

```
In [1]: import pandas as pd
import numpy as np
import requests
import bs4 as bs
import urllib.request
```

Movie 2022

```
In [2]: link = "https://en.wikipedia.org/wiki/List_of_American_films_of_2022"
```

```
In [3]: source = urllib.request.urlopen(link).read()
soup = bs.BeautifulSoup(source, 'lxml')
```

```
In [4]: tables = soup.find_all('table', class_='wikitable sortable')
```

```
In [5]: df1 = pd.read_html(str(tables[0]))[0]
df2 = pd.read_html(str(tables[1]))[0]
df3 = pd.read_html(str(tables[2]))[0]
df4 = pd.read_html(str(tables[3]).replace("'1\"\\'", "'1'"))[0]
```

```
In [6]: df = df1.append(df2.append(df3.append(df4, ignore_index=True), ignore_index=True), i
```

In [7]: df

Out[7]:

	Opening	Opening.1	Title	Production company	Cast and crew	.mw-parser-output .tooltip-dotted{border-bottom:1px dotted;cursor:help}Ref.
0	JANUARY	7	The 355	Universal Pictures / Perfect World Pictures / ...	Simon Kinberg (director/screenplay); Theresa R...	[1]
1	JANUARY	14	Scream	Paramount Pictures / Spyglass Media Group	Matt Bettinelli-Olpin, Tyler Gillett (director...	[2]
2	JANUARY	14	Deep Water	20th Century Studios / Regency Enterprises	Adrian Lyne (director); Zach Helm, Sam Levinso...	[3]
3	JANUARY	14	The Man from Toronto	Columbia Pictures / Bron Creative / Escape Art...	Patrick Hughes (director); Robbie Fox (screenp...	[4]
4	JANUARY	21	Five Eyes	STXfilms / Miramax	Guy Ritchie (director); Ivan Atkinson, Marn Da...	[5]
...
62	DECEMBER	16	Avatar 2	20th Century Studios / Lightstorm Entertainment	James Cameron (director/screenplay); Josh Frie...	NaN
63	DECEMBER	16	Aquaman and the Lost Kingdom	Warner Bros. Pictures / DC Films / Atomic Mons...	James Wan (director), David Leslie Johnson-McG...	NaN
64	DECEMBER	23	I Wanna Dance with Somebody	TriStar Pictures	Kasi Lemmons (director); Anthony McCarten (scr...	NaN
65	DECEMBER	23	The Nightingale	TriStar Pictures	Mélanie Laurent (director), Dana Stevens (scre...	NaN
66	DECEMBER	25	Babylon	Paramount Pictures	Damien Chazelle (director/screenplay); Brad Pi...	NaN

67 rows × 7 columns



```
In [54]: df_2022 = df[['Title', 'Cast and crew']]
```

```
In [65]: df_2022=pd.read_csv('check2.csv')
df_2022
```

Out[65]:

	Title	Cast and crew
0	The 355	Simon Kinberg (director/screenplay); Theresa R...
1	Scream	Matt Bettinelli-Olpin, Tyler Gillett (director...
2	Deep Water	Adrian Lyne (director); Zach Helm, Sam Levinso...
3	The Man from Toronto	Patrick Hughes (director); Robbie Fox (screenp...
4	Five Eyes	Guy Ritchie (director); Ivan Atkinson, Marn Da...
...
61	Avatar 2	James Cameron (director/screenplay); Josh Frie...
62	Aquaman and the Lost Kingdom	James Wan (director), David Leslie Johnson-McG...
63	I Wanna Dance with Somebody	Kasi Lemmons (director); Anthony McCarten (scr...
64	The Nightingale	Mélanie Laurent (director), Dana Stevens (scre...
65	Babylon	Damien Chazelle (director/screenplay); Brad Pi...

66 rows × 2 columns

```
In [66]: from tmdbv3api import TMDb
import json
import requests
tmdb = TMDb()
tmdb.api_key = 'b7e7aeea608257af93a032205bfada32'
```

```
In [67]: from tmdbv3api import Movie
tmdb_movie = Movie()
def get_genre(x):
    genres = []
    result = tmdb_movie.search(x)
    movie_id = result[0].id
    response = requests.get('https://api.themoviedb.org/3/movie/{}?api_key={}'.format(movie_id, tmdb.api_key))
    data_json = response.json()
    if data_json['genres']:
        genre_str = ""
        for i in range(0, len(data_json['genres'])):
            genres.append(data_json['genres'][i]['name'])
        return genre_str.join(genres)
    else:
        return np.NaN
```

```
In [68]: df_2022['genres'] = df_2022['Title'].map(lambda x: get_genre(str(x)))
```

In [69]: df_2022

Out[69]:

	Title	Cast and crew	genres
0	The 355	Simon Kinberg (director/screenplay); Theresa R...	Action Thriller
1	Scream	Matt Bettinelli-Olpin, Tyler Gillett (director...	Crime Horror Mystery
2	Deep Water	Adrian Lyne (director); Zach Helm, Sam Levinso...	Thriller Romance Drama
3	The Man from Toronto	Patrick Hughes (director); Robbie Fox (screenp...	Romance Comedy
4	Five Eyes	Guy Ritchie (director); Ivan Atkinson, Marn Da...	Action Thriller
...
61	Avatar 2	James Cameron (director/screenplay); Josh Frie...	Action Adventure Science Fiction Fantasy
62	Aquaman and the Lost Kingdom	James Wan (director), David Leslie Johnson-McG...	Adventure Action Fantasy
63	I Wanna Dance with Somebody	Kasi Lemmons (director); Anthony McCarten (scr...	Drama
64	The Nightingale	Mélanie Laurent (director), Dana Stevens (scre...	Drama Thriller
65	Babylon	Damien Chazelle (director/screenplay); Brad Pi...	Action Adventure Science Fiction Thriller

66 rows × 3 columns

```
In [70]: def get_director(x):
         if "(director)" in x:
             return x.split(" (director)") [0]
         elif "(directors)" in x:
             return x.split(" (directors)") [0]
         else:
             return x.split(" (director/screenplay)") [0]
```

```
In [71]: df_2022['director_name'] = df_2022['Cast and crew'].map(lambda x: get_director(st
```

```
In [72]: def get_actor1(x):
         return ((x.split("screenplay); ")[-1]).split(", ")[0])
```

```
In [73]: df_2022['actor_1_name'] = df_2022['Cast and crew'].map(lambda x: get_actor1(str(>
```

```
In [74]: def get_actor2(x):  
         if len((x.split("screenplay); ")[-1]).split(", ")) < 2:  
             return np.NaN  
         else:  
             return ((x.split("screenplay); ")[-1]).split(", ")[1])
```

```
In [75]: df_2022['actor_2_name'] = df_2022['Cast and crew'].map(lambda x: get_actor2(str(x)))
```

```
In [76]: def get_actor3(x):  
         if len((x.split("screenplay); ")[-1]).split(", ")) < 3:  
             return np.NaN  
         else:  
             return ((x.split("screenplay); ")[-1]).split(", ")[2])
```

```
In [77]: df_2022['actor_3_name'] = df_2022['Cast and crew'].map(lambda x: get_actor3(str(x)))
```

In [78]: df_2022

Out[78]:

	Title	Cast and crew	genres	director_name	actor_1_name	actor_2_name	actor_3
0	The 355	Simon Kinberg (director/screenplay); Theresa R...	Action Thriller	Simon Kinberg	Jessica Chastain	Lupita Nyong'o	Penélop
1	Scream	Matt Bettinelli-Olpin, Tyler Gillett (director...	Crime Horror Mystery	Matt Bettinelli-Olpin, Tyler Gillett	Neve Campbell	David Arquette	Courten
2	Deep Water	Adrian Lyne (director); Zach Helm, Sam Levinso...	Thriller Romance Drama	Adrian Lyne	Ben Affleck	Ana de Armas	Trac
3	The Man from Toronto	Patrick Hughes (director); Robbie Fox (screenp...	Romance Comedy	Patrick Hughes	Kevin Hart	Woody Harrelson	Kaley
4	Five Eyes	Guy Ritchie (director); Ivan Atkinson, Marn Da...	Action Thriller	Guy Ritchie	Jason Statham	Aubrey Plaza	Bugzy
...
61	Avatar 2	James Cameron (director/screenplay); Josh Frie...	Action Adventure Science Fiction Fantasy	James Cameron	Sam Worthington	Zoe Saldana	Stephe
62	Aquaman and the Lost Kingdom	James Wan (director), David Leslie Johnson-McG...	Adventure Action Fantasy	James Wan	Jason Momoa	Amber Heard	Patrick
63	I Wanna Dance with Somebody	Kasi Lemmons (director); Anthony McCarten (scr...	Drama	Kasi Lemmons	Naomi Ackie	Ashton Sanders	
64	The Nightingale	Mélanie Laurent (director), Dana Stevens (scre...	Drama Thriller	Mélanie Laurent	Dakota Fanning	Elle Fanning	
65	Babylon	Damien Chazelle (director/screenplay); Brad Pi...	Action Adventure Science Fiction Thriller	Damien Chazelle	Brad Pitt	Margot Robbie	Ka Wa

66 rows × 7 columns



In [79]: df_2022 = df_2022.rename(columns={'Title':'movie_title'})

In [80]: new_df22 = df_2022.loc[:,['director_name','actor_1_name','actor_2_name','actor_3_

In [81]: new_df22

Out[81]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title
0	Simon Kinberg	Jessica Chastain	Lupita Nyong'o	Penélope Cruz	Action Thriller	The 355
1	Matt Bettinelli-Olpin, Tyler Gillett	Neve Campbell	David Arquette	Courteney Cox	Crime Horror Mystery	Scream
2	Adrian Lyne	Ben Affleck	Ana de Armas	Tracy Letts	Thriller Romance Drama	Deep Water
3	Patrick Hughes	Kevin Hart	Woody Harrelson	Kaley Cuoco	Romance Comedy	The Man from Toronto
4	Guy Ritchie	Jason Statham	Aubrey Plaza	Bugzy Malone	Action Thriller	Five Eyes
...
61	James Cameron	Sam Worthington	Zoe Saldana	Stephen Lang	Action Adventure Science Fiction Fantasy	Avatar 2
62	James Wan	Jason Momoa	Amber Heard	Patrick Wilson	Adventure Action Fantasy	Aquaman and the Lost Kingdom
63	Kasi Lemmons	Naomi Ackie	Ashton Sanders	NaN	Drama	I Wanna Dance with Somebody
64	Mélanie Laurent	Dakota Fanning	Elle Fanning	NaN	Drama Thriller	The Nightingale
65	Damien Chazelle	Brad Pitt	Margot Robbie	Katherine Waterston	Action Adventure Science Fiction Thriller	Babylon

66 rows × 6 columns

In [82]: ' + new_df22['actor_2_name'] + ' ' + new_df22['actor_3_name'] + ' ' + new_df22['di

In [83]: new_df22 = new_df22.dropna(how='any')

In [84]: new_df22['movie_title'] = new_df22['movie_title'].str.lower()

<ipython-input-84-bd701421fe8d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

new_df22['movie_title'] = new_df22['movie_title'].str.lower()

In [85]: new_df22

26	Pierre Perifel	Sam Rockwell	Awkwafina	Anthony Ramos	Animation	the bad guys
27	Tom Gormican	Nicolas Cage	Pedro Pascal	Sharon Horgan	Comedy	the unbearable weight of massive talent
28	Taika Waititi	Chris Hemsworth	Natalie Portman	Tessa Thompson	Action Adventure Fantasy Comedy Science Fiction	thor: love and thunder
29	Scott Beck, Bryan Woods	Adam Driver	Ariana Greenblatt	Chloe Coleman	Drama	65
30	Jared Stern	Sam Levine (co-director): Dwavne	Kevin Hart	...	Kate Animation Adventure	dc league of

In [88]: old_df = pd.read_csv('main_data.csv')

In [90]: old_df

Out[90]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	
0	John Lasseter	Tom Hanks	Tim Allen	Don Rickles	Animation Comedy Family	toy story	Tom Tim Alle Rickles Lass
1	Joe Johnston	Robin Williams	Jonathan Hyde	Kirsten Dunst	Adventure Fantasy Family	jumanji	W Jor Hyde k Dunst
2	Howard Deutch	Walter Matthau	Jack Lemmon	Ann-Margret	Romance Comedy	grumpier old men	Matthau Lemmon M How
3	Forest Whitaker	Whitney Houston	Angela Bassett	Loretta Devine	Comedy Drama Romance	waiting to exhale	W Hc / B L Dev
4	Charles Shyer	Steve Martin	Diane Keaton	Martin Short	Comedy	father of the bride part ii	Steve Diane k Martin Ch
...	
40203	Lana Wachowski	Keanu Reeves	Carrie-Anne Moss	Yahya Abdul- Mateen II	Action Adventure Thriller Science Fiction	the matrix resurrections	I R Carrie Moss Abdul-M
40204	Garth Jennings	Matthew McConaughey	Reese Witherspoon	Scarlett Johansson	Family Comedy Animation Drama Music	sing 2	Ma McCona I Wither Sc
40205	Matthew Vaughn	Ralph Fiennes	Gemma Arterton	Rhys Ifans	War Action Thriller Adventure	the king's man	Ralph Fi G Arterton Ifans Ma
40206	Erwin brothers	Zachary Levi	Anna Paquin	Ser'Darius Blain	Drama	american underdog: the kurt warner story	Zachar Anna F Ser'l Blain
40207	Joe Wright	Peter Dinklage	Haley Bennett	Kelvin Harrison Jr.	Drama Comedy History Romance	cyrano	Dii Haley B Harris

40208 rows × 7 columns



```
In [91]: final_df = old_df.append(new_df22, ignore_index=True)
```

In [92]: final_df

Out[92]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	con
0	John Lasseter	Tom Hanks	Tim Allen	Don Rickles	Animation Comedy Family	toy story	Tom Han Tim All Don Rickl Jo Lasseter
1	Joe Johnston	Robin Williams	Jonathan Hyde	Kirsten Dunst	Adventure Fantasy Family	jumanji	Rot Williar Jonath Hy Kirst Dunst Joe
2	Howard Deutch	Walter Matthau	Jack Lemmon	Ann-Margret	Romance Comedy	grumpier old men	Wal Matth Ja Lemm Ar Marg Howard
3	Forest Whitaker	Whitney Houston	Angela Bassett	Loretta Devine	Comedy Drama Romance	waiting to exhale	Whitn Houst Anger Bass Lore Devine
4	Charles Shyer	Steve Martin	Diane Keaton	Martin Short	Comedy	father of the bride part ii	Ste Mar Dia Keat Mar Sh Charles
...
40257	Maria Schrader	Carey Mulligan	Zoe Kazan	Patricia Clarkson	Comedy Crime Drama Mystery	she said	Car Mullig Zoe Kaz Patric Clarks Ma
40258	Michael B. Jordan	Michael B. Jordan	Tessa Thompson	Phylicia Rashad	Drama Action	creed iii	Michael Jord Tes Thomps Phylic Rash
40259	James Cameron	Sam Worthington	Zoe Saldana	Stephen Lang	Action Adventure Science Fiction Fantasy	avatar 2	Se Worthingt Z Salda Steph La James

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	con
40260	James Wan	Jason Momoa	Amber Heard	Patrick Wilson	Adventure Action Fantasy	aquaman and the lost kingdom	Jas Mom Amt Hea Patri Wils James V
40261	Damien Chazelle	Brad Pitt	Margot Robbie	Katherine Waterston	Action Adventure Science Fiction Thriller	babylon	Brad F Març Robt Katheri Waterst Dæ

40262 rows × 7 columns



```
In [93]: final_df.to_csv('main_data.csv', index=False)
```

```
In [ ]:
```