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

# Problem Domain

We are going to study the growth of Netflix over the years. This might include the movies produced by Netflix or the number of TV show accessible in the platform. We find that this subject may be very interesting to understand the how our way of consuming content, movies or TV show, has change over time. Since Netflix is one of the main players of this game and we wish to understand how.

# 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 catalog. Specifically, we will explore the following key questions:

## Example Questions

1. **Top Performing Content**: *What are the top 10 movies or TV shows of all time on Netflix, and in which years were they released?*
2. **Netflix Original Productions**: *How many movies and TV shows are released each year, and what trends are evident in their production volume and reception?*
3. **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?*
4. **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?*
5. **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?*

# 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

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris quis tellus bibendum, aliquam mi et, fermentum neque. Sed et lectus at purus consequat accumsan eget nec risus. Suspendisse in luctus enim, vel fermentum mi. Etiam egestas efficitur mauris, sit amet iaculis purus euismod eu. Suspendisse eget ligula ut lacus ullamcorper posuere sit amet eget purus. Nullam et nunc et sem facilisis suscipit. Cras scelerisque congue ex, vitae luctus lectus elementum sed. Duis auctor, turpis quis dictum congue, justo purus tempus arcu, eu elementum quam tortor id sem. Suspendisse maximus, lorem sed auctor euismod, odio lorem tincidunt libero, ut viverra velit leo ut lectus.

## Selected/Derived Data

# Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris quis tellus bibendum, aliquam mi et, fermentum neque. Sed et lectus at purus consequat accumsan eget nec risus. Suspendisse in luctus enim, vel fermentum mi. Etiam egestas efficitur mauris, sit amet iaculis purus euismod eu. Suspendisse eget ligula ut lacus ullamcorper posuere sit amet eget purus.

## 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']”,* 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

# Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris quis tellus bibendum, aliquam mi et, fermentum neque. Sed et lectus at purus consequat accumsan eget nec risus. Suspendisse in luctus enim, vel fermentum mi. Etiam egestas efficitur mauris, sit amet iaculis purus euismod eu. Suspendisse eget ligula ut lacus ullamcorper posuere sit amet eget purus.

## 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 Availability:** 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.