PREDICTIVE MODELING ASSIGNMENT REPORT

Krishnan CS

GREAT LEARNING

Contents

1 Problem Statement

1.1 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.

1.2 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.

1.3 Data Description

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

- 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

The views_content column would be the dependent variable. The rest of the columns would be treated as independent variables when creating the linear regression model for analysis.

2 Exploratory Data Analysis

2.1 Problem Definition, Questions to be answered

We have been provided with data related to viewership, on the OTT platform owned by ShowTime. The driving factors for first-day content viewership needs to be determined so that the viewership of the content on the platform could be improved. This analysis would also be used to answer the following questions:

1. What does the distribution of content views look like?
2. What does the distribution of genres look like?
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?
4. How does the viewership vary with the season of release?
5. What is the correlation between trailer views and content views?

2.2 Data background and contents

A quick overview of the data suggests that there are 1000 rows in the dataset with no null values. There are 4 numeric columns – *visitors*, *ad_impressions*, *views_trailer* and *views_content*. The *major_sports_event* column is a Yes/No field. A summary of the numerical columns is shown in Table 1.

	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

Figure 1 Summary of numerical columns

There are also 3 other columns, which are categorical in nature.

There are no null values in any of the columns. There are no duplicates either.

Genre: This has 8 possible values – Comedy, thriller, drama, romance, sci-fi, horror, action, and others. 255 rows are classified as others, while the counts of the other 7 are in between 101 and 114.

Season: This has 4 values – winter, fall, spring, and summer. The count of the categories ranges from 244 to 257.

Day of week: All 7 days of week are shown. Friday has the maximum frequency at 369 and Tuesday has only 23.

A more detailed study of the variables will be performed in the Univariate analysis.

2.3 Univariate analysis

Let us quickly check the categorial columns first before proceeding on to the numerical columns.

Season: There are 4 seasons, winter, fall, spring, and summer. There are more rows for winter 257, compared to summer 244. The number of releases does not vary a lot between seasons. More analysis is required to understand if seasonality plays an important role in viewership.

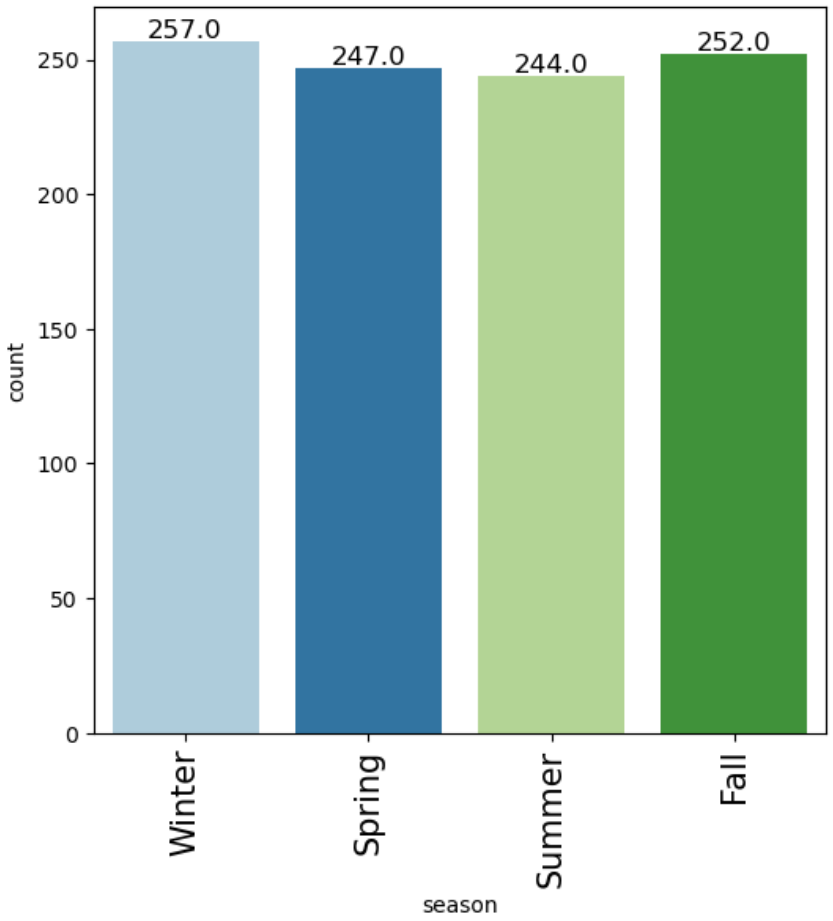

Figure 2 Count plot for season

Days of week:

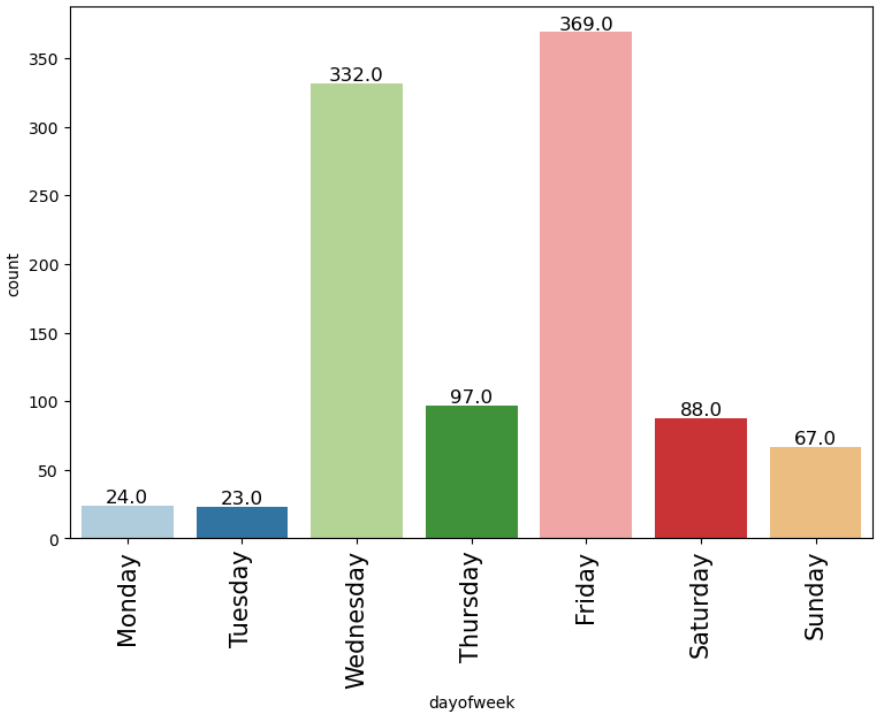

Figure 3 Count plot for days of week

The frequency is maximum for Fridays at 369, closely followed by Wednesdays at 332. Mondays and Tuesdays have a very low count at 24 and 23 respectively. A lot of the content is released on Fridays and Wednesdays and very little on Mondays and Tuesdays. A more detailed study is required to check if this is the correct approach.

