Stat 240 Final Draft

2023-04-27

The film and television industry adds $186 billion dollars to the U.S. GDP alone. [1] Flourishing film industries across the world add economic value to their countries and joy to the lives of their viewers. But when you want to watch something, how do you choose what to watch? Ratings are an invaluable tool for deciding what movie you will give an hour and a half of your life to. We were curious about film ratings in different countries around the world, and wanted to know if average movie ratings differed across regions (where the evaluation of average rating would be separated by genre.) *For this project, we decided to focus on the United States and India, as they were the two countries with the highest sample sizes, and we thought the comparison of Hollywood and Bollywood would be interesting. Do movie ratings differ between the U.S. and India? More specifically, are there differences in movie ratings within genres?*

We will demonstrate that when considering all genres together, the average movie ratings do not differ much between India and the U.S. However, we will also show that there are significant differences in the proportion of “not bad” or better ratings (≥ 6) in all top 10 genres (except in Biography and Documentary.)

To explore our thesis, we will analyze datasets from IMDb–the Internet Movie Database–to explore the tastes of India and the U.S. in movie genres and uncover patterns and trends within the worldwide entertainment industry.[2] We chose IMDb’s datasets owing to its prominence as an online platform which provides legitimate information about movies and the entertainment industry. Our data is directly sourced from IMDb’s website where we have downloaded three TSV files that we have merged into one CSV file [3]. The datasets we will be working with contain crucial information such as genres, regions, title identifiers for movies, number of votes, and average ratings in which users can rate movies on a scale from one to ten based on their quality. These select variables provide us insight into audience preferences and the performance of movies across different regions and/or cultures.

To understand our thesis statement, we will investigate the background information surrounding movie preferences and how these datasets relate to our overall analysis. IMDb scores are rated on a 1 to 10 scale, with 1 regarded as “Do Not Want” and 10 as “Masterpiece”. [4] Additionally, the scores are a weighted average between all scores of a movie, which helps to reduce the impact of outlier ratings. The data directly relates to our question as we are exploring movie ratings between the United States and India. The IMDb dataset is crucial in getting each movie’s ratings and is therefore relevant.

Although IMDb is known as the “world’s most popular and authoritative movie source for media,” it is possible that user input data might be biased or skewed.[5] Given the large size of our data files, we will filter them down by region, year, etc., which may affect the interpretation of our results. Another factor that may impact our interpretations is the unequal sample size difference between the United States and India. According to our data, after 2000, the United States had more distinct movie releases compared to India. Each rating for an Indian movie carries more weight for its respective genre. There may also be an issue with the selection of Indian movies because not all Indian movies are reported on IMDb. Selection bias would occur if only the best Indian movies are reported on IMDb.

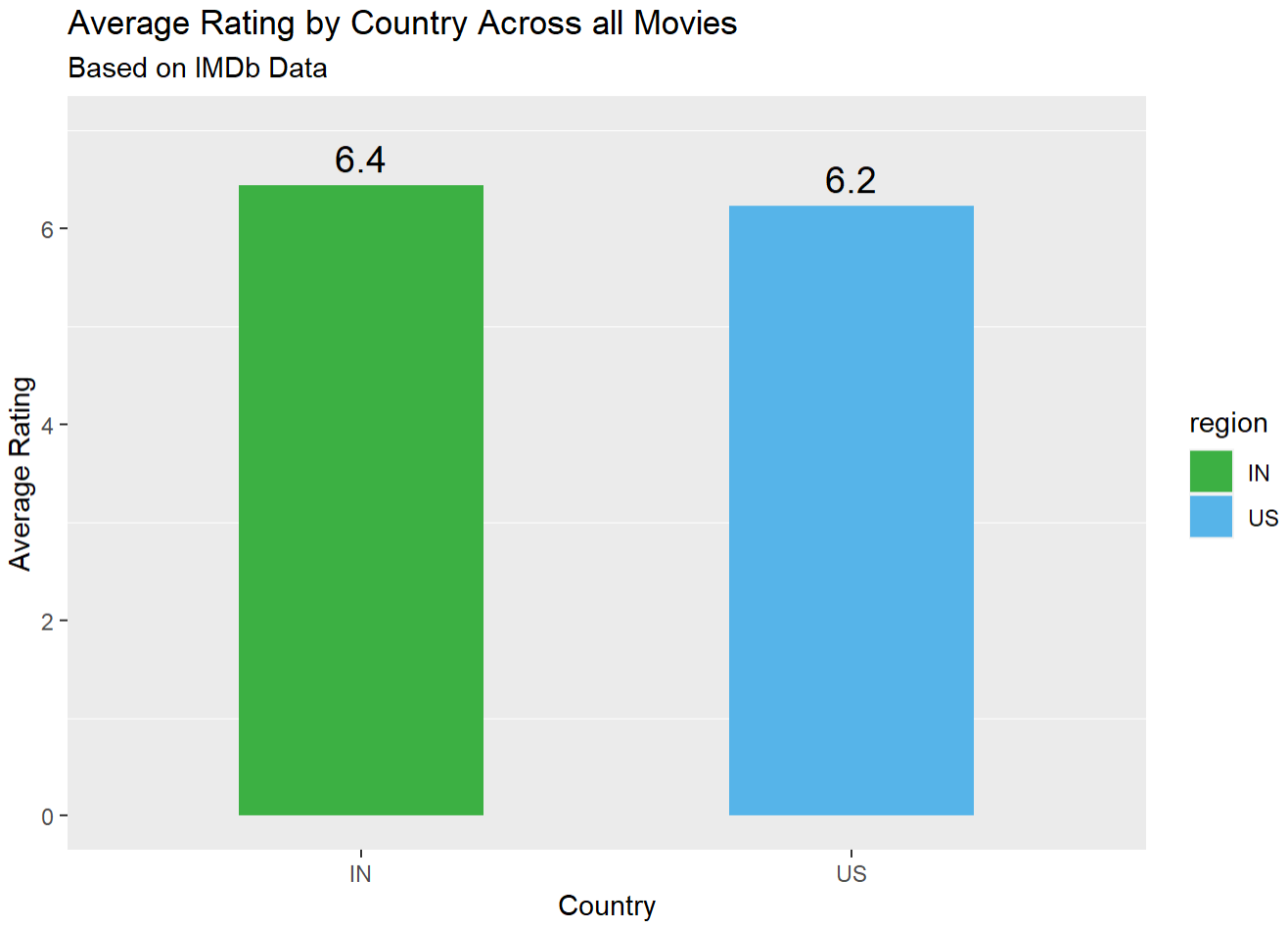
For the rest of our project, we intend to utilize the Lubridate and Tidyverse packages to visualize and analyze our datasets. We will visualize bar graphs to demonstrate how ratings differ between numerous genres and also density plots to showcase the likelihood of certain ratings for movies. We will use this analysis to see whether there are statistically significant differences in average movie ratings between the U.S. and India. As a preliminary step to this analysis of differences, we will visualize the mean, median, and spread of ratings for each country and genre. This first step will help us understand our data before we begin statistical calculations. We would like to do this analysis for all genres grouped together and for separated genres. After this, we will visualize the proportion of movies produced by genre in each country to determine any national specialties. In addition, we will conduct hypothesis tests to determine if the average ratings differ between countries by comparing means of ratings of movies.

Data:

```
## # A tibble: 11 × 3
##    variable_name   description                                          type
##    <chr>           <chr>                                                <chr>
##  1 tconst          alphanumeric unique identifier of the title          str
##  2 ordering        a number to uniquely identify rows for a given title ID int
##  3 title           the localized title                                  str
##  4 region          the region for this version of the title             str
##  5 language        the language of the title                            str
##  6 types           An attribute for this alt title with 'imdbDisplay'   arr
##  7 attributes      Additional terms to describe this alternative title  arr
##  8 isOriginalTitle 0: not original title; 1: original title             bool
##  9 averageRating   weighted average of all the individual user ratings  int
## 10 numVotes        number of votes the title has received               int
## 11 genres          includes the primary genre associated with the title arr
```

Mean of Average Rating Across All Genres

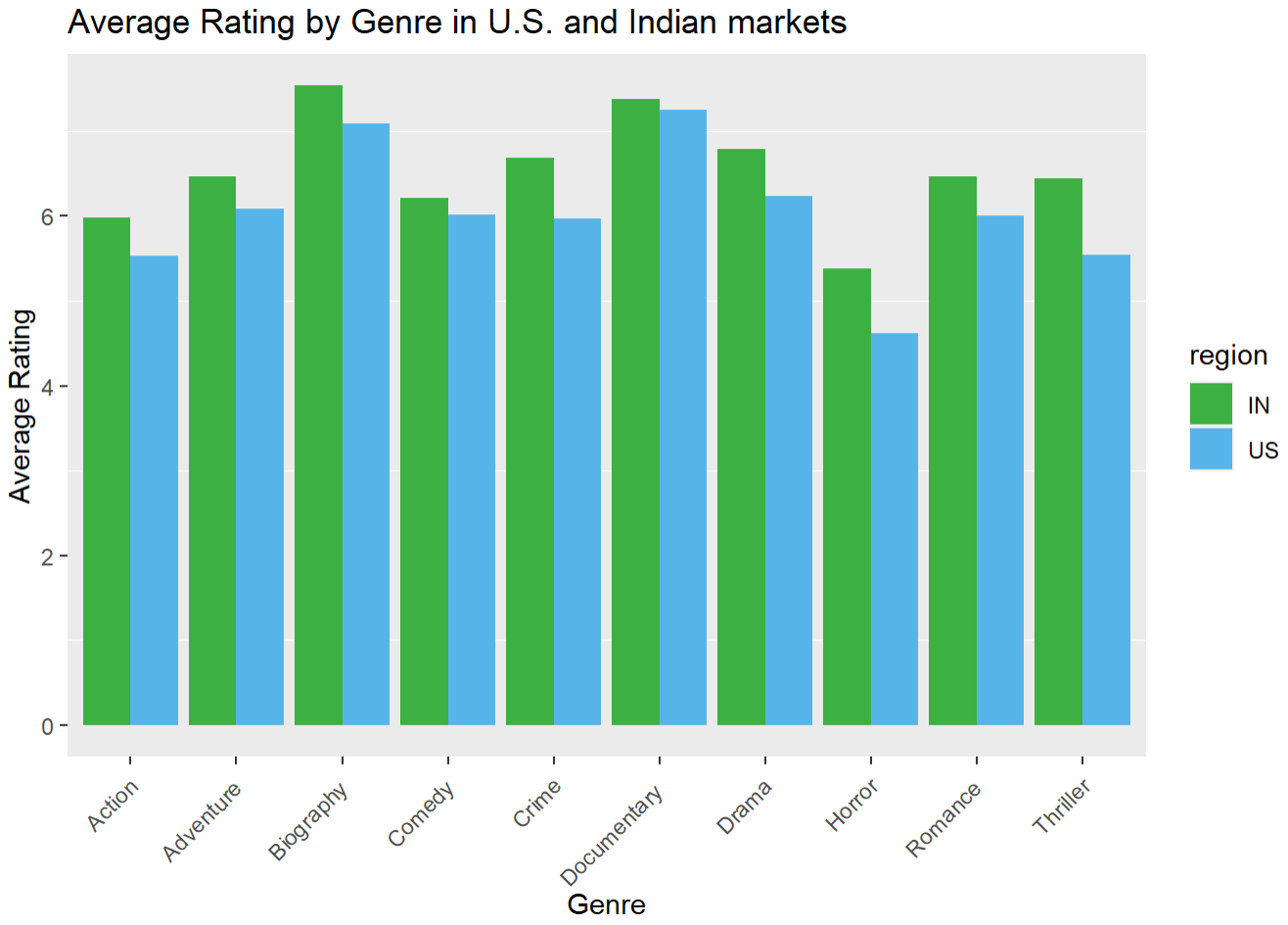
This graph simply shows the average ratings across the ten most popular genres common to both countries.



From the graph, we can see the baseline rating for India is slightly higher than the U.S., which is important to keep in mind while analyzing the rest of the data. We also wondered whether or not the average ratings in India and the US would differ between genres. For that reason, we created a graph with the mean average rating separated by genre and country.

Grouped Bar Graph of Mean Average Ratings in Different Genres

This graph is similar to the previous one, but broken down by genre, giving a more detailed look into the preliminary differences between the two countries.

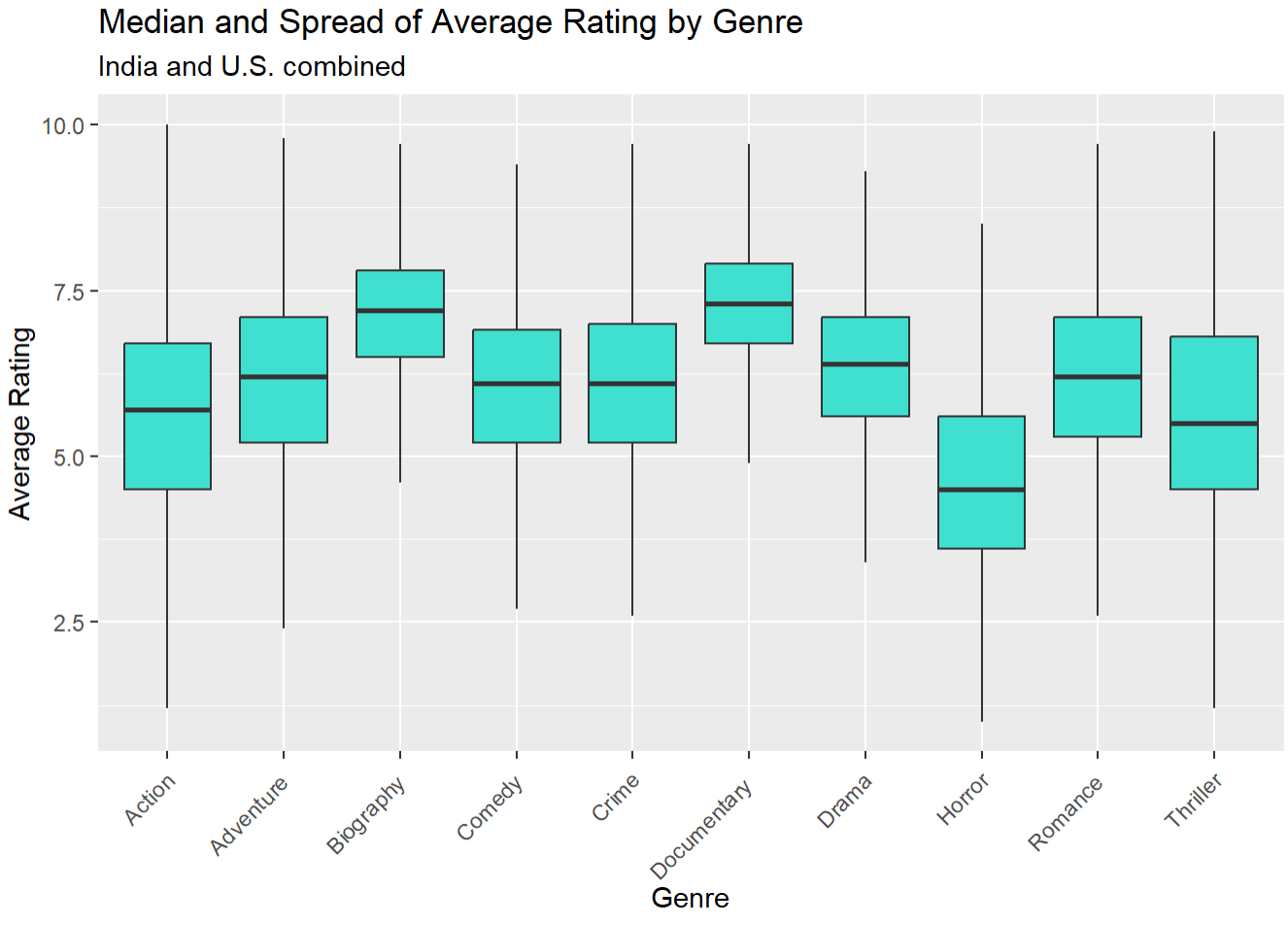


Note that the average score for India is consistently slightly higher. Generally, the mean average movie rating for a genre is similar in both India and the US. We noticed, however, that some genres have a greater difference in average movie ratings between the two countries. In particular, the horror and thriller genres have a larger gap, which we will look into further in our box-and-whisker and density plots.

