Predictive Modelling Coded PROJECT

ShowTime OTT Data Analysis Business Report

Submitted by: Yash Juneja

Submitted to: Great Learning

Table of Contents

Context……………………………………………………………………….……3
Objective……………………………………………………………………….…….3
Data Description…………………………………………………………………...…3
Data Dictionary……………………………………………………………………….3-4
Understanding the Structure of data…………………………………………………4
Exploratory Data Analysis……………………………………………………...……5

- Distribution of Contents…………………………………………………..…………5
- Distribution of Genres………………………………………………….……………5
- Viewership by Day of Release…………………………………….…………………5
- Viewership by Season of Release……………………………………………………7
- Correlation between Trailer Views and Content Views…………………….…………7

Univariate Analysis…………………………………………………………..….…….8-11
Bivariate Analysis………………………………………………………………………12-20
Data PreProcessing…………………………………………………………..…………20

- Duplicate Value check…………………………………………………………..……20
- Missing Values check……………………………………………………………..…20-21
- Outlier treatment (Boxplot for Content Views)……………………………….………21
- Feature Engineering……………………………………………………………...…21-22
- Data Preparation for modeling…………………………………………………...…22

Model building – Linear Regression………………………………………….……………22

- Splitting the data into Training and Test sets………………………………………..22
- Checking the shape of Training and Test sets………………………………………22
- Model coefficients and statistics……………………………………………………..22-23

Testing the assumptions of linear regression model……………….…………………23

- Test for Linearity and Independence…………………………………………………23
- Test for Normality………………………………………………………………….……24-25
- Test for Homoscedasticity……………………………………………..………………25

Model Performance Evaluation……………………………………………………………26-27
Actionable Insights & Recommendations……………………………………………..27-29

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 behavior, 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 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 spend, content timing clashes, weekends and holidays, etc. They have hired you as a Data Scientist, shared the data of the current content in their platform, and asked you to analyze the data and come up with a linear regression model to determine the driving factors for first-day viewership.

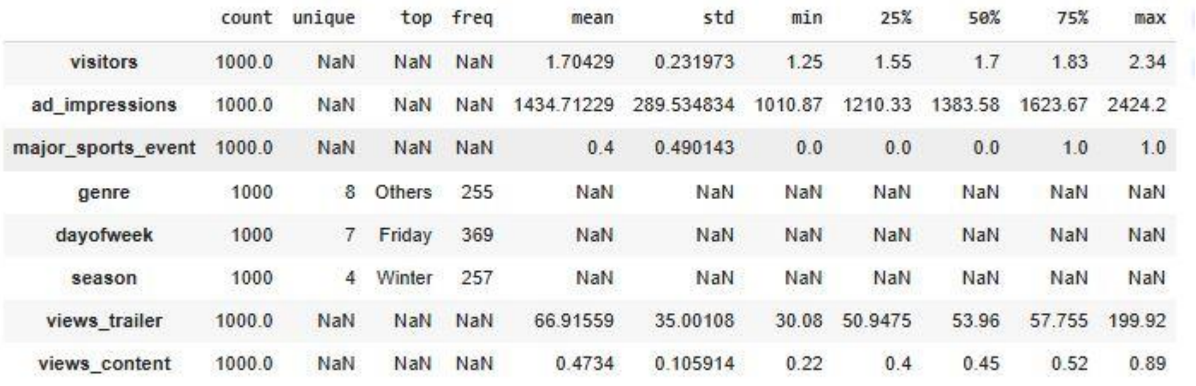
Data Description

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

Data Dictionary:

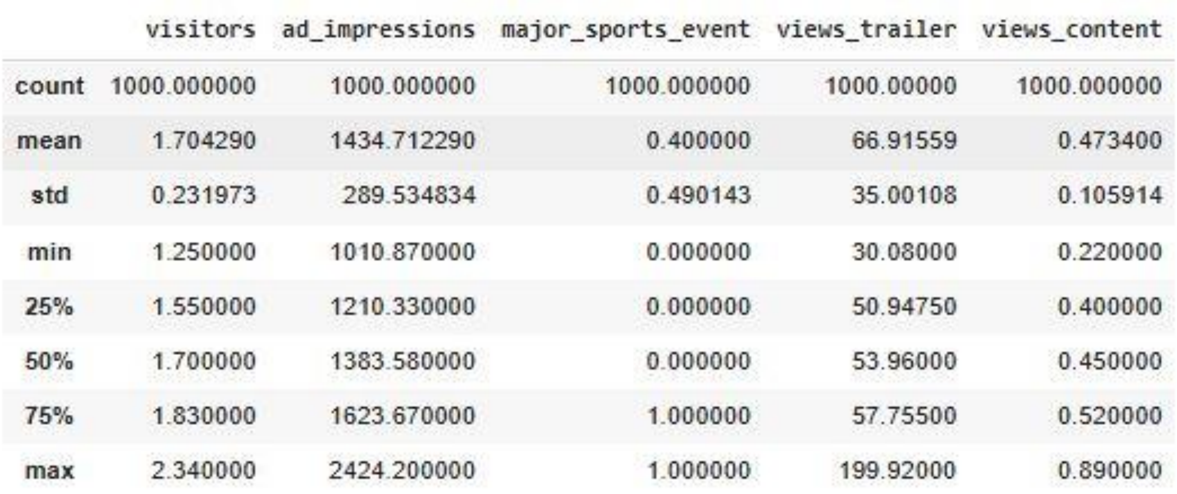
- visitors: Average number of visitors, in millions, to the platform in the past week
- ad_impressions: Number of ad impressions, in millions, across all ad campaigns for the content (running and completed)
- major_sports_event: Any major sports event on the day
- genre: Genre of the content
- dayofweek: 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

Understanding the Structure of 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 : 1



	visitors	ad_impressions	major_sports_event	views_trailer	views_content
count	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000
mean	1.704290	1434.712290	0.400000	66.91559	0.473400
std	0.231973	289.534834	0.490143	35.00108	0.105914
min	1.250000	1010.870000	0.000000	30.08000	0.220000
25%	1.550000	1210.330000	0.000000	50.94750	0.400000
50%	1.700000	1383.580000	0.000000	53.96000	0.450000
75%	1.830000	1623.670000	1.000000	57.75500	0.520000
max	2.340000	2424.200000	1.000000	199.92000	0.890000

Fig : 2

The Provided data structure is a dataset related to an OTT (Over The Top) media service, which is a streaming service that offers online content to users. The dataset contains 5 rows, each representing a single observation or record and a 8 columns, each representing a feature or variable.

• The Dataset has been loaded properly.

- Dataset consists of several columns displaying the various attributes related to an OTT (Over-the-top) media service, which is a streaming service that offers online content to users.

visitors: This column represents the number of visitors to the OTT platform.

major_sports_event: This column is a binary indicator (0 or 1) representing whether if a major sports event occurred on the corresponding day. It is a relevant information for analysing the impact of sports events on user engagement.

views_trailer: This column represents the number of views for trailers on the platform.

The day of the week and season might have an impact on user engagement.

The Trailer views are relatively high as compared to content views, which might indicate that users are more interested in previewing any content before watching it.

Exploratory Data Analysis

Problem definition, questions to be answered - Data background and contents - Univariate analysis - Bivariate analysis - Answers to the key questions provided - Insights based on EDA

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

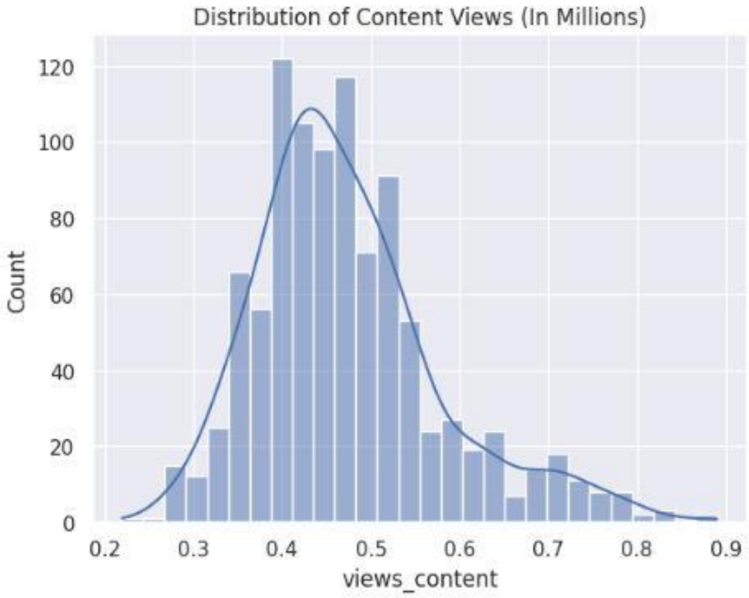


Fig : Distribution of Contents

This Graph shows a Right Skewed Distribution of Content Views, With Most views Count is around 0.4 – 0.5 Millions

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

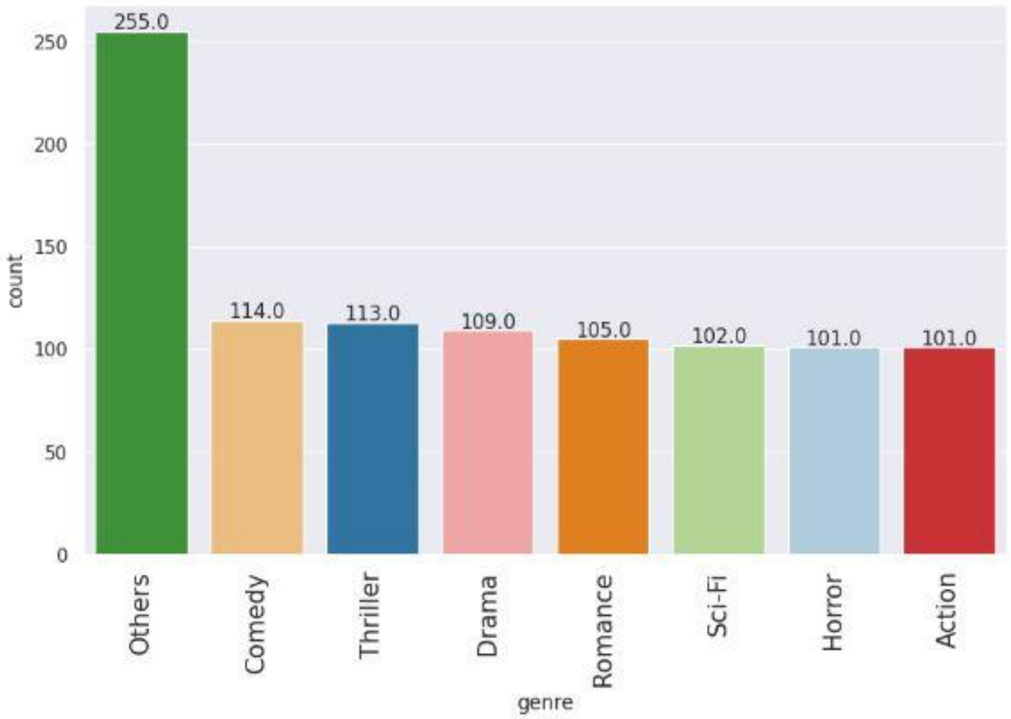


Fig : Distribution of Genres

As per graph, other Category is the most viewed and Comedy is the most popular genre, followed by Thriller and Drama.

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?

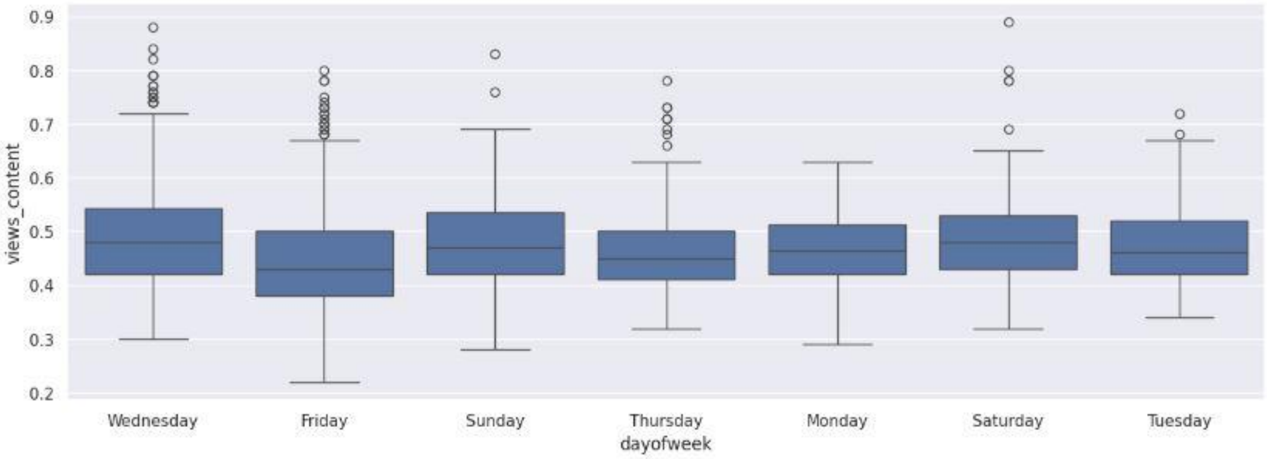


Fig : Viewers by Day of Release

This Boxplot shows that viewers are highest on Wednesday and Saturday, Lowest on Friday.

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

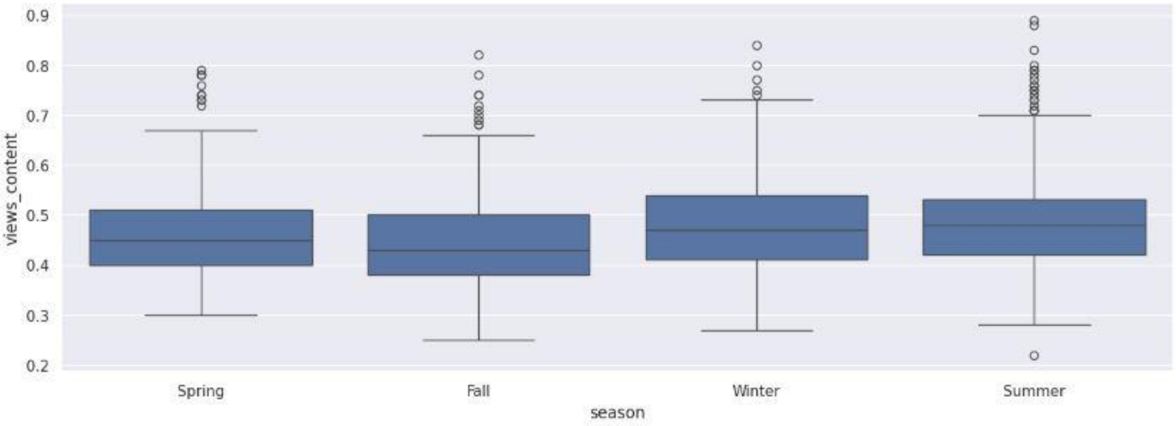


Fig : Viewership by Season of release

This Boxplot shows that viewership is highest in summers, Lowest in Fall Season

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

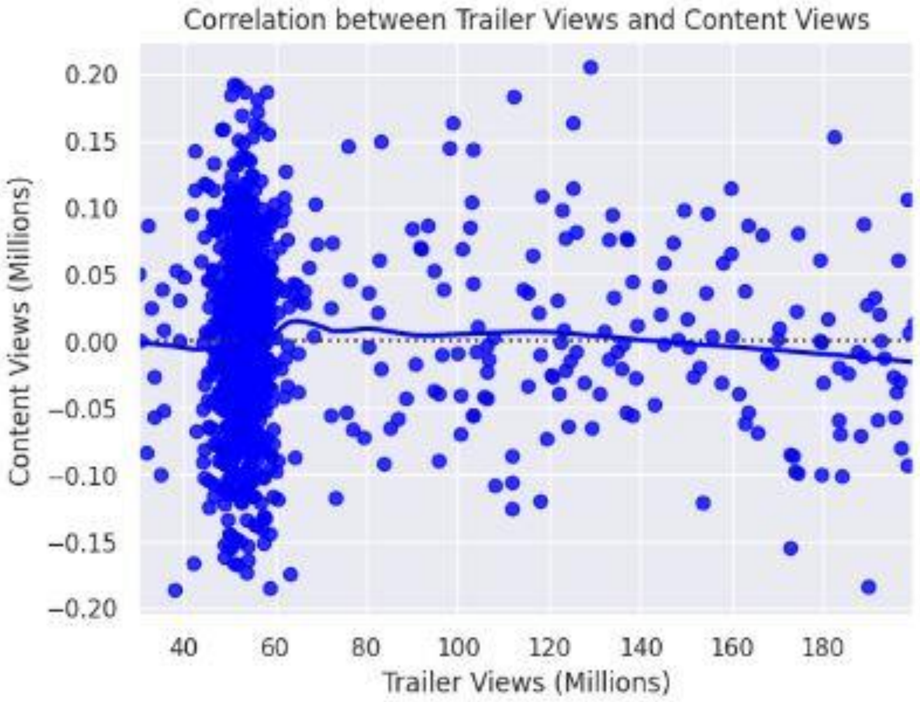


Fig : Correlation between Trailer Views and Content Views

This Correlation in figure of Scatter Plot shows Not a Strong Positive Relation Between Trailer Views and Content Views, It’s a Normal Correlation or somewhere we can say Negative relation.

Univariate analysis

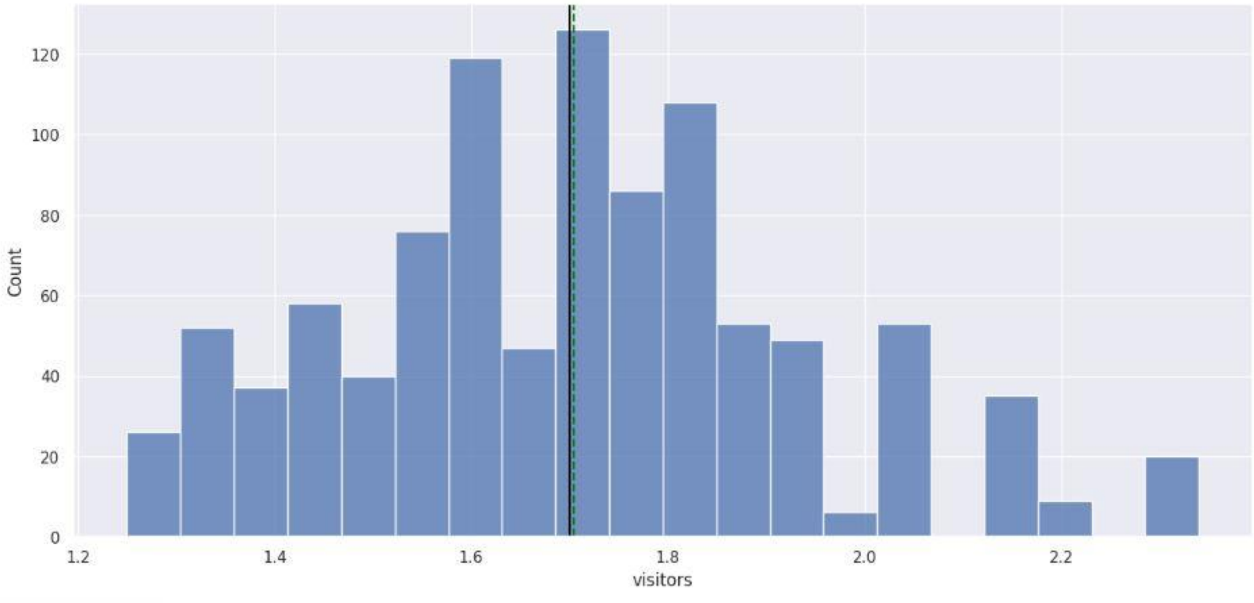


Fig 1: Visitors Vs Counts

Figure 1: shows that Average number of visitors are around 1.7 millions on platform in the past week.

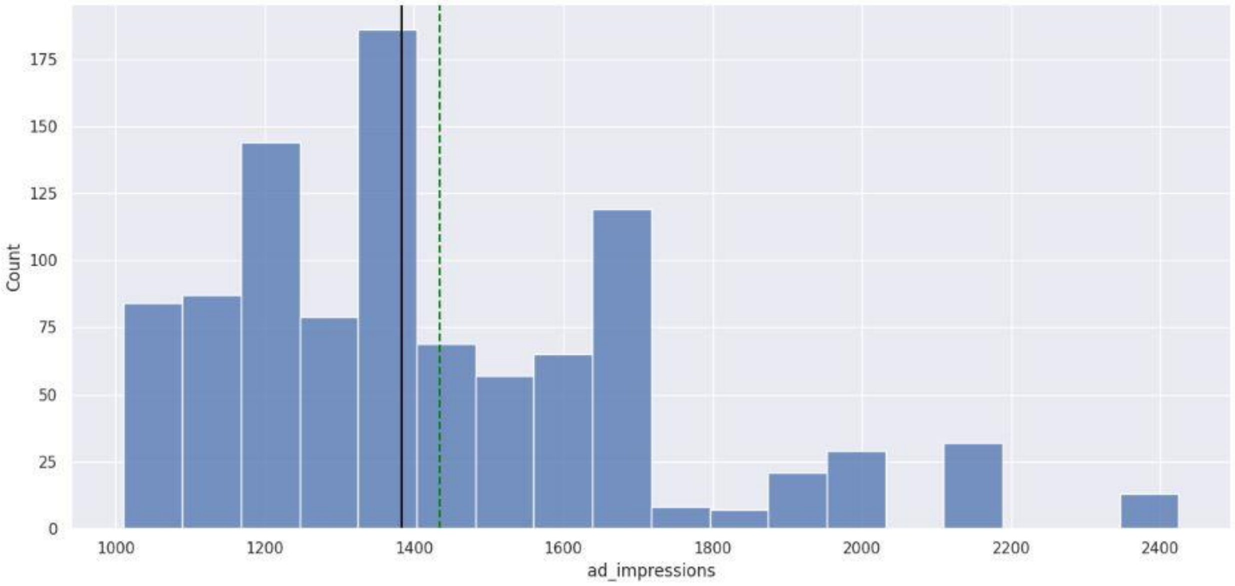


Fig 2: Ad Impressions Vs Counts

Figure 2: shows most of the counts come on 1400 millions Ad impressions, which is still not good Count, Need to focus on targeted customer properly for Good and Better counts on Particular Ad impressions.

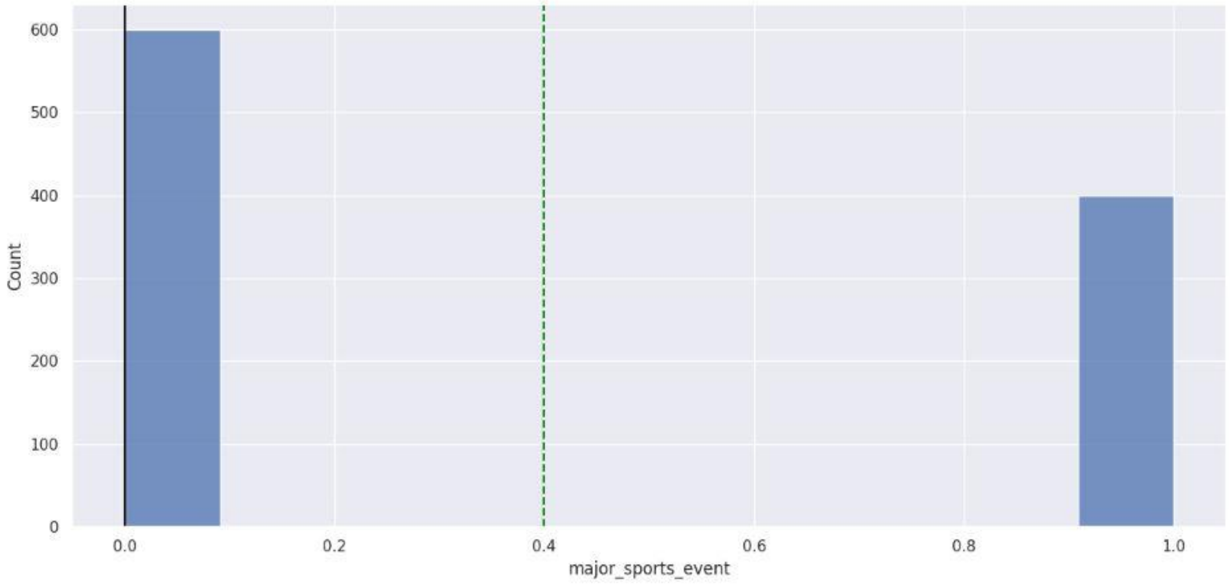


Fig 3: Major Sports Events Vs Counts

Figure 3: Shows most of the counts and Visitors come when there is no sports events, so it is clear that, Teaser/Trailer and Series/Content should be uploaded or release on that days where no sports event are taking place, Because Sports events are effecting our visitors and Viewers.

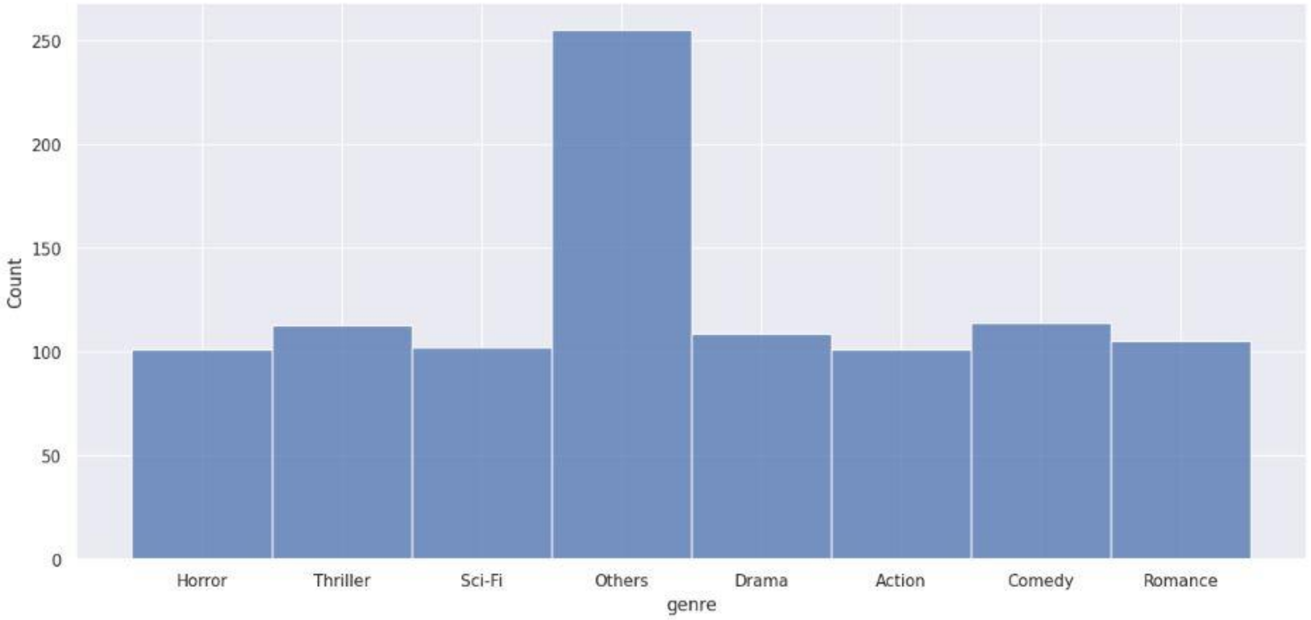


Fig 4: Genre Vs Counts

Figure 4: Shows, Apart from other categories, Comedy and Thriller are most popular and Viewed categories, So it is clear that, should focus on those categories and promote as per strategies.

