PREDICTIVE MODELLING PROBLEM STATEMENT – CODED PROJECT

By: BENITA MERLIN.E
PGP-Data Science and Business Analytics.
BATCH: PGP DSBA. O. MAY24.A

Contents

1. BACKGROUND 6
2. OBJECTIVE 6
3. DATA DESCRIPTION 6
3.1 DATA DICTIONARY: 7
3.2 DATA INFORMATION: 7
4. STATISTICAL ANALYSIS: 8
5. EXPLORATORY DATA ANALYSIS (EDA 9
5.1.2. Views Contents: 10
5.1.3. Views_Trailer: 11
5.1.4. Ad_impressions: 12
5.1.5. Major sports events: 13
5.1.6. Genre: 14
5.1.7. Day of Week: 14
5.1.8. Season: 15
5.2. BIVARIATE ANALYSIS: 15
5.2.1. Boxplot of Genre Vs Views_Content: 15
5.2.2. Boxplot of Genre Vs Views_Trailer: 16
5.2.3 Boxplot of Day of Week Vs Views_Content: 16
5.2.4. Boxplot of Day of Week Vs Visitors: 17
5.2.5. Boxplot of View Content Vs Seasons: 17
5.2.6. Barplot of Views_Content Vs Major_Sports_Event: 18
5.2.7 Boxplot of Views Trailer Vs Seasons: 18
5.2.8. Boxplot of Ad_Impressions Vs Seasons: 19
5.2.9. Barplot of Genre Vs Ad_Impressions: 19
5.3. MULTIVARIATE ANALYSIS 20
5.3.1. Relationship among Genre, Visitors and Season: 20
5.3.2. Relationship among Genre, View Content and Season: 21
5.3.3. Relationship among Ad_Impression, Major Sports Event and Season: 22
5.3.4 Relationship among Genre, Major Sports Event and Views_Content: 22
5.3.5. Relationship among Views Trailer, Day of Week and Views_Content: 23
5.3.6. Relationship among Visitor, Major Sports Event and Season 24
5.3.7. Relationship among Visitors, Major Sports Event and Genre: 24
5.3.9. Pair plot of Numeric Data 26
6. QUESTION AND ANSWER: 27
Question 1: What does the distribution of content views look like? 27

Question 2. What does the distribution of genres look like? 28
Question 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? 29
Question 4: How does the viewership vary with the season of release? 30
Question 5: What is the correlation between trailer views and content views? 31
7. DATA PREPROCESSING 31
7.1 Outlier Treatment 32
7.2. Feature Engineering 33
8. MODEL BUILDING - LINEAR REGRESSION 34
9. CHECKING LINEAR REGRESSION ASSUMPTIONS 37
9.1 Test for Multicollinearity 37
9.2. Linearity and Independence of Variables 43
9.3. Normality of error terms 44
9.4. Test for Homoscedasticity 46
10. PREDICTIONS ON TEST DATA 46
11. FINAL MODEL 47
12. ACTIONABLE INSIGHTS AND RECOMMENDATIONS 48
12.1 Conclusion on predictor significance 49
12.2 Key takeaways for the business 49

List of Figures

Title	Page number
Fig 1. Statistical summary	8
Fig 2: Univariate Visitor Analysis	9
Fig 3: Univariate Content View Analysis	10
Fig 4: Univariate Trailer View Analysis	11
Fig 5: Univariate Ad Impression Analysis	12
Fig 6: Univariate Major Sports Event	13
Fig 7: Univariate Genre Counts	14
Fig 8: Univariate day of content release	15
Fig 9: Univariate Seasons	15
Fig 10: Genre vs Content Views	16
Fig 11: Genre vs Trailer Views	16
Fig 12: Content Views vs day of release	17
Fig 13: Visitors vs day of release	17
Fig 14: Box plot for Season vs View content	18
Fig 15: Bar plot for Major sports event vs View content	18
Fig 16: Boxplot for Season vs View Trailer	19
Fig 17: Boxplot for Season vs ad-impression	19
Fig 18: Bar plot for Genre vs Ad-impression	20
Fig 19: Bar plot of Day of week vs View trailer	20
Fig 20: Bar plot for Genre, Visitor and Season	21
Fig 21: Boxplot for Genre, View Content and Season	22
Fig 22: Boxplot for ad-impression, Major sports event and Season	22
Fig 23: Violin plot for Genre, Major Sports Event and Views Content	23
Fig 24: Content vs Trailer Views w. r. t. day of content release	23
Fig 25: Boxplot for Visitor, Major Sports Event and Season	24
Fig 26: Boxplot for Visitor, Major Sports Event and Genre	25
Fig 27: Heat Map for Numeric Data	25
Fig 28: Pair Plot for Numeric Data	26
Fig 29: Distribution of Content Views	27
Fig 30: Distribution of Genre	28
Fig 31: Average viewership by day of release	29
Fig 32: Average viewership by season of release	30
Fig 33: correlation between trailer views and content views	31
Fig 34: Before outlier Treatment	32
Fig 35: After outlier Treatment	32
Fig 36: Replacement of yes and no in column Major sports event	33
Fig 37: OLS Model with Multicollinearity and high p-values	34
Fig 38: OLS Model After Removal of Multicollinearity	40

Fig 39: OLS Model without Multicollinearity and high p-values	42
Fig 40: Linearity and Independence of Variables	43
Fig 41: Normality of Residuals	44
Fig 42: Q-Q plot of residuals	45
Fig 43: Final OLS Model	47

List of Table

Title	Page number
Table 1. Data Dictionary	7
Table 2. Data Information	7
Table 3: Training and Test Performance of existing variables	36
Table 4: VIF (Variance Inflation factor) before Removal of Multicollinearity	38
Table 5: VIF (Variance Inflation factor) After Removal of Multicollinearity	39
Table 6: Training and Test Performance after removal of variables	42
Table 7: Actual vs Predicted value of built model	46
Table 8: Training and Test Performance of Final Model	47

1. BACKGROUND

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.

2. OBJECTIVE

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

3. DATA DESCRIPTION

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

Dataset Name: ottdata

3.1 DATA DICTIONARY:

This data dictionary provides a clear and concise reference for understanding the dataset's structure and the meaning of each column, ensuring accurate and consistent data handling in analysis.

SI.NO	VARIABLE	DESCRIPTION
1	visitors	Average number of visitors, in millions, to the platform in the past week
2	ad_impressions	Number of ad impressions, in millions, across all ad campaigns for the content (running and completed)
3	major_sports_event	Any major sports event on the day
4	genre	Genre of the content
5	dayofweek	Day of the release of the content
6	season	Season of the release of the content
7	views_trailer	Number of views, in millions, of the content trailer
8	views_content	Number of first-day views, in millions, of the content

Table 1. Data Dictionary

3.2 DATA INFORMATION:

```
<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
```

Table 2. Data Information

Observation:

- There are in total 1000 rows and 8 columns present.
- There are 5 numeric (float and int type) and 3 string (object type) columns in the data.
- The target variable is the views_content (first day views), which is a float type.
- In total,4 seasons observed.

- There are no duplicates.
- There are no missing values in all the columns.

4. STATISTICAL ANALYSIS:

The statistical analysis provides a summary of the key metrics for the numerical columns in the dataset. This includes measures such as the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), and maximum values for each column.

Fig 1. Statistical summary

	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

Observation:

- The distribution of visitors has a mean of 1.70429 and ranges from 1.25 to 2.34, indicating a relatively narrow spread around the mean.
- The ad impressions have a mean of 1434.71229 with a significant range from 1010.87 to 2424.2, indicating a wider spread and potential variability in impressions.
- This binary variable (presumably indicating whether a major sports event is happening or not)
- It has a mean of 0.4, suggesting that about 40% of the observations involve a major sports event.
- There are 8 unique genres, with 'Others' being the most frequent genre, occurring 255 times in the dataset.
- There are 7 unique days of the week with 'Friday' being the most common, occurring 369 times, indicating a possible trend or preference for Fridays.
- There are 4 unique seasons with 'Winter' being the most frequent, occurring 257 times.
- The views for trailers have a mean of 66.91559, with values ranging from 30.08 to 199.92, indicating a wide range of viewership.
- The views for content have a mean of 0.4734, with values ranging from 0.22 to 0.89, suggesting a narrower range and possibly a more consistent viewership.

5. EXPLORATORY DATA ANALYSIS (EDA)
5.1. UNIVARIATE ANALYSIS:

5.1.1. Visitors:

Fig 2: Univariate Visitor Analysis

Observation:

Box Plot (Top Plot):

- Median: The central line within the box represents the median (50th percentile) of the data.
- Interquartile Range (IQR): The box itself represents the interquartile range (IQR), which contains the middle 50% of the data (from the 25th to the 75th percentile).
- Whiskers: The lines extending from the box (whiskers) show the range of the data, excluding outliers. They typically extend to 1.5 times the IQR.
- Outliers: The points outside the whiskers are considered outliers. In this plot, there appears to be an outlier on the right side.

Histogram (Bottom Plot):

- Distribution Shape: The histogram shows the frequency distribution of the visitor data. It seems to be slightly right-skewed, indicating that there are more data points clustered towards the lower values with a tail extending towards higher values.
- Peak: The highest bars in the histogram represent the mode, which is the most frequent range of visitor counts. Here, it seems that the most common visitor counts are around 1.7.

- Spread: The distribution of visitors ranges from approximately 1.2 to 2.2, with most values concentrated between 1.4 and 1.8.
- Mean and Median: The dashed vertical line in the histogram represents the mean, which appears to be close to the median in the box plot, indicating a relatively symmetric distribution.

5.1.2. Views Contents:

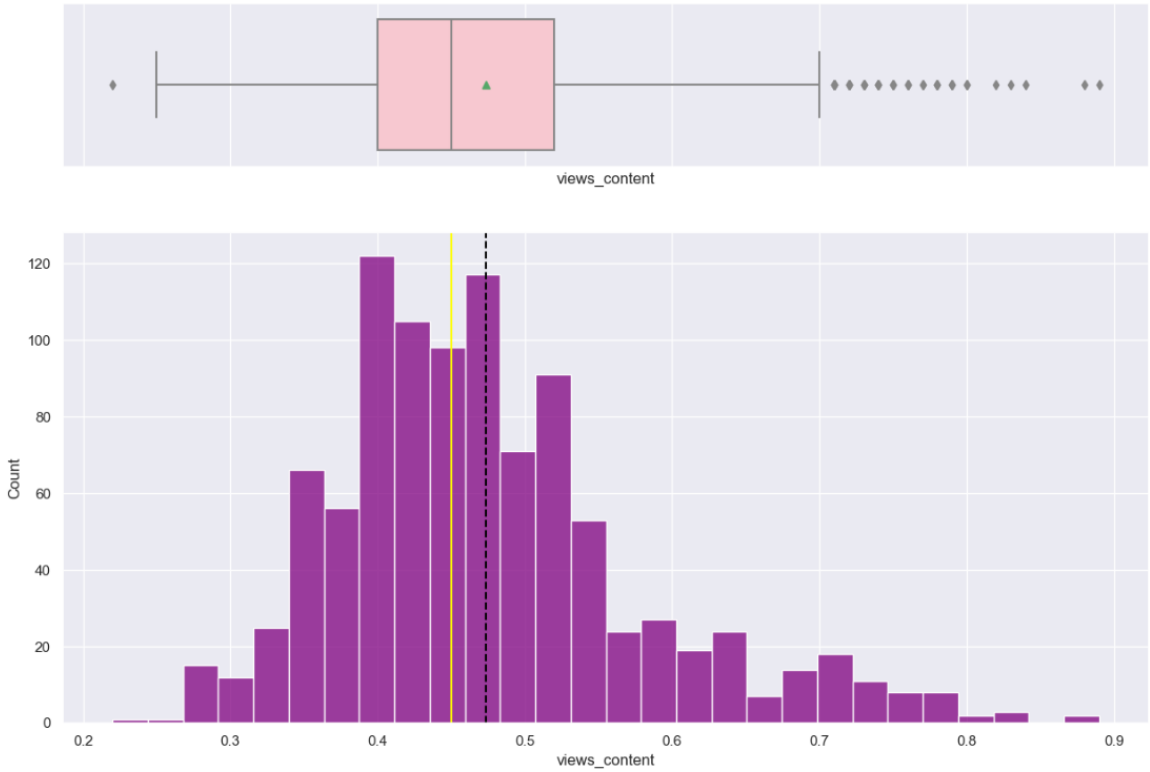


Fig 3: Univariate Content View Analysis

Observation:

Box Plot Analysis

- **Median**: The median of **views_content** is approximately 0.45, indicating that half of the observations have values below this point.
- **Interquartile Range (IQR)**: The box (spanning from Q1 to Q3) shows that the middle 50% of the data falls between approximately 0.4 and 0.52.
- **Whiskers and Outliers**:
- The lower whisker extends to approximately 0.22, indicating the minimum non-outlier value.
- The upper whisker extends to around 0.7, but there are several outliers beyond this point, reaching up to 0.89.
- **Mean**: The green triangle indicates the mean, which is slightly higher than the median, suggesting a right-skewed distribution.

Histogram Analysis

- **Distribution Shape**: The histogram reveals a right-skewed distribution, with a higher frequency of lower values and a tail extending towards higher values.

- **Peak**: The mode of views_content is around 0.45, corresponding with the median in the box plot.
- **Spread**: The data ranges from approximately 0.2 to 0.9, with the highest concentration of values between 0.4 and 0.5.
- **Frequency**: The highest bar represents the most frequent value range, confirming that most content views cluster around the median.

5.1.3. Views_Trailer:

Fig 4: Univariate Trailer View Analysis

Observation:

Box Plot Analysis

- Median: The median of views_content is approximately 59, indicating that half of the observations have values below this point.
- Interquartile Range (IQR): The box (spanning from Q1 to Q3) shows that the middle 50% of the data falls between approximately 50 and 65.
- Whiskers and Outliers:
- The lower whisker extends to approximately 40, indicating the minimum non-outlier value.
- The upper whisker extends to around 70, but there are several outliers beyond this point, reaching up to 200.
- Mean: The green triangle indicates the mean, which is slightly higher than the median, suggesting a right-skewed distribution.

Histogram Analysis

- Distribution Shape: The histogram reveals a right-skewed distribution, with a higher frequency of lower values and a tail extending towards higher values.
- Peak: The mode of views_trailer is around 55, corresponding with the median in the box plot.
- Spread: The data ranges from approximately 30 to 200, with the highest concentration of values between 45 and 70.
- Frequency: The highest bar represents the most frequent value range, confirming that most trailer views cluster around the median.

5.1.4. Ad_impressions:

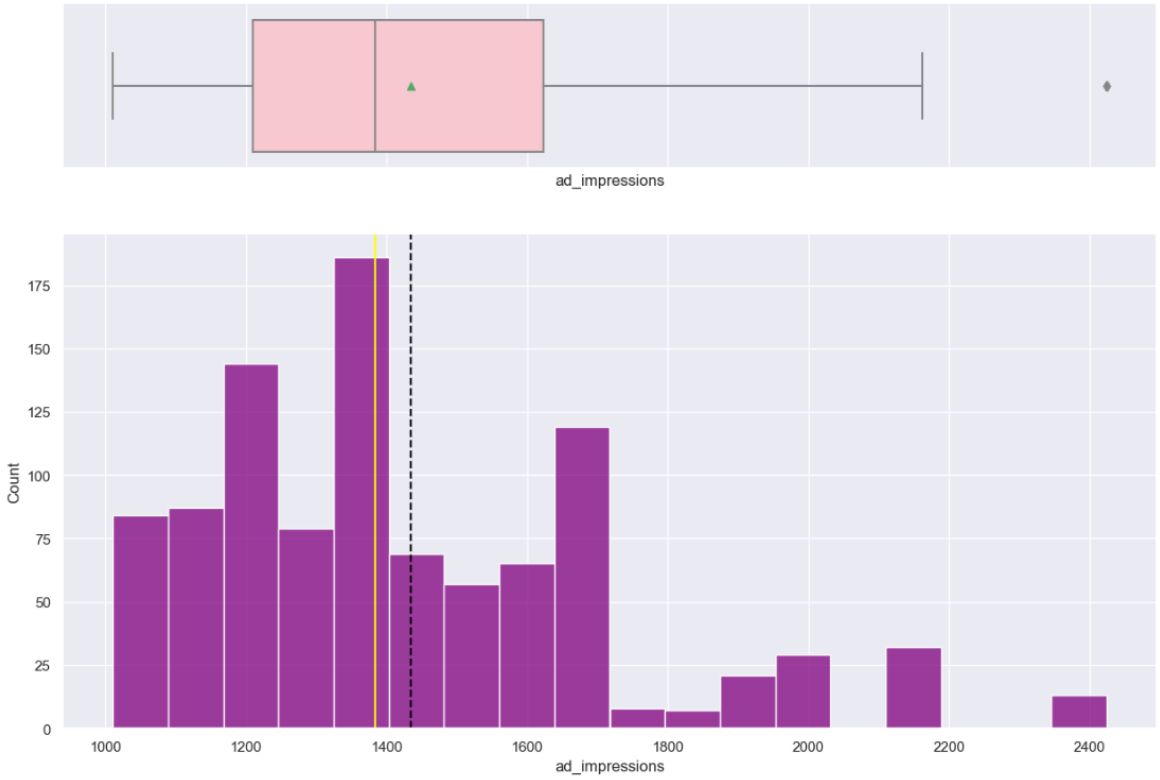


Fig 5: Univariate Ad Impression Analysis

Boxplot

- Median: The median (central line in the box) of ad_impressions appears to be around 1400.
- Interquartile Range (IQR): The box represents the IQR, which shows that the middle 50% of ad_impressions values lie between approximately 1200 and 1600.
- Whiskers: The whiskers extend to the minimum and maximum values within 1.5 times the IQR from the quartiles. The lower whisker ends around 1000, and the upper whisker ends around 2000.
- Outliers: There is one outlier on the higher end beyond 2400.

