TASK-2

PROBLEM STATEMENT - Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice.Explore the relationships between variables and identify patterns and trends in the data.

Source:-Kaggle

Description:- Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original contents. This dataset is a cleaned version of the original version which can be found here. The data consist of contents added to Netflix from 2008 to 2021. The oldest content is as old as 1925 and the newest as 2021.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
import seaborn as sns
```

```
data=pd.read_csv('netflix1.csv')
data
```

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

0 rows × 10 columns

```
df = pd.DataFrame(data)



# Treat the duplicates
df = df.drop_duplicates()

# Populate missing rows
df.fillna("Unknown", inplace=True)


# Drop unnecessary columns if they exist
if "director" in df.columns:
    df = df.drop(columns=["director"])

print(df)
```

```
     date_added  release_year rating   duration  \
0     9/25/2021          2020  PG-13     90 min   
1     9/24/2021          2021  TV-MA   1 Season   
2     9/24/2021          2021  TV-MA   1 Season   
3     9/22/2021          2021  TV-PG     91 min   
4     9/24/2021          1993  TV-MA    125 min   
...         ...           ...    ...        ...   
8785  1/17/2017          2016  TV-PG  2 Seasons   
8786  9/13/2018          2016  TV-Y7  3 Seasons   
8787 12/15/2016          2012  TV-PG   1 Season   
8788  6/23/2018          2016   TV-Y   1 Season   
8789   6/7/2018          2016  TV-Y7   1 Season   

                                              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   
...                                                 ...   
8785                    International TV Shows, TV Dramas   
8786                                           Kids' TV   
8787  International TV Shows, Romantic TV Shows, TV ...   
8788                                           Kids' TV   
8789                                           Kids' TV   

                    category1                category2  \
0               Documentaries                  Unknown   
1              Crime TV Shows   International TV Shows   
2                   TV Dramas                TV Horror   
3     Children & Family Movies                 Comedies   
4                      Dramas       Independent Movies   
...                       ...                      ...   
8785   International TV Shows                TV Dramas   
8786                 Kids' TV                  Unknown   
8787   International TV Shows        Romantic TV Shows   
8788                 Kids' TV                  Unknown   
8789                 Kids' TV                  Unknown   

                  category3  
0                   Unknown  
1      TV Action & Adventure  
2              TV Mysteries  
3                   Unknown  
4      International Movies  
...                     ...  
8785                Unknown  
8786                Unknown  
8787              TV Dramas  
8788                Unknown  
8789                Unknown  

[8790 rows x 12 columns]
```

Double-click (or enter) to edit

```
Exploratory Data Analysis (EDA)
```

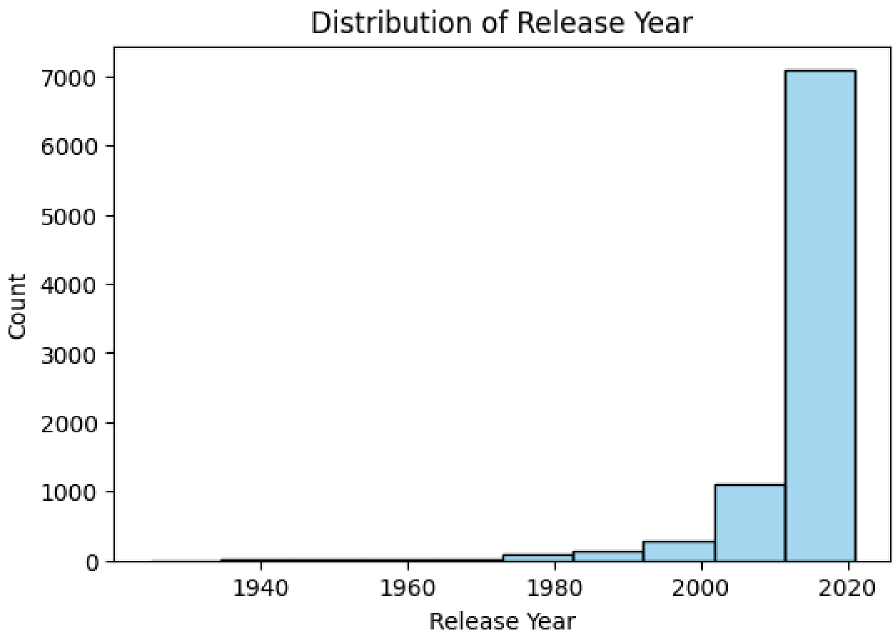
```
# Display basic information about the dataset
print("Basic information about the dataset:")
print(df.info())

# Summary statistics for numerical columns
print("\nSummary statistics for numerical columns:")
print(df.describe())
```

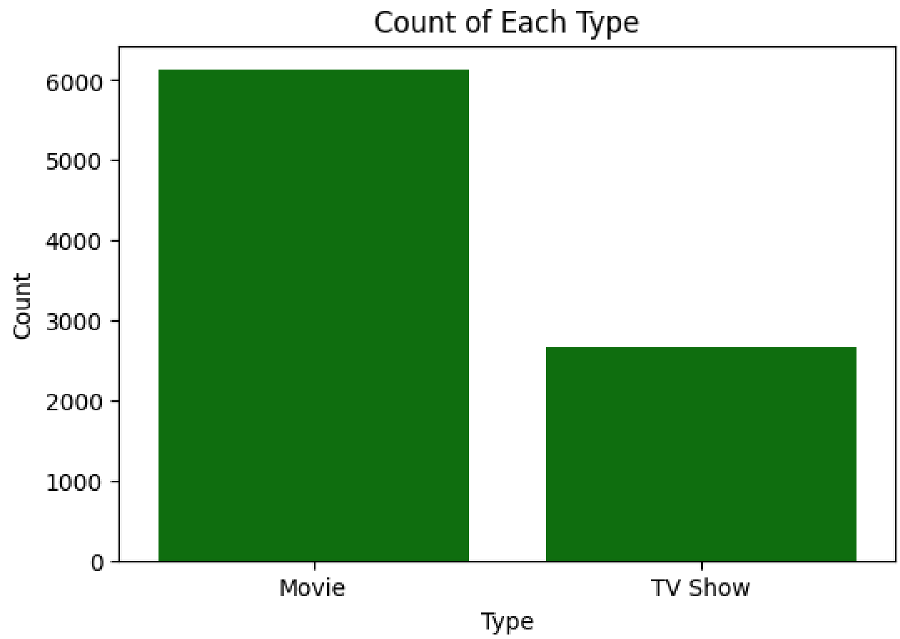
```
Basic information about the dataset:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8790 entries, 0 to 8789
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   country       8790 non-null   object
 4   date_added    8790 non-null   object
 5   release_year  8790 non-null   int64 
 6   rating        8790 non-null   object
 7   duration      8790 non-null   object
 8   listed_in     8790 non-null   object
 9   category1     8790 non-null   object
 10  category2     8790 non-null   object
 11  category3     8790 non-null   object
dtypes: int64(1), object(11)
memory usage: 892.7+ KB
None

Summary statistics for numerical columns:
       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
```

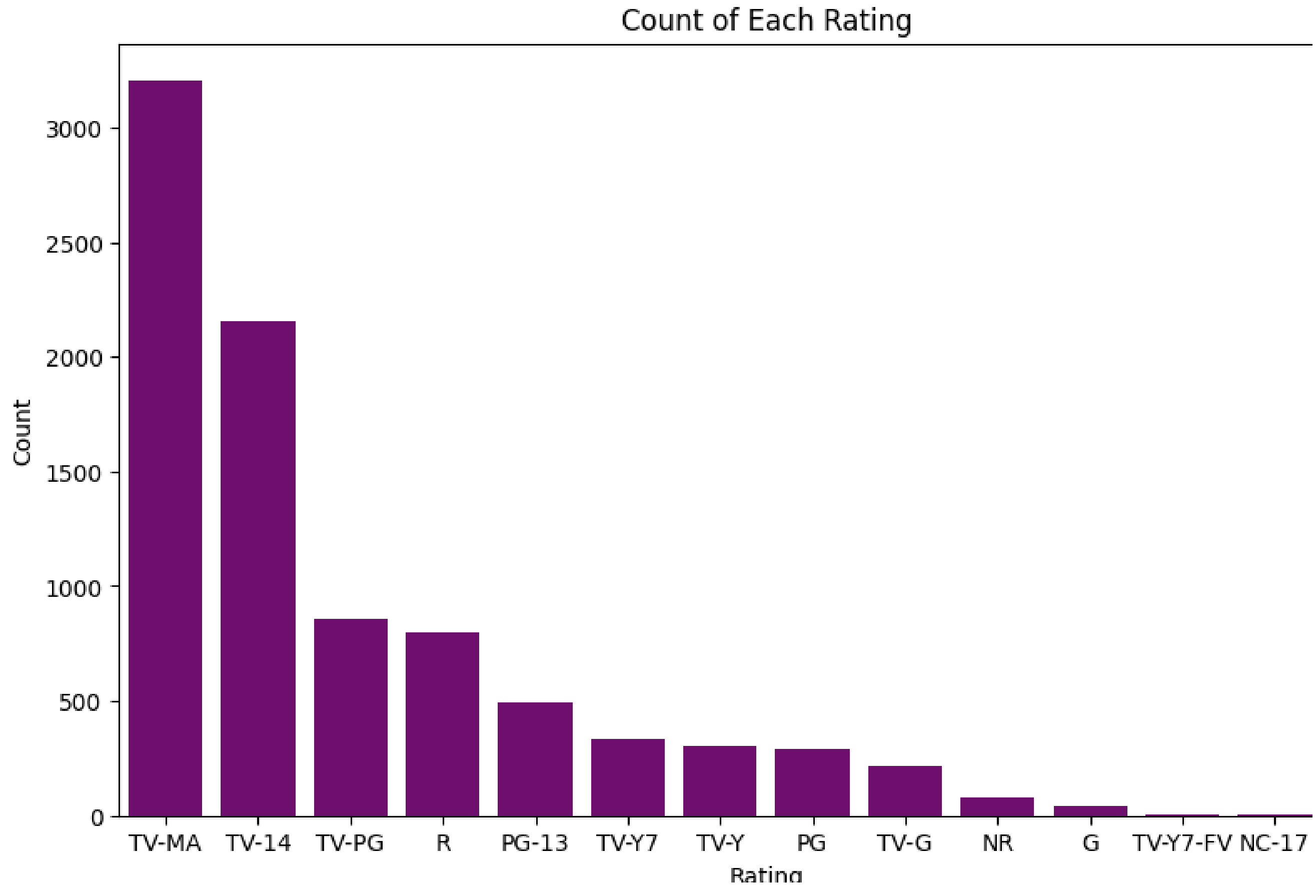
```
# Plotting the distribution of 'release_year'
plt.figure(figsize=(6, 4))
sns.histplot(data=df, x='release_year', bins=10,color='skyblue')
plt.title('Distribution of Release Year')
plt.xlabel('Release Year')
plt.ylabel('Count')
plt.show()
```



```
# Plotting the count of each type (Movie vs. TV Show)
plt.figure(figsize=(6, 4))
sns.countplot(data=df, x='type',color='g')
sns.countplot(data=df, x='type',color='g')
plt.title('Count of Each Type')
plt.xlabel('Type')
plt.ylabel('Count')
plt.show()
```



```
# Plotting the count of each rating
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='rating', order=df['rating'].value_counts().index,color='purple')
plt.title('Count of Each Rating')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.show()
```


Count of Each Rating
Count
TV-MA
TV-14
TV-PG
R
PG-13
TV-Y7
TV-Y
PG
TV-G
NR
G
TV-Y7-FV
NC-17
Rating