

```
import pandas as pd

df = pd.read_csv("Netflix_movies_and_tv_shows_clustering.csv")

print(df.info())
print(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4808 entries, 1 to 7783
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   show_id                4808 non-null   object
1   type                  4808 non-null   object
2   title                 4808 non-null   object
3   director              4808 non-null   object
4   cast                  4808 non-null   object
5   country               4808 non-null   object
6   date_added            4808 non-null   object
7   release_year          4808 non-null   int64
8   rating                4808 non-null   object
9   duration              4808 non-null   object
10  listed_in             4808 non-null   object
11  description            4808 non-null   object
dtypes: int64(1), object(11)
memory usage: 488.3+ KB
None
```

	show_id	type	title	director
1	s2	movie	7:19	jorge michel grau
2	s3	movie	23:59	gilbert chan
3	s4	movie	9	shane acker
4	s5	movie	21	robert luketic
5	s6	tv show	46	serdar akar

	cast	country
1	demián bichir, héctor bonilla, oscar serrano, ...	mexico
2	tedd chan, stella chung, henley hii, lawrence ...	singapore
3	elijah wood, john c. reilly, jennifer connelly...	united states
4	jim sturgess, kevin spacey, kate bosworth, aar...	united states
5	erdal beşikçioğlu, yasemin allen, melis birkan...	turkey

	date_added	release_year	rating	duration
1	december 23, 2016	2016	tv-ma	93 min
2	december 20, 2018	2011	r	78 min
3	november 16, 2017	2009	pg-13	80 min
4	january 1, 2020	2008	pg-13	123 min
5	july 1, 2017	2016	tv-ma	1 season

listed_in \

```

1          dramas, international movies
2          horror movies, international movies
3  action & adventure, independent movies, sci-fi...
4          dramas
5  international tv shows, tv dramas, tv mysteries

```

```

description
1  after a devastating earthquake hits mexico cit...
2  when an army recruit is found dead, his fellow...
3  in a postapocalyptic world, rag-doll robots hi...
4  a brilliant group of students become card-coun...
5  a genetics professor experiments with a treatm...

```

```
print(df.isnull().sum()) #identify the null values
```

```

show_id      0
type         0
title        0
director    2389
cast        718
country     507
date_added   10
release_year  0
rating       7
duration     0
listed_in    0
description  0
dtype: int64

```

```
df.dropna(inplace=True) #drop rows having missing values
```

```
df.drop_duplicates(inplace=True) # remove duplicate rows
```

```

text_columns = df.select_dtypes(include='object').columns
for col in text_columns:
    df[col] = df[col].astype(str).str.strip().str.lower()

```

```
print(df.head())
```

```

show_id  type  title  director \
1      s2  movie  7:19  jorge michel grau
2      s3  movie  23:59  gilbert chan
3      s4  movie      9  shane acker
4      s5  movie     21  robert luketic
5      s6  tv show    46  serdar akar

```

```

cast  country \
1  demián bichir, héctor bonilla, oscar serrano, ...  mexico
2  tedd chan, stella chung, henley hii, lawrence ...  singapore

```

```

3 elijah wood, john c. reilly, jennifer connelly... united states
4 jim sturgess, kevin spacey, kate bosworth, aar... united states
5 erdal beşikçioğlu, yasemin allen, melis birkan... turkey

```

```

      date_added  release_year rating  duration \
1 december 23, 2016          2016  tv-ma    93 min
2 december 20, 2018          2011      r    78 min
3 november 16, 2017          2009  pg-13    80 min
4  january 1, 2020          2008  pg-13   123 min
5    july 1, 2017          2016  tv-ma    1 season

```

```

                                listed_in \
1                                dramas, international movies
2                                horror movies, international movies
3  action & adventure, independent movies, sci-fi...
4                                dramas
5  international tv shows, tv dramas, tv mysteries

```

```

                                description
1  after a devastating earthquake hits mexico cit...
2  when an army recruit is found dead, his fellow...
3  in a postapocalyptic world, rag-doll robots hi...
4  a brilliant group of students become card-coun...
5  a genetics professor experiments with a treatm...

```

```

if 'date_added' in df.columns: #converting date into proper format
    df['date_added'] = pd.to_datetime(df['date_added'],
errors='coerce')
    df['date_added'] = df['date_added'].dt.strftime('%d-%m-%Y')

```

<ipython-input-21-11ef0551eb15>:2: UserWarning: Could not infer format, so each element will be parsed individually, falling back to `dateutil`. To ensure parsing is consistent and as-expected, please specify a format.

```
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```

```
print(df.head)
```

```

<bound method NDFrame.head of      show_id    type    title
director \
1         s2    movie    7:19  jorge michel grau
2         s3    movie   23:59    gilbert chan
3         s4    movie     9    shane acker
4         s5    movie    21    robert luketic
5         s6  tv show    46    serdar akar
...
7778    s7779    movie  zombieland    ruben fleischer
7780    s7781    movie     zoo    shlok sharma
7781    s7782    movie    zoom    peter hewitt
7782    s7783    movie    zozo    josef fares

```

7783	s7784	movie	zubaan	mozez singh	
				cast	\
1				demián bichir, héctor bonilla, oscar serrano, ...	
2				tedd chan, stella chung, henley hii, lawrence ...	
3				elijah wood, john c. reilly, jennifer connelly...	
4				jim sturgess, kevin spacey, kate bosworth, aar...	
5				erdal beşikçioğlu, yasemin allen, melis birkan...	
...				...	
7778				jesse eisenberg, woody harrelson, emma stone, ...	
7780				shashank arora, shweta tripathi, rahul kumar, ...	
7781				tim allen, courteney cox, chevy chase, kate ma...	
7782				imad creidi, antoinette turk, elias gergi, car...	
7783				vicky kaushal, sarah-jane dias, raaghav chanan...	
				country	date_added \
1				mexico	23-12-2016
2				singapore	20-12-2018
3				united states	16-11-2017
4				united states	01-01-2020
5				turkey	01-07-2017
...			
7778				united states	01-11-2019
7780				india	01-07-2018
7781				united states	11-01-2020
7782				sweden, czech republic, united kingdom, denmar...	19-10-2020
7783				india	02-03-2019
				release_year	rating duration \
1				2016	tv-ma 93 min
2				2011	r 78 min
3				2009	pg-13 80 min
4				2008	pg-13 123 min
5				2016	tv-ma 1 season
...			
7778				2009	r 88 min
7780				2018	tv-ma 94 min
7781				2006	pg 88 min
7782				2005	tv-ma 99 min
7783				2015	tv-14 111 min
				listed_in	\
1				dramas, international movies	
2				horror movies, international movies	
3				action & adventure, independent movies, sci-fi...	
4				dramas	
5				international tv shows, tv dramas, tv mysteries	
...				...	
7778				comedies, horror movies	
7780				dramas, independent movies, international movies	

```

7781         children & family movies, comedies
7782         dramas, international movies
7783         dramas, international movies, music & musicals

```

```

description
1  after a devastating earthquake hits mexico cit...
2  when an army recruit is found dead, his fellow...
3  in a postapocalyptic world, rag-doll robots hi...
4  a brilliant group of students become card-coun...
5  a genetics professor experiments with a treatm...
...
7778 looking to survive in a world taken over by zo...
7780 a drug dealer starts having doubts about his t...
7781 dragged from civilian life, a former superhero...
7782 when lebanon's civil war deprives zozo of his ...
7783 a scrappy but poor boy worms his way into a ty...

```

```
[4808 rows x 12 columns]>
```

```

df.columns = [col.strip().lower().replace(" ", "_") for col in
df.columns] #column

```

```
print(df.head)
```

```

<bound method NDFrame.head of          show_id    type    title
director \
1         s2      movie    7:19  jorge michel grau
2         s3      movie   23:59    gilbert chan
3         s4      movie     9      shane acker
4         s5      movie    21    robert luketic
5         s6  tv show    46      serdar akar
...
7778    s7779      movie  zombieland    ruben fleischer
7780    s7781      movie     zoo      shlok sharma
7781    s7782      movie    zoom    peter hewitt
7782    s7783      movie    zozo      josef fares
7783    s7784      movie   zubaan      mozez singh

```

```

cast \
1  demián bichir, héctor bonilla, oscar serrano, ...
2  tedd chan, stella chung, henley hii, lawrence ...
3  elijah wood, john c. reilly, jennifer connelly...
4  jim sturgess, kevin spacey, kate bosworth, aar...
5  erdal beşikçioğlu, yasemin allen, melis birkan...
...
7778  jesse eisenberg, woody harrelson, emma stone, ...
7780  shashank arora, shweta tripathi, rahul kumar, ...
7781  tim allen, courteney cox, chevy chase, kate ma...
7782  imad creidi, antoinette turk, elias gergi, car...
7783  vicky kaushal, sarah-jane dias, raaghav chanan...

```

	country	date_added	\
1	mexico	23-12-2016	
2	singapore	20-12-2018	
3	united states	16-11-2017	
4	united states	01-01-2020	
5	turkey	01-07-2017	
...	
7778	united states	01-11-2019	
7780	india	01-07-2018	
7781	united states	11-01-2020	
7782	sweden, czech republic, united kingdom, denmar...	19-10-2020	
7783	india	02-03-2019	

	release_year	rating	duration	\
1	2016	tv-ma	93 min	
2	2011	r	78 min	
3	2009	pg-13	80 min	
4	2008	pg-13	123 min	
5	2016	tv-ma	1 season	
...	
7778	2009	r	88 min	
7780	2018	tv-ma	94 min	
7781	2006	pg	88 min	
7782	2005	tv-ma	99 min	
7783	2015	tv-14	111 min	

	listed_in	\
1	dramas, international movies	
2	horror movies, international movies	
3	action & adventure, independent movies, sci-fi...	
4	dramas	
5	international tv shows, tv dramas, tv mysteries	
...	...	
7778	comedies, horror movies	
7780	dramas, independent movies, international movies	
7781	children & family movies, comedies	
7782	dramas, international movies	
7783	dramas, international movies, music & musicals	

	description
1	after a devastating earthquake hits mexico cit...
2	when an army recruit is found dead, his fellow...
3	in a postapocalyptic world, rag-doll robots hi...
4	a brilliant group of students become card-coun...
5	a genetics professor experiments with a treatm...
...	...
7778	looking to survive in a world taken over by zo...
7780	a drug dealer starts having doubts about his t...
7781	dragged from civilian life, a former superhero...

```
7782  when lebanon's civil war deprives zozo of his ...
7783  a scrappy but poor boy worms his way into a ty...

[4808 rows x 12 columns]>

if 'release_year' in df.columns:
    df['release_year'] = pd.to_numeric(df['release_year'],
errors='coerce').astype('Int64')

df.to_csv("cleaned_netflix_data.csv", index=False) #saving it
```