Business Report
PM Coded Project

PGPDSBA

Chithira Raj

Table of Contents

List of Tables

List of Figures

1. Context

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

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

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

2. Objective

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

3. Data Dictionary

S.No.	Variables	Description
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

Table 1: Data Dictionary

4. Data Overview

4.1. Import libraries and load the data

	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

Figure 1: Data Overview

4.2. Check the structure of data

Shape of the dataset: 1000 rows and 8 columns

4.3. Check the types of the data

```
<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
```

Figure 2: Datatypes

4.4. Check for and treat (if needed) missing values

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

Figure 3: Missing values check

4.5. Data Duplicates

There are no duplicate rows.

4.6. Statistical Summary

	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

Figure 4: Statistical Summary

4.7. Insights

- The number of visitors is fairly consistent, with a mean of about 1.7 and a relatively small standard deviation of 0.23, indicating low variability. The median matches the mean, suggesting a symmetric distribution.
- Ad impressions show more variability with a higher standard deviation of 289.53. The mean is around 1434.71, and the median is slightly lower at 1383.58, suggesting a possible right skew.
- Views for trailers have a mean of about 66.92 with a standard deviation of 35. The median is much lower at 53.96, indicating a right skew with some high outliers.
- Content views have a mean of 0.4734 with a standard deviation of 0.1059, indicating moderate variability. The median is 0.45, close to the mean, suggesting a roughly symmetric distribution.

