

Arkaprava Mazumder

Date of Submission: 13 July, 2025

Table of Contents

Problem Definition

ShowTime, an OTT (Over-The-Top) media service provider, aims to enhance the performance of its content by improving first-day viewership. With the growing competitiveness in the streaming industry and the shift in consumer behavior toward on-demand digital content, understanding the key factors that influence how content performs on the first day of its release has become critical.

In this project, we are tasked with identifying the **driving factors that significantly impact first-day content viewership** on the ShowTime platform. By analyzing the available data on content performance—including marketing metrics, platform engagement, release timing, and genre—we aim to build a **linear regression model** that can quantify the influence of these variables on viewership.

This analysis will help ShowTime:

- Recognize patterns and trends in successful content releases,

- Optimize release strategies (e.g., timing, promotions),

- Improve marketing and content planning efforts,

- Ultimately boost user engagement and platform retention.

Data Overview

The dataset contains 1000 records of content metadata with the following fields:

- visitors: Weekly platform visitors (in millions)

- ad_impressions: Number of ad impressions (in millions)

- major_sports_event: 0/1 indicating major sports clash on the release day

- genre: Genre of the content (categorical)

- dayofweek: Release day of the week (categorical)

- season: Season of release (categorical)

- views_trailer: Trailer views (in millions)

- views_content: Target variable - First-day content views (in millions)

Exploratory Data Analysis (EDA)

In this section, we conduct a thorough exploratory analysis to understand the key patterns, distributions, and relationships within the dataset provided by ShowTime, an OTT content provider. The main objective of this step is to gain preliminary insights into the data that can guide model building and feature selection later in the process.

The EDA begins with a summary of the dataset's structure and variables, followed by univariate analysis to examine the distribution of individual features, especially the target variable (`views_content`). Next, we perform bivariate analysis to explore the relationships between features, such as how trailer views, genres, day of release, and seasonality impact content viewership. Key questions posed by the business are answered using relevant plots and statistical summaries.

Insights drawn from this section help in identifying trends, patterns, and possible transformations required, and also highlight which variables might be important drivers of content performance.

Key Questions to be answered:

1. What does the distribution of content views look like?
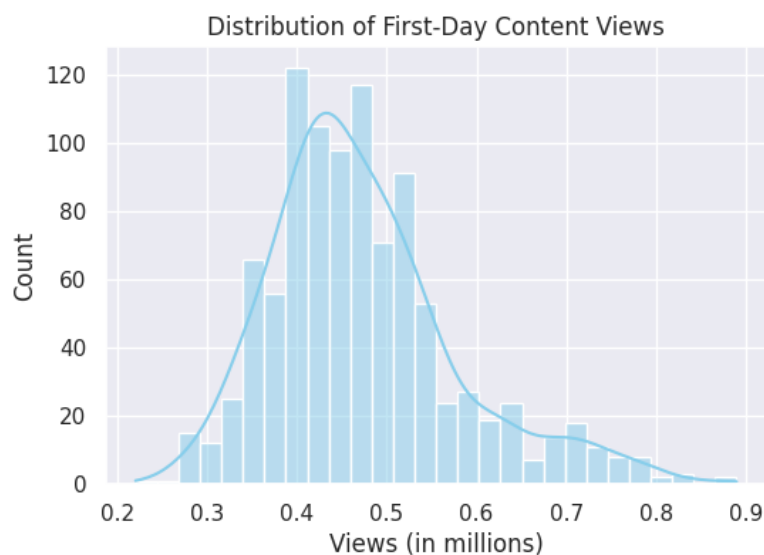
Fig. 1.1

The histogram of first-day content views appears roughly normal but slightly right-skewed, with most content receiving between **0.35 to 0.55 million** views. This indicates that while most content garners moderate first-day attention, a few titles achieve significantly higher viewership.

Insights:

- Most content clusters around mid-range viewership, suggesting consistent audience engagement.

- A few outlier contents show high demand or hype pre-launch.

2. What does the distribution of genres look like?
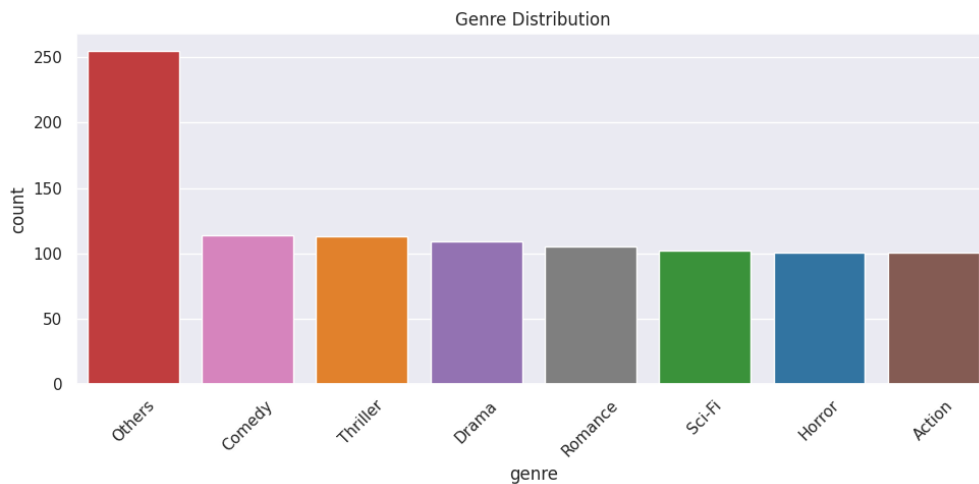
Fig. 1.2

The bar plot shows that the **"Others" genre** dominates the distribution, while Comedy, Thriller, Drama, Romance, Sci-Fi, Horror, and Action are more evenly distributed among the remaining data.

Insights:

- The imbalance due to the "Others" category could influence modeling unless handled correctly.

- Diverse genres allow for capturing a range of audience preferences.

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?

First-Day Viewership by Day of Release
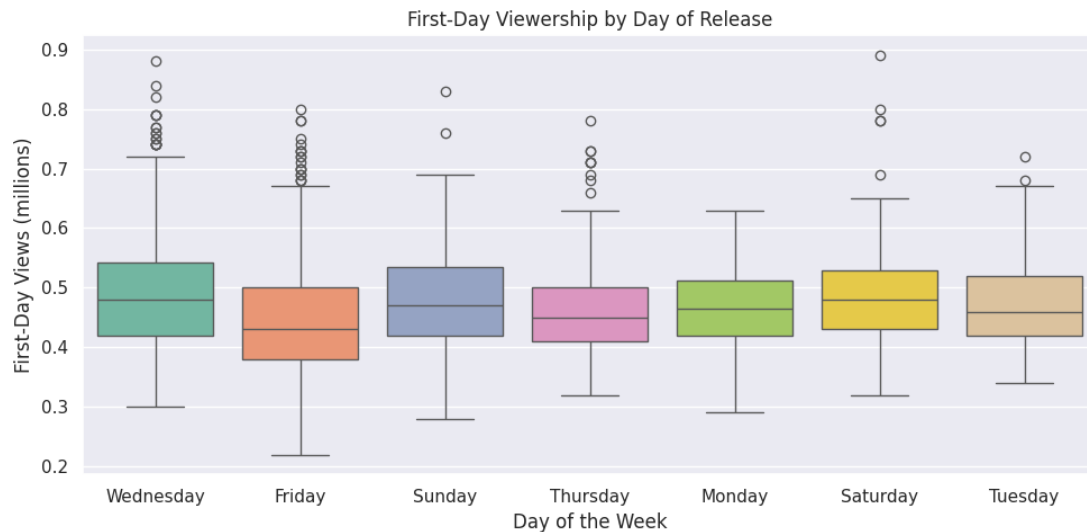
Fig. 1.3

Boxplots indicate that content released on **Saturday, Wednesday, and Sunday** tends to have higher median first-day views. **Friday releases** show more variation, suggesting uneven popularity among content released on that day.

Insights:

- Midweek and weekend releases seem to perform better in general.

- High variability on Fridays could be due to blockbuster and niche releases combined.

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

Viewership by Release Season
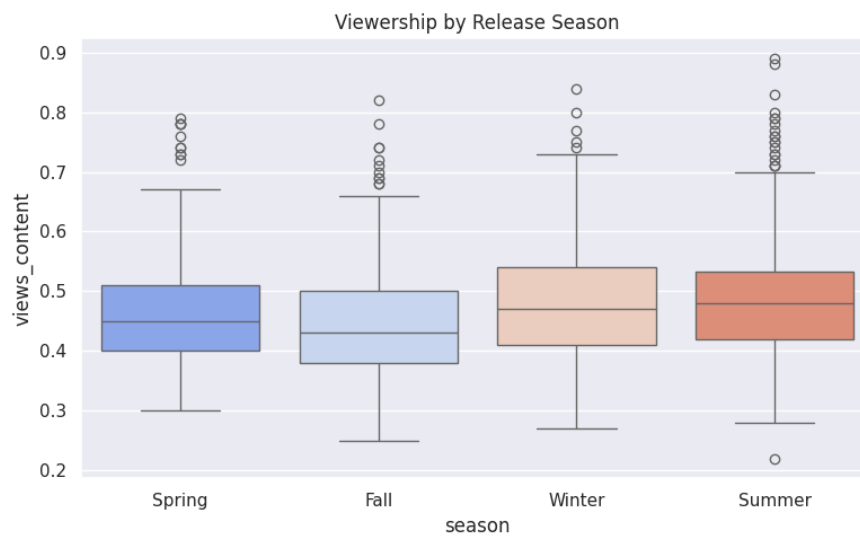
Fig. 1.4

Boxplots show that **Summer and Winter** releases have slightly higher median views compared to Spring or Fall. Summer, in particular, has a wider range of viewership, possibly due to school vacations or seasonal marketing.

Insights:

- Seasonal timing plays a notable role in viewership, likely due to changes in audience availability.

- Strategic release timing could boost first-day views.

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

The scatter plot reveals a **strong positive linear correlation** between trailer views and first-day content views, suggesting that well-promoted content attracts more initial attention.
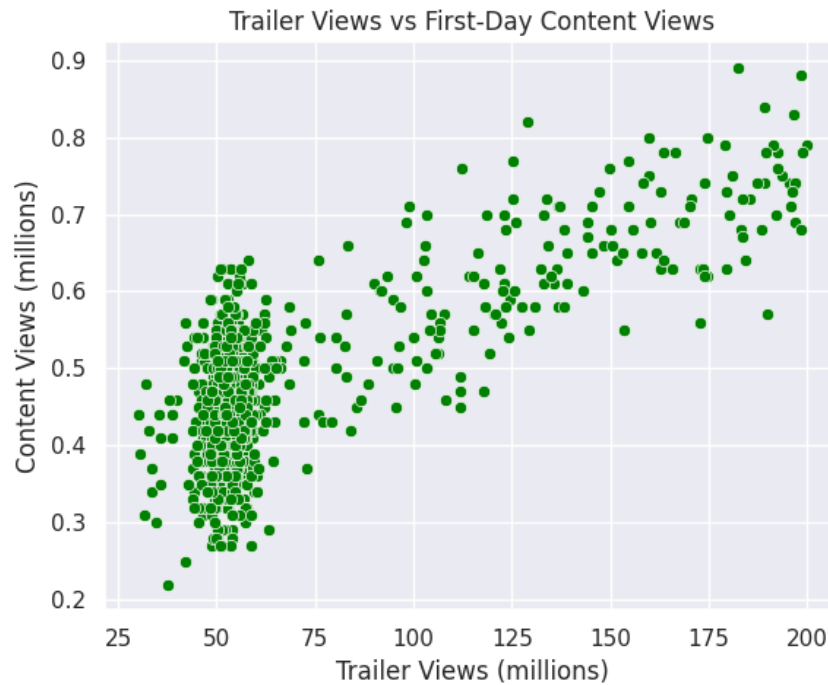
Fig. 1.5

Insights:

- Trailer performance is a strong indicator of content success.

- Marketing efforts before release can substantially impact viewership metrics.

Correlation Matrix
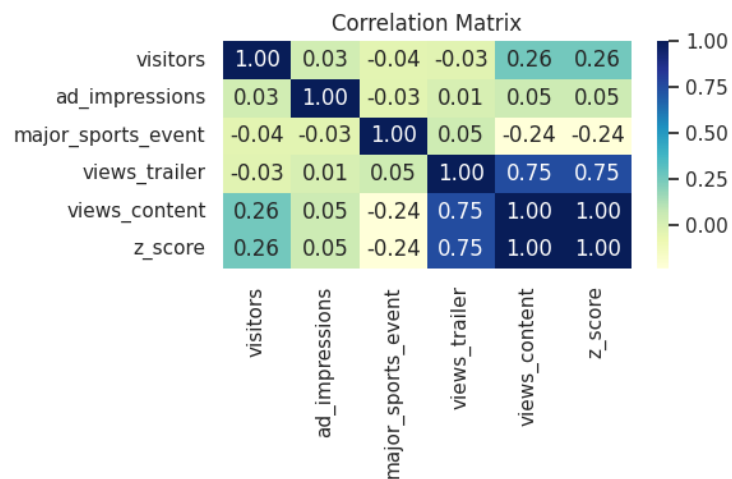
Fig. 1.6

- **views_trailer** shows a **strong positive correlation** (r ≈ 0.75) with **views_content**, indicating that trailer performance is a key predictor of content performance.

- **visitors** have a moderate correlation (r ≈ 0.26) with first-day views, suggesting platform traffic plays a role.

- **major_sports_event** has a **negative correlation** (r ≈ -0.24) with first-day views, hinting at reduced content consumption during sports events.

- **ad_impressions** has minimal correlation (r ≈ 0.05), suggesting its effect is weak on first-day views in isolation.

Pairwise Relationships

Pairwise Relationships and Correlations
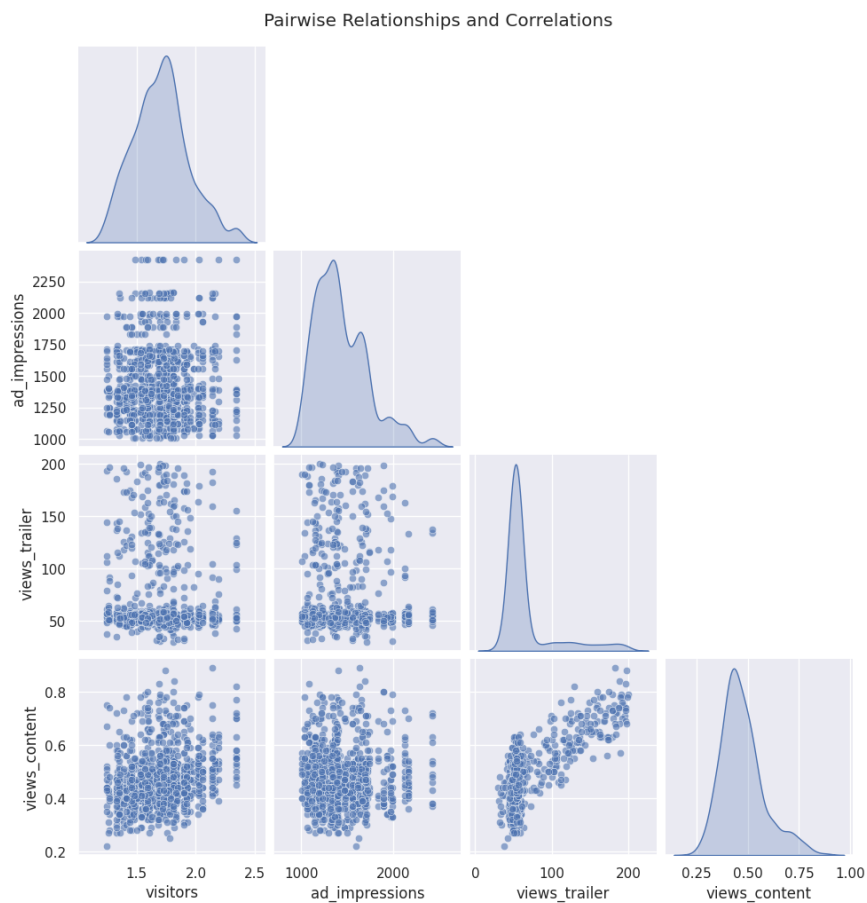
Fig. 1.7

- There is **visible linearity** between views_trailer and views_content, reinforcing the strength of this relationship.
- ad_impressions and visitors show **scattered relationships**, less predictive but still important in multivariate context.
- The diagonal histograms show distribution skews: views_trailer is right-skewed, while views_content is nearly normal with some tailing.

Viewership vs. Major Sports Event
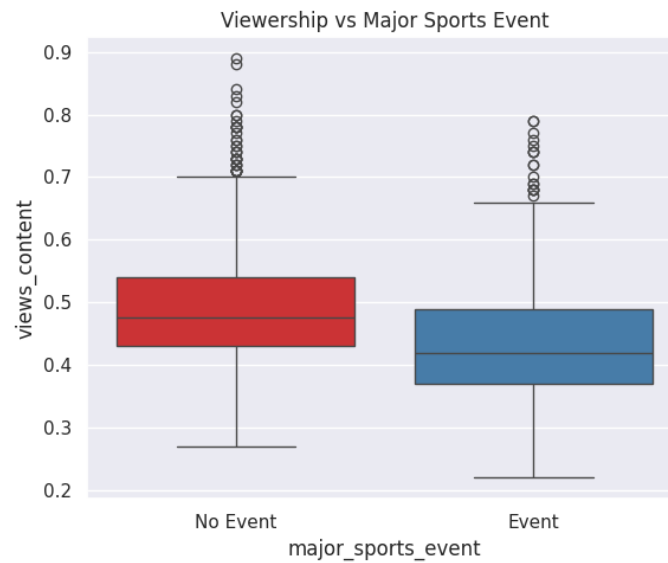
Fig. 1.8

- Content released **without a sports event** tends to receive higher **median and upper quartile views**.

- There are **more high-performing outliers** when no sports event occurs.

- Presence of a sports event **likely diverts audience attention**, thus reducing first-day content performance.

Data preprocessing

Data Description

	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

The data contains 1000 rows and 8 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   visitors            1000 non-null   float64
 1   ad_impressions      1000 non-null   float64
 2   major_sports_event  1000 non-null   int64
 3   genre               1000 non-null   object
 4   dayofweek           1000 non-null   object
 5   season              1000 non-null   object
 6   views_trailer       1000 non-null   float64
 7   views_content       1000 non-null   float64
dtypes: float64(4), int64(1), object(3)
memory usage: 62.6+ KB
```

From the 8 columns 8 of them are numeric and other 3 are the categorical columns. For Linear Regression model we only need numeric columns.

```
oData.describe(include='all').T
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
visitors	1000.0	NaN	NaN	NaN	1.70429	0.231973	1.25	1.55	1.7	1.83	2.34
ad_impressions	1000.0	NaN	NaN	NaN	1434.71229	289.534834	1010.87	1210.33	1383.58	1623.67	2424.2
major_sports_event	1000.0	NaN	NaN	NaN	0.4	0.490143	0.0	0.0	0.0	1.0	1.0
genre	1000	8	Others	255	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dayofweek	1000	7	Friday	369	NaN	NaN	NaN	NaN	NaN	NaN	NaN
season	1000	4	Winter	257	NaN	NaN	NaN	NaN	NaN	NaN	NaN
views_trailer	1000.0	NaN	NaN	NaN	66.91559	35.00108	30.08	50.9475	53.96	57.755	199.92
views_content	1000.0	NaN	NaN	NaN	0.4734	0.105914	0.22	0.4	0.45	0.52	0.89

Key Initial Insights from the Data description

- Average first-day views are around 0.47 million, with a wide range (from 0.22M to 0.89M), indicating that some content performs much better than others.

- Platform visitors (last week) range from 1.25M to 2.34M — higher traffic may lead to higher first-day views.

- Ad impressions (marketing spend) vary significantly — more impressions might drive more awareness and better first-day performance.

- Major sports events occurred on 40% of release days — these could divert audience attention and reduce viewership.

- Trailer views show large variation (30M to 200M) — likely a strong predictor of content interest and first-day success.

- Genre distribution is skewed (e.g., "Others" is the most common), but still includes enough variety to compare viewership across genres.

- Friday is the most common release day, followed by weekends — release timing might influence viewership due to user availability.

- Winter and Summer releases are slightly more frequent — possibly linked to holidays or leisure periods that affect viewer engagement.

Missing Values

	0
visitors	0
ad_impressions	0
major_sports_event	0
genre	0
dayofweek	0
season	0
views_trailer	0
views_content	0

dtype: int64

No missing values were found in the dataset. No imputation or deletion required for data cleaning.

Duplicate Check

There are no duplicate rows found in the dataset.

Outlier detection & treatment

We used the Z-Score method to detect potential outliers in the target variable views_content. A z-score > 3 or < -3 is typically considered an outlier.

	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content	z_score
0	1.67	1113.81	0	Horror	Wednesday	Spring	56.70	0.51	0.345735
1	1.46	1498.41	1	Thriller	Friday	Fall	52.69	0.32	-1.449064
2	1.47	1079.19	1	Thriller	Wednesday	Fall	48.74	0.39	-0.787822
3	1.85	1342.77	1	Sci-Fi	Friday	Fall	49.81	0.44	-0.315507
4	1.46	1498.41	0	Sci-Fi	Sunday	Winter	55.83	0.46	-0.126581

	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content	z_score
124	1.82	1690.43	0	Sci-Fi	Wednesday	Winter	189.00	0.84	3.463018
141	1.74	1398.05	0	Action	Wednesday	Summer	198.31	0.88	3.840870
489	1.81	1890.97	0	Romance	Saturday	Winter	159.73	0.80	3.085166
825	2.34	1629.94	0	Others	Wednesday	Fall	128.97	0.82	3.274092
854	1.59	1890.97	0	Thriller	Friday	Summer	174.62	0.80	3.085166
937	1.69	1079.19	0	Romance	Sunday	Summer	196.42	0.83	3.368555
987	2.14	1629.94	0	Romance	Saturday	Summer	182.33	0.89	3.935334

A total of 7 records have z-scores greater than 3 (extreme high values), indicating that they are potential outliers.
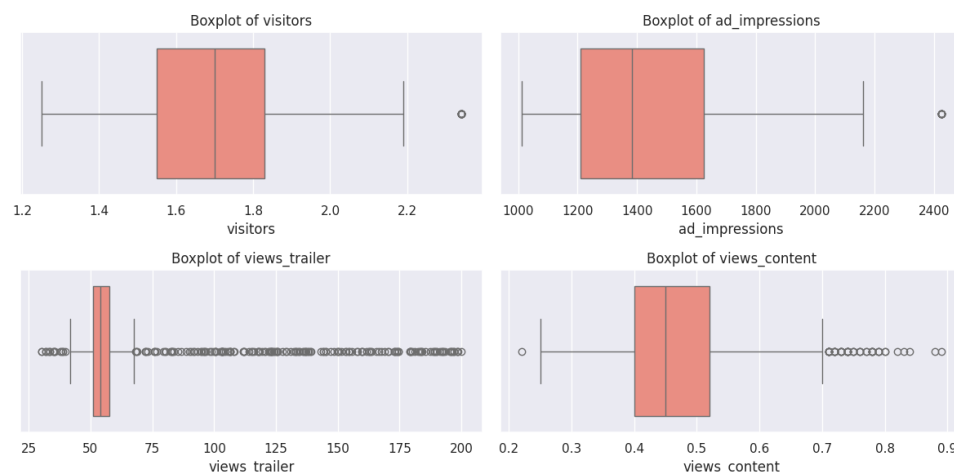
Fig. 2.1

views_trailer and views_content show the presence of some right-skewed outliers. These values are not extreme enough to significantly distort the distribution. Most views_content values lie within ±3 standard deviations. There are few mild outliers, mostly on the higher end (e.g., > 0.75 million views). These points may reflect popular releases or marketing success. Since the dataset is reasonably clean and linear regression is robust to a few outliers, we choose not to remove them.

Feature Engineering

1) The z_score column was added earlier for outlier detection but is derived from the target variable and **not a genuine predictor**, so it was **dropped** before modeling.

2) The categorical columns genre, dayofweek, and season were **converted to numeric format using one-hot encoding**.

To avoid the **dummy variable trap**, we used drop_first=True.

This resulted in the creation of new binary columns like: `genre_Thriller`, `dayofweek_Sunday`, `season_Winter`, etc.

```
   visitors  ad_impressions  major_sports_event  views_trailer  genre_Comedy  \
0      1.67         1113.81                 0.0          56.70           0.0
1      1.46         1498.41                 1.0          52.69           0.0
2      1.47         1079.19                 1.0          48.74           0.0
3      1.85         1342.77                 1.0          49.81           0.0
4      1.46         1498.41                 0.0          55.83           0.0

   genre_Drama  genre_Horror  genre_Others  genre_Romance  genre_Sci-Fi  \
0          0.0           1.0           0.0            0.0           0.0
1          0.0           0.0           0.0            0.0           0.0
2          0.0           0.0           0.0            0.0           0.0
3          0.0           0.0           0.0            0.0           1.0
4          0.0           0.0           0.0            0.0           1.0

   genre_Thriller  dayofweek_Monday  dayofweek_Saturday  dayofweek_Sunday  \
0             0.0               0.0                 0.0               0.0
1             1.0               0.0                 0.0               0.0
2             1.0               0.0                 0.0               0.0
3             0.0               0.0                 0.0               0.0
4             0.0               0.0                 0.0               1.0

   dayofweek_Thursday  dayofweek_Tuesday  dayofweek_Wednesday  season_Spring  \
0                 0.0                0.0                  1.0            1.0
1                 0.0                0.0                  0.0            0.0
2                 0.0                0.0                  1.0            0.0
3                 0.0                0.0                  0.0            0.0
4                 0.0                0.0                  0.0            0.0

   season_Summer  season_Winter
0            0.0            0.0
1            0.0            0.0
2            0.0            0.0
3            0.0            0.0
4            0.0            1.0
0    0.51
1    0.32
2    0.39
3    0.44
4    0.46
Name: views_content, dtype: float64
```

3) Defined Feature Matrix (X) and Target Variable (y)

- The **target variable** is views_content.
- All other variables form the **feature matrix**, and we converted them to float for compatibility with statsmodels:

Data preparation for Modeling:

1) Train-Test Split:

The dataset was split into training and testing sets using a **70-30 split**.

2) Added Constant for Intercept Term (Statsmodels requirement)

- A constant column was added to the training data so that the regression model includes an intercept.

```
     visitors  ad_impressions  major_sports_event  views_trailer  genre_Comedy  \
0        1.67         1113.81                 0.0          56.70           0.0
1        1.46         1498.41                 1.0          52.69           0.0
2        1.47         1079.19                 1.0          48.74           0.0
3        1.85         1342.77                 1.0          49.81           0.0
4        1.46         1498.41                 0.0          55.83           0.0

   genre_Drama  genre_Horror  genre_Others  genre_Romance  genre_Sci-Fi  \
0          0.0           1.0           0.0            0.0           0.0
1          0.0           0.0           0.0            0.0           0.0
2          0.0           0.0           0.0            0.0           0.0
3          0.0           0.0           0.0            0.0           1.0
4          0.0           0.0           0.0            0.0           1.0

   genre_Thriller  dayofweek_Monday  dayofweek_Saturday  dayofweek_Sunday  \
0             0.0               0.0                 0.0               0.0
1             1.0               0.0                 0.0               0.0
2             1.0               0.0                 0.0               0.0
3             0.0               0.0                 0.0               0.0
4             0.0               0.0                 0.0               1.0

   dayofweek_Thursday  dayofweek_Tuesday  dayofweek_Wednesday  season_Spring  \
0                 0.0                0.0                  1.0            1.0
1                 0.0                0.0                  0.0            0.0
2                 0.0                0.0                  1.0            0.0
3                 0.0                0.0                  0.0            0.0
4                 0.0                0.0                  0.0            0.0

   season_Summer  season_Winter
0            0.0            0.0
1            0.0            0.0
2            0.0            0.0
3            0.0            0.0
4            0.0            1.0
0    0.51
1    0.32
2    0.39
3    0.44
4    0.46
Name: views_content, dtype: float64
```

x

y

Training data (x_train, y_train)

Model building- Linear Regression

We defined:

- **Target variable (y)**: views_content (first-day content views in millions)

- **Independent variables (X)**: All other numeric and dummy-encoded features including:

    o Platform metrics: visitors, ad_impressions, views_trailer

    o Event indicator: major_sports_event

    o Encoded categorical variables: genre, dayofweek, season

Regression Model building

To identify the key factors driving first-day content viewership on the ShowTime OTT platform, I developed a **Linear Regression model using the Ordinary Least Squares (OLS)** method from the statsmodels library. This method estimates the relationship between the target variable and several predictors by minimizing the sum of squared residuals.

```
                          OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.784
Method:                 Least Squares   F-statistic:                     128.1
Date:                Sat, 12 Jul 2025   Prob (F-statistic):           7.57e-215
Time:                        15:01:28   Log-Likelihood:                 1125.0
No. Observations:                 700   AIC:                            -2208.
Df Residuals:                     679   BIC:                            -2112.
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
```

1) R-squared:

- This means that **79.1% of the variance in the target variable (views_content) is explained by the independent variables** in our model.
- It indicates a **very good fit** — our model is capturing most of the trends and patterns in the data.

2) Adjusted R-squared:

- Adjusted R-squared accounts for the **number of predictors** used in the model.
- It slightly penalizes the addition of unnecessary variables (those that don't improve the model).
- Still **very close to R-squared (0.784 vs. 0.791)**, so most of the included predictors are useful and relevant.

3) F-statistic:

- The F-statistic tests whether our model provides a better fit than a model with no predictors (i.e., intercept-only model).
- A **high F-statistic with a very low p-value (< 0.05)** means our model is **statistically significant overall**.
- In simple terms: **at least one predictor is significantly associated with the target** variable.

Variable	Coefficient	p-value	Significance
visitors	0.1320	0.000	Strong positive influence
views_trailer	0.0023	0.000	Strong positive influence
major_sports_event	-0.0620	0.000	Significant negative impact
dayofweek_Saturday	0.0572	0.000	Higher viewership for Saturday releases
season_Summer	0.0474	0.000	Seasonal impact is strong

The variables like ad_impressions, genre_Comedy, genre_Romance, etc., have **higher p-values** ($p > 0.05$), suggesting they are **not statistically significant** individually.

Significant Predictors from OLS Regression

- **visitors**: Strong positive impact on content views ($p < 0.001$)

- **major_sports_event**: Negative impact; content released during major events tends to perform worse ($p < 0.001$)

- **views_trailer**: Strong positive relationship with content views ($p < 0.001$)

- **genre_Drama**: Slight positive effect ($p = 0.039$)

- **genre_Thriller**: Moderate positive effect ($p = 0.013$)

- **dayofweek_Monday**, **Saturday**, **Sunday**, **Thursday**, **Tuesday**, **Wednesday**: All show statistically significant positive effects on viewership ($p < 0.05$)

- **season_Spring**, **Summer**, **Winter**: Seasonal timing significantly affects viewership ($p < 0.001$)

These features are the most relevant drivers of first-day content viewership and will be used for further analysis and recommendations.

Testing the assumptions of linear regression model

Before interpreting the results of a linear regression model, it is essential to verify that its key assumptions hold true. Violations of these assumptions can lead to biased or inefficient estimates, reducing the model's reliability and predictive power.

In this section, we test the following standard assumptions of linear regression:

- **No Multicollinearity:** Independent variables should not be highly correlated with each other.

- **Linearity of Variables:** There should be a linear relationship between independent variables and the target variable.

- **Independence of Error Terms:** Residuals (errors) should be independent of each other.

- **Normality of Error Terms:** Residuals should be normally distributed.

- **Homoscedasticity (Constant Variance):** The variance of residuals should remain constant across all levels of the independent variables.

Test for Multicollinearity

```
              feature         VIF
0               const  102.778999
8        genre_Others    2.721951
11     genre_Thriller    1.932870
5        genre_Comedy    1.907395
6         genre_Drama    1.896472
9       genre_Romance    1.871581
10       genre_Sci-Fi    1.859077
7        genre_Horror    1.856066
19      season_Summer    1.564355
20      season_Winter    1.547323
18      season_Spring    1.513794
17  dayofweek_Wednesday   1.286838
15   dayofweek_Thursday   1.152524
13   dayofweek_Saturday   1.140913
14     dayofweek_Sunday   1.122056
16    dayofweek_Tuesday   1.052625
12     dayofweek_Monday   1.050543
3     major_sports_event   1.040329
4        views_trailer    1.018468
1             visitors    1.016457
2        ad_impressions    1.016221
```

Interpretation:

- All VIF values are well below the common threshold of 5.

- This indicates **no significant multicollinearity** exists among the predictor variables.

- The constant term (const) often shows a high VIF and is not of concern in this context.

The assumption of **No Multicollinearity** is satisfied, allowing us to interpret the model coefficients without concern for redundant information among predictors.

Test for Linearity and Independence
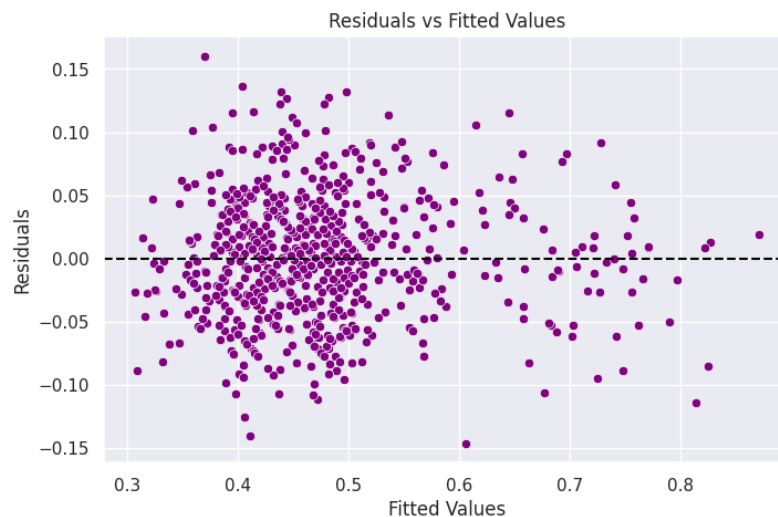
Fig. 3.1

Residuals vs. Fitted Values Plot

This diagnostic plot helps evaluate two key assumptions of linear regression:

- **Linearity**: The relationship between predictors and the outcome should be linear. This is tested by checking whether residuals are centered around zero without any distinct curve.

- **Independence of Errors**: Residuals should not display any specific pattern (e.g., clusters, trends, or sequences).

Interpretation:

- The residuals appear fairly **randomly scattered** around the horizontal line at 0.

- There is **no obvious curvature** or "funnel shape" (which would suggest non-linearity or heteroscedasticity).

- No grouping or sequence of residuals is visible, implying **independence** of the error terms.

The assumptions of **linearity** and **independence of residuals** are reasonably met. This supports the validity of using a linear regression model for the given data.
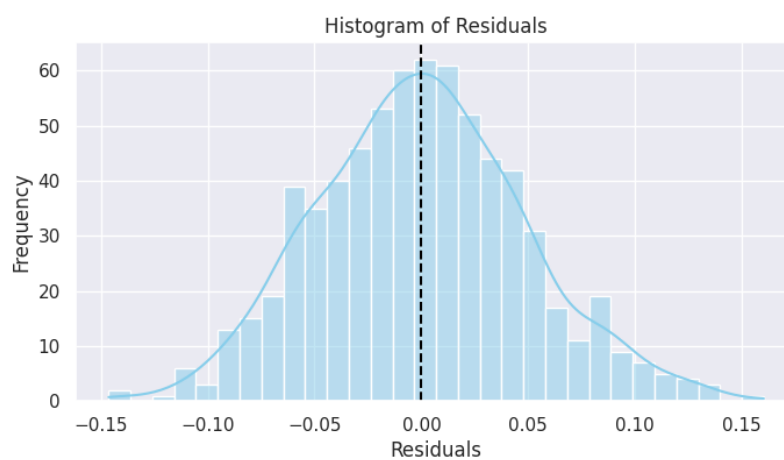
Test for Normality of Residuals

