PM CODED PROJECT

Business Report

DSBA

Submitted By: Maheep Singh
Batch : PGP-DSBA (PGPDSBA.O.AUG24.A)

Table of Contents

List of Figures

List of Tables

Business Context & Data Dictionary

Context

An over-the-top (OTT) media service is a media service offered directly to viewers via the internet. The term is most synonymous with subscription-based video-on-demand services that offer access to film and television content, including existing series acquired from other producers, as well as original content produced specifically for the service. They are typically accessed via websites on personal computers, apps on smartphones and tablets, or televisions with integrated Smart TV platforms.
Presently, OTT services are at a relatively nascent stage and are widely accepted as a trending technology across the globe. With the increasing change in customers' social behaviour, which is shifting from traditional subscriptions to broadcasting services and OTT on-demand video and music subscriptions every year, OTT streaming is expected to grow at a very fast pace. The global OTT market size was valued at $121.61 billion in 2019 and is projected to reach $1,039.03 billion by 2027, growing at a CAGR of 29.4% from 2020 to 2027. The shift from television to OTT services for entertainment is driven by benefits such as on-demand services, ease of access, and access to better networks and digital connectivity.

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

Objective

ShowTime is an OTT service provider and offers a wide variety of content (movies, web shows, etc.) for its users. They want to determine the driver variables for first-day content viewership so that they can take necessary measures to improve the viewership of the content on their platform. Some of the reasons for the decline in viewership of content would be the decline in the number of people coming to the platform, decreased marketing spends, 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 analyse the data and come up with a linear regression model to determine the driving factors for first-day viewership.

Data Description

The data contains the different factors to analyse 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

Rubric Question 1: Exploratory Data Analysis

Data Overview

- **Load dataset** & display top 5 rows: -

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

Figure 1: Top 5 rows of the dataset

- There are **1000 rows & 8 columns** in the dataset
- **Checking datatypes**: -

```
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   visitors            1000 non-null   float64
 1   ad_impressions      1000 non-null   float64
 2   major_sports_event  1000 non-null   int64
 3   genre               1000 non-null   object
 4   dayofweek           1000 non-null   object
 5   season              1000 non-null   object
 6   views_trailer       1000 non-null   float64
 7   views_content       1000 non-null   float64
dtypes: float64(4), int64(1), object(3)
```

Figure 2: Datatypes in the Dataset

- ✓ We have 3 object (string/category) type, 4 float (numeric) & 1 Int datatypes in the dataset.
- ✓ Since 'major_sports_event' represents a category with value either '1' or '0', we convert it into object datatype.

```
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   visitors            1000 non-null   float64
 1   ad_impressions      1000 non-null   float64
 2   major_sports_event  1000 non-null   object
 3   genre               1000 non-null   object
 4   dayofweek           1000 non-null   object
 5   season              1000 non-null   object
 6   views_trailer       1000 non-null   float64
 7   views_content       1000 non-null   float64
```

Figure 3: Datatypes in the Dataset, post datatype-treatment

- **Statistical Summary of the dataset**: -

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
visitors	1000.0	NaN	NaN	NaN	1.70429	0.231973	1.25	1.55	1.7	1.83	2.34
ad_impressions	1000.0	NaN	NaN	NaN	1434.71229	289.534834	1010.87	1210.33	1383.58	1623.67	2424.2
major_sports_event	1000.0	2.0	0.0	600.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
genre	1000	8	Others	255	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dayofweek	1000	7	Friday	369	NaN	NaN	NaN	NaN	NaN	NaN	NaN
season	1000	4	Winter	257	NaN	NaN	NaN	NaN	NaN	NaN	NaN
views_trailer	1000.0	NaN	NaN	NaN	66.91559	35.00108	30.08	50.9475	53.96	57.755	199.92
views_content	1000.0	NaN	NaN	NaN	0.4734	0.105914	0.22	0.4	0.45	0.52	0.89

Figure 4: Statistical Summary of the Dataset

Type	Columns	Observations & Insights
Numerical	visitors	✓ Average weekly visitors range between 1.25 million to 2.34 million. ✓ Mean & Median, both, are 1.7 million. ✓ Standard Deviation is 0.23 million.
Numerical	ad_impressions	✓ Number of ad impressions across all ad campaigns for the content range between 1010.87 million to 2424.2 million. ✓ Mean is 1434.71 million and Median is 1383.58 million. ✓ Standard Deviation is 289.53 million.
Categorical	major_sports_ event	✓ Bivariate variable with 2 unique values signifying whether there was a major sporting event on the day. Data is divided between '1' & '0', representing 'Yes' & 'No' respectively. ✓ '0' is the most dominant value occurring 600/1000 data, implying no major sports event on majority of days.
Categorical	genre	✓ Multivariate variable with 8 unique values. ✓ 'Others' category is the most dominant value occurring 255/1000 in the data, implying majority of the genre is miscellaneous.
Categorical	dayofweek	✓ Multivariate variable with 7 unique values representing seven days of the week. ✓ 'Friday' is the most dominant value occurring 369/1000 in the data, implying majority of the content is released on the last weekday (just before the weekend).
Categorical	season	✓ Multivariate variable with 4 unique values representing four seasons in a year. ✓ 'Winter' is the most dominant value occurring 257/1000 in the data, implying majority of the content is released in cold weather when one prefers to stay inside.
Numerical	views_trailer	✓ No. of content trailer views range between 30.08 million to 199.92 million. ✓ Mean is 66.92 million & Median is 53.96 million. ✓ Standard Deviation is 35 million.
Numerical	views_content	✓ No. of first-day content views range between 0.22 million to 0.89 million. ✓ Mean is 0.47 million & Median is 0.45 million. ✓ Standard Deviation is 0.11 million.

Table 1: Statistical Summary – Observations

- **Check & Treat Duplicate, Missing & Error Values**: -
  - ✓ Upon checking, neither duplicate nor missing values were found. Hence, no treatment required.

```
Missing Values:-

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

 Duplicated Values:  0
```

Figure 5: Missing/Duplicate Value-check

✓ Upon checking unique values of all categorical variables, no error/mistypes were found in the data. Hence, no treatment required.

```
Index(['visitors', 'ad_impressions', 'major_sports_event', 'genre',
       'dayofweek', 'season', 'views_trailer', 'views_content'],
      dtype='object')

----------------------------------------------------------------------------------------------------
major_sports_event :-

Unique Values = [0 1]

Column Value Count major_sports_event
0    600
1    400
Name: count, dtype: int64

----------------------------------------------------------------------------------------------------
genre :-

Unique Values = ['Horror' 'Thriller' 'Sci-Fi' 'Others' 'Drama' 'Action' 'Comedy' 'Romance']

Column Value Count genre
Others      255
Comedy      114
Thriller    113
Drama       109
Romance     105
Sci-Fi      102
Horror      101
Action      101
Name: count, dtype: int64

----------------------------------------------------------------------------------------------------
dayofweek :-

Unique Values = ['Wednesday' 'Friday' 'Sunday' 'Thursday' 'Monday' 'Saturday' 'Tuesday']

Column Value Count dayofweek
Friday       369
Wednesday    332
Thursday      97
Saturday      88
Sunday        67
Monday        24
Tuesday       23
Name: count, dtype: int64

----------------------------------------------------------------------------------------------------
season :-

Unique Values = ['Spring' 'Fall' 'Winter' 'Summer']

Column Value Count season
Winter    257
Fall      252
Spring    247
Summer    244
Name: count, dtype: int64
```

Figure 6: Error-value check in the Dataset

- **Replacing values in 'major_sports_event'** – '0' to be replaced with 'No' and '1' with 'Yes', as it is a categorical variable.

```
major_sports_event :-

Unique Values = ['No' 'Yes']

Column Value Count major_sports_event
No     600
Yes    400
Name: count, dtype: int64
```

Figure 7: Feature Engineering – 'major_sports_event'

Univariate & Bivariate Analysis

- **Perform Univariate Analysis** – Use Histograms & Boxplots to analyse each numerical variable, followed by Barplots for categorical variables: -
  - Distribution of visitors: -

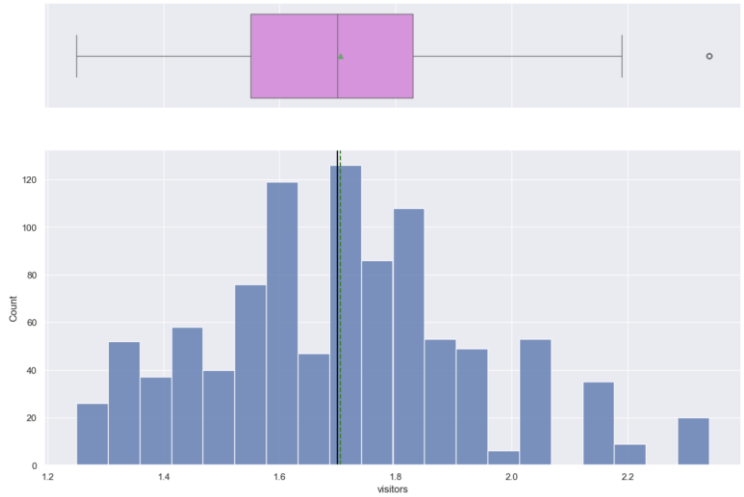


Figure 8: Univariate Analysis – visitors

✓ **Observations & Insights** can be summarized below: -
- Distribution of visitors seem to be slightly right-skewed.
- Multimodal distribution having multiple peaks.
- Few Outliers observed: -

```
Lower Wishker at 1.13 | Upper Whisker at 2.25
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 20
Total Outlier Count= 20
Outlier Percentage in visitors= 2.0 %
```

Figure 9: Outlier Inspection – visitors

- Distribution of ad_impressions: -

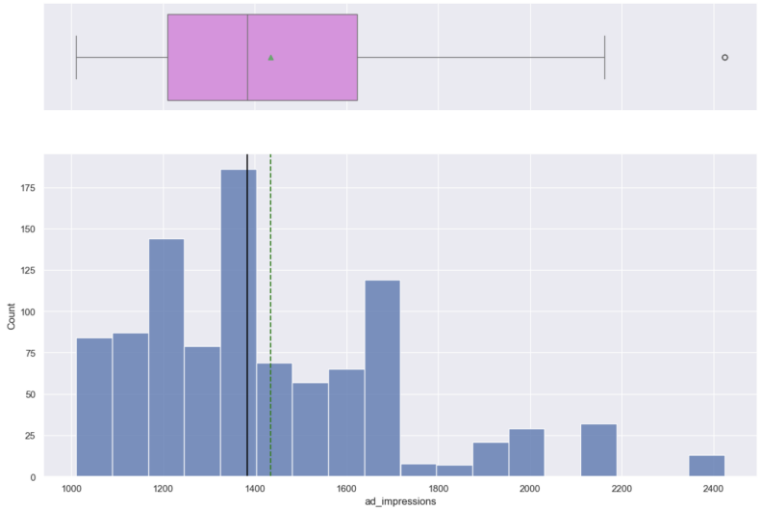


Figure 10: Univariate Analysis – ad_impressions

✓ **Observations & Insights** can be summarized below: -

- Distribution of ad_impressions seem to be right-skewed.
- Multimodal distribution having multiple peaks.
- Few Outliers observed: -

```
Lower Wishker at 590.3199999999997 | Upper Whisker at 2243.6800000000003
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 13
Total Outlier Count= 13
Outlier Percentage in ad_impressions= 1.3 %
```

Figure 11: Outlier Inspection – ad_impressions

- Distribution of views_trailer: -

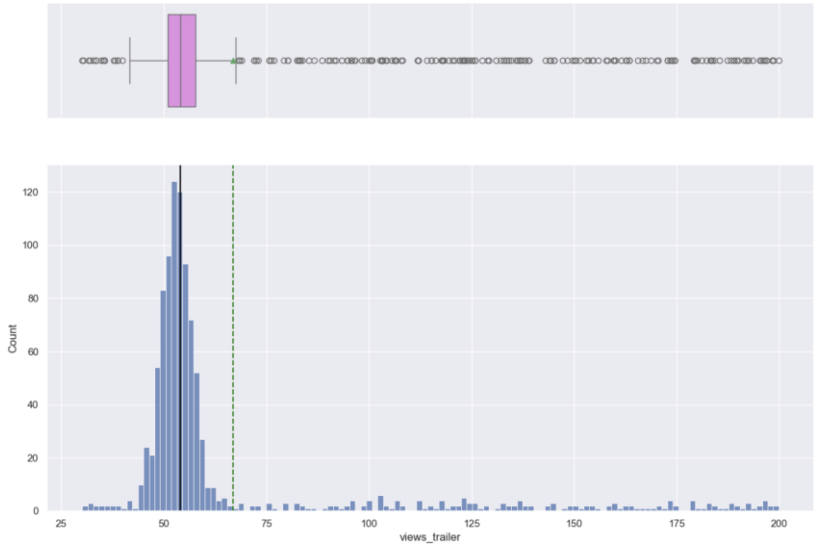


Figure 12: Univariate Analysis – views_trailer

✓ **Observations & Insights** can be summarized below: -

- Distribution of views_trailer seem to be highly right-skewed.
- Unimodal distribution having single peak.
- Outliers observed: -

```
Lower Wishker at 40.73625000000002 | Upper Whisker at 67.96624999999997
Lower Whisker Outlier Count = 16
Upper Whisker Outlier Count = 173
Total Outlier Count= 189
Outlier Percentage in views_trailer= 18.9 %
```

Figure 13: Outlier Inspection – views_trailer

- Distribution of views_content: -

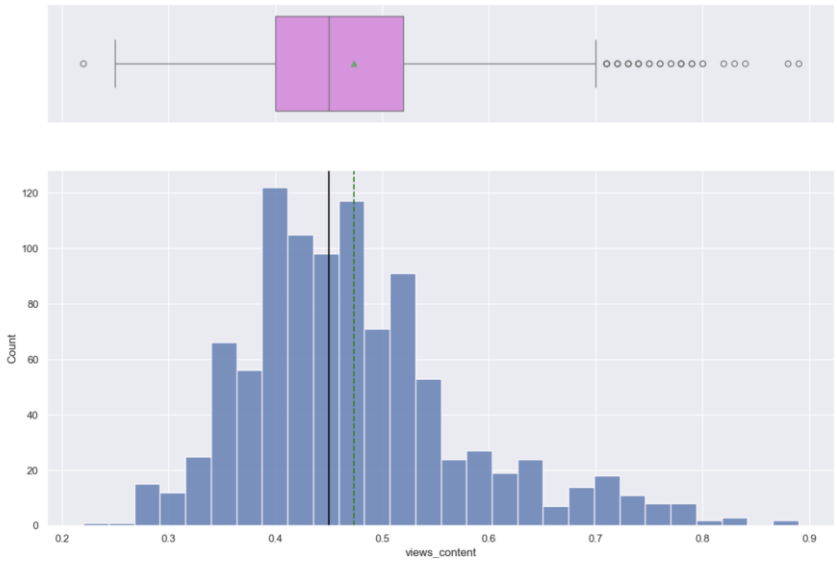


Figure 14: Univariate Analysis – views_content

✓ **Observations & Insights** can be summarized below: -

- Distribution of views_content seem to be right-skewed.
- Multimodal distribution having multi peaks.
- Outliers observed: -

```
Lower Wishker at 0.22000000000000003 | Upper Whisker at 0.7
Lower Whisker Outlier Count = 1
Upper Whisker Outlier Count = 46
Total Outlier Count= 47
Outlier Percentage in views_content= 4.7 %
```

Figure 15: Outlier Inspection – views_content

- Distribution of major_sports_event: -

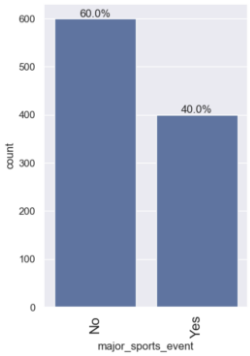


Figure 16: Univariate Analysis – major_sports_event

✓ **Observations & Insights** can be summarized below: -

- For 60% cases, there was no major sports event on the day of release/first-view of content
- For 40% cases, there was a major sports event on the day of release/first-view of content. However, whether the impact of this event on content views is statistically significant, remains to be seen (post conducting Regression Analysis).

- Distribution of genre: -

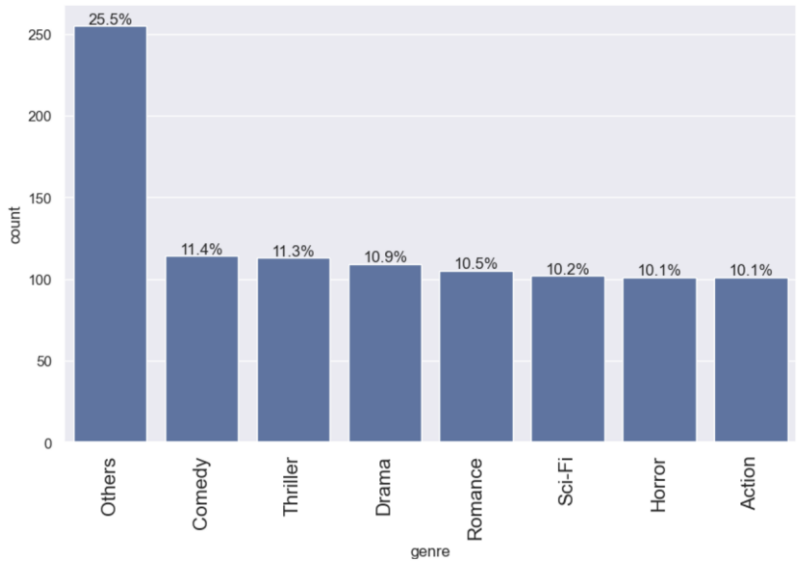


Figure 17: Univariate Analysis – genre

✓ **Observations & Insights** can be summarized below: -

- Majority of the data (25.5%) is represented by miscellaneous category, while all other genre categories equitably contribute 10-11% each.
- There is no clear trend that is being implied with genre data, whether any preference is being made to a particular category.
- Whether this variable contributes to content views can be further evaluated during Regression Analysis.

- Distribution of dayofweek: -

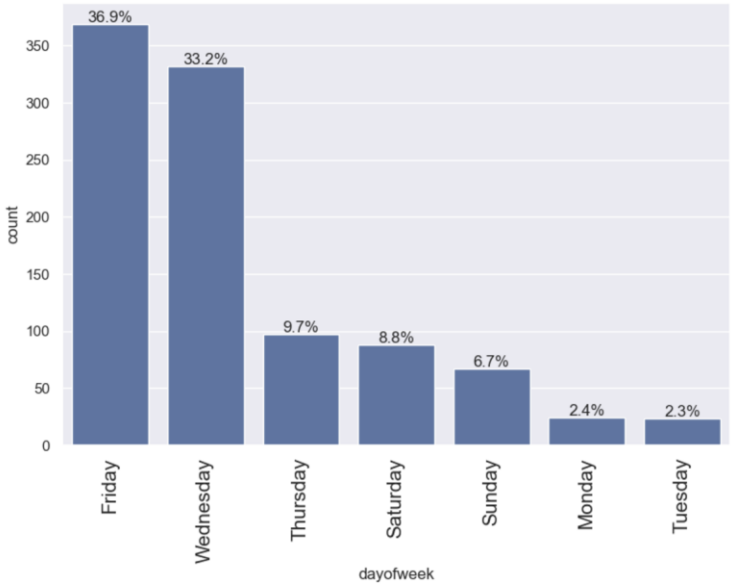


Figure 18: Univariate Analysis – dayofweek

✓ **Observations & Insights** can be summarized below: -

- Majority of the data (36.9% + 33.2%) is recorded either for a Friday or a Wednesday, followed by Thursday, Saturday & Sunday with 7-10%.
- Monday & Tuesday contribute lowest at 2-2.5%
- Whether this variable contributes to content views can be further evaluated during Regression Analysis.

- Distribution of season: -

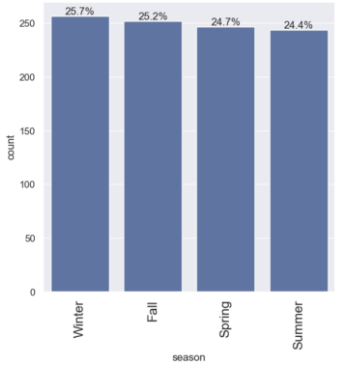


Figure 19: Univariate Analysis – season

✓ **Observations & Insights** can be summarized below: -

- No clear trend visible as all seasons contribute more-or-less equitably (24-26%).
- Whether this variable contributes to content views can be further evaluated during Regression Analysis.

▪ **Perform Bivariate Analysis** – Use Pairplot & Heatmap to carry out bivariate analysis between numerical variables, followed by Boxplots & Barplots between numerical & categorical variables: -

- **Numerical Variables**: -

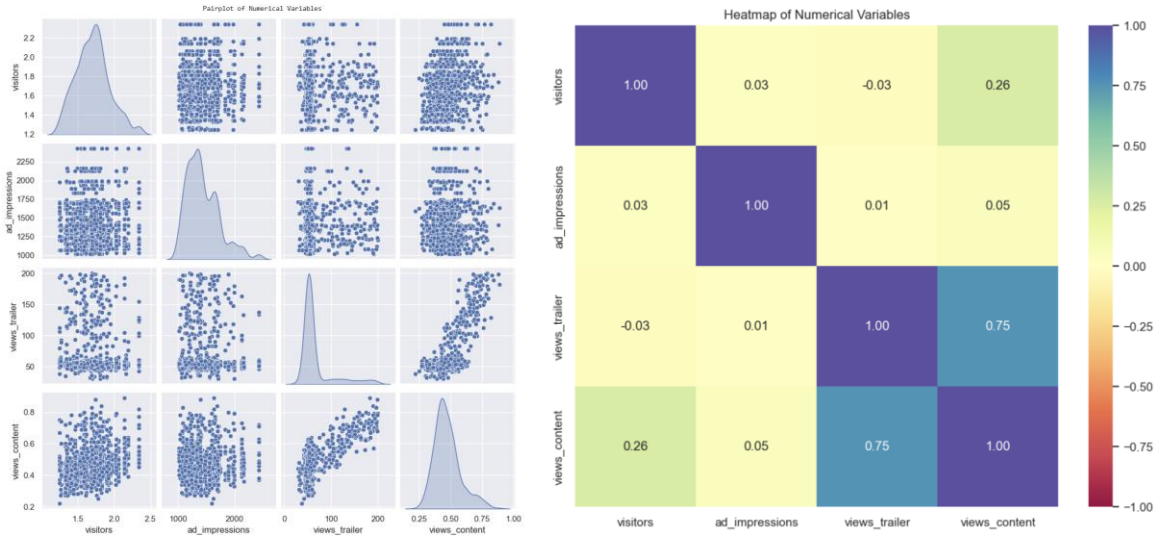


Figure 20: Bivariate Analysis – Numerical Variables

✓ **Observations & Insights** can be summarized below: -

- High correlation seen between views_content and views_trailer in the Heatmap. A positive linear relationship is visible in the Pairplot as well.
- Apart from above, no clear relationship visible between any other variables.
- To further confirm the statistical significance of the relationship, we will carry out Regression Analysis in the subsequent sections.

- **Content Viewership vs Categorical Variables**: -
  (as content viewership is the subject of interest so we analyse the interaction between views_content & other categorical variables & choose to skip analysis of interaction between categorical variables and other numerical variables)

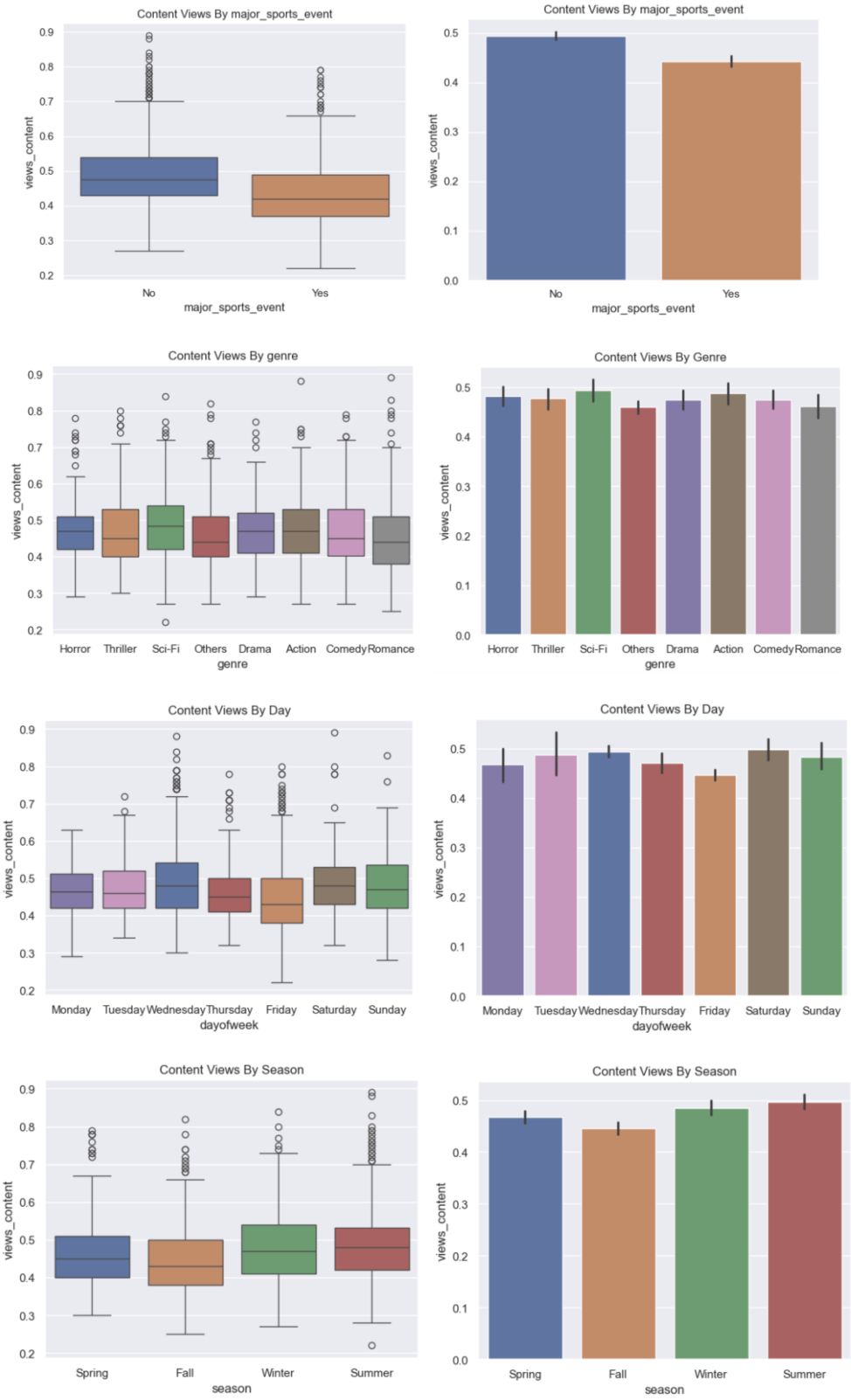


Figure 21: Bivariate Analysis – Categorical Variables

- ✓ **Observations & Insights** can be summarized below: -

  - ➢ **Major Sports Event vs. Views Content**
    - When a major sports event is occurring, it appears that content viewership is generally lower compared to when there is no major sports event.
    - This suggests that during major sports events, users may prefer watching live sports rather than other content, leading to reduced views for regular content.
  - ➢ **Genre vs. Views Content**
    - Sci-Fi seems to have the highest content viewership compared to other genres.
    - This indicates that viewers might prefer certain genres (like Sci-Fi), and tend to draw more views than others.
    - There is also variability within each genre, suggesting that individual content performance can vary significantly.
  - ➢ **Day of the Week vs. Views Content**
    - The weekend (e.g., Sunday and Saturday) seems to have slightly higher content viewership indicating that viewers are more likely to watch content during weekends.
  - ➢ **Season vs. Views Content**
    - Summer & Winter seem to have slightly higher content viewership compared to others.
    - This might indicate that extreme weather conditions seasons lead to more indoor activities like watching content.

P.S. – Answers to the key questions provided in the Problem statement and Insights based on EDA are already covered in the analysis above (as part of 'Observations & Insights' section after each analysis). There is no separate section for the same.

Rubric Question 2: Data Preprocessing

Duplicate & Missing/Error Value-check

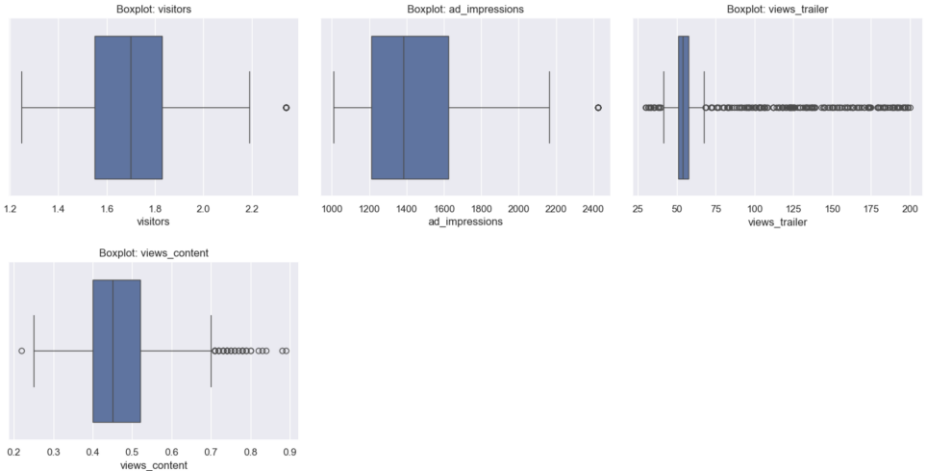
- Please refer Check & Treat Duplicate, Missing & Error Values section (Figure 5 & Figure 6).
- No treatment required as explained in the section above.

Feature Engineering

- Please refer Replacing values in 'major_sports_event' section (Figure 7).
- Values have been changed from '0' to 'No' and '1' to 'Yes', as it is a categorical variable, for better representation of the data.
- We may change it back to binary form (Float-type) later before we start regression analysis.

Outlier Treatment

- Please refer Univariate Analysis section for Outlier inspection for each numerical variable (Figure 9, Figure 11, Figure 13, Figure 15).
- Below is a summary of the histograms & outlier information for each numerical variable: -




```
Lower Wishker at 1.13 | Upper Whisker at 2.25
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 20
Total Outlier Count= 20
Outlier Percentage in visitors= 2.0 %
```

```
Lower Wishker at 590.3199999999997 | Upper Whisker at 2243.6800000000003
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 13
Total Outlier Count= 13
Outlier Percentage in ad_impressions= 1.3 %
```

```
Lower Wishker at 40.73625000000002 | Upper Whisker at 67.96624999999997
Lower Whisker Outlier Count = 16
Upper Whisker Outlier Count = 173
Total Outlier Count= 189
Outlier Percentage in views_trailer= 18.9 %
```

```
Lower Wishker at 0.22000000000000003 | Upper Whisker at 0.7
Lower Whisker Outlier Count = 1
Upper Whisker Outlier Count = 46
Total Outlier Count= 47
Outlier Percentage in views_content= 4.7 %
```

Figure 22: Outlier Inspection Summary

- The outlier count/percentage for each variable is significant and, therefore, cannot be removed from the dataset.
- 'view_trailer' outlier count/percentage is very high (19%), followed by 'view_content' (5%); but, there is a high chance that these outliers can occur for some of the content due to viral marketing. It is better to keep them as is.
- Otherwise, the outlier values seem to be genuine & not entered incorrectly.
- Given the business case, we choose to skip outlier treatment as it may lead to loss of information.

Data Preparation for Modelling

- As a data scientist, we want analyse the data and come up with a Linear Regression model to determine the driving factors for first-day viewership.
  - Before we proceed to build a model, we will encode categorical features.
  - We will split the data into Train and Test datasets to be able to evaluate the model that we build on the Train data. We will build a Linear Regression model using the Train dataset and then check its performance on Test dataset.
- Let's define the Dependent (y) and Independent (x) variables from the dataset: -

	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer
0	1.67	1113.81	No	Horror	Wednesday	Spring	56.70
1	1.46	1498.41	Yes	Thriller	Friday	Fall	52.69
2	1.47	1079.19	Yes	Thriller	Wednesday	Fall	48.74
3	1.85	1342.77	Yes	Sci-Fi	Friday	Fall	49.81
4	1.46	1498.41	No	Sci-Fi	Sunday	Winter	55.83

Figure 23: Independent Variables Dataset (i.e. x)

0	0.51
1	0.32
2	0.39
3	0.44
4	0.46

Figure 24: Dependent Variable Dataset (i.e. y)

- Adding Intercept to the dataset: -

	const	visitors	ad_impressions	major_sports_event	genre	dayofweek	season	views_trailer
0	1.0	1.67	1113.81	No	Horror	Wednesday	Spring	56.70
1	1.0	1.46	1498.41	Yes	Thriller	Friday	Fall	52.69
2	1.0	1.47	1079.19	Yes	Thriller	Wednesday	Fall	48.74
3	1.0	1.85	1342.77	Yes	Sci-Fi	Friday	Fall	49.81
4	1.0	1.46	1498.41	No	Sci-Fi	Sunday	Winter	55.83

Figure 25: Independent Variable Dataset post adding Intercept

- Creating Dummy Variables & convert values to Float datatype: -
  - **One Hot Encoding** is a method for converting categorical variables into a binary format. It creates new binary columns (0s and 1s) for each category in the original variable. Each category in the original column is represented as a separate column, where a value of 1 indicates the presence of that category, and 0 indicates its absence.
  - It takes 'True' or 'False' as its values and is used to get (k-1) dummies out of k categorical levels (sorted in the ascending order of the alphabet) by removing the first level
  - We have 4 categorical variables (major_sports_event, genre, dayofweek, season) where this encoding will be applicable: -
    - ✓ major_sports_event – 2 categorical levels | 1 dummy created| Reference category: 'No'
    - ✓ genre – 8 categorical levels | 7 dummies created| Reference category: 'Action
    - ✓ dayofweek – 7 categorical levels | 6 dummies created | Reference category: 'Friday'
    - ✓ season – 4 categorical levels | 3 dummies created | Reference category: 'Fall'
  - Before we run the dataset through Linear Regression Model, we convert all the values to Float Datatype.

- Below is the transpose of the ‘x – Independent Variable' dataset post dummy-variable creation: -

	0	1	2	3	4	5	6	7	8	9	...
const	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	...
visitors	1.67	1.46	1.47	1.85	1.46	1.61	1.80	1.58	1.70	1.78	...
ad_impressions	1113.81	1498.41	1079.19	1342.77	1498.41	1588.38	1311.96	1690.43	1498.41	1336.16	...
views_trailer	56.70	52.69	48.74	49.81	55.83	49.72	48.15	56.11	51.91	48.22	...
major_sports_event_Yes	0.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	...
genre_Comedy	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
genre_Drama	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	...
genre_Horror	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	...
genre_Others	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	...
genre_Romance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
genre_Sci-Fi	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	...
genre_Thriller	0.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	...
dayofweek_Monday	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
dayofweek_Saturday	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
dayofweek_Sunday	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	...
dayofweek_Thursday	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	...
dayofweek_Tuesday	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
dayofweek_Wednesday	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	...
season_Spring	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
season_Summer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	...
season_Winter	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	...

Figure 26: Dataset post Dummy Variable Creation

- Split ‘x’ & ‘y’ into Train & Test datasets in 70:30 ratio: -
  - In order to train a model properly, we require Train & Test datasets such that the model can be trained using the Train dataset and can be tested on the unseen Test dataset to get a better understanding of how the model is performing.
  - If we have only one dataset provided, we need to split it into Train and Test datasets.
  - Post Splitting the datasets, below are the rows for each of the Train & Test datasets: -

```
Number of rows in Train dataset = 700
Number of rows in Test dataset = 300
```

Figure 27: Row count post Dataset-split into Train & Test sets

Rubric Question 3: Model building- Linear Regression

Build Linear Regression Model & Displace Summary

- Train Dataset is passed through the OLS (Ordinary Least Squares) Regression model.
- Below is the summary of the model output: -

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:	Mon, 25 Nov 2024	Prob (F-statistic):	1.32e-215
Time:	21:10:04	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.

Figure 28: Linear Regression Model Summary – ver1 (Interim)

Model (Interim) Coefficients with Column-names & Linear Regression Equation

- Below are the coefficients with column-names: -

const	0.060157
visitors	0.129451
ad_impressions	0.000004
views_trailer	0.002330
major_sports_event_Yes	-0.060326
genre_Comedy	0.009352
genre_Drama	0.012625
genre_Horror	0.009862
genre_Others	0.006325
genre_Romance	0.000551
genre_Sci-Fi	0.013143
genre_Thriller	0.008708
dayofweek_Monday	0.033662
dayofweek_Saturday	0.057887
dayofweek_Sunday	0.036321
dayofweek_Thursday	0.017289
dayofweek_Tuesday	0.022837
dayofweek_Wednesday	0.047376
season_Spring	0.022602
season_Summer	0.044203
season_Winter	0.027161

Figure 29: Coefficients with Column-names of the Interim Regression Equation

- Below is the Interim Linear Regression Equation: -

```
views_content = 0.0601565517494681 + 0.1294505744423475 * ( visitors ) +  3.6231483229941107e-06 * ( ad_impressions ) +  0.0023296732638245334 * ( views_
trailer ) +  -0.0603260919574294 * ( major_sports_event_Yes ) +  0.009352182723248822 * ( genre_Comedy ) +  0.012625081979422553 * ( genre_Drama ) +  0.0
09862438834606128 * ( genre_Horror ) +  0.006325042184391273 * ( genre_Others ) +  0.0005509751584567429 * ( genre_Romance ) +  0.013143426559714788 * (
genre_Sci-Fi ) +  0.008707503270430399 * ( genre_Thriller ) +  0.033661631397537355 * ( dayofweek_Monday ) +  0.05788749657271275 * ( dayofweek_Saturday
) +  0.03632087548608467 * ( dayofweek_Sunday ) +  0.01728908621675788 * ( dayofweek_Thursday ) +  0.022836918485820683 * ( dayofweek_Tuesday ) +  0.0473
7607155228765 * ( dayofweek_Wednesday ) +  0.022602339842472906 * ( season_Spring ) +  0.0442034460193336 * ( season_Summer ) +  0.027161220661786764 * (
season_Winter )
```

Figure 30: Interim Linear Regression Equation

Model (Interim) Statistics Observations & Insights

- General Model Information: -
  - **Dependent Variable**: views_content
    - ✓ This is the target variable that the model aims to predict.
  - **Model**: Ordinary Least Squares (OLS) Regression.
  - **R-squared**: **0.792**
    - ✓ This indicates that **79.2%** of the variability in the dependent variable (views_content) is explained by the independent variables included in the model. A high value suggests a strong model fit on Training set.
  - **Adj. R-squared: 0.785**
    - ✓ This adjusts the R-squared value for the number of predictors in the model. It penalizes the inclusion of less meaningful predictors.
    - ✓ 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 **0.722**, which is good.
  - **F-statistic**: **129.0** with p-value **1.32e-215**
    - ✓ This tests the null hypothesis that all regression coefficients are equal to zero. The small p-value suggests that the model is statistically significant as a whole. The regression model is applicable as the test is statistically significant.

- Coefficients Table: -
  - Each row represents an independent variable in the regression model.
  - **coef**: The estimated coefficient of the variable – It represents the change in the dependent variable for a one-unit change in the independent variable, holding all else constant.
  - **std err:** The standard error of the coefficient estimate.
  - **t:** The t-statistic, calculated as the coefficient divided by its standard error.
  - **P>|t|:** The p-value associated with the t-statistic. It tests the null hypothesis that the coefficient is zero – If the p-value is less than 0.05, the coefficient is considered statistically significant at **5%** level.
  - **Confidence Intervals:** The [0.025, 0.975] column shows the **95%** confidence interval for the coefficient.

- Key Variable Insights (Please note that these are temporary insights as we still not have checked for multicollinearity & other assumptions yet. There may be room for more optimization): -

Variable	Coefficient	P-value	Observations & Insights
const	0.0602	0.001 (Significant)	✓ This is the intercept, representing the baseline value of views_content when all independent variables are 0. In this context, it reflects the expected number of views when there are no visitors, ad impressions, views of trailers, or other effects
visitors	0.1295	0.000 (Significant)	✓ For every additional visitor, views_content increases by 0.1295 units, holding all other variables constant. ✓ This variable has the highest practical effect among the predictors, emphasizing that the number of visitors is a strong driver of views_content.
ad_impressions	3.623e-06	0.582 (Not Significant)	✓ The coefficient is effectively 0, indicating that ad_impressions have no meaningful impact on views_content. This variable may not contribute much to predicting views.
views_trailer	0.5245	0.000 (Significant)	✓ For every additional trailer view, views_content increases by 0.5245 units. This is a strong and significant predictor, suggesting that trailers effectively drive content viewership.
major_sports_event_Yes	-0.0603	0.000 (Significant)	✓ If a major sports event is occurring, views_content decreases by 0.0603 units compared to when no sports event is happening (i.e. reference category). This implies sports events may divert viewers away from other content.
genre_Comedy	0.0099	0.582 (Not Significant)	✓ 'Comedy' content sees a small increase in views_content (+0.0099) compared to the 'Action' (reference category), but this is not statistically significant.
genre_Drama	0.0126	0.121 (Not Significant)	✓ 'Drama content sees a small increase in views_content (+0. 0126) compared to the 'Action' (reference category), but this is not statistically significant.
genre_Horror	0.0099	0.228 (Not Significant)	✓ 'Horror' content sees a small increase in views_content (+0.0099) compared to the 'Action' (reference category), but this is not statistically significant.

genre_Others	0.0063	0.370 (Not Significant)	✓ 'Others' content sees a small increase in views_content (+0. 0063) compared to the 'Action' (reference category), but this is not statistically significant.
genre_Romance	0.0006	0.948 (Not Significant)	✓ 'Others' content sees a small increase in views_content (+0. 0006) compared to the 'Action' (reference category), but this is not statistically significant.
genre_Sci-Fi	0.0131	0.110 (Not Significant)	✓ 'Sc-Fi' content sees a small increase in views_content (+0. 0131) compared to the 'Action' (reference category), but this is not statistically significant (if we consider 5% as level of significance)
genre_Thriller	0.0087	0.281 (Not Significant)	✓ 'Thriller' content sees a small increase in views_content (+0. 0087) compared to the 'Action' (reference category), but this is not statistically significant (if we consider 5% as level of significance)
dayofweek_Monday	0.0337	0.012 (Significant)	✓ On Mondays, views_content increases by 0.0337 units compared to Fridays (Reference category).
dayofweek_Saturday	0.0579	0.000 (Significant)	✓ On Saturdays, views_content increases significantly by 0.0579 units compared to Fridays (Reference category), suggesting Saturdays are a strong day for viewership.
dayofweek_Sunday	0.0363	0.000 (Significant)	✓ On Sundays, views_content increases significantly by 0.0363 units compared to Fridays (Reference category), suggesting Sundays are a strong day for viewership.
dayofweek_Thursday	0.0173	0.000 (Significant	✓ On Thursdays, views_content increases marginally by 0.0173 units compared to Fridays (Reference category).
dayofweek_Tuesday	0.0228	0.281 (Not Significant)	✓ Tuesdays have a very small increase in views compared to Fridays, but it is not statistically significant.
dayofweek_Wednesday	0.0474	0.000 (Significant)	✓ On Wednesdays, views_content increases significantly by 0.0474 units compared to Fridays (Reference category), suggesting Wednesdays are a strong day for viewership.
season_Spring	0.0226	0.000 (Significant)	✓ In Spring, views_content increases by 0.0226 units compared to Fall (Reference category).
season_Summer	0.0442	0.000 (Significant)	✓ In Summer, views_content increases by 0.0442 units compared to Fall (Reference category).
season_Winter	0.0272	0.000 (Significant)	✓ In Winter, views_content increases by 0.0272 units compared to Fall (Reference category).

Table 2: Regression Model Variable Summary – Interim

Model (Interim) Performance Check

- Let's evaluate model performance using metrics RMSE, MAE, R-Squared.
- Below is the summary for both Training & Test datasets: -

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.04853	0.038197	0.791616	0.785162	8.55644

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.050603	0.040782	0.766447	0.748804	9.030464

Figure 31: Interim Regression Model Performance Metrics

- Summary: -
  - The Training R-squared is 0.76, 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 content views within a mean error of 0.04 on the test data
  - MAPE of 9.03 on the test data means that we are able to predict within 9.03% of the content views

- A more detailed Model Performance Evaluation would be carried out in the end once we derive the final model post assumptions-check & optimizations.

Rubric Question 4: Testing the Assumptions of Linear Regression Model

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.**
- **Variance Inflation factor**: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is 'inflated' by the existence of correlation among the predictor variables in the model.
- **General Rule of Thumb while interpreting VIF: -**
  - If VIF is 1, then there is no correlation among the k^{th} predictor and the remaining predictor variables, and hence, the variance of β_k is not inflated at all.
  - If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.
- Below is the VIF for all independent variables: -

	feature	VIF
0	const	99.679317
1	visitors	1.027837
2	ad_impressions	1.029390
3	views_trailer	1.023551
4	major_sports_event_Yes	1.065689
5	genre_Comedy	1.917635
6	genre_Drama	1.926699
7	genre_Horror	1.904460
8	genre_Others	2.573779
9	genre_Romance	1.753525
10	genre_Sci-Fi	1.863473
11	genre_Thriller	1.921001
12	dayofweek_Monday	1.063551
13	dayofweek_Saturday	1.155744
14	dayofweek_Sunday	1.150409
15	dayofweek_Thursday	1.169870
16	dayofweek_Tuesday	1.062793
17	dayofweek_Wednesday	1.315231
18	season_Spring	1.541591
19	season_Summer	1.568240
20	season_Winter	1.570338

Figure 32: VIF for Independent Variables (Multicollinearity-check)

- Summary of Multicollinearity-check with further actions to fix Multicollinearity (if any): -
  - We will ignore the VIF for intercept.
  - Since, **none of the VIF > 5, clearly, none of the features are correlated** with one or more independent features. Hence, **no treatment required to fix Multicollinearity.**
  - In case any of the VIF was greater than 5 (in our case, below steps are not required as none of VIF>5), we would have followed the below steps: -
    I. Drop every column one by one that has a VIF score greater than 5.
    II. Look at the adjusted R-squared and RMSE of all these models.
    III. Drop the variable that makes the least change in adjusted R-squared.
    IV. Check the VIF scores again.
    V. Continue till you get all VIF scores under 5.

- Multicollinearity affects only those independent variables that are correlated. In this case, we can trust the p-values of all the features (independent variables) since none of the independent variables are correlated.
- **Dropping high p-value variables**: -
  - We will drop the predictor variables having a p-value greater than 0.05 as they do not significantly impact the target variable.
  - But sometimes p-values change after dropping a variable. So, we won't drop all variables at once.
  - Instead, we will do the following:
    I. Build a model, check the p-values of the variables, and drop the column with the highest p-value.
    II. Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
    III. Repeat the above two steps till there are no columns with p-value > 0.05.
  - By following the above steps, we ran **9 iterations to arrive at the final model**. Below is the summary of each iteration showing which variables with what p-values were removed: -

```
------- Iteration  1  -------
Removed  genre_Romance  with p value  0.948
------- Iteration  2  -------
Removed  ad_impressions  with p value  0.582
------- Iteration  3  -------
Removed  genre_Others  with p value  0.298
------- Iteration  4  -------
Removed  genre_Thriller  with p value  0.409
------- Iteration  5  -------
Removed  genre_Comedy  with p value  0.427
------- Iteration  6  -------
Removed  genre_Horror  with p value  0.48
------- Iteration  7  -------
Removed  genre_Drama  with p value  0.267
------- Iteration  8  -------
Removed  genre_Sci-Fi  with p value  0.285
------- Iteration  9  -------
Removed  dayofweek_Tuesday  with p value  0.075

Features after dropping high p_value variables:-
 ['const', 'visitors', 'views_trailer', 'major_sports_event_Yes', 'dayofweek_Monday', 'dayofweek_Saturday', 'dayofweek_Sunday', 'dayofweek_Thursday', 'da
yofweek_Wednesday', 'season_Spring', 'season_Summer', 'season_Winter']
```

Figure 33: Features Removed with High p-value

- Updated Linear Regression Model post removing non-significant features: -

```
                            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:                Mon, 25 Nov 2024   Prob (F-statistic):          7.03e-224
Time:                        21:10:05   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.
```

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.048841	0.038385	0.788937	0.785251	8.595246

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051109	0.041299	0.761753	0.751792	9.177097

Figure 34: Linear Regression Model Post Multicollinearity-check – ver2.

- Observation – Model seems to be optimized given the Model Performance metrics (metrics comparable between Train & Test datasets with a good R-squared metric). However, final analysis to be outlined in the end post checking other assumptions.

Test for Linearity and Independence

- Linearity describes a straight-line relationship between two variables. Independent variables must have a linear relation with the dependent variable.
- The independence of the error-terms/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.
- **Test** – we plot a Residual vs Fitted values chart & if no pattern is visible, it means the independent variables have a linear relation.

	Actual Values	Fitted Values	Residuals
731	0.40	0.445434	-0.045434
716	0.70	0.677403	0.022597
640	0.42	0.433999	-0.013999
804	0.55	0.562030	-0.012030
737	0.59	0.547786	0.042214

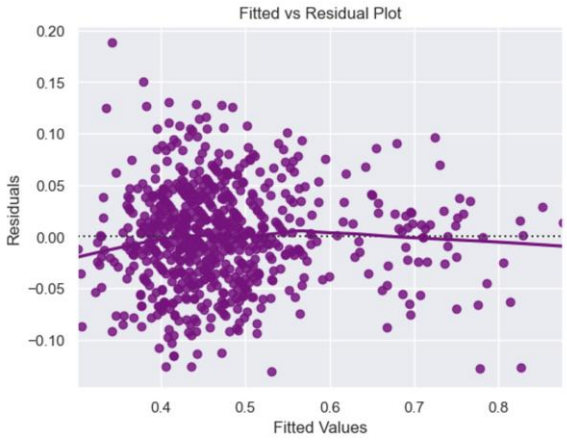


Figure 35: Residual Plot to test Linearity & Independence

- We see **no pattern** in the plot above. Hence, the **assumptions of Linearity and Independence are satisfied**.

Test for Normality

- Error terms/Residuals should be normally distributed.
- If the Residuals are not normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.
- Presence of non-normality suggests that there are a few unusual data points which must be studied closely to make a better model.
- **Test** – (1) Histogram of Residuals (2) QQ Plot (3) Shapiro-Wilk test
  - **Histogram of Residuals**: It is evident from the plot below that the Residuals are normally distributed.

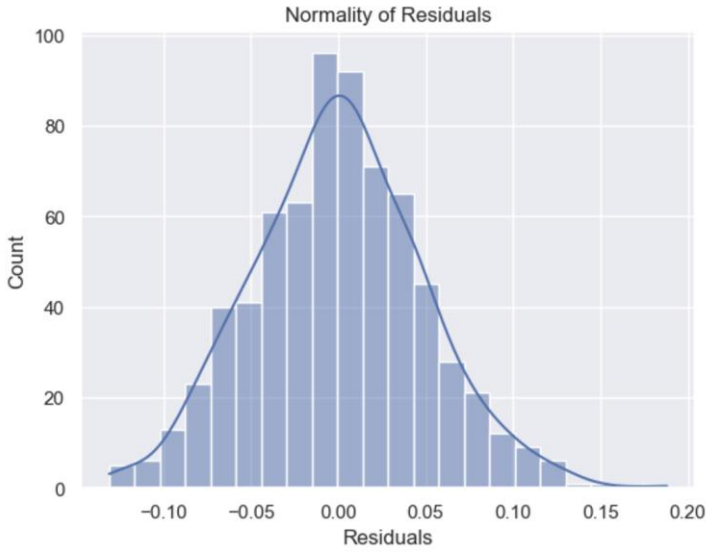


Figure 36: Histogram of Residuals to test Normality

- **QQ Plot**: It is evident from the plot below that most of the points are lying on the straight line in QQ plot, implying, the Residuals are normally distributed.

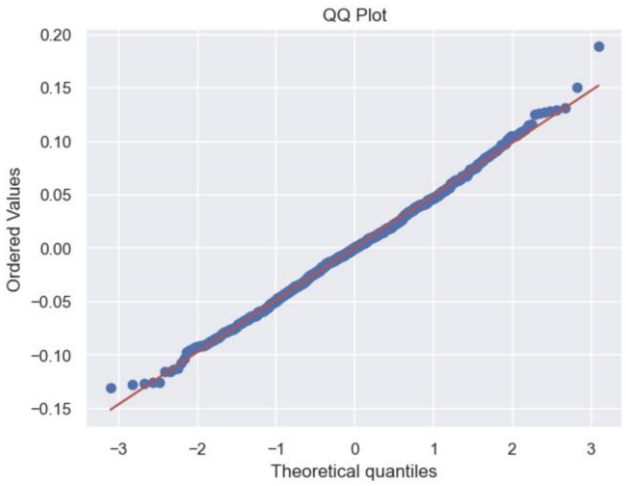


Figure 37: QQ Plot of Residuals to test Normality

- **Shapiro-Wilk test**:
  - ✓ Null hypothesis: Residuals are normally distributed
  - ✓ Alternate hypothesis: Residuals are not normally distributed
    - ➢ Below is the output of the Shapiro-Wilk test: -

```
ShapiroResult(statistic=0.9973155427169234, pvalue=0.31085896470043806)
```

Figure 38: Shapiro-Wilk Test to test Normality

    - ➢ Since the **p-value > 0.05,** we fail to reject the Null hypothesis. Therefore, **Residuals are normally distributed**.

Test for Homoscedasticity

- **Homoscedastic** - If the variance of the Residuals is symmetrically distributed across the regression line, then the data is said to homoscedastic.
- **Heteroscedastic** - If the variance is unequal for the Residuals across the regression line, then the data is said to be heteroscedastic. In this case the residuals can form an arrow shape or any other non-symmetrical shape.
- **Test** - Goldfeldquandt Test: -
  - Null hypothesis: Residuals are homoscedastic
  - Alternate hypothesis: Residuals have heteroscedasticity
    - ✓ Below is the output of the Goldfeldquandt Test: -

```
[('F statistic', 1.131361290420075), ('p-value', 0.12853551819087372)]
```

Figure 39: Goldfeldquandt Test to test Homoscedasticity

    - ✓ Since the **p-value > 0.05,** we fail to reject the Null hypothesis. Therefore, **Residuals are homoscedastic.**

Rubric Question 5: Model Performance Evaluation

Final Model Summary

- Now that we are done with all the assumptions, lets recreate the final model and print its summary to gain insights: -

```
                            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:                Mon, 25 Nov 2024   Prob (F-statistic):          7.03e-224
Time:                        21:10:07   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.
```

Figure 40: Linear Regression Model Summary – Final

- Below are the randomly chosen 10 Predicted-values against their Actuals from the Test dataset. The output looks quite comparable: -

	Actual	Predicted
983	0.43	0.43
194	0.51	0.50
314	0.48	0.43
429	0.41	0.49
267	0.41	0.49
746	0.68	0.68
186	0.62	0.60
964	0.48	0.50
676	0.42	0.49
320	0.58	0.56

Figure 41: Predicted vs Actuals from Test Dataset

- Below are the final model-coefficients with their columns/features, followed by the Final Linear Regression Equation: -

```
const                     0.074671
visitors                  0.129096
views_trailer             0.002331
major_sports_event_Yes   -0.060555
dayofweek_Monday          0.032066
dayofweek_Saturday        0.057029
dayofweek_Sunday          0.034386
dayofweek_Thursday        0.015449
dayofweek_Wednesday       0.046495
season_Spring             0.022605
season_Summer             0.043391
season_Winter             0.028231
```

Figure 42: Final Model-coefficients with their Features

```
views_content = 0.07467052053721267 + 0.12909581825894126 * ( visitors ) +  0.002330816786164013 * ( views_trailer ) +  -0.06055507818137332 * ( major_sp
orts_event_Yes ) +  0.03206580679023629 * ( dayofweek_Monday ) +  0.057028596601651195 * ( dayofweek_Saturday ) +  0.034386229923625 * ( dayofweek_Sunday
) +  0.01544944176997319 * ( dayofweek_Thursday ) +  0.04649480366984812 * ( dayofweek_Wednesday ) +  0.022604915818118004 * ( season_Spring ) +  0.04339
100263609978 * ( season_Summer ) +  0.028230557183976823 * ( season_Winter )
```

Figure 43: Final Linear Regression Equation

Evaluate Model Performance Metrics

- Let's evaluate model performance using metrics RMSE, MAE, R-Squared.
- Below is the Performance Metrics summary for both Training & Test datasets: -

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.048841	0.038385	0.788937	0.785251	8.595246

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.051109	0.041299	0.761753	0.751792	9.177097

Figure 44: Final Linear Regression Model Performance Metrics

- High-level summary: -
  - ✓ The model is able to explain ~79% of the variation in the Train dataset & ~76% in the Test dataset.
  - ✓ The Train and Test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.
  - ✓ The MAPE on the Test dataset suggests we can predict within 9.2% of the view contents.
  - ✓ Hence, we can conclude the model is good for prediction as well as inference purposes

- Model Performance deep-dive: -

Metric Comparison

Metric	Training Set	Test Set	Difference (Train - Test)
RMSE	0.0488	0.0511	-0.0023
MAE	0.0384	0.0413	-0.0029
R-squared	0.7889	0.7618	-0.0271
Adjusted R-squared	0.7853	0.7518	-0.0335
MAPE (%)	8.5952	9.1771	-0.5819

Figure 45: Final Linear Regression Model Performance Metrics Comparison

I. **Root Mean Square Error (RMSE)**

- RMSE measures the average magnitude of the error between the predicted values and the actual values.
- RMSE quantifies the difference between the predicted values and actual observations. It does this by calculating the square root of the average of squared errors.
- By squaring the errors, RMSE penalizes larger errors more than smaller ones.
- RMSE is often used to compare the predictive accuracy of different models. A lower RMSE value suggests better model performance, making it a useful metric for choosing the best model among multiple options.
- It helps evaluate the **bias-variance trade-off**: -
  - A much lower RMSE on the training set compared to the test set suggests overfitting (model learns noise rather than the underlying trend).
  - Similar RMSE values on training and test sets indicate that the model generalizes well and is less prone to overfitting or underfitting
  - A high RMSE on the test data relative to the training data is an indicator of a model that doesn't generalize well. This is particularly important when evaluating whether a model will perform adequately on new, unseen data.
- **Evaluation of the Model above: -**
  - RMSE for both Train & Test dataset are very low, suggesting good model performance
  - RMSE for both Train & Test dataset are comparable, implying that the model is neither overfitted nor underfitted.

II. **Mean Absolute Error (MAE)**

- MAE represents the average absolute difference between the actual values and the predicted values.
- MAE is expressed in the same units as the dependent variable, it is easy to interpret. It tells, on average, how much model's predictions deviate from the actual values.
- A lower MAE value indicates better accuracy in predicting the target variable.
- If RMSE is significantly larger than MAE, it suggests that the model has some larger errors or outliers (since RMSE penalizes large errors more heavily due to squaring).
- If RMSE and MAE are close in value, it suggests a more consistent model with errors of similar magnitude.
- **Evaluation of the Model above: -**
  - MAE for both Train & Test dataset are very low, suggesting model's predictions are close to the actual values, and the average magnitude of prediction error is quite small.
  - MAE for both Train & Test dataset are comparable, implying that the model performs consistently across both datasets.

III. **R-squared**

- Also known as the coefficient of determination, it is a key metric that indicates how well the independent variables explain the variance in the dependent variable. It provides an overall measure of how well the regression model fits the data.
- Value of R-squared ranges from 0 to 1.
- A higher R-squared value means the model better fits the data, indicating that the independent variables do a good job of explaining the variability in the dependent variable.
- A low R-squared value suggests that the model is not capturing much of the underlying trend and that the independent variables may not be well-chosen or sufficient.
- A very high R-squared value (close to 1) in the training set can sometimes indicate overfitting, where the model fits the training data too closely, including noise and random fluctuations, rather than capturing the underlying pattern. Overfitting typically results in poor performance on test data. Comparing the R-squared values of training and test sets is crucial to evaluate whether the model generalizes well.
- **Evaluation of the Model above: -**
  - The model is able to explain **~79%** of the variation in the Train dataset & **~76%** in the Test dataset, which suggest that it is a good model.
  - To establish there is no overfitting, we can observe that that R-squared is not close to 1 & the values are comparable for Train & Test datasets.

IV. **Adjusted R-squared**

- Adjusted R-squared is designed to overcome one of the main drawbacks of the standard R-squared: the tendency to always increase when new predictors are added to the model, regardless of their relevance. This can make R-squared misleading when comparing models of different complexities.
- Adjusted R-squared only increases if a new predictor improves the model beyond what would be expected by chance. It decreases if the added predictor does not have a meaningful effect, helping to prevent overfitting by penalizing excessive model complexity.

- Higher Adjusted R-squared indicates that a greater proportion of variance is being explained by the model without including irrelevant variables. This means that most of the predictors are making meaningful contributions.
- If training adjusted R-squared is much higher than the test adjusted R-squared, it may indicate overfitting.
- If both values are close, it suggests that the model generalizes well to unseen data.
- **Evaluation of the Model above: -**
  - Adjusted R-squared indicates that **79%** of the proportion of variance is being explained by the model without including irrelevant variables in the Train dataset & **~75%** in the Test dataset, which suggest that it is a good model.
  - Adjusted R-squared of Train & Test dataset (0.79 & 0.75 respectively) are comparable, suggesting the model generalizes well to unseen data.

V. Mean Absolute Percentage Error MAPE

- MAPE represents errors as a percentage of the actual values. It tells, on average, how much error there is between the actual and predicted values, expressed as a percentage.
- AMPE indicates how far off, in percentage terms, the model's predictions are from actual values.
- MAPE allows for easy comparison of models applied to different datasets, even if the data scales are different.
- A low MAPE indicates a highly accurate model.
- **Evaluation of the Model above: -**
  - MAPE values indicate that the model's predictions are, on average, **8.6% off** from actual values in the **Train** dataset and **9.2% off** in the **Test** dataset
  - The low MAPE values indicate a good predictive model, with the average deviation from the actual values being relatively small. This means that the model's predictions are generally accurate.
  - The small increase in MAPE from Train dataset to Test dataset shows that the model **generalizes well to unseen data** without significant overfitting.

Rubric Question 6: Actionable Insights & Recommendations

Final Model Summary

- Please refer the Final Model Summary section for the final Linear Regression Model output.

Insights & Recommendations

- General Model Information: -
  - **Dependent Variable**: views_content
    - ✓ This is the target variable that the model aims to predict.
  - **Model**: Ordinary Least Squares (OLS) Regression.
  - **R-squared**: **0.789**
    - ✓ This indicates that **78.9%** of the variability in the dependent variable (views_content) is explained by the independent variables included in the model. This is a high value, suggesting the model fits well.
  - **Adj. R-squared: 0.786**
    - ✓ Adjusted R-squared values generally range from 0 to 1, indicating that the predictors collectively do a good job explaining the variance. A higher value indicates a good fit. In our case, the value for adj. R-squared is **0.786**, which is good.
  - **F-statistic**: **233.8** with p-value **7.03e-224**
    - ✓ This tests the null hypothesis that all regression coefficients are equal to zero. The small p-value suggests that the model is statistically significant as a whole. The overall model is statistically significant, meaning at least one of the independent variables contributes significantly to explaining views_content.

- Delving into each feature insights from the final Regression Model: -

Variable	Coefficient	P-value	Observations	Insights & Recommendations
const	0.0747	0.000 (Significant)	✓ This is the intercept, representing the baseline value of views_content when all predictors are zero. It is significant, indicating that the intercept is an important component of the model.	✓ NA
visitors	0.1291	0.000 (Significant)	✓ For each additional visitor, content views **increase** by **0.1291** units. This is highly significant, showing that attracting more visitors has a strong positive impact on content views.	✓ Focus on strategies to increase the number of visitors, such as improving marketing efforts or optimizing user acquisition channels.
views_trailer	0.0023	0.000 (Significant)	✓ Each additional trailer view **increase** content views by **0.0023** units. It is significant, suggesting that trailer views positively impact content views.	✓ Increase promotion of trailers to boost content views. ✓ Trailers can be embedded in popular platforms and can be highlighted in social media campaigns, or used as targeted ads.
major_sports_event_Yes	-0.0606	0.000 (Significant)	✓ When there is a major sports event, content views **decrease** by **0.0606** units. This is significant and suggests that sports events have a negative effect on content views, likely because viewers are focused on the sports event instead.	✓ During major sports events, consider adjusting the release schedule or content promotions to minimize the impact of reduced viewership. Consider running campaigns before or after these events.
dayofweek_Monday	0.0321	0.006 (Significant)	✓ On Mondays, content views **increase** by **0.0321** units compared to the reference day (Friday). This is significant.	✓ Viewership is higher on weekends (especially Saturday) and certain weekdays like Wednesdays. ✓ Plan to release new content or run promotions on these high-viewership days to maximize reach and engagement.
dayofweek_Saturday	0.0570	0.000 (Significant)	✓ On Saturdays, views **increase** by **0.0570** units compared to the reference day (Friday). This is highly significant, suggesting higher engagement during weekends.	
dayofweek_Sunday	0.0344	0.000 (Significant)	✓ On Sundays, views **increase** by **0.0344** units compared to the reference day (Friday). This is highly significant, suggesting higher engagement during weekends.	
dayofweek_Thursday	0.0154	0.021 (Significant)	✓ On Thursdays, views **increase** by **0.0154** units compared to the reference day (Friday). This is marginally significant.	
dayofweek_Wednesday	0.0465	0.000 (Significant)	✓ On Wednesdays, content views **increase** by **0.0465** units compared to the reference day (Friday). This is significant.	

season_Spring	0.0226	0.000 (Significant)	✓ During Spring, content views **increase** by **0.0226** units compared to reference season (Fall). This is significant.	✓ Viewership is higher during **summer** and **winter**, possibly due to vacation times when people have more free time to consume content. ✓ Focus content releases and marketing efforts during these seasons for maximum engagement.
season_Summer	0.0434	0.000 (Significant)	✓ During Summer, content views **increase** by **0.0434** units compared to reference season (Fall). This is significant.	
season_Winter	0.0282	0.000 (Significant)	✓ During Winter, content views **increase** by **0.0282** units compared to reference season (Fall). This is significant.	

Table 3: Regression Model Feature Summary – Final

- **Summary of Recommendations: -**
  - ✓ **Increase Visitor Traffic –** Invest in strategies that drive traffic to your platform & improve user acquisition through: -
    - ➢ Increase paid advertising.
    - ➢ Utilize social media marketing.
    - ➢ Optimize SEO (Search Engine Optimization).
  - ✓ **Boost Trailer Engagement: –** Increase visibility of trailers through: -
    - ➢ Social media campaigns.
    - ➢ Embedding trailers on popular websites and partner platforms.
    - ➢ Using trailers in email marketing campaigns.
  - ✓ **Adjust Scheduling During Major Sports Events: –** Since major sports events negatively impact viewership, adjust content scheduling to minimize this effect. Focus on: -
    - ➢ Releasing content before or after major events.
    - ➢ Running promotions that tie in with the sports event (i.e., content related to the event).
  - ✓ **Optimize Content Release Days: –** Content views are higher on weekends (especially, Saturdays) and on Wednesdays: -
    - ➢ Schedule important content releases or promotional campaigns on these days to maximize viewership and engagement.
  - ✓ **Seasonal Marketing Strategy: –** Content views are relatively higher in Summers & Winters: -
    - ➢ Viewership is higher in summer and winter. Allocate more marketing budget during these seasons and align content releases to take advantage of increased viewer engagement.
    - ➢ Consider running seasonal promotions or special offers during these high-viewership periods. Schedule important content releases or promotional campaigns on these days to maximize viewership and engagement.
  - ✓ **Content Optimization by Day and Season: –** Leverage the insights into day of the week and seasonal effects to plan targeted campaigns and release schedules: -
    - ➢ Launch major series or special features on Saturdays to leverage higher viewership.
    - ➢ Utilize summer and winter holidays for promoting binge-worthy content when audiences are more likely to engage.