**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.**

**Milestone 2:**  EDA & Feature Engineering (Week 3 & 4)

**1. Exploratory Data Analysis (EDA)**

After cleaning and normalizing the dataset, Exploratory Data Analysis (EDA) was conducted to understand key patterns, distributions, and relationships in the Netflix catalogue. The analysis focused on content growth trends, genre composition, rating distribution, and geographic contribution.

**1. Content Growth Over Time**

* The number of titles increased steadily from 2010, with the sharpest growth observed between **2015 and 2020**.
* **TV Shows** saw rapid expansion after **2017**, reflecting Netflix’s strategic shift toward serialized and original content.

**2. Content Type Distribution**

* **Movies** constitute approximately **70%** of the total titles.
* **TV Shows** make up the remaining **30%**, showing significant growth in recent years.

**3. Genre Analysis**

* The most frequent genres are **International Movies**, **Dramas**, **Comedies**, and **Action & Adventure**.
* The dominance of international and dramatic content highlights Netflix’s global audience targeting.

**4. Ratings Distribution**

* The most common rating is **TV-MA**, followed by **TV-14**, **TV-PG**, and **R**.
* This indicates that the majority of Netflix content caters to **mature audiences**, with limited family-oriented content.

**5. Country-Level Contribution**

* **Top countries** by content count:
  1. United States (~2,800 titles)
  2. India (~900 titles)
  3. United Kingdom (~650 titles)
  4. Canada (~400 titles)
  5. Japan (~350 titles)
* The U.S. remains the primary content producer, while India and the U.K. are strong emerging contributors.

**6. Duration Trends**

* Most **Movies** range between **90–120 minutes**.
* Most **TV Shows** have **1–2 Seasons**, indicating limited-series formats are common.

**2. Feature Engineering**

After EDA, **feature engineering** was performed to create new variables and improve the dataset’s predictive power. These derived features capture additional insights and relationships not directly visible in raw data.

**a. Date-based Features**

* **Date Added** was split into separate columns for **Year**, **Month**, and **Day** to analyze seasonal patterns in content addition.
* These components were converted into **integer values** for modeling compatibility.

**b. Duration-based Features**

* The original duration column was standardized into:
  + **duration\_int** → Numeric value (e.g., 90, 2, etc.)
  + **duration\_type** → Category (min or Season)

This separation allows clearer comparisons between movie lengths and show seasons.

**c. Content-Length Category**

* A new derived column **Content\_Length\_Category** was created using the duration\_int feature:
  + **Short** → less than 60 minutes
  + **Medium** → between 60–120 minutes
  + **Long** → more than 120 minutes  
    This helps group titles into meaningful viewing duration segments.

**d. Licensing and Production Type (Derived Feature)**

* A binary feature **Original\_vs\_Licensed** was derived based on whether the title’s description or metadata indicated **“Netflix Original”**.
  + 1 → Netflix Original Content
  + 0 → Licensed / Third-Party Content  
    This feature helps distinguish Netflix’s in-house productions from externally licensed titles.

**e. Genre Simplification**

* For multi-genre titles, the **first genre** was extracted as **main\_genre** for analysis and frequency encoding.
* This improves clarity when visualizing or modeling categorical data.

**f. Country Normalization**

* For titles with multiple countries, the **first country** was taken as the **primary production country** (main\_country).
* This ensures uniformity in country-level analysis.

**3. Visualization**

Visualizations were created to represent:

* **Content growth trend** (line plot)
* **Genre and rating distributions** (bar and pie charts)
* **Country-wise contributions** (bar chart)
* **Content type ratio (Movies vs TV Shows)** (pie chart)

These visualizations helped validate feature relationships and trends observed during analysis.

**Outcome of Feature Engineering**

* Dataset now includes multiple derived and encoded columns for better interpretability and modeling.
* Enhanced features such as Content\_Length\_Category, Original\_vs\_Licensed, and split date attributes provide deeper analytical insights.
* The processed dataset is ready for machine learning tasks such as recommendation systems, trend forecasting, or clustering analysis.