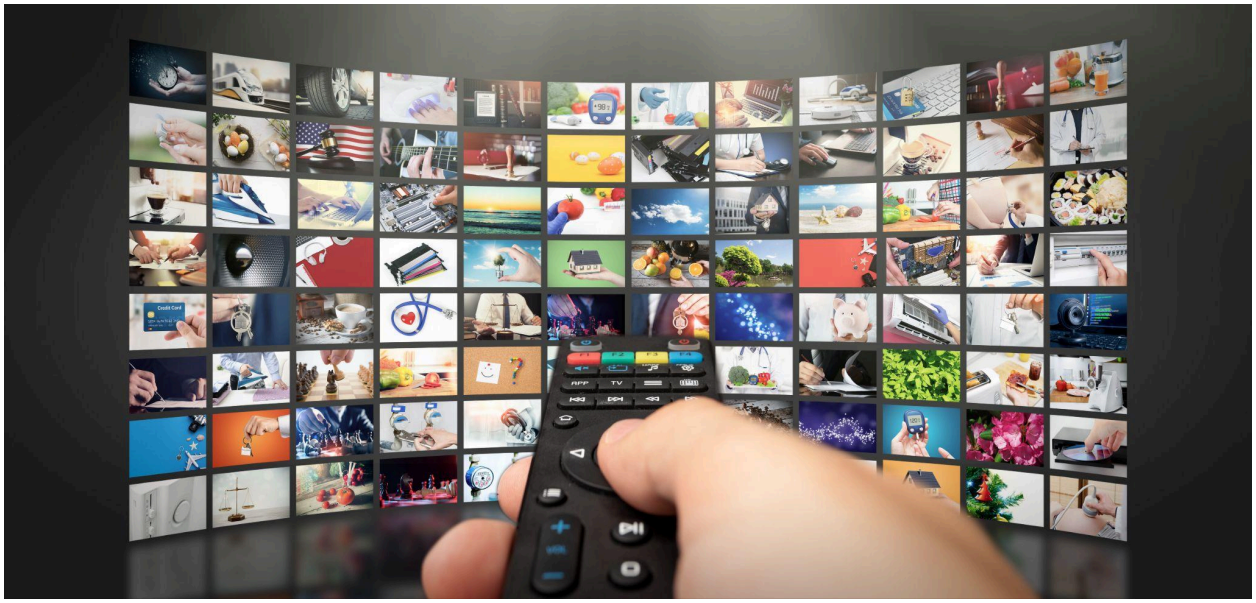
Business Report

ShowTime

OTT Data Analysis Business Report



- Ruchi Rikta

Contents

List of Figures

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 in 2019 and is projected to reach $1,039.03 billion by 2027, growing at a CAGR of 29.4% from 2020 to 2027. The shift from television to OTT services for entertainment is driven by benefits such as on-demand services, ease of access, and access to better networks and digital connectivity.

With the outbreak of COVID-19, OTT services are striving to meet the growing entertainment appetite of viewers, with some platforms already experiencing a 46% increase in consumption and subscriber count as viewers seek fresh content. With innovations and advanced transformations, which will enable the customers to access everything they want in a single space, OTT platforms across the world are expected to increasingly attract subscribers on a concurrent basis.

Objective

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

Data Description

The data contains the different factors to analyze for the content. The detailed data dictionary is given below.

Data Dictionary:

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

Data Overview

- **Data Type**

```
<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
```

- **1st 5 rows of the Dataset**

	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

- **Statistical Summary**

	visitors	ad_impressions	major_sports_event	views_trailer	views_content
count	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000
mean	1.704290	1434.712290	0.400000	66.91559	0.473400
std	0.231973	289.534834	0.490143	35.00108	0.105914
min	1.250000	1010.870000	0.000000	30.08000	0.220000
25%	1.550000	1210.330000	0.000000	50.94750	0.400000
50%	1.700000	1383.580000	0.000000	53.96000	0.450000
75%	1.830000	1623.670000	1.000000	57.75500	0.520000
max	2.340000	2424.200000	1.000000	199.92000	0.890000

- **Null 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

● Categorical Statistical Summary

	count	unique	top	freq
genre	1000	8	Others	255
dayofweek	1000	7	Friday	369
season	1000	4	Winter	257

● Observations and Insights

- The data set contains information about first-day views of the content based on the day of week, genre, and season.
- The data set contains 1000 rows with 8 attributes.
- There are 5 numeric (4 float and 1 int) and 3 string (object) type data in the columns of the dataset.
- There are 8 genres, 4 seasons, and all days of the week mentioned in the data.
- The mean and standard deviation of the Views of the content are 0.47 and 0.105, respectively.

Exploratory Data Analysis

1. Univariate Data Analysis

1.1. Distribution of Visitors

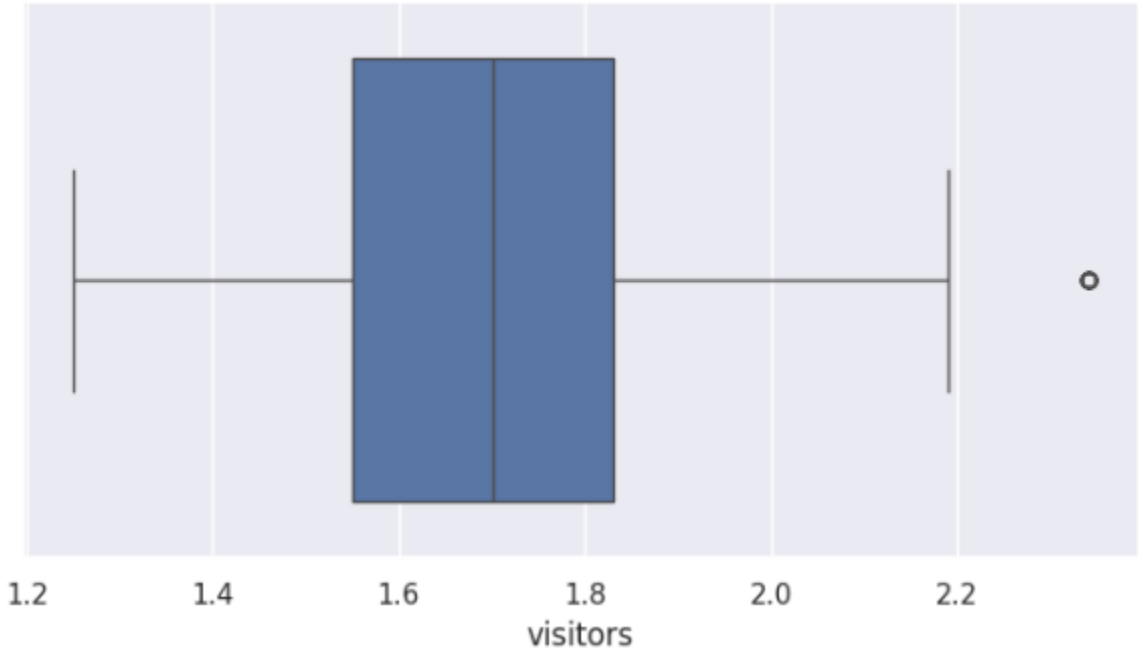


Fig.1.1: Distribution of Visitors (Boxplot)

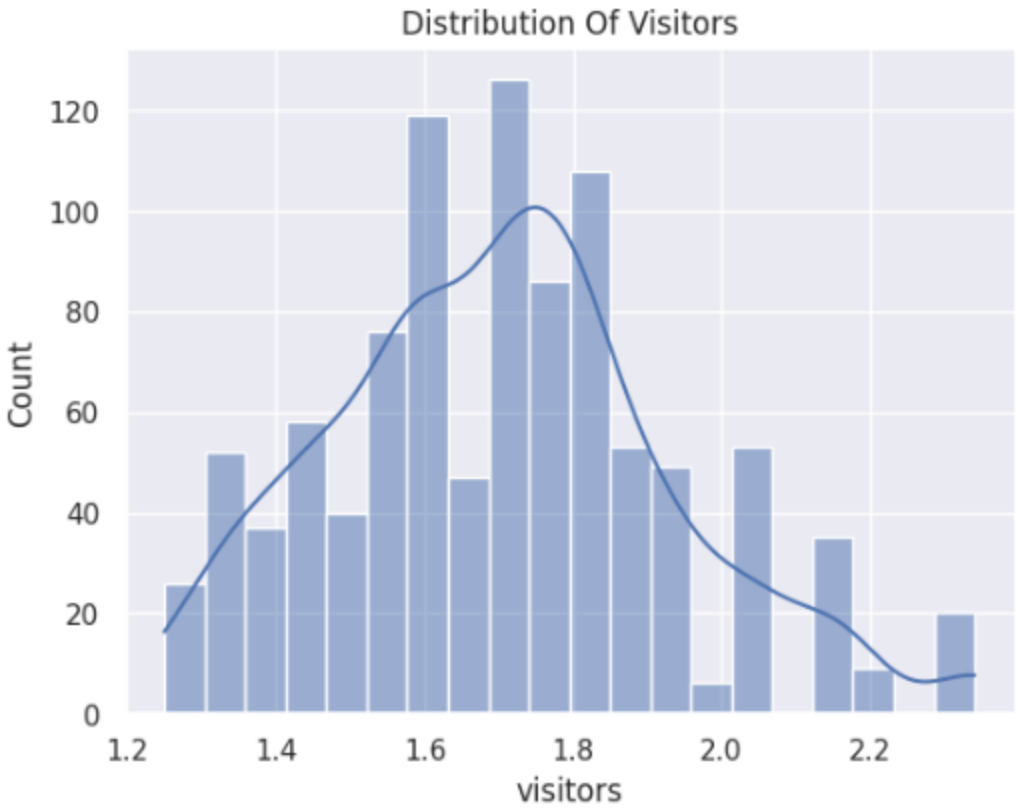


Fig.1.2: Distribution of Visitors (Histogram)

- The histogram is right-skewed, and it is a normal distribution.
- Very few outliers are observed.

