ShowTime

(Business Report)

Report By:

Naman Srivastava

10 Aug, 2025

TABLE OF CONTENT

List of Figures

CONTEXT

An over-the-top (OTT) media service is a media service offered directly to viewers via the internet. The term is most synonymous with subscription-based video-on-demand services that offer access to film and television content, including existing series acquired from other producers, as well as original content produced specifically for the service. They are typically accessed via websites on personal computers, apps on smartphones and tablets, or televisions with integrated Smart TV platforms.

Presently, OTT services are at a relatively nascent stage and are widely accepted as a trending technology across the globe. With the increasing change in customers' social 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 increasingly attract subscribers on a concurrent basis.

OBJECTIVE

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

DATA DESCRIPTION

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

Variable	*Description*
visitor	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

Questions

The following questions need to be answered as part of the EDA section of the project:

1. What does the distribution of content views look like?
2. What does the distribution of genres look like?
3. The day of the week on which content is released generally plays a key role in the viewership. How does the viewership vary with the day of release?
4. How does the viewership vary with the season of release?
5. What is the correlation between trailer views and content views?

Kindly ensure to thoroughly examine all variables of the data in EDA, going beyond the above questions.

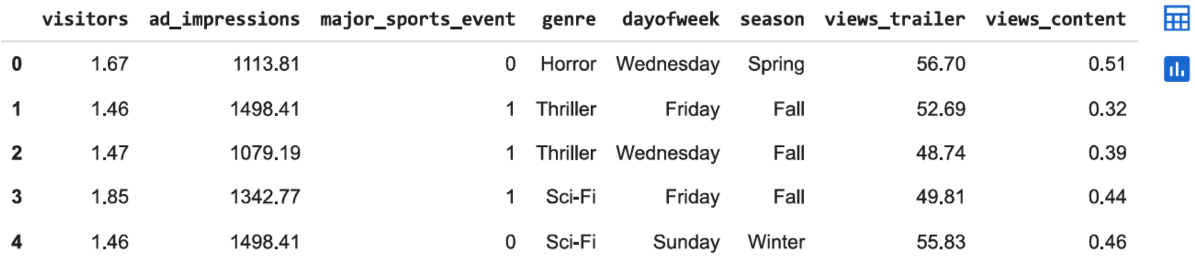
Exploratory Data Analysis

Exploratory Data Analysis was done on the provided data set (ottdata.csv) using Python tools on google colab. The objective of entire process was done to make data more understandable, reliable for meaningful decision making.

Data Import and Cleaning

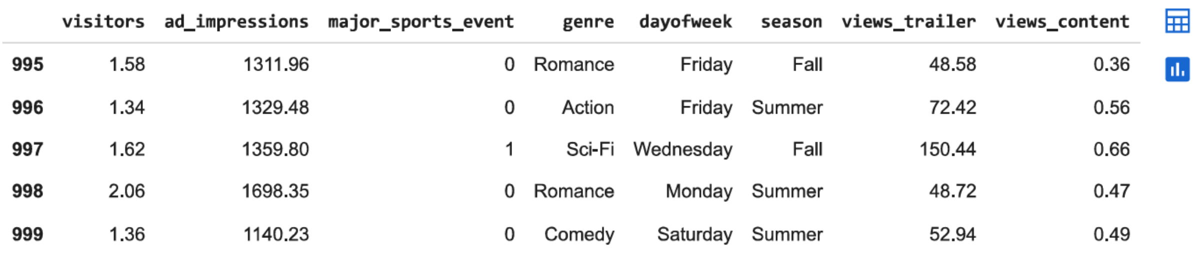
After successfully loading the data on the google colab notebook, and importing all the required libraries, we found out that on initial checking, that the data consists of **1000 entries, and 8 features**.

- Loading head and tail of the data



	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

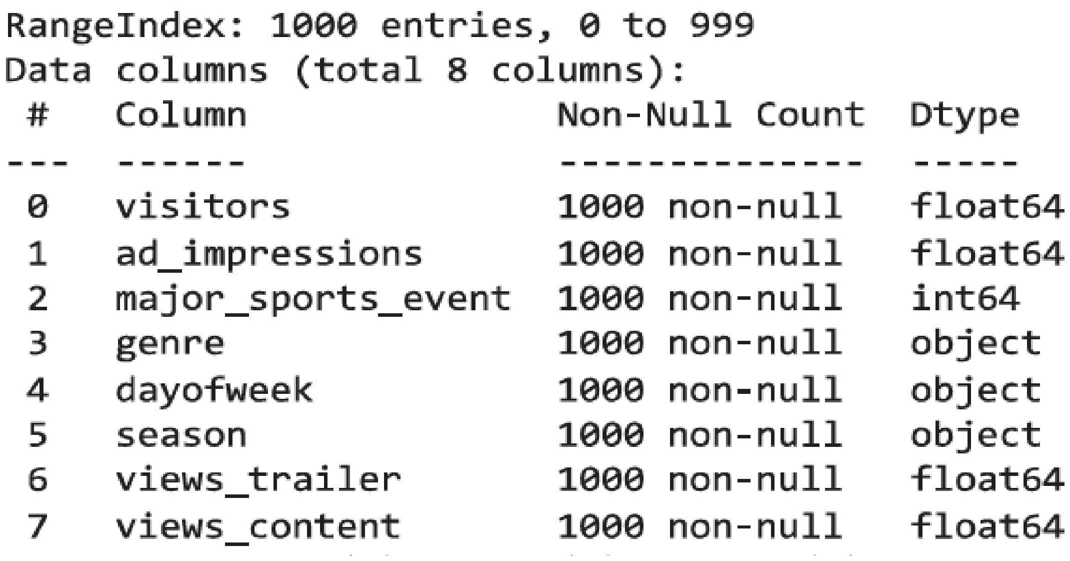
fig 1: Head



	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content
995	1.58	1311.96	0	Romance	Friday	Fall	48.58	0.36
996	1.34	1329.48	0	Action	Friday	Summer	72.42	0.56
997	1.62	1359.80	1	Sci-Fi	Wednesday	Fall	150.44	0.66
998	2.06	1698.35	0	Romance	Monday	Summer	48.72	0.47
999	1.36	1140.23	0	Comedy	Saturday	Summer	52.94	0.49

fig 2: Tail

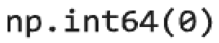
- Getting info of the data



```
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
```

fig 3: info

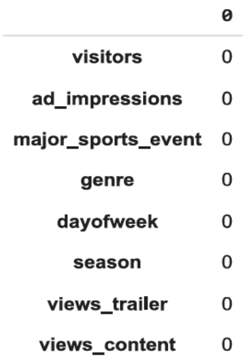
- Checking for duplicated values



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

fig 4: Duplicated values

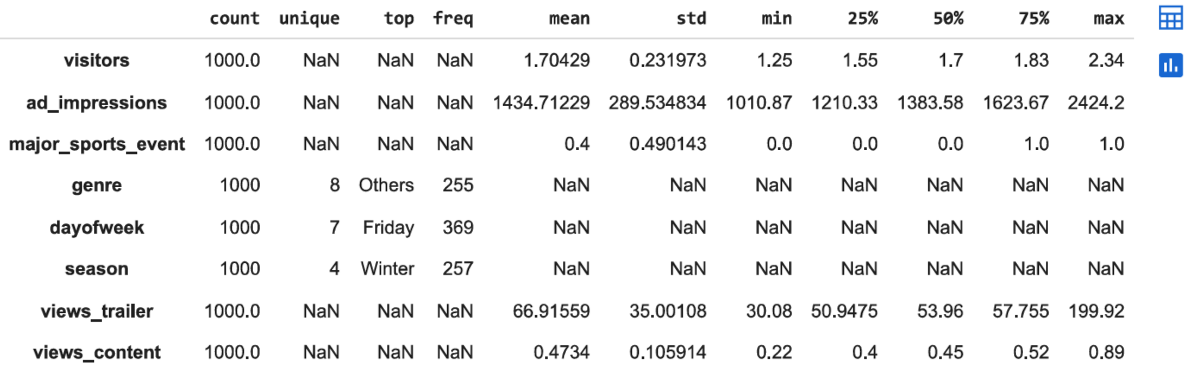
- Investigating null values



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

fig 5: Null values

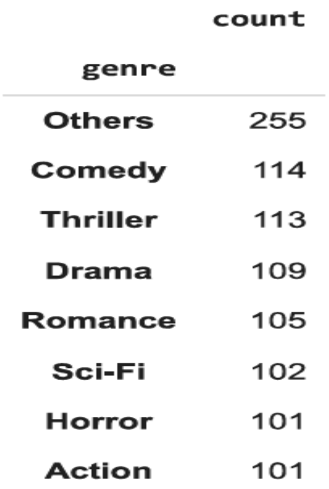
- Getting description of the data



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
visitors	1000.0	NaN	NaN	NaN	1.70429	0.231973	1.25	1.55	1.7	1.83	2.34
ad_impressions	1000.0	NaN	NaN	NaN	1434.71229	289.534834	1010.87	1210.33	1383.58	1623.67	2424.2
major_sports_event	1000.0	NaN	NaN	NaN	0.4	0.490143	0.0	0.0	0.0	1.0	1.0
genre	1000	8	Others	255	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dayofweek	1000	7	Friday	369	NaN	NaN	NaN	NaN	NaN	NaN	NaN
season	1000	4	Winter	257	NaN	NaN	NaN	NaN	NaN	NaN	NaN
views_trailer	1000.0	NaN	NaN	NaN	66.91559	35.00108	30.08	50.9475	53.96	57.755	199.92
views_content	1000.0	NaN	NaN	NaN	0.4734	0.105914	0.22	0.4	0.45	0.52	0.89

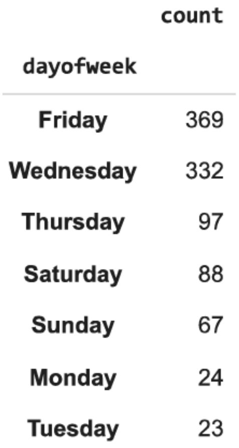
fig 6: Describe

- Checking the frequency in the different categorical data



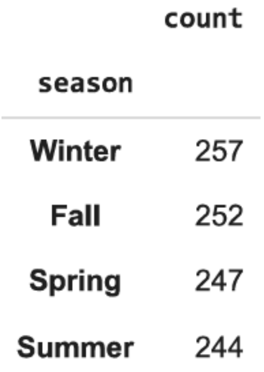
	count
genre	
Others	255
Comedy	114
Thriller	113
Drama	109
Romance	105
Sci-Fi	102
Horror	101
Action	101

fig 7: genre value count



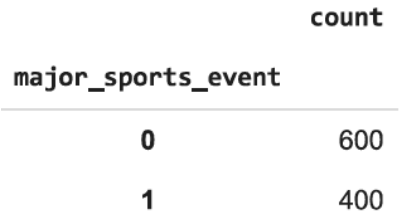
	count
dayofweek	
Friday	369
Wednesday	332
Thursday	97
Saturday	88
Sunday	67
Monday	24
Tuesday	23

fig 8: day of week value count



	count
season	
Winter	257
Fall	252
Spring	247
Summer	244

fig 9: season value count



	count
major_sports_event	
0	600
1	400

fig 10: major sports event value count

Now we have sufficient idea about our data and we can proceed towards further analysis.

Univariate Analysis

In univariate analysis we analyse single variable individually.

- Visitors:

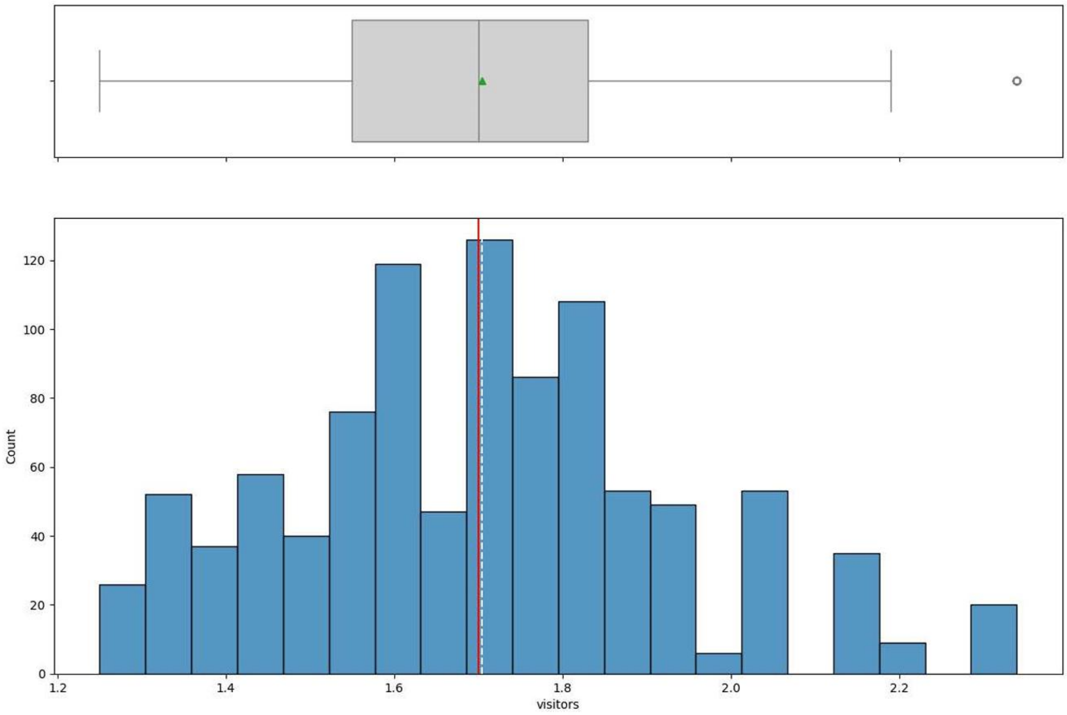


fig 11: Visitors

- Ad impressions:

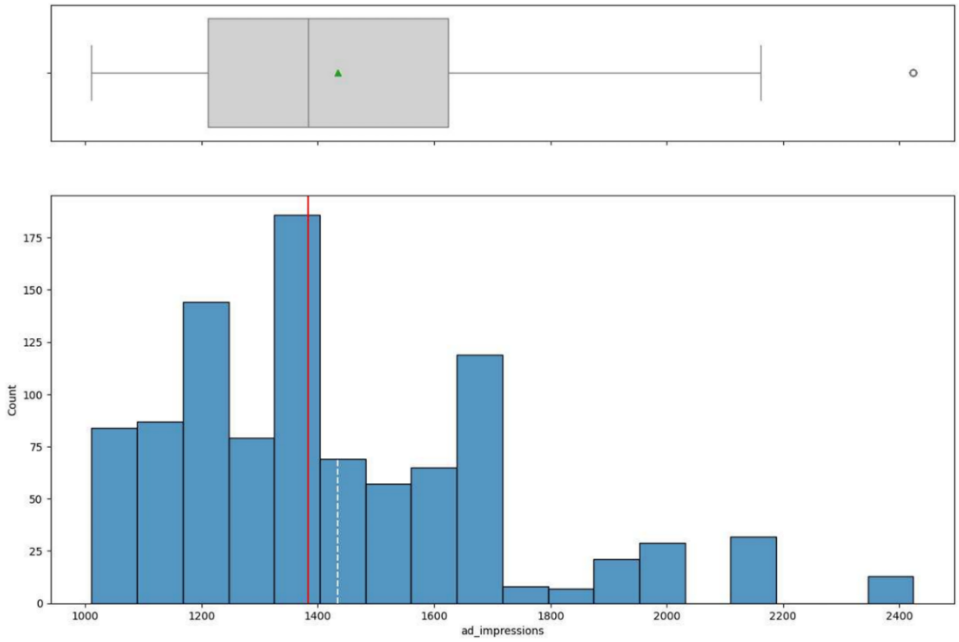


fig 12: Ad impressions

- Views Trailer:

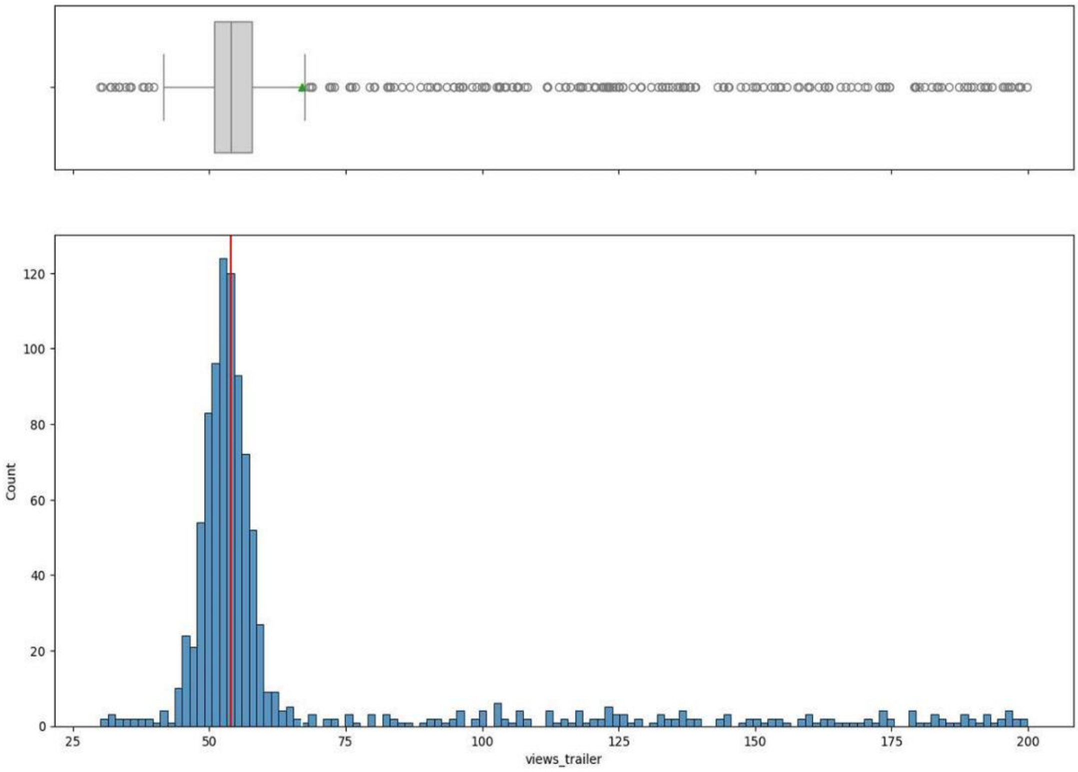


fig 13: views trailer

- Views content:

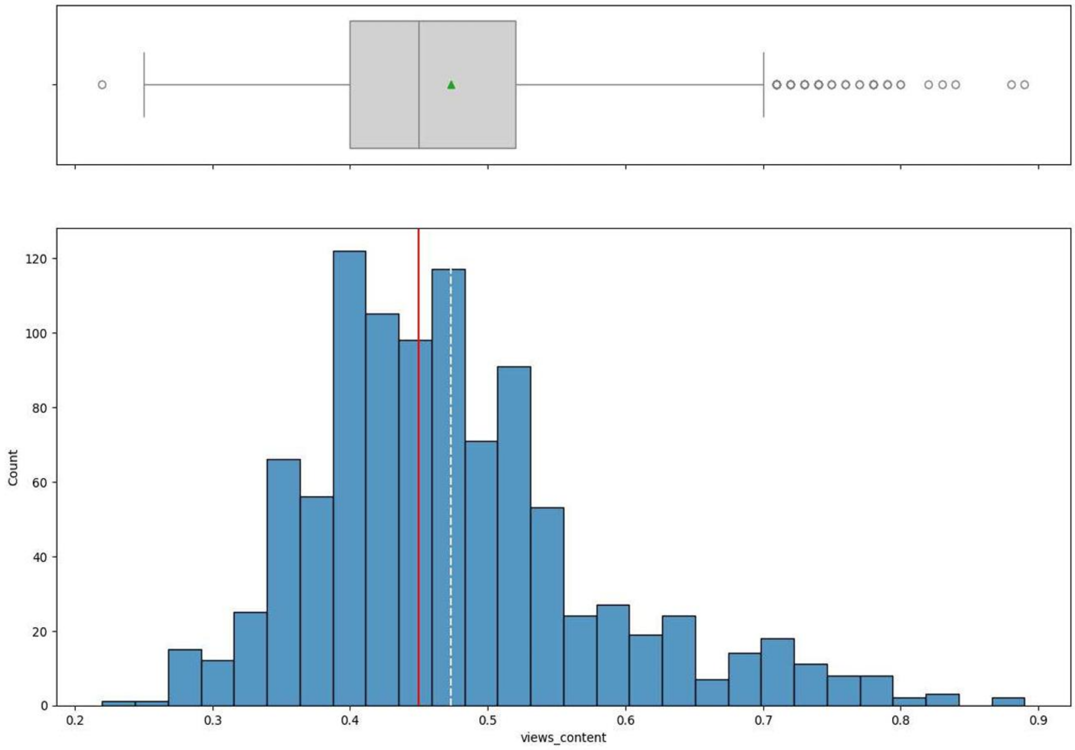


fig 14: views content

- Genre:

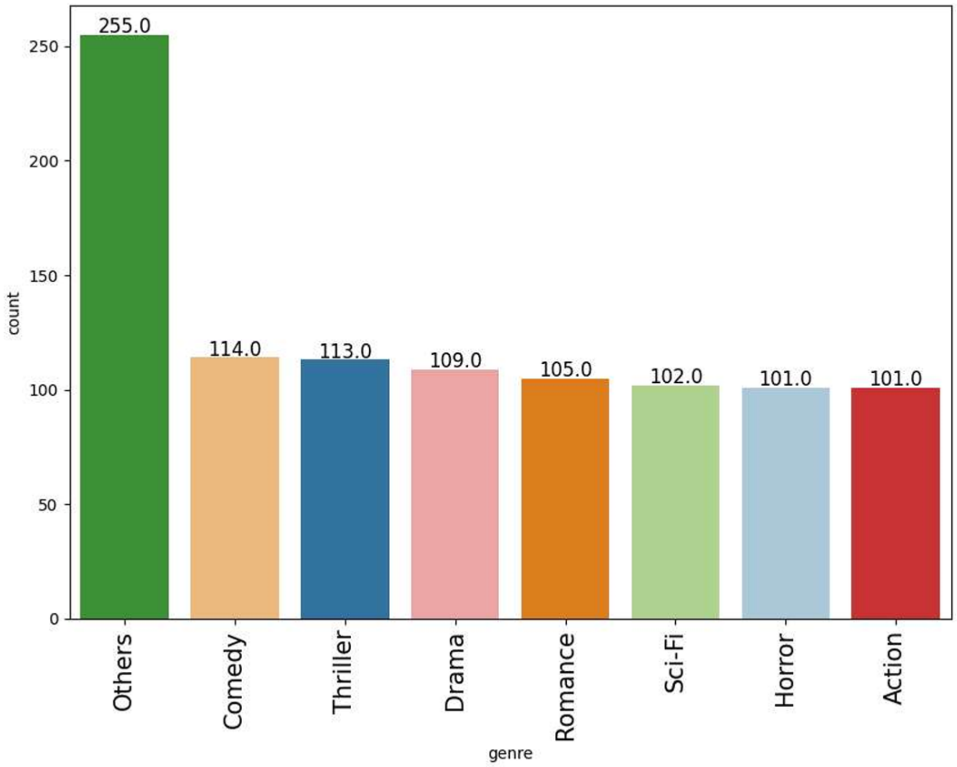


fig 15: Genre

- Day of week:

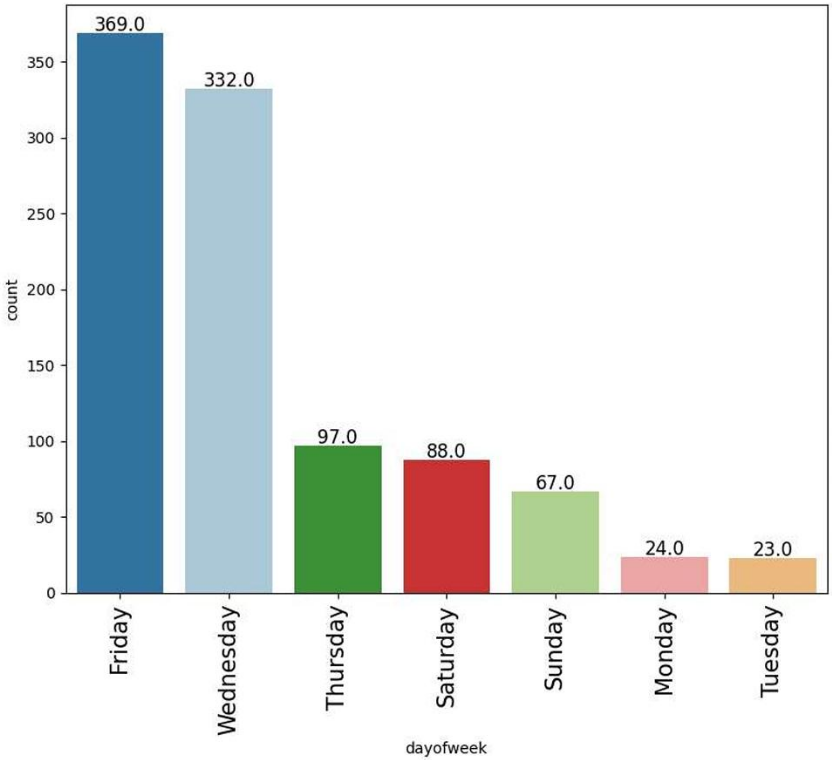


fig 16: Day of week

- Season:

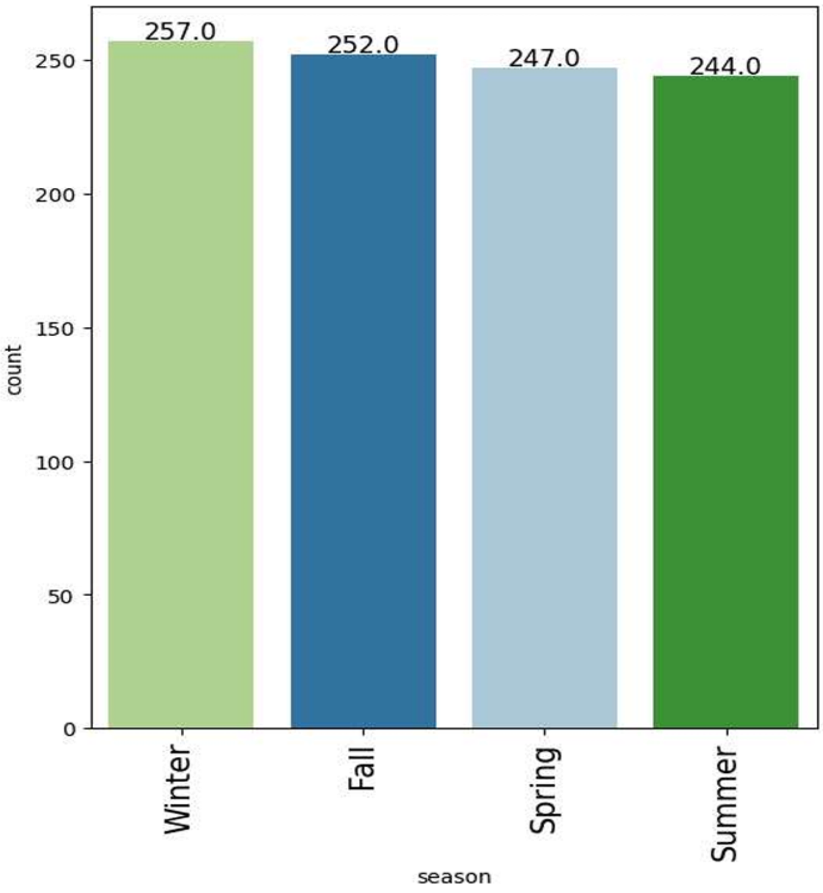


fig 17: Season

- Major Sports Event:

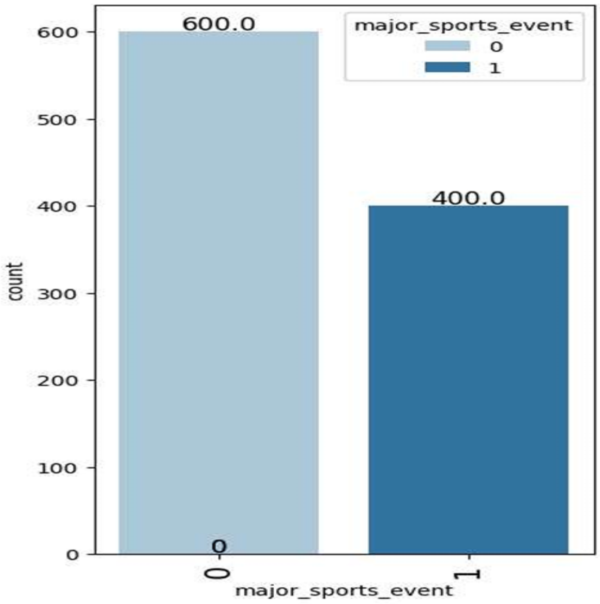


fig 18: Major sports event

Bivariate Analysis

In Bivariate Analysis, we analyse the relationship between 2 variables, which could be numerical vs numerical, categorical vs numerical data.

- Visitor vs views content:

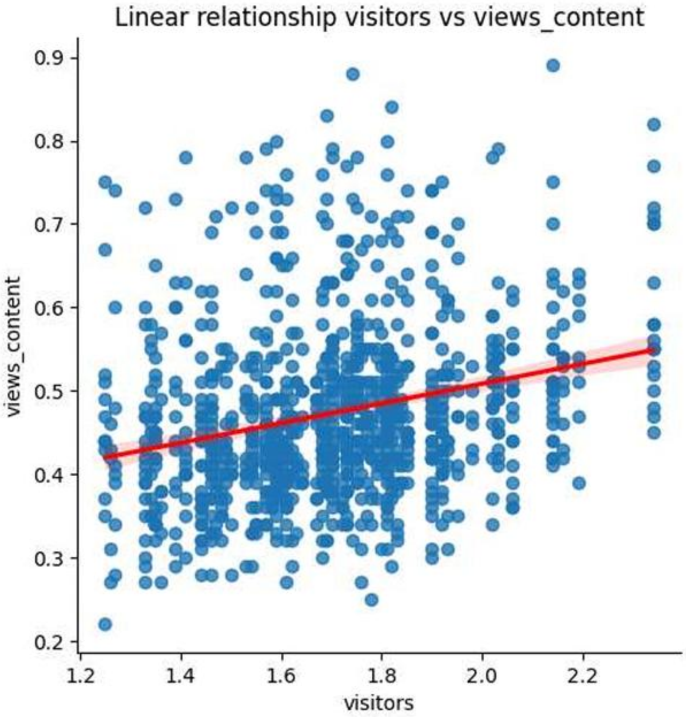


fig 19: Visitors vs views content

- Ad impression vs views content:

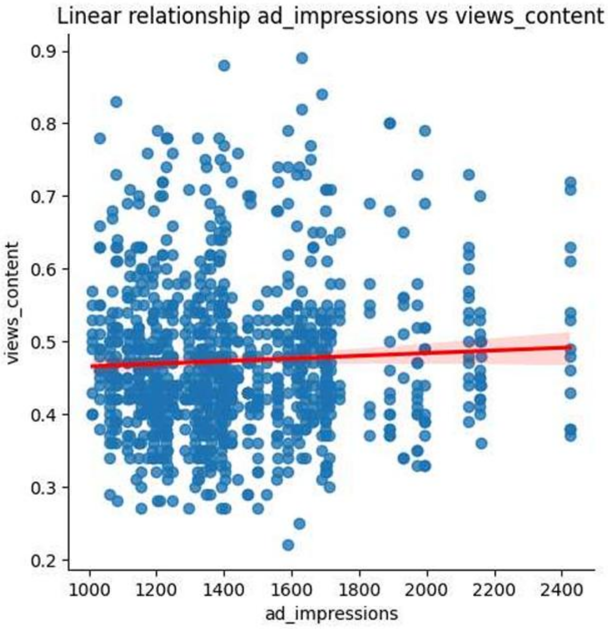


fig 20: Ad impressions vs views content

- Views trailer vs views content:

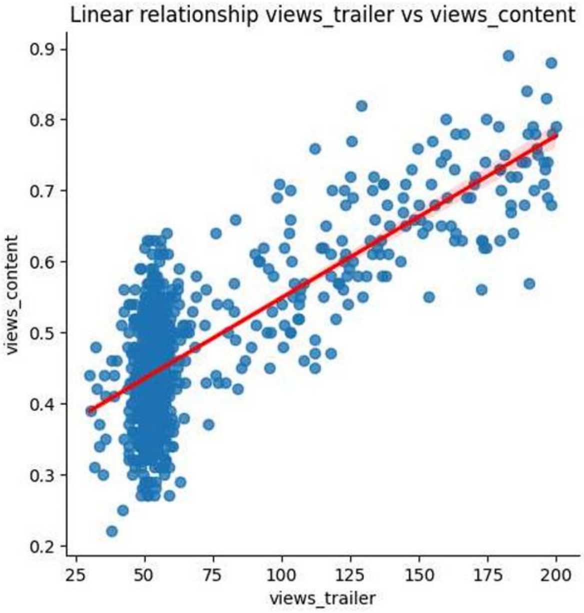


fig 21: Views trailer vs views content

- Major sports event vs views content:

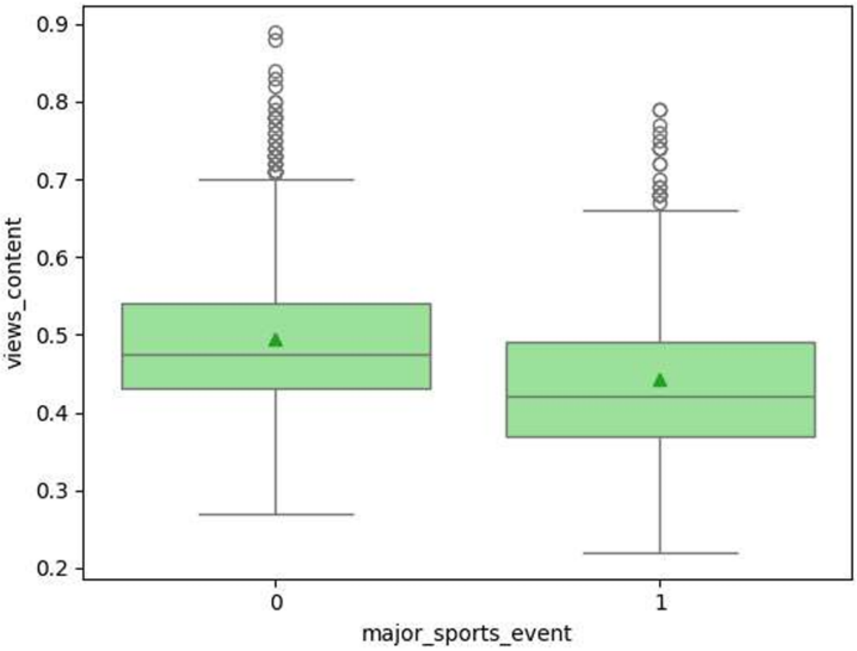


fig 22: Major sports event vs views content

- Season vs views content:

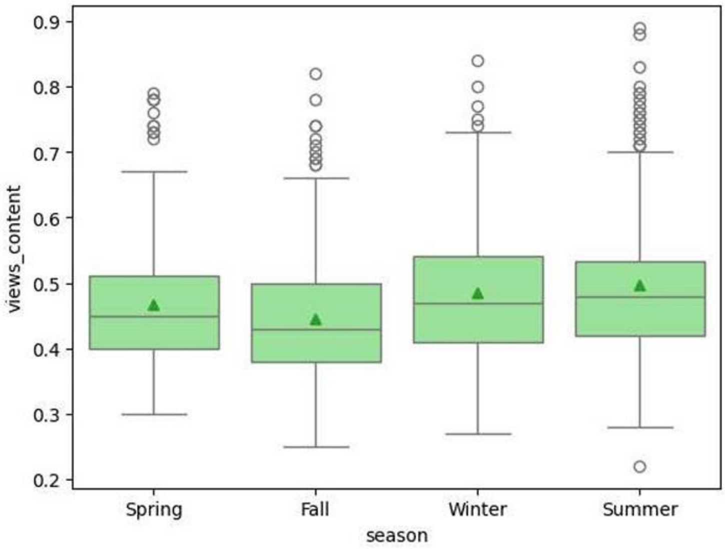


fig 23: Season vs views content

- Genre vs views content:

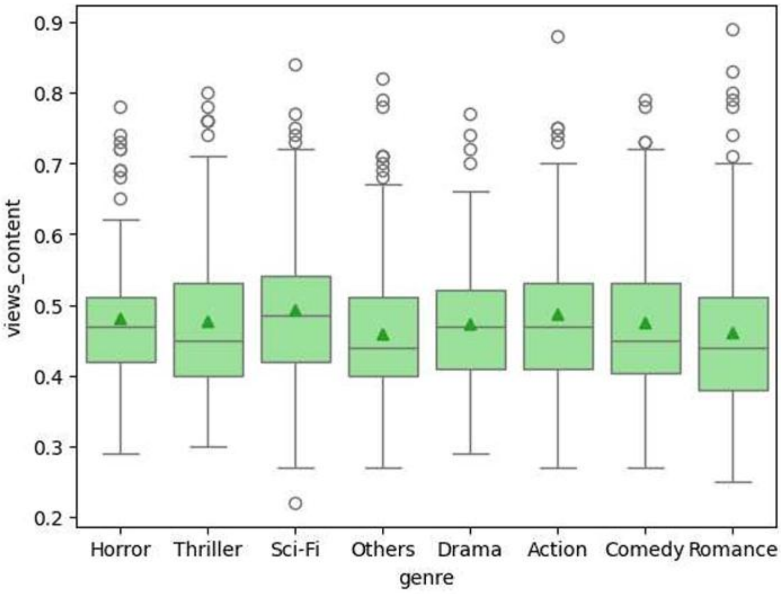


fig 24: Genre vs views content

- Day of week vs views content:

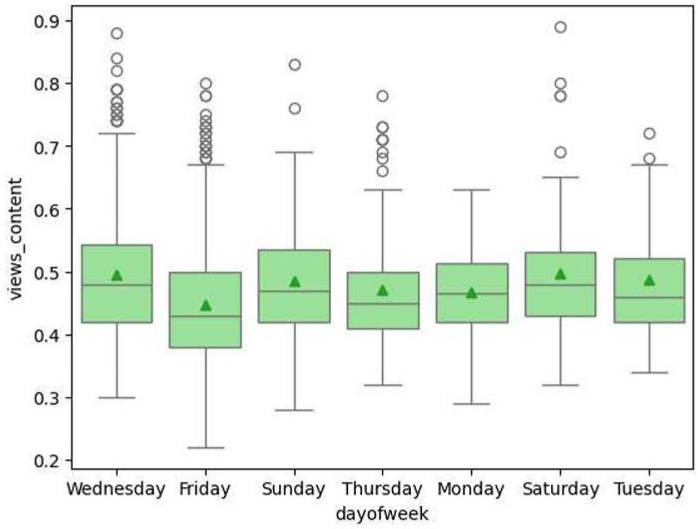


fig 25: Day of week vs views content

Heatmap and Pairplot:

For better understanding of the numerical vs numerical data, we created Heatmap and Pairplot.

- **Heatmap:**
  In the heatmap we can see the corelation between different numerical data and here we can see that "views_trailer" is significantly corelated to "views_content" and "visitors" are mildly corelated to "views_content". While for "ad_impressions" there is not much corelation with the "views_content".

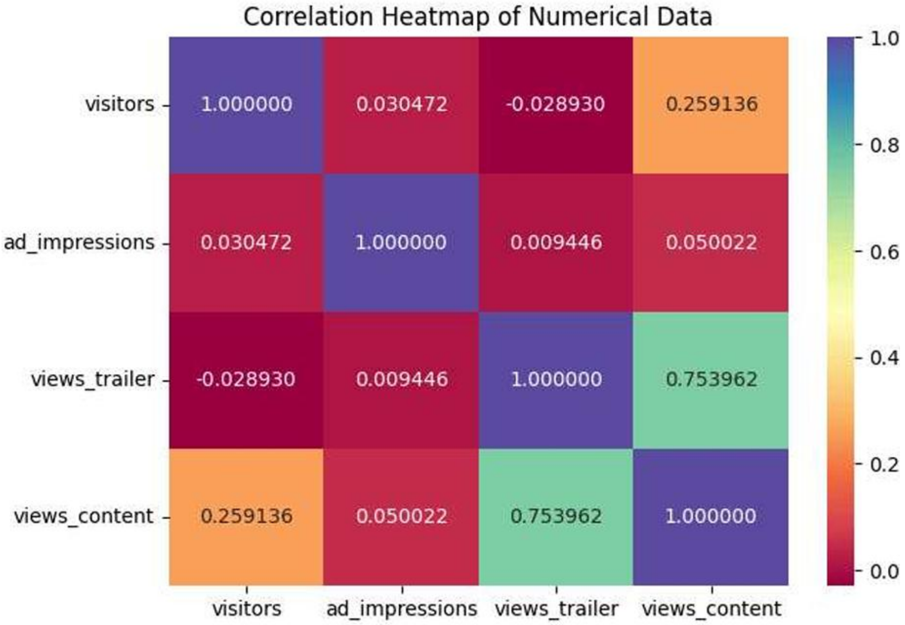


fig 26: Heatmap

- **Pairplot:**
  Here we can visualize the numerical vs numerical data in a scatter plot all at once.

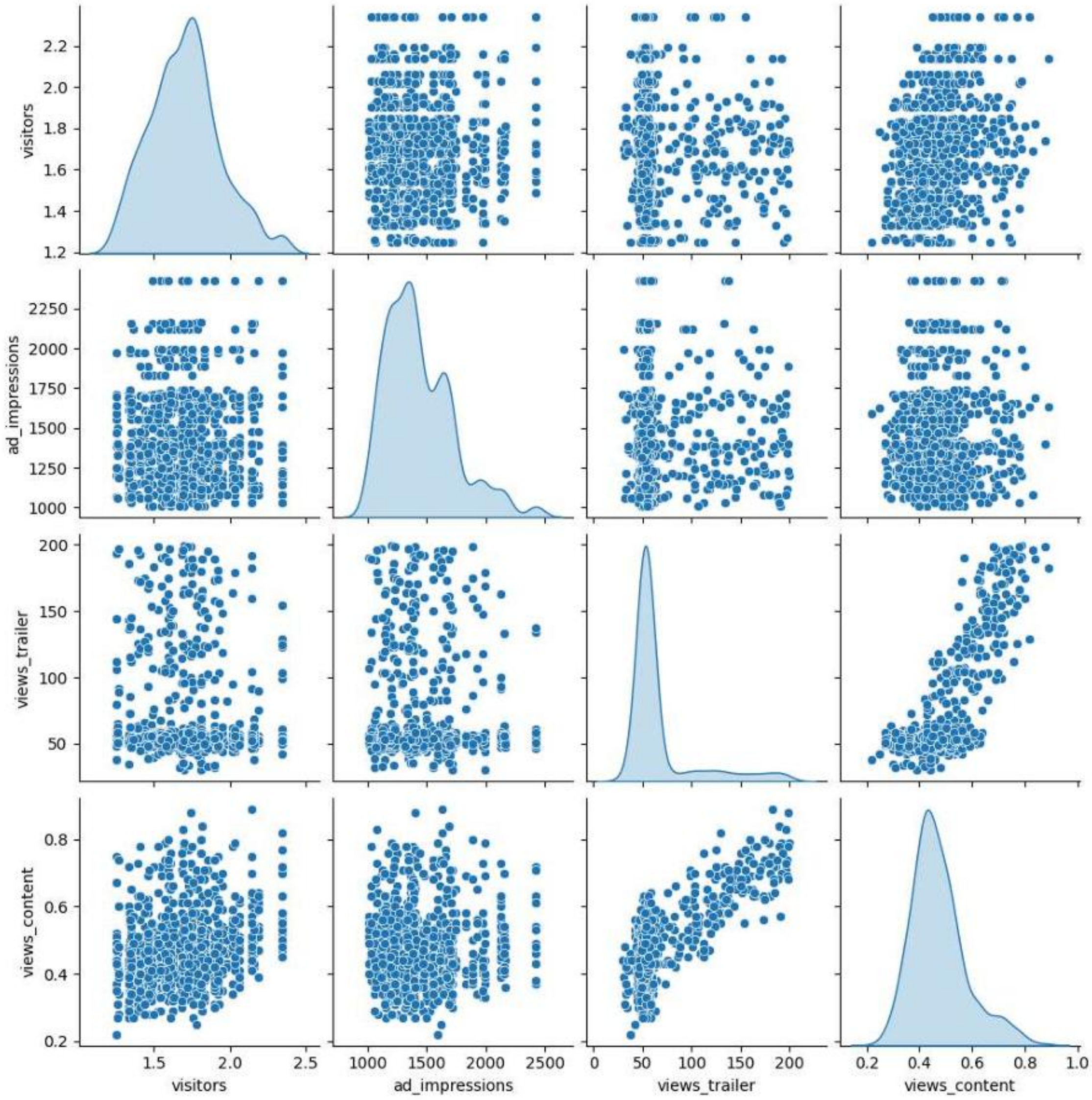


fig 27: Pairplot

QUESTION AND ANSWERS

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

➢ The distribution of the content views looks like bell shaped but skewed towards right i.e. positively skewed. It does have the outliers but the median and mean are not very far off from each other which means outliers are the genuine ones.

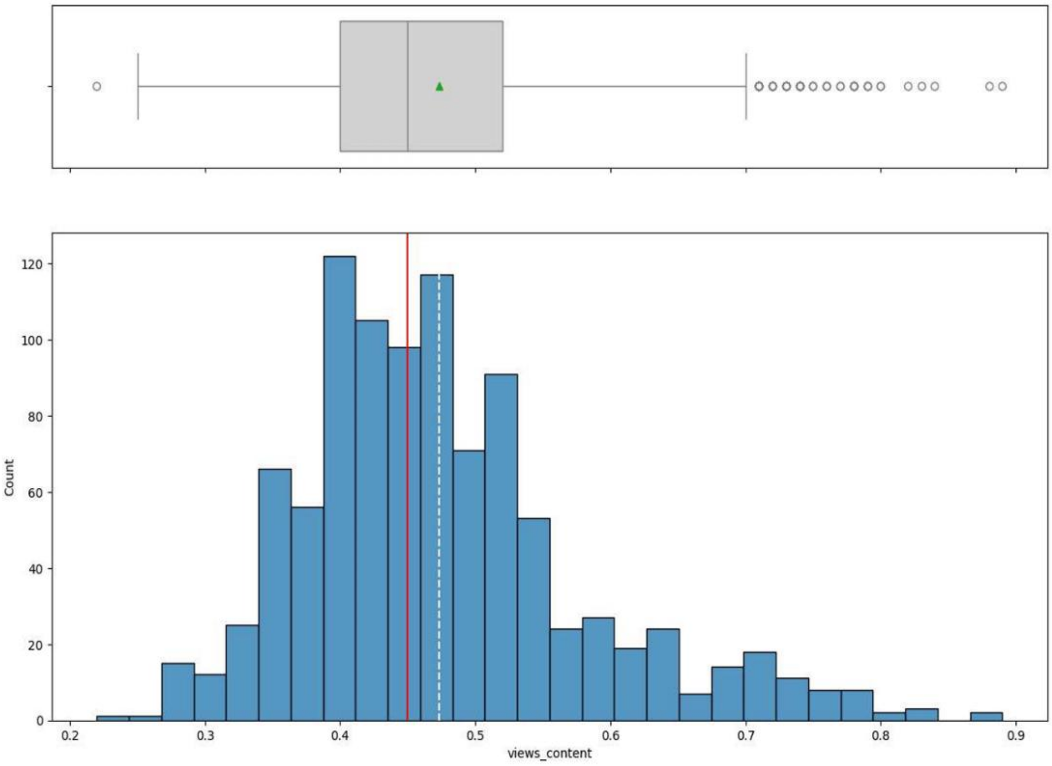


fig 28: Distribution of content views

2. What does the distribution of genres look like?

- Comedy leads with the highest number of titles, Actions and Horror being the least ones. Although it can be said that the difference between the highest and the lowest is not significant, if we exclude Others.

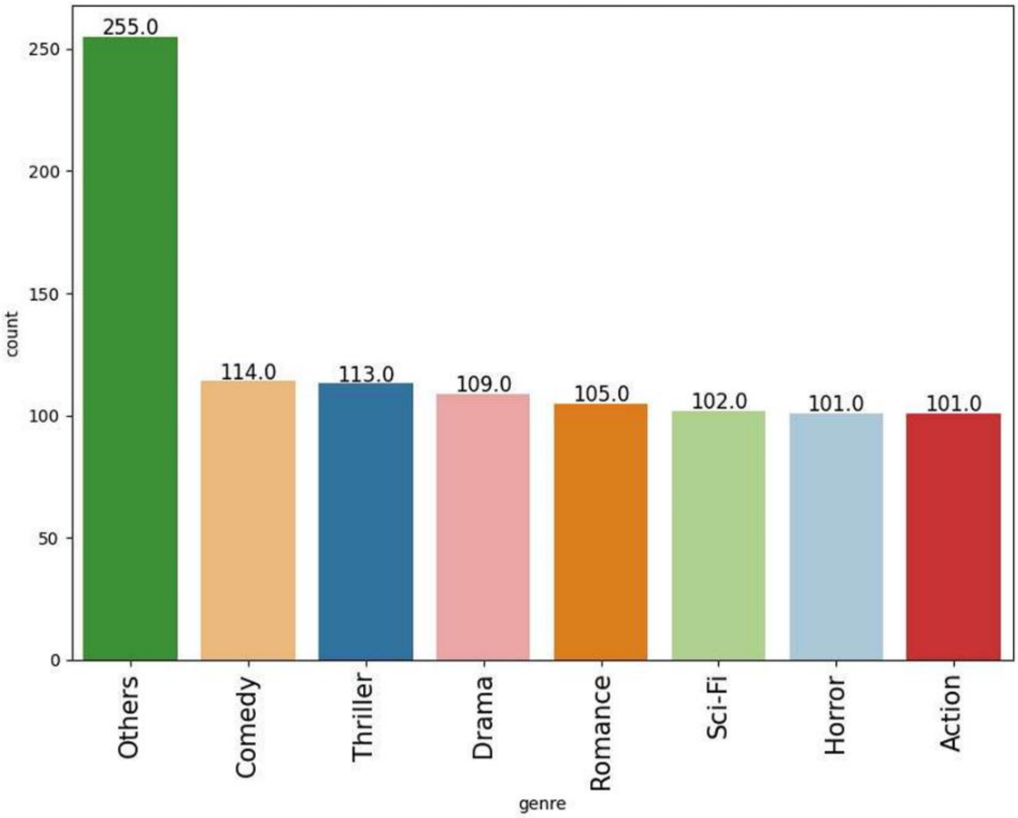


fig 29: Distribution of genre

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

- Wednesday seems to have highest viewership. On the other hand, the median viewership of Friday is lowest. Also, both the days have high number of outliers on the higher side. That could mean people prefer to watch content more on these days compared to other days.

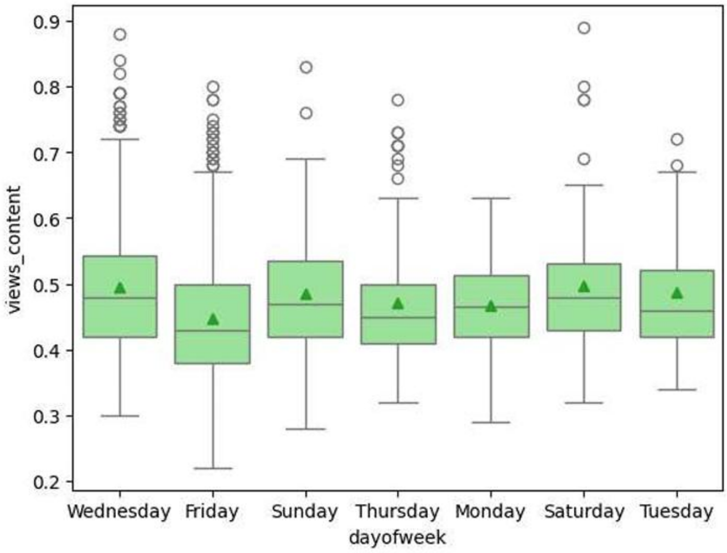


fig 30: Day of week vs views content

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

- Summers have highest amount of viewership on day 1 of the content release. This is followed by Winter.

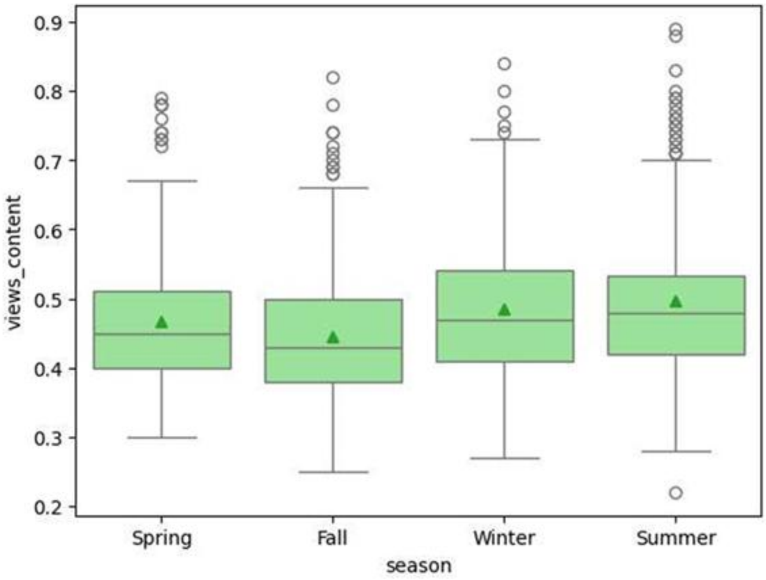


fig 31: Season vs views content

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

- There is a significant and positive correlation between trailer view and content views, with a corelation coefficient of 0.75 as per the heatmap. This means a higher number of trailer views is a strong indicator of higher first-day content viewership.

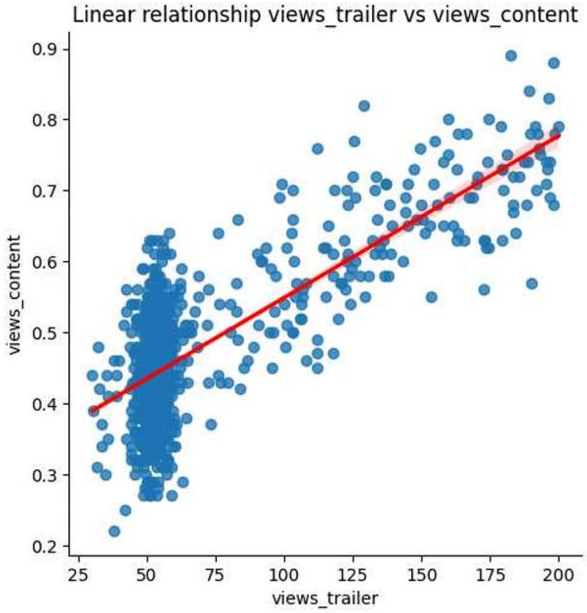


fig 32: Views trailer vs views content

KEY INSIGHTS AS PER EDA

- Views on content is directly and strongly linked to views on trailer
- Visitor are also positively corelated to the views on content
- On the day of any major sporting event, viewership dips a bit
- Different genres available on the OTT platform is more or less similar in numbers.
- Summers have the highest number of the viewership on content release day.
- Wednesdays have the highest number of the viewership on content release day.

Outliers Treatment

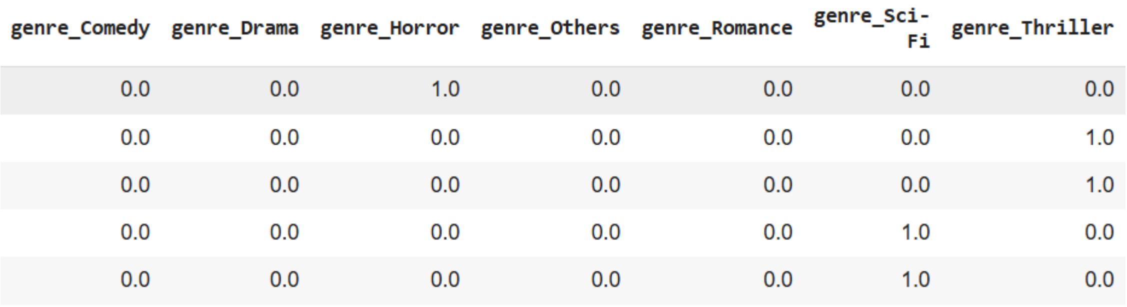
We have detected outliers in the data set, but those outliers seem to be genuine, therefore, we are not going to treat the outliers. We will keep the outliers in the data and will move ahead with the same.

Feature Engineering

We need to apply One-Hot Encoding in the categorical data, i.e. for Genre, Season and Day of Week. This process converts each category into a new binary column. For example, the genre column becomes multiple new columns, like genre_comedy, genre_drama, etc., with a value of 1 if the data point belongs to that category and 0 otherwise.

Example:

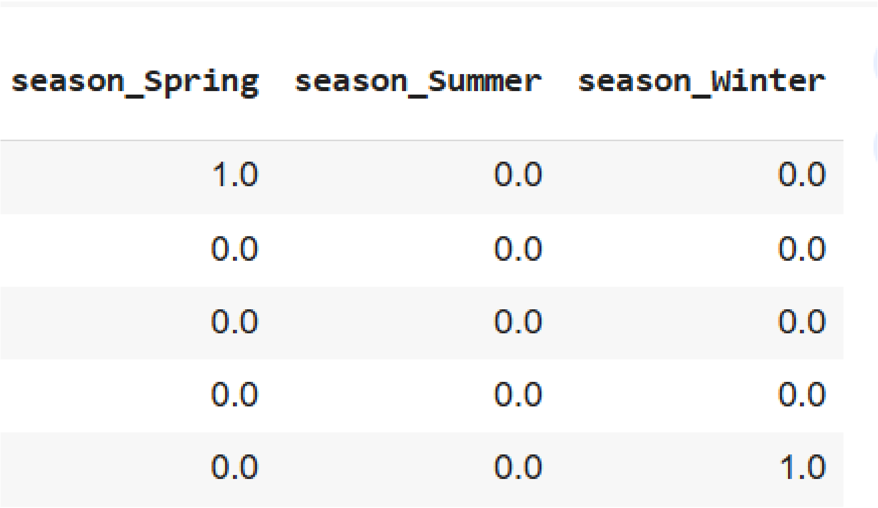
- Genre



genre_Comedy	genre_Drama	genre_Horror	genre_Others	genre_Romance	genre_Sci-Fi	genre_Thriller
0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0

fig 33: Genre

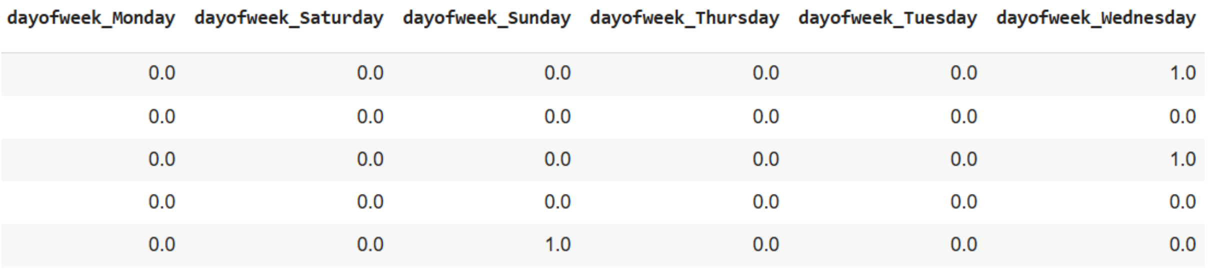
- Season



season_Spring	season_Summer	season_Winter
1.0	0.0	0.0
0.0	0.0	0.0
0.0	0.0	0.0
0.0	0.0	0.0
0.0	0.0	1.0

fig 34: Season

- Day of week



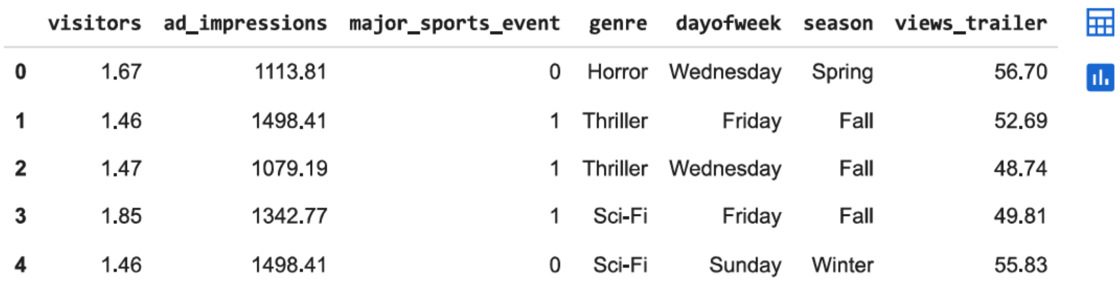
dayofweek_Monday	dayofweek_Saturday	dayofweek_Sunday	dayofweek_Thursday	dayofweek_Tuesday	dayofweek_Wednesday
0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0

fig 35: Day of Week

DATA PREPARATION FOR MODELING

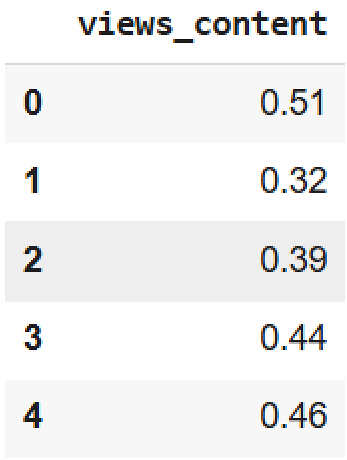
We need to prepare our data for modelling.

