


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

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

netflix_titles=raw_data_csv.copy()
```

netflix_titles



	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...
...

netflix_titles.shape

(8807, 12)

netflix_titles.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
```

netflix_titles.isnull().sum().sum()

4307

netflix_titles.isnull().sum()/len(netflix_titles)*100

```
show_id      0.000000
type         0.000000
```

```
title            0.000000
director         29.908028
cast             9.367549
country          9.435676
date_added       0.113546
release_year     0.000000
rating           0.045418
duration         0.034064
listed_in        0.000000
description      0.000000
dtype: float64

duplicate_netflix_titles = netflix_titles[netflix_titles.duplicated(keep='first')]

duplicate_netflix_titles
```

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

```
netflix_titles.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
```

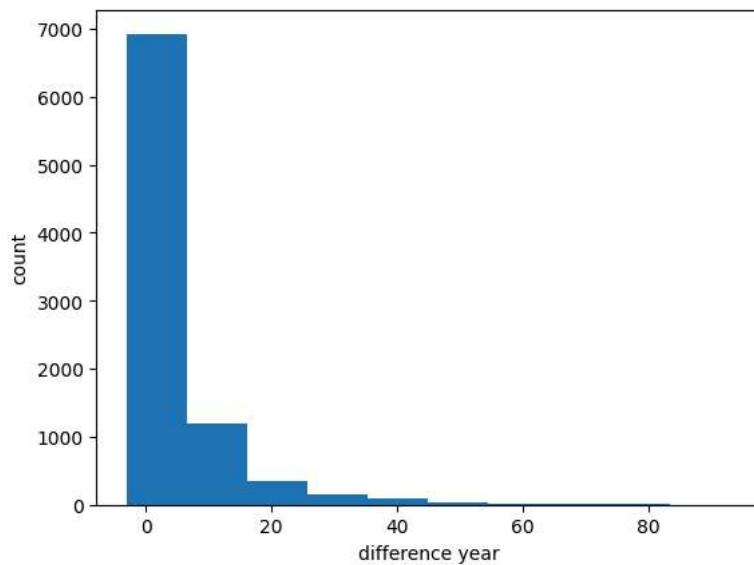
```
netflix_titles['date_added'] = pd.to_datetime(netflix_titles['date_added'])
```

```
netflix_titles.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	2021-09-25	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	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...
		TV		Julien	Sami Bouajila, Tracy						Crime TV Shows	To protect his family from a

```
netflix_titles['date_added-released_year'] = netflix_titles['date_added'].dt.year - netflix_titles['release_year']
```

```
plt.hist(netflix_titles['date_added-released_year'])
plt.xlabel('difference year')
plt.ylabel('count')
plt.show()
```



```
fill_nan = pd.to_datetime(netflix_titles['release_year'].map(str)+'-01-01')
```

```
netflix_titles['date_added'].isna().sum()
```

```
10
```

```
netflix_titles['date_added'] = netflix_titles['date_added'].fillna(netflix_titles['release_year'])
```

```
netflix_titles['date_added'].isna().sum()
```

```
0
```

```
netflix_titles[netflix_titles['show_id']=='s194']
```

show_id	type	title	director	cast	country	date_added	release_year	rating	d
				Jung Hae-in.					

```
netflix_titles['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

netflix_titles = netflix_titles.assign(country = netflix_titles['country'].str.split(',')).explode('country')
netflix_titles['country'] = netflix_titles['country'].str.strip()

netflix_titles['country'].value_counts()

United States      3690
India              1046
United Kingdom      806
Canada              445
France              393
...
Ecuador             1
Armenia              1
Mongolia             1
Bahamas              1
Montenegro           1
Name: country, Length: 123, dtype: int64

netflix_titles['country'].isna().sum()

831

netflix_titles['country'].fillna('unknown',inplace=True)

netflix_titles['country'].isna().sum()

0

netflix_titles[netflix_titles['show_id']=='s194']
```

show_id	type	title	director	cast	country	date_added	release_year	rating	d
193	s194	TV Show	D.P.	NaN	Jung Hae-in, Koo Kyo-hwan, Kim	2021-08-27 00:00:00	2021	TV-MA	1

```
netflix_titles.drop(columns=['date_added-released_year'],inplace=True)

netflix_titles['director'].nunique()

