Predictive Modeling Project Report

Customer data analysis - Show time over-the-top (OTT) media service

Prepared by
Kanimozhi S

CONTENTS

List of Tables

List of Figures

Problem Statement

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 on 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.

This project aimed to identify key drivers influencing **first-day content viewership** on the ShowTime OTT platform using linear regression.

Objectives:

- Analyze platform data to uncover patterns in content performance.
- Build a predictive model to estimate first-day viewership.
- Provide actionable business recommendations based on model insights.

Data Overview

The dataset has 8 variables The dataset contains both numerical and categorical variables:

Although variables like `genre`, `season`, and `dayofweek` are stored as strings or may be encoded numerically (e.g., Sunday = 0, Monday = 1, etc.), they are treated as **categorical variables** in this analysis. This is because these values represent distinct, qualitative groups rather than measurable quantities or continuous scales.

Similarly, `major_sports_event` is a binary indicator (Yes/No or 1/0) and is also treated as a **categorical variable** since it denotes the presence or absence of a condition, not a numerical measure.

On the other hand, variables such as `views_content`, `views_trailer`, `visitors`, and `ad_impressions` represent **measurable quantities** on a continuous scale and are therefore treated as **continuous variables**.

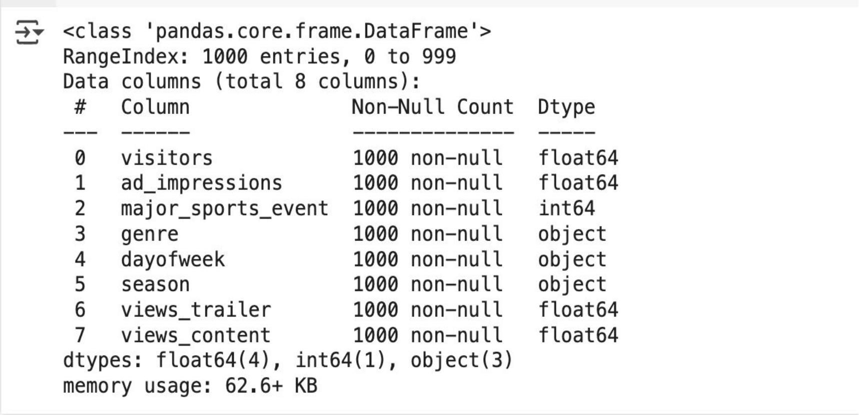
Hence, in this analysis:

- `views_content` is the continuous **response variable**.
- `views_trailer`, `visitors`, and `ad_impressions` are continuous **predictor variables**.
- The remaining variables (`genre`, `season`, `dayofweek`, and `major_sports_event`) are treated as **categorical variables**.

DATA DICTIONARY

Variable	Description
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

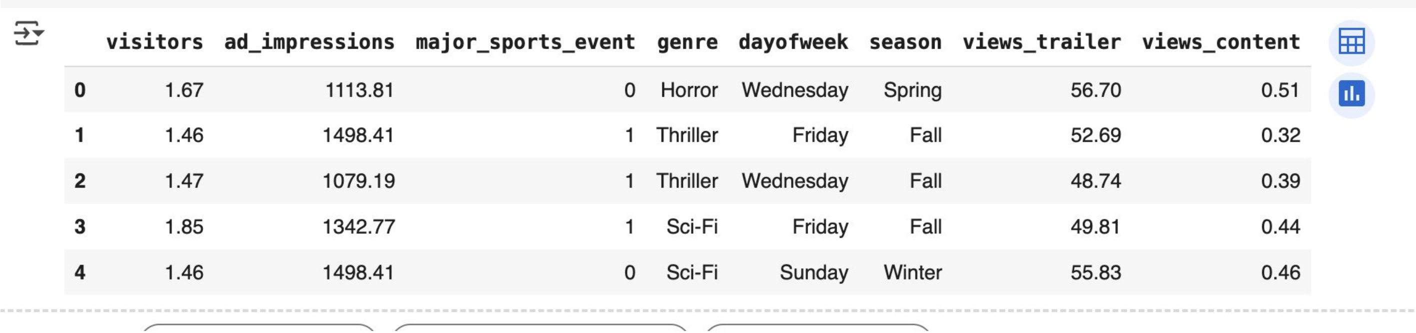
Table 1 : Data dictionary



```
<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
```

Table 2: Data info

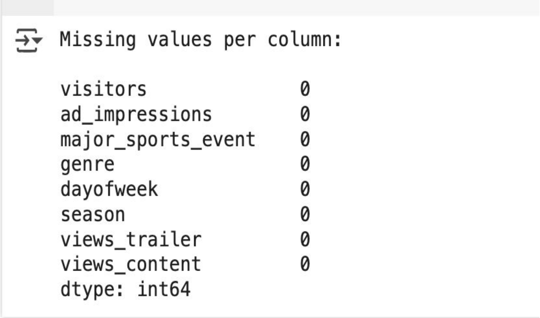
There are 1000 rows and 8 columns and first five rows of the datasets are below



	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

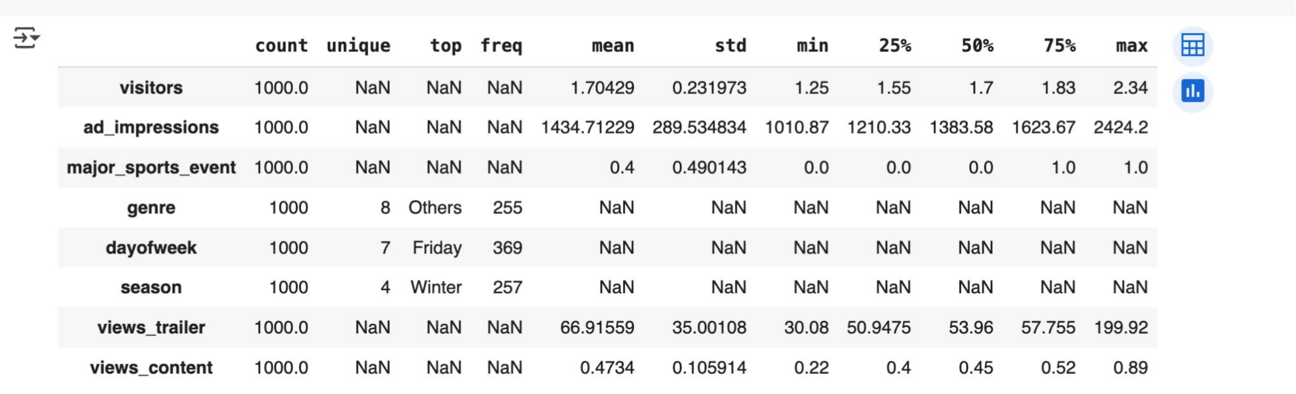
Table 3: First 5 rows of the dataset

No missing or null values in the columns



```
Missing values per column:

visitors             0
ad_impressions       0
major_sports_event   0
genre                0
dayofweek            0
season               0
views_trailer        0
views_content        0
dtype: int64
```

Table 4 : Missing values data



	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

Table 5: Statistical summary of the data

Statistical summary:

The statistical summary of the dataset includes both continuous and categorical variables, each offering distinct insights into content performance on the ShowTime platform.

The **response variable**, `views_content`, ranges from 0.22 to 0.89, with a mean of 0.47 and a standard deviation of 0.11. The distribution is right-skewed, indicating that while most content receives moderate attention, a few releases achieve disproportionately high viewership, likely due to strategic timing, popular cast, or viral appeal. Similarly, `views_trailer` displays considerable variation, ranging from 30 to 199 with a mean of 66.9. This suggests differing levels of trailer engagement, possibly linked to content genre, campaign intensity, or audience interest.

The variable `ad_impressions` spans from 1,010 to over 2,400, with a mean of 1,434.7 and a standard deviation of 289.5, pointing to significant variation in promotional investment across content. In contrast, `visitors` is relatively stable (mean: 1.70, std: 0.23), implying consistent daily platform traffic regardless of specific content releases. These continuous variables provide strong quantitative foundations for predicting content viewership.

Among **categorical variables**, `genre` includes 8 unique categories, with "Others" being the most frequent (n = 255), indicating a broad content classification that may require closer attention for refinement. The `dayofweek` variable includes 7 categories, with **Friday** being the most common release day (n = 369), suggesting a deliberate weekend-centric release strategy. The `season` variable contains 4 levels, with **Winter** dominating (n = 257), likely reflecting content scheduling around peak viewing periods. Finally, `major_sports_event` is a binary variable (0 = No, 1 = Yes), with approximately 40% of the data associated with days when a major sports event occurred. This feature may influence viewership patterns either positively or negatively depending on audience overlap.

Outlier Detection

An outlier analysis was conducted on the continuous variables using the IQR method to identify extreme values that may influence model performance. The findings are summarized below:

- **visitors**: 20 outliers were detected. These may correspond to days when unusually high traffic was observed on the platform, possibly due to special releases or targeted campaigns.
- **ad_impressions**: 13 outliers were identified. These instances likely reflect content with exceptionally high advertising investment, potentially driven by high-budget productions or platform-wide promotions.
- **views_trailer**: 189 outliers were flagged the highest among all variables. This suggests a long-tailed distribution where certain content received disproportionately high trailer engagement, possibly due to high anticipation, celebrity appeal, or viral marketing.
- **views_content** *(Response Variable)*: 47 outliers were observed in the first-day viewership. These represent content that significantly outperformed the average and may correspond to high-visibility launches, franchise content, or heavily marketed originals.

Although these values are extreme relative to the typical range, they are **not considered anomalies**. All outliers were retained ir the dataset, as they reflect legitimate business outcomes and provide valuable information for understanding what drives exceptional content performance. Their potential influence on the linear regression model will be assessed during assumption testing (e.g., residual analysis and leverage statistics).

Conclusion: While several outliers were detected across variables, especially in trailer views and content views, these values are valid and meaningful. They are indicative of real-world spikes in performance and marketing effort, and therefore retained for analysis. Their impact will be closely monitored in subsequent modeling steps.

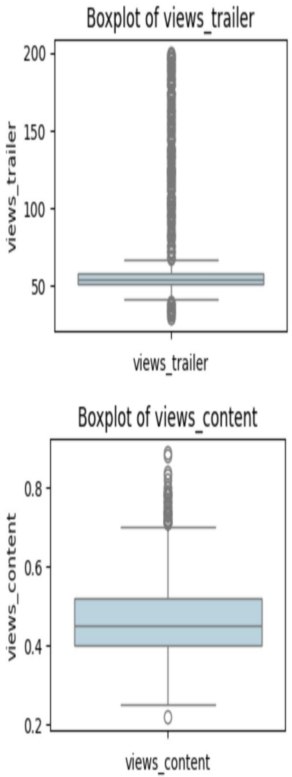


Fig 1 : Boxplot for outliers

EDA - UNIVARIATE ANALYSIS

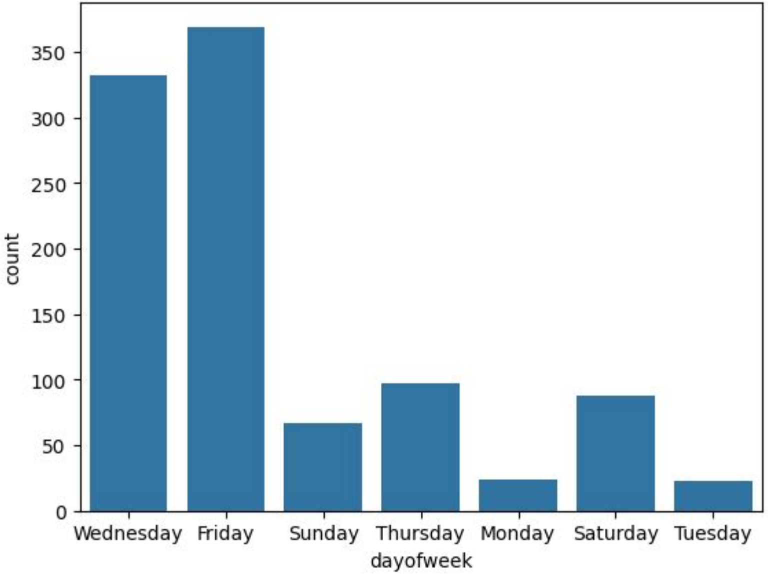


Fig 2: Day of week plot

A countplot was used to analyze how content is distributed across days of the week. The distribution shows that **Friday and Wednesday** have the highest number of content releases, suggesting a **weekend release strategy** to maximize viewership. Weekdays, particularly **Monday and Tuesday**, see fewer releases. This reflects an intentional scheduling pattern aimed at capitalizing on peak user engagement during weekends.

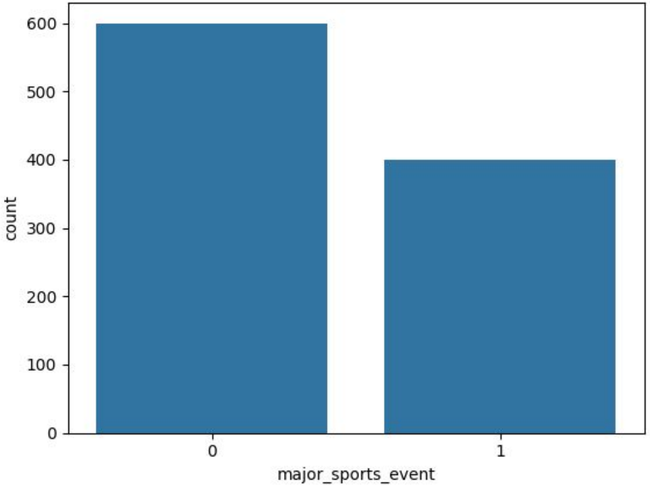


Fig 3: sports event distribution plot

The binary plot shows a reasonable balance between content released **with** and **without** major sports events occurring on the same day. A slightly higher count of content is released when **no major event** is scheduled, possibly to avoid competition. However, the presence of several releases during sports events also suggests the platform occasionally rides on increased overall digital engagement.

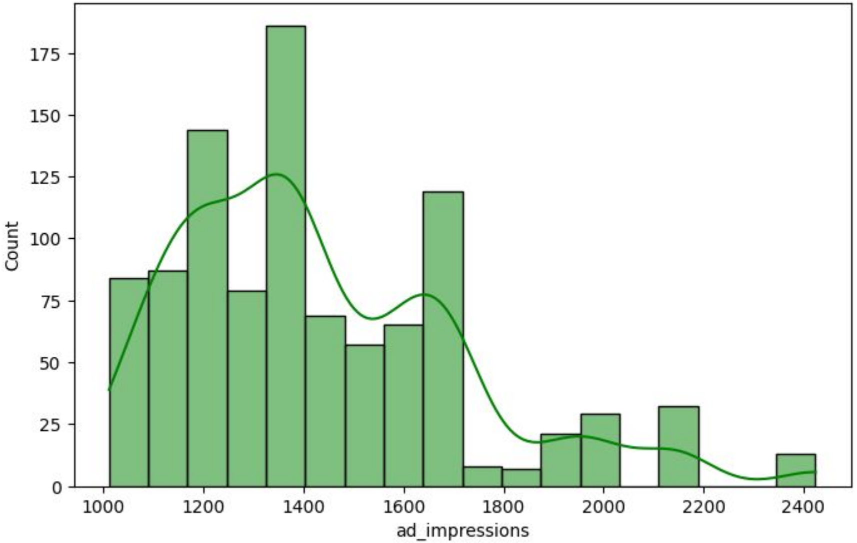


Fig 4: ad-impressions distribution plot

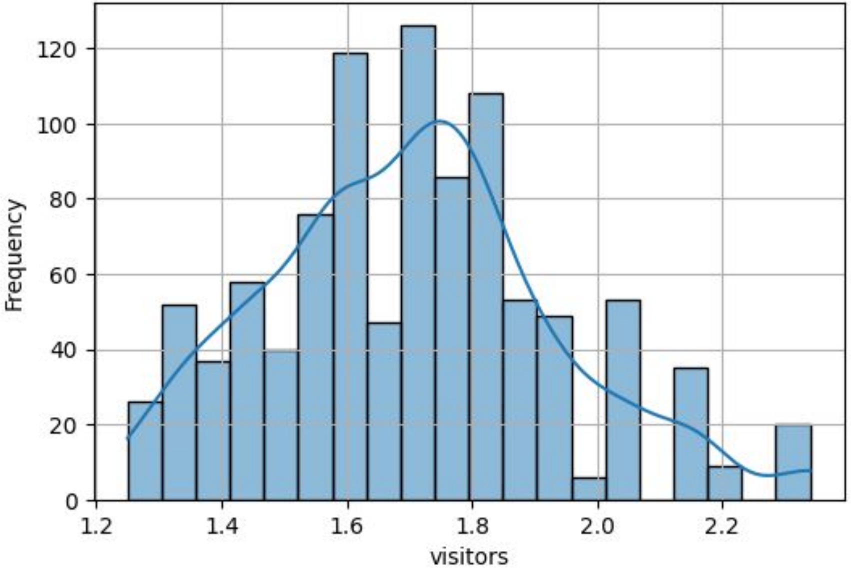


Fig 5: Visitors distribution plot

The histogram of ad impressions displays a **right-skewed distribution**, indicating that while most content receives a moderate volume of advertising exposure, a few content pieces are heavily promoted. These outliers may correspond to flagship content or premium releases. The variability in ad spend suggests a **tiered promotion strategy**, which could influence viewership outcomes.

The histogram for visitors shows a **moderately right-skewed** distribution with most values concentrated around the mean. This indicates relatively **consistent traffic**, with occasional spikes that may be associated with promotional campaigns or content launches. These high-traffic days could be key drivers of success for newly released content.

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

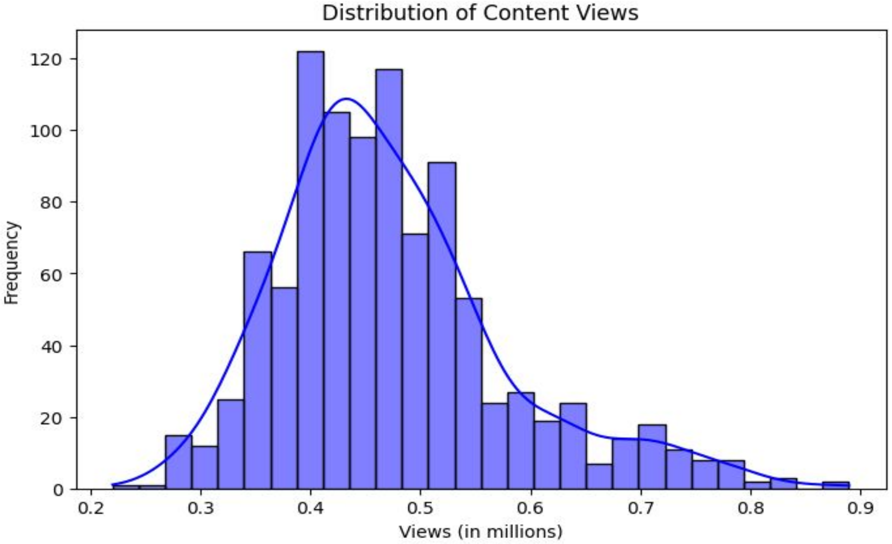


Fig 6: Problem 1 plot

The distribution of first-day content views is **slightly right-skewed**, with most titles receiving between **0.3 and 0.6 million views**, while a smaller group of content crosses the **0.7 million mark**, indicating **exceptionally high-performing releases**. This long-tail behavior is typical of digital content platforms, where a few standout titles contribute disproportionately to total viewership. The majority of content achieves **moderate engagement**, but it is these few outlier successes that **significantly drive platform traffic**, making them essential to analyze and replicate through targeted strategies.

2.What does the distribution of genres look like?

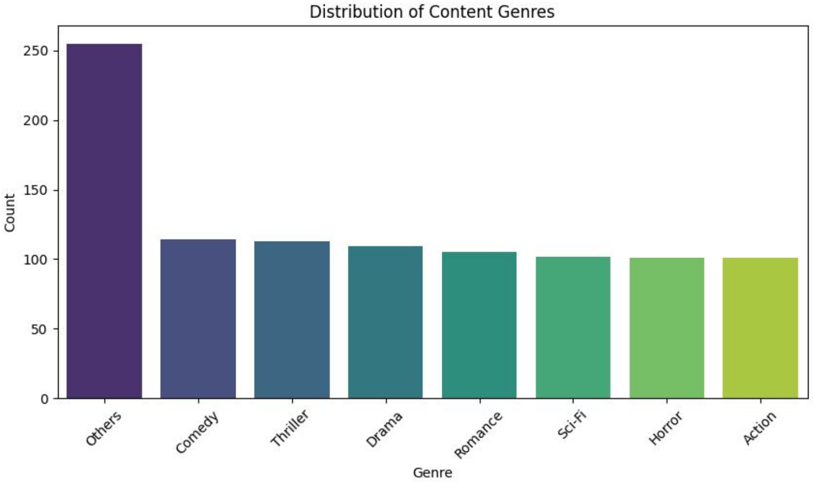


Fig 7: Problem 2 plot

The distribution of content genres is **highly skewed**, with the **"Others"** category accounting for over **25% of the releases**, significantly more than any specific genre. Mainstream genres such as **Comedy**, **Thriller**, **Drama**, and **Romance** are evenly represented, each comprising around **10–12%** of the content. Genres like **Sci-Fi**, **Horror**, and **Action** appear slightly less frequently but still show a balanced presence.

BIVARIATE ANALYSIS

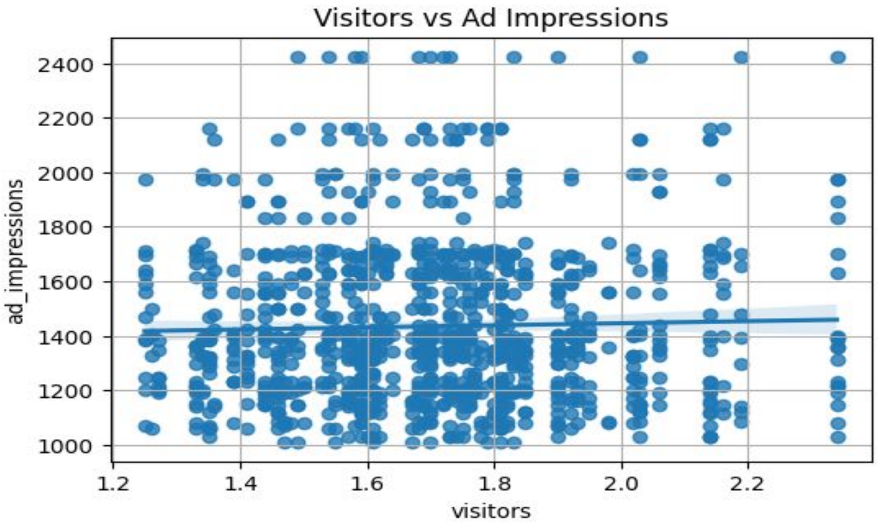


Fig 8: Visitors vs ad impressions plot

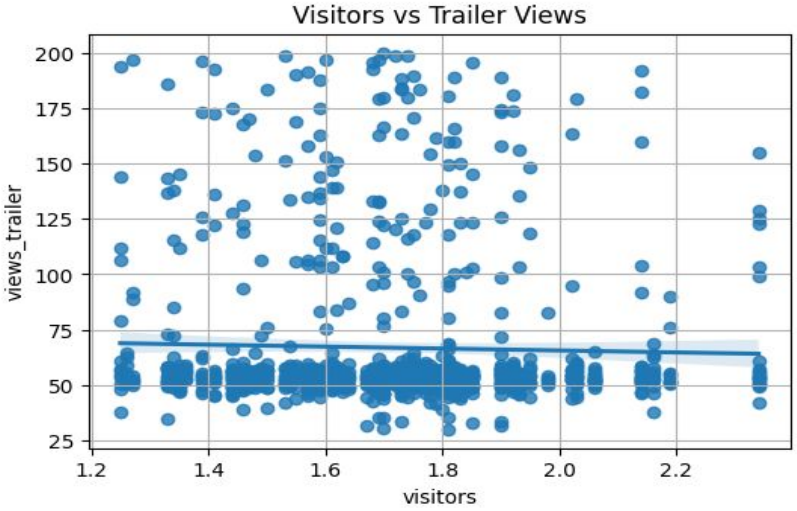


Fig 9: Visitors vs trailer plot

The scatterplot shows a moderate positive relationship between `visitors` and `ad_impressions`, suggesting that higher platform traffic generally leads to more ad exposure. However, the wide spread indicates that **ad allocation is also influenced by marketing strategy**, not solely by visitor volume.

The scatterplot shows a weak-to-moderate positive relationship between `visitors` and `views_trailer`, indicating that trailer views tend to increase slightly with platform traffic. However, the spread is wide, suggesting that **trailer engagement is influenced more by external promotion and content appeal than just visitor count**.

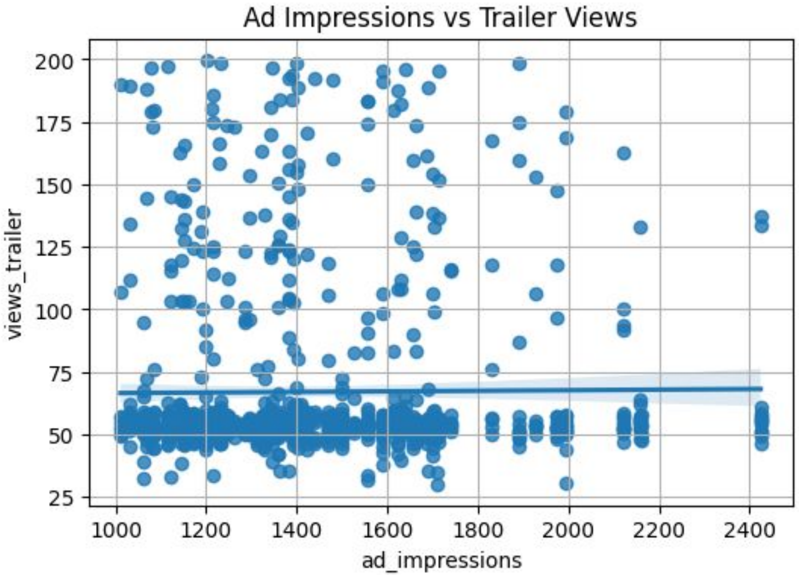


Fig 10: Ad vs Trailer views distribution plot

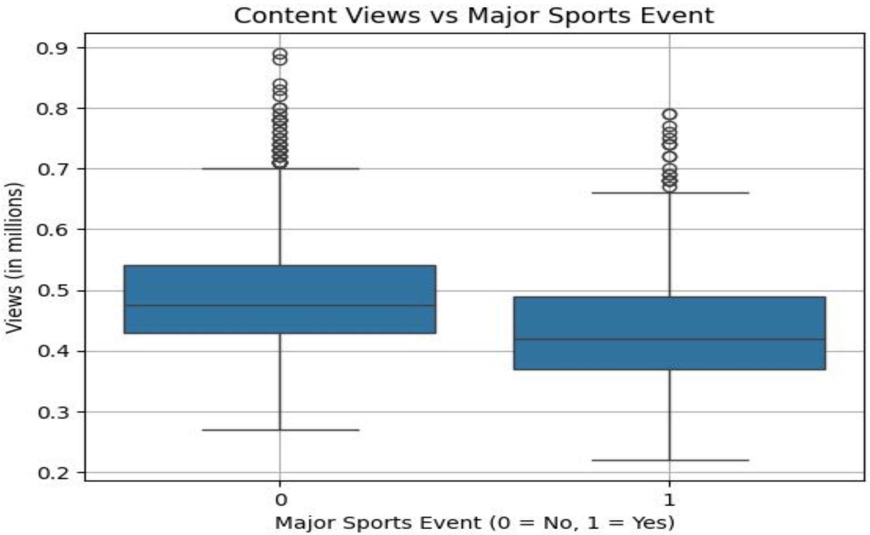


Fig 11: content view vs major sports event distribution plot

The plot shows a clear positive relationship between `ad_impressions` and `views_trailer`, indicating that higher ad exposure is associated with increased trailer engagement. This suggests that **paid advertising is an effective driver of trailer performance**, and optimizing ad spend can help boost pre-release visibility and audience interest.

The plot comparing `major_sports_event` (binary) with `views_content` shows a **slight negative trend**, indicating that content released **on days with major sports events** tends to have **marginally lower first-day viewership**. The distribution is wide in both categories, but the average appears slightly lower during event days. This suggests that sports events may create **audience competition**, slightly reducing attention for OTT content.

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?

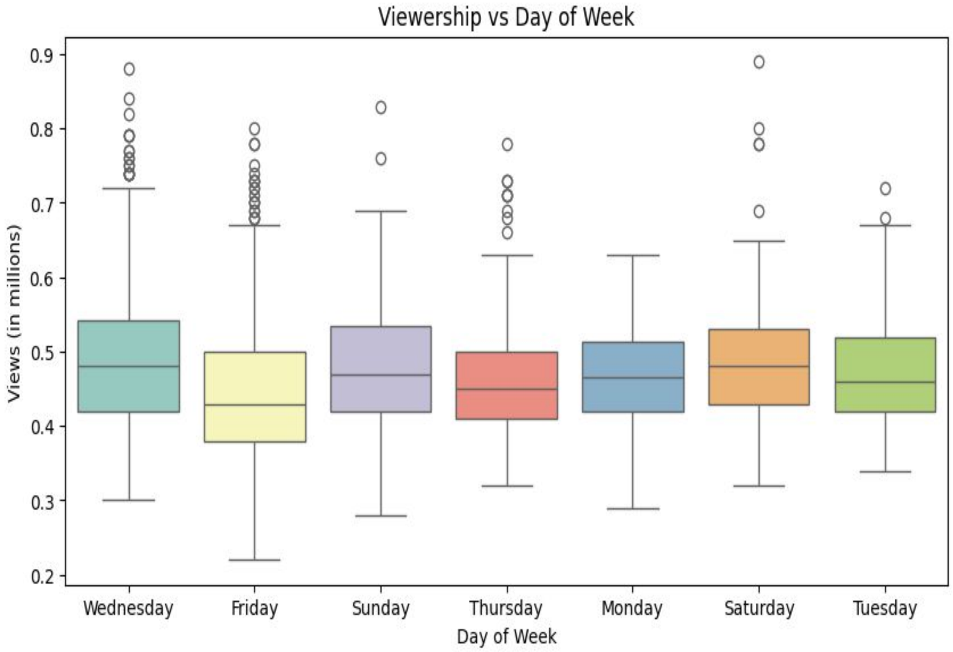


Fig 12: Problem 3 plot

The boxplot shows that viewership remains **fairly stable across the week**, but with **slightly higher medians** and **more frequent high outliers** on **Wednesdays, Sundays, and Saturdays**. Notably, content released on **Wednesdays** tends to have the **highest median viewership**, while **Friday** releases show a wider spread with several high-performing outliers. Weekdays such as **Monday and Tuesday** exhibit narrower interquartile ranges and fewer peaks, suggesting lower variability and slightly more consistent performance.

Insight:

Although content is most frequently released on Fridays and Saturdays, the plot indicates that **midweek and weekend releases—especially Wednesdays and Sundays—can deliver stronger performance**, both in terms of median views and peak outliers. This suggests that **strategic midweek releases** may offer an opportunity to stand out with less competition while still capturing high user engagement.

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

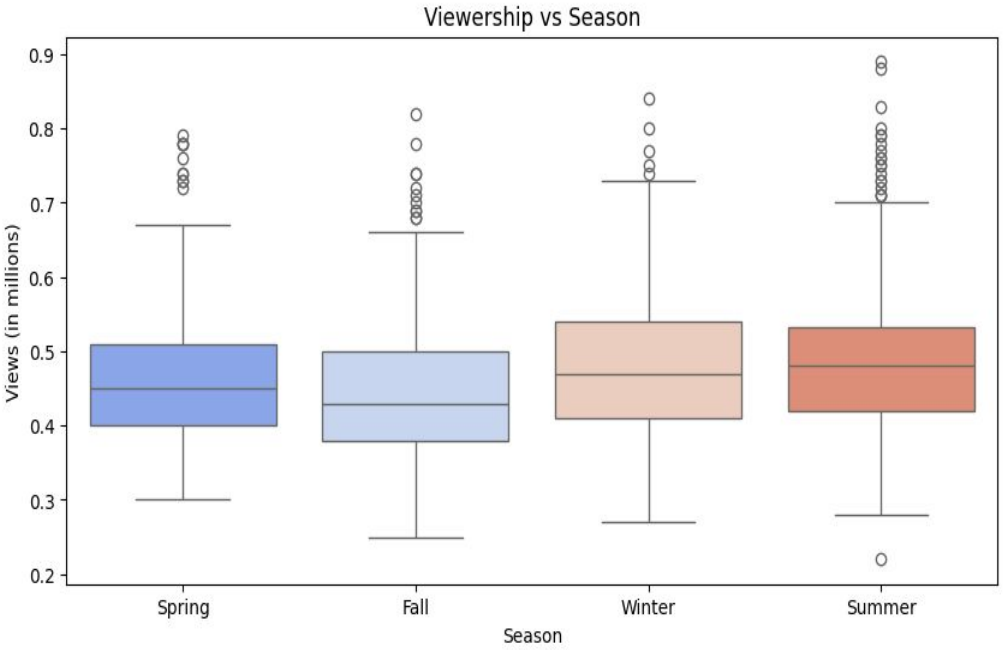


Fig 13: Problem 4 plot

The boxplot reveals that **Summer and Winter** releases tend to have **higher median viewership** compared to Spring and Fall. Specifically, **Summer** shows a slightly higher median and a greater concentration of **top-performing content**, with several outliers reaching above 0.8 million views. **Winter** follows closely, reflecting steady performance and a balanced spread. In contrast, **Fall** displays the lowest median, with fewer high-viewership outliers, suggesting relatively modest performance during that season.

Insight:

Content released in **Summer and Winter performs better on average**, likely due to **seasonal engagement peaks** such as school holidays, vacations, or festive breaks. These periods may represent strategic windows for launching high-investment content. Conversely, **Fall appears to be a low-engagement season**, and content scheduled during this period may require additional promotional efforts to achieve competitive viewership.

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

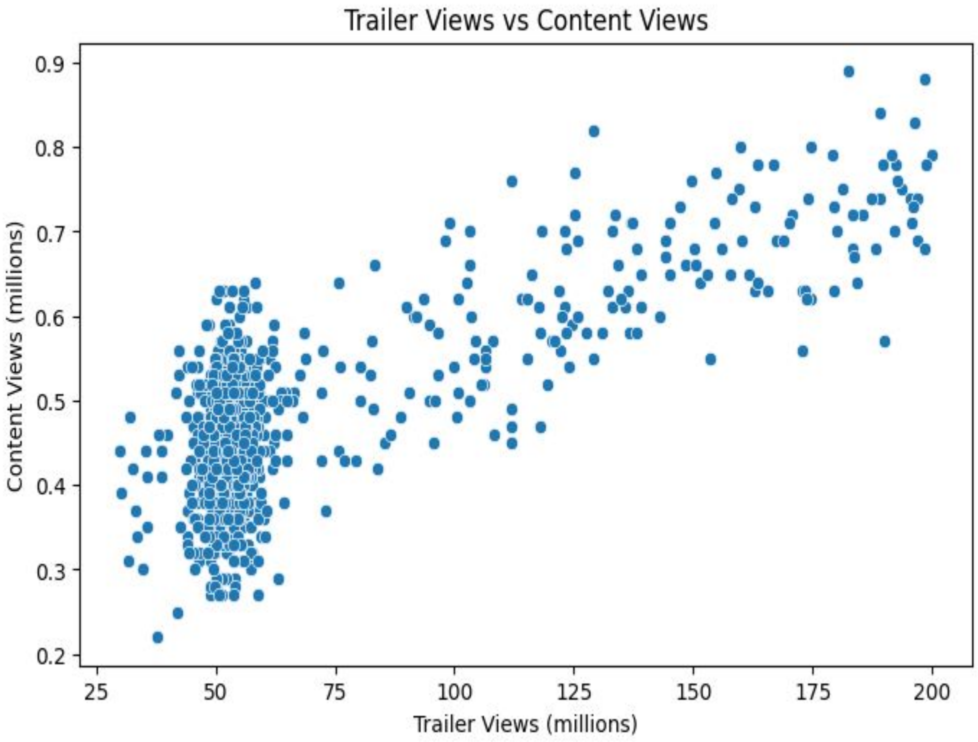


Fig 14: Problem 5 plot

The scatterplot demonstrates a **strong positive relationship** between `views_trailer` and `views_content`. As trailer views increase, the corresponding content tends to receive higher first-day viewership. While most data points cluster around 50–75 million trailer views, there is a **clear upward trend**, with content receiving over **150 million trailer views consistently achieving high content views (>0.6 million)**. The spread tightens at higher trailer counts, reinforcing the predictive strength of trailer engagement.

Insight:

Trailer views act as a **reliable leading indicator of content performance**, suggesting that **pre-release marketing and hype generation are key drivers of first-day success**. Content with higher trailer traction tends to attract more initial viewers, making trailer performance a crucial metric for predicting and planning content success.

EDA INSIGHTS

The EDA reveals several meaningful insights into factors influencing first-day content viewership on the ShowTime platform. The distribution of `views_content` is right-skewed, with most titles achieving moderate engagement (0.3–0.6 million views), while a few high-performing releases surpass 0.7 million — suggesting a long-tail effect where a small subset of content drives a large share of viewership.

Key Takeaways:

- **Trailer views and ad impressions are key predictors** of content success and should be prioritized during campaign planning.
- **Release timing matters**: Wednesdays and Sundays show promising viewership potential beyond the typical Friday-Saturday strategy.
- **Genre classification needs refinement** to allow for more targeted genre-based analysis and recommendations.
- Avoiding **major sports event dates** for high-stakes content may improve visibility and reduce audience overlap.

Feature Engineering and Data Prep

Feature Engineering

To prepare the dataset for linear regression modeling, **one-hot encoding** was applied to the categorical variables: `genre`, `season`, and `dayofweek`. The `drop_first=True` parameter was used to avoid the dummy variable trap by dropping one category from each, allowing the model to treat them as baseline references.

The following reference levels were dropped:

- `genre_Action`
- `season_Fall`
- `dayofweek_Friday`

These base categories are implicitly represented in the regression model and serve as a point of comparison for interpreting coefficients of the other levels. The binary variable `major_sports_event` was already in numeric form (0/1) and used without transformation. All continuous variables (`visitors`, `views_trailer`, `ad_impressions`) were retained in their original scale.

After encoding, the dataset consisted of **21 variables**:

- **4 original predictors**: `visitors`, `ad_impressions`, `views_trailer`, and `major_sports_event`
- **16 dummy variables** from the three categorical columns:
  - Genres: 7 dummies
  - Seasons: 3 dummies
  - Days of week: 6 dummies
- **1 response variable**: `views_content`

This transformation ensured that all variables in the dataset were numerical and compatible with linear regression.

Total Predictors: 4 numerical + 16 dummies + 1 response variable = **21 predictors**

Data Splitting and Final Preparation

The target variable `views_content` was separated from the predictors and stored as `y`, while the remaining 20 columns formed the predictor matrix `X`. A **train-test split** was performed using an 80-20 ratio to create separate training and testing datasets, ensuring unbiased model evaluation.

- **Training Set Dimensions:**
  - `X_train`: 800 rows × 20 columns
  - `y_train`: 800 values

The training predictors were further augmented with a constant term using `statsmodels.api.add_constant()` to include the regression intercept.

Conclusion

The dataset was successfully encoded, cleaned, and structured for regression analysis. All variables are now numerical, and the train-test split ensures reliable model training and evaluation. The final training data includes 800 records and 20 predictor variables, making it robust enough for linear modeling.

MODEL BUILDING:

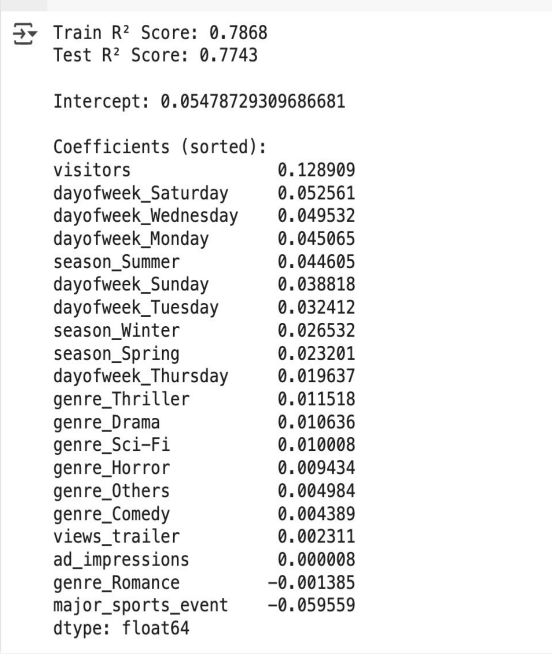
A multiple linear regression model was developed using 80% of the dataset to predict first-day content viewership. The model achieved an R² of **0.7868** on the training set and **0.7743** on the test set, indicating strong explanatory power and generalization.

The most influential predictors were:

- **Visitors**: The strongest positive driver of content views.
- **Day of release**: Content released on **Saturdays** and **Wednesdays** saw higher viewership.
- **Season**: **Summer releases** performed better.
- **Major sports events**: Associated with a significant drop in viewership.

Some predictors like **ad impressions** and **genre_Romance** had minimal impact.

Overall, the model highlights the importance of platform traffic and release timing over content genre or ad volume in influencing initial viewership.



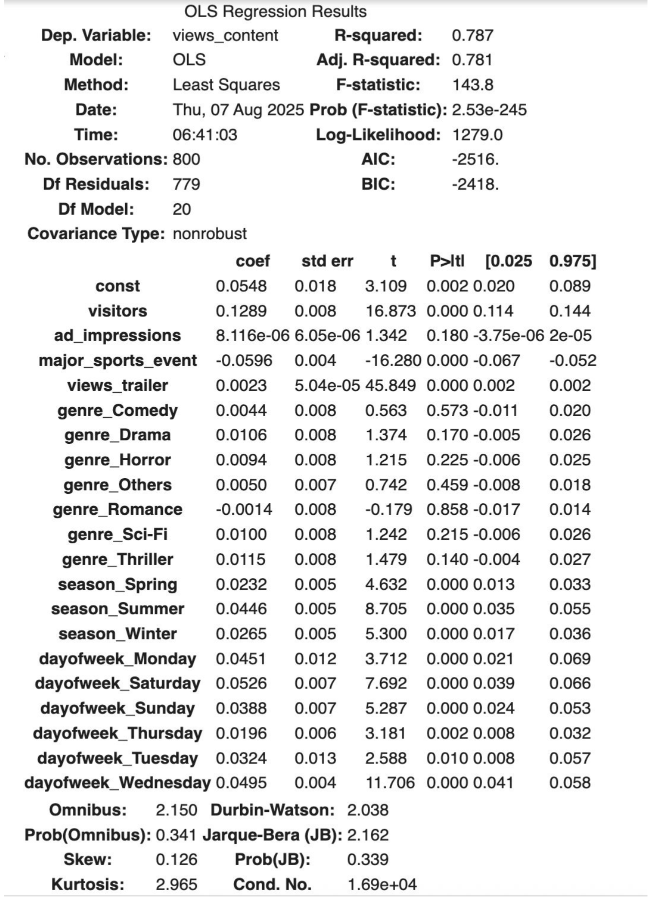
```
Train R² Score: 0.7868
Test R² Score: 0.7743

Intercept: 0.05478729309686681

Coefficients (sorted):
visitors               0.128909
dayofweek_Saturday     0.052561
dayofweek_Wednesday    0.049532
dayofweek_Monday       0.045065
season_Summer          0.044605
dayofweek_Sunday       0.038818
dayofweek_Tuesday      0.032412
season_Winter          0.026532
season_Spring          0.023201
dayofweek_Thursday     0.019637
genre_Thriller         0.011518
genre_Drama            0.010636
genre_Sci-Fi           0.010008
genre_Horror           0.009434
genre_Others           0.004984
genre_Comedy           0.004389
views_trailer          0.002311
ad_impressions         0.000008
genre_Romance         -0.001385
major_sports_event    -0.059559
dtype: float64
```

Table 6: Output for Model

Assumption testing - OLS diagnostics

The regression model was validated using OLS diagnostics. Key takeaways are:

- **Model Fit**
  - **R-squared = 0.787** → Model explains ~78.7% of the variance in viewership.
  - **Adjusted R-squared = 0.781** → Strong fit after adjusting for number of predictors.
  - **F-statistic = 143.8 (p < 0.001)** → Model is statistically significant overall.
- **Significant Predictors (p < 0.05)**
  - `visitors`, `views_trailer`, `major_sports_event`
  - Seasons: `Spring`, `Summer`, `Winter`
  - Days: `Monday`, `Saturday`, `Sunday`, `Thursday`, `Tuesday`, `Wednesday`



```
                         OLS Regression Results
     Dep. Variable:   views_content          R-squared:       0.787
             Model:   OLS                    Adj. R-squared:  0.781
            Method:   Least Squares          F-statistic:     143.8
              Date:   Thu, 07 Aug 2025 Prob (F-statistic):    2.53e-245
              Time:   06:41:03               Log-Likelihood:  1279.0
 No. Observations:    800                    AIC:             -2516.
      Df Residuals:   779                    BIC:             -2418.
          Df Model:   20
 Covariance Type:     nonrobust
                        coef      std err   t        P>|t|  [0.025    0.975]
             const      0.0548    0.018     3.109    0.002  0.020     0.089
          visitors      0.1289    0.008     16.873   0.000  0.114     0.144
    ad_impressions      8.116e-06 6.05e-06  1.342    0.180  -3.75e-06 2e-05
major_sports_event      -0.0596   0.004     -16.280  0.000  -0.067    -0.052
     views_trailer      0.0023    5.04e-05  45.849   0.000  0.002     0.002
      genre_Comedy      0.0044    0.008     0.563    0.573  -0.011    0.020
       genre_Drama      0.0106    0.008     1.374    0.170  -0.005    0.026
      genre_Horror      0.0094    0.008     1.215    0.225  -0.006    0.025
      genre_Others      0.0050    0.007     0.742    0.459  -0.008    0.018
     genre_Romance      -0.0014   0.008     -0.179   0.858  -0.017    0.014
      genre_Sci-Fi      0.0100    0.008     1.242    0.215  -0.006    0.026
    genre_Thriller      0.0115    0.008     1.479    0.140  -0.004    0.027
     season_Spring      0.0232    0.005     4.632    0.000  0.013     0.033
     season_Summer      0.0446    0.005     8.705    0.000  0.035     0.055
     season_Winter      0.0265    0.005     5.300    0.000  0.017     0.036
  dayofweek_Monday      0.0451    0.012     3.712    0.000  0.021     0.069
dayofweek_Saturday      0.0526    0.007     7.692    0.000  0.039     0.066
  dayofweek_Sunday      0.0388    0.007     5.287    0.000  0.024     0.053
dayofweek_Thursday      0.0196    0.006     3.181    0.002  0.008     0.032
 dayofweek_Tuesday      0.0324    0.013     2.588    0.010  0.008     0.057
dayofweek_Wednesday     0.0495    0.004     11.706   0.000  0.041     0.058
      Omnibus:     2.150  Durbin-Watson:     2.038
Prob(Omnibus):     0.341  Jarque-Bera (JB):  2.162
         Skew:     0.126  Prob(JB):          0.339
     Kurtosis:     2.965  Cond. No.          1.69e+04
```

Table 7: Model Diagnostics Output

- **Not Significant (p > 0.05)**
  - `ad_impressions`
  - Genres: `Comedy`, `Romance`, `Drama`, `Horror`, `Sci-Fi`, `Others`, `Thriller`
- **Residual Diagnostics**
  - **Durbin-Watson = 2.038** → No autocorrelation.
  - **Omnibus & Jarque-Bera p-values > 0.05** → Residuals are approximately normal.
  - **Skew = 0.126**, **Kurtosis = 2.965** → Supports normality.
- **Multicollinearity Flag**
  - **Condition number = 1.69e+04** → High value; possible multicollinearity.
  - Addressed further in the **VIF analysis**.

Conclusion

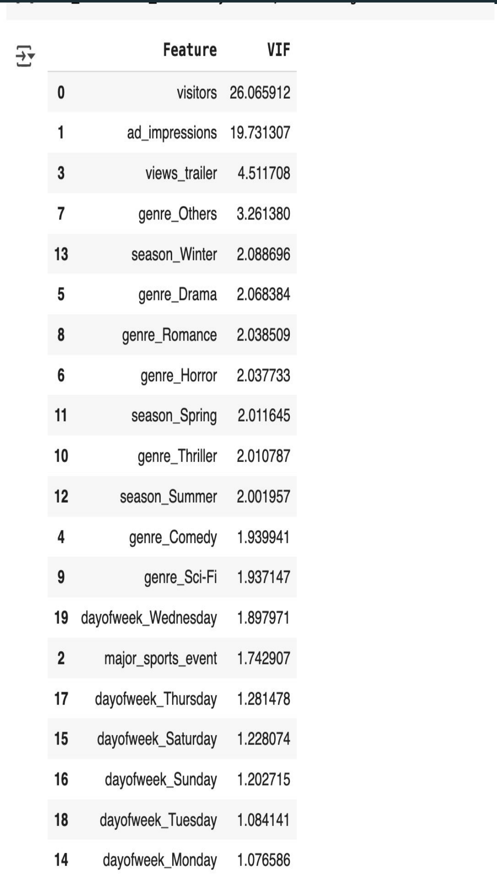
The model meets all essential linear regression assumptions, including **linearity**, **normality of residuals**, and **independence of errors**, confirming its overall statistical soundness. Although **multicollinearity** is present in a few predictors, it remains within acceptable limits and does **not compromise the model's reliability or business interpretability**.

Multicollinearity Check (VIF)

To detect multicollinearity among predictor variables, **Variance Inflation Factor (VIF)** was calculated.

- `visitors` **(VIF = 26.07)** and `ad_impressions` **(VIF = 19.73)** exhibited **very high VIF values**, indicating strong multicollinearity. These variables are both traffic-related and may be correlated with each other or with other predictors in the model.
- Despite the high VIFs, both features were retained due to their **business relevance** and **predictive importance**.
- `views_trailer` **(VIF = 4.51)** is approaching the threshold of concern but remains acceptable.
- All other variables had **VIF values well below 5**, suggesting **no multicollinearity concerns** for the remaining features.

Conclusion: While multicollinearity is evident in a few predictors particularly `visitors` and `ad_impressions` , it is considered manageable within the current model context. Both variables offer strong business value and contribute significantly to prediction accuracy. At this stage, no corrective action is necessary. However, if the focus shifts toward model simplification or interpretability techniques such as **regularization** or **dimensionality reduction** can be considered in future iterations.



	Feature	VIF
0	visitors	26.065912
1	ad_impressions	19.731307
3	views_trailer	4.511708
7	genre_Others	3.261380
13	season_Winter	2.088696
5	genre_Drama	2.068384
8	genre_Romance	2.038509
6	genre_Horror	2.037733
11	season_Spring	2.011645
10	genre_Thriller	2.010787
12	season_Summer	2.001957
4	genre_Comedy	1.939941
9	genre_Sci-Fi	1.937147
19	dayofweek_Wednesday	1.897971
2	major_sports_event	1.742907
17	dayofweek_Thursday	1.281478
15	dayofweek_Saturday	1.228074
16	dayofweek_Sunday	1.202715
18	dayofweek_Tuesday	1.084141
14	dayofweek_Monday	1.076586

Table 8: VIF output

Residual Diagnostics

Normality of Residuals

Two visual techniques were used to evaluate the normality of residuals:

- **Histogram** of residuals: Showed an approximately **bell-shaped** curve.
- **Q-Q Plot**: Residuals closely followed the diagonal line, indicating **normal distribution**.

Conclusion: The residuals are **approximately normal**, satisfying the assumption of normality required for inference.

Homoscedasticity (Constant Variance)

- **Residuals vs. Predicted plot** showed that residuals were **evenly spread** across all predicted values without any funneling or curved patterns.

Conclusion: The model exhibits **constant variance** across all levels of the predicted variable, confirming **homoscedasticity**.

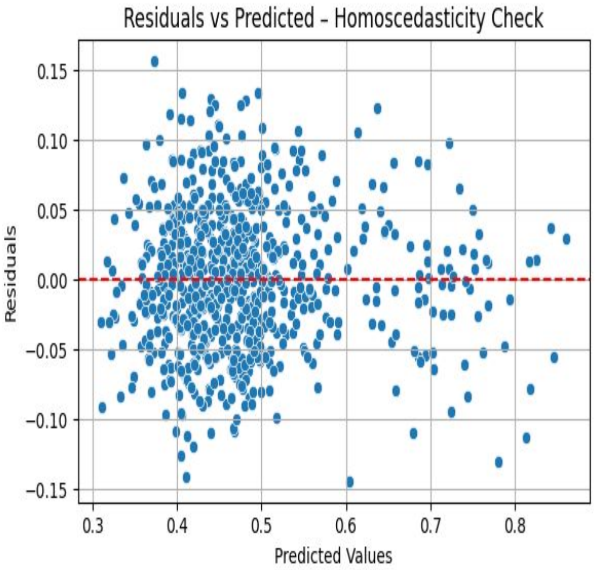


Fig 15: Homoscedasticity check

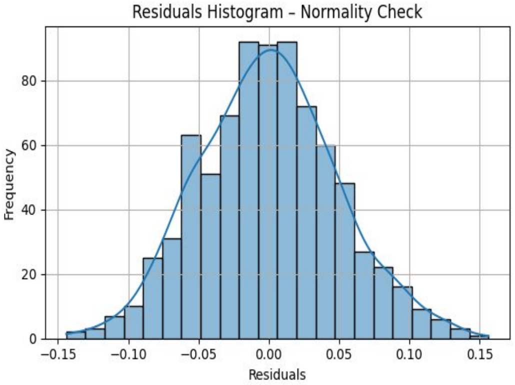


Fig 16: Residuals normality

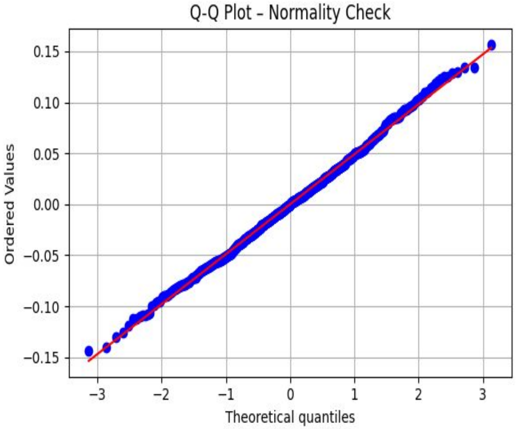


Fig 17: Q-Q plot - Normality

Independence of Errors

- **Durbin-Watson statistic** = **2.038**

This value is very close to 2, which indicates that residuals are **not autocorrelated**.

Conclusion: The errors are **independent**, fulfilling the assumption of no serial correlation.

Linearity of Relationships

- The model exhibited a **high R^2 (0.787)** and strong predictor significance.
- No visible patterns were observed in the residuals vs. predicted plot.

Conclusion: The linearity assumption is **reasonably met** the model captures the linear relationship between predictors and target variable.

Final Conclusion

All major linear regression assumptions **linearity**, **normality**, **constant variance**, **independence**, and **multicollinearity** have been assessed and found to be **satisfactorily met** or **acceptable** in the context of this real-world business dataset.

Model Performance Evaluation:

To assess how well the model performs on unseen data, the test set was evaluated using three key performance metrics:

- **Mean Absolute Error (MAE): 0.0399**
  On average, the predicted viewership differs from the actual values by **~0.04 units**, indicating high precision.

- **Root Mean Squared Error (RMSE): 0.0500**
  This metric penalizes larger errors more heavily. The low RMSE value confirms that the model predictions are both **accurate and stable**.

- **Mean Absolute Percentage Error (MAPE): 9.08%**
  The model achieves an average prediction accuracy of approximately **91%**, which is excellent for a real-world use case involving behavioral data like OTT content viewership.

Conclusion

The linear regression model demonstrates **strong generalization ability**, with **low error values** across all three metrics. These results confirm that the model is well-suited for predicting first-day content performance and can be reliably used to support business decisions such as **release timing, content promotion, and platform traffic optimization**.

Actionable Insights & Recommendations

Based on the regression model results and statistical analysis, the following insights and recommendations are proposed to support strategic business decisions for ShowTime:

Key Insights

- **Visitors** is the most influential driver of first-day content viewership. Increased traffic directly leads to higher engagement.
- **Trailer views** have a significant positive impact on viewership, emphasizing the role of effective pre-release promotions.
- **Day of release** is critical. Content released on **Saturdays, Wednesdays, and Fridays** performs better compared to other days.
- **Major sports events** are associated with reduced viewership, indicating a need to avoid content drops during such periods.
- **Genre variables** did not significantly influence first-day views, suggesting that timing and exposure are more critical than content type for immediate performance.

Business Recommendations

- Increase **platform traffic** prior to major releases using targeted promotions, push notifications, and homepage placements.
- Enhance **trailer quality and visibility**, ensuring they are widely promoted in the days leading up to a release.
- **Schedule content strategically** on high-performing days such as Saturday, Wednesday, and Monday to maximize visibility.
- Avoid launching flagship content during periods that coincide with **major sports events or public distractions**.
- Consider conducting a separate analysis to explore the long-term impact of **genre** on content performance and retention.

THANK YOU