In [2]:
```
#Khai báo thư viện
import numpy as np
import pandas as pd
```

In [3]:
```
#Đọc tập dữ liệu credits của các phim từ file csv
credits = pd.read_csv('credits.csv')
```

In [4]:
```
#Print
#Tập dữ liệu 45476 dòng và 3 cột
credits
```

Out[4]:

	cast	crew	id
0	[{'cast_id': 14, 'character': 'Woody (voice)',...	[{'credit_id': '52fe4284c3a36847f8024f49', 'de...	862
1	[{'cast_id': 1, 'character': 'Alan Parrish', '...	[{'credit_id': '52fe44bfc3a36847f80a7cd1', 'de...	8844
2	[{'cast_id': 2, 'character': 'Max Goldman', 'c...	[{'credit_id': '52fe466a9251416c75077a89', 'de...	15602
3	[{'cast_id': 1, 'character': "Savannah 'Vannah...	[{'credit_id': '52fe44779251416c91011acb', 'de...	31357
4	[{'cast_id': 1, 'character': 'George Banks', '...	[{'credit_id': '52fe44959251416c75039ed7', 'de...	11862
...	...	...	...
45471	[{'cast_id': 0, 'character': '', 'credit_id': ...	[{'credit_id': '5894a97d925141426c00818c', 'de...	439050
45472	[{'cast_id': 1002, 'character': 'Sister Angela...	[{'credit_id': '52fe4af1c3a36847f81e9b15', 'de...	111109
45473	[{'cast_id': 6, 'character': 'Emily Shaw', 'cr...	[{'credit_id': '52fe4776c3a368484e0c8387', 'de...	67758
45474	[{'cast_id': 2, 'character': '', 'credit_id': ...	[{'credit_id': '533bccebc3a36844cf0011a7', 'de...	227506
45475	[]	[{'credit_id': '593e676c92514105b702e68e', 'de...	461257

45476 rows × 3 columns

In [5]:
```
#Đọc tập dữ liệu movies_metadata là một tập siêu dữ liệu bao gồm tất cả về 1 bộ p
meta = pd.read_csv('movies_metadata.csv')
```

```
C:\Users\vohan\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:316
5: DtypeWarning: Columns (10) have mixed types.Specify dtype option on import o
r set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

In [6]:
```
meta['release_date'] = pd.to_datetime(meta['release_date'], errors='coerce')
```

In [7]:
```
meta['year'] = meta['release_date'].dt.year
```

In [8]:
```
#Lọc sắp xếp dữ liệu đếm tổng số phim qua các năm
meta['year'].value_counts().sort_index()
```

Out[8]:
```
1874.0       1
1878.0       1
1883.0       1
1887.0       1
1888.0       2
           ...
2015.0    1905
2016.0    1604
2017.0     532
2018.0       5
2020.0       1
Name: year, Length: 135, dtype: int64
```

In [11]:
```
# Do tập dữ liệu có các phim ở năm từ 2018-2020 chỉ có 6 phim nên em chỉ lấy từ 2
new_meta = meta.loc[meta.year <= 2017,['genres','id','title','year']]
```

In [12]:
```
new_meta
```

Out[12]:

	genres	id	title	year
0	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	862	Toy Story	1995.0
1	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	8844	Jumanji	1995.0
2	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	15602	Grumpier Old Men	1995.0
3	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	31357	Waiting to Exhale	1995.0
4	[{'id': 35, 'name': 'Comedy'}]	11862	Father of the Bride Part II	1995.0
...	...	...	...	...
45460	[{'id': 18, 'name': 'Drama'}, {'id': 28, 'name...	30840	Robin Hood	1991.0
45462	[{'id': 18, 'name': 'Drama'}]	111109	Century of Birthing	2011.0
45463	[{'id': 28, 'name': 'Action'}, {'id': 18, 'nam...	67758	Betrayal	2003.0
45464	[]	227506	Satan Triumphant	1917.0
45465	[]	461257	Queerama	2017.0

45370 rows × 4 columns

In [14]:
```
#Gán kiểu dữ liệu cho id là kiểu int
new_meta['id'] = new_meta['id'].astype(int)
```

In [15]:
```
#Merge 2 tập dữ liệu credits và meta_data
data = pd.merge(new_meta, credits, on='id')
```

In [16]:
```
pd.set_option('display.max_colwidth', 75)
data
```

Out[16]:

	genres	id	title	year	cast	
0	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': ...	862	Toy Story	1995.0	[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3...	['52fe4284c3a36847 'department': 'Di
1	[{'id': 12, 'name': 'Adventure'}, {'id': 14, 'name': 'Fantasy'}, {'id':...	8844	Jumanji	1995.0	[{'cast_id': 1, 'character': 'Alan Parrish', 'credit_id': '52fe44bfc3a3...	['52fe44bfc3a36847 'department': 'Pro
2	[{'id': 10749, 'name': 'Romance'}, {'id': 35, 'name': 'Comedy'}]	15602	Grumpier Old Men	1995.0	[{'cast_id': 2, 'character': 'Max Goldman', 'credit_id': '52fe466a92514...	['52fe466a9251416c7 'department': 'Di
3	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 1074...	31357	Waiting to Exhale	1995.0	[{'cast_id': 1, 'character': "Savannah 'Vannah' Jackson", 'credit_id': ...	['52fe44779251416c5 'department': 'Di
4	[{'id': 35, 'name': 'Comedy'}]	11862	Father of the Bride Part II	1995.0	[{'cast_id': 1, 'character': 'George Banks', 'credit_id': '52fe44959251...	['52fe44959251416c7 'department': 'Soun
...	...	...	...	...	...	
45440	[{'id': 18, 'name': 'Drama'}, {'id': 28, 'name': 'Action'}, {'id': 1074...	30840	Robin Hood	1991.0	[{'cast_id': 1, 'character': 'Sir Robert Hode', 'credit_id': '52fe44439...	['52fe44439251416c9 'department': 'Di
45441	[{'id': 18, 'name': 'Drama'}]	111109	Century of Birthing	2011.0	[{'cast_id': 1002, 'character': 'Sister Angela', 'credit_id': '52fe4af1...	['52fe4af1c3a36847 'department': 'Di
45442	[{'id': 28, 'name': 'Action'}, {'id': 18, 'name': 'Drama'}, {'id': 53, ...	67758	Betrayal	2003.0	[{'cast_id': 6, 'character': 'Emily Shaw', 'credit_id': '52fe4776c3a368...	['52fe4776c3a368484 'department': 'Di
45443	[]	227506	Satan Triumphant	1917.0	[{'cast_id': 2, 'character': '', 'credit_id': '52fe4ea59251416c7515d7d5...	['533bccebc3a36844 'department': 'Di

	genres	id	title	year	cast	
45444	[]	461257	Queerama	2017.0	[]	[{'593e676c92514105b 'department': 'Di

45445 rows × 6 columns

In [17]:
```
# Đánh dấu từng thuộc tính trong csdl và tách thể loại ra thành danh sách thể loạ
import ast
data['genres'] = data['genres'].map(lambda x: ast.literal_eval(x))
data['cast'] = data['cast'].map(lambda x: ast.literal_eval(x))
data['crew'] = data['crew'].map(lambda x: ast.literal_eval(x))
```

In [18]:
```
def make_genresList(x):
    gen = []
    st = " "
    for i in x:
        if i.get('name') == 'Science Fiction':
            scifi = 'Sci-Fi'
            gen.append(scifi)
        else:
            gen.append(i.get('name'))
    if gen == []:
        return np.NaN
    else:
        return (st.join(gen))
```

In [19]:
```
data['genres_list'] = data['genres'].map(lambda x: make_genresList(x))
```

In [20]:
```
data['genres_list']
```

Out[20]:
```
0         Animation Comedy Family
1        Adventure Fantasy Family
2                 Romance Comedy
3            Comedy Drama Romance
4                         Comedy
                   ...
45440        Drama Action Romance
45441                       Drama
45442       Action Drama Thriller
45443                         NaN
45444                         NaN
Name: genres_list, Length: 45445, dtype: object
```

In [21]:
```
#Tách tên diễn viên từ data và lưu từng diễn viên theo id
def get_actor1(x):
    casts = []
    for i in x:
        casts.append(i.get('name'))
    if casts == []:
        return np.NaN
    else:
        return (casts[0])
```

In [22]:
```
data['actor_1_name'] = data['cast'].map(lambda x: get_actor1(x))
```

In [23]:
```
def get_actor2(x):
    casts = []
    for i in x:
        casts.append(i.get('name'))
    if casts == [] or len(casts)<=1:
        return np.NaN
    else:
        return (casts[1])
```

In [24]:
```
data['actor_2_name'] = data['cast'].map(lambda x: get_actor2(x))
```

In [27]:
```
data['actor_2_name']
```

Out[27]:
```
0               Tim Allen
1            Jonathan Hyde
2              Jack Lemmon
3           Angela Bassett
4             Diane Keaton
                  ...
45440          Uma Thurman
45441          Perry Dizon
45442         Adam Baldwin
45443    Nathalie Lissenko
45444                  NaN
Name: actor_2_name, Length: 45445, dtype: object
```

In [28]:
```
def get_actor3(x):
    casts = []
    for i in x:
        casts.append(i.get('name'))
    if casts == [] or len(casts)<=2:
        return np.NaN
    else:
        return (casts[2])
```

In [29]:
```
data['actor_3_name'] = data['cast'].map(lambda x: get_actor3(x))
```

In [31]:
```
data['actor_3_name']
```

Out[31]:
```
0            Don Rickles
1          Kirsten Dunst
2            Ann-Margret
3         Loretta Devine
4           Martin Short
               ...
45440    David Morrissey
45441      Hazel Orencio
45442      Julie du Page
45443       Pavel Pavlov
45444                NaN
Name: actor_3_name, Length: 45445, dtype: object
```

In [32]:
```
#Tách tên đạo diễn
def get_directors(x):
    dt = []
    st = " "
    for i in x:
        if i.get('job') == 'Director':
            dt.append(i.get('name'))
    if dt == []:
        return np.NaN
    else:
        return (st.join(dt))
```

In [33]:
```
data['director_name'] = data['crew'].map(lambda x: get_directors(x))
```

In [38]:
```
data['director_name']
```

Out[38]:
```
0           John Lasseter
1            Joe Johnston
2           Howard Deutch
3         Forest Whitaker
4           Charles Shyer
                ...
45440          John Irvin
45441            Lav Diaz
45442      Mark L. Lester
45443    Yakov Protazanov
45444       Daisy Asquith
Name: director_name, Length: 45445, dtype: object
```

In [37]:
```
movie = data.loc[:,['director_name','actor_1_name','actor_2_name','actor_3_name',
```

In [39]:
```
#Tập dữ liệu gồm tên đạo diễn, diễn viên 1,2,3 thể loại, tiêu đề (tên phim)
movie
```

Out[39]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres_list	title
0	John Lasseter	Tom Hanks	Tim Allen	Don Rickles	Animation Comedy Family	Toy Story
1	Joe Johnston	Robin Williams	Jonathan Hyde	Kirsten Dunst	Adventure Fantasy Family	Jumanji
2	Howard Deutch	Walter Matthau	Jack Lemmon	Ann-Margret	Romance Comedy	Grumpier Old Men
3	Forest Whitaker	Whitney Houston	Angela Bassett	Loretta Devine	Comedy Drama Romance	Waiting to Exhale
4	Charles Shyer	Steve Martin	Diane Keaton	Martin Short	Comedy	Father of the Bride Part II
...	...	...	...	...	...	...
45440	John Irvin	Patrick Bergin	Uma Thurman	David Morrissey	Drama Action Romance	Robin Hood
45441	Lav Diaz	Angel Aquino	Perry Dizon	Hazel Orencio	Drama	Century of Birthing
45442	Mark L. Lester	Erika Eleniak	Adam Baldwin	Julie du Page	Action Drama Thriller	Betrayal
45443	Yakov Protazanov	Iwan Mosschuchin	Nathalie Lissenko	Pavel Pavlov	NaN	Satan Triumphant
45444	Daisy Asquith	NaN	NaN	NaN	NaN	Queerama

45445 rows × 6 columns

In [40]:
```
#Thống kê tổng đạo diễn, diễn viên 1,2,3 thể loại, tiêu đề (tên phim)
movie.isna().sum()
```

Out[40]:
```
director_name     835
actor_1_name     2354
actor_2_name     3683
actor_3_name     4593
genres_list      2384
title               0
dtype: int64
```

In [41]:
```
#Loại bỏ các giá trị n.a
movie = movie.dropna(how='any')
```

In [42]:
```
movie.isna().sum()
```

Out[42]:
```
director_name    0
actor_1_name     0
actor_2_name     0
actor_3_name     0
genres_list      0
title            0
dtype: int64
```

In [43]:
```
movie = movie.rename(columns={'genres_list':'genres'})
movie = movie.rename(columns={'title':'movie_title'})
```

In [44]:
```
movie['movie_title'] = movie['movie_title'].str.lower()
```

In [45]:
```
movie['comb'] = movie['actor_1_name'] + ' ' + movie['actor_2_name'] + ' '+ movie[
```

In [121]:
```
movie
```

Out[121]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	c
0	John Lasseter	Tom Hanks	Tim Allen	Don Rickles	Animation Comedy Family	toy story	H Tim A Ric Lass Anima Com Fa
1	Joe Johnston	Robin Williams	Jonathan Hyde	Kirsten Dunst	Adventure Fantasy Family	jumanji	R Willi Jona H Kir Dunst John Adver Fa

In [122]:
```
#Xóa các phim trùng lặp
movie.drop_duplicates(subset ="movie_title", keep = 'last', inplace = True)
```

In [123]:
```
movie
```

Out[123]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	comb
0	John Lasseter	Tom Hanks	Tim Allen	Don Rickles	Animation Comedy Family	toy story	Tom Hanks Tim Aller Dor Rickles John Lassete Animatior Comedy Family
1	Joe Johnston	Robin Williams	Jonathan Hyde	Kirsten Dunst	Adventure Fantasy Family	jumanji	Robir Williams Jonathar Hyde Kirster Dunst Joe Johnstor Adventure Fanta..
2	Howard Deutch	Walter Matthau	Jack Lemmon	Ann-Margret	Romance Comedy	grumpier old men	Walte Matthau Jack Lemmor Ann Margre Howarc Deutch Romance Comedy
3	Forest Whitaker	Whitney Houston	Angela Bassett	Loretta Devine	Comedy Drama Romance	waiting to exhale	Whitney Houstor Angela Basset Loretta Devine Fores Whitake Comedy Dr..
4	Charles Shyer	Steve Martin	Diane Keaton	Martin Short	Comedy	father of the bride part ii	Steve Martir Diane Keator Martir Shor Charles Shye Comedy
...	...	...	...	...	...	...	..
45438	Ben Rock	Monty Bane	Lucy Butler	David Grammer	Horror	the burkittsville 7	Monty Bane Lucy Butle Davic Gramme Ben Rock Horro

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	comb
45439	Aaron Osborne	Lisa Boyle	Kena Land	Zaneta Polard	Sci-Fi	caged heat 3000	Lisa Boyle Kena Land Zaneta Polard Aaron Osborne Sci-F
45440	John Irvin	Patrick Bergin	Uma Thurman	David Morrissey	Drama Action Romance	robin hood	Patrick Bergin Uma Thurman David Morrissey John Irvin Drama Action Romance
45441	Lav Diaz	Angel Aquino	Perry Dizon	Hazel Orencio	Drama	century of birthing	Ange Aquino Perry Dizon Haze Orencio Lav Diaz Drama
45442	Mark L. Lester	Erika Eleniak	Adam Baldwin	Julie du Page	Action Drama Thriller	betrayal	Erika Eleniak Adam Baldwin Julie du Page Mark L Leste Action Drama Th..

36341 rows × 7 columns

In [46]:
```
#Xuất file movie
movie.to_csv('movie.csv',index=False)
```

In [ ]:
```
```