DATA ANALYSIS ON LATEST NETFLIX TV SHOWS AND MOVIES

INTRODUCTION

This dataset offers a comprehensive overview of movies and TV shows available on a popular streaming platform, providing information about titles, directors, cast, country of origin, release year, ratings, genres, and more. With over 4,000 entries, it serves as a valuable resource to explore global entertainment trends, uncovering patterns related to movie categories, content durations, and the distribution of genres across different regions. The primary objective of this report is to analyze these factors, identifying key insights that can inform decisions about content offerings, marketing, and regional targeting for streaming platforms.

Through the analysis, we will investigate various aspects, such as the relationship between movie categories (recent, classic) and release years, the popularity of genres in different countries, and the distribution of content ratings and age suitability across movie types. Additionally, the report will delve into content durations, exploring trends in movie and TV show lengths, and examine the influence of directors and actors on the success of movies. Ultimately, this analysis aims to provide actionable insights that can help content providers enhance their libraries, better align with regional preferences, and optimize content strategies based on audience demand.

AIM

The aim of this analysis is to extract actionable insights from the dataset of movies and TV shows available on a streaming platform. By examining various attributes such as genres, ratings, durations, countries of origin, and content types, this analysis seeks to uncover patterns and trends that can inform decisions about content strategy, audience targeting, and platform optimization.

Specifically, the goal is to understand regional preferences in genres, assess the distribution of age suitability and content ratings, identify prolific directors and popular actors, and explore the relationship between content duration and

audience engagement. These findings aim to guide streaming platforms in curating their libraries, personalizing user experiences, and enhancing their competitive edge in the global entertainment market.

STEPS INVOLVED IN THE ANALYSIS

1. Data Cleaning and Preprocessing
    - Convert object data types (e.g., dates) to appropriate formats for analysis.
    - Handle missing values, duplicates, and inconsistencies to ensure data quality.
    - Split and clean multi-valued fields like cast and genres for meaningful exploration.
2. Feature Engineering
    - Create new columns like year_diff to calculate the difference between release year and addition date.
    - Categorize content into groups such as Movie_Cat (e.g., Recent, Classics) and Age_Suitability based on ratings.
    - Extract and aggregate data for more detailed insights (e.g., top genres by country, prolific directors).
3. Exploratory Data Analysis (EDA)
    - Visualize content distribution across various attributes like genres, countries, and age suitability.
    - Identify top contributors (e.g., directors, actors, and countries) using bar charts and count plots.
    - Analyze duration trends for movies and TV shows and their impact on user engagement.
    - Explore rating distributions and their alignment with content types.

4. Insight Derivation
    - Compare the popularity of genres by region and type to tailor content recommendations.
    - Identify high-performing directors, actors, and content to prioritize partnerships or acquisitions.
    - Assess the diversity of offerings across countries and audience suitability to enhance inclusivity.
5. Visualization and Reporting
    - Use visual tools like bar charts, pie charts, scatter plots, and histograms for impactful storytelling.
    - Summarize key findings with actionable recommendations for optimizing the streaming platform's content strategy.

OVERVIEW OF THE DATASET COLUMNS

The dataset comprises several key columns, each offering unique insights into the content library:

- Show_ID**:** Id of the movie/TV show
- Type: Categorizes content as either "Movie" or "TV Show."
- Title: Lists the names of the available content.
- Director : Director of the movie/TV show
- Cast: actors/actress who have acted in the movie/TV show
- Country: Identifies the region or origin of the content.
- Date Added: Indicates when the content was made available on the platform.
- Release Year: Shows the year the content was originally released.
- Rating: Specifies age suitability and content ratings like TV-MA, R, and PG-13.

- Duration: Indicates the length of movies (in minutes) or TV shows (number of seasons).
- Genres: Describes the content's categories, such as Action, Comedy, or Drama.
- Description: Summary of the movie
- Derived Features: Columns like year_diff, Movie_Cat, and Age_Suitability add analytical depth by categorizing or calculating trends.

Each column provides crucial information, enabling a comprehensive analysis of content patterns and audience preferences.

DATA OVERVIEW

type	Object
title	Object
director	Object
cast	Object
country	Object
date_added	Datetime
release_year	Int
rating	Object
duration	Int
genres	Object
year_diff	Int
Movie_Cat	Object
Age_Suitability	Object

DATA ANALYSIS

- This dataset consists of over 4,000 entries representing movies and TV shows available on a popular streaming platform.
- It contains 13 columns.

FEATURE ENGINEERING

Movie Category (Movie_Cat):

- This column categorizes movies based on the difference between the release year and the year they were added to the platform (year_diff).
- Using a conditional logic:
    - Movies with year_diff >= 10 are classified as Classics, representing older, timeless films.
    - Movies with year_diff <= 2 are labeled as Moderate, indicating recent releases.
    - All other movies fall under the Recent category, covering moderately aged films.

Insight: This feature provides a temporal perspective on the catalog, helping identify trends in content availability based on its age**.**

```
df['Movie_Cat']=df['year_diff'].apply(lambda x:'Classics' if x>=10 else 'Moderate' if x<=2 else 'Recent')
```

Age Suitability (Age_Suitability):

- This feature maps each movie's or TV show's **rating** to an age-appropriate category using predefined guidelines:
    - TV-MA, R, NC-17 → **18+** (suitable for adults only).
    - PG-13 → **13+** (suitable for teens and older).
    - TV-PG → **10+** (suitable for older children).
    - G → **All Ages** (family-friendly content).

