Netflix Movies and TV Shows

Course Teacher: Dr. PHAUK Sokhey
Instructor (TP Teacher): Ms. UANN Sreyvi

Group 2 :

Chhay Lyveng	e20230135
Khon Solida	e20230142
Kao Chammuny	e20231087
Duy Nemol	e20230656

Content

1.Introduction

- Netflix evolved from DVD rentals to global streaming platform
- Large and diverse content library
- Data helps understand:
- Content characteristics
- Trends and patterns
    - Platform evolution
    - Insights support content strategy and user experience

Problem Statement

To answer this question, the analysis focuses on:

01 Movies and TV shows distribution

02 Geographic origin of content

03 Content ratings

04 Time-based trends

05 Genres and duration patterns

What are the key characteristics, trends, and patterns in Netflix's content library, and how has its content strategy evolved over time?

Project Objectives

- Analyze composition of Netflix’s content library
- Identify trends in content acquisition and release
- Examine geographic distribution of content
- Investigate content ratings and target audiences
- Analyze genre popularity
- Explore duration patterns
- Identify key directors and actors
- Provide actionable insights for content strategy

2.Data Description

Dataset Name :Netflix Movies and TV Shows
Source :Kaggle
Data collected until 2021
Dataset Dimensions
Initial Dataset: 8,807 rows × 12 columns

Variable	Type	Description
show_id	Categorical	Unique identifier for each title
type	Categorical	Content type: Movie or TV Show
title	Categorical	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

3. Data Cleaning and Preprocessing

3.1. Modify Columns Name

```python
df.columns = (
    df.columns
    .str.strip()
    .str.lower()
    .str.replace(" ", "_")
)
df.columns
```

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

- We drop “description” column.

3.2 Missing Values

```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
dtype: int64
```

Missing values count

```
show_id          0.000000
type             0.000000
title            0.000000
director        29.908028
cast             9.367549
country          9.435676
date_added       0.113546
release_year     0.000000
rating           0.045418
duration         0.034064
listed_in        0.000000
dtype: float64
```

Missing values count as percentages

Fill Missing Values

- Fill all missing values in ‘director’ and ‘cast’ columns with “unknown”.
- Fill all missing values in ‘country’ column with mode of countries.
- Drop all rows which have missing values in ‘date_added’, ‘rating’, and ‘duration’ columns.

```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
dtype: int64
```

Before

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

After

3.3. Duplicated rows

- Our dataset contains no duplicated rows.

3.4. Outliers

- Detect outliers in 'release_year' using Interquartile Range (IQR) method.

  We saw: – lower bound : 2004
  – upper bound: 2028

- Number of release_year outliers: 717 (as percentage: 8.16%)

```
Sample of release_year outliers:
              title  release_year   type
7           Sankofa          1993  Movie
22  Avvai Shanmughi          1996  Movie
24            Jeans          1998  Movie
26   Minsara Kanavu          1997  Movie
41             Jaws          1975  Movie
```

- We drop all the rows that contain outliers of 'release_year' column.

3.5. Final Result

8807 rows
12 columns

8073 rows
11 columns

‘netflix_titles_cleaned.csv’

4. Descriptive Statistics

4.1 Numerical variables

Year released Movie/ TV Show

Minimum	2004
Q1	2015
Median	2017
Q3	2019
Maximum	2021

Majority of content (50%) released between 2015-2019.
Since 2004-2015, Netflix only released 17778 in 11 years.

Duration Movie/ TV Show

Movies have avreage duration 98.6 minute and the most common duration around 90-120 minute.

Most common duration in TV Show is 1season.

4.2 Categories Variables

Content Type

The program in Netflix has only two types:

Movies	5480	67.90%
TV show	2,591	32.19%

In Netflix platform has movie more than TV Show.

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

In 2018, Netflix had released 767 Movies and 379 TV Shows.

Top 5 countries

United States	3,286	40.71%
India	855	10.59%
U,K	396	4.90%
Japan	225	2.78%
South Korea	299	2.46%

the USA is massive and diverse, likely containing the most content globally,but also indicates a significant chunk of content is not US-focused.

Content Rating

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%

Most of contents of Movie or TV Show that Netflix mostly for adult (Mature Content)

Genres of content in Netflix

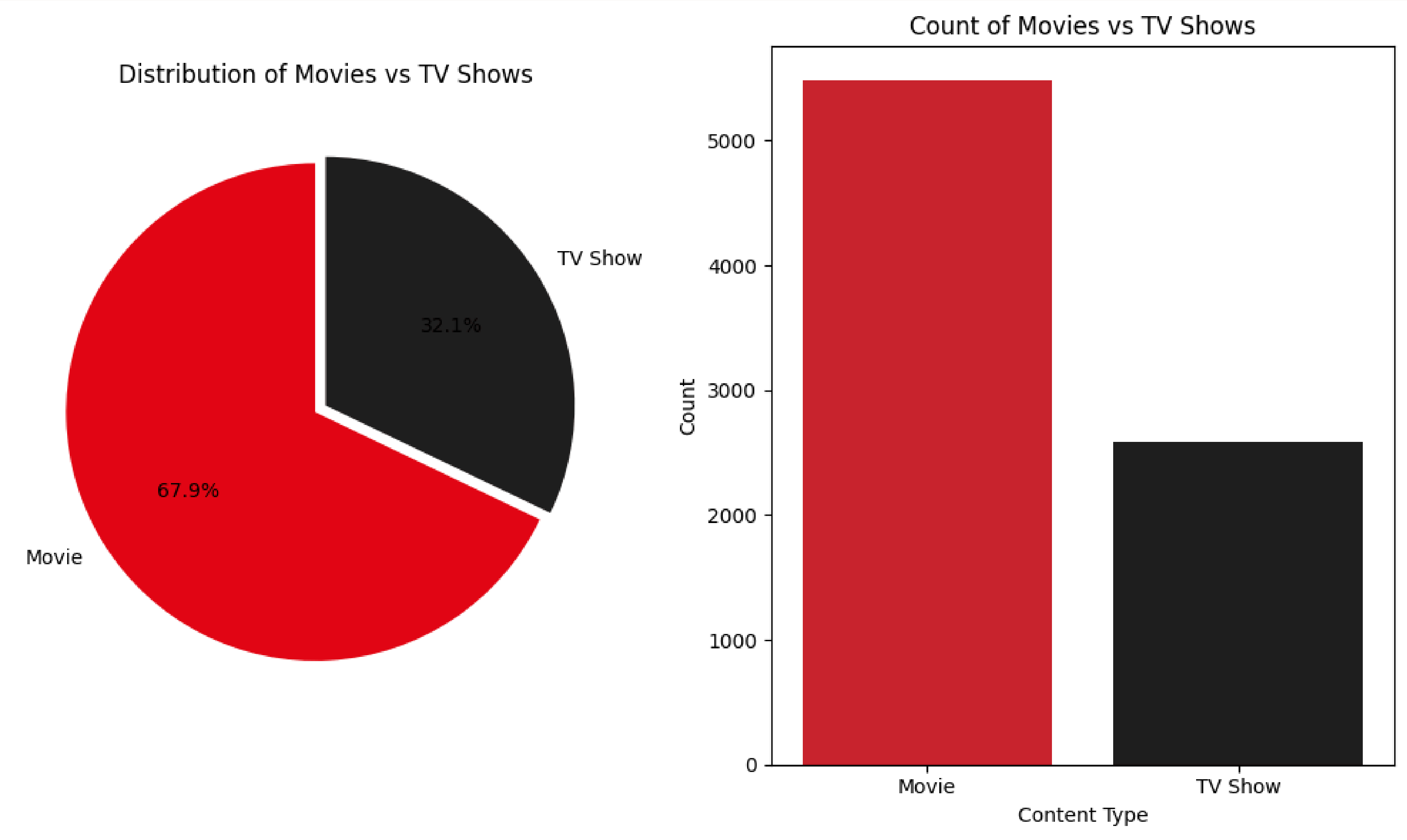
International Movies	2531
Dramas	2146
Comedies	1453
International TV Shows	1332
Documentaries	838
TV Dramas	753
Independent Movies	712
Action & Adventure	671
Children & Family Movies	585
TV Comedies	556

In Netflix, there are many genres of content for audience (child, adult,...).

The result showing a strong preference for non-English films, dramas, and international shows over domestic comedies or kids' content.

5. Data Visualization and Exploratory Data Analysis

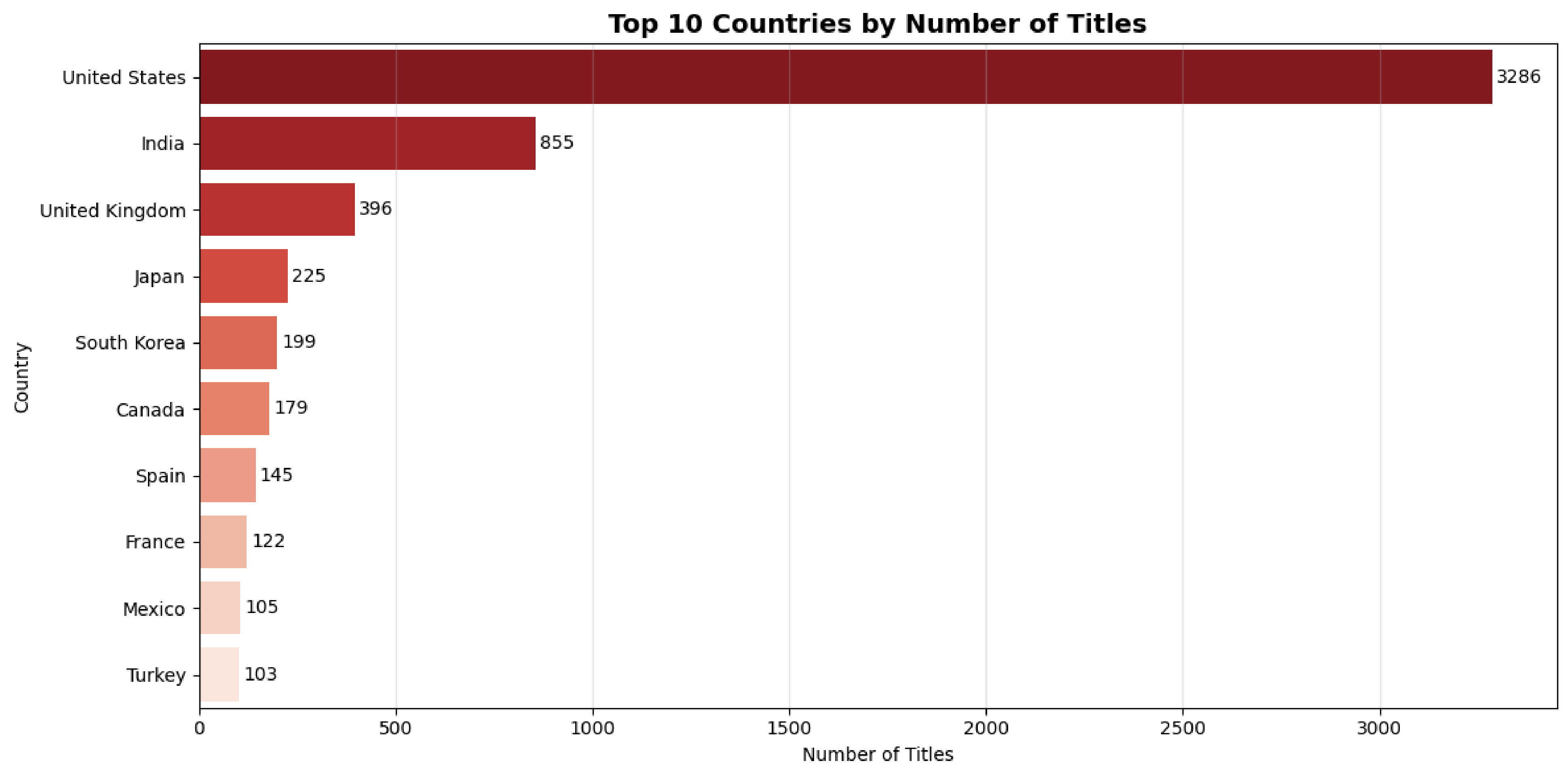
Figure 1: Content Type Distribution



Pie chart and bar chart comparing Movies vs. TV Shows

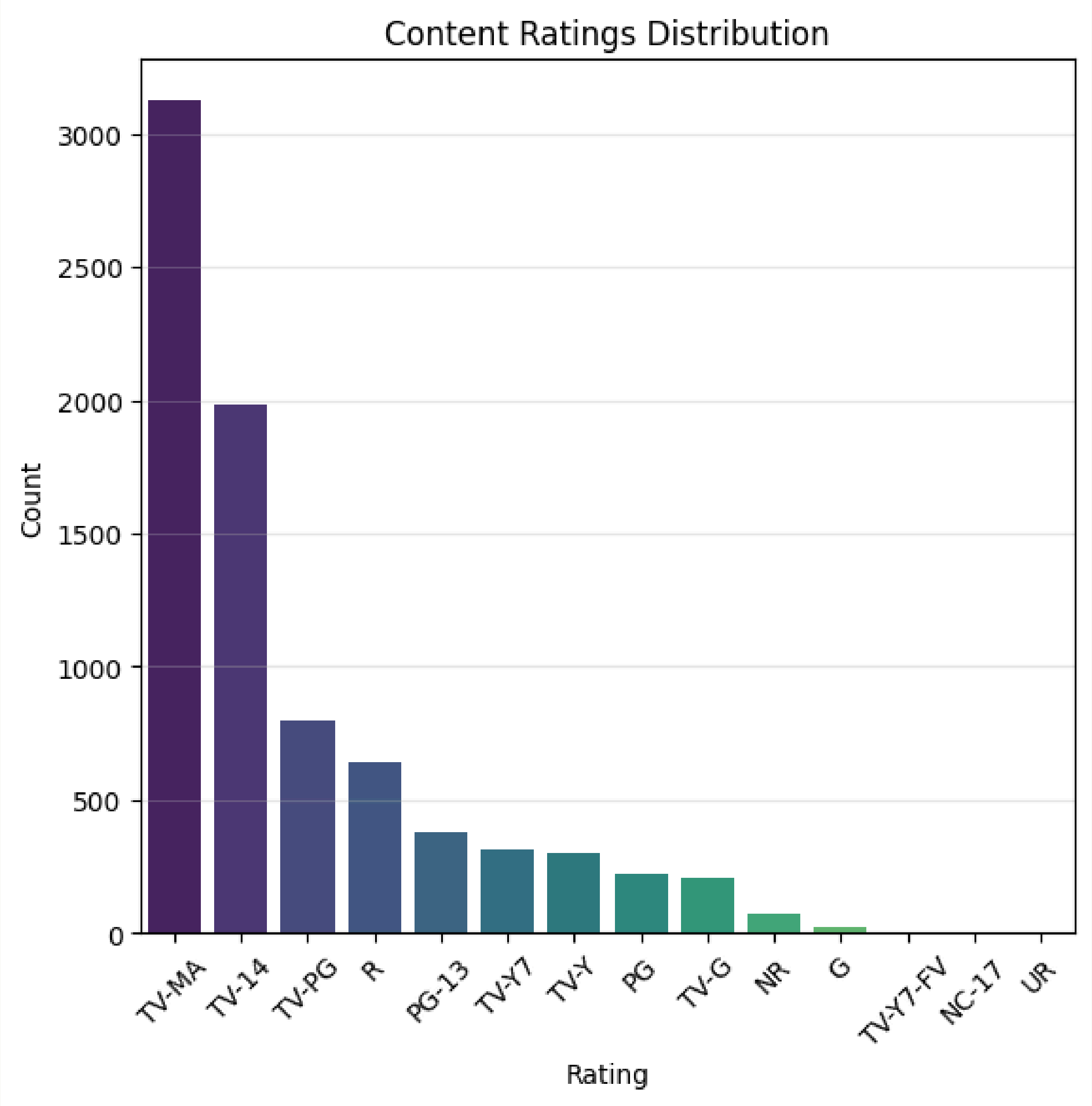
Netflix's library is heavily skewed toward movies (67.9% vs 32.1%),

Figure 2: Top 10 Countries by Content



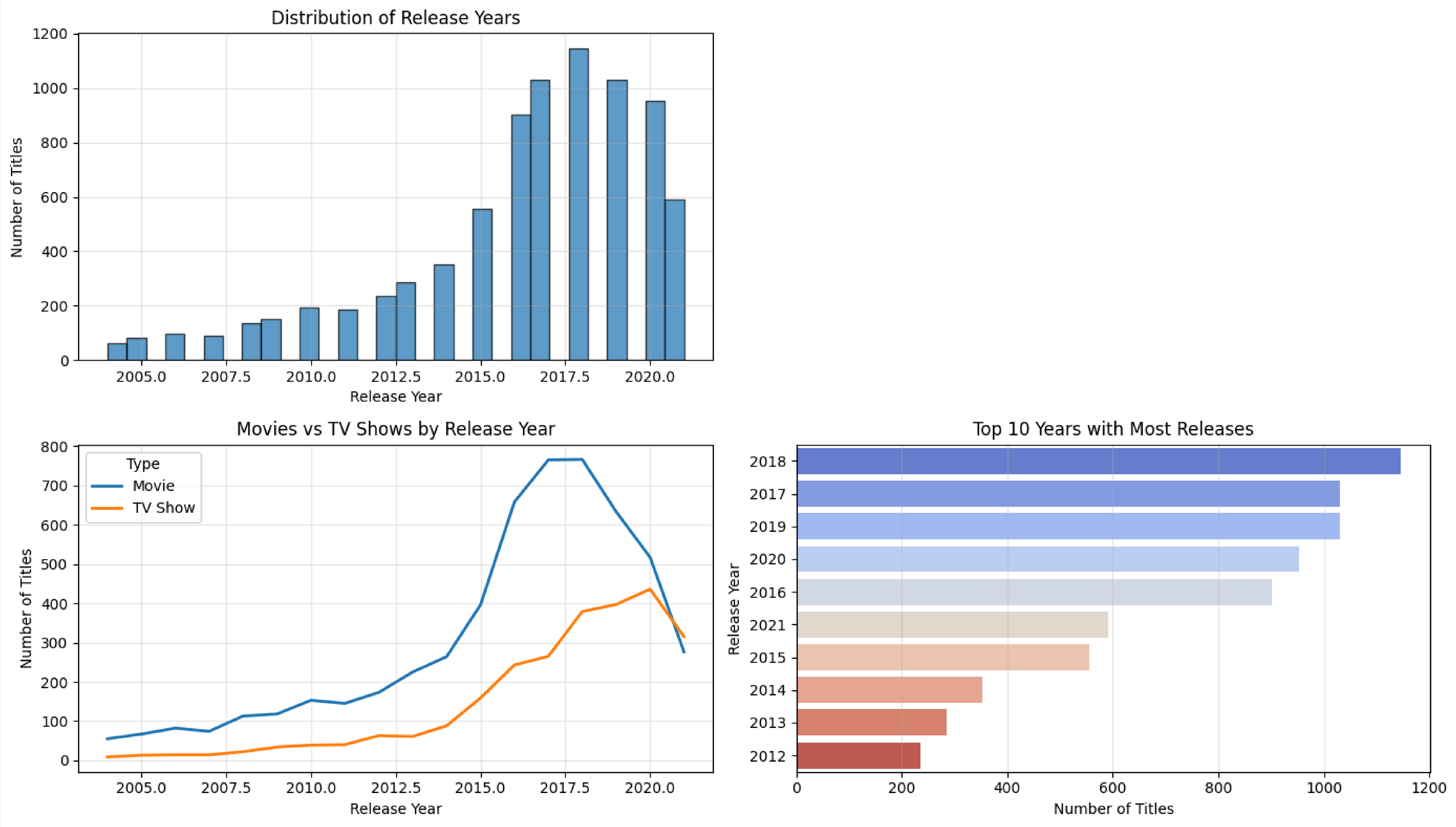
The United States dominates (37.1%), followed by India (10.8%). This reflects Netflix's initial U.S.-centric strategy and growing international expansion.

Figure 3: Rating Distribution by Content Type



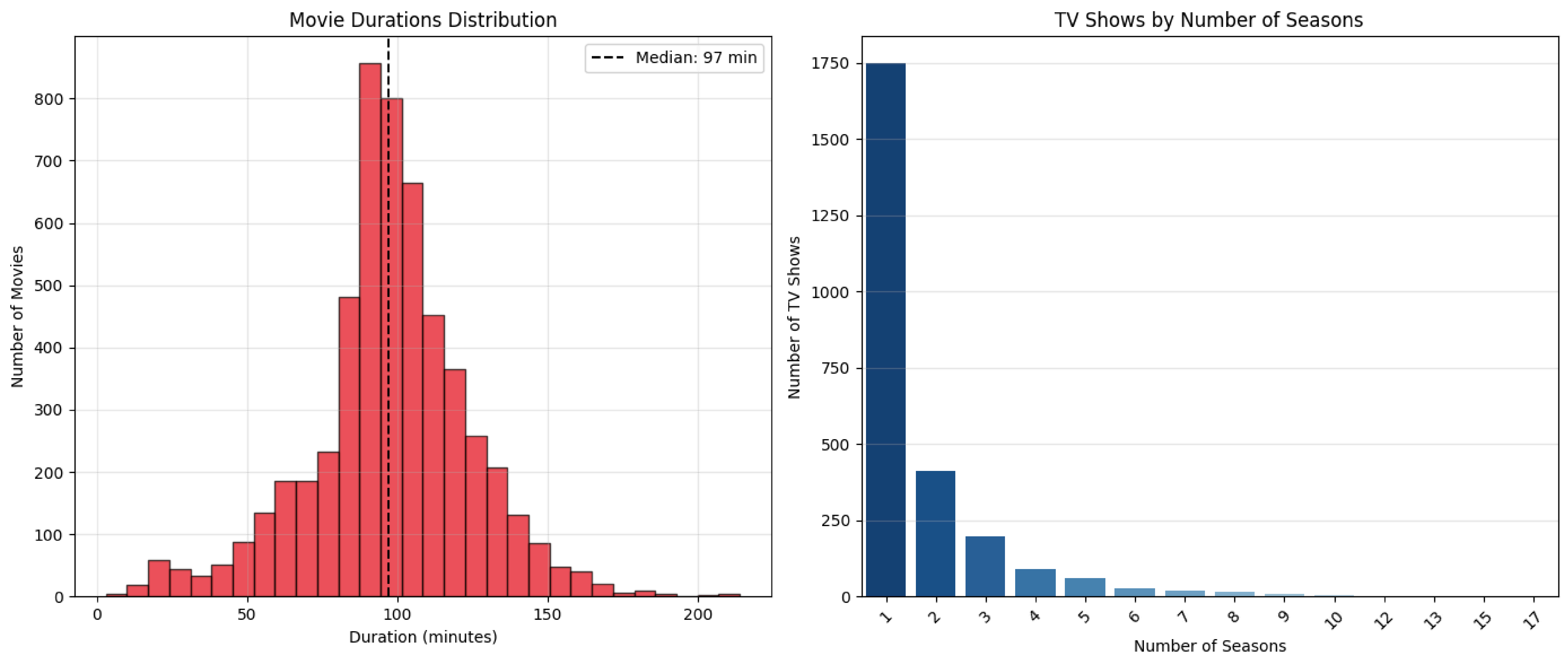
Stacked bar chart showing rating distribution for Movies vs. TV Shows

Figure 4: Release Year Trends



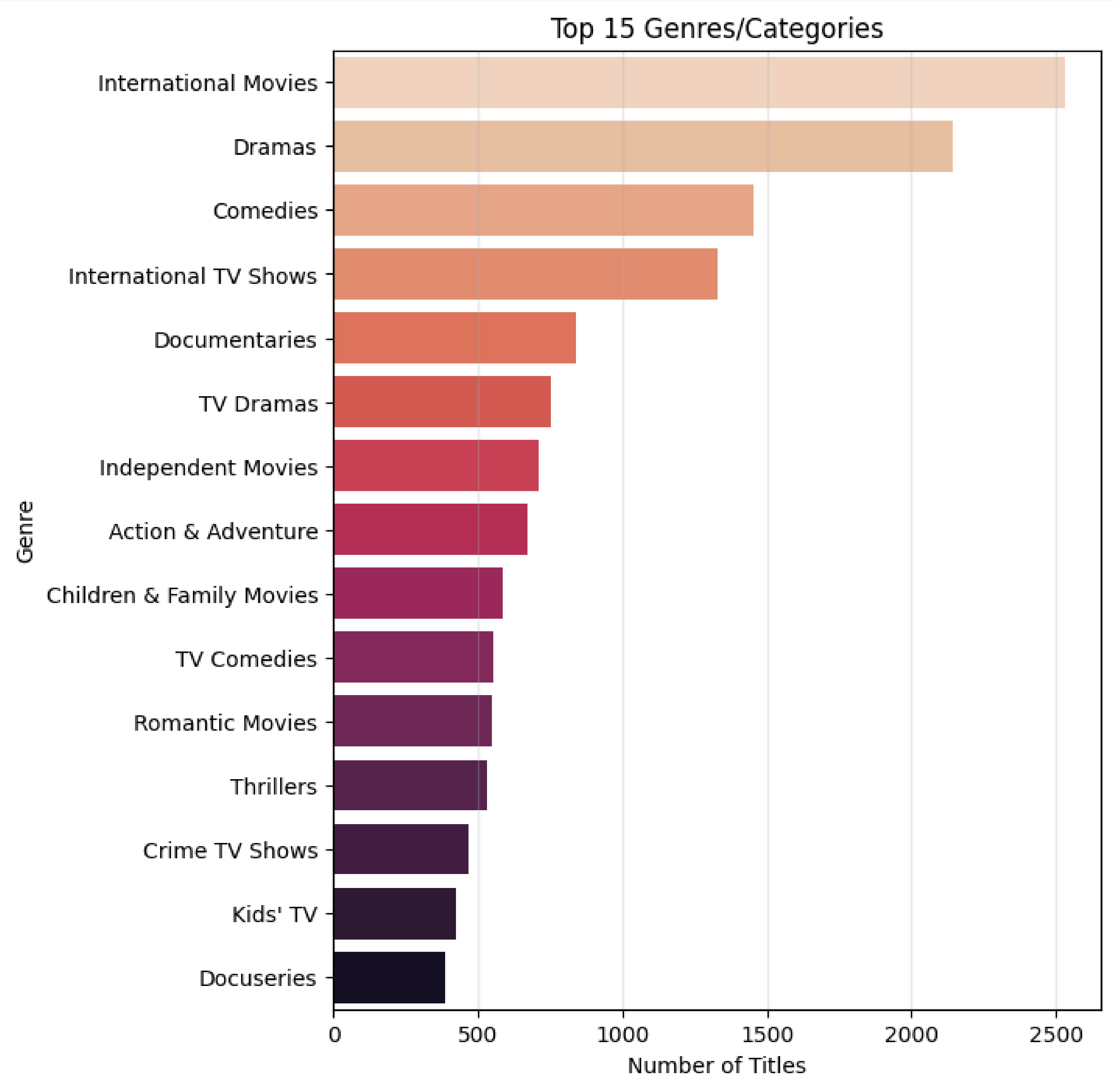
Multiple plots showing distribution of release years and trends over time

Figure 5: Duration Analysis



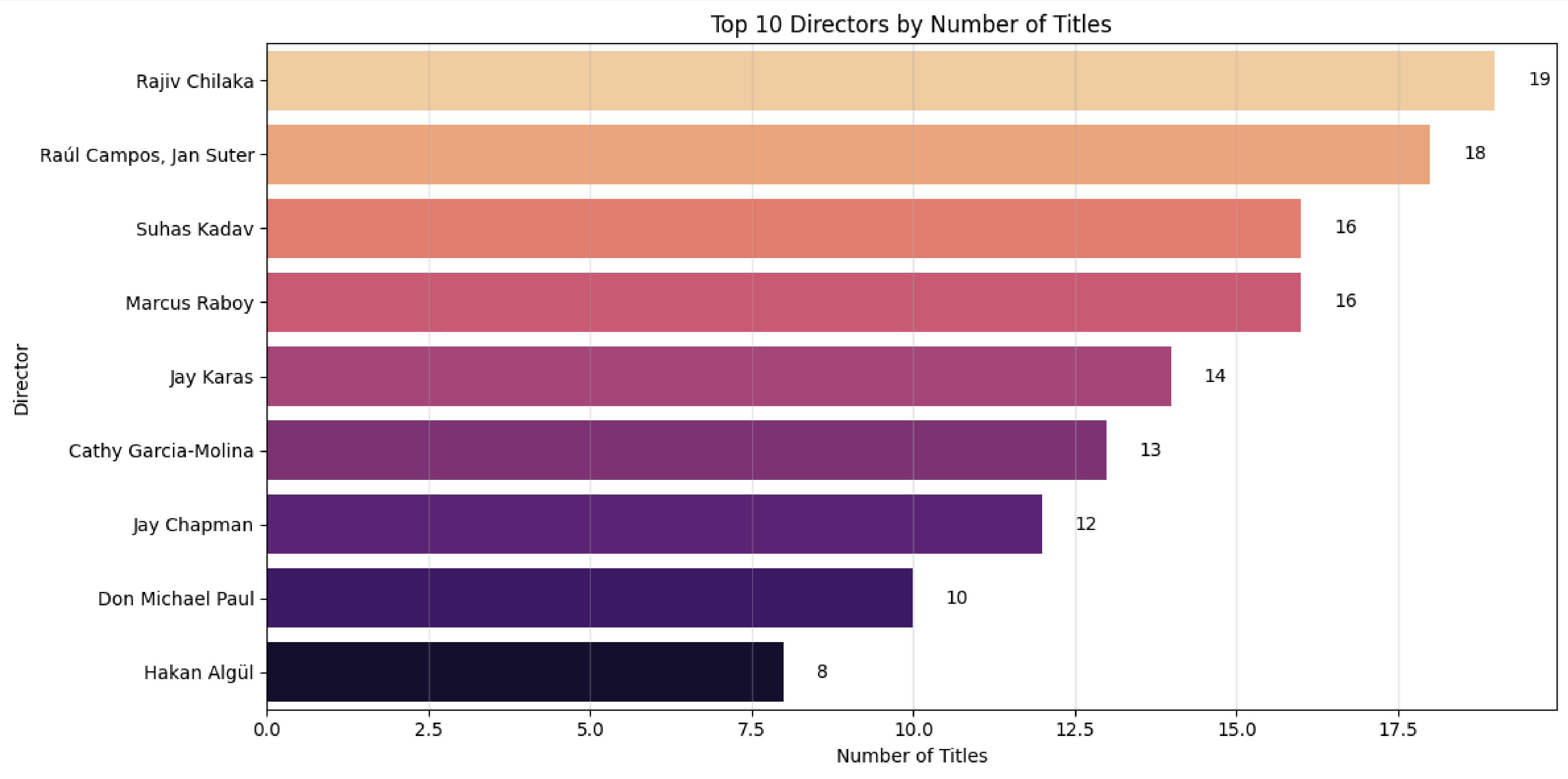
Histogram of movie durations and bar chart of TV show seasons

Figure 6: Top Genres/Categories



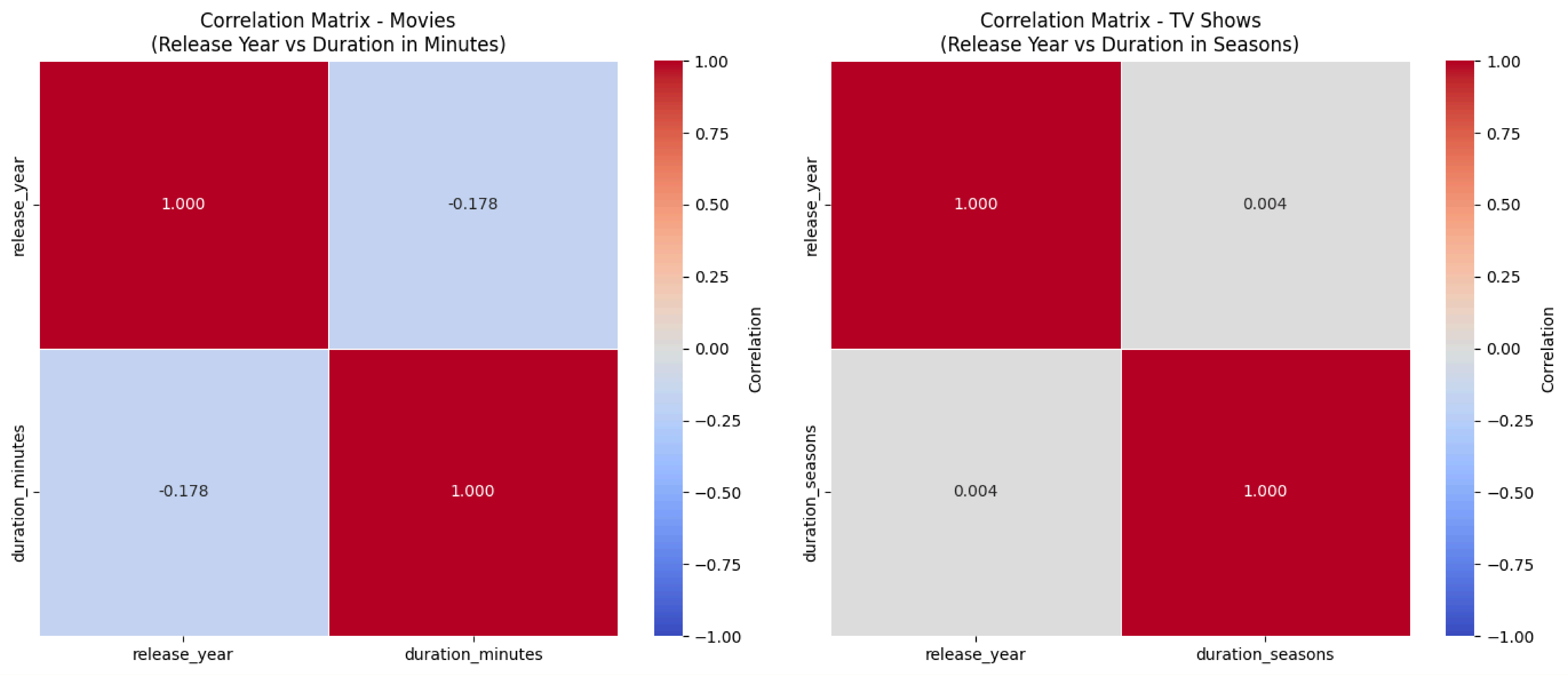
Horizontal bar chart of most common genres

Figure 7: Directo Analysis



A diverse set of directors contributes to Netflix's catalog

Figure 9: Correlation release_year and duration



There is a slight tendency for more recent movies (release_year larger) to be shorter in duration (duration_minutes smaller), but the relationship is weak

This correlation is essentially zero (negligible correlation).
This means there is no linear relationship between the release year of a TV show and how many seasons it lasts.

6. Machine Learning

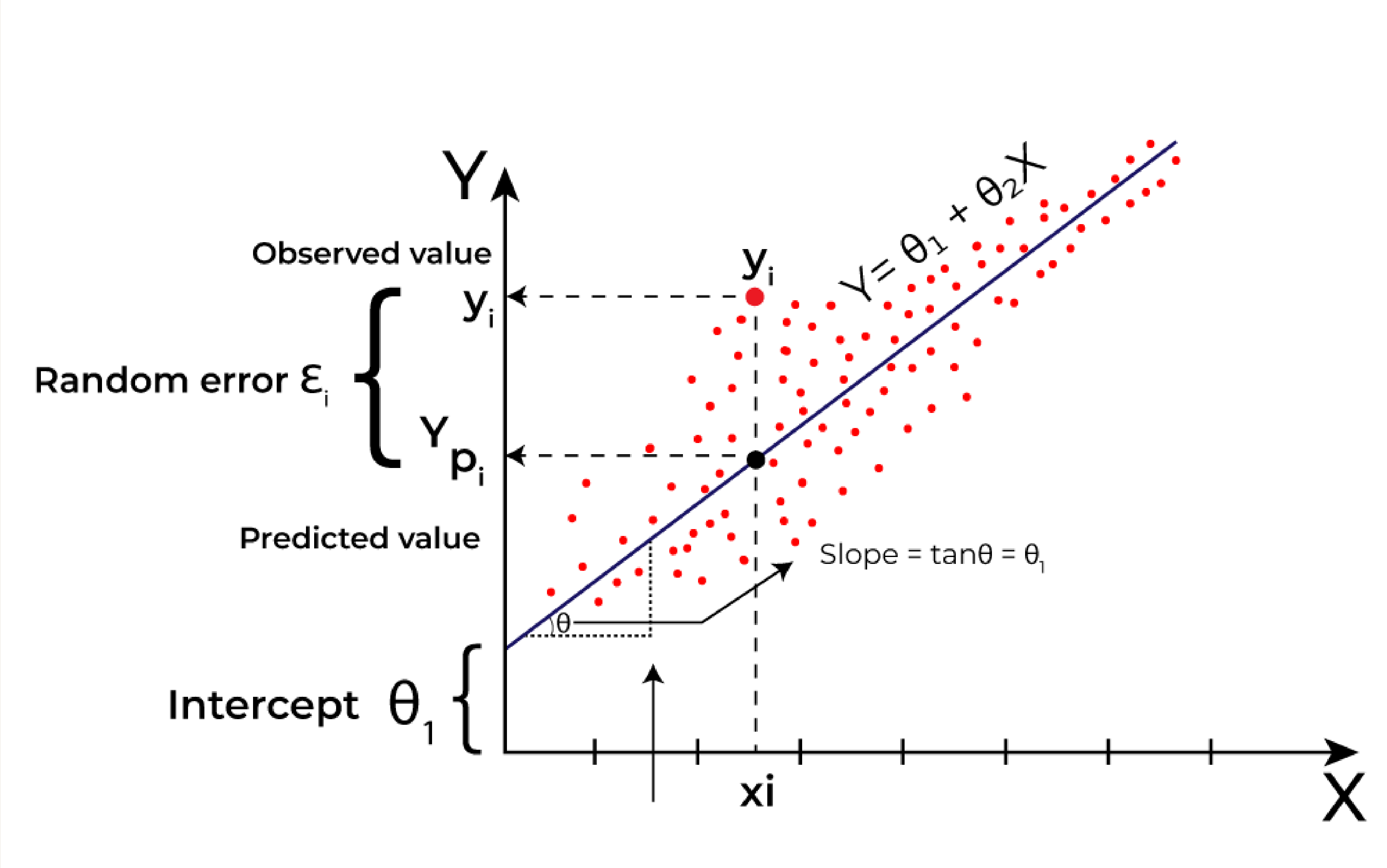
This section details the implementation and results of the predictive modeling phase, where Linear Regression was employed to understand the relationship between content metadata and duration.

6.1. Purpose of Linear Regression in Machine Learningata Preparation and Feature Engineering

- Predicting a Continuous Outcome.
- Finding Relationships Between Variables

6.2 Purpose of Evaluating Linear Regression

1. Measure Prediction Accuracy
2. Identify Model Performance Issues
3. Select the Best Model



6.3. Predict Movie durations base on selected features

```
# Prepare Movies dataset
movies_ml = movies_df.copy()
```

Creates a copy of the original movies_df dataset.

```
# Encode categorical variables
le_rating = LabelEncoder()
```

LabelEncoder converts these strings into numeric labels, e.g., "G" → 0, "PG-13" → 1, "R" → 2

Movies ML Dataset Shape: (5482, 20)

```
X_movies = movies_ml[['release_year', 'year_added', 'month_added',
                      'genre_count', 'country_count', 'rating_encoded', 'country_encoded']]
y_movies = movies_ml['duration_minutes']

```

'release_year', 'year_added', 'month_added', 'genre_count', 'country_count', 'rating_encoded', 'country_encoded are featured for predict target Varible'

duration_minutes is **target Varible**

```
X_train_m, X_test_m, y_train_m, y_test_m = train_test_split(
    X_movies, y_movies, test_size=0.2, random_state=42
)
```

Split data for training and testing

Training set size: 4385

Test set size: 1097

Train Linear Regression

```python
lr_movies = LinearRegression()
lr_movies.fit(X_train_m_scaled, y_train_m)
y_pred_m = lr_movies.predict(X_test_m_scaled)
```

Evaluate

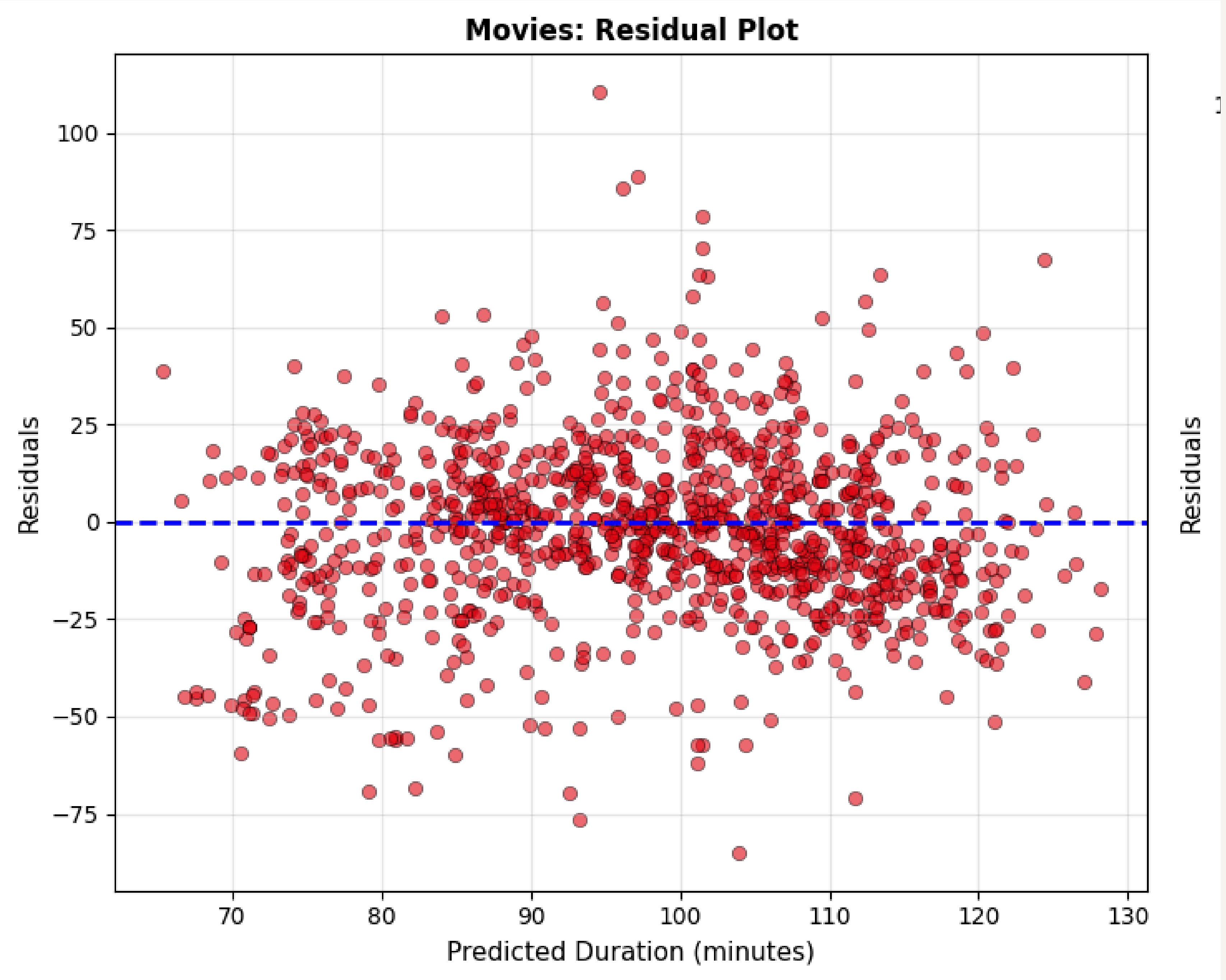
```python
mse_m = mean_squared_error(y_test_m, y_pred_m)
rmse_m = np.sqrt(mse_m)
mae_m = mean_absolute_error(y_test_m, y_pred_m)
r2_m = r2_score(y_test_m, y_pred_m)

```

```
============================================================
LINEAR REGRESSION - MOVIES DURATION PREDICTION
============================================================
Root Mean Squared Error (RMSE): 22.35 minutes
Mean Absolute Error (MAE): 16.94 minutes
R² Score: 0.2677
Accuracy: 26.77%

Feature Coefficients:
            Feature  Coefficient
3       genre_count     9.333470
5    rating_encoded    -5.854074
6   country_encoded    -3.751460
0      release_year    -3.256031
1        year_added     2.496451
2       month_added     0.612233
4     country_count    -0.450619
```

Movies: Residual Plot



The model is not very strong (R^2 = 0.2677), but it shows:

- Genre count is the strongest positive predictor of movie duration.
- Rating and country have notable negative effects.
- Predictions are off by ~17 minutes on average (MAE).

Thank You