Data cleaning on Netflix dataset

By Prince Ogwu

In [2]:
```
# Import the neccessary library

import pandas as pd
```

In [72]:
```
#get data from an external source

url = 'https://raw.githubusercontent.com/kedeisha1/Challenges/main/netflix_titles.c
data = pd.read_csv(url)

data.head(5)
```

Out[72]:

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

In [84]:
```
#Saving our data

data.to_csv('netflix.csv', index=False)
```

In [3]:
```
df =pd.read_csv('netflix.csv')
df.tail(10)
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_ye
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	September 13, 2018	20
8798	s8799	Movie	Zed Plus	Chandra Prakash Dwivedi	Adil Hussain, Mona Singh, K.K. Raina, Sanjay M...	India	December 31, 2019	20
8799	s8800	Movie	Zenda	Avadhoot Gupte	Santosh Juvekar, Siddharth Chandekar, Sachit P...	India	February 15, 2018	20
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	December 15, 2016	20
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	March 9, 2016	20
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	20
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	20
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	20

	show_id	type	title	director	cast	country	date_added	release_ye
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	20
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	20

In [75]:
```
# Explore our data

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 [87]:
```
# Check for missing values

df.isna().any()
```

Out[87]:
```
show_id         False
type            False
title           False
director         True
cast             True
country          True
date_added       True
release_year    False
rating           True
duration         True
listed_in       False
description     False
dtype: bool
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [5]:
```
#Drop missing vaues
df.dropna(inplace=True)
```

In [105…
```
df.shape
```

Out[105]:
```
(8807, 12)
```

In [78]:
```
# drop Missing values
df.dropna(inplace=True)
```

In [80]:
```
# Checking to see we've dropped the missing records

df.isna().any().sum()
```

Out[80]:
```
0
```

In [81]:
```
# Check for duplicate records

df.duplicated().any()
```

Out[81]:
```
False
```

In [83]:
```
df.shape
```

Out[83]:
```
(5332, 12)
```

In [ ]: