PREDICTIVE MODELING PROJECT BUSINESS REPORT

PGP-(PGPDS.O. MAR25.A)

BY PAPPOPPULA AKHILESH

TABLE OF CONTENTS

LIST OF FIGURES

1. PROBLEM STATEMENT:

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.

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

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

2. OBJECTIVE:

ShowTime is an OTT service provider and offers a wide variety of content (movies, web shows, etc.) for its users. They want to determine the driver variables for first-day content viewership so that they can take necessary measures to improve the viewership of the content on their platform.

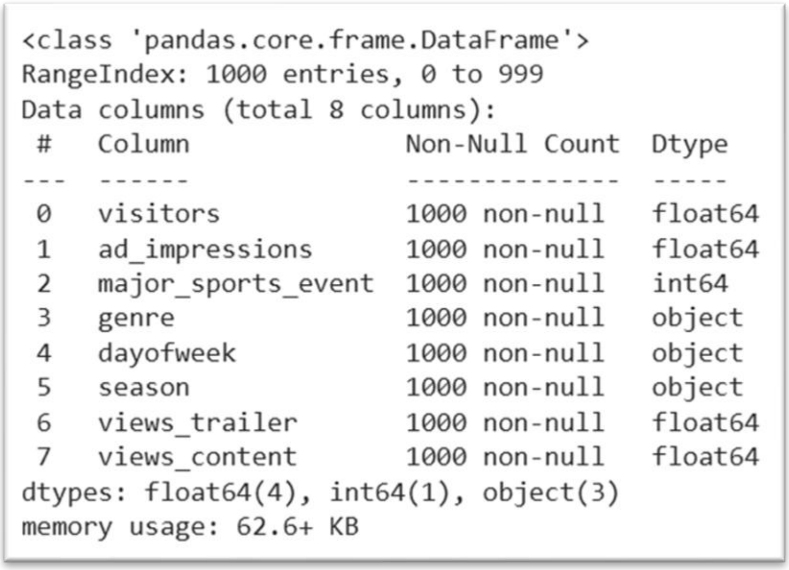
Some of the reasons for the decline in viewership of content would be the decline in the number of people coming to the platform, decreased marketing spend, content timing clashes, weekends and holidays, etc. They have hired you as a Data Scientist, shared the data of the current content on their platform, and asked you to analyze the data and come up with a linear regression model to determine the driving factors for first-day viewership.

