**Netflix Data: Cleaning, Analysis & Visualization**

**Tools Used:** Python (Pandas, Seaborn, Matplotlib, WordCloud), Excel

**Domain:** Data Analytics & Data Science

**Difficulty Level:** Intermediate

**1. Project Overview**

Netflix is one of the world’s largest streaming platforms, offering thousands of movies, TV shows, and original content.

**The aim of this project is to:**

Perform data cleaning to prepare a reliable dataset.

Conduct exploratory data analysis (EDA) to uncover patterns and insights.

Create visualizations for a clear understanding of Netflix’s content library trends.

**2. Dataset Description**

* Total Records: 8,790
* Columns: 10
* show\_id: Unique identifier for each title
* type: Movie or TV Show
* title: Title of the content
* director: Director(s) of the content
* country: Country of production
* date\_added: Date when content was added to Netflix
* release\_year: Year of original release
* rating: Audience rating (e.g., PG, TV-MA)
* duration: Duration in minutes or number of seasons
* listed\_in: Genres/categories

The dataset covers content added from 2008 to 2021, with some movies as old as 1925.

**3. Data Cleaning Steps**

• Removed duplicate records to ensure unique entries.  
• Converted 'date\_added' to datetime format and extracted year and month for trend analysis.  
• Filled missing values in columns like director, country, and cast with 'Unknown'.  
• Split the 'listed\_in' column to derive multiple genres for each title.  
• Extracted numerical values from 'duration' to separate duration and type (minutes/seasons).

**4. Exploratory Data Analysis (EDA)**

**4.1 Content Type Distribution**

Netflix has more Movies compared to TV Shows.

**4.2 Most Common Genres**

Top genres include International Movies, Dramas, Comedies, and Documentaries.

**4.3 Content Added Over Time**

There is a sharp increase in content addition after 2015, coinciding with Netflix’s global expansion and original productions.

**4.4 Top 10 Directors**

Raúl Campos and Jan Suter lead with the most titles on Netflix.

**4.5 Word Cloud of Movie Titles**

Common title keywords suggest trends in naming (e.g., “Love”, “World”, “Life”).

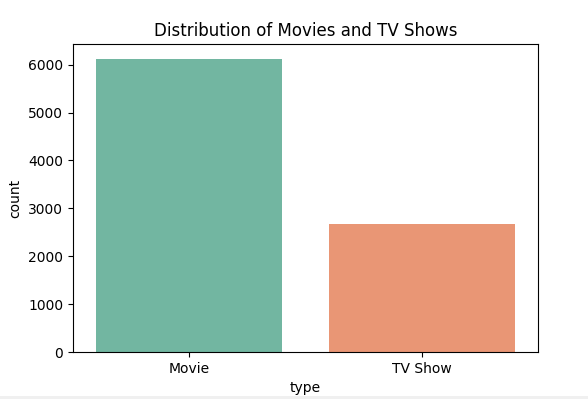
**5. Insights Derived**

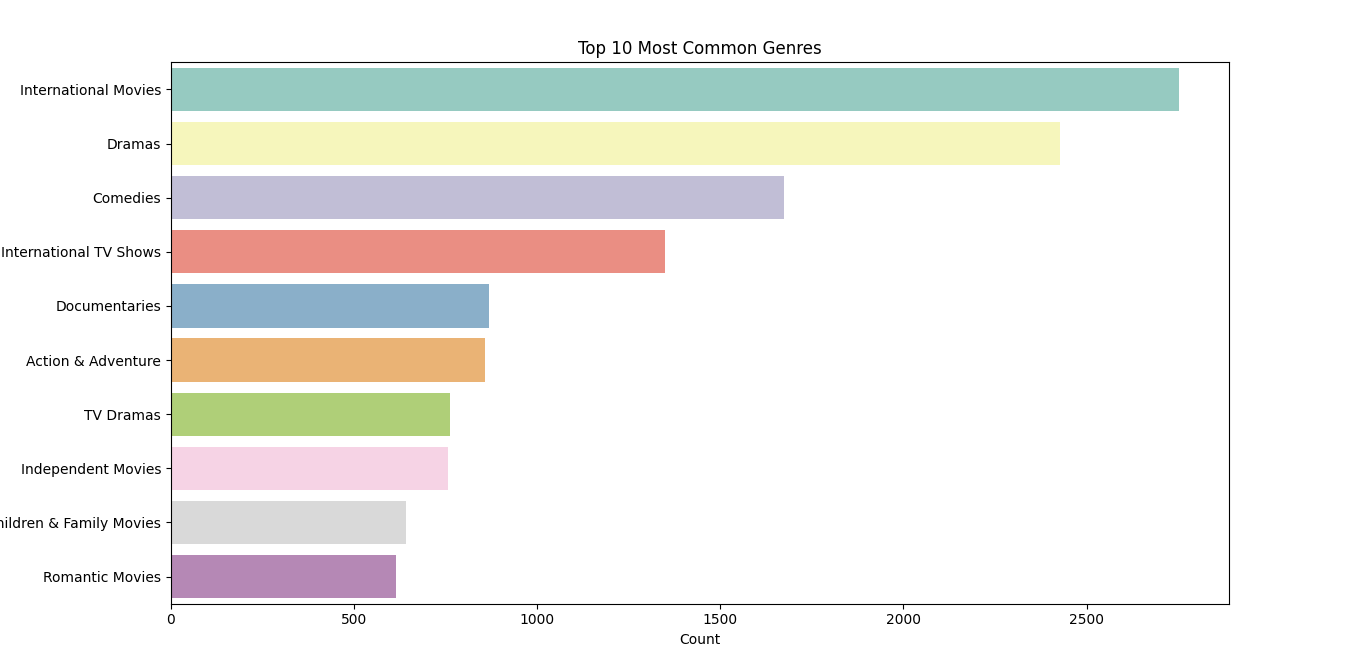
* Movies dominate Netflix’s library, showing its preference for feature films.
* Genre diversity is a key strength, with strong representation of international content.
* Post-2015 expansion significantly increased Netflix’s content base.
* Directors with frequent collaborations have higher title counts, often linked to original content.

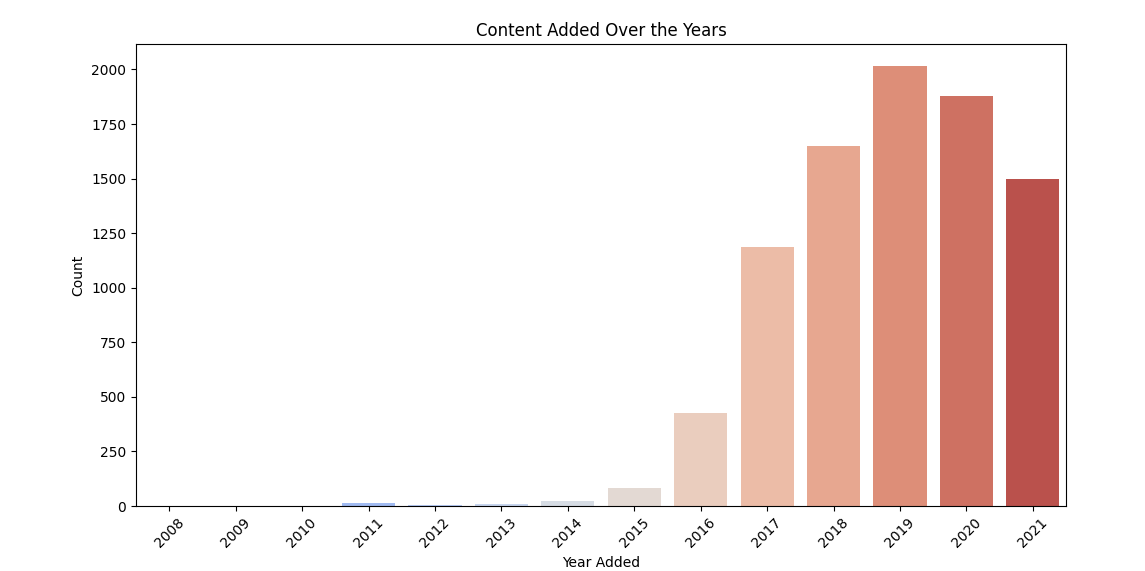
**6. Recommendations & Next Steps**

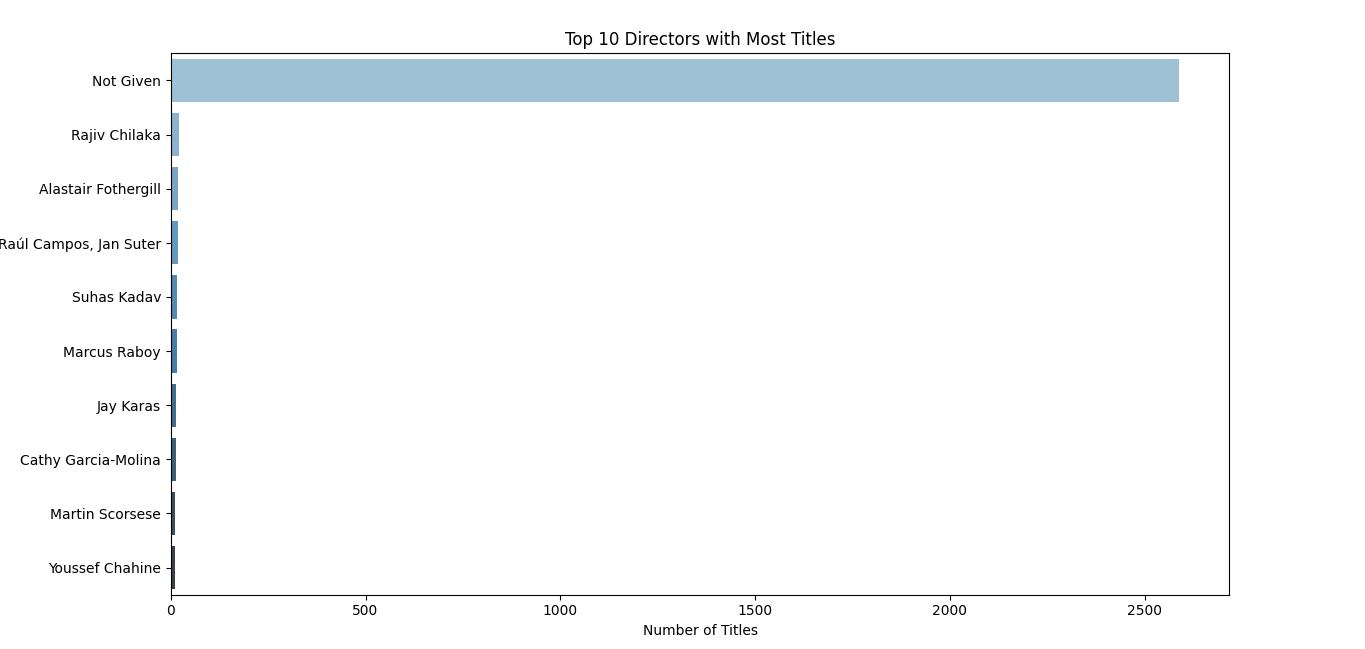
* Feature Engineering: Add derived features like number of genres per title and numeric rating mapping.
* Machine Learning: Build recommendation models based on genre similarity.
* Visualization: Develop an interactive dashboard in Tableau or Power BI for dynamic analysis.
* Data Enrichment: Combine with external datasets (e.g., IMDb ratings) for deeper insights.

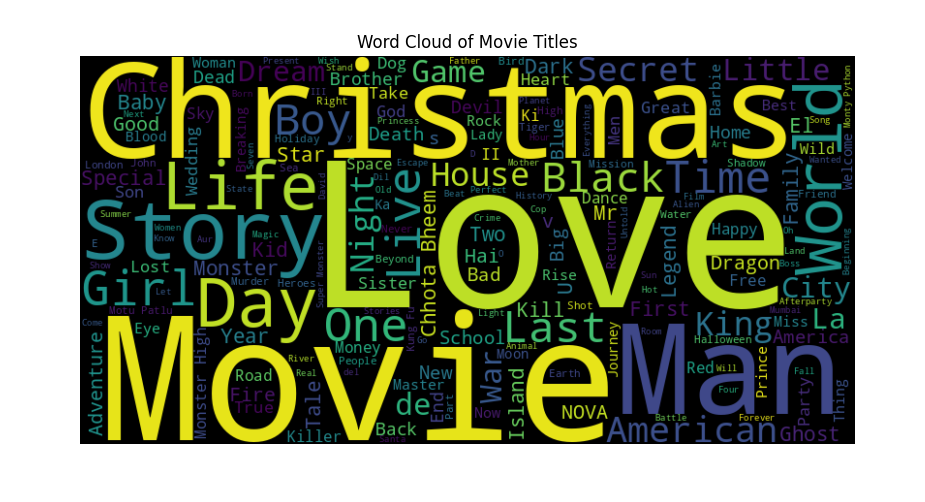
1. **Output Files**











**8. Conclusion**

* This project demonstrated an end-to-end approach to data cleaning, exploration, and visualization using Python.
* The findings revealed content trends, popular genres, and key contributors in Netflix’s catalog.
* These insights can support content strategy, personalization, and future predictive analytics.