Project Title: Netflix Content Analysis using Data Analytics
Name: Vignesh A
Batch: 10 Feb 2025 to 10 May 2025
UNID: UMIP277374

Abstract

This project provides a comprehensive analysis of Netflix's content catalog to uncover insights into distribution trends, geographic presence, popular genres, and production patterns. Through data cleaning, exploratory data analysis (EDA), and visualization techniques, the project reveals viewer preferences and platform strategies, helping understand Netflix's global expansion and content strategy.

1. Introduction

Netflix, a global leader in streaming entertainment, offers thousands of titles spanning various genres and countries. Analyzing its content data can help understand platform focus areas, production volume, and market reach. This project aims to:

- Study content type distribution
- Identify top contributing countries, directors, and genres
- Explore release year and platform trends
- Analyze descriptions, genres, and cast using visualizations like WordClouds

2. Tools and Technologies Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, WordCloud
- **Visualization Tools:** Matplotlib, Seaborn
- **Platform:** Jupyter Notebook

3. Dataset Description

- **Source:** Netflix title dataset (`netflix1.csv`)
- **Columns:**
  - `title` – Name of the show/movie
  - `director`, `cast`, `country` – People and location details
  - `date_added`, `release_year` – Temporal info
  - `rating`, `duration` – Censorship and content length
  - `listed_in` – Genres
  - `description` – Summary of the content

Data Cleaning Performed:

- Dropped duplicate entries
- Handled missing values in `director` and `country`
- Parsed `date_added` and created `year`, `month`, `day` columns

---

4. Exploratory Data Analysis (EDA)

4.1 Content Type Distribution

- Bar plot reveals that **Movies** dominate the catalog over **TV Shows**.

4.2 Country-Wise Content

- The **United States**, **India**, and **United Kingdom** are the top contributors.
- A bar graph was used to visualize country-wise content volume.

4.3 Top Directors and Cast (WordClouds)

- WordClouds identify frequent directors like **Raúl Campos**, **Marcus Raboy**.
- Top actors include **Anupam Kher**, **David Attenborough**, and **Noah Centineo**.

4.4 Popular Genres

- WordCloud from the `listed_in` field shows that **Dramas**, **Comedies**, and **Documentaries** are highly represented.

4.5 Release Trend Over Years

- Bar chart displays a rise in content from **2015 to 2019**, peaking right before the pandemic.
- Few titles were added during early years (pre-2010), indicating platform growth post-2015.

---

5. Key Insights

- **Movies are more prevalent** than TV Shows on Netflix.
- The **U.S. and India** are key content producers.
- Directors and actors with recurring appearances hint at preferred creators.
- There’s been a **content boom post-2015**, possibly due to Netflix’s global expansion.
- Genres like **Drama and Comedy** dominate viewership.

---

6. Limitations and Scope

- Dataset does not include viewership or user ratings.
- Some fields like `cast` and `director` had missing values.
- Future studies could integrate watch-time or engagement metrics for deeper insights.

---

7. Conclusion

The analysis reveals how Netflix's content strategy has evolved over the years, focusing heavily on movies and expanding into various global markets. Insights from genre popularity and creator frequency can assist in understanding how Netflix caters to its audience.

---

8. Future Scope

- Integrate Netflix viewership data for popularity-based analysis
- Apply clustering to group similar content types
- Build a recommendation engine using content metadata
- Deploy findings in an interactive dashboard using Power BI or Streamlit

---

9. Project Link

https://github.com/vignesh-a-09/Unified-Mentor-Internship-2025/blob/main/Netflix%20data%20analysis/Netflixdata.ipynb