Insight: This classification aids in understanding the content's target audience and supports creating filters for specific age groups or parental control features

```
age_map = {'TV-MA': '18+', 'R': '18+', 'PG-13': '13+', 'TV-PG': '10+', 'NC-17': '18+', 'G': 'All Ages'} df['Age_Suitability'] = df['rating'].map(age_map)
```

Year Difference (year_diff):

- This column calculates the number of years between the **release year** (release_year) and the year the content was **added to the platform** (date_added).
- This feature helps measure how quickly movies or TV shows are made available on the platform after their original release.

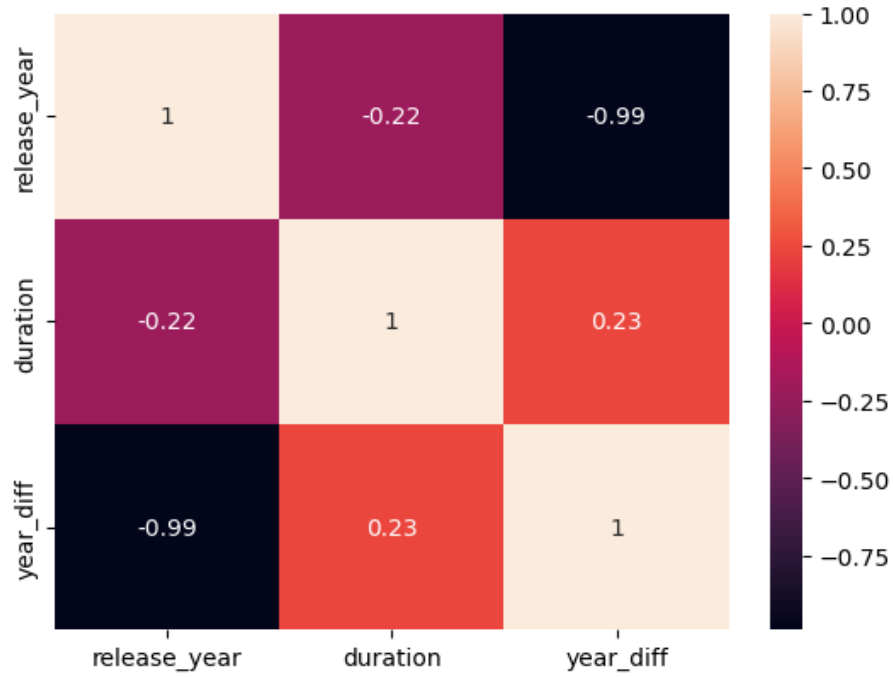
Insight: year_diff provides a lens to examine the freshness of the catalog, allowing analysis of content acquisition strategies over time.

```
df['year_diff'] = pd.to_datetime(df['date_added']).dt.year - df['release_year']
```

These additional columns enrich the dataset by providing meaningful classifications and metrics that enhance the analysis. They allow for nuanced insights into movie recency, age suitability, and platform acquisition trends, making them vital for understanding the content library's dynamics.

VISUALIZATION

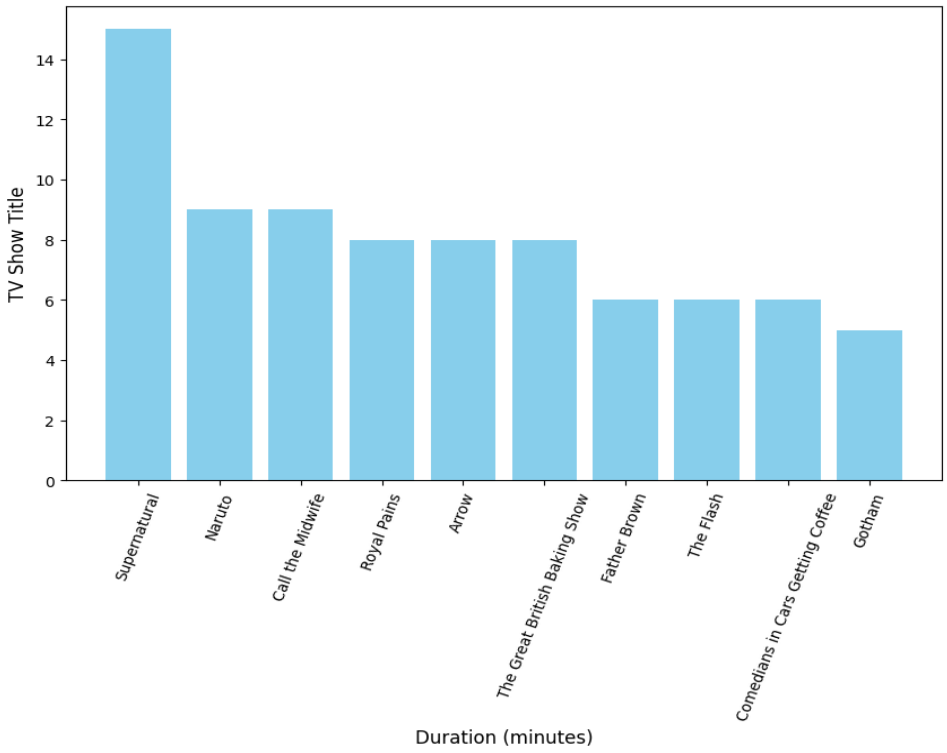
HEATMAP



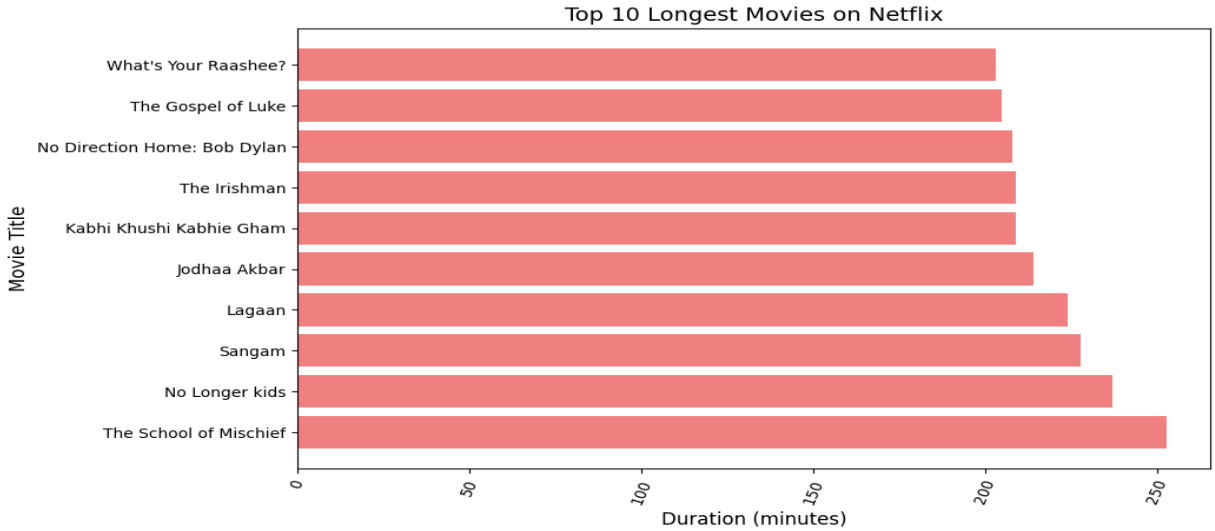
The heatmap depicts the correlation between three variables: release year, duration, and year difference. The darker the color, the stronger the

correlation. We can see that release year has a strong negative correlation with year difference, while duration has a weak positive correlation with year difference.

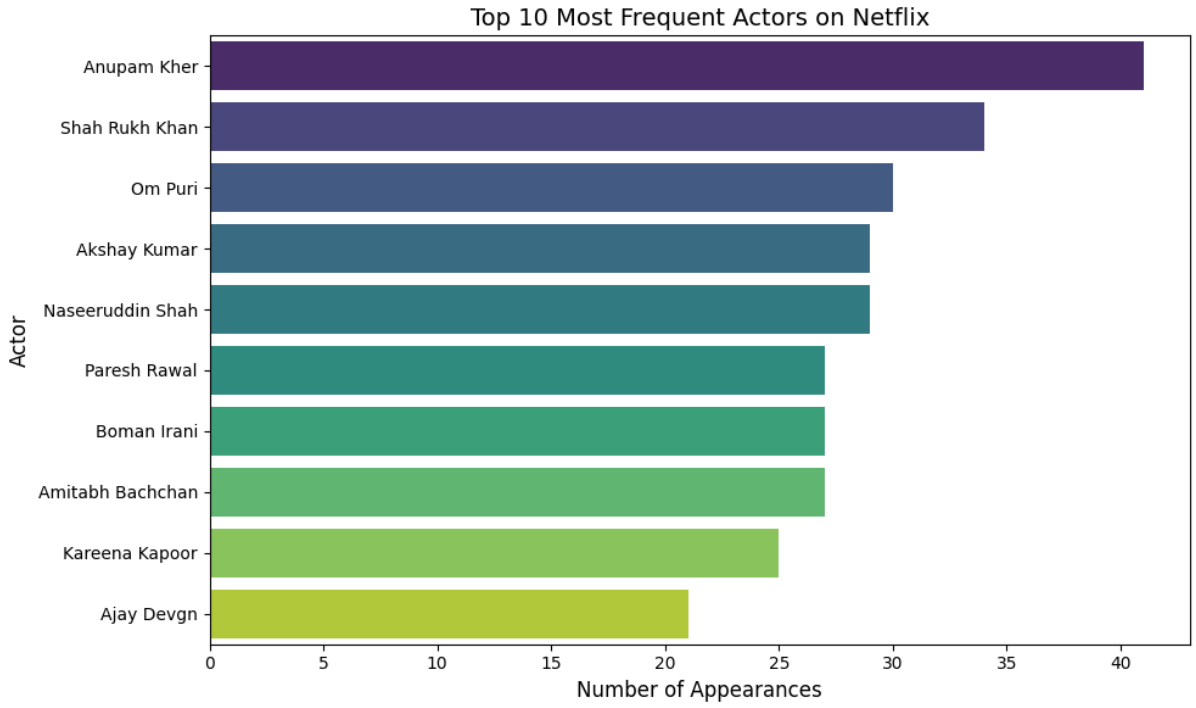
BARPLOT



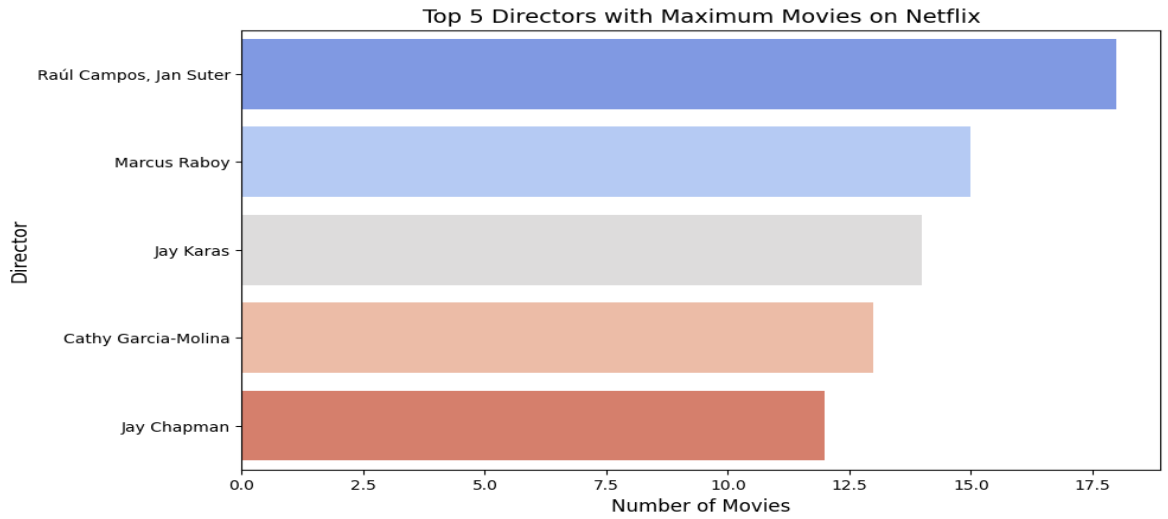
The bar chart shows the duration of 10 TV shows. Supernatural has the longest duration (14 minutes), while Gotham has the shortest (5 minutes). The majority of the shows have durations between 6 and 8 minutes.



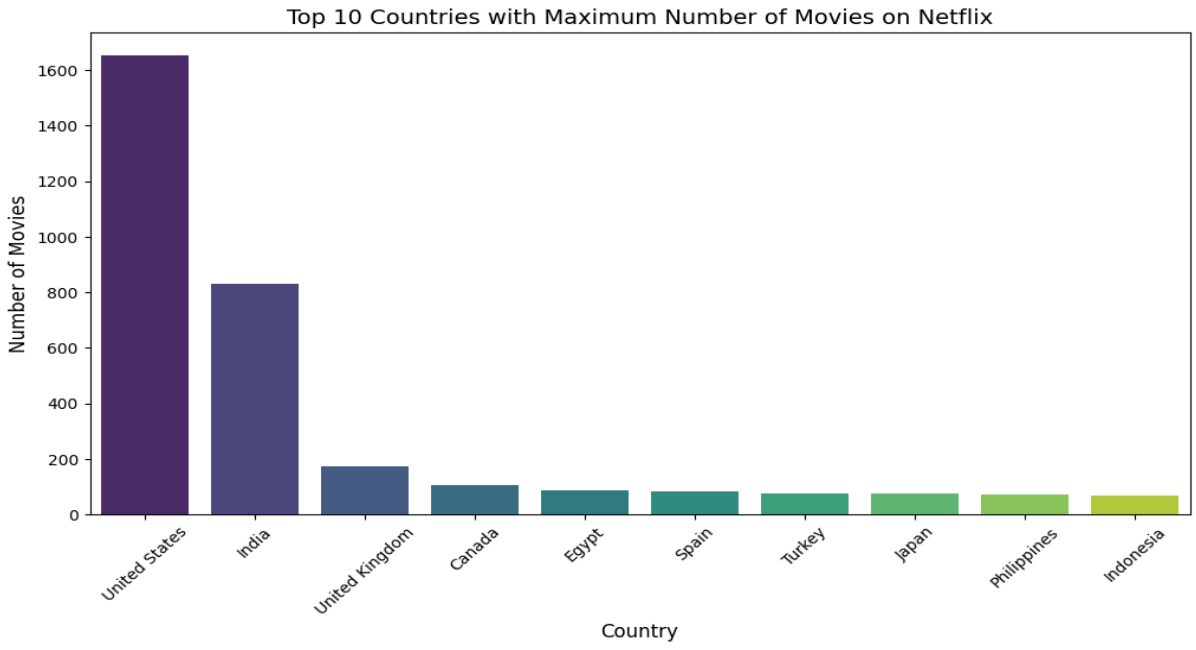
The bar chart displays the top 10 longest movies on Netflix. "The School of Mischief" is the longest movie, clocking in at over 250 minutes. The length of movies ranges from approximately 180 minutes to over 250 minutes.



The bar chart presents the top 10 most frequent actors on Netflix. Anupam Kher leads with the highest number of appearances, followed by Shah Rukh Khan. The remaining actors, including Om Puri, Akshay Kumar, and Naseeruddin Shah, also have a significant presence on the platform. The chart highlights the popularity and frequent casting of these actors in Netflix productions.

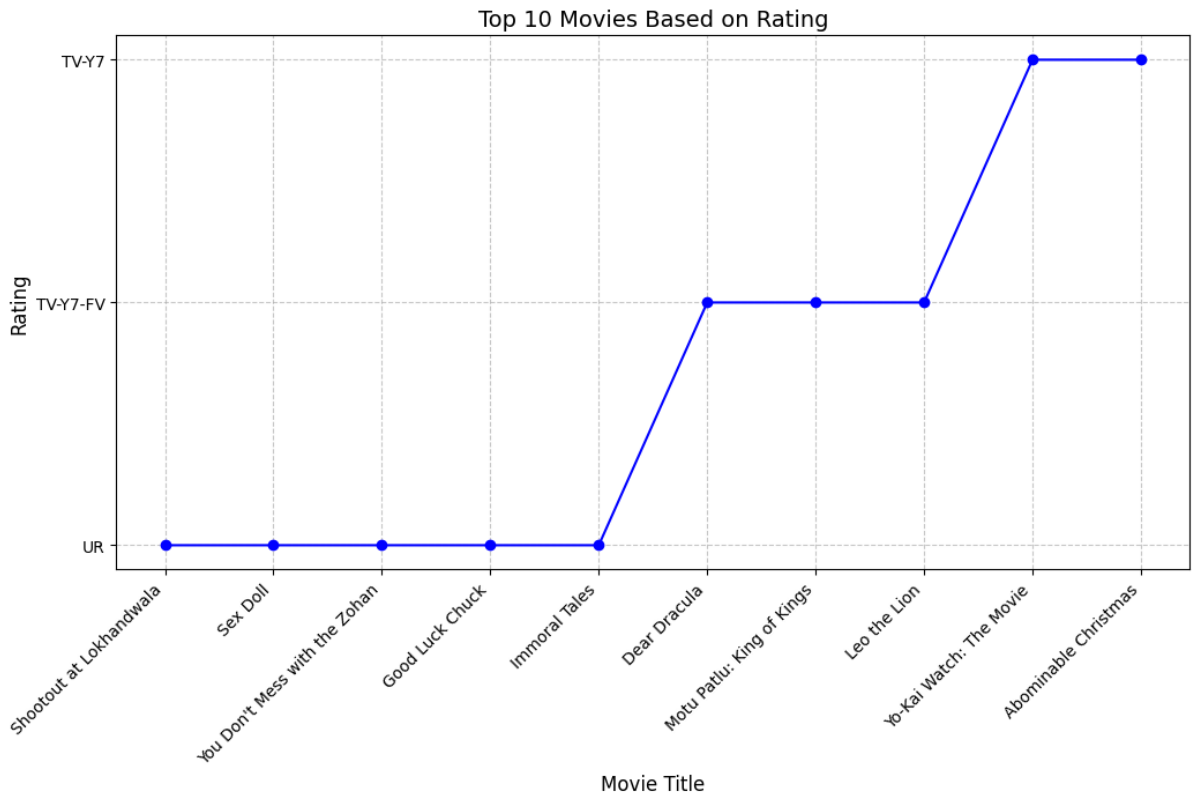


The bar chart presents the top 5 directors with the maximum number of movies on Netflix. Raul Campos, Jan Suter takes the lead with the highest number of movies, followed by Marcus Raboy and Jay Karas. The chart highlights the prolific presence of these directors in Netflix's content library.



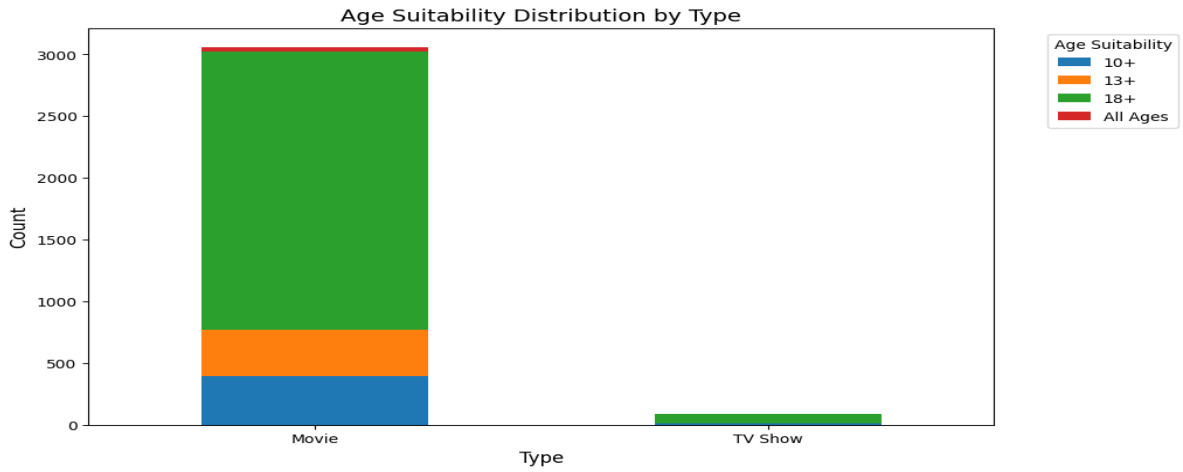
The bar chart presents the top 10 countries with the maximum number of movies on Netflix. The United States leads with the highest number of movies, followed by India and the United Kingdom. The remaining countries, including Canada, Egypt, and Spain, also have a significant presence in Netflix's movie collection.

LINEPLOT



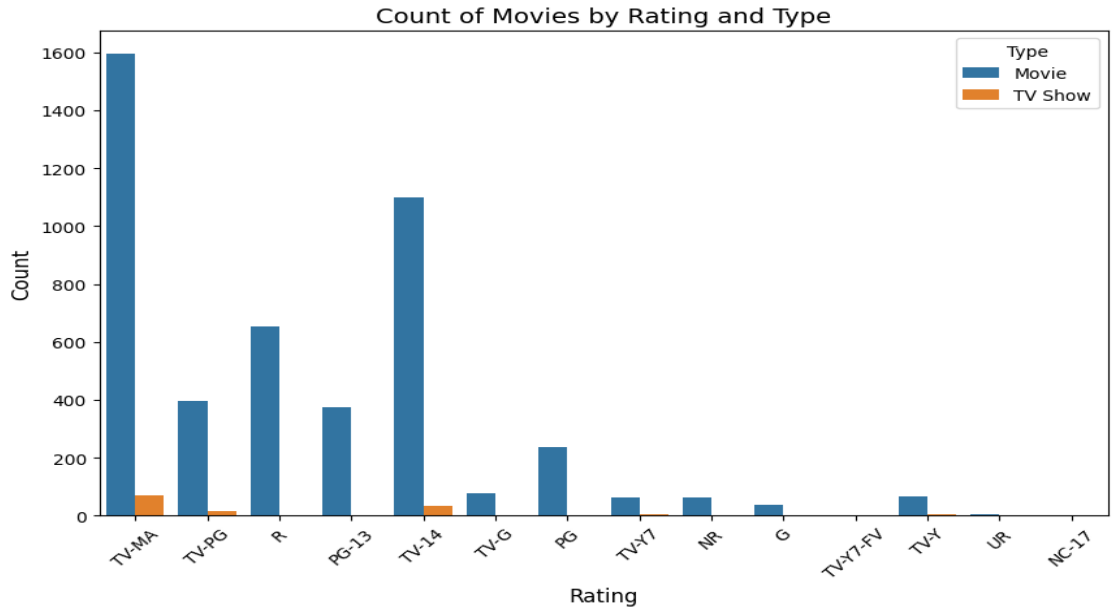
The line chart presents the top 10 movies based on their rating. The movies are arranged in ascending order of their rating, with "Shootout at Lokhandwala" having the lowest rating and "Abominable Christmas" having the highest rating. The chart reveals a significant jump in ratings between "Immoral Tales" and "Dear Dracula," suggesting a notable difference in the rating level for these movies. It also shows a plateau in ratings from "Dear Dracula" to "Yo-Kai Watch: The Movie," indicating a group of movies with similar high ratings.

HISTPLOT



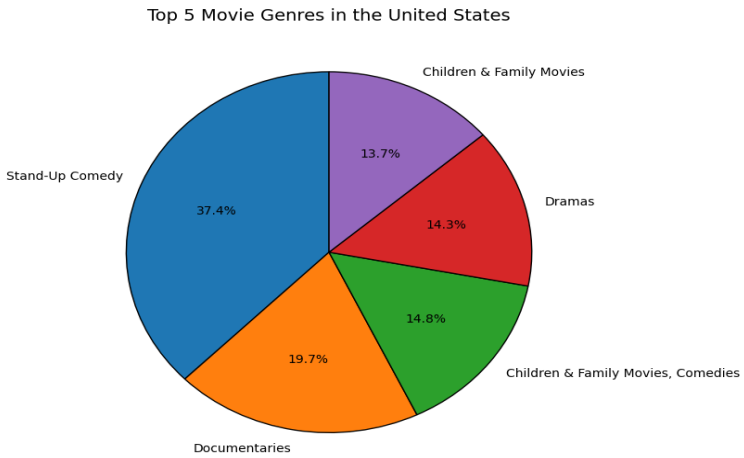
The chart shows the distribution of age suitability ratings for movies and TV shows on Netflix. Movies are more likely to be rated for all ages or 18+ compared to TV shows, which are more likely to be rated for 10+ or 13+.

COUNTPLOT

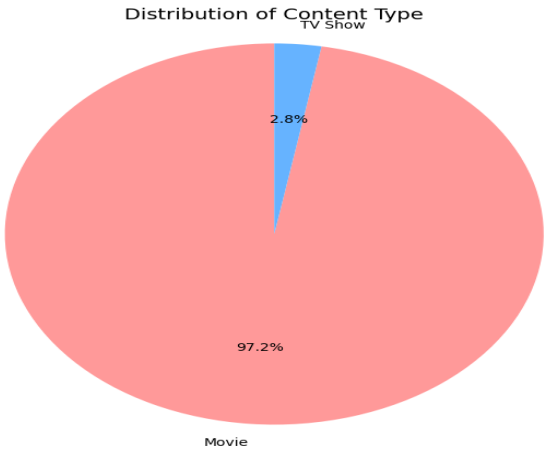


This shows the distribution of movies and TV shows on Netflix by rating. Movies are more likely to be rated TV-MA, R, and PG-13, while TV shows are more likely to be rated TV-PG, TV-14, and TV-G. This suggests that movies on Netflix tend to have more mature content compared to TV shows.

PIEPLOT

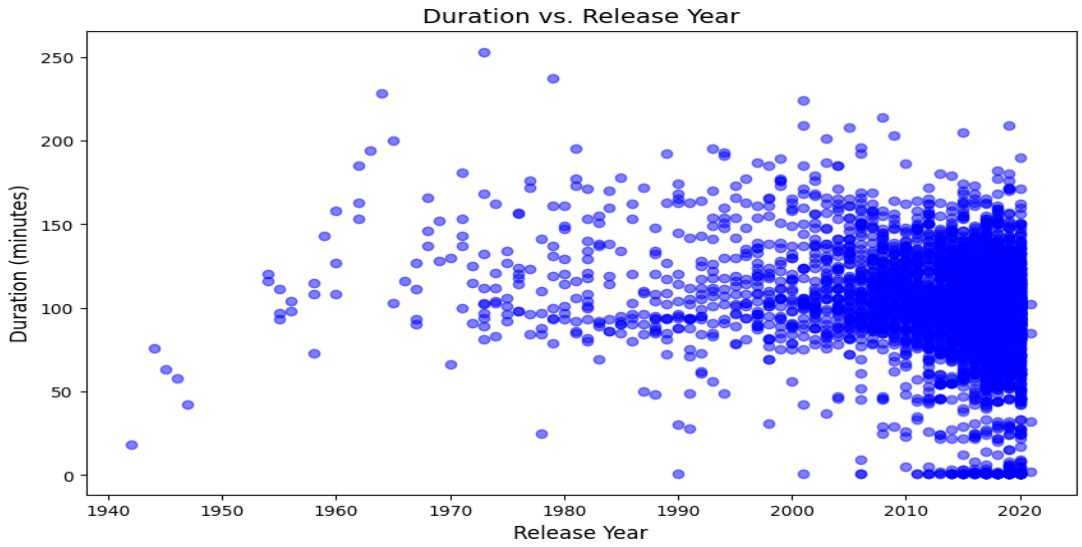


The pie chart illustrates the top 5 movie genres in the United States. Stand-Up Comedy holds the largest share with 37.4%, followed by Documentaries with 19.7%. The remaining genres, Children & Family Movies, Dramas, and Children & Family Movies, Comedies, have smaller but significant proportions in the US movie market.

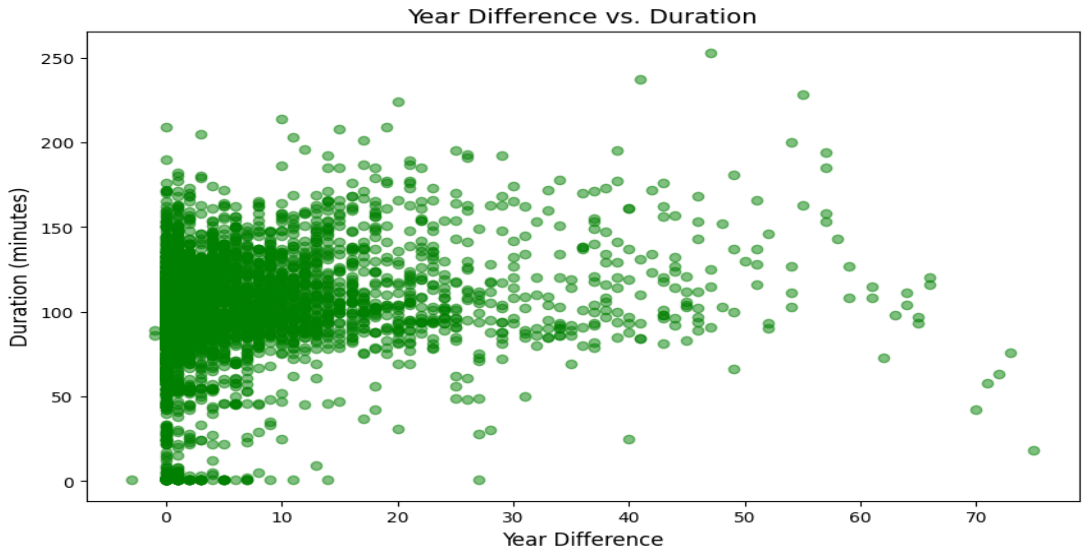


The pie chart shows that Movies make up the majority of content on Netflix, accounting for 97.2% of the total, while TV shows represent a much smaller portion at 2.8%.

SCATTERPLOT

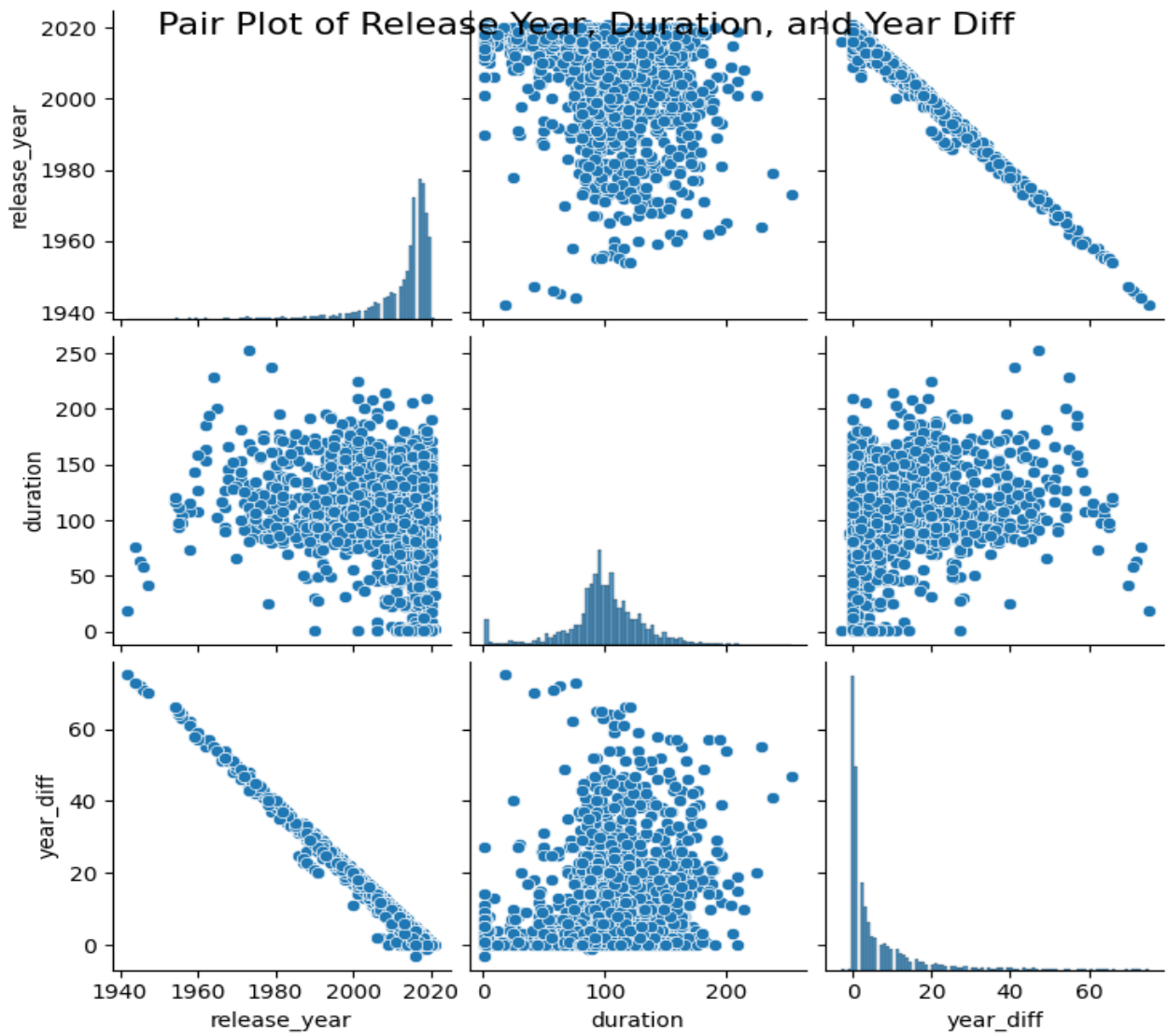


The scatter plot depicts the relationship between movie duration and release year. There is no clear trend between the two variables, indicating that movie duration has not changed significantly over time. However, there is a slight increase in the number of movies with shorter durations in recent years.



The scatter plot displays the relationship between movie duration and release year. There is no clear trend between the two variables, suggesting that movie duration has not changed significantly over time.

PAIRPLOT



The pair plot visualizes the relationship between release year, duration, and year difference. It reveals a negative correlation between release year and year difference, and a weak positive correlation between duration and year difference. The histograms on the diagonal show the distribution of each variable.

CONCLUSION

1. **Correlation Insights:**
    - Release year has a strong negative correlation with year difference, indicating that newer releases are added to the platform sooner.
    - Duration exhibits a weak positive correlation with year difference, suggesting a slight tendency for older movies to be longer.
2. **Content Duration:**
    - TV Shows: "Supernatural" has the longest duration (14 minutes), while most shows range between 6–8 minutes.
    - Movies: "The School of Mischief" is the longest movie on Netflix, exceeding 250 minutes, with other movies ranging from 180–250 minutes.
3. **Actors and Directors:**
    - Actors: Anupam Kher is the most frequently cast actor, followed by Shah Rukh Khan and Om Puri, highlighting their popularity on Netflix.
    - Directors: Raul Campos and Jan Suter are the most prolific directors on Netflix, followed closely by Marcus Raboy and Jay Karas.
4. **Geographic Trends:**
    - The United States leads with the highest number of movies on Netflix, followed by India and the United Kingdom, demonstrating strong regional content production.
    - The top 5 movie genres in the United States are dominated by Stand-Up Comedy, with Documentaries and Children & Family Movies also being significant.

5. **Content Ratings and Suitability:**
    - Movies: Tend to have more mature content, with higher occurrences of TV-MA, R, and PG-13 ratings.
    - TV Shows: Cater more to family audiences, often rated TV-PG, TV-14, or TV-G.
    - Age suitability ratings reveal that movies are more likely rated for all ages or 18+, while TV shows often target younger audiences (10+, 13+).
6. **Ratings and Trends:**
    - Movies with higher ratings exhibit a plateau, indicating consistent quality in top-rated content. However, there is a significant jump in ratings among certain titles.
7. **Content Composition:**
    - Movies constitute the majority of content (97.2%) on Netflix, with TV shows accounting for only a small share (2.8%).
8. **Genre Preferences:**
    - Stand-Up Comedy leads as the most popular genre in the United States, followed by Documentaries and Children & Family Movies.
9. **Duration and Release Year Trends:**
    - The scatter plot reveals no significant change in movie durations over time, although shorter durations are more frequent in recent years.
10. **Pair Plot Observations:**
    - A negative correlation exists between release year and year difference, reaffirming that newer releases are added sooner.
    - A weak positive correlation is observed between duration and year difference, with older content tending to have slightly longer durations.