Netflix Project — Final Report

1) Basic Details
- Name: Vaibhav Parmar
- Email: vaibhavparmar.180708@gmail.com
- Institution: Atmiya University
- Project Title: Netflix Data — Cleaning, Analysis & Visualization
- Tech Stack: Python, Pandas, Matplotlib

2) Problem Statement
Briefly explain what you wanted to do with the Netflix dataset (clean it, explore trends, and make simple charts).

3) Dataset
- Source: Netflix titles dataset
- Size: 8790 (Total rows after cleaning)
- Time coverage: 2008–2021

4) What I Did
- Loaded the data and checked its shape and columns.
- Removed duplicates and trimmed whitespaces.
- Converted date_added to date and created year_added & month_added.
- Extracted duration_minutes for movies and seasons for TV shows.
- Split listed_in into a list of genres for counting.
- Made simple charts: content type, ratings, top countries, top genres, top directors, top actors, and yearly trends.

5) Insights
After exploring the Netflix dataset, I found that Movies are more common than TV Shows on the platform. This means Netflix still releases more films overall compared to series, though both have a strong presence.

In terms of content ratings, I noticed that TV-MA and TV-14 appear the most. These ratings usually represent content meant for mature audiences or teenagers, which suggests Netflix has a lot of content aimed at older viewers rather than small kids.

When checking which countries contribute the most titles, the United States is clearly at the top. Following that, countries like India, United Kingdom, and Canada also produce a large amount of Netflix content, showing that Netflix's library is quite global.

The most common genres I found were Documentaries, Dramas, and Comedies. These genres appear very frequently, which makes sense because they appeal to a wide audience and Netflix invests heavily in them.

Looking at yearly trends, the number of titles on Netflix has grown a lot over the years, especially after 2015. This shows that Netflix has expanded its content library rapidly in recent years to grow its user base.

Overall, Netflix seems to focus heavily on mature content, and most of its library consists of movies, with strong representation from countries like the US and India.

6) Challenges
While working on this project, I faced a few small challenges.
First, the dataset had many missing values in important columns like the director, cast, and country fields. I handled this by replacing them with "Unknown" so that the analysis could continue smoothly.
Another challenge was that the duration column had mixed formats — sometimes in minutes and sometimes in seasons. I needed to separate these into two new columns to make the data easier to

understand.
Also, the country and genre fields sometimes contained multiple values in one cell, so I had to clean and split them properly to count them correctly.
Apart from these points, the rest of the process — loading, cleaning, and visualizing the data — went smoothly.

7) Conclusion
In this project, I cleaned and analyzed the Netflix dataset to understand how the platform's content is distributed. I found that movies make up most of the titles on Netflix, and ratings like TV-MA and TV-14 are the most frequent. The United States produces the most content, followed by countries like India and the UK.
Popular genres include documentaries, dramas, and comedies, showing that Netflix focuses on diverse and widely appealing categories.
Overall, the platform has grown a lot over the years, adding more titles especially after 2015. This project helped me understand how to work with real-world data, clean it properly, and extract simple insights through visualizations.

8) How to Run
- python netflix_project.py
- Outputs saved inside outputs/ folder.

9) Project Links
- GitHub repo link: https://github.com/vaibhavparmar96/Internship-projects