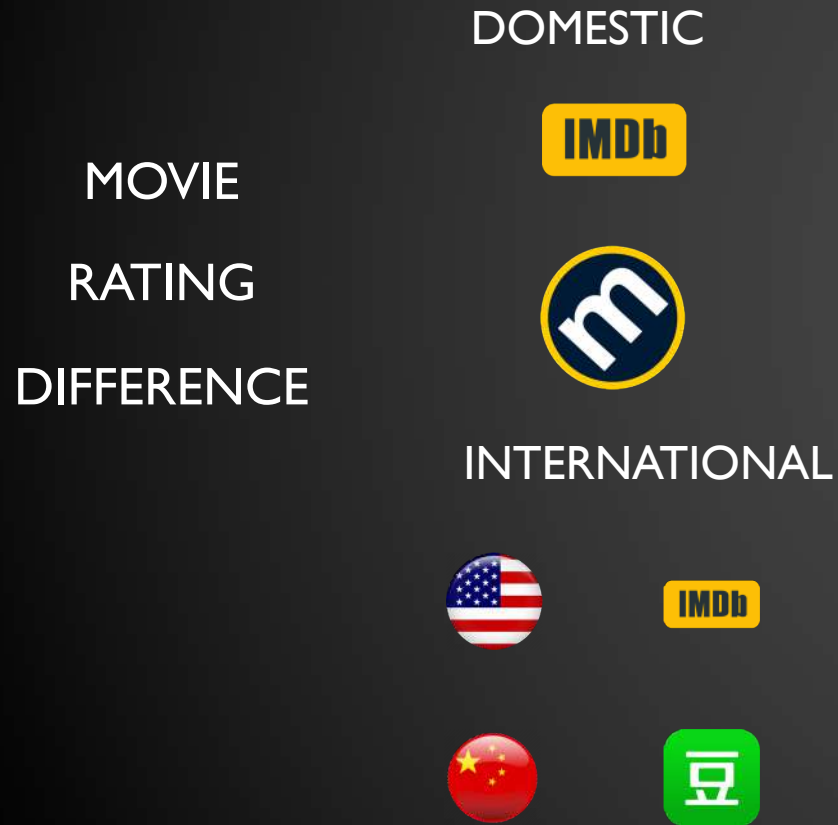


A PEEK OF GROUP DIFFERENCES THROUGH MOVIE RATING PREFERENCE

JIALAN ZHU

PROBLEM STATEMENT



WORK FLOW

WEB SCRAPING
(SCRAPY)

DATA MANIPULATION

DATA VISUALIZATION



PYTHON3

NUMPY

PANDAS

RE

PYTHON3

SEABORN

MATPLOTLIB

IMDb

+



WEB SCRAPING

IMDb

IMDb

Find Movies, TV shows, Celebrities and more...

All

Q

Movies, TV & Showtimes

Celebs, Events & Photos

News & Community


Watchlist

Most Voted Feature Films Released 2018-01-01 to 2018-12-31

1 to 50 of 14,402 titles | Next »

View Mode: Compact | Detailed

Sort by: Popularity | Alphabetical | IMDb Rating | Number of Votes | US Box Office | Runtime | Year | Release Date



1. **Avengers: Infinity War** (2018)


PG-13 | 149 min | Action, Adventure, Fantasy

★ 8.6 ☆ Rate this 68 Metascore

The Avengers and their allies must be willing to sacrifice all in an attempt to defeat the powerful Thanos before his blitz of devastation and ruin puts an end to the universe.

Directors: Anthony Russo, Joe Russo | Stars: Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans

Votes: 503,935 | Gross: \$678.82M



2. **Black Panther** (2018)

PG-13 | 134 min | Action, Adventure, Sci-Fi

★ 7.4 ☆ Rate this 88 Metascore

T'Challa, heir to the hidden but advanced kingdom of Wakanda, must step forward to lead his people into a new future and must confront a challenger from his country's past.

