1. Project Overview

The purpose of this project is to analyze Netflix's content library to extract meaningful insights about its movies and TV shows. This includes understanding content distribution by type, genre, country, release year, ratings, and duration. The project demonstrates data cleaning, feature engineering, exploratory data analysis (EDA), visualization, SQL querying, and business insights.

2. Dataset Description

The dataset consists of 8807 titles available on Netflix, with 12 columns.

Columns:

- show_id: Unique identifier

- type: Movie or TV Show

- title: Title name

- director: Director(s)

- cast: Main cast

- country: Producing country

- date_added: Date added to Netflix

- release_year: Year of original release

- rating: Viewer rating category

- duration: Runtime or seasons

- listed_in: Genres

- description: Summary text

3. Exploratory Data Analysis using Python

We began the analysis by preparing, cleaning, and transforming the Netflix dataset in Python. The steps included:

- Data Loading

Imported the dataset using pandas.

- Initial Exploration

Performed preliminary analysis using:

- df.info() — to understand column data types and non-null counts
- df.head() — to preview the first few rows
- df.isnull().sum() — to identify missing values
- df.describe() — for summary statistics (numeric fields)

- Missing Data Handling

Checked the dataset for missing values across all columns.

Applied cleaning steps:

- Filled missing text-based fields (director, cast, country, rating) with "Unknown".
- Replaced missing date_added with a placeholder date "2000-01-01".
- Filled missing duration values with "Unknown".

After imputation, verified that there were **no remaining null values** in the essential fields.

- Feature Engineering

Created multiple new columns to support richer analysis:

- **year_added**
  Extracted the year from date_added.
- **month_added**
  Converted the date into a human-readable month name.
- **movie_minutes**
  Parsed total minutes for movies using regular expressions (e.g., “90 min → 90”).
- **seasons**
  Extracted the number of seasons for TV shows (e.g., “3 Seasons → 3”).

These engineered features enabled granular trend analysis such as content growth per year and duration-based insights.

- Date Standardization

Converted date_added into proper datetime format using pd.to_datetime() to ensure consistent time-based analysis.

- Genre Normalization

Processed the listed_in column by:

- Splitting comma-separated genres
- Removing extra whitespace
- Exploding the list to count individual genre frequencies

This helped identify the **top-performing and most common content categories** on Netflix.

- Data Consistency Checks

Performed integrity checks such as:

- Validating extracted durations
- Ensuring movie titles and types aligned with newly engineered fields
- Verifying no invalid date formats remained after parsing

- Database Integration

Connected Python to a **MySQL database** using SQLAlchemy and PyMySQL.

Steps included:

- Creating a connection engine
- Exporting the cleaned DataFrame into a MySQL table named **netflix**
- Running SQL queries for deeper analytical insights

4. Data Analysis using SQL

1. Total number of Movies and TV Shows available on Netflix

type	total_titles
Movie	6131
TV Show	2676

2. Top 10 countries contributing the highest amount of Netflix content.

country	total_titles
United States	2818
India	972
United Kingdom	419
Japan	245
South Korea	199
Canada	181
Spain	145
France	124
Mexico	110
Egypt	106

3. Most common genres available on the Netflix platform.

genre	occurrences
International Movies	2752
Dramas	2427
Comedies	1674
International TV Shows	1351
Documentaries	869
Action & Adventure	859
TV Dramas	763
Independent Movies	756
Children & Family Movies	641
Romantic Movies	616

4. Year-wise trend of titles added to Netflix.

year	total_titles
2000	98
2008	2
2009	2
2010	1
2011	13
2012	3
2013	10
2014	23
2015	73
2016	418
2017	1164
2018	1625
2019	1999
2020	1878
2021	1498

5. Distribution of titles across Netflix rating categories.

rating	total_titles
TV-MA	3207
TV-14	2160
TV-PG	863
R	799
PG-13	490
TV-Y7	334
TV-Y	307
PG	287
TV-G	220
NR	80
G	41
TV-Y7...	6
Unkn...	4
NC-17	3
UR	3
74 min	1
84 min	1
66 min	1

6. Top 10 directors with the highest number of titles on Netflix.

director	total_titles
Rajiv Chilaka	19
Raúl Campos, Jan Suter	18
Suhas Kadav	16
Marcus Raboy	16
Jay Karas	14
Cathy Garcia-Molina	13
Jay Chapman	12
Youssef Chahine	12
Martin Scorsese	12
Steven Spielberg	11

7. Year-wise distribution of movie releases on Netflix.

release_year	total_movies
2017	767
2018	767
2016	658
2019	633
2020	517
2015	398
2021	277
2014	264
2013	225
2012	173
2010	154
2011	145
2009	118
2008	113
2006	82
2007	74
2005	67
2004	55

8. Countries producing the highest number of TV Shows on Netflix.

country	total_tv_shows
United States	760
United Kingdom	213
Japan	169
South Korea	158
India	79
Taiwan	68
Canada	59
France	49
Australia	48
Spain	48

9. Distribution of movie durations on Netflix

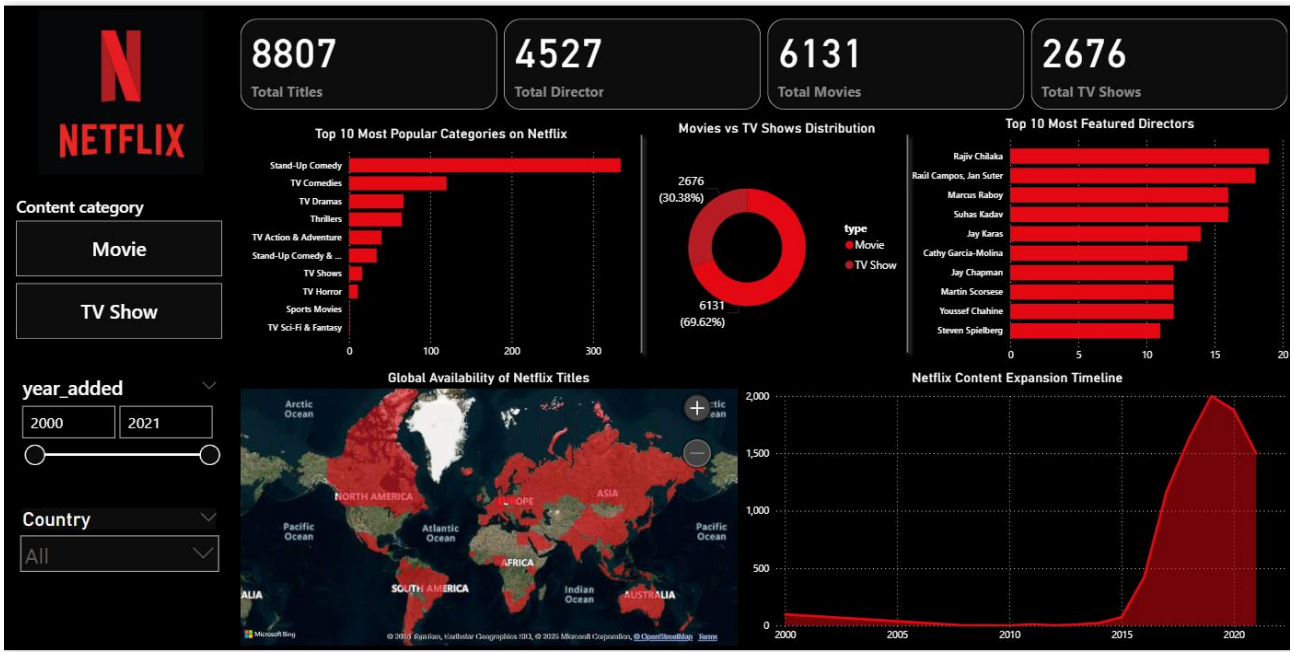
movie_minutes	frequency
90	152
94	146
97	146
93	146
91	144
95	137
96	130
92	129
102	122
98	120

10. Monthly trend of new content additions on Netflix.

month_added	total_added
January	825
July	819
December	797
September	765
April	759
October	755
August	749
March	734
June	724
November	697
May	626
February	557

5. Dashboard in Power BI

Finally, we built an interactive dashboard in Power BI to present insights visually.



6. Business Insights & Recommendations

- Movies dominate the catalog → expand TV shows for higher engagement.

- Popular genres can guide future content investments.

- High-production countries can be targeted for partnerships.

- Additions by month can inform marketing timelines.

- Frequent directors can be approached for exclusives.

- Movie duration insights help align with viewer preferences.

7. Conclusion

This project demonstrates full-cycle data analysis, including cleaning, EDA, SQL querying, and business insight generation. The findings support strategic decisions for streaming platforms.