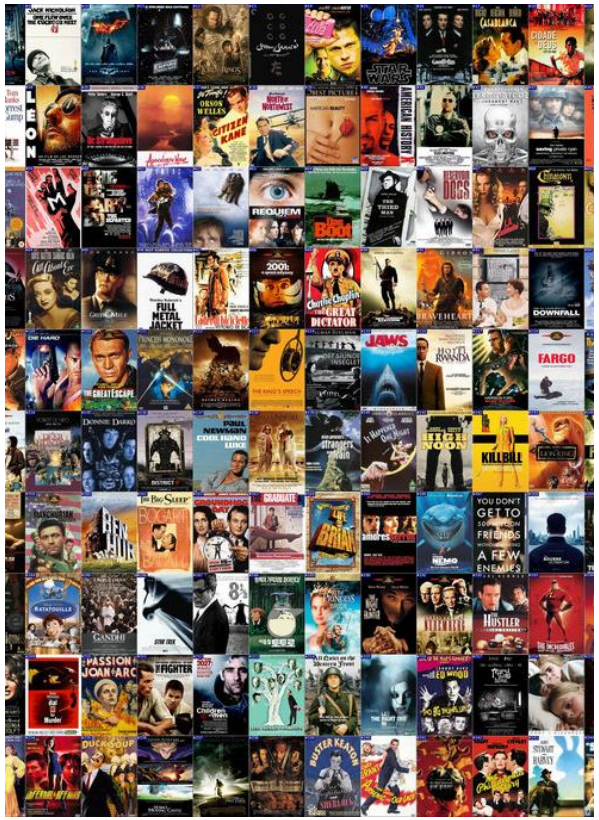
IMDB.COM

WEB SCRAPING PROJECT

BY STEPHEN SHAFER

CONTENTS

WHAT IS IMDB?


CASABLANCA
STAR WARS
CITIZEN KANE
REQUIEM
FULL METAL JACKET
THE GREAT DICTATOR
JAWS
FARGO
KILL BILL
HIGH NOON
GRADUATE
HUSTLER
DUCK SOUP

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

```
item = ImdbItem()
item['run_time'] = run_time
item['genre'] = genre
item['title'] = title
item['imdb_rating'] = imdb_rating
item['meta_rating'] = meta_rating
item['MPAA_rating'] = MPAA_rating
item['release_date'] = release_date
item['director'] = director
item['actors'] = actors
item['male_teen_rating'] = male_teen_rating
item['male_youngAdult_rating'] = male_youngAdult_rating
item['male_adult_rating'] = male_adult_rating
item['male_elder_rating'] = male_elder_rating
item['male_ratingCount'] = male_ratingCount
item['female_teen_rating'] = female_teen_rating
item['female_youngAdult_rating'] = female_youngAdult_rating
item['female_adult_rating'] = female_adult_rating
item['female_elder_rating'] = female_elder_rating
item['female_ratingCount'] = female_ratingCount
item['non_USusers'] = non_USusers
item['non_UScount'] = non_UScount
item['us_users'] = us_users
item['us_count'] = us_count


yield item
```

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)]
```

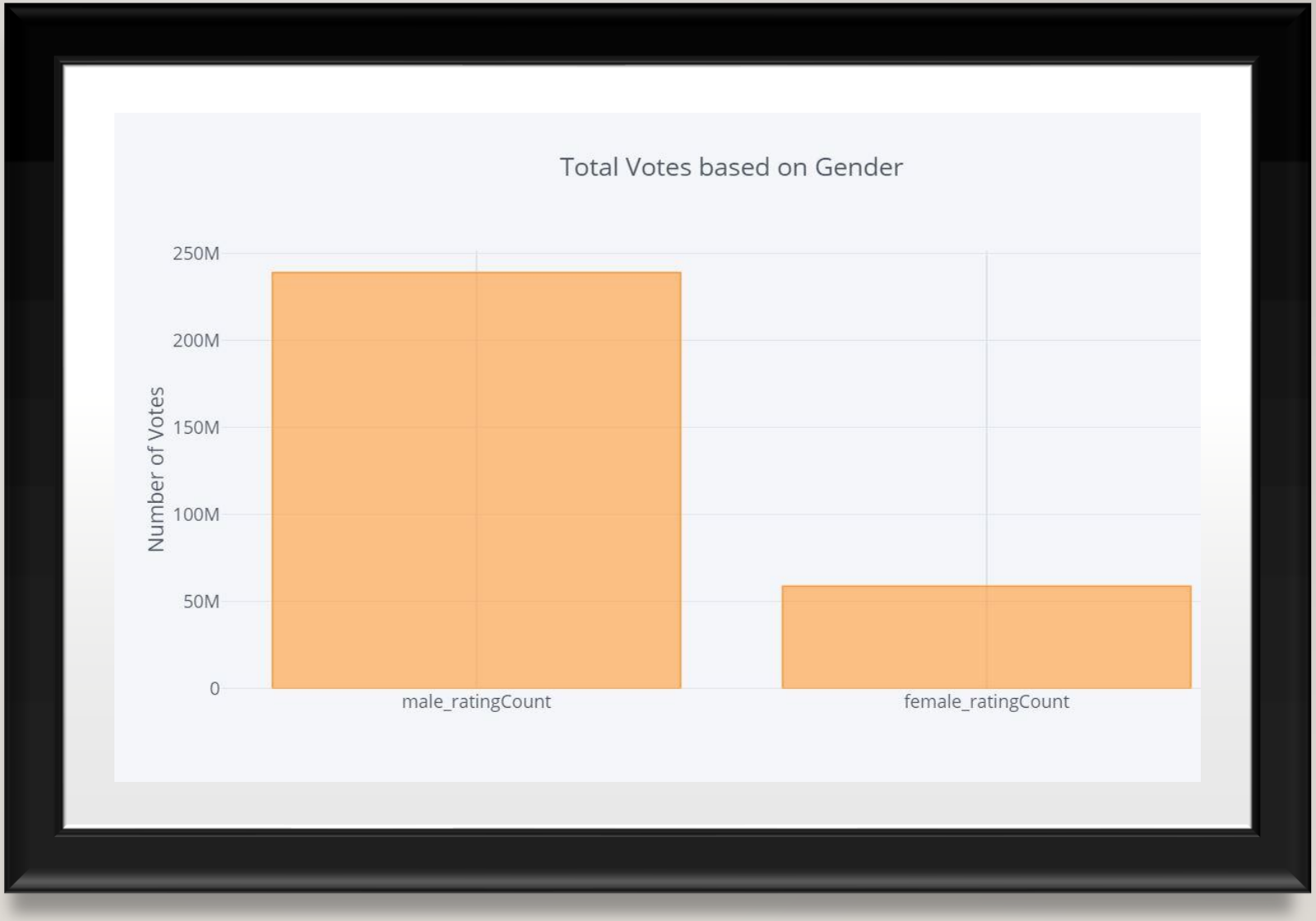
In [7]:
```
# changing rating counts into int type
imdb_test.loc[:,"female_ratingCount"] = imdb_test.loc[:,"female_ratingCount"].astype(int)
imdb_test.loc[:,"male_ratingCount"] = imdb_test.loc[:,"male_ratingCount"].astype(int)
imdb_test.loc[:,"meta_rating"] = imdb_test.loc[:,"meta_rating"].astype(int)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

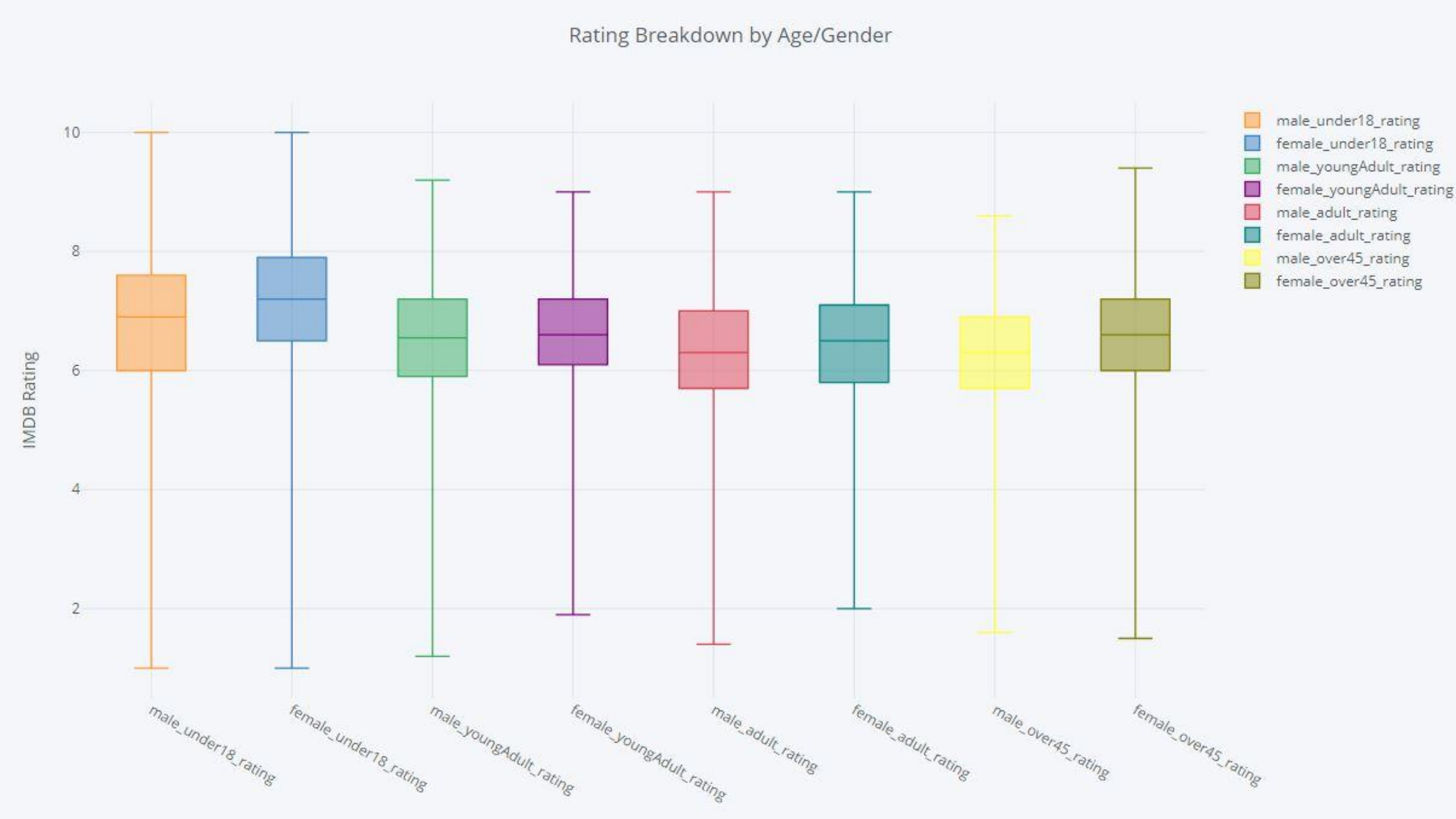
In [8]:
```
# converting release_date into datetime format
import re

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'])
```

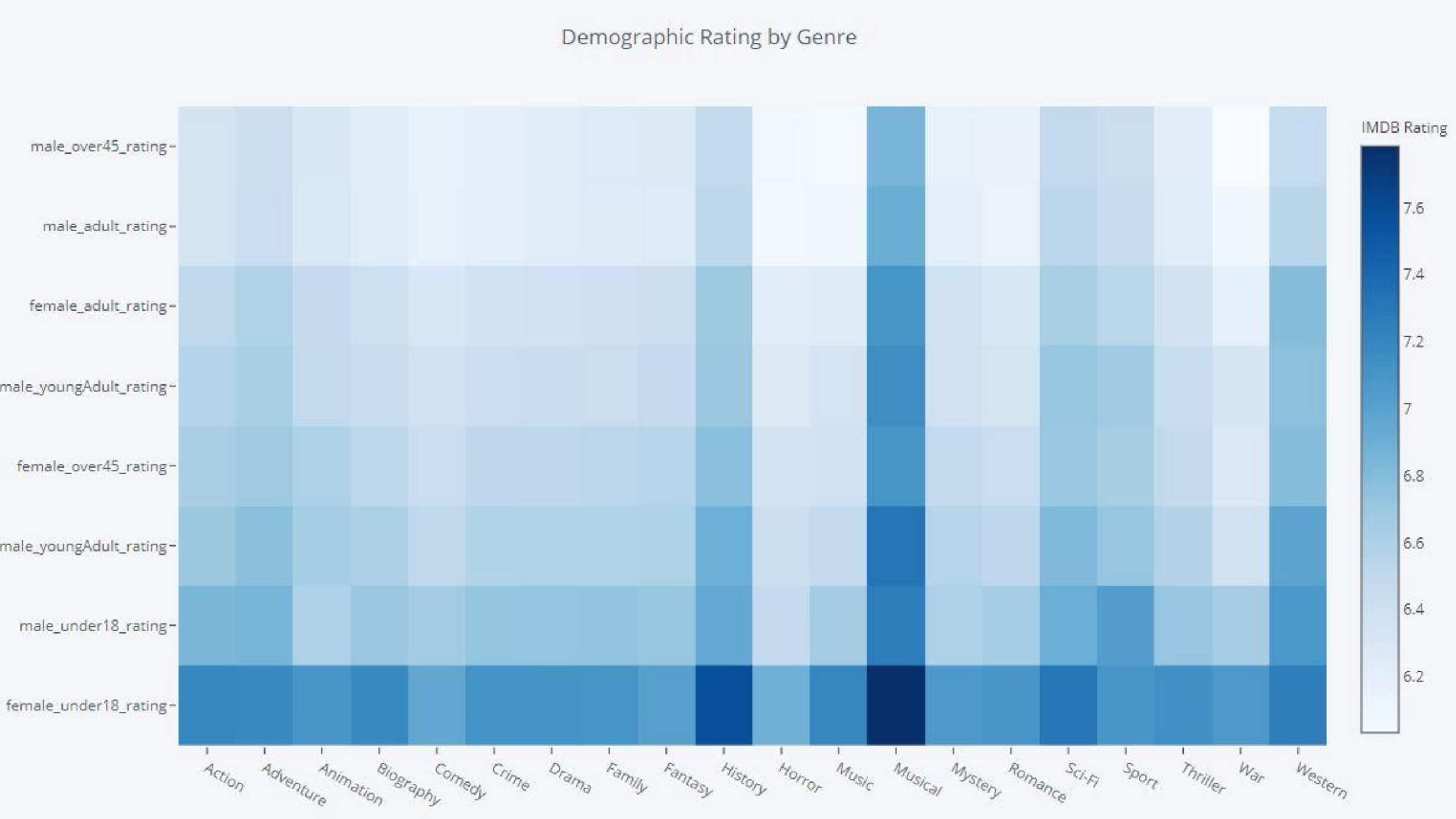
In [9]:
```
# giving year its own column
imdb_test["release_year"] = imdb_test["release_date"].apply (lambda x : x.year)
imdb_test = imdb_test.dropna(subset=["release_year"])
imdb_test.release_year = imdb_test.release_year.astype(int)
```

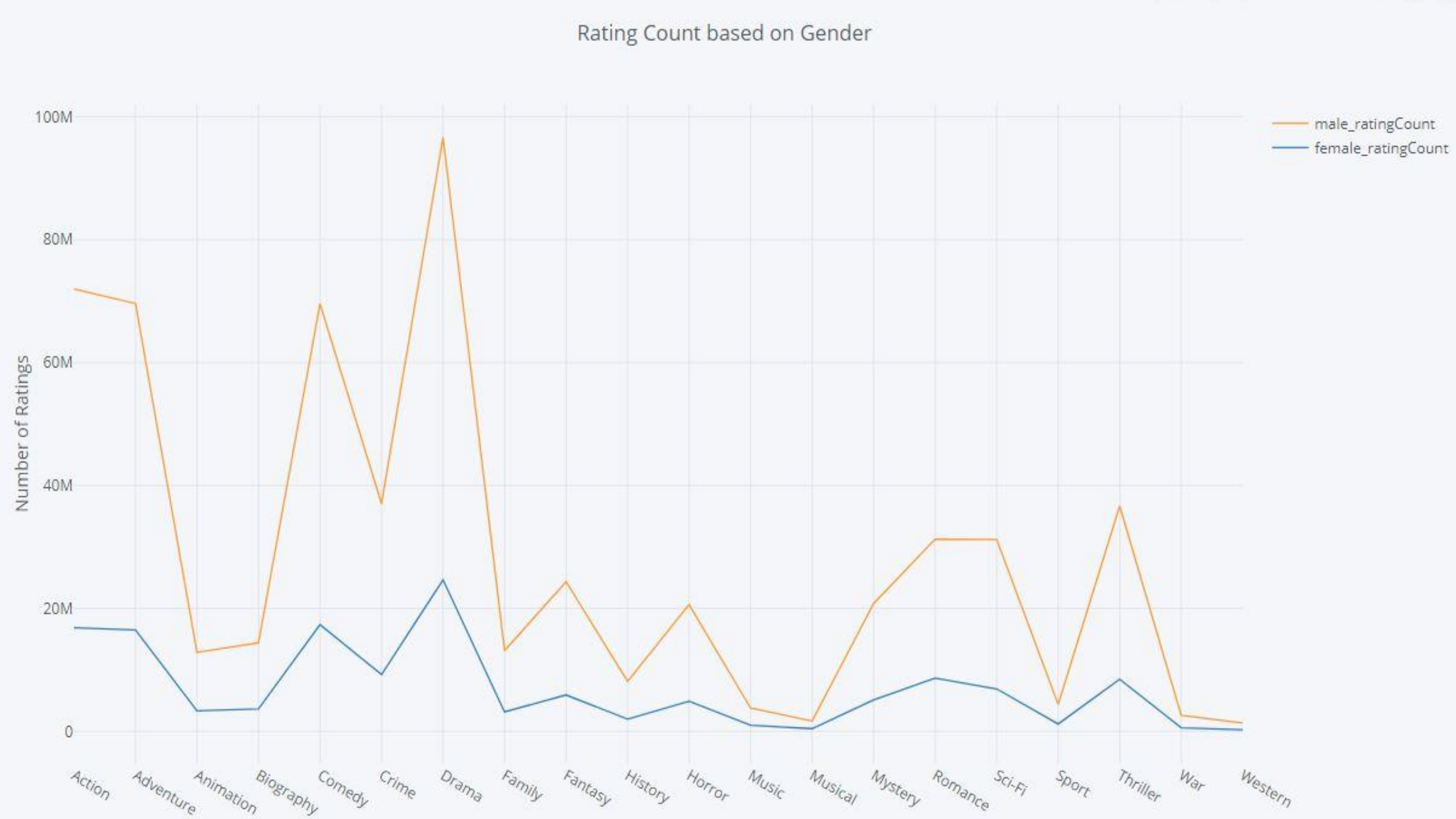
In [10]:
```
# giving month its own column
imdb_test["release_month"] = imdb_test["release_date"].apply (lambda x : x.month)
imdb_test = imdb_test.dropna(subset=["release_month"])
imdb_test.release_month = imdb_test.release_month.astype(int)
```

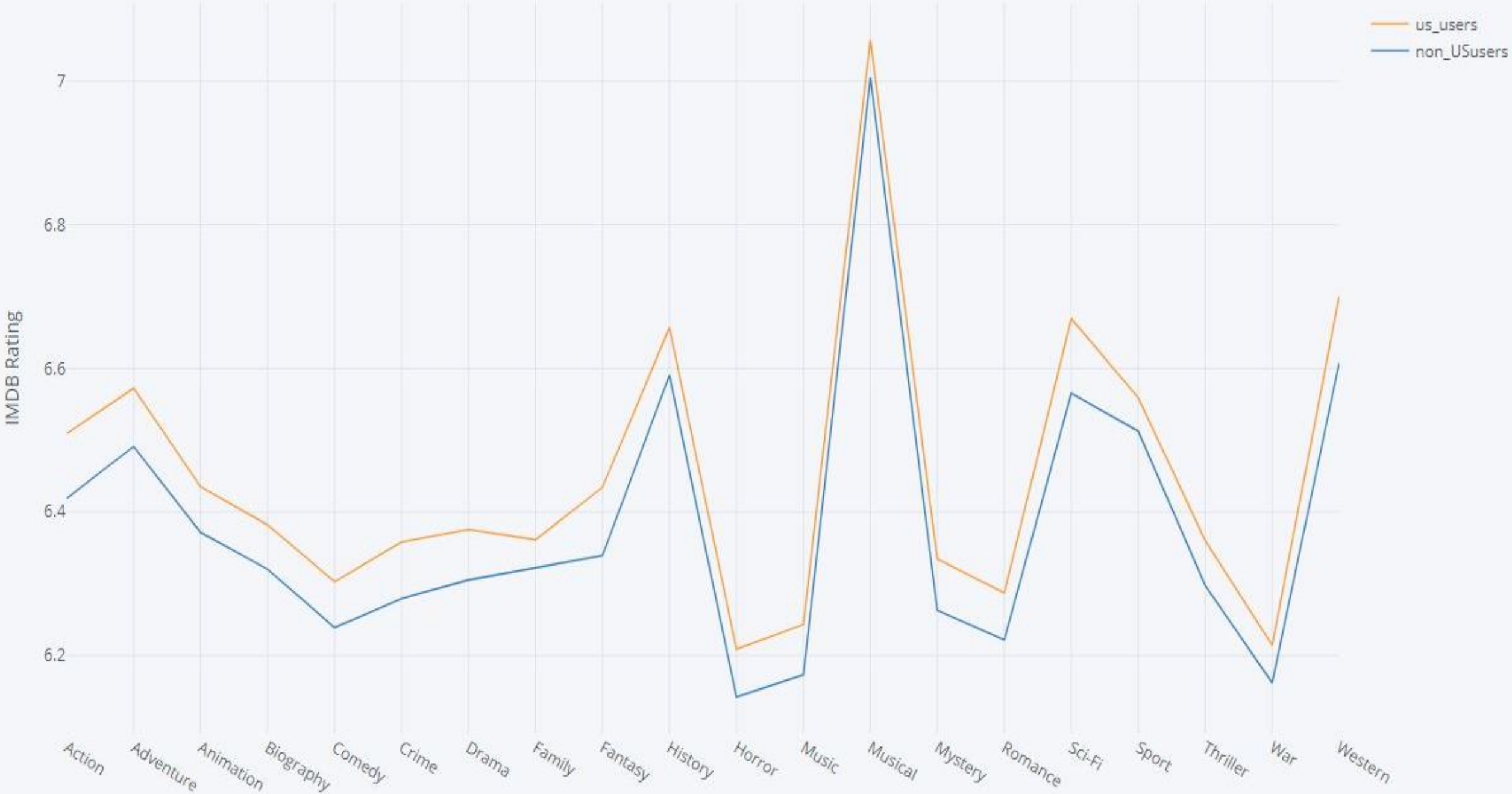
ANALYSIS & VISUALIZATION

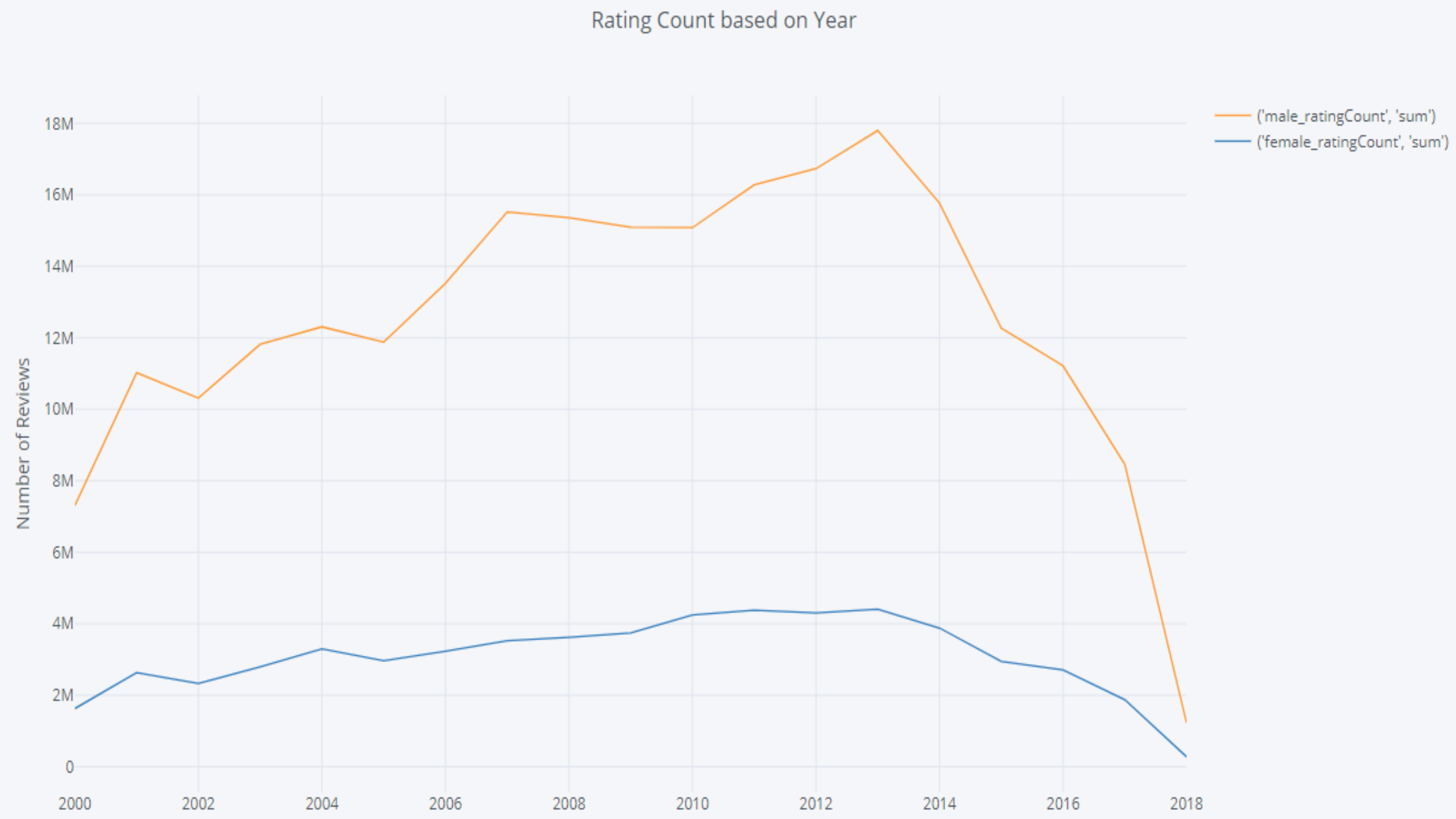



Rating Breakdown by Age/Gender
IMDB Rating
10
8
6
4
2
male_under18_rating
female_under18_rating
male_youngAdult_rating
female_youngAdult_rating
male_adult_rating
female_adult_rating
male_over45_rating
female_over45_rating
male_under18_rating
female_under18_rating
male_youngAdult_rating
female_youngAdult_rating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

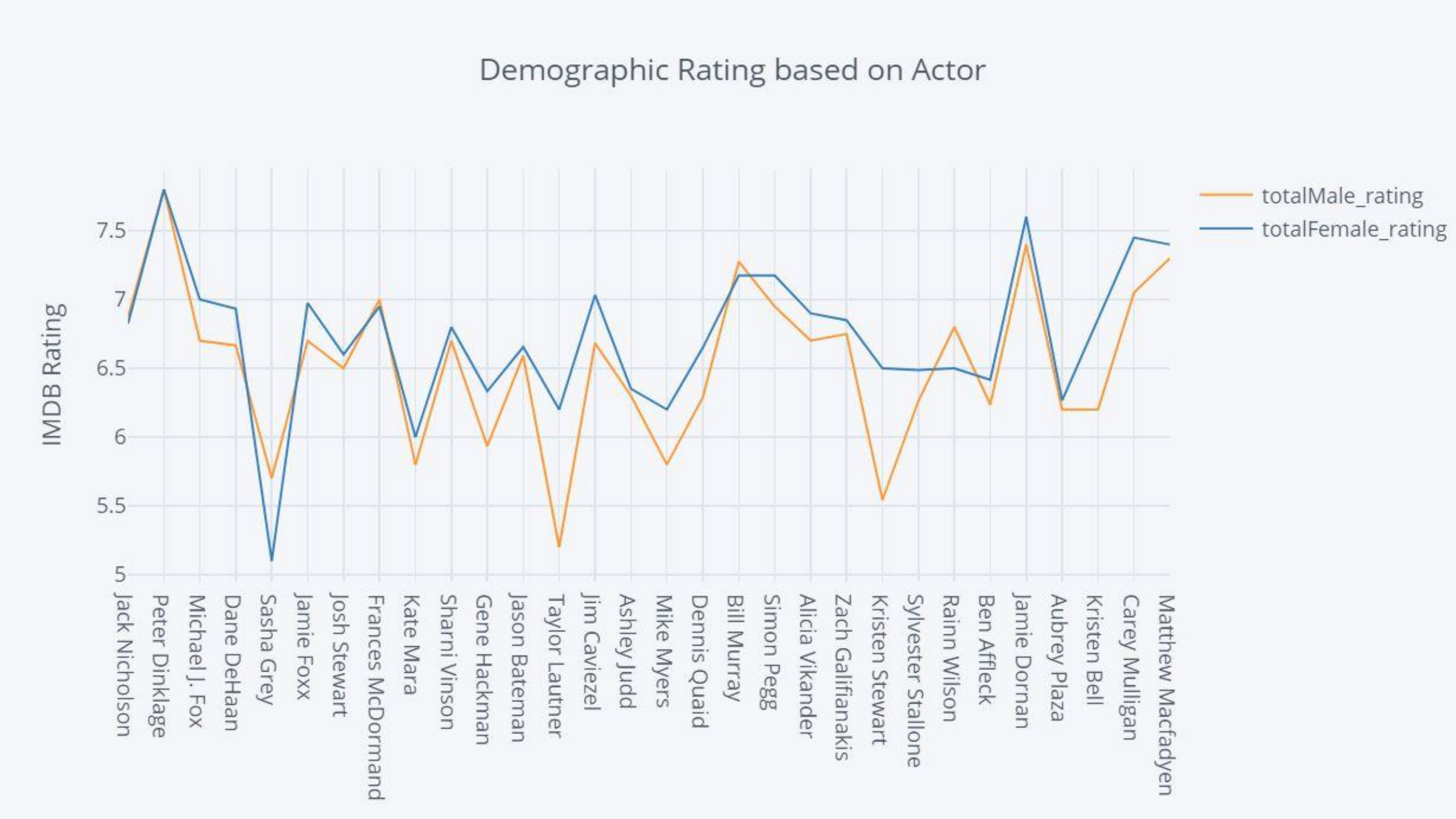

Rating Count based on Gender
male_ratingCount
female_ratingCount
Number of Ratings
100M
80M
60M
40M
20M
0
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western


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')


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