Netflix Data: Cleaning, Analysis, and Visualization

Introduction

This project focuses on cleaning, analyzing, and visualizing a dataset of Netflix content spanning from 2008 to 2021. The dataset includes a wide range of movies and TV shows, with the goal of deriving insights into content distribution, popular genres, and trends over time.

Tools and Technologies

- **Programming Languages**: Python
- **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, WordCloud
- **Database**: PostgreSQL for data cleaning
- **Visualization Tool**: Power Bi(for potential future visualization)

Dataset Overview

The dataset contains information about Netflix titles, including:

- **Columns**: show_id, type, title, director, country, date_added, release_year, rating, dur ation, listed_in.
- **Content Range**: Titles added from 1925 to 2021.

Data Cleaning Process

1. **Handling Missing Values**: Identified and treated null values.
2. **Removing Duplicates**: Ensured data integrity by dropping duplicate entries.
3. **Data Type Corrections**: Converted date columns to appropriate formats.
4. **Splitting Columns**: Processed multi-value columns for better analysis.

Exploratory Data Analysis (EDA)

1. **Content Type Distribution**:
    - Analyzed the distribution of Movies vs. TV Shows.
    - Visualized using bar plots and pie charts.

2. **Most Common Genres**:
    - Split the listed_in column to count occurrences of each genre.
    - Created visualizations to display the top genres.
3. **Content Added Over Time**:
    - Extracted year and month from the date_added column.
    - Visualized trends in content addition over the years.
4. **Top Directors with Most Titles**:
    - Identified directors with the highest number of titles.
    - Visualized results using bar charts.
5. **Word Cloud of Movie Titles**:
    - Generated a word cloud to visualize the frequency of movie titles

Key Insights:

- The dataset revealed trends in content addition over time, highlighting significant growth in specific genres.
- Popular genres were identified, providing insights into viewer preferences.
- The analysis also pointed out key directors contributing to Netflix's library.

Next Steps:

1. **Feature Engineering**: Explore additional features such as genre counts per title.
2. **Machine Learning Applications**: Utilize cleaned data for predictive analytics or recommendation systems.
3. **Advanced Visualization**: Consider creating interactive dashboards for deeper insights.

Conclusion

This project serves as a foundational exercise in data cleaning and visualization techniques using Python and its libraries. The insights gained can inform future content strategies for streaming services like Netflix.

Steps Done Using Jupyter Notebook

Step1: Import Required Libraries

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

Step2: Load the Dataset

```
data = pd.read_csv('netflix1.csv')
        print(data.head())
```

```
  show_id     type                                title         director  \
0      s1    Movie              Dick Johnson Is Dead  Kirsten Johnson
1      s3  TV Show                          Ganglands  Julien Leclercq
2      s6  TV Show                      Midnight Mass    Mike Flanagan
3     s14    Movie  Confessions of an Invisible Girl    Bruno Garotti
4      s8    Movie                            Sankofa     Haile Gerima

         country  date_added  release_year rating  duration  \
0  United States   9/25/2021          2020  PG-13    90 min
1         France   9/24/2021          2021  TV-MA  1 Season
2  United States   9/24/2021          2021  TV-MA  1 Season
3         Brazil   9/22/2021          2021  TV-PG    91 min
4  United States   9/24/2021          1993  TV-MA   125 min

                                          listed_in
0                                    Documentaries
1  Crime TV Shows, International TV Shows, TV Act...
2                 TV Dramas, TV Horror, TV Mysteries
3                 Children & Family Movies, Comedies
4   Dramas, Independent Movies, International Movies
5
```

Step 3: Data Cleaning

```
print(data.isnull().sum())
```

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

```
data.drop_duplicates(inplace=True)
```

```
data.dropna(subset=['director',   'country'], inplace=True)
```

```
data['date_added'] = pd.to_datetime(data['date_added'])
```

