CSE 487 Assignment-1

Strategic Content Optimization for Amazon Prime Video

Aditya Srivatsav Lolla(alolla |50559685)
Sai Krishna Goud Valdas(svaldas | 50560726)

Problem Statement:

This database related to Amazon Prime Video can be a great source of information for studying viewer preferences, content types, and ration trends. Studying this data allows us to learn what types of content people tune into, what genres sell, as well as what ratings clients will tune into and in what specific demographic group. This analysis can be incredibly beneficial in finding patterns and trends to shape content creation, acquisition strategies, and marketing. This understanding is key to optimizing viewer satisfaction and engagement, leading to increased subscription growth and reduced churn.

The main intention of this problem statement is to fetch the viewer interest's clarity of data using data analysis methodologies of Amazon Prime Video. By leveraging content types and ratings trends, we aim to answer what kind of content best strikes the chord with the viewers and whether the ratings accurately reflect viewer delight. This insight can be used to personalize the content being recommended, personalize marketing communication, and ultimately enhance the user experience. The goal is to use data-driven findings to improve the platform's content strategy and align it more closely with the developing needs of a viewer population.

This analysis has a quite broad scope, as can be seen from the example uses mentioned. As for content strategy, user data can inform what new content to create or licence to ensure that investments are placed in areas of potential high interest and likely to result in high viewer satisfaction. On the marketing side, analytics from viewer preferences can help determine where groups of subscribers are and build marketing campaigns around popular content to attract more subscribers or retain existing ones. With reported user-experience, understanding rating trends would further enable the understanding of trends in recommendations, making them more personalized and fine-tuned. This multifaceted strategy promises to give Amazon Prime Video a major advantage in the streaming marketplace.

The analysis is based on some of the toughest parts of managing a media streaming service. The largest challenge is comprehending the diverse and massive user tastes around the globe. Knowing what trends and preferences are present helps Amazon Prime Video find the best way to cater to the different kinds of viewers. This provides an additional layer of content spend optimization by making sure that the content assets invested in are those that bring, from a viewership perspective, the highest level of impact and engagement, hence viewer satisfaction. In addition, the findings aid in solving the issues of pattern detection of subscribers with content that continuously corresponds to current viewer expectations and preferences. In conclusion, this analysis not only drives your content strategy and marketing efforts but also solves fundamental problems regarding a better streaming service.

Overview of dataset :

The dataset pertaining to Amazon Prime Video provides insights, into the content on its platform. Each entry contains information about a variety of shows and movies offering a glimpse into preferences, content trends and market dynamics. Lets now explore the dataset in detail:

Total rows : 9,668
Total columns : 12
Link: https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows/data

Categories and Details :

Show_id : Unique identifier for each show.
Type : Category of content Title : Name of
the show or movie.
Director : Filmmaker behind the content
Cast : The people who acted in the movie or TV Show
Country: It tells about the specific Country does the movie or TV Show belongs to
Date_added: The release date
Release_year: The release year
Rating: The rating of the movie as per the content Family, Adult, Teen and others
Duration: Time duration of the movie
Listed_in: It tells about genre of the movie
Description: The description of the movie

Data Cleaning / Processing :

Remove the Duplicates

We eliminated duplicates based on the show_id column to guarantee that every entry in the dataset is distinct. This stage ensures data integrity for next studies by confirming that each program or film is represented uniquely.

Handling missing values

We rectified the missing values by substituting the proper values for them. In particular, ‘Unknown’ was used to fill in the missing data in the director, cast, nation, and date_added fields. This method stops important documents from being lost because of missing information.

Standardise the Text Data

In this we standardised the entries in the director and country columns to address inconsistencies in textual data. This required eliminating any leading or trailing whitespace and changing every entry to title case. By ensuring consistency in naming conventions, this makes reliable analysis possible.

Handling Missing Ratings with Mode

Here the column mode was used to fill in this missing values in the rating column. The overall rating patter is preserved when the most frequent rating is used, as this guarantees that the imputation is based on the dataset intrinsic distribution.

Convert Datatype for the Datetime

To enable time-based analysis, we changed the data_aded field to a datetime format. This phase permits consistent data manipulations and additional data-related actions, despite format inference warnings.

Extract Year from data_added

In this step we took the year out of the data_added column and made a new column called year_added to make it easier to analyse patterns over time. Zeros were used to fill in any missing data, and the column was changed to an integer type.

Standardising 'duration' Field to Numeric

In order to express duration in minutes, we took numerical values out of the duration column and standardised them. The quantitative study of content length is supported by this conversion to numeric representation.

Categorise the Content Based to Ratings

Based on ratings, we divided the content into four categories : "Family," "Teen," "Adult," and "Others". By putting material into more general categories, this categorization makes audience targeting and content recommendation tactics easier.

Identification and Removal of Outliers

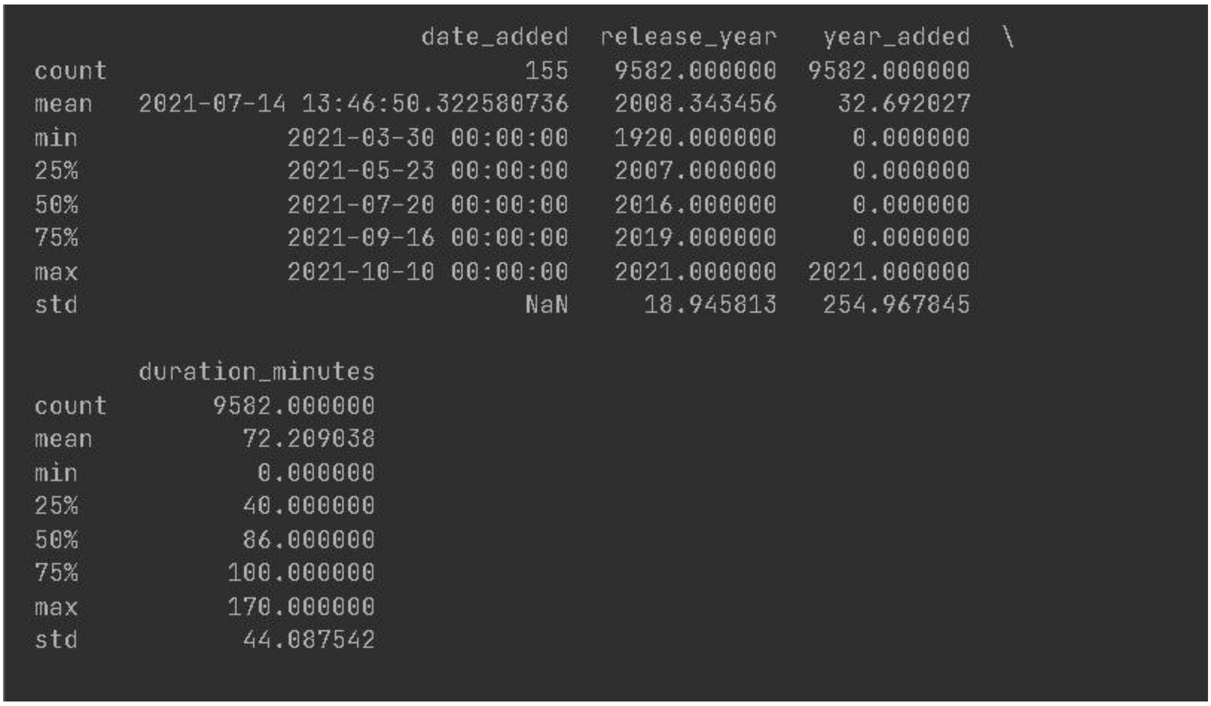
We eliminated the entries whose duration_minutes above the 99th percentile in order to address excessive values in content length. By taking this precaution, outliers are prevented from distorting the study and giving a more realistic depiction of average content lengths.

Identification of the Primary Genre

The primary genre is assigned from the Amazon data from the overall list and a separate column is created. Examples are Action, Drama and Thriller. The primary genre is Action.

Statistics and Analysis:

The dataset covers the average release year of 2008, the dataset spans a wide range of release years from 1920 to 2021 and features a combination of vintage and modern content. The items have an average duration of approximately 72 minutes, which suggest a significant variation in the length of the material. Despite its limitations, the ‘data_added’ data exhibits a regular pattern, with the majority of additions taking place in 2021, notably in the middle of july.



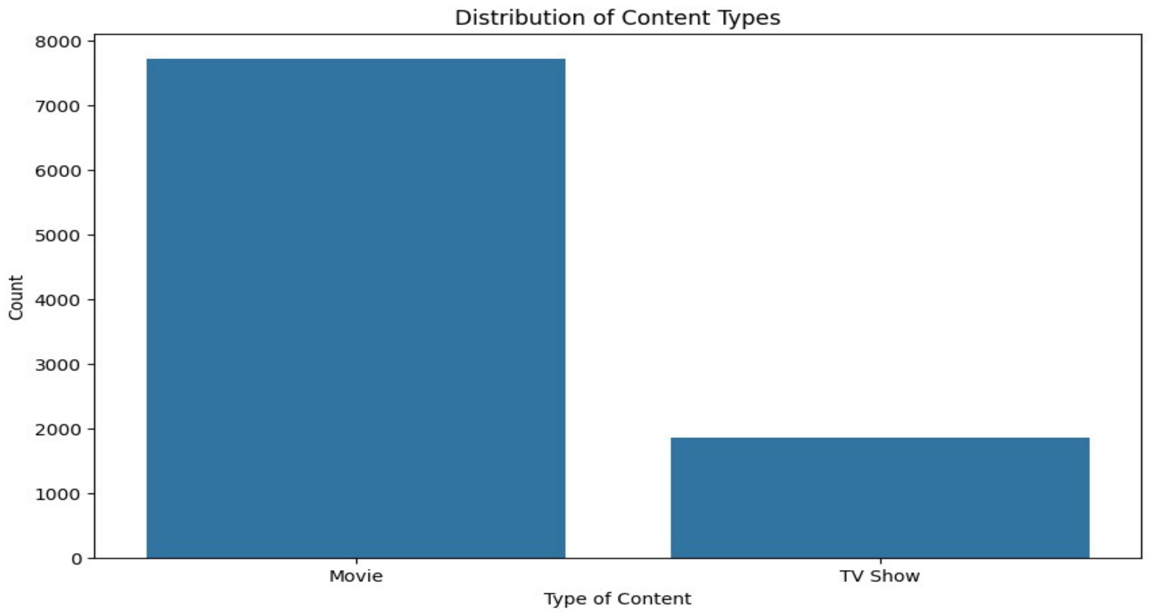
```
                         date_added  release_year   year_added  \
count                           155   9582.000000  9582.000000
mean  2021-07-14 13:46:50.322580736   2008.343456    32.692027
min             2021-03-30 00:00:00   1920.000000     0.000000
25%             2021-05-23 00:00:00   2007.000000     0.000000
50%             2021-07-20 00:00:00   2016.000000     0.000000
75%             2021-09-16 00:00:00   2019.000000     0.000000
max             2021-10-10 00:00:00   2021.000000  2021.000000
std                             NaN     18.945813   254.967845

      duration_minutes
count      9582.000000
mean         72.209038
min           0.000000
25%          40.000000
50%          86.000000
75%         100.000000
max         170.000000
std          44.087542
```

Exploratory Data Analysis

Exploratory data Analysis is the approach which identifies the data. It is the first step of the Analysis.

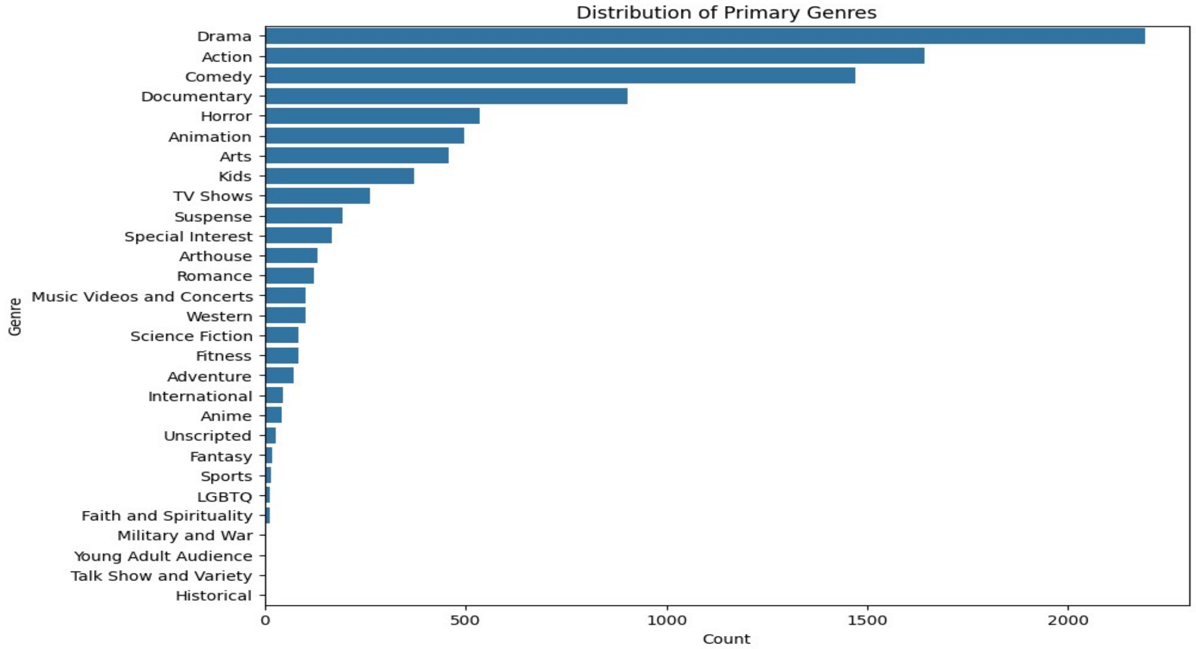
Content Type Distribution :

Examined how different content genres were distributed (TV series Versus movies.)



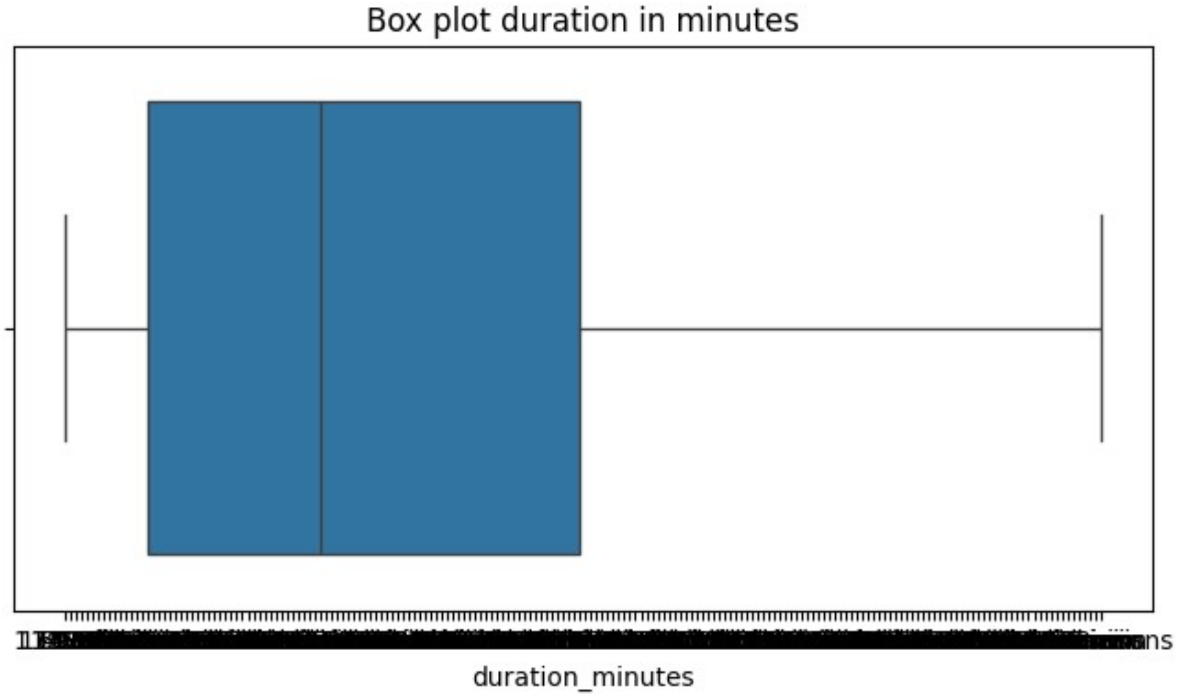
Primary Geners Distribution :

A bar plot was used to visualised the primary genre distribution.



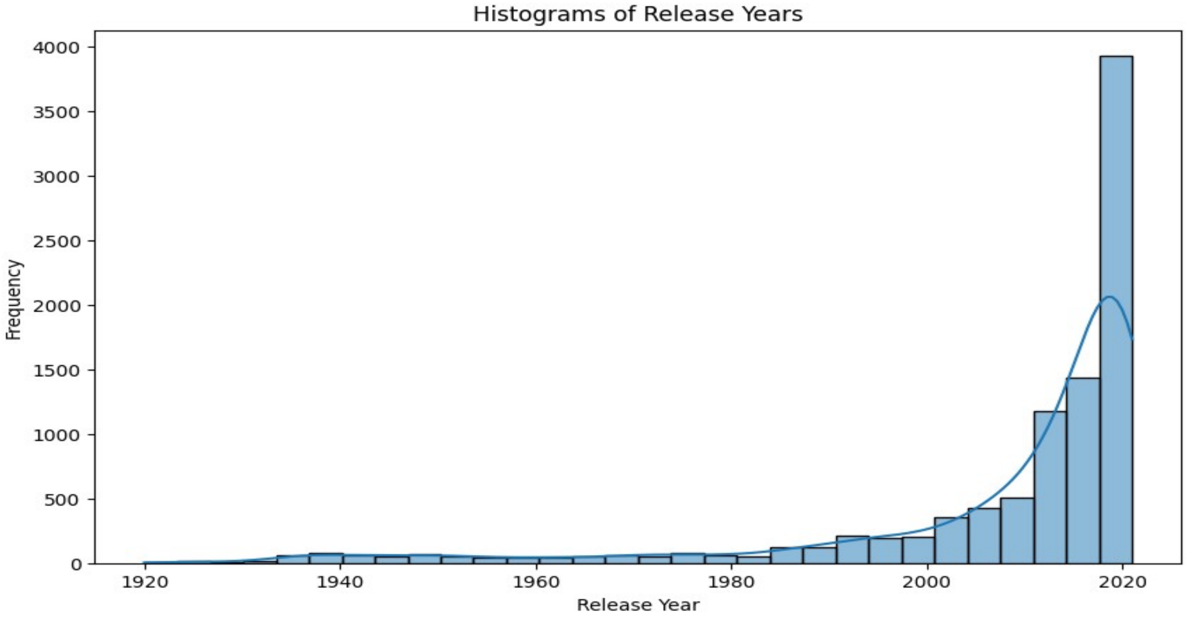
Duration Analysis :

Box and violin plots were used to analyse the length of the content.



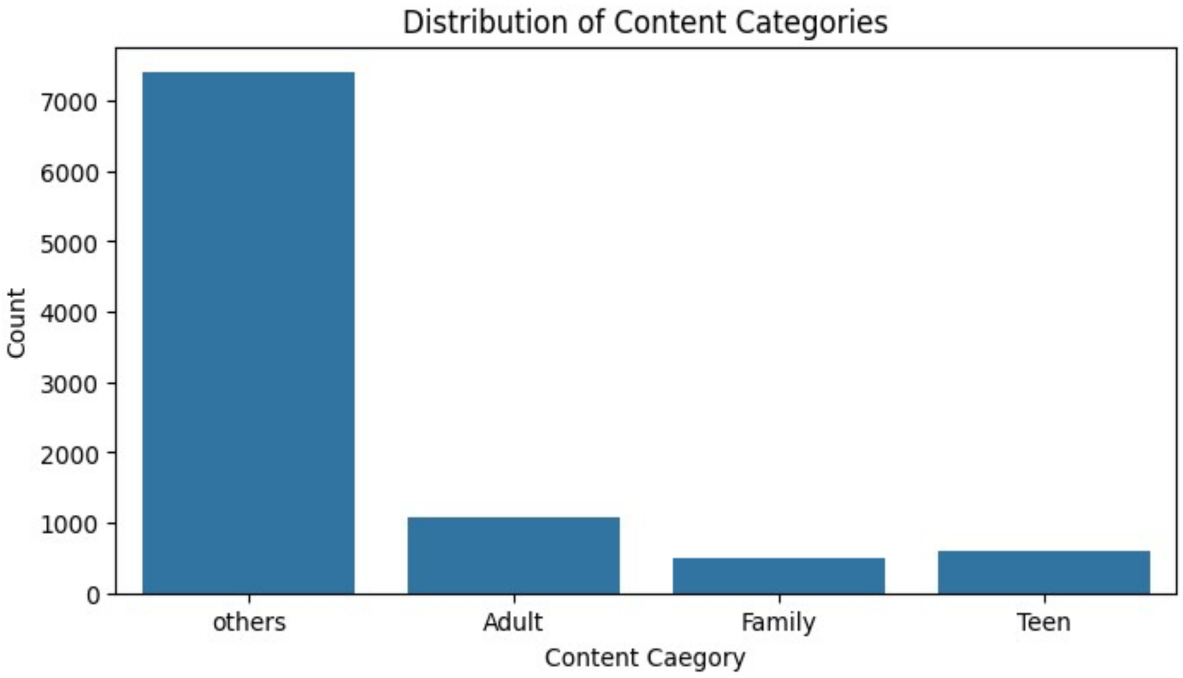
Release Year Analysis :

A histogram was used to analyse the distribution of release years.



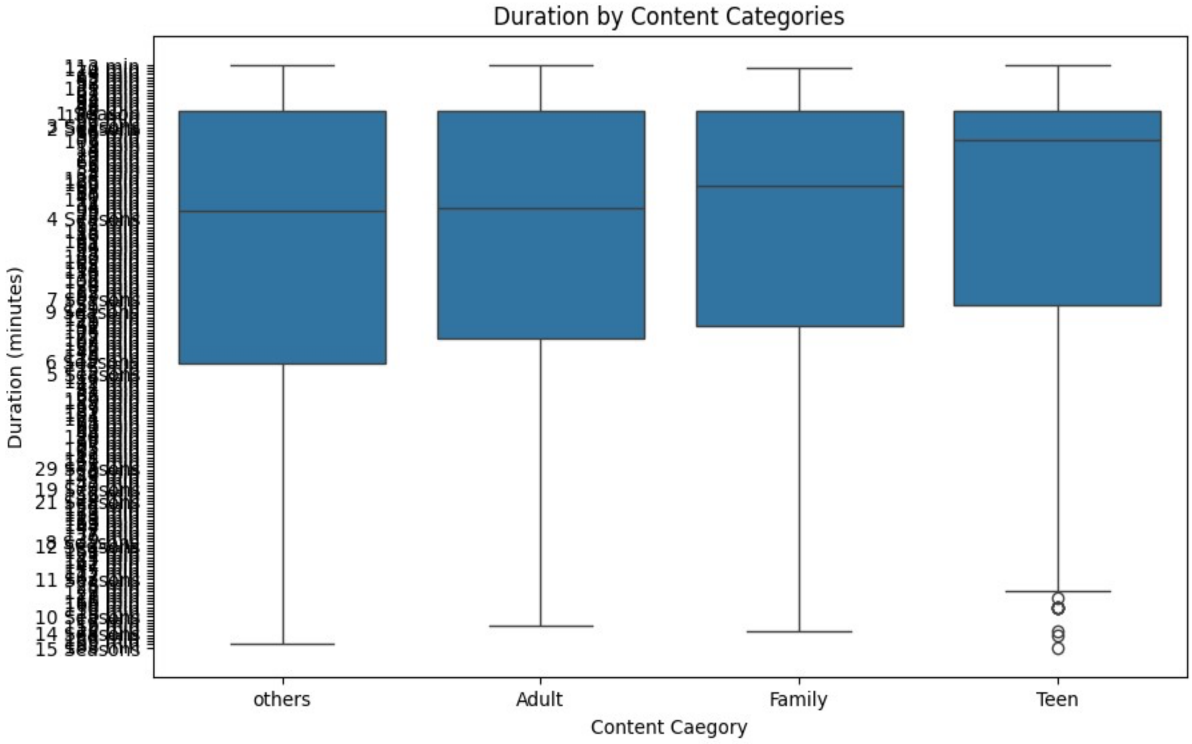
Category-Based Content Distribution :

Count plots were used to visualise the distribution of content by categories.



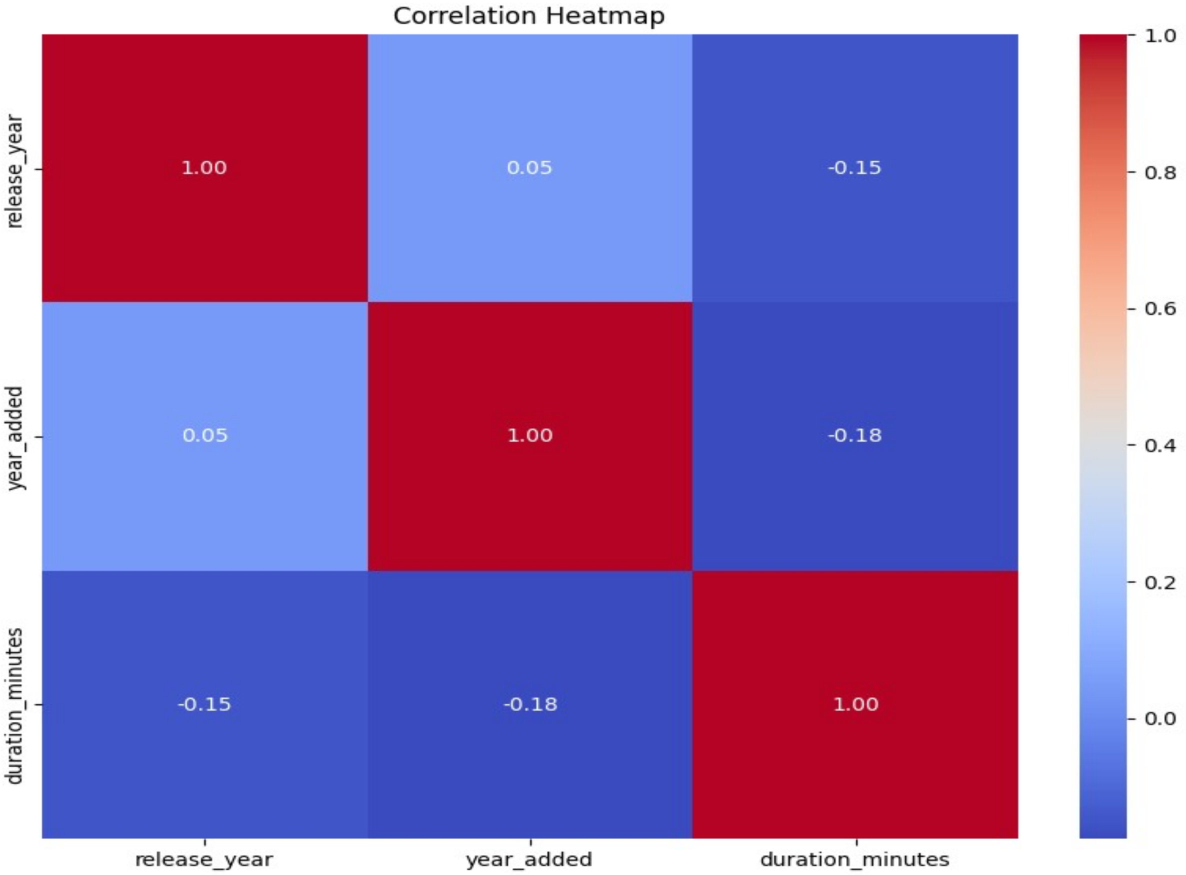
Duration vs Content Category :

