Trends in film and television –Netflix

AUTHOR
Yifan Chen, Xinran Bi, Haibo Sun, Ruiwen Zhang

PUBLISHED
December 4, 2024

Introduction

Major Questions

This project aims to analyze Netflix's content dataset to answer key questions about its trends and composition. Specifically, we are seeking answers to the following questions:
- How has the dominance of genres changed by the year?
- How are the top 5 genres distributed across the top 5 countries with the most Netflix titles?
• What are the trends in TV shows and movies over the years? What might be contributing to this trend?

Distribution of Genres (Pastel Morandi Colors)
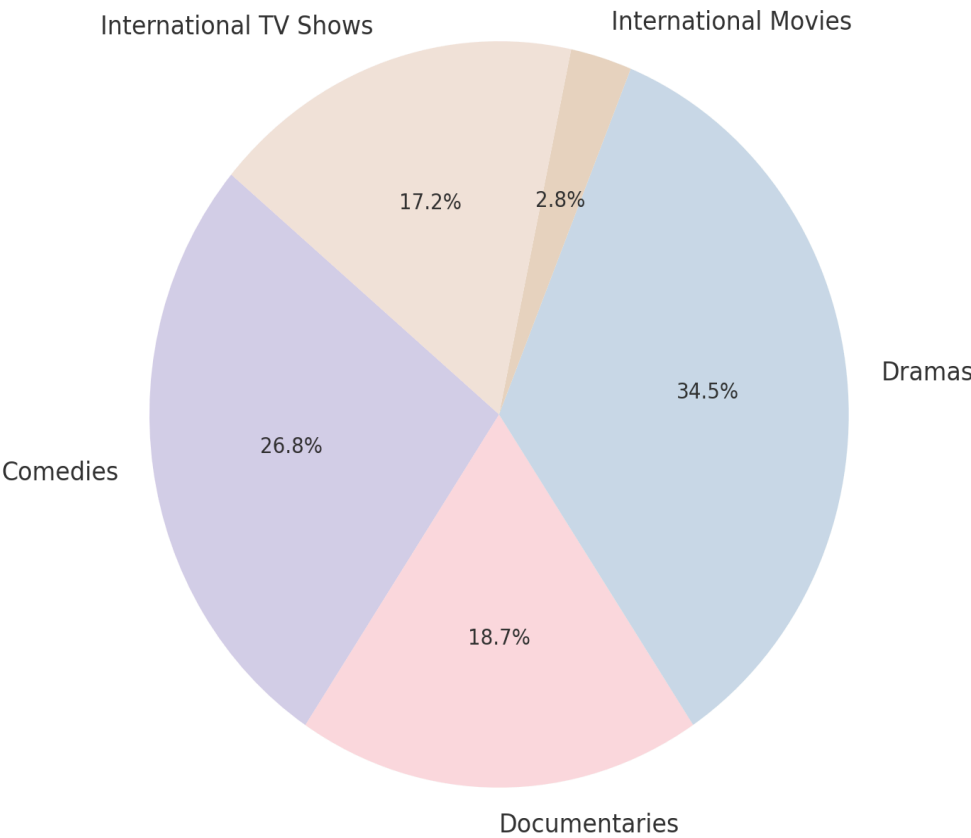
Data Source

To explore the above questions, we used a publicly available dataset from Kaggle:

1. Netflix movies and TV shows:
   https://www.kaggle.com/datasets/shivamb/netflix-shows2
2. Netflix Dataset:
   https://www.kaggle.com/datasets/imtkaggleteam/netflix

Both datasets provide structured information about the content available on Netflix, allowing us to explore trends in the diversity and volume of titles on Netflix.

Ethical Questions and Limitations

Data privacy:

These datasets are publicly available and aggregate information about Netflix's catalog, so they do not violate user privacy. However, this data may not always be up to date or accurate. #### Limitations: Data sets may not reflect real-time data or provide complete regional information, as Netflix frequently updates its productions. #### Analytical bias: Certain regions or types may be underrepresented in the data, which may affect the results of the analysis.

3 Questions with Interactive Visualizations

Loading Data

```
library(ggplot2)
library(plotly)
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ─────────────────────── tidyverse 2.0.0 ──
✓ forcats    1.0.0      ✓ stringr    1.5.1
✓ lubridate 1.9.3      ✓ tibble     3.2.1
✓ purrr      1.0.2      ✓ tidyr      1.3.1
✓ readr      2.1.5

── Conflicts ────────────────────────────────── tidyverse_conflicts() ──
✗ dplyr::filter() masks plotly::filter(), stats::filter()
✗ dplyr::lag()     masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```r
library(stringr)
netflix_df <- read.csv(("C:/Users/86130/Desktop/作业/info201/cleand_data.csv"))
head(netflix_df)
```

```
     type                               title         director
1 TV Show                                3%
2   Movie                               1920     Vikram Bhatt
3   Movie                         3 Heroines   Iman Brotoseno
4   Movie Blue Mountain State: The Rise of Thadland   Lev L. Spiro
5 TV Show                       Blue Planet II
6   Movie                          Blue Ruin Jeremy Saulnier

cast
1 João Miguel, Bianca Comparato, Michel Gomes, Rodolfo Valente, Vaneza Oliveira, Rafael
Lozano, Viviane Porto, Mel Fronckowiak, Sergio Mamberti, Zezé Motta, Celso Frateschi
2                           Rajneesh Duggal, Adah Sharma, Indraneil Sengupta,
Anjori Alagh, Rajendranath Zutshi, Vipin Sharma, Amin Hajee, Shri Vallabh Vyas
3
Reza Rahadian, Bunga Citra Lestari, Tara Basro, Chelsea Islan
4                 Alan Ritchson, Darin Brooks, James Cade, Rob Ramsay, Chris Romano,
Frankie Shaw, Omari Newton, Ed Marinaro, Dhani Jones, Ed Amatrudo, Jimmy Tatro
5
David Attenborough
6         Macon Blair, Devin Ratray, Amy Hargreaves, Kevin Kolack, Eve Plumb, David W.
Thompson, Brent Werzner, Stacy Rock, Sidné Anderson, Sandy Barnett, Bonnie Johnson
           country date_added release_year rating duration.x
1           Brazil  14-Aug-20         2020  TV-MA          4
2            India  15-Dec-17         2008  TV-MA        143
3        Indonesia   5-Jan-19         2016  TV-PG        124
4    United States   1-Mar-16         2016      R         90
5   United Kingdom   3-Dec-18         2017   TV-G          1
6 United States, France  25-Feb-19     2013      R         90
                                           genres
1 International TV Shows, TV Dramas, TV Sci-Fi & Fantasy
2        Horror Movies, International Movies, Thrillers
3          Dramas, International Movies, Sports Movies
4                                          Comedies
5      British TV Shows, Docuseries, Science & Nature TV
6                      Independent Movies, Thrillers
```

Question 1

Question: How has the dominance of genres changed by the year?

Code and Output:

```r
analyze_genres <- function(data, genres) {
  netflix_df %>%
    separate_rows(!!sym(genres), sep = ", ") %>%
    count(!!sym(genres), sort = TRUE) %>%
    rename(Genre = !!sym(genres), Count = n)
}
genre_analysis <- analyze_genres(netflix_df, "genres")
head(genre_analysis)
```

```
# A tibble: 6 × 2
  Genre                 Count
  <chr>                 <int>
1 International Movies    2437
2 Dramas                 2106
3 Comedies               1471
4 International TV Shows  1199
5 Documentaries           786
6 Action & Adventure      721
```

```r
netflix_df <- netflix_df %>%
  mutate(genres = str_split(genres, ", ") %>% map_chr(~ .x[1]))

netflix_df_filtered <-
  netflix_df |>
  filter(genres == "International Movies"|genres == "Dramas"|genres == "Comedies"|genres == "I
genre_year_data <- netflix_df_filtered %>%
  filter(!is.na(release_year) & !is.na(genres)) %>%
  group_by(release_year, genres) %>%
  summarise(count = n(), .groups = "drop")
genre_year <- plot_ly(
  genre_year_data,
  x = ~release_year,
  y = ~count,
  color = ~genres,
  type = 'scatter',
  mode = 'lines+markers'
) %>%
  layout(
    title = "Trends of Genres Over the Years",
    xaxis = list(title = "Released Year"),
    yaxis = list(title = "Count of Genres"),
    legend = list(title = list(text = "Genres"))
  )
genre_year
```

Trends of Genres Over the Years

200   Genres
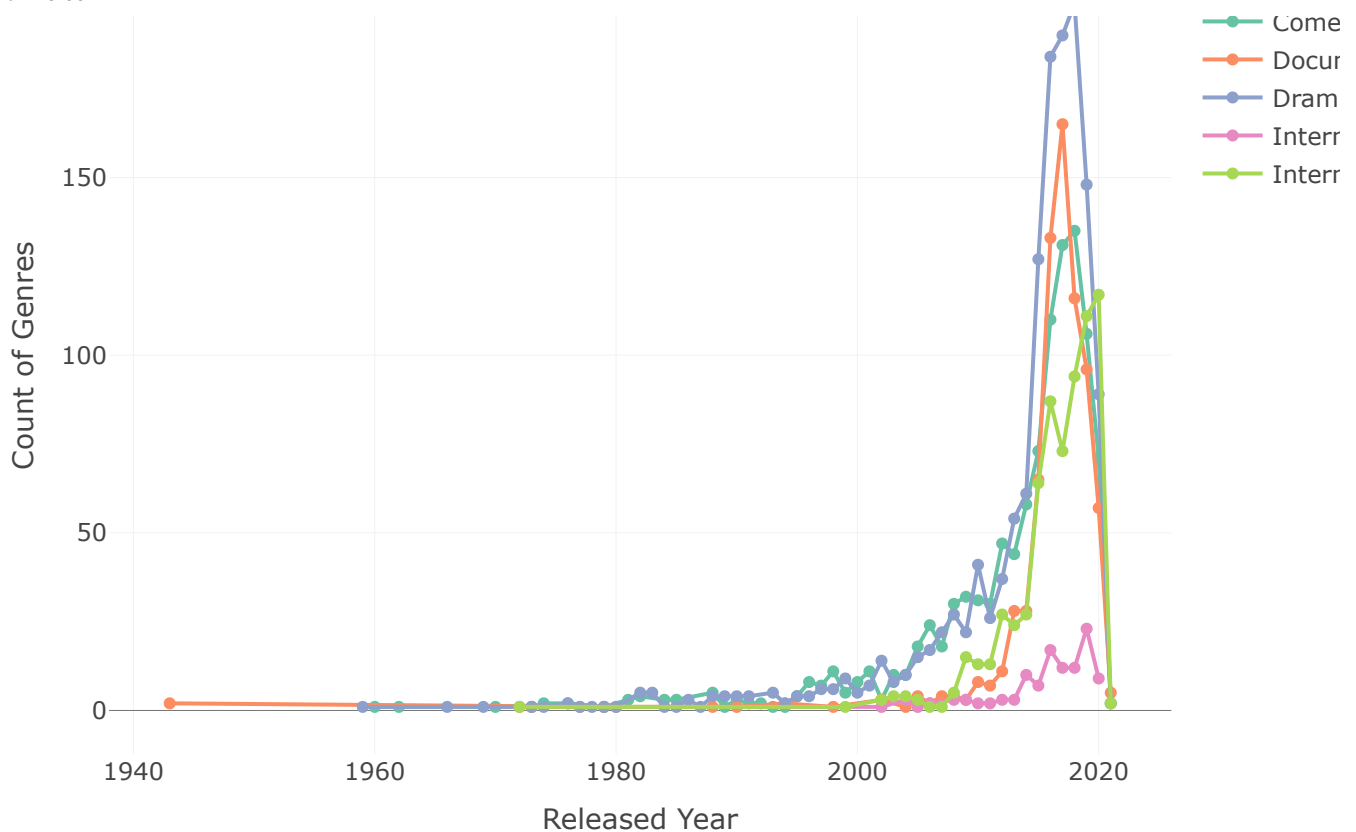
Brife Description:

We first made a new table that looked at the amount of all genres in the data and then selected the top five most representative genres based on the table.

From this interactive visualization graph, the comedies in 2012 are dominant compared to other genres on Netflix. Then, the dominant genre changed in 2018, dramas have dominance on Netflix. For dramas, it gradually increased from around 1995. This graph highlights how the popularity of different genres changed over the years. From 2008, the count of the top 5 genres has all rapidly boosted. Also, the trend of documentaries surged rapidly since the requirements for this type of genre's demand increased. The count of documentaries, dramas, and comedies is rapidly alleviated because of the pandemic covid-19 and also the lack of thoroughness of data collected from 2018.

The reason why we decided to make a new interactive visualization graph for the top five genres is that it can showcase a clearer and more concise trend of the diversification of the count of the top 5 genres because of the release year.

Question 2

Question: How are the top 5 genres distributed across the top 5 countries with the most Netflix titles?

Code and Output:

```
library(dplyr)
library(plotly)

# Step 1: Specify the top 5 genres globally
```

```r
top_5_genres <- c("International Movies", "Dramas", "Comedies", "International TV Shows", "Docu

# Step 2: Filter data for the top 5 countries and the specified top 5 genres
top_countries <- netflix_df %>%
  mutate(country = str_trim(country)) %>%
  filter(!is.na(country) & country != "") %>%
  count(country, sort = TRUE) %>%
  slice(1:5)

country_genre_data <- netflix_df %>%
  filter(
    country %in% top_countries$country &
      genres %in% top_5_genres
  ) %>%
  count(country, genres, name = "count")

# Step 3: Create interactive stacked bar chart
plot_ly(
  country_genre_data,
  x = ~country,
  y = ~count,
  color = ~genres,
  type = 'bar',
  text = ~paste("Country:", country, "<br>Genre:", genres, "<br>Count:", count),
  hoverinfo = "text"
) %>%
  layout(
    title = "Top 5 Countries with Most Netflix Titles by Top 5 Genres",
    xaxis = list(title = "Country"),
    yaxis = list(title = "Number of Titles"),
    barmode = 'stack',
    legend = list(title = list(text = "Genres"))
  )
```
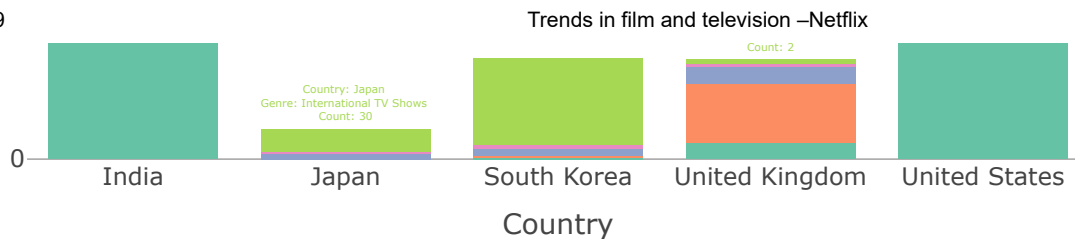
Top 5 Countries with Most Netflix Titles by Top 5 Genres

Brife Description:

The chart shows how Netflix titles are distributed across the top five countries with the most titles, divided into five main genres: International Movies, Dramas, Comedies, International TV Shows, and Documentaries. The United States has the most titles in all genres, showing its significant role in Netflix's library. India comes second, with many International Movies and TV Shows showing Netflix's focus on regional content in that country. Japan, South Korea, and the United Kingdom have smaller libraries, but their titles are spread across the genres.