Director: Ryan Coogler | Stars: Chadwick Boseman, Michael B. Jordan, Lupita Nyong'o, Danai Gurira

Votes: 401,012 | Gross: \$700.06M

SCRAPING DATA:

1. MOVIE NAME
2. GENRE
3. IMDB RATING
4. METAScore
5. NUMBER OF VOTES
6. GROSS

(6301 X 7)

	genre	gross	imdb_rating	meta_rating	movie_name	movie Rated	n_of_votes
0	Drama	NaN	5.6	NaN	In My Room	NaN	96
1	Thriller	NaN	5.4	NaN	The Wrong Cruise	TV-14	96
2	Comedy	NaN	8.0	NaN	Vadhaiyaan Ji Vadhaiyaan	NaN	96
3	Biography, Comedy, Music	NaN	7.3	NaN	Ana e Vitória	NaN	96
4	Drama	NaN	6.7	NaN	L'animale	NaN	96
5	Horror, Thriller	NaN	4.7	NaN	The Witch Files	NaN	95

WEB SCRAPING



豆瓣电影 搜索电影、电视剧、综艺、影人

影讯&购票 选电影 电视剧 排行榜 分类 影评 2017年度榜单 2017观影报告

电影标签: 美国电影 + 2018

在结果中找: 科幻 动作 动画 喜剧 犯罪 奇幻 惊悚 漫画改编 爱情 (更多)

综合排序 / 评分排序 / 日期排序 / 标注次数排序

蚁人2: 黄蜂女现身 / 蚁侠2: 黄蜂女现身(港) / 蚁人2

2018-07-06(美国) / 2018-08-24(中国大陆) / 保罗·路德 / 伊万杰琳·莉莉 / 迈克尔·佩纳 / 汉娜·乔恩-卡门 / 沃尔顿·戈金斯 / 鲍比·坎纳瓦尔 / 朱迪·格雷尔 / T.I. / 大卫·达斯马齐连 / 艾比·莱德·弗特森 / 兰道尔·朴 / 迈克尔·菲佛 / 劳伦斯·菲什伯恩...

★★★★☆ 7.4 (174188人评价)

钢铁侠2 / 铁甲奇侠2(港) / 钢铁人2(台) [可播放]

2010-05-07(中国大陆) / 2010-05-07(美国) / 小罗伯特·唐尼 / 格温妮斯·帕特洛 / 基里安·墨菲 / 斯嘉丽·约翰逊 / 山姆·洛克威尔 / 唐·钱德尔 / 塞缪尔·杰克逊 / 乔恩·费儒 / 保罗·贝坦尼 / 克拉克·格雷格 / 凯特·玛拉 / 约翰·斯拉特里 / 美国...

★★★★☆ 7.3 (206809人评价)

蚁人2: 黄蜂女现身 Ant-Man and the Wasp (2018)

豆瓣评分 7.4 (174188人评价)

5星 13.4%
4星 48.2%
3星 34.8%
2星 3.2%
1星 0.3%

好于 75% 科幻片
好于 77% 动作片

导演: 佩顿·里德
编剧: 克里斯·麦克纳 / 埃里克·萨默斯 / 保罗·路德 / 安德鲁·贝伦 / 加百利·法拉利
主演: 保罗·路德 / 伊万杰琳·莉莉 / 迈克尔·佩纳 / 汉娜·乔恩-卡门 / 沃尔顿·戈金斯 / 更多...
类型: 动作 / 科幻 / 冒险
官方网站: www.marvel.com/antman
制片国家/地区: 美国
语言: 英语
上映日期: 2018-08-24(中国大陆) / 2018-07-06(美国)
片长: 118分钟 / 119分钟(中国大陆)
又名: 蚁侠2: 黄蜂女现身(港) / 蚁人2 / 蚁人与黄蜂女 / Ant-Man and the Wasp

403 ERROR!

```
headers= {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64; rv:48.0) Gecko/20100101 Firefox/48.0'}  
Request(url, headers=headers)
```

(698 X 3)

	movie_name_douban	movie_rating_douban	n_of_vote_douban
0	漫威影业: 宇宙集结 Marvel Studios: Assembling a Universe	8.5	2999.0
1	极品老妈 第三季 Mom Season 3	8.9	3595.0
2	甜蜜与卑微 Sweet and Lowdown	7.8	3509.0
3	罗尼和我 Ronny & I	8.6	4212.0
4	主厨的餐桌 第一季 Chef's Table Season 1	9.2	4986.0

SCRAPING DATA:

1. MOVIE NAME
2. DOUBAN RATING
3. NUMBER OF VOTES

DATA MANIPULATION & VISUALIZATION

IMDb

(6301 X 7)

	genre	gross	imdb_rating	meta_rating	movie_name	movie_rated	n_of_votes
0	Drama	NaN	5.6	NaN	In My Room	NaN	96
1	Thriller	NaN	5.4	NaN	The Wrong Cruise	TV-14	96
2	Comedy	NaN	8.0	NaN	Vadhayiyaan Ji Vadhayiyaan	NaN	96
3	Biography, Comedy, Music	NaN	7.3	NaN	Ana e Vitória	NaN	96
4	Drama	NaN	6.7	NaN	L'animale	NaN	96
5	Horror, Thriller	NaN	4.7	NaN	The Witch Files	NaN	95

1. DROP NA
2. CONVERT RATING SYSTEM
3. CONVERT STRING TYPE NUMBER TO NUMERIC
4. REMOVE NON-ALPHABIC CHARACTER IN MOVIE NAMES
5. CREATE COLUMN FOR MATCHING

IMDb

(406 X 8)



	genre	gross	imdb_rating	meta_rating	movie_name	movie_rated	n_of_votes	movie_name_for_match
0	drama	NaN	5.6	NaN	in my room	NaN	96	in my room
1	thriller	NaN	5.4	NaN	the wrong cruise	TV-14	96	the wrong cruise
2	comedy	NaN	8.0	NaN	vadhayiyaan ji vadhayiyaan	NaN	96	vadhayiyaan ji vadhayiyaan
3	biography, comedy, music	NaN	7.3	NaN	ana e vitória	NaN	96	ana e vit ria
4	drama	NaN	6.7	NaN	l'animale	NaN	96	l animale

IMDb

(3289 X 8)

1. INNER MERGE ON MOVIE NAME
2. REMOVE UNNECESSARY COLUMN
3. REMOVE DUPLICATE MOVIES



(35 X 10)

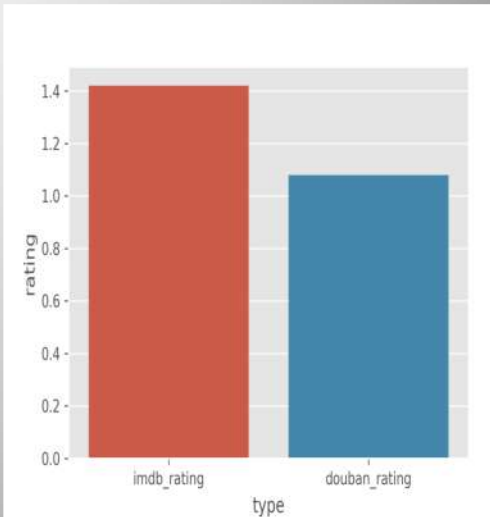
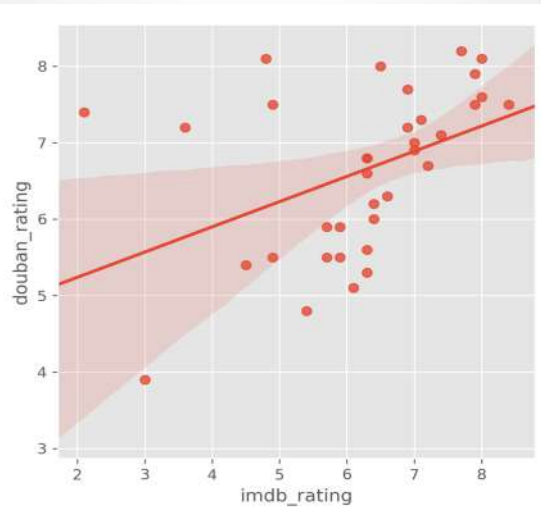
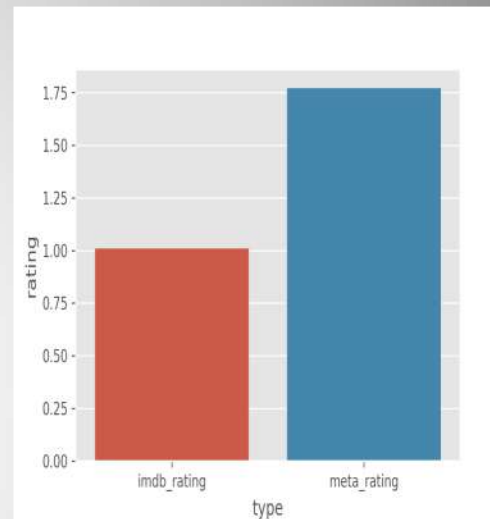
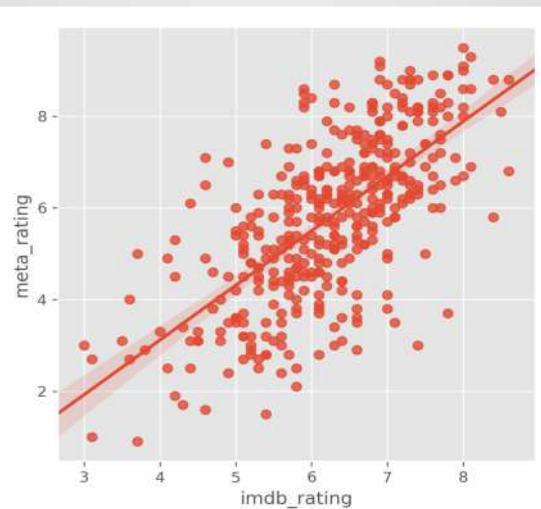
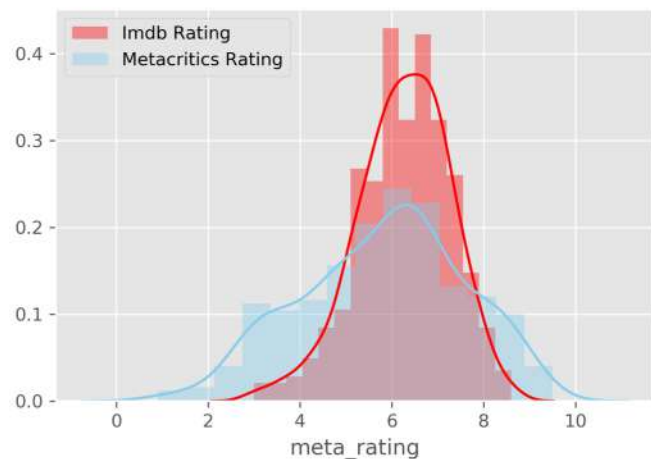


	genre	gross	imdb_rating	meta_rating	movie_name	movie_rated	n_of_votes	movie_name_for_match	movie_rating_douban	n_of_vote_douban
0	horror, thriller	NaN	6.9	NaN	sabrina	NaN	111	sabrina	7.7	16205.0
1	drama	NaN	3.6	NaN	macbeth	NaN	118	macbeth	7.2	9710.0
2	drama, family	\$1.29M	4.8	40.0	little women	PG-13	207	little women	8.1	11498.0
3	action, thriller	NaN	4.9	NaN	genius	NaN	434	genius	7.5	31606.0
4	biography, drama	NaN	8.0	91.0	the favourite	R	907	the favourite	7.6	146.0

豆

(599 X 4)

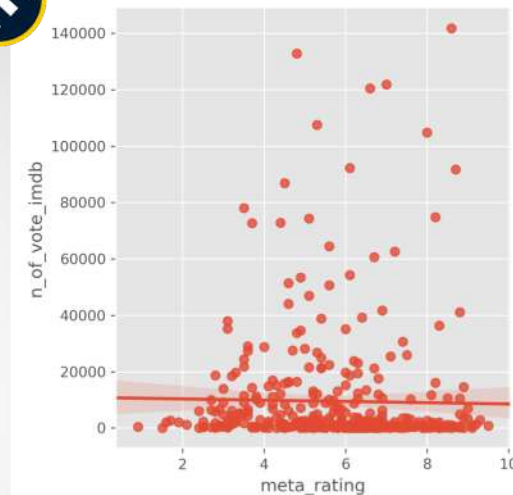
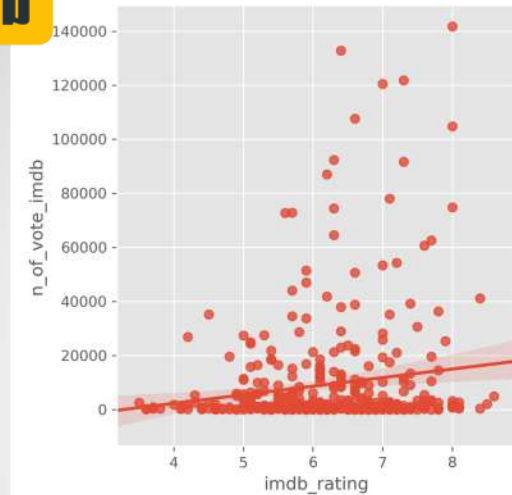
DATA MANIPULATION & VISUALIZATION



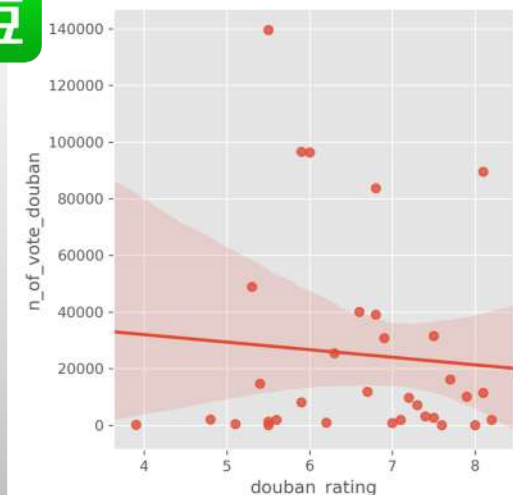
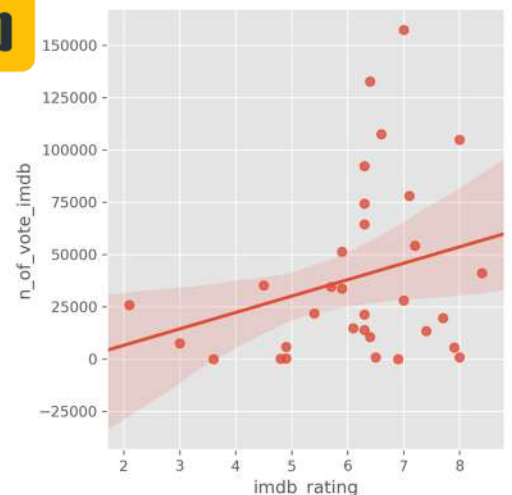
DATA MANIPULATION & VISUALIZATION

POPULARITY INFLUENCE

IMDb



IMDb



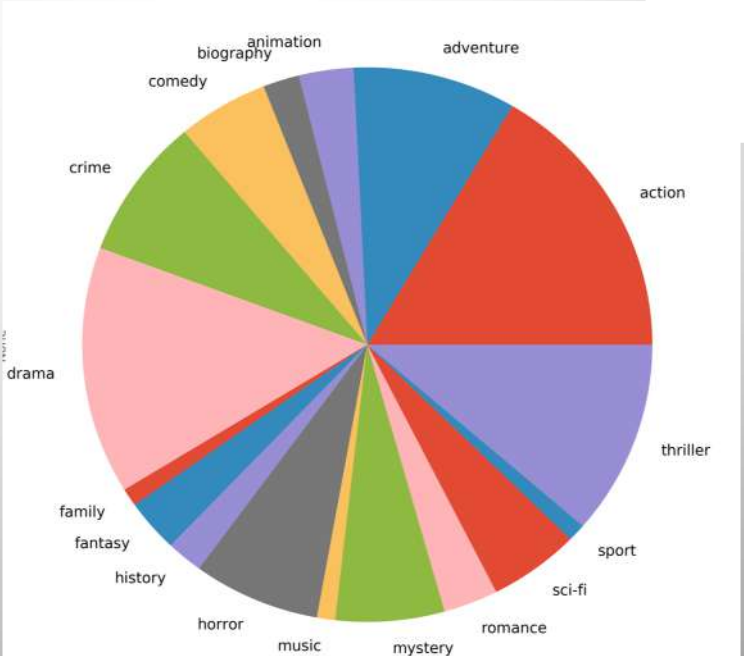
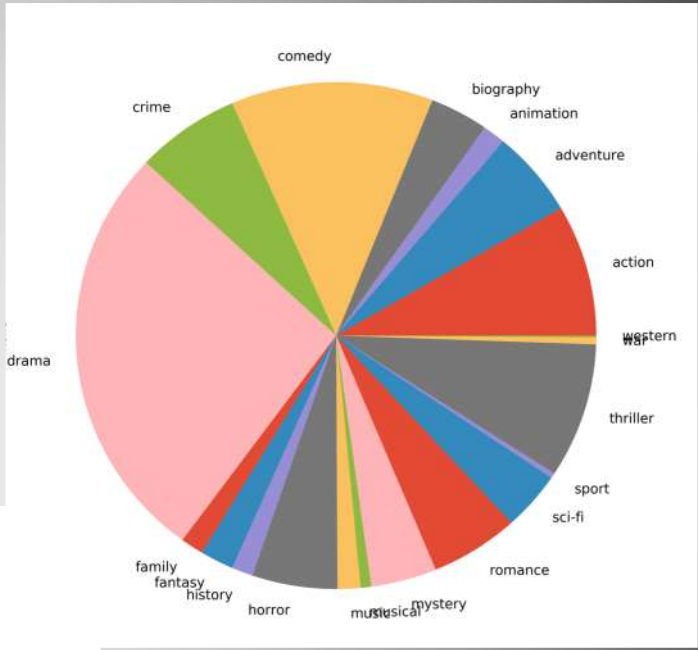
DATA MANIPULATION & VISUALIZATION

GROUP BY GENRE

genre
drama
thriller
comedy
biography, comedy, music
drama
horror, thriller

MANIPULATIONS

- 1. REMOVE COMMA
- 2. STACK THE DATA FRAMES
- 3. GROUP BY GENRE



DATA MANIPULATION & VISUALIZATION



IMDb

movie_name

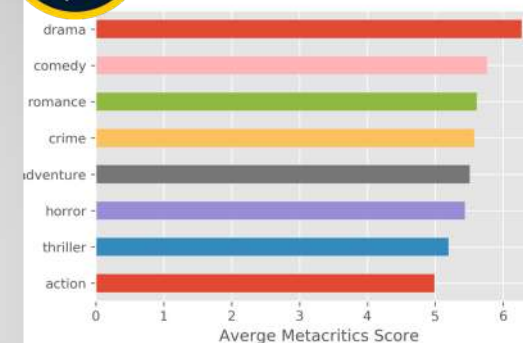
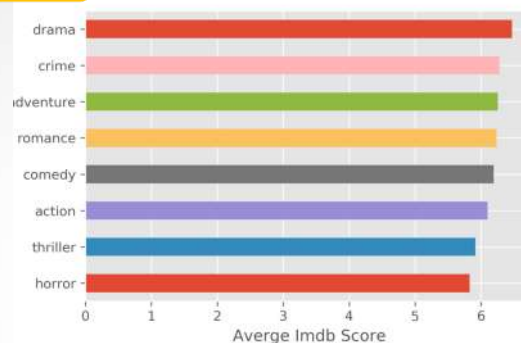
the wild pear tree
avengers: infinity war
thunder road
a star is born
dragged across concrete



movie_name

roma
shoplifters
sunday's illness
ray & liz
the favourite

IMDb



IMDb

movie_name

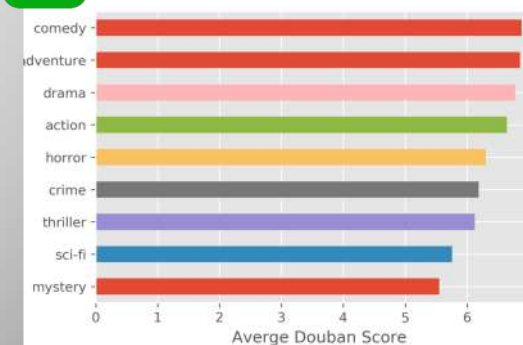
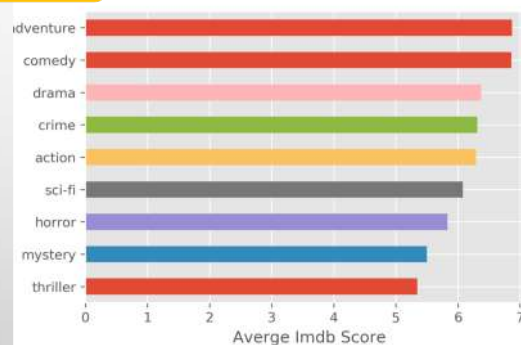
deadpool 2
annihilation
solo: a star wars story
tomb raider
red sparrow



movie_name

annihilation
pacific rim: uprising
deadpool 2
the meg
tomb raider

IMDb



Question?