Business Report - 3

PG Program in Data Science and Business Analytics

submitted by

Sangram Keshari Patro

BATCH:PGPDSBA.O.AUG24.B

Contents

List of Figures

List of Tables

1 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.

2 Data Description

The data contains the different factors to analyze for the content. The detailed data dictionary is ws, in millions, of the content

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 (both running and completed).
3. major_sports_event: Indicates if there was any major sports event on the day.
4. genre: Genre of the content.
5. dayofweek: Day of the week on which the content was released.
6. season: Season during which the content was released.
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.

3 Data Overview

3.1 Importing necessary libraries and the dataset

The dataframe is printed. It has 1000 rows & 8 columns.

	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: Dataframe

3.2 Structure and type of data

Data is explored further. Data doesn't have any duplicate rows.

```
<class 'pandas.core.frame.DataFrame'>
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)
```

Figure 2: Table depicting the datatype and Non-Null values in each column.

3.3 Statistical summary

	count	mean	std	min	25%	50%	75%	max
visitors	1000.0	1.69331	0.182420	1.41	1.5500	1.70	1.830	1.97
ad_impressions	1000.0	1411.04096	236.672877	1010.87	1210.3300	1383.58	1623.670	1830.34
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 3: Statistical summary of the data

From this table we can observe that there are outliers in the columns.

4 Exploratory Data Analysis

4.1 Univariate Analysis

4.1.1 Numerical columns

- **views_content**

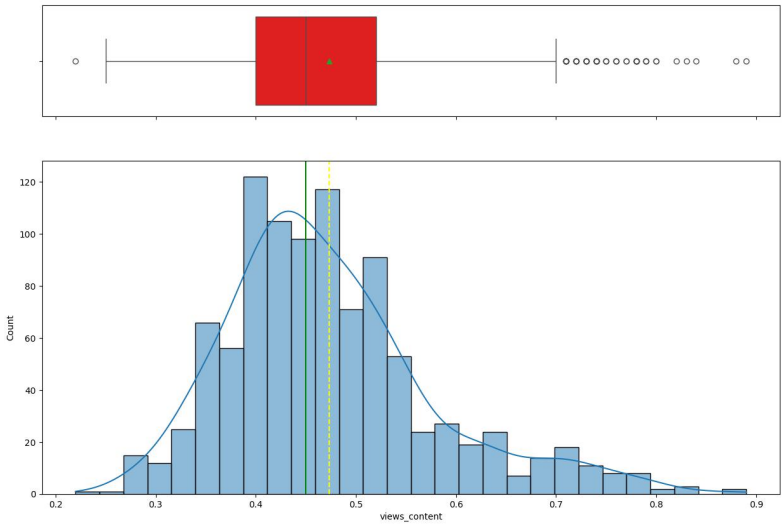


Figure 4: Histogram and boxplot of 'views_content' column

Observations:

The distribution of first-day views (views_content) is approximately normal with a slight positive skew. The mean first-day views are around 0.47 million. The median is close to the mean, indicating a roughly symmetric distribution. The range of first-day views spans from approximately 0.2 million to 0.9 million, with some content achieving exceptionally high viewership.

Insights:

Most content has moderate viewership on the first day, with a few notable outliers. ShowTime could analyze high-performing content to identify factors driving higher viewership, such as genre or marketing, to apply these insights to other content.

- **views_trailer**

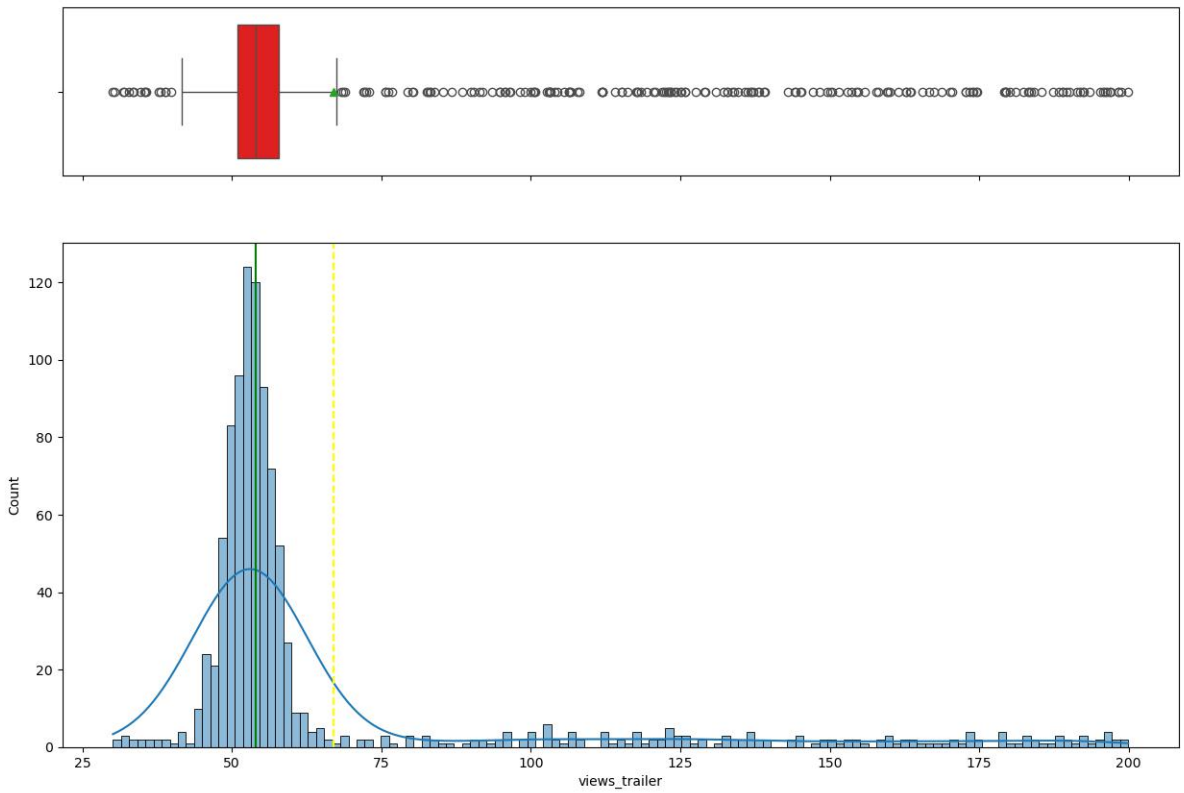


Figure 5: Histogram and boxplot of 'views_trailer' column

Observations:

The distribution of trailer views is positive skewed. Mean trailer views are around 67 million, with the median around 54 million which is lower than the mean. Trailer views range from 30 million to about 200 million, with most values between 40 and 70 million.

Insights:

The consistency in trailer views suggests that trailers generate stable interest among viewers. Increasing the visibility or promotion of trailers may help drive higher anticipation and, consequently, first-day views.

- **visitors**

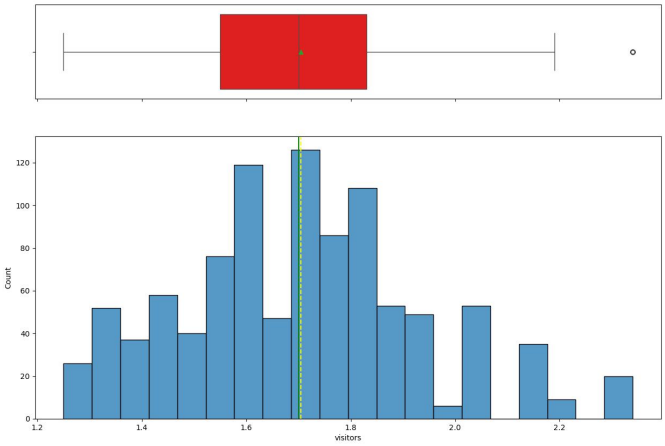


Figure 6: Histogram and boxplot of "visitors" column

Observations:

The visitors distribution has a mean of approximately 1.7 million to the platform in the past week. The median is close to the mean. Visitor counts range from 1.4 million to 2 million, suggesting consistent platform traffic in the past week. There are no outliers in the dataset.

Insights:

ShowTime has a steady visitor base, but not all visitors engage with new content. Encouraging these visitors to view new releases, possibly through notifications or recommendations, could significantly increase first-day views.

- **ad_impression**

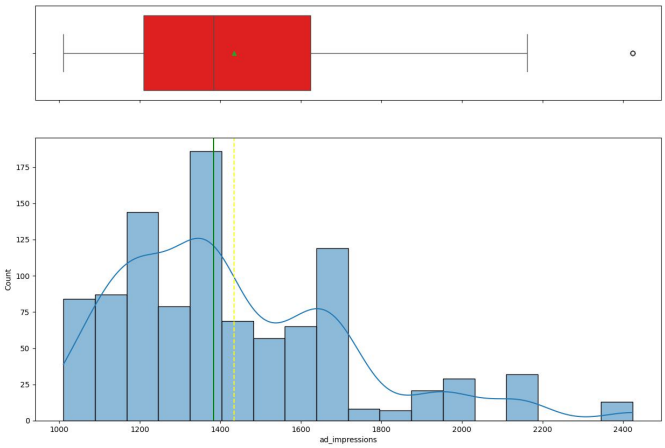
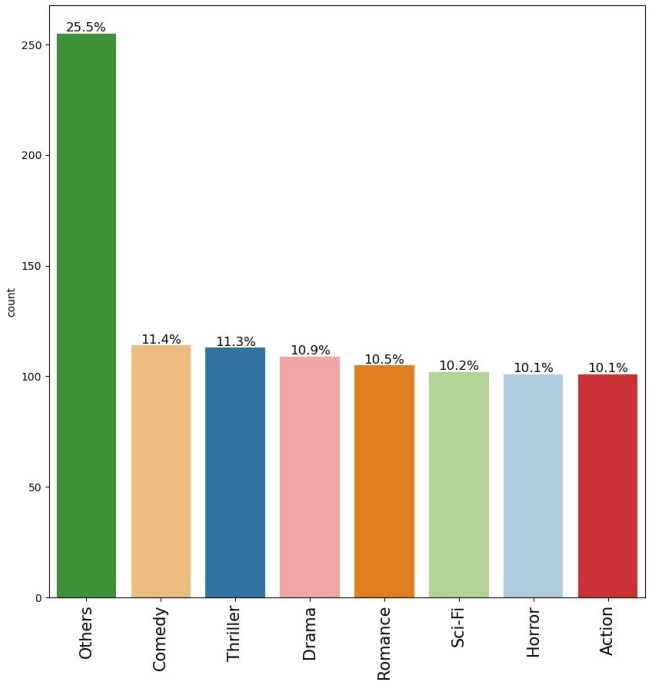


Figure 7: Histogram and boxplot of '**ad_impression**' column

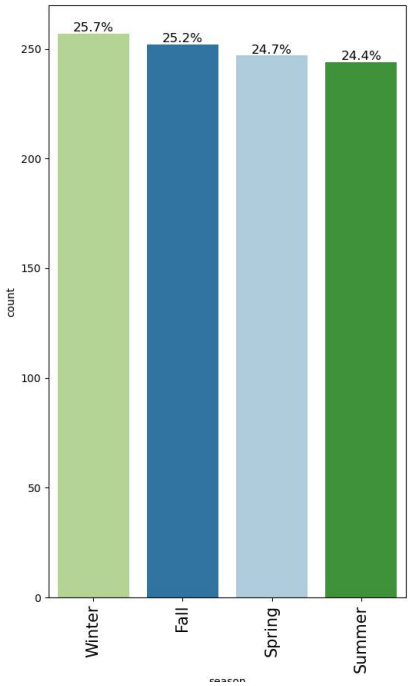
Observations:

The ad impression distribution has a mean of approximately 1411 million to the platform in the past week. The median is close to the mean. There are no outliers in the dataset.

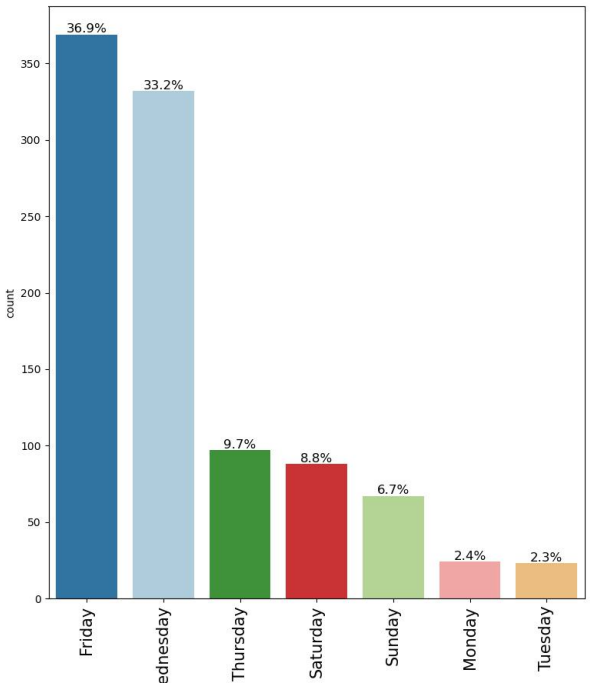
4.1.2 Categorical columns



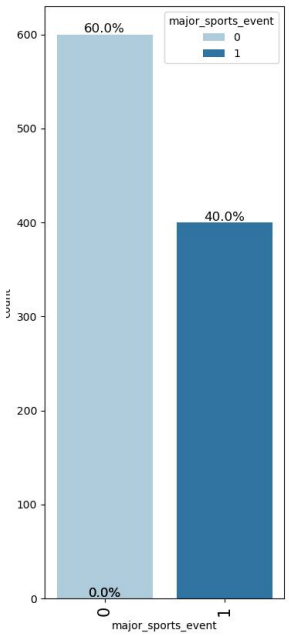
(a) '**genre**'



(b) '**season**'



(c) '**dayofweek**'



(d) '**major_sports_event**'

Figure 8: Barchart of '**genre**' , '**season**', '**dayofweek**' and '**major_sports_event**' column

Observations

1. The **day of the week bar chart** indicates that **Friday** has the highest count, followed by **Wednesday**. **Tuesday** and **Monday** have the lowest counts i.e. Most of the contents are released on **Friday** and**Tuesday**.
2. The **season bar chart** shows relatively even distribution across all seasons, with **Winter** having the highest count, closely followed by **Fall**, **Spring**, and **Summer**.
3. The contents by ShowTime feature various genres, all of which have equal weightage, with Comedy content being released the most frequently.

Insights

1. The company may consider focusing marketing efforts more heavily on weekends too apart Fridays to maximize audience engagement.
2. The even distribution of counts across seasons indicates that activities or events are spread throughout the year. This suggests that seasonal trends have minimal impact on the data, so the company may maintain a consistent approach throughout the year.
3. The content released on days when major sports events occur should be minimized, as it currently accounts for 40

4.2 Bivariate Analysis

4.2.1 Numerical variables

- **Heatmap**

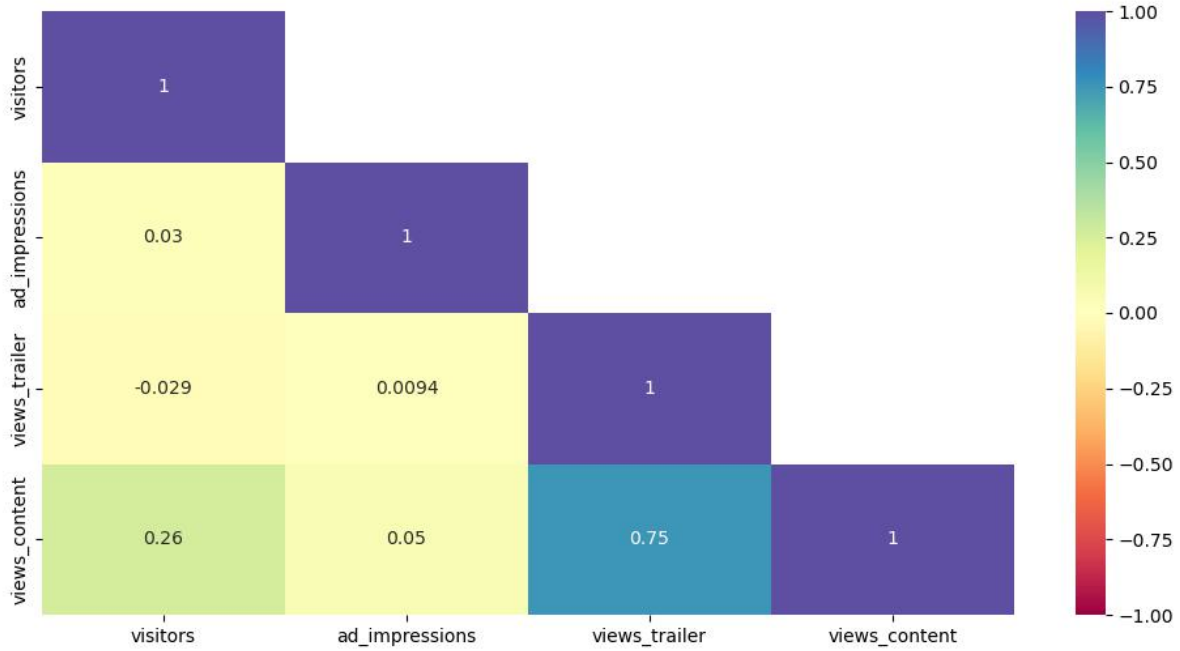


Figure 9: Heatmap of all numerical variables

Key Insights: -

(a) **Views_content** has a strong positive correlation with **views_trailer** (0.75), suggesting that customers who watch trailers are highly likely to watch the content as well.

