This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

```
import pandas as pd
import numpy as np

df = pd.read_csv("/content/netflix_titles.csv")
df.head(10)
```

```
  show_id     type                                 title  \
0      s1    Movie                  Dick Johnson Is Dead
1      s2  TV Show                          Blood & Water
2      s3  TV Show                              Ganglands
3      s4  TV Show                 Jailbirds New Orleans
4      s5  TV Show                           Kota Factory
5      s6  TV Show                          Midnight Mass
6      s7    Movie  My Little Pony: A New Generation
7      s8    Movie                                Sankofa
8      s9  TV Show     The Great British Baking Show
9     s10    Movie                           The Starling

                         director  \
0                  Kirsten Johnson
1                              NaN
2                  Julien Leclercq
3                              NaN
4                              NaN
5                    Mike Flanagan
6  Robert Cullen, José Luis Ucha
7                     Haile Gerima
8                  Andy Devonshire
9                   Theodore Melfi

                                                cast  \
0                                                NaN
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...
3                                                NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...
5  Kate Siegel, Zach Gilford, Hamish Linklater, H...
6  Vanessa Hudgens, Kimiko Glenn, James Marsden, ...
7  Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...
8  Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...
9  Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...
```

```
                                                country
date_added  \
0                                        United States  September 25,
2021
1                                         South Africa  September 24,
2021
2                                                  NaN  September 24,
2021
3                                                  NaN  September 24,
2021
4                                                India  September 24,
2021
5                                                  NaN  September 24,
2021
6                                                  NaN  September 24,
2021
7  United States, Ghana, Burkina Faso, United Kin...  September 24,
2021
8                                       United Kingdom  September 24,
2021
9                                        United States  September 24,
2021

   release_year rating   duration  \
0          2020  PG-13     90 min
1          2021  TV-MA  2 Seasons
2          2021  TV-MA   1 Season
3          2021  TV-MA   1 Season
4          2021  TV-MA  2 Seasons
5          2021  TV-MA   1 Season
6          2021     PG     91 min
7          1993  TV-MA    125 min
8          2021  TV-14  9 Seasons
9          2021  PG-13    104 min

                                           listed_in  \
0                                        Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                            Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...
5                 TV Dramas, TV Horror, TV Mysteries
6                           Children & Family Movies
7   Dramas, Independent Movies, International Movies
8                        British TV Shows, Reality TV
9                                   Comedies, Dramas

                                         description
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...
5  The arrival of a charismatic young priest brin...
6  Equestria's divided. But a bright-eyed hero be...
7  On a photo shoot in Ghana, an American model s...
8  A talented batch of amateur bakers face off in...
9  A woman adjusting to life after a loss contend...
```

###1- how many rows and columns in this dataset?

```
print(f'the number of rows is: {df.shape[0]}, and the number of
columns is: {df.shape[1]}')

the number of rows is: 8807, and the number of columns is: 12
```

###2- try seeing some information about the data and check if there is nulls

```
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

# a description of your data

for i in list(df.columns):

    print("\n ************ "+i+" ************\n")
    print("\n",df[i].value_counts())
    print("\n",df[i].describe(),"\n")


 ************ show_id ************
```

```
 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

 count     8807
unique    8807
top         s1
freq         1
Name: show_id, dtype: object


 ************ type ************


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

 count      8807
unique        2
top       Movie
freq       6131
Name: type, dtype: object


 ************ title ************


 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

 count                     8807
unique                    8807
top       Dick Johnson Is Dead
freq                         1
Name: title, dtype: object


 ************ director ************


 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

 count              6173
unique             4528
top       Rajiv Chilaka
freq                 19
Name: director, dtype: object


 ************ cast ************


 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

 count                   7982
unique                  7692
top       David Attenborough
freq                      19
Name: cast, dtype: object


 ************ country ************


 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

 count              7976
unique              748
top       United States
freq               2818
Name: country, dtype: object


 ************ date_added ************


 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

 count                 8797
unique                1767
top       January 1, 2020
freq                   109
Name: date_added, dtype: object


 ************ release_year ************


 2018    1147
2017    1032
2019    1030
2020     953
2016     902
        ...
1959       1
1925       1
1961       1
1947       1
1966       1
Name: release_year, Length: 74, dtype: int64

 count    8807.000000
mean     2014.180198
std         8.819312
min      1925.000000
25%      2013.000000
50%      2017.000000
75%      2019.000000
max      2021.000000
Name: release_year, dtype: float64


 ************ rating ************


 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

 count      8803
unique       17
top       TV-MA
freq       3207
Name: rating, dtype: object


 ************ duration ************


 1 Season     1793
2 Seasons     425
3 Seasons     199
90 min        152
94 min        146
             ...
16 min          1
186 min         1
193 min         1
189 min         1
191 min         1
Name: duration, Length: 220, dtype: int64

 count         8804
unique          220
top       1 Season
freq           1793
Name: duration, dtype: object


 ************ listed_in ************


 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

 count                              8807
unique                              514
top       Dramas, International Movies
freq                                362
Name: listed_in, dtype: object


 ************ description ************


 Paranormal activity at a lush, abandoned property alarms a group
eager to redevelop the site, but the eerie events may not be as
unearthly as they think.    4
Challenged to compose 100 songs before he can marry the girl he loves,
a tortured but passionate singer-songwriter embarks on a poignant
musical journey.    3
A surly septuagenarian gets another chance at her 20s after having her
photo snapped at a studio that magically takes 50 years off her life.
3
Multiple women report their husbands as missing but when it appears
they are looking for the same man, a police officer traces their
cryptic connection.     3
Secrets bubble to the surface after a sensual encounter and an
unforeseen crime entangle two friends and a woman caught between them.
2

..
Sent away to evade an arranged marriage, a 14-year-old begins a
harrowing journey of sex work and poverty in the slums of Accra.
1
When his partner in crime goes missing, a small-time crook’s life is
transformed as he dedicates himself to raising the daughter his friend
left behind.     1
During 1962’s Cuban missile crisis, a troubled math genius finds
himself drafted to play in a U.S.-Soviet chess match — and a deadly
game of espionage.      1
A teen's discovery of a vintage Polaroid camera develops into a darker
tale when she finds that whoever takes their photo with it dies soon
```

```
afterward.       1
A scrappy but poor boy worms his way into a tycoon's dysfunctional
family, while facing his fear of music and the truth about his past.
1
Name: description, Length: 8775, dtype: int64
```

```
 count                                                    8807
unique                                                   8775
top       Paranormal activity at a lush, abandoned prope...
freq                                                        4
Name: description, dtype: object
```

Some Questions you should ask yourself about. 1- Is there any duplicates? . 2-What about the nulls?. 3-Does all columns has the a correct format in its values? if its not how should you make it better? 4-Datatypes? 5- Before starting , after seeing some info about the dataset and from the first look on the dataset , what columns you think will not be necessary in our dataset? (io: what columns you think dropping it will be better?) feel free to wirte only their names in the next cell

unnecessary columns:

1. show_id
2. title
3. director
4. cast
5. description

###3- show the number of duplicates here

```
df.duplicated().sum()
```

```
0
```

###4- show number of nulls

```
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
```

```
# the cast column is full with nulls , replace the nulls with
"UnKnown"

df['cast'].fillna('UnKnown',inplace=True)

# make a new column that describes the number of people in the cast
(io :Hom many people in the cast? if it is unknown make it 0)
# hint --> make an external function and use apply method
def count_cast_members(cast):
    if cast=='UnKnown':
        return 0
    else:
      return len(cast.split(','))

df['cast_members'] = df['cast'].apply(count_cast_members)
df['cast_members']
```

```
0        0
1       19
2        9
3        0
4        8
        ..
8802    10
8803     0
8804     7
8805     9
8806     8
Name: cast_members, Length: 8807, dtype: int64
```

now time to get rid of these nulls

```
# let's start with date_added , see the number of nulls in it ,
replace these nulls with the mode of this column , and in the end
# convert this column to be in suitable format date time ,(hint -->
use fillna method)
df['date_added'].fillna(df['date_added'].mode()[0],inplace=True)
df['date_added']= pd.to_datetime(df['date_added'])
df['date_added'].info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 8807 entries, 0 to 8806
Series name: date_added
Non-Null Count  Dtype
--------------  -----
8807 non-null   datetime64[ns]
```

```
dtypes: datetime64[ns](1)
memory usage: 68.9 KB
```

```
# now time for country
# when you look closer at the dataset you will find that most of null values in country has the value "Anime" in listed_in column
# make a function that checks if Anime is in listed_in column
# and if it is then replace the null in country column of this row with "Japan"
# if it is not then replace the null with the most frequented value (io : mode)
# i will give you a first structue
df['country'].mode()
```

```
0    United States
Name: country, dtype: object
```

```
def country_null(x):
    if pd.isnull(x["country"]):
        # Check if "Anime" is in the "listed_in" column
        if "Anime" in x["listed_in"]:
            return "Japan"
        else:
            # Replace null with the mode (most frequent value) of the "country" column
            mode_country = df['country'].mode()[0]
            return mode_country
    else:
        return x["country"]

# Assuming df is your DataFrame
df['country'] = df.apply(country_null, axis=1)
df['country']
```

```
0       United States
1        South Africa
2       United States
3       United States
4               India
            ...
8802    United States
8803    United States
8804    United States
8805    United States
8806            India
Name: country, Length: 8807, dtype: object
```

```
# director column , duration and rating , fill with mode

df["director"].fillna(df["director"].mode()[0], inplace=True)
df["duration"].fillna(df["duration"].mode()[0], inplace=True)
```

```
df["rating"].fillna(df["rating"].mode()[0], inplace=True)
df.isnull().sum()
```

```
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
cast_members    0
dtype: int64
```

```
# in the rating column , UR and NR is the same where Unrated and
Notrated , so fix this
df['rating'].replace('UR','Not Rated',inplace= True)
df['rating'].replace('NR','Not Rated',inplace= True)

df.drop(df.index[df['rating'] =='74 min'],inplace=True)
df.drop(df.index[df['rating'] =='84 min'],inplace=True)
df.drop(df.index[df['rating'] =='66 min'],inplace=True)
df['rating'].value_counts()
```

```
TV-MA        3211
TV-14        2160
TV-PG         863
R             799
PG-13         490
TV-Y7         334
TV-Y          307
PG            287
TV-G          220
Not Rated      83
G              41
TV-Y7-FV        6
NC-17           3
Name: rating, dtype: int64
```

###in the end of this notebook some columns is strange ,

###what do you think we should do with something like show_id column

###feel free to do the same for the columns you thought it is not necessary

###and please write an explanation why do you think it is not important

```
# We should drop the show_id same with the title, director, cast, and
description columns because it is unnecessary and will not do good for
our machine learning project
# because of the number of unique values in it which will just put a
load on the learning process
df= df.drop('show_id',axis=1)
df= df.drop('title',axis=1)
df= df.drop('director',axis=1)
df= df.drop('cast',axis=1)
df= df.drop('description',axis=1)

df.head()
```

```
      type         country date_added  release_year rating
duration  \
0    Movie   United States 2021-09-25          2020  PG-13     90 min

1  TV Show    South Africa 2021-09-24          2021  TV-MA  2 Seasons

2  TV Show   United States 2021-09-24          2021  TV-MA   1 Season

3  TV Show   United States 2021-09-24          2021  TV-MA   1 Season

4  TV Show           India 2021-09-24          2021  TV-MA  2 Seasons


                                          listed_in  cast_members
0                                    Documentaries             0
1    International TV Shows, TV Dramas, TV Mysteries            19
2  Crime TV Shows, International TV Shows, TV Act...             9
3                            Docuseries, Reality TV              0
4  International TV Shows, Romantic TV Shows, TV ...             8
```