Netflix Movies and TV Shows Clustering

Introduction & Business Goal

The primary objective of this project was to apply unsupervised machine learning techniques to the Netflix Movies and TV Shows dataset. By analyzing features such as genre, rating, and duration, the goal was to segment the extensive content library into distinct, meaningful clusters. The identification of these similar content groups serves several key business purposes: enhancing personalized content recommendations, identifying niche content categories, understanding market trends, and assisting production houses in recognizing content gaps.

This analysis successfully segmented the Netflix library into three primary, interpretable clusters: a large collection of **Modern Movies**, a comprehensive group of **TV Shows**, and a smaller, niche library of **Classic Films**. This report details the methodology, from data preprocessing to model evaluation, and discusses the insights derived from these findings.

Data Cleaning and Feature Engineering

The initial dataset consisted of approximately 7,787 entries and 12 columns, containing a mix of numerical and text-based data about movies and TV shows available on Netflix. The first step in the analysis was a thorough data cleaning and preprocessing phase to handle inconsistencies and prepare the data for modeling.

Significant numbers of **missing values** were identified in the `director`, `cast`, and `country` columns. To preserve the dataset's size, these were handled by imputing a placeholder string ("Unknown") for `director` and `cast`, and filling the missing `country` values with the dataset's mode ("United States"). A few rows with missing `rating` and `date_added` values were dropped due to their small number. Additionally, data type errors, such as leading whitespace in the `date_added` column, were corrected to ensure proper date conversion.

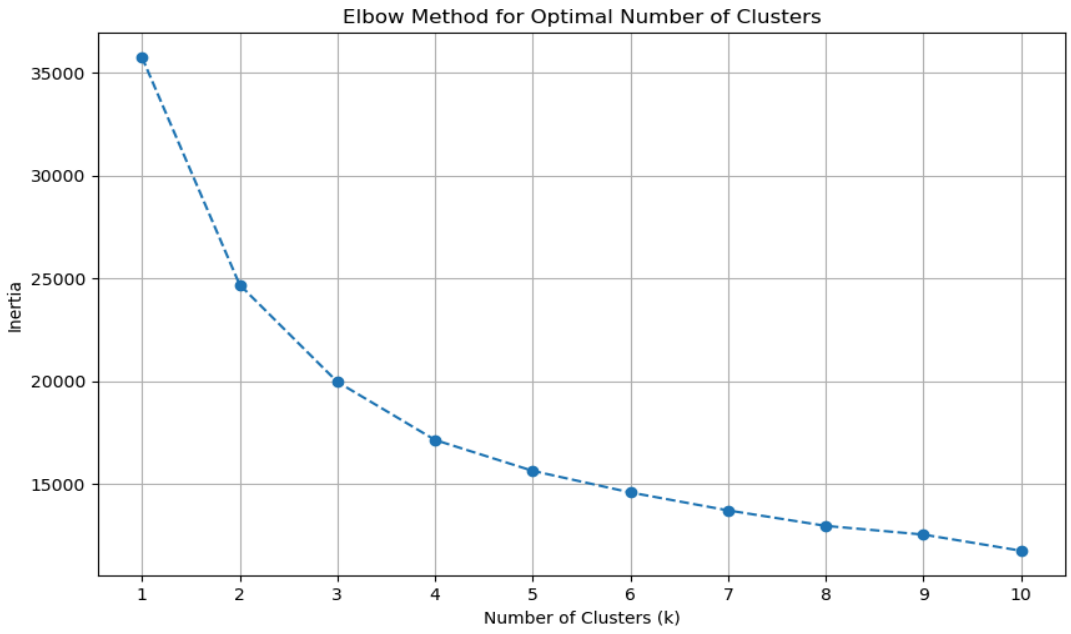
Following the cleaning process, **feature engineering** was performed to create more valuable features for the clustering model. The `duration` column, which contained mixed formats (e.g., "93 min", "2 Seasons"), was transformed into two distinct numerical columns: `duration_min` for movies and `duration_seasons` for TV shows. A new

feature, `content_age`, was also created by calculating the difference between the content's `release_year` and a modern reference year, providing a clear numerical measure of its age. The original `duration` and `release_year` columns were then dropped, resulting in a clean and feature-rich dataset ready for the final preprocessing steps.

Modeling and Evaluation

With the data cleaned and features engineered, the final step before modeling was to convert all features into a purely numerical format. The **`listed_in`** (genre) column was processed using **TF-IDF Vectorization**, while the **`rating`** column was transformed using **one-hot encoding**. All numerical features were then scaled using a **StandardScaler** to ensure uniform influence on the model. This resulted in a final feature matrix with 61 dimensions.

K-Means clustering was selected as the primary algorithm for this analysis. To determine the optimal number of clusters, the **Elbow Method** was employed. By running the K-Means algorithm for a range of cluster counts (1 through 10) and plotting the corresponding inertia, it was clear that the point of diminishing returns—the "elbow"—occurred at **k=3**. This suggested that three distinct clusters was the most appropriate grouping for this dataset.



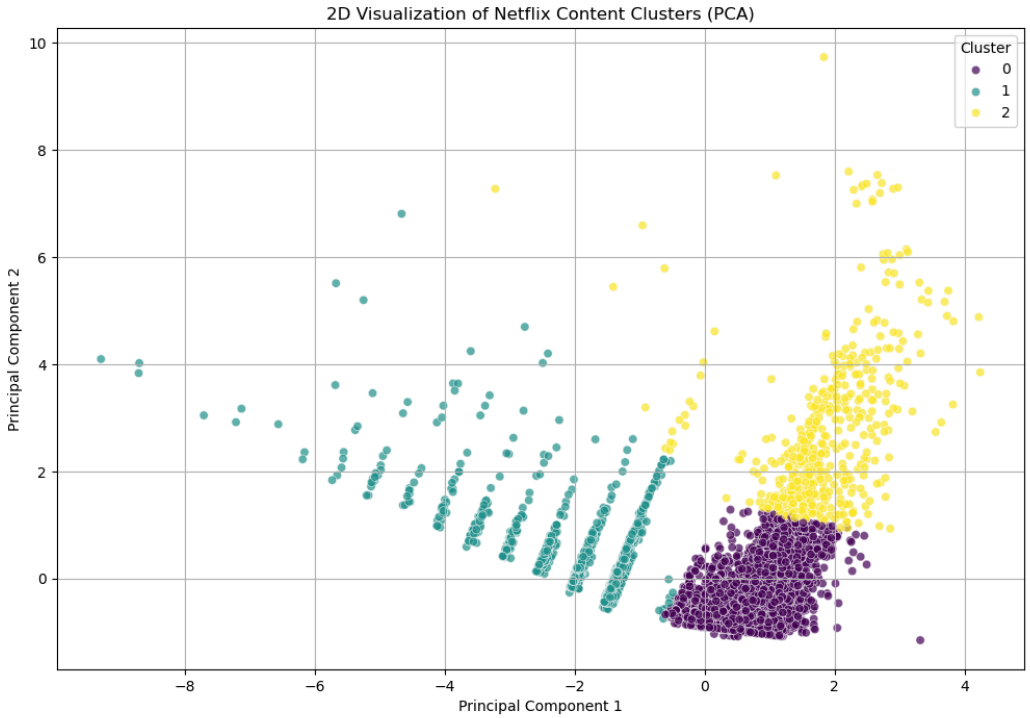
The performance of the trained K-Means model was assessed using two standard evaluation metrics:

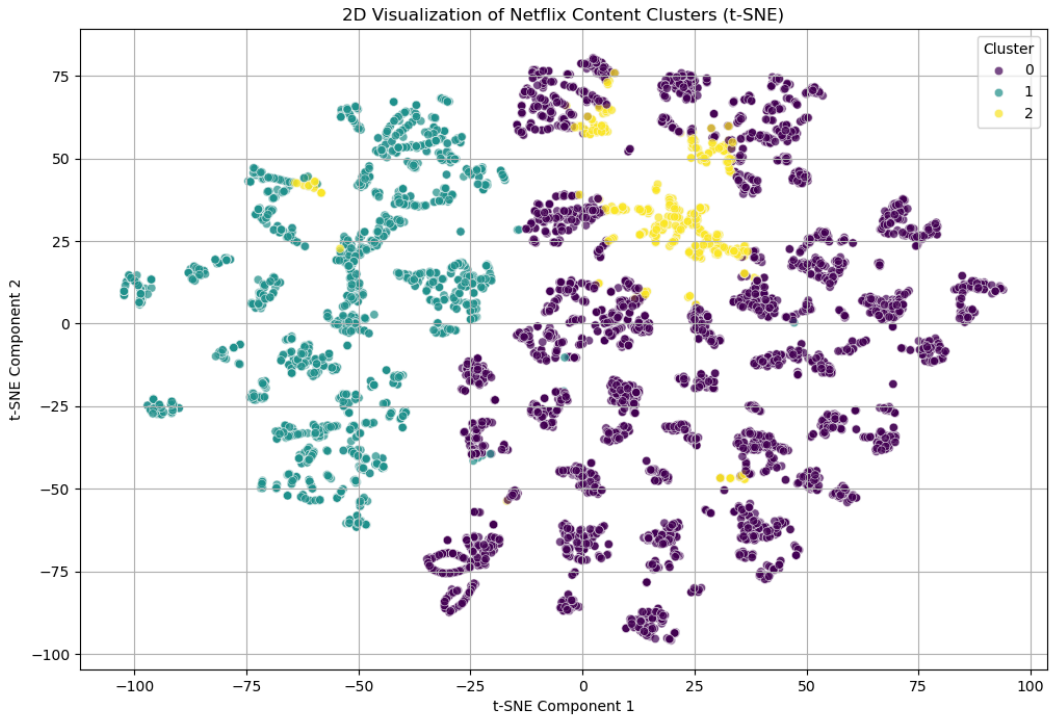
- **Silhouette Score**: 0.3530 (This positive score indicates that the clusters are reasonably dense and well-separated.)
- **Davies-Bouldin Index**: 1.1235 (A relatively low score, further supporting the validity of the cluster separation.)

To validate these findings, a second model, **Hierarchical Clustering**, was also trained on the same data. It produced an identical **Silhouette Score of 0.3530**, which provides strong confirmation that the three identified clusters represent a stable and meaningful structure within the data.

Results and Cluster Profiles

The evaluation metrics suggested that the K-Means model successfully identified a meaningful structure within the data. This is visually confirmed by reducing the 61-dimensional feature space into two dimensions using both **Principal Component Analysis (PCA)** and **t-SNE**. The t-SNE plot, in particular, shows a clear and distinct separation between the three clusters.





Cluster 0: The Modern Movie Collection

This is the largest cluster, containing **4,903 titles**. It consists almost exclusively of **movies** with an average runtime of **98 minutes**. The average `content_age` is approximately **9.7 years**, indicating that this cluster represents the core of Netflix's library of modern films from the last decade.

Cluster 1: The TV Show Universe

This is the second-largest group, with **2,389 titles**. This cluster is defined by its format, consisting entirely of **TV Shows**. On average, these are the most recently added titles, with a `content_age` of **8.5 years**, and an average length of **1.75 seasons**.

Cluster 2: The Classic Film Library

This is the smallest and most distinct cluster, containing **478 titles**. It is a niche collection of significantly older **movies**, with an average `content_age` of **38.5 years**. These films also tend to be slightly longer than their modern counterparts, with an average runtime of **111 minutes**.

Conclusion & Business Implications

This project successfully demonstrated the power of unsupervised machine learning to uncover meaningful patterns within the Netflix content library. By systematically cleaning the data, engineering relevant features, and applying clustering algorithms, the analysis segmented the dataset into three stable and distinct groups: a large **Modern Movie Collection**, a comprehensive **TV Show Universe**, and a niche **Classic Film Library**. The validity of these clusters was confirmed through strong evaluation metrics and the consistent results produced by both K-Means and Hierarchical Clustering models.

The insights gained from this analysis directly address the project's initial business goals. The identified clusters can be used to significantly enhance **personalized recommendations** by suggesting content to users from within the same clusters they frequently watch. The discovery of the "Classic Film Library" serves as a key example of **identifying a niche category**, allowing for targeted marketing and content curation for that specific audience. Furthermore, understanding the size and characteristics of these content groups provides valuable insights into **market trends**, assisting in advertising strategies and helping Netflix and its production partners identify potential **content gaps** in the library.

Ultimately, this clustering model provides a valuable framework for understanding the Netflix library not just as a collection of individual titles, but as a structured ecosystem of content categories.