Men like romantic movies too?

Zara Waheed

12/10/2021

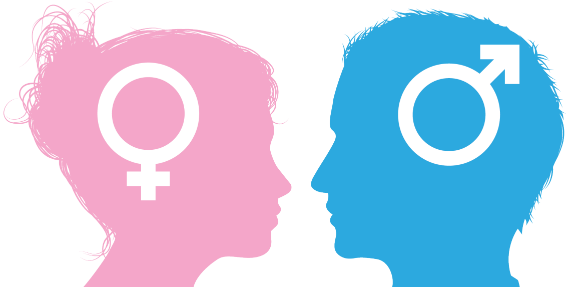


Figure 1: The difference between male and female Romance movie ratings

Abstract

Men are known to protest when they're forced to sit down for a new "chick flick" but can this be backed by numbers. In this report, I analyzed the relationship gender and reactions to romance films and compare these to other genres. The data set I used is a compilation of IMDb ratings, subset down to release dates between 1990 to 2020. I built a multilevel model to try to understand these behaviors in a little more depth. Even though the evidence points to the fact that women prefer these movies more than men, I went into detail on what could potentially affects these ratings.

Introduction

According to a report from 2016 on theatrical market statistics, the difference between romance movie tickets sold to both genders was only 5% the top romance movies that year, but the stereotype is that these movies are watched predominantly by women. I wanted to look into this and find statistical evidence, as well as explore the relationship and figure out what variables impacted it.

Do men like Romance movies more than women? Do men watch more of these movies than women? Is there any impact if the budget of these movies is higher? Do young audiences appreciate these movies more than older ones? Are critically acclaimed movies more preferred by them? Do these trends vary with different genres? I wanted to test the impact of these variables on the difference between men and women and I used a multilevel model to achieve that result and the relevant variables are as follows:

column names	explanation
title	The title of the movie
male_rating	Average rating from 1-10 by males
genre	The main category of movie
budget	Approximate total budget of the movie (in million $)
young	Proportion of voters under 30

column names	explanation
critics	Number of reviews the movie received from critics
reviews	Number of reviews the movie received from watchers
year	The year the movie came out
total_votes	Total votes received by the movie
duration	Total duration of the movie (in minutes)

Method

Data Cleaning and Processing

IMDb is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

My initial thought was to find a data set of the number of views of movies with gender demographics, or individual level survey data from watchers. However, since I was not able to find my ideal data set, I chose to base the study on two IMDb data sets that I sourced from Kaggle, one on information about the movies themselves and one on the ratings.

To begin exploring the data I needed to sort out the data sets and make assumptions about the subset I would be looking at. After importing the data into R, the main important steps I took are as follows:

1. Wrangled the movie details data set and the ratings data set into one.
2. Did some initial exploratory data analysis to explore the variables that would be useful to include.
3. Filtered the data down to the top ten genres.
4. Subset the data down to just USA, only keeping movies with release dates from 1990 to 2020 (I could not find data for 2021).
5. Created new variables by combining together ones that were not fit to add to the model on their own, but could bring potential value to the model.
6. Subset the data down to only the relevant columns.

Initially I had 85855 movies and 77 variables and after cleaning and sorting, I ended up with 2940 movies with full data, with 15 potential variables from which I will chose some relevant variables to work with.

EDA

Once the data had been cleaned and subset I decided to test the variables and began to explore relationships I found interesting. As an avid movie watcher, I tried to keep factors I, or people around me, would consider important in mind.

The initial question to answer was whether male ratings were higher than female ratings in the majority of romance movies. That can evidently be seen in the plot below.

After figuring out that the ratings were higher in females, I wanted to explore what factors affected this relationship. I also wanted to see how these results differed across different movie genres to make a stronger case about how the effect on Romance movie rating differed from other ones. The plot below shows the impact of young voter proportion on

```
## `geom_smooth()` using formula 'y ~ x'
```

I selected the variables I would be modelling, and after trying a few basic models to see which ones were significant, I realized I would need to re-scale my variables since most of the values were much larger than the continuous variable. After some experimentation, using z-scores for the continuous variables seemed like the best way to go.

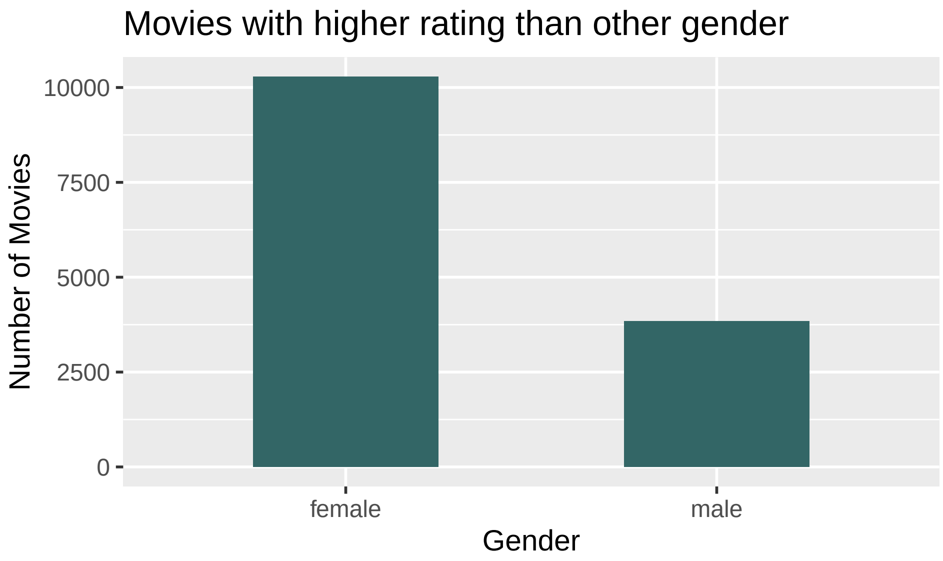


Figure 2: The number of movies where each gender has higher rating

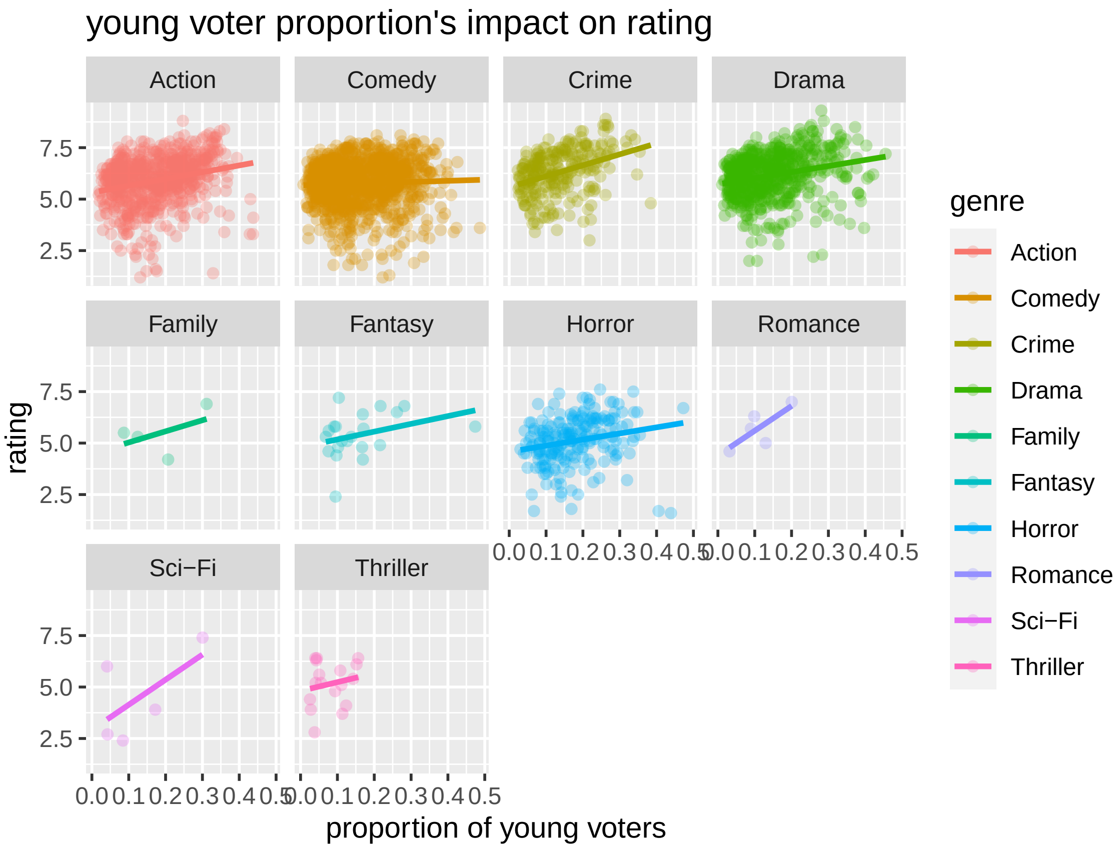


Figure 3: Including Young voter's proportion as a variable

budget and total_votes had a very large range and the plots below show how the rescaling impacted the relationship.

It was interesting to see that even though the number of total votes affected male ratings, a change in the number of reviews did not seem to have much of an impact. Another relationship that stood out to me was the proportion of young voters young. Without looking at the data, that had seemed like the most significant variable to include but even though the impact on male rating was be not too small, it had too large a variance to be considered seriously. For more detail on the selected variables, more EDA plots can be found in the appendix.

Model fitting and Validation

I fit a multilevel model with varying intercepts for different genres and varying slopes for budget, total_votes and critics as they seemed to impact the outcome the most. After some trial and error and some exploration, it seemed that the best model, all things considered was the one mentioned below:

```
model <- lmer(male_rating ~ budget_z + total_votes_z + critics_z + duration_z + year_z
              + reviews_z +  young + (1 + budget_z|genre) + (1 + total_votes_z|genre) +
                (1 + critics_z|genre), data = imdb_clean)
```

To see how well the fit of the model was, we can look at the table below where it seems that most of the variables are significant at alpha = 0.05 level.

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>
(Intercept)	5.64323	0.13578	6.84844	41.561	1.73e-09 ***
budget_z	-0.30505	0.04764	5.95434	-6.403	0.000705 ***
total_votes_z	0.34959	0.07128	5.63021	4.904	0.003222 **
critics_z	0.51316	0.03997	7.37739	12.839	2.59e-06 ***
duration_z	0.31686	0.01976	2019.23487	16.035	< 2e-16 ***
year_z	-0.32535	0.02037	2602.78085	-15.968	< 2e-16 ***
reviews_z	-0.04101	0.02603	2266.44536	-1.576	0.115230
young	0.26939	0.25274	2509.45580	1.066	0.286586

In this model, the budget, total votes and critic reviews coefficients differ for each genre since it is a varying slope model. The coefficients of each variable seem to be relatively small considering they are all z scores, but that is because the variance between them is large and most values deviate from the mean by huge values. For example, the mean of budget is $26751666 but a lot of movie budgets are three or four times that size or even larger and similarly there are many that are way smaller. Each of the z-score coefficients can be interpreted as the increase (if + sign of coefficient) or decrease (if negative sign of coefficient) of that amount with the increase of 1 standard deviation. e.g. With each increase of one standard deviation from the mean of number of critic reviews, the male rating for that movie will increase by 0.51.

The plots below give us an idea of the fit of the model. The Q-Q plot does not align fully with the line but the fit is not bad. The leverage plot shows that there may be a leverage point.

Conclusion

Limitations

There were many limitations that I had to keep in mind when doing this analysis. The main one , IMDB might be a well known voting platform but it is not the only popular one. Rotten tomatoes and Metacritic are also very reliable sources to checking ratings of movies and many people might just be voting on those instead.

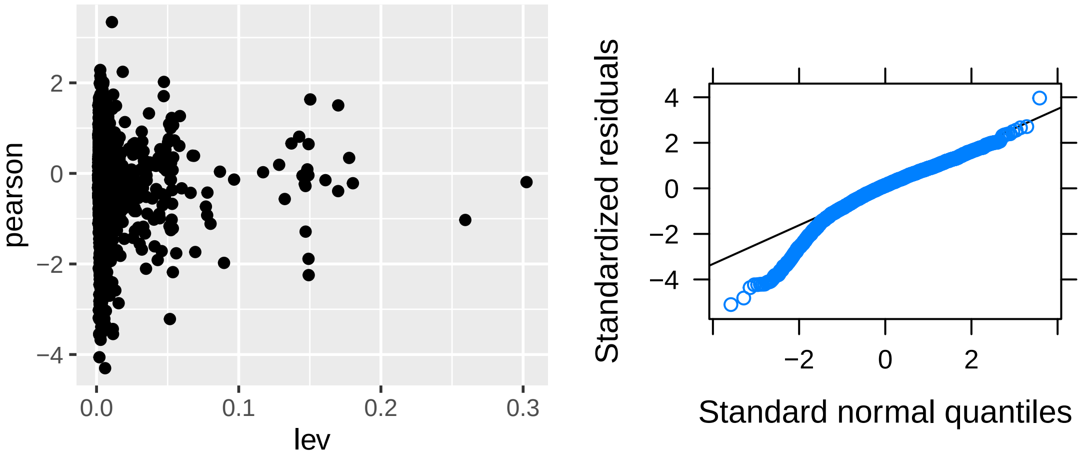


Figure 4: Q-Q plot.

Some of the limitations of this model come from the variables themselves. The three variables that lead to the most concern would be `total_votes`, `young` and `reviews`.

`reviews` was a far fetched variable to use because even though it seems to show a relationship with the dependent variable, more views do not necessarily signify a movie's popularity and it is difficult to interpret the coefficient due to this issue. A movie could very likely have more reviews because viewers have strong negative feelings towards it.

Additionally, I had some concerns when using of `young` and `total_votes` together. Even though they did not come from the same values, they were both related to votes and could be correlated. However, they were both variables that I was interested in observing and I feel brought value to the analysis. `total_votes` showed more significance in the model before I added `young`.

Another key restriction was the population estimates - Even though imdb is a popular film rating website, most movie watchers, such as myself, do not cast votes at all, and there is no accountability in this analysis for the gender demographics of that part of the population. Based on the stereotype that it is a taboo for males to like Romance films, the sample being used is most likely an overestimate, especially for males, despite trying to account for it by keeping number of votes as a variable. So it's difficult to make solid conclusions from this analysis.

Looking at the fixed effects not all of the predictors used are Even though all the predictors don't have p-values less than 0.05, they were an interesting addition to the model. The estimates seem to be reasonable enough. The intercept is close to the midpoint of the scale and coefficients are generally small because of the huge variance from the mean especially for `budget` and `total_votes`. However the EDA plots seemed to show a more significant relationship between male rating and that was not mirrored in the model.

Keeping all the above in mind, in the future it may be beneficial to attempt a model without `young` and `reviews`, as well as using movies with similar or equal number of votes and popularity for a more balanced analysis.

Citations

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv

https://www.kaggle.com/unanimad/disney-plus-shows

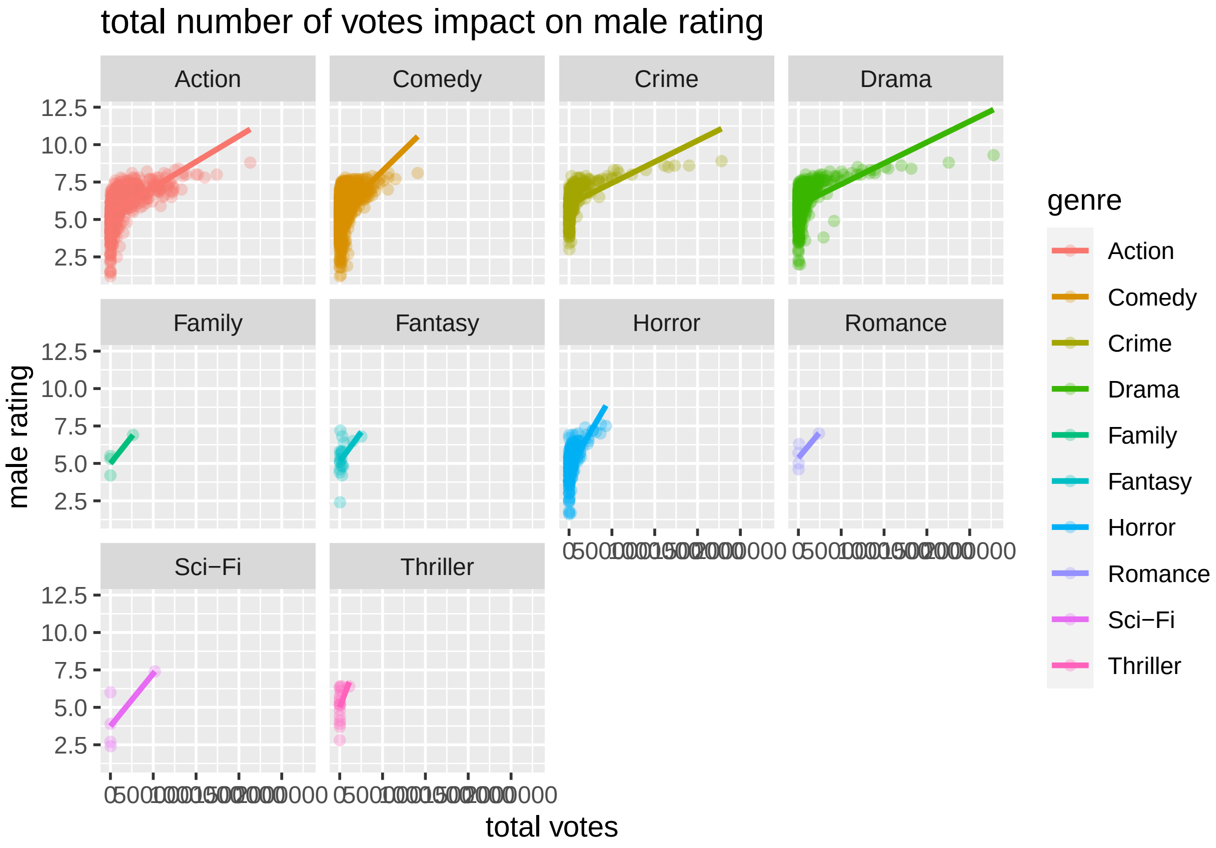
https://www.motionpictures.org/wp-content/uploads/2018/03/MPAA-Theatrical-Market-Statistics-2016_Final-1.pdf

Appendix

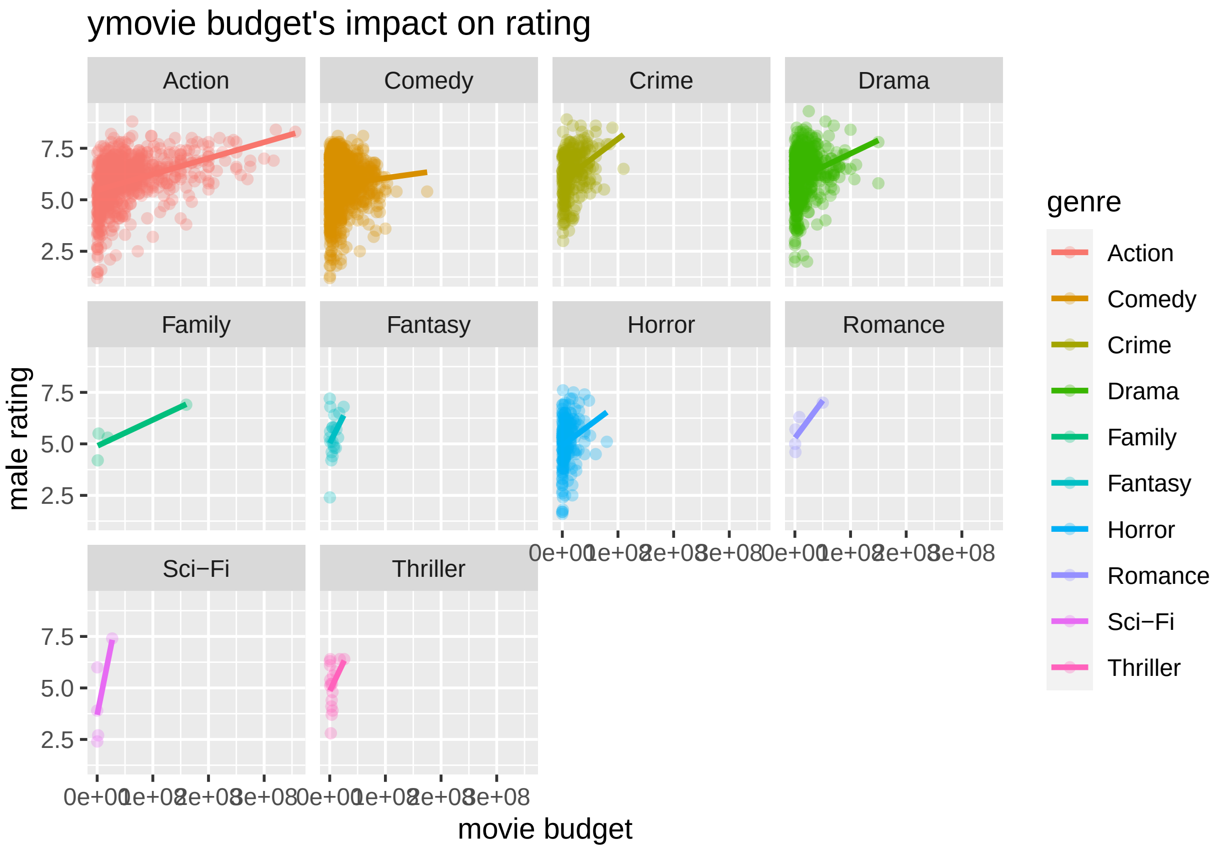
EDA and details of fit

Other variables impact on rating plots

```
## `geom_smooth()` using formula 'y ~ x'
```



```
## `geom_smooth()` using formula 'y ~ x'
```



The below model fit slightly better and had less variables but I didn't use that as the final one because I wanted to see the effects of those additional variables.

```
## $genre
##          (Intercept)   budget_z total_votes_z critics_z duration_z     year_z
## Action      5.735363 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Comedy      5.902453 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Crime       6.075833 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Drama       6.083662 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Family      5.691180 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Fantasy     5.541250 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Horror      5.295520 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Romance     5.705888 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Sci-Fi      5.161613 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
## Thriller    5.681139 -0.2817958     0.2000342 0.5437912  0.3046249 -0.3135926
##
## attr(,"class")
## [1] "coef.mer"
```

Model Validation

Complete Results

Random effects of model

Fixed effects of model

Coefficients of model