MovieLens Insights

Exploratory Data Analysis

Case Study 3

Group 3

Atharva P. Kulkarni (akulkarni@wpi.edu)
Rutuja Dongre(rdongre@wpi.edu)
Shubham Wagh(swagh@wpi.edu)
Yash Malviya(ymalviya@wpi.edu)

Master's in Data Science

Worcester Polytechnic Institute

DS 501: Introduction to Data Science

Professor Dr. Torumoy Ghoshal (tghoshal@wpi.edu)
Teaching Assistant: Himan Namdari (hnamdari@wpi.edu)

November 2, 2023

Our Approach:

The three files that included the data that we downloaded were movies.dat, users.dat, and ratings.dat. The first step was converting the .dat files to .csv files. To do this, we saved each file with a.csv extension and added commas to every type of separator found in the.dat files. Next, we imported the data into our Python environment using pandas (pd.read_csv). The following step was combining all of the data frames into one (movielens_df). Following this, we needed to gather some data, such as the number of films with an average rating of 4.5 overall and the average rating of 4.5 for both men and women. Additionally, we discovered the average score for any film in which the age of every guy and woman was over 30.

Movies with average rating over 4.5: 21
Movies with average rating over 4.5 among men: 23
Movies with average rating over 4.5 among women: 51
Movies with median rating over 4.5 among men over age 30: 86
Movies with median rating over 4.5 among women over age 30: 149

Ten most popular movies:

American Beauty (1999): 3428
Star Wars: Episode IV - A New Hope (1977): 2990
Star Wars: Episode V - The Empire Strikes Back (1980): 2990
Star Wars: Episode VI - Return of the Jedi (1983): 2883
Jurassic Park (1993): 2672
Saving Private Ryan (1998): 2652
Terminator 2: Judgment Day (1991): 649
Matrix The (1999): 590
Back to the Future (1985): 2582
Silence of the Lambs The (1991): 2578
Average rating by people over 30: 3.65
Average rating by people 30 and under: 3.53

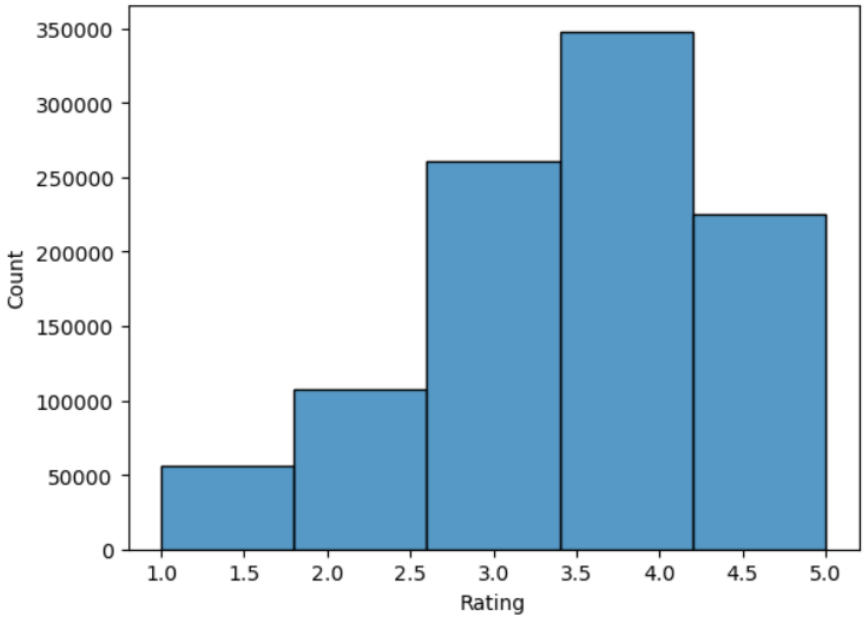
Conjecture is false: Older people are not necessarily harder to please.

The purpose of all these exercises was to gain an understanding of how age influences movie ratings and how various genders have varied options.

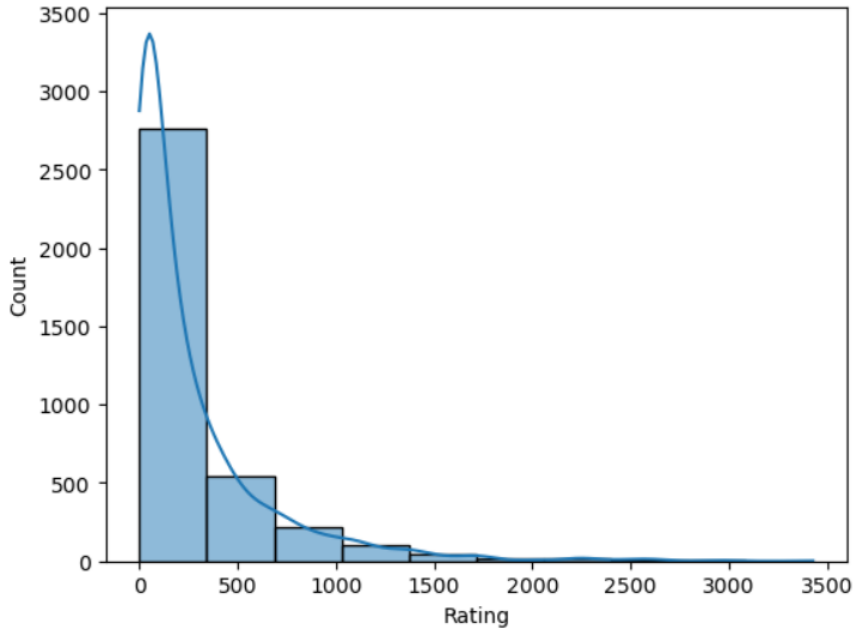
Making a histogram out of some data was the next challenge at hand. We were able to determine the distribution of the data from this. The total number of ratings each movie

received, the average rating per movie, and the overall ratings in the data were all presented as histograms.

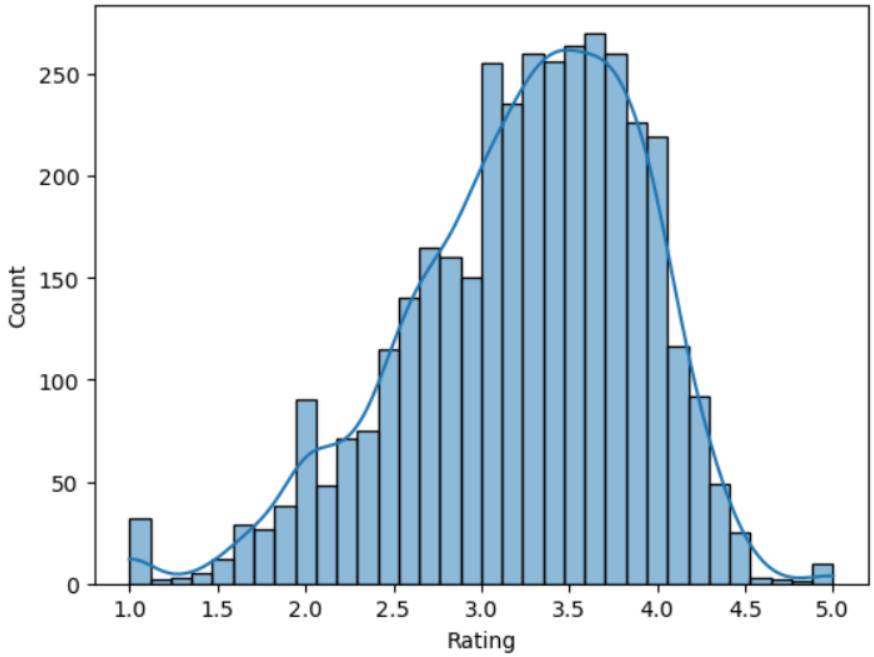
Overall Ratings given by users:



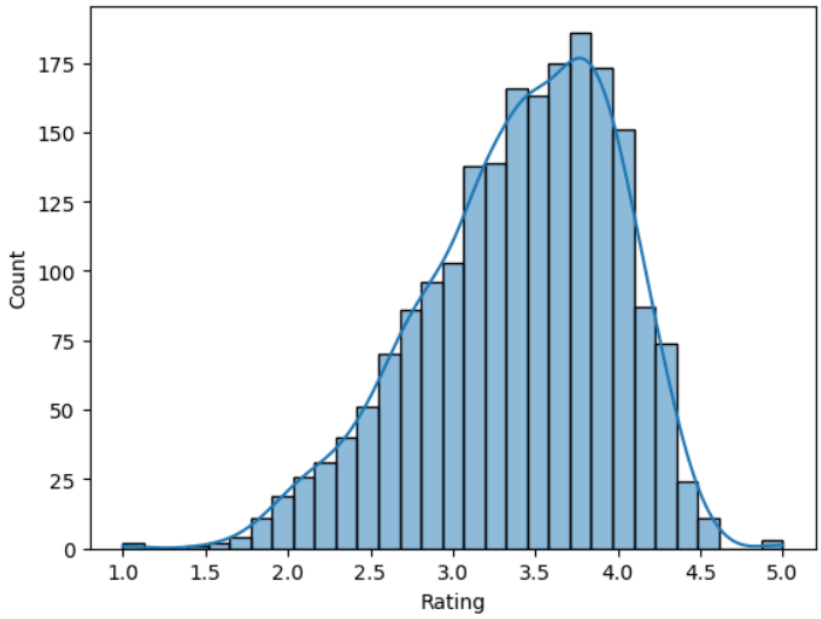
Ratings received by each movie:



