Netflix

About

Netflix is a subscription-based popular streaming service that offers a vast catalog of movies, TV shows, and original contents.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.

Importing Required Modules

1- importing numpy for mathematical operation on arrays and dataframe.

2- importing pandas for reading data and data manipualtion.

3- importing matplotlib and seaborn to show the insights and visualization from the dataset.

4- importing warnings for Warning messages that are typically issued in dataframe where it is useful to alert the user of some condition in a program, where that condition (normally) doesn t warrant raising an exception and terminating the program.

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

Reading Dataset

1- using pandas read_csv to load the data

2- After loading it is important to check the complete information of data as it can indication many of the hidden infomation such as null values in a column or a row

In [2]:

```
pd.read_csv("netflix1.csv")
```

Out[2]:

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

8790 rows × 10 columns

In [3]:

```
# Assigning it to a variable df

df=pd.read_csv("netflix1.csv")
```

Finding the basic information

In [4]:

```
df.head()
```

Out[4]:

	show_id	type	title	director	country	date_added	release_year	rating	duratio
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 mi
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	Seaso
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	Seaso
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 mi
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 mi

In [5]:

```
# Finding the rows and columns

df.shape
```

Out[5]:

```
(8790, 10)
```

In [6]:

```
# Column information

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
```

In [7]:

```
# Basic statistical analysis of integer column

df.describe()
```

Out[7]:

	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 [8]:

```
# Basic statistical analysis of categorical column

df.describe(include=object)
```

Out[8]:

	show_id	type	title	director	country	date_added	rating	duration	listed_in
count	8790	8790	8790	8790	8790	8790	8790	8790	8790
unique	8790	2	8787	4528	86	1713	14	220	513
top	s1	Movie	9-Feb	Not Given	United States	1/1/2020	TV-MA	1 Season	Dramas, International Movies
freq	1	6126	2	2588	3240	110	3205	1791	362

In [9]:

```
# Finding the null values

df.isnull().sum()
```

Out[9]:

```
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 [10]:

```
# As date_added column is object datatype we need to convert it into datetime datatype

df["date_added"]=pd.to_datetime(df["date_added"])
```

In [11]:

```
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   datetime64[ns]
 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: datetime64[ns](1), int64(1), object(8)
memory usage: 686.8+ KB
```

In [12]:

```
# Changing columns name for simplicity

df.columns
```

Out[12]:

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

In [13]:

```
df.columns = ["Show_id","Type","Title","Director","Country","Added_date","Year_of_releas
              "Duration","Catalogued"]
```

In [14]:

```
df.head()
```

Out[14]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating	Du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	S
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA	S
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG	
4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA	1

In [15]:

```
df["Type"].unique()
```

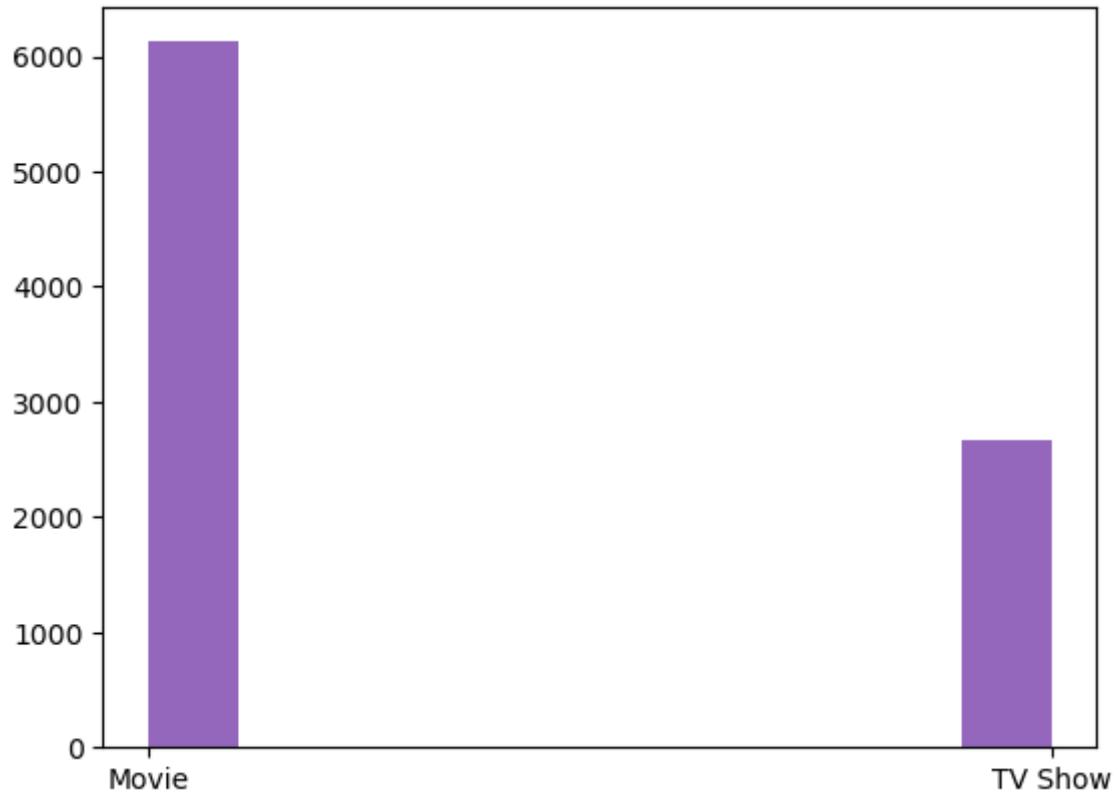
Out[15]:

```
array(['Movie', 'TV Show'], dtype=object)
```

Are movies more or TV shows

In [16]:

```
plt.hist(df["Type"],color="#9467bd");
```



In [17]:

```
L=[]
for i in df['Catalogued']:
    L.append(i.split(',')[0])
df['Basic category']=L
```

In [18]:

```
df['Basic category'].nunique()
```

Out[18]:

```
36
```

In [19]:

```
df[df["Type"]=="Movie"]
```

Out[19]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG
4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA
6	s10	Movie	The Starling	Theodore Melfi	United States	2021-09-24	2021	PG-13
7	s939	Movie	Motu Patlu in the Game of Zones	Suhas Kadav	India	2021-05-01	2019	TV-Y7
...	...	...	...	...	...	...	...	...
8702	s8232	Movie	The Bund	Not Given	Hong Kong	2018-09-20	1983	TV-14
8707	s8269	Movie	The Darkest Dawn	Not Given	United Kingdom	2018-06-23	2016	TV-MA
8716	s8331	Movie	The Great Battle	Not Given	South Korea	2019-04-08	2018	TV-MA
8763	s8648	Movie	Twisted Trunk, Big Fat Body	Not Given	India	2017-01-15	2015	TV-14
8783	s8785	Movie	Yoko and His Friends	Not Given	Russia	2018-06-23	2015	TV-Y

6126 rows × 11 columns

In [20]:

```
# Assigning this to a variable

mv=df[df["Type"]=="Movie"]
```

In [21]:

```
mv["Type"].value_counts()
```

Out[21]:

```
Movie    6126
Name: Type, dtype: int64
```

In [22]:

```
df[df["Type"]=="TV Show"]
```

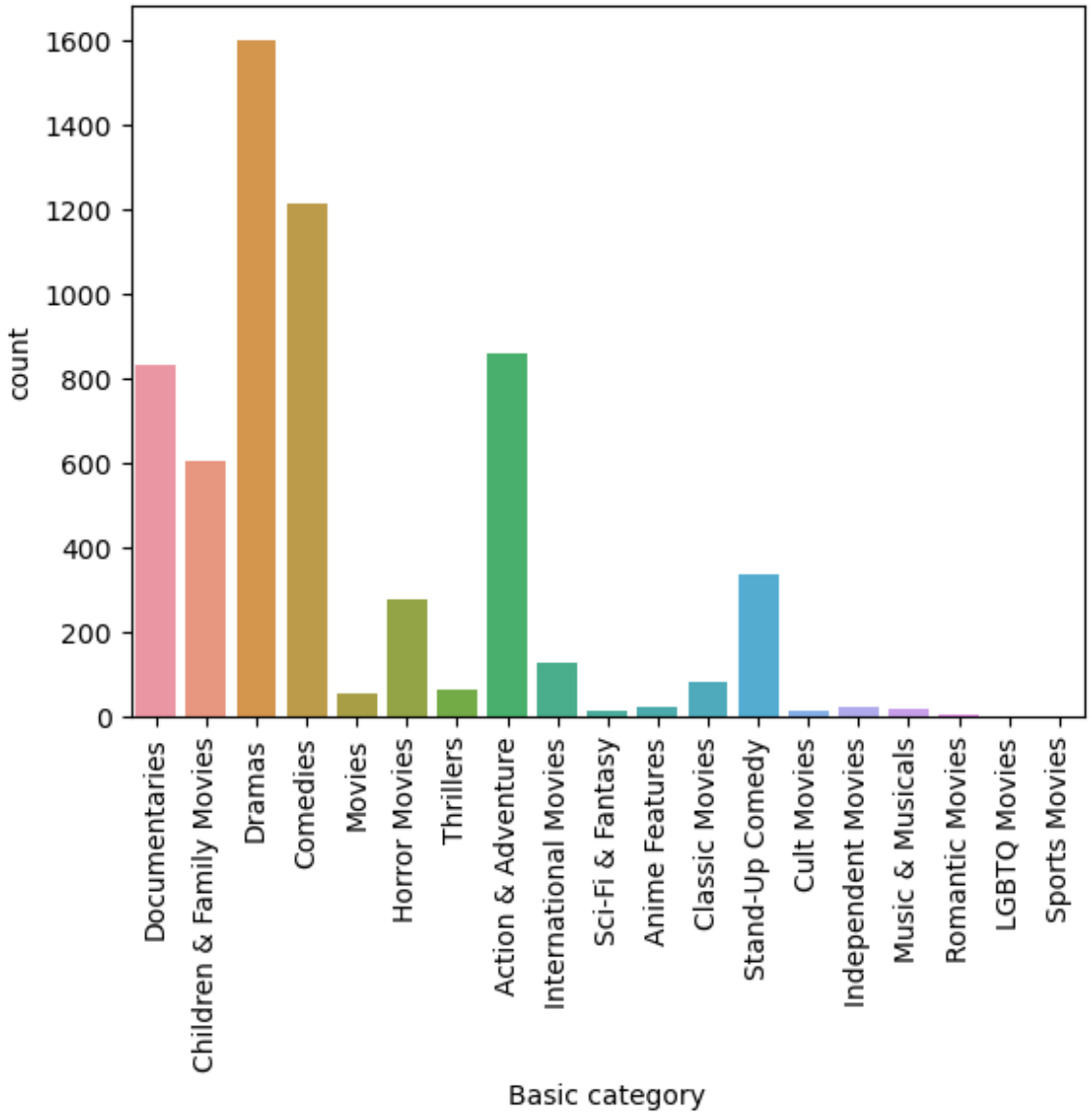
Out[22]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA
5	s9	TV Show	The Great British Baking Show	Andy Devonshire	United Kingdom	2021-09-24	2021	TV-14
17	s4	TV Show	Jailbirds New Orleans	Not Given	Pakistan	2021-09-24	2021	TV-MA
18	s15	TV Show	Crime Stories: India Detectives	Not Given	Pakistan	2021-09-22	2021	TV-MA
...	...	...	...	...	...	...	...	...
8785	s8797	TV Show	Yunus Emre	Not Given	Turkey	2017-01-17	2016	TV-PG
8786	s8798	TV Show	Zak Storm	Not Given	United States	2018-09-13	2016	TV-Y7
8787	s8801	TV Show	Zindagi Gulzar Hai	Not Given	Pakistan	2016-12-15	2012	TV-PG
8788	s8784	TV Show	Yoko	Not Given	Pakistan	2018-06-23	2016	TV-Y
8789	s8786	TV Show	YOM	Not Given	Pakistan	2018-06-07	2016	TV-Y7

2664 rows × 11 columns

In [23]:

```
sns.countplot(mv['Basic category'])
plt.xticks(rotation=90);
```



In [24]:

```
# Assigning it to a variable

tv=df[df["Type"]=="TV Show"]
```

In [25]:

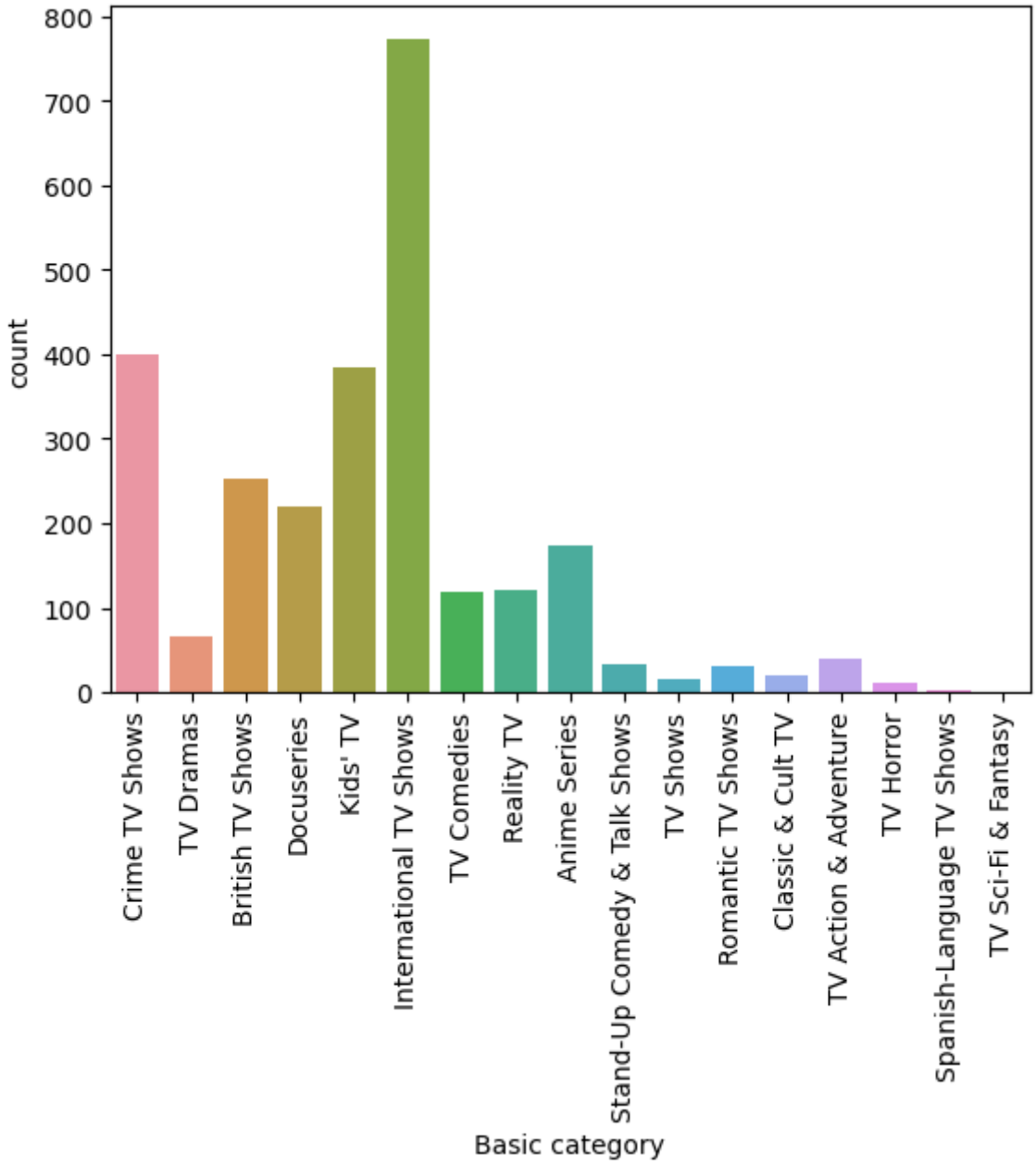
```
tv["Type"].value_counts()
```

Out[25]:

```
TV Show    2664
Name: Type, dtype: int64
```

In [26]:

```
sns.countplot(tv['Basic category'])
plt.xticks(rotation=90);
```



Indian TV shows and movies on netflix

In [27]:

```
# 1057 tv shows and movies are available on netflix

df[df["Country"]=="India"]
```

Out[27]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating
7	s939	Movie	Motu Patlu in the Game of Zones	Suhas Kadav	India	2021-05-01	2019	TV-Y7
9	s940	Movie	Motu Patlu in Wonderland	Suhas Kadav	India	2021-05-01	2013	TV-Y7
10	s941	Movie	Motu Patlu: Deep Sea Adventure	Suhas Kadav	India	2021-05-01	2014	TV-Y7
11	s942	Movie	Motu Patlu: Mission Moon	Suhas Kadav	India	2021-05-01	2013	TV-Y7
29	s25	Movie	Jeans	S. Shankar	India	2021-09-21	1998	TV-14
...	...	...	...	...	...	...	...	...
8715	s8322	TV Show	The Golden Years with Javed Akhtar	Not Given	India	2017-06-01	2016	TV-G
8721	s8350	TV Show	The House That Made Me	Not Given	India	2017-03-31	2015	TV-PG
8724	s8374	TV Show	The Jungle Book	Not Given	India	2019-05-11	2010	TV-Y7
8763	s8648	Movie	Twisted Trunk, Big Fat Body	Not Given	India	2017-01-15	2015	TV-14
8781	s8776	TV Show	Yeh Meri Family	Not Given	India	2018-08-31	2018	TV-PG

1057 rows × 11 columns

US and UK all types on netflix

In [28]:

```
df[(df["Country"]=="United States") | (df["Country"]=="United Kingdom")]
```

Out[28]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA
4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA
5	s9	TV Show	The Great British Baking Show	Andy Devonshire	United Kingdom	2021-09-24	2021	TV-14
6	s10	Movie	The Starling	Theodore Melfi	United States	2021-09-24	2021	PG-13
...	...	...	...	...	...	...	...	...
8777	s8748	TV Show	Winsanity	Not Given	United States	2018-12-15	2016	TV-G
8779	s8756	TV Show	Women Behind Bars	Not Given	United States	2016-11-01	2010	TV-14
8780	s8759	TV Show	World's Busiest Cities	Not Given	United Kingdom	2019-02-01	2017	TV-PG
8782	s8781	TV Show	Yo-Kai Watch	Not Given	United States	2016-04-01	2015	TV-Y7
8786	s8798	TV Show	Zak Storm	Not Given	United States	2018-09-13	2016	TV-Y7

3878 rows × 11 columns

Supernatural tv show catalogued

In [29]:

```
df[df["Title"]=="Supernatural"]
```

Out[29]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating
1971	s2424	TV Show	Supernatural	Phil Sgriccia	United States	2020-06-05	2019	TV-14

Finding the ratings

In [30]:

```
df["Rating"].unique()
```

Out[30]:

```
array(['PG-13', 'TV-MA', 'TV-PG', 'TV-14', 'TV-Y7', 'TV-Y', 'PG', 'TV-G',
       'R', 'G', 'NC-17', 'NR', 'TV-Y7-FV', 'UR'], dtype=object)
```

In [31]:

```
# TV-MA rating level (which includes dark humor and intense violence) tv shows and movie

r=df.groupby(["Type"])["Rating"].value_counts()
```

In [32]:

```
r
```

Out[32]:

```
Type     Rating
Movie    TV-MA       2062
         TV-14       1427
         R            797
         TV-PG        540
         PG-13        490
         PG           287
         TV-Y7        139
         TV-Y         131
         TV-G         126
         NR            75
         G             41
         TV-Y7-FV       5
         NC-17          3
         UR             3
TV Show  TV-MA       1143
         TV-14        730
         TV-PG        321
         TV-Y7        194
         TV-Y         175
         TV-G          94
         NR             4
         R              2
         TV-Y7-FV       1
Name: Rating, dtype: int64
```

Movies added in lockdown period(2020)

In [33]:

```
# Extracting year from date added

df["Added_year"]=df["Added_date"].dt.year
```

In [34]:

```
df.head(1)
```

Out[34]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating	Duratic
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 m

In [35]:

```
df[df["Added_year"]==2020]
```

Out[35]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating
226	s1472	Movie	Best of Stand-Up 2020	Not Given	Pakistan	2020-12-31	2020	TV-MA
231	s1483	TV Show	Transformers: War for Cybertron: Earthrise	Not Given	Pakistan	2020-12-30	2020	TV-Y7
232	s1586	TV Show	Manhunt: Deadly Games	Not Given	Pakistan	2020-12-07	2020	TV-14
233	s1597	TV Show	The Great British Baking Show: Holidays	Not Given	Pakistan	2020-12-04	2020	TV-MA
234	s1669	TV Show	Heart & Soul	Not Given	Pakistan	2020-11-20	2019	TV-14
...	...	...	...	...	...	...	...	...
8673	s8063	TV Show	Space Racers	Not Given	United States	2020-03-31	2014	TV-Y
8685	s8126	TV Show	Super Wings	Not Given	United States	2020-12-01	2020	TV-Y
8687	s8133	TV Show	Surviving R. Kelly Part II: The Reckoning	Not Given	United States	2020-04-13	2020	TV-MA
8722	s8367	TV Show	The Investigator: A British Crime Story	Not Given	United Kingdom	2020-03-05	2018	TV-MA
8762	s8647	TV Show	Twirlywoos	Not Given	United Kingdom	2020-05-15	2018	TV-Y

1879 rows × 12 columns

In [36]:

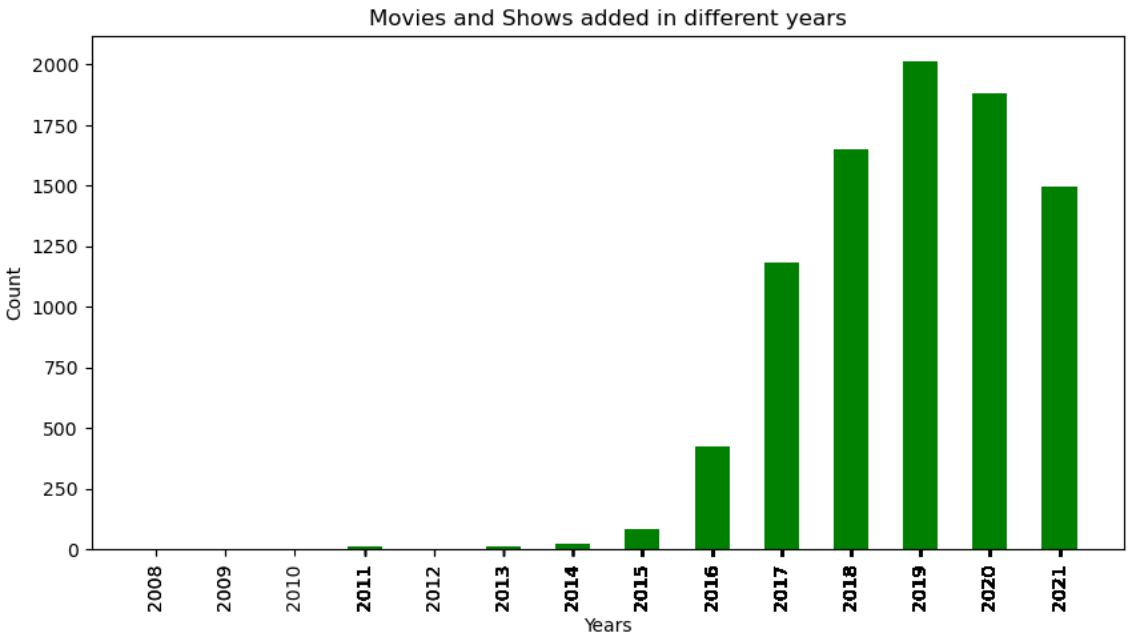
```
df[["Added_year"]].value_counts()
```

Out[36]:

```
Added_year
2019          2016
2020          1879
2018          1648
2021          1498
2017          1185
2016           426
2015            82
2014            24
2011            13
2013            11
2012             3
2008             2
2009             2
2010             1
dtype: int64
```

In [37]:

```
x=[2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019,2020,2021]
y=[2,2,1,13,3,11,24,82,426,1185,1648,2016,1879,1498]
plt.figure(figsize=(10,5))
plt.bar(x,y,color="g",width=0.5)
plt.xlabel("Years")
plt.ylabel("Count")
plt.xticks(df["Added_year"],rotation=90)
plt.title("Movies and Shows added in different years")
plt.show()
```



In [38]:

```
df[(df["Added_year"]==2008) | (df["Added_year"]==2009) | (df["Added_year"]==2010)]
```

Out[38]:

	Show_id	Type	Title	Director	Country	Added_date	Year_of_release	Rating	D
4251	s5956	Movie	Splatter	Joe Dante	United States	2009-11-18	2009	TV-MA	
4252	s5957	Movie	Just Another Love Story	Ole Bornedal	Denmark	2009-05-05	2007	TV-MA	
4253	s5958	Movie	To and From New York	Sorin Dan Mihalcescu	United States	2008-01-01	2006	TV-MA	
5381	s7371	Movie	Mad Ron's Prevues from Hell	Jim Monaco	United States	2010-11-01	1987	NR	
8423	s6612	TV Show	Dinner for Five	Not Given	United States	2008-02-04	2007	TV-MA	

Different countries tv shows/movies availability on netflix

In [39]:

```
df["Country"].unique()
```

Out[39]:

```
array(['United States', 'France', 'Brazil', 'United Kingdom', 'India',
       'Germany', 'Pakistan', 'Not Given', 'China', 'South Africa',
       'Japan', 'Nigeria', 'Spain', 'Philippines', 'Australia',
       'Argentina', 'Canada', 'Hong Kong', 'Italy', 'New Zealand',
       'Egypt', 'Colombia', 'Mexico', 'Belgium', 'Switzerland', 'Taiwan',
       'Bulgaria', 'Poland', 'South Korea', 'Saudi Arabia', 'Thailand',
       'Indonesia', 'Kuwait', 'Malaysia', 'Vietnam', 'Lebanon', 'Romania',
       'Syria', 'United Arab Emirates', 'Sweden', 'Mauritius', 'Austria',
       'Turkey', 'Czech Republic', 'Cameroon', 'Netherlands', 'Ireland',
       'Russia', 'Kenya', 'Chile', 'Uruguay', 'Bangladesh', 'Portugal',
       'Hungary', 'Norway', 'Singapore', 'Iceland', 'Serbia', 'Namibia',
       'Peru', 'Mozambique', 'Ghana', 'Zimbabwe', 'Israel', 'Finland',
       'Denmark', 'Paraguay', 'Cambodia', 'Georgia', 'Soviet Union',
       'Greece', 'West Germany', 'Iran', 'Venezuela', 'Slovenia',
       'Guatemala', 'Jamaica', 'Somalia', 'Croatia', 'Jordan',
       'Luxembourg', 'Senegal', 'Belarus', 'Puerto Rico', 'Cyprus',
       'Ukraine'], dtype=object)
```

In [40]:

```
df1=df.groupby(["Type"])["Country"].value_counts()
```

In [41]:

```
df1
```

Out[41]:

```
Type     Country
Movie    United States           2395
         India                    976
         United Kingdom           387
         Not Given                257
         Canada                   187
                                 ...
TV Show  Puerto Rico                1
         Senegal                    1
         Switzerland                1
         United Arab Emirates       1
         Uruguay                    1
Name: Country, Length: 138, dtype: int64
```

Finding the categories

In [42]:

```
df["Basic category"].unique()
```

Out[42]:

```
array(['Documentaries', 'Crime TV Shows', 'TV Dramas',
       'Children & Family Movies', 'Dramas', 'British TV Shows',
       'Comedies', 'Movies', 'Docuseries', "Kids' TV", 'Horror Movies',
       'Thrillers', 'International TV Shows', 'TV Comedies', 'Reality TV',
       'Anime Series', 'Action & Adventure', 'International Movies',
       'Sci-Fi & Fantasy', 'Anime Features', 'Classic Movies',
       'Stand-Up Comedy', 'Stand-Up Comedy & Talk Shows', 'TV Shows',
       'Romantic TV Shows', 'Cult Movies', 'Independent Movies',
       'Classic & Cult TV', 'Music & Musicals', 'Romantic Movies',
       'LGBTQ Movies', 'TV Action & Adventure', 'TV Horror',
       'Spanish-Language TV Shows', 'TV Sci-Fi & Fantasy',
       'Sports Movies'], dtype=object)
```

In [43]:

```
k=df["Basic category"].value_counts().reset_index()
k
```

Out[43]:

	index	Basic category
0	Dramas	1599
1	Comedies	1210
2	Action & Adventure	859
3	Documentaries	829
4	International TV Shows	773
5	Children & Family Movies	605
6	Crime TV Shows	399
7	Kids' TV	385
8	Stand-Up Comedy	334
9	Horror Movies	275
10	British TV Shows	252
11	Docuseries	220
12	Anime Series	174
13	International Movies	128
14	Reality TV	120
15	TV Comedies	119
16	Classic Movies	80
17	TV Dramas	67
18	Thrillers	65
19	Movies	53
20	TV Action & Adventure	39
21	Stand-Up Comedy & Talk Shows	34
22	Romantic TV Shows	32
23	Anime Features	21
24	Independent Movies	20
25	Classic & Cult TV	20
26	Music & Musicals	18
27	TV Shows	16
28	Sci-Fi & Fantasy	13
29	Cult Movies	12
30	TV Horror	11
31	Romantic Movies	3
32	Spanish-Language TV Shows	2
33	LGBTQ Movies	1
34	TV Sci-Fi & Fantasy	1
35	Sports Movies	1

In [44]:

```
top=k.iloc[0:11]
```

In [45]:

```
top
```

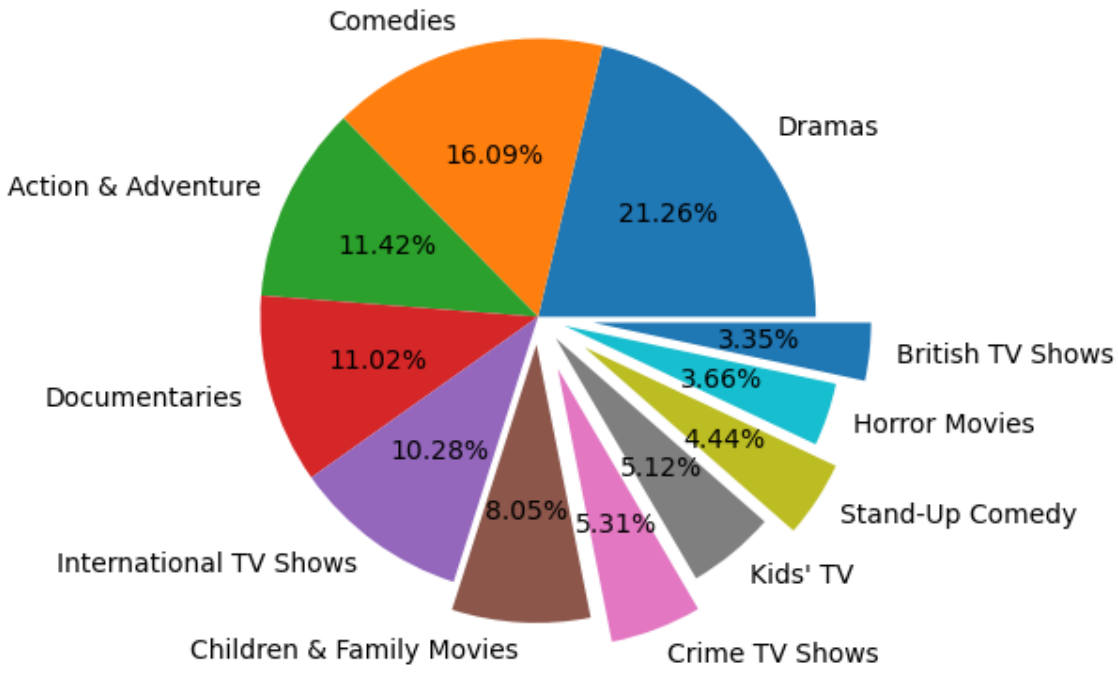
Out[45]:

	index	Basic category
0	Dramas	1599
1	Comedies	1210
2	Action & Adventure	859
3	Documentaries	829
4	International TV Shows	773
5	Children & Family Movies	605
6	Crime TV Shows	399
7	Kids' TV	385
8	Stand-Up Comedy	334
9	Horror Movies	275
10	British TV Shows	252

In [47]:

```
# Shows that people mostly prefers Drama,Comedy,Action & Adventure,Documentaries etc.

y=[0,0,0,0,0,0.1,0.2,0.1,0.2,0.1,0.2]
plt.pie(top["Basic category"],labels=top["index"],autopct='%0.2f%%',explode=y);
```



In [ ]:

```

```

In [ ]:

```

```