Genre:

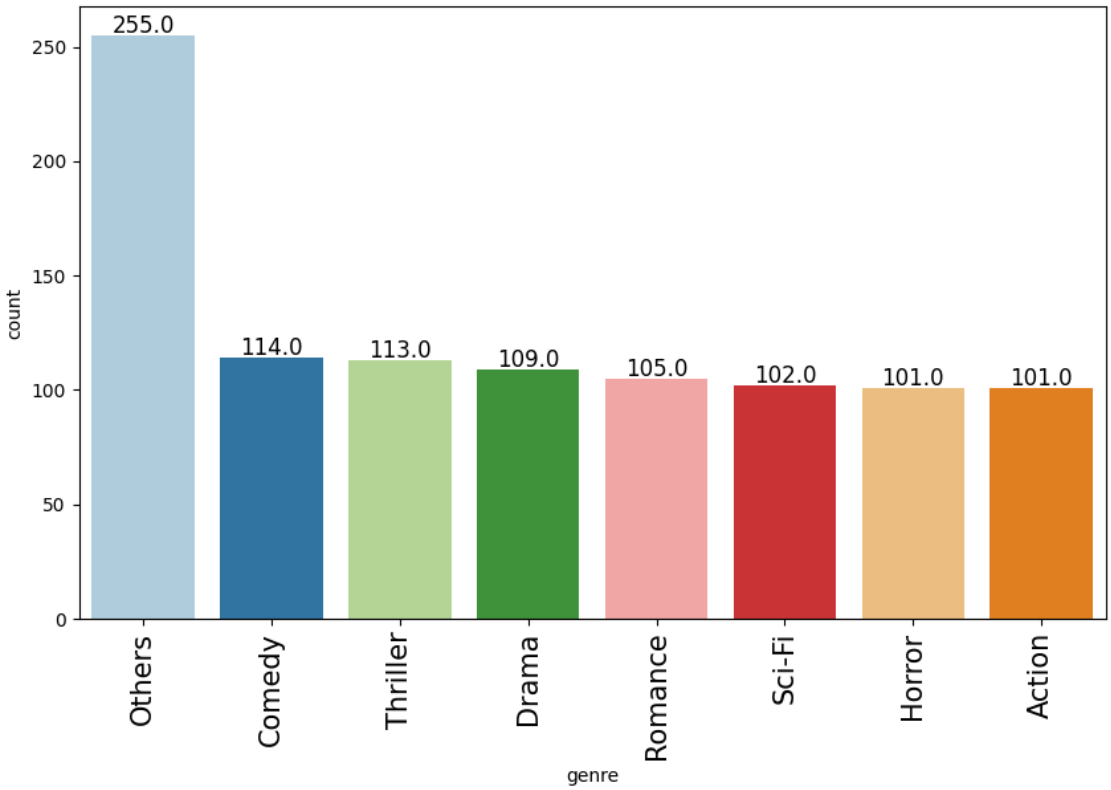

Figure 4 Count plot for genre

All the genres, except Others, are evenly distributed. From just the count of the different genres, it is hard to come to any conclusion. A bivariate analysis is required to understand which genres impact viewership.

Major_sports_event:

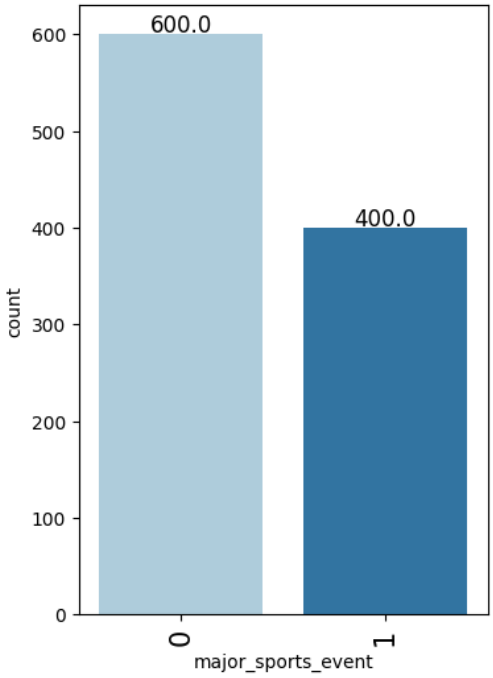

Figure 5 Count plot for major sports event

It is very much possible that a major sports event affects viewership. On 600 occasions, there were no major sporting events on the day of release of new content and on 400 occasions, the new content clashed with a major sporting event. The correlation needs to be checked to confirm if a major sporting event reduces viewership.

Visitors:

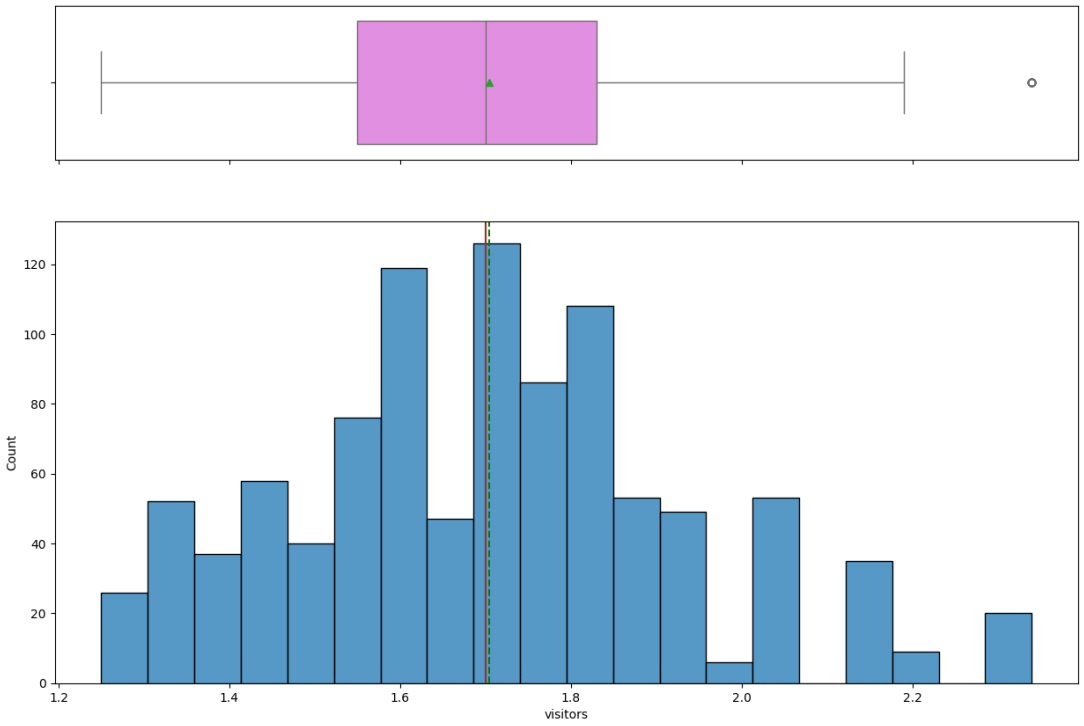

Figure 6 Histogram and box plot for visitors

The distribution of number of visitors to the platform in the past week ranges from 1.25 to 2.34 million. The mean and the median are very close to each other, which indicates low skewness. The box plot indicates that the values beyond 2.2 million are outliers. However, they look like genuine values and will not be treated.

Ad_impressions:

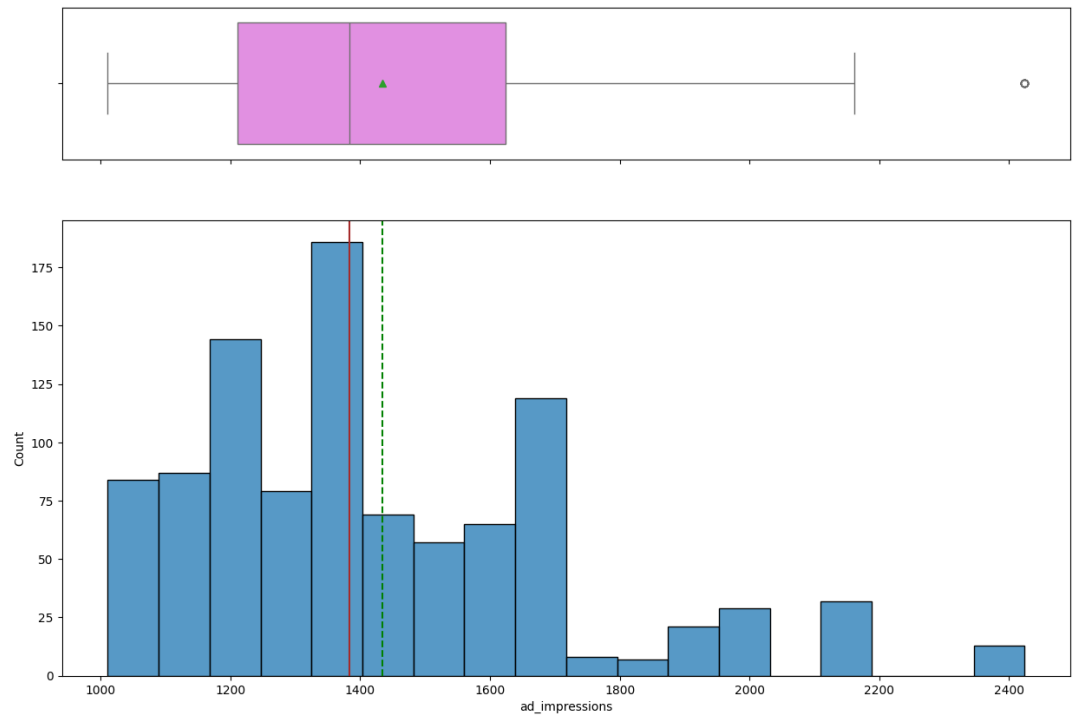

Figure 7 Histogram and box plot for ad impressions

The distribution of number of ad impressions ranges from 1010 to 2424 million. The mean is about 50 million greater than the median. The data is positively skewed and there are a few large values, which has influenced the mean value. The box plot indicates that the values beyond 2200 million are outliers. The outlier values are not unreasonable and they will not be treated.

Views_trailer:

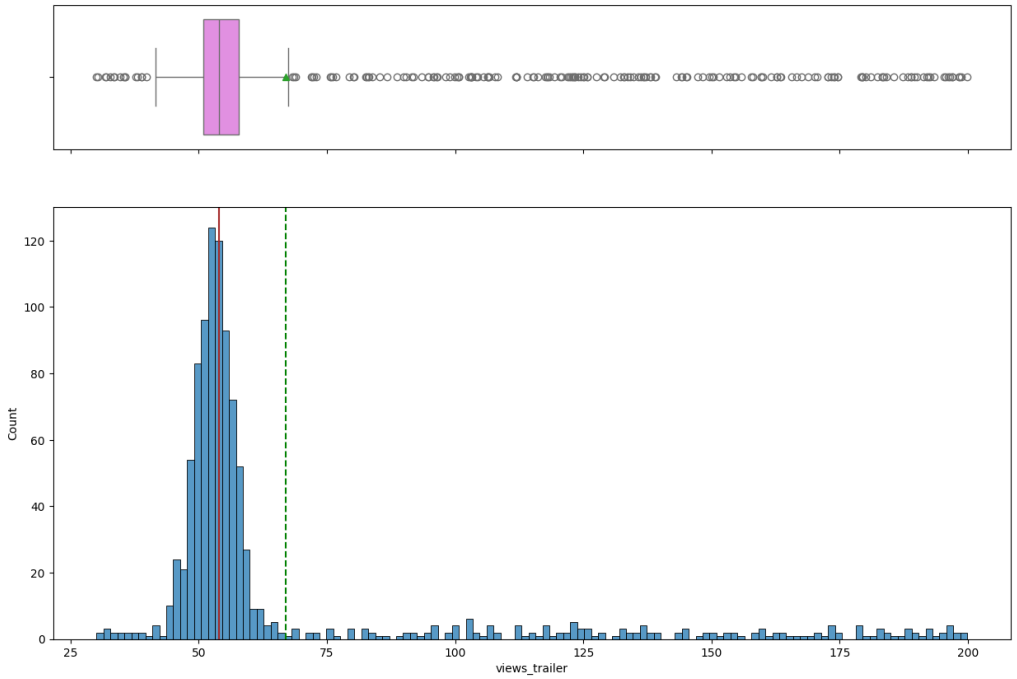

Figure 8 Histogram and box plot for trailer views

The number of trailer views ranges from 30 to 200 million. The mean is roughly 13 million more than the median. Both the histogram and the box plot suggest that the distribution is right skewed. 50 % of the data lies between 50.94 and 57.75 million and there are a lot of outliers on either side of it. This fluctuation is possible. The outliers will be retained as they are, in the analysis.

Views_content:

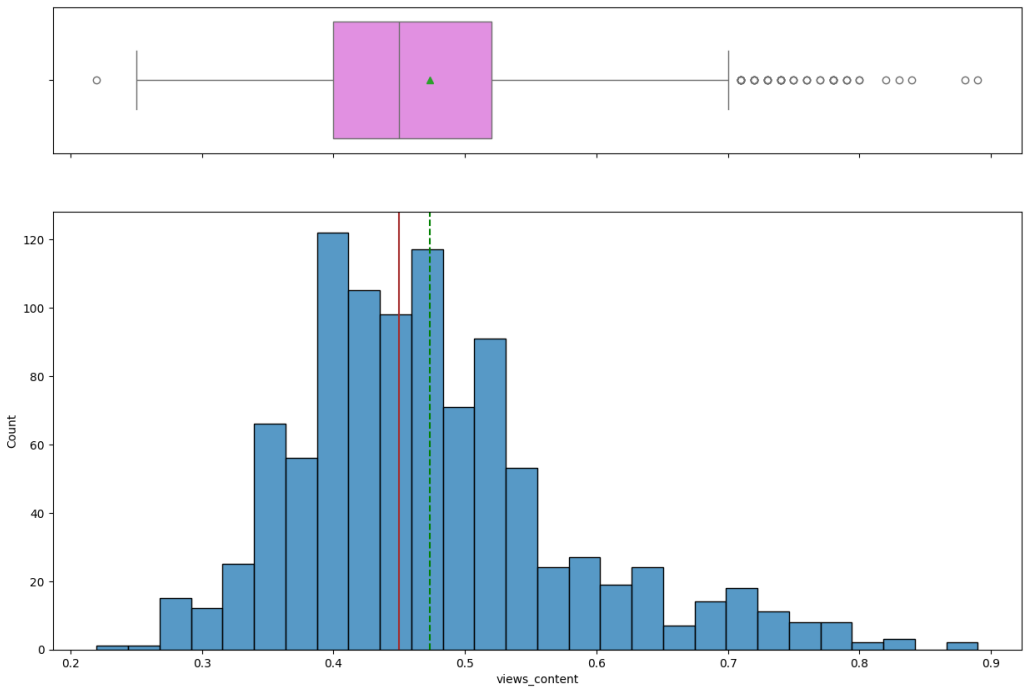

Figure 9 Histogram and box plot for content views

The number of content views ranges from 0.22 to 0.89 million. These numbers are much lower than the trailer views. The mean is roughly 0.02 million more than the median, suggesting a slight right skew. 50 % of the data lies between 0.4 and 0.52 million and there are a few outliers on either side of it. The viewership can be high or low depending on the content. The outliers will not be treated for this column as well.

2.4 Bivariate Analysis

Pair plot for numerical columns:

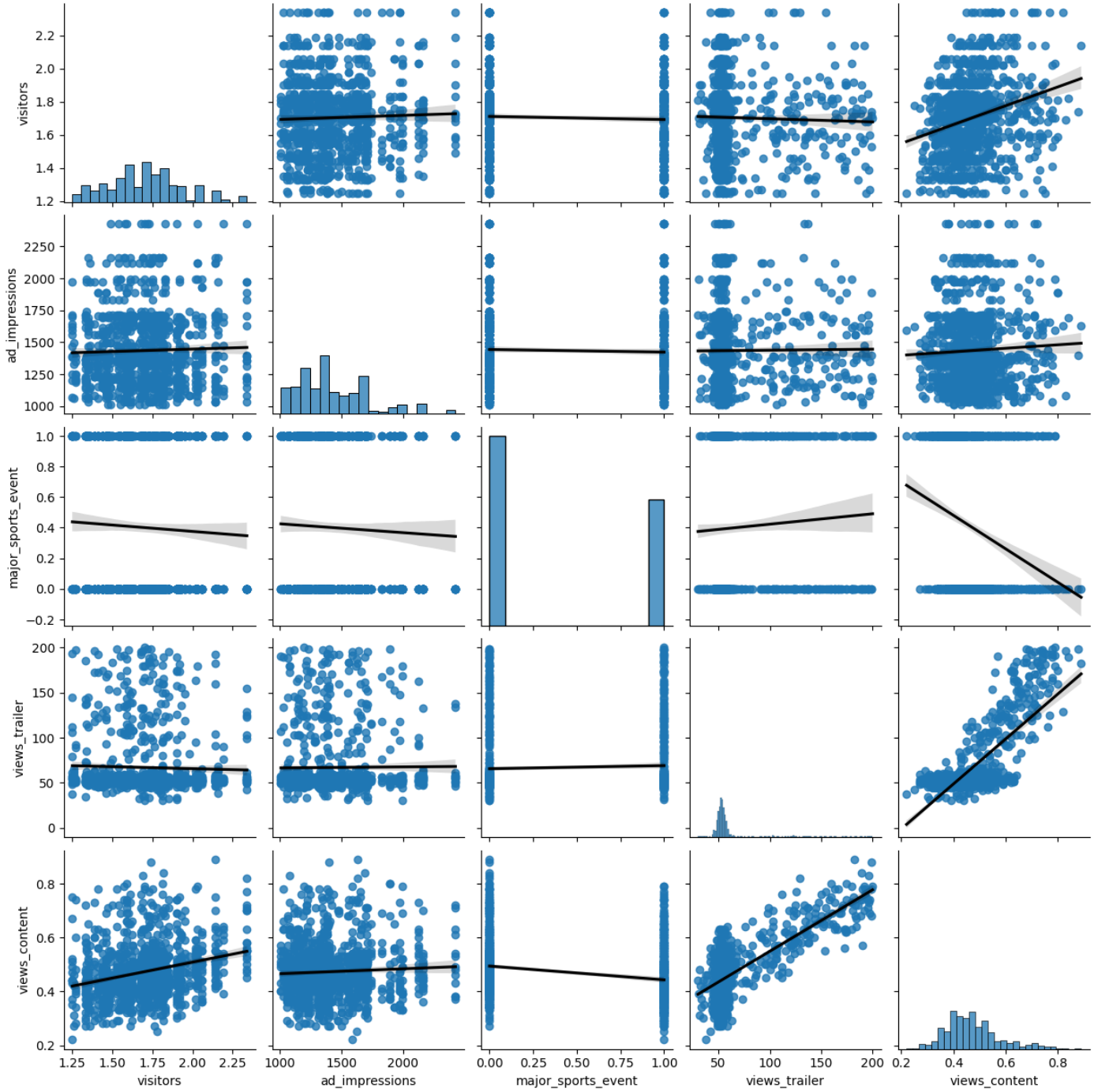

Figure 10 Scatter plot for all numerical columns

The plot in Figure 10 shows the relationship between the various numerical columns in the data. Our focus is to study the interaction between *views_content* and the other variables. The last row of plots in the above figure gives us a quick summary.

- The *views_content* seems to increase along with *visitors* and *views_trailer*.
- The number of *ad_impressions* does not seem to have much of an effect on *views_content*.
- Finally, a *major_sports_event* seems to decrease the number of *views_content*.
- A correlation heat map can be plotted to confirm this behaviour.

Correlation between numerical columns:

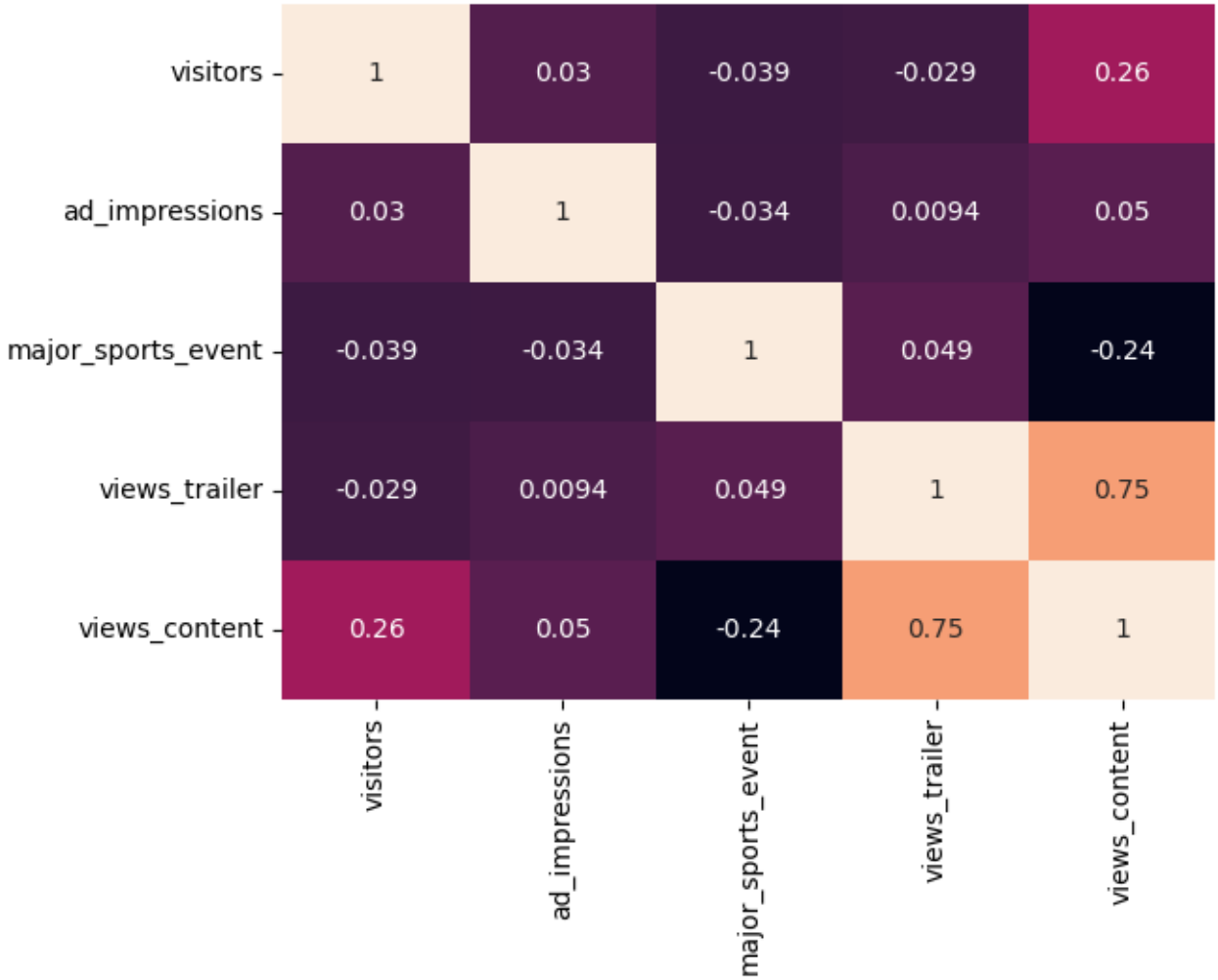

Figure 11 Correlation heat map for numerical columns

The correlation heat map indicates how strong the relationships between the variables are. Since *views_content* is the most important parameter, the correlation to that column will be described.

- There is a strong positive correlation between *views_content* and *views_trailer*. There is a good chance that the viewers who watch the trailer also watch the content.
- There is a weak positive correlation between *views_content* and *visitors*. However, there is very little correlation between *visitors* and *ad_impressions* or between *visitors* and *views_trailer*. It is more likely that the new content increases the number of visitors, rather than the other way around.
- There is a weak negative correlation between a *major_sports_event* and *views_content*. The presence of a *major_sports_event* could keep users away from the content.
- The number of *ad_impressions* does not seem to have much of an effect on *views_content*. This is similar to what was observed from the scatter plot.

Relationship between numerical and categorical variables

The focus of the analysis is to determine the driving factors for first-day viewership. In this section, the plots for *views_content* grouped by the different categorical columns, namely *season*, *dayofweek* and *genre*, will be shown. The rest of the combinations will not be checked, unless there is a requirement to go deeper into them.

Views_content by Season:

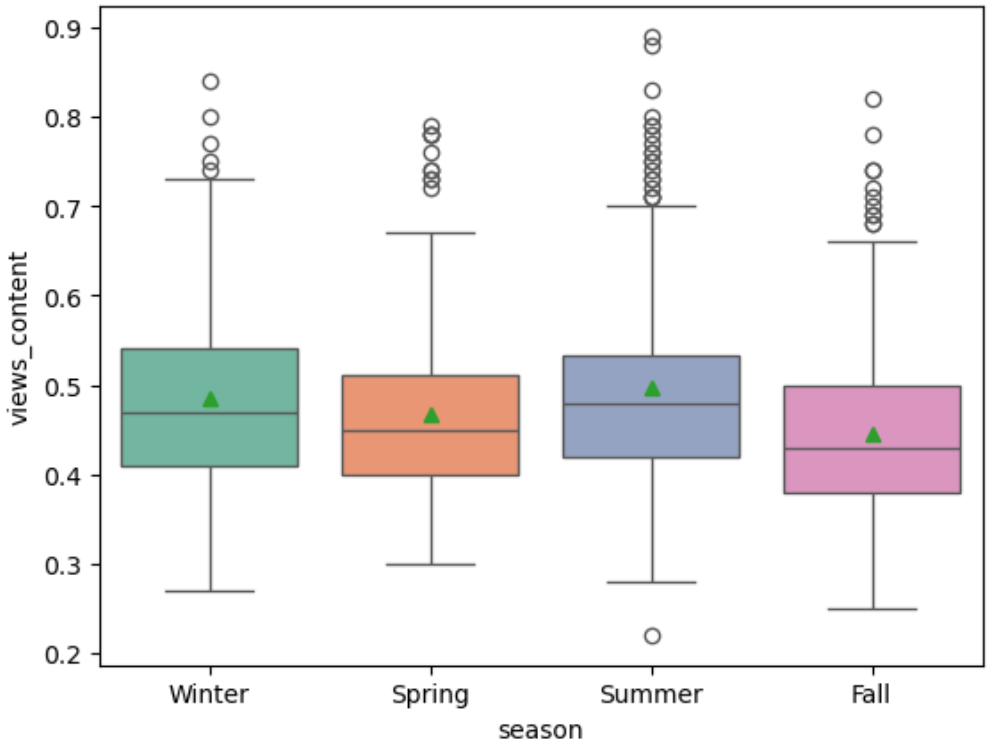

Figure 12 Variation of content views by season

The box plot for the viewership grouped by the season is shown above. The mean and median viewership count is the highest for Summer and the lowest for Fall. There are a lot of outliers above the upper whisker value. These were also noticed during the univariate analysis of *views_content*.

Views_content by Day of week:

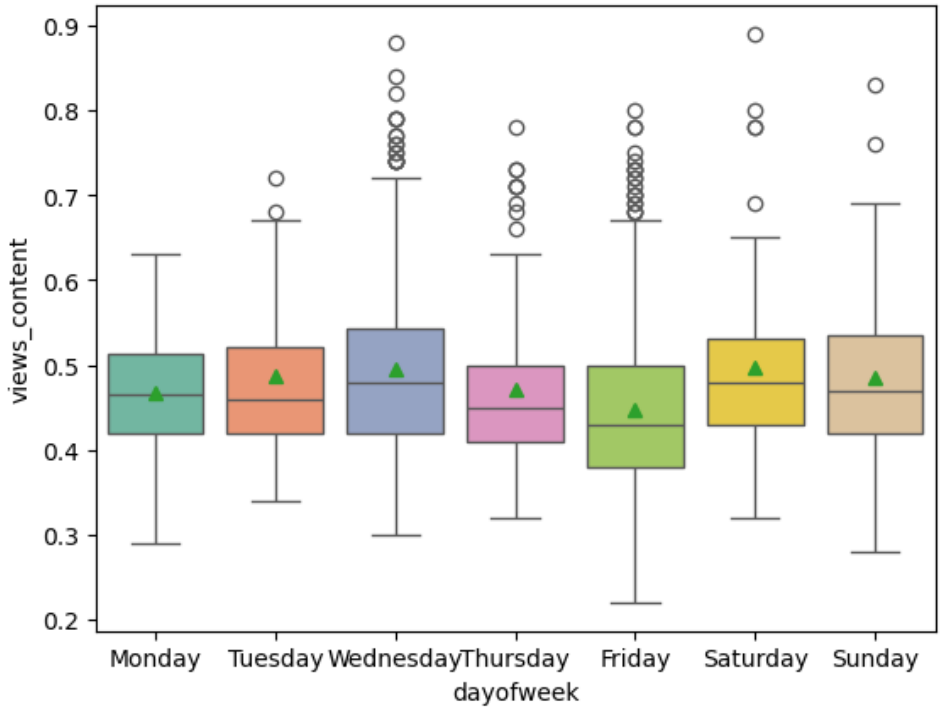

Figure 13 Variation of content views by day of week

The box plot for the viewership grouped by the day of week is shown above. The mean and median viewership count is higher for Wednesday and Saturday and lower for Friday. There are a lot of outliers above the upper whisker value. These were also noticed during the univariate analysis of *views_content*.

Views_content by genre:

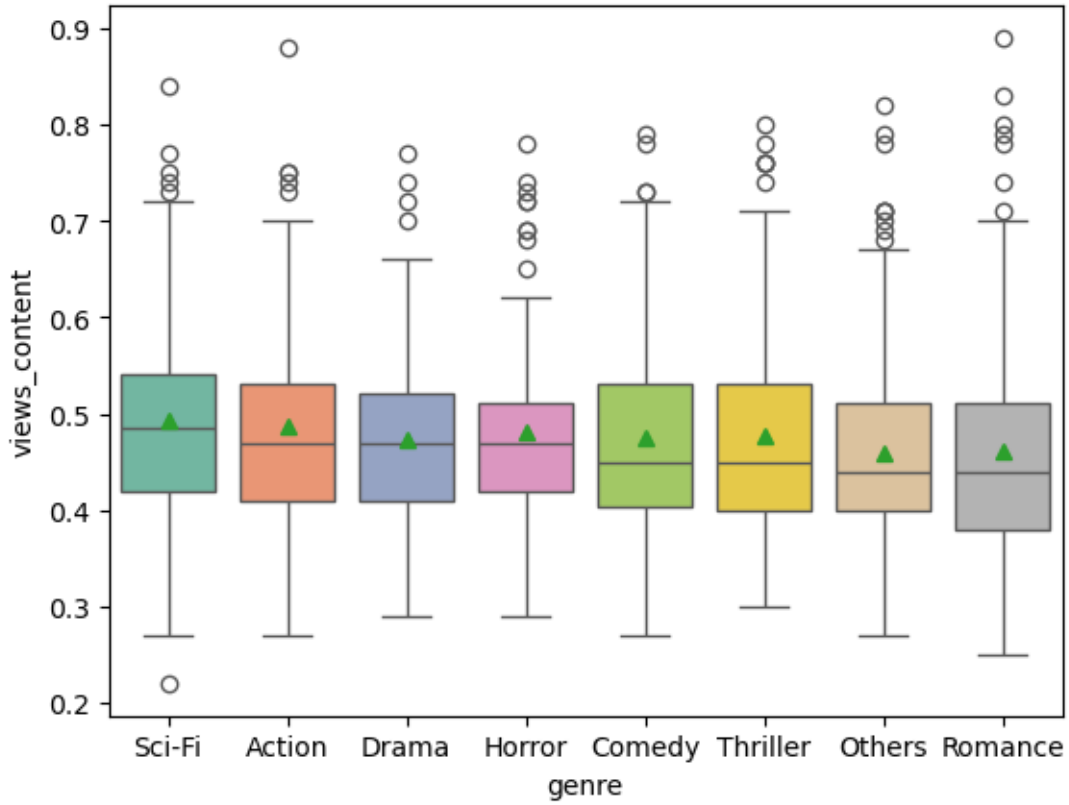

Figure 14 Views_content based on genre

The box plot for the viewership grouped by genre and sorted by median is shown above. The mean and median viewership is the highest for Sci-Fi. The mean viewership is the lowest for Others while the lowest median viewership is tied between Others and Romance. There are a lot of outliers above the upper whisker value. These were also noticed during the univariate analysis of *views_content.*

2.5 Answer to the key questions

1. What does the distribution of content views look like?
   The distribution is shown in Figure 9. The number of content views ranges from 0.22 to 0.89 million. The mean is roughly 0.02 million more than the median, suggesting a slight right skew. 50 % of the data lies between 0.4 and 0.52 million and there are a few outliers on either side of it.