Box plots were used to analysed the length of material in various categories.



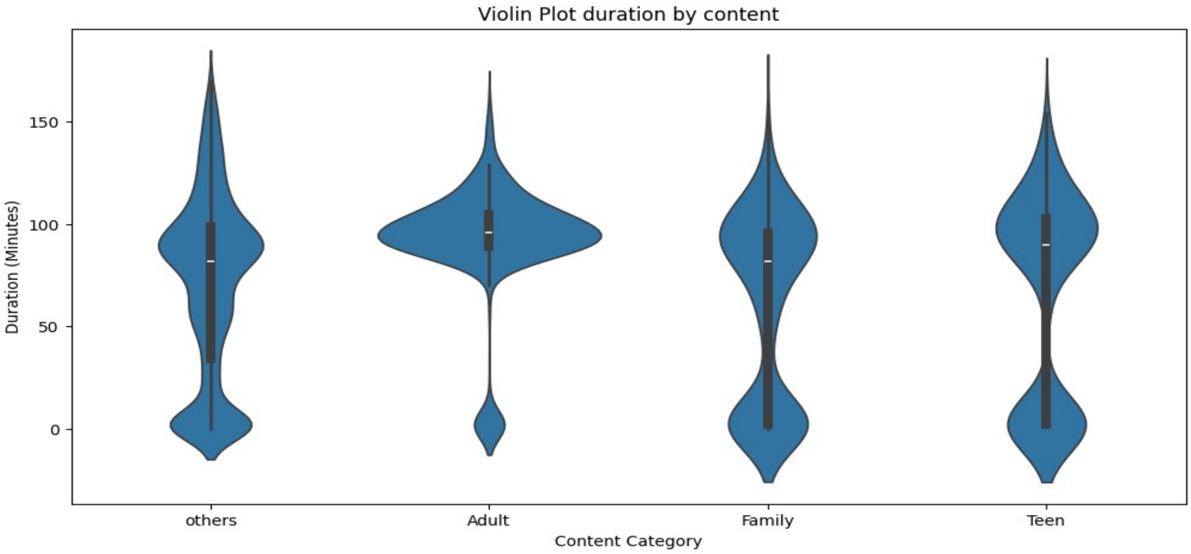
Correlation Heatmap :

A Heatmap is used to visualise correlation between numerical variables.



Violin Plot Analysis of Content Duration by Category :

This violin plot shows how the length of content varies in the following categories : Teen, Adult, Family, and Others. It demonstrates that there is a wider variety of durations with distinct peaks in the Others and Adult groups. The Family and Teen categories, on the other hand, have less diversity in their lengths with more of their durations centred around a core point.



- Please write the names of your group members.

Group member 1 : Lolla Aditya Srivatsav

Group member 2 : Sai Krishna Goud Valdas

Group member 3 :

- Rate each groupmate on a scale of 5 on the following points, with 5 being HIGHEST and 1 being LOWEST.

Evaluation Criteria	**Group member 1**	**Group member 2**	**Group member 3**
How effectively did your group mate work with you?	5	5	
Contribution in writing the report	5	5	
Demonstrates a cooperative and supportive attitude.	5	5	
Contributes significantly to the success of the project .	5	5	
TOTAL	20	20	

Also please state the overall contribution of your teammate in percentage below, with total of all the three members accounting for 100% (33.33+33.33+33.33 ~ 100%) :

Group member 1 : 50%

Group member 2 : 50%

Group member 3 :