Internship Project Documentation

Project Title: *DV: StreamScope – Netflix Content Strategy Analyzer*

Intern Name: AR Darshan

Internship Duration: 8 Weeks

Institution: Sri Eshwar College of Engineering

Mentor : Swaraj Alla

---

1. Project Overview

1.1 Introduction

In today's digital era, data drives entertainment decisions. Streaming platforms like Netflix rely heavily on data analytics to understand viewer preferences, optimize content, and plan global expansions.

This project, titled DV: StreamScope - Netflix Content Strategy Analyzer, focuses on analyzing Netflix's catalog to uncover hidden patterns, such as:

- The growth of content over time
- Genre and rating trends
- Regional content variations
- Audience-oriented content preferences

The aim is to gain insights from data -*not by guessing*, but through data-driven storytelling and analysis.

1.2 Objective

To build a data analytics system that processes and visualizes Netflix's catalog data to understand:

- What kind of content Netflix produces most
- Which genres and countries dominate
- How its content has evolved through years
- What patterns exist between content type, rating, and release

---

1.3 Expected Outcomes

- A cleaned and well-prepared dataset for analysis
- Insightful exploratory data analysis (EDA) with visual trends
- Grouping and classification of content using analytics methods
- An interactive dashboard showcasing Netflix's content strategy

---

2. Dataset Understanding

2.1 Source of Data

The dataset was obtained from Kaggle — titled *"Netflix Movies and TV Shows Dataset."*
It contains over 8,000 titles available on Netflix, including details about movies and TV shows from different countries and time periods.

2.2 Attributes in Dataset

Each record in the dataset provides rich metadata about a title:

- show_id: Unique identifier
- type: Movie or TV Show
- title: Name of the content
- director: Director of the title
- cast: Main actors
- country: Country of origin
- date_added: When it was added to Netflix
- release_year: When it was originally released
- rating: Age rating (e.g., TV-MA, PG, R)
- duration: Movie length or number of seasons
- listed_in: Genres or categories
- description: Short summary of the title

2.3 Dataset Insights

- The dataset gives both content-level and metadata-level information.
- It provides a global perspective, not limited to one country or language.

- It’s an excellent dataset for learning data cleaning, analysis, and visualization in the context of media and entertainment analytics.

---

3. Data Preprocessing and Cleaning

Before analyzing, the dataset was cleaned to ensure accuracy and consistency.

3.1 Handling Duplicates

Duplicate records were removed to maintain uniqueness.

3.2 Managing Missing Values

Several columns like *director*, *cast*, and *country* contained missing data.

- Replaced missing director and cast names with “Unknown”.
- For missing *ratings*, replaced with “Not Rated”.
- Checked and ensured there were no null values left.

3.3 Data Consistency

- Standardized categorical values (e.g., “TV Show” vs “TV show”).
- Ensured all year and duration fields were in uniform format.
- Saved the cleaned dataset for further analysis.

Result:
A clean, structured, and complete dataset ready for deep exploration.

---

4. Exploratory Data Analysis (EDA)

4.1 Understanding Content Composition

Analysis began by examining the types of content available.
Netflix has a clear focus on Movies, making up about 70% of titles, while TV Shows form the remaining 30%.
This indicates Netflix’s balance between short-form entertainment and long-running series.

4.2 Growth Over the Years

By observing the *release_year* column, it was clear that:

- Netflix’s content library grew steadily until 2015.

- Between 2016–2020, the growth accelerated rapidly, aligning with Netflix's expansion into new countries and the rise of Netflix Originals.

This shows Netflix's strategic scaling period and investment in global content.

4.3 Genre and Category Analysis

The listed_in column revealed which genres dominate the platform:

- *Dramas*, *Comedies*, and *Action & Adventure* are the most popular.
- Documentary content and thrillers also showed steady growth.

This suggests that Netflix targets a wide audience — from casual comedy lovers to serious drama viewers.

4.4 Regional Insights

The country column analysis showed that:

- The United States leads in content production.
- India and United Kingdom are also major contributors.
- A notable rise in non-English titles reflects Netflix's localization strategy.

This reflects how Netflix adapts its library to regional interests.

4.5 Ratings and Audience Targeting

The rating field indicates the audience maturity level of each title.
Most content is rated *TV-MA* and *TV-14*, meaning Netflix's majority of titles cater to adult and teen audiences, respectively.

4.6 Duration and Format

Movies were analyzed based on their length (Short / Medium / Long), while TV shows were grouped by number of seasons.
This helped differentiate between quick entertainment and long-term series engagement strategies.

---

5. Feature Engineering & Advanced Insights

Feature engineering was performed to create new perspectives from existing data:

- Duration Category: Classified as *Short (<60 mins)*, *Medium (60–120 mins)*, and *Long (>120 mins)*.
- Year Grouping: Grouped into decades (e.g., 1990s, 2000s, 2010s).
- Original vs Licensed: Differentiated titles that are Netflix Originals vs those acquired from other studios.

These derived features helped make the dataset richer and ready for advanced analysis.