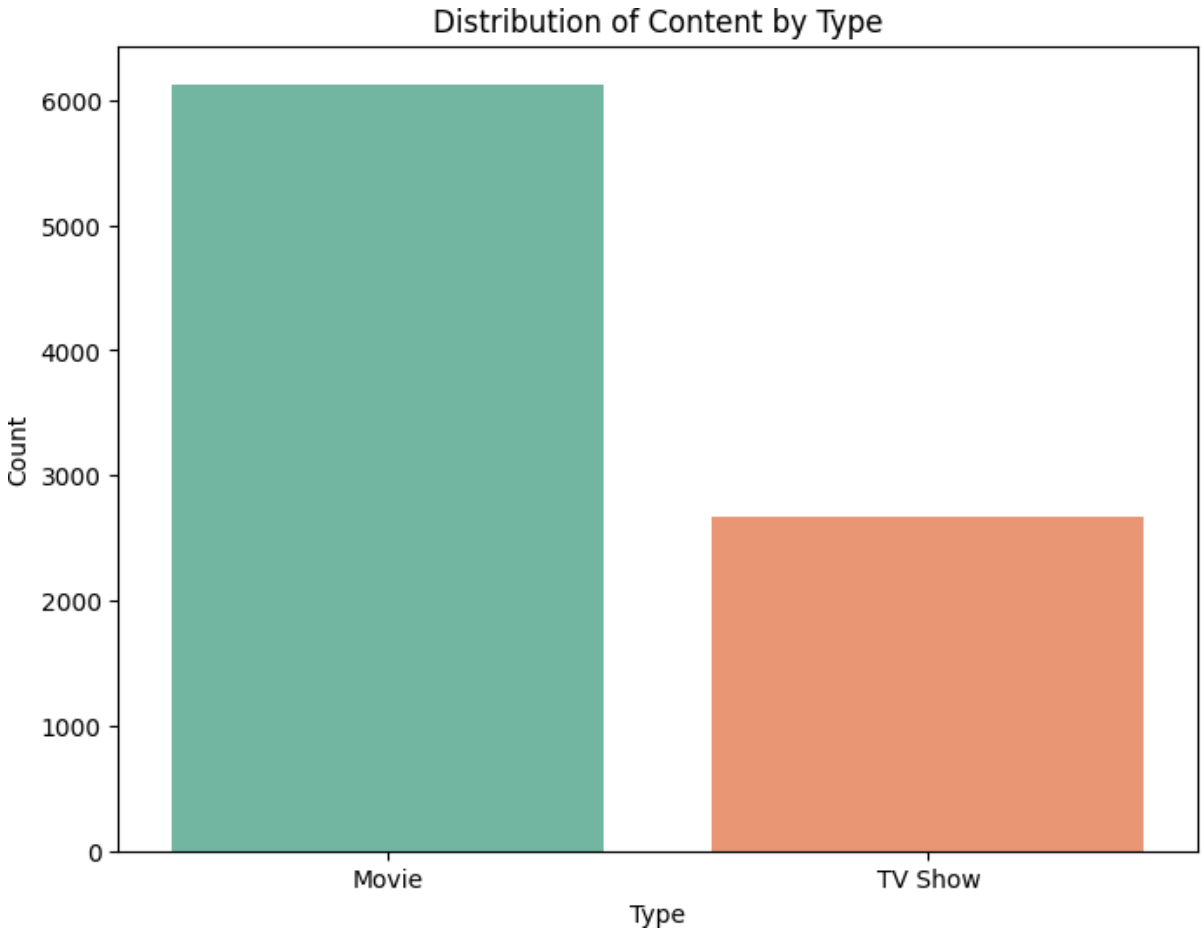
```
print(data.dtypes) print(data.info())
```

```
show_id                  object
type                     object
title                    object
director                 object
country                  object
date_added       datetime64[ns]
release_year              int64
rating                   object
duration                 object
listed_in                object
dtype: object
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   datetime64[ns]
 6   release_year  8790 non-null   int64
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: datetime64[ns](1), int64(1), object(8)
memory usage: 686.8+ KB
None
```

Step 4: Exploratory Data Analysis (EDA)