International Movies and TV Shows are the most popular genres globally, showing Netflix's effort to provide content for diverse audiences. Dramas and Documentaries are also crucial, especially in the U.S. and India. This chart highlights Netflix's focus on the U.S. and India as key markets while showing room for growth in places like Japan and South Korea, which have strong entertainment industries.

Question 3

Question:

Code and Output:

```
trend_data <- netflix_df %>%
  filter(!is.na(release_year)) %>%
  group_by(release_year, type) %>%
  summarise(count = n(), .groups = 'drop')

trend_plot <- plot_ly(
  trend_data,
  x = ~release_year,
  y = ~count,
  color = ~type,
  type = 'scatter',
  mode = 'lines+markers',
  text = ~paste("Year:", release_year, "<br>Type:", type, "<br>Count:", count),
  hoverinfo = "text"
) %>%
  layout(
    title = "Trends of TV Shows and Movies Over the Years",
    xaxis = list(title = "Release Year"),
    yaxis = list(title = "Count"),
    legend = list(title = list(text = "Type"))
  )

trend_plot
```

```
Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested
palette with 3 different levels
Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested
palette with 3 different levels
```

Trends of TV Shows and Movies Over the Years
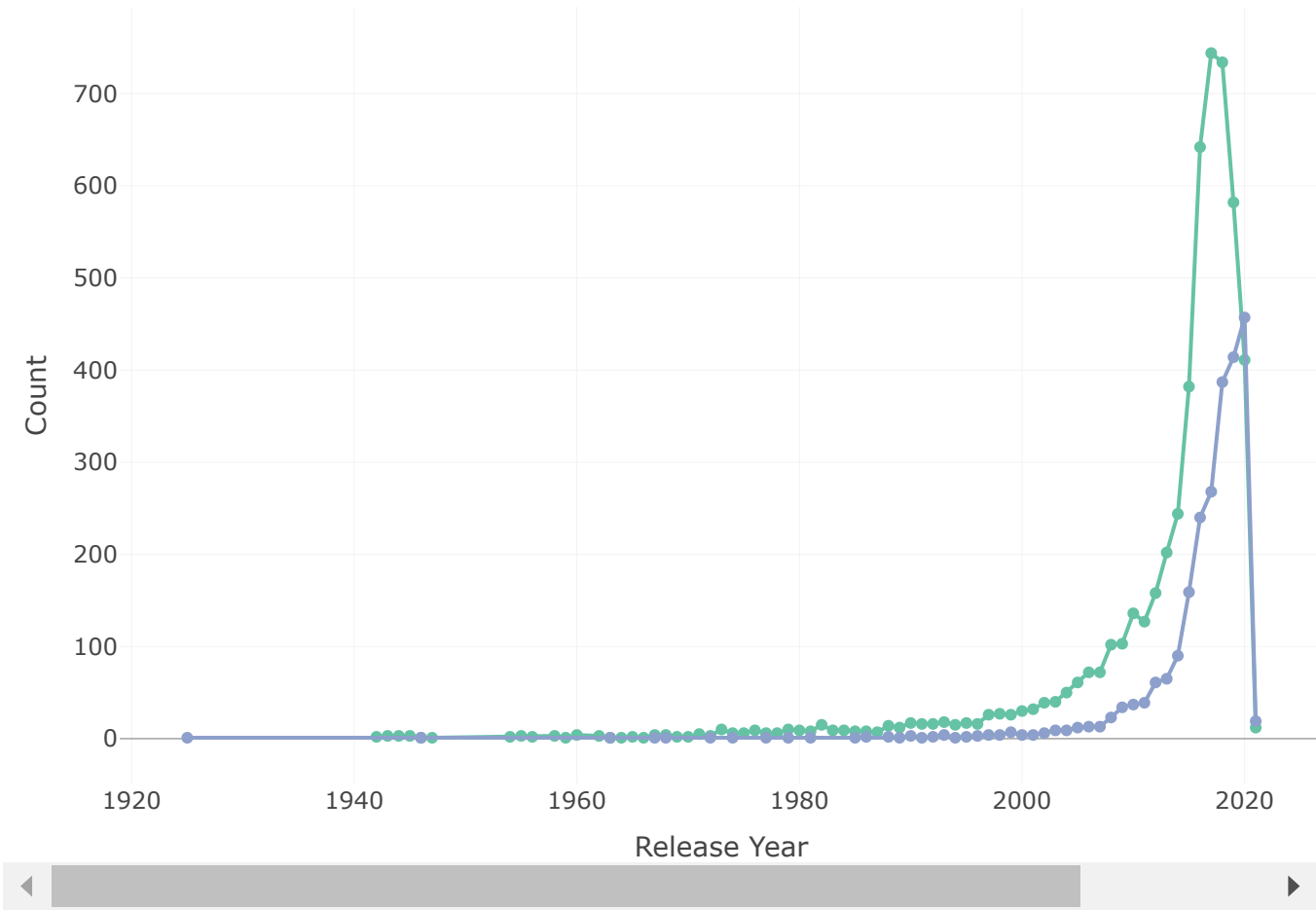
Brife Description:

Over the years, the production trends of TV shows and movies have undergone significant changes. Before 2000, the output of both remained relatively low and stable. After 2000, and particularly after 2010, there was a rapid increase in production. Movies peaked around 2015 before experiencing a slight decline, while TV shows continued to grow. Between 1920 and 1990, movies were the dominant form of entertainment, but after 2010, TV shows caught up and even surpassed movies in certain years. Post 2020, both TV shows and movies saw a notable decline, likely influenced by events such as the COVID-19 pandemic.

These changes can be attributed to several factors. The rise of streaming platforms like Netflix and advancements in production technology have lowered barriers, fueling growth. Audience preferences have shifted toward serialized content, driving the rise of TV shows. Additionally, TV shows' relatively lower production costs have made them an attractive option for creators and investors. Global events like the pandemic disrupted traditional production and distribution, accelerating these trends.

Summary

By analyzing the problems we have raised, we can draw the following conclusions.

First, Netflix's content is mainly focused on five genres: drama, comedy, documentary, international film and international TV series.

Looking at the time trend, the number of shows in these genres grew very slowly before 2000, and after 2010, perhaps with the popularity of streaming technology and Netflix's investment in original content, the number of shows in these genres increased rapidly. Documentaries and dramas in particular have shown the most significant growth, indicating their enduring appeal among audiences. In addition, the growth of comedy and international genre content is also very evident, reflecting the increased demand for diverse and localized content from global audiences.

From the perspective of regional preference, different regions have significant differences in their preferences for program types. For example, American audiences tend to prefer comedies and documentaries, while Indian audiences prefer international dramas and dramas, and South Korean audiences show a greater interest in international dramas than other countries. This difference may be related to the cultural background and entertainment habits of each region, but also closely related to Netflix's marketing strategy.

From the changing trend of TV shows and movies, the number of movies dominated in the early days, but in recent years, the growth rate of TV shows has obviously exceeded that of movies. This trend shows that TV shows are playing a bigger role in attracting users and increasing subscriber stickiness as viewers' demand for long-form, continuous content increases.

In conclusion, our research reveals Netflix's diversified content strategy in the streaming space and its global expansion trend. By analyzing program types, geographic preferences, and time shifts, documentaries and dramas dominate content, while the rapid growth of international content reflects global audiences' demand for localized and diverse programming. The rise of TV shows is a further indication of users' appetite for long-form content.