NETFLIX CONTENT ANALYSIS USING DATA SCIENCE

AARYA SHUKLA

UNIFIED MENTOR

NETFLIX CONTENT ANALYSIS USING DATA SCIENCE

Abstract

With the rapid growth of streaming services, Netflix has become a global entertainment giant. This project analyzes the Netflix dataset to uncover insights into content trends, genre distribution, release patterns, and the evolution of content across regions and time. The analysis includes EDA, preprocessing, and visualization techniques using Python and data science libraries.

1. Introduction

The entertainment industry has undergone a revolution due to streaming platforms like Netflix. Understanding the composition and evolution of Netflix's content can help creators, marketers, and analysts identify trends and gaps in global media consumption. This notebook uses Netflix's publicly available dataset to explore content types, genre trends, release year distribution, and more.

2. Dataset Overview

The dataset (netflix_titles.csv) includes the following attributes:

- show_id: Unique ID of the show

- **type: Movie or TV Show**

- **title: Name of the content**

- **director: Director(s) name**

- **cast: Main cast members**

- **country: Production country**

- **date_added: Date when added to Netflix**

- **release_year: Year content was released**

- **rating: Age classification**

- **duration: Length (mins or seasons)**

- **listed_in: Genres or categories**

- **description: Brief summary**

3. Methodology

Data Cleaning & Preprocessing:

- **Removed null entries (e.g., missing directors, cast)**

- **Split and standardized multi-value fields like listed_in, cast, and country**

- **Converted date_added to datetime format**

Exploratory Data Analysis (EDA):

- **Count of Movies vs. TV Shows**

- **Most frequent directors and actors**

- **Top countries producing Netflix content**

- **Year-wise release trends**

- Genre distribution

Tools Used:

- **Pandas** for data manipulation

- **Matplotlib** and **Seaborn** for visualization

- **WordCloud** for visualizing genre popularity

4. Key Analysis and Visualizations

- **Content Type Distribution**:

    o **Movies dominate the dataset (~70%), with the rest being TV shows**

- **Top Countries**:

    o **USA, India, UK are the highest producers of Netflix content**

- **Year-wise Release Trend**:

    o **Peak in 2017–2020, with consistent yearly additions**

- **Top Genres**:

    o **International Movies, Dramas, Comedies appear most frequently**

- **Frequent Directors and Actors**:

    o **Count plots revealed individuals with the most Netflix entries**

- **Ratings Analysis**:

    o **Distribution of content across TV-MA, PG-13, R, and other age groups**

5. Insights & Observations

- **Growth Pattern**:
  - Netflix's content library grew exponentially between 2015–2020
  - TV Shows saw a significant increase after 2016
- **Genre Trends**:
  - "Documentaries" and "Children's Shows" are strong contributors
  - Increasing diversity in genre combinations
- **Geographic Diversity**:
  - U.S.-centric but expanding internationally (notably India, UK, Canada)
- **Family and Adult Content Split**:
  - Ratings analysis shows balanced offerings for both age groups

6. Conclusion

Netflix has evolved into a globally diverse content provider with an extensive catalog of movies and TV shows. The insights from this analysis can be useful for:

- Content strategy teams to understand gaps or strengths

- **Market researchers to analyze viewing trends**
- **Recommendation engines for user personalization**

7. Future Scope

- **Build a recommendation system using content-based filtering**
- **Perform sentiment analysis on descriptions**
- **Explore viewership statistics if available**
- **Integrate with IMDB or TMDB datasets for rating comparisons**
- **Use network analysis to map actor-director collaborations**

8. References

- **Netflix Titles Dataset (Kaggle)**
- **Python libraries: Pandas, NumPy, Seaborn, Matplotlib, WordCloud**