NETFLIX

EXPLORATORY DATA ANALYTICS

- Data summarizing
- Data cleaning
- Data visulization
- Data insight

info:
Avinash Kumar
avinash969658@gmail.com

2 About the Netflix

Netflix is a global online video streaming platform where users can watch movies, TV shows, web series, documentaries, and more on the internet. Founded in 1997 in the United States, Netflix originally started as a DVD rental service but shifted to online streaming in 2007.

Today, Netflix operates in 190+ countries and is one of the world's largest OTT (Over-The-Top) platforms. It produces its own exclusive shows and films called Netflix Originals, including popular titles like Stranger Things, Money Heist, Dark, and Squid Game.

Netflix works on a subscription model, offering ad-free streaming in HD and 4K quality across mobile, laptops, and smart TVs. The platform uses machine learning and big data to give personalized recommendations to each user.

In short, Netflix is a major streaming service known for its huge content library, high-quality originals, and smooth viewing experience.

3 Objective

Purpose of the Dataset:

The goal of the Netflix EDA project is to conduct a comprehensive exploration and analysis of Netflix's content dataset. This includes understanding the data structure, ensuring data integrity by handling missing values and duplicates, deriving descriptive statistics, and visualizing content distribution across genres and release years.

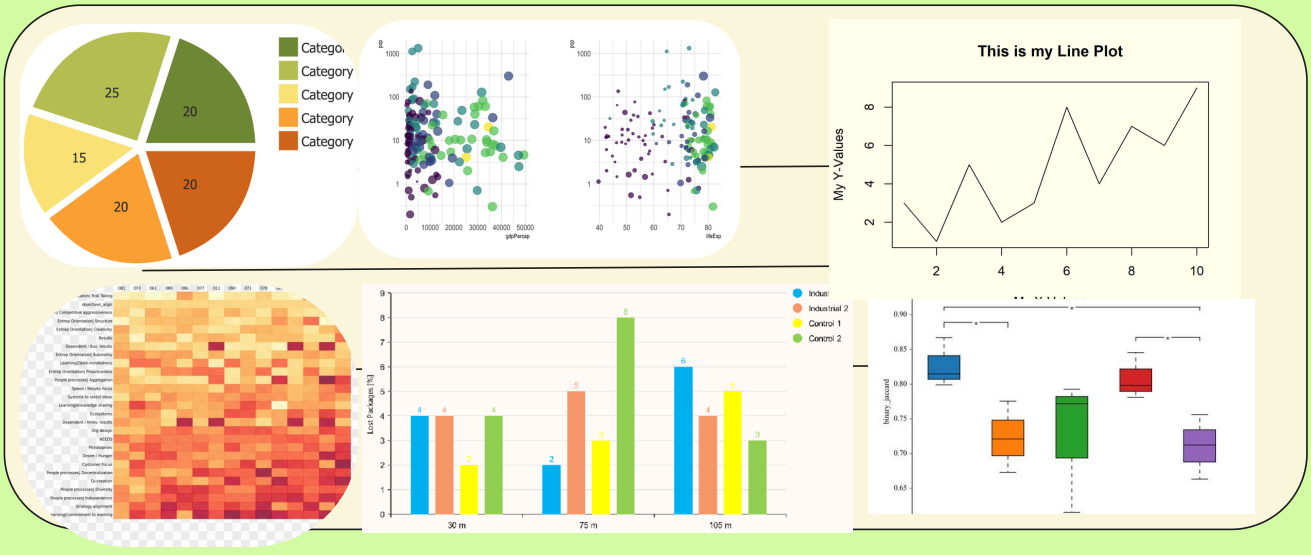
Additionally, the project aims to identify temporal trends, analyze content attributes like ratings and duration, and assess audience engagement metrics. By synthesizing these insights, the project aims to draw meaningful conclusions and provide actionable recommendations to enhance Netflix's content offerings and user experience.

