Netflix Data Cleaning, Analysis and Visualization

Presented by: Yash Sahu

Table of Contents

1. Introduction & Problem Statement
2. Dataset Description
3. Data Preprocessing
4. Exploratory Data Analysis (EDA) & Insights
5. Data Visualization
6. Conclusion

Introduction & Problem Statement

Introduction

- The project aims to analyze and visualize Netflix data to uncover trends in content distribution, genres, and popularity.

Problem Statement

- How is Netflix’s content distributed by year, genre, and country?
- What are the most common content types and trends over time?

Dataset Description

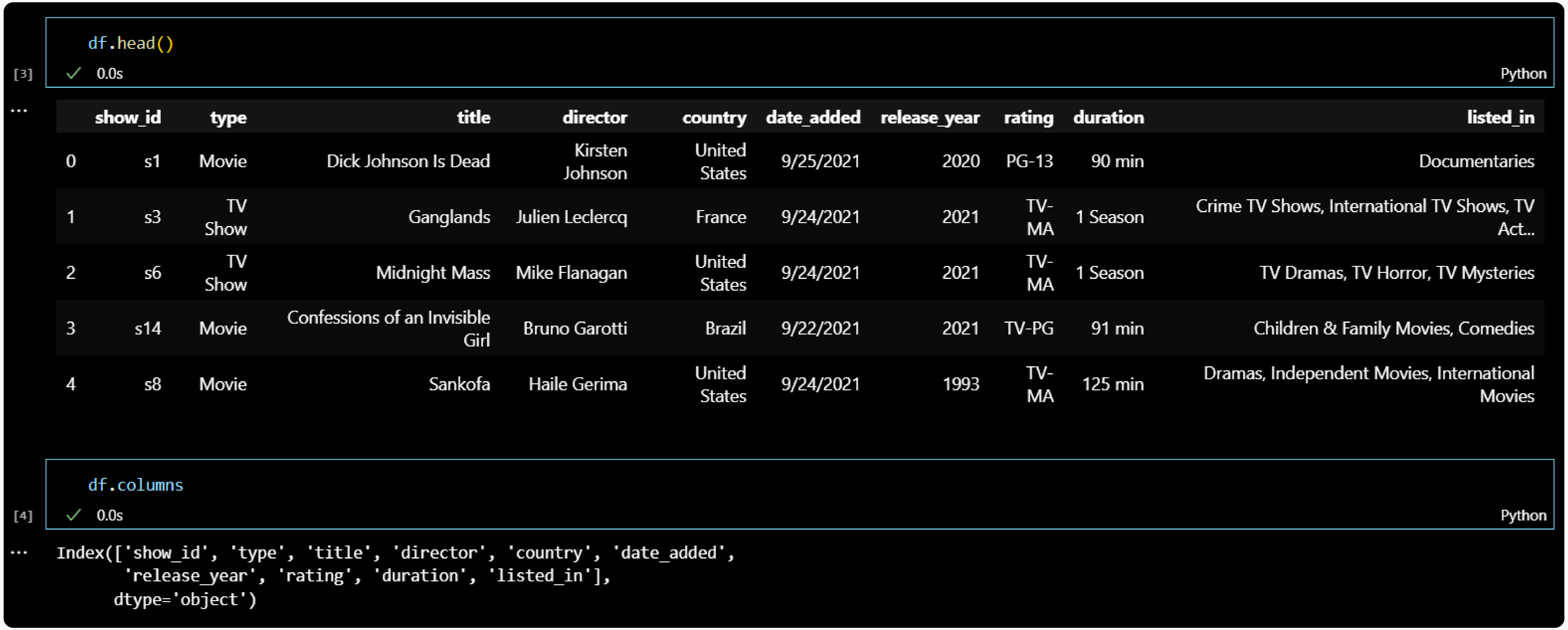
Overview

- The dataset contains information on Netflix movies and TV shows, including:
- Title, Genre, Release Year, Country, Duration and Ratings.

Data Insights

- Total Records: (8790 rows and 10 columns)
- Missing Values: (No missing data)

Dataset Description



```
[3]  df.head()
     ✓ 0.0s                                                          Python
```

	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 Movies

```
[4]  df.columns
     ✓ 0.0s                                                          Python
```

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

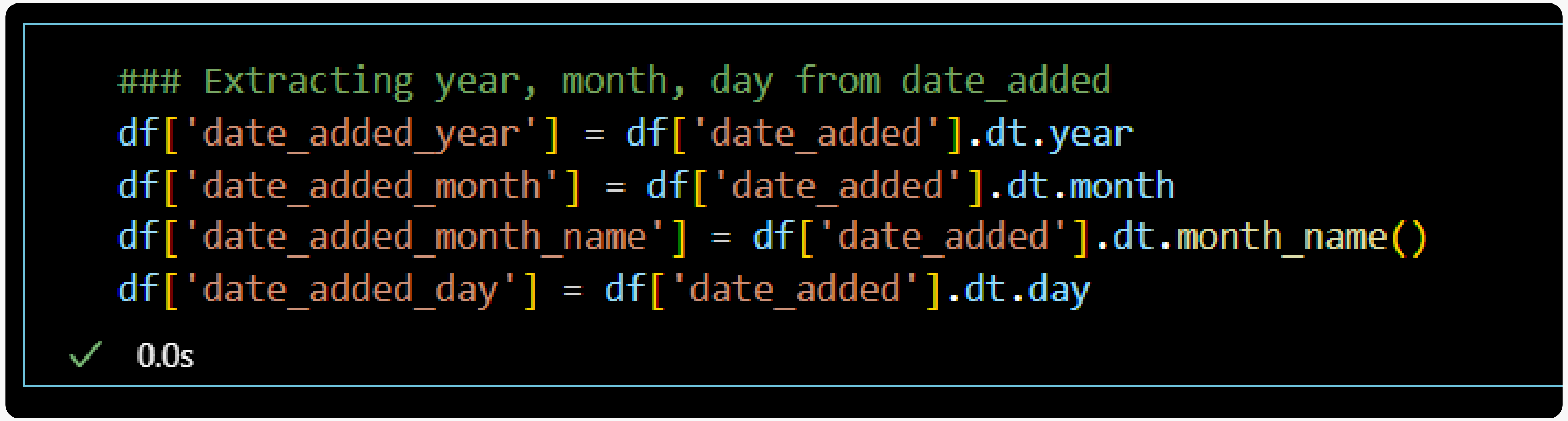
This is the output of df.head() first five rows of dataset.

List of columns in dataset.

Data Preprocessing

Steps:

- Handling Missing Values: Replaced or dropped missing values in relevant columns.
- Date Formatting: Converted release dates into a standardized format for analysis using pandas.to_datetime()
- Feature Engineering: Extracted new features (e.g., Year and Month of Release).



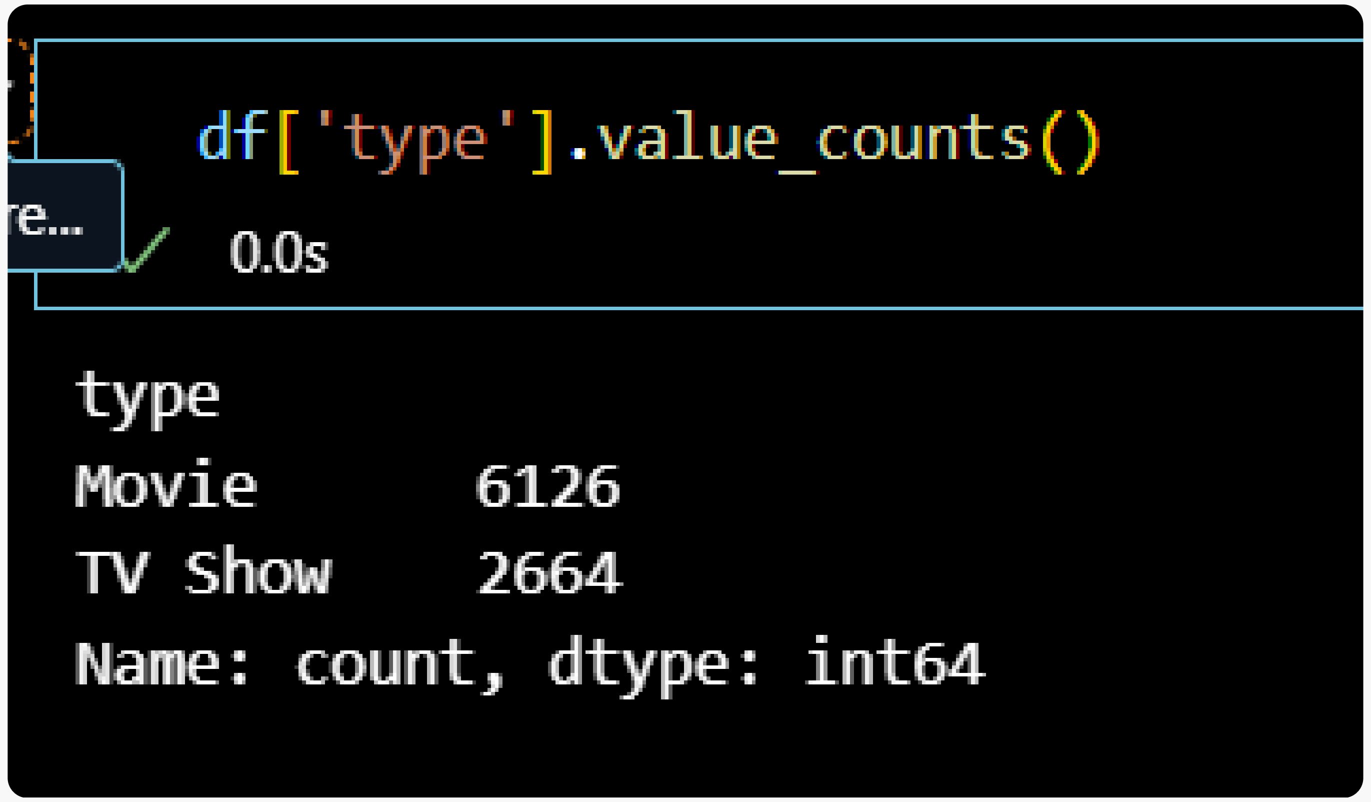
```
### Extracting year, month, day from date_added
df['date_added_year'] = df['date_added'].dt.year
df['date_added_month'] = df['date_added'].dt.month
df['date_added_month_name'] = df['date_added'].dt.month_name()
df['date_added_day'] = df['date_added'].dt.day
```

✓ 0.0s

Exploratory Data Analysis (EDA) & Insights

Key Questions Explored:

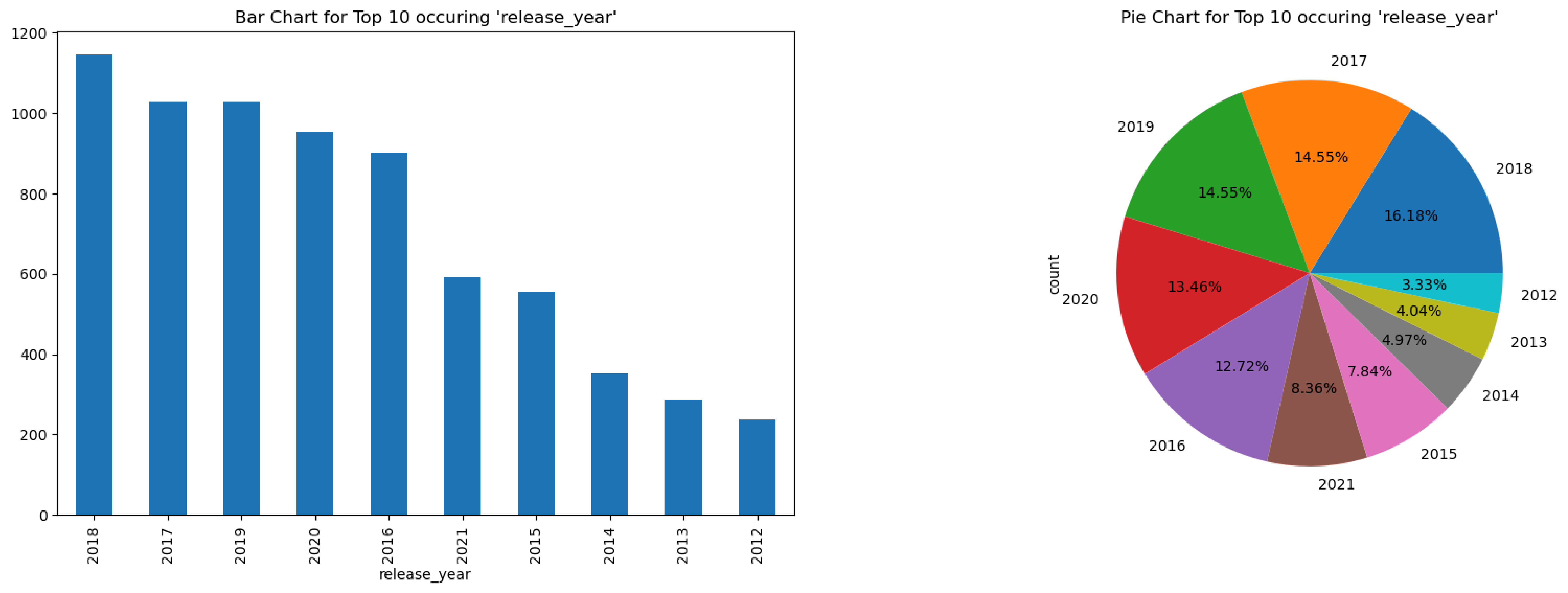
- What is the distribution of movies vs. TV shows?
- What are the most popular genres on Netflix?
- Which countries contribute the most content?
- How has content production changed over the years?



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

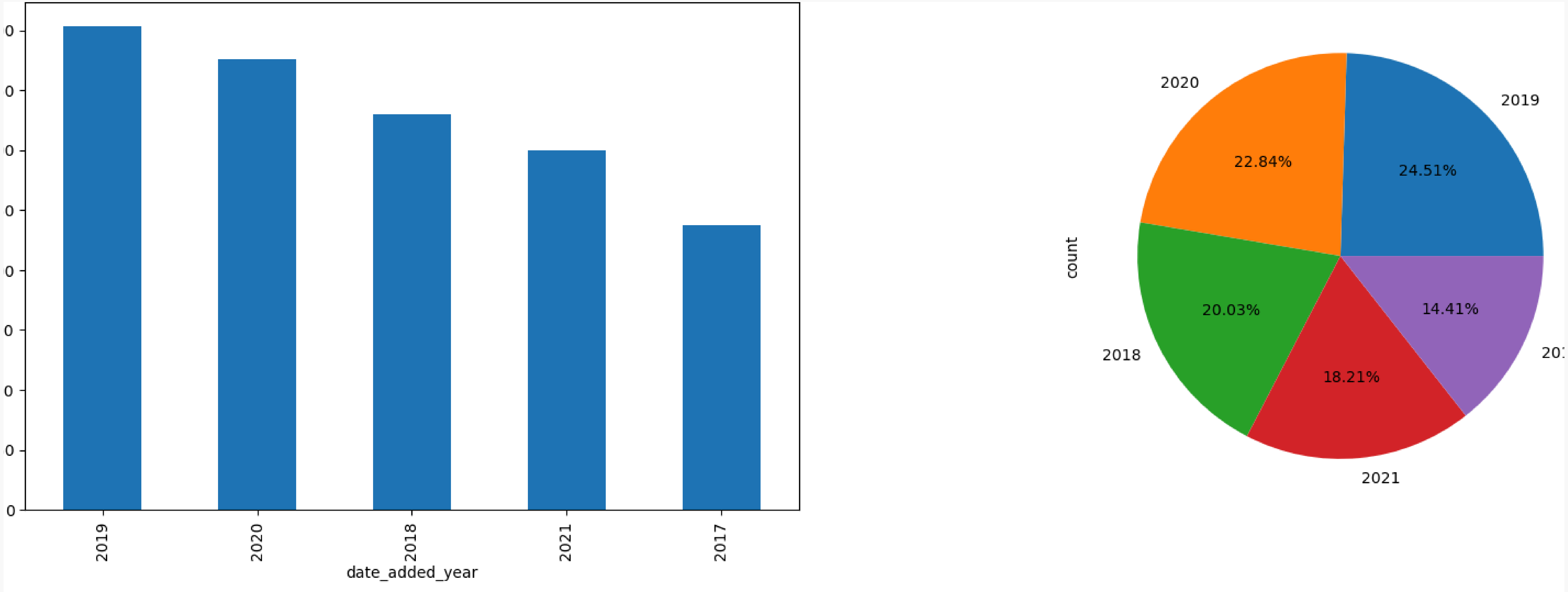
```
type
Movie      6126
TV Show    2664
Name: count, dtype: int64
```

Data Visualization



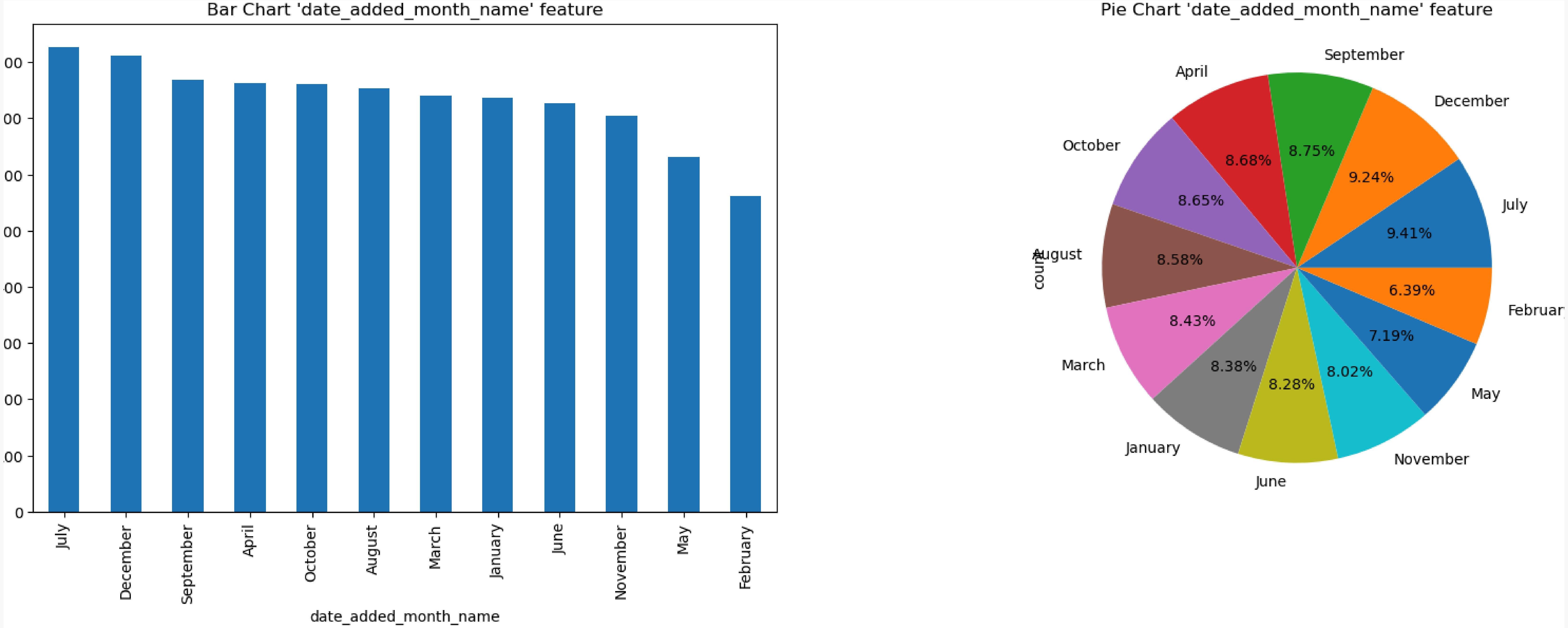
Bar chart and Pie chart for Top 10 years in ‘release_year’

Data Visualization



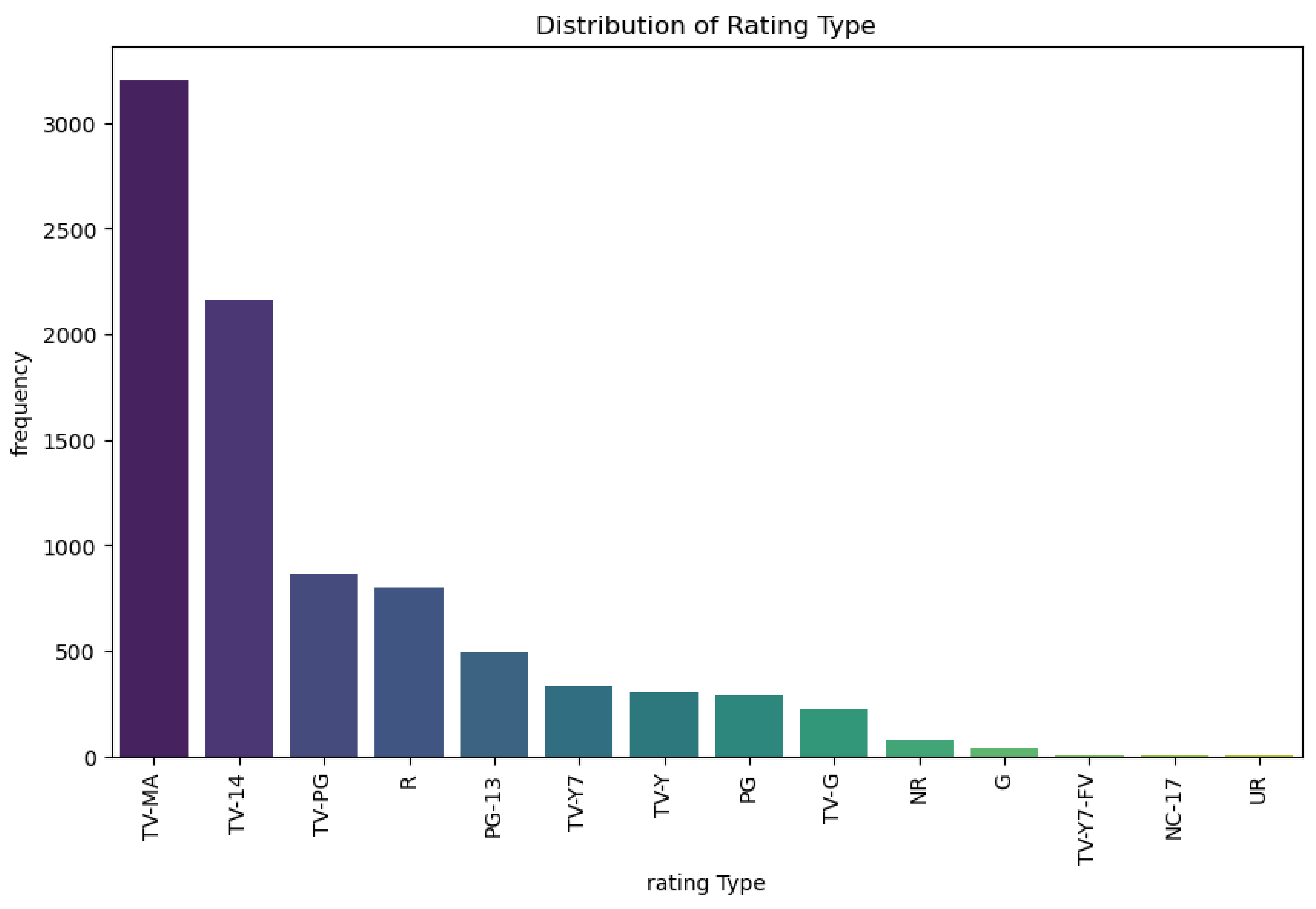
Distribution of year in date_added feature

Data Visualization



Distribution of month in date_added feature

Data Visualization



Conclusion

Conclusion:

- This analysis provided insights into Netflix’s content trends, genre distribution, and country-wise contributions.
- Visualizations helped in understanding data patterns and trends over time.

Thank You!