The data and its source –

The dataset is a collection of Netflix data, which provides insights into the content available on the Netflix streaming platform. This dataset offers a comprehensive view of various aspects of the shows and movies available on Netflix, and included following fields-

- **SHOW ID:** Unique ID of each show
- **TYPE:** Show category. Could be either a Movie or a TV Show.
- **TITLE:** Name of the show
- **DIRECTOR:** Name of the director(s) of the show
- **CAST:** Names of actors/actresses in the show
- **COUNTRY:** Countries where the show is available to watch on Netflix
- **DATE ADDED:** Date when the show was added on Netflix
- **RELEASE YEAR:** Release year of the show
- **RATING:** Show rating on Netflix
- **DURATION:** Time duration of the show
- **LISTED IN:** Genre of the show
- **DESCRIPTION:** Brief insight into what the show is about

It appears that most of the columns in my dataset contain categorical or alphanumeric data, with the exception of the RELEASE YEAR column, which contains numeric data representing the release years of the shows.

The dataset is collected from website named Kaggle.com

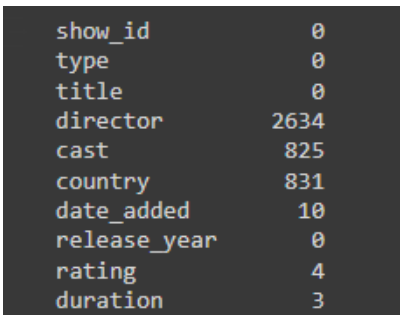
Data exploration and data cleaning steps

Data exploration involved a series of steps to understand the dataset better. This included displaying the first and last rows, checking for null values, identifying and handling missing data, detecting duplicate rows, and obtaining basic statistics about the dataset.

Data Exploration:

1. Displayed the first and last rows of the dataset to get an initial look at the data.

2. Checked for null values in the dataset, identifying columns with the most null values.



show_id	0
type	0
title	0
director	2634
cast	825
country	831
date_added	10
release_year	0
rating	4
duration	3

3. Decided to handle missing values differently for different columns based on their data type and nature.

4. Checked for duplicate rows in the dataset.

5. Checked the dimensions of the dataset, finding that it has 8,807 rows and 12 columns.

6. Obtained basic information about the dataset using the .info() method to display column names, non-null counts, and data types.

7. Checked the distribution of the data in the 'type' column to understand the balance between "Movie" and "TV Show" categories.

8. Explored the distribution of show ratings in the 'rating' column to identify the most common rating categories on Netflix.

9. Investigated the cast members who appeared in the most content on Netflix, which provides insights into popular actors or actresses.

10. Analyzed the duration of shows available on Netflix to identify the most common show durations.

11. Identified the top values in each of these analyses, such as the most frequent content type, rating category, cast members, and show durations.

Data Cleaning:

1. Filled missing values in the 'director', 'cast', 'country', and 'duration' columns with the string 'No Data Available,' which is a reasonable approach for handling missing values in categorical and text data.

2. Filled missing values in the 'date added' and 'rating' columns using the mode (most frequent) value from the same column to ensure that missing values are replaced with common values.

3. No duplicates were found in the dataset.

Two clearly stated comparison questions with the unit of analysis, the comparison values and how they are computed.

Comparison Question 1: What is the popular choice for most directors when choosing a genre for TV shows or movies?

Unit of Analysis: Directors

Comparison Values: Popular genres for directors.

How They Are Computed: The code counts the number of movies or TV shows made by directors, groups the data by director and genre, and computes the most popular genre choices.

Result of Analysis –

- The result shows that "Stand-Up Comedy" is the popular choice for most directors, with 18 directors choosing this genre.
- "Children & Family Movies" is also a popular choice among directors, with 18 directors selecting it.
- The data provides a comprehensive list of directors and their genre preferences, showcasing the diversity of genre choices among directors.

