PM Coded Project Report

Prepared By: Parthasarathi Behura

CONTENTS:

Over-the-top (OTT) media service Page

LIST OF FIGURES:

LIST OF TABLES:

PM Project

Business Context

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

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

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

Objective

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

Data Description

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

Data Dictionary:

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

- views_trailer: Number of views, in millions, of the content trailer
- views_content: Number of first-day views, in millions, of the content

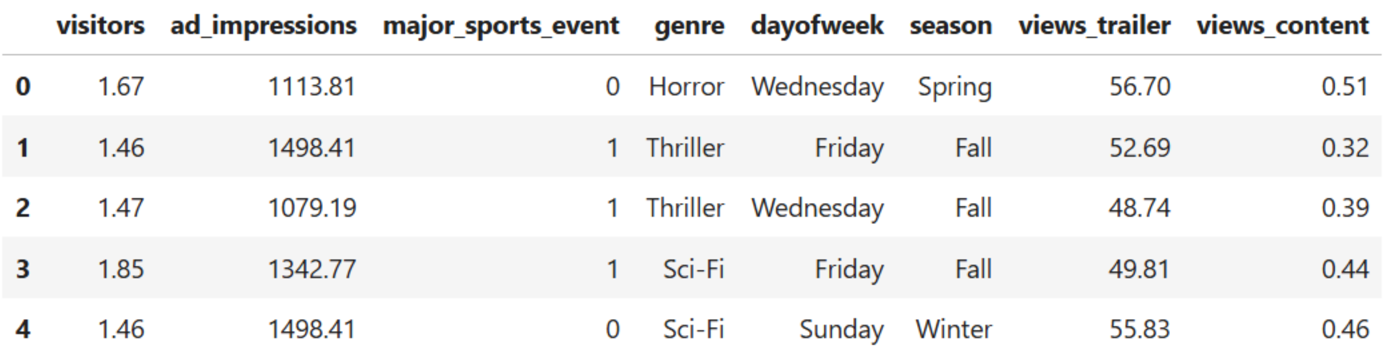
Imported the libraries for the Data are

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn.model_selection
- Sklearn.linear_model
- Sklearn.preprocessing
- sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
- statsmodels.api as sm
- variance_inflation_factor

DATA PROCESSING:

1. There are some information about the dataset, decision makers should have a look.

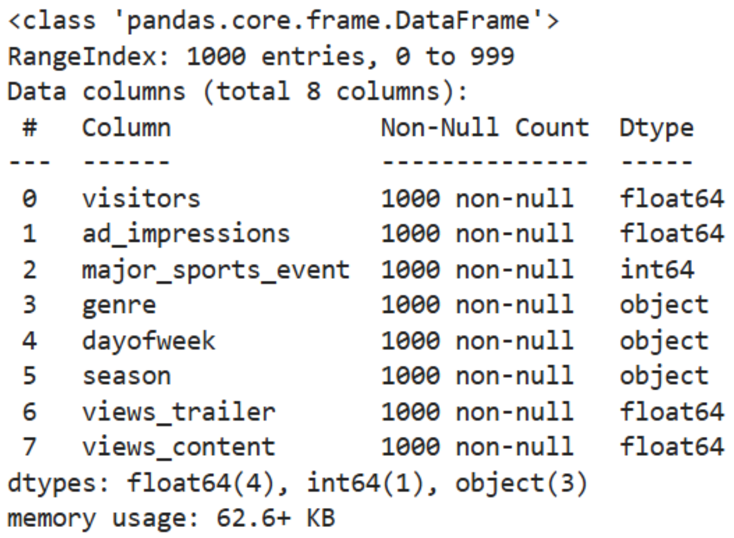
- The dataset is having 8 columns.
- There is a look on the 5 sample rows to check the data type.



	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer	views_content
0	1.67	1113.81	0	Horror	Wednesday	Spring	56.70	0.51
1	1.46	1498.41	1	Thriller	Friday	Fall	52.69	0.32
2	1.47	1079.19	1	Thriller	Wednesday	Fall	48.74	0.39
3	1.85	1342.77	1	Sci-Fi	Friday	Fall	49.81	0.44
4	1.46	1498.41	0	Sci-Fi	Sunday	Winter	55.83	0.46

Table 1: Top five rows of the dataset

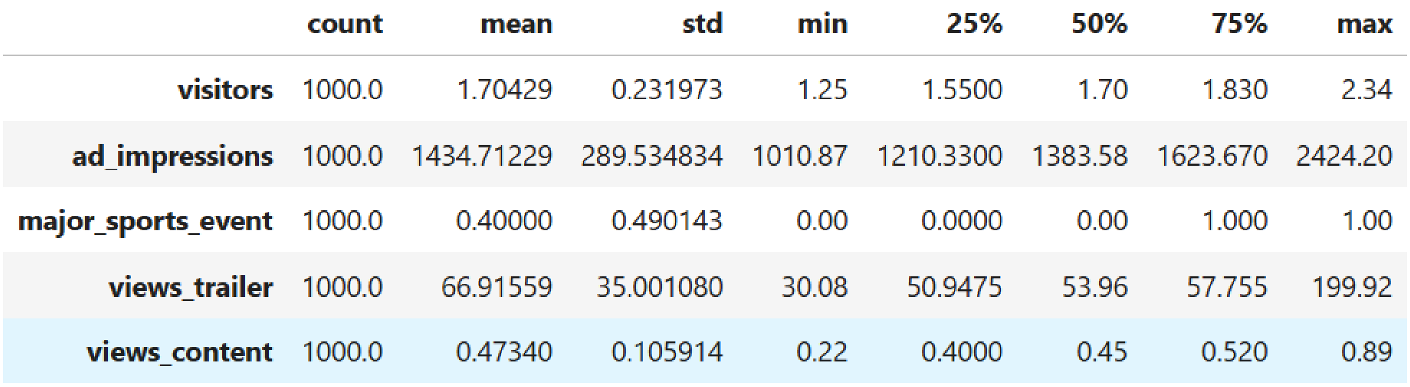
2. While having a look on the data set information, it is found that there are 5 numerical and 3 categorial variables. The below table contains the same information.



```
<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: Basic information of the data type

3. Checking the statistical summary of the data.

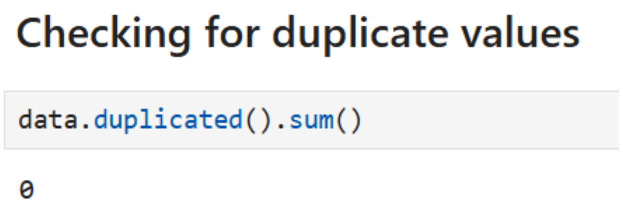


	count	mean	std	min	25%	50%	75%	max
visitors	1000.0	1.70429	0.231973	1.25	1.5500	1.70	1.830	2.34
ad_impressions	1000.0	1434.71229	289.534834	1010.87	1210.3300	1383.58	1623.670	2424.20
major_sports_event	1000.0	0.40000	0.490143	0.00	0.0000	0.00	1.000	1.00
views_trailer	1000.0	66.91559	35.001080	30.08	50.9475	53.96	57.755	199.92
views_content	1000.0	0.47340	0.105914	0.22	0.4000	0.45	0.520	0.89

Table 3: Statistical summary of the data

4. Checking the duplicate values.

Ans:- There are no duplicate values present in the data.



Checking for duplicate values

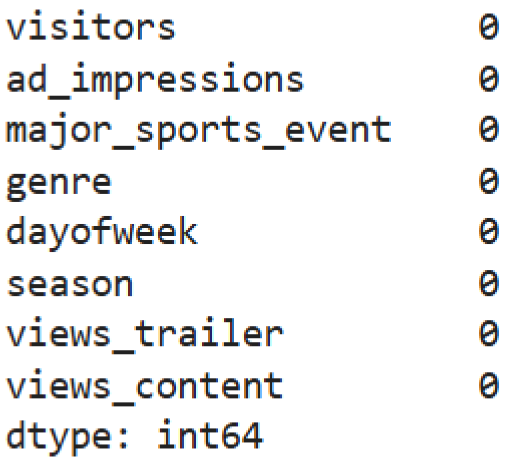
```
data.duplicated().sum()
```