(b) The average number of **visitors** to the platform last week shows a moderate positive correlation (0.24) with **Views_content**.

(c) Interestingly, **ad_impressions** does not show a strong correlation with **Views_content**, highlighting the need to reconsider the type or platform of advertisements being used.

- **Pairplot**

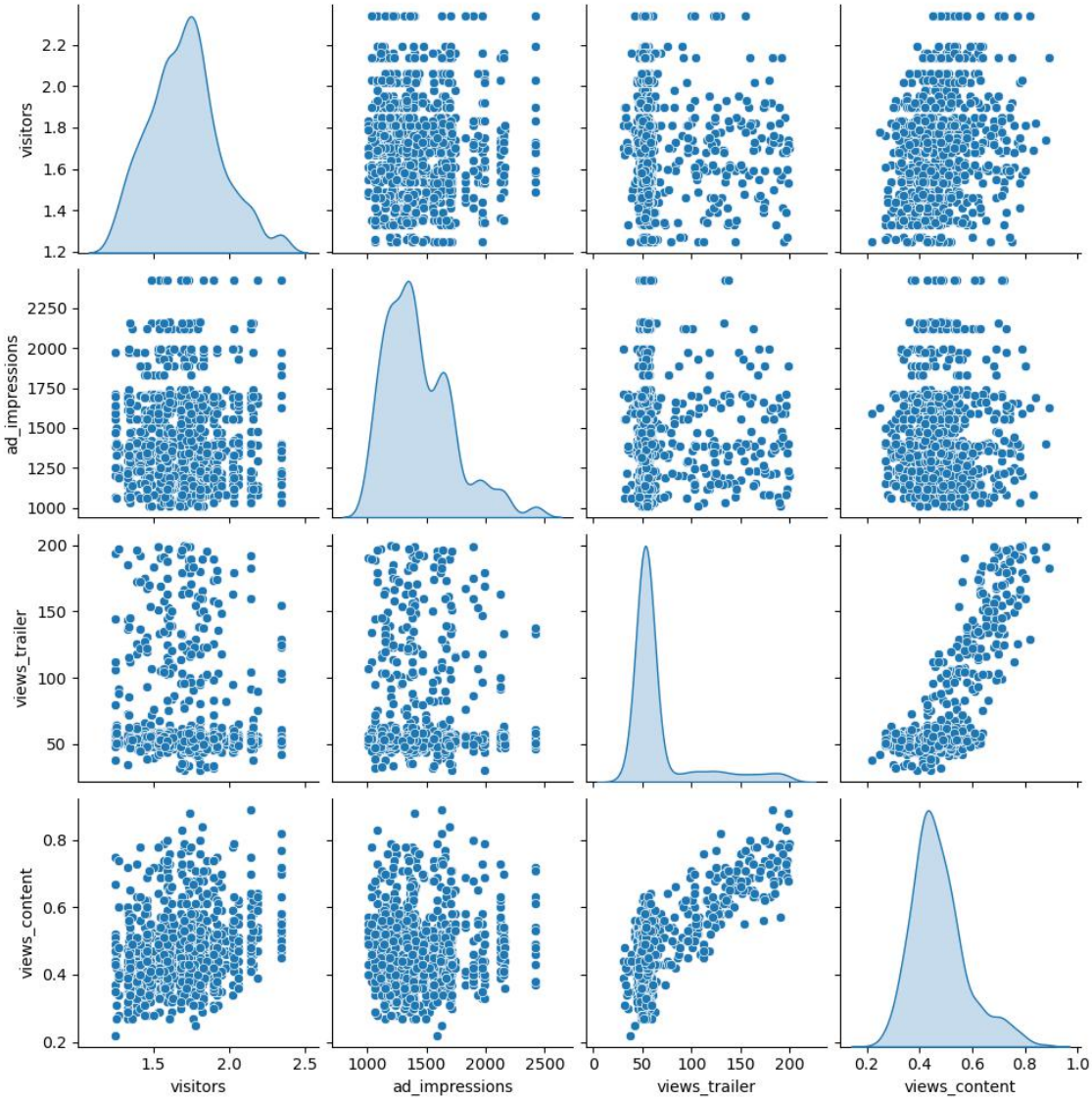
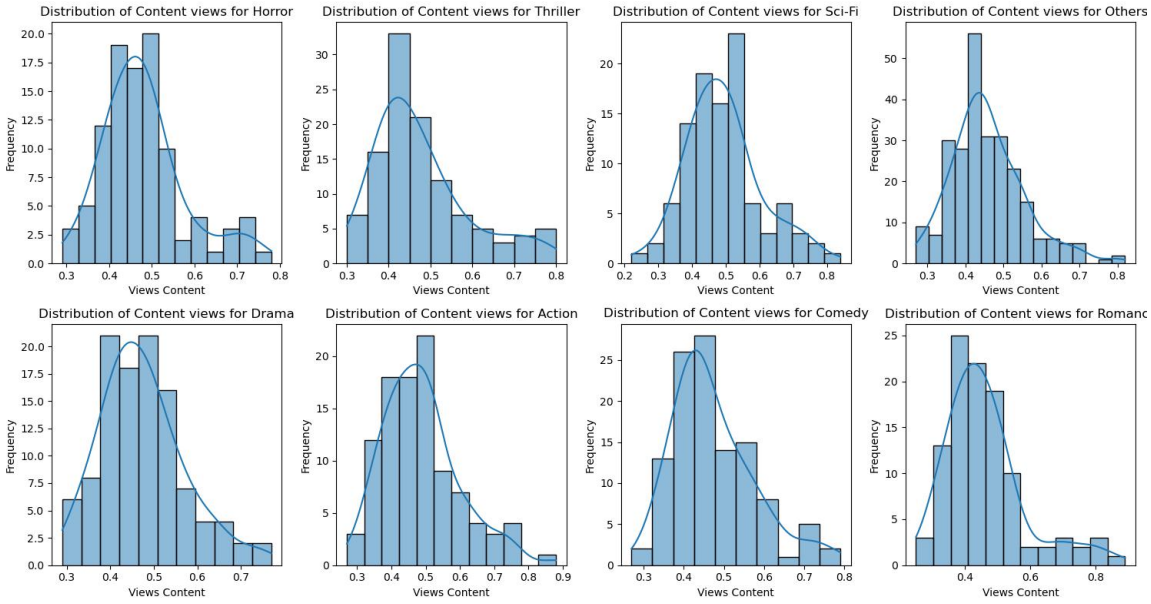


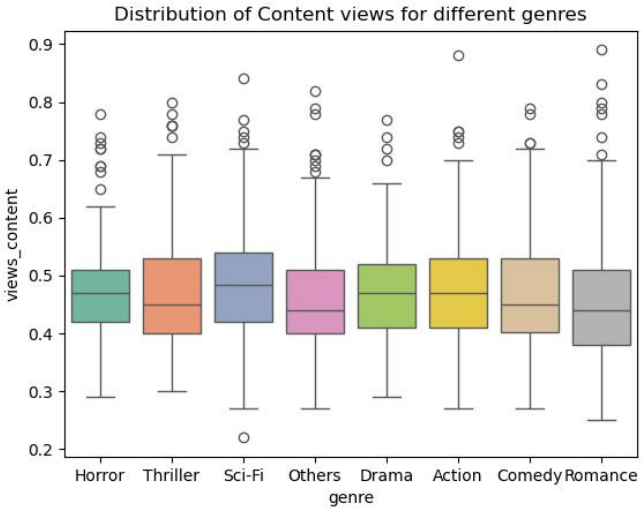
Figure 10: Pairplot of all numerical variables

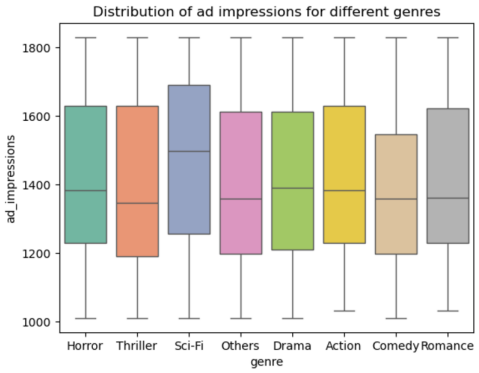
This graph explains that customers watching trailer mostly prefer to watch the content too as the correlation is high.

4.2.2 Categorical vs numerical variables

- **'genre' vs 'views_content'**







genre	ad_impressions max	ad_impressions mean	ad_impressions min	views_content max	views_content mean	views_content min	views_trailer max	views_trailer mean	views_trailer min	visitors max	visitors mean	visitors min
Action	1830.340000	1430.026040	1031.850000	0.880000	0.487327	0.270000	198.310000	71.224950	33.510000	1.970000	1.720198	1.410000
Comedy	1830.340000	1381.720000	1010.870000	0.790000	0.474211	0.270000	192.320000	63.592368	43.870000	1.970000	1.683772	1.410000
Drama	1830.340000	1416.390367	1010.870000	0.770000	0.473945	0.290000	190.020000	69.286330	42.590000	1.970000	1.689725	1.410000
Horror	1830.340000	1406.766634	1010.870000	0.780000	0.481089	0.290000	196.110000	66.217426	30.340000	1.970000	1.702970	1.410000
Others	1830.340000	1395.845137	1010.870000	0.820000	0.459765	0.270000	196.950000	62.445137	31.720000	1.970000	1.682510	1.410000
Romance	1830.340000	1404.416762	1031.850000	0.890000	0.461238	0.250000	199.920000	67.690762	35.550000	1.970000	1.708571	1.410000
Sci-Fi	1830.340000	1474.692157	1010.870000	0.840000	0.493235	0.220000	189.000000	71.201765	30.080000	1.970000	1.692157	1.410000
Thriller	1830.340000	1405.304425	1010.870000	0.800000	0.476903	0.300000	198.720000	70.252655	45.270000	1.970000	1.684956	1.410000

Figure 11: Boxplot and histplot for various aspects across different **genre**

Observations:

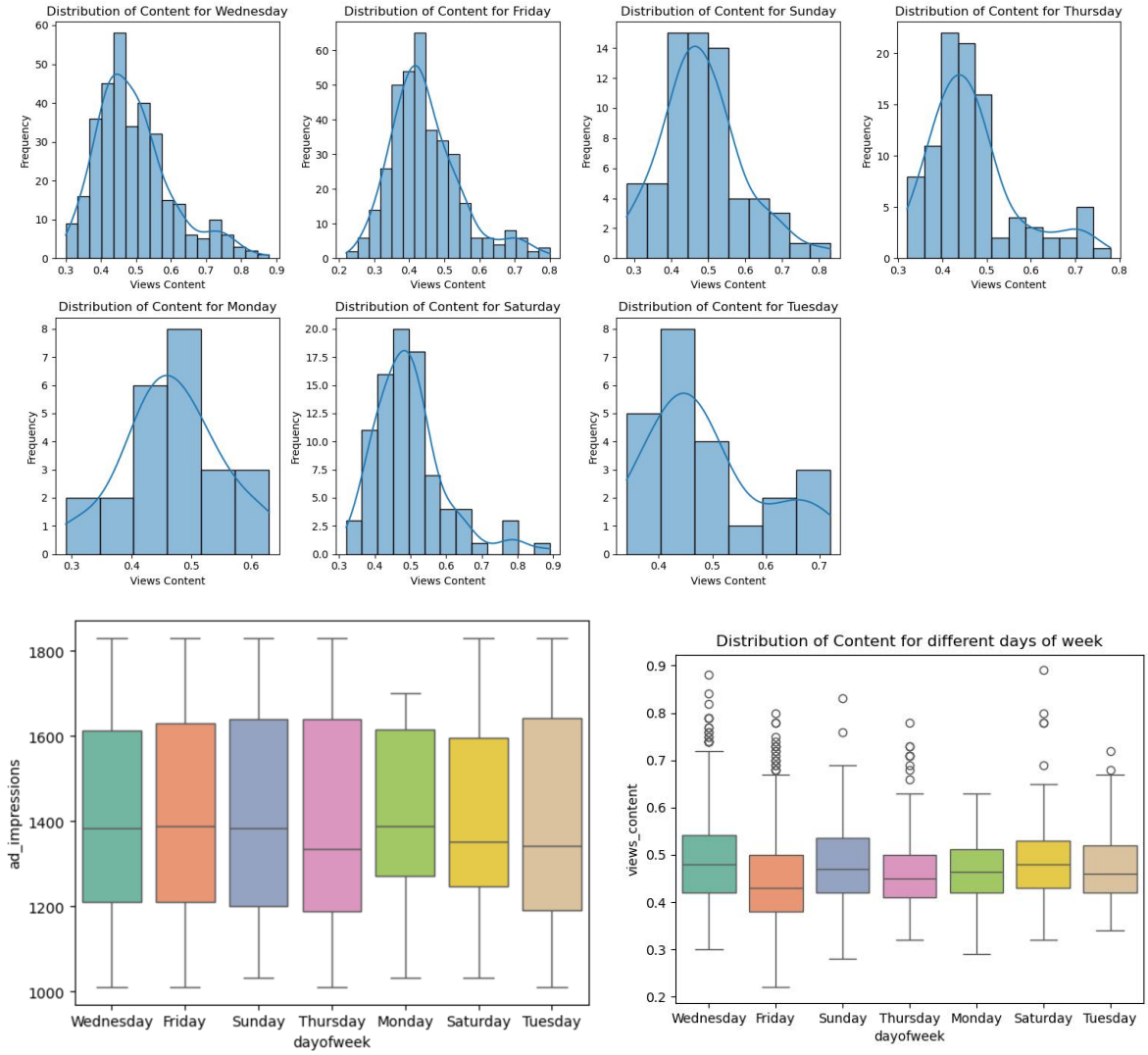
The distributions across different genres are quite similar, exhibiting a right-skewed pattern. These distributions generally peak around 0.4 to 0.5 million view counts, with a range between 0.3 and 0.8 million view counts. Among all genres, Sci-Fi has the highest mean and median content views, while the Thriller genre has the lowest content views.

Insights:

ShowTime should focus on adding more Sci-Fi content to their platform, as it attracts higher view-

ership, and consider producing similar content. Additionally, since ad impressions are lowest for the Thriller genre, resulting in minimal content views, ShowTime should work on strategies to boost the viewership of Thriller movies on their platform. Ad impressions for Sci-Fi are significantly higher compared to other genres, resulting in increased view counts. The pattern of median content views aligns closely with ad impressions across genres, except for Comedy and Romance. This suggests that increasing ad impressions for Comedy genre as compared to Romance genre could help boost content views for Comdedy genre more effectively.

- **'dayofweek' vs 'views_content'**



	ad_impressions			views_content			views_trailer			visitors		
	max	mean	min	max	mean	min	max	mean	min	max	mean	min
dayofweek												
Friday	1830.340000	1419.750921	1010.870000	0.800000	0.446694	0.220000	198.720000	65.999566	30.080000	1.970000	1.692304	1.410000
Monday	1700.040000	1410.471667	1031.850000	0.630000	0.467917	0.290000	138.030000	62.680000	46.010000	1.970000	1.720000	1.410000
Saturday	1830.340000	1414.546705	1031.850000	0.890000	0.497955	0.320000	192.320000	62.818864	35.580000	1.970000	1.695568	1.410000
Sunday	1830.340000	1419.865672	1031.850000	0.830000	0.484179	0.280000	196.420000	69.039254	44.280000	1.970000	1.705970	1.410000
Thursday	1830.340000	1391.906186	1010.870000	0.780000	0.470619	0.320000	196.110000	67.447216	38.100000	1.970000	1.681134	1.410000
Tuesday	1830.340000	1390.571739	1010.870000	0.720000	0.487826	0.340000	174.580000	71.657826	44.610000	1.970000	1.726522	1.410000
Wednesday	1830.340000	1405.699970	1010.870000	0.880000	0.494608	0.300000	199.920000	68.413343	30.340000	1.970000	1.690602	1.410000

Figure 12: Boxplot and histplot for various aspects across different **days of week**

Observations:

The distributions across different days of the week vary significantly, with a right-skewed pattern. Content views are higher on Wednesdays, Saturdays, and Sundays. Ad impressions are higher on Wednesdays, Fridays, Mondays, and Sundays.

Insights:

ShowTime should consider increasing advertisements on Saturdays, as these days attract higher viewership. Additionally, ad impressions are lowest on Thursdays, leading to minimal content views. While content views on Fridays are relatively low, a significant amount of money is being spent on advertisements on that day. This should be adjusted, with more advertisements allocated to Wednesdays, Saturdays, and Sundays to optimize impact.

