Insight on Netflix Data Analysis

Gordon Oboh

Department of Data Science, Monroe College, King Graduate School

CS628-151HY: Data Science

Professor Sanjay Upadhyay

October 27, 2024

Insight on Netflix Data Analysis

Introduction

This report presents an in-depth analysis of a Netflix dataset comprising 8,807 titles. The primary objective is to uncover trends, patterns, and insights Through comprehensive data cleaning, preprocessing, and exploratory data analysis, this report sheds light on various aspects of Netflix's content library, including content types, director and cast prominence, geographic distribution, release trends, ratings, genres, and more.

Data Description

The dataset under examination contains 12 columns with the following attributes:

- show_id: Unique identifier for each title (object)
- type: Specifies whether the title is a Movie or TV Show (object)
- title: Name of the title (object)
- director: Director(s) of the title (object)
- cast: Main cast members (object)
- country: Country or countries where the title was produced (object)
- date_added: Date when the title was added to Netflix (object)
- release_year: Year the title was originally released (int64)
- rating: Content rating (e.g., TV-MA, TV-14) (object)
- duration: Length of the title (e.g., minutes for movies, seasons for TV shows) (object)
- listed_in: Genres/categories the title belongs to (object)
- description: Brief synopsis of the title (object)

The release_year column was converted from an integer type to a datetime64[ns] type and subsequently split into three separate columns:

- month_added: Month when the title was added to Netflix (object)
- day_added: Day when the title was added to Netflix (object)
- year_added: Year when the title was added to Netflix (int64)

Data Cleaning and Preprocessing

During the initial data inspection, the following preprocessing steps were undertaken:

- Handling Missing Data: A total of 4,307 missing entries were identified across various columns. These missing values were imputed with the placeholder 'Not Specified' and '0' for release_year, the aim is to maintain dataset integrity and facilitate subsequent analyses.
- Duplicate Removal: An assessment revealed no duplicate records within the dataset, ensuring each row uniquely represents a distinct title.
- Date Conversion: The date_added column was transformed from an object type to a datetime format to enable temporal analyses. This facilitated the extraction of month, day, and year components for more granular insights.

Data Analysis and Findings

The dataset encompasses 12 columns and 8,807 rows, representing a diverse array of content available on Netflix. The breadth of data allows for multifaceted analyses across various dimensions, including content type, director and cast involvement, geographic distribution, and temporal trends.

Content Distribution

- **Type of Content:** The dataset reveals that movies constitute approximately 70% of the titles, nearly double the representation of TV shows, which account for the remaining 30%. This indicates a strategic emphasis on expanding the movie library.

Director and Cast Insights

- **Top Directors:** Analysis of the number of titles directed by individuals highlights a set of prolific directors who have significantly contributed to Netflix's library. For instance:
  - **Stand-Up Comedy:** Raúl Campos stands out with 18 titles, dominating this genre.
  - **Children & Family Movies:** Rajiv Chilaka and Suhas Kadav lead with 22 & 16 titles respectively.
  - **Dramas:** Youssef Chahine is notable with 12 titles.

  Identifying these key directors can inform future collaborations and content development strategies tailored to specific genres.
- **Top Cast Members:** The top 20 cast members are predominantly of Asian ethnicity, suggesting a strong representation and possible focus on Asian talent within Netflix's content strategy. This diversity may cater to a global audience and reflect broader industry trends.
- **Top Cast vs. Genres:**
  - **Action and Adventure:** Amitabh Bachchan and Shah Rukh Khan lead with 11 and 10 titles respectively.
  - **Comedies:** Adam Sandler, Shah Rukh Khan, and Akshay Kumar top the list with 18, 14, and 13 titles respectively.
  - **Docuseries and Science & Nature TV:** David Attenborough is the biggest name.
  - **Dramas:** Shah Rukh Khan, Amitabh Bachchan, Akshay Kumar, and Aamir Khan appear in 19, 16, 13, and 12 titles respectively.

  Obseervation: Shah Rukh Khan, Amitabh Bachchan, Akshay Kumar, and David Attenborough consistently appear across a variety of titles and genres, highlighting their versatility and broad appeal.

Geographic Distribution

- **Country of Origin:** The United States leads by a substantial margin in terms of the number of titles produced, followed by India, which contributes less than one-third of the US output. This dominance underscores the influence of the US entertainment industry on Netflix's offerings.

Release and Addition Trends

- **Release Years:** A significant portion of titles were originally released between 2016 and 2020, indicating a preference for relatively recent content. This may be aimed at appealing to contemporary audience tastes and maintaining relevance in a competitive market.
- **Addition of Titles:** Netflix exhibited peak addition of titles between 2018 and 2021, suggesting a period of aggressive expansion and content diversification.

Ratings and Content Length

- **Content Ratings:** The majority of titles are rated TV-MA (Mature Audience) and TV-14, implying a focus on adult-oriented content rather than pre-teen friendly programming.
- **Duration:** Most TV shows consist of a single season, while movies predominantly fall within the 88-106 minute range. This consistency in duration may align with audience viewing preferences and production efficiencies.

Genre Analysis

- **Popular Genres:** Dramas, Comedies, Documentaries, and Action & Adventure emerge as the most prevalent genres. This diversity caters to a wide range of viewer interests and supports Netflix's strategy of offering varied content to attract and retain subscribers.

Temporal Patterns in Content Addition

- **Monthly Trends:** February and May are identified as months with fewer title additions, whereas July and December experience peaks. This seasonal variation could align with strategic release schedules to maximize viewership during holidays and summer breaks.
- **Weekly Trends:** The majority of titles are added on Fridays, followed by Thursdays in a 7:4 ratio. This timing likely aims to capitalize on weekend viewership, encouraging subscribers to engage with new content during leisure periods.

Correlation Analysis

- **Type vs. Duration:** A negative correlation of -0.38 exists between content type (Movie vs. TV Show) and duration, indicating that as the type shifts from TV shows to movies, the duration tends to increase.
- **Release Year vs. Rating:** A positive correlation of +0.25 suggests that newer titles are more likely to receive higher maturity ratings, possibly reflecting evolving content standards and audience expectations.
- **Type vs. TV Ratings:** A positive correlation of +0.29 between content type and TV ratings implies that the type of content influences its rating, with certain types (e.g., movies vs. TV shows) being more associated with specific rating categories.

Keyword Analysis in Descriptions

- **Frequent Terms:** A word count analysis, excluding stop words, identified "life," "young," "new," "family," "world," "man," "two," "love," "woman," and "friends" as the top 10 most frequent words in title descriptions. These terms highlight recurring themes and focal points in Netflix's content, such as relationships, personal growth, and familial narratives.

Top Cast vs. Top Directors

Collaborations:

- Rajiv Chilak and Vatsal Dubey have collaborated on 16 titles together.
- Suhas Kadav and Saurav Chakraborty have worked on 8 titles.
- Toshiya Shinohara and Kappei Yammaguchi have collaborated on 7 titles.
- S.S. Rajamouli and Prabhas have also worked together on 7 titles.

Conclusion

The analysis of the Netflix dataset provides valuable insights into the platform's content strategy and offerings. Key findings include a predominant focus on movies over TV shows, significant representation of Asian cast members, and a strong emphasis on content from the United States and India. Temporal trends indicate strategic content additions aligned with peak viewing periods, while genre diversity caters to a broad audience base. Additionally, the prominence of certain directors and cast members across multiple genres highlights their influence and versatility within the Netflix ecosystem.

Recommendations

Based on the findings, the following recommendations are proposed:

- **Diversify Content Types:** While movies dominate the library, increasing the proportion of TV shows could attract viewers who prefer serialized content.
- **Expand Geographic Representation:** Enhancing content from countries beyond the USA and India may cater to a more global audience and tap into emerging markets. Specifically, targeting the Chinese, Korean, and Japanese markets may help attract a larger and more diverse subscriber base.
  - **Chinese Market:** Developing and acquiring titles from China can open doors to one of the world's largest streaming audiences.
  - **Korean Market:** Korean dramas, TV shows, and movies have garnered global popularity. By investing in high-quality Korean content, Netflix can leverage the existing fan base and continue to expand its reach.
  - **Japanese Market:** Japanese anime is a genre with a dedicated and passionate global following.
- **Balanced Rating Spectrum:** Introducing more content with lower maturity ratings could attract younger audiences and families, broadening the subscriber base.
- **Strategic Release Scheduling:** Continue leveraging peak months like July and December for major releases, while exploring opportunities to bolster content additions in traditionally slower months like February and May.