Home About

Netflix Content Analysis Using Python

NETFLIX

Tools Used: Python, pandas, Matplotlib, Seaborn

-Presented by : SANA KHAN

ABOUT NETFLIX

Netflix is an American streaming service founded in 1997 by Reed Hastings and Marc Randolph. It started as a DVD rental service and transitioned into a global leader in video- on-demand streaming. Today, Netflix offers a vast library of films, documentaries, TV shows, and original productions across various genres and languages, reaching millions of subscribers worldwide.

PROJECT OBJECTIVES

This project aims to:

Determine the balance between Movies and TV Shows

Track content volume added over the years and months

Identify top content-producing countries

Explore the relationship between content types and ratings

Discover the most common genres on Netflix

Analyze the most prolific directors

Calculate the average duration of Movies and TV Shows

PROBLEM STATEMENTS

1. What is the distribution of Movies vs TV Shows on Netflix?

2. How has content addition changed over the years?

3. Which countries produce the most content on Netflix?

4. How are ratings distributed across content types?

5. Who are the most prolific directors?

6. What are the most popular genres?

7. What is the trend of monthly and yearly content additions?

8. What is the average duration of Movies and TV Shows?

TOOLS AND LIBRARIES USED

Python Pandas Matplotlib Seaborn jupyter Nootbook

Python: For data handling and scripting

pandas: For data manipulation and analysis

Matplotlib & Seaborn: For data visualization

Jupyter Notebook: For running the project in an interactive environment

DATASET OVERVIEW

Import the required libraries and load the dataset for finding missing values

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

# Display settings
pd.set_option('display.max_columns', None)
```

```
# Load the dataset
data = pd.read_csv("netflix1.csv")

# Display first few rows
data.head()
```

Dataset Source: Provided by Unified Mentor as part of internship project

📥 Load the Dataset

Let's load the Netflix dataset and check its basic structure, including missing values.

```
[ ]: # Load the dataset
data = pd.read_csv("netflix1.csv")

# Display first few rows
data.head()
```

:

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Dramas, Independent Movies, International

Data cleaning

```
print(data.columns)
```

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

Checked for null values and handled missing data appropriately

Removed duplicates from the dataset

Converted date formats for date_added

Split duration into numeric minutes and seasons for better analysis

```
]: data.info()
```

```
<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
```

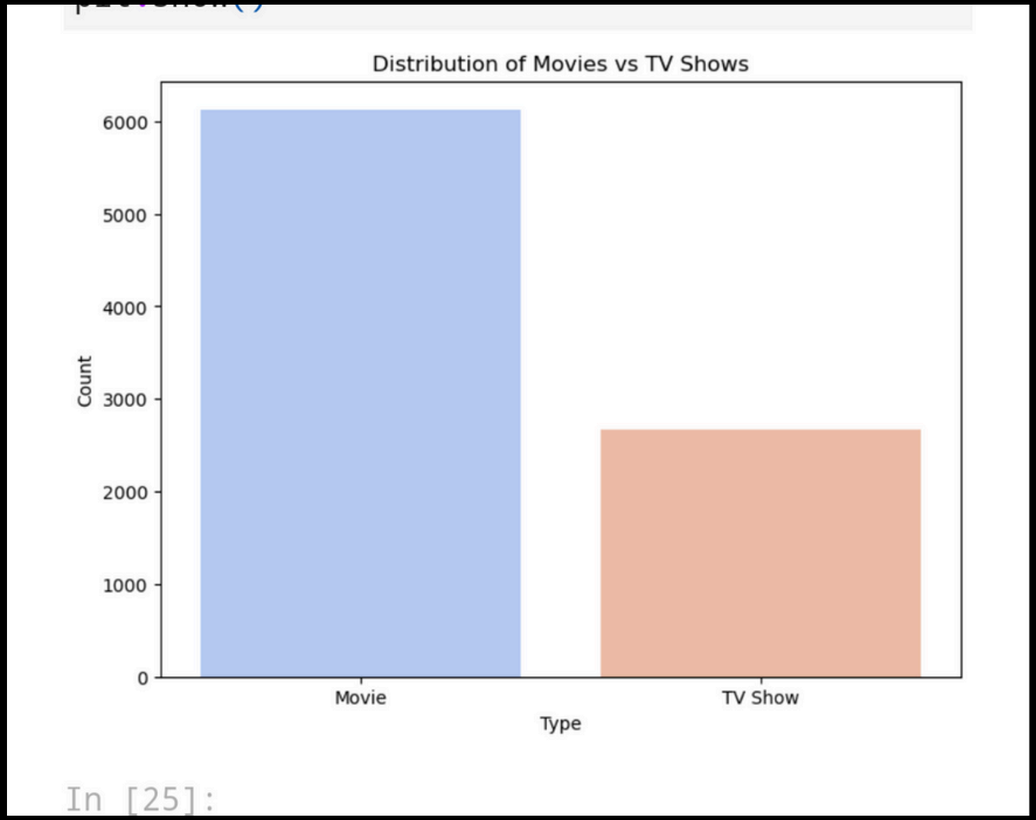
EXPLORATORY DATA ANALYSIS

