EDA Report

Dataset Name: **Netflix Movies and TV Shows**

Date: **Jun 6, 2025**

Loading dataset	import kagglehub path = kagglehub.dataset_download("shivamb/netflix-shows") df = pd.read_csv(f"{path}/netflix_titles.csv")
Exploring dataset	1. Finding number of rows and columns present in dataset 2. Getting all column's names 3. Extracting overall data information - tells not null count and data type of each column 4. Getting statistical information for numeric column 5. Finding count of total null values in all columns 6. Checking if any duplicated row is present
Cleaning approach	1. Removing duplicates 2. Converting to proper data type 3. Filling up missing values in some columns
Drawing Important Insights	1. Netflix Content Breakdown by Type 2. Number of Netflix Titles Released Each Year 3. Most Popular Content Genres on Netflix 4. Leading Countries Producing Content for Netflix 5. Actors with the Most Appearances in Netflix Content 6. Growth of Movies vs TV Shows on Netflix Over the Years 7. Netflix Content Distribution Over the Decades

Loading Dataset

Source : Netflix Movies and TV Shows

In order to use it's code one must have kagglehub library installed in their pc

To install it , run : **pip install kagglehub**

Exploring Dataset

1.Finding number of rows and columns present in dataset

```
print("Total rows , columns = " , df.shape)
```

```
Total rows , columns =  (8807, 12)
```

2.Getting all column’s names

```
print("Column names = " , df.columns)
```

```
Column names =  Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

3.Extracting overall information

```
[24]: print(df.info()) #To check for not null value's count and data type of the column
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 13 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      8807 non-null   object        
 4   cast          8807 non-null   object        
 5   country       8807 non-null   object        
 6   date_added    8709 non-null   datetime64[ns]
 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        
 12  decade        8807 non-null   int64         
dtypes: datetime64[ns](1), int64(2), object(10)
memory usage: 894.6+ KB
None
```

4.Statistical Information of numeric column

```
print(df.describe())
```

```
       release_year
count   8807.000000
mean    2014.180198
std        8.819312
min     1925.000000
25%     2013.000000
50%     2017.000000
75%     2019.000000
max     2021.000000
```

5.Count of null values in each column

```
print("Null values = \n" , df.isnull().sum())
```

```
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
listed_in           0
description         0
dtype: int64
```

6.Checking for duplicated rows

```
print("Duplicated rows = " , df.duplicated().sum())
```

```
Duplicated rows =  0
```

Cleaning Dataset

1.Removing Duplicates (if any by chance)

```
#Removing duplicates
df.drop_duplicates(inplace=True)
```

2.Converting “date_added” column’s data type

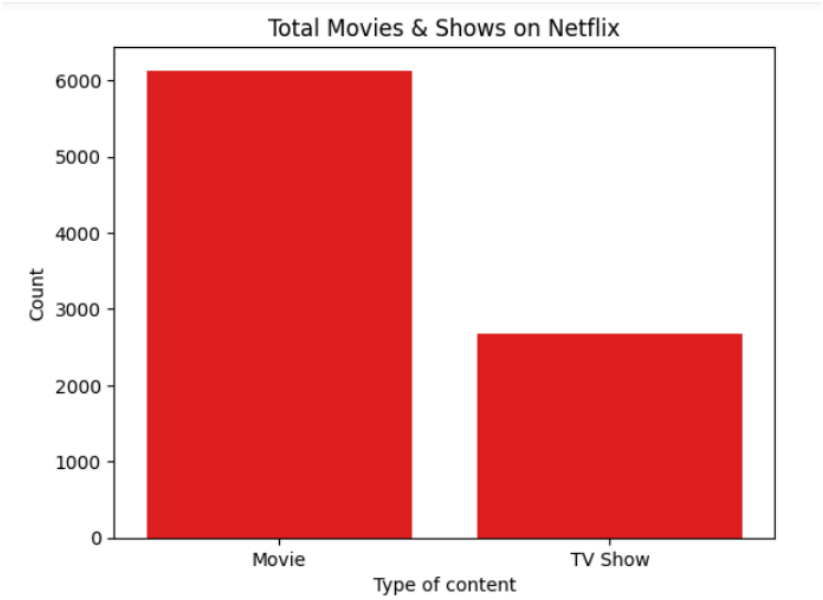
```
df["date_added"] = pd.to_datetime(df["date_added"] , errors = "coerce")
```

3.Filling out missing values

```
#Fillig null values in some columns
df["country"] = df["country"].fillna("Unknown")
df["cast"] = df["cast"].fillna("Unknown")
df["director"] = df["director"].fillna("Unknown")
```

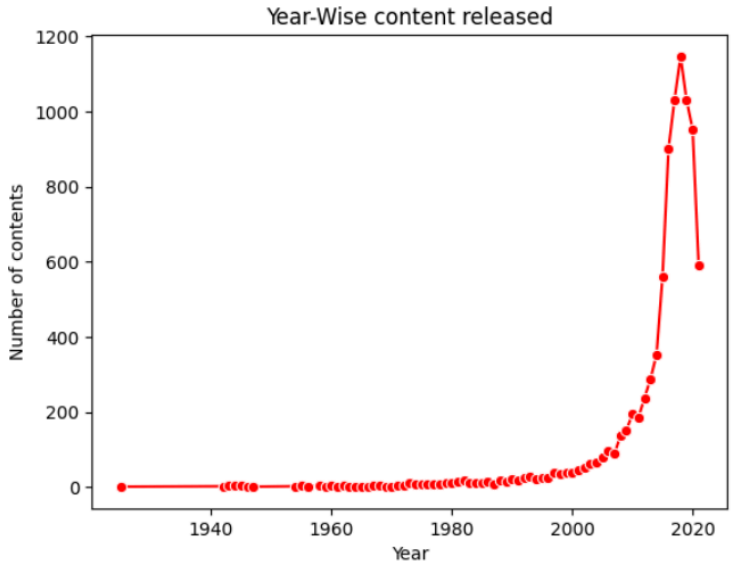
Visualization

1.Netflix Content Breakdown by Type



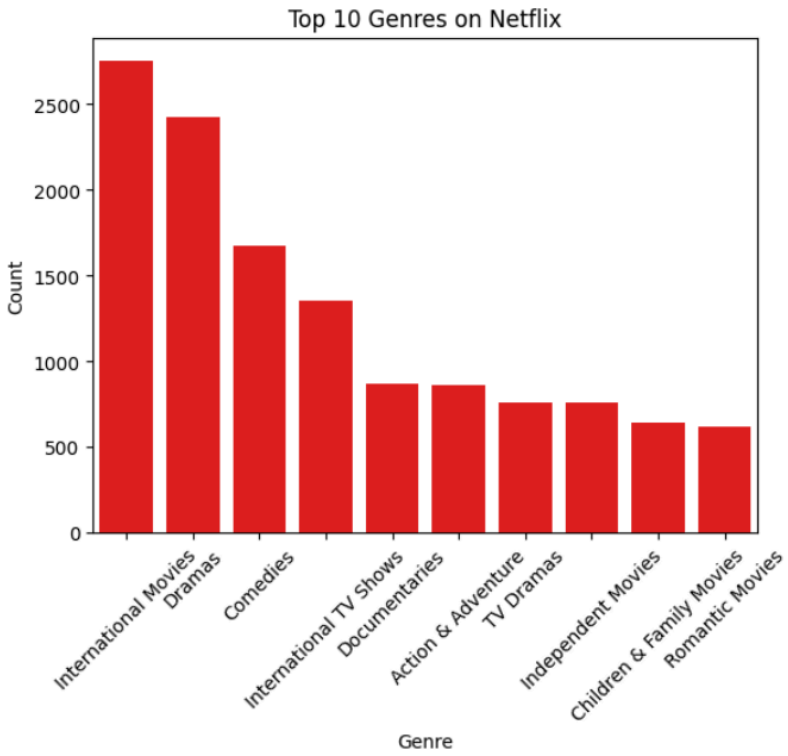
Netflix features a significantly larger number of movies than TV shows, suggesting its strategy leans toward offering more film-based content. This could be aimed at attracting viewers who prefer quick, complete stories in one sitting, rather than committing to multi-episode series.

2.Number of Netflix Titles Released Each Year



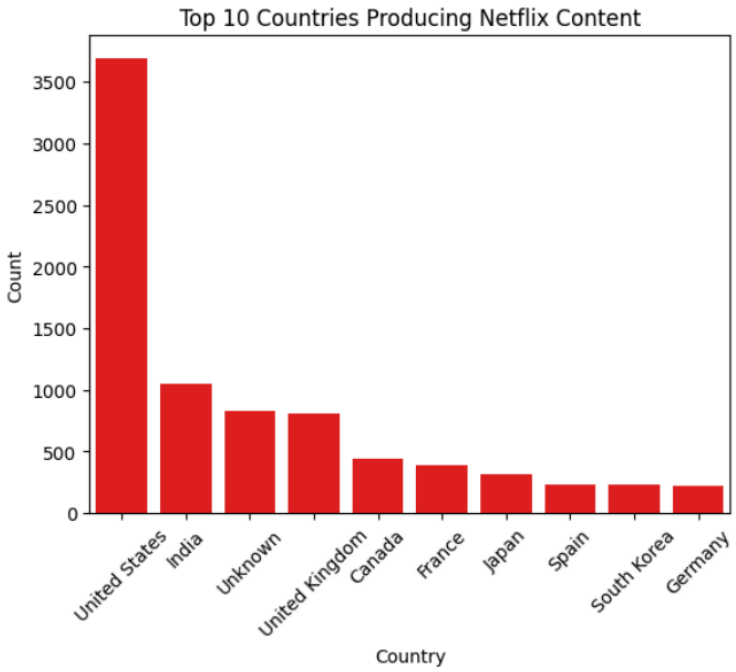
Netflix saw a sharp rise in content releases after 2010, peaking around 2019. This reflects its rapid expansion, though releases slightly dropped post-2019, likely due to global disruptions.

3.Most Popular Content Genres on Netflix



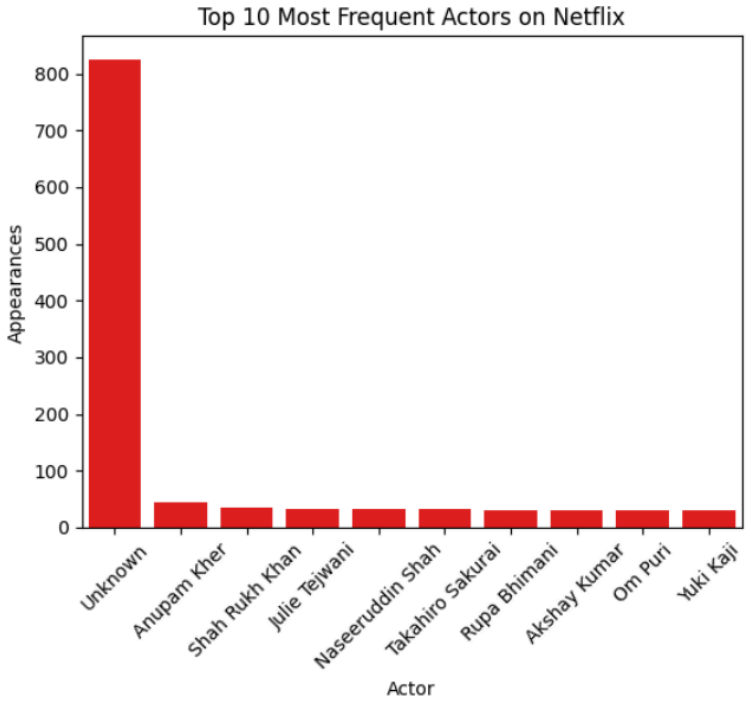
"International Movies" and "Dramas" are the most prevalent genres on Netflix, indicating a strong global and emotional content appeal. Comedies and International TV Shows also hold significant presence, while genres like Romantic Movies and Children & Family Movies have relatively lower counts but are a part of top 10 genres on Netflix.

4.Leading Countries Producing Content for Netflix



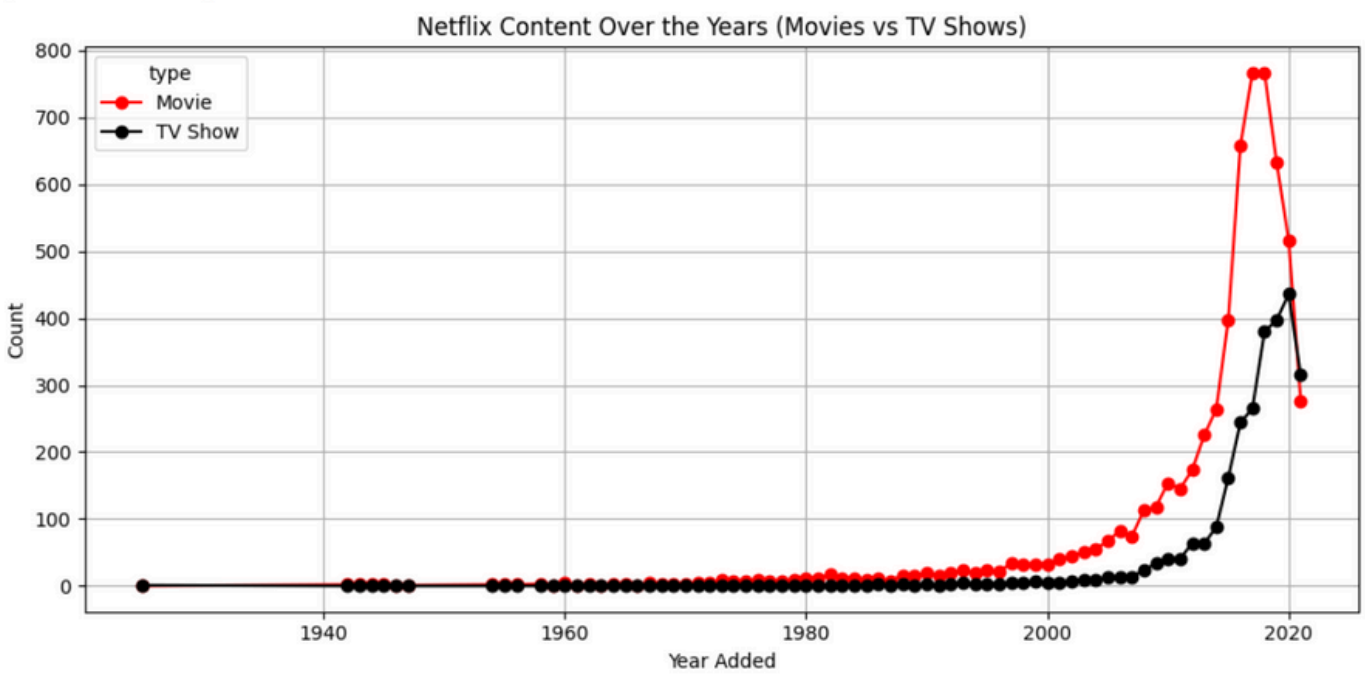
The United States dominates Netflix content production by a wide margin, followed by India and the UK.
Other countries like Canada, France, and Japan contribute modestly, while South Korea and Germany appear at the lower end among the top 10 producers.

5.Actors with the Most Appearances in Netflix Content



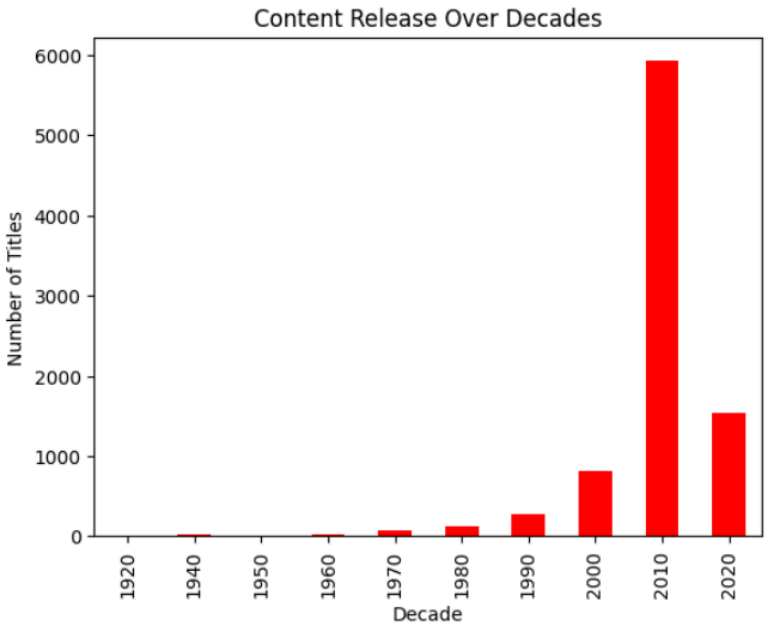
A significant portion of actor data on Netflix is listed as "Unknown," indicating missing metadata.
Among identified actors, Anupam Kher, Shah Rukh Khan, and Julie Tejwani are the most frequently appearing, reflecting strong representation from Indian cinema.

6.Growth of Movies vs TV Shows on Netflix Over the Years



Netflix content, especially movies, saw a sharp rise after 2015, peaking around 2018–2019.
TV shows also grew significantly during the same period, although at a slower pace, with both formats seeing a slight decline post-2020.

7.Netflix Content Distribution Over the Decades



Content production on Netflix grew steadily over the decades, peaking sharply in the 2010s.
The 2020s show a dip in releases, indicating a possible shift in production trends or external challenges.