4528

netflix_titles['director'].unique()
```

```
array(['Kirsten Johnson', nan, 'Julien Leclercq', ..., 'Majid Al Ansari',
      'Peter Hewitt', 'Mozez Singh'], dtype=object)

netflix_titles['director'].value_counts()

Rajiv Chilaka      19
Raúl Campos, Jan Suter  18
Martin Scorsese    18
Steven Spielberg   18
Youssef Chahine    17
..
Camille Shooshani  1
Vijay Kumar        1
Phanindra Narsetti  1
Gideon Raff        1
Mozez Singh        1
Name: director, Length: 4528, dtype: int64

netflix_titles=netflix_titles.assign(director=netflix_titles['director'].str.split(',')).explode('director')
netflix_titles['director'] = netflix_titles['director'].str.strip()

netflix_titles['director'].nunique()

4993

netflix_titles['director'].unique()

array(['Kirsten Johnson', nan, 'Julien Leclercq', ..., 'Majid Al Ansari',
      'Peter Hewitt', 'Mozez Singh'], dtype=object)

netflix_titles['director'].value_counts()

Rajiv Chilaka      22
Jan Suter          21
Raúl Campos        19
Martin Scorsese    18
Steven Spielberg   18
..
YC Tom Lee         1
Manu Ashokan       1
Anita Udeep        1
Alberto Arnaut Estrada  1
Mozez Singh        1
Name: director, Length: 4993, dtype: int64

netflix_titles['director'].isna().sum()

2970

netflix_titles[netflix_titles['director'].isna()].head(2)
```

show_id	type	title	director	cast	country	date_added	release_year	rating
1	s2 TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail	South Africa	2021-09-24 00:00:00	2021	TV-M

```
mode_director_by_country = netflix_titles.groupby('country')['director'].apply(lambda a: a.mode())

mode_director_by_country = mode_director_by_country.reset_index().drop(columns='level_1')

mode_director_by_country.drop_duplicates(subset=['country'],keep = 'first',inplace = True)
```

```
mode_director_by_country = pd.Series(mode_director_by_country['director'].values,index = mode_director_by_country['country'])
```

```
mode_director_by_country.index = mode_director_by_country.index.str.strip()
```

```
mode_director_by_country
```

```
country
Afghanistan      Denis Do
Albania          Pieter-Jan De Pue
Algeria          Antonio Morabito
Angola           Maïwenn
Venezuela       Chris Roland
Vietnam          ...
West Germany    Aaron Woolf
Zimbabwe        Bao Nhan
unknown         Christian Herrendoerfer
Length: 119, dtype: object
```

```
netflix_titles['director'] = netflix_titles.apply(lambda x: mode_director_by_country[x['country']] if pd.isnull(x['director']) and x['country'] in mode_director_by_country else x['director'], axis=1)
```

```
netflix_titles['director'].fillna('Unknown_director', inplace=True)
```

```
netflix_titles['director'].isna().sum()
```

```
0
```

```
netflix_titles[netflix_titles['director'].isna()]
```

```
show_id  type  title  director  cast  country  date_added  release_year  rating  duration
0  S001  TV  The Shawshank Redemption  Frank Darabont  Tim Robbins, Morgan Freeman  USA  1993-09-23  1993  9.3  143
```

```
netflix_titles.isna().sum()/len(netflix_titles)*100
```

```
show_id      0.000000
type         0.000000
title        0.000000
director     0.000000
cast        10.002517
country      0.000000
date_added   0.000000
release_year  0.000000
rating       0.033565
duration     0.025174
listed_in    0.000000
description  0.000000
dtype: float64
```

```
netflix_titles=netflix_titles.assign(cast=netflix_titles['cast'].str.split(',')).explode('cast')
```

```
netflix_titles['cast'] = netflix_titles['cast'].str.strip()
```

```
netflix_titles['cast'].nunique()
```

```
36439
```

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

```
Alfred Molina      85
Liam Neeson        82
John Krasinski     67
Frank Langella     66
Salma Hayek        66
..
Ray Cordova        1
Jonathan Braylock  1
Shawtane Bowen     1
Otávio Martins     1
Chittaranjan Tripathy 1
Name: cast, Length: 36439, dtype: int64
```

```
netflix_titles['cast'].fillna('Unknown_cast', inplace = True)
```

```
netflix_titles.isna().sum()
```

```
show_id      0
type         0
title        0
director     0
cast         0
country      0
date_added   0
release_year  0
rating       38
duration     3
listed_in    0
description  0
dtype: int64
```

