Netflix Data Analysis 🎬

Introduction

This project is an exploratory data analysis of a Netflix titles dataset. The goal is to uncover key trends and patterns in content distribution, release timelines, and overall content library growth using Python and popular data science libraries.

Data Cleaning 🧼

Data cleaning is a crucial first step in any data analysis project. For this dataset, we performed the following key cleaning tasks to ensure the data is accurate and reliable:

- **Handling Null Values:** Null (`NaN`) values were identified in columns such as `director`, `cast`, and `country`. These were addressed by either filling them with a placeholder (e.g., "Unknown") or removing them, depending on the specific analysis being performed.
- **Finding and Removing Duplicates:** Duplicate rows were located and removed from the dataset to ensure each entry is unique. This is essential for accurate counts and analyses.
- **Missing Values:** Columns like `director`, `cast`, and `country` had missing values (`NaN`). For relevant analyses, these were handled by either filling them with a placeholder (e.g., "Unknown") or by simply filtering them out to prevent errors in visualizations.

Key Questions & Analysis ❓

This analysis aims to answer the following questions:

- **Release Trends:** Which years had the most movie and TV show releases on Netflix?
- **Content Distribution:** What is the ratio of Movies to TV Shows in the dataset?
- **Top Contributors:** Who are the most prolific directors and actors with the most titles?

Key Findings ✅

Based on our analysis, we found some interesting insights:

- **Release Peak:** The year **2018** was the most active year for content releases, with a significantly higher number of titles added than any other year.

- **Content Type:** The dataset contains a much larger number of **Movies** than TV Shows, indicating a strong focus on films in the content library.