GREAT LEARNING

PREDICTION MODELING OVER VIEWERSHIP ON OTT

Project designer: *DURGESH KUMAR JHA*

Session: 2024-25

From the business perspective we have to find the driving factors which influences the trends for the OTT content on their platform, and analyse the data so we can provide the recommendation to the company.

Problem Statement

With the outbreak of COVID19, OTT services are striving to meet the growing entertainment appetite of viewers, with some platforms already experiencing a 46% increase in consumption and subscriber count as viewers seek fresh content. With innovations and advanced transformations, which will enable the customers to access everything they want in a single space, OTT platforms across the world are expected to increasingly attract subscribers on a concurrent basis.

Objective

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

Data Description

SI. NO	Column Name	Description
1	***Visitors***	Average number of visitors, in millions, to the platform in the past week

2	**Add_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

List of Image

Data Overview

	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

Image 1

- ***Above image shows us the top 5 rows of the data set.***
- We have data set with different Genres.

```
<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
```

Image 2

- ***Above image shows that we have total 1000 rows and 8 columns.***
- There are ***5 Numeric*** (float and int type) and ***3 string*** (object type) columns in data set.
- All the value in the data set given***, we have not any null value***.
- Our target variable is viewing content which is of float type.

	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

Image 3

- ***Above Image us statistical summary of the data set.***

- We can see the view content vary between 0.22 to 0.89 which shows the ***view content is on scale of 0-1.***

```
Number of duplicate rows =  0
```

Image 4

- We have ***no duplicate values*** in our data set

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

Image 5

- Above image shows that ***we do not have any missing value*** in our data set.

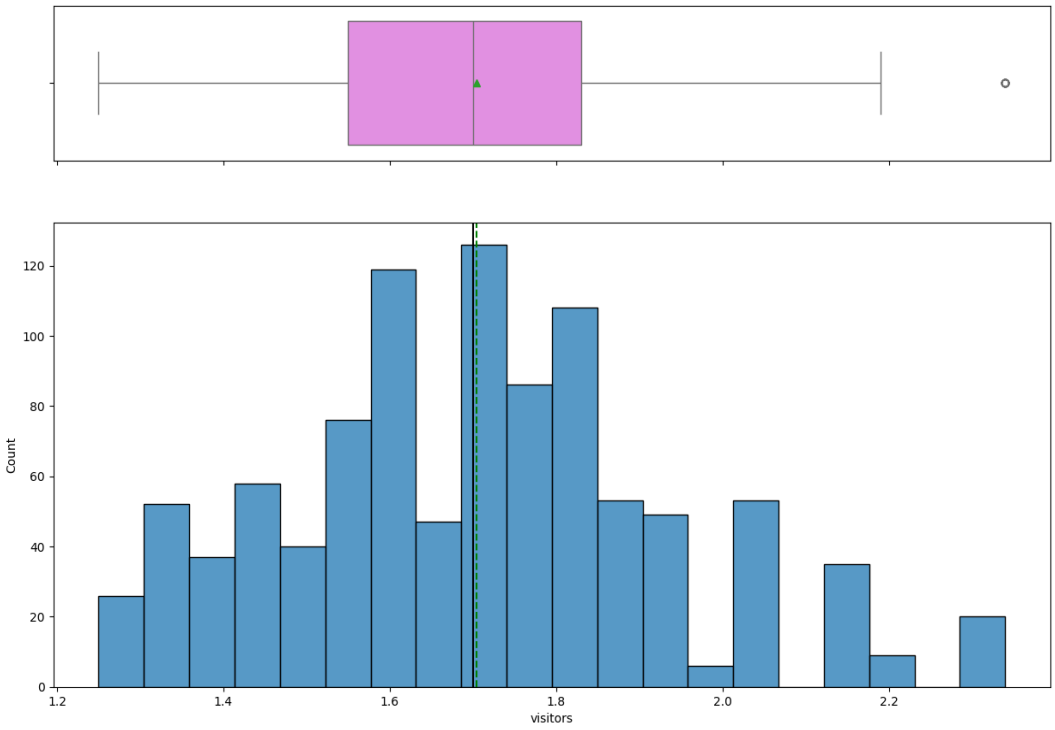


Image 6

- ***Above image shows that the number of the visitors are almost normally distributed.***
- And the ***mean*** value for the visitors is around ***1.7***.

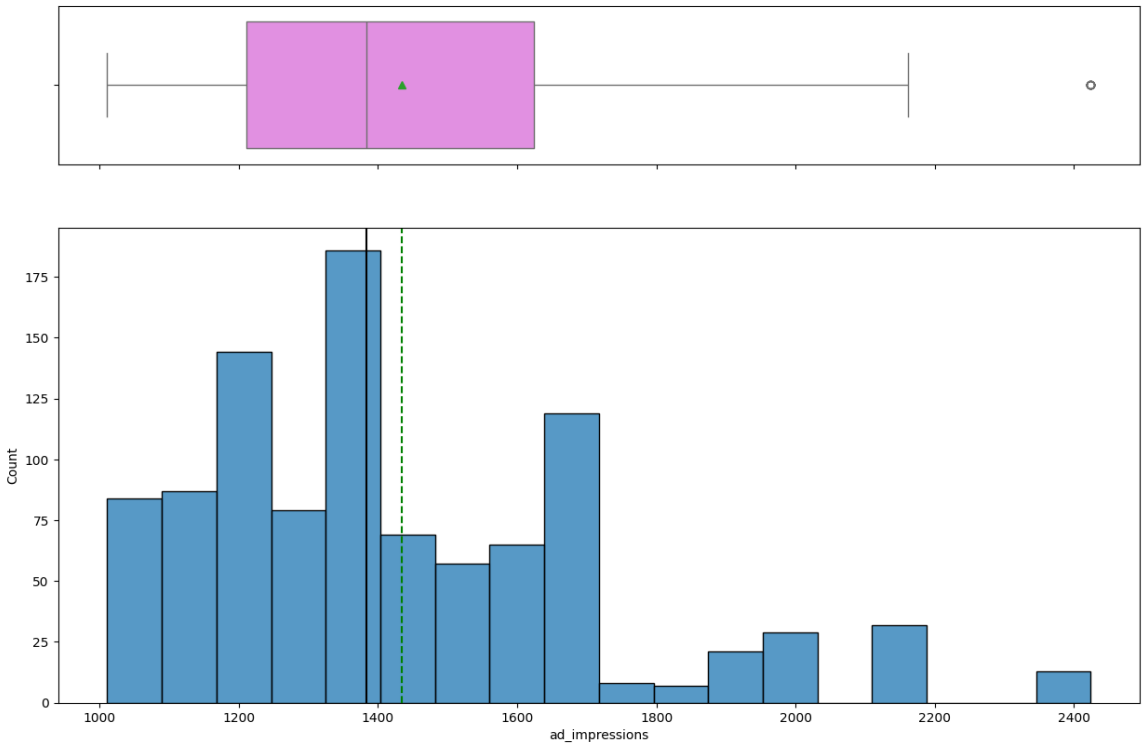


Image 7

- ***Above image shows our Ad impression data is Right skewed as the peak of our graph lies in Left side of our data set.***
- ***Mean*** of the Ad Impression is nearby to ***1424***.

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

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

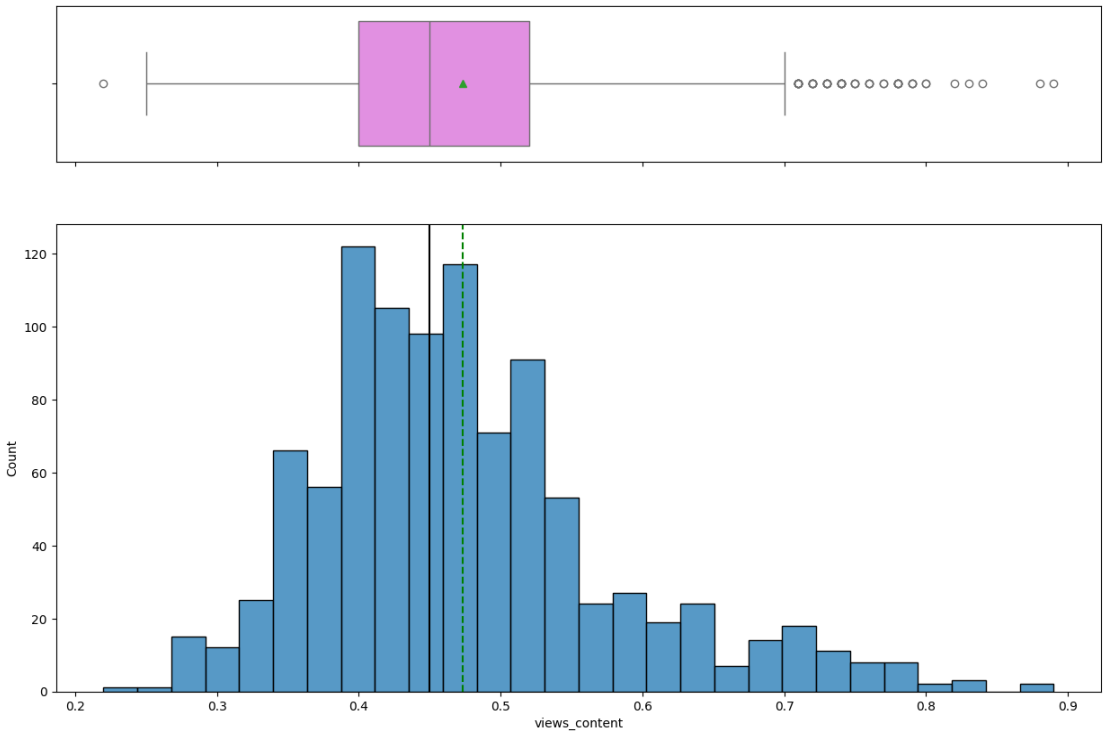


Image 8

- ***Above Image shows the distribution of the View content column in our data set, which is right skewed.***
- ***Our 50% of data is below the 0.47.***

