Data Exploration and Visualisation — Netflix

Presented By :: Suman Kumar Nandi

1. Importing Libraries -

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

2. Reading the .csv file

```
df = pd.read_csv('/content/netflix.csv')
```

```
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	des
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	25-Sep-21	2020	PG-13	90 min	Documentaries	As nea
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	24-Sep-21	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	pa
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	24-Sep-21	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	

Basic Operations on the dataset

INSIGHT 1 : Total count of rows is 8807 and no of columns is 12
INSIGHT 2 : ALL THE Data are Non numeric Columns
INSIGHT 3 : THERE ARE 4307 NULL VALUES FOUND
INSIGHT 4 : THERE ARE NO DUPLICATE VALUES
INSIGHT 5 : DESCRIPTIVE STATS OF THE CATEGORICAL COLUMNS

Observations -

- There are total 4307 missing values in the entire data set.
- Columns that contains Null Values are:

```
director        2634
cast             825
country          831
date_added        10
rating             4
duration           3
```

```
df.shape
```

```
(8807, 12)
```

```
df.info()
```

```
<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   category
 2   title         8807 non-null   object
 3   director      6304 non-null   object
 4   cast          8707 non-null   object
 5   country       8048 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   category
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: category(2), int64(1), object(9)
memory usage: 706.2+ KB
```

```
df.isnull().sum().sum()
```

```
4307
```

```
df.duplicated().sum()
```

```
0
```

```
df.describe()
```

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

df.describe() :: give the insight of

1. Count: 8,807 entries
2. Mean: Around the year 2014
3. Standard Deviation: Approximately 8.82 years
4. Minimum: Year 1925
5. 25th Percentile (Q1): Year 2013
6. Median (50th Percentile): Year 2017
7. 75th Percentile (Q3): Year 2019
8. Maximum: Year 2021

```
df.describe(include = 'object')
```

	show_id	title	director	cast	country	date_added	duration	listed_in	description
count	8807	8807	6304	8707	8048	8797	8804	8807	8807
unique	8807	8804	4564	7118	113	1767	220	514	8775
top	s1	15-Aug	Rajiv Chilaka	Julie Tejwani	United States	1-Jan-20	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	2	22	22	2894	109	1793	362	4

3.a) **Handling Nan/Missing Values**

```
df.isnull().sum()
```

```
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
```

```
df.isnull().sum().sum()
```

```
4307
```

```
df[df.isnull().any(axis=1)]
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_i
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	25-Sep-21	2020	PG-13	90 min	Documentarie
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	24-Sep-21	2021	TV-MA	2 Seasons	Internationa TV Shows, T Dramas, T Mysterie
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	24-Sep-21	2021	TV-MA	1 Season	Crime T Shows Internationa TV Shows, T Act.
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	24-Sep-21	2021	TV-MA	1 Season	Docuseries Reality T
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	24-Sep-21	2021	TV-MA	2 Seasons	Internationa TV Shows Romantic T Shows, TV .
...	...	...	...	...	...	...	...	...	...	...	.
8795	s8796	TV Show	Yu-Gi-Oh! Arc-V	NaN	Mike Liscio, Emily Bauer, Billy Bob Thompson, ...	Japan, Canada	1-May-18	2015	TV-Y7	2 Seasons	Anime Series Kids' T
8796	s8797	TV Show	Yunus Emre	NaN	Gökhan Atalay, Payidar Tüfekçioglu, Baran Akbu...	Turkey	17-Jan-17	2016	TV-PG	2 Seasons	Internationa TV Shows, T Drama
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George,	United States, France, South Korea,	13-Sep-18	2016	TV-Y7	3 Seasons	Kids' T

3.b) **Imputation of Missing values:**

Replacing the missing value of perticular column looking at above info, we can impute the cells of columns "director

```
# fill missing values in Director Column with "No_Director"
df.director.fillna("No_Director", inplace = True)
# fill missing values in Cast Column with "No_Cast"
df.cast.fillna("No_Cast", inplace = True)
# fill missing values in Country Column with most frequent country apreared in that column (mode of the 'Country Column')
df['country'].fillna(df['country'].mode()[0], inplace = True)
# Dropping irrelevant rows with minimal null values
df.dropna(subset = ["date_added", 'rating','duration'], inplace = True)
# Re-check the data
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	de
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No_Cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries	A ne
1	s2	TV Show	Blood & Water	No_Director	Ama Qamata	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	p
2	s3	TV Show	Ganglands	Julien Leclercq	Khosi Ngema	United States	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	
3	s4	TV Show	Jailbirds New Orleans	No_Director	Gail Mabalane	United States	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	ar
4	s5	TV Show	Kota Factory	No_Director	Thabang Molaba	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	