Tools

4

matplotlib pandas seaborn NumPy plotly

Graphs



5 Description

- Loading and Preparing Netflix Dataset for Analysis in Google Colab

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

Mounted at /content/drive
```

- in this analysis , we import Pandas for data manipulation ,Numpy for numerical operations ,Matplotlib for creating visualization ,and Seaborn and plotly for enhanced statistical graphic.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
```

- We import the Google drive module to mount Google `Drive`,enabiling access to files and dataset in the for our analysis.

```
df=pd.read_csv("/content/drive/MyDrive/data set/Copy of netflix_titles.csv")
```

- We use pandas to Read the CSV file containing big Netflix from Google Drive, loading it into a DATAFRAME Name ‘df’ for analysis.

6 Data overview

```
df
```

	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 comprises of 8790 rows and 15 columns.

```
f"Rows:{df.shape[0]},Columns:{df.shape[1]}"
```

```
'Rows:8790,Columns:15'
```

- The data set drop duplicate columns

```
df.drop_duplicates()
```

Description

show_id:
A unique identifier assigned to each movie or TV show in the dataset.

type:
Indicates whether the entry is classified as a Movie or a TV Show.

7 Description

title:
The official name of the movie or TV show.

director:
The name of the director responsible for the creation of the title.

cast:
A list of actors and actresses who performed in the movie or TV show.

country:
The country or countries where the movie or TV show was produced.

date_added:
The date on which the content was added to the streaming platform.

release_year:
The year in which the movie or TV show was originally released.

rating:
The age restriction or maturity classification assigned to the content (e.g., PG-13, TV-MA).

8 Description

duration: Represents how long the content is:
For movies → total minutes
For TV shows → number of seasons

listed_in: Categories or genres associated with the movie or TV show.

description: A short summary or synopsis explaining what the movie or TV show is about.

Describe

```
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
```

9 Describe

Column Name	Non-Null Values	Description
show_id	8807	Unique ID for each show
type	8807	Movie or TV Show
title	8807	Title of the show
director	6173	Director name (many missing values)
cast	7982	Actors (some missing values)
country	7976	Country of origin
date_added	8797	Date added to Netflix
release_year	8807	Release year (only integer column)
rating	8803	Age rating (PG, TV-MA, etc.)
duration	8804	Duration in minutes or seasons
listed_in	8807	Genre/category
description	8807	Summary of the show

10 Data Cleaning & Pre-Processing

```
d[d>0]
```

	0
director	2634
cast	825
country	831
date_added	10
rating	4
duration	3

We fill the missing data in the director ,cast,country,data_added,rating and duration

```
df["director"].fillna("not Present",inplace=True)

df["cast"].fillna("not present",inplace=True)

df["country"].fillna("not present",inplace=True)

df.dropna(inplace=True)
```

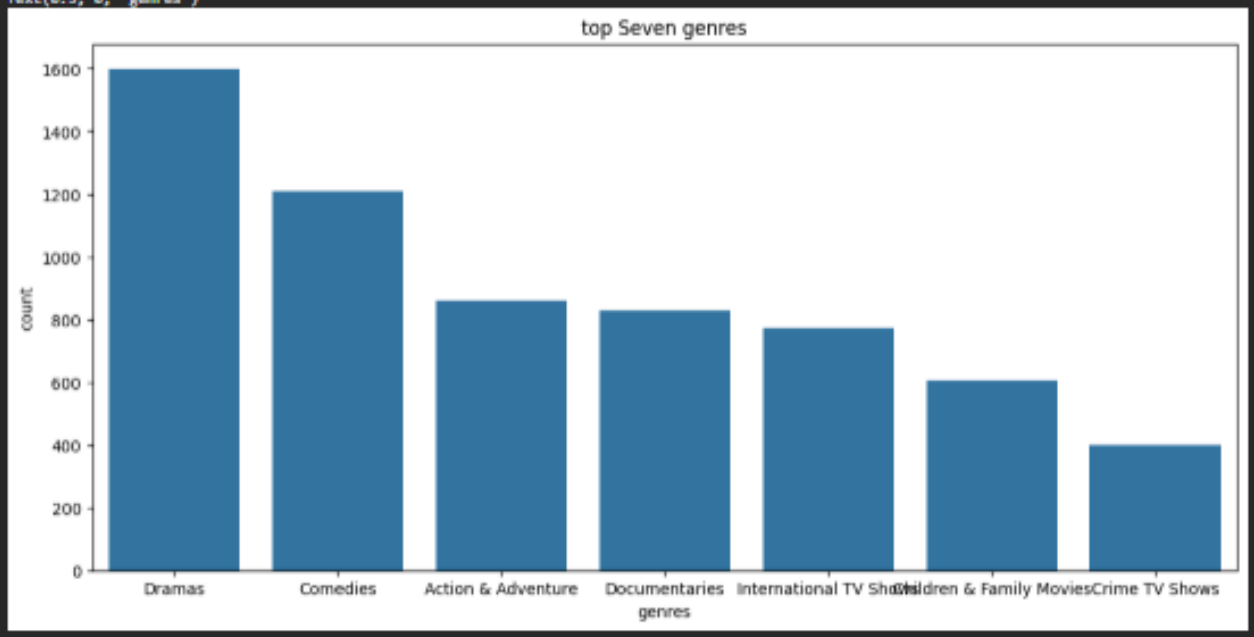
type	0
title	0
director	0
cast	0
country	0
date_added	0
release_year	0
rating	0
duration	0
listed_in	0