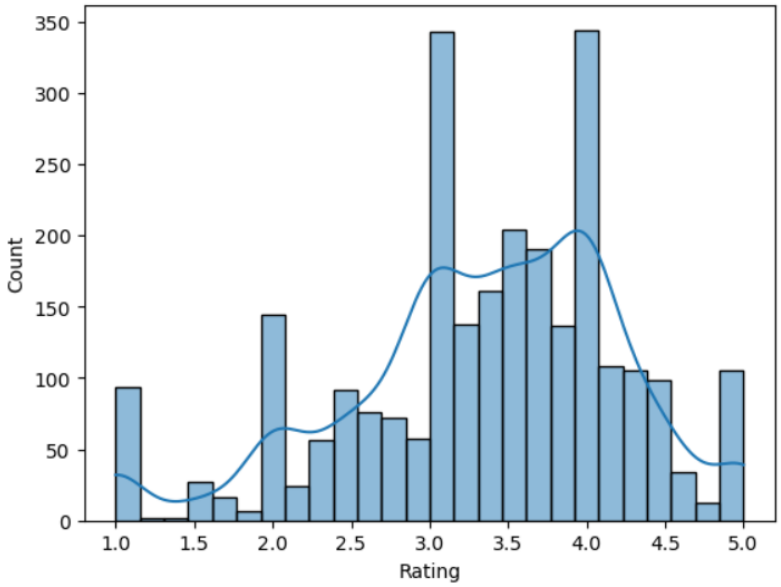
Average ratings received by each movie:



Following this, we had to remove some data and only take into account the films that received a total rating of more than 100, as there were a lot of films with ratings of less than ten. We were not going to consider these kinds of films because the limited number of ratings does not allow us to make any inferences. We were able to identify a change in the histogram after eliminating these. The histogram began to show no interspikes and a much more regularly distributed appearance.

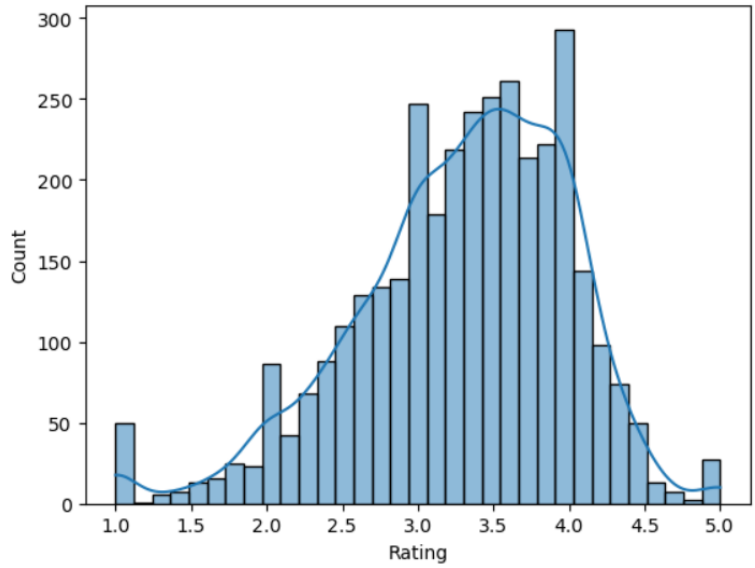


Next, we only looked at the data that included users who were between the ages of 1 and 10. This was carried out in order to identify any patterns in the way that kids rated the films. Based on this, we discovered that more people in this age range gave severe scores, such as 1 or 5. However, in this age range, three or four was the greatest number of evaluations.

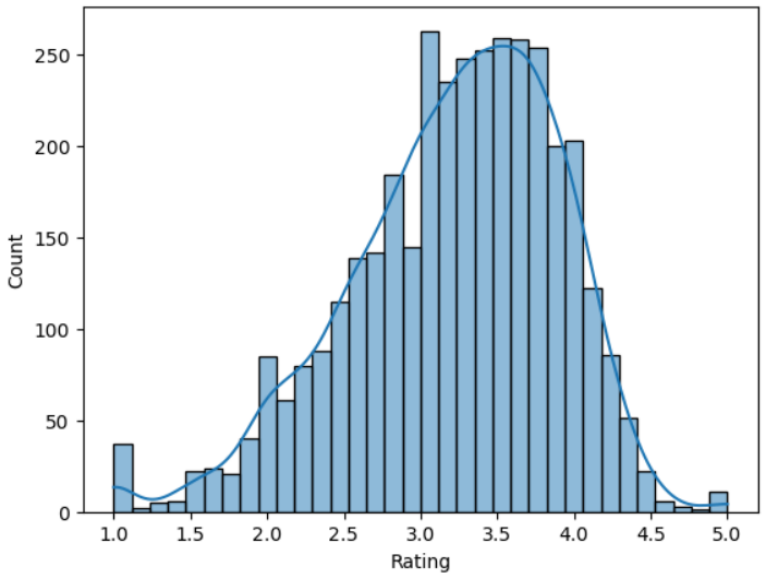


We also considered the average of men and women for each movie.

Average Rating by women.

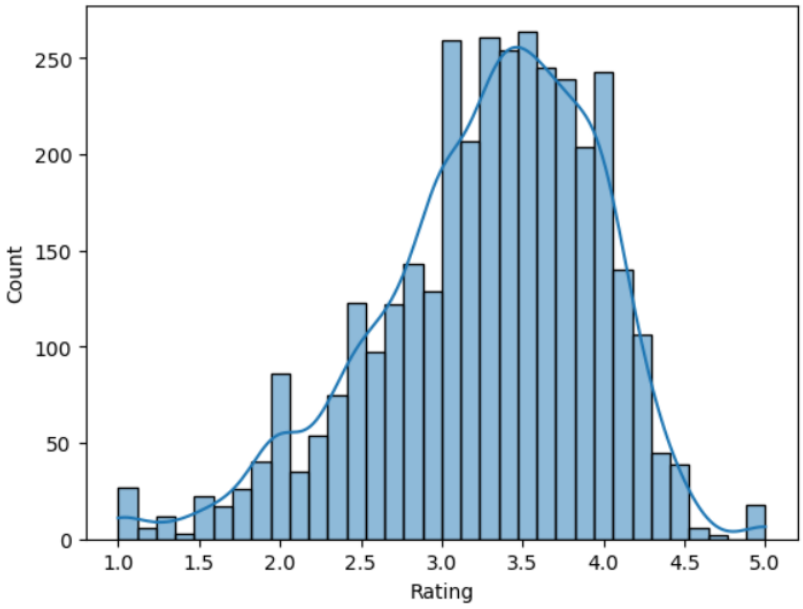


Average rating by men:

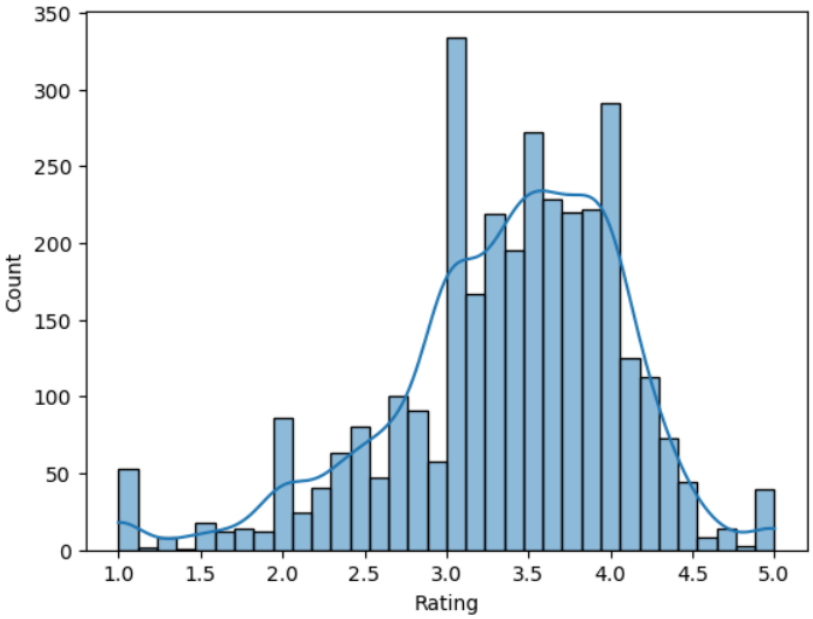


We also considered the average rating given for each movie by men and women over the age of 30.

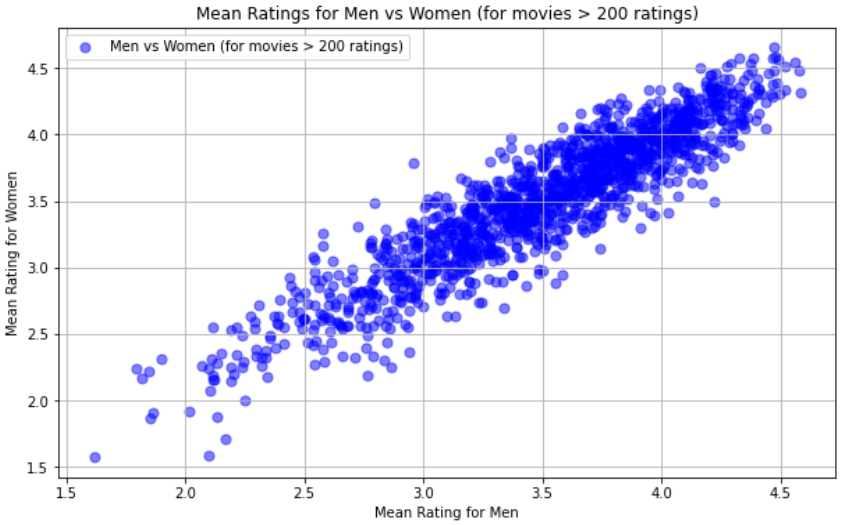
Average rating by men over 30:

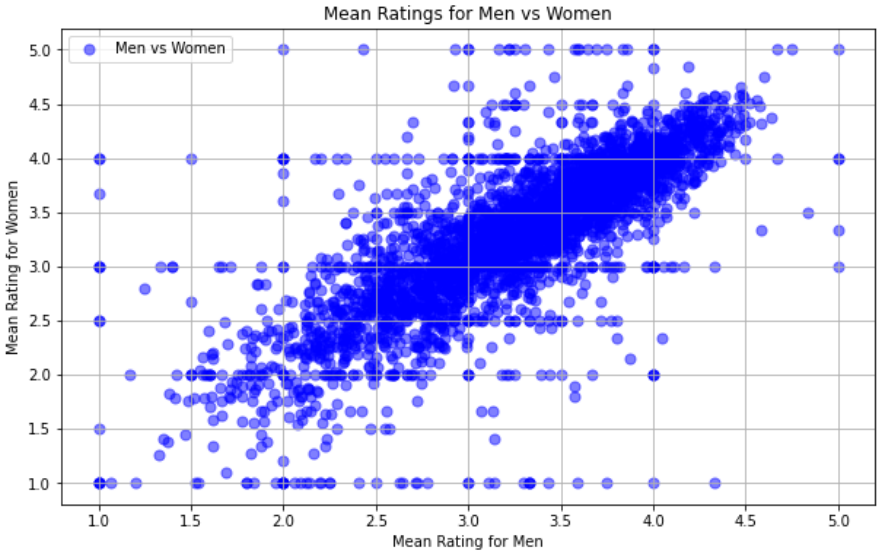


Average rating by women over 30:



We also did some analysis and compared the data where average ratings for each movie were compared and plotted some scatter plots for them.





Additionally, we discovered a link between the average ratings that men and women gave each movie. About 0.7 was the correlation coefficient. This suggests that the average ratings provided by men and women are positively correlated.

About the data

There were 998080 rows and 10 columns in the dataset. This makes a total of 9980800 data points in the dataset.

The columns in the data set were: UserID, MovieID, Rating, Timestamp, Age, Gender, Occupation, Zipcode, Title and Genres.

	UserID	MovieID	Rating	Timestamp	Gender	Age	Occupation	Zip Code	Title	Genres
0	2	1357	5	978298709	M	56	16	70072	Shine (1996)	Drama|Romance
1	8	1357	4	978230800	M	25	12	11413	Shine (1996)	Drama|Romance
2	10	1357	5	978227625	F	35	1	95370	Shine (1996)	Drama|Romance
3	18	1357	5	978156876	F	18	3	95825	Shine (1996)	Drama|Romance
4	33	1357	4	978109488	M	45	3	55421	Shine (1996)	Drama|Romance

The data type for each column is as follows:

1. UserID: int64
2. MovieID: int64
3. Rating: int64
4. Timestamp: int64
5. Gender: object
6. Age: int64
7. Occupation: int64
8. Zip Code: object
9. Title: object
10. Genres: object

The mean, median, standard deviation for all the integer values was found to be as follows:

	UserID	MovieID	Rating	Timestamp	Age	Occupation
count	998080.000000	998080.000000	998080.000000	9.980800e+05	998080.000000	998080.000000
mean	3024.608884	1869.434325	3.580357	9.722469e+08	29.744052	8.035970
std	1728.273821	1093.878797	1.117300	1.215794e+07	11.749987	6.531805
min	2.000000	2.000000	1.000000	9.567039e+08	1.000000	0.000000
25%	1506.000000	1034.000000	3.000000	9.653027e+08	25.000000	2.000000
50%	3070.000000	1845.000000	4.000000	9.730190e+08	25.000000	7.000000
75%	4476.000000	2772.000000	4.000000	9.752211e+08	35.000000	14.000000
max	6040.000000	3952.000000	5.000000	1.046455e+09	56.000000	20.000000

What business decision do you think this data could help answer? Why?

1. **Movie recommendations**: Based on user ratings, preferences and their demographics, similar movies can be recommended to the users. Thus analyzing genre and other factors can help build effective recommendation algorithms and can help personalize and improve the user experience leading to more business revenue.

2. **Genre-specific or Age-specific marketing**: Understanding which genres are preferred by different user segments based on different demographic groups can help forming targeted marketing strategies. This can lead to more efficient marketing spending and can increase ticket sales or streaming subscriptions.

3. **Type of Content Production**: What types of movies should the company invest in or acquire to meet user demand can be deduced by analyzing genre preferences, user demographics, and ratings. It can guide with decisions on which types of content to produce or acquire. This can lead to more successful and profitable content strategies.

4. **Aging Movie Catalog Management:**
   When should older movies be removed from the catalog, and when should they be promoted or remastered can be inferred by analyzing how movie ratings change over time can help in deciding when to promote older films, remaster classics, or remove less popular content. This can optimize the content library and improve user retention.

5. **Collaborations & Partnerships:**
   Business decisions like, based on user preferences, which directors, studios, or actors should the company work with can be determined by finding out whether partnerships and collaborations will be well-received by the public. Analyzing user preferences will result in successful joint ventures and co-productions.

- **What conjectures did you make and how did you support or disprove them using data?**

- Conjecture: Older people (age above 30) are harder to please (lower average ratings)
    - Regarding the conjecture about different age groups and their satisfaction levels, the provided data contradicts the assumption that older people are harder to please. In this case, the average rating by people over 30 is slightly higher (3.65) compared to the average rating by people 30 and under (3.53). It suggests that older audiences might be as, if not more, pleased with the movies they watch compared to younger audiences.

- Conjecture : age range (1-10) given more extreme ratings? Can we think children are more or less likely to rate a movie 1 or 5?

    - Based on the histogram, I think that younger audiences are more likely to give extreme ratings (1 or 5 stars) than older audiences. This is because the histogram shows that there are two peaks in the distribution of ratings, one at 3.5 stars and one at 4.5 stars. This suggests that there are two groups of movies: those that are generally liked, and those that are generally disliked. Younger audiences may be more likely to be influenced by their peers or by the popularity of a movie, and they may be less likely to have a nuanced understanding of what makes a movie good or bad. As a result, they may be more likely to give a movie a rating of 1 or 5 stars, depending on whether they liked it or disliked it.

- Conjecture : Men, as a group, tend to provide their own distinct set of ratings for movies within the MovieLens dataset. Their average ratings for various films might reflect certain preferences, genre inclinations, or specific storytelling elements that cater to or align with their tastes and interests.

    - The histogram shows the distribution of average ratings for movies by men. The histogram is skewed to the right, with a long tail towards higher ratings. This suggests that most movies are rated favorably by men, but there is a small number of movies that are rated very highly.this conjecture is true from above analysis.

- Conjectures : Female reviewers within the MovieLens dataset offer a unique perspective on movies, resulting in varying average ratings across different films. Their ratings may reflect distinct genre preferences, storytelling elements, or cinematic aspects that particularly resonate with this demographic group, potentially influencing the overall perception and reception of these movies.

    - The histogram shows the distribution of average ratings for movies by women. The histogram is also skewed to the right, with a long tail towards higher ratings. This suggests that most movies are rated favorably by women, but there is a small number of movies that are rated very highly.Women are more likely to rate movies favorably than men. This is supported by the fact that the histogram is skewed to the right, with a long tail towards higher ratings. Women are more likely to be interested in movies that are romantic, comedies, or dramas. These genres of movies are typically rated higher by women than by men.

- Conjecture : under what circumstances the rating given by one gender can be used to predict the rating given by the other gender. For example, are men and women more similar when they are younger or older?

    - Thus we can observe that the correlation coefficient of ratings of men and women below thirty is higher than that of those above 30. So younger males and females have more similar ratings than the older ones

- conjecture :If there's a high positive correlation between the ratings given by both genders for a specific type or genre of movies, it suggests a strong relationship. In such cases, the rating given by one gender might reasonably predict the rating given by the other gender. For instance, if men and women consistently give similar ratings to action movies, it might be possible to predict one gender's rating based on the other's.

    - The insights derived from the correlation coefficients explicitly confirm and reinforce your conjecture. The evidence provided by the correlations for drama and comedy movies distinctly aligns with the pattern and degree of agreement in how men and women rate these specific genres,

validating the conjecture's key points regarding the levels of consistency, alignment, and shared appreciation across genders for each genre.

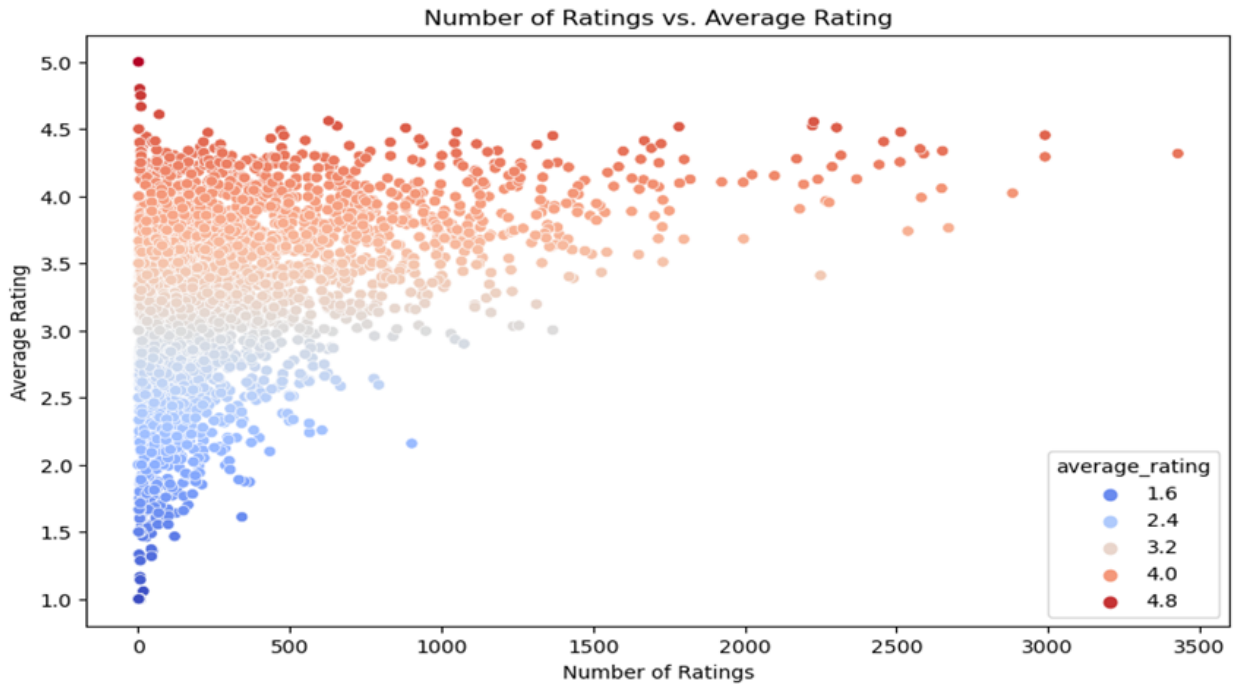
Surprising Discoveries in Movie Ratings Analysis

MovieLens Data Exploration

As part of our extensive analysis of the MovieLens dataset, which encompasses a wealth of user ratings and demographic data, we have stumbled upon several intriguing and counterintuitive trends. These findings could potentially reshape our understanding of viewer preferences and movie popularity.

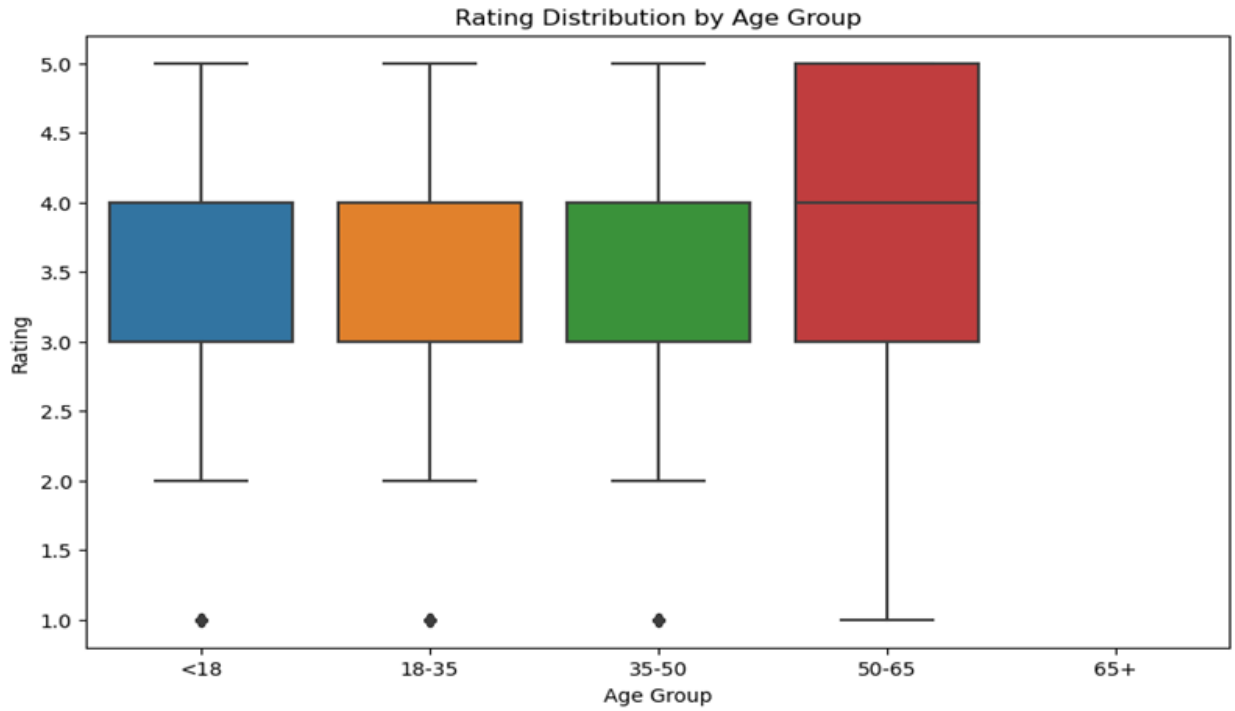
Surprising Popular Movies with Lower Average Ratings

One of the most unexpected discoveries was the identification of films that, despite having amassed a significant number of ratings, maintain a surprisingly low average rating. This phenomenon contradicts the common assumption that high viewership correlates with high ratings. For instance, some blockbuster movies have garnered massive audience numbers yet their average ratings hover around the modest 3.0 mark. This discrepancy may suggest a distinction between the concepts of 'watchability' and 'quality', where certain movies are widely consumed due to factors such as star power, marketing, and social phenomena, but do not necessarily satisfy the audience to a corresponding degree.



Demographic Anomalies

A further surprising pattern emerged when examining the demographic distribution of ratings. Conventional wisdom might suggest that younger audiences are less critical and more likely to rate movies generously. However, the data tells a different story. Users in the 18-24 age bracket were unexpectedly harsh in their ratings, often rating movies lower than their counterparts in the 50+ age range. This insight challenges the narrative that youthful audiences are less discerning and prompts a reevaluation of target demographic groups for certain film genres



These unexpected findings underscore the complex nature of viewer preferences and the multifaceted factors that drive movie popularity. While certain trends align with industry perceptions, others invite a deeper investigation into the dynamics of movie consumption. The implication for stakeholders in the entertainment industry is clear: data-driven insights must be nuanced and contextual, with an acknowledgment that numbers alone do not tell the whole story.

Story of our group:

Our group effort involved investigating movie data. We encountered difficulties preparing the data for analysis. Unexpectedly, some really well-liked films received poor reviews, while younger audiences proved to be harsher reviewers than anticipated. Additionally, despite some discrepancies, we discovered that men and women often agree when it comes to movie evaluations. We learned from this project that data can be complex and that the story is not always told by the numbers.We found that firms in the movie industry can benefit from data by using it to understand customer preferences and make better decisions when it comes to movie recommendations. It was illuminating and demonstrated to us how data analytics may provide hidden information about what audiences find appealing in films.

Task Distribution Amongst the Group Members

Data Gathering and Merging: Rutuja Dongre
Problem 1: Shubham Wagh
Problem 2: Atharva Kulkarni
Problem 3: Rutuja Dongre
Problem 4: Yash Malviya
Conjectures and Business Insights: Yash Malviya
Report: Atharva Kulkarni, Shubham Wagh
Presentation: Atharva Kulkarni, Rutuja Dongre, Shubham Wagh and Yash Malviya