Netflix Data Analysis

Goal: This report aims to analyze the available Netflix dataset, which contains detailed information on various shows and movies, including their type, release year, duration, rating, cast, and country of production. By leveraging this data, we seek to uncover insights and trends that can help Netflix optimize its content production strategy, improve customer satisfaction, and identify opportunities for growth.

Approach: Data-driven insights using a clean, structured analysis process.

1. Data Cleaning And Exploration

1.1. Data Exploration

Code:

```
df.info()
df.describe()
```

Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

Insights:

1. The dataset consists of 8807 rows (entries) and 12 columns (features).
2. There are various data types including object and int64. 'object' represents string data types and 'int64' represents integer data types.
3. 'release_year' is the only numerical column in the dataset.

1.2. Checking for missing Values in dataset.

Code:

```
import pandas as pd

df = pd.read_csv('netflix.csv')
df.isnull().sum()
```

Output:

```
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
description      0
dtype: int64
```

Insights:
Director: A significant number of rows (2634) are missing director information.
Cast: A smaller number of rows (825) are missing cast information.
Country: There are 831 rows missing country information.
Date_added: 10 rows are missing the date added to Netflix.
Rating: 4 rows are missing the rating.
Duration: 3 rows are missing the duration.

1.3. Handling the missing values in datset.

```
df.loc[:, 'director'] = df['director'].fillna('Unknown')
df.loc[:, 'cast'] = df['cast'].fillna('Unknown')
df.loc[:, 'country'] = df['country'].fillna('Unknown')
df.loc[:, 'date_added'] = df['date_added'].fillna(df['date_added'].mode()[0])
df.head()
```

Output:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	Unknown	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	Unknown	Unknown	Unknown	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	Unknown	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Insights:

These replaces missing values in the 'director', 'cast', and 'country' columns with the string 'Unknown' and fills missing values in the 'date_added' column with the most frequent value (mode) in that column. Lastly, it displays the first 5 rows of the cleaned DataFrame.

2. Non Graphical Analysis

1.Value Counts & Unique Attributes:

Code:

```
print(df['type'].value_counts())
print(df['country'].value_counts())
print(df['listed_in'].value_counts())
```

Output:

```
type
Movie      6131
TV Show    2676
Name: count, dtype: int64
country
United States                                2818
India                                         972
Unknown                                       831
United Kingdom                                419
Japan                                         245
                                             ...
Romania, Bulgaria, Hungary                      1
Uruguay, Guatemala                              1
France, Senegal, Belgium                        1
Mexico, United States, Spain, Colombia          1
United Arab Emirates, Jordan                    1
Name: count, Length: 749, dtype: int64
listed_in
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
                                                     ...
Kids' TV, TV Action & Adventure, TV Dramas              1
TV Comedies, TV Dramas, TV Horror                       1
Children & Family Movies, Comedies, LGBTQ Movies        1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows      1
Cult Movies, Dramas, Thrillers                          1
Name: count, Length: 514, dtype: int64
```

Insights:

1.Type Distribution:

Movies dominate the dataset with 6,131 entries, while **TV Shows** have 2,676 entries. This suggests a greater focus on movies compared to TV shows in the dataset.

2. Country Distribution:

The **United States** is the most represented country with 2,818 entries, indicating that a significant portion of the content is from the US.

India follows with 972 entries, suggesting a considerable amount of content from India.

There are many entries with multiple countries listed (e.g., "Romania, Bulgaria, Hungary"), and some with "Unknown" country information (831 entries), which may need further investigation or data cleaning.

2. Category Distribution:
Dramas and **International Movies** are the most common genres listed, with **362** and **359** entries, respectively.
Stand-Up Comedy and a combination of **Comedies, Dramas, and International Movies** are also prominent, indicating a variety of content types with a heavy presence of drama and comedy.
There are many unique combinations of genres, with **514** distinct categories, suggesting a diverse range of content types available.

Recommendations:
Focus on US and Indian content: Since these countries have the highest representation, consider analyzing viewer preferences or trends specific to these regions.
Investigate 'Unknown' country entries: Addressing the "Unknown" entries could improve the dataset's completeness and accuracy.
Explore genre popularity: Given the diversity in genre combinations, analyzing viewer preferences for specific genres or combinations might provide insights into content trends and potential gaps.

2. Top 10 countries for movies.

Code:

```
# Filtering only movies and grouping by country
top_10_movie_countries = df[df['type'] == 'Movie'].groupby('country')['title'].nunique().nlargest(10)
print(top_10_movie_countries)
```

Output:

```
country
United States     2058
India              893
United Kingdom     206
Canada             122
Spain               97
Egypt               92
Nigeria             86
Indonesia           77
Japan               76
Turkey              76
Name: title, dtype: int64
```

Insights:
United States Dominance: The United States leads by a substantial margin with **2058** movie titles on Netflix, showing that a significant portion of Netflix's movie library is U.S.-produced.
India's Strong Presence: India ranks second with **893** movies, reflecting a strong focus on Indian content. This suggests Netflix is catering to the large Indian audience, which is a growing market for the platform.
Western-Centric Content: Countries like the **United Kingdom**, **Canada**, and **Spain** also feature in the top 10, showing Netflix's preference for content from Western countries.

Recommendations:
Should continue to **diversify its content portfolio** by expanding its investment in emerging markets such as **Nigeria, Egypt, and Indonesia**. These regions are showing significant growth in content production and demand, and by increasing local content production, Netflix can strengthen its presence and appeal to a broader global audience.

3. Top 10 countries for TV Shows.

Code:

```
# Filtering only TV shows and grouping by country
top_10_tv_countries = df[df['type'] == 'TV Show'].groupby('country')['title'].nunique().nlargest(10)
print(top_10_tv_countries)
```

Output:

```
country
United States     760
United Kingdom    213
Japan             169
South Korea       158
India              79
Taiwan             68
Canada             59
France             49
Australia          48
Spain              48
Name: title, dtype: int64
```

Insights:

1.United States Dominance: The United States leads the list again with **760 TV shows**, indicating that a significant portion of Netflix's TV show library originates from the U.S., reflecting its strong production capabilities and global demand for U.S. content.
2.United Kingdom's Strong Presence: The United Kingdom ranks second with **213 TV shows**, showing Netflix's focus on British content, which is popular both domestically and internationally.
3.Asian Content Growth: Japan (**169 shows**) and **South Korea** (**158 shows**) have a strong representation in Netflix's TV show library, reflecting the global rise in popularity of anime, K-dramas, and other Asian content.
4.India's Moderate Contribution: India, while a major player in the movie segment, has a smaller contribution with **79 TV shows**. This suggests that Netflix has more focus on movies than TV shows in the Indian market.
5.Emerging Markets and Global Content: Taiwan, **Canada**, **France**, **Australia**, and **Spain** all contribute to Netflix's diverse catalog of TV shows, indicating Netflix's efforts to source and distribute content from various global regions.

Recommendations:

Netflix should continue to **expand its TV show content from Asian markets**, especially in countries like **Japan** and **South Korea**, as these regions are showing strong performance and global appeal, particularly in genres like anime and K-dramas. Investing more in **original content** from these markets can further strengthen Netflix's foothold in Asia and attract a broader international audience.

Additionally, Netflix can **increase its focus on India for TV shows**, given the strong demand for streaming content in the country. A strategy of boosting **original series and local content** in India could replicate its success with movies and tap into the growing viewer base for serialized content. Similarly, nurturing content in emerging markets like **Taiwan** and **Spain** will help Netflix diversify its global offerings and cater to various audience segments worldwide.

4. Top 10 Directors.

Code:

```
# Group by director and count unique titles
top_10_directors = df.groupby('director')['title'].nunique().nlargest(10)
print(top_10_directors)
```

Output:

```
director
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Marcus Raboy              16
Suhas Kadav               16
Jay Karas                 14
Cathy Garcia-Molina       13
Jay Chapman               12
Martin Scorsese           12
Youssef Chahine           12
Steven Spielberg          11
Name: title, dtype: int64
```

Insights:

Rajiv Chilaka leads with 19 titles, mainly from his popular Indian animated series *Chhota Bheem*. Raúl Campos and Jan Suter follow closely with 18 titles, known for their Latin American documentaries and specials. Marcus Raboy and Suhas Kadav, each with 16 titles, contribute to stand-up comedy and Indian animation, respectively. Jay Karas (14 titles) spans various genres, while Cathy Garcia-Molina (13 titles) focuses on Filipino romantic comedies. Legendary directors Martin Scorsese (12 titles) and Steven Spielberg (11 titles) further elevate Netflix's content. This list showcases Netflix's collaboration with directors from diverse regions and genres, ensuring a wide range of content for global audiences.

Recommendations:

Netflix should continue to collaborate with **both local and international directors** to maintain its wide range of content. Expanding relationships with directors like **Rajiv Chilaka and Suhas Kadav** can help Netflix strengthen its offering of animated content, particularly for younger audiences. Simultaneously, collaborating more with directors like **Martin Scorsese and Steven Spielberg** will enhance its appeal to viewers who enjoy critically acclaimed and high-quality films.

4. Top 10 Actors.

Code:

```
# Spliting the cast column and explode to count individual actors
df['cast'] = df['cast'].str.split(', ')
df_exploded = df.explode('cast')

# Group by cast and count unique titles
top_10_actors = df_exploded.groupby('cast')['title'].nunique().nlargest(10)
print(top_10_actors)
```

Output:

```
cast
Anupam Kher          43
Shah Rukh Khan       35
Julie Tejwani        33
Naseeruddin Shah     32
Takahiro Sakurai     32
Rupa Bhimani         31
Akshay Kumar         30
Om Puri              30
Yuki Kaji            29
Amitabh Bachchan     28
Name: title, dtype: int64
```

Insights:

Anupam Kher Leads with 43 Titles: His extensive presence in Netflix's library showcases his prominence in both Bollywood and international films.

Bollywood Dominance: Shah Rukh Khan, Naseeruddin Shah, Akshay Kumar, Om Puri, and Amitabh Bachchan highlight the strong demand for Bollywood content, popular among global audiences, especially the Indian diaspora.

Voice Actors in Animation: Julie Tejwani and Rupa Bhimani, with 33 and 31 titles respectively, reflect Netflix's focus on animated content for younger viewers.

International Appeal: Japanese voice actors Takahiro Sakurai and Yuki Kaji signify Netflix's growing influence in the anime market, catering to global anime enthusiasts.

Recommendations:

Strengthen Bollywood Partnerships: Continue investing in high-quality Indian content with stars like Anupam Kher, Shah Rukh Khan, and Akshay Kumar to attract global audiences.

Capitalize on Animation and Anime: Expand the animation and anime library by collaborating with prominent **voice actors like Julie Tejwani and Takahiro Sakurai.**

Region-Specific Content: Focus on expanding Bollywood for India and anime for Japan to boost regional appeal in high-growth markets.

Actor-Based Marketing: Create targeted campaigns highlighting popular actors' extensive work to increase viewership.

Explore New Genres: Launch cross-genre projects blending Bollywood and anime to attract diverse global audiences.

5. Best Time to Launch TV Shows and Movies.

Code:

```
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
# rows with invalid dates (optional)
invalid_dates = df[df['date_added'].isna()]
print(f"Invalid dates: {len(invalid_dates)}")
# 'week_added' column for analysis
df['week_added'] = df['date_added'].dt.isocalendar().week
# Group by week_added and count titles for TV shows and movies
best_week_tv = df[df['type'] == 'TV Show'].groupby('week_added')['title'].count().idxmax()
best_week_movie = df[df['type'] == 'Movie'].groupby('week_added')['title'].count().idxmax()

print(f"Best week for TV shows: {best_week_tv}")
print(f"Best week for movies: {best_week_movie}")
```

Output:

```
Invalid dates: 98
Best week for TV shows: 27
Best week for movies: 1
```

Insights:

Invalid dates: 98: There are 98 rows where the date_added could not be converted into a valid date format. These rows were excluded from the analysis, which is a normal outcome when dealing with real-world data.
Best week for TV shows: 27: This means that the 27th week of the year is the best week for releasing TV shows, based on the count of TV shows added in that week.
Best week for movies: 1: This indicates that the 1st week of the year (first week of January) is the best week for releasing movies.

Recommendations:

The data shows that there were **98 invalid dates**, which is quite a large number. These dates couldn't be processed, potentially meaning they were either missing or incorrectly formatted in the dataset. It raises questions about the quality of the data itself. Are these invalid entries important releases that were lost in the analysis? There's something unsettling about data being ignored, no matter how "invalid" it seems.

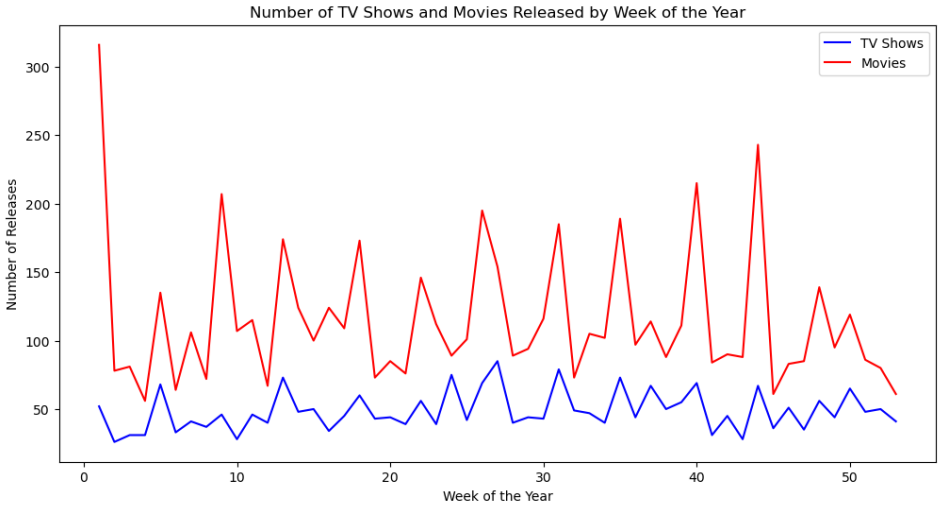
As for the release timing, the **27th week** emerges as the best for TV shows. This lands in late June or early July, a time when summer viewing might peak. But why is that? Do people really crave TV shows as summer heat bears down, or is it something else? Meanwhile, movies seem to shine in the **first week of January**—a time when everyone is either making resolutions or fighting post-holiday blues. It's a curious thing to note, almost as if the beginning of a new year calls for escapism.

6. Distribution of releases by week. (Better visual understanding of the distribution of releases throughout the year by week.)

Code:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Group by week and count TV shows and movies
tv_shows_by_week = df[df['type'] == 'TV Show'].groupby('week_added')['title'].count()
movies_by_week = df[df['type'] == 'Movie'].groupby('week_added')['title'].count()

plt.figure(figsize=(12, 6))

# Plot TV shows count by week
sns.lineplot(x=tv_shows_by_week.index, y=tv_shows_by_week.values, label='TV Shows', color='blue')

# Plot movies count by week
sns.lineplot(x=movies_by_week.index, y=movies_by_week.values, label='Movies', color='red')

plt.xlabel('Week of the Year')
plt.ylabel('Number of Releases')
plt.title('Number of TV Shows and Movies Released by Week of the Year')
plt.legend()
plt.show()
```

Output:



Insights:

1. Seasonal Patterns: Both TV shows and movies exhibit seasonal patterns in their release schedules.
2. Peak in December: The number of both movies and TV shows released peaks in the last week of December, likely due to the holiday season and increased viewership during that time.
3. Consistent Releases: Throughout the year, there are consistent releases of both movies and TV shows, with some fluctuations.
4. Movie Dominance: In most weeks, the number of movies released exceeds the number of TV shows, indicating a stronger focus on movies.

Recommendations:

Leverage Seasonal Trends: Consider releasing more content during peak viewing periods, especially around the holidays, to maximize viewership.

Diversify Release Schedule: While seasonal patterns exist, explore opportunities to release content throughout the year to maintain a steady stream of new offerings.

Balance Movie and TV Show Releases: While movies dominate, aim for a more balanced release schedule to cater to diverse audience preferences.

7. Best Month to Release TV Shows/Movies

Code:

```
# Converting date_added to datetime, handling mixed formats
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')

# Checkiing for rows with invalid dates (optional)
invalid_dates = df[df['date_added'].isna()]
print(f"Invalid dates: {len(invalid_dates)}")

# Createing 'month_added' column for analysis
df['month_added'] = df['date_added'].dt.month

# Group by month_added and count titles for TV shows and movies
best_month_tv = df[df['type'] == 'TV Show'].groupby('month_added')['title'].count().idxmax()
best_month_movie = df[df['type'] == 'Movie'].groupby('month_added')['title'].count().idxmax()

print(f"Best month for TV shows: {best_month_tv}")
print(f"Best month for movies: {best_month_movie}")
```

Output:

```
Invalid dates: 98
Best month for TV shows: 7.0
Best month for movies: 7.0
```

Insights:

1. July is a Popular Release Month: Netflix can use this information to plan their content releases and maximize viewership during July.

2. Seasonal Trends: The popularity of July suggests that there may be seasonal trends in viewer behavior and preferences. Netflix can further investigate these trends to optimize their content strategy.

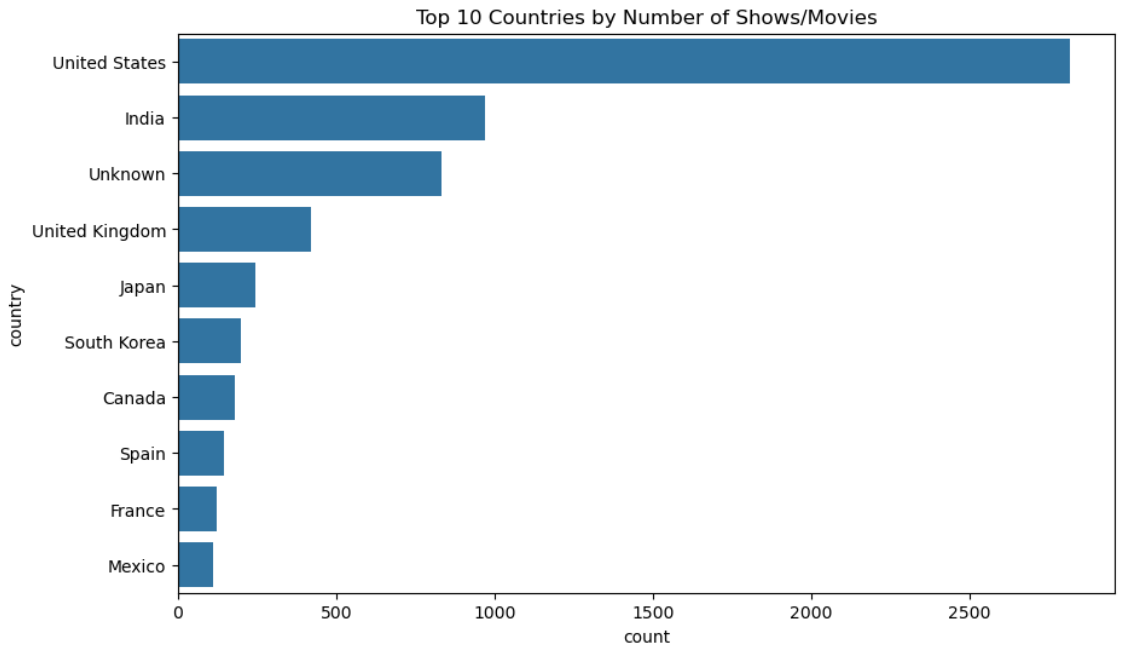
3. Competitive Analysis: By understanding the best month for releases, Netflix can monitor and analyze the release strategies of competitors to identify opportunities and potential challenges.

3. Visual Exploration

1.What type of content is available in different countries?
Code:

```
plt.figure(figsize=(10, 6))
sns.countplot(y='country', data=df, order=df['country'].value_counts().head(10).index)
plt.title('Top 10 Countries by Number of Shows/Movies')
plt.show()
```

Output:



Insights:

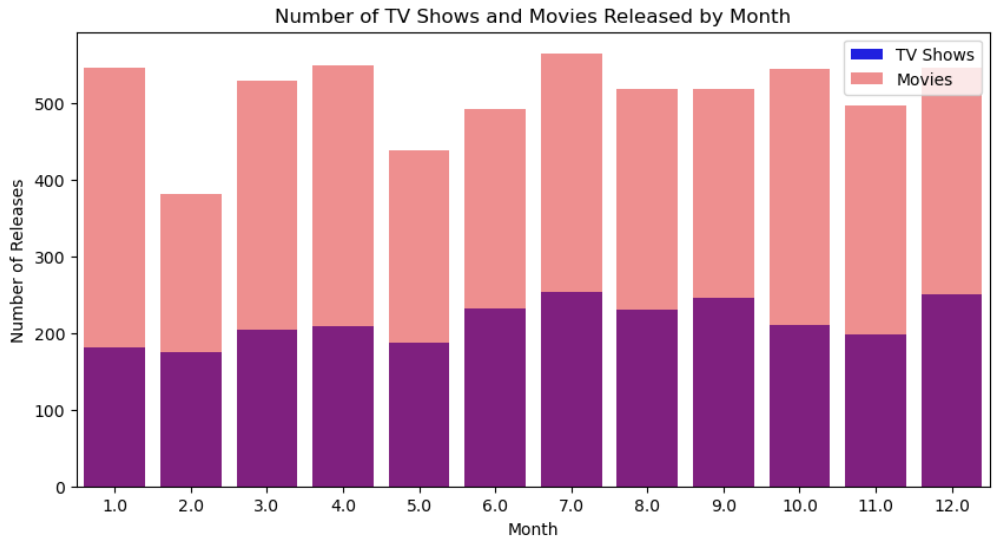
United States and India: These two countries have a significantly higher number of shows/movies compared to others, indicating a strong focus on content from these regions.
Western Countries: The majority of the top 10 countries are Western, suggesting a bias towards content from these regions.
Emerging Markets: While India is represented, there may be opportunities to expand into other emerging markets with growing subscriber bases.

2. Number of releases for each month:

Code:

```
import seaborn as sns
import matplotlib.pyplot as plt
# Plot TV shows count by month
tv_shows_by_month = df[df['type'] == 'TV Show'].groupby('month_added')['title'].count()
movies_by_month = df[df['type'] == 'Movie'].groupby('month_added')['title'].count()
plt.figure(figsize=(10, 5))
# Plot TV shows
sns.barplot(x=tv_shows_by_month.index, y=tv_shows_by_month.values, color='blue', label='TV Shows')
# Plot movies
sns.barplot(x=movies_by_month.index, y=movies_by_month.values, color='red', alpha=0.5, label='Movies')
plt.xlabel('Month')
plt.ylabel('Number of Releases')
plt.title('Number of TV Shows and Movies Released by Month')
plt.legend()
plt.show()
```

Output:



Insights:

Seasonal Trends: There seems to be a seasonal pattern in the number of releases, with a peak in July for both TV shows and movies.

Movie Dominance: Overall, there are more movies released compared to TV shows throughout the year.

Recommendations:

Optimize Release Strategy: Leverage the knowledge of peak release months (July in this case) to strategically schedule content releases and potentially maximize viewership during these periods.

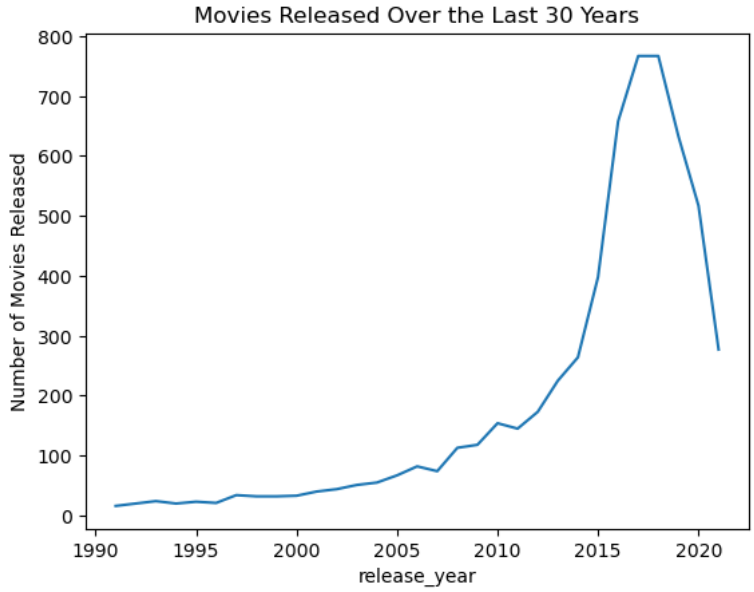
Content Diversification: While July is a strong month for releases, consider exploring ways to distribute content more evenly throughout the year to maintain audience engagement.

3. How has the number of movies released per year changed over the last 20-30 years?

Code:

```
releases_per_year = df[df['type'] == 'Movie'].groupby('release_year').size()
releases_per_year_last_30 = releases_per_year[releases_per_year.index >= (2021 - 30)]
releases_per_year_last_30.plot(kind='line', title='Movies Released Over the Last 30 Years')
plt.ylabel('Number of Movies Released')
plt.show()
```

Output:



Insights:

Steady Growth: The number of movies released per year has generally increased over the past 30 years, with a significant spike in the mid-2010s.

Recent Plateau: However, the number of releases seems to have plateaued in recent years, possibly due to factors like the COVID-19 pandemic or changes in production and distribution models.

Market Dynamics: The increasing number of movies released in recent years suggests a highly competitive market with a growing demand for content.

Mid-2010s Surge: The most notable growth occurred between 2010 and 2015, potentially driven by factors like the rise of streaming services, increased global production, and the popularity of franchises and sequels.

Recommendations:

Exclusive Content: Continue to invest in high-quality original content that is exclusive to Netflix.

Regional Focus: Produce content tailored to specific regions to expand market reach and cater to local tastes.

Diverse Offerings: Ensure a wide variety of genres and content formats to cater to different audience preferences.

Global Appeal: Focus on creating content with global appeal that can resonate with audiences worldwide.

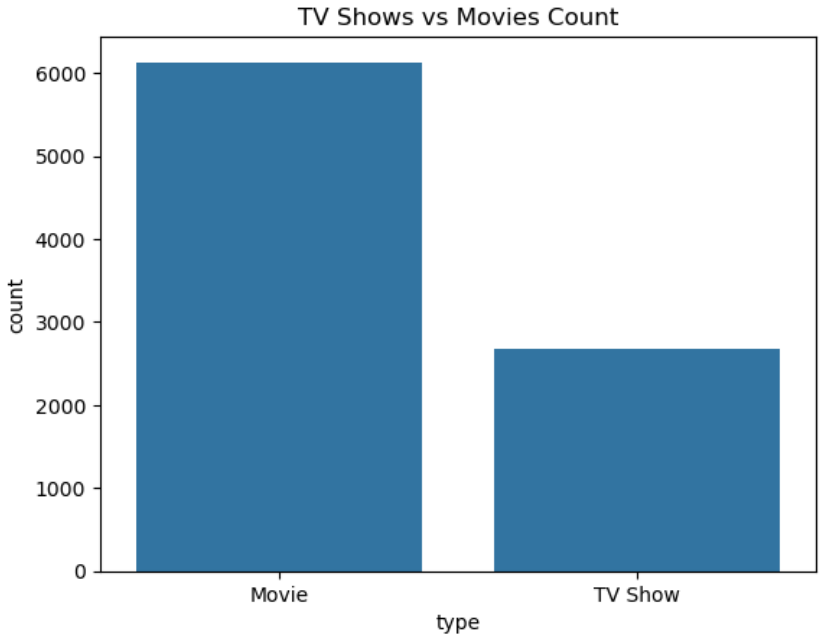
Strategic Partnerships: Form strategic partnerships with production companies and studios to secure exclusive licensing rights.

4. Comparison of TV Shows vs. Movies:

Code:

```
sns.countplot(x='type', data=df)
plt.title('TV Shows vs Movies Count')
plt.show()
```

Output:



Insights:

Movies: The number of movies on Netflix is significantly higher than the number of TV shows. This suggests that movies are a more prominent part of Netflix's content library.

Audience Preferences: The dominance of movies might indicate that Netflix's audience has a preference for movies over TV shows, or it could be due to factors like licensing agreements and production costs.

Recommendations:

Content Strategy: Understanding the audience's preference for movies can help Netflix optimize their content strategy and allocate resources accordingly.

Curated Collections: Create curated collections that highlight both movies and TV shows, making it easier for users to discover new content.

Personalized Recommendations: Improve the recommendation algorithm to ensure that both movies and TV shows are prominently featured in users' recommendations.

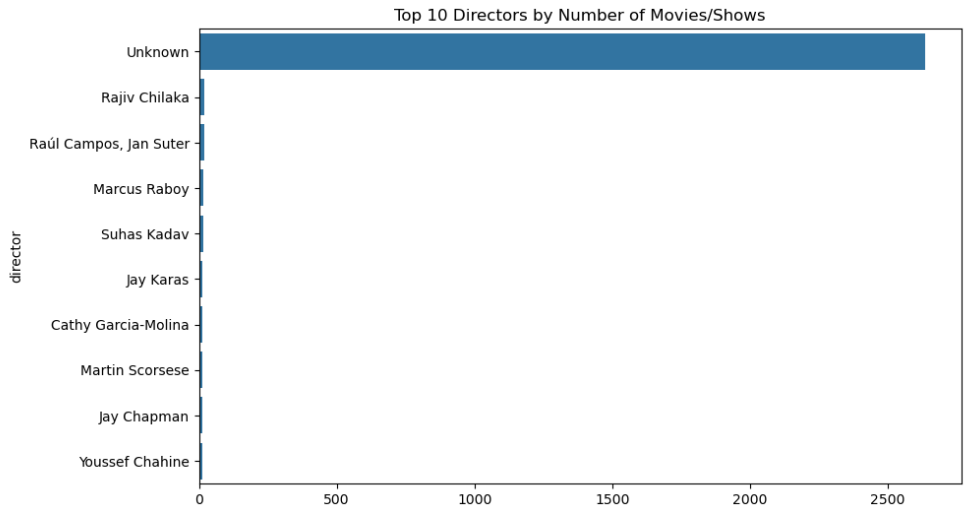
By this Netflix can achieve a more balanced library of movies and TV shows, catering to a wider range of audience preferences and ensuring long-term success.

5. Top 10 directors by number of movies and shows

Code:

```
df_directors = df.groupby('director').size().sort_values(ascending=False).head(10)
df_cast = df.groupby('cast').size().sort_values(ascending=False).head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=df_directors.values, y=df_directors.index)
plt.title('Top 10 Directors by Number of Movies/Shows')
plt.show()
```

Output:



Insights:

Unknown: The director "Unknown" is at the top of the list, indicating that a significant number of movies/shows on Netflix do not have credited directors. This could be due to various reasons, also while checking null values there are 2634 null values present in directors. So it could be one of them are dominating.

Rajiv Chilaka: Rajiv Chilaka is the second most prolific director, suggesting that he has been involved in a large number of Netflix productions.

Indian Directors: Several Indian directors are featured in the top 10, indicating a strong focus on Indian content on Netflix

International Representation: The presence of international directors like Martin Scorsese and Youssef Chahine suggests that Netflix is also acquiring or producing content from various regions.

Recommendations:

Diverse Genre Offerings: Ensure a wide variety of genres and subgenres to cater to different audience preferences.

Original Content: Invest in high-quality original content that is exclusive to Netflix.

Global Appeal: Create content with global appeal that can resonate with audiences worldwide.

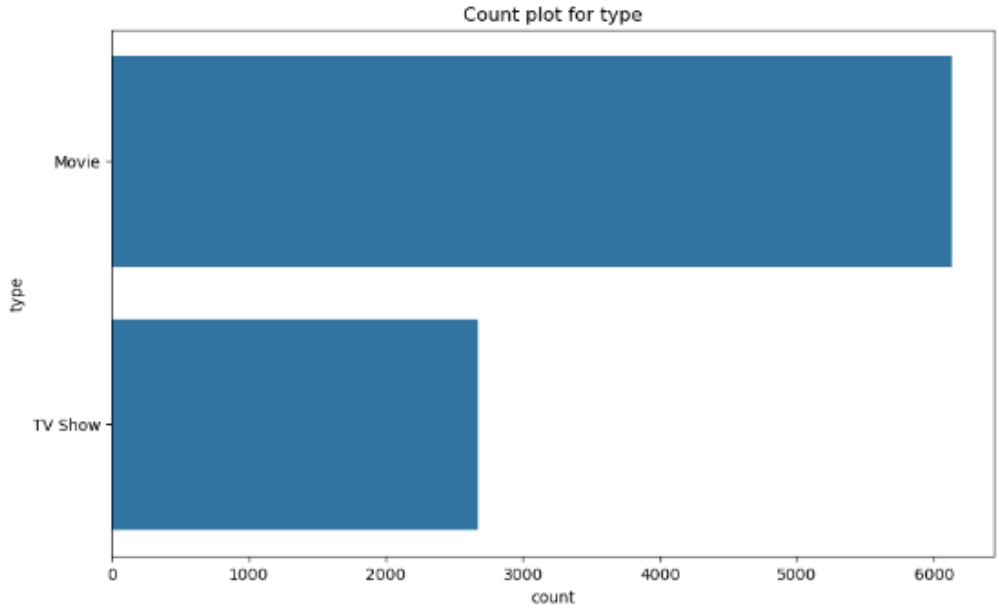
Local Content: Produce or acquire localized content to cater to specific regional markets and build local audiences.

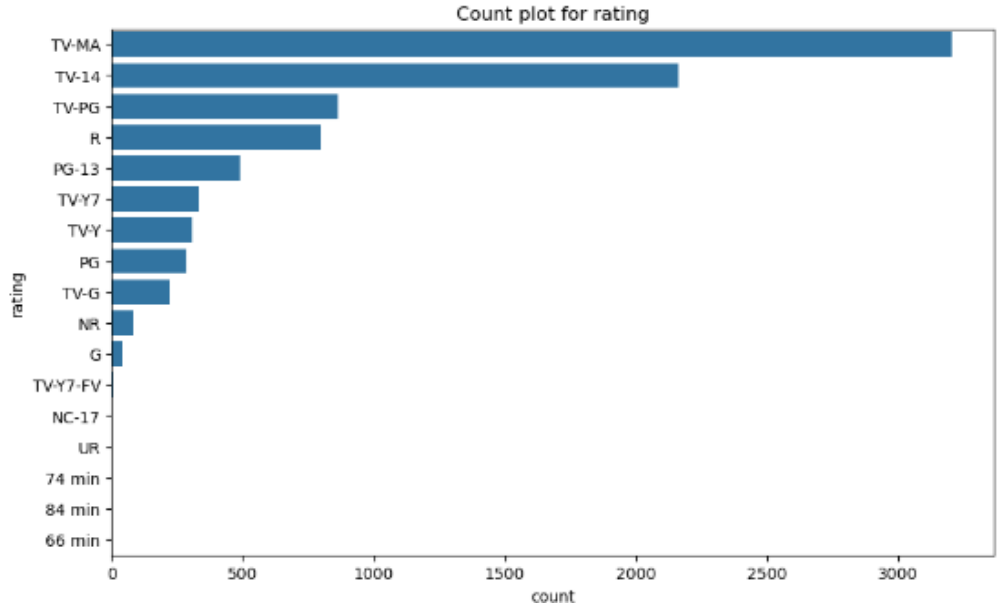
6. Count plot for type and ratings.

Code:

```
# Graphical analysis of categorical variables
for col in ['type','rating']:
    plt.figure(figsize=(10, 6))
    sns.countplot(data=df, y=col, order=df[col].value_counts().index)
    plt.title(f"Count plot for {col}")
    plt.show()
```

Output:





Insights:

Type of Content:

Movies dominate the Netflix library, with approximately **6000+ titles**, while **TV shows have around 2000+ titles**. This indicates that Netflix has a stronger focus on movies, making up around **75%** of the content offering compared to TV shows.

Ratings Distribution:

1. TV-MA (Mature Audience) is the most common rating, with around **3000+ titles**, suggesting that Netflix has a substantial amount of content aimed at adults.

2. TV-14 and **TV-PG** are the next most frequent ratings, indicating that Netflix also targets teenagers and families with appropriate content. However, these are significantly lower in number than TV-MA.

3. There are fewer titles with ratings like **PG-13**, **TV-Y7**, and **TV-Y** (for younger audiences), implying that Netflix's focus is more on mature content rather than content for children or families.

4. Ratings like **G**, **NR (Not Rated)**, **TV-Y7-FV**, and **NC-17** are minimally represented, showing limited focus on general audiences and certain niche categories.

Recommendations:

1. Balanced Content Strategy: Netflix should consider **increasing the production of TV shows** to balance the current heavy focus on movies. TV shows tend to engage users over longer periods, fostering subscription retention.

2. Diversifying Audience Target: While Netflix caters predominantly to adults with its TV-MA content, it would benefit from expanding its **family-friendly and children's content** (e.g., PG, TV-Y, TV-G). This could attract a broader audience, including families with young children.

3. Focus on International Expansion: Netflix could explore creating or acquiring more content with **PG-13 and family-friendly ratings**, which may appeal to a global audience where such ratings are widely accepted. This will help in increasing its user base in countries that value family-oriented content.

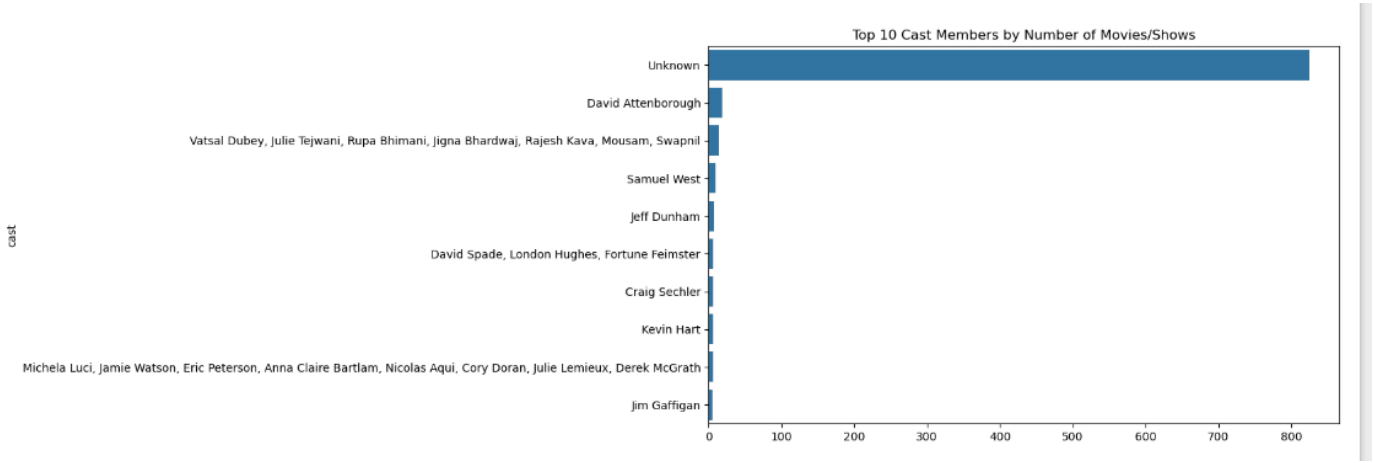
4. Recommendation System Enhancement: Given the heavy distribution of TV-MA content, Netflix should **improve its recommendation system** to ensure viewers from other segments (like families or younger viewers) are shown appropriate content based on their preferences.

6. Top 10 cast members by number of movies/show.

Code:

```
plt.figure(figsize=(10, 6))
sns.barplot(x=df_cast.values, y=df_cast.index)
plt.title('Top 10 Cast Members by Number of Movies/Shows')
plt.show()
```

Output:



Insights:

Unknown: The cast member "Unknown" is at the top of the list, indicating that a significant number of movies/shows on Netflix do not have credited cast members. This could be due to various reasons.

David Attenborough: David Attenborough is the second most frequent cast member, suggesting his involvement in a large number of Netflix productions, likely nature documentaries.

Regional Focus: Netflix could consider acquiring or producing more content from regions with underrepresented cast members to diversify their offerings.

Overall, the visualization highlights the dominance of certain cast members and the potential for regional bias in Netflix's content library.

Recommendations:

Actor-Specific Metrics: Track the performance of movies/shows featuring different actors to identify trends and areas for improvement.

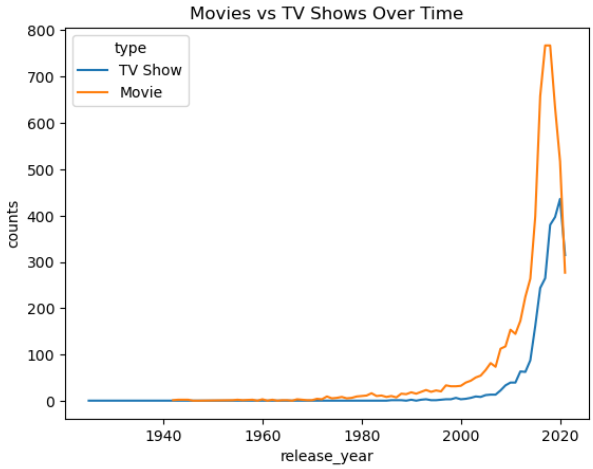
Content Acquisition: Consider acquiring content featuring popular actors to attract viewers.

7. Is there a shift in Netflix's content strategy towards TV shows?

Code:

```
yearly_content = df.groupby(['release_year', 'type']).size().reset_index(name='counts')
sns.lineplot(x='release_year', y='counts', hue='type', data=yearly_content)
plt.title('Movies vs TV Shows Over Time')
plt.show()
```

Output:



Insights:

Movie Growth: The number of movies released has steadily increased over the years, with a significant spike in the late 2000s and early 2010s.

TV Show Growth: The number of TV shows released has also increased, but at a slower pace compared to movies. There was a notable surge in TV show releases in the late 2000s and early 2010s, coinciding with the rise of streaming platforms.

Movie Dominance: Throughout the majority of the time period, movies have outnumbered TV shows.

Recent Shift: In recent years, there has been a slight increase in the number of TV shows released relative to movies, indicating a growing trend towards TV series.

Streaming Impact: The rise of streaming platforms has likely contributed to the increase in both movie and TV show releases, as these platforms provide a wider distribution channel and demand for original content.

Audience Preferences: The increasing number of TV shows suggests that audiences may be shifting their preferences towards TV series.

Recommendations:

Balanced Approach: Maintain a balance between movies and TV shows to cater to diverse audience preferences.

Genre Exploration: Explore a wider range of genres and subgenres within both movies and TV shows.

Original Content: Invest in high-quality original content in both movie and TV show formats.

Emerging Technologies: Stay updated on emerging technologies (e.g., VR, AR) and explore opportunities to incorporate them into content production and distribution.

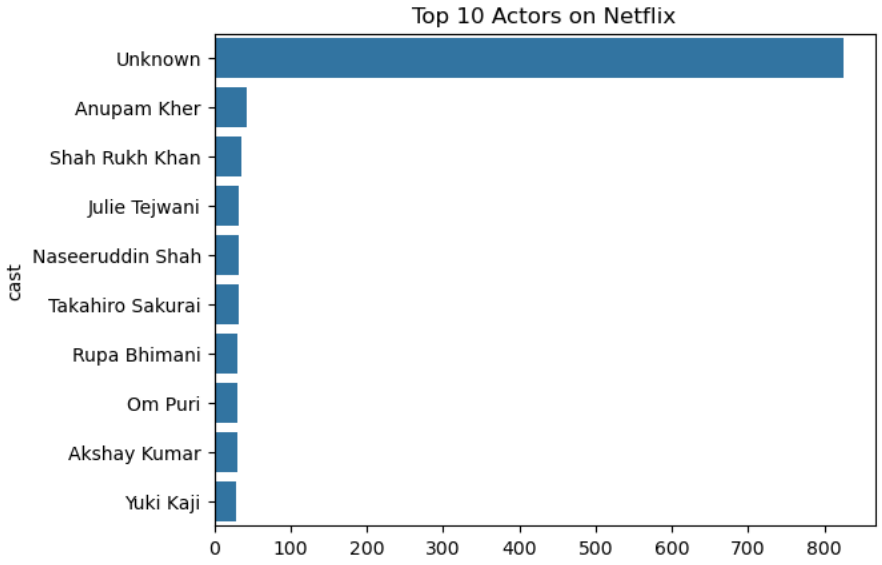
Market Analysis: Continuously monitor market trends and consumer preferences to adapt strategies accordingly.

8. Top 10 Actors on Netflix

Code:

```
df['cast'] = df['cast'].str.split(', ')
cast_counts = df.explode('cast').groupby('cast').size().sort_values(ascending=False)
sns.barplot(y=cast_counts.index[:10], x=cast_counts.values[:10])
plt.title('Top 10 Actors on Netflix')
plt.show()
```

Output:



Insights:

Unknown: The actor "Unknown" is at the top of the list, indicating that a significant number of movies/shows on Netflix do not have credited actors. This could be due to various reasons.

Anupam Kher: Anupam Kher is the second most frequent actor and then followed by Shah Rukh Khan, suggesting their involvement in a large number of Netflix productions.

Indian Actors: Several Indian actors are featured in the top 10, indicating a strong focus on Indian content on Netflix.

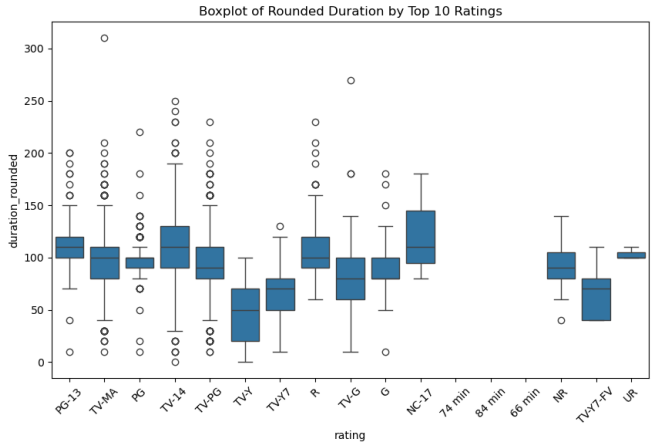
Overall, the visualization highlights the dominance of certain actors and the potential for regional bias in Netflix's content library.

9. The distribution of rounded durations for different ratings categories.

Code:

```
df['duration_numeric'] = df['duration'].apply(lambda x: int(x.split(' ')[0]) if 'min' in str(x) else None)
# Round the durations to the nearest 10 for clearer visual representation
df['duration_rounded'] = df['duration_numeric'].apply(lambda x: round(x, -1))
# Plot the boxplot with rounded durations
plt.figure(figsize=(10, 6))
sns.boxplot(x='rating', y='duration_rounded', data=df)
plt.xticks(rotation=45)
plt.title('Boxplot of Rounded Duration by Top 10 Ratings')
plt.show()
```

Output:



Insights:

Median Duration: The median duration for most ratings categories falls within the range of 100-150 minutes.
Variation: The variation in duration (as indicated by the box plots and whiskers) is relatively high for some ratings categories, such as PG-13 and TV-MA, suggesting a wider range of movie lengths.
Outliers: There are some outliers (individual data points outside the whiskers) present in several ratings categories, indicating movies with significantly longer or shorter durations compared to the majority.
Overall, the boxplot provides a visual representation of the distribution of movie durations across different ratings categories.

Recommendations:

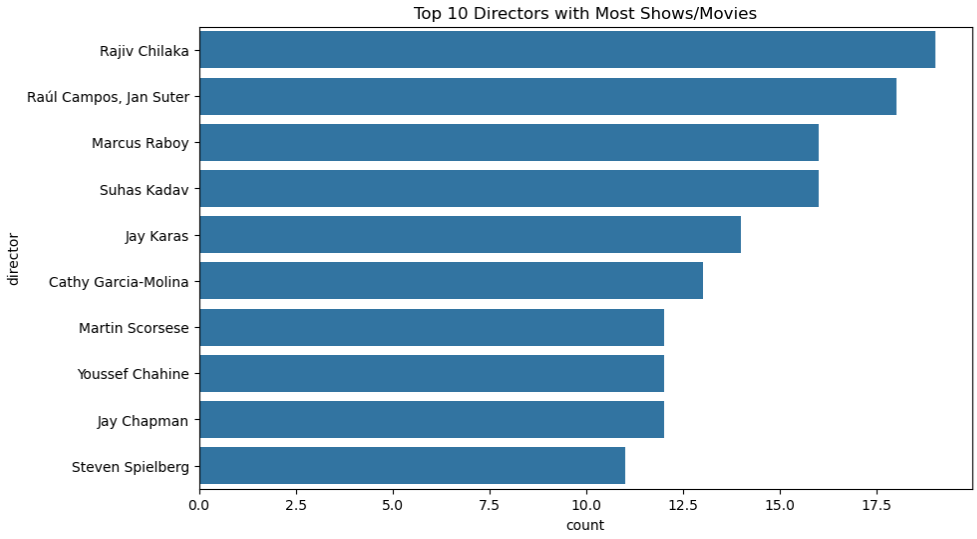
Target Audience: Consider the target audience for each rating category and tailor content accordingly. For example, movies rated PG-13 might appeal to a broader audience, while NC-17 movies might have a more niche appeal.
Content Length: Be mindful of the typical duration for different ratings categories. For example, if a movie is rated TV-G, it might be appropriate to keep the duration shorter to cater to a younger audience.
Audience Engagement: Evaluate how duration affects audience engagement and retention. Shorter movies might be more suitable for binge-watching, while longer movies could offer a more immersive experience.
Target Audience: Tailor marketing and distribution strategies based on the target audience for each rating category.

10. Most common directors.

Code:

```
top_directors = df['director'].value_counts().index[:10]
filtered_directors = df[df['director'].isin(top_directors)]

plt.figure(figsize=(10, 6))
sns.countplot(y='director', data=filtered_directors, order=top_directors)
plt.title('Top 10 Directors with Most Shows/Movies')
plt.show()
```

Output:



Insights:

Bar plot representing the top 10 directors with the most shows/movies on a platform (presumably Netflix, based on the previous context).

Rajiv Chilaka: Rajiv Chilaka leads the list, indicating that he has been involved in a significant number of productions.

Raul Campos, Jan Suter: The duo of Raul Campos and Jan Suter is the second most prolific, suggesting their frequent collaboration.

Indian Directors: Several Indian directors are featured in the top 10, suggesting a strong focus on Indian content.

Recommendations:

Production Companies: Examining the production companies associated with these directors could reveal any patterns or biases in terms of content acquisition or production.

Director-Specific Metrics: Track the performance of movies/shows directed by different directors to identify trends and areas for improvement.

Audience Preferences: Use data analytics to understand audience preferences for specific directors and their styles.

Content Acquisition: Consider acquiring content featuring popular directors to attract viewers.

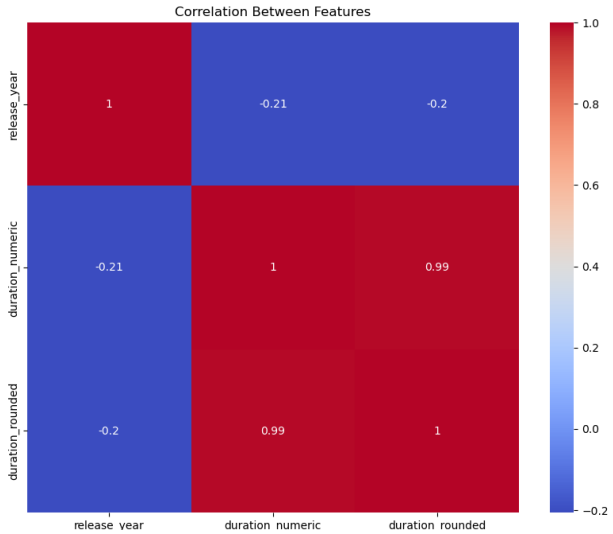
4. Correlation Analysis

1.Correlation Between Features:

Code:

```
plt.figure(figsize=(10, 8))
corr_matrix = df.corr(numeric_only=True)
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Between Features')
plt.show()
```

Output:



Insights:

Strong Correlation Between duration numeric and duration rounded:

The correlation coefficient between duration numeric and duration rounded is **0.99**, which is almost perfect. This is expected since duration rounded is simply the rounded version of duration numeric. This means there is a very high linear relationship between these two features, but it doesn't provide new information. It suggests that one of the two columns can be dropped in further analysis to avoid redundancy.

Negative Correlation Between release year and duration:

There is a moderate negative correlation between release year and duration numeric (-0.21), indicating that as the release year increases (i.e., newer shows), the duration tends to decrease. This might imply that more recent content on Netflix tends to have shorter durations compared to older content.

Correlation Between release year and duration rounded:

The correlation between release year and duration rounded is slightly lower but still negative (-0.2). This further supports the observation that newer releases tend to be shorter in duration.

Recommendations:

Focus on Shorter Content for Newer Releases:

Given the negative correlation between release year and duration, it seems that newer content tends to be shorter. This aligns with the modern trend of shorter, more digestible content such as mini-series and short films.

Target Older Content with Longer Durations for Audience Segmentation:

Since older content tends to have longer durations, Netflix could use this insight to create targeted campaigns for audiences who enjoy classic, long-form content.

Overall Conclusion:

The analysis of Netflix's dataset reveals significant trends in content production, audience preferences, and regional focus. The dominance of the United States in both movies and TV shows reflects Netflix's strong base in Western content. However, the growing presence of countries like India, Japan, and South Korea indicates Netflix's expansion into emerging markets and its efforts to cater to global audiences.

The steady growth of movie releases, especially in the mid-2010s, suggests a competitive market driven by the rise of streaming platforms. However, recent plateaus and shifts towards shorter, more frequent releases in newer content reflect evolving production and consumption trends. Notably, movies outnumber TV shows across most years, indicating that Netflix has a stronger focus on films, although the popularity of TV shows has been rising in recent year

There is also a moderate negative correlation between the release year and the duration of content, suggesting that more recent productions tend to be shorter. This could reflect changing consumption habits as users prefer more concise content in a highly competitive market. The duration correlation analysis further revealed redundancy between certain columns, providing an opportunity for data optimization.

Strong Recommendation to Netflix :

Recommendation: Netflix should consider expanding its content library into other emerging markets beyond India, such as Southeast Asia and Africa, where there is a growing demand for streaming content. By investing in local productions and regional partnerships, Netflix can not only increase its global subscriber base but also diversify its content offerings, appealing to a wider audience and strengthening its position as a leader in global streaming. This approach will help the platform remain competitive, particularly as new streaming services enter these markets.

Expand Regional Content Offerings: Netflix should continue to diversify its content by focusing on underrepresented regions, particularly in emerging markets. While the U.S. and India dominate, countries like Japan and South Korea also show significant potential, especially with the rising popularity of anime and K-dramas.

Leverage Bollywood and International Stars: With actors like Anupam Kher and Shah Rukh Khan leading the list of most featured actors, Netflix can tap further into the Bollywood industry to appeal to its global audience. Collaboration with other international stars will also expand its appeal.

Focus on Animated and Family Content: Directors like Rajiv Chilaka and voice actors involved in animated series signal a strong demand for family and animated content. Netflix should invest more in this genre to engage younger audiences and families.

Optimize Release Strategies for Seasonal Peaks: The insights show that December and July are peak release months. Netflix should capitalize on these periods with high-impact content, while also exploring opportunities to spread out releases throughout the year.

Balance Movie and TV Show Offerings: While movies have historically outnumbered TV shows, the recent surge in TV show popularity indicates a shift in audience preference. Netflix should aim for a balanced content strategy by increasing investments in TV series to capture audience interest.

Focus on Shorter, Engaging Content: The negative correlation between release year and content duration suggests that newer releases are becoming shorter. Netflix should focus on producing more concise and engaging content, which aligns with modern viewer preferences.

Jupyter Notebook Analysis

For a detailed view of the full analysis, including code, visualizations, and insights, please refer to the complete Jupyter notebook available in the PDF format. The notebook documents each step of the analysis process, from data exploration to the final recommendations.

You can access the Jupyter notebook PDF through the **following link**:

Jupyter Notebook