3. DATA DESCRIPTION

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

- **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

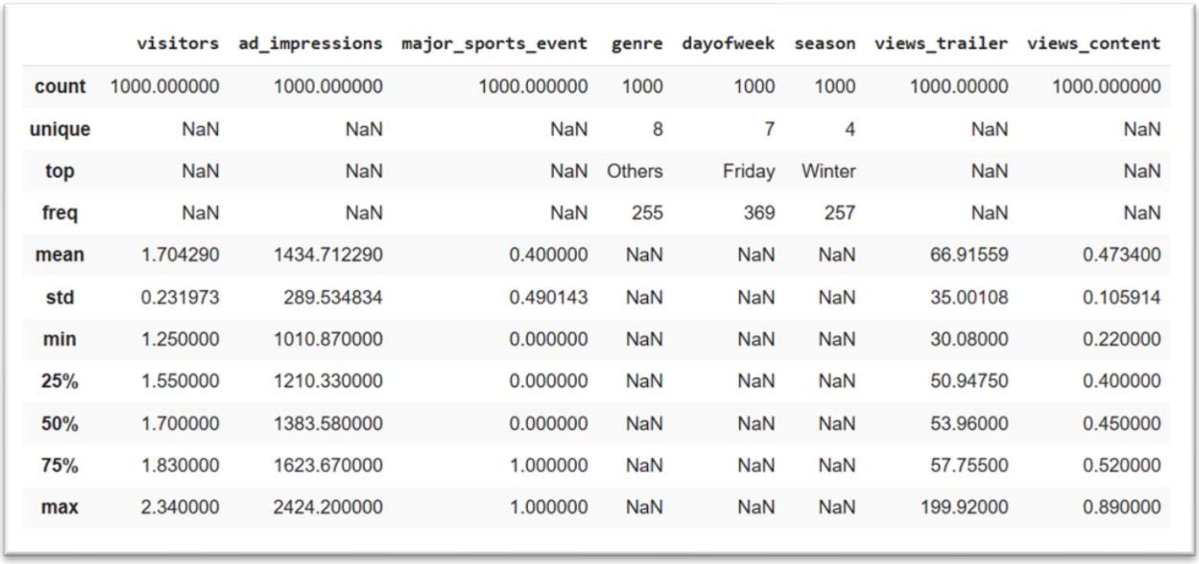
4. DATA OVERVIEW:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   visitors            1000 non-null   float64
 1   ad_impressions      1000 non-null   float64
 2   major_sports_event  1000 non-null   int64  
 3   genre               1000 non-null   object 
 4   dayofweek           1000 non-null   object 
 5   season              1000 non-null   object 
 6   views_trailer       1000 non-null   float64
 7   views_content       1000 non-null   float64
dtypes: float64(4), int64(1), object(3)
memory usage: 62.6+ KB
```

Figure 1 (Data types of the column for the dataset)

- The data set contains 1000 rows and 8 columns.
- There are 5 numeric (float and int type) and 3 string (object type) columns in the data
- The target variable is Views content, which is of float type.



	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content
count	1000.000000	1000.000000	1000.000000	1000	1000	1000	1000.00000	1000.000000
unique	NaN	NaN	NaN	8	7	4	NaN	NaN
top	NaN	NaN	NaN	Others	Friday	Winter	NaN	NaN
freq	NaN	NaN	NaN	255	369	257	NaN	NaN
mean	1.704290	1434.712290	0.400000	NaN	NaN	NaN	66.91559	0.473400
std	0.231973	289.534834	0.490143	NaN	NaN	NaN	35.00108	0.105914
min	1.250000	1010.870000	0.000000	NaN	NaN	NaN	30.08000	0.220000
25%	1.550000	1210.330000	0.000000	NaN	NaN	NaN	50.94750	0.400000
50%	1.700000	1383.580000	0.000000	NaN	NaN	NaN	53.96000	0.450000
75%	1.830000	1623.670000	1.000000	NaN	NaN	NaN	57.75500	0.520000
max	2.340000	2424.200000	1.000000	NaN	NaN	NaN	199.92000	0.890000

Figure 2 (Statistical summary of the data set)

- The average number of visitors is approximately 1.704 with a standard deviation of 0.232 in millions.
- The mean number of ad impressions is about 143.71, ranging from a minimum of 1010.87 to a maximum of 2424.20 in millions.
- There are 8 unique genres; the most frequent genre is “others” with 255 occurrences.
- There are 7 unique days and 4 unique seasons in the dataset.
- The average views for the trailer are approximately 66.91 million, with a standard deviation of 35.00 in millions.
- The minimum number of views is 30.08, and the maximum is 199.92 in millions.