```
0
```

Table 4: Duplicate values in the dataset

5. Checking the null values:

Ans:- There are no missing values present in the dataset.



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

Table 5: Inspecting missing values in the dataset

6. Exploratory Data Analysis

Univariate Analysis: From the below hist plot, we can see that there are outliers present in the “visitors”, “ad_impressions”, “views_trailer”, “views_content”.

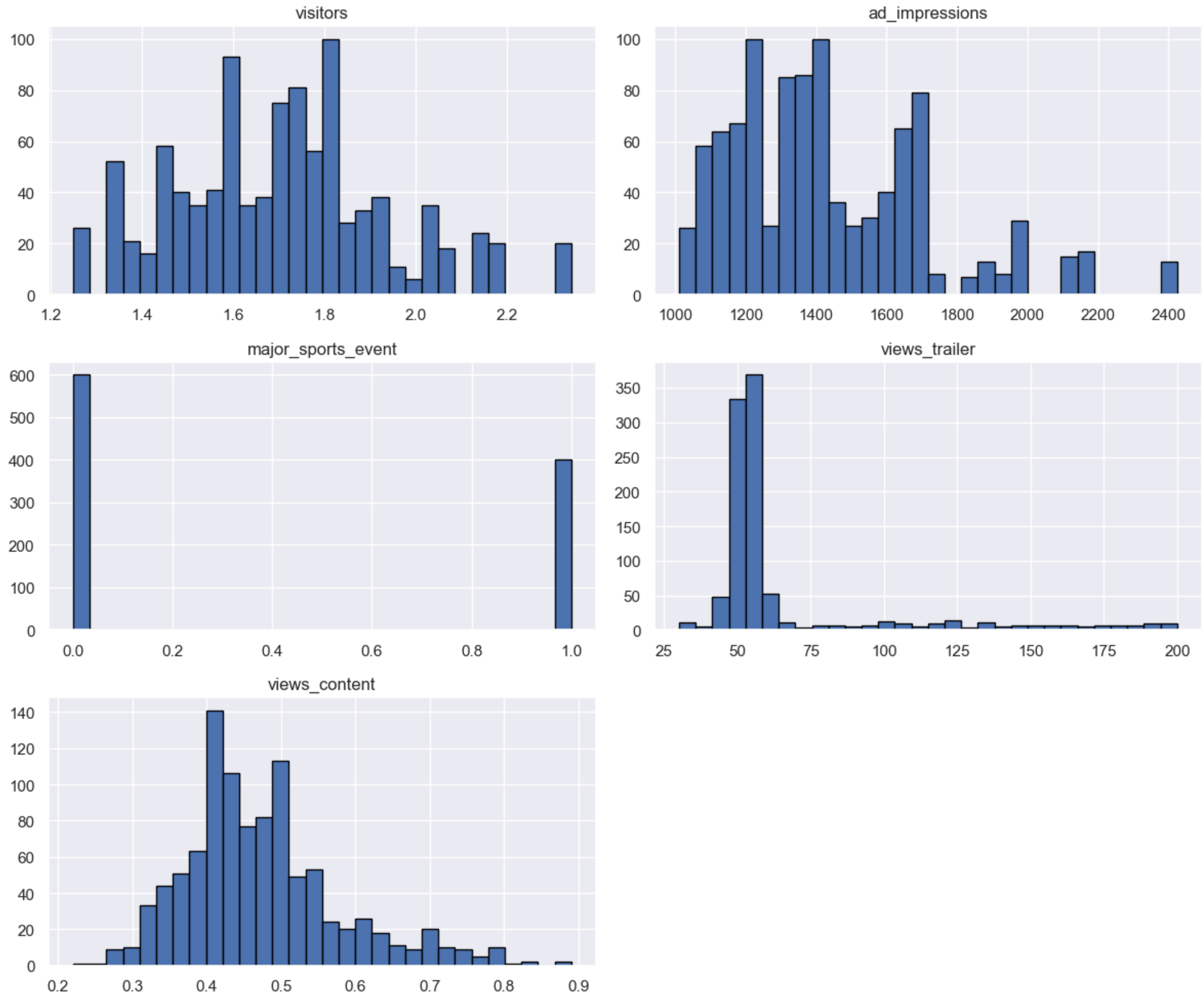


Figure-1 : Hist plots of numerical variables

For further analysis of the “views_content” we plot a histogram to check the presence of outliers. And we found the same close figure from this plot again.

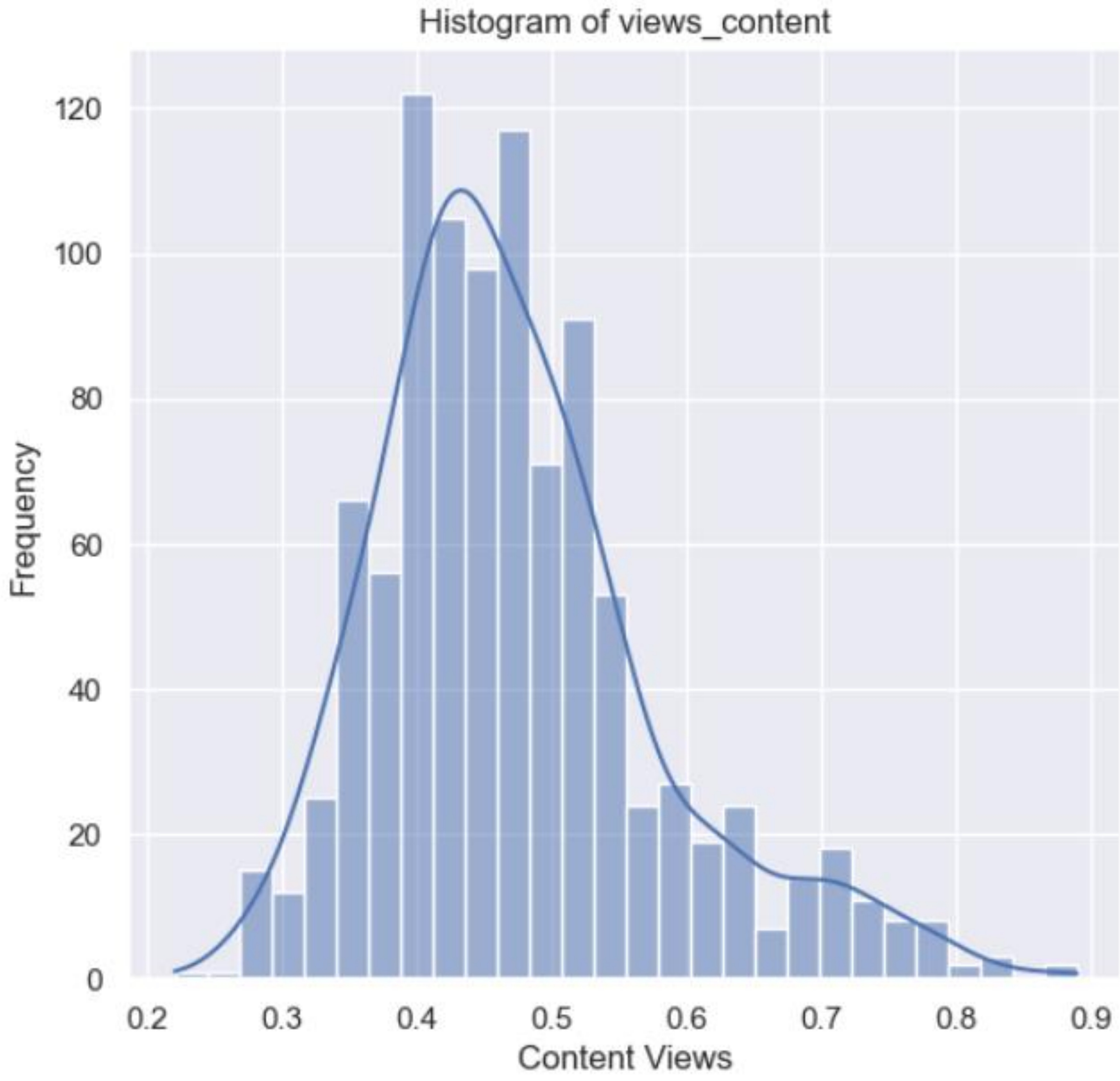


Figure-2 : Hist plots of views_content

Checking the univariate analysis of categorial variables.

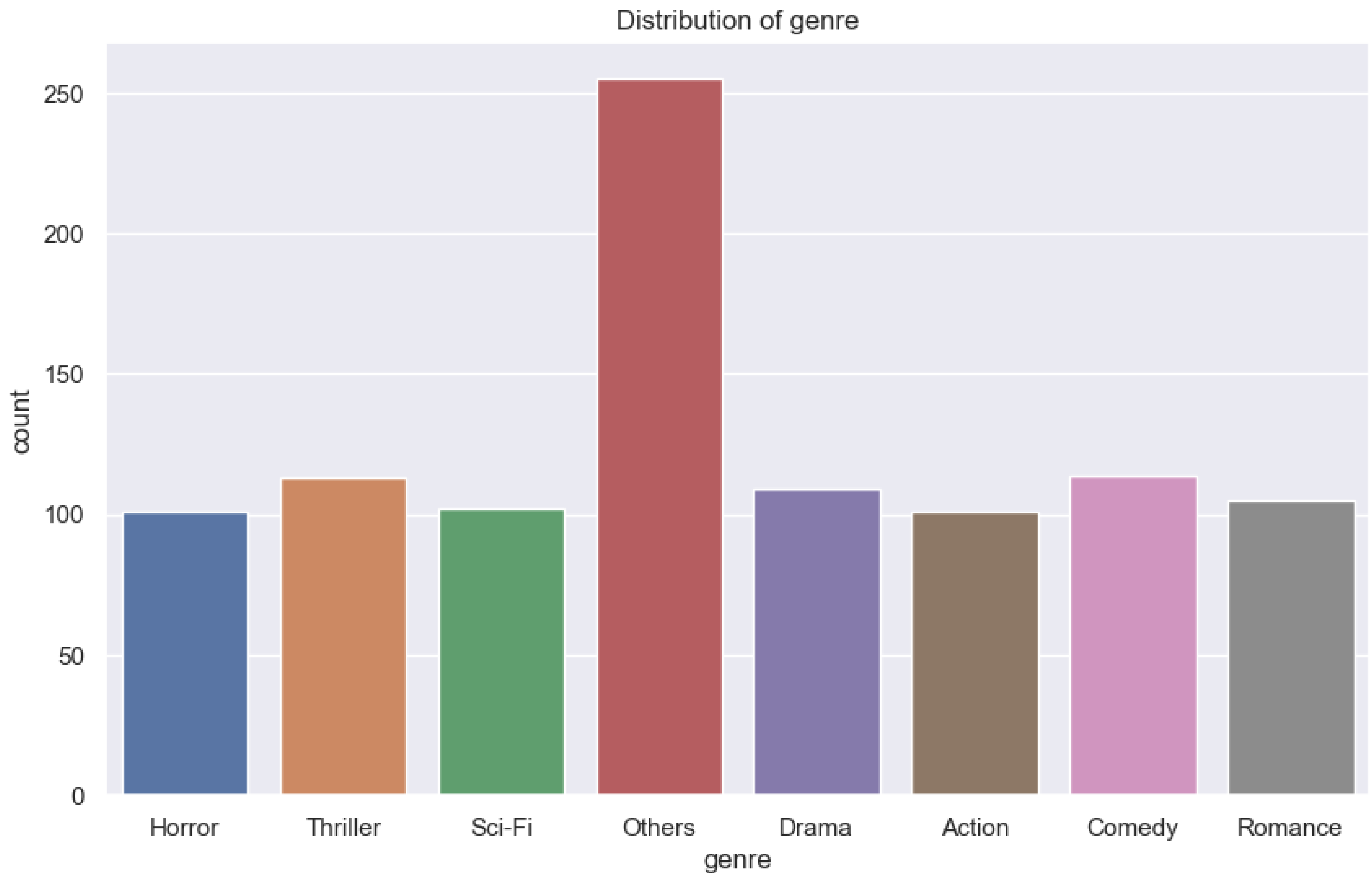


Figure-3 : Count plots of genre

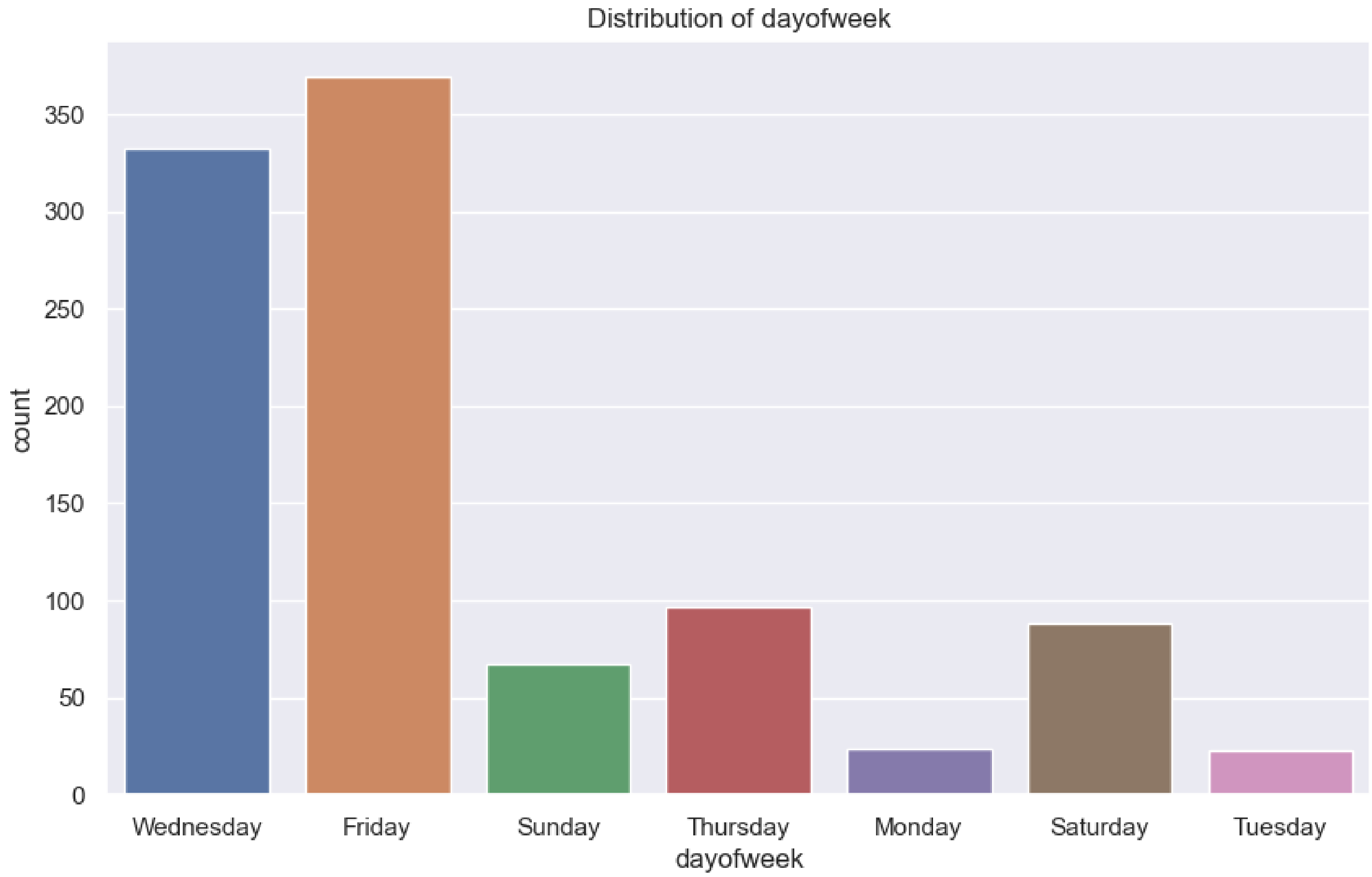


Figure-4 : Count plots of distribution of day_of_week

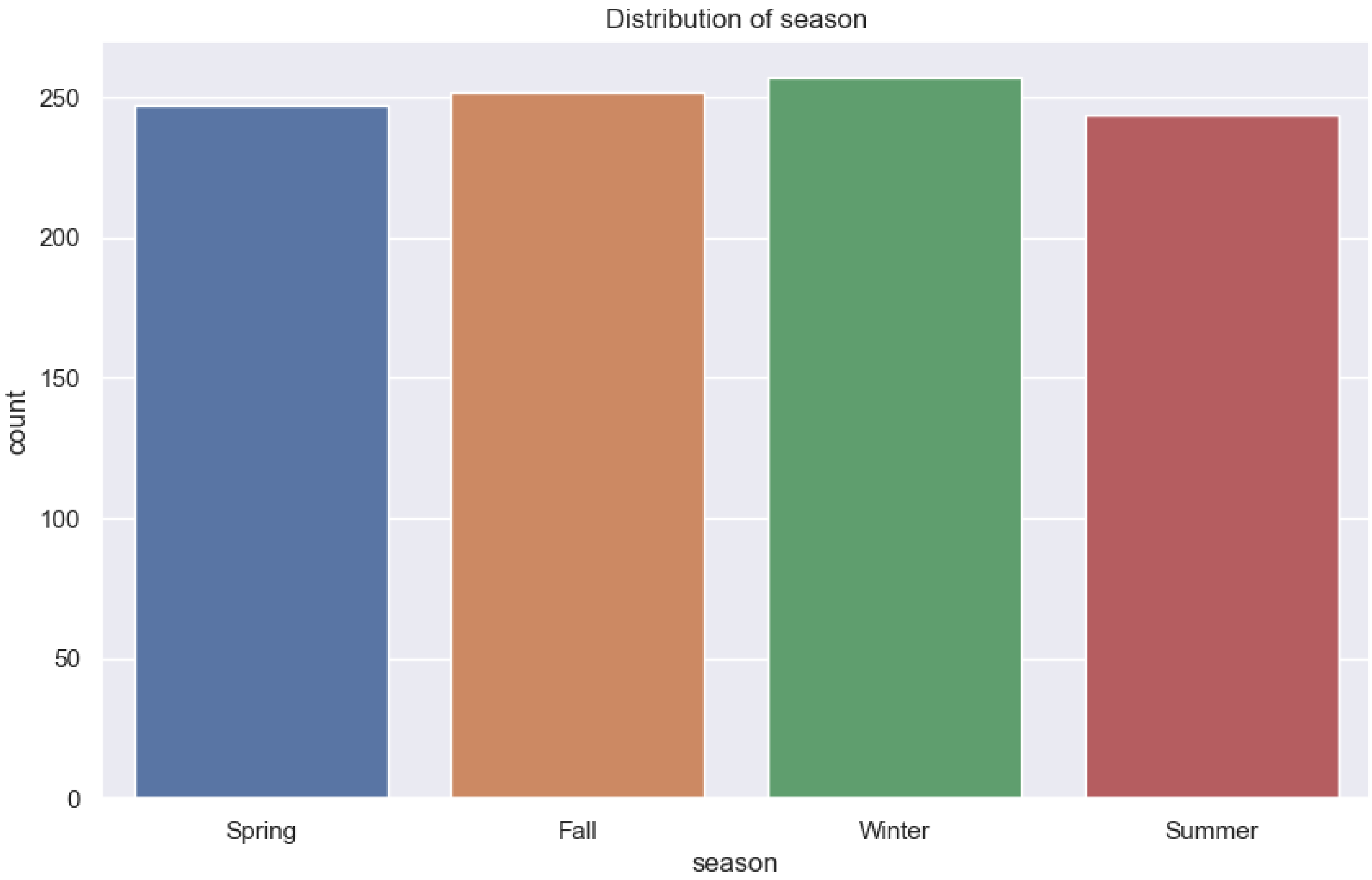


Figure-5 : Count plots of distribution of season

Outliers treatment of the numerical variables needs to be done for the smooth interpretation of data fitting to the model.

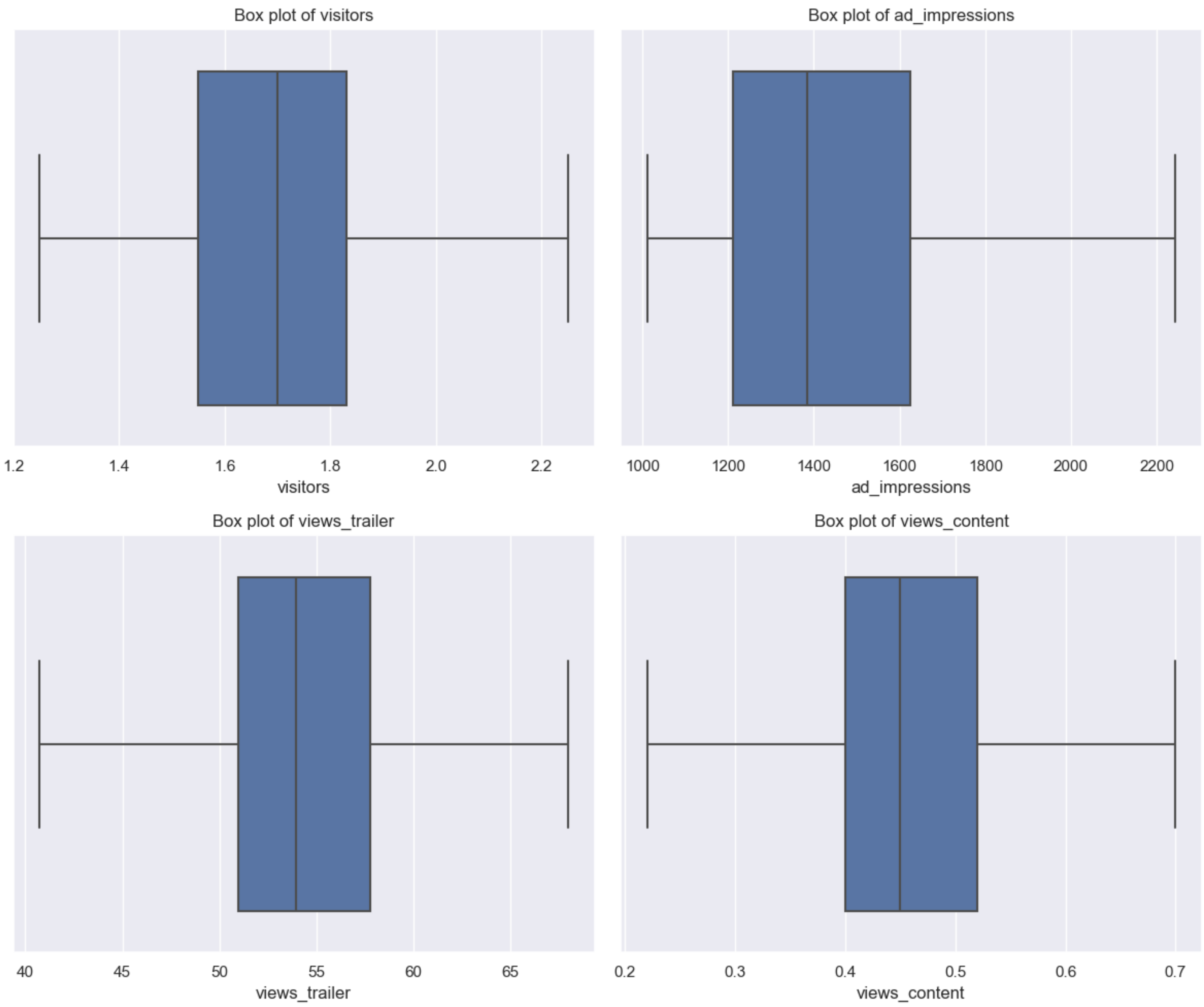


Figure-6 : Box plots of numeric variables after treatment of outliers

Bivariate Analysis:

Correlation Check-

Correlation check between all the numeric variables found to be not a sing presence of co-linearity between the numerical variables.

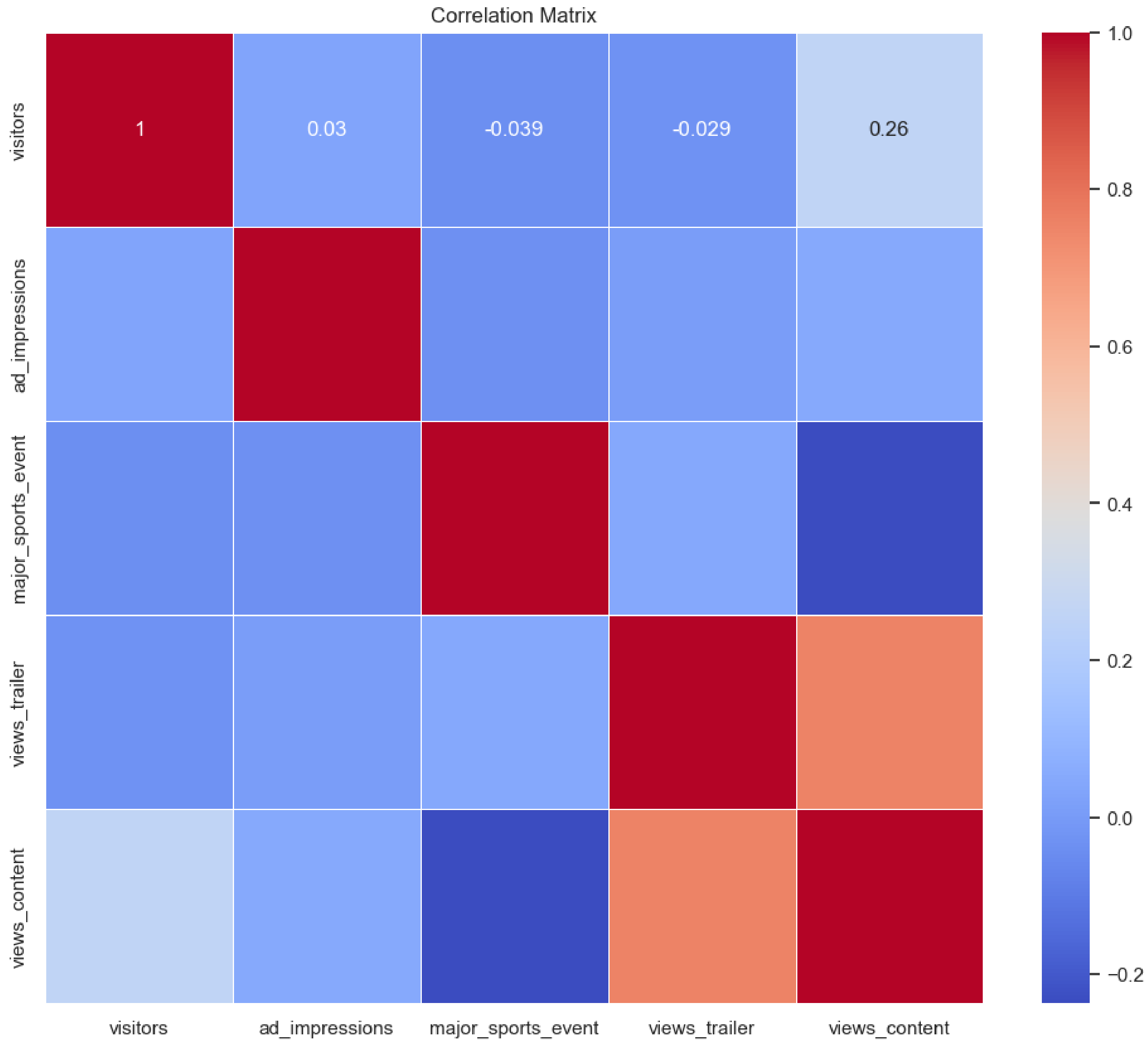


Figure-7 : Heat map of numeric variables

Inspecting the pair plot between “major_sports_events” and other numeric variables.

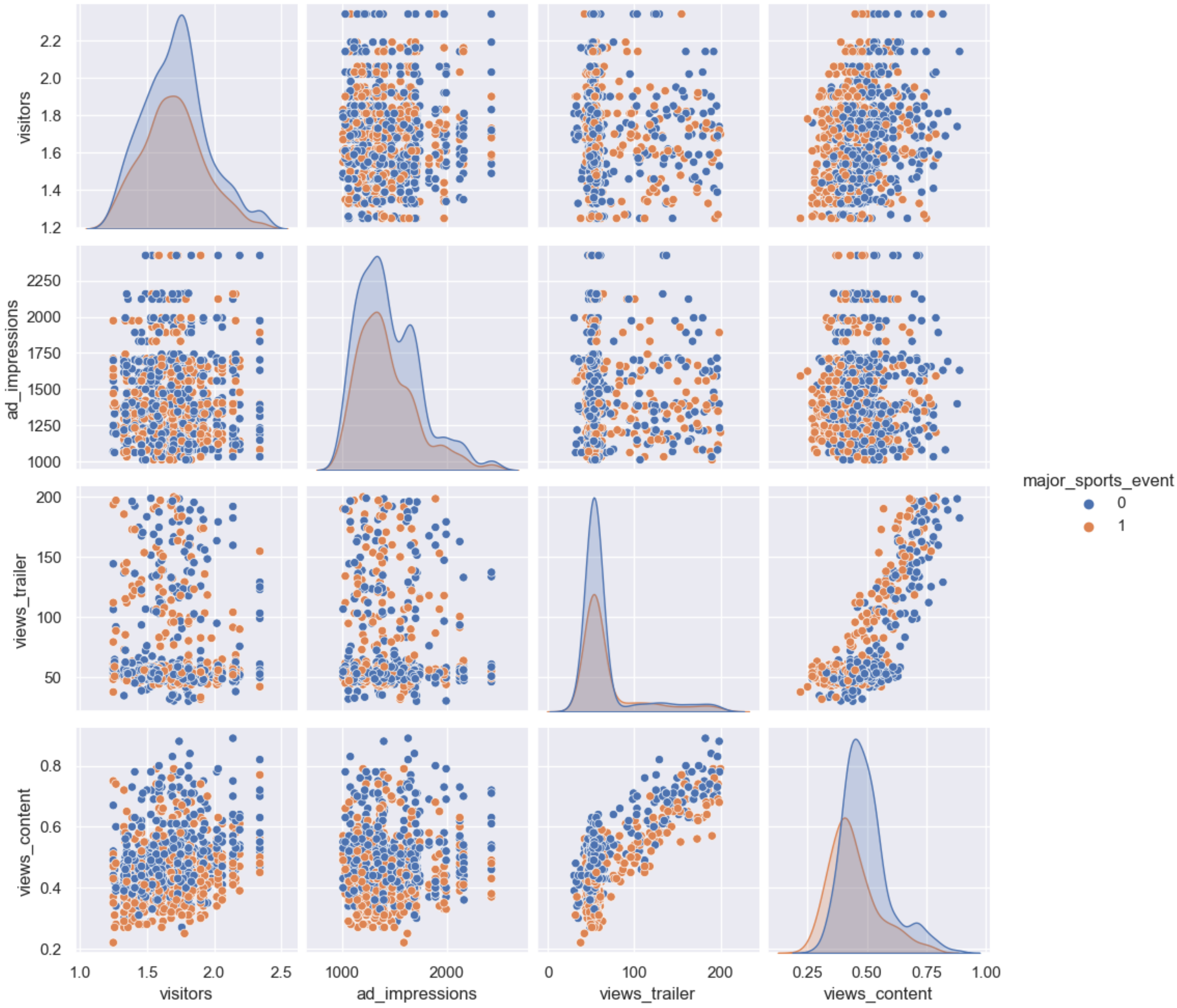


Figure-8 : Pair plot of major_sports_event & numeric variables

Inferences from EDA:

Q1. What does the distribution of content views look like?

Ans: From the Fig. 2, we can clearly see that the content views is having a right skewed data plot. That means,

The distribution appears to be approximately normal (bell-shaped), with most content views clustered around the mean value.