```
netflix_titles = netflix_titles.assign(listed_in = netflix_titles['listed_in'].str.split(',')).explode('listed_in')
netflix_titles['listed_in'] = netflix_titles['listed_in'].str.strip()
```

```
netflix_titles['listed_in'].value_counts()
```

```
Dramas                29806
International Movies   28243
Comedies              20829
International TV Shows 12845
Action & Adventure    12216
Independent Movies     9834
Children & Family Movies 9771
TV Dramas             8942
Thrillers             7107
Romantic Movies       6412
TV Comedies           4963
Crime TV Shows        4733
Horror Movies         4571
Kids' TV              4568
Sci-Fi & Fantasy      4037
Music & Musicals       3077
Romantic TV Shows     3049
Documentaries         2409
Anime Series          2313
TV Action & Adventure  2288
Spanish-Language TV Shows 2126
British TV Shows      1808
Sports Movies         1531
Classic Movies        1443
TV Mysteries          1281
Korean TV Shows       1122
Cult Movies           1077
TV Sci-Fi & Fantasy    1045
Anime Features        1045
TV Horror             941
Docuseries            845
LGBTQ Movies          838
TV Thrillers          768
Teen TV Shows         742
Reality TV            735
Faith & Spirituality   719
Stand-Up Comedy       540
Movies                412
TV Shows              337
Classic & Cult TV      272
Stand-Up Comedy & Talk Shows 268
Science & Nature TV    157
Name: listed_in, dtype: int64
```

```
netflix_titles['listed_in'].isna().sum()
```

```
0
```

```
netflix_titles
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown_cast	United States	2021-09-25 00:00:00	2020
1	s2	TV Show	Blood & Water	Adze Ujah	Ama Qamata	South Africa	2021-09-24 00:00:00	2021
1	s2	TV Show	Blood & Water	Adze Ujah	Ama Qamata	South Africa	2021-09-24 00:00:00	2021
1	s2	TV Show	Blood & Water	Adze Ujah	Ama Qamata	South Africa	2021-09-24 00:00:00	2021
1	s2	TV Show	Blood & Water	Adze Ujah	Khosi Ngema	South Africa	2021-09-24 00:00:00	2021
...

```
netflix_titles['rating'].isna().sum()

67

netflix_titles['rating'].value_counts()

TV-MA      73915
TV-14      43957
R           25860
PG-13      16246
TV-PG      14926
PG          10919
TV-Y7       6304
TV-Y        3665
TV-G        2779
NR           1573
G            1530
NC-17        149
TV-Y7-FV     86
UR            86
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```



```
netflix_titles['rating'].fillna('NR',inplace = True)

netflix_titles['rating'].isna().sum()

0

netflix_titles[netflix_titles['rating']=='NR'].head()
```

	show_id	type	title	director	cast	country	date_added	release_year
5971	s5972	Movie	(T)ERROR	Lyric R. Cabral	Unknown_cast	United States	2016-06-30 00:00:00	2016
5971	s5972	Movie	(T)ERROR	David Felix Sutcliffe	Unknown_cast	United States	2016-06-30 00:00:00	2016
5987	s5988	Movie	13 Cameras	Victor Zarcoff	PJ McCabe	United States	2016-08-13 00:00:00	2016

```
netflix_titles.drop(columns = 'description',inplace = True)

netflix_titles.to_csv('cleaned_preprocessed_netflix_titles.csv', sep = ',', index = False)

netflix_titles_new = pd.read_csv('/content/cleaned_preprocessed_netflix_titles.csv')

netflix_titles_new['country'].fillna('Unknown', inplace = True)

netflix_titles_new.to_csv('cleaned_preprocessed_netflix_titles.csv', sep = ',', index = False)

netflix_titles.head(1)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	1	TV-14	Dick	Robert B. Weide	Robert B. Weide	United States	2004-08-05 00:00:00	2004	NR

```
netflix_titles.isna().sum()

show_id      0
type         0
title        0
director     0
cast         0
country      0
date_added   0
release_year 0
rating       0
duration     3
listed_in    0
dtype: int64

netflix_titles_new.isna().sum()
```

```
show_id      0
type         0
title        0
director     0
cast         0
country      0
date_added   0
release_year  0
rating       0
duration     3
listed_in    0
dtype: int64
```