Capstone Final Report:
Netflix Recommendation System

Problem Statement:

In an era of overwhelming content choices, Netflix users often struggle to find shows and movies that align with their preferences, leading to user dissatisfaction and potential churn. This project aims to develop a recommendation system for Netflix that effectively suggests titles based on user preferences and viewing habits. Leveraging content-based filtering techniques, the system will analyze rich metadata including genre, director, cast, and user historical data to provide personalized recommendations. By enhancing the user experience through tailored content suggestions, the goal is to increase viewer satisfaction, engagement, and retention on the platform.

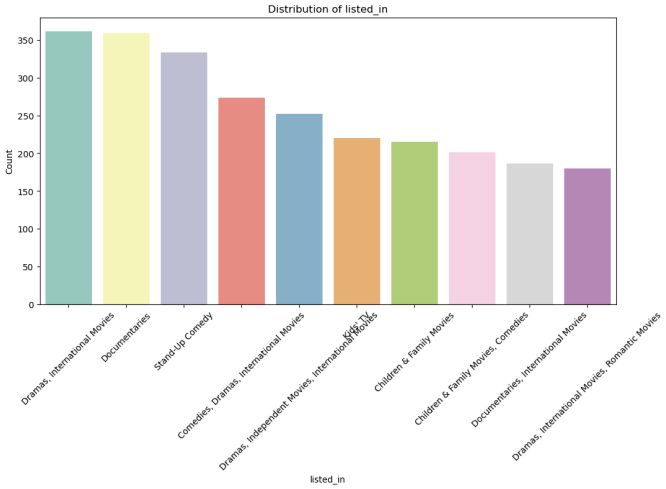
This project focuses on implementing a content-based filtering approach within Netflix's recommendation system. Content-based filtering recommends titles by identifying similarities in attributes such as genres, actors, directors, and other metadata available in the Netflix dataset. By analyzing these features, the system aims to suggest movies and TV shows that align closely with each user's preferences and viewing history, thereby enhancing the personalized viewing experience and improving user satisfaction on the platform.

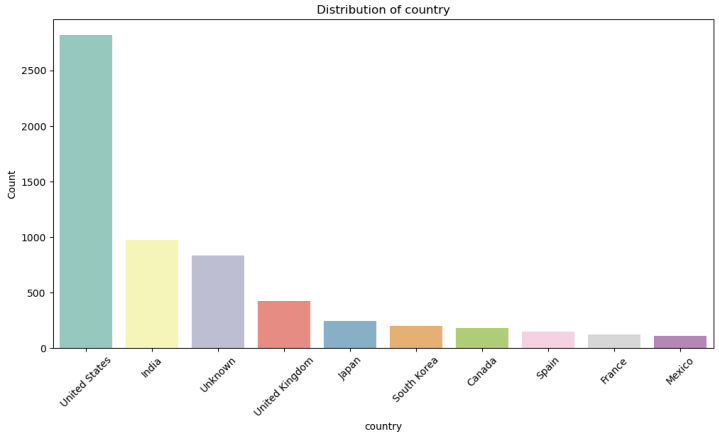
Data Cleaning:

In the data cleaning process, several steps were taken to ensure the Netflix dataset was prepared for analysis. Initially, missing values were identified across columns using df.isnull().sum(). Notably, columns such as 'director', 'cast', 'country', 'date_added', 'rating', and 'duration' had varying degrees of missing data. To address this, missing values were filled in categorical columns with appropriate placeholders: 'Unknown' for 'director', 'cast', and 'country', while 'Unknown' was also used for 'rating' and 'duration'. For the 'date_added' column, missing values were replaced with a default date ('2024-01-01').

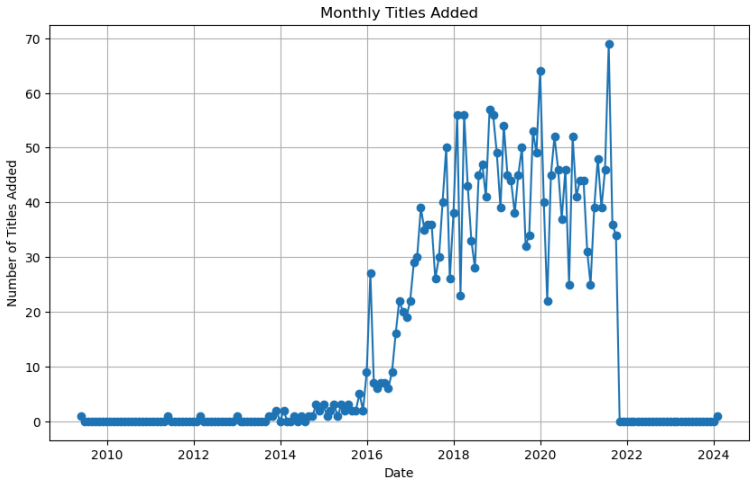
To further refine the dataset, rows with multiple critical missing values or any missing values in essential columns ('director', 'cast', 'country', 'date_added', 'rating', 'duration') were dropped using a conditional selection (drop_condition). This ensured that only rows with sufficient data across these essential attributes were retained for analysis. After cleaning, the dataset remained consistent with its original size, confirming that no significant data loss occurred through these operations. This meticulous cleaning process ensures that subsequent analysis and modeling efforts on the Netflix dataset are based on a robust and complete foundation of data.

Exploratory Data Analysis:





The analysis of categorical features reveals several key insights into Netflix's content distribution and audience targeting strategies. Movies significantly dominate the dataset, comprising a substantial majority compared to TV shows, suggesting Netflix prioritizes a diverse selection of films to cater to a broad audience seeking varied movie options. In terms of user ratings, TV-MA, TV-14, and TV-PG are the most prevalent, indicating a strong focus on content suitable for mature audiences and older children, while less common ratings like G, NC-17, and UR suggest fewer titles with extreme content restrictions or targeted towards very young viewers. Geographically, the United States leads in the number of titles, followed by India, the United Kingdom, and Canada, underscoring Netflix's global content reach and significant production or acquisition efforts from these regions. Finally, popular genres such as Dramas, Comedies, and Documentaries highlight high viewer demand, alongside efforts in International Movies and Kids TV genres, reflecting Netflix's commitment to diverse global content and family-oriented programming



Analysis of Netflix's yearly additions shows a consistent upward trend, indicating continuous expansion of its content library to meet growing viewer demand. Notable spikes, such as in 2020 during the COVID-19 pandemic, suggest strategic efforts to enhance content offerings in response to increased streaming activity. These trends reflect Netflix's dynamic approach to content acquisition, aimed at maintaining competitiveness and satisfying diverse viewer preferences.

The sparsity analysis of our simulated user-item interaction matrix reveals a high degree of emptiness, with a calculated sparsity of 0.9501. This indicates that approximately 95% of the entries in the matrix are zeros, illustrating that most users have not interacted with the vast majority of Netflix titles in our dataset. Such high sparsity underscores the challenge of recommending personalized content based on user interactions alone, as the data is heavily skewed towards non-interaction. Effective recommendation systems must leverage advanced techniques such as collaborative filtering or content-based filtering to mitigate the impact of sparse data and accurately predict user preferences. Understanding and addressing this sparsity is crucial for enhancing the recommendation accuracy and user experience on platforms like Netflix.

Model Selection:

The modeling phase of the Netflix Recommendation System project commenced with comprehensive data preprocessing and feature engineering to optimize dataset quality for subsequent analysis. Key steps included removing stop words and special characters from text columns such as 'director', 'cast', 'country', and 'listed_in', ensuring the cleanliness and relevance of textual data. Using the CountVectorizer from scikit-learn, categorical text features were transformed into binary token matrices, facilitating efficient representation of attributes like directors, actors, countries, and genres. These binary matrices were then integrated into DataFrames, with rows

representing individual movies and columns encoding specific features. The cosine similarity matrix, derived from these binary-encoded features, quantified similarity between movies based on shared attributes, thereby laying a robust foundation for accurate and personalized movie recommendations tailored to user preferences and content similarities.

The movie and TV show recommendation models utilize a content-based filtering approach to suggest relevant titles based on similarities with the queried movie or TV show. Each model begins by verifying the presence of the entered title in its respective dataset. Upon confirmation, the model calculates similarity scores using precomputed matrices that capture attributes such as genre, director, cast, and description. These scores are then used to identify and rank other titles in descending order of similarity, excluding the queried title itself. The models return the top 5 recommended movies or TV shows along with their respective similarity scores, ensuring each recommendation includes essential details like type, director, cast, country, date added, release year, rating, duration, genres, and a brief description. By leveraging content similarities rather than user viewing history, these models provide personalized viewing suggestions that align closely with individual preferences, enhancing user engagement and satisfaction by delivering tailored recommendations based on the entered title.

```
movie_recommendation('The Interview')
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	popularity_score	similiarity
1	s2710	Movie	Coffee & Kareem	Michael Dowse	Ed Helms, Taraji P. Henson, Terrence Little Ga...	United States	April 3, 2020	0.659930	TV-MA	88 min	Action & Adventure, Comedies	An inept Detroit cop must team up with his gir...	0.96	0.35634
2	s4783	Movie	The Legacy of a Whitetail Deer Hunter	Jody Hill	Josh Brolin, Danny McBride, Montana Jordan, Sc...	United States	July 6, 2018	0.433143	TV-14	83 min	Action & Adventure, Comedies, Dramas	A star of hunting videos strives to bond with ...	0.63	0.32232
3	s296	Movie	The Paper Tigers	Quoc Bao Tran	Alain Uy, Ron Yuan, Mykel Shannon Jenkins, Jae...	United States	August 7, 2021	0.773324	PG-13	111 min	Action & Adventure, Comedies	After reuniting as middle-aged men, three kung...	0.89	0.30860
4	s2837	Movie	Spenser Confidential	Peter Berg	Mark Wahlberg, Winston Duke, Alan Arkin, Bokee...	United States	March 6, 2020	0.659930	R	111 min	Action & Adventure, Comedies	Spenser, an ex-cop and ex-con, teams up with a...	0.93	0.30860
5	s4171	Movie	Polar	Unknown	Unknown	United States, Germany	January 25, 2019	0.546537	TV-MA	119 min	Action & Adventure, International	An assassin on the verge of retirement	0.65	0.30304

```
TV_show_recommendation('New Girl')
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	popularity_score	similiarity
1	s4571	TV Show	Hot Date	Unknown	Emily Axford, Brian Murphy	United States	October 1, 2018	0.433143	TV-MA	1 Season	Romantic TV Shows, TV Comedies	Interconnected sketches and performances skewe...	0.48	0.516398
2	s512	TV Show	Chelsea	Unknown	Unknown	United States	July 6, 2021	0.319749	TV-MA	2 Seasons	Stand-Up Comedy & Talk Shows, TV Comedies	It's not her first talk show, but it is a firs...	0.16	0.424264
3	s1440	TV Show	History of Swear Words	Unknown	Nicolas Cage	United States	January 5, 2021	0.773324	TV-MA	1 Season	Docuseries, TV Comedies	Nicolas Cage hosts this proudly profane, funny...	0.14	0.424264
4	s1531	TV Show	Schulz Saves America	Alexx Media	Andrew Schulz	United States	December 17, 2020	0.659930	TV-MA	1 Season	Stand-Up Comedy & Talk Shows, TV Comedies	Comedian Andrew Schulz takes on the year's mos...	0.12	0.424264
5	s2069	TV Show	Felipe Esparza: Bad Decisions	Unknown	Felipe Esparza	United States	September 1, 2020	0.659930	TV-MA	1 Season	Stand-Up Comedy & Talk Shows, TV Comedies	Two live performances, one in English and one ...	0.19	0.424264

Conclusion:

The Netflix Recommendation System project was designed to create a highly personalized engine for suggesting movies and TV shows based on user preferences and content similarities. Key stages included meticulous data cleaning to remove noise and enhance data quality, particularly in text-based columns such as directors, cast, and genres. This step ensured that subsequent analyses were based on accurate and meaningful data.

Feature extraction utilized the CountVectorizer to transform categorical text data into binary token matrices, enabling the representation of attributes like director names and genres as structured binary features. This approach facilitated efficient similarity calculations, essential for generating tailored recommendations. Binary encoding further refined the data into a format where each row represented a unique title and each column indicated the presence or absence of specific attributes.

By employing these methodologies, the recommendation system was able to deliver relevant and engaging suggestions to users, enhancing their viewing experience by focusing on content similarities rather than relying on historical viewing patterns. This approach not only personalized recommendations but also accommodated the diverse and extensive Netflix dataset, demonstrating its effectiveness in providing curated content recommendations based on user preferences