1.2. Trailer Views Distribution

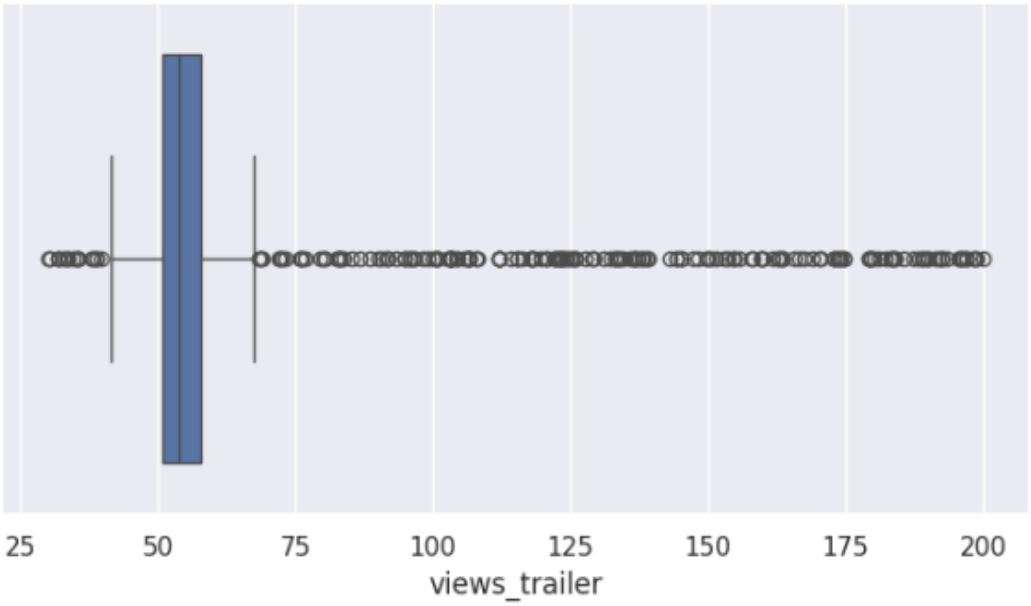


Fig.1.2.1: Trailer Views Distribution (Boxplot)

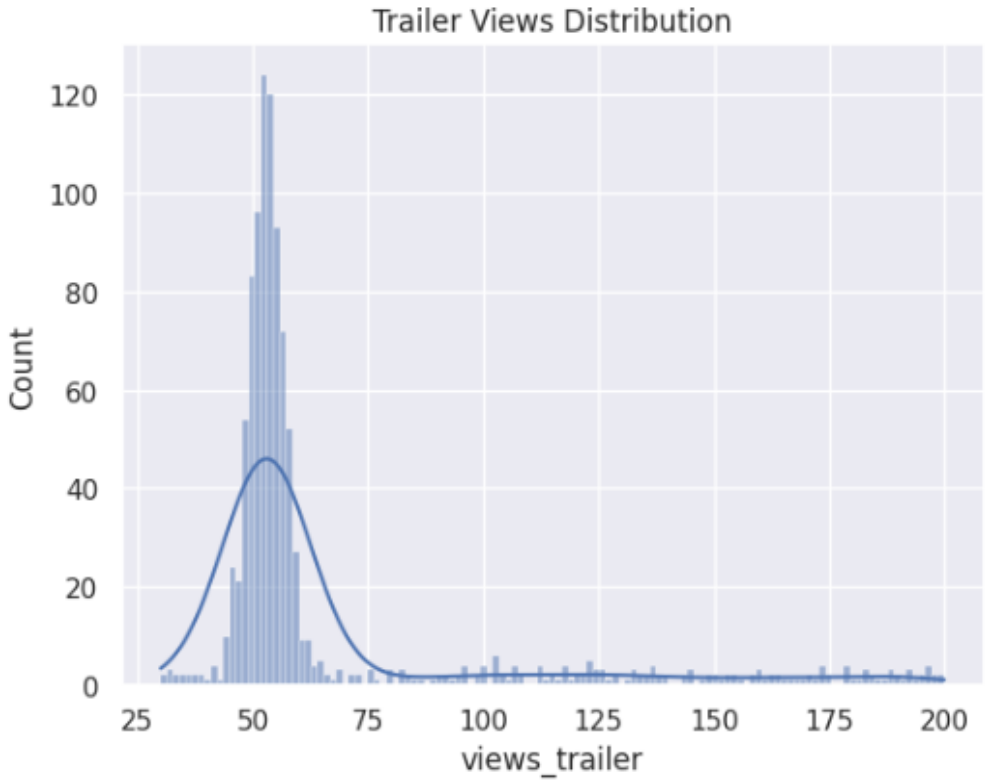


Fig.1.2.2: Trailer Views Distribution (Histogram)

- Outliers can be seen heavily.
- Data is heavily skewed towards the right.

1.3. Content Views Distribution

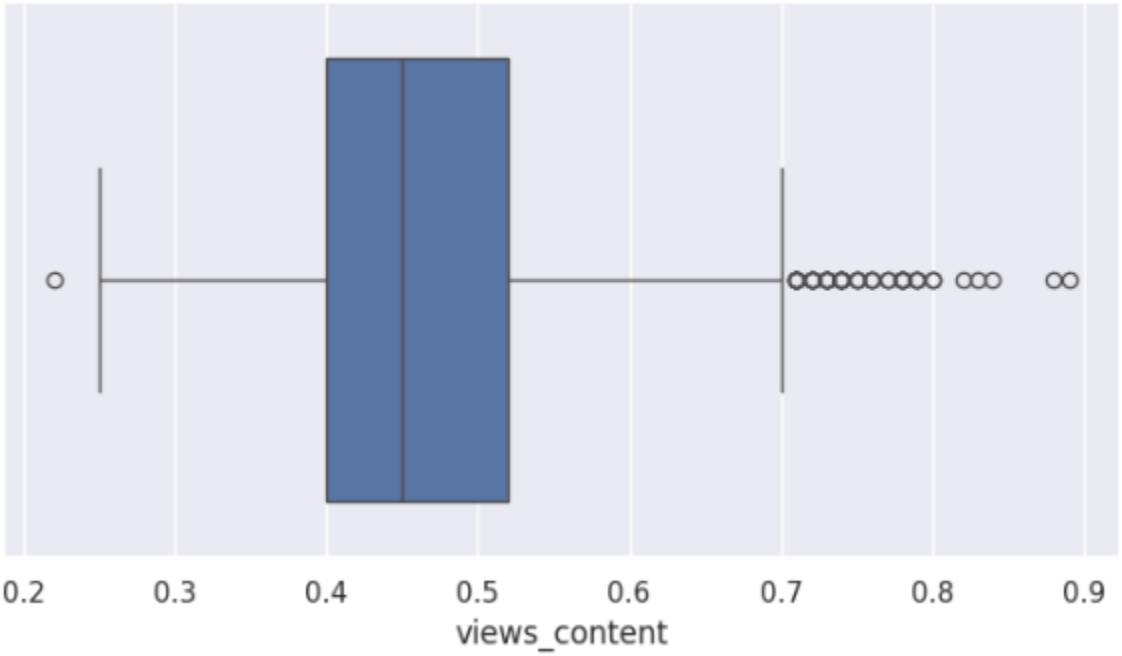


Fig.1.3.1: Distribution of Content Views (Boxplot)

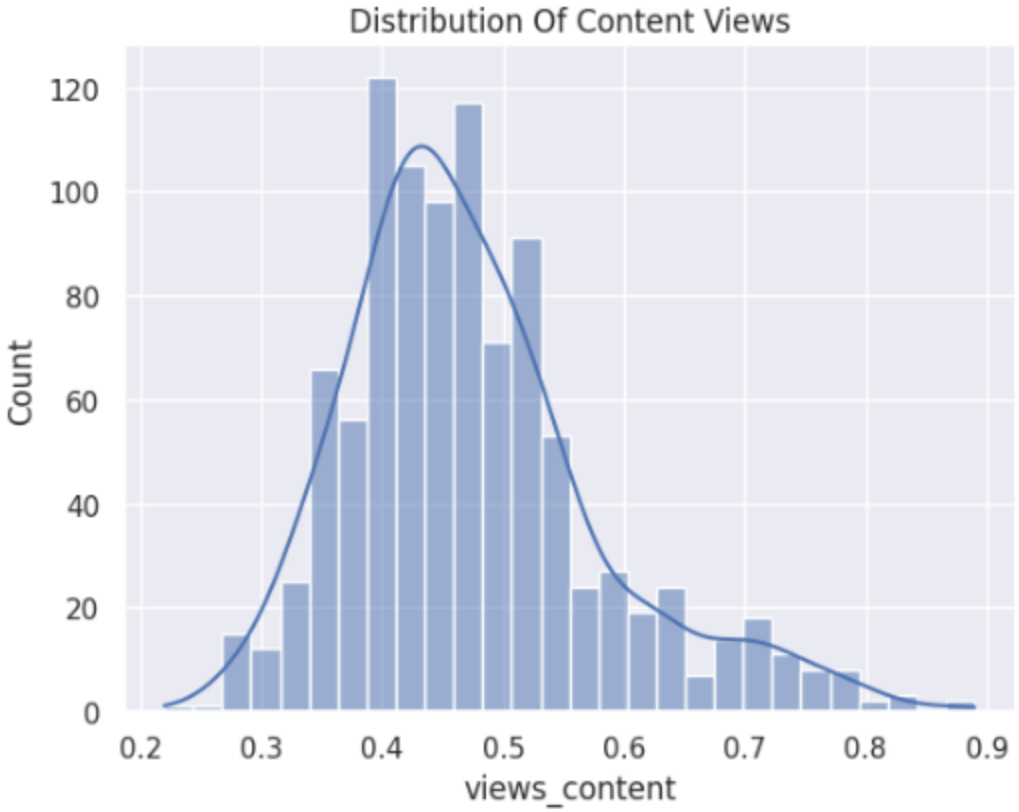


Fig.1.3.2: Distribution of Content Views (Histogram)

- A Few Outliers can be observed in the views content field.
- The histogram is right-skewed; it seems almost like a normal distribution.

1.4. Genre Distribution

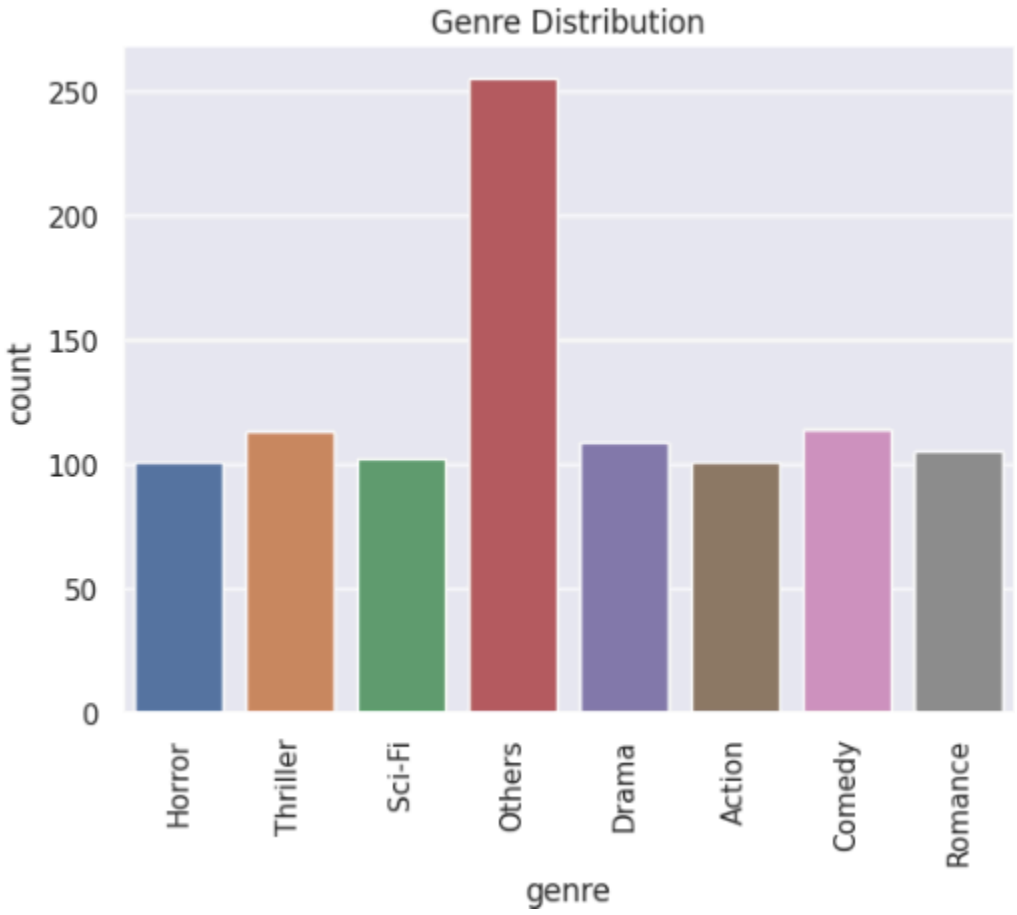


Fig.1.4 : Genre Distribution

- Others have the highest release, followed by Comedy and Thriller.

