

```
#1_import libraries and load dataset
import pandas as pd
import numpy as np
df=pd.read_csv("/content/netflix_titles.csv.zip")
```

```
#2_initial dataset overviews
df.info()
df.head()
```

<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

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	desc
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As h r e life
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	p party
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 pi fami (
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	flirta toil dov
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 ...	Ir

Next steps:

Generate code with df

☒ View recommended plots

New interactive sheet

```
#3_handling missing values
print(df.isnull().sum())
```

show_id
type
title
director
cast
country

0
0
0
2634
825
831

◆

What can I help you build?

⊕

▶

```

date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description       0
dtype: int64

```

```

# Drop rows where 'title' is missing (critical column
print(df.dropna(subset=["title"]))

```

```

↗
   show_id  type      title      director \
0         s1  Movie  Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show      Blood & Water             NaN
2         s3  TV Show      Ganglands      Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans             NaN
4         s5  TV Show      Kota Factory             NaN
...      ...    ...      ...      ...
8802    s8803  Movie      Zodiac      David Fincher
8803    s8804  TV Show      Zombie Dumb             NaN
8804    s8805  Movie      Zombieland  Ruben Fleischer
8805    s8806  Movie      Zoom      Peter Hewitt
8806    s8807  Movie      Zubaan      Mozez Singh

   cast      country \
0      NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...  South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...      NaN
3      NaN      NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...      India
...      ...      ...
8802  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8803      NaN      NaN
8804  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8805  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8806  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...      India

   date_added  release_year  rating  duration \
0  September 25, 2021      2020  PG-13    90 min
1  September 24, 2021      2021  TV-MA  2 Seasons
2  September 24, 2021      2021  TV-MA    1 Season
3  September 24, 2021      2021  TV-MA    1 Season
4  September 24, 2021      2021  TV-MA  2 Seasons
...      ...      ...      ...
8802  November 20, 2019      2007    R    158 min
8803    July 1, 2019      2018  TV-Y7  2 Seasons
8804  November 1, 2019      2009    R    88 min
8805  January 11, 2020      2006   PG    88 min
8806    March 2, 2019      2015  TV-14   111 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 ...
...      ...
8802      Cult Movies, Dramas, Thrillers
8803  Kids' TV, Korean TV Shows, TV Comedies
8804      Comedies, Horror Movies
8805  Children & Family Movies, Comedies
8806  Dramas, International Movies, Music & Musicals

   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...

```

```

# # Fill missing 'country' with mode
print(df['country'].fillna(df['country'].mode()[0], inplace=True))

```

```


↗ None
/tmp/ipython-input-26-3300334354.py:1: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series thr
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[c

```

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

```
# Fill missing 'director' and 'cast' with 'Not Specified'
df['director'].fillna('Not Specified', inplace=True)
df['cast'].fillna('Not Specified', inplace=True)
```

 /tmp/ipython-input-27-3608875680.py:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through inplace=... The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we


For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[c

```
df['director'].fillna('Not Specified', inplace=True)
/tmp/ipython-input-27-3608875680.py:3: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we
```

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[c

```
df['cast'].fillna('Not Specified', inplace=True)
```

```
# Fill missing 'date_added' with placeholder or drop (depending on case)
df['date_added'].fillna('Unknown', inplace=True)
```

 /tmp/ipython-input-28-3515460210.py:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[c

```
df['date_added'].fillna('Unknown', inplace=True)
```

#5. Standardizing Text Data

```
# Trim whitespaces and lowercase column names
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')
df.columns
```

```
# Example: Standardize 'type' column
df['type'] = df['type'].str.title()
df['type']
```

 **type**

0 Movie

1 Tv Show

2 Tv Show

3 Tv Show

4 Tv Show

...

8802 Movie

8803 Tv Show

8804 Movie

8805 Movie

8806 Movie

8807 rows × 1 columns

dtype: object

```
#6. Date Format Consistency
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
df['date_added']
```



	date_added
0	2021-09-25
1	2021-09-24
2	2021-09-24
3	2021-09-24
4	2021-09-24
...	...
8802	2019-11-20
8803	2019-07-01
8804	2019-11-01
8805	2020-01-11
8806	2019-03-02

8807 rows × 1 columns

dtype: datetime64[ns]

#7. Fixing Data Types

Ensure 'release_year' is integer

```
df['release_year'] = df['release_year'].astype(int)
df['release_year']
```



	release_year
0	2020
1	2021
2	2021
3	2021
4	2021
...	...
8802	2007
8803	2018
8804	2009
8805	2006
8806	2015

8807 rows × 1 columns

dtype: int64

Split duration into numeric + unit (optional)

```
df[['duration_num', 'duration_type']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')
df['duration_num'] = pd.to_numeric(df['duration_num'], errors='coerce')
df['duration_num']
```



	duration_num
0	90.0
1	2.0
2	1.0
3	1.0
4	2.0
...	...
8802	158.0
8803	2.0
8804	88.0
8805	88.0
8806	111.0

8807 rows × 1 columns

dtype: float64