

IMDB.COM

WEB SCRAPING PROJECT

BY STEPHEN SHAFER

CONTENTS

OVERVIEW

WEB SCRAPING

DATA CLEANING


ANALYSIS/VISUALIZATION

[illegible]

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 demographics differ based on who the lead actor is?

WEB SCRAPING


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

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist (246)

Most Popular Feature Films Released 2000-01-01 to 2018-05-31 With At Least 10000 Votes and Country of Origin United States

1 to 50 of 3,414 titles | [Next »](#) View Mode: [Compact](#) | [Detailed](#)

Sort by: [Popularity](#) | [Alphabetical](#) | [IMDb Rating](#) | [Number of Votes](#) | [US Box Office](#) | [Runtime](#) | [Year](#) | [Release Date](#) | [Your Rating](#) | [Date of Your Rating](#)



1. Avengers: Infinity War (2018)


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

★ **8.8** ☆ [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: 313,508 | Gross: \$548.09M





2. Black Panther (2018)

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

★ **7.6** ☆ [8](#) 88 Metascore

T'Challa, the King of Wakanda, rises to the throne in the isolated, technologically

PG-13 | 2h 4min | Action, Adventure, Sci-Fi | 22 July 2011 (USA)



2:32 | [Trailer](#) 16 VIDEOS | 233 IMAGES

[Watch Now](#)
From \$2.99 (SD) on Prime Video

ON TV ON DISC ALL

Steve Rogers, a rejected military soldier transforms into Captain America after taking a dose of a "Super-Soldier serum". But being Captain America comes at a price as he attempts to take down a war monger and a terrorist organization.

Director: Joe Johnston
Writers: Christopher Markus (screenplay), Stephen McFeely (screenplay) | [2 more credits »](#)
Stars: Chris Evans, Hugo Weaving, Samuel L. Jackson | [See full cast & crew »](#)

66 **Metascore**
From metacritic.com

Reviews
728 user | [526 critic](#)

Popularity
57 (● 11)

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	
7.0 43		6.9 808		7.2 92,342	
				6.7 250,976	

CLEANING

Dropping NAs and cleaning Data

```
In [6]: # dropping 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

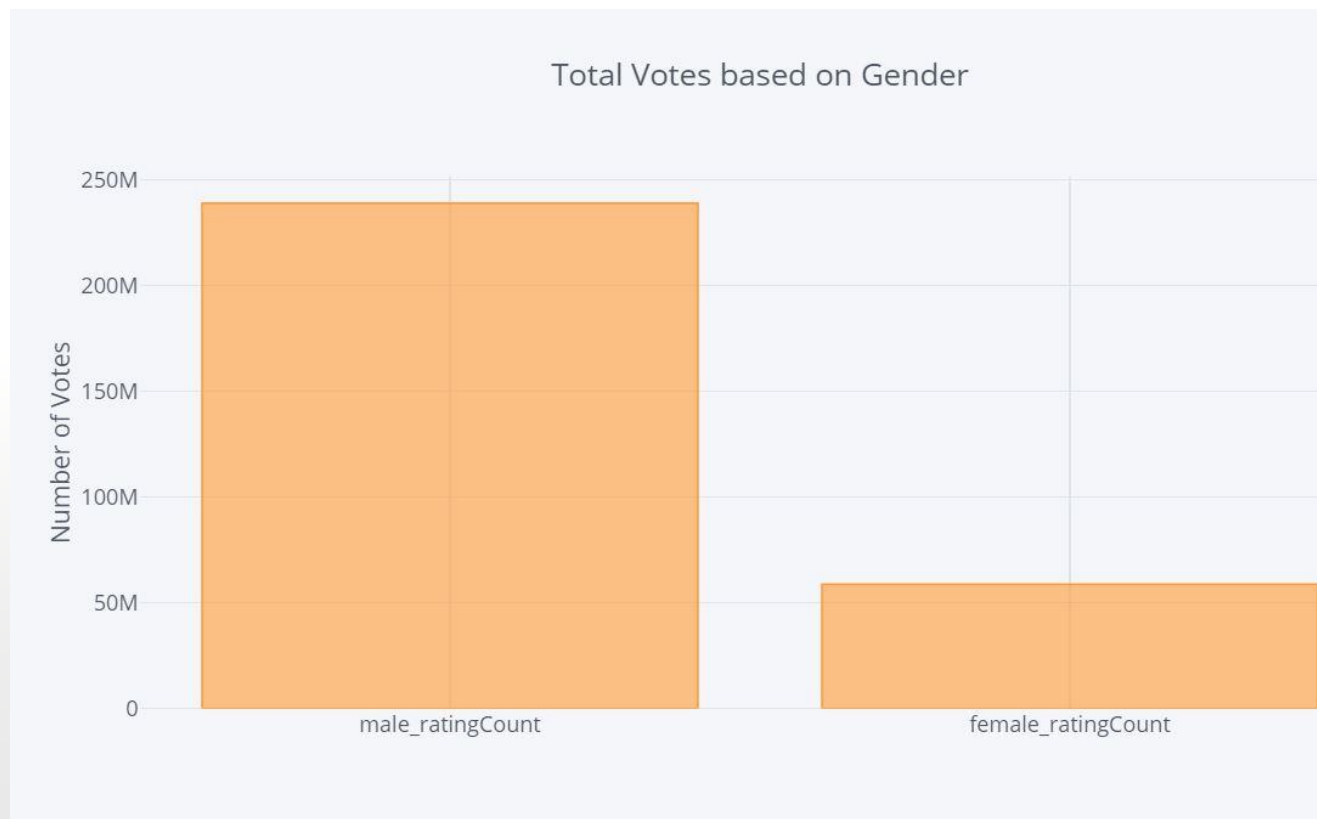
def split_it(year):
    return re.findall('(\d+ \w+ \d+)', year)

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'])
```

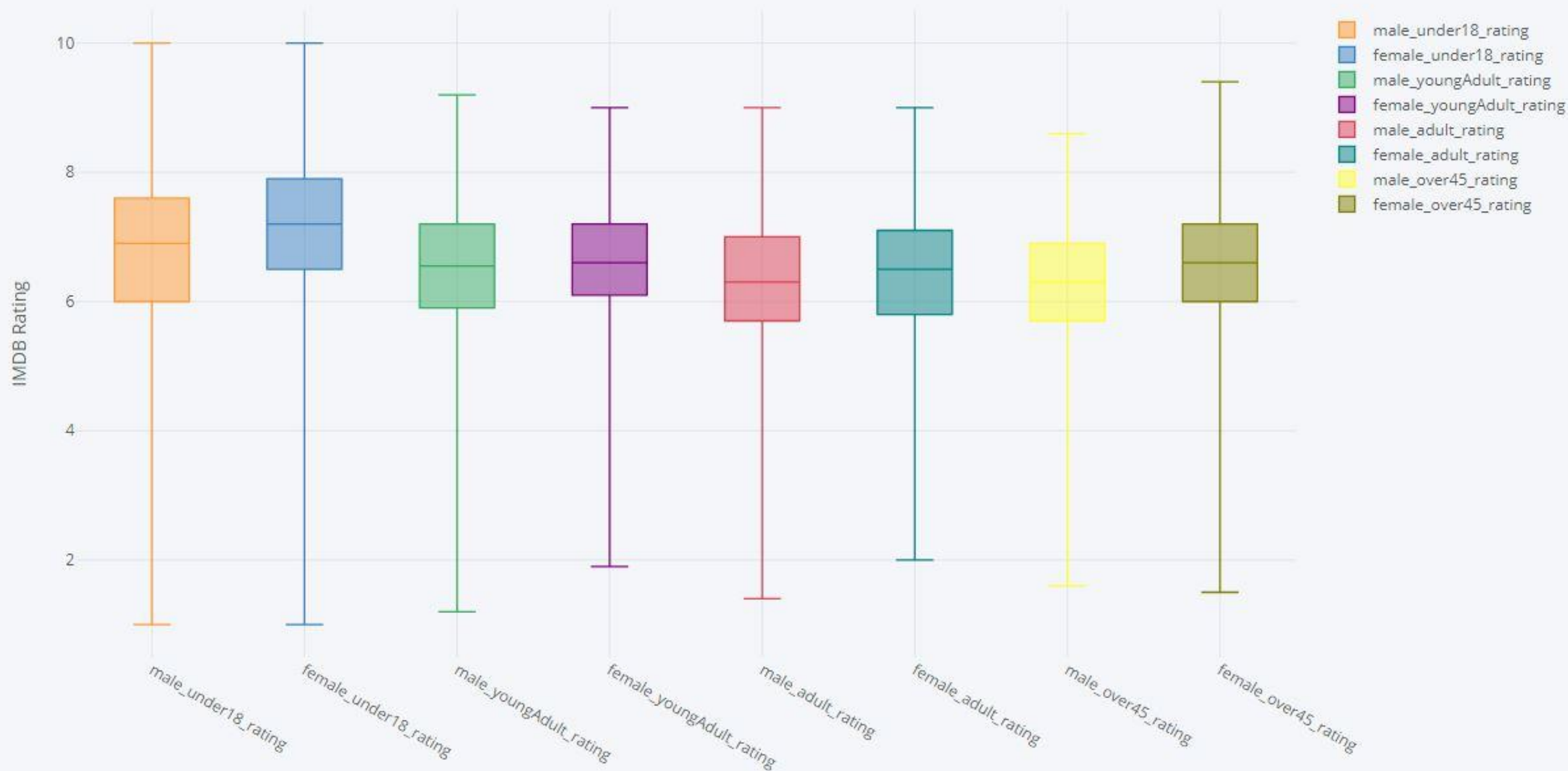
```
In [9]: # giving year its own column
imdb_test["release_year"] = imdb_test["release_date"].apply(lambda x : x.year)
imdb_test = imdb_test.dropna(subset=["release_year"])
imdb_test.release_year = imdb_test.release_year.astype(int)
```

```
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)
```

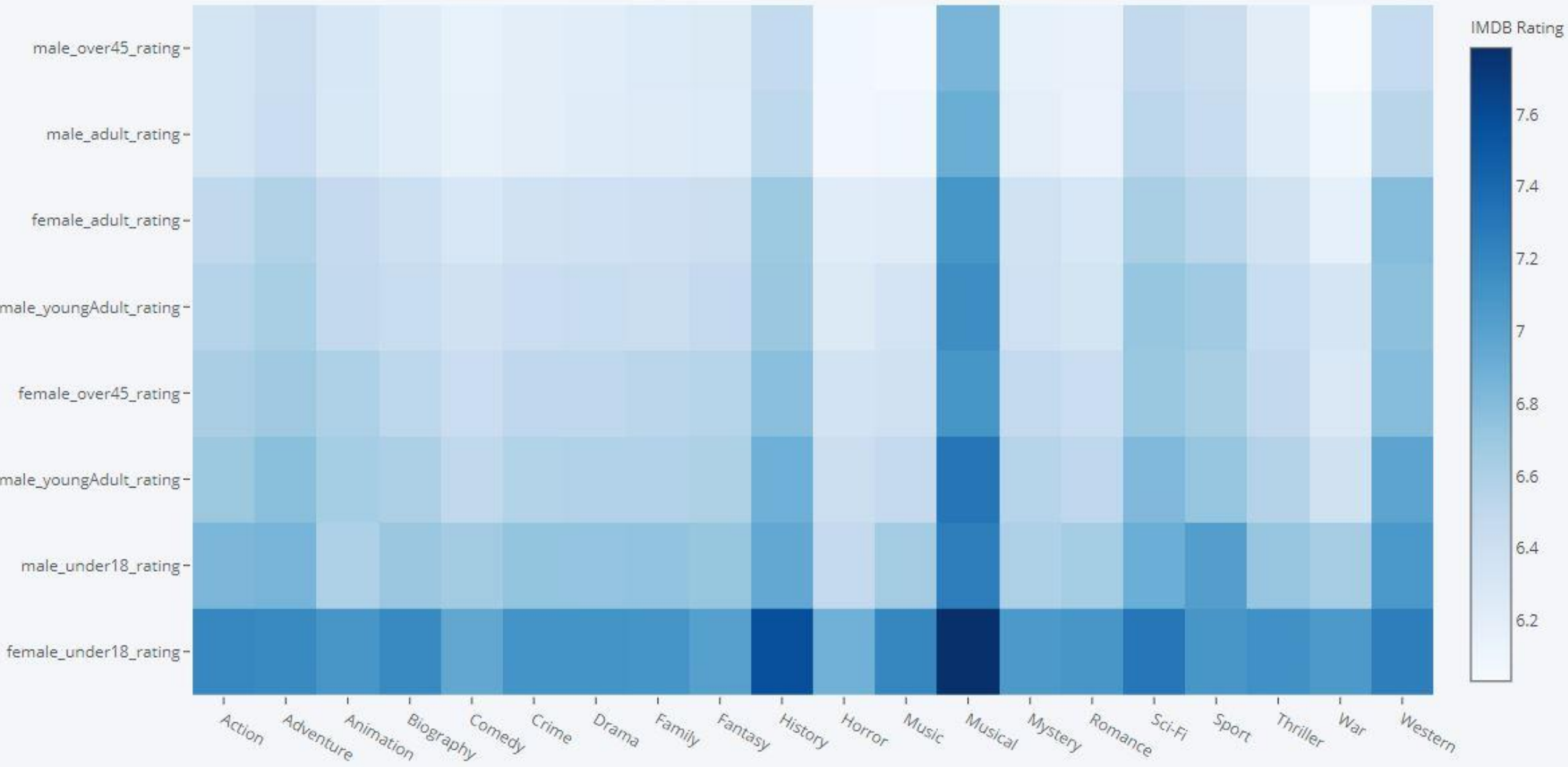
ANALYSIS & VISUALIZATION



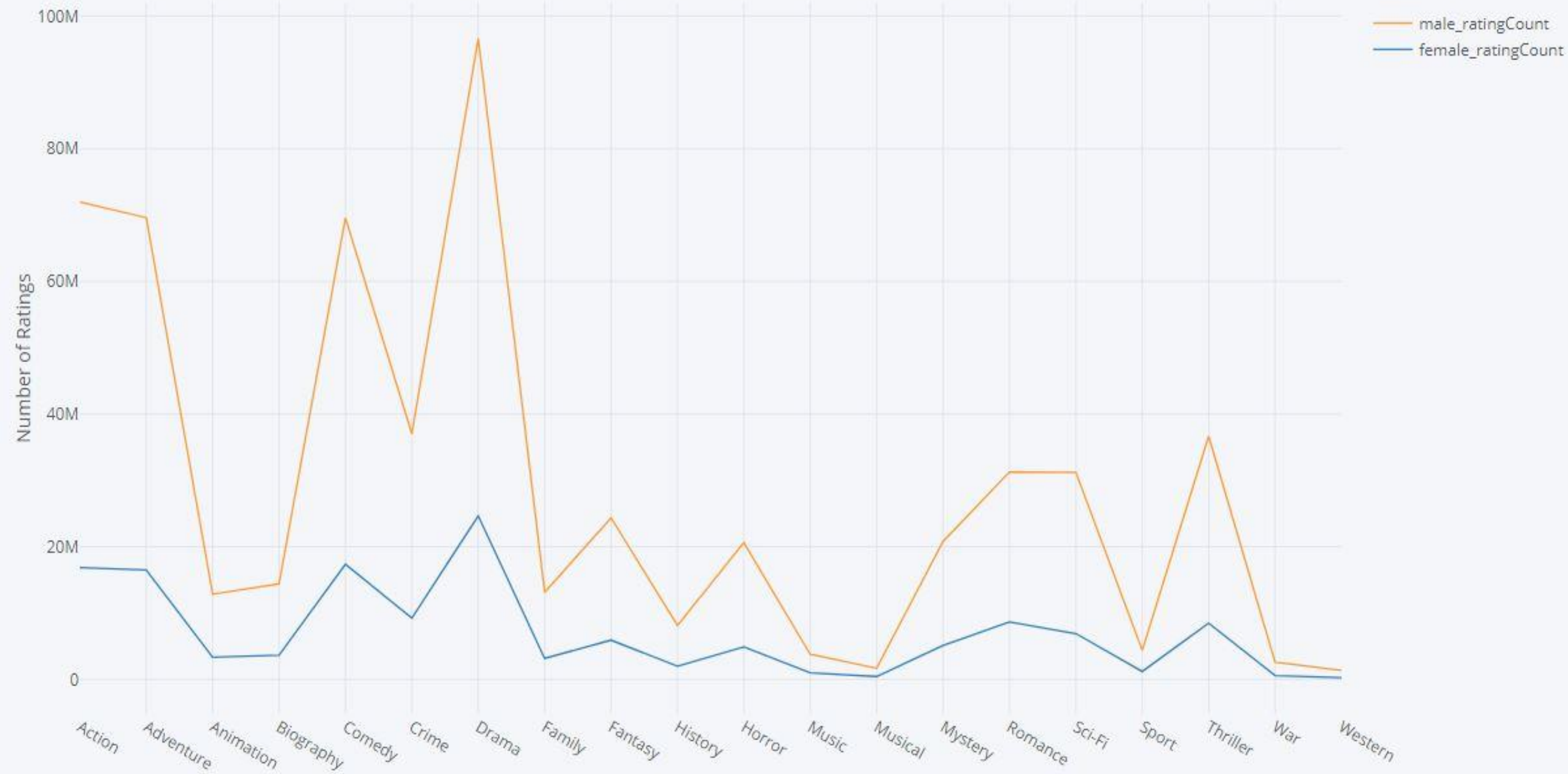
Rating Breakdown by Age/Gender

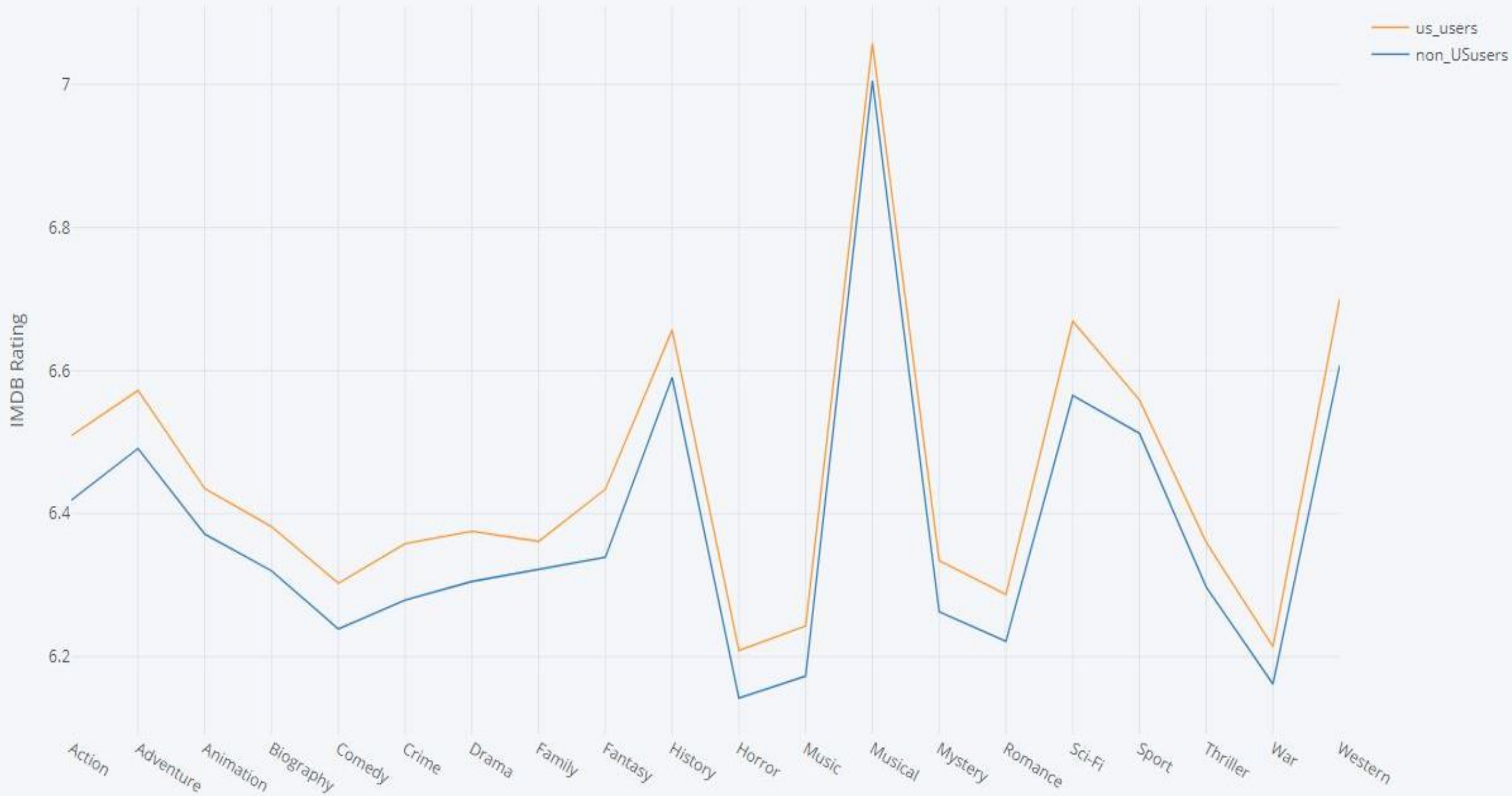


Demographic Rating by Genre

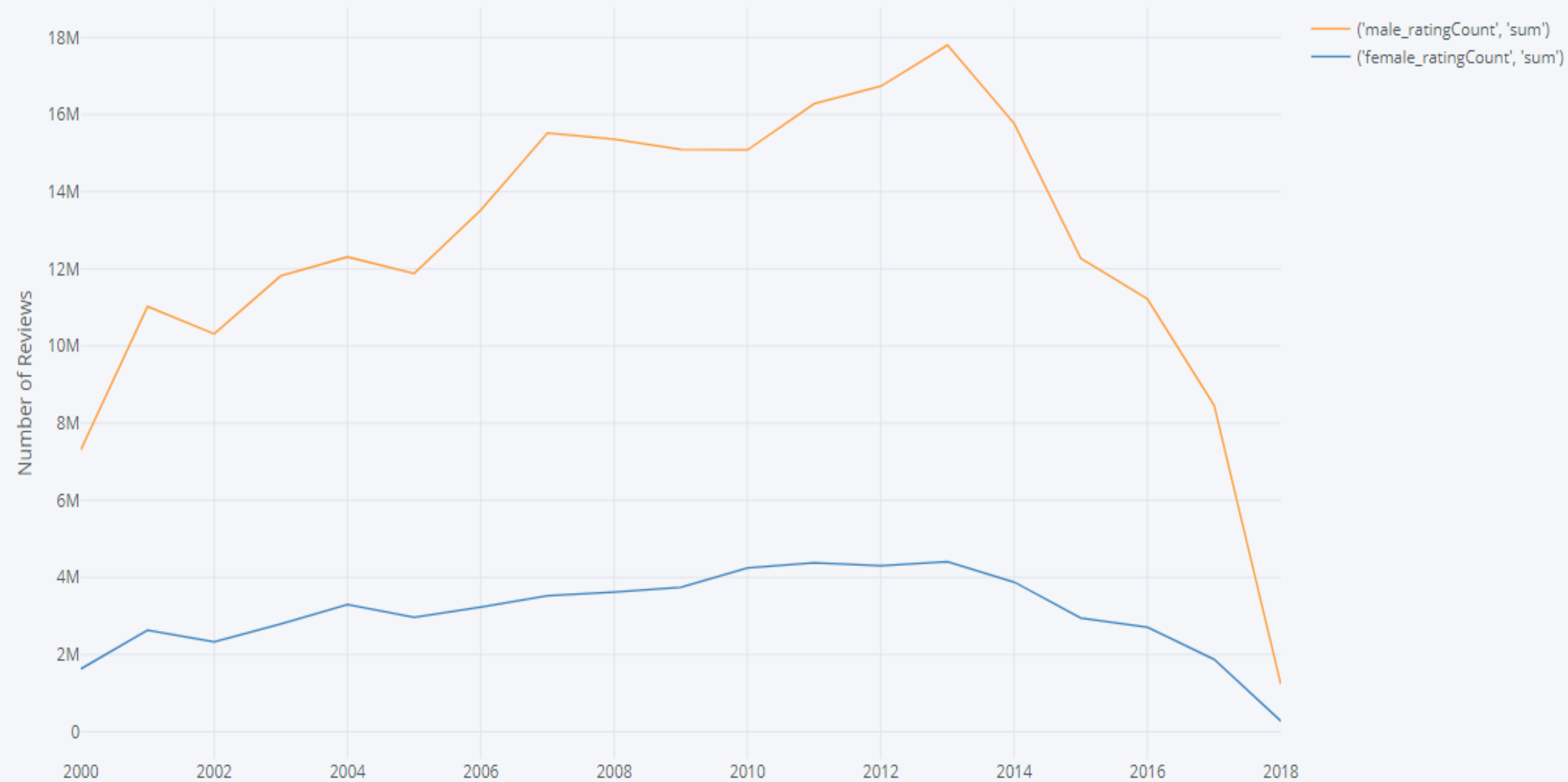


Rating Count based on Gender





Rating Count based on Year



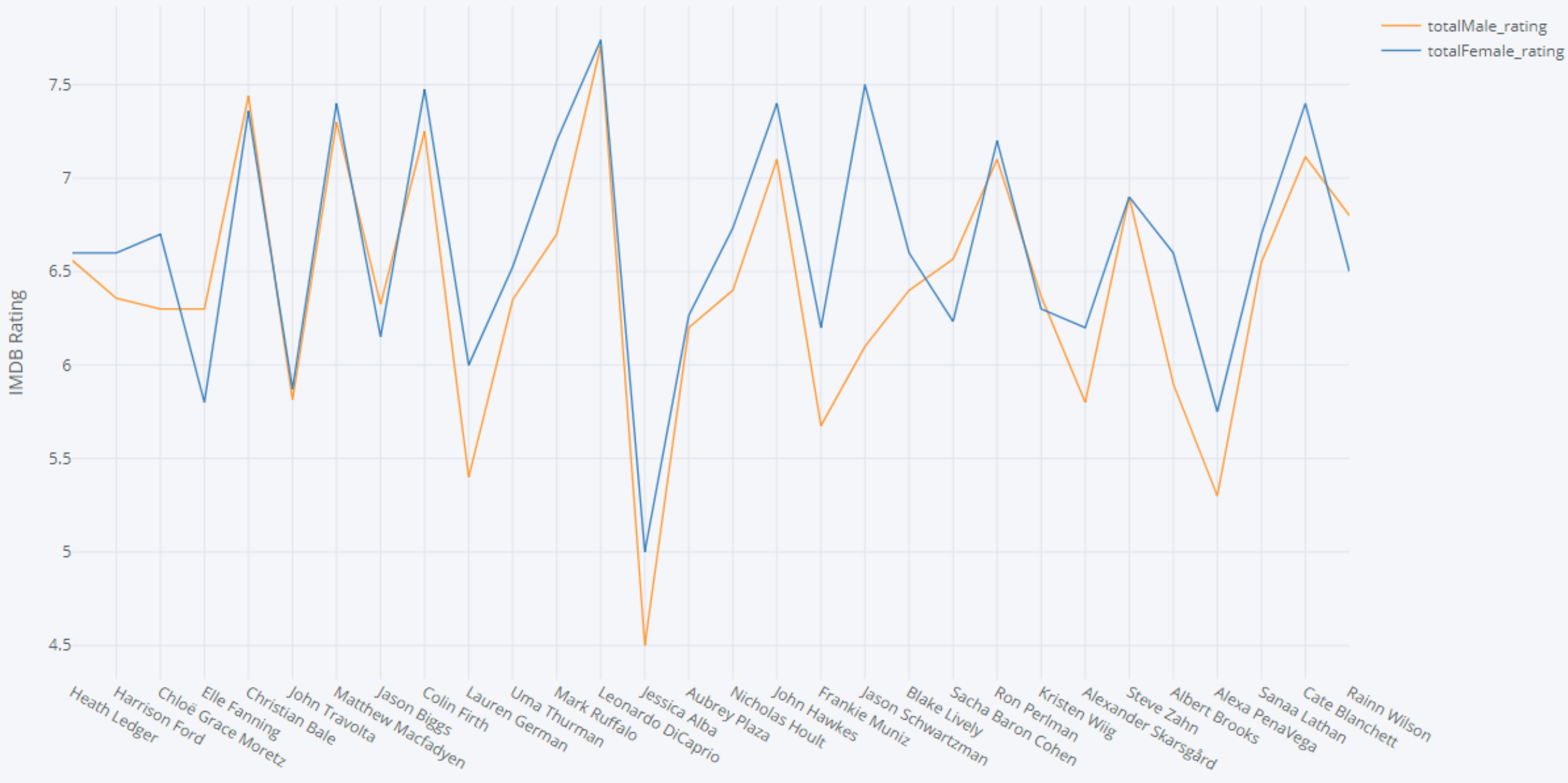
Mean Rating by Year



Mean Rating by Month



Ratings Based on Actor



Demographic Rating based on Actor



CONCLUSION