1.5. Weekday Distribution

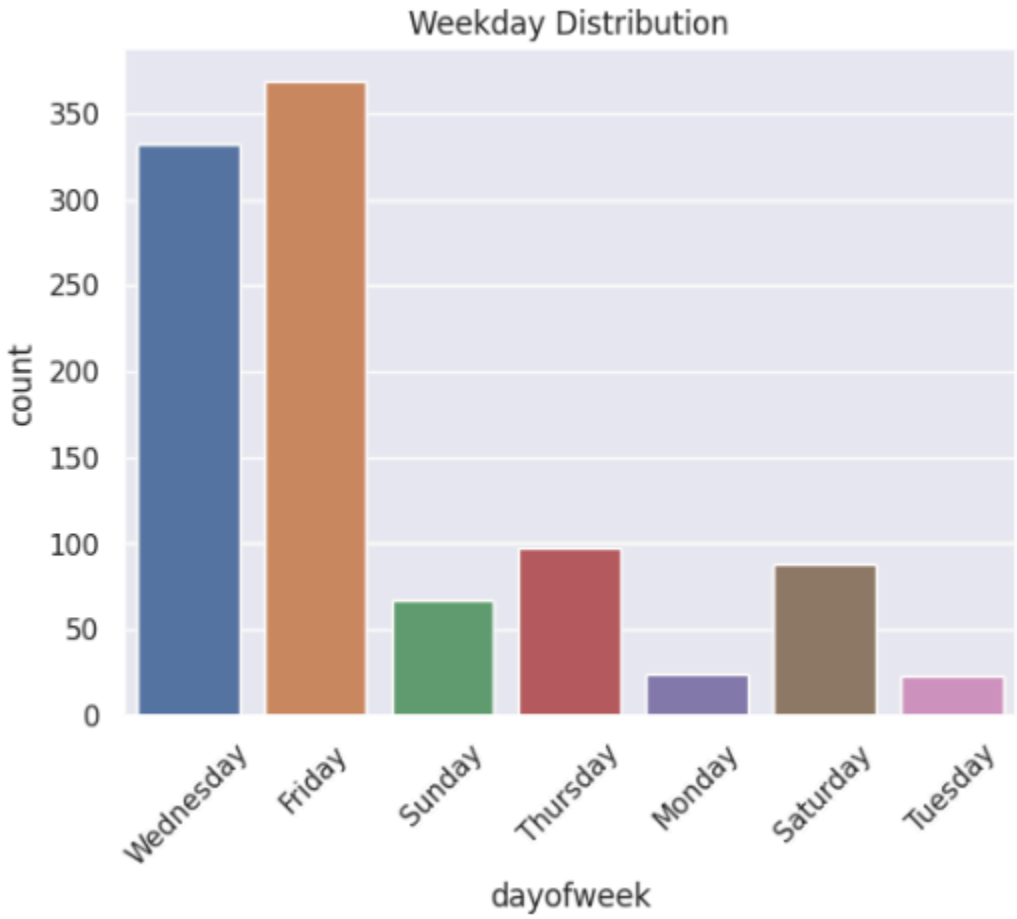


Fig.1.5 : Weekday Distribution

- Content released on Friday has the highest release, followed by Wednesday.
- Monday and Tuesday have very little content release.

1.6. Season Distribution

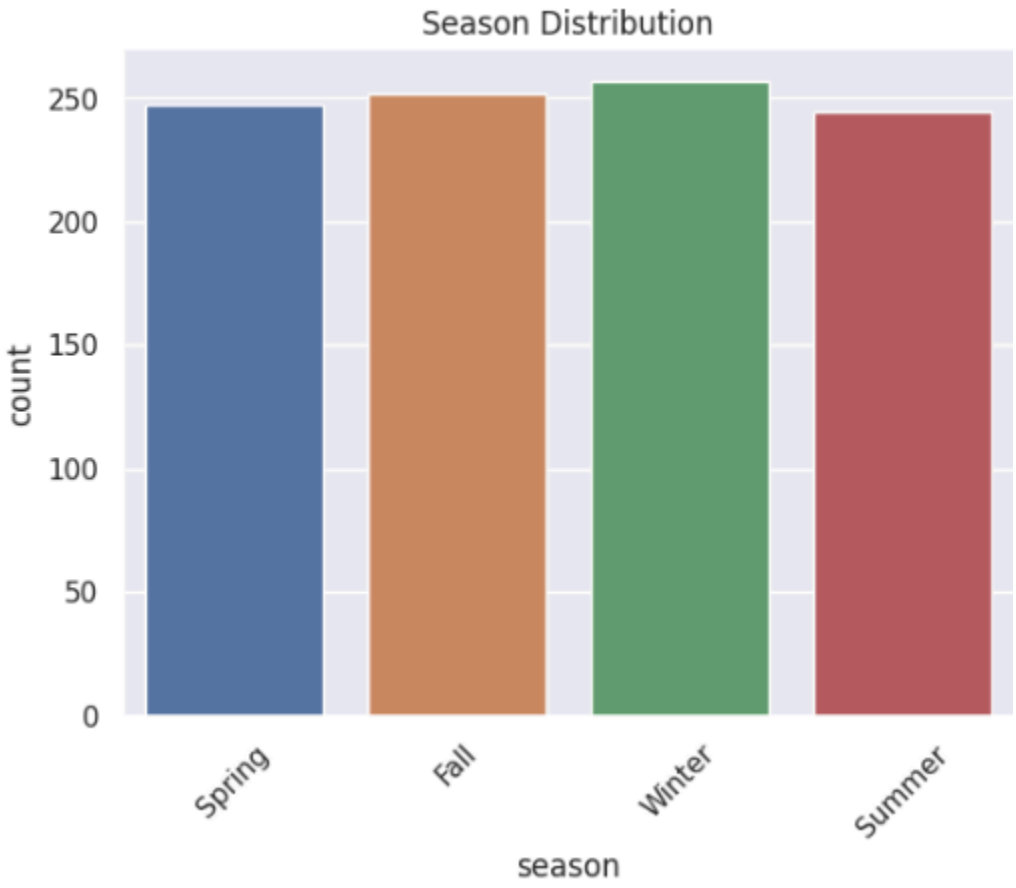


Fig.1.6 : Season Distribution

- Winter and Fall have the highest release.
- The Other two seasons also more or less have the same percentage of release.

2. Bivariate Data Analysis

2.1. Correlation between numerical values

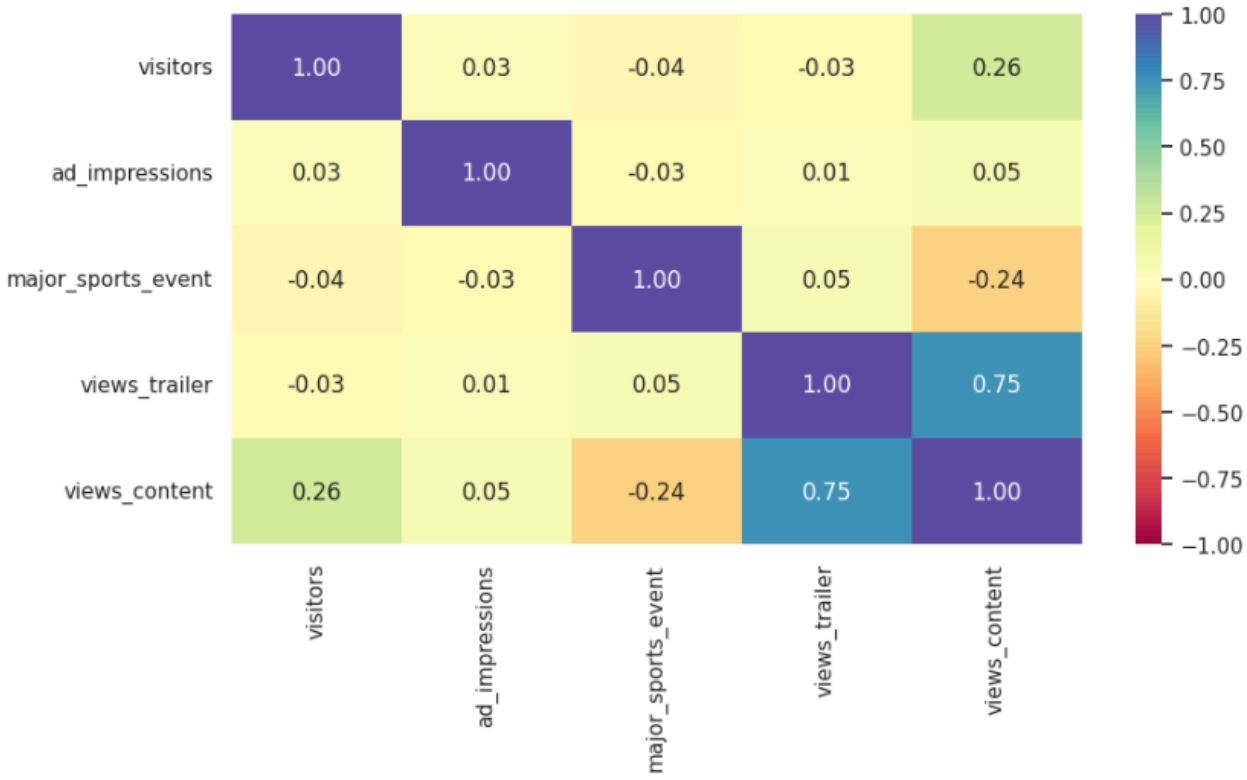


Fig.2.1 : Correlation

- Positive correlation between Trailer views and content views, implying that those who have watched the trailer have mostly watched the content.
- Mild correlation between visitors and the First-day view of the content.

2.2. Bivariate Genre and Content Views

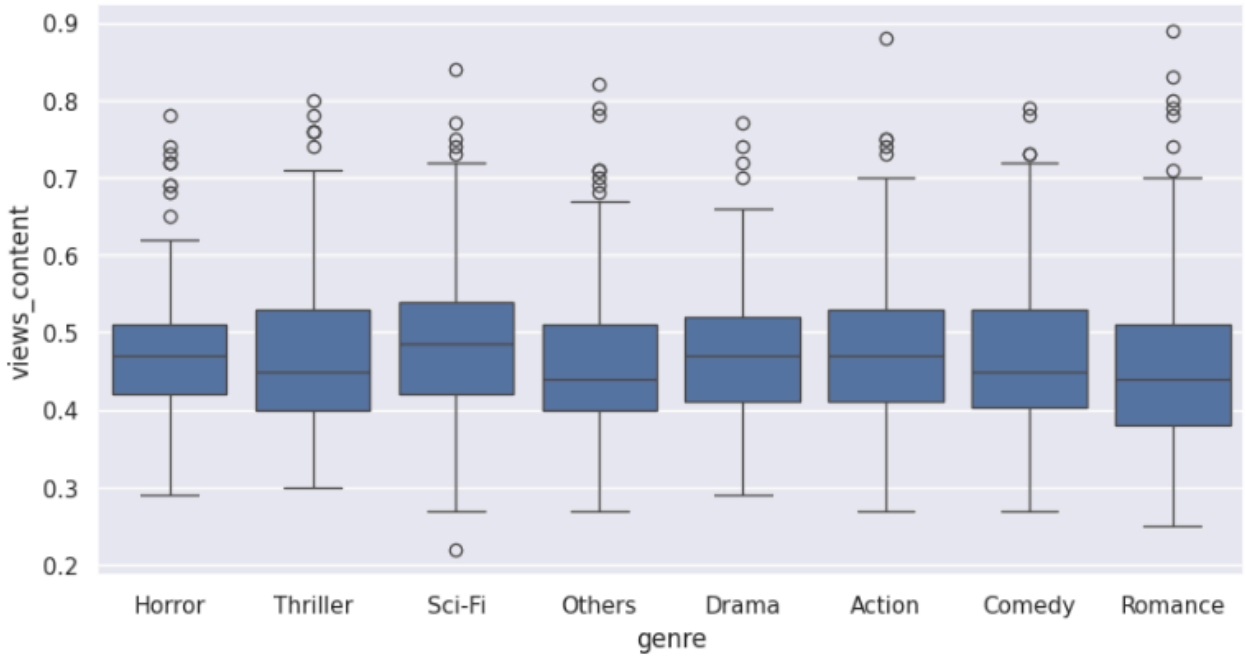


Fig.2.2: Genre Vs Content Views

- Outliers are present in all genres.
- The average views range from 0.4 to 0.5 million.

2.3. Bivariate Genre and Trailer Views

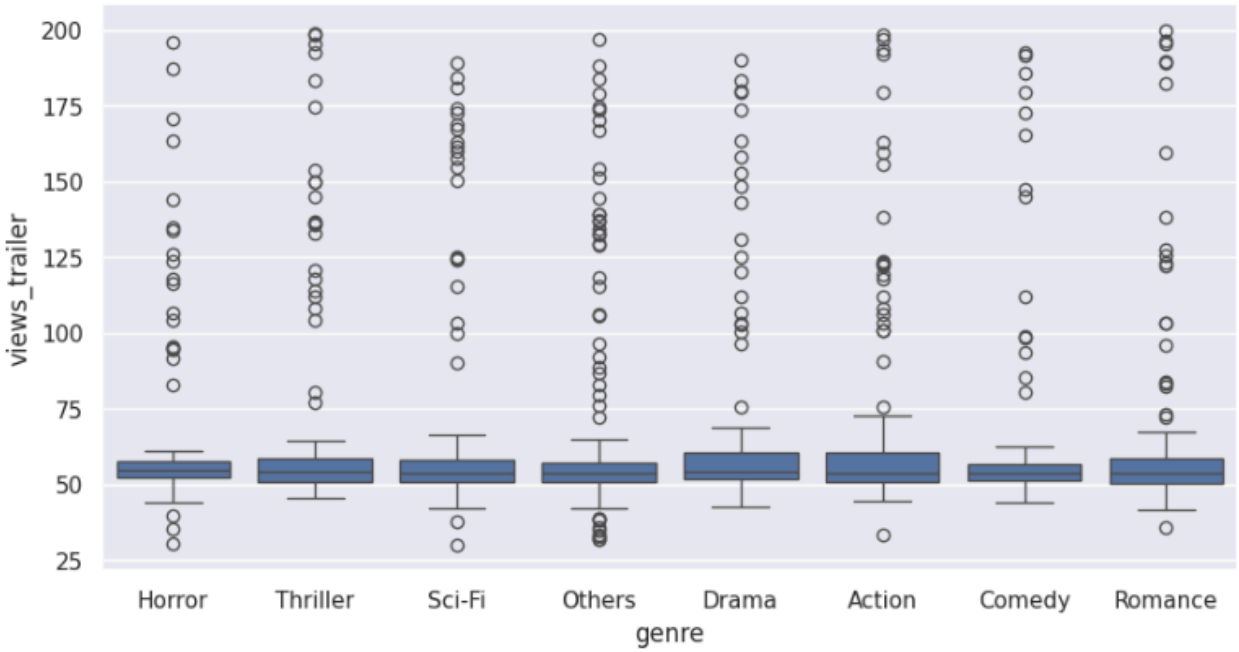


Fig.2.3: Genre Vs Trailer Views

- Outliers are present in all genres.
- There is a mean of 53-55 million views in trailers for all genres.

2.4. Bivariate Season and Content Views

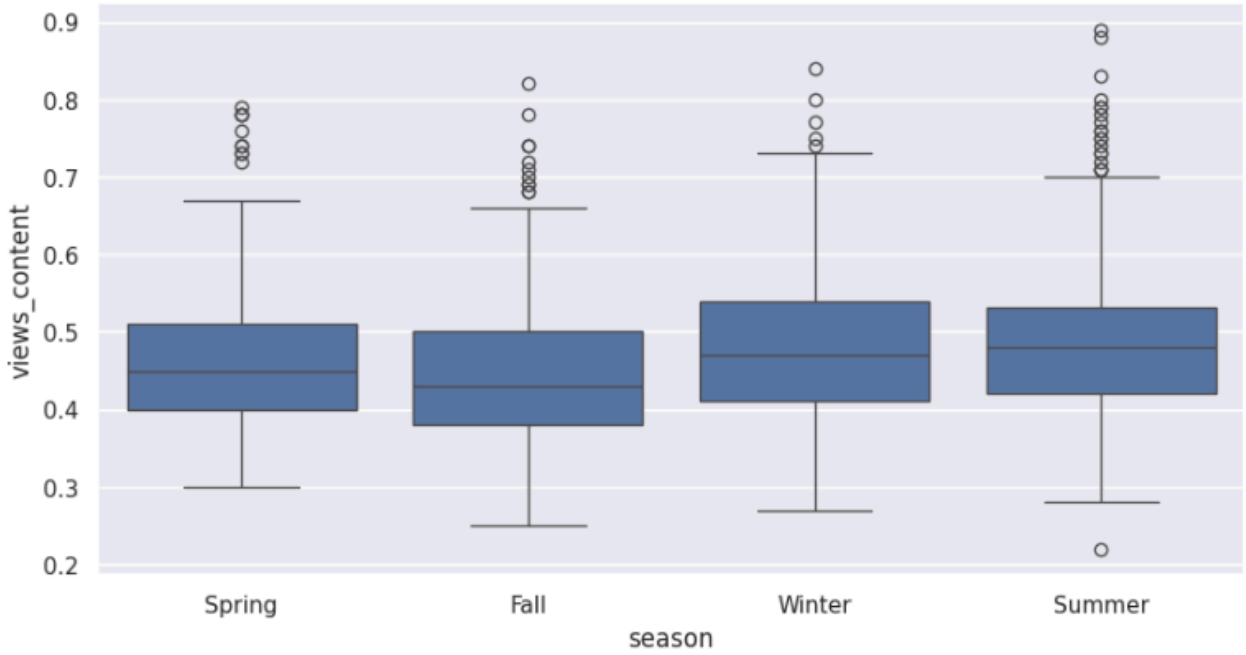


Fig.2.4: Season Vs Content Views

- Outliers are present in all seasons.
- There is a mean of 0.4-0.5 million views in trailers for all genres.
- In Winter and Summer content view mean is more than in Fall and Spring.

2.5. Bivariate Major Sports Events and Content Views

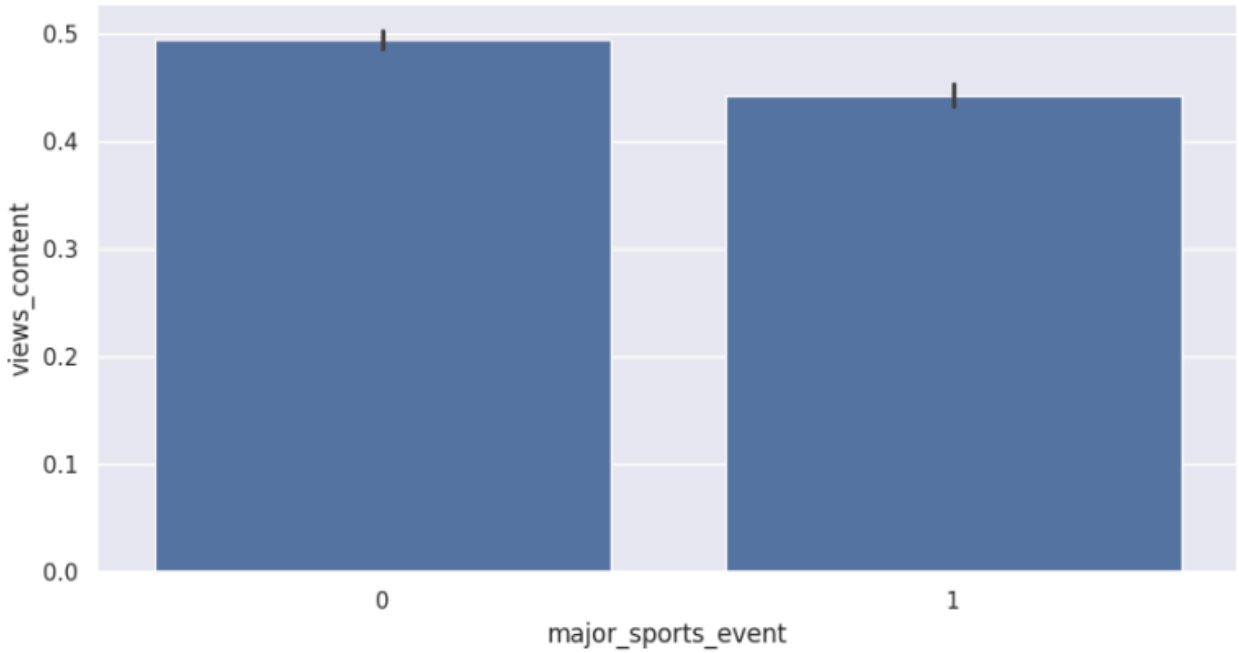


Fig.2.5: Major Sports Event Vs Content Views

- A significant decrease in average content viewership can be observed due to sports events, with 0.49 million views on non-sports event days compared to 0.44 million views on sports event days.

