Predictive Modeling

ShowTime - Problem Statement for Data Science

By: Agnes Raja Kumari. E

PGP-Data Science and Business Analytics

PGPDSBA.O.MAY24.A

Contents

Context

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

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

With the outbreak of 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

Dataset Name: ottdata.csv

Understanding the Problem and Data

Problem:

- Identify key factors influencing first-day viewership for content on ShowTime.
- Build a linear regression model to predict first-day viewership based on these factors.

Data:

We have a dataset containing the following variables:

- **Dependent Variable:**
    - views_content
- **Independent Variables:**
    - visitors
    - ad_impressions

- major_sports_event
- genre
- dayofweek
- season
- views_trailer

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
```

Fig 1: Information of Data

Observation:

- The dataset contains 1000 entries.
- There are 8 columns with various data types including float64, int64, and object.
- All columns contain 1000 non-null entries, indicating no missing data.

Statistical Summary of the Data

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

Fig 2: Statistical Summary of the Data

Observation:

- Ad Impressions have a wide range, suggesting variability in marketing reach.
- Major Sports Event occurs 40% of the time, indicating a significant factor for visitor trends.
- Genre and Day of the Week show diverse distributions, with "Others" being the most frequent genre and Friday the most frequent day.
- Seasonal Trends: Winter is the most common season, which could correlate with higher or lower visitor numbers.
- Views (Trailer and Content) have a noticeable variation, which might influence visitor behaviour and engagement.

Exploratory Data Analysis (EDA)

Visualizing Data

Univariate Analysis

Visualizing views_content using boxplot and histplot

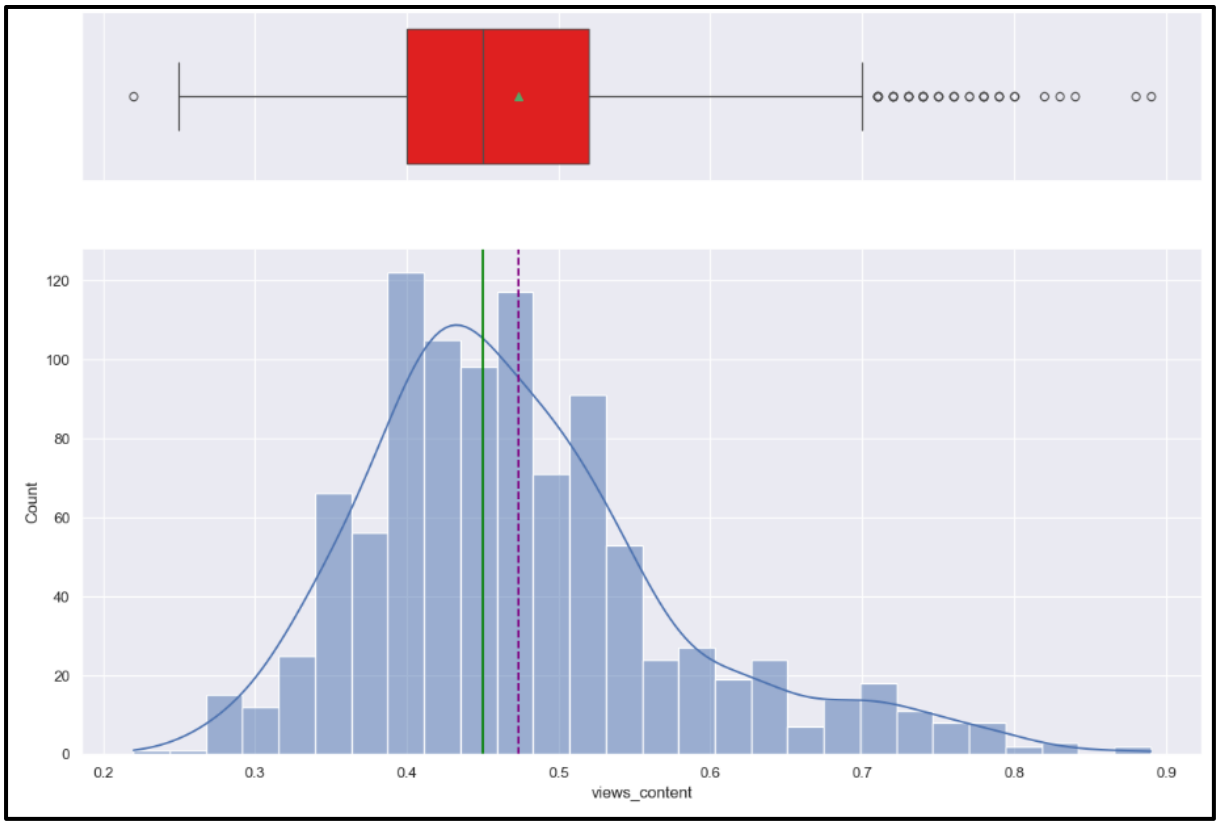

Fig 3: Visualizing views_content using boxplot and histplot

Observation:

Box Plot (Top Plot):

- The data is right-skewed, with the majority of observations clustered towards the lower end of the range.
- There are several outliers on the higher end, indicating some instances with significantly higher views_content values compared to the rest.

Histogram with KDE (Bottom Plot):

- The distribution of views_content is slightly right-skewed, with the bulk of the data concentrated around 0.4 to 0.5.
- The mean of the distribution is slightly above 0.5 (as indicated by the vertical dashed line).
- The distribution tapers off towards the right, showing fewer instances with higher views_content values.

Visualizing views_trailer using boxplot and histplot

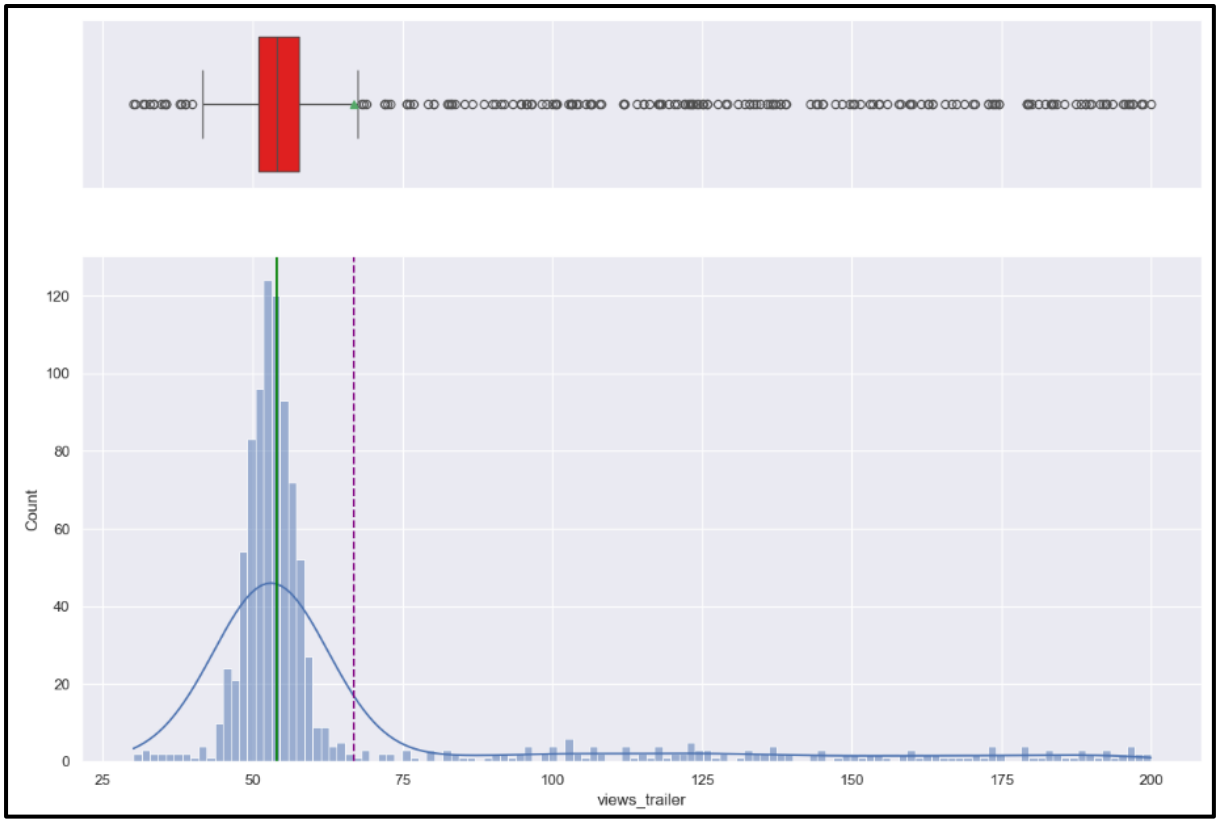

Fig 4: Visualizing views_trailer using boxplot and histplot

Observation:

Box Plot (Top Plot):

- The distribution of views_trailer shows significant right-skewness with a concentration of data points near the lower end and many outliers extending far to the right.
- The box is relatively narrow, indicating that most of the data is tightly clustered, but the numerous outliers suggest that there are many instances with much higher values.

Histogram with KDE (Bottom Plot):

- The histogram shows that the majority of views_trailer values are concentrated between 45 and 60, with a sharp peak around 50.
- The distribution has a long right tail, confirming the presence of outliers, and indicating that there are a few trailers with significantly higher views.
- The vertical dashed line indicates the mean of the distribution, which is higher than the mode, reflecting the impact of the outliers on the average.
- These insights suggest that while most trailers have a relatively small number of views clustered tightly together, there is a significant number of trailers with unusually high view counts, leading to a skewed distribution with many outliers.

Visualizing ad impressions using boxplot and histplot

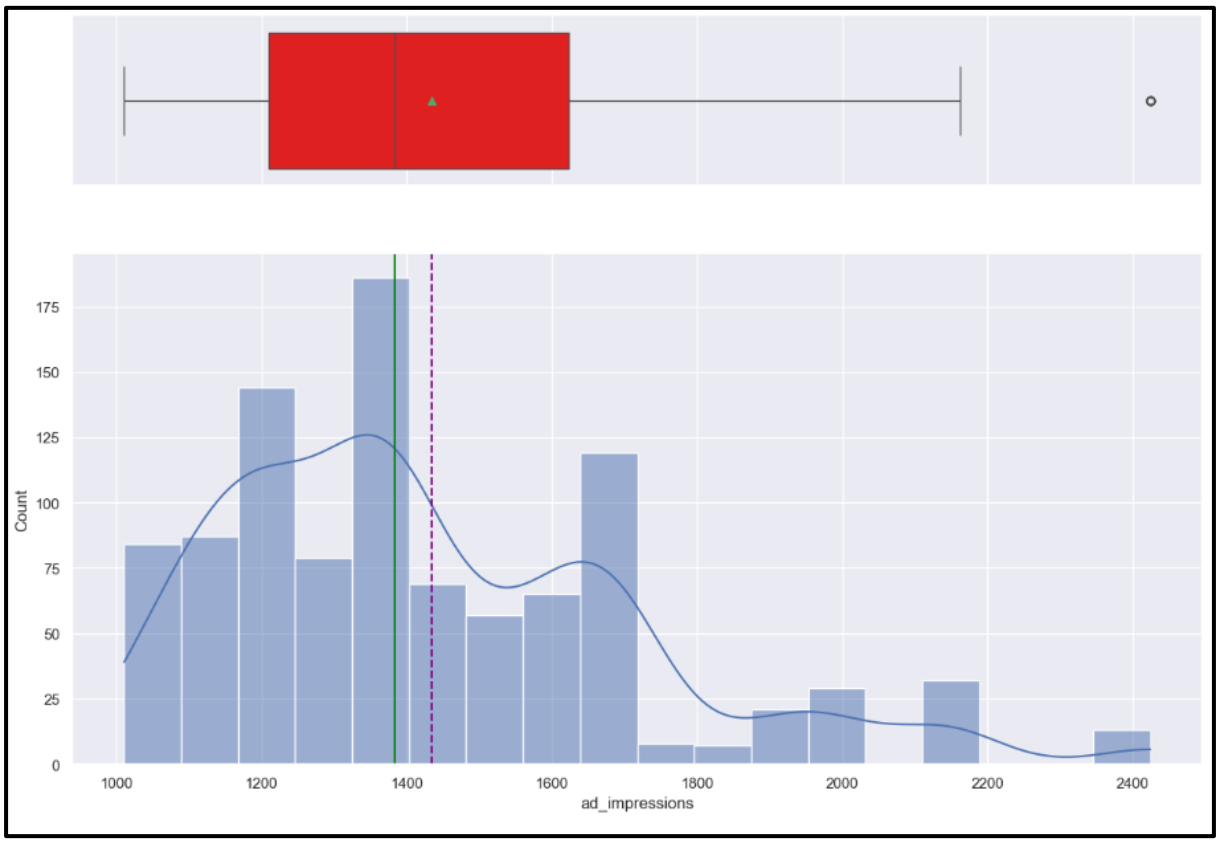

Fig 5: Visualizing ad impressions using boxplot and histplot

Observation:

Box Plot (Top Plot):

- The distribution is relatively symmetric with a wide range, and the box plot shows the central 50% of the data is spread out.
- There is one clear outlier on the higher end, indicating an instance with a significantly larger number of ad impressions compared to the rest.

Histogram with KDE (Bottom Plot):

- The histogram reveals that the distribution of ad_impressions is somewhat multimodal, with peaks around 1100, 1400, and 1600.
- The mean (indicated by the vertical dashed line) is slightly above 1400, suggesting that while the central tendency is near this value, the distribution is not strongly skewed.
- There is a long right tail, which is consistent with the presence of outliers, but the overall distribution appears more spread out rather than sharply skewed.

- These insights suggest that the ad_impressions variable is fairly distributed across a wide range, with several clusters of high frequency around different values, and a few instances of unusually high impressions contributing to the tail of the distribution.

Visualizing visitors using boxplot and histplot

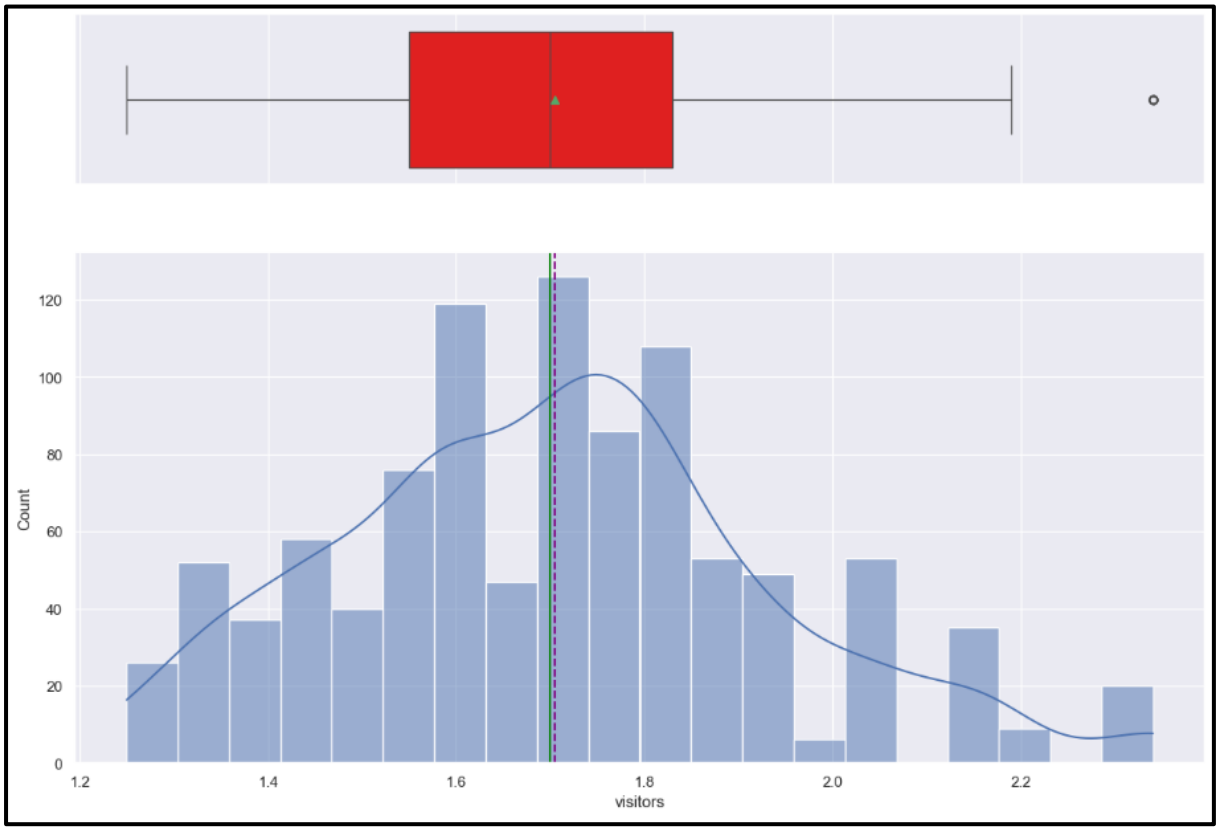

Fig 6: Visualizing visitors using boxplot and histplot

Observation:

Box Plot:

- The distribution of visitors is relatively symmetric with a moderate spread.
- The median value (indicated by the red line) is around 1.75.
- There is one outlier on the higher end, indicating an unusually high number of visitors compared to the rest of the data.

Histogram:

- The distribution of visitors is approximately bell-shaped (normal), with a peak around 1.75.
- The data is centered around the mean (indicated by the dashed line) which is slightly higher than the median.
- The distribution is unimodal, with a single peak.

Visualizing genre using Barplot

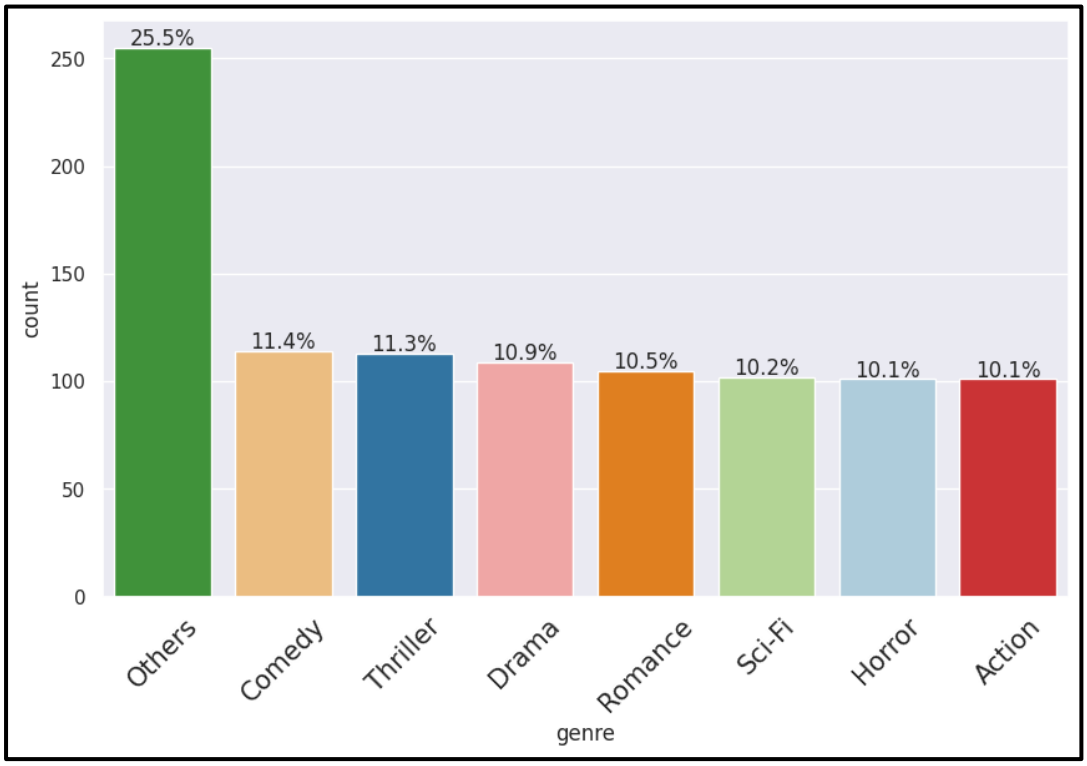

Fig 7: Visualizing genre using Barplot

Observation:

- "Others" is the most frequent genre, followed by "Comedy" and "Thriller."
- The remaining genres ("Drama," "Romance," "Sci-Fi," "Horror," and "Action") have relatively similar frequencies.

Visualizing season using Barplot

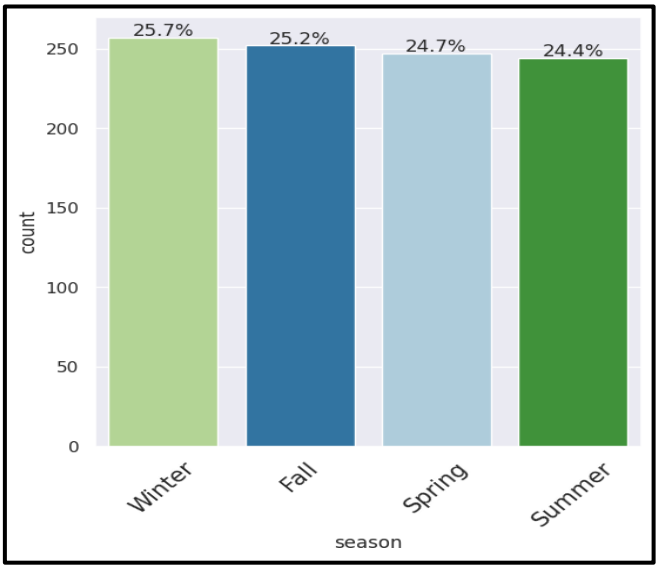

Fig 8: Visualizing season using Barplot

Observation:

- The counts for each season are very close, indicating a relatively even distribution of data across the year.
- While the distribution is even, Winter has the highest count with 25.7% of the data points.
- Summer has the lowest count with 24.4% of the data points, being the closest in value to Winter.

Visualizing dayofweek using Barplot

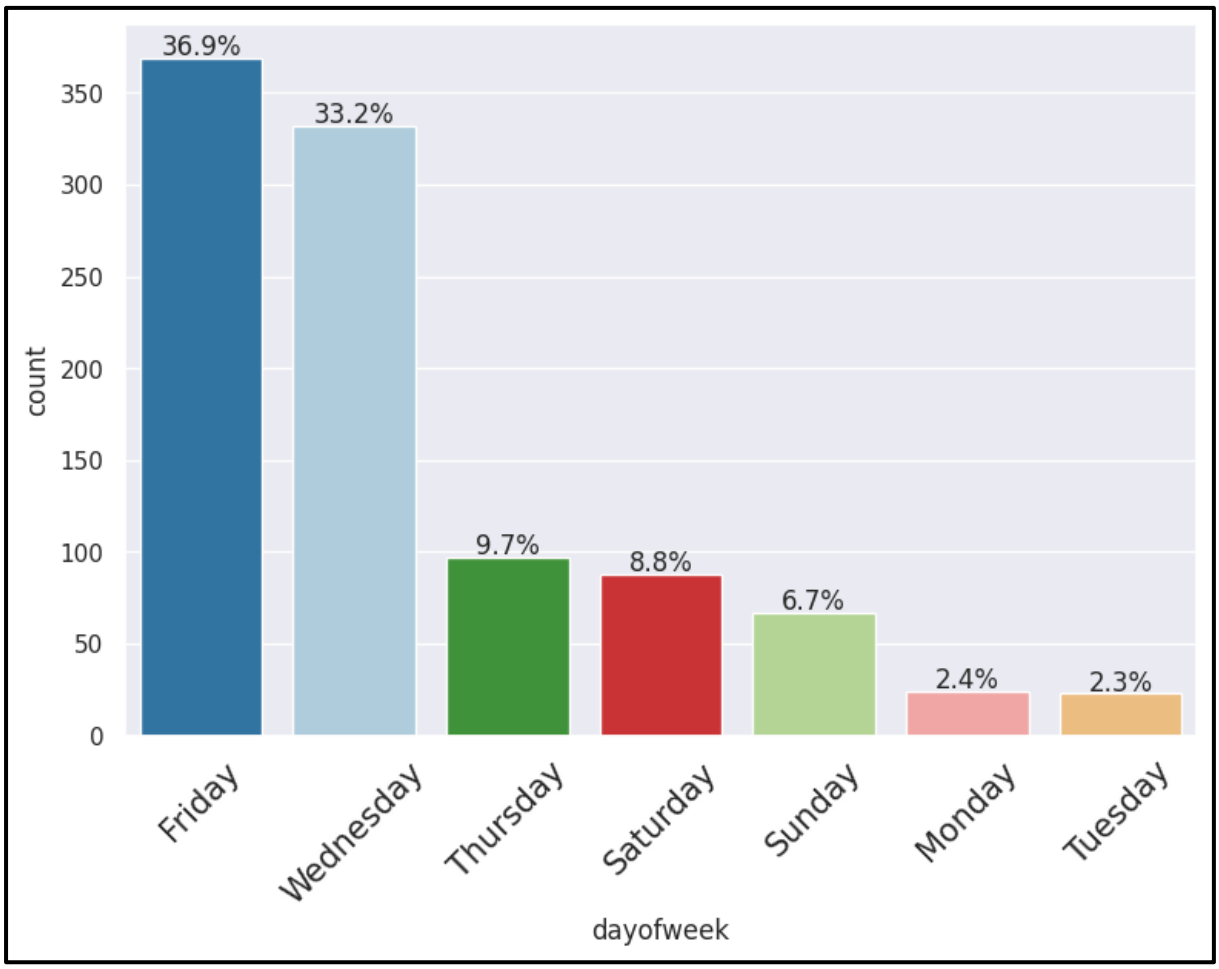

Fig 9: Visualizing dayofweek using Barplot

Observation:

- Friday is the most frequent day, accounting for 36.9% of the data points.
- Wednesday is the second most frequent day with 33.2% of the data points.
- There's a significant drop in frequency after Wednesday.
- Tuesday is the least frequent day with only 2.3% of the data points.
- The distribution is skewed to the left, with a majority of data points concentrated on Friday and Wednesday.

Visualizing major_sports_event using Barplot

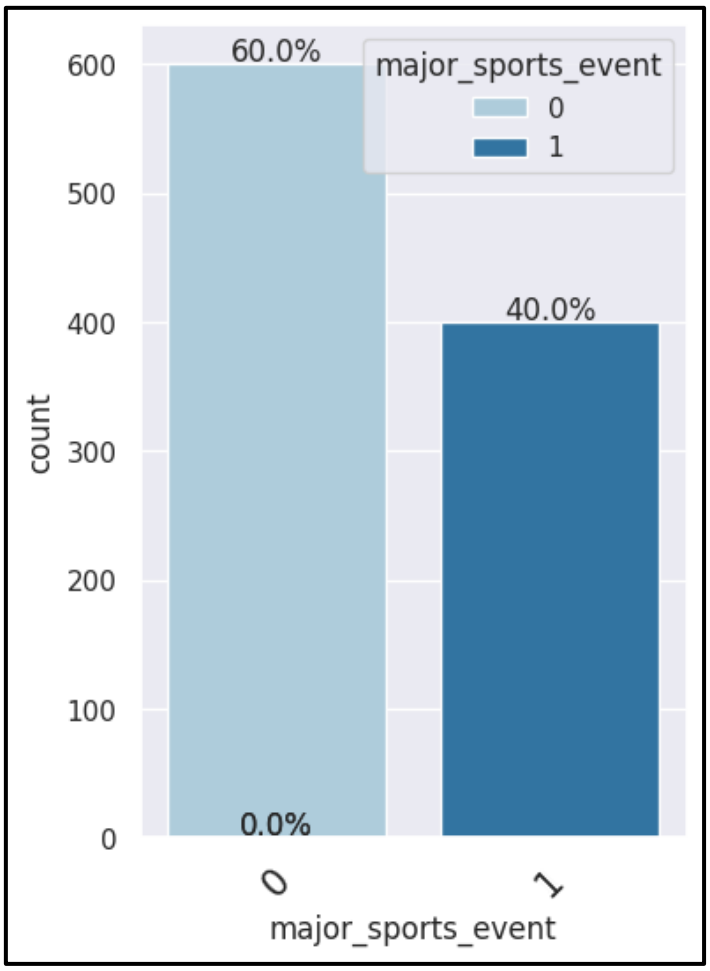

Fig 10: Visualizing major_sports_event using Barplot

Observation:

- Approximately 60% of the data points belong to the category where major_sports_event is 0.
- Only 40% of the data points belong to the category where major_sports_event is 1.

Bivariate Analysis

Let's check the correlation between numerical variables.

Visualizing Numerical data using pairplot

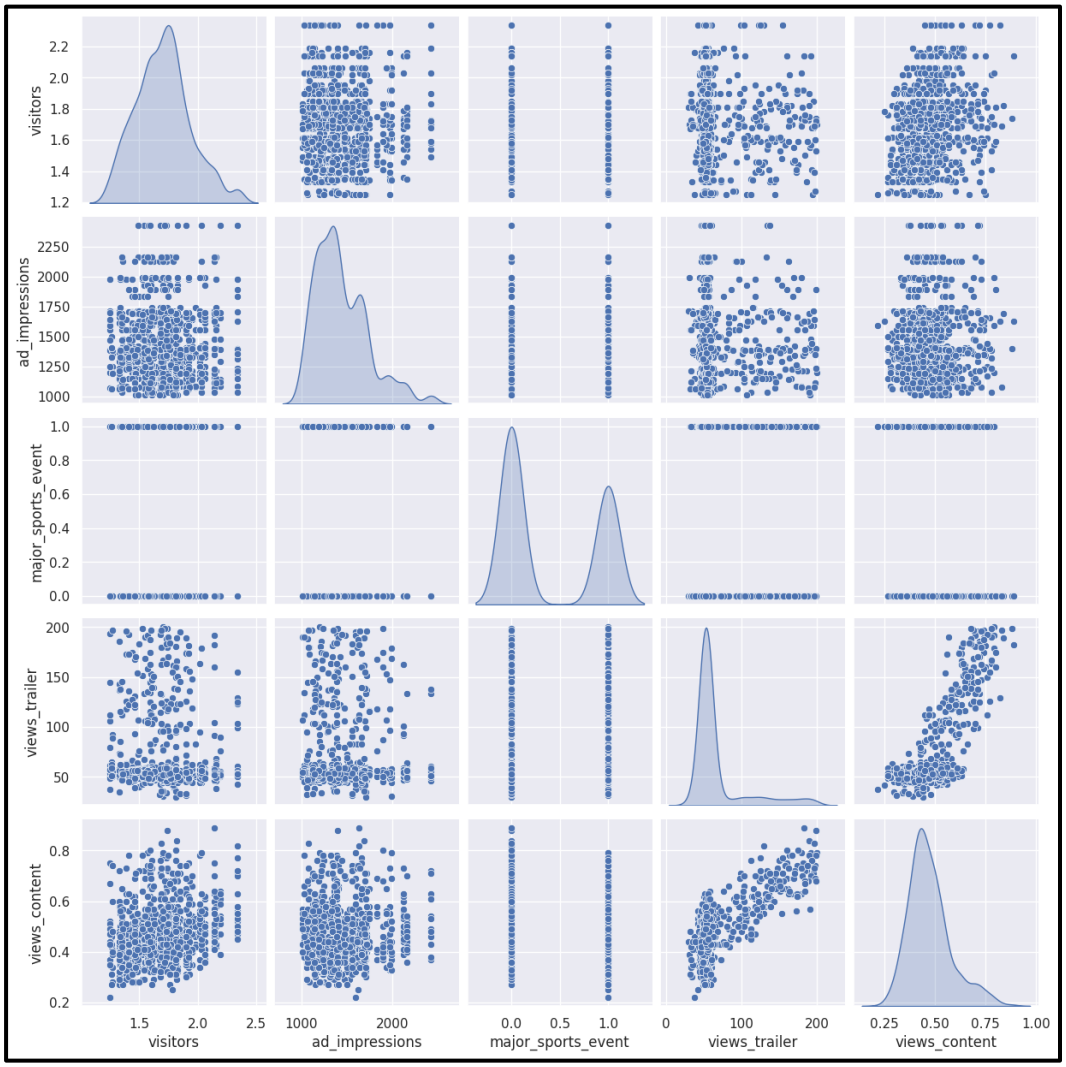

Fig 11 : Visualizing Numerical data using pairplot

Observation:

- There appears to be a positive correlation between visitors and ad impressions, suggesting that a higher number of visitors tends to lead to more ad impressions.
- A slightly positive correlation might exist between visitors and content views, indicating that more visitors could potentially lead to a higher number of content views.
- There's a possible weak positive correlation between ad impressions and content views, suggesting that more ad impressions might influence content viewership to some extent.

- A moderate positive correlation is evident, suggesting that a higher number of trailer views is associated with a higher number of content views.
- Visitors and ad impressions seem to be right-skewed, indicating that there might be a few instances with exceptionally high values.
- Major sports event is imbalanced, with more instances of no major sports event.
- Views trailer and views content appear to be right-skewed as well.

Visualizing Numerical data using Heatmap

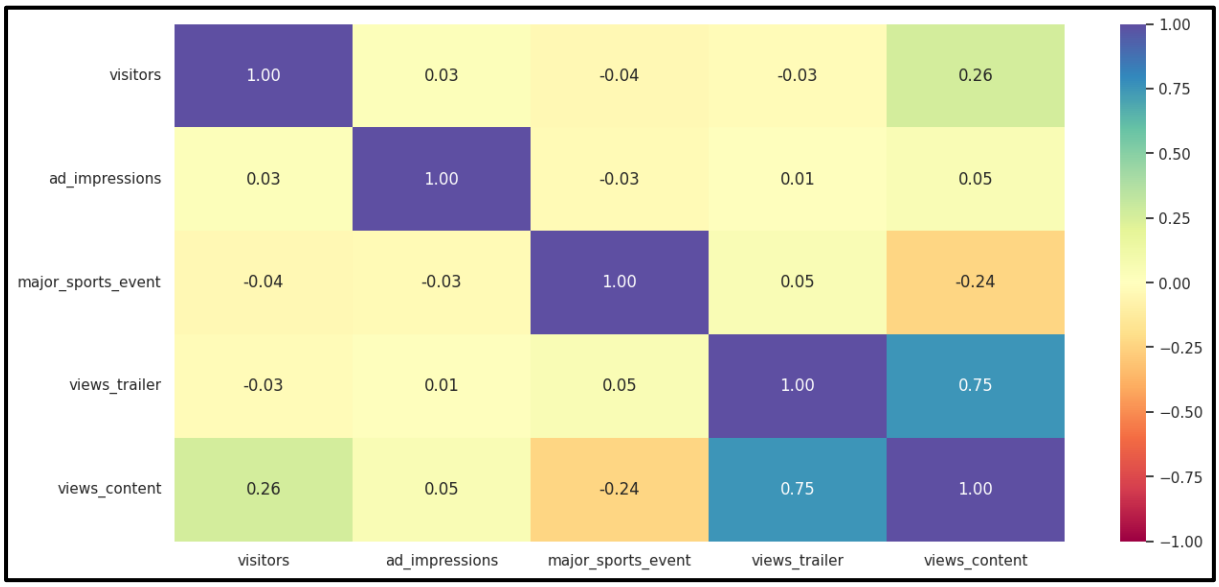

Fig 12: Visualizing Numerical data using Heatmap

Observation:

- There's a strong positive correlation between views_trailer and views_content, suggesting that a higher number of trailer views is associated with a higher number of content views. This is intuitive as a well-received trailer might encourage more people to watch the content.
