# DV : StreamScope

## Netflix Content Strategy Analyzer: Insights into Global Streaming Trends

## Objective

To develop a data-driven system that analyzes Netflix’s content catalog using the Kaggle Netflix Movies and TV Shows dataset, focusing on data preparation and preprocessing during Week 1–2.

## Outcomes (Week 1–2)

* Load and inspect the Netflix dataset.
* Clean the dataset by handling missing values and removing duplicates.
* Normalize categorical features for downstream analysis.
* Prepare the dataset for Week 3–4 EDA and feature engineering.

## Dataset

**Netflix Movies and TV Shows Dataset (Kaggle)**

* Contains details of 8,000+ titles including type (Movie/TV Show), director, cast, country, release year, rating, duration, and listed genres.

## Data Cleaning & Preprocessing Steps

**1. Load Dataset:**

**import** pandas **as** pd  
df = pd.read\_csv("/Volumes/workspace/default/datasets/netflix\_titles.csv")

**2. Inspect Dataset:**

* Check column info, data types, and missing values.

df.info()  
df.isnull().sum()

**3. Handle Missing Values:**

* Fill missing categorical columns with ‘Unknown’:

df['director'].fillna('Unknown', inplace=True)  
df['cast'].fillna('Unknown', inplace=True)  
df['country'].fillna('Unknown', inplace=True)  
df['date\_added'].fillna(df['date\_added'].mode()[0], inplace=True)  
df['rating'].fillna('Not Rated', inplace=True)

**4. Remove Duplicates:**

df.drop\_duplicates(inplace=True)

**5. Normalize Categorical Features:**

* Standardize country, rating, and listed\_in columns (strip whitespace, lowercase).
* Convert duration into numeric values for movies (minutes) and TV Shows (number of episodes).

## Summary of Week 1–2 Progress

* Dataset loaded successfully with ~8,000+ titles.
* Missing values handled using filling strategies, preserving as much data as possible.
* Duplicate rows removed.
* Categorical features normalized for easier analysis in Week 3–4.
* Dataset is now clean, structured, and ready for exploratory data analysis.

## Tools & Tech Stack

* **Python**: pandas, numpy
* **IDE**: Databricks
* **Version Control**: GitHub