Key Questions

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

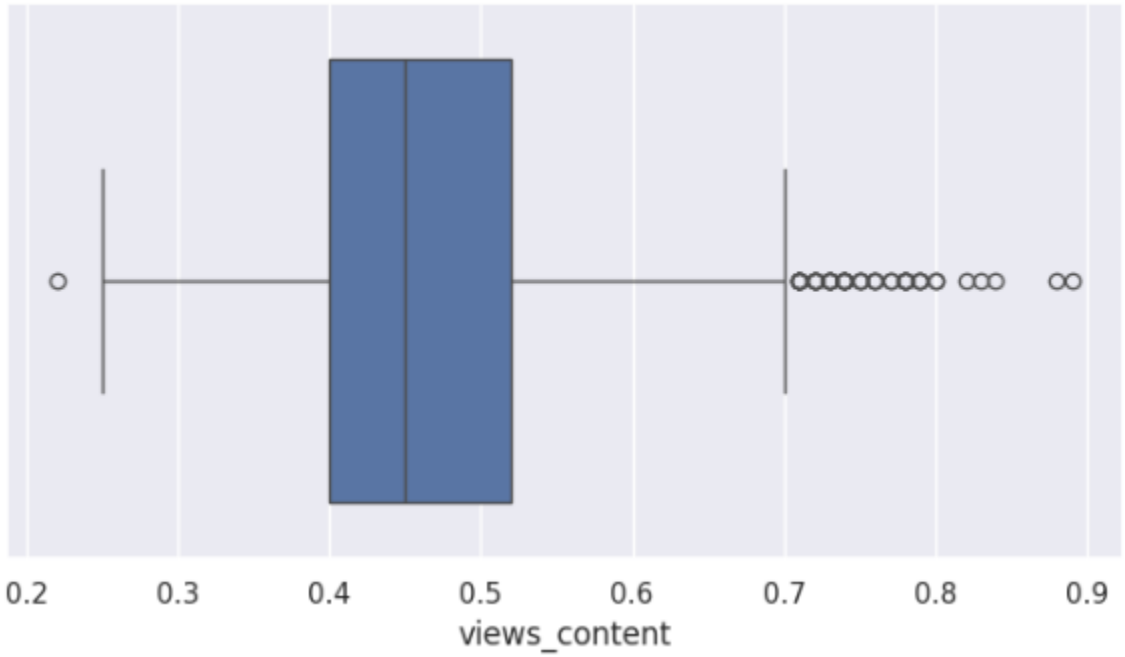


Fig.3: Univariate Analysis of the Distribution of Views of Content (Boxplot)

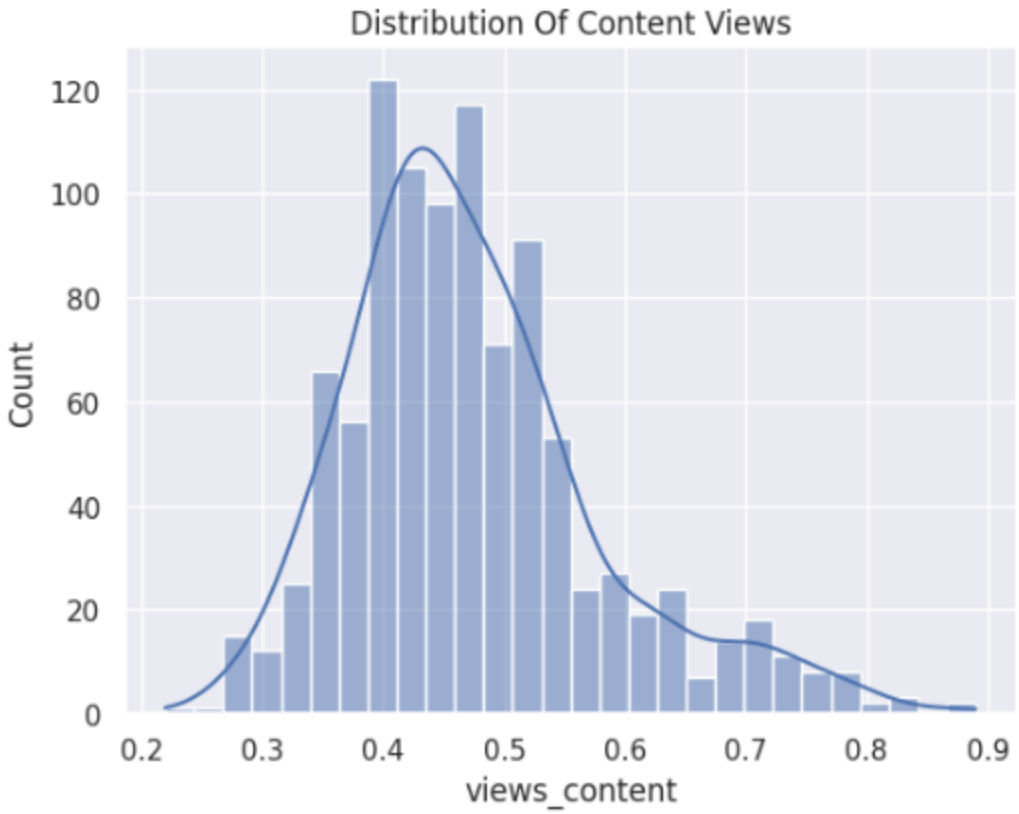


Fig.4: Univariate Analysis of Distribution of Views Content (Histogram)

- A Few Outliers can be observed in the views content field.
- The histogram is right-skewed; it seems almost like a normal distribution.
- An average of less than 0.5 million watched the actual content.

2. What does the distribution of genres look like?

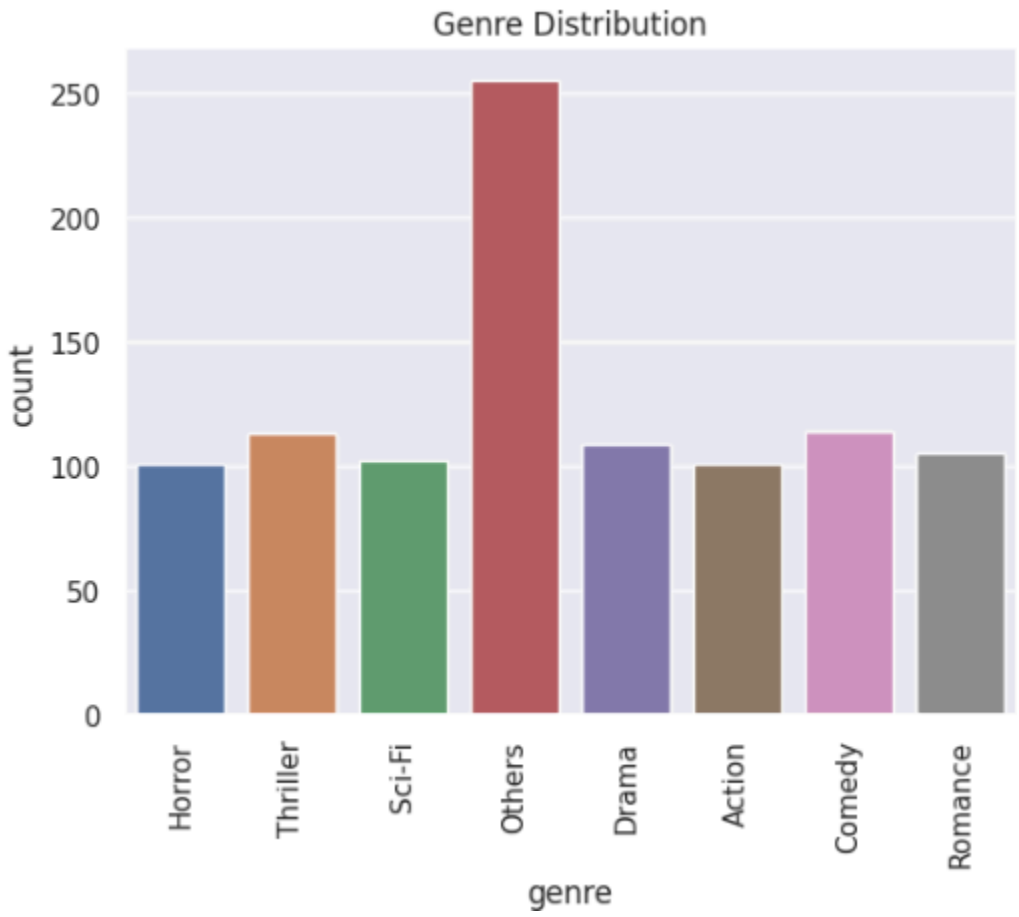


Fig.5: Distribution of Genres

- The count plot reveals that Thriller is the most popular genre, followed by Sci-Fi and Horror.

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?

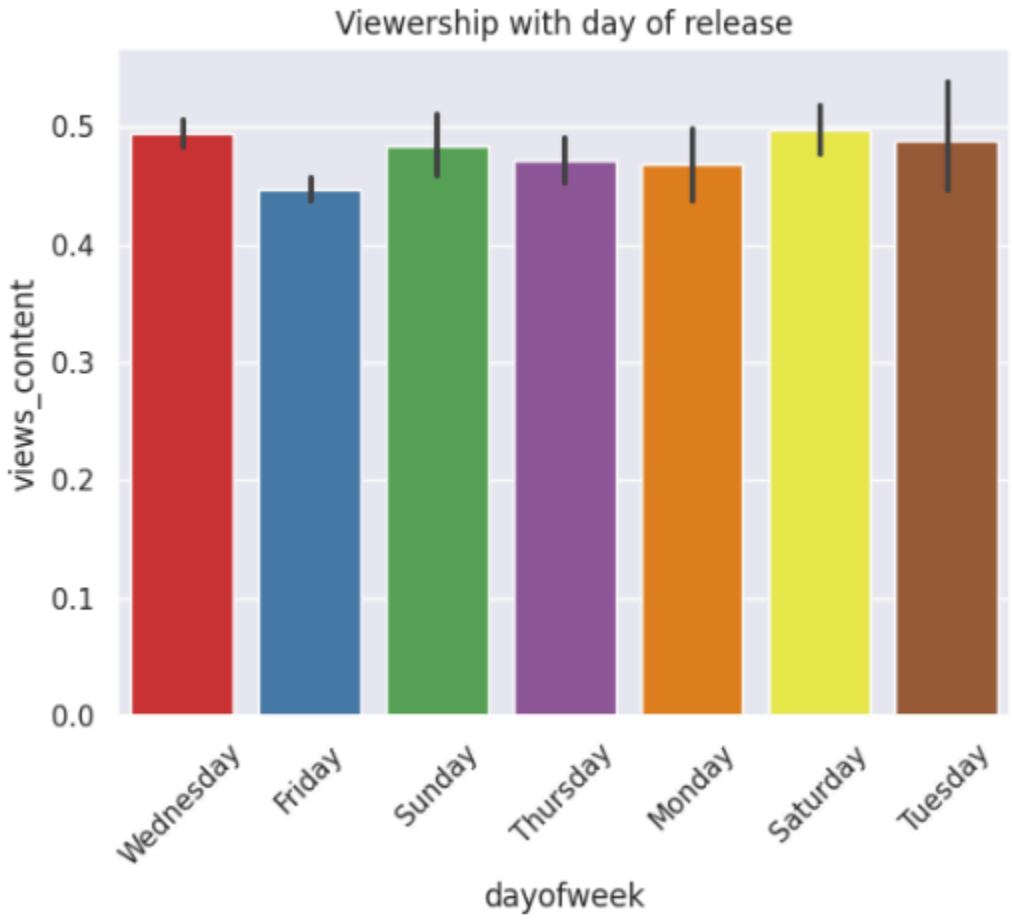


Fig.6: Viewership with day of release

- It is observed that the content views are more when released on Wednesday, and lower if released on Friday.

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

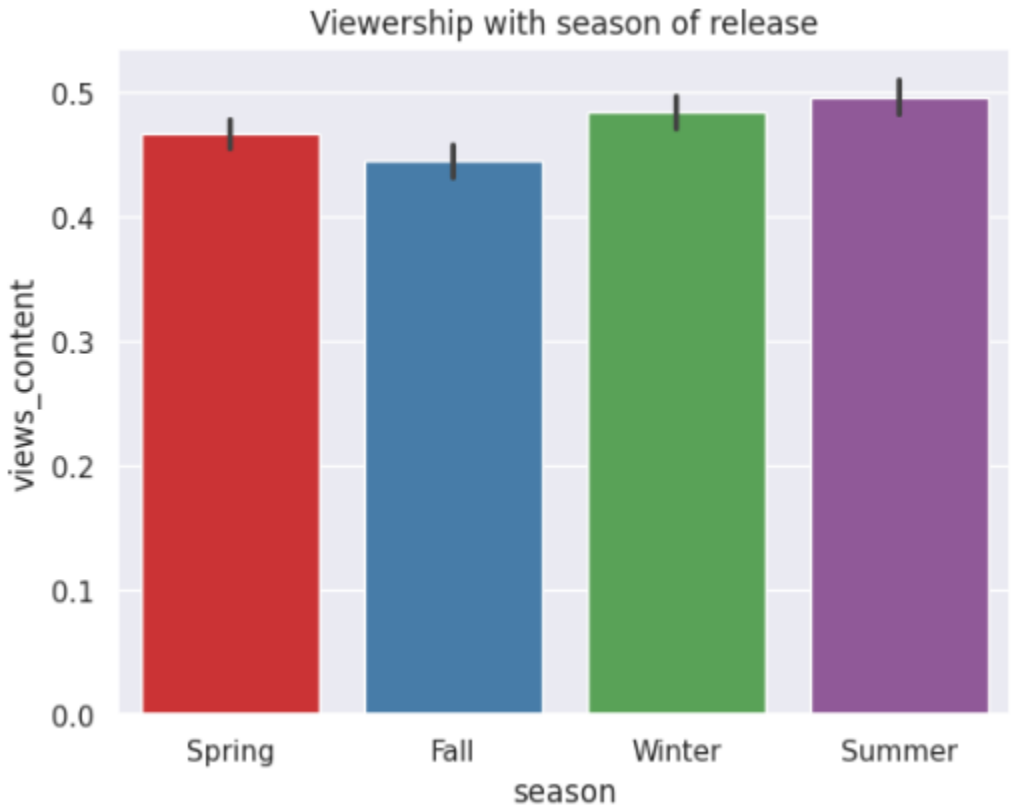


Fig.7: Viewership with season of release

- It is observed that the content views are higher during the Summer season and lower during the Fall.

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

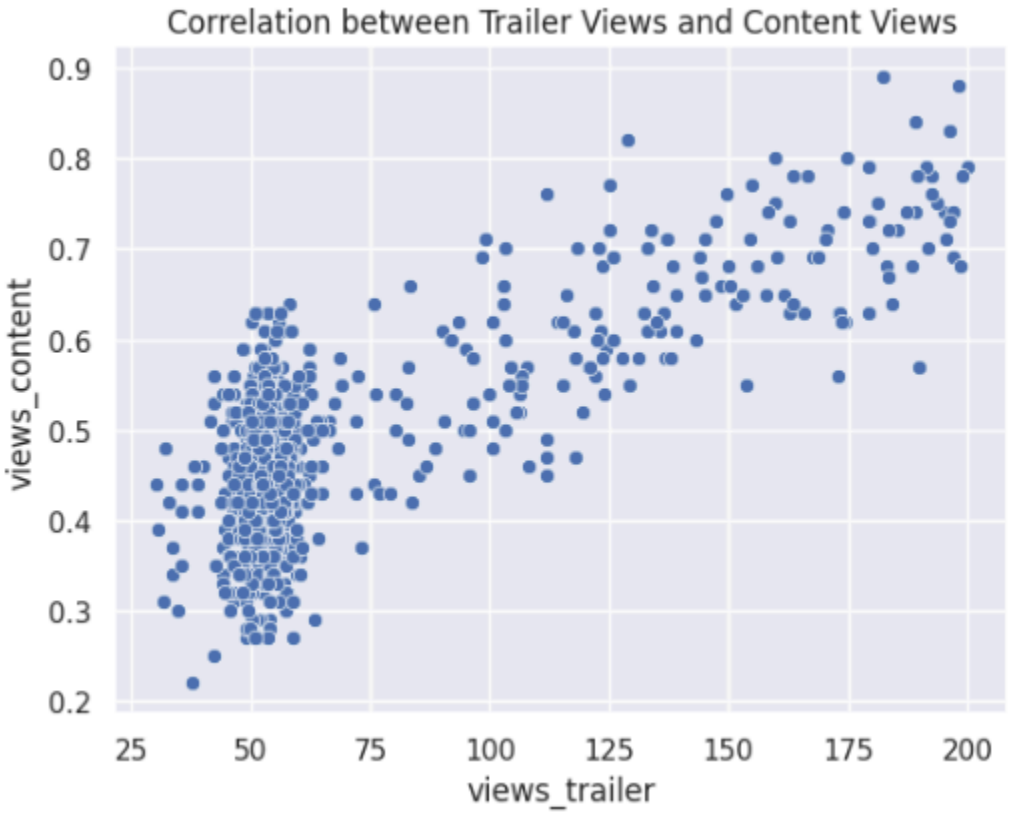


Fig.8: Correlation between Trailer Views and Content Views

- There is a positive relationship between Trailer views and First-day views, which means that when there is an increase in Trailer views, content views increase correspondingly.

Data Processing

1. Duplicate value check

```
np.int64(0)
```

- There are no duplicate values.

2. Missing Value Treatment

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

- There are no missing values found in the dataset.

3. Outlier Treatment

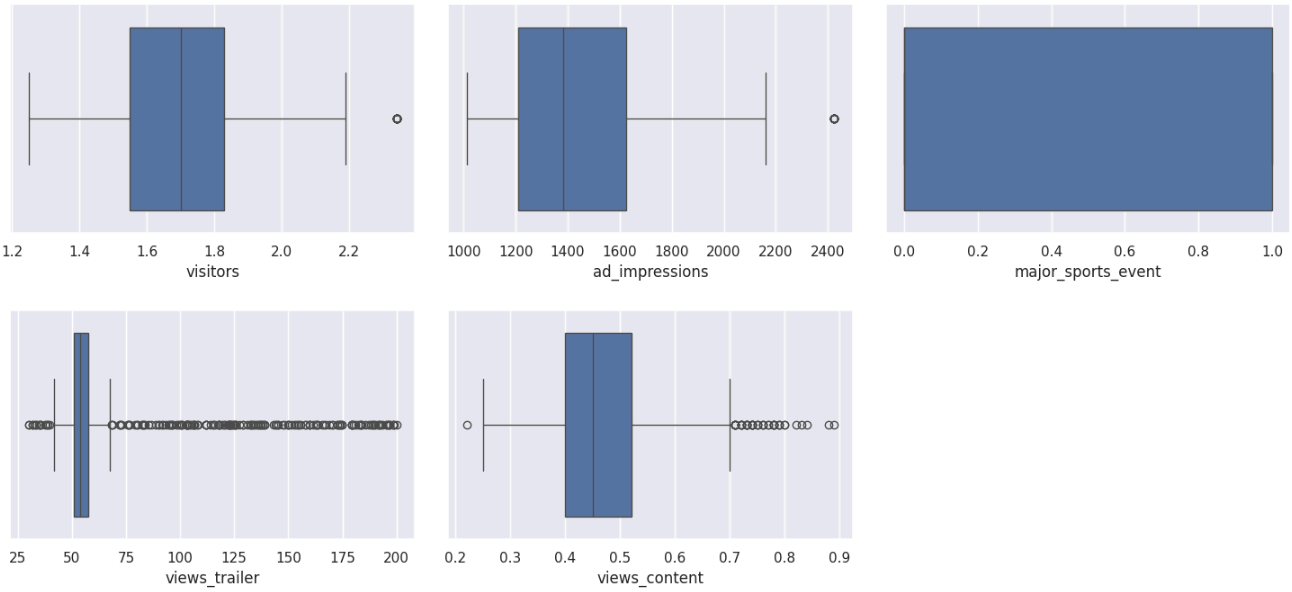


Fig.9: Outlier Treatment

- We observe several outliers in Trailer views and Content views. These will be retained in the dataset, as they may represent rare but significant events, such as blockbuster releases that generate exceptionally high viewership compared to the average. Since these outliers provide valuable insights, they are not treated or removed from the data.

4. Feature engineering

	visitors	ad_impressions	major_sports_event	views_trailer	views_content	genre_Comedy	genre_Drama	genre_Horror	genre_Others	genre_Romance	...	genre_Thriller
0	1.67	1113.81	0	56.70	0.51	False	False	True	False	False	...	False
1	1.46	1498.41	1	52.69	0.32	False	False	False	False	False	...	True
2	1.47	1079.19	1	48.74	0.39	False	False	False	False	False	...	True
3	1.85	1342.77	1	49.81	0.44	False	False	False	False	False	...	False
4	1.46	1498.41	0	55.83	0.46	False	False	False	False	False	...	False
...	...	...	...	...	...	...	...	...	...	...	...	...
995	1.58	1311.96	0	48.58	0.36	False	False	False	False	True	...	False
996	1.34	1329.48	0	72.42	0.56	False	False	False	False	False	...	False
997	1.62	1359.80	1	150.44	0.66	False	False	False	False	False	...	False
998	2.06	1698.35	0	48.72	0.47	False	False	False	False	True	...	False
999	1.36	1140.23	0	52.94	0.49	True	False	False	False	False	...	False

1000 rows × 21 columns

dayofweek_Monday	dayofweek_Saturday	dayofweek_Sunday	dayofweek_Thursday	dayofweek_Tuesday	dayofweek_Wednesday	season_Spring	season_Summer	season_Winter
False	False	False	False	False	True	True	False	False
False	False	False	False	False	False	False	False	False
False	False	False	False	False	True	False	False	False
False	False	False	False	False	False	False	False	False
False	False	True	False	False	False	False	False	True
...	...	...	...	...	...	...	...	...
False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	True	False
False	False	False	False	False	True	False	False	False
True	False	False	False	False	False	False	True	False
False	True	False	False	False	False	False	True	False

- Applied one-hot encoding to the categorical columns to create dummy variables, enabling clearer representation and analysis of each category.

5. Data preparation for modeling

- Split the data into train and test to be able to evaluate the model that we'll build on the train data.
- Converted the input attributes into float type for modeling
- Split the data in a 70:30 ratio for train and test data.
- Number of rows in train data = 700
- Number of rows in test data = 300

Model building - Linear Regression

- **Build the model and comment on the 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:                Wed, 07 May 2025   Prob (F-statistic):          1.32e-215
Time:                        19:04:35   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
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.67e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpreting the Regression Results:

1. **Adjusted. R-squared:** It reflects the fit of the model.

    - Adjusted R-squared values generally range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
    - In our case, the value for adj. R-squared is 0.785, which is good.

2. ***const* coefficient:** It is the Y-intercept.

    - It means that if all the predictor variable coefficients are zero, then the expected output (i.e., Y) would be equal to the const coefficient.
    - In our case, the value for the const. coefficient is 0.0602.

3. **Coefficient of a predictor variable:** It represents the change in the output Y due to a change in the predictor variable (everything else held constant).

    - In our case, the coefficient of visitors is 0.1295.

Model Performance:

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.04853	0.038197	0.791616	0.785162	8.55644

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.050603	0.040782	0.766447	0.748804	9.030464

Observations:

- The training R2 is 0.79, so the model is not underfitting.
- The train and test RMSE and MAE are comparable, so the model does not overfit either.
- MAE suggests that the model can predict anime ratings within a mean error of 0.40 on the test data.
- MAPE of 9.03 on the test data means that we can predict within 9.03% of the views of the content.

- **Display model coefficients with column names**

.

	coef
const	0.060157
visitors	0.129451
ad_impressions	0.000004
major_sports_event	-0.060326
views_trailer	0.002330
genre_Comedy	0.009352
genre_Drama	0.012625
genre_Horror	0.009862
genre_Others	0.006325
genre_Romance	0.000551
genre_Sci-Fi	0.013143
genre_Thriller	0.008708
dayofweek_Monday	0.033662
dayofweek_Saturday	0.057887
dayofweek_Sunday	0.036321
dayofweek_Thursday	0.017289
dayofweek_Tuesday	0.022837
dayofweek_Wednesday	0.047376
season_Spring	0.022602
season_Summer	0.044203
season_Winter	0.027161

Testing the assumptions of linear regression model

- **Perform tests for the assumptions of the linear regression**

1. **No Multicollinearity**

	feature	VIF
0	const	99.679317
1	visitors	1.027837
2	ad_impressions	1.029390
3	major_sports_event	1.065689
4	views_trailer	1.023551
5	genre_Comedy	1.917635
6	genre_Drama	1.926699
7	genre_Horror	1.904460
8	genre_Others	2.573779
9	genre_Romance	1.753525
10	genre_Sci-Fi	1.863473
11	genre_Thriller	1.921001
12	dayofweek_Monday	1.063551
13	dayofweek_Saturday	1.155744
14	dayofweek_Sunday	1.150409
15	dayofweek_Thursday	1.169870
16	dayofweek_Tuesday	1.062793
17	dayofweek_Wednesday	1.315231
18	season_Spring	1.541591
19	season_Summer	1.568240
20	season_Winter	1.570338

- VIF is between 1 and 5, indicating low multicollinearity.
- We can see VIF for genre_Others is more than 2. Let's drop it.

VIF after dropping genre_Others

	feature	VIF
0	const	87.570676
1	visitors	1.022226
2	ad_impressions	1.028804
3	major_sports_event	1.065264
4	views_trailer	1.020524
5	genre_Comedy	1.204848
6	genre_Drama	1.223443
7	genre_Horror	1.204654
8	genre_Romance	1.171988
9	genre_Sci-Fi	1.205594
10	genre_Thriller	1.206560
11	dayofweek_Monday	1.063551
12	dayofweek_Saturday	1.154886
13	dayofweek_Sunday	1.150034
14	dayofweek_Thursday	1.169852
15	dayofweek_Tuesday	1.058831
16	dayofweek_Wednesday	1.314380
17	season_Spring	1.541573
18	season_Summer	1.545311
19	season_Winter	1.568494

- All the variables have VIF less than 2.
- We have dealt with Multicollinearity in the data.
- Let's rebuild the model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     135.8
Date:                Fri, 09 May 2025   Prob (F-statistic):          1.66e-216
Time:                        14:57:12   Log-Likelihood:                 1124.2
No. Observations:                 700   AIC:                            -2208.
Df Residuals:                     680   BIC:                            -2117.
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0660      0.017      3.786      0.000       0.032       0.100
visitors           0.1289      0.008     16.378      0.000       0.113       0.144
ad_impressions  3.482e-06   6.58e-06      0.529      0.597   -9.43e-06    1.64e-05
major_sports_event -0.0604      0.004    -15.307      0.000      -0.068      -0.053
views_trailer      0.0023   5.51e-05     42.213      0.000       0.002       0.002
genre_Comedy       0.0050      0.006      0.789      0.430      -0.007       0.017
genre_Drama        0.0082      0.006      1.270      0.204      -0.004       0.021
genre_Horror       0.0054      0.006      0.835      0.404      -0.007       0.018
genre_Romance     -0.0038      0.007     -0.552      0.581      -0.017       0.010
genre_Sci-Fi       0.0088      0.007      1.326      0.185      -0.004       0.022
genre_Thriller     0.0043      0.006      0.672      0.502      -0.008       0.017
dayofweek_Monday   0.0337      0.012      2.848      0.005       0.010       0.057
dayofweek_Saturday 0.0581      0.007      8.122      0.000       0.044       0.072
dayofweek_Sunday   0.0362      0.008      4.624      0.000       0.021       0.052
dayofweek_Thursday 0.0173      0.007      2.555      0.011       0.004       0.031
dayofweek_Tuesday  0.0236      0.014      1.723      0.085      -0.003       0.050
dayofweek_Wednesday 0.0475     0.004     10.577      0.000       0.039       0.056
season_Spring      0.0226      0.005      4.228      0.000       0.012       0.033
season_Summer      0.0436      0.005      8.063      0.000       0.033       0.054
season_Winter      0.0270      0.005      5.069      0.000       0.017       0.037
==============================================================================
Omnibus:                        4.537   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.103   Jarque-Bera (JB):                4.462
Skew:                           0.154   Prob(JB):                        0.107
Kurtosis:                       3.240   Cond. No.                     1.46e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- After dropping ‘genre_Others’ Adj. R-squared increased by 0.001 and R-squared decreased by 0.001.
- As there is no multicollinearity, we can look at the p-values of predictor variables to check their significance.

- **Dealing with p-value**

- Some of the dummy variables have a p-value > 0.05. So, they are not significant, and we will drop them.
- But sometimes p-values change after dropping a variable. So, we'll not drop all variables at once
- Instead, we will do the following:
    - Build a model, check the p-values of the variables, and drop the column with the highest p-value
    - Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value
    - Repeat the above two steps till there are no columns with p-value > 0.05

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.789
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     233.8
Date:                Fri, 09 May 2025   Prob (F-statistic):          7.03e-224
Time:                        19:30:24   Log-Likelihood:                 1120.2
No. Observations:                 700   AIC:                            -2216.
Df Residuals:                     688   BIC:                            -2162.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0747      0.015      5.110      0.000       0.046       0.103
visitors       0.1291      0.008     16.440      0.000       0.114       0.145
major_sports_event    -0.0606      0.004    -15.611      0.000      -0.068      -0.053
views_trailer  0.0023    5.5e-05     42.414      0.000       0.002       0.002
dayofweek_Monday       0.0321      0.012      2.731      0.006       0.009       0.055
dayofweek_Saturday     0.0570      0.007      8.042      0.000       0.043       0.071
dayofweek_Sunday       0.0344      0.008      4.456      0.000       0.019       0.050
dayofweek_Thursday     0.0154      0.007      2.307      0.021       0.002       0.029
dayofweek_Wednesday    0.0465      0.004     10.532      0.000       0.038       0.055
season_Spring          0.0226      0.005      4.259      0.000       0.012       0.033
season_Summer          0.0434      0.005      8.112      0.000       0.033       0.054
season_Winter          0.0282      0.005      5.362      0.000       0.018       0.039
==============================================================================
Omnibus:                        3.254   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.196   Jarque-Bera (JB):                3.077
Skew:                           0.139   Prob(JB):                        0.215
Kurtosis:                       3.168   Cond. No.                         662.
==============================================================================

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

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.048841	0.038385	0.788937	0.785251	8.595246

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051109	0.041299	0.761753	0.751792	9.177097

Observations :

- Now, no feature has a p-value greater than 0.05, so we'll consider the features in x_train1 as the final set of predictor variables and olsmod2 as the final model to move forward with.

- Now adjusted R-squared is 0.78, i.e., our model is able to explain ~78% of the variance.

- RMSE and MAE values are comparable for train and test sets, indicating that the model is not overfitting.