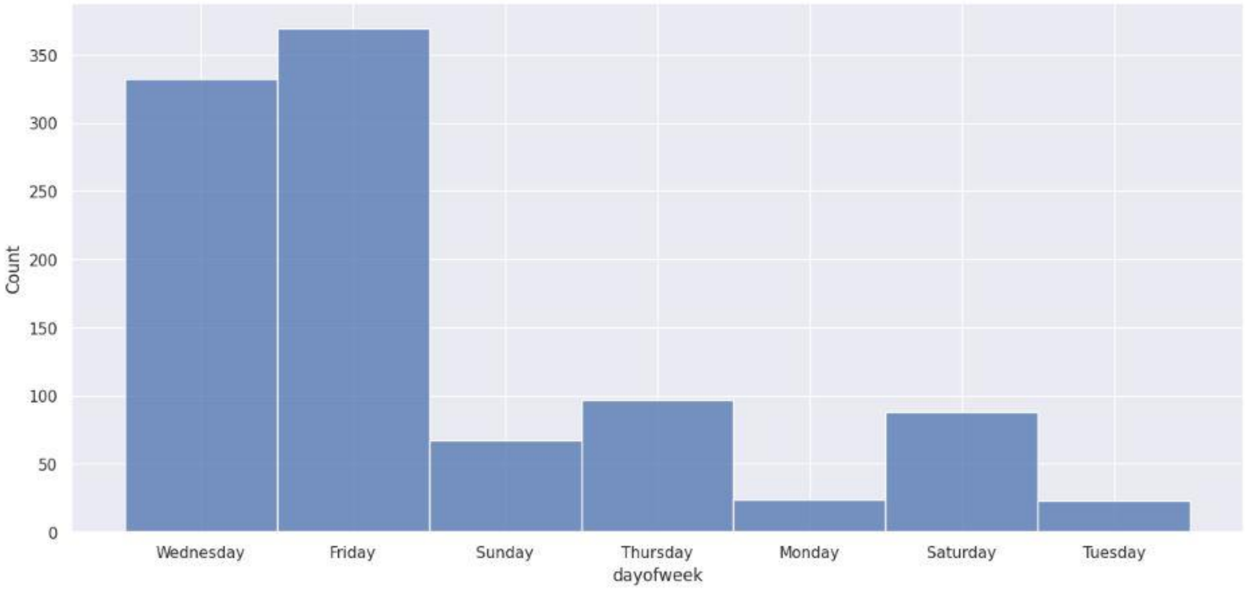


Fig 5: Days of Week Vs Counts

Figure 5: States that, most Viewers are Preferring Wednesday and Friday for Watching content, due to many other reasons, So when ad campaigns are set, at that time we need to set campaign and boost our organic and inorganic strategies both on Wednesday and Friday, more

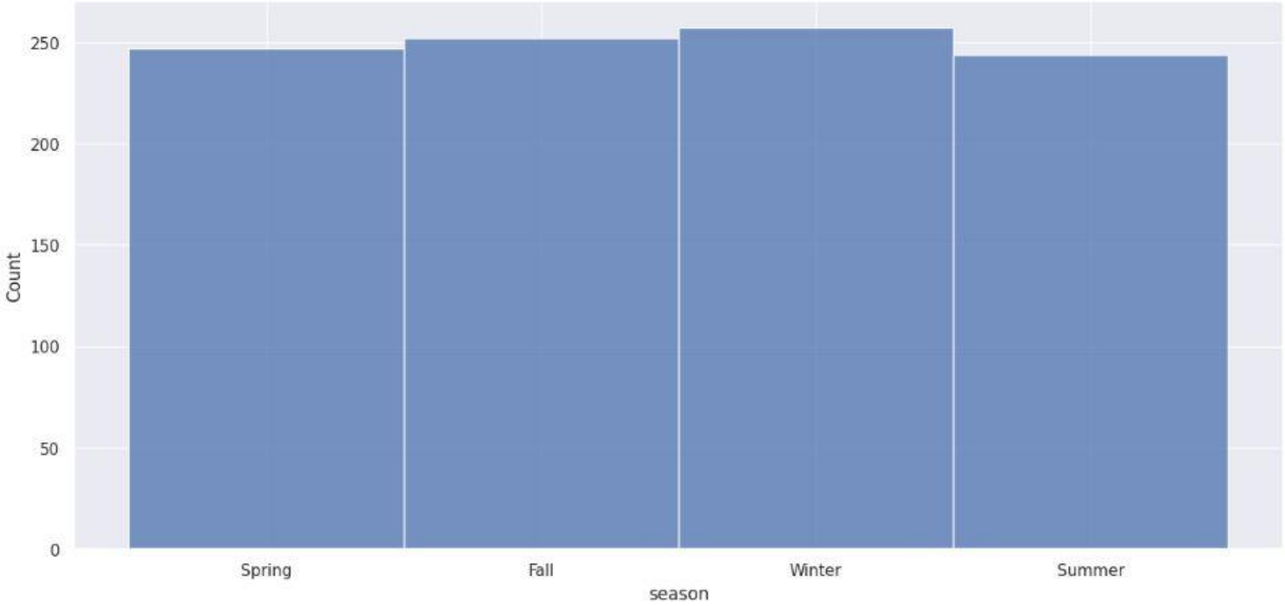


Fig 6: Season Vs Counts

Figure 6: States that, most content views are in winters, but if we see properly, there is not major difference, so focus should be on all seasons, all the year equally.

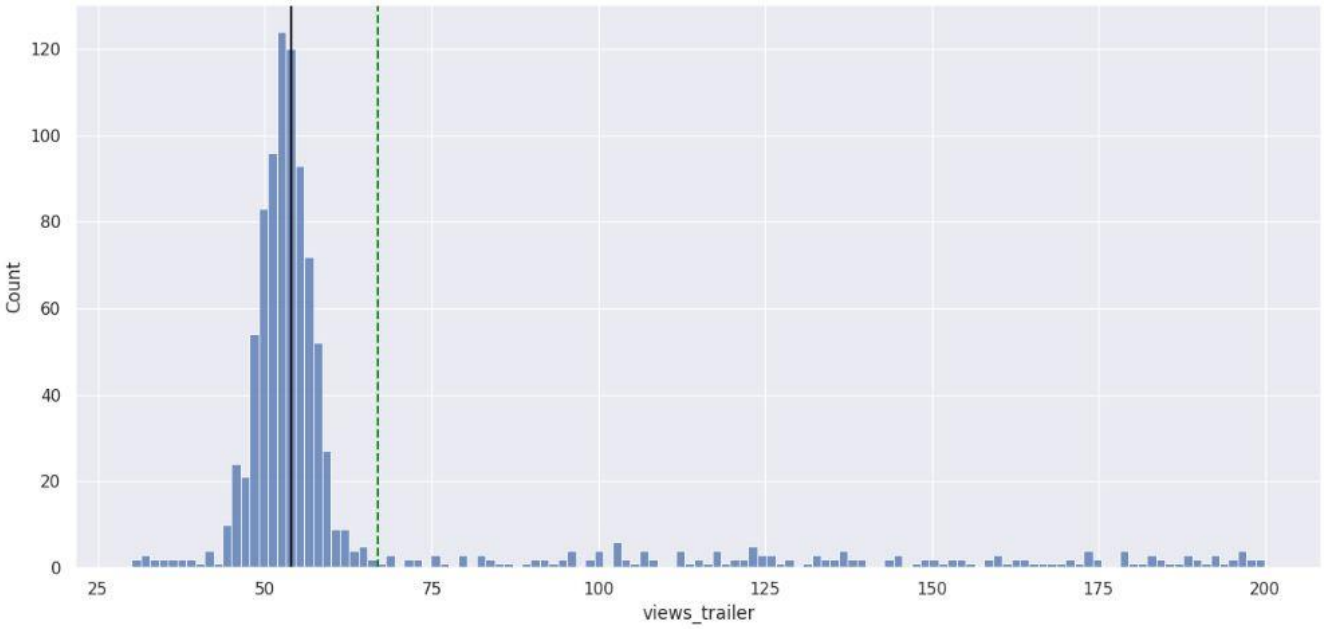


Fig 7: Trailer Views Vs Counts

Figure 7: States that, most trailer views are around 50-55 Millions, which is decent, We can target more customers,

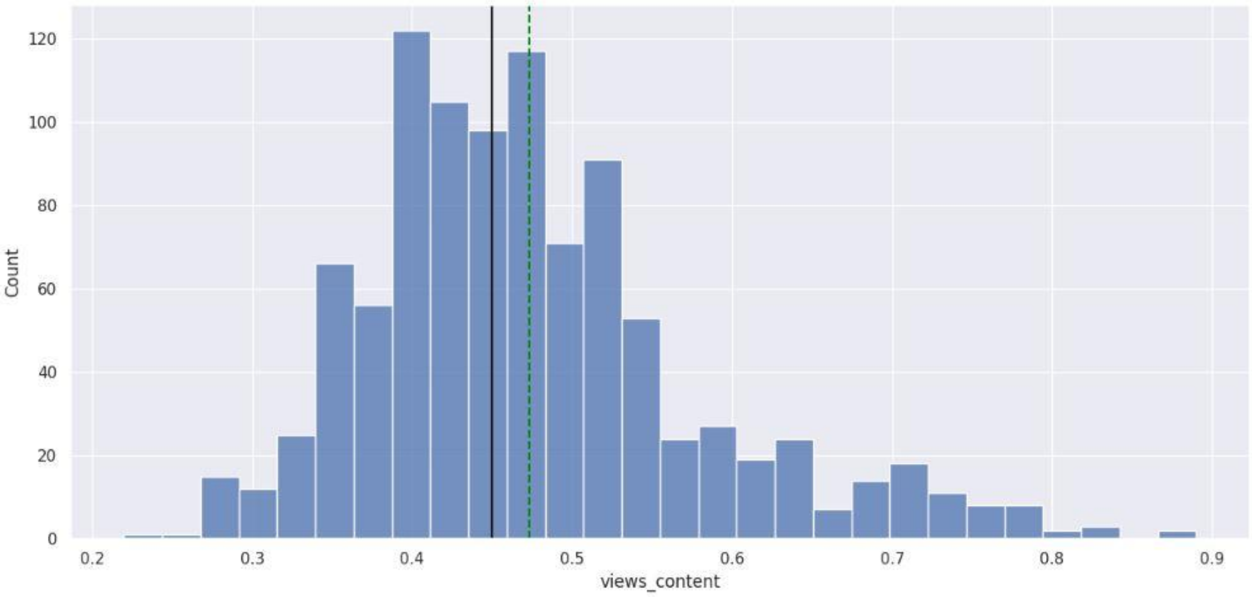


Fig 8:Content Views Vs Counts

Figure 8: States that, Most of the Content Views are between 0.4-0.5 Millions, As stated in Figure 7: If we compare, according to Trailer Views and Content views, there is much difference, We need to focus on more Quality Content, It seems like

People are watching trailer, but they are not liking the content, that’s why content views are not upto the mark which could be as compared to Trailer views.

Bivariate analysis

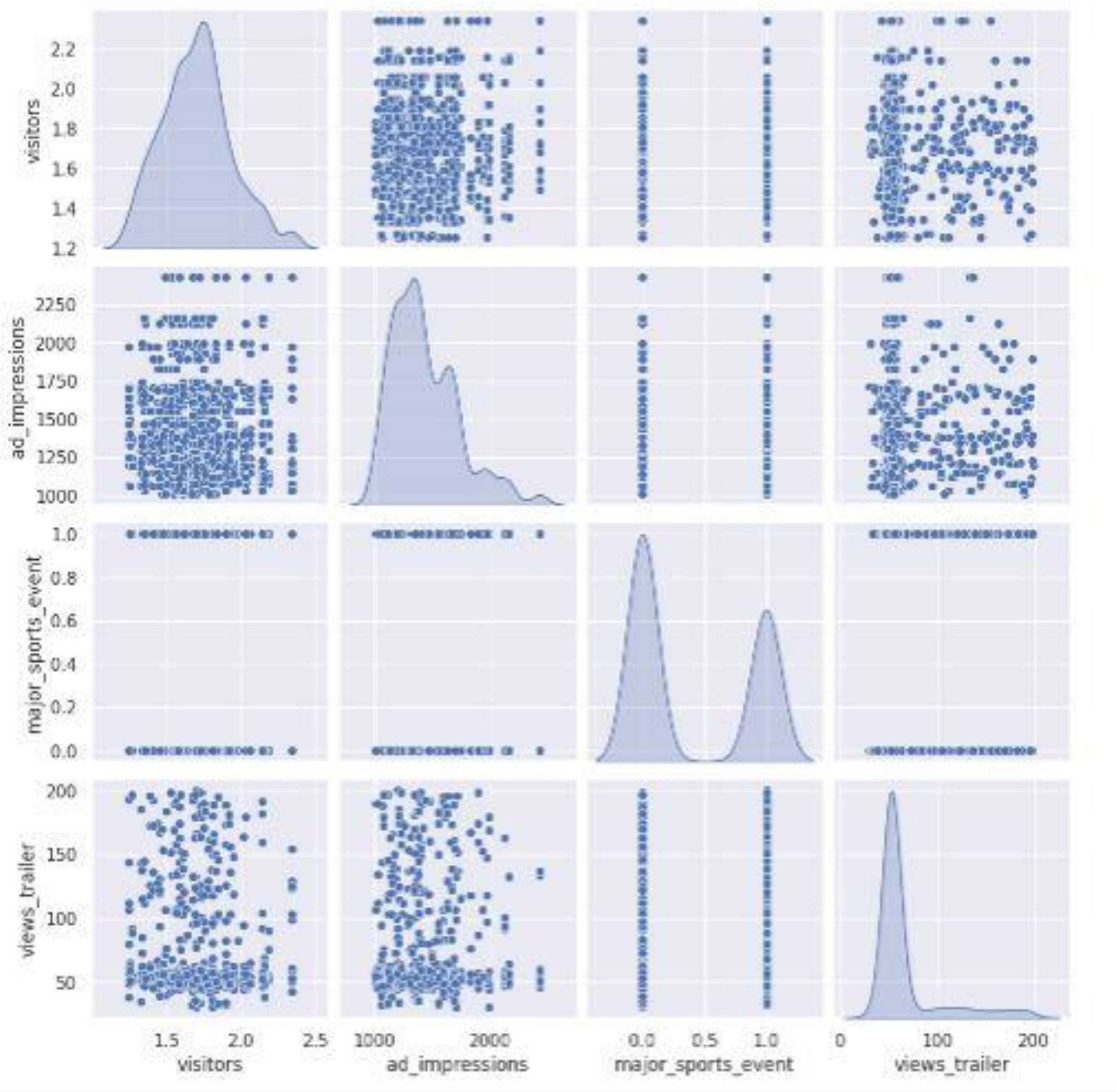


Fig 1 : All in one Relation Graph

Fig 1: Shows Visitors are having a Normal Relation with respect to Trailer views, Ad Impressions, We can’t say positive because sometimes it goes up and sometimes down, But it’s giving negative impact when there is any major sports event.

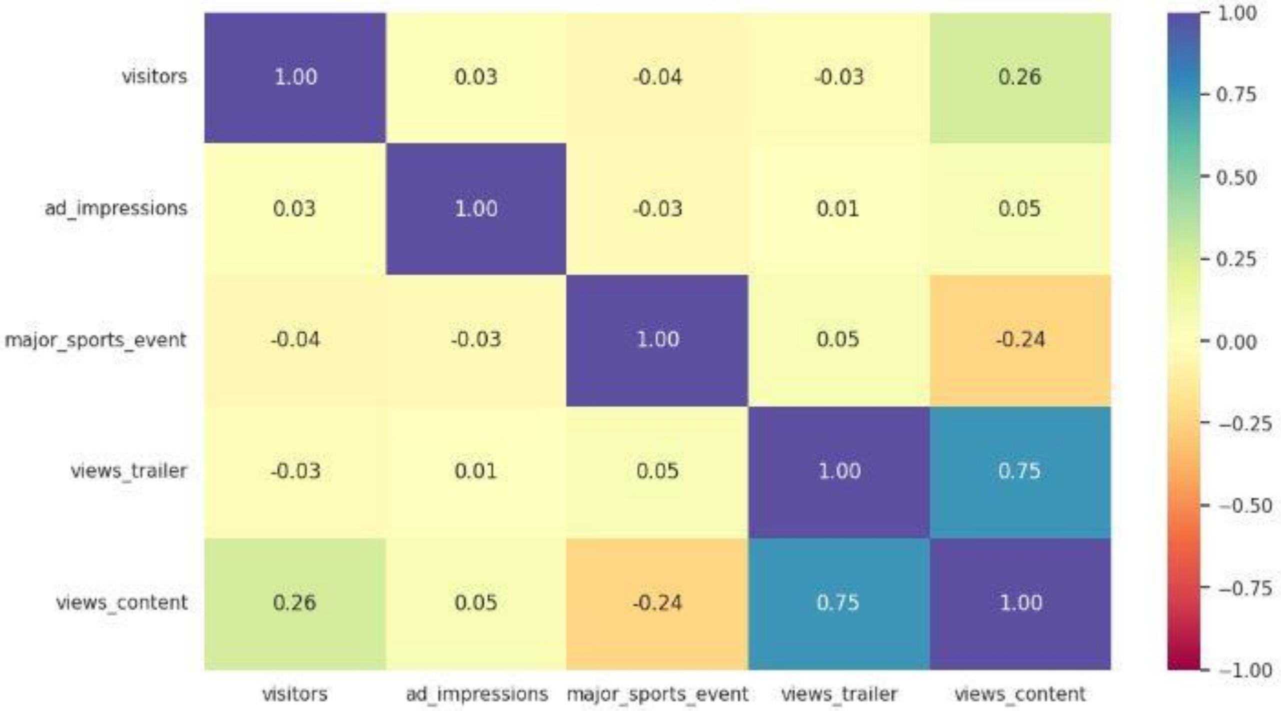


Fig 2 : All in one Relation Graph

Fig 2: Shows that

- Visitors are having positive relation with Ad impressions and Content Views, but Negative relation with Trailer Views and Major Sports event, So Trailer should be promoted well by looking for the interest genre of the audience.
- Ad Impressions Looks Positive for Trailer Views and Content views.
- Major sports event anyway giving negative impact to content, but positive impact to Trailer, Because Trailer is hardly 2-3 Minutes, anyone can watch and make time to that, but content length is more, So Viewers are not preferring to watch content in those days, when there is any major sports event happening
- Trailer and Content should be release only when there is no major sports event, which will increase visibility, visitors, Trailer views and Content views also.

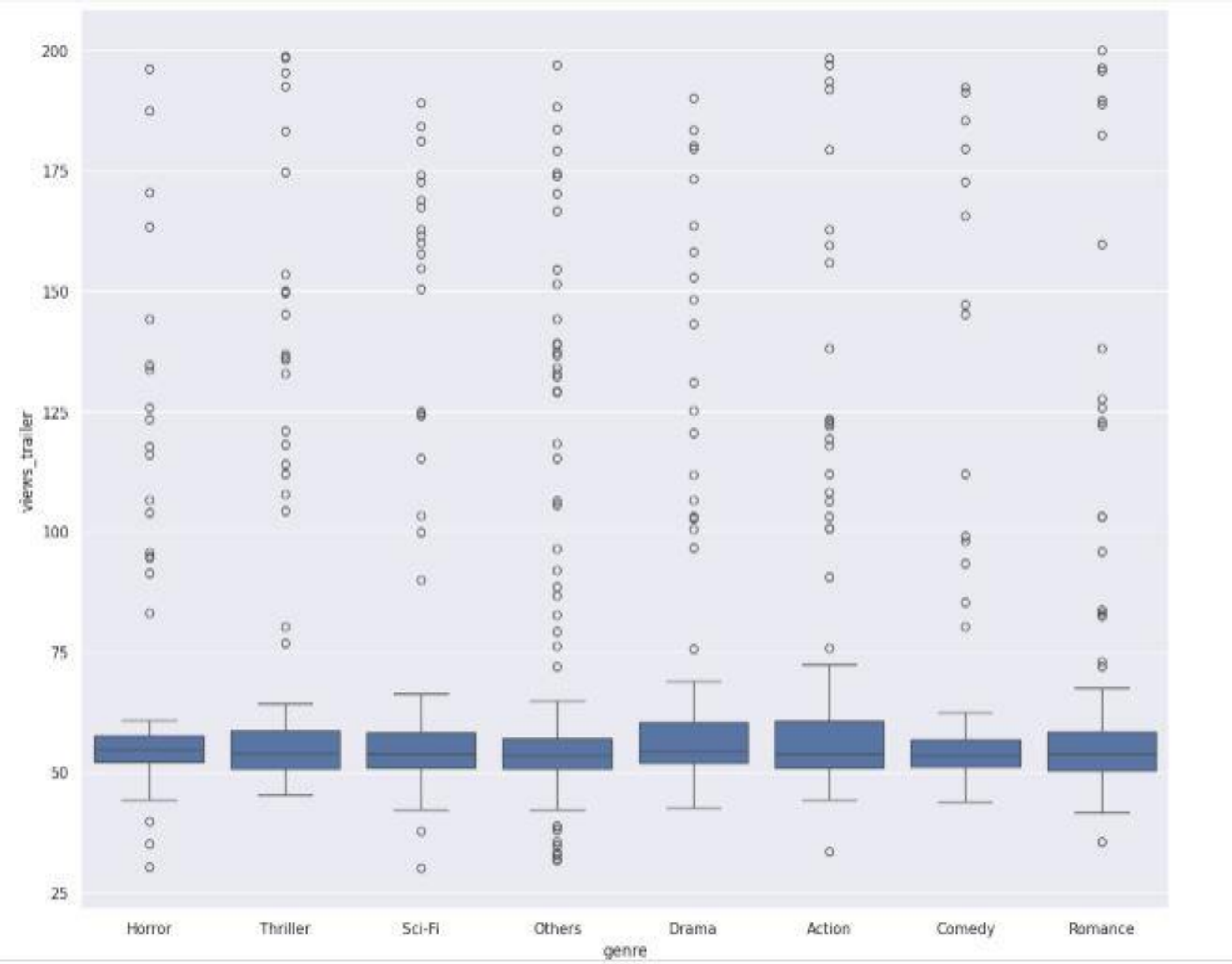


Fig 3 : Genre Vs Trailer Views

Fig 3: Refers that Trailer of Shows Horror and Drama are most Viewed as compared to others.

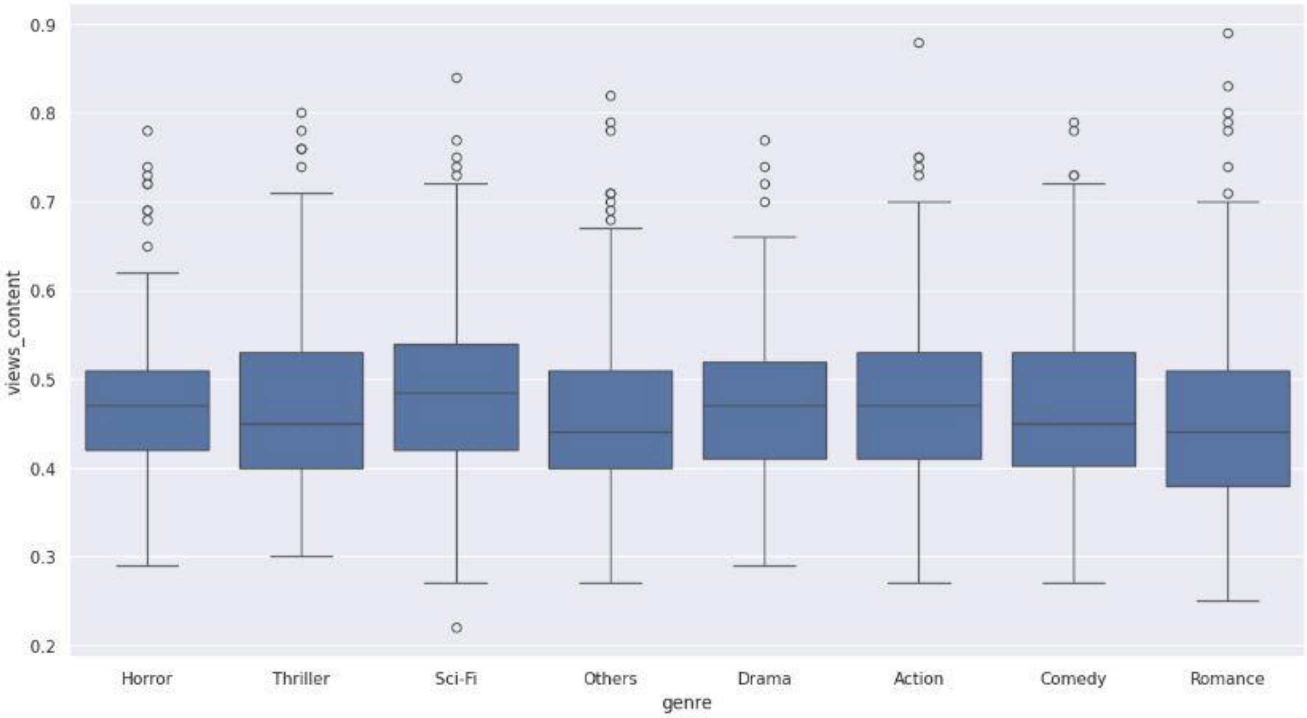


Fig 4 : Genre Vs Content Views

Fig 4: Refers that Sci-Fi is the most Viewed content followed by Horror, Drama and Action.

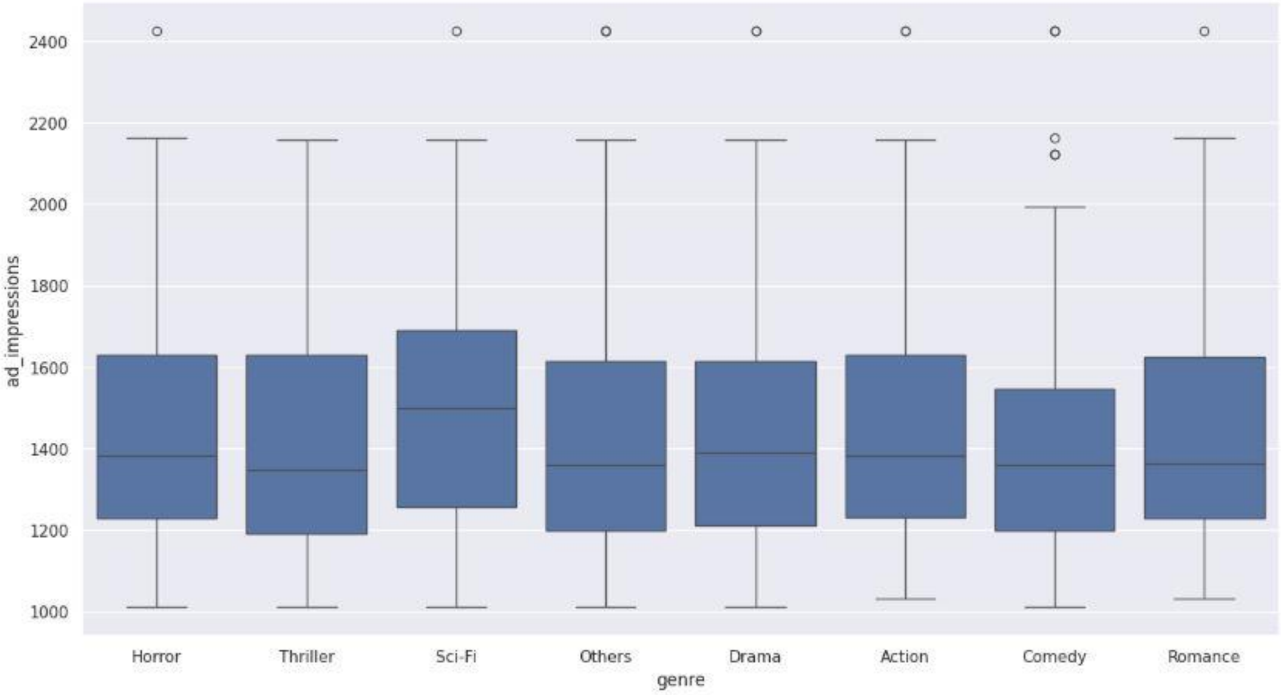


Fig 5 : Genre Vs Ad Impressions

Fig 5: Shows Sci-Fi Category Content has most Impressions as compared to others, May be due to owner is targeting and running more Ad campaigns on Sci-Fi category Content, or may be there is any other more reasons.

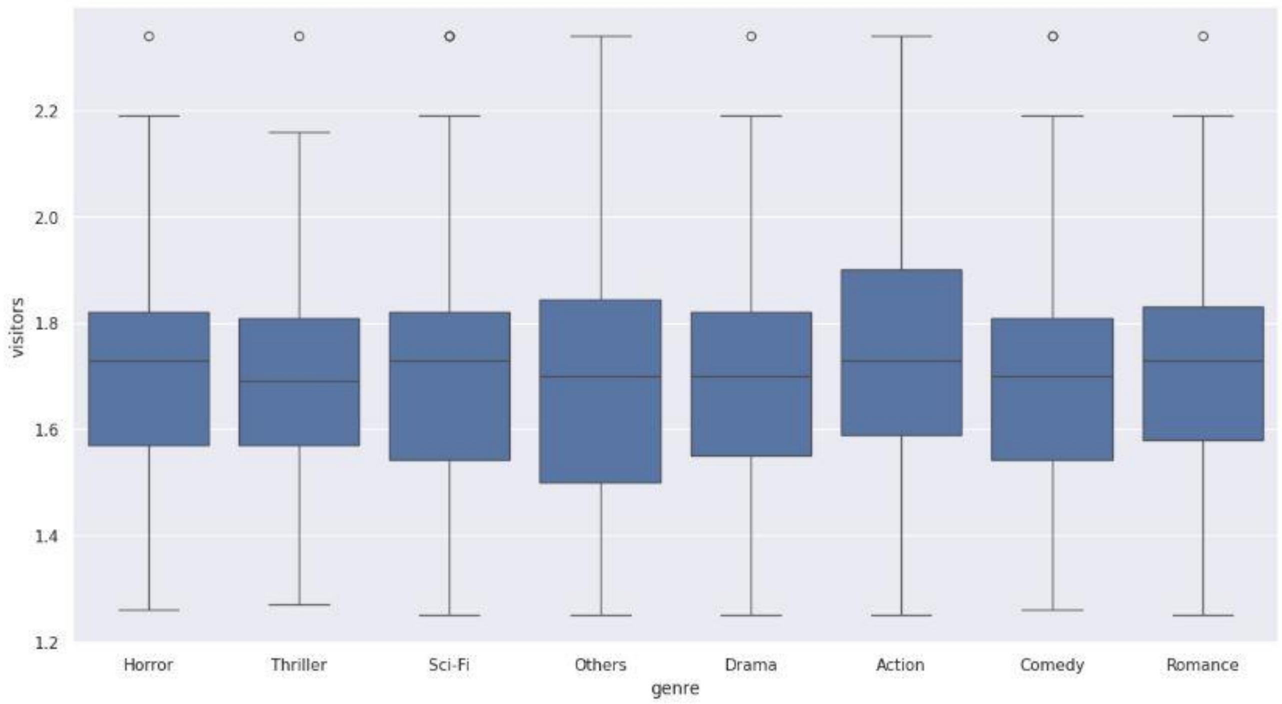


Fig 6 : Genre Vs Visitors

Fig 6: Shows more visitors are Clicking and Preferring to watch Sci-fi, Horror and Action Content.

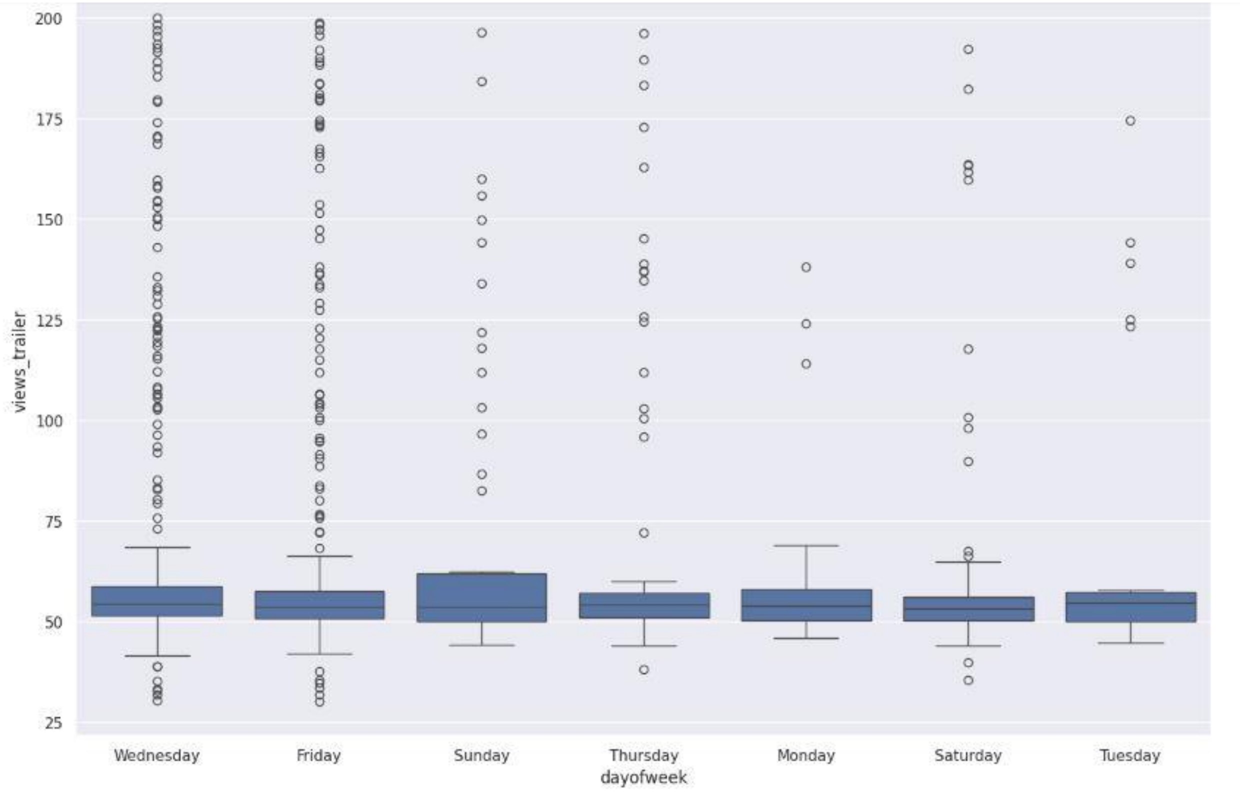


Fig 7 : Day of Week Vs Trailer Views

Fig 7: Shows Trailer Views are almost equal on all 7 days, Very few dip on Friday, Saturday and Sundays, But It Doesn't mean nothing, May be visitors are more preferring to watch content on Weekends, or may be hangout somewhere outside, Because Weekends are off for Salaried/Job Person

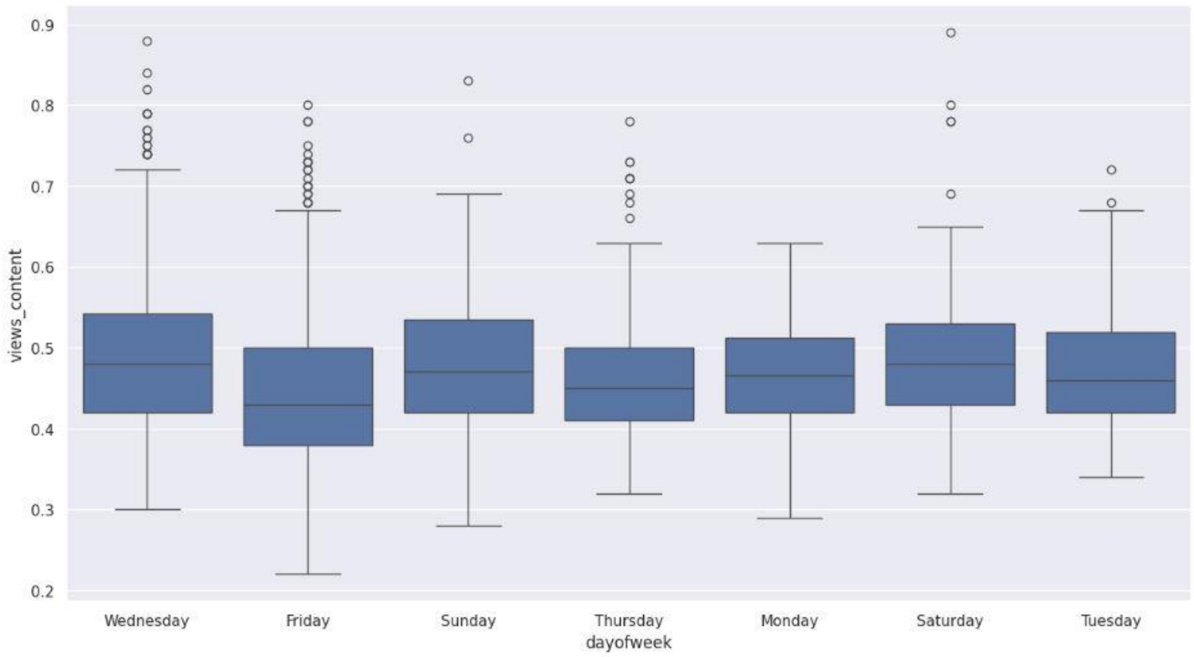


Fig 8 : Day of Week Vs Content Views

Fig 8: Shows Content Views are more on Wednesday Saturday and Sunday.

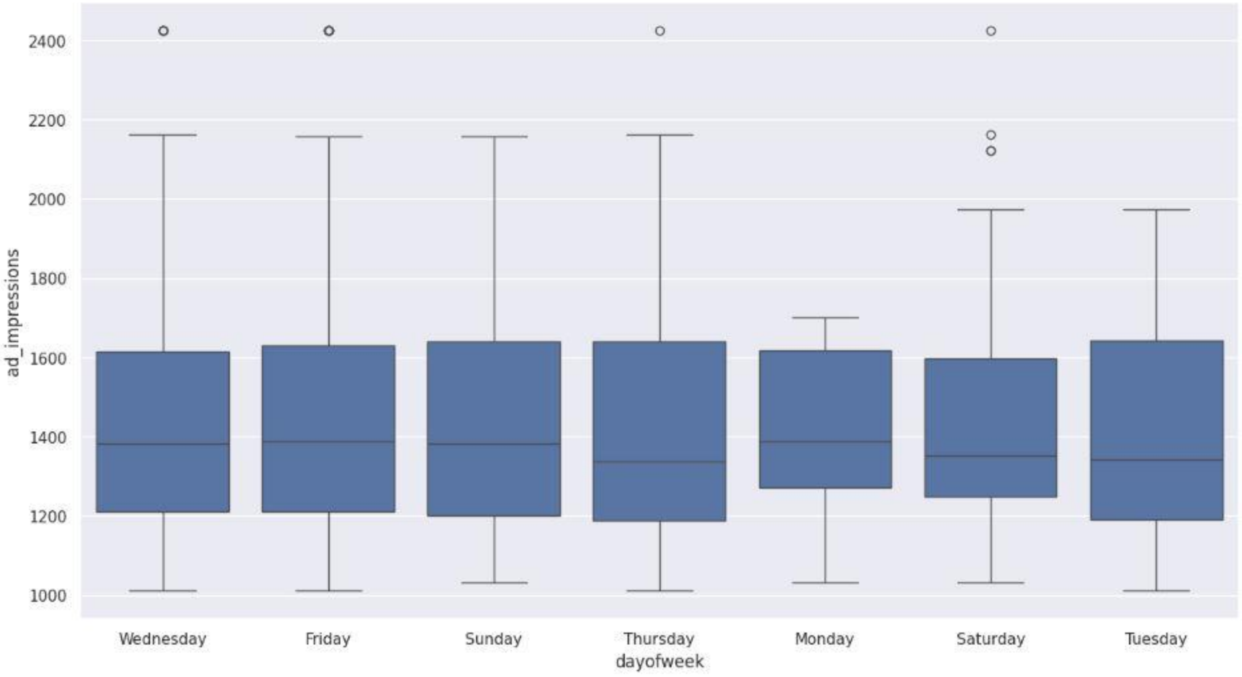


Fig 9 : Day of Week Vs Ad Impressions

Fig 9: Shows Almost equal Ad impressions on Monday, Wednesday, Friday and Sunday.

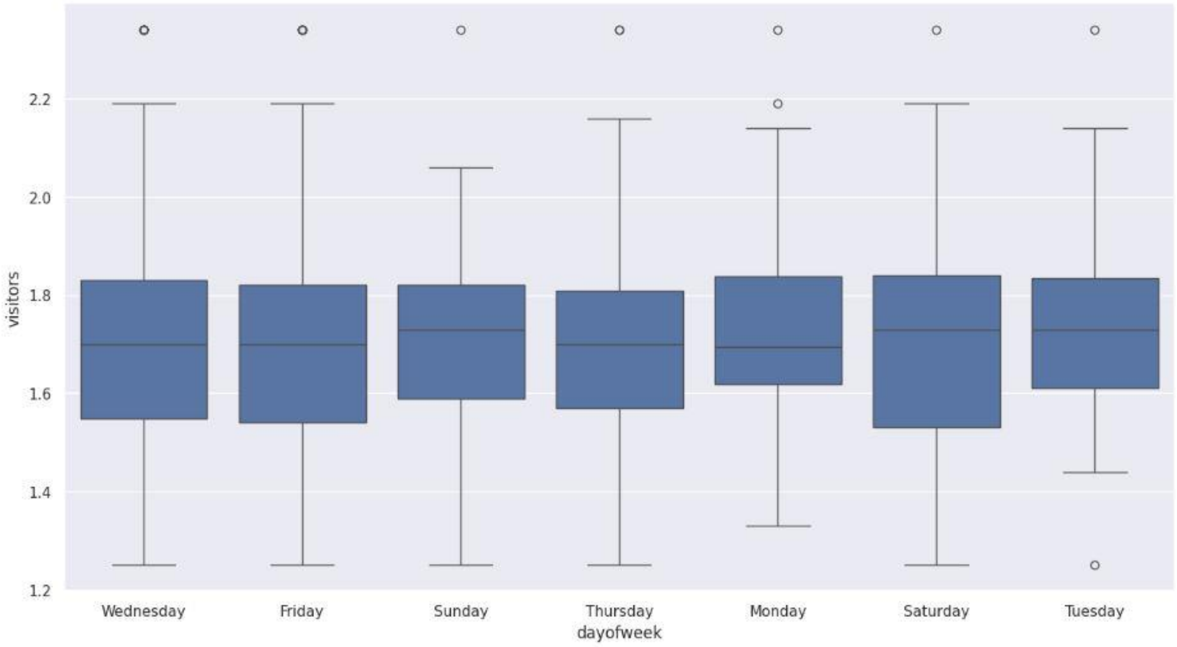


Fig 10 : Day of Week Vs Visitors

Fig 10: Shows most visitors came on Saturday and Sunday, Somewhere Tuesday is also same as well, Which is okay just because Job/Corporate person has only 2 days to view content.

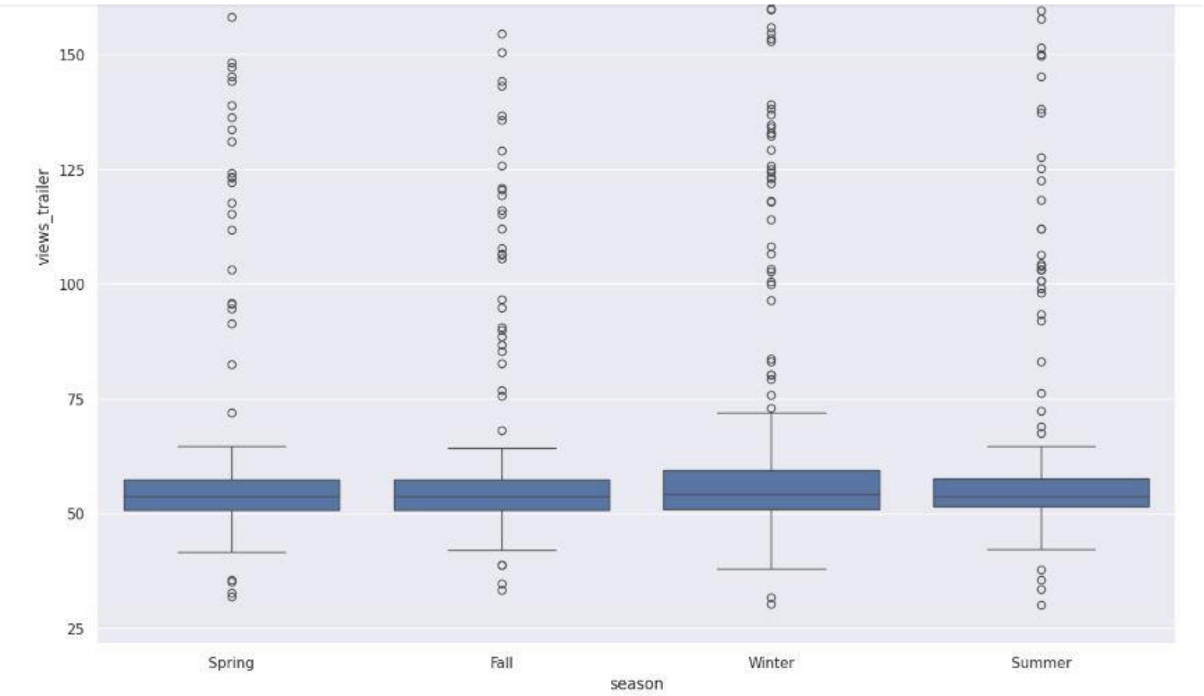


Fig 11 : Season Vs Trailer Views

Fig 11: Shows there is no impact or very minimal impact in summers for Trailer Views

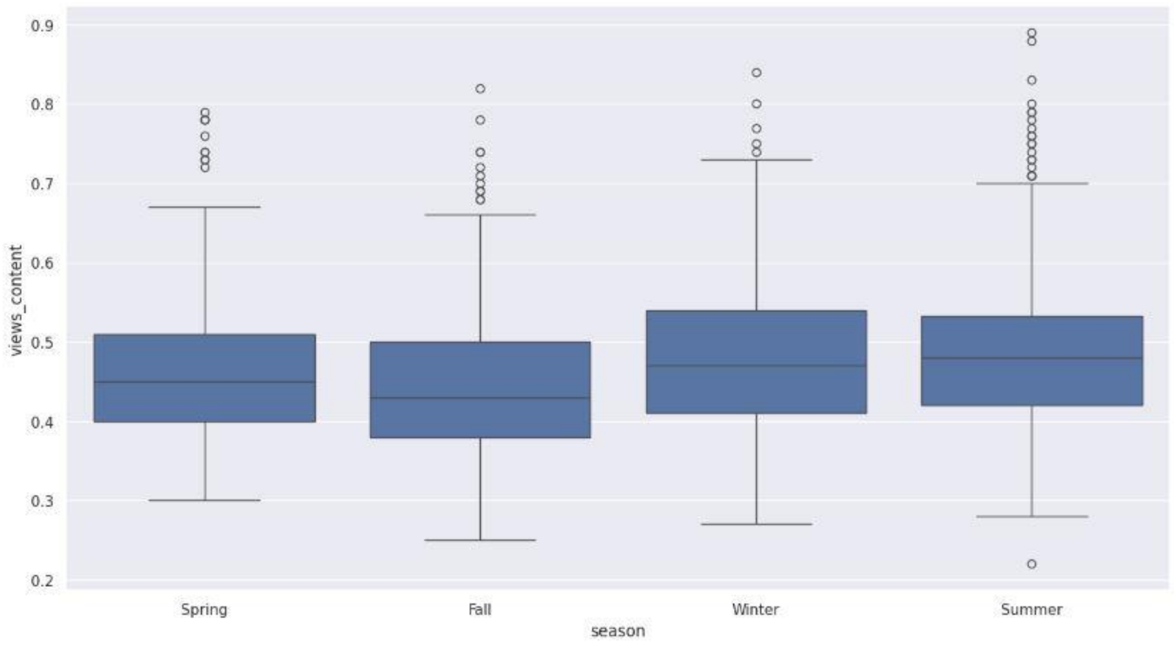


Fig 12 : Season Vs Content Views

Fig 12: Shows there is More content views in Summer and less views in Fall season, but this can be ignore because Difference is minimum.

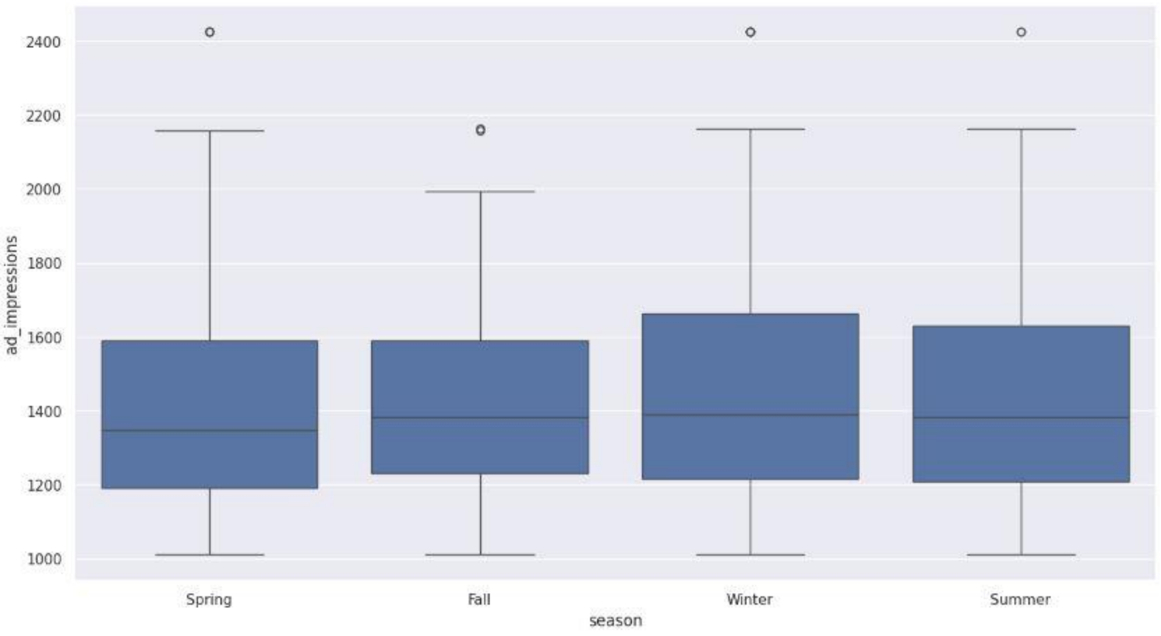


Fig 13: Season Vs Ad Impressions

Fig 13: Shows there is More Ad Impressions in Winter and Less Ad impressions in Spring season

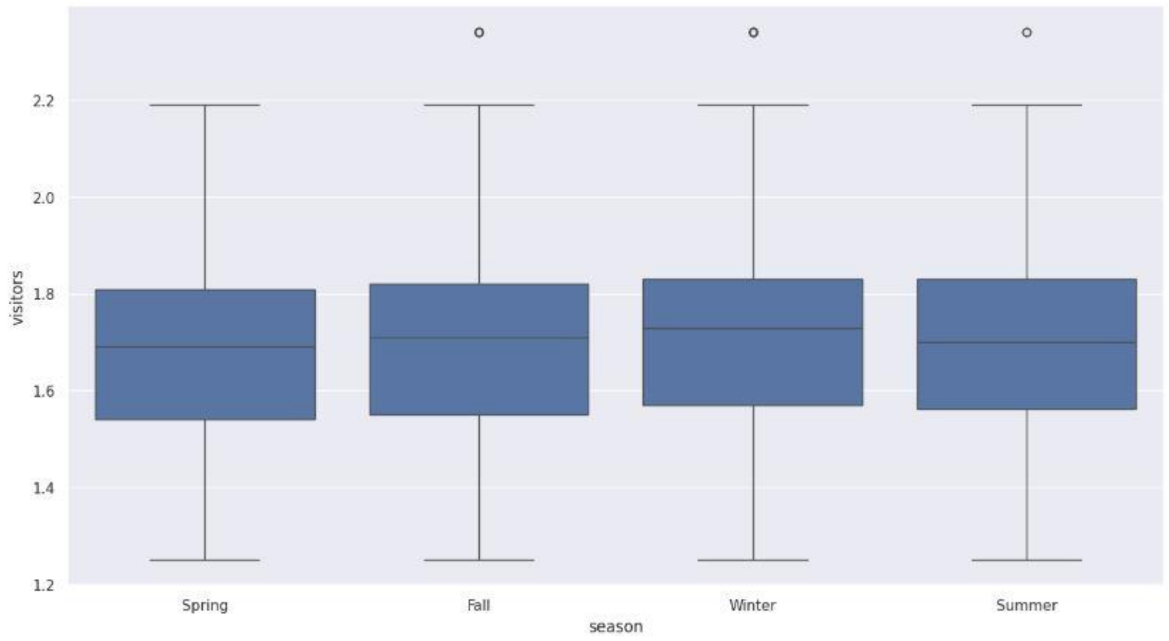


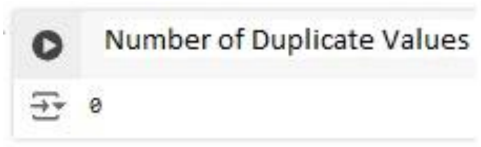
Fig 14: Season Vs Visitors

Fig 14: Shows there is almost equal Visitors in all season, there is very Negligible difference

Data preprocessing

Duplicate value check - Missing value treatment - Outlier treatment - Feature engineering - Data preparation for modelling

1.) Duplicate Value Check

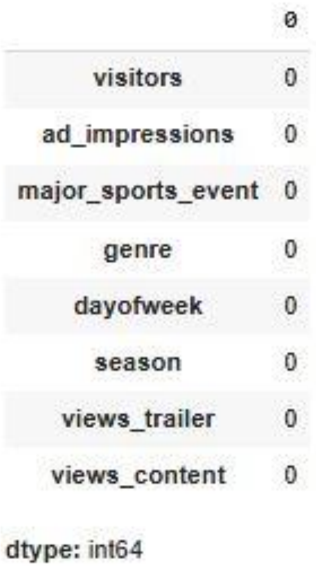


Number of Duplicate Values

```
0
```

No Duplicate Value found

2.) Checking for Missing Values



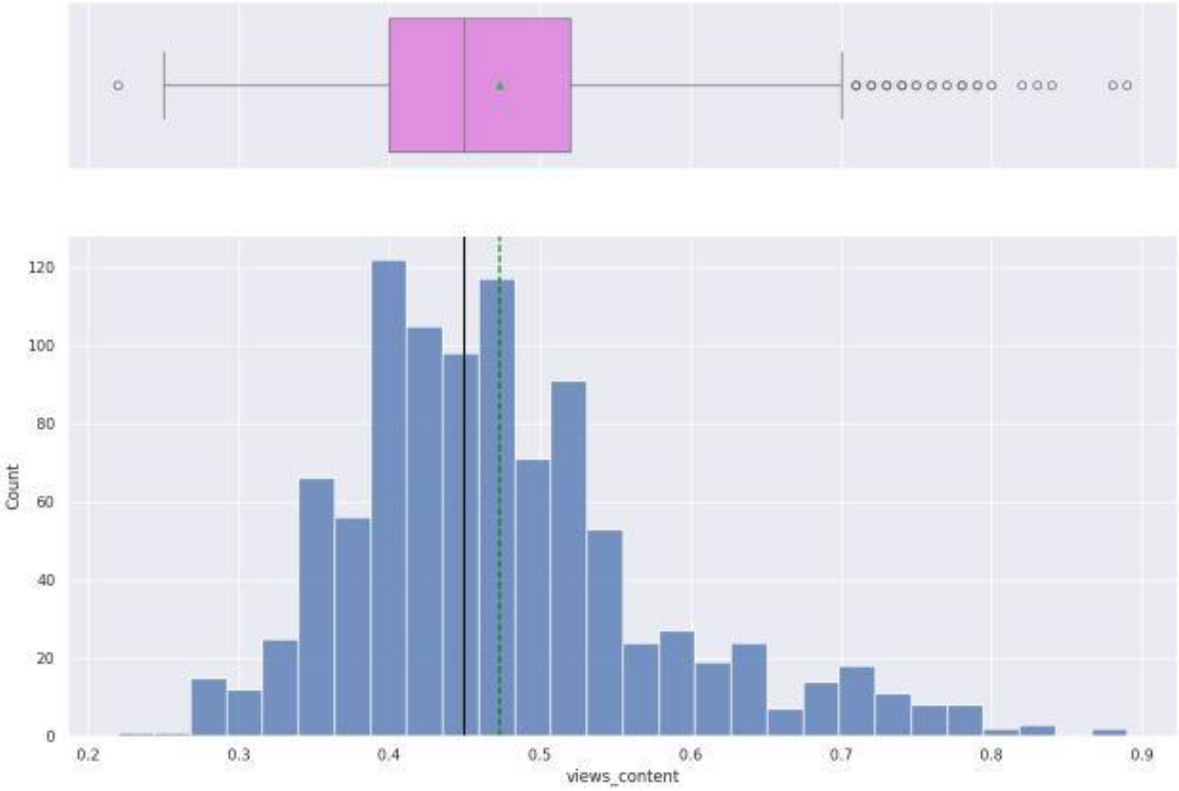
	0
visitors	0
ad_impressions	0
major_sports_event	0
genre	0
dayofweek	0
season	0
views_trailer	0
views_content	0

dtype: int64

No Missing values found

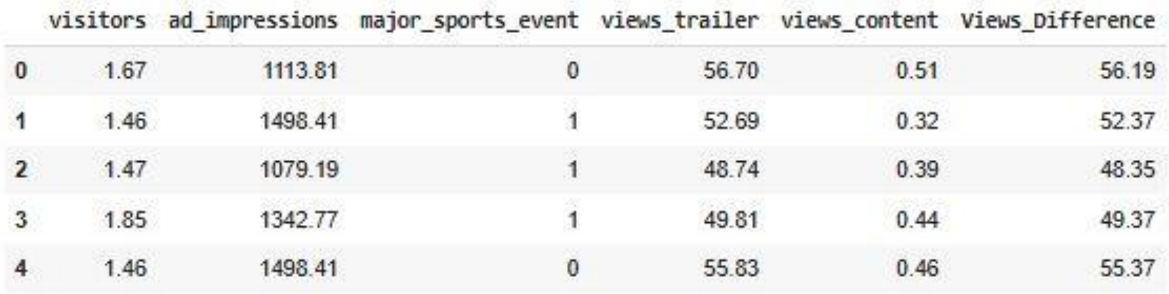
3.) Outlier Treatment

This Boxplot Graph Shows that no outlier have been detected in the content views for the given dataset



4.) Feature Engineering

We have done encoding and dropped Categorical variables (Genre, dayofweek, Season)



	visitors	ad_impressions	major_sports_event	views_trailer	views_content	Views_Difference
0	1.67	1113.81	0	56.70	0.51	56.19
1	1.46	1498.41	1	52.69	0.32	52.37
2	1.47	1079.19	1	48.74	0.39	48.35
3	1.85	1342.77	1	49.81	0.44	49.37
4	1.46	1498.41	0	55.83	0.46	55.37

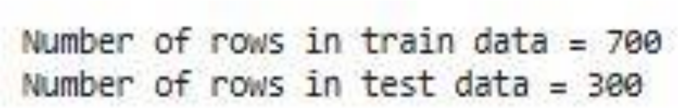
5.) Data Preparation for Modeling

Model building - Linear Regression

Build the model and comment on the model statistics - Display model coefficients with column names

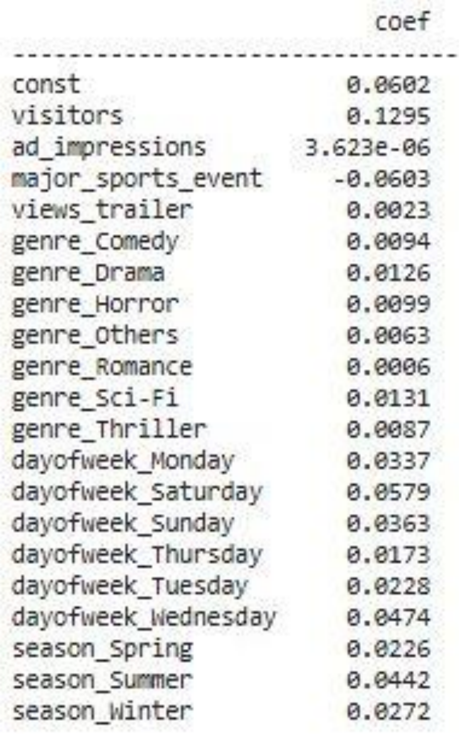
Building a linear regression model to predict content views based on the pre-processed data and Displaying Model coefficients and Statistics.

1.) Splitting the Data into Training and Test Sets.
2.) Checking the Shape of Training and Test Sets.



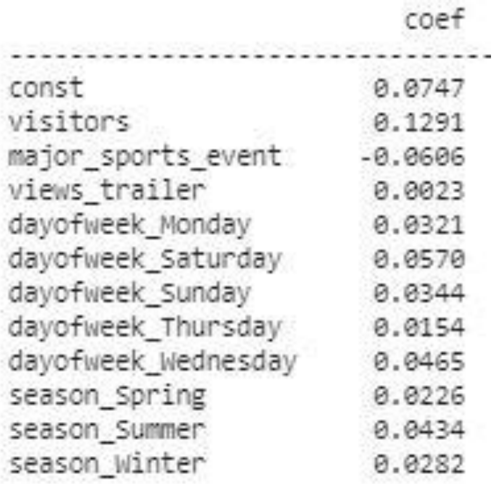
```
Number of rows in train data = 700
Number of rows in test data = 300
```

3.) Model coefficients and Statistics



```
                          coef
-------------------------------
const                   0.0602
visitors                0.1295
ad_impressions       3.623e-06
major_sports_event     -0.0603
views_trailer           0.0023
genre_Comedy            0.0094
genre_Drama             0.0126
genre_Horror            0.0099
genre_Others            0.0063
genre_Romance           0.0006
genre_Sci-Fi            0.0131
genre_Thriller          0.0087
dayofweek_Monday        0.0337
dayofweek_Saturday      0.0579
dayofweek_Sunday        0.0363
dayofweek_Thursday      0.0173
dayofweek_Tuesday       0.0228
dayofweek_Wednesday     0.0474
season_Spring           0.0226
season_Summer           0.0442
season_Winter           0.0272
```

Fig 1 : Model coefficients and Statistics (Actual)



```
                              coef
-------------------------------------
const                       0.0747
visitors                    0.1291
major_sports_event         -0.0606
views_trailer               0.0023
dayofweek_Monday            0.0321
dayofweek_Saturday          0.0570
dayofweek_Sunday            0.0344
dayofweek_Thursday          0.0154
dayofweek_Wednesday         0.0465
season_Spring               0.0226
season_Summer               0.0434
season_Winter               0.0282
```

Fig 2 : **Final Model coefficients and Statistics (After Alteration)**

Testing the assumptions of linear regression model

Perform tests for the assumptions of the linear regression - Comment on the findings from the tests

1.) **TEST FOR LINEARITY AND INDEPENDENCE**

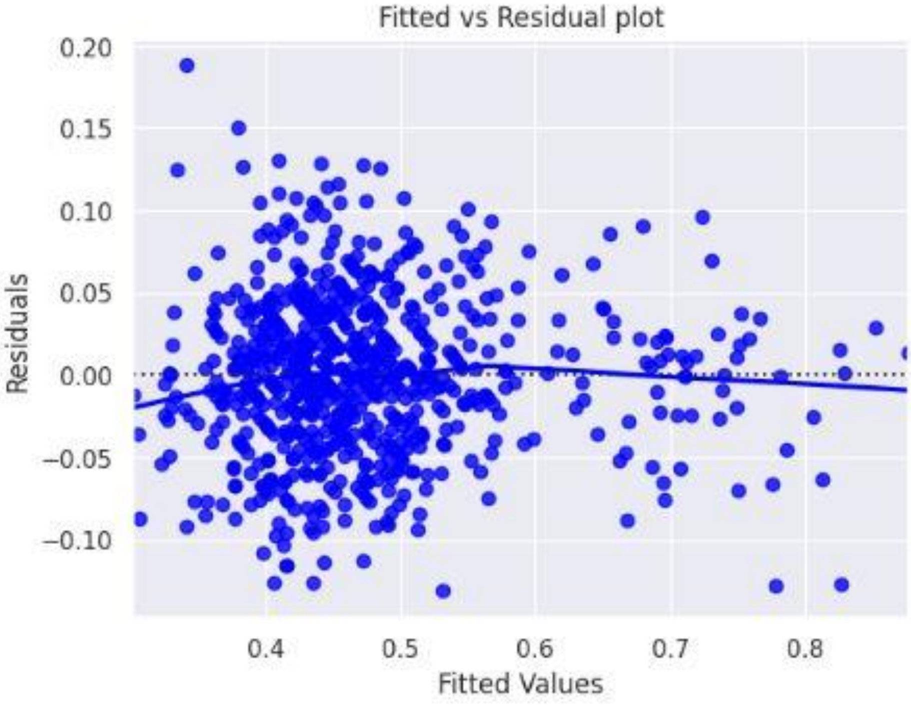


Fig : Residuals Values Vs Fitted Values

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

- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.
- We can observe a pattern in the residual vs fitted values, hence we will try to transform the continous variables in the data.

2.) **TEST FOR NORMALITY**

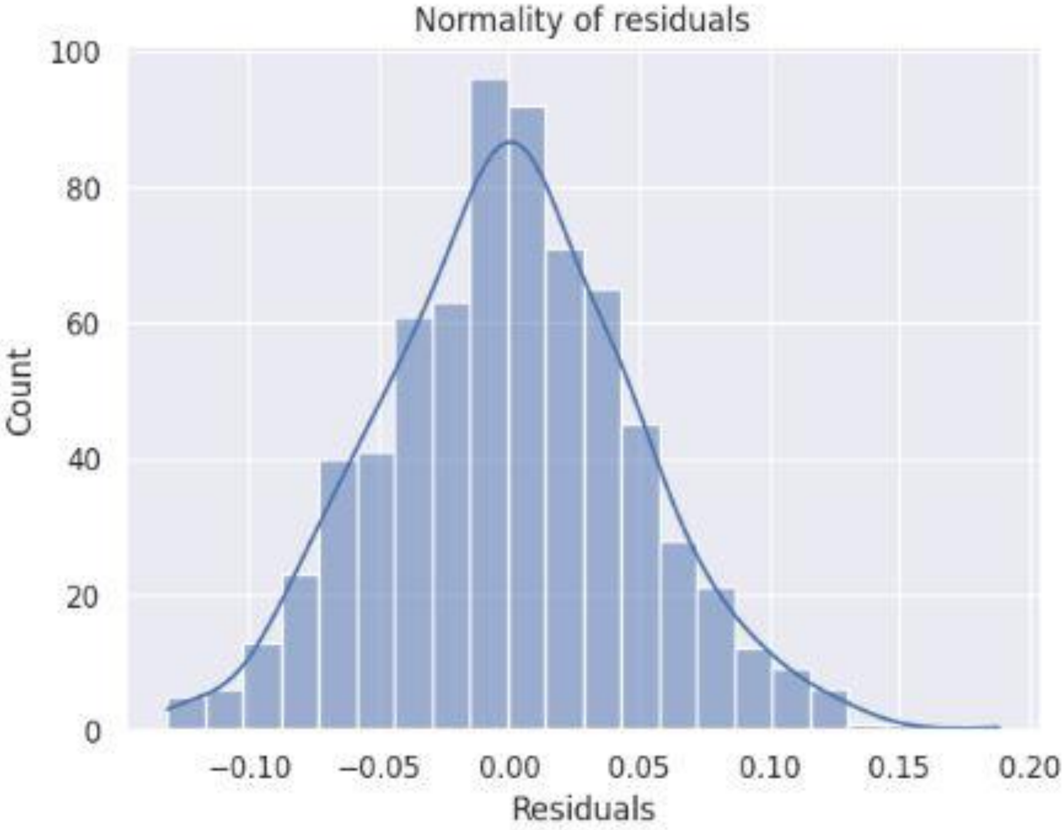
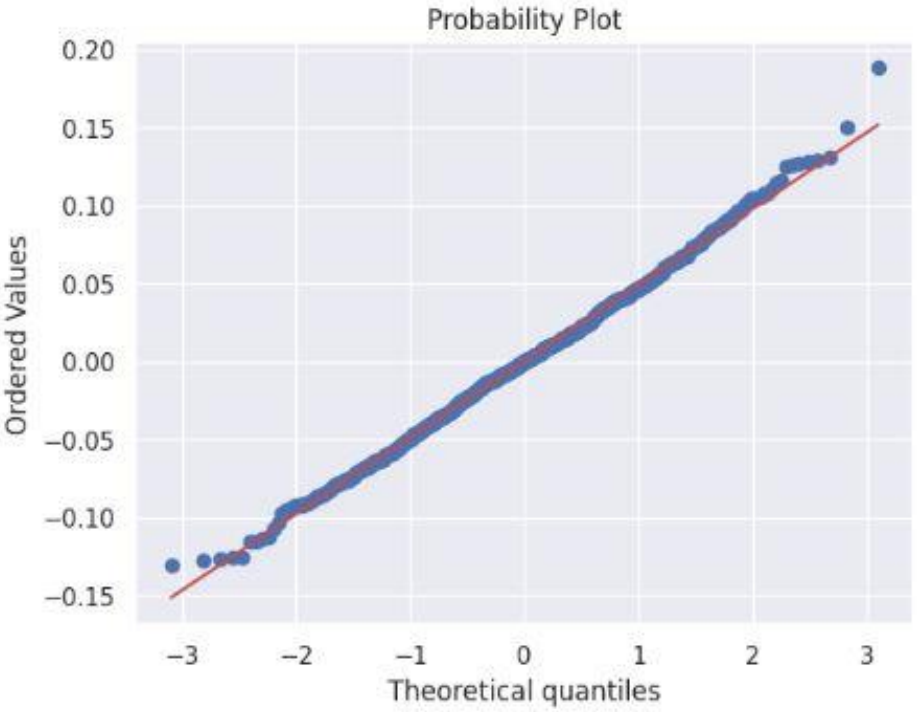


Fig : Count Vs Residuals

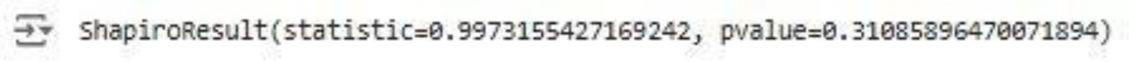
- The Above Graph Shows that The residual terms are normally distributed

Lets check Q-Q Plot

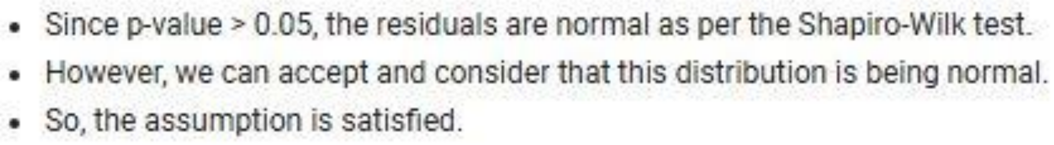


Observation :

- The residuals almost follow a straight line.
- Let's check the results of the Shapiro-Wilk test

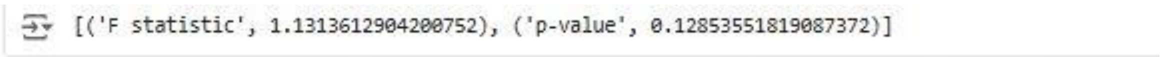


```
ShapiroResult(statistic=0.9973155427169242, pvalue=0.31085896470071894)
```

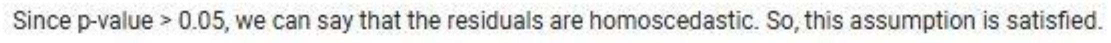


- Since p-value > 0.05, the residuals are normal as per the Shapiro-Wilk test.
- However, we can accept and consider that this distribution is being normal.
- So, the assumption is satisfied.

3.) **TEST FOR HOMOSCEDASTICITY**



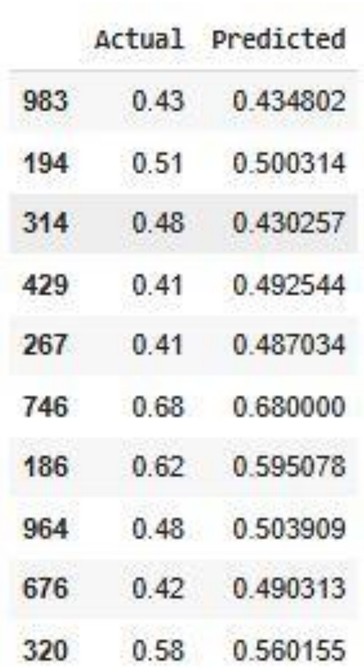
```
[('F statistic', 1.1313612904200752), ('p-value', 0.12853551819087372)]
```



Since p-value > 0.05, we can say that the residuals are homoscedastic. So, this assumption is satisfied.

Model performance evaluation

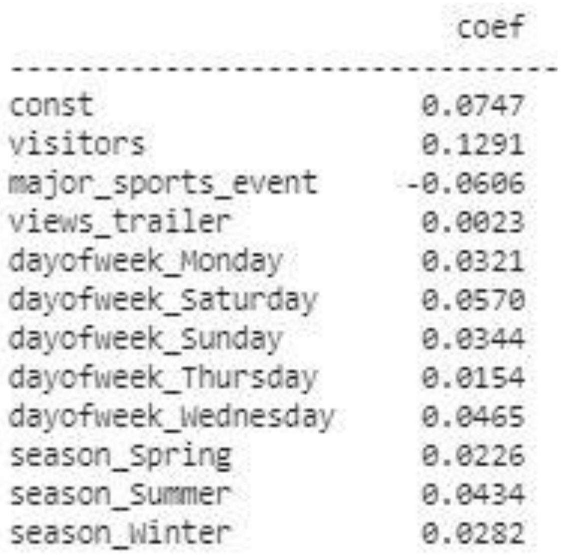
Evaluate the model on different performance metrics



	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

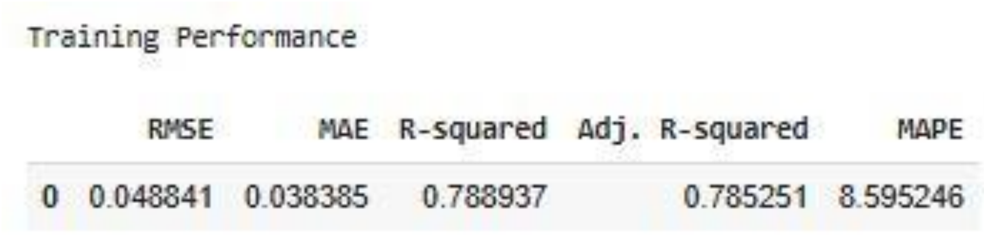
Fig : Data of Actual and Predicted Values

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



```
                              coef
--------------------------------
const                       0.0747
visitors                    0.1291
major_sports_event         -0.0606
views_trailer               0.0023
dayofweek_Monday            0.0321
dayofweek_Saturday          0.0570
dayofweek_Sunday            0.0344
dayofweek_Thursday          0.0154
dayofweek_Wednesday         0.0465
season_Spring               0.0226
season_Summer               0.0434
season_Winter               0.0282
```

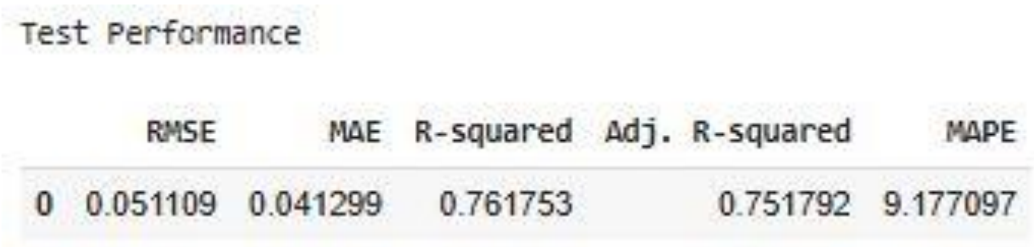
Fig : Model coefficients and Statistics of Final Model



```
Training Performance

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

Fig : Training Performance



```
Test Performance

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

Fig : Test Performance

- The model is able to explain ~78% of the variation in the data
- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting
- The MAPE on the test set suggests we can predict within 9% of the Viewership
- Hence, we can conclude the model olsmodel_final is good for prediction as well as inference purposes
- The model is able to explain ~78% of the variation in the data and within 9.1% of the Content Viewership on the test data, which is good, this indicates that the model is good for prediction as well as inference purposes

Actionable Insights & Recommendations

Comments on significance of predictors - Key takeaways for the business

Below are the insights and recommendations based on the analysis:

Significance of Predictors: The coefficients table shows that ad impressions, trailer view, and certain genres (e.g., Thriller) are significant predictors of content views.

According to Univariate Analysis

Ad impressions and visitors are high, but content views are less, So Digital Marketing/Online Marketing should be done by targeting Audience of their Category interest, which results in more Trailer and Content Views

It has been stated that, Content views less when there is any Major sports event happening at the time of release of Content so, Optimize content release schedules, and avoid clashes with major sports events.

Visitors are more often clicking the Application on Wednesday and Friday, So Teaser/Trailer/Song/Content, Anything from this should be releases on that days to give organic boost

Season has no impact on Trailer/Content Views, So it is going best as per data.

According to Bivariate Analysis

Analysis of Genre

- Trailer Views are high in Horror and drama categories
- Content views are high in Sci – fi , Horror , drama
- Ad impressions are high in sci-fi
- Visitors are high in Sci-fi, Horror and action

As Ad impressions, Trailer Views and Content Views are high in Sci-Fi, Horror and Drama, but Visitors are also considering and Liking Action Category also, So it is recommended to Make more content on Action, Advertise that content organically and inorganically both, So it will give more New and Unique visitors

Day of week

- Trailer views are Almost equal in All Categories
- Content views are more on Wednesday, Saturday & Sunday
- Ad impressions are High on Monday, Wednesday, Friday & Sunday
- Visitors are more on Saturday and Sunday

As According to Day of week data, it is recommended to go with the Current Strategy whatever is running as all the situation is exactly matching to what we need.

Analysis of Season

- Trailer views has No/negligible impact on any season
- Content views are more in Summer but there is very Minimum difference
- Ad impressions are more in winter & less In spring, which can be ignored as difference is Negligible
- Visitors are almost Equal in all Season

According to data, Ad impressions, Visitors, Trailer and Content views are almost equal in all season, there is hardly any minor difference, But main point to kept in mind is to show the Content related to interest of the Visitors/Viewer, through Digital Campaign strategies

Key Recommendations:

- Increase marketing spend to improve ad impressions.
- Teaser/Trailer should be made with main cuts from the Series, which don’t indicates or elaborates the Whole story, Make it suspense till the whole series will be watched.
- Promote Teaser/trailers more to increase first-day viewership and Launch Them in Live Stage Platform.
- We can do Strong PR before Teaser/Trailer Release to create Buzz among viewers.

Overall, this report provides a comprehensive analysis of the OTT data, identifying key patterns, relationships, and predictors of content views. The insights and recommendations generated can inform business decisions to improve the OTT service's performance.