Analysis of User Engagement on Netflix

1. Introduction

1.1 Project Overview

This report explores factors driving user engagement on Netflix, focusing on metrics such as content type, genre, and regional preferences.

1.2 Goals and Objectives

- **Identify Key Factors Influencing User Engagement**: Determine significant elements affecting viewer engagement.
- **Analyse Regional Variations**: Understand user engagement differences across regions.
- **Provide Recommendations**: Suggest improvements based on analysis results.

1.3 Roadmap

1. Introduction
2. Background
3. Specifications and Design
4. Data Overview
5. Implementation and Execution
6. Analysis
7. Key Findings
8. Conclusions and Recommendations
9. Future Work and Further Analysis

2. Background

The aim of this project is to explore the key factors that influence user engagement on netflix. (define, ratings, and watchtime on Netflix. By analysing these factors, we aim to gain deeper insights into user engagement, which can ultimately enhance content recommendations and improve viewer satisfaction.

This project provides insights into consumer preferences, enabling production teams to allocate funding and resources more effectively and distribute budgets in a manner that aligns with viewer interests.

The insights from this project can help marketing teams understand user engagement trends, allowing them to craft more impactful advertising campaigns tailored to a larger demographic, as well as Individuals responsible for selecting and creating content for the streaming platform, as well as analysts who aim to understand trends in content popularity and the factors influencing global media distribution.

3. Specifications and design

3.1 Requirements

- **Technical Requirements:**
    - A clean dataset with comprehensive data on Netflix content and user engagement metrics.
    - Tools for data cleaning, preprocessing, and analysis (e.g., Python, Pandas, NumPy, Matplotlib).
    - Integration with TMDb API for additional metadata and ratings using Python.

- **Non-Technical Requirements:**
    - Clear understanding of the goals and objectives.
    - Effective collaboration and communication among team members.
    - Consistent documentation of the analysis process and findings.

3.2 Design and Architecture

The design and architecture of the project involve:

- Data collection from multiple sources including Netflix engagement reports, Kaggle datasets, and TMDb API.
- Data cleaning and preprocessing to ensure the datasets are ready for analysis.

- Analytical processes to derive insights, including statistical analysis, visualization, and comparative analysis.

4. Data Overview

4.1 Information needed

- User engagement metrics from Netflix (e.g., viewings amount).
- Metadata about Netflix content (e.g., genres, release dates, country of origin).
- IMDb ratings for comparative analysis.

4.2 Information Available - Sources

- **Official Netflix Engagement Report:**
  Detailed watch hours from January to June 2023 from the Netflix engagement official report. While a dataset spanning a longer time period would have been more beneficial, the availability of watch time data was highly advantageous for our analysis.
  Netflix - What We Watched

- **Kaggle Netflix Titles Dataset:**
  Comprehensive dataset from Kaggle on Netflix content. Downloaded and integrated into the analysis pipeline. We leveraged a large dataset from Kaggle that offered information on Netflix content. This dataset included valuable details such as genres, release dates, countries of origin, and other important metrics. These elements were crucial for exploring the diversity and distribution of Netflix's content library.
  Netflix - Movies and TV Shows

- **TMDb API:**
  Used to fetch a vast library of movies and TV shows along with their IMDb ratings. This integration enabled us to perform comparative analyses by leveraging the extensive metadata and ratings provided by TMDb. By incorporating the exploratory analysis of ratings, we were able to agree on certain conclusions for the role of imdb ratings and user engagement with content.
  The Movie Database (TMDB) API

5. Implementation and Execution

5.1 Development Approach and Team Member Roles

The development approach involved iterative meetings to decide the topic according to the data we could find.

Analysis and refinement, with team members assigned specific tasks:

- Researching Data: Charlotte, Deena and Evelyn
- API integration to fetch data: Evelyn, Deena, Charlotte and Dosa, each of them worked in a different endpoint.
- Jupyter Notebook: Deena
- Data Cleaning: Deena.
- Analysis: Everyone.
- Report: Everyone.
- Presentation: Charlotte.

  Activity log

5.2 Tools and Libraries

The project utilized the following tools and libraries:

- Jupiter Notebook: For organisation and display of manipulation and analysis
- Pandas: For data manipulation and analysis.
- NumPy: For numerical computations.
- TMDb API: For retrieving movie and TV show metadata. (Python)
- Matplotlib and Seaborn: For data visualization.
- Github: For storing the repository, tracking, and collaboration of the project.

5.3 Implementation Process

Achievements:

- Successfully integrated and cleaned datasets from multiple sources.
- Identified key trends and insights regarding user engagement.

Challenges:

- Data Research, Finding the relevant data was the biggest challenge, as Netflix do not have a public API anymore, so we had to rely on datasets found in Kaggle.
- Handling missing data in critical columns like "Director" and "Cast".
- Ensuring consistency in genre categorization between Netflix and IMDb data.
- Different schedules and availability among members of the team.

Decision to Change Something:

- We decided to work in the entertainment industry, and we were thinking at the beginning about comparing Cinema releases versus Streaming platforms, however we were unable to find the relevant data to make that analysis.
- Modified the data imputation strategy for missing values to improve data integrity. Using mode, 'Unknown' to fill empty rows within data and have non null values.

Implementation challenges

- The API Key didn't work, so we had to use a token, the API Data was too big, the endpoint for Movies and Tv Shows popularity was designed to call per page, it was a big challenge to understand how many pages were and how many we can get per call.

6. Analysis

6.1 Data Cleaning and Preprocessing

The dataset contains 8,809 rows and 26 columns of data. Out of these 26 columns, 14 were completely blank and unnecessary for the analysis, so we removed them. This leaves us with 12 columns, consisting of integer, datetime and object/string type columns.

We encountered several missing values in the dataset:

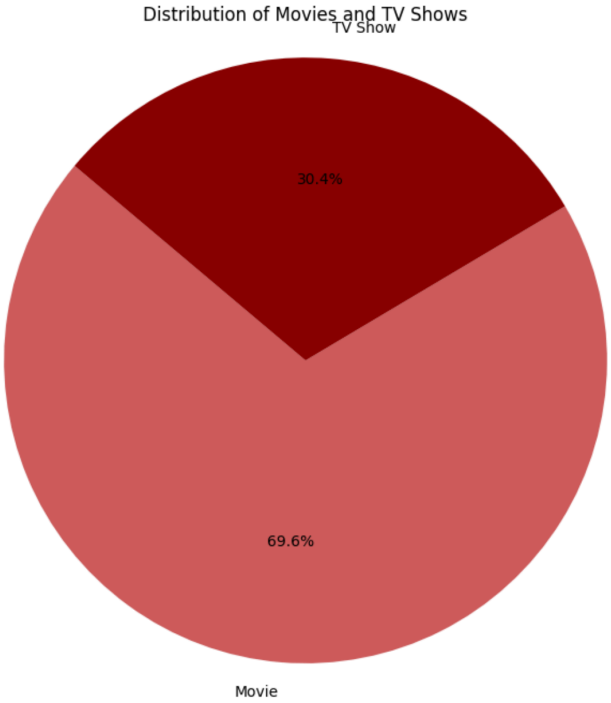
- The "Director" column had 30% missing data, which we filled with 'Unknown'.
- The "Cast" and "Country" columns each had 9% missing data, also replaced by 'Unknown'.

- The "Date Added" column had 0.12% missing data, which we filled using a combination of the "release_year" for the year and random values for the month and date.

7. Key Findings

7.1 Distribution of Content Types

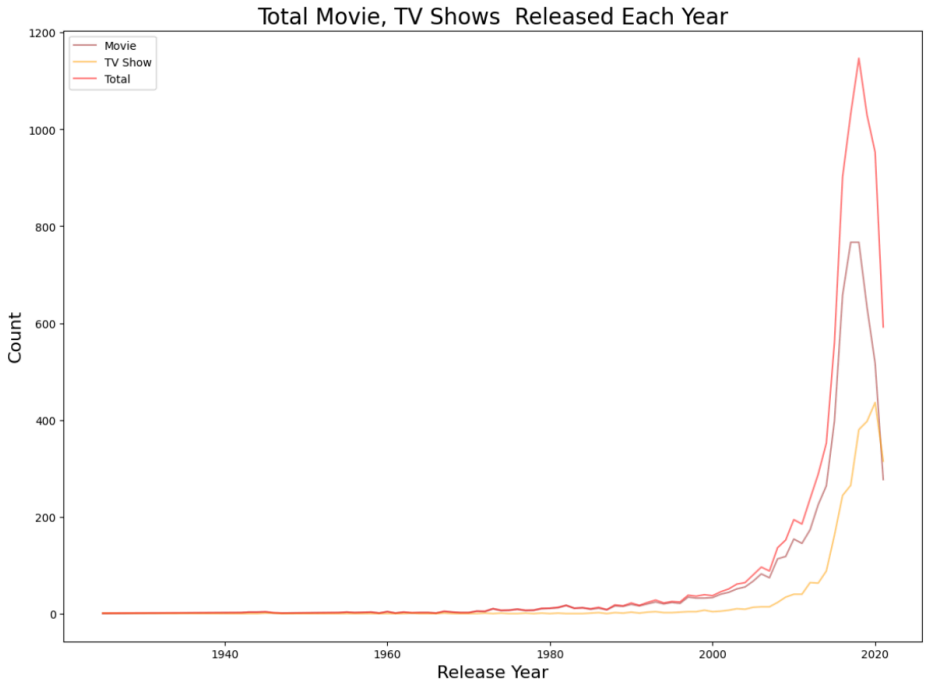
- **Movies**: 69.6% of the content.
- **TV Shows**: 30.4% of the content.



7.2 Popular Genres

- **Top Genres**: ‘Action & Adventure, Sci-Fri & Fantasy’, ‘Children & Family Movies, Comedies’ and ‘TV Dramas’
- **Insight**: Diversifying content in popular genres can enhance viewer engagement.

7.3 Release Year Trends



- **Observation**: Increasing number of releases in years leading up to 2019, but then a significant drop in 2020.
- **Insight**: Continuous growth in content production, indicating a strong content acquisition strategy. Clear impact of COVID-19 on content production in 2020.

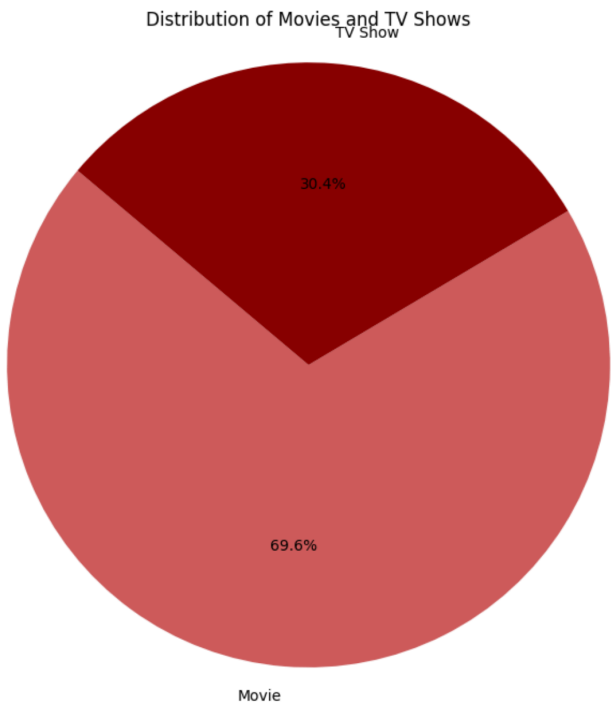
7.4 Regional Distribution

- **Top Countries**: United States, India, United Kingdom.
- **Insight**: Content produced by these counties can be more accessible globally and their popularity is further helped by their large populations. Focused content acquisition and marketing strategies in these regions can also drive higher engagement.

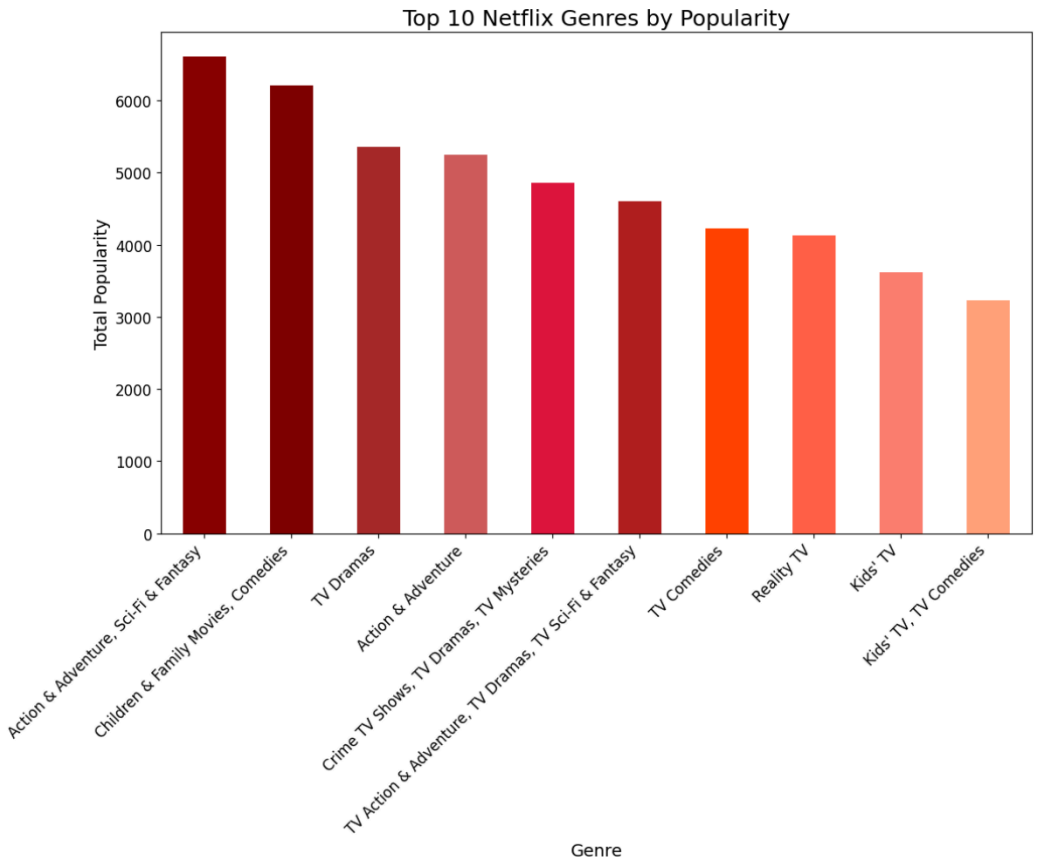
8. Conclusions and Recommendations

8.1 Key Findings

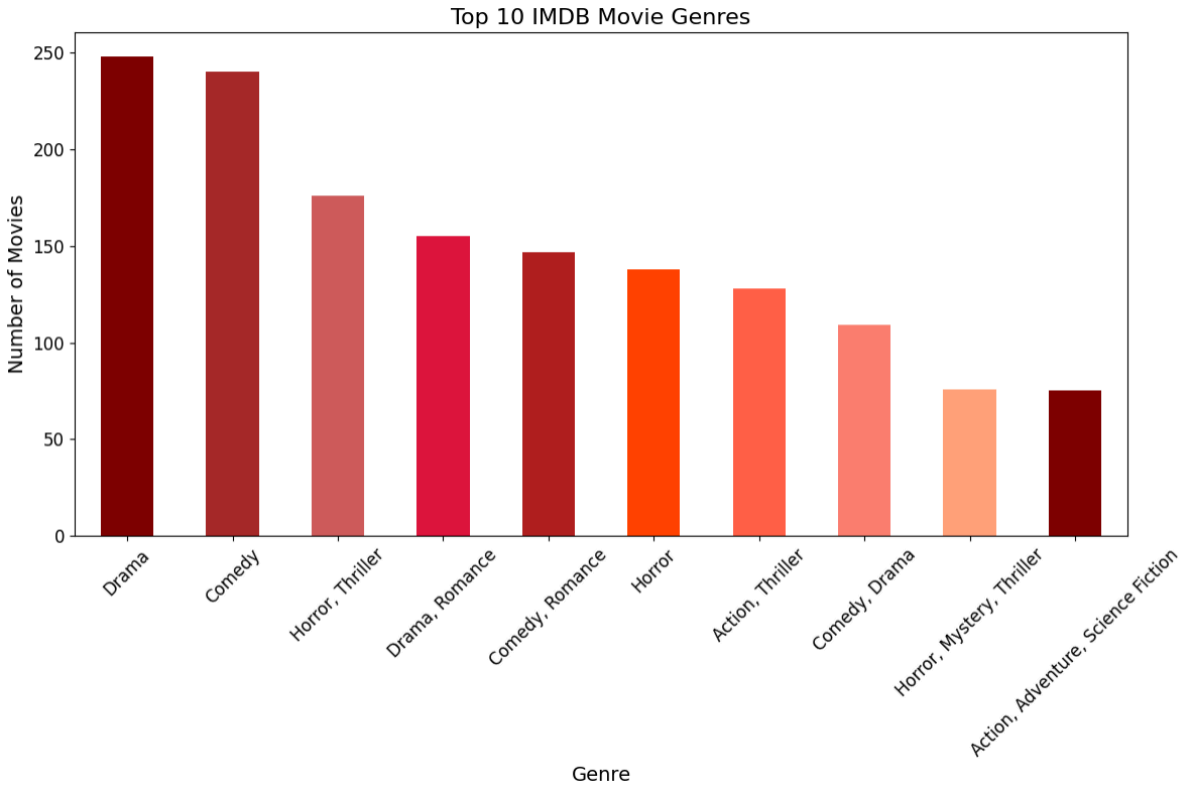
- **Content Types**: Movies dominate the content library.



- **Genres**: ‘Action & Adventure, Sci-Fi & Fantasy’, ‘Children & Family Movies, Comedies’ and ‘TV Dramas’ are the top three most popular Netflix Genres. While ‘Drama’, ‘Comedy’ and ‘Horror, Thriller’ are the top 3 for IMDB Movie genres.
    - Important to note: there is some overlap in how the genres are categorised in the Netflix and IMDB data. For example, in the Netflix bar graph, TV Dramas is mentioned once more in the 5th category ‘Crime TV Shows, TV Dramas, TV Mysteries’, however the close repetition of the genres further should only highlight their popularity even further, so it should impact the conclusion we deduce.
    - IMDb and Netflix do not share the same naming of categories which could impact the reliability of comparing these two datapoints together.
    - However, both charts share a high popularity in ‘Drama’ and ‘Comedy’ categories.

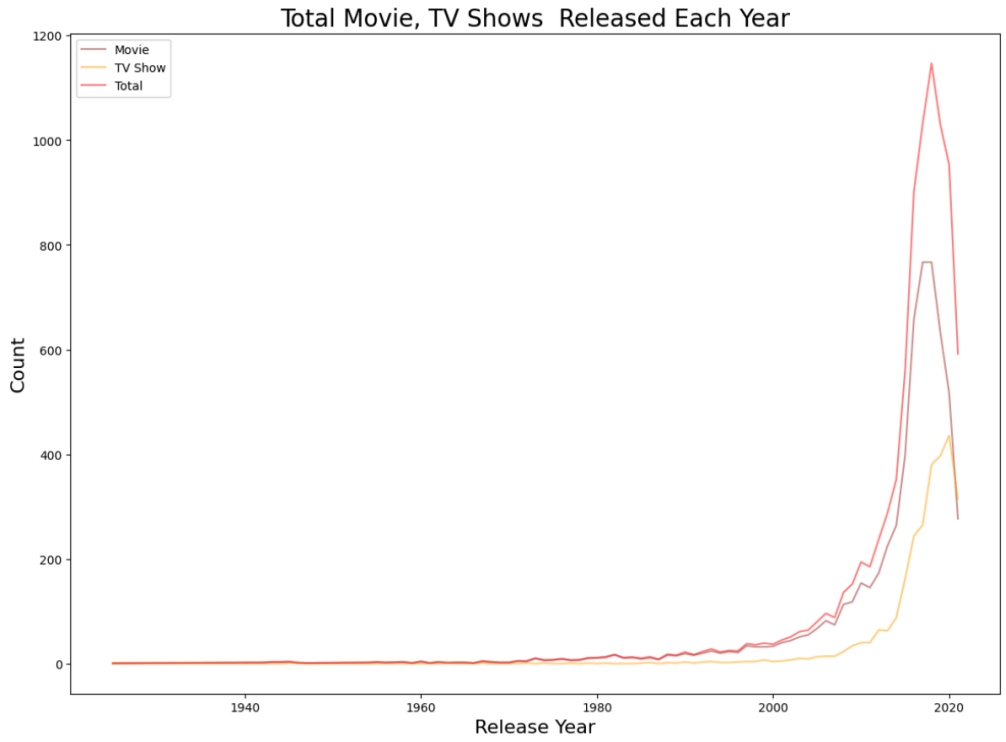

Top 10 Netflix Genres by Popularity
Total Popularity
0
1000
2000
3000
4000
5000
6000
Action & Adventure, Sci-Fi & Fantasy
Children & Family Movies, Comedies
TV Dramas
Action & Adventure
Crime TV Shows, TV Dramas, TV Mysteries
TV Action & Adventure, TV Dramas, TV Sci-Fi & Fantasy
TV Comedies
Reality TV
Kids' TV
Kids' TV, TV Comedies
Genre

-


Top 10 IMDB Movie Genres
Number of Movies
0
50
100
150
200
250
Drama
Comedy
Horror, Thriller
Drama, Romance
Comedy, Romance
Horror
Action, Thriller
Comedy, Drama
Horror, Mystery, Thriller
Action, Adventure, Science Fiction
Genre

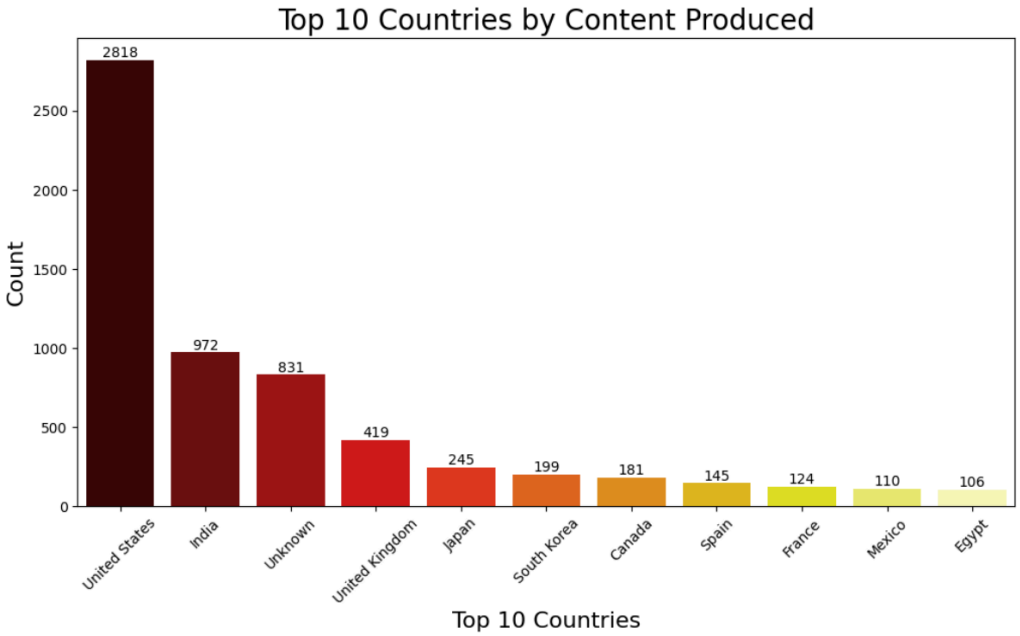
-

- **Release Trends**: An increase in releases, especially just prior to 2020 and a drop in 2020.

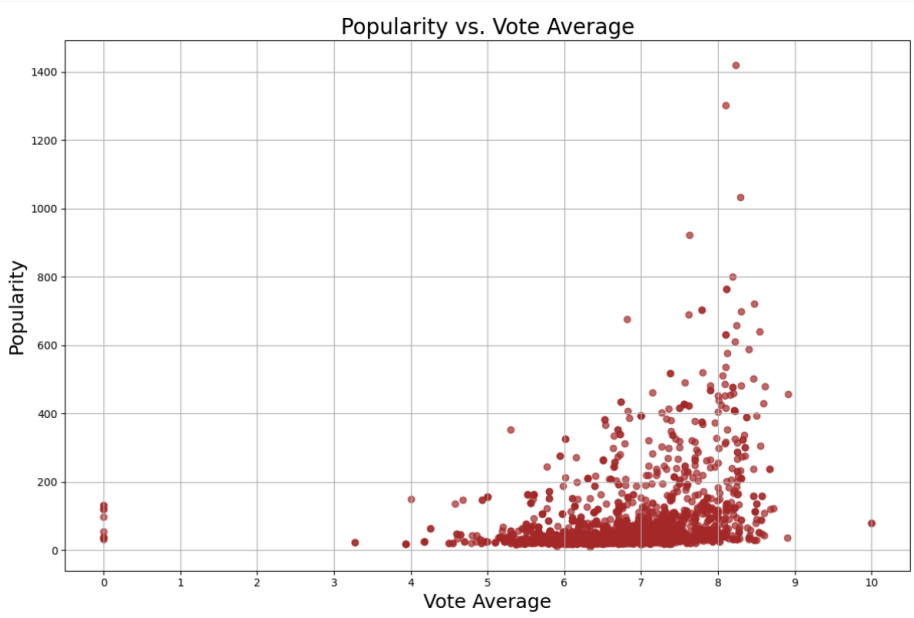


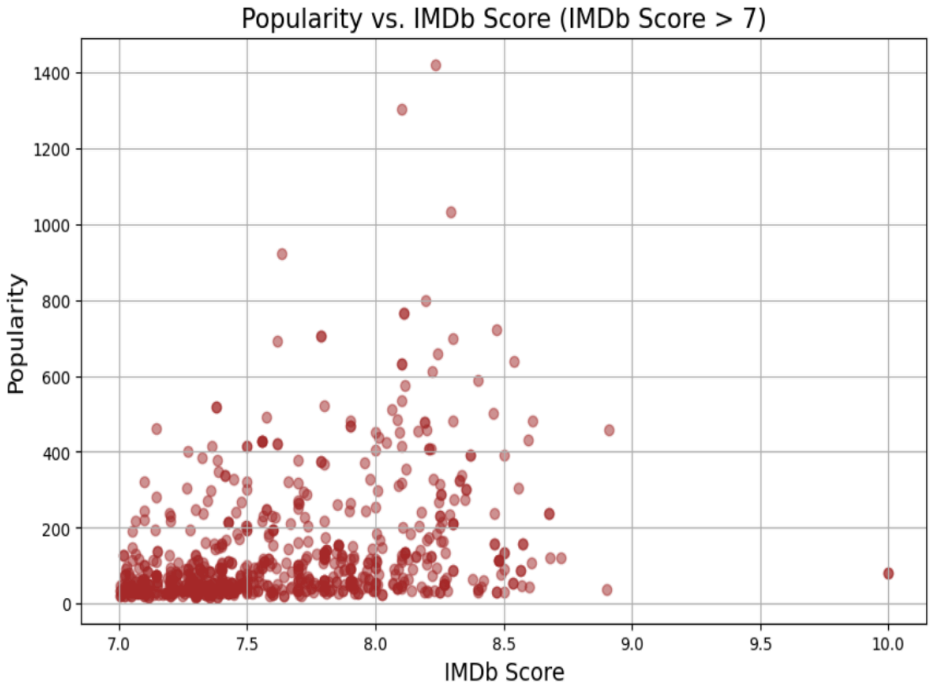
- **Regional Preferences**: The United States and India have the highest content counts. However, when looking at the data there is a significant number of content that does not include the country of production. Assuming that this content spans across multiple countries, we can categorize the content produced by the United Kingdom as producing a high amount of content in comparison to

other countries also.



- **IMDb rating:** Content with an IMDb rating of 7 or above are more popular.





Observing the scatter plot of popularity against IMDb scores, it becomes apparent that there's a noteworthy trend. Specifically, points on the plot where IMDb scores exceed 7 tend to exhibit higher levels of popularity. This observation underscores the significance of IMDb scores in determining content popularity. By filtering the data to include only points where the IMDb score exceeds 7 and plotting them on the scatter plot, we can clearly visualize the relationship between higher IMDb scores and increased popularity.

This insight sheds light on viewer preferences, suggesting that content with higher IMDb ratings tends to garner greater attention and engagement.

8.2 Recommendations

- **Personalised Recommendations**: Tailor content suggestions based on user preferences and regions.
- **Content Acquisition**: Focus on popular genres and emerging trends. For example, focusing on movies instead of TV shows considering their higher popularity, and tailoring to show more family friendly content to reach a wider audience (following the popularity of 'Children & Family Movies, Comedies'))
- **Engagement Strategies**: Implement features like notifications for new releases in favourite genres.

8.3 Future Work

- **Enhanced Models**: Incorporate more data sources and advanced techniques.
- **Regional Analysis**: Conduct deeper analysis of user preferences across different regions.

9. Future Work and Further Analysis

If more time were available, several additional analyses could be performed to gain deeper insights into the Netflix dataset. Here are some ideas:

9.1 Detailed Genre Analysis

- **Objective**: Understand the popularity and distribution of different genres within movies and TV shows.
- **Steps**:
    - Analyse the distribution of genres within each content type.
    - Identify the most and least common genres.
    - Plot the genre distribution using bar charts or word clouds.

9.2 Temporal Analysis

- **Objective**: Analyse the release trends over time.
- **Steps**:
    - Plot the number of movies and TV shows released each year.
    - Identify trends such as peaks in release years or changes in content type preferences over time.
    - Examine the relationship between content release dates and Netflix's business milestones.
    - Identify the changes in genre popularity and duration of content prior to 2020 and during/after 2020.
    - Consider the drop in content creation in 2020 and what content was being watched to ‘replace’ this. E.g. Was there a resurgence of older TV content? If so, from when was the most popular content produced in 2020?

9.3 Duration Analysis

- **Objective**: Investigate the typical duration of movies and TV shows.
- **Steps**:
    - Calculate summary statistics (mean, median, mode) for the duration of movies.
    - Examine the distribution of TV show seasons/episodes.
    - Identify any trends or patterns in content duration over time.

9.4 Content Rating Analysis

- **Objective**: Explore the distribution of content ratings (e.g., PG, R, TV-MA) and their implications.
- **Steps**:
    - Analyse the distribution of different content ratings.
    - Identify which types of content (movies vs. TV shows) dominate each rating category.
    - Investigate any patterns or trends in content ratings over time.

9.5 Comparative Analysis

- **Objective**: Compare Netflix's content with other streaming platforms such as Amazon Prime, Disney etc.
- **Steps**:
    - Gather data from other streaming services.

- Perform comparative analyses on content type, genre distribution, release trends, etc.
- Identify Netflix's unique strengths and areas for improvement.