In [43]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
import matplotlib.pyplot as plt
```

In [44]:
```
import seaborn as sns
import os
plt.style.use('seaborn-whitegrid')
sns.set_style('whitegrid')
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

In [45]:
```
df = pd.read_csv("D:\\netflix\\netflix_titles_nov_2019.csv")
```

In [46]:
```
df.head()
print("Done Reading!")
```

```
Done Reading!
```

In [47]:
```
def data_inv(df):
    print('netflix movies and shows: ',df.shape[0])
    print('dataset variables: ',df.shape[1])
    print('-'*10)
    print('dateset columns: \n')
    print(df.columns)
    print('-'*10)
    print('data-type of each column: \n')
    print(df.dtypes)
    print('-'*10)
    print('missing rows in each column: \n')
    c=df.isnull().sum()
    print(c[c>0])
data_inv(df)
```

```
netflix movies and shows:  5837
dataset variables:  12
----------
dateset columns:

Index(['show_id', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description',
       'type'],
      dtype='object')
----------
data-type of each column:

show_id          int64
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
type            object
dtype: object
----------
missing rows in each column:

director      1901
cast           556
country        427
date_added     642
rating          10
dtype: int64
```

In [48]:
```
#Data Cleaning
dups=df.duplicated(['title','country','type','release_year'])
df[dups]
```

Out[48]:

show_id	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descrip

	show_id	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descrip
1134	80175351	Kakegurui	NaN	Saori Hayami, Minami Tanaka, Tatsuya Tokutake,...	Japan	NaN	2019	TV-14	2 Seasons	Anime Series, International TV Shows, TV Thril...	High r Yur Jabami p to c ho
1741	81072516	Sarkar	A.R. Murugadoss	Vijay, Varalakshmi Sarathkumar, Keerthi Suresh...	India	March 2, 2019	2018	TV-MA	162 min	Action & Adventure, Dramas, International Movies	A rut businessn missi expose

In [49]:
```
df=df.drop_duplicates(['title','country','type','release_year'])
df=df.drop('show_id',axis=1)
df['cast']=df['cast'].replace(np.nan,'Unknown')
def cast_counter(cast):
    if cast=='Unknown':
        return 0
    else:
        lst=cast.split(', ')
        length=len(lst)
        return length
df['number_of_cast']=df['cast'].apply(cast_counter)
df['cast']=df['cast'].replace('Unknown',np.nan)
```

In [50]:
```
df=df.reset_index()
df['rating']=df['rating'].fillna(df['rating'].mode()[0])
df['date_added']=df['date_added'].fillna('January 1, {}'.format(str(df['release_year'].mode()[0])))
```

In [51]:
```
for i,j in zip(df['country'].values,df.index):
    if i==np.nan:
        if ('Anime' in df.loc[j,'listed_in']) or ('anime' in df.loc[j,'listed_in']):
                df.loc[j,'country']='Japan'
        else:
            continue
    else:
        continue
```

In [52]:
```
import re
months={
    'January':1,
    'February':2,
    'March':3,
    'April':4,
    'May':5,
    'June':6,
    'July':7,
    'August':8,
    'September':9,
    'October':10,
    'November':11,
    'December':12
}
date_lst=[]
for i in df['date_added'].values:
    str1=re.findall('([a-zA-Z]+)\s[0-9]+\,\s[0-9]+',i)
    str2=re.findall('[a-zA-Z]+\s([0-9]+)\,\s[0-9]+',i)
    str3=re.findall('[a-zA-Z]+\s[0-9]+\,\s([0-9]+)',i)
    date='{}-{}-{}'.format(str3[0],months[str1[0]],str2[0])
    date_lst.append(date)
```

In [53]:
```
df['date_added_cleaned']=date_lst
df=df.drop('date_added',axis=1)
df['date_added_cleaned']=df['date_added_cleaned'].astype('datetime64[ns]')
```

In [54]:
```
#EDA Exploratory Data Analysis
#Understand every category in rating column(Google it)
#Understanding what content is available in different countries.
```

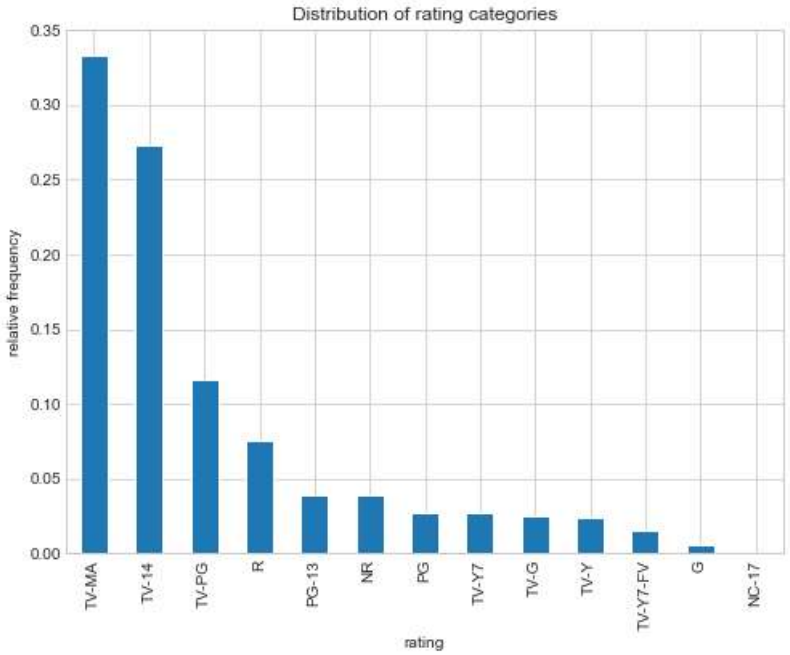
```
#Is Netflix has increasingly focusing on TV rather than movies in recent years.
#The most observed rating categories in TV-shows and Movies
#Identifying similar content by matching text-based features
#How many content its release year differ from its year added
```

In [55]:
```
for i in df.index:
    if df.loc[i,'rating']=='UR':
        df.loc[i,'rating']='NR'
```

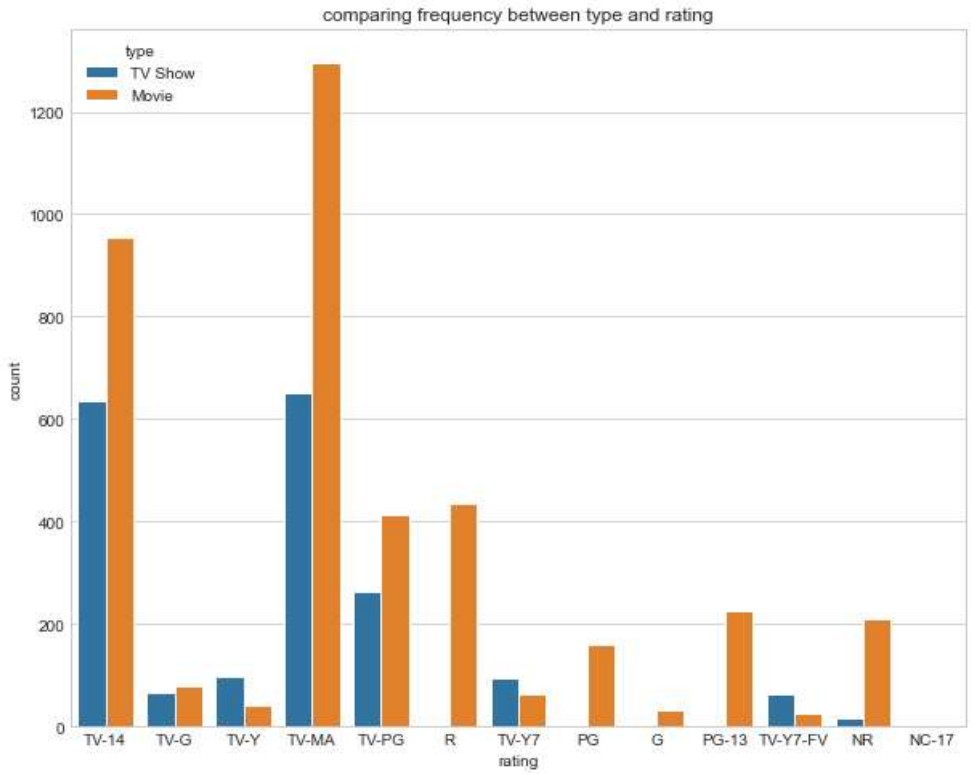
In [56]:
```
#fixing this problem by replace UR category by NR.
```

In [57]:
```
#{TV-MA:This program is specifically designed to be viewed by adults and therefore may be unsuitable for chil
#TV-14:This program contains some material that many parents would find unsuitable for children under 14 year
#TV-PG:This program contains material that parents may find unsuitable for younger children.
##R:Under 17 requires accompanying parent or adult guardian,Parents are urged to learn more about the film be
#PG-13:Some material may be inappropriate for children under 13. Parents are urged to be cautious. Some mater
#NR or UR:If a film has not been submitted for a rating or is an uncut version of a film that was submitted
#PG:Some material may not be suitable for children,May contain some material parents might not like for their
#TV-Y7:This program is designed for children age 7 and above.
#TV-G:This program is suitable for all ages.
#TV-Y:Programs rated TV-Y are designed to be appropriate for children of all ages. The thematic elements port
#TV-Y7-FV:is recommended for ages 7 and older, with the unique advisory that the program contains fantasy vio
#G:All ages admitted. Nothing that would offend parents for viewing by children.
#NC-17:No One 17 and Under Admitted. Clearly adult. Children are not admitted.
#here we discover that UR and NR is the same rating(unrated,Not rated)
#Uncut/extended versions of films that are labeled "Unrated" also contain warnings saying that the uncut vers
#so we have the fix this
```

In [58]:
```
plt.figure(figsize=(8,6))
df['rating'].value_counts(normalize=True).plot.bar()
plt.title('Distribution of rating categories')
plt.xlabel('rating')
plt.ylabel('relative frequency')
plt.show()
```



In [59]:
```
plt.figure(figsize=(10,8))
sns.countplot(x='rating',hue='type',data=df)
plt.title('comparing frequency between type and rating')
plt.show()
```

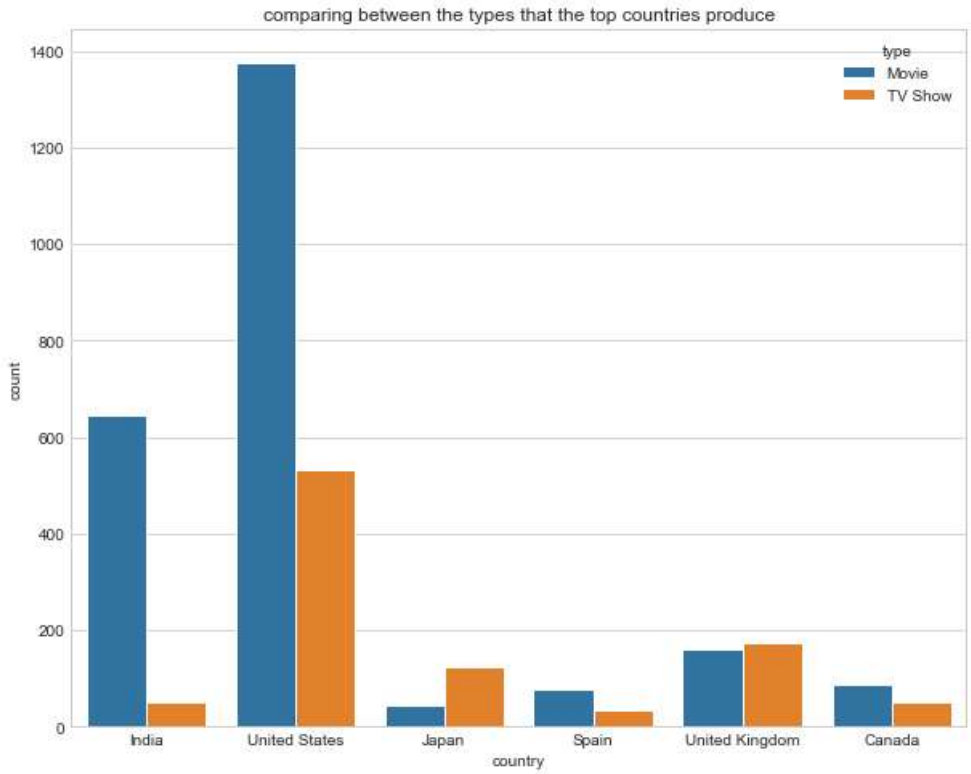


In [60]:
```
#We can say that Movies is the majority category in every rating category on Netflix, except(TV-Y,TV-Y7,TV-Y7

df['country'].value_counts().sort_values(ascending=False)
```

Out[60]:
```
United States                                 1907
India                                          696
United Kingdom                                 336
Japan                                          167
Canada                                         139
                                              ...
United Kingdom, United States, France, Germany   1
Venezuela, Colombia                              1
Germany, United Kingdom, United States           1
United States, United Kingdom, Canada, Japan     1
Italy, Germany                                   1
Name: country, Length: 527, dtype: int64
```

In [61]:
```
top_productive_countries=df[(df['country']=='United States')|(df['country']=='India')|(df['country']=='United
                            (df['country']=='Canada')|(df['country']=='Spain')]
plt.figure(figsize=(10,8))
sns.countplot(x='country',hue='type',data=top_productive_countries)
plt.title('comparing between the types that the top countries produce')
plt.show()
```



UK and Japan produces TV-Shows more than Movies

In [62]:
```
for i in top_productive_countries['country'].unique():
    print(i)
    print(top_productive_countries[top_productive_countries['country']==i]['rating'].value_counts(normalize=T
    print('-'*10)
```

```
India
TV-14       53.160920
TV-MA       23.419540
TV-PG       17.097701
NR           2.873563
TV-G         1.005747
TV-Y7        0.718391
PG-13        0.574713
PG           0.431034
TV-Y7-FV     0.287356
R            0.287356
TV-Y         0.143678
Name: rating, dtype: float64
----------
United States
TV-MA       31.515469
TV-14       17.409544
R           12.165705
TV-PG       10.854746
PG-13        7.079182
PG           4.876770
NR           4.824331
TV-G         3.198741
TV-Y7        2.621919
TV-Y7-FV     2.097535
TV-Y         2.045097
G            1.258521
NC-17        0.052438
Name: rating, dtype: float64
----------
Japan
TV-14       46.107784
TV-MA       25.748503
TV-PG       10.179641
NR           7.185629
TV-Y7        5.389222
PG-13        1.796407
TV-Y7-FV     1.197605
TV-Y         1.197605
TV-G         0.598802
PG           0.598802
Name: rating, dtype: float64
----------
```

```
Spain
TV-MA    64.601770
TV-14    19.469027
NR        7.079646
TV-PG     3.539823
PG        1.769912
R         1.769912
TV-G      0.884956
TV-Y      0.884956
Name: rating, dtype: float64
----------
United Kingdom
TV-MA       40.773810
TV-14       22.321429
TV-PG       18.452381
R            7.142857
TV-G         6.250000
NR           2.380952
TV-Y7        0.892857
TV-Y         0.595238
PG           0.595238
G            0.297619
TV-Y7-FV     0.297619
Name: rating, dtype: float64
----------
Canada
TV-MA       31.654676
TV-14       17.985612
TV-PG       15.107914
R            7.913669
PG           5.035971
TV-Y7        4.316547
TV-G         4.316547
TV-Y         4.316547
NR           3.597122
TV-Y7-FV     3.597122
PG-13        1.438849
G            0.719424
Name: rating, dtype: float64
----------
```

In [63]:
```
df['year_added']=df['date_added_cleaned'].dt.year
df['type'].value_counts(normalize=True)
```

Out[63]:
```
Movie      0.674893
TV Show    0.325107
Name: type, dtype: float64
```

In [64]:
```
df.groupby('year_added')['type'].value_counts(normalize=True)*100
```

Out[64]:
```
year_added  type
2008        Movie       50.000000
            TV Show     50.000000
2009        Movie      100.000000
2010        Movie      100.000000
2011        Movie      100.000000
2012        Movie       57.142857
            TV Show     42.857143
2013        Movie       66.666667
            TV Show     33.333333
2014        Movie      100.000000
2015        Movie       78.378378
            TV Show     21.621622
2016        Movie       64.077670
            TV Show     35.922330
2017        Movie       77.111486
            TV Show     22.888514
2018        Movie       56.872247
            TV Show     43.127753
2019        Movie       74.158523
            TV Show     25.841477
Name: type, dtype: float64
```

In [65]:
```
#We can say that Netflix begin to focus on TV-Shows,but Movies still has the lead in evey year.

dups=df.duplicated(['title'])
df[dups]['title']
```

Out[65]:
```
212                                  Drive
511                                 Tunnel
1243                             Supergirl
1286                             Limitless
```

```
1706                                Shadow
2360                            Oh My Ghost
2410                               Love O2O
2765                                 Bleach
2801                                One Day
2871                           The Innocents
2932                         The Birth Reborn
3048                            Oh My Ghost
3212                            Us and Them
3433                                   Troy
3483                              Locked Up
3583                                   Love
3585                            The Outsider
3591                                   Benji
3658                                   Solo
3669                            The Silence
3670                            The Silence
3882                               Lovesick
4013                             The Secret
4133                                Top Boy
4223                                    Zoo
4239                                Charmed
4268                               The Code
4301                                Manhunt
4337                                   Love
4377                                 Maniac
4391                      She's Gotta Have It
4442                                Persona
4443                           The Iron Lady
4453                               The Saint
4511                             The In-Laws
4576                                Hostages
4589                               The Oath
4659                                 Tunnel
4689                                 Prince
4857                         Rosario Tijeras
4866                           We Are Family
4900                                Lavender
4908                                   Skins
4949                             Blood Money
5055                                   Tiger
5061                      People You May Know
5122                               The Lovers
5182                                Aquarius
5337    Little Baby Bum: Nursery Rhyme Friends
5531                                    Deep
5536                                Limitless
5595                           Frank and Cindy
5619                               Retribution
5732                  Wet Hot American Summer
5755                                    Life
Name: title, dtype: object
```

In [66]:

```
for i in df[dups]['title'].values:
    print(df[df['title']==i][['title','type','release_year','country']])
    print('-'*40)
```

```
     title   type  release_year        country
101  Drive  Movie          2011  United States
212  Drive  Movie          2019          India
----------------------------------------
      title     type  release_year      country
303   Tunnel  TV Show          2019          NaN
511   Tunnel  TV Show          2017  South Korea
4659  Tunnel    Movie          2016  South Korea
----------------------------------------
         title     type  release_year                        country
492   Supergirl    Movie          1984  United Kingdom, United States
1243  Supergirl  TV Show          2019                  United States
----------------------------------------
         title     type  release_year        country
474   Limitless    Movie          2017          India
1286  Limitless    Movie          2011  United States
5536  Limitless  TV Show          2016  United States
----------------------------------------
      title     type  release_year           country
154   Shadow    Movie          2018  China, Hong Kong
1706  Shadow  TV Show          2019               NaN
----------------------------------------
           title     type  release_year      country
775   Oh My Ghost  TV Show          2015  South Korea
2360  Oh My Ghost  TV Show          2018     Thailand
3048  Oh My Ghost    Movie          2009     Thailand
----------------------------------------
        title     type  release_year country
1563  Love O2O    Movie          2016   China
```

```
2410  Love O2O  TV Show          2016   China
---------------------------------------
      title     type  release_year country
2428  Bleach  TV Show          2006   Japan
2765  Bleach    Movie          2018   Japan
---------------------------------------
       title   type  release_year                       country
2225  One Day  Movie          2011  United States, United Kingdom
2801  One Day  Movie          2016                      Thailand
---------------------------------------
             title     type  release_year         country
2170  The Innocents    Movie          2016  France, Poland
2871  The Innocents  TV Show          2018  United Kingdom
---------------------------------------
               title   type  release_year country
2689  The Birth Reborn  Movie          2018     NaN
2932  The Birth Reborn  Movie          2013  Brazil
---------------------------------------
           title     type  release_year      country
775   Oh My Ghost  TV Show          2015  South Korea
2360  Oh My Ghost  TV Show          2018     Thailand
3048  Oh My Ghost    Movie          2009     Thailand
---------------------------------------
           title   type  release_year         country
3115  Us and Them  Movie          2017  United Kingdom
3212  Us and Them  Movie          2018           China
---------------------------------------
     title     type  release_year  \
510   Troy    Movie          2004
3433  Troy  TV Show          2018

                                          country
510               United States, Malta, United Kingdom
3433  United Kingdom, South Africa, Australia, Unite...
---------------------------------------
         title     type  release_year        country
538   Locked Up  TV Show          2019          Spain
3483  Locked Up    Movie          2017  United States
---------------------------------------
     title     type  release_year          country
2341  Love    Movie          2008        Indonesia
3583  Love  TV Show          2018    United States
4337  Love    Movie          2015  France, Belgium
---------------------------------------
             title   type  release_year        country
591   The Outsider  Movie          2019  United States
3585  The Outsider  Movie          2018  United States
---------------------------------------
      title   type  release_year                             country
3542  Benji  Movie          2018  United Arab Emirates, United States
3591  Benji  Movie          1974                        United States
---------------------------------------
     title   type  release_year country
2050  Solo  Movie          2018   Spain
3658  Solo  Movie          2017   India
---------------------------------------
           title   type  release_year  country
1514  The Silence  Movie          2019  Germany
3669  The Silence  Movie          2017    India
3670  The Silence  Movie          2015    India
---------------------------------------
           title   type  release_year  country
1514  The Silence  Movie          2019  Germany
3669  The Silence  Movie          2017    India
3670  The Silence  Movie          2015    India
---------------------------------------
         title     type  release_year         country
2807  Lovesick  TV Show          2014             NaN
3882  Lovesick  TV Show          2018  United Kingdom
---------------------------------------
          title   type  release_year                   country
2761  The Secret  Movie          2018                      NaN
4013  The Secret  Movie          2006  Australia, United States
---------------------------------------
        title     type  release_year         country
611   Top Boy  TV Show          2019  United Kingdom
4133  Top Boy  TV Show          2011  United Kingdom
---------------------------------------
     title     type  release_year        country
3178   Zoo    Movie          2018          India
4223   Zoo  TV Show          2017  United States
---------------------------------------
        title     type  release_year        country
1240  Charmed  TV Show          2019  United States
4239  Charmed  TV Show          2005  United States
---------------------------------------
         title     type  release_year         country
```

```
3773  The Code  TV Show          2011  United Kingdom
4268  The Code  TV Show          2014       Australia
---------------------------------------
       title     type  release_year           country
3317  Manhunt    Movie          2017  China, Hong Kong
4301  Manhunt  TV Show          2017     United States
---------------------------------------
     title     type  release_year          country
2341  Love    Movie          2008        Indonesia
3583  Love  TV Show          2018    United States
4337  Love    Movie          2015  France, Belgium
---------------------------------------
      title     type  release_year        country
2732  Maniac  TV Show          2018  United States
4377  Maniac  TV Show          2015         Norway
---------------------------------------
                   title     type  release_year        country
1252  She's Gotta Have It  TV Show          2018  United States
4391  She's Gotta Have It    Movie          1986  United States
---------------------------------------
       title     type  release_year      country
1510  Persona  TV Show          2019  South Korea
4442  Persona  TV Show          2015          NaN
---------------------------------------
             title     type  release_year                 country
950   The Iron Lady    Movie          2011  United Kingdom, France
4443  The Iron Lady  TV Show          2009                     NaN
---------------------------------------
         title   type  release_year        country
671   The Saint  Movie          1997  United States
4453  The Saint  Movie          2017  United States
---------------------------------------
          title     type  release_year                        country
1023  The In-Laws    Movie          2003  United States, Germany, Canada
4511  The In-Laws  TV Show          2011                             NaN
---------------------------------------
        title     type  release_year country
3782  Hostages    Movie          2017     NaN
4576  Hostages  TV Show          2016  Israel
---------------------------------------
        title     type  release_year    country
4012  The Oath    Movie          2016    Iceland
4589  The Oath  TV Show          2011  Singapore
---------------------------------------
      title     type  release_year      country
303   Tunnel  TV Show          2019          NaN
511   Tunnel  TV Show          2017  South Korea
4659  Tunnel    Movie          2016  South Korea
---------------------------------------
      title   type  release_year country
3490  Prince  Movie          2010   India
4689  Prince  Movie          1969   India
---------------------------------------
              title     type  release_year   country
1265  Rosario Tijeras  TV Show          2018    Mexico
4857  Rosario Tijeras  TV Show          2010  Colombia
---------------------------------------
             title   type  release_year           country
2697  We Are Family  Movie          2010  India, Australia
4866  We Are Family  Movie          2016   France, Belgium
---------------------------------------
        title     type  release_year                country
2025  Lavender  TV Show          2002                 Taiwan
4900  Lavender    Movie          2016  Canada, United States
---------------------------------------
     title     type  release_year         country
4479  Skins  TV Show          2013  United Kingdom
4908  Skins    Movie          2017           Spain
---------------------------------------
           title   type  release_year        country
3712  Blood Money  Movie          2017  United States
4949  Blood Money  Movie          2012          India
---------------------------------------
     title   type  release_year           country
2692  Tiger  Movie          2016             India
5055  Tiger  Movie          2016  Argentina, Italy
---------------------------------------
                  title   type  release_year              country
3687  People You May Know  Movie          2017         United States
5061  People You May Know  Movie          2016  United States, Spain
---------------------------------------
          title   type  release_year                   country
161   The Lovers  Movie          2017             United States
5122  The Lovers  Movie          2015  Belgium, India, Australia
---------------------------------------
        title     type  release_year         country
4629  Aquarius  TV Show          2016   United States
```