11 Data Visualization and Insights

1. **Create visualizations to represent the distribution of content over different genres.**

```
plt.figure(figsize=(13,6))
sns.barplot(df["genres"].value_counts().head(7))
plt.title("top Seven genres")
plt.ylabel("count")
plt.xlabel("genres")
```

```
Text(0.5, 0, 'genres')
```



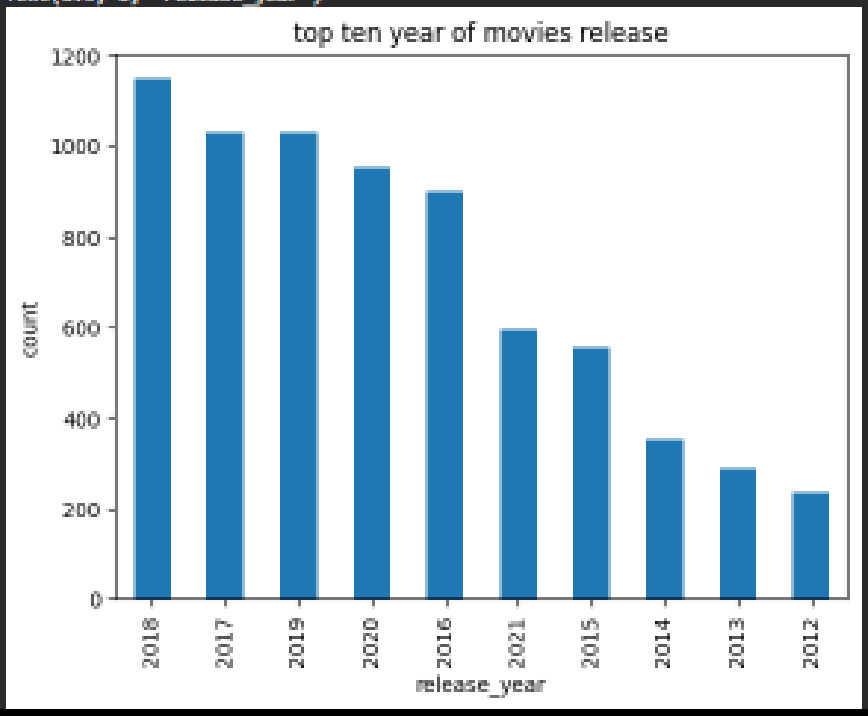
- **Dominant Genre: Dramas lead by a significant margin, indicating a strong preference or abundance of dramatic content in the dataset.**
- **Popular Categories: Following dramas, Comedies and Action & Adventure are the next most frequent, suggesting these genres also have wide appeal.**
- **Diverse Interests: The presence of Documentaries, International TV Shows, Children & Family Movies, and Crime TV Shows reflects a mix of educational, global, family-friendly, and suspenseful content.**
- **Strategic Implication: If you're building a recommendation engine or curating content, focusing on the top three genres could cover a large portion of user interest.**

12 Data Visualization and Insights

- Visualize the distribution of content across release years.

```
df["release_year"].value_counts().head(10).plot(kind="bar")
plt.title("top ten year of movies release")
plt.ylabel("count")
plt.xlabel("release_year")
```

```
Text(0.5, 0, 'release_year')
```



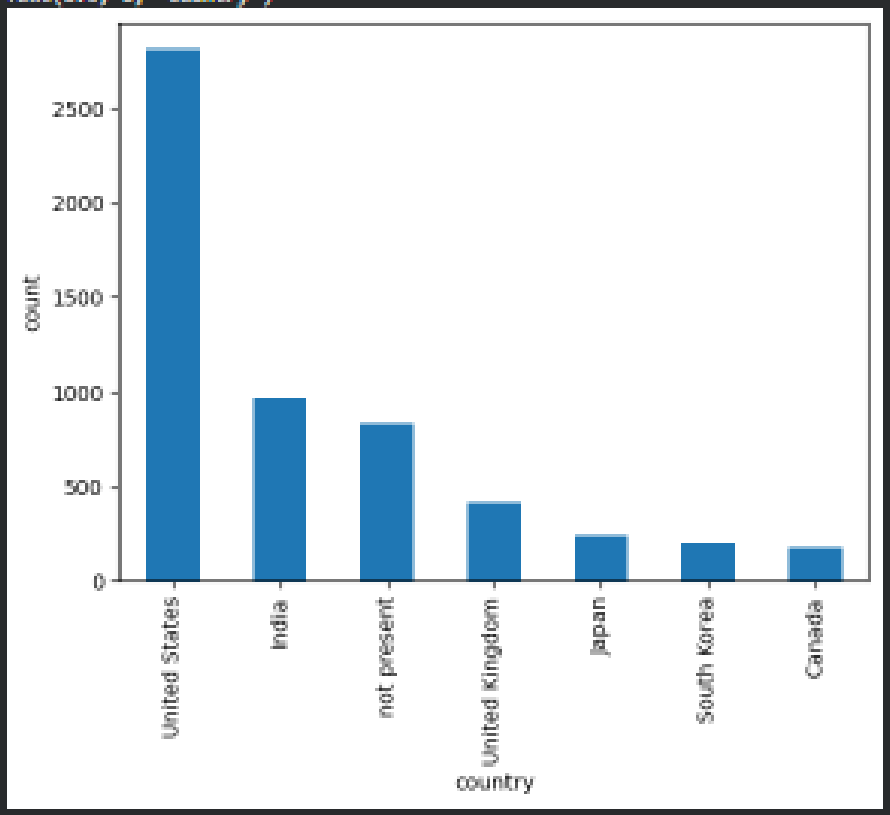
- Peak Year: 2016 stands out as the most prolific year for movie releases in the dataset, followed closely by 2017 and 2015.
- Recent Dominance: All top ten years fall between 2012 and 2018, indicating a strong concentration of content from the last decade.
- Content Boom: This trend suggests a surge in production or platform acquisitions during this period, possibly driven by streaming growth and global content expansion.
- Strategic Use: If you're analyzing viewer preferences or platform evolution, focusing on this 2012–2018 window could yield the richest insights.

13 Data Visualization and Insights

- Explore the geographical distribution of content (if applicable).

```
df["country"].value_counts().head(7).plot(kind="bar")
plt.ylabel("count")
plt.xlabel("country")
```

```
Text(0.5, 0, 'country')
```



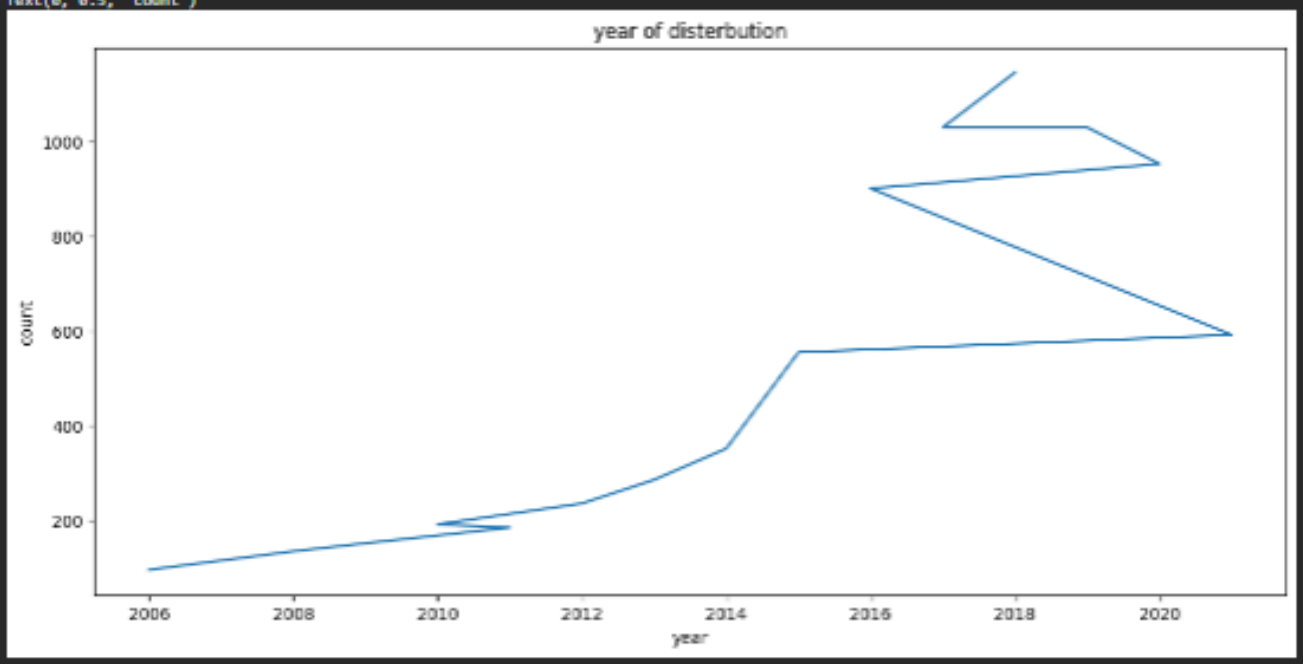
- **Dominant Producer: The United States overwhelmingly leads in content production, reflecting its global influence in entertainment.**
- **Strong Contributor: India ranks second, showcasing its prolific film and TV industry, especially in regional and Bollywood content.**
- **Data Gaps: The presence of "not present" as the third most frequent value highlights a significant portion of missing country data — a key issue to address during preprocessing.**
- **Global Spread: Other top contributors include the United Kingdom, Japan, South Korea, and Canada, indicating a diverse international catalog.**

14 Data Visualization and Insights

- If there's a temporal component, perform time series analysis to identify trends and patterns over time

```
plt.figure(figsize=(13,6))
df["release_year"].value_counts().head(15).plot(kind="line")
plt.title("year of disterbution")
plt.xlabel("year")
plt.ylabel("count")
```

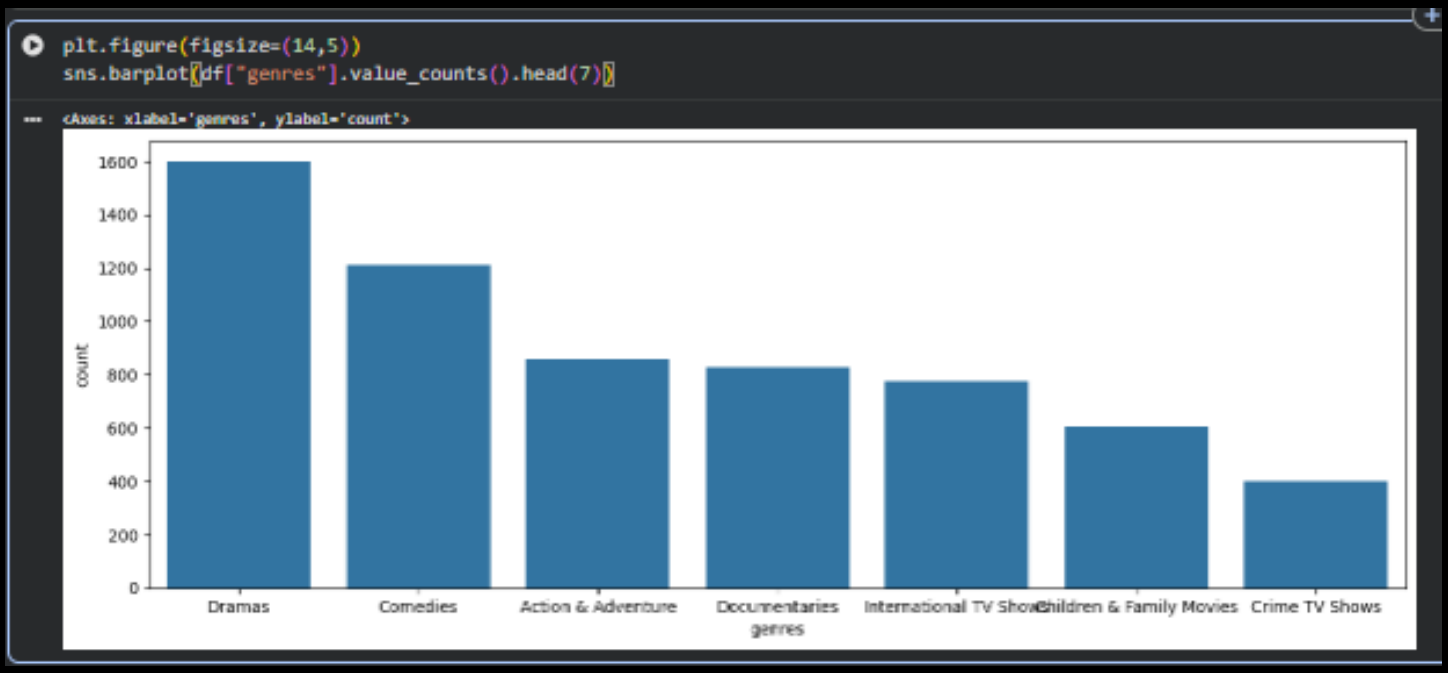
```
Text(0, 0.5, 'count')
```



- **Upward Trend: There's a clear rise in content releases starting around 2014, peaking near 2019, which likely reflects the streaming boom and increased global production.**
- **Content Surge: The sharp incline suggests platforms were rapidly expanding their libraries during this period, possibly due to competition and demand for original content.**
- **Strategic Window: The 2014–2019 range appears to be the most active and relevant for trend analysis, viewer behavior studies, or platform growth evaluation.**

15 Data Visualization and Insights

- Analyze the distribution of content ratings.



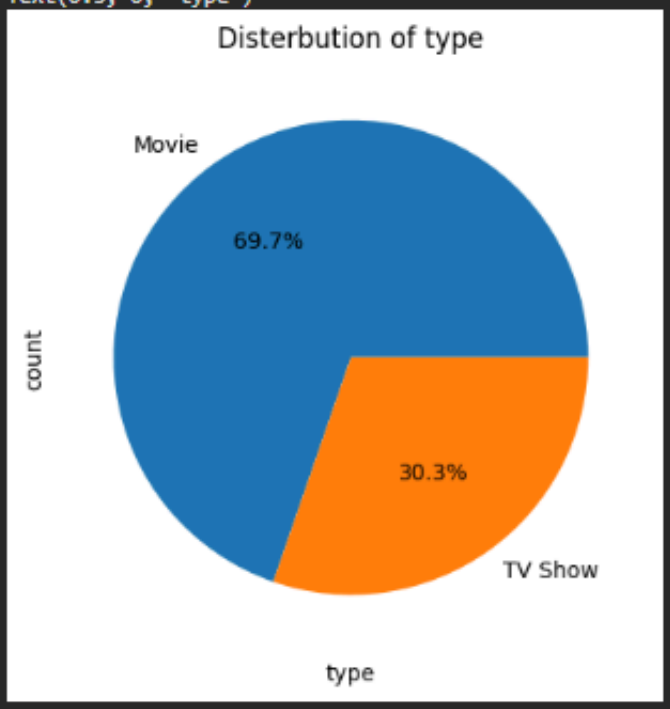
- Genre Leader: Dramas dominate the dataset, with over 1400 entries — a clear favorite across regions.
- Comedy & Action Appeal: Comedies and Action & Adventure follow closely, reflecting strong global demand for entertainment and thrill.
- Diverse Offerings: The presence of Documentaries, International TV Shows, and Children & Family Movies highlights a well-rounded catalog catering to education, cultural variety, and family audiences.
- Niche Interest: Crime TV Shows, while less frequent, still hold a notable share, suggesting a dedicated viewer base for suspense and mystery.

16 Data Visualization and Insights

- Explore the length of movies or episodes and identify any trends.

```
df["type"].value_counts().plot(kind="pie",autopct="%1.1f%%")
plt.title("Disterbution of type")
plt.xlabel("type")
```

```
Text(0.5, 0, 'type')
```



- Movies Dominate: With 69.7%, movies form the bulk of the dataset, indicating a strong emphasis on cinematic content.
- TV Shows Hold Ground: 30.3% of entries are TV shows, reflecting a substantial share and growing interest in episodic storytelling.
- Strategic Implication: If you're analyzing viewer habits or platform focus, this split suggests prioritizing movie-related features while still catering meaningfully to TV audiences.

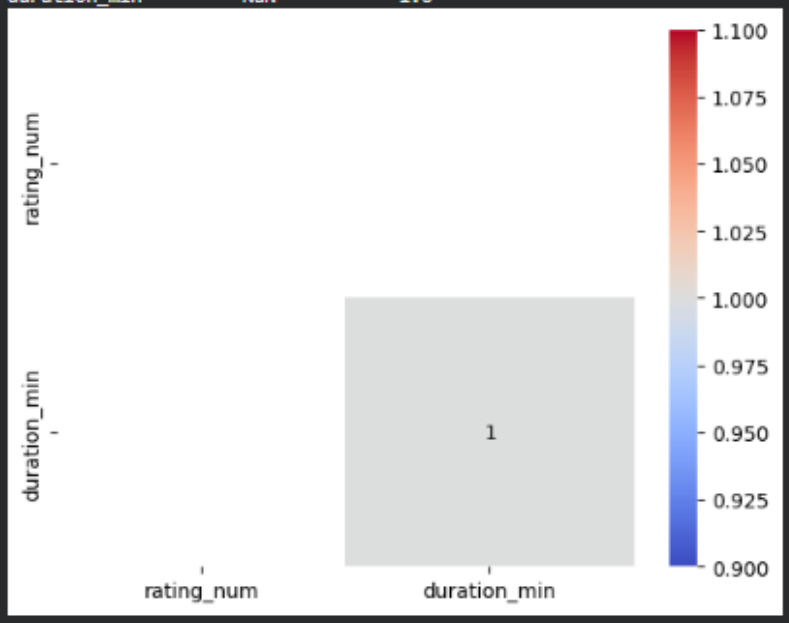
17 Data Visualization and Insights

- **Correlation Analysis: Investigate potential correlations between variables (e.g., ratings and duration).**

```
corr = df[["rating_num", "duration_min"]].corr()
print(corr)

sns.heatmap(corr, annot=True, cmap="coolwarm")
plt.show()
```

```
              rating_num  duration_min
rating_num           NaN           NaN
duration_min         NaN           1.0
```



- **No Relationship Detected: The correlation value between rating_num and duration_min is NaN, meaning it's undefined — likely due to missing or non-numeric data in one or both columns.**
- **Implication: This suggests that either: One of the columns contains insufficient valid data for correlation calculation.**
- **The data types or preprocessing steps (e.g., conversion to numeric) may not have been properly handled.**
- **Next Step: Investigate the rating_num and duration_min columns for: Null or non-numeric entriesInconsistent formattingNeed for imputation or type conversion**

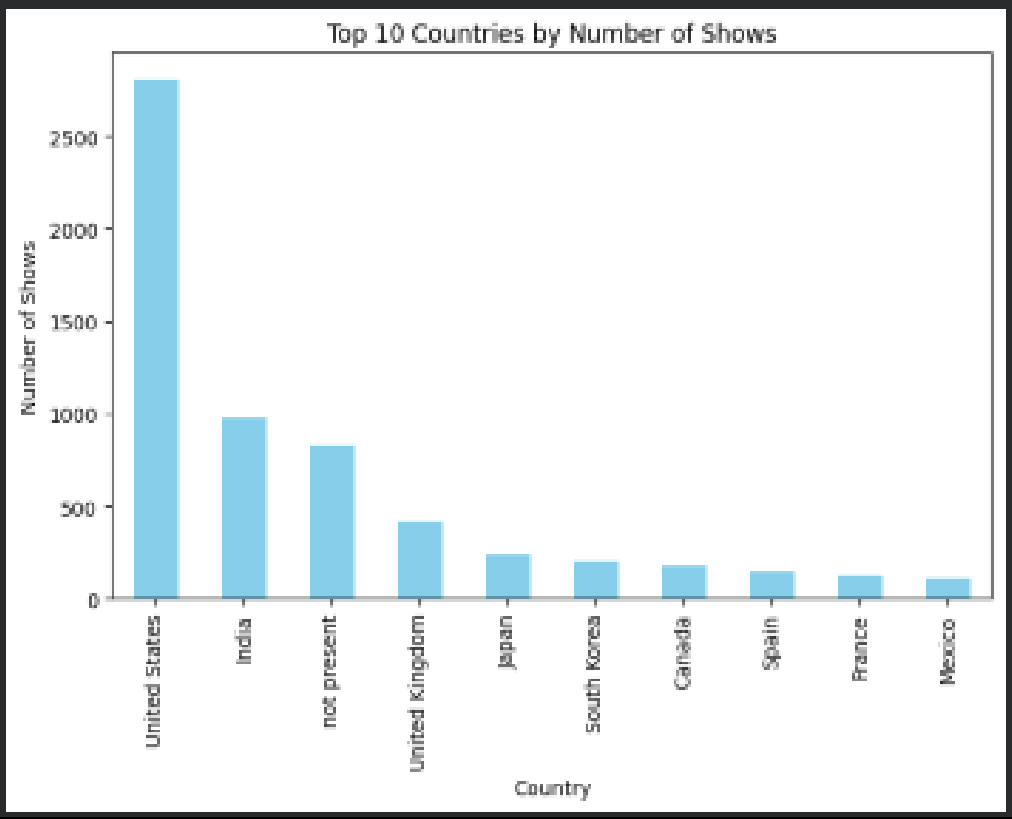
18 Data Visualization and Insights

- **Geographical Analysis:**

Further explore the distribution of content across different countries and regions.

```
country_counts = df['country'].value_counts().head(10)  # top 10 countries

# Bar plot
country_counts.plot(kind='bar', figsize=(8,5), color='skyblue')
plt.title("Top 10 Countries by Number of Shows")
plt.xlabel("Country")
plt.ylabel("Number of Shows")
plt.show()
```



- **United States Leads: With the highest number of shows, the U.S. dominates the dataset, reflecting its massive entertainment output and global reach.**
- **India in Second Place: India's strong showing highlights its vibrant film and television industry, especially in regional and Bollywood content.**
- **Data Gaps Noted: The third most frequent entry is "not present", indicating a significant portion of missing country data — a key issue to address during data cleaning.**
- **Diverse Contributors: Countries like the United Kingdom, Japan, South Korea, Canada, Spain, France, and Mexico round out the top 10, showcasing a rich international mix of content.**

Conclusion

1. Content Type Distribution
   - Movies dominate the dataset with nearly 70%, while TV Shows make up the remaining 30%.
   - This indicates a platform or catalog heavily skewed toward cinematic content.

2. Genre Popularity
   - Dramas are the most frequent genre, followed by Comedies and Action & Adventure.
   - The genre mix reflects a balance of emotional depth, humor, and thrill — appealing to a wide audience.

3. Release Year Trends
   - Most content was released between 2012 and 2019, with a peak around 2016–2017.
   - This suggests a surge in production or acquisitions during the streaming boom.

4. Country Representation
   - The United States leads in content volume, followed by India, UK, Japan, and South Korea.
   - A significant portion of entries have missing country data ("not present"), which should be addressed in preprocessing.

5. Correlation Analysis
   - No valid correlation was found between rating_num and duration_min, likely due to missing or improperly formatted data.
   - This highlights the need for data cleaning before deeper statistical analysis.

Thanks for reading

Thank you

For coding part.....

Google Colab
co google.com