DATA ANALYSIS ON NETFLIX DATASET.

BY

RIDDHIMA KAURA

Table of Contents

Tip: Click on the text and land on the page you want to visit.

Tip: Click on the text and land on the page you want to visit.

Tip: Click on the text and land on the page you want to visit.

Tip: Click on the text and land on the page you want to visit.

PROJECT SUMMARY

The Netflix EDA project is focused on conducting exploratory data analysis on a dataset related to Netflix content, which includes details of movies and TV shows accessible on the platform. The primary objective of this project is to uncover valuable insights, detect trends, and draw conclusions from the data. To accomplish this, the dataset is analyzed using various techniques, including data cleaning, visualizations, and statistical analysis. The ultimate goal is to gain a comprehensive understanding of the data, which can aid in making informed decisions and formulating effective strategies.

RIDDHIMA KAURA

ALL ABOUT THE PROJECT

The Netflix EDA project is a comprehensive analysis of a dataset related to Netflix content, which includes information about movies and TV shows available on the platform. The dataset contains various attributes such as show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description. The analysis is performed using Python programming language and the Pandas library to explore and manipulate the data.The dataset contains 8807 records with 12 columns, with show_id, type, title, and listed_in being the most unique values. The dataset has been explored for missing values, duplicates, and data types.

RIDDHIMA KAURA

The unique values in the listed_in column suggest that the dataset covers various genres of content, including documentaries, TV shows, crime, action, adventure, and comedy.The dataset has been analyzed to extract meaningful insights and draw conclusions. For instance, the analysis reveals that the dataset has a significant number of records with missing values in the director and cast columns. Additionally, the dataset has some duplicates, which have been removed to ensure the accuracy of the analysis.The analysis also includes visualizations to explore the distribution of the data and identify trends. For instance, the distribution of release_years reveals that the dataset covers content from various years, with a significant number of records from recent years.

Similarly, the analysis of the duration column reveals that the dataset contains content with varying durations, ranging from a few minutes to several hours.Overall, the Netflix EDA project aims to provide a comprehensive analysis of the dataset to aid in making informed decisions and formulating effective strategies. The analysis can help in understanding the trends and patterns in the Netflix content, identifying the most popular genres, and evaluating the quality of the content.

RK
RIDDHIMA KAURA

WHAT IS THE SCOPE OF THE PROJECT

The scope of the project involves performing an exploratory data analysis on a dataset related to Netflix content. The dataset includes information about movies and TV shows available on the platform, with attributes such as show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description. The project aims to extract meaningful insights and draw conclusions from the dataset by performing various data analysis tasks, such as:

RIDDHIMA KAURA

- Exploring the data: The project begins by exploring the dataset to understand its size, dimensions, and structure. The first few records are displayed to get a sense of the data.
- Summary of the dataset: A detailed summary of the dataset is provided, including the number of non-null values, data types, and memory usage.
- Unique values: The number of unique values for each column is calculated to identify the diversity of the data.
- Checking for duplicates: The project checks for any duplicate records in the dataset and removes them to ensure data accuracy.
- Data cleaning: The project fills in missing values in the director, cast, and country columns with a placeholder value of "Unknown_Value" to maintain the integrity of the analysis.

- Data validation: The project drops any records with missing values in the date_added, rating, and duration columns to ensure complete data.
- Data analysis: The project analyzes the unique values in the title, type, cast, director, and listed_in columns to understand the diversity of the dataset.
- Data visualization: The project visualizes the distribution of the release_year, rating, and duration columns to identify trends and patterns in the data.

Overall, the project aims to provide a comprehensive analysis of the Netflix dataset to aid in making informed decisions and formulating effective strategies for content creation, distribution, and marketing.

LIMITATIONS OF THE DATASET

From the data analysis on the Netflix dataset, the limitations of this project could include:

Missing Data: The dataset contains missing values in several columns, such as director, cast, country, date_added, rating, and duration. These missing values could impact the accuracy and completeness of the analysis.

Incomplete Information: Some columns have incomplete information, which may limit the depth of analysis that can be performed on certain aspects of the dataset.

Data Quality: The quality of the data, including the accuracy of the information provided in the dataset, could affect the reliability of the analysis and conclusions drawn from it.

Sample Size: The dataset contains 8,807 records, which may not be representative of the entire Netflix library, potentially leading to biased conclusions.

WERE THERE ANY DATA QUALITY ISSUES THAT AFFECTED THE RESULTS OF THE DATA ANALYSIS ON THE NETFLIX DATASET

There were indeed data quality issues that could have affected the results of the data analysis. Some of the key data quality issues include:

Missing Values: The dataset contains missing values in columns such as director, cast, country, and date_added. These missing values could impact the accuracy and completeness of the analysis, especially when analyzing specific attributes like director or cast members.

Inconsistent Data Entry: The dataset may have inconsistencies in data entry, leading to discrepancies or errors in the dataset. Inconsistent data entry could affect the reliability of the analysis results and lead to incorrect conclusions.

Data Format Issues: The data types of certain columns may not be appropriate for the analysis conducted. For instance, if date_added is not in a standardized date format, it could affect date-based analyses.

Duplicate Entries: The dataset may contain duplicate entries, which could skew the analysis results, especially when counting unique titles or genres.

Data Integrity: The presence of missing values and inconsistencies in the dataset could compromise the overall data integrity, potentially leading to biased or inaccurate analysis results.

These data quality issues should be considered when interpreting the results of the data analysis on the Netflix dataset, as they could have influenced the conclusions drawn from the analysis.

RIDDHIMA KAURA

WERE THERE ANY MEASURES TAKEN TO ADDRESS THE DATA QUALITY ISSUES IN THE DATA ANALYSIS ON NETFLIX DATASET

The following measures were taken to address data quality issues in the data analysis on the Netflix dataset:

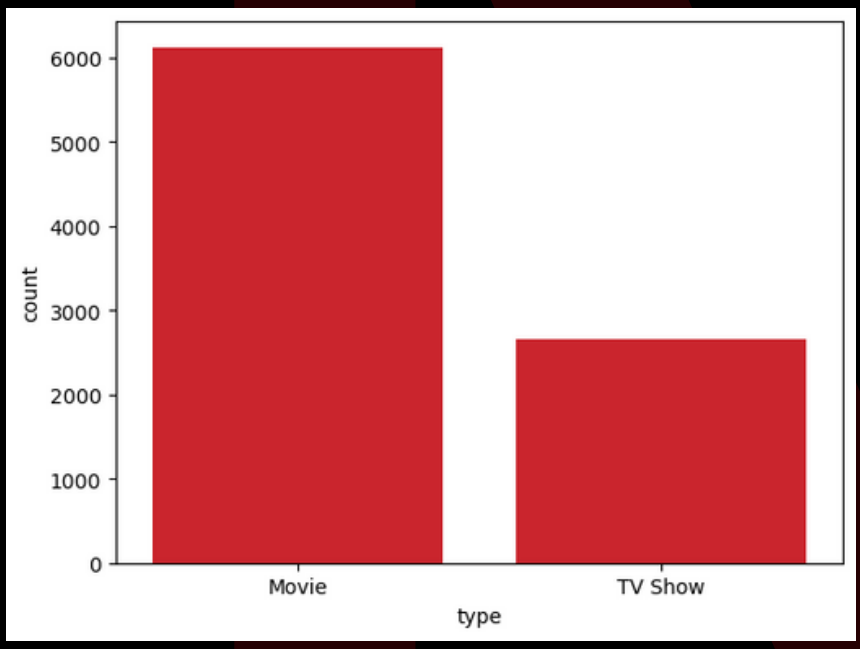
Handling Missing Values: The dataset contained missing values in several columns, including director, cast, country, date_added, rating, and duration. To address this issue, the data analysis used various techniques to handle missing values, such as removing rows with missing values, filling missing values with appropriate values, or using methods like forward-fill, backward-fill, or mean imputation.

Data Cleaning: The dataset contained inconsistent data entry, such as different formats for the duration. To address this issue, the data analysis used data cleaning techniques to standardize the format of these columns, such as converting duration to a consistent format.

Data Integrity: The dataset contained duplicate entries, which could affect the accuracy of the analysis. To address this issue, the data analysis used techniques to remove duplicate entries, such as using the drop_duplicates() function in pandas.

By using these measures, the data analysis on the Netflix dataset aimed to improve the quality of the data and ensure that the results were accurate and reliable.

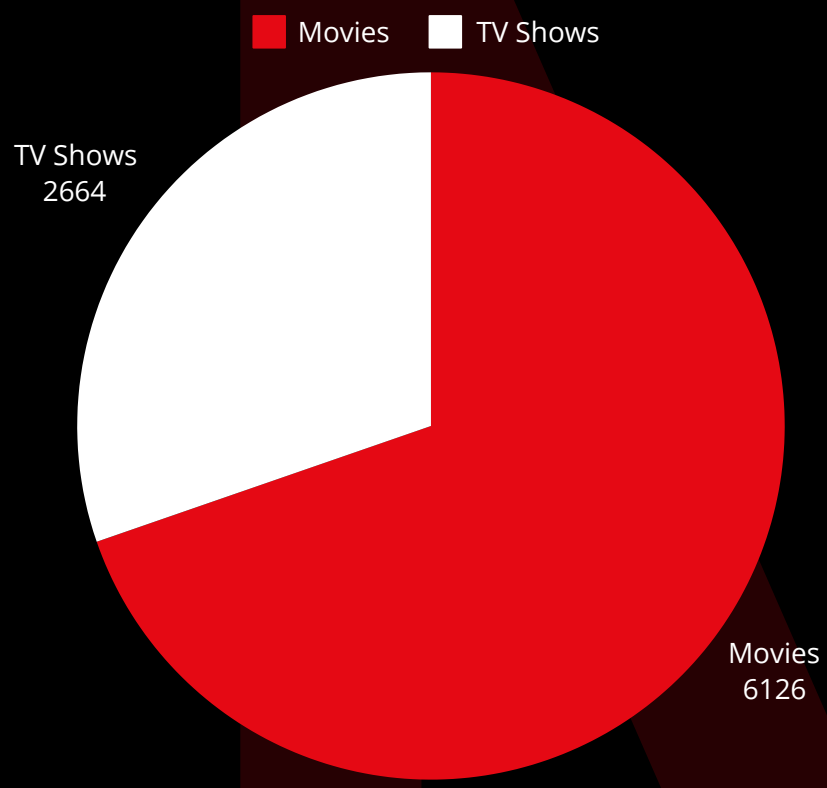
ASSESSING THE MOST-OCCURRING AND LEAST-OCCURRING TYPE AVAILABLE ON NETFLIX USING COUNTPLOT



-The above countplot depicts that Netflix subsume more movies as compared to TV Shows.

RIDDHIMA KAURA

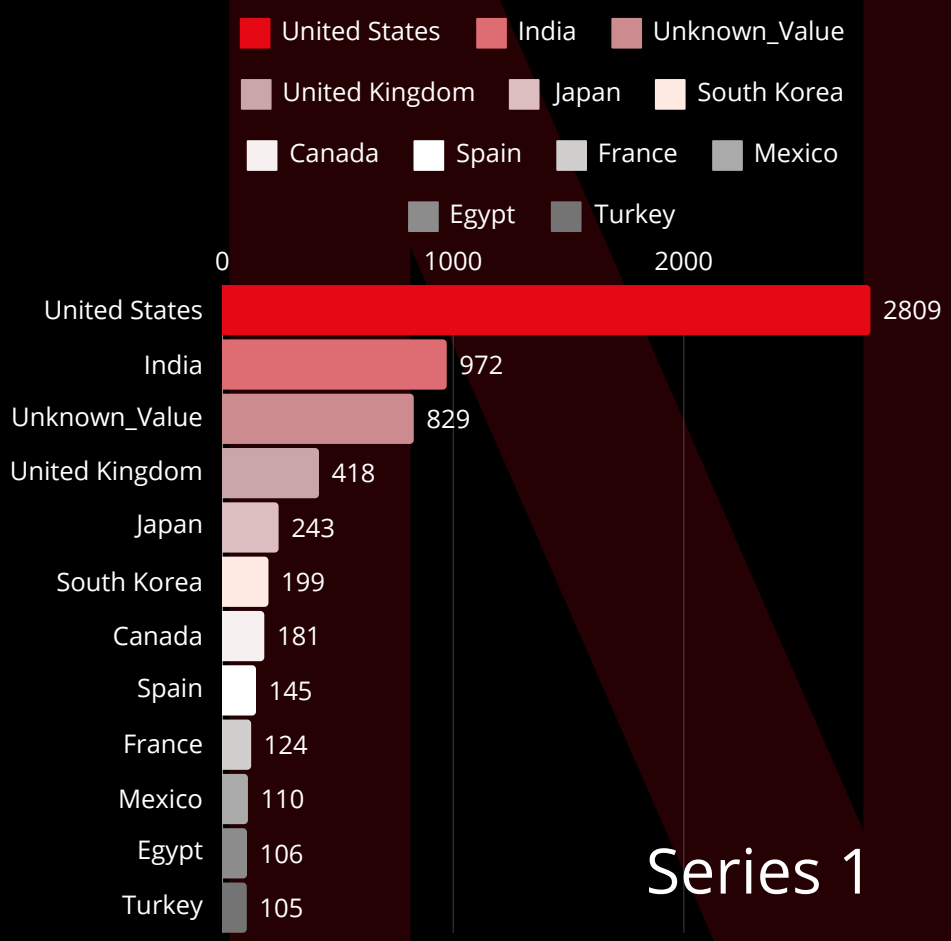
PROVIDING THE OCCURENCES OF EACH UNIQUE VALUE.



As is presented by using value_counts movies available on Netflix are 6126 and TV Shows are 2664.

RIDDHIMA KAURA

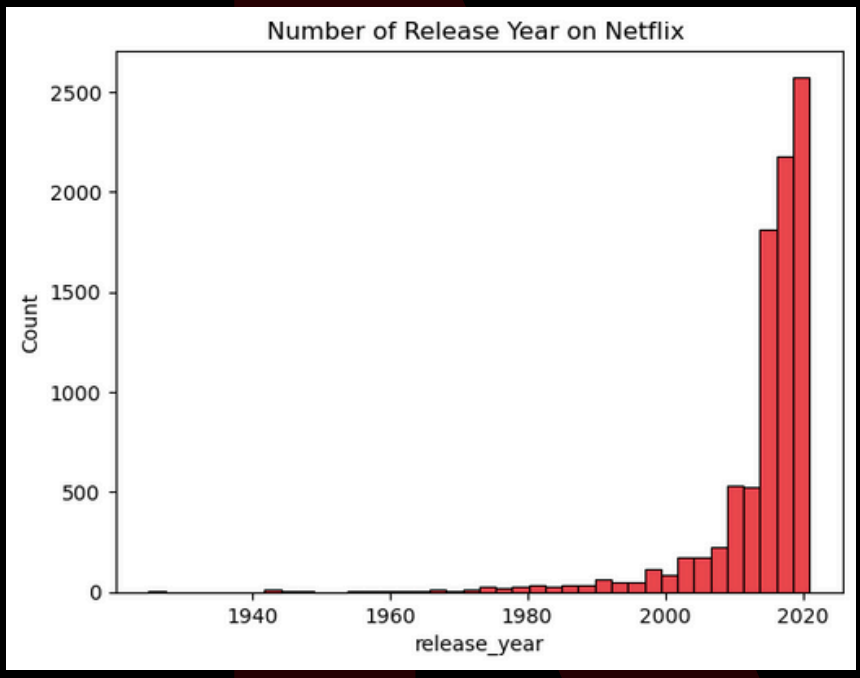
12 COUNTRIES THAT HIT THE PEAK IN TERMS OF CONTENT



The supplied graph gives information on 12 Countries that are on top in terms of content. It could be plainly viewed that the United States is actively using Netflix and has the most available worldwide content. On the other hand, Turkey has the least titles in top 12.

RIDDHIMA KAURA

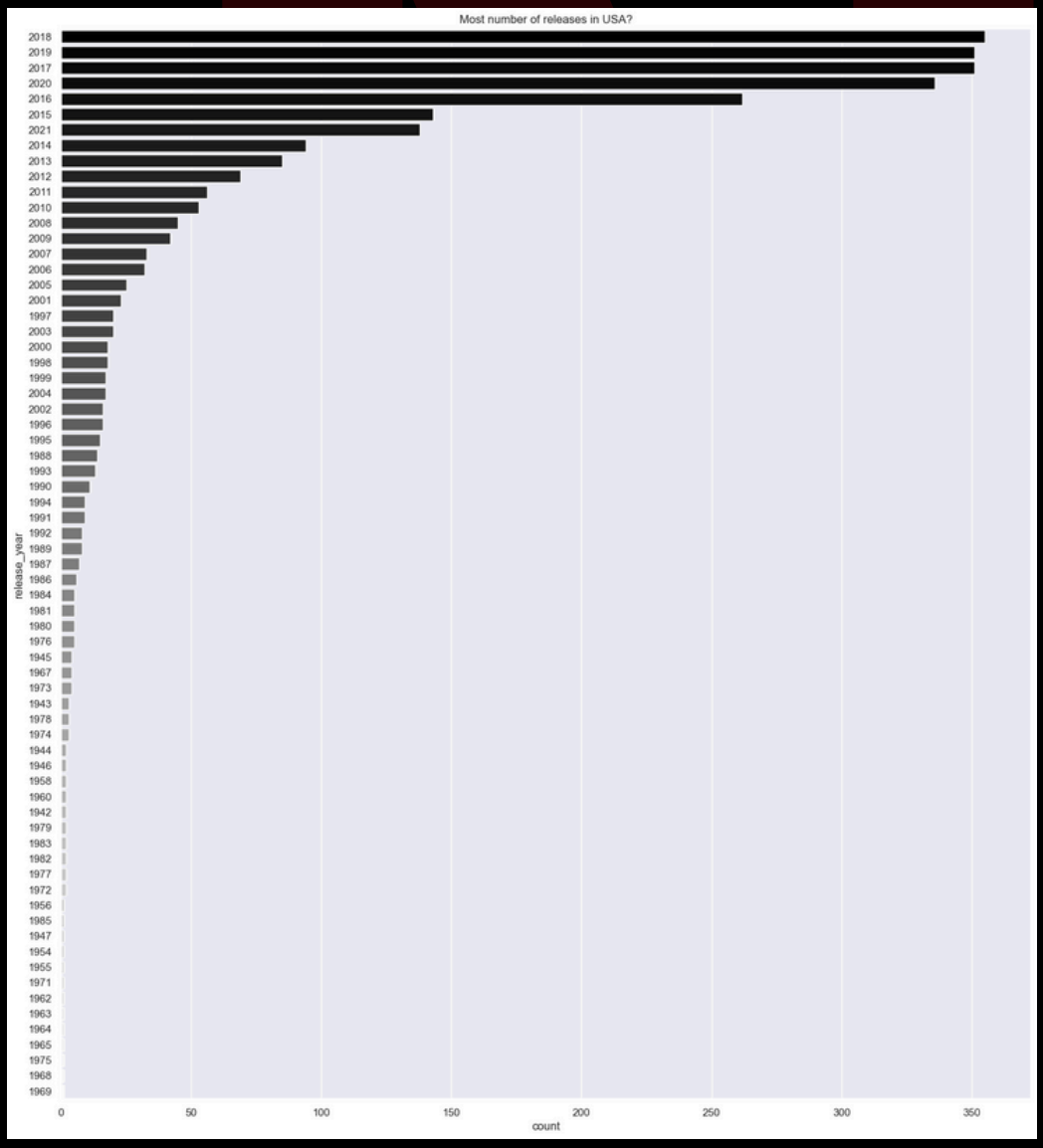
WHEN WAS NETFLIX GETTING POPULAR?



It is explicitly observed that Netflix had large amount of streaming traffic during the year 2016-21 with a significant number of movies and TV shows being added to its platform during that period. This influx of content suggests a time when Netflix was gaining popularity and expanding its library to cater to a diverse audience.

RIDDHIMA KAURA

IN WHICH YEAR DID NETFLIX GET POPULAR WITH THE MOST NUMBER OF RELEASES IN USA?



RIDDHIMA KAURA

The countplot presents the Netflix data which shows the number of releases over 5 decades, commencing from 1969. As is observed, In year 2018 Netflix was getting popular with the most number of releases in the USA. In the years 2017 and 2019, the number of releases was precisely the same i.e., 350. While an actual increasing trend of Netflix began in 2016 in the USA and after that, it kept increasing.

RIDDHIMA KAURA

LIST OF MOVIES FROM USA ON NETFLIX HOLLYWOOD

There are a total of 2055 Movies from the USA on Netflix.

Here are the movies from the USA on Netflix Hollywood:

An Unfinished Life (2005)
Barbie Big City Big Dreams (2021)
Blade Runner: The Final Cut (1982)
Catch Me If You Can (2002)
Cloudy with a Chance of Meatballs (2009)
Deep Blue Sea (1999)
Ferris Bueller's Day Off (1986)
Freedomland (2006)
Friday Night Lights (2010)
Magnolia (1999)

RIDDHIMA KAURA

Major Payne (1995)
My Girl (1991)
My Girl 2 (1994)
Open Season 2 (2008)
Osmosis Jones (2001)
Pineapple Express (2008)
Planet 51 (2009)
Poms (2019)
Seabiscuit (2003)
Space Cowboys (2000)
The Edge of Seventeen (2016)
The Machinist (2004)

These movies span various genres, from drama and comedy to action and animation, providing a diverse range of viewing options for Netflix subscribers.

RK
RIDDHIMA KAURA

LIST OF TV SHOWS FROM USA ON NETFLIX HOLLYWOOD

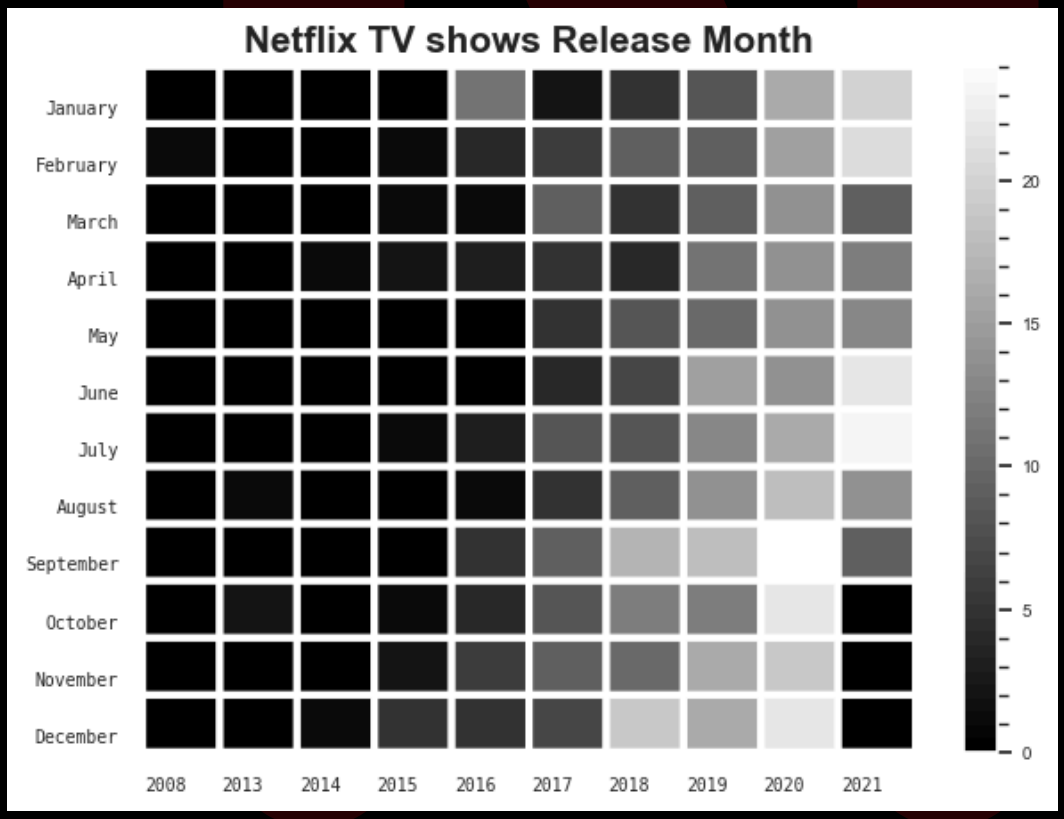
There are a total of 754 TV Shows from the USA on Netflix. Some TV shows from the USA on Netflix:

Saved by the Bell (1994)
The Defeated (2020)
Too Hot To Handle: Latino (2021)
The Flash (2021)
The Snitch Cartel: Origins (2021)
30 Rock (2012)
Grace and Frankie (2021)
Cooking With Paris (2021)
Top Secret UFO Projects: Declassified (2021)
Two Fathers (2013)

These shows offer a range of genres, from comedy and drama to action and adventure, providing something for everyone to enjoy.

RIDDHIMA KAURA

IN WHICH MONTH OF THE YEAR COMMENCING FROM 2008 NETFLIX RELEASED NEW TV SHOWS THE MOST IN USA?



RIDDHIMA KAURA

Potentially, a new Netflix TV Show could be released any month of the year. Regardless of how the presented heatmap shows which month of the year Netflix released the most new TV Shows in the United States starting in 2008.

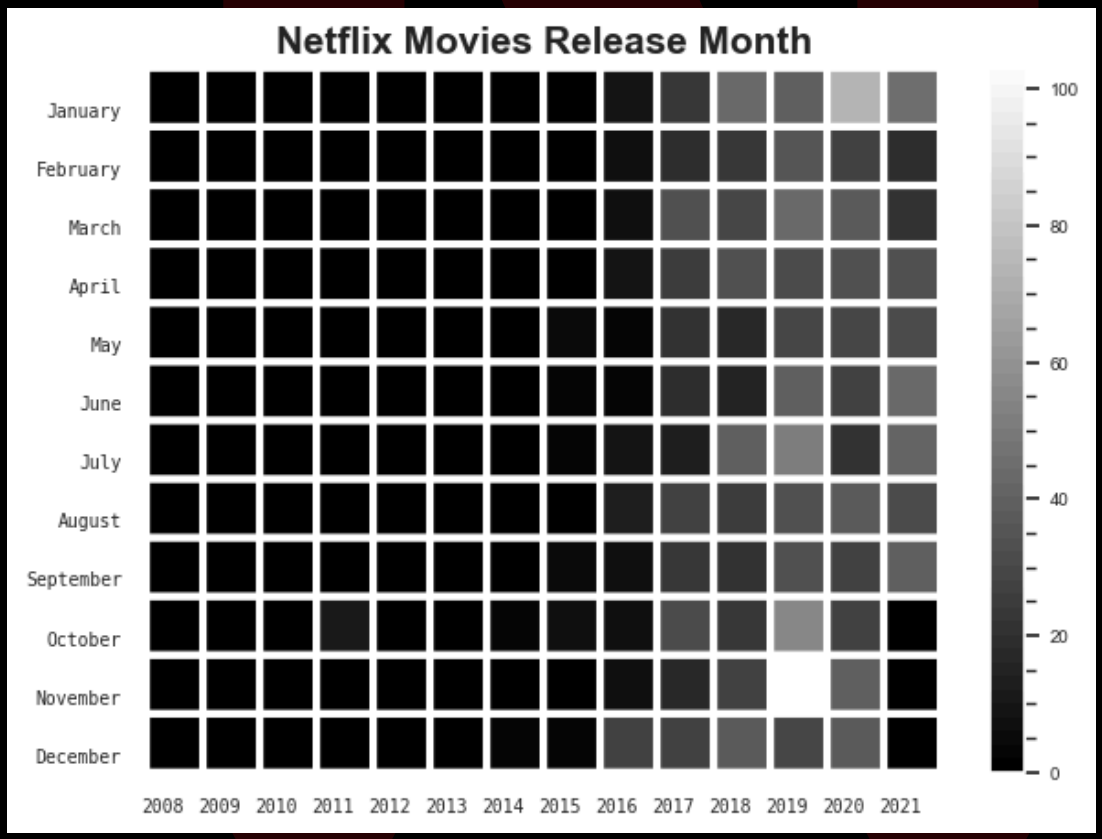
As is obvious, the number of releases in the four mentioned years i.e., 2008, 2013, 2014, and 2015 shows almost the same pattern and barely any release in sight.

In the year 2016, it is obvious that most of the releases were in January, and in the year 2017, scarcely any release was in sight in January in the rest of the month the pattern changes intermittently, and few releases are observed.

During 2018 and 2019, Netflix started getting popular with many releases in every month of the year mainly in the first half of the year.

It is interesting to note that most of the releases can be seen in the year 2020 attributable to COVID-19. while in the year 2021, many releases can be seen in every month except Oct, Nov, and December.
To put it simply, the trend of releasing content on Netflix mainly started between 2017 and 2018 and became prominent from 2020.

IN WHICH MONTH OF THE YEAR COMMENCING FROM 2008 NETFLIX RELEASED NEW MOVIES THE MOST IN USA?

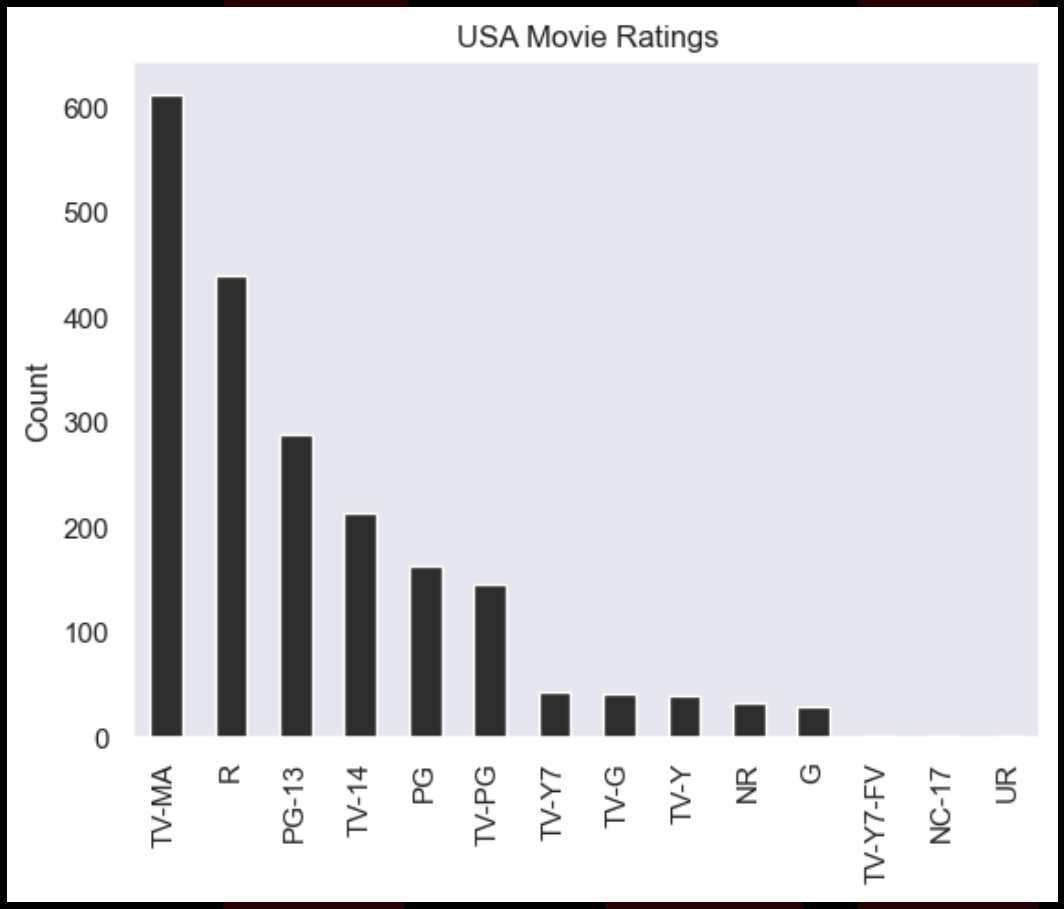


RIDDHIMA KAURA

It could be viewed that between 2008 and 2016 shows the same pattern and barely any release in sight. Starting from December 2016, the trend of releasing movies on Netflix started.

Most of the releases are in 2019 and 2020.

USA MOVIE RATINGS



As is observed, TV-MA is exactly 600. Generally speaking, TV-MA is a rating meant for mature audiences, designed for adults, and potentially inappropriate for individuals under 17, as it may include explicit language, sexual content, or violent scenes.

Just under TV-MA, rating R is roughly 400. The film with a rating of R is suitable for viewers aged 17 and above without the need for a parent or guardian. However, it contains certain mature content, and parents are advised to familiarize themselves with the film's content before taking their younger children to watch it.

PG-13 is a little less than 300. Parents are advised to exercise caution due to some content that may not be suitable for children under 13 years of age.

TV-14 is nearly around 200. Parents are strongly cautioned as this program includes content that may not be suitable for children under 14 years old.

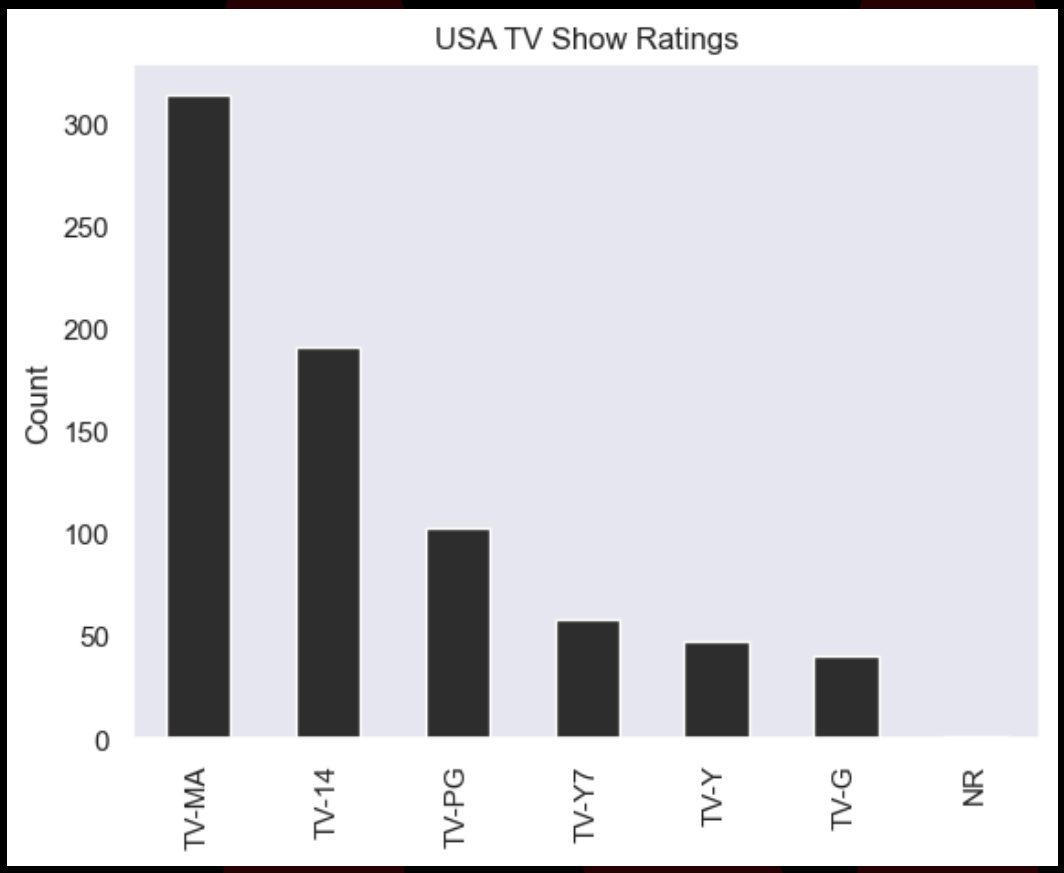
PG and TV-PG more or less are same. This program contains minimal violence, no explicit language, and minimal sexual content or dialogue.

TV-PG: Parental guidance is suggested, as some parents may deem the content unsuitable for their younger children. It is recommended that parents consider watching the program with their children to ensure its appropriateness.

TV-Y7, TV-G, TV-Y, NR, and G are almost the same.

Below G, a few movies are rated under TV-Y7, NC-17, and UR.

USA TV SHOW RATINGS



It can be seen that TV-MA is a little more than 300, TV-14 is just under 200, TV-PG is about 100, TV-Y7 is a little more than 50, TV-Y and TV-G are more or less the same and few TV Shows are rated under NR.

TOP 10 COUNTRIES WITH HIGHEST RATED CONTENT

COUNTRY	RATING
01. UNITED STATES	UR
02. UNITED KINGDOM, FRANCE	UR
03. FRANCE	UR
04. UNITED STATES, ITALY	TV-Y7-FV
05. INDIA	TV-Y7-FV
06. CANADA	TV-Y7-FV

07.	DENMARK, CHINA	TV-Y7-FV
08.	UNITED STATES, UNITED KINGDOM, AUSTRALIA	TV-Y7
09.	FINLAND	TV-Y7
10.	NETHERLANDS, GERMANY, ITALY, CANADA	TV-Y7
11.	UNITED STATES, SOUTH KOREA	TV-Y7

TOP 12 HIGHEST RATED MOVIES

MOVIES

01. YOU DON'T MESS WITH THE ZOHAN

02. SEX DOLL

03. IMMORAL TALES

04. LEGO NINJAGO: MASTERS OF SPINJITZU: DAY OF THE...

05. MOTU PATLU: KING OF KINGS

06. LEO THE LION

RIDDHIMA KAURA

07. LITTLE SINGHAM AUR KAAL KA MAHAJAAL

08. DEAR DRACULA

09. MONSTER HIGH: FRIGHT ON!

10. TOM AND JERRY: THE MAGIC RING

11. MONSTER HIGH: FRIGHTS, CAMERA, ACTION!

12. GARFIELD'S FUN FEST

TOP 12 HIGHEST RATED TV SHOWS

TV SHOWS

01. OH NO! IT'S AN ALIEN INVASION

02. ZOMBIE DUMB

03. POKÉMON THE SERIES

04. LITTLEST PET SHOP: A WORLD OF OUR OWN

05. WE THE PEOPLE

06. SHAUN THE SHEEP: ADVENTURES FROM MOSSY BOTTOM

RIDDHIMA KAURA

07. ZIG & SHARKO

08. H2O: MERMAID ADVENTURES

09. DINO GIRL GAUKO

10. SAILOR MOON CRYSTAL

11. ARCHIBALD'S NEXT BIG THING

12. HORRID HENRY

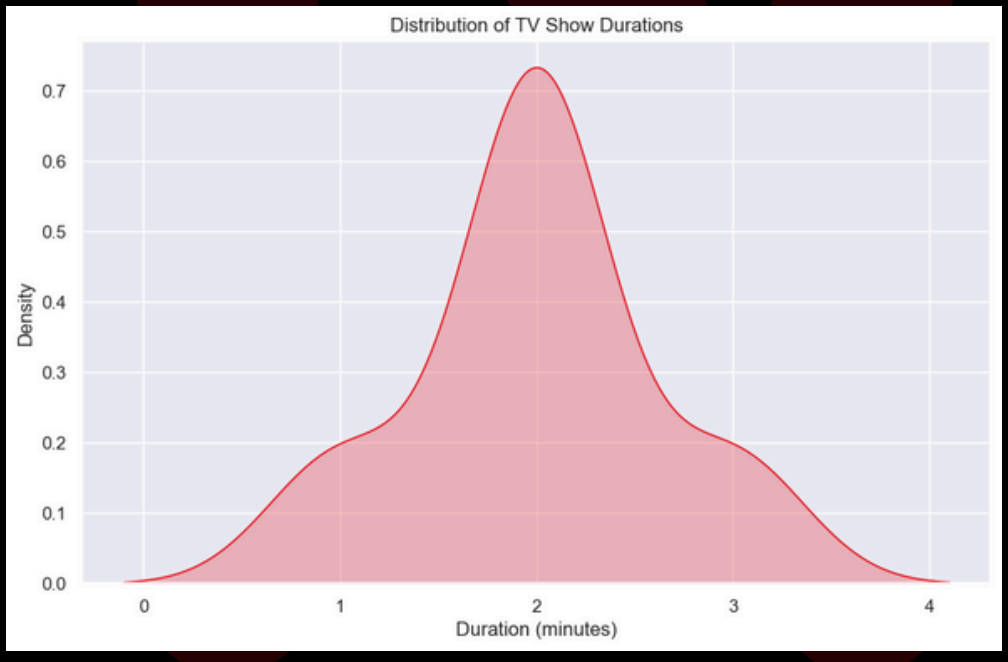
RK
RIDDHIMA KAURA

TOP 12 HIGHEST RATED TV SHOWS DURATION ANALYSIS

AVERAGE DURATION OF TV SHOWS: 2.0

LONGEST TV SHOW DURATION: 3

SHORTEST TV SHOW DURATION: 1



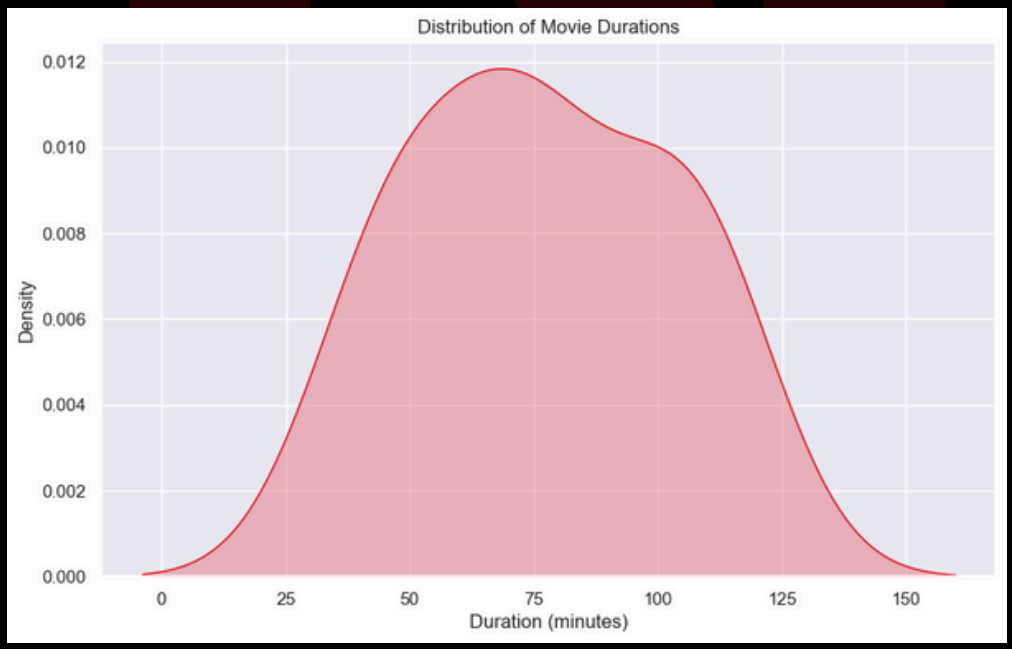
RIDDHIMA KAURA

TOP 12 HIGHEST-RATED MOVIES DURATION ANALYSIS

AVERAGE DURATION OF MOVIES:
76.66666666666667

LONGEST MOVIE DURATION: 113

SHORTEST MOVIE DURATION: 43



RIDDHIMA KAURA

COMPARATIVE ANALYSIS OF MOVIE DURATION BY GENRE, COUNTRY AND RELEASE YEAR

Average Duration of Movies by Genre:

listed_in	
Action & Adventure	104.890625
Action & Adventure, Anime Features	84.000000
Action & Adventure, Anime Features, Children & Family Movies	91.750000
Action & Adventure, Anime Features, Classic Movies	119.500000
Action & Adventure, Anime Features, Horror Movies	96.000000
...	...
Sci-Fi & Fantasy	95.000000
Sci-Fi & Fantasy, Thrillers	107.583333
Sports Movies	87.000000
Stand-Up Comedy	66.913174
Thrillers	99.953846

Average Duration of Movies by Country:

country	
, France, Algeria	103.000000
Argentina	85.789474
Argentina, Brazil, France, Poland, Germany, Denmark	96.000000
Argentina, Chile	95.000000
Argentina, Chile, Peru	100.000000
...	...
Venezuela	119.000000
Venezuela, Colombia	82.000000
Vietnam	106.285714
West Germany	150.000000
Zimbabwe	100.000000

RIDDHIMA KAURA

```
Average Duration of Movies by Release Year:
release_year
1942    35.000000
1943    62.666667
1944    52.000000
1945    51.333333
1946    58.000000
          ...
2017    95.611765
2018    96.185137
2019    93.466035
2020    92.141199
2021    96.444043
```

The Comparative Analysis of Movie Duration by Genre, Country, and Release Year involves examining the duration of movies across different genres, countries, and release years to identify patterns and trends. By analyzing this data, one can gain insights into how movie durations vary based on these factors.

Genre: This analysis would involve comparing the average duration of movies within different genres. It can reveal if certain genres tend to have longer or shorter films, providing an understanding of audience preferences and industry standards.

Country: By comparing movie durations across different countries, one can explore cultural influences on film length. This analysis may uncover whether movies from specific countries tend to be longer or shorter, reflecting storytelling traditions or audience expectations.

Release Year: Analyzing movie durations over various release years can show trends in filmmaking practices. It can indicate if movies have become longer or shorter over time, potentially reflecting changes in audience attention spans, technological advancements, or shifts in storytelling techniques.

Overall, this comparative analysis can offer valuable insights into the relationship between movie duration, genre, country of origin, and release year, providing a comprehensive view of how these factors influence the length of films in the global cinematic landscape.

TRENDS IN TV SHOW GENRES

	release_year	listed_in	count
0	1925	TV Shows	1
1	1945	TV Shows	1
2	1946	TV Shows	1
3	1963	Classic & Cult TV, TV Sci-Fi & Fantasy	1
4	1967	Classic & Cult TV, TV Comedies	1
...	...	...	...
984	2021	TV Dramas, TV Mysteries, TV Sci-Fi & Fantasy	1
985	2021	TV Dramas, TV Sci-Fi & Fantasy	1
986	2021	TV Dramas, Teen TV Shows	1
987	2021	TV Horror, Teen TV Shows	1
988	2021	TV Shows	2

Group the TV shows by genre and year of release, then calculate the number of TV shows in each genre for each year. This analysis can help identify which genres are becoming more popular or less popular over time.

RIDDHIMA KAURA

TOP 10 CAST WITH HIGHEST RATINGS

	cast	rating
3038	Lise Danvers, Fabrice Luchini, Charlotte Alexa...	UR
80	Adam Sandler, John Turturro, Emmanuelle Chriqu...	UR
1832	Hafsia Herzi, Ash Stymest, Karole Rocher, Paul...	UR
1140	Daniel Amerman, John Cygan, Matthew Mercer, Am...	TV-Y7-FV
5159	Unknown_Value	TV-Y7-FV
3681	Nathan Gamble, Ray Liotta, Emilio Estevez, Ari...	TV-Y7-FV
4549	Saurav Chakrabarty, Vinay Pathak	TV-Y7-FV
4765	Sonal Kaushal, Rupa Bhimani, Jigna Bharadhwaj,...	TV-Y7
3360	Matthew Wolf, Rick Gomez, Tara Strong, Alistai...	TV-Y7
4551	Saurav Chakraborty, Ganesh Diweker, Arpita Vor...	TV-Y7
4552	Saurav Chakraborty, Omee, Sankalp, Rajesh, Vin...	TV-Y7
4787	Sourav Chakraborty, Mayur Vyas, Anil Datt	TV-Y7

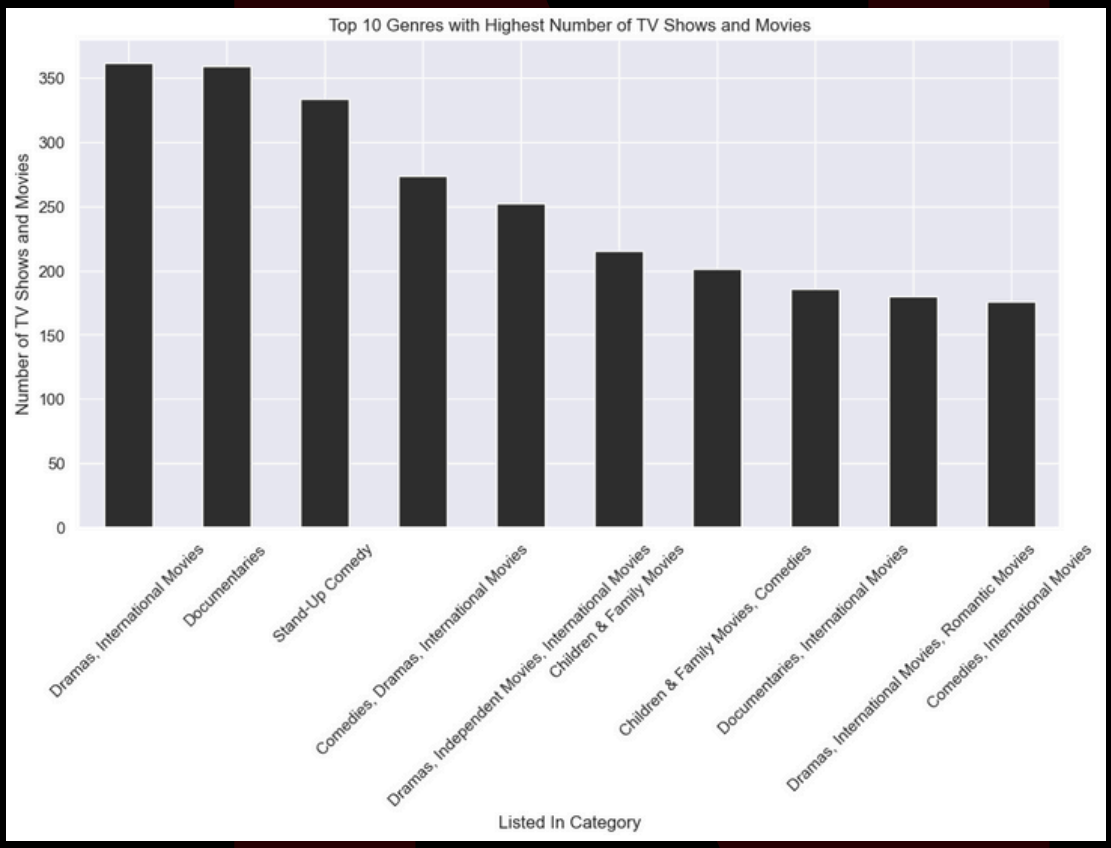
TOP 10 DIRECTORS WITH HIGHEST RATINGS

DENNIS DUGAN	UR
WALERIAN BOROWCZYK	UR
SYLVIE VERHEYDE	UR
CHAD VAN DE KEERE	TV-Y7-FV
PRAKASH SATAM	TV-Y7-FV
MARIO CAMBI	TV-Y7-FV

RIDDHIMA KAURA

SUHAS KADAV	TV-Y7-FV
STEVE BALL, ANDREW DUNCAN	TV-Y7
THOMAS ASTRUC	TV-Y7
TETSUO YAJIMA	TV-Y7
CORY EDWARDS	TV-Y7

TOP 10 GENRES WITH HIGHEST NUMBER OF TV SHOWS AND MOVIES



These genres have the highest number of titles in the data.

RIDDHIMA KAURA

TV SHOWS WITH THE HIGHEST NUMBER OF SEASONS

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
32	s33	TV Show	Sex Education	NaN	Asa Butterfield, Gillian Anderson, Ncuti Gatwa...	United Kingdom	September 17, 2021	2020	TV-MA	3 Seasons	British TV Shows, International TV Shows, TV C...	Insecure Otis has all the answers when it come...
39	s40	TV Show	Chhota Bheem	NaN	Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jig...	India	September 16, 2021	2021	TV-Y7	3 Seasons	Kids' TV	A brave, energetic little boy with superhuman ...
95	s96	TV Show	The Circle	NaN	Michelle Buteau	United States, United Kingdom	September 8, 2021	2021	TV-MA	3 Seasons	Reality TV	Status and strategy collide in this social exp...
124	s125	TV Show	Pororo - The Little Penguin	NaN	NaN	South Korea	September 2, 2021	2013	TV-Y7	3 Seasons	Kids' TV, Korean TV Shows	On a tiny island, Pororo the penguin has fun a...
154	s155	TV Show	Kuroko's Basketball	NaN	Kensho Ono, Yuki Ono, Chiwa Saito, Yoshimasa H...	Japan	September 1, 2021	2015	TV-MA	3 Seasons	Anime Series, International TV Shows, Teen TV ...	Five middle school basketball stars went to se...
...	...	...	...	...	...	...	...	...	...	...	...	...
8599	s8600	TV Show	Toast of London	Michael Cumming	Matt Berry, Robert Bathurst, Doon Mackichan, S...	United Kingdom	September 1, 2017	2015	TV-MA	3 Seasons	British TV Shows, Classic & Cult TV, TV Comedies	After a divorce and fatal career move, a class...
8605	s8606	TV Show	Top Grier	NaN	NaN	United States	December 31, 2018	2018	TV-MA	3 Seasons	Reality TV	Social media star Hayes Grier returns to North...
8684	s8685	TV Show	Vroomiz	NaN	Joon-seok Song, Jeong-hwa Yang,	South Korea	August 1, 2017	2016	TV-Y	3 Seasons	Kids' TV, Korean TV Shows	For these half-car, half-animal

There are total 199 TV Shows with highest number of seasons.

RIDDHIMA KAURA

WERE THERE ANY SPECIFIC DECISIONS MADE BASED ON THE DATA ANALYSIS ON NETFLIX DATASET

Based on the information from the data analysis on the Netflix dataset, specific decisions could have been made based on the insights gained from the analysis. Some potential decisions that could have been made include:

Content Strategy: Based on the analysis of the most popular genres and titles, decisions could be made regarding the acquisition or production of new content to cater to the preferences of the audience.

RIDDHIMA KAURA

Regional Focus: Analysis of the distribution of titles across different countries could inform decisions on expanding content offerings in specific regions or tailoring content to regional preferences.

Rating Considerations: Understanding the distribution of ratings among titles could influence decisions on content categorization, viewer recommendations, or parental controls.

Duration Optimization: Analysis of the duration of titles could guide decisions on the creation of content with optimal lengths to maximize viewer engagement.

Cast and Director Selection: Insights into the impact of cast members or directors on the popularity of titles could influence decisions on casting choices or collaborations with specific directors.

Release Timing: Analysis of release years could inform decisions on the timing of content releases to align with viewer preferences or trends.

Genre Expansion: Identification of underrepresented genres could lead to decisions on diversifying the content library to attract a broader audience.

Marketing Strategies: Insights from the analysis could guide decisions on targeted marketing campaigns based on genre preferences, viewer demographics, or content characteristics.

These decisions could be crucial for content acquisition, production, marketing, and overall content strategy on the Netflix platform, aiming to enhance viewer satisfaction, engagement, and platform performance.

PYTHON CODE

RIDDHIMA KAURA

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

# # DATA ANALYSIS ON NETFLIX DATASET

# 1. Import the Libraries

# In[1]:


import ...
warnings.filterwarnings('ignore')
import plotly.express as px


# 2. Load Dataset

# In[2]:


df = pd.read_csv('netflix_titles.csv')


# In[3]:


df

```

```
# 3. Exploring the Data

# -The dimension or size of the data

# In[4]:


df.shape


# -Displaying first few records

# In[5]:


df.head()


# -Detailed summary about the Dataset

# In[6]:


df.info()


#
# -Checking for unique values
```

```
# In[7]:


df.nunique()


# -Checking for duplicates if any

# In[8]:


df[df.duplicated()]


# -Checking and Handling Missing Values

# In[9]:


df.isnull().sum()


# In[10]:


df['director'].fillna("Unknown_Value",inplace=True)
df['cast'].fillna("Unknown_Value",inplace=True)
df['country'].fillna("Unknown_Value",inplace=True)

```

```
df['country'].fillna("Unknown_Value", inplace=True)


# In[11]:


df.dropna(subset=['date_added'], inplace=True)
df.dropna(subset=['rating'], inplace=True)
df.dropna(subset=['duration'], inplace=True)


# In[12]:


df.isna().sum()


# In[13]:


df['title'].unique()


# In[14]:


df['type'].unique()

```

```
122  df['cast'].unique()
123
124
125  # In[16]:
126
127
128  df['director'].unique()
129
130
131  # In[17]:
132
133
134  df['country'].unique()
135
136
137  # In[18]:
138
139
140  df['release_year'].unique()
141
142
143  # In[19]:
144
145
146  df['date_added'].unique()
147
```

```
df['rating'].unique()


# In[21]:


df['duration'].unique()


# In[22]:


df['description'].unique()


# In[23]:


df['listed_in'].unique()


# 4. Analysing and Visualizing the Data

# In[24]:

```

```
df.head(10)

# # -Assessing the most-occurring and least-occurring type available on Netflix using Countplot

# In[25]:


sns.countplot(x='type',data=df,color='#E50914')
plt. grid(False)
plt.figure(figsize=(10,5))
plt.show()


# -The above countplot depicts that Netflix subsume more movies as compared to TV Shows.

# In[26]:


#providing the occurences of each unique value in a column.
df['type'].value_counts()


# -As is presented by using value_counts movies available on Netflix are 6126 and TV Shows are 2664.
```

```
# # -12 Countries that hit the peak in terms of content

# In[27]:


df.country.value_counts().head(12)


# In[28]:


top_countries=df.country.value_counts().head(12)


# In[29]:


top_countries.plot(kind='bar', x='country', y=top_countries.values, color='#E50914')
plt. grid(False)
plt.ylabel("Count")
plt.title("12 Countries that hit the peak in terms of content")
plt.show()


# The supplied graph gives information on 12 Countries that are on top in terms of content. It could be plainly viewed that t
```

```
# # When was Netflix getting popular?

# In[30]:


df['release_year'].value_counts()


# In[31]:


sns.histplot(df['release_year'],bins=40, color='#E50914')
plt.title("Number of Release Year on Netflix ")
plt.show()


# It is explicitly observed that Netflix had large amount of streaming traffic during the year 2016-20.

# # Most Content Creating Countries

# In[32]:


country=df['country'].value_counts().sort_values(ascending=False)
country=pd.DataFrame(country)
topcountries=country[0:13]
topcountries
```

```
# In[33]:


import plotly.graph_objects as go


# # List of movies from USA on Netflix hollywood

# In[34]:


Movies_from_USA= df[(df['type']=='Movie')&(df['country']=='United States')]['title']
Movies_from_USA=pd.DataFrame(Movies_from_USA)
Movies_from_USA
fig = go.Figure(data=[go.Table(header=dict(values=['Title'],fill_color='#E50914', font_color='white', align='left', font_size=16),
                 cells=dict(values=[Movies_from_USA['title']],fill_color='black',align='left',font_color='white'))
                     ])
fig.show()
```

```
# # List of TV Shows from USA on Netflix hollywood

# In[35]:


TV_shows_from_USA= df[(df['type']=='TV Show')&(df['country']=='United States')]['title']
TV_shows_from_USA=pd.DataFrame(TV_shows_from_USA)
TV_shows_from_USA
fig = go.Figure(data=[go.Table(header=dict(values=['Title'],fill_color='#E50914', font_color='white', align='left', font_size=16),
                 cells=dict(values=[TV_shows_from_USA['title']],fill_color='black',align='left',font_color='white'))
                     ])
fig.show()
```

```
# # In which year Netflix was getting popular with the most number of releases in

# In[36]:


us_data=df[df['country']=='United States']
us_data_yearwise=us_data.sort_values(by='release_year')[0:8807]
fig = go.Figure(data=[go.Table(header=dict(values=['Title', 'Release Year'],fill_color='#E50914', font_color='white', align='left',
                 cells=dict(values=[us_data_yearwise['title'],us_data_yearwise['release_year']],fill_color='black',align='left',f
                     ])

fig.show()
```

DATA ANALYSIS ON NETFLIX DATASET.ipynb

```
# In[37]:


plt.figure(figsize=(18,20))
sns.set(style="darkgrid")
plt.title("Most number of releases in USA?")
ax = sns.countplot(y="release_year", data=us_data_yearwise, palette="gist_gray", order=us_data_yearwise['release_year'].value_count

# The countplot presents the Netflix data which shows the number of releases over 5 decades, commencing from 1969. As is observed,

# In[38]:


df

# In[39]:


df.count

# In[40]:
```

```
# In[40]:


import plotly.graph_objects as go
us_data=df[df['country']=='United States']
us_data_yearwise=us_data.sort_values(by='release_year')[0:8807]
us_data_datewise=us_data.sort_values(by='date_added')[0:8807]

fig = go.Figure(data=[go.Table(header=dict(values=['Title','date_added', 'Release Year'],fill_color='#E50914', font_color='white',
                 cells=dict(values=[us_data_yearwise['title'],us_data_datewise['date_added'],us_data_yearwise['release_year']],fill
                     ])

fig.show()
```

```
# # In which month of the year commencing from 2008 Netflix released new TV shows and Movies the most in USA?
#

# In[41]:


us_data


# In[42]:


us_data.count()
netflix_series=us_data[us_data['type']=='TV Show']

date = netflix_series[['date_added']].dropna()
date['year'] = date['date_added'].apply(lambda x_: x.split(', ')[-1])
date['month'] = date['date_added'].apply(lambda x_: x.lstrip().split(' ')[0])

month_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'Dece
us_dataaa = date.groupby('year')['month'].value_counts().unstack().fillna(0)[month_order].T
plt.figure(figsize=(7, 5), dpi=100)
plt.pcolor(us_dataaa, cmap='gist_gray', edgecolors='white', linewidths=3)
plt.xticks(np.arange(0.3, len(us_dataaa.columns), 1),us_dataaa.columns, fontsize=7, fontfamily='Monospace')
plt.yticks(np.arange(0.3, len(us_dataaa.index), 1),us_dataaa.index, fontsize=7, fontfamily='Monospace')

plt.title('Netflix TV shows Release Month', fontsize=15, fontweight='bold', position=(0.50, 1.0+0.02))
cbar = plt.colorbar()
```

```
plt.yticks(np.arange(0.3, len(us_dataaa.index), 1),us_dataaa.index, fontsize=7, fontfamily='Monospace')

plt.title('Netflix TV shows Release Month', fontsize=15, fontweight='bold', position=(0.50, 1.0+0.02))
cbar = plt.colorbar()

cbar.ax.tick_params(labelsize=7)
cbar.ax.minorticks_on()

plt.show()


# Potentially, a new Netflix TV Show could be released any month of the year. Regardless of how the presented heatmap shows which m
#
# As is obvious, the number of releases in the four mentioned years i.e., 2008, 2013, 2014, and 2015 shows almost the same pattern
#
# In the year 2016, it is obvious that most of the releases were in January, and in the year 2017, scarcely any release was in sigh
#
# During 2018 and 2019, Netflix started getting popular with many releases in every month of the year mainly in the first half of t
#
# It is interesting to note that most of the releases can be seen in the year 2020 attributable to COVID-19. while in the year 2021
#
# To put it simply, the trend of releasing content on Netflix mainly started between 2017 and 2018 and became prominent in 2020.
#

