Netflix Movies and TV Shows Clustering – Project Report

Objective

The goal of this project is to apply clustering techniques on the Netflix Movies and TV Shows dataset to identify meaningful groups of content based on numerical and textual features.

Dataset Summary

Total records: 7,787
Key feature used:

- **Numerical:** release_year, duration_num, genre_count, content_age
- **Categorical:** type, rating (encoded)
- **Text:** description, listed_in (genres)

Data Preprocessing

- Missing values were handled using simple strategies (mode / derived values).
- Duration was converted into a numeric format.
- New features created:
  - **Content age** = current year − release year
  - **Genre count** = number of genres per title

Text columns were converted into numerical form using **TF-IDF Vectorization**.

Feature Engineering & Reduction

- TF-IDF was applied to `description` and `listed_in`.
- All features were combined into a single feature matrix.
- **Truncated SVD** was used to reduce dimensionality and improve clustering performance.
- Features were scaled using **StandardScaler**.

Clustering Approach

Algorithm Used

- **K-Means Clustering**
- Initialization: **k-means++**
- Number of initializations: **n_init = 10**

Choosing Number of Clusters

- **Elbow Method** showed a clear bend at **k = 3**.

- **Silhouette Score** was highest at **k = 3**.

Hence, **3 clusters** were selected.

Clustering Evaluation Metrics

Best Model (Selected)

- **Algorithm:** K-Means
- **Initialization:** k-means++
- **Number of clusters (k):** 3
- **Number of initializations:** 10