- The average views for the content are approximately 0.47 with a standard deviation of 0.106 in millions.
- The minimum views are 0.22 and the maximum is 0.89 in millions.

5. EXPLORATORY DATA ANALYSIS:

5.1 UNIVARIATE ANALYSIS:

- Creating a Histogram and boxplot for all the numerical columns.

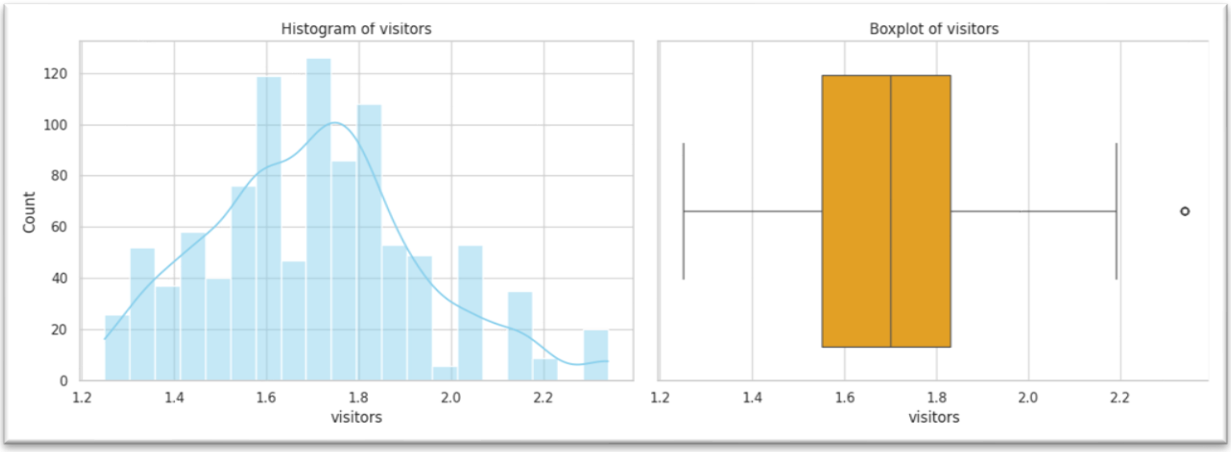


Figure 3 (Histogram and Boxplot on Visitors column)

Insights:

- The data suggests that most of the visitors fall within the range of 1.6 to 1.8 million, with a few outliers.
- The median visitor value is around 1.7 million.

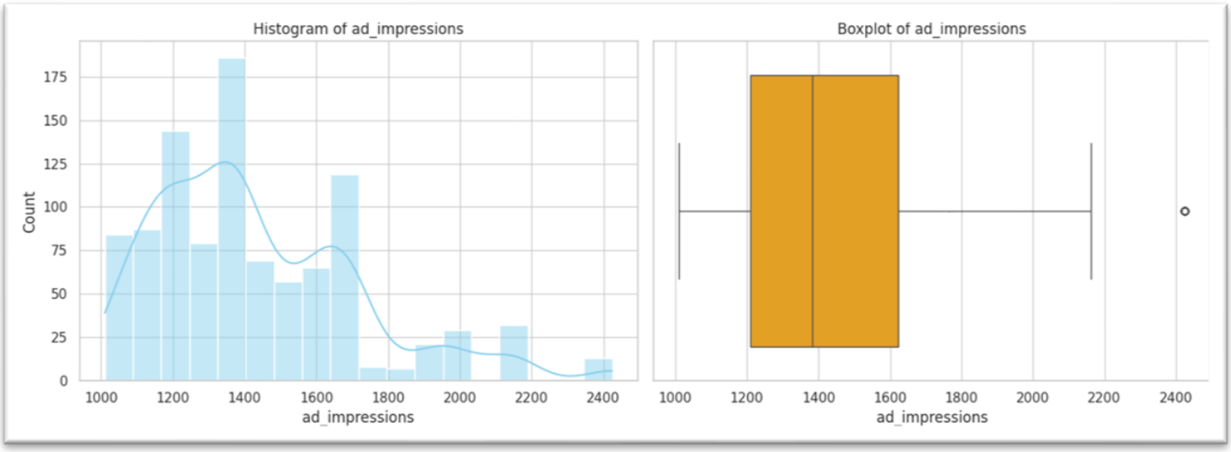


Figure 4(Histogram and Boxplot on ad_impressions column)

Insights

- The distribution appears to be right-skewed, peaking around 1200-1400 ad impressions with a couple of outliers.
- The median appears to be around 1200 million, which suggests that half of the ads receive more than 1600 million impressions and half receive less.

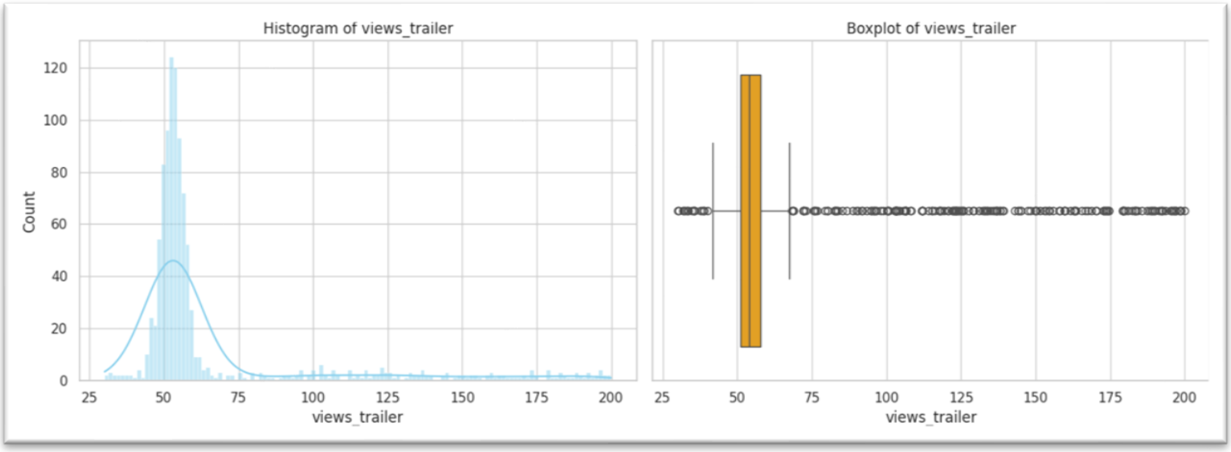


Figure 5 (Histogram and Boxplot on views_trailer column)

Insights:

- The distribution is highly skewed to the right, with many outliers on the higher end of the view range.
- This suggests that most of the trailers have significantly fewer views, with a small number of trailers having a higher number of views.

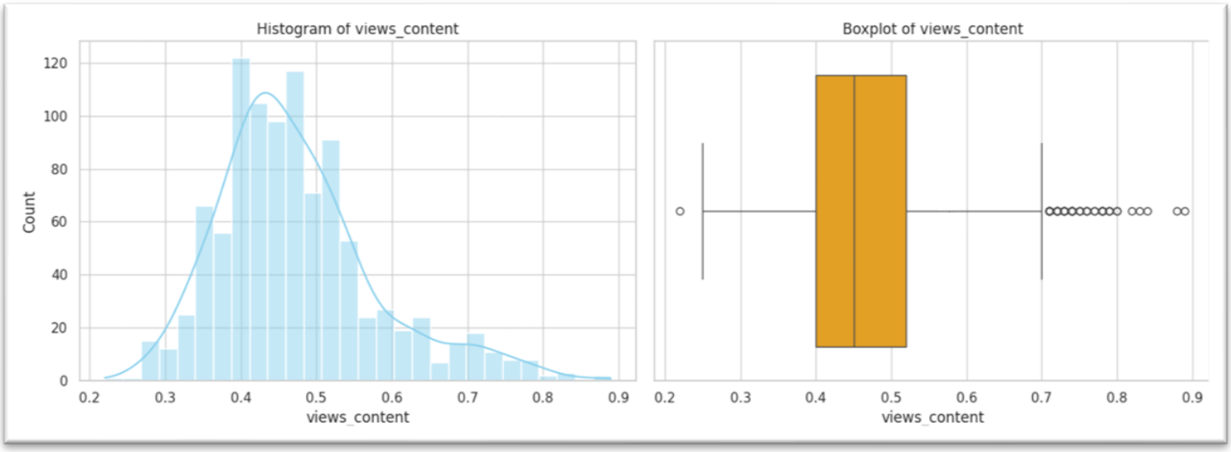


Figure 6 (Histogram and Boxplot on Views_content column)

Insights:

- The data is skewed towards the right side, indicating that most of the values are concentrated on the lower end.
- The peak of the distribution appears to be around 0.4 to 0.5
- The median value is around 0.45.
- There are several outliers at the higher end of the distribution.

- Now creating a Count plot on all categorical columns.

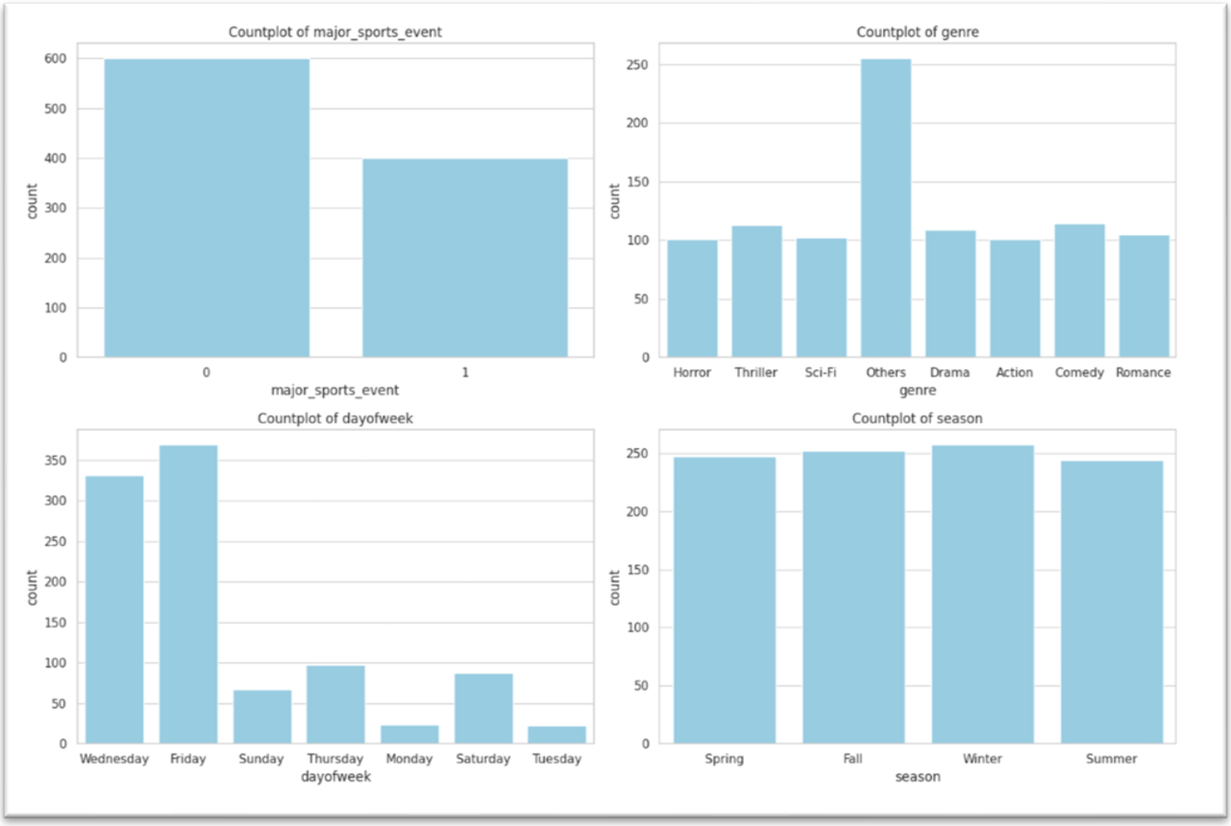


Figure 7 (Count plot on all categorical columns)

Insights:

1. **Major Sports Event:** This plot shows the count of any major sports events on the content release day, having a count of approximately 400.

2. **Day of the Week**: This plot displays the counts for different days of the week. Wednesday and Friday have a higher count of content release (around 300-350), while Monday, Tuesday have a very low count of content release.

3. **Genre:** The “others” category has the highest count, followed by “Comedy”, “Drama”, “Thriller”, and “Romance.” While “Sci-Fi,” “Horror,” and “Action” have lower counts.

4. **Season:** “Winter” and “Fall” have the highest count of the content release, around 250, followed by “Spring” and “Summer”, which have the least count of the content release.

5.2 BIVARIATE ANALYSIS:

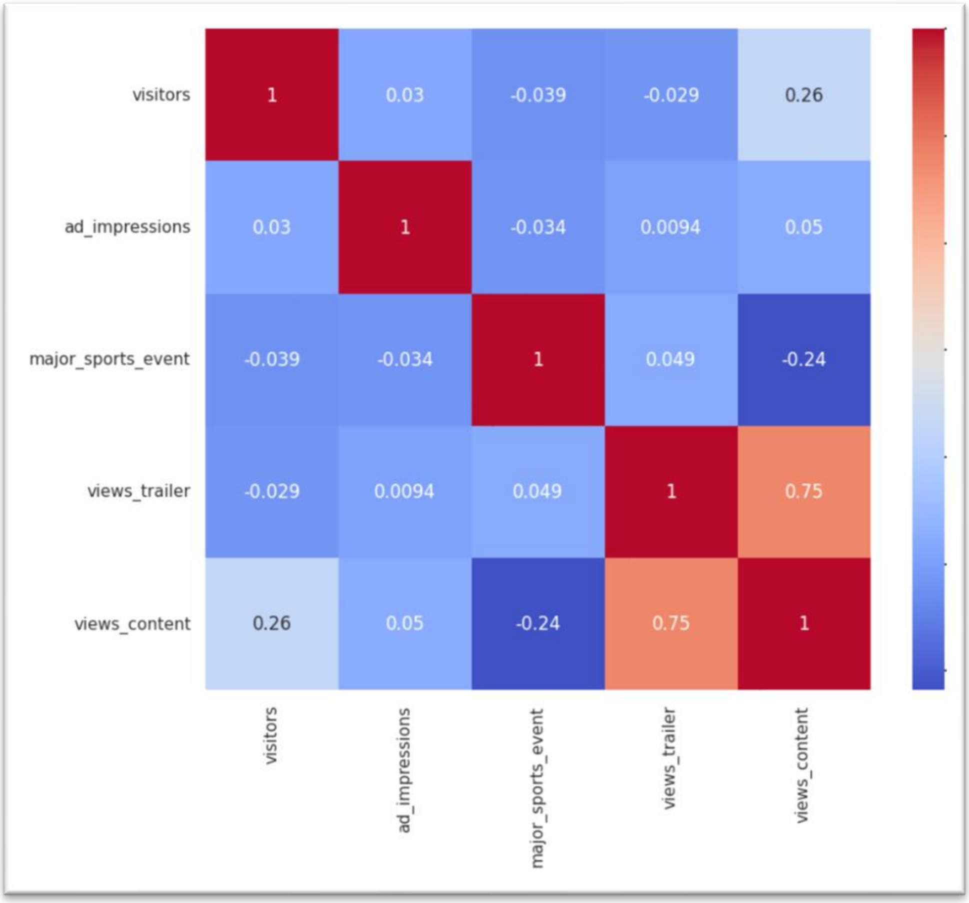


Figure 8 (Correlation Heatmap on all numerical columns)

<u>Insights:</u>

1. **Strong Positive Correlations:** There is a strong positive correlation (0.75) between "views_trailer" and "views_content." This suggests that when there are more trailer views, there are also more content views, and vice versa.

2. **Moderate Positive Correlations**: There is a moderate positive correlation (0.26) between "visitors" and "views_content." This indicates that more visitors tend to correlate with more content views.

3. **Weak/No Correlations:** The correlation between "visitors" and "ad_impressions" is very weak (0.03), meaning that there is almost no linear relationship between the number of visitors and the number of ad impressions.

4. **Negative Correlations:** There are negative correlations between "major_sports_event" and "views_content" (-0.24) and between "major_sports_event" and "visitors" (-0.039). This suggests that when there are more major sports events, there are fewer content views and visitors.

5.3 KEY QUESTIONS:

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

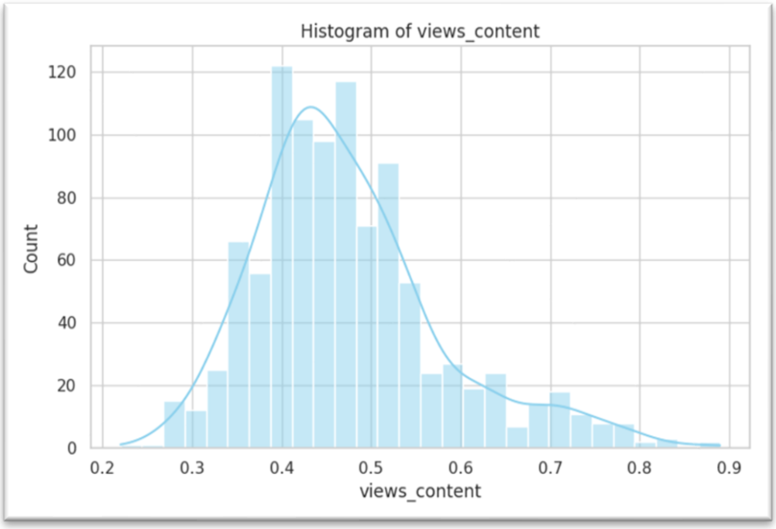


Figure 9 (Histogram of Views_content)

- The content views are **right-skewed**, meaning most content receives relatively lower viewership, while a few pieces of content receive significantly higher views.

2. **What does the distribution of genres look like?**

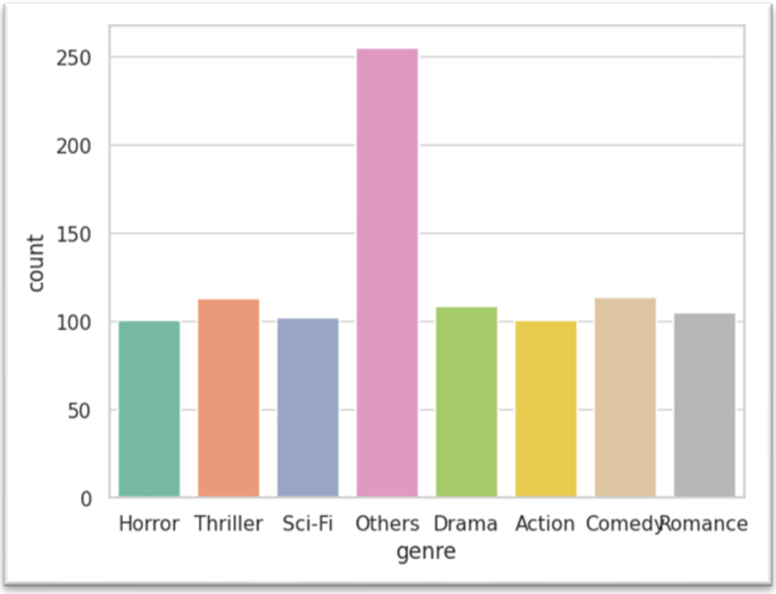


Figure 10 (Count plot of Genre)

- The most frequent genres in the dataset include **Thriller, Drama, and Comedy.**
- **Genres like Action, Sci-fi, and Comedy** appear less compared to other genres.

3. How does the viewership vary with the day of release?

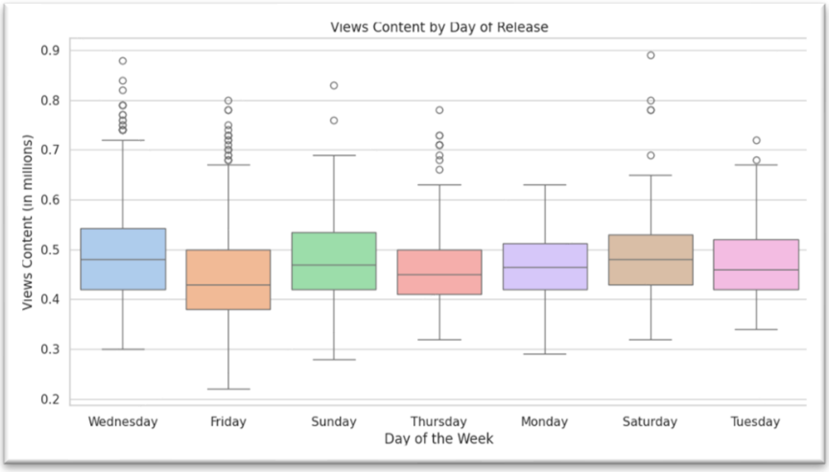


Figure 11 (Boxplot on views of content by the day of release)

- The **median (middle line)** for Wednesday is the **highest** among all days.
- This suggests that content released on **Wednesdays generally performs better** in terms of average viewership.
- **Friday** has many outliers, many exceeding 0.7 million views, Although the median viewership is moderate, Friday presents a high potential for standout performance.
- In contrast**, Saturday and Sunday** have narrower interquartile ranges, suggesting more consistent and stable viewership, even if not the highest."

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

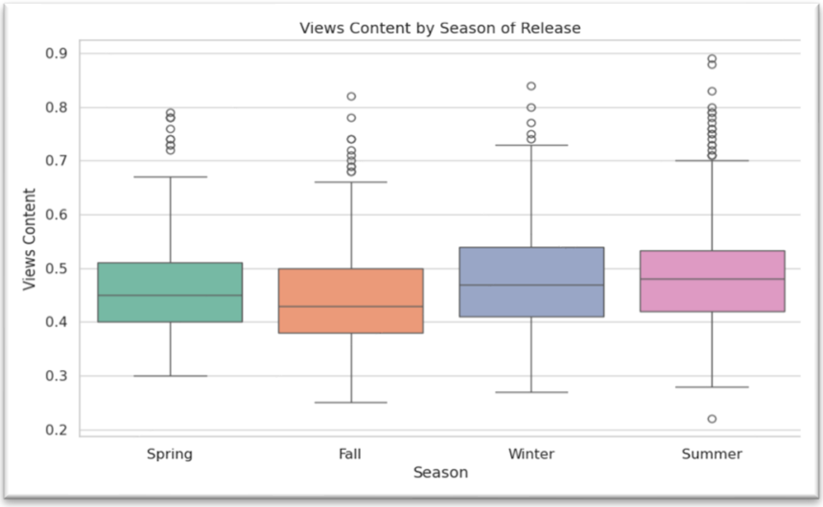


Figure 12 (Views of the content by season)

- The median view count appears to be relatively similar across all seasons, with summer having a slightly higher median than the others.
- All seasons show outliers, with summer having significantly more outliers than the other seasons.

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

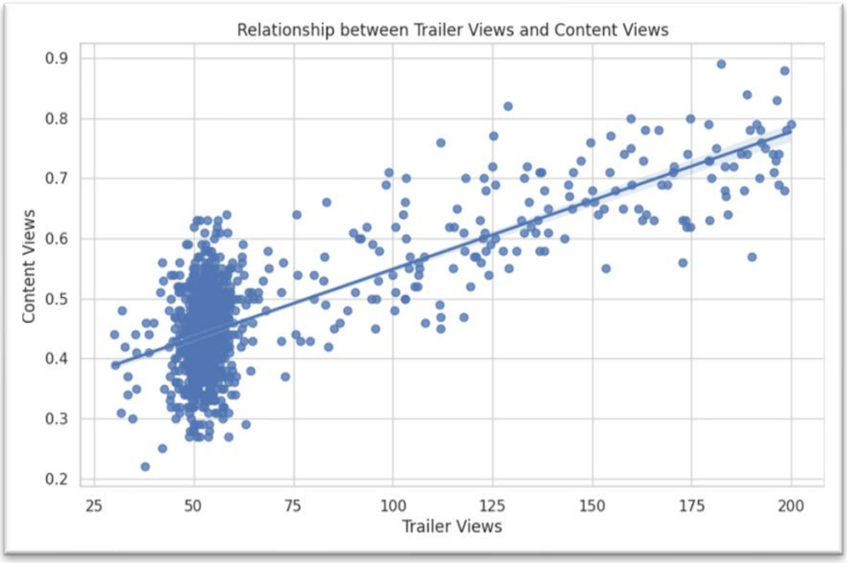


Figure 13 (Relationship between Trailer views and Content views)

- There is a **strong positive correlation** between trailer views and content views.
- This indicates that **higher trailer engagement is likely to lead to higher actual content viewership**, highlighting the importance of promotional efforts.

6. DATA PREPROCESSING:

(Checking for missing values, Duplicate entries & outlier detection, and treatment)

- The data set contains no duplicate entries and has no missing values.
- There are quite a few outliers in the dataset.
- However, we will not treat them as they are genuine values

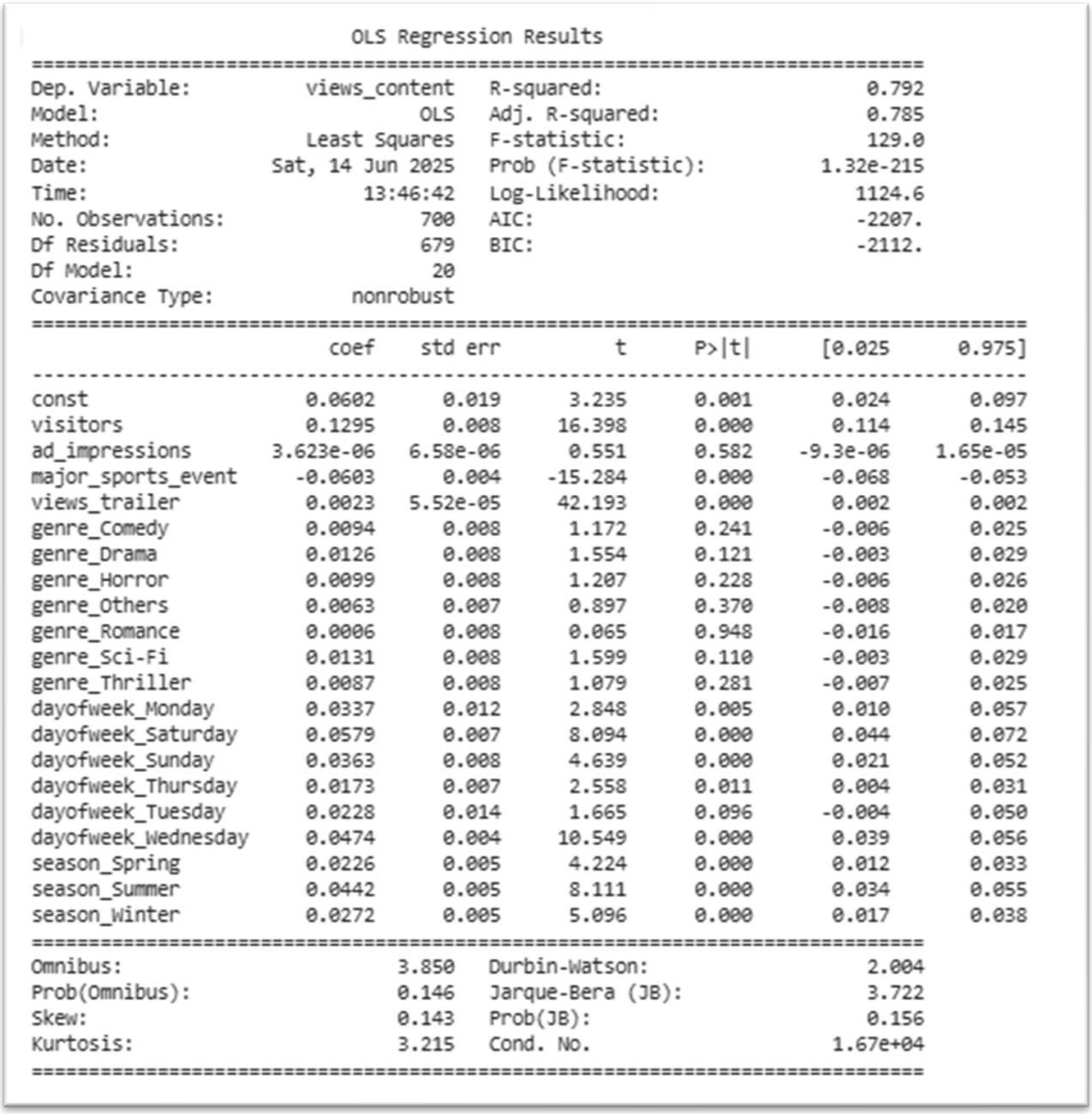
7. DATA PREPARATION FOR MODELLING:

We need to identify the key factors that influence first-day viewership.

- We have defined our independent (X) and dependent (y) variables.
- The X variable consists of data that excludes the 'views_content' column.
- The target variable, y, contains only the 'views_content' column.
- We have encoded the categorical features prior to building the model.
- All columns have been converted to the float64 data type for modeling.
- We split the data into training and testing sets in a 70:30 ratio to evaluate the model built on the training data.

- Number of rows in train data = 700
- Number of rows in test data = 300
- Now we will build a Linear Regression model using the training data and then check its performance

8. MODEL BUILDING – LINEAR REGRESSION:



```
                            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:                Sat, 14 Jun 2025   Prob (F-statistic):          1.32e-215
Time:                        13:46:42   Log-Likelihood:                 1124.6
No. Observations:                 700   AIC:                            -2207.
Df Residuals:                     679   BIC:                            -2112.
Df Model:                          20
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.0602      0.019      3.235      0.001       0.024       0.097
visitors                0.1295      0.008     16.398      0.000       0.114       0.145
ad_impressions       3.623e-06   6.58e-06      0.551      0.582    -9.3e-06    1.65e-05
major_sports_event     -0.0603      0.004    -15.284      0.000      -0.068      -0.053
views_trailer           0.0023   5.52e-05     42.193      0.000       0.002       0.002
genre_Comedy            0.0094      0.008      1.172      0.241      -0.006       0.025
genre_Drama             0.0126      0.008      1.554      0.121      -0.003       0.029
genre_Horror            0.0099      0.008      1.207      0.228      -0.006       0.026
genre_Others            0.0063      0.007      0.897      0.370      -0.008       0.020
genre_Romance           0.0006      0.008      0.065      0.948      -0.016       0.017
genre_Sci-Fi            0.0131      0.008      1.599      0.110      -0.003       0.029
genre_Thriller          0.0087      0.008      1.079      0.281      -0.007       0.025
dayofweek_Monday        0.0337      0.012      2.848      0.005       0.010       0.057
dayofweek_Saturday      0.0579      0.007      8.094      0.000       0.044       0.072
dayofweek_Sunday        0.0363      0.008      4.639      0.000       0.021       0.052
dayofweek_Thursday      0.0173      0.007      2.558      0.011       0.004       0.031
dayofweek_Tuesday       0.0228      0.014      1.665      0.096      -0.004       0.050
dayofweek_Wednesday     0.0474      0.004     10.549      0.000       0.039       0.056
season_Spring           0.0226      0.005      4.224      0.000       0.012       0.033
season_Summer           0.0442      0.005      8.111      0.000       0.034       0.055
season_Winter           0.0272      0.005      5.096      0.000       0.017       0.038
==============================================================================
Omnibus:                        3.850   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.146   Jarque-Bera (JB):                3.722
Skew:                           0.143   Prob(JB):                        0.156
Kurtosis:                       3.215   Cond. No.                     1.67e+04
==============================================================================
```

Figure 14 (Regression Results)

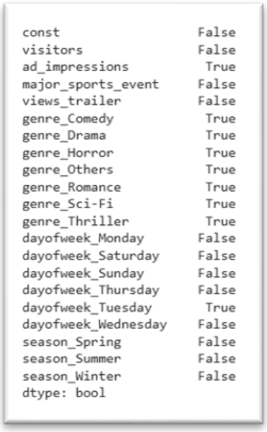
OLS Regression results:

- **R-squared = 0.792, Adj. R-squared = 0.785** → The model explains about 79% of the variance in views_content. This is a strong fit.
- The constant coefficient is **0.0602,** which means that if visitors, views_trailer, day-of-week dummies, genre dummies, ad impressions, season dummies, and major_sports_event are all zero → the baseline predicted views_content is **0.0602**.
- There are some variables which are having p-value> 0.05 which need to be dropped.
- The ad_impression does not have any statistically significant relationship with content views.
- Each additional visitor is associated with 0.1295 million views.
- Major sports events reduce views by ~0.06 million views
- More trailer views strongly drive content views.
- Sunday is the strongest weekend day for views.

Summary:

- **Strong positive drivers:** visitors, trailer views, weekends, Winter.
- **Negative impact**: major sports events.
- **No effect:** ad impressions.
- **Seasonality** and **weekends** matter significantly.

DEALING WITH HIGH P-VALUE VARIABLES:



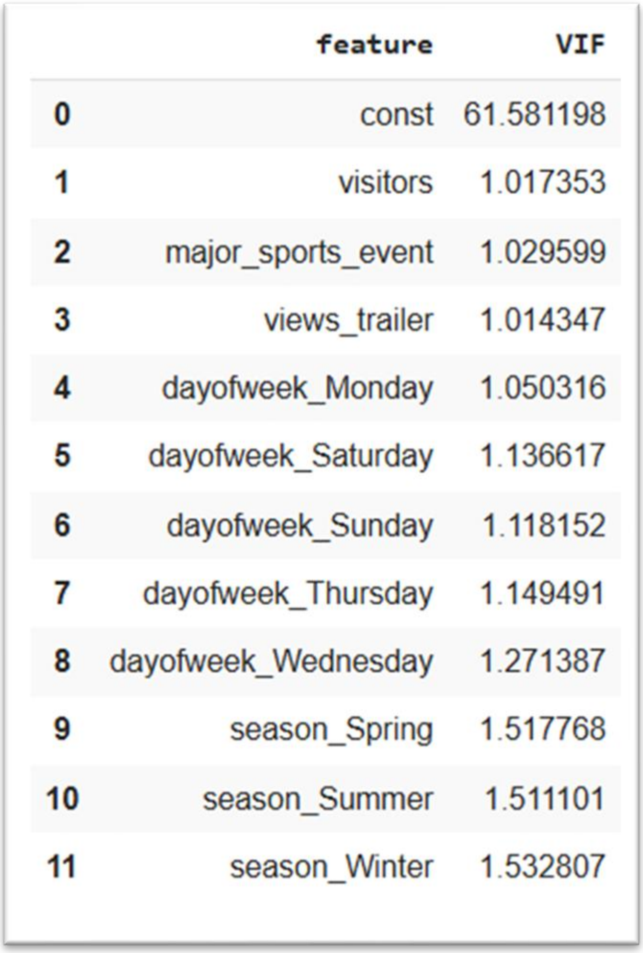
```
const                 False
visitors              False
ad_impressions         True
major_sports_event    False
views_trailer         False
genre_Comedy           True
genre_Drama            True
genre_Horror           True
genre_Others           True
genre_Romance          True
genre_Sci-Fi           True
genre_Thriller         True
dayofweek_Monday      False
dayofweek_Saturday    False
dayofweek_Sunday      False
dayofweek_Thursday    False
dayofweek_Tuesday      True
dayofweek_Wednesday   False
season_Spring         False
season_Summer         False
season_Winter         False
dtype: bool
```

Figure 15 (List of variables with high P-values)

- We have dropped the variables which ever are having values greater than 0.05 (=True).
- Now, no feature has a p-value greater than 0.05, so we'll consider the features in x_trainnew and x_testnew as the final set of predictor variables and olsmod2 as the final model to move forward with.

9. TESTING THE ASSUMPTIONS OF THE LINEAR REGRESSION MODEL:

9.1 TEST FOR MULTICOLLINEARITY:



	feature	VIF
0	const	61.581198
1	visitors	1.017353
2	major_sports_event	1.029599
3	views_trailer	1.014347
4	dayofweek_Monday	1.050316
5	dayofweek_Saturday	1.136617
6	dayofweek_Sunday	1.118152
7	dayofweek_Thursday	1.149491
8	dayofweek_Wednesday	1.271387
9	season_Spring	1.517768
10	season_Summer	1.511101
11	season_Winter	1.532807

Figure 16 (VIF Table)

- All the predictor variables have a VIF value less than 5.
- We will ignore the VIF values for the constant (intercept) and dummy variables.
- So, all the predictor variables do not have any multicollinearity.
- Assumption is satisfied.

9.2 TEST FOR LINEARITY AND INDEPENDENCE:

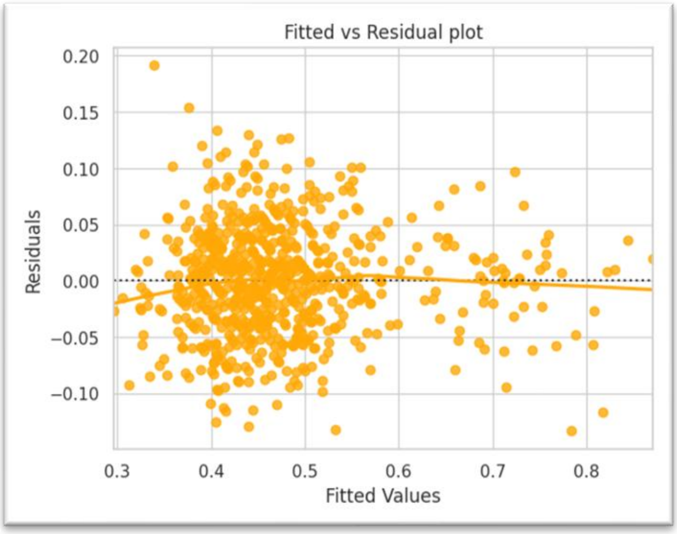
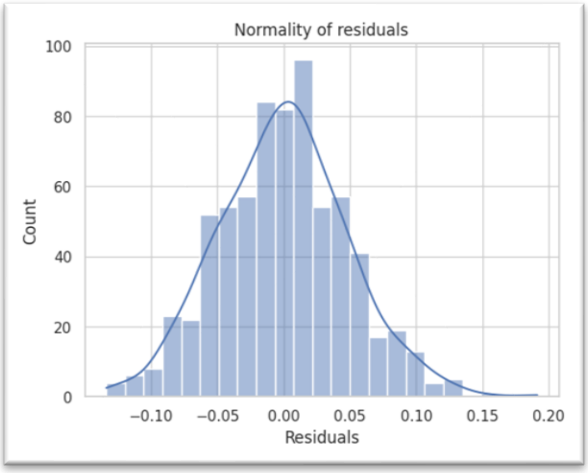


Figure 17 (Fitted VS Residual plot)

- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- In this plot, the residuals appear to be randomly distributed around zero, suggesting a reasonable fit. So the assumption is satisfied,

9.3 TEST FOR NORMALITY



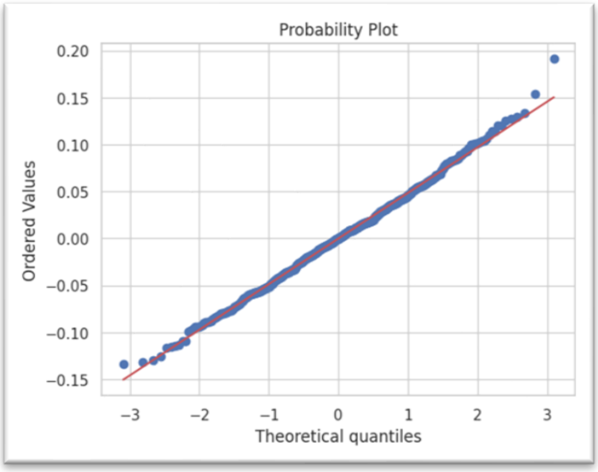


Figure 18 (Histogram of Residuals)

Figure 19 (Q-Q plot)

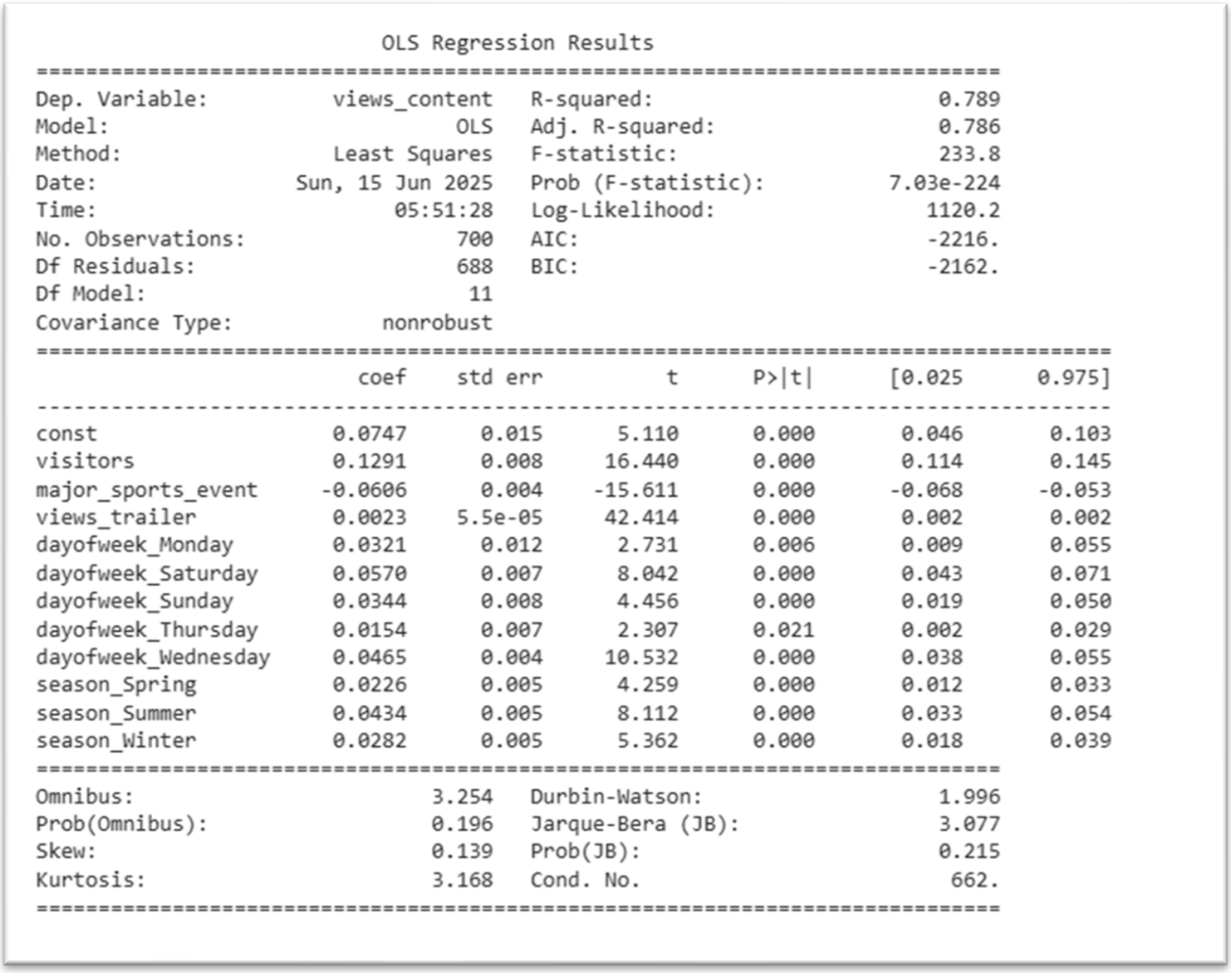
- The histogram of residuals has having bell-shaped curve.
- In the Q-Q plot, the residuals follow a straight line.
- The Shapiro Test p-value is also less than 0.05
- So, the assumption is satisfied.

9.4 TEST FOR HOMOSCEDASTICITY:

- Our p-value = (0.11134047820446640)
- Since **p-value > 0.05,** we can say that the residuals are homoscedastic. So, this assumption is satisfied.

Since **all the assumptions were satisfied,** we will build the final model and gain insights.

10. FINAL MODEL:



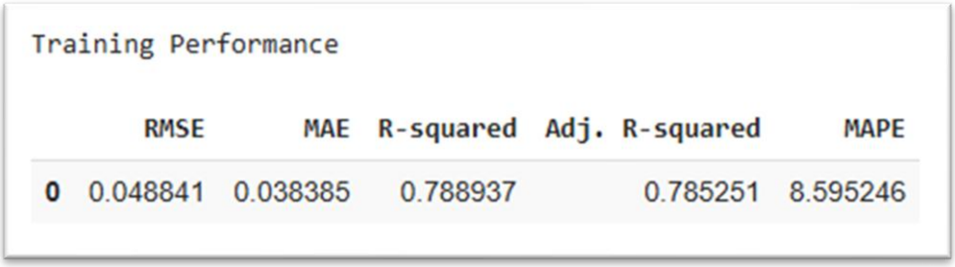
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.789
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     233.8
Date:                Sun, 15 Jun 2025   Prob (F-statistic):          7.03e-224
Time:                        05:51:28   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.
==============================================================================
```

Figure 20 (Final regression model results)

- **R^2 = 0.789, Adj. R^2 = 0.786**. The model explains ~79% of the variance in content views.

10.1 MODEL PERFORMANCE EVALUATION:

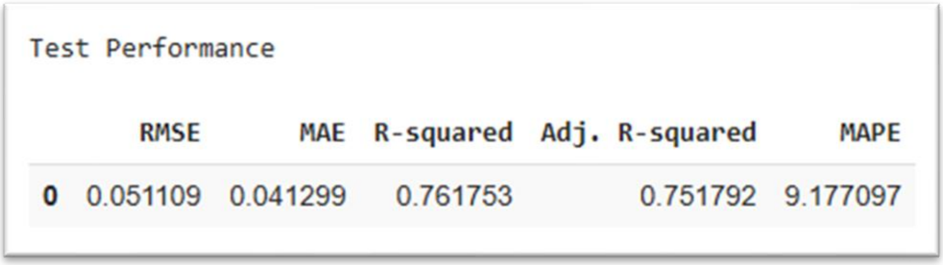
Figure 21 (Training Performance)



Training Performance

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

Figure 22 (Test Performance)



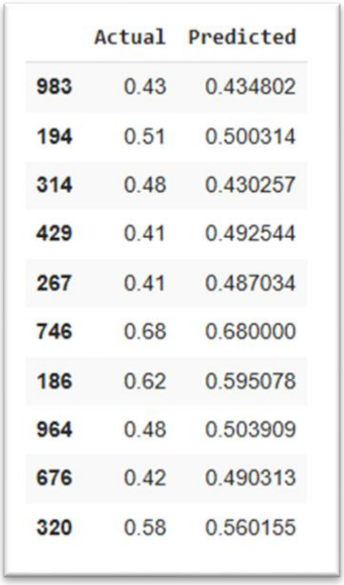
Test Performance

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

10.2 INSIGHTS:

1. The model performs similarly on the training and test sets.
2. **Training R²:** 78.9%
3. **Test R²:** 76.2%
4. 79% variance explained on training, and 76% on test, reliable predictions.
5. **Training RMSE:** 0.0488; **Test RMSE:** 0.0511
6. **Training MAE:** 0.0384; **Test MAE:** 0.0413
7. Both RMSE and MAE are low on training and test data, indicating good predictive accuracy.

10.3 PREDICTIONS ON TEST SET:



	Actual	Predicted
983	0.43	0.434802
194	0.51	0.500314
314	0.48	0.430257
429	0.41	0.492544
267	0.41	0.487034
746	0.68	0.680000
186	0.62	0.595078
964	0.48	0.503909
676	0.42	0.490313
320	0.58	0.560155

Figure 23 (Actual vs Predicted values on Test set)

- We can observe here that our model has returned pretty good prediction results, and the actual and predicted values are comparable

11. KEY TAKEAWAYS FOR THE BUSINESS:

1. **Positive drivers:** visitors, trailers, weekends, summer season.
2. **Negative driver**: major sports events.
3. **Visitors are the primary driver**
4. More visitors directly lead to more first-day views.
5. Each additional visitor increases views by ~0.13m ($p < 0.001$).
6. **Trailer views significantly boost first-day content views**
7. Each trailer view adds ~0.0023m to content views ($p < 0.001$).
8. **When a major sports event happens, views drop** by ~0.061 ($p < 0.001$),
9. **The day of the week and the season have a significant impact on viewership on the first day.**
10. Saturdays (+0.057) and Sundays (+0.034) → significant weekend boost.
11. Wednesdays (+0.047) → surprisingly strong midweek uplift.
12. Mondays (+0.032) and Thursdays (+0.015) → modest positive effects.
13. Summer (+0.043) → largest seasonal boost to views.
14. Winter (+0.028) and Spring (+0.023) → positive, but smaller compared to Summer.

12. RECOMMENDATIONS:

1. **Prioritize pre-launch trailer campaigns:** Focus on building trailer view counts, especially in the final days before release.

2. **Align release dates with high-viewership days:** Target Saturdays, Sundays, and Wednesdays for first-day launches where feasible.

3. **Avoid major sports event dates:** Integrate sports calendars into release planning to minimize viewership cannibalization.

4. **Leverage seasonal peaks:** Plan major releases in Summer, and second-tier releases in Winter/Spring.

5. **Invest in visitor acquisition:** Since visitors are a core driver, campaigns that increase site/app visitors (ads, influencer collabs, partnerships) will directly enhance first-day views.