```
]: plt.figure(figsize=(8, 6))
   sns.barplot(x=type_counts.index, y=type_counts.values, palette='coolwarm', errorbar=None)
   plt.title('Distribution of Movies vs TV Shows')
   plt.xlabel('Type')
   plt.ylabel('Count')
   plt.show()
```

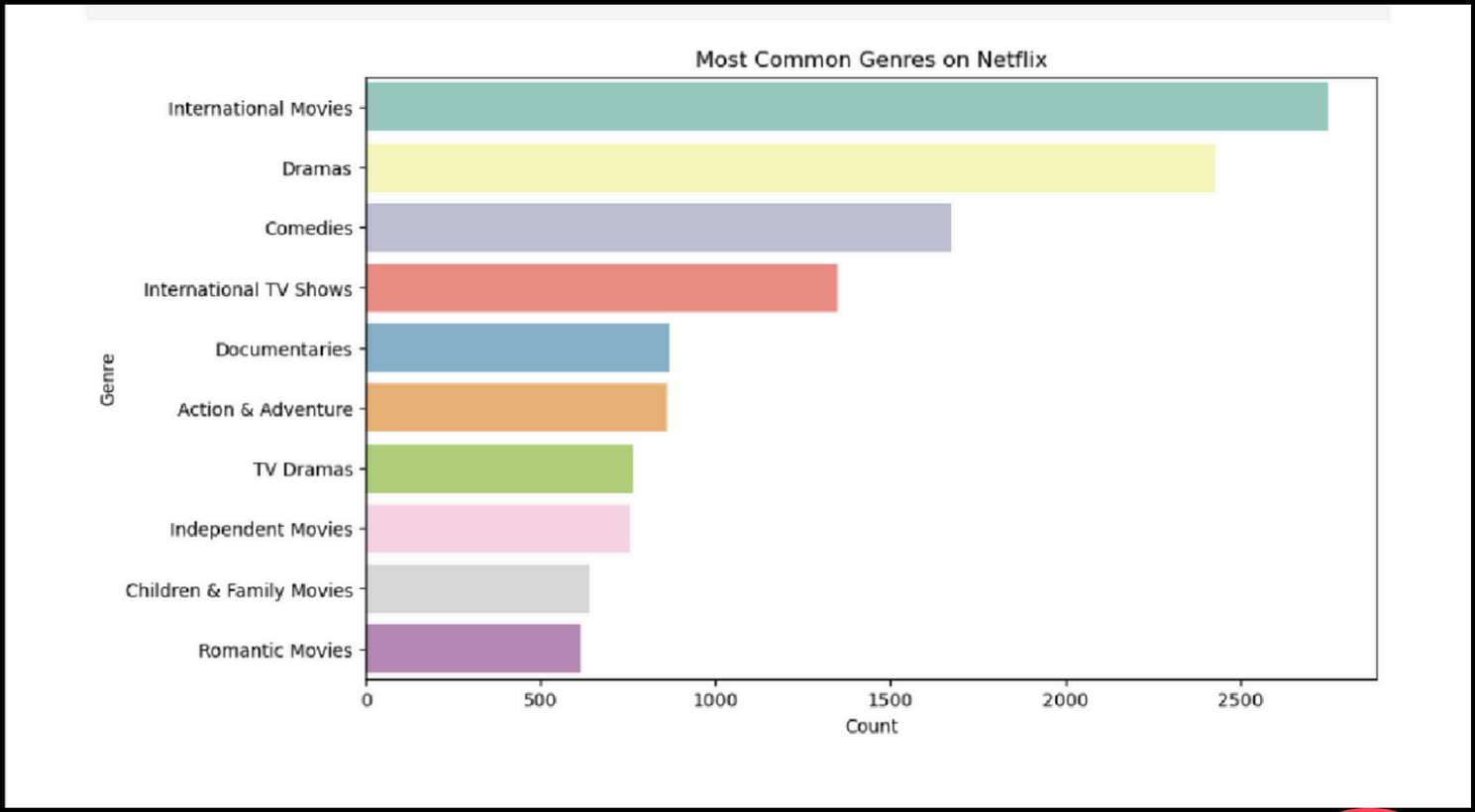
```
[25]: # Split 'listed_in' column into genres
      data['genres'] = data['listed_in'].apply(lambda x: x.split(', ') if isinstance(x, str) else [])

      # Flatten list and count genres
      all_genres = sum(data['genres'], [])
      genre_counts = pd.Series(all_genres).value_counts().head(10)

      # Plot
      plt.figure(figsize=(10, 6))
      sns.barplot(y=genre_counts.index,
                  x=genre_counts.values,
                  palette='Set3',
                  hue=genre_counts.index,
                  legend=False)
      plt.title('Most Common Genres on Netflix')
      plt.xlabel('Count')
      plt.ylabel('Genre')
      plt.show()
```

Movie vs TV Show Distribution

Majority of the content on Netflix consists of Movies
TV Shows form a smaller proportion



In [25]:



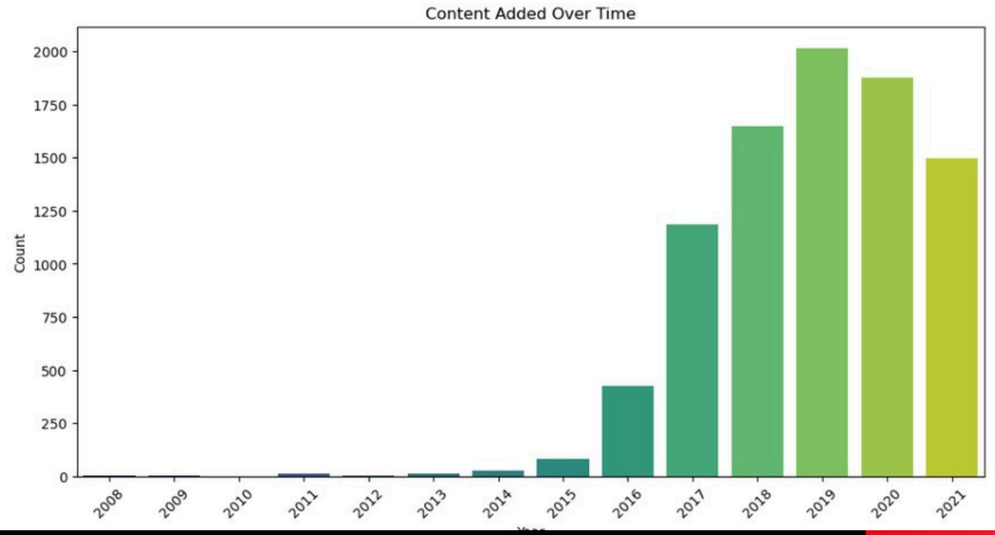
CONTENT ADDED OVER THE YEARS

- **Significant growth in content after 2015**

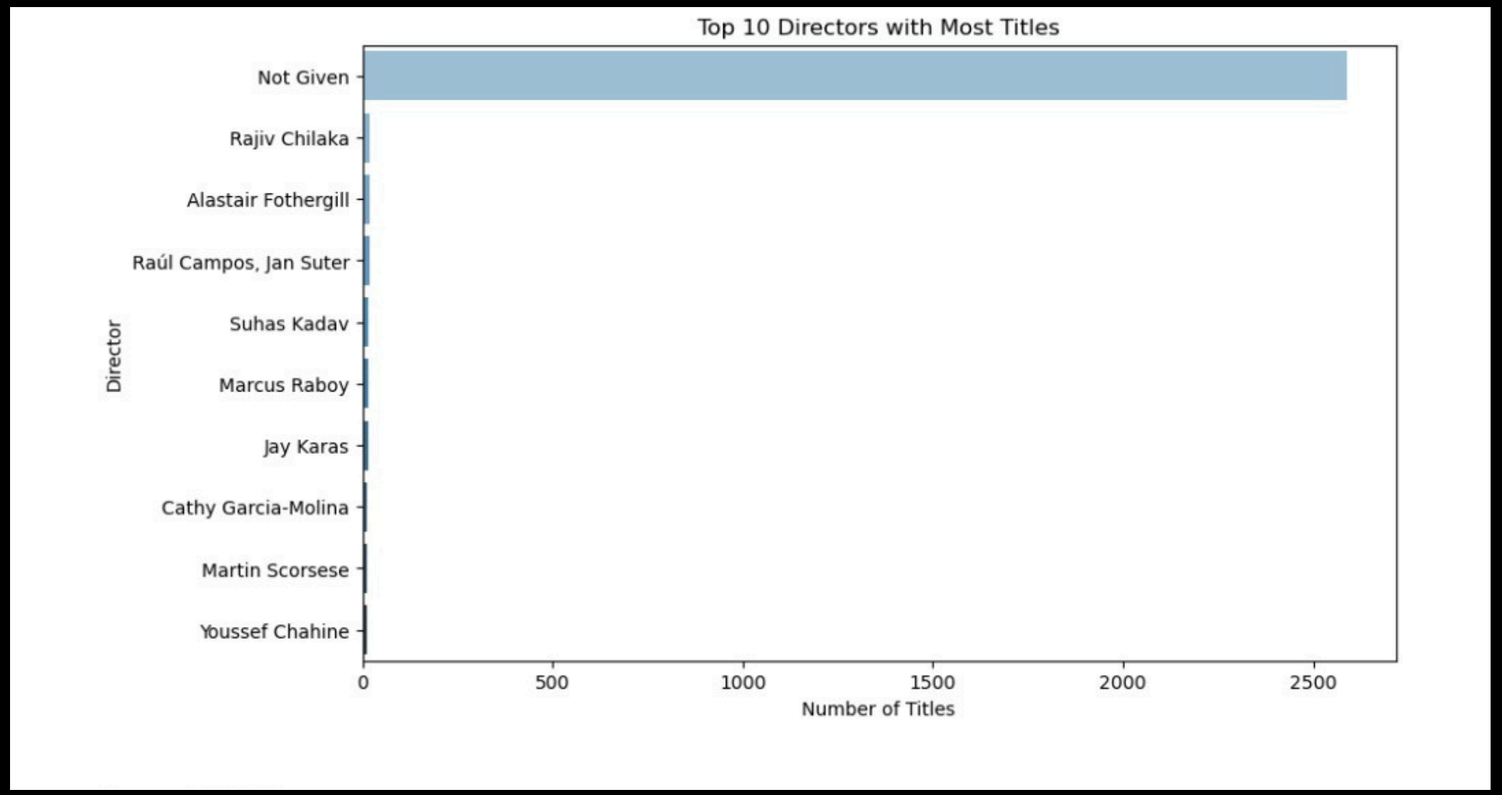
- **Peaks observed in 2018 and 2020**

- **Drop in additions after 2020 possibly due to pandemic-related slowdowns**

```
# Convert 'date_added' to datetime format
data['date_added'] = pd.to_datetime(data['date_

# Extract year from 'date_added'
data['year_added'] = data['date_added'].dt.yea

# Plot content added over the years
plt.figure(figsize=(12, 6))
sns.countplot(x='year_added',
              data=data,
              palette='viridis')
plt.title('Content Added Over Time')
plt.xlabel('Year')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```



Top Prolific Directors on Netflix



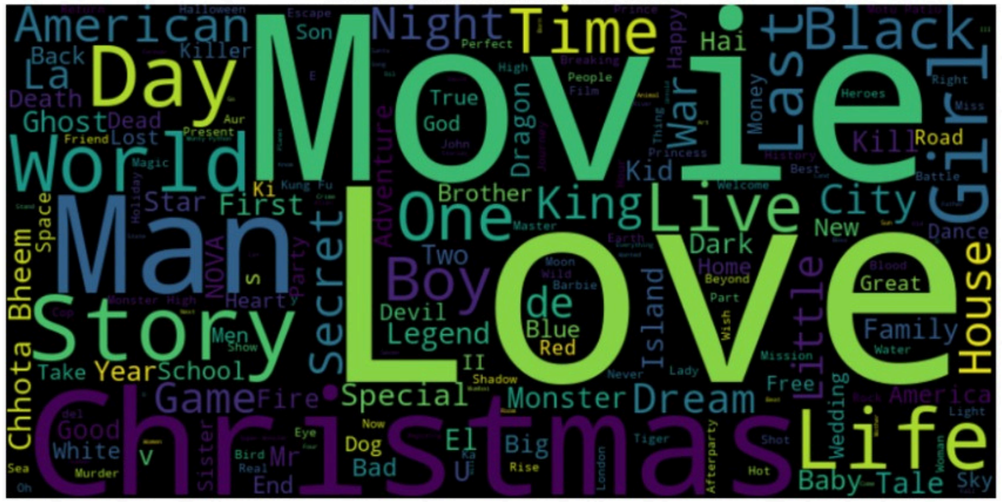
Rajiv Chilaka leads with the highest number of directed titles

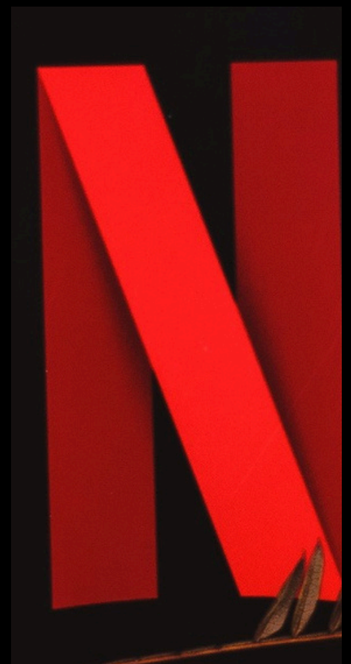
Other notable names include Alastair Fothergill, Raúl Campos, and Martin Scorsese

Word Cloud of Movie Titles

```
# Generate word cloud
movie_titles = data[data['type'] == 'Movie']['t
wordcloud = WordCloud(width=800, height=400, ba

# Plot word cloud
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud of Netflix Movie Titles')
plt.show()
```

Word Cloud of Netflix Movie Titles





KEY INSIGHTS

Movies make up a larger share of Netflix’s content

TV-MA is the most common rating

Content additions peaked around 2018–2020

July sees the highest number of content additions

USA, India, and UK dominate content production

Top genres include Drama, Documentaries, and Comedy

Average movie duration is ~99 minutes

Conclusion & Learnings

Learned to analyze and clean real-world datasets using pandas

Created visualizations to uncover trends in Netflix content

Understood how to explore categorical and time-based data

Gained hands-on experience in presenting data-driven insights

Contact

See You Next

THANK YOU

2025 Data Analysis Presentation