To assess whether the residuals follow a normal distribution (a key assumption of linear regression), we use:

a) Histogram of Residuals

b) Q-Q (Quantile-Quantile) Plot
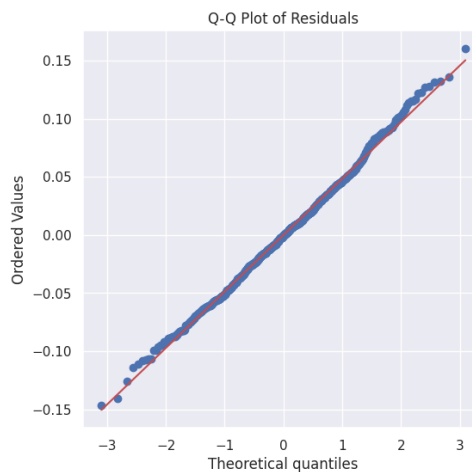
Interpretation:

- **Histogram**:

Fig. 3.2

The distribution of residuals appears roughly symmetric and bell-shaped, centered around zero. This suggests that the residuals are approximately normally distributed.

- **Q-Q Plot**:

Q-Q Plot of Residuals

Fig. 3.3

Most of the residuals fall along the 45-degree reference line. There is only a slight deviation at the tails, which is acceptable in practical scenarios.

The residuals satisfy the **normality assumption** reasonably well. This supports the reliability of statistical inference (e.g., p-values and confidence intervals) from the regression model.

Test for Homoscedasticity (Equal Variance of Residuals)

To verify whether the residuals exhibit constant variance (homoscedasticity), we conducted the **Goldfeld-Quandt Test**.

Test Summary:

- **F-statistic**: 1.0068

- **p-value**: 0.4755

Interpretation:

- Since the p-value is **greater than 0.05**, we **fail to reject** the null hypothesis.

- The null hypothesis assumes that the residuals have **equal variance** (homoscedasticity).

- Therefore, **no significant evidence of heteroscedasticity** is present in the model.

The assumption of **homoscedasticity holds true**, supporting the validity of model estimations and ensuring consistent standard error estimates across predictors.

Model Performance Evaluation

After training the linear regression model on **70% of the data**, we evaluated its performance on the **test set (30%)** using standard regression metrics.

```
Model Performance on Test Set:

R-squared       : 0.764
Mean Absolute Error (MAE): 0.041
Mean Squared Error (MSE): 0.003
Root MSE (RMSE) : 0.051
```

Model Performance on Test Set:

- R-squared value: Indicates that **76.4% of the variance** in content viewership is explained by the model on unseen data.

- MAE (Mean Absolute Error): On average, the predicted views deviate from the actual views by 0.041 units.

- MSE (Mean Squared Error): The average of the squared differences is 0.003, showing a low overall error.

- RMSE (Root Mean Squared Error): The square root of MSE is 0.051, providing an interpretable error metric in original units.

Overall, the model generalizes well on the test data with high R-squared and low prediction errors.

```
       Actual   Predicted
521     0.42    0.395047
737     0.59    0.550833
740     0.32    0.414984
660     0.69    0.722816
411     0.28    0.325739
678     0.51    0.475848
626     0.35    0.419176
513     0.43    0.462366
859     0.45    0.453925
136     0.45    0.500262
```

We can observe here that our model has returned pretty good prediction results, and the actual and predicted values are comparable.

Actionable Insights & Recommendations

Based on the linear regression model and the statistical significance of predictors, we derive the following insights and business recommendations for **ShowTime**:

Predictor	Coefficient	Impact on Views	Significance
views_trailer	+0.0023	Strong positive influence	Highly significant
visitors	+0.1320	More platform traffic → more content views	Highly significant
major_sports_event	−0.0620	Reduces viewership	Highly significant
dayofweek_Saturday	+0.0572	Saturday releases perform better	Significant
season_Summer	+0.0474	Summer boosts content views	Significant

Key Takeaways for the Business:

- **Boost Trailer Marketing:**
  Trailer views are the strongest predictor of first-day content views. Invest more in **pre-release promotions** and trailer visibility to build hype and drive interest.

- **Increase Platform Traffic:**
  Weekly visitor count directly influences content views. Focus on **retention strategies, personalized notifications**, and exclusive content to boost returning user engagement.

- **Avoid Releases on Major Sports Days:**
  Major sports events negatively impact viewership. Plan releases away from significant matches or tournaments to **avoid audience competition**.

- **Leverage Weekends (Especially Saturdays):**
  Content released on **Saturdays and Sundays** shows significantly better performance. Prioritize premium content for **weekend drops**.

- **Utilize Seasonal Trends:**
  Summer releases outperform others. Consider launching **flagship shows or high-budget films during summer** for better reach.