Box and Whisker Plots of Ratings by Genre

With Both Regions Combined

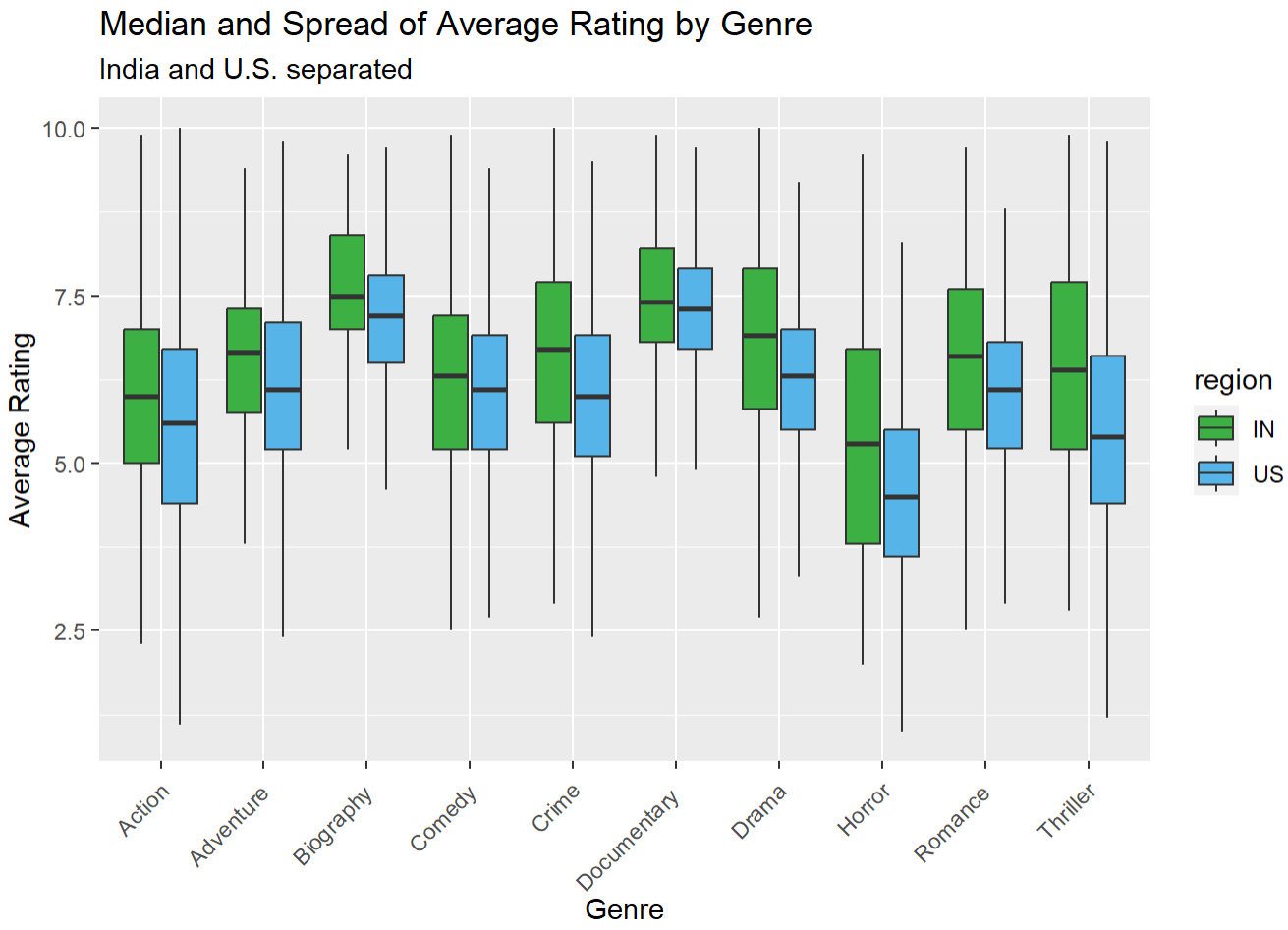
This graph considers the median and spread of ratings in various genres for both India and U.S. combined. We realized that it is important to understand not only a measure of central location for the average rating but also how spread out the data is. We begin with a graph similar to the first mean graph. We consider the median, first quartile, and third quartile for each movie genre, where the ratings are considered for both countries combined.



The median rating and the spread of ratings can vary widely between genres. This suggests that both countries think some genres are better than others. The opinions of a genre can very widely. Some genres have very high AND very low ratings (e.g. Thriller) while some genres only vary from mediocre to very good (e.g. Biography). While this graph is informative about the median ratings and the spread of ratings in each genre, it tells us nothing about international differences. Given that the mean ratings in each genre vary between India and the US, we wondered what the median and spread of each genre in both countries was.

With Both Regions Separated

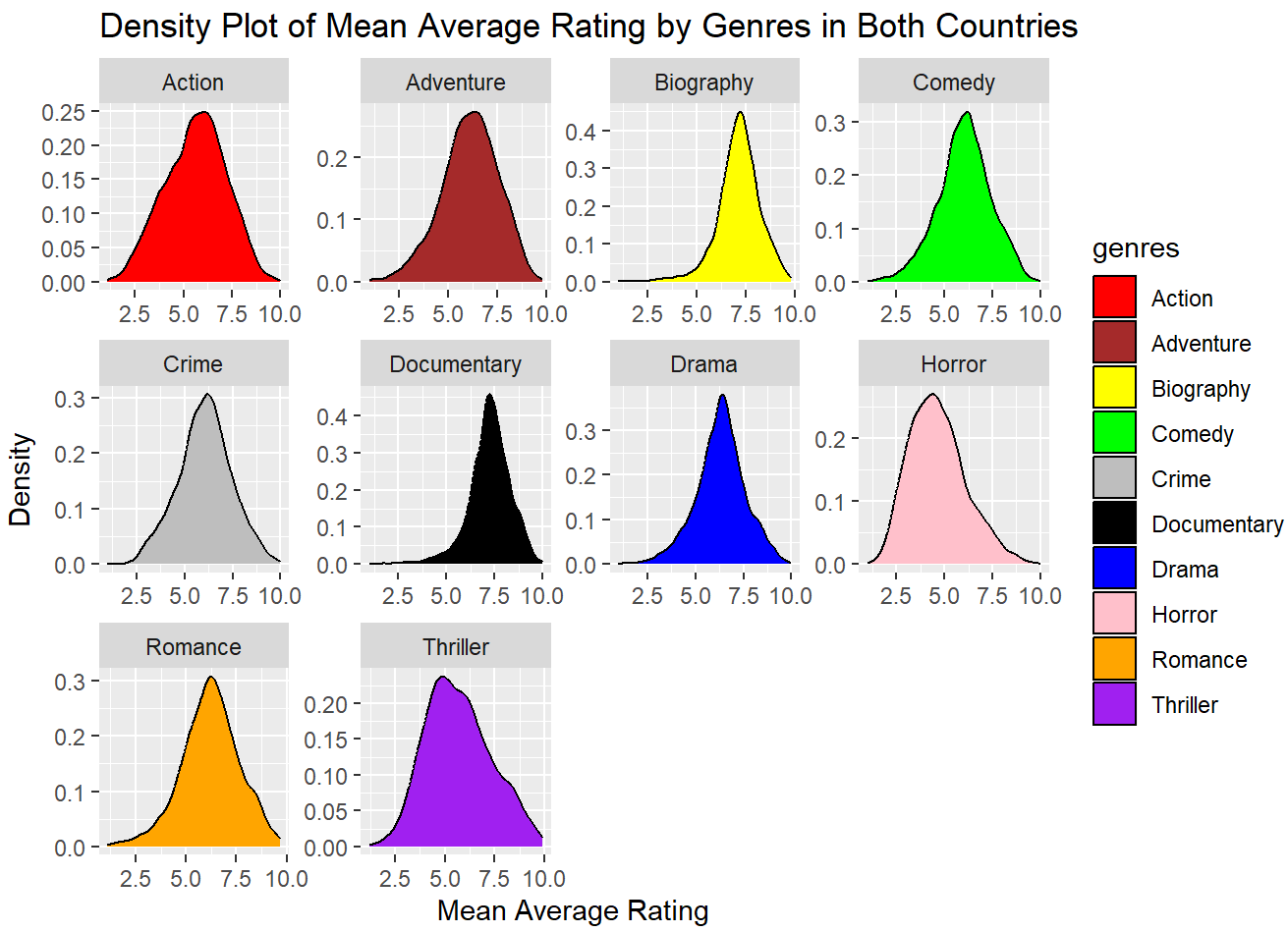
This graph considers whether the median and spread of ratings in a genre vary between India and the US. We generated a boxplot with the same genres as above, but used two box and whisker plots for each genre (green for India, blue for the US.)



The US generally rates their movies slightly lower than India rates their movies. As we mentioned in the introduction, this may be due in part to selection bias. IMDb does not contain all Indian movies, and the movies posted to IMDb may favor the most acclaimed movies. If selection bias is present, it would make sense to see Indian movie ratings that are systematically higher than American movie ratings. Though there are some differences in Indian and American ratings in a genre, they are generally quite similar. The medians are close to one another and the spreads are similar in magnitude. To visualize the difference in genre ratings between India and the US in another way, we decided to use density plots.

Density Graph for Average Rating by Genre in Both Countries

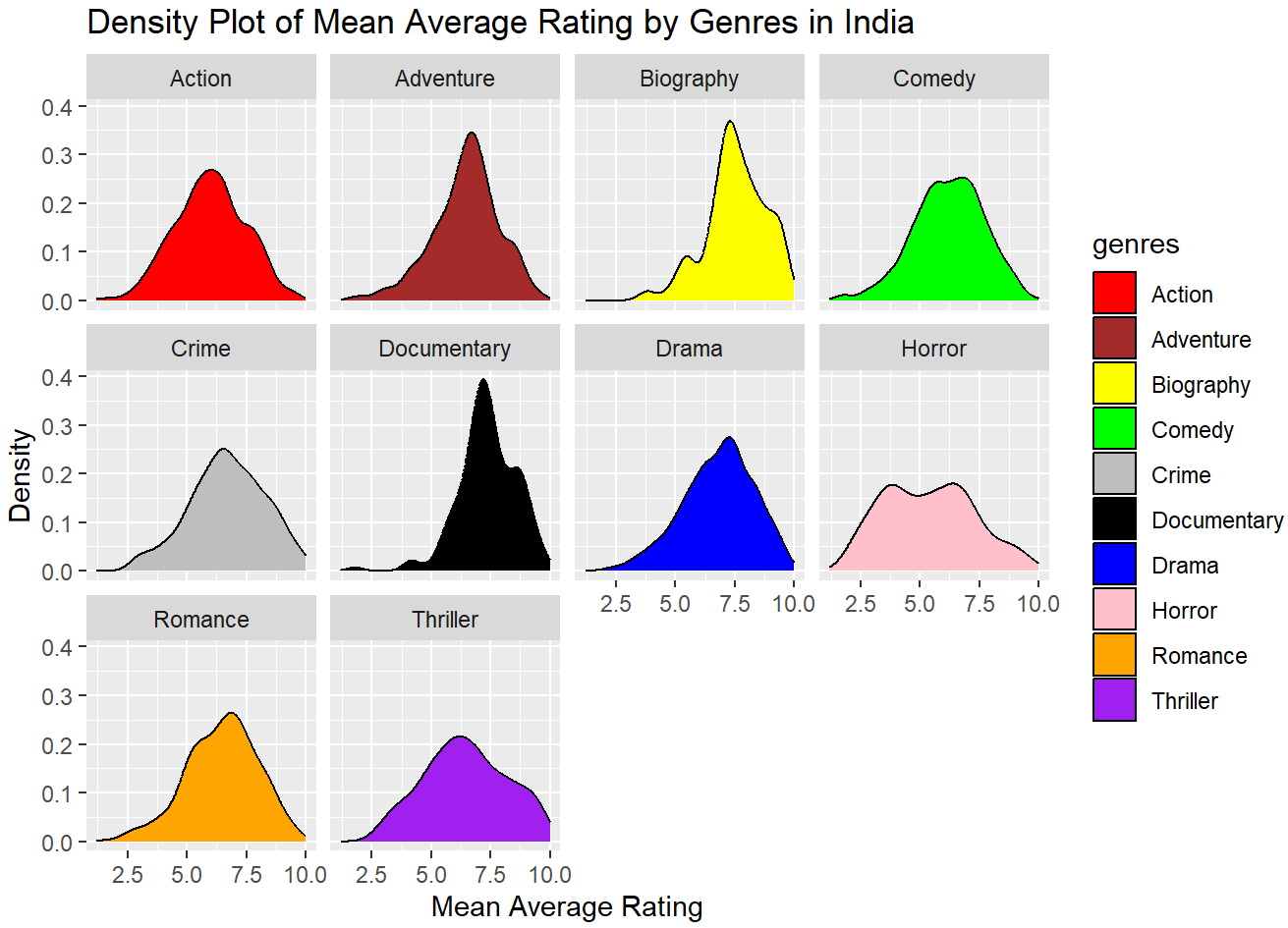
Expanding on the insights provided by the previous graph, which establishes the similarities in median ratings and spreads for Indian and American audiences, the subsequent density plot is essential for a deeper understanding of genre preferences within these two countries.



The density plot reveals that Biography and Documentary genres receive the most consistently high ratings, reflecting their strong appeal to audiences. On the other hand, Horror and Thriller genres are found to have consistently low ratings, indicating a general preference for other genres over these two.

Density Graph for Average Rating by Genre in India

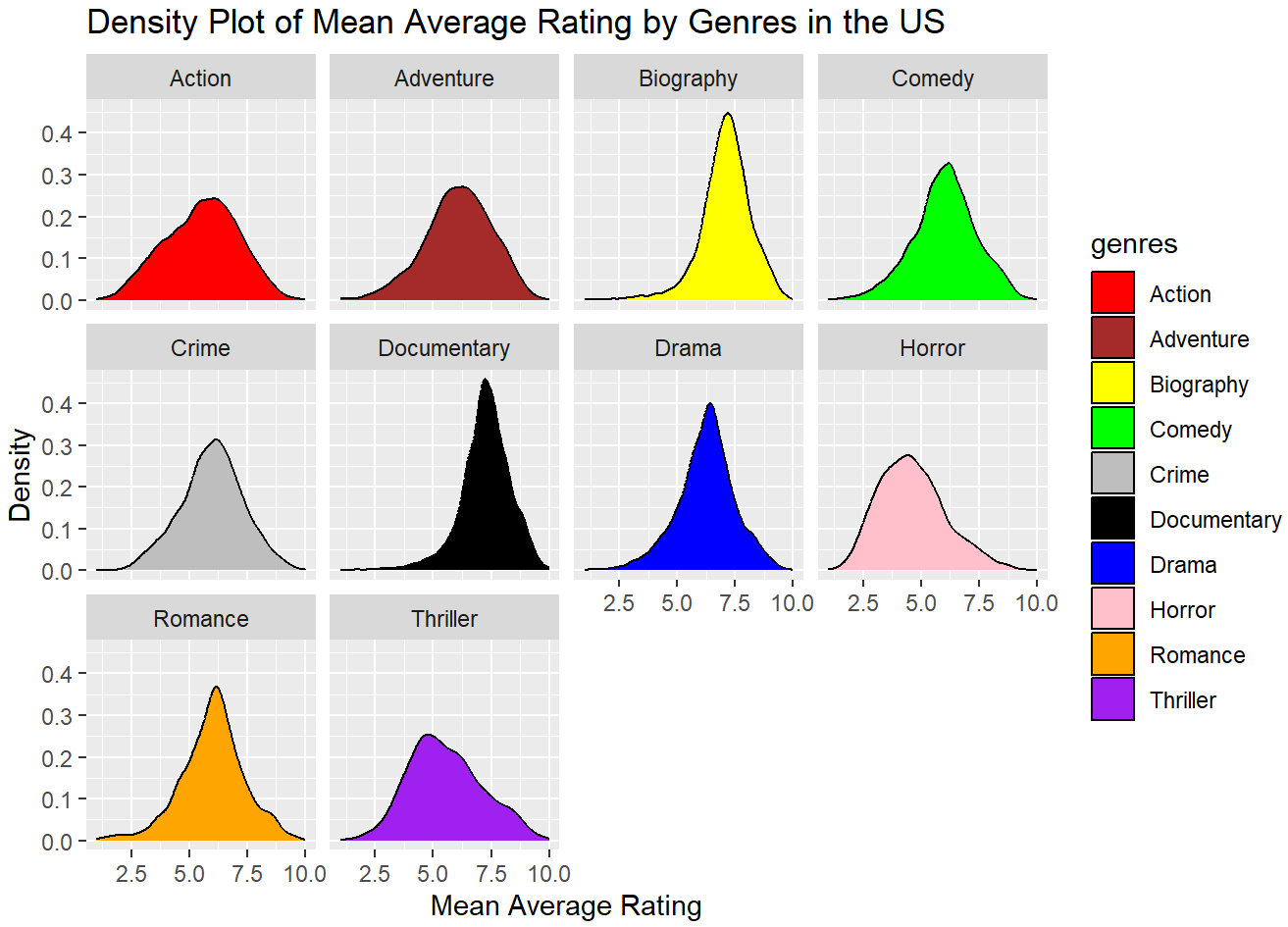
By presenting a density plot that illustrates the distribution of average ratings for various film genres in India, we can delve deeper into the patterns and trends specific to the Indian market. This graph is essential for attaining a comprehensive understanding of the genre preferences of Indian moviegoers.



The data clearly illustrates that biography and documentary films consistently receive the highest ratings with both genres maintaining a similar average rating. In contrast, horror films in India appear to be the least favored genre.

Density Graph for Average Rating by Genre in US

By presenting a density plot that illustrates the distribution of average ratings for various film genres in the United States, we can observe trends specific to the American market. This graph is particularly informative because it allows us to pinpoint the most highly rated genres in the US and compare them with those in India, offering valuable insight into the similarities and differences between the preferences of these two audiences.



This graph highlights the consistent popularity of biography and documentary films and the most highly rated genres. Conversely, the horror genre stands out as the most consistently low-rated genre, reflecting a clear preference among American audiences.

Proportions Inference

We want to investigate the differences in the proportions using a confidence interval. This will aid in discovering whether reviewers rate one country's movies higher than the other. The interval will have a 95% confidence level and use the Agresti-Coull method to improve accuracy. We will be using comedy movies to create the interval because both countries have a similar proportion of all movies made which are comedy.

We are 95% confident that the ratings difference between comedy movies released in India and not the US and movies released in the US and not India will be between -0.0696 and -0.0178.

With this confidence interval, it seems as though reviewers will be harsher on American-only distributed movies compared to their Indian counterparts concerning comedy movies. This aligns with what is shown in the average mean ratings for each genre (separated by country) graph. To determine whether or not the quality of movies within each genre between India and the United States were the same, we need to use a proportions hypothesis test. We will use the comedy genre as a comparison as it has the smallest gap in proportions.

For each sample, we will be looking at the probability that a comedy movie distributed in the domestic country but not in the foreign country will have an average rating greater than or equal to 6.0.

We are assuming that the status quo is that the success probability between the two samples are the same. We are testing that against the idea that they are different. To do so, we need to look at the probability difference based on our data. When testing, we are under the assumption that the success probability is the same.

