Exploring Movie Ratings and User Preferences in Business Intelligence

Submitted By:

Nitya Phani Oruganty Santosh

Sreeram Marimuthu

Chithramvel Sanarpalayam Selvamuthukumar

Index:

1. Abstract:

In the era of data-driven decisions, understanding user preferences and behaviour is critical for businesses, especially in the entertainment industry. This report explores the MovieLens dataset to analyse how various factors, such as occupation, gender, movie genres, and age groups, impact movie ratings. Through data analysis and conjecture testing, we aim to uncover valuable insights from a business intelligence perspective.

2. Introduction:

In the realm of Business Intelligence (BI), data-driven insights are invaluable for making informed decisions. One fascinating domain for such analysis is the entertainment industry, where user preferences play a pivotal role. The MovieLens dataset, a collection of movie ratings, provides a rich source for exploring user behaviours and their impact on movie ratings.

3. Data Analysis:

3. a. Data Preparation:

To begin our analysis, we carefully prepared the dataset from the 3 datasets provided with us. By combining them all we created a single file (dataframe) MovieLens in HDF5 format.

3.b. Exploratory Data Analysis (EDA):

We employed EDA techniques to gain an initial understanding of the dataset. This included statistical summaries, distribution plots, and the identification of any anomalies.

4. Task 1: Conjectures:

4.a. Conjecture 1: Users with the occupation "writer" give lower ratings on average.

Our first conjecture aimed to determine whether individuals with the occupation "writer" tend to give lower movie ratings on average. We calculated the average ratings for various occupations and found that writers do not give lower ratings on average.

4.b. Conjecture 2: Men are more generous with high ratings (4 or 5) than women.

The second conjecture explored the gender-based differences in movie ratings. By calculating the proportions of ratings 4 or 5 given by men and women, we discovered that women are more generous with high ratings (4 or 5) compared to men.

4.c. Conjecture 3: Movies in the "Action" genre tend to receive higher average ratings than movies in the "Romance" genre.

This conjecture revolved around the genre-specific impact on average ratings. We found that movies in the "Action" genre indeed tend to receive higher average ratings than movies in the "Romance" genre.

4.d. Conjecture 4: Users have a preference for specific movie genres, and their ratings may vary depending on the genre.

To support this conjecture, we calculated the average rating for each genre and identified the genres with the highest average ratings. This analysis suggested that users have genre preferences that influence their ratings.

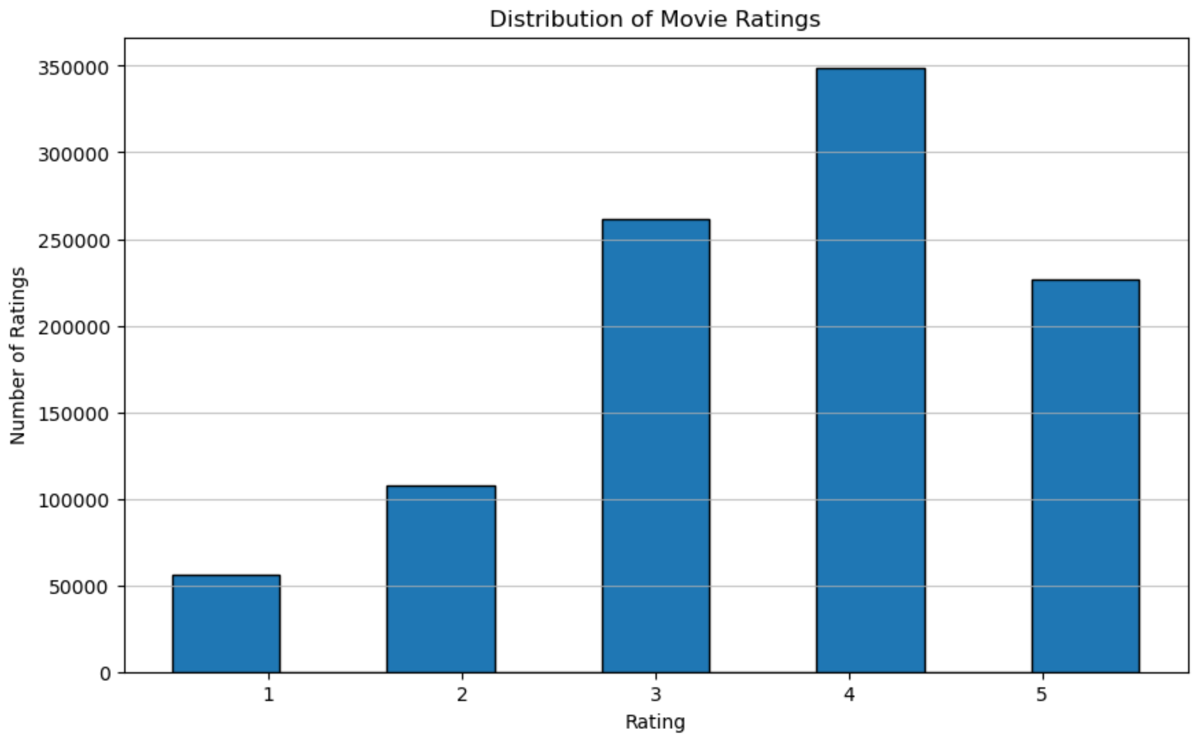
4.e. Conjecture 5: The choice of movie genres and the average rating given by users may vary based on their age group.

We introduced age groups to the analysis and calculated the average rating for each combination of age group and genre. Our findings supported the notion that different age groups have varying preferences for movie genres, reflected in their average ratings.

5. Analysis of Movie Ratings and User Preferences

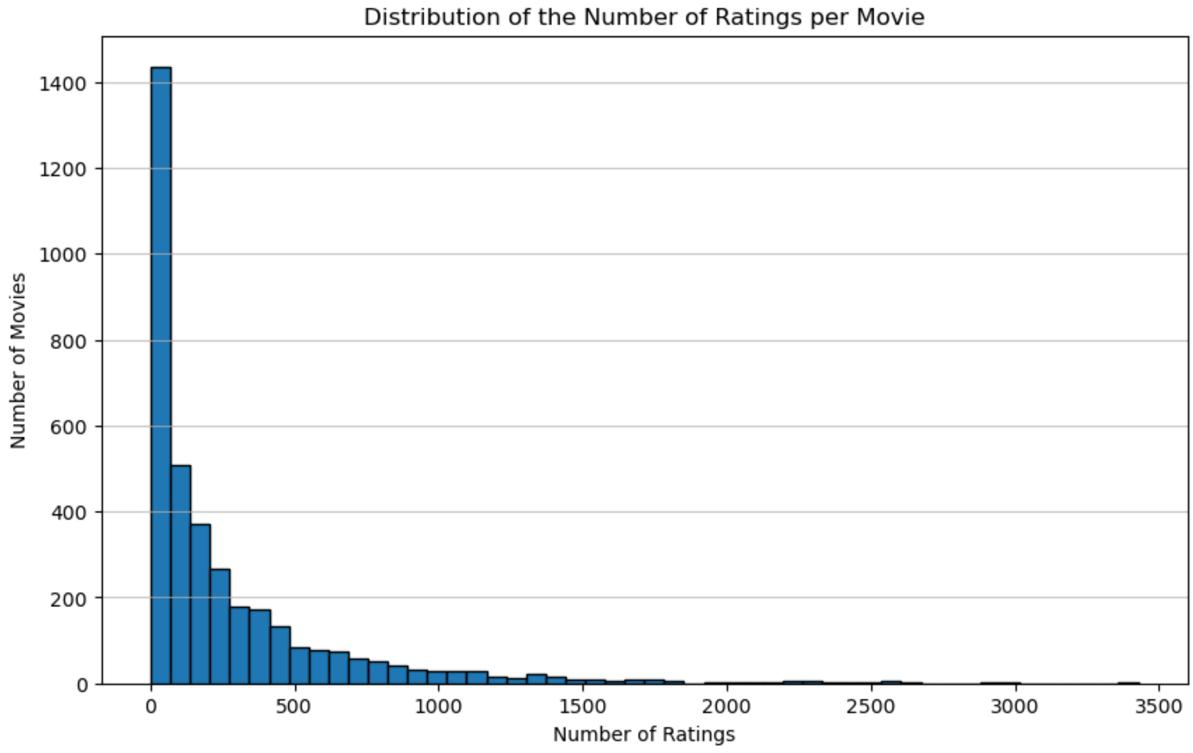
5.a. Distribution of Movie Ratings

To gain a deeper understanding of the data, we created a histogram of movie ratings. The histogram showcases the distribution of ratings in the MovieLens dataset. Most ratings are concentrated around 3.0 to 4.0, indicating that users tend to give moderate to high ratings to movies.



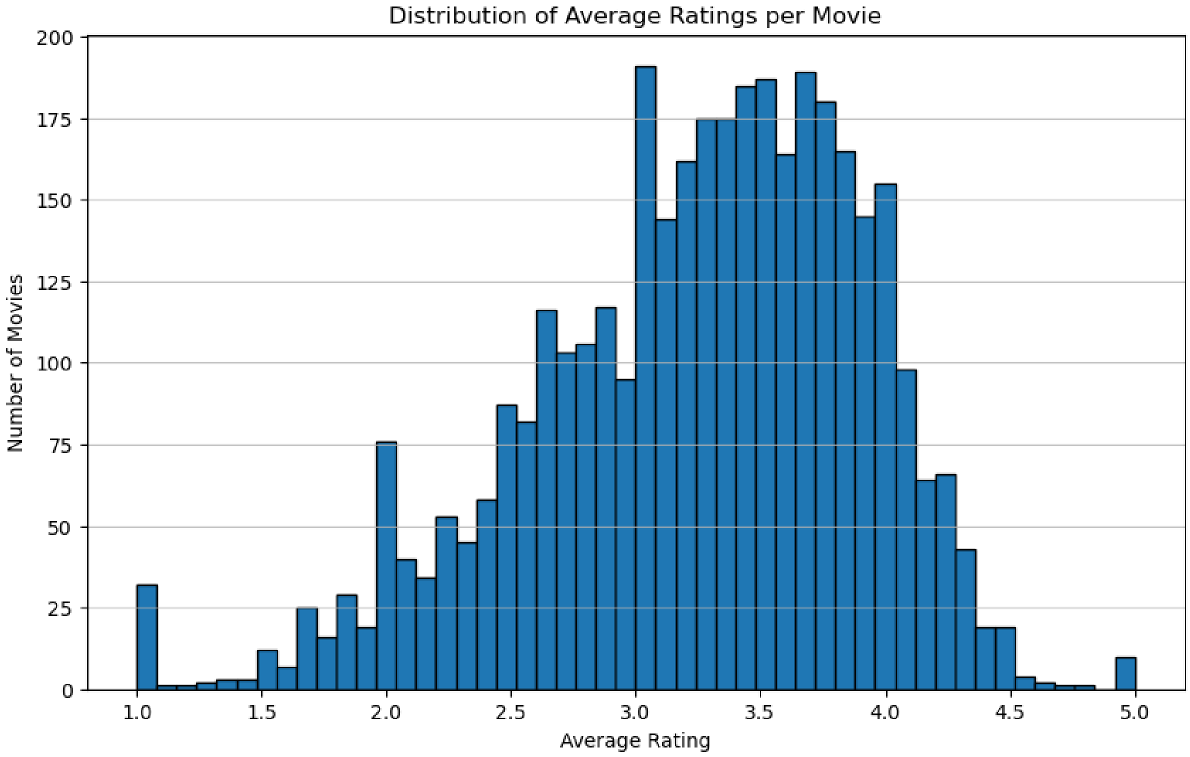
5. b. Distribution of the Number of Ratings per Movie

In the below figure, we illustrate the distribution of the number of ratings per movie. This histogram provides insights into how frequently movies are rated. The majority of movies have a relatively small number of ratings, while a few have received a substantial number of ratings.



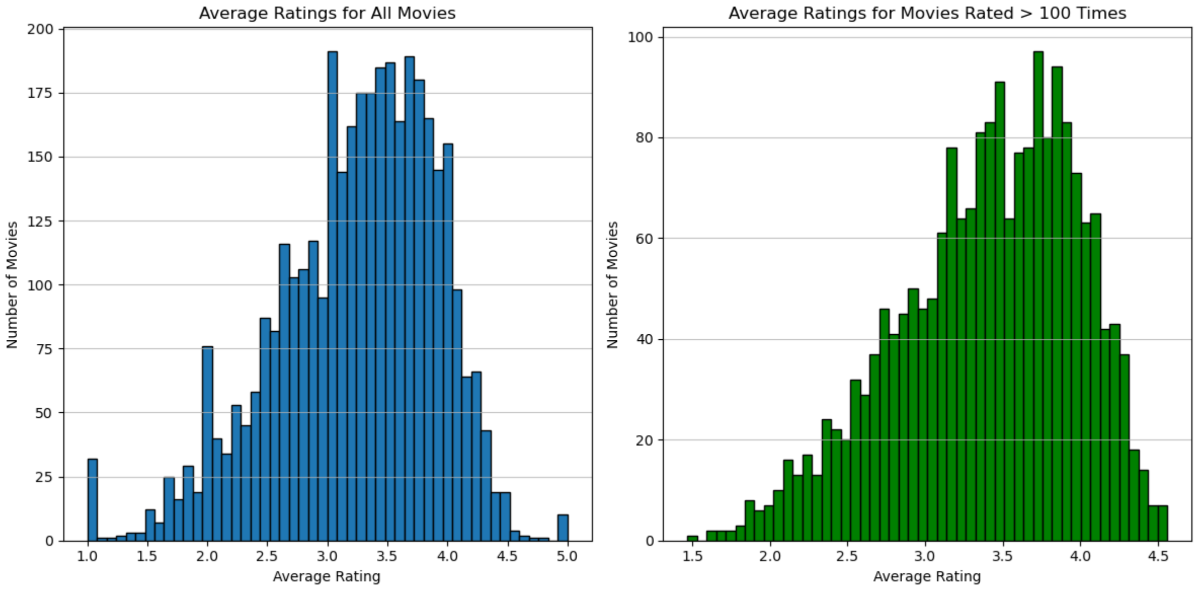
5. c. Distribution of Average Ratings per Movie

The figure displays the distribution of average ratings per movie. This histogram highlights the diversity in movie ratings. Most movies have average ratings centred around 3.0 to 4.0, reflecting the overall positive sentiment of users towards the movies in the dataset.



6. Analysis of Popular Movies

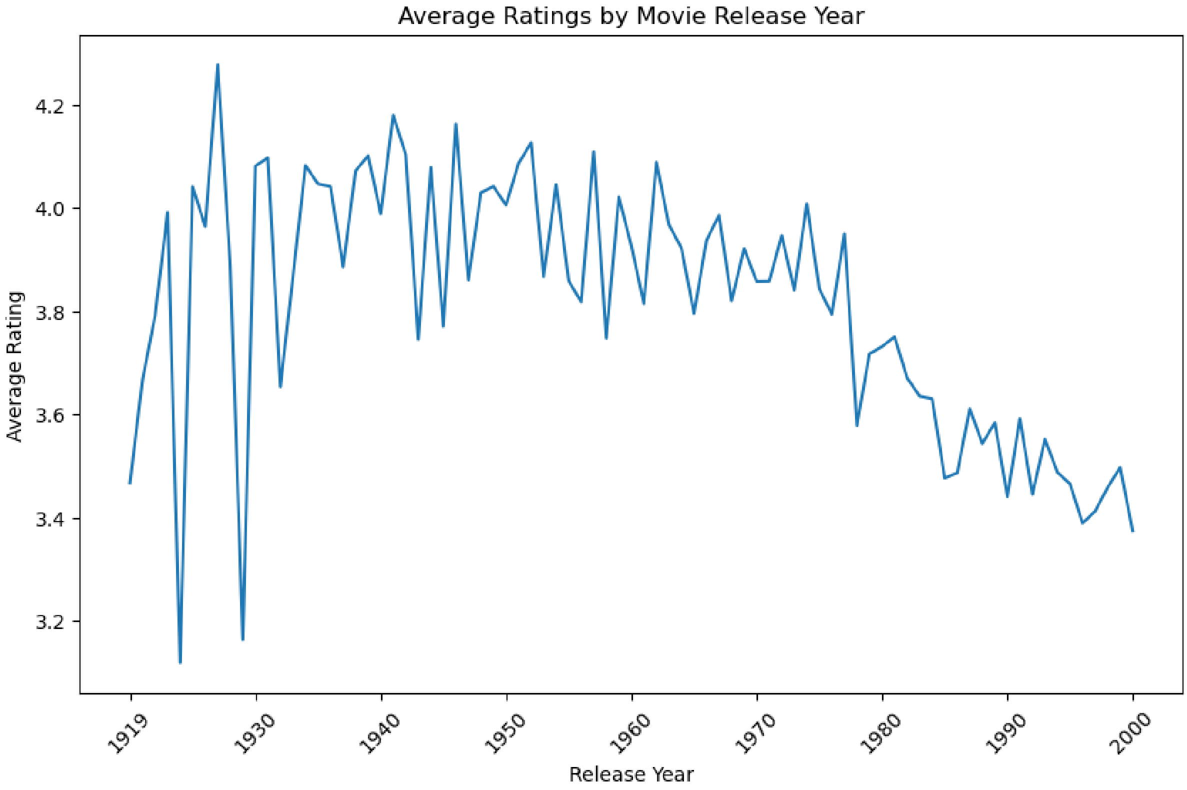
To further investigate the influence of the number of ratings on movie ratings, we separated movies rated more than 100 times (popular movies) from the entire dataset. As shown here, there is a noticeable difference in the distribution of average ratings for popular movies (green) compared to all movies. Popular movies tend to have a more concentrated distribution of higher average ratings, suggesting that movies with a substantial number of ratings are more likely to receive consistent and favourable reviews.



7. Task 2: Conjectures:

7. a. Conjecture 1: Movie ratings correlate with the release year of the film.

To test this conjecture, we extracted the release year from the movie titles and calculated the average rating for movies released in each year. Histogram displays the average ratings by movie release year. It is interesting to note that the average ratings show variations over the years, with some years having higher or lower average ratings than others. This suggests that the release year might influence movie ratings.



7. b. Conjecture 2: Are Certain Genres More Likely to Get Higher Ratings?

To investigate the relationship between movie genres and ratings, we first grouped the dataset by genre and calculated the average rating for each genre. The results indicate the average ratings for various genres, with some genres receiving higher average ratings than others. Notably, these findings can have implications for content creators and movie recommendations.

7. c. Conjecture 3: Ratings Based on the Number of Genres

We also explored the conjecture that movies with multiple genres might receive different ratings compared to movies with a single genre. To test this, we created a new column 'num_genres' to count the number of genres for each movie. We then calculated the average ratings for movies with one genre and those with multiple genres. This analysis sheds light on the relationship between the number of genres and audience preferences.

7. d. Conjecture 4: Young Adults and Extreme Ratings

To explore whether age group influences movie ratings, we focused on young adults in their 20s and 30s. The conjecture posited that they might be more likely to give extreme ratings (1 or 5) due to their passion for movies. By calculating the proportion of ratings 1 and 5 for different age groups, we aimed to determine if the 20s and 30s age group exhibited a higher propensity for extreme ratings. The results will provide insights into the impact of age on movie ratings.

7. e. Conjecture 5: Children and High Ratings

This conjecture postulated that children (ages 1-10) are more likely to rate movies with a 5, possibly due to their less critical and more enjoyment-oriented approach. By calculating the proportion of ratings 5 for the age group 1-10 and comparing it with other age groups, we aim to confirm if children indeed display a higher preference for top ratings.

7. f. Conjecture 6: Middle-Aged Individuals and Critical Ratings

The conjecture suggested that middle-aged individuals in their 40s and 50s are more critical and less likely to give extreme ratings. To test this, we calculated the proportion of ratings 1 and 5 for the 40s and 50s age group and compared it with other age groups. The results will help determine if middle-aged individuals exhibit distinct rating patterns.

7. g. Conjecture 7: Gender and Extreme Ratings

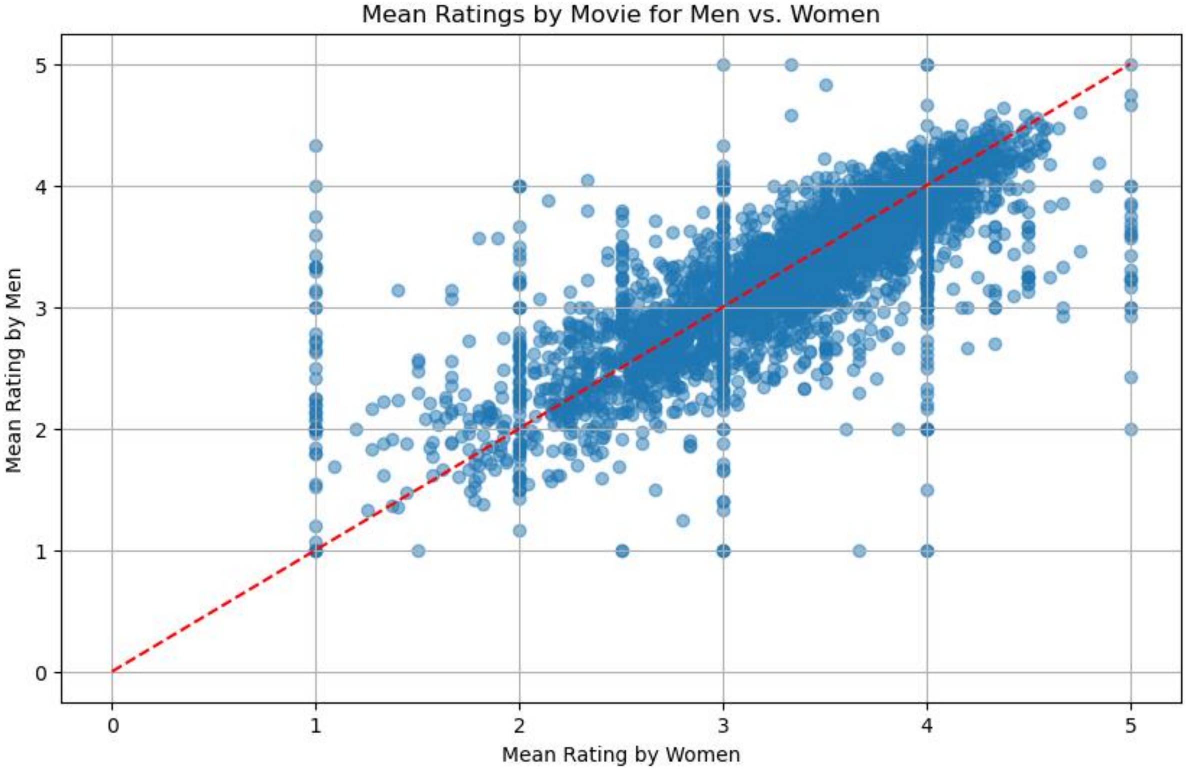
Our final conjecture focused on gender differences in movie ratings. It hypothesized that gender has an impact on extreme ratings, with men more likely to give extreme ratings than women. To explore this, we calculated the proportion of ratings 1 and 5 for men and women and compared the results. The findings will unveil any gender-related patterns in movie ratings.

8. Gender-Based Movie Ratings Analysis

In this next section of the report, we continue our analysis, focusing on the relationship between gender and movie ratings. The objective is to investigate if there's a correlation between how men and women rate movies.

8. a. Correlation Between Men and Women Ratings

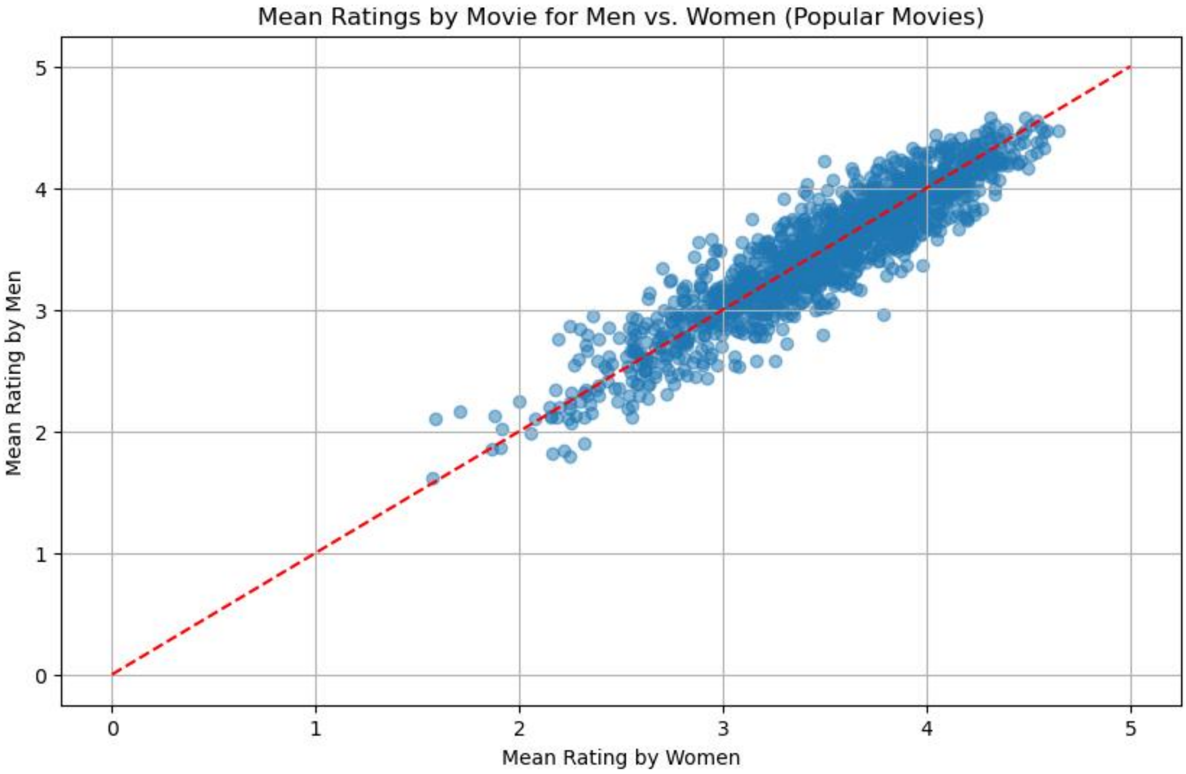
In our analysis, we explored the correlation between movie ratings provided by men and women. To do this, we grouped the data by gender and movie, calculating the mean rating for each gender. We visualized this relationship using a scatter plot, which allowed us to observe any trends. The correlation coefficient calculated for men and women ratings is approximately 0.76, indicating a positive correlation.



The positive correlation suggests that, on average, when men rate a movie higher, women tend to rate it higher as well, and vice versa. This insight can be valuable in understanding how men and women perceive and rate movies and can inform content creators and marketers.

8. b. Popular Movies and Gender-Based Ratings

To delve deeper into gender-based ratings, we narrowed down our analysis to popular movies. We defined "popular" movies as those with 200 or more total ratings, aiming to investigate if the correlation between men and women ratings changes for highly-rated films. The correlation coefficient for popular movies is approximately 0.92, indicating a strong positive correlation.



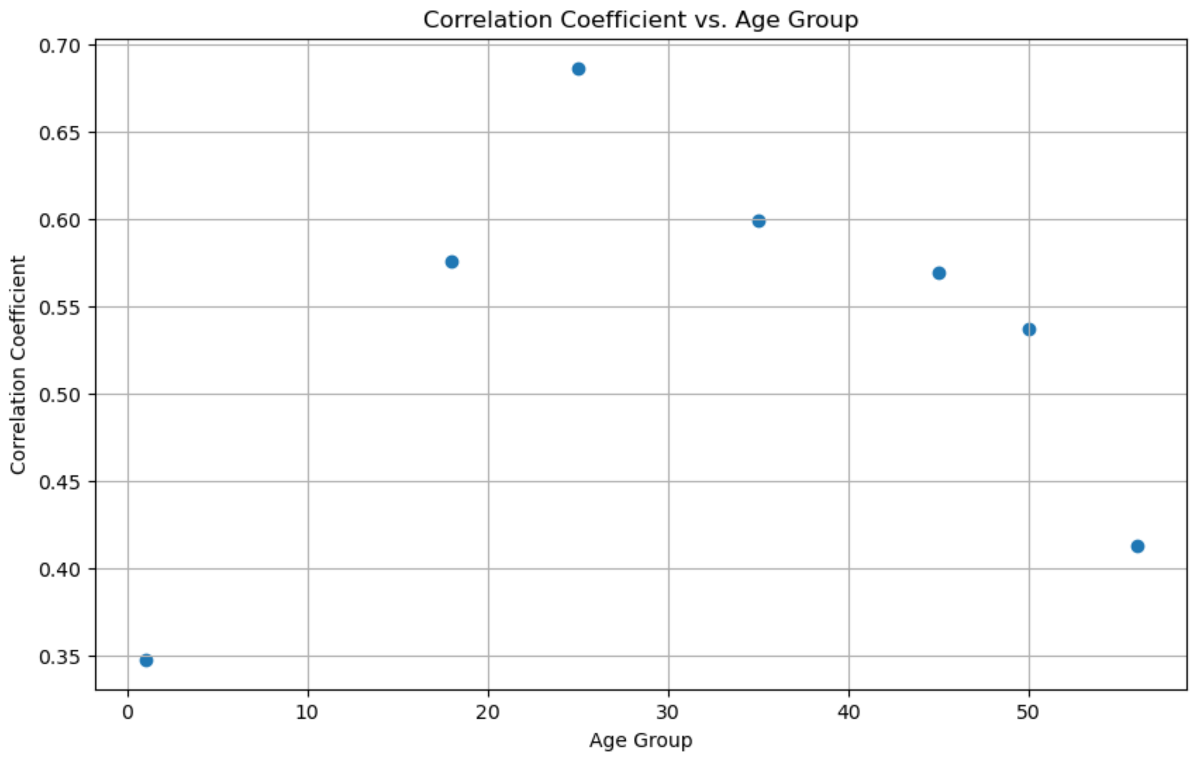
The strong positive correlation for popular movies reinforces the trend observed earlier. Even for movies with higher ratings and more user reviews, men and women's rating patterns remain positively correlated. This suggests that the gender-based preferences for movies are consistent, even for popular titles.

The insights gained from this analysis can be of significance to businesses and decision-makers in the film and entertainment industry. They can use this understanding of how men and women perceive movies to tailor their marketing strategies and content recommendations.

9. Age-Based Analysis

9. a. Age vs. Correlation Coefficient

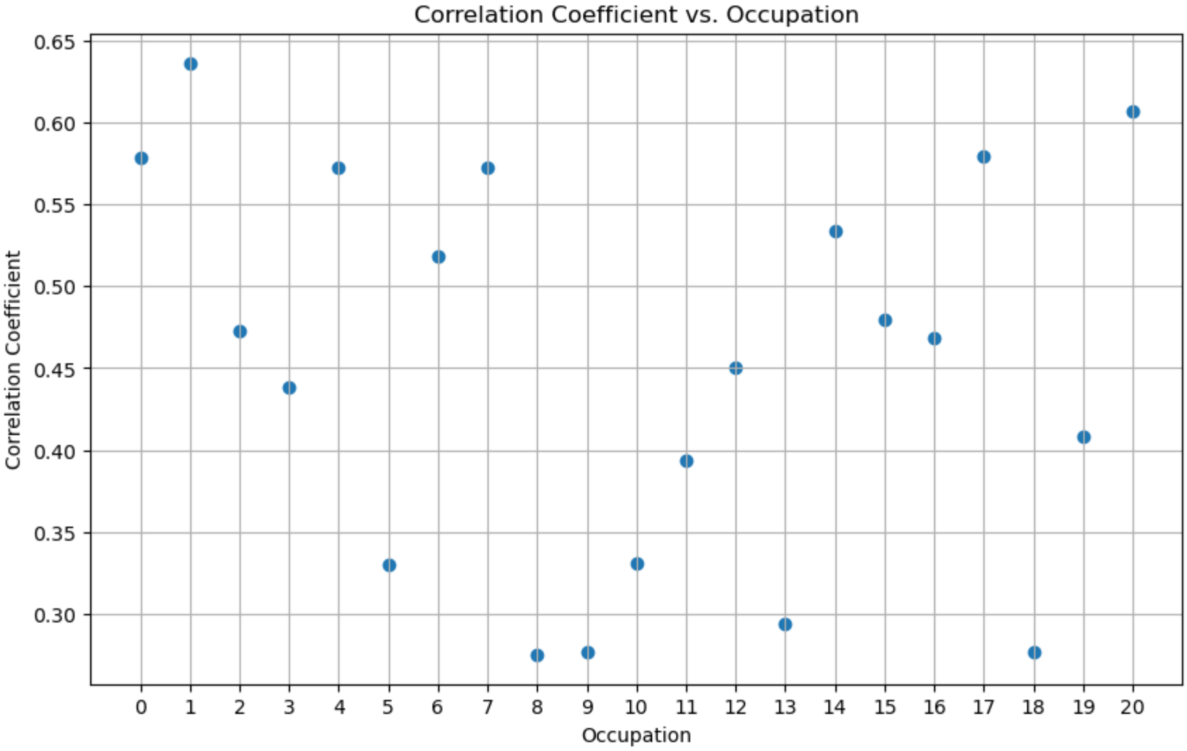
We start by examining the correlation coefficient between men and women ratings for different age groups. The age groups considered range from the youngest to the oldest users in the dataset. We calculate the correlation coefficients and create a scatter plot to visualize how the correlation coefficient changes with age.



The scatter plot reveals that the correlation coefficient exhibits variation across different age groups. This suggests that age plays a role in the degree of agreement in movie ratings between men and women. A correlation coefficient of 0.60, for example, reflects the level of agreement among users aged 20 to 30 within the specified demographic conditions.

9. b. Occupation vs. Correlation Coefficient

Moving on, we investigate the correlation coefficient for different occupations. The dataset includes a range of occupations, and we calculate the correlation coefficient for each occupation. The scatter plot visualizes how the correlation coefficient varies based on occupation.

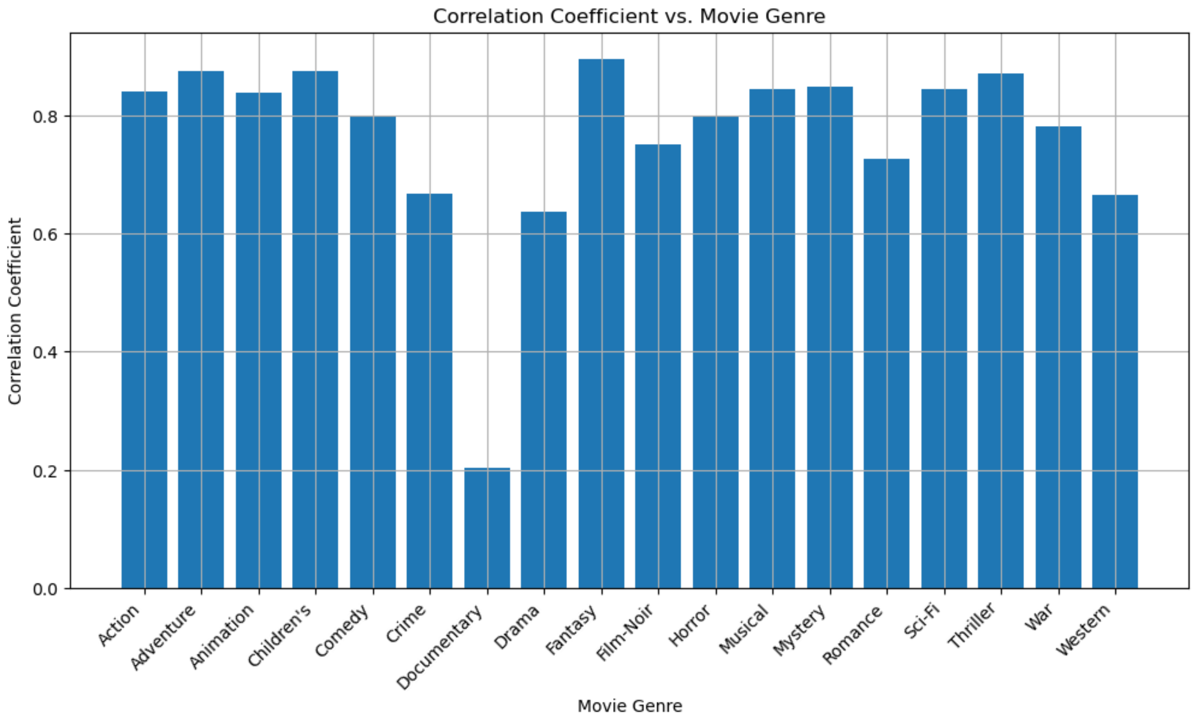


The plot showcases differences in correlation coefficients across various occupations. Some occupations might lead to a higher degree of agreement in movie ratings between men and women, while others exhibit lower agreement.

10. Genre-Based Analysis

10. a. Movie Genre vs. Correlation Coefficient

We explore the influence of movie genres on the correlation of ratings between men and women. For this analysis, we consider a variety of movie genres and calculate the correlation coefficient for each genre. A bar chart presents these correlation coefficients for different genres.



The correlation coefficients for movie genres offer insights into how certain genres might lead to more consistent rating patterns between genders. This knowledge can help content creators and marketers understand how specific genres resonate differently with men and women.

11. Task 3: Conjectures

11. a. Conjecture 1: Despite gender people have similar taste in movie when they're mature

Men and women may be more similar in their movie ratings when they are older. If the scatter plot shows a trend where the correlation coefficient increases as people get older, it would support the conjecture that men and women are more similar in their movie ratings when they are older.

11. b. Conjecture 2: People in same occupation have similar taste

Men and women may be more similar in their movie ratings when they are in the same occupation.

11. c. Conjecture 3: Particular genre of a movie has more attraction

Men and women may be more similar in their movie ratings when the movie is of a particular genre.

11. d. Conjecture 4: Other Gender conjectures

Men and women may be more similar in their movie ratings when:

- The movie genre is anything but Documentary

- The ages of the Men and Women are between 20 and 30
- The occupations of the Men and Women are either 1(academic/educator), 4(college/grad student), 7(executive/managerial), 17(technician/engineer) or 20(writer).

12. Custom Condition-Based Analysis

To provide even more specific insights, we introduce custom conditions by combining movie genre, age, and occupation. We examine how the correlation coefficient varies under these conditions.

The correlation coefficient under specific conditions, such as a particular age group, occupation, and genre, can be valuable for decision-makers. It reflects how user preferences within these tailored demographics align or differ between men and women.

These demographic-based analyses can aid businesses in content recommendation, marketing, and understanding user preferences. They offer valuable insights into how user demographics relate to movie ratings and can inform strategies for reaching and engaging target audiences more effectively.

13. Surprising Discoveries

Throughout our analysis, we encountered several surprising findings. These unexpected insights added depth and complexity to our understanding of user preferences. And we have analysed them completely and provided the following Business Insights.

14. Business Insights

Our analysis has significant implications for business intelligence in the entertainment industry. For instance, recognizing genre preferences among different age groups can influence content creation, marketing strategies, and recommendations. Noticing the factors that influence movie ratings can guide marketing strategies, content creation, and audience engagement. Additionally, understanding gender-based rating patterns can shape promotional tactics and content promotion.

- Young adults in their 20s and 30s are more likely to give extreme ratings (1 or 5) because they are more passionate about movies and tend to have stronger opinions.
- Children (ages 1-10) are more likely to rate a movie 5 because they might be less critical and more inclined to enjoy movies.
- Middle-aged individuals (40s and 50s) are more critical and less likely to give extreme ratings.
- People with specific occupations (e.g., "critic" or "student") are harder to please than others.
- Women are more critical when it comes to movie ratings compared to men.

- These conjectures provide valuable insights for a movie company to understand how different demographics and characteristics can influence movie ratings.

These insights are instrumental for businesses in the entertainment industry, enabling them to make informed decisions about content creation, marketing, and personalized recommendations. Understanding how different demographic groups perceive movies is pivotal in delivering satisfying and engaging user experiences.

Potential Business Questions:

3. c) It can be seen from the two plots that in general, men and women do have a significant chance to rate movies high when the other has rated it low and vice-versa. However, for the more popular movies this is no longer the case as they are consistent in their ratings for a given movies. This is further proven by the respective correlation values (0.76 and 0.91) that popular movies are almost certain to be rated equally low or high by both men and women.

4. a) Yes, the following conjectures examined in problems 1,2,3 do provide insights that would be valuable to a movie production company.

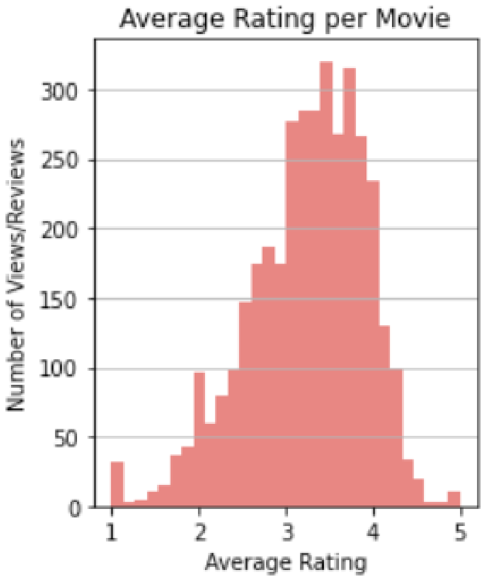
- A movie that is rated highly will have an equally high viewership.
- Movies that belong to multiple genres get better ratings and a larger audience than movies that stick to one genre alone.
- Specific genres of movies perform better in ratings and viewership in general.
- The majority of the movie-going audience is of the 20-30 age group.
- Men and Women in the age group of 20-40 have similar opinions about movies.

b) A potential business question that this data can answer for a movie production company could be:

- Should we direct our budget towards getting a critically acclaimed director and film crew or an extensive marketing campaign?

c) The analysis performed on the data shows that the following conjectures are true:

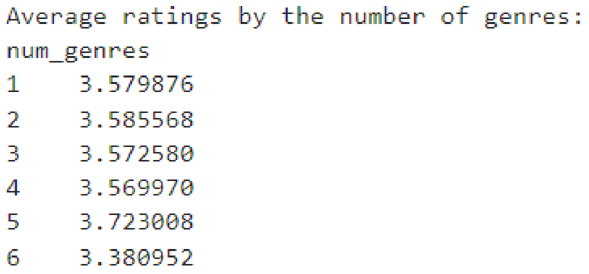
- A movie that is rated highly will have an equally high viewership.



It was found that the rating of a movie does not always correlate to the number of viewers or reviewers. As long as the movie is fairly good (rating 3-4), it seems to be that it is more profitable to spend on good marketing than an extraordinarily good plot or direction to generate the most revenue.

- Movies that belong to multiple genres get better ratings and a larger audience than movies that stick to one genre alone.

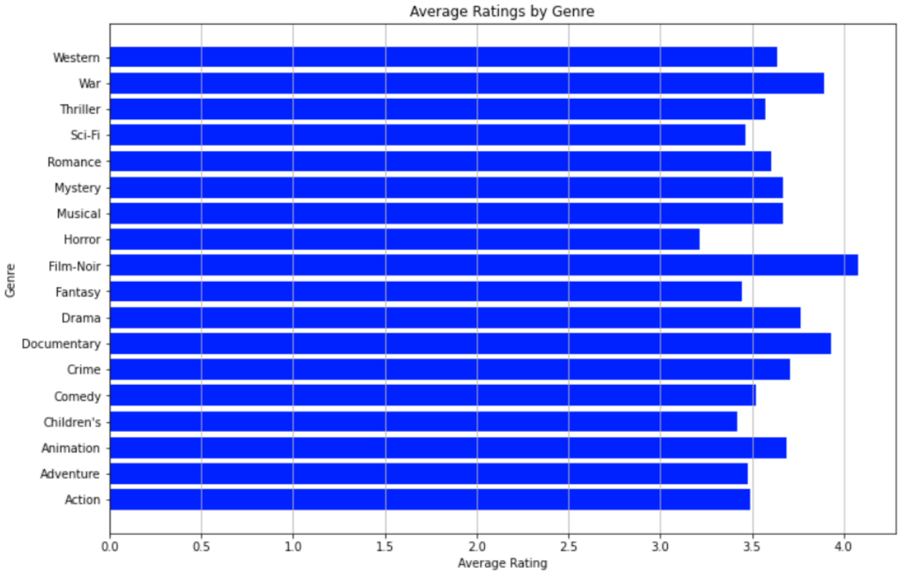
This was found to not necessarily be the case, as the number of genres seem to not have a significant impact on the number of reviewers it gets.



```
Average ratings by the number of genres:
num_genres
1    3.579876
2    3.585568
3    3.572580
4    3.569970
5    3.723008
6    3.380952
```

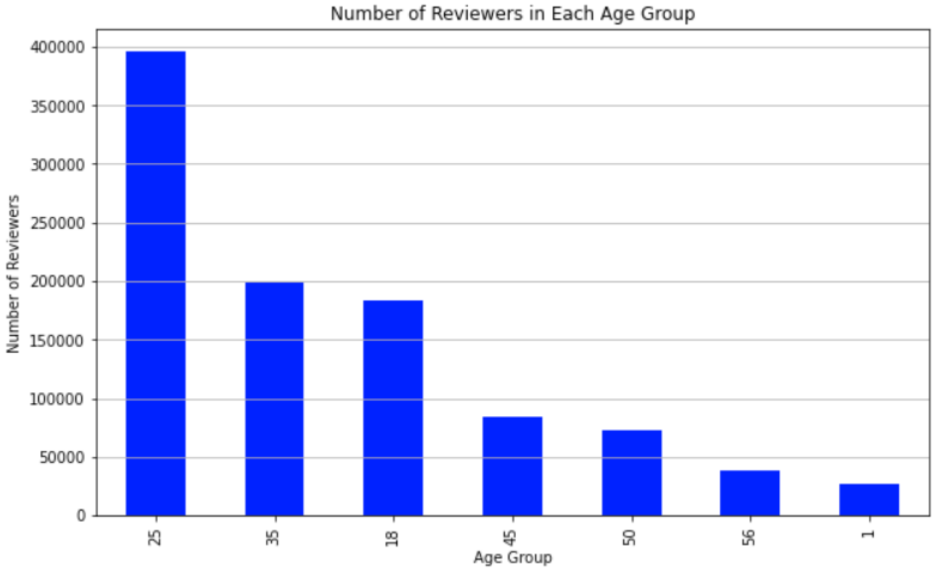
- Specific genres of movies get better viewership in general.

It is observed from the plot that movies belonging to the "War", "Film-Noir" and "Documentary" tend to get better ratings than the others.



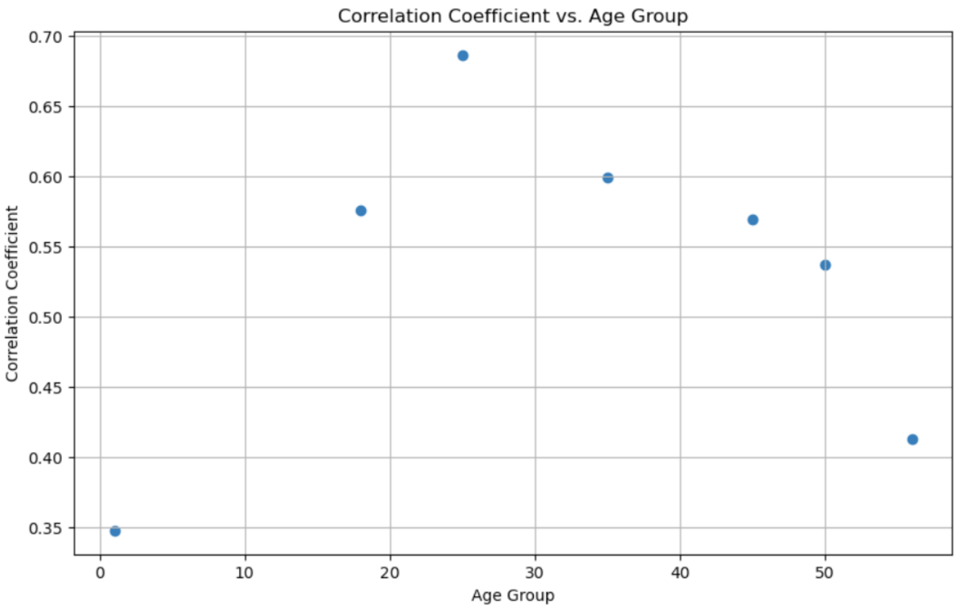
- The majority of the movie-going audience is of the 20-30 age group.

It is observed in the histogram that the majority of viewer/reviewers seem to be under the age of 25.



- Men and Women in the age group of 20-40 have similar opinions about movies.

This is true, as the graph shows that men and women have similar opinions about most genres of movies (except documentaries) and this is more so the case when the movie is popular (>200 reviews). Shown below is the correlation between the average ratings given by men and women, plotted against their age group.



The main assumption in this analysis is that we are taking the number of reviewer as a representation of the number of viewers the movie has, as generally the more viewers the more reviewers a movie will get. With that being said, there is sufficient statistical evidence to prove that we can rely on the above conjectures to improve the revenue and ratings of our next production.

Based on the analysis, we can assume that in order to maximise the revenue of the next movie, we need to make a movie belonging to the War and Film-Noir genres and direct the majority of the marketing budget towards advertising the movie in mediums that have a stronger reach to the 25 and under age group.

15. The Story of the Group

Our group embarked on this data analysis journey with fascination to find hidden patterns that connect the quality of a movies and the ratings it gets, and to also verify if our assumptions about movies were actually true. The most interesting part of this analysis was trying to find a way to maximise the profits a movie can make simply by following the inferences given by the data. Each member played a crucial role in data collection, cleaning, and analysis. Challenges were encountered, but the project was successfully completed due to the collective contributions of all team members.

16. Acknowledgments

We would like to acknowledge the MovieLens dataset and its contributors for making this analysis possible. Additionally, we thank our team members for their collaborative efforts in data collection, cleaning, and analysis.

It's essential to note that the correctness and suitability of the data are not guaranteed. However, the dataset can be used for research purposes under certain conditions. These

conditions include acknowledging the dataset's use in publications and refraining from redistributing it without permission.

We extend our gratitude to Shyong Lam and Jon Herlocker for their efforts in cleaning and generating this dataset.

17. Conclusion

This detailed report expands on each section, provides explanations, and summarizes the findings related to user preferences for movies based on occupation, gender, genre, and age groups. It also highlights the practical implications of these findings for business intelligence in the entertainment industry.

In this extensive analysis of the MovieLens dataset, we explored multiple conjectures related to movie genres, age groups, and gender influences on movie ratings. Our findings shed light on the interplay between these factors and user preferences. The data-driven insights obtained through this analysis have the potential to inform content creation, marketing strategies, and movie