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 9 action movies or TV shows between 2014-2016 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.
2. **Netflix Original Productions:**
   Question: How many Netflix movies and Netflix TV shows were released in 2018?
   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 years. 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.
3. **Trends in Content Addition Over Time:**
   Question: How many romantic movies or romantic TV show were added between 2017-2019?
   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.
4. **Evolution of Country Content:**
   Question: Does the USA have the most action movies and TV shows available on Netflix?
   Type: *geolocation, trends and distribution*
   Description: We all notice when we go on holiday to a country that the movies or TV shows available on Netflix where different. Does one country have more content available than another? Is the content well distributed? With this map we will be able to answer these kinds of questions at a glance.

5. **Comparison between countries**

   Question: Do northern European countries have more comedy Netflix produced content on Netflix than southern countries?

   Type: *geolocation and distribution*

   Description: With this question we intend to show how the Netflix content is distributed around the world. Is there country that are left out?

Data

We decided to choose *"Latest Netflix data with 26+ joined attributes"* 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 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. 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.
Actors	Nominal	The main actors featured in the title.
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', 'Writer']"* those appeared to be not necessary as of now.

Then we **dropped** all lines that were partially blanks or empty, apart in three columns which are ["IMDb Score", "Rotten Tomatoes Score", "Metacritic Score"] because we applied a different formula here, we do not directly use those columns but we use **Average Score** which was created by doing : *"df_cleaned['Average Score'] = df_cleaned[['IMDb Score', 'Rotten Tomatoes Score', 'Metacritic Score']].mean(axis=1, skipna=True)"*

Which does exactly what we wanted, taking the avrage of those 3 columns but skipping NA, or blank values. After computing this value, we removed the three which were no longer needed.

So, at the end our data set contains 9'614 lines, this will help our visualization to react way more faster!

Data Abstraction

For this data set we created two derived measures: 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.

We also created Days Until Netflix Release which is the number of days between the original release date and the Netflix release date.

Mapping (Data sample/Questions)

Top Performing Content:
What are the top 9 action movies or TV shows between 2014-2016 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", "Genre". 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 Netflix movies and Netflix TV shows were released in 2018?

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:

How many romantic movies or romantic TV show were added between 2017-2019?

We will use the "Release Date", "Netflix Release Date", "Genre" and "Series or Movie" fields.

Ratio of Movies to TV Shows Over Time:

How does the ratio of movies to TV shows change over the 2015-2020 time period?

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

Evolution of Country Contributions:
Does the USA have the most action movies and TV shows available on Netflix?

We will use the “Country availability”, “Genre” and “Netflix release date”