Histogram

- Distribution Shape: The histogram shows a right-skewed distribution of ad_impressions. There are more values clustered towards the lower range.
- Mode: The highest peak occurs around 1400, indicating this is the most frequent value range for ad_impressions.
- Spread: The distribution ranges from just below 1000 to above 2400.

Comparison with Mean and Median:

- Mean: The yellow line represents the mean, which appears slightly higher than the median, confirming the right-skewed nature.
- Median: The black dashed line represents the median, which aligns with the central peak of the histogram.

5.1.5. Major sports events:

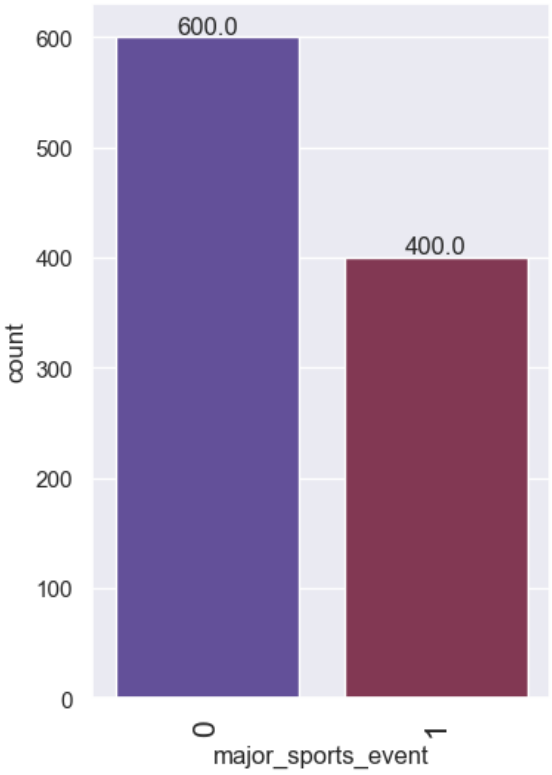


Fig 6: Univariate Major Sports Event

- Only 400 people have watched major_sports_event
- 600 people are not interested in major_sports_event

5.1.6. Genre:

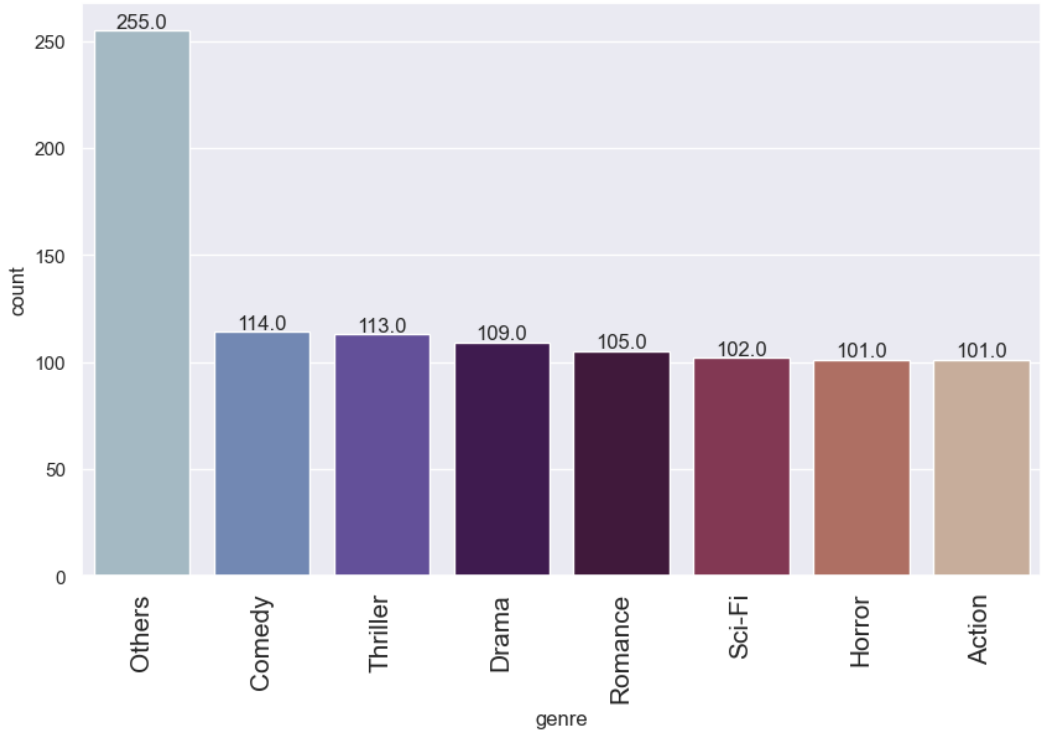


Fig 7: Univariate Genre Counts

- Genre Others have the highest percentage 25.5.

5.1.7. Day of Week:

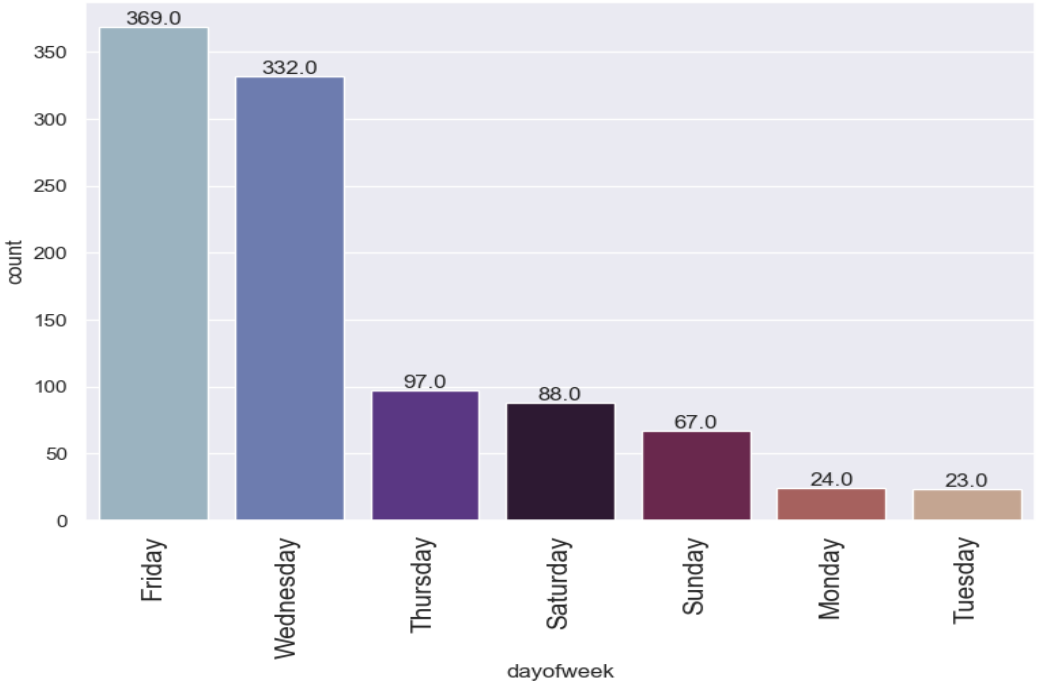


Fig 8: Univariate day of content release

- Most of the content have released on Friday followed by Wednesday.
- Very few contents have released on Tuesday.

5.1.8. Season:

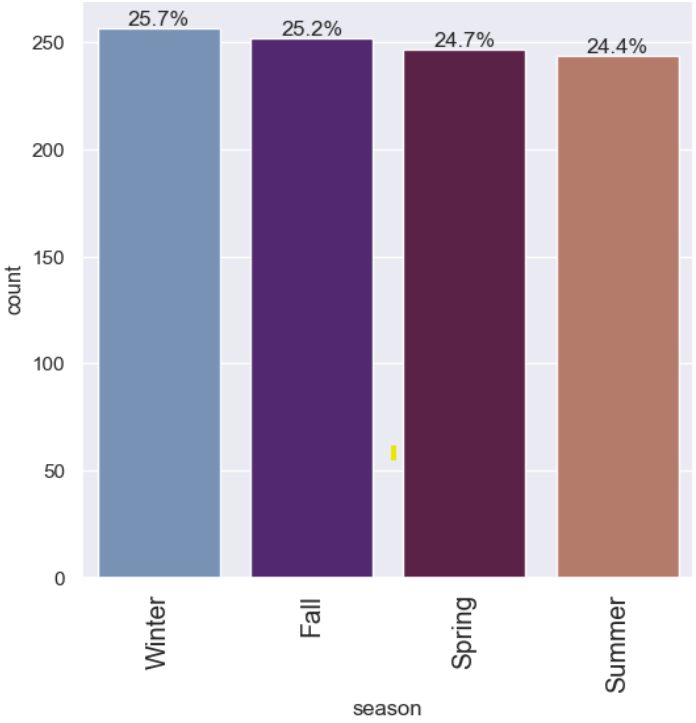


Fig 9: Univariate Seasons

- There are four seasons almost all of them contains one fourth of the data.

5.2. BIVARIATE ANALYSIS:

5.2.1. Boxplot of Genre Vs Views_Content:

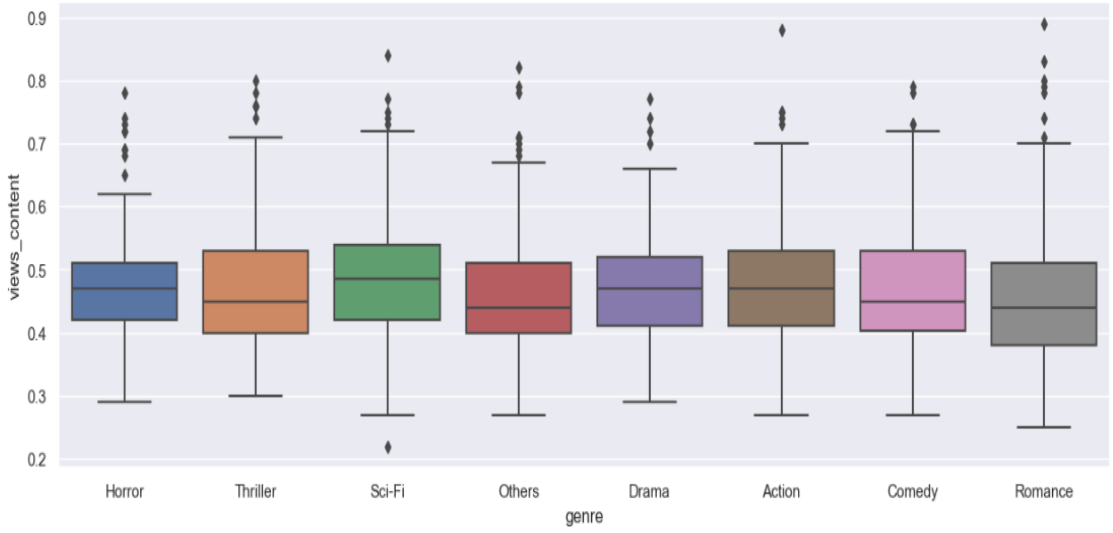


Fig 10: Genre vs Content Views

- Sci-Fi has higher variability in views, suggesting that while some content in these genres is very popular, there is also a significant amount of less-viewed content.
- Horror and Romance have multiple high outliers, indicating that certain content in these genres is particularly popular.

5.2.2. Boxplot of Genre Vs Views_Trailer:

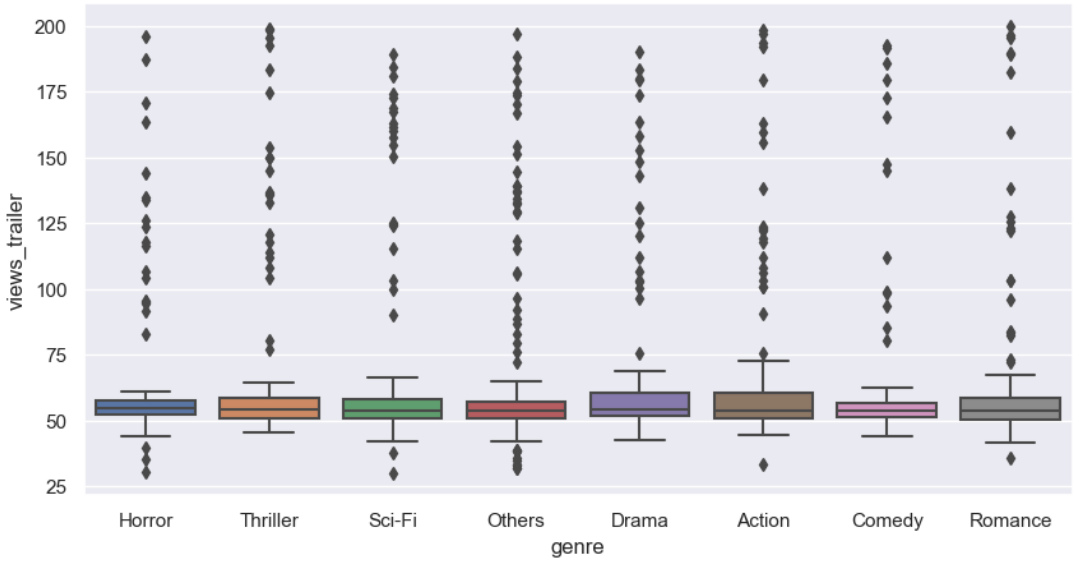


Fig 11: Genre vs Trailer Views

- Horror, Romance, Action, others and Thriller: These genres have the highest outliers, suggesting that trailers for these genres tend to attract more views.
- Drama and Comedy: Despite having high outliers, these genres have a slightly narrower range of typical trailer views, indicating more consistency.

5.2.3 Boxplot of Day of Week Vs Views_Content:

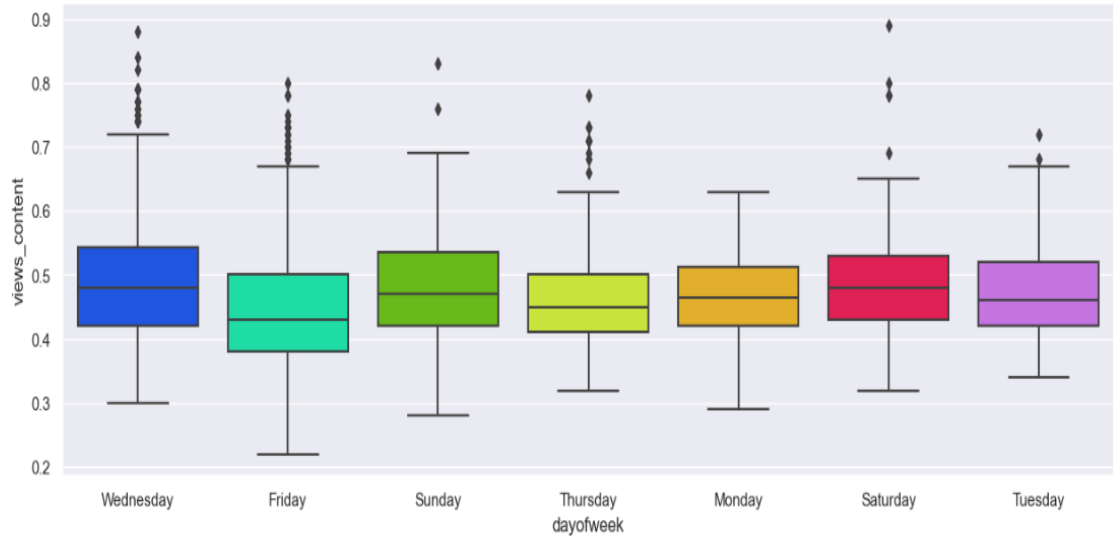


Fig 12: Content Views vs day of release

- Saturday and then Wednesday have high viewership of content.
- less viewership of content on Friday.

5.2.4. Boxplot of Day of Week Vs Visitors:

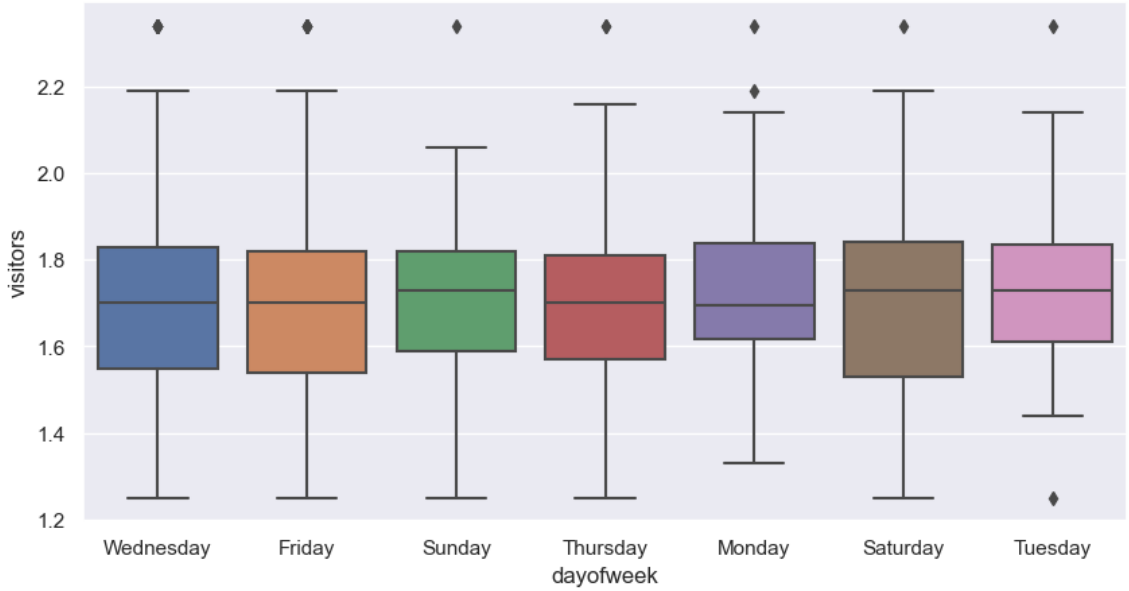


Fig 13: Visitors vs day of release

- More visitors have seen content release on Saturday, Tuesday and Sunday.
- Outliers are in all the day of the week. Some content release has high response from visitors.

5.2.5. Boxplot of View Content Vs Seasons:

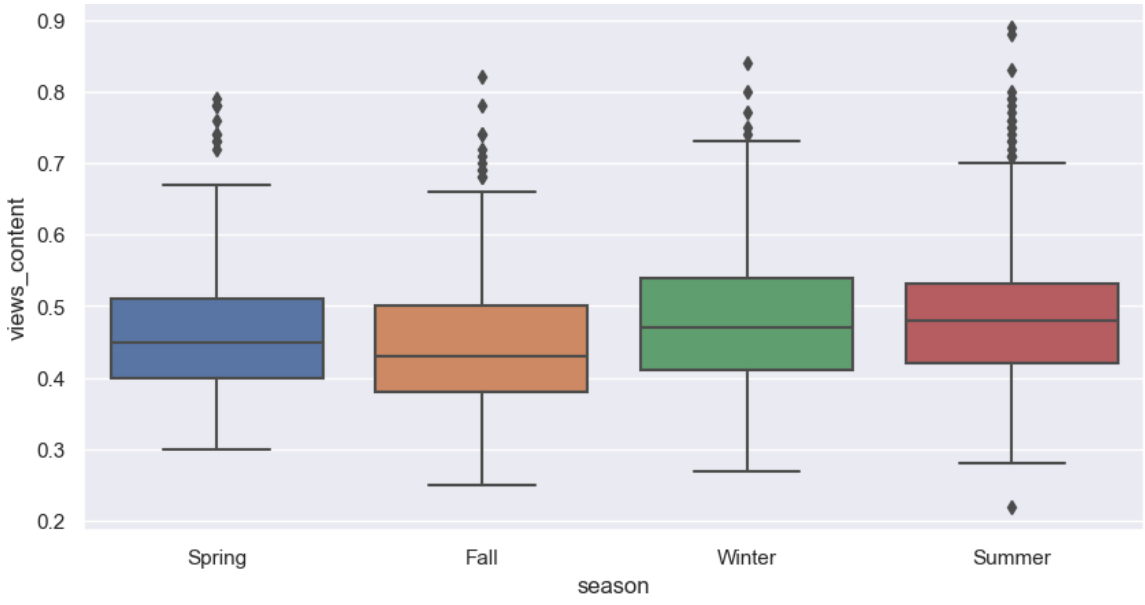


Fig 14: Box plot for Season vs View content

- More viewership on content was found in summer.
- All season has outlier. Some content in all the seasons was attracted by viewers.

5.2.6. Barplot of Views_Content Vs Major_Sports_Event:

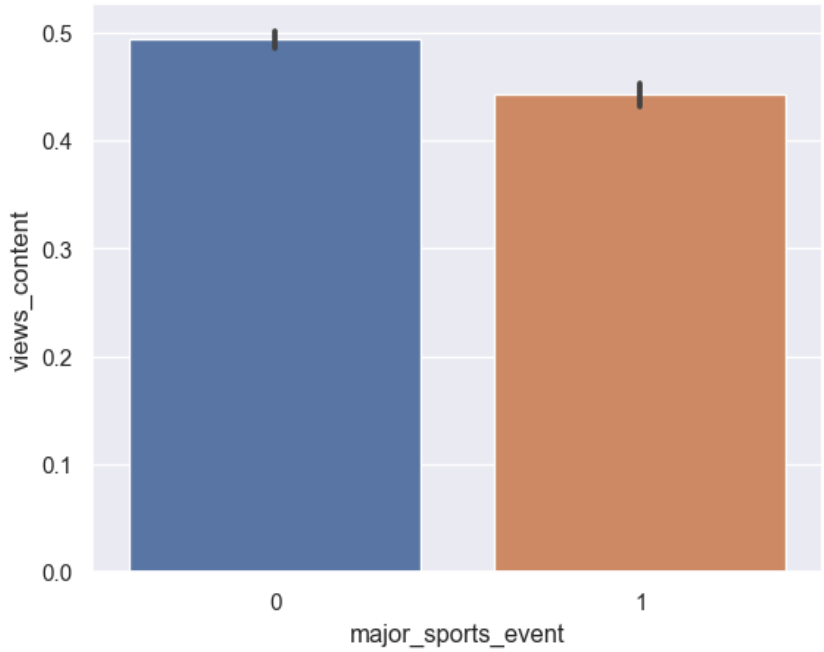


Fig 15: Bar plot for Major sports event vs View content

- A significant decrease in average content viewership can be observed due to sports events, with 0.49 million views on non-sports event days compared to 0.44 million views on sports event days.

5.2.7 Boxplot of Views Trailer Vs Seasons:

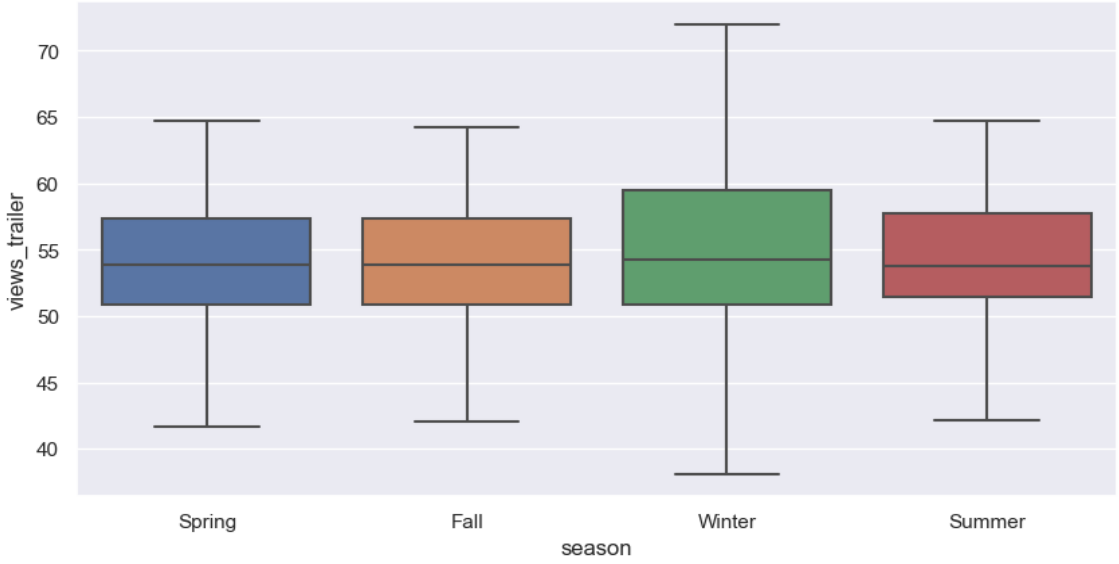


Fig 16: Boxplot for Season vs View Trailer

- The mean trailer views are almost at same number for all the season release of content.
- However, for winter the third quartile is higher than all the season realease.

5.2.8. Boxplot of Ad_Impressions Vs Seasons:

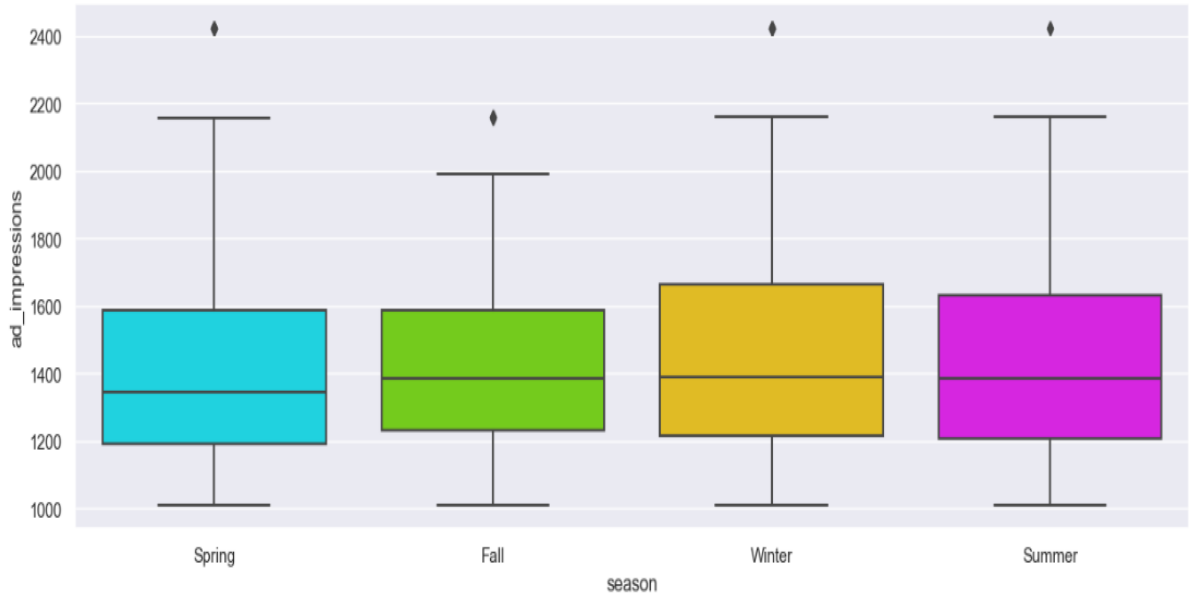


Fig 17: Boxplot for Season vs ad-impression

5.2.9. Barplot of Genre Vs Ad_Impressions:

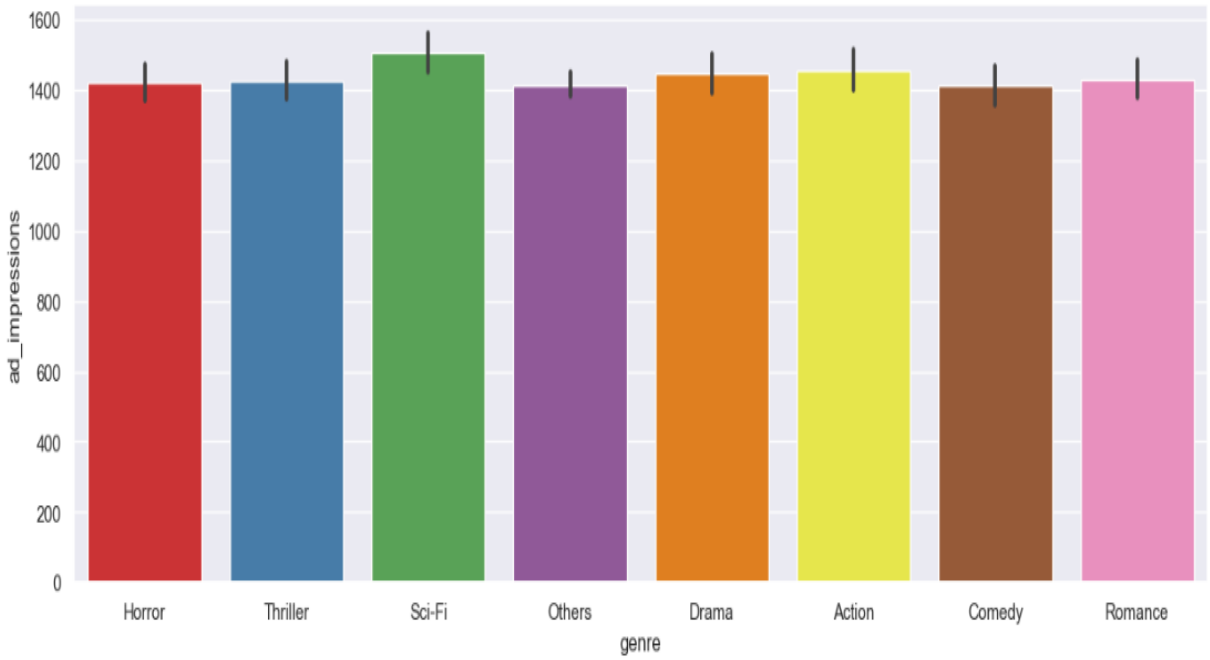


Fig 18: Barplot for Genre vs Ad-impression

- Ad-impression is little bit high for the Sci-Fi than another genre.

5.2.10. Barplot of Day of Week Vs Views_Trailer:

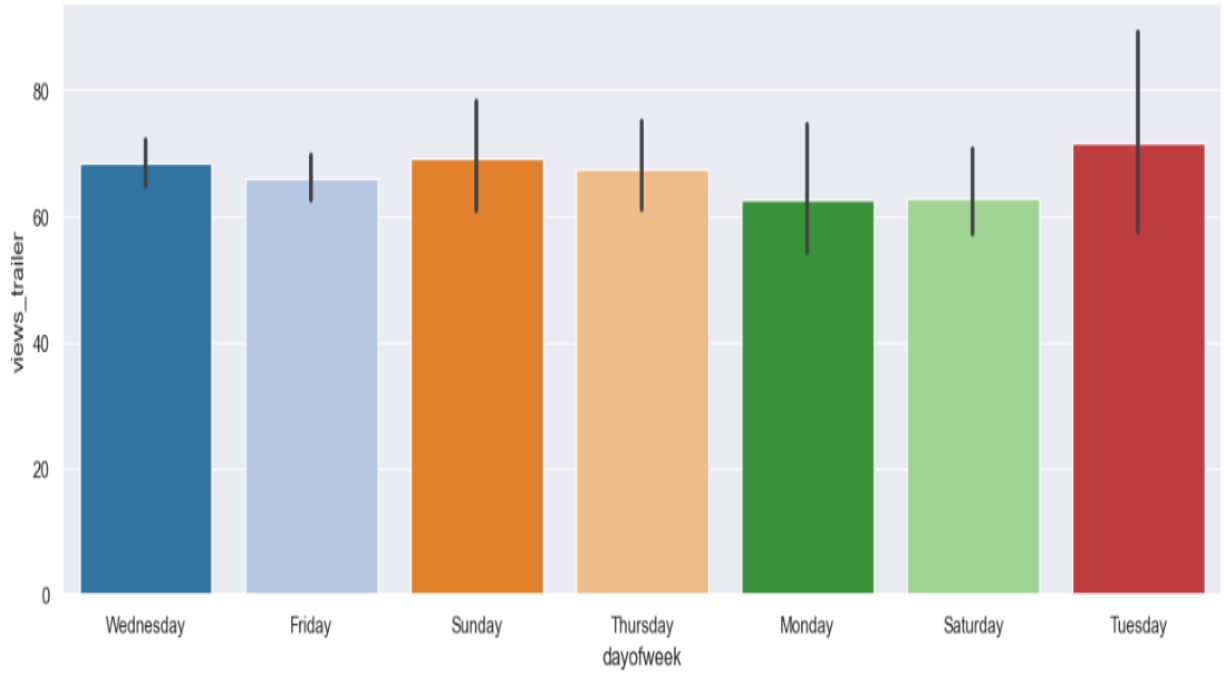


Fig 19: Barplot of Day of week vs View trailer

- Tuesday and Sunday have the high views for trailer
- The mean trailer views are almost similar (53 to 55 millions) for all the day of the release of the content.
- In all the days of content release outliers can be observed.

5.3. MULTIVARIATE ANALYSIS:

5.3.1. Relationship among Genre, Visitors and Season:

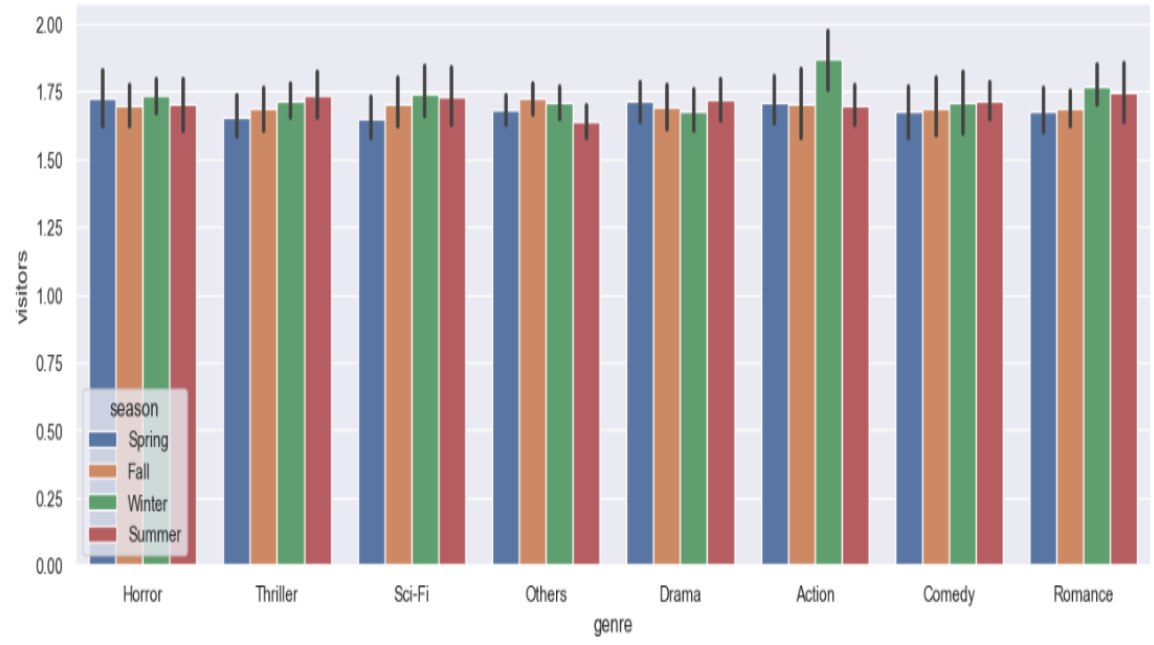


Fig 20: Barplot for Genre, Visitor and Season

- Overall, Action content has high views on winter season comparatively
- For Horror,winter season has high visitors.
- For Thriller content,summer has the high visitors.
- For Sci-Fi,winter season has high views.
- For others.fall season has high visitors.
- For Drama,summer season has high visitors.
- Romance has high views on winter and summer season.

5.3.2. Relationship among Genre, View Content and Season:

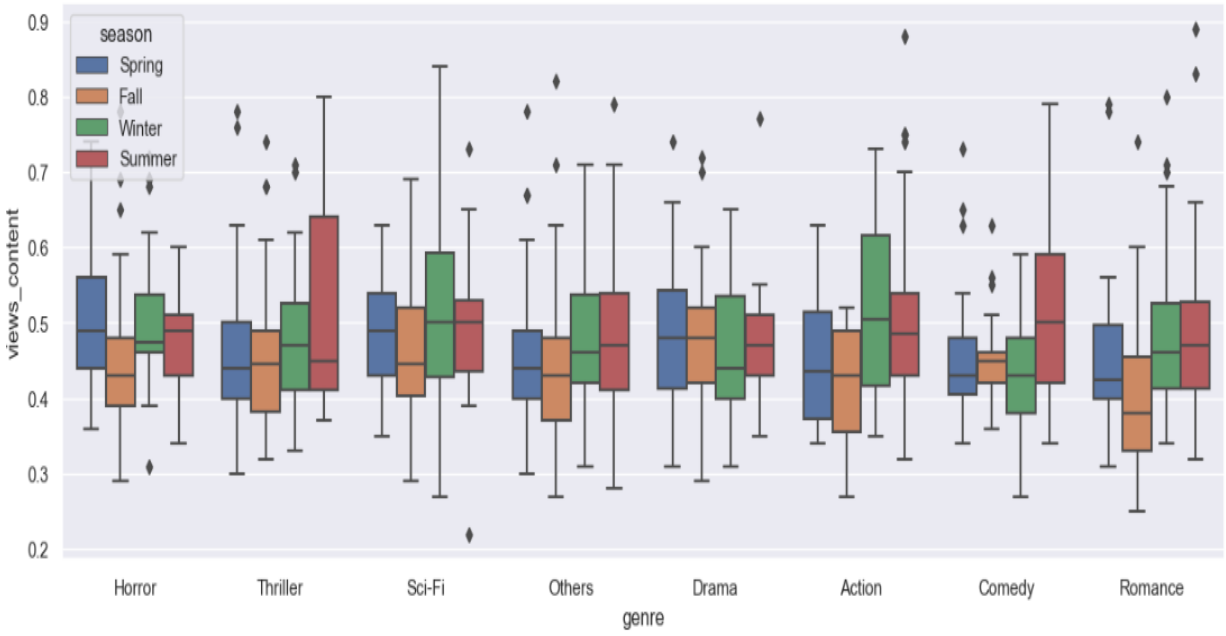


Fig 21: Boxplot for Genre, View Content and Season

- Overall, Action content has high views on winter season comparatively

5.3.3. Relationship among Ad_Impression, Major Sports Event and Season:

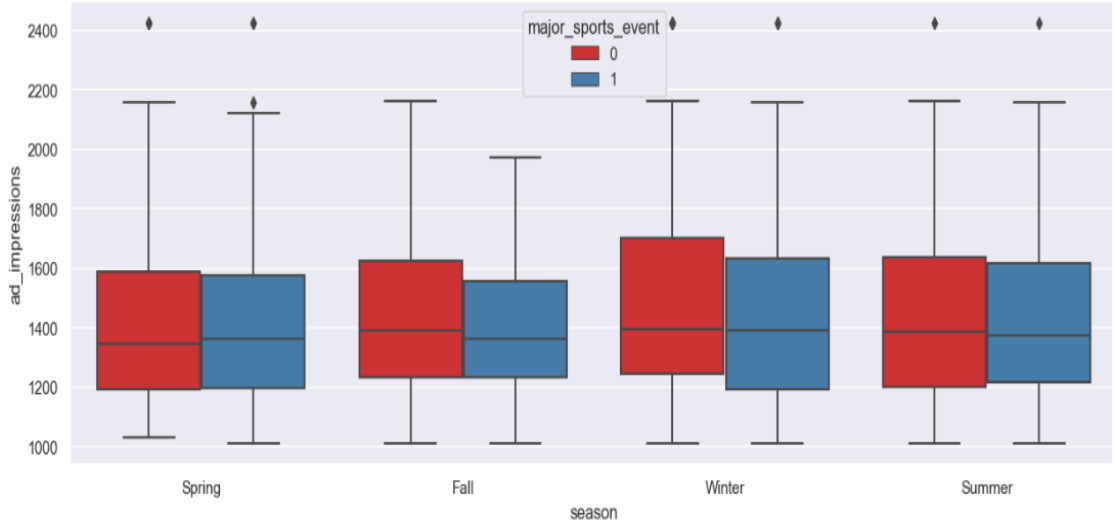


Fig 22: Boxplot for ad-impression, Major sports event and Season

- Ad-impression has less views on all the season during major-sport-event except in spring season.

5.3.4 Relationship among Genre, Major Sports Event and Views_Content:

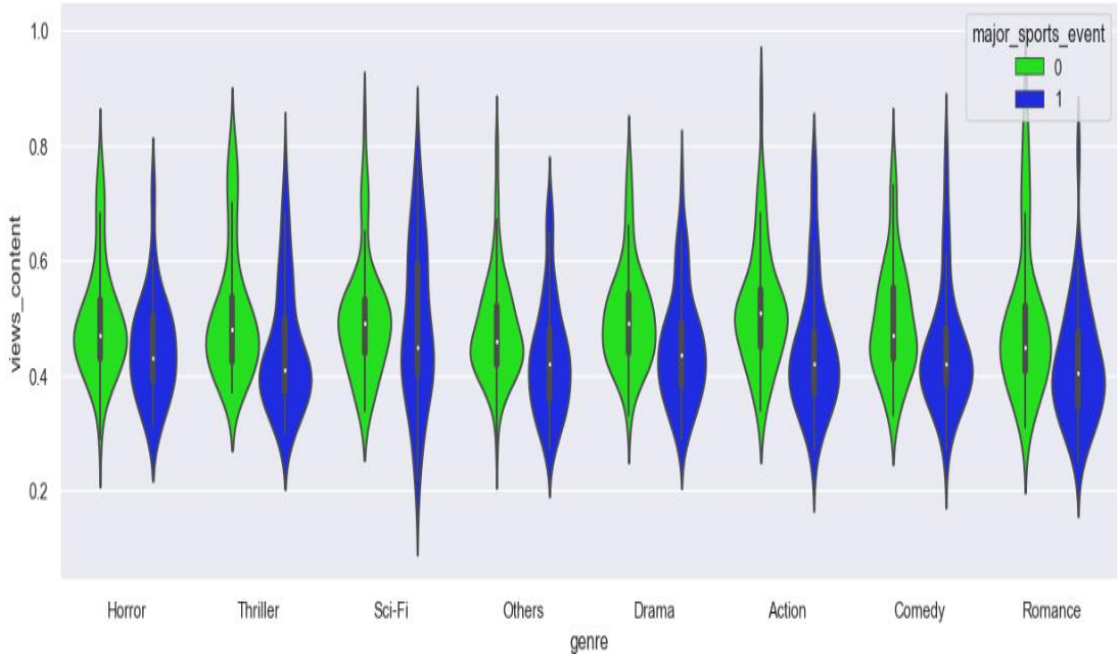


Fig 23: Violin plot for Genre, Major Sports Event and Views Content

- All the genre has less viewership on content during the major-sports event.

5.3.5. Relationship among Views Trailer, Day of Week and Views_Content:

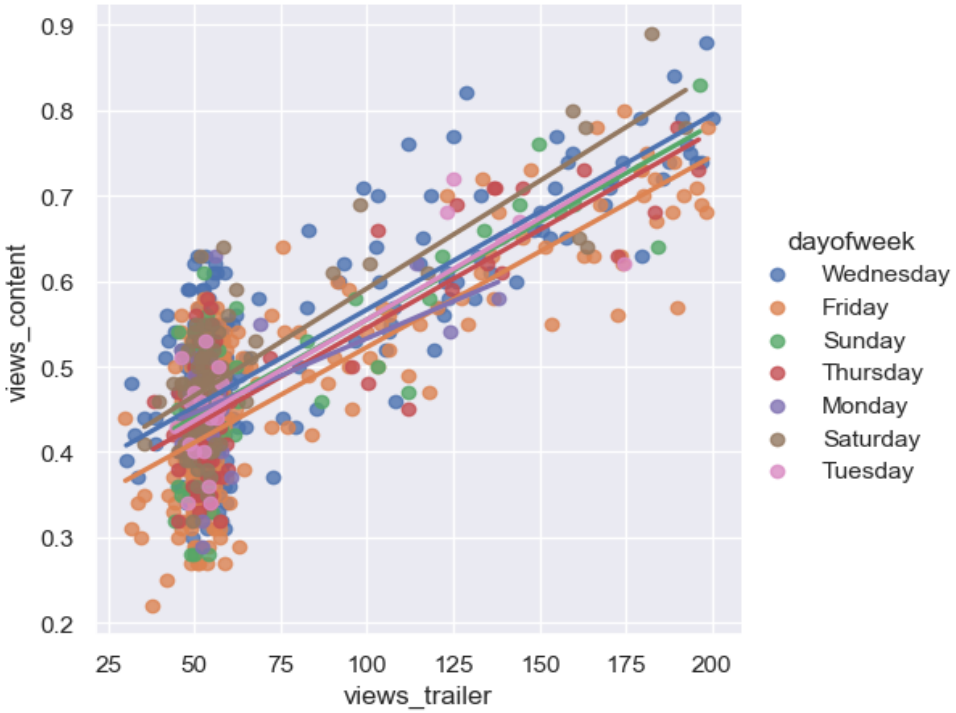


Fig 24: Content vs Trailer Views w.r.t. day of content release

- Content viewership is positively correlated with trailer viewership.
- On Saturdays, there is a higher number of content viewers compared to other weekdays.
- Interestingly, Sundays show relatively low viewers, which could be further analyzed by examining the data for different hours of the day.

5.3.6. Relationship among Visitor, Major Sports Event and Season:

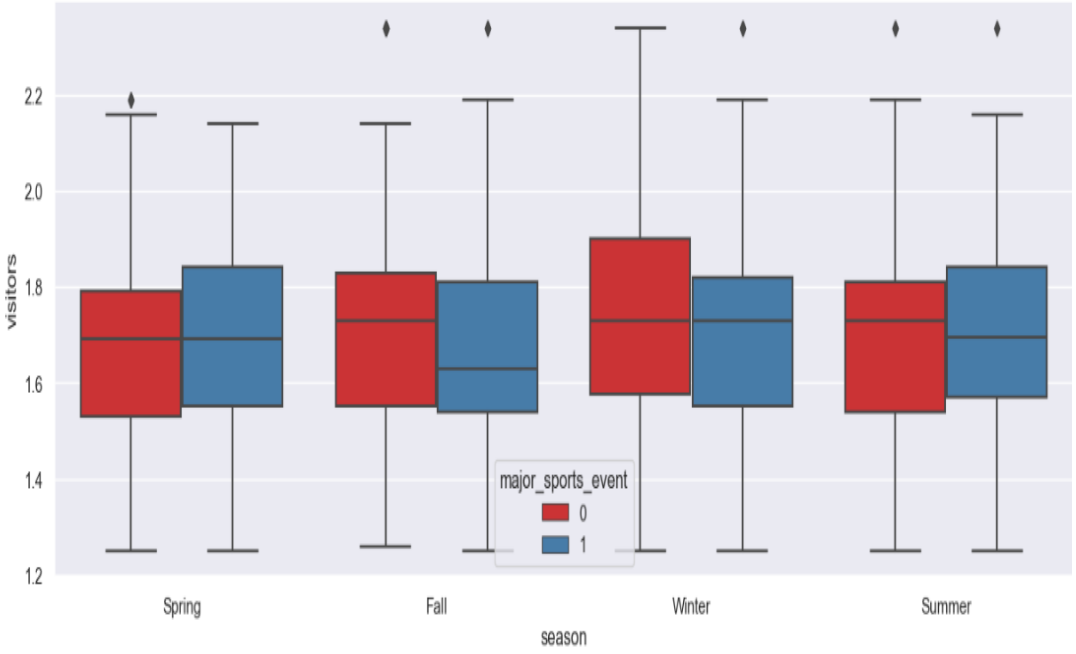


Fig 25: Boxplot for Visitor, Major Sports Event and Season

- Fall and summer season have less visitors during major sports event.
- For other seasons, there is no much variation due to major sports event.

5.3.7. Relationship among Visitors, Major Sports Event and Genre:

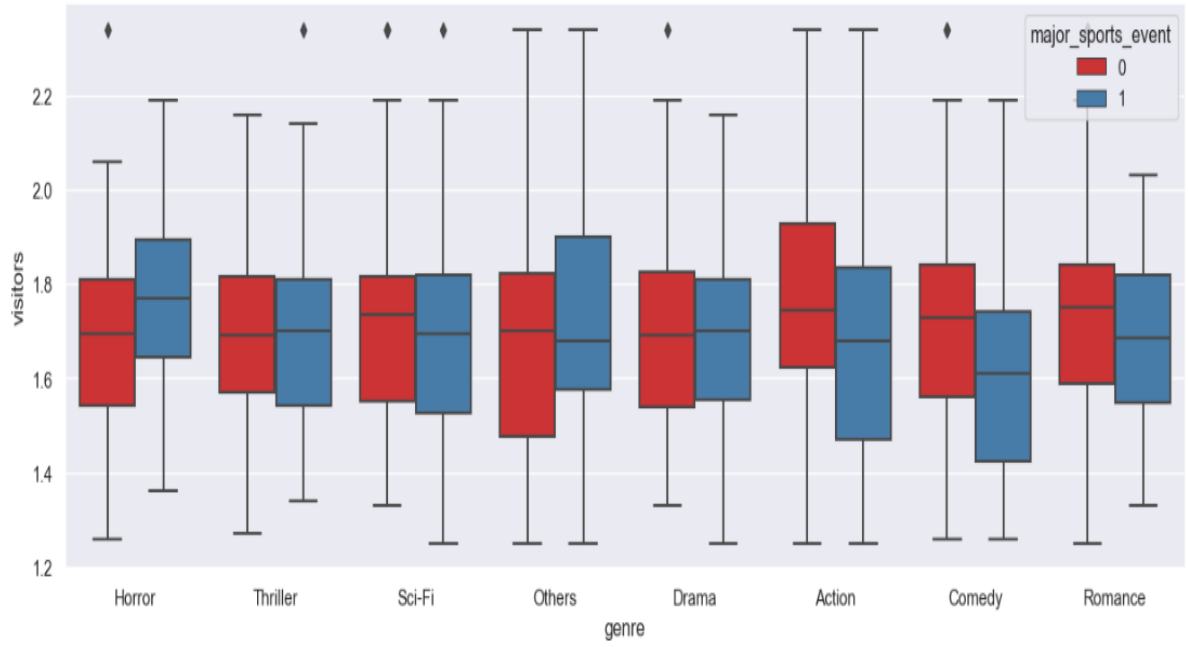


Fig 26: Boxplot for Visitor, Major Sports Event and Genre

- Horror content has high visitors irrespective of major sports event.
- Other genre has less visitors during major sports event.

5.3.8. Heat Map for Numeric Data:

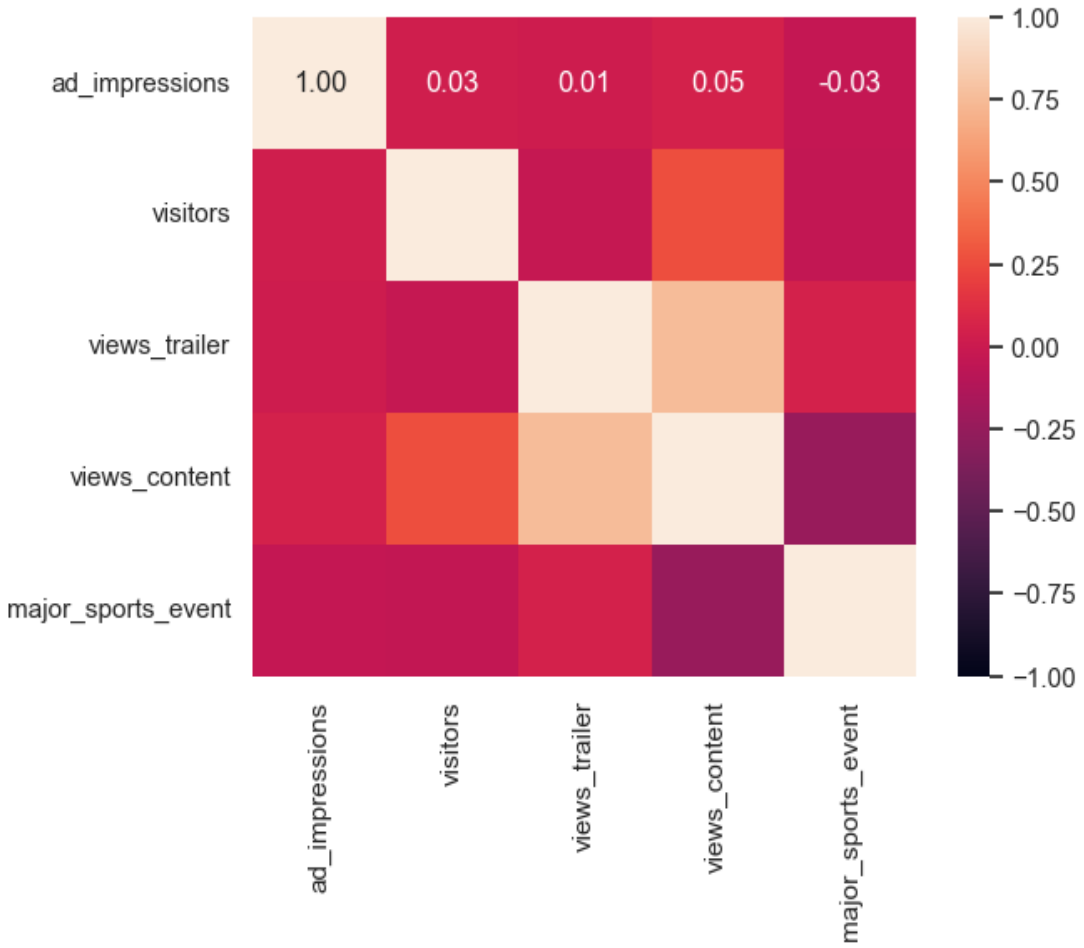


Fig 27: Heat Map for Numeric Data

- High Positive correlation can be observed between Views content and views trailer.
- A negative correlation can be observed between major sprots event and content views.
- A negative correlation can be observed between major sprots event and views-trailer.
- A negative correlation can be observed between major sprots event and ad-impression

5.3.9. Pair plot of Numeric Data

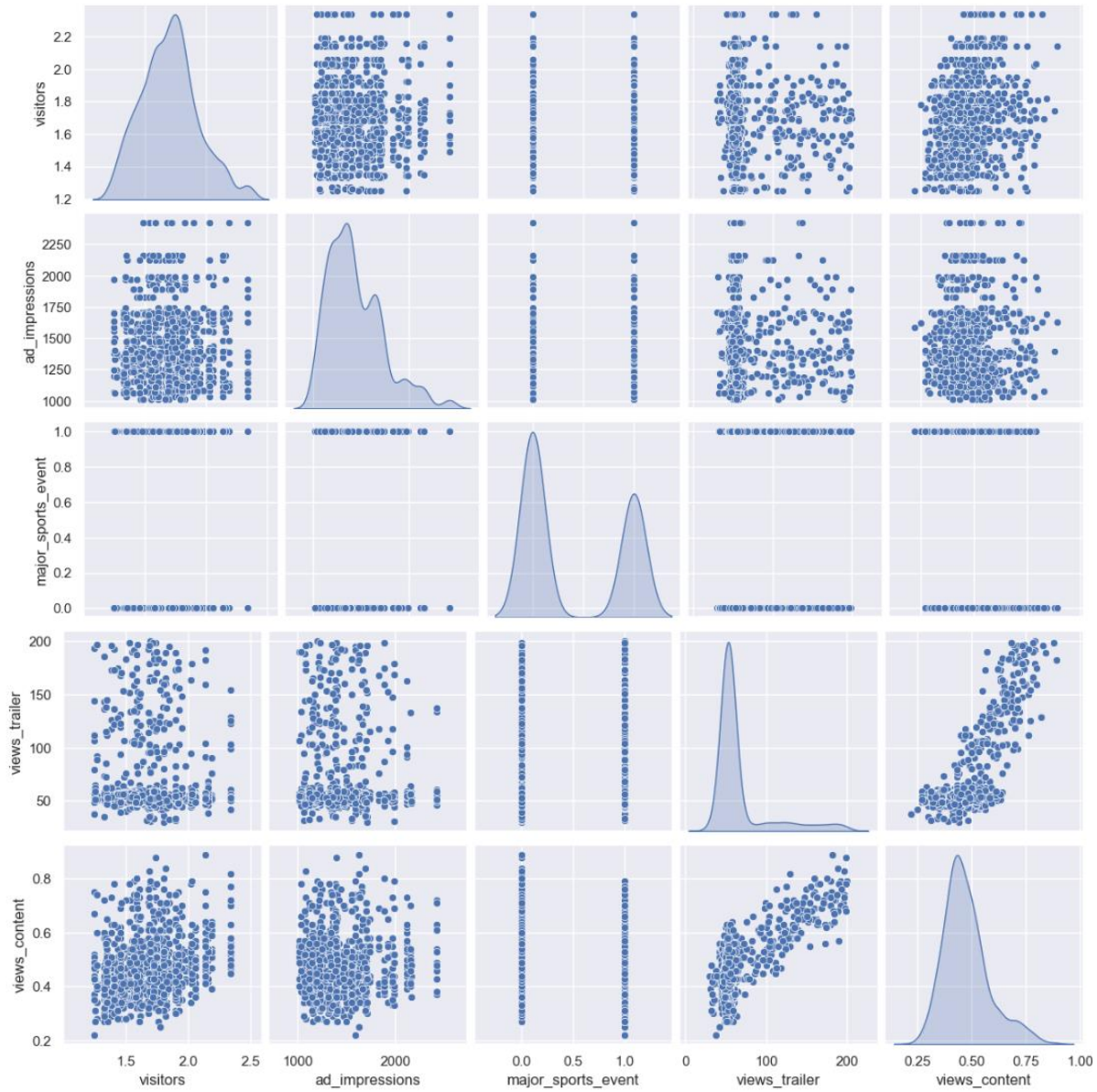


Fig 28: Pair Plot for Numeric Data

6. QUESTION AND ANSWER:

Question 1: What does the distribution of content views look like?

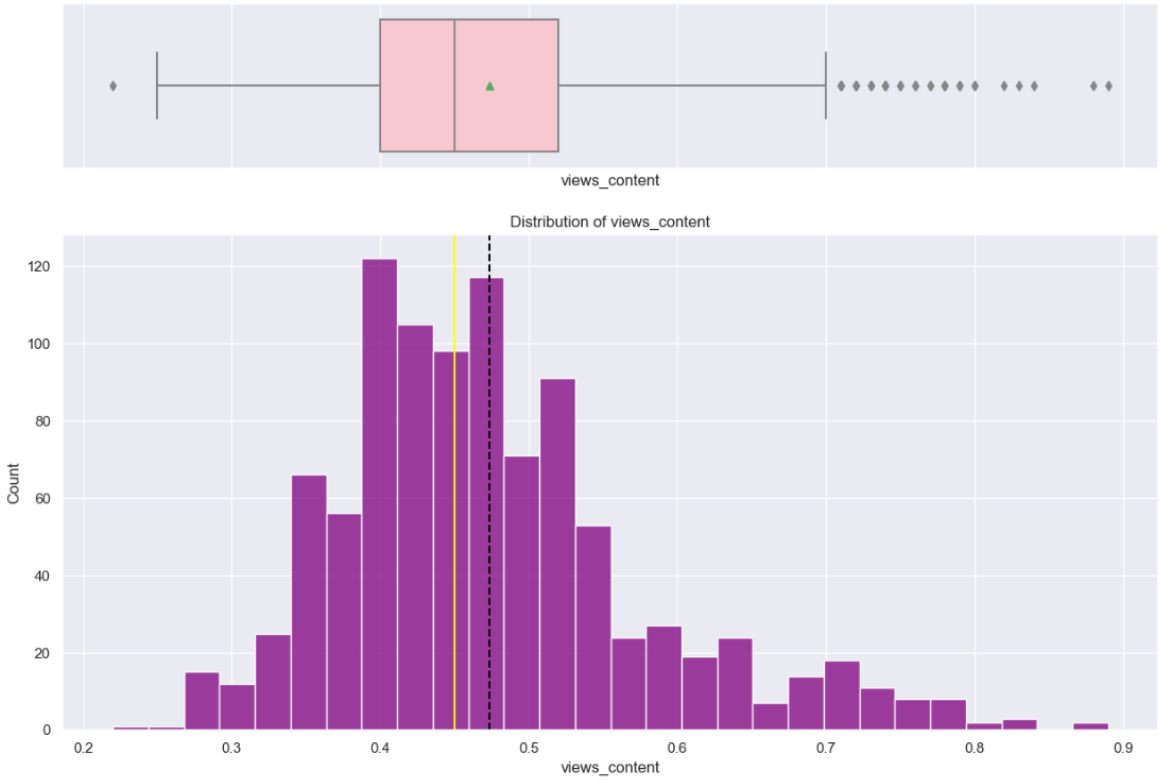


Fig 29: Distribution of Content Views

Observation:

Box Plot Analysis

- **Median**: The median of **views_content** is approximately 0.45, indicating that half of the observations have values below this point.
- **Interquartile Range (IQR)**: The box (spanning from Q1 to Q3) shows that the middle 50% of the data falls between approximately 0.4 and 0.52.
- **Whiskers and Outliers**:
- The lower whisker extends to approximately 0.22, indicating the minimum non-outlier value.
- The upper whisker extends to around 0.7, but there are several outliers beyond this point, reaching up to 0.89.
- **Mean**: The green triangle indicates the mean, which is slightly higher than the median, suggesting a right-skewed distribution.

Histogram Analysis

- **Distribution Shape**: The histogram reveals a right-skewed distribution, with a higher frequency of lower values and a tail extending towards higher values.
- **Peak**: The mode of views_content is around 0.45, corresponding with the median in the box plot.

- **Spread**: The data ranges from approximately 0.2 to 0.9, with the highest concentration of values between 0.4 and 0.5.
- **Frequency**: The highest bar represents the most frequent value range, confirming that most content views cluster around the median.
- Average less than 0.5 million watched the actual content

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

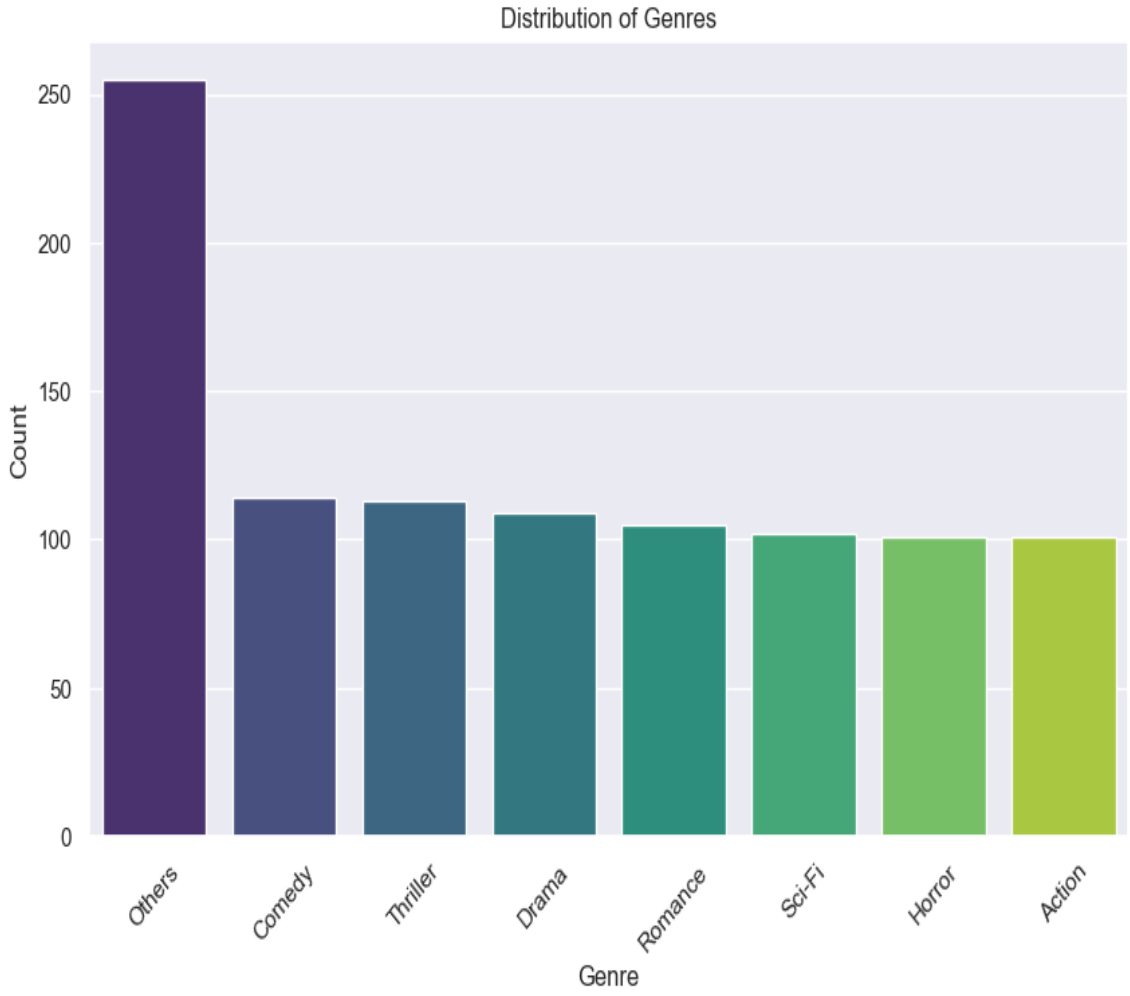


Fig 30: Distribution of Genre

- From this plot we can say, Genre others have the highest distribution.
- Genre others is almost one fourth of total genre.

Question 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?

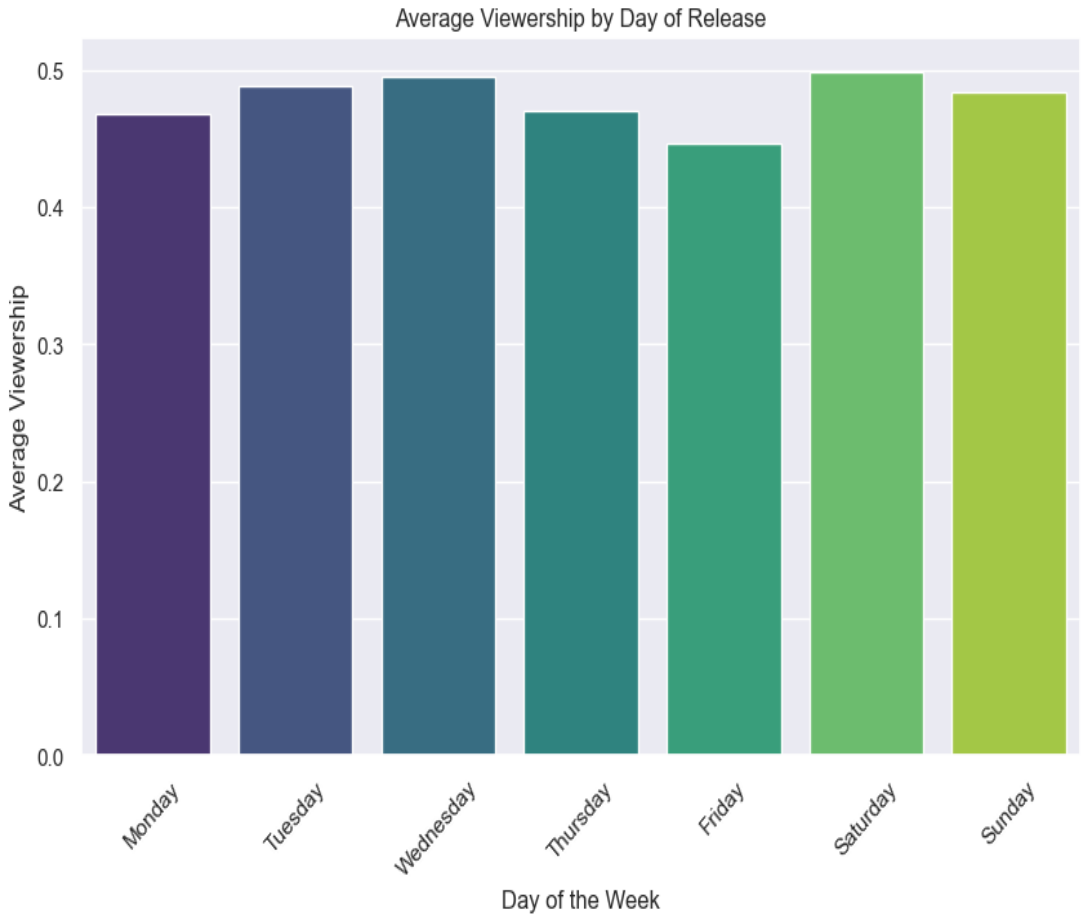


Fig 31: Average viewership by day of release

- The mean content views decrease for Friday content release with respect to other days.
- Saturday has the highest average viewership nearly 0.5.
- Wednesday is the second highest.

Question 4: How does the viewership vary with the season of release?

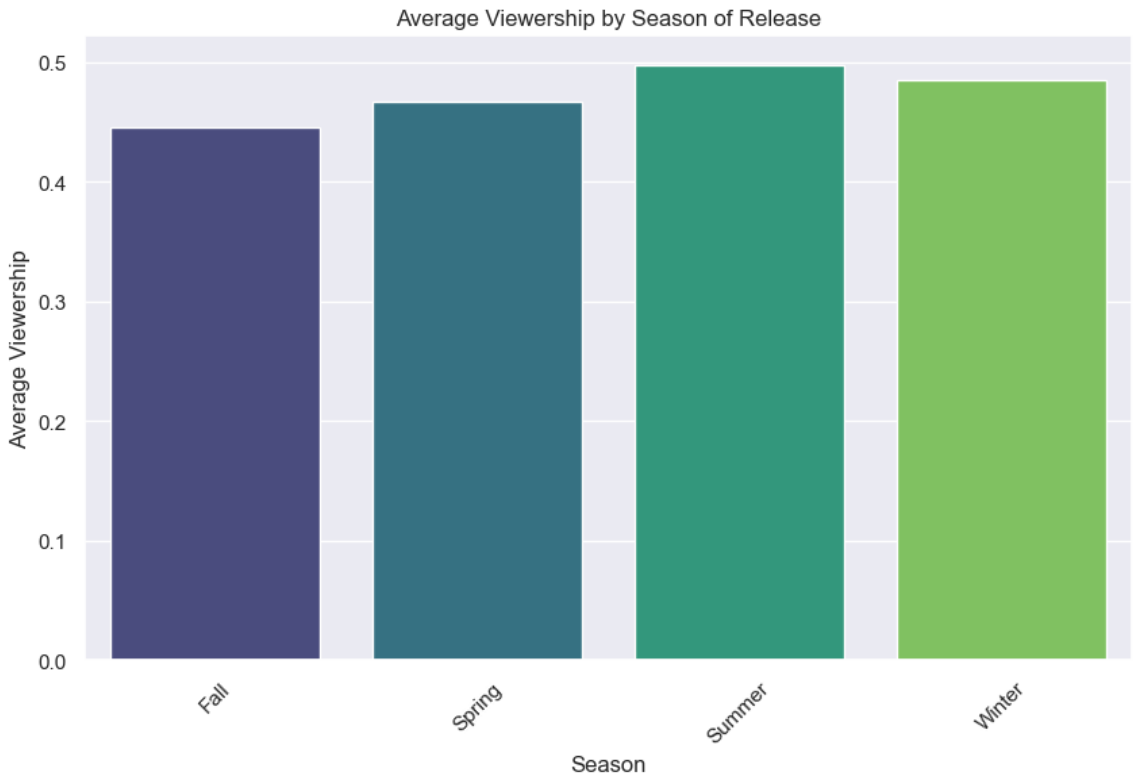


Fig 32: Average viewership by season of release

- From this plot we can say, summer has the highest average viewership nearly 0.5, while Fall has the lowest which is 0.45 average viewership.
- In the winter and summer the mean content views increases than the fall and spring season.
- Average content views for all the seasons looks in between 0.43 to 0.5 millions.

Question 5: What is the correlation between trailer views and content views?

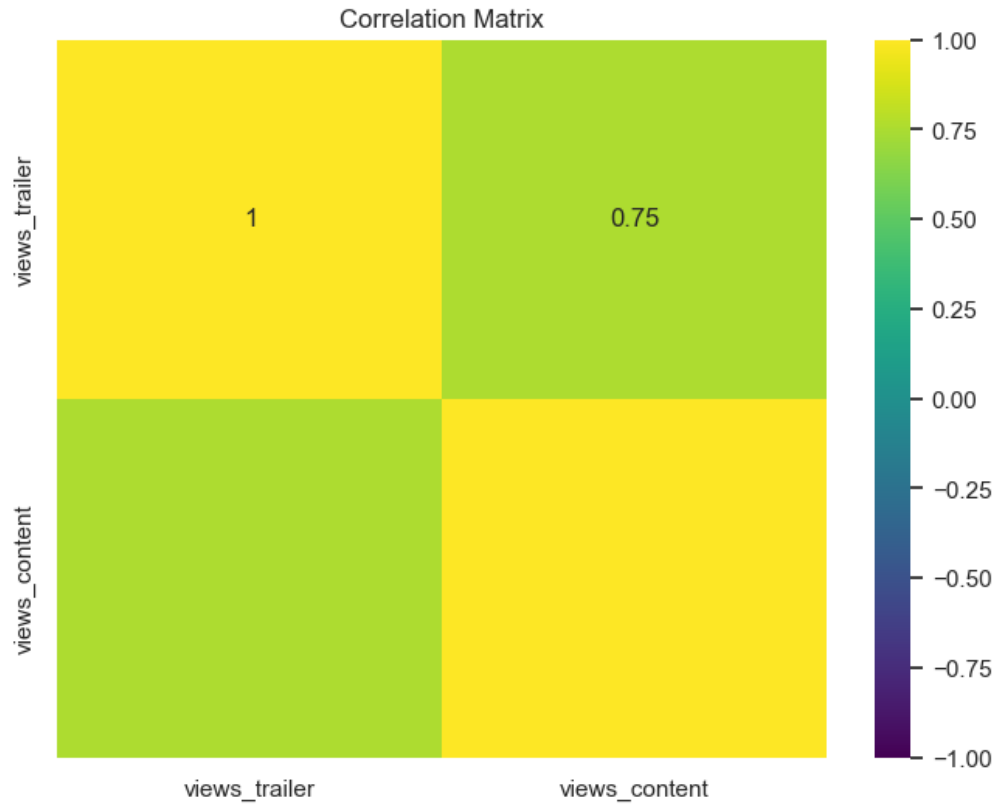


Fig 33: correlation between trailer views and content views

- Trailer views and content views have a positive correlation.
- Correlation coefficient between trailer views and content views: 0.75

7. DATA PREPROCESSING

- There are no duplicated values in the data
- There are no missing values in the data
- Outliers are detected.

7.1 Outlier Treatment

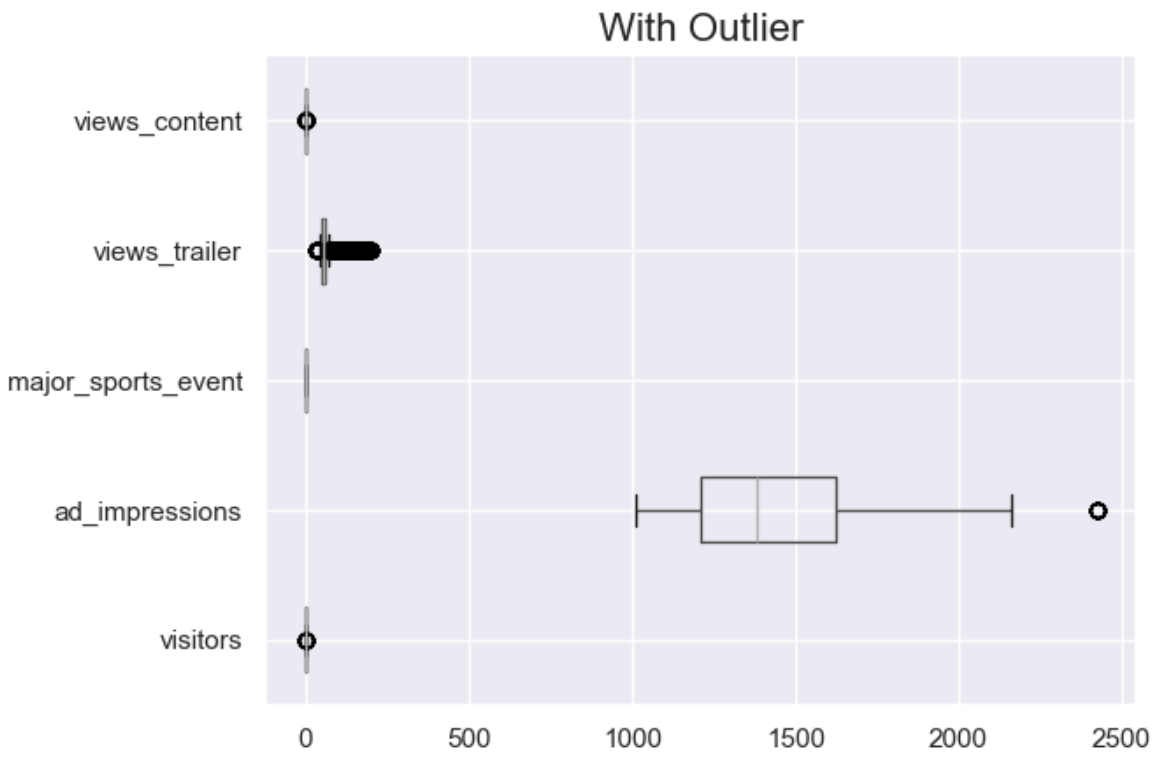


Fig 34: Before outlier Treatment

AFTER REMOVAL OF OUTLIERS

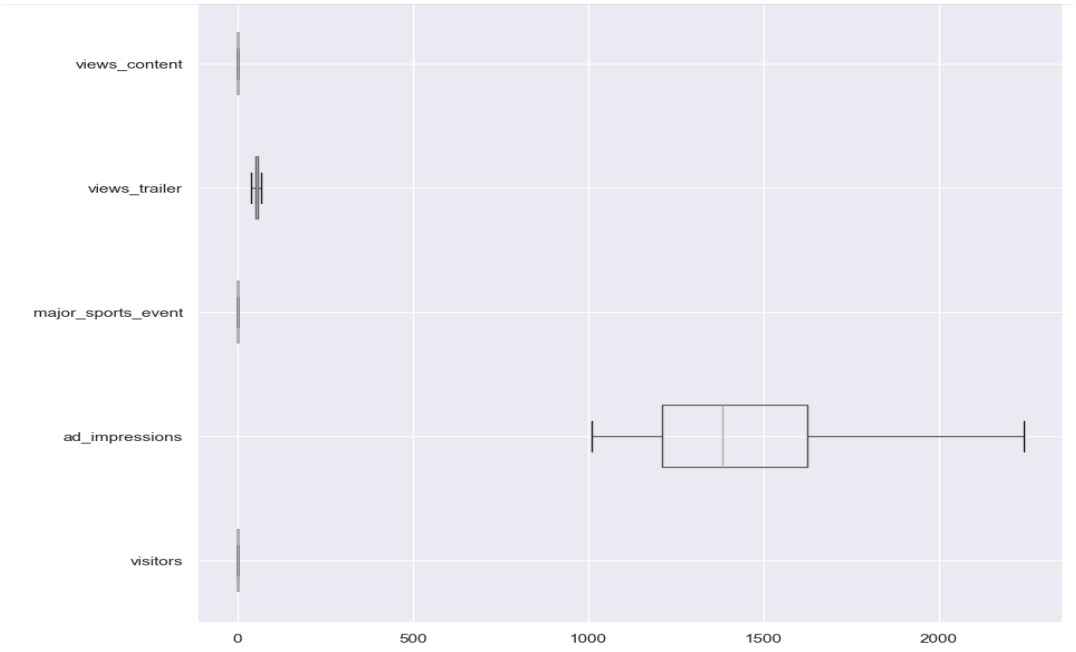


Fig 35: After outlier Treatment

7.2. Feature Engineering

- For the column major_sports_event, replace the 1 value with 'yes' and 0 value with 'no'

	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content
0	1.67	1113.81	no	Horror	Wednesday	Spring	56.70	0.51
1	1.46	1498.41	yes	Thriller	Friday	Fall	52.69	0.32
2	1.47	1079.19	yes	Thriller	Wednesday	Fall	48.74	0.39
3	1.85	1342.77	yes	Sci-Fi	Friday	Fall	49.81	0.44
4	1.46	1498.41	no	Sci-Fi	Sunday	Winter	55.83	0.46

Fig 36: Replacement of yes and no in column Major sports event

- For the column major_sports_event, replaced the 1 value with '***yes***' and 0 values with '***no***'.
- Created dummy variables for the columns (major sports event, genre, season, day of week release of content) mainly having datatype object and category.
- Splatted the data y with content views and x with rest of the 7 columns.
- Splitting the data in 70:30 ratio for train to test data.
- Number of rows in train data = 700
- Number of rows in test data = 300

8. MODEL BUILDING - LINEAR REGRESSION

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.641
Model:	OLS	Adj. R-squared:	0.631
Method:	Least Squares	F-statistic:	60.73
Date:	Fri, 02 Aug 2024	Prob (F-statistic):	2.15e-136
Time:	16:41:32	Log-Likelihood:	986.21
No. Observations:	700	AIC:	-1930.
Df Residuals:	679	BIC:	-1835.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2730	0.029	-9.329	0.000	-0.330	-0.216
visitors	0.1161	0.010	11.818	0.000	0.097	0.135
ad_impressions	6.227e-06	8.26e-06	0.754	0.451	-9.99e-06	2.24e-05
views_trailer	0.0093	0.000	27.084	0.000	0.009	0.010
major_sports_event_yes	-0.0619	0.005	-12.846	0.000	-0.071	-0.052
genre_Comedy	-0.0017	0.010	-0.177	0.860	-0.021	0.017
genre_Drama	-0.0039	0.010	-0.397	0.692	-0.023	0.015
genre_Horror	-0.0019	0.010	-0.192	0.848	-0.021	0.018
genre_Others	-0.0045	0.009	-0.520	0.603	-0.021	0.012
genre_Romance	-0.0106	0.010	-1.032	0.302	-0.031	0.010
genre_Sci-Fi	0.0158	0.010	1.574	0.116	-0.004	0.035
genre_Thriller	6.335e-05	0.010	0.006	0.995	-0.019	0.019
dayofweek_Monday	0.0241	0.014	1.673	0.095	-0.004	0.052
dayofweek_Saturday	0.0532	0.009	6.099	0.000	0.036	0.070
dayofweek_Sunday	0.0383	0.010	4.016	0.000	0.020	0.057
dayofweek_Thursday	0.0190	0.008	2.311	0.021	0.003	0.035
dayofweek_Tuesday	0.0473	0.017	2.835	0.005	0.015	0.080
dayofweek_Wednesday	0.0412	0.005	7.515	0.000	0.030	0.052
season_Spring	0.0272	0.007	4.172	0.000	0.014	0.040
season_Summer	0.0433	0.007	6.528	0.000	0.030	0.056
season_Winter	0.0297	0.006	4.569	0.000	0.017	0.042

Omnibus:	6.645	Durbin-Watson:	1.988
Prob(Omnibus):	0.036	Jarque-Bera (JB):	6.559
Skew:	0.234	Prob(JB):	0.0376
Kurtosis:	3.075	Cond. No.	1.97e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.97e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Fig 37: OLS Model with Multicollinearity and high p-values

Interpretation of Regression results:

1. Adjusted. R-squared: It reflects the fit of the model.

- Adjusted R-squared values generally range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
- In our case, the value for adj. R-squared is 0.641, which is good.

2. Const coefficient: It is the Y-intercept.

- It means that if all the predictor variable coefficients are zero, then the expected output (i.e., Y) would be equal to the const coefficient.
- In our case, the value for const coefficient is -0.2730

3. Coefficient of a predictor variable: It represents the change in the output Y due to a change in the predictor variable (everything else held constant).

- In our case, the coefficient of Visitors is 0.1161.

Interpretation of Coefficients

- The coefficients tell us how one unit change in X can affect y.
- The sign of the coefficient indicates if the relationship is positive or negative.
- In this data set, for example, presence of sports event on same day occurs with a 0.0619 decrease in viewership, and increase in visitors by single person occurse with a 0.1161 increase in the content viewership.
- Earlier in the heat map, we observed that the relationship between major sports events and content viewership is negatively correlated (as sports events increase, content viewership decreases, and vice versa), while the relationship between the number of visitors and content viewership is positively correlated (as the number of visitors increases, content viewership also increases, and vice versa). Consequently, the signs of the coefficients align with these relationships, suggesting a low probability of multicollinearity in our data.
- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

Interpretation of p-values (P > |t|)

For each predictor variable there is a null hypothesis and alternate hypothesis.

Null hypothesis: Predictor variable is not significant Alternate hypothesis: Predictor variable is significant (P > |t|) gives the p-value for each predictor variable to check the null hypothesis.

If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.

However, due to the presence of multicollinearity in our data, the p-values will also change.

We need to ensure that there is no multicollinearity in order to interpret the p-values.

How to check for Multicollinearity

- There are different ways of detecting (or testing) multicollinearity. One such way is Variation Inflation Factor.

- Variance Inflation factor: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient Bk is "inflated" by the existence of correlation among the predictor variables in the model.

General Rule of Thumb:

- If VIF is 1, then there is no correlation among the kth predictor and the remaining predictor variables, and hence, the variance of Bk is not inflated at all.
- If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.
- The purpose of the analysis should dictate which threshold to use.

Model Performance Check

Let's check the performance of the model using different metrics.

- We will be using metric functions defined in sklearn for RMSE, MAE, and R - Squared
- We will define a function to calculate MAPE and adjusted R -Squared
- The mean absolute percentage error (MAPE) measures the accuracy of predictions as a percentage, and can be calculated as the average absolute percent error for each predicted value minus actual values divided by actual values. It works best if there are no extreme values in the data and none of the actual values are 0.
- We will create a function which will print out all the above metrics in one go.

PERFORMANCE	RMSE	MAE	R-squared	Adj. R -squared	MAPE
Training performance	0.059142	0.046708	0.64141	0.630303	10.199322
Testing performance	0.065309	0.052897	0.561854	0.528757	11.506957

Table 3: Training and Test Performance of existing variables

Observations

- The training R- squared is 0.64, so the model is not underfitting

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

- MAE suggests that the model can predict view_contents within a mean error of 0.05 on the test data

- MAPE of 11.50 on the test data means that we are able to predict within 11.50% of the viewership of content.

9. CHECKING LINEAR REGRESSION ASSUMPTIONS

We will be checking the following Linear Regression assumptions:

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

9.1 Test for Multicollinearity

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the correlation between variables is high, it can cause problems when we fit the model and interpret the results. When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.
- There are different ways of detecting (or testing) multicollinearity. One such way is by using the Variance Inflation Factor, or VIF.
- Variance Inflation Factor (VIF): Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.
- If VIF is 1, then there is no correlation among the k th predictor and the remaining predictor variables, and hence, the variance of βk is not inflated at all.
- General Rule of thumb:
- If VIF is between 1 and 5, then there is low multicollinearity.
- If VIF is between 5 and 10, we say there is moderate multicollinearity.
- If VIF is exceeding 10, it shows signs of high multicollinearity.

	feature	VIF
0	const	166.186883
1	visitors	1.026119
2	ad_impressions	1.029638
3	views_trailer	1.028166
4	major_sports_event_yes	1.070634
5	genre_Comedy	1.913797
6	genre_Drama	1.921646
7	genre_Horror	1.902867
8	genre_Others	2.567207
9	genre_Romance	1.753305
10	genre_Sci-Fi	1.864765
11	genre_Thriller	1.920255
12	dayofweek_Monday	1.063674
13	dayofweek_Saturday	1.155999
14	dayofweek_Sunday	1.150747
15	dayofweek_Thursday	1.169813
16	dayofweek_Tuesday	1.059785
17	dayofweek_Wednesday	1.318664
18	season_Spring	1.541006
19	season_Summer	1.566296
20	season_Winter	1.570520

Table 4: VIF (Variance Inflation factor) before Removal of Multicollinearity

We can see the VIF for genre_Others is more than 2. Let's drop that.

- On dropping 'genre_Others', There is no major changes in adj. R-squared and R-Squared.

```
const                       153.669019
visitors                      1.020698
ad_impressions                1.028897
views_trailer                 1.027870
major_sports_event_yes        1.070178
genre_Comedy                  1.204448
genre_Drama                   1.223159
genre_Horror                  1.205554
genre_Romance                 1.171698
genre_Sci-Fi                  1.205336
genre_Thriller                1.206277
dayofweek_Monday              1.063674
dayofweek_Saturday            1.155123
dayofweek_Sunday              1.150353
dayofweek_Thursday            1.169773
dayofweek_Tuesday             1.056174
dayofweek_Wednesday           1.317860
season_Spring                 1.540998
season_Summer                 1.542795
season_Winter                 1.568307
dtype: float64
```

Table 5: VIF (Variance Inflation factor) After Removal of Multicollinearity

- All the variables have VIF less than 2.
- We have dealt with multicollinearity in the data

Let's rebuild the model using the updated set of predictors variables

Now that we do not have multicollinearity in our data, the p-values of the coefficients have become reliable and we can remove the non-significant predictor variables all together having p value > 0.05

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.641
Model:                            OLS   Adj. R-squared:                  0.631
Method:                 Least Squares   F-statistic:                     63.98
Date:                Fri, 02 Aug 2024   Prob (F-statistic):          3.02e-137
Time:                        16:41:33   Log-Likelihood:                 986.07
No. Observations:                 700   AIC:                            -1932.
Df Residuals:                     680   BIC:                            -1841.
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                     -0.2771      0.028     -9.855      0.000      -0.332      -0.222
visitors                   0.1165      0.010     11.894      0.000       0.097       0.136
ad_impressions          6.342e-06   8.25e-06      0.769      0.442   -9.86e-06    2.25e-05
views_trailer              0.0093      0.000     27.111      0.000       0.009       0.010
major_sports_event_yes    -0.0619      0.005    -12.845      0.000      -0.071      -0.052
genre_Comedy               0.0014      0.008      0.176      0.860      -0.014       0.016
genre_Drama               -0.0008      0.008     -0.105      0.917      -0.016       0.015
genre_Horror               0.0012      0.008      0.154      0.878      -0.014       0.017
genre_Romance             -0.0075      0.008     -0.897      0.370      -0.024       0.009
genre_Sci-Fi               0.0189      0.008      2.343      0.019       0.003       0.035
genre_Thriller             0.0032      0.008      0.408      0.683      -0.012       0.018
dayofweek_Monday           0.0241      0.014      1.674      0.095      -0.004       0.052
dayofweek_Saturday         0.0530      0.009      6.090      0.000       0.036       0.070
dayofweek_Sunday           0.0384      0.010      4.029      0.000       0.020       0.057
dayofweek_Thursday         0.0191      0.008      2.315      0.021       0.003       0.035
dayofweek_Tuesday          0.0468      0.017      2.811      0.005       0.014       0.080
dayofweek_Wednesday        0.0411      0.005      7.509      0.000       0.030       0.052
season_Spring              0.0272      0.007      4.173      0.000       0.014       0.040
season_Summer              0.0438      0.007      6.645      0.000       0.031       0.057
season_Winter              0.0298      0.006      4.594      0.000       0.017       0.043
==============================================================================
Omnibus:                        6.121   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.047   Jarque-Bera (JB):                6.025
Skew:                           0.225   Prob(JB):                       0.0492
Kurtosis:                       3.070   Cond. No.                     1.85e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.85e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig 38: OLS Model After Removal of Multicollinearity

Observations

- We can see that there is no change in adj. R-squared, which shows that the dropped columns did not have much effect on the model
- As there is no multicollinearity, we can look at the p-values of predictor variables to check their significance

Interpreting the Regression Results:

4. std err: It reflects the level of accuracy of the coefficients.

- The lower it is, the higher is the level of accuracy.

5. P>|t|: It is p-value.

- For each independent feature, there is a null hypothesis and an alternate hypothesis. Here Bi is the coefficient of the ith independent variable.

H0: Independent feature is not significant (Bi = 0) Ha : Independent feature is that it is significant (Bi != 0)

- (P>|t|) gives the p-value for each independent feature to check that null hypothesis. We are considering 0.05 (5%) as significance level.

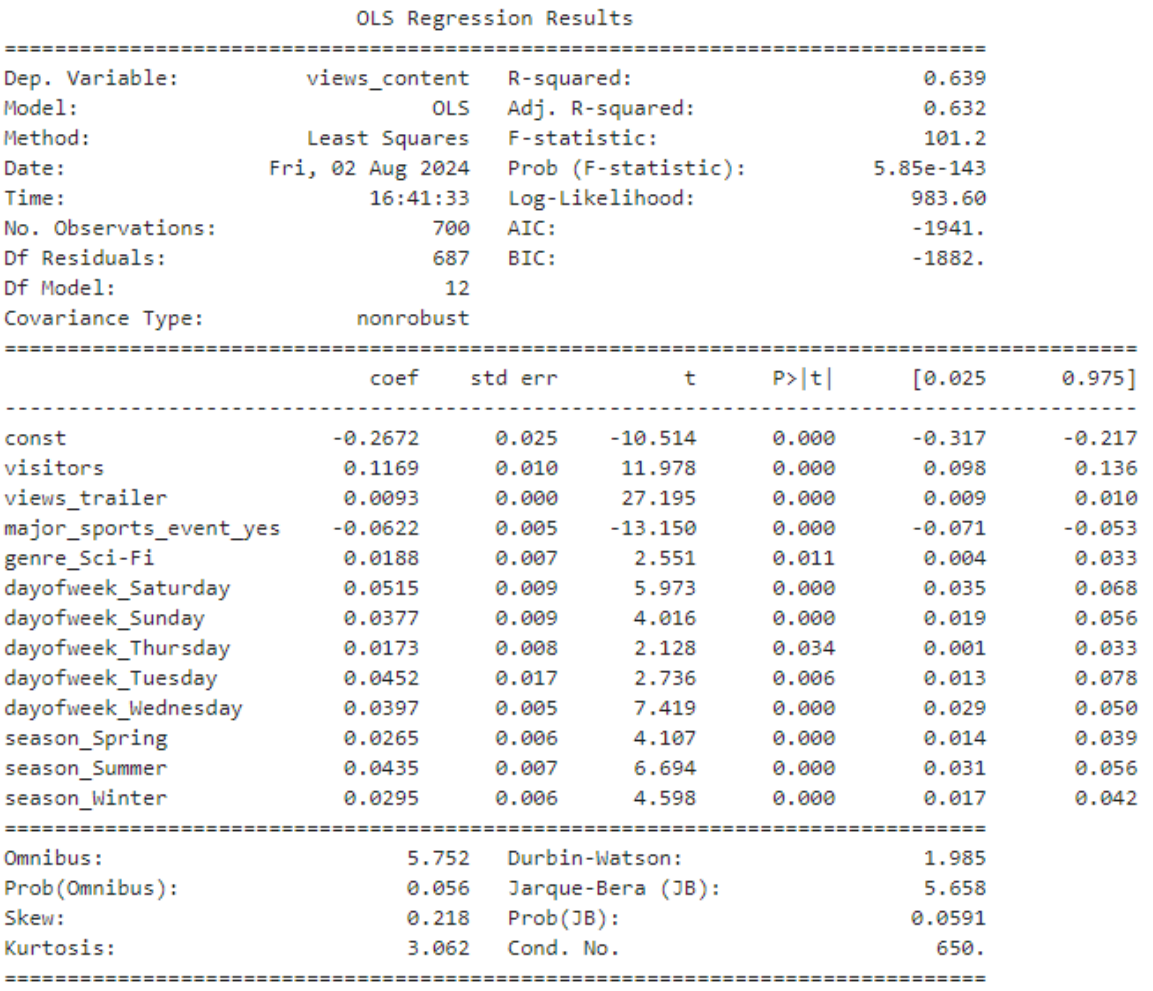
- A p-value of less than 0.05 is considered to be statistically significant.

6. Confidence Interval: It represents the range in which our coefficients are likely to fall (with a likelihood of 95%).

Dealing with high p-value variables

- Some of the dummy variables in the data have p-value > 0.05. So, they are not significant and we'll drop them
- But sometimes p-values change after dropping a variable. So, we'll not drop all variables at once
- Instead, we will do the following:
- Build a model, check the p-values of the variables, and drop the column with the highest p-value
- Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value
- Repeat the above two steps till there are no columns with p-value > 0.05

Note: The above process can also be done manually by picking one variable at a time that has a high p-value, dropping it, and building a model again. But that might be a little tedious and using a loop will be more efficient.



OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.639
Model:	OLS	Adj. R-squared:	0.632
Method:	Least Squares	F-statistic:	101.2
Date:	Fri, 02 Aug 2024	Prob (F-statistic):	5.85e-143
Time:	16:41:33	Log-Likelihood:	983.60
No. Observations:	700	AIC:	-1941.
Df Residuals:	687	BIC:	-1882.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2672	0.025	-10.514	0.000	-0.317	-0.217
visitors	0.1169	0.010	11.978	0.000	0.098	0.136
views_trailer	0.0093	0.000	27.195	0.000	0.009	0.010
major_sports_event_yes	-0.0622	0.005	-13.150	0.000	-0.071	-0.053
genre_Sci-Fi	0.0188	0.007	2.551	0.011	0.004	0.033
dayofweek_Saturday	0.0515	0.009	5.973	0.000	0.035	0.068
dayofweek_Sunday	0.0377	0.009	4.016	0.000	0.019	0.056
dayofweek_Thursday	0.0173	0.008	2.128	0.034	0.001	0.033
dayofweek_Tuesday	0.0452	0.017	2.736	0.006	0.013	0.078
dayofweek_Wednesday	0.0397	0.005	7.419	0.000	0.029	0.050
season_Spring	0.0265	0.006	4.107	0.000	0.014	0.039
season_Summer	0.0435	0.007	6.694	0.000	0.031	0.056
season_Winter	0.0295	0.006	4.598	0.000	0.017	0.042

Omnibus:	5.752	Durbin-Watson:	1.985
Prob(Omnibus):	0.056	Jarque-Bera (JB):	5.658
Skew:	0.218	Prob(JB):	0.0591
Kurtosis:	3.062	Cond. No.	650.

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

Fig 39: OLS Model without Multicollinearity and high p-values

PERFORMANCE	RMSE	MAE	R-squared	Adj. R -squared	MAPE
Training performance	0.059363	0.046772	0.638723	0.631877	10.214645
Testing performance	0.065566	0.053145	0.558408	0.538336	11.586675

Table 6: Training and Test Performance after removal of variables

Observations

- Now no feature has p-value greater than 0.05, so we'll consider the features in x_train2 as the final set of predictor variables and olsmod3 as the final model to move forward with

- Now adjusted R-squared is 0.632, i.e., our model is able to explain ~63% of the variance
- The adjusted R-squared in olsmod2 (where we considered the variables without multicollinearity) was 0.631
- This shows that the variables we dropped were not much affecting the model
- RMSE and MAE values are comparable for train and test sets, indicating that the model is not overfitting.

Now we'll check the rest of the assumptions on olsmod3.

2. Linearity of variables
3. Independence of error terms
4. Normality of error terms
5. No Heteroscedasticity

9.2. Linearity and Independence of Variables

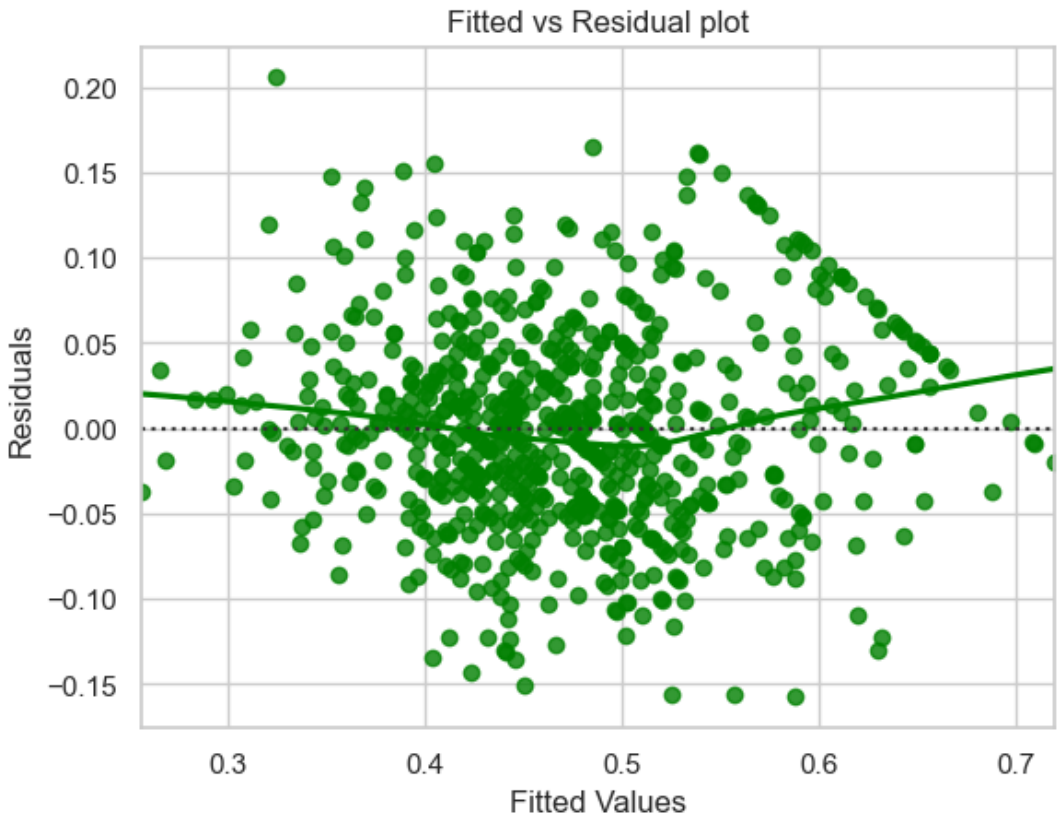


Fig 40: Linearity and Independence of Variables

We see no pattern in the plot above. Hence, the assumptions of linearity and independence are satisfied.

9.3. Normality of error terms

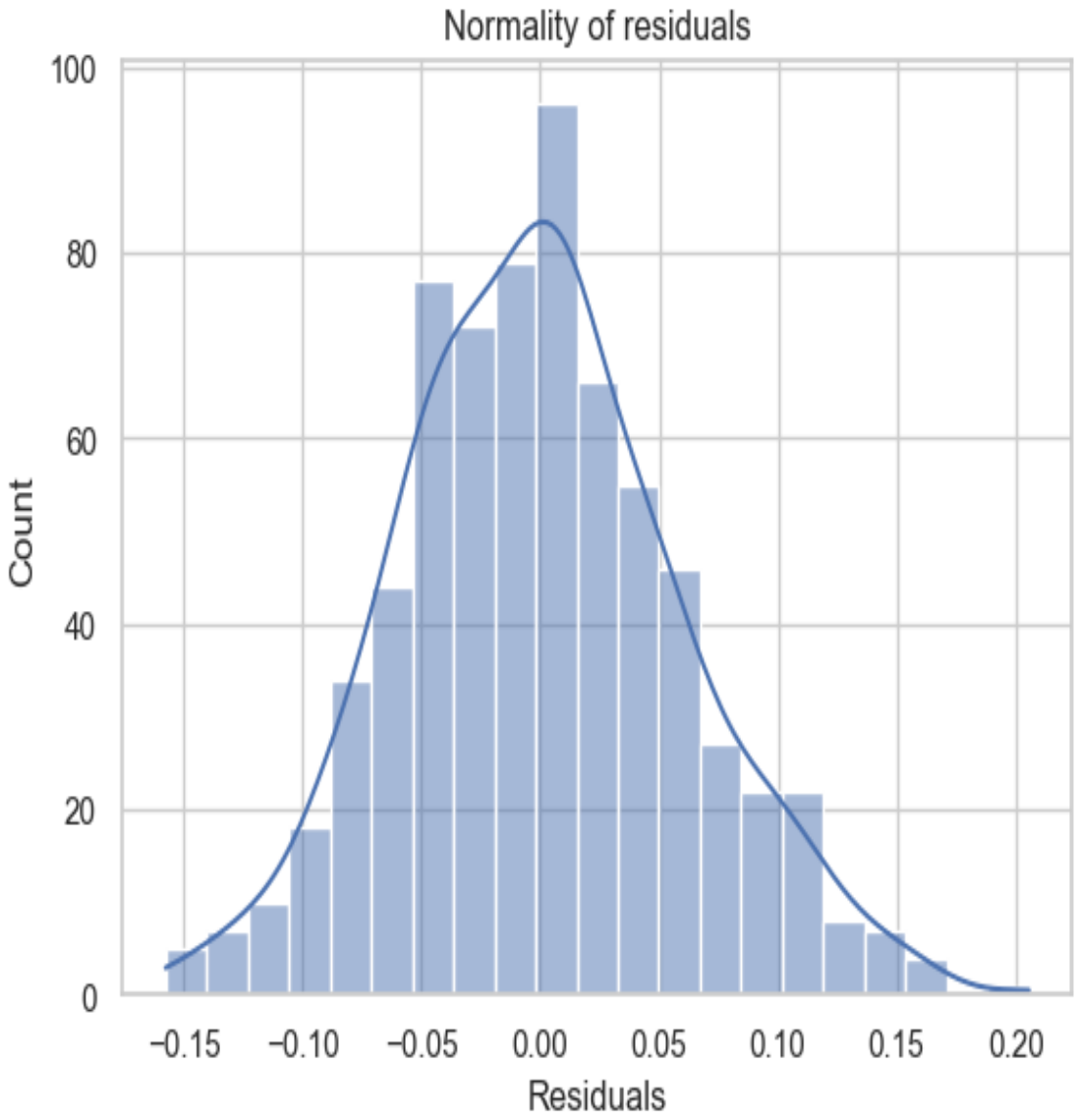


Fig 41: Normality of Residuals

- The residual terms are normally distributed

Q-Q PLOT:

The Q-Q plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

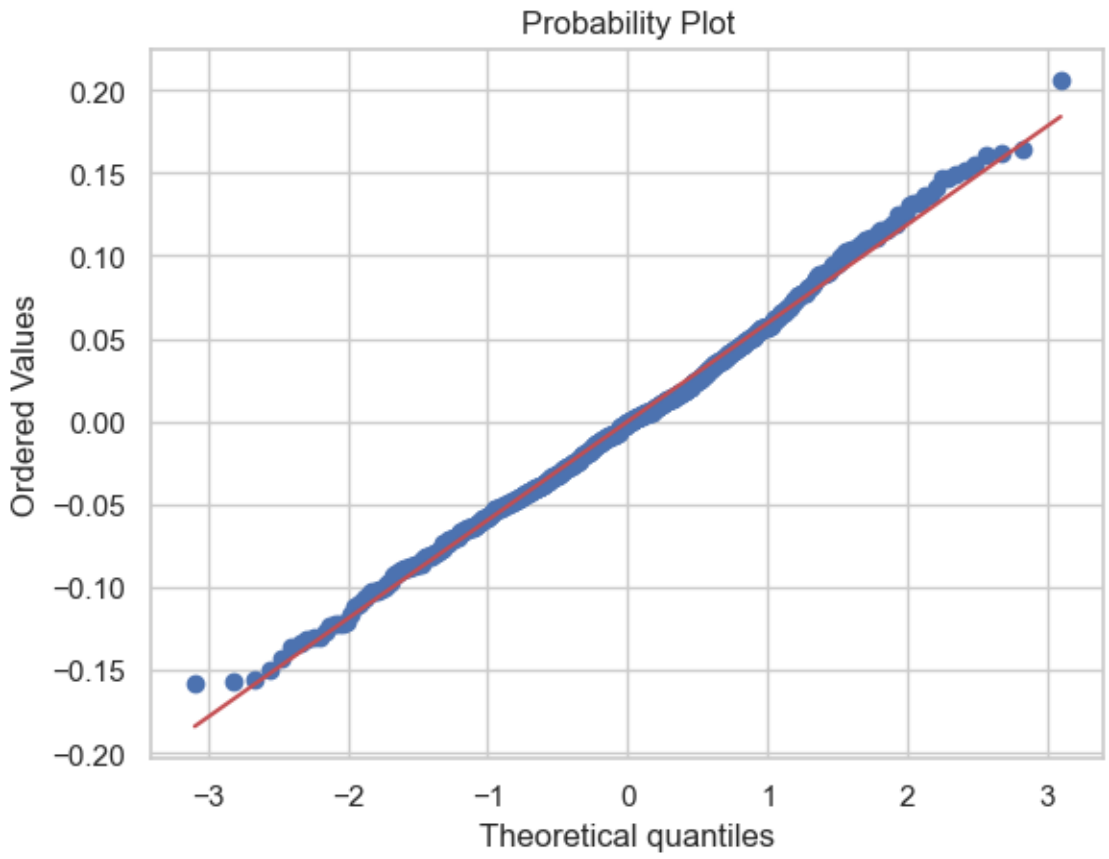


Fig 42: Q-Q plot of residuals

- Most of the points are lying on the straight line in Q-Q plot except the tails.

The **Shapiro-Wilk** test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

• Null hypothesis - Data is normally distributed.
• Alternate hypothesis - Data is not normally distributed.

```
ShapiroResult(statistic=0.9955450296401978, pvalue=0.04188501089811325)
```

• Since p-value < 0.05, the residuals are not normal as per Shapiro- wilk test.

- While the test statistic is close to 1, indicating that the data is near normal, the p-value suggests that deviations from normality are statistically significant just because of the outliers. This means we need to consider alternative methods or transformations if normality is a crucial assumption for the analysis or modelling. But here we have already done Q-Q plot of residuals, most of the points are lying on the straight line. Hence, we can say the data is normally distributed.

9.4. Test for Homoscedasticity

- Homoscedacity - If the variance of the residuals are symmetrically distributed across the regression line, then the data is said to homoscedastic.

The null and alternate hypotheses of the goldfeldquandt test are as follows:

- Null hypothesis: Residuals are homoscedastic
- Alternate hypothesis: Residuals have hetroscedasticity

```
[('F statistic', 1.1616405144400677), ('p-value', 0.08477906249012608)]
```

- Since p-value > 0.05 we can say that the residuals are homoscedastic.

10. PREDICTIONS ON TEST DATA

	Actual	Predicted
983	0.43	0.479335
194	0.51	0.615628
314	0.48	0.461251
429	0.41	0.496161
267	0.41	0.473426
746	0.68	0.520500
186	0.62	0.588667
964	0.48	0.496081
676	0.42	0.491408
320	0.58	0.605350

Table 7: Actual vs Predicted value of built model

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

11. FINAL MODEL

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.639
Model:	OLS	Adj. R-squared:	0.632
Method:	Least Squares	F-statistic:	101.2
Date:	Fri, 02 Aug 2024	Prob (F-statistic):	5.85e-143
Time:	17:18:34	Log-Likelihood:	983.60
No. Observations:	700	AIC:	-1941.
Df Residuals:	687	BIC:	-1882.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2672	0.025	-10.514	0.000	-0.317	-0.217
visitors	0.1169	0.010	11.978	0.000	0.098	0.136
views_trailer	0.0093	0.000	27.195	0.000	0.009	0.010
major_sports_event_yes	-0.0622	0.005	-13.150	0.000	-0.071	-0.053
genre_Sci-Fi	0.0188	0.007	2.551	0.011	0.004	0.033
dayofweek_Saturday	0.0515	0.009	5.973	0.000	0.035	0.068
dayofweek_Sunday	0.0377	0.009	4.016	0.000	0.019	0.056
dayofweek_Thursday	0.0173	0.008	2.128	0.034	0.001	0.033
dayofweek_Tuesday	0.0452	0.017	2.736	0.006	0.013	0.078
dayofweek_Wednesday	0.0397	0.005	7.419	0.000	0.029	0.050
season_Spring	0.0265	0.006	4.107	0.000	0.014	0.039
season_Summer	0.0435	0.007	6.694	0.000	0.031	0.056
season_Winter	0.0295	0.006	4.598	0.000	0.017	0.042

Omnibus:	5.752	Durbin-Watson:	1.985
Prob(Omnibus):	0.056	Jarque-Bera (JB):	5.658
Skew:	0.218	Prob(JB):	0.0591
Kurtosis:	3.062	Cond. No.	650.

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

Fig 43: Final OLS Model

PERFORMANCE	RMSE	MAE	R-squared	Adj. R -squared	MAPE
Training performance	0.059363	0.046772	0.638723	0.631877	10.214645
Testing performance	0.065566	0.053145	0.558408	0.538336	11.586675

Table 8: Training and Test Performance of Final Model

Observation:

- The model is able to explain ~63% 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 11.58% of the viewership of content.
- Hence, we can conclude the model olsmodel_final is good for prediction as well as inference purposes

12. ACTIONABLE INSIGHTS AND RECOMMENDATIONS

- The model's R-squared value is approximately 0.639, and the adjusted R-squared is 0.632, indicating that the model can explain about 63% of the variance in the data. This is quite satisfactory.
- This suggests that the model is suitable for both prediction and inference purposes.
- Major Sports Event: Negative coefficient (-0.0622) with a high t-value (-13.150) and a p-value of 0.000, indicating that the presence of major sports events negatively impacts first-day viewership.
- A major sports event will lead to a 0.0604 unit decrease in content viewership, assuming all other variables remain constant.
- To improve content viewership, it is recommended to avoid releasing content on days when major sports events are happening.
- Visitors: Positive coefficient (0.1169) with a very high t-value (11.978) and a p-value of 0.000, indicating a strong positive relationship with first-day viewership. More visitors to the platform lead to higher first-day viewership.
- An increase of one unit in the 'visitors' variable results in a 0.1169 unit increase in content viewership, with all other variables held constant.
- The client should provide more detailed information to identify the reasons behind increases in viewership.
- Views of Trailer: Positive coefficient (0.0093) with a high t-value (27.195) and a p-value of 0.000, indicating that more trailer views significantly increase first-day viewership.
- An increase of one unit in trailer viewers will result in a 0.0093 unit increase in content viewership, all other variables held constant.
- Genre (Sci-Fi): Positive coefficient (0.0188) with a t-value (2.551) and a p-value of 0.011, suggesting Sci-Fi content has a positive impact on first-day viewership.

- Releasing content on specific days of the week will increase viewership: Saturday (0.0515 units), Tuesday (0.0452 units), Wednesday (0.0397 units), Sunday (0.0377 units), and Thursday (0.0173 units), with all other variables held constant.

- Therefore, releasing content on Saturdays and Tuesday will boost viewership, provided no major sports events occur on those days.

- Seasons: Significant positive coefficients for Spring (0.0265), Summer (0.0435), and Winter (0.0295), suggesting these seasons see higher viewership compared to the baseline (probably Fall).

- The summer season can result in a 0.0435 unit increase in content viewership, with all other variables held constant.

- Releasing content during the summer season can enhance viewership.

12.1 Conclusion on predictor significance

The analysis reveals that all predictors included in the model are statistically significant, with p-values below the 0.05 threshold. This means that each of these variables has a meaningful impact on the first-day viewership of content on ShowTime's platform. The most influential factors include the number of visitors, views of the trailer, presence of major sports events, the Sci-Fi genre, specific days of the week, and certain seasons.

12.2 Key takeaways for the business

- Implement strategies to increase visitors, such as promotional offers, partnerships, improved user experience, and personalized recommendations.
- Invest in marketing campaigns to promote trailers. Utilize social media, email marketing, and in-app notifications to increase trailer visibility.
- Schedule major content releases to avoid clashes with major sports events to minimize the negative impact on viewership.
- Focus on producing and promoting Sci-Fi content to attract a dedicated audience. Consider creating special Sci-Fi series or movies to leverage this trend.
- Plan major content releases on these days to maximize viewership. Avoid releasing content on less favourable days (likely Monday and Friday).
- Align major content releases with these seasons. Consider creating seasonal special content to attract more viewers during these times.
- Continuously monitor viewership data and adjust strategies based on real-time insights.