GOWTHAM J
220901027

PYTHON MACHINE LEARNING PROJECT

Visualizing and Analyzing Netflix Titles: Insights from Data

Abstract:

This project explores and visualizes the Netflix Titles dataset to uncover insights about genre distribution, country contributions, release trends, and more. The dataset is cleaned and processed to handle missing values and ensure accuracy. The project employs Python libraries such as Pandas and Matplotlib to generate graphical representations that highlight patterns in the streaming platform's catalog. The findings provide valuable perspectives for understanding Netflix's offerings and customer preferences.

Introduction

Netflix is one of the most popular streaming platforms worldwide, offering a vast catalog of shows and movies. Understanding the characteristics of its content is crucial for content creators, marketers, and researchers. This project focuses on cleaning and visualizing the Netflix Titles dataset to derive actionable insights into the platform's offerings. The analysis covers genres, countries of production, ratings, release year trends, and durations.

Data Overview

The Netflix Titles dataset contains information about movies and TV shows available on Netflix. Key columns include:

- **Title**: The name of the movie or show.
- **Director**: The director's name.
- **Cast**: Leading actors.
- **Country**: Country of production.
- **Date Added**: The date content was added to Netflix.
- **Release Year**: The year of release.
- **Rating**: Audience rating (e.g., PG-13, TV-MA).
- **Duration**: Duration of content (e.g., minutes for movies or seasons for shows).
- **Genre (Listed In)**: Genres assigned to the content.

The dataset required significant cleaning, as many columns had missing values.

Challenges Addressed:

- **Missing Values**: Key columns like `country`, `director`, `rating`, and `cast` contained missing values. Strategies like filling with "Unknown" or defaults were employed.
- **String Parsing**: Certain columns (e.g., `duration`) needed string extraction and conversion to numeric formats for analysis.
- **Data Distribution**: Some distributions (e.g., genres and countries) were heavily skewed, requiring careful visualization to avoid misrepresentation.

Methodology:

1. **Data Cleaning**:
    - Handled missing values with appropriate replacements (e.g., "Unknown").
    - Parsed and transformed data where necessary (e.g., `duration` column extraction).
2. **Visualization**:
    - Used `Matplotlib` for graphical representations.
    - Explored the following aspects:
        - Genre distribution.
        - Top contributing countries.
        - Audience rating patterns.
        - Trends in release years.
        - Duration distributions.
    - Designed a dynamic visualization function to iterate over updated data.
3. **Live Analysis**:
    - Implemented a loop to refresh the visualizations in real time (optional for further development).

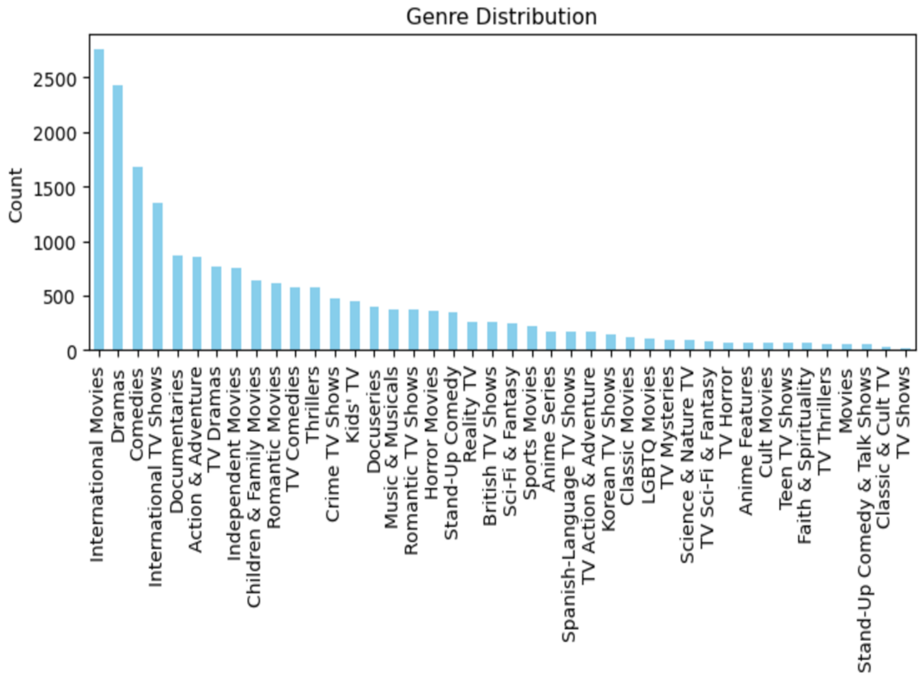
Code:

```
# Importing Required Libraries

import pandas as pd

import matplotlib.pyplot as plt

from IPython.display import clear_output

# Load Data

file_path = r"C:\Users\D E L L\Downloads\netflix_titles.csv"

data = pd.read_csv(file_path)

# Data Cleaning

data_cleaned = data.copy()

data_cleaned['country'].fillna('Unknown', inplace=True)
```

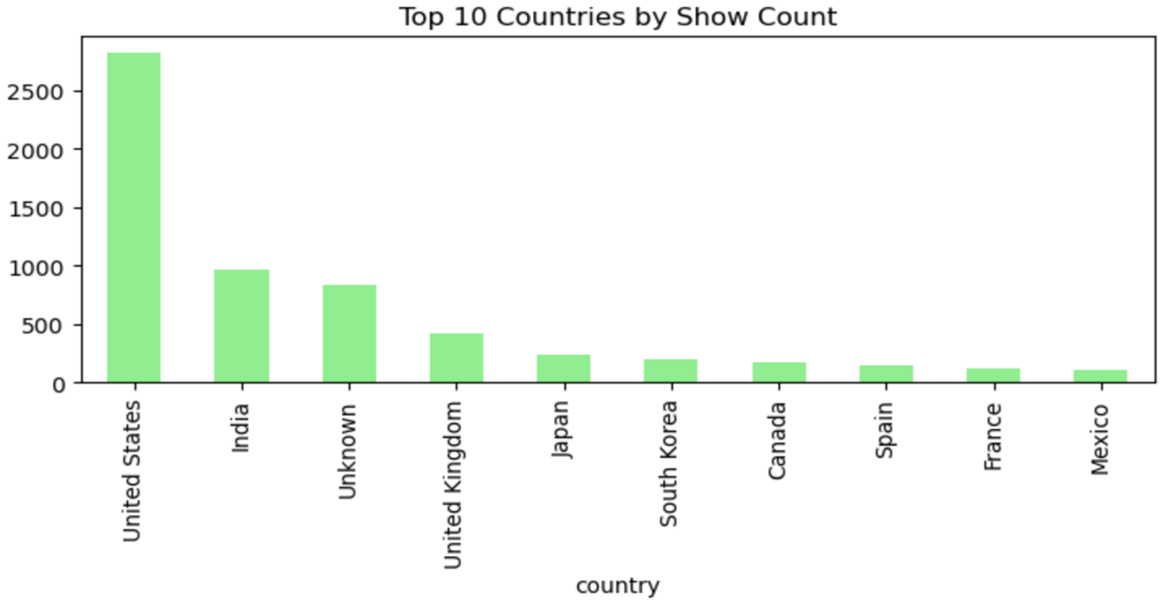
```
data_cleaned['rating'].fillna('Unrated', inplace=True)

data_cleaned['director'].fillna('Unknown', inplace=True)

data_cleaned['cast'].fillna('Unknown', inplace=True)

data_cleaned['date_added'].fillna('Unknown', inplace=True)

data_cleaned['duration'].fillna('Unknown', inplace=True)

data_cleaned['description'].fillna('No description available.', inplace=True)

# Graphical Representations

def plot_graphs(df):

    clear_output(wait=True)

   plt.figure(figsize=(15, 15))

 # Subplot 1: Genre Distribution

    plt.subplot(3, 2, 1)

    genres = df['listed_in'].str.split(', ').explode().value_counts()

    genres.plot(kind='bar', title="Genre Distribution", color='skyblue')

    plt.ylabel("Count")

   # Subplot 2: Top 10 Countries by Show Count

    plt.subplot(3, 2, 2)

    countries = df['country'].value_counts().head(10)

    countries.plot(kind='bar', title="Top 10 Countries by Show Count", color='lightgreen')

    plt.ylabel("Count")

   # Subplot 3: Rating Distribution

    plt.subplot(3, 2, 3)

    ratings = df['rating'].value_counts()

    ratings.plot(kind='bar', title="Rating Distribution", color='lightcoral')
    plt.ylabel("Count")
 # Subplot 4: Release Year Trends
    plt.subplot(3, 2, 4)
    release_years = df['release_year'].value_counts().sort_index()
    release_years.plot(kind='line', title="Release Year Trends", color='orange')
    plt.ylabel("Count")
    plt.xlabel("Year")
   # Subplot 5: Top 10 Directors by Show Count
    plt.subplot(3, 2, 5)
    directors = df['director'].value_counts().head(10)
    directors.plot(kind='bar', title="Top 10 Directors by Show Count", color='lightblue')
    plt.ylabel("Count")
   # Subplot 6: Duration Distribution
    plt.subplot(3, 2, 6)
    df['duration'] = df['duration'].str.extract('(\d+)').astype(float)
    df['duration'].dropna().plot(kind='hist', bins=30, title="Duration Distribution", color='violet')
    plt.xlabel("Duration (minutes)")
    plt.tight_layout()
    plt.show()
# Main loop to continuously display graphical representations
while True:
    plot_graphs(data_cleaned)
    print("Exiting program...")
    break
```

OUTPUT;

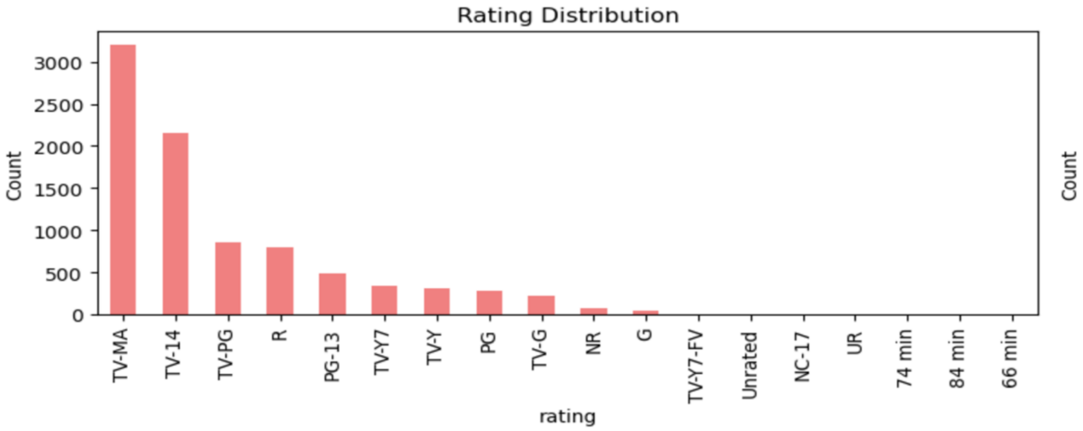
Genre Distribution



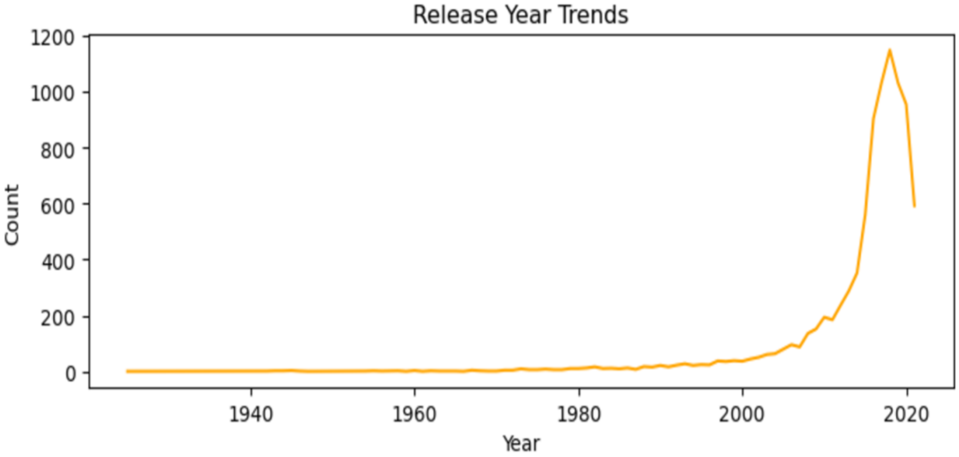
Top 10 Countries by Show Count



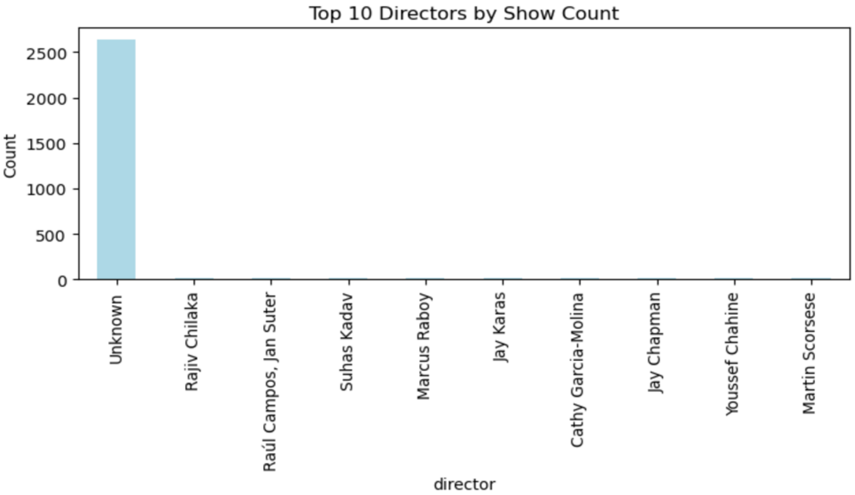
Rating Distribution



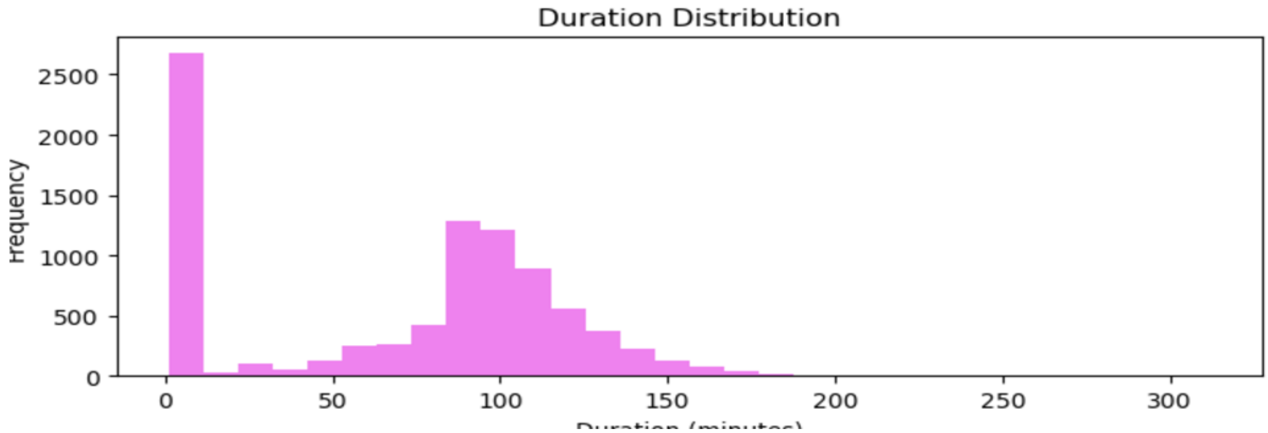
Release Year Trends



Top 10 Directors by Show Count



Duration distribution



Link for dataset;

netflix_titles.csv

CONCLUSION;

This analysis highlights key aspects of Netflix's catalog, such as popular genres, top countries of production, and trends in ratings. The cleaning and visualization process makes the data accessible and insightful, offering a foundation for further studies or strategic decisions. Future work could include deeper genre-specific analyses, regional trends, or predictions based on release patterns.