```
df.shape
```

```
(8790, 16)
```

```
df.isnull().any()
```

```
show_id         False
type            False
title           False
director        False
cast            False
country         False
date_added      False
release_year    False
rating          False
duration        False
listed_in       False
description     False
dtype: bool
```

4.a) **Non Graphical Analysis**

1)Convert categorical attributes to 'category' data type if required

2) Value counts for key attributes

3) Unique attributes for key columns

1. **Type**

- Count: 8,807
- Unique Values: 2 (Movie, TV Show)
- Most Frequent: Movie
- Frequency: 6,131

2. **Country**

- Count: 7,976 (some missing values)
- Unique Values: 748
- Most Frequent: United States

- Frequency: 2,818

3. **Rating**

- Count: 8,803 (some missing values)
- Unique Values: 17
- Most Frequent: TV-MA
- Frequency: 3,207

1) Convert categorical attributes to 'category' data type if required

```
# Convert categorical attributes to 'category' data type if required
categorical_columns = ['type', 'country', 'rating']
df[categorical_columns] = df[categorical_columns].astype('category')

# After conversion data types
after_conversion_data_types = df.dtypes
# Missing value detection
missing_values = df.isnull().sum()


after_conversion_data_types
```

```
show_id          object
type           category
title            object
director         object
cast             object
country        category
date_added       object
release_year      int64
rating         category
duration         object
listed_in        object
description      object
dtype: object
```

```
missing_values
```

```
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
description     0
dtype: int64
```

2) Value counts for key attributes

```
value_counts_type = df['type'].value_counts()
value_counts_type
```

```
type
Movie      6126
TV Show    2664
Name: count, dtype: int64
```

```
value_counts_country = df['country'].value_counts()  # Top 10 countries
value_counts_country.head(10)
```

```
country
United States     3630
India              851
United Kingdom     588
Canada             344
France             314
Japan              290
Spain              210
South Korea        204
Germany            185
Mexico             144
Name: count, dtype: int64
```

```
value_counts_rating = df['rating'].value_counts()
value_counts_rating
```

```
rating
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
```

```
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
UR             3
NC-17          3
74 min         1
84 min         1
66 min         1
Name: count, dtype: int64
```

```
value_counts_release_year = df['release_year'].value_counts().head(10)  # Top 10 release years
value_counts_release_year
```

```
release_year
2018    1146
2017    1030
2019    1030
2020     953
2016     901
2021     592
2015     555
2014     352
2013     286
2012     236
Name: count, dtype: int64
```

3) Unique attributes for key columns

```
# Unique attributes for key columns
unique_type = df['type'].nunique()
unique_country = df['country'].nunique()
unique_rating = df['rating'].nunique()
unique_title = df['title'].nunique()
unique_director = df['director'].nunique()
unique_cast = df['cast'].nunique()
```

```
unique_type, unique_country, unique_rating, unique_title, unique_director, unique_cast
```

```
(2, 748, 14, 8787, 4527, 7679)
```

b) Graphical Analysis -

Pre-processing of the data

```
# Converting the dataatype of date column
df['date_added'] = pd.to_datetime(df['date_added'], errors = 'coerce')

# Extracting month, month name, year and day from the 'date_added' column
df['month_added'] = df['date_added'].dt.month
df['month_name_added'] = df['date_added'].dt.month_name()
df['year_added'] = df['date_added'].dt.year
df['day_added'] = df['date_added'].dt.day
df['Week_No'] = df['date_added'].dt.strftime('w%V')
```

```
# Un-nesting the Cast, Country and Director Columns

df_exp = df

df_exp['cast'] = df['cast'].str.split(', ').explode('cast').to_frame()
df_exp['country'] = df['country'].str.split(', ').explode('country').to_frame()
df_exp['director'] = df['director'].str.split(', ').explode('director').to_frame()
df_exp['listed_in'] = df['listed_in'].str.split(', ').explode('listed_in').to_frame()


df.head(10)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	c
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No_Cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries	r
1	s2	TV Show	Blood & Water	No_Director	Ama Qamata	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows	
2	s3	TV Show	Ganglands	Julien Leclercq	Khosi Ngema	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas	
3	s4	TV Show	Jailbirds New Orleans	No_Director	Gail Mabalane	United States	2021-09-24	2021	TV-MA	1 Season	TV Mysteries	a
4	s5	TV Show	Kota Factory	No_Director	Thabang Molaba	India	2021-09-24	2021	TV-MA	2 Seasons	Crime TV Shows	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Dillon Windvogel	United States	2021-09-24	2021	TV-MA	1 Season	International TV Shows	a
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen	Natasha Thahane	United States	2021-09-24	2021	PG	91 min	TV Action & Adventure	c
7	s8	Movie	Sankofa	José Luis Ucha	Arno Greeff	United States	2021-09-24	1993	TV-MA	125 min	Docuseries	
8	s9	TV Show	The Great British Baking Show	Haile Gerima	Xolile Tshabalala	Ghana	2021-09-24	2021	TV-14	9 Seasons	Reality TV	
9	s10	Movie	The Starling	Andy Devonshire	Getmore Sithole	Burkina Faso	2021-09-24	2021	PG-13	104 min	International TV Shows	

```
df.shape
```

```
(8790, 17)
```

```
df_exp.shape
```

```
(8790, 17)
```

Observations -

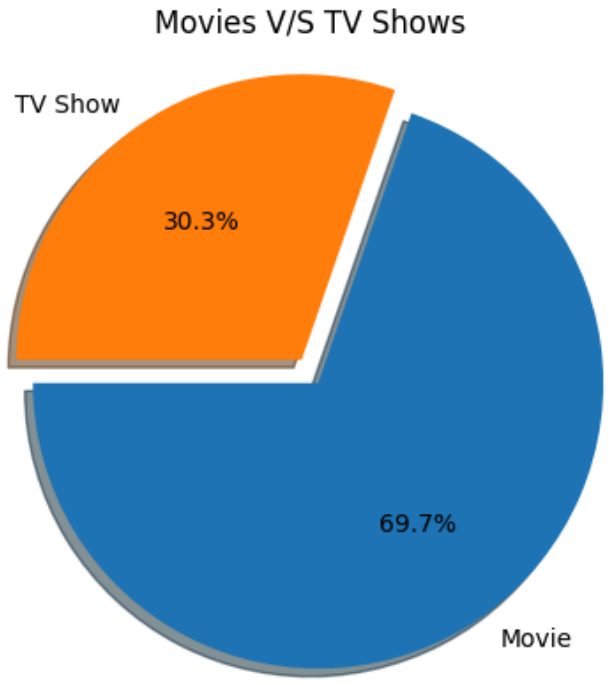
Un-nesting of the coulmns increased the total number of rows in the data frame The 3 new columns are added in the data frame assiciated with date and time

Visualization

Univariate Analysis : Pie Plot for Type of Content

Let’s start by understanding the distribution of Movies vs. TV Shows on Netflix.

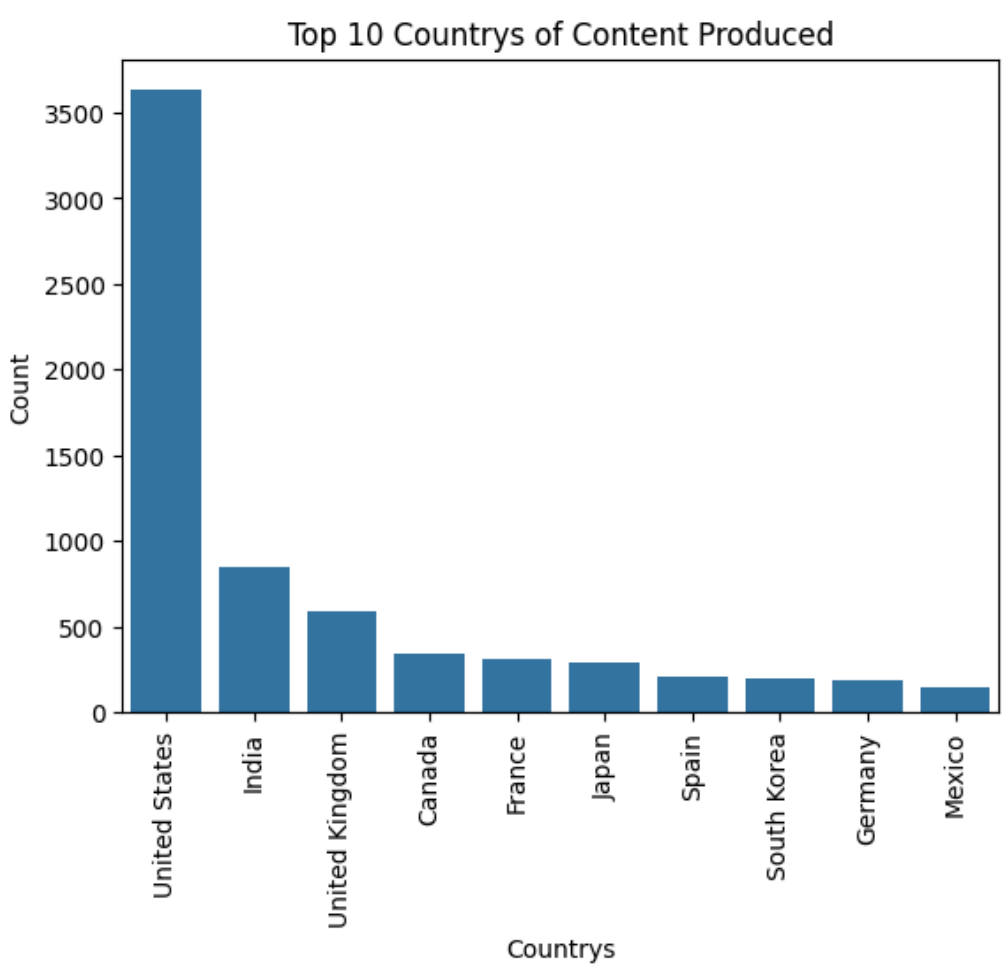
```
 # Comparision of Type of Content Produced

labels = value_counts_type.index
sizes = value_counts_type.values
explode = (0.1 ,0)
plt.pie(sizes, labels=labels, explode = explode, shadow = True, autopct='%1.1f%%', startangle = 180)
plt.title('Movies V/S TV Shows')
plt.axis('equal')
plt.show()
```



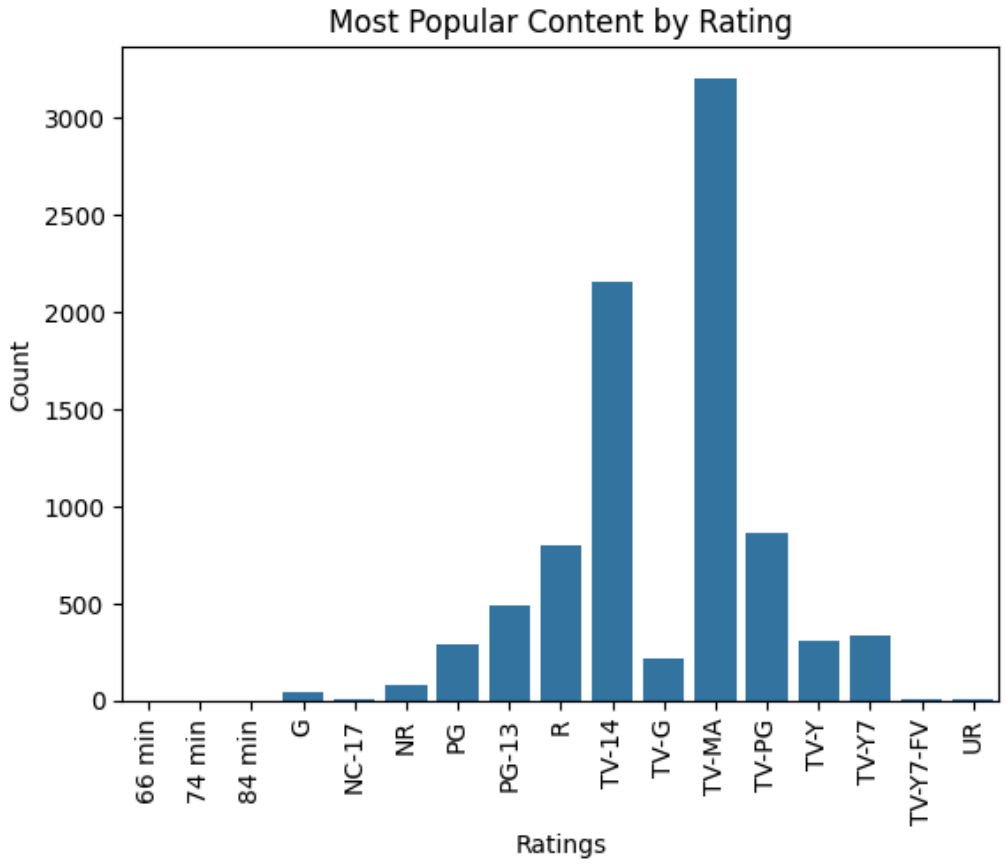
Observations -

Nearly 70% content on the Netflix is of Movies adn remaining 30% is of TV-Shows

```
# Top 10 Countrys of Content Produced
top_countrys = value_counts_country.head(10)

x_count = top_countrys.index
y_count = top_countrys.values

sns.barplot(x = x_count, y= y_count)
plt.xticks(rotation = 90)
plt.xlabel('Countrys')
plt.ylabel('Count')
plt.title('Top 10 Countrys of Content Produced')
plt.show()
```

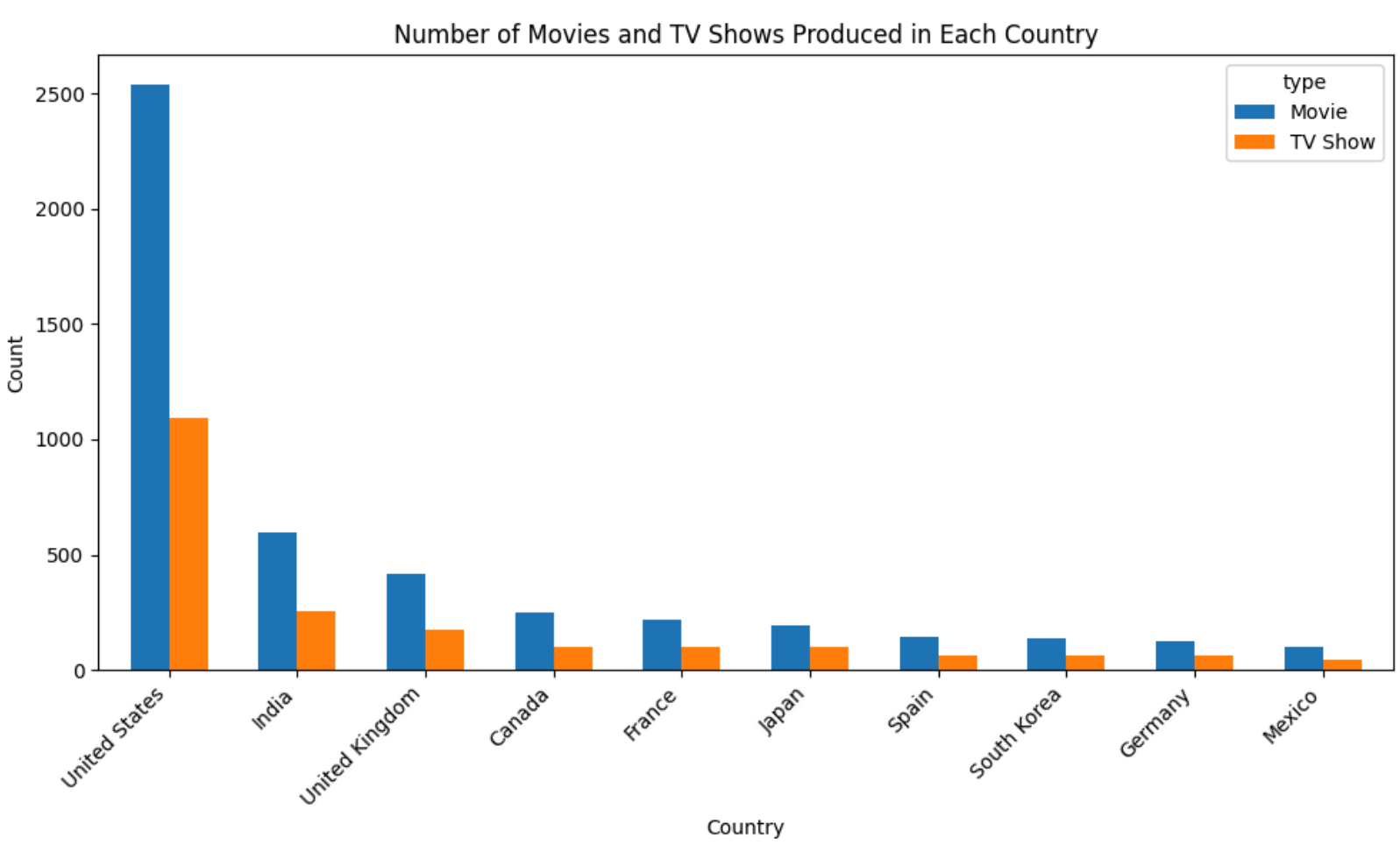



```
# Most popular Content by Rating
x_r = value_counts_rating.index
y_r = value_counts_rating.values
sns.barplot(x = x_r, y= y_r)
plt.xticks(rotation = 90)
plt.xlabel('Ratings')
plt.ylabel('Count')
plt.title('Most Popular Content by Rating')
plt.show()
```



5) **Comparison of TV Shows and Movies**

```
# No of Movies vs No of TV-Shows Produced in top 10 Contries

# grouping the dataset by country and type and count the number of rows in each group
df_grouped = df_exp.groupby(['country', 'type']).size().reset_index(name='count')

# filtering the dataset to only include movies and TV shows
df_filtered = df_grouped[df_grouped['type'].isin(['Movie', 'TV Show'])]

# pivot the dataset to create a table with countries as rows and movie and TV show counts as columns
df_pivoted = df_filtered.pivot(index='country', columns='type', values='count')

# sort the table by the total number of movies and TV shows
df_sorted = df_pivoted.sort_values(by=['Movie', 'TV Show'], ascending=False)

# pick the top 10 countries
df_top_10 = df_sorted.head(10)

# plot the results as a dodged bar plot
ax = df_top_10.plot(kind='bar', stacked=False, figsize=(10, 6), width=0.6)
ax.set_title('Number of Movies and TV Shows Produced in Each Country')
ax.set_xlabel('Country')
ax.set_ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



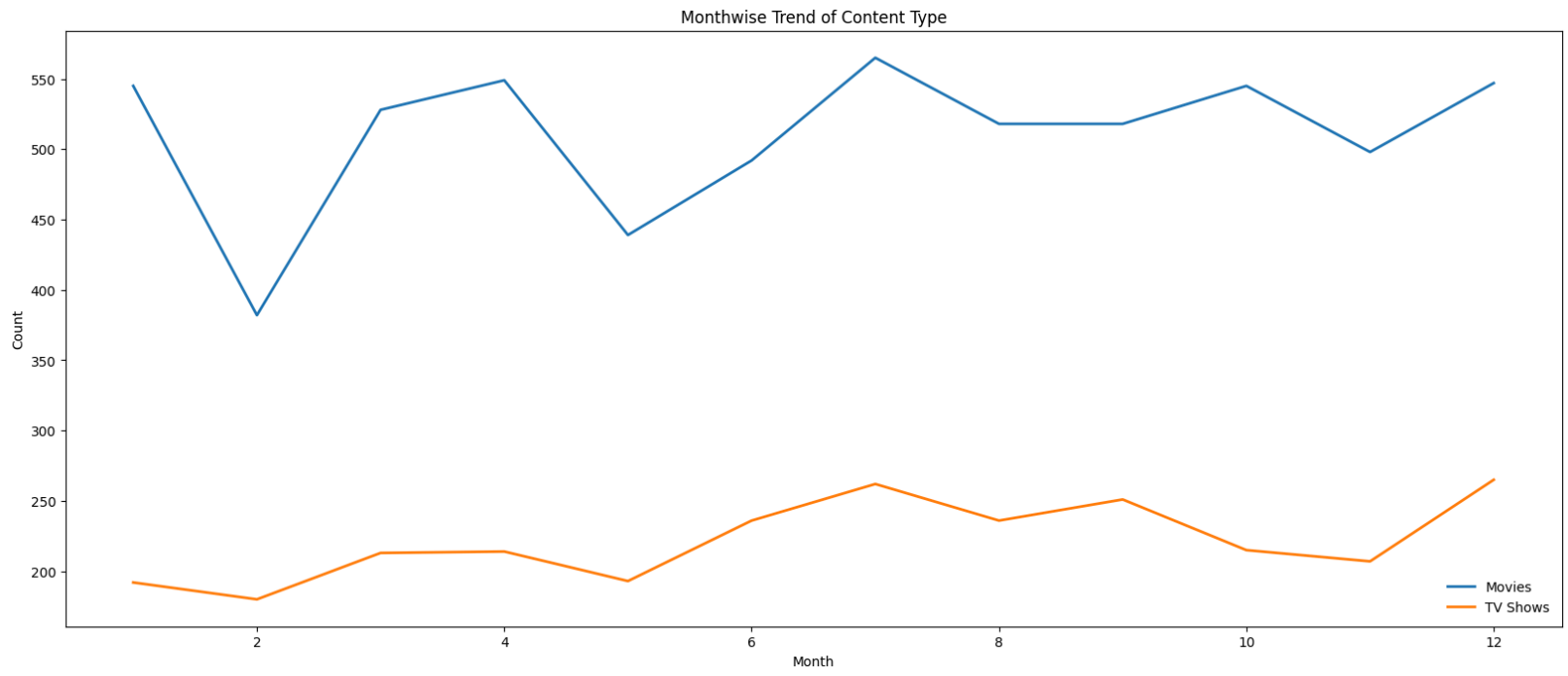

```
df_top_10
```

type	Movie	TV Show
country		
United States	2540	1090
India	599	252
United Kingdom	416	172
Canada	247	97
France	217	97
Japan	193	97
Spain	146	64
South Korea	139	65
Germany	124	61
Mexico	99	45

6. What is the best month of the year to launch a Movie and TV show?

```
# grouping the dataset by month and type and count the number of rows in each group
df_mnth_grp = df_exp.groupby(['month_added', 'type']).size().reset_index(name='count')

# filtering the dataset to only include movies and TV shows
df_filt_mnth = df_mnth_grp[df_grouped['type'].isin(['Movie', 'TV Show'])]

# pivot the dataset to create a table with countries as rows and movie and TV show counts as columns
df_mnth = df_filt_mnth.pivot(index='month_added', columns='type', values='count')

# sort the table by the total number of movies and TV shows
df_mnth_sort = df_mnth.sort_values(by=['month_added','Movie', 'TV Show'], ascending=False)

plt.figure(figsize=(20, 8))
sns.lineplot(data = df_mnth_sort, x=df_mnth_sort.index, y=df_mnth_sort['Movie'], label = 'Movies', linewidth = 2)
sns.lineplot(data = df_mnth_sort, x=df_mnth_sort.index, y=df_mnth_sort['TV Show'], label = 'TV Shows', linewidth = 2)
plt.legend(loc='lower right', frameon = False)
plt.xlabel('Month')
plt.ylabel('Count')
plt.title('Monthwise Trend of Content Type')
plt.show()
```

```
<ipython-input-166-9d70a3a93204>:5: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  df_filt_mnth = df_mnth_grp[df_grouped['type'].isin(['Movie', 'TV Show'])]
```

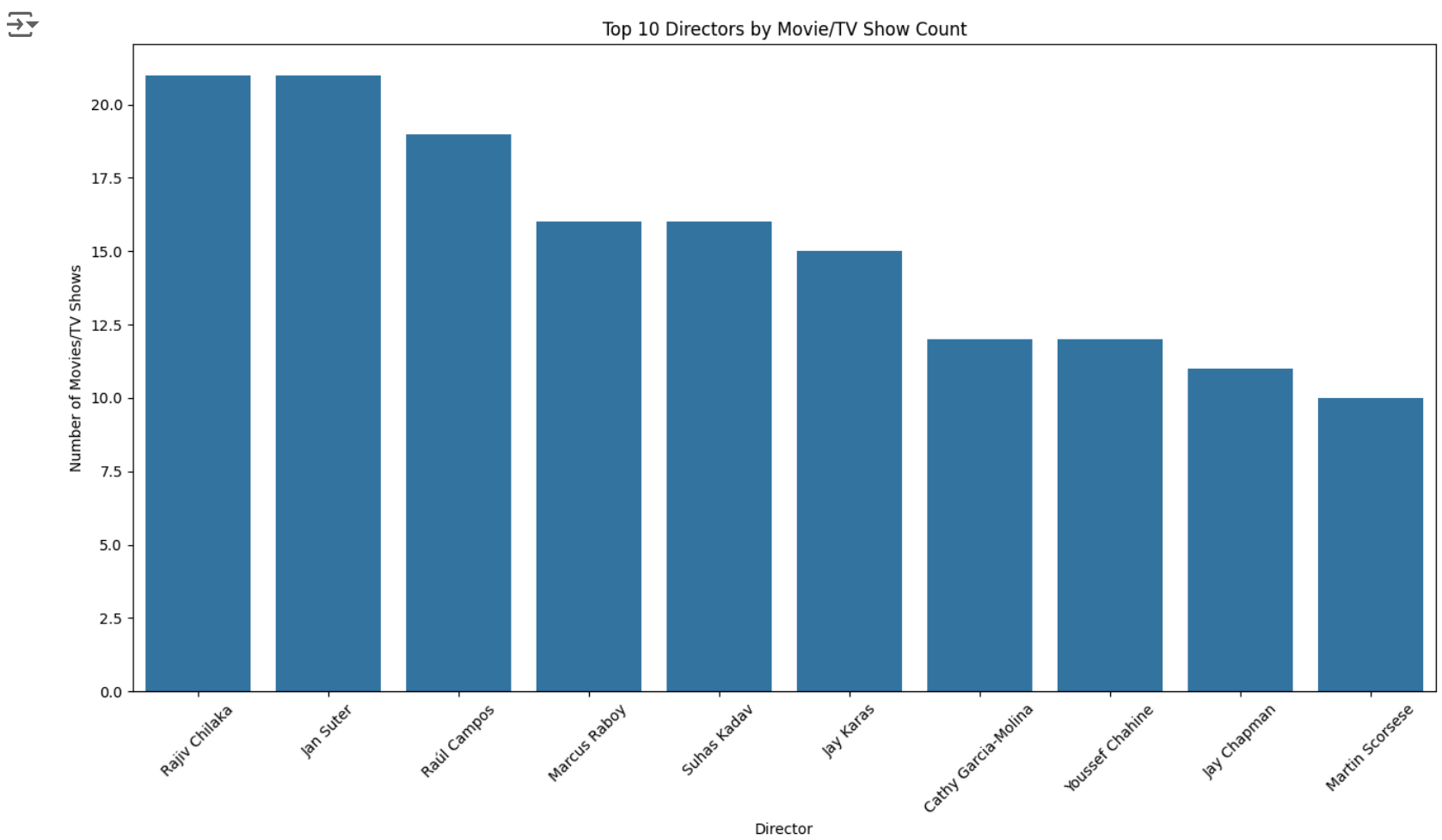



```
df_mnth_sort
```

type	Movie	TV Show
month_added		
12	547	265
11	498	207
10	545	215
9	518	251
8	518	236
7	565	262
6	492	236
5	439	193
4	549	214
3	528	213
2	382	180
1	545	192

7. Top 10 Directors of Movies/TV-Shows

```
# Count the occurrences of each director
director_counts = df_exp['director'].value_counts()[1:]

# Select the top 10 directors
top_10_directors = director_counts.head(10)

plt.figure(figsize=(16, 8))

bar_plot = sns.barplot(x=top_10_directors.index, y=top_10_directors.values)

plt.xticks(rotation = 45)
plt.xlabel('Director')
plt.ylabel('Number of Movies/TV Shows')
plt.title('Top 10 Directors by Movie/TV Show Count')


plt.show()
```



Observations - Jaun Suter and Rajiv Chilaka both are the top most dirctor who have directed more than 20 number of movies and TV-Shows

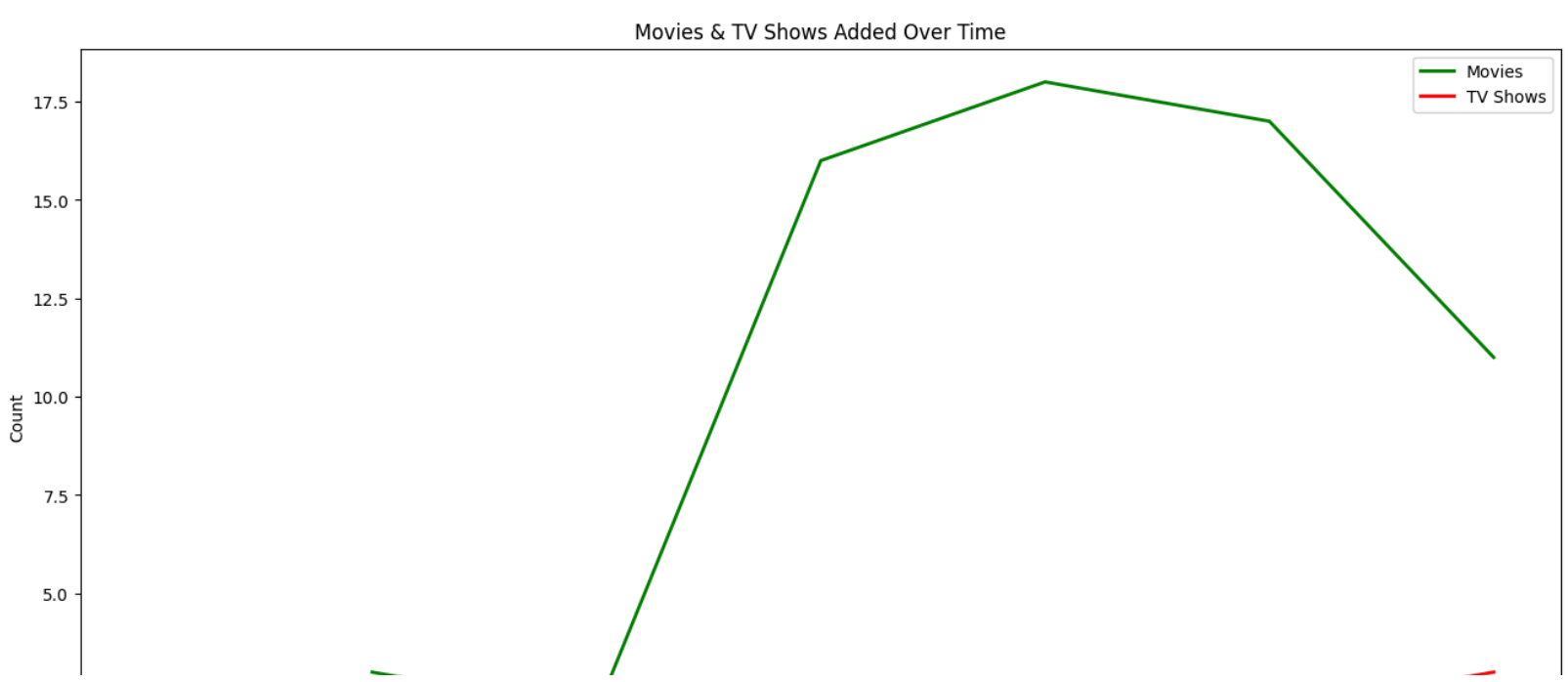
```
from wordcloud import WordCloud
df_gen = df[df['type'] == 'Movie']
# Concatenate all the titles into a single string
text = ' '.join(df_gen['listed_in'])

wordcloud = WordCloud(width = 600, height = 600,
                background_color ='white',
                min_font_size = 5).generate(text)

# plot the WordCloud image
plt.figure(figsize = (5, 5), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

Observations - Most Popular Genres are "Intrnational Movies" , "International Dramas" , "Action" and " Adventure"

```
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	de
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No_Cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries	A ne
1	s2	TV Show	Blood & Water	No_Director	Ama Qamata	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows	p
2	s3	TV Show	Ganglands	Julien Leclercq	Khosi Ngema	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas	
3	s4	TV Show	Jailbirds New Orleans	No_Director	Gail Mabalane	United States	2021-09-24	2021	TV-MA	1 Season	TV Mysteries	ar
4	s5	TV Show	Kota Factory	No_Director	Thabang Molaba	India	2021-09-24	2021	TV-MA	2 Seasons	Crime TV Shows	

```
df['year_diff'] = df['year_added'] - df['release_year']

# Filter the DataFrame to include only Movies and TV Shows
df_movies = df[(df['type'] == 'Movie') & (df['release_year'].isin({2005:2021}))]
df_tv_shows = df[(df['type'] == 'TV Show') & (df['release_year'].isin({2005:2021}))]

# Group the data by year and count the number of Movies and TV Shows
# added in each year
movies_count = df_movies['year_diff'].value_counts().sort_index()
tv_shows_count = df_tv_shows['year_diff'].value_counts().sort_index()

# Create a line chart to visualize the trends over time
plt.figure(figsize=(16, 8))
plt.plot(movies_count.index, movies_count.values, color='g',
label='Movies', linewidth=2)
plt.plot(tv_shows_count.index, tv_shows_count.values, color='r',
label='TV Shows', linewidth=2)

# Customize the plot
plt.xlabel('Year_Diff')
plt.ylabel('Count')
plt.title('Movies & TV Shows Added Over Time')
plt.legend()

# Show the plot
plt.show()
```



Data-Backed Business Insights

1. **Content Diversity** Quantifiable Insight: Netflix's catalog is diversified with productions from 748 unique countries and covers a wide array of genres. The top three countries contributing to the content are the United States (2,818 titles), India (972 titles), and the United Kingdom (419 titles). Business Interpretation: This broad geographical and genre-based diversity suggests that Netflix is well-positioned to cater to a global audience with varied tastes. This is a strong asset for market penetration and customer retention.
2. **Focus on Recent Content** Quantifiable Insight: A significant chunk of Netflix's content has been released in recent years. For instance, the years 2018, 2017, and 2019 collectively account for 3,209 titles, making up approximately 36.4% of the total catalog. Additionally, the median release year for TV Shows is more recent compared to Movies. Business Interpretation: This focus on newer content likely aligns with current viewer preferences for fresh and relevant material. It also indicates that Netflix is actively keeping its content up-to-date, which is essential for maintaining subscriber interest and attracting new customers.
3. **Ratings and Target Demographic** Quantifiable Insight: The ratings 'TV-MA' and 'TV-14' dominate the content on Netflix, with 3,207 and 2,160 titles respectively. These two ratings alone make up around 61.2% of all content. Business Interpretation: The predominance of these ratings suggests that Netflix's primary target demographic is mature and teen audiences. Content strategies targeting these demographics are likely to be more successful.

Data-Backed Recommendations