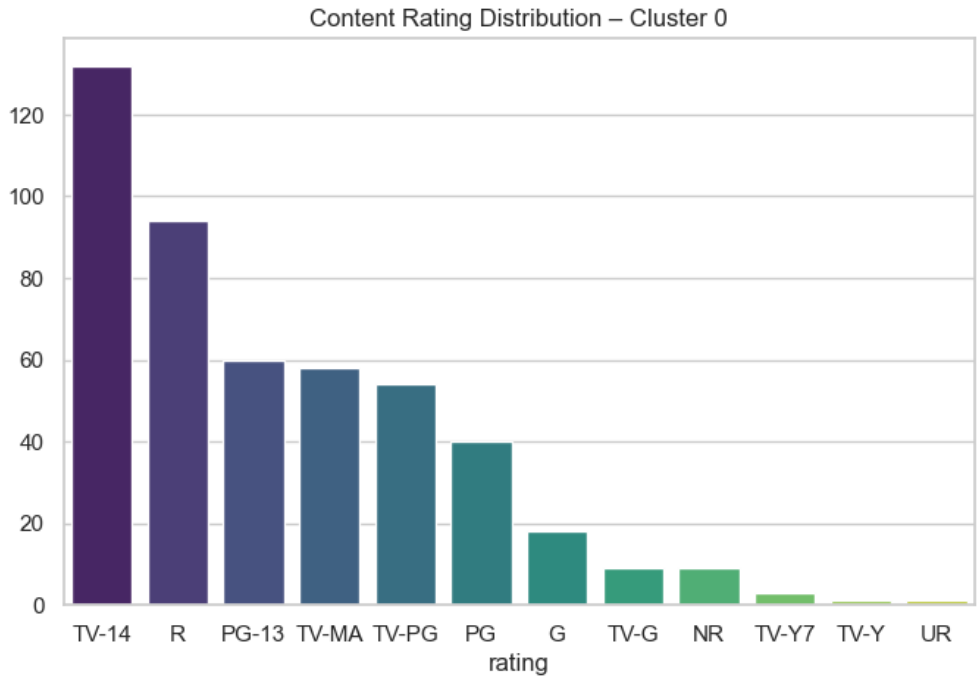
Metric Values

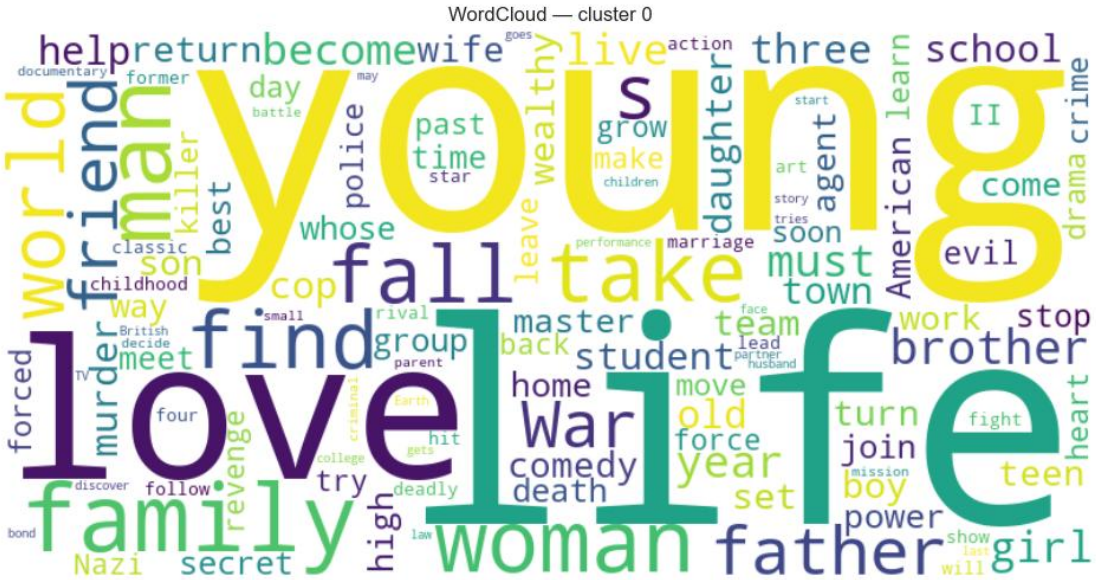
- **Silhouette Score: 0.36** → Good cluster separation
- **Inertia: 26,944.79** → Compact clusters without overfitting
- **Calinski–Harabasz Index: 3,663.24** → Strong cluster structure
- **Davies–Bouldin Index: 1.06** → Lower value indicates better clustering

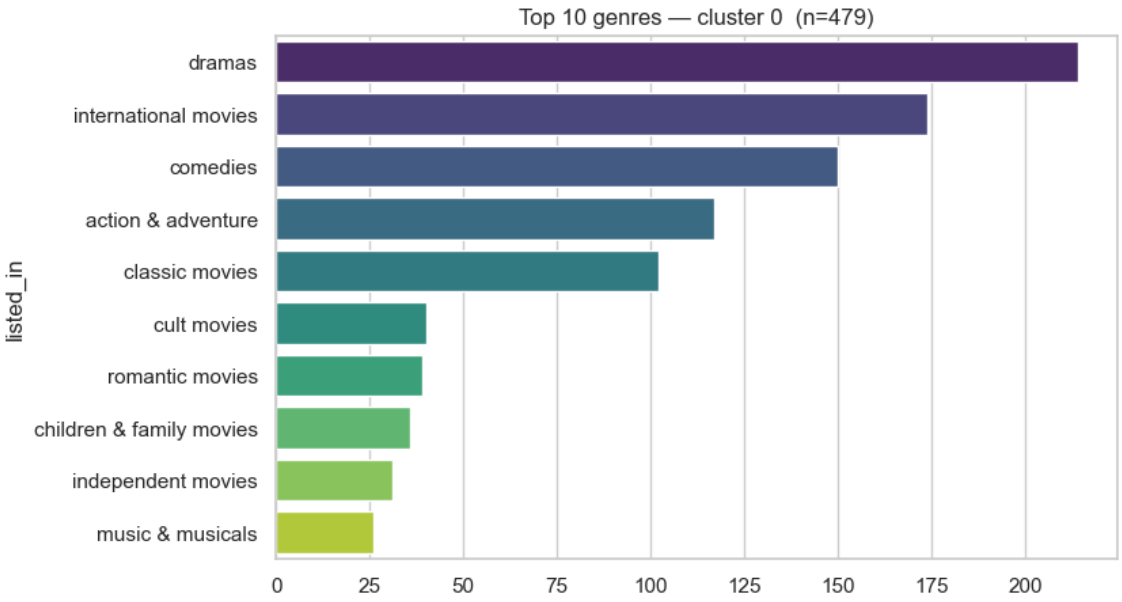
Results

- The chosen model produced well-separated clusters.
- PCA visualization confirmed clear cluster separation.
- Evaluation metrics supported the selected configuration.

Cluster 0: Older, long-duration content





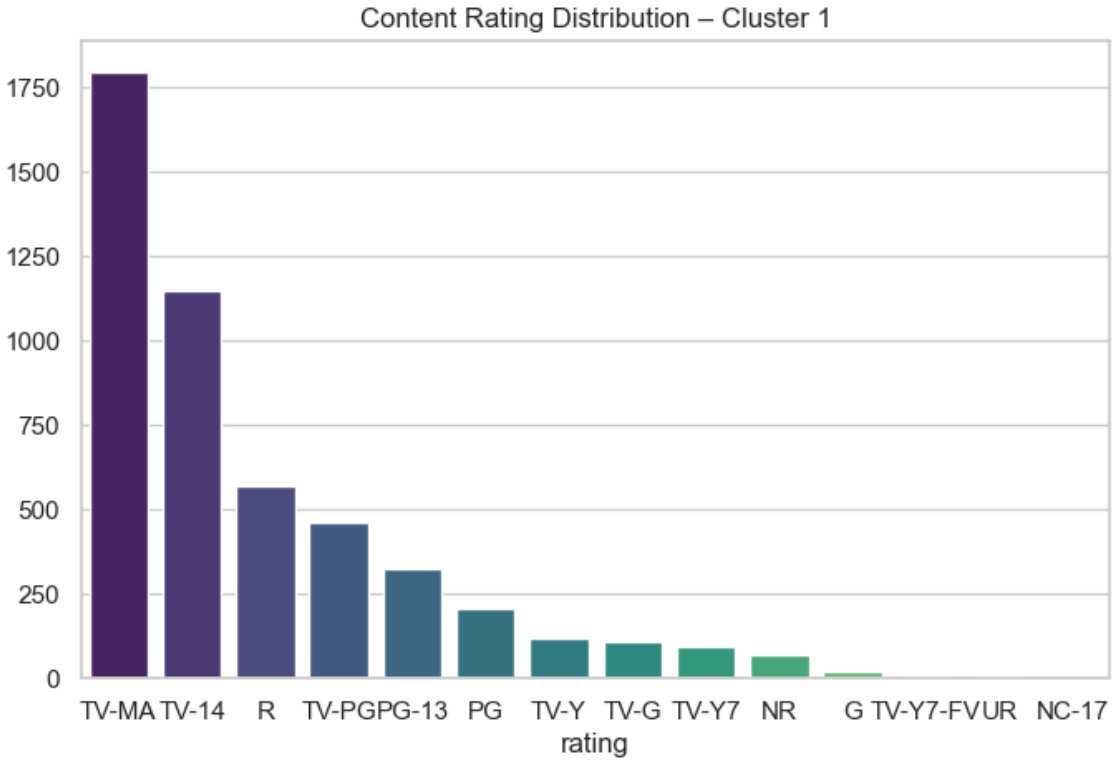


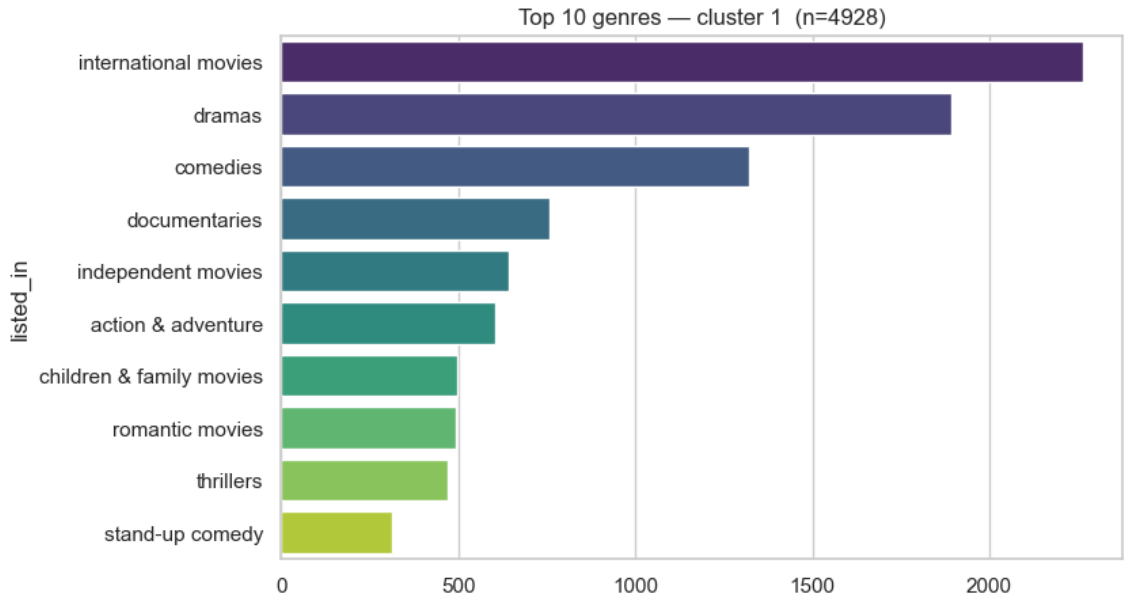
Cluster 0 mainly contains drama-based and classic content, with dramas, international movies, and comedies being the most common genres.

The WordCloud shows emotional and story-driven themes such as love, family, war, and relationships, indicating serious and narrative-focused content.

Most content is rated TV-14, R, and PG-13**,** showing that this cluster is targeted toward teen and adult audiences**,** not children.

Cluster 1: Modern mainstream Netflix content



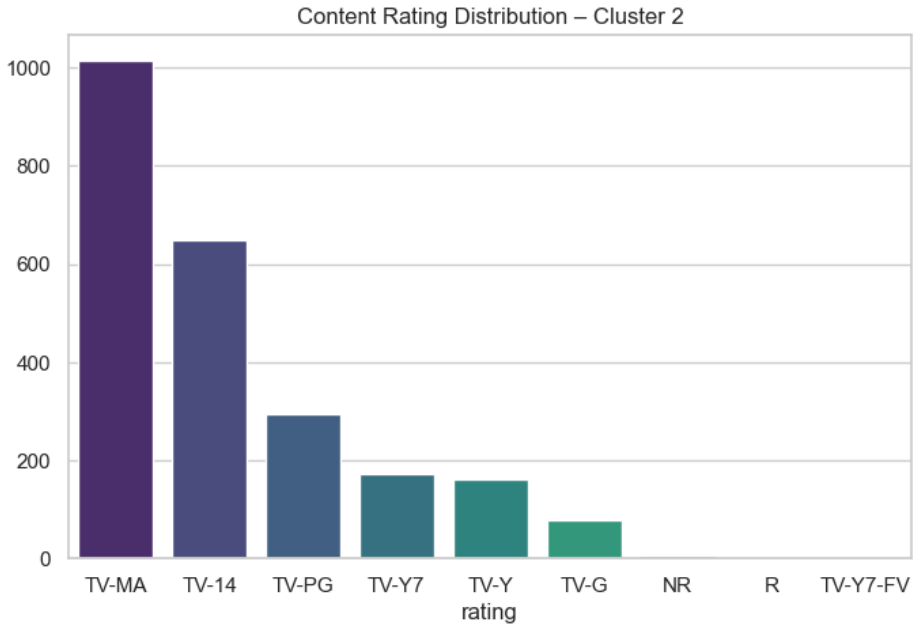


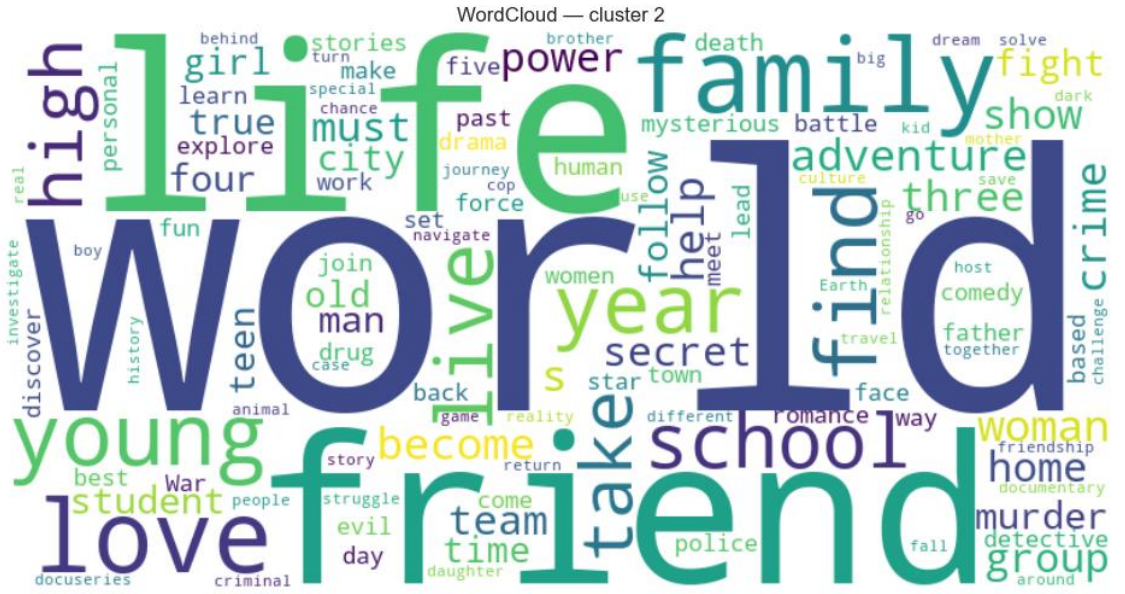
Cluster 1 represents the main Netflix catalog, dominated by international movies, dramas, and comedies.

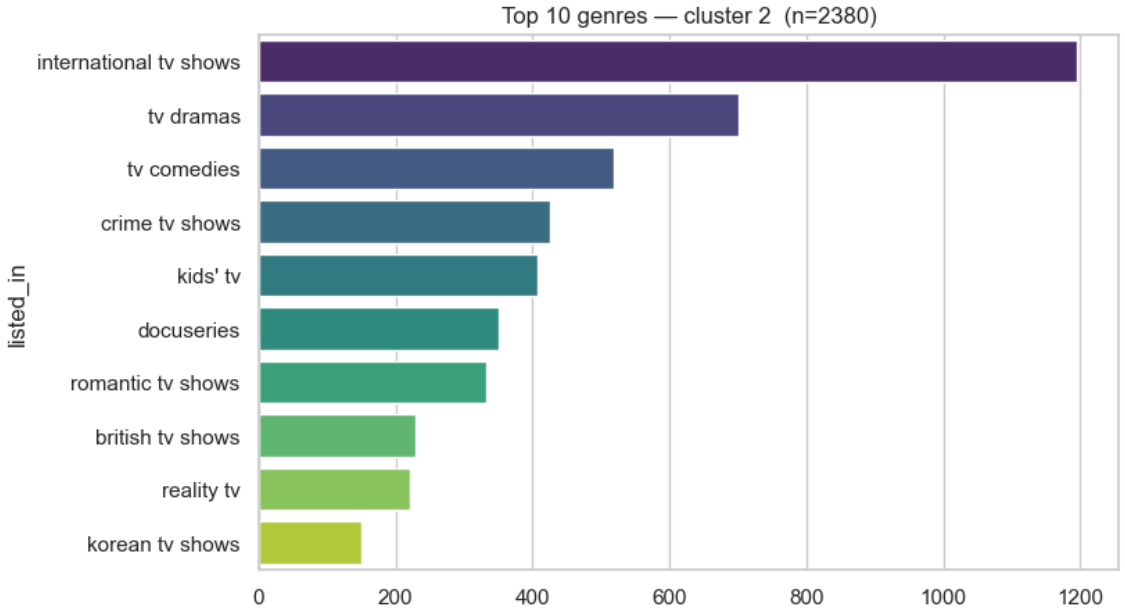
The WordCloud shows everyday and relationship-based themes such as life, family, friends, love, and youth, indicating general audience content.

Most content is rated TV-MA and TV-14, showing this cluster targets teen and adult viewers with a wide variety of content types.

Cluster 2: Very recent content, short duration







Cluster 2 is mainly TV-show focused, with international TV shows, TV dramas, and TV comedies being the most common genres.

The WordCloud highlights family, friends, school, and youth-related themes**,** indicating ongoing series and character-driven storytelling.

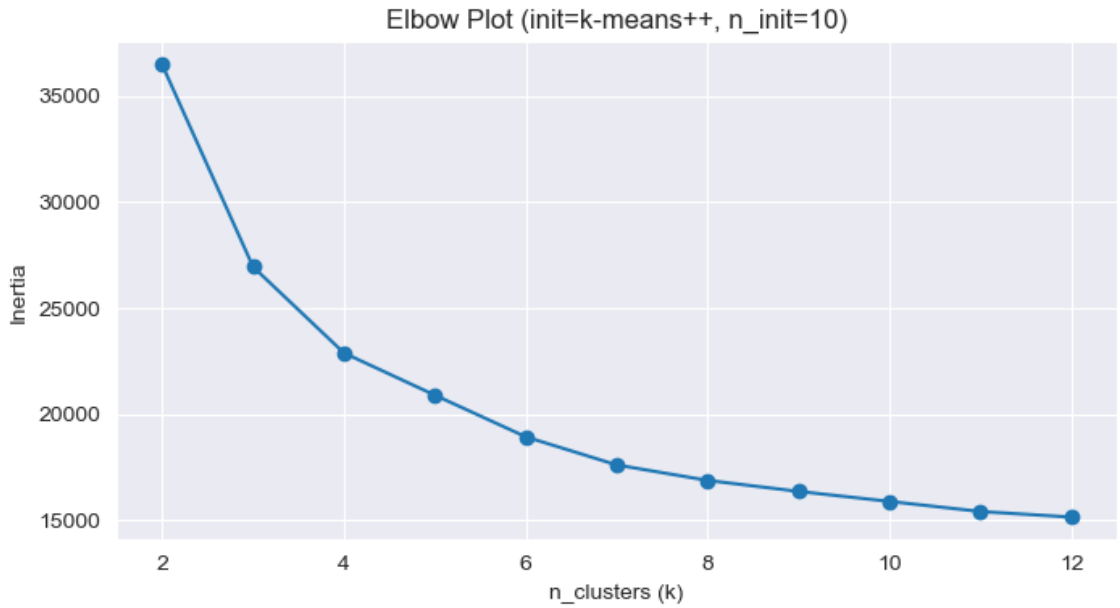
Most content is rated TV-MA and TV-14**,** showing this cluster targets teen and adult audiences, with some kids' TV also present.

Key Observations

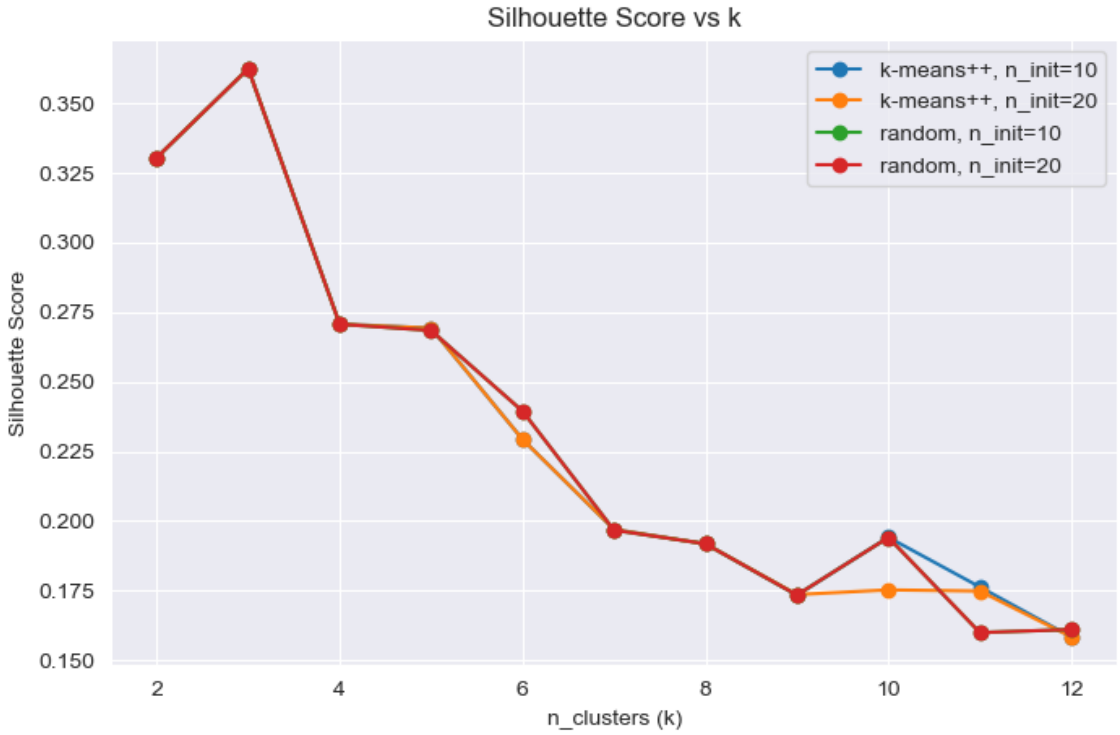
- Release year and content age strongly influence clustering.
- Duration helps separate long-form and short-form content.

Visual Analysis

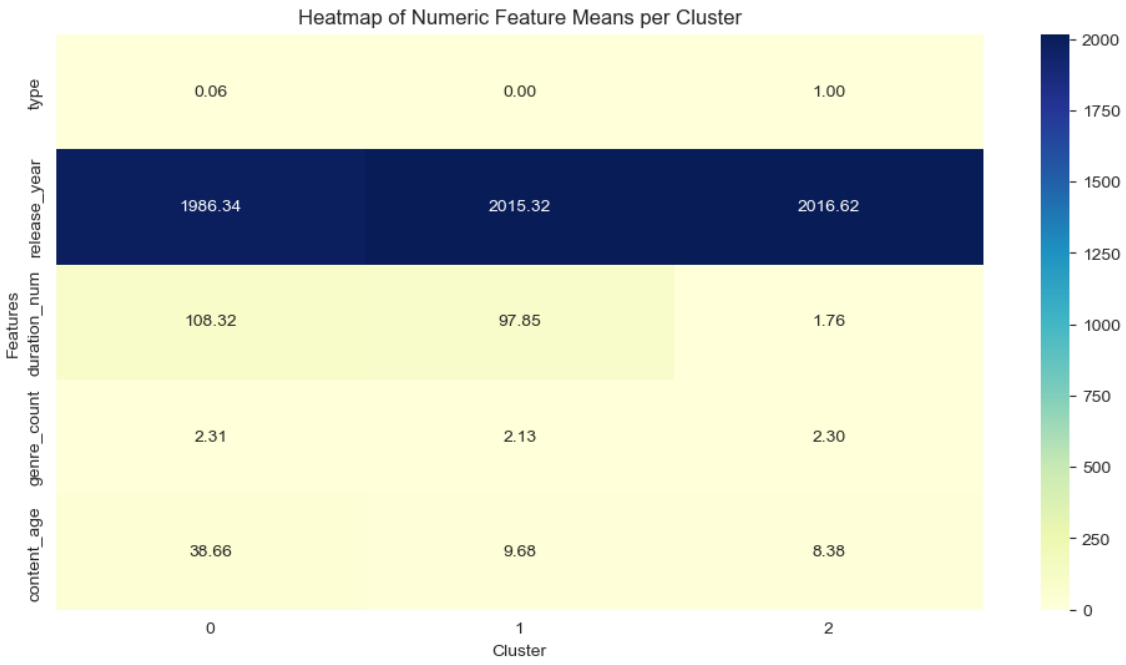
Elbow plot: Helped identify optimal k



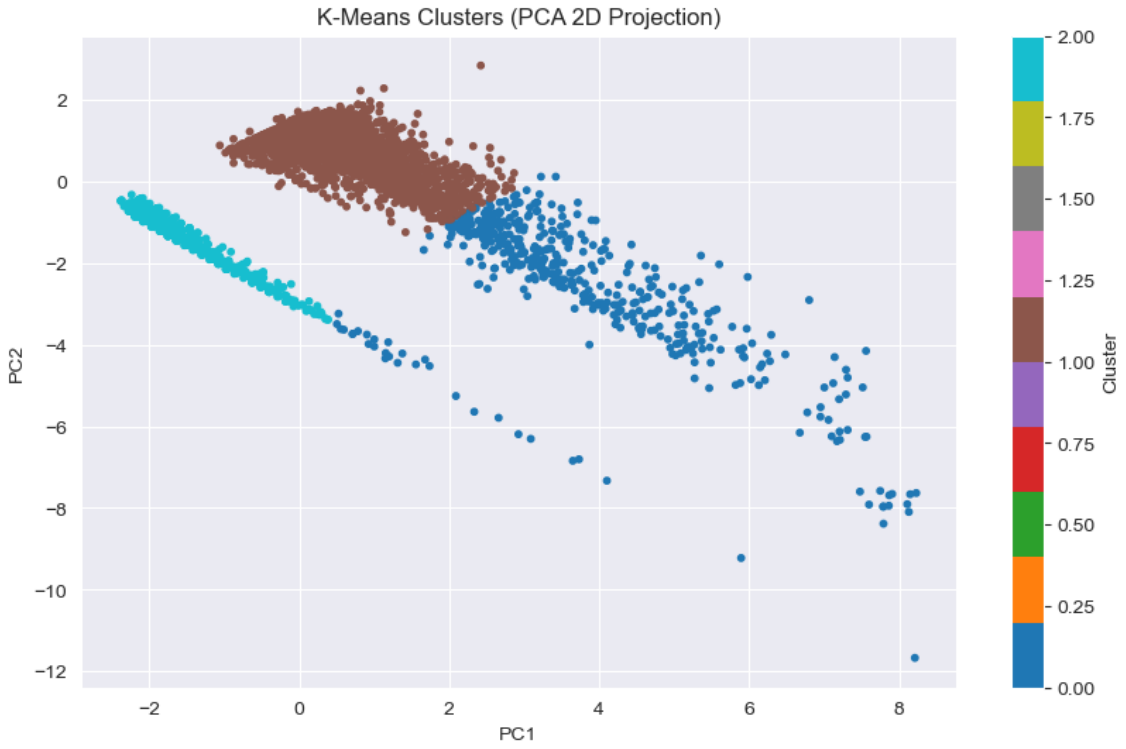
Silhouette plot: Validated cluster quality



Heatmap: Showed numerical difference across clusters



PCA scatter plots: Visualized cluster separation in 2D



Conclusion

The clustering approach successfully grouped Netflix content into three meaningful clusters, providing insights into content age, duration, and structure.