- Splitting the data into features (X) and the target (y).



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

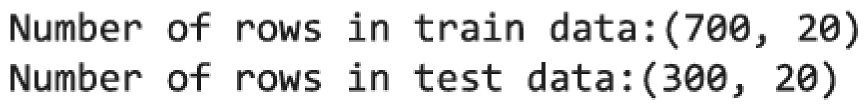
fig 36: Feature(X)



	views_content
0	0.51
1	0.32
2	0.39
3	0.44
4	0.46

fig 37: Target(y)

- Feature Engineering as discussed above.
- Splitting data in 70:30 ratio, where 70% data is used for training and 30% data will be used for testing.



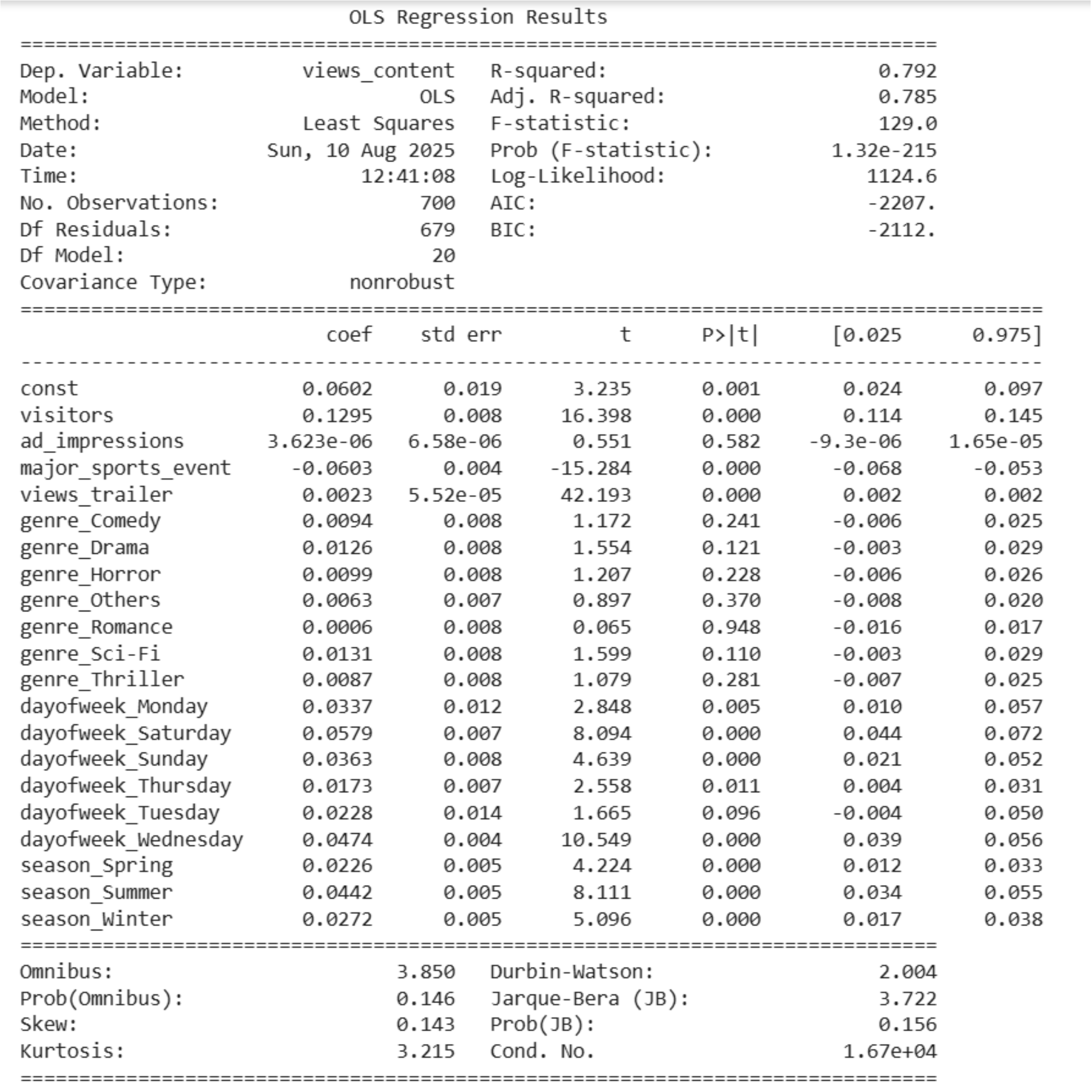
```
Number of rows in train data:(700, 20)
Number of rows in test data:(300, 20)
```

fig 38: Split of train and test data

BUILDING MODEL- LINEAR REGRESSION

OLS Model 1

Below, is our 1st OLS (Ordinary Least Squares) Regression model. Model is trained on 700 entries.



```
                            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:                Sun, 10 Aug 2025   Prob (F-statistic):          1.32e-215
Time:                        12:41:08   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
==============================================================================
```

fig 39: OLS model 1

Summary:

R squared: The R-squared value is 0.792. It means that approximately 79.2% of the variance in views_content can be explained by the independent variables included in the model.

Adj. R squared: The adjusted R-squared is 0.785, which is very close to the R-squared. This value generally ranges between 0 and 1 and higher the value better the fit.

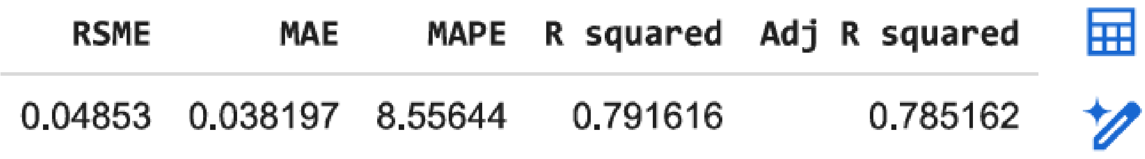
P-values: There are some significant features, and some not so significant features. We'll be removing not so significant features in further analysis. Features having p value more than 0.05 will be considered as not so significant. As, when p>0.05, we will fail to reject H_0 (i.e. H_0= There is no significant relationship between the feature and the target).

Coefficients: The coefficients in the OLS regression model represent the change in the target (views_content) for every one-unit change in a feature, while holding all other features as constant.

Constant: It is the Y constant, here the constant value is 0.0602.

Performance of OLS Model 1

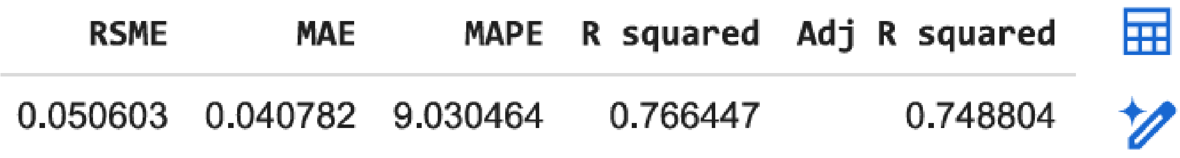
Performance of **training data** for ols model 1



RSME	MAE	MAPE	R squared	Adj R squared
0.04853	0.038197	8.55644	0.791616	0.785162

fig 40: Performance of training data OLS model 1

Performance of **testing data** for ols model 1



RSME	MAE	MAPE	R squared	Adj R squared
0.050603	0.040782	9.030464	0.766447	0.748804

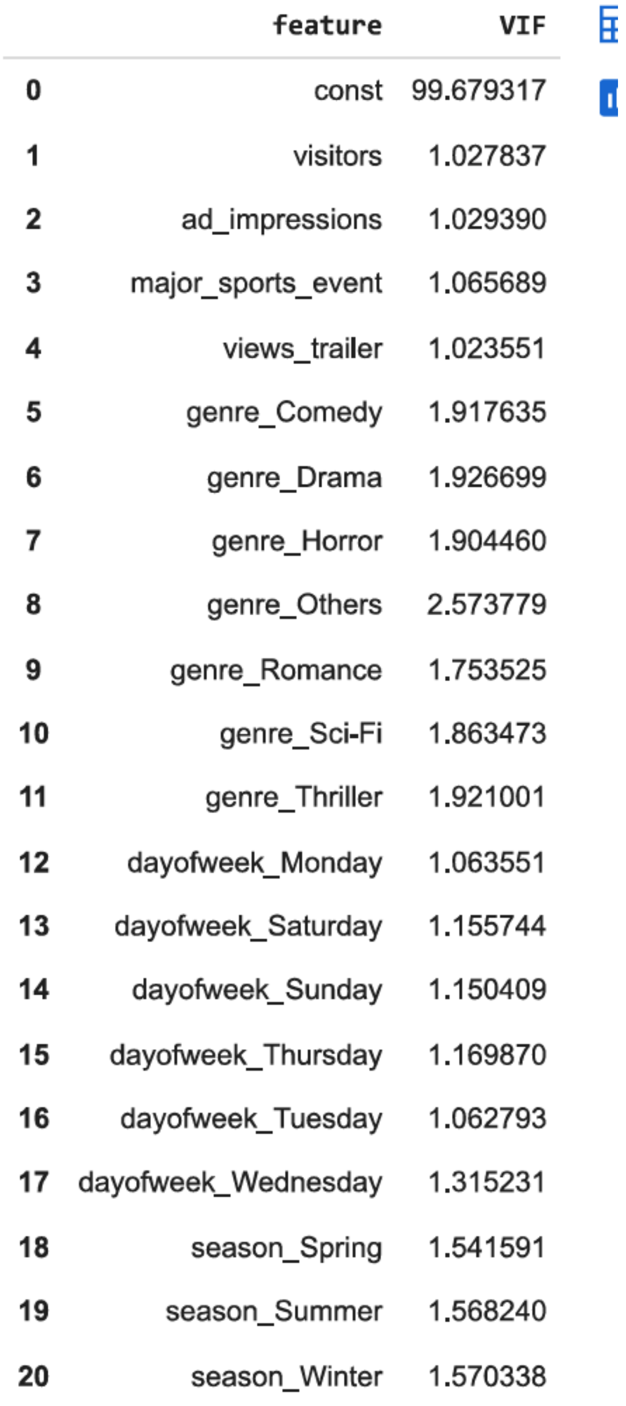
fig 41: Performance of testing data OLS model 1

After comparing the above 2 results we can say that our model is not overfitting or underfitting. As the values above are quite similar, especially, Adj R squared and MAPE values.

But we did have p values greater than 0.05 for different features, so we'll need to remove them.

TEST OF MULTICOLLINEARITY

Here, we'll be testing VIF (Variance Inflation Factor). This is also, one of the assumptions for Linear Regression. Our VIF for features is less than 5 which means there is no multicollinearity. **Our model satisfies this assumption**.

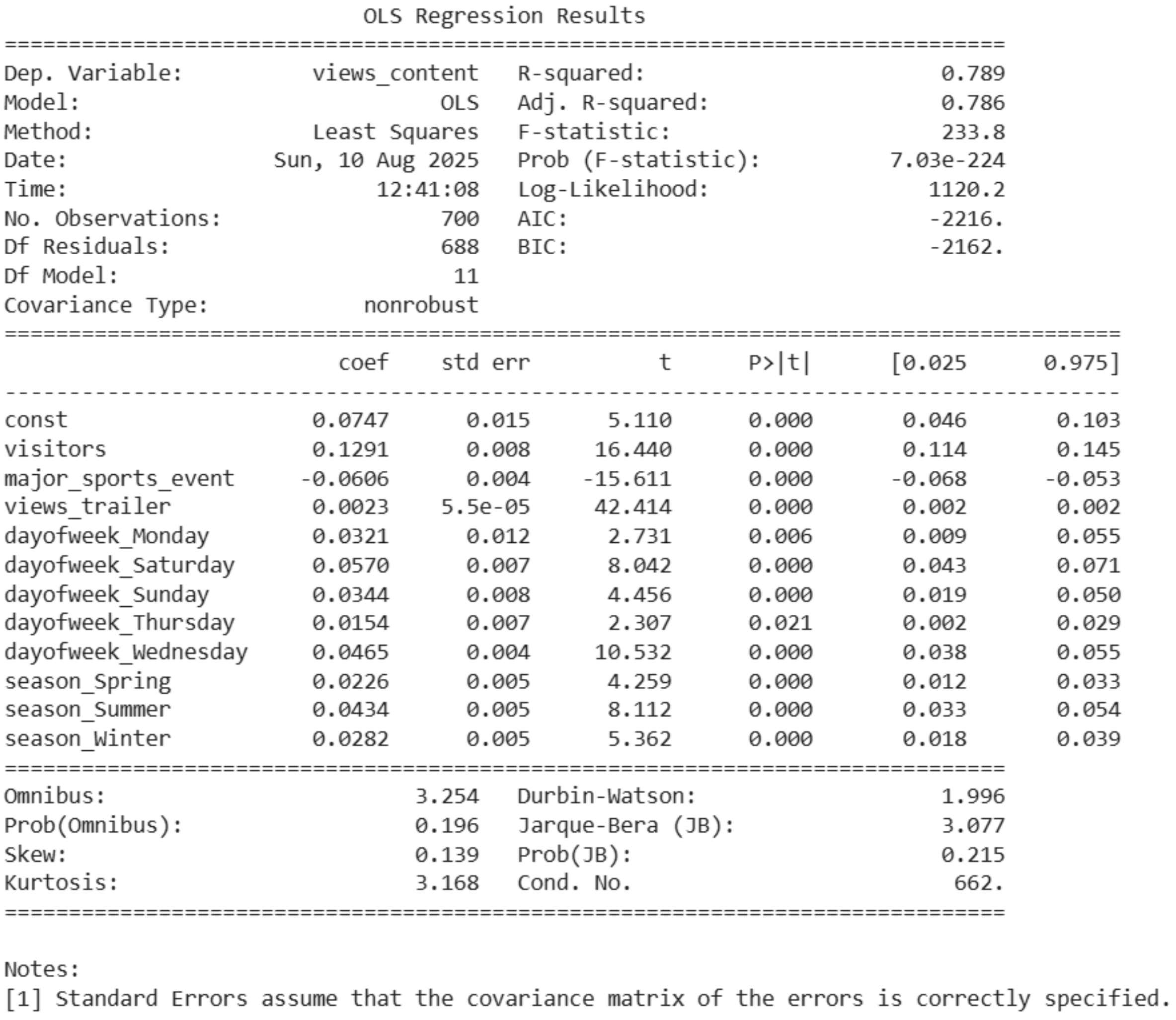


	feature	VIF
0	const	99.679317
1	visitors	1.027837
2	ad_impressions	1.029390
3	major_sports_event	1.065689
4	views_trailer	1.023551
5	genre_Comedy	1.917635
6	genre_Drama	1.926699
7	genre_Horror	1.904460
8	genre_Others	2.573779
9	genre_Romance	1.753525
10	genre_Sci-Fi	1.863473
11	genre_Thriller	1.921001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

fig 42: VIF Test

Removal Of Not So Significant Features

As we know that we have p values greater than 0.05 for different features. So, those particular feature are now insignificant for our model. So, we'll now remove them and test the model again with the name of **OLS model 2.**

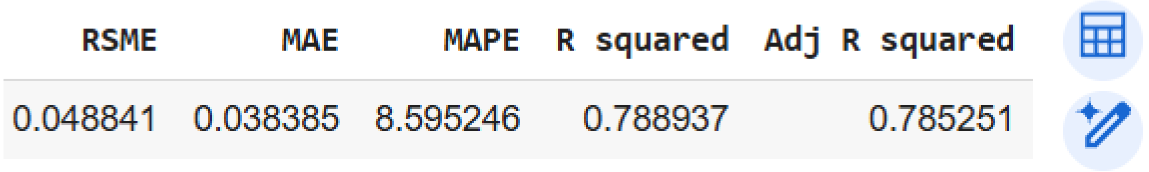


```
                            OLS Regression Results
==============================================================================
Dep. Variable:           views_content   R-squared:                       0.789
Model:                             OLS   Adj. R-squared:                  0.786
Method:                  Least Squares   F-statistic:                     233.8
Date:                 Sun, 10 Aug 2025   Prob (F-statistic):          7.03e-224
Time:                         12:41:08   Log-Likelihood:                 1120.2
No. Observations:                  700   AIC:                            -2216.
Df Residuals:                      688   BIC:                            -2162.
Df Model:                           11
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.0747      0.015      5.110      0.000       0.046       0.103
visitors                0.1291      0.008     16.440      0.000       0.114       0.145
major_sports_event     -0.0606      0.004    -15.611      0.000      -0.068      -0.053
views_trailer           0.0023    5.5e-05     42.414      0.000       0.002       0.002
dayofweek_Monday        0.0321      0.012      2.731      0.006       0.009       0.055
dayofweek_Saturday      0.0570      0.007      8.042      0.000       0.043       0.071
dayofweek_Sunday        0.0344      0.008      4.456      0.000       0.019       0.050
dayofweek_Thursday      0.0154      0.007      2.307      0.021       0.002       0.029
dayofweek_Wednesday     0.0465      0.004     10.532      0.000       0.038       0.055
season_Spring           0.0226      0.005      4.259      0.000       0.012       0.033
season_Summer           0.0434      0.005      8.112      0.000       0.033       0.054
season_Winter           0.0282      0.005      5.362      0.000       0.018       0.039
==============================================================================
Omnibus:                         3.254   Durbin-Watson:                   1.996
Prob(Omnibus):                   0.196   Jarque-Bera (JB):                3.077
Skew:                            0.139   Prob(JB):                        0.215
Kurtosis:                        3.168   Cond. No.                         662.
==============================================================================

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

fig 43: OLS Model 2

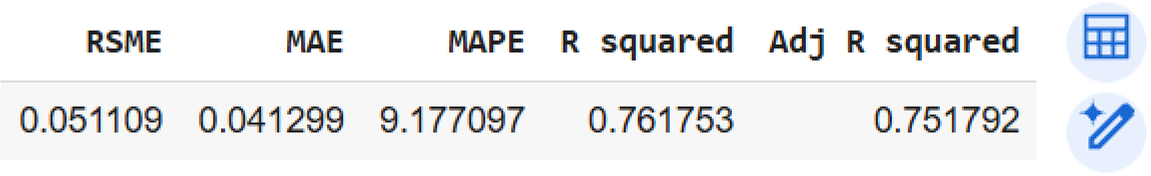
Now, p value<0.05.

Now we'll check the performance again.



RSME	MAE	MAPE	R squared	Adj R squared
0.048841	0.038385	8.595246	0.788937	0.785251

fig 44: Performance of training data OLS Model 2



RSME	MAE	MAPE	R squared	Adj R squared
0.051109	0.041299	9.177097	0.761753	0.751792

fig 45: Performance of testing data OLS Model 2

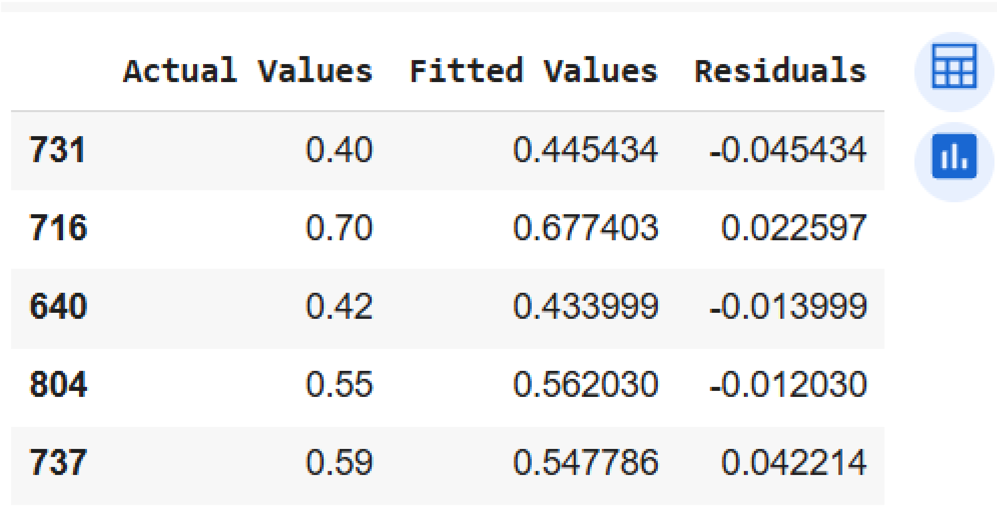
Test of Assumptions

Following needs to be satisfied for the test of assumptions for linear regression

1. No Multicollinearity
2. Linearity of variables
3. Independence of error terms
4. Normality of error terms
5. No heteroscedasticity

No Multicollinearity: As discussed above our data is free from multicollinearity.

Linearity of variables and Independence of error terms: This is tested by fitted vs residual plot.



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

fig 46: Residual values

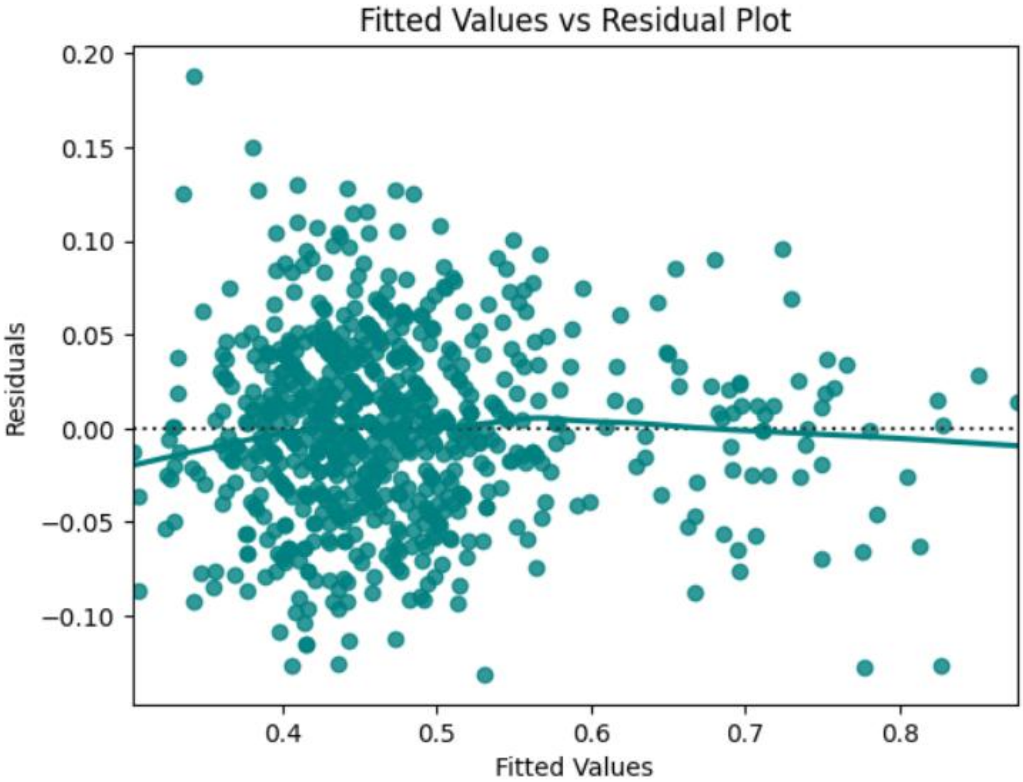


fig 47: Fitted values vs Residuals plot

In the above graph we don't see any type of pattern. Hence, the **test of linearity and independence is satisfied.**

Normality of error terms: This can be tested by the shape of histogram of residuals, QQ plot of residual and Shapiro-Wilk test (statistically).

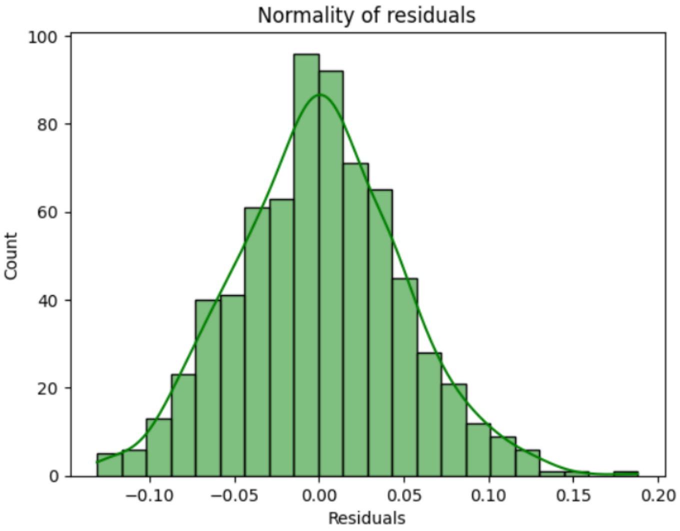


fig 48: Histogram of residuals

The above graph looks like a normal curve.

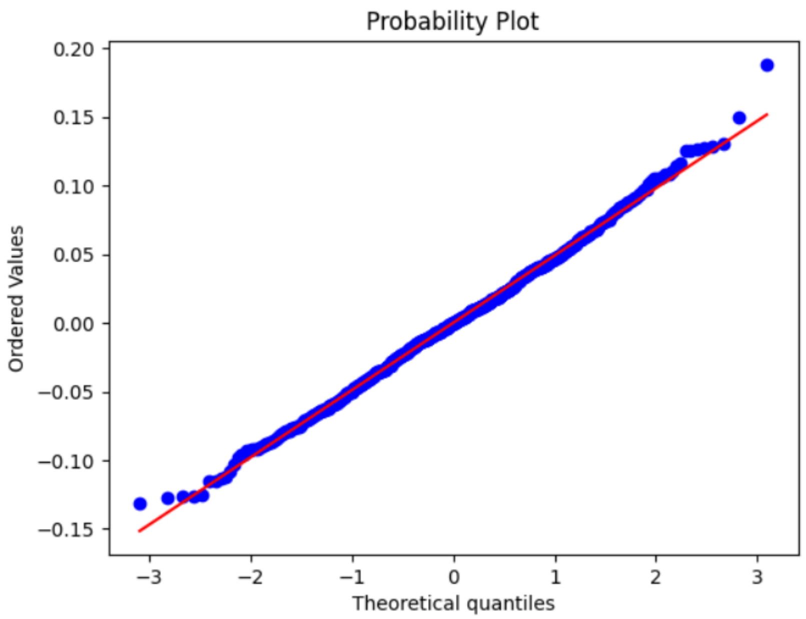
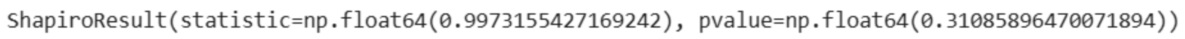


fig 49: Q-Q Plot of Residuals

Above, Q-Q plot is mostly normal except in the trails.

Shapiro-Wilk Test: Our p value more than 0.05, therefore, we fail to reject the null hypothesis (H_0: Residuals are normally distributed).

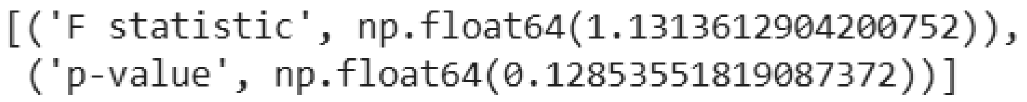


```
ShapiroResult(statistic=np.float64(0.9973155427169242), pvalue=np.float64(0.31085896470071894))
```

fig 50: Shapiro Wilk Result

Test of Heteroscedasticity: If the variance of residual is symmetrically distributed across the regression line, then the data is said to be homoscedastic. Otherwise, it is termed as heteroscedastic.

This is determined by p value. If p value is more than 0.05 then the test of heteroscedasticity is satisfied. As this confirms our data is Homoscedastic.



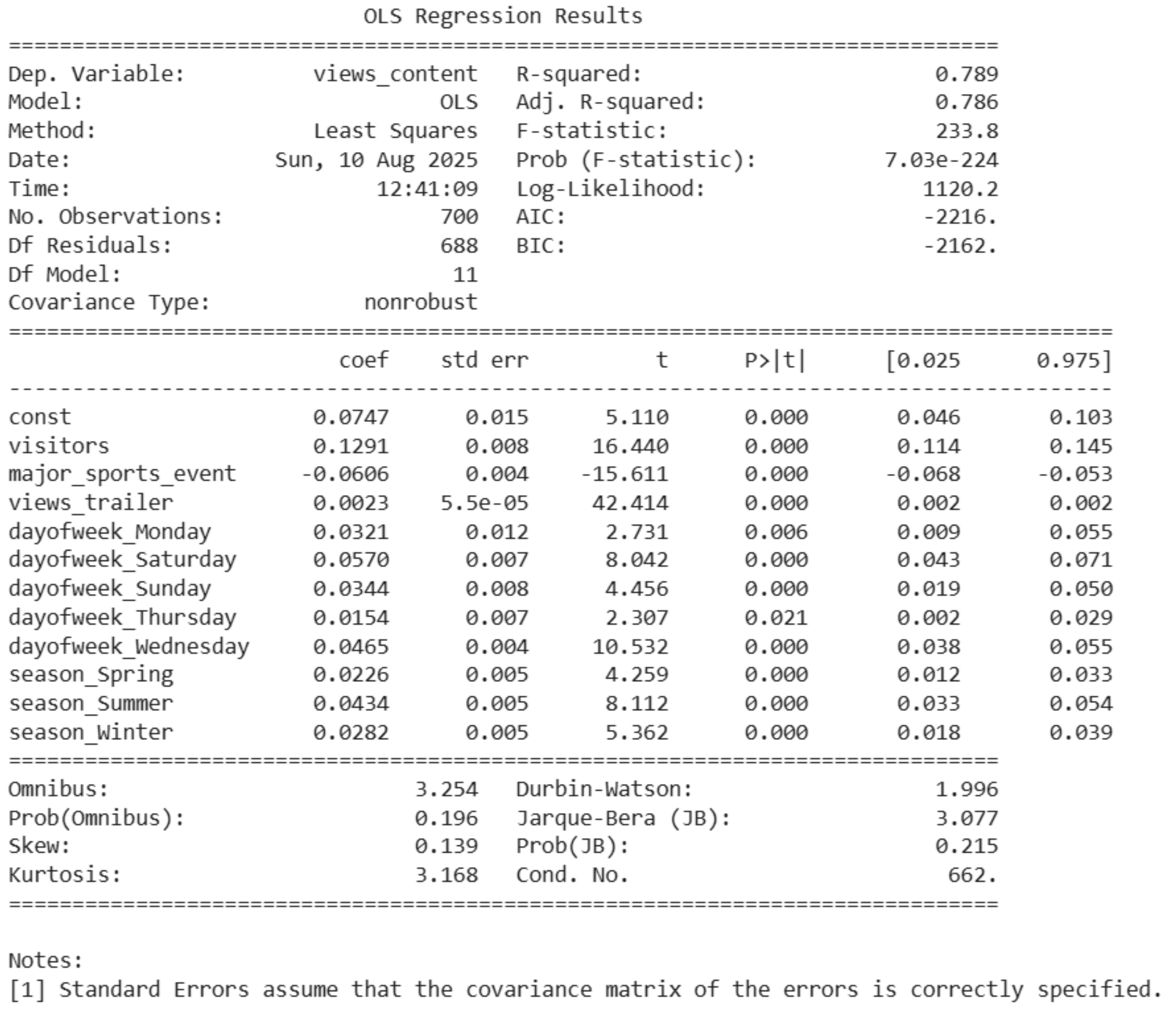
```
[('F statistic', np.float64(1.1313612904200752)),
 ('p-value', np.float64(0.12853551819087372))]
```

fig 51: Test of Heteroscedasticity

Now, we'll build the final OLS model.

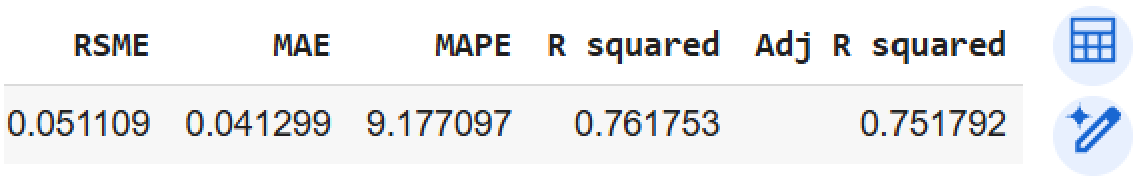
FINAL OLS MODEL

After all the testing above, now we'll be building final OLS Model.



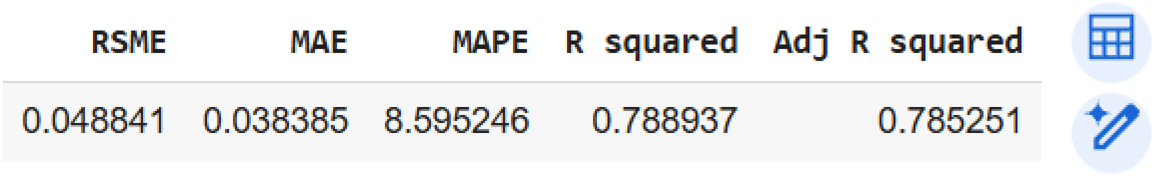
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           views_content   R-squared:                       0.789
Model:                             OLS   Adj. R-squared:                  0.786
Method:                  Least Squares   F-statistic:                     233.8
Date:                 Sun, 10 Aug 2025   Prob (F-statistic):          7.03e-224
Time:                         12:41:09   Log-Likelihood:                 1120.2
No. Observations:                  700   AIC:                            -2216.
Df Residuals:                      688   BIC:                            -2162.
Df Model:                           11
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.0747      0.015      5.110      0.000       0.046       0.103
visitors                0.1291      0.008     16.440      0.000       0.114       0.145
major_sports_event     -0.0606      0.004    -15.611      0.000      -0.068      -0.053
views_trailer           0.0023    5.5e-05     42.414      0.000       0.002       0.002
dayofweek_Monday        0.0321      0.012      2.731      0.006       0.009       0.055
dayofweek_Saturday      0.0570      0.007      8.042      0.000       0.043       0.071
dayofweek_Sunday        0.0344      0.008      4.456      0.000       0.019       0.050
dayofweek_Thursday      0.0154      0.007      2.307      0.021       0.002       0.029
dayofweek_Wednesday     0.0465      0.004     10.532      0.000       0.038       0.055
season_Spring           0.0226      0.005      4.259      0.000       0.012       0.033
season_Summer           0.0434      0.005      8.112      0.000       0.033       0.054
season_Winter           0.0282      0.005      5.362      0.000       0.018       0.039
==============================================================================
Omnibus:                         3.254   Durbin-Watson:                   1.996
Prob(Omnibus):                   0.196   Jarque-Bera (JB):                3.077
Skew:                            0.139   Prob(JB):                        0.215
Kurtosis:                        3.168   Cond. No.                         662.
==============================================================================

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

fig 52: Final OLS Model



RSME	MAE	MAPE	R squared	Adj R squared
0.051109	0.041299	9.177097	0.761753	0.751792

fig 53: Performance of Final OLS Model (training data)



RSME	MAE	MAPE	R squared	Adj R squared
0.048841	0.038385	8.595246	0.788937	0.785251

fig 54: Performance of Final OLS Model (testing data)

This will be our final model as it satisfies all the assumptions. All the p values are less than 0.05. MAPE values don't differ much between training and testing data. Adj R squared values are also quite similar for training and testing data.

The model successfully identifies several significant drivers of first-day content viewership, explaining approximately **76% of the variance** in viewership.

KEY TAKEAWAYS FOR THE BUSINESS

- The model successfully identifies several significant drivers of first-day content viewership, explaining approximately **76% of the variance** in viewership.
- The key factors are **platform visitors, trailer views, and strategic content release timing** related to the day of the week and season.
- Critically, the model also reveals that **major sporting events negatively impact viewership**, and that **genre and ad impressions are not statistically significant** drivers.
- **Focus on Driving Platform Visitors:** The variable visitors have the highest coefficient (0.1291), making it the single most influential factor.
- **Leverage Trailer Performance:** ShowTime should prioritize creating high-quality, engaging trailers and ensure they are widely promoted to maximize their impact.
- **Summer** and **Winter** are the best seasons for content releases, as the model shows a significant positive impact on viewership during these times.
- ShowTime should **avoid** releasing the content at the time of **major sporting events.**
- Genres and Ad impressions are **not significant drivers** for the growth of first day viewership of the content.
- ShowTime should release its content on **Saturday** and **Wednesday** as it is showing positive impact on the Day 1 viewership.