5. Exploratory Data Analysis

5.1. Univariate Analysis

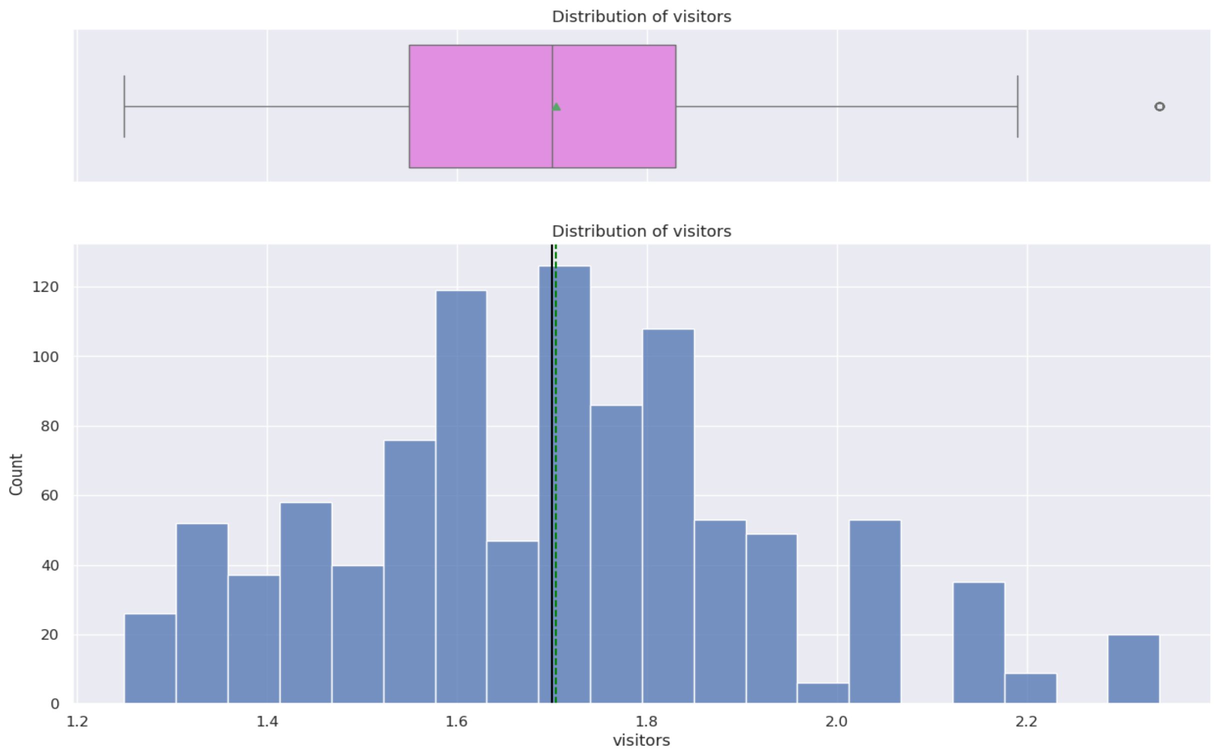

Figure 5: Visitors

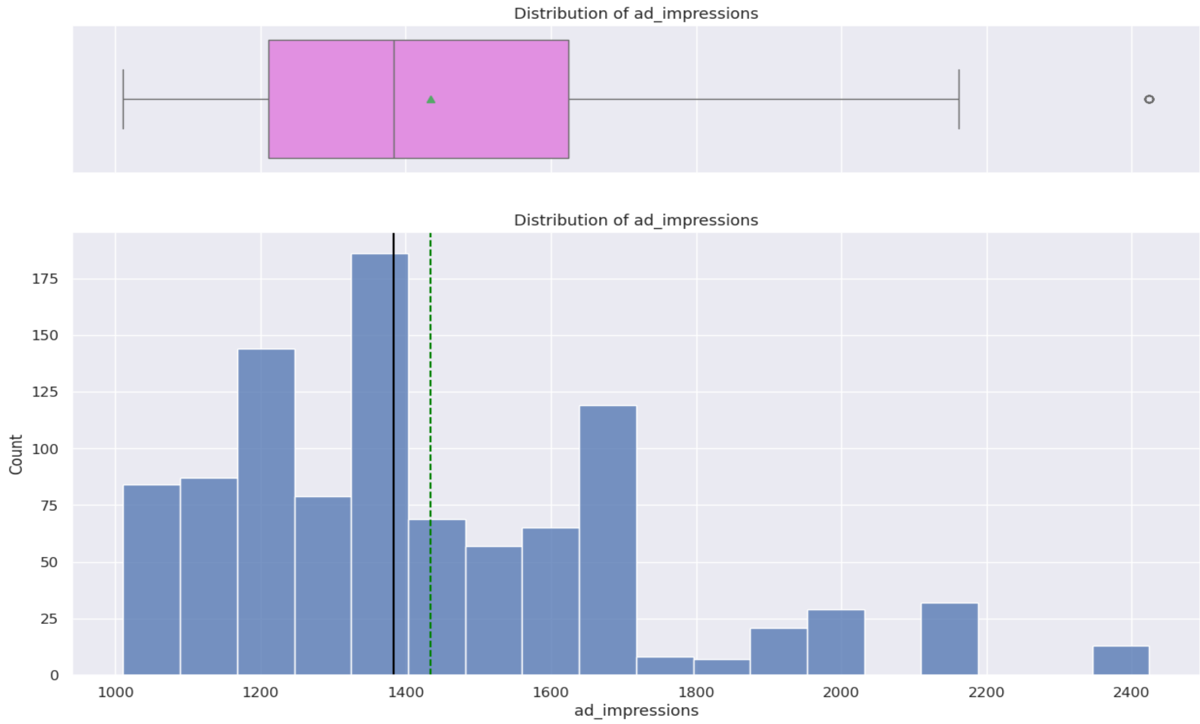

Figure 6: ad_impressions

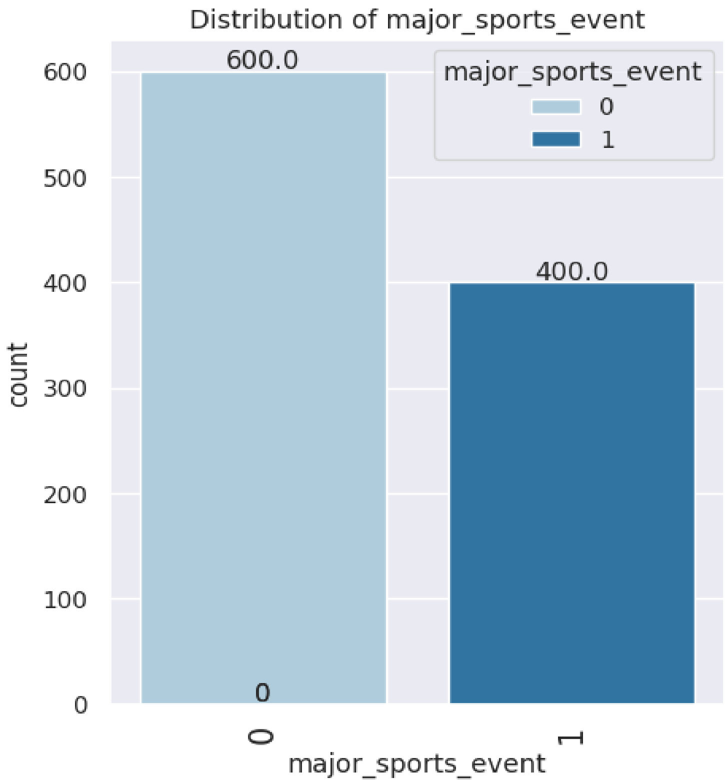

Figure 7: major_sports_event

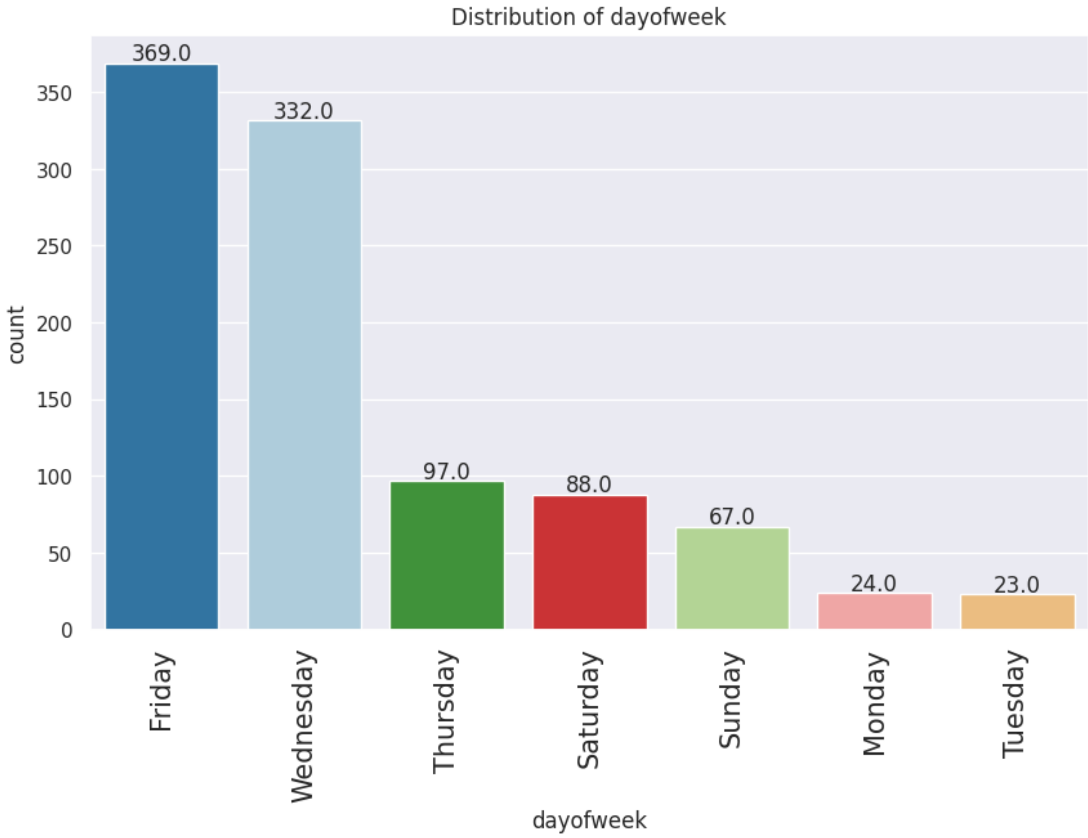

Figure 8: dayofweek

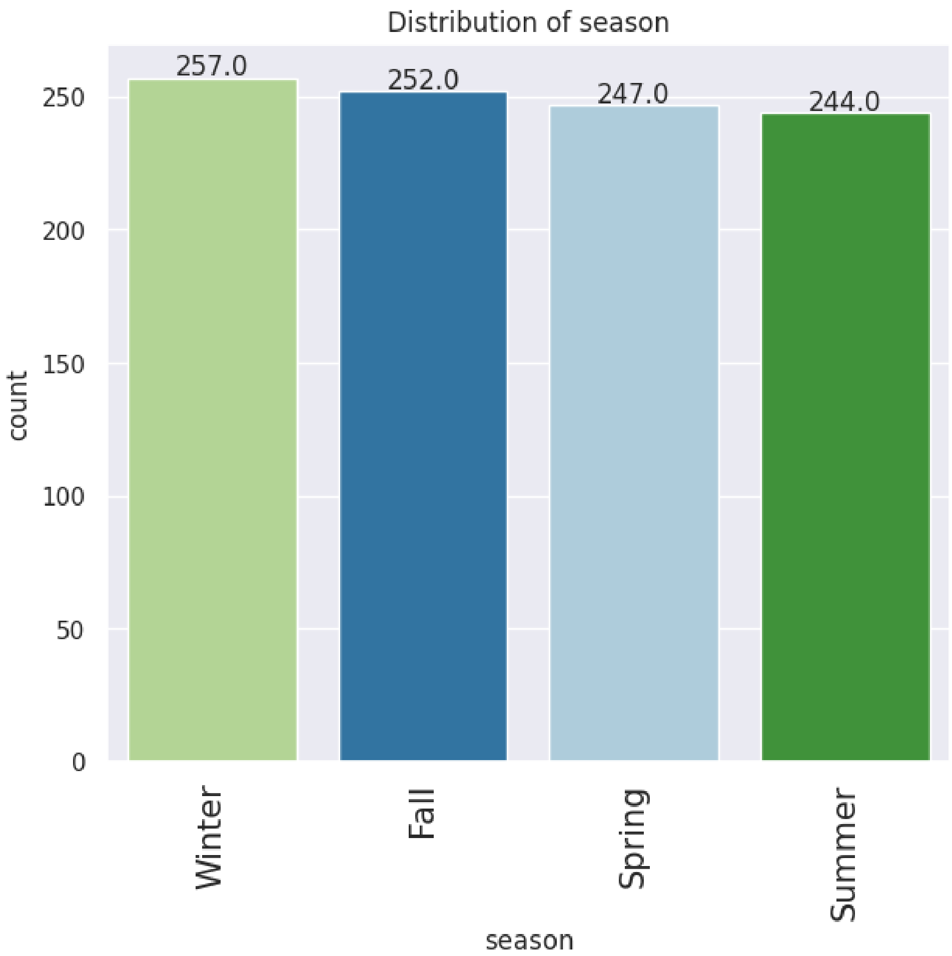

Figure 9: season

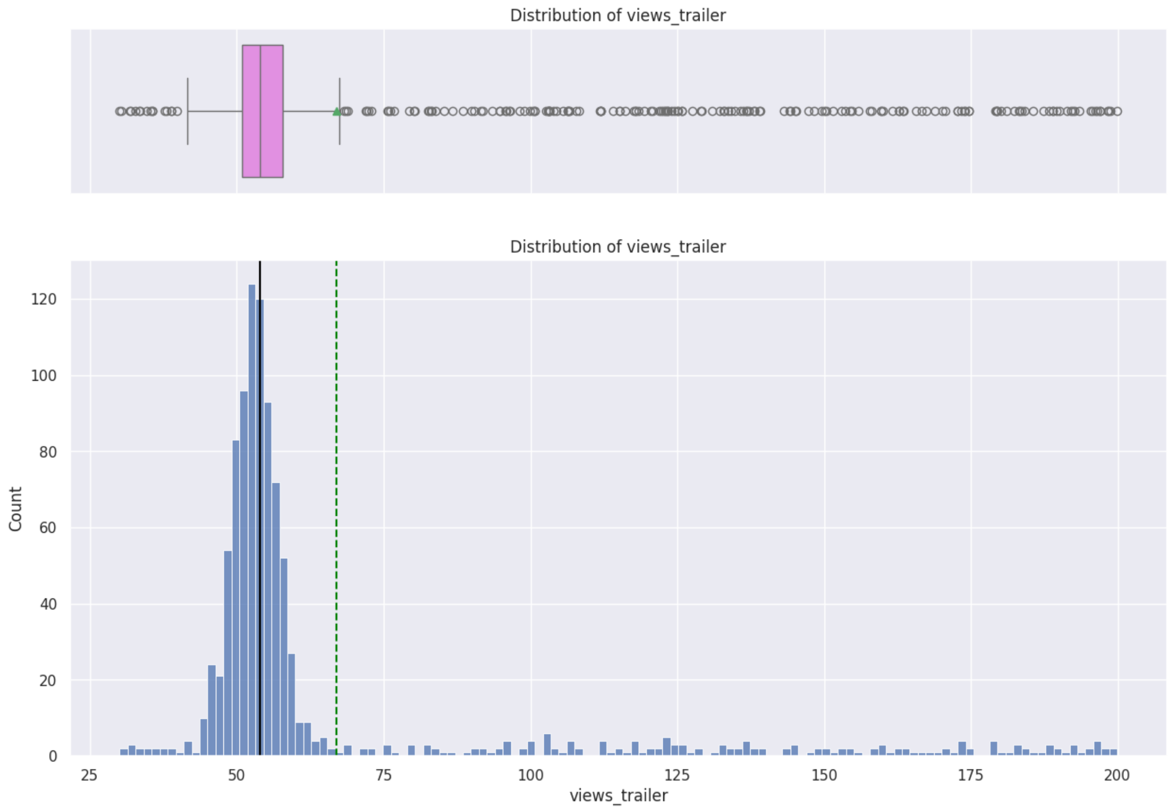

Figure 10: views_trailer

Question 1: What does the distribution of content views look like?

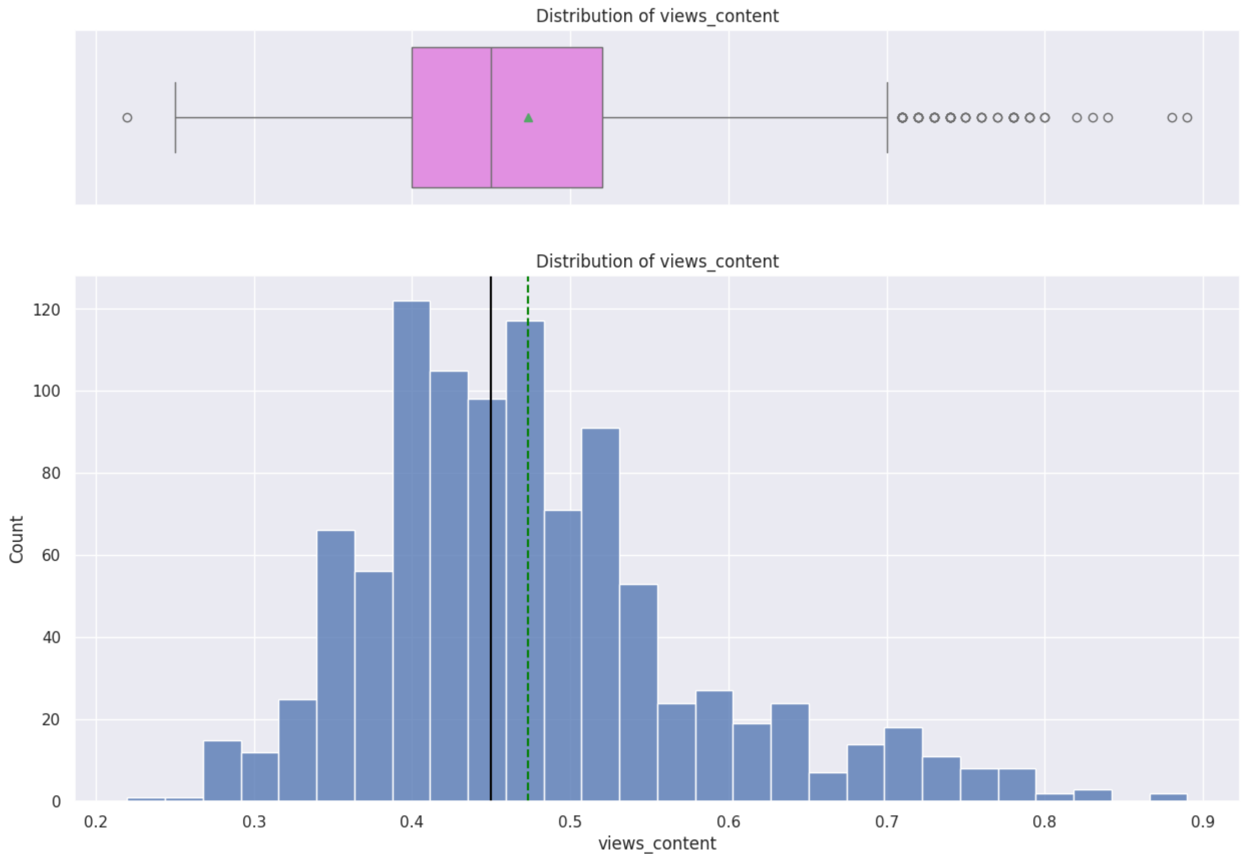

Figure 11: views_content

Question 2: What does the distribution of genres look like?

