greatlearning
Learning for Life

Business report

Over The Top (OTT)

Media services

DATA SCIENCE AND BUSINESS ANALYTICS

greatlearning
Learning for Life

Content:

1 .Exploratory Data Analysis

- 1.1. Problem definition
- 1.2. Data background and contents
- 1.3. Univariate analysis
- 1.4. Bivariate analysis
- 1.5. Answers to the key questions provided
- 1.6. Insights based on EDA

2. Data preprocessing

- 2.1. Duplicate value check
- 2.2. Missing value treatment
- 2.3. Outlier treatment
- 2.4. Feature engineering
- 2.5. Data preparation for modeling

3. Model building - Linear Regression

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

4. Testing the assumptions of linear regression model

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

5. Model performance evaluation

- 5.1. Evaluate the model on different performance metrics

6. Actionable Insights & Recommendations

- 6.1. Comments on significance of predictors
- 6.2. Key takeaways for the business

1. Exploratory Data Analysis

1.1. Problem definition:

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.

1.2. Data background and content.

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 121.61billion 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.

Data Description

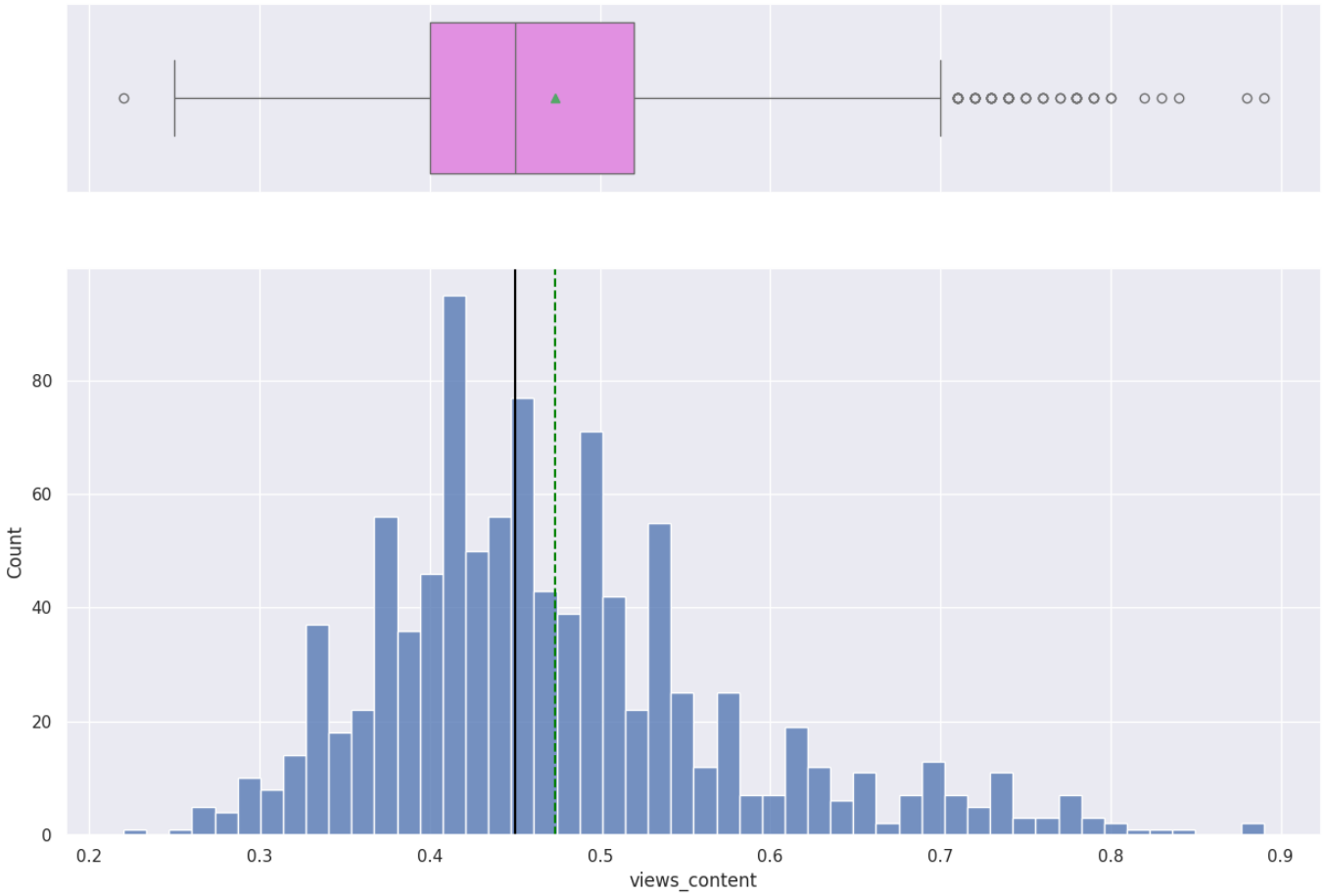
The data contains the different factors to analyze for the content. The detailed data dictionary is given below.

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

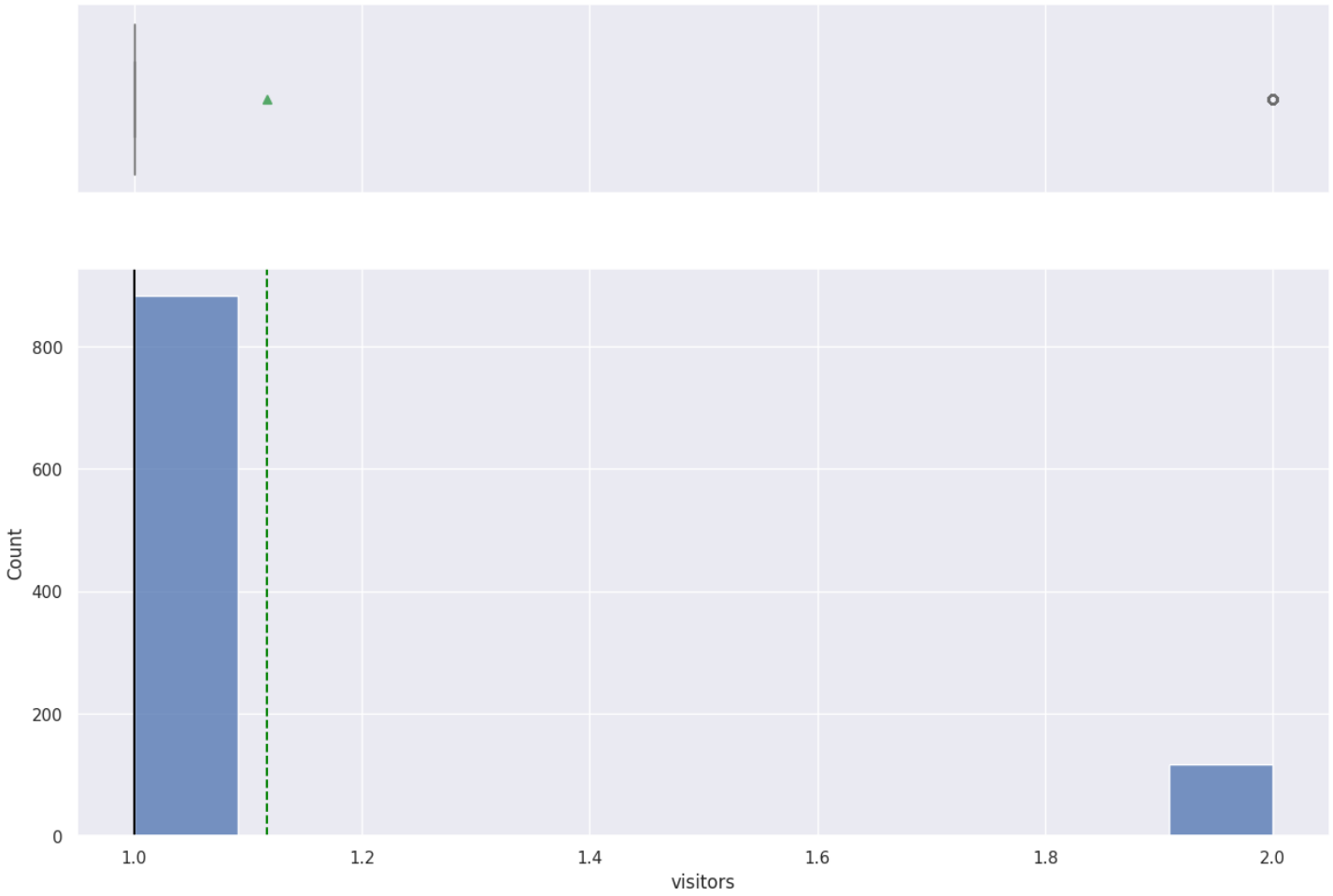
1.3. Univariate analysis.

- **VIEWS CONTENT:**



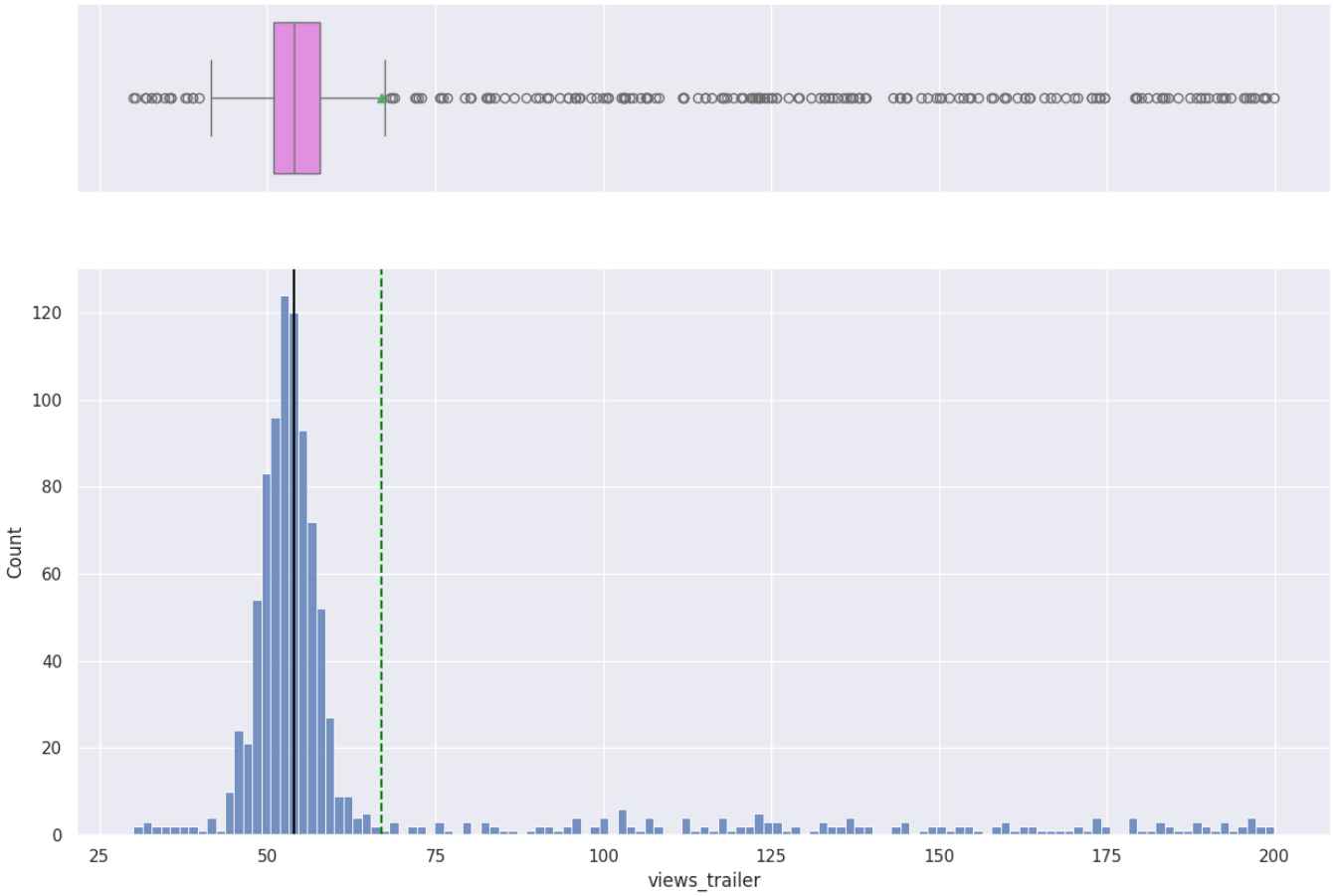
- Viewership of content is slightly normally distributed with right tail
- Average and the 50% of the viewership content is between 0.4 to 0.5.

- **VISITORS:**



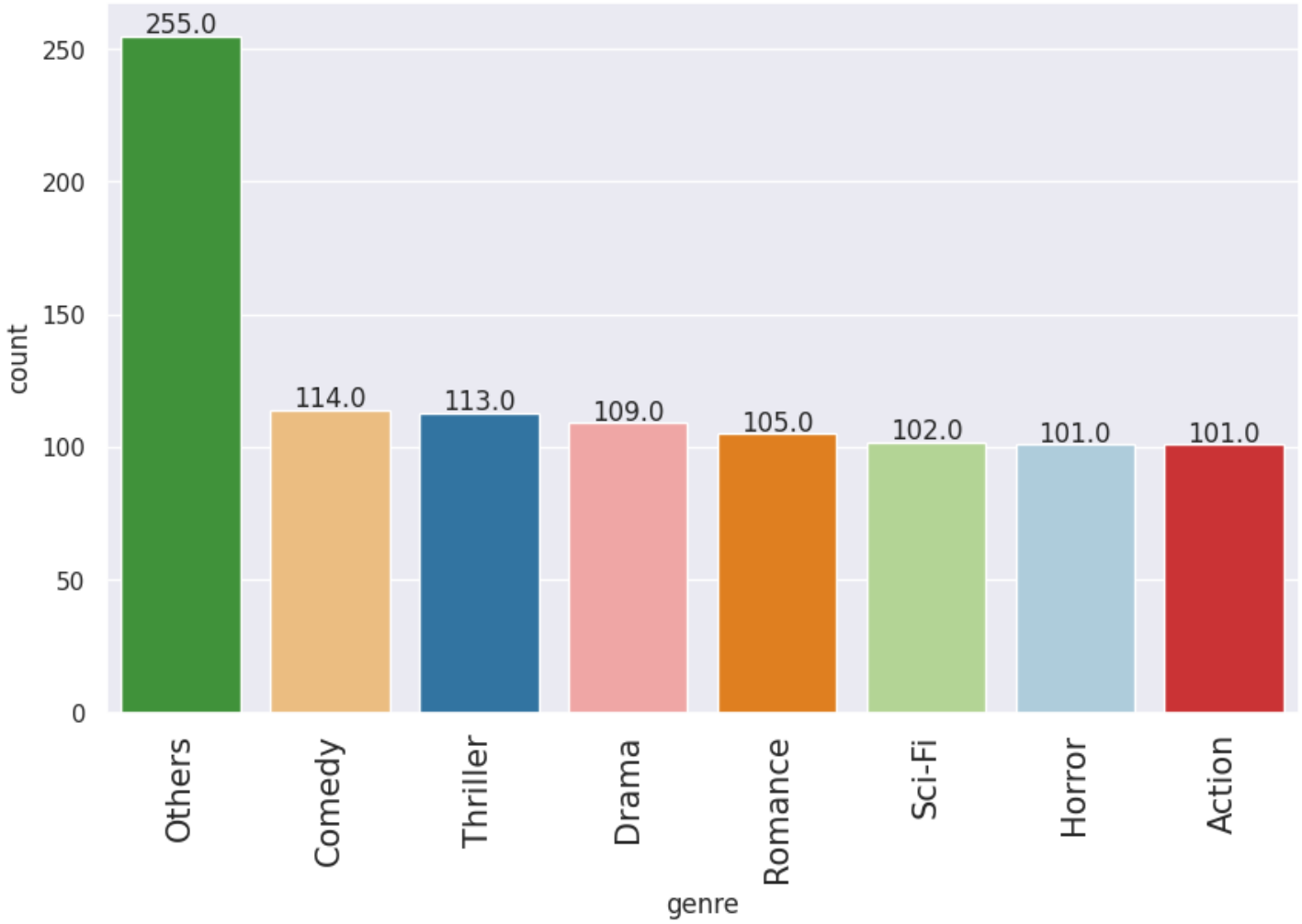
- There is a single outlier present in the visitors column
- The mean and the median are at 1.7.

- **VIEW TRAILER:**



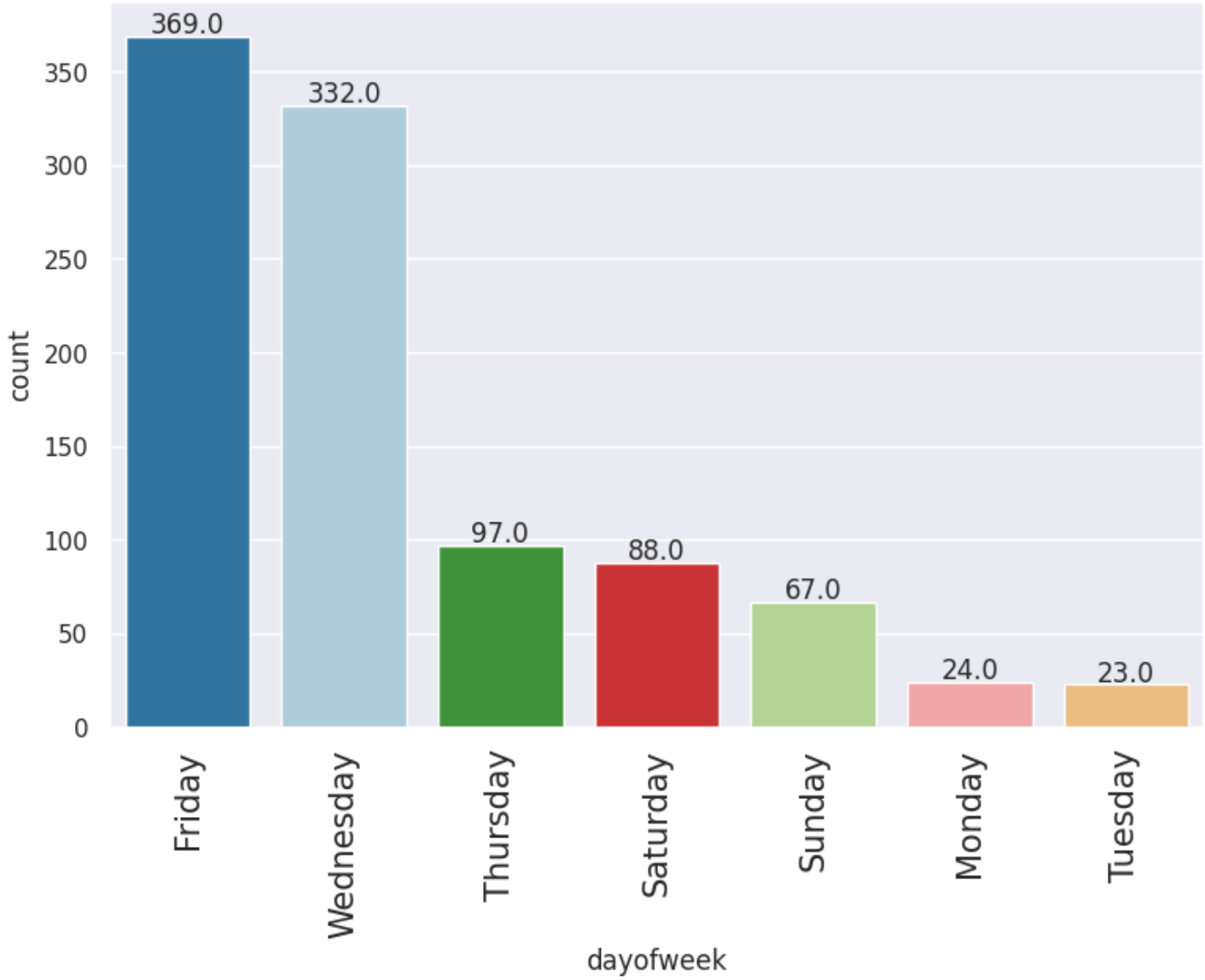
- The distribution of views trailer is highly right skewed with so many outliers in it
- Most of number of views on trailer is around 50.

- **GENRE:**



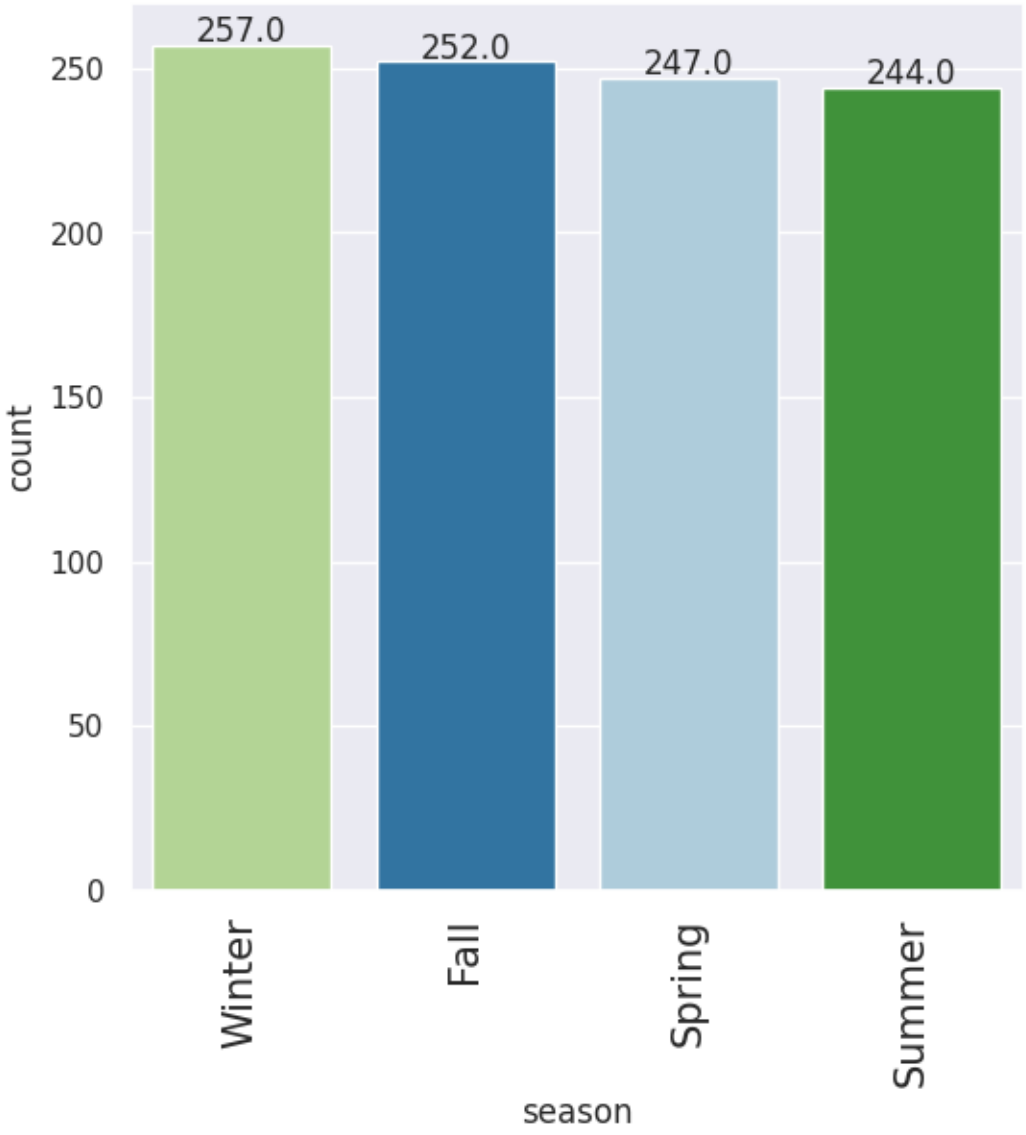
- Genre of viewers is more in others.
- Comedy is preferred by most of the people.

- **DAY OF WEEK:**



- According to the above plot most of the movies had been watched on Fridays even more than Sunday.
- Second preferable day for most of the viewers is Wednesday.

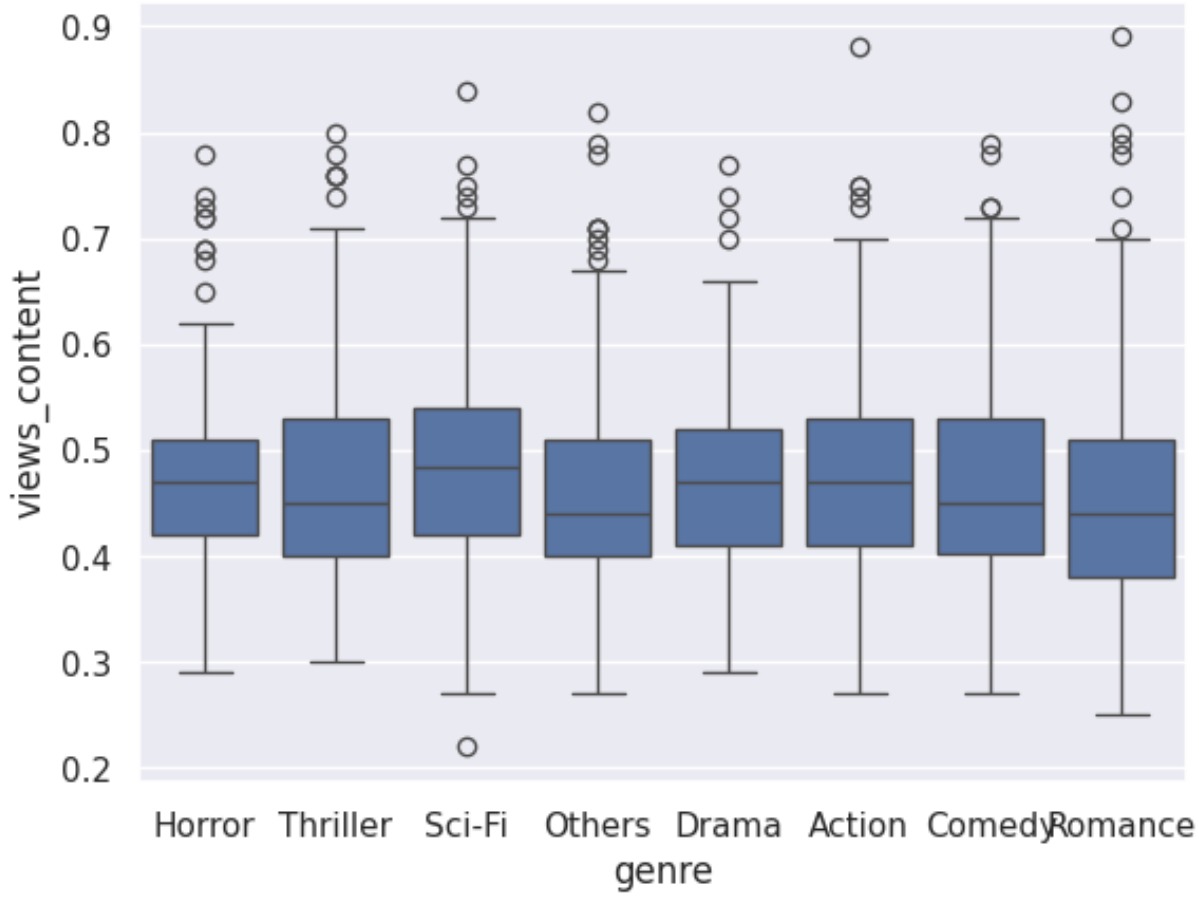
- **SEASON:**



- Most common season where people prefer to watch movies is winter.
- Later on most of the people watch movies on fall.

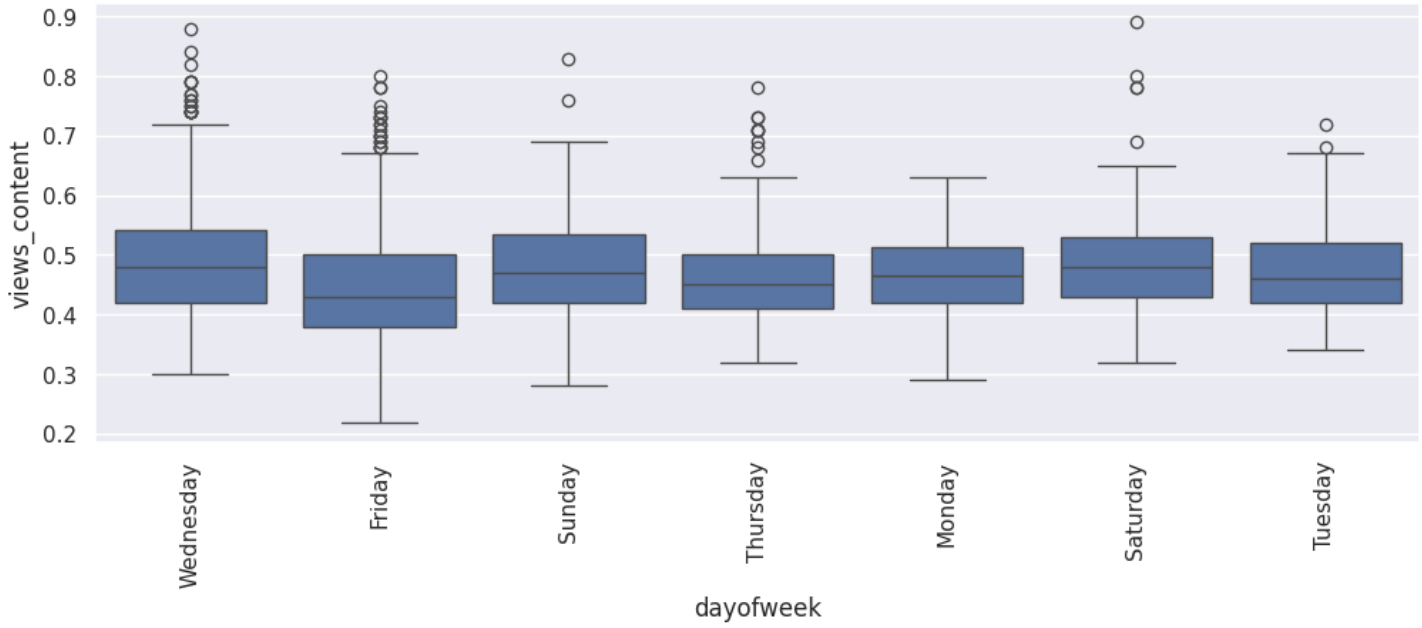
1.4. BIVARIATE ANALYSIS

- **GENRE VS VIEWS CONTENT:**



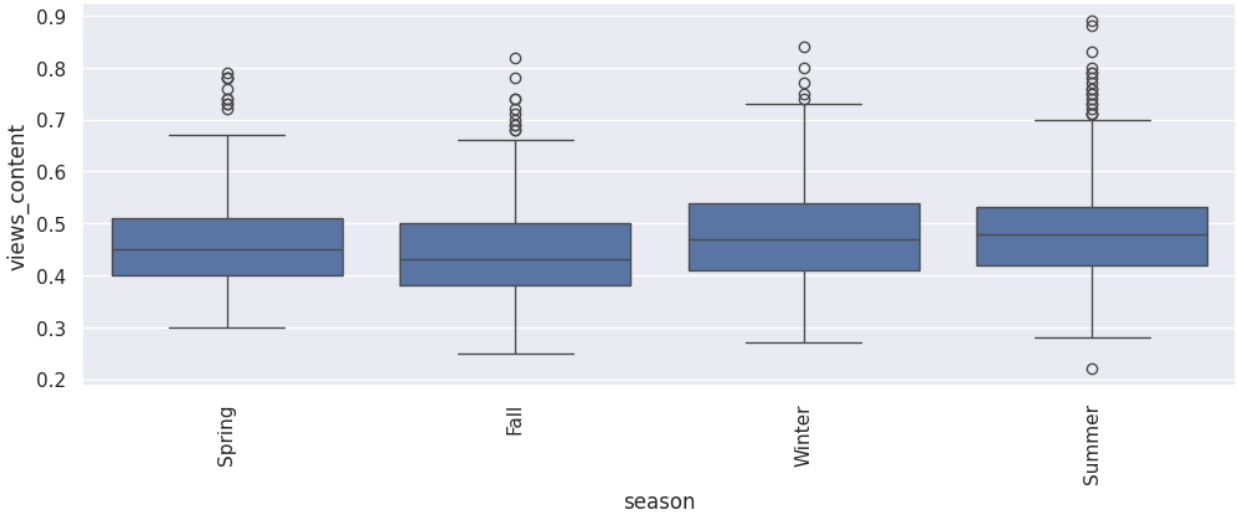
- View content on romance and action has more outliers
- Sci-Fi has outliers in both the sides.

- **DAY OF WEEK VS VIEWS CONTENT:**



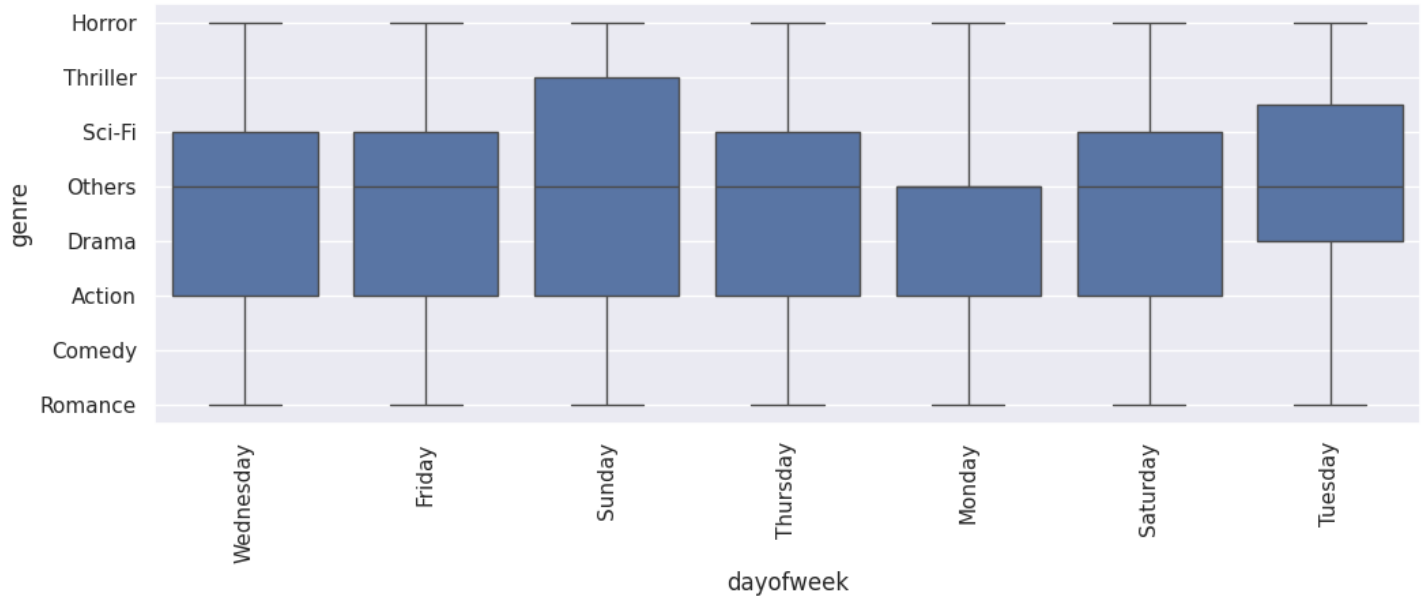
➢ Wednesday has more views content and second largest is Friday.

- **SEASON VS VIEWS CONTENT:**



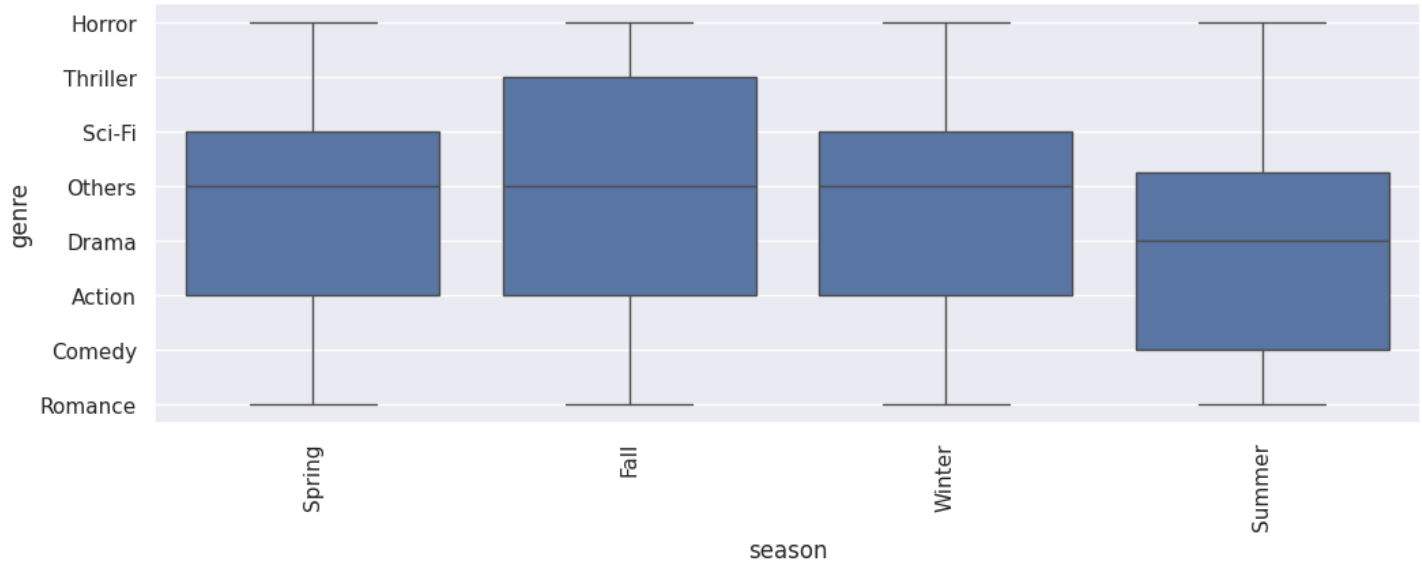
➢ Views content has more on summer.

- **DAY OF WEEK VS GENRE:**



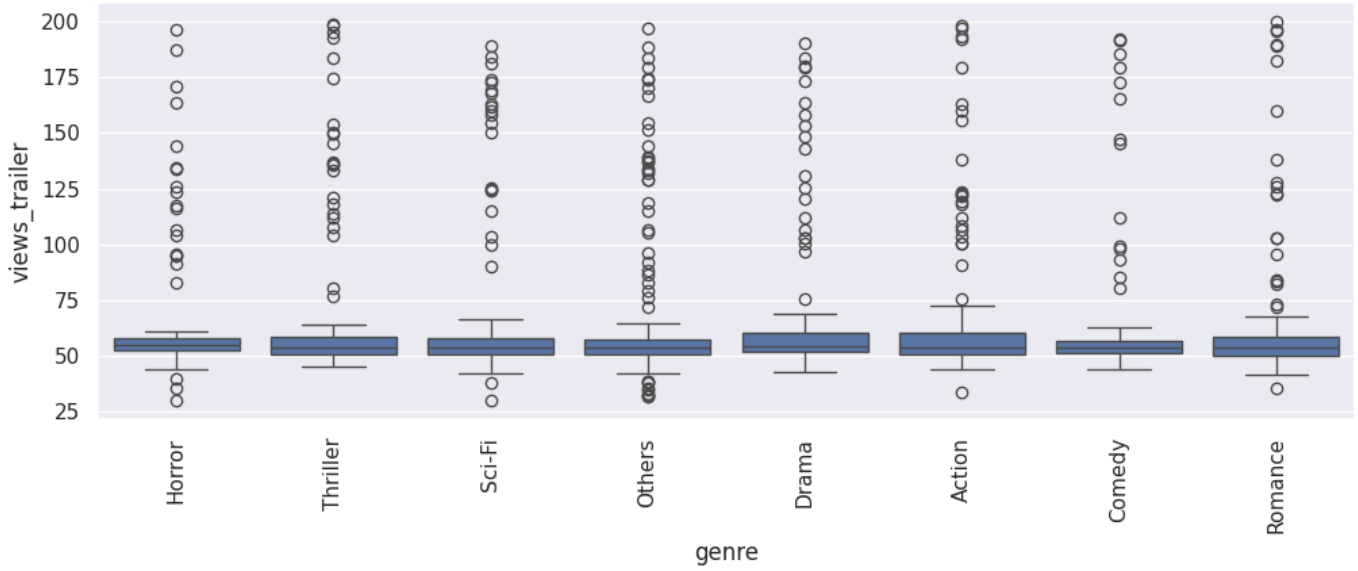
➢ On Sunday thriller had been watched more.

- **SEASON VS GENRE:**



➢ In Fall sci-fi ,thriller has watched more.

- **GENRE VS VIEWS TRAILER:**

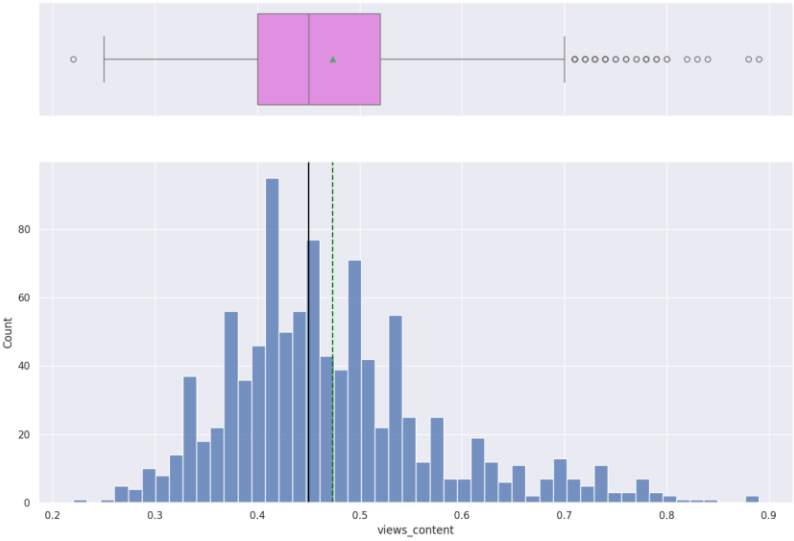


➢ In views trailer other has more outliers.

1.5. Answers to the key questions provided:

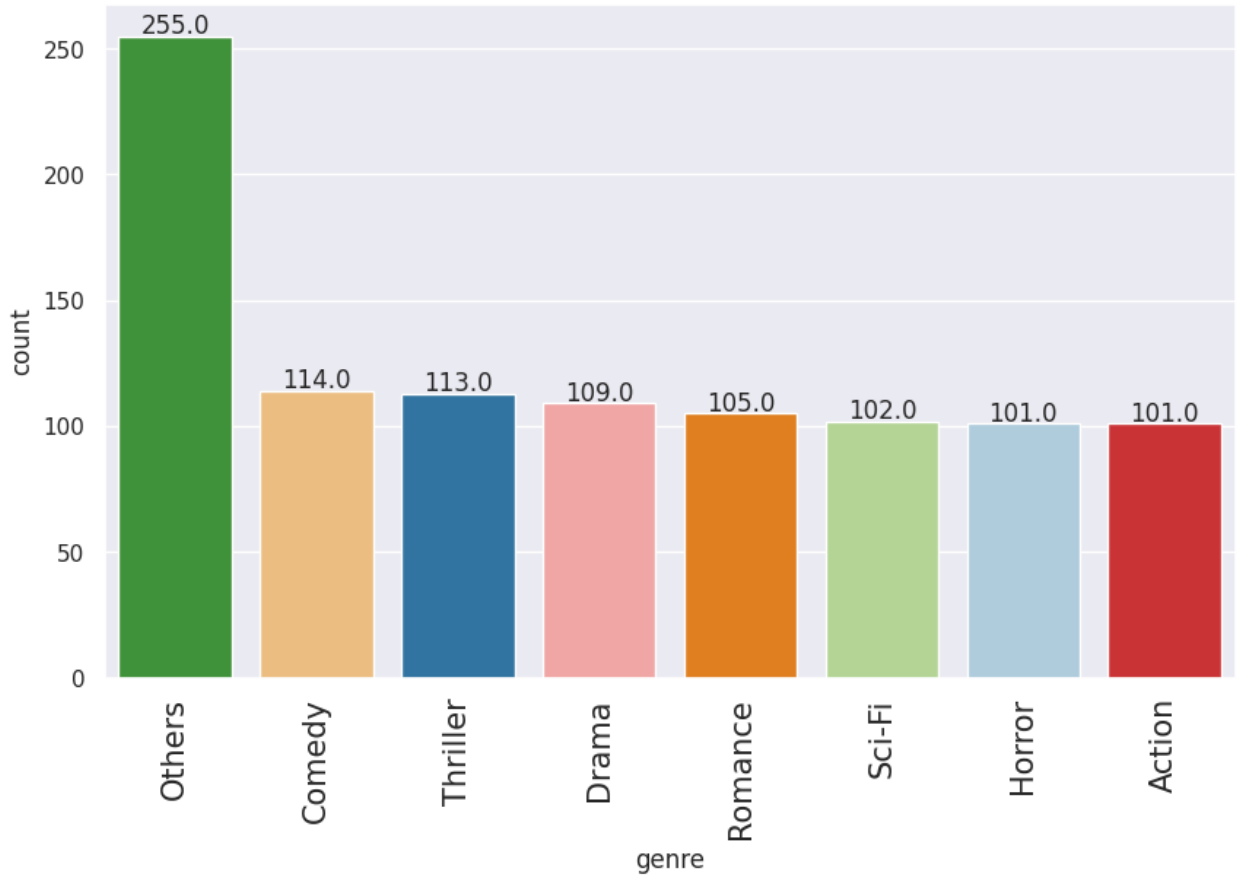
The following questions need to be answered:

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



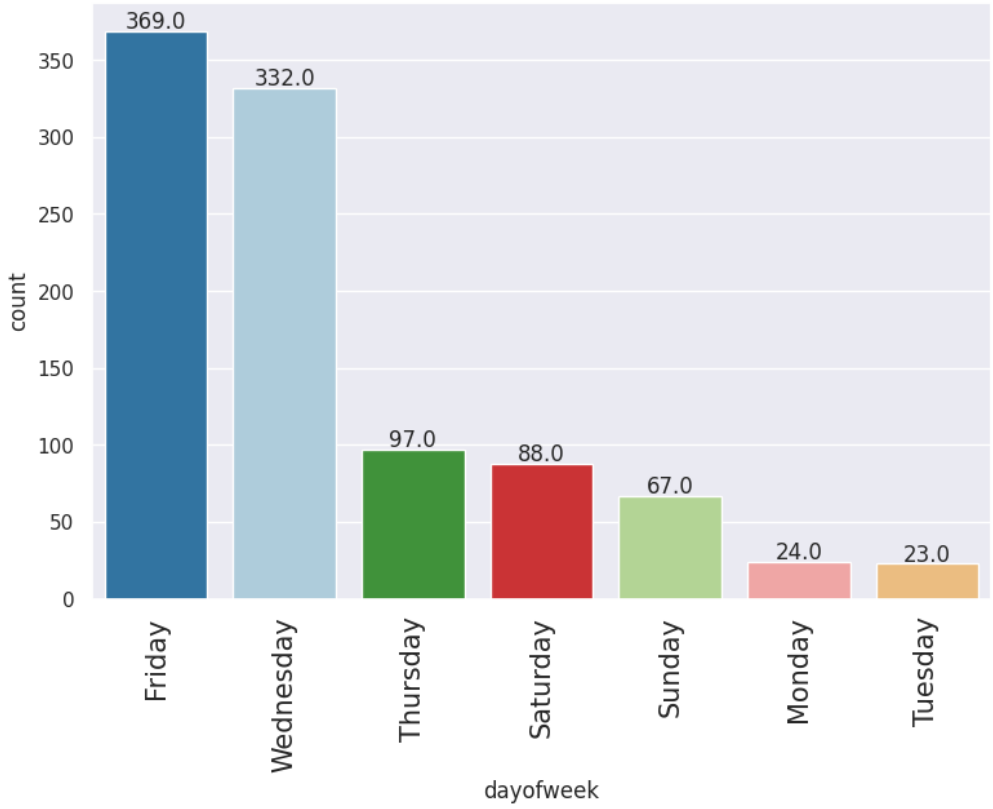
- Viewership of content is slightly normally distributed with right tail
- Average and the 50% of the viewership content is between 0.4 to 0.5.
- There are lots of outliers in the right and a single outliers in the left.

2)What does the distribution of genres look like?



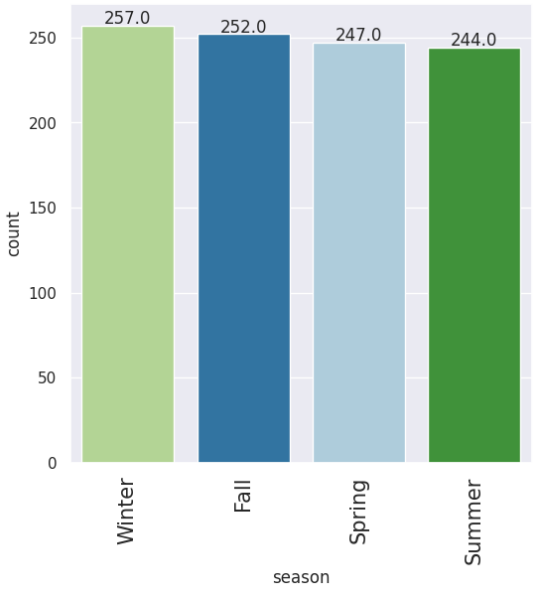
- Genre of viewers is more in others.
- Comedy, Thriller, Drama are preferred by most of the people.

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?



- According to the above plot most of the movies had been watched on Fridays even more than Sunday.
- Second preferable day for most of the viewers is Wednesday.

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



- Most common season where people prefer to watch movies is winter.
- Later on most of the people watch movies on fall.

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

- The views of trailer and views content have high correlation which means they are highly depended on each other.
- It has the highest correlation among all the variables.
- ad_impressions has little correlation with visitors ,views trailer and views content but has negatively correlation with major sports event.
- Major sports event has little negative correlation with visitors, ad_impressions and views content whereas views trailer is positively correlated.
- Visitors have little correlation with ad impressions and negatively correlated with major sports event and views trailer, whereas visitors are correlated with views content.

1.6. Insights based on EDA:

- Viewership of content is slightly normally distributed with right tail
- Average and the 50% of the viewership content is between 0.4 to 0.5.
- There is a single outlier present in the visitors column
- The mean and the median are at 1.7 in the visitors column.
- The distribution of views trailer is highly right skewed with so many outliers in it
- Most of number of views on trailer is around 50.
- Genre of viewers is more in others.
- Comedy is preferred by most of the people.
- According to the plot most of the movies had been watched on Fridays even more than Sunday and Second preferable day for most of the viewers is Wednesday.
- Most common season where people prefer to watch movies is winter.
- Later on most of the people watch movies on fall.
- View content on romance and action has more outliers
- On Sunday, thriller had been watched more.
- In Fall sci-fi, thriller has watched more.

2. DATA PREPROCESSING

2.1. Checking for duplicate entries in the dataset

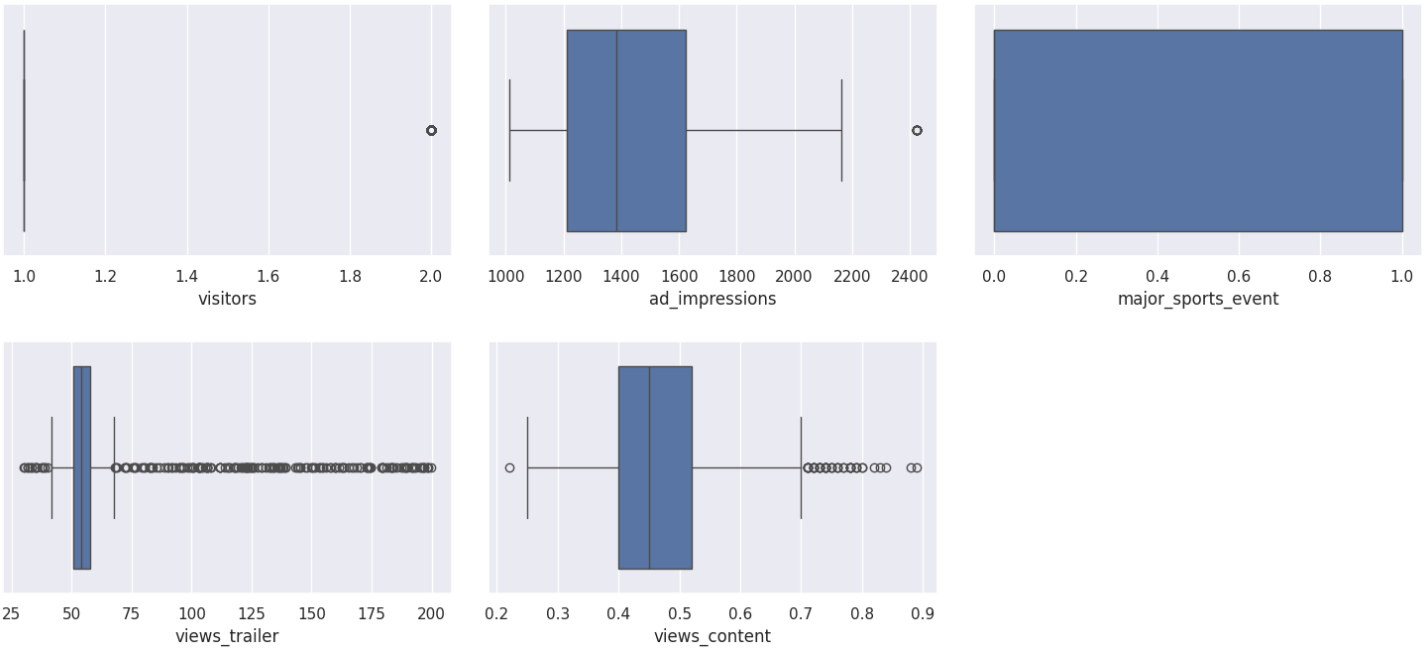
- There are no duplicate entries in the dataset.

2.2. Checking for missing values in the dataset.

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

There are no null values in the dataset.

2.3. Outlier checking and treatment



- There are quite a few outliers in the data
- If we treat it will result in loss of data.
- Outliers in the column seem to be genuine.
- However, we will not treat them as they are proper value.

2.4. Feature engineering

- **Ad_impression per visitors:**

```
0      1113.810
1      1498.410
2      1079.190
3      1342.770
4      1498.410
         ...
995    1311.960
996    1329.480
997    1359.800
998     849.175
999    1140.230
Length: 1000, dtype: float64
```

- **Views trailer per visitors:**

```
0       56.70
1       52.69
2       48.74
3       49.81
4       55.83
        ...
995     48.58
996     72.42
997    150.44
998     24.36
999     52.94
Length: 1000, dtype: float64
```

- **Major sports event per ad_impression:**

	0
0	0.00
1	1498.41
2	1079.19
3	1342.77
4	0.00
...	...
995	0.00
996	0.00
997	1359.80
998	0.00
999	0.00

2.5. Data preparation for modeling

- We want to predict the views content.
- Before we proceed to build a model, we'll have to encode categorical features
- 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

➢ **Dropping views content as we want to predict views content:**

Remaining columns:

	visitors	ad_impressions	major_sports_event	genre	dayofweek	season
0	1	1113.81	0	Horror	Wednesday	Spring
1	1	1498.41	1	Thriller	Friday	Fall
2	1	1079.19	1	Thriller	Wednesday	Fall
3	1	1342.77	1	Sci-Fi	Friday	Fall
4	1	1498.41	0	Sci-Fi	Sunday	Winter

	views_trailer
0	56.70
1	52.69
2	48.74
3	49.81
4	55.83
0	0.51
1	0.32
2	0.39
3	0.44
4	0.46

➢Creating dummy variables:

	visitors	ad_impressions	major_sports_event	views_trailer	genre_Comedy	genre_Drama	genre_Horror	genre_Others	genre_Romance	...	genre_Thriller	dayofweek_Monday	dayofweek_Saturday	dayofweek_Sunday	dayofweek_Thursday	dayofweek_Tuesday	dayofweek_Wednesday	season_Spring	season_Summer	season_Winter	
0	1.0	1	1113.81	0	56.70	False	False	True	False	...	False	False	False	False	False	False	False	True	True	False	False
1	1.0	1	1498.41	1	52.69	False	False	False	False	...	False	True	False	False	False	False	False	False	False	False	False
2	1.0	1	1079.19	1	48.74	False	False	False	False	...	False	True	False	False	False	False	False	True	False	False	False
3	1.0	1	1342.77	1	49.81	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False	False
4	1.0	1	1498.41	0	55.83	False	False	False	False	...	False	False	False	False	True	False	False	False	False	False	True

➢ Converting dummy variables into float:

	visitors	ad_impressions	major_sports_event	views_trailer	genre_Commedy	genre_Drama	genre_Horror	genre_Others	genre_Romance	...	genre_Thriller	dayofweek_Monday	dayofweek_Saturday	dayofweek_Sunday	dayofweek_Thursday	dayofweek_Tuesday	dayofweek_Wednesday	season_Spring	season_Summer	season_Winter	...
0	1.0	1.0	1113.81	0.0	56.70	0.0	0.0	1.0	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	1.0	1.0	1498.41	1.0	52.69	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	1.0	1.0	1079.19	1.0	48.74	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	1.0	1.0	1342.77	1.0	49.81	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	1.0	1.0	1498.41	0.0	55.83	0.0	0.0	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

➢ Splitting data into 70:30 ratio for train to test data:

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

greatlearning
Learning for Life

3. Model building - linear regression:

3.1. Build the model and Display model coefficients with column names

OLS Regression Results

Dep. Variable:	views_content	R-squared:	0.756
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	105.1
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	1.91e-192
Time:	01:41:11	Log-Likelihood:	1069.2
No. Observations:	700	AIC:	-2096.
Df Residuals:	679	BIC:	-2001.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1971	0.016	12.609	0.000	0.166	0.228
visitors	0.0735	0.006	11.406	0.000	0.061	0.086
ad_impressions	5.606e-06	7.12e-06	0.787	0.431	-8.38e-06	1.96e-05
major_sports_event	-0.0606	0.004	-14.179	0.000	-0.069	-0.052
views_trailer	0.0023	5.98e-05	38.916	0.000	0.002	0.002
genre_Comedy	0.0007	0.009	0.079	0.937	-0.016	0.018
genre_Drama	0.0104	0.009	1.187	0.236	-0.007	0.028
genre_Horror	0.0101	0.009	1.138	0.255	-0.007	0.027
genre_Others	0.0023	0.008	0.304	0.761	-0.013	0.017
genre_Romance	0.0004	0.009	0.042	0.966	-0.018	0.018
genre_Sci-Fi	0.0117	0.009	1.311	0.190	-0.006	0.029
genre_Thriller	0.0075	0.009	0.854	0.393	-0.010	0.025

```
dayofweek_Monday        0.0325      0.013      2.535      0.011       0.007
0.058
dayofweek_Saturday      0.0596      0.008      7.701      0.000       0.044
0.075
dayofweek_Sunday        0.0434      0.008      5.123      0.000       0.027
0.060
dayofweek_Thursday      0.0170      0.007      2.322      0.021       0.003
0.031
dayofweek_Tuesday       0.0265      0.015      1.785      0.075      -0.003
0.056
dayofweek_Wednesday     0.0476      0.005      9.790      0.000       0.038
0.057
season_Spring           0.0221      0.006      3.814      0.000       0.011
0.033
season_Summer           0.0443      0.006      7.504      0.000       0.033
0.056
season_Winter           0.0289      0.006      5.018      0.000       0.018
0.040
==============================================================================
Omnibus:                        2.208   Durbin-Watson:                   2.078
Prob(Omnibus):                  0.332   Jarque-Bera (JB):                2.221
Skew:                           0.136   Prob(JB):                        0.329
Kurtosis:                       2.960   Cond. No.                     1.48e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.48e+04. This might indicate that there are
   strong multicollinearity or other numerical problems.
```

Comment on the model statistics/Interpreting the Regression Results:

1. **Adjusted. R-squared**: It reflects the fit of the model.
    - Adjusted R-squared values generally range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
    - In our case, the value for adj. R-squared is **0.785**, 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 duration is **0.0123**.

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 R2.
- We will define a function to calculate MAPE and adjusted R2.
    - 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.

Training Performance:

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.052528	0.042369	0.755872	0.748311	9.53836

Test Performance:

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051737	0.042037	0.755858	0.737415	9.323966

Observations:

- The training R2 is 0.72, so the model is not under fitting
- The train and test RMSE and MAE are comparable, so the model is not over fitting either
- MAE suggests that the model can predict anime ratings within a mean error of 0.34 on the test data
- MAPE of 12.6 on the test data means that we are able to predict within 12.6% of the anime ratings

greatlearning
Learning for Life

4. Testing the assumption of linear regression model

4.1. Perform tests for the assumptions of the linear regression:

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**

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 kth 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.

Comment on 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 β_i is the coefficient of the ith independent variable.
        - H_o : Independent feature is not significant ($\beta_i = 0$)
        - H_a : Independent feature is that it is significant ($\beta_i \neq 0$)
    - (P>|t|) gives the p-value for each independent feature to check that null hypothesis. We are considering 0.05 (5%) as significance level.
        - A p-value of less than 0.05 is considered to be statistically significant.

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

Observations

- 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:	0.753
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	190.3
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	2.98e-200
Time:	01:41:11	Log-Likelihood:	1064.6
No. Observations:	700	AIC:	-2105.
Df Residuals:	688	BIC:	-2051.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2122	0.010	22.103	0.000	0.193	0.231
visitors	0.0729	0.006	11.384	0.000	0.060	0.085
major_sports_event	-0.0609	0.004	-14.496	0.000	-0.069	-0.053
views_trailer	0.0023	5.95e-05	39.195	0.000	0.002	0.002
dayofweek_Monday	0.0307	0.013	2.416	0.016	0.006	0.056
dayofweek_Saturday	0.0582	0.008	7.573	0.000	0.043	0.073
dayofweek_Sunday	0.0415	0.008	4.972	0.000	0.025	0.058
dayofweek_Thursday	0.0150	0.007	2.063	0.039	0.001	0.029
dayofweek_Wednesday	0.0462	0.005	9.663	0.000	0.037	0.056
season_Spring	0.0216	0.006	3.754	0.000	0.010	0.033
season_Summer	0.0430	0.006	7.424	0.000	0.032	0.054
season_Winter	0.0301	0.006	5.275	0.000	0.019	0.041

Omnibus:	1.735	Durbin-Watson:	2.069
Prob(Omnibus):	0.420	Jarque-Bera (JB):	1.817
Skew:	0.109	Prob(JB):	0.403
Kurtosis:	2.879	Cond. No.	486.

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

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.052876	0.042822	0.752621	0.7483	9.641289

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.052059	0.042188	0.752817	0.742481	9.384755

Comment:

- 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 *olsmod2* as the final model to move forward with
- Now adjusted R-squared is 0.7483, i.e., our model is able to explain ~75% of the variance
- The adjusted R-squared in *olsmod* (where we considered the variables without multicollinearity) was 0.749
    - 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

Now we'll check the rest of the assumptions on *olsmod2*.

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

greatlearning
Learning for Life

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.
- The independence of the error terms (or residuals) is important. If the residuals are not independent, then the confidence intervals of the coefficient estimates will be narrower and make us incorrectly conclude a parameter to be statistically significant.

How to check linearity and independence?

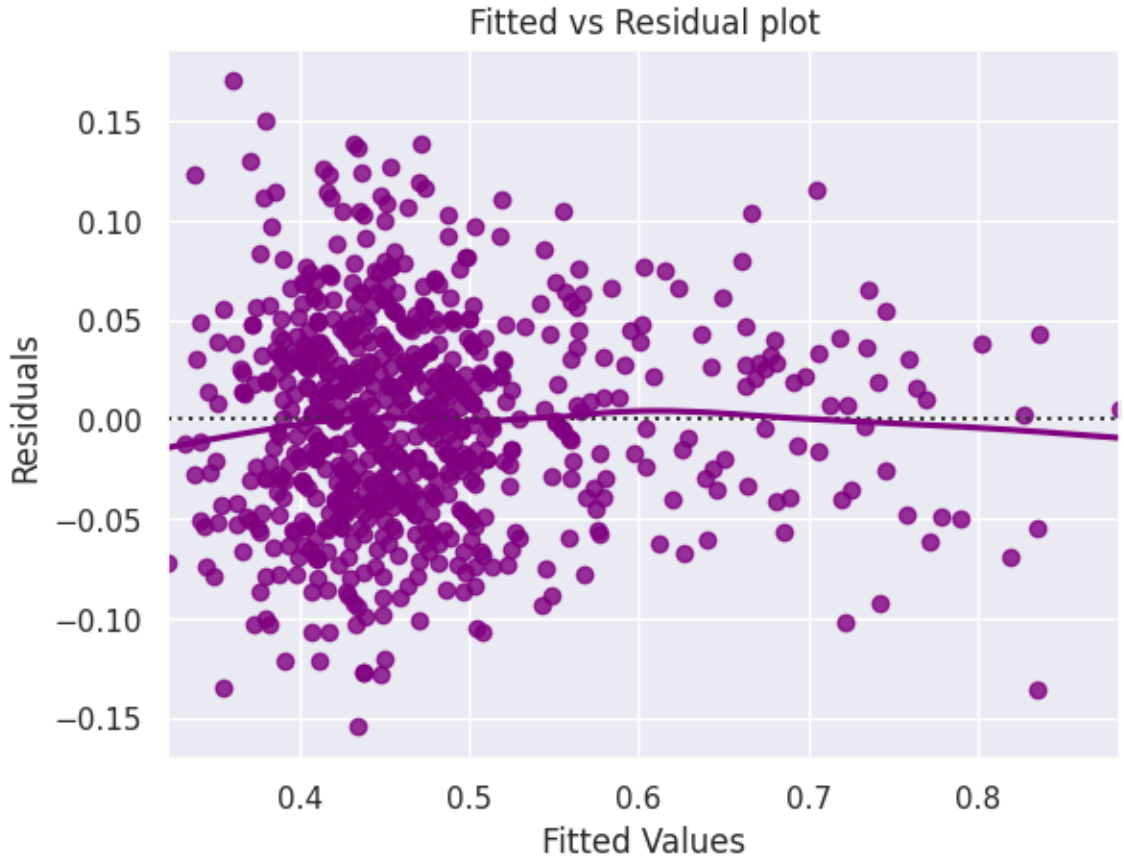
- Make a plot of fitted values vs residuals.
- If they don't follow any pattern, then we say the model is linear and residuals are independent.
- Otherwise, the model is showing signs of non-linearity and residuals are not independent.

How to fix if this assumption is not followed?

- We can try to transform the variables and make the relationships linear.

Actual Values	Fitted Values	Residuals	
731	0.40	0.443818	-0.043818
716	0.70	0.671362	0.028638
640	0.42	0.454337	-0.034337
804	0.55	0.579947	-0.029947
737	0.59	0.547080	0.042920

Fitted vs residual plot:



Comment:

- The scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.

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

greatlearning
Learning for Life

TEST FOR NORMALITY:

Why the test?

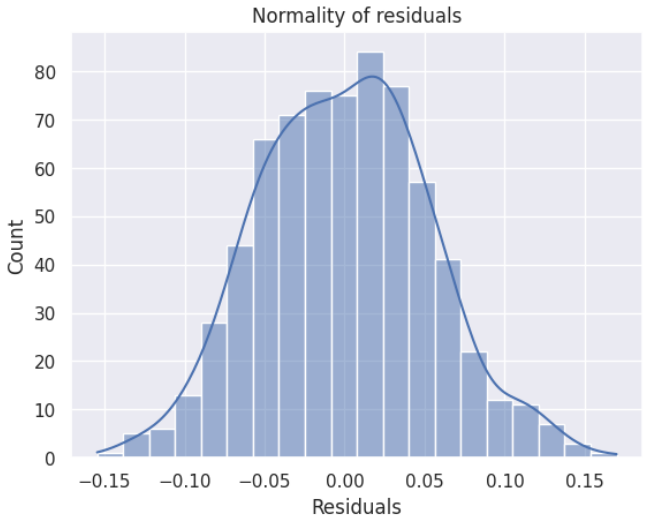
- Error terms, or residuals, should be normally distributed. If the error terms are not normally distributed, confidence intervals of the coefficient estimates may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Non-normality suggests that there are a few unusual data points that must be studied closely to make a better model.

How to check normality?

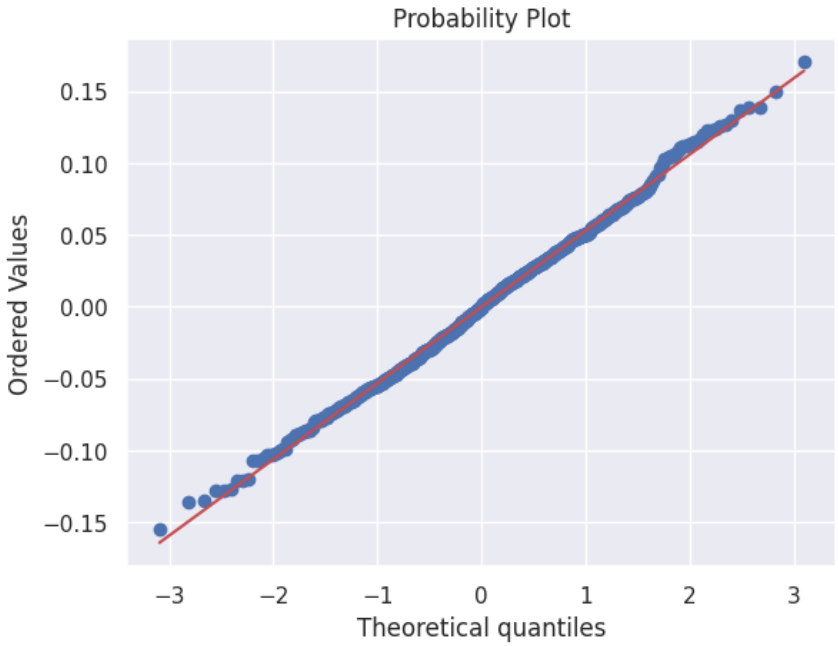
- The shape of the histogram of residuals can give an initial idea about the normality.
- It can also be checked via a Q-Q plot of residuals. If the residuals follow a normal distribution, they will make a straight line plot, otherwise not.
- Other tests to check for normality include the Shapiro-Walk test.
    - Null hypothesis: Residuals are normally distributed
    - Alternate hypothesis: Residuals are not normally distributed

How to fix if this assumption is not followed?

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



- The histogram of residuals does have a bell shape.
- Let's check the Q-Q plot.



- The residuals more or less follow a straight line except for the tails.
- Let's check the results of the Shapiro-Wilk test.

Comment:

- Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test.
- Strictly speaking, the residuals are not normal.
- However, as an approximation, we can accept this distribution as close to being normal.
- **So, the assumption is satisfied.**

greatlearning
Learning for Life

TEST FOR HOMOSCEDASTICITY

- **Homoscedascity**: If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic.
- **Heteroscedascity**: If the variance is unequal for the residuals across the regression line, then the data is said to be heteroscedastic.

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 for homoscedasticity?

- The residual vs fitted values plot can be looked at to check for homoscedasticity. In the case of heteroscedasticity, the residuals can form an arrow shape or any other non-symmetrical shape.
- The goldfeldquandt test can also be used. If we get a p-value > 0.05 we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.
    - Null hypothesis: Residuals are homoscedastic
    - Alternate hypothesis: Residuals have heteroscedasticity

How to fix if this assumption is not followed?

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

Comment:

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

greatlearning
Learning for Life

5.Model performance evaluation

5.1. Evaluate the model on different performance metrics:

Prediction:

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

	Actual	Predicted
983	0.43	0.421975
194	0.51	0.527895
314	0.48	0.417427
429	0.41	0.465917
267	0.41	0.506276
746	0.68	0.666523
186	0.62	0.600098
964	0.48	0.494030
676	0.42	0.476133
320	0.58	0.551945

- 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:                       0.753
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     190.3
Date:                Thu, 12 Dec 2024   Prob (F-statistic):          2.98e-200
Time:                        01:41:13   Log-Likelihood:                 1064.6
No. Observations:                 700   AIC:                            -2105.
Df Residuals:                     688   BIC:                            -2051.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--
const                     0.2122      0.010     22.103      0.000       0.193
0.231
visitors                  0.0729      0.006     11.384      0.000       0.060
0.085
major_sports_event       -0.0609      0.004    -14.496      0.000      -0.069      -
0.053
views_trailer             0.0023   5.95e-05     39.195      0.000       0.002
0.002
dayofweek_Monday          0.0307      0.013      2.416      0.016       0.006
0.056
dayofweek_Saturday        0.0582      0.008      7.573      0.000       0.043
0.073
dayofweek_Sunday          0.0415      0.008      4.972      0.000       0.025
0.058
dayofweek_Thursday        0.0150      0.007      2.063      0.039       0.001
0.029
dayofweek_Wednesday       0.0462      0.005      9.663      0.000       0.037
0.056
season_Spring             0.0216      0.006      3.754      0.000       0.010
0.033
season_Summer             0.0430      0.006      7.424      0.000       0.032
0.054
season_Winter             0.0301      0.006      5.275      0.000       0.019
0.041
==============================================================================
Omnibus:                        1.735   Durbin-Watson:                   2.069
Prob(Omnibus):                  0.420   Jarque-Bera (JB):                1.817
Skew:                           0.109   Prob(JB):                        0.403
Kurtosis:                       2.879   Cond. No.                         486.
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.052876	0.042822	0.752621	0.7483	9.641289

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.052059	0.042188	0.752817	0.742481	9.384755

- The model is able to explain ~75% of the variation in the data
- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from over fitting
- The MAPE on the test set suggests we can predict within 9.38% of the views content
- Hence, we can conclude the model *olsmodel_final* is good for prediction as well as inference purposes

6.Actionable insights and recommendations:

6.1. Comments on significance of predictors

1. The model is able to explain ~75% of the variation in the data and within 9.38% of the views content on the test data, which is good
    - This indicates that the model is good for prediction as well as inference purposes
2. If the visitors of view content increases by one unit, then its view increases by 0.0729 units, all other variables held constant
3. If the major sports event increases by one unit, then its views decreases 0.069 by units, all other variables held constant
4. If the views of the trailer increases by one unit, then its views increases by 0.0023 units, all other variables held constant
5. The views on content on Saturday increases, then its view increases by 0.0582 units.
6. As the views content increase with an increase in the number of visitors, the company can improve its marketing activities to promote their content
7. As the views content increase with a decrease in major sports event.

6.2. Key takeaways for the business:

- Streamist can look to increase the number of content under the Drama , Horror ,Sci-fi genres as they are the most watched on the platform.
- Streamist can gather data about their users like age, gender, geographical location, occupation, etc. to better understand the kind of web series and movies different users like.
