Filmography Trends Analysis

Vanesa Avoyan, Victoria Makaryan, Artyom Muradyan, Gor Parisakoyan,

Data Analysis Group Project
DS 120: Programming for Data Science, Section A
Professor: Liana Harutyunyan
College of Engineering and Science, AUA

Table of contents

Introduction

By examining historical data from IMDb and Netflix, we seek to identify underlying patterns, trends, and correlations that affect user preferences, watching habits, and content engagement.

This project aims to reveal tendencies and trends among filmmakers and watchers:

- Genre popularity among film/TV show producers and viewers (“The Best of the best”)
- Correlation between movie success and runtime
- TV shows analysis

We used 4 different CSV (and TSV) files as part of the project, generating 3 pairwise independent analyses containing 17 different plots in total.

A more detailed overview of the project and team members’ contributions can be found on our GitHub repository page

Data Selection and Description

Three primary datasets were used in the course of the project.

- Netflix Titles, retrieved from Kaggle, uploaded by Shivam Bansal
- 2 official IMDb non-profit datasets

In addition, for purposes of one plot, a dataset of all Academy Award nominations was used, retrieved from Kaggle as well.

Netflix Titles dataset contains ≈**8,000** entries with features such as:

- **Type**: Denotes the type of content, distinguishing between movies and TV shows.
- **Cast:** Actors involved in the movie/show
- **Genre**: The genre of the movie or show, providing additional insights into content categorization and viewer preferences.
- **Release Year**: The actual release year of the movie or show, providing temporal context to the content.
- **Rating**: The rating assigned to the movie or show reflecting its suitability for different audience demographics.
- **Total Duration**: Represents the duration of the content, measured either in minutes for movies or the number of seasons for TV shows, aiding in content categorization and analysis.

IMDb data is separated into multiple datasets, each containing specific data. We used 2 - title.basics and title.ratings.

Title.basics contains over **10 million** entries, which were later cleaned into a significantly smaller dataset. It's publicly collected data features movies from the 19th century up to 2024. Features include:

- **Title Type** – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- **Primary Title** – the title used by the filmmakers on promotional materials at the point of release
- **Original Title** - original title, in the original language
- **Start Year** – represents the release year of a title. In the case of TV Series, it is the series start year
- **Runtime Minutes** – primary runtime of the title, in minutes

- **Genres (string array)** – includes up to three genres associated with the title

Title.ratings includes average rating and number of voters.

The dataset for Academy Awards includes features such as title, year of release and nomination category.

Data Cleaning

Netflix Titles dataset did not require cleaning, as it was already in a usable condition. IMDb datasets, however, required significant cleaning and merging. This process is described in detail in `./imdb/imdb_cleaning.ipynb`. Due to the shear size of the original IMDb datasets, they were not included in the final Zip Archive, only the result of cleaning. Following the instructions in the file, it is possible to download the compressed datasets to verify cleaning.

Analysis 1

All the scripts for generating these plots can be found in the file `./Netlifx data.rmd`, with an HTML-generated report in `./Netflix data.nb.html`. Here is the description of manipulations and interpretations of the findings.

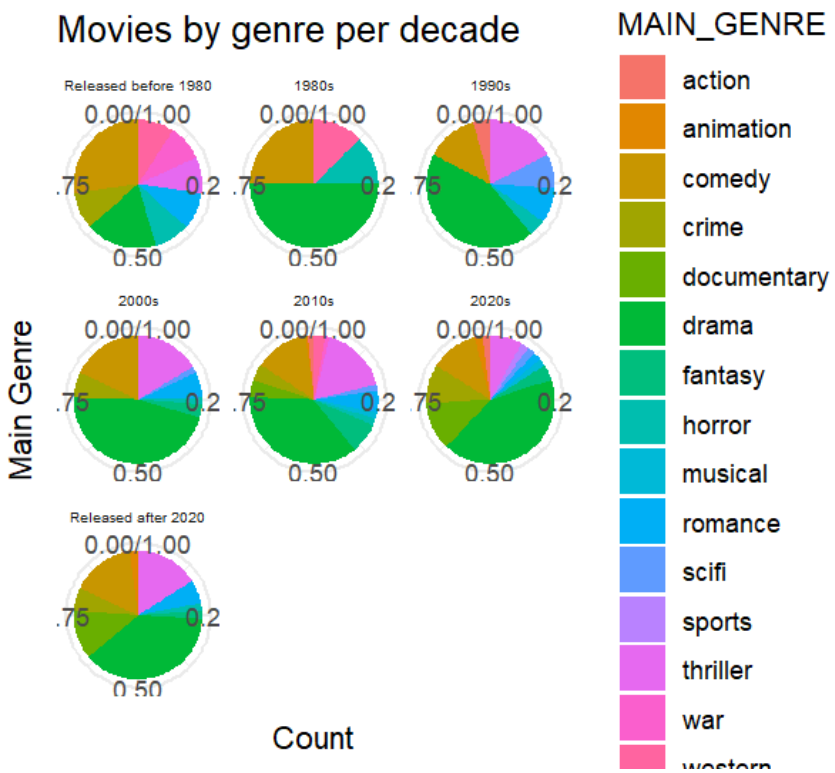
The plot illustrates the relationship between movie genres and their average ratings on Netflix. Each genre is represented along the x-axis, while the y-axis depicts the number of movies released in that genre assigned to movies within each decade.

Pie charts represent the popularity of genres among producers based on their quantity in each decade.

Notable mention: The release date starts from 1969, and the decade named “Released before 1980” contains movies and shows from 1969 to 1980. Another mention is that “Released after 2020” contains movies and shows of two years.

This study proposes the utilization of pie charts to explore the evolving popularity of movie genres across different decades. Each pie chart will be dedicated to a specific ten-year period, with its slices meticulously segmented to represent the proportional distribution of films belonging to various genres. This visual representation will serve to illuminate the dominant genres that have captivated producers throughout time. Furthermore, a particular emphasis will be placed on the position of Drama as the "absolute champion," aiming to investigate its prevalence and potential reasons for enduring popularity compared to other genres.

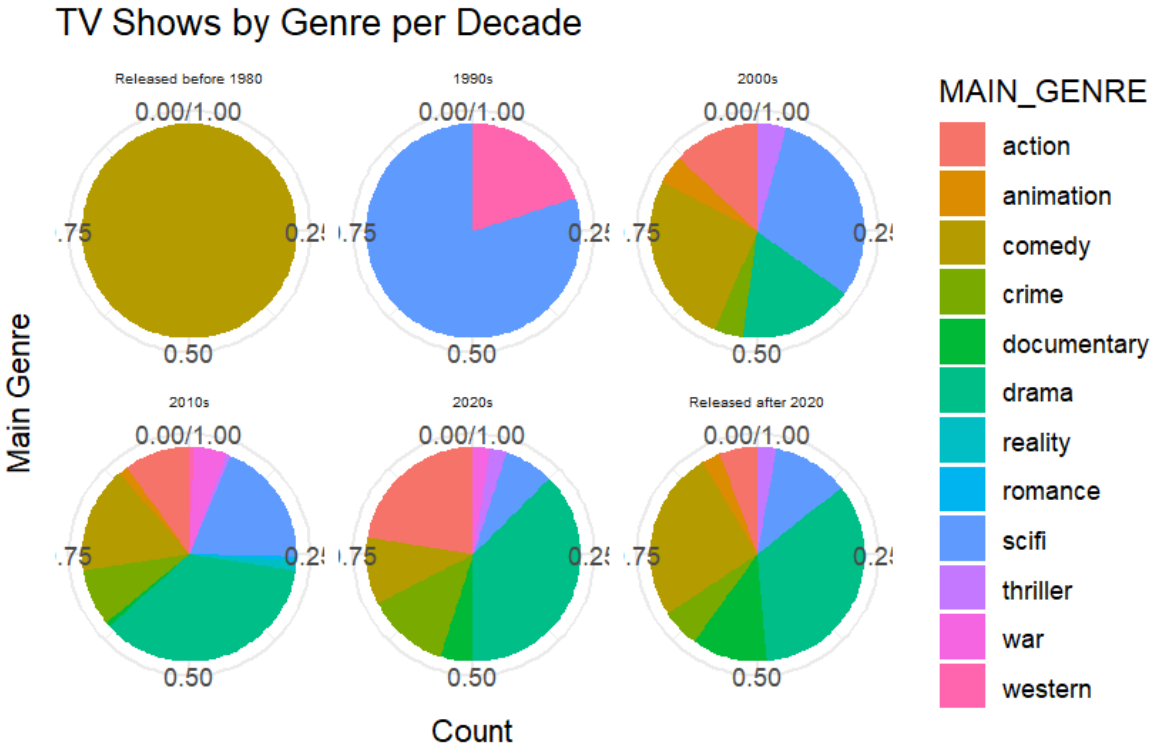
Figure 1.



While the provided data offers a glimpse into movie genre popularity across decades, limitations hinder a comprehensive analysis. The focus on Drama across all decades suggests its potential dominance, but the lack of complete data for other genres prevents definitive conclusions.

Further investigation with complete data sets could reveal fascinating trends. Analyzing the rise and fall of specific genres, the potential for genre convergence or divergence, and the influence of cultural shifts on genre preferences are all exciting avenues for future exploration.

Figure 2.



The chart does hint at a dynamic landscape for audience preferences. Genres like Comedy and Drama appear to have traded places in popularity throughout the time frames shown. Further investigation with complete data sets, including the total number of shows and the percentage each genre represents within a decade, would be necessary to draw more substantial conclusions about genre popularity and how it has evolved over time.

Upon observation of the pie charts, several trends and shifts in genre preferences become apparent across different decades. In earlier decades, such as the 1980s and 1990s, genres like Drama and Comedy often dominated the landscape, reflecting the cinematic trends and cultural preferences of the time. However, as the decades progress, there is a noticeable diversification in

genre representation, with the emergence of genres such as Romance and Thriller gaining prominence.

Figure 3.

MAIN_GENRE <fctr>	AVERAGE_VOTERS <dbl>	SCORE <dbl>
scifi	547524.4	8.8
western	395729.9	8.4
horror	219894.2	8.1
crime	183700.5	8.4

4 rows

Figure 4.

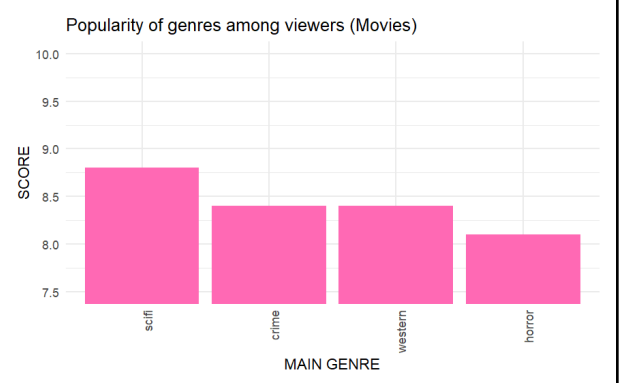
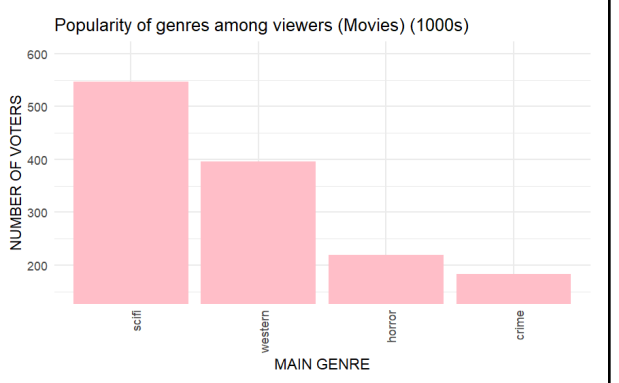


Figure 5.



This analysis, employing viewer votes as the primary metric of popularity, identified science fiction, westerns, horror, and crime as the top four most popular film genres. This genre diversity suggests a multifaceted audience drawn to a spectrum of cinematic experiences.

Science fiction's prominence reflects a fascination with exploring imaginative worlds and the potential of the future. Westerns, on the other hand, tap into a historical vein, offering narratives steeped in a bygone era's social and cultural landscapes. Horror's popularity speaks to the human desire for thrills and a confrontation with the macabre. Finally, crime films cater to viewers seeking suspenseful narratives that delve into the complexities of human behavior and the underbelly of society.

It is crucial to acknowledge that this study solely focused on audience votes. Future research could incorporate additional variables for a more comprehensive picture. These might include factors like critical reception, awards recognition, box office performance, or streaming

platform viewership data. Furthermore, delving deeper into subgenres within each category would provide a richer understanding of audience preferences.

Figure 6.

MAIN_GENRE <fctr>	AVERAGE_VOTERS <dbl>	SCORE <dbl>
scifi	139158.1	9.3
action	128604.6	9.1
thriller	111789.0	8.1
drama	105265.2	9.5

4 rows

Figure 7.

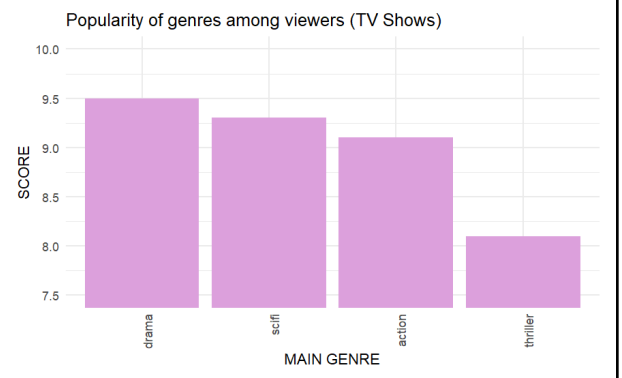
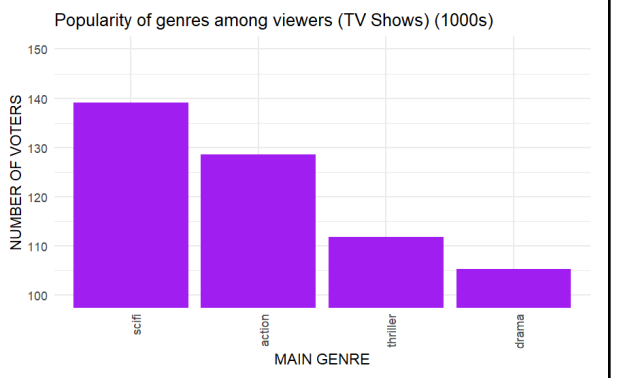


Figure 8.



This analysis, which utilized the variable of viewer votes to gauge popularity, revealed science fiction, action, thriller, and drama as the top four most favored television show genres. This diversity underscores a broad audience interest that transcends specific genre conventions.

The popularity of science fiction points to a fascination with imaginative worlds and the exploration of the unknown. Action and thriller genres, on the other hand, cater to viewers seeking excitement and suspense. Finally, drama's prominence reflects the enduring appeal of narratives that delve into the complexities of human experience.

It is important to acknowledge that this analysis focused solely on viewer votes, and further research could explore additional variables such as critical reception, awards recognition, or streaming platform viewership data. Furthermore, a deeper examination of subgenres within each category would provide a more nuanced understanding of audience preferences.

Sheir ability to resonate with audiences across generations.

Analysis 2

Details of the analysis of the correlation between runtime and success can be found in `./runtime-analysis/runtime-analysis.ipynb`. The notebook includes the scripts, plots, and interpretations of the findings.

Analysis 3

Details of the analysis of TV shows and their genre composition can be found in `./TV show analysis.ipynb`. The notebook includes the scripts, plots, and interpretations of the findings.

Conclusion and Recommendations

Film Genre Analysis Across Decades: An examination of film genres across various decades reveals a dominance of dramas. Dramas consistently hold the highest quantity of "champion" titles, suggesting a strong audience preference for this genre. Interestingly, science fiction films boast the highest average viewer votes and the genre with the most films achieving the maximum score.

Runtime and Success: The data demonstrates the clear connection between longers runtimes and higher ratings and success rates. However, different genres have vastly different typical lengths, which have changed dramatically over the years, allowing us to trace and better understand the evolution of mainstream filmography. At last, the analysis outputs the potential optimal lengths that movie producers today can use to maximise their success.

Television Genre Preferences: Television genres exhibit a more dynamic landscape compared to films. Audience favorites have fluctuated throughout the decades, with comedy, science fiction, and drama all taking turns at the top. When analyzing scoring comparisons, science fiction again emerges as the leader in terms of average viewer votes, followed by action, thriller, and drama. However, in television shows, the genre achieving the highest maximum score is drama, indicating a potential association between critically acclaimed television and dramatic storytelling.