2. What does the distribution of genres look like?

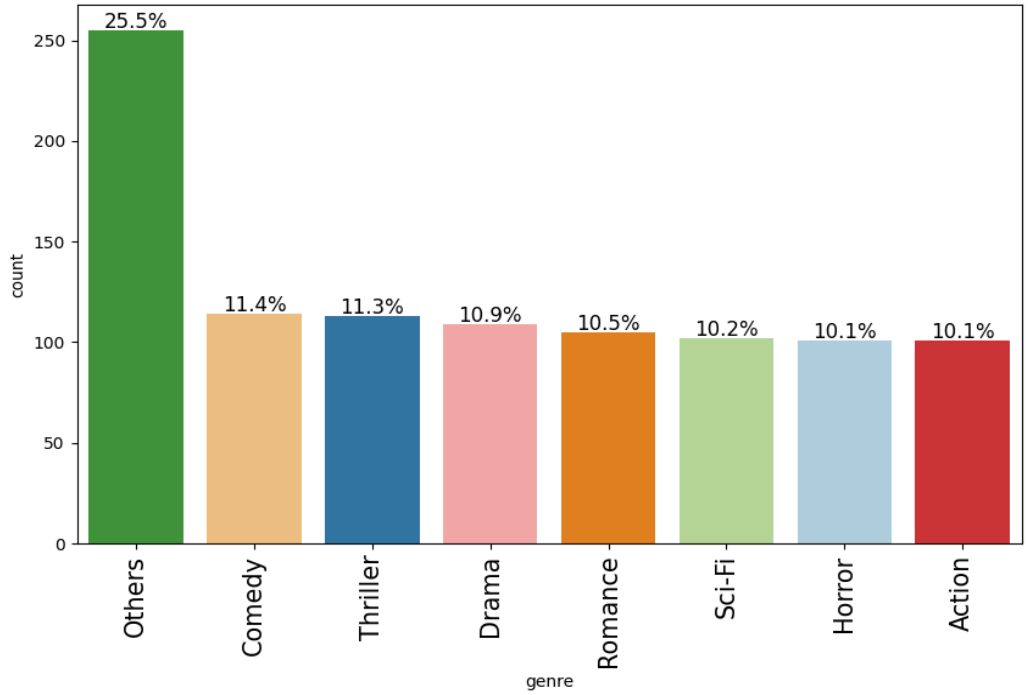


Image 9

- ***Above Image shows the distribution of the Genre.***
- ***Most famous Genre is Others about ¼ of the total Genres***, then Comedy and so on.
- ***Least popular Genres*** are Horror and Action with ***least 10.1%.***

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?

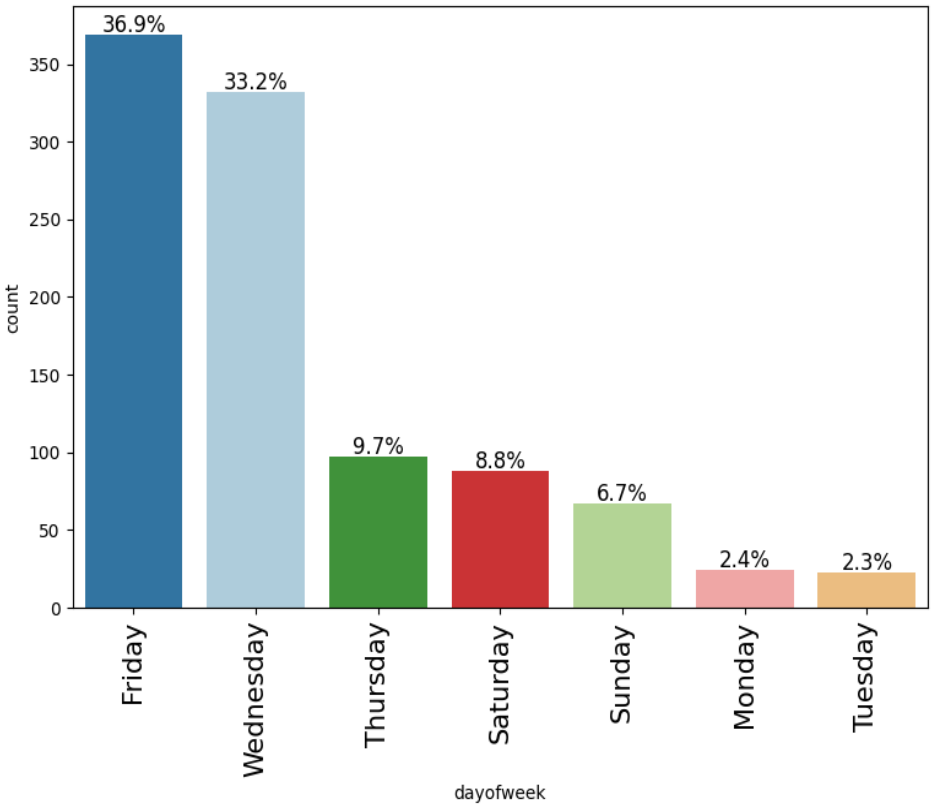


Image 10

- ***Above Image shows the effect of content release on different days.***
- If a content is released on ***Friday which covers more than 1/3 of total viewership,*** then the other days.
- If a content is released on ***Tuesday, has the least*** viewership, as it is a non-weekend.
- The content release on ***Saturday has only 8.8%*** effect on viewership is a little surprising as it is a weekend.

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

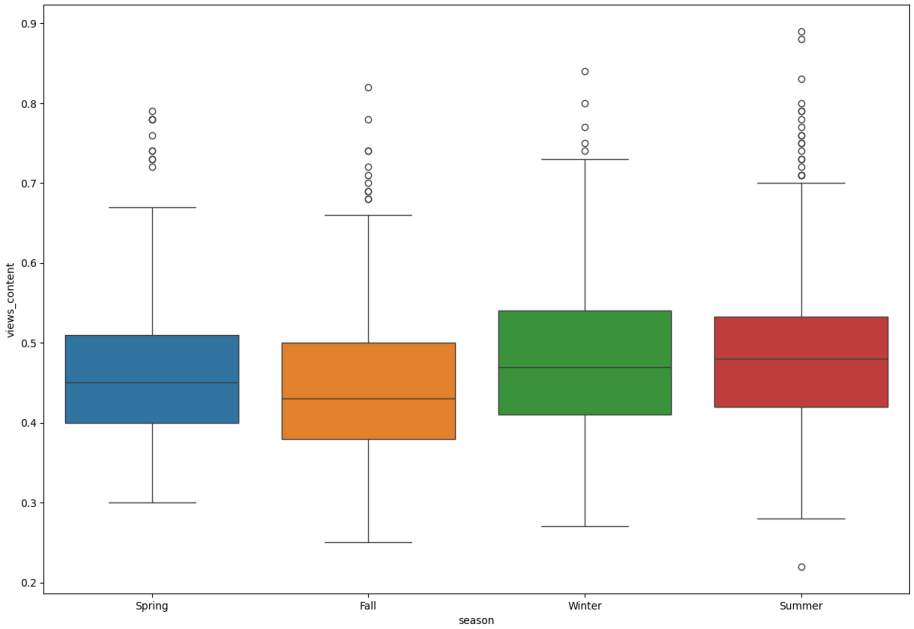


Image 11

- ***Above image show the variation in viewership with the change in season***

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

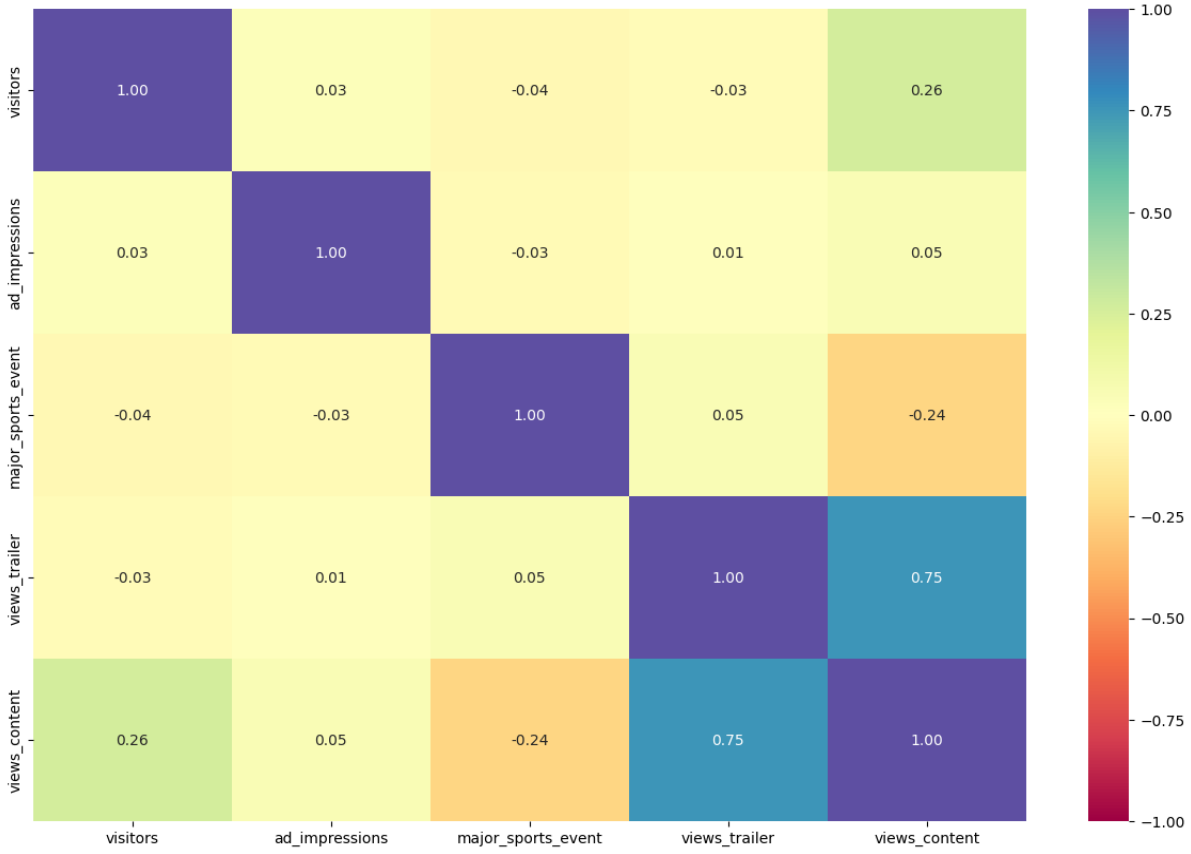


Image 12

- ***Above Image shows us the Correlation between the numerical variables.***
- Above Image shows the co-relation between ***view trailer and the view content is 0.75*** which shows that there is a ***strong co-relation b\w two variable.***
- Above Image shows there is ***weak co-relation b\w the view content and the ad impression*** which shows there is not much effect of the ad on the persons who view content.

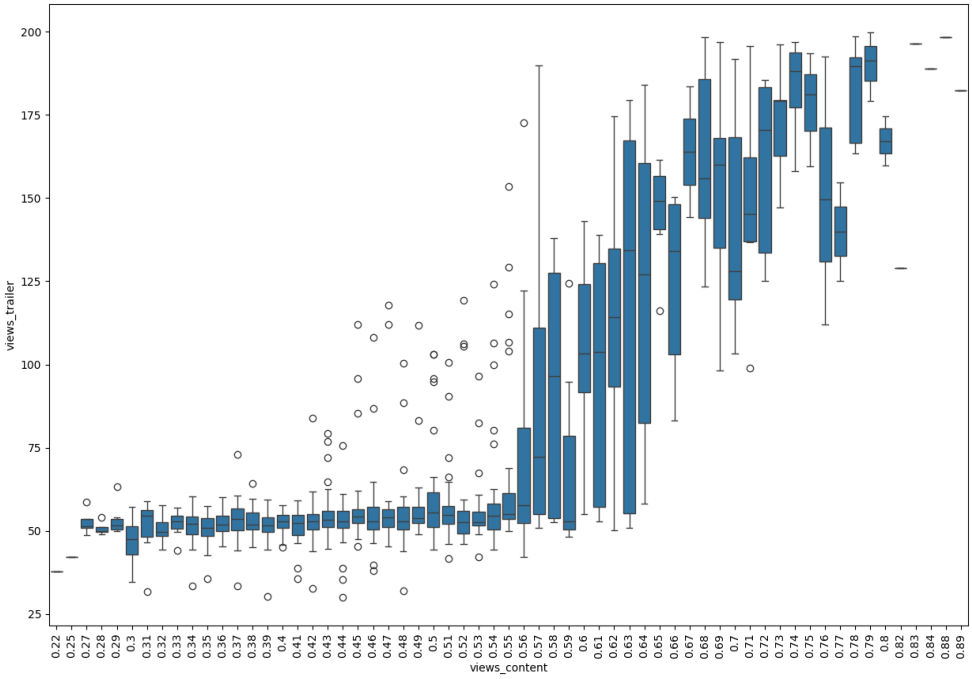


Image 13

- ***Above candlestick show the relation between view trailer and view content***

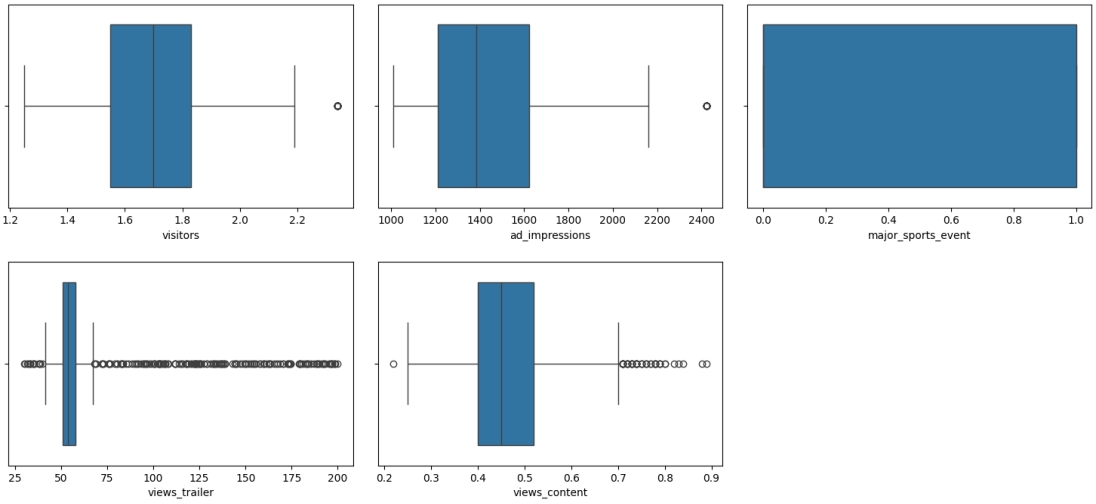


Image 14

- ***Above boxplot shows the outliers present in our data set in numerical variable.***
- As ***we don't have many outliers*** in the independent variable, as our dependent variable is Views content.
- ***We will not treat outliers*** as we have ***already found*** that there is strong ***co-relation b\w the view trailer and view content, and these are proper values***

```
   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
5      1.61         1588.38                   1  Thriller     Sunday    Fall
6      1.80         1311.96                   1    Others   Thursday    Fall

   views_trailer
0          56.70
1          52.69
2          48.74
3          49.81
4          55.83
5          49.72
6          48.15
0    0.51
1    0.32
2    0.39
3    0.44
4    0.46
Name: views_content, dtype: float64
```

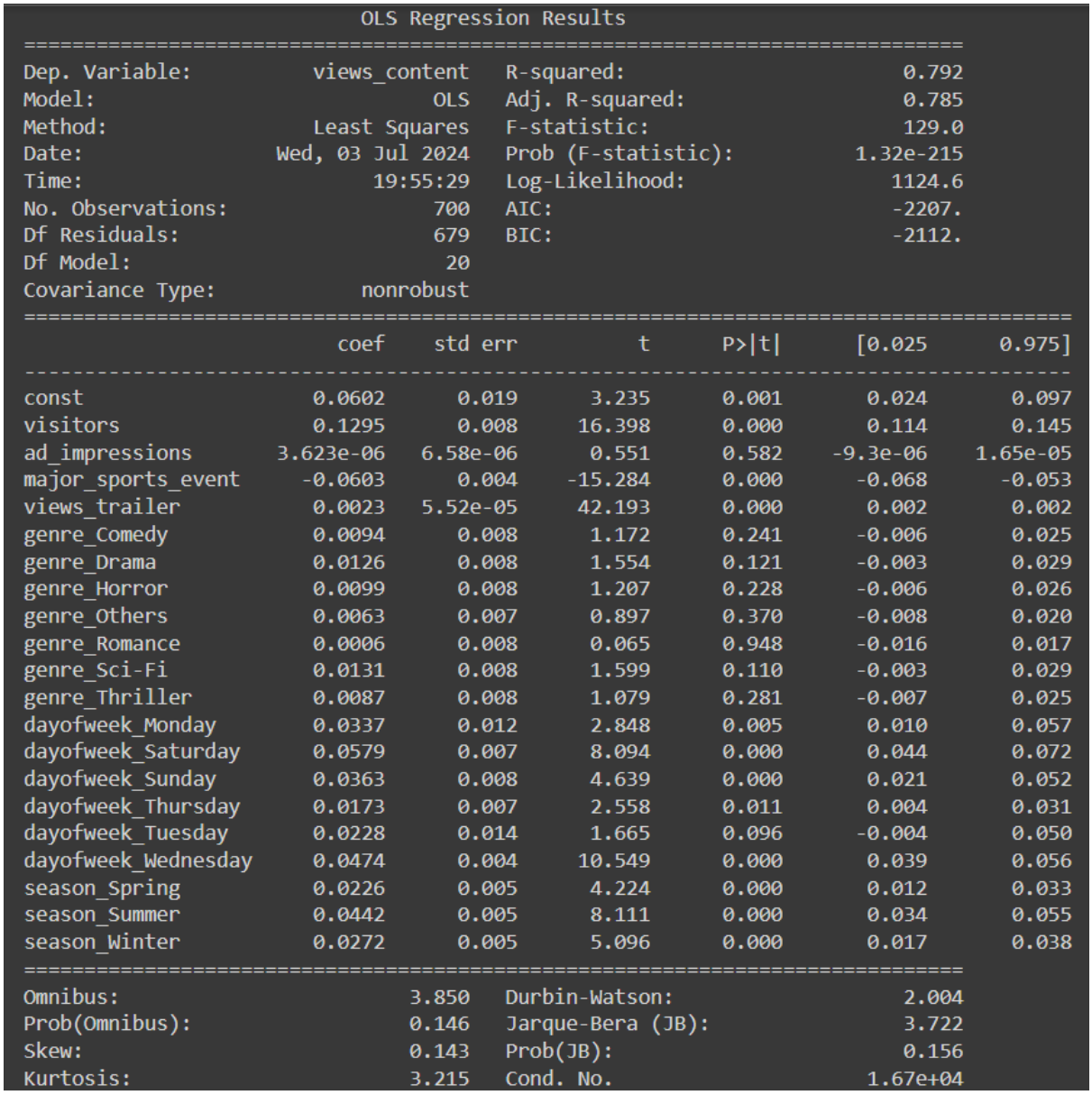
Image 15

- ***Above image shows the top 7 rows of the train data and 5 rows of test data.***

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

Image 16

- ***Above image tells that We have taken 700 rows for train data, 300 rows for test data.***



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:	Wed, 03 Jul 2024	Prob (F-statistic):	1.32e-215
Time:	19:55:29	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

Image 17

- ***Above image shows the performance of our model we have built with the data available.***

Interpreting the Regression Results:

- **Adjusted. R-squared**: It reflects the fit of the model, with the range from 0 to 1.
  - In our case ***we have 0.785 as Adjusted R-square*** which ***is good***.
- ***const* coefficient**: It is the Y-intercept.
  - In our case ***the value for constant coefficient is 0.0602.***

Model Performance Check

Training Performance

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

Test Performance

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

Image 18

- ***Above image shows the variation between training and test performance***

Observations

The training ***R^2 is 0.79***, so the model ***is not underfitting***.

The train and test data RMSE & MAE are comparable, so the model ***is not overfitting either***

MAE suggest that the model predicts the view content with in the ***mean error of 0.40*** on test data

MAPE of 9.03 shows on test data shows ***we can predict*** within ***9.03% of view content***.

Checking Linear Regression Assumptions

TEST FOR MULTICOLLINEARITY

Variance Inflation Factor (VIF)

	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

Image 19

We will have to drop all the columns which has values with VIF > 5

We don't have any columns which has high VIF, which shows in our data set we don't have the multi-collinearity in our data set.

Dealing with high p-value variables

As per the model we have built shows we have many dummy variables in data have p-value > 0.05, as they are not very significant so we will drop them

```
                            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:                Wed, 03 Jul 2024   Prob (F-statistic):          7.03e-224
Time:                        19:55:51   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.
```

Image 20

Training Performance

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

Test Performance

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

Image 21

Observations

As now we don't have any p-value which is greater then 0.05 so we will consider the features in x_train2 as final set of predictor variable and olsmod2 as final model to move forward

Now we have adjusted R^2 is 0.78 so model is able to explain 79% of variance

It shows the variables we drop is not affecting the model

RMSE & MAE are comparable for train and test data shows model is not overfitting

Now we'll check the rest of the assumptions on *olsmod2*.

TEST FOR LINEARITY AND INDEPENDENCE

	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

Image 22

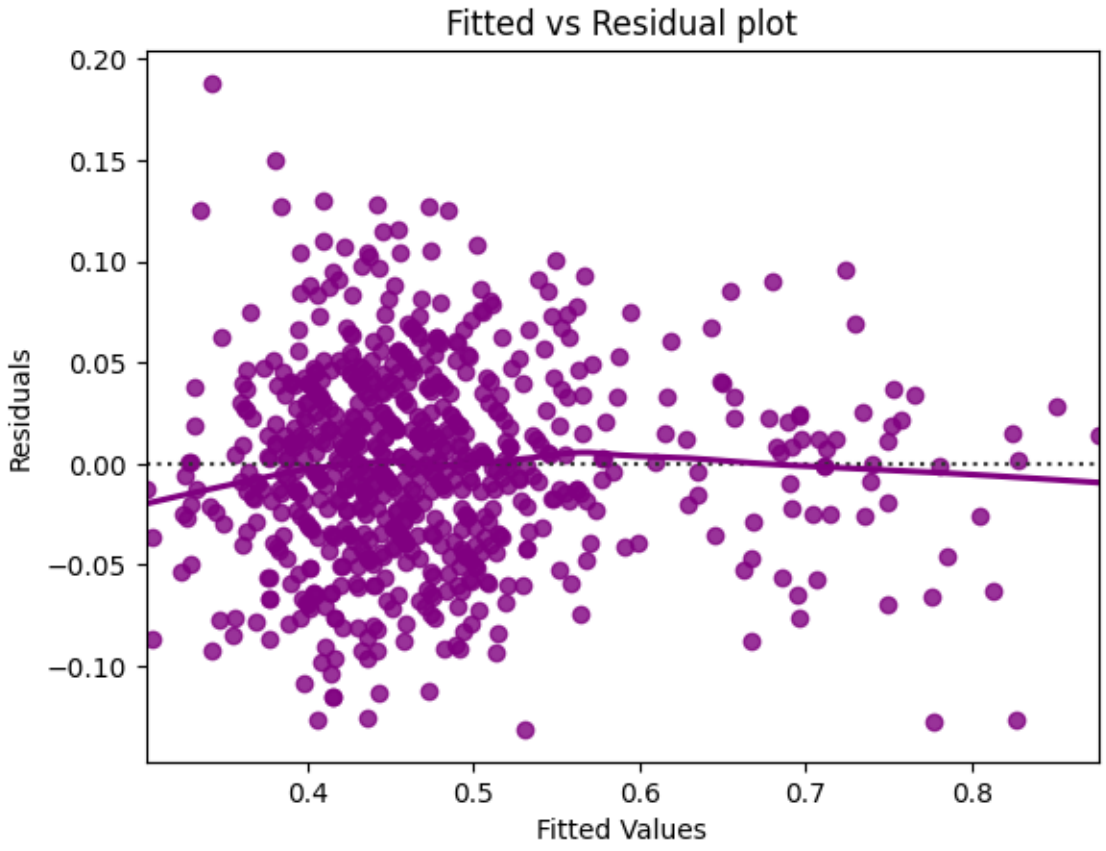


Image 23

The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values)

As we not see any pattern in the above plot shows, the assumptions of linearity and independence are satisfied.

TEST FOR NORMALITY

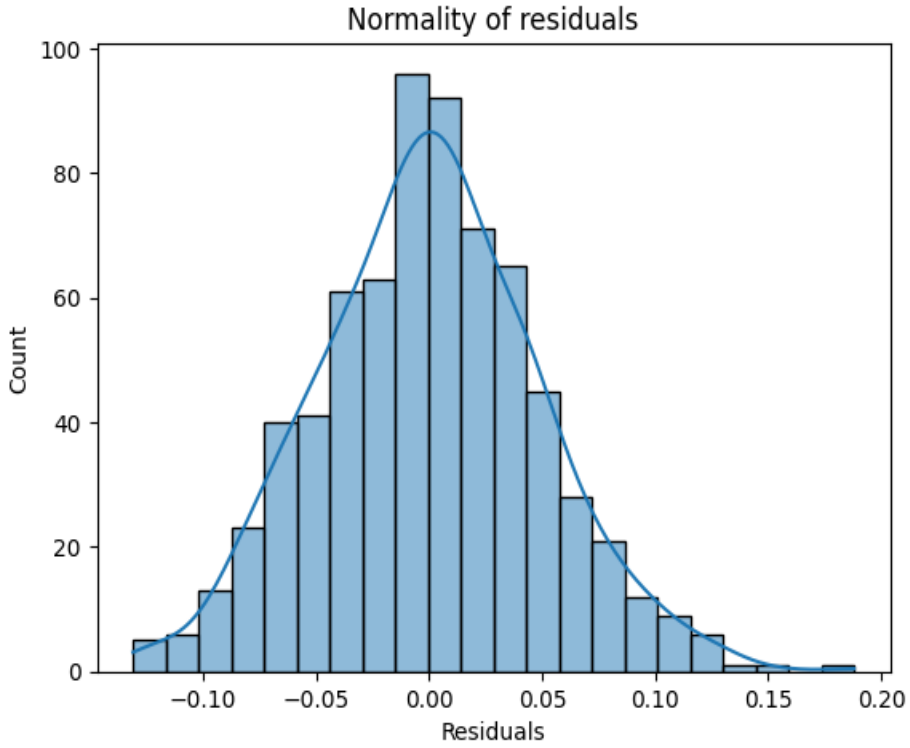


Image 24

The histogram of residuals has a bell shape

Let's check Q-Q plot.

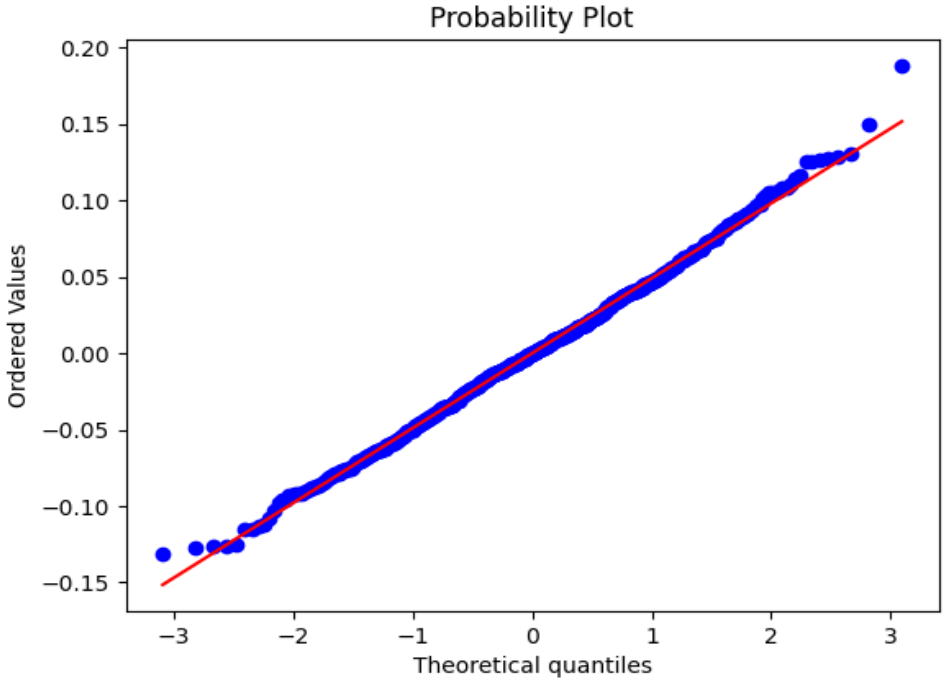


Image 25

The residuals more or less follow straight line except the tails

Lest check the results with the Shapiro test

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

Image 26

As, the **p-value > 0.05**, so the residuals are normal as per Shapiro-Wilk test

Residuals are about normal.

Assumption is satisfied.

TEST FOR HOMOSCEDASTICITY

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

Image 27

As the p-value > 0.05, we can say residuals are homoscedasticity. So, assumption is satisfied.

Predictions on test data

	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

Image 28

- We found that the model has returned good result, as the actual and predicted values are comparable.

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:                Wed, 03 Jul 2024   Prob (F-statistic):          7.03e-224
Time:                        19:56:30   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.
```

Image 29

Training Performance

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

Test Performance

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

Image 30

- The model is able to explain ~78% of variation in data

- The train and the test MAE & RMSE are low and comparable. So, **our model is not either overfitting and underfitting.**
- The ***MAPE*** of test data shows we can predict the ***9.18% of content view in first day***.
- So, our final model is ***Olsmodel_final*** is good for prediction and for inference purposes

Conclusions and Recommendations

- The model is able to explain ~ 78% of the variation in data set with 9.1% of the view content in test data

- Model shows that if content release on weekends, then it do less business then if content released on Friday.so it is better to release content on Friday or Wednesday.
- Ad content doesn't affect viewership to much so don't spend too much on ad content
- Trailer viewership highly affect the content viewership so make trailer reach as high as possible by finding new way.
- There is approximately equal number of audience for different genre of films.so one can chose to make film on any genre content risk is very low to flop if content is good