**KGiSL MICRO COLLEGE, COIMBATORE**

PYTHON REINFORCEMENT PROJECT-1

**NAME:ABIDEV c s**

**COURSE:DATA ANALYSIS&DATA SCIENCE**

**TITLE:NETFLIX DATA SET**

ABSTRACT

INTRODUCTION

METHODOLOGY

IMPORTED LIBRARY AND THEIR USES

DATA CLEANING

DATA VISUALIZATION

BOX PLOT

UNIVARIANT ANALYSIS

HISTORY PLOT

BIVARIANT ANALYSIS

MULTIVARIANT ANALYSIS

CONCLUSION

**ABSTRACT**

This project focuses on analyzing Netflix’s dataset to uncover patterns, trends, and insights regarding content distribution, genres, and audience preferences. The study employs **Exploratory Data Analysis (EDA)** techniques, statistical methods, and visualizations to interpret key attributes such as content type (Movies vs TV Shows), genre popularity, release year trends, and geographical contributions. The findings can guide **content strategy, recommendation systems, and business decisions** for streaming platforms. The analysis concludes with actionable insights and potential directions for **predictive modeling** and personalization.

**INTRODUCTION**

The rise of streaming services has transformed how audiences consume content. Netflix, being a global leader, offers a wide range of movies and TV shows. Analysing its catalog provides insights into content diversity, audience targeting, and trends over time.

**Problem Statement**  
Despite having vast content, Netflix faces challenges in **content recommendation**, **regional preferences**, and **retention strategies**. Understanding historical trends and user preferences is essential for better decision-making.

**Objectives**

* To analyze the Netflix dataset and identify patterns in content distribution.
* To examine genre popularity across countries.
* To investigate the temporal distribution of releases.
* To provide insights that can assist recommendation systems and strategic planning.

**Methodology**

* **Dataset Description**  
  The dataset contains attributes like **title, type, director, cast, country, date added, release year, rating, duration, and listed genres**.
* **Preprocessing Steps**
  + Handling missing values in fields like *director, cast, country*.
  + Converting *date added* to datetime format for temporal analysis.
  + Splitting multiple genres and countries for detailed insights.
* **Analysis Approach**
  + **Descriptive Statistics** to summarize data.
  + **EDA** using Python libraries like pandas, matplotlib, seaborn for trend visualization.
  + **Hypothesis Testing** for significant differences between content types.

**IMPORTED LIBRARY AND THEIR USES:**

* Pandas - It is denoted as pd, which is used for data manipulation and analysis.
* numpy - It is denoted ad np, which is the core library for numerical computations in Python.
* matplotlib.pyplot - It is denotd as plt, which is a fundamental library for data visualization.
* seaborn - It is debnotd as sns, which is built on top of matplotlib and specializes in statistical data visualization.
* warnings - filterwarning("ignore"),The warnings module manages warning messages.

# Netflix Data Cleaning

# 1. Import Libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import datetime as dt

* The file has been read and identified the total number of rows and columns bore starting the data cleaning .

# Load the dataset

df = pd.read\_csv("netflix1.csv")

# Display the updated DataFrame

df.head()

**DATA CLEANING:**

Data cleaning is the process of preparing raw data for analysis by correcting or removing inaccurate, incomplete, or irrelevant parts. By using

            \* isnull() - detect NAN values in dataframe.

            \*dropna()  - remover rows or columns with missing values.

            \*fillna()  - replacing missing values with a specified values or method.

            \*dropduplicate() - remove duplicate rows .

            \*info() - displays a concise summary: names,non-null counts, and data types.

            \*describe() - provides statistical summary for numerical columns: count,mean,std,min,max,etc...

            \*unique - returns unique values in a series

            \*shape - returns a tuple of (rows & columns) in the dataframe

            \*head() - shows the 1st half or upper rows and columns in the dataset

            \*tail() -  shows the last half or lower part in the dataset

**DATA VISUALIZATION**

* Matplotlib and seaborn are two popular python libraries for data visualization.
* Matplotlib commonly used for simple line plots, bar charts, histograms, etc.
* Seaborn statistical plots, built on top of matplotlib
* Plotly interactive visualizations in python

**BOXPLOT**

 A boxplot is a graphical representation of the distribution of dataset. It is useful for spotting outlier, uderstanding spread, and comparing medians across differnt groups.

**The below code helps to combain all the columns for plotting boxplot:**

num\_column=df.select\_dtypes(include=['number'])

len(num\_column.columns)

4

plt.figure(figsize=(10,8))

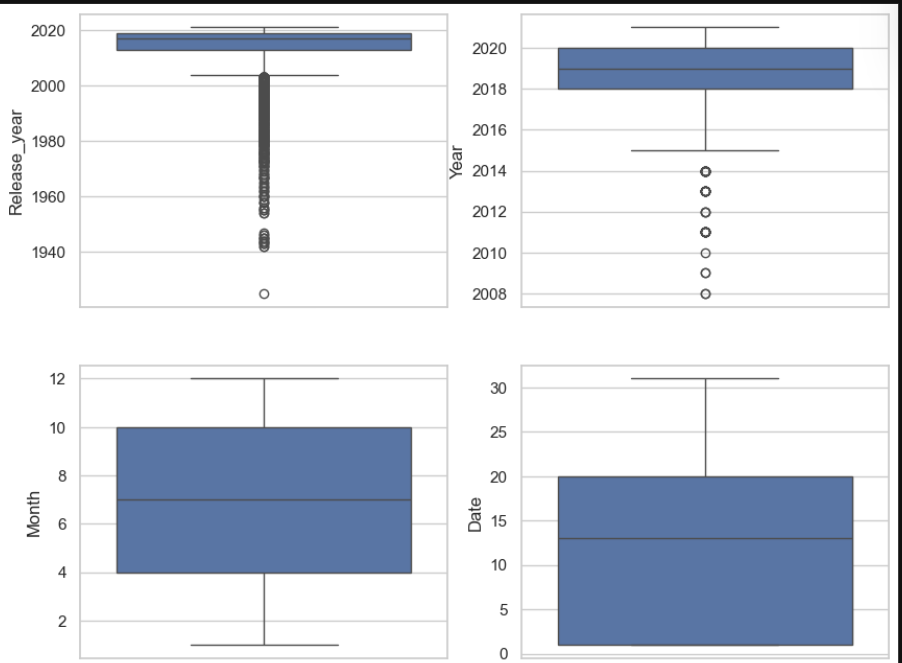
for i in range(len(num\_column.columns)):

    plt.subplot(2,2,i+1)

    sns.boxplot(y=num\_column.columns[i],data=df)

plt.show()

OUTPUT:



**UNIVARIATE ANALYSE**

Univariate analysis examines a single variable to understand its distribution, central tendency, and spread.

**COUNTPLOT FOR TYPE COLUMN**

plt.figure(figsize=(6,4))

sns.countplot(data=df, x="Type")

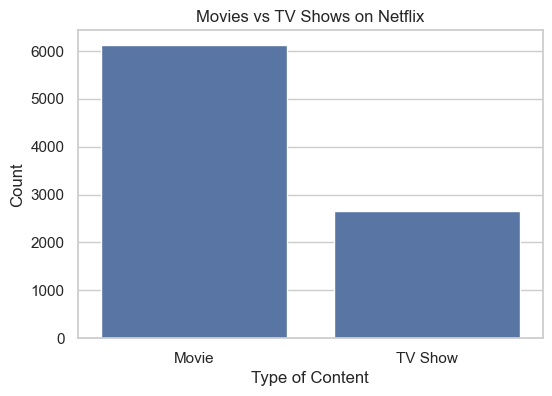
plt.title("Movies vs TV Shows on Netflix")

plt.xlabel("Type of Content")

plt.ylabel("Count")

plt.show()

OUTPUT:



**INSIGHTS**

- Movies dominate Netflix’s library → The number of movies is more than double the number of TV shows.

- TV Shows are fewer but growing → While movies still lead, Netflix has been increasing its investment in original TV shows in recent years.

- User preference reflection → This imbalance may reflect Netflix’s strategy to offer a wide variety of films to attract global audiences.

- TV Shows are strategic for engagement → Though smaller in number, TV shows often keep users subscribed longer because of multiple episodes/seasons.

- Overall insight → Netflix is primarily a movie-heavy platform, but TV shows play a crucial role in retaining subscribers and driving long-term engagement.

**COUNTPLOT FOR RATING**

* TV-MA (Mature Audience) is the most common rating with over 3,000 titles, meaning Netflix has a strong focus on adult-oriented content.
* TV-14 comes second, showing a large share of content suitable for teenagers and young adults.
* TV-PG and R-rated titles form the next tier, catering to family audiences and mature movie watchers.
* Ratings like TV-Y, TV-Y7, PG, and TV-G are much fewer, indicating limited children’s content compared to adult and teen content.
* Rare ratings such as NC-17, UR, and G are almost negligible, suggesting Netflix avoids extremely restricted or overly niche categories.

plt.figure(figsize=(10,6))

sns.countplot(data=df, y="Rating", order=df["Rating"].value\_counts().index, )

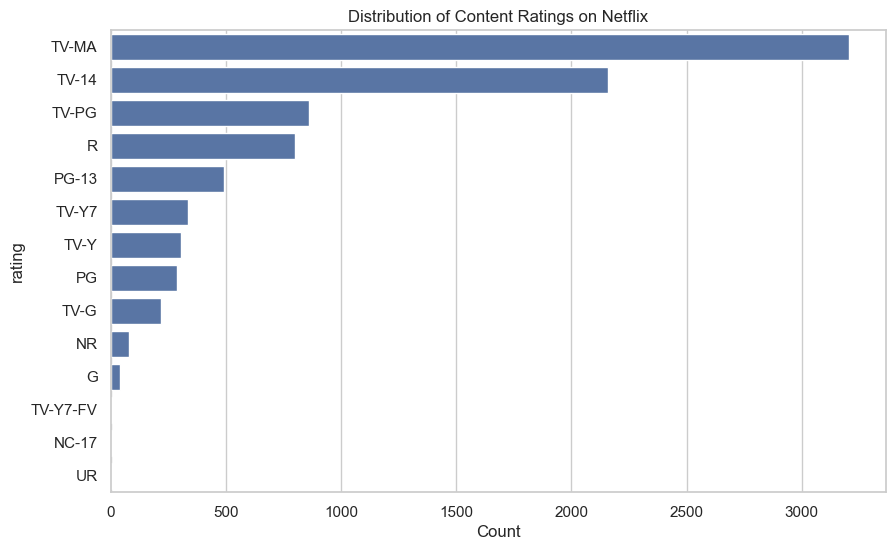
plt.title("Distribution of Content Ratings on Netflix")

plt.xlabel("Count")

plt.ylabel("rating")

plt.show()

OUTPUT:



**KEY PLOT FOR RELEASE YEAR**

* Very few titles before 1980 – Netflix’s library includes only a handful of very old/classic movies.
* Gradual rise after 1990s – Content availability slowly increases from the 1990s onward.
* Sharp growth after 2000 – The density starts climbing rapidly after 2000, reflecting the digital and streaming boom.
* Peak around 2015–2020 – Most Netflix titles are concentrated in this period, showing the platform’s aggressive expansion.
* Slight decline post-2020 – The density drops after 2020, possibly due to the COVID-19 pandemic affecting production.

plt.figure(figsize=(10,6))

sns.kdeplot(data=df, x="Release\_year", fill=True, color="blue", alpha=0.4)

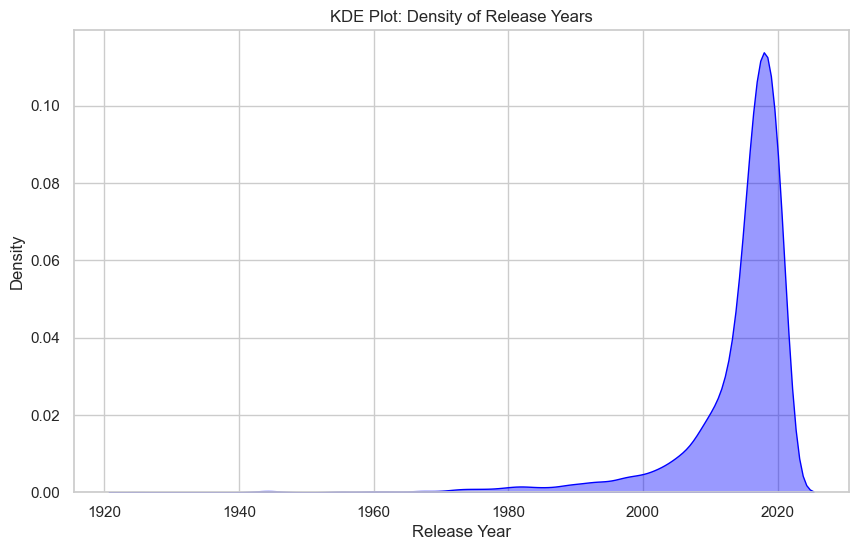
plt.title("KDE Plot: Density of Release Years")

plt.xlabel("Release Year")

plt.ylabel("Density")

plt.show()

OUTPUT:



**HISTORY PLOT FOR YEAR**

plt.figure(figsize=(8,5))

sns.histplot(x='Year',data=df, bins=30, kde=True, color="orange")

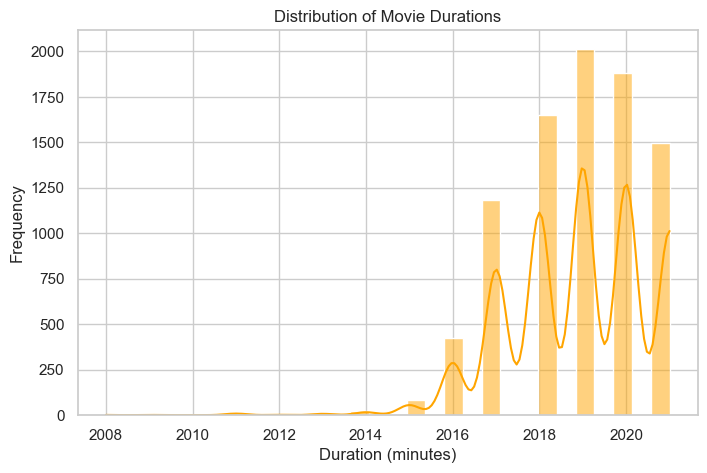
plt.title("Distribution of Movie Durations")

plt.xlabel("Duration (minutes)")

plt.ylabel("Frequency")

plt.show()

OUTPUT:



**BIVARIATE ANALYSE**

**United States dominates with the highest number of Netflix titles (by a big margin). India is in second place, showing its strong role in Netflix’s international expansion. United Kingdom follows, contributing a significant number of titles. Countries like Pakistan, Canada, Japan, South Korea, France, and Spain also appear, but with much smaller contributions compared to the US. The presence of “Not Given” indicates missing country data in the dataset.**

**BARPLOT FOR TOP COUNTEIES**

plt.figure(figsize=(10,6))

top\_countries = df["Country"].value\_counts().head(10)

sns.barplot(x=top\_countries.values, y=top\_countries.index,)

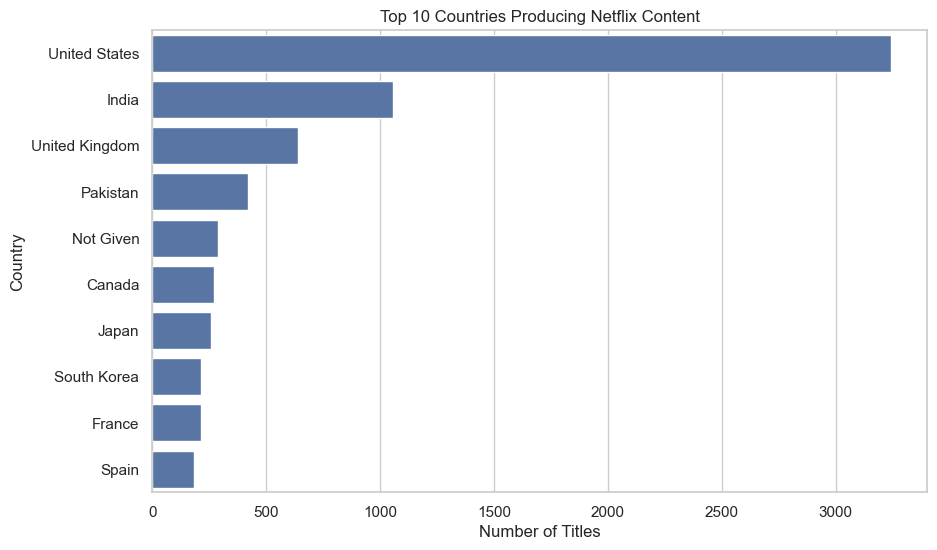
plt.title("Top 10 Countries Producing Netflix Content")

plt.xlabel("Number of Titles")

plt.ylabel("Country")

plt.show()

OUTPUT:



**BARPLOT FOR GENRE COUNTS**

International Movies dominate, showing Netflix’s strong focus on global/local content to appeal to different regions. Dramas and Comedies are next — these are evergreen genres with wide appeal. International TV Shows also rank high, reinforcing Netflix’s strategy of diversifying content beyond Hollywood. Documentaries have a strong presence, reflecting growing interest in real-life stories. Genres like Action & Adventure, TV Dramas, Independent Movies, Children & Family Movies, and Romantic Movies are present but comparatively smaller.

from collections import Counter

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Extract genres

genres = []

for entry in df["Listed\_In"].dropna():

    for g in entry.split(", "):

        genres.append(g)

# Count and sort top genres

genre\_counts = pd.Series(Counter(genres)).sort\_values(ascending=False).head(10)

# Plot

plt.figure(figsize=(10, 6))

sns.barplot(x=genre\_counts.values, y=genre\_counts.index, palette="magma")

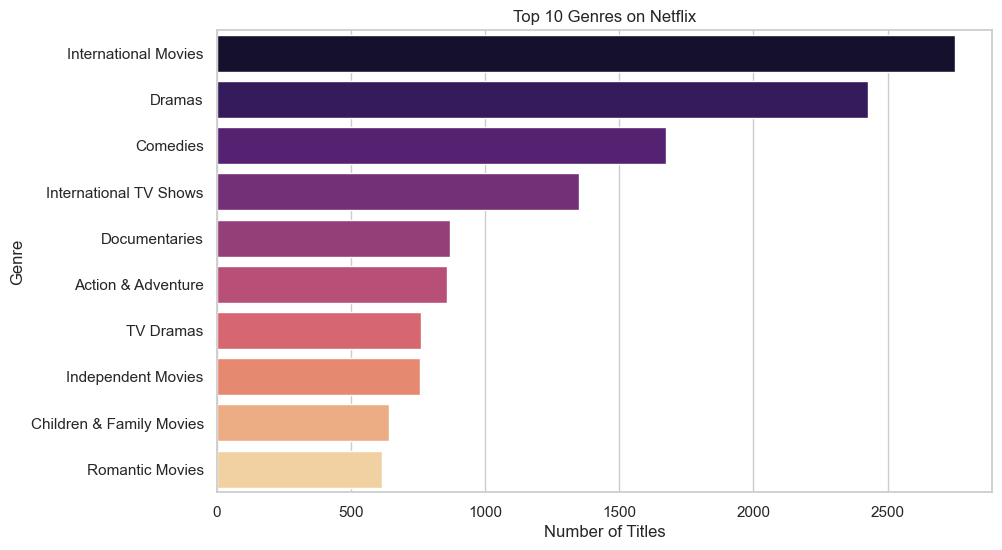
plt.title("Top 10 Genres on Netflix")

plt.xlabel("Number of Titles")

plt.ylabel("Genre")

plt.show()

OUTPUT:



**HISTORY PLOT FOR CONTENT ADDED TO NETFLIX PER YEAR**

**Sharp Growth After 2015 Netflix added very little content before 2014. From 2015 onwards, there’s a sharp rise, reflecting Netflix’s global expansion. Peak Years Around 2017–2019, the number of new titles surged, peaking close to 2000 titles per year. This was the period when Netflix heavily invested in original content and international licensing. Slight Decline After 2020 After 2019, the trend shows a slight decline in content addition. Possible reasons: Market saturation. COVID-19 pandemic affecting production. Netflix focusing on quality originals rather than bulk additions.**

# Netflix Data Visualization

# Optional: if your file is raw, make sure to clean first

# (date\_added, year/month, duration\_int, etc.)

# Assuming you already have the cleaned version with these columns.

# Set seaborn style for better visuals

sns.set(style="whitegrid")

plt.rcParams["figure.figsize"] = (12,6)

