|  |  |  |
| --- | --- | --- |
| Checkpoint I | Checkpoint I: Project Proposal | |
| Group: | 30 |
| Date: | 2024/09/17 |
|  |  |

# Problem Domain

We will study the growth of Netflix content over the years. This could include the movies produced by Netflix or the number of TV shows available on the platform. We think this topic can be very interesting to understand how our way of consuming content, be it movies or TV shows, has changed over time. Creating this visualization will help us understand the role Netflix plays in this growing trend of streaming.

# Task Abstraction

Our project aims to dive into various aspects of Netflix's evolution over the years, focusing on both the platform's user base and its content catalogue. Specifically, we will explore the following key questions:

## Example Questions

1. **Top Performing Content**:

**Question**: What are the top 10 movies or TV shows of all time on Netflix, and in which years were they released?

**Type**: *rank over time*

**Description:** The aim of this question is to show in a concise way which content performs better on the platform and to be able to tune our search to specific years or specific categories. This makes for a very interesting and interactive table and can answer many questions, such as what are the top 10 action movies of 1999.

1. **Netflix Original Productions:**

**Question**: How many Netflix movies and Netflix TV shows are released each year, and what trends are evident in their production volume and reception?

**Type**: *trends, time intervals*

**Description**: We thought about this question because we are almost overwhelmed by the amount of content that is released every year, every month, almost every week! So, this question is to put into perspective the rush of Netflix production that goes on throughout the year. More and more content is being created by Netflix and we are not able to watch it all. Is this a reality? We want to find out.

1. **Trends in Content Addition Over Time:**

**Question:** What are the trends in the number of movies and TV shows added to Netflix over time? Are there periods of significant increase or decrease, and what might be the reasons behind them?

**Type:** *trends/time intervals*

**Description:** We noticed that the content added to Netflix is not uniform, some 1990 movies can be added in 2024, what is happening here? Does Netflix wait to pay less to have the right to diffuse the movie or is there another reason behind it, this visualization could help us understand this phenomenon better.

1. **Ratio of Movies to TV Shows Over Time:**

**Question:** How does the ratio of movies to TV shows change over time? Does this ratio correlate with shifts in viewer preferences or strategic decisions by Netflix?

**Type:** type *proportion, correlation and trends*

**Description:** With this question in mind, we want to put into perspective what kind of content Netflix is putting forward the most. What makes more sense to them, is a short TV show better than a good film? And does their strategy change over the years?

1. **Evolution of Country Contributions:**

**Question:** Which countries have contributed the most content to Netflix over the years, and how has this trend changed? What does this indicate about Netflix's global expansion and content acquisition strategies?

**Type:** *geolocation, trends and distribution*

**Description:** With this question we expect to see the US come first thanks to Hollywood but is there a new trendy film or TV show production location, like maybe Bollywood. We want to see if the TV show are also mostly produced in the US or if it is not the case then where?

# Data

We decided to choose [*“Latest Netflix data with 26+ joined attributes”*](https://www.kaggle.com/datasets/ashishgup/netflix-rotten-tomatoes-metacritic-imdb) available on Kaggle, it currently provides lots of data and allows us to answer mainly all of our questions.

## Initial Dataset

## The initial dataset is quite extensive, containing 15,071 unique values and comprising 29 different attributes. The size of the dataset exceeds 14MB, and it encompasses every Netflix movie or TV show added to the platform from 2015 to 2021, so this is a static dataset. Given the vast amount of data, not all elements are useful for our analysis. For instance, attributes such as the link for an image do not contribute to answering our research questions or enhancing our visualizations.

## To optimize our analysis and ensure that our visualizations are as responsive and informative as possible, we conducted a cleanup of the dataset. This process involved identifying and removing any extraneous or irrelevant elements. By focusing on the most pertinent attributes, we aim to create a more manageable and insightful dataset that will facilitate effective analysis and visualization. This refined dataset will allow us to explore trends, patterns, and other insights related to Netflix content over the specified period.

## Selected/Derived Data

The final dataset is a **tabular dataset**, where each row represents a unique movie or series, and each column represents an attribute associated with that title. The dataset includes a mix of nominal, continuous, and ratio variables, making it suitable for various types of analysis, including statistical computations and time-based evaluations.

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Semantics |
| Title | Nominal | The name of the movie or series. |
| Genre | Nominal | The genre(s) the movie or series belongs to (e.g., Drama, Comedy). |
| Languages | Nominal | The languages in which the title is available. |
| Series or Movie | Nominal | Indicates whether the title is a movie or a series. |
| Country Availability | Nominal | The countries where the title is available on Netflix. |
| Director | Nominal | The director(s) of the movie or series. |
| Writer | Nominal | The writer(s) of the movie or series. |
| Actors | Nominal | The main actors featured in the title. |
| IMDb Score | Ratio | The IMDb rating of the title, ranging from 0 to 10. |
| Rotten Tomatoes Score | Ratio | The Rotten Tomatoes rating as a percentage, ranging from 0% to 100%. |
| Metacritic Score | Ratio | The Metacritic score of the title, ranging from 0 to 100. |
| Release Date | Continuous | The original release date of the title, represented as a point on a continuous time scale. |
| Netflix Release Date | Continuous | The date when the title became available on Netflix, represented as a point on a continuous time scale. |
| Average Score | Ratio | The average of the IMDb, Rotten Tomatoes, and Metacritic scores, providing a unified measure of critical reception. |
| Days Until Netflix Release | Ratio | The number of days between the original release date and the Netflix release date, indicating the time lag in availability on the platform. |

## Data Processing

To clean it we used pandas and removed the following columns “*['Tags', 'Runtime', 'View Rating', 'Production House', 'Netflix Link', 'IMDb Link', 'Summary', 'IMDb Votes', 'Poster', 'TMDb Trailer', 'Trailer Site', 'Hidden Gem Score', 'Awards Received', 'Awards Nominated For', 'Boxoffice']”* those appeared to be not necessary as of now. We might delete few more columns, but we are still thinking of visualization ideas so for example the Poster of the movie might be useful if we plan on doing a line chart with a onMouseClick event where we could display some movie statistics.

We are also facing an issue using blank values

## Data Abstraction

For this data set we created only one derived measure: the average score. This data is useful because we don’t always have the data for all scores (IMDb Score, Rotten Tomatoes Score or Metacritic Score) and with this average score we are able to answer questions about score even if some data is missing. For now we don’t see the use of other derived measure but we will create new one if needed.

## Mapping (Data sample/Questions)

**Top Performing Content**: *What are the top 10 movies or TV shows of all time on Netflix, and in which years were they released?*

For this question we will need to use the “Title”m “Series or Movie”, “IMDB Score”, “Rotten Tomatoes Score”, “Release Date”, “Netflix Release Date”. Note that we could use the “Image Field” to visualize some other statisitcs of the movies, such as the Director, Actors and others statistics such as the country avalaibility.

**Netflix Original Productions::** How many produced movies and TV shows are available in each country, and what trends are evident in their availability and reception across different regions?

For this question we will use the following fields “Genre”, “Country Availability”, “Series or Movie”, again we might want to add other fields for vizualization purposes.

**Trends in Content Addition Over Time**: *What are the trends in the number of movies and TV shows added to Netflix over time? Are there periods of significant increase or decrease, and what might be the reasons behind them?*

We will use the “Release Date”, “Netflix Release Date” and “Series or Movie” fields.

**Ratio of Movies to TV Shows Over Time**: *How does the ratio of movies to TV shows change over time? Does this ratio correlate with shifts in viewer preferences or strategic decisions by Netflix?*

We will use “Release Date”, “Netflix Release Date” and “Series or Movie” fields.

**Evolution of Country Contributions::** Which countries have contributed the most content to Netflix over the years, and how has this trend changed? What does this indicate about Netflix's global expansion and content acquisition strategies?