In [112]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Data Preprocessing

Data Import

In [113]:
```
df = pd.read_csv('netflix_titles.csv')
df
```

Out[113]:

	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...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

Check Null Values and Deal with Missing Values

```
In [114]: 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
```

```
In [115]: df.isnull().sum()
```

```
Out[115]: 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
```

```
In [116]: for i in df.columns:
              null_rate = df[i].isna().sum() / len(df) * 100
              print(f"{i}'s null rate : {null_rate}%")
```

```
show_id's null rate : 0.0%
type's null rate : 0.0%
title's null rate : 0.0%
director's null rate : 29.908027705234474%
cast's null rate : 9.367548540933349%
country's null rate : 9.435676166685592%
date_added's null rate : 0.11354604292040424%
release_year's null rate : 0.0%
rating's null rate : 0.04541841716816169%
duration's null rate : 0.034063812876121265%
listed_in's null rate : 0.0%
description's null rate : 0.0%
```

Thus it is obvious that null values exist in director, cast, country, data_added, ratings, and duration columns. Moreover, the director, cast, and country columns have the more null values compared to the rest of three.

The supposed reasons for large amount of null values are different for each column.
Director: Some TV shows have multiple directors contribute to the work which makes it difficult to select one person as a main director.
Cast: Might be assumed to 0 as there are too many people to be input in the cast.
Country: Some movies and TV shows might be difficult to determine the country it belongs to because it can be produced by multiple countries.

```
In [117]: df['cast'].replace(np.nan, 'No Data',inplace  = True)
          df['director'].replace(np.nan, 'No Data',inplace  = True)
          df['country'].replace(np.nan, 'No Data',inplace  = True)
          df['rating'].replace(np.nan, 'No Rated',inplace  = True)

          # Drop other null values
          df.dropna(inplace=True)
```

```
In [118]: df.isnull().sum()
```

```
Out[118]: show_id         0
          type            0
          title           0
          director        0
          cast            0
          country         0
          date_added      0
          release_year    0
          rating          0
          duration        0
          listed_in       0
          description     0
          dtype: int64
```

Check Unique Values for Features

```
In [119]: df["type"].unique()
```

```
Out[119]: array(['Movie', 'TV Show'], dtype=object)
```

```
In [120]: df["type"].value_counts()
```

```
Out[120]: Movie      6128
          TV Show    2666
          Name: type, dtype: int64
```

```
In [121]: df["rating"].unique()
```

```
Out[121]: array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
                 'TV-G', 'G', 'NC-17', 'NR', 'No Rated', 'TV-Y7-FV', 'UR'],
                dtype=object)
```

```
In [122]: df["rating"].nunique()
```

```
Out[122]: 15
```

```
In [123]: df['rating'].replace('NR', 'No Rated',inplace  = True)
          df['rating'].replace('UR', 'No Rated',inplace  = True)
          df["rating"].unique()
```

```
Out[123]: array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
                 'TV-G', 'G', 'NC-17', 'No Rated', 'TV-Y7-FV'], dtype=object)
```

```
In [124]: df["country"].nunique()
```

```
Out[124]: 749
```

```
In [125]: df["release_year"].min()
```

```
Out[125]: 1925
```

```
In [126]: df["release_year"].max()
```

```
Out[126]: 2021
```

```
In [127]: df["duration"].nunique()
```

```
Out[127]: 220
```

```
In [128]: df["duration"].unique()
```

```
Out[128]: 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',
```

Separate the time data for date_added column

In [129]:
```
df["date_added"] = pd.to_datetime(df['date_added'])

df['day_added']=df['date_added'].dt.day
df['month_added']=df['date_added'].dt.month_name()
df['year_added'] = df['date_added'].dt.year

df.head(5)
```

Out[129]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	day_added	month_added	y
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Data	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	25	September	
1	s2	TV Show	Blood & Water	No Data	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	24	September	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	No Data	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	24	September	
3	s4	TV Show	Jailbirds New Orleans	No Data	No Data	No Data	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...	24	September	
4	s5	TV Show	Kota Factory	No Data	Mayur More, Jitendra Kumar, Ranjan	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV	In a city of coaching centers known to	24	September	

Download the cleaned data in a csv file ¶

In [130]:
```
df.to_csv('cleaned_data.csv')
```