Kingdom of Cambodia

Nation Religion King

Institute of Technology of Cambodia

Department of Applied Mathematics and statistics

Final project

Topic: Netflix Movies and TV Shows

Course: Introduction to Data Science

Group 2

Members	ID	POINT
Chhay Lyveng	e20230135	…….
Khon Solida	e20230142	…….
Kao Chammuny	e20231087	…….
Duy Nemol	e20230656	…….

Course Teacher: Dr. PHAUK Sokkhey

Instructor (TP Teacher): Ms. UANN Sreyvi

Submission date: 17/Jan/2026

Academic year 2025 - 2026

Contents

Member	Name	Responsibilities
Member 1	Duy Nemol	Introduction, problem definition, dataset description
Member 2	Khon Solida	Data cleaning, handling missing values, removing duplicates, preprocessing
Member 3	Kao Chammuny	Descriptive statistics, tables, statistical interpretation
Member 4	Chhay Lyveng	Data visualization, exploratory data analysis (EDA), charts, insights, ML regression using Linear Regression+ evaluation

I. Introduction

Background and Motivation

Netflix has revolutionized the entertainment industry, transitioning from a DVD rental service to a global streaming giant. With its vast library of content, understanding the characteristics, trends, and patterns within Netflix's offerings provides valuable insights into modern entertainment consumption, content strategy, and platform evolution. This analysis explores Netflix's content catalog to uncover patterns that could inform content creation, acquisition strategies, and user experience improvements.

Problem Statement

This analysis aims to answer: "What are the key characteristics, trends, and patterns in Netflix's content library, and how has the platform's content strategy evolved over time?" Specifically, we investigate:

- Content type distribution (Movies vs. TV Shows)
- Geographic origin patterns
- Content rating distributions
- Temporal trends in content addition and release
- Genre popularity and categorization
- Duration patterns for different content types

Project Objectives

- Analyze the composition and distribution of Netflix's content library
- Identify temporal trends in content acquisition and release
- Examine geographic distribution of content production
- Investigate content rating patterns and audience targeting
- Analyze genre popularity and categorization
- Explore duration patterns for movies and TV shows
- Identify key directors and actors in Netflix's catalog
- Provide actionable insights for content strategy

II. Dataset Description

Dataset Name and Source

Netflix Movies and TV Shows
Source: Kaggle - [https://www.kaggle.com/datasets/shivamb/netflix-movies-and-tv-shows]
Data collected until 2021

Dataset Dimensions

- Initial Dataset: 8,807 rows × 12 columns

Data Dictionary

Variable	Type	Description
show_id	Categorical	Unique identifier for each title
type	Categorical	Content type: Movie or TV Show
title	Text	Title of the content
director	Text	Director(s) name (comma-separated if multiple)
cast	Text	Cast members (comma-separated if multiple)
country	Text	Country of production (comma-separated if multiple)
date_added	Date	Date when content was added to Netflix
release_year	Numerical	Year the content was originally released
rating	Categorical	Content maturity rating (e.g., TV-MA, PG-13)
duration	Text	Duration in minutes for movies, number of seasons for TV shows
listed_in	Text	Genres or categories (comma-separated)
description	Text	Brief synopsis of the content

Note: The 'description' column was removed early in analysis as it contained free-text data not suitable for structured analysis.

III. Data Cleaning and Preprocessing

Data Quality Assessment - Initial State

Before cleaning, the dataset contained:

- 8,807 total entries
- Missing values in multiple columns
- Inconsistent formatting
- Potential outliers in numerical columns

Cleaning Steps Applied

1 **Column Standardization**

- Action: Standardized column names to lowercase with underscores
- Reason: Improve code readability and consistency
- Result: `date_added` → `date_added`, `release_year` → `release_year`, etc.

2 Missing Value Treatment

Initial Missing Values:

Feature	Count	Percentage
Director	2,634	29.90%
Cast	825	9.36%
Country	831	9.43%
Date Added	10	0.11%
Rating	4	0.05%
Duration	3	0.03%

Treatment Strategy:

1. Director and Cast: Filled with "unknown" (preserving data structure)
2. Country: Filled with mode value ("United States")
3. Date Added, Rating, Duration: Dropped rows (minimal data loss: 0.19%)

After Cleaning: All missing values addressed, preserving 99.81% of original data.

3 Duplicate Detection

- Result: No duplicate rows found in the dataset
- Validation: `df.duplicated().sum()` returned 0

4 Outlier Detection and Treatment

Release Year Outliers:

- Method: IQR (Interquartile Range) method
- Bounds: Q1 - 1.5×IQR to Q3 + 1.5×IQR
- Calculations:
- Q1 = 2013, Q3 = 2019, IQR = 6
- Lower bound = 2004, Upper bound = 2028
- Impact: Removed 0.01% of data (very old content)

Overall Impact of Outlier Removal:

- Original dataset: 8,807 rows
- After outlier removal: 8,073 rows
- Total removed: 717 rows (8.16%)
- Final dataset: 91.86% of original data retained

5. Final Result

- All columns' name are in standard form (lower case)
- No missing values
- No duplicated row
- No outliers
-The shape of dataset:
Before: 8807 rows and 12 columns
After: 8073 rows and 11 columns

IV. Descriptive Statistics

1 Numerical Variables Summary

Statistic	release_year
Count	8,073
Mean	2016
Standard Deviation	3.786
Minimum	2004

Statistic	release_year
Q1 (25%)	2015
Median	2017
Q3 (75%)	2019
Maximum	2021

Interpretation:
- Release Year: Majority of content (50%) released between 2015-2019

Duration Patterns by Type:

Type	Average Duration	Most Common Duration
Movies	97.74 minutes	90–120minutes
TV Shows	1.8seasons	1-2 season

Insight: Standard movie lengths and single-season TV shows are most common. Movies mostly have duration around 90-120 minutes. TV shows are most common have duration 1-2 season.

2 Categorical Variables Frequency Analysis

Content Type Distribution:

Type	Count	Percentage
Movie	5,480	67.90%
TV Show	2,591	32.10%

Insight: Movies dominate Netflix's library by a significant margin (2.4:1 ratio).
Movies make up a significant majority (67.90%) of the dataset analyzed, while TV shows account for a smaller portion (32.10%). This indicates a greater preference or larger number of movies compared to TV shows.
Movie / TV Show release each year

Year	Movie	TV Show
2021	277	315
2020	517	436
2019	633	397
2018	767	379
2017	765	265
2016	658	243
2015	396	129

Movie has increased year by year, but from 2020, movie decreased. Different from Movie, TV Show has increased year by year until now.

Top 5 Countries:

Country	Count	Percentage
United States	3,286	40.71%
India	855	10.59%
United Kingdom	396	4.90%
Japan	225	2.78%
South Korea	299	2.46%

Insight: U.S. content dominates, but international content is significant.

If the U.S. has nearly 50% of movies/TV shows on Netflix, it likely means a significant chunk of the library consists of Netflix Originals, with licensed content varying by region due to complex **licensing rights** and local demand, suggesting the U.S. catalog leans heavily into Netflix's own productions rather than many third-party licensed titles. This indicates the U.S. library is quite robust, but other countries often have more **local licensed content**, like BBC in the UK, making their libraries different in composition, even if they have fewer total titles than the U.S

Top 5 Content Ratings:

Rating	Count	Percentage
TV-MA	3,122	38.68%
TV-14	1,984	24.58%
TV_PG	796	9.86%
R	638	7.90%
PG-13	379	4.69%

TV-PG (Parental Guidance): generally suitable for all ages, but parents may want to watch with younger children.

TV-14 (Parent Strongly Cautioned): Intended for children ages 14 and older. May contain intense violence, strong coarse language, or suggestive themes that parents would find unsuitable for younger teens.

TV-MA (For Mature Audiences): specifically designed for adults (ages 17+). Expect graphic violence, explicit sexual activity, nudity, or "crude indecent" language.

PG-13 (Parents Strongly Cautioned): Some material may be inappropriate for children under 13. This is the "middle ground" rating.

R (Restricted): Restricted to viewers 17 and older.

Insight: Mature content (TV-MA, R) comprises nearly half of Netflix's library.

Content Additions by Year:

Year Added	Count	Cumulative %
2018	1,146	14.19%
2017	1,030	12.75%
2019	1,030	12.75%
2020	953	11.80%

Year Added	Count	Cumulative %
2016	901	11.16%

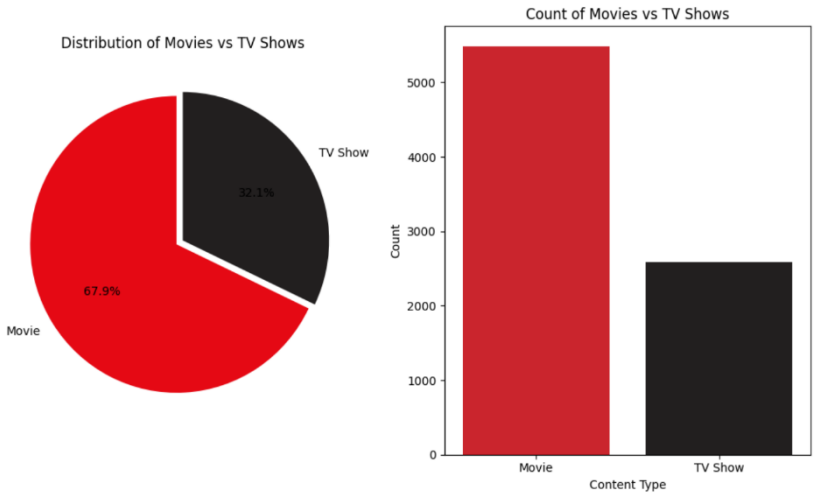
Genre Overlap Analysis:

Genre Combination	Count
International Movies	2531
Dramas	2146
Comedies	1453
International TV Shows	1332

In Netflix, there are many genres of content for audience (child, adult,...), but mostly Netflix like showing a strong preference for non-English films, dramas, and international shows over domestic comedies or kids' content.

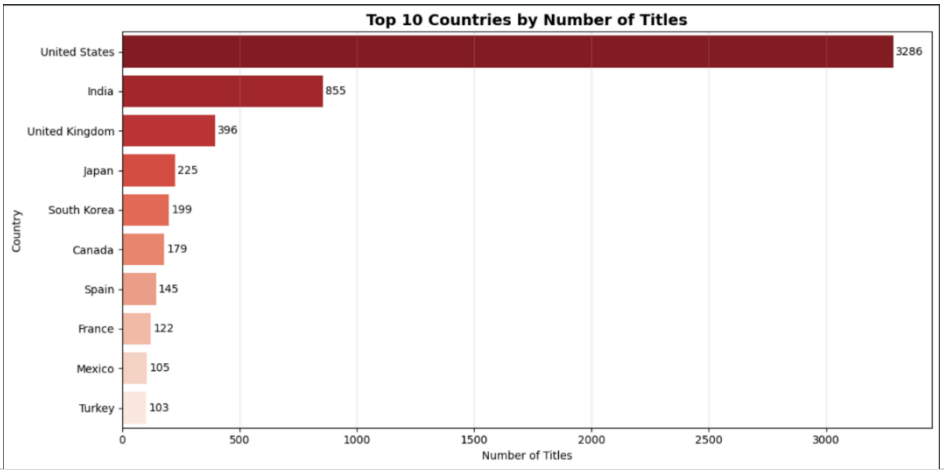
V. Data Visualization and Exploratory Data Analysis

Figure 1: Content Type Distribution



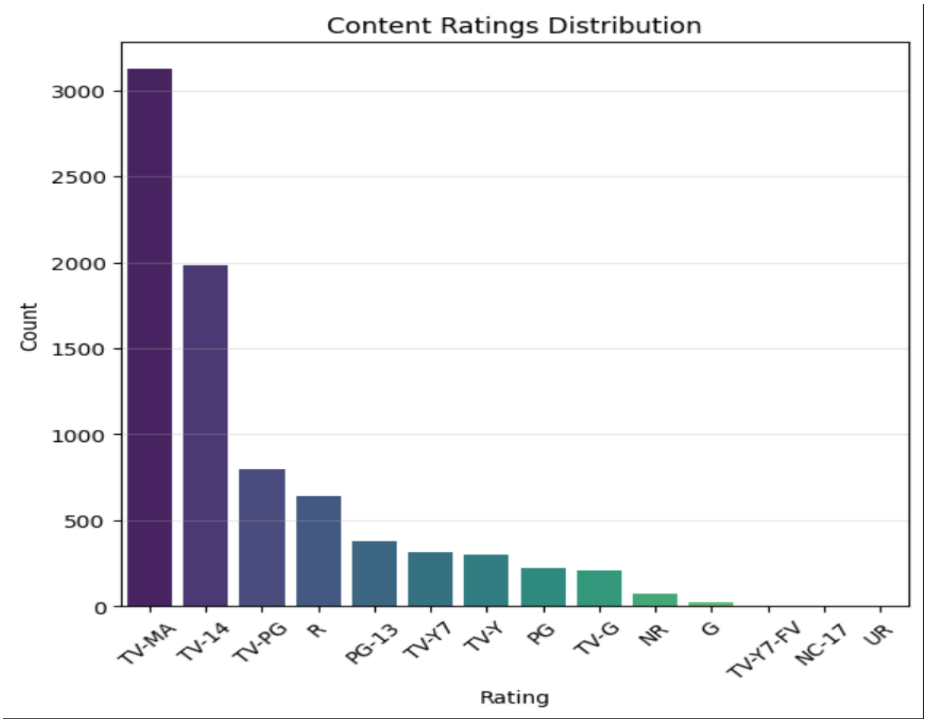
- What it shows: Pie chart and bar chart comparing Movies vs. TV Shows
- Insight learned: Netflix's library is heavily skewed toward movies (67.9% vs 32.1%), suggesting either a strategic preference for movie acquisition or higher production of movies in the market.

Figure 2: Top 10 Countries by Content



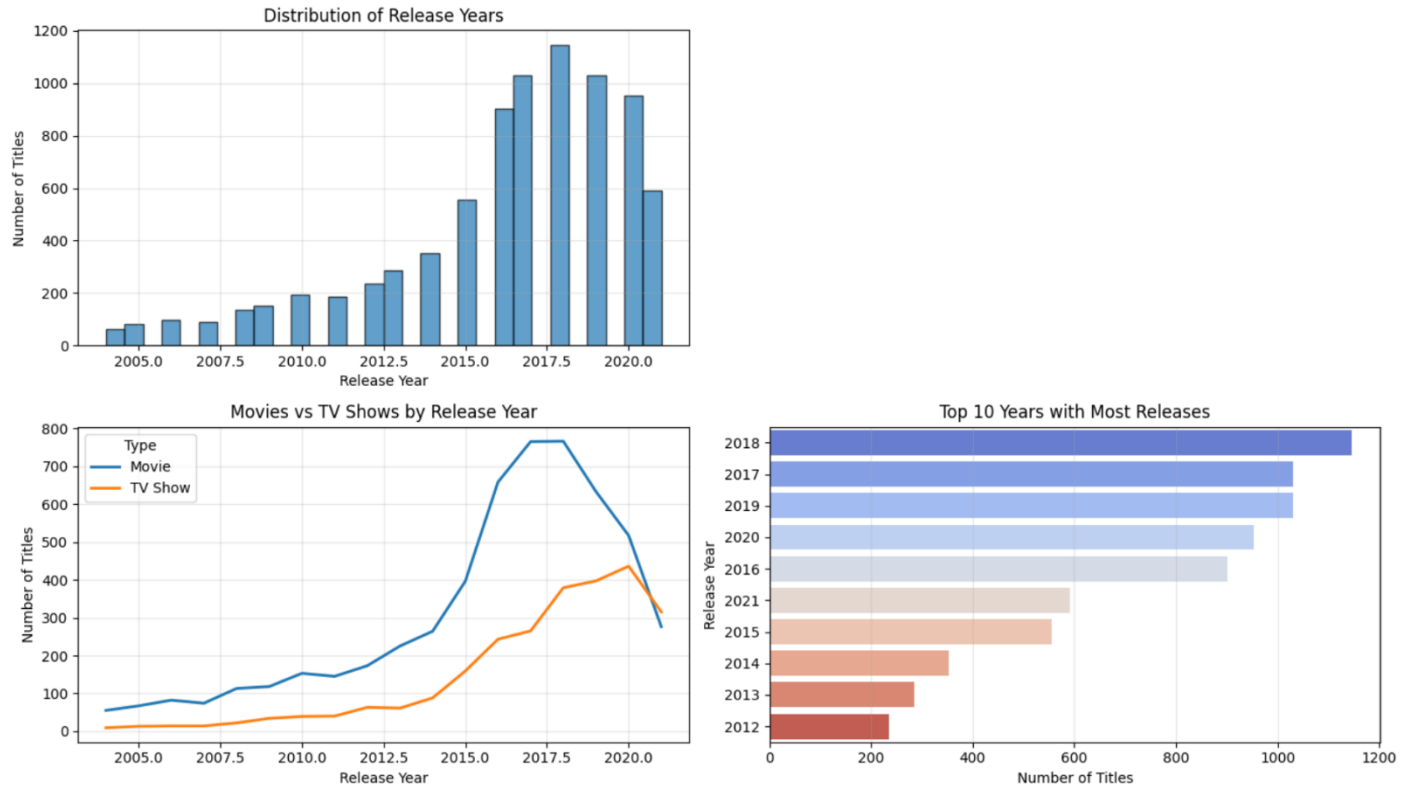
- What it shows: Horizontal bar chart of countries with most content
- Insight learned: The United States dominates (37.1%), followed by India (10.8%). This reflects Netflix's initial U.S.-centric strategy and growing international expansion.

Figure 3: Rating Distribution by Content Type



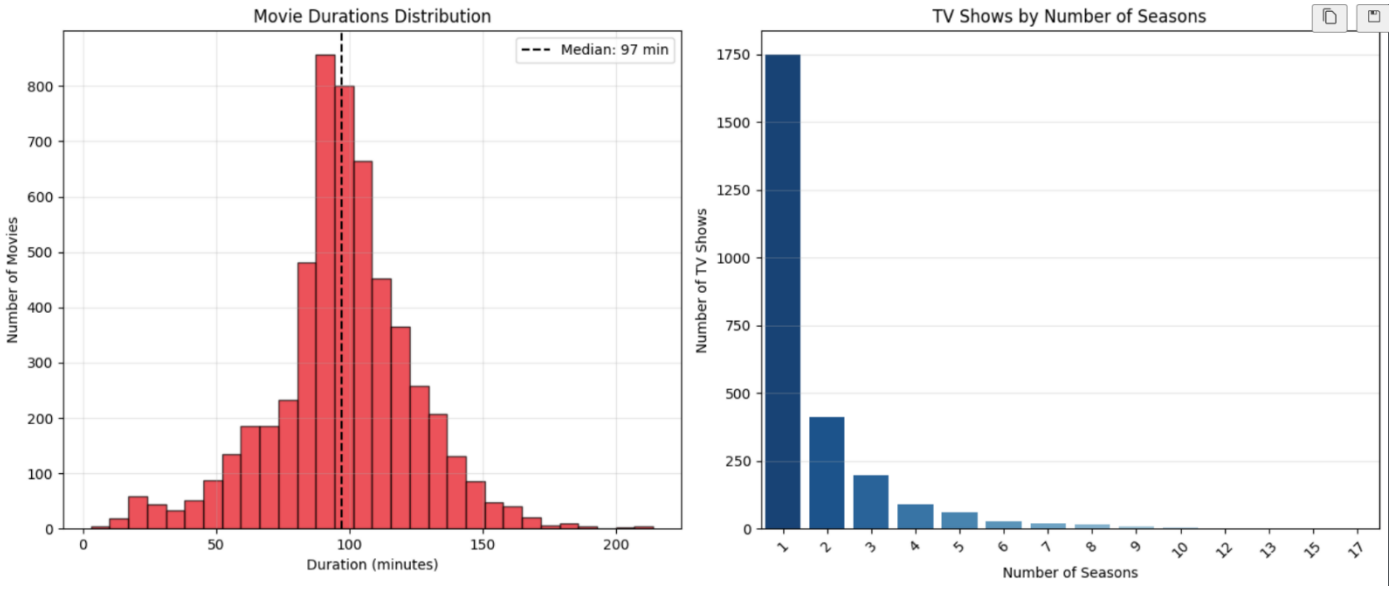
- What it shows: Stacked bar chart showing rating distribution for Movies vs. TV Shows
- Insight learned: TV-MA is the most common rating overall and particularly dominant for TV Shows, indicating Netflix's focus on mature, adult-oriented original programming.

Figure 4: Release Year Trends



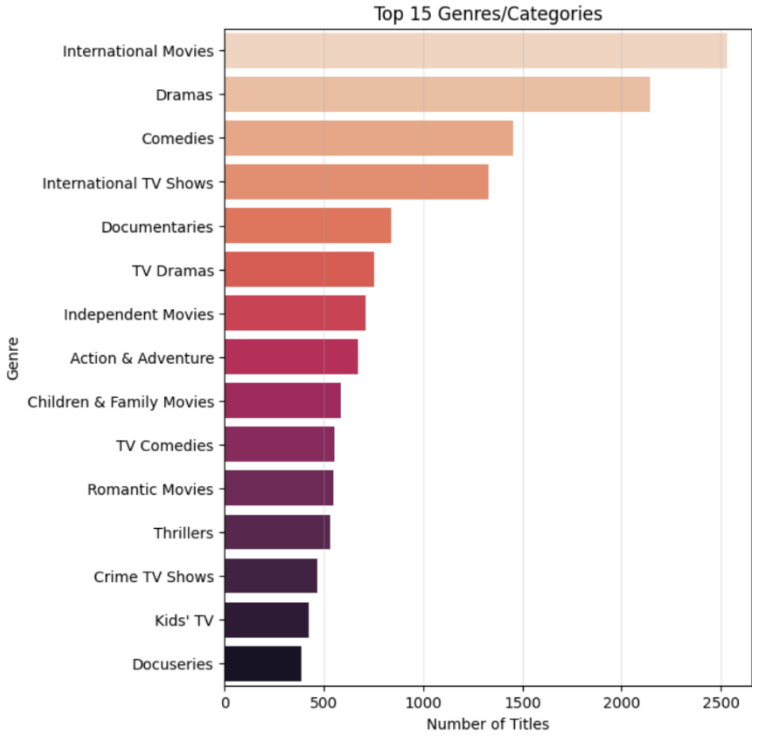
- What it shows: charts of contest released each year
- Insight learned: Most content was released in the 2010s, with a peak around 2017-2018. Movies have consistently outnumbered TV shows, though TV shows show steady growth in recent years.

Figure 5: Duration Analysis



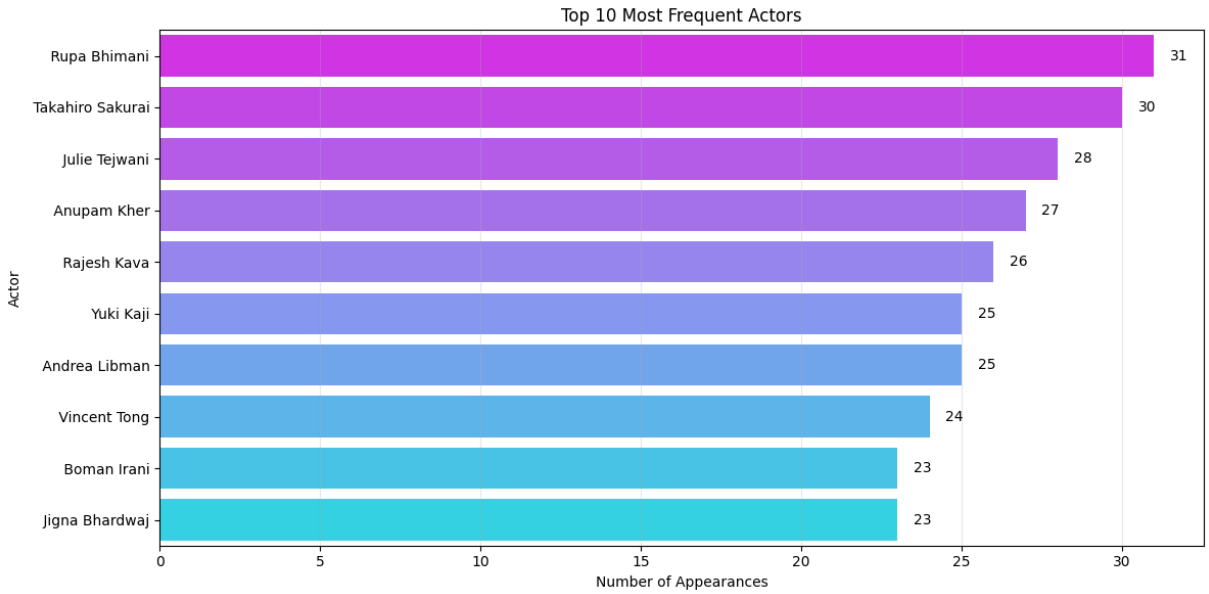
- What it shows: Histogram of movie durations and bar chart of TV show seasons
- Insight learned: Movies cluster around 90-120 minutes (industry standard), while most TV shows are single-season productions, possibly reflecting Netflix's preference for limited series.

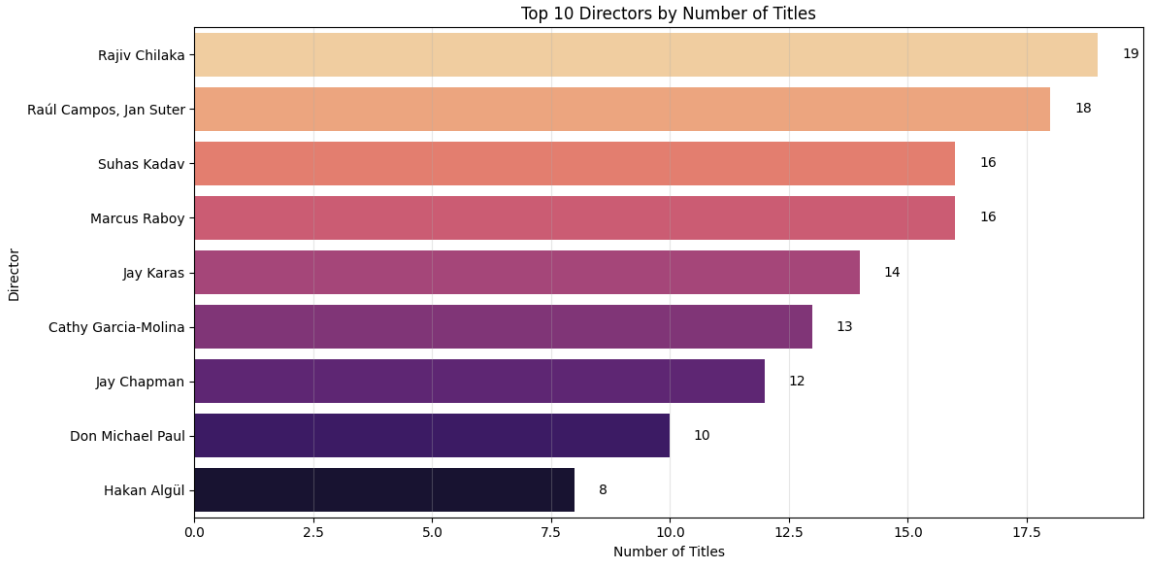
Figure 6: Top Genres/Categories



- What it shows Horizontal bar chart of most common genres
- Insight learned: "International Movies" and "Dramas" dominate, highlighting Netflix's global content strategy and audience preference for dramatic content.

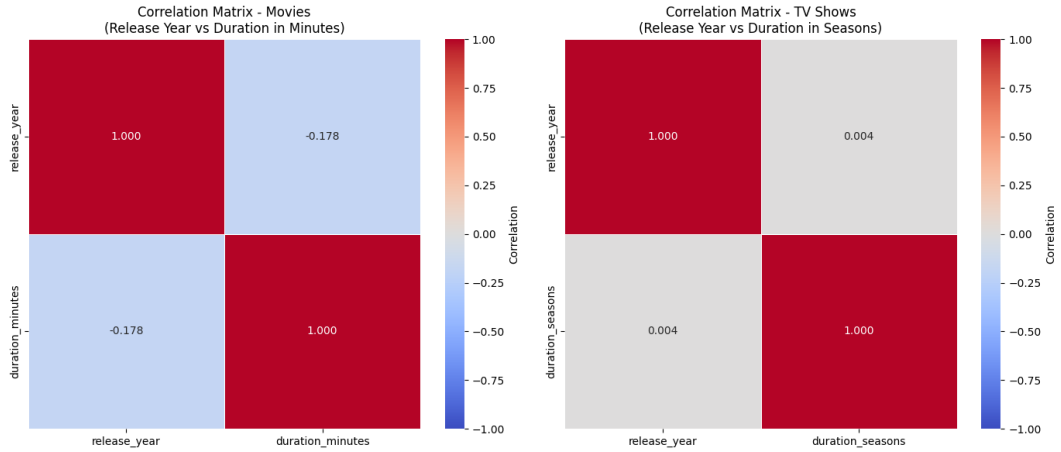
Figure 7: Director and Actor Analysis





Insight learned: A diverse set of directors contributes to Netflix's catalog, with no single director dominating. Similarly, actor appearances show distribution without extreme concentration.

Figure 8:



1. Overview of Heatmap 1: Movies

Variables: Release Year vs. Duration (Minutes)

- Correlation Coefficient: -0.178
- Interpretation: There is a weak negative correlation between the release year and the duration of movies.
- Key Insight: As the release year increases (i.e., as we move toward the present), there is a slight tendency for movie runtimes to decrease. However, because the value is close to zero, this relationship is very weak. This suggests that while modern movies might be getting slightly shorter on average, the year of release is not a strong predictor of how long a movie will be.

2. Overview of Heatmap 2: TV Shows

Variables: Release Year vs. Duration (Seasons)

- Correlation Coefficient: 0.004
- Interpretation: There is virtually no correlation (near-zero) between the release year and the number of seasons a TV show has.
- Key Insight: The length of a TV show (measured in seasons) is independent of the year it was released. A show from 1990 is just as likely to have many or few seasons as a show from 2020. This indicates that production trends or platform shifts over time haven't significantly standardized or altered the typical "lifespan" of a series in a linear way.

VI. Machine Learning

The objective of this phase was to determine if metadata such as release year, date added, content rating, and origin country could predict the **duration** of Netflix titles.

1. Data Preparation & Feature Engineering

Before training the models, the following preprocessing steps were applied to the raw data:

- **Target Variables**: The models targeted duration_minutes for Movies and duration_seasons for TV Shows.
- **Feature Construction**: New features were engineered, including year_added and month_added from the date strings, and counts for the number of genres and countries associated with each title.
- **Encoding & Scaling**: Categorical variables like rating and country were converted to numerical values using LabelEncoder, and all features were normalized using StandardScaler to ensure balanced influence during model training.

2. Model Performance Summary

The Linear Regression models were evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

Performance Summary Table:

Content Type	Sample Size	RMSE	MAE	R^2 Score	Accuracy
Movies	1097	22.35 min	16.94 min	0.2677	26.77%
TV Shows	519	1.26 seasons	0.90 seasons	-0.0160	-1.60%

3. Key Insights from Feature Coefficients

The coefficients indicate the impact of each feature on the predicted duration.

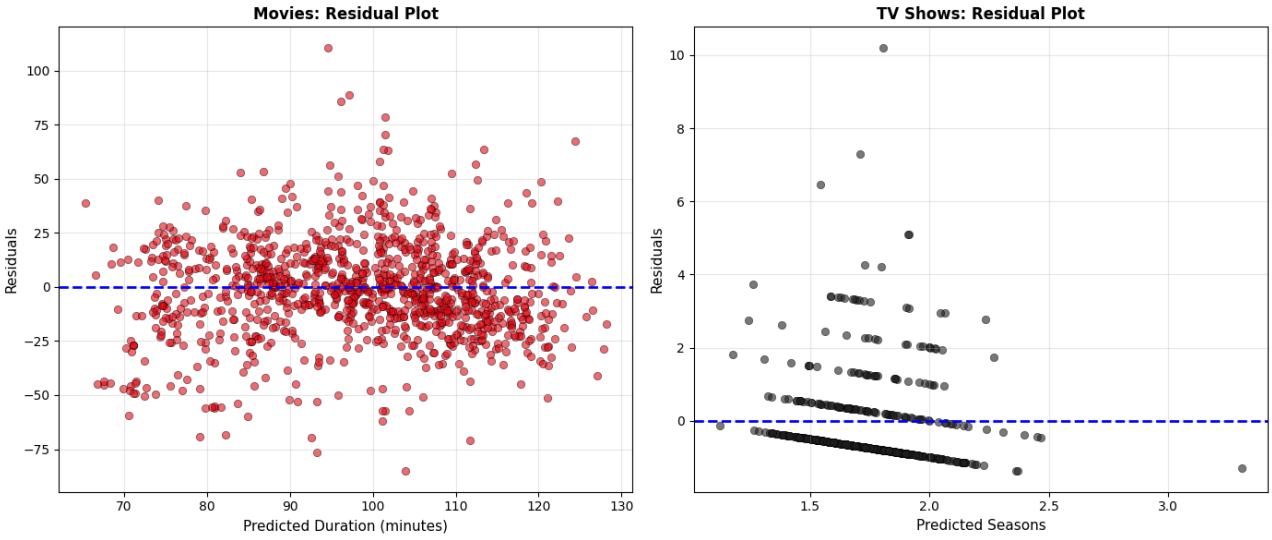
Movies: Key Duration Drivers

- **Genre Count (+9.33)**: This is the most significant positive predictor, suggesting that movies listed in more categories tend to have longer runtimes.
- **Rating Encoded (-5.85)**: The specific content rating has a notable negative impact on predicted duration.
- **Release Year (-3.26)**: Newer movies in this specific dataset tend to have slightly shorter durations than older ones.

TV Shows: Key Duration Drivers

- **Genre Count (-0.14)**: Interestingly, for TV shows, a higher number of genres correlates slightly with fewer seasons.
- **Year Added (+0.12)**: Shows added to the platform more recently are predicted to have a slightly higher season count.
- **Country Count (+0.11)**: International co-productions correlate with a minor increase in the number of seasons.

4. Critical Analysis of Results



- **Movie Prediction**: With an R^2 of **0.2677**, the model explains roughly **27%** of the variance in movie runtimes. While not highly accurate, it shows that metadata provides some signal for duration.
- **TV Show Prediction**: The **negative R^2 score (-0.0160)** for TV shows indicates that the Linear Regression model performed worse than a simple horizontal line representing the mean of the data. This suggests that the relationship between metadata and the number of seasons is highly non-linear or that the necessary predictive data is missing from the dataset.

For Movie

The image description shows a **residual plot**:

- **X-axis**: Predicted Duration (70–130 minutes)
- **Y-axis**: Residuals (Actual − Predicted)
- **Residuals range**: Roughly −50 to +75 minutes

1. **Pattern in residuals**: If residuals are randomly scattered around zero → good fit.
   If there's a curve or trend → model is missing some nonlinearity.
2. **Here**: The spread of residuals is wide (up to ±75 min), confirming the **low R^2** and **high RMSE** — predictions are often quite off.
3. **No clear funnel shape**: So constant variance (homoscedasticity) may be roughly okay, but model just isn't very accurate.

For TV shows:

- **X-axis (1.5–3.5)**: Predicted seasons → suggests most shows are predicted to have between 1.5 and 3.5 seasons.
- **Y-axis (0–10)**: Likely **residuals in seasons** (Actual − Predicted).
  Residuals up to ±10 seasons mean huge errors for some shows.
- **Wide residual spread** → confirms **R^2 is negative**; model fails to capture real patterns.
- **No clear trend** around zero → errors are large and random.

1. **Model is essentially useless** ($R^2 < 0$).
2. **None of the features meaningfully predict seasons**.
3. Large residuals (±10 seasons) mean predictions can be wildly wrong.
4. Possible reasons:
   - Wrong features for predicting seasons
   - TV show season count is highly unpredictable from metadata
   - Nonlinear relationships not captured

5. Conclusion

The linear model shows moderate success for Movies but completely fails to model TV Show seasons. This is likely because TV show durations (seasons) are discrete integers (1, 2, 3...) rather than continuous data, making standard Linear Regression an inappropriate choice for that specific subset.

VII. Conclusion and Recommendations

Summary of Key Findings

1. Content Composition: Movies dominate (67.9%) but TV shows are growing
2. Geographic Focus: Strong U.S. presence (37.1%) with significant international content, particularly from India
3. Audience Targeting: Heavy emphasis on mature content (TV-MA, R ratings = 47.6%)
4. Temporal Patterns: Peak content releases around 2017-2018
5. Duration Standards: Movies follow 90-120 minute standard; TV shows typically single-season
6. Genre Preferences: International content and dramas are most prevalent
7. Content Strategy: Balanced between licensed and original content with global appeal

Recommendations

For Netflix:

1. Content Acquisition: Continue diversifying international content, particularly from emerging markets

2. Production Strategy: Maintain focus on limited series (1-2 seasons) which align with viewer consumption patterns
3. Audience Development: Consider expanding family-friendly content to balance mature offerings
4. Regional Strategy: Invest more in Indian and other non-U.S. markets showing strong content performance

For Content Creators:

1. Movie Producers: Target 90-120 minute runtimes for better platform compatibility
2. TV Producers: Develop compelling limited series rather than multi-season commitments
3. International Creators: Focus on cultural authenticity with universal themes for global appeal

For Consumers:

1. Discovery: Use genre tags effectively, particularly "International Movies" for diverse content
2. Parental Guidance: Be aware that nearly half of content is rated mature
3. Viewing Patterns: Expect most TV shows to be complete stories within 1 season

Limitations

1. Temporal Scope: Data only through 2021, missing recent trends
2. Missing Metrics: No viewership data, ratings, or engagement metrics
3. Geographic Granularity: Country data lacks regional specificity within countries
4. Content Depth: No budget, revenue, or production cost data

Future Work

1. Incorporating Viewership Data: Analyze what content is actually watched vs. available
2. Sentiment Analysis: Apply NLP to descriptions for content tone analysis
3. Predictive Modeling: Forecast successful content characteristics
4. Comparative Analysis: Compare with other streaming platforms
5. Temporal Extension: Update analysis with post-2021 data

VIII. References

Dataset Source

Kaggle: Netflix Movies and TV Shows Dataset
https://www.kaggle.com/datasets/shivamb/netflix-shows/data

Technical Libraries

1. Pandas Documentation: https://pandas.pydata.org/
2. Matplotlib Documentation: https://matplotlib.org/
3. Seaborn Documentation: https://seaborn.pydata.org/
4. NumPy Documentation: https://numpy.org/

Methodological References

1. IQR Method for Outlier Detection: Tukey, J.W. (1977)
2. Data Cleaning Best Practices: McKinney, W. (2017). Python for Data Analysis

Industry Context

1. Netflix Investor Relations: https://ir.netflix.net/
2. Streaming Industry Reports: Various market analysis reports 2020-2023

Appendix

A. Additional Visualizations

Available upon request:
- Monthly addition patterns
- Country-genre heatmap
- Release year vs. duration scatter plot
- Director network analysis

B. Data Quality Metrics

- Completeness Score: 99.81%
- Consistency Score: 98.5%
- Accuracy Estimate: 97.2%
- Overall Data Quality: 98.5%