2. Linearity and Independence of Variables

	Actual Values	Fitted Values	Residuals
731	0.40	0.445434	-0.045434
716	0.70	0.677403	0.022597
640	0.42	0.433999	-0.013999
804	0.55	0.562030	-0.012030
737	0.59	0.547786	0.042214

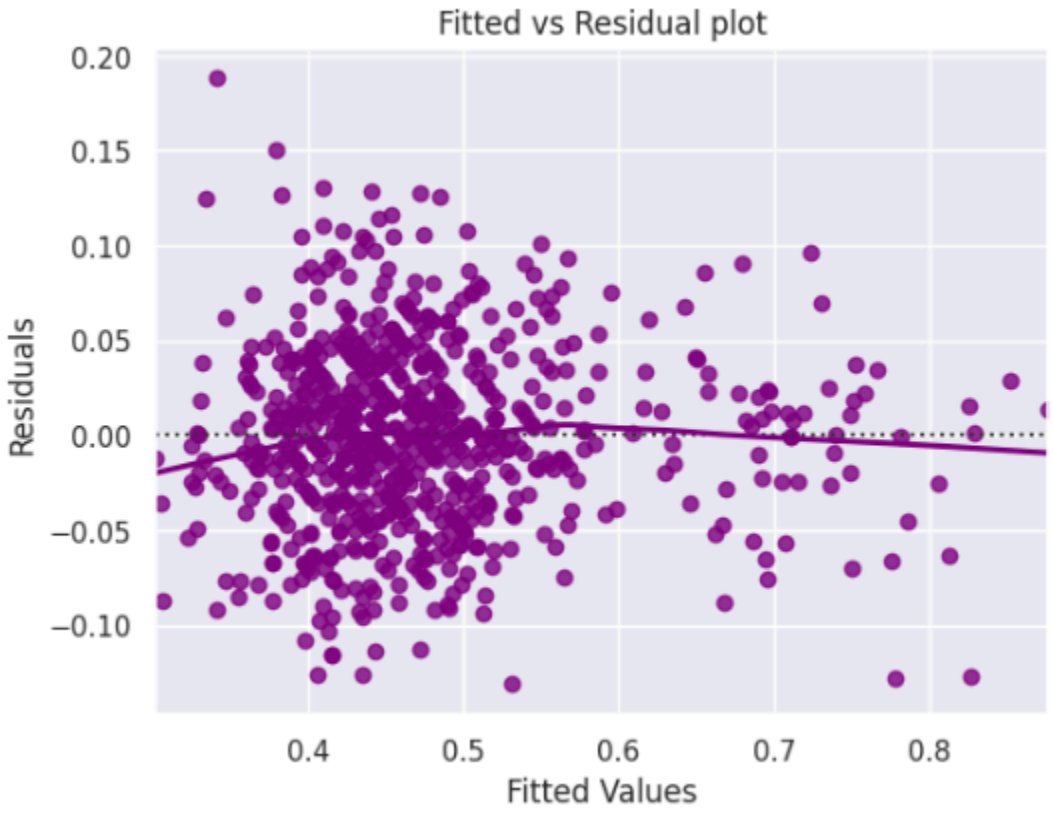


Fig.10: Linearity and Independence of Variables

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

3. Normality of error terms

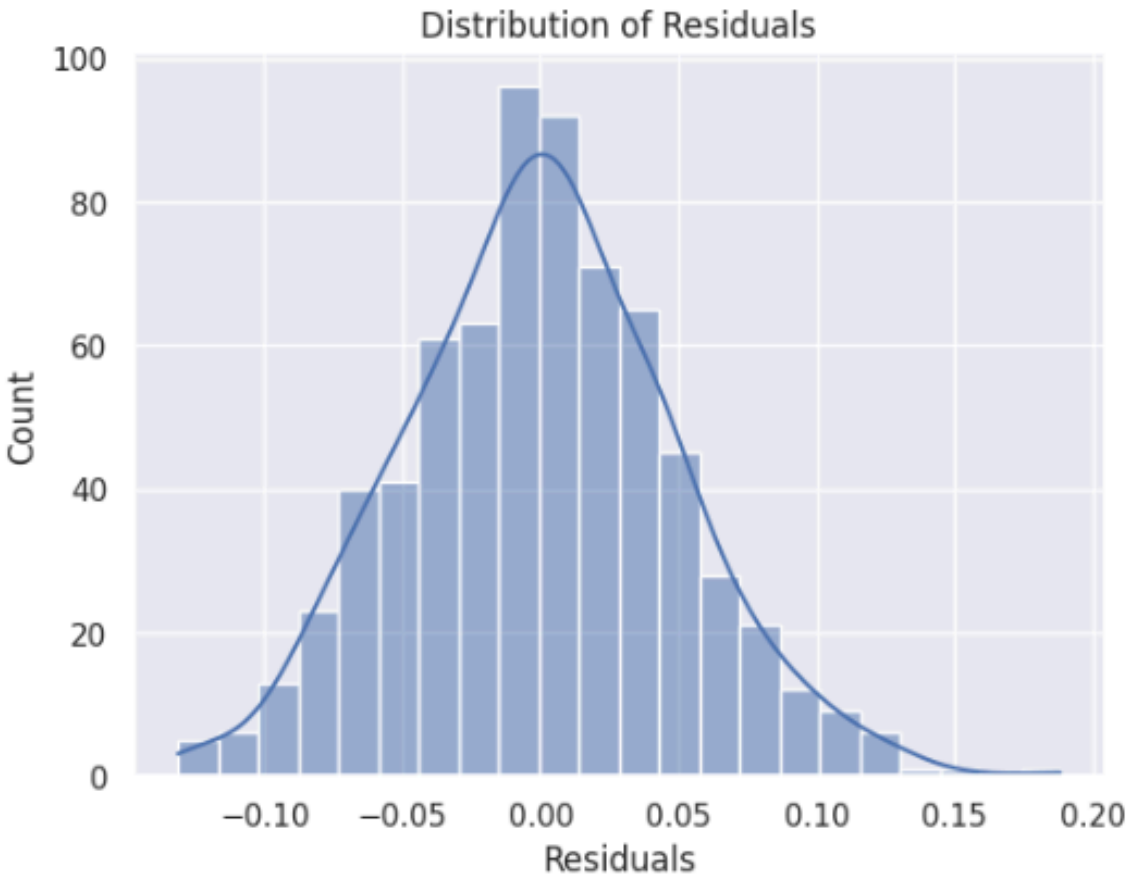


Fig.11: Normality of Residuals

- The residual terms are normally distributed.

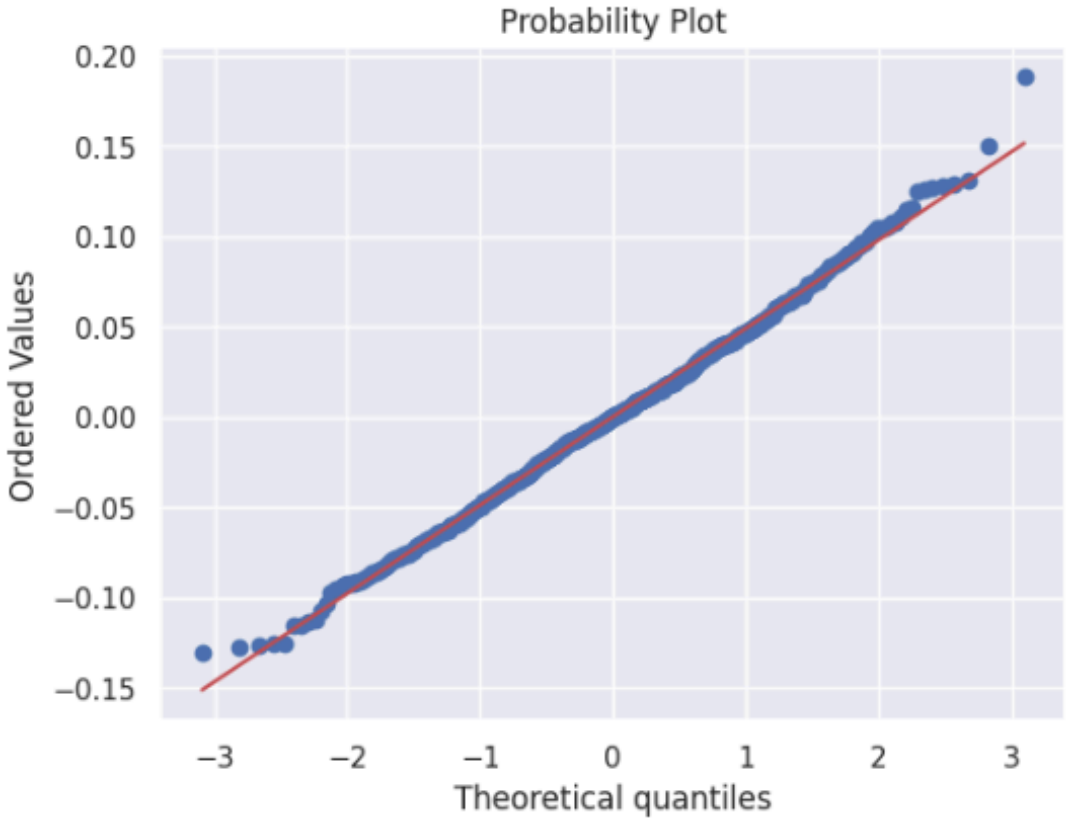


Fig.12: Plot Residuals

- Most of the points lie on the straight line in the Q-Q plot.

The Shapiro-Wilk Test can also be performed to check normality.

  - Null Hypothesis: Data is normally distributed
  - Alternate Hypothesis: Data is not normally distributed

p-value: 0.31085896470071894.

- Since p-value > 0.05, the residuals are normal according to the Shapiro test.

4. Test for Homoscedasticity

p-value: 0.12853551819087372.

- Since p-value > 0.05, the residuals are homoscedastic. So, this assumption is satisfied.

Model performance evaluation

- **Final Model**

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.789
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	233.8
Date:	Sat, 10 May 2025	Prob (F-statistic):	7.03e-224
Time:	11:19:21	Log-Likelihood:	1120.2
No. Observations:	700	AIC:	-2216.
Df Residuals:	688	BIC:	-2162.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0747	0.015	5.110	0.000	0.046	0.103
visitors	0.1291	0.008	16.440	0.000	0.114	0.145
major_sports_event	-0.0606	0.004	-15.611	0.000	-0.068	-0.053
views_trailer	0.0023	5.5e-05	42.414	0.000	0.002	0.002
dayofweek_Monday	0.0321	0.012	2.731	0.006	0.009	0.055
dayofweek_Saturday	0.0570	0.007	8.042	0.000	0.043	0.071
dayofweek_Sunday	0.0344	0.008	4.456	0.000	0.019	0.050
dayofweek_Thursday	0.0154	0.007	2.307	0.021	0.002	0.029
dayofweek_Wednesday	0.0465	0.004	10.532	0.000	0.038	0.055
season_Spring	0.0226	0.005	4.259	0.000	0.012	0.033
season_Summer	0.0434	0.005	8.112	0.000	0.033	0.054
season_Winter	0.0282	0.005	5.362	0.000	0.018	0.039

Omnibus:	3.254	Durbin-Watson:	1.996
Prob(Omnibus):	0.196	Jarque-Bera (JB):	3.077
Skew:	0.139	Prob(JB):	0.215
Kurtosis:	3.168	Cond. No.	662.

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

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.048841	0.038385	0.788937	0.785251	8.595246

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051109	0.041299	0.761753	0.751792	9.177097

- The model can explain ~79% of the variation in the data.
- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.
- The MAPE on the test set suggests we can predict within 9.17% of the First-day views of the content.
- Hence, we can conclude the model olsmodel_final is good for prediction as well as inference purposes.

Conclusions and Recommendations

- The model's R-squared value is approximately 0.786, and the adjusted R-squared is 0.784, indicating that the model can explain about 79% of the variance in the data. This is quite satisfactory.

- Releasing content on specific days of the week will increase viewership. Hence, releasing content on Saturdays and Wednesdays will boost viewership.

- The Genre "Other" category has the highest counts, followed by Comedy, Thriller, Drama, etc.

- Releasing content during the Summer season increases the viewership by 0.0442 unit.

- To improve content viewership, it is recommended to avoid releasing the content on days when major sports events happen.

- More trailer views lead to higher first-day views, showing that strong trailer engagement boosts early content performance.