CAPSTONE PROJECT

NETFLIX ORIGINAL

INTRODUCTION

In the era of digital entertainment, understanding viewer preferences, content trends, and platform strategies is essential. This project delves into a comprehensive analysis of Netflix's dataset to explore the dynamics of movies and TV shows on one of the world's leading streaming platforms.

Netflix's dataset offers a wealth of information on content type, release years, ratings, and regional availability. Our goal is to analyze these data points, extract meaningful insights, and uncover trends that define Netflix's vast library.

Through our exploration, we aim to identify key content trends, popular genres, and factors influencing content distribution. This will provide a deeper understanding of what drives audience engagement and the evolution of digital streaming.

This report highlights our journey into Netflix data analytics, presenting valuable insights that reflect the streaming giant's content strategies and audience preferences in an ever-evolving entertainment industry.

OBJECTIVE

Content Insights:

Analyze the distribution of movies and TV shows, identifying trends in content production across different years and genres.

Regional Analysis:

Identify the top countries producing Netflix content and understand regional content preferences.

User Preference Trends:

Examine content ratings, genres, and release patterns to understand audience preferences.

Growth Over Time:

Study the trend of content additions over the years to assess Netflix's expansion strategy.

Strategic Decision-Making:

Provide data-driven insights to help Netflix optimize its content strategy, improve recommendations, and enhance user engagement.

Libraries Used for Netflix Data Analysis

1. **NumPy (import numpy as np)** – Supports numerical operations with multi-dimensional arrays and mathematical functions.
2. **Pandas (import pandas as pd)** – Used for data manipulation and analysis with structured data like tables.
3. **Seaborn (import seaborn as sns)** – Simplifies statistical data visualization with aesthetically pleasing charts.
4. **Matplotlib (import matplotlib.pyplot as plt)** – A versatile 2D plotting library for creating static visualizations.
5. **Plotly Express (import plotly.express as px)** – Enables interactive web-based visualizations with minimal code.
6. **Warnings (import warnings**; warnings.filterwarnings('ignore')) – Suppresses unnecessary warning messages for a cleaner output.

DATA OVERVIEW

Loading Dataset

```
# loading the dataset
df = pd.read_csv("C:\\Users\\Himani Sharma\\OneDrive\\Desktop\\projects skillcircle\\netflix_titles.csv")
```

loads the dataset netflix_titles.csv from the specified file path using the pandas library. It reads the CSV file into a DataFrame (df), which is then used for data analysis and visualization in the Netflix content analysis project.

Analysing Data structure

```
#inspecting the structure of the dataset
df.head(10)
```

	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	NaN	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	NaN	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...	NaN	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	NaN	NaN	NaN	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	NaN	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...

The dataset contains 8,807 entries of Movies and TV Shows available on Netflix, with 12 attributes such as show ID, type, title, director, cast, country, date added, release year, rating, duration, listed genres, and description. Some columns like director, cast, and country have missing values. Movies have duration in minutes, whereas TV Shows are measured in seasons. This dataset can be used to analyze content trends, genre distribution, and regional availability on Netflix.

DATA OVERVIEW

Checking Basic Information

```
# data information
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   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
```

- The dataset contains 8,807 records with 12 columns related to Netflix content.
- Most columns are of type object (string), except release_year, which is an integer.
- Some columns have missing values, particularly in director, cast, and country, which need handling.
- The dataset includes important metadata such as title, type (Movie/TV Show), director, cast, release year, rating, and genres (listed_in).
- The date_added column represents when the content was added to Netflix and can be useful for time-based analysis.
- The duration column specifies the length of movies or the number of seasons for TV shows.

DATA OVERVIEW

Statistical Summary Of Numerical Data

```
#summary statistics of a dataset
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

The dataset spans release years from 1925 to 2021, with a mean of 2014 and a median of 2017. Most content falls between 2013 and 2019, indicating a focus on recent productions. The standard deviation of 8.82 suggests moderate variation in release years.

DATA OVERVIEW

Handling Missing Values

```
#number of null values in Dataset
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
```

The dataset contains missing values, with the highest in the "director" column (2634 nulls), followed by "cast" (825) and "country" (831). Other columns like "date_added," "rating," and "duration" have minimal missing values, which may require handling through imputation or removal based on analysis needs.

Imputing & Dropping Null Values

```
#imputing null values
df.director=df.director.fillna("unknown")
df.cast = df.cast.fillna('Not known')
df.country=df.country.fillna('unknown')
df.rating = df.rating.fillna("No Rating")
```

```
# Dropping rows with null values
df.dropna(inplace=True)
```

DATA OVERVIEW

The missing values in the dataset are handled using two approaches. The "director," "cast," "country," and "rating" columns are filled with placeholder values like "unknown" and "Not known." For other missing values, rows with null values are removed using dropna(inplace=True), ensuring a cleaner dataset for analysis.

Checking The datatype

```
# checking the datatype of columns in the dadtset
df.dtypes
```

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

The first step involves using df.dtypes to display the data types of each column in the dataset. This helps identify columns that may need type conversion for accurate analysis.

Fixing Datatype

```
#changing the datatype of"date_added" to datetime
df['date_added'] = pd.to_datetime(df['date_added'], format='%B %d, %Y', errors='coerce')
```

The second step attempts to convert the date_added column to a datetime format using pd.to_datetime(). This is crucial for time-based analysis, ensuring dates are correctly interpreted and manipulated.

DATA OVERVIEW

Null values arose due to inconsistent or missing data during type conversion

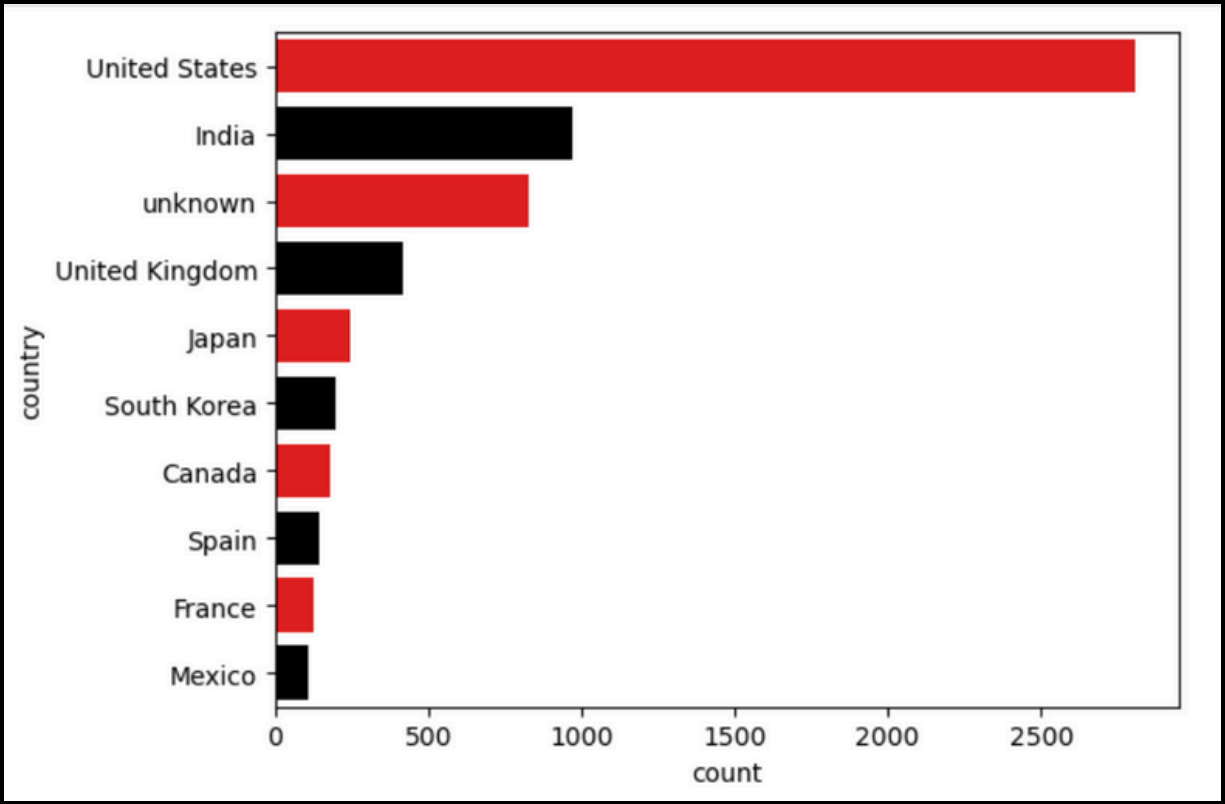
```
#Null values arose due to inconsistent or missing data during type conversion
df.isnull().sum()
```

```
show_id          0
type             0
title            0
director         0
cast             0
country          0
date_added      88
release_year     0
rating           0
duration         0
listed_in        0
description      0
dtype: int64
```

Some values in the date_added column were converted to null (NaT) due to incorrect formats, missing values, or non-date entries. Further data cleaning may be required to handle these missing values.

VISUALIZATION

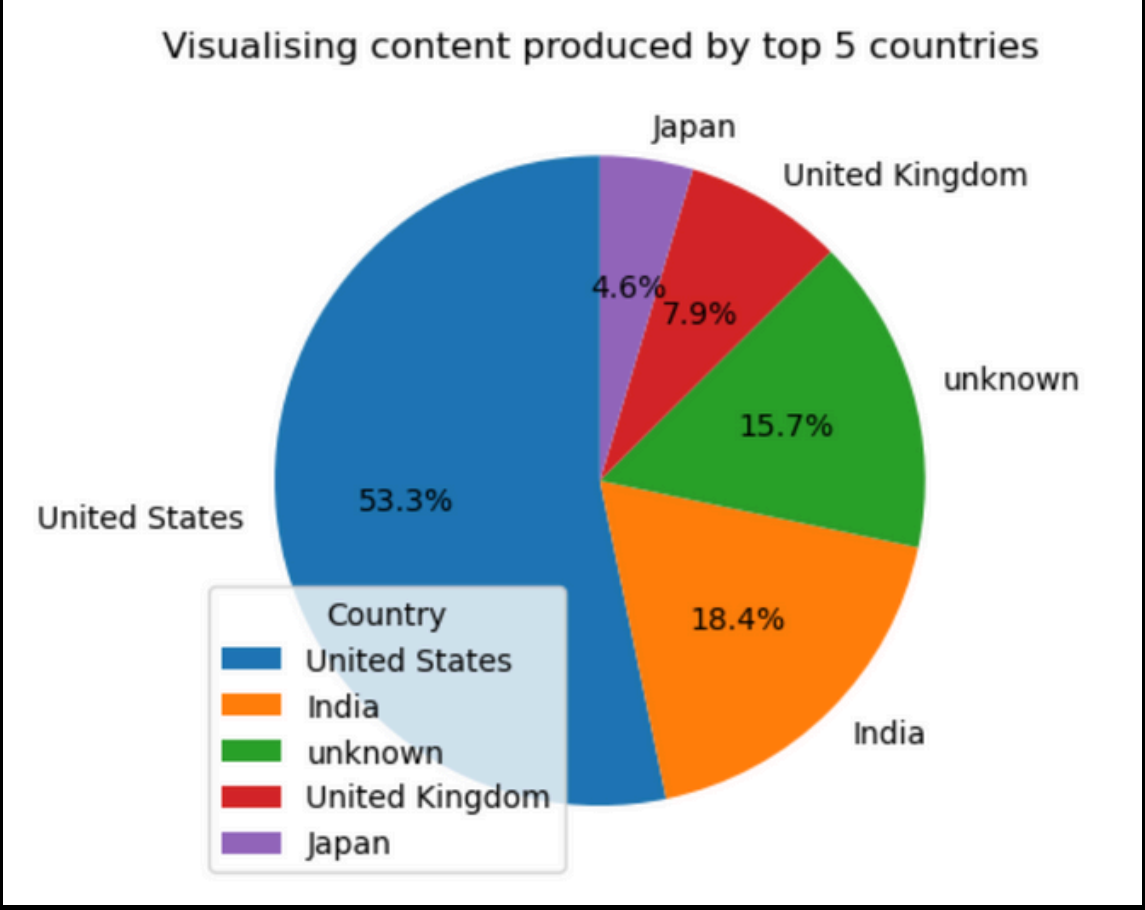
Top 10 Countries Producing Netflix Content



I've visualized the top 10 countries with the highest content count using a seaborn bar plot (sns.barplot). The x-axis represents the number of content entries, and the y-axis displays the respective country names. This visualization offers a clear comparison of the content counts for different countries, providing insights into the distribution and popularity of content based on geographical origin in the dataset.

VISUALIZATION

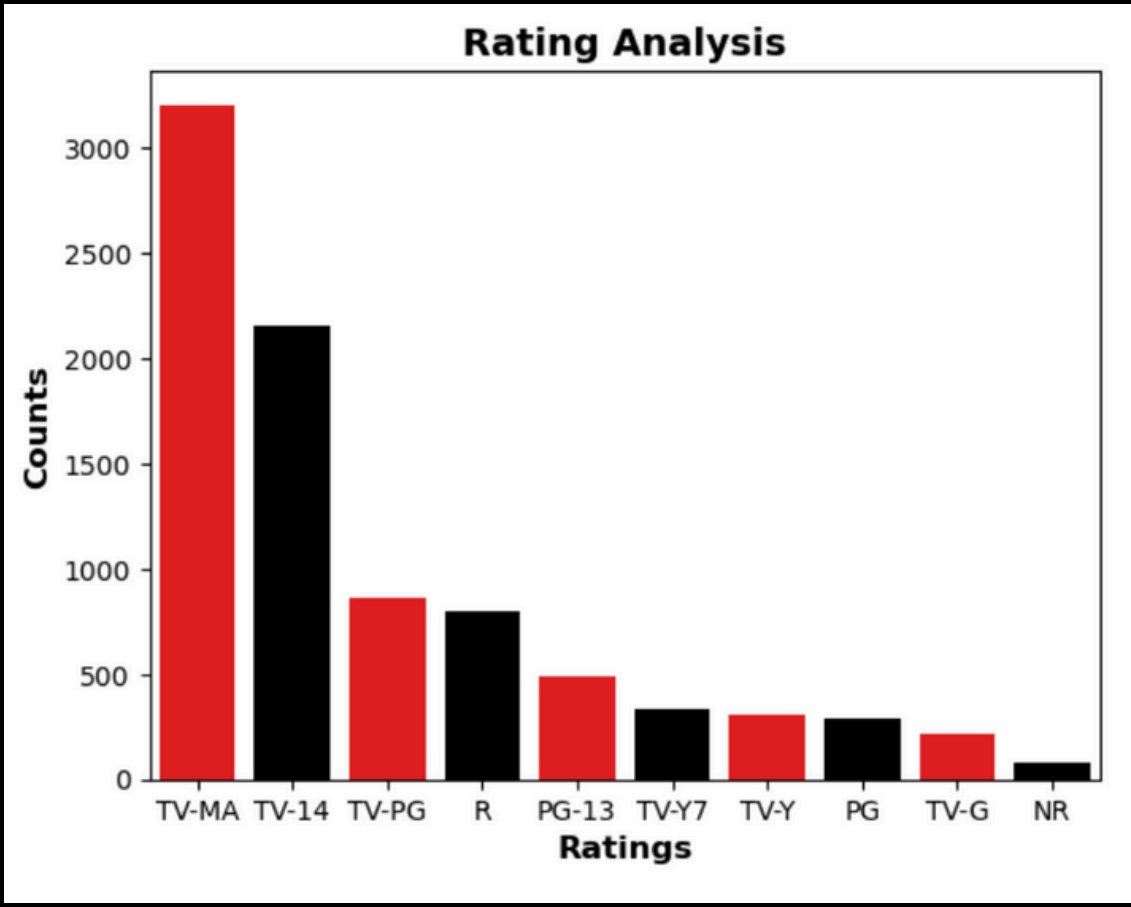
Distribution of Content Production by Top 5 Countries



I've visualized the content production share of the top 5 countries using a pie chart. The chart highlights the percentage of content contributed by each country, with the United States leading at 53.3%, followed by India at 18.4%. The United Kingdom, Japan, and an "Unknown" category also represent smaller portions of the content. This visualization provides a clear understanding of the dominant contributors in the dataset and underscores the significant role of the United States and India in content production. The "Unknown" category indicates missing data, which may require further attention for completeness.

VISUALIZATION

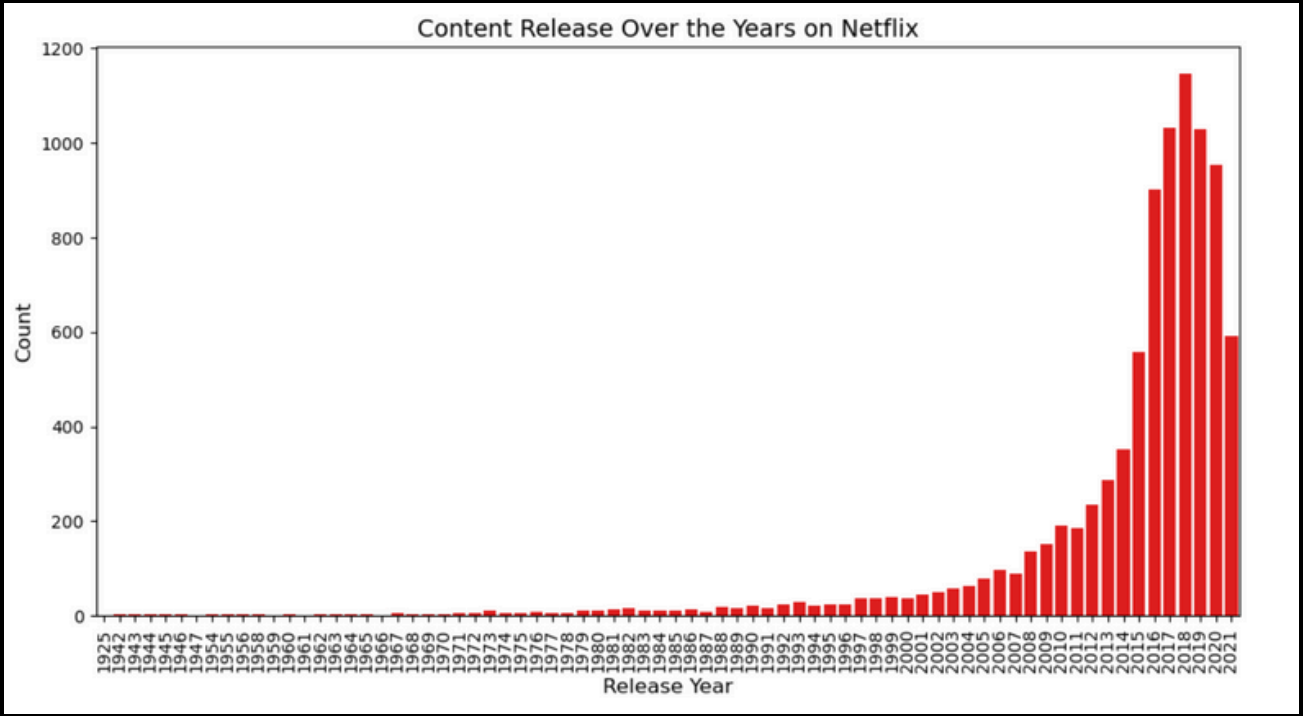
Rating analysis



I've visualized the distribution of content across different rating categories using a bar plot. The x-axis represents various ratings such as TV-MA, TV-14, TV-PG, R, PG-13, TV-Y7, TV-Y, PG, TV-G, and NR. The y-axis shows the count of content for each rating. TV-MA has the highest count, indicating it is the most common rating in the dataset. This visualization helps in understanding the prevalence of different content ratings, which can be useful for audience targeting and content strategy analysis.

VISUALIZATION

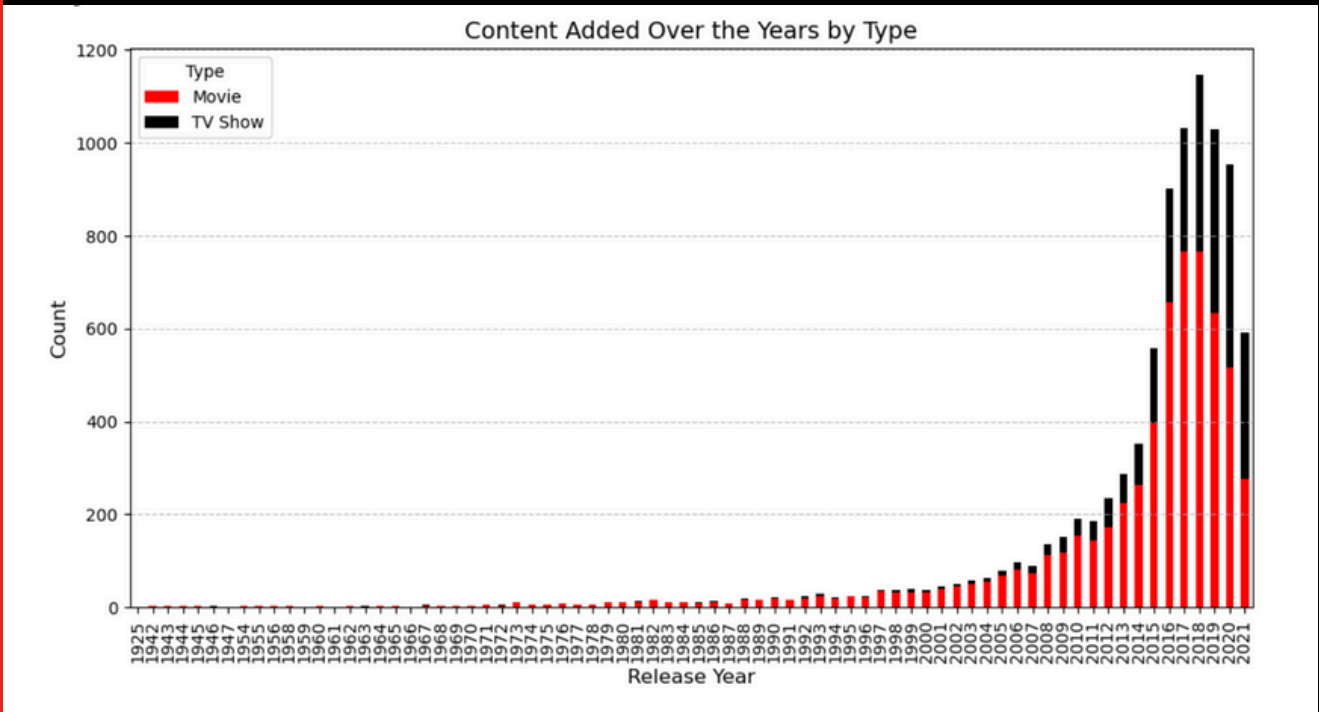
Content Release Trends Over the Years on Netflix



I've visualized the trend of content releases on Netflix over the years using a line plot. The x-axis represents the years, and the y-axis shows the count of content released each year. The plot indicates a significant increase in content production, especially from 2015 onwards, with a peak around 2020-2021. This visualization helps in understanding the growth trajectory of Netflix's content library and highlights the platform's expansion and investment in new content over time.

VISUALIZATION

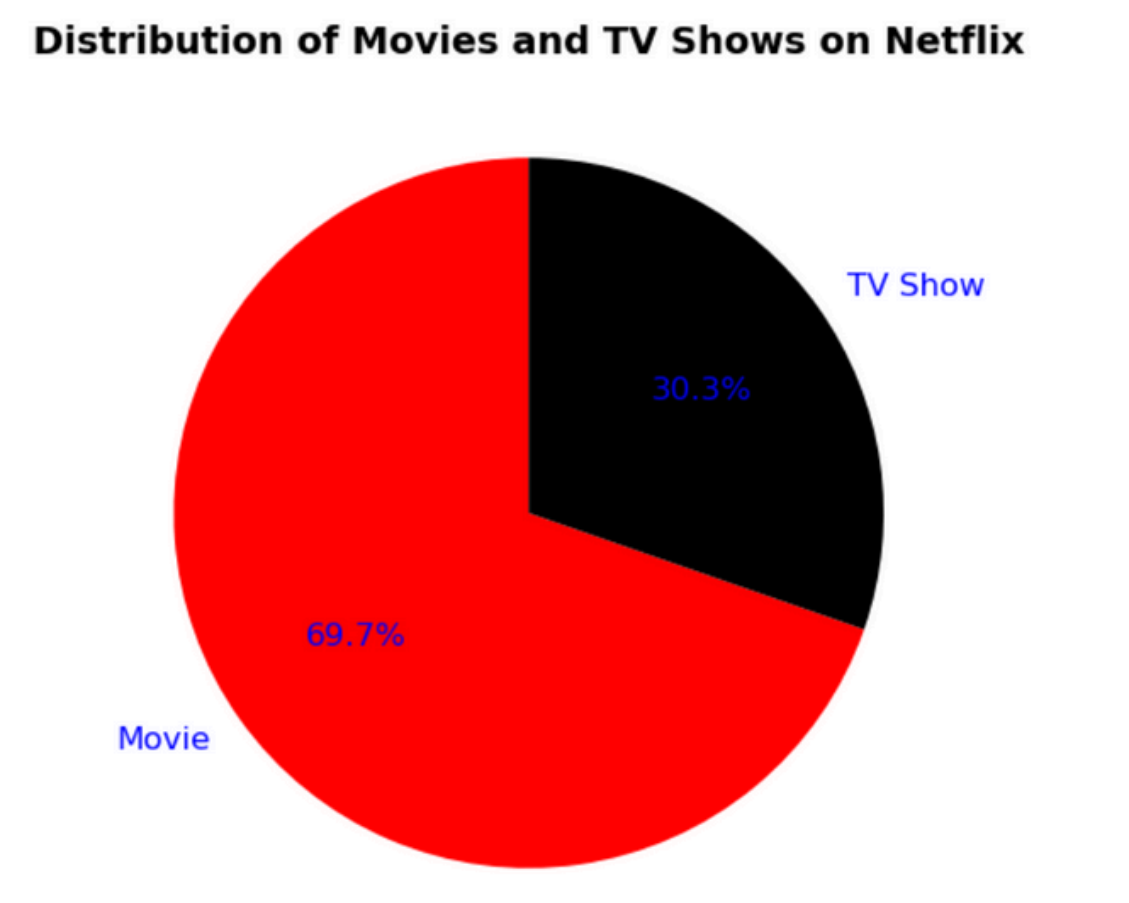
Content Added Over the Years by Type on Netflix



I've visualized the distribution of content added over the years by type (Movie and TV Show) using a grouped bar plot or line plot. The x-axis represents the release years, and the y-axis shows the count of content added. The plot compares the number of movies and TV shows released each year, providing insights into Netflix's content strategy. This visualization helps identify trends in content production, such as whether Netflix is focusing more on movies or TV shows in different years, and highlights shifts in content preferences over time.

VISUALIZATION

Distribution of Movies and TV Shows on Netflix



I've visualized the distribution of content types on Netflix using a pie chart. The chart shows that movies make up 69.7% of the content, while TV shows account for 30.3%. This visualization provides a clear overview of the proportion of movies versus TV shows available on the platform, highlighting Netflix's stronger focus on movies. Understanding this distribution is crucial for analyzing content strategy and audience preferences.

RECOMMENDATIONS

1. **Content Strategy Optimization:**
   - Given the dominance of movies (69.7%) over TV shows (30.3%), Netflix could consider balancing the content library by investing more in high-quality TV shows, especially in genres that are currently underrepresented.
   - Focus on producing more content in countries like India and the United Kingdom, which are significant contributors but still have room for growth compared to the United States.

2.Addressing Missing Data:

- Improve data collection processes to reduce missing values, especially in columns like "director," "cast," and "country." This will enhance the accuracy of future analyses and provide more comprehensive insights.

3.Regional Content Focus:

- Tailor content to regional preferences by analyzing genre popularity in different countries. For example, if certain genres are more popular in specific regions, Netflix can produce more content in those genres to cater to local audiences.

RECOMMENDATIONS

4.Rating-Based Recommendations:

- Since TV-MA is the most common rating, Netflix could explore creating more content for other rating categories to attract a broader audience, including families and younger viewers.

5.Time-Based Content Release:

- Leverage the trend of increasing content releases, especially post-2015, to plan strategic releases during peak engagement periods, such as holidays or special events.

CONCLUSION

- This exploratory data analysis (EDA) of Netflix's content library has provided valuable insights into the platform's content strategy, regional contributions, and audience preferences. The United States leads in content production, with movies dominating the library. However, there is significant potential for growth in other regions and content types.
- By addressing missing data, optimizing content strategy, and focusing on regional preferences, Netflix can further enhance its global appeal and user engagement. The insights gained from this analysis can guide strategic decisions, helping Netflix maintain its position as a leading streaming platform in the ever-evolving digital entertainment industry.