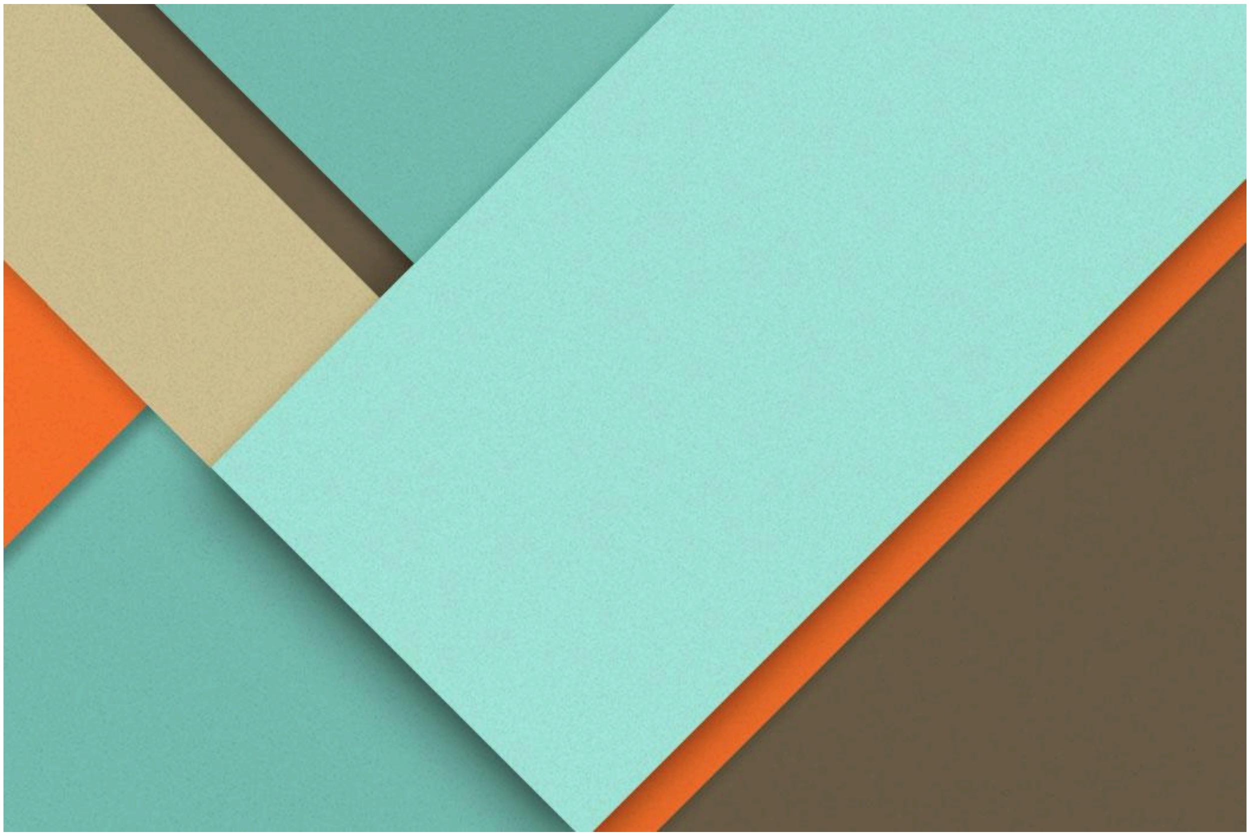

DV : StreamScope

Netflix Content Strategy Analyzer: Insights into Global Streaming Trends

Milestone 1-Netflix Data Cleaning & Insights Report

1. Importing the Dataset

- The dataset (netflix_titles.csv) was imported using pandas (pd.read_csv()).
- The initial display(df) helped to understand the structure of the data, the number of columns, and a preview of rows.

2. Handling Duplicates

- Function used: drop_duplicates()
- Before and after comparison of rows was done using df.shape[0].
- Insight: There were no duplicate rows in the dataset (rows before = rows after).

3. Identifying Missing Values

- Functions used:
  - ○ df.info() → To check datatypes and non-null counts.
  - ○ df.isnull().sum() → To count missing values in each column.
- Insight: Some columns like Director, Cast, Country, Date Added, Rating, and Duration had missing values.

4. Handling Missing Values

- Missing data was handled by filling with placeholders:
  - ○ Director → "Unknown"

  - Cast → "Not Available"
  - Country → "Unknown"
  - Date Added → "Unknown"
  - Rating → "Unrated"
  - Duration → "Unknown"
- Function used: fillna()
- Insight: After filling, the dataset had 0 missing values.

5. Cleaning Column Names

- Function used: df.columns.str.title()
- Changed column names so that the first letter of each word is capitalized (e.g., show_id → Show_Id).
- Makes the dataset more readable and presentation-friendly.

6. Cleaning Text Columns

- Function used: str.strip()
- Removed leading/trailing spaces in text-based columns like Title, Director, Cast, Country, Description, Listed_in.

7.Key Insights & Next Steps

- Data Quality Check: No duplicate rows, but several missing values were found and fixed.
- Standardization: Column names and text values were cleaned for consistency.
- Now the dataset is prepared for deeper analysis.