Analysis of the content available on OTT giant : NETFLIX

Importing the Data

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df=pd.read_csv("movies.csv")
```

In [3]:
```
df.head(10)
```

Out[3]:

	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...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	PG	91 min	Children & Family Movies	Equestria's divided. But a bright-eyed hero be...
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model s...
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amateur bakers face off in...
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	Comedies, Dramas	A woman adjusting to life after a loss contend...

Goal - Growth of business

- can be done in multiple ways:-
  - Increasing customer base (Getting new users).
  - Decreasing customer churn rate (Preventing users to leave the platform, to increase customer stickyness).
  - So, we can find insights related to shows/movies to produce, to increase the busineess.

Problem Statement: - Finding which type of shows/movies to produce and how they can grow the business in different countries

- Questions that can be asked for finding types of shows/movies to produce: -
  - Which are popular genres, actors, directors?

- TV/Movie rating (PG-13, NC-17, etc.)
- Global and coutry specific trend
- Demographic distribution of the users (based on age)
- Country based popularity analysis
- Watch time (bing worthy or not).
- Evergreeness of the shows (Time a user watched)
- Movies/Shows sequels prefereable or not.

Observing the data

In [4]:
```
df.head()
```

Out[4]:

	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...

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

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

In [6]:
```
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
```

- Total 8807 data entries are there, of which: -

ountry, data_added, rating, and **duration**.

- Data type of **type** and **rating** to be changed into category data type.

In [7]:
```
df.describe(include='all')
```

Out[7]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8807.000000	8803	8804	8807	8807
unique	8807	2	8807	4528	7692	748	1767	NaN	17	220	514	8775
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	NaN	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	1	19	19	2818	109	NaN	3207	1793	362	4
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.180198	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.819312	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN	NaN

- Data ranges from Year **1925** to **2021**.

In [8]:
```
df.loc[df.duplicated()]
```

Out[8]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description

- No duplication of enitre row

In [9]:
```
# Checking duplicates in columns = titles : If any movies data is added twice in the database.
df.loc[df['title'].duplicated()]
```

Out[9]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description

- No movie is repeated/duplicated.

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
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
```

Column-wise data observation

Show_id

In [11]:
```
df['show_id'].unique
```

Out[11]:
```
<bound method Series.unique of 0          s1
1          s2
2          s3
3          s4
4          s5
         ...
8802    s8803
8803    s8804
8804    s8805
8805    s8806
8806    s8807
Name: show_id, Length: 8807, dtype: object>
```

In [12]:
```
df['show_id'].value_counts()
```

Out[12]:
```
s1       1
s5875    1
s5869    1
s5870    1
s5871    1
        ..
s2931    1
s2930    1
s2929    1
s2928    1
s8807    1
Name: show_id, Length: 8807, dtype: int64
```

- Unique show_id for each entry of a movie/show.
- This column is fine but not useful for the analysis, So, remove this column.

type

In [13]:
```
df['type'].value_counts()
```

Out[13]:
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

- Categorical column for a show
- Convert the column into category dtype

title

In [14]:
```
df['title'].value_counts()
```

Out[14]:
```
Dick Johnson Is Dead                    1
Ip Man 2                                1
Hannibal Buress: Comedy Camisado        1
Turbo FAST                              1
Masha's Tales                           1
                                       ..
Love for Sale 2                         1
ROAD TO ROMA                            1
Good Time                               1
Captain Underpants Epic Choice-o-Rama   1
Zubaan                                  1
Name: title, Length: 8807, dtype: int64
```

- There are all unique titles.

director

In [15]:
```
df['director'].value_counts()
```

Out[15]:
```
Rajiv Chilaka                     19
Raúl Campos, Jan Suter            18
Marcus Raboy                      16
Suhas Kadav                       16
Jay Karas                         14
                                  ..
Raymie Muzquiz, Stu Livingston     1
Joe Menendez                       1
Eric Bross                         1
Will Eisenberg                     1
Mozez Singh                        1
Name: director, Length: 4528, dtype: int64
```

In [16]:
```
df.loc[df['director'].str.len()>25,'director']
```

Out[16]:
```
6                          Robert Cullen, José Luis Ucha
16         Pedro de Echave García, Pablo Azorín Williams
30     Ashwiny Iyer Tiwari, Abhishek Chaubey, Saket C...
68     Hanns-Bruno Kammertöns, Vanessa Nöcker, Michae...
84                           JJC Skillz, Funke Akindele
                             ...
8714        Stephen Donnelly, Olly Reid, Jun Falkenstein
8728                    Heidi Brandenburg, Mathew Orzel
8737              Milla Harrison-Hansley, Alicky Sussman
8739                          Frank Capra, Anatole Litvak
8765    Jovanka Vuckovic, Annie Clark, Roxanne Benjami...
Name: director, Length: 463, dtype: object
```

- There are NaN values in the director column.
- There are nested names of the director in single rows.

cast

In [17]:
```
df['cast'].value_counts()
```

Out[17]:
```
David Attenborough                                                                                                                                 19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil                                                            14
Samuel West                                                                                                                                        10
Jeff Dunham                                                                                                                                         7
David Spade, London Hughes, Fortune Feimster                                                                                                        6
                                                                                                                                                   ..
Michael Peña, Diego Luna, Tenoch Huerta, Joaquin Cosio, José María Yazpik, Matt Letscher, Alyssa Diaz                                               1
Nick Lachey, Vanessa Lachey                                                                                                                         1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayashi, Ken Yasuda, Arata Furuta, Suzuki Matsuo, Koichi Yamadera, Arata Iura, Chikako Kaku, Kotaro Yoshida    1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Agu, Dele Odule, Femi Adebayo, Bayray McNwizu, Biodun Stephen    1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranjan Tripathy               1
Name: cast, Length: 7692, dtype: int64
```

In [18]:
```
df['cast'].isnull().sum()
```

Out[18]:
```
825
```

- There are nested lists of casts in each row.
- There are 825 null values in cast column

country

In [19]:
```
df['country'].value_counts()
```

Out[19]:
```
United States                             2818
India                                      972
United Kingdom                             419
Japan                                      245
South Korea                                199
                                          ...
Romania, Bulgaria, Hungary                   1
Uruguay, Guatemala                           1
France, Senegal, Belgium                     1
Mexico, United States, Spain, Colombia       1
United Arab Emirates, Jordan                 1
Name: country, Length: 748, dtype: int64
```

In [20]:
```
df['country'].isnull().sum()
```

Out[20]:
```
831
```

- There are null values in the country column.
- There are nested values in rows of country column.

date_added

In [21]:
```
df['date_added'].value_counts()
```

Out[21]:
```
January 1, 2020      109
November 1, 2019      89
March 1, 2018         75
December 31, 2019     74
October 1, 2018       71
                     ...
December 4, 2016       1
November 21, 2016      1
November 19, 2016      1
November 17, 2016      1
January 11, 2020       1
Name: date_added, Length: 1767, dtype: int64
```

In [22]:
```
df['date_added'].isnull().sum()
```

Out[22]:
```
10
```

- The date_added column data type needs to be changed to datetime.
- There are null values in date_added column.

release_year

In [23]:
```
x=df['release_year'].unique()
np.sort(x)
```

Out[23]:
```
array([1925, 1942, 1943, 1944, 1945, 1946, 1947, 1954, 1955, 1956, 1958,
       1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969,
       1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980,
       1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991,
       1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002,
       2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013,
       2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021], dtype=int64)
```

- Release_year column is fine

rating

In [24]:
```
df['rating'].value_counts()
```

Out[24]:
```
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
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```

Out[25]: 4

- There are three values of duration column coming in rating column.
- There are 4 null values in rating column.

duration

In [26]:
```
df['duration'].unique()
```

Out[26]:
```
array(['90 min', '2 Seasons', '1 Season', '91 min', '125 min',
       '9 Seasons', '104 min', '127 min', '4 Seasons', '67 min', '94 min',
       '5 Seasons', '161 min', '61 min', '166 min', '147 min', '103 min',
       '97 min', '106 min', '111 min', '3 Seasons', '110 min', '105 min',
       '96 min', '124 min', '116 min', '98 min', '23 min', '115 min',
       '122 min', '99 min', '88 min', '100 min', '6 Seasons', '102 min',
       '93 min', '95 min', '85 min', '83 min', '113 min', '13 min',
       '182 min', '48 min', '145 min', '87 min', '92 min', '80 min',
       '117 min', '128 min', '119 min', '143 min', '114 min', '118 min',
       '108 min', '63 min', '121 min', '142 min', '154 min', '120 min',
       '82 min', '109 min', '101 min', '86 min', '229 min', '76 min',
       '89 min', '156 min', '112 min', '107 min', '129 min', '135 min',
       '136 min', '165 min', '150 min', '133 min', '70 min', '84 min',
       '140 min', '78 min', '7 Seasons', '64 min', '59 min', '139 min',
       '69 min', '148 min', '189 min', '141 min', '130 min', '138 min',
       '81 min', '132 min', '10 Seasons', '123 min', '65 min', '68 min',
       '66 min', '62 min', '74 min', '131 min', '39 min', '46 min',
       '38 min', '8 Seasons', '17 Seasons', '126 min', '155 min',
       '159 min', '137 min', '12 min', '273 min', '36 min', '34 min',
       '77 min', '60 min', '49 min', '58 min', '72 min', '204 min',
       '212 min', '25 min', '73 min', '29 min', '47 min', '32 min',
       '35 min', '71 min', '149 min', '33 min', '15 min', '54 min',
       '224 min', '162 min', '37 min', '75 min', '79 min', '55 min',
       '158 min', '164 min', '173 min', '181 min', '185 min', '21 min',
       '24 min', '51 min', '151 min', '42 min', '22 min', '134 min',
       '177 min', '13 Seasons', '52 min', '14 min', '53 min', '8 min',
       '57 min', '28 min', '50 min', '9 min', '26 min', '45 min',
       '171 min', '27 min', '44 min', '146 min', '20 min', '157 min',
       '17 min', '203 min', '41 min', '30 min', '194 min', '15 Seasons',
       '233 min', '237 min', '230 min', '195 min', '253 min', '152 min',
       '190 min', '160 min', '208 min', '180 min', '144 min', '5 min',
       '174 min', '170 min', '192 min', '209 min', '187 min', '172 min',
       '16 min', '186 min', '11 min', '193 min', '176 min', '56 min',
       '169 min', '40 min', '10 min', '3 min', '168 min', '312 min',
       '153 min', '214 min', '31 min', '163 min', '19 min', '12 Seasons',
       nan, '179 min', '11 Seasons', '43 min', '200 min', '196 min',
       '167 min', '178 min', '228 min', '18 min', '205 min', '201 min',
       '191 min'], dtype=object)
```

In [27]:
```
df['duration'].isnull().sum()
```

Out[27]: 3

- There are 3 null values present for which the values are there in correspending rows of rating column.
- There are two type of durations - movie duration and TV shows duration. We need to put them in separate columns.
- Also, need to convert the two columns into simple integer for further analysis.

listed_in

In [28]:
```
df['listed_in'].value_counts()
```

Out[28]:
```
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
                                                     ...
Kids' TV, TV Action & Adventure, TV Dramas              1
TV Comedies, TV Dramas, TV Horror                       1
Children & Family Movies, Comedies, LGBTQ Movies        1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows      1
Cult Movies, Dramas, Thrillers                          1
Name: listed_in, Length: 514, dtype: int64
```

In [29]:
```
df['listed_in'].isnull().sum()
```

Out[29]:
```
0
```

- Need to unnest the genres.
- Rename the column to genres.

description

In [30]:
```
df['description']
```

Out[30]:
```
0       As her father nears the end of his life, filmm...
1       After crossing paths at a party, a Cape Town t...
2       To protect his family from a powerful drug lor...
3       Feuds, flirtations and toilet talk go down amo...
4       In a city of coaching centers known to train I...
                              ...
8802    A political cartoonist, a crime reporter and a...
8803    While living alone in a spooky town, a young g...
8804    Looking to survive in a world taken over by zo...
8805    Dragged from civilian life, a former superhero...
8806    A scrappy but poor boy worms his way into a ty...
Name: description, Length: 8807, dtype: object
```

- This column is not useful for analysis. Drop the column.

Assessment of the data

Quality issues with the Data

```
- Missing values in the director, cast, country, data_added. rating, and duration.
- Date format needs to be changed to datetime format.
- Data type of type and rating to be changed into category data type.
- Duration col format needs to be in integer format only.
- Rename the column listed_in : genre, rating : certification
```

Structural issues with the Data

- director, Cast, listed_in (genre) info to be changed into separate rows.
- TV shows duration should have separate column, same with movies.
- Three values of duration column coming in rating column.
- Remove unnecessary columns -> description, show_id

Data Cleaning

In [31]:
```
df.sample(5)
```

Out[31]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
1964	s1965	Movie	High & Low The Movie	Shigeaki Kubo	Takanori Iwata, Akira, Sho Aoyagi, Hiroomi Tos...	Japan	September 20, 2020	2016	TV-MA	129 min	Action & Adventure, International Movies	The five rival gangs ruling the SWORD district...
6568	s6569	TV Show	Darr Sabko Lagta Hai	NaN	Bipasha Basu	India	March 1, 2018	2015	TV-MA	1 Season	International TV Shows, TV Horror, TV Thrillers	In this chilling horror anthology series, actr...
2599	s2600	TV Show	El señor de los Cielos	NaN	Rafael Amaya, Ximena Herrera, Robinson Díaz, R...	United States, Mexico, Colombia	April 30, 2020	2019	TV-MA	7 Seasons	Crime TV Shows, International TV Shows, Spanis...	Only Aurelio Casillas can fill Pablo Escobar's...
3440	s3441	TV Show	Rhythm + Flow	NaN	Cardi B, Chance The Rapper, T.I.	United States	October 9, 2019	2019	TV-MA	1 Season	Reality TV	In this music competition show, judges Tip "T....
5086	s5087	Movie	Tom Segura: Disgraceful	Jay Karas	Tom Segura	United States	January 12, 2018	2018	TV-MA	71 min	Stand-Up Comedy	Tom Segura gives voice to the sordid thoughts ...

1. Removing unnecessary columns

In [32]:
```
df.drop(columns=['description','show_id'],inplace=True)
```

In [33]:
```
df.head()
```

Out[33]:

	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries
1	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
2	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...
3	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV
4	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 ...

2. Changing data type of columns

a. To category dtype - rating, type

In [34]:
```
df['type']=df['type'].astype('category')
```

In [35]:
```
df['rating'].value_counts()
```

Out[35]:
```
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
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```

- Rating column needs to be fixed with null values and the duration values then only can be converted into category dtype

b. Changing date format to datetime dtype

In [36]:
```
df['date_added']=pd.to_datetime(df['date_added'])
```

In [37]:
```
df.head()
```

Out[37]:

	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

3. Renaming column

In [38]:
```
df.rename(columns={'listed_in':'genre','rating':'certification'},inplace=True)
```

In [39]:
```
df.head()
```

Out[39]:

	type	title	director	cast	country	date_added	release_year	certification	duration	genre
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

ation column values

In [40]:
```
df['certification'].value_counts()
```

Out[40]:
```
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
74 min         1
84 min         1
66 min         1
Name: certification, dtype: int64
```

In [41]:
```
idx=df[df['certification'].str.contains('min')==True].index
```

In [42]:
```
for i in idx:
    df.loc[i,'duration']=df.loc[i,'certification']
    df.loc[i,'certification']=np.NaN
df.loc[idx,:]
```

Out[42]:

	type	title	director	cast	country	date_added	release_year	certification	duration	genre
5541	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	NaN	74 min	Movies
5794	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	NaN	84 min	Movies
5813	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	NaN	66 min	Movies

5. Fixing Missing Values

In [43]:
```
df.isnull().sum()
```

Out[43]:
```
type              0
title             0
director       2634
cast            825
country         831
date_added       10
release_year      0
certification     7
duration          0
genre             0
dtype: int64
```

In [44]:
```
df.shape
```

Out[44]:
```
(8807, 10)
```

Column - certification

- As, the TV certificaitons and movie certifications are different.
- Therefore, we will add mode value to TV and movie correspondingly.

In [45]:
```
df.loc[df['type']=='TV Show','certification'].value_counts()
```

Out[45]:
```
TV-MA       1145
TV-14        733
TV-PG        323
TV-Y7        195
TV-Y         176
TV-G          94
NR             5
R              2
TV-Y7-FV       1
Name: certification, dtype: int64
```

In [46]:
```
# Adding TV-MA to the NaN values of the TV shows
df.loc[df['type']=='TV Show','certification']=df.loc[df['type']=='TV Show','certification'].fillna('TV-MA')
```

In [47]:
```
df.loc[df['type']=='Movie','certification'].value_counts()
```

Out[47]:
```
TV-MA       2062
TV-14       1427
R            797
TV-PG        540
PG-13        490
PG           287
TV-Y7        139
TV-Y         131
TV-G         126
NR            75
G             41
TV-Y7-FV       5
NC-17          3
UR             3
Name: certification, dtype: int64
```

In [48]:
```
df['certification'].value_counts()
```

Out[48]:
```
TV-MA       3209
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: certification, dtype: int64
```

In [49]:
```
# Adding TV-MA to the NaN values of the TV shows
df.loc[df['type']=='Movie','certification']=df.loc[df['type']=='Movie','certification'].fillna('TV-MA')
```

Out[50]:
```
0
```

Column - date_added

In [51]:
```
df['date_added'].isnull().sum()
```

Out[51]:
```
10
```

- As, there are only 10 values i.e., 0.1% of the total values, therefore we can add mode of date_added according to type of the content (to add more granularity to mode).

In [52]:
```
def clean_date(x):
    x['date_mode']=x['date_added'].mode()[0]
    return x
```

In [53]:
```
df=df.groupby('type').apply(clean_date)
```

In [54]:
```
df['date_added'].fillna(df['date_mode'],inplace=True)
```

In [55]:
```
df.isnull().sum()
```

Out[55]:
```
type               0
title              0
director        2634
cast             825
country          831
date_added         0
release_year       0
certification      0
duration           0
genre              0
date_mode          0
dtype: int64
```

In [56]:
```
df.drop('date_mode',axis=1,inplace=True)
```

Column - country

In [57]:
```
df['country'].isnull().sum()
```

Out[57]:
```
831
```

In [58]:
```
df['country'].value_counts()
```

Out[58]:
```
United States                                2818
India                                         972
United Kingdom                                419
Japan                                         245
South Korea                                   199
                                             ...
Romania, Bulgaria, Hungary                      1
Uruguay, Guatemala                              1
France, Senegal, Belgium                        1
Mexico, United States, Spain, Colombia          1
United Arab Emirates, Jordan                    1
Name: country, Length: 748, dtype: int64
```


lues, thus we can't drop or add mode value as it will make the data skewed.

- So, To tackle this, we will replace NaN with 'Unknown country'.

In [59]:
```
df['country']=df['country'].fillna('Unknown country')
```

In [60]:
```
df.isnull().sum()
```

Out[60]:
```
type               0
title              0
director        2634
cast             825
country            0
date_added         0
release_year       0
certification      0
duration           0
genre              0
dtype: int64
```

Column - cast

In [61]:
```
df['cast'].value_counts()
```

Out[61]:
```
David Attenborough                                                                                                                                    19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil                                                               14
Samuel West                                                                                                                                           10
Jeff Dunham                                                                                                                                            7
David Spade, London Hughes, Fortune Feimster                                                                                                           6
                                                                                                                                                      ..
Michael Peña, Diego Luna, Tenoch Huerta, Joaquin Cosio, José María Yazpik, Matt Letscher, Alyssa Diaz                                                  1
Nick Lachey, Vanessa Lachey                                                                                                                            1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayashi, Ken Yasuda, Arata Furuta, Suzuki Matsuo, Koichi Yamadera, Arata Iura, Chikako Kaku, Kotaro Yoshida    1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Agu, Dele Odule, Femi Adebayo, Bayray McNwizu, Biodun Stephen       1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranjan Tripathy                  1
Name: cast, Length: 7692, dtype: int64
```

- As, there are almost ~10% missing values, thus we can't drop or add mode value as it will make the data skewed.
- So, To tackle this, we will replace NaN with 'Unknown cast'.

In [62]:
```
df['cast']=df['cast'].fillna('Unknown cast')
```

In [63]:
```
df.isnull().sum()
```

Out[63]:
```
type               0
title              0
director        2634
cast               0
country            0
date_added         0
release_year       0
certification      0
duration           0
genre              0
dtype: int64
```

Column - director

In [64]:
```
df['director'].isnull().sum()
```

```
2634
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

- As, there are almost ~30% missing values, thus we can't drop or add mode value as it will make the data skewed.
- So, To tackle this, we will replace NaN with 'Unknown director'.

In [65]:
```
df['director']=df['director'].fillna('Unknown Director')
```

In [66]:
```
df.isnull().sum()
```

Out[66]:
```
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
certification   0
duration        0
genre           0
dtype: int64
```

In [67]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8807 entries, 0 to 8806
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   type           8807 non-null   category
 1   title          8807 non-null   object
 2   director       8807 non-null   object
 3   cast           8807 non-null   object
 4   country        8807 non-null   object
 5   date_added     8807 non-null   datetime64[ns]
 6   release_year   8807 non-null   int64
 7   certification  8807 non-null   object
 8   duration       8807 non-null   object
 9   genre          8807 non-null   object
dtypes: category(1), datetime64[ns](1), int64(1), object(7)
memory usage: 954.8+ KB
```

6. Fixing Duration column

In [68]:
```
df.head()
```

Out[68]:

	type	title	director	cast	country	date_added	release_year	certification	duration	genre
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	Unknown Director	Unknown cast	Unknown country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

For analysis of duration, we are keeping a separate copy of df called duration.

In [69]:
```
duration=df.copy(deep=True)
```

In [70]:
```
duration_movie=duration[duration['type']=='Movie'].copy(deep=True)
```

In [71]:
```
duration_movie['duration']=duration_movie['duration'].str[:-4]
```

In [72]:
```
duration_movie.head()
```

Out[72]:

	type	title	director	cast	country	date_added	release_year	certification	duration	genre
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	2021-09-25	2020	PG-13	90	Documentaries
6	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Unknown country	2021-09-24	2021	PG	91	Children & Family Movies
7	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	2021-09-24	1993	TV-MA	125	Dramas, Independent Movies, International Movies
9	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	2021-09-24	2021	PG-13	104	Comedies, Dramas
12	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	2021-09-23	2021	TV-MA	127	Dramas, International Movies

In [73]:
```
duration_movie['duration']=duration_movie['duration'].astype('int')
```

Creation similar dataframe for TV Shows

In [74]:
```
duration_tv=duration[duration['type']=='TV Show'].copy(deep=True)
```

In [75]:
```
del(duration)
```

In [76]:
```
duration_tv.duration.value_counts()
```

Out[76]:
```
1 Season      1793
2 Seasons      425
3 Seasons      199
4 Seasons       95
5 Seasons       65
6 Seasons       33
7 Seasons       23
8 Seasons       17
9 Seasons        9
10 Seasons       7
13 Seasons       3
15 Seasons       2
12 Seasons       2
11 Seasons       2
17 Seasons       1
Name: duration, dtype: int64
```

In [77]:
```
duration_tv['duration'].str[:-7].value_counts()
```

Out[77]:
```
1      1793
2       425
3       199
4        95
5        65
6        33
7        23
8        17
9         9
10        7
13        3
15        2
12        2
11        2
17        1
Name: duration, dtype: int64
```

In [78]:
```
duration_tv['duration']=duration_tv['duration'].str[:-7]
```

In [79]:
```
duration_tv.rename({'duration':'seasons'},axis=1,inplace=True)
```

In [80]:
```
duration_tv['seasons']=duration_tv['seasons'].astype('int')
```

In [81]:
```
duration_tv.head()
```

Out[81]:

	type	title	director	cast	country	date_added	release_year	certification	seasons	genre
1	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021	TV-MA	1	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	Unknown Director	Unknown cast	Unknown country	2021-09-24	2021	TV-MA	1	Docuseries, Reality TV
4	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2	International TV Shows, Romantic TV Shows, TV ...
5	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	Unknown country	2021-09-24	2021	TV-MA	1	TV Dramas, TV Horror, TV Mysteries

> For the analysis of duration, we would be using the **duration_movie** & **duration_tv**.

7. Unnesting the director, cast, country and genre column

In [82]:
```
df.head()
```

Out[82]:

	type	title	director	cast	country	date_added	release_year	certification	duration	genre
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	Unknown Director	Unknown cast	Unknown country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

> Creating Unnested **director** table.

In [83]:
```
direct=df[['title','director']].copy(deep=True)
```

In [84]:
```
direct['director']=direct['director'].str.split(', ')
```

In [85]:
```
direct=direct.explode('director',ignore_index=True)
```

In [86]:
```
direct['director'].value_counts()
```

Out[86]:
```
Unknown Director    2634
Rajiv Chilaka         22
Jan Suter             21
Raúl Campos           19
Suhas Kadav           16
                    ...
Raymie Muzquiz         1
Stu Livingston         1
Joe Menendez           1
Eric Bross             1
Mozez Singh            1
Name: director, Length: 4994, dtype: int64
```

> Creating Unnested **cast** table.

In [87]:
```
cast=df[['title','cast']].copy(deep=True)

cast['cast']=cast['cast'].str.split(', ')

cast=cast.explode('cast',ignore_index=True)

cast['cast'].value_counts()
```

Out[87]:
```
Unknown cast               825
Anupam Kher                 43
Shah Rukh Khan              35
Julie Tejwani               33
Naseeruddin Shah            32
                          ...
Melanie Straub               1
Gabriela Maria Schmeide      1
Helena Zengel                1
Daniel Valenzuela            1
Chittaranjan Tripathy        1
Name: cast, Length: 36440, dtype: int64
```

> Creating unnested **country** column

In [88]:
```
cntry=df[['title','country']].copy(deep=True)

cntry['country']=cntry['country'].str.split(', ')

cntry=cntry.explode('country',ignore_index=True)

cntry['country'].value_counts()
```

Out[88]:
```
United States      3689
India              1046
Unknown country     831
United Kingdom      804
Canada              445
                   ...
Bermuda               1
Ecuador               1
Armenia               1
Mongolia              1
Montenegro            1
Name: country, Length: 128, dtype: int64
```

> Creating unnested **genre** column

In [89]:
```
genre=df[['title','genre']].copy(deep=True)

genre['genre']=genre['genre'].str.split(', ')

genre=genre.explode('genre',ignore_index=True)

genre['genre'].value_counts()
```

Out[89]:
```
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
TV Comedies                     581
Thrillers                       577
Crime TV Shows                  470
Kids' TV                        451
Docuseries                      395
Music & Musicals                375
Romantic TV Shows               370
Horror Movies                   357
Stand-Up Comedy                 343
Reality TV                      255
British TV Shows                253
Sci-Fi & Fantasy                243
Sports Movies                   219
Anime Series                    176
Spanish-Language TV Shows       174
TV Action & Adventure           168
Korean TV Shows                 151
Classic Movies                  116
LGBTQ Movies                    102
TV Mysteries                     98
Science & Nature TV              92
TV Sci-Fi & Fantasy              84
TV Horror                        75
Anime Features                   71
Cult Movies                      71
Teen TV Shows                    69
Faith & Spirituality             65
TV Thrillers                     57
Movies                           57
Stand-Up Comedy & Talk Shows     56
Classic & Cult TV                28
TV Shows                         16
Name: genre, dtype: int64
```

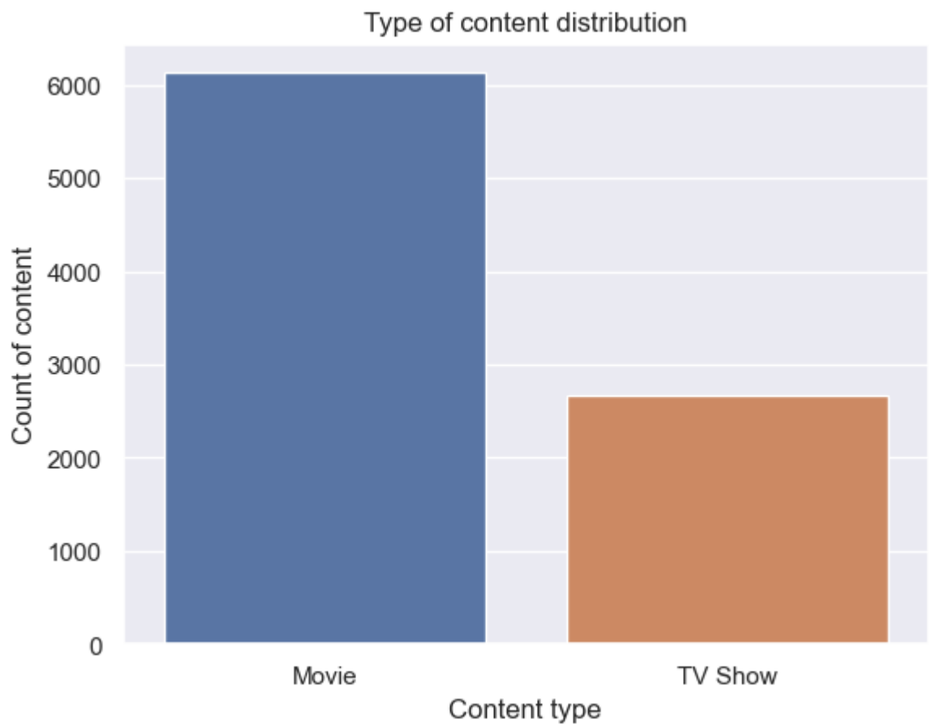
Univariate Analysis

In [90]:
```
df.head()
```

Out[90]:

	type	title	director	cast	country	date_added	release_year	certification	duration	genre
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	TV Show	Blood & Water	Unknown Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
3	TV Show	Jailbirds New Orleans	Unknown Director	Unknown cast	Unknown country	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
4	TV Show	Kota Factory	Unknown Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...

In [91]:
```
sns.set_theme(style="darkgrid")
sns.countplot(data=df,x='type')
plt.title('Type of content distribution')
plt.xlabel('Content type')
plt.ylabel('Count of content')
plt.show()
```

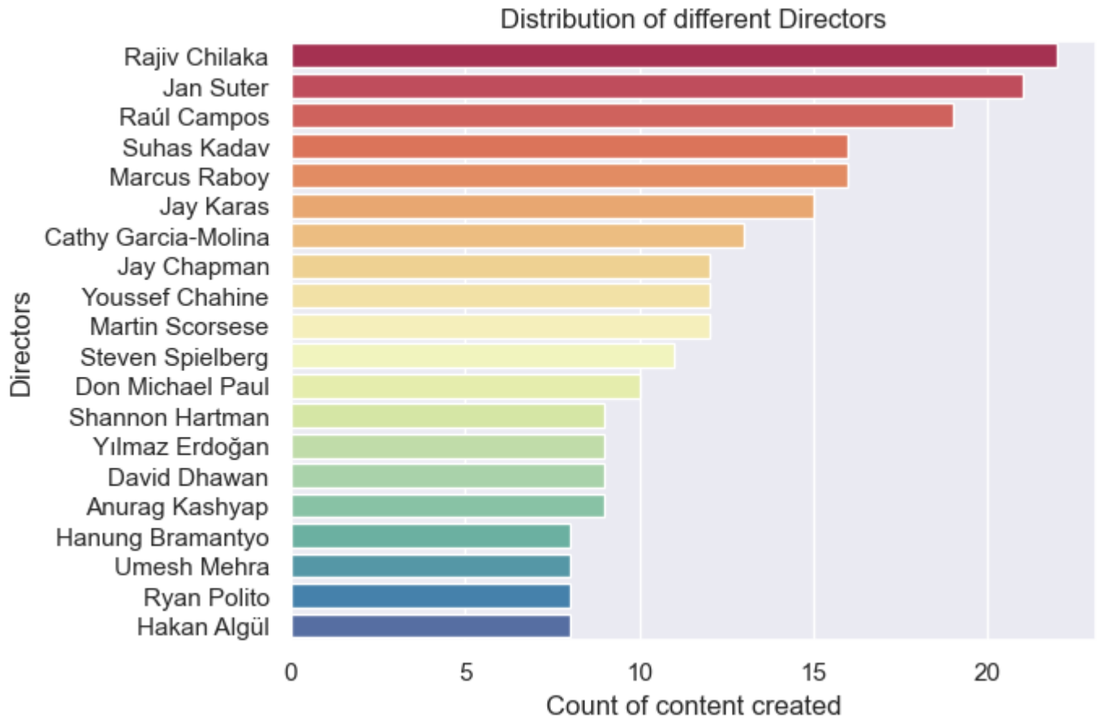


- **Insights** :

  > Majorly, movies are produced.

Column : **director**

In [92]:
```
directors=direct.loc[direct['director']!='Unknown Director','director'].value_counts().to_frame().head(20)
sns.barplot(data=directors,y=directors.index, x='director',palette='Spectral')
plt.xlabel('Count of content created')
plt.ylabel('Directors')
plt.title('Distribution of different Directors')
plt.show()
```

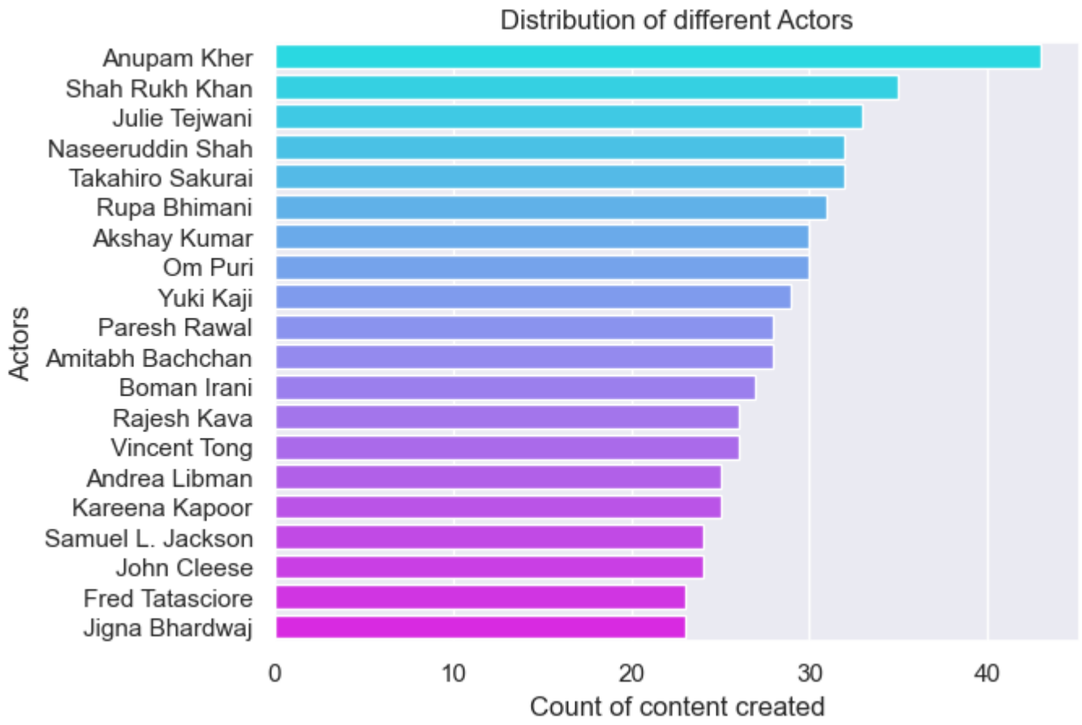


- **Insights** :

> These are the Top 20 directors who made most movies/TV Shows. **Rajiv Chilaka** is the director with the most movies.

Column : **cast**

In [93]:
```
casts=cast.loc[cast['cast']!='Unknown cast','cast'].value_counts().to_frame().head(20)
sns.barplot(data=casts,y=casts.index, x='cast',palette='cool')
plt.xlabel('Count of content created')
plt.ylabel('Actors')
plt.title('Distribution of different Actors')
plt.show()
```

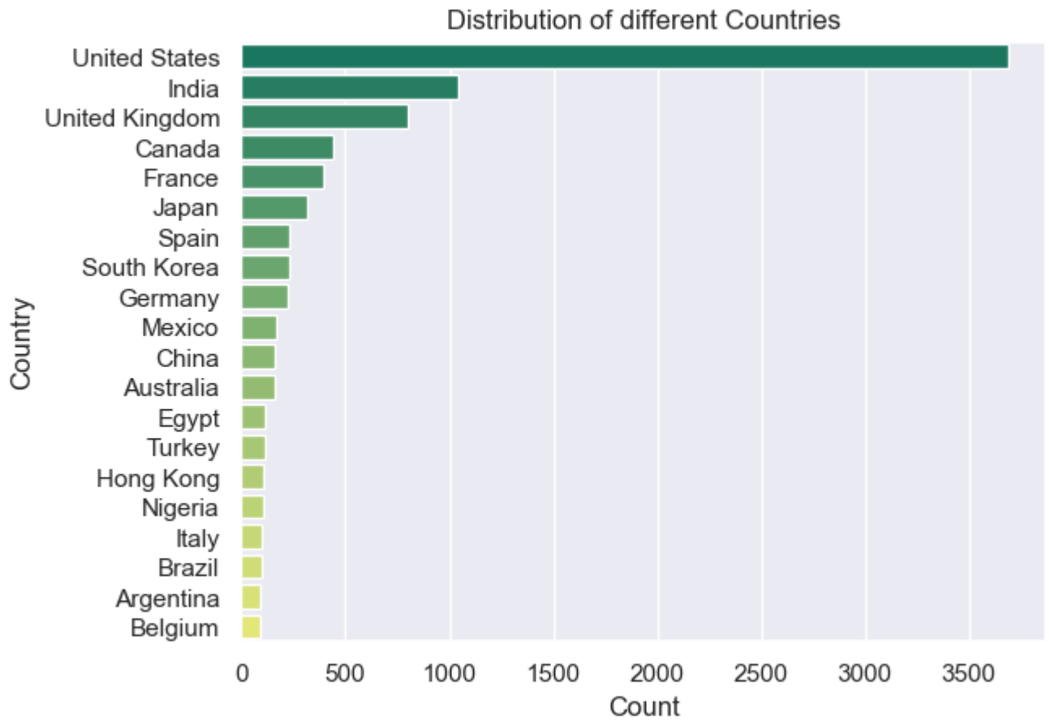


- **Insights** :

> These are the Top 20 actors who w orked on most movies/TV Shows in the dataset. **Anupam Kher** is the actor with the most movies.

Column : **country**

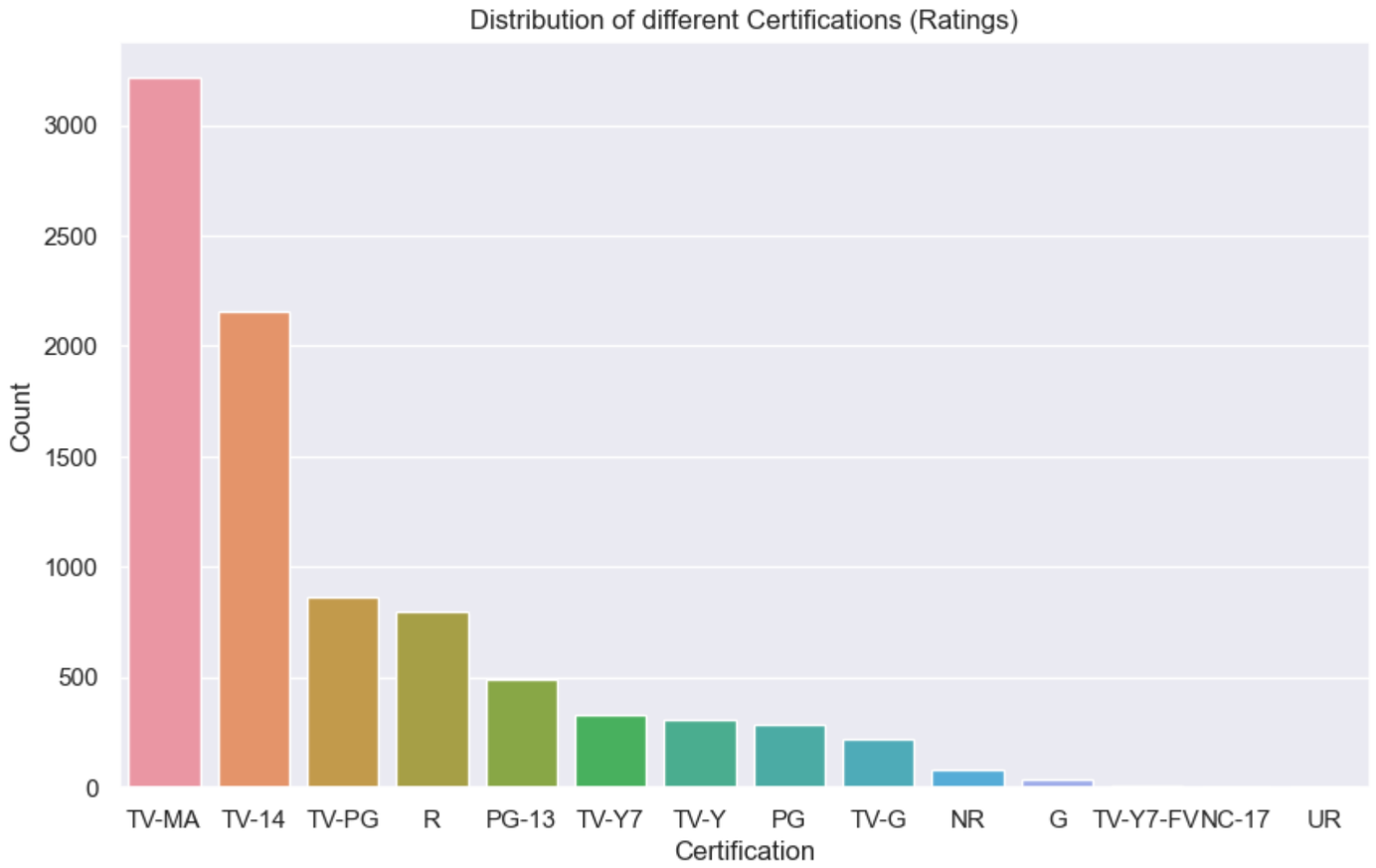
In [94]:
```
countries=cntry.loc[cntry['country']!='Unknown country','country'].value_counts().to_frame().head(20)
sns.barplot(data=countries,y=countries.index, x='country',palette='summer')
plt.xlabel('Count')
plt.ylabel('Country')
plt.title('Distribution of different Countries')
plt.show()
```



- **Insights** :
  - The coutry-wise distribution of the data.
  - **United States** has produced most of the movies.
  - Top 5 countries in terms of producing content: - **US, India, UK, Canada, France**

Column : **Certification**

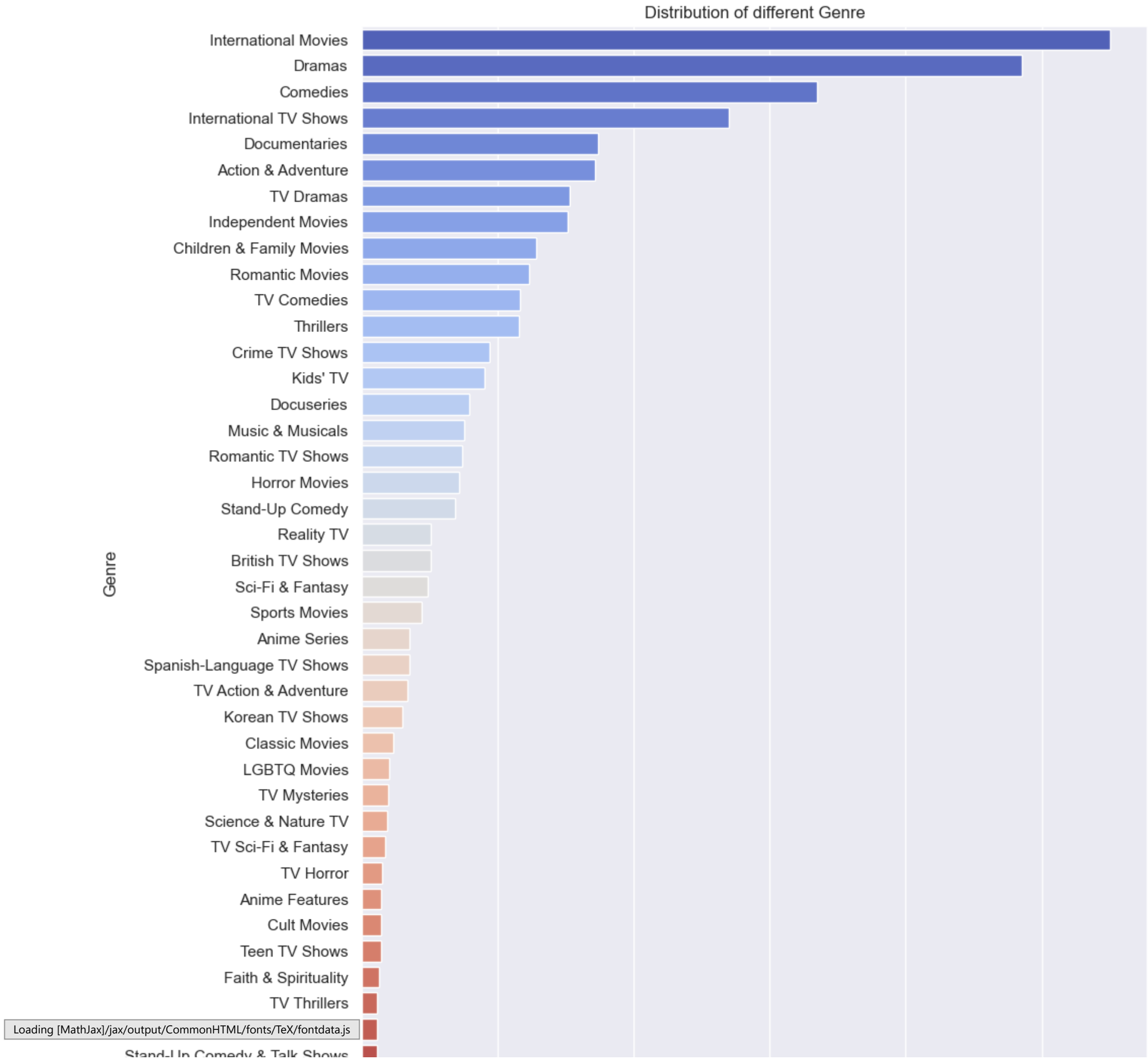
In [95]:
```
plt.figure(figsize=(10,6))
order=df['certification'].value_counts()
sns.countplot(data=df,x='certification',order=order.index)
plt.ylabel('Count')
plt.xlabel('Certification')
plt.title('Distribution of different Certifications (Ratings)')
plt.show()
```

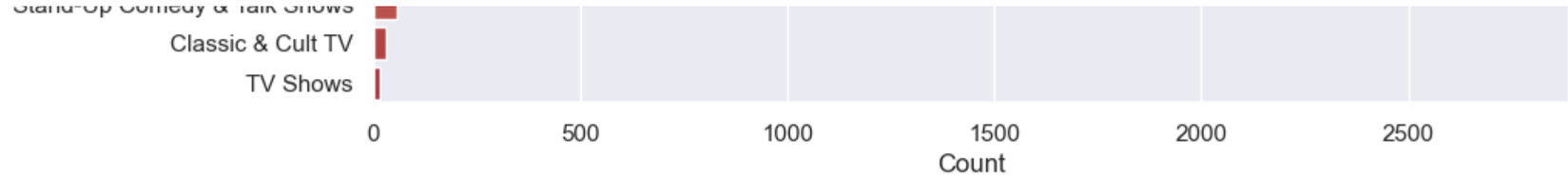


- **Insights** :
  - The certificaiton wise distribution of the data.
  - **TV-MA** certification type movies have been produced the most.

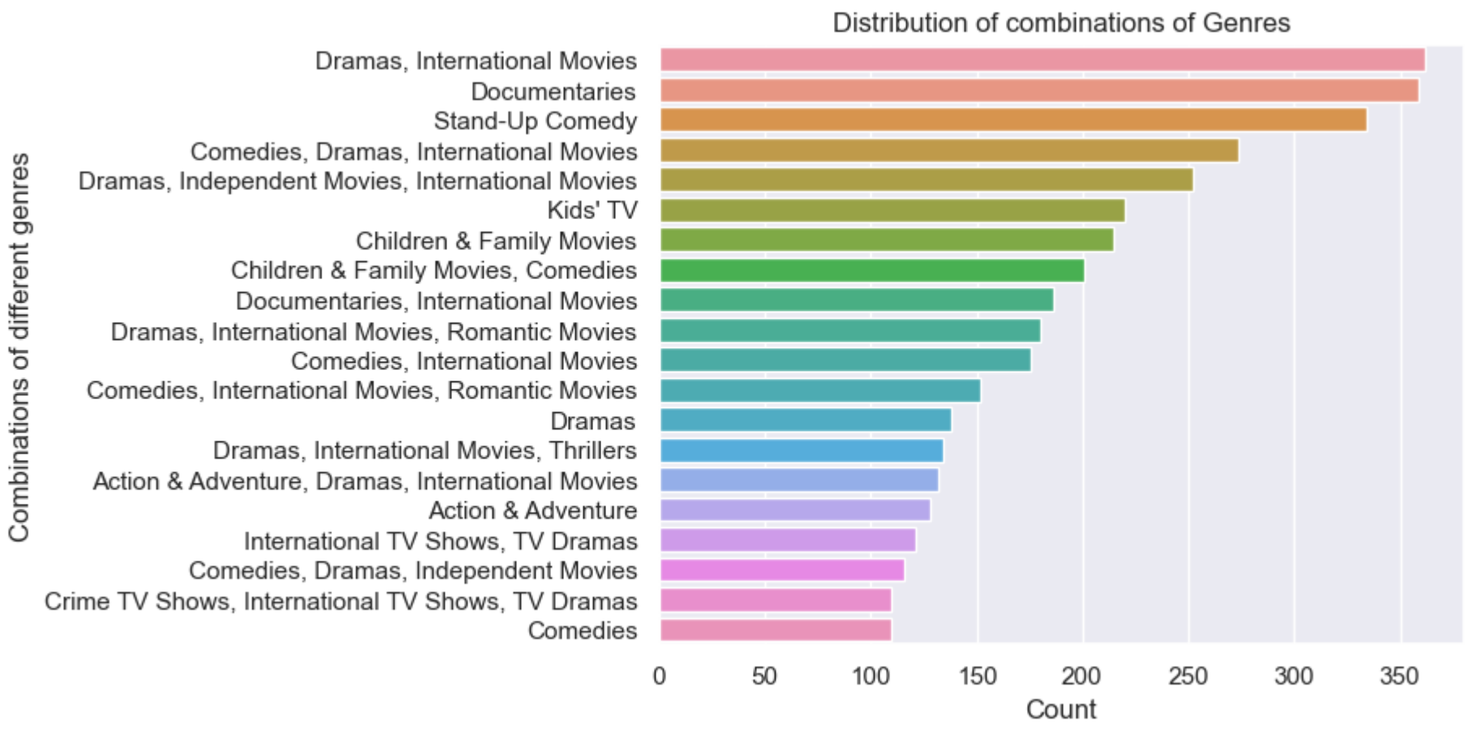
Column : **genre**

In [96]:
```
plt.figure(figsize=(10,14))
genres=genre['genre'].value_counts().to_frame().reset_index()
sns.barplot(data=genres,y='index', x='genre',palette='coolwarm')
plt.xlabel('Count')
plt.ylabel('Genre')
plt.title('Distribution of different Genre')
plt.show()
```





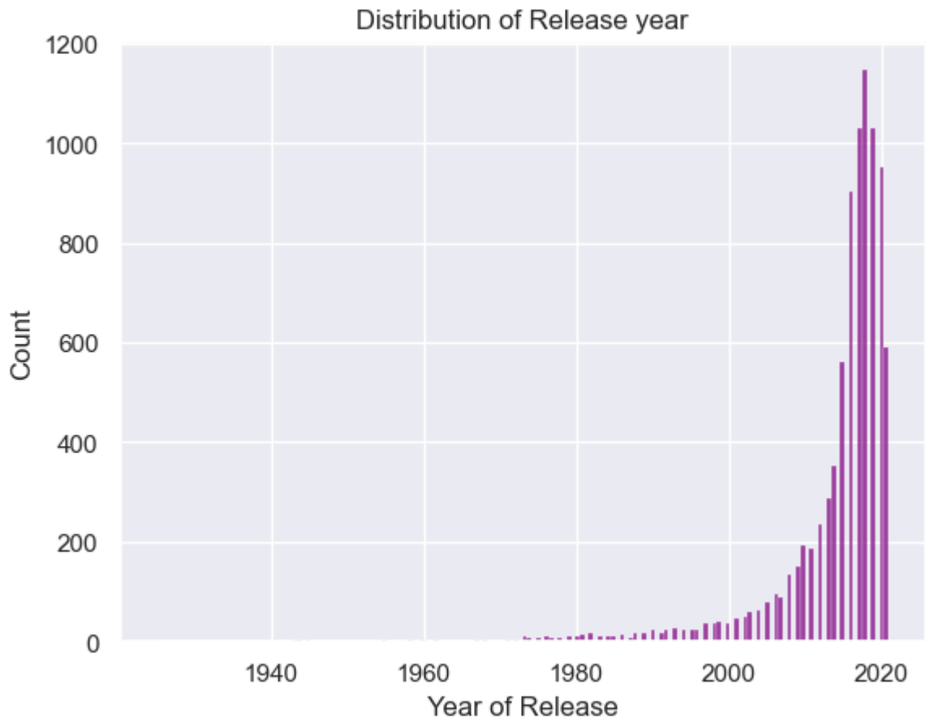
In [97]:
```
gen=df['genre'].value_counts().reset_index().head(20)
sns.barplot(data=gen,y='index',x='genre')
plt.ylabel('Combinations of different genres')
plt.xlabel('Count')
plt.title('Distribution of combinations of Genres')
plt.show()
```



- **Insights** :
  - The genre-wise distribution of the data.
  - **International movies** are the most produced movies.
  - Top 5 countries in terms of producing content: - **International movies, Dramas, Comedies, International TV shows, Documentries**
  - Most preferred Combination of genres - **International Dramas**

Column : **Release year**

In [98]:
```
sns.histplot(data=df,x='release_year',color='purple')
plt.ylabel('Count')
plt.xlabel('Year of Release')
plt.title('Distribution of Release year')
plt.show()
```

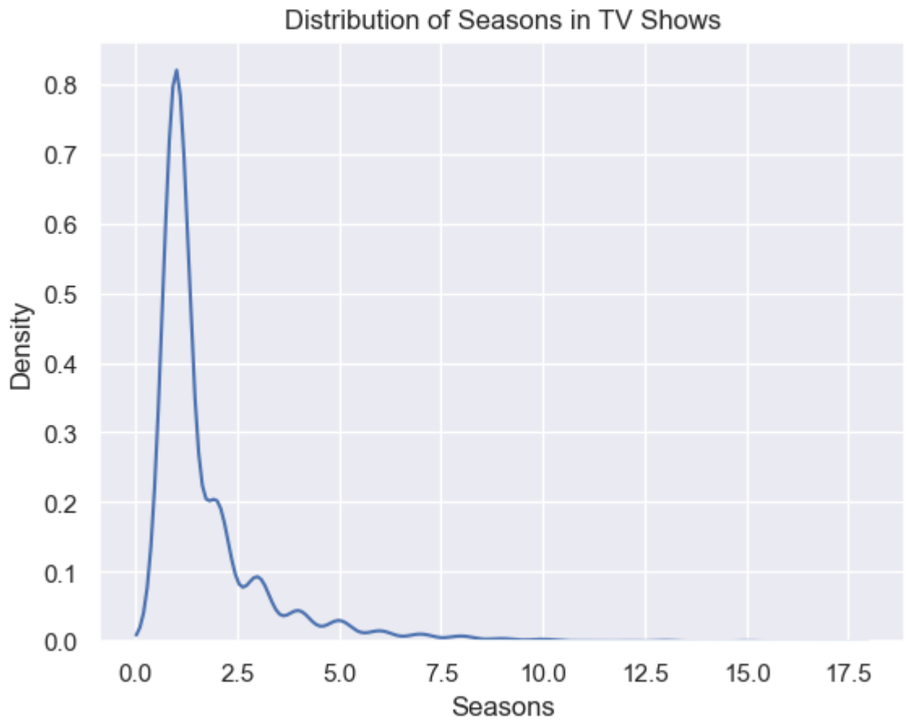


- **Insights** :
  - The release_year distribution of the data.
  - Most movies are of 2000 or later in the dataset.

Duration distribution

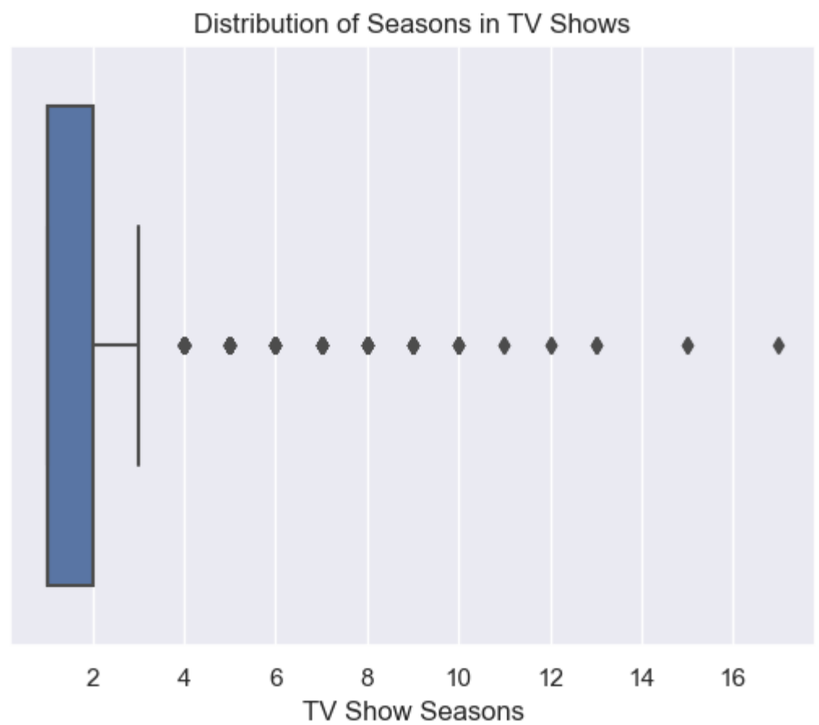
- **TV Shows**

In [99]:
```
sns.kdeplot(data=duration_tv,x='seasons')
plt.xlabel('Seasons')
plt.title('Distribution of Seasons in TV Shows')
plt.show()
```



In [100…

```
sns.boxplot(data=duration_tv,x='seasons')
plt.xlabel('TV Show Seasons')
plt.title('Distribution of Seasons in TV Shows')
plt.show()
```

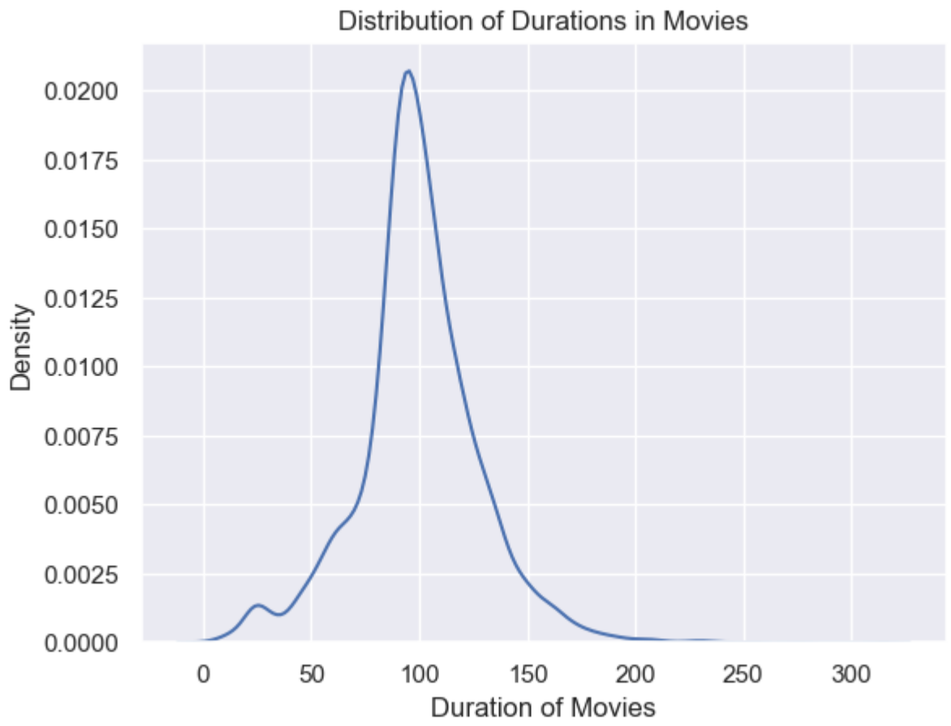


- **Insights** :
  - Most of the TV shows are having <= 2 Seasons.

- **Movies**

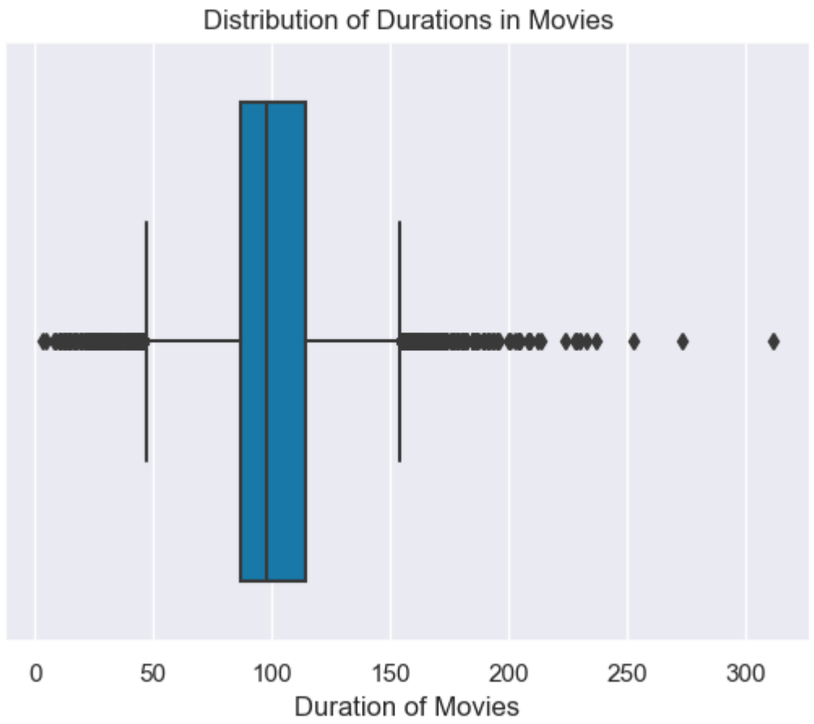
In [101…

```
sns.kdeplot(data=duration_movie,x='duration')
plt.xlabel('Duration of Movies')
plt.title('Distribution of Durations in Movies')
plt.show()
```



In [102…

```
sns.boxplot(data=duration_movie,x='duration',palette='winter')
plt.xlabel('Duration of Movies')
plt.title('Distribution of Durations in Movies')
plt.show()
```

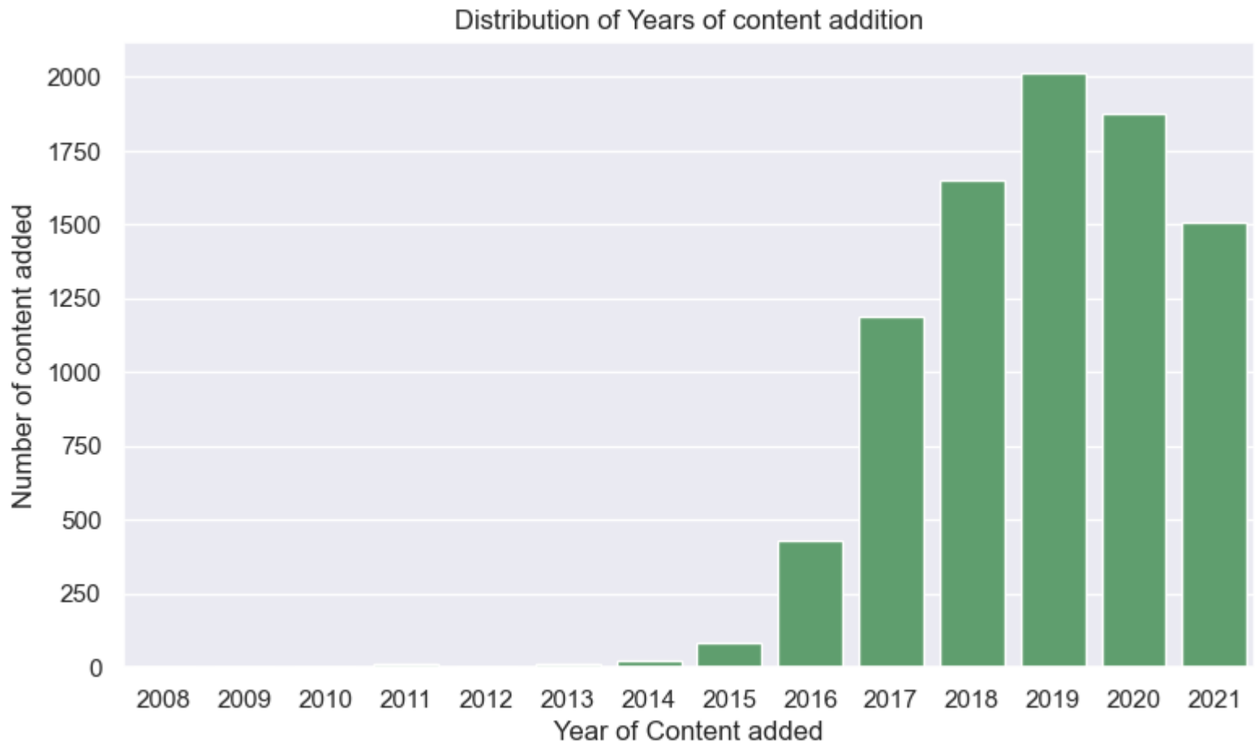


- **Insights** :
  - Most of the movies are in the range of duration: - 40 min to 160 min with median of **95 min**.

Column : **date_added**

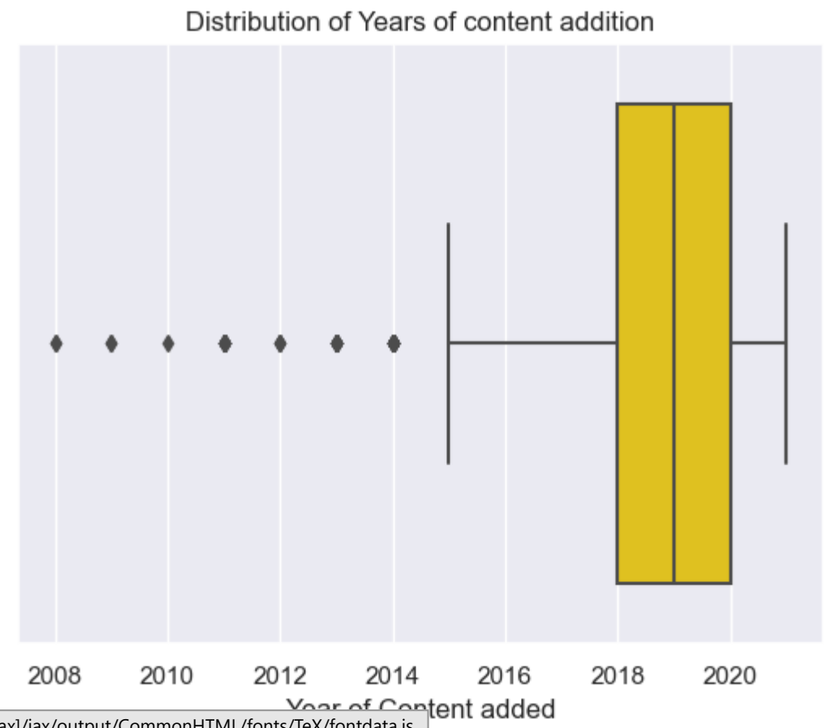
In [103…

```
# Yearly distribution of the content added
plt.figure(figsize=(9,5))
sns.countplot(data=df,x=df['date_added'].dt.year,color='g')
plt.ylabel('Number of content added')
plt.xlabel('Year of Content added')
plt.title('Distribution of Years of content addition')
plt.show()
```



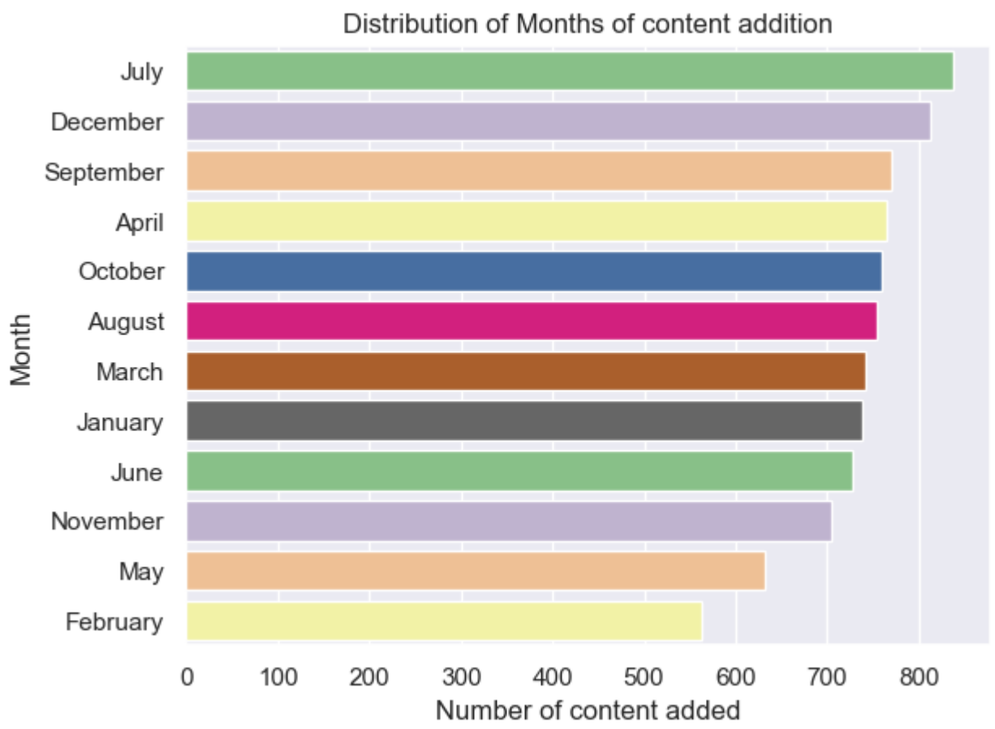
In [104…

```
sns.boxplot(data=df,x=df['date_added'].dt.year,palette='prism')
plt.xlabel('Year of Content added')
plt.title('Distribution of Years of content addition')
plt.show()
```



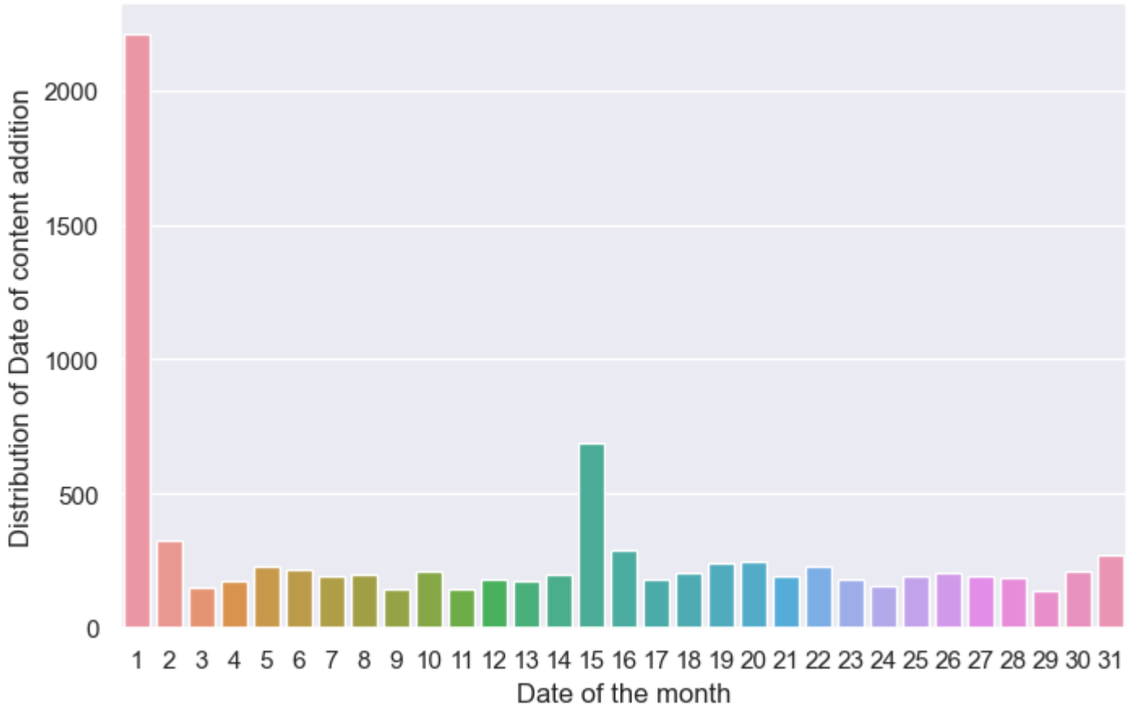
- **Insights** :
  - Most of the content added is from **2014- 2021**, median of which is **2019**.

In [105…
```
#Monthly distribution of the date, the content is added.
mnt=df['date_added'].dt.month_name().value_counts().to_frame().reset_index()
sns.barplot(data=mnt,y='index',x='date_added',palette='Accent')
plt.ylabel('Month')
plt.xlabel('Number of content added')
plt.title('Distribution of Months of content addition')
plt.show()
```



- **Insights** :
  - Most content is added in **July** and **December**.
  - Least content is added in **February** and **May**.

In [106…
```
#Date-wise distribution of the content added.
plt.figure(figsize=(8,5))
sns.countplot(data=df,x=df['date_added'].dt.day)
plt.xlabel('Date of the month')
plt.ylabel('Distribution of Date of content addition')
plt.title('')
plt.show()
```



- **Insights** :
  - Most of the content added on platform on **1st of the Month**.

Bivariate analysis

Pair-plot

In [107…

```
df_pair=df[['type','date_added','release_year']].copy(deep=True)
df_pair['month']=df['date_added'].dt.month
df_pair['date']=df['date_added'].dt.day
df_pair['year']=df['date_added'].dt.year
df_pair.drop('date_added',axis=1,inplace=True)
df_pair
```

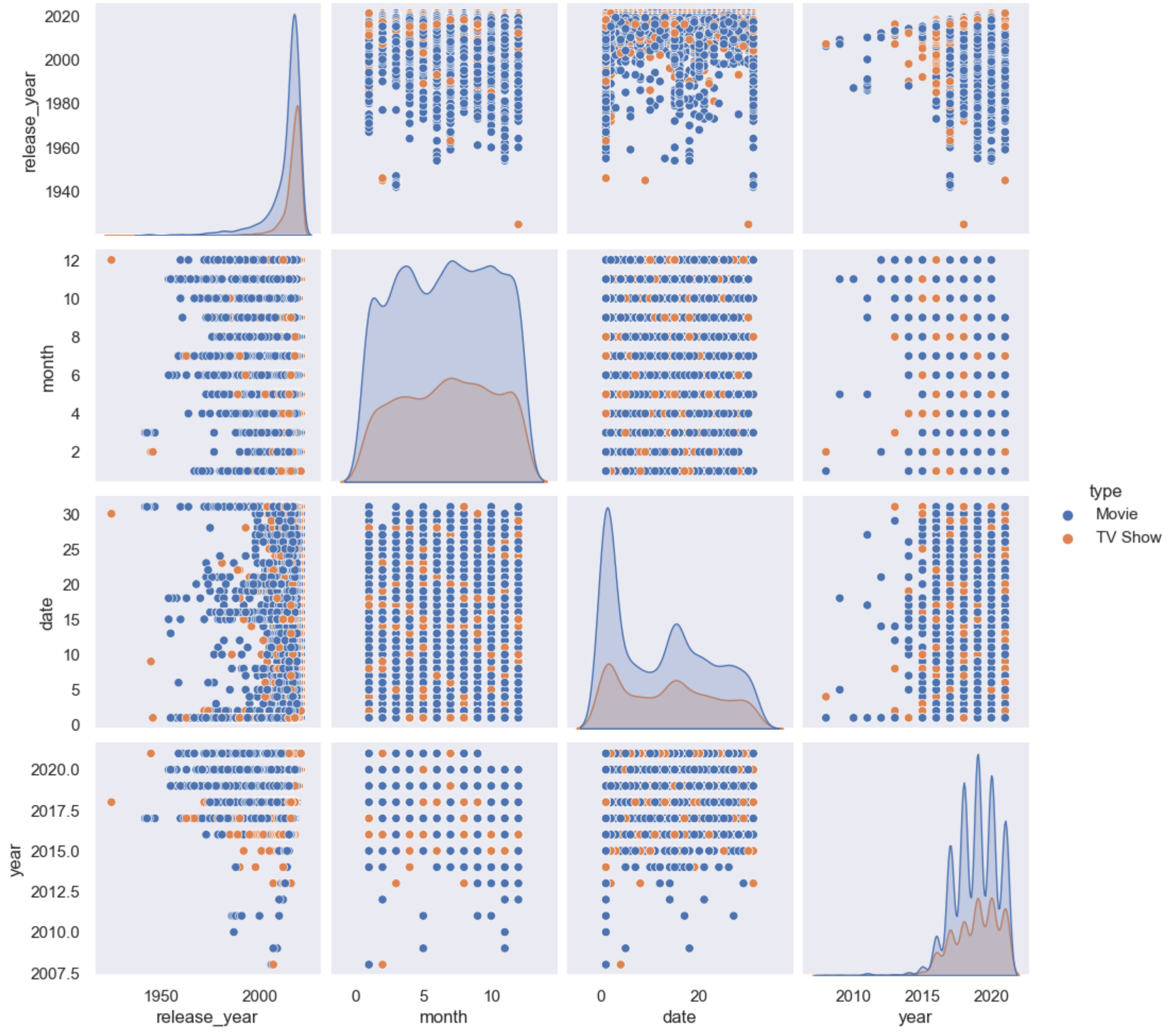
Out[107]:

	type	release_year	month	date	year
0	Movie	2020	9	25	2021
1	TV Show	2021	9	24	2021
2	TV Show	2021	9	24	2021
3	TV Show	2021	9	24	2021
4	TV Show	2021	9	24	2021
...	...	...	...	...	...
8802	Movie	2007	11	20	2019
8803	TV Show	2018	7	1	2019
8804	Movie	2009	11	1	2019
8805	Movie	2006	1	11	2020
8806	Movie	2015	3	2	2019

8807 rows × 5 columns

In [108…

```
sns.set_theme(style="dark")
sns.pairplot(df_pair,hue='type')
plt.show()
```



- Most of the content has been added post 2015.
- Almost >50% movies are added on 1st day of the month.

Country wise Analysis

- We will be analysing for top 3 Countries i.e., US, India, and UK

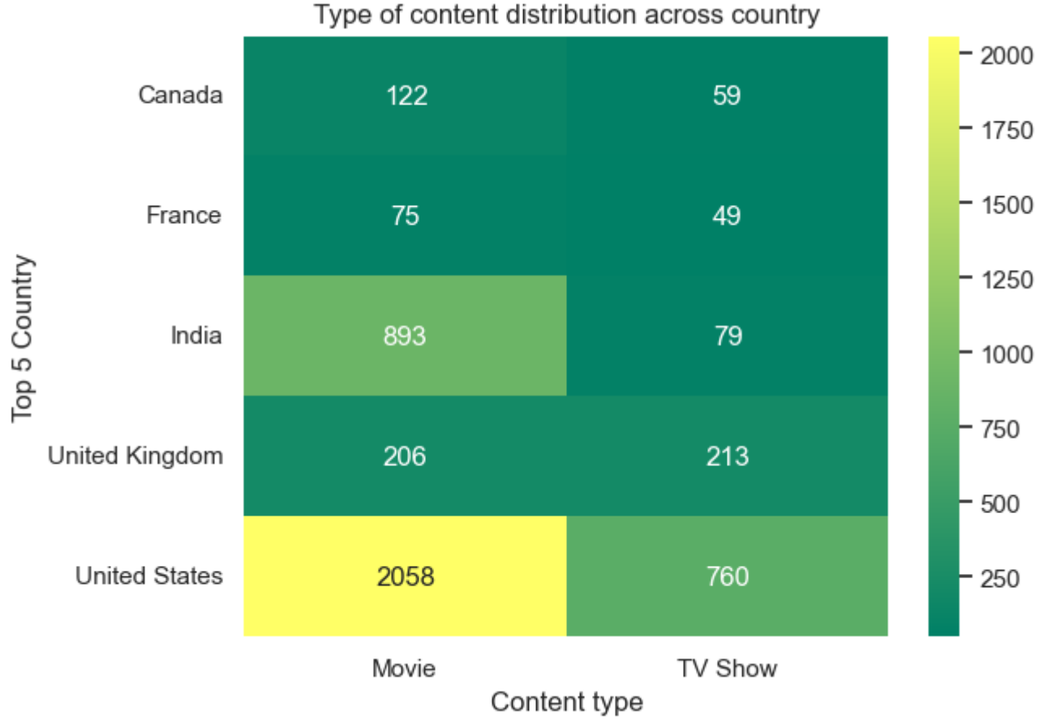
a) Country wise type of content analysis

In [109…
```
top5_cntry=cntry.loc[cntry['country']!='Unknown country','country'].value_counts().iloc[:5].index
```

In [110…
```
cntry_type=cntry.merge(df,how='inner')
cntry_type=cntry_type[cntry_type['country']!='Unknown country']
cntry_type=cntry_type[cntry_type['country'].isin(top5_cntry)==True]
cntry_type=cntry_type.groupby(['country','type']).size().reset_index(name='count')
```

In [111…
```
cntry_type=cntry_type.pivot(index='country',columns='type',values='count')
```

In [112…
```
sns.set_theme(style="darkgrid")
sns.heatmap(data=cntry_type,cmap='summer',fmt='d',annot=True)
plt.title('Type of content distribution across country')
plt.xlabel('Content type')
plt.ylabel('Top 5 Country')
plt.show()
```



Insights: -

- In **India**, Movies are on top.
- In **UK**, TV Shows are on top.

b). Country-wise genre distribution

In [113…

```
top5_cntry=cntry.loc[cntry['country']!='Unknown country','country'].value_counts().iloc[:5].index
top20_genre=genre['genre'].value_counts().iloc[:20].index
top5_genre=genre['genre'].value_counts().iloc[:5].index #For future use
print(top5_cntry,top10_genre)
```

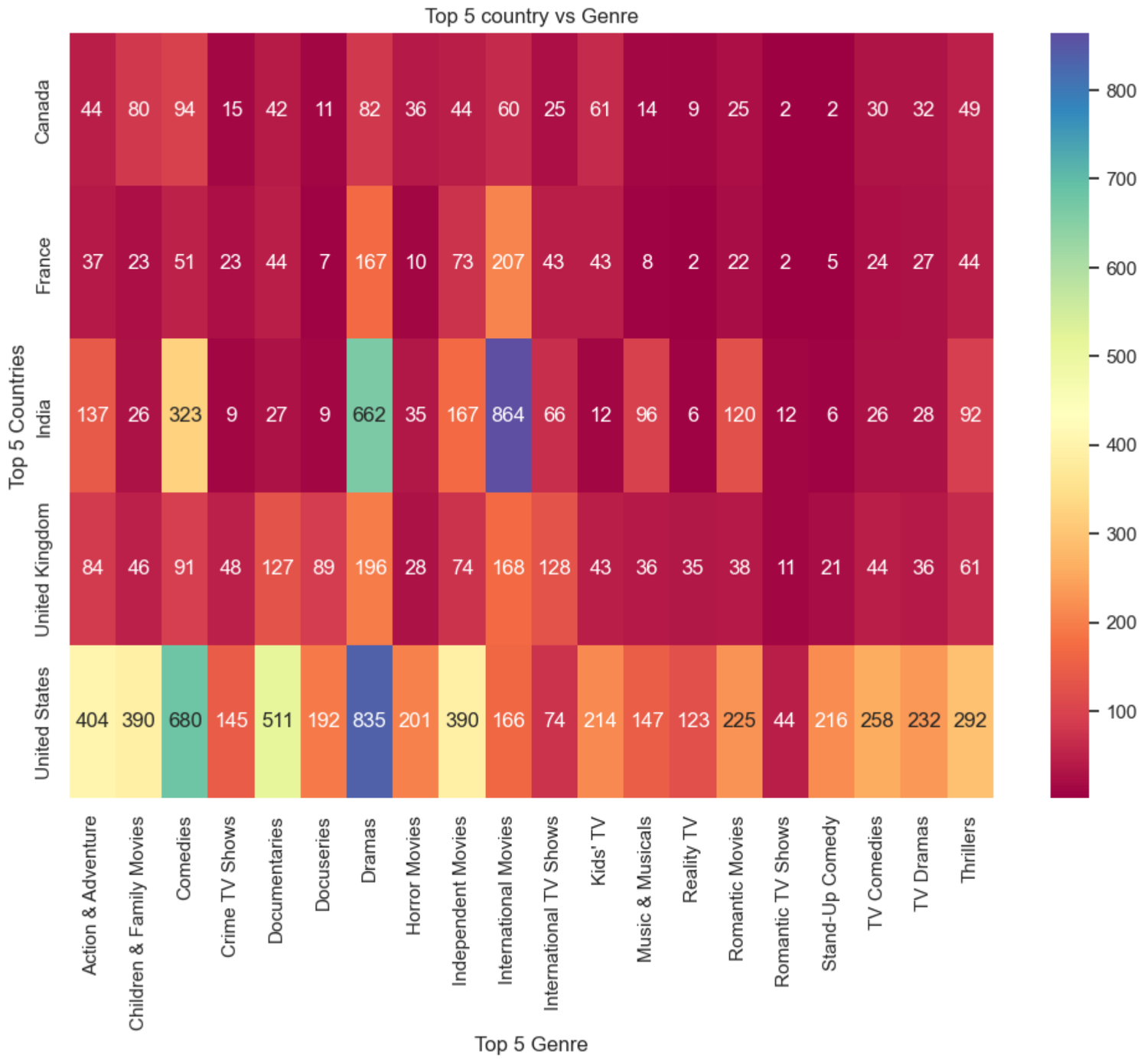
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_9020\545445382.py in <module>
      2 top20_genre=genre['genre'].value_counts().iloc[:20].index
      3 top5_genre=genre['genre'].value_counts().iloc[:5].index #For future use
----> 4 print(top5_cntry,top10_genre)

NameError: name 'top10_genre' is not defined
```

In [114…

```
cntry_genre=cntry.merge(genre,how='inner')
cntry_genre=cntry_genre.groupby(['country','genre']).size().reset_index(name='count')
cntry_genre=cntry_genre[cntry_genre['country'].isin(top5_cntry)==True]
cntry_genre=cntry_genre[cntry_genre['genre'].isin(top20_genre)==True]
cntry_genre_pivot=cntry_genre.pivot(index='country',columns='genre',values='count')
```

In [115…

```
plt.figure(figsize=(12,8))
sns.heatmap(cntry_genre_pivot,cmap='Spectral',annot=True,fmt='d')
plt.title('Top 5 country vs Genre')
plt.xlabel('Top 5 Genre')
plt.ylabel('Top 5 Countries')
plt.yticks(rotation = 90)
plt.show()
```



Insights:

- In USA, top genres are: - **DRAMAS**, **Comedies**, **Documentries**.
- In India, top genres are: - **INTERNATIONAL MOVIES**, **Dramas**, **Comedies**
- In UK, top genres are: - **DRAMAS**, **International movies**, **International TV shows** and **Documenteries**

stribution

In [116…
```
top25_direct=direct.loc[direct['director']!='Unknown Director','director'].value_counts().iloc[:25].index
top25_direct
```

Out[116]:
```
Index(['Rajiv Chilaka', 'Jan Suter', 'Raúl Campos', 'Suhas Kadav',
       'Marcus Raboy', 'Jay Karas', 'Cathy Garcia-Molina', 'Jay Chapman',
       'Youssef Chahine', 'Martin Scorsese', 'Steven Spielberg',
       'Don Michael Paul', 'Shannon Hartman', 'Yılmaz Erdoğan', 'David Dhawan',
       'Anurag Kashyap', 'Hanung Bramantyo', 'Umesh Mehra', 'Ryan Polito',
       'Hakan Algül', 'Fernando Ayllón', 'Johnnie To', 'Quentin Tarantino',
       'Troy Miller', 'Justin G. Dyck'],
      dtype='object')
```

In [117…
```
cntry_direct=direct.merge(cntry,how='inner')
cntry_direct=cntry_direct[(cntry_direct['director']!='Unknown Director')&(cntry_direct['country']!='Unknown country')]
cntry_direct=cntry_direct[['director','country']]
cntry_direct=cntry_direct.groupby(['country','director']).size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_direct=cntry_direct[cntry_direct['director'].isin(top25_direct)]
cntry_direct=cntry_direct[cntry_direct['country'].isin(top5_cntry)]
cntry_direct
```

Out[117]:

	country	director	count
5241	United States	Jay Karas	15
5707	United States	Marcus Raboy	15
5753	United States	Martin Scorsese	12
5240	United States	Jay Chapman	12
6394	United States	Steven Spielberg	11
4909	United States	Don Michael Paul	10
6315	United States	Shannon Hartman	9
1777	India	David Dhawan	9
1706	India	Anurag Kashyap	9
529	Canada	Justin G. Dyck	8
6531	United States	Troy Miller	8
2308	India	Umesh Mehra	8
6237	United States	Ryan Polito	8
6071	United States	Quentin Tarantino	7
2092	India	Rajiv Chilaka	5
1315	France	Youssef Chahine	4
2252	India	Steven Spielberg	3
4112	United Kingdom	Martin Scorsese	2
4268	United Kingdom	Steven Spielberg	2
4477	United States	Anurag Kashyap	1
5491	United States	Justin G. Dyck	1
1179	France	Martin Scorsese	1
669	Canada	Steven Spielberg	1
			1

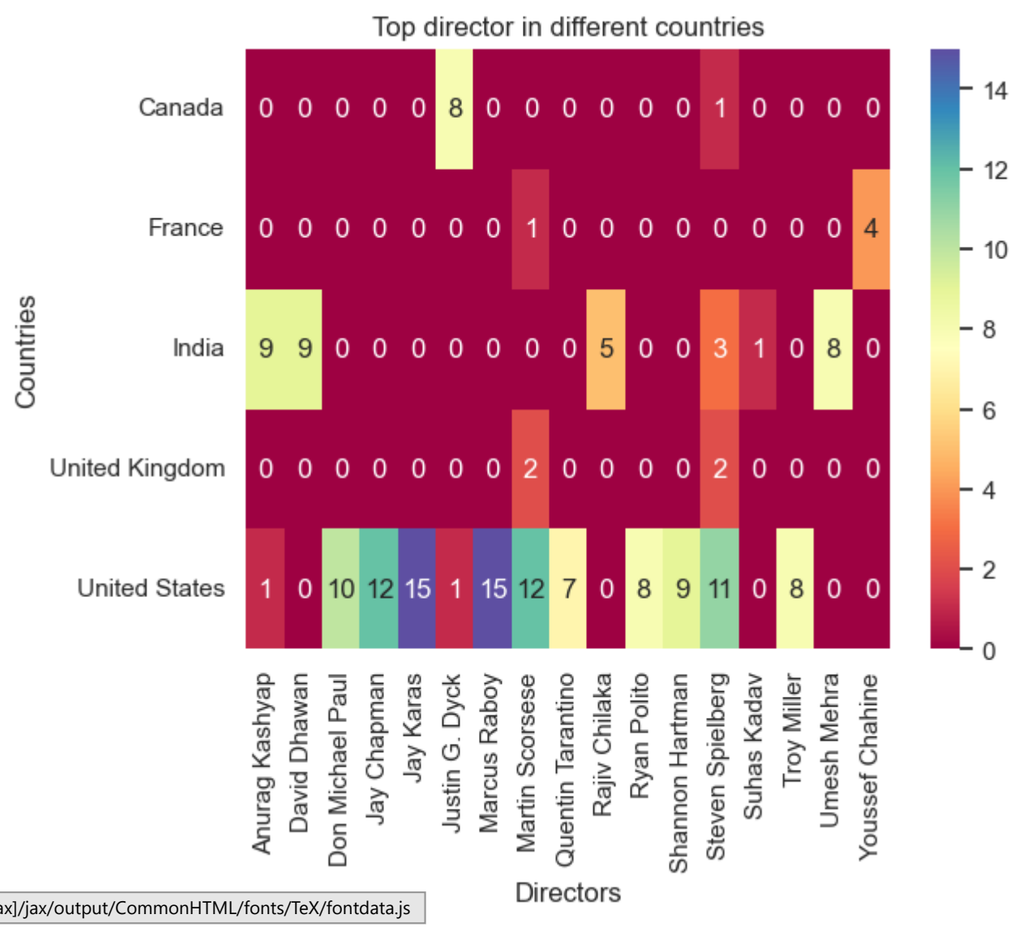
In [118…]
```
cntry_direct=cntry_direct.pivot(index='country',columns='director',values='count')
cntry_direct.fillna(0,inplace=True)
```

In [119…]
```
cntry_direct
```

Out[119]:

director	Anurag Kashyap	David Dhawan	Don Michael Paul	Jay Chapman	Jay Karas	Justin G. Dyck	Marcus Raboy	Martin Scorsese	Quentin Tarantino	Rajiv Chilaka	Ryan Polito	Shannon Hartman	Steven Spielberg	Suhas Kadav	Troy Miller	Umesh Mehra	Youssef Chahine
country																	
Canada	0.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
France	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0
India	9.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	3.0	1.0	0.0	8.0	0.0
United Kingdom	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
United States	1.0	0.0	10.0	12.0	15.0	1.0	15.0	12.0	7.0	0.0	8.0	9.0	11.0	0.0	8.0	0.0	0.0

In [120…]
```
sns.heatmap(cntry_direct, cmap='Spectral',annot=True)
plt.title('Top director in different countries')
plt.xlabel('Directors')
plt.ylabel('Countries')
plt.show()
```



Insights: -

- Top directors in US:- **Marcus Raboy**, **Jay Chapman**, **Martin Scorsese**, **Steven Spilberg**,and **Don Michael paul**
- Top directors in India: - **Anurag Kashyap**, **David Dhawan**,and **Umesh Mehra**
- Top director in Canada: - **Justing G. Dyck**

d). Country-wise actors distribution

In [121…
```
top20_cast=cast.loc[cast['cast']!='Unknown cast','cast'].value_counts().iloc[:20].index
```

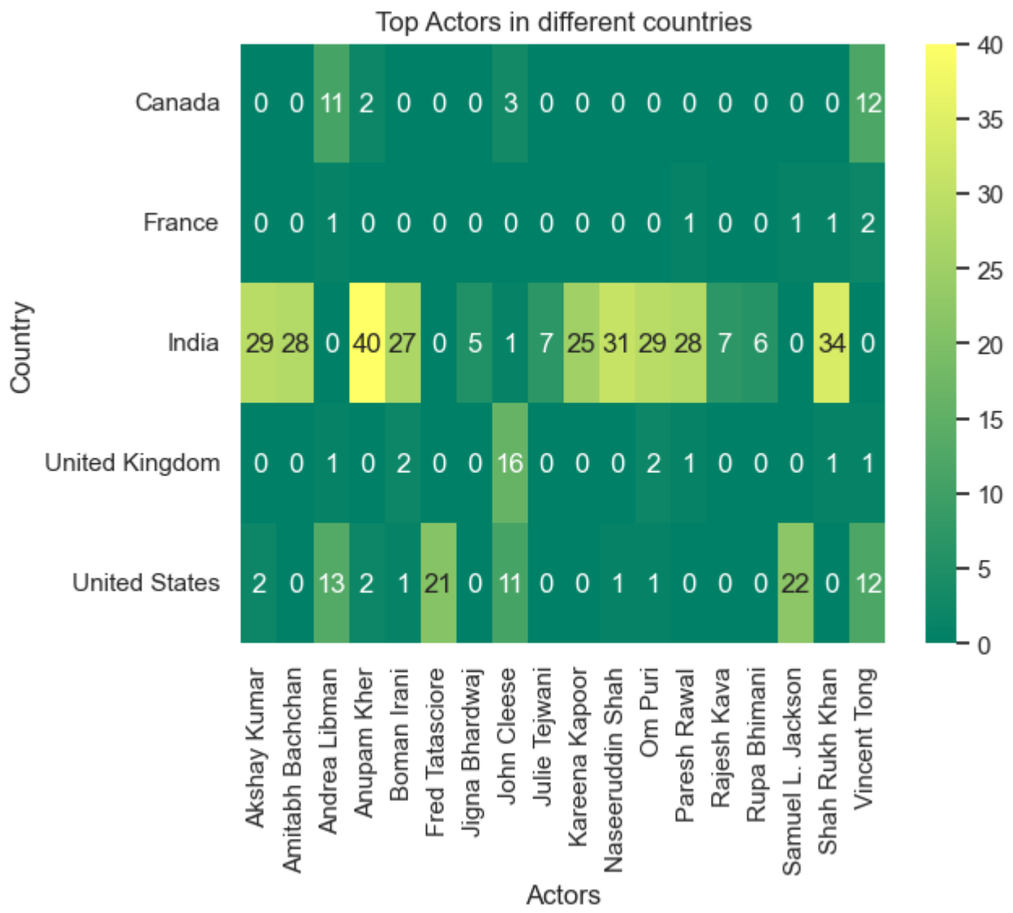
In [122…
```
cntry_cast=cast.merge(cntry,how='inner')
cntry_cast=cntry_cast[(cntry_cast['country']!='Unknown country') & (cntry_cast['cast']!='Unknown cast')]
cntry_cast=cntry_cast.groupby(['country','cast']).size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_cast=cntry_cast[(cntry_cast['cast'].isin(top20_cast)==True) & cntry_cast['country'].isin(top5_cntry)]
cntry_cast.head()
```

Out[122]:

	country	cast	count
14214	India	Anupam Kher	40
16854	India	Shah Rukh Khan	34
15860	India	Naseeruddin Shah	31
13982	India	Akshay Kumar	29
16010	India	Om Puri	29

In [123…
```
cntry_cast=cntry_cast.pivot(index='country',columns='cast',values='count')
cntry_cast.fillna(0,inplace=True)
```

In [124…
```
sns.heatmap(cntry_cast,cmap='summer',annot=True)
plt.xlabel('Actors')
plt.ylabel('Country')
plt.title('Top Actors in different countries')
plt.show()
```



Insights: -

- This heatmap not giving a clear picture of Actors with recpect to the country. Thus, we need to analyse the cast country wise separatly.

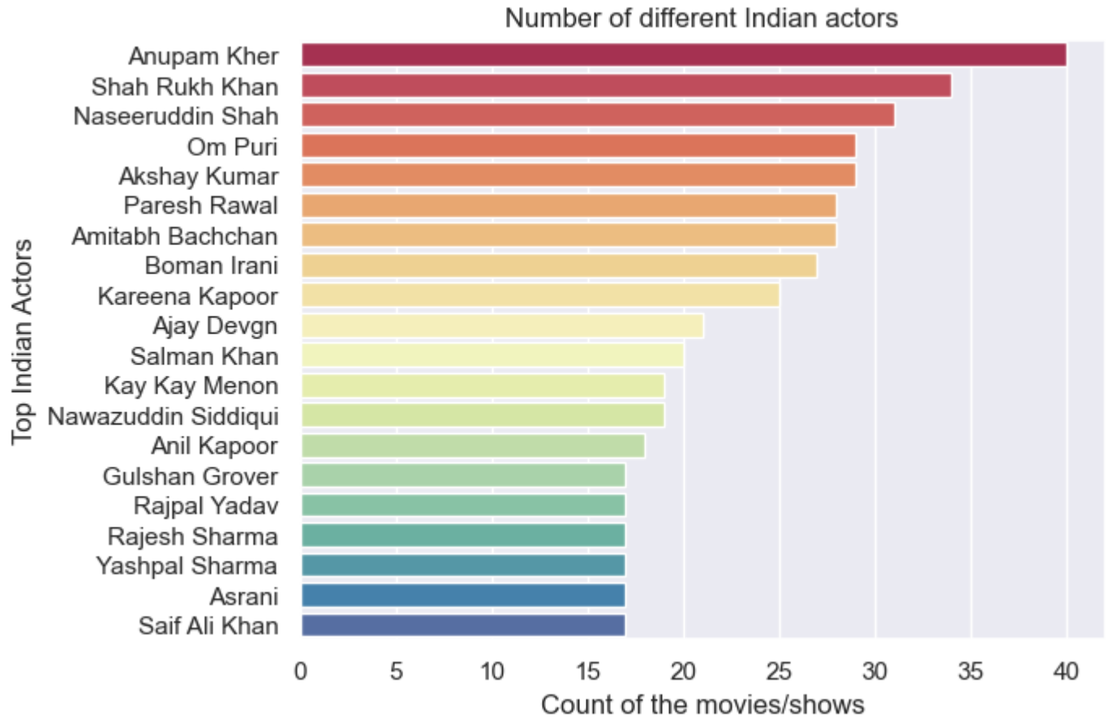
India

Actors & Directors distribution in india

1. Actors distribution with genre (India)

In [125…
```
cntry_cast=cast.merge(cntry,how='inner')
cntry_cast=cntry_cast[(cntry_cast['country']=='India') & (cntry_cast['cast']!='Unknown cast')]
cntry_cast=cntry_cast.groupby('cast').size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_cast=cntry_cast.iloc[:20]
```

In [126…
```
sns.barplot(data=cntry_cast,y='cast',x='count',palette='Spectral')
plt.xlabel('Count of the movies/shows')
plt.ylabel('Top Indian Actors')
plt.title('Number of different Indian actors')
plt.show()
```



Insights:-

- Top actors in India: - **Anupam Kher**, **Shah Rukh Khan**, **Naseeruddin Shah**, **Late Mr. Om Puri**, **Akshay Kumar** and **Paresh Rawel**.

Genre wise Indian actors distribution

In [127…

```
top5_indian_actors=cntry_cast.iloc[:5]
top5_indian_actors
```

Out[127]:

	cast	count
383	Anupam Kher	40
3023	Shah Rukh Khan	34
2029	Naseeruddin Shah	31
2179	Om Puri	29
151	Akshay Kumar	29

In [128…

```
indian_cast_genre=genre.merge(cast,how='inner')
indian_cast_genre=indian_cast_genre[indian_cast_genre['cast'].isin(top5_indian_actors['cast'])==True]
indian_cast_genre=indian_cast_genre.groupby(['cast','genre']).size().reset_index(name='count').sort_values(by=['cast','count'],ascending=False)
indian_cast_genre.head()
```

Out[128]:

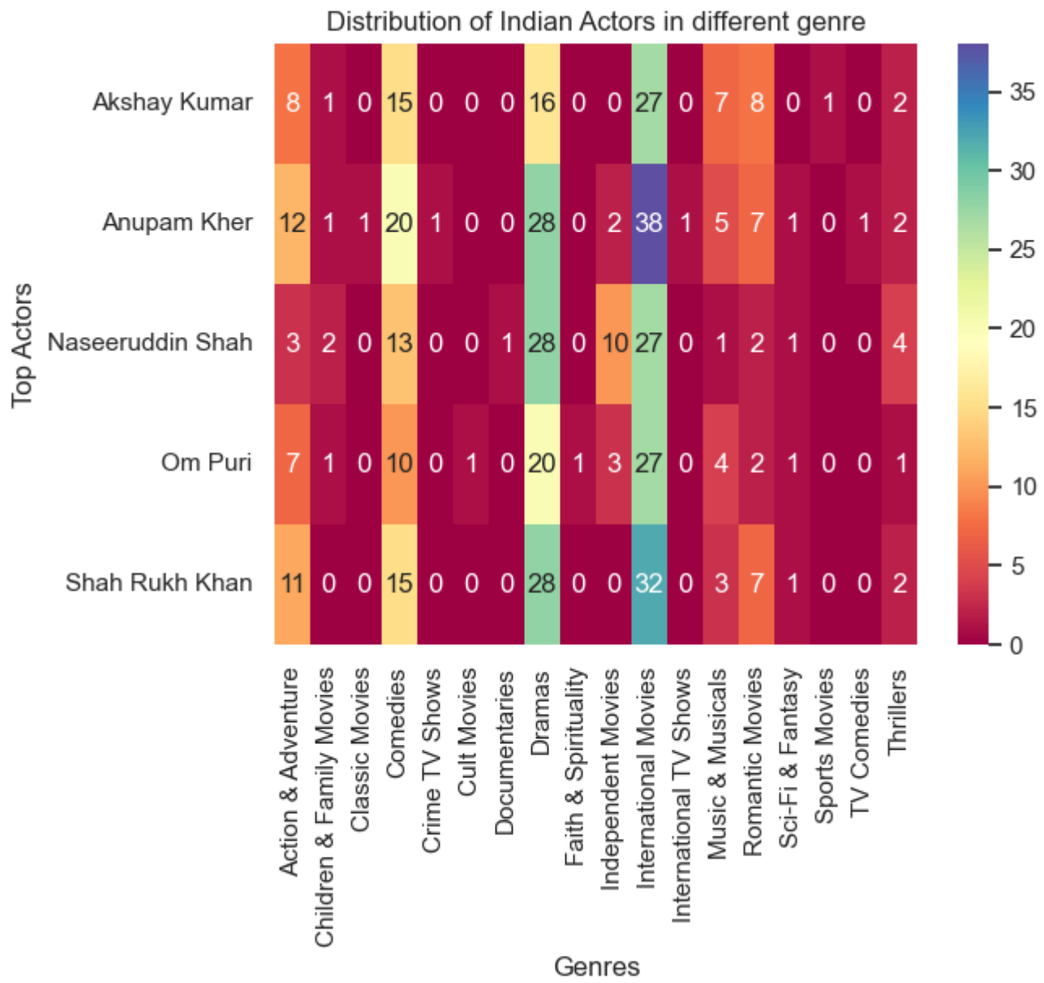
	cast	genre	count
49	Shah Rukh Khan	International Movies	32
48	Shah Rukh Khan	Dramas	28
47	Shah Rukh Khan	Comedies	15
46	Shah Rukh Khan	Action & Adventure	11
51	Shah Rukh Khan	Romantic Movies	7

In [129…

```
indian_cast_genre=indian_cast_genre.pivot(index='cast',columns='genre',values='count')
indian_cast_genre.fillna(0,inplace=True)
```

In [130…

```
sns.heatmap(indian_cast_genre,cmap='Spectral',annot=True)
plt.xlabel('Genres')
plt.ylabel('Top Actors')
plt.title('Distribution of Indian Actors in different genre')
plt.show()
```



Insights: -

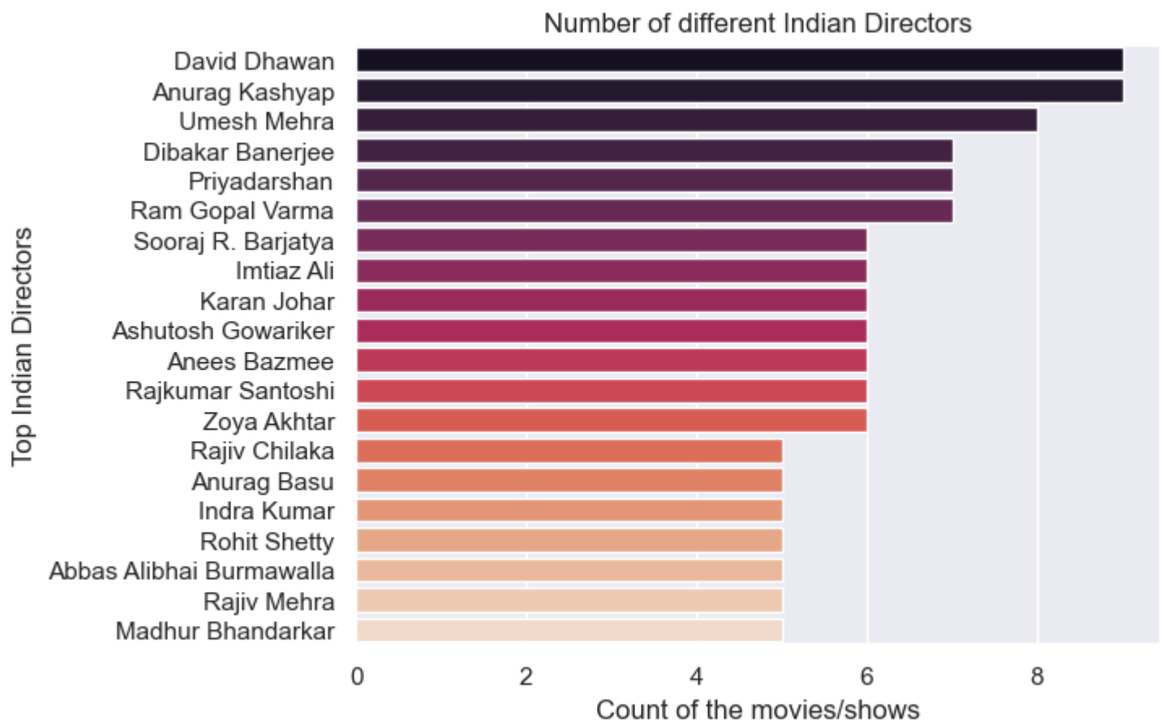
Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js **upam Kher** and **Shahrukh Khan** are top.

- Among **Dramas** -> **Anupam Kher**, **Naseeruddin shah** and **Shahrukh Khan** are top.
- Among **Comedies** -> **Anupam Kher**, **Akshay Kumar**, and **Shahrukh Khan** are top.

2. Directors distribution with genre (India)

In [131…
```
cntry_direct=direct.merge(cntry,how='inner')
cntry_direct=cntry_direct[(cntry_direct['country']=='India') & (cntry_direct['director']!='Unknown Director')]
cntry_direct=cntry_direct.groupby('director').size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_direct=cntry_direct.iloc[:20]
```

In [132…
```
sns.barplot(data=cntry_direct,y='director',x='count',palette='rocket')
plt.xlabel('Count of the movies/shows')
plt.ylabel('Top Indian Directors')
plt.title('Number of different Indian Directors')
plt.show()
```



Insights:-

- Top directors in India: - **David Dhawan**, **Anurag Kashyap**, **Umesh Mehra**, **Dibakar Banerjee**, **Priyadarshan** and **Ram Gopal Verma**.

Genre wise Indian director distribution

In [133…
```
top5_indian_direct=cntry_direct.iloc[:5]
top5_indian_direct
```

Out[133]:

	director	count
151	David Dhawan	9
80	Anurag Kashyap	9
682	Umesh Mehra	8
168	Dibakar Banerjee	7
427	Priyadarshan	7

In [134…

```
indian_direct_genre=genre.merge(direct,how='inner')
indian_direct_genre=indian_direct_genre[indian_direct_genre['director'].isin(top5_indian_direct['director'])==True]
indian_direct_genre=indian_direct_genre.groupby(['director','genre']).size().reset_index(name='count').sort_values(by=['director','count'],ascending=False)
indian_direct_genre.head()
```

Out[134]:

	director	genre	count
32	Umesh Mehra	International Movies	8
31	Umesh Mehra	Dramas	6
30	Umesh Mehra	Action & Adventure	3
33	Umesh Mehra	Music & Musicals	2
34	Umesh Mehra	Romantic Movies	1

In [135…

```
indian_direct_genre=indian_direct_genre.pivot(index='director',columns='genre',values='count')
indian_direct_genre.fillna(0,inplace=True)
```

In [136…

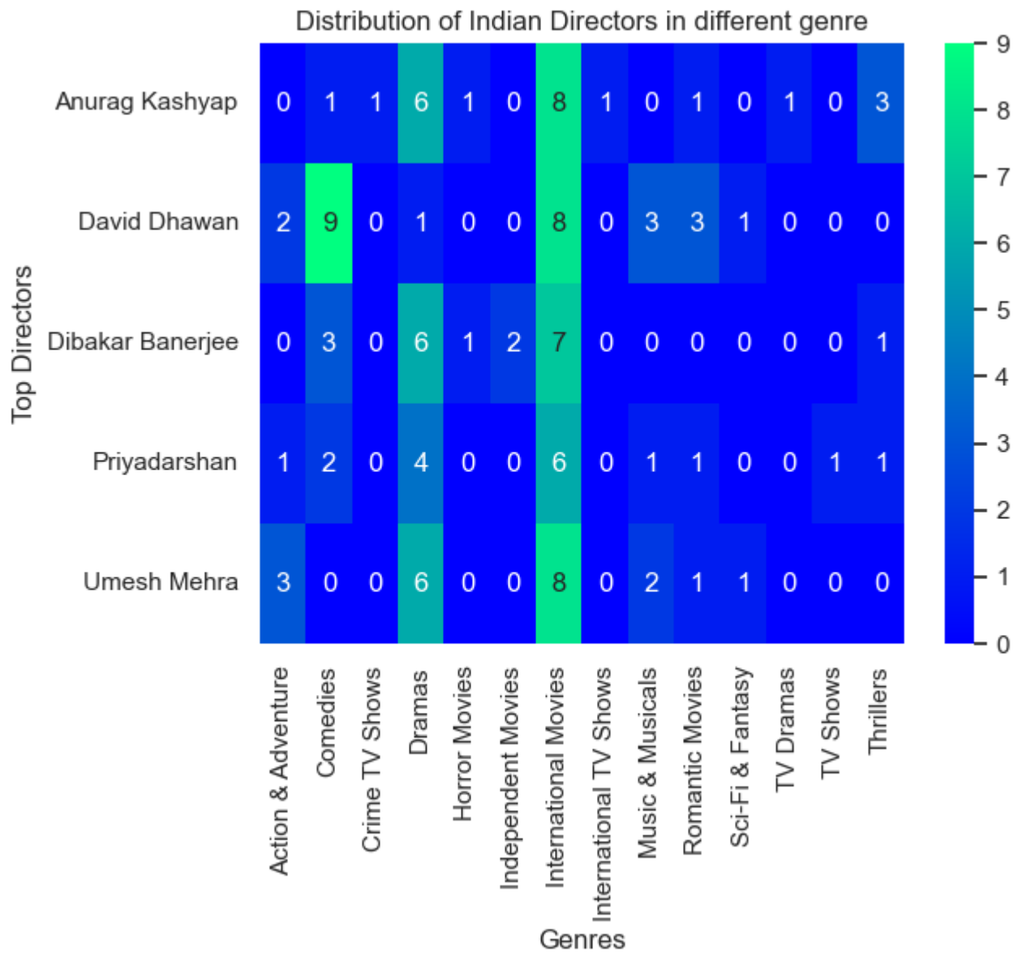
```
indian_direct_genre
```

Out[136]:

genre	Action & Adventure	Comedies	Crime TV Shows	Dramas	Horror Movies	Independent Movies	International Movies	International TV Shows	Music & Musicals	Romantic Movies	Sci-Fi & Fantasy	TV Dramas	TV Shows	Thrillers
director														
Anurag Kashyap	0.0	1.0	1.0	6.0	1.0	0.0	8.0	1.0	0.0	1.0	0.0	1.0	0.0	3.0
David Dhawan	2.0	9.0	0.0	1.0	0.0	0.0	8.0	0.0	3.0	3.0	1.0	0.0	0.0	0.0
Dibakar Banerjee	0.0	3.0	0.0	6.0	1.0	2.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Priyadarshan	1.0	2.0	0.0	4.0	0.0	0.0	6.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0
Umesh Mehra	3.0	0.0	0.0	6.0	0.0	0.0	8.0	0.0	2.0	1.0	1.0	0.0	0.0	0.0

In [137…

```
sns.heatmap(indian_direct_genre,annot=True,cmap='winter')
plt.xlabel('Genres')
plt.ylabel('Top Directors')
plt.title('Distribution of Indian Directors in different genre')
plt.show()
```



Insights: -

- Among **Comedies** -> **David Dhawan** on top.
- Among **Dramas** -> **Dibakar Banerjee**, **Umesh Mehra**, and **Anurag Kashyap** on top.
- Among **Thriller** -> **Anurag Kashyap** on top.
- Among **Musicals** -> **David Dhawan** on top.
- Among **Romantics** -> **David Dhawan** on top.

Actors & Directors distribution in USA

1. Actors distribution

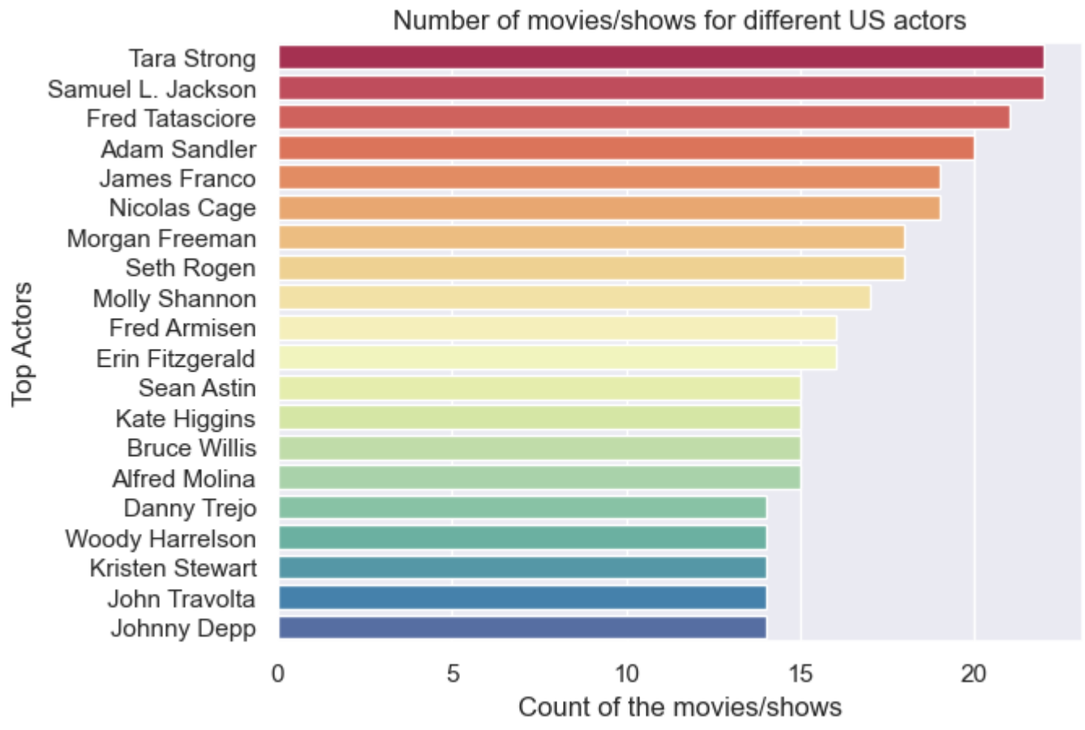
In [138…

```
cntry_cast=cast.merge(cntry,how='inner')
cntry_cast=cntry_cast[(cntry_cast['country']=='United States') & (cntry_cast['cast']!='Unknown cast')]
cntry_cast=cntry_cast.groupby('cast').size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_cast=cntry_cast.iloc[:20]
```

In [139…

```
sns.barplot(data=cntry_cast,y='cast',x='count',palette='Spectral')
plt.xlabel('Count of the movies/shows')
```

```
plt.title('Number of movies/shows for different US actors')
plt.show()
```



Insights:-

- Top actors in USA: - **Tara strong**, **Sam L Jackson**, **Fred Tatasciore**, **Adam Sandler**, **James Franco** and **Nicolas Cage**.

Genre wise US actors distribution.

In [140…
```
top5_us_actors=cntry_cast.iloc[:5]
top5_us_actors
```

Out[140]:

	cast	count
13823	Tara Strong	22
12748	Samuel L. Jackson	22
4871	Fred Tatasciore	21
136	Adam Sandler	20
6082	James Franco	19

In [141…
```
us_cast_genre=genre.merge(cast,how='inner')
us_cast_genre=us_cast_genre[us_cast_genre['cast'].isin(top5_us_actors['cast'])==True]
us_cast_genre=us_cast_genre.groupby(['cast','genre']).size().reset_index(name='count').sort_values(by=['cast','count'],ascending=False)
us_cast_genre.head()
```

Out[141]:

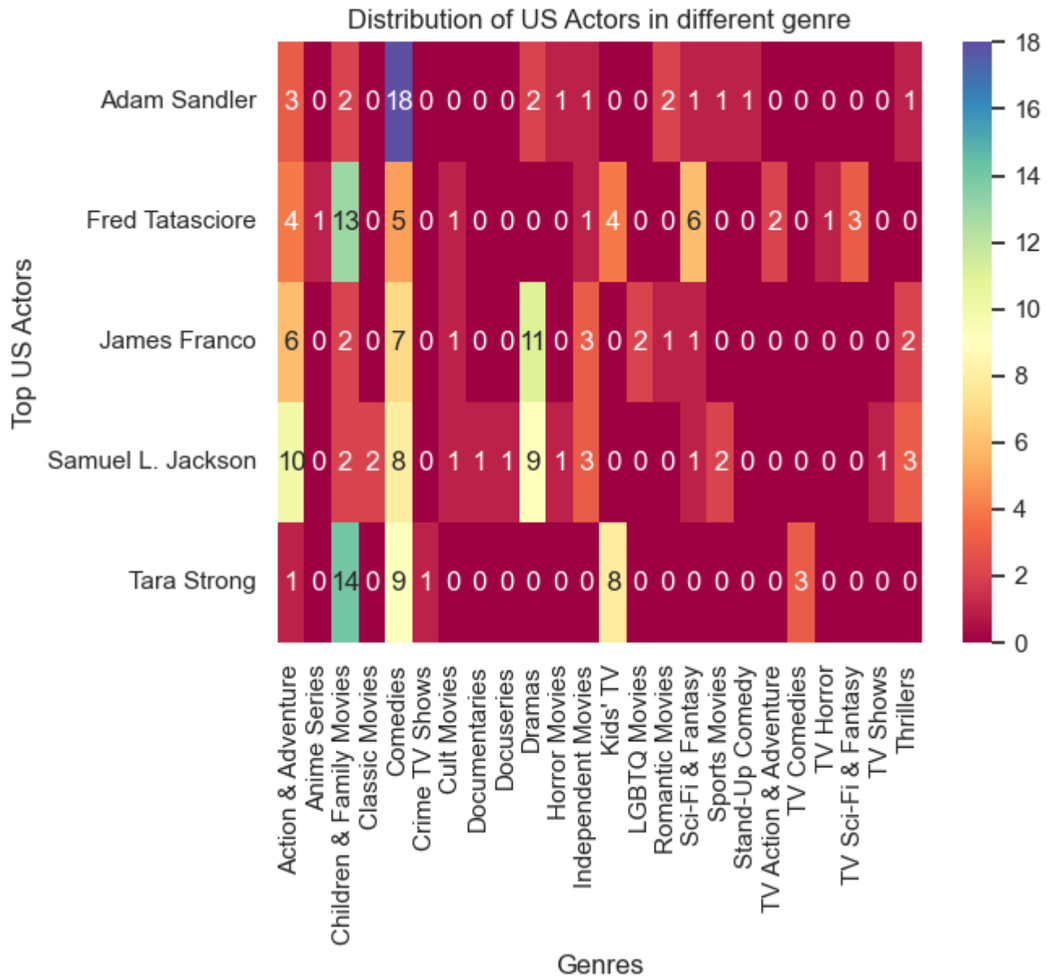
	cast	genre	count
47	Tara Strong	Children & Family Movies	14
48	Tara Strong	Comedies	9
50	Tara Strong	Kids' TV	8
51	Tara Strong	TV Comedies	3
46	Tara Strong	Action & Adventure	1

In [142…

```
us_cast_genre=us_cast_genre.pivot(index='cast',columns='genre',values='count')
us_cast_genre.fillna(0,inplace=True)
```

In [143…

```
sns.heatmap(us_cast_genre,cmap='Spectral',annot=True)
plt.xlabel('Genres')
plt.ylabel('Top US Actors')
plt.title('Distribution of US Actors in different genre')
plt.show()
```



Insights: -

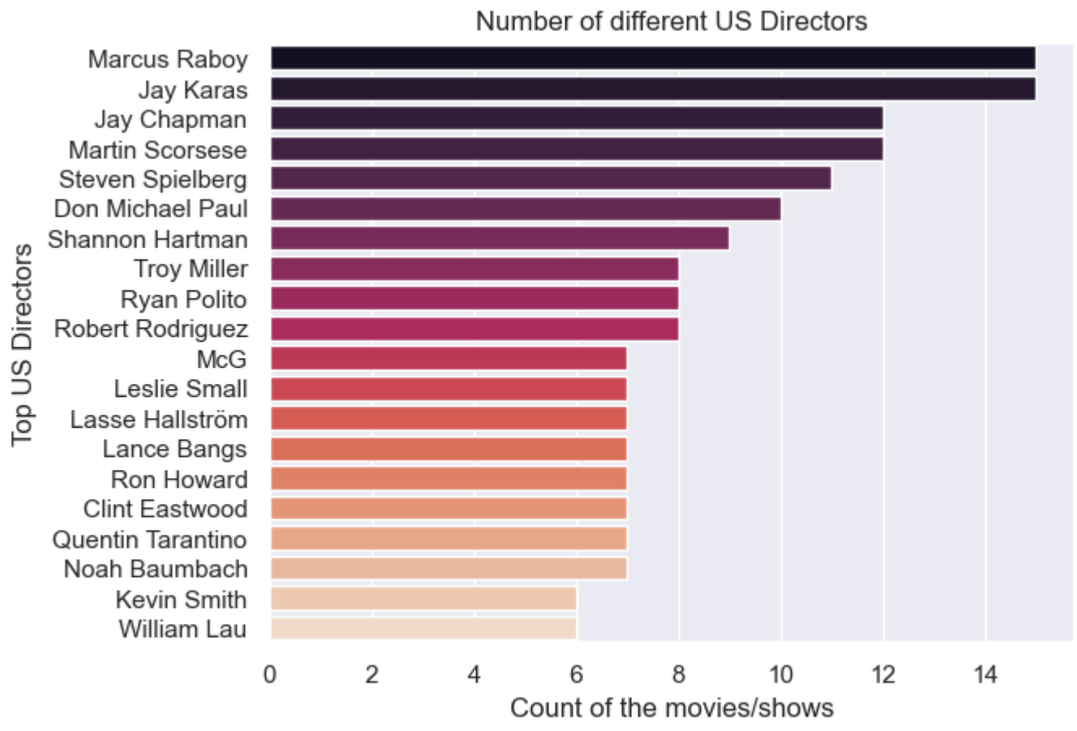
 on top.

- Among **Children & Family movies** -> **Fred Tatasciore** on top.
- Among **Dramas** -> **James Franco** and **Sam L Jackson** on top.
- Among **Kid's TV** -> **Tara Strong** on top.
- Among **Action & Adventure** -> **Samuel L. Jackson** on top.

2. Directors distribution in US

In [144…
```
cntry_direct=direct.merge(cntry,how='inner')
cntry_direct=cntry_direct[(cntry_direct['country']=='United States') & (cntry_direct['director']!='Unknown Director')]
cntry_direct=cntry_direct.groupby('director').size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_direct=cntry_direct.iloc[:20]
```

In [145…
```
sns.barplot(data=cntry_direct,y='director',x='count',palette='rocket')
plt.xlabel('Count of the movies/shows')
plt.ylabel('Top US Directors')
plt.title('Number of different US Directors')
plt.show()
```



Insights:-

- Top directors in US: - **Marcus Raboy**, **Jay Karas**, **Jay Chapman**, **Martin Scorsese**, **Steven Spielberg** and **Ddon Michael Paul**.

Genre wise US directors distribution

In [146…
```
top5_us_direct=cntry_direct.iloc[:10]
top5_us_direct
```

Out[146]:

	director	count
1391	Marcus Raboy	15
925	Jay Karas	15
924	Jay Chapman	12
1437	Martin Scorsese	12
2078	Steven Spielberg	11
593	Don Michael Paul	10
1999	Shannon Hartman	9
2215	Troy Miller	8
1921	Ryan Polito	8
1860	Robert Rodriguez	8

In [147…

```
us_direct_genre=genre.merge(direct,how='inner')
us_direct_genre=us_direct_genre[us_direct_genre['director'].isin(top5_us_direct['director'])==True]
us_direct_genre=us_direct_genre.groupby(['director','genre']).size().reset_index(name='count').sort_values(by=['director','count'],ascending=False)
us_direct_genre.head()
```

Out[147]:

	director	genre	count
35	Troy Miller	Stand-Up Comedy	7
34	Troy Miller	Comedies	1
30	Steven Spielberg	Children & Family Movies	6
32	Steven Spielberg	Dramas	6
29	Steven Spielberg	Action & Adventure	5

In [148…

```
us_direct_genre=us_direct_genre.pivot(index='director',columns='genre',values='count')
us_direct_genre.fillna(0,inplace=True)
```

In [149…

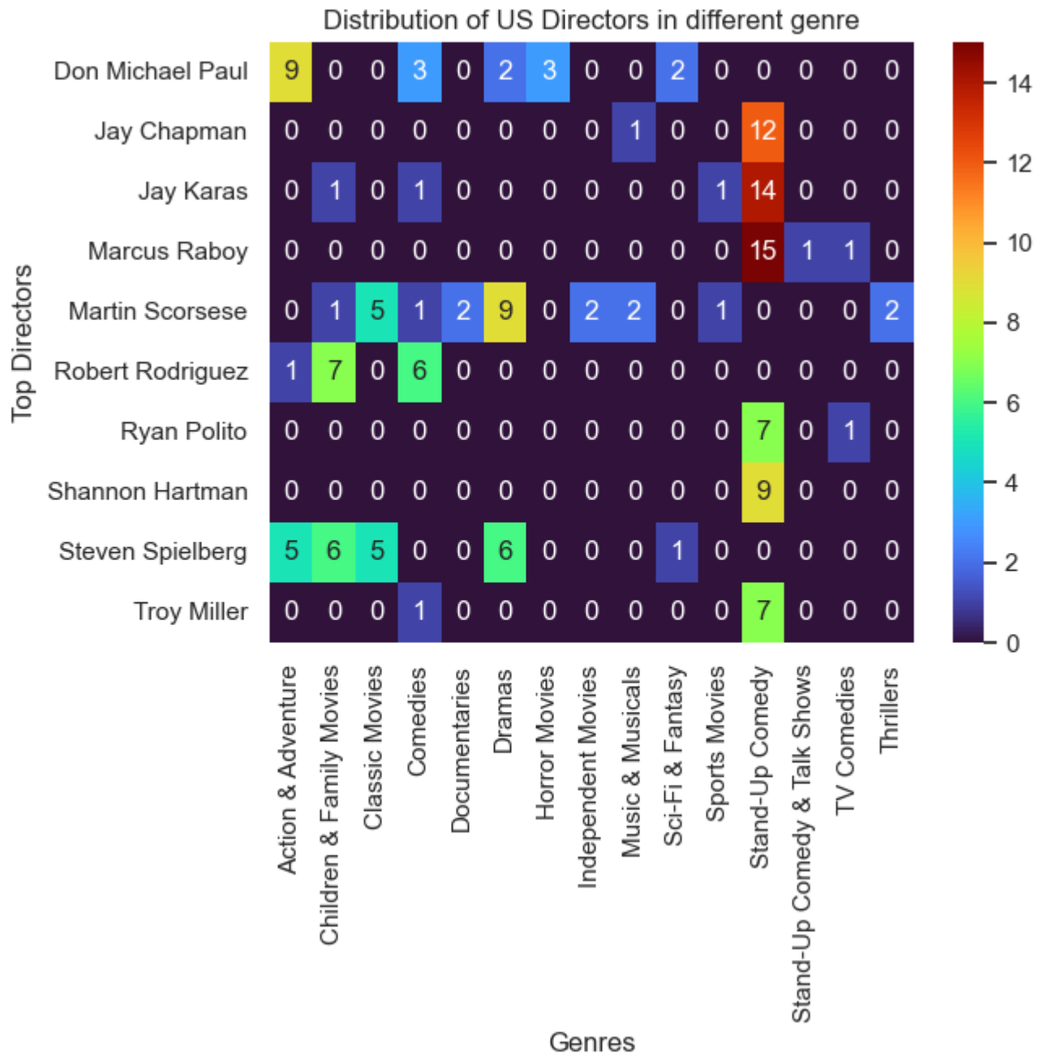
```
us_direct_genre
```

Out[149]:

director	genre	Action & Adventure	Children & Family Movies	Classic Movies	Comedies	Documentaries	Dramas	Horror Movies	Independent Movies	Music & Musicals	Sci-Fi & Fantasy	Sports Movies	Stand-Up Comedy	Stand-Up Comedy & Talk Shows	TV Comedies	Thrillers
Don Michael Paul		9.0	0.0	0.0	3.0	0.0	2.0	3.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
Jay Chapman		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	12.0	0.0	0.0	0.0
Jay Karas		0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	14.0	0.0	0.0	0.0
Marcus Raboy		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.0	1.0	1.0	0.0
Martin Scorsese		0.0	1.0	5.0	1.0	2.0	9.0	0.0	2.0	2.0	0.0	1.0	0.0	0.0	0.0	2.0
Robert Rodriguez		1.0	7.0	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ryan Polito		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.0	1.0	0.0
Shannon Hartman		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	0.0	0.0	0.0
Steven Spielberg		5.0	6.0	5.0	0.0	0.0	6.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Troy Miller		0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0

In [150…

```
sns.heatmap(us_direct_genre,annot=True,cmap='turbo')
plt.xlabel('Genres')
plt.ylabel('Top Directors')
plt.title('Distribution of US Directors in different genre')
plt.show()
```



Insights: -

- Among **Dramas** -> **Martin Scorsese**, and **Steven Spielberg**.
- Among **Action & Adventure, Children & Family movies, Dramas** -> **Steven Spielberg**.
- Among **Comedies** -> **Robert Rodriguez**

Actors & Directors distribution in UK

1) Actors distribution

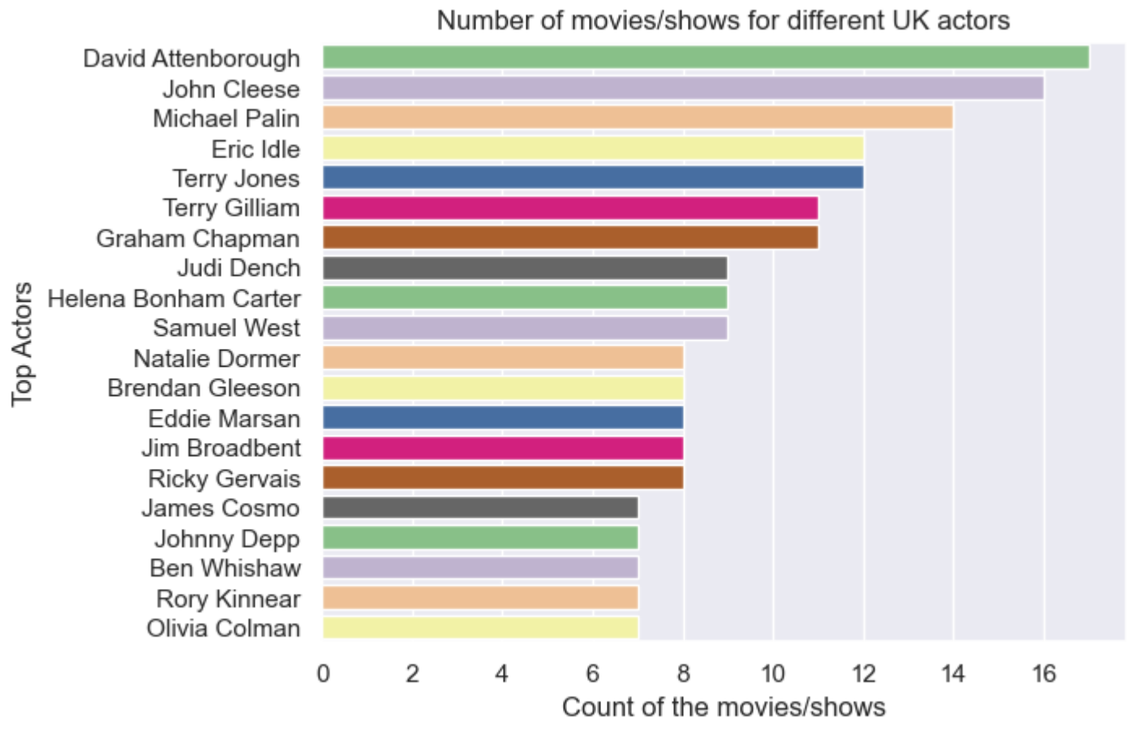
In [151…

```
cntry_cast=cast.merge(cntry,how='inner')
cntry_cast=cntry_cast[(cntry_cast['country']=='United Kingdom') & (cntry_cast['cast']!='Unknown cast')]
cntry_cast=cntry_cast.groupby('cast').size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_cast=cntry_cast.iloc[:20]
```

In [152…

```
sns.barplot(data=cntry_cast,y='cast',x='count',palette='Accent')
                                          s')
plt.ylabel( 'Top Actors' )
```

```
plt.title('Number of movies/shows for different UK actors')
plt.show()
```



Insights:-

- Top actors in UK: - **David Attenborough**, **John Cleese**, **Micheal Palin**, **Eric Idle**, **Terry Jones** and **Terry Gilliam**.

Genre wise UK Top actors distribution

In [153…

```
top20_uk_actors=cntry_cast.iloc[:20]
top20_uk_actors
```

Out[153]:

	cast	count
840	David Attenborough	17
1775	John Cleese	16
2504	Michael Palin	14
1125	Eric Idle	12
3579	Terry Jones	12
3578	Terry Gilliam	11
1312	Graham Chapman	11
1875	Judi Dench	9
1392	Helena Bonham Carter	9
3228	Samuel West	9
2621	Natalie Dormer	8
496	Brendan Gleeson	8
1026	Eddie Marsan	8
1701	Jim Broadbent	8
3047	Ricky Gervais	8
1529	James Cosmo	7
1812	Johnny Depp	7
419	Ben Whishaw	7
3127	Rory Kinnear	7
2739	Olivia Colman	7

In [154…

```
uk_cast_genre=genre.merge(cast,how='inner')
uk_cast_genre=uk_cast_genre[uk_cast_genre['cast'].isin(top20_uk_actors['cast'])==True]
uk_cast_genre=uk_cast_genre.groupby(['cast','genre']).size().reset_index(name='count').sort_values(by=['cast','count'],ascending=False)
uk_cast_genre.head()
```

Out[154]:

	cast	genre	count
212	Terry Jones	Comedies	7
209	Terry Jones	British TV Shows	4
214	Terry Jones	Documentaries	4
217	Terry Jones	TV Comedies	3
211	Terry Jones	Classic Movies	2

In [155…

```
uk_cast_genre=uk_cast_genre.pivot(index='cast',columns='genre',values='count')
uk_cast_genre.fillna(0,inplace=True)
uk_cast_genre
```

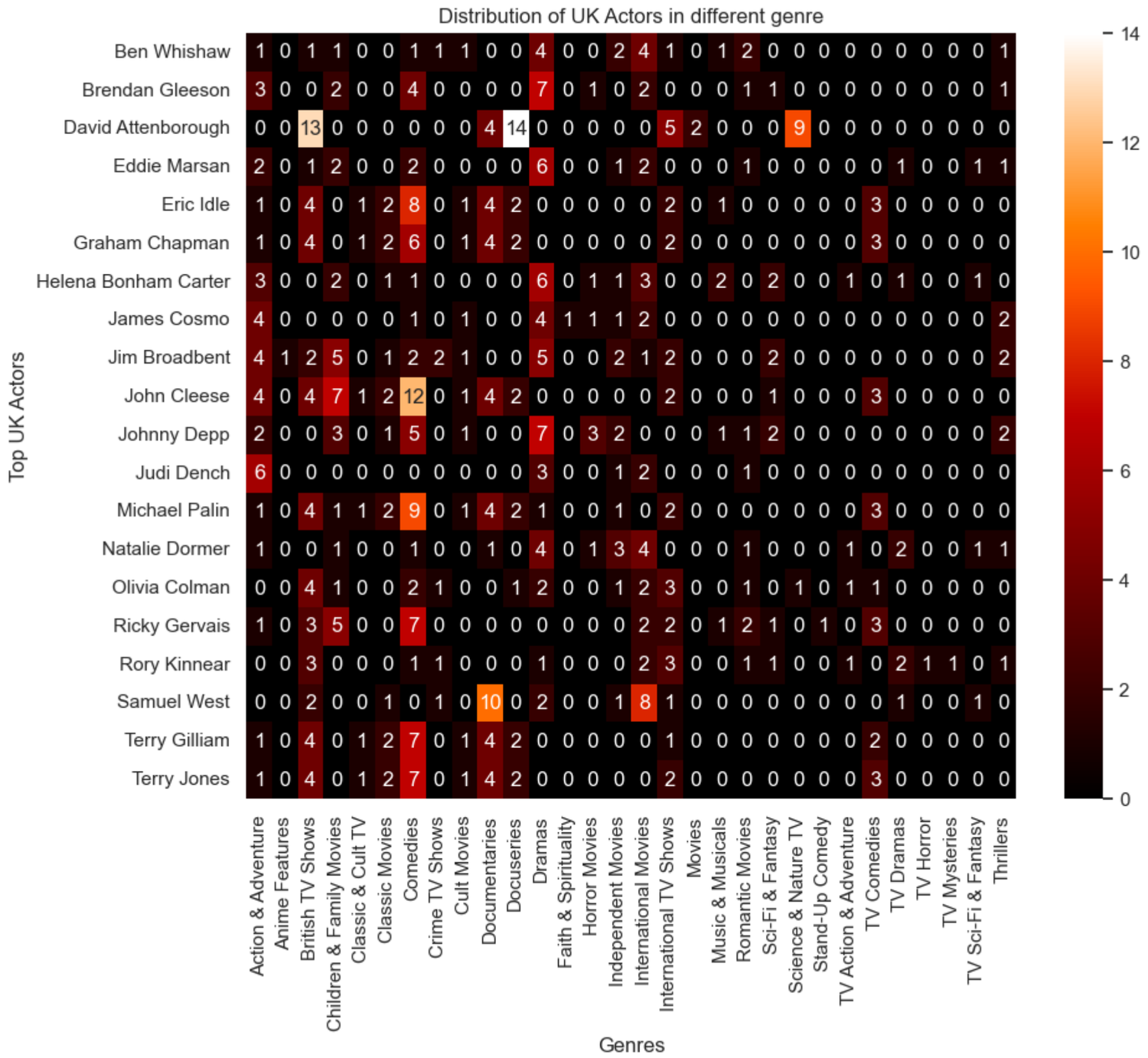
Out[155]:

cast	Action & Adventure	Anime Features	British TV Shows	Children & Family Movies	Classic & Cult TV	Classic Movies	Comedies	Crime TV Shows	Cult Movies	Documentaries	...	Sci-Fi & Fantasy	Science & Nature TV	Stand-Up Comedy	TV Action & Adventure	TV Comedies	TV Dramas	TV Horror	TV Mysteries	TV Sci-Fi & Fantasy	Thrillers
genre																					
Ben Whishaw	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Brendan Gleeson	3.0	0.0	0.0	2.0	0.0	0.0	4.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
David Attenborough	0.0	0.0	13.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	...	0.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Eddie Marsan	2.0	0.0	1.0	2.0	0.0	0.0	2.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0
Eric Idle	1.0	0.0	4.0	0.0	1.0	2.0	8.0	0.0	1.0	4.0	...	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0
Graham Chapman	1.0	0.0	4.0	0.0	1.0	2.0	6.0	0.0	1.0	4.0	...	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0
Helena Bonham Carter	3.0	0.0	0.0	2.0	0.0	1.0	1.0	0.0	0.0	0.0	...	2.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0
James Cosmo	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
Jim Broadbent	4.0	1.0	2.0	5.0	0.0	1.0	2.0	2.0	1.0	0.0	...	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
John Cleese	4.0	0.0	4.0	7.0	1.0	2.0	12.0	0.0	1.0	4.0	...	1.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0
Johnny Depp	2.0	0.0	0.0	3.0	0.0	1.0	5.0	0.0	1.0	0.0	...	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
Judi Dench	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Michael Palin	1.0	0.0	4.0	1.0	1.0	2.0	9.0	0.0	1.0	4.0	...	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0
Natalie Dormer	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	...	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	1.0	1.0
Olivia Colman	0.0	0.0	4.0	1.0	0.0	0.0	2.0	1.0	0.0	0.0	...	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
Ricky Gervais	1.0	0.0	3.0	5.0	0.0	0.0	7.0	0.0	0.0	0.0	...	1.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0
Rory Kinnear	0.0	0.0	3.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...	1.0	0.0	0.0	1.0	0.0	2.0	1.0	1.0	0.0	1.0
Samuel West	0.0	0.0	2.0	0.0	0.0	1.0	0.0	1.0	0.0	10.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
Terry Gilliam	1.0	0.0	4.0	0.0	1.0	2.0	7.0	0.0	1.0	4.0	...	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
Terry Jones	1.0	0.0	4.0	0.0	1.0	2.0	7.0	0.0	1.0	4.0	...	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0

20 rows × 30 columns

In [156…

```
plt.figure(figsize=(10,8))
sns.heatmap(uk_cast_genre,cmap='gist_heat',annot=True)
plt.xlabel('Genres')
plt.ylabel('Top UK Actors')
plt.title('Distribution of UK Actors in different genre')
plt.show()
```



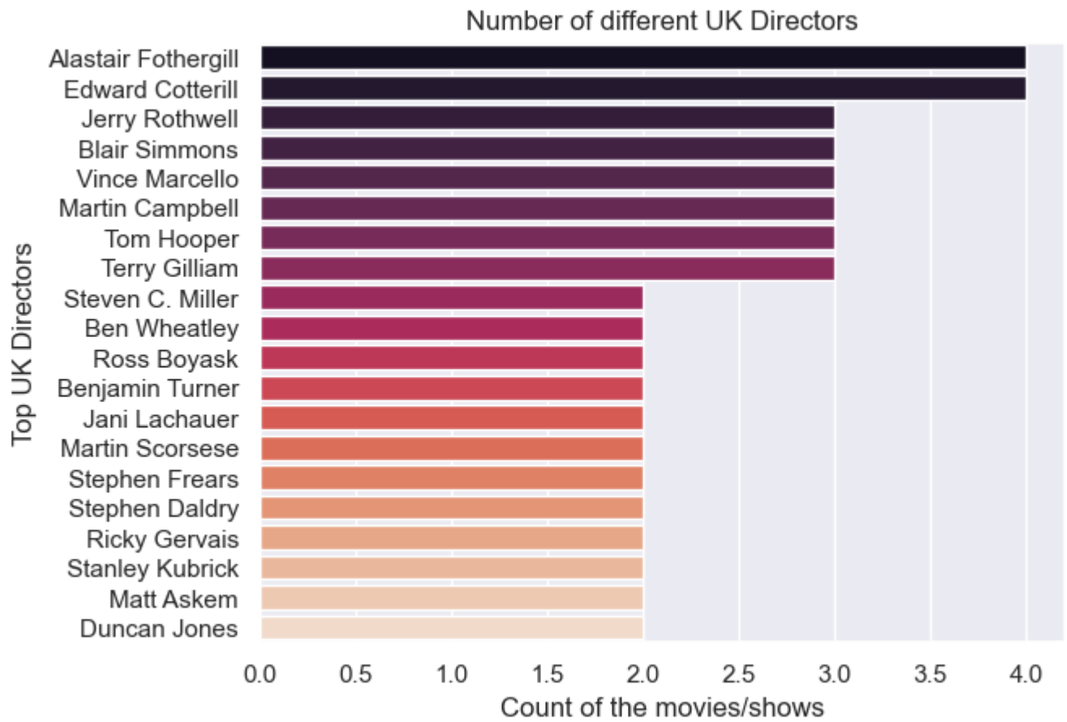
Insights: - Actors distribution in Top Genres in UK

- Among **Dramas** -> **Johnny Depp**, **Brendan Gleeson**,**Eddie Marson**, and **HElena Bonham Carter**
- Among **International Movies** -> **Samuel West**
- Among **International TV Shows** -> **Olivia Colnman**, and **Rorry Kinnear**
- Among **Documentaries** -> **David Attenborough** on top.

2) Directors distribution in United Kingdom

In [157…
```
cntry_direct=direct.merge(cntry,how='inner')
cntry_direct=cntry_direct[(cntry_direct['country']=='United Kingdom') & (cntry_direct['director']!='Unknown Director')]
cntry_direct=cntry_direct.groupby('director').size().reset_index(name='count').sort_values(by='count',ascending=False)
cntry_direct=cntry_direct.iloc[:20]
```

In [158…
```
sns.barplot(data=cntry_direct,y='director',x='count',palette='rocket')
plt.xlabel('Count of the movies/shows')
plt.ylabel('Top UK Directors')
plt.title('Number of different UK Directors')
plt.show()
```



Insights:-

- Top directors in UK in India: - **Alastair Fothergill**, **Edward cotterill**, **Jerry Rothwell**, **Blair Simmons**, **Vince Marcello**, **Matin campbell**, and **Tom Hooper**.

Genre wise UK directors distribution

In [159…
```
top5_uk_direct=cntry_direct.iloc[:20]
top5_uk_direct
```

Out[159]:

	director	count
11	Alastair Fothergill	4
120	Edward Cotterill	4
201	Jerry Rothwell	3
59	Blair Simmons	3
503	Vince Marcello	3
312	Martin Campbell	3
491	Tom Hooper	3
476	Terry Gilliam	3
466	Steven C. Miller	2
51	Ben Wheatley	2
431	Ross Boyask	2
54	Benjamin Turner	2
192	Jani Lachauer	2
313	Martin Scorsese	2
462	Stephen Frears	2
461	Stephen Daldry	2
411	Ricky Gervais	2
459	Stanley Kubrick	2
317	Matt Askem	2
115	Duncan Jones	2

In [160…

