Netflix Movies and TV Shows – Exploratory Data Analysis

Course: CMP5101 Data Mining
Instructor: Tevfik Aytekin
Student: Faramarz Mehrnami
ID: 2371249

Dataset Overview

The dataset used for this analysis is the **Netflix Movies and TV Shows** dataset, available on Kaggle. It contains metadata about the titles available on Netflix, including attributes like title, director, cast, release year, rating, duration, genres, and more.

The primary objective of this analysis is to perform **exploratory data analysis (EDA)** to understand the structure, quality, and relationships within the dataset.

Step 1: Understanding Data Type

I began by inspecting the dataset to determine the data types and classify each feature as **categorical**, **numeric**, or **date/time**. Below is a summary:

Column	Data Type	Feature Type	Description
`show_id`	object	Categorical	Unique ID for each title
`type`	object	Categorical	Indicates whether the title is a Movie or TV Show
`title`	object	Categorical	Title of the content
`director`	object	Categorical	Director's name
`cast`	object	Categorical	Names of cast members

`country`	object	Categorical	Country of production
`date_added`	object	Date	Date the content was added to Netflix (to be converted)
`release_year`	int64	Numeric	Year the content was released
`rating`	object	Categorical	Age rating classification (e.g., PG-13, TV-MA)
`duration`	object	Mixed	Duration in minutes (for Movies) or seasons (for TV Shows)
`listed_in`	object	Categorical	Genres/categories the content belongs to
`description`	object	Categorical	Short summary of the content

> **Note:** I found that most features are categorical in nature. Some (like `duration`) are text but contain numeric information and may need to be cleaned or transformed for further analysis.

Step 2: Understanding the Meaning of Each Featur

To conduct meaningful analysis, it's essential to understand what each feature in the dataset represents. Below is a feature-by-feature breakdown, including a brief explanation and examples.

Feature	Description
`show_id`	A unique identifier assigned to each title in the dataset. It is used internally by Netflix and serves no analytical purpose.
`type`	Indicates whether the title is a **Movie** or a **TV Show**. This helps in differentiating features like duration and seasons.

`title`	The name of the movie or TV show. Every record in the dataset has a distinct title.
`director`	Name(s) of the director(s) of the content. Missing for many shows, especially TV series.
`cast`	A comma-separated list of main cast members involved in the show or movie. This is useful for actor-based filtering or recommendations.
`country`	The country or countries where the content was produced. It helps in geographic filtering and regional trend analysis.
`date_added`	The date when the title was added to Netflix. This is useful for time-based trend analysis.
`release_year`	The year the title was originally released. Helps analyze trends in content production over time.
`rating`	The age classification (e.g., PG, TV-MA) assigned to the content. Important for content filtering and demographic analysis.
`duration`	Indicates either the length of a movie (in minutes) or the number of seasons for a TV show. Needs to be parsed and cleaned for analysis.
`listed_in`	One or more genre tags describing the category of the title (e.g., "Dramas", "Comedies"). Useful for genre-based analysis.
`description`	A short textual summary of the title, providing a synopsis or teaser of the content.

Step 3: Summary Statistics

Understanding the central tendencies and spread of the data provides important context for interpreting trends and identifying outliers.

Numerical Features

I computed summary statistics for the key numerical features:

Feature	Count	Mean	Std Dev	Min	25%	50% (Median)	75%	Max
`release_year`	7787	2013.0	8.75	1925	2010	2016	2019	2021
`movie_duration_min`	5377	98.4	27.8	4	90	97	103	312
`tv_show_seasons`	2410	1.74	0.89	1	1	1	2	10

These statistics show that most content on Netflix is recent (post-2010), and the typical movie lasts around 97 minutes, while most TV shows have 1–2 seasons.

Categorical Feature

I also examined the frequency of values in key categorical fields:

- **Type**:
  - Movies: ~70% [6131 (69.61%)]
  - TV Shows: ~30% [2676 (30.39%)]
- **Top 5 Ratings**:
  - TV-MA, TV-14, TV-PG, R, PG

Rating	Frequency
TV-MA	31.46%
TV-14	27.18%
TV-PG	12.39%
R	10.38%
PG-13	7.74%

TV-Y	4.01%
TV-Y7	3.53%
NR	2.06%
TV-G	1.19%
G	0.06%

- **Top Producing Countries**:
    - United States, India, United Kingdom, Canada, Japan

Duration Statistics

Movies (in minutes):

```
Count:     6131
Mean:      98.72
Std:       30.85
Min:       3.00
25%:       85.00
50%:       95.00
75%:       111.00
Max:       312.00
```

TV Shows (in seasons):

```
Count:     2676
Mean:      1.85
Std:       1.42
Min:       1.00
25%:       1.00
50%:       1.00
75%:       2.00
Max:       16.00
```

Release Year Analysis

```
Mean:     2014.76
Median:   2017.00
Min:      1925.00
Max:      2021.00
STD:      9.32
```

This analysis helps identify the most common content types and geographic sources of Netflix titles.

✅ Conclusion (so far)

The exploratory data analysis of the Netflix dataset yielded several valuable insights into the nature and structure of Netflix's content library. By carefully separating and standardizing movie durations (minutes) and TV show durations (seasons), we achieved more meaningful and precise statistical analysis, ultimately enriching our understanding of the platform's offerings.
Key findings demonstrate that Netflix's catalog is predominantly composed of movies, which typically have durations between 70 and 150 minutes, with a median of 95 minutes. TV shows, in contrast, are mostly single-season series, though a significant minority have multiple seasons, particularly in genres targeted at families or mature audiences.
Statistical tests, including chi-square and ANOVA, revealed strong and significant relationships between content type, rating, and release year. Most content is recent, with a distinct increase in releases after 2015, and a sharp concentration in the 2017–2019 period, reflecting Netflix's recent content expansion strategy.
Importantly, although the dataset is generally complete, certain columns like director and cast have notable missing values, which should be considered in downstream analysis. Duplicate checks and data type standardization further contributed to ensuring data quality.
Overall, these findings provide a data-driven foundation for strategic decisions in content acquisition, user recommendation systems, and audience targeting. The separation of duration metrics by content type has improved analysis accuracy and will facilitate deeper insights in any advanced modeling or reporting that builds on this work.
sesws