- **'season' vs 'views_content'**

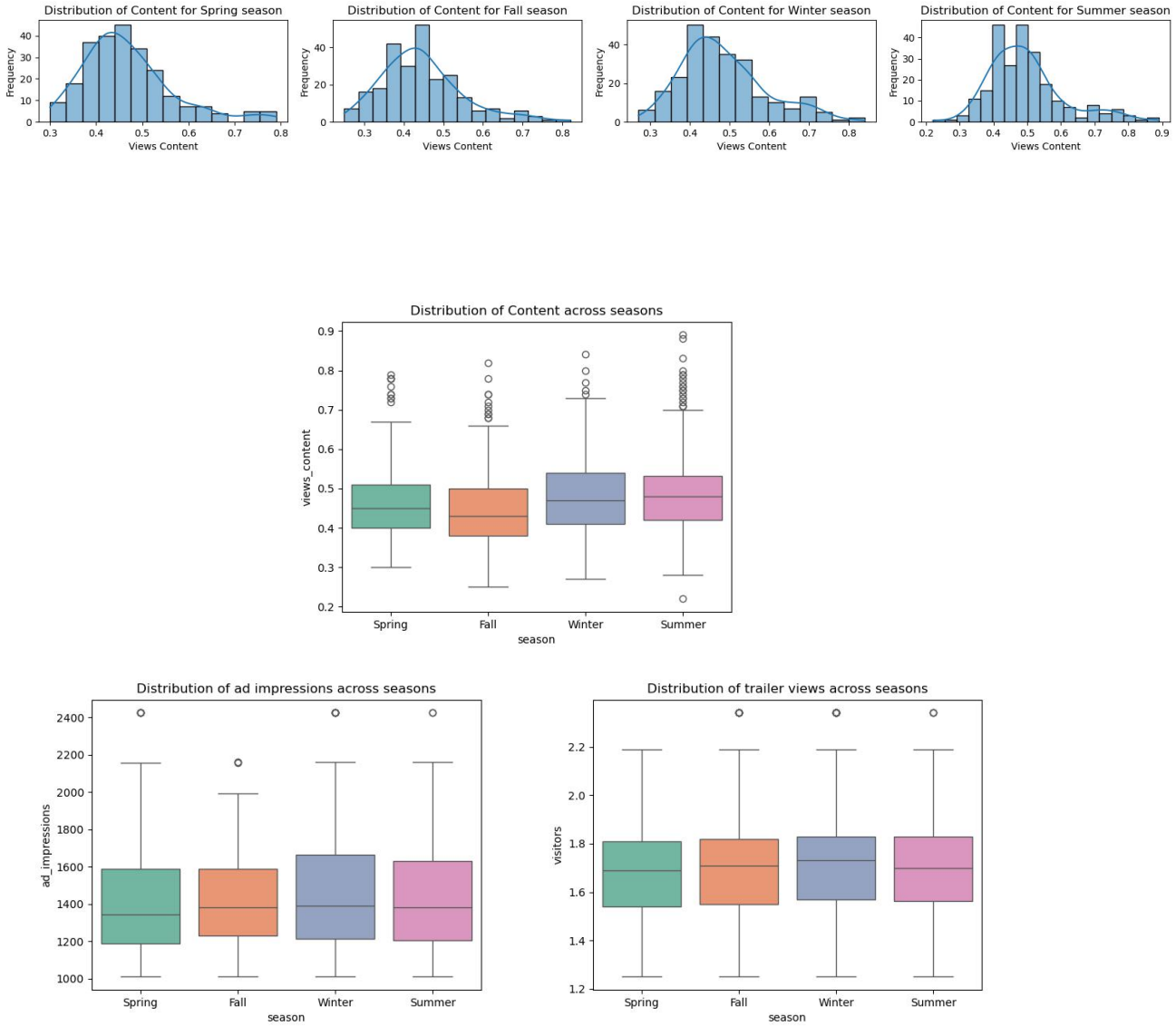


Figure 13: Boxplot and histplot for various aspects across different **season**

Observations:

- The distributions across different season are quite similar, exhibiting a right-skewed pattern.

- Content views are higher on summer and winter seasons. Ad impressions are higher on fall, summer and winter seasons.

Insights:

- ShowTime should consider increasing advertisements on Spring seaons, as it attracts higher viewership.
- While content views on Fall are relatively low, a significant amount of money is being spent on advertisements on that season. This should be adjusted, with more advertisements allocated to summer and winter seasons to optimize impact.
- The number of visitors during summer in the past week was lower than in winter, yet content views remained high. This suggests that allocating more budget to advertisements in summer compared to winter could be beneficial.

5 Data preprocessing and Model building

The dataset contains no missing or duplicate values. Outliers in the "visitors" and "ad impressions" columns have been addressed, but outliers in "trailer views" and "content views" have not been treated, as doing so would negatively impact the model's performance. Removing outliers from these two columns resulted in an R^2 value of approximately 0.5, whereas retaining the outliers produced an R^2 value of around 0.78.

We have splited the data for training and testing purposes. The model summary is as follows.

5.1 Model 1

	feature	VIF
0	const	145.408592
1	visitors	1.027710
2	ad_impressions	1.039782
3	views_trailer	1.035394
4	major_sports_event_yes	1.083391
5	genre_Comedy	1.921503
6	genre_Drama	1.911886
7	genre_Horror	1.845480
8	genre_Others	2.549471
9	genre_Romance	1.761568
10	genre_Sci-Fi	1.817227
11	genre_Thriller	1.888895
12	dayofweek_Monday	1.059507
13	dayofweek_Saturday	1.151651
14	dayofweek_Sunday	1.147846
15	dayofweek_Thursday	1.185335
16	dayofweek_Tuesday	1.071421
17	dayofweek_Wednesday	1.333738
18	season_Spring	1.575481
19	season_Summer	1.592312
20	season_Winter	1.602733

Dep. Variable:	views_content	R-squared:	0.775
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	108.3
Date:	Sat, 16 Nov 2024	Prob (F-statistic):	1.85e-188
Time:	16:42:36	Log-Likelihood:	1024.2
No. Observations:	650	AIC:	-2006.
Df Residuals:	629	BIC:	-1912.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0258	0.024	1.070	0.285	-0.022	0.073
visitors	0.1548	0.011	14.143	0.000	0.133	0.176
ad_impressions	-8.752e-07	8.64e-06	-0.101	0.919	-1.78e-05	1.61e-05
views_trailer	0.0023	6.04e-05	38.707	0.000	0.002	0.002
major_sports_event_yes	-0.0613	0.004	-14.393	0.000	-0.070	-0.053
genre_Comedy	0.0114	0.008	1.350	0.178	-0.005	0.028
genre_Drama	0.0116	0.009	1.333	0.183	-0.006	0.029
genre_Horror	0.0074	0.009	0.847	0.397	-0.010	0.025
genre_Others	0.0056	0.008	0.739	0.460	-0.009	0.020
genre_Romance	-0.0020	0.009	-0.220	0.826	-0.019	0.016
genre_Sci-Fi	0.0101	0.009	1.149	0.251	-0.007	0.027
genre_Thriller	0.0048	0.009	0.550	0.582	-0.012	0.022
dayofweek_Monday	0.0306	0.014	2.236	0.026	0.004	0.057
dayofweek_Saturday	0.0580	0.008	7.533	0.000	0.043	0.073
dayofweek_Sunday	0.0345	0.008	4.091	0.000	0.018	0.051
dayofweek_Thursday	0.0169	0.007	2.399	0.017	0.003	0.031
dayofweek_Tuesday	0.0250	0.015	1.628	0.104	-0.005	0.055
dayofweek_Wednesday	0.0456	0.005	9.432	0.000	0.036	0.055
season_Spring	0.0234	0.006	4.060	0.000	0.012	0.035
season_Summer	0.0443	0.006	7.514	0.000	0.033	0.056
season_Winter	0.0291	0.006	5.065	0.000	0.018	0.040

(a) Model summary

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.050051	0.039293	0.774969	0.767445	8.830589

(b) Model Performance on train data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051085	0.040841	0.769931	0.755201	8.914241

(c) Model Performance on test data

```
View_counts = 0.025755954218990795 + (0.15479798531328332)*(visitors) + (-8.751800999103212e-07)*(ad_impressions) + (0.00233706181227939)*(views_trailer)
+ (-0.061317774474314676)*(major_sports_event_yes) + (0.011428201612989995)*(genre_Comedy) + (0.01164571413860957)*(genre_Drama) + (0.007411579149033405)
*(genre_Horror) + (0.005552040053479907)*(genre_Others) + (-0.0019530201076793374)*(genre_Romance) + (0.010099086222761115)*(genre_Sci-Fi) + (0.004751552
092982891)*(genre_Thriller) + (0.03059292899060472)*(dayofweek_Monday) + (0.0579709959930875)*(dayofweek_Saturday) + (0.03446085468072264)*(dayofweek_Sun
day) + (0.01692323992153946)*(dayofweek_Thursday) + (0.0249761721449454)*(dayofweek_Tuesday) + (0.04560173218269493)*(dayofweek_Wednesday) + (0.023368820
57626128)*(season_Spring) + (0.04430825834540765)*(season_Summer) + (0.029065021735611256)*(season_Winter)
```

Figure 14: Model 1

The R-squared value tells us that our model can explain 77.5% of the variance in the training set.

- The coefficients tell us how one unit change in X can affect y.
- The sign of the coefficient indicates if the relationship is positive or negative.
- In this data set, for example, an increase of 1 visitor occurs with a 0.1548 increase in view count or in other words 1 million view counts can be increased by $\frac{1}{.1548} \approx 6.45 \sim 6.5$ million visitors visiting the platform in the past week and a unit increase in major sports event occurs with a 0.0568 million decrease in the view count. Similarly, the same explanation applies to the other coefficients as well.
- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

(a) **Null hypothesis** - Contribution of the column i.e. the coefficient is zero.

(b) **Alternate hypothesis** - Contribution of the column i.e. the coefficient is not zero.

If p-value is > 0.05 then we accept our null hypothesis.

5.2 Testing the assumptions of linear regression model

We will be checking the following Linear Regression assumptions:

- No Multicollinearity
- Linearity of variables
- Independence of error terms
- Normality of error terms
- No Heteroscedasticity

5.2.1 No Multicollinearity

Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is "inflated" by the existence of correlation among the predictor variables in the model. We can clearly observe the vif of all the factors are less than 3. Hence we can look into the p-values and then drop the columns which are not statistically significant.

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	196.6
Date:	Sat, 16 Nov 2024	Prob (F-statistic):	1.41e-196
Time:	16:59:09	Log-Likelihood:	1020.2
No. Observations:	650	AIC:	-2016.
Df Residuals:	638	BIC:	-1963.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0335	0.019	1.718	0.086	-0.005	0.072
visitors	0.1537	0.011	14.149	0.000	0.132	0.175
views_trailer	0.0023	5.98e-05	39.037	0.000	0.002	0.002
major_sports_event_yes	-0.0616	0.004	-14.772	0.000	-0.070	-0.053
dayofweek_Monday	0.0305	0.014	2.247	0.025	0.004	0.057
dayofweek_Saturday	0.0571	0.008	7.485	0.000	0.042	0.072
dayofweek_Sunday	0.0324	0.008	3.896	0.000	0.016	0.049
dayofweek_Thursday	0.0154	0.007	2.212	0.027	0.002	0.029
dayofweek_Wednesday	0.0452	0.005	9.528	0.000	0.036	0.054
season_Spring	0.0239	0.006	4.191	0.000	0.013	0.035
season_Summer	0.0439	0.006	7.559	0.000	0.033	0.055
season_Winter	0.0304	0.006	5.349	0.000	0.019	0.041

	feature	VIF
0	const	95.414752
1	visitors	1.013861
2	views_trailer	1.018334
3	major_sports_event_yes	1.040906
4	dayofweek_Monday	1.043717
5	dayofweek_Saturday	1.135693
6	dayofweek_Sunday	1.119728
7	dayofweek_Thursday	1.161694
8	dayofweek_Wednesday	1.285127
9	season_Spring	1.544629
10	season_Summer	1.548045
11	season_Winter	1.570438

(a) Model summary

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.05048	0.039486	0.771101	0.767155	8.845579

(b) Model Performance on train data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051637	0.041431	0.764936	0.757286	9.055561

(c) Model Performance on test data

```
View_counts = 0.03345832276267558 + (0.15368466239808792)*(visitors) + (0.0023354539674109785)*(views_trailer) + (-0.061631103365330564)*(major_sports_ev
ent_yes) + (0.030481686264103303)*(dayofweek_Monday) + (0.05714656661505086)*(dayofweek_Saturday) + (0.032380950465093754)*(dayofweek_Sunday) + (0.015433
114921408614)*(dayofweek_Thursday) + (0.0451748635632742)*(dayofweek_Wednesday) + (0.0238643166479003)*(season_Spring) + (0.043911296611955544)*(season_S
ummer) + (0.030353496722740486)*(season_Winter)
```

Figure 15: Model 2

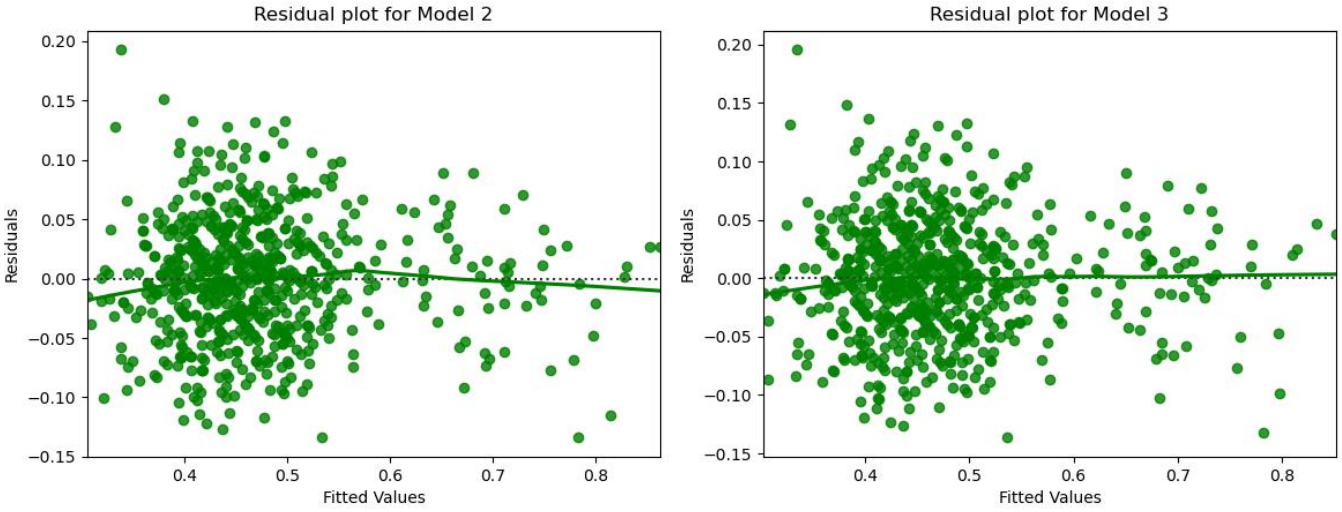
- Dropping the high p-value predictor variables has not adversely affected the model performance as R^2 is unchanged.
- This shows that these variables do not significantly impact the target variable.

5.2.2 TEST FOR LINEARITY AND INDEPENDENCE

Why the test?

- Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable. How to check linearity?
- Make a plot of fitted values vs residuals. If they don't follow any pattern (the curve is a straight line), then we say the model is linear otherwise model is showing signs of non-linearity. How to fix if this assumption is not followed?

We can try different transformations. I have plotted the pairplot between all the parameters to observe any pattern in the data and observed that **'views_content'** is non-linear to **'views_trailer'.** Hence we try taking squareroot of 1 column and re-consider the model again.



(a) Model 2 (b) Model 3

Figure 16: Residual plots

Residual Analysis

Model 2:

- The residuals show a visible curve (non-linear pattern), especially towards higher fitted values.
- This suggests that Model 2 might be missing some non-linear relationships or has some other specification issues.

Model 3:

- The residuals appear more randomly scattered with a lesser tendency to show a trend.
- There is slight heteroscedasticity, but it is not very pronounced.

AIC and BIC

1. Akaike Information Criterion (AIC)

Definition: Measures model quality, penalizing complexity to avoid overfitting. **Formula:**

$$AIC = -2\ln(L) + 2k$$

Where:

- $\ln(L)$: Log-likelihood of the model.
- k: Number of parameters (including intercept).

Interpretation: Lower AIC is better. Penalizes complex models to ensure simplicity.

2. Bayesian Information Criterion (BIC)

Definition: Similar to AIC but applies a stronger penalty for complexity, based on Bayesian principles.
Formula:

$$BIC = -2\ln(L) + k\ln(n)$$

Where:

- $\ln(L)$: Log-likelihood.
- k: Number of parameters.
- n: Number of observations.

Interpretation: Lower BIC is better. Stronger penalty than AIC, especially for large n.

3. Comparison

- **AIC:** Penalizes complexity less, works well with small datasets.
- **BIC:** Stronger penalty, prefers simpler models with large datasets.

4. When to Use

- **AIC:** Focus on prediction; small sample size.
- **BIC:** Focus on simplicity; large sample size.

Model Comparison

Model	AIC	BIC
Model 2	-2016.42	-1962.70
Model 3	-2021.22 (lower AIC)	-1963.02 (lower BIC)

Table 1: Comparison of Models based on AIC and BIC

Conclusion: Model 3 is preferred due to its lower AIC and BIC, indicating a better trade-off between fit and complexity.

5.2.3 TEST FOR NORMALITY for Model 3

What is the test?

- Error terms/residuals should be normally distributed.
- If the error terms are not normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.

What does non-normality indicate?

- It suggests that there are a few unusual data points which must be studied closely to make a better model. How to check the Normality?
- It can be checked via QQ Plot - residuals following normal distribution will make a straight line plot, otherwise not.
- Another test to check for normality is the Shapiro-Wilk test.
- How to Make residuals normal?
- We can apply transformations like log, exponential, arcsinh, etc as per our data.

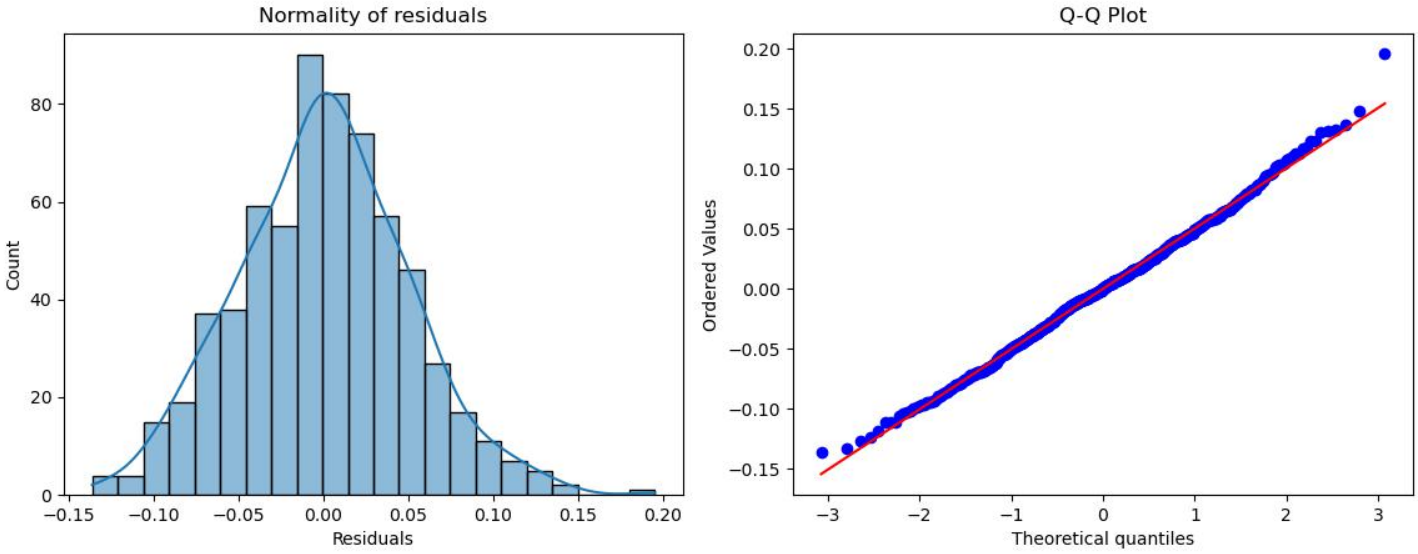


Figure 17: Test for normality for Model 3

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

Null hypothesis - Data is normally distributed.

Alternate hypothesis - Data is not normally distributed.

The Shapiro-Wilk test yields a p-value of 0.295, which is greater than 0.05. Thus, we can conclude that the residuals follow a normal distribution.

5.2.4 TEST FOR HOMOSCEDASTICITY

- **Homoscedacity** - If the variance of the residuals are symmetrically distributed across the regression line , then the data is said to homoscedastic.
- **Heteroscedacity** - If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic. In this case the residuals can form an arrow shape or any other non symmetrical shape.

Why the test?

- The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers. How to check if model has Heteroscedasticity?
- Can use the goldfeldquandt test. If we get p-value > 0.05 we can say that the residuals are homoscedastic, otherwise they are heteroscedastic. How to deal with Heteroscedasticity?

Can be fixed via adding other important features or making transformations.

The p-value for the Goldfeld-Quandt test is 0.134, indicating that we fail to reject the null hypothesis. This suggests that the residuals exhibit homoscedasticity.

5.3 Model 3

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.775
Model:	OLS	Adj. R-squared:	0.770
Method:	Least Squares	F-statistic:	182.3
Date:	Sun, 17 Nov 2024	Prob (F-statistic):	7.33e-197
Time:	15:59:49	Log-Likelihood:	1023.6
No. Observations:	650	AIC:	-2021.
Df Residuals:	637	BIC:	-1963.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1330	0.067	-1.980	0.048	-0.265	-0.001
visitors	0.1533	0.011	14.175	0.000	0.132	0.175
views_trailer	0.0005	0.001	0.701	0.483	-0.001	0.002
major_sports_event_yes	-0.0629	0.004	-15.037	0.000	-0.071	-0.055
dayofweek_Monday	0.0287	0.014	2.119	0.034	0.002	0.055
dayofweek_Saturday	0.0572	0.008	7.523	0.000	0.042	0.072
dayofweek_Sunday	0.0326	0.008	3.938	0.000	0.016	0.049
dayofweek_Thursday	0.0147	0.007	2.121	0.034	0.001	0.028
dayofweek_Wednesday	0.0448	0.005	9.488	0.000	0.036	0.054
season_Spring	0.0248	0.006	4.359	0.000	0.014	0.036
season_Summer	0.0444	0.006	7.676	0.000	0.033	0.056
season_Winter	0.0301	0.006	5.336	0.000	0.019	0.041
views_trailer_sq	0.0363	0.014	2.588	0.010	0.009	0.064

Omnibus:	2.867	Durbin-Watson:	2.016
Prob(Omnibus):	0.239	Jarque-Bera (JB):	2.680
Skew:	0.132	Prob(JB):	0.262
Kurtosis:	3.172	Cond. No.	2.58e+03

(a) Model summary

	feature	VIF
0	const	1145.116186
1	visitors	1.014056
2	views_trailer	145.511376
3	major_sports_event_yes	1.054886
4	dayofweek_Monday	1.046542
5	dayofweek_Saturday	1.135698
6	dayofweek_Sunday	1.119829
7	dayofweek_Thursday	1.163413
8	dayofweek_Wednesday	1.286290
9	season_Spring	1.550343
10	season_Summer	1.549801
11	season_Winter	1.570749
12	views_trailer_sq	145.740291

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.050101	0.039362	0.774526	0.769918	8.820493

(b) Model Performance on train data

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.050358	0.040247	0.776434	0.767784	8.833591

(c) Model Performance on test data

```
View_counts = -0.13300112523075014 + (0.1532960908166131)*(visitors) + (0.0004991639606274113)*(views_trailer) + (-0.06287701873377341)*(major_sports_eve
nt_yes) + (0.02866310022717379)*(dayofweek_Monday) + (0.05718528728778898)*(dayofweek_Saturday) + (0.03258383648259437)*(dayofweek_Sunday) + (0.014741522
633178503)*(dayofweek_Thursday) + (0.044807260634074926)*(dayofweek_Wednesday) + (0.02475667172646963)*(season_Spring) + (0.04441538891477993)*(season_Su
mmer) + (0.030147521579935166)*(season_Winter) + (0.03633200300044395)*(views_trailer_sq)
```

(d) Equation of the model

Figure 18: Model 3

Despite the high VIF values for **'views_trailer_sq'** and **'views_trailer'** both variables have been retained in the model. This decision considers the observed patterns in the pairplot, the lower AIC and BIC values of Model 3, and the overall improvement in model performance metrics such as RMSE, MAE, MAPE, and Adjusted R^2.We also note that **'views_trailer'** has a p-value of 0.483. Therefore, we exclude this column to develop Model 4.

5.4 Final model(model 4) performance evaluation

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.774
Model:	OLS	Adj. R-squared:	0.770
Method:	Least Squares	F-statistic:	199.0
Date:	Sun, 17 Nov 2024	Prob (F-statistic):	6.48e-198
Time:	17:07:52	Log-Likelihood:	1023.4
No. Observations:	650	AIC:	-2023.
Df Residuals:	638	BIC:	-1969.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1777	0.021	-8.399	0.000	-0.219	-0.136
visitors	0.1532	0.011	14.171	0.000	0.132	0.174
major_sports_event_yes	-0.0632	0.004	-15.215	0.000	-0.071	-0.055
dayofweek_Monday	0.0282	0.014	2.087	0.037	0.002	0.055
dayofweek_Saturday	0.0572	0.008	7.528	0.000	0.042	0.072
dayofweek_Sunday	0.0326	0.008	3.946	0.000	0.016	0.049
dayofweek_Thursday	0.0146	0.007	2.099	0.036	0.001	0.028
dayofweek_Wednesday	0.0447	0.005	9.476	0.000	0.035	0.054
season_Spring	0.0250	0.006	4.414	0.000	0.014	0.036
season_Summer	0.0446	0.006	7.712	0.000	0.033	0.056
season_Winter	0.0301	0.006	5.334	0.000	0.019	0.041
views_trailer_sq	0.0461	0.001	39.306	0.000	0.044	0.048

	feature	VIF
0	const	113.683358
1	visitors	1.013792
2	major_sports_event_yes	1.041906
3	dayofweek_Monday	1.043740
4	dayofweek_Saturday	1.135694
5	dayofweek_Sunday	1.119713
6	dayofweek_Thursday	1.162031
7	dayofweek_Wednesday	1.285236
8	season_Spring	1.544140
9	season_Summer	1.547473
10	season_Winter	1.570713
11	views_trailer_sq	1.019936

(a) Model summary

RMSE	MAE	R-squared	Adj. R-squared	MAPE
0.05012	0.039397	0.774352	0.770101	8.819924

(b) Model Performance on train data

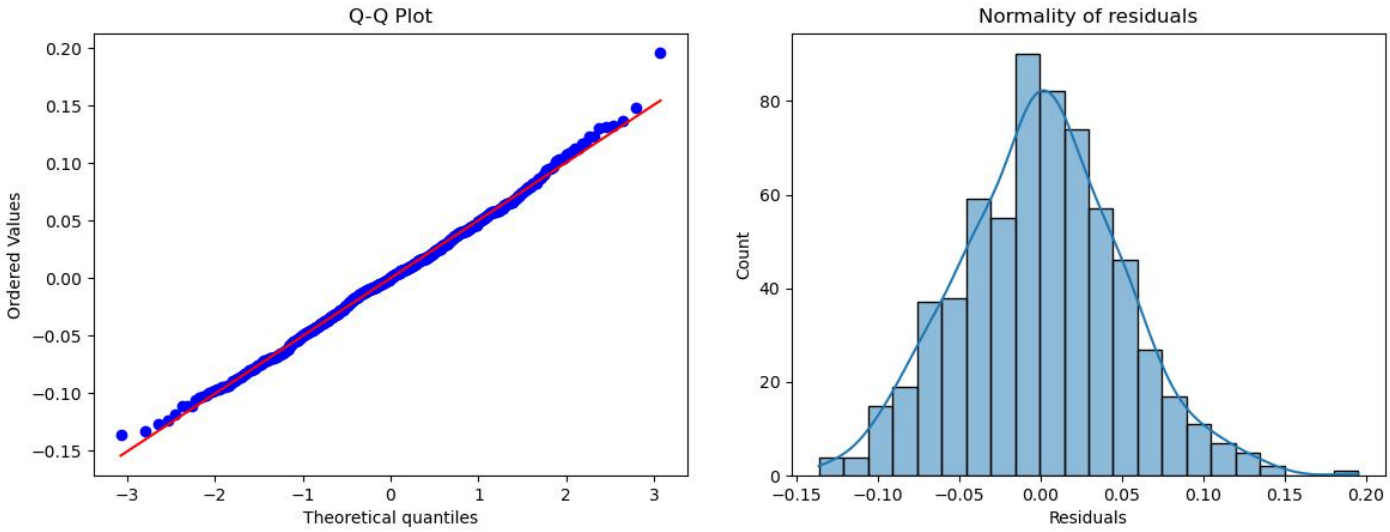
RMSE	MAE	R-squared	Adj. R-squared	MAPE
0.050238	0.040141	0.777499	0.769576	23.772465

(c) Model Performance on test data

View_counts = -0.1776975693750753 + (0.15317356250470612)*(visitors) + (-0.06320221055399805)*(major_sports_event_yes) + (0.028172425094917342)*(dayofweek_Monday) + (0.057195094985482 17)*(dayofweek_Saturday) + (0.032642763092627 01)*(dayofweek_Sunday) + (0.014573583386866318)*(dayofweek_Thursday) + (0.044712483106538 3)*(dayofweek_Wednesday) + (0.025008523634690312)*(season_Spring) + (0.04457259558897844)*(season_Summer) + (0.030128374512302 68)*(season_Winter) + (0.046139118233662894)*(views_trailer_sq)

(d) Equation of the model

Figure 19: Model 4



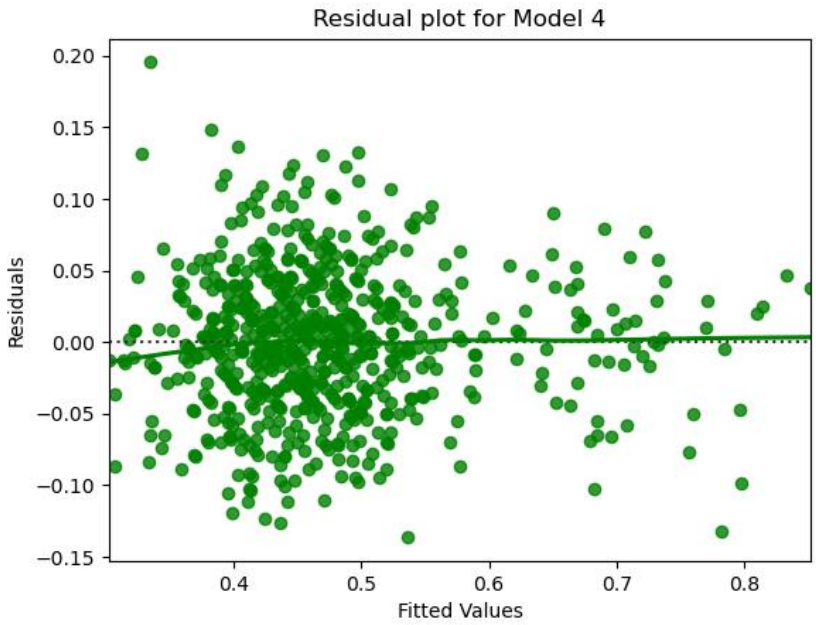


Figure 20: Test for **normality** and **homoscedacity** for final model 4

Model	AIC	BIC
Model 2	-2016.42	-1962.70
Model 3	-2021.22 (lower AIC than Model 2)	-1963.02 (lower BIC than Model 2)
Model 4	-2022.72 (lowest AIC)	-1968.99 (lowest BIC)

Table 2: Comparison of Models based on AIC and BIC

Conclusion

Model 4 is the preferred model as it achieves the lowest AIC and BIC values, indicating the best balance between model fit and complexity among the compared models.

This is our final model with every condition satisfied along with the model assumptions. This decision considers the observed patterns in the pairplot, the lower AIC and BIC values of Model 4, and the overall improvement in model performance metrics such as RMSE, MAE, MAPE, and Adjusted R^2.This model satifies both Shapiro and Goldfeld-Quandt test as well.

- We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
- MAE indicates that our current model is able to predict view counts within a mean error of 0.04 units on the test data.
- Hence, we can conclude the model "Model4" is good for prediction as well as inference purposes.

6 Actionable Insights & Recommendations

(a) With our linear regression model, we have been able to capture ~77.4% of the variation in our data.

(b) The model indicates that the most significant predictors of the number of first-day views, in millions, of the content are the following:

- visitors

- existance of major_sports_event (on the day of content release).
- dayofweek (Monday,Wednesday,Thursday ,Saturday and Sunday)
- Season (Spring,Summer,Winter)
- views_trailer

(**The p-values for these predictors are less than 0.05 in our final model.**)

(a) 3. an increase of 1 visitor occurs with a 0.1532 increase in view count or in other words 1 million view counts can be increased by $\frac{1}{.1532} \approx 6.53 \sim 6.5$ million visitors visiting the platform in the past week

It is important to note here that the predicted values are square(views_trailer) and therefore coefficients have to be converted accordingly to understand their influence on view counts. It is important to note here that correlation is not equal to causation.

4. If there is any one major sports event on the day of content release then there is 0.063 million decrease in view counts

5.The categorical variables are a little hard to interpret. It can be seen that all the dayofweek_category variables in the dataset have a positive relationship with the view counts, and the magnitude of this positive relationship is high for Wednesday,Sunday and Saturday as already visualized through a boxplot(12).

6.It can be seen that all the season_category variables in the dataset have a positive relationship with the view counts, and the magnitude of this positive relationship is high for Summer,Winter and Spring respectively following the same pattern as already visualized through the boxplot(**??**).

4. As the number of views, in millions, of the content trailer increases, the number of first-day views, in millions, of the content also increases.

Our final Linear Regression model has a MAPE of 23% on the test data, which means that we are able to predict within 23% of the content views. This is a very good model and we can use this model for the benefit of the company.