# 2. Content Added Per Year

plt.figure(figsize=(12,6))

df["Year"].value\_counts().sort\_index().plot(kind="bar", color="skyblue")

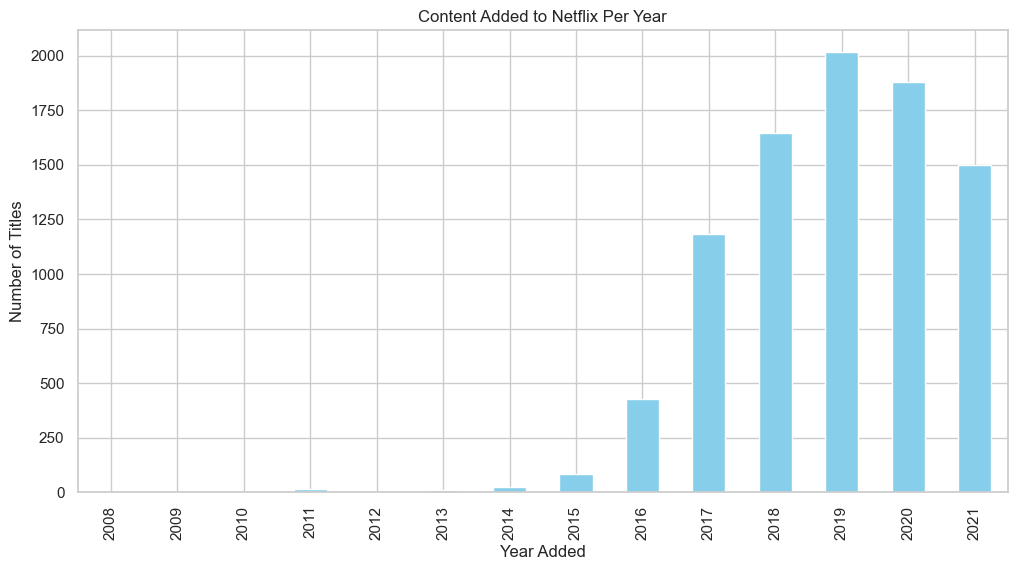
plt.title("Content Added to Netflix Per Year")

plt.xlabel("Year Added")

plt.ylabel("Number of Titles")

plt.show()

OUTPUT:



United States clearly dominates, producing the largest share of Netflix content. India comes second, highlighting Netflix’s strong push into the Indian entertainment market. United Kingdom is third, reflecting its well-established TV and film industry. Countries like Pakistan, Canada, Japan, South Korea, France, and Spain also contribute, but in smaller volumes. The presence of “Not Given” again shows some missing metadata in the dataset.

# 3. Top 10 Countries Producing Content

plt.figure(figsize=(10,6))

top\_countries = df["Country"].value\_counts().head(10)

sns.barplot(x=top\_countries.values, y=top\_countries.index, palette="viridis")

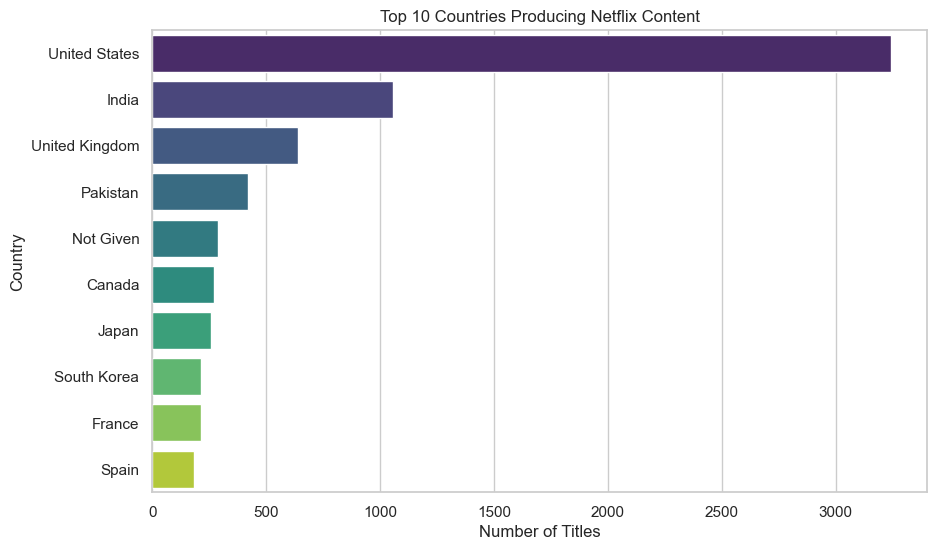
plt.title("Top 10 Countries Producing Netflix Content")

plt.xlabel("Number of Titles")

plt.ylabel("Country")

plt.show()

OUTPUT:



**Most Common Ratings** → The majority of titles are rated **TV-MA** (Mature Audience) and **TV-14** (suitable for teens and above).

* This shows Netflix’s strong focus on **teen and adult audiences**.

**Family/Kids Content** → Ratings like **TV-PG, TV-Y, TV-Y7, G, and PG** are present but form a smaller proportion, meaning **children’s content is less dominant**.

**Movies vs. TV Shows** → Movies are more often rated **PG, R, PG-13**, while TV shows tend to dominate **TV-MA, TV-14** categories.

**Global Appeal** → Having a spread across many ratings categories allows Netflix to **serve diverse demographics worldwide**

# 4. Distribution of Ratings

plt.figure(figsize=(10,6))

sns.countplot(data=df, y="Rating", order=df["Rating"].value\_counts().index, palette="coolwarm")

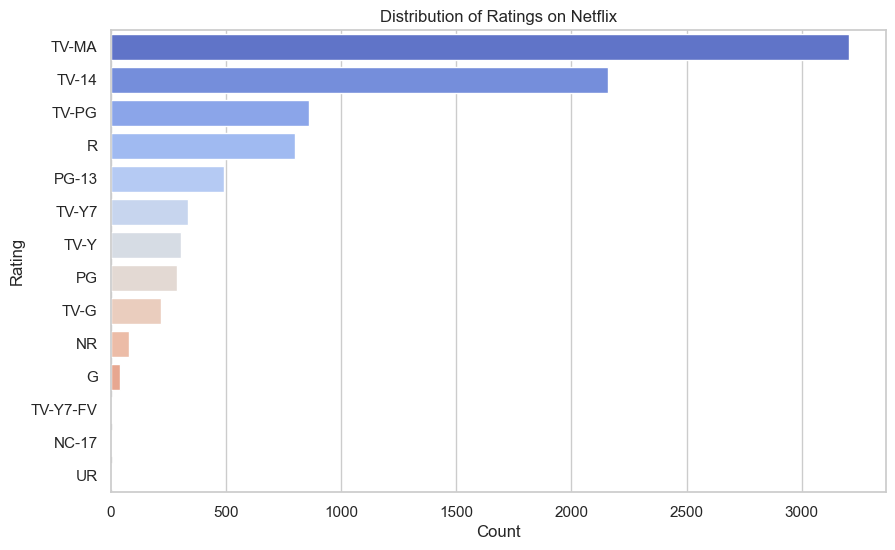
plt.title("Distribution of Ratings on Netflix")

plt.xlabel("Count")

plt.ylabel("Rating")

plt.show()

OUTPUT:



**Most Dominant Genres** → The top positions are typically **International Movies, Dramas, and Comedies**, confirming Netflix’s focus on global and story-driven content.

**Strong Global Presence** → International content (both movies and TV shows) forms a big chunk, showing Netflix’s strategy of catering to worldwide audiences.

**Balanced Mix** → Along with mainstream genres (Dramas, Comedies), Netflix invests in **Documentaries, Action & Adventure, and Family content**, ensuring variety for different audiences.

**Niche Genres** → Genres like **Romantic Movies, Independent Movies, and Children & Family** appear in the top 10 but with fewer titles compared to dominant ones.

# 5. Top 10 Genres

genres = []

for entry in df["Listed\_in"].dropna():

    for g in entry.split(", "):

        genres.append(g)

genre\_counts = pd.Series(Counter(genres)).sort\_values(ascending=False).head(10)

plt.figure(figsize=(10,6))

sns.barplot(x=genre\_counts.values, y=genre\_counts.index, palette="magma")

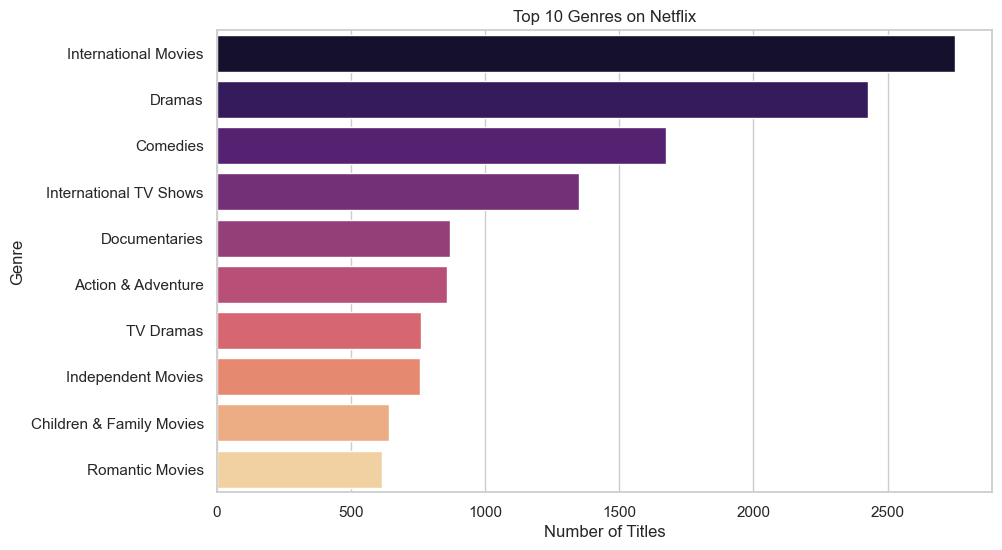
plt.title("Top 10 Genres on Netflix")

plt.xlabel("Number of Titles")

plt.ylabel("Genre")

plt.show()

OUTPUT:



# 3. Horizontal Bar Chart → Top 10 Countries

plt.figure(figsize=(10,6))

top\_countries = df["Country"].value\_counts().head(10)

top\_countries.plot(kind="barh", color="purple")

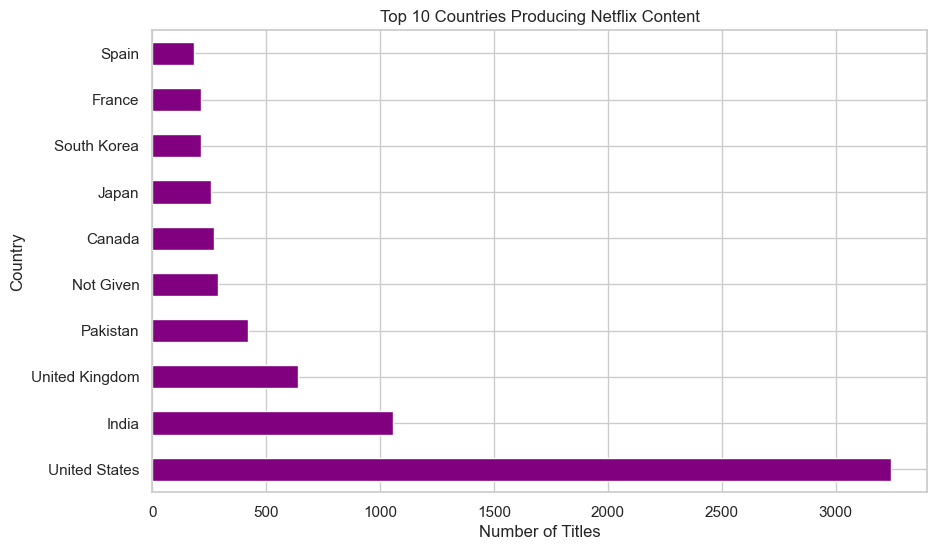
plt.title("Top 10 Countries Producing Netflix Content")

plt.xlabel("Number of Titles")

plt.ylabel("Country")

plt.show()

OUTPUT:



**Early Years (2008–2014)** → Relatively fewer titles were added, reflecting Netflix’s initial transition into streaming.

**Sharp Growth (2015–2019)** → Rapid increase in content additions, showing Netflix’s aggressive expansion into original productions and global licensing.

**Peak Period** → Around 2018–2019, the number of new titles added reached its highest point.

**Recent Years (2020 onwards)** → Slight decline or stabilization in growth, possibly due to:

* COVID-19 production delays,
* Shift in strategy towards **quality over quantity**,

# 2. Line Chart → Content Added Over Years

plt.figure(figsize=(10,5))

df["Year"].value\_counts().sort\_index().plot(kind="line", marker="o", color="green")

plt.title("Trend of Content Added Over Years")

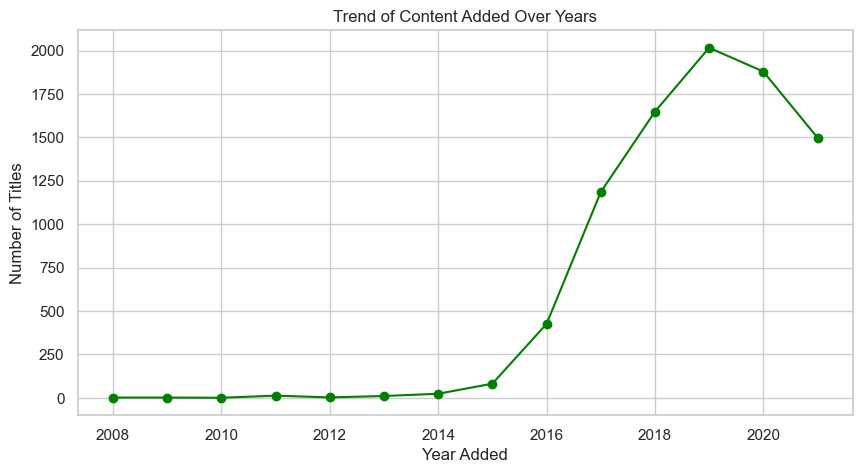
plt.xlabel("Year Added")

plt.ylabel("Number of Titles")

plt.grid(True)

plt.show()

OUTPUT:



**MULTIVARIANT ANALYIZE**

**HEATMAP:**

* A heatmap is a data visualization technique that uses color gradients to represent values in a matrix (rows × columns).
* In this we used correlation heatmap.

numeric\_df=df.select\_dtypes(include="number")

corr\_matrix=numeric\_df.corr()

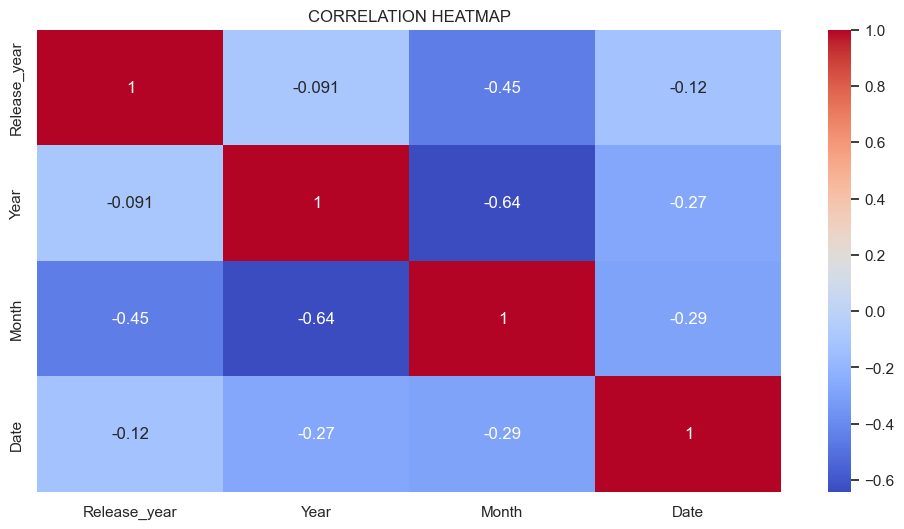
# CORRELATION HEATMAP

sns.heatmap(corr\_matrix.corr(), annot=True , cmap="coolwarm")

plt.title("CORRELATION HEATMAP")

plt.show()

OUTPUT :



**PAIR PLOT:**

 A pair plot (also called a scatterplot matrix) is a visualization that shows the pairwise relationships between multiple variables in a dataset.

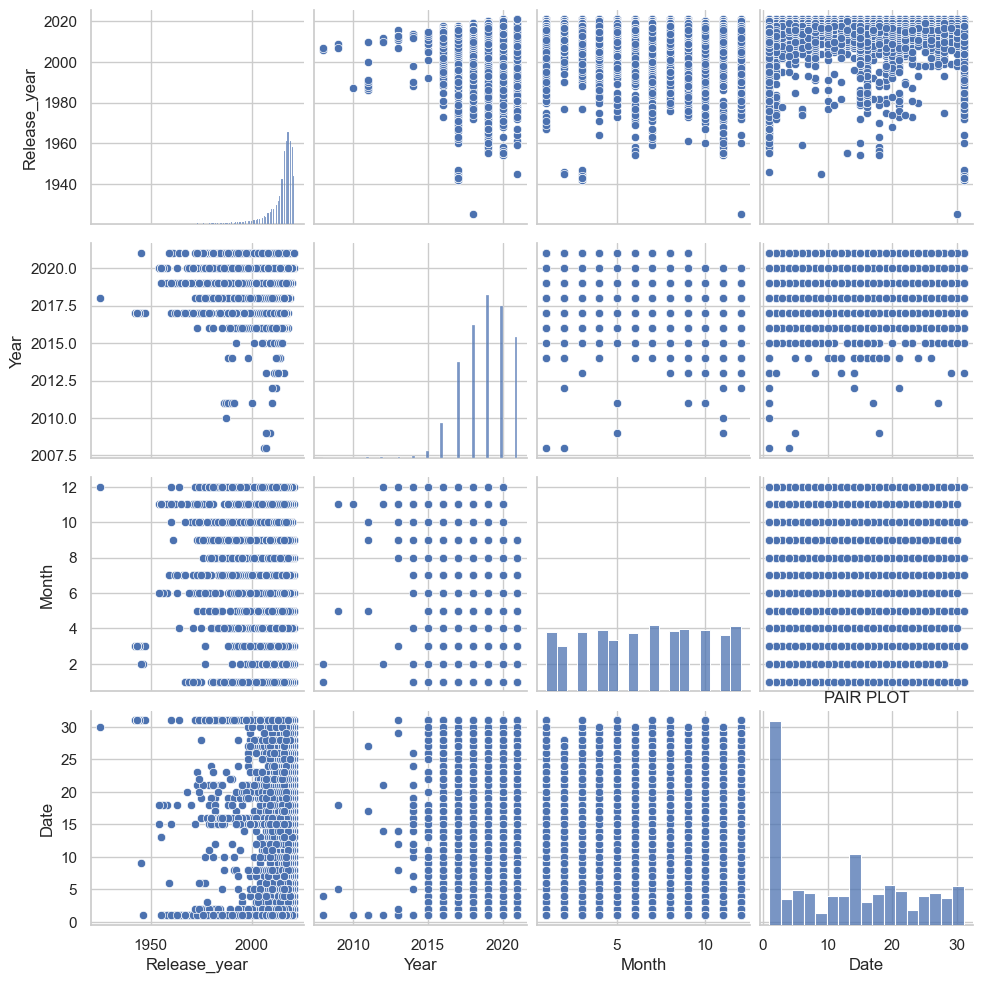
# PAIR PLOT

sns.pairplot(df)

plt.title("PAIR PLOT")

plt.show()

OUTPUT :



CONCLUSION

Netflix’s content library is highly dominated by a few popular categories. Genres such as **International Movies, Dramas, and Comedies** appear most frequently, highlighting Netflix’s focus on diverse storytelling and global reach. While niche genres are present, the majority of titles fall into mainstream categories that appeal to wider audiences. This indicates that Netflix prioritizes genres with broad popularity while still offering variety to cater to different viewer preferences.

**Conclusion :**

Netflix offers a diverse content library: movies with standardized lengths, and TV shows for extended viewing.

Genres highlight a global strategy, focusing on international and drama-based storytelling.

Statistical tests confirm consistency in movie duration and clear distinction between content formats.

**Recommendations :**

Expand Underserved Genres → Invest more in independent, family, and documentary content to attract niche audiences.

Content Duration Strategy → Maintain the ~100-minute standard for movies but explore slightly longer films for premium storytelling.

TV Show Data → Track episode-level durations for better audience engagement insights instead of only “seasons.”

Global Content Push → Strengthen international content as it already dominates, appealing to diverse audiences worldwide.