IMDB.COM

WEB SCRAPING PROJECT

BY STEPHEN SHAFER

CONTENTS

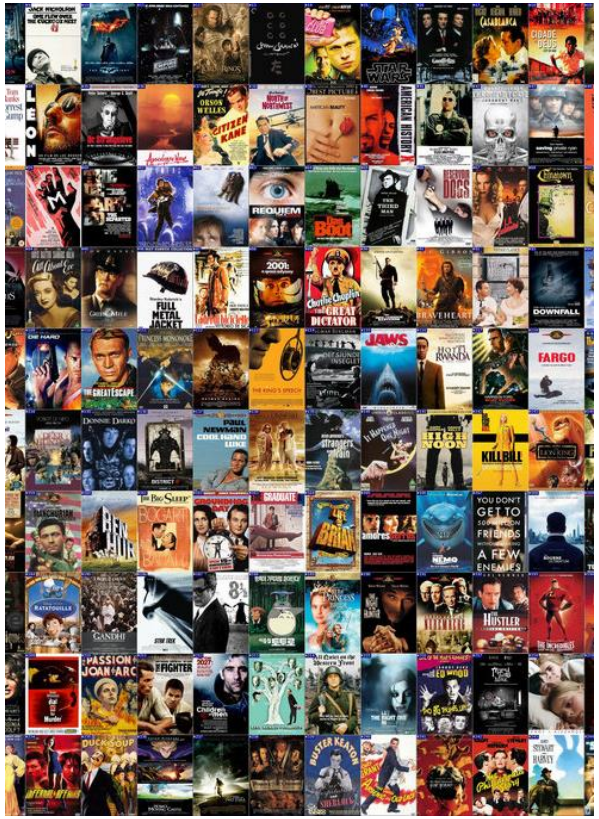
OVERVIEW

WEB SCRAPING

DATA CLEANING

ANALYSIS/VISUALIZATION

WHAT IS IMDB?

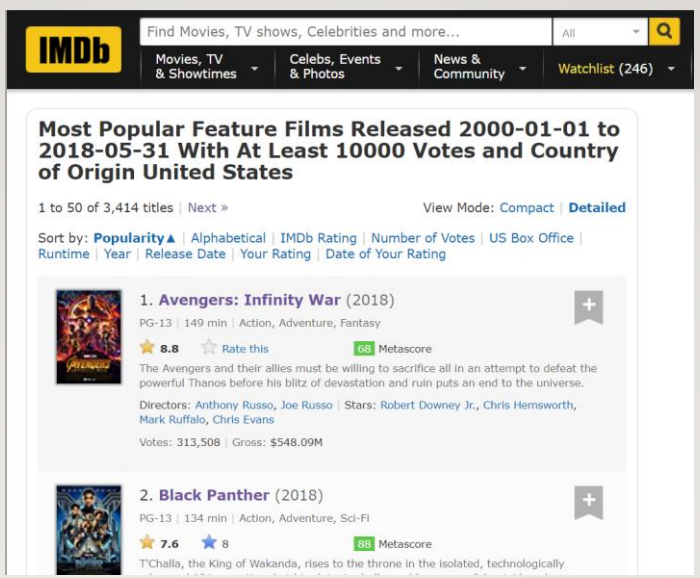


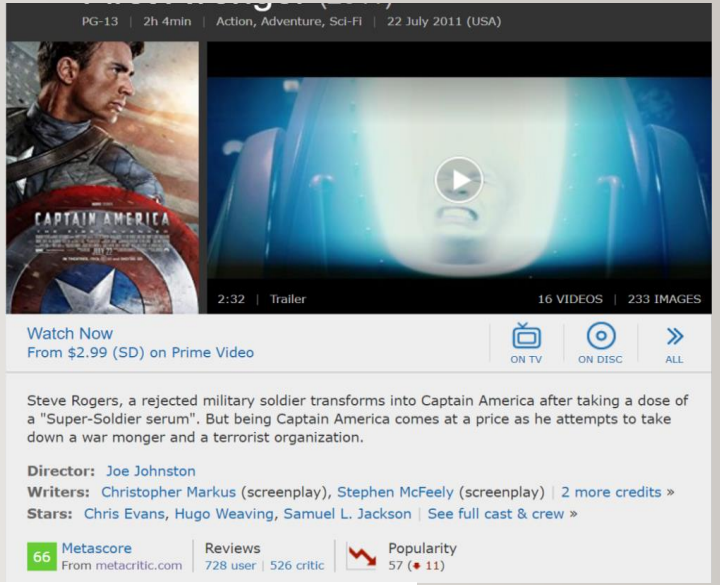
LÉON
CITIZEN KANE
REQUIEM
FULL METAL JACKET
DOWNFALL
JAWS
FARGO
KILLBILL
YOU DON'T GET TO SOUND OF FRIENDS A FEW ENEMIES
HUSTLER
PASSION JOAN of ARC

QUESTIONS ASKED GOING IN

- What is the general makeup of IMDb's users?
- Do users rate movies differently based on their demographic?
- When are the highest rated movies released?
- Does voting between these demographic differ based on who the lead actor is?

WEB SCRAPING





Rating By Demographic

	All Ages	<18	18-29	30-44	45+
All	6.9 598,397	7.2 1,934	6.9 212,165	6.7 193,525	7.0 37,829
Males	6.8 397,489	7.1 1,557	6.8 169,441	6.7 165,644	7.0 31,942
Females	7.1 77,652	7.8 364	7.2 40,679	7.0 25,311	7.3 5,285

IMDb Staff	Top 1000 Voters	US Users	Non-US Users
7.0 43	6.9 808	7.2 92,342	6.7 250,976

CLEANING

Dropping NAs and cleaning Data

In [6]:
```
# droping all rows with NAs in meta_rating and male_teen_rating
imdb_test = imdb_test.dropna(subset=["male_under18_rating"])
imdb_test = imdb_test.dropna(subset=["meta_rating"])
#imdb_test[imdb_test.isnull().any(axis=1)]
```

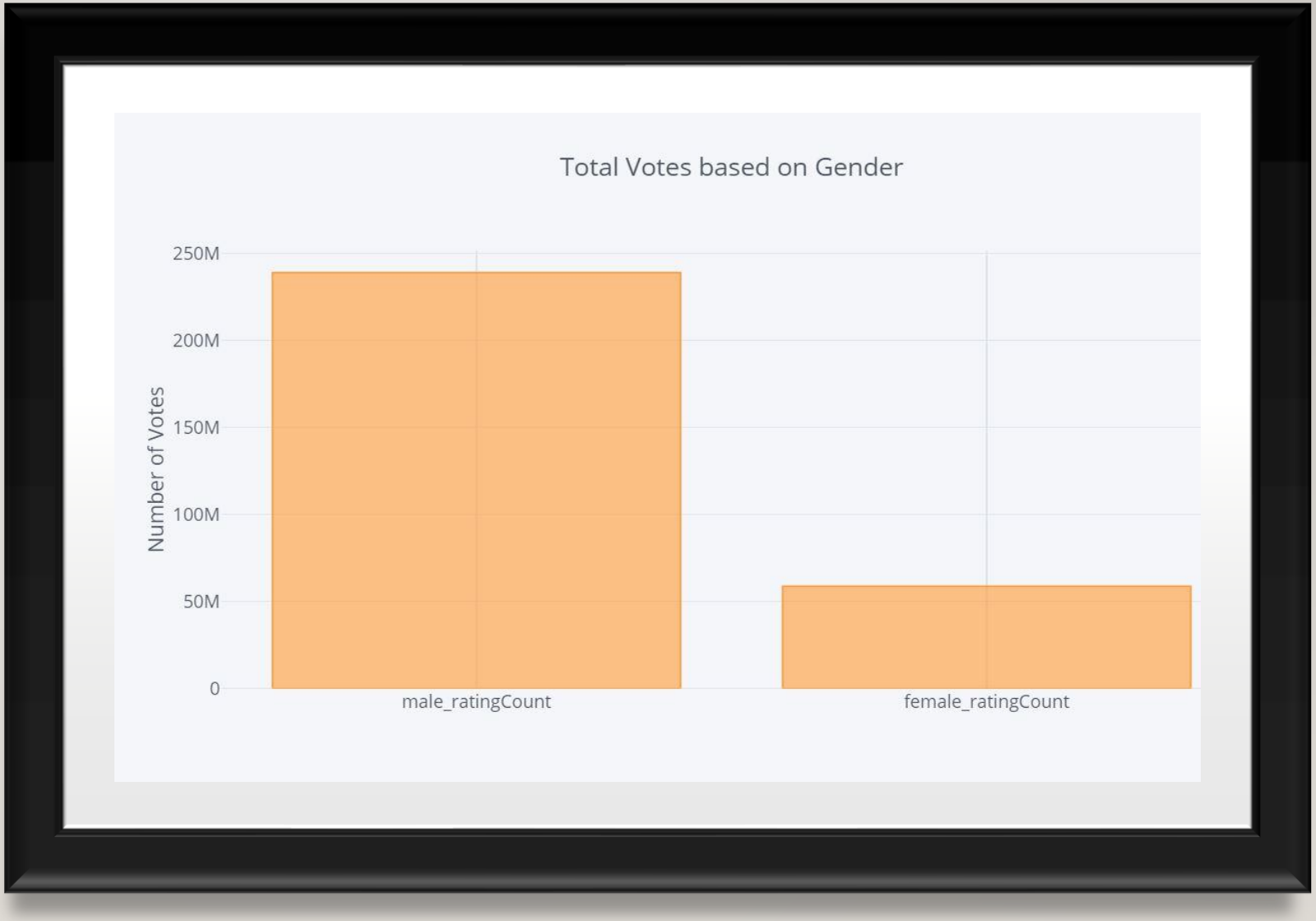
In [7]:
```
# changing rating counts into int type
imdb_test.loc[:,"female_ratingCount"] = imdb_test.loc[:,"female_ratingCount"].astype(int)
imdb_test.loc[:,"male_ratingCount"] = imdb_test.loc[:,"male_ratingCount"].astype(int)
imdb_test.loc[:,"meta_rating"] = imdb_test.loc[:,"meta_rating"].astype(int)
```

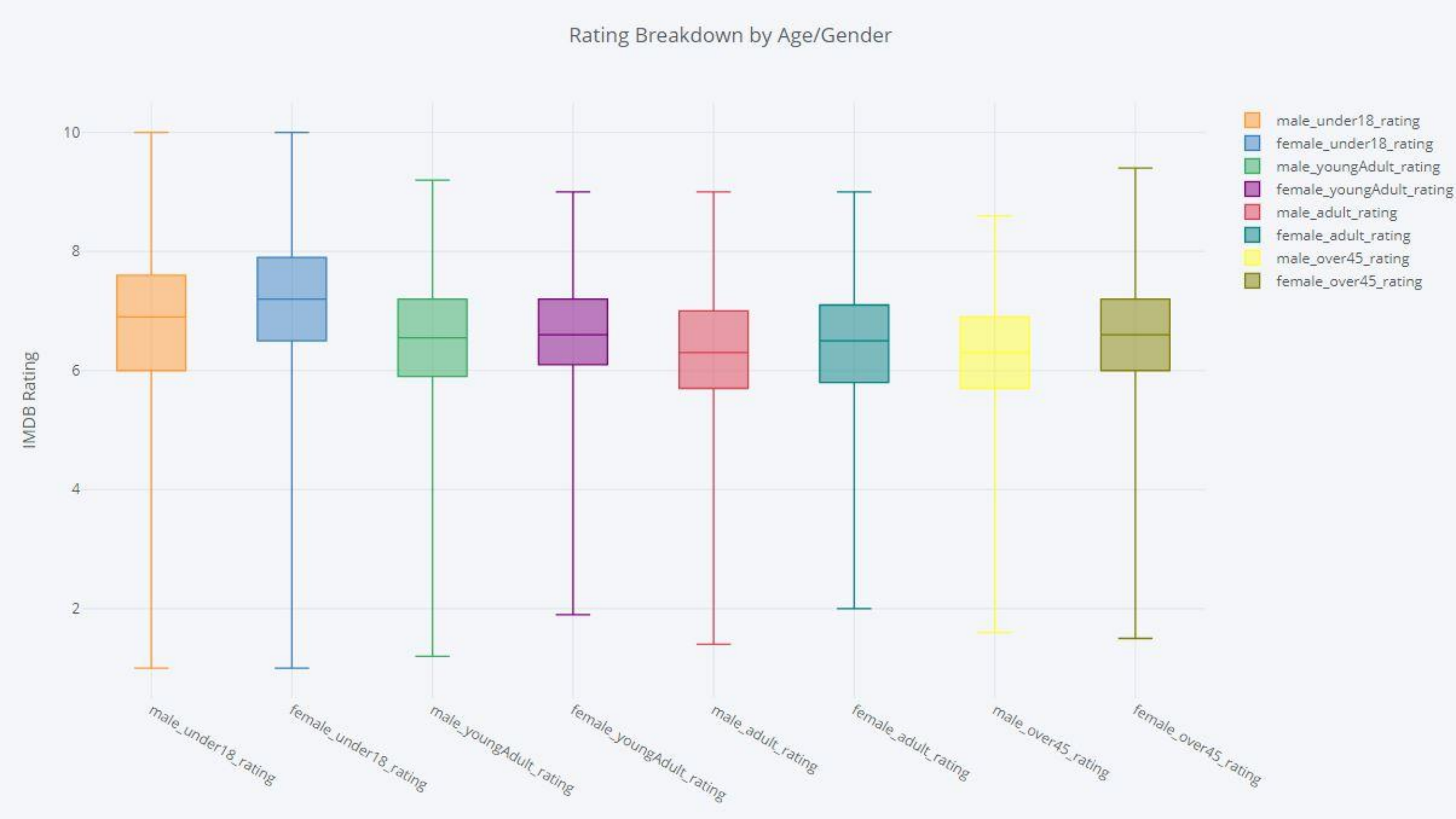
In [8]:
```
# converting release_date into datetime format
import re

def split_it(year):
    return re.findall('(\d+ \w+ \d+)', year)

imdb_test['release_date'] = imdb_test['release_date'].apply(split_it)
imdb_test['release_date'] = imdb_test['release_date'].apply(lambda x: ','.join(map(str, x)))
imdb_test['release_date']= pd.to_datetime(imdb_test['release_date'])
```

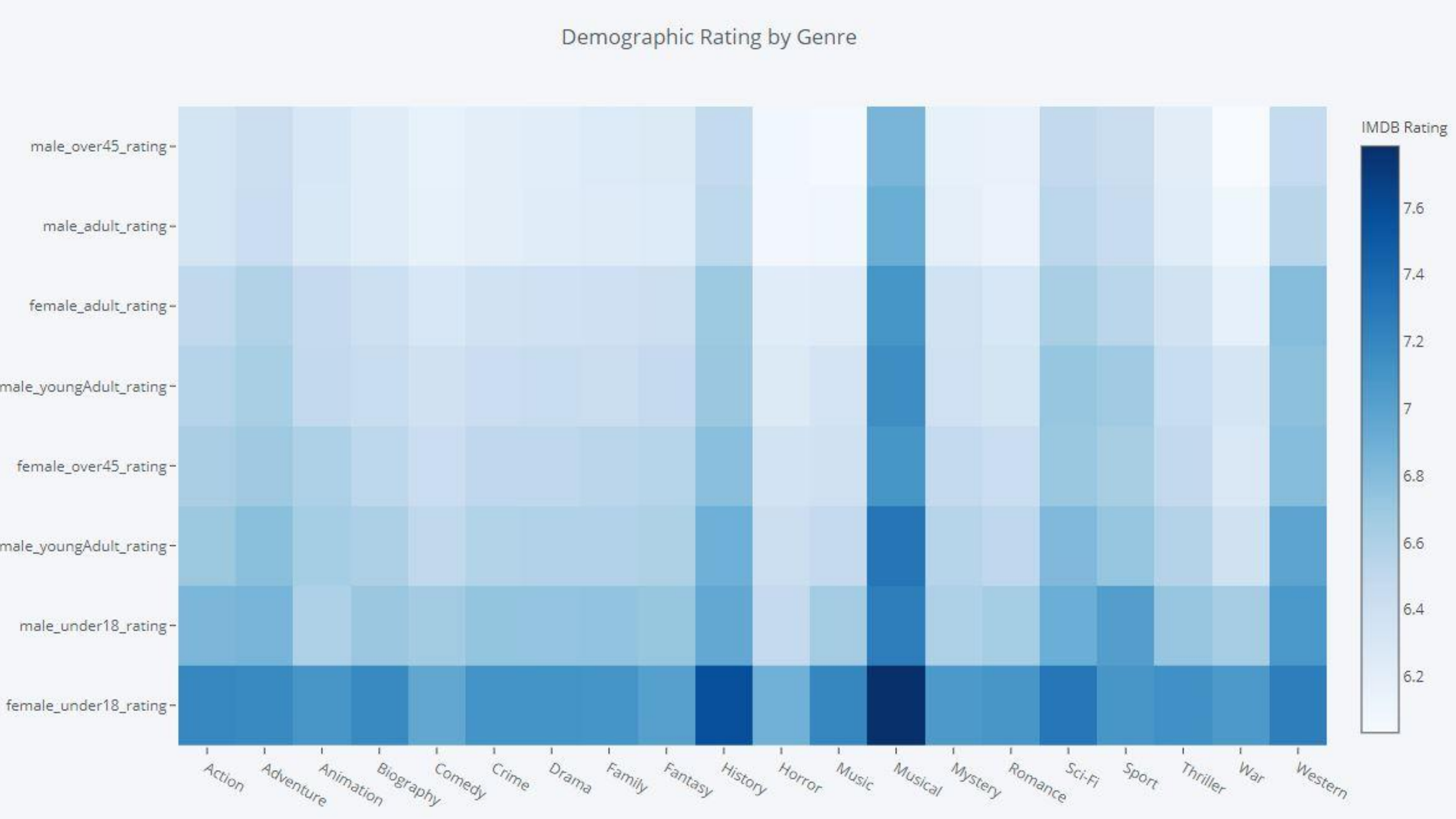
In [9]:
```
# giving year its own column
imdb_test["release_year"] = imdb_test["release_date"].apply (lambda x : x.year)
imdb_test = imdb_test.dropna(subset=["release_year"])
imdb_test.release_year = imdb_test.release_year.astype(int)
```

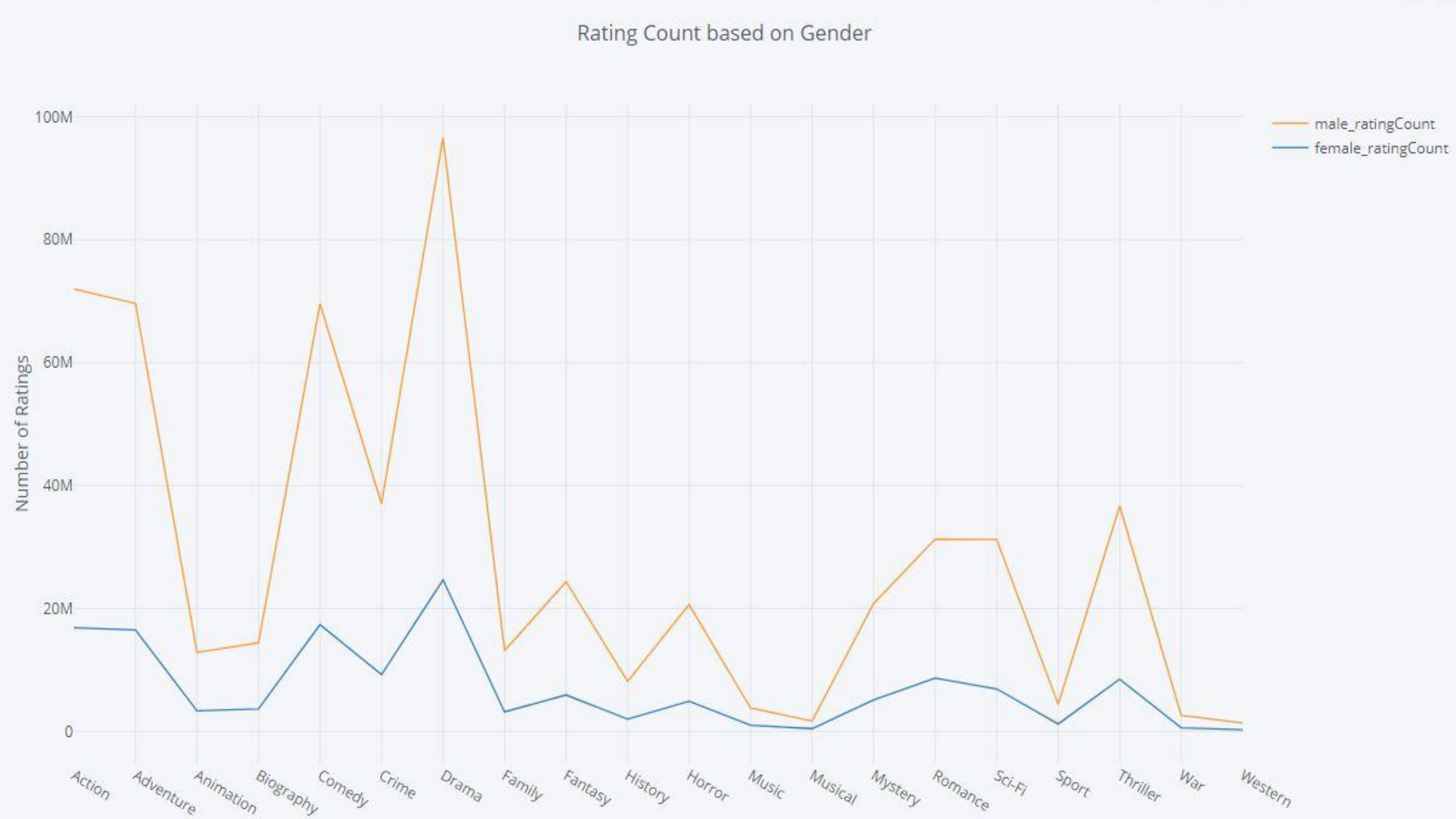
In [10]:
```
# giving month its own column
imdb_test["release_month"] = imdb_test["release_date"].apply (lambda x : x.month)
imdb_test = imdb_test.dropna(subset=["release_month"])
imdb_test.release_month = imdb_test.release_month.astype(int)
```

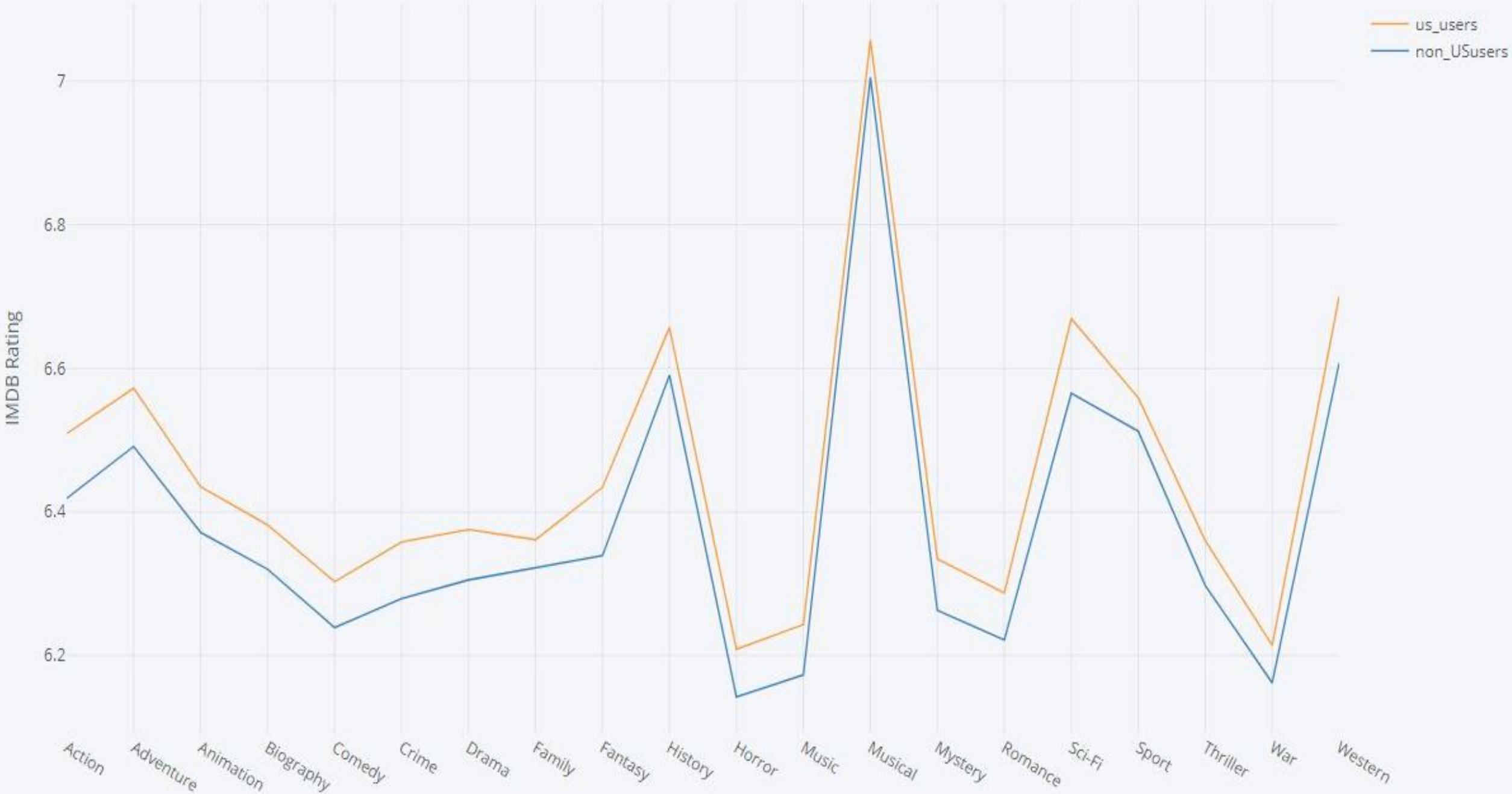
ANALYSIS & VISUALIZATION

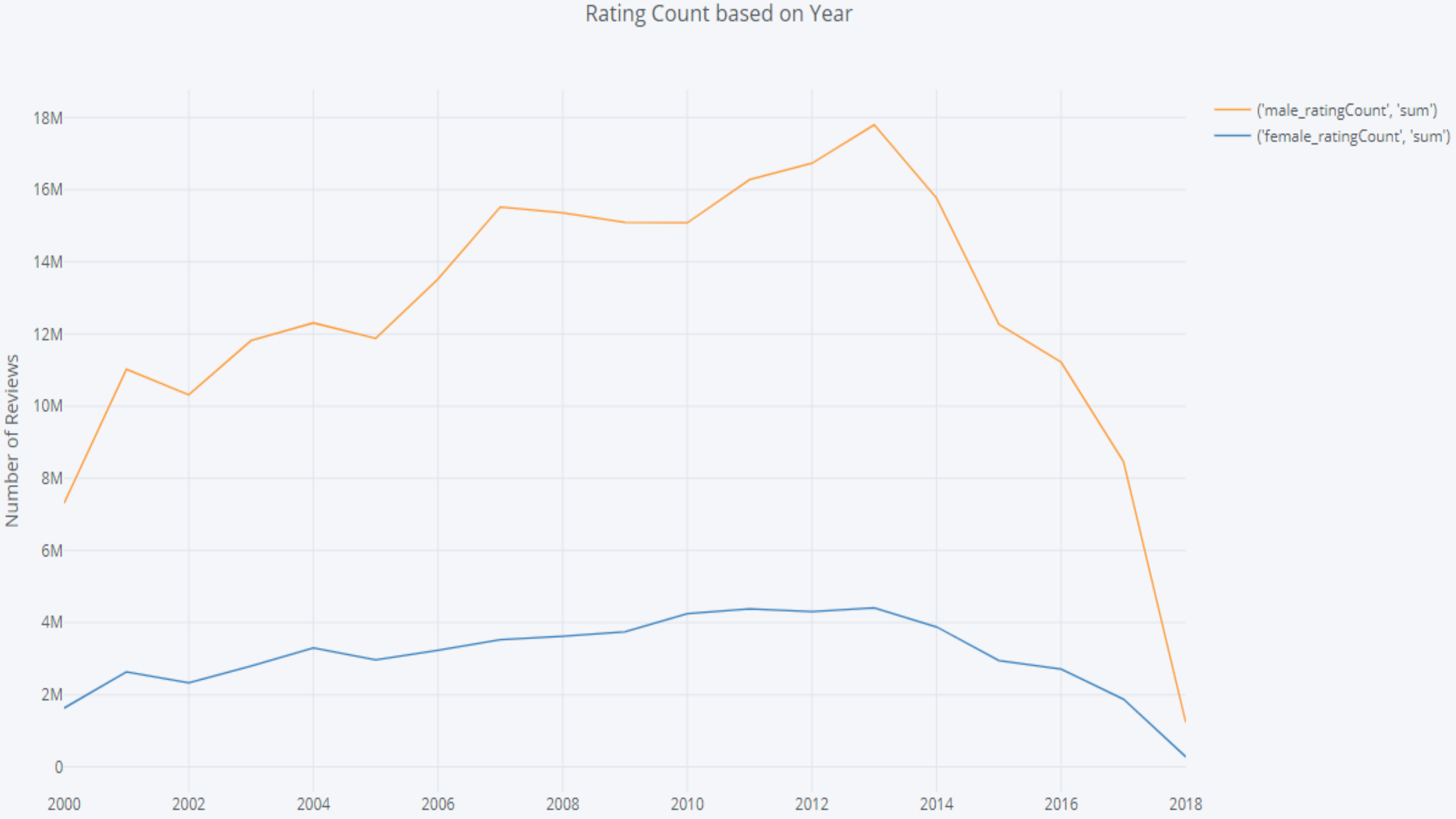



Rating Breakdown by Age/Gender
IMDB Rating
10
8
6
4
2
male_under18_rating
female_under18_rating
male_youngAdult_rating
female_youngAdult_rating
male_adult_rating
female_adult_rating
male_over45_rating
female_over45_rating
male_under18_rating
female_under18_rating
male_youngAdult_rating
female_youngAdult_rating
male_adult_rating
female_adult_rating
male_over45_rating
female_over45_rating


Demographic Rating by Genre
male_over45_rating
male_adult_rating
female_adult_rating
male_youngAdult_rating
female_over45_rating
male_youngAdult_rating
male_under18_rating
female_under18_rating
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western
IMDB Rating
7.6
7.4
7.2
7
6.8
6.6
6.4
6.2

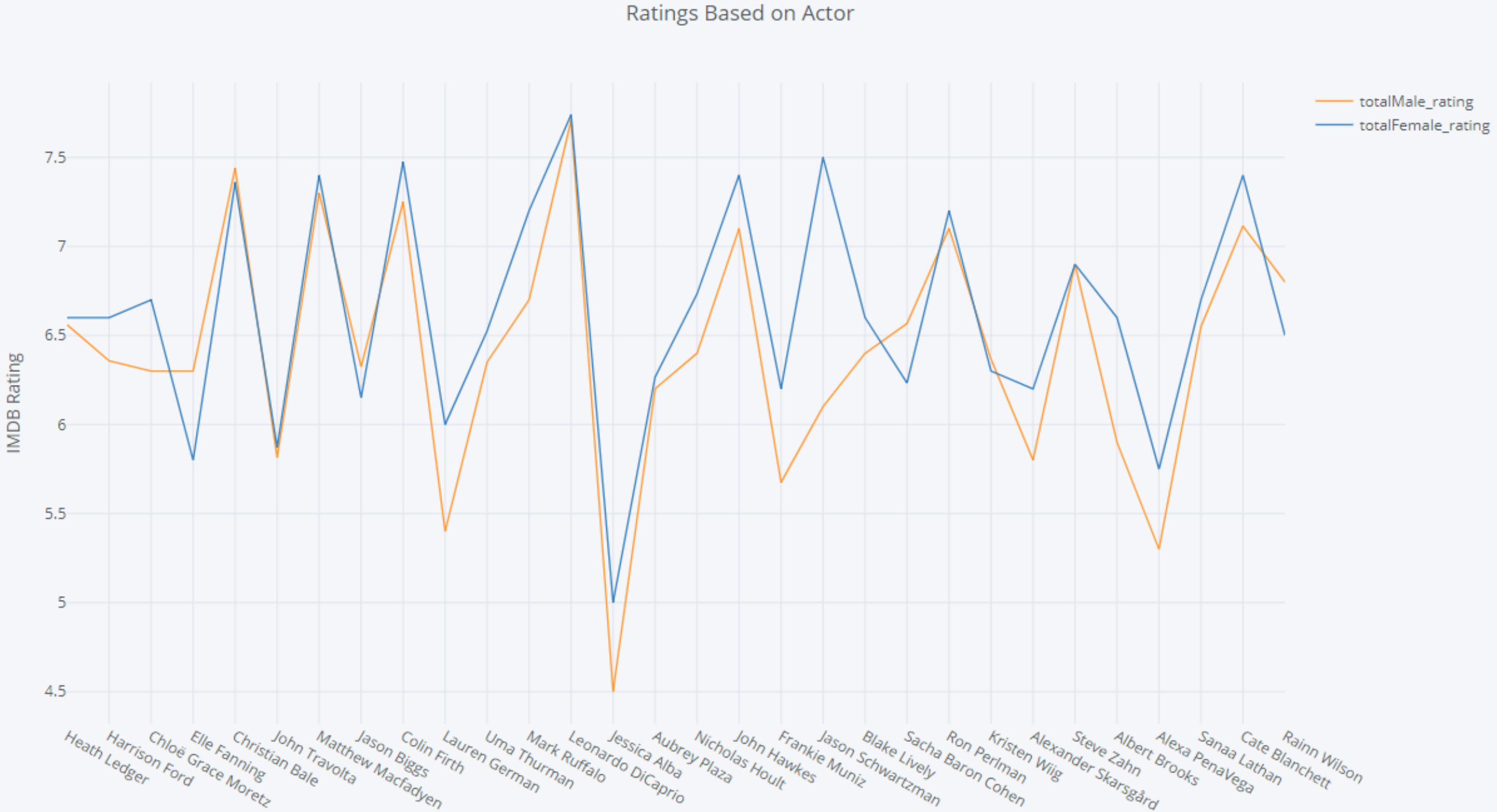

Rating Count based on Gender
male_ratingCount
female_ratingCount
Number of Ratings
100M
80M
60M
40M
20M
0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western

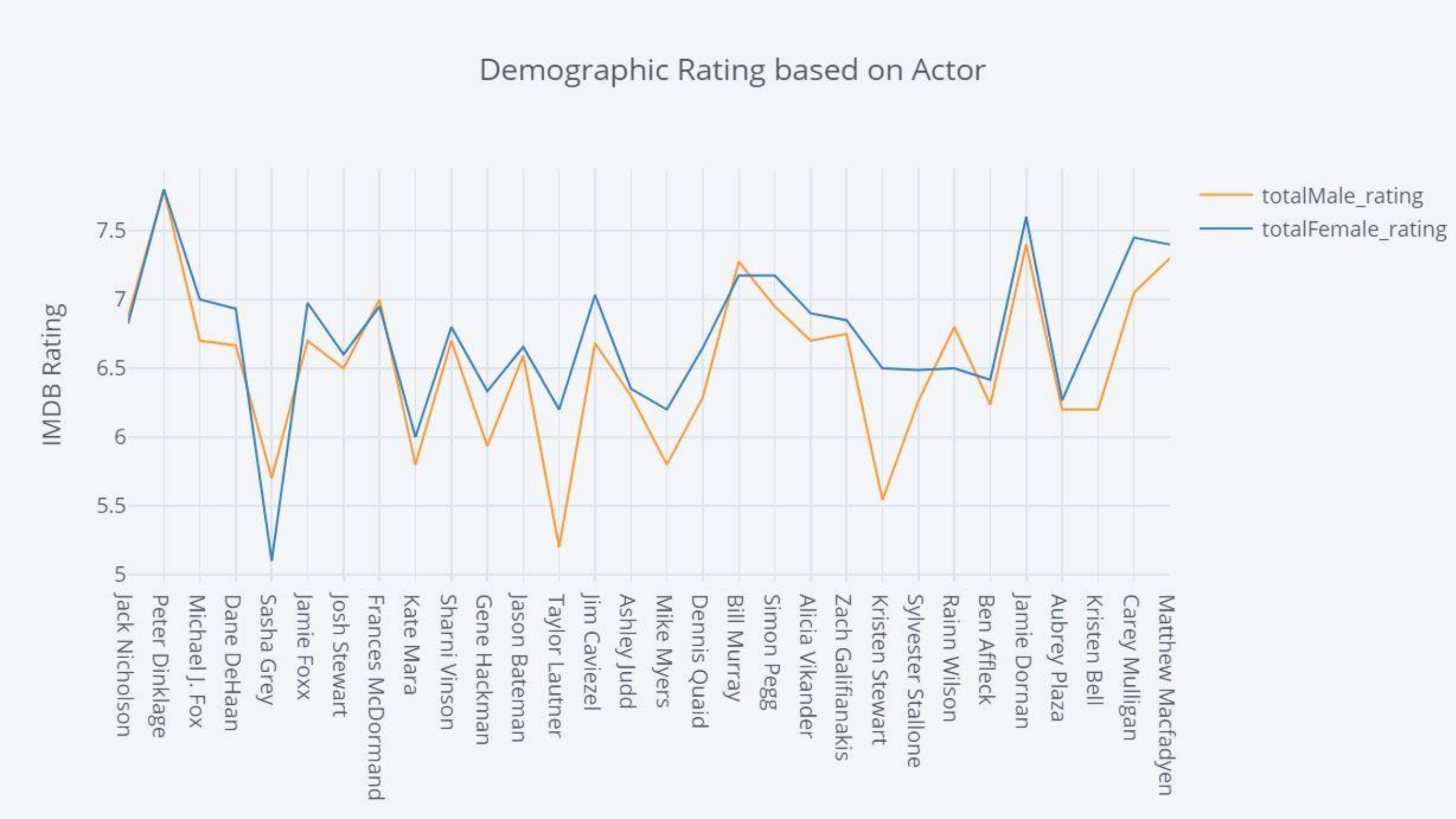

us_users
non_USusers
IMDB Rating
7
6.8
6.6
6.4
6.2
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western


Rating Count based on Year
('male_ratingCount', 'sum')
('female_ratingCount', 'sum')
Number of Reviews
18M
16M
14M
12M
10M
8M
6M
4M
2M
0
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

Mean Rating by Year
('totalMale_rating', 'mean')
('totalFemale_rating', 'mean')
IMDB Rating
7
6.9
6.8
6.7
6.6
6.5
6.4
6.3
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

Mean Rating by Month
IMDB Rating
6.9
6.8
6.7
6.6
6.5
6.4
6.3
6.2
2
4
6
8
10
12
('totalMale_rating', 'mean')
('totalFemale_rating', 'mean')


Ratings Based on Actor
totalMale_rating
totalFemale_rating
IMDB Rating
7.5
7
6.5
6
5.5
5
4.5
Heath Ledger
Harrison Ford
Chloë Grace Moretz
Elle Fanning
Christian Bale
John Travolta
Matthew Macfadyen
Jason Biggs
Colin Firth
Lauren German
Uma Thurman
Mark Ruffalo
Leonardo DiCaprio
Jessica Alba
Aubrey Plaza
Nicholas Hoult
John Hawkes
Frankie Muniz
Jason Schwartzman
Blake Lively
Sacha Baron Cohen
Ron Perlman
Kristen Wiig
Alexander Skarsgård
Steve Zahn
Albert Brooks
Alexa PenaVega
Sanaa Lathan
Cate Blanchett
Rainn Wilson


Demographic Rating based on Actor
IMDB Rating
7.5
7
6.5
6
5.5
5
totalMale_rating
totalFemale_rating
Jack Nicholson
Peter Dinklage
Michael J. Fox
Dane DeHaan
Sasha Grey
Jamie Foxx
Josh Stewart
Frances McDormand
Kate Mara
Sharni Vinson
Gene Hackman
Jason Bateman
Taylor Lautner
Jim Caviezel
Ashley Judd
Mike Myers
Dennis Quaid
Bill Murray
Simon Pegg
Alicia Vikander
Zach Galifianakis
Kristen Stewart
Sylvester Stallone
Rainn Wilson
Ben Affleck
Jamie Dornan
Aubrey Plaza
Kristen Bell
Carey Mulligan
Matthew Macfadyen

CONCLUSION