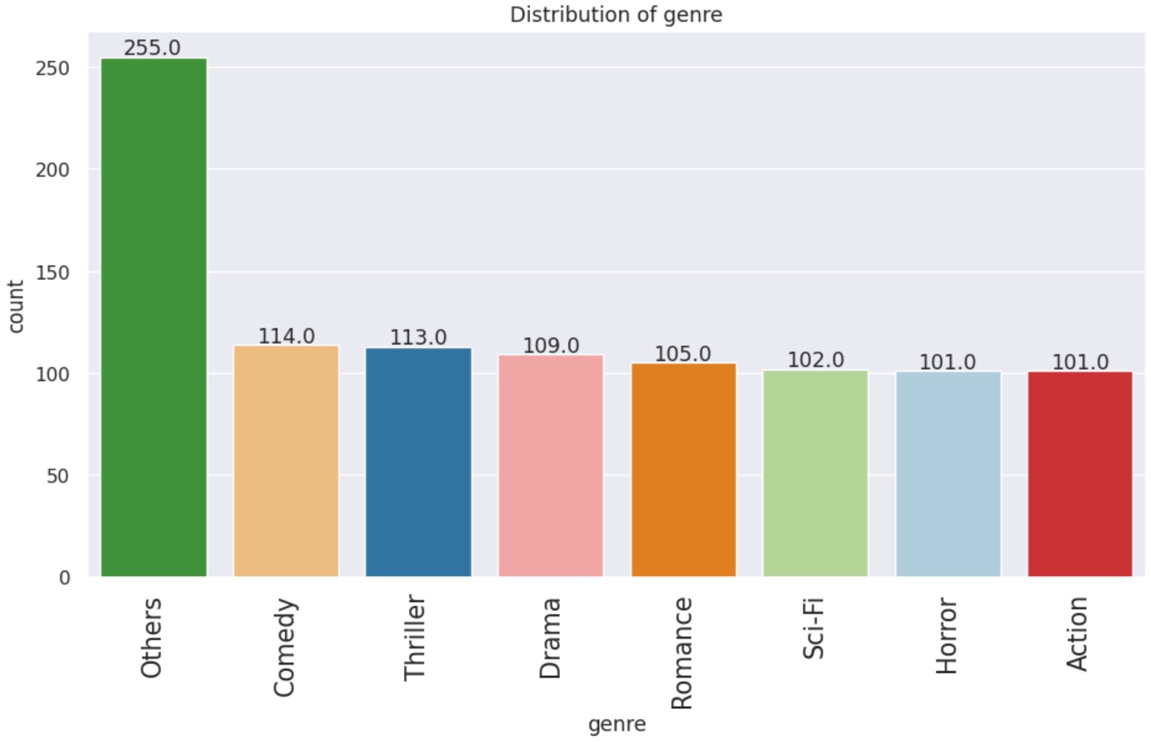

Figure 12: genre

5.2. Bivariate Analysis

Correlation Check

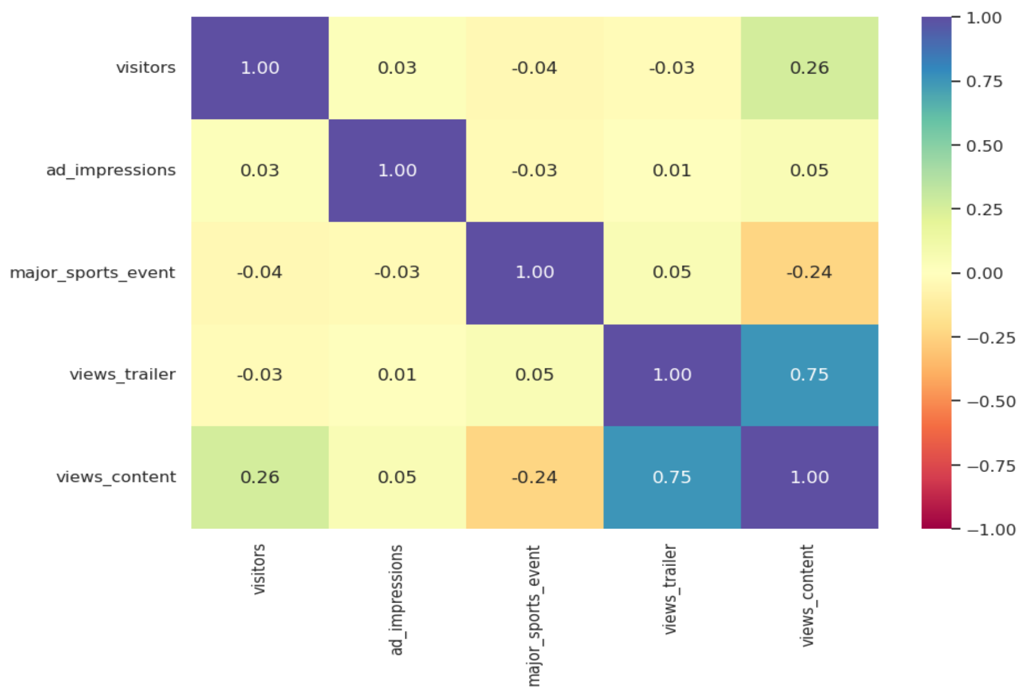

Figure 13: Heatmap

Insights

- There is a positive correlation between views of the trailer and views of the content.
- A major sports event is slightly negatively correlated with views of the content.
- There is a slight positive correlation between the number of visitors and views of the content.

Question 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?

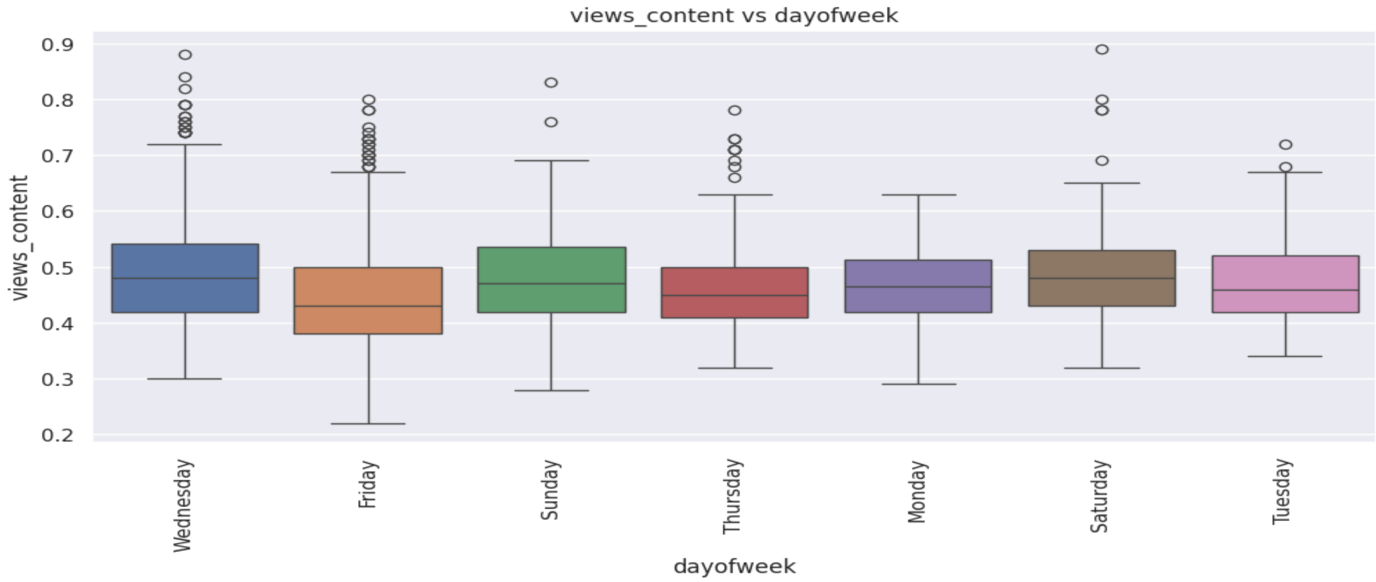

Figure 14: views_content vs dayofweek

views_content vs genre

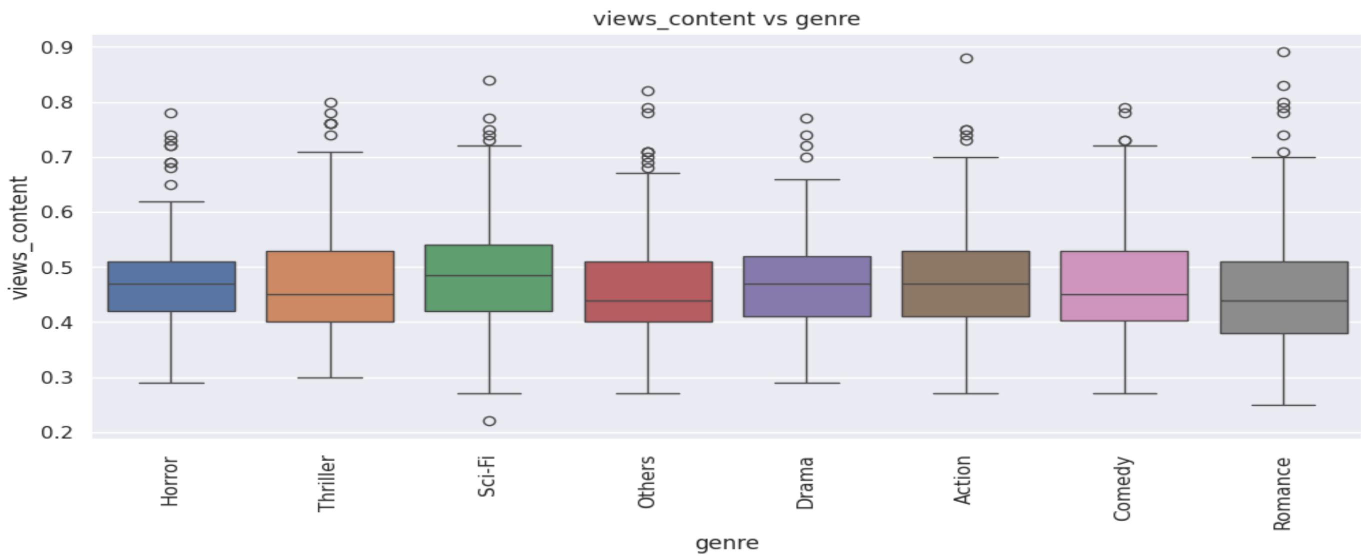

Figure 15: views_content vs genre

Question 4: How does the viewership vary with the season of release?

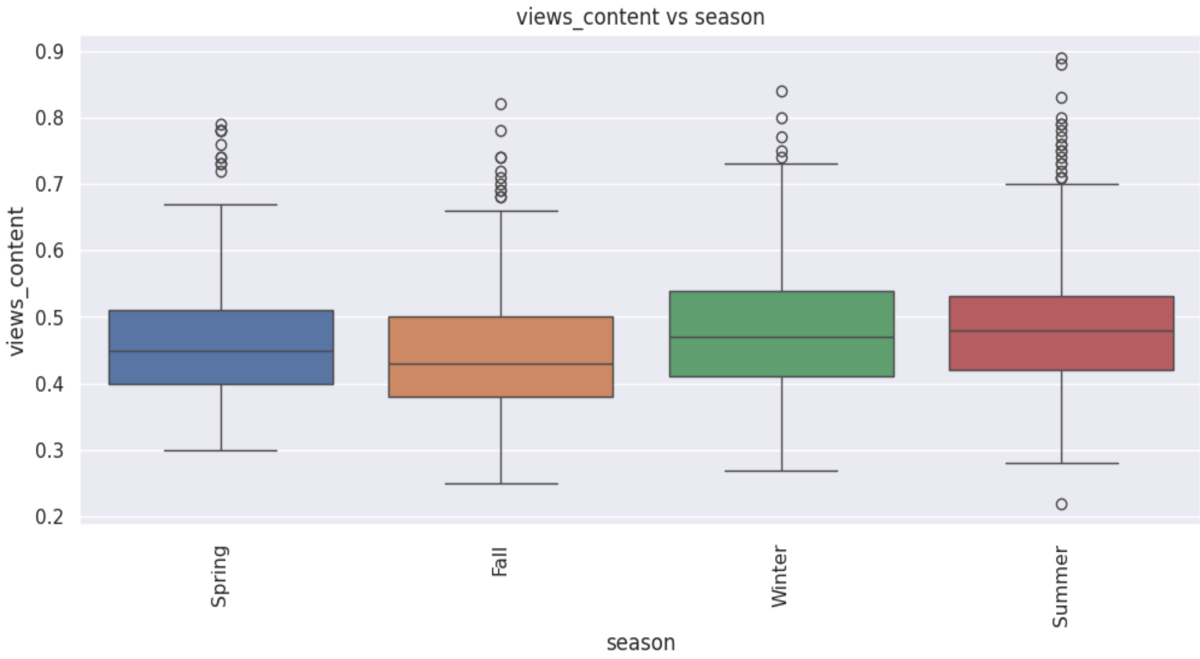

Figure 16: views_content vs season

Question 5: What is the correlation between trailer views and content views?

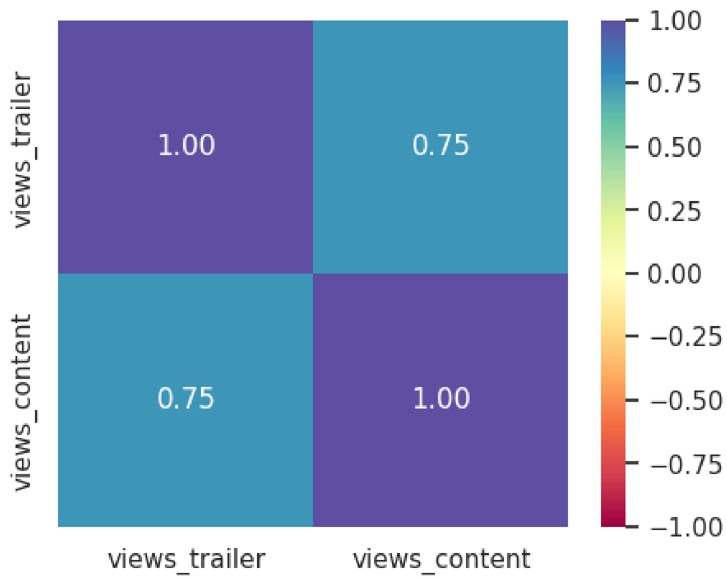

Figure 17: views_content vs views_trailer

6. Data Preprocessing

6.1. Missing Value treatment

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

Figure 18: Missing values check

6.2. Duplicate value check

There are no duplicate rows.

6.3. Feature Engineering

We could perform feature engineering by generating new columns from the existing ones. However, since the dataset is very small and lacks date-like features, it is best to keep the columns as they are.

6.4. Outlier Detection

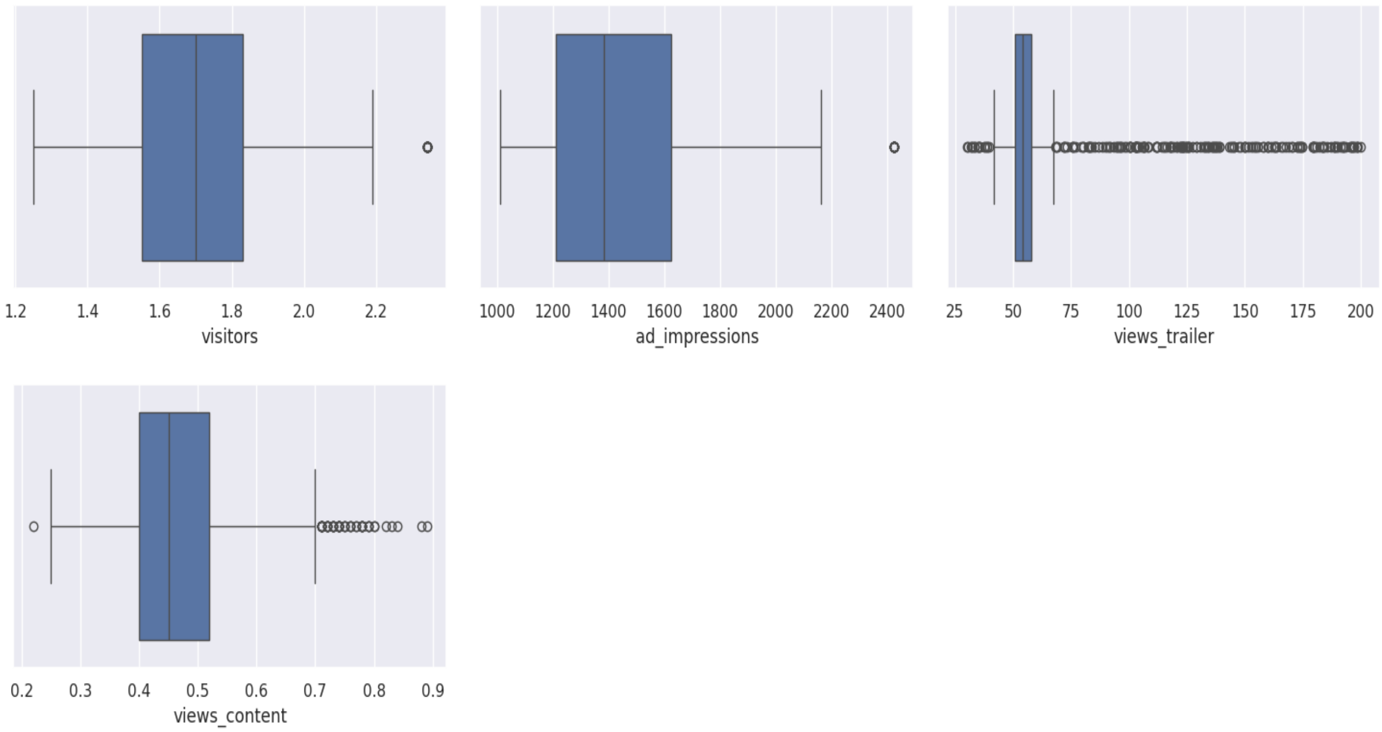

Figure 19: Outliers

There are outliers in the views_trailer column. We have a few options for handling these outliers:

- Use the IQR (Interquartile Range) to determine the lower and upper bounds of the column and either replace or remove the outliers.
- However, since we lack additional information from a subject matter expert, we may decide not to treat these outliers for now.

6.5. Data Preparation for Modeling

To predict the first-day viewership, we'll follow these steps:

1. **Encode categorical features**: Transform any categorical variables into numerical representations suitable for model input.
2. **Split the data**: Divide the dataset into training and testing sets to evaluate the model's performance effectively.
3. **Build a Linear Regression model**: Train the model using the training data.
4. **Evaluate the model**: Assess the model's performance on the testing data.

6.5.1. Encoding Categorical Features

	const	visitors	ad_impressions	major_sports_event	views_trailer	genre_Comedy	genre_Drama	genre_Horror	genre_Others	genre_Romance	...	genre_Thriller	dayofweek_Monday
0	1.0	1.67	1113.81	0	56.70	False	False	True	False	False	...	False	False
1	1.0	1.46	1498.41	1	52.69	False	False	False	False	False	...	True	False
2	1.0	1.47	1079.19	1	48.74	False	False	False	False	False	...	True	False
3	1.0	1.85	1342.77	1	49.81	False	False	False	False	False	...	False	False
4	1.0	1.46	1498.41	0	55.83	False	False	False	False	False	...	False	False

5 rows × 21 columns

Figure 20: Encoding

6.5.2. Train – Test Split

- Number of rows in train data = 700
- Number of rows in test data = 300

7. Model building- Linear Regression

7.1. Model Statistics

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.792
Model:                            OLS   Adj. R-squared:                  0.785
Method:                 Least Squares   F-statistic:                     129.0
Date:                Thu, 04 Jul 2024   Prob (F-statistic):          1.32e-215
Time:                        01:22:54   Log-Likelihood:                 1124.6
No. Observations:                 700   AIC:                            -2207.
Df Residuals:                     679   BIC:                            -2112.
Df Model:                          20
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   0.0602      0.019      3.235      0.001       0.024       0.097
visitors                0.1295      0.008     16.398      0.000       0.114       0.145
ad_impressions       3.623e-06   6.58e-06      0.551      0.582    -9.3e-06    1.65e-05
major_sports_event     -0.0603      0.004    -15.284      0.000      -0.068      -0.053
views_trailer           0.0023   5.52e-05     42.193      0.000       0.002       0.002
genre_Comedy            0.0094      0.008      1.172      0.241      -0.006       0.025
genre_Drama             0.0126      0.008      1.554      0.121      -0.003       0.029
genre_Horror            0.0099      0.008      1.207      0.228      -0.006       0.026
genre_Others            0.0063      0.007      0.897      0.370      -0.008       0.020
genre_Romance           0.0006      0.008      0.065      0.948      -0.016       0.017
genre_Sci-Fi            0.0131      0.008      1.599      0.110      -0.003       0.029
genre_Thriller          0.0087      0.008      1.079      0.281      -0.007       0.025
dayofweek_Monday        0.0337      0.012      2.848      0.005       0.010       0.057
dayofweek_Saturday      0.0579      0.007      8.094      0.000       0.044       0.072
dayofweek_Sunday        0.0363      0.008      4.639      0.000       0.021       0.052
dayofweek_Thursday      0.0173      0.007      2.558      0.011       0.004       0.031
dayofweek_Tuesday       0.0228      0.014      1.665      0.096      -0.004       0.050
dayofweek_Wednesday     0.0474      0.004     10.549      0.000       0.039       0.056
season_Spring           0.0226      0.005      4.224      0.000       0.012       0.033
season_Summer           0.0442      0.005      8.111      0.000       0.034       0.055
season_Winter           0.0272      0.005      5.096      0.000       0.017       0.038
==============================================================================

Omnibus:                        3.850   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.146   Jarque-Bera (JB):                3.722
Skew:                           0.143   Prob(JB):                        0.156
Kurtosis:                       3.215   Cond. No.                     1.67e+04
==============================================================================
```

Figure 21: Model Statistics

Interpreting the Regression Results:

- Adjusted R-squared: This value reflects the model's fit. Adjusted R-squared values range from 0 to 1, with higher values indicating a better fit, assuming certain conditions are met. In our case, the adjusted R-squared value is 0.785, which suggests a good fit.
- Constant Coefficient: This is the Y-intercept of the model. It indicates that if all predictor variable coefficients are zero, the expected output (Y) would equal the constant coefficient. In our case, the constant coefficient is 0.0602.
- Coefficient of a Predictor Variable: This represents the change in the output (Y) due to a change in the predictor variable, holding all other variables constant. In our case, the coefficient of visitors is 0.1295, indicating that an increase in visitors results in an increase of 0.1295 in the predicted output.

7.2. Model Coefficients with column names

Variables	Coefficient Estimate
const	0.060
visitors	0.129
ad_impressions	0.000
major_sports_event	-0.060
views_trailer	0.002
genre_Comedy	0.009
genre_Drama	0.013
genre_Horror	0.010
genre_Others	0.006
genre_Romance	0.001
genre_Sci-Fi	0.013
genre_Thriller	0.009
dayofweek_Monday	0.034
dayofweek_Saturday	0.058
dayofweek_Sunday	0.036
dayofweek_Thursday	0.017
dayofweek_Tuesday	0.023
dayofweek_Wednesday	0.047
season_Spring	0.023
season_Summer	0.044
season_Winter	0.027

Figure 22: Coefficients

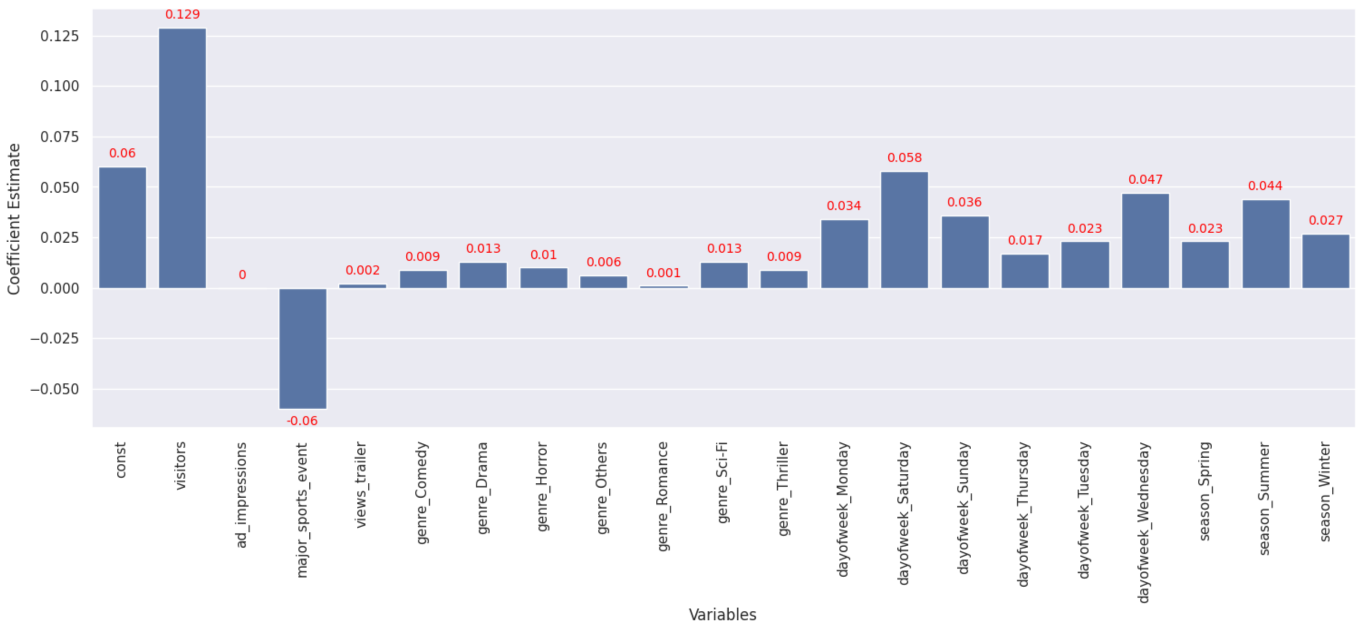

Figure 23: Model Coefficients

8. Testing the assumptions of linear regression model

We will be checking the following Linear Regression assumptions:

1. No Multicollinearity
2. Linearity of variables
3. Independence of error terms
4. Normality of error terms
5. No Heteroscedasticity

8.1. TEST FOR MULTICOLLINEARITY

Multicollinearity occurs when predictor variables in a regression model are correlated, which poses a problem because predictor variables should ideally be independent. High correlation between variables can lead to unreliable coefficients when fitting and interpreting the model.

One method to detect multicollinearity is the Variance Inflation Factor (VIF):

- Variance Inflation Factor (VIF): This measures the inflation in the variances of regression parameter estimates due to collinearities among the predictors. VIF quantifies how much the variance of a regression coefficient (βk) is inflated due to correlation among the predictor variables.

  - If VIF = 1, there is no correlation among the kth predictor and the other predictors, indicating no inflation in the variance of βk.
  - General Rule of Thumb:
    - VIF between 1 and 5 indicates low multicollinearity.
    - VIF between 5 and 10 indicates moderate multicollinearity.
    - VIF exceeding 10 indicates high multicollinearity.

feature	VIF
const	99.679317
genre_Others	2.573779
genre_Drama	1.926699
genre_Thriller	1.921001
genre_Comedy	1.917635
genre_Horror	1.904460
genre_Sci-Fi	1.863473
genre_Romance	1.753525
season_Winter	1.570338
season_Summer	1.568240
season_Spring	1.541591
dayofweek_Wednesday	1.315231
dayofweek_Thursday	1.169870
dayofweek_Saturday	1.155744
dayofweek_Sunday	1.150409
major_sports_event	1.065689
dayofweek_Monday	1.063551
dayofweek_Tuesday	1.062793
ad_impressions	1.029390
visitors	1.027837
views_trailer	1.023551

Figure 24: VIF

- Except for the constant term, all other features have VIF values below 5, suggesting that multicollinearity is not a significant issue for these predictors.
- The highest VIF among the predictors is for "genre_Others" (2.573779), which is within acceptable limits.
- This means the predictors are relatively independent, and the regression coefficients should be reliable.
- We will ignore the VIF values for dummy variables and the constant (intercept)

8.2. TEST FOR LINEARITY AND INDEPENDENCE

- Linearity: This describes a straight-line relationship between two variables. Predictor variables must have a linear relationship with the dependent variable for the model to be valid.
- Independence of Error Terms: It is important that the residuals (errors) are independent. If the residuals are not independent, the confidence intervals of the coefficient estimates will be narrower, potentially leading to incorrect conclusions about a parameter's statistical significance.

  How to Check Linearity and Independence?

- Plot of Fitted Values vs. Residuals: Create a plot of the fitted values against the residuals.
- If the plot shows no pattern, the model is linear and the residuals are independent.
- If there is a pattern, it indicates non-linearity and that the residuals are not independent.

  How to Fix if This Assumption is Not Followed?

- Transform the Variables: If the relationship is not linear or the residuals are not independent, try transforming the variables to achieve linearity and independence. This can help meet the assumptions required for the regression model to be valid.

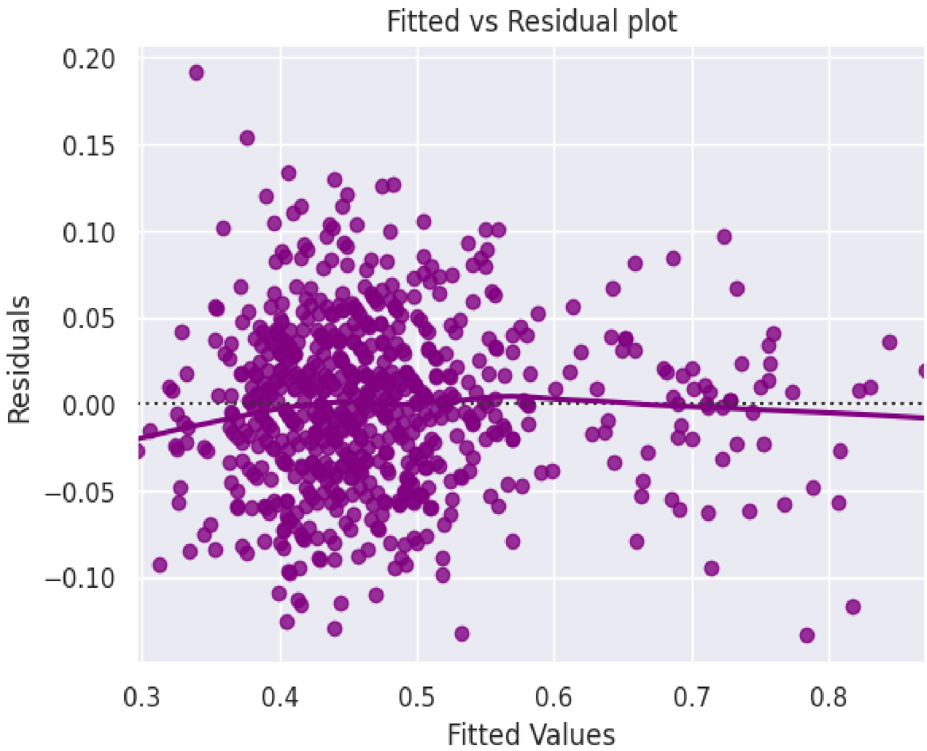

Figure 25: Fitted vs Residual

We see no pattern in the plot above. Hence, the assumptions of linearity and independence are satisfied.

8.3. TEST FOR NORMALITY

Why the Test?

Error terms, or residuals, should be normally distributed. If the residuals are not normally distributed, the confidence intervals of the coefficient estimates may become too wide or narrow, making the confidence intervals unstable and leading to difficulties in estimating coefficients using the minimization of least squares. Non-normality suggests the presence of unusual data points that should be studied closely to improve the model.

How to Check Normality?

- Histogram of Residuals: The shape of the histogram of residuals can give an initial idea about their normality.
- Q-Q Plot of Residuals: If the residuals follow a normal distribution, they will form a straight line in a Q-Q plot. Deviations from this line indicate non-normality.
- Shapiro-Wilk Test: This test can be used to formally check normality.
  - Null Hypothesis: Residuals are normally distributed.
  - Alternate Hypothesis: Residuals are not normally distributed.

How to Fix if This Assumption is Not Followed?

Apply Transformations: Transform the data using functions like log, exponential, arcsinh, etc., as appropriate for the data.

Histogram

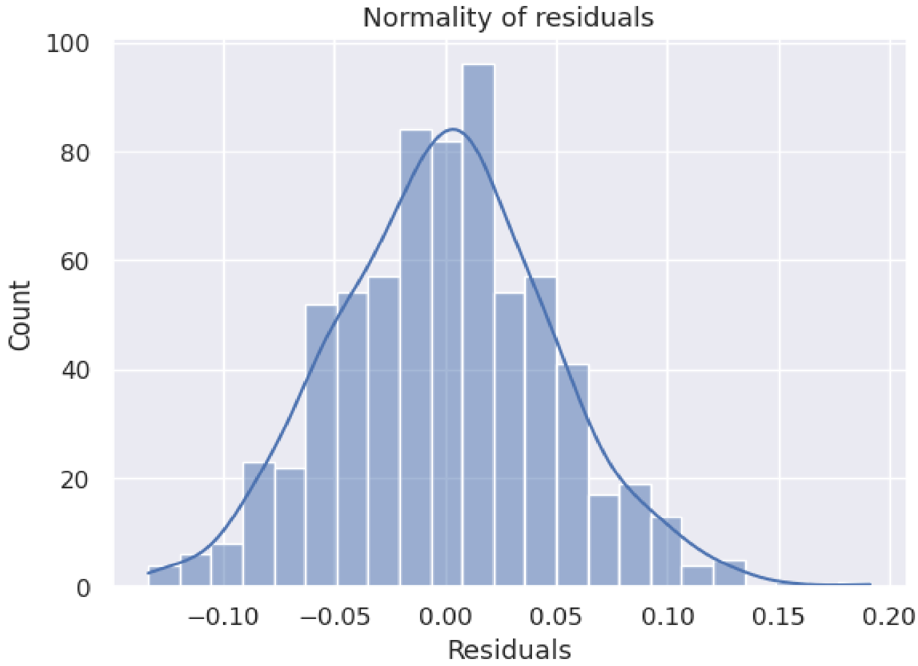

Figure 26: Histogram of residuals

The histogram of residuals has a bell shape, which suggests that the residuals are normally distributed. This indicates that the normality assumption is likely satisfied, supporting the validity of our regression model.

Q-Q plot

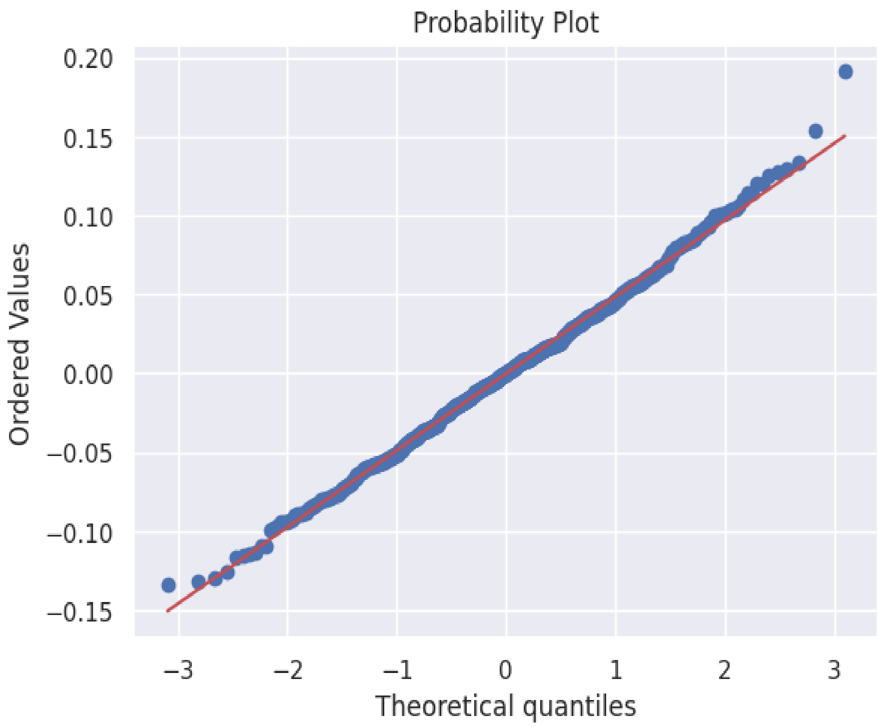

Figure 27: Q-Q plot

The residuals more or less follow a straight line except for the tails, indicating that the residuals are normally distributed.

Shapiro-Wilk test

Null Hypothesis: The residuals are normally distributed.

Alternate Hypothesis: The residuals are not normally distributed.

statistic=0.9972

pvalue=0.2841

Since the p-value is greater than 0.05, the Shapiro-Wilk test indicates that the residuals are normally distributed. Therefore, the assumption of normality is satisfied.

8.4. TEST FOR HOMOSCEDASTICITY

Homoscedasticity: If the variance of the residuals is symmetrically distributed across the regression line, the data is said to be homoscedastic.

Heteroscedasticity: If the variance of the residuals is unequal across the regression line, the data is said to be heteroscedastic.

Why the Test?

- The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in the presence of outliers.

How to Check for Homoscedasticity?

1. Residual vs. Fitted Values Plot:

    - Examine the plot of residuals versus fitted values. If the residuals form an arrow shape or any other non-symmetrical pattern, it indicates heteroscedasticity.

2. Goldfeld-Quandt Test:

    - Perform the Goldfeld-Quandt test. If the p-value is greater than 0.05, we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.
    - Null Hypothesis: Residuals are homoscedastic.
    - Alternate Hypothesis: Residuals exhibit heteroscedasticity.

How to Fix if This Assumption is Not Followed?

- Adding Important Features: Include other relevant features that might explain the variance.
- Transformations: Apply transformations to stabilize the variance, such as log, square root, or other suitable transformations.

Goldfeld-Quandt test

Null Hypothesis: Homoscedasticity

Alternate Hypothesis: Heteroscedasticity

F statistic = 1.1444

p-value = 0.1108

Since the p-value is greater than 0.05, we can conclude that the residuals are homoscedastic. Therefore, this assumption is satisfied.

9. Model performance evaluation

Train Performance

RMSE	MAE	R-squared	Adj. R-squared	MAPE
0.04853	0.038197	0.791616	0.785162	8.55644

Figure 28: Train data evaluation

Test Performance

RMSE	MAE	R-squared	Adj. R-squared	MAPE
0.050603	0.040782	0.766447	0.748804	9.030464

Figure 29: Test data evaluation

Insights

- The model explains approximately 79% of the variation in the data.
- The train and test RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) are low and comparable, indicating that the model is not suffering from overfitting.
- The MAPE (Mean Absolute Percentage Error) on the test set suggests that we can predict within 9% of the first-day viewership.
- Therefore, we can conclude that the model is good for both prediction and inference purposes.

10. Actionable Insights & Recommendations

The model is able to explain ~79% of the variation in the data and within 9% of the first-day viewership on the test data, which is good. This indicates that the model is good for prediction as well as inference purposes

Significant Predictors

- Visitors: The coefficient for visitors (0.1295, $p < 0.0001$) is positive and highly significant. This suggests that an increase in the number of visitors strongly correlates with an increase in views on the content.
- Major Sports Event: The coefficient is negative (-0.0603, $p < 0.0001$), indicating that major sports events tend to reduce views on the content. This insight can help in planning content release schedules to avoid clashing with major sports events to minimize the impact on views.
- Views Trailer: The coefficient for views trailer (0.0023, $p < 0.0001$) is positive and highly significant. Increasing trailer views can lead to a substantial increase in content views.
- Day of the Week: Several days of the week have significant positive coefficients. This indicates that content views are higher on these days, especially on weekends.
- Seasons: The coefficients for all seasons (Spring: 0.0226, Summer: 0.0442, Winter: 0.0272) are positive and significant, with Summer having the highest impact.

Non-Significant Predictors:

- Ad Impressions: The coefficient for ad impressions (3.623e-06, $p = 0.582$) is not significant. This suggests that the number of ad impressions does not have a strong direct impact on content views.
- Genres: Most genres (Comedy, Drama, Horror, Others, Romance, Sci-Fi, Thriller) have non-significant coefficients. This indicates that the genre of content does not significantly impact the number of views.

Key Takeaways for the Business:

- Focus on Increasing Visitors: Invest in strategies to attract more visitors, as this has the highest impact on content views.
- Avoid Major Sports Events: Schedule important content releases outside of major sports events to avoid viewership dips.
- Promote Trailers: Increase efforts to promote trailers as their views strongly correlate with content views.
- Optimal Release Days: Plan significant content releases or special promotions on Mondays, Wednesdays, Thursdays, Saturdays, and Sundays to maximize viewership.
- Seasonal Planning: Align major content strategies with seasonal trends, focusing on summer for the highest impact.
- Review Ad Strategies: Reassess the current ad impression strategies as they do not show a significant impact on views. Consider optimizing or exploring other advertising methods.