Data Analysis Report: Netflix Content Catalog

Project: DV: StreamScope - Netflix Content Strategy Analyzer
Module: Milestone 1: Data Preparation and Initial Insights (Week 1–2)

WEEK 1

1. Introduction and Project Objective

The primary objective of the **StreamScope** project is to develop a data-driven system that analyzes Netflix's vast content catalog. By focusing on data preparation and initial insights, this foundational report ensures the data is clean, reliable, and ready for advanced analysis in subsequent project phases.

The goal of this phase was to:

1. Quantify and clean the dataset.
2. Handle all missing and duplicate records.
3. Derive the first set of key insight metrics, specifically focusing on content distribution by genre.

2. Dataset Overview: Quantification of Shape

The analysis is performed on the **Netflix Movies and TV Shows Dataset (Kaggle)**, which details over 8,000 titles.

Metric	Before Preprocessing (Initial State)	After Preprocessing (Clean State)
Total Number of Titles (Rows)	8,807	**8,787**
Total Number of Features (Columns)	12	**14**
Data Reduction	N/A	20 total records removed (10 duplicates, 10 NaNs in date_added)

Insight: The cleaning process successfully retained approximately **99.77%** of the original data. A total of 20 non-essential records were safely removed (including duplicates and rows with critical missing dates), and two new features were engineered, providing a strong,

standardized dataset foundation with **14 features**.

3. Step-wise Data Cleaning & Preprocessing

Data cleaning is essential to address missing values and inconsistencies that can skew analytical results. The following steps were executed to prepare the dataset.

Step 3.1: Inspection and Missing Value Identification

The first action was to check which columns contained missing (null) data using a simple code snippet.

Reference Code Snippet (Python Pandas):

```
df.isnull().sum()
```

Statistical Values: Missing Data Count (Before Treatment)

Column Name	Type	Count of Missing Values	Handling Strategy
director	Categorical	2,634	Fill with 'Not Available'
cast	Categorical	825	Fill with 'Not Available'
country	Categorical	831	Fill with 'Unknown'
date_added	Categorical	10	**Remove Row (Drop NaN)**
rating	Categorical	4	Fill with Mode (most frequent rating)
duration	Categorical/Numeric	3	Fill with '0' (Placeholder)

Step 3.2: Handling Missing Values

Missing values in key categorical columns were replaced using specific placeholders (like 'Not Available' or 'Unknown') or by statistical imputation (Mode). Critical missing values in

date_added were handled by removing the corresponding rows to ensure temporal features are accurate.

Reference Code Snippet (Python Pandas):

```
# Fills director and cast with 'Not Available'
df['director'] = df['director'].fillna('Not Available')
df['cast'] = df['cast'].fillna('Not Available')

# Fills country with 'Unknown'
df['country'] = df['country'].fillna('Unknown')

# Row Removal for 'date_added': Removes 10 rows missing a critical addition date
df = df.dropna(subset=['date_added'])

# Mode Imputation for 'rating': Replaces missing values with the most frequent rating
mode_rating = df['rating'].mode()[0]
df['rating'] = df['rating'].fillna(mode_rating)

# Fills duration with '0' as a placeholder for further processing
df['duration'] = df['duration'].fillna('0')
```

Step 3.3: Duplicate Removal

Duplicate entries must be removed to ensure each row represents a unique title, occurring *after* the initial dataset load and *before* final cleaning, leading to the removal of 10 rows.

Reference Code Snippet (Python Pandas):

```
df.drop_duplicates(inplace=True) # Removes 10 duplicate rows
```

Step 3.4: Feature Normalization & Engineering

Categorical features (like country, rating, and listed_in) were standardized by converting text to a consistent case (e.g., lowercase) and removing extra spaces. This phase also involved **Feature Engineering**, resulting in the creation of 2 new columns, bringing the total feature count to **14**.

4. Week 1 Insight Metrics: Genre Distribution

To gain immediate understanding of Netflix's content library, we analyzed the distribution of content based on the primary genre combinations listed for each title. This provides insight

into the strategic focus of the catalog.

Rank	Primary Genre Combination	Count of Titles	Percentage of Total Catalog (approx.)
1	**Dramas, International Movies**	362	4.12%
2	**Documentaries**	359	4.08%
3	**Stand-Up Comedy**	334	3.80%
4	**Comedies, Dramas, International Movies**	274	3.12%
5	**Dramas, Independent Movies, International Movies**	252	2.87%

Key Insight:
The analysis reveals a heavy strategic reliance on Dramas and International Movies, often in combination, indicating that Netflix's content acquisition and production strategy is highly focused on global narratives and serious/compelling storytelling. Documentaries and Stand-Up Comedy also form major pillars of the library, reflecting specific, high-demand content niches.