Netflix Dataset Analysis Report

1. Introduction

This report summarizes the exploratory data analysis (EDA) performed on the Netflix dataset (netflix1.csv). The dataset contains metadata about Movies and TV Shows available on Netflix, including their type, title, director, country, release year, rating, duration, and listed genres.

2. Dataset Overview

• Total Records (Rows): 8,790
• Columns: 10
• Size: ~87,900 data points
• Memory Usage: ~687 KB

Columns Description

show_id – Unique identifier
type – Movie or TV Show
title – Name of the show/movie
director – Director of the content
country – Country of origin
date_added – Date added on Netflix
release_year – Release year
rating – Audience rating (e.g., TV-MA, PG-13)
duration – Duration (minutes for movies / seasons for shows)
listed_in – Genres

3. Exploratory Analysis

Content Type Distribution: Movies outnumber TV Shows on Netflix.
Country-wise Distribution: United States, India, and UK dominate Netflix’s catalog.
Release Year Distribution: Majority of content is from 2015–2021.
Genre Analysis: Top genres include International Movies, Dramas, Comedies, and Documentaries.
Ratings Distribution: TV-MA and TV-14 dominate, highlighting focus on adult audiences.
Movie Duration Analysis: Most movies are between 90–120 minutes.

4. Data Preprocessing

• Cleaned show_id by removing prefix 's'.
• Applied Label Encoding on categorical columns for ML readiness.

5. Key Insights

• US and India dominate Netflix’s content library.
• Movies are more common than TV Shows.
• Netflix strongly focuses on adult-oriented content (TV-MA, TV-14).
• Most content is recent, aligning with Netflix’s rapid global expansion.
• Dramas, comedies, and international films form the backbone of Netflix’s catalog.
• Standard feature film length (~100 mins) is most common for movies.

6. Conclusion

Netflix offers a diverse catalog dominated by movies, mature ratings, and recent releases. The platform emphasizes global content, with the US, India, and UK being major contributors.

Future Directions

• Time-series analysis of content additions.
• Deeper genre clustering for audience segmentation.
• Predictive modeling on popularity vs. content features.