1.Defining Problem Statement

Problem Statement :

The Netflix dataset poses challenges related to understanding and leveraging the vast amount of data available from Netflix's streaming platform. The dataset comprises various types of information, such as Director,Cast, Ratings, Genre(Listed_in), Movies and TV shows. The goal is to extract meaningful insights for analyzing the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries.

Importing libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading Dataset

In [2]:
```
df = pd.read_csv(r"F:\data_set\netflix_dataset.csv")
```

In [3]:
```
df.head(5)
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	9
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

1. Defining Problem Statement and Analysing basic metrics

- Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to --'category',missing value detection, statistical summary

Exploratory Data Analysis

- Exploratory data analysis: Why It's Important ?

- For solving any business case,there are no short cuts. After collecting the data, we cannot immediately go on to the Visualisations or model-building phase. The exploratory data analytics (EDA) stage is crucial in helping us outline our strategy in an organised way. Exploratory data analysis is the process of examining data, coming up with new ideas, and spotting hidden patterns.

Non-Graphical Analysis: Value Counts and Unique attributes

- So based on the result we can also confirm that Netflix dataset have 12 columns.Now we will dig more deeper to know about the dataset and values

In [4]:
```
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
```

1.Shape of the Data

In [5]:
```
df.shape
```

Out[5]:
```
(8807, 12)
```

2.Rename the "listed_in" into genre

In [6]:
```
df.rename(columns={"listed_in": "genre"},inplace=True)
df.head(5)
```

Out[6]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	9
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

3. Removing "S" from show_id

In [7]:

```
df["show_id"]=df["show_id"].apply(lambda x: x.replace("s",""))
df.head(5)
```

Out[7]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	9
1	2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 S
4	5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

4. Describing the String values

In [8]:
```
df.describe(include="object").T
```

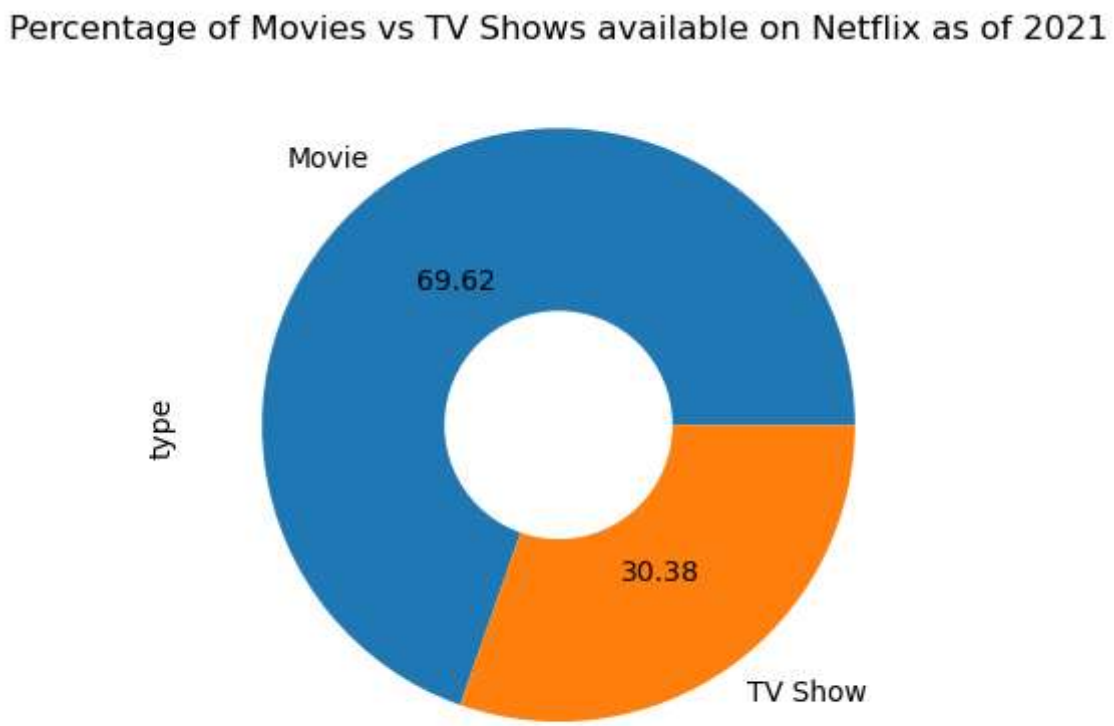
Out[8]:

	count	unique	top	freq
show_id	8807	8807	1	1
type	8807	2	Movie	6131
title	8807	8807	Dick Johnson Is Dead	1
director	6173	4528	Rajiv Chilaka	19
cast	7982	7692	David Attenborough	19
country	7976	748	United States	2818
date_added	8797	1767	January 1, 2020	109
rating	8803	17	TV-MA	3207
duration	8804	220	1 Season	1793
genre	8807	514	Dramas, International Movies	362
description	8807	8775	Paranormal activity at a lush, abandoned prope...	4

Movies vs TV Shows

In [9]:

```
plt.title("Percentage of Movies vs TV Shows available on Netflix as of 2021")
df['type'].value_counts().plot(kind = "pie", autopct="%.2f")
circle = plt.Circle((0, 0), 0.38, color='white')
plt.gcf().gca().add_artist(circle)
plt.show()
```



- Nearly 2/3rd of the content on netflix are movies and remaining 1/3rd of them are TV Show.

Before Un-nesting the "Director" column

In [10]:
```
df["director"].value_counts().head(10)
```

Out[10]:
```
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Marcus Raboy              16
Suhas Kadav               16
Jay Karas                 14
Cathy Garcia-Molina       13
Martin Scorsese           12
Youssef Chahine           12
Jay Chapman               12
Steven Spielberg          11
Name: director, dtype: int64
```

After un-nesting the "Director" columns

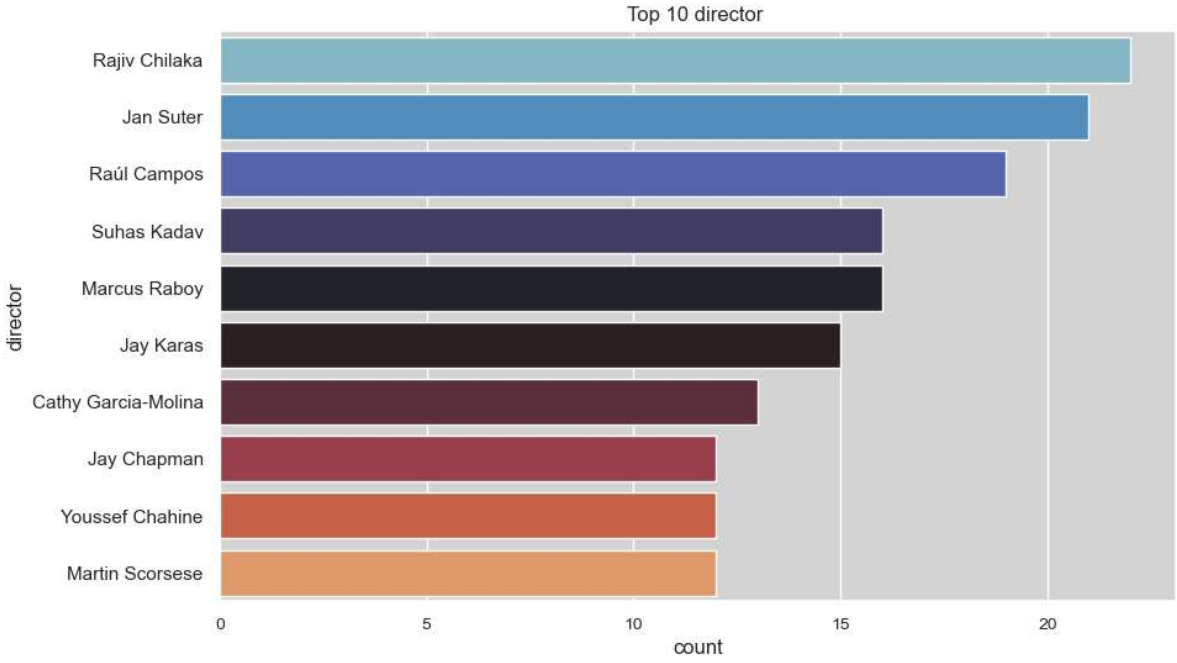
In [11]:
```
df_dir=df[["title","director"]]
df_dir["director"]=df_dir["director"].str.split(", ")
df_dir= df_dir.explode("director")
```

Top 10 directors

In [12]:
```
df_dir["director"].value_counts().head(10)
```

Out[12]:
```
Rajiv Chilaka          22
Jan Suter              21
Raúl Campos            19
Suhas Kadav            16
Marcus Raboy           16
Jay Karas              15
Cathy Garcia-Molina    13
Jay Chapman            12
Youssef Chahine        12
Martin Scorsese        12
Name: director, dtype: int64
```

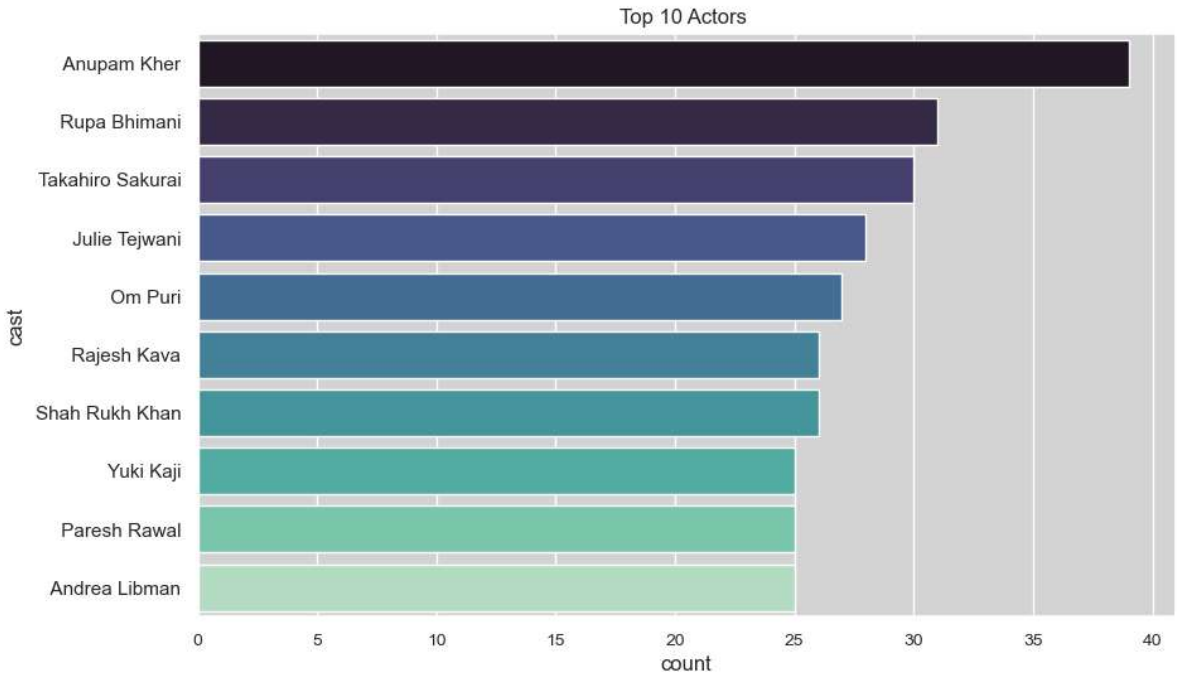
In [13]:
```
sns.set_theme(style="whitegrid")
sns.set(rc={"axes.facecolor":"lightgrey"})
plt.figure(figsize= (10,6))
sns.countplot(palette="icefire",y=df_dir["director"],order=pd.value_counts(df_dir[
plt.xticks( fontsize=10)
plt.title("Top 10 director ")
plt.show()
```



Top 10 Actors

In [14]:
```
df_actor=df[["title","cast"]]
df_actor["cast"]=df_actor["cast"].str.split(",")
df_actor= df_actor.explode("cast")
```

In [15]:
```
sns.set_theme(style="whitegrid")
sns.set(rc={"axes.facecolor":"lightgrey"})
plt.figure(figsize= (10,6))
sns.countplot(palette="mako",y=df_actor["cast"],order=pd.value_counts(df_actor["cas
plt.xticks( fontsize=10)
plt.title("Top 10 Actors ")
plt.show()
```



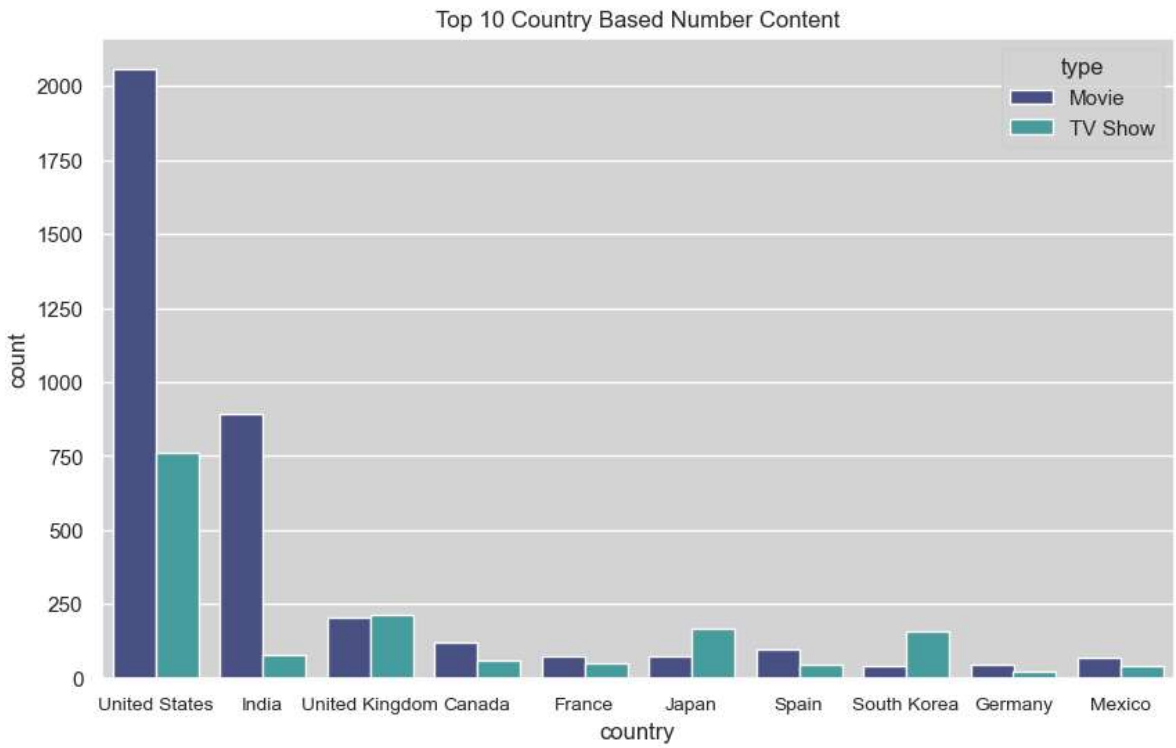
Top 10 Countries Content

In [16]:
```
df_country=df[["country"]]
df_country["country"]=df_country["country"].str.split(", ")
df_country= df_country.explode("country")
```

In [17]:
```
df_country["country"].value_counts().head(10)
```

Out[17]:
```
United States     3689
India             1046
United Kingdom     804
Canada             445
France             393
Japan              318
Spain              232
South Korea        231
Germany            226
Mexico             169
Name: country, dtype: int64
```

In [18]:
```
sns.set_theme(style="whitegrid")
sns.set(rc={"axes.facecolor":"lightgrey"})
plt.figure(figsize= (10,6))
sns.countplot(palette="mako",x=df["country"],hue=df["type"],order=pd.value_counts(
plt.xticks( fontsize=10)
plt.title("Top 10 Country Based Number Content")
plt.show()
```



- The above top 10 countries contribute 85.76% of the total content.
- United States, India, United Kingdom, Canada and France contribute 72.41% of the total content.
- United States alone contributes to 41.9% of the total content on Netflix.

Percentage Contribution Top 10 Countries to the total content of Netfix

In [19]:
```
total_count = df["title"].value_counts().sum()
top_10 = df_country["country"].value_counts().head(10).sum()
print(round((top_10/total_count)*100,2),"%")
```

```
85.76 %
```

Percentage Contribution Top 5 Countries to the total content of Netfix

In [20]:
```
total_count = df["title"].value_counts().sum()
top_10 = df_country["country"].value_counts().head(5).sum()
print(round((top_10/total_count)*100,2),"%")
```

```
72.41 %
```

Percentage Contribution Top 1 Countries to the total content of Netfix

In [21]:
```
total_count = df["title"].value_counts().sum()
top_10 = df_country["country"].value_counts().head(1).sum()
print(round((top_10/total_count)*100,2),"%")
```

```
41.89 %
```

- Nearly 2/3rd of the content on netflix are movies and remaining 1/3rd of them are TV Show. We can see that US, India, United Kingdom, Canada and France contribute 72.4% of the top 5 countries.

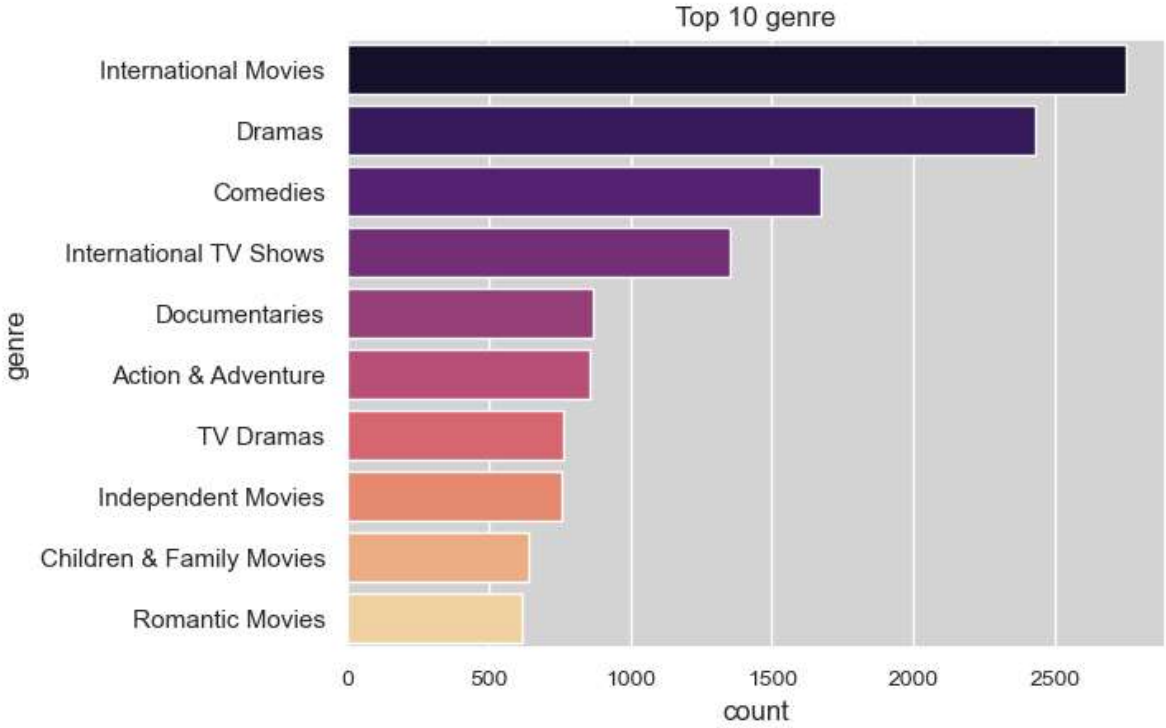
Top 10 Genres

In [22]:
```
df_genre=df[["genre"]]
df_genre["genre"]=df_genre["genre"].str.split(", ")
df_genre= df_genre.explode("genre")
```

In [23]:
```
df_genre["genre"].value_counts().head(10)
```

Out[23]:
```
International Movies        2752
Dramas                      2427
Comedies                    1674
International TV Shows      1351
Documentaries                869
Action & Adventure           859
TV Dramas                    763
Independent Movies           756
Children & Family Movies     641
Romantic Movies              616
Name: genre, dtype: int64
```

In [24]:
```
sns.set_theme(style="whitegrid")
sns.set(rc={"axes.facecolor":"lightgrey"})
sns.countplot(palette="magma",y=df_genre["genre"],order=pd.value_counts(df_genre["
plt.xticks( fontsize=10)
plt.title("Top 10 genre")
plt.show()
```



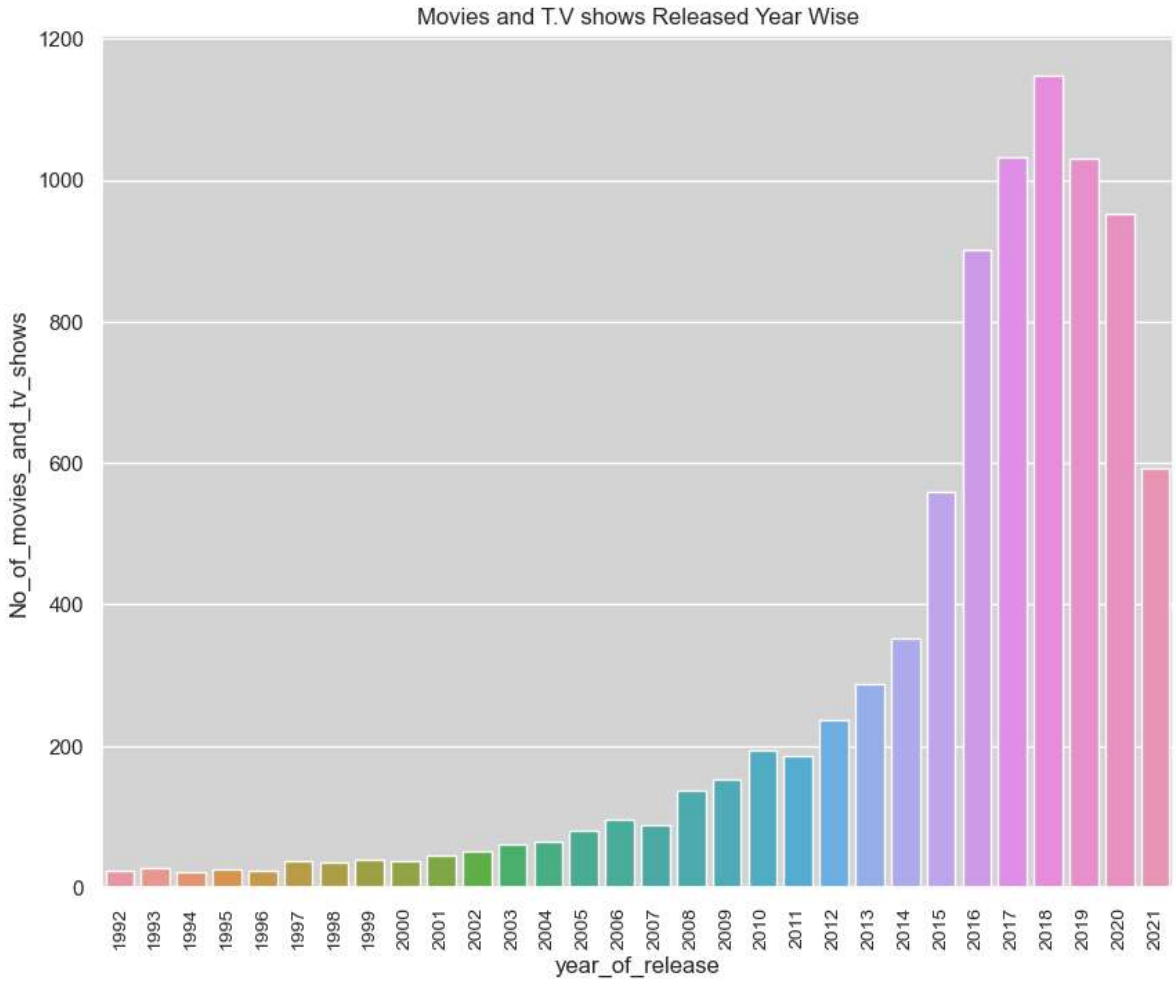
Content Released Year wise

In [25]:
```
df["release_year"].value_counts().head(10)
```

Out[25]:
```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
2021     592
2015     560
2014     352
2013     288
2012     237
Name: release_year, dtype: int64
```

In [26]:
```
year_wise =pd.value_counts(df["release_year"]).reset_index().head(30)
year_wise.rename(columns={"index":"year_of_release","release_year":"No_of_movies_a
```

In [27]:
```
sns.set_theme(style="whitegrid")
sns.set(rc={"axes.facecolor":"lightgrey"})
plt.figure(figsize=(10,8))
sns.barplot(data=year_wise,x="year_of_release",y="No_of_movies_and_tv_shows")
plt.xticks(rotation=90 ,fontsize=10)
plt.title("Movies and T.V shows Released Year Wise")
plt.show()
```



Movies and TV shows Released Year Wise

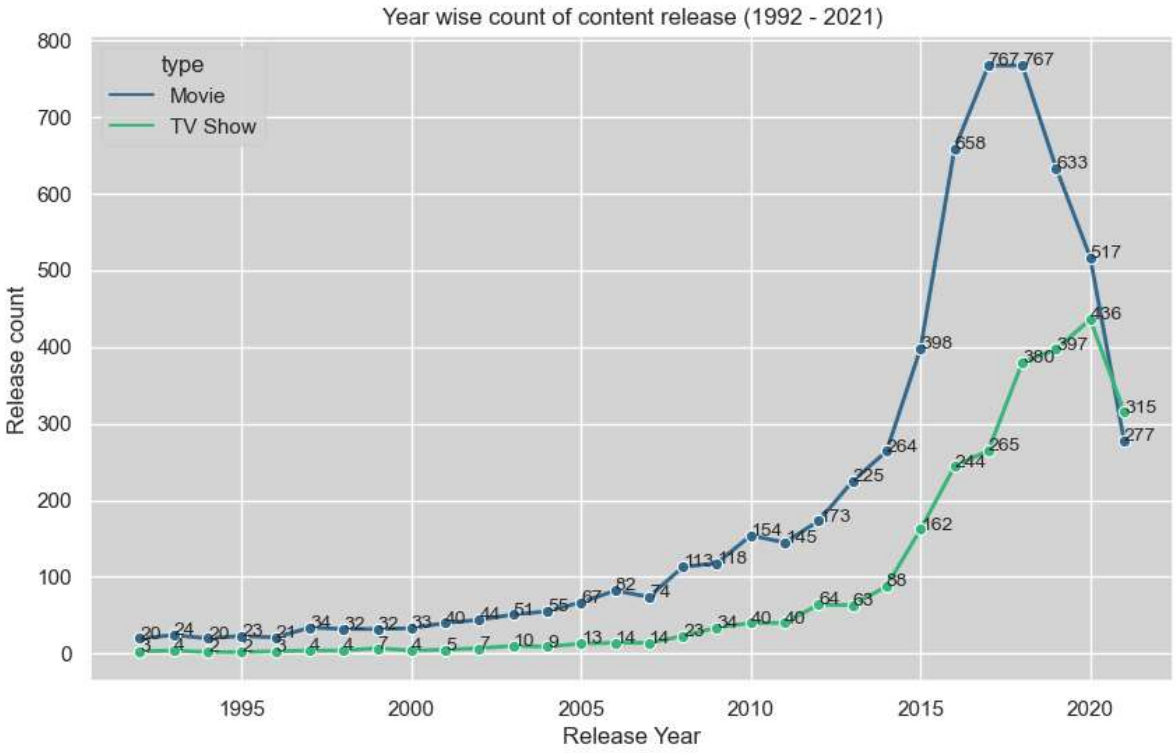
In [28]:
```
year_wise = pd.value_counts(df["release_year"]).reset_index().head(30)
year_wise.rename(columns={"index":"Year of release","release_year":"No. of titles"

df_mov_tv_rel = df[['type', 'release_year']]

df_mov_tv_rel = df_mov_tv_rel.value_counts().reset_index()
df_mov_tv_rel.rename(columns  = {0: 'count'}, inplace = True)

df_mov_tv_rel = df_mov_tv_rel[df_mov_tv_rel["release_year"]>=1992]

plt.figure(figsize = (10, 6))
a = sns.lineplot(data=df_mov_tv_rel, x="release_year", y="count", legend='auto', pa
for i, j in zip(df_mov_tv_rel["release_year"], df_mov_tv_rel["count"]):
    a.text(i, j, str(j), fontsize = 10)
plt.title("Year wise count of content release (1992 - 2021)")
a.set_ylabel('Release count')
a.set_xlabel('Release Year')
plt.show()
```



- There is a Rapid Growth of Movies/Tv shows from the Year 2014 to 2018 .
- There is an Rapid Decline of Movies/Tv shows from the Year 2019 to 2021 .

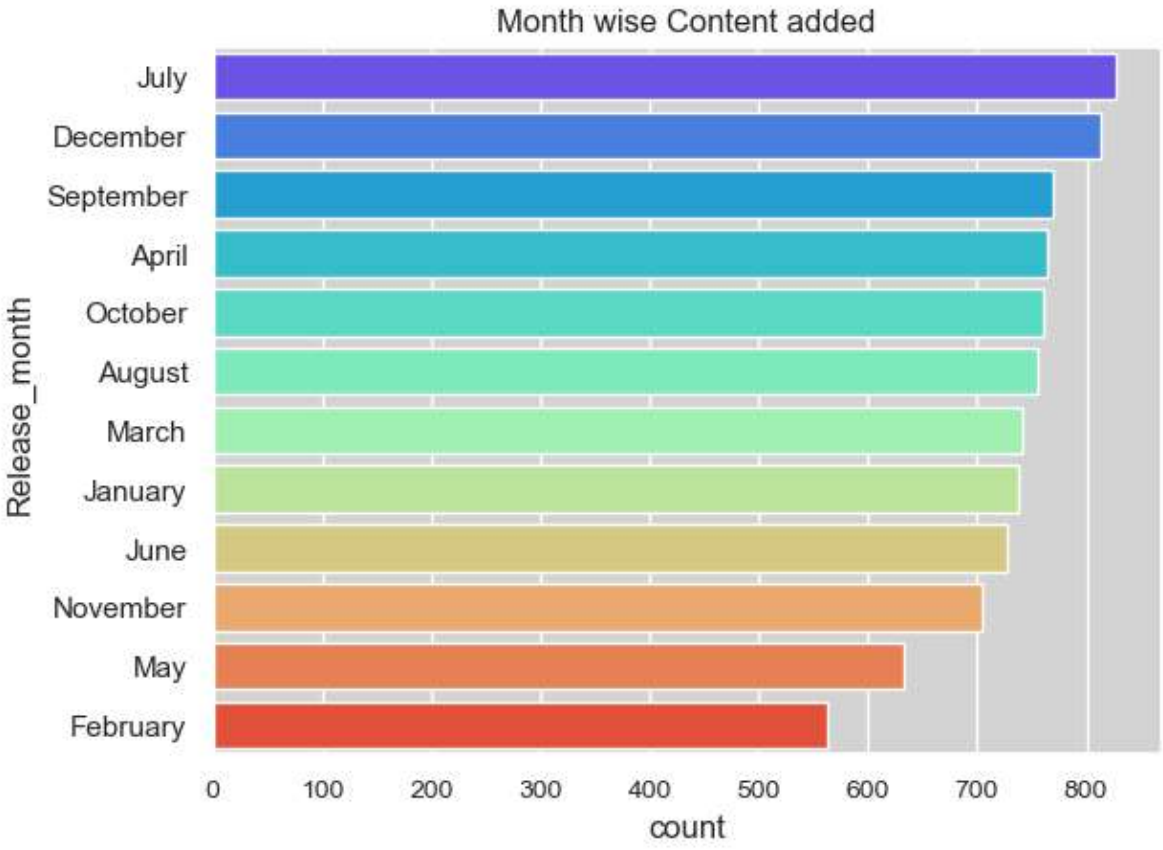
Content Added To Netflix Month Wise

In [29]:
```
Month = df[["date_added"]].replace(np.nan,"No_date")
Month["Release_month"] = Month["date_added"].apply(lambda x:x.lstrip().split(" ")[(
```

In [30]:
```
Month["Release_month"].value_counts().head(12)
```

Out[30]:
```
July         827
December     813
September    770
April        764
October      760
August       755
March        742
January      738
June         728
November     705
May          632
February     563
Name: Release_month, dtype: int64
```

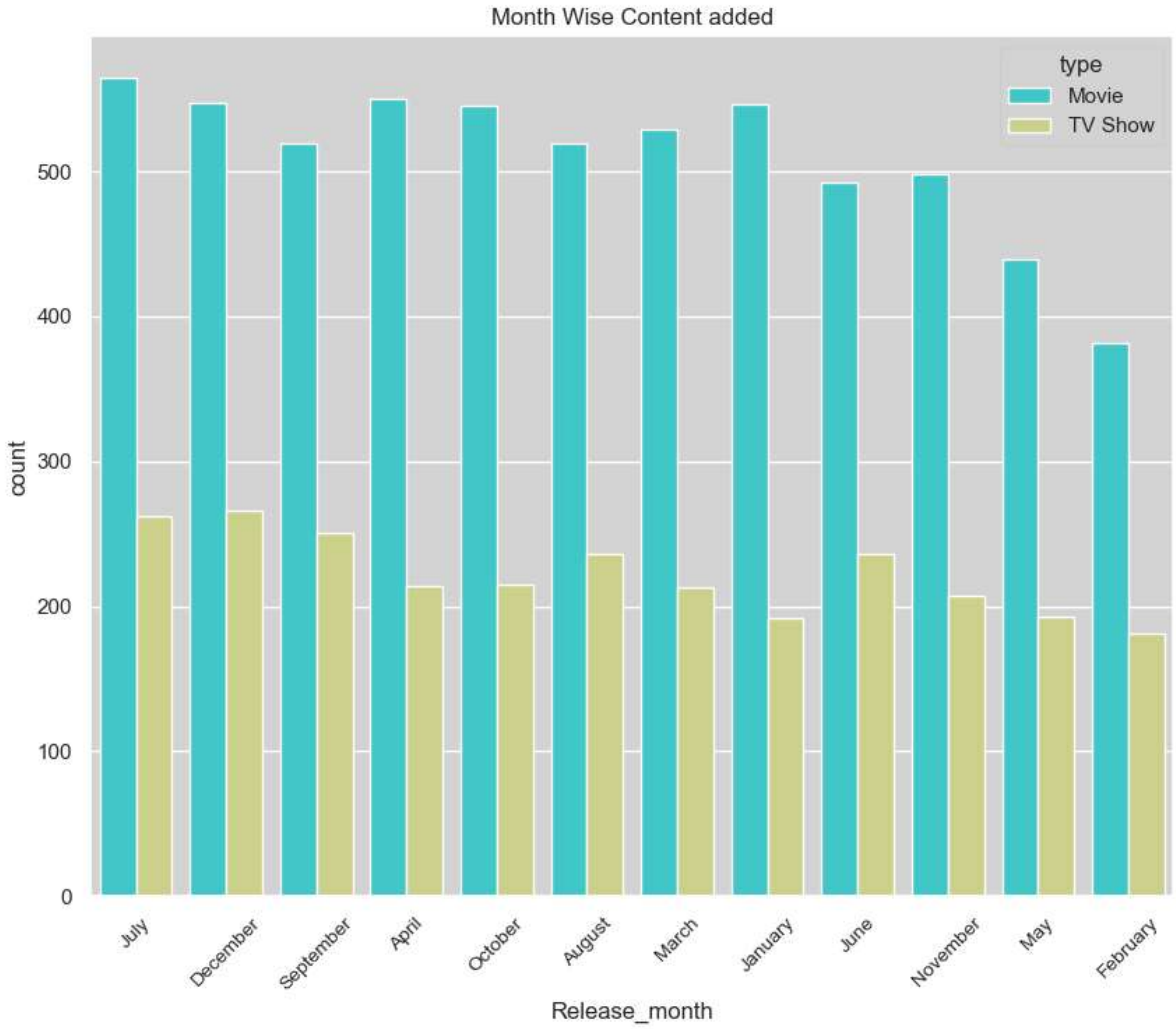
In [31]:
```
sns.set_theme(style="whitegrid")
sns.set(rc={"axes.facecolor":"lightgrey"})
sns.countplot(palette="rainbow",y= Month["Release_month"],order=pd.value_counts(Mo
plt.xticks(fontsize=10)
plt.title("Month wise Content added")
plt.show()
```



- The growth in contents are higher in the month of July ,December, september
- Least number of contents are added in the month of February.

Month Wise Movies and TV Show added to Netflix

In [32]:
```
sns.set_theme(style="whitegrid")
plt.figure(figsize=(10,8))
sns.set(rc={"axes.facecolor":"lightgrey"})
sns.countplot(palette="rainbow",x= Month["Release_month"],hue=df["type"],order=pd.
plt.xticks(rotation = 45,fontsize=10)
plt.title("Month Wise Content added")
plt.show()
```



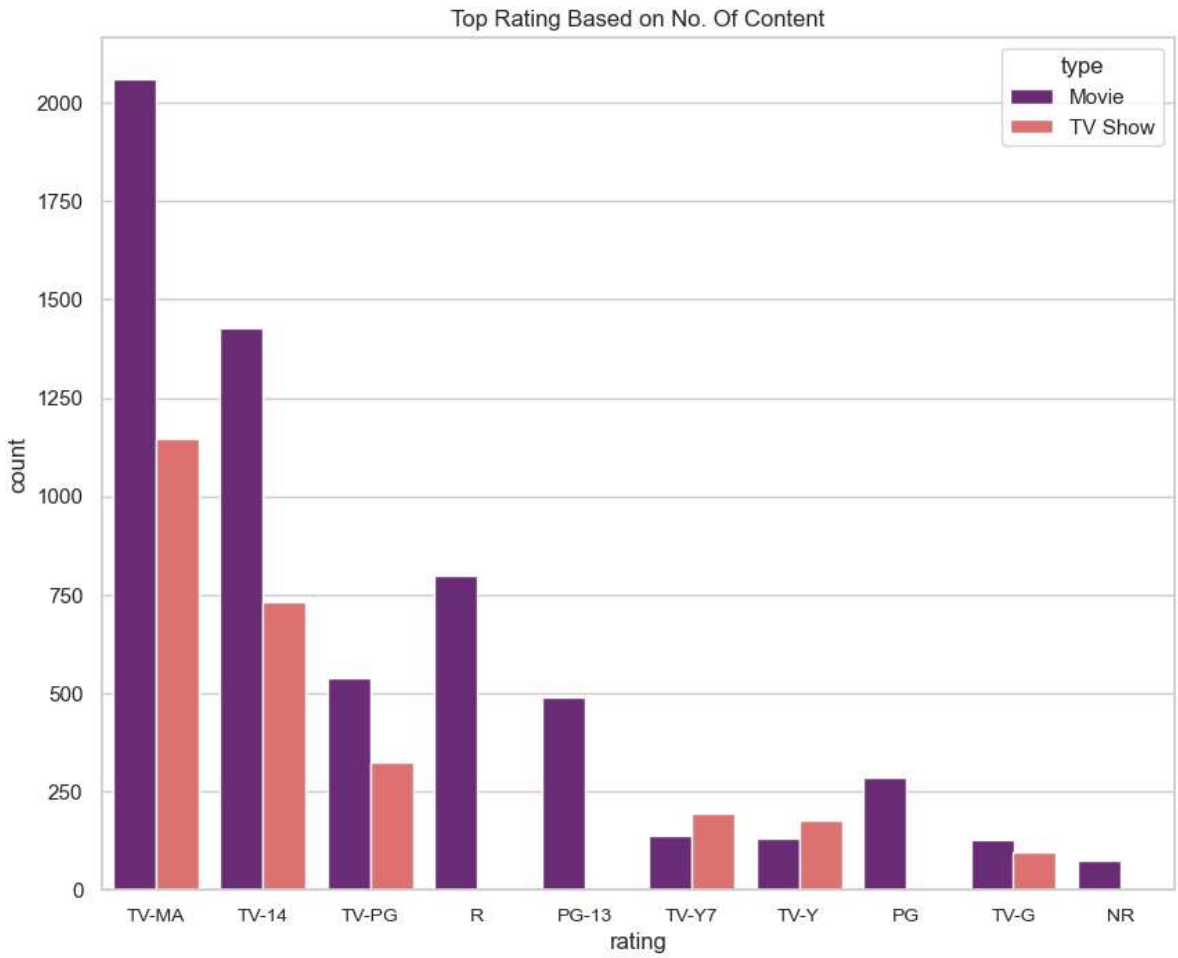
- Most movies are Added on Netfilx in the Month of July,January,April, october and December.
- Least Movies of are Added on Netfilx in the month of February, May and June.
- Most Tv Shows are Added on Netflix in the Month of July, December,September.
- Least Tv Shows are Added on Netflix in the Month of February, May, and January

Ratings According to the Movies and TV Show

In [33]:
```
df["rating"].value_counts()[:-3]
```

Out[33]:
```
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
NC-17          3
UR             3
Name: rating, dtype: int64
```

In [34]:
```
sns.set_theme(style="whitegrid")
plt.figure(figsize=(10,8))
sns.countplot(palette="magma",x=df["rating"],hue=df["type"],order=pd.value_counts(
plt.xticks( fontsize=10)
plt.title("Top Rating Based on No. Of Content")
plt.show()
```



- TV-Y :Designed to be appropriate for all children
- TV-Y7 :Suitable for ages 7 and up
- G :Suitable for General Audiences
- TV-G :Suitable for General Audiences
- PG: Parental Guidance suggested
- TV-PG: Parental Guidance suggested
- PG-13:Parents strongly cautioned. May be inappropriate for ages under 13.
- TV-14 : Parents strongly cautioned. May not be suitable for ages under 14.
- R : Restricted. May be inappropriate for ages under 17.
- TV-MA : For Mature Audiences
- NC-17 : Inappropriate for ages 17 and under

- Most of the Movies and Tv Show are for Mature Audiencesfollowed by TV-14 rating

Movies Vs Tv show

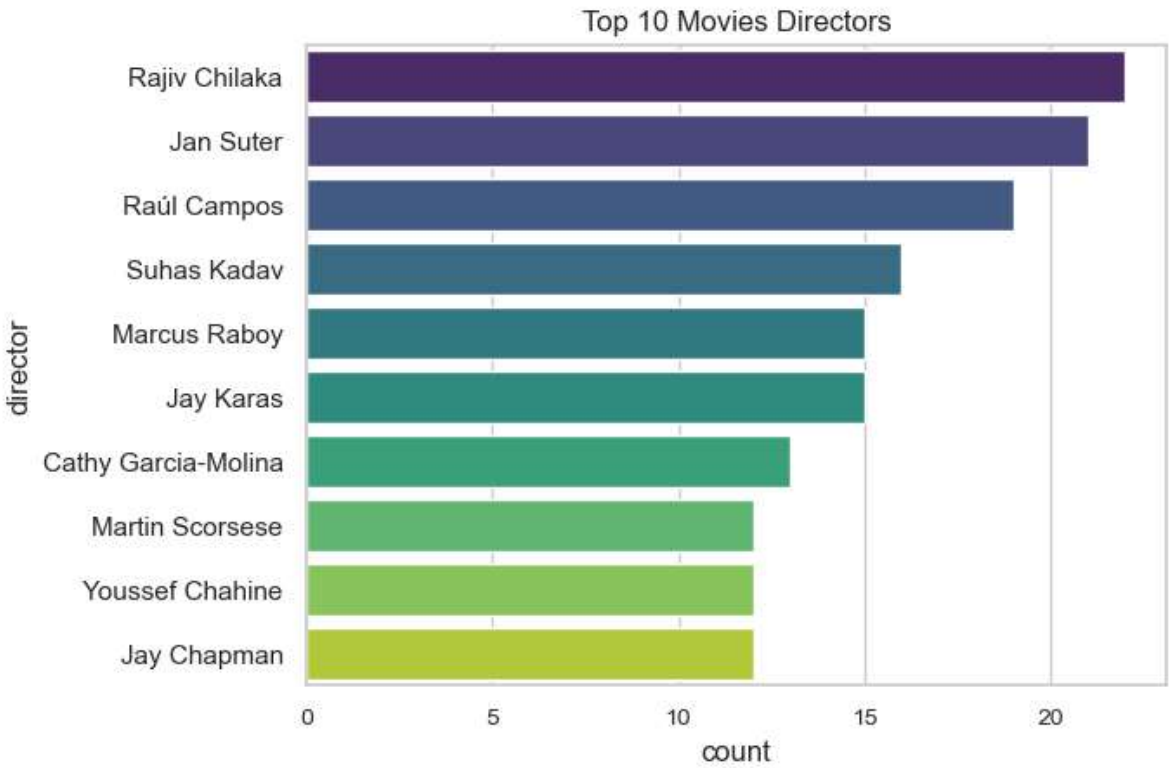
Netflix_Movies

In [35]:
```
movies_df=df.loc[(df["type"]=="Movie")]
movies_df.reset_index()
movies_df.columns.name=None
```

Top 10 Movies Directors

In [36]:
```
df_movie_dir=movies_df[["director"]]
df_movie_dir["director"]=df_movie_dir["director"].str.split(", ")
df_movie_dir= df_movie_dir.explode("director")
```

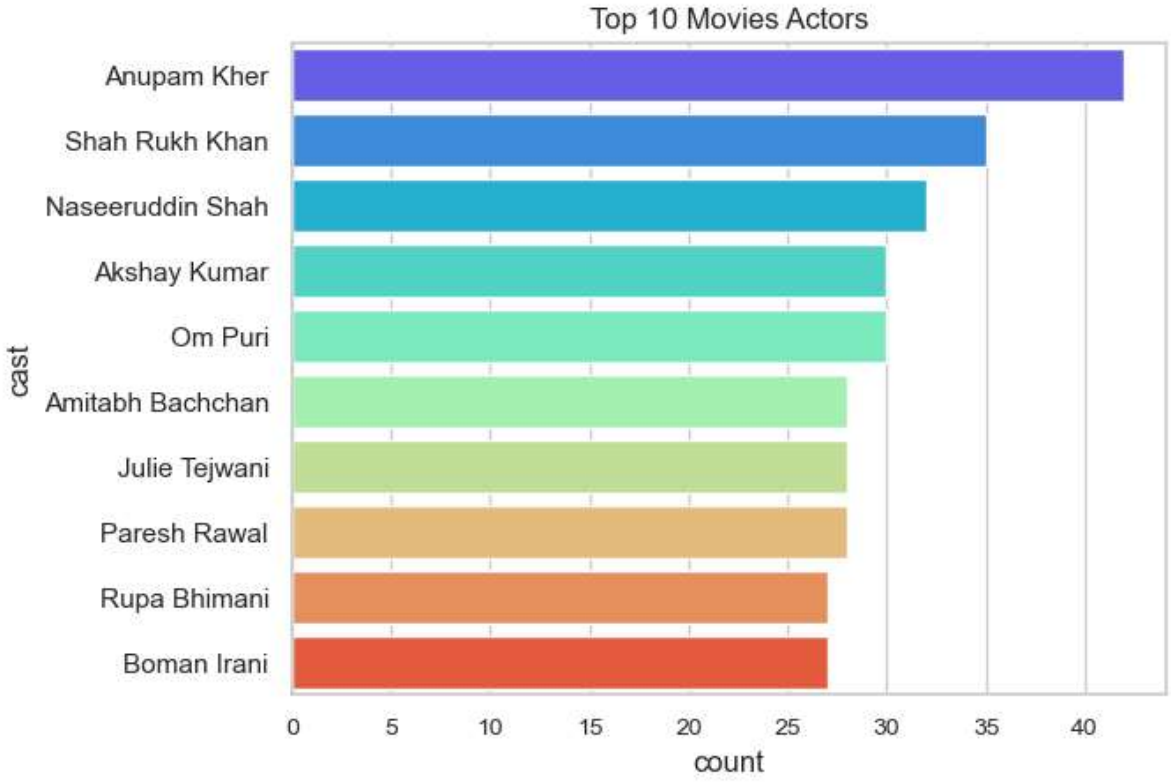
In [37]:
```
sns.set_theme(style="whitegrid")
sns.countplot(palette="viridis",y=df_movie_dir["director"],order=pd.value_counts(d
plt.xticks( fontsize=10)
plt.title("Top 10 Movies Directors")
plt.show()
```



Top 10 Movies Movies Actors

In [38]:
```
df_movie_cast=movies_df[["cast"]]
df_movie_cast["cast"]=df_movie_cast["cast"].str.split(", ")
df_movie_cast= df_movie_cast.explode("cast")
```

In [39]:
```
sns.set_theme(style="whitegrid")
sns.countplot(palette="rainbow",y=df_movie_cast["cast"],order=pd.value_counts(df_m
plt.xticks( fontsize=10)
plt.title("Top 10 Movies Actors")
plt.show()
```



- Top 10 Actors who Acted in Movie all are from India

Movies Duration in Boxplot

In [40]:
```
movies_df['duration'] = movies_df['duration'].astype(str).str.replace(' min', '')
movies_df.head(5)
```

Out[40]:

	show_id	type	title	director	cast	country	date_added	release_year	rati
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-
6	7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	
7	8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	T N
9	10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-
12	13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	T N

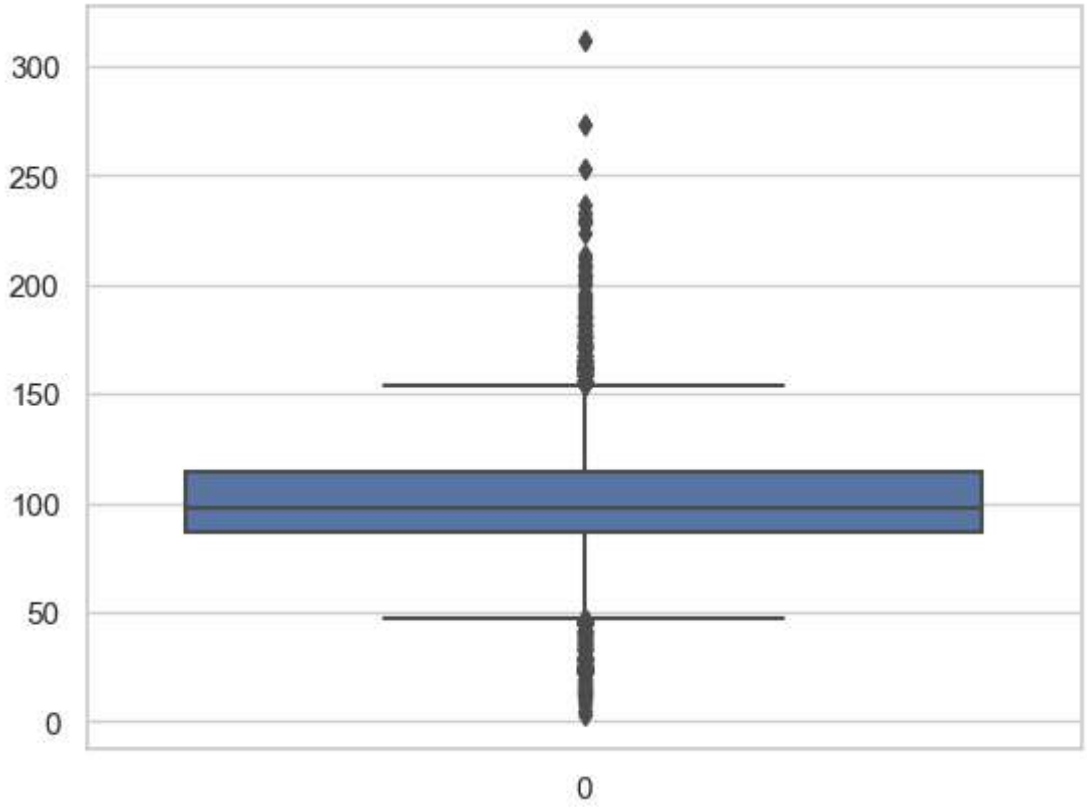
In [41]:
```
movies_df.dropna(subset=["duration"],axis=0,inplace=True)
```

In [42]:
```
movies_df["duration"]=movies_df["duration"].astype(float)
```

In [43]:
```
movies_df["duration"].describe()
```

Out[43]:
```
count    6128.000000
mean       99.577187
std        28.290593
min         3.000000
25%        87.000000
50%        98.000000
75%       114.000000
max       312.000000
Name: duration, dtype: float64
```

In [44]:
```
sns.boxplot(movies_df['duration'])
plt.show()
```

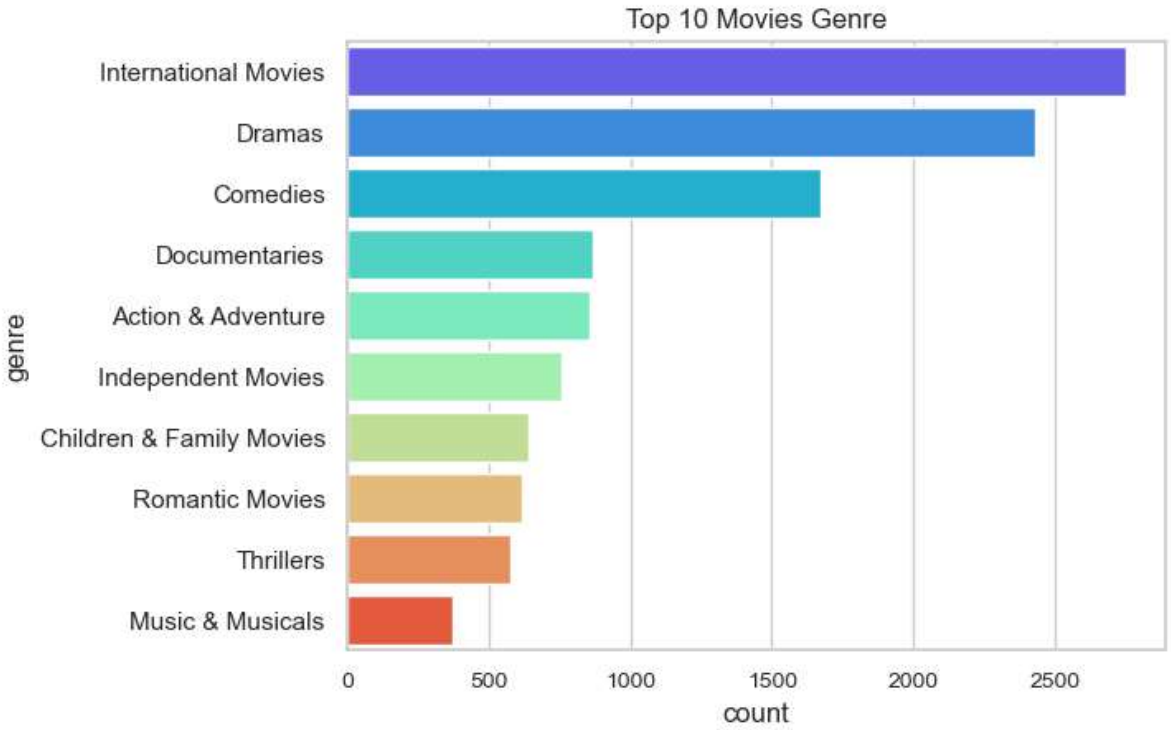


- we analyzed that there are few outliers in Duration of Movies
- The average duration of Movies is between 87 to 114 minutes
- The minumum duration of Movies is 50 minutes and Maximun Duration is 150 minutes as represented by wisker
- The minmum outliers below 50 and the maximum outliers are above 150

Top 10 Movies Movies Genre

In [45]:
```
df_movie_genre=movies_df[["genre"]]
df_movie_genre["genre"]=df_movie_genre["genre"].str.split(", ")
df_movie_genre= df_movie_genre.explode("genre")
```

In [46]:
```
sns.set_theme(style="whitegrid")
sns.countplot(palette="rainbow",y=df_movie_genre["genre"],order=pd.value_counts(df_
plt.xticks( fontsize=10)
plt.title("Top 10 Movies Genre")
plt.show()
```



- The top Genre are international Movies followed by Dramas and Comedies

Netfix_Tv_shows

In [47]:
```
tv_show_df=df.loc[(df["type"]=="TV Show")]
tv_show_df.reset_index()
tv_show_df.columns.name=None
```

In [48]:
```
tv_show_df.head(5)
```

Out[48]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
1	2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 S
4	5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se
5	6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 S

In [49]:

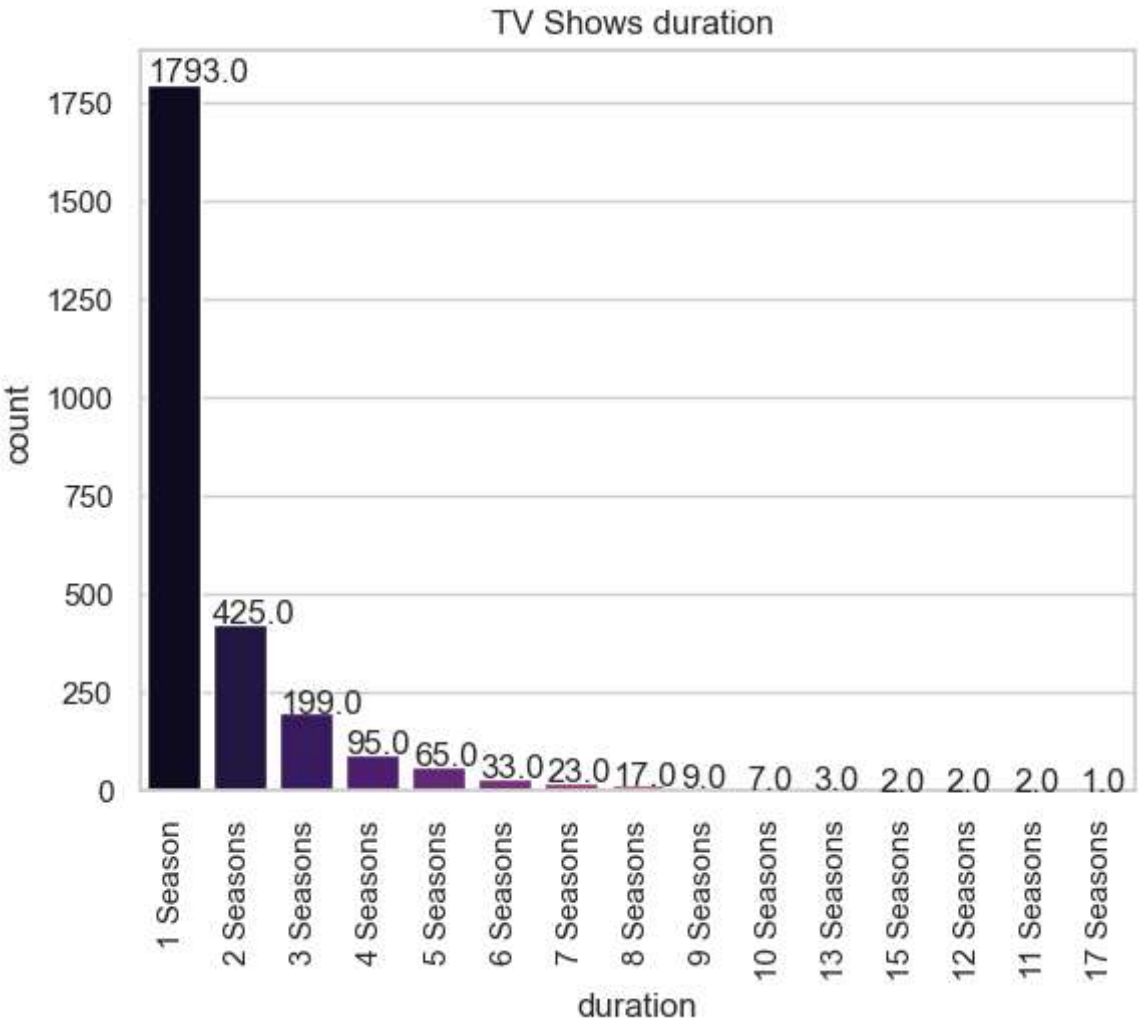
```
tv_show_df['duration'].astype(str)
```

Out[49]:

```
1        2 Seasons
2         1 Season
3         1 Season
4        2 Seasons
5         1 Season
           ...
8795     2 Seasons
8796     2 Seasons
8797     3 Seasons
8800      1 Season
8803     2 Seasons
Name: duration, Length: 2676, dtype: object
```

In [50]:

```
sns.set_theme(style="whitegrid")
ax=sns.countplot(palette="magma",x=tv_show_df['duration'],order=pd.value_counts(tv_
plt.xticks(rotation = 90)
for p in ax.patches:
        ax.annotate(str(p.get_height()), (p.get_x() * 1.005, (p.get_height() * 1.0
plt.title("TV Shows duration ")
plt.show()
```

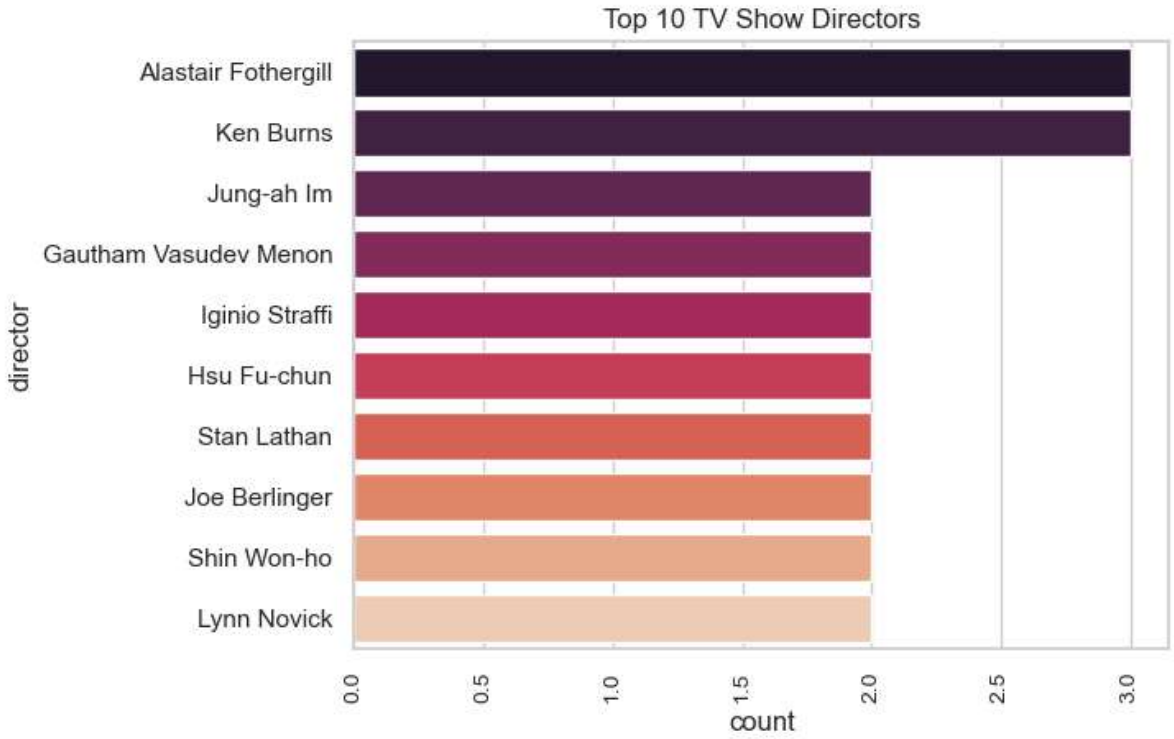


- we analyze the Tv show data and came to know 1793 shows have only 1 Season
- there only few shows which have more than 10 seasons
- there is one show which has 17 Seasons

Top TV Shows Directors

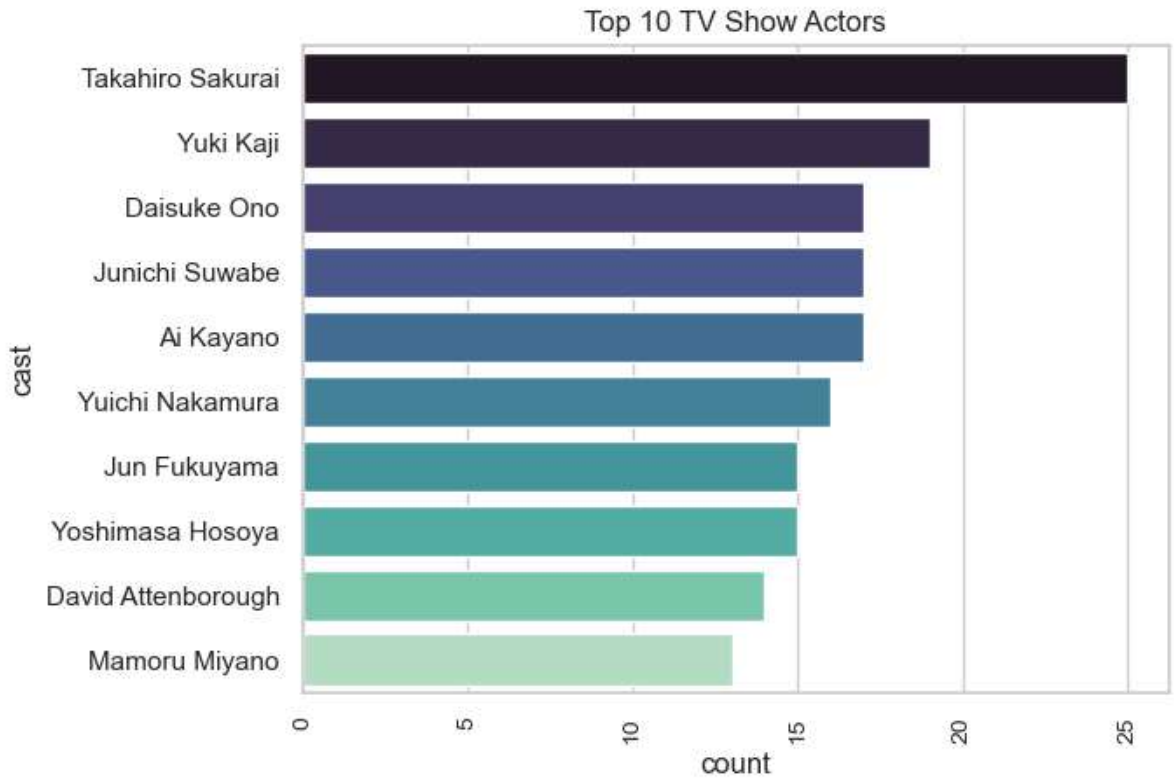
In [51]:
```
df_tv_dir=tv_show_df[["director"]]
df_tv_dir["director"]=df_tv_dir["director"].str.split(", ")
df_tv_dir= df_tv_dir.explode("director")
```

In [52]:
```
sns.set_theme(style="whitegrid")
sns.countplot(palette="rocket",y=df_tv_dir["director"],order=pd.value_counts(df_tv_
plt.xticks(rotation=90, fontsize=10)
plt.title("Top 10 TV Show Directors")
plt.show()
```



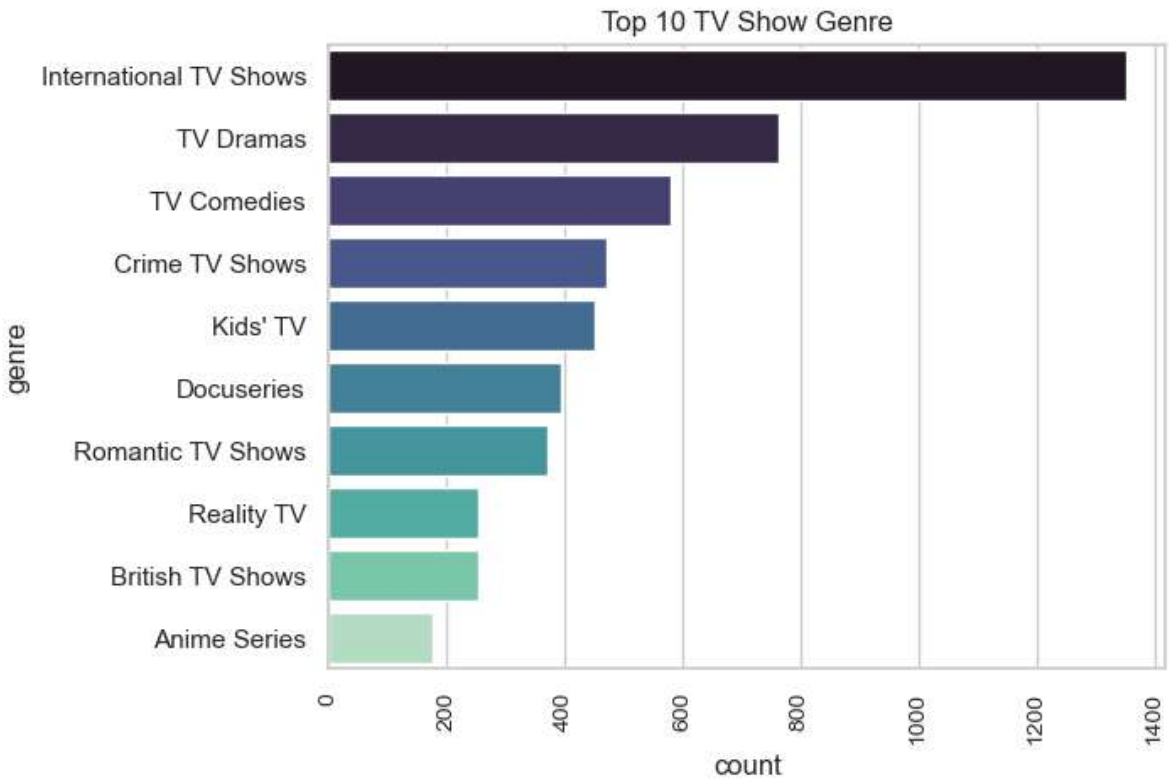
In [53]:
```
df_tv_cast=tv_show_df[["cast"]]
df_tv_cast["cast"]=df_tv_cast["cast"].str.split(", ")
df_tv_cast= df_tv_cast.explode("cast")
```

In [54]:
```
sns.set_theme(style="whitegrid")
sns.countplot(palette="mako",y=df_tv_cast["cast"],order=pd.value_counts(df_tv_cast
plt.xticks(rotation=90, fontsize=10)
plt.title("Top 10 TV Show Actors")
plt.show()
```



In [55]:
```
df_tv_genre=tv_show_df[["genre"]]
df_tv_genre["genre"]=df_tv_genre["genre"].str.split(", ")
df_tv_genre= df_tv_genre.explode("genre")
```

In [56]:
```
sns.set_theme(style="whitegrid")
sns.countplot(palette="mako",y=df_tv_genre["genre"],order=pd.value_counts(df_tv_ge
plt.xticks(rotation=90, fontsize=10)
plt.title("Top 10 TV Show Genre")
plt.show()
```



Days on Which Content is added Every Month

In [57]:
```
Date = df[['date_added']].dropna()
Date['date'] = Date['date_added'].apply(lambda x : x.lstrip().split(' ')[1].replac
Date['months'] = Date['date_added'].apply(lambda x : x.lstrip().split(' ')[0])
Date['years'] = Date['date_added'].apply(lambda x : x.split(', ')[-1])
Order_Of_Month = ['January', 'February', 'March', 'April', 'May','June', 'July', '
df8 = Date.groupby('years')['months'].value_counts().unstack().fillna(0)[Order_Of_
```

In [58]:
```
Date['date'].value_counts().sort_values(ascending=False).reset_index().head()
```

Out[58]:

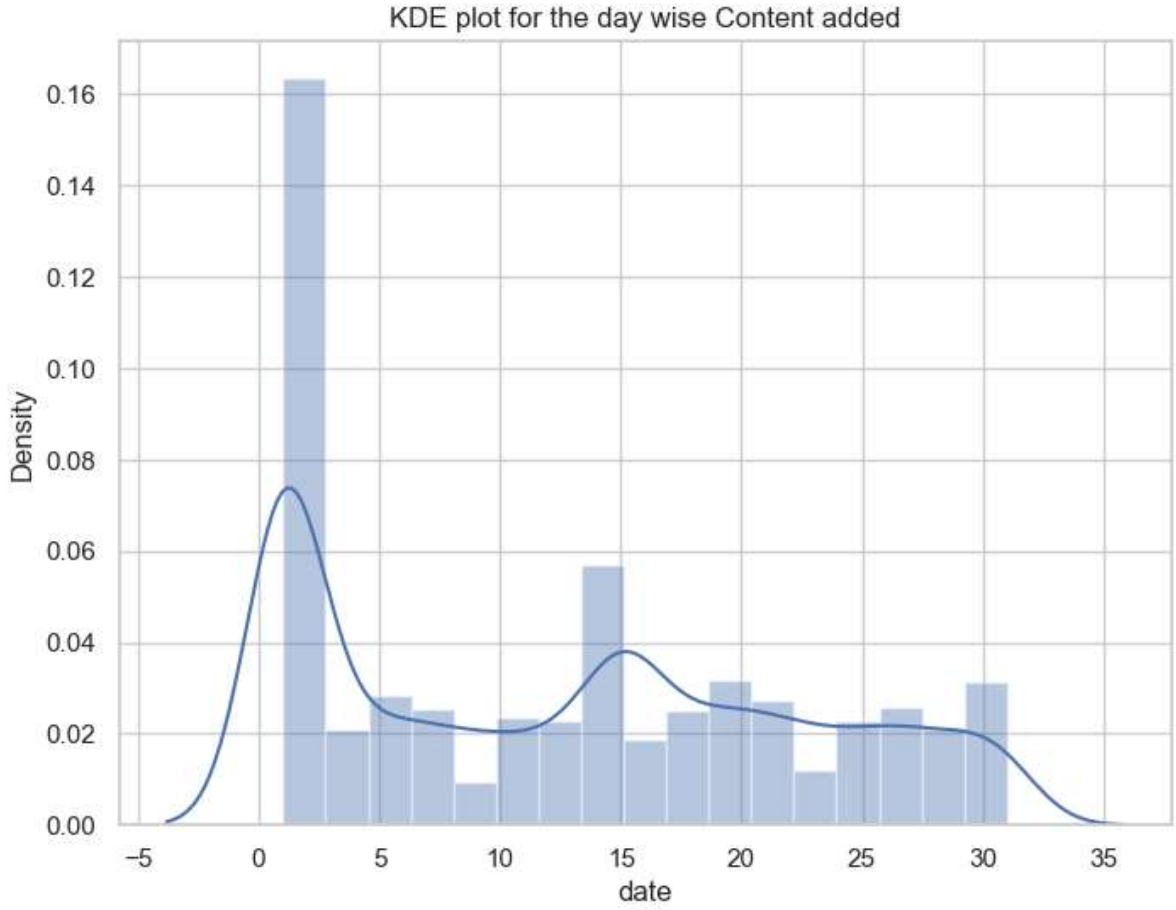
	index	date
0	1	2212
1	15	687
2	2	325
3	16	289
4	31	274

In [59]:
```
perc_1 = (Date['date'].value_counts()[30]/Date['date'].value_counts().sum())*100
print((round(perc_1,2)),"%")
```

```
1.6 %
```

In [60]:
```
plt.figure(figsize=(8,6))
sns.distplot(Date['date'])
```

```
plt.title('KDE plot for the day wise Content added')
plt.show()
```

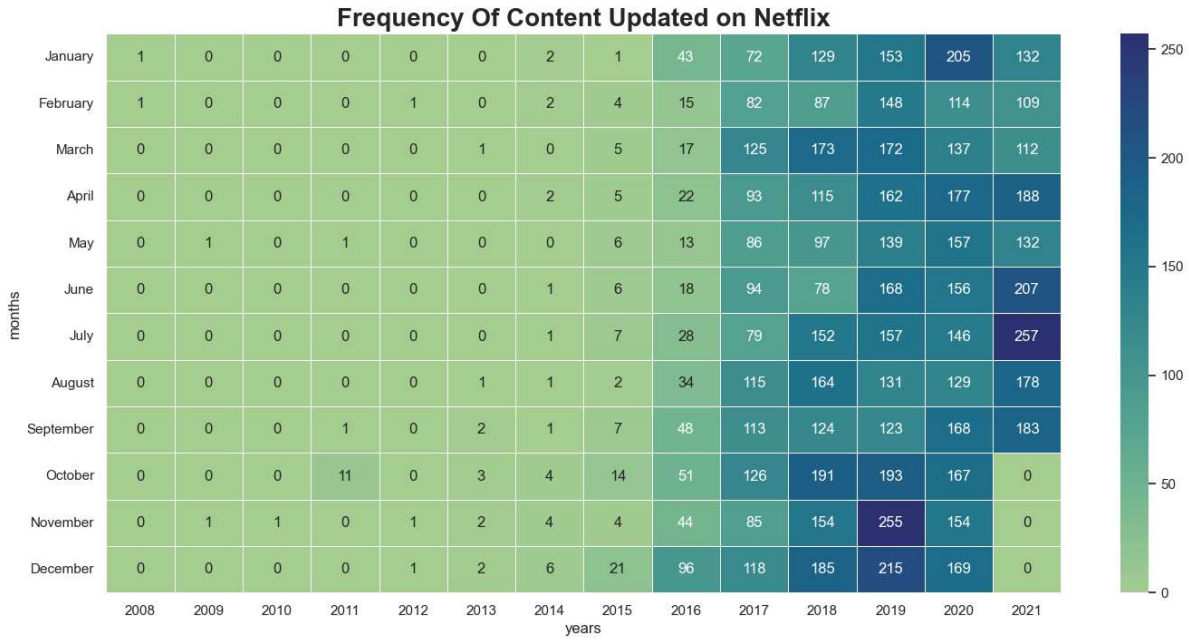


- We analyzed that content is added to the Netfix on 1st date of the month nearly 25% of the times follwed by 15th date of the month nearly 7.8% of the time
- We analyzed that content is added to the Netflix on 29th date is the lowest among all the days is 1.6%

Correlation of the Year and Months of the content added

In [61]:
```
Date = df[['date_added']].dropna()
Date['months'] = Date['date_added'].apply(lambda x : x.lstrip().split(' ')[0])
Date['years'] = Date['date_added'].apply(lambda x : x.split(', ')[-1])
Order_Of_Month = ['January', 'February', 'March', 'April', 'May','June', 'July', '
df8 = Date.groupby('years')['months'].value_counts().unstack().fillna(0)[Order_Of_
```

In [62]:
```
plt.figure(figsize=(17,8))
sns.heatmap(data=df8,annot=True,linewidth=0.5,fmt=".0f",cmap="crest")
plt.title('Frequency Of Content Updated on Netflix', fontsize=20,fontfamily='Arial
plt.show()
```



- Most amount of content was added in july 2021 (257) followed by November 2019(255).
- There is a significant growth in content added on Netflix from 2016.

5. MISSING VALUE TREATMENT (Dropping Null Values)

In [63]:
```
df.dropna(subset=["duration", "rating", "date_added"], inplace = True)
```

In [64]:
```
round((df.isnull().sum()/df.shape[0]*100),2).sort_values(ascending = False)
```

Out[64]:
```
director        29.82
country          9.43
cast             9.39
show_id          0.00
type             0.00
title            0.00
date_added       0.00
release_year     0.00
rating           0.00
duration         0.00
genre            0.00
description      0.00
dtype: float64
```

In [65]:
```
# Replacing NaN values in cast column with "No Cast"

df["cast"].replace(np.NaN, "No Cast", inplace = True)
```

In [66]:
```
# Replacing NaN values in country column with "Unknown"

df["country"].replace(np.NaN, "Unknown", inplace = True)
```

In [67]:
```
round((df.isna().sum()/df.shape[0]*100),2).sort_values(ascending = False)
```

Out[67]:
```
director        29.82
show_id          0.00
type             0.00
title            0.00
cast             0.00
country          0.00
date_added       0.00
release_year     0.00
rating           0.00
duration         0.00
genre            0.00
description      0.00
dtype: float64
```

Insights based on Non-Graphical and Visual Analysis

- Nearly 2/3rd of the content on netflix are movies and remaining 1/3rd of them are TV Show.
- We can see that US, India, United Kingdom, Canada and France contribute 72.4% of the Top10 countries.
- About 1200 new movies were released in both 2018 and 2019.
- The growth in content started from 2014.
- The growth in contents are higher in the first three months and the last three months of the year.
- Least number of contents are added in the month of February.
- Most movies added in the months of July, December, April, october and January.
- Least movies added in the months of June, May and February.
- Most TV Shows added in the months of July and December.
- Least TV Shows added in the months of January and February.
- Analysed that content is added to Netflix on 1st date of the month.
- Most amount of content was added in july 2021 (257) followed by November 2019(255).
- There is a significant growth in content added on Netflix from 2016.
- TV-MA rating has the highest number of Movies and TV Shows available followed by TV-14 rating.
- we analyzed that there are few outliers in Duration of Movies.
- The average duration of Movies is between 87 to 114 minutes.
- The minumum duration of Movies is 50 minutes and Maximun Duration is 150 minutes as represented by wisker.
- The minmum outliers below 50 and the maximum outliers are above 150.
- When we analyze the Tv show data we came to know 1793 shows have only 1 Season.
- where as we there only few shows which have more than 10 seasons.
- there is one show which has 17 Seasons.
- Anupam Kher has acted in the most number of Movies followed by Shah Rukh Khan and Naseeruddin Shah.
- All of the top 10 actors based on number of Movies are from India.

Business Insights

- 72% of the content available on Netflix belongs to the movie category while Tv shows consists of the rest 30%.
- Rajiv Chilaka, Jan Suter and Raul campos have directed most amount of titles among Movies and TV Shows.
- Anupam Kher has acted in the most number of movies followed by Shah Rukh Khan and Naseeruddin Shah.
- All of the top 10 actors based on number of movies are of India.
- United States is the largest country producing movies on Netflix.
- International Movies is most watched Genre.
- In the year 2018, most movies and Tv shows were added.
- TV-MA rating has the highest number of rating for Movies and TV Shows.

Recommendations

1. Which countries to target?

- United States, India, United Kingdom, Canada and France contribute 72.41% of the total content. So including content produced from these countries is recommended. #### 2. Which Movie/TV Show genre to add?
- International Movies, Dramas, Comedies are the top 3 genres for Movies on Netflix.
- International TV Shows, TV Dramas, TV Comedies are the top 3 genres for TV Shows on Netflix.

3. Average duration of the Movie/TV shows?

- Around 87 to 114 minutes for the movies
- 67% of the total number of TV Shows have only 1 Season. Not more than 3 seasons is recommended.

4. In which month to add a Movie/TV show?

- For Movies, most months are suitable however, February, May and June should be looked into as they have the least amount of Movies added.
- For TV Shows, most months are suitable however, January, February and May should be looked into as they have the least amount of TV Shows added.

5. Which directors can be preferred for Movies/TV Shows?

- Rajiv Chilaka, Jan Suter, Raul Campos, Suhas Kadav and Marcus Raboy are top 5 most recurring directors that can be preferred.

6. Which actors can be considered Movies/TV Shows?

- Anupam Kher, Shah Rukh Khan, Naseeruddin Shah, Akshay Kumar and Om Puri are the top 5 most recurring Movie actors.
- Takahiro Sakurai, Yuji Kaji, Junichi Suwabe, Daisuke Ono and Ai Kayano are the top 5 most recurring TV Shows actors.

7. When to add new content?

- Usually 1st day of the month is most recurring day when new content is added however, 15th day of the month can also be considered as a second option which is considerably less than 1st day of the month.

8. Which rating to consider?

- TV-MA is the most recurring rating for Movies as well as TV-Shows, so it is most recommended.

Additional point to be considered:

- There is a significant drop in the amount of movies added to the Netflix after 2018 and for TV shows there is a drop after
  2020, which must be looked into while making future business regarding adding content to Netflix.

In [ ]:

```
```