**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 8 thousand 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.
* Provide **insights and metrics** to summarize key trends.

**2. Success Metrics**

* All missing and inconsistent values are handled.
* The dataset is free of duplicates and formatted consistently.
* Categorical variables such as **content\_type**, **genres**, **country**, **rating**, and **duration\_type** are encoded into numeric representations.
* Numerical features such as **duration\_int**, **release\_year**, and **date\_added** are scaled or frequency-encoded.
* The final normalized dataset is stored as a CSV file (netflix\_normalized\_full.csv) ready for analysis or modeling.
* Descriptive insights and metrics highlight meaningful patterns in the data.

**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 original dataset was downloaded from Kaggle using:

!kaggle datasets download -d shivamb/netflix-shows -p /workspace/netflix --unzip

It contains the following main columns:

* show\_id, type, title, director, cast, country, date\_added, release\_year, rating, duration, listed\_in, description.

**3.2** **Removed unnecessary columns**: title, director, cast, description, and show\_id.

**3.3** **Renamed columns**:

* type → content\_type
* listed\_in → genres

**3.4** **Handled missing values**:

* country → filled with "Unknown"
* rating → filled with "Not Rated"
* duration → filled with "Unknown"

**3.5** **Standardized duration column**:

* Unified "Seasons" → "Season"
* Split into duration\_int (numeric value) and duration\_type (unit: min or Season)

**3.6** **Removed duplicates** and stripped whitespace from categorical fields.

**3.7** Reset the index after cleaning.

**4. Dataset Comparison**

**Before Cleaning**

* Columns had missing values in country, rating, and duration.
* Duration was inconsistent (e.g., “1 Season”, “2 Seasons”, “90 min”).
* Contained duplicate rows and unnecessary columns.

| **Metric** | **Value (Example)** |
| --- | --- |
| Total rows | ~ 7,787 |
| Missing in country | ~ 831 |
| Missing in rating | ~ 4 |
| Missing in duration | ~ 3 |
| Duplicate rows | Present |

**After Cleaning**

* No missing values in key columns.
* All duplicates removed.
* Duration standardized and split into numeric & type columns.

| **Metric** | **Value (Example)** |
| --- | --- |
| Total rows | ~ 7,774 |
| Missing in country | 0 |
| Missing in rating | 0 |
| Missing in duration | 0 |
| Duplicate rows | 0 |

**5. Normalization**

We normalized categorical and numeric features for better modeling:

* **Categorical Normalization:**
  + content\_type → **Label Encoded** (Movie=0, TV Show=1)
  + country → **Frequency Encoded** (# of titles from each country)
  + rating → **Ordinal Encoded** (e.g., G=0, PG=1, …, TV-MA=10, Not Rated=-1)
  + genres → **Frequency Encoded** (based on main/first genre)
  + duration\_type → **Label Encoded** (min=0, Season=1)
* **Numeric Normalization:**
  + release\_year → Frequency Encoded (# of titles per year)
  + duration\_int → **Min-Max Scaled** (0–1)
  + date\_added → Converted to number of days since earliest date and scaled.

**After Normalization**

* All columns are numeric and machine-learning ready.
* Preserved essential information but standardized formats.

| **Feature Example** | **Transformation** |
| --- | --- |
| content\_type | Movie → 0, TV Show → 1 |
| rating | TV-MA → 10, G → 0, etc. |
| country | US → 2800, UK → 650 |
| duration\_int | Scaled between 0–1 |
| genres | Frequency of genre |

**6. Key Insights & Metrics**

Here are some descriptive insights derived from the cleaned dataset:

**1. Content Distribution**

* **Movies:** ~70% of total titles
* **TV Shows:** ~30% of total titles

**2. Top Countries with Most Titles**

| **Rank** | **Country** | **Titles Count** |
| --- | --- | --- |
| 1 | United States | ~2,800 |
| 2 | India | ~900 |
| 3 | United Kingdom | ~650 |
| 4 | Canada | ~400 |
| 5 | Japan | ~350 |

The majority of titles are from the **United States**, followed by **India** and the **United Kingdom**.

**3. Ratings Distribution**

* **Most common rating:** **TV-MA** (suitable for mature audiences)
* Other frequent ratings: **TV-14**, **TV-PG**, **R**

**4. Genre Insights**

* Top genres include **International Movies**, **Dramas**, **Comedies**, and **Action & Adventure**.
* Genre frequency encoding helps identify popular categories for analysis.

**5. Release Year Trends**

* Peak content production was between **2015 – 2020**, especially for **Movies**.
* TV Shows saw rapid growth after 2017.

**6. Duration Insights**

* Most movies have a duration between **90–120 minutes**.
* TV Shows usually have **1–2 Season.**

**7. Conclusion of Milestone 1**

* Defined a clear preprocessing pipeline for the Netflix dataset.
* Produced clean and normalized data for further analytics or predictive modeling.
* Extracted key insights such as **content distribution, top-producing countries, popular ratings, and genre trends**.