# In[43]:


us_data.count()
netflix_movies=us_data[us_data['type']=='Movie']
```

```
us_data.count()
netflix_movies=us_data[us_data['type']=='Movie']

date = netflix_movies[['date_added']].dropna()
date['year'] = date['date_added'].apply(lambda x_: x.split(', ')[-1])
date['month'] = date['date_added'].apply(lambda x_: x.lstrip().split(' ')[0])

month_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'Dece
us_dataa = date.groupby('year')['month'].value_counts().unstack().fillna(0)[month_order].T
plt.figure(figsize=(7, 5), dpi=100)
plt.pcolor(us_dataa, cmap='gist_gray', edgecolors='white', linewidths=3)
plt.xticks(np.arange(0.3, len(us_dataa.columns), 1), us_dataa.columns, fontsize=7, fontfamily='Monospace')
plt.yticks(np.arange(0.3, len(us_dataa.index), 1), us_dataa.index, fontsize=7, fontfamily='Monospace')

plt.title('Netflix Movies Release Month', fontsize=15, fontweight='bold', position=(0.50, 1.0+0.02))
cbar = plt.colorbar()

cbar.ax.tick_params(labelsize=7)
cbar.ax.minorticks_on()

plt.show()


# It could be viewed that between 2008 and 2016 shows the same pattern and barely any release in sight. Starting from December 201

# # USA Movie Ratings

# In[44]:
```

```
# # USA Movie Ratings

# In[44]:


usa_movie_rating=netflix_movies.rating.value_counts()


# In[45]:


usa_movie_rating.plot(kind='bar', x='rating', y=usa_movie_rating.values, color='#2F2F2F')
plt. grid(False)
plt.ylabel("Count")
plt.title("USA Movie Ratings")
plt.show()


# As is observed, TV-MA is exactly 600. Generally speaking, TV-MA is a rating meant for mature audiences, designed for adults, and
#
# Just under TV-MA, rating R is roughly 400. The film with a rating of R is suitable for viewers aged 17 and above without the need
#
# PG-13 is a little less than 300. Parents are advised to exercise caution due to some content that may not be suitable for childre
#
# TV-14 is nearly around 200. Parents are strongly cautioned as this program includes content that may not be suitable for childre
#
# PG and TV-PG more or less are same. This program contains minimal violence, no explicit language, and minimal sexual content or
#
```

```
# TV-Y7, TV-G, TV-Y, NR, and G are almost the same.
#
# Below G, a few movies are rated under TV-Y7, NC-17, and UR.

# # USA TV Show Ratings

# In[46]:


usa_TVShow_rating=netflix_series.rating.value_counts()


# In[47]:


usa_TVShow_rating.plot(kind='bar', x='rating', y=usa_TVShow_rating.values, color='#2F2F2F')
plt. grid(False)
plt.ylabel("Count")
plt.title("USA TV Show Ratings")
plt.show()


# It can be seen that TV-MA is a little more than 300, TV-14 is just under 200, TV-PG is about 100, TV-Y7 is a little more than 50

# In[48]:
```

```
# It can be seen that TV-MA is a little more than 300, TV-14 is just under 200, TV-PG is about 100, TV-Y7 is a

# In[48]:


df


# # Countries with highest rated content

# In[49]:


top_country_ratings= df.groupby("country")["rating"].max().reset_index()
top_country_ratings=top_country_ratings.sort_values(by='rating', ascending=False).head(12)


# In[50]:


top_country_ratings
```

```
489
490 top_country_ratings
491
492
493 # # Top 12 Highest Rated Movies
494
495 # In[51]:
496
497
498 movies_df = df[df['type']=='Movie']
499 sorted_df = movies_df.sort_values(by='rating', ascending=False)
500 top_10_movies = sorted_df.head(12)
501
502
503 # In[52]:
504
505
506 top_10_movies
507
```

```
509 # # Top 12 Highest Rated TV Shows
510
511 # In[53]:
512
513
514 tvshows_df = df[df['type']=='TV Show']
515 sorted_dftv = tvshows_df.sort_values(by='rating', ascending=False)
516 top_10_tvshows = sorted_dftv.head(12)
517
518
519 # In[54]:
520
521
522 top_10_tvshows
523
```

```
# In[55]:


# Extract TV show durations
tv_show_durations = top_10_tvshows[top_10_tvshows['type'] == 'TV Show']['duration']

# Convert duration to numeric values
def convert_duration(duration):
    if 'Season' in duration:
        return int(duration.split()[0])
    elif 'min' in duration:
        return int(duration.split()[0])
    else:
        return None

tv_show_durations = tv_show_durations.apply(convert_duration)

# Analyze the data
average_duration = tv_show_durations.mean()
longest_show = tv_show_durations.max()
shortest_show = tv_show_durations.min()

print("Average Duration of TV Shows:", average_duration)
print("Longest TV Show Duration:", longest_show)
print("Shortest TV Show Duration:", shortest_show)


# In[56]:
```

```
# Plot the distribution of movie durations using a KDE plot
plt.figure(figsize=(10, 6))
sns.kdeplot(data=tv_show_durations, color='#E50914', fill=True)
plt.xlabel('Duration (minutes)')
plt.ylabel('Density')
plt.title('Distribution of TV Show Durations')
plt.show()


# # Top 12 Highest Rated Movies Duration Analysis

# In[57]:


```

```python
# Extract movie durations
movie_durations = top_10_movies[top_10_movies['type'] == 'Movie']['duration']

# Convert duration to numeric values
def convert_duration(duration):
    if 'min' in duration:
        return int(duration.split()[0])
    elif 'h' in duration:
        return int(duration.split()[0]) * 60
    else:
        return None

movie_durations = movie_durations.apply(convert_duration)

# Analyze the data
average_duration = movie_durations.mean()
longest_movie = movie_durations.max()
shortest_movie = movie_durations.min()

print("Average Duration of Movies:", average_duration)
print("Longest Movie Duration:", longest_movie)
print("Shortest Movie Duration:", shortest_movie)
```

```python
# In[58]:


# Plot the distribution of movie durations using a KDE plot
plt.figure(figsize=(10, 6))
sns.kdeplot(data=movie_durations, color='#E50914', fill=True)
plt.xlabel('Duration (minutes)')
plt.ylabel('Density')
plt.title('Distribution of Movie Durations')
plt.show()
```

```
# # Comparative Analysis of Movie Duration by Genre, Country and Release Year

# In[59]:


# Cleaning and preprocessing the data
df = df[df['duration'].str.contains('min')]  # Keep only rows with 'min'
df['duration'] = df['duration'].str.replace(' min', '').astype(int)  # Remove 'min' and convert to int

# Compare movie durations by genre
genre_durations = df[df['type'] == 'Movie'].groupby('listed_in')['duration'].mean()
print("Average Duration of Movies by Genre:")
print(genre_durations)

# Compare movie durations by country
country_durations = df[df['type'] == 'Movie'].groupby('country')['duration'].mean()
print("\nAverage Duration of Movies by Country:")
print(country_durations)

# Compare movie durations by release year
year_durations = df[df['type'] == 'Movie'].groupby('release_year')['duration'].mean()
print("\nAverage Duration of Movies by Release Year:")
print(year_durations)
```

```
# # Time Series Analysis of Movies by Release Year

# In[60]:


# Convert 'duration' column to string type
df['duration'] = df['duration'].astype(str)

# Analyzing trends in movie durations over time
time_series = df[df['type'] == 'Movie'].groupby('release_year')['duration'].mean()
time_series
```

```
# # Trends in TV Show Genres: Group the TV shows by genre and year of release, then calculate the number of TV sh

# In[61]:


dfs = pd.read_csv('netflix_titles.csv')


# In[62]:


dfs


# In[63]:


# Filter only TV Shows
tv_showsf = dfs[dfs['type'] == 'TV Show']

# Group TV shows by genre and year of release
genre_year_grouped = tv_showsf.groupby(['release_year', 'listed_in']).size().reset_index(name='count')
```

```
# Display the resulting table
genre_year_grouped


# In[64]:


# Extract the list of actors and actresses from the 'cast' column
cast_list = df['cast'].tolist()

# Split the list of actors and actresses into individual names
actors = []
for cast in cast_list:
    actors += cast.split(',')

# Remove any empty strings from the list of actors
actors = [actor.strip() for actor in actors if actor.strip()]

# Count the number of TV shows or movies each actor has appeared in
actor_counts = pd.Series(actors).value_counts()

# Display the resulting Series
actor_counts
```

```


# # Top 10 Cast with Highest Ratings

# In[65]:


cast_ratings= df.groupby("cast")["rating"].max().reset_index()
cast_ratings=cast_ratings.sort_values(by='rating', ascending=False).head(12)

cast_ratings


# # Top 10 directors with Highest Ratings

# In[66]:


director_ratings= df.groupby("director")["rating"].max().reset_index()
director_ratings=director_ratings.sort_values(by='rating', ascending=False).head(12)

director_ratings



```

```
# # Top 10 Genres with Highest Number of TV Shows and Movies
#

# In[67]:


# Filter the dataset to include only TV shows and movies
df = df[df["type"].isin(["TV Show", "Movie"])]

# Group the data by the listed_in column and count the number of TV shows and movies
listed_in_counts = df["listed_in"].value_counts().nlargest(10)

# Plot the top 10 highest listed_in categories
plt.figure(figsize=(12, 6))
listed_in_counts.plot(kind="bar", color="#2F2F2F")
plt.xlabel("Listed In Category")
plt.ylabel("Number of TV Shows and Movies")
plt.title("Top 10 Genres with Highest Number of TV Shows and Movies")
plt.xticks(rotation=45)
plt.show()



# # TV Shows with Highest Number of Seasons

# In[68]:


dfs


# In[69]:


dfs.duration.value_counts()


# In[70]:


dfs[dfs.duration == '3 Seasons']


# In[71]:


```