

# f-imdb-top-movies-from-each-genre

November 16, 2024

## 1 CREATE JSON FILES OF IMDB TOP RATED MOVIES FROM EACH GENRE

The primary source for an IMDb all movies dataset based on genre would be IMDb, the world's most popular and authoritative source for movie, TV, and celebrity content. IMDb has an extensive database of movies that is constantly updated with new titles and information. The IMDb all movies dataset based on genre provides a wealth of insights and opportunities for analysis.

The aim of the project is to clean all the datasets and create JSON files of top rated movies from each genre from IMDb movies datasets. We have datasets of following genres: \* Action \* Adventure \* Crime \* Family \* Fantasy \* Film-noir \* History \* Horror \* Mystery \* Sci-fi \* Sports \* Thriller \* War

### Each Dataset Consists of following columns

- movie\_id - IMDB Movie ID
- movie\_name - Name of the movie
- year - Release year
- certificate - Certificate of the movie
- run\_time - Total movie run time
- genre - Genre of the movie
- rating - Rating of the movie
- description - Description of the movie
- director - Director of the movie
- director\_id - IMDB id of the director
- star - Star of the movie
- star\_id - IMDB id of the star
- votes - Number of votes in IMDB website
- gross(in \$) - Gross Box Office of the movie

[ ]:

### 1.0.1 Importing Libraries

```
[241]: import os
import pandas as pd
import numpy as np
import re
```

```
import seaborn as sns
import json
sns.set_style('darkgrid')
```

## 1.0.2 Concatenating all the genre dataframes into one

```
[242]: files = ['data/'+i for i in os.listdir('data')]
main_df = []
for i in os.listdir('data'):
    data = pd.read_csv(f'data/{i}')
    main_genre = i.split('.')[0].capitalize()
    data['main_genre'] = main_genre
    main_df.append(data)

movies = pd.concat(main_df)
```

```
[243]: movies.head()
```

```
[243]:
```

	movie_id	movie_name	year	certificate	runtime	\
0	tt9114286	Black Panther: Wakanda Forever	2022	PG-13	161 min	
1	tt1630029	Avatar: The Way of Water	2022	PG-13	192 min	
2	tt5884796	Plane	2023	R	107 min	
3	tt6710474	Everything Everywhere All at Once	2022	R	139 min	
4	tt5433140	Fast X	2023	NaN	NaN	

	genre	rating	\
0	Action, Adventure, Drama	6.9	
1	Action, Adventure, Fantasy	7.8	
2	Action, Thriller	6.5	
3	Action, Adventure, Comedy	8.0	
4	Action, Crime, Mystery	NaN	

	description	\
0	The people of Wakanda fight to protect their h...	
1	Jake Sully lives with his newfound family form...	
2	A pilot finds himself caught in a war zone aft...	
3	A middle-aged Chinese immigrant is swept up in...	
4	Dom Toretto and his family are targeted by the...	

	director	director_id	\
0	Ryan Coogler	/name/nm3363032/	
1	James Cameron	/name/nm0000116/	
2	Jean-François Richet	/name/nm0724938/	
3	Dan Kwan, \nDaniel Scheinert	/name/nm3453283/	
4	Louis Leterrier	/name/nm0504642/	

```
star \
```

```

0 Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1 Sam Worthington, \nZoe Saldana, \nSigourney We...
2 Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3 Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4 Vin Diesel, \nJordana Brewster, \nTyrese Gibso...

```

```

                                star_id    votes  gross(in $)  \
0  /name/nm4004793/,/name/nm2143282/,/name/nm1775...  204835.0      NaN
1  /name/nm0941777/,/name/nm0757855/,/name/nm0000...  295119.0      NaN
2  /name/nm0124930/,/name/nm1591496/,/name/nm0001...   26220.0      NaN
3  /name/nm3215397/,/name/nm0000706/,/name/nm3513...  327858.0      NaN
4  /name/nm0004874/,/name/nm0108287/,/name/nm0879...    NaN      NaN

```

```

main_genre
0      Action
1      Action
2      Action
3      Action
4      Action

```

### 1.0.3 Duplicate Values

```
[244]: movies.shape
```

```
[244]: (298975, 15)
```

```
[245]: movies.duplicated(subset='movie_name').sum()
```

```
[245]: 124117
```

```
[246]: movies = movies.drop_duplicates(subset='movie_name')
```

```
[247]: movies.shape
```

```
[247]: (174858, 15)
```

```
[248]: movies.duplicated().sum()
```

```
[248]: 0
```

```
[249]: movies.shape
```

```
[249]: (174858, 15)
```

## 1.0.4 Handling Null Values

```
[250]: movies.isnull().sum()
```

```
[250]: movie_id          0
      movie_name        0
      year            32271
      certificate      135551
      runtime          60296
      genre            0
      rating           75371
      description       0
      director         14703
      director_id      14703
      star             31247
      star_id          27484
      votes            75369
      gross(in $)      166747
      main_genre        0
      dtype: int64
```

```
[251]: movies['certificate'].unique()
```

```
[251]: array(['PG-13', 'R', nan, 'Not Rated', 'TV-MA', 'PG', 'TV-14', '18+',
      'Approved', '13+', 'G', 'Unrated', 'M/PG', 'M', 'TV-PG', 'Passed',
      'TV-Y7', 'GP', 'TV-G', 'NC-17', '16+', 'X', 'TV-Y7-FV', 'TV-Y',
      '12', 'E', 'T', 'E10+', 'Open', 'TV-13', 'MA-13', 'AO', 'U', 'A',
      'MA-17', '18', 'UA', '13', 'UA 7+', 'UA 13+', '16', 'UA 16+'],
      dtype=object)
```

Certificates that are Null can be replaced by ‘Non Rated’

```
[252]: movies['certificate'].fillna('Not Rated', inplace=True)
```

```
[253]: movies['certificate'].isnull().sum()
```

```
[253]: 0
```

```
[254]: movies.head(1)
```

```
[254]:   movie_id          movie_name  year certificate  runtime \
0  tt9114286  Black Panther: Wakanda Forever  2022      PG-13  161 min

      genre  rating \
0  Action, Adventure, Drama    6.9

      description          director \
0  The people of Wakanda fight to protect their h...  Ryan Coogler
```

```

        director_id                                star \
0  /name/nm3363032/  Letitia Wright, \nLupita Nyong'o, \nDanai Guri...

                                star_id    votes  gross(in $) \
0  /name/nm4004793/,/name/nm2143282/,/name/nm1775...  204835.0      NaN

    main_genre
0      Action

```

```
[255]: movies['gross(in $)'].isnull().sum()
```

```
[255]: 166747
```

As majority of the data in gross column is null we will drop the gross columns and work on ratings

```
[256]: movies.drop('gross(in $)',axis = 1, inplace=True)
```

```
[257]: movies.head()
```

```
[257]:
    movie_id      movie_name  year certificate  runtime \
0  tt9114286  Black Panther: Wakanda Forever  2022      PG-13  161 min
1  tt1630029      Avatar: The Way of Water  2022      PG-13  192 min
2  tt5884796                      Plane  2023           R   107 min
3  tt6710474  Everything Everywhere All at Once  2022           R   139 min
4  tt5433140                      Fast X  2023  Not Rated      NaN

```

```

                                genre  rating \
0  Action, Adventure, Drama      6.9
1  Action, Adventure, Fantasy      7.8
2  Action, Thriller              6.5
3  Action, Adventure, Comedy      8.0
4  Action, Crime, Mystery        NaN

```

```

                                description \
0  The people of Wakanda fight to protect their h...
1  Jake Sully lives with his newfound family form...
2  A pilot finds himself caught in a war zone aft...
3  A middle-aged Chinese immigrant is swept up in...
4  Dom Toretto and his family are targeted by the...

```

```

                                director      director_id \
0  Ryan Coogler  /name/nm3363032/
1  James Cameron  /name/nm0000116/
2  Jean-François Richet  /name/nm0724938/
3  Dan Kwan, \nDaniel Scheinert  /name/nm3453283/

```

```
4 Louis Leterrier /name/nm0504642/
```

```
star \
0 Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1 Sam Worthington, \nZoe Saldana, \nSigourney We...
2 Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3 Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4 Vin Diesel, \nJordana Brewster, \nTyrese Gibso...
```

	star_id	votes	main_genre
0	/name/nm4004793/,/name/nm2143282/,/name/nm1775...	204835.0	Action
1	/name/nm0941777/,/name/nm0757855/,/name/nm0000...	295119.0	Action
2	/name/nm0124930/,/name/nm1591496/,/name/nm0001...	26220.0	Action
3	/name/nm3215397/,/name/nm0000706/,/name/nm3513...	327858.0	Action
4	/name/nm0004874/,/name/nm0108287/,/name/nm0879...	NaN	Action

```
[258]: movies.isnull().sum()
```

```
[258]: movie_id      0
movie_name      0
year           32271
certificate      0
runtime         60296
genre           0
rating          75371
description      0
director        14703
director_id     14703
star            31247
star_id         27484
votes           75369
main_genre      0
dtype: int64
```

### Removing all null values

```
[259]: movies.dropna(inplace=True)
```

```
[260]: movies.shape
```

```
[260]: (90617, 14)
```

```
[261]: movies.isnull().sum()
```

```
[261]: movie_id      0
movie_name      0
year           0
```

```

certificate    0
runtime        0
genre          0
rating         0
description    0
director       0
director_id    0
star           0
star_id        0
votes          0
main_genre     0
dtype: int64

```

```
[262]: movies = movies.reset_index(drop = True)
```

### 1.0.5 Formatting Feature Data

```
[263]: movies.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90617 entries, 0 to 90616
Data columns (total 14 columns):
#   Column          Non-Null Count  Dtype
---  -
0   movie_id        90617 non-null  object
1   movie_name      90617 non-null  object
2   year            90617 non-null  object
3   certificate      90617 non-null  object
4   runtime         90617 non-null  object
5   genre           90617 non-null  object
6   rating          90617 non-null  float64
7   description     90617 non-null  object
8   director        90617 non-null  object
9   director_id     90617 non-null  object
10  star            90617 non-null  object
11  star_id         90617 non-null  object
12  votes           90617 non-null  float64
13  main_genre      90617 non-null  object
dtypes: float64(2), object(12)
memory usage: 9.7+ MB

```

```
[264]: movies.describe()
```

```

[264]:
           rating      votes
count  90617.000000  9.061700e+04
mean         5.770457  7.901064e+03
std         1.341005  5.471597e+04

```

min	1.000000	5.000000e+00
25%	4.900000	3.700000e+01
50%	5.900000	1.560000e+02
75%	6.700000	8.620000e+02
max	10.000000	2.675531e+06

## 1.1 CLEANING THE DATA

Cleaning movie\_id column - to get only the numerical id value

```
[265]: movies.head(1)['movie_id']
```

```
[265]: 0    tt9114286
      Name: movie_id, dtype: object
```

```
[266]: def filter_movie_id(data):
      return int(data[2:])
```

```
[267]: movies['movie_id'] = movies['movie_id'].apply(filter_movie_id)
```

```
[268]: movies.head()
```

```
[268]:
```

	movie_id	movie_name	year	certificate	runtime \
0	9114286	Black Panther: Wakanda Forever	2022	PG-13	161 min
1	1630029	Avatar: The Way of Water	2022	PG-13	192 min
2	5884796	Plane	2023	R	107 min
3	6710474	Everything Everywhere All at Once	2022	R	139 min
4	10954600	Ant-Man and the Wasp: Quantumania	2023	PG-13	125 min

	genre	rating \
0	Action, Adventure, Drama	6.9
1	Action, Adventure, Fantasy	7.8
2	Action, Thriller	6.5
3	Action, Adventure, Comedy	8.0
4	Action, Adventure, Comedy	6.6

	description \
0	The people of Wakanda fight to protect their h...
1	Jake Sully lives with his newfound family form...
2	A pilot finds himself caught in a war zone aft...
3	A middle-aged Chinese immigrant is swept up in...
4	Scott Lang and Hope Van Dyne, along with Hank ...

	director	director_id \
0	Ryan Coogler	/name/nm3363032/
1	James Cameron	/name/nm0000116/
2	Jean-François Richet	/name/nm0724938/
3	Dan Kwan, \nDaniel Scheinert	/name/nm3453283/



```
4 Peyton Reed /name/nm0715636/
```

```
star \
0 Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1 Sam Worthington, \nZoe Saldana, \nSigourney We...
2 Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3 Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4 Paul Rudd, \nEvangeline Lilly, \nJonathan Majo...
```

```
star_id votes main_genre
0 /name/nm4004793/,/name/nm2143282/,/name/nm1775... 204835.0 Action
1 /name/nm0941777/,/name/nm0757855/,/name/nm0000... 295119.0 Action
2 /name/nm0124930/,/name/nm1591496/,/name/nm0001... 26220.0 Action
3 /name/nm3215397/,/name/nm0000706/,/name/nm3513... 327858.0 Action
4 /name/nm0748620/,/name/nm1431940/,/name/nm3718... 5396.0 Action
```

Cleaning runtime column - To get only runtime in integer value and remove 'min' suffix.

```
[269]: def filter_runtime(data):
        return int(data.split(' ')[0].replace(',', ''))
```

```
[270]: movies['runtime'] = movies['runtime'].apply(filter_runtime)
```

```
[271]: movies.head()
```

```
[271]: movie_id      movie_name  year certificate  runtime \
0    9114286  Black Panther: Wakanda Forever  2022      PG-13    161
1    1630029      Avatar: The Way of Water  2022      PG-13    192
2    5884796                Plane  2023          R     107
3    6710474  Everything Everywhere All at Once  2022          R    139
4   10954600  Ant-Man and the Wasp: Quantumania  2023      PG-13    125
```

```
genre rating \
0  Action, Adventure, Drama    6.9
1  Action, Adventure, Fantasy    7.8
2                Action, Thriller    6.5
3  Action, Adventure, Comedy    8.0
4  Action, Adventure, Comedy    6.6
```

```
description \
0 The people of Wakanda fight to protect their h...
1 Jake Sully lives with his newfound family form...
2 A pilot finds himself caught in a war zone aft...
3 A middle-aged Chinese immigrant is swept up in...
4 Scott Lang and Hope Van Dyne, along with Hank ...
```

	director	director_id \
0	Ryan Coogler	/name/nm3363032/
1	James Cameron	/name/nm0000116/
2	Jean-François Richet	/name/nm0724938/
3	Dan Kwan, \nDaniel Scheinert	/name/nm3453283/
4	Peyton Reed	/name/nm0715636/

	star \
0	Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1	Sam Worthington, \nZoe Saldana, \nSigourney We...
2	Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3	Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4	Paul Rudd, \nEvangeline Lilly, \nJonathan Majo...

	star_id	votes	main_genre
0	/name/nm4004793/,/name/nm2143282/,/name/nm1775...	204835.0	Action
1	/name/nm0941777/,/name/nm0757855/,/name/nm0000...	295119.0	Action
2	/name/nm0124930/,/name/nm1591496/,/name/nm0001...	26220.0	Action
3	/name/nm3215397/,/name/nm0000706/,/name/nm3513...	327858.0	Action
4	/name/nm0748620/,/name/nm1431940/,/name/nm3718...	5396.0	Action

Cleaning Genre column- get all the genres included in the movie in a list

```
[272]: def filter_genre(data):
        return data.split(',')

```

```
[273]: movies['genre'] = movies['genre'].apply(filter_genre)

```

```
[274]: movies.head()

```

	movie_id	movie_name	year	certificate	runtime \
0	9114286	Black Panther: Wakanda Forever	2022	PG-13	161
1	1630029	Avatar: The Way of Water	2022	PG-13	192
2	5884796	Plane	2023	R	107
3	6710474	Everything Everywhere All at Once	2022	R	139
4	10954600	Ant-Man and the Wasp: Quantumania	2023	PG-13	125

	genre	rating \
0	[Action, Adventure, Drama]	6.9
1	[Action, Adventure, Fantasy]	7.8
2	[Action, Thriller]	6.5
3	[Action, Adventure, Comedy]	8.0
4	[Action, Adventure, Comedy]	6.6

	description \
0	The people of Wakanda fight to protect their h...
1	Jake Sully lives with his newfound family form...

```

2 A pilot finds himself caught in a war zone aft...
3 A middle-aged Chinese immigrant is swept up in...
4 Scott Lang and Hope Van Dyne, along with Hank ...

```

```

              director      director_id \
0              Ryan Coogler /name/nm3363032/
1              James Cameron /name/nm0000116/
2              Jean-François Richet /name/nm0724938/
3 Dan Kwan, \nDaniel Scheinert /name/nm3453283/
4              Peyton Reed /name/nm0715636/

```

```

                                star \
0 Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1 Sam Worthington, \nZoe Saldana, \nSigourney We...
2 Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3 Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4 Paul Rudd, \nEvangeline Lilly, \nJonathan Majo...

```

```

                                star_id      votes main_genre
0 /name/nm4004793/,/name/nm2143282/,/name/nm1775... 204835.0      Action
1 /name/nm0941777/,/name/nm0757855/,/name/nm0000... 295119.0      Action
2 /name/nm0124930/,/name/nm1591496/,/name/nm0001... 26220.0      Action
3 /name/nm3215397/,/name/nm0000706/,/name/nm3513... 327858.0      Action
4 /name/nm0748620/,/name/nm1431940/,/name/nm3718... 5396.0      Action

```

### Cleaning director column - get director names in a specific way

```
[275]: def filter_director(data):
        return ' - '.join(data.replace(' \n', '').split(','))
```

```
[276]: movies['director'] = movies['director'].apply(filter_director)
```

```
[277]: movies.head()
```

```
[277]:  movie_id      movie_name  year certificate  runtime \
0    9114286  Black Panther: Wakanda Forever  2022      PG-13    161
1    1630029      Avatar: The Way of Water  2022      PG-13    192
2    5884796                      Plane  2023          R    107
3    6710474  Everything Everywhere All at Once  2022          R    139
4   10954600  Ant-Man and the Wasp: Quantumania  2023      PG-13    125

```

```

              genre  rating \
0  [Action, Adventure, Drama]    6.9
1  [Action, Adventure, Fantasy]    7.8
2          [Action, Thriller]    6.5
3  [Action, Adventure, Comedy]    8.0
4  [Action, Adventure, Comedy]    6.6

```

```

                                description \
0 The people of Wakanda fight to protect their h...
1 Jake Sully lives with his newfound family form...
2 A pilot finds himself caught in a war zone aft...
3 A middle-aged Chinese immigrant is swept up in...
4 Scott Lang and Hope Van Dyne, along with Hank ...

```

```

                                director      director_id \
0              Ryan Coogler /name/nm3363032/
1              James Cameron /name/nm0000116/
2              Jean-François Richet /name/nm0724938/
3 Dan Kwan - Daniel Scheinert /name/nm3453283/
4              Peyton Reed /name/nm0715636/

```

```

                                star \
0 Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1 Sam Worthington, \nZoe Saldana, \nSigourney We...
2 Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3 Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4 Paul Rudd, \nEvangeline Lilly, \nJonathan Majo...

```

```

                                star_id      votes main_genre
0 /name/nm4004793/,/name/nm2143282/,/name/nm1775... 204835.0    Action
1 /name/nm0941777/,/name/nm0757855/,/name/nm0000... 295119.0    Action
2 /name/nm0124930/,/name/nm1591496/,/name/nm0001... 26220.0     Action
3 /name/nm3215397/,/name/nm0000706/,/name/nm3513... 327858.0    Action
4 /name/nm0748620/,/name/nm1431940/,/name/nm3718... 5396.0      Action

```

**Cleaning director\_id column - extract only director id from the column values.**

```
[278]: def filter_directorid(data):
        return int(re.findall('\d+', data)[0])
```

```
[279]: movies['director_id'] = movies['director_id'].apply(filter_directorid)
```

```
[280]: movies.head()
```

```
[280]:
movie_id      movie_name  year certificate  runtime \
0    9114286  Black Panther: Wakanda Forever  2022      PG-13      161
1    1630029    Avatar: The Way of Water  2022      PG-13      192
2    5884796                Plane  2023          R       107
3    6710474  Everything Everywhere All at Once  2022          R       139
4   10954600  Ant-Man and the Wasp: Quantumania  2023      PG-13      125

```

```

                                genre  rating \
0    [Action, Adventure, Drama]      6.9

```

1	[Action, Adventure, Fantasy]	7.8
2	[Action, Thriller]	6.5
3	[Action, Adventure, Comedy]	8.0
4	[Action, Adventure, Comedy]	6.6

	description \
0	The people of Wakanda fight to protect their h...
1	Jake Sully lives with his newfound family form...
2	A pilot finds himself caught in a war zone aft...
3	A middle-aged Chinese immigrant is swept up in...
4	Scott Lang and Hope Van Dyne, along with Hank ...

	director	director_id \
0	Ryan Coogler	3363032
1	James Cameron	116
2	Jean-François Richet	724938
3	Dan Kwan - Daniel Scheinert	3453283
4	Peyton Reed	715636

	star \
0	Letitia Wright, \nLupita Nyong'o, \nDanai Guri...
1	Sam Worthington, \nZoe Saldana, \nSigourney We...
2	Gerard Butler, \nMike Colter, \nTony Goldwyn, ...
3	Michelle Yeoh, \nStephanie Hsu, \nJamie Lee Cu...
4	Paul Rudd, \nEvangeline Lilly, \nJonathan Majo...

	star_id	votes	main_genre
0	/name/nm4004793/,/name/nm2143282/,/name/nm1775...	204835.0	Action
1	/name/nm0941777/,/name/nm0757855/,/name/nm0000...	295119.0	Action
2	/name/nm0124930/,/name/nm1591496/,/name/nm0001...	26220.0	Action
3	/name/nm3215397/,/name/nm0000706/,/name/nm3513...	327858.0	Action
4	/name/nm0748620/,/name/nm1431940/,/name/nm3718...	5396.0	Action

Cleaning Star column- get all the actors names in a list

```
[281]: def filter_star(data):
        return data.replace(' \n', '').split(',')
```

```
[282]: movies['star'] = movies['star'].apply(filter_star)
```

```
[283]: movies.head()
```

	movie_id	movie_name	year	certificate	runtime \
0	9114286	Black Panther: Wakanda Forever	2022	PG-13	161
1	1630029	Avatar: The Way of Water	2022	PG-13	192
2	5884796	Plane	2023	R	107
3	6710474	Everything Everywhere All at Once	2022	R	139

4	10954600	Ant-Man and the Wasp: Quantumania	2023	PG-13	125
---	----------	-----------------------------------	------	-------	-----

		genre	rating	\
0		[Action, Adventure, Drama]	6.9	
1		[Action, Adventure, Fantasy]	7.8	
2		[Action, Thriller]	6.5	
3		[Action, Adventure, Comedy]	8.0	
4		[Action, Adventure, Comedy]	6.6	

		description	\
0		The people of Wakanda fight to protect their h...	
1		Jake Sully lives with his newfound family form...	
2		A pilot finds himself caught in a war zone aft...	
3		A middle-aged Chinese immigrant is swept up in...	
4		Scott Lang and Hope Van Dyne, along with Hank ...	

		director	director_id	\
0		Ryan Coogler	3363032	
1		James Cameron	116	
2		Jean-François Richet	724938	
3		Dan Kwan - Daniel Scheinert	3453283	
4		Peyton Reed	715636	

		star	\
0		[Letitia Wright, Lupita Nyong'o, Danai Gurira,...	
1		[Sam Worthington, Zoe Saldana, Sigourney Weave...	
2		[Gerard Butler, Mike Colter, Tony Goldwyn, Yos...	
3		[Michelle Yeoh, Stephanie Hsu, Jamie Lee Curti...	
4		[Paul Rudd, Evangeline Lilly, Jonathan Majors,...	

		star_id	votes	main_genre
0	/name/nm4004793/,/name/nm2143282/,/name/nm1775...	204835.0	Action	
1	/name/nm0941777/,/name/nm0757855/,/name/nm0000...	295119.0	Action	
2	/name/nm0124930/,/name/nm1591496/,/name/nm0001...	26220.0	Action	
3	/name/nm3215397/,/name/nm0000706/,/name/nm3513...	327858.0	Action	
4	/name/nm0748620/,/name/nm1431940/,/name/nm3718...	5396.0	Action	

```
[284]: movies['star'][0]
```

```
[284]: ['Letitia Wright', 'Lupita Nyong\'o', 'Danai Gurira', 'Winston Duke']
```

Cleaning star\_id column- extract all the actors id's in a list

```
[285]: def filter_starid(data):
        return [int(re.findall('\d+', i)[0]) for i in data.split(',')]
```

```
[286]: movies['star_id'] = movies['star_id'].apply(filter_starid)
```

```
[287]: movies.head()
```

```
[287]:  movie_id      movie_name  year certificate  runtime  \
0    9114286  Black Panther: Wakanda Forever  2022      PG-13    161
1    1630029      Avatar: The Way of Water  2022      PG-13    192
2    5884796                Plane  2023          R    107
3    6710474  Everything Everywhere All at Once  2022          R    139
4   10954600  Ant-Man and the Wasp: Quantumania  2023      PG-13    125

      genre  rating  \
0  [Action, Adventure, Drama]    6.9
1  [Action, Adventure, Fantasy]    7.8
2                [Action, Thriller]    6.5
3  [Action, Adventure, Comedy]    8.0
4  [Action, Adventure, Comedy]    6.6

      description  \
0  The people of Wakanda fight to protect their h...
1  Jake Sully lives with his newfound family form...
2  A pilot finds himself caught in a war zone aft...
3  A middle-aged Chinese immigrant is swept up in...
4  Scott Lang and Hope Van Dyne, along with Hank ...

      director  director_id  \
0          Ryan Coogler    3363032
1          James Cameron      116
2      Jean-François Richet    724938
3  Dan Kwan - Daniel Scheinert    3453283
4          Peyton Reed    715636

      star  \
0  [Letitia Wright, Lupita Nyong'o, Danai Gurira,...
1  [Sam Worthington, Zoe Saldana, Sigourney Weave...
2  [Gerard Butler, Mike Colter, Tony Goldwyn, Yos...
3  [Michelle Yeoh, Stephanie Hsu, Jamie Lee Curti...
4  [Paul Rudd, Evangeline Lilly, Jonathan Majors,...

      star_id  votes  main_genre
0  [4004793, 2143282, 1775091, 6328300]  204835.0    Action
1          [941777, 757855, 244, 2332]  295119.0    Action
2      [124930, 1591496, 1282, 5256788]   26220.0    Action
3  [3215397, 706, 3513533, 130, 702841]  327858.0    Action
4   [748620, 1431940, 3718007, 1105980]   5396.0    Action
```

## Cleaning Votes column

```
[288]: def filter_votes(data):
        return int(data)
```

```
[289]: movies['votes'] = movies['votes'].apply(filter_votes)
```

```
[290]: movies.head()
```

```
[290]:
```

	movie_id	movie_name	year	certificate	runtime \
0	9114286	Black Panther: Wakanda Forever	2022	PG-13	161
1	1630029	Avatar: The Way of Water	2022	PG-13	192
2	5884796	Plane	2023	R	107
3	6710474	Everything Everywhere All at Once	2022	R	139
4	10954600	Ant-Man and the Wasp: Quantumania	2023	PG-13	125

	genre	rating \
0	[Action, Adventure, Drama]	6.9
1	[Action, Adventure, Fantasy]	7.8
2	[Action, Thriller]	6.5
3	[Action, Adventure, Comedy]	8.0
4	[Action, Adventure, Comedy]	6.6

	description \
0	The people of Wakanda fight to protect their h...
1	Jake Sully lives with his newfound family form...
2	A pilot finds himself caught in a war zone aft...
3	A middle-aged Chinese immigrant is swept up in...
4	Scott Lang and Hope Van Dyne, along with Hank ...

	director	director_id \
0	Ryan Coogler	3363032
1	James Cameron	116
2	Jean-François Richet	724938
3	Dan Kwan - Daniel Scheinert	3453283
4	Peyton Reed	715636

	star \
0	[Letitia Wright, Lupita Nyong'o, Danai Gurira,...
1	[Sam Worthington, Zoe Saldana, Sigourney Weave...
2	[Gerard Butler, Mike Colter, Tony Goldwyn, Yos...
3	[Michelle Yeoh, Stephanie Hsu, Jamie Lee Curti...
4	[Paul Rudd, Evangeline Lilly, Jonathan Majors,...

	star_id	votes	main_genre
0	[4004793, 2143282, 1775091, 6328300]	204835	Action
1	[941777, 757855, 244, 2332]	295119	Action
2	[124930, 1591496, 1282, 5256788]	26220	Action
3	[3215397, 706, 3513533, 130, 702841]	327858	Action



```
4    [748620, 1431940, 3718007, 1105980]    5396    Action
```

### 1.1.1 Creating some new useful features

Extracting lead\_actor from each movie- The Lead Actor

```
[291]: def extract_main_data(data):  
        return data[0]
```

```
[292]: movies['lead_actor'] = movies['star'].apply(extract_main_data)
```

Calculating total\_rating\_as\_per\_votes - Product of average\_rating and total votes

```
[293]: movies['total_rating_as_per_votes'] = movies['votes'] * movies['rating']
```

### 1.1.2 Certification that yields the highest rated films

```
[294]: movies.head()
```

```
[294]:  movie_id      movie_name  year  certificate  runtime  \  
0    9114286  Black Panther: Wakanda Forever  2022      PG-13      161  
1    1630029    Avatar: The Way of Water  2022      PG-13      192  
2    5884796                Plane  2023          R      107  
3    6710474  Everything Everywhere All at Once  2022          R      139  
4   10954600  Ant-Man and the Wasp: Quantumania  2023      PG-13      125
```

```
      genre  rating  \  
0  [Action, Adventure, Drama]    6.9  
1  [Action, Adventure, Fantasy]    7.8  
2          [Action, Thriller]    6.5  
3  [Action, Adventure, Comedy]    8.0  
4  [Action, Adventure, Comedy]    6.6
```

```
      description  \  
0  The people of Wakanda fight to protect their h...  
1  Jake Sully lives with his newfound family form...  
2  A pilot finds himself caught in a war zone aft...  
3  A middle-aged Chinese immigrant is swept up in...  
4  Scott Lang and Hope Van Dyne, along with Hank ...
```

```
      director  director_id  \  
0          Ryan Coogler    3363032  
1          James Cameron      116  
2    Jean-François Richet    724938  
3  Dan Kwan - Daniel Scheinert  3453283  
4          Peyton Reed    715636
```

```
      star  \  
0
```

```

0 [Letitia Wright, Lupita Nyong'o, Danai Gurira,...
1 [Sam Worthington, Zoe Saldana, Sigourney Weave...
2 [Gerard Butler, Mike Colter, Tony Goldwyn, Yos...
3 [Michelle Yeoh, Stephanie Hsu, Jamie Lee Curti...
4 [Paul Rudd, Evangeline Lilly, Jonathan Majors,...

```

	star_id	votes	main_genre	lead_actor \
0	[4004793, 2143282, 1775091, 6328300]	204835	Action	Letitia Wright
1	[941777, 757855, 244, 2332]	295119	Action	Sam Worthington
2	[124930, 1591496, 1282, 5256788]	26220	Action	Gerard Butler
3	[3215397, 706, 3513533, 130, 702841]	327858	Action	Michelle Yeoh
4	[748620, 1431940, 3718007, 1105980]	5396	Action	Paul Rudd

	total_rating_as_per_votes
0	1413361.5
1	2301928.2
2	170430.0
3	2622864.0
4	35613.6

```
[295]: certificates_rating = movies.groupby('certificate')['rating'].mean().
      ↪sort_values(ascending=False)[1:][:10]
```

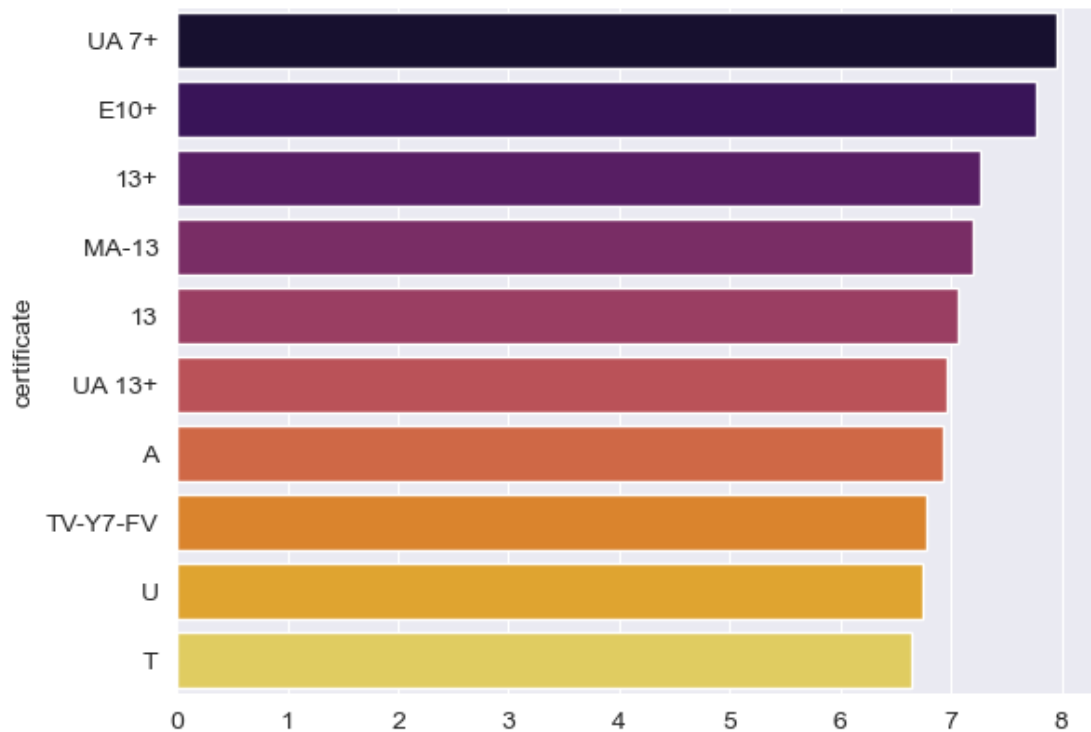
```
[296]: certificates_rating
```

```
[296]: certificate
UA 7+      7.950000
E10+      7.766667
13+       7.266667
MA-13     7.200000
13        7.060000
UA 13+    6.966667
A         6.930000
TV-Y7-FV  6.780000
U         6.754118
T         6.650000
Name: rating, dtype: float64
```

## HIGHEST RATED MOVIES CERTIFICATIONS

```
[297]: sns.barplot(x = certificates_rating.values , y = certificates_rating.index,
      ↪palette='inferno')
```

```
[297]: <AxesSubplot:ylabel='certificate'>
```



### 1.1.3 Creating a json report of the top movies in each category based on the rating and votes

```
[298]: def get_top_movies(movies):
        for genre in movies['main_genre'].unique():
            report = {}
            data = movies[movies['main_genre'] == genre]
            result = data.groupby('movie_id')['total_rating_as_per_votes'].sum().
            ↪sort_values(ascending = False)[:5]
            movie_data = []
            for idx in result.index:
                movie = data[data['movie_id'] == idx]
                movie_data.append(movie)
            top_movies = pd.concat(movie_data)
            top_movies.drop('total_rating_as_per_votes',axis = 1, inplace = True)
            report[genre] = top_movies.to_dict(orient='record')
            with open(f'reports/{genre}.json' , 'w') as file:
                json.dump(report, file, indent = 4)
```

```
[299]: get_top_movies(movies)
```

C:\Users\User\AppData\Local\Temp\ipykernel\_25876\2795265343.py:12:

FutureWarning: Using short name for 'orient' is deprecated. Only the options:

('dict', list, 'series', 'split', 'records', 'index') will be used in a future version. Use one of the above to silence this warning.

```
report[genre] = top_movies.to_dict(orient='record')
```

#### 1.1.4 Average runtime of movies from 2010-2020 using pivot.

```
[312]: avg_runtime = pd.pivot_table(data = movies, index = 'year', values='runtime',  
    ↪aggfunc='mean')[-15:-4]
```

```
[313]: avg_runtime
```

```
[313]:      runtime  
year  
2010  96.589942  
2011  95.532265  
2012  95.237148  
2013  96.292948  
2014  97.164211  
2015  97.358143  
2016  97.390386  
2017  97.822471  
2018  98.523096  
2019  99.378026  
2020  94.851186
```