Comparison Question 2: How does the distribution of content types (Movies vs. TV Shows) change over the years in the United States, and what are the most prevalent content categories on Netflix in the USA?

Unit of Analysis: Years and Content Types

Comparison Values:

- Distribution of content types over the years.
- Most prevalent content categories in the USA.

How They Are Computed

- The code first checks the number of movies and TV shows released each year from 1950 to 2020.
- It then creates a subset of data for the USA to focus on content available in the United States.
- The code visualizes the trends over the years using Seaborn for movies and TV shows
- Additionally, the most popular content categories in the USA are determined and visualized.

Result of Analysis –

- It shows that the number of both movies and TV shows released has been increasing in recent years, with a significant uptick in the number of TV shows released starting in the 2000s.
- In 1950, there were approximately 250 movies released, and this number increased over the years, reaching over 700 movies by 2000.
- Similarly, the number of TV shows released also increased significantly over time, surpassing 400 TV shows in 2020.
- Overall, the graph demonstrates the increasing trend in the number of movies and TV shows released each year, with TV shows gaining prominence over time.
- Documentary is the most popular movie and TV show category in the United States followed by Stand-up comedy

Description of the Program:

The program uses libraries such as Pandas, NumPy, Matplotlib, and Seaborn to load, explore, clean, and analyze the Netflix dataset. It includes code for data loading, data exploration, data cleaning, answering the comparison questions, and visualizing the results using various charts.

Program includes following sections-

Data Loading:

The program begins by connecting to Google Drive to access the dataset, demonstrating the ability to import data from external sources.

Data Reading:

It utilizes Pandas to read the dataset and load it into a DataFrame for further analysis.

Data Cleaning:

Data cleaning is a crucial step, and the program handles missing values effectively.It replaces missing values in the 'director', 'cast', 'country', and 'duration' columns with the string 'No Data Available.'It fills missing values in the 'date_added' and 'rating' columns with the mode (most frequent value) from the same column.

The program checks for duplicate values, but in this case, no duplicates were found.

The result is a clean dataset ready for exploration and analysis.

Data Exploration:

The program explores the dataset through several components:

It displays the distribution of content types, specifically "Movie" and "TV Show," present in the 'type' column.

It examines the distribution of show ratings and show durations, providing insights into the content's characteristics.

A quick statistical summary of numeric columns is generated to provide an overview of numeric data.

The program answers two key comparison questions:

Question 1 - Director's Genre Preferences:

It analyzes the popular choices of directors when selecting genres for TV shows or movies.

Counts the number of movies or TV shows made by directors and groups the data by director and genre.

Missing values are filled with 0, and the result is sorted in descending order based on counts.

A graph is plotted to show the count of genres chosen by directors.

Question 2 - Growth in Production and Content Categories:

It compares the growth in production of movies and TV shows on Netflix over the years for the United States.

Identifies the prevalent content categories on Netflix in the USA.

The program creates a subset of data for the USA and checks the number of movies and TV shows released over the years.

The program effectively combines data loading, cleaning, exploration, and analysis steps to gain insights into Netflix's content and the viewing preferences of directors and viewers. It provides a clear and structured approach to working with the dataset, ensuring that data is handled efficiently, and results are visualized for better comprehension.

Description of the Result of the Analysis:

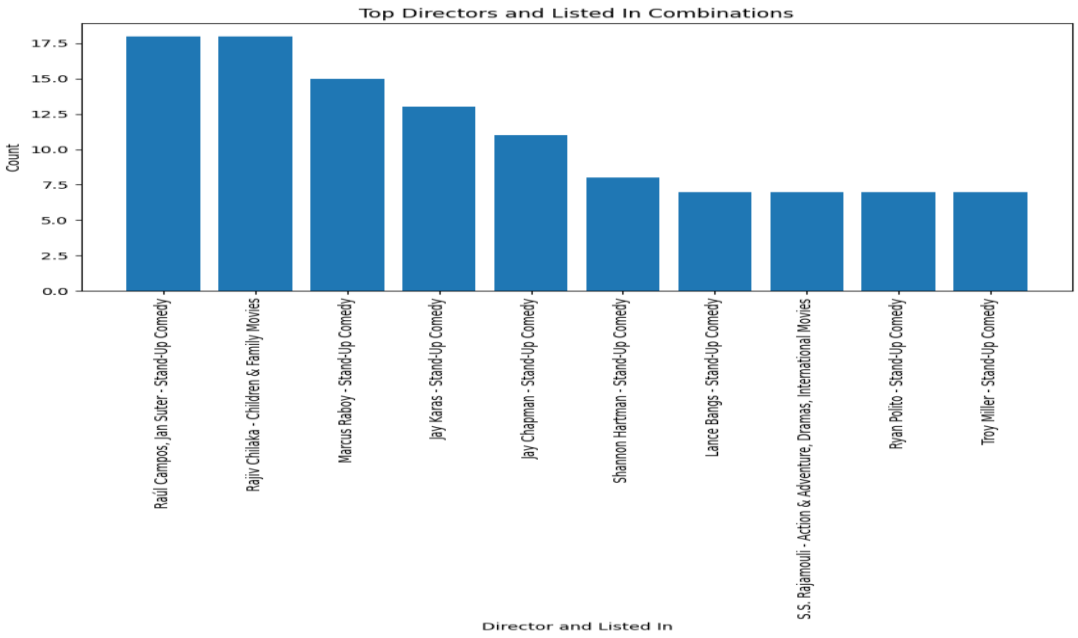
The analysis provides insights into the Netflix dataset, such as the popularity of content genres among directors, changes in content types (Movies vs. TV Shows) over the years in the United States, and the most popular content categories in the USA. The results are visualized through bar charts, pie charts, and line plots to make the insights more accessible and understandable.

Following are the key takeaways from the analysis -

- Content Type Distribution: Netflix produces more movies compared to TV shows, indicating that movies are more prevalent on the platform.
- Popular Ratings: Most of the content on Netflix is rated as TV-MA and TV-14, suggesting that the content caters to a mature audience.
- Content Release Trends: The analysis reveals that Netflix has released a significantly higher number of content items around 2020, marking a period of increased production.
- Top Contributing Countries: The United States is the top contributor to Netflix content, followed by India, making it the second most influential country in terms of content contribution.
- Growth Trend: The analysis suggests that Netflix has shown substantial growth over the years, with an increasing number of content releases as time progresses. This trend indicates a growing presence of Netflix in the entertainment industry.
- Show Duration: Most of the shows on Netflix have a duration of one season, indicating that single-season shows are prevalent on the platform.
- Top Cast Member: David Attenborough has the most cast appearances in shows on Netflix, highlighting his significant role in Netflix content.
- Content Release Peaks: Netflix has witnessed the highest number of content releases around the year 2018, marking a period of peak production.

Result of Analysis of Q1 –

- The result shows that "Stand-Up Comedy" is the popular choice for most directors, with 18 directors choosing this genre.
- "Children & Family Movies" is also a popular choice among directors, with 18 directors selecting it.
- The data provides a comprehensive list of directors and their genre preferences, showcasing the diversity of genre choices among directors.



Result of Analysis of Q2 –

- It shows that the number of both movies and TV shows released has been increasing in recent years, with a significant uptick in the number of TV shows released starting in the 2000s.
- In 1950, there were approximately 250 movies released, and this number increased over the years, reaching over 700 movies by 2000.
- Similarly, the number of TV shows released also increased significantly over time, surpassing 400 TV shows in 2020.
- Overall, the graph demonstrates the increasing trend in the number of movies and TV shows released each year, with TV shows gaining prominence over time.
- Documentary is the most popular movie and TV show category in the United States followed by Stand-up comedy.