With our calculation of the standard error, we are able to find the p-value using normal approximation.

There is very strong evidence to suggest that rankings of comedy movies distributed in the US and not in India are different from comedy movies distributed in India but not the US, hence we reject the null hypothesis ($p = 0.0009$, z-test for difference in proportions).

We can simulate the hypothesis test by taking 10000000 results from a Binomial distribution. Comedy's p-value is very low.

Given the statistical significance of the comedy proportions, we want to further investigate whether the proportion of quality movies is similar for all of the top ten genres. Through testing, we can find whether or not the unequal proportions are genres-specific or qualifies for all genres. We will use multiple testing to explore.

When conducting similar hypothesis test for all genres (alpha = [0.05 / 10], z-test for difference in proportions), the p-value results are as follow:

genres	ci_lower	ci_upper	p_value	alpha	hypo
Biography	-0.0459774	0.0688373	0.6963642	0.005	TRUE
Documentary	-0.0330027	0.0553725	0.6198155	0.005	TRUE
Romance	-0.1404791	-0.0104011	0.0230019	0.005	TRUE
Adventure	-0.2306442	-0.0529017	0.0017681	0.005	FALSE
Comedy	-0.0696142	-0.0178262	0.0009354	0.005	FALSE
Crime	-0.2219385	-0.1115464	0.0000000	0.005	FALSE
Action	-0.1148573	-0.0586937	0.0000000	0.005	FALSE
Thriller	-0.2720801	-0.1533711	0.0000000	0.005	FALSE
Horror	-0.2883417	-0.1812838	0.0000000	0.005	FALSE
Drama	-0.1164261	-0.0762649	0.0000000	0.005	FALSE

We fail to see any differences in the rating proportions for Biography, Romance and Documentary genres, but we see dissimilar rating proportions for Adventure, Comedy, Crime, Action, Thriller, Horror, and Drama genres.

Means Inference

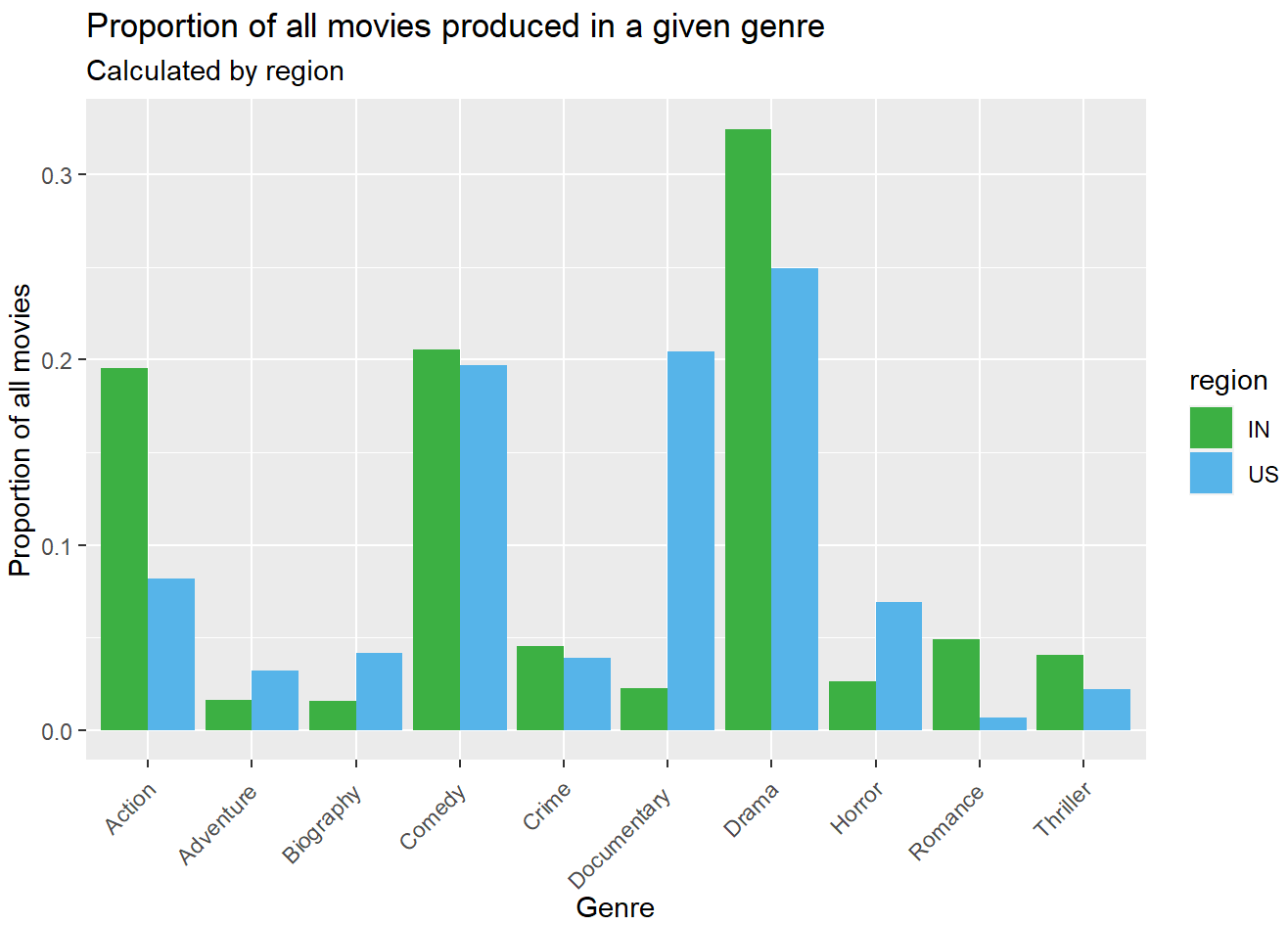
To encompass a bigger picture of movie quality in the US and India, we want to compare the difference in the mean between these independent samples. This will better demonstrate whether there's a difference in average ratings for both countries. We have two independent samples. We have the sample of movie ratings from the United States, classified as X, and the sample of the movie ratings from India, classified as Y. We are under the assumption that the null hypothesis is true, the mean ratings are the same in both countries. We are testing against the idea that they are different. It is most appropriate to use a t-test as the ratings for all genres follow an approximately normally distributed. This is shown in the density graph. The most appropriate t-test to use is Welch Two Sample t-test as the samples are independent of each other.

We have found that the p-value equals 8*10^-39.

There is very strong evidence to suggest that the mean IMDb score for movies released in the U.S. and not in India and for movies released in India and not the U.S. are different, and therefore the null hypothesis is rejected (p=8*10^-39, two-sided t-test, unequal variances). The statistically significant result in the t-test confirms that movie ratings differ between countries. This compounds with the evidence shown throughout the exploration.

Proportion of movies released in a given genre

If one country has a specialty genre, they should have a higher proportion of that movie type. The proportion of movies in a genre are calculated using the original dataset (that way proportions would consider the number of movies in a genre out of all movies, and not just all movies in top 10 genres.)



Note: calculated p may be more variable in India because of the smaller sample size. For example, an additional horror movie would impact the horror movie proportion in India more than it would impact the horror movie proportion in the U.S.

The proportion of total movies in the Drama and Action genres is higher in India. The proportion of total movies in the Horror genre is higher in the U.S.

The specialty genres produced in the United States and India differ wildly. Movies produced in India are likely to be either drama, action, or comedies, while the U.S. produces mostly drama, comedy, and documentary films. The production of these movie genres could be interpreted as movie viewers' preferences, since companies produce movies viewers want to watch. Additionally, the intersection with the comedy and drama movies could mean that those types of movies are easier and cheaper to produce compared to other genres, which allows them to be produced more often. Reviewers in both countries did not appreciate horror movies, while both countries liked biographies and documentaries the most. Neither country had a particular genre they excelled in. This could mean that the specific audiences viewing movies with these genres have leaner and stricter standards. The similarities in genre ratings could also mean India and the U.S. share similar attitudes towards certain genres. Broadly, it could mean that public response to tropes and themes common in movies correspond more with the human experience compared to countries' cultural preference. The statistically significant results found in the proportions and mean hypothesis tests demonstrate that there is a difference in ratings when considering genres both separately and combined. This may mean that critics in the two countries have equally strict standards and the movie quality differs. Another interpretation is that reviewers India and the United States have a different value to number ratings. Reviewers in one country may view a 6 as good while the reviewers in the other country see it as mediocre.

A shortcoming in the analysis is the use of movies that were only released in one country but not the other. While this metric does help narrow down domestic film industries, it doesn't filter solely domestic films. Additionally, with the process of globalization, publishing films outside of movies is ever so prevalent. Limiting down film genres to one, while useful in doing statistical analysis, forego the diversity of films in which secondary or tertiary genres are appropriate descriptions. Our results may have changed if we used Stringr methods to best include all of these genres. There is also selection bias present, as indie and small films in the United States have a greater title eligibility chance compared to their Indian counterparts. There is also the linguistic diversity between India and the United States. It is difficult for critics who don't speak a non-English language, which acts as the lingua franca, to review movies in non-English. This is particular for India, which has a high linguistic diversity. This matters as there's likely to be less reviewers, which could potentially both prevent the movie from showing up on IMDb and lessen the movie review accuracy. Finally, the use of movies from after the year 2000 is limiting as it fails to capture the potential dynamism in movie ratings while also limiting our ability to analyze the trends across both countries' film industries.

In the future, we could compare the rating between countries besides the United States and India. Various countries have comparable movie industries. It would be interesting to compare countries that are relatively more culturally similar (e.g. Anglophone Countries, countries in a specific continent, etc.) to test whether or not they rate movies in the same genre similarly. We also could explore subsets of movies that are ranked below a 6 to determine if some genres have a similar proportion concerning less-quality movies. The original dataset has a plethora of information that this exploration did not focus on, such as movie runtime and age ratings. These new methods would allow for a better understanding of how countries' movie ratings, not just India and the United States, compare and contrast.

To refine our understanding of countries' genre taste, it would be helpful to analyze differences in audiences. Two noteworthy categories would be critics' scores versus general audiences' scores. It would be interesting to explore whether critics from other countries think more alike than general audiences. Additionally, it would be interesting to compare data sources that explore specific audiences, i.e. mainly Americans or mainly Indians.

As aforementioned, the data found on the IMDb website hosts a plethora of other variables. If we wanted to compare visual media as a whole rather than only movies, we could analyze shorts and TV shows. We could synergize the IMDb data with data from other sources to enhance the data accuracy. For example, comparing Rotten Tomatoes ratings with IMDb ratings and averaging their results. Rotten Tomatoes primarily focuses on professional critics' perceptions. We could also compare it to gross box office revenue to see the average amount of money made by a movie in a genre.

We have found that the proportions of "not bad" ratings between India and the United States when considering respective genres are similar. However, when comparing the mean ratings of all genres and focusing only on quality movies, the United States and India are mostly different. Only Documentaries and Biographies share a similar proportion of high-quality movies. There are more areas for exploration that would add to our analysis.

1. "Driving Economic Growth," Motion Picture Association, January 23, 2023, https://www.motionpictures.org/what-we-do/driving-economic-growth/ (https://www.motionpictures.org/what-we-do/driving-economic-growth/).↩
2. IMDb (IMDb.com), accessed April 21, 2023, https://www.imdb.com/interfaces/ (https://www.imdb.com/interfaces/).↩
3. "IMDb Data Files Available for Download." Datasets.imdbws.com, 2023. https://datasets.imdbws.com/ (https://datasets.imdbws.com/)↩
4. Drramyamin, "How to Rate Movies on IMDb.," IMDb (IMDb.com, April 1, 2015), https://www.imdb.com/list/ls076459507/ (https://www.imdb.com/list/ls076459507/).↩
5. "Press Room," IMDb (IMDb.com), accessed April 23, 2023, https://www.imdb.com/pressroom/about/ (https://www.imdb.com/pressroom/about/).↩