Aim: Clean the dataset replace missing values, remove outliers

In [2]:
```
#the required dataset is already stored in the codes on bytes project
```

In [3]:
```
#importing the libraries is the initial step in the process of data preprocessing
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Replacing the Missing Values

In [4]:
```
df=pd.read_csv('netflix.csv')
print(df.head(5))
```

```
  show_id     type                              title         director  \
0      s1    Movie              Dick Johnson Is Dead  Kirsten Johnson
1      s3  TV Show                         Ganglands  Julien Leclercq
2      s6  TV Show                     Midnight Mass    Mike Flanagan
3     s14    Movie  Confessions of an Invisible Girl    Bruno Garotti
4      s8    Movie                           Sankofa     Haile Gerima

         country date_added  release_year rating  duration  \
0  United States  9/25/2021          2020  PG-13    90 min
1         France  9/24/2021          2021  TV-MA  1 Season
2  United States  9/24/2021          2021  TV-MA  1 Season
3         Brazil  9/22/2021          2021  TV-PG    91 min
4  United States  9/24/2021          1993  TV-MA   125 min

                                        listed_in
0                                   Documentaries
1  Crime TV Shows, International TV Shows, TV Act...
2                TV Dramas, TV Horror, TV Mysteries
3                Children & Family Movies, Comedies
4   Dramas, Independent Movies, International Movies
```

let's explore the data with the with respect of columns , data types as well as checking the null values.

In [5]:
```
df.dtypes
```

Out[5]:
```
show_id         object
type            object
title           object
director        object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
dtype: object
```

In [6]:
```
print(r" the size of the data is ", df.size)
print(r" and it's shape is " ,df.shape)
```

```
the size of the data is  87900
and it's shape is  (8790, 10)
```

From the above data out of 10 columns 9 are belongs to object and only is the integer type data.

In [47]:
```
df.columns
```

Out[47]:
```
Index(['show_id', 'type', 'title', 'director', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in'],
      dtype='object')
```

In [48]:
```
print(df.isnull().sum())
```

```
show_id         0
type            0
title           0
director        0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64 
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: int64(1), object(9)
memory usage: 686.8+ KB
```

we can obeserve here that in the data frame there are no null values are present.

But many people often treats the Null values and the missing values are same.

although in general context, they used for one another but deep down they are different from each other, for example:

the null values represents absence of data field in the data frame where as the

missing values are placed with the Placeholders such as "NA" or "NaN".

SInce the data doesn't consists of null values let's explore for the missing values

In [50]:
```
# unique function helps to find the unique values in every column.
df
```

Out[50]:

	show_id	type	title	director	country	date_added	release_year	rating	duration
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min
...	...	...	...	...	...	...	...	...	...
8785	s8797	TV Show	Yunus Emre	Not Given	Turkey	1/17/2017	2016	TV-PG	2 Seasons
8786	s8798	TV Show	Zak Storm	Not Given	United States	9/13/2018	2016	TV-Y7	3 Seasons
8787	s8801	TV Show	Zindagi Gulzar Hai	Not Given	Pakistan	12/15/2016	2012	TV-PG	1 Season
8788	s8784	TV Show	Yoko	Not Given	Pakistan	6/23/2018	2016	TV-Y	1 Season
8789	s8786	TV Show	YOM	Not Given	Pakistan	6/7/2018	2016	TV-Y7	1 Season

8790 rows × 10 columns

In [51]:
```
#let's see unique value numbers withrespect to the columns
print(df.nunique())
```

```
show_id         8790
type               2
title           8787
director        4528
country           86
date_added      1713
release_year      74
rating            14
duration         220
listed_in        513
dtype: int64
```

In [37]:
```
#let's see value counts for all the data
def unix(i):
    print(df[i].value_counts())
for i in df:
    unix(i)
```

```
s1       1
s7990    1
s7982    1
s7984    1
s7986    1
        ..
s3960    1
s3958    1
s3956    1
s3955    1
s8786    1
Name: show_id, Length: 8790, dtype: int64
Movie      6126
TV Show    2664
Name: type, dtype: int64
9-Feb                   2
15-Aug                  2
22-Jul                  2
Dick Johnson Is Dead    1
SGT. Will Gardner       1
                       ..
Mercy Black             1
The Trap                1
Pinky Memsaab           1
Love O2O                1
YOM                     1
Name: title, Length: 8787, dtype: int64
Not Given                          2588
Rajiv Chilaka                        20
Alastair Fothergill                  18
Raúl Campos, Jan Suter               18
Suhas Kadav                          16
                                    ...
Matt D'Avella                         1
Parthiban                             1
Scott McAboy                          1
Raymie Muzquiz, Stu Livingston        1
Mozez Singh                           1
Name: director, Length: 4528, dtype: int64
United States     3240
India             1057
United Kingdom     638
Pakistan           421
Not Given          287
                  ...
Iran                 1
West Germany         1
Greece               1
Zimbabwe             1
Soviet Union         1
Name: country, Length: 86, dtype: int64
1/1/2020      110
11/1/2019      91
3/1/2018       75
12/31/2019     74
10/1/2018      71
              ...
6/26/2015       1
6/23/2015       1
6/1/2015        1
5/29/2015       1
4/1/2014        1
Name: date_added, Length: 1713, dtype: int64
2018    1146
```

```
2017    1030
2019    1030
2020     953
2016     901
        ...
1966       1
1959       1
1925       1
1947       1
1961       1
Name: release_year, Length: 74, dtype: int64
TV-MA       3205
TV-14       2157
TV-PG        861
R            799
PG-13        490
TV-Y7        333
TV-Y         306
PG           287
TV-G         220
NR            79
G             41
TV-Y7-FV       6
NC-17          3
UR             3
Name: rating, dtype: int64
1 Season      1791
2 Seasons      421
3 Seasons      198
90 min         152
97 min         146
              ...
5 min            1
16 min           1
186 min          1
193 min          1
11 Seasons       1
Name: duration, Length: 220, dtype: int64
Dramas, International Movies                                  362
Documentaries                                                 359
Stand-Up Comedy                                               334
Comedies, Dramas, International Movies                        274
Dramas, Independent Movies, International Movies              252
                                                             ...
Anime Features                                                  1
Action & Adventure, Horror Movies, Independent Movies           1
Action & Adventure, Classic Movies, International Movies        1
Cult Movies, Independent Movies, Thrillers                      1
Classic & Cult TV, Crime TV Shows, TV Dramas                    1
Name: listed_in, Length: 513, dtype: int64
```

In this process i have identified that there is a filed named as "Not Given" which is the missing values we have to handle.

let's see how many missing values lies in each of the attribute or column.

In [42]:
```
def che(i):
    t=0
    for j in df[i]:
        if j=='Not Given':
            t+=1
    print(i,t)
```

```
print("Not Given details")
for i in df:
    che(i)
```

```
Not Given details
show_id 0
type 0
title 0
director 2588
country 287
date_added 0
release_year 0
rating 0
duration 0
listed_in 0
```

we can observe that for 2588 movies or shows doesn't provide their director name out of 8790 movies/shows. So in the process of dealing with missing values we are going to drop the column named "director", since it occupies a large part compared to that of total shows.

In [52]:
```
df=df.drop('director',axis=1)
df
```

Out[52]:

	show_id	type	title	country	date_added	release_year	rating	duration	liste
0	s1	Movie	Dick Johnson Is Dead	United States	9/25/2021	2020	PG-13	90 min	Document
1	s3	TV Show	Ganglands	France	9/24/2021	2021	TV-MA	1 Season	Crim Sh Internat TV Show
2	s6	TV Show	Midnight Mass	United States	9/24/2021	2021	TV-MA	1 Season	TV Drama Horro Myst
3	s14	Movie	Confessions of an Invisible Girl	Brazil	9/22/2021	2021	TV-PG	91 min	Childr Family Mo Come
4	s8	Movie	Sankofa	United States	9/24/2021	1993	TV-MA	125 min	Dra Indepen Mo Internat M
...	...	...	...	...	...	...	...	...	
8785	s8797	TV Show	Yunus Emre	Turkey	1/17/2017	2016	TV-PG	2 Seasons	Internat TV Show Dra
8786	s8798	TV Show	Zak Storm	United States	9/13/2018	2016	TV-Y7	3 Seasons	Kid
8787	s8801	TV Show	Zindagi Gulzar Hai	Pakistan	12/15/2016	2012	TV-PG	1 Season	Internat TV Sh Romant Shows,
8788	s8784	TV Show	Yoko	Pakistan	6/23/2018	2016	TV-Y	1 Season	Kid
8789	s8786	TV Show	YOM	Pakistan	6/7/2018	2016	TV-Y7	1 Season	Kid

8790 rows × 9 columns

but if we consider the "country" it also occupies nearly 1/30 th part of data is "Not Given". Here handling it is a bit odd because we can't fill the mean or median since it is categorical data.

Here we can follow many approaches such as filling it with mode or Dropping Rows or Creating a "Missing" Category.

but for applying mode the data is not suitable since there are many coutries it leads to unskewd data , similarly as above

mentioned the data is very small part so we can't drop the entire column.

so we are going to leave it as same as "Not Given " column

we can also apply many approaches such as Use a Machine Learning Algorithm,K-Nearest Neighbors (K-NN) Imputation,Probabilistic Imputation(by Z-test) etc, but let's focus on remaining it in the columun only.

In [53]:
```
df
```

Out[53]:

	show_id	type	title	country	date_added	release_year	rating	duration	liste
0	s1	Movie	Dick Johnson Is Dead	United States	9/25/2021	2020	PG-13	90 min	Document
1	s3	TV Show	Ganglands	France	9/24/2021	2021	TV-MA	1 Season	Crim Sh Internat TV Show
2	s6	TV Show	Midnight Mass	United States	9/24/2021	2021	TV-MA	1 Season	TV Drama Horro Myst
3	s14	Movie	Confessions of an Invisible Girl	Brazil	9/22/2021	2021	TV-PG	91 min	Childr Family Mo Come
4	s8	Movie	Sankofa	United States	9/24/2021	1993	TV-MA	125 min	Dra Indepen Mo Internat M
...	...	...	...	...	...	...	...	...	
8785	s8797	TV Show	Yunus Emre	Turkey	1/17/2017	2016	TV-PG	2 Seasons	Internat TV Show Dra
8786	s8798	TV Show	Zak Storm	United States	9/13/2018	2016	TV-Y7	3 Seasons	Kid
8787	s8801	TV Show	Zindagi Gulzar Hai	Pakistan	12/15/2016	2012	TV-PG	1 Season	Internat TV Sh Romant Shows,
8788	s8784	TV Show	Yoko	Pakistan	6/23/2018	2016	TV-Y	1 Season	Kid
8789	s8786	TV Show	YOM	Pakistan	6/7/2018	2016	TV-Y7	1 Season	Kid

8790 rows × 9 columns

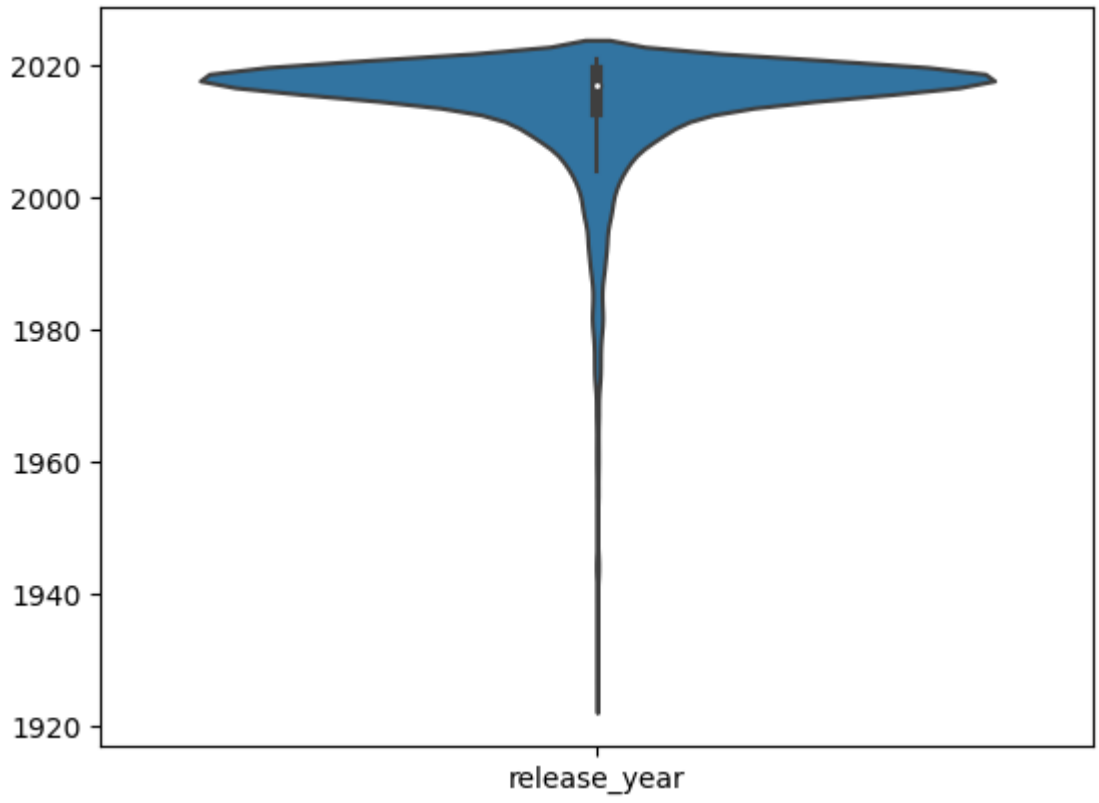
Removing Outliers

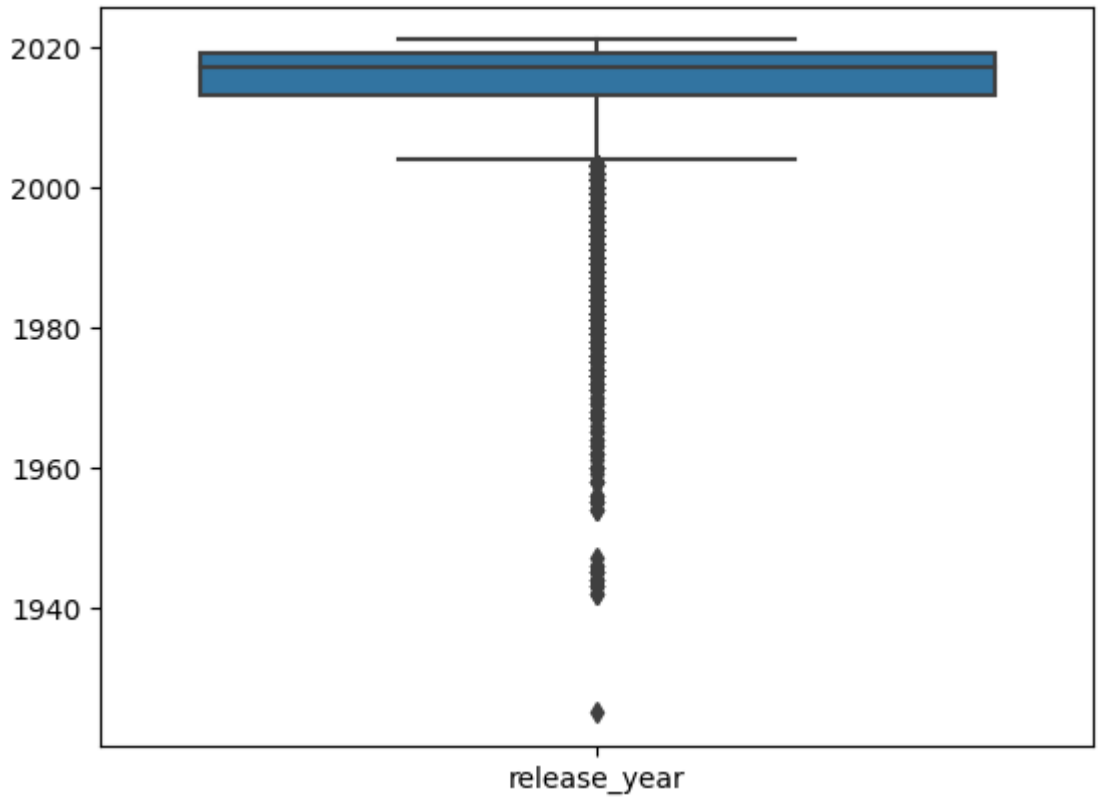
In [45]:
```
#let's see the outliers with the help of plotting.
#importing seaborn library
import seaborn as sn
```

In [65]:
```
#here in the data release year is only in the numerical form so let's visualize it.
#print(df['release_year'].describe())
print(df.describe())
```

```
       release_year
count   8790.000000
mean    2014.183163
std        8.825466
min     1925.000000
25%     2013.000000
50%     2017.000000
75%     2019.000000
max     2021.000000
```

In [72]:
```
sn.violinplot(data=df)
plt.show()
sn.boxplot(data=df)
plt.show()
```





here we are going to remove the outliers in the data by using Z-static

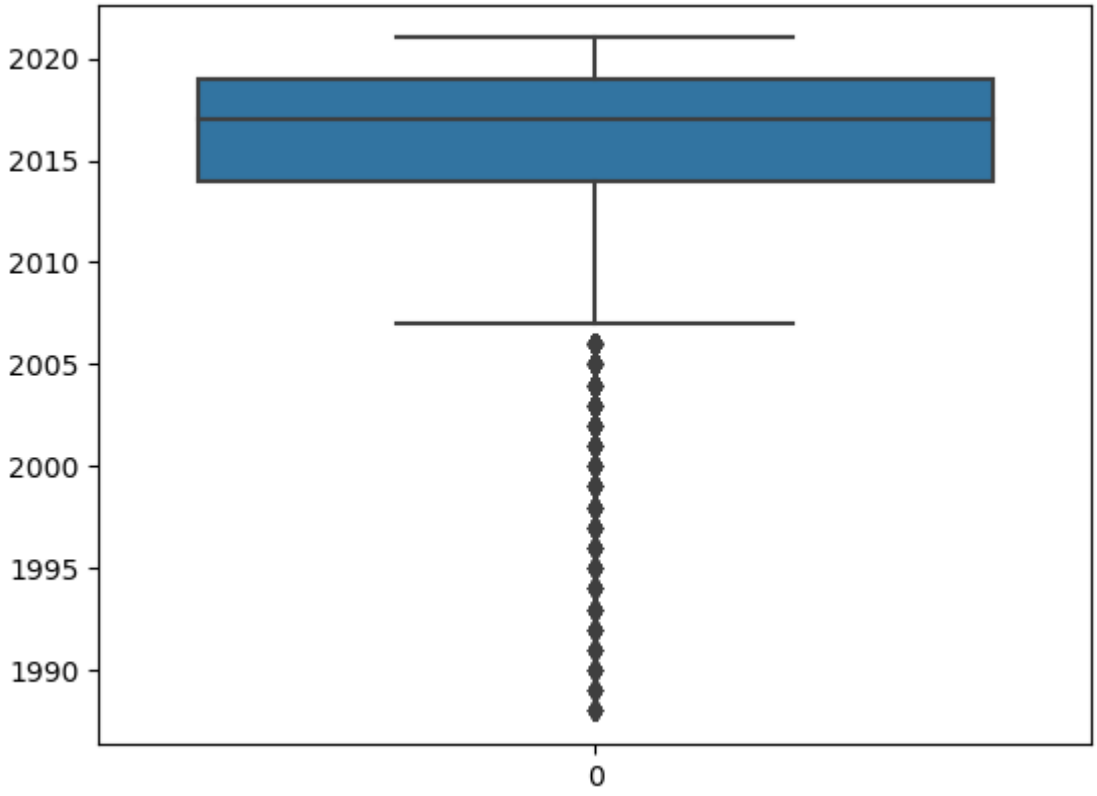
In [73]:
```
df1=df
```

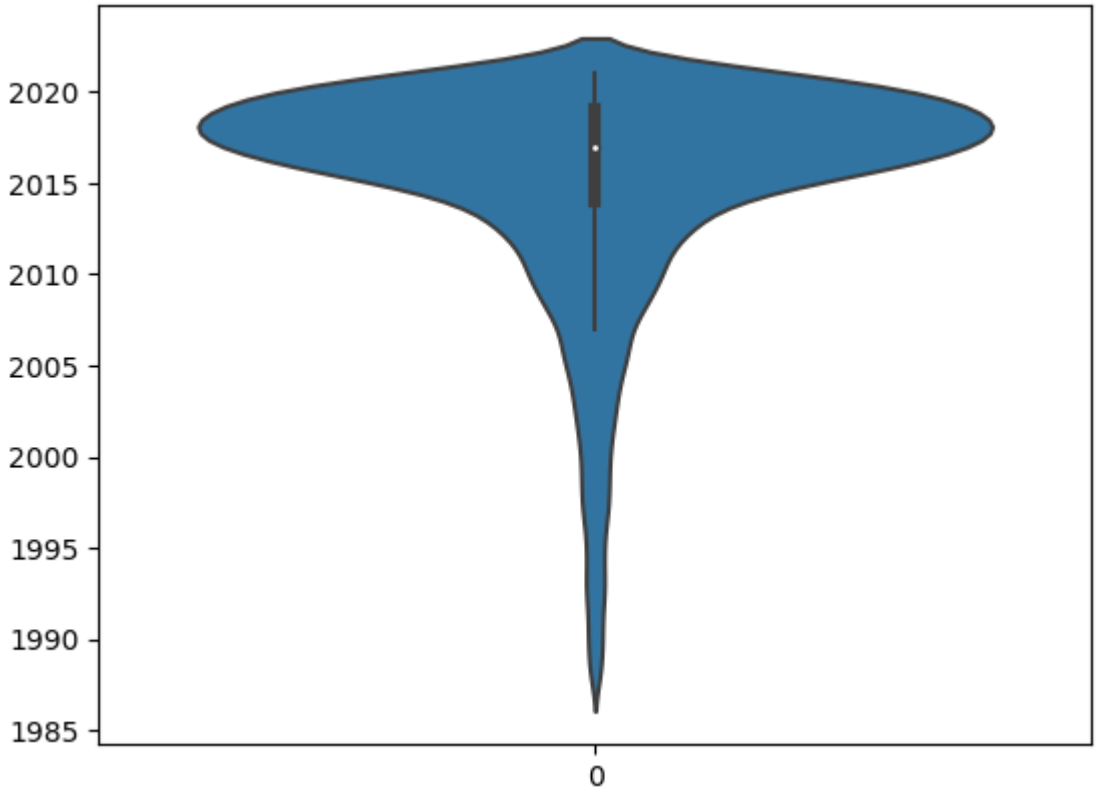
In [96]:
```
from scipy import stats
z_scores = stats.zscore(df1['release_year'])
outliers = df1[(z_scores < -3) | (z_scores > 3)]
print(len(outliers))
m=list(df1['release_year'])
for i in outliers['release_year']:
    if i in m:
        m.remove(i)

# print(m)
print(len(m))
print(min(m))
print(max(m))
```

```
217
8573
1988
2021
```

In [92]:
```
sn.boxplot(data=m)
plt.show()
sn.violinplot(data = m)
plt.show()
```





here we have completed the removing of outliers. so, our task is completed.

In [ ]: