COGS9: Introduction to Data Science

Final Project - The Cinematic Bias Analysis: Unraveling IMDb's Top Movie Rankings (download version)

Question

How do factors or biases, other than the subjective quality of the film, affect the ratings and rankings of IMDb top movie titles?

Hypothesis

There are many factors which affect the ratings of the top 250 IMDb movie titles but here are some biases that we think affect the IMDb ratings of these top movies and maybe the main reasons why they have the rating that they do:

Genre

We hypothesize that genre plays an important and determining role in the rating that IMDb top movie titles have. We think that a genre bias exists where the titles that fall under the drama genre on average have a higher rating than the titles that do not fall under this genre. We compare drama vs not drama due to the presence of multiple genres for some titles. This bias can probably be attributed to the distribution of titles by genre as drama is one of the most predominant genres we see in IMDb movie titles. While it may not be the main genre of many titles, it is a very common genre. That in part justifies why most movies with a high rating have drama as one of their genres.

The top 3 Genres (Comedy, Drama, and Documentary) take a share of more than 5% but all other genres account for nothing more than 3%. Additionally, 186 of the 250 top movie titles have Drama as one of the genres.

Gender

We hypothesize that gender bias also plays a role in determining the rating for top IMDb movie titles as, in our dataset, we see that most of the votes that were registered were cast by males and votes by females make up for a smaller percentage of total votes than the votes by males do. Additionally, this bias can also be attributed to the target audience of these films. Most of the movies in the top IMDb titles either belong to a group of movies where titular roles are portrayed by males or belong to a group in which the target audience mostly consists of male moviegoers. Additionally, most of the crewmembers being males on many of these titles may also justify why these movies have the ratings they do from males.

For instance, the top 3 IMDb titles consisted of star casts which were mostly male: The Shawshank Redemption, The Godfather, and The Godfather Part 2. The movies that we are talking about being mostly male-dominated could probably be attributed to the fact that lots of film writing strategies and principles come from males and hence are very male-dominant. One example of this is that a lot of movies for a long time followed the framework of A Hero’s Journey by Joseph Campbell. It is not called A Heroine’s Journey, is it? That is reflected in the history of the cinema that we are all very fond of, where women taking the

director's helm or being the main weight of the movie, unfortunately, took place later than it should have. Only 5 women have been nominated for best director in the Oscars history.

Age

While it is clear that the demographic that votes for IMDb titles might majority consist of adults, it is not the only thing due to which we see the top IMDb movie titles that are on the list right now. We think that the age brackets by which the IMDb votes are classified are very important. Most of the voters on IMDb are adults, according to our dataset. This bias is very interesting to see as only 48 of the top 250 titles are R-rated movies and a total of only 1.8% of all IMDb titles are X-rated. But the bias that we see, may be because of the material being suitable for mostly people who are above 13. That is why we think that most of the titles being rated by adults play a role in the determining top 250 titles. For instance, Toy Story has a higher average rating for the age bracket 0-18, than it does for any other age bracket. If IMDb had a higher concentration of minor voters, it probably would have been higher on the top 250 ranking list.

Background Information

IMDb (Internet Movie Database) is an online database that stores data on a vast number of titles for film, television, home video, and other streaming content. The online database organises these titles in terms of reviews, cast and crew, biographies and much more. Reviews are graded on a scale of 1-10 and are voted in by registered users. Presently, IMDb ranks its titles in terms of a weighted average (Weighted Average Ratings - Help Center, IMDb) of all the votes cast by its registered users. Although not much is known about how this weighted mean is calculated, IMDb assures users that this same weight is applied to all their titles to reflect a more accurate voting average (by accounting for vote stuffing). Some titles however may have additional weightages if unusual voting activity is detected (Ratings FAQ - Help Center, IMDb).

IMDb has a list of the top 250 movies ranked according to their registered users' ratings. The list is calculated with the help of a formula which provides a true "Bayesian estimate," (Ratings FAQ - Help Center, IMDb) that considers the minimum votes required to be eligible for the list, the number of votes for each title, and the mean vote for all titles. The formula is given by:

$$Weighted\ Rating\ =\ (v \div (v + m)) \times R + (m \div (v + m)) \times C$$

V - number of votes for the movie
M - minimum number of votes needed to be eligible for the top 250 (25,000 votes)
R - average rating for the movie
C - mean rating across the whole report

The formula attempts to prevent new movies that just become eligible for the top 250 list from having artificially high ratings.

Data

The perfect dataset that we are looking for contains the top 250 IMDb movies ranked according to the average rating. These ratings must be broken down into various sub-groups to test our hypothesis such as votes by different genders and age groups. Along with this, we are also looking for a breakdown of votes by genre.

We found a close-to-perfect dataset on Kaggle: (https://www.kaggle.com/stefanoleone992/IMDb-extensive-dataset). This data set has nearly 86,000 movies with their names, ratings, and voter demographic breakdown. The names, ratings, movies, and title principles are all in different CSVs. We had to merge them in a Jupyter Notebook so that we could sort the movies in descending order while filtering out movies with less than 25,000 votes. After doing this, we saw some irregularities in the data. Hababam Sınıfı was the highest rated movie which did not match up with the official rankings available on IMDb's website. After trying various data-wrangling techniques, we still did not get an accurate dataset. Thus, we took the top 300 movies instead of the top 250 so that the actual top 250 movies get included in the data set since the problem occurred throughout our data set.

Ethical Considerations

Ethical considerations include the user's privacy (confidentiality) as well as data privacy. IMDb allows users to adjust their privacy settings according to what they want to share. In addition, it is important to consider discrimination in data collection for our question because certain factors such as genre, age and gender affect the outcome rating and can cause a bias if it is not an even spread (which is the case for most ratings). Transparency in the data is another ethical consideration because it is important to note to the users which of their data will be used and how it will be used, especially in this case where their gender and age group are being revealed.

In our research question specifically, we encounter ethical considerations mainly with bias issues including sampling bias, age bias, and gender bias. Since we are looking at data provided by IMDb, we have chosen to look at just the top 300 movies which presents a sampling bias as we are not accounting for the other lists and movie titles (top 500, etc). Age bias and gender bias exist in votes where the votes tend to skew towards male and younger demographics.

A. Data Collection

- A.1 Informed consent: If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?
    - IMDb is a public platform and our data analysis does not use any personal data. Along with this, there are no restrictions on the reproducibility of data.
    - However, on IMDb itself, users can choose whether to submit a rating or not and are given an explanation of which data and how the data will be used.
- A.2 Collection bias: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?

    - IMDb weighs the votes of newly created accounts less because they don't want new votes to heavily influence the ratings
        - IMDb is biased against new voters
    - Sampling bias could be introduced because some members of the total voters have a lower/higher sampling probability due to the targeted audience of each movie on the list we have chosen.
        - We have tried to mitigate this by analyzing the list of the top 250 movies which would be more representative of all demographics vs. a list of the top 10 or 20
        - Additionally in our sample we took the top 300 to be sure that the sample included the top 250 and to mitigate any sources of bias
    - Voter region could be another source of bias because we are not considering votes outside of the U.S.
        - However, we are not focusing on Non-U.S. votes because they are a smaller percentage of the total rankings and are not broken down by country or region
- A.3 Limit PII exposure: Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?
    - Not mentioning any person who worked on the movie.
    - No exposure at all.
    - Not accountable for the data since IMDb is the one to post the data.
- A.4 Downstream bias mitigation: Have we considered ways to enable testing downstream results for biased outcomes (e.g., collecting data on protected group status like race or gender)?
    - Not relevant to our project

B. Data Storage

- B.1 Data security: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
    - Since the data is publicly available on IMDb, there is no need on our part to apply any sort of data security.
- B.2 Right to be forgotten: Do we have a mechanism through which an individual can request their personal information be removed?
    - IMDb allows users to delete or change their ratings at any time. Additionally, users can adjust their privacy settings at any time as well.
    - If this does happen, we will incorporate this change into our project.
- B.3 Data retention plan: Is there a schedule or plan to delete the data after it is no longer needed?
    - We plan to keep this data after this project because several different projects can be worked on using this data.

C. Analysis

- C.1 Missing perspectives: Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?
    - Not relevant
- C.2 Dataset bias: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?
    - Most voters on IMDb are American, which leads to the list of top 250 movies being mostly comprised of American movies.
    - Gender Bias: We'll start with every film that's eligible for IMDb's Top 250 list. A film needs 25,000 ratings from regular IMDb voters to qualify for the list. As of Feb. 14, that was 4,377 titles. Of those movies, only 97 had more ratings from women than men. The other 4,280 films were mostly rated by men, and it wasn't even close to all but a few films. In 3,942 cases (90 percent of all eligible films), the men outnumbered the women by at least 2-to-1. In 2,212 cases (51 percent), men outnumbered women more than 5-to-1. And in 513 cases (12 percent), the men outnumbered the women by at least 10-to-1. (Article by FiveThirtyEight)
- C.3 Honest representation: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?
    - Bar chart for age bias and bar plot for gender bias
- C.4 Privacy in analysis: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?
    - Yes, data that would include any personally identifiable information has not been included or displayed in our visualizations.
- C.5 Auditability: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?
    - Yes, we have ensured that the accurate steps and outline are listed thoroughly and that if needed it could be reproducible in the case that we discover issues and/or need to generate the analysis again.

D. Modeling

- D.1 Proxy discrimination: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
    - We are including gender, genre, and age as our variables, thereby ensuring that we have the least proxy discrimination possible. Therefore the model does not rely on any variable that could be seen as unfairly discriminatory. In addition, any personally identifiable information has been omitted as well.
- D.2 Fairness across groups: Have we tested model results for fairness concerning different affected groups (e.g., tested for disparate error rates)?

    - Initially our model was only going to consider US votes, however after realizing that a lot of brilliant non-Hollywood movies were being left out, we incorporated worldwide votes.
- D.3 Metric selection: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?
    - Our only metric is the number of weighted average votes
- D.4 Explainability: Can we explain in understandable terms a decision the model made in cases where a justification is needed?
    - Yes, with our visualizations and readable model it is possible to easily explain any cases where it might not be clear why the decision was made because the justification would be prominent in the model.
- D.5 Communicate bias: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?
    - We have listed the relevant shortcomings, limitations, and biases present in the model that could be understood by any viewer including relevant stakeholders.

E. Deployment

- E.1 Redress: Have we discussed with our organization a plan for a response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?
    - The nature of our project is such that the project won't harm users. In addition, no personally identifiable information is used so users would remain protected in the case that there was a disagreement amongst the results.
- E.2 Rollback: Is there a way to turn off or roll back the model in production if necessary?
    - Yes, we can simply delete the model if necessary.
- E.3 Concept drift: Do we test and monitor for concept drift to ensure the model remains fair over time?
    - Yes, since the IMDb list is dynamic, we plan to monitor the project so that it remains fair over time. As the IMDb list evolves or changes, the model can adapt and do the same to stay accurate and fair.
- E.4 Unintended use: Have we taken steps to identify and prevent unintended uses and abuse of the model and do we have a plan to monitor these once the model is deployed?
    - Yes, inappropriate or offensive reviews, false information and/or false reviews would be reviewed and taken down if need be. Because the model uses variables that are about the user there is not much concern about unintended uses or abuse of this model.

Analysis Proposal

Data Wrangling and Data Collection

First, we downloaded the above-mentioned data from Kaggle (As .csv files). Then, we imported the data to JupyterHub and used pandas and numpy libraries in Python to clean the data according to our requirements. We used .merge to join the different tables since a single table did not contain all the variables we needed. Every movie in our data has a unique 'title id', through which we were able to join the different CSV files. After joining the required files, we first filtered out all the movies with a low number of votes from regular voters as IMDb's algorithm considers only the regular voters in determining the top 250 titles. Then, we sorted all the movies by their weighted average vote ratings. After sorting, we got the top 300 titles. We took 300 titles instead of 250 to get rid of all the discrepancies such as the dataset having multiple entries with the same number of average votes. After obtaining the top 300 movies according to their weighted average vote ratings, we removed all the unnecessary columns (not needed for our study). After we obtained the final data set to work with, we converted that table to an Excel file, so that we could continue with our exploratory analysis. Here we used data wrangling to clean and unify large, messy data sets to make them easier for access and analysis.

```
In [18]: import numpy as np
         import pandas as pd

In [19]: rating = pd.read_csv('IMDb ratings.csv').set_index("imdb_title_id")

In [20]: movies = pd.read_csv('IMDb movies.csv').set_index("imdb_title_id")

In [21]: title_principals = pd.read_csv('IMDb title_principals.csv').set_index("imdb_title_id")
```

```
In [7]: title = pd.read_csv('IMDb movies.csv').set_index("imdb_title_id")

In [8]: dataset = title.merge(rating, left_on='imdb_title_id', right_on='imdb_title_id')

In [9]: filtered_with_votes = dataset[(dataset.get('top1000_voters_votes') > 200)]

In [10]: top_300 = filtered_with_votes.head(300)

In [11]: top_300.get(['original_title', 'weighted_average_vote', 'us_voters_rating', 'males_allages_avg_vote',
                      'males_allages_votes', 'males_0age_avg_vote', 'males_0age_votes',
                      'males_18age_avg_vote', 'males_18age_votes', 'males_30age_avg_vote',
                      'males_30age_votes', 'males_45age_avg_vote', 'males_45age_votes',
                      'females_allages_avg_vote', 'females_allages_votes',
                      'females_0age_avg_vote', 'females_0age_votes', 'females_18age_avg_vote',
                      'females_18age_votes', 'females_30age_avg_vote', 'females_30age_votes',
                      'females_45age_avg_vote', 'females_45age_votes'])
```

```
In [13]: with_gender = filtered_with_votes.get(['original_title', 'weighted_average_vote', 'us_voters_rating', 'males_allages
               'males_allages_votes', 'males_0age_avg_vote', 'males_0age_votes',
               'males_18age_avg_vote', 'males_18age_votes', 'males_30age_avg_vote',
               'males_30age_votes', 'males_45age_avg_vote', 'males_45age_votes',
               'females_allages_avg_vote', 'females_allages_votes',
               'females_0age_avg_vote', 'females_0age_votes', 'females_18age_avg_vote',
               'females_18age_votes', 'females_30age_avg_vote', 'females_30age_votes',
               'females_45age_avg_vote', 'females_45age_votes', 'genre'])

In [14]: gender_votes_2 = with_gender.sort_values(by='us_voters_rating', ascending = False).head(300)

In [15]: gender_votes_2
```

ages_votes	males_0age_avg_vote	males_0age_votes	males_18age_avg_vote	males_18age_votes	males_30age_avg_vote	...	females_allages_votes	females_0age_avg_vote	females_
1392803.0	9.3	1327.0	9.3	389793.0	9.3	...	274168.0	9.0	
1004808.0	9.2	869.0	9.3	260811.0	9.2	...	151729.0	8.9	
1409165.0	9.2	1544.0	9.3	422587.0	9.0	...	250634.0	8.8	
712572.0	9.1	532.0	9.1	178943.0	9.0	...	96023.0	9.0	
712228.0	9.1	481.0	9.0	172138.0	8.9	...	158875.0	8.8	

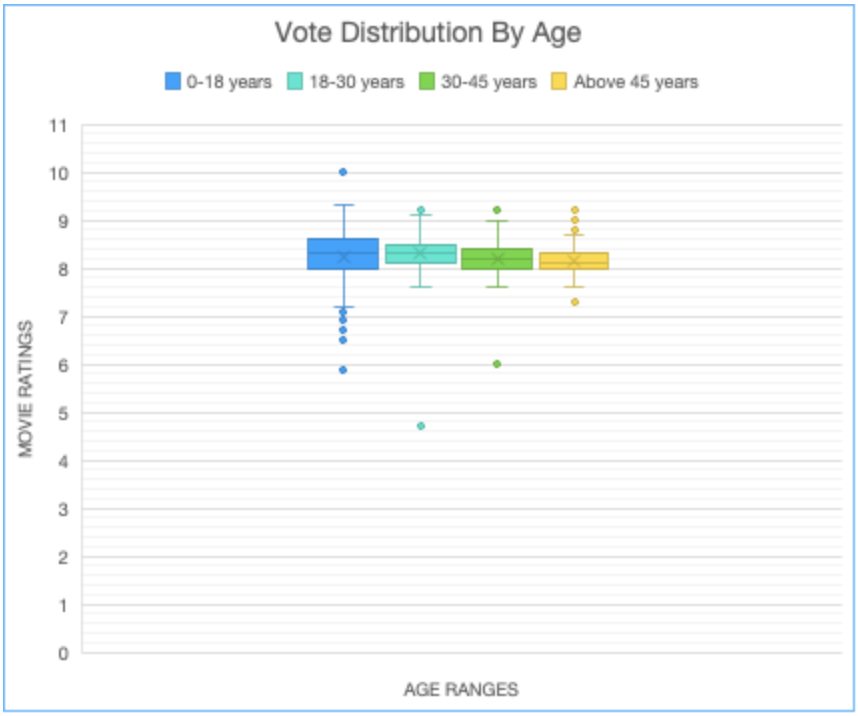
```
In [16]: gender_votes_2.to_csv(r'with_genre_2.csv', index = False)
```

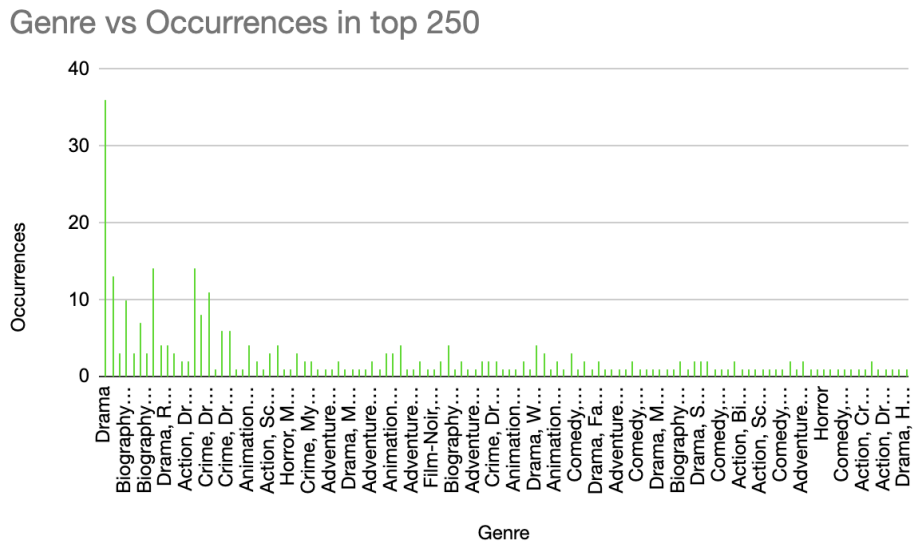
Descriptive & Exploratory

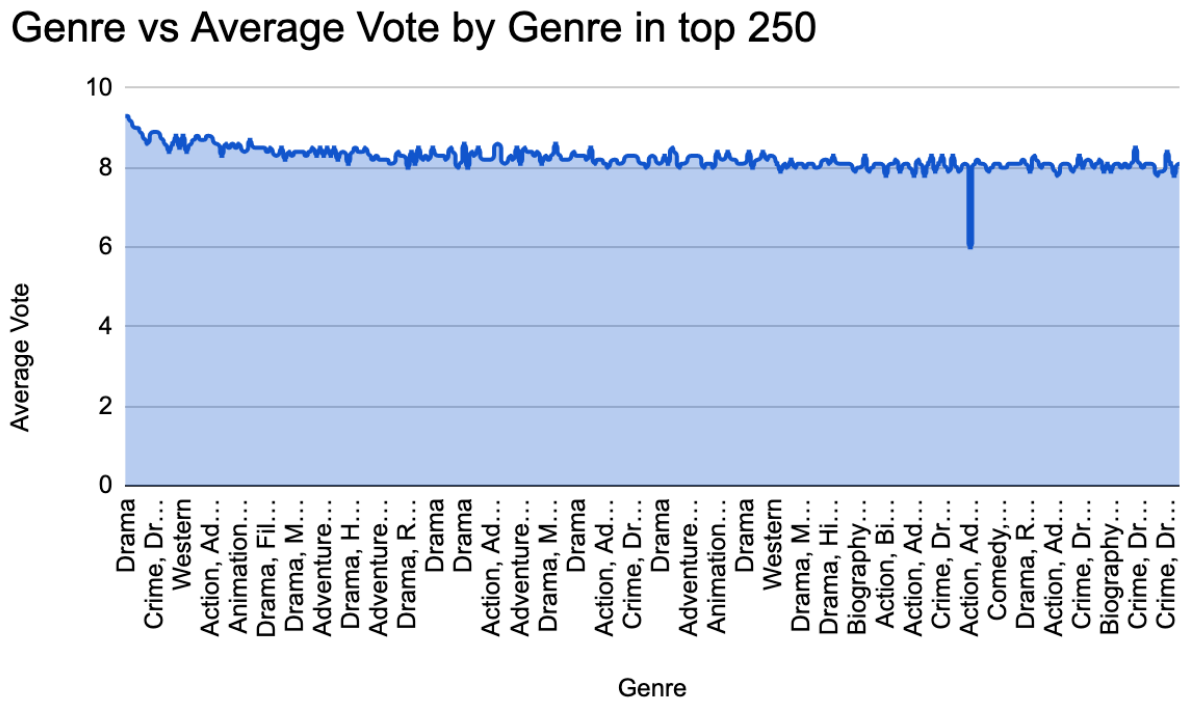
We stated in our hypothesis that we think that movies with the genre drama as one of their genres would have a higher chance of being in the top 250 IMDb titles as the number of titles with drama as their genre is just higher than any other genre. We checked on this claim by carrying out an exploratory analysis of the data set. We made a bar plot showing the distribution of the top titles by their genre and another bar plot of average rating by the genre where we could see the distribution of the titles by their average rating and their genre. Additionally, we hypothesized that the tiles with a majority male voter base would belong to a certain genre. By doing this we show a correlation between two of the variables that we looked at and their effect on the weighted average rating generally. Additionally, we hypothesized that there would be a gender bias in place. We checked on this claim by categorizing the titles by their star cast's gender distribution and plotting them against the average ratings from different genders.

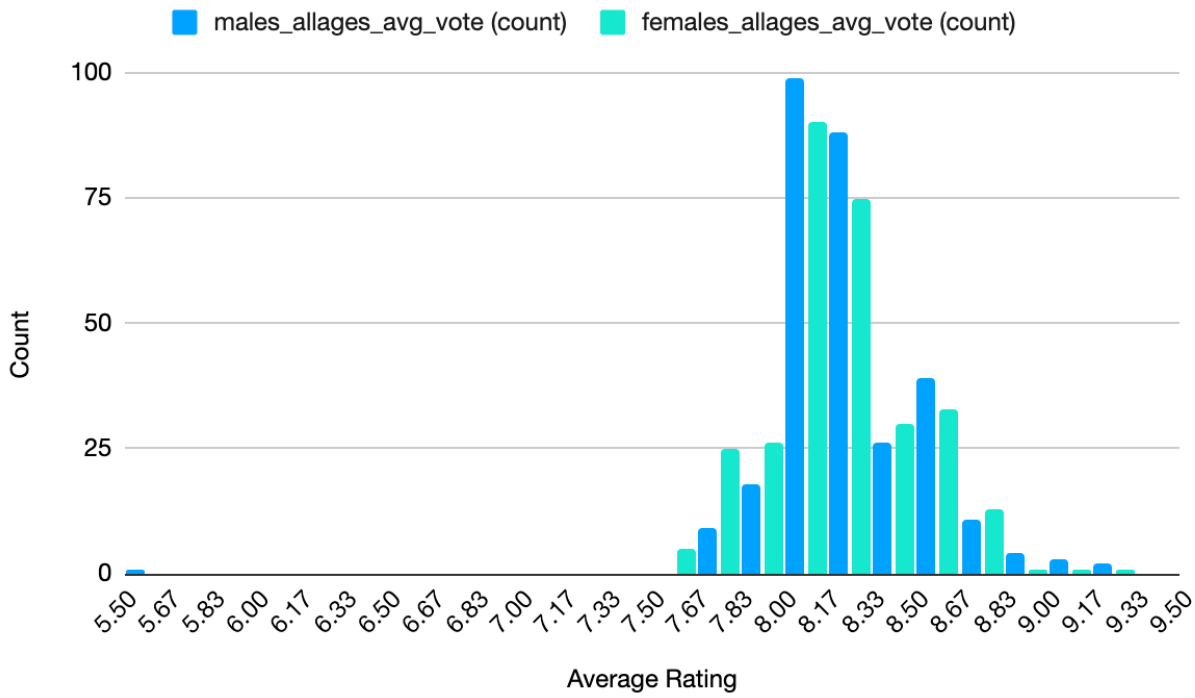
While carrying out this analysis we made sure not to manipulate, change, delete, or add any of the entries to maintain the authenticity of our results and to make sure that we were not twisting the data to fit our needs so that we get the most desirable results possible. We kept the data at its peak accuracy to ensure the same in our results.

Data Visualization









In our visualizations, we plan to describe the results from our analysis in terms of the three variables: age, genre, and gender. The box plot that shows the vote distribution for different age brackets was drawn to get an indication of how the different age groups voted and to answer our claim of adults voting higher for

these movies due to a majority of movies being PG-13 in nature. Additionally, we also hypothesized that the adult age brackets preferred genres like drama and due to the abundance of drama titles adults would have more concentrated votes for these titles and that was the indication we were looking for in the box plot that was graphed.

We chose a column chart to investigate our hypothesis about the dominance of drama. Such a chart indicates the count of each genre. Along with this, we also found out the average rating for each genre to see if any genre dominates over other genres after accounting for the count of each genre. We chose to keep films with multiple genres such as “action, drama” separate from single genres such as both “action” and “drama” since that would simply inflate the dominance of certain genres.

Lastly, we chose a histogram to test our hypothesis about the dominance of males in votes, since most films are male-pandering to some extent. We compared the average ratings by each gender and their counts to see if there were any stark differences between the genders.

Along with this, in our visualizations, we chose to select colors keeping in mind colorblind individuals and as such, chose not to have red and green in the same graph.

Statistical Analysis

For our statistical analysis, we plan to use A|B testing to check whether there is a genre bias present in the data or not. We will use A|B testing to figure out whether the two numerical samples come from the same distribution or not. We plan to check whether there is a significant difference between the ratings of Drama movies vs. the ratings of movies from other categories.

To perform A|B testing, we will only need two columns from our dataset, Genre and Rating. To separate the Drama genre from every other genre, it would be a good idea to change the Genre column to a boolean value column (column consisting of only True and False; True if the movie genre contains the keyword ‘Drama’, and False otherwise). Now our Genre column only consists of True and False values, with True corresponding to Drama movies and False corresponding to ‘Not Drama’ movies.

To perform A|B testing, we will need to formulate the following null and alternative hypotheses:
Null Hypothesis: There is no difference between the ratings of Drama (True) movies and movies of other (False) genres i.e the ratings of Drama movies and movies of other genres come from the same underlying distribution, which means that any difference between the two is due to chance
Alternative Hypothesis: There is a significant difference between the ratings of Drama (True) movies and movies of other (False) genres.

Now, we have to decide on a test statistic through which we can compare the two groups. For our study, our test statistic would be: **Mean rating of Drama movies - Mean rating of ‘Not Drama’ movies**

According to the test statistic, extreme values would tend to support the alternative hypothesis, as this implies that there is a large difference between the ratings of Drama and ‘Not Drama’ movies. First, we

calculate the test statistic from the sample that we have collected. This would be our 'observed test statistic.' Knowing this value is not of much help. We need to know whether this observed test statistic is even significant or not. To check whether our observed test statistic is 'statistically significant' or not, we would need to simulate our test statistic under the null hypothesis and compare that with our 'observed test statistic.'

To carry out the simulation, we would have to make use of random permutations. In random permutations, we will randomly shuffle all the ratings between Drama and 'Not Drama' movies. After shuffling all the ratings, we will obtain completely new, random two columns. We will now compute the value of our test statistic for this randomly permuted sample. Since all the ratings have been shuffled, we will obtain a different test statistic than our 'observed test statistic.' We repeat this process until we have a significant amount of test statistics. The main idea behind this is to simulate the value of our test statistic under the null hypothesis, many times, and then finally compare those values with our 'observed test statistic.'

Now, to conclude the obtained data, it would be helpful to graph a histogram of all the values of our test statistic that we obtained through our simulation. Intuitively, the histogram would look like a symmetric graph, centered around 0 (This is because the difference between the average ratings of 'Drama' and 'Not Drama' ratings would not be much as those differences are due to random chance i.e. under the null hypothesis).

The next step for us is to calculate the p-value and our significance level. The p-value is essentially the probability of obtaining a value as extreme as our 'observed test statistic,' assuming the null hypothesis is true. So, for example, if we simulated 10,000 randomly permuted test statistics and there are 500 of those values which are as extreme as our 'observed test statistic,' our p-value would be (500/10000 =) 0.05. Thus, the lower the p-value, the more statistically significant our 'observed test statistic' would be. If the p-value that we obtain is less than the significance level of our test, we can reject the null hypothesis and conclude that there is a significant difference between the ratings of 'Drama' and 'Not Drama.' If we assume a 5% significance level for our hypothesis test, a p-value of less than 0.05 would lead us to reject the null hypothesis.

This approach does have some limitations. As we have talked about this quarter, we would have to make sure that we don't p-hack the dataset to get the results that we want according to our biases.

Discussion

Results Interpretation

From our analyses we saw that our results were more or less in line with what we hypothesized about the biases that are present in determining the ratings of the top 250 ranked IMDb movies. We saw through visual, statistical, and exploratory analyses that age, gender, and genre all play an important role in determining the rating for IMDb movie titles.

We observed that a majority of the concentrated high votes were made by the age bracket of 18-30 years old voters. While the age bracket 0-18 also voted around the same for the most part, the voters in this age bracket had more scattered votes indicating that this age bracket's votes were more all over the place and was indicative a certain, for the lack of a better word, disagreement in votes while the other age brackets were more united and had a shared opinion of the titles that they voted for.

We also saw the gender bias that we hypothesized about when we carried out A|B testing by changing the variable, votes by a certain gender. We visualized these results in the 4th graph that is shown above and we saw that more males cast votes that were higher in value than females did and vice-versa. This clearly shows that the movies that are highly rated are more preferred by males than females and this can be attributed to the prominence of males in cinema, as we hypothesized before, and the prominence of male voters on the IMDb database.

Additionally, as we see in the second graph a majority of the movies in our dataset belonged to the drama genre and we hypothesized that these titles would have a higher average rating as compared to the other genres due to their prominence in cinema itself and we see that through exploratory analysis and A|B testing we saw our claim stand as the movies with drama as one of their genres had the highest peaks in the graph, meaning that the highest average votes out of all genres were given to titles belonging to these genres with a drama element.

Additionally, we saw links between the three factors in our statistical and exploratory analyses as we saw that certain demographics preferred certain genres and this was reflected in the votes for their titles. To carry out these analyses and draw the results that we eventually obtained we had to go through a detailed data wrangling and data cleaning process.

Limitations

There are a few limitations concerning bias in our data sources including sampling bias and bias that exists within IMDb. Sampling bias could be introduced because some members of the total voters have a lower/higher sampling probability due to the targeted audience of each movie on the list we have chosen. In addition, voter region could be another source of bias because initially, we chose to only look at votes within the U.S. but after we decided to include non-U.S. votes as well. As said before, there is also a bias that already exists in IMDb because they value the votes of newly created accounts less than existing accounts.

We have tried to mitigate these limitations by analyzing the list of top 300 movies which would get rid of any discrepancies in the weighted average and be more representative of all demographics in addition to making sure the top 250 titles are included. Also, by viewing votes outside of the U.S. we are addressing any bias issues related to voter region by eliminating further potential pitfalls by including the percentage of total rankings of worldwide votes. Additionally, the sources of our crowds do affect our outcome because we are using information from the users who have accounts with IMDb so our data relies on them.

Our analysis and question only take into account the top-rated IMDb movies, so we have to be careful and not draw conclusions about all the movies on IMDb. In other words, this study is not representative of all the IMDb movies, but only the top ones. In our study, we have tried to limit the effects of confounders by working with only a few variables for each bias. However, it is also important to note that there might be some other factors at play. This is the reason why we have not established causation in our study, but rather significant differences which exist. Additionally, the data wrangling and cleaning process that we went through could be seen as manipulating/torturing the data and it would increase the sampling bias that is present in our studies but we were careful enough to not let that hinder our results.We took three factors into account, while many other factors such as the geographical setting of the production house, movie, and the voters also play a role in determining ratings for IMDb titles so that also is a limitation to our study but we did consider the top three factors that we as watchers and users of the database thought were most important ones.

Societal and/or ethical implications

In our proposed project, there are a few possible societal and/or ethical implications that could impact the users and viewers of our model. One of them includes the possibility of viewers concluding all movies on IMDb as well as how the viewer's movie-going or movie-watching experience is influenced. The overall ranking and score of a movie title could potentially affect whether or not someone who views the ranking decides to watch that movie which could then impact how well a movie does and how many people see it. Additionally, disagreements could arise because our proposed project deals with people's personal opinions and rankings. Each individual has their own opinion on the movie title and scores it according to their personal view so some may think a movie is great while others believe the same movie is not so good, this leads to disagreements amongst viewers and users on the resulting score which can be skewed higher or lower than what they believe the movie deserves. This disagreement can lead to the societal implication of a decline in IMDb users as well as those who come to IMDb to view rankings on movie titles, etc because we are analyzing other people's views which are subjective to them. However, these ethical considerations can be addressed by providing clear explanations of how each movie is reviewed, what the score breakdown is, and the calculation of the final weighted average. This information would help address these issues because then people are less likely to immediately conclude movie titles after knowing how the score is calculated.

We believe that

.

.

.

.

.

.

“Data Science and Films are two things that can hugely help advance human imagination and we love chasing our imagination despite knowing we’ll never catch it.”

.

.

.

.

.

.

- Aryanka Thaker, Amol Khanna, Shouvik Guha, Akshat Mittal, Khyat Doshi