The peak of the histogram indicates that the most common content views value is around 0.4.

Most of the content views values lie between 0.3 and 0.6. The distribution tapers off at both ends, with fewer content views values below 0.3 and above 0.6.

Overall, the content views data is relatively normally distributed with a slight right skew, indicating a typical range of content views values around the mean.

Q2. What does the distribution of genres look like?

Ans: From Fig. 3, we can see that the “others” is the highest in the number of 256.

Q3. 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?

Ans:

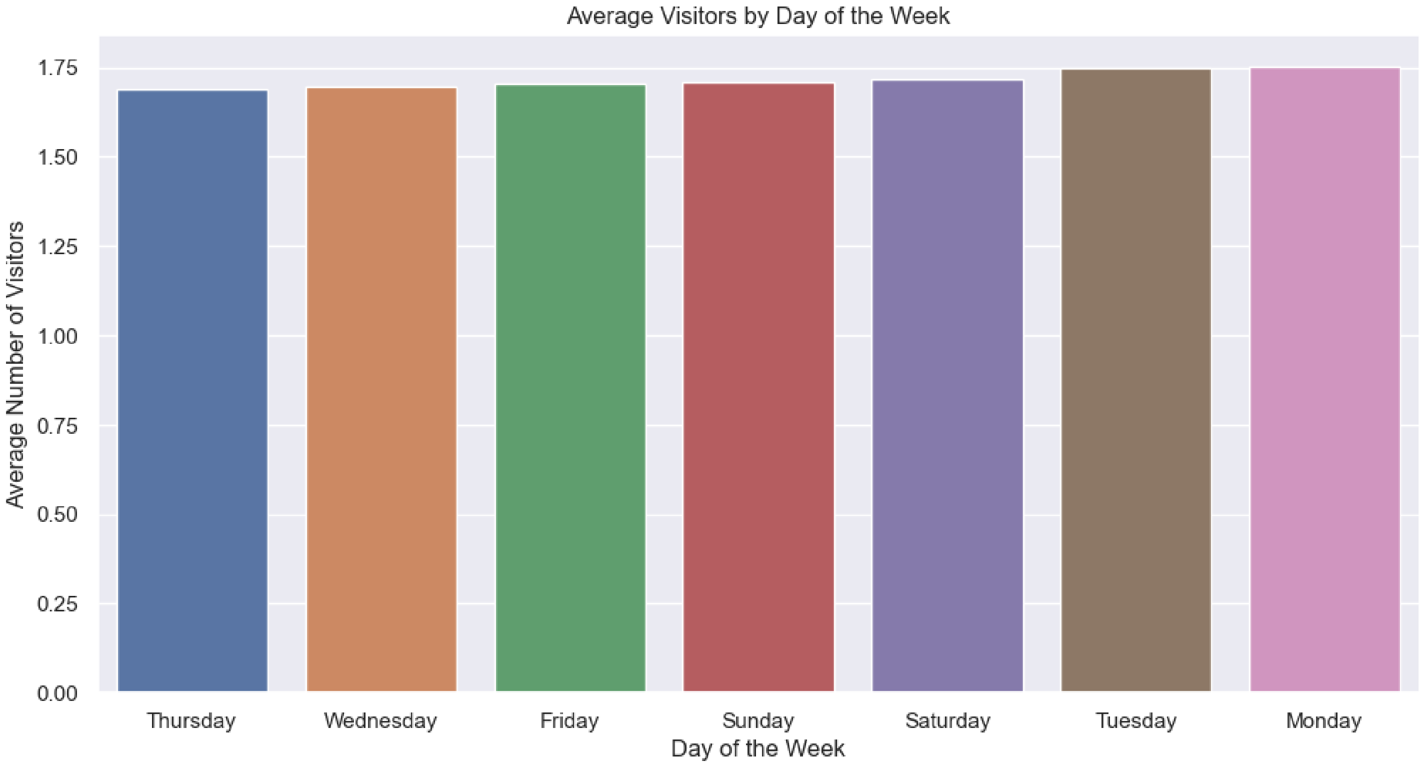


Figure-9 : Bar plot of number of visitors & day of the week

From Fig. 8, we can state that there is no proper pattern to the number of viewers with the day of the week released. But we can see that on Tuesday and Monday week days the number of viewers are comparatively more than the other days.

Q4. How does the viewership vary with the season of release?

Ans:

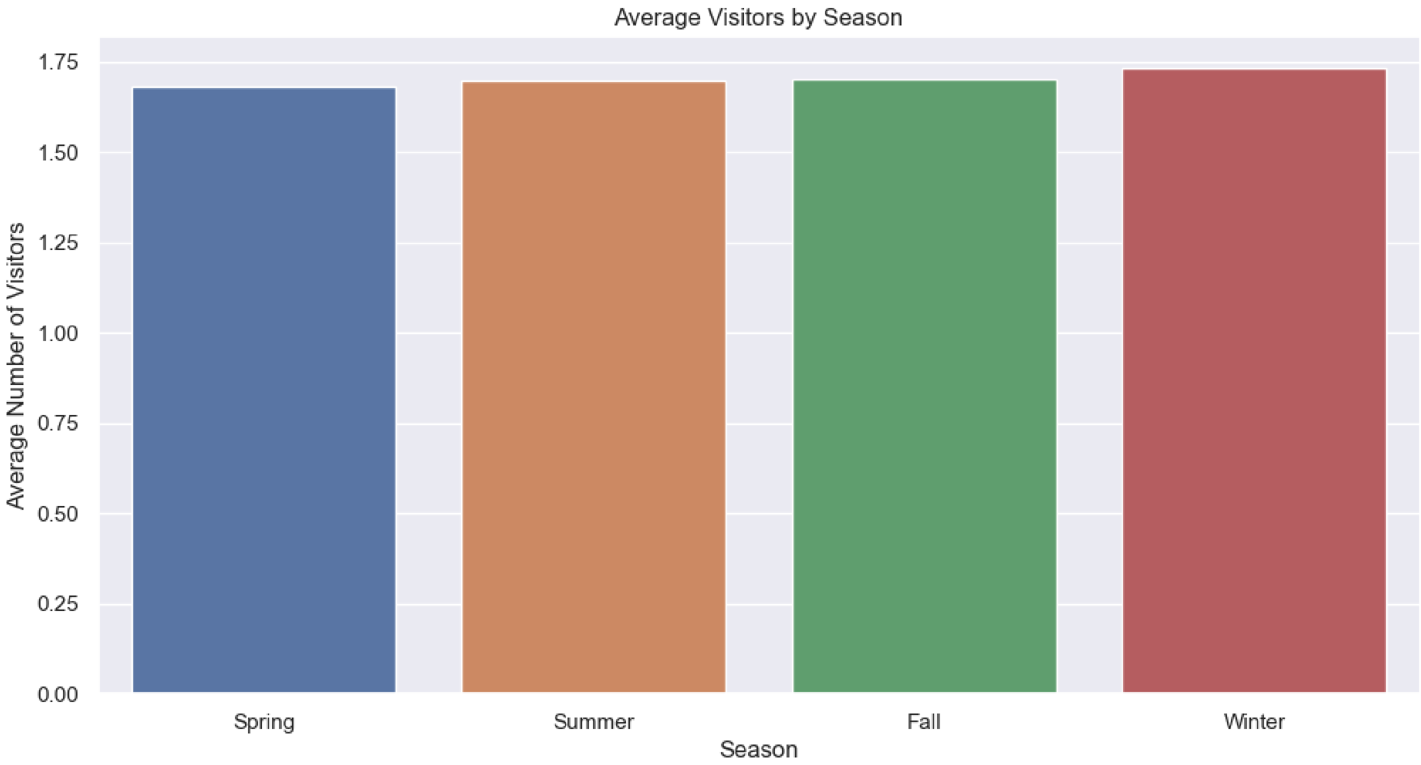


Figure-10 : Bar plot of number of visitors & season

From the Fig. 9, we can clearly see that the average number of viewers are higher in the winter season. Although there is not consisten pattern, but the winter is having more viewers compared to other seasons.

Q5. What is the correlation between trailer views and content views?

Ans: Correlation between trailer views and content views: 0.7539622148205366

From this above correlation point, we can find a strong correlation between the trailer views and content views, as the correlation value is near to 1.

8. Now we'll split the data into train and test to be able to evaluate the model that we build on the train data.

We will build a Linear Regression model using the train data and then check it's performance.

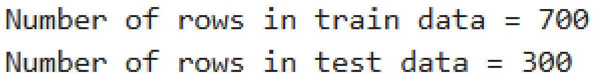
We'll build a model to predict visitors based on other variables.

ad_impressions major_sports_event views_trailer views_contentgenre_Comedy
genre_Drama genre_Horror genre_Others genre_Romance genre_Sci-Fi genre_Thriller

```
        dayofweek_Monday  dayofweek_Saturday  dayofweek_Sunday   dayofweek_Thursday
        dayofweek_Tuesday  dayofweek_Wednesday       season_Spring season_Summer
        season_Winter

0       1113.81      0.0     56.70  0.51   0.0    0.0    1.0    0.0    0.0    0.0    0.0    0.0    0.0
        0.0     0.0  0.0     1.0    1.0    0.0    0.0

1       1498.41      1.0     52.69  0.32   0.0    0.0    0.0    0.0    0.0    0.0    1.0    0.0    0.0
        0.0     0.0  0.0     0.0    0.0    0.0    0.0

2       1079.19      1.0     48.74  0.39   0.0    0.0    0.0    0.0    0.0    0.0    1.0    0.0    0.0
        0.0     0.0  0.0     1.0    0.0    0.0    0.0

3       1342.77      1.0     49.81  0.44   0.0    0.0    0.0    0.0    0.0    1.0    0.0    0.0    0.0
        0.0     0.0  0.0     0.0    0.0    0.0    0.0

4       1498.41      0.0     55.83  0.46   0.0    0.0    0.0    0.0    0.0    1.0    0.0    0.0    0.0
        1.0     0.0  0.0     0.0    0.0    0.0    1.0
```

Table 6: Feature Engineering- Changing data types

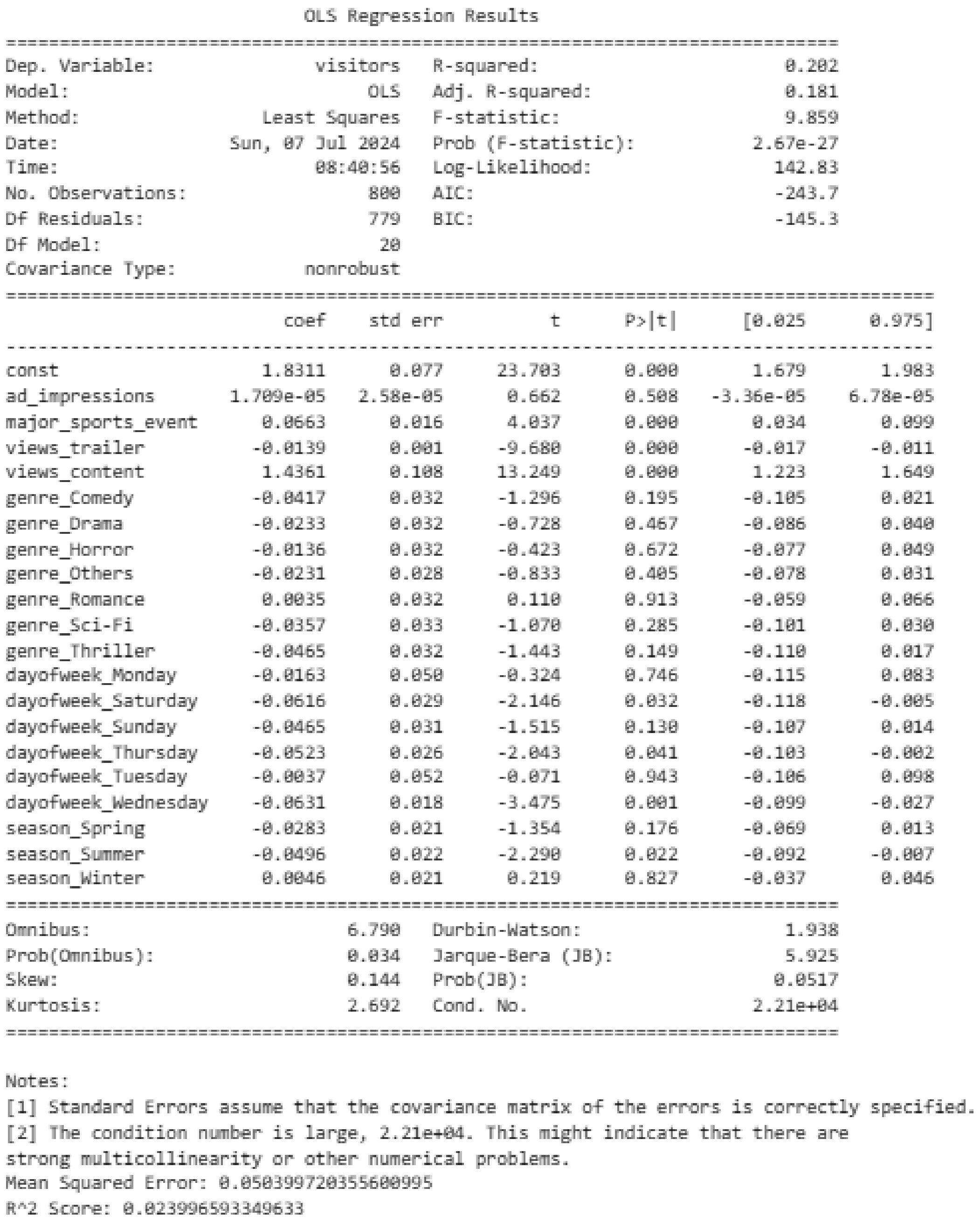
9. splitting the data in 70:30 ratio for train to test data



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

Table 7: Ratio change to train and test data

10. Regression fitting model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               visitors   R-squared:                       0.202
Model:                            OLS   Adj. R-squared:                  0.181
Method:                 Least Squares   F-statistic:                     9.859
Date:                Sun, 07 Jul 2024   Prob (F-statistic):           2.67e-27
Time:                        08:40:56   Log-Likelihood:                 142.83
No. Observations:                 800   AIC:                            -243.7
Df Residuals:                     779   BIC:                            -145.3
Df Model:                          20
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      1.8311      0.077     23.703      0.000       1.679       1.983
ad_impressions          1.709e-05   2.58e-05      0.662      0.508   -3.36e-05    6.78e-05
major_sports_event         0.0663      0.016      4.037      0.000       0.034       0.099
views_trailer             -0.0139      0.001     -9.680      0.000      -0.017      -0.011
views_content              1.4361      0.108     13.249      0.000       1.223       1.649
genre_Comedy              -0.0417      0.032     -1.296      0.195      -0.105       0.021
genre_Drama               -0.0233      0.032     -0.728      0.467      -0.086       0.040
genre_Horror              -0.0136      0.032     -0.423      0.672      -0.077       0.049
genre_Others              -0.0231      0.028     -0.833      0.405      -0.078       0.031
genre_Romance              0.0035      0.032      0.110      0.913      -0.059       0.066
genre_Sci-Fi              -0.0357      0.033     -1.070      0.285      -0.101       0.030
genre_Thriller            -0.0465      0.032     -1.443      0.149      -0.110       0.017
dayofweek_Monday          -0.0163      0.050     -0.324      0.746      -0.115       0.083
dayofweek_Saturday        -0.0616      0.029     -2.146      0.032      -0.118      -0.005
dayofweek_Sunday          -0.0465      0.031     -1.515      0.130      -0.107       0.014
dayofweek_Thursday        -0.0523      0.026     -2.043      0.041      -0.103      -0.002
dayofweek_Tuesday         -0.0037      0.052     -0.071      0.943      -0.106       0.098
dayofweek_Wednesday       -0.0631      0.018     -3.475      0.001      -0.099      -0.027
season_Spring             -0.0283      0.021     -1.354      0.176      -0.069       0.013
season_Summer             -0.0496      0.022     -2.290      0.022      -0.092      -0.007
season_Winter              0.0046      0.021      0.219      0.827      -0.037       0.046
==============================================================================
Omnibus:                        6.790   Durbin-Watson:                   1.938
Prob(Omnibus):                  0.034   Jarque-Bera (JB):                5.925
Skew:                           0.144   Prob(JB):                       0.0517
Kurtosis:                       2.692   Cond. No.                     2.21e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.21e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
Mean Squared Error: 0.050399720355600995
R^2 Score: 0.023996593349633
```

Table 8: Regression fit model

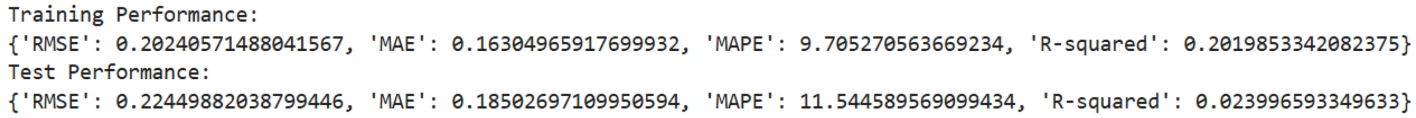
Observations:-

Both the R-squared and Adjusted R squared of our model are low. This is an indication that we have been able to create a good model that is able to explain the number of visitors (viewership) for content on the OTT platform.

The model is not an underfitting model.

To be able to make statistical inferences from our model, we will have to test that the linear regression assumptions are followed.

11. Performance Evaluation: -



```
Training Performance:
{'RMSE': 0.20240571488041567, 'MAE': 0.16304965917699932, 'MAPE': 9.705270563669234, 'R-squared': 0.2019853342082375}
Test Performance:
{'RMSE': 0.22449882038799446, 'MAE': 0.18502697109950594, 'MAPE': 11.544589569099434, 'R-squared': 0.023996593349633}
```

Table 9: Performance Evaluation

Inferences:

The R-squared value for the training set is 0.202, which indicates that approximately 20.2% of the variance in the target variable is explained by the model. This is relatively low, suggesting that the model may not be capturing much of the underlying pattern in the data.

The R-squared value for the test set is even lower at 0.024, indicating that only 2.4% of the variance in the target variable is explained by the model on unseen data. This further suggests that the model is not generalizing well to new data.

12. Testing Multicoolinearity

	feature	VIF
0	const	112.368582
1	ad_impressions	1.030484
2	major_sports_event	1.301241
3	views_trailer	1.953696
4	views_content	2.312916
5	genre_Comedy	1.907253
6	genre_Drama	1.917251
7	genre_Horror	1.900879
8	genre_Others	2.559604
9	genre_Romance	1.755691
10	genre_Sci-Fi	1.860019
11	genre_Thriller	1.914637
12	dayofweek_Monday	1.068045
13	dayofweek_Saturday	1.218036
14	dayofweek_Sunday	1.176927
15	dayofweek_Thursday	1.175783
16	dayofweek_Tuesday	1.072332
17	dayofweek_Wednesday	1.414615
18	season_Spring	1.566477

19	season_Summer	1.636245
20	season_Winter	1.616897

Table 10: Testing Multicollinearity

Observation:

As there is no multicollinearity, we can look at the p-values of predictor variables to check their significance.

13. Linearity of variables test.

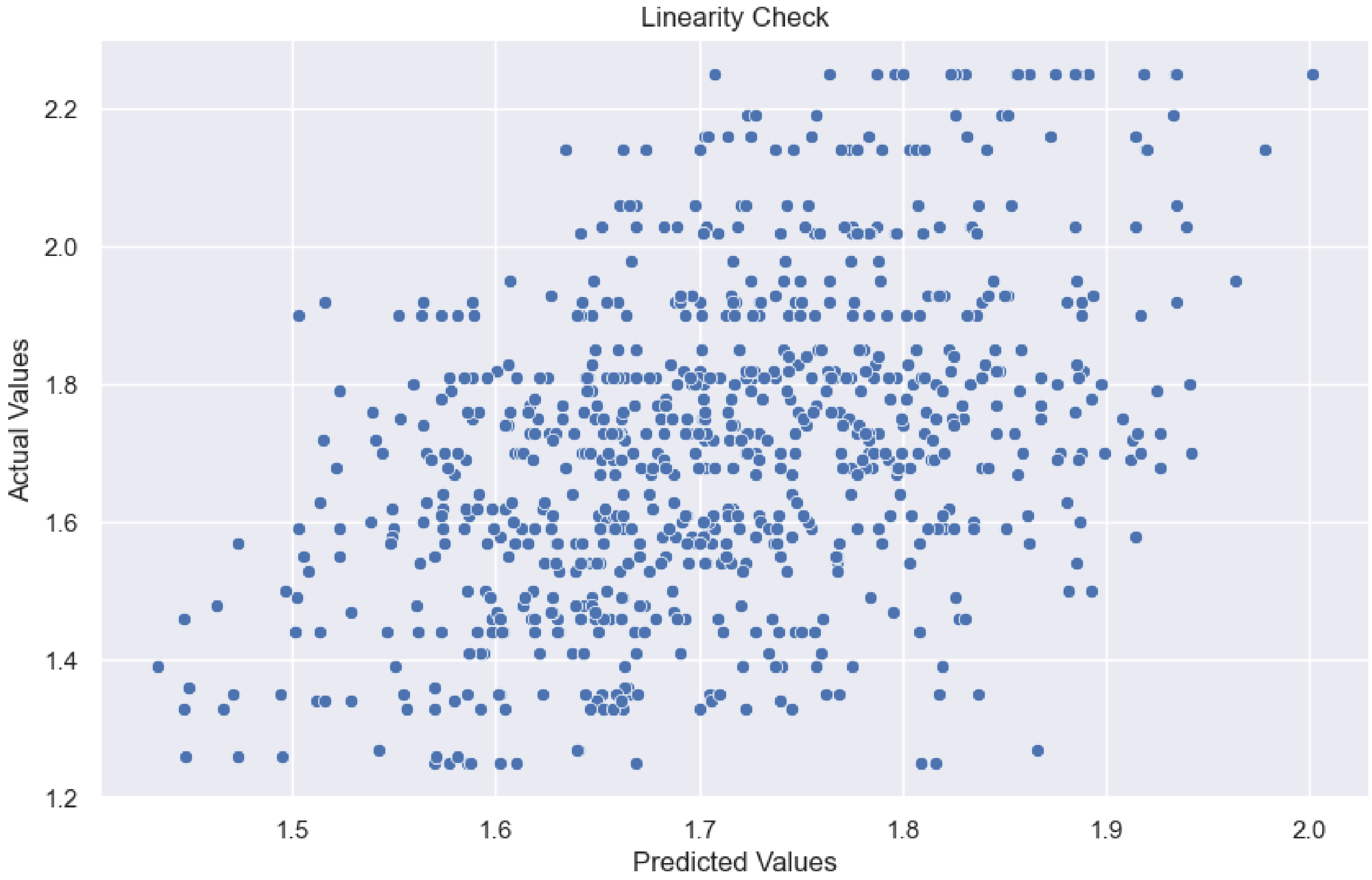
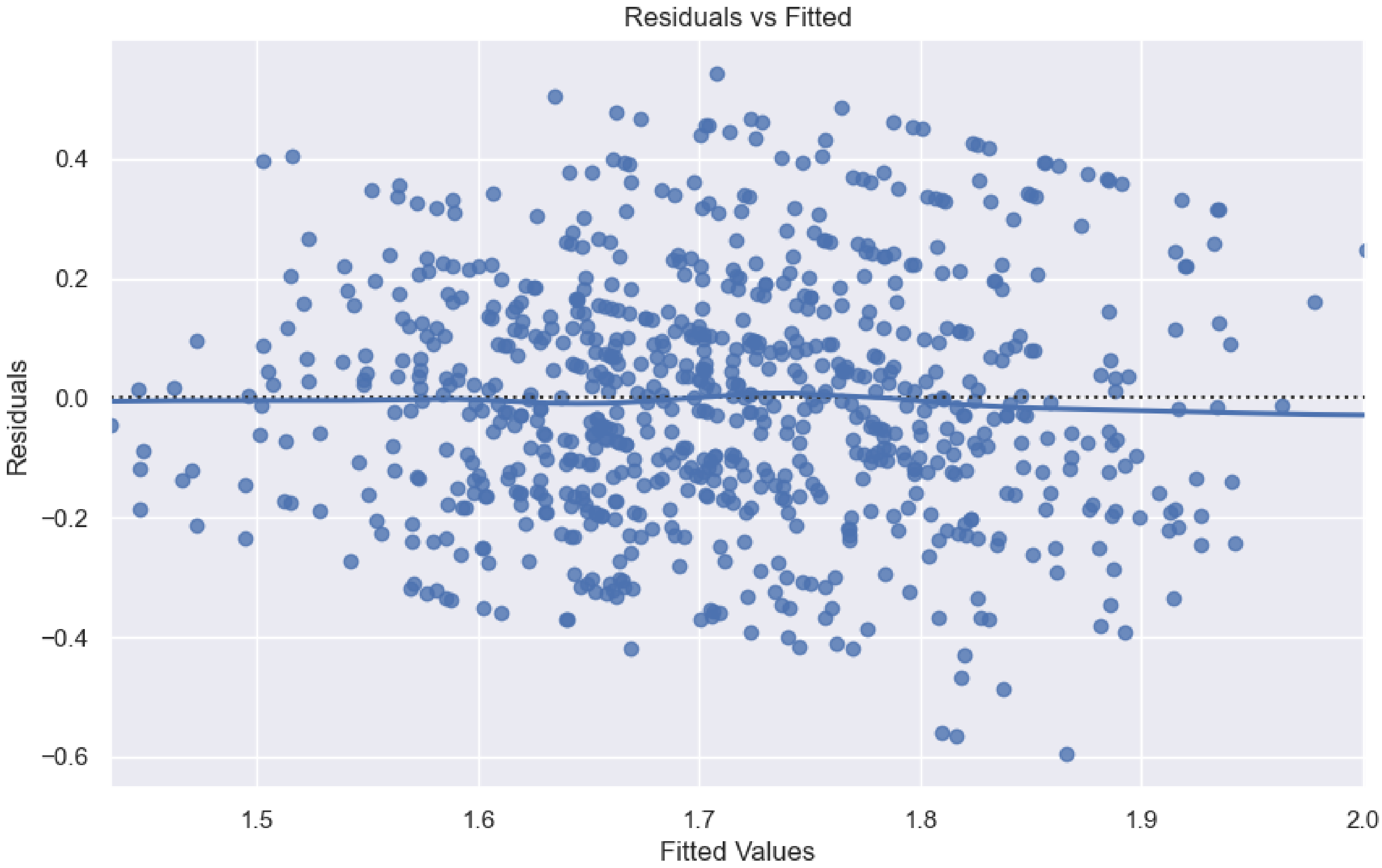


Figure-11 : Scatter plot of predicted and actual values

Observation:

The residuals are very largely scattered



.

Figure-12 : Scatter plot of residual & fitted values

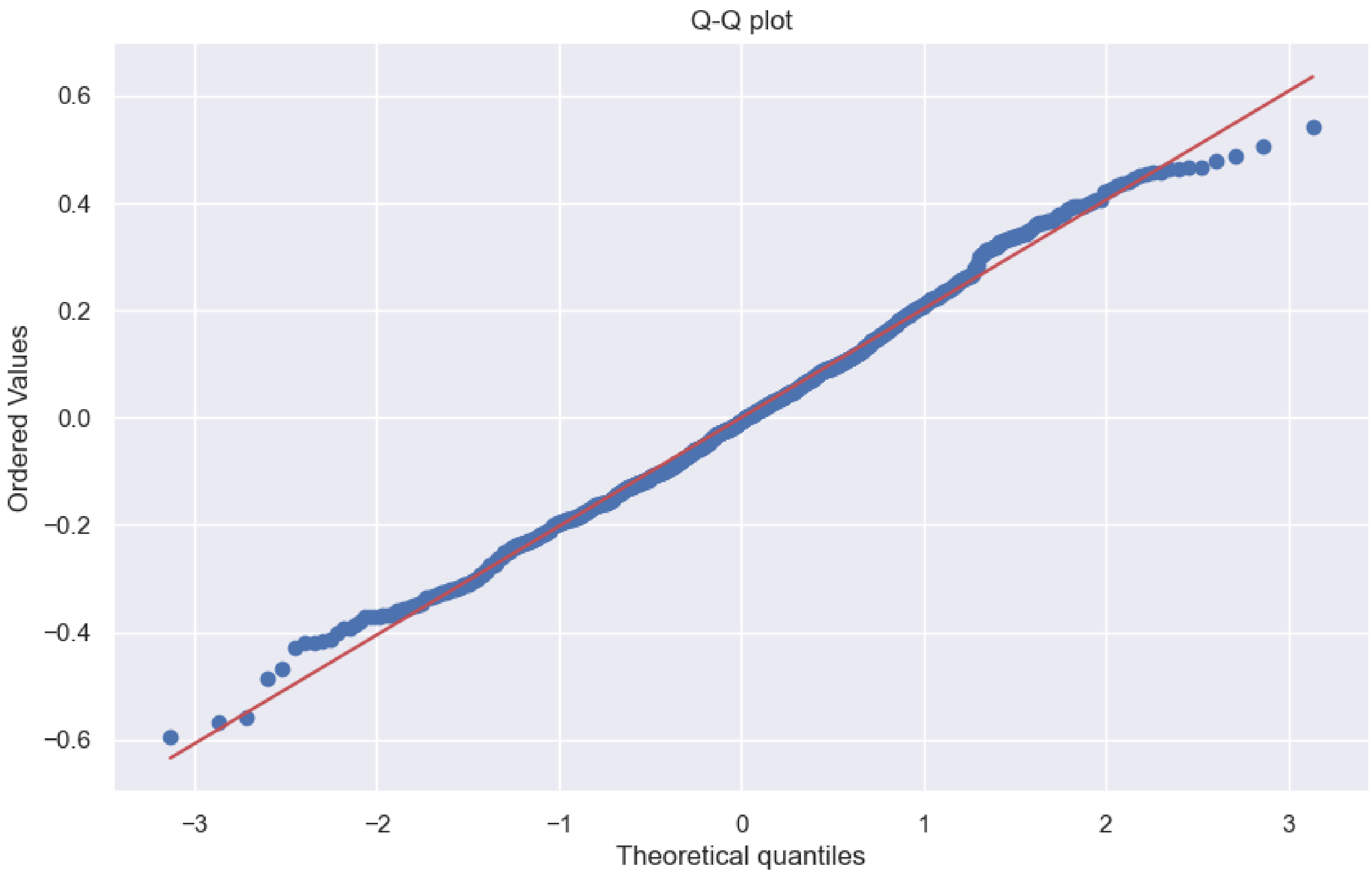
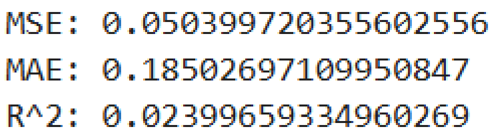


Figure-13 : Q-Q plot

Observation:

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

14. Model Performance Evaluation



```
MSE: 0.050399720355602556
MAE: 0.18502697109950847
R^2: 0.02399659334960269
```

Table 11: Model performance Evaluation

Observatioins:

Mean Squared Error (MSE):

MSE is 0.0504. MSE measures the average of the squares of the errors—that is, the average squared difference between the actual and predicted values. A lower MSE indicates better model performance.

This MSE value is consistent with the earlier RMSE value (since RMSE is the square root of MSE). Both metrics indicate the model's prediction errors are relatively low, but not negligible.

Mean Absolute Error (MAE):

MAE is 0.1850. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The MAE value indicates that, on average, the model's predictions are off by about 0.185 units from the actual values. This is consistent with the previous MAE value observed.

R-squared (R^2):

R^2 is 0.0240. R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating better model performance. An R^2 value of 0.0240 indicates that only 2.4% of the variance in the target variable is explained by the model. This suggests that the model is not capturing much of the underlying pattern in the data, which is consistent with the earlier analysis.

Performance Evaluation

Model Accuracy: The low R^2 value indicates poor model accuracy. The model explains only a small fraction of the variance in the dependent variable. Error Magnitude: The MSE and MAE values, while relatively low, suggest that the model makes errors in its predictions. However, the low R^2 value shows that these errors do not capture much of the variability in the actual data.

15. Actionable Insights:

Significant Predictors:

Identify the significant predictors from the model summary (those with p-values less than 0.05). These predictors have a strong influence on the number of visitors. For example, if views_trailer, dayofweek, and genre are significant predictors, focus on these variables for strategic decisions.

Impact of Genre:

Certain genres might attract more visitors. If the model indicates that genres like Drama or Comedy have higher coefficients, consider producing more content in these genres.

Day of the Week:

The day of release significantly impacts viewership. If the model shows that weekends or specific days have higher coefficients, schedule releases on these days to maximize viewership.

Trailer Views:

There is a positive correlation between trailer views and content views. This suggests that effective marketing and promotion of trailers can significantly boost viewership.

Seasonal Trends:

As the model shows certain seasons (e.g., winter Mondays & Tuesdays) have higher viewership, plan major releases or marketing campaigns around these times.

Business Recommendations:-

Content Production and Acquisition:

Focus on producing and acquiring content in genres that attract higher viewership. Utilize the insights from the genre coefficients to guide content strategy. Regularly update the genre preferences by periodically retraining the model with new data to adapt to changing viewer tastes.

Marketing and Promotion:

Invest in promoting trailers, especially for content with high predicted viewership. Utilize social media and targeted advertising to maximize trailer views. Leverage the positive correlation between trailer views and content views by running teaser campaigns and releasing trailers well in advance of the content release.

Scheduling Strategy:

Schedule content releases on days with higher predicted viewership. Use insights from the day-of-week coefficients to optimize the release calendar. Consider special releases and promotions on high-viewership days to attract more viewers.

Seasonal Campaigns:

Plan major content releases during high-viewership seasons. For instance, launch major series or movies during holidays or other peak seasons identified by the model. Run seasonal marketing campaigns to capitalize on the increased viewership during these periods.

Data-Driven Decision Making:

Continuously monitor the model's predictions and performance. Update the model with new data to ensure it remains accurate and reflective of current trends. Use the model's insights to inform decisions across various departments, including content strategy, marketing, and scheduling.

User Engagement and Retention:

Utilize insights from the model to improve user engagement and retention. For example, recommend popular genres or content to users based on their viewing history. Implement personalized content recommendations using significant predictors to enhance user experience and keep viewers engaged.
By implementing these recommendations, the OTT platform can optimize its content strategy, marketing efforts, and scheduling to maximize viewership and improve overall business.

END