Unified Mentor - Netflix Content Analysis Report

- Adimalla Nithin Siddhartha

Introduction

This report summarizes the initial analysis performed on the Netflix content dataset. The dataset contains a collection of information about movies and TV shows available on the Netflix streaming platform. The primary objective of this project is to conduct an exploratory data analysis (EDA) to understand the characteristics of the content library, identify trends in content addition, explore popular genres and directors, and gain insights that can inform further analysis or potential applications like content recommendation systems.

Data Loading and Preprocessing

The analysis commenced with loading the dataset from the provided CSV file, named netflix1.csv. The dataset was loaded into a pandas DataFrame, a fundamental data structure for data manipulation and analysis in Python.

- **Data Loading:** The pd.read_csv() function was used to load the data into a DataFrame named df. To get an initial understanding of the data structure and content, the head of the DataFrame (the first few rows) was displayed. This provided a preview of the columns and the types of data they contain.
- **Data Cleaning:** A crucial step in data analysis is handling missing values. A check for missing values was performed across all columns using df.isnull().sum(). The output of this operation revealed that the dataset is remarkably clean, with no missing values in any of the columns. This simplifies the preprocessing stage significantly, as no imputation or removal of missing data is required.
- **Data Type Conversion:** For time-based analysis, the date_added column, which was initially loaded as an object type, was converted to datetime objects using pd.to_datetime(). This conversion allows for easy extraction of temporal components like year and month. Subsequently, new columns, year_added and month_added, were extracted from the date_added column. These new features are essential for analyzing trends in content addition over time.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to delve deeper into the characteristics and patterns within the dataset. Various aspects of the Netflix content library were explored through visualizations and statistical summaries.

- **Content Type Distribution:** The distribution of content types, specifically Movies versus TV Shows, was analyzed by counting the occurrences of each type in the 'type' column using the value_counts() method. A bar plot was generated to visualize this distribution. The visualization clearly showed that the Netflix library in this dataset contains a significantly higher number of Movies compared to TV Shows. This initial insight highlights the platform's focus on movies within this dataset.

- **Most Common Genres:** The 'listed_in' column often contains multiple genres for a single title, separated by commas. To analyze the popularity of individual genres, the 'listed_in' column was split by the comma and space delimiter, and the resulting individual genres were stacked into a single Series. The value_counts() method was then applied to this Series to count the occurrences of each genre. The top 10 most common genres were identified and visualized using a bar plot with a 'viridis' color palette. The plot revealed that 'International Movies' and 'Dramas' are the most prevalent genres, followed by 'Comedies' and 'International TV Shows'. This provides valuable insight into the popular content categories available on the platform.

- **Content Added Over Time:** To understand how the Netflix content library has grown over the years, the number of titles added each month was calculated. This was achieved by grouping the DataFrame by the extracted year_added and month_added columns and counting the number of entries in each group using the size() method, which was then reset as a DataFrame with a 'count' column. A new 'date' column was created by combining the year and month information and converting it to datetime objects. A line plot was generated to visualize the trend of content addition over time, with the 'date' on the x-axis and the 'count' on the y-axis. The line plot clearly illustrated a significant increase in content added to Netflix starting around 2016, with a peak in content additions observed in recent years within the dataset's timeframe.

- **Top 10 Directors with the Most Titles:** To identify the most prolific directors on Netflix within this dataset, the 'director' column was analyzed. Entries marked as 'Not Given', which represent missing director information, were excluded from the analysis. The value_counts() method was applied to the filtered 'director' column to count the number of titles for each director. The top 10 directors with the highest number of titles were identified and displayed. A bar plot was created to visualize this distribution, using a 'viridis' palette. The plot highlighted directors who have contributed a large number of titles to the Netflix library, such as Rajiv Chilaka and the duo Raúl Campos, Jan Suter.

- **Word Cloud of Movie Titles:** To gain a visual understanding of the most frequent words used in movie titles on Netflix, a word cloud was generated. First, the DataFrame was filtered to include only entries of 'Movie' type. The titles of these movies were then concatenated into a single large string. The WordCloud library was used to generate the word cloud from this text, with parameters set for width, height, and background color. The resulting word cloud image was displayed, with the size of each word in the cloud being proportional to its frequency in the movie titles. The word cloud visually emphasized common themes and keywords present in Netflix movie titles, such as "Movie", "Love", "Story", and "Life".

Next Steps

Based on the initial analysis and the insights gained from the EDA, there are several potential next steps to further explore the Netflix content dataset:

- **Further Feature Engineering:** Extracting more granular information from existing columns. For example, analyzing the primary country of production from the 'country' column, or separating the duration for movies (in minutes) and the number of seasons for TV shows.
- **Advanced Visualization:** Creating more complex and insightful visualizations to explore relationships between different features. This could include visualizing the distribution of genres over time, analyzing the distribution of ratings by content type, or exploring the relationship between release year and the year the content was added to Netflix.
- **Content Recommendation System:** Leveraging the insights gained from the EDA and potentially incorporating additional features to build a content recommendation system. This could involve using techniques like collaborative filtering or content-based filtering to suggest titles to users based on their viewing history or the characteristics of the content.
- **In-depth analysis of specific features:** Conducting a deeper dive into the 'country', 'rating', and 'release_year' columns to uncover more specific trends, patterns, and potential correlations with other features. For instance, analyzing the content distribution by country, the popularity of different ratings, or the trend of content release years over time.
- **Statistical Modeling:** Applying statistical models to analyze the data, such as time series analysis on content addition trends or regression analysis to explore factors influencing content popularity.

- **Text Analysis on Descriptions:** If available, performing text analysis on content descriptions to identify common themes, keywords, and sentiment associated with different content types or genres.