# **Netflix Data: Cleaning, Analysis, and Visualization**

## **1. Introduction**

### **1.1 Overview**

Netflix has evolved into one of the world's leading entertainment platforms, offering an extensive library of movies, TV shows, documentaries, and original content. Understanding how Netflix's content catalog is structured, which genres dominate, and how content has evolved over time can provide valuable insights for content strategy, marketing, and user engagement.

This project focuses on cleaning, analyzing, and visualizing the Netflix content dataset. By applying data cleaning techniques and exploratory data analysis (EDA), we aim to discover patterns and trends that shape the Netflix content landscape.

### **1.2 Problem Statement**

The Netflix dataset contains information on thousands of titles, including movies and TV shows, available on the platform. The objective of this project is to:

* Clean the dataset by handling missing values, duplicates, and incorrect data types.
* Perform exploratory data analysis to extract insights about content distribution, genre popularity, top contributors (directors, actors, etc.), and content trends over time.
* Visualize the findings to better understand the content offering on Netflix.

### **1.3 Importance of Data Analysis in Streaming Platforms**

Analyzing data from streaming platforms like Netflix is crucial for decision-making in areas such as:

* **Content Strategy**: Understanding which genres and types of content perform best can inform future content acquisition and creation.
* **User Engagement**: Analysis of viewership patterns can help tailor recommendations to users, increasing engagement.
* **Content Investment**: Identifying trends over time helps in determining where to invest resources for new content.

## **2. Tools and Technologies**

### **2.1 Programming Language**

* **Python**: Python was chosen due to its robust ecosystem of libraries for data analysis, visualization, and machine learning.

### **2.2 Libraries Used**

* **pandas**: To handle data manipulation and cleaning tasks, including missing value handling and feature extraction.
* **numpy**: For numerical computations and data handling.
* **matplotlib** & **seaborn**: For data visualization to help uncover patterns and trends.
* **wordcloud**: Used to generate visual representations of frequently occurring words, such as popular movie titles and genres.
* **PostgreSQL**: For advanced data cleaning in cases where large datasets need to be stored and queried efficiently.

### **2.3 Tools**

* **Jupyter Notebook**: For developing, testing, and iterating through code efficiently.
* **Visual Studio Code**: As a versatile and user-friendly Integrated Development Environment (IDE) for Python development.
* **Tableau**: Used for creating interactive visual dashboards for business stakeholders to explore insights independently.

## **3. Dataset Overview**

### **3.1 Data Source**

* **Netflix Titles Dataset**: The dataset includes all Netflix content available from 2008 through 2021, with metadata such as title, type, director, cast, country, date added, release year, rating, duration, and genres.

### **3.2 Columns Overview**

* **show\_id**: Unique identifier for each title.
* **type**: Specifies whether the title is a Movie or TV Show.
* **title**: The title of the movie or show.
* **director**: Name of the director of the content.
* **cast**: List of lead actors in the title.
* **country**: The country where the content was produced.
* **date\_added**: Date when the title was added to Netflix.
* **release\_year**: The year the title was released.
* **rating**: Content rating (e.g., PG, TV-MA).
* **duration**: Duration of the movie (in minutes) or TV show (number of seasons).
* **listed\_in**: Categories/genres that the title falls under.

### **3.3 Initial Data Review**

Before we start analyzing the dataset, it’s important to review its structure, check for missing values, and assess data quality.

python

import pandas as pd

# Load the Netflix dataset

data = pd.read\_csv('netflix\_titles.csv')

# Display first few rows

print(data.head())

# Show basic info about the dataset

print(data.info())

# Check for missing values in each column

print(data.isnull().sum())

## **4. Data Cleaning**

### **4.1 Handling Missing Values**

Many columns in the dataset, such as `director`, `cast`, and `country`, contain missing values. These fields are critical for analysis, especially when looking at content creators and country-level content distribution. To address this:

* **Drop rows** where critical information like `director` or `country` is missing.
* **Fill missing values** where appropriate for other columns, such as `rating` or `date\_added`.

python

# Drop rows with missing critical information like director, cast, and country

data\_cleaned = data.dropna(subset=['director', 'cast', 'country'])

# Fill missing 'rating' values with a default value

data\_cleaned['rating'].fillna('Not Rated', inplace=True)

# Fill missing 'date\_added' values with the most frequent date

most\_frequent\_date = data\_cleaned['date\_added'].mode()[0]

data\_cleaned['date\_added'].fillna(most\_frequent\_date, inplace=True)

### **4.2 Duplicates and Data Type Correction**

We removed duplicate entries to ensure each title is unique. Furthermore, the `date\_added` column was converted from string to a **datetime** format for accurate analysis over time.

python

# Remove duplicates based on the show\_id column

data\_cleaned.drop\_duplicates(subset=['show\_id'], inplace=True)

# Convert 'date\_added' column to datetime format

data\_cleaned['date\_added'] = pd.to\_datetime(data\_cleaned['date\_added'])

### **4.3 Feature Engineering**

To enhance the dataset, we created new features:

* **Year Added**: Extracted from the `date\_added` column to analyze trends over time.
* **Duration in Minutes**: Parsed the `duration` column for movies, and split it for TV shows (season count).

python

# Extract year from 'date\_added'

data\_cleaned['year\_added'] = data\_cleaned['date\_added'].dt.year

# Split 'duration' into numeric values for Movies (minutes) and TV Shows (seasons)

data\_cleaned['duration'] = data\_cleaned['duration'].apply(lambda x: int(x.split()[0]))

## **5. Exploratory Data Analysis (EDA)**

### **5.1 Distribution of Content Type**

Analyzing the breakdown between Movies and TV Shows reveals how Netflix's library is weighted toward different content types.

python

import matplotlib.pyplot as plt

import seaborn as sns

# Count the number of Movies and TV Shows

type\_counts = data\_cleaned['type'].value\_counts()

# Plot the content type distribution

plt.figure(figsize=(8,6))

sns.barplot(x=type\_counts.index, y=type\_counts.values, palette='Set2')

plt.title('Distribution of Netflix Content: Movies vs. TV Shows')

plt.xlabel('Content Type')

plt.ylabel('Count')

plt.show()

### **5.2 Content Added Over Time**

This analysis tracks the growth of Netflix’s library by year, showing how the platform's content offering has increased.

python

# Count the number of titles added per year

plt.figure(figsize=(12,6))

sns.countplot(x='year\_added', data=data\_cleaned, palette='coolwarm')

plt.title('Content Added to Netflix Over Time')

plt.xlabel('Year')

plt.ylabel('Number of Titles Added')

plt.xticks(rotation=45)

plt.show()

### **5.3 Top Genres on Netflix**

By analyzing the genres, we can understand Netflix's focus areas and which genres dominate their content offering.

python

# Split genres and count occurrences

data\_cleaned['genres'] = data\_cleaned['listed\_in'].apply(lambda x: x.split(', '))

all\_genres = sum(data\_cleaned['genres'], [])

# Count the most common genres

genre\_counts = pd.Series(all\_genres).value\_counts().head(10)

# Plot the most popular genres

plt.figure(figsize=(10, 6))

sns.barplot(x=genre\_counts.values, y=genre\_counts.index, palette='Set3')

plt.title('Top 10 Genres on Netflix')

plt.xlabel('Count')

plt.ylabel('Genre')

plt.show()

### **5.4 Top Directors and Contributors**

The analysis identifies the directors with the highest number of titles on Netflix.

python

# Count titles by director

top\_directors = data\_cleaned['director'].value\_counts().head(10)

# Plot the top directors with the most titles

plt.figure(figsize=(10, 6))

sns.barplot(x=top\_directors.values, y=top\_directors.index, palette='Blues\_d')

plt.title('Top 10 Directors on Netflix')

plt.xlabel('Number of Titles')

plt.ylabel('Director')

plt.show()

### **5.5 Word Cloud of Popular Titles**

Using a word cloud, we visualize the most frequent words in Netflix titles to highlight trends in naming and focus areas.

python

from wordcloud import WordCloud

# Generate word cloud for titles

movie\_titles = data\_cleaned[data\_cleaned['type'] == 'Movie']['title']

wordcloud = WordCloud(width=800, height=400, background\_color='black').generate(' '.join(movie\_titles))

# Display the word cloud

plt.figure(figsize=(10, 6))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis('off')

plt.show()

## **6. Data Visualization Using Tableau**

Using Tableau, we created interactive visual dashboards to:

* Explore the content distribution across different countries.
* Analyze the trends in content acquisition over time.
* Create visualizations that stakeholders can use to track the evolution of Netflix's library.

Key visualizations include:

* **Year-wise addition of content** to Netflix, visualized as an interactive line chart.
* **Top genres and content types**, broken down by country and year.
* **Geo-distribution of content**, where viewers can filter by country and see the number of titles Netflix has available.

## **7. Conclusion and Insights**

### **7.1 Key Findings**

* **Content Type**: Movies dominate Netflix's content library, accounting for approximately 70% of the total titles. TV Shows make up the remaining 30%.
* **Content Growth**: The number of titles added annually has seen exponential growth, especially after 2015, reflecting Netflix's aggressive content acquisition strategy.
* **Popular Genres**: Dramas, documentaries, and stand-up comedies are among the most popular genres on Netflix.
* **Directors and Contributors**: A small group of directors are responsible for a significant portion of the content, indicating a high concentration of creators producing Netflix content.
* **Content Distribution**: The United States and India lead the world in producing content available on Netflix, contributing a significant portion of the titles in its library.

### **7.2 Conclusion**

Through cleaning, analysis, and visualization of the Netflix dataset, we gained important insights into the platform's content strategy. Our analysis highlighted the dominant genres, the growing trend of TV shows, and key contributors shaping Netflix’s content.

### **7.3 Future Work**

* **Advanced Analysis**: Predictive models can be built to forecast content trends, such as predicting which genres will become popular in the coming years.
* **Content Recommendations**: By applying machine learning algorithms, we could develop a recommendation system based on user behavior and content preferences.
* **User Engagement Analysis**: Linking this dataset with user viewership data could help analyze which types of content drive the most engagement.