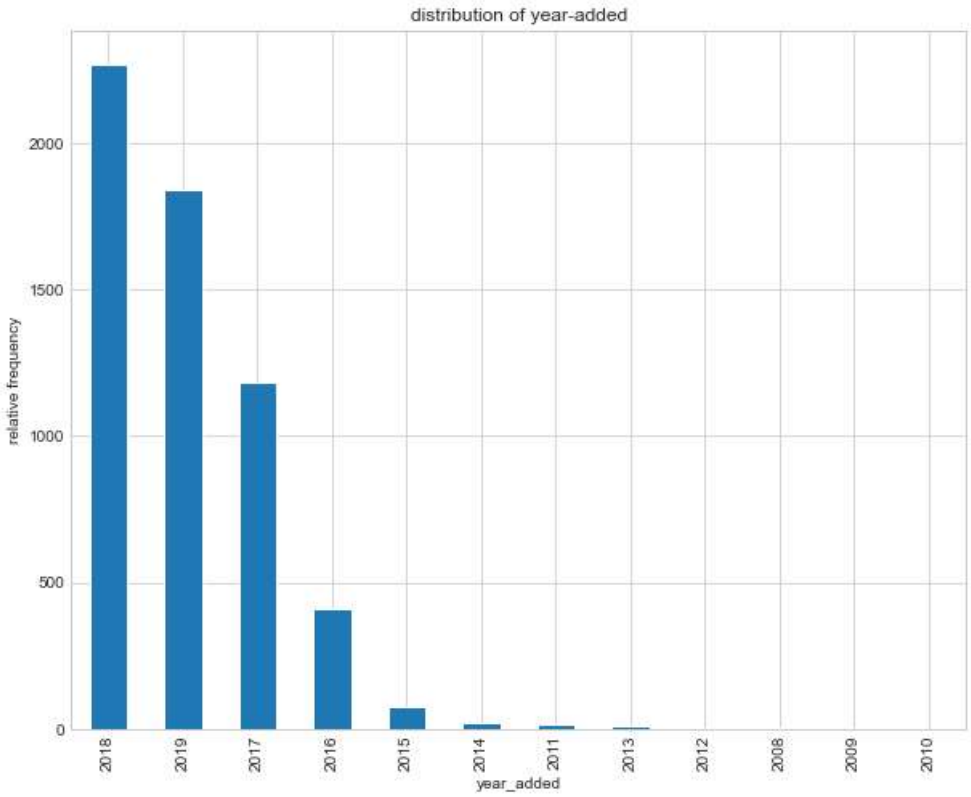
```
5182  Aquarius    Movie          2016  Brazil, France
----------------------------------------
                                    title     type  release_year country
357   Little Baby Bum: Nursery Rhyme Friends  TV Show          2019     NaN
5337  Little Baby Bum: Nursery Rhyme Friends    Movie          2016     NaN
----------------------------------------
     title     type  release_year  \
4059  Deep    Movie          2017
5531  Deep  TV Show          2016

                                          country
4059  Spain, Belgium, Switzerland, United States, Ch...
5531                                         France
----------------------------------------
         title     type  release_year        country
474   Limitless    Movie          2017          India
1286  Limitless    Movie          2011  United States
5536  Limitless  TV Show          2016  United States
----------------------------------------
               title   type  release_year        country
5594  Frank and Cindy  Movie          2007  United States
5595  Frank and Cindy  Movie          2015  United States
----------------------------------------
           title     type  release_year         country
3779  Retribution  TV Show          2016  United Kingdom
5619  Retribution    Movie          2015           Spain
----------------------------------------
                      title     type  release_year        country
1778  Wet Hot American Summer    Movie          2001  United States
5732  Wet Hot American Summer  TV Show          2015  United States
----------------------------------------
     title     type  release_year  \
2777  Life  TV Show          2018
5755  Life  TV Show          2009

                                          country
2777                                  South Korea
5755  United Kingdom, United States, Greece, Italy, ...
----------------------------------------
```

In [67]:

```
plt.figure(figsize=(10,8))
df['year_added'].value_counts().plot.bar()
plt.title('distribution of year-added')
plt.ylabel('relative frequency')
plt.xlabel('year_added')
plt.show()
```



2018 is considered to be the most remarkable year for

netflix

In [68]:
```
counts=0
for i,j in zip(df['release_year'].values,df['year_added'].values):
    if i!=j:
        counts+=1
print('number of contents that its release year differ from the year added to netflix are ',str(counts))
```

```
number of contents that its release year differ from the year added to netflix are  3971
```

In [ ]:

In [ ]: