BUSINESS REPORT

Predictive Modelling

FEBRUARY 8, 2025

SAMRUDDHI POKALE
pokalesw@gmail.com

Table of Contents

Table of Figures

Problem Statement

I. Context

OTT media services, offering on-demand films and shows via the internet, are rapidly growing as a trending technology globally. Valued at $121.61 billion in 2019, the market is projected to reach $1,039.03 billion by 2027, with a CAGR of 29.4%. Factors like ease of access, better connectivity, and changing viewer preferences are driving this shift from traditional TV. The COVID-19 pandemic further accelerated growth, with some platforms seeing a 46% rise in consumption as viewers seek fresh content. Innovations continue to make OTT platforms more attractive to subscribers.

II. Objective

The primary objective is to analyse the data and build a linear regression model to determine the key factors that drive first-day content viewership on ShowTime's OTT platform. Some specific goals are:

- Determine which variables have the greatest impact on first-day viewership.
- Analyse potential reasons for the decline in first-day viewership.
- Build and evaluate a linear regression model to quantify the influence of each variable and predict future first-day viewership.
- Provide business recommendations based on observations and insights for ShowTime.

III. Data Description

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

Data Overview

I. Structure of data

```
Structure of the data:
   visitors  ad_impressions  major_sports_event     genre  dayofweek  season  \
0      1.67         1113.81                   0    Horror  Wednesday  Spring
1      1.46         1498.41                   1  Thriller     Friday    Fall
2      1.47         1079.19                   1  Thriller  Wednesday    Fall
3      1.85         1342.77                   1    Sci-Fi     Friday    Fall
4      1.46         1498.41                   0    Sci-Fi     Sunday  Winter

   views_trailer  views_content
0          56.70           0.51
1          52.69           0.32
2          48.74           0.39
3          49.81           0.44
4          55.83           0.46

Column Names:
Index(['visitors', 'ad_impressions', 'major_sports_event', 'genre',
       'dayofweek', 'season', 'views_trailer', 'views_content'],
      dtype='object')

Data Shape:
Number of rows =  1000
Number of columns =  8
```

Figure 1 Structure of data

- Here we can see the first five entries in the ShowTime dataset.
- There are total 1000 entries in 8 columns.
- There are no duplicates.

II. Types of Data

```
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   visitors            1000 non-null   float64
 1   ad_impressions      1000 non-null   float64
 2   major_sports_event  1000 non-null   int64
 3   genre               1000 non-null   object
 4   dayofweek           1000 non-null   object
 5   season              1000 non-null   object
 6   views_trailer       1000 non-null   float64
 7   views_content       1000 non-null   float64
dtypes: float64(4), int64(1), object(3)
memory usage: 62.6+ KB
```

Figure 2 Types of data

- There are three different datatypes in this dataset:
  Integer(int64), Float(float64) and Object

III. Statistical Summary

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

Figure 3 Statistical Summary

- Numerical Columns:
    1. **count**: Number of non-missing values.
    2. **mean**: Average value.
    3. **std**: Standard deviation.
    4. **min**: Minimum value.
    5. **25%, 50%, 75%:** Percentile values (quartiles).
    6. **max**: Maximum value.
- Categorical Columns:
    1. **count**: Number of non-missing values.
    2. **unique**: Number of unique categories.
    3. **top**: Most frequent category.
    4. **freq**: Frequency of the most frequent category.

IV. Univariate Analysis

Numerical Variables

- "Visitor's" skewness is 0.37, suggests a fairly symmetric distribution, though slightly positively skewed. Most values are centred around the mean, with a small tail on the right.
- A positive skew of 1.03 of Ad Impressions indicates a longer right tail. This suggests that while most values are clustered toward the lower range, there are some high outliers increasing the mean.
- Similar to "Visitors", "Major_sports_even"t shows a mild positive skew 0f 0.4, indicating a slight rightward tail but still relatively symmetric.
- "Views_Trailer" has a skewness of 2.37 which represents a strong positive skew, suggesting a highly asymmetrical distribution. Most values are low, but there are a few very high values that significantly impact the mean.

- “Views_content” has moderate positive skew of 0.94, this distribution is less symmetric than “Visitors” and “Major_Sports_Event” with higher values being more frequent but some large outliers stretching the tail.

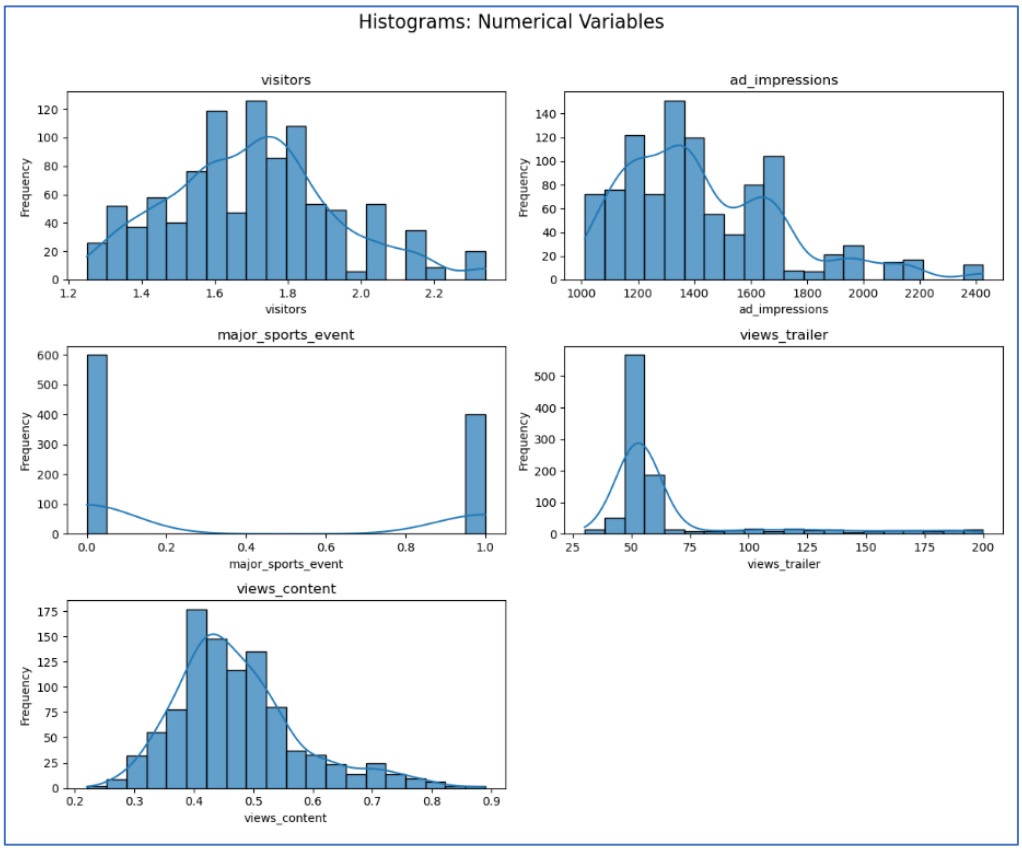


Figure 4 Histograms: Univariate Variables

Categorical Variables

Unique value counts for categorical variables:

- Genre
  1. Others 255
  2. Comedy 114
  3. Thriller 113
  4. Drama 109
  5. Romance 105
  6. Sci-Fi 102
  7. Horror 101
  8. Action 101
- Dayofweek
  1. Friday 369
  2. Wednesday 332
  3. Thursday 97
  4. Saturday 88
  5. Sunday 67
  6. Monday 24
  7. Tuesday 23
- Season
  - Winter 257
  - Fall 252
  - Spring 247
  - Summer 244

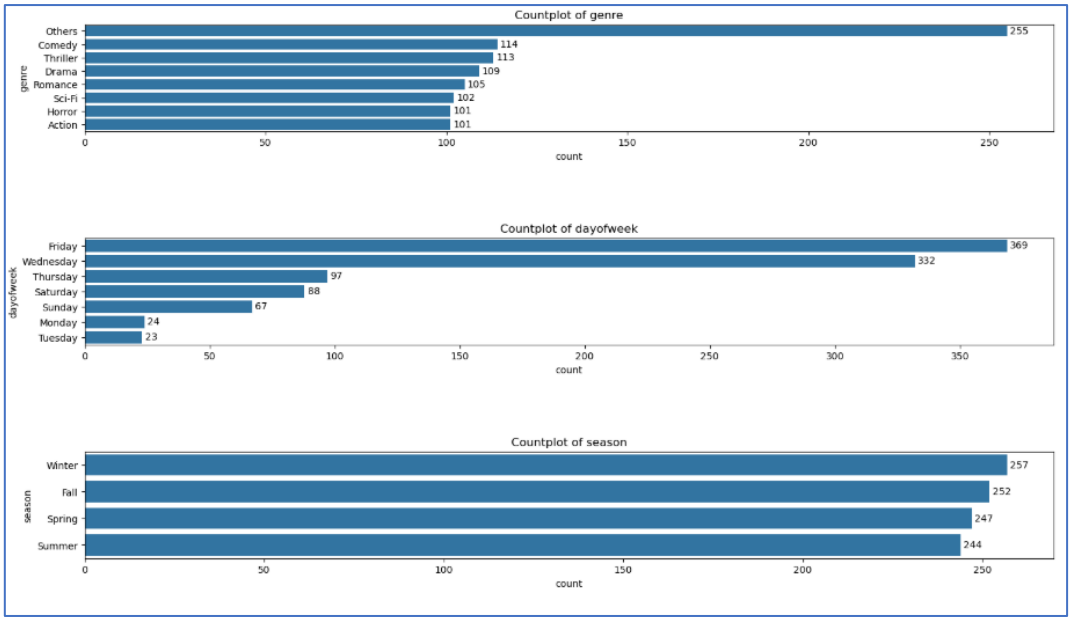


Figure 5 Countplot : Categorical Variables

V. Bivariate Analysis

Relationship between all numerical variables

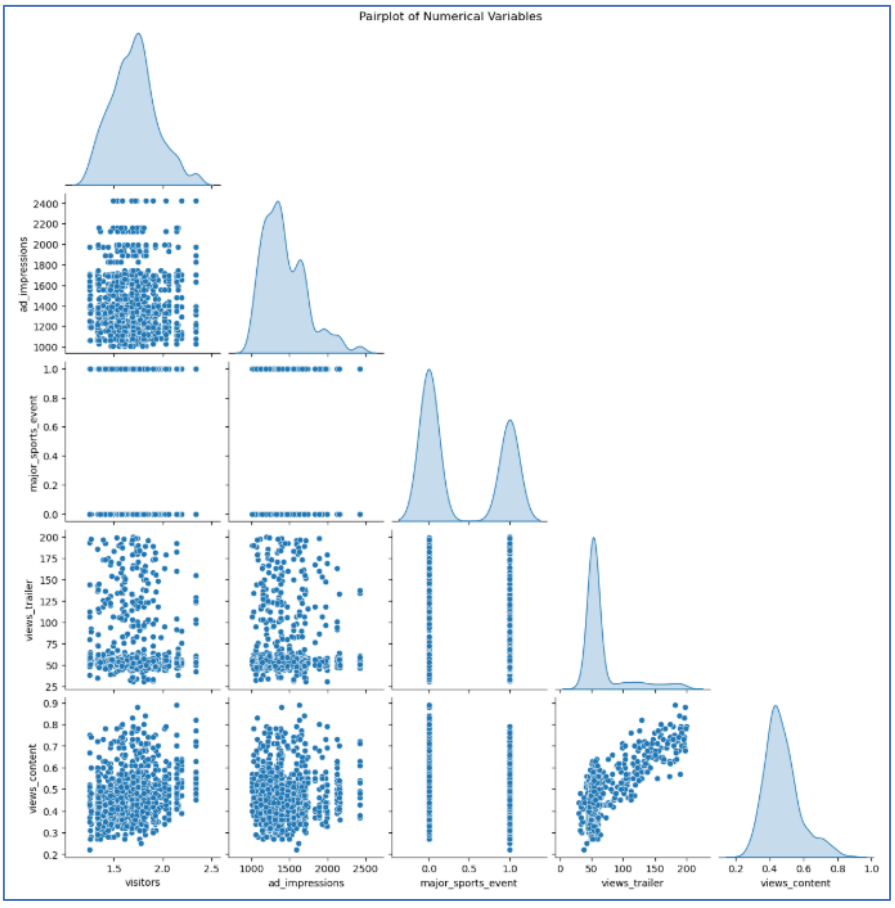


Figure 6 Pair plot: Numerical Variables

Correlation between numerical values

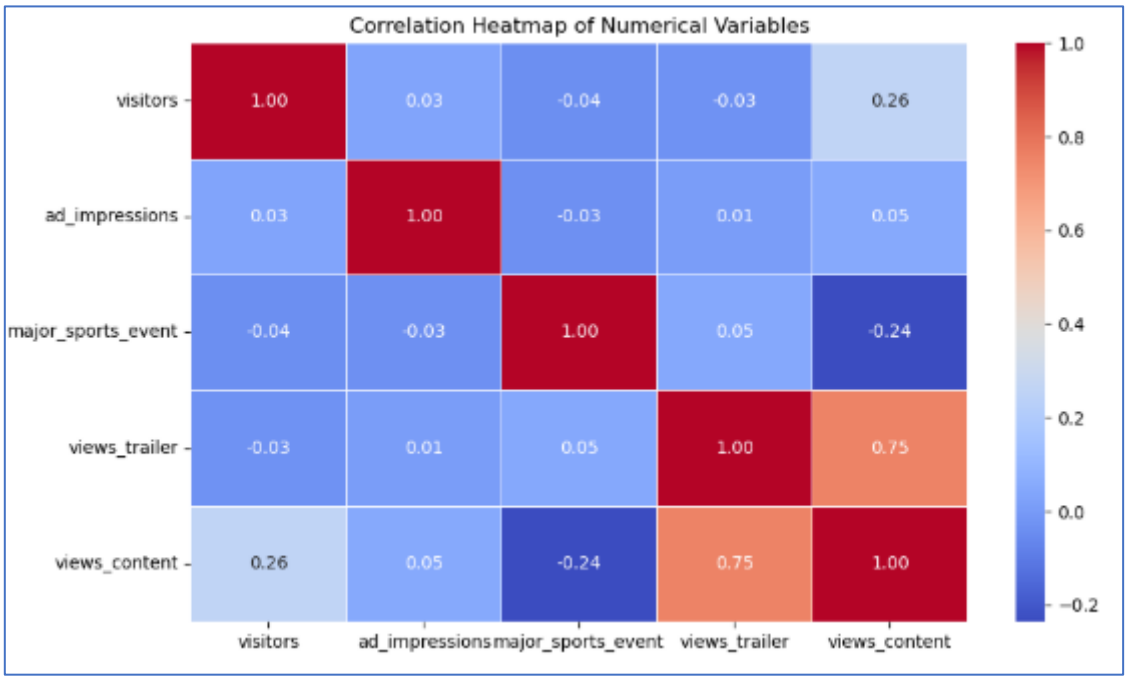


Figure 7 Correlation heatmap

Strong Positive Correlations:

1. **Views Trailer and Views Content (0.75):** Higher trailer views strongly correspond to higher content views, indicating a significant relationship.

Moderate Positive Correlations:

1. **Visitors and Views Content (0.26):** A moderate relationship, an increase in visitors slightly increases content views.

Weak Positive Correlations:

1. **Ad Impressions and Views Content (0.05):** A very weak relationship indicates ad impressions have minimal influence on content views.
2. **Major Sports Event and Views Trailer (0.05):** A weak correlation means sports events are marginally influenced by trailer views.

Weak Negative Correlations:

1. **Major Sports Event and Views Content (-0.24):** Fewer content views are observed during major sports events, as there is weak negative relationship.
2. **Visitors and Major Sports Event (-0.04):** A negligible negative correlation suggests almost no relationship between these variables.

Relationship between categorical vs numerical variables

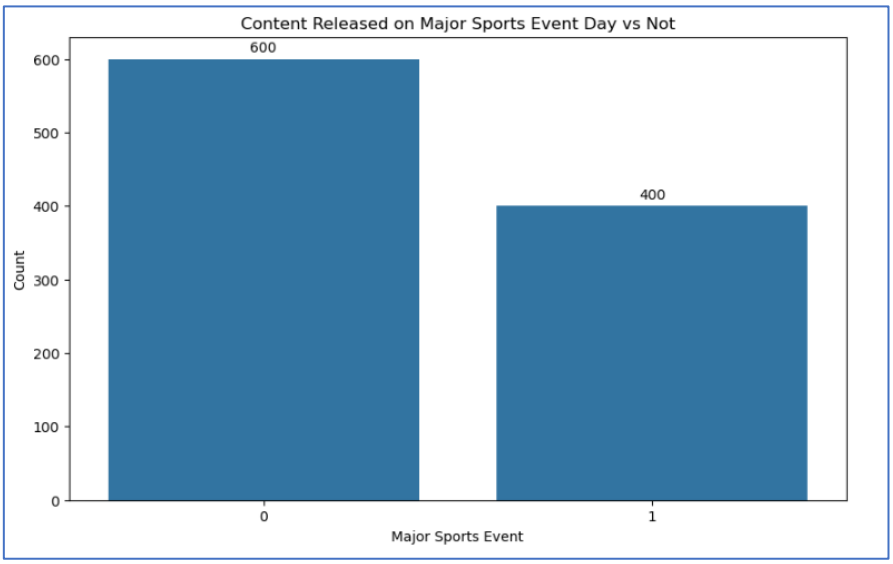


Figure 8 Content Released on Major Sports Day

- More Content is released when there is no major sports event.

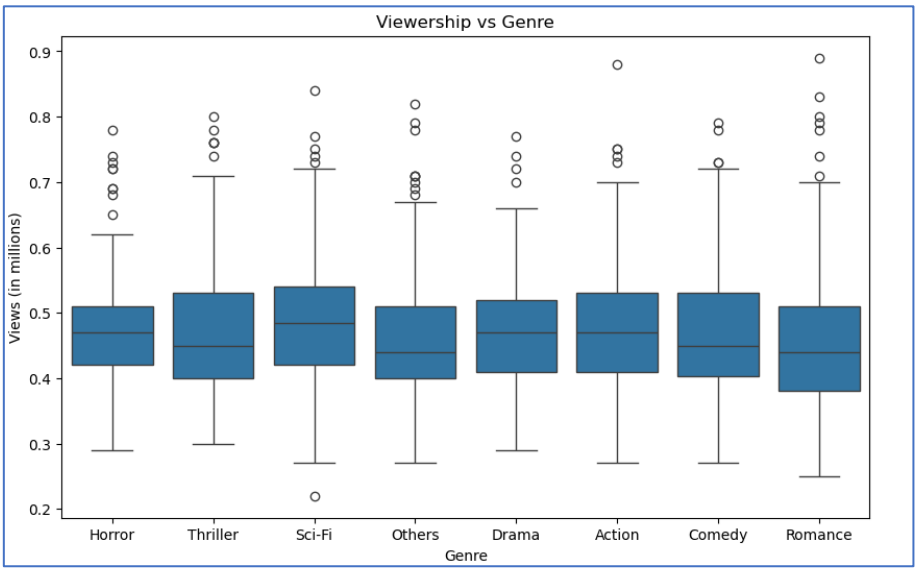


Figure 9 Viewership vs Genre

- Content viewership is consistent across all Genres with a mean of approximately 0.3 million views.

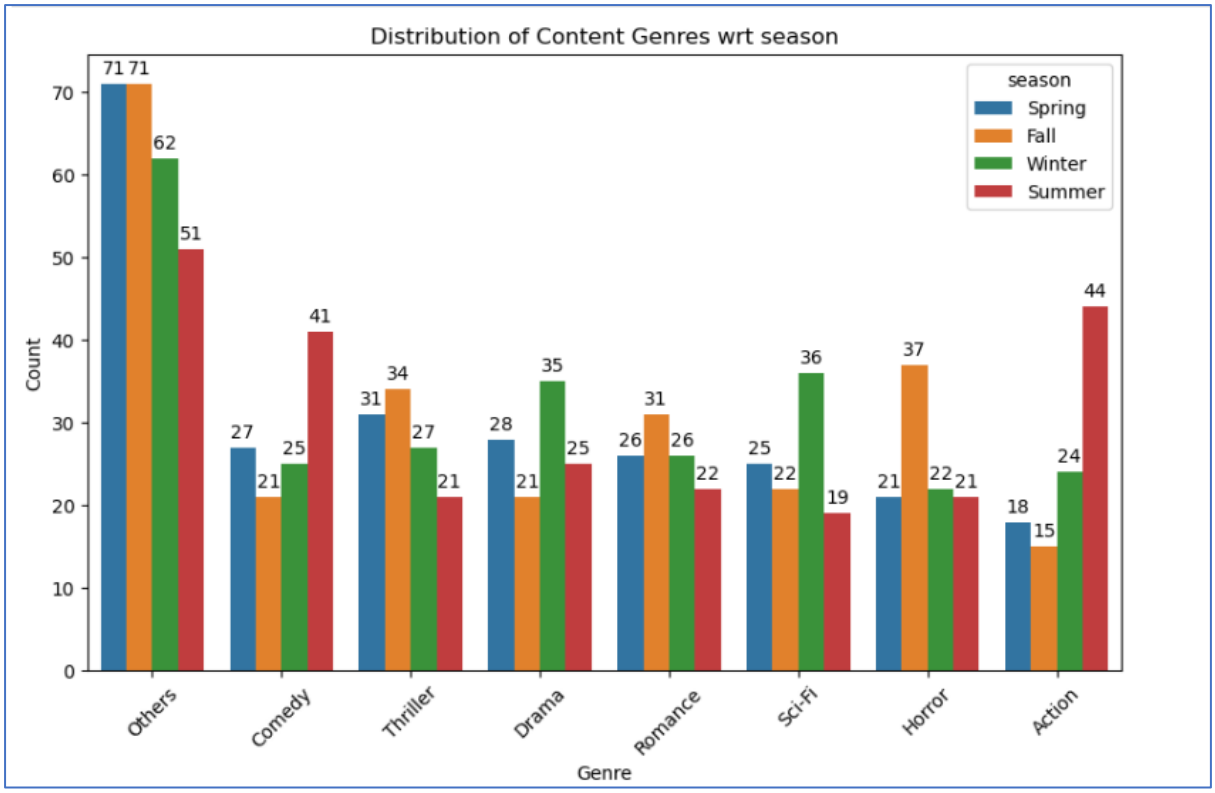


Figure 10 Distribution of Genres wrt season

- The greatest number of movies are released in “Others” category.
- All other genres have almost same number of contents.
- “Comedy” genre content is released the most in summer.
- “Thriller” is mostly released in Fall followed by Spring.
- “Drama” genre content is mostly released in winter.
- Maximum content release of “Romance” is in Fall while that of “Sci-Fi” is in winter.
- “Action” genre is released mostly in Summer and least in Fall.

VI. Key Questions

What does the distribution of content views look like?

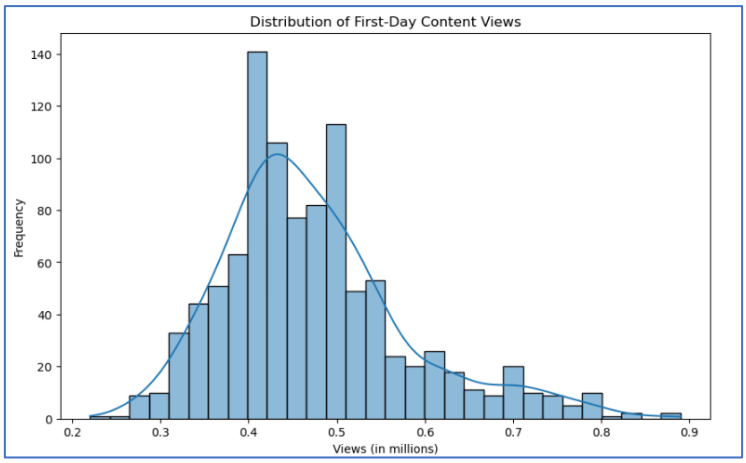


Figure 11 Distribution of Content

- The data is right-skewed, meaning that some of the higher values are falling off and the most of the first day content views are focused towards the lower values.The data is between 0.2 million and 0.9 million views.
- The highest bar in the histogram indicates that the peak frequency of content views is about 0.4 million views.
- On the first day, most of the content seems to get between 0.3 and 0.6 million views.
- The smooth density curve that is superimposed on histogram points to the multimodal and nonuniform distribution which can also be bimodal.

What does the distribution of genres look like?

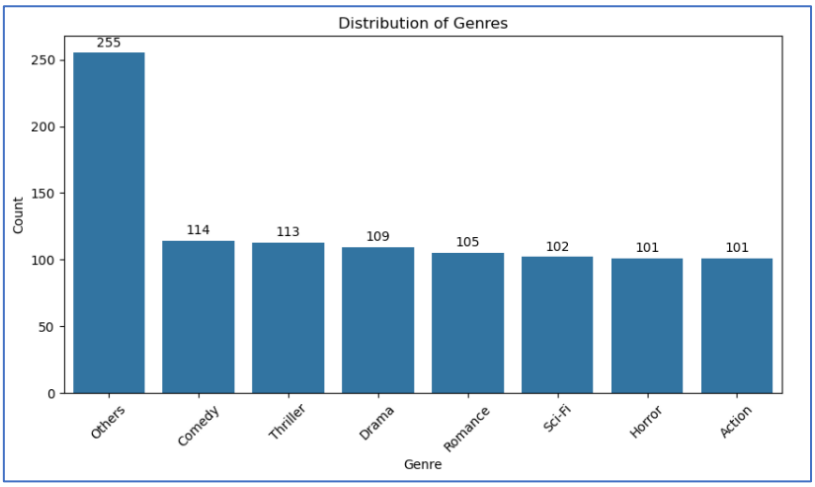


Figure 12 Distribution of content Genres

- The distribution of genres in your dataset shows that the "Others" category has the highest count (255), significantly more than any specific genre. Among the specific genres, Comedy (114), Thriller (113), and Drama (109) are the most common, while Action and Horror have the lowest counts (101 each).

The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?

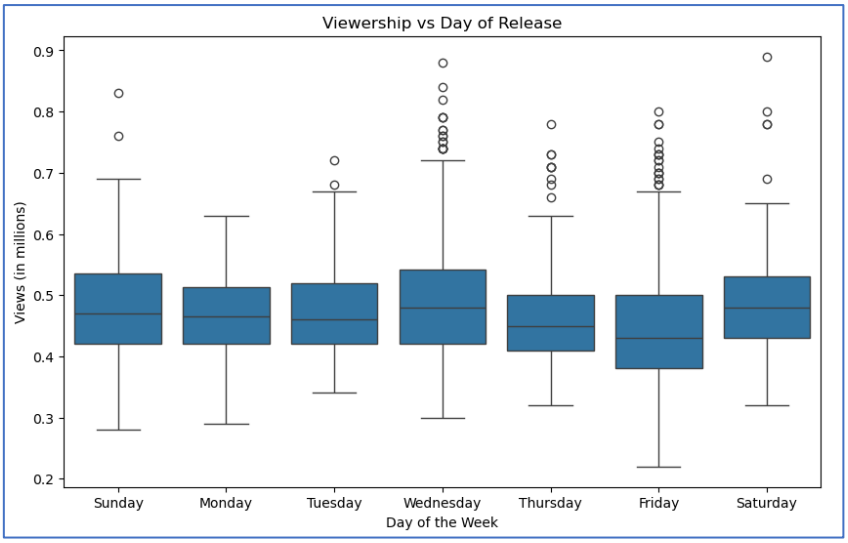


Figure 13 Viewership vs Day of Release

- Most days having similar median views around 0.5 million.
- Some outliers, like Wednesday, Thursday and Friday, indicate occasional high-performing releases.
- Least number of views are on Monday.

How does the viewership vary with the season of release?

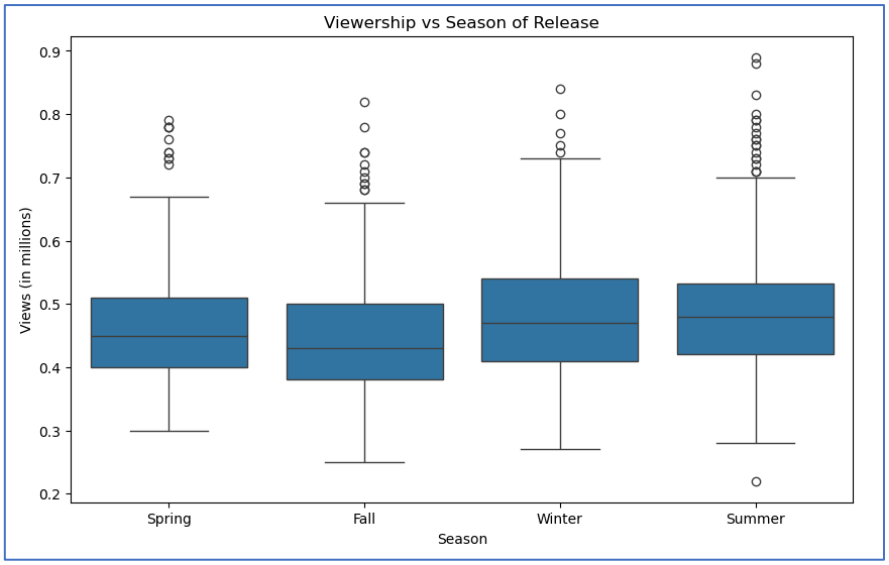


Figure 14 Viewership vs Season of release

- The median viewership is almost identical across all seasons ,suggesting a steady popularity across all seasons.
- There are many outliers in every season, with Summer having the most, indicating that some content gets much more views.

What is the correlation between trailer views and content views?

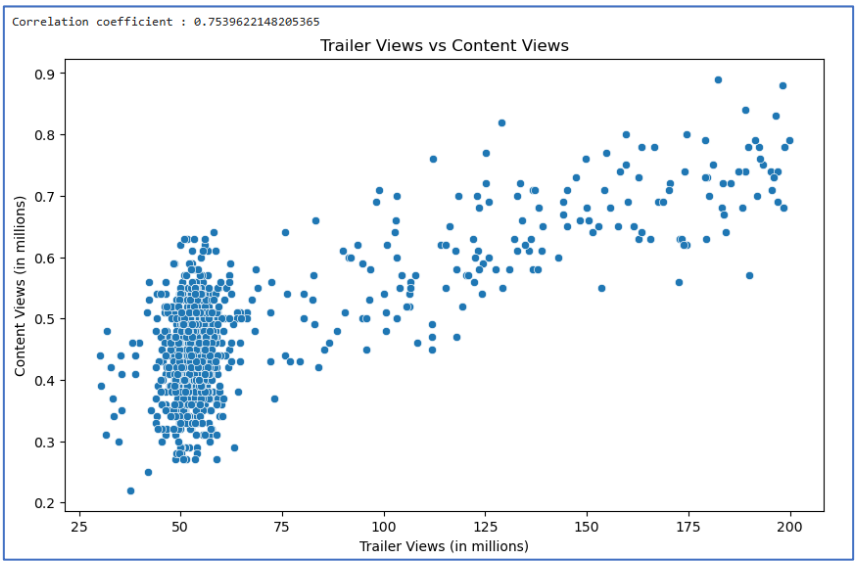


Figure 15 Trailer views vs content views

- The correlation coefficient is 0.75.
- There is a high positive correlation between trailer views and content views as content views is increasing with the increase in trailer views.

VII. Observations and insights

- Visitors range from 1.25 to 2.34, with a mean of 1.70.
- Ad Impressions range from 1010.87 to 2424.20, with a mean of 1434.71.
- Major Sports Event is binary (0 or 1), with 40% indicating an event.
- Views Trailer range from 30.08 to 199.92, with a mean of 66.92.
- Views Content range from 0.22 to 0.89, with a mean of 0.47.
- Most content is released just before the weekend to get more viewership.
- More number of views of the content trailer indicates a greater number of first-day views of the content.
- Visitors have no connection with occurrence major sports event.
- The viewership across days of the week is fairly consistent.

Data Pre-processing

I. Value check

- There are no duplicate values.
- There are no missing values.

```
Missing values per column:
visitors                0
ad_impressions          0
major_sports_event      0
genre                   0
dayofweek               0
season                  0
views_trailer           0
views_content           0
dtype: int64
```

Figure 16 Missing values

II. Outlier detection and treatment

- Outliers are detected using IQR method and extreme values are replaced with lower or upper bounds in the column.
- Formula for calculating lower and upper bounds:
  Q1 = data [column_name]. quantile (0.25)
  Q3 = data [column_name]. quantile (0.75)
  IQR = Q3 - Q1
  lower_bound = Q1 - 1.5 * IQR
  upper_bound = Q3 + 1.5 * IQR

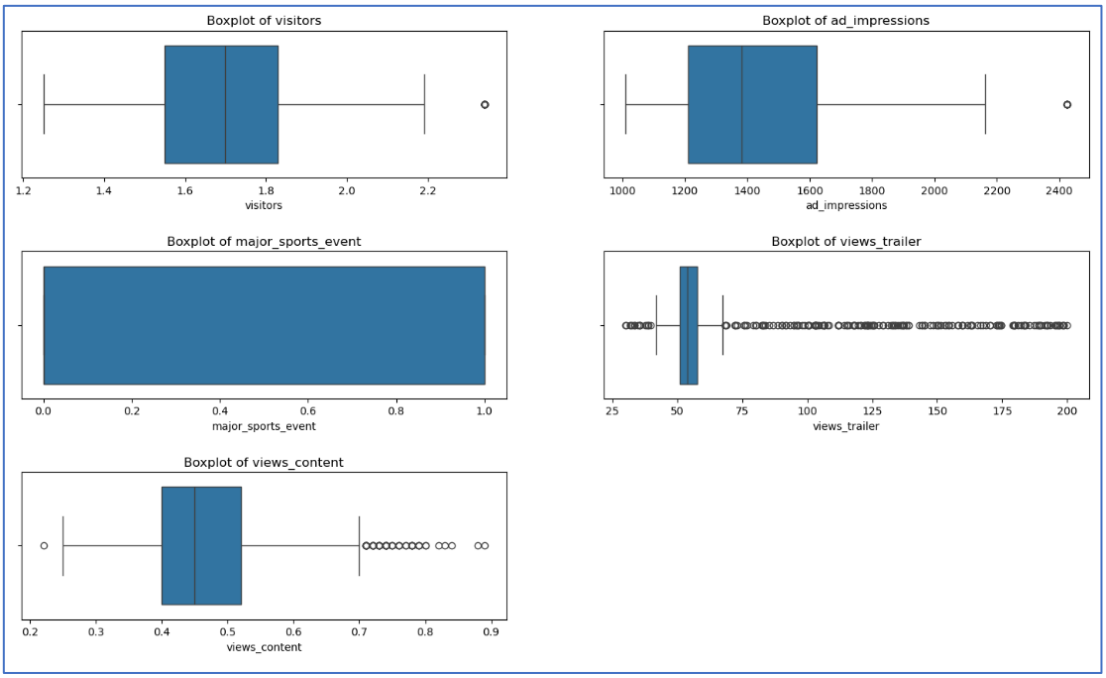


Figure 17 Boxplot for Outlier detection

- visitors: 20 outliers detected
- ad_impressions: 13 outliers detected
- major_sports_event: 0 outliers detected
- views_trailer: 189 outliers detected
- views_content: 47 outliers detected
- Outliers are not treated as Visitors, ad impressions, views_trailer, and views_content are numerical variables where extreme values reflect genuine data.

III. Feature Engineering

Aggregation Features
Ratios that reveal relative relationships:

- views_content / views_trailer (content engagement ratio).
- ad_impressions / visitors (ad exposure per visitor).

One-Hot Encoding:

- Categorical columns (genre, season, dayofweek, major_sports_event) are converted to dummy variables, excluding the first category (drop_first=True) to avoid multicollinearity.

Columns:

['visitors', 'ad_impressions', 'views_trailer', 'views_content','content_to_trailer_ratio', 'ad_per_visitor', 'genre_Comedy', 'genre_Drama', 'genre_Horror', 'genre_Others', 'genre_Romance', 'genre_Sci-Fi', 'genre_Thriller', 'season_Spring', 'season_Summer', 'season_Winter', 'dayofweek_Monday', 'dayofweek_Saturday', 'dayofweek_Sunday', 'dayofweek_Thursday', 'dayofweek_Tuesday', 'dayofweek_Wednesday', 'major_sports_event_1']

	visitors	ad_impressions	views_trailer	views_content	content_to_trailer_ratio	ad_per_visitor	genre_Comedy	genre_Drama	genre_Horror	genre_Others	...	season_Sp
0	1.67	1113.81	56.70	0.51	0.008995	666.951696	False	False	True	False	...	
1	1.46	1498.41	52.69	0.32	0.006073	1026.307516	False	False	False	False	...	
2	1.47	1079.19	48.74	0.39	0.008002	734.142358	False	False	False	False	...	
3	1.85	1342.77	49.81	0.44	0.008834	725.821229	False	False	False	False	...	
4	1.46	1498.41	55.83	0.46	0.008239	1026.307516	False	False	False	False	...	

5 rows × 23 columns

Figure 18 Data after Feature engineering

This preprocessing helps the dataset by:

- Capturing interactions between variables.
- Making categorical variables usable for the Linear Regression model.

IV. Data preparation for modelling

- X: Independent variables
- Y: Dependent variable(views_content)
- All the input attributes are converted into float type for modelling.
- The data is spilt in 70:30 ratio for train to test data.

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

Figure 19 Data split

Model building- Linear Regression

I. Model building

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.928
Model:	OLS	Adj. R-squared:	0.925
Method:	Least Squares	F-statistic:	394.9
Date:	Mon, 27 Jan 2025	Prob (F-statistic):	0.00
Time:	09:48:14	Log-Likelihood:	1495.1
No. Observations:	700	AIC:	-2944.
Df Residuals:	677	BIC:	-2840.
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2630	0.034	-7.802	0.000	-0.329	-0.197
visitors	0.0899	0.018	4.868	0.000	0.054	0.126
ad_impressions	-5.007e-05	2.16e-05	-2.320	0.021	-9.24e-05	-7.7e-06
views_trailer	0.0039	5.44e-05	71.373	0.000	0.004	0.004
content_to_trailer_ratio	39.2687	1.102	35.621	0.000	37.104	41.433
ad_per_visitor	8.78e-05	3.61e-05	2.430	0.015	1.69e-05	0.000
genre_Comedy	0.0055	0.005	1.169	0.243	-0.004	0.015
genre_Drama	0.0103	0.005	2.139	0.033	0.001	0.020
genre_Horror	0.0070	0.005	1.450	0.147	-0.002	0.016
genre_Others	0.0047	0.004	1.119	0.263	-0.004	0.013
genre_Romance	0.0059	0.005	1.181	0.238	-0.004	0.016
genre_Sci-Fi	0.0033	0.005	0.689	0.491	-0.006	0.013
genre_Thriller	0.0089	0.005	1.876	0.061	-0.000	0.018
season_Spring	0.0021	0.003	0.658	0.511	-0.004	0.008
season_Summer	0.0130	0.003	3.902	0.000	0.006	0.020
season_Winter	0.0079	0.003	2.482	0.013	0.002	0.014
dayofweek_Monday	0.0148	0.007	2.114	0.035	0.001	0.029
dayofweek_Saturday	0.0213	0.004	4.894	0.000	0.013	0.030
dayofweek_Sunday	0.0133	0.005	2.849	0.005	0.004	0.023
dayofweek_Thursday	0.0091	0.004	2.281	0.023	0.001	0.017
dayofweek_Tuesday	0.0065	0.008	0.796	0.426	-0.009	0.022
dayofweek_Wednesday	0.0180	0.003	6.490	0.000	0.013	0.023
major_sports_event_1	-0.0156	0.003	-5.887	0.000	-0.021	-0.010

Omnibus:	124.855	Durbin-Watson:	1.958
Prob(Omnibus):	0.000	Jarque-Bera (JB):	851.068
Skew:	-0.598	Prob(JB):	1.56e-185
Kurtosis:	8.268	Cond. No.	1.71e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.71e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 20 OLS-1 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.
   - The value for adj. R-squared is **0.925**, which is high.

2. ***const* coefficient**:
   - It is the Y-intercept.
   - If all the predictor variable coefficients are zero, then the expected output (i.e., Y) would be equal to the *const* coefficient.
   - The value for const coefficient is -**0.2630**
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).
   - The coefficient of visitors is **0.0899**
4. The condition number is large, 1.71e+06. This might indicate that there are strong multi-collinearity or other numerical problems.

II. Model Statistics

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.028585	0.019667	0.927703	0.925243	4.04905

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.036069	0.024198	0.88134	0.871452	4.920487

Figure 21 Model Statistics – 1

1. **Good Model Performance**: The model performs well on both training and test sets.
2. **Low Error & High R^2**: Strong predictive capability with low RMSE, MAE, and MAPE.
3. **Slight Overfitting**: Training R^2 (92.77%) is higher than Test R^2 (88.13%).

III. Model Coefficients

Model Coefficients:

	Feature	Coefficient
const	const	-0.262965
visitors	visitors	0.089928
ad_impressions	ad_impressions	-0.000050
views_trailer	views_trailer	0.003881
content_to_trailer_ratio	content_to_trailer_ratio	39.268667
ad_per_visitor	ad_per_visitor	0.000088
genre_Comedy	genre_Comedy	0.005504
genre_Drama	genre_Drama	0.010254
genre_Horror	genre_Horror	0.006999
genre_Others	genre_Others	0.004658
genre_Romance	genre_Romance	0.005894
genre_Sci-Fi	genre_Sci-Fi	0.003347
genre_Thriller	genre_Thriller	0.008935
season_Spring	season_Spring	0.002118
season_Summer	season_Summer	0.013012
season_Winter	season_Winter	0.007922
dayofweek_Monday	dayofweek_Monday	0.014785
dayofweek_Saturday	dayofweek_Saturday	0.021277
dayofweek_Sunday	dayofweek_Sunday	0.013330
dayofweek_Thursday	dayofweek_Thursday	0.009116
dayofweek_Tuesday	dayofweek_Tuesday	0.006451
dayofweek_Wednesday	dayofweek_Wednesday	0.018038
major_sports_event_1	major_sports_event_1	-0.015570

Figure 22 Model Coefficients

Assumptions of linear regression model

I. Test For Multicollinearity

1. Test for multicollinearity using VIF:
   - If VIF is 1 then there is no correlation between the kth predictor and the remaining predictor variables.
   - If VIF exceeds 5 or is close to exceeding 5, we say there is moderate multicollinearity.
   - If VIF is 10 or exceeding 10, it shows signs of high multicollinearity.

	feature	VIF
0	const	941.121103
1	visitors	16.176316
2	ad_impressions	31.801598
3	views_trailer	2.852850
4	content_to_trailer_ratio	3.675638
5	ad_per_visitor	45.349930
6	genre_Comedy	1.920192
7	genre_Drama	1.928064
8	genre_Horror	1.910773
9	genre_Others	2.574680
10	genre_Romance	1.763650
11	genre_Sci-Fi	1.869680
12	genre_Thriller	1.923344
13	season_Spring	1.602620
14	season_Summer	1.687297
15	season_Winter	1.617590
16	dayofweek_Monday	1.069681
17	dayofweek_Saturday	1.227417
18	dayofweek_Sunday	1.180626
19	dayofweek_Thursday	1.175723
20	dayofweek_Tuesday	1.066376
21	dayofweek_Wednesday	1.447462
22	major_sports_event_1	1.375180

VIF after dropping ['ad_per_visitor']

	feature	VIF
0	const	139.707281
1	visitors	1.282755
2	ad_impressions	1.029631
3	views_trailer	2.852799
4	content_to_trailer_ratio	3.675614
5	genre_Comedy	1.918813
6	genre_Drama	1.927154
7	genre_Horror	1.905249
8	genre_Others	2.574163
9	genre_Romance	1.754643
10	genre_Sci-Fi	1.869578
11	genre_Thriller	1.921003
12	season_Spring	1.595437
13	season_Summer	1.686186
14	season_Winter	1.617242
15	dayofweek_Monday	1.069602
16	dayofweek_Saturday	1.225798
17	dayofweek_Sunday	1.174514
18	dayofweek_Thursday	1.174097
19	dayofweek_Tuesday	1.066300
20	dayofweek_Wednesday	1.444886
21	major_sports_event_1	1.374932

Figure 23 VIF values

2. Ad_impressions, visitors and ad_per_visitor show high multicollinearity.
3. To remove multicollinearity:
   - Drop every column one by one that has a VIF score greater than 5.
   - Look at the adjusted R-squared and RMSE of all these models.
   - Drop the variable that makes the least change in adjusted R-squared.
   - Check the VIF scores again.
   - Continue till you get all VIF scores under 5.

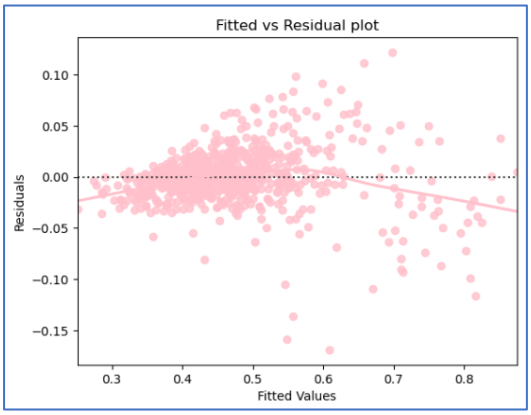
	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	ad_impressions	0.924871	0.029161
1	ad_per_visitor	0.924813	0.029172
2	visitors	0.922855	0.029549

Figure 24 Adj R and RMSE after dropping columns

4. Dropping ad_impressions and ad_per_visitor doesn't show much impact on adjusted r-squared value while dropping visitors show a significant impact. Therefore, we will drop, ad_per_visitor column.
5. We will drop the predictor variables having a p-value greater than 0.05 as they do not significantly impact the target variable.
   - 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.
6. Features selected : ['const', 'visitors', 'views_trailer', 'content_to_trailer_ratio', 'season_Summer', 'season_Winter', 'dayofweek_Monday', 'dayofweek_Saturday', 'dayofweek_Sunday', 'dayofweek_Thursday', 'dayofweek_Wednesday', 'major_sports_event_1']

II. Test for Linearity and Independence

1. Test for Linearity and independence:
   - Make a plot of fitted values vs residuals and checking for patterns.
   - If there is no pattern, then the model is linear and residuals are independent.
   - If the model is showing signs of non-linearity and residuals are not independent.



	Actual Values	Fitted Values	Residuals
731	0.40	0.418448	-0.018448
716	0.70	0.647475	0.052525
640	0.42	0.416235	0.003765
804	0.55	0.524635	0.025365
737	0.59	0.588346	0.001654

Figure 25 Fitted vs Residual values

2. The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
3. We see no pattern in the plot above. Hence, the assumptions of linearity and independence are satisfied.

III. Test For Normality

1. Test for Normality:
   - Check the Q-Q plot of residuals and by use the Shapiro-Wilk test.
   - If the residuals follow a normal distribution, they will make a straight line plot, otherwise not.
   - If the p-value of the Shapiro-Wilk test is greater than 0.05, we can say the residuals are normally distributed.

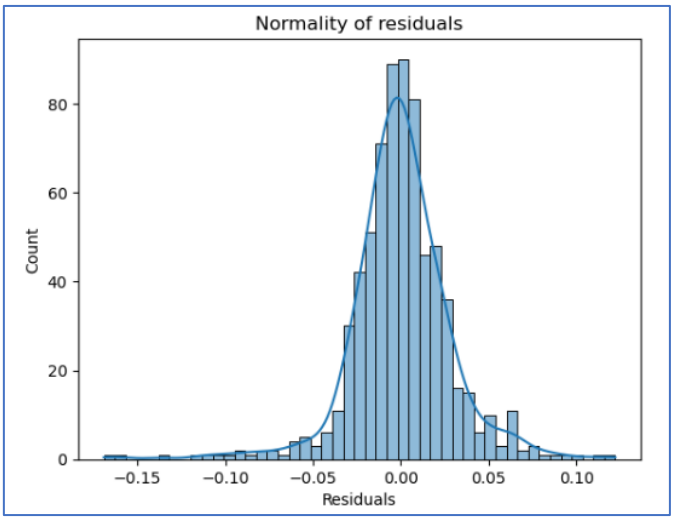


Figure 26 Normality of Residuals

2. The histogram of residuals almost has a bell shape structure.
3. The residuals more or less follow a straight line except for the tails.
4. Shapiro Result (statistic=0.9216209109466189, pvalue=1.3294500584587767e-18):
   - Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test.
   - However, as an approximation, we accept this distribution as close to being normal.
   - So, the assumption is satisfied.

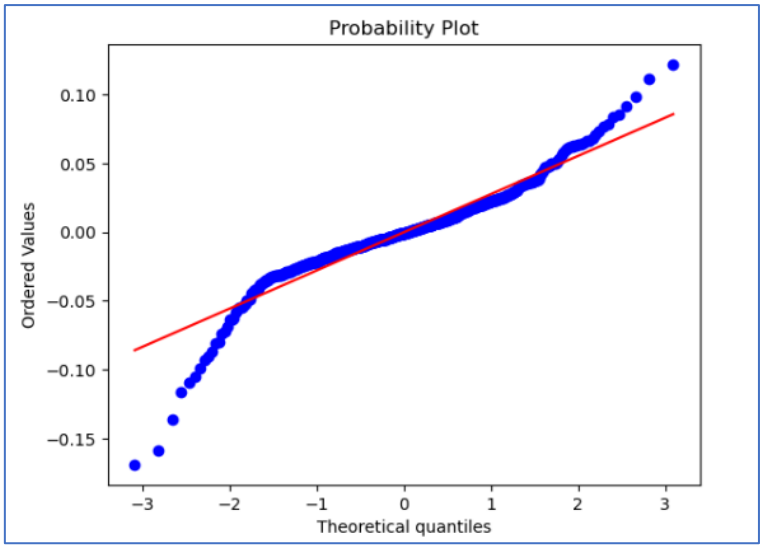


Figure 27 Probability Plot

IV. Test For Homoscedasticity

1. **Homoscedascity**: If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic.
2. **Heteroscedascity**: If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic
3. Test for homoscedasticity:
   - Use the goldfeldquandt test.
   - Null hypothesis: Residuals are homoscedastic
   - Alternate hypothesis: Residuals have heteroscedasticity
4. F statistic= 1.0822768095980726, p-value=0.2338634525925889:
   - Since p-value > 0.05, the residuals are homoscedastic.
   - So, the assumption is satisfied.

Model performance evaluation

I. Final Model Statistics

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.028859	0.019676	0.92631	0.925023	4.038724

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.036424	0.024655	0.878994	0.873935	5.039697

Figure 28 Model Statistics – 2

1. The model explains **92.63% (Train) and 87.90% (Test)** of the variance in content viewership, indicating a **good fit** but a slight drop in test performance.
2. **RMSE and MAE values** remain low, suggesting accurate predictions.
3. **MAPE** indicates that the model's percentage error is relatively small, making it reliable for decision-making.

II. Comparison between initial and final models

Training performance comparison:

	Linear Regression (initial)	Linear Regression (final)
RMSE	0.028585	0.028859
MAE	0.019667	0.019676
R-squared	0.927703	0.926310
Adj. R-squared	0.925243	0.925023
MAPE	4.049050	4.038724

Figure 29 Training performance comparison

1. **Training performance comparison:**
   - The final model shows a negligible change in training performance, with slightly higher RMSE and lower R-squared, indicating minimal improvement.
   - The MAPE slightly decreased, suggesting better percentage error handling.
   - Overall, the model remains stable, but further refinement may be needed to optimize generalization.

2. **Testing performance comparison:**
   - The final model shows a slight increase in RMSE and MAE, indicating marginally higher error on the test set.
   - R-squared dropped slightly, suggesting a minor decline in explanatory power.
   - The MAPE increased, indicating slightly worse percentage error. Overall, the model's performance remains similar.

Test performance comparison:

	Linear Regression (initial)	Linear Regression (final)
RMSE	0.036069	0.036424
MAE	0.024198	0.024655
R-squared	0.881340	0.878994
Adj. R-squared	0.871452	0.873935
MAPE	4.920487	5.039697

Figure 30 Test Performance comparison

III. Inferences from the Model

1. A unit increase in visitors increases the median content views by 0.0460 units, all other variables held constant.
2. A unit increase in views_trailer increases the median content views by 0.0039 units, all other variables held constant.
3. A unit increase in content_to_trailer_ratio increases the median content views by 39.4164 units, all other variables held constant.
4. A unit increase in the season being summer increases the median content views by 0.0113 units, all other variables held constant.
5. A unit increase in the season being Winter increases the median content views by 0.0072 units, all other variables held constant.
6. A unit increase in the day of the week being Monday increases the median content views by 0.0146 units, all other variables held constant.
7. A unit increase in the day of the week being Saturday increases the median content views by 0.0202 units, all other variables held constant.
8. A unit increase in the day of the week being Sunday increases the median content views by 0.0117 units, all other variables held constant.

9. A unit increase in the day of the week being Thursday increases the median content views by 0.0083 units, all other variables held constant.
10. A unit increase in the day of the week being Wednesday increases the median content views by 0.0173 units, all other variables held constant.
11. A house located in an area with a major sports event (major_sports_event_1) decreases the median content views by 0.0152 units, all other variables held constant.

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.926
Model:	OLS	Adj. R-squared:	0.925
Method:	Least Squares	F-statistic:	786.2
Date:	Sun, 02 Feb 2025	Prob (F-statistic):	0.00
Time:	19:01:10	Log-Likelihood:	1488.5
No. Observations:	700	AIC:	-2953.
Df Residuals:	688	BIC:	-2898.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1789	0.011	-16.030	0.000	-0.201	-0.157
visitors	0.0460	0.005	8.946	0.000	0.036	0.056
views_trailer	0.0039	5.34e-05	72.704	0.000	0.004	0.004
content_to_trailer_ratio	39.4164	1.079	36.531	0.000	37.298	41.535
season_Summer	0.0113	0.003	4.023	0.000	0.006	0.017
season_Winter	0.0072	0.003	2.650	0.008	0.002	0.012
dayofweek_Monday	0.0146	0.007	2.100	0.036	0.001	0.028
dayofweek_Saturday	0.0202	0.004	4.694	0.000	0.012	0.029
dayofweek_Sunday	0.0117	0.005	2.550	0.011	0.003	0.021
dayofweek_Thursday	0.0083	0.004	2.092	0.037	0.001	0.016
dayofweek_Wednesday	0.0173	0.003	6.349	0.000	0.012	0.023
major_sports_event_1	-0.0152	0.003	-5.816	0.000	-0.020	-0.010

Omnibus:	123.839	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	885.936
Skew:	-0.574	Prob(JB):	4.18e-193
Kurtosis:	8.390	Cond. No.	7.35e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.35e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 31 OLS-Final Regression Results

Actionable Insights & Recommendations

I. Significance of predictors

1. All independent variables have p-values < 0.05, meaning they significantly influence views_content.
2. The most impactful predictor is content_to_trailer_ratio (coef = 39.4164, p < 0.001), suggesting that a higher trailer-to-content ratio is strongly associated with increased content views.
3. views_trailer (coef = 0.0039, p < 0.001) also has a strong effect, showing that more trailer views contribute positively to content views.
4. major_sports_event_1 (coef = -0.0152, p < 0.001) negatively impacts views_content, suggesting that during major sports events, content views tend to decrease.
5. Seasons: Content views tend to increase during Summer (coef = 0.0113, p < 0.001) and Winter (coef = 0.0072, p = 0.008) compared to the reference season (likely Spring).
6. Days of the Week:
   - Higher views on Saturday (coef = 0.0202, p < 0.001) and Wednesday (coef = 0.0173, p < 0.001) suggest that these days have significantly higher engagement.
   - Monday, Sunday, and Thursday also show a slight positive effect.

II. Key takeaways for the business

1. **Visitors' Influence:** Increase the telephone number of visitant significantly increases content position. This means that place in high spirits visitor figure could now contribute to increase in views, making it crucial to focus on drive more traffic.
2. **Trailer Views**: More views on trailers are powerfully correlate with more than content views. Promoting trailers or puzzle content effectively could accept a significantly impingement on overall viewership.
3. **Content-to-Trailer Ratio:** A higher content-to-trailer ratio greatly impacts purview. Ensure that content is employ and well-aligned with laggard could lead to increase in content consumption.
4. **Seasonal Impact**: Both Summer and Winter seasons feature prescribed consequence on viewership suggesting that mental object pulmonary tuberculosis might top out during these periods. Preparation marketing movement and content departure during this meter could maximise views.
5. **Day of the Week:** Viewership tends to be higher on specific days, such as Saturday, Wednesday, and Monday. Programming mental object releases or promotional cause around these days could enhance engagement.
6. **Major Sports Events**: The comportment of a major sport event negatively touch content views, possibly because witness is distracted or occupied with the event. It's all-important to set content vent schedule around such upshot to optimise visibility and engagement.

III. Recommendations

1. **Increase Visitor Traffic:** Since the routine of visitant has a significant positive impact on sight, the business should clothe in strategies to drive more visitors to the political program. This could admit improving SEO, unravel targeted advert political campaign, leverage influencers, and optimize drug user experience to increase retention.
2. **Promote Trailers Effectively:** Impart that dawdler views have an unassailable correlation with content views, the business should centre on push prevue to a greater extent strategically. This could take well locating on the homepage, leveraging social media program for poke teasers, and optimizing drone content to beguile interest.
3. **Optimize Content-to-Trailer Ratio:** The depicted object-to-trailer ratio significantly affects viewership. The business should pore on ensuring that trailers match the content well in terms of calibre and thematic component. Additionally, the business organisation could experiment with message previews to exert audience involvement and engagement.
4. **Leverage Seasonal Trends:** With mellow viewership in the Summer and Winter seasons, the stage business should consider schedule major content releases or crusade during these periods to maximise engagement. Seasonal marketing strategy and theme contentedness can as well align with audience interests during these times.
5. **Focus on Weekdays with Higher Engagement:** Saturdays, Wednesdays, and Mondays appear to make higher viewership. The business should focus on releasing or encourage content more aggressively on these days. Analyse audience behaviour on these Clarence Day could help in further optimize depicted object release schedules.
6. **Set Strategy for Major Sports Events:** Since major sports consequence negatively impact viewership, the business should deliberate avoid cognitive content dismission or major promotional material during this issue. Instead, they could target audiences before or after these events to capitalize on a possible cutpurse in competition.
7. **Enhance User Segmentation**: The business should consider segment its user base far to tailor content and promotions found on factor such as seasonality, viewing drug abuse, and engagement form. This sectionalization could lead to more personalised and effective marketing strategies.