2. What does the distribution of genres look like?
   The distribution is shown in Figure 4. The count for all the genres, except Others, is roughly the same. From just the count and the distribution of the different genres, it is hard to come to any conclusion.

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?
   The box plot for the viewership grouped by the *dayofweek* is shown in Figure 13. Saturdays and Wednesdays have a higher mean and median viewership. The mean and median of the viewership count on Friday is a little lower compared to all other days of the week.

dayofweek	Saturday	Wednesday	Tuesday	Sunday	Thursday	Monday	Friday
mean	0.497955	0.494608	0.487826	0.484179	0.470619	0.467917	0.446694
median	0.480000	0.480000	0.460000	0.470000	0.450000	0.465000	0.430000

Figure 15 Mean and median viewership for each day of the week

4. How does the viewership vary with the season of release?
   The box plot for the viewership grouped by the *season* is shown in Figure 12. The mean and median of the viewership count is the highest for Summer and the lowest for Fall.

season	Summer	Winter	Spring	Fall
mean	0.496803	0.484669	0.467166	0.445357
median	0.480000	0.470000	0.450000	0.430000

Figure 16 Mean and median viewership for each season

5. What is the correlation between trailer views and content views?
   There is a strong positive correlation between *views_content* and *views_trailer*. There is a good chance that the viewers who watch the trailer also watch the content. The scatter plot is shown in Figure below.

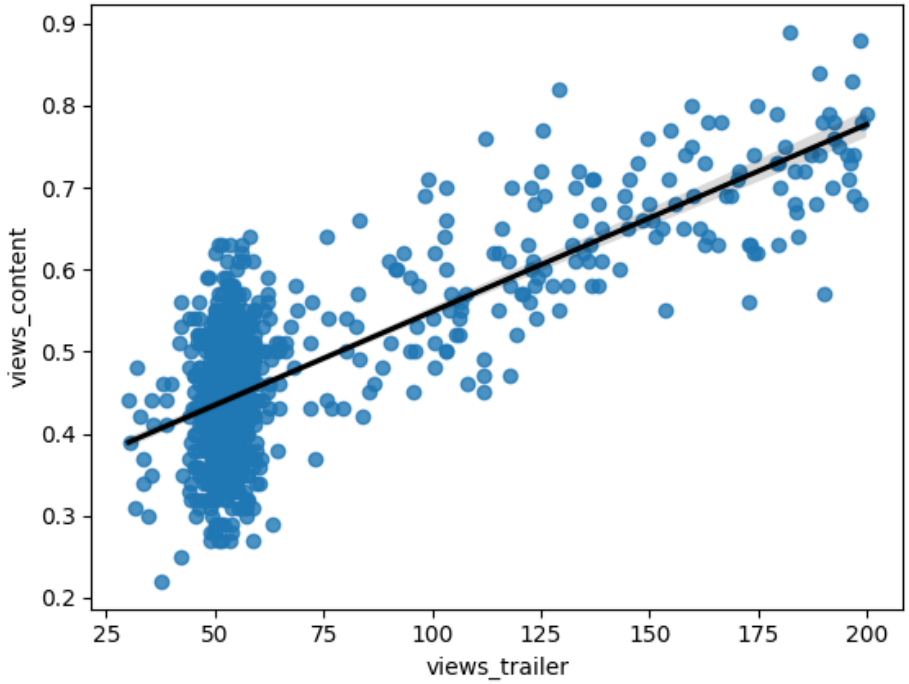

Figure 17 Scatter plot between trailer views and content views

2.6 Insights from EDA

- After getting an overview of the data, univariate and bivariate analysis of the numerical and categorical columns have been performed.
- There are a lot of outliers in the *visitors*, *ad_impressions*, *views_trailer* and *views_content*. Linear regression is very sensitive to outliers. It would be interesting to observe how the model shapes up.
- ShowTime releases a lot of content on Wednesdays and Fridays and very few content on Mondays and Tuesdays. Analysis is required to understand if this strategy is justified.
- *Views_content* has a strong positive correlation with *views_trailer*, a weak positive correlation with *visitors*. *Ad_impressions* seems to have very little impact. *Major_sports_event* has a weak negative correlation. The linear regression model is expected to capture these effects.
- The effects of other factors like *genre* and *season* on *views_content* will also be studied.

3 Data Preprocessing

3.1 Duplicate value check

There are no duplicate values in the data.

3.2 Missing value treatment

There are no missing values in the data.

3.3 Outlier treatment

All the values in all the columns look reasonable. Outlier treatment is not necessary.

3.4 Feature Engineering

There are 3 categorical columns – *season, dayofweek* and *genre*. To perform a linear regression, they must be converted using dummy variables and one hot encoding. This process creates a new column for each value in that category. These values are called factors. In linear regression analysis, one of the factors for each feature will be dropped. This is important to ensure that there is no linear dependency between the columns.

The factor that is dropped from each category would form the baseline for determining the effect of the category. There is no rule regarding which factor should be dropped. It is mainly a question of interpreting the results and how significant the baseline is.

In this scenario, let us drop '**Others'** from *genre*, since it is not a proper genre. In *dayofweek*, **Wednesday** could be dropped and set as the baseline since both the count and the mean *view_count* is high on that day. For *season*, **Winter** will be used as the baseline since it has the maximum number of new releases with a good mean viewership.

3.5 Data preparation for modelling

In this step, the dependent variable column, *views_content*, is separated out from the dataset. This will be called '**y**'. A second dataset is created with all the features except *views_content*. This will be represented as '**X**'. In the X dataset, a new constant column is introduced and the whole dataset is converted to float. Both these datasets will be further split into training dataset and test dataset.

The training dataset will comprise of roughly 70% of the observations. This will be used for model building. The remaining 30% of the data will be classified as the test dataset. Test data is not seen by the model and it will be used to evaluate how good our model is.

4 Model Building

4.1 Linear Regression

The linear regression is performed using the OLS (Ordinary Least Squares) method. The X and y training datasets will be used as input for this process. The model is built using the results of linear regression.

Arriving at the final model is an iterative process. After getting the summary of the values, two factors need to be continually evaluated. The first is the variance inflation factor to ensure that there is no multicollinearity.

Table 1 Variance Inflation Factor for the initial model

Variable	Variance inflation factor
const	93.841
Visitors	1.028
ad_impressions	1.029
major_sports_event	1.066
views_trailer	1.024
season_Fall	1.542
season_Spring	1.536
season_Summer	1.534
dayofweek_Friday	1.328
dayofweek_Monday	1.063
dayofweek_Saturday	1.164
dayofweek_Sunday	1.153
dayofweek_Thursday	1.183
dayofweek_Tuesday	1.062
genre_Action	1.340
genre_Comedy	1.349
genre_Drama	1.358
genre_Horror	1.331
genre_Romance	1.291
genre_Sci-Fi	1.336
genre_Thriller	1.341

In the table showing variance inflation factor, all the values are less than 2. This indicates that there is no multicollinearity between the variables.

In the next stage of model building, the values from the **P > |t|** column of the summary table are considered. Here, the level of significance has been set as 0.05. If the p-value for any of the variable is greater than the level of significance, a new model will be built without the variable. If there is no big drop in R-squared and Adjusted R-squared values, that variable will not be part of the final model. This process is carried out until all the p-values are less than 0.05.

The statistics from the model summary table is shown in the next section.

4.2 Model statistics

The model was built using linear regression and the summary table is shown below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          views_content   R-squared:                       0.788
Model:                            OLS   Adj. R-squared:                  0.785
Method:                 Least Squares   F-statistic:                     320.8
Date:                Wed, 26 Feb 2025   Prob (F-statistic):          7.75e-227
Time:                        10:56:33   Log-Likelihood:                 1118.4
No. Observations:                 700   AIC:                            -2219.
Df Residuals:                     691   BIC:                            -2178.
Df Model:                           8
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      0.1444      0.015      9.951      0.000       0.116       0.173
visitors                   0.1281      0.008     16.372      0.000       0.113       0.143
major_sports_event        -0.0603      0.004    -15.569      0.000      -0.068      -0.053
views_trailer              0.0023   5.48e-05     42.477      0.000       0.002       0.002
season_Fall               -0.0254      0.005     -5.553      0.000      -0.034      -0.016
season_Summer              0.0191      0.005      4.127      0.000       0.010       0.028
dayofweek_Friday          -0.0441      0.004    -10.533      0.000      -0.052      -0.036
dayofweek_Saturday         0.0139      0.007      1.990      0.047       0.000       0.028
dayofweek_Thursday        -0.0272      0.007     -4.129      0.000      -0.040      -0.014
==============================================================================
Omnibus:                        2.179   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.336   Jarque-Bera (JB):                1.999
Skew:                           0.110   Prob(JB):                        0.368
Kurtosis:                       3.142   Cond. No.                         658.
==============================================================================

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

Figure 18 Summary of the final model

For this model, the R-squared value is 0.788. It means that the model explains 78.8% of the variance in the training set.

The coefficients tell us how one unit change in X can affect y and the sign of the coefficient indicates if the relationship is positive or negative. For example, if visitors increase by 1 unit, rating increases by 0.1281 unit, provided all other features are kept constant.

For categorical variables like the *dayofweek* and *season*, the value represents the difference in the effect from the baseline. For example, a value of 0.0191 for season_Summer indicates that the views_content for Summer will be 0.0191 units higher that of Winter (season_Winter was set as the baseline).

(P > |t|) gives the p-value for each predictor variable to check the null hypothesis. If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant. Multicollinearity must be tested before reading into the p-values.

The std err column is the standard deviation of the coefficient. The 0.025 and 0.975 columns together represent the 95% confidence interval for predicting the true value of the coefficient. For example, we could say with 95% confidence that the 'God-given true' coefficient for visitors would like between 0.113 and 0.143.

4.3 Model coefficients with column names

The linear regression model for views_content can be calculated as:

*views_content = 0.144 + 0.128*visitors - 0.06*major_sports_event + 0.002*views_trailer - 0.025*season_Fall + 0.019 *season_Summer - 0.044*dayofweek_Friday + 0.014*dayofweek_Saturday - 0.027*dayofweek_Thursday*

5 Testing the assumptions of linear regression model

The following assumptions should be tested before we draw inferences regarding the model estimates.

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

5.1 Test for Linearity and Independence

Linearity describes a straight-line relationship between two variables, predictor variables must have a linear relation with the dependent variable. This is done observing the plot between the fitted values and the residuals.

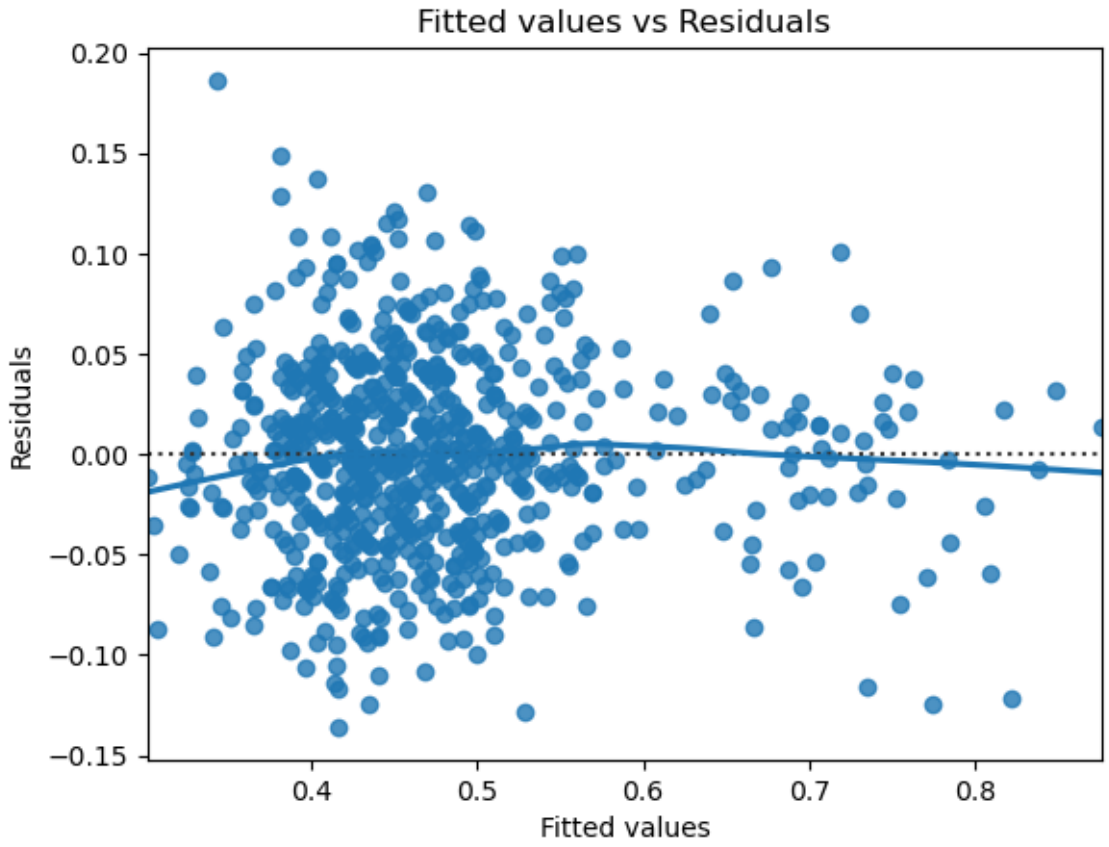

Figure 19 Fitted values vs Residuals - Test for linearity and independence

This plot does not follow any pattern. Hence, the model is linear and the residuals are independent.

5.2 Test for normality of error terms

If the error terms are not normally distributed, confidence intervals may become too wide or narrow. Once the confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. There are three tests that can be performed to ensure that the residuals are normally distributed.

5.2.1 Normality Test 1: Plot the histogram

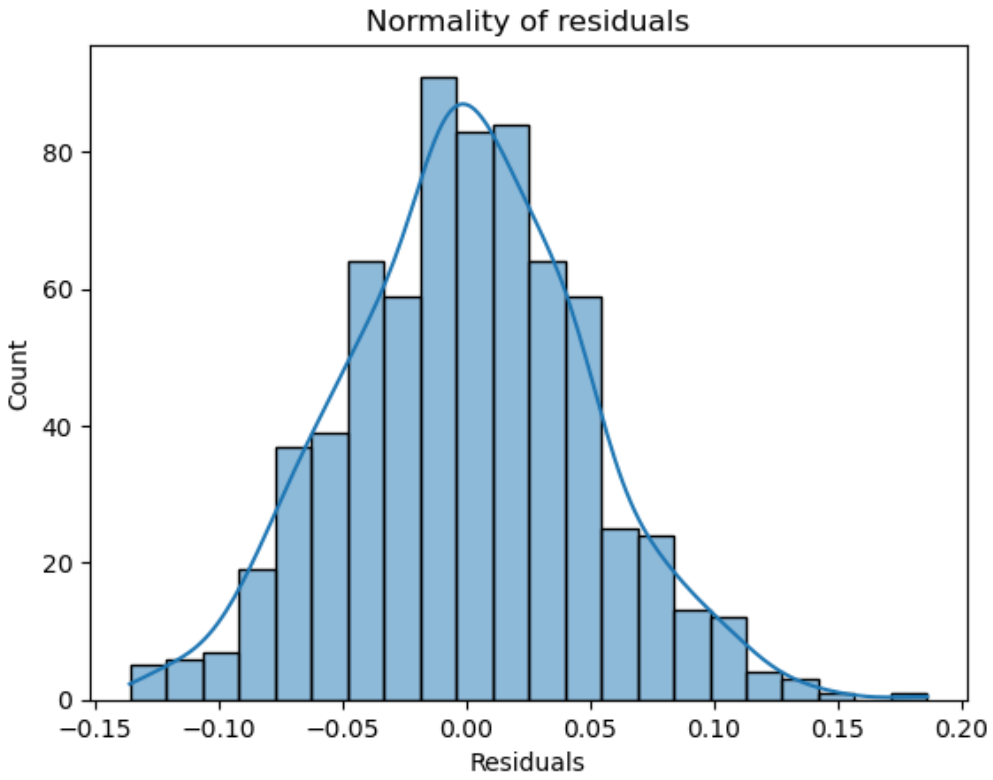

Figure 20 Histogram of residuals - Test for normality

On visual inspection, the values appear to be normally distributed.

5.2.2 Normality Test 2: Q-Q plot of residuals

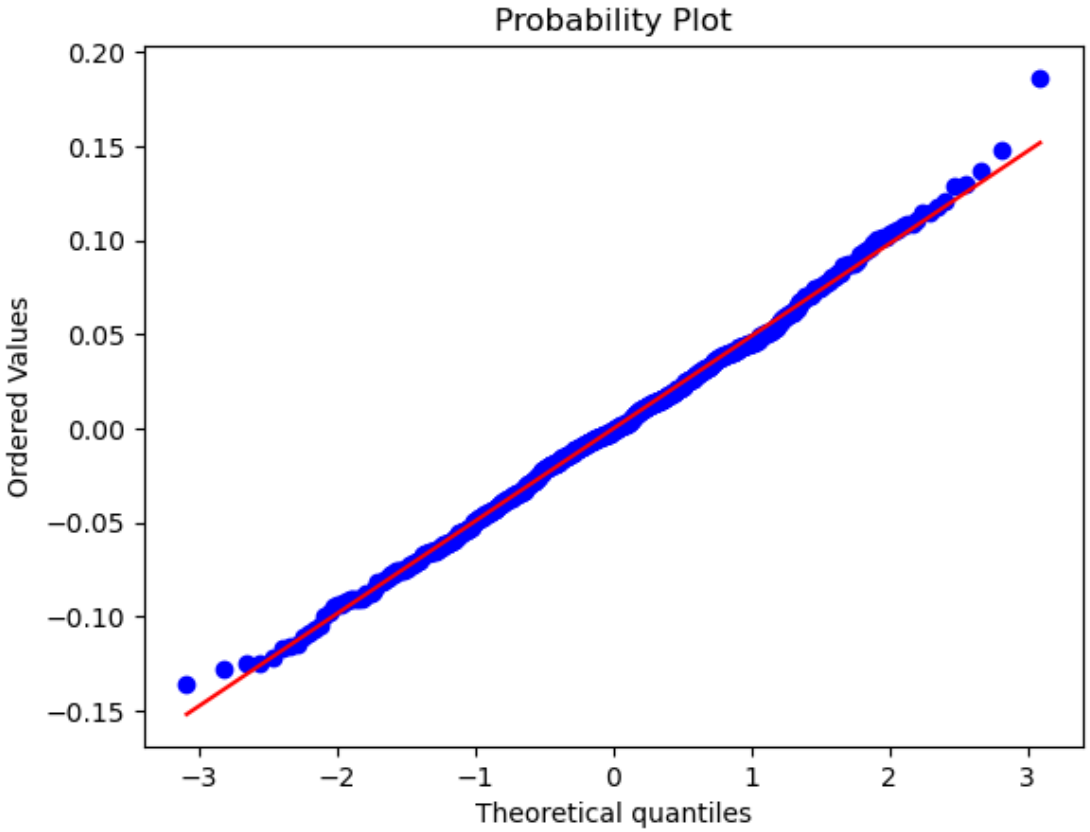

Figure 21 Q-Q plot to test normality

Most of the points like on the straight line. So, the residuals are close to a normal distribution.

5.2.3 Normality Test 3: Shapiro-Wilk test

This is a statistical test, available in python statsmodels. This produced a value of 0.391, which is greater than the level of significance of 0.05. This also confirms that the distribution is normal.

5.3 Test for homoscedasticity

If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic. Otherwise, the data is heteroscedastic.

Apart from checking the plot, there is also a statistical test called the goldfeldquandt test. This produced a value of 0.162, which is greater than the level of significance of 0.05. This confirms that the data is homoscedastic.

6 Model performance evaluation

The model performance will be evaluated by predicting the values on the test dataset. The root mean square error (RMSE), the mean absolute error (MAE) and the R2_Score can be computed on the training and test datasets.

Table 2 Model performance evaluation metrics

	Training data	Test data
RMSE	0.049	0.0511
MAE	0.0385	0.0413
R2_Score	0.7879	0.7621

The RMSE and MAE values are comparable for train and test data, indicating that the model is not overfitting.

7 Actionable Insights & Recommendations

Our final model is

*views_content = 0.144 + 0.128*visitors – 0.06*major_sports_event + 0.002*views_trailer – 0.025*season_Fall + 0.019*season_Summer – 0.044*dayofweek_Friday + 0.014*dayofweek_Saturday – 0.027*dayofweek_Thursday*

7.1 Significance of predictors

- R-squared of the model is 0.788 and adjusted R-squared is 0.785, which shows that the model can explain ~79% variance in the data. This is quite good.
- 1 unit increase in *visitors* will increase the views_content by 0.128 units, when all the other variables are held constant.
- 1 unit increase in *views_trailer* will increase the views_content by 0.002 units, when all the other variables are held constant.
- 1 unit increase in *major_sports_event* will decrease the views_content by 0.06 units, when all the other variables are held constant.
- The *views_content* will be 0.025 units lower in Fall compared to Winter. It will be 0.019 units higher in Summer compared to Winter, when all other variables are held constant.
- The *views_content* will be 0.044 units lower on Fridays compared to Wednesdays. It will be 0.014 units higher on Thursday compared to Wednesdays and 0.027 units lower on Thursdays, when all other variables are held constant.

7.2 Key takeaways for the business

- Ad impressions do not have much of an impact on the first day content views. It would be better to focus on trailers and visitors as they both have a positive impact on *views_content*.
- The values of 0.128 units change for visitors and 0.002 units change for trailers give the impression that visitors are more important than trailers. However, this is not the case. The mean values for visitors and trailers are 1.7 and 66.9. Considering the scale, trailers will also have an equally significant effect on the content views compared to visitors.
- The presence of a major sports event reduces the first day content views. It would be best to avoid releasing content on such days.
- The genre of the content does not play much of a role in the first day content views.
- For the seasons, Winter was chosen as the baseline. Compared to this baseline, Summer produced better content views, Fall had worse and Spring was about the same. ShowTime could release more content in Summer and less in Fall.
- Among the days of the week, Wednesday was set at the baseline. Thursday and Friday produced worse results than the baseline, while Saturday had better results. More new content could be planned on Saturdays and less on Thursdays and Fridays.