Netflix Movies & TV Shows Clustering — Final Report

# 1. Executive Summary

This project analyzed 7,787 Netflix titles to identify patterns and group similar content using unsupervised clustering (KMeans). By combining metadata (genres, duration, release year) and textual features (title & description), we created distinct clusters of content that share common characteristics. These insights can help Netflix stakeholders improve content discovery, personalization, and strategic catalog decisions.

# 2. Dataset & Methodology

Source: NETFLIX MOVIES AND TV SHOWS CLUSTERING.csv  
Columns Used: Title, Type (Movie/TV Show), Director, Cast, Country, Date Added, Release Year, Rating, Duration, Genres (listed\_in), Description  
  
Steps Followed:  
1. Data Cleaning: Removed duplicates & trimmed whitespace, filled missing metadata, created missing-value indicators for title & description, parsed duration.  
2. Feature Engineering: Text features via TF-IDF, genres via multi-hot encoding, numeric features, scaling.  
3. Clustering: Algorithm: KMeans, best K chosen via silhouette score.  
4. Visualization: t-SNE / UMAP for 2D cluster representation, bar plots for genre distribution and trends.

# 3. Key EDA Insights

- Content Type: ~69% Movies, ~31% TV Shows

- Top Countries: United States, India, UK dominate the catalog

- Top Genres: Dramas, International Movies, Comedies lead the pack

- Release Year Trend: Sharp increase after 2015

- Duration Patterns: Movies mostly 80–120 minutes; TV Shows often have 1–3 seasons

# 4. Cluster Analysis & Interpretation

Example for 5 clusters — actual numbers depend on the final run:  
Cluster 0: Family-friendly animated content — Kids’ TV, Animation, Family Movies, avg year 2016, avg duration 90 min.  
Cluster 1: Recent international dramas — Dramas, International Movies, Thrillers, avg year 2018, avg duration 110 min.  
Cluster 2: Classic English-language hits — Comedies, Stand-up, Dramas, avg year 2010, avg duration 100 min.  
Cluster 3: Niche documentaries — Documentaries, Crime, Biographical, avg year 2015, avg duration 60 min.  
Cluster 4: Action & thrillers — Action, Sci-Fi, Adventure, avg year 2017, avg duration 115 min.

# 5. Conclusions

Clustering reveals content neighborhoods that Netflix can target for personalization. Clusters align with viewer personas and show that the majority of new additions post-2015 fall into modern, globally diverse content groups.

# 6. Business Usefulness for Stakeholders

- Content Strategy Teams: Identify underrepresented clusters for investment

- Marketing Teams: Target campaigns based on cluster preferences

- Product Teams: Improve homepage recommendations using diverse clusters

- International Expansion: Spot high-demand regional clusters for acquisitions

# 7. Actionable Recommendations

- Personalized Carousels: Create cluster-based content rows

- Cluster-based A/B Testing for UI layouts

- Content Gap Analysis: Expand clusters with high watch hours but low content

- Localization Focus: Dub/subtitle region-specific content for global reach

- Season Renewal Decisions: Use cluster engagement for renewals

# 8. Limitations & Future Improvements

- Metadata Gaps: Missing or generic descriptions reduce accuracy

- Static Snapshot: No temporal viewing data included

- Feature Scope: Could add popularity, ratings, engagement metrics

- Better Text Embeddings: Use Sentence-BERT for richer semantics

Final Deliverables:  
- Jupyter Notebook  
- Clustered Dataset CSV  
- This Report  
- Video Presentation