```
uk_direct_genre=genre.merge(direct,how='inner')
uk_direct_genre=uk_direct_genre[uk_direct_genre['director'].isin(top5_uk_direct['director'])==True]
uk_direct_genre=uk_direct_genre.groupby(['director','genre']).size().reset_index(name='count').sort_values(by=['director','count'],ascending=False)
uk_direct_genre.head()
```

Out[160]:

	director	genre	count
71	Vince Marcello	Comedies	5
73	Vince Marcello	Romantic Movies	3
70	Vince Marcello	Children & Family Movies	2
72	Vince Marcello	Dramas	1
66	Tom Hooper	Dramas	3

In [161…

```
uk_direct_genre=uk_direct_genre.pivot(index='director',columns='genre',values='count')
uk_direct_genre.fillna(0,inplace=True)
```

In [162…

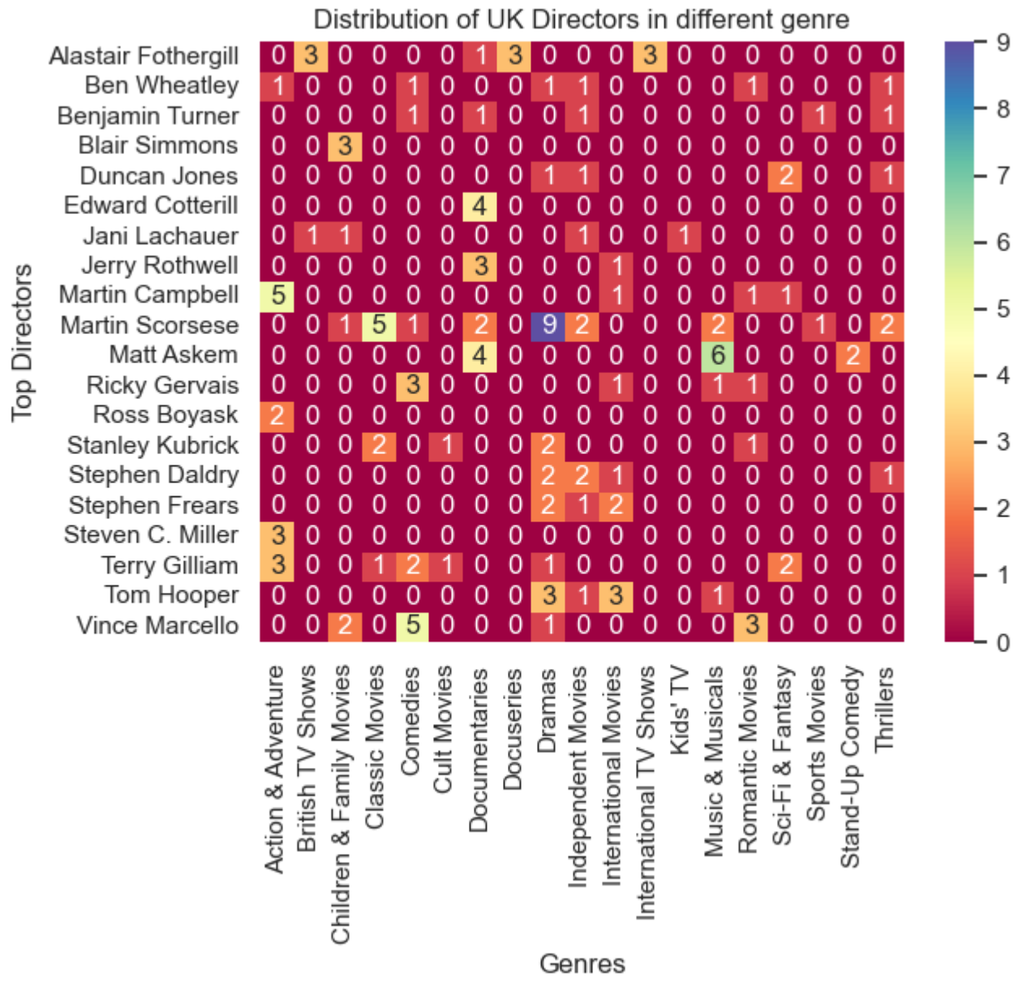
```
uk_direct_genre
```

Out[162]:

genre	Action & Adventure	British TV Shows	Children & Family Movies	Classic Movies	Comedies	Cult Movies	Documentaries	Docuseries	Dramas	Independent Movies	International Movies	International TV Shows	Kids' TV	Music & Musicals	Romantic Movies	Sci-Fi & Fantasy	Sports Movies	Stand-Up Comedy	Thrillers
director																			
Alastair Fothergill	0.0	3.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ben Wheatley	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
Benjamin Turner	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
Blair Simmons	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Duncan Jones	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0
Edward Cotterill	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Jani Lachauer	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Jerry Rothwell	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Martin Campbell	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
Martin Scorsese	0.0	0.0	1.0	5.0	1.0	0.0	2.0	0.0	9.0	2.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	2.0
Matt Askem	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	2.0	0.0
Ricky Gervais	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
Ross Boyask	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Stanley Kubrick	0.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
Stephen Daldry	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Stephen Frears	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Steven C. Miller	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Terry Gilliam	3.0	0.0	0.0	1.0	2.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
Tom Hooper	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	1.0	3.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Vince Marcello	0.0	0.0	2.0	0.0	5.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0

In [163]:
```
sns.heatmap(uk_direct_genre,annot=True,cmap='Spectral')
plt.ylabel('Top Directors')
```

```
plt.title('Distribution of UK Directors in different genre')
plt.show()
```



Insights: -

- Among **Dramas** -> **Martin Scorsese** on top.
- Among **International movies** -> **Top Hooper** on top.
- Among **Internation TV Shows** -> **Alastair Fothergill** on top.
- Among **Documentaries** -> **Edward Cotterill** on top.

- Till now, We have analysed the directors and actors distribution with different genre in case of Top 3 countries - US, India, UK.

Certification analysis

1) Coutry-wise certification distribution

In [164...

```
cntry_cert=cntry.merge(df)
try','certification']).size().reset_index(name='count').sort_values(by=['country','count'],ascending=(True,False))
```

In [165…
```
cntry_cert
```

Out[165]:

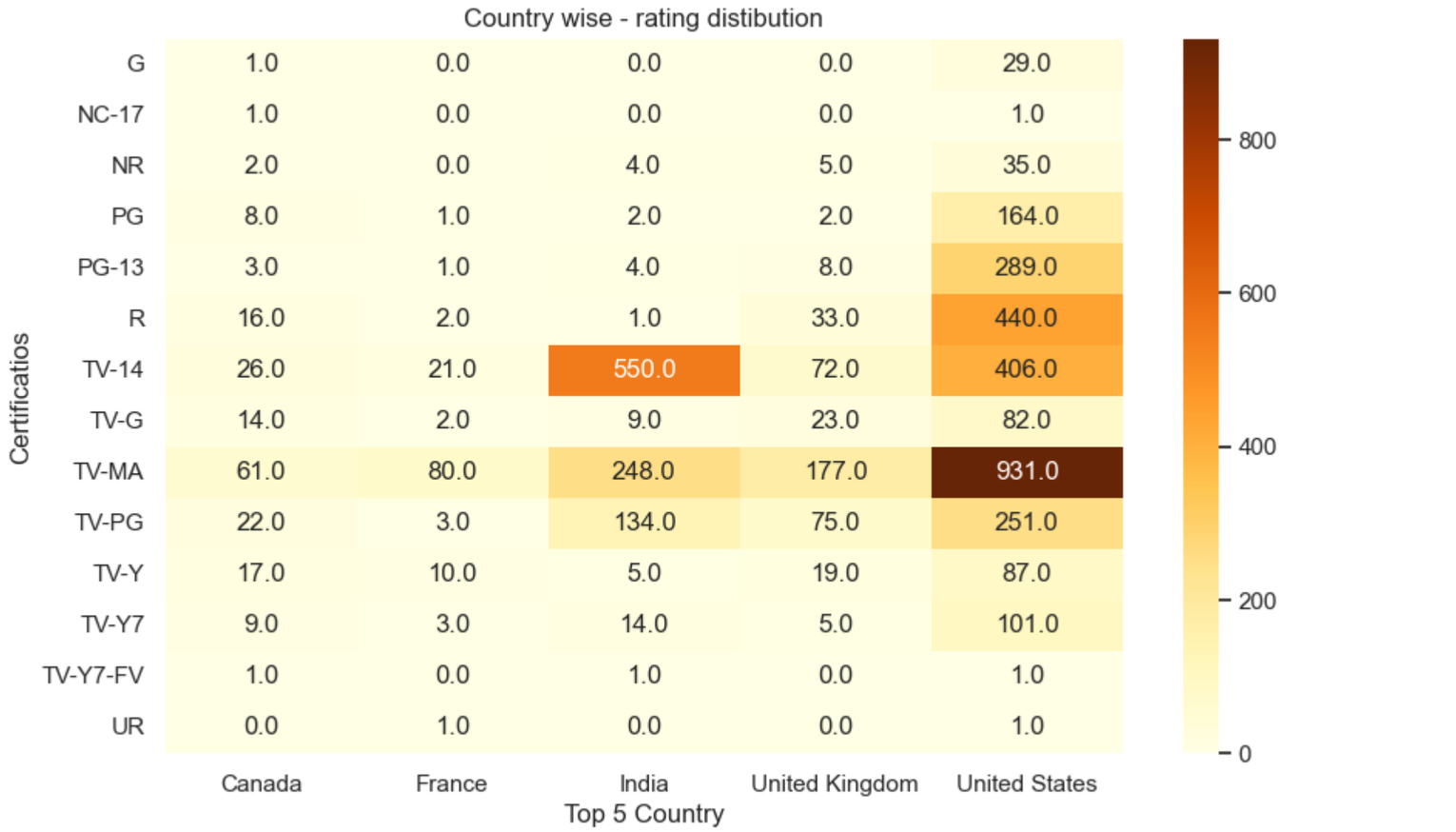
	country	certification	count
4	Argentina	TV-MA	39
2	Argentina	TV-14	7
5	Argentina	TV-PG	3
0	Argentina	NR	2
3	Argentina	TV-G	2
...	...	...	...
293	Vietnam	TV-14	3
295	Vietnam	TV-MA	3
294	Vietnam	TV-G	1
296	West Germany	TV-MA	1
297	Zimbabwe	TV-G	1

298 rows × 3 columns

In [166…
```
cntry_cert=cntry_cert[cntry_cert['country'].isin(top5_cntry)==True]
```

In [167…
```
cntry_cert=cntry_cert.pivot(index='country',columns='certification',values='count')
cntry_cert.fillna(0,inplace=True)
```

In [168…
```
plt.figure(figsize=(10,6))
sns.heatmap(cntry_cert.T,cmap='YlOrBr',annot=True,fmt=".1f")
plt.xlabel('Top 5 Country')
plt.ylabel('Certificatios')
plt.title('Country wise - rating distibution')
plt.show()
```



Insights: -

- In all the top countries Mature Adult content i.e., **TV-MA** rating is popular.
- In **India**, The most popular content is of >14 rating Year old ie., **TV-14**. After that **TV-MA** and **TV-PG** rating.

2) Genre-wise certificaiton distribution

In [169…
```
genre_cert=genre.merge(df)
genre_cert=genre_cert.groupby(['genre','certification']).size().reset_index(name='count').sort_values(by=['genre','count'],ascending=(True,False))
```

In [170…
```
genre_cert=genre_cert[genre_cert['genre'].isin(top5_genre)]
```

In [171…
```
genre_cert=genre_cert.pivot(index='genre',columns='certification',values='count')
genre_cert.fillna(0,inplace=True)
```

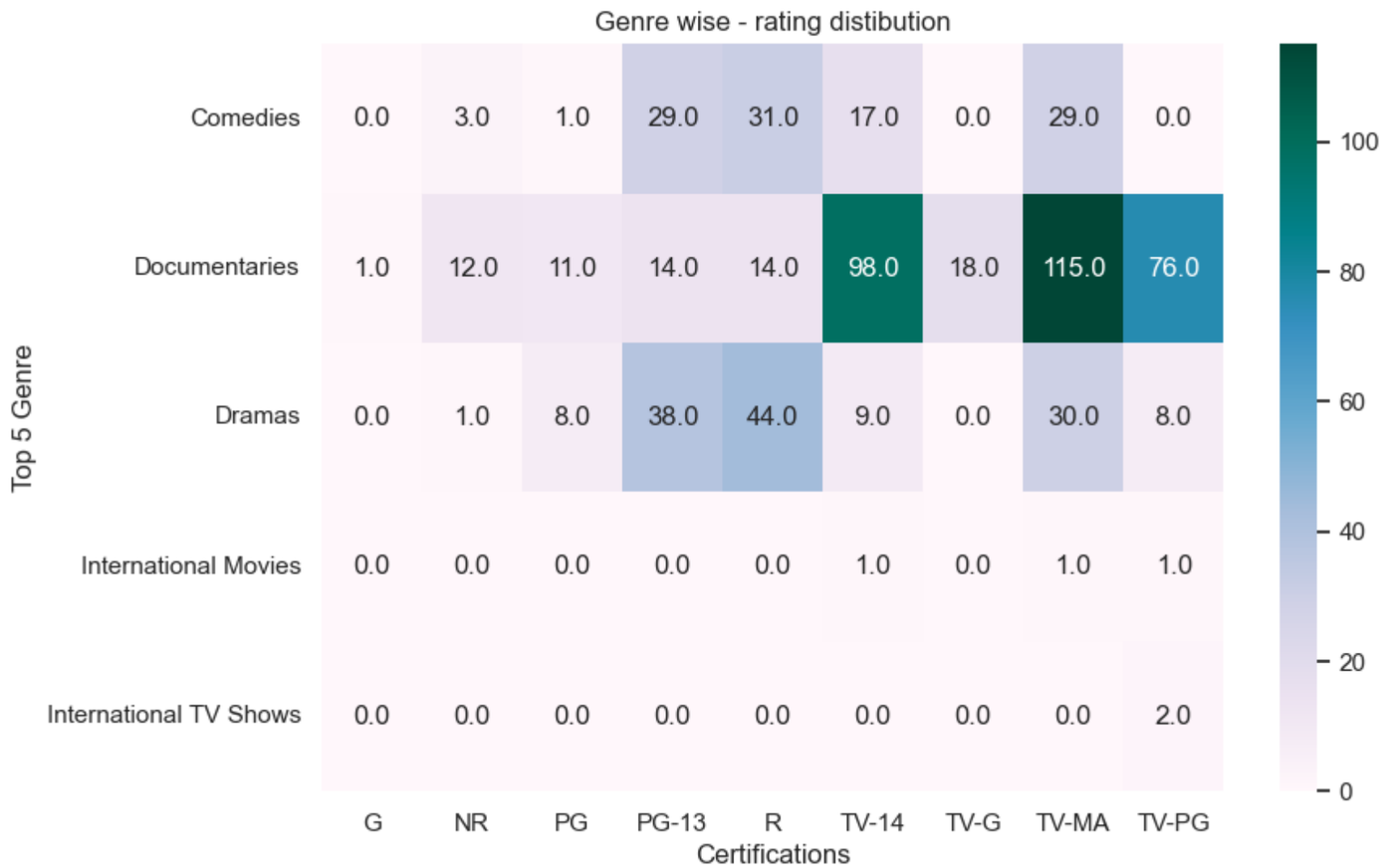
In [172…
```
genre_cert
```

Out[172]:

certification genre	G	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG
Comedies	0.0	3.0	1.0	29.0	31.0	17.0	0.0	29.0	0.0
Documentaries	1.0	12.0	11.0	14.0	14.0	98.0	18.0	115.0	76.0
Dramas	0.0	1.0	8.0	38.0	44.0	9.0	0.0	30.0	8.0
International Movies	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
International TV Shows	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0

In [173…

```
plt.figure(figsize=(9,6))
sns.heatmap(genre_cert,cmap='PuBuGn',annot=True,fmt=".1f")
plt.ylabel('Top 5 Genre')
plt.xlabel('Certifications')
plt.title('Genre wise - rating distibution')
plt.show()
```



Insights: -

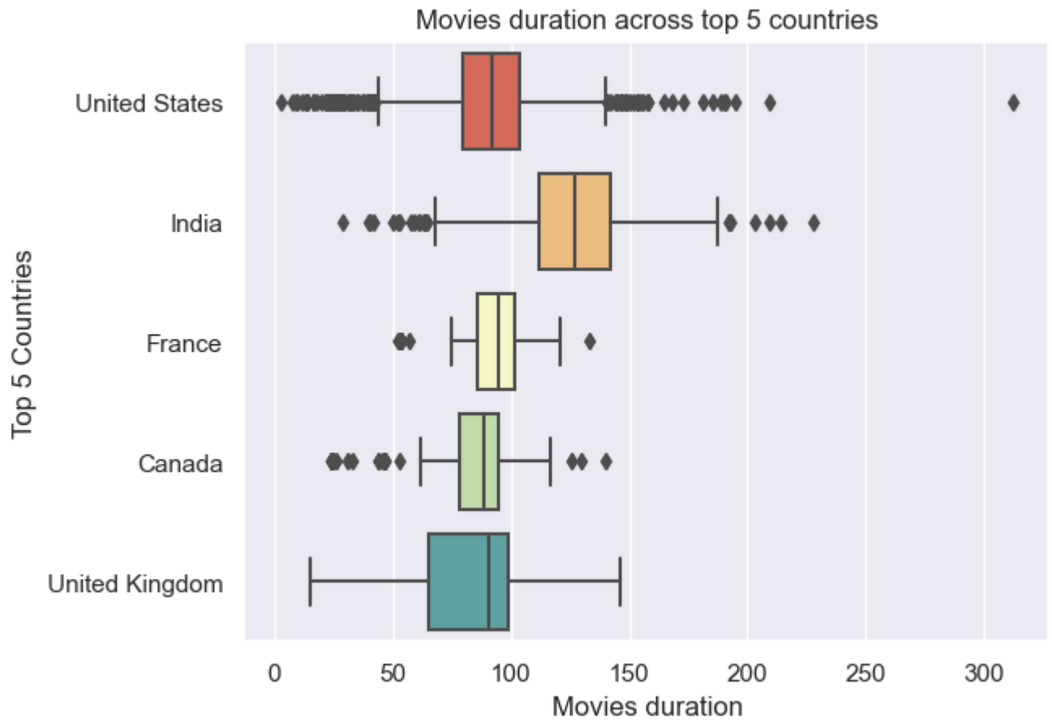
- In **Documenteries**, most popular certifications: - **TV-MA**, **TV-14**, and **TV-PG**.
- In **Dramas** and **Comedies**, most popular certifications: - **R**, **PG-13**, and **TV-MA**

Duration analysis

1) Movies duration in Top 5 countries

In [174…
```
cntry_duration_m=cntry.merge(duration_movie,how='inner')
cntry_duration_m=cntry_duration_m[['title','country','duration']]
cntry_duration_m=cntry_duration_m[cntry_duration_m['country'].isin(top5_cntry)]
```

In [175…
```
sns.boxplot(data=cntry_duration_m,x='duration', y='country',palette='Spectral')
plt.ylabel('Top 5 Countries')
plt.xlabel('Movies duration')
plt.title('Movies duration across top 5 countries')
plt.show()
```



Insights: -

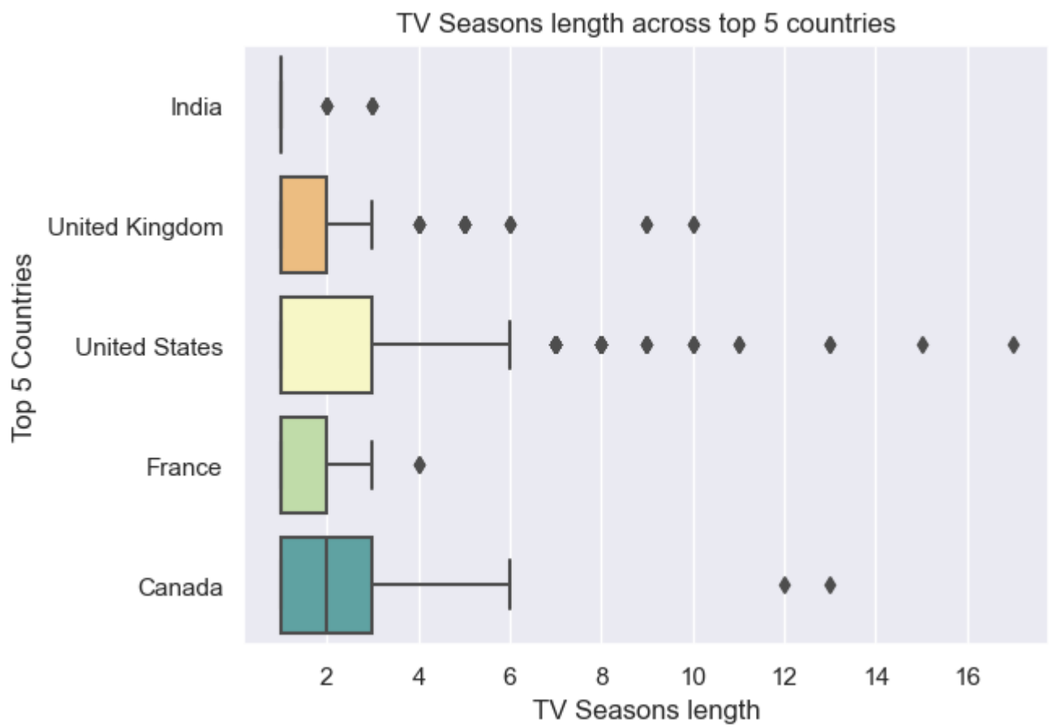
- In **India** median movie duration is **~2 Hours (120 min)**.
- In **US, UK, France and Canada** median movie duration is **~90 min**.

2) Tv shows duration in Top 5 countries

In [176…
```
cntry_duration_tv=cntry.merge(duration_tv,how='inner')
cntry_duration_tv=cntry_duration_tv[['title','country','seasons']]
cntry_duration_tv=cntry_duration_tv[cntry_duration_tv['country'].isin(top5_cntry)]
```

In [177…
```
sns.boxplot(data=cntry_duration_tv,x='seasons', y='country',palette='Spectral')
plt.ylabel('Top 5 Countries')
```

```
plt.title('TV Seasons length across top 5 countries')
plt.show()
```



- We can't infer anything from the visualization, as the higher number of seasons of a particular show defines it's success, and there are very few shows which runs for long. Thus, we can't analyse the season instead in future we can analyse the duration of each episodes and total episodes in a season for each show.

Date analysis

- Date on which which TV Shows/Movies being added to the Platform.

1) Date of content addition on platform in Top 5 countries

In [178…

```
cntry_date=cntry.merge(df,how='inner')
cntry_date=cntry_date[cntry_date['country'].isin(top5_cntry)==True]
cntry_date=cntry_date[['country','date_added','type']]
cntry_date
```

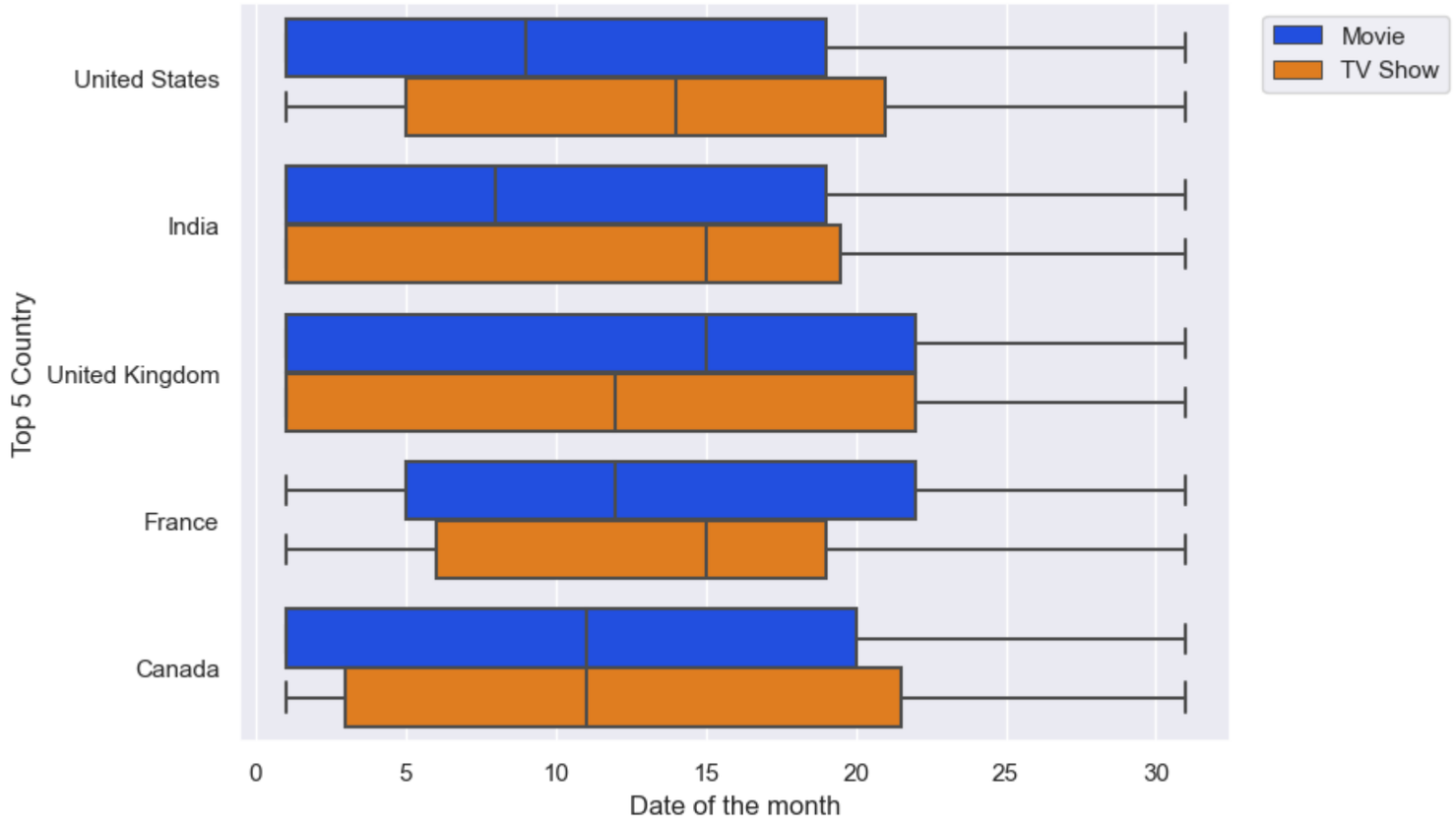
Out[178]:

	country	date_added	type
0	United States	2021-09-25	Movie
4	India	2021-09-24	TV Show
7	United Kingdom	2021-09-24	TV Show
8	United States	2021-09-24	Movie
13	United States	2021-09-22	TV Show
...	...	...	...
7484	India	2018-02-15	Movie
7486	United States	2019-11-20	Movie
7488	United States	2019-11-01	Movie
7489	United States	2020-01-11	Movie
7490	India	2019-03-02	Movie

4514 rows × 3 columns

In [179…

```
plt.figure(figsize=(8,6))
sns.boxplot(data=cntry_date,x=cntry_date['date_added'].dt.day,y=cntry_date['country'],hue='type',palette='bright')
plt.ylabel('Top 5 Country')
plt.xlabel('Date of the month')
plt.legend(loc='upper left',bbox_to_anchor=(1.02, 1))
plt.show()
```



Insights: -

- In **India** and **US** median date of release on platform is **~8th of the Month : Movies** and **~15th of the Month : TV Shows**.
- In **Canada** median date of release on platform is **~11th of the Month : both Movies and TV Shows**.
- In **UK**, median date of release on platform is **15th : Movies** and **12th : TV Shows**.
- In **France** median date of release on platform is **11th : Movies** and **15th : TV Shows**.

2) Month of content addition on platform in Top 5 countries

In [180…

```
cntry_date['month']=cntry_date['date_added'].dt.month_name()
cntry_date
```

Out[180]:

	country	date_added	type	month
0	United States	2021-09-25	Movie	September
4	India	2021-09-24	TV Show	September
7	United Kingdom	2021-09-24	TV Show	September
8	United States	2021-09-24	Movie	September
13	United States	2021-09-22	TV Show	September
...	...	...	...	...
7484	India	2018-02-15	Movie	February
7486	United States	2019-11-20	Movie	November
7488	United States	2019-11-01	Movie	November
7489	United States	2020-01-11	Movie	January
7490	India	2019-03-02	Movie	March

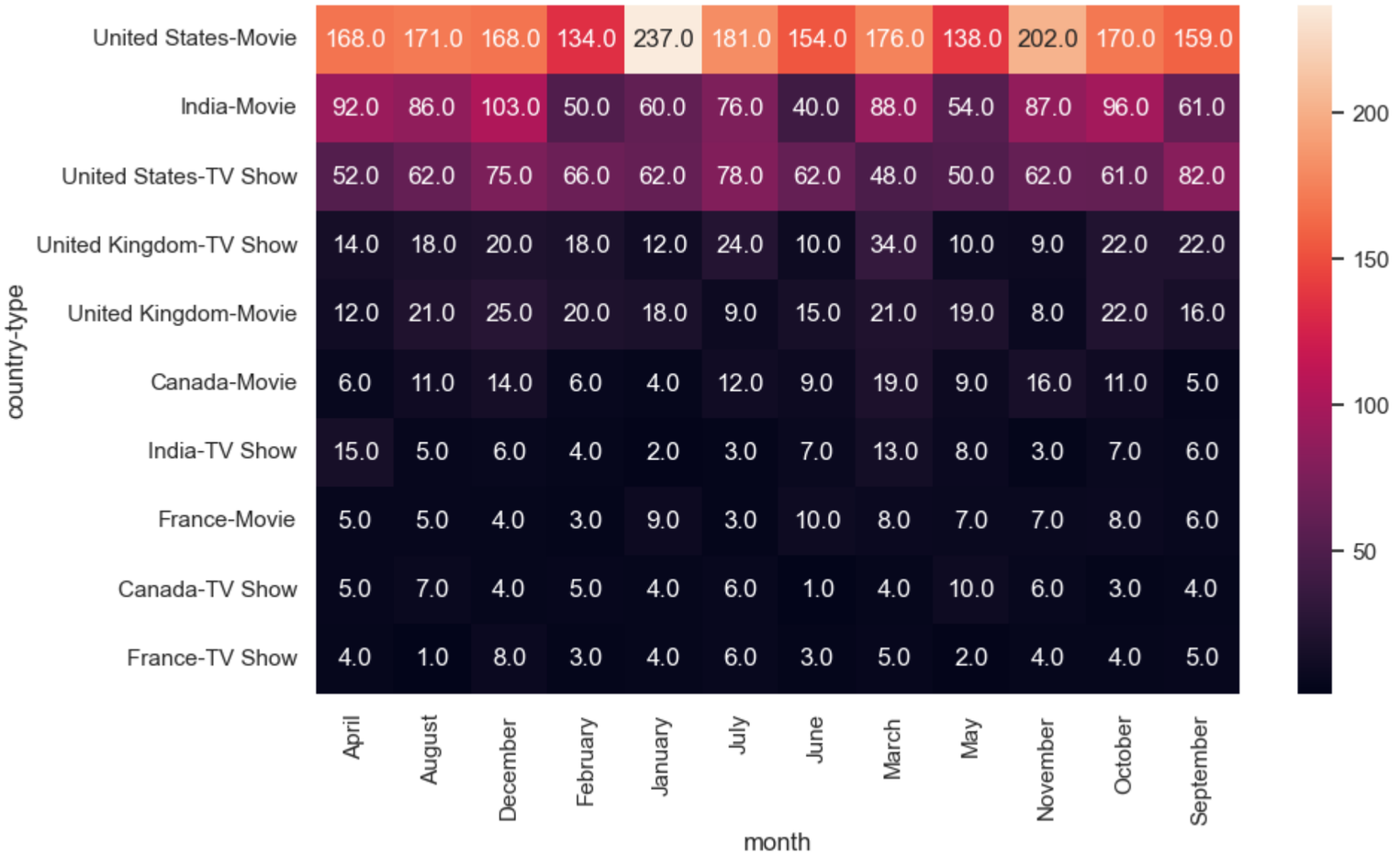
4514 rows × 4 columns

In [181…
```
cntry_date.drop('date_added',axis=1,inplace=True)
```

In [182…
```
cntry_date=cntry_date.groupby(['country','type','month']).size().reset_index(name='count').sort_values(by=['count'],ascending=False)
```

In [183…
```
cntry_date=cntry_date.pivot(index='month',columns=['country','type'],values='count')
```

In [184…
```
plt.figure(figsize=(10,6))
sns.heatmap(cntry_date.T,annot=True,fmt=".1f")
plt.show()
```



Insights: -

- In **India**,
  - **Movies : December, October, April** (Months of festivals : Holi, Diwali, Christmas and New-Year).
  - **TV Shows : April, March**
- In **US**,
  - **Movies : December, January, July** (Month of Holidays : New-year, Christmas and Independence day).
  - **TV Shows : September, July, December**
- In **UK**,
  - **Movies : December** and **TV Shows : March, July**
- In **Canada**,
  - **Movies : March and November** and **TV Shows : May**
- In **France**,
  - **Movies : June, January** and **TV Shows : December**

Country wise Genre and rating distribution

In [185…

```
genre_cert=genre.merge(df,how='left',left_on='title',right_on='title')
e_x','certification']]
```

In [186…
```
genre_cert_cntry=cntry.merge(genre_cert,how='left',left_on='title',right_on='title')
```

In [187…
```
genre_cert_cntry.rename({'genre_x':'genre'},axis=1,inplace=True)
```

In [188…
```
genre_cert_cntry=genre_cert_cntry.groupby(['country','genre','certification']).size().reset_index(name='count').sort_values(by='count',ascending=False)
```

In [189…
```
top3_cntry=cntry.loc[cntry['country']!='Unknown country','country'].value_counts().iloc[:3].index
top3_cntry
```

Out[189]:
```
Index(['United States', 'India', 'United Kingdom'], dtype='object')
```

In [190…
```
genre_cert_cntry=genre_cert_cntry[genre_cert_cntry['country'].isin(top3_cntry)==True]
```

In [191…
```
#Considering at least 25 movies/tv shows per country per genre per certification
genre_cert_cntry=genre_cert_cntry[genre_cert_cntry['count']>25]
```

In [192…
```
genre_cert_cntry=genre_cert_cntry.pivot(index='genre',columns=['country','certification'],values='count')
```

In [193…
```
genre_cert_cntry=genre_cert_cntry.sort_index(axis=1,level=['country','certification'])
```

In [194…
```
genre_cert_cntry.fillna(0,inplace=True)
```

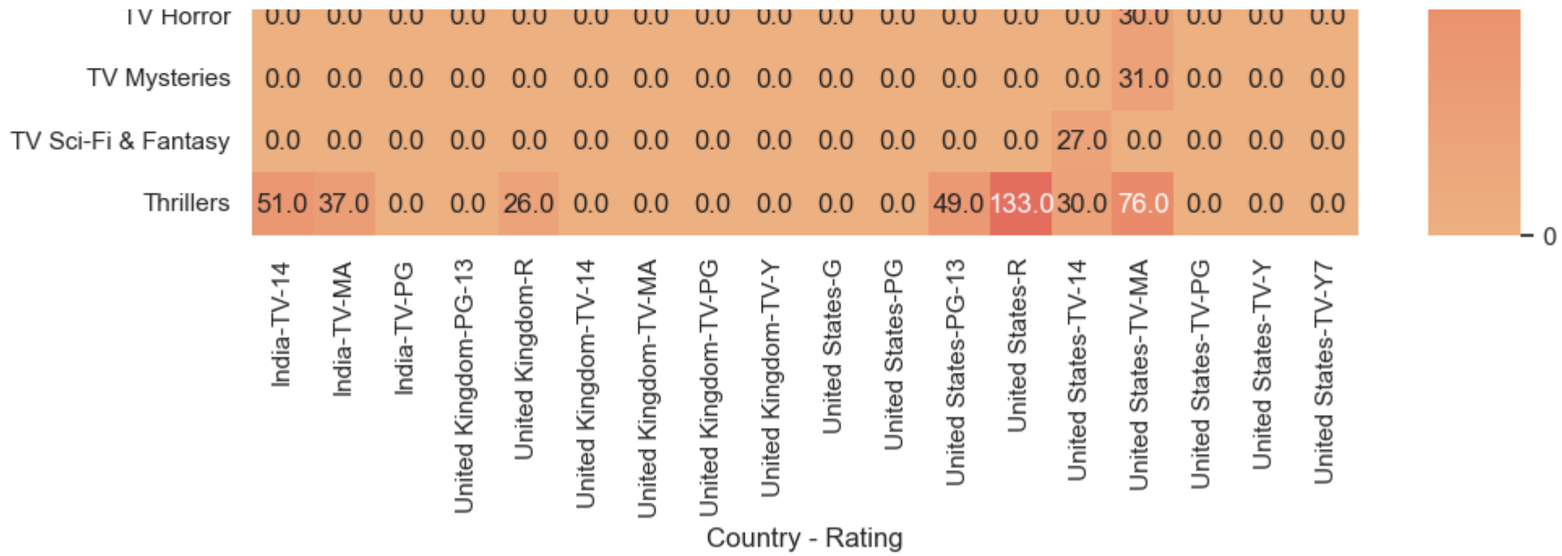
In [195…
```
genre_cert_cntry=genre_cert_cntry.astype('int')
```

In [200…
```
plt.figure(figsize=(11,15))
ax=sns.heatmap(data=genre_cert_cntry,annot=True,fmt=".1f",cmap='flare')
plt.xlabel('Country - Rating')
plt.ylabel('Genres')
plt.title('Country wise : Genre and rating distribution')
plt.show()
```

Country wise : Genre and rating distribution

Genres																		
Action & Adventure	104.0	0.0	0.0	26.0	42.0	0.0	0.0	0.0	0.0	0.0	27.0	139.0	169.0	0.0	44.0	0.0	0.0	0.0
British TV Shows	0.0	0.0	0.0	0.0	0.0	41.0	95.0	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Children & Family Movies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.0	169.0	0.0	0.0	0.0	0.0	52.0	47.0	50.0
Classic Movies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	0.0	0.0	0.0	0.0	0.0
Comedies	199.0	51.0	65.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	134.0	149.0	156.0	46.0	100.0	26.0	0.0	0.0
Crime TV Shows	0.0	0.0	0.0	0.0	0.0	0.0	33.0	0.0	0.0	0.0	0.0	0.0	0.0	34.0	109.0	0.0	0.0	0.0
Cult Movies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	31.0	0.0	0.0	0.0	0.0	0.0
Documentaries	0.0	0.0	0.0	0.0	0.0	39.0	35.0	29.0	0.0	0.0	0.0	28.0	0.0	120.0	194.0	93.0	0.0	0.0
Docuseries	0.0	0.0	0.0	0.0	0.0	0.0	27.0	34.0	0.0	0.0	0.0	0.0	0.0	48.0	97.0	40.0	0.0	0.0
Dramas	379.0	167.0	88.0	44.0	78.0	0.0	39.0	0.0	0.0	0.0	61.0	153.0	311.0	64.0	185.0	36.0	0.0	0.0
Horror Movies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.0	85.0	0.0	60.0	0.0	0.0	0.0
Independent Movies	67.0	74.0	0.0	0.0	38.0	0.0	0.0	0.0	0.0	0.0	0.0	26.0	169.0	29.0	146.0	0.0	0.0	0.0
International Movies	508.0	213.0	118.0	27.0	38.0	0.0	58.0	0.0	0.0	0.0	0.0	0.0	32.0	29.0	52.0	0.0	0.0	0.0
International TV Shows	0.0	32.0	0.0	0.0	0.0	0.0	69.0	26.0	0.0	0.0	0.0	0.0	0.0	0.0	48.0	0.0	0.0	0.0
Kids' TV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	29.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	75.0	91.0
LGBTQ Movies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	34.0	0.0	0.0	0.0
Music & Musicals	63.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	63.0	0.0	0.0	0.0
Reality TV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	38.0	39.0	34.0	0.0	0.0
Romantic Movies	72.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	80.0	46.0	0.0	33.0	0.0	0.0	0.0
Sci-Fi & Fantasy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	84.0	37.0	0.0	0.0	0.0	0.0	0.0
Sports Movies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.0	0.0	0.0	0.0
Stand-Up Comedy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	175.0	0.0	0.0	0.0
Stand-Up Comedy & Talk Shows	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.0	0.0	0.0	0.0
TV Action & Adventure	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.0	0.0	0.0	0.0
TV Comedies	0.0	0.0	0.0	0.0	0.0	0.0	31.0	0.0	0.0	0.0	0.0	0.0	0.0	54.0	112.0	38.0	0.0	0.0
TV Dramas	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	77.0	138.0	0.0	0.0	0.0

500
400
300
200
100



Insights:-

- In **US** : -
  - Among **Dramas** -> R rated
  - Among **Comedies** -> R-Rated
  - Among **Documenteries** -> TV-MA Rated
- In **India** : -
  - Among **Internation Movies** -> TV-14 rated
  - Among **Dramas** -> TV-14 -Rated
  - Among **Comedies** -> TV-14 Rated
- In **UK**
  - Among **Dramas** -> R-Rated
  - Among **Internation Movies** -> Mature-adult (TV-MA) rated
  - Among **International TV shows** -> TV-14 -Rated
  - Among **Documenteries** -> TV-14 & TV-MA Rated

Insights

Basic Analysis

- Majorly, **movies** are produced.
- **Rajiv Chilaka** is the director with the most movies.
- **Anupam Kher** is the actor with the most movies.
- **United States** has produced most of the movies.
- Top 5 countries in terms of producing content: - **US, India, UK, Canada, France**
- Mature Audiance only (**TV-MA**) certification type movies have been produced the most.
- Top 5 countries in terms of producing content: - **International movies, Dramas, Comedies, International TV shows, Documentries**.
- Most preferred Combination of genres - **International Dramas**

- Most present content is released after 2000's.

Streaming date wise insights

- Most of the content added is from **2014- 2021**, median of which is **2019**.
- Most content is added in **July** and **December**.
- Least content is added in **February** and **May**.
- Most of the content added on platform on **1st of the Month**.
- In **India** median date of release on platform is **~8th of the Month : Movies** and **~15th of the Month : TV Shows**.
  - **Movies : December, October, April** (Months of festivals : Holi, Diwali, Christmas and New-Year).
  - **TV Shows : April, March**
- In **US** median date of release on platform is **~8th of the Month : Movies** and **~15th of the Month : TV Shows**.
  - **Movies : December, January, July** (Month of Holidays : New-year, Christmas and Independence day).
  - **TV Shows : September, July, December**
- In **UK**, median date of release on platform is **15th : Movies** and **12th : TV Shows**.
  - **Movies : December**
  - **TV Shows : March, July**

Rating wise insights

- In all the top countries Mature Adult content i.e., **TV-MA** rating is popular.
- In **India**, The most popular content is of >14 rating Year old ie., **TV-14**. After that **TV-MA** and **TV-PG** rating.
- In **Documenteries**, most popular certifications: - **TV-MA**, **TV-14**, and **TV-PG**.
- In **Dramas** and **Comedies**, most popular certifications: - **R**, **PG-13**, and **TV-MA**

Duration wise insights

- Most of the TV shows are having <= 2 Seasons.
- Most of the movies are in the range of duration: - 40 min to 160 min with median of **95 min**.
- In **India** median movie duration is **~2 Hours (120 min)**.
- In **US, UK, France and Canada** median movie duration is **~90 min**.

Country wise Insights (Top 3 countries)

US

- Type of content on top: - **MOVIES**
- Top genres are: - **DRAMAS**, **Comedies**, **Documentries**.

- Top actors in USA: - **Tara strong**, **Sam L Jackson**, **Fred Tatasciore**, **Adam Sandler**, **James Franco** and **Nicolas Cage**.

  - Among **Comedies** -> **Adam Sandler** on top.
  - Among **Children & Family movies** -> **Fred Tatasciore** on top.
  - Among **Dramas** -> **James Franco** and **Sam L Jackson** on top.
  - Among **Kid's TV** -> **Tara Strong** on top.
  - Among **Action & Adventure** -> **Samuel L. Jackson** on top.
- Top directors in US: - **Marcus Raboy**, **Jay Karas**, **Jay Chapman**, **Martin Scorsese**, **Steven Spielberg** and **Ddon Michael Paul**.

sese, and **Steven Spielberg**.

    - Among **Action & Adventure, Children & Family movies, Dramas** -> **Steven Spielberg**.
    - Among **Comedies** -> **Robert Rodriguez**
- Certifications on top: -
    - Among **Dramas** -> R rated
    - Among **Comedies** -> R-Rated
    - Among **Documenteries** -> TV-MA Rated

India

- Type of content on top: - **MOVIES**
- Top genres are: - **INTERNATIONAL MOVIES**, **Dramas**, **Comedies**.
- Top actors in India: - **Anupam Kher**, **Shah Rukh Khan**, **Naseeruddin Shah**, **Late Mr. Om Puri**, **Akshay Kumar** and **Paresh Rawel**.
    - Among **International movies** -> **Anupam Kher** and **Shahrukh Khan** are top.
    - Among **Dramas** -> **Anupam Kher**, **Naseeruddin shah** and **Shahrukh Khan** are top.
    - Among **Comedies** -> **Anupam Kher**, **Akshay Kumar**, and **Shahrukh Khan** are top.
- Top directors in India: - **David Dhawan**, **Anurag Kashyap**, **Umesh Mehra**, **Dibakar Banerjee**, **Priyadarshan** and **Ram Gopal Verma**.
    - Among **Comedies** -> **David Dhawan** on top.
    - Among **Dramas** -> **Dibakar Banerjee**, **Umesh Mehra**, and **Anurag Kashyap** on top.
    - Among **Thriller** -> **Anurag Kashyap** on top.
    - Among **Musicals** -> **David Dhawan** on top.
    - Among **Romantics** -> **David Dhawan** on top.
- Certifications on top: -
    - Among **Internation Movies** -> TV-14 rated
    - Among **Dramas** -> TV-14 -Rated
    - Among **Comedies** -> TV-14 Rated

UK

- Type of content on top: - **TV Shows**
- Top genres are: - **DRAMAS**, **International movies**, **International TV shows** and **Documenteries**
- Top actors in UK: - **David Attenborough**, **John Cleese**, **Micheal Palin**, **Eric Idle**, **Terry Jones** and **Terry Gilliam**.
    - Among **Dramas** -> **Johnny Depp**, **Brendan Gleeson**,**Eddie Marson**, and **HElena Bonham Carter**
    - Among **International Movies** -> **Samuel West**
    - Among **International TV Shows** -> **Olivia Colnman**, and **Rorry Kinnear**
    - Among **Documentaries** -> **David Attenborough** on top.
- Top directors in UK in India: - **Alastair Fothergill**, **Edward cotterill**, **Jerry Rothwell**, **Blair Simmons**, **Vince Marcello**, **Matin campbell**, and **Tom Hooper**.
    - Among **Dramas** -> **Martin Scorsese** on top.
    - Among **International movies** -> **Top Hooper** on top.
    - Among **Internation TV Shows** -> **Alastair Fothergill** on top.
    - Among **Documentaries** -> **Edward Cotterill** on top.
- Certifications : -

- Among **Dramas** -> R-Rated
- Among **Internation Movies** -> Mature-adult (TV-MA) rated
- Among **International TV shows** -> TV-14 -Rated
- Among **Documenteries** -> TV-14 & TV-MA Rated

Recommendations

US

- Produce **Movies**
  - **Dramas** with
    - Directors like: - **Martin Scorsese** and **Steven Spielberg**
    - Actors like: - **James Franco** and **Sam L Jackson**
    - Having duration of **~90 mins** and **R-rated** certification.
  - **Comedies** with
    - Director: - **Robert Rodriguez**
    - Actor like: - **Adam Sandler**
    - Having duration of **~90 mins** and **R-rated** certification.
- Stream it in December, January, July (Month of Holidays : New-year, Christmas and Independence day) on 15th of the Month.

India

- Produce **Movies**
  - **International and Drama** movies with
    - Directors like: - **Dibakar Banerjee, Umesh Mehra, and Anurag Kashyap**
    - and actors like: - **Anupam Kher, Naseeruddin shah and Shahrukh Khan**
  - **Comedies** with
    - Director: - **David Dhawan**
    - and actors like: - **Anupam Kher, Akshay Kumar, and Shahrukh Khan**
- Having duration of **~120 mins** and **TV-14** certification.
- Stream them in Holiday season month like - **December, October or April** on **8th** of the Month.

UK

- Produce **TV Show**
  - **Dramas** with
    - Directors like: - **Martin scorsese**
    - Actors like: - **Johnny Depp, Brendan Gleeson,Eddie Marson, and Helena Bonham Carter**
  - International TV Show with

    - Director: - Alastair Fothergill
    - Actors like: - Olivia Colnman, and Rorry Kinnear
    - Having TV-14 rated certification.
- Stream them in March and July on 12th of the Month.

THE END
