Predictive Modelling Report

Presented by :
Sanjay Rajan J

TABLE OF CONTENTS

LIST OF FIGURES

1. DATA OVERVIEW

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 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.

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 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.

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

- Shape:

  There are 1000 rows and 8 columns in this dataset.

- Basic Info:

Fig.1 Data Info

```
<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
```

- Null values Check:

Fig.2 Null values check

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

- Numerical Statistics:

Fig.3 Numerical Statistics

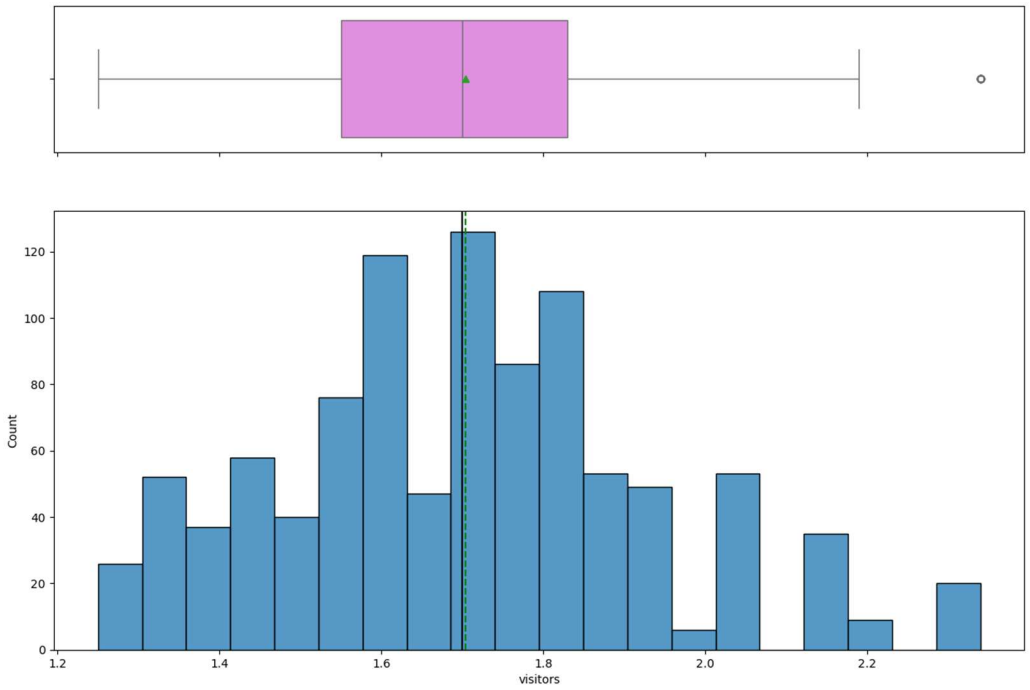
	count	mean	std	min	25%	50%	75%	max
visitors	1000.0	1.70429	0.231973	1.25	1.5500	1.70	1.830	2.34
ad_impressions	1000.0	1434.71229	289.534834	1010.87	1210.3300	1383.58	1623.670	2424.20
major_sports_event	1000.0	0.40000	0.490143	0.00	0.0000	0.00	1.000	1.00
views_trailer	1000.0	66.91559	35.001080	30.08	50.9475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

2. EXPLORATORY DATA ANALYSIS

2.1 UNIVARIATE ANALYSIS:

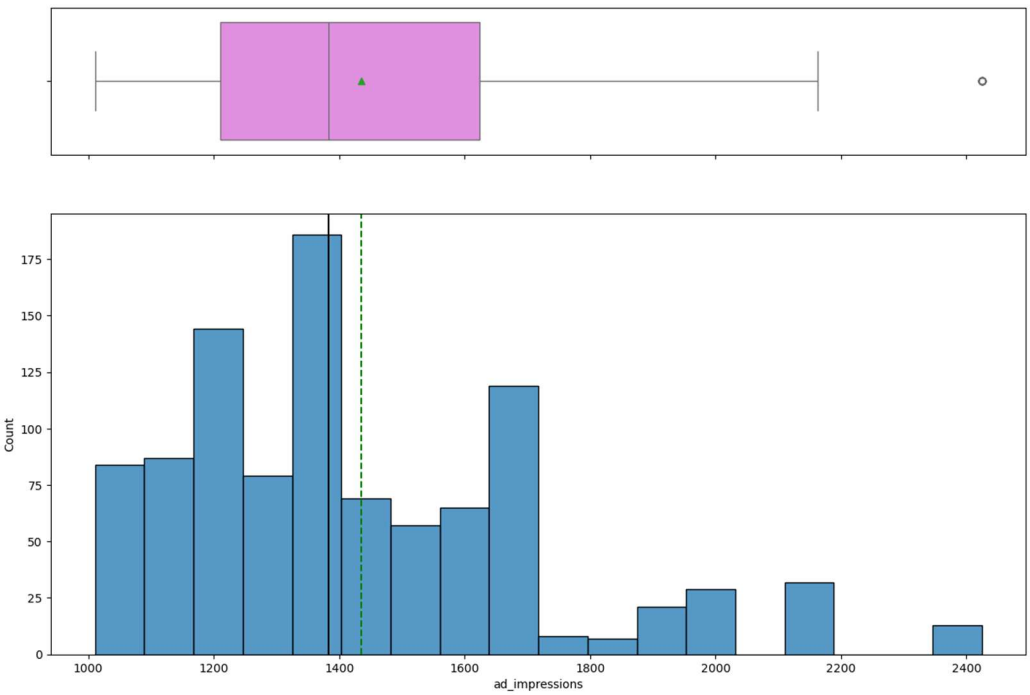
- Distribution of Visitors:

Fig.4 Visitors distribution



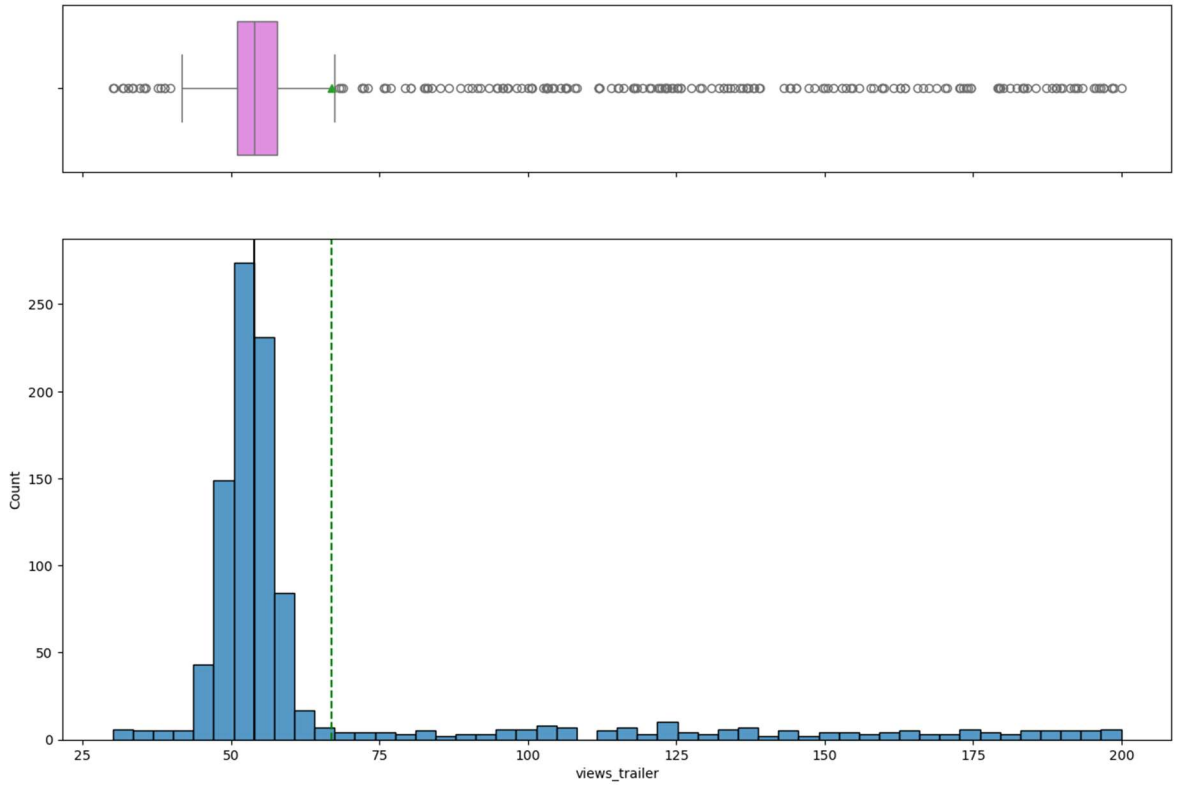
- Distribution of Ad Impressions:

Fig.5 Ad Impressions distribution



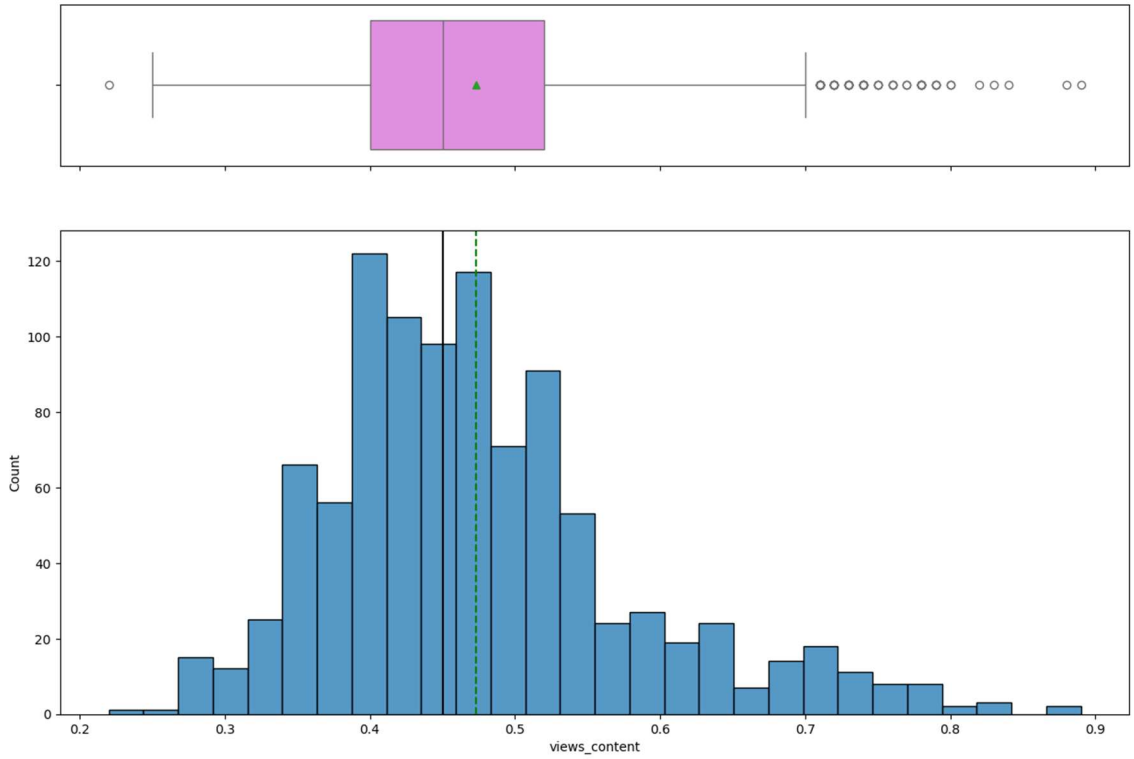
➢ Distribution of Trailer views:

Fig.6 Trailer views distribution



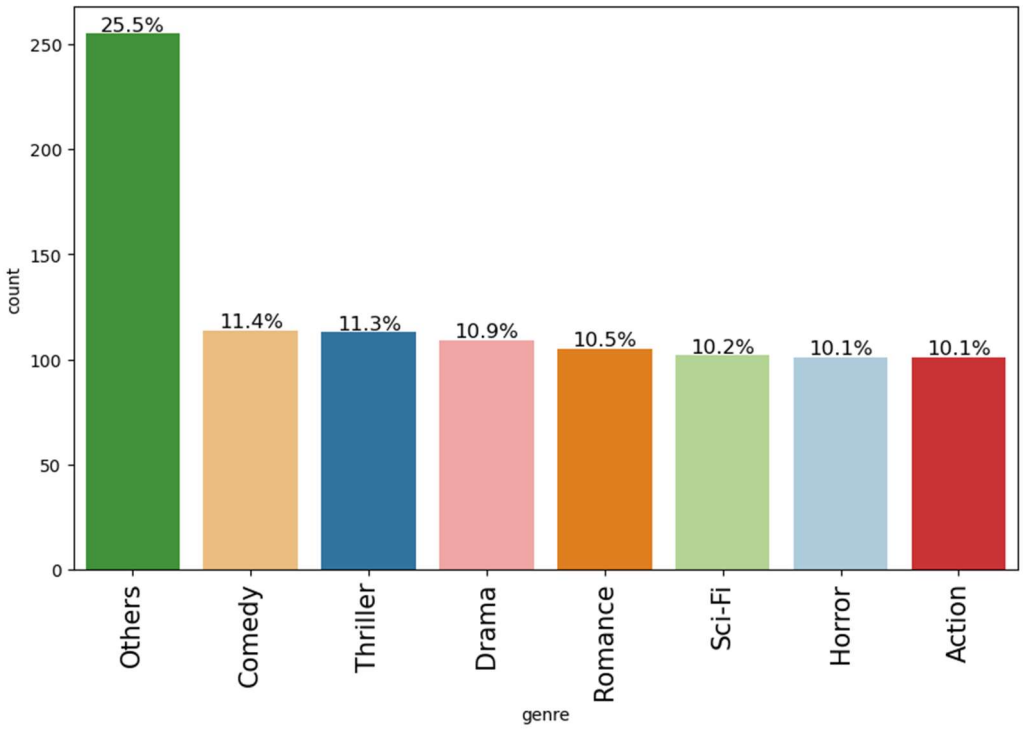
➢ Distribution of Content views:

Fig.7 Content views distribution



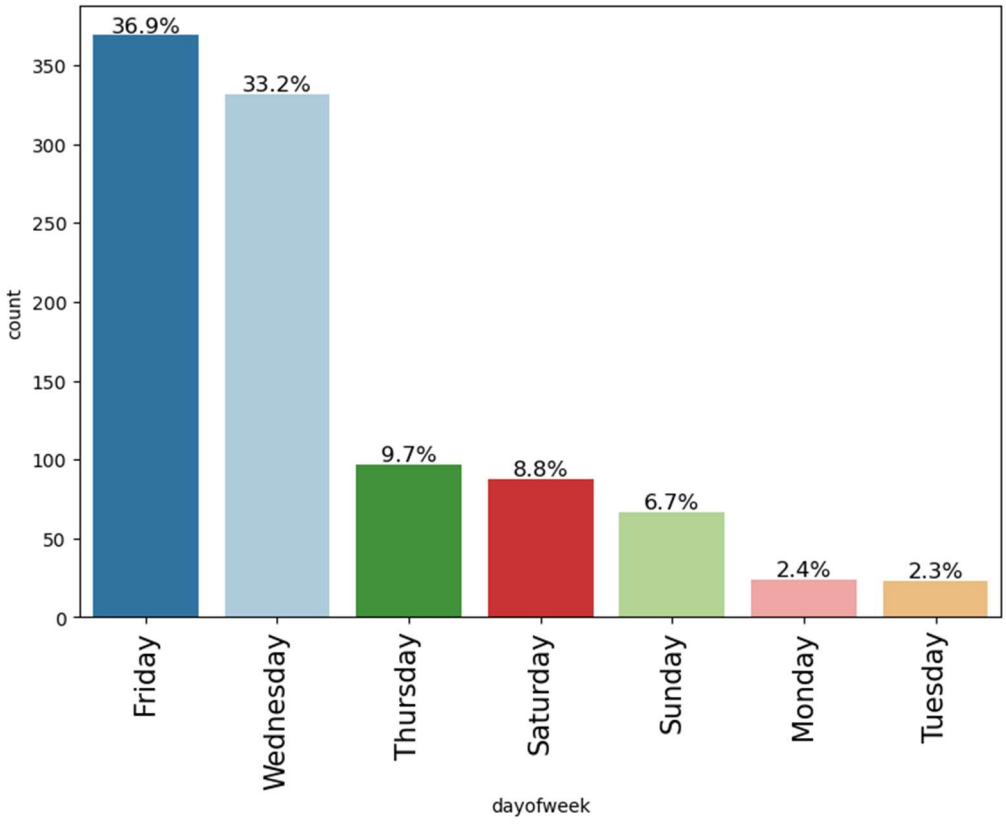
➢ Distribution of Genres:

Fig.8 Genre distribution



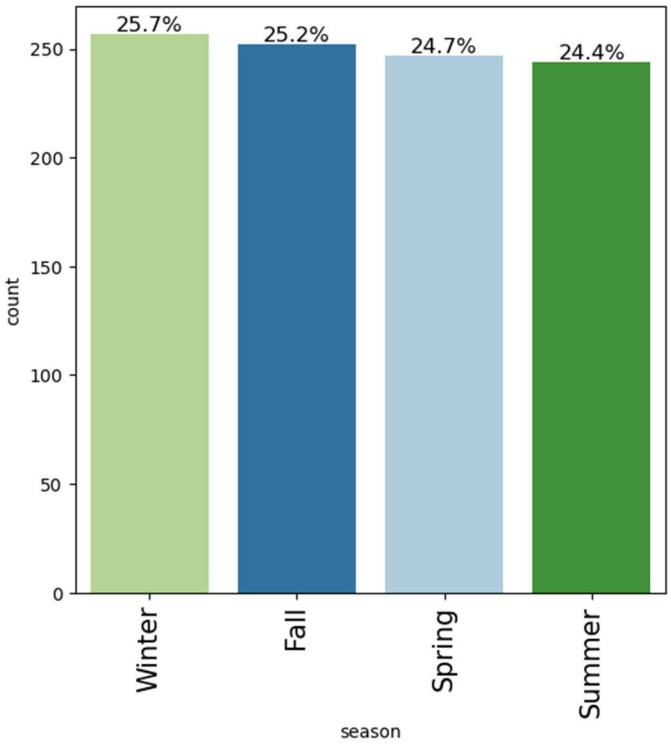
➢ Distribution of Days of week:

Fig.9 Daysofweek distribution



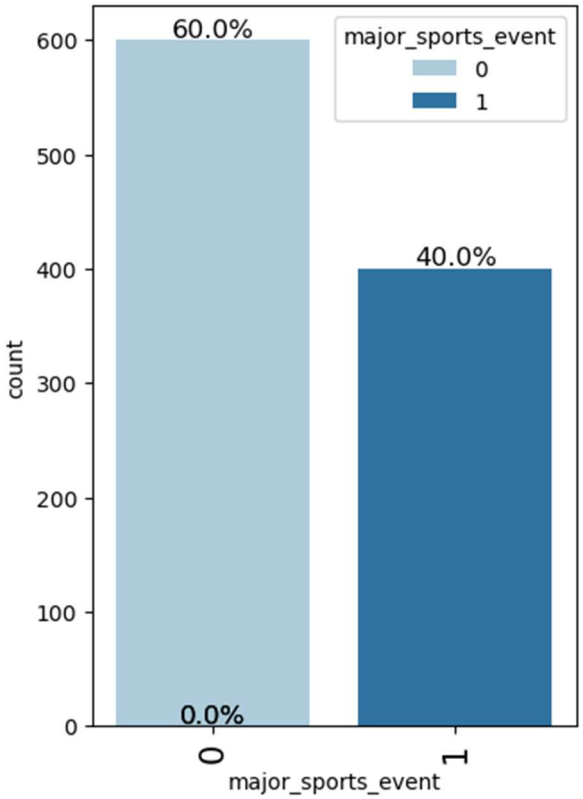
- Distribution of Seasons:

Fig.10 Seasons distribution



- Distribution of Major sport event:

Fig.11 Major sport event distribution



Inferences:

- Distribution of the Ad_impressions is right skewed.
- Majority of the contents are released in Friday followed by Wednesday.
- Tuesday has the least amount of releases.
- The distribution of the seasons is approximately equal.
- The largest category is "Others" with 25.5% of the total, indicating that a significant portion of the dataset includes genres outside the main specified categories.
- The distribution of the views_count approximately follows normal distribution.
- The distribution of the views_trailer is heavily right-skewed and there are too many outliers.
- The distribution of the visitors approximately follows normal distribution with only one outlier.

2.2 BIVARIATE ANALYSIS:

Fig.12 Heatmap

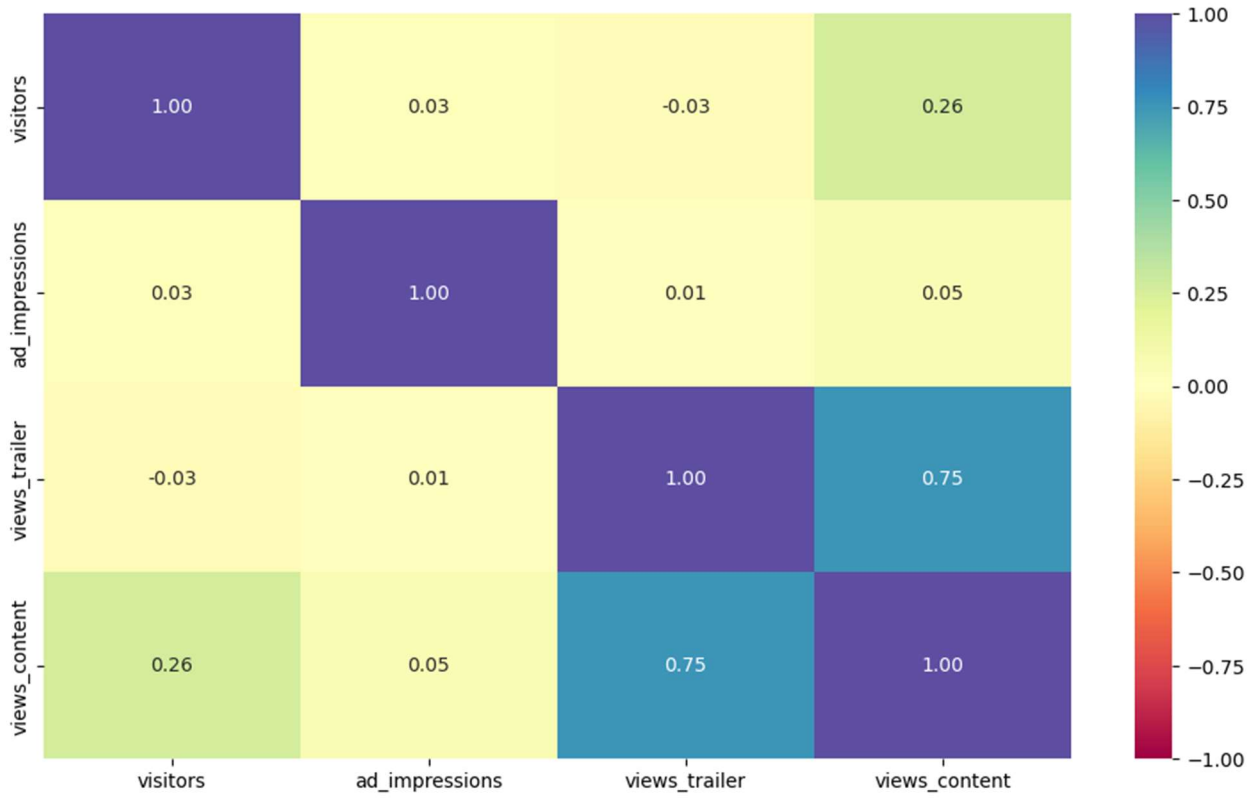


Fig.13 Views_trailer vs Views_Content

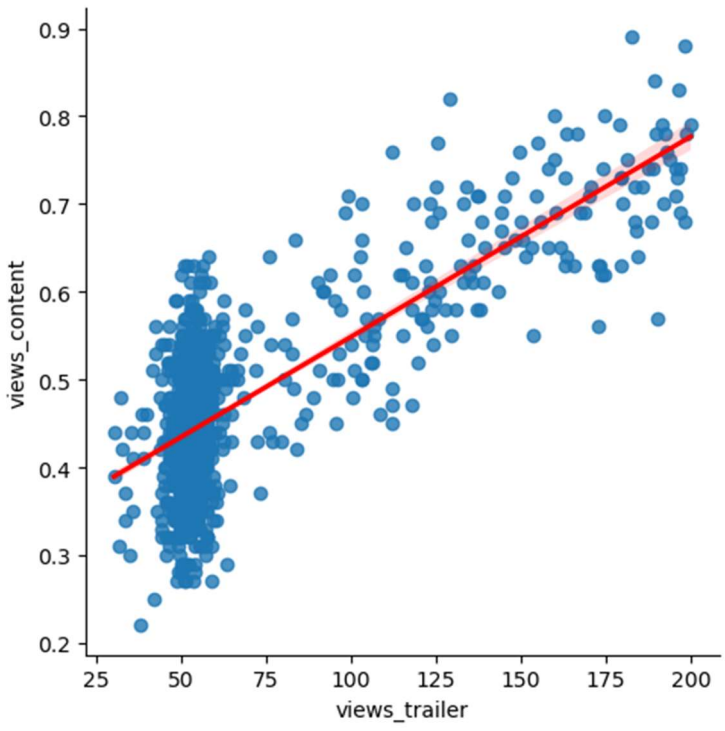


Fig.14 Visitors vs Views_Content

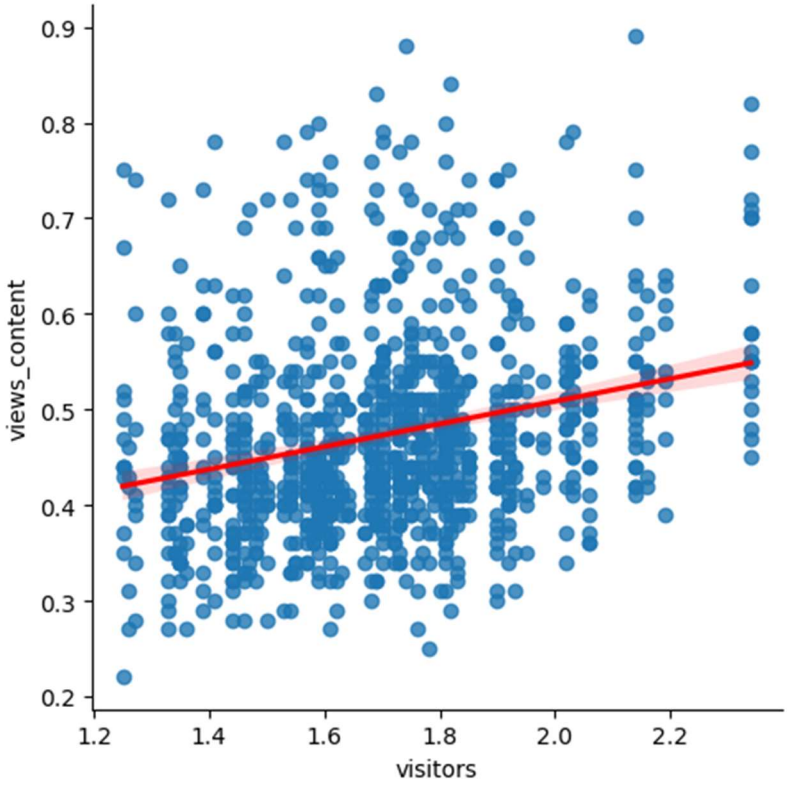


Fig.15 Ad_impressions vs Views_Content

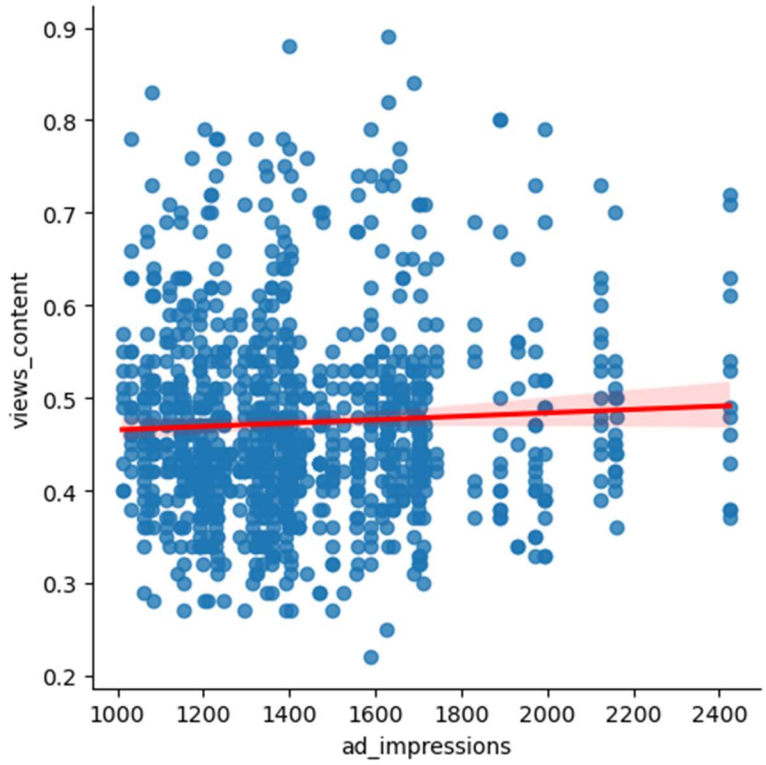


Fig.16 Seasons vs Views_Content

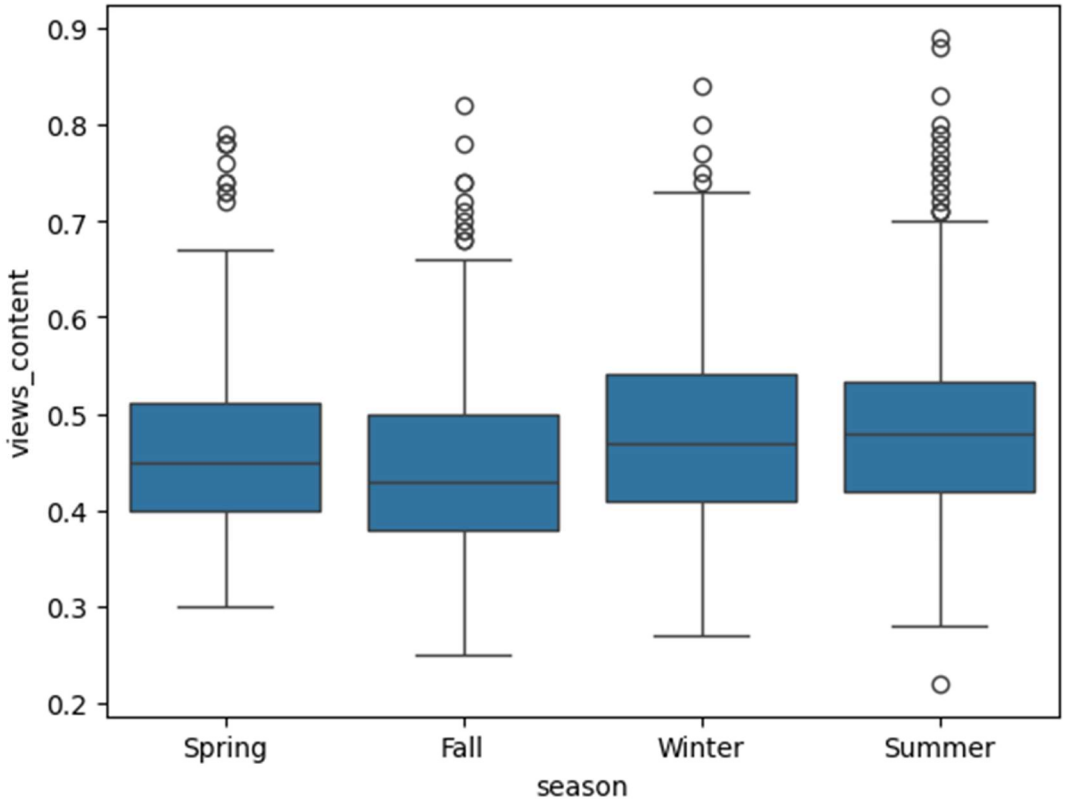


Fig.17 Dayofweek vs Views_Content

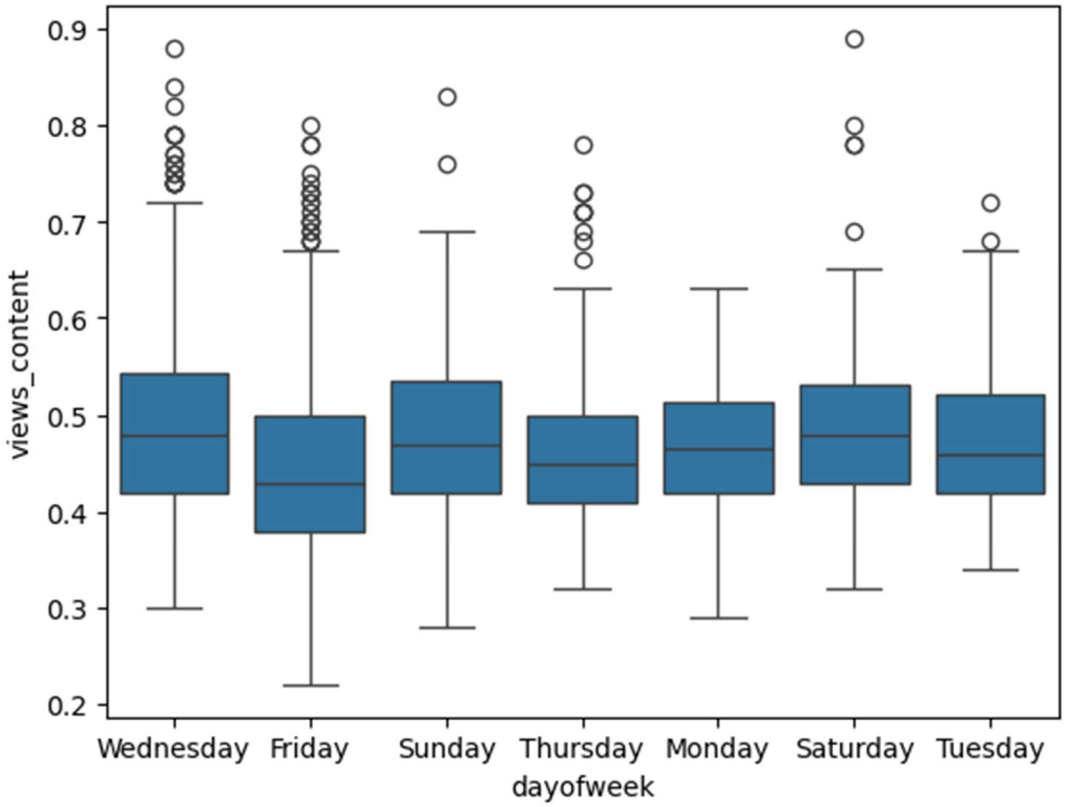


Fig.18 Genre vs Views_Content

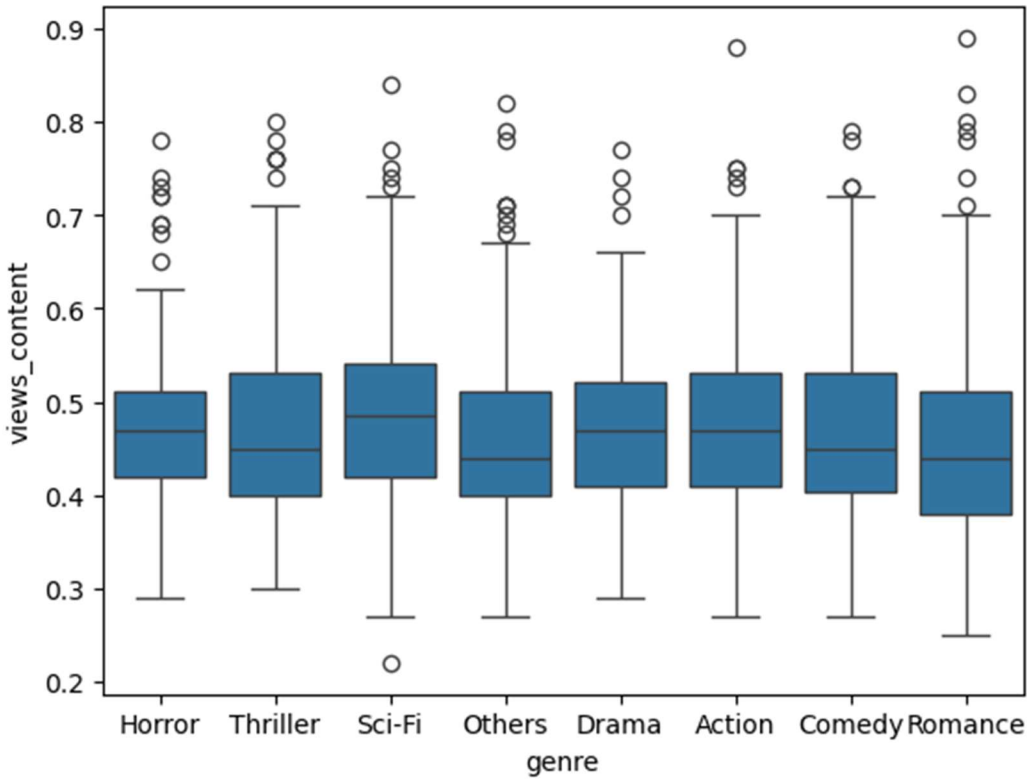


Fig.19 Majort_sports_event vs Views_Content

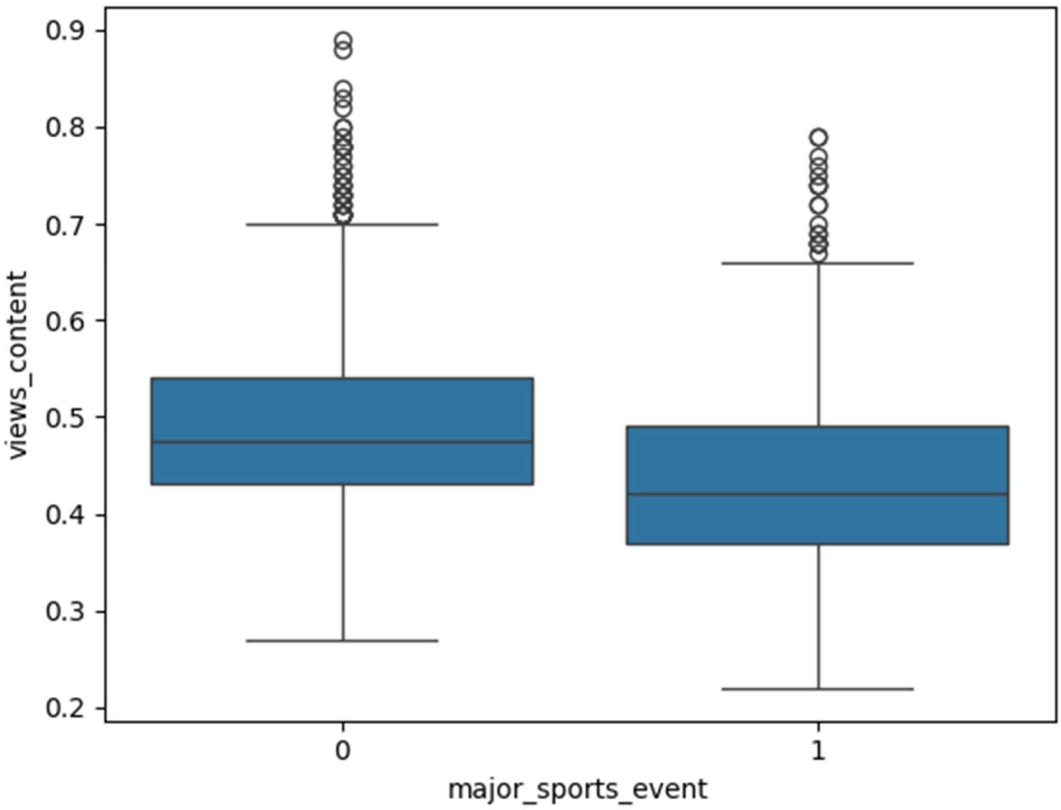


Fig.20 Genre vs Views_Content vs Sports_event

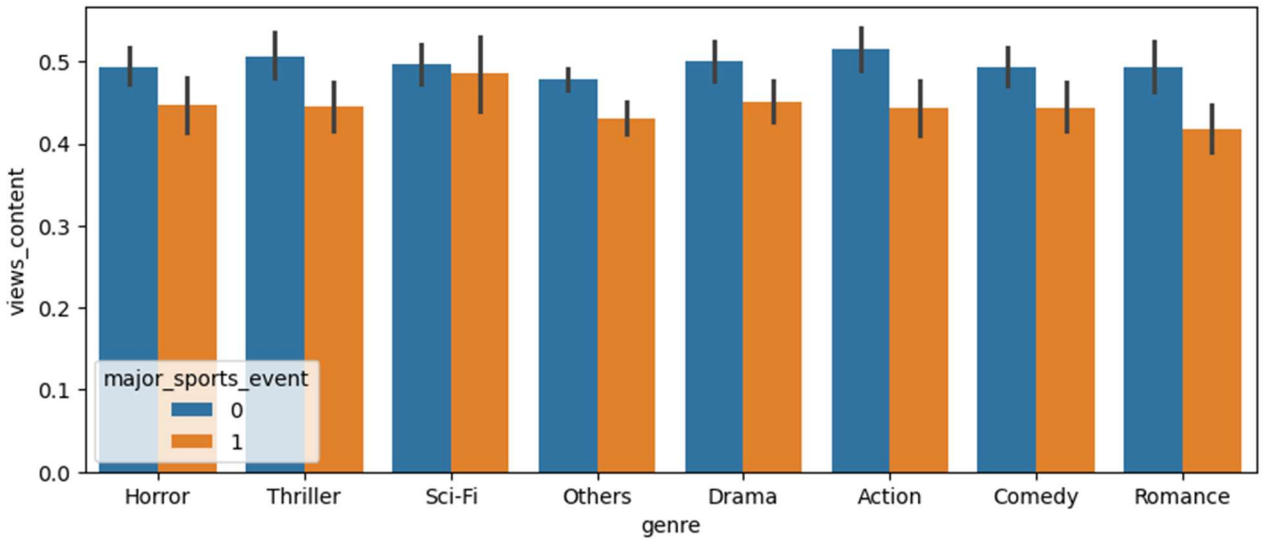
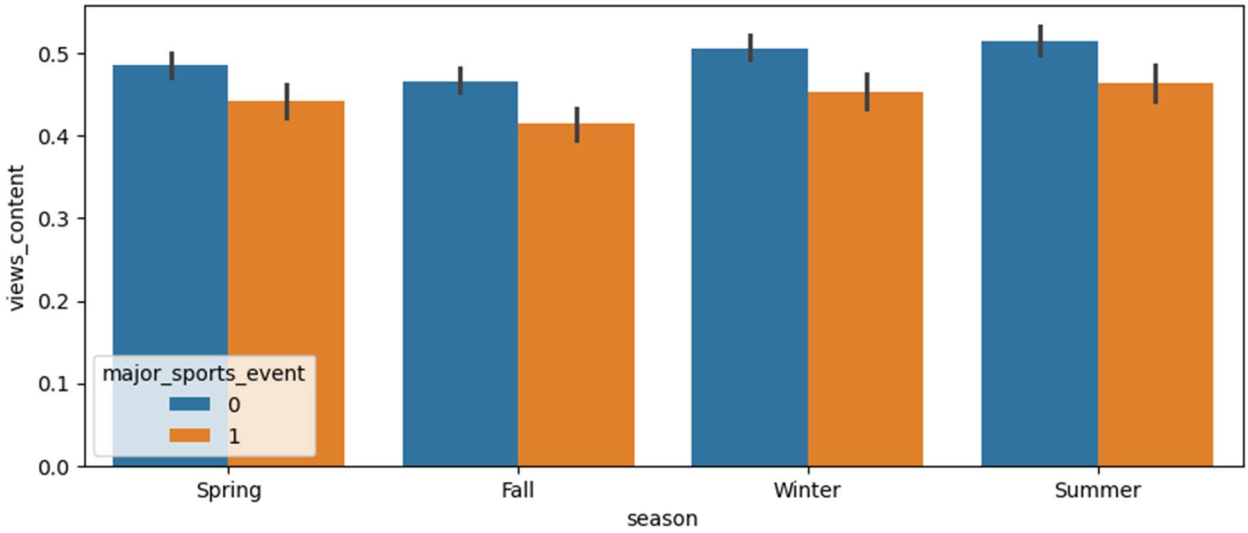


Fig.21 Season vs Views_Content vs Sports_event



Inferences:

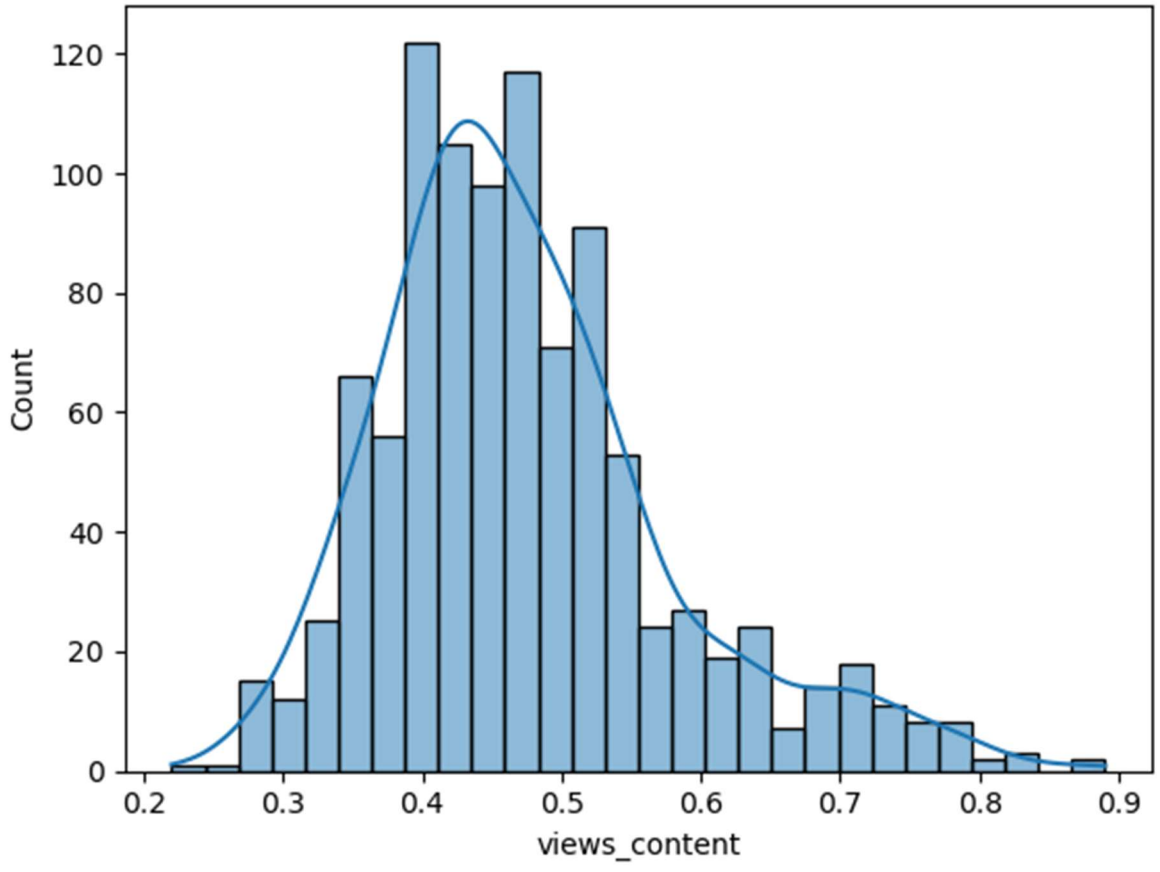
- The scatter plot shows a positive correlation between two variables: “views_trailer” and “views_content”.
- There is a very weak positive relationship between "ad_impressions" and "views_content".
- There is a positive correlation between two variables: “visitors” and “views_content”.

- The viewership is approximately similar across all the days.
- Friday and Thursday have high number of outliers, while Tuesday has the lowest number of outliers.
- The data range of Winter and Fall are widespread.
- Summer season has the highest number of outliers.
- The spread for major_sports_event = 0 (no sports event) is slightly wider, indicating more variability in content views when there isn't any major sports event.

2.3 BUSINESS QUESTIONS:

2.3.1 What does the distribution of content views look like?

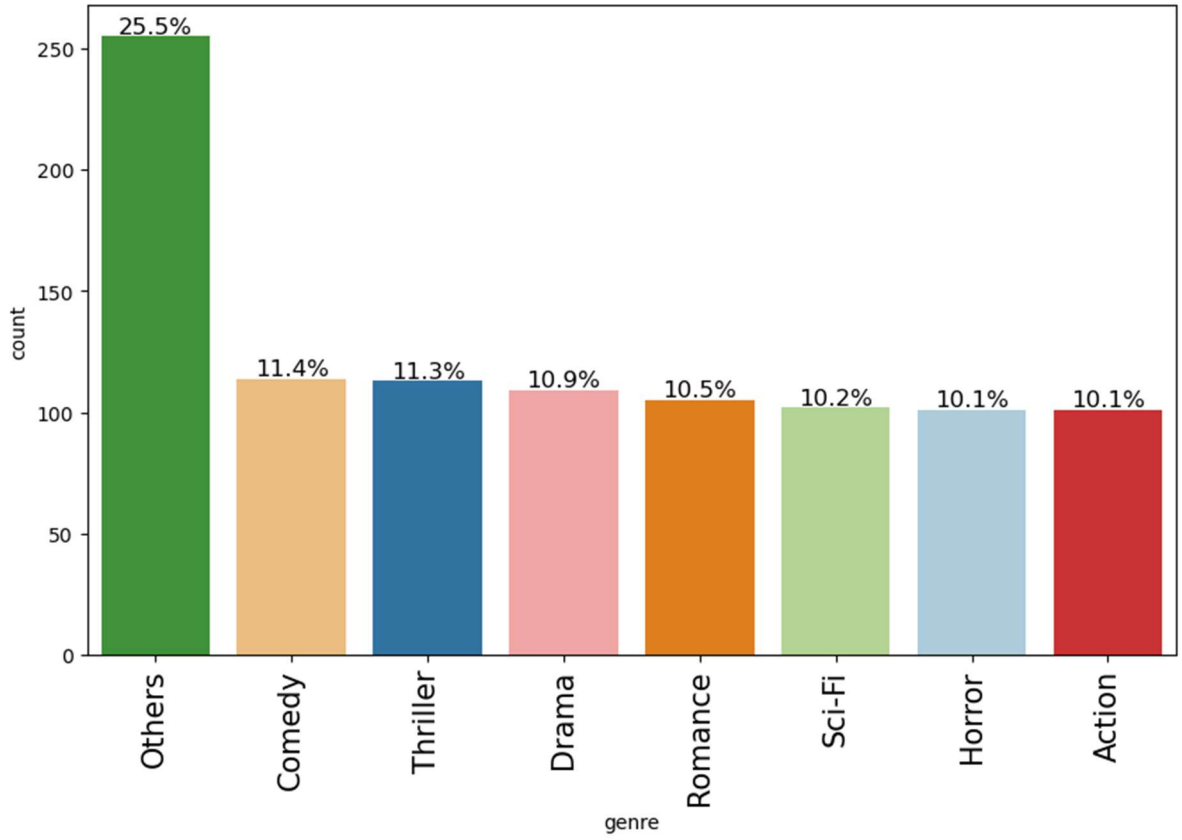
Fig.22 Views_count distribution



- The range of the views_content variable is approximately from 0.2 to 0.9, indicating that the values are spread out but concentrated mainly between 0.3 and 0.6.
- The distribution appears to be slightly right-skewed.

2.3.2 What does the distribution of genres look like?

Fig.23 Genre distribution

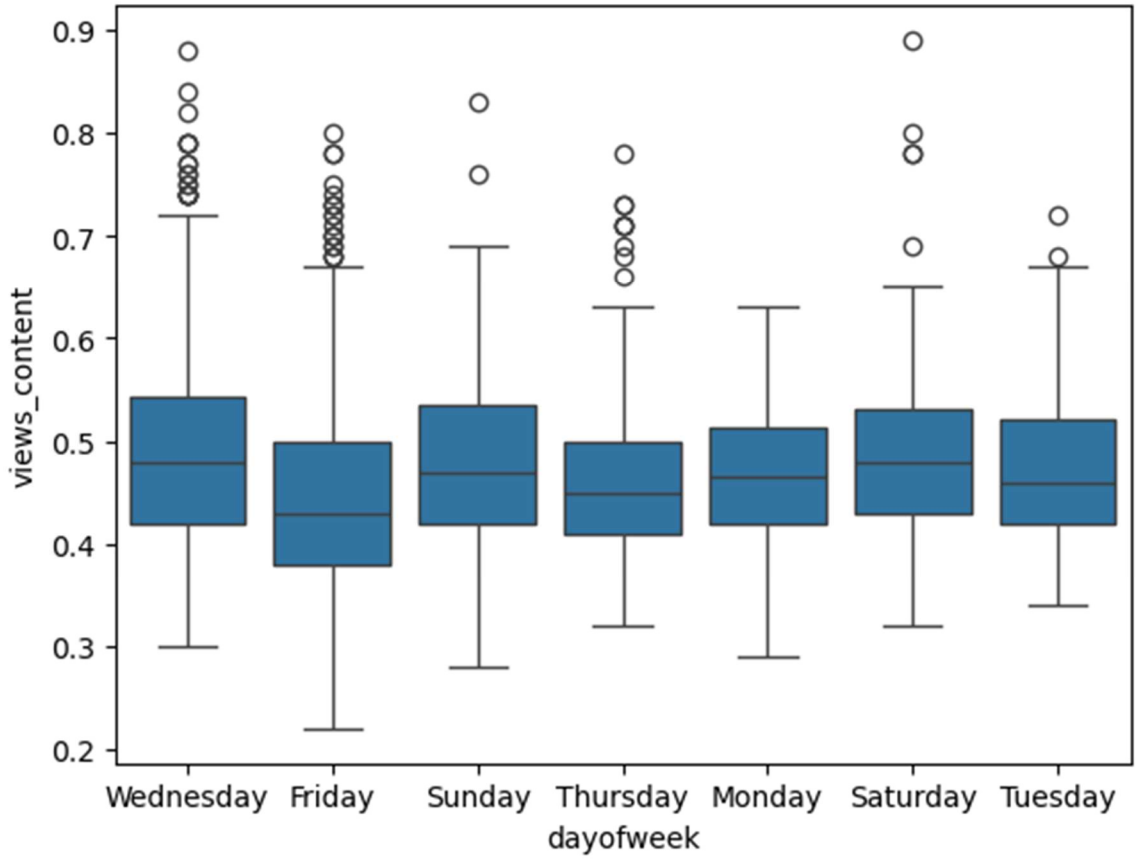


- The largest category is "Others" with 25.5% of the total, indicating that a significant portion of the dataset includes genres outside the main specified categories.

- The rest of the genres are approximately equally distributed, indicating a wide range of genres with broad audience.

2.3.3 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.24 Boxplot of dayofweek

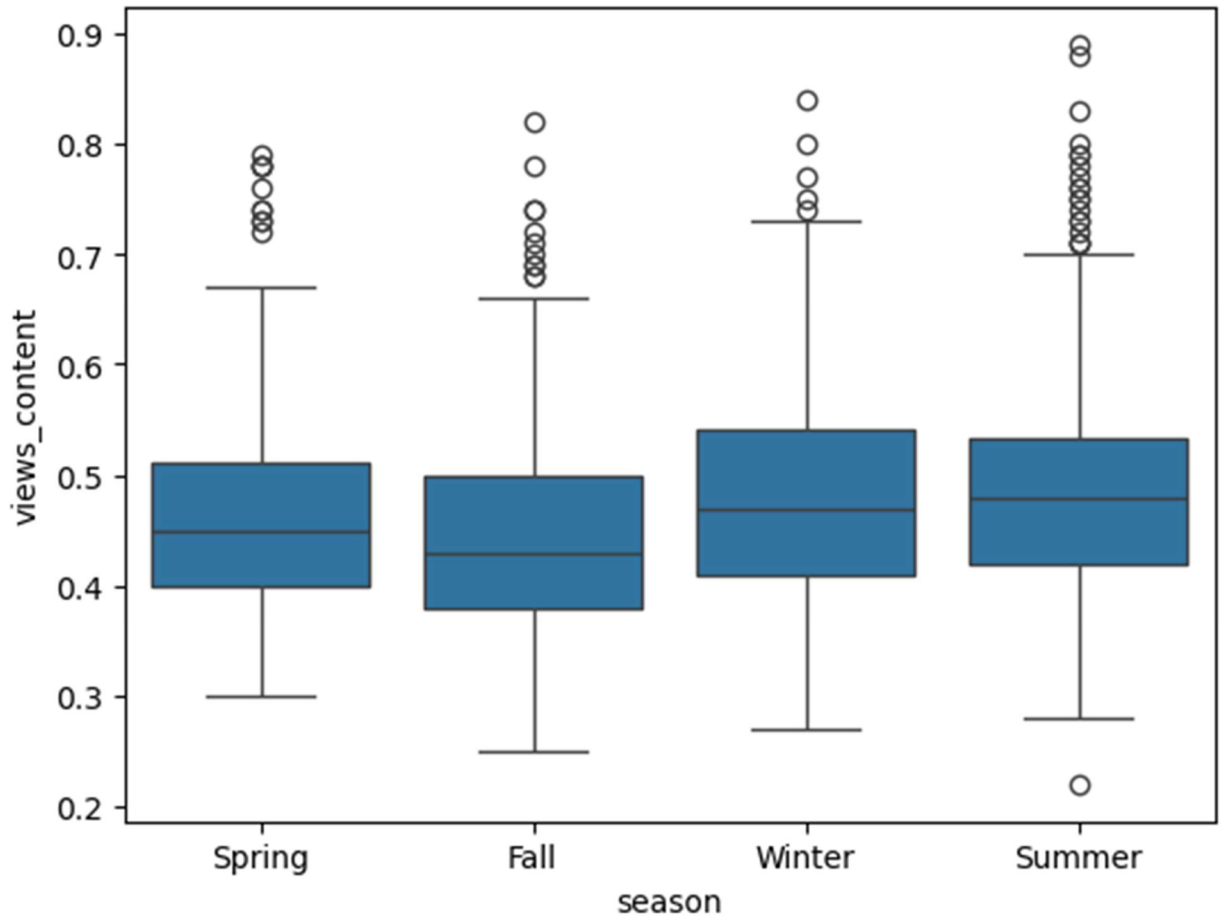


- The median values for all the days are fairly consistent and centered around the 0.45 to 0.5 million range. The Interquartile range(IQR) is also similar for all the days.
- This indicates that the viewership is approximately similar across all the days.

- Friday and Thursday have high number of outliers, while Tuesday has the lowest number of outliers.

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

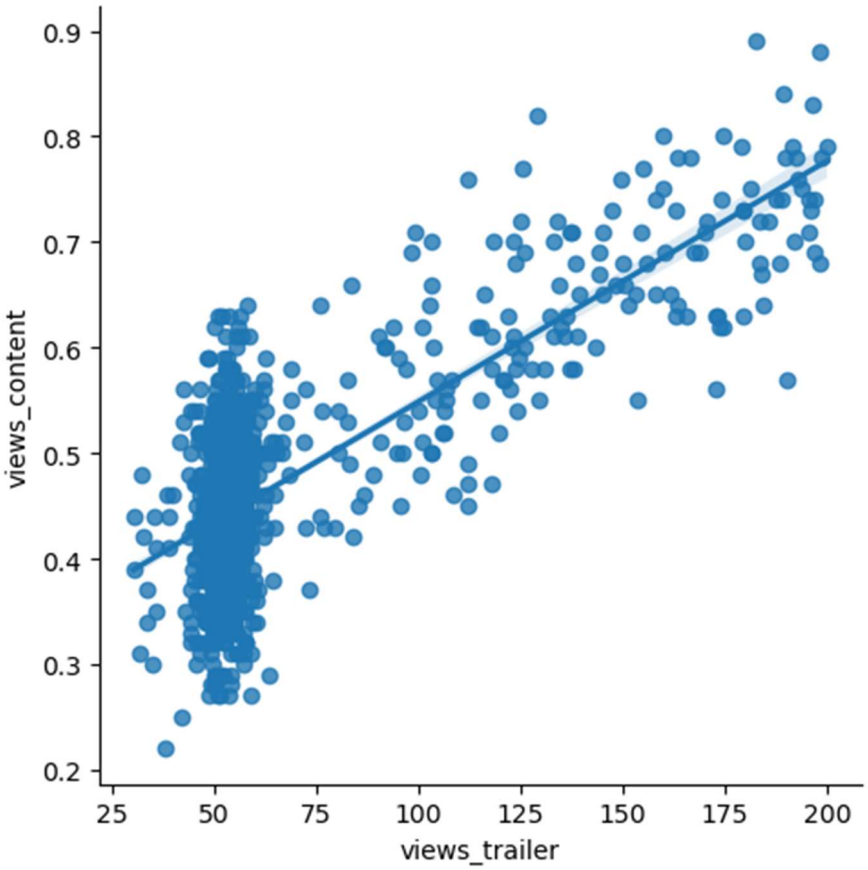
Fig.25 Boxplot of seasons



- The median values of all the seasons are approximately between 0.4 to 0.5 million.
- The data range of Winter and Fall are widespread.
- Summer season has the highest number of outliers.

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

Fig.26 Views_trailer vs Views_content 2



- The scatter plot shows a positive correlation between two variables: “views_trailer” and “views_content”.
- As the value of “views_trailer” increases, the corresponding “views_content” tends to increase as well.
- This suggests that viewership of trailers is associated with higher viewership of content.

3. DATA PREPROCESSING

- DUPLICATE VALUE CHECK:

There are no duplicates in the given dataset.

- MISSING VALUE TREATMENT:

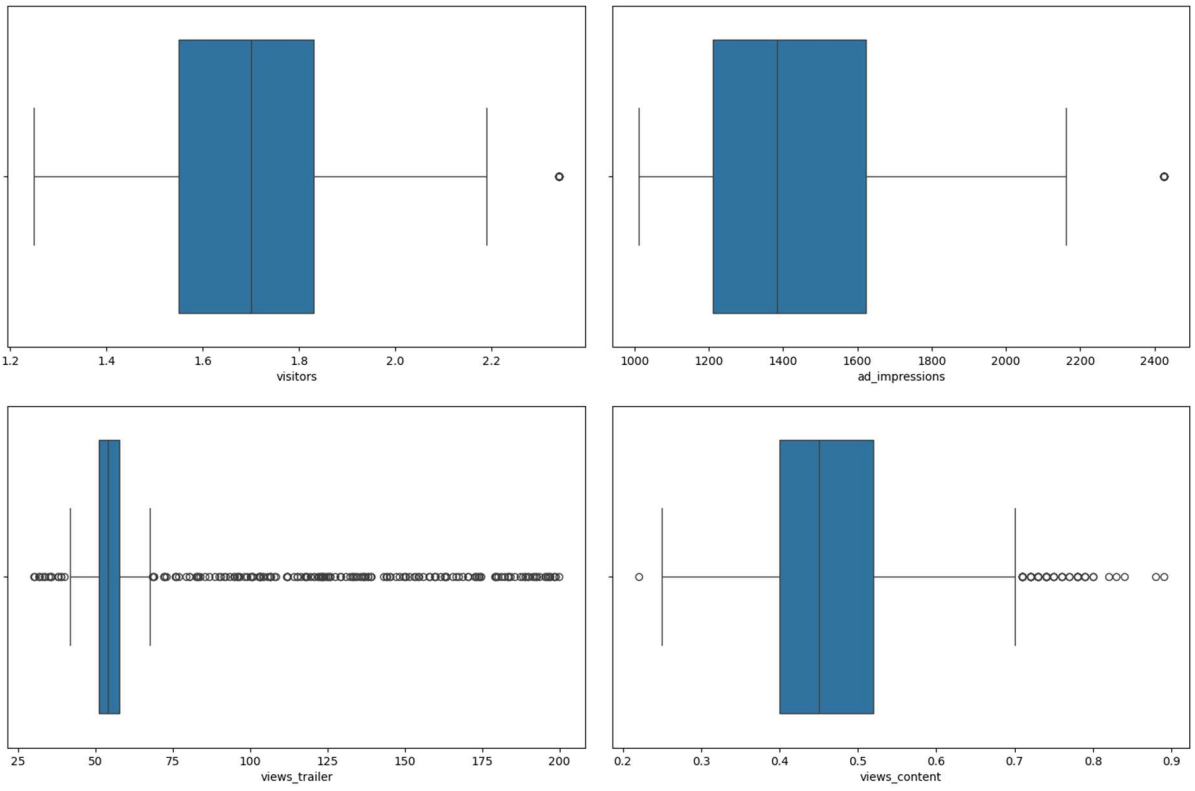
Fig.27 Null values summary

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

There are no null-values or missing values in the given dataset.

- OUTLIER TREATMENT:

Fig.28 Outliers check



- As we can clearly see, there are many outliers in the "views_trailer" and "views_content" columns.
- These outliers are meaningful and we will not treat them because each content has a distinct and huge audience.

➢ DATA PREPARATION FOR MODELLING:

- Splitting the Dataset into X and y

Fig.29 X dataset

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

Fig.30 y dataset

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

- Categorical variables are converted into numerical data using One-hot Encoding by creating dummy variables.
- These categorical columns are converted from Boolean datatype to Float type for modelling.

Fig.31 Columns after One-hot encoding

```
Columns after one-hot encoding:
const
visitors
ad_impressions
major_sports_event
views_trailer
genre_Comedy
genre_Drama
genre_Horror
genre_Others
genre_Romance
genre_Sci-Fi
genre_Thriller
dayofweek_Monday
dayofweek_Saturday
dayofweek_Sunday
dayofweek_Thursday
dayofweek_Tuesday
dayofweek_Wednesday
season_Spring
season_Summer
season_Winter
```

- The dataset is split into the ratio of 70:30.

Fig.32 No. of elements

```
Number of rows in train dataset = 700
Number of rows in test dataset = 300
```

4. MODEL BUILDING – LINEAR REGRESSION

This is the initial model summary before assumptions and treating multicollinearity.

Fig.33 Initial Model summary

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, 03 Oct 2024	Prob (F-statistic):	1.32e-215
Time:	20:22:12	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.

Column names and their coefficients(initial model):

Fig.34 Coeffecients 1

	coef
const	0.0602
visitors	0.1295
ad_impressions	3.623e-06
major_sports_event	-0.0603
views_trailer	0.0023
genre_Comedy	0.0094
genre_Drama	0.0126
genre_Horror	0.0099
genre_Others	0.0063
genre_Romance	0.0006
genre_Sci-Fi	0.0131
genre_Thriller	0.0087
dayofweek_Monday	0.0337
dayofweek_Saturday	0.0579
dayofweek_Sunday	0.0363
dayofweek_Thursday	0.0173
dayofweek_Tuesday	0.0228
dayofweek_Wednesday	0.0474
season_Spring	0.0226
season_Summer	0.0442
season_Winter	0.0272

➢ TRAINING PERFOMANCE:

Fig.35 Training Perf

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

➢ TEST PERFOMANCE:

Fig.36 Test Perf

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

OBSERVATIONS:

- The training *R*-squared is 0.79, so the model is not underfitting
- The train and test RMSE and MAE are comparable, so the model is not overfitting either
- MAE suggests that the model can predict content views within a mean error of ~0.04 on the test data
- MAPE of 9.03 on the test data means that we are able to predict within 9.03% of the content views.

5. TESTING THE ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

Assumptions of the Linear Regression Model:

- Treating Multicollinearity
- Test for Linearity
- Test for Independence
- Test for Normality
- Test for Homoscedasticity

5.1 TREATING MULTICOLLINEARITY:

Fig.37 VIF values

	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

- If VIF is between 1 and 5, then there is low multicollinearity.
- As we can clearly observe from the above table, the VIF values of all the columns are below 5.
- So, there is very low multicollinearity and negligible.

Dealing with high p-value variables:

- After treating multicollinearity in the data, we can deal the variables having high p-values.
- Some variables have p-value greater than 0.05 and can be dropped one by one, as they are not significant in predicting “view_content”.

Model Summary after treating multicollinearity and dealing p-values:

Fig.38 Summary 2

OLS Regression Results

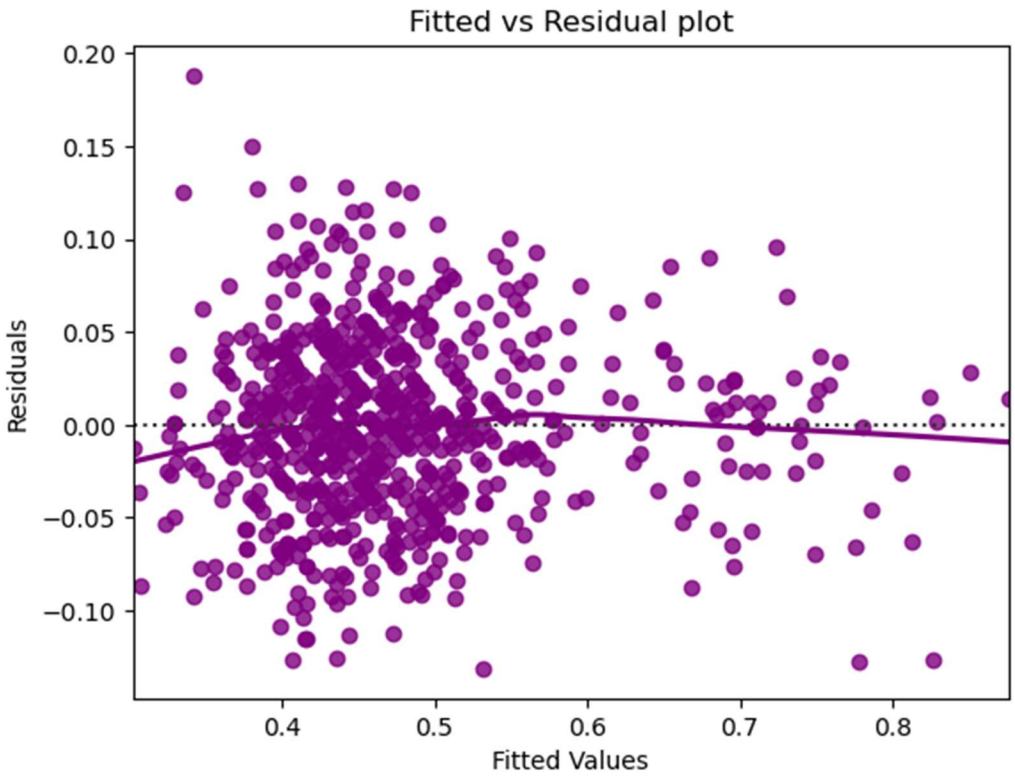
Dep. Variable:	views_content	R-squared:	0.789
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	233.8
Date:	Sun, 06 Oct 2024	Prob (F-statistic):	7.03e-224
Time:	04:30:36	Log-Likelihood:	1120.2
No. Observations:	700	AIC:	-2216.
Df Residuals:	688	BIC:	-2162.
Df Model:	11		
Covariance 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.

5.2 TEST FOR LINEARITY AND INDEPENDENCE:

Fig.39 Fitted vs Residual plot



- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.
- We see no pattern in the plot above.
- **Hence, the assumptions of linearity and independence are satisfied**.

5.3 TEST FOR NORMALTIY:

Fig.40 Distribution of Residuals

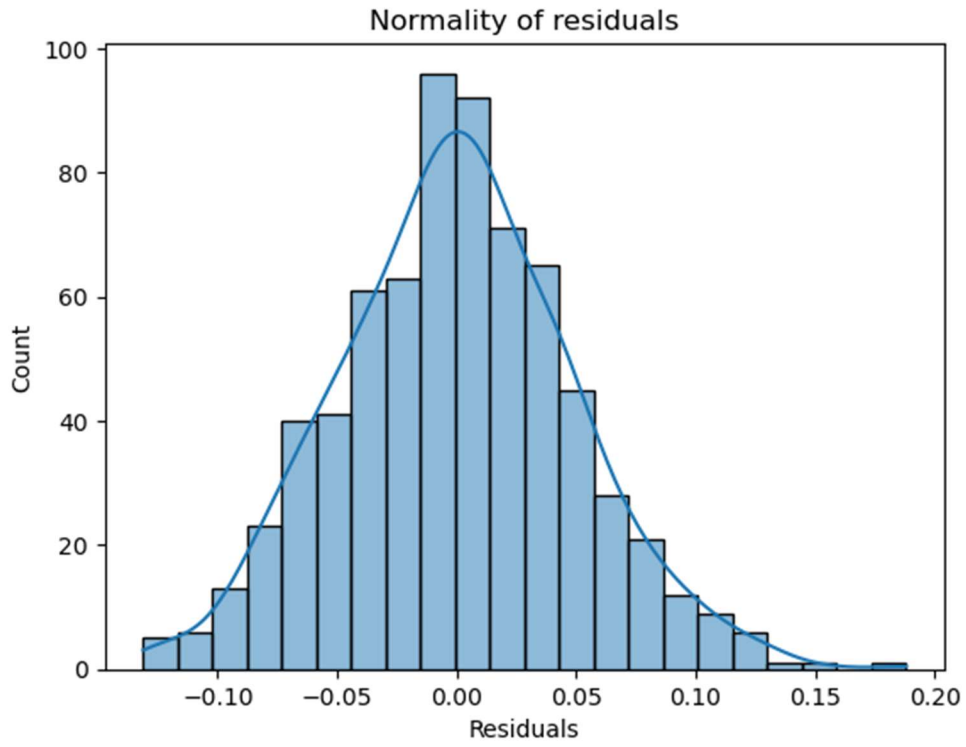
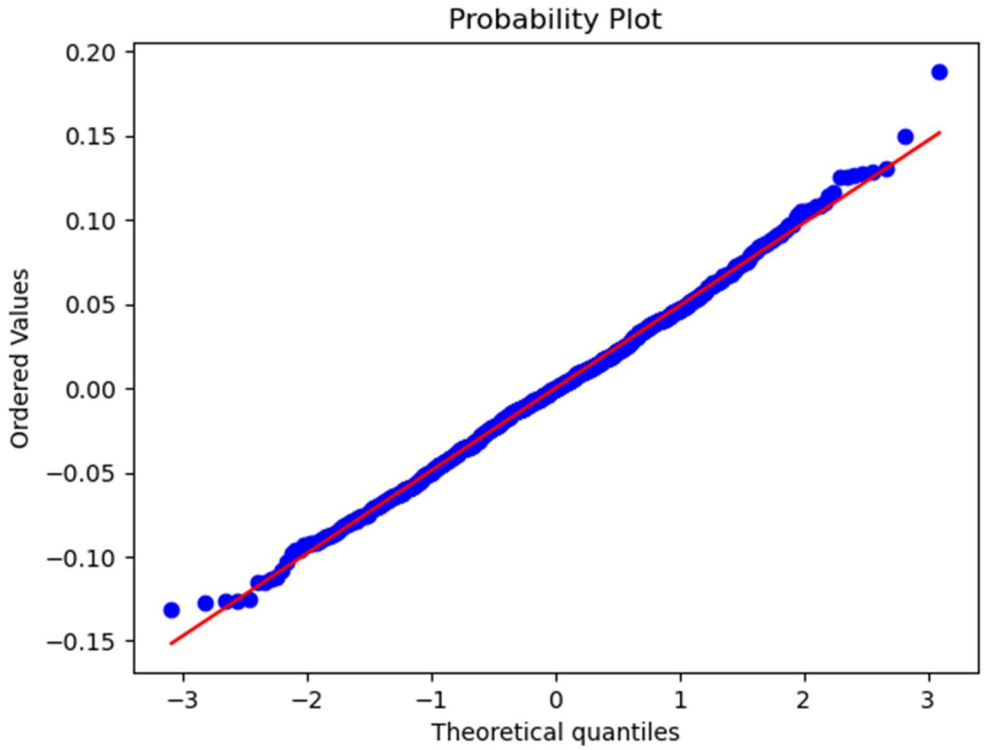


Fig.41 Q-Q plot of Residuals



OBSERVATIONS:

- The Histogram plot of Residuals have a bell shape and resembles a normal distribution.
- In the Q-Q plot, the Residuals follow a straight line except for the very end of the tails.

SHAPIRO-WILK TEST:

Null Hypothesis: Residuals follows normal distribution.

Alternate Hypothesis: Residuals doesn't follow normal distribution.

Fig.42 Shapiro-wilk test result

```
ShapiroResult(statistic=0.9973143339157104, pvalue=0.3104695975780487)
```

- Since the p-value(0.3104) is greater than 0.05, we fail to reject the null hypothesis.
- **The Residuals are normally distributed as per Shapiro-Wilk test.**

5.4 TEST FOR HOMOSCEDASTICITY:

GOLDFELDQUANDT TEST:

Null Hypothesis: Residuals are homoscedastic.

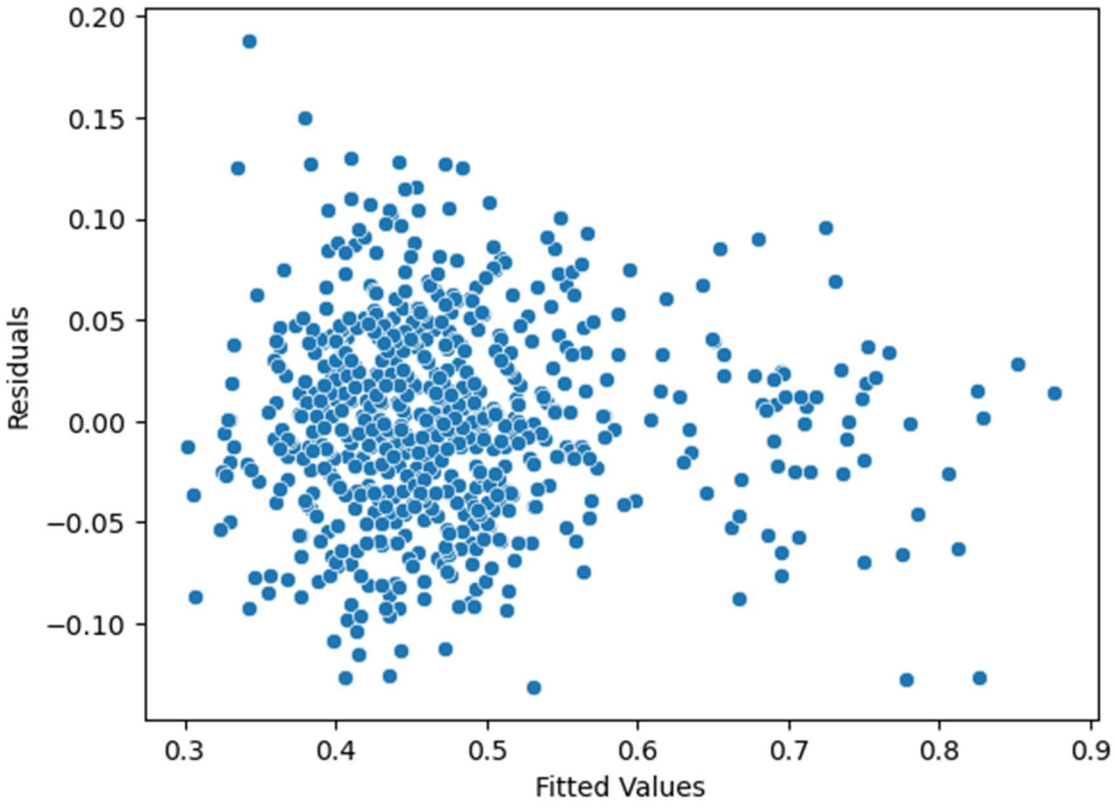
Alternate Hypothesis: Residuals have heteroscedasticity.

Fig.43 Goldfeldquandt test result

```
[('F statistic', 1.1313612904200754), ('p-value', 0.12853551819086995)]
```

- Since the p-value(0.128) is greater than 0.05, we fail to reject the null hypothesis.
- **The Residuals are homoscedastic as per the Goldfeldquandt test**.

Fig.44 Fitted vs Residuals



- **As we can clearly see, the points in the residual plot are randomly scattered.**

6. MODEL PERFOMANCE EVALUATION

Final Model Summary:

Fig.45 Final Model Summary

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:	Sun, 06 Oct 2024	Prob (F-statistic):	7.03e-224
Time:	17:03:46	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.

- TRAINING PERFOMANCE:

Fig.46 Final Training Perf

Training Performance

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

➢ TESTING PERFOMANCE:

Fig.47 Final Testing Perf

Test Performance

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

OBSERVATIONS:

- The training *R*-squared is 0.78, so the model is not underfitting
- The model is able to explain ~78% of the variation in the data.
- The train and test RMSE and MAE are comparable, so the model is not overfitting either.
- MAE suggests that the model can predict views count within a mean error of ~0.04 on the test data
- MAPE of 9.17 on the test data means that we are able to predict within 9.17% of the views count.

Fig.48 Final Model Coefficients

	coef
const	0.0747
visitors	0.1291
major_sports_event	-0.0606
views_trailer	0.0023
dayofweek_Monday	0.0321
dayofweek_Saturday	0.0570
dayofweek_Sunday	0.0344
dayofweek_Thursday	0.0154
dayofweek_Wednesday	0.0465
season_Spring	0.0226
season_Summer	0.0434
season_Winter	0.0282

ACTIONABLE INSIGHTS & RECOMMENDATIONS

- The model is able to explain ~78% of the variation in the data and within 9.17% of the views count on the test data, which is good. This indicates that the model is good for prediction as well as inference.
- The baseline value for "views_content" is 0.074 when all other variables are set to zero.
- If the “visitors” count increases by one unit, then the “view_count” increases by 0.1291 units. This is the largest positive coefficient in the equation, suggesting that the number of visitors has a strong and direct effect on the views.
- Specific days of the week (especially Saturdays and Wednesdays) can moderately increase the views_count.
- Summer has the most significant positive effect on views among seasons, followed by Winter and Spring.
- When a major sports event occurs, the views_content decreases by approximately 0.0606, holding all other factors constant. This suggests that major sports events can detract from content views, possibly because people are more focused on the sports event.
- As the views count increase with an increase in the trailer views, the company can improve its marketing activities to promote their trailers.
- To increase the first day viewership, the content can be released in Summer season and specific days like Saturdays and Wednesdays.
- As we can clearly see, the ‘visitors’ is the largest positive coefficient, the OTT service provider can implement various marketing strategies to increase the visitors.