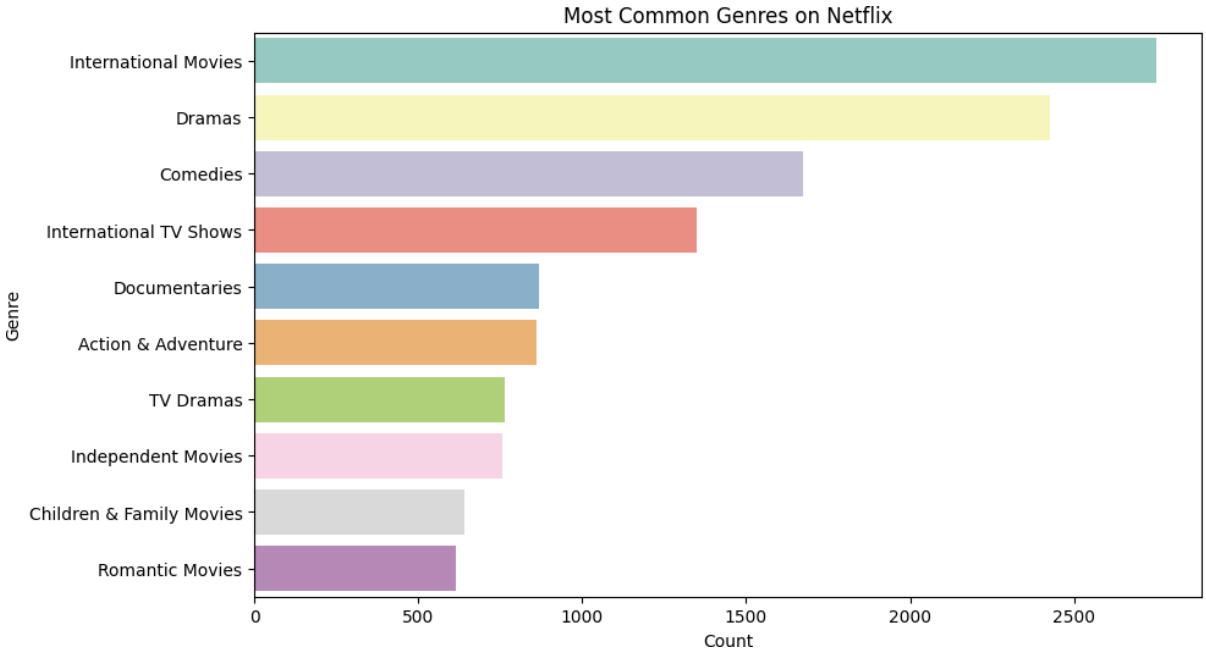
1. Content Type Distribution (Movies vs. TV Shows)

```
type_counts = data['type'].value_counts()
plt.figure(figsize=(8, 6))
sns.barplot(x=type_counts.index, y=type_counts.values, hue=type_counts.index, palet
plt.title('Distribution of Content by Type')
plt.xlabel('Type')
plt.ylabel('Count')
plt.show()
```



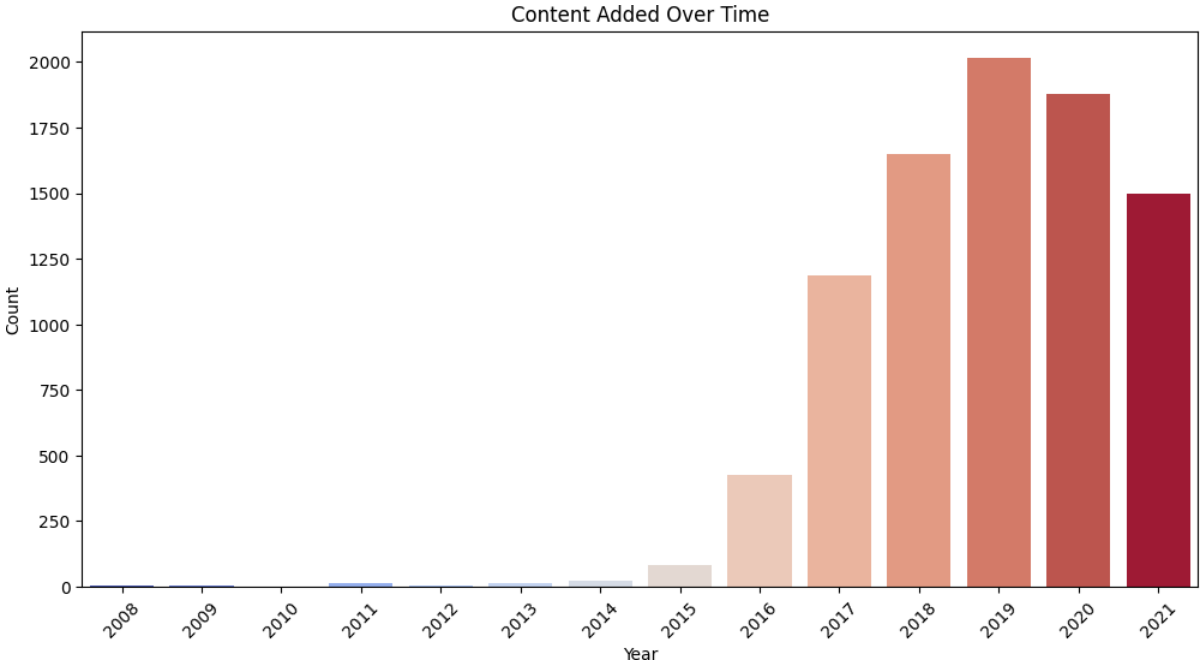
2. Most Common Genres

```
data['genres'] = data['listed_in'].apply(lambda x: x.split(', '))
all_genres = sum(data['genres'], [])
genre_counts = pd.Series(all_genres).value_counts().head(10)

plt.figure(figsize=(10, 6))
sns.barplot(x=genre_counts.values, y=genre_counts.index, hue=genre_counts.index, pa
plt.title('Most Common Genres on Netflix')
plt.xlabel('Count')
plt.ylabel('Genre')
plt.show()
```



4. Content Added Over Time

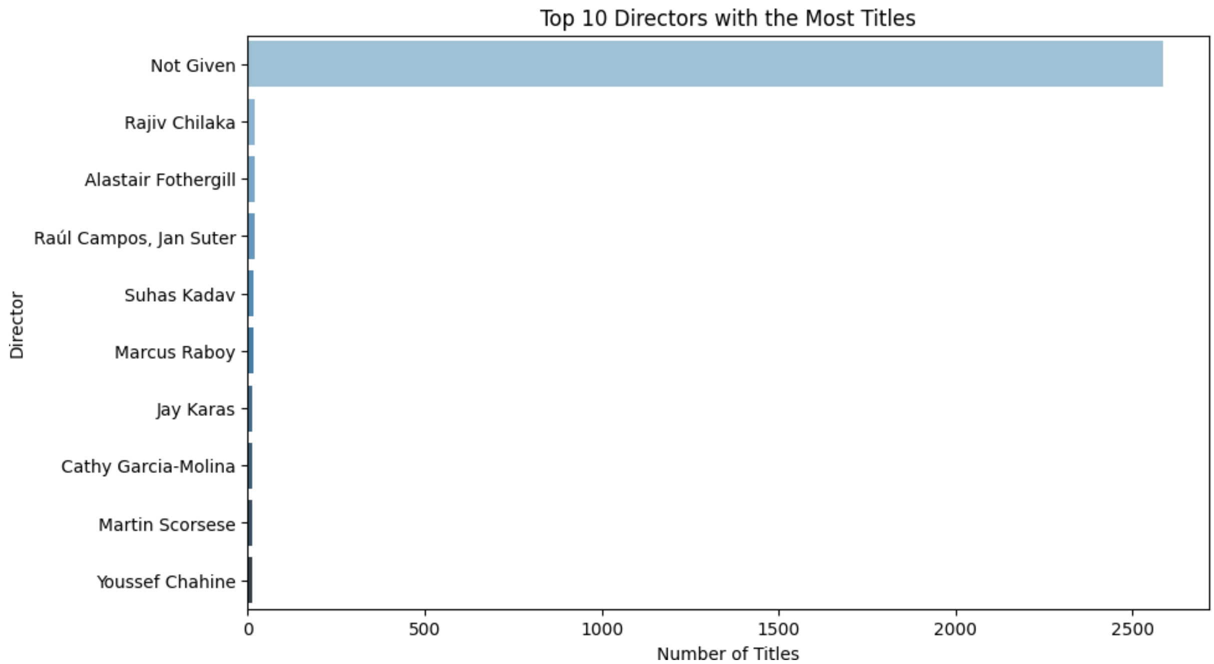
```
data['year_added'] = data['date_added'].dt.year

plt.figure(figsize=(12, 6))
sns.countplot(x='year_added', data=data, hue='year_added', palette='coolwarm', lege
plt.title('Content Added Over Time')
plt.xlabel('Year')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```



4. Top 10 Directors with the Most Titles

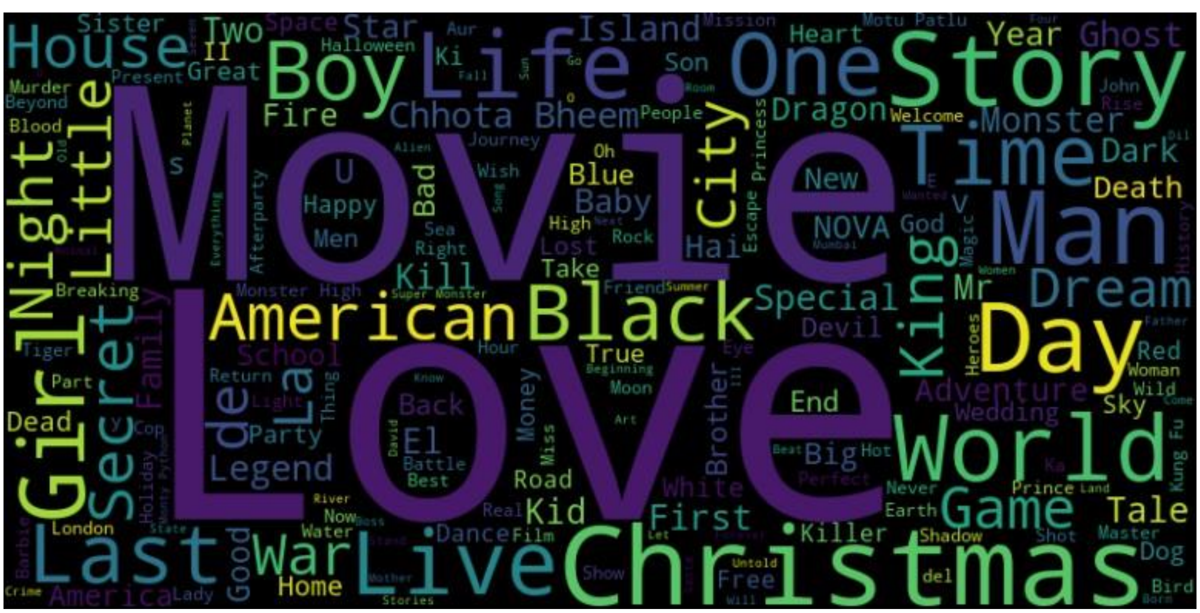
```
top_directors = data['director'].value_counts().head(10)

plt.figure(figsize=(10, 6))
sns.barplot(x=top_directors.values, y=top_directors.index, hue=top_directors.index,
```

```
plt.title('Top 10 Directors with the Most Titles')
plt.xlabel('Number of Titles')
plt.ylabel('Director')
plt.show()
```



5. Word Cloud of Movie Titles

```
movie_titles = data[data['type'] == 'Movie']['title']
wordcloud = WordCloud(width=800, height=400, background_color='black').generate('  '

plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



Netflix Dashboard

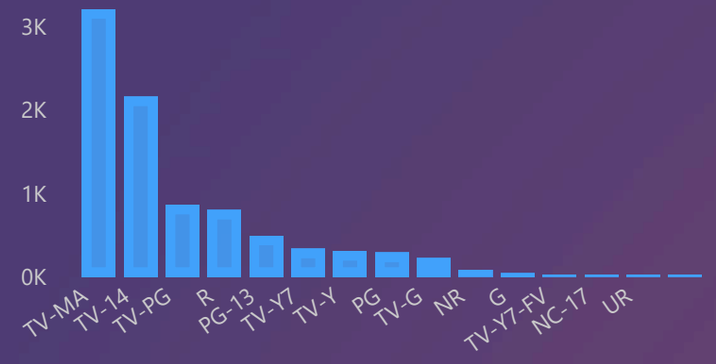
Type

Movie | TV Show

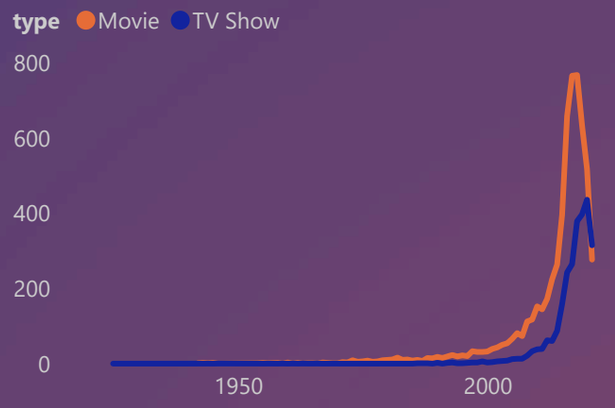
Release_Year: All

Rating: All

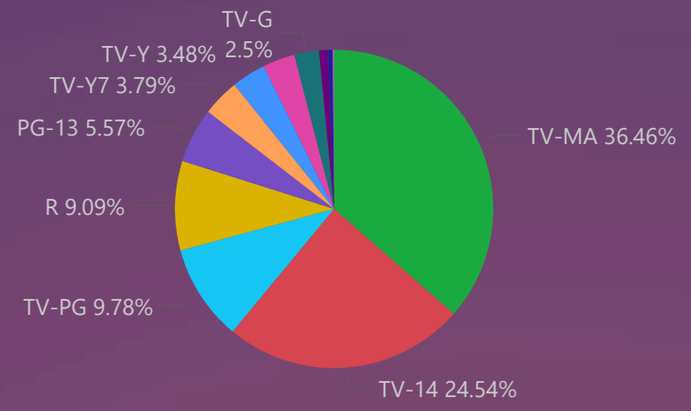
Rating on Netflix



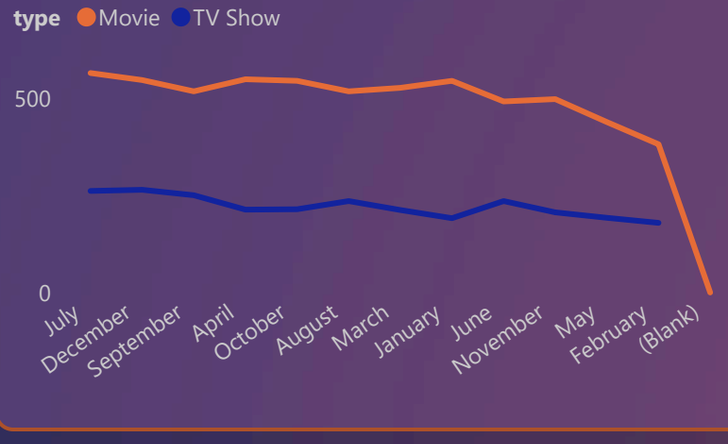
Yearly Released Movies and Tv Shows



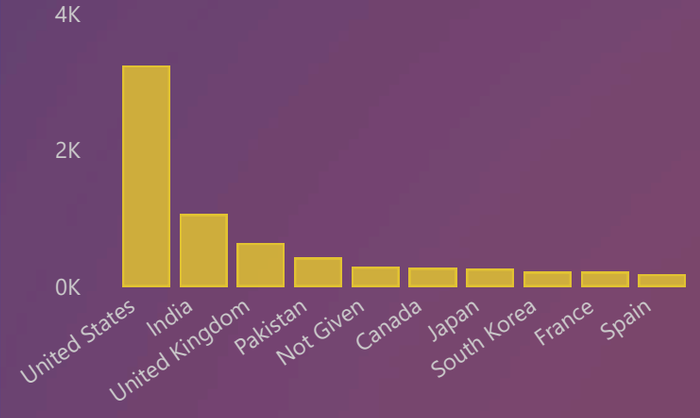
Rating On Netflix



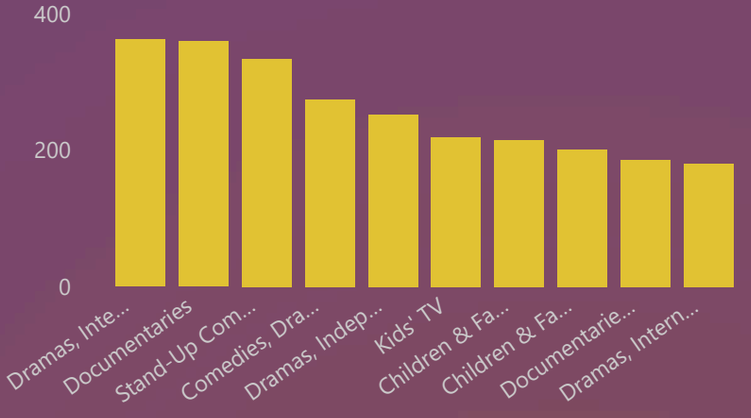
Montly Released Movies and Tv Shows

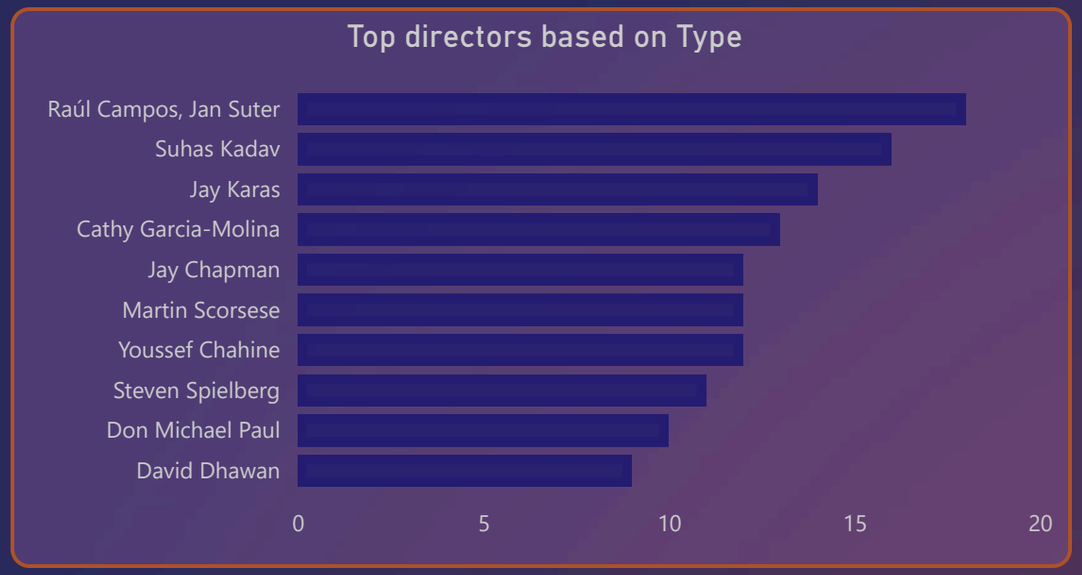


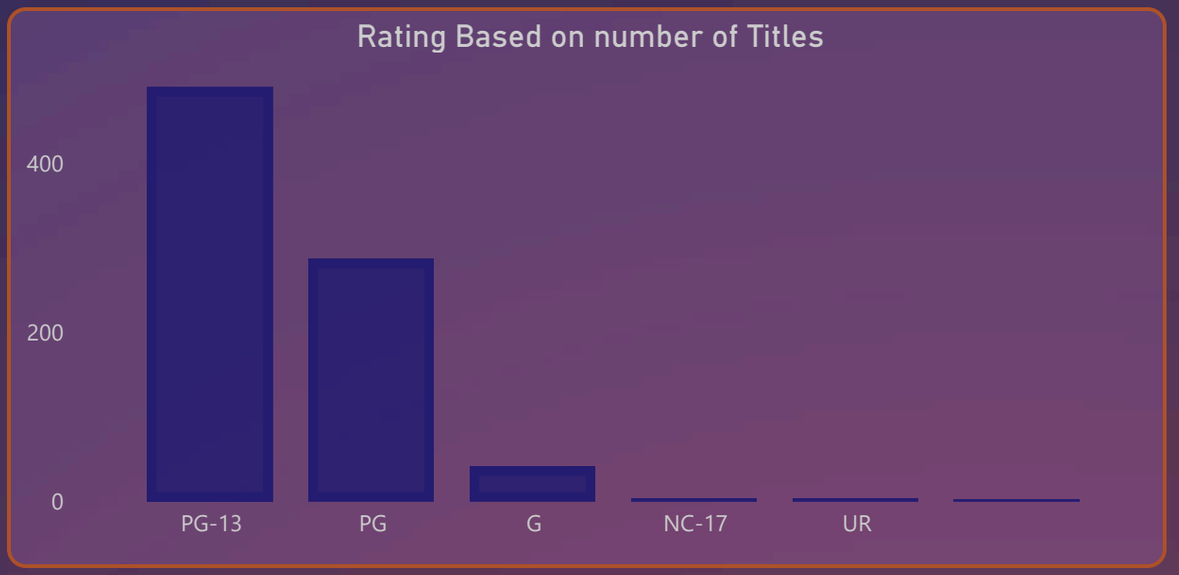
Top 10 Country with most Content

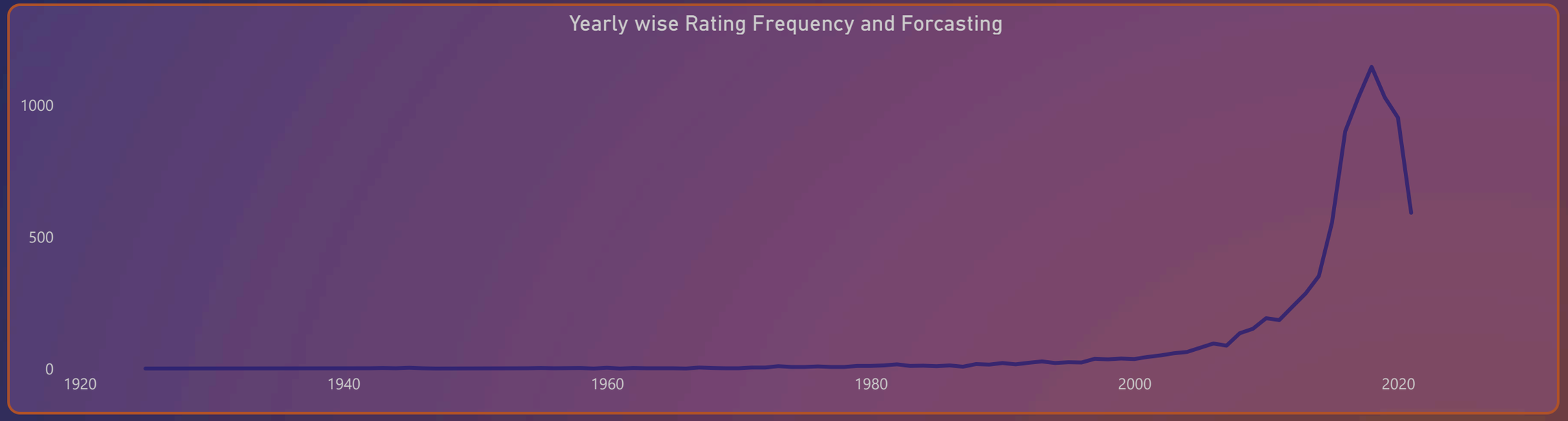


Top 10 Popular Genres




Top directors based on Type
Raúl Campos, Jan Suter
Suhas Kadav
Jay Karas
Cathy Garcia-Molina
Jay Chapman
Martin Scorsese
Youssef Chahine
Steven Spielberg
Don Michael Paul
David Dhawan
0
5
10
15
20


Rating Based on number of Titles
400
200
0
PG-13
PG
G
NC-17
UR


Yearly wise Rating Frequency and Forcasting
1000
500
0
1920
1940
1960
1980
2000
2020


the
of
A
Love
and
with
School
Days
City
Rise
Club
Marvel
Christmas
All
Things
Out
Explained
Miss
Seven
Away
Age
Universe
Prince
Beyblade
Rescue
Bheem
Team
Is
Space
America
Kingdom
African
Rangers
Go
Magic
Dragon
Game
Luna
Legend
no
Way
Online
Friends
From
by
Dangerous
Girl
Ghost
Show
Ever
Secret
Tales
My
D
Story
Hotel
Battle
An
Last
Legends
New
Food
France
Transformers
Made
Secrets
Day
Back
How
Stories
British
Great
Life
Beyond
At
Black
Series
First
Files
Kill
Super
For
World
Code
Nailed
Inside
Events
to
In
de
El
del
la
Oh
Behind
Force
Live
Us
Good
Deep
Planet
Get
Crime
Case
Monsters
Adventures
Queen
American
Evil
100
House
Dream
Squad
Mind
Man
Street
Criminal
Heart
Home
Family
Academy
High
on
Park
Blood
Dad
You
Girls
Bad
Two
King
Do
Empire
Art
Time
War
vs
That
Chronicles
Chef
History
Your
Lost
Be
Bangkok
1
Power
Mighty
Heaven
Kids
We
Big
This
Happy
Me
Call
LEGO
Cinta
Murder
Dark
Earth
Burns
Comedy
Nature
Mexico
Boys
Monty
Holiday
DreamWorks
Boy
Little
Money
Mr
Masters
Baby
Night
Diaries
Dino
True
Friday
Top
One
Up
or
Next
Killer
Best
Fate
It
People
Ken
Dead
Death