


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

```
df = pd.read_csv("netflix_data.csv")
df
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train l...
...

```
df.head()
```

	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...
		TV		Julien	Sami Bouajila, Tracy		September				Crime TV Shows	To protect his family from a

```
df.shape
```

(8807, 12)

```
df.columns
```

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

```
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
```

```
df.describe()
```

```

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

```
df.isnull().sum().sort_values(ascending=False)
```

```

director      2634
country        831
cast          825
date_added     10
rating          4
duration        3
show_id         0
type            0
title           0
release_year    0
listed_in       0
description     0
dtype: int64
```

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

```

director      29.91
country        9.44
cast          9.37
date_added     0.11
rating         0.05
duration       0.03
show_id        0.00
type           0.00
title          0.00
release_year   0.00
listed_in      0.00
description    0.00
dtype: float64
```

```
df["director"].value_counts()
```

```

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
Mozes Singh          1
Name: director, Length: 4528, dtype: int64

```

```
df.type.value_counts()
```

```

Movie      6131
TV Show    2676
Name: type, dtype: int64

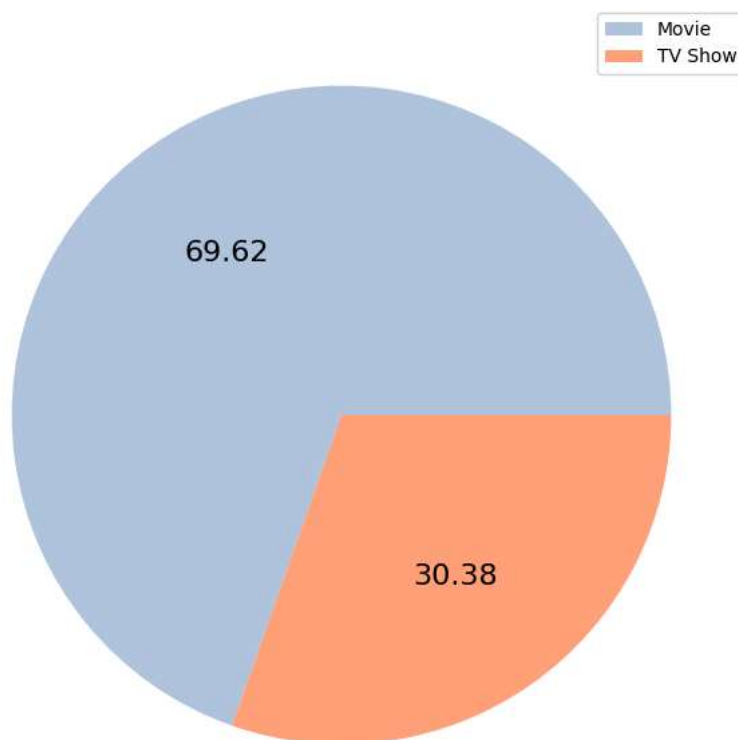
```

```
plt.figure(figsize=(10,8))
```

```

plt.pie(df.type.value_counts(),
        labels = df.type.value_counts().index,
        labeldistance = None, autopct="%.2f",
        textprops = {'fontsize': 16,},
        colors = ['lightsteelblue','lightsalmon' ] )
plt.legend()
plt.show()

```



```
df.rating.value_counts()
```

```

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

```

```
df.country.value_counts()
```

```

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

```

```
df.country.value_counts().head(10)
```

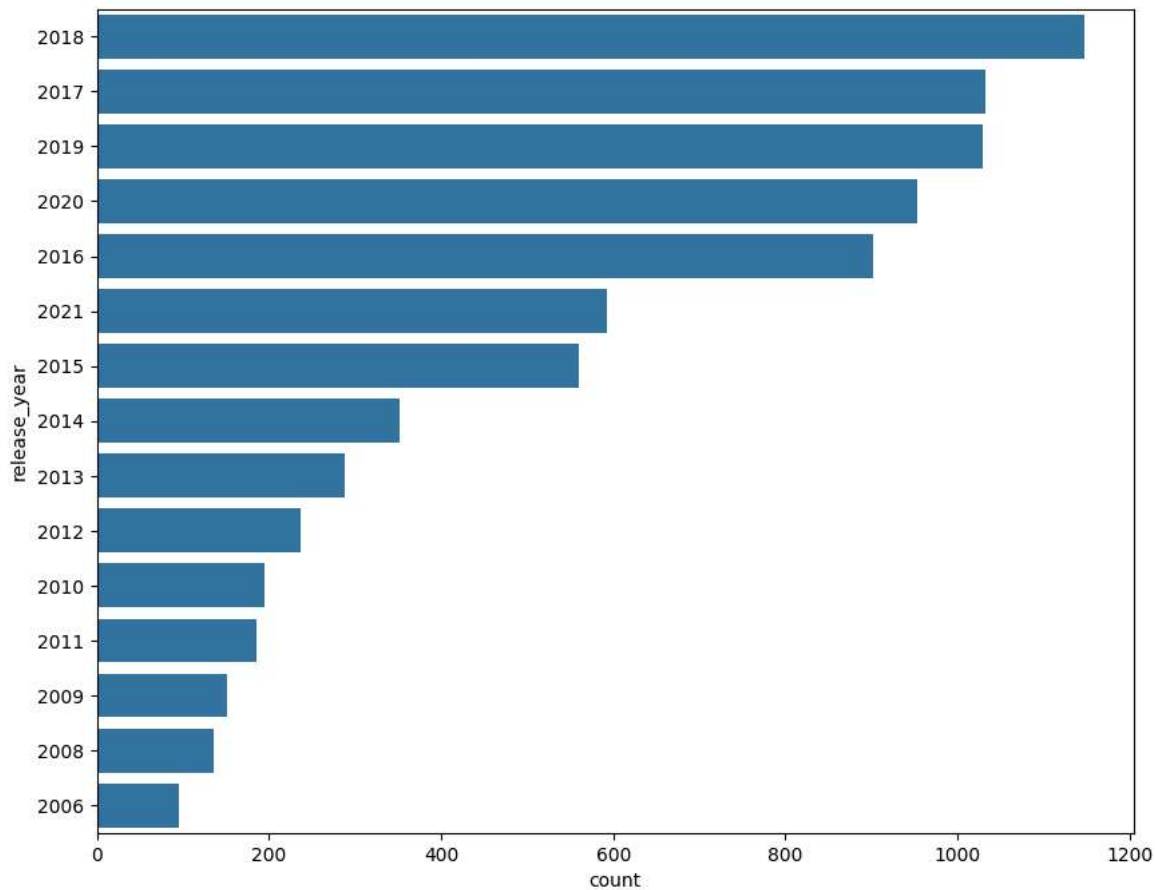
```

United States      2818
India              972
United Kingdom     419
Japan              245
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Egypt              106
Name: country, dtype: int64

```

```
plt.figure(figsize=(10,8))
```

```
ax = sns.countplot(y="release_year", data=df, order=df.release_year.value_counts().index[0:15])
```



```
df.director.value_counts().head(10)
```

```

Rajiv Chilaka      19
Raúl Campos, Jan Suter  18
Marcus Raboy       16
Suhas Kadav        16
Jay Karas          14
Cathy Garcia-Molina 13
Martin Scorsese     12

```

```

Youssef Chahine      12
Jay Chapman          12
Steven Spielberg     11
Name: director, dtype: int64

```

```
df.listed_in.value_counts().head(10)
```

```

Dramas, International Movies      362
Documentaries                    359
Stand-Up Comedy                   334
Comedies, Dramas, International Movies 274
Dramas, Independent Movies, International Movies 252
Kids' TV                          220
Children & Family Movies          215
Children & Family Movies, Comedies 201
Documentaries, International Movies 186
Dramas, International Movies, Romantic Movies 180
Name: listed_in, dtype: int64

```

```
df.listed_in.value_counts().tail()
```

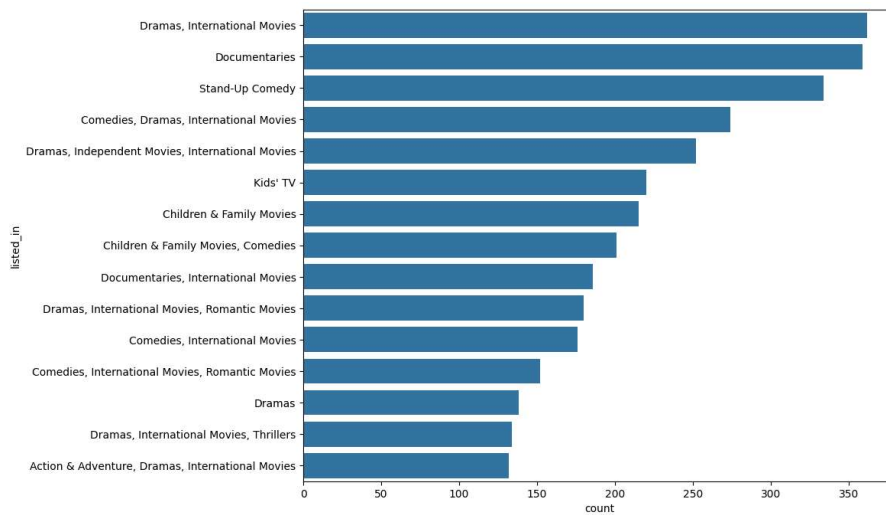
```

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, dtype: int64

```

```
plt.figure(figsize=(10,8))
```

```
ax = sns.countplot(y="listed_in", data=df, order=df.listed_in.value_counts().index[0:15])
```



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

```

director      29.91
country       9.44
cast          9.37
date_added    0.11
rating        0.05
duration      0.03
show_id       0.00
type          0.00
title         0.00

```

```
release_year    0.00
listed_in       0.00
description     0.00
dtype: float64
```

```
round(df.isnull().sum())
```

```
show_id        0
type           0
title          0
director      2634
cast          825
country       831
date_added     10
release_year   0
rating         4
duration       3
listed_in      0
description    0
dtype: int64
```

```
df.dropna(subset=["rating", "duration"], axis=0, inplace=True)
```

```
df.shape
```

```
(8800, 12)
```

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

```
director      29.90
country       9.43
cast          9.38
date_added    0.11
show_id       0.00
type          0.00
title         0.00
release_year  0.00
rating        0.00
duration      0.00
listed_in     0.00
description   0.00
dtype: float64
```

```
df.dropna(subset=["date_added"], axis=0, inplace=True)
```

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

```
director      29.82
country       9.43
cast          9.39
show_id       0.00
type          0.00
title         0.00
date_added    0.00
release_year  0.00
rating        0.00
duration      0.00
listed_in     0.00
description   0.00
dtype: float64
```

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

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

```
director      29.82
cast          9.39
show_id       0.00
type          0.00
title         0.00
country       0.00
date_added    0.00
release_year  0.00
rating        0.00
duration      0.00
listed_in     0.00
```

```
description      0.00
dtype: float64
```

```
df.country.value_counts().head()
```

```
United States    2809
India            972
Unknown          829
United Kingdom   418
Japan            243
Name: country, dtype: int64
```

```
df.cast.value_counts().head(10)
```

```
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
Kevin Hart               6
Craig Sechler            6
Michela Luci, Jamie Watson, Eric Peterson, Anna Claire Bartlam, Nicolas Aqui, Cory Doran, Julie Lemieux, Derek McGrath  6
Iliza Shlesinger         5
Jim Gaffigan             5
Name: cast, dtype: int64
```

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

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

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

```
show_id      0.0
type         0.0
title        0.0
director     0.0
cast         0.0
country      0.0
date_added   0.0
release_year 0.0
rating       0.0
duration     0.0
listed_in    0.0
description   0.0
dtype: float64
```

```
df["title"]
```

```
0      Dick Johnson Is Dead
1      Blood & Water
2      Ganglands
3      Jailbirds New Orleans
4      Kota Factory
...
8802      Zodiac
8803      Zombie Dumb
8804      Zombieland
8805      Zoom
8806      Zubaan
Name: title, Length: 8790, dtype: object
```

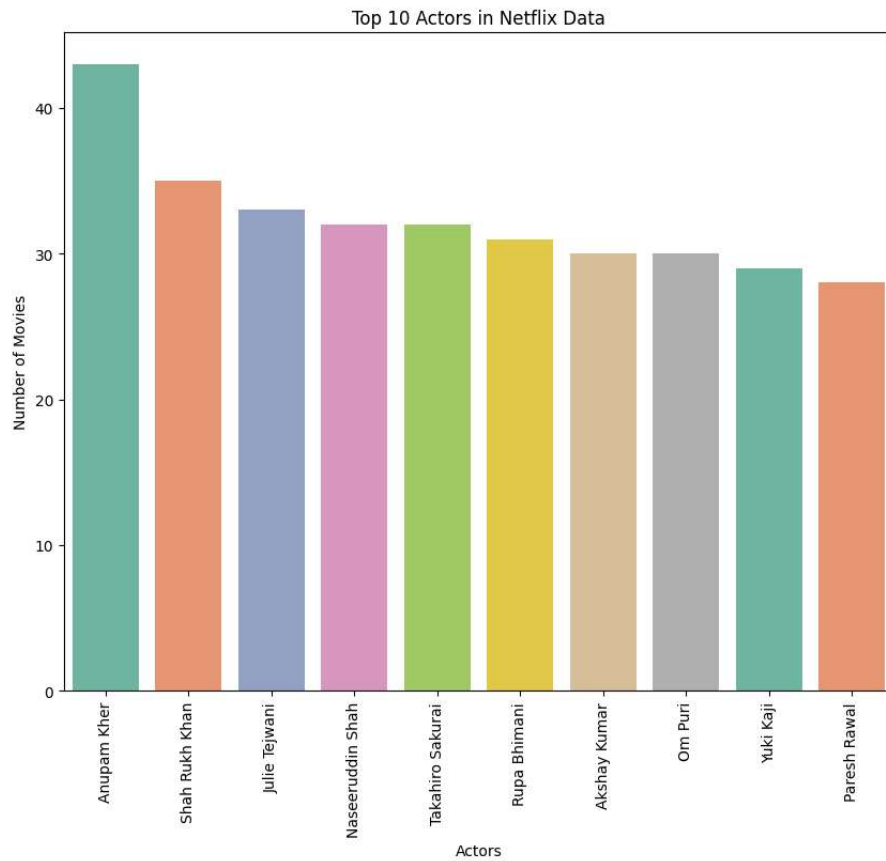
```
cast_shows = df[df["cast"] != "No Cast"].set_index("title").cast.str.split(", ", expand=True).stack().reset_index(level=1, drop=True)
top_10_actors = cast_shows.value_counts().head(10)
```

```
plt.figure(figsize=(10, 8))
sns.barplot(x=top_10_actors.index, y=top_10_actors.values, palette='Set2')
plt.xticks(rotation=90)
plt.xlabel("Actors")
plt.ylabel("Number of Movies")
plt.title("Top 10 Actors in Netflix Data")
plt.show()
```

```
<ipython-input-37-2f2bf9865dc4>:5: FutureWarning:
```

```
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
```

```
sns.barplot(x=top_10_actors.index, y=top_10_actors.values, palette='Set2')
```



```
movies_df = df.loc[(df["type"] == "Movie")]
movies_df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Unknown	September 24, 2021	2021
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmiike Ogunlano, Alexandra	United States, Ghana, Burkina Faso, ...	September 24, 2021	1993


```
show_df = df.loc[(df["type"] == "TV Show")]
show_df.head()
```

show_id	type	title	director	cast	country	date_added	release_year	rat
1	s2 TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mababane, Thaban...	South Africa	September 24, 2021	2021	TV
2	s3 TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel	Unknown	September 24, 2021	2021	TV

```
movies_df.duration = movies_df.duration.apply(lambda x: x.replace(" min", "") if 'min' in x else x)
movies_df.head()
```

<ipython-input-40-145cfd8505d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
movies_df.duration = movies_df.duration.apply(lambda x: x.replace(" min", "") if 'min'

show_id	type	title	director	cast	country	date_added	release_year
0	s1 Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020
6	s7 Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Unknown	September 24, 2021	2021
7	s8 Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra	United States, Ghana, Burkina Faso, ...	September 24, 2021	1993

```
movies_df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 6126 entries, 0 to 8806
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  -
0   show_id         6126 non-null   object
1   type            6126 non-null   object
2   title           6126 non-null   object
3   director        6126 non-null   object
4   cast            6126 non-null   object
5   country         6126 non-null   object
6   date_added      6126 non-null   object
7   release_year    6126 non-null   int64
8   rating          6126 non-null   object
9   duration        6126 non-null   object
10  listed_in       6126 non-null   object
11  description      6126 non-null   object
dtypes: int64(1), object(11)
memory usage: 751.2+ KB

movies_df.loc[:,["duration"]] = movies_df.loc[:,["duration"]].apply(lambda x: x.astype('int64'))
movies_df.describe()

<ipython-input-42-8d86ad9fba4d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
movies_df.loc[:,["duration"]] = movies_df.loc[:,["duration"]].apply(lambda x: x.astype('int64'))
<ipython-input-42-8d86ad9fba4d>:1: DeprecationWarning: In a future version, `df.iloc[:, i] = newvals` will attempt to set the values in place if the values are non-numeric.
movies_df.loc[:,["duration"]] = movies_df.loc[:,["duration"]].apply(lambda x: x.astype('int64'))
```

	release_year	duration
count	6126.000000	6126.000000
mean	2013.120144	99.584884
std	9.681723	28.283225
min	1942.000000	3.000000
25%	2012.000000	87.000000
50%	2016.000000	98.000000
75%	2018.000000	114.000000
max	2021.000000	312.000000

```
shortest_movies = movies_df.loc[(movies_df['duration']==np.min(movies_df.duration))]
shortest_movies
```

show_id	type	title	director	cast	country	date_added	release_year	rating
Limbort								

```
longest_movies = movies_df.loc[(movies_df['duration']==np.max(movies_df.duration))]
longest_movies
```

show_id	type	title	director	cast	country	date_added	release_year
Fionn Whitehead							

```
show_df.duration = show_df.duration.apply(lambda x: x.replace(" Season", "") if 'Season' in x else x)
show_df.head(3)
```

```
<ipython-input-53-35f7f7f8bf45>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
 show_df.duration = show_df.duration.apply(lambda x: x.replace(" Season", "") if 'Seasc

show_id	type	title	director	cast	country	date_added	release_year	rating	
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mababalane, Thabane...	South Africa	September 24, 2021	2021	TV

```
show_df.duration = show_df.duration.apply(lambda x: x.replace("Seacon", "") if 'Seacon' in x else x)
show_df.head(3)
```

```
<ipython-input-57-d589482c786d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/user>
 show_df.duration = show_df.duration.apply(lambda x: x.replace("Seacon", "") if 'Seacon'

show_id	type	title	director	cast	country	date_added	release_year	rat	
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mababalane, Thaban...	South Africa	September 24, 2021	2021	TV

```
show_df.loc[:,["duration"]] = show_df.loc[:,["duration"]].apply(lambda x: x.astype('int64'))
show_df.describe()
```

```
<ipython-input-58-f52c766124a1>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c

```
show_df.loc[:,["duration"]] = show_df.loc[:,["duration"]].apply(lambda x: x.astype('int64'))
<ipython-input-58-f52c766124a1>:1: DeprecationWarning: In a future version, `df.iloc[:, i] = newvals` will attempt to set the values in
show_df.loc[:,["duration"]] = show_df.loc[:,["duration"]].apply(lambda x: x.astype('int64'))
```

	release_year	duration
count	2664.000000	2664.000000
mean	2016.627628	1.751877
std	5.735194	1.550622
min	1925.000000	1.000000
25%	2016.000000	1.000000
50%	2018.000000	1.000000
75%	2020.000000	2.000000
max	2021.000000	17.000000

```
show_df.duration.value_counts().tail(10)
```

```
6    33
7    23
```

```

8      17
9      9
10     6
13     2
15     2
12     2
17     1
11     1
Name: duration, dtype: int64

```

```

longest_shows = show_df.loc[(show_df["duration"] >= 13)]
longest_shows

```

	show_id	type	title	director	cast	country	date_added	release_year
	548	s549 TV Show	Grey's Anatomy	No Director	Ellen Pompeo, Sandra Oh, Katherine Heigl, Just...	United States	July 3, 2021	2005
	1354	s1355 TV Show	Heartland	No Director	Amber Marshall, Michelle Morgan, Graham Wardle...	Canada	February 1, 2021	2007

```
longest_shows.rating.value_counts()
```

```

TV-14      4
TV-MA      1
Name: rating, dtype: int64

```

▼ Release Year

```

last_decade = df[["type", "release_year"]]
last_decade = last_decade.rename(columns = {"release_year" : "Release Year"})
last_decade = last_decade[last_decade["Release Year"]>=2010]
last_decade

```

	type	Release Year
0	Movie	2020
1	TV Show	2021
2	TV Show	2021
3	TV Show	2021
4	TV Show	2021
...
8798	Movie	2014
8800	TV Show	2012
8801	Movie	2015
8803	TV Show	2018
8806	Movie	2015

7458 rows × 2 columns

```
last_decade_df = last_decade.groupby("Release Year")["type"].size().reset_index()
last_decade_df = pd.DataFrame(last_decade_df)
last_decade_df
```

	Release Year	type
0	2010	192
1	2011	185
2	2012	236
3	2013	286
4	2014	352
5	2015	555
6	2016	901
7	2017	1030
8	2018	1146
9	2019	1030
10	2020	953
11	2021	592

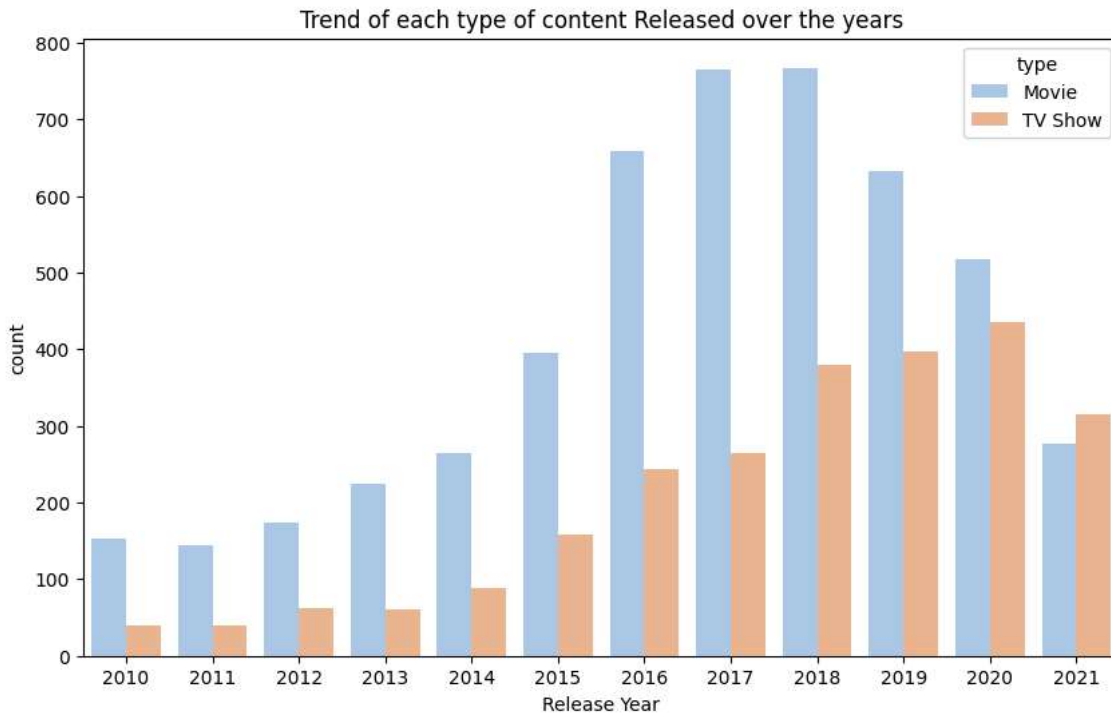
```
last_decade_df.rename(columns = {"type": "Total Content"}, inplace = True)
```

```
last_decade.groupby("Release Year")["type"].value_counts()
```

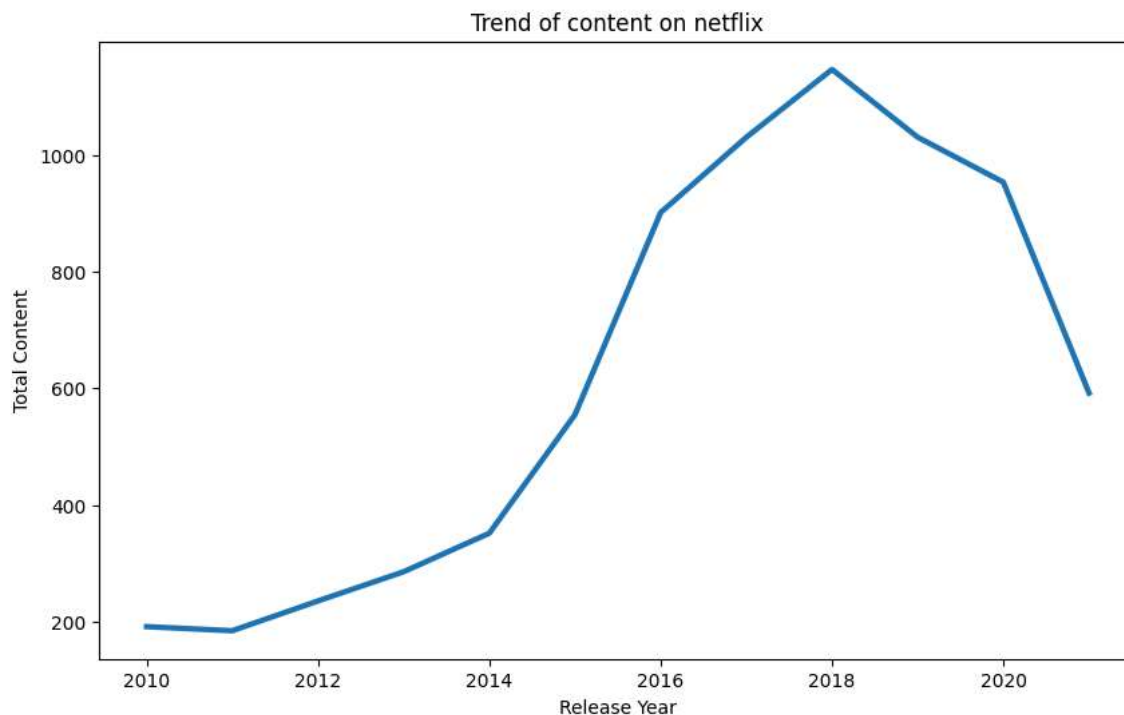
Release Year	type	
2010	Movie	153
	TV Show	39
2011	Movie	145
	TV Show	40
2012	Movie	173
	TV Show	63
2013	Movie	225
	TV Show	61
2014	Movie	264
	TV Show	88
2015	Movie	396
	TV Show	159
2016	Movie	658
	TV Show	243
2017	Movie	765
	TV Show	265
2018	Movie	767
	TV Show	379
2019	Movie	633
	TV Show	397
2020	Movie	517
	TV Show	436
2021	TV Show	315

Movie 277
Name: type, dtype: int64

```
plt.figure(figsize = (10,6))
count_plot = sns.countplot(x = "Release Year", data = last_decade, hue="type",
                             palette= "pastel")
count_plot.set(title = "Trend of each type of content Released over the years")
plt.show()
```



```
plt.figure(figsize = (10,6))
plot_total_content= sns.lineplot(x= "Release Year", y = "Total Content", data = last_decade_df,
                                  linewidth = 3)
plot_total_content.set(xlabel = "Release Year", ylabel = "Total Content",
                       title = "Trend of content on netflix")
plt.show()
```



✓ Countries

```
top_10_countries= df.country.value_counts().head(10)
top_10_countries = pd.DataFrame(top_10_countries)
top_10_countries
```

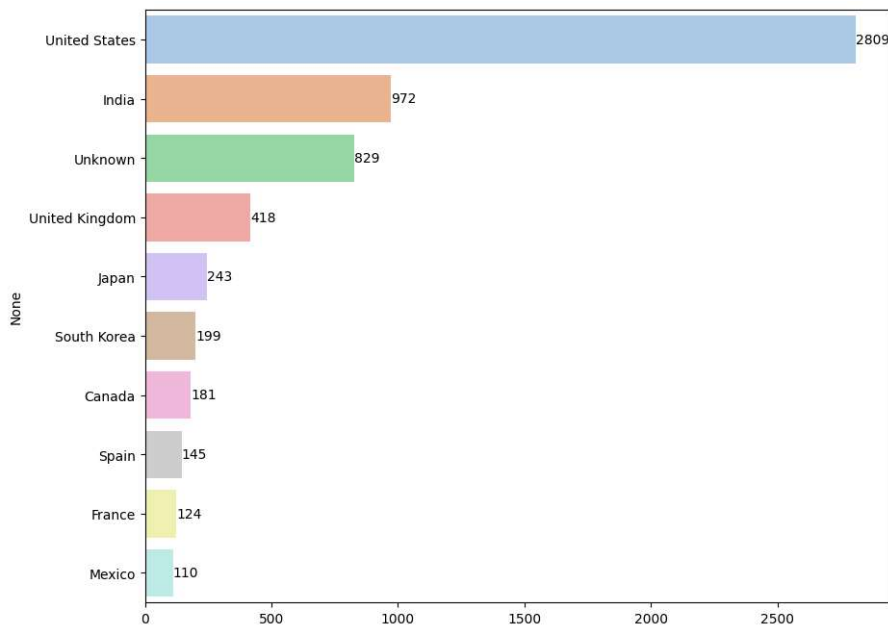
	country
United States	2809
India	972
Unknown	829
United Kingdom	418
Japan	243
South Korea	199
Canada	181
Spain	145
France	124
Mexico	110

```
plt.figure(figsize = (10,8))
country_plot = sns.barplot(x = df.country.value_counts()[ :10].values,
                           y= df.country.value_counts()[ :10].index,palette = "pastel")
for i in country_plot.containers:
    country_plot.bar_label(i)
```

<ipython-input-75-cab6b87401bf>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.

```
country_plot = sns.barplot(x = df.country.value_counts()[ :10].values,
```



Rating

```
df.rating.unique()
```

```
array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',  
      'TV-G', 'G', 'NC-17', 'NR', 'TV-Y7-FV', 'UR'], dtype=object)
```

```
new_catgs = {
```

```
    'TV-PG': 'Parental Guidance',
```

```
    'TV-MA' : 'Mature Audience',
```

```
    'TV-Y7-FV': 'Teens',
```

```
    'TV-Y7': 'Teens',
```

```
    'TV-14': 'Teens',
```

```
    'R': 'Mature Audience',
```

```
    'TV-Y': 'General Audience',
```

```
    'NR': 'Mature Audience',
```

```
    'PG-13': 'Teens',
```

```
    'TV-G': 'General Audience',
```

```
    'PG': 'Teens',
```

```
    'G': 'General Audience',
```

```
    'UR': 'Mature Audience',
```

```
    'NC-17': 'Mature Audience'
```

```
}
```

```
df['rating']=df['rating'].replace(new_catgs)
```

```
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	r
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang...	South Africa	September 24, 2021	2021	I Au
					Sami Bouajila, Tessa...				

```
plt.figure(figsize= (10,6))
sns.countplot(x="rating", data=df, hue="rating")
plt.title("count of Rating by Movie and Shows")
plt.show()
```