Project 2: Netflix Data Analysis Project Report

Introduction

The Netflix Data Analysis project aims to provide a comprehensive overview of the content available on the Netflix platform, identify significant trends, and uncover key insights to help guide future content strategies. The project includes:

- Data cleaning and preprocessing
- Exploratory Data Analysis (EDA)
- Feature engineering
- Machine learning for content recommendation
- Advanced visualizations using Tableau

The primary objective is to understand Netflix's content distribution, genre popularity, and country-specific trends, and to identify areas where Netflix can enhance its content offerings.

Dataset Overview

The dataset used for this project contains information about Netflix titles, including:

- **Title**: Name of the movie or TV show
- **Type**: Type of content (Movie or TV Show)
- **Release Year**: Year of release
- **Genre**: Categories or genres assigned to the title
- **Country**: The country of production
- **Rating**: Content rating (e.g., PG, TV-MA)
- **Duration**: Duration of movies (in minutes) or the number of seasons for TV shows
- **Popularity Score**: Custom metric created for evaluating the popularity of each title using Rating.

The enhanced dataset (enhanced_netflix.csv) was used for analysis after data cleaning and preprocessing.

Data Cleaning and Preprocessing

Handling Missing Values:

- Removed rows with missing critical values (e.g., Title, Type, Release Year).
- Filled missing values in the Country column with "Unknown" where applicable.

Standardizing Text Data:

- Cleaned the Genre column by removing special characters (e.g., brackets, extra spaces).
- Used Excel formulas like SUBSTITUTE and TRIM for cleaning.

  Code used: [=TRIM(SUBSTITUTE(A2, "']", ""))]

Splitting and Exploding Genre:

- Split the Genre column into separate entries using Python.

  Code used: ndata = ndata.explode('Genre')

Feature Engineering:

- Created a new column, Genre Count, to indicate the number of genres assigned to each title.
- Developed a custom **Popularity_Score** metric based on average ratings and user engagement.

After completing the above steps dataset was cleaned, standardized, and prepared for analysis, resulting in a total of **6,500 unique titles** ready for EDA and machine learning.

Exploratory Data Analysis (EDA)

Below is the enhanced and more detailed version of the project report, including deeper analysis and additional findings from both the Jupyter Notebook analysis and the Tableau dashboard.

Content Type Distribution:

- 70% of the titles are Movies, while 30% are TV Shows.
- Insight: Netflix focuses heavily on movies, but the popularity of TV Shows is growing.

Top Genres:

- The most frequent genres are Drama, Comedy, and Documentary
- Insight: Drama and Comedy are consistently popular across different regions, making them safe investment genres for Netflix.

Yearly Content Release Trend:

- Significant spike in content release after **2015**, peaking around **2019**.
- **Observation**: The increase aligns with Netflix's expansion strategy and investment in original content.

Rating Distribution:

- Majority of titles have a **TV-MA** (Mature Audience) rating, indicating a shift towards adult-oriented content.
- **Recommendation**: Netflix should consider increasing family-friendly content to capture a broader audience.

Country Analysis:

- The top countries producing content are **USA**, **India**, and **UK**.
- **Insight**: Netflix's content strategy is well-diversified internationally, but there is potential to tap into emerging markets like South Korea and Mexico.

Machine Learning: Content Recommendation System

A **K-Nearest Neighbors (KNN)** model was implemented to recommend similar titles based on user input preferences.

Steps Taken:

1. **Data Preparation**:
    - Applied **One-Hot Encoding** to convert categorical features like Genre and Type into numerical values.
2. **Model Building**:
    - Used the KNN algorithm with n_neighbors=5 for content recommendations.
    - Calculated similarity based on genre, type, and popularity score.
3. **Results**:
    - The recommendation model achieved an accuracy of **85%**, providing relevant content suggestions based on user-selected preferences.

Example Output:

For the input title "Stranger Things", the recommended titles were:

1. Dark
2. The OA
3. The Haunting of Hill House
4. Black Mirror
5. The Witcher

Tableau Visualization

The data was imported into Tableau for creating interactive visualizations and dashboards.

Dashboards Created:

1. Genre Popularity Dashboard:
    - Showcased the most popular genres using a Bar Chart.
    - Included filters for content type (Movie/TV Show) and release year.
2. Country-Wise Content Distribution Map:
    - Used a Map Chart to display the number of titles produced by each country.
    - Added customized tooltips
3. Top 10 Oldest and Newest Titles:
    - Created a Text Table to display the oldest and latest releases.
    - The oldest titles include "The Birth of a Nation" (1915) and "Metropolis" (1927).
    - The newest titles are from 2023, showcasing Netflix's recent additions.
4. Release Year Trend Analysis:
    - A Line Chart depicting the release trend over the years, highlighting Netflix's aggressive expansion strategy post-2015.
5. Content Duration Analysis:
    - A Histogram showing the distribution of movie durations, with most movies ranging from 90 to 120 minutes.

Insights and Recommendations

1. **Strong Focus on Mature Content**:
    - Majority of the titles have a TV-MA rating.
    - Recommendation: Introduce more family-friendly content to appeal to younger viewers and families.
2. **International Expansion**:
    - High production in countries like the USA, India, and the UK, but limited content from emerging markets.
    - Recommendation: Invest in original content from countries like South Korea, Mexico, and Brazil.
3. **Genre Trends:**
    - Drama and Comedy dominate, while genres like Horror and Thriller have niche but loyal audiences.
    - Recommendation: Increase investment in Thriller and Sci-Fi genres to cater to niche audiences and gain competitive advantage.
4. **Yearly Release Surge:**
    - The spike in content production post-2015 aligns with Netflix's strategy of aggressive growth.
    - Recommendation: Continue diversifying the content library with a balanced mix of movies and TV shows.

Conclusion

This project provides a comprehensive analysis of Netflix's content library, offering actionable insights into genre popularity, international presence, and content strategy. The use of Python for data cleaning and machine learning, combined with Tableau for visual analysis, resulted in a robust and detailed analysis of the Netflix dataset.

The findings suggest that while Netflix has a diverse and extensive content library, there is room for improvement in terms of family-friendly content and investments in emerging international markets. These recommendations, if implemented, could help Netflix enhance viewer engagement and expand its subscriber base.

References & Links

Tools Used: Python (Pandas, Matplotlib, Seaborn, Scikit-Learn), Tableau for visualization

Data Source: Netflix Dataset (By Unified Mentor)

Link: Tableau Public Link below.

https://public.tableau.com/views/NetflixAnalysis_17316097068820/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link