Predictive Modelling Project on OTT Media Service Provider Data

Apr B Sunday Onkar 10:30 AM Batch

Arindam Saha

Table of Contents

List of Figures

List of Tables

1. Background

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

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

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

2. Objective

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

Variables	Description
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

Table 1:Variables and Descriptions

3. Data Information

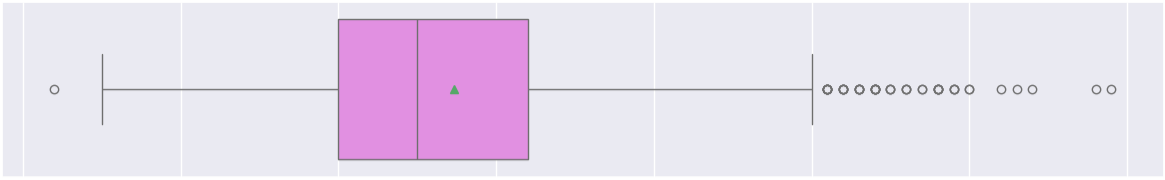
Note: -

- There are in total 1000 rows and 8 columns present.
- There are 5 numeric (float and int type) and 3 string (object type) columns in the data.
- The target variable is the views_content (first day views), which is a float type.
- In total 4 seasons observed.
- Views trailer & ad impressions seem to have outliers.
- The mean views content has mean of 0.47 with standard deviation of 0.1
- Mean is almost similar or very low differnce to median for views content.
- There are no duplicates.
- There are no missing values in all the columns.
- For the column major_sports_event**,** we have replaced the 1 values with '***yes***' and 0 values with '***no***'.

4. Exploratory Data Analysis (EDA)

4.1. Univariate Analysis

4.1.1. Univariate content views



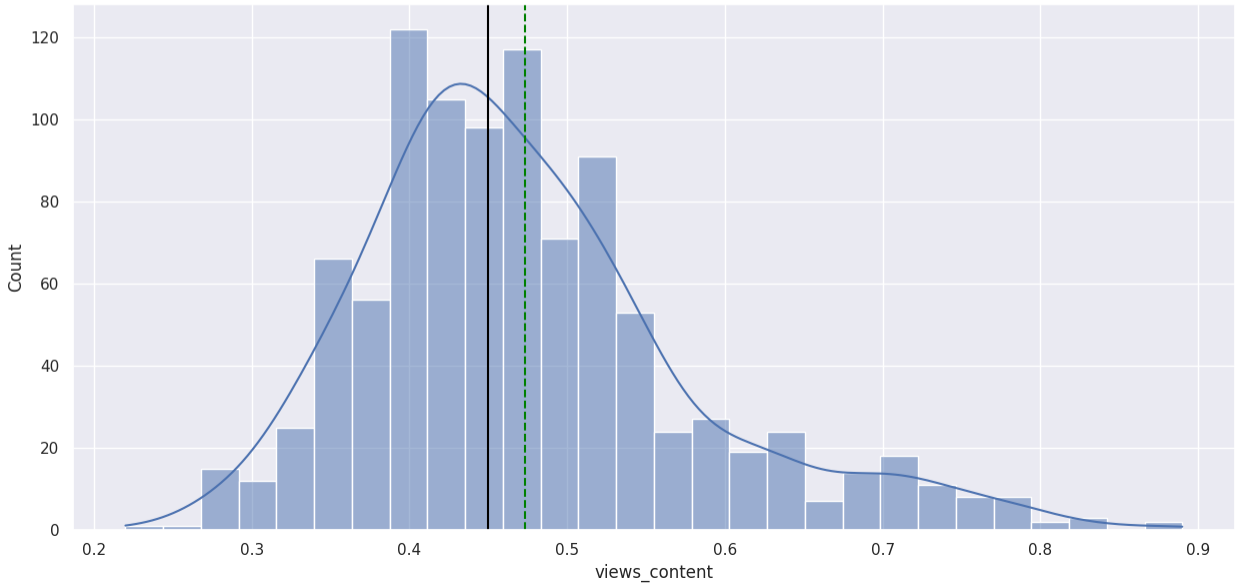


Fig 1: Univariate Content View Analysis

- Few outliers can be observed in the views content field and distribution is right skewed. However, it seems almost like a normal distribution.
- Avg less than 0.5 million watched the actual content.

4.1.2. Univariate Trailer Views

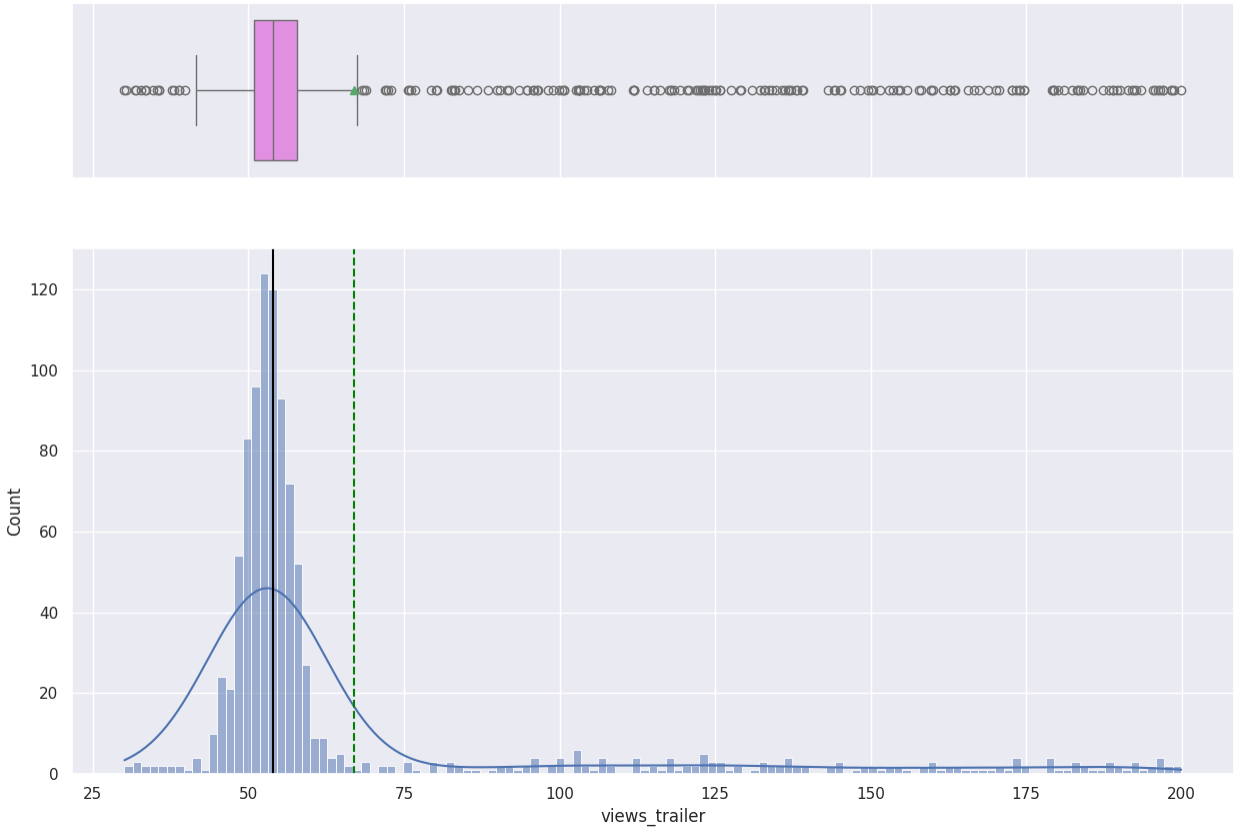


Fig 2: Univariate Trailer View Analysis

- Heavily right-skewed data. But distribution is normal.
- Too many outliers can be observed at both the tails.
- Mean less than 70 million viewed the trailer. Mean and median differ in their location.

4.1.3. Univariate ad_impression

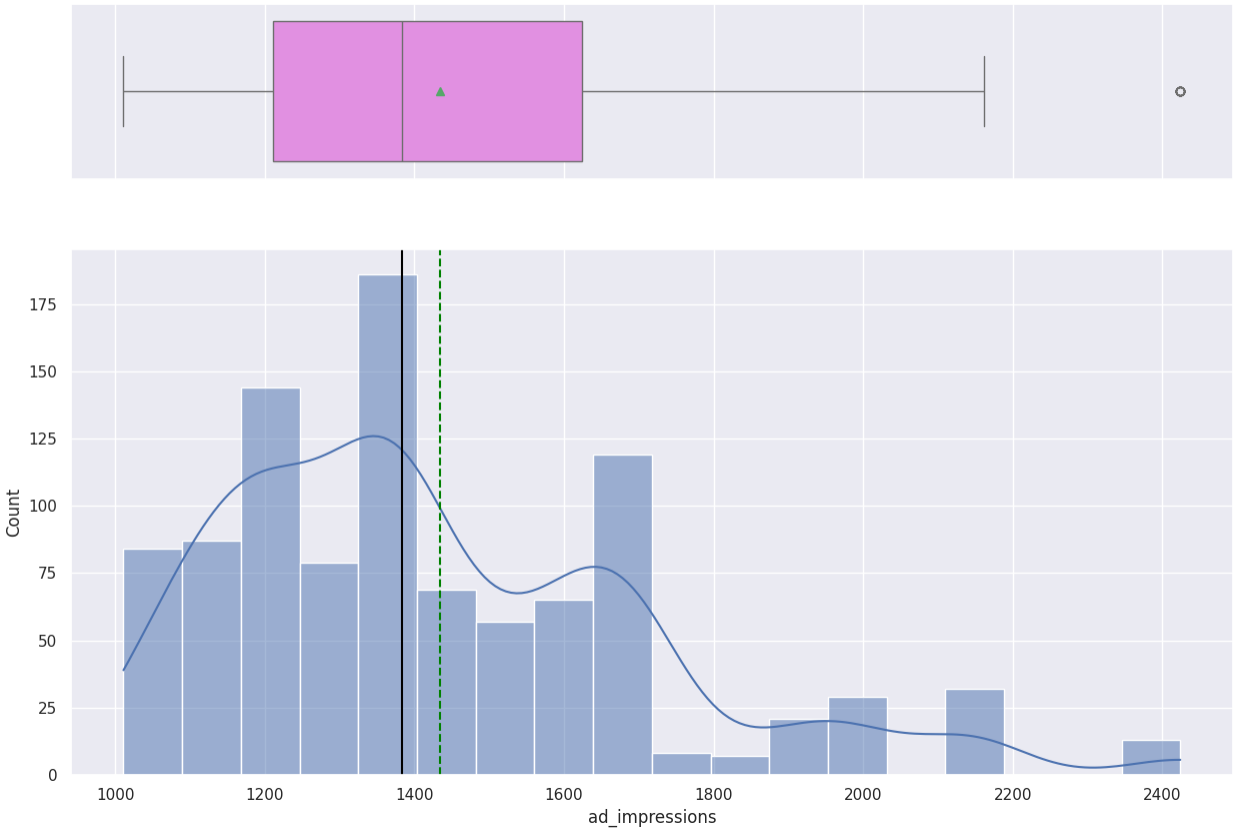


Fig 3: Univariate Ad Impression Analysis

- One outlier can be observed in the ad impression.
- Distribution seems to be right skewed and it's not following a normal distribution.
- Mean ad impression is less than 1400.

4.1.4. Univariate Visitors

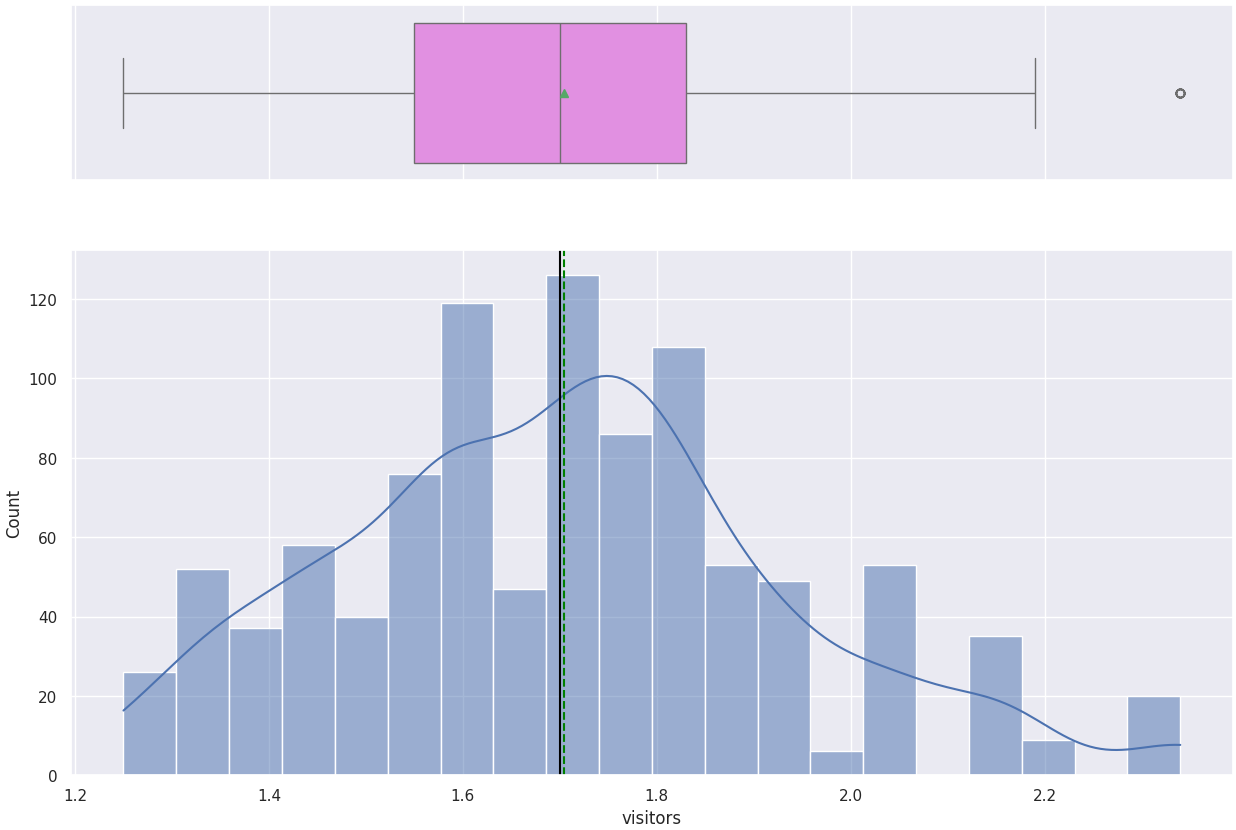


Fig 4:Univariate Visitor Analysis

- Almost a normal distribution with only single outlier can be observed.
- Mean and median follows almost similar at 1.7 million visitors for previous week.

4.1.5. Univariate Genre

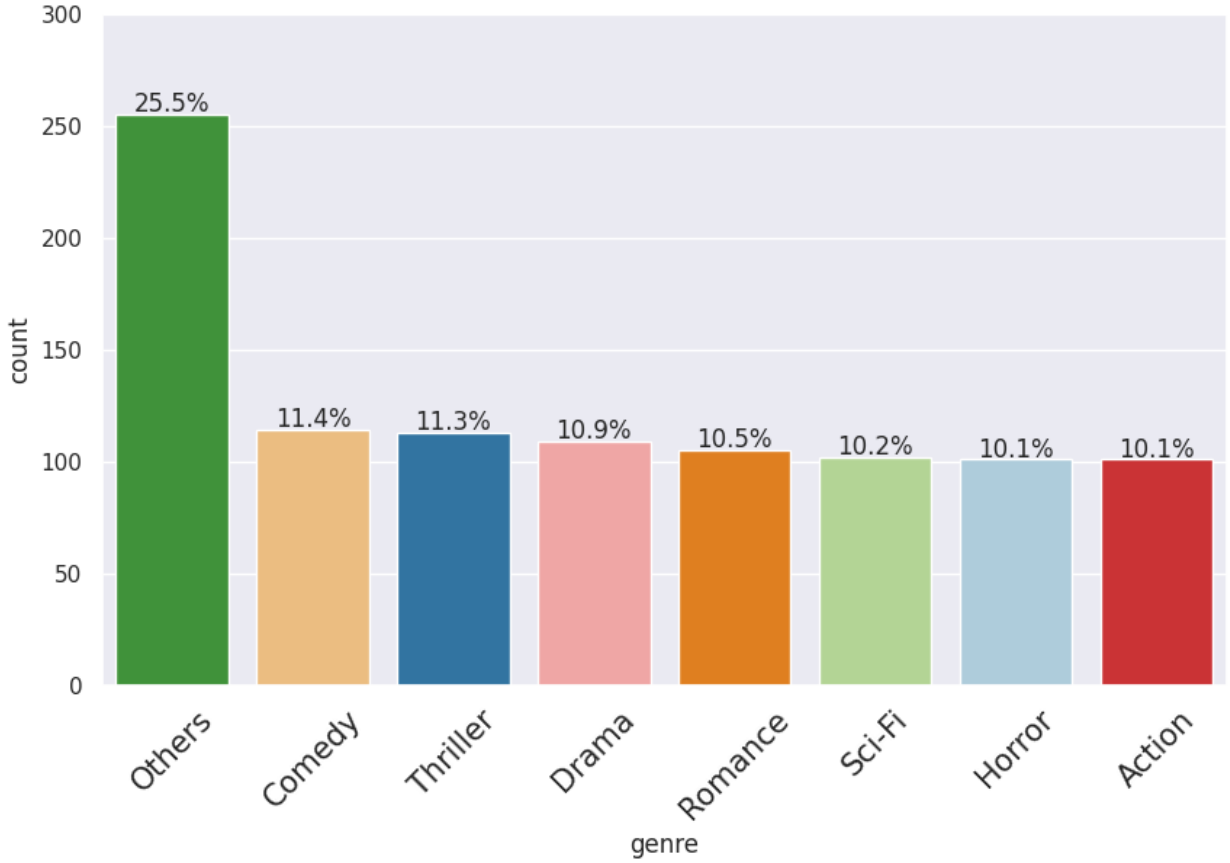


Fig 5:Univariate Genre Counts

- Genre others is almost one fourth of total genre.

4.1.6. Univariate Season

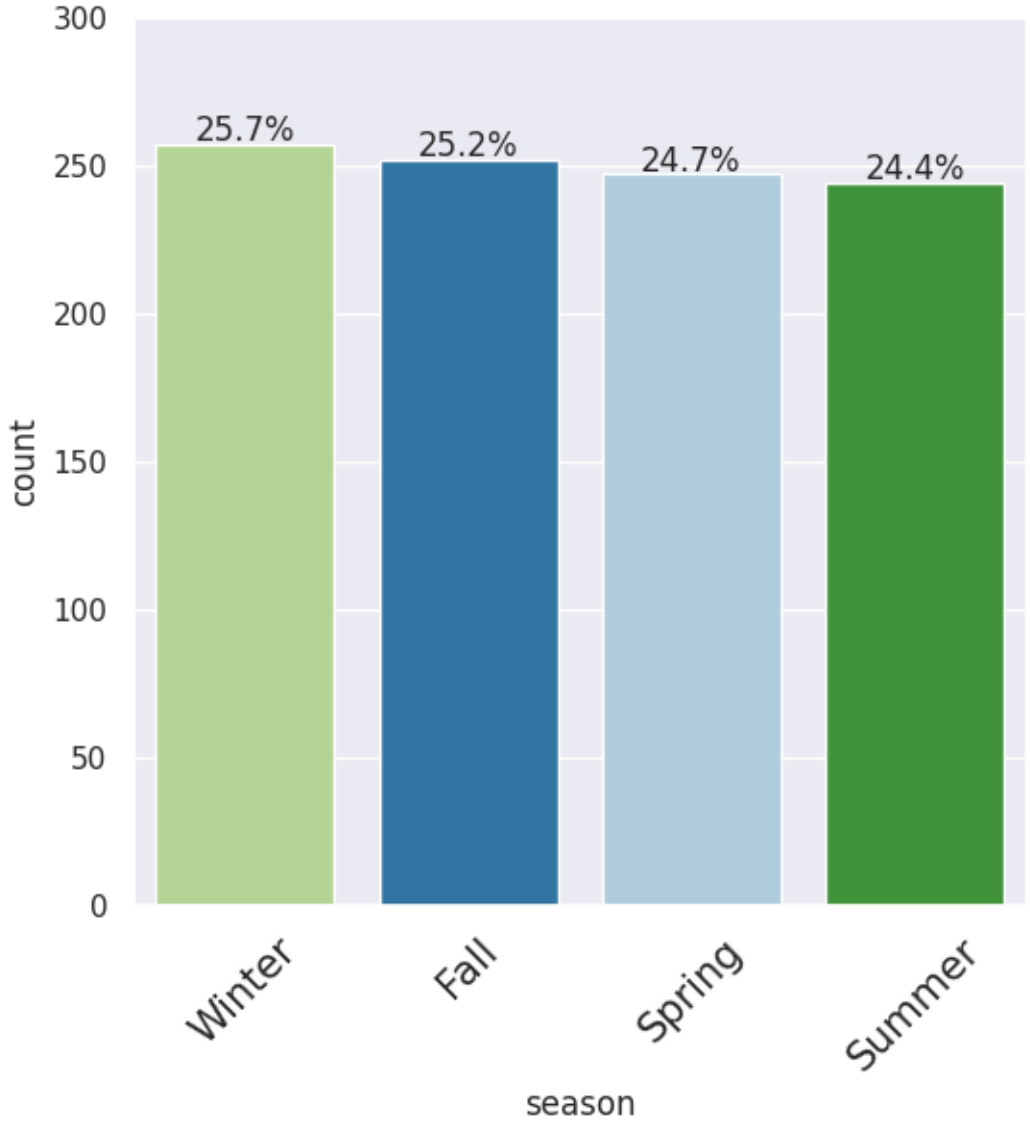


Fig 6:Univariate Seasons

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

4.1.7. Univariate day of the week

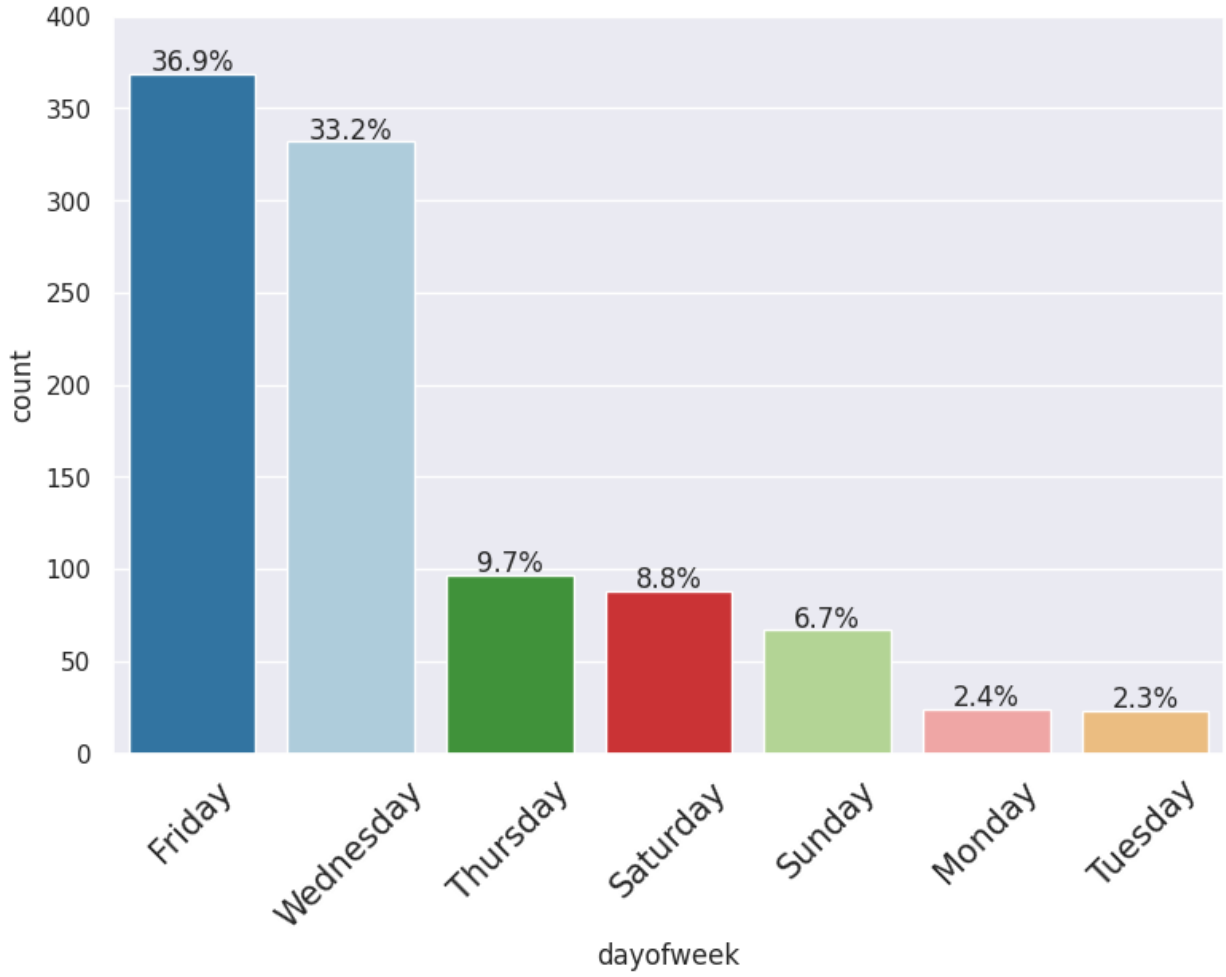


Fig 7: Univariate day of content release

- Most of the content seems to be released on Friday followed by Wednesday.
- Very few contents get released on Wednesday.

4.2. Bivariate Analysis

4.2.1. Pair plot and Heat map

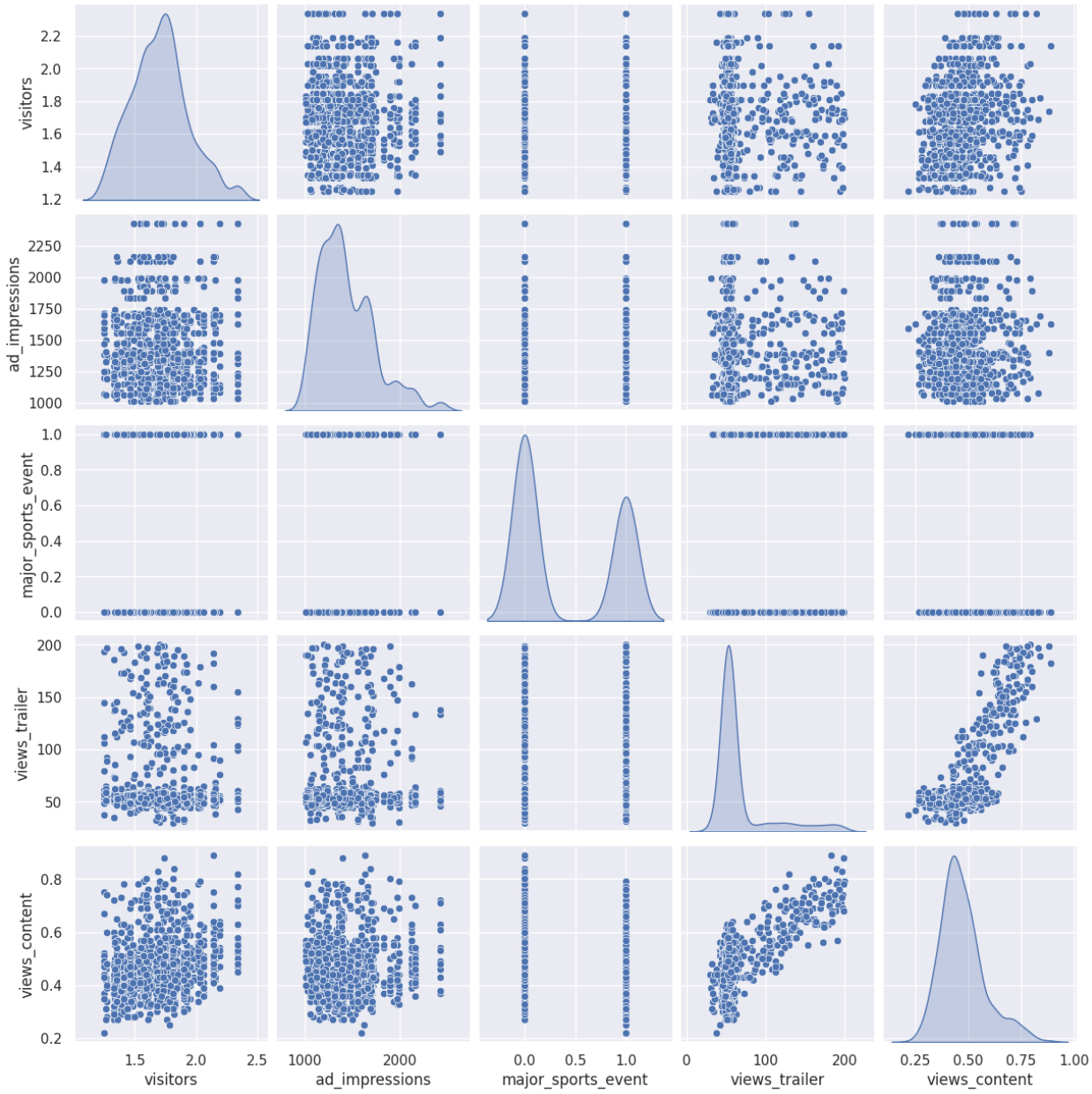


Fig 8: Pair Plot

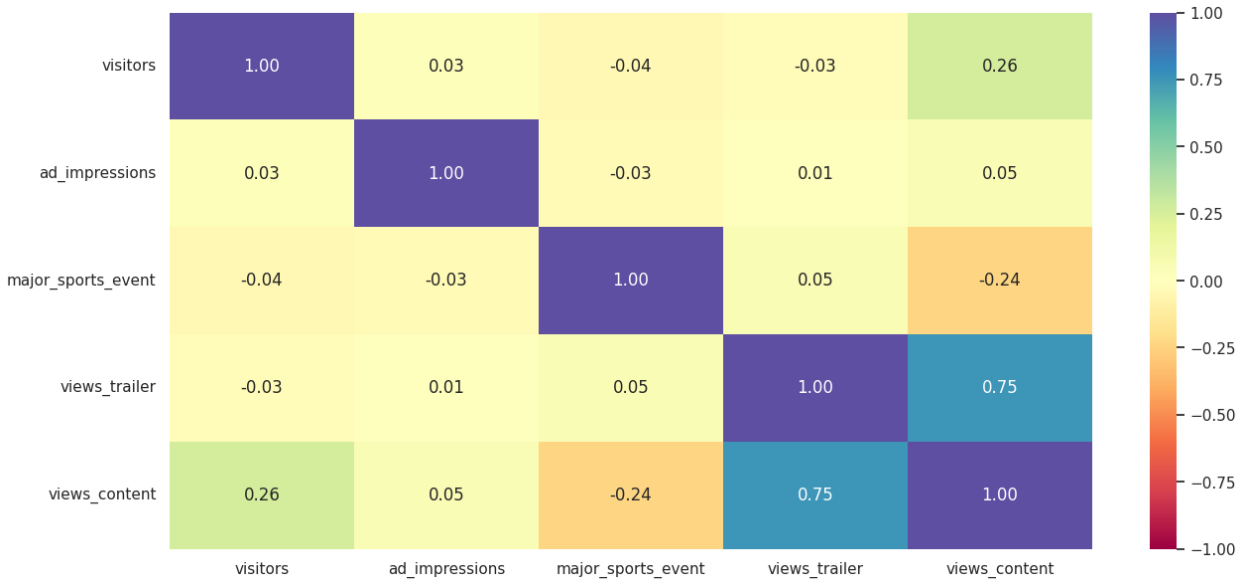


Fig 9: Heat Map

- Positive correlation can be observed between views content and views trailer.
- A little negative correlation can be observed between major sprots event and content views.
- Positive effect of visitor towards the content view can be observed.

4.2.2. Bivariate genre and content views

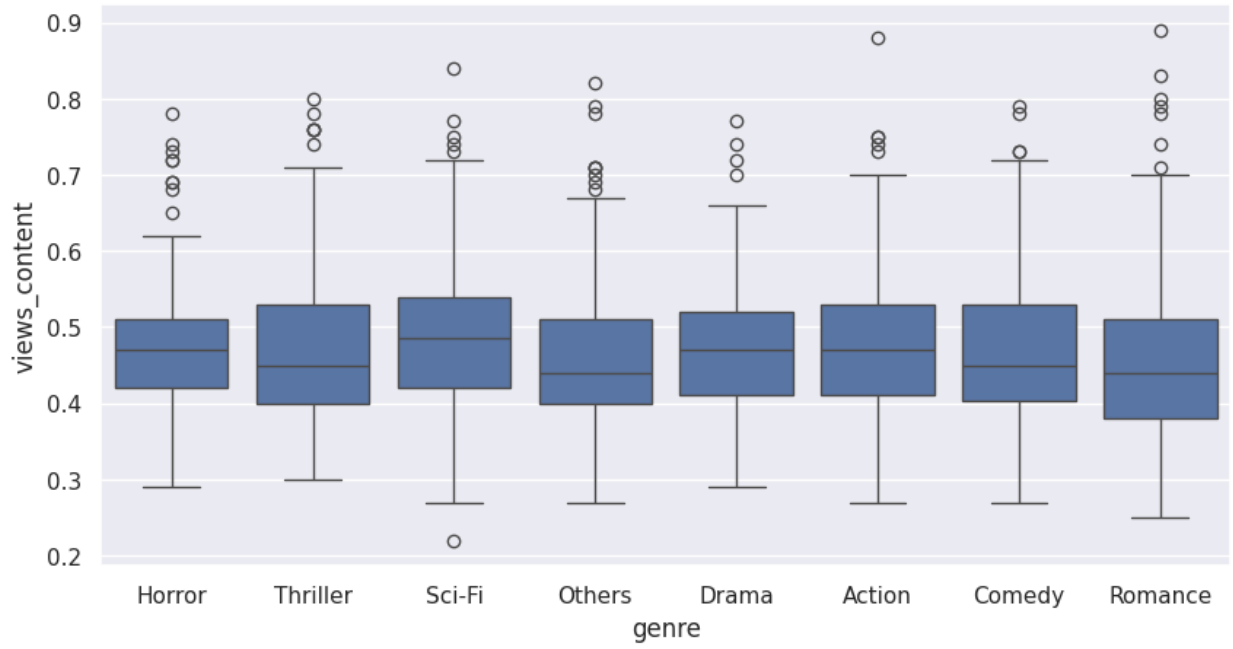


Fig 10: Genre vs Content Views

- With respect to different genre mean observed views of content looks in between 0.43 to 0.5 million.
- However, there seems to be presence of outliers for all the genre types.

4.2.3. Bivariate genre and trailer views

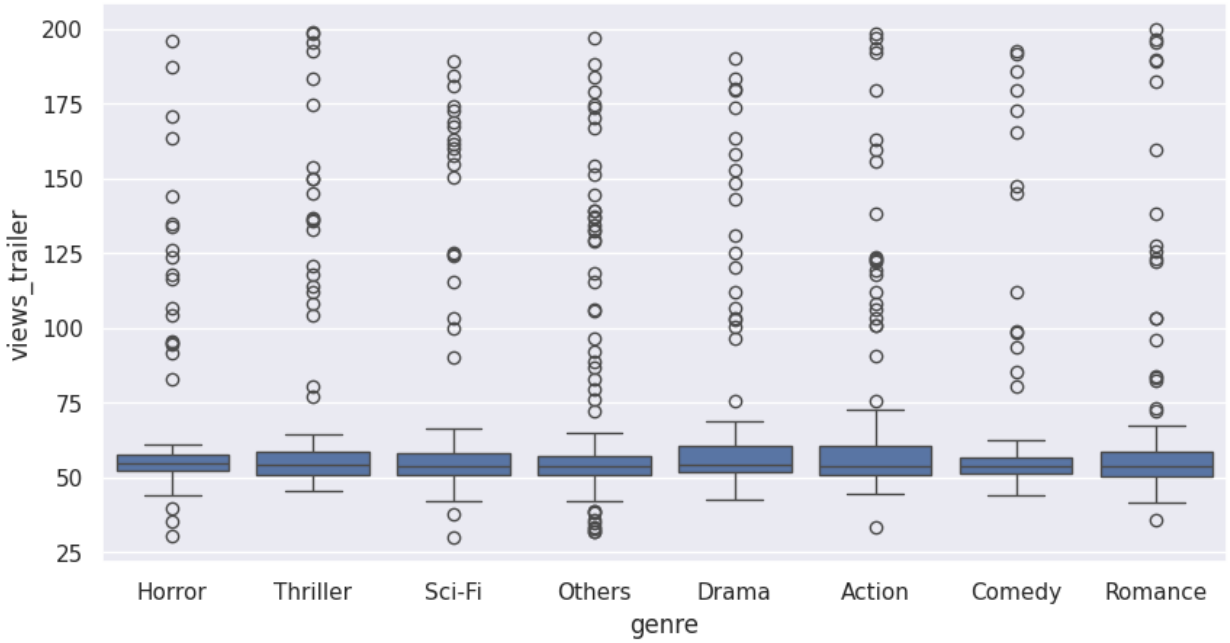


Fig 11: Genre vs Trailer Views 1

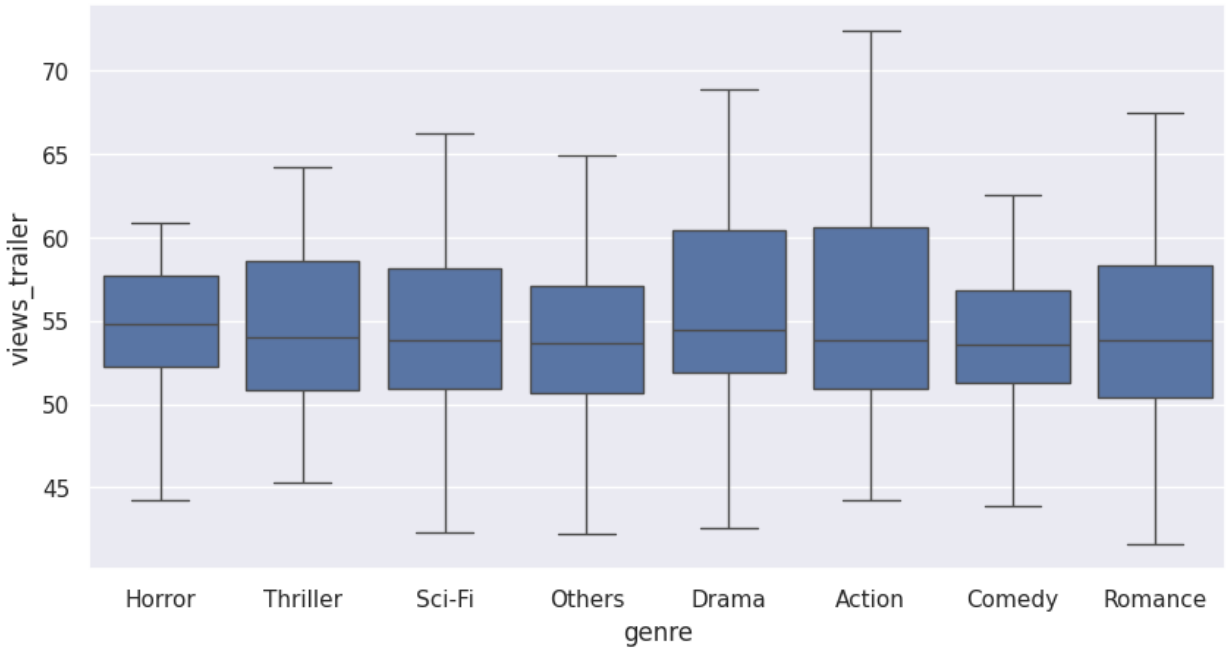


Fig 12: Genre vs Trailer Views 2

- Outliers can be observed at the right.
- Seems to have almost mean of 53 - 55 million views in trailer for all the genre.

4.2.4. Bivariate genre and ad impression

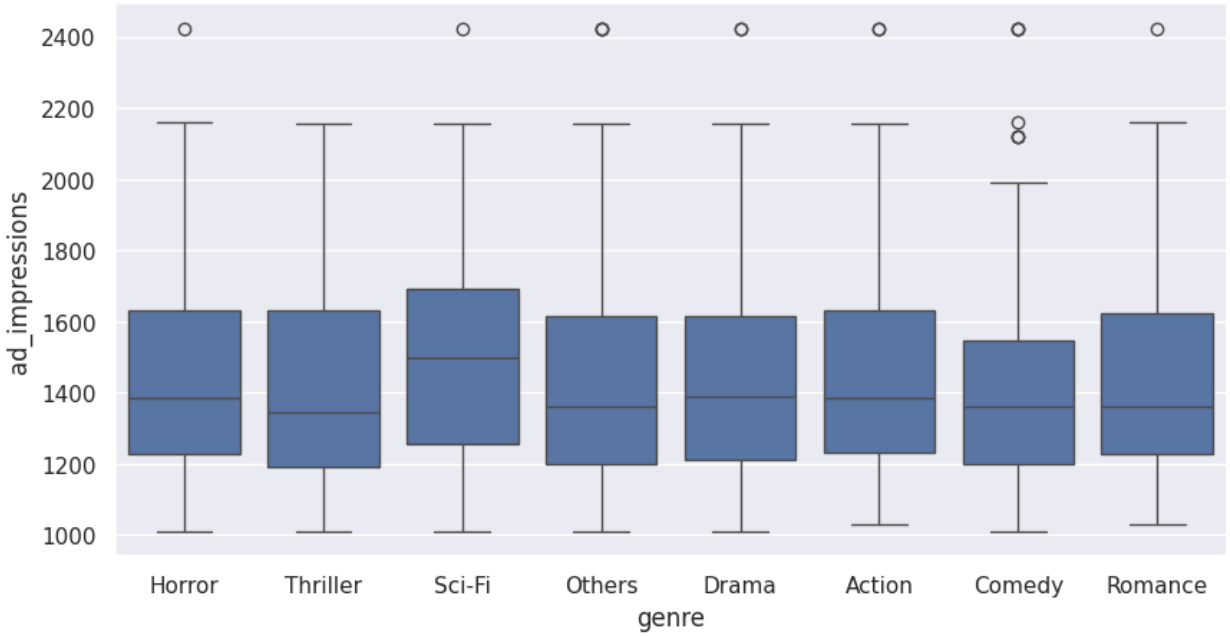


Fig 13: Genre vs Ad Impressions

- Mean ad impression is little bit high for the Sci-Fi than other genres.

4.2.5. Bivariate visitors and day of week

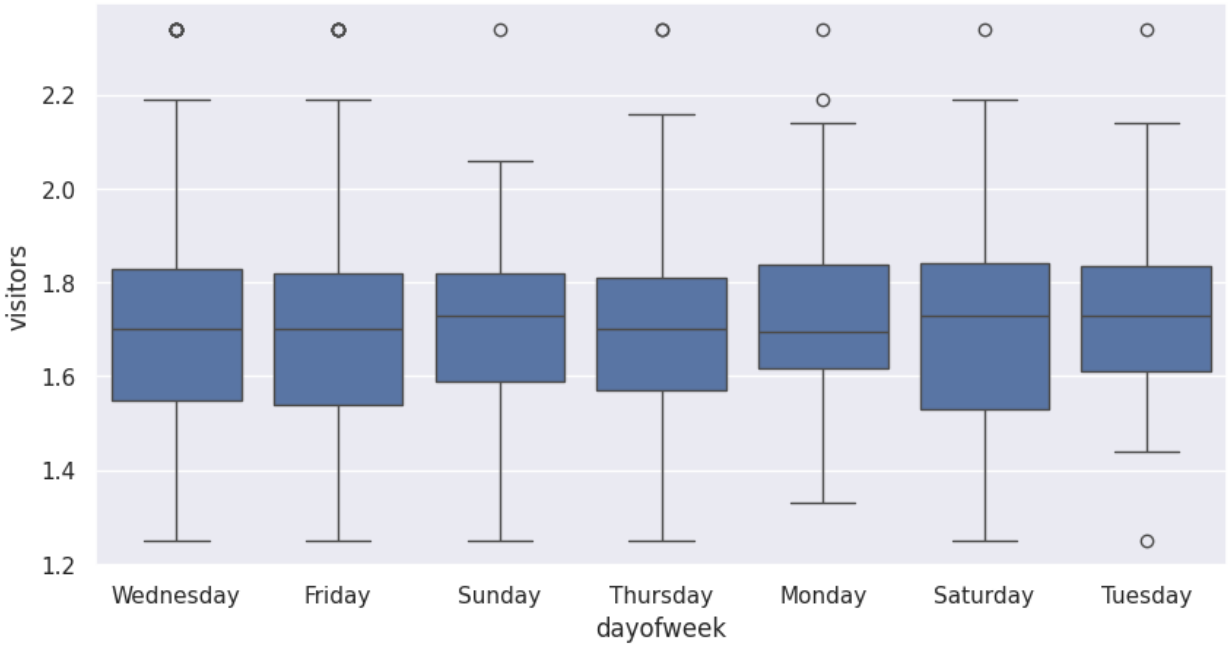


Fig 14: Visitors vs Day of Content release

- In all the days of release average visitors seems to be in between 1.675 to 1.730 million.
- Outliers can be observed.

4.2.6. Bivariate day of week vs view content and trailer

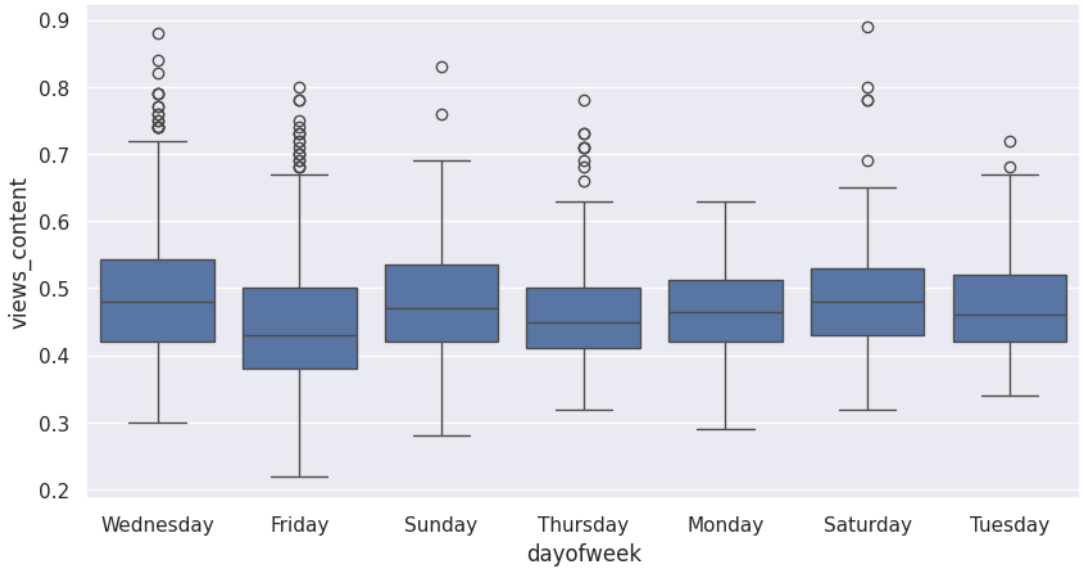


Fig 15: Content Views vs day of release

- The mean content views decrease for Friday content release with respect to other days.
- Slightly more content views can be observed on Wednesday may be due to maximum content release day.
- There are outliers present for all the days of content release at right tail (positive skewed).

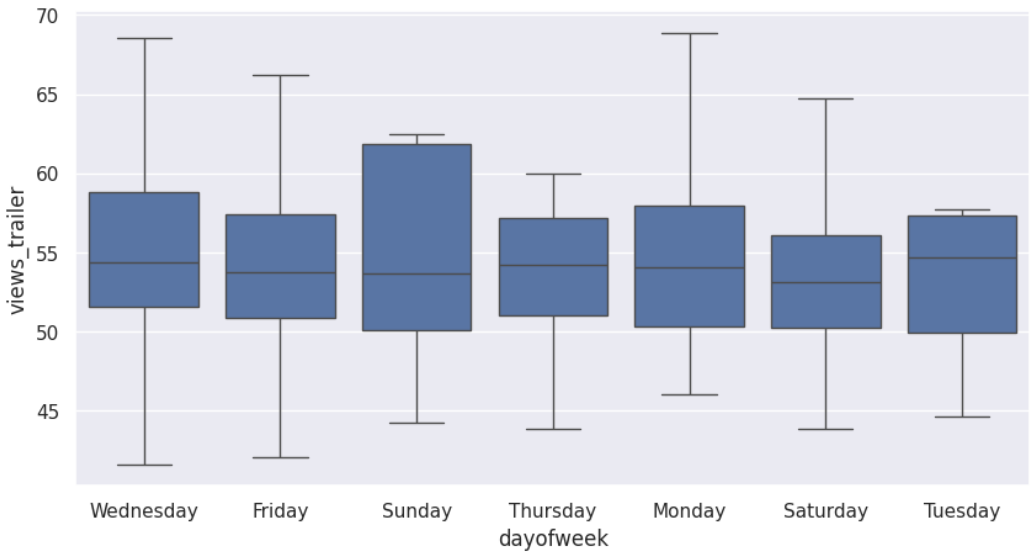


Fig 16: Trailer Views vs day of content release

- The mean trailer views are almost similar (53 to 55 million) for all the day of the release of the content.

- However, for Sunday third quartile shows higher than others, nearly around at 62 - 63 million.
- In all the days of content release outliers can be observed.

4.2.7. Bivariate season vs content and trailer views

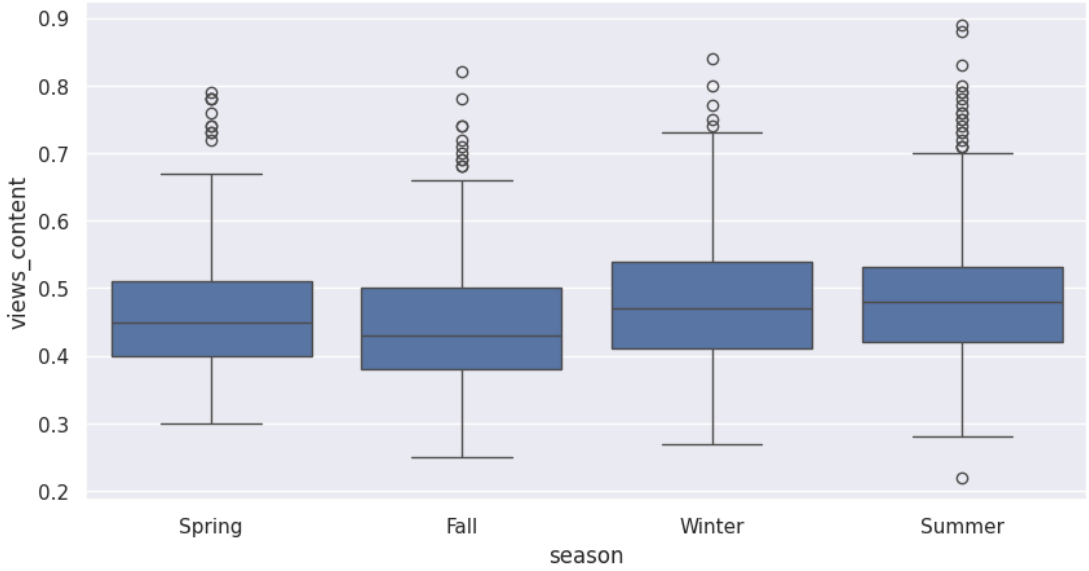


Fig 17: Season vs content views

- 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 million.
- There are outliers present for all the season at the right tail.

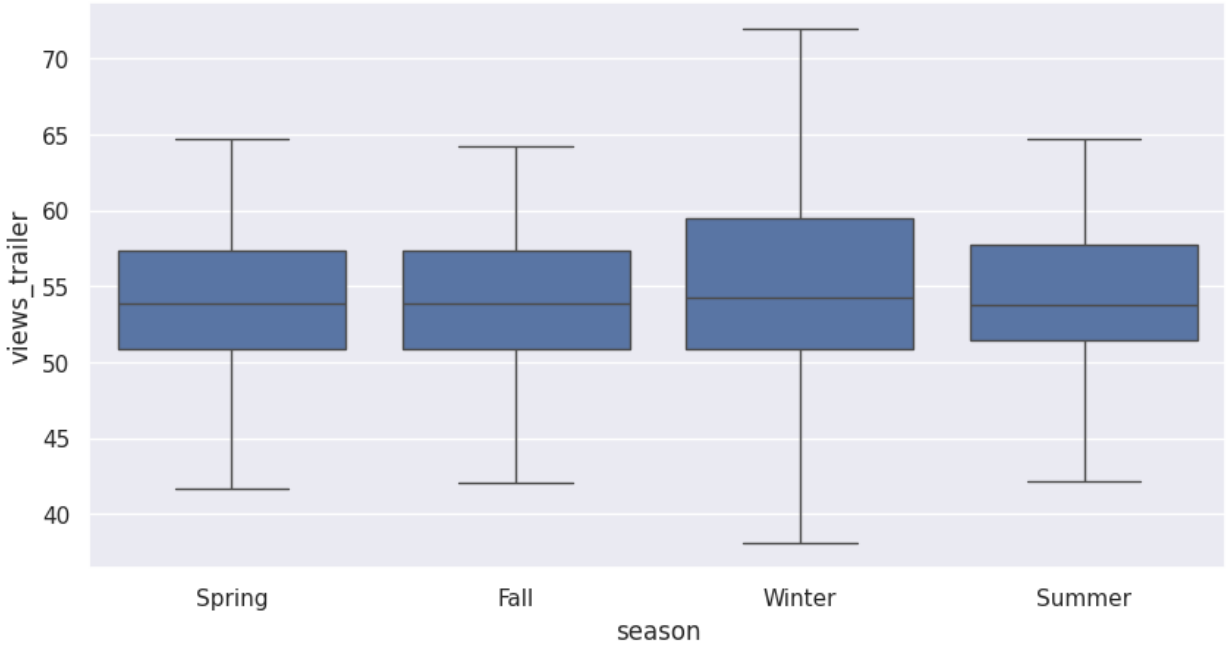


Fig 18: Season vs Trailer Views

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

4.2.8. Bivariate Major sports event and content views

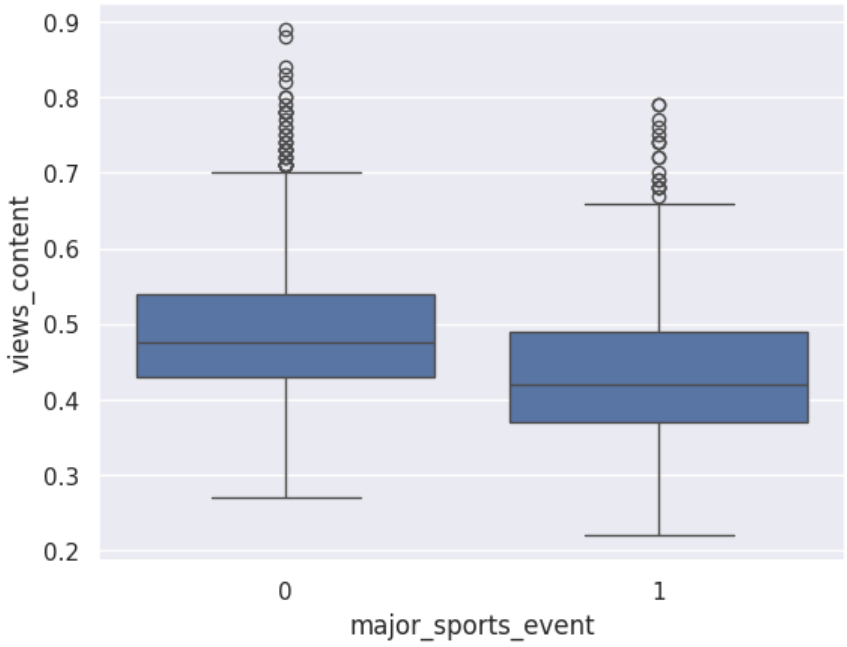


Fig 19: Major sports events vs content views

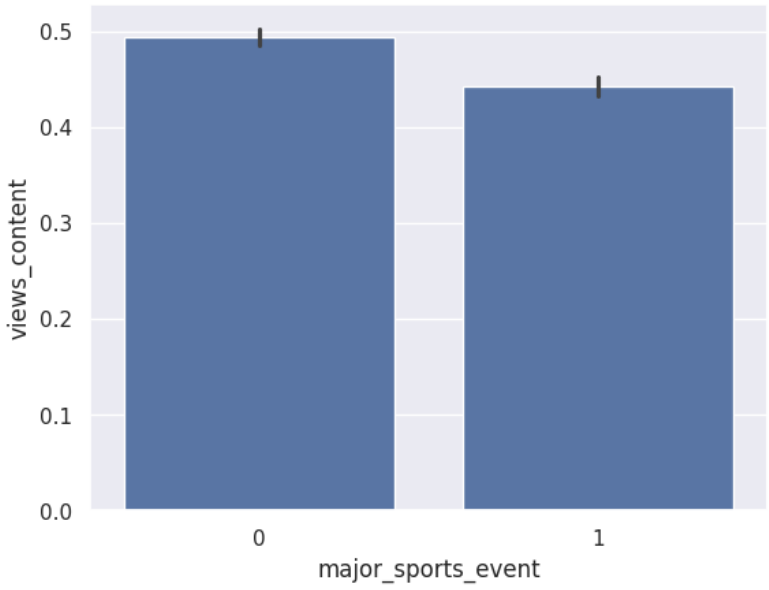


Fig 20: Major sports event vs Trailer Views

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

4.2.9. Bivariate content views and trailer views

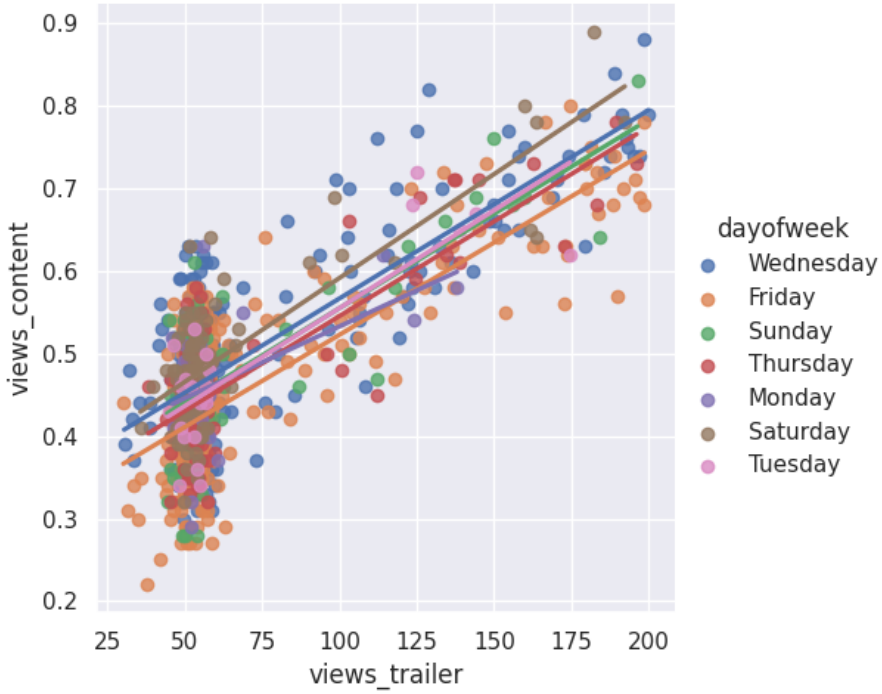


Fig 21: Content vs Trailer Views w.r.t. day of content release

- Content viewership is positively correlated with trailer viewership. On Saturdays, there is a higher number of content viewers compared to other weekdays.
- Interestingly, Sundays show relatively low viewers, which could be further analyzed by examining the data for different hours of the day.
- Also, the second day is Wednesday where most of the viewer of both content and trailer can be observed due to maximum content release day observed earlier.

5. Data Preprocessing

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

6. Model Building – Linear Regression

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.785
Method:	Least Squares	F-statistic:	129.0
Date:	Sat, 29 Jun 2024	Prob (F-statistic):	1.32e-215
Time:	07:03:27	Log-Likelihood:	1124.6
No. Observations:	700	AIC:	-2207.
Df Residuals:	679	BIC:	-2112.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0602	0.019	3.235	0.001	0.024	0.097
visitors	0.1295	0.008	16.398	0.000	0.114	0.145
ad_impressions	3.623e-06	6.58e-06	0.551	0.582	-9.3e-06	1.65e-05
views_trailer	0.0023	5.52e-05	42.193	0.000	0.002	0.002
major_sports_event_yes	-0.0603	0.004	-15.284	0.000	-0.068	-0.053
genre_Comedy	0.0094	0.008	1.172	0.241	-0.006	0.025
genre_Drama	0.0126	0.008	1.554	0.121	-0.003	0.029
genre_Horror	0.0099	0.008	1.207	0.228	-0.006	0.026
genre_Others	0.0063	0.007	0.897	0.370	-0.008	0.020
genre_Romance	0.0006	0.008	0.065	0.948	-0.016	0.017
genre_Sci-Fi	0.0131	0.008	1.599	0.110	-0.003	0.029
genre_Thriller	0.0087	0.008	1.079	0.281	-0.007	0.025
dayofweek_Monday	0.0337	0.012	2.848	0.005	0.010	0.057
dayofweek_Saturday	0.0579	0.007	8.094	0.000	0.044	0.072
dayofweek_Sunday	0.0363	0.008	4.639	0.000	0.021	0.052
dayofweek_Thursday	0.0173	0.007	2.558	0.011	0.004	0.031
dayofweek_Tuesday	0.0228	0.014	1.665	0.096	-0.004	0.050
dayofweek_Wednesday	0.0474	0.004	10.549	0.000	0.039	0.056
season_Spring	0.0226	0.005	4.224	0.000	0.012	0.033
season_Summer	0.0442	0.005	8.111	0.000	0.034	0.055
season_Winter	0.0272	0.005	5.096	0.000	0.017	0.038

Omnibus:	3.850	Durbin-Watson:	2.004
Prob(Omnibus):	0.146	Jarque-Bera (JB):	3.722
Skew:	0.143	Prob(JB):	0.156
Kurtosis:	3.215	Cond. No.	1.67e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.67e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Fig 22: OLS Model with Multicollinearity and high p-values

1. **Adjusted. R-squared**: It reflects the fit of the model.
    - Adjusted R-squared values generally range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
    - In our case, the value for adj. R-squared is **0.785**, which is good.

2. ***const* coefficient**: It is the Y-intercept.
    - It means that if all the predictor variable coefficients are zero, then the expected output (i.e., Y) would be equal to the *const* coefficient.

- In our case, the value for const coefficient is **0.0602**

3. **Coefficient of a predictor variable**: It represents the change in the output Y due to a change in the predictor variable (everything else held constant).
    - In our case, the coefficient of Visitors is **0.1295**

Interpretation of Coefficients

- The coefficients tell us how one unit change in X can affect y.
- The sign of the coefficient indicates if the relationship is positive or negative.
- In this data set, for example, presence of sports event on same day occurs with a 0.0603 decrease in viewership, and increase in visitors by single person occurs with a 0.1295 increase in the content viewership.
- Earlier in the heat map, we observed that the relationship between major sports events and content viewership is negatively correlated (as sports events increase, content viewership decreases, and vice versa), while the relationship between the number of visitors and content viewership is positively correlated (as the number of visitors increases, content viewership also increases, and vice versa). Consequently, the signs of the coefficients align with these relationships, suggesting a low probability of multicollinearity in our data.
- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

Training Performance				
RMSE	MAE	R-squared	Adj. R-squared	MAPE
0.04853	0.038197	0.791616	0.785162	8.55644
Test Performance				
0.050603	0.040782	0.766447	0.748804	9.030464

Table 2: Training and Test Performance of existing variables

Observations

- The training R2 is 0.79, so the model is not underfitting.
- The train and test RMSE and MAE are comparable, so the model is not overfitting either.
- MAE suggests that the model can predict anime ratings within a mean error of 0.40 on the test data.
- MAPE of 9.03 on the test data means that we are able to predict within 9.03% of the anime ratings.

7. Checking linear Regression Assumptions

We will be checking the following Linear Regression assumptions:

1. **No Multicollinearity**
2. **Linearity of variables**

3. **Independence of error terms**
4. **Normality of error terms**
5. **No Heteroscedasticity**

7.1. Checking for Multicollinearity

Variables	VIF
const	99.67932
visitors	1.027837
ad_impressions	1.02939
views_trailer	1.023551
major_sports_event_yes	1.065689
genre_Comedy	1.917635
genre_Drama	1.926699
genre_Horror	1.90446
genre_Others	2.573779
genre_Romance	1.753525
genre_Sci-Fi	1.863473
genre_Thriller	1.921001
dayofweek_Monday	1.063551
dayofweek_Saturday	1.155744
dayofweek_Sunday	1.150409
dayofweek_Thursday	1.16987
dayofweek_Tuesday	1.062793
dayofweek_Wednesday	1.315231
season_Spring	1.541591
season_Summer	1.56824
season_Winter	1.570338

Table 3: VIF (Variance Inflation factor) before Removal of Multicollinearity

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

On dropping 'genre_Others', adj. R-squared increased by 0.001 and R-Square decreased by 0.001

- VIF after dropping genre others has all the values less than 2.

Variables	VIF
const	87.57068
visitors	1.022226
ad_impressions	1.028804
views_trailer	1.020524
major_sports_event_yes	1.065264
genre_Comedy	1.204848

genre_Drama	1.223443
genre_Horror	1.204654
genre_Romance	1.171988
genre_Sci-Fi	1.205594
genre_Thriller	1.20656
dayofweek_Monday	1.063551
dayofweek_Saturday	1.154886
dayofweek_Sunday	1.150034
dayofweek_Thursday	1.169852
dayofweek_Tuesday	1.058831
dayofweek_Wednesday	1.31438
season_Spring	1.541573
season_Summer	1.545311
season_Winter	1.568494

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

- All the variables have VIF less than 2.
- We have dealt with multicollinearity in the data
- Let's rebuild the model using the updated set of predictors variables

Check the Model after dropping the column genre others.

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.791
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	135.8
Date:	Sun, 30 Jun 2024	Prob (F-statistic):	1.66e-216
Time:	11:19:51	Log-Likelihood:	1124.2
No. Observations:	700	AIC:	-2208.
Df Residuals:	680	BIC:	-2117.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0660	0.017	3.786	0.000	0.032	0.100
visitors	0.1289	0.008	16.378	0.000	0.113	0.144
ad_impressions	3.482e-06	6.58e-06	0.529	0.597	-9.43e-06	1.64e-05
views_trailer	0.0023	5.51e-05	42.213	0.000	0.002	0.002
major_sports_event_yes	-0.0604	0.004	-15.307	0.000	-0.068	-0.053
genre_Comedy	0.0050	0.006	0.789	0.430	-0.007	0.017
genre_Drama	0.0082	0.006	1.270	0.204	-0.004	0.021
genre_Horror	0.0054	0.006	0.835	0.404	-0.007	0.018
genre_Romance	-0.0038	0.007	-0.552	0.581	-0.017	0.010
genre_Sci-Fi	0.0088	0.007	1.326	0.185	-0.004	0.022
genre_Thriller	0.0043	0.006	0.672	0.502	-0.008	0.017
dayofweek_Monday	0.0337	0.012	2.848	0.005	0.010	0.057
dayofweek_Saturday	0.0581	0.007	8.122	0.000	0.044	0.072
dayofweek_Sunday	0.0362	0.008	4.624	0.000	0.021	0.052
dayofweek_Thursday	0.0173	0.007	2.555	0.011	0.004	0.031
dayofweek_Tuesday	0.0236	0.014	1.723	0.085	-0.003	0.050
dayofweek_Wednesday	0.0475	0.004	10.577	0.000	0.039	0.056
season_Spring	0.0226	0.005	4.228	0.000	0.012	0.033
season_Summer	0.0436	0.005	8.063	0.000	0.033	0.054
season_Winter	0.0270	0.005	5.069	0.000	0.017	0.037

Omnibus:	4.537	Durbin-Watson:	2.002
Prob(Omnibus):	0.103	Jarque-Bera (JB):	4.462
Skew:	0.154	Prob(JB):	0.107
Kurtosis:	3.240	Cond. No.	1.46e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Fig 23: OLS Model After Removal of Multicollinearity

Observations

- We can see that adj. R-squared has dropped from 0.792 to 0.791, 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.

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.789
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	233.8
Date:	Sun, 30 Jun 2024	Prob (F-statistic):	7.03e-224
Time:	11:34:28	Log-Likelihood:	1120.2
No. Observations:	700	AIC:	-2216.
Df Residuals:	688	BIC:	-2162.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0747	0.015	5.110	0.000	0.046	0.103
visitors	0.1291	0.008	16.440	0.000	0.114	0.145
views_trailer	0.0023	5.5e-05	42.414	0.000	0.002	0.002
major_sports_event_yes	-0.0606	0.004	-15.611	0.000	-0.068	-0.053
dayofweek_Monday	0.0321	0.012	2.731	0.006	0.009	0.055
dayofweek_Saturday	0.0570	0.007	8.042	0.000	0.043	0.071
dayofweek_Sunday	0.0344	0.008	4.456	0.000	0.019	0.050
dayofweek_Thursday	0.0154	0.007	2.307	0.021	0.002	0.029
dayofweek_Wednesday	0.0465	0.004	10.532	0.000	0.038	0.055
season_Spring	0.0226	0.005	4.259	0.000	0.012	0.033
season_Summer	0.0434	0.005	8.112	0.000	0.033	0.054
season_Winter	0.0282	0.005	5.362	0.000	0.018	0.039

Omnibus:	3.254	Durbin-Watson:	1.996
Prob(Omnibus):	0.196	Jarque-Bera (JB):	3.077
Skew:	0.139	Prob(JB):	0.215
Kurtosis:	3.168	Cond. No.	662.

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

Fig 24: OLS Model without Multicollinearity and high p-values

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

Test Performance				
0.051109	0.041299	0.761753	0.751792	9.177097

Table 5: Training and Test Performance after removal of variables

Observations

- Now no feature has p-value greater than 0.05, so we'll consider the features in *x_train2* as the final set of predictor variables and *olsmod3* as the final model to move forward with
- Now adjusted R-squared is 0.788, i.e., our model is able to explain ~79% of the variance
- The adjusted R-squared in *olsmod2* (where we considered the variables without multicollinearity) was 0.786
    - This shows that the variables we dropped were not affecting the model
- RMSE and MAE values are comparable for train and test sets, indicating that the model is not overfitting

After dropping the features causing strong multicollinearity and the statistically insignificant ones, our model performance hasn't dropped sharply (adj. R-squared has dropped from 0.792 to 0.789). This shows that these variables did not have much predictive power.

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

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

7.2. Linearity and Independence of Variables

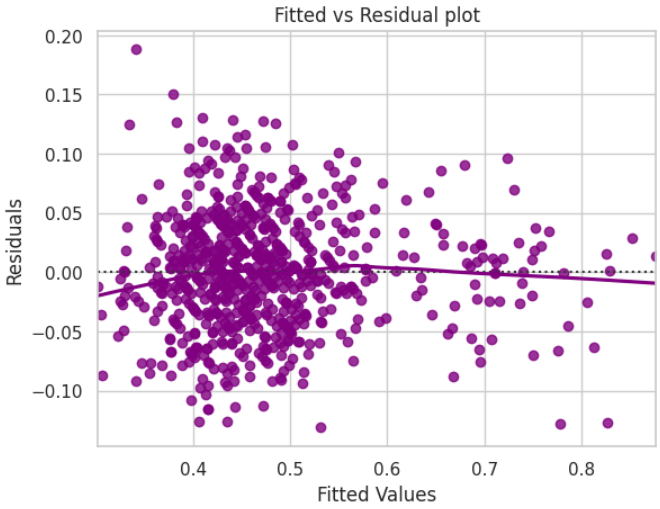


Fig 25: Linearity and Independence of Variables

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

7.3. Normality of error terms

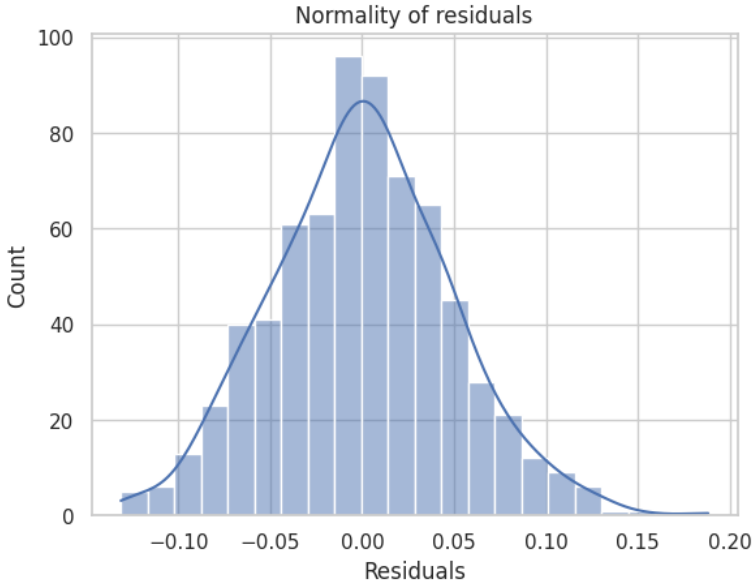


Fig 26: Normality of Residuals

- The residual terms are normally distributed

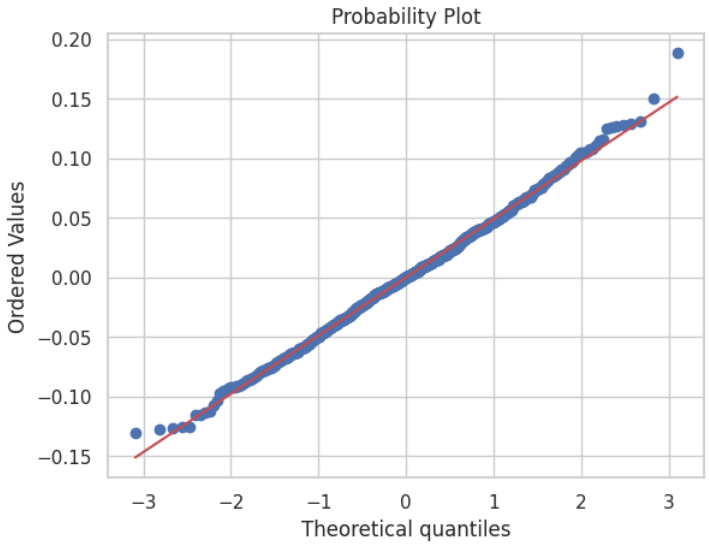


Fig 27: Q-Q plot of residuals

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

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

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

pvalue=0.3104695975780487

- Since p-value > 0.05, the residuals are normal as per shapiro test.

7.4. Test for Homoscedasticity

Pvalue : 0.12853551819087372

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

8. Predictions on test data

	Actual	Predicted
983	0.43	0.434802
194	0.51	0.500314
314	0.48	0.430257
429	0.41	0.492544
267	0.41	0.487034
746	0.68	0.68
186	0.62	0.595078
964	0.48	0.503909
676	0.42	0.490313
320	0.58	0.560155

Fig 28: Actual vs Predicted value of built model

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

9. Final Model

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.789
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	233.8
Date:	Sun, 30 Jun 2024	Prob (F-statistic):	7.03e-224
Time:	12:15:56	Log-Likelihood:	1120.2
No. Observations:	700	AIC:	-2216.
Df Residuals:	688	BIC:	-2162.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0747	0.015	5.110	0.000	0.046	0.103
visitors	0.1291	0.008	16.440	0.000	0.114	0.145
views_trailer	0.0023	5.5e-05	42.414	0.000	0.002	0.002
major_sports_event_yes	-0.0606	0.004	-15.611	0.000	-0.068	-0.053
dayofweek_Monday	0.0321	0.012	2.731	0.006	0.009	0.055
dayofweek_Saturday	0.0570	0.007	8.042	0.000	0.043	0.071
dayofweek_Sunday	0.0344	0.008	4.456	0.000	0.019	0.050
dayofweek_Thursday	0.0154	0.007	2.307	0.021	0.002	0.029
dayofweek_Wednesday	0.0465	0.004	10.532	0.000	0.038	0.055
season_Spring	0.0226	0.005	4.259	0.000	0.012	0.033
season_Summer	0.0434	0.005	8.112	0.000	0.033	0.054
season_Winter	0.0282	0.005	5.362	0.000	0.018	0.039

Omnibus:	3.254	Durbin-Watson:	1.996
Prob(Omnibus):	0.196	Jarque-Bera (JB):	3.077
Skew:	0.139	Prob(JB):	0.215
Kurtosis:	3.168	Cond. No.	662.

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

Fig 29: Final OLS Model

Training Performance				
RMSE	MAE	R-squared	Adj. R-squared	MAPE
0.048841	0.038385	0.788937	0.785251	8.595246
Test Performance				
0.051109	0.041299	0.761753	0.751792	9.177097

Fig 30: Training and Test Performance of Final Model

- The model is able to explain ~79% of the variation in the data
- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting
- The MAPE on the test set suggests we can predict within 9.17% of the anime ratings
- Hence, we can conclude the model *olsmodel_final* is good for prediction as well as inference purposes

10. Conclusions and Recommendations

- The model's R-squared value is approximately 0.79, and the adjusted R-squared is 0.786, indicating that the model can explain about 79% of the variance in the data. This is quite satisfactory.
    - This suggests that the model is suitable for both prediction and inference purposes.
- A major sports event will lead to a 0.0604 unit decrease in content viewership, assuming all other variables remain constant.
    - To improve content viewership, it is recommended to avoid releasing content on days when major sports events are happening.
- An increase of one unit in the 'visitors' variable results in a 0.1291 unit increase in content viewership, with all other variables held constant.
    - The client should provide more detailed information to identify the reasons behind increases in viewership.
- An increase of one unit in trailer viewers will result in a 0.0023 unit increase in content viewership, all other variables held constant.
- Releasing content on specific days of the week will increase viewership: Saturday (0.0583 units), Wednesday (0.0478 units), Sunday (0.0357 units), Monday (0.0334 units), Tuesday (0.0242 units), and Thursday (0.0168 units), with all other variables held constant.
    - Therefore, releasing content on Saturdays and Wednesdays will boost viewership, provided no major sports events occur on those days.
- The summer season can result in a 0.0438 unit increase in content viewership, with all other variables held constant.
    - Releasing content during the summer season can enhance viewership.

11. Final Linear Regression Equation

Content Views = 0.07467052053721143 + 0.12909581825894134 * (visitors) + 0.0023308167861640127 * (views_trailer) + -0.06055507818137333 * (major_sports_event_yes) + 0.03206580679023641 * (dayofweek_Monday) + 0.05702859660165112 * (dayofweek_Saturday) + 0.03438622992362503 * (dayofweek_Sunday) + 0.015449441769973173 * (dayofweek_Thursday) + 0.04649480366984801 * (dayofweek_Wednesday) + 0.022604915818117893 * (season_Spring) + 0.04339100263609974 * (season_Summer) + 0.028230557183976775 * (season_Winter)