**Netflix Content Analysis Report**

**Project**: Netflix Content Strategy Analyzer

**Milestone 1**: Requirements & Dataset Preparation (Week 1&2)

**1.Introduction and Project Overview**

The objective of this project is to analyze the Netflix Movies and TV Shows dataset obtained from Kaggle. The dataset contains detailed information about thousands of titles, including movies and TV shows available on Netflix. The primary focus of this analysis is on data preprocessing, cleaning, and extracting meaningful insights from the dataset.

The aim of this milestone is:

* Define project scope and success metrics.
* Load the Netflix Kaggle dataset.
* Clean the dataset (handle missing values, remove duplicates).
* Normalize categorical features such as genre, rating, and country.

**2. Dataset Overview**

The dataset comprises 8000+ of titles with multiple attributes (Movie/TV Show), director, cast, country, release year, rating, duration, and listed genres .Before and after preprocessing, the dataset shape changed slightly due to removal of redundant and inconsistent records. Additionally, new features were engineered to strengthen analysis.

**3. Data Cleaning & Preprocessing**

Data cleaning is a crucial step to improve dataset quality and reduce noise and ensures that the dataset is ready for analysis. The following procedures were applied:

**3.1 Loading the dataset**  
The dataset was imported directly from Kaggle into the working environment. Pandas’ read\_csv() function was used to load the data into a DataFrame, which provided a convenient structure for analysis and cleaning. Initial exploration (df.info() and df.head()) helped identify missing values, column types, and overall data quality.

**3.2 Deleting redundant columns**  
Columns that did not add analytical value or contained too many missing values were dropped. For example, if two columns carried duplicate information (like show\_id not being needed for analysis), they were removed. This reduced dataset complexity and improved processing efficiency.

**3.3 Renaming columns**  
To maintain consistency and readability, column names were standardized. For instance, spaces in column names such as "listed\_in" were kept lowercase with underscores for easier reference in code (e.g. release\_year, duration).

**3.4 Removing duplicates**  
Duplicate rows were identified using df.duplicated() and removed to avoid double counting of the same title.

**3.5 Handling NaN values**  
Missing values were carefully handled depending on the column type:

* For categorical fields like director, cast, and country, missing entries were replaced with "Unknown" to preserve rows.
* For critical columns like type, release\_year, and rating, rows with missing values were dropped to maintain accuracy.  
  This ensured the dataset remained both comprehensive and reliable.

**4. Netflix Content Metrics and Insights**

Analysis of content distribution by genre combination highlighted Netflix’s strategic focus on certain categories:

4.1 Genre Focus & Distribution

* Dramas + International Movies dominate the library with 362 titles (≈4.12%), reflecting Netflix’s strong investment in cross-border storytelling and culturally diverse narratives.
* Documentaries are nearly at par with 359 titles (≈4.08%), showing Netflix’s emphasis on non-fiction and educational content alongside entertainment.
* Stand-up Comedy contributes 334 titles (≈3.8%), highlighting Netflix’s push into original specials and global stand-up talent.
* Titles blending Comedies, Dramas, and International Movies (274 titles, ≈3.12%) suggest that hybrid genres are an important strategy to appeal to multiple audience segments at once.

4.2 Content Strategy Insights

* Global-first strategy: The frequent pairing of International Movies across top categories indicates Netflix’s reliance on diverse cultural content to expand and retain global subscribers.
* Balanced Portfolio: While dramas dominate, Netflix balances this with strong representation of comedy and documentaries, which together account for over 11% of the total catalog.
* Audience Retention Mix: The mix of Drama-heavy and Comedy-focused titles suggests Netflix is targeting both binge-watchers (dramas) and casual, short-session viewers (stand-up comedy).

4.3 Performance Metrics

* Genre Concentration: Top 5 genre combinations together represent ~18% of the catalog, meaning a relatively small set of categories drive a significant portion of Netflix’s offering.
* Variety vs. Depth: With nearly equal weightage to Dramas (fiction-heavy) and Documentaries (non-fiction-heavy), Netflix is hedging content risk by catering to both escapism and reality-driven viewing preferences.
* Originals Impact: Genres like Stand-up Comedy and Documentaries are often Netflix originals, hinting that a big slice of their exclusive IP lies in these formats.

**5 Cleaning & Normalization**

Categorical features like Country, Rating, and Genre often come in inconsistent formats (uppercase, lowercase, extra spaces, duplicates). To normalize them, the following steps are applied:

* Case Standardization → Converting all titles first letter to uppercase .
* Whitespace Removal → Removing trailing/multiple spaces ("drama " → "drama")
* Category Merging → Combining synonymous or overlapping categories
* Feature Engineering → Splitting multi-valued columns (e.g., "Drama, International Movies") into separate tags for better analysis.

**Benefits of Cleaning & Normalization**

* Consistency: Prevents the same category from being treated as different due to spelling/case differences.
* Accuracy in Analysis: Reduces noise and ensures reliable statistics (e.g., "Drama" and "dramas" are counted together).
* Better Model Performance: Machine learning models handle standardized data more efficiently.
* Improved Comparisons: Easier to compare across genres, ratings, or countries without misclassification.
* Reduces Redundancy: Merges duplicate or similar labels, simplifying insights.

**📊 Conclusions**

* Content Focus → Netflix’s catalog is heavily driven by dramas and international movies, showing its global-first approach.
* Balanced Strategy → Equal weightage to fiction (dramas) and non-fiction (documentaries) reflects diversification in storytelling.
* Comedy Strength → Stand-up comedy is a unique differentiator, reinforcing Netflix’s brand as an entertainment hub beyond traditional shows and films.
* Indie Support → A smaller but notable share of independent films shows Netflix’s positioning as a platform for niche and artistic content.
* Strategic Mix → The top 5 genre clusters together form ~18% of the total catalog, showing that while Netflix offers variety, a few genres dominate viewer attention.
* Data Cleaning Impact → Without cleaning and normalization, these insights would be fragmented or misleading, since genres and categories could be split incorrectly.