# Task-2 Unified-Mentor-Netflix

Netflix Data Analysis Project Report (Task 2)

## Introduction

This report summarizes my work on the "Netflix Data: Cleaning, Analysis, and Visualization" project for Task 3 with Unified Mentors Pvt. Ltd. The goal was to clean and analyze a Netflix dataset containing movies and TV shows from 2008 to 2021, then visualize insights to understand trends. I used Google Colab with Python to process the data and create visualizations, testing my skills in data cleaning, analysis, and visualization.

## Dataset Overview

The dataset includes details about Netflix content with the following columns:

show_id: Unique identifier for each title

type: Movie or TV Show

title: Name of the content

director: Director of the content

country: Country of origin

date_added: Date added to Netflix

release_year: Year of release

rating: Content rating (e.g., PG-13, TV-MA)

duration: Length of content (e.g., minutes or seasons)

listed_in: Genres of the content

**The dataset had 8,790 rows with no missing values or duplicates, making it ready for analysis after minor cleaning.**

**Steps Performed**

1. Data Loading and Initial Checks

* Loaded Data: I loaded the Netflix dataset from a CSV file using Python in Google Colab.

* Checked Data: I examined the dataset's structure (8,790 rows, 10 columns), data types, and confirmed no missing values or duplicates using df.info() and df.duplicated().sum().

2. Data Cleaning

* I cleaned the dataset to prepare it for analysis:

* Converted Data Types:

* Changed date_added to datetime format to enable date-based analysis (e.g., extracting years and months).

* Confirmed release_year as an integer for numerical analysis.

**Processed Genres:**

- Created a genres column by splitting listed_in (e.g., "Drama, Comedy" became ["Drama", "Comedy"]).
- Added genre_count (number of genres per title) and primary_genre (first genre listed) for additional insights.
- Exploded the genres column to create a new DataFrame (df_exploded) where each genre is a separate row, enabling accurate genre counting.

**Extracted Date Features:**

- Created year_added and month_added from date_added to analyze content addition trends.

### 3. Exploratory Data Analysis (EDA)

- I analyzed the dataset to uncover trends using Python libraries (pandas, matplotlib, seaborn, wordcloud):
- Total Content: Counted all titles to show the dataset's size.
- Rating Distribution: Analyzed the proportion and count of ratings (e.g., TV-MA, PG-13).
- Top Countries: Identified the top 10 countries contributing content.
- Monthly Releases: Examined monthly release patterns for Movies and TV Shows.
- Yearly Releases: Analyzed content addition trends over years.
- Top Genres for Movies: Found the top 10 genres for Movies.
- Top Genres for TV Shows: Found the top 10 genres for TV Shows.

Word Cloud: Visualized common words in titles.

### 4. Visualizations

I created clear visualizations to present findings:

Pie Chart (Rating Distribution): Showed the percentage of each rating, highlighting dominant ratings.

Bar Chart (Rating Distribution): Displayed the count of titles per rating for comparison.

Bar Chart (Top 10 Countries): Showed countries with the most content (e.g., United States).

Stacked Bar Chart (Monthly Releases): Compared Movie and TV Show releases by month, showing seasonal trends.

Line Plot (Yearly Releases): Plotted content additions over years, highlighting peak years.

Bar Chart (Top Movie Genres): Showed the most popular genres for Movies (e.g., Dramas, Comedies).

Bar Chart (Top TV Show Genres): Showed the most popular genres for TV Shows (e.g., TV Dramas, Reality TV).

Word Cloud: Visualized frequent words in titles, indicating popular themes.

**5. Tools Used**

* Google Colab: Platform for running Python code and generating visualizations.

Python Libraries:

- pandas: For data cleaning and analysis.
- matplotlib and seaborn: For creating charts (pie, bar, line).
- wordcloud: For generating the word cloud of titles.

**Key Findings**

Total Content: The dataset contains 8,790 titles (Movies and TV Shows).

Ratings: TV-MA and PG-13 are among the most common ratings, indicating content for mature and teen audiences.

Countries: The United States contributes the most content, followed by countries like India and the UK.

Monthly Releases: Releases vary by month, with some months (e.g., December) showing higher activity.

Yearly Releases: Content additions peaked around 2019–2020, showing Netflix's growth.

Movie Genres: Popular genres include International Movies, Dramas, and Comedies.

TV Show Genres: Popular genres include TV Dramas, Reality TV, and Kids' TV.

Word Cloud: Common title words like "Love," "Life," and "Story" suggest popular themes.

**Conclusion**

This project helped me build essential data analysis skills. I successfully:

Loaded and cleaned the Netflix dataset by converting data types, processing genres, and extracting date features.

Analyzed trends in content types, ratings, countries, releases, and genres.

Created visualizations (pie charts, bar charts, line plots, word cloud) to present insights clearly.

**Next Steps**

Feature Engineering: Extract numerical duration (e.g., minutes for Movies, seasons for TV Shows) for further analysis.

Machine Learning: Build a recommendation system using genres or ratings.

Advanced Visualization: Create an interactive Tableau dashboard to explore trends.

**Challenges and Learnings**

Challenge: Splitting and exploding the listed_in column to analyze genres accurately.

Learning: Gained hands-on experience with Python for data cleaning, analysis, and visualization in Google Colab.

Skill Development: Improved my ability to process data, create meaningful visualizations, and interpret trends.

This project was a valuable exercise that prepared me for advanced data analysis tasks.