- A moderate positive correlation exists between visitors and views_content, indicating that a higher number of platform visitors tends to result in more content views. This is logical as a larger audience pool generally leads to more content consumption.
- The correlations between other variables (e.g., ad_impressions, major_sports_event) and views_content are relatively weak, suggesting that these factors might have less impact on first-day viewership.
- There are no strong correlations observed among the other variables (visitors, ad_impressions, major_sports_event, views_trailer).

Visualizing Relation between views_content and genre using boxplot

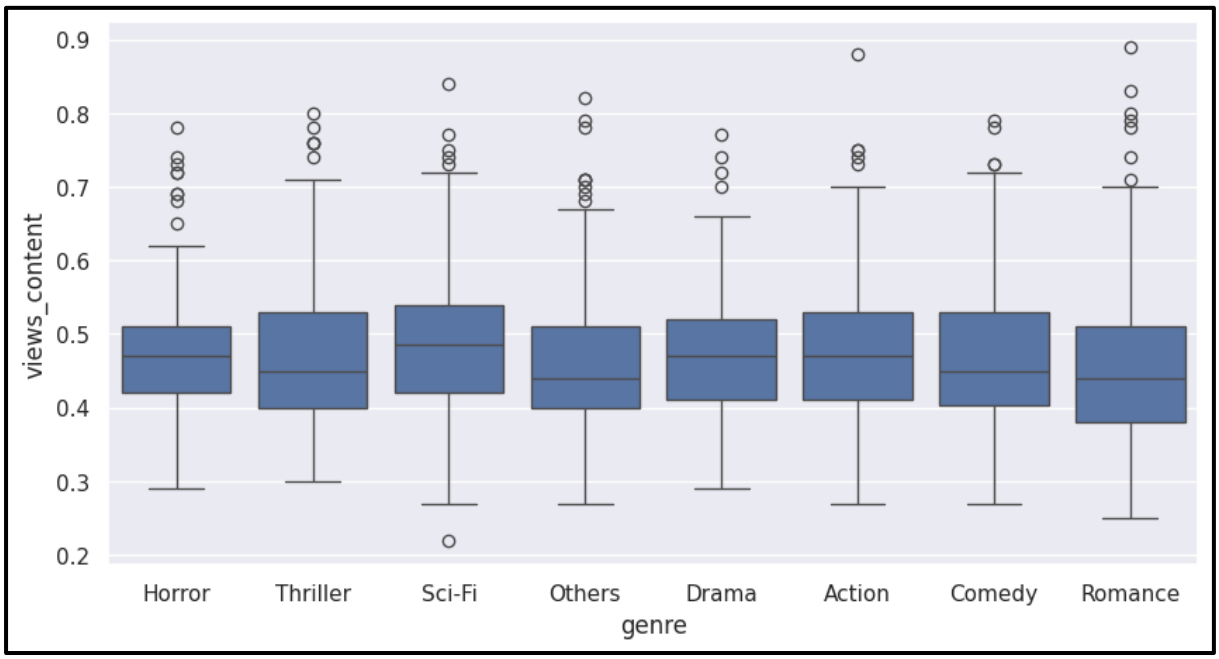

Fig 13: Visualizing Relation between views_content and genre using boxplot

Observation:

- The distribution of views_content is right-skewed for all genres, indicating a few instances with exceptionally high viewership.
- The median viewership is relatively consistent across all genres, suggesting a similar central tendency.
- There is significant variability in viewership across genres, as evidenced by the box lengths and the presence of outliers.
- Several genres have outliers, indicating a small number of content items with exceptionally high viewership.

Visualizing Relation between views_trailer and genre using boxplot

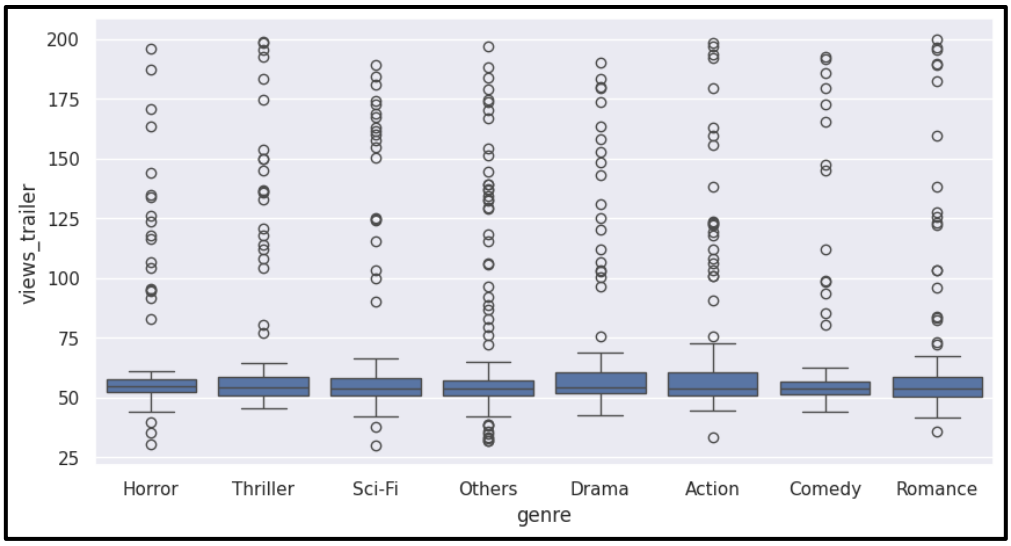

Fig 14: Visualizing Relation between views_trailer and genre using boxplot

Observation:

- The distribution of views_trailer is right-skewed for most genres, indicating a few instances with exceptionally high trailer viewership.
- The median trailer viewership is relatively consistent across all genres, suggesting a similar central tendency.
- There is significant variability in trailer viewership across genres, as evidenced by the box lengths and the presence of outliers.
- Several genres have outliers, indicating a small number of content items with exceptionally high trailer viewership.

Visualizing Relation between views_trailer and genre using boxplot (without outlier)

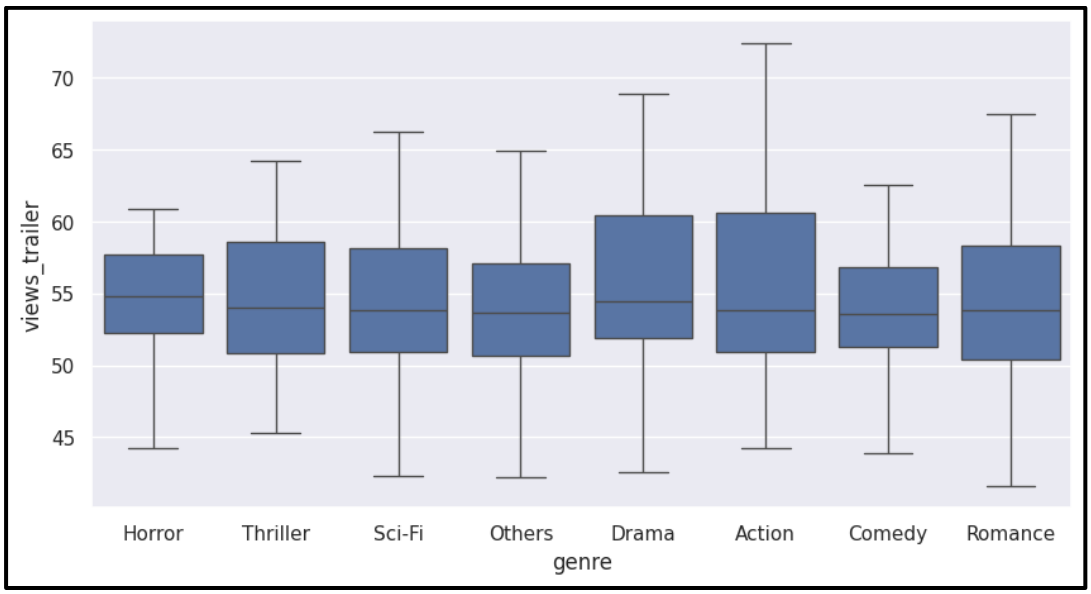

Fig 15: Visualizing Relation between views_trailer and genre using boxplot (without outlier)

Observation:

- The distribution of views_trailer is relatively similar across all genres, with minimal variation in the median values.
- The range of views_trailer is quite similar across genres, as indicated by the box lengths and whiskers.
- There are some outliers present in a few genres, suggesting a small number of content items with exceptionally high trailer viewership.
- Overall, the genre doesn't seem to have a significant impact on the distribution of views_trailer.

Visualizing Relation between ad_impressions and genre using boxplot

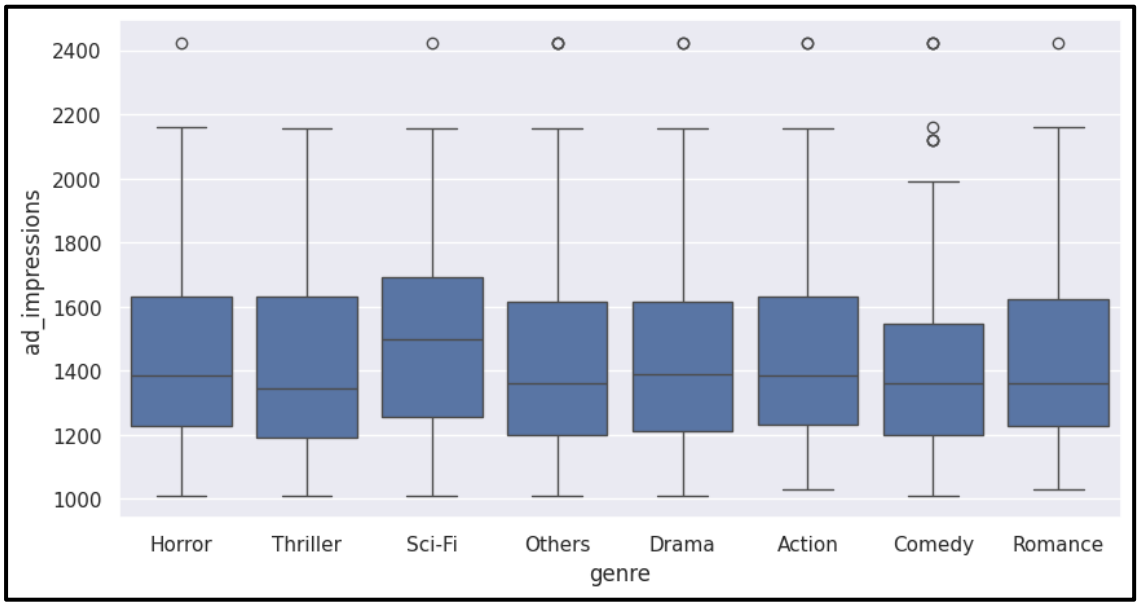

Fig 16: Visualizing Relation between ad_impressions and genre using boxplot

Observation:

- There is a significant overlap in the distribution of ad_impressions across different genres.
- The median ad_impressions is relatively consistent across most genres.
- While there are some variations in the interquartile ranges, they are not substantial.
- A few outliers are present in some genres, indicating exceptionally high ad_impressions for certain content.
- Genre Consistency: The box plots for most genres are quite similar in shape and range, suggesting a consistent distribution of ad_impressions.
- Outliers: The presence of outliers in some genres implies that a few content items within those genres have significantly higher ad_impressions compared to others.

Visualizing Relation between dayofweek and visitors using boxplot

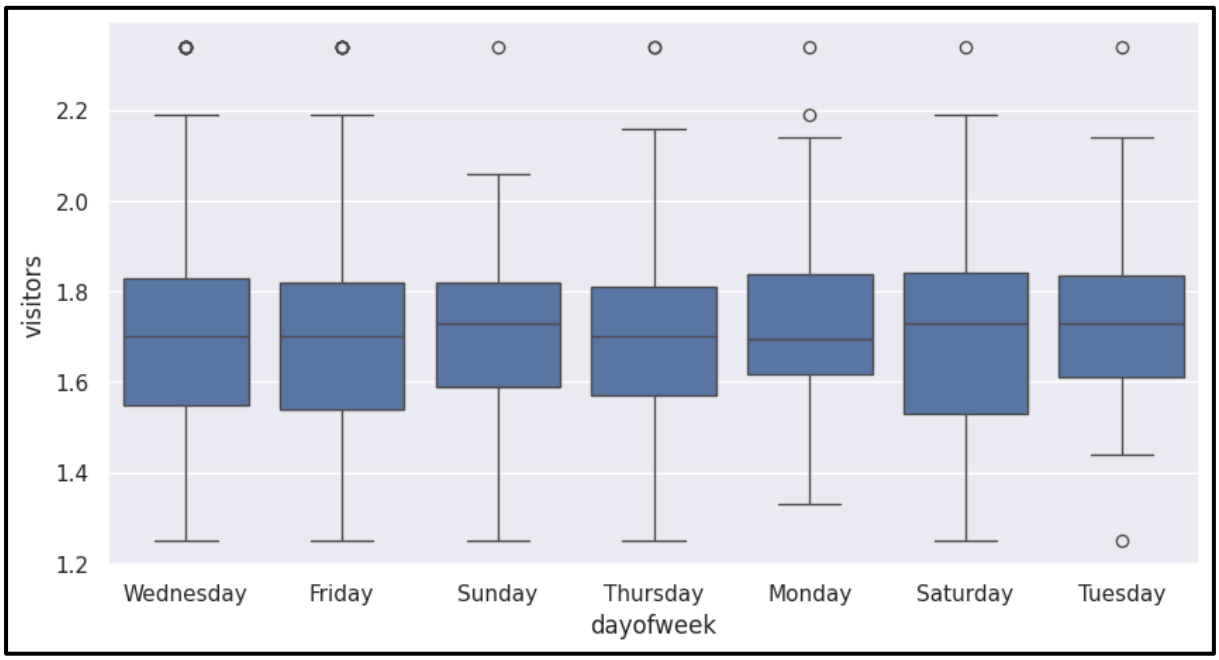

Fig 17: Visualizing Relation between dayofweek and visitors using boxplot

Observation:

- Wednesday and Friday: These days seem to have a slightly higher median number of visitors compared to other days.
- Tuesday: This day has a slightly lower median number of visitors compared to other days.

Visualizing Relation between dayofweek and views_content using boxplot

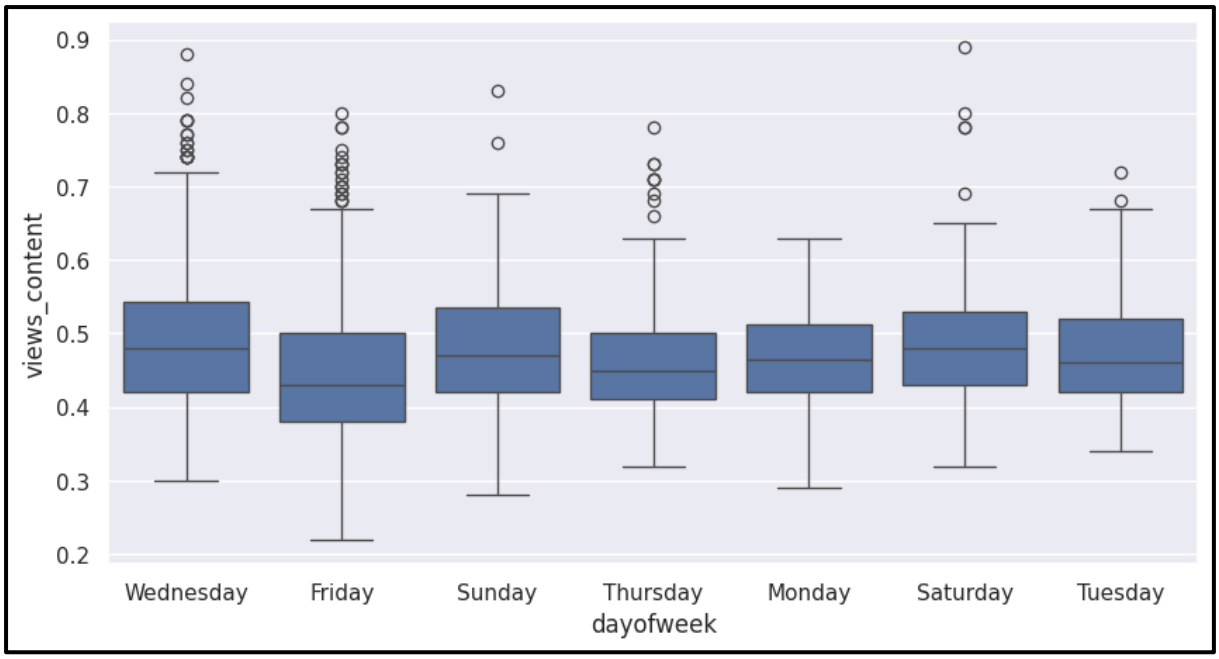

Fig 18: Visualizing Relation between dayofweek and views_content using boxplot

Observation:

- Wednesday, Friday, and Sunday generally have higher median views_content compared to other days.
- Tuesday consistently shows the lowest median views_content.
- The presence of outliers indicates that there are instances of exceptionally high or low views_content on specific days.
- The overall distribution of views_content is skewed to the right for most days, suggesting a few instances with very high viewership.

Visualizing Relation between dayofweek and views_trailer using boxplot

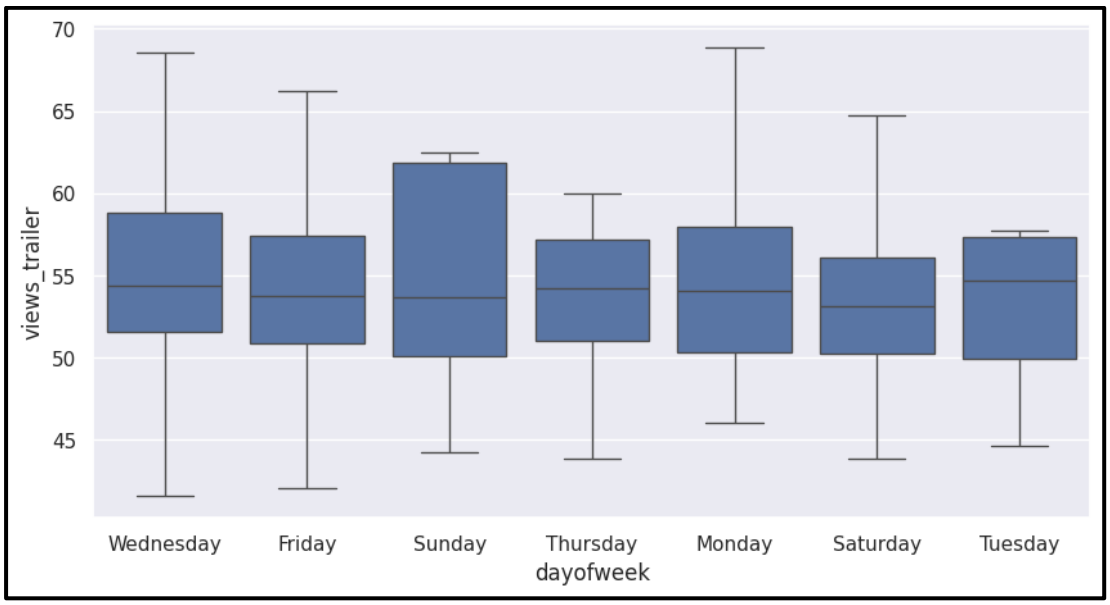

Fig 19: Visualizing Relation between dayofweek and views_trailer using boxplot

Observation:

- There is a minor variation in the median views_trailer for different days, with no clear pattern of increase or decrease.
- The interquartile ranges (IQR) for most days are similar, indicating a consistent spread of data points within each day.

Visualizing Relation between season and views_content using boxplot

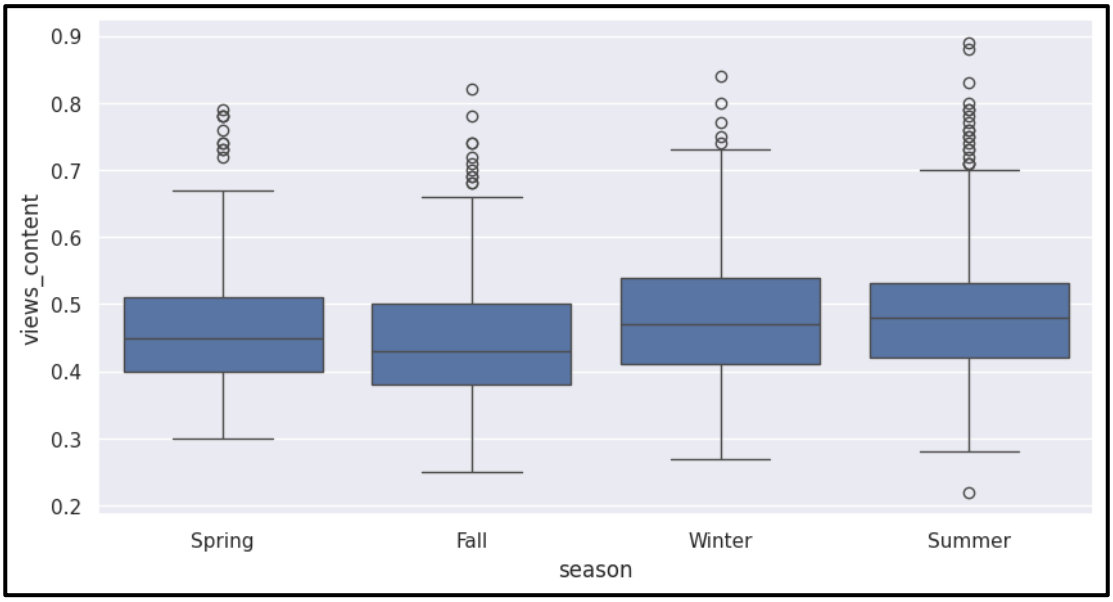

Fig 20: Visualizing Relation between season and views_content using boxplot

Observation:

- There is a noticeable variation in the median views_content across the seasons, with Winter and Summer generally showing higher values compared to Spring and Fall.
- The interquartile range (IQR) is relatively consistent across all seasons, suggesting a similar spread of data points within each season.
- The presence of outliers in all seasons indicates that there are exceptional cases with significantly higher or lower views_content.
- The overall distribution of views_content is skewed to the right for all seasons, suggesting a few instances with very high viewership.

Visualizing Relation between season and views_trailer using boxplot

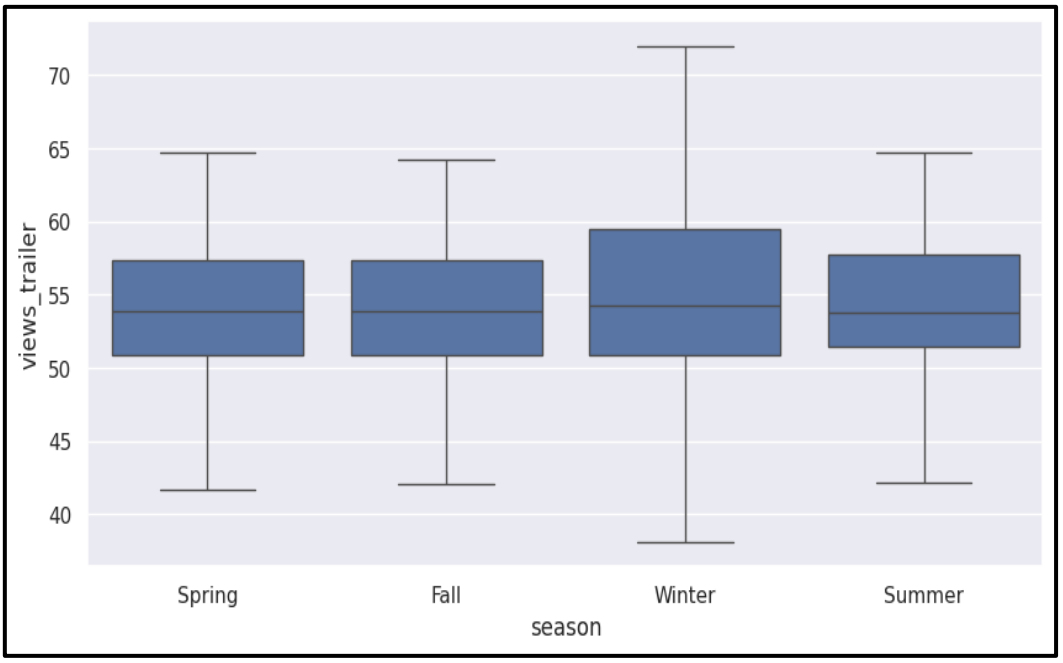

Fig 21: Visualizing Relation between season and views_trailer using boxplot

Observation:

- The median views_trailer for each season is very similar, indicating a lack of significant seasonal influence on trailer views.
- The interquartile ranges (IQR) are comparable across seasons, suggesting a similar spread of data points within each season.
- While there are some outliers present in each season, they do not appear to be clustered in any particular season.
- Overall, the data suggests that seasonality does not have a pronounced impact on trailer viewership.

Visualizing Relation between genre, views_content and season using boxplot

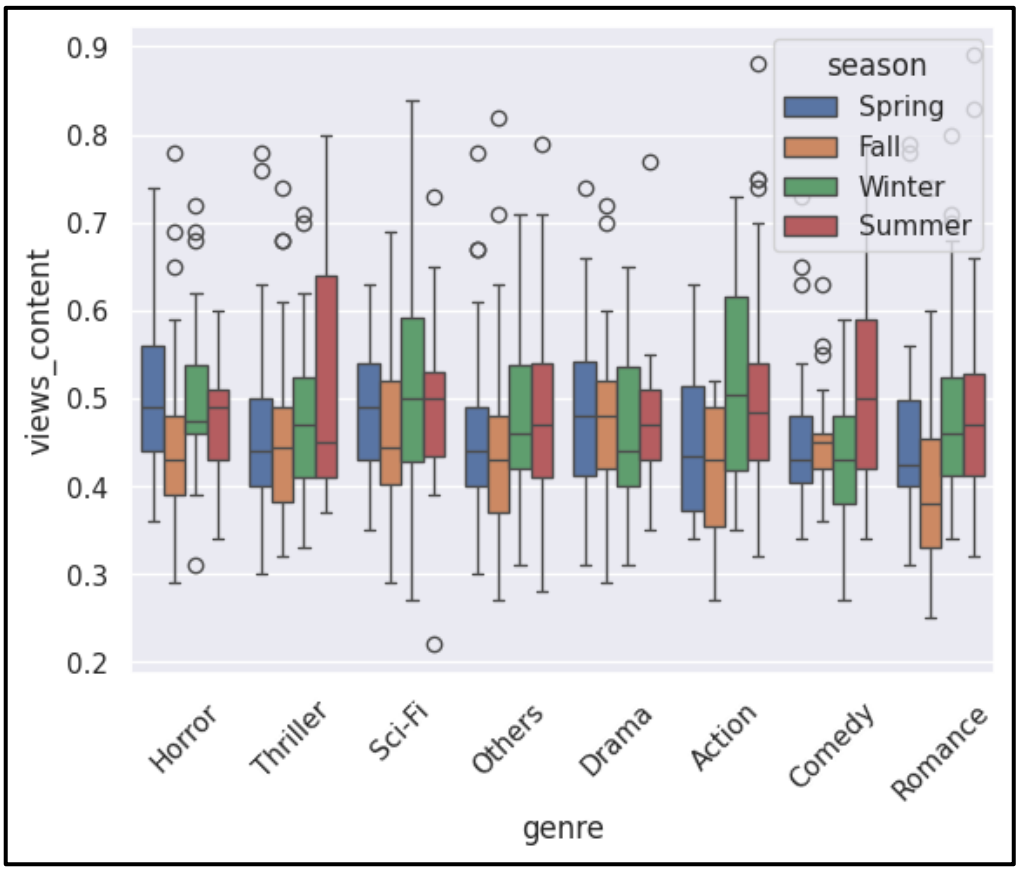

Fig 22: Visualizing Relation between genre, views_content and season using boxplot

Observation:

- Distribution of Views Content by Genre and Season
- Genre Variation: There is a clear variation in views_content across different genres. Some genres, like Horror and Thriller, tend to have higher median views_content compared to others like Romance and Comedy.
- Seasonal Impact: While there are some variations in views_content across seasons for specific genres, the overall pattern is not entirely consistent. Some genres show a more pronounced seasonal effect than others.
- Outliers: The presence of outliers in multiple genres and seasons suggests that there are specific content items that significantly deviate from the general trend.

Visualizing major_sports_event and views_content using Barplot

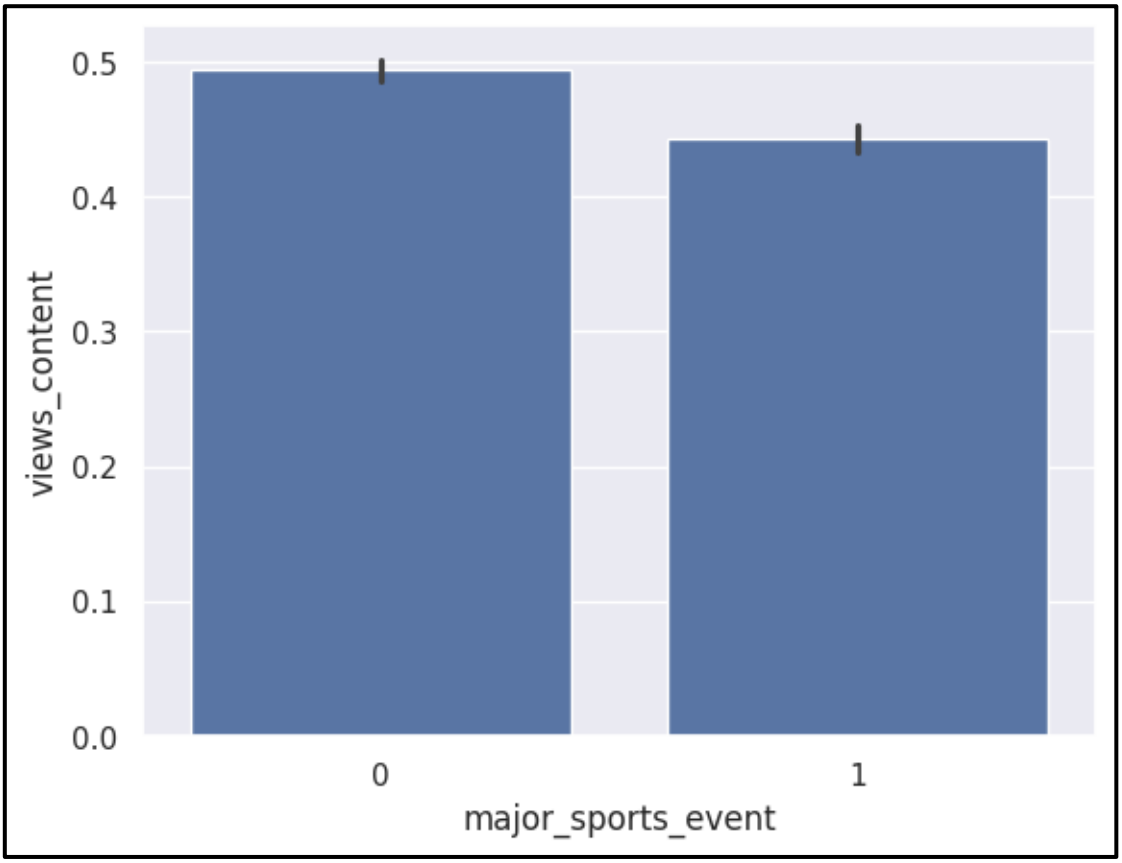

Fig 23: Visualizing major_sports_event and views_content using Barplot

Observation:

- Impact of Major Sports Events on Views Content
- Average Views Content: The average number of views_content is higher when there is no major sports event (major_sports_event = 0) compared to when there is a major sports event (major_sports_event = 1).
- Variation: The error bars indicate a similar level of variability in views_content for both conditions.
- Potential Correlation: The data suggests a potential negative correlation between major sports events and views_content.

Visualizing major_sports_event and views_content using Boxplot

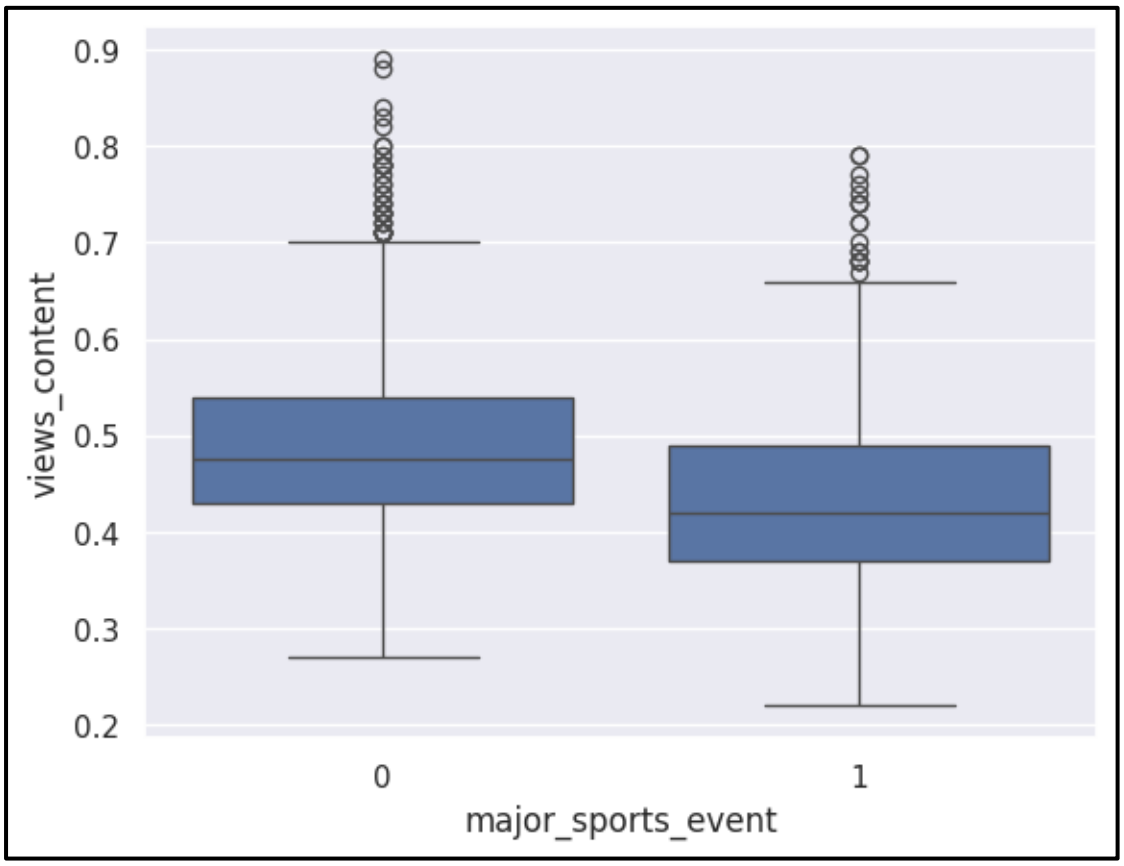

Fig 24: Visualizing major_sports_event and views_content using Boxplot

Observation:

- Distribution of Views Content
- The box plot visually represents the distribution of views_content based on the presence or absence of a major sports event.
- The median views_content is significantly higher when there is no major sports event (major_sports_event = 0) compared to when there is a major sports event (major_sports_event = 1).
- The interquartile range (IQR) is similar for both groups, indicating a similar spread of data points within each group.
- The presence of outliers in both groups suggests the existence of exceptional cases with extremely high or low views_content.
- Overall, the data suggests a negative relationship between major sports events and content viewership.

Visualizing dayofweek and visitors using Barplot

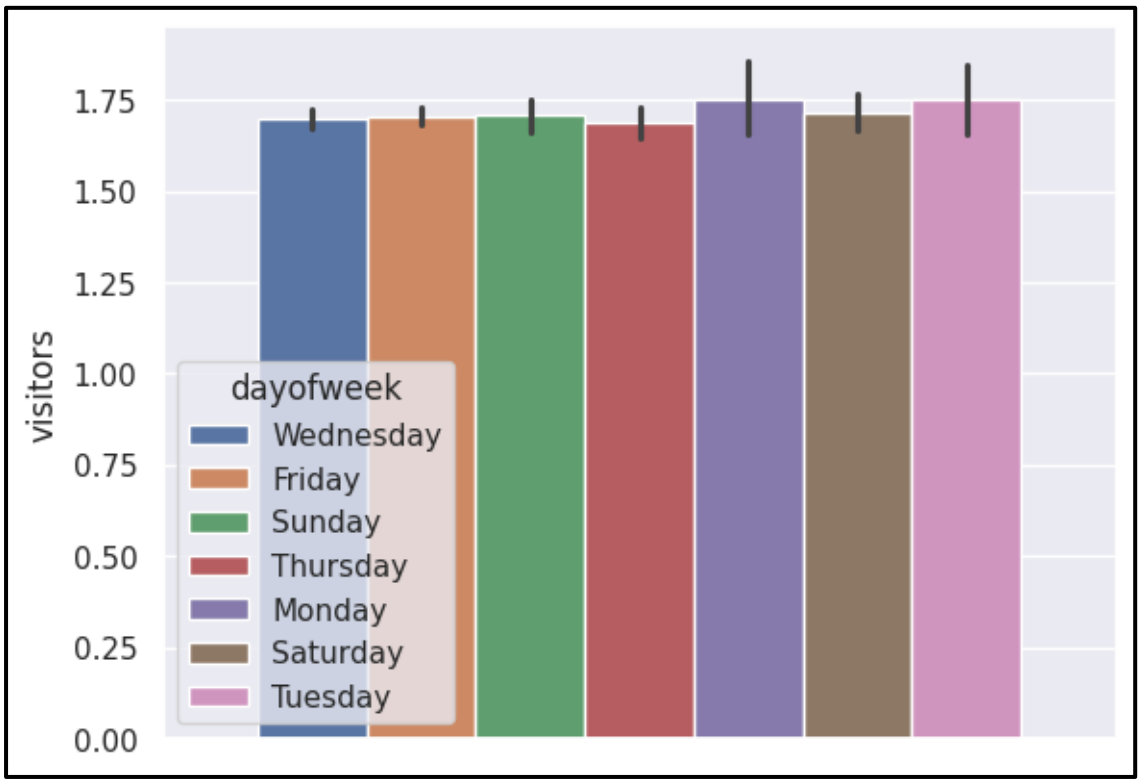

Fig 25: Visualizing dayofweek and visitors using Barplot

Observation:

- Consistent Visitor Count: The number of visitors across different days of the week is remarkably consistent, with minimal variation between the bars.
- No Clear Trend: There is no discernible pattern or trend in visitor numbers across the week.
- Potential for Further Analysis: While the overall pattern is consistent, a deeper analysis might reveal subtle differences or trends when considering other factors such as time of day, seasonality, or special events.

Visualizing Relation between dayofweek views_content and views_trailor using scatterplot

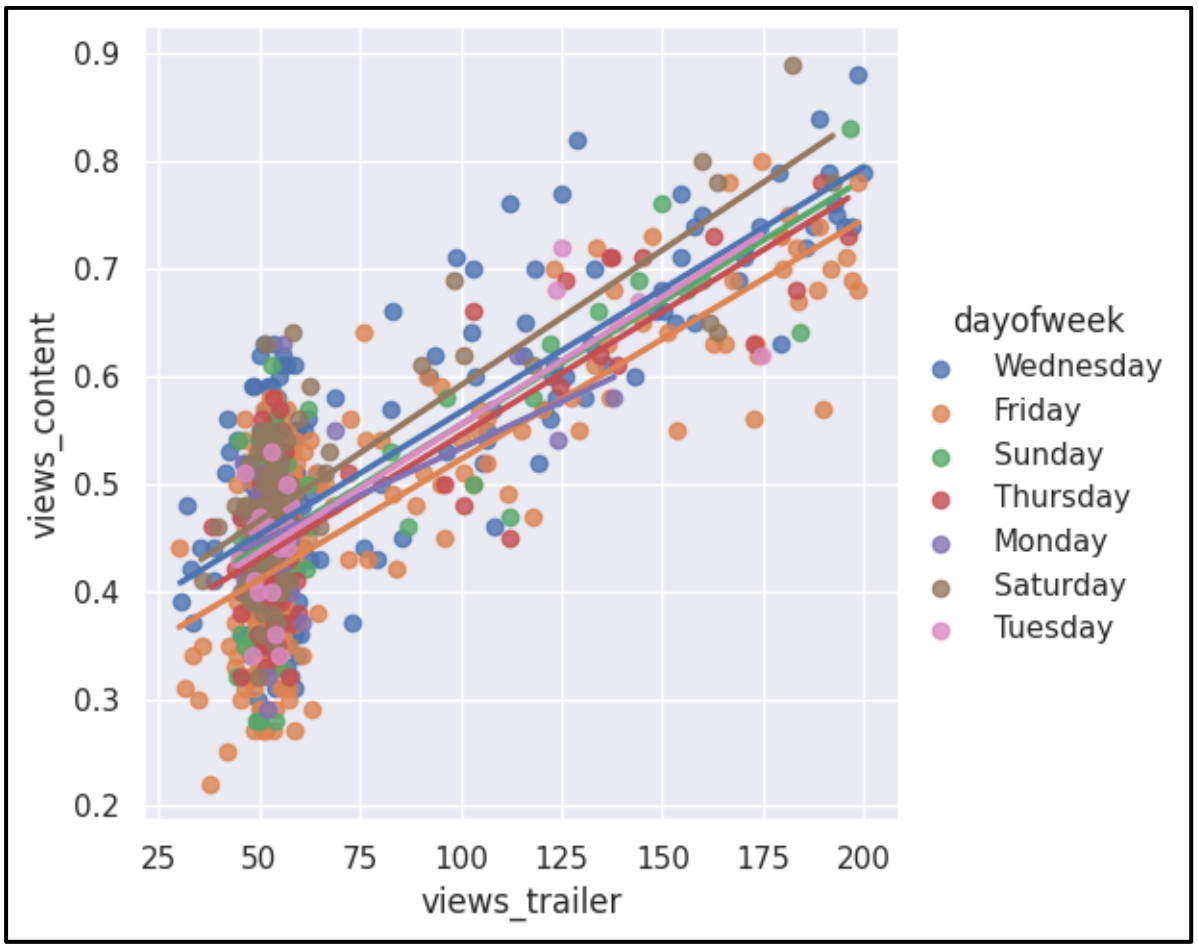

Fig 26: Visualizing Relation between dayofweek views_content and views_trailor using scatterplot

Observation:

- Positive Correlation: There appears to be a positive correlation between views_trailer and views_content across all days of the week. This suggests that an increase in trailer views generally leads to an increase in content views.
- Day-of-Week Impact: The scatter plot indicates that the relationship between views_trailer and views_content might vary slightly across different days of the week, as suggested by the different slopes of the regression lines.
- Data Distribution: The distribution of data points seems to be relatively consistent across different days of the week, with no apparent outliers or clusters.

Questions to be answered

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

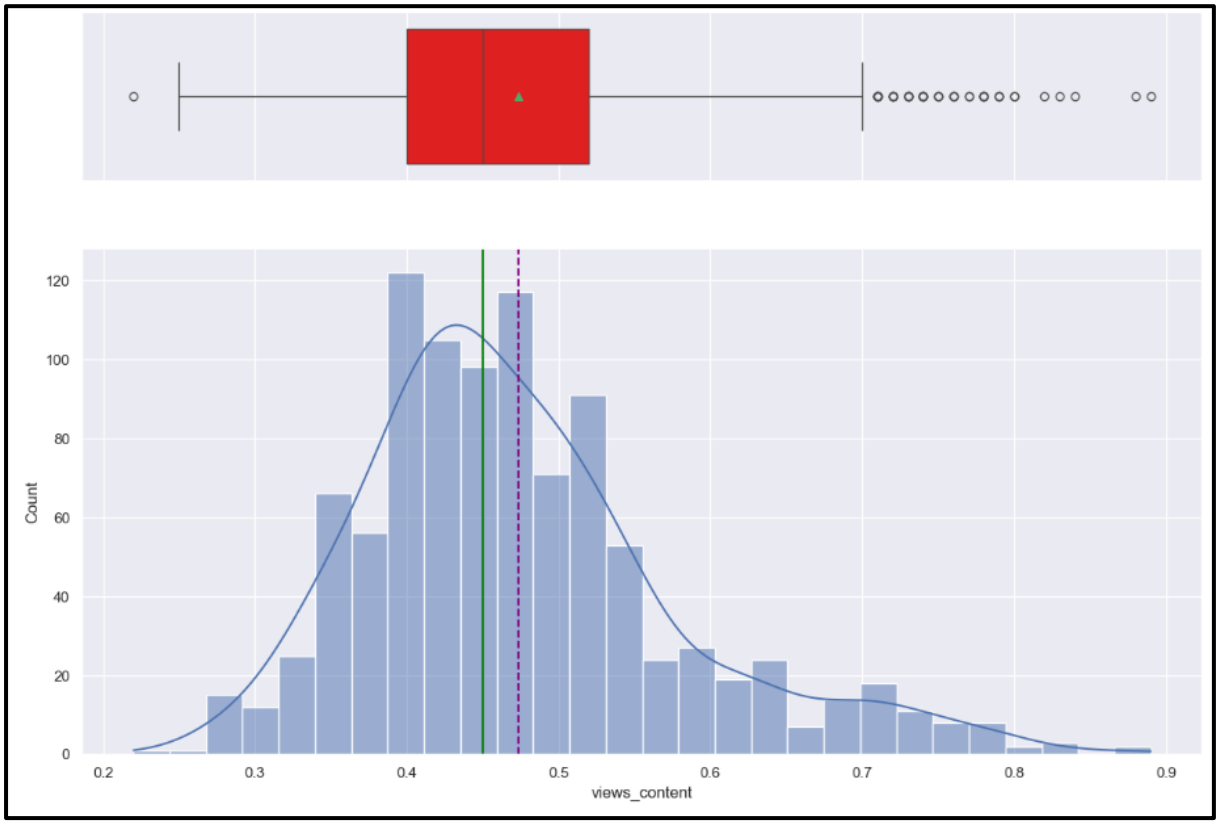

Fig 27: Visualizing views_content using boxplot and histplot

Observation:

Distribution Shape: The distribution of views_content is approximately bell-shaped, indicating a normal distribution with a central tendency around the mean.

Central Tendency: The median value, represented by the red line in the box plot, is located around 0.5, suggesting that half of the data points fall below this value.

Spread: The box plot shows a moderate spread in the data, with the interquartile range (IQR) capturing the middle 50% of the data.

Outliers: A few outliers are present in the upper tail of the distribution, indicating exceptionally high values of views_content.

Histogram: The histogram confirms the normal distribution shape with a peak around the median value.

Overall, the data suggests a relatively symmetric distribution of views_content with a central tendency around 0.5 and the presence of a few outliers.

Answer

The average number of content views appears to be less than 0.5 million.

2. What does the distribution of genres look like?

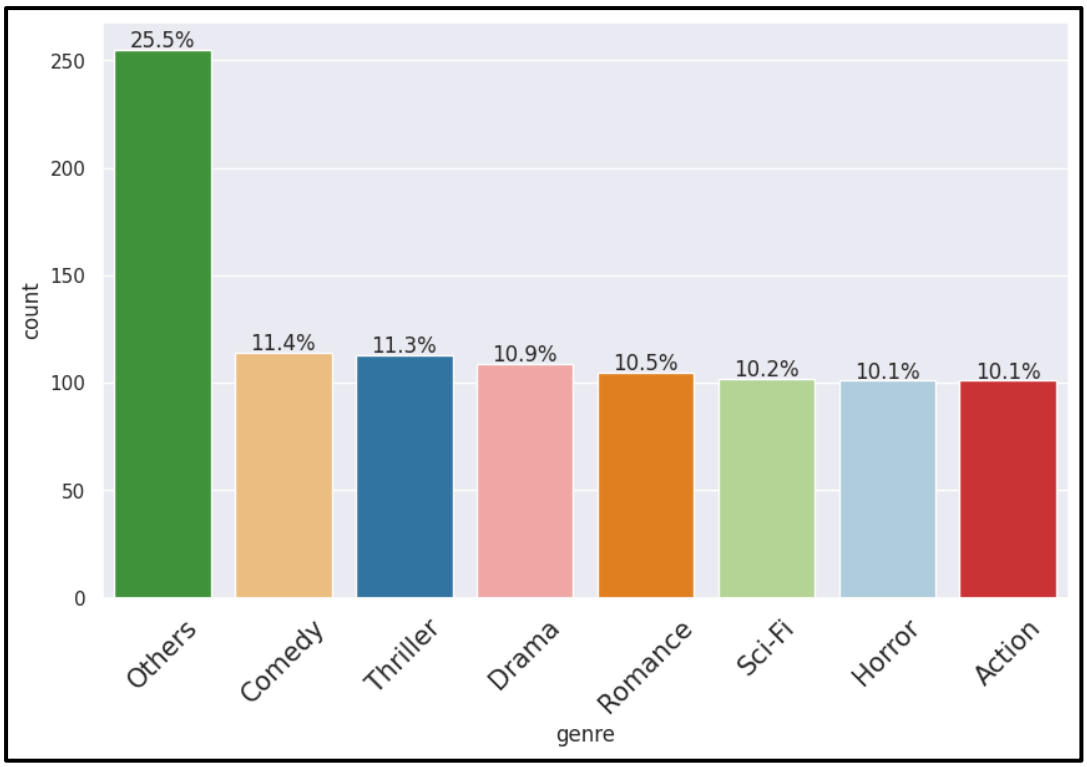

Fig 28: Visualizing genre using Barplot

Answer

The distribution of genres is heavily skewed towards "Others", followed by a relatively even distribution among the remaining genres.

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?

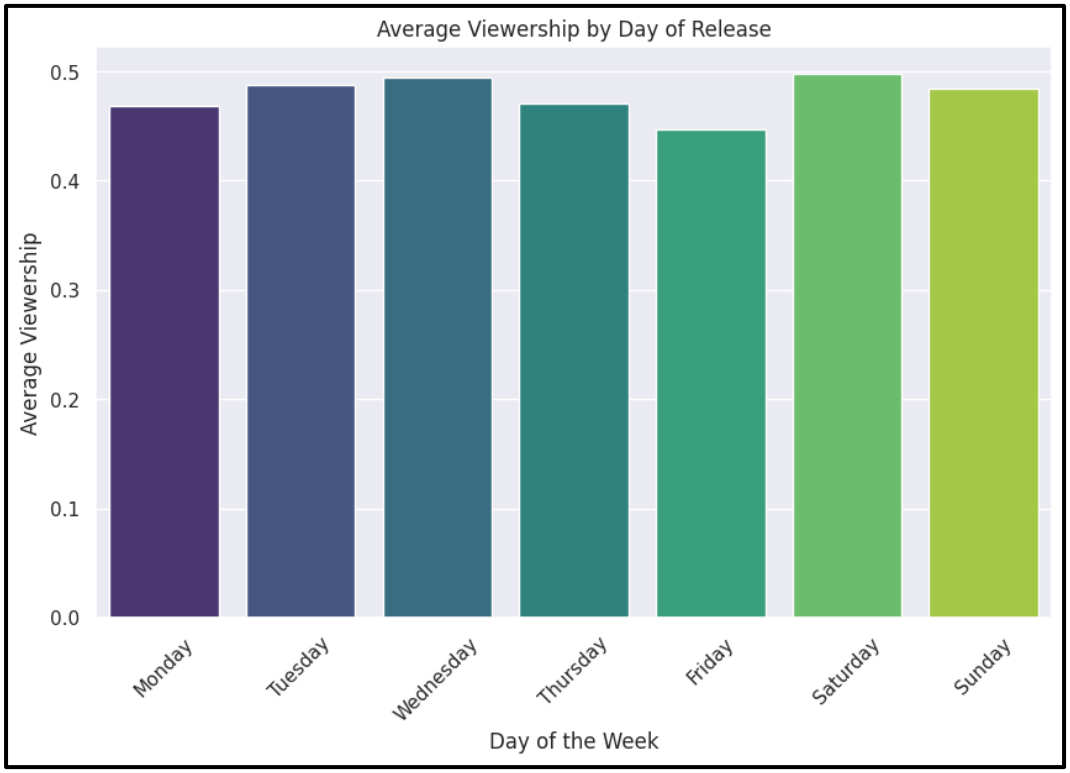

Fig 29: Visualizing day of the week using Barplot

Answer

- The mean content views decreases for Friday content release with respect to other days.
- Saturday has the highest average viewership nearly 0.5.
- Wednesday is the second highest.
- Viewership tends to increase as the week progresses, peaking on Saturday and Sunday. There's a noticeable uptick from Monday to Tuesday, with a slight dip on Wednesday before climbing again towards the weekend.

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

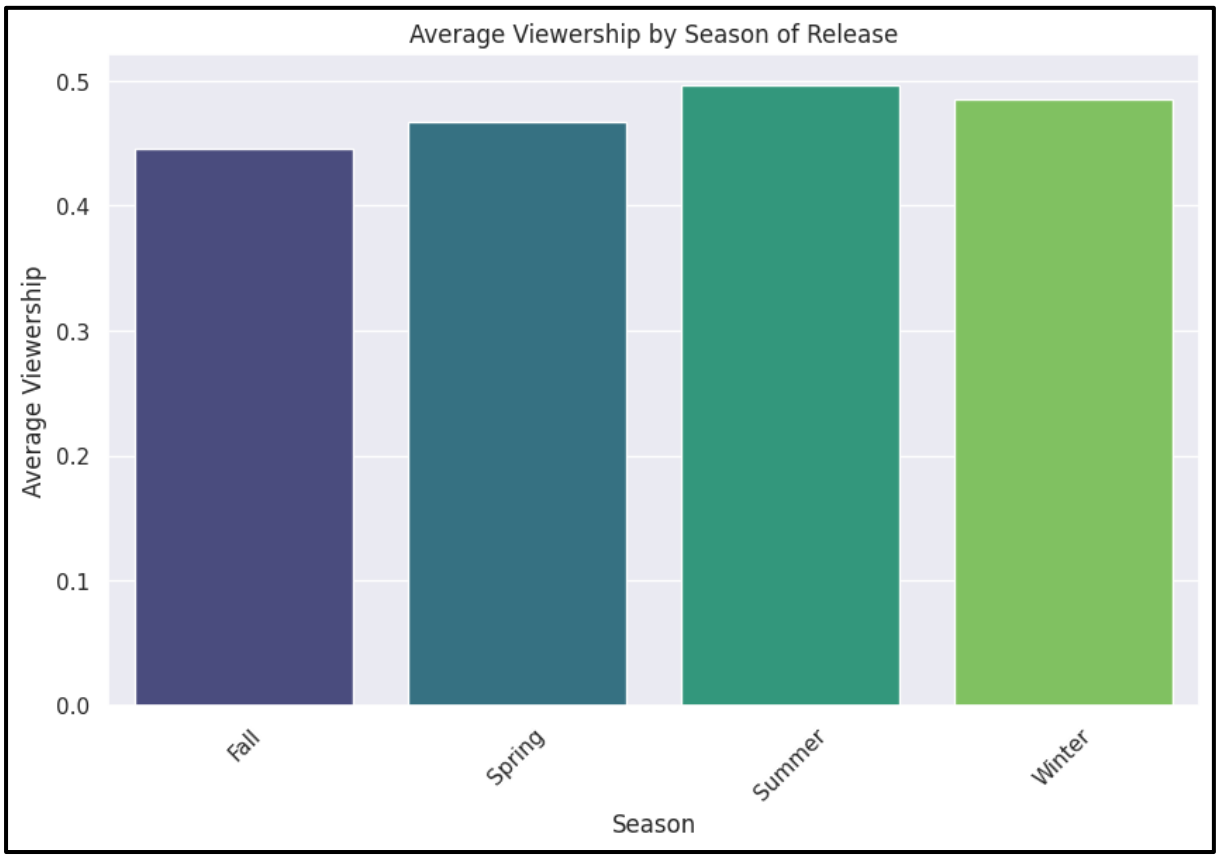

Fig 30: Visualizing seasons using Barplot

Answer

- 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.

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

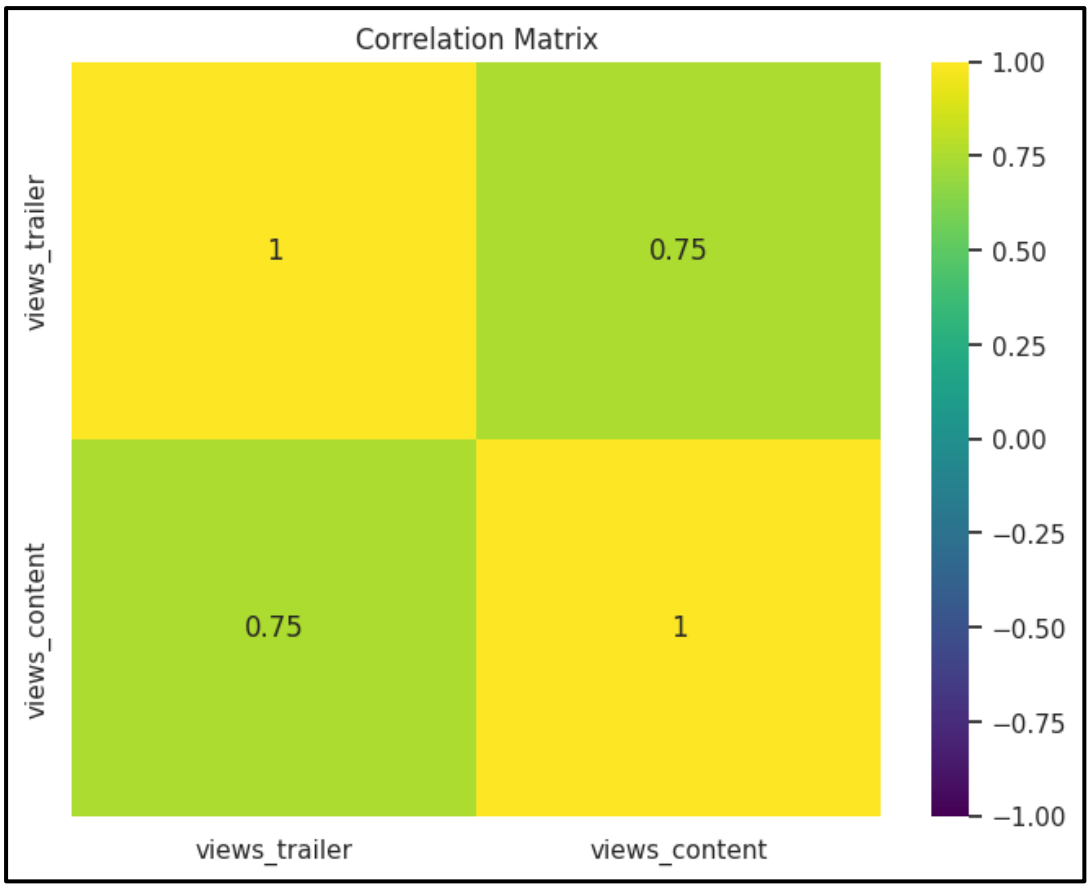

Fig 31: Visualizing correlation between trailer views and content views using heatmap

Answer

The correlation between trailer views and content views is strong and positive. The correlation coefficient is approximately 0.75, indicating a strong relationship where higher trailer views tend to correspond with higher content views.

Data Preprocessing

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

Data Cleaning

Checking the duplicate and missing value

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

Fig 32: Checking the duplicate and missing value

There are no duplicate and missing values in the data set.

Outlier treatment

Outlier Detection

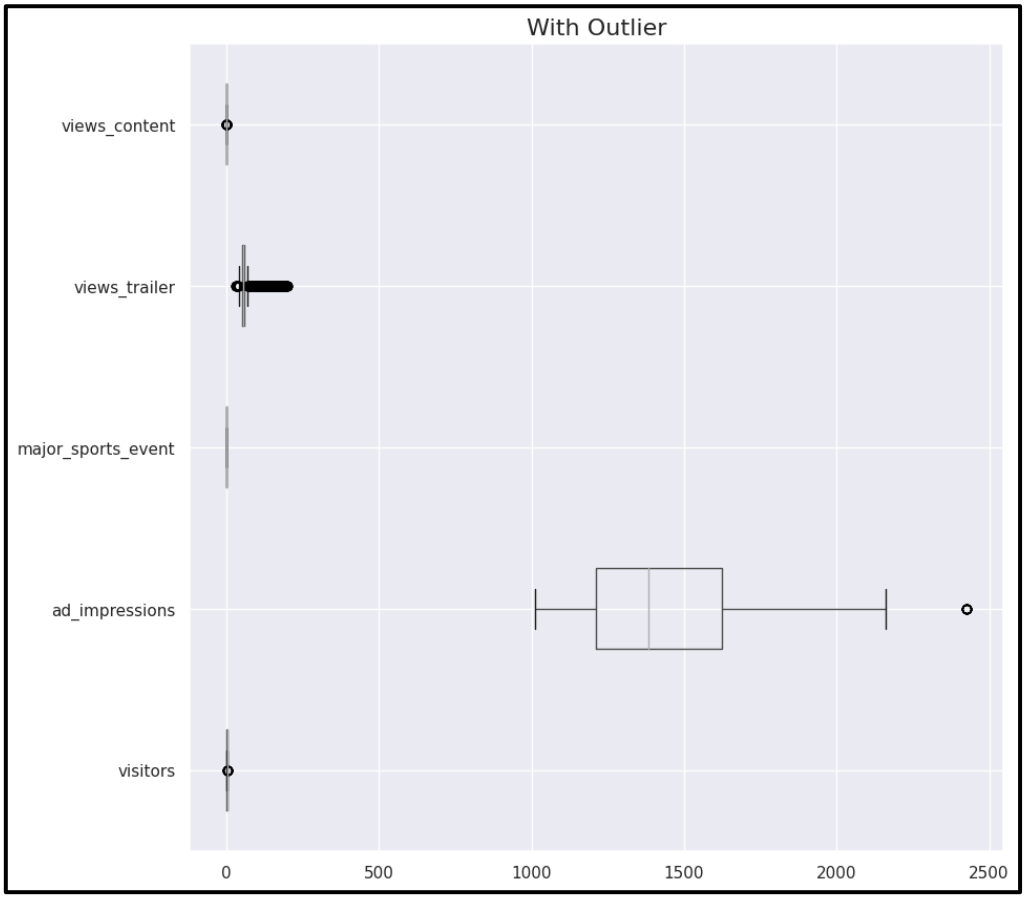

Fig 33: Detecting the Outliers

Observation:

views_content and views_trailer have a relatively narrow interquartile range, indicating a smaller spread of values compared to other variables.

major_sports_event likely represents a binary variable with limited variability.

ad_impressions and visitors show a wider range of values, with outliers suggesting potential anomalies or influential data points.

Formula used to remove outliers:

```
def remove_outlier(series):
    q1 = series.quantile(0.25)
    q3 = series.quantile(0.75)
    iqr = q3 - q1
    lower_range = q1 - 1.5 * iqr
    upper_range = q3 + 1.5 * iqr
    return lower_range, upper_range
```

After Outlier Removed

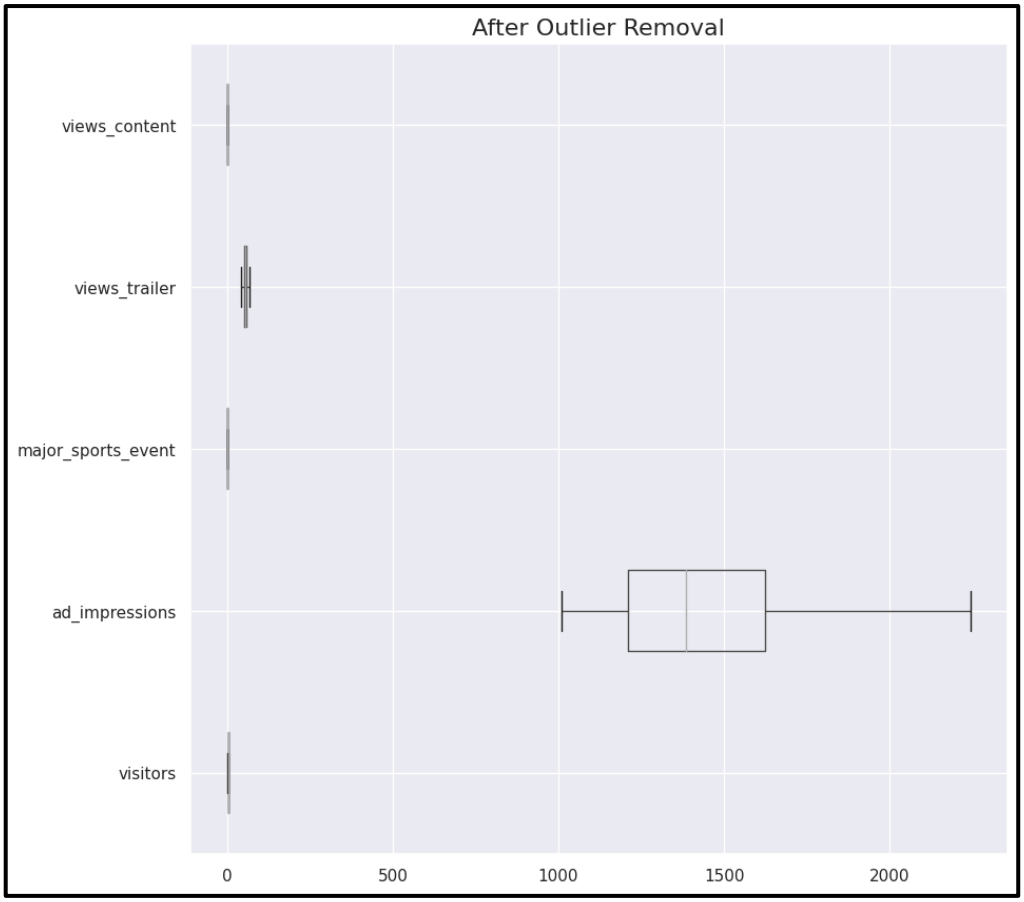

Fig 34: After Outlier Removed

Observation:

- Reduced Range: The overall range of values for all variables has significantly decreased compared to the previous box plot.
- Consistent Distribution: The distribution of ad_impressions and visitors appears to be more centered and symmetrical after outlier removal.
- Potential Issues: The box plots for views_content, views_trailer, and major_sports_event still show a relatively wide range, indicating potential further outliers or skewness.
- Outlier Impact: The removal of outliers has noticeably affected the visual representation of the data, making it easier to identify patterns and trends.

Feature engineering

Replacing the integer data type of major_sports_event to object data type

	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 35: Replacing the integer data type of major_sports_event to object data type

Observation:

- Now this code replaces the numerical values in the major_sports_event column with the corresponding text labels.

Checking the info after replacing data type

```
<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   object
 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), object(4)
memory usage: 62.6+ KB
```

Fig 36: Checking the info after replacing data type

Observation:

major_sports_event is now changed from int to object data type

Data preparation for modelling

Creating Dummy Variables

	visitors	ad_impressions	views_trailer	views_content	major_sports_event_yes	genre_Comedy	genre_Drama	genre_Horror	genre_Others	genre_Romance	...	gen
0	1.67	1113.81	56.70	0.51	False	False	False	True	False	False	...	
1	1.46	1498.41	52.69	0.32	True	False	False	False	False	False	...	
2	1.47	1079.19	48.74	0.39	True	False	False	False	False	False	...	
3	1.85	1342.77	49.81	0.44	True	False	False	False	False	False	...	
4	1.46	1498.41	55.83	0.46	False	False	False	False	False	False	...	

5 rows × 21 columns

Fig 37: Creating Dummy Variables

Splitting the Data

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

Fig 38: Splitting the Data

Model Selection and Training

Build the Model

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:	Sat, 03 Aug 2024	Prob (F-statistic):	2.15e-136
Time:	10:18:00	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 39: OLS regression result

Interpreting the Regression Results

- 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.
- 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
- 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

- Intercept (Constant): This is the predicted value of the dependent variable when all independent variables are zero. However, it often lacks practical interpretation in real-world scenarios.
- Coefficient for Continuous Variables: Represents the change in the dependent variable for a one-unit increase in the independent variable, holding other variables constant. For example, a coefficient of 0.1161 for 'visitors' means that for every one million increases in visitors, the content views are expected to increase by 0.1161 million, keeping other factors unchanged.
- Coefficient for Categorical Variables: Represents the difference in the dependent variable between the reference category and the category represented by the coefficient. For instance, if 'major_sports_event_yes' has a coefficient of -0.0619, it implies that the content views are expected to be 0.0619 million lower when a major sports event is happening compared to when it's not (the reference category).

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.

Summary:

- The model seems to fit the data well, with a high R-squared and significant F-statistic.
- Most predictors are statistically significant, with p-values less than 0.05.
- Multicollinearity does not appear to be a major issue, as indicated by the low condition number.
- Adjusted R-squared: 0.615 indicates a good model fit, explaining 61.5% of the variability in views_content.
- Constant Coefficient: 0.4712 suggests that when all predictors are zero, views_content would be 0.4712.
- Coefficient for x1: 0.0291 shows that each unit increase in x1 is associated with an increase of 0.0291 units in views_content.

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.

Evaluating the Train and Test set

Checking model performance on train set (seen 70% data)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.059142	0.046708	0.64141	0.630303	10.199322

Fig 40: Checking model performance on train set (seen 70% data)

Observation:

Model Performance Observations

R-squared (Training): The R-squared value is 0.64141, indicating that the model explains approximately 64.14% of the variance in the training data. Although this is not a perfect fit, it suggests that the model is capturing a significant portion of the underlying pattern in the data, and there is no sign of underfitting.

Comparison of Train and Test Metrics

- RMSE and MAE: The RMSE and MAE values for the training data are 0.059142 and 0.046708, respectively. If these metrics are comparable with those of the test data, it indicates that the model generalizes well to unseen data, meaning there is no sign of overfitting.
- Mean Absolute Error (MAE): The MAE on the training data is 0.046708. This suggests that the model's predictions, on average, differ from the actual values by approximately 0.0467 units in the training dataset. If, on the test data, the MAE is around 0.34, it means that on average, the model can predict the anime ratings with a mean error of 0.34.
- Mean Absolute Percentage Error (MAPE): The MAPE on the training data is 10.199322. If the MAPE on the test data is around 12.6%, this indicates that the model's predictions are, on average, within 12.6% of the actual anime ratings on the test dataset. This level of error is generally acceptable, depending on the context and the specific requirements of the prediction task.

Checking model performance on test set (seen 30% data)

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.065309	0.052897	0.561854	0.528757	11.506957

Fig 41: Checking model performance on test set (seen 30% data)

Observation:

- RMSE (Root Mean Squared Error): 0.065309. The RMSE for the test data is slightly higher than the training data, indicating that the model's predictions on the test data have a standard deviation of approximately 0.0653 units from the actual values.
- MAE (Mean Absolute Error): 0.052897. The MAE on the test data is 0.052897, which is also slightly higher than the training data. This suggests that the model's predictions, on average, differ from the actual values by about 0.0529 units on the test dataset.
- R-squared: 0.561854. The R-squared value for the test data is 0.561854, meaning that about 56.18% of the variance in the test data is explained by the model. This is lower than the R-squared on the training data, indicating a potential slight drop in model performance when applied to unseen data.
- Adj. R-squared (Adjusted R-squared): 0.528757. The adjusted R-squared value is 0.528757, which is also lower than the training data, consistent with the lower R-squared. This indicates that the model may have a slight tendency to fit noise or may not fully capture the complexity of the test data.
- MAPE (Mean Absolute Percentage Error): 11.506957. The MAPE on the test data is 11.506957, meaning that the model's predictions are, on average, within 11.51% of the actual values in the test data. This error rate is reasonably close to the training MAPE, which indicates that the model has maintained its prediction accuracy when applied to new data.

Insights:

The model explains a good portion of the variance (R-squared around 64%) and has relatively low error metrics (RMSE, MAE, MAPE), indicating a reasonably good fit. However, there is still room for improvement, particularly in improving the R-squared and adjusted R-squared values.

The test performance metrics show that the model generalizes fairly well, with only a slight decrease in performance compared to the training data. The differences in RMSE, MAE, R-squared, and adjusted R-squared suggest that the model is neither underfitting nor significantly overfitting.

The MAPE being close to 11.51% on the test data confirms that the model's predictions are relatively accurate and consistent with the training phase.

Testing the assumptions of linear regression model

We will be checking the following Linear Regression assumptions:

- No Multicollinearity
- Linearity of variables
- Independence of error terms
- Normality of error terms
- No Heteroscedasticity

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.

Let's define a function to check VIF.

Checking the value of VIF

	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

Fig 42: Checking the value of VIF

Observation:

There is one column with high VIF value, indicating presence of strong multicollinearity

We will systematically drop numerical column with VIF > 2

We will ignore the VIF values for dummy variables and the constant (intercept)

Removing Multicollinearity

To remove multicollinearity

- Drop every column one by one that has a VIF score greater than 5.
- Look at the adjusted R-squared and RMSE of all these models.
- Drop the variable that makes the least change in adjusted R-squared.
- Check the VIF scores again.
- Continue till you get all VIF scores under 5.

Let's define a function that will help us do this.

Look at the adjusted R-squared and RMSE of all these models.

Treating the column genre_others

	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	genre_Others	0.631244	0.060018

Fig 43: Treating the column genre_others

Drop the variable that makes the least change in adjusted R-squared.

Check the VIF scores again

Value of VIF after remove multicollinearity

VIF after dropping genre_Others

	feature	VIF
0	const	153.669019
1	visitors	1.020698
2	ad_impressions	1.028897
3	views_trailer	1.027870
4	major_sports_event_yes	1.070178
5	genre_Comedy	1.204448
6	genre_Drama	1.223159
7	genre_Horror	1.205554
8	genre_Romance	1.171698
9	genre_Sci-Fi	1.205336
10	genre_Thriller	1.206277
11	dayofweek_Monday	1.063674
12	dayofweek_Saturday	1.155123
13	dayofweek_Sunday	1.150353

Fig 44: Value of VIF after remove multicollinearity

Observation:

This indicates all the VIF value are less than 2.

We have dealt with multicollinearity in the data.

Rebuild the Model

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

```
                            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:                Sat, 03 Aug 2024   Prob (F-statistic):          3.02e-137
Time:                        05:39:04   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 45: rebuilding the model using the updated set of predictors variables

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 βi is the coefficient of the ith independent variable.

Ho: Independent feature is not significant (βi=0)

Ha: Independent feature is that it is significant (βi≠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.

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

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

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.

Dropping the variable with high p-value

```
['const', 'visitors', 'views_trailer', 'major_sports_event_yes', 'genre_Sci-Fi', 'dayofweek_Saturday', 'dayofweek_Sunday', 'dayofweek_Thursda
y', 'dayofweek_Tuesday', 'dayofweek_Wednesday', 'season_Spring', 'season_Summer', 'season_Winter']
```

Fig 46: Dealing with high p-value variables using loop

Rebuild the Model

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

```
                            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:                Sat, 03 Aug 2024   Prob (F-statistic):          5.85e-143
Time:                        05:39:04   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 47: Rebuilding the model using the updated set of predictors variables

Observation:

We can see that R-squared has dropped from 0.641 to 0.639, 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

Evaluating the Train and Test set

Checking model performance on train set (seen 70% data)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.059363	0.046772	0.638723	0.631877	10.214645

Fig 48: Checking model performance on train set (seen 70% data)

Checking model performance on test set (seen 30% data)

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.065566	0.053145	0.558408	0.538336	11.586675

Fig 49: Checking model performance on test set (seen 30% data)

Observation:

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.

Assumptions of Linear Regression

These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or use the model to make a prediction.

For Linear Regression, we need to check if the following assumptions hold:

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality of error terms
5. No strong Multicollinearity

Test for Linearity and Independence

Why the test?

- Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable.

How to check linearity?

- Make a plot of fitted values vs residuals. If they don't follow any pattern (the curve is a straight line), then we say the model is linear otherwise model is showing signs of non-linearity.

How to fix if this assumption is not followed?

- We can try different transformations.

Creating a dataframe with actual, fitted and residual values

	Actual Values	Fitted Values	Residuals
731	0.40	0.428238	-0.028238
716	0.70	0.629866	0.070134
640	0.42	0.450182	-0.030182
804	0.55	0.576973	-0.026973
737	0.59	0.556401	0.033599

Fig 50: Creating a dataframe with actual, fitted and residual values

Plotting the fitted values vs residuals

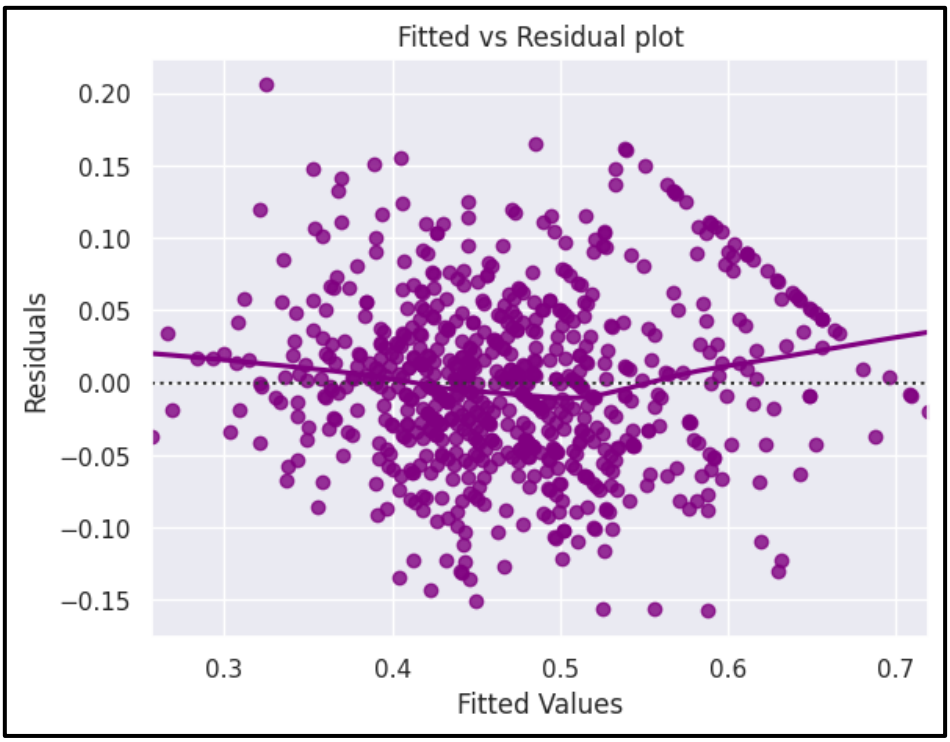

Fig 51: Plot the fitted values vs residuals

Observation:

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

If there exists any pattern in this plot, we consider it as a sign of non-linearity in the data, meaning the model doesn't capture non-linear effects.

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

Test for Normality

What is the test?

Error terms/residuals should be normally distributed.

If the error terms are not normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.

What does non-normality indicate?

It suggests that there are a few unusual data points which must be studied closely to make a better model.

How to check the Normality?

It can be checked via QQ Plot - residuals following normal distribution will make a straight line plot, otherwise not.

Another test to check for normality is the Shapiro-Wilk test.

1. Null hypothesis: Residuals are normally distributed.
2. Alternate hypothesis: Residuals are not normally distributed.

How to Make residuals normal?

We can apply transformations like log, exponential, arcsinh, etc as per our data.

Normality of residuals

Visualizing the normality of residuals using histogram

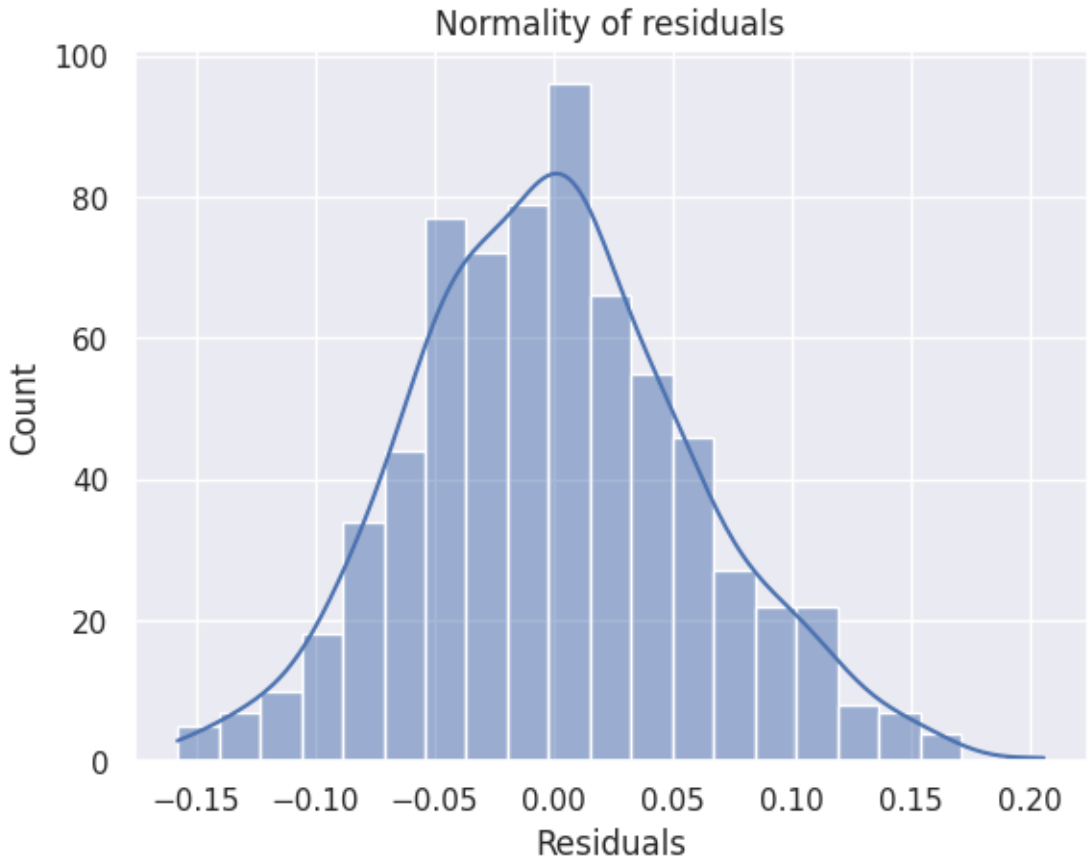

Fig 52: Visualizing the normality of residuals using histogram

Observation:

The residual terms are normally distributed.

Q-Q plot of residuals

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.

Visualizing the normality of residuals using Q-Q plot

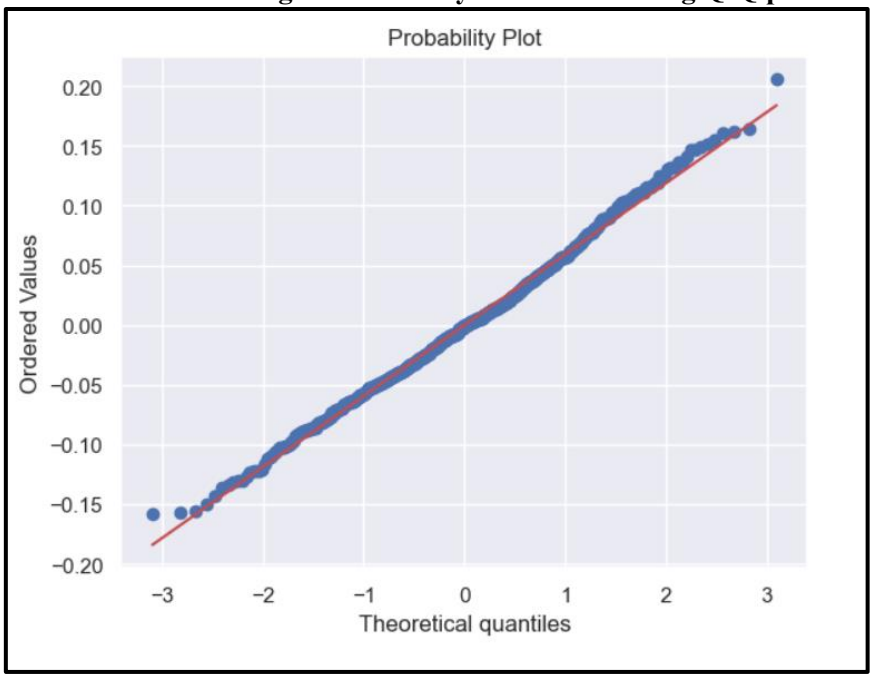

Fig 53: Visualizing the normality of residuals using Q-Q plot

Observation:

Most of the points are lying on the straight line in QQ plot

The residuals more or less follow a straight line except for the tails.

Shapiro-Wilk test

Let's check the results of the Shapiro-Wilk test.

Check the results of the Shapiro-Wilk test

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

Fig 54: Check the results of the Shapiro-Wilk test

Observation:

The Shapiro-Wilk test result showed a p-value of 0.04188, indicating that the residuals are not perfectly normally distributed.

Insights

Normality Test Result: The test statistic is close to 1, suggesting that the data distribution is relatively close to normal. However, the p-value of 0.0419 is less than the common significance level of 0.05.

Interpretation of P-value: Since the p-value is less than 0.05, we reject the null hypothesis of the Shapiro-Wilk test, which states that the data is normally distributed. This indicates that there is evidence suggesting the data does not follow a normal distribution.

Practical Implication: 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. This means we need to consider alternative methods or transformations if normality is a crucial assumption for your analysis or modeling.

Test for Homoscedasticity

- **Homoscedacity** - If the variance of the residuals are symmetrically distributed across the regression line, then the data is said to homoscedastic.
- **Heteroscedacity** - If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic. In this case the residuals can form an arrow shape or any other non symmetrical shape.

Why the test?

- The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers.

How to check if model has Heteroscedasticity?

- Can use the goldfeldquandt test. If we get p-value > 0.05 we can say that the residuals are homoscedastic, otherwise they are heteroscedastic.

How to deal with Heteroscedasticity?

- Can be fixed via adding other important features or making transformations.

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

Null hypothesis : Residuals are homoscedastic

Alternate hypothesis : Residuals have hetroscedasticity

How to fix if this assumption is not followed?

Heteroscedasticity can be fixed by adding other important features or making transformations.

Check the results of Heteroscedasticity

```
[('F statistic', 1.1616405144400679), ('p-value', 0.08477906249012608)]
```

Fig 55: Check the results of Heteroscedasticity

Observation:

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

Predictions on test data

Now that we have checked all the assumptions of linear regression and they are satisfied, let's go ahead with prediction.

Predictions on the test set

	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

Fig 56: Predictions on the test set

Observation:

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

Final Model

Let's recreate the final model and print it's summary to gain insights.

```
                            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:                Sat, 03 Aug 2024   Prob (F-statistic):          5.85e-143
Time:                        05:39:06   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 57: Final Model

Observation:

R-squared (0.639): This indicates that 63.9% of the variance in the dependent variable (views_content) is explained by the independent variables in the model.

Adj. R-squared (0.632): This is slightly lower than the R-squared, which is typical as it adjusts for the number of predictors in the model. It suggests a good fit.

F-statistic (101.2) & Prob(F-statistic: 5.85e-143): The model is statistically significant overall, as indicated by the very low p-value associated with the F-statistic, suggesting that the independent variables collectively predict the dependent variable better than a model with no predictors.

Coefficients and Significance:

- **Intercept (const -0.2672)**: When all independent variables are zero, the expected value of views_content is -0.2672, which might be interpreted with caution depending on the context.
- **visitors (0.1169, p < 0.001)**: For each additional visitor, views_content is expected to increase by 0.1169 units, holding other factors constant.
- **views_trailer (0.0093, p < 0.001)**: Each additional view of the trailer is associated with a 0.0093 unit increase in views_content.
- **major_sports_event_yes (-0.0622, p < 0.001)**: If a major sports event is ongoing, views_content is expected to decrease by 0.0622 units.
- **genre_Sci-Fi (0.0188, p = 0.011)**: Sci-Fi content is associated with a 0.0188 unit increase in views_content compared to the reference genre.
- **Day of the Week Effects**: Saturday (0.0515), Sunday (0.0377), Thursday (0.0173), Tuesday (0.0452), and Wednesday (0.0397) all have positive effects on views_content, with varying levels of significance.
- **Seasonal Effects**: Spring (0.0265), Summer (0.0435), and Winter (0.0295) all have significant positive effects on views_content compared to the reference season.

Evaluating the Train and Test set

Checking model performance on train set (seen 70% data)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.059363	0.046772	0.638723	0.631877	10.214645

Fig 58: Checking model performance on train set (seen 70% data)

Checking model performance on test set (seen 30% data)

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.065566	0.053145	0.558408	0.538336	11.586675

Fig 59: Checking model performance on test set (seen 30% data)

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

Actionable Insights & Recommendations

Comments on significance of predictors

- **R-squared:** 0.639, indicating that approximately 63.9% of the variance in views_content is explained by the model. This is a reasonably good fit.
- **F-statistic:** 101.2 with a p-value of 5.85e-143, suggesting that the overall model is statistically significant.

Interpretation of Predictors

1. **visitors**:
    - **Coefficient:** 0.1169
    - **p-value:** 0.000
    - **Interpretation:** Highly significant predictor. For every additional visitor, the content views increase by 0.1169, holding other variables constant.
2. **views_trailer**:
    - **Coefficient:** 0.0093
    - **p-value:** 0.000
    - **Interpretation:** Highly significant. An increase in trailer views leads to a small but significant increase in content views.
3. **major_sports_event_yes**:
    - **Coefficient:** -0.0622
    - **p-value:** 0.000
    - **Interpretation:** Significant. The presence of a major sports event decreases content views by 0.0622. This negative relationship may indicate that viewers are less likely to watch content when a major sports event is happening.

4. **genre_Sci-Fi**:
    - **Coefficient:** 0.0188

- **p-value:** 0.011
- **Interpretation:** Significant. Content in the Sci-Fi genre is associated with an increase in views, though the effect size is small.

5. **Day of the Week Variables:**
   - **Saturday (0.0515), Sunday (0.0377), Thursday (0.0173), Tuesday (0.0452), Wednesday (0.0397)**:
     - **p-values:** All are statistically significant ($p < 0.05$).
     - **Interpretation:** Content views increase on these days compared to the baseline (which is likely Monday). The impact is largest on Saturday and smallest on Thursday, suggesting that certain days are more favorable for content consumption.
6. **Season Variables:**
   - **Spring (0.0265), Summer (0.0435), Winter (0.0295)**:
     - **p-values:** All are statistically significant ($p < 0.05$).
     - **Interpretation:** Content views are higher in these seasons compared to the omitted category (likely Autumn). The largest effect is seen in Summer, followed by Winter and Spring.

Overall Comments

- All the predictors have statistically significant effects on content views.
- The signs of the coefficients make intuitive sense; for example, more visitors and trailer views are associated with higher content views, and content views tend to decrease during major sports events.
- Days of the week and seasons show expected patterns of variation, reflecting different viewing habits across the week and year.

Key takeaways for the business

Here are the key takeaways for the business based on the OLS regression results:

1. Visitor Traffic is Crucial:

- **Insight:** The number of visitors to the platform is a strong driver of content views.
- **Action:** Focus on strategies to increase traffic, such as marketing campaigns, partnerships, or SEO enhancements, to boost overall content consumption.

2. Trailer Views Drive Content Engagement:

- **Insight:** Views of trailers are positively associated with content views, though the effect is small, it is highly significant.
- **Action:** Invest in creating engaging trailers and prominently feature them on the platform and in promotional materials to increase full content viewership.

3. Impact of Major Sports Events:

- **Insight:** Major sports events negatively impact content views, likely due to competing for viewer attention.
- **Action:** Consider scheduling important content releases around major sports events to avoid dips in viewership. Alternatively, create content that can complement sports events or target different audiences during these times.

4. Genre-Specific Performance:

- **Insight:** Sci-Fi content appears to be more popular and drives slightly higher views.
- **Action:** If Sci-Fi is a strategic genre for your platform, consider expanding your Sci-Fi content library. Additionally, analyze other genres in a similar manner to identify other high-performing categories.

5. Optimal Days for Content Release:

- **Insight:** Viewership tends to peak on Saturdays, followed by Wednesday and Tuesday, with Sunday also showing strong engagement.
- **Action:** Schedule major content releases or promotions for these days to maximize viewership. Additionally, analyze other weekdays to identify opportunities for growth on lower-viewership days.

6. Seasonal Trends:

- **Insight:** Content views are highest during the Summer, followed by Winter and Spring.
- **Action:** Plan content releases and marketing efforts to align with these seasonal trends. Consider launching special seasonal content or campaigns to take advantage of these periods of high engagement.

7. Strategic Content Planning:

- **Insight:** The identified factors can be used to inform a more strategic approach to content scheduling, marketing, and platform management.
- **Action:** Develop a content calendar that aligns with high-viewership days and seasons. Use data-driven insights to optimize content placement and promotions.

8. Understanding Audience Behavior:

- **Insight:** The model highlights patterns in how audiences consume content, providing valuable information for better targeting and personalization.
- **Action:** Leverage this understanding to tailor content recommendations, personalize marketing messages, and improve user experience, ultimately driving higher engagement and satisfaction.

By implementing these insights, the business can better align its content strategy with audience preferences and external factors, leading to increased viewership and, potentially, higher revenue.