In [3]:
```
import numpy as np
import pandas as pd
```

In [46]:
```
df=pd.read_csv('D:\\Scaler\\Scaler\\Python\\Dataset\\netflix.csv')
```

In [130…
```
df.head()
```

Out[130]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Cast not specified	United States	2021-09-25	2020	PG-13	9
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	2021-09-24	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	Director not specified	Cast not specified	United States	2021-09-24	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	Director not specified	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	Se

1. Understanding/Inspecting the data-set

In [21]:
```
df.isnull().sum()
```

Out[21]:
```
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 [22]:
```
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 [23]:
```
df.head()
```

Out[23]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	9
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

In [24]:
```
df.tail()
```

Out[24]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14

In [25]:
```
df.shape
```

Out[25]:
```
(8807, 12)
```

In [26]:
```
df.columns
```

Out[26]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [27]:
```
df.describe()
```

Out[27]:

	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

In [28]:
```
df.isnull().sum().sort_values(ascending=False)
```

Out[28]:
```
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
```

In [29]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[29]:
```
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
```

2. Data Cleaning

Handling Missing Values

2.a As 'Director' Column contains missing (NaaN) values, the same is replaced by "Director not specified"

In [48]:
```
df['director']=df['director'].fillna('Director not specified')
```

In [49]:
```
df.head()
```

Out[49]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	9
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	Director not specified	NaN	NaN	September 24, 2021	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	Director not specified	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

In [50]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[50]:
```
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
director        0.00
release_year    0.00
listed_in       0.00
description     0.00
dtype: float64
```

2.b As 'Cast' column contains missing values, the same is replaced by 'Cast not specified'

In [51]:
```
df['cast']=df['cast'].fillna('Cast not specified')
```

In [52]:
```
df.head()
```

Out[52]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Cast not specified	United States	September 25, 2021	2020	PG-13	9
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	Director not specified	Cast not specified	NaN	September 24, 2021	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	Director not specified	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

In [53]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[53]:
```
country         9.44
date_added      0.11
rating          0.05
duration        0.03
show_id         0.00
type            0.00
title           0.00
director        0.00
cast            0.00
release_year    0.00
listed_in       0.00
description     0.00
dtype: float64
```

2.c As Rating, duration and date_added columns contain missing values and their number is negligible, those columns are being dropped

In [57]:
```
df.dropna(subset=['rating', 'duration', 'date_added'], axis=0, inplace=True)
```

In [58]:
```
df.shape
```

Out[58]:
```
(8790, 12)
```

In [59]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[59]:
```
country         9.43
show_id         0.00
type            0.00
title           0.00
director        0.00
cast            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
```

In [63]:
```
country_counts = (df['country'].value_counts() / df.shape[0] * 100).round(2)
country_counts.head()
```

Out[63]:
```
United States     31.96
India             11.06
United Kingdom     4.76
Japan              2.76
South Korea        2.26
Name: country, dtype: float64
```

2.d As country contains a good number of missing values and we can see that United States contributes 32%, we can replace the missing values with United states

In [64]:
```
df['country'].fillna(df['country'].mode()[0], inplace=True)
```

In [65]:
```
df.head()
```

Out[65]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Cast not specified	United States	September 25, 2021	2020	PG-13	9
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	September 24, 2021	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	Director not specified	Cast not specified	United States	September 24, 2021	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	Director not specified	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Se

In [68]:
```
country_counts = (df['country'].value_counts() / df.shape[0] * 100).round(2)
country_counts.head()
```

Out[68]:
```
United States     41.39
India             11.06
United Kingdom     4.76
Japan              2.76
South Korea        2.26
Name: country, dtype: float64
```

All the missing values handled so far

In [66]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[66]:
```
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
```

3. Data Preparation

3.1 Managing 'date_added' column

In [70]:
```
df.dtypes
```

Out[70]:
```
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object
```

Type of 'date_added' column needs to be converted into date type

In [71]:
```
df['date_added'] = pd.to_datetime(df['date_added'])
```

In [72]:
```
df.dtypes
```

Out[72]:
```
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
description             object
dtype: object
```

Additional columns named 'month', 'year' and 'week' are added which might help in data analysis

In [73]:
```
df['month'] = df['date_added'].dt.month
df['year'] = df['date_added'].dt.year
df['week'] = df['date_added'].dt.week
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_20472\3376655636.py:3: FutureWarning: Series.dt.weekofyear and Series.dt.week have been deprecated. Please use Series.dt.isocalendar().week instead.
  df['week'] = df['date_added'].dt.week
```

In [74]:
```
df.head()
```

Out[74]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Cast not specified	United States	2021-09-25	2020	PG-13	9
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	2021-09-24	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	Director not specified	Cast not specified	United States	2021-09-24	2021	TV-MA	1 S
4	s5	TV Show	Kota Factory	Director not specified	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	Se

3.2 Managing 'type' column - Divided them into 'Movie' and 'TV Show' types so that we can do the 'duration' analysis separately in terms of minutes and seasons respectively

In [81]:
```
movies_df=df.loc[(df['type']=='Movie')]
movies_df.head(2)
```

Out[81]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	dur
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Cast not specified	United States	2021-09-25	2020	PG-13	9
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	United States	2021-09-24	2021	PG	9

In [82]:
```
tvshow_df=df.loc[(df['type']=='TV Show')]
tvshow_df.head(2)
```

Out[82]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	dur
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	United States	2021-09-24	2021	TV-MA	1 S

3.3 Managing 'duration' column

Data type of 'duration' column needs to be converted into numerical type to do the required data analysis

In [90]:
```
movies_df = movies_df.copy()
movies_df['duration'] = movies_df['duration'].apply(lambda x: x.replace(' min', ''
movies_df.duration
```

Out[90]:
```
0        90
6        91
7       125
9       104
12      127
       ...
8801     96
8802    158
8804     88
8805     88
8806    111
Name: duration, Length: 6126, dtype: object
```

In [91]:
```
movies_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6126 entries, 0 to 8806
Data columns (total 15 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   datetime64[ns]
 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
 12  month         6126 non-null   int64
 13  year          6126 non-null   int64
 14  week          6126 non-null   int64
dtypes: datetime64[ns](1), int64(4), object(10)
memory usage: 765.8+ KB
```

In [96]:
```
movies_df['duration'] = movies_df['duration'].apply(lambda x: int(x) if isinstance

movies_df.describe()
```

Out[96]:

	release_year	duration	month	year	week
count	6126.000000	6126.000000	6126.000000	6126.000000	6126.000000
mean	2013.120144	99.584884	6.609370	2018.851126	26.223310
std	9.681723	28.283225	3.452541	1.561173	15.116616
min	1942.000000	3.000000	1.000000	2008.000000	1.000000
25%	2012.000000	87.000000	4.000000	2018.000000	13.000000
50%	2016.000000	98.000000	7.000000	2019.000000	26.000000
75%	2018.000000	114.000000	10.000000	2020.000000	39.000000
max	2021.000000	312.000000	12.000000	2021.000000	53.000000

In [97]:
```
movies_df.dtypes
```

Out[97]:
```
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                 int64
listed_in               object
description             object
month                    int64
year                     int64
week                     int64
dtype: object
```

In [101…
```
tvshow_df = tvshow_df.copy()
tvshow_df['duration'] = tvshow_df['duration'].apply(lambda x: x.replace(' Season',
tvshow_df['duration'] = tvshow_df['duration'].apply(lambda x: x.replace('s', '') i
tvshow_df.duration
```

Out[101]:
```
1       2
2       1
3       1
4       2
5       1
       ..
8795    2
8796    2
8797    3
8800    1
8803    2
Name: duration, Length: 2664, dtype: object
```

In [102…
```
tvshow_df['duration'] = tvshow_df['duration'].apply(lambda x: int(x) if isinstance

tvshow_df.describe()
```

Out[102]:

	release_year	duration	month	year	week
count	2664.000000	2664.000000	2664.000000	2664.000000	2664.000000
mean	2016.627628	1.751877	6.762763	2018.925300	27.827327
std	5.735194	1.550622	3.396231	1.600804	14.790648
min	1925.000000	1.000000	1.000000	2008.000000	1.000000
25%	2016.000000	1.000000	4.000000	2018.000000	15.000000
50%	2018.000000	1.000000	7.000000	2019.000000	28.000000
75%	2020.000000	2.000000	10.000000	2020.000000	40.000000
max	2021.000000	17.000000	12.000000	2021.000000	53.000000

In [103…
```
tvshow_df.dtypes
```

Out[103]:
```
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                 int64
listed_in               object
description             object
month                    int64
year                     int64
week                     int64
dtype: object
```

3.4 Managing Multiple values in a single record/cell

a. 'cast' column

In [ ]:

In [ ]:

In [ ]:

In [105…
```
constraint=df['cast'].apply(lambda x: str(x).split(', ')).tolist()
```

In [106…
```
cast_df=pd.DataFrame(constraint,index=df['title'])
```

In [107…
```
cast_df=cast_df.stack()
cast_df=pd.DataFrame(cast_df)
cast_df.reset_index(inplace=True)
cast_df=cast_df[['title',0]]
cast_df.columns=['title','cast']
```

In [108…
```
cast_df.head()
```

Out[108]:

	title	cast
0	Dick Johnson Is Dead	Cast not specified
1	Blood & Water	Ama Qamata
2	Blood & Water	Khosi Ngema
3	Blood & Water	Gail Mabalane
4	Blood & Water	Thabang Molaba

3.4 Managing Multiple values in a single record/cell -

b. 'director' column

In [111…
```
Cons=df['director'].apply(lambda x: str(x).split(', ')).tolist()
```

In [113…
```
director_df=pd.DataFrame(Cons,index=df['title'])
```

In [114…
```
director_df=director_df.stack()
director_df=pd.DataFrame(director_df)
director_df.reset_index(inplace=True)
director_df=director_df[['title',0]]
director_df.columns=['title','director']
```

In [115…
```
director_df.head()
```

Out[115]:

	title	director
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	Director not specified
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	Director not specified
4	Kota Factory	Director not specified

In [116…
```
director_count=director_df.groupby(['director']).size().reset_index(name='count')
```

In [117…
```
director_count
```

Out[117]:

	director	count
0	A. L. Vijay	2
1	A. Raajdheep	1
2	A. Salaam	1
3	A.R. Murugadoss	2
4	Aadish Keluskar	1
...	...	...
4987	Éric Warin	1
4988	Ísold Uggadóttir	1
4989	Óskar Thór Axelsson	1
4990	Ömer Faruk Sorak	3
4991	Şenol Sönmez	2

4992 rows × 2 columns

In [118…
```
director_count = director_count.sort_values(by=['count'], ascending = False)
```

In [119…
```
director_count
```

Out[119]:

	director	count
1206	Director not specified	2621
3748	Rajiv Chilaka	22
1906	Jan Suter	21
3799	Raúl Campos	19
2865	Marcus Raboy	16
...	...	...
2293	Jovanka Vuckovic	1
634	Brandon Camp	1
2295	Juan Antin	1
2296	Juan Antonio de la Riva	1
2956	María Jose Cuevas	1

4992 rows × 2 columns

In [120…
```
director_count = director_count[director_count.director != 'Director not specified
```

In [121…
```
director_count
```

Out[121]:

	director	count
3748	Rajiv Chilaka	22
1906	Jan Suter	21
3799	Raúl Campos	19
2865	Marcus Raboy	16
4456	Suhas Kadav	16
...	...	...
2293	Jovanka Vuckovic	1
634	Brandon Camp	1
2295	Juan Antin	1
2296	Juan Antonio de la Riva	1
2956	María Jose Cuevas	1

4991 rows × 2 columns

3.4 Managing Multiple values in a single record/cell -

a. 'cast' column

In [125…
```
Const=df['listed_in'].apply(lambda x: str(x).split(', ')).tolist()
```

In [126…
```
genre_df=pd.DataFrame(Const,index=df['title'])
```

In [127…
```
genre_df=genre_df.stack()
#genre_df=genre_df.stack()
genre_df=pd.DataFrame(genre_df)
genre_df.reset_index(inplace=True)
genre_df=genre_df[['title',0]]
genre_df.columns=['title','genre']
```

In [128…
```
genre_df.head()
```

Out[128]:

	title	genre
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows

In [129…
```
choco install pandoc
```

```
  Cell In[129], line 1
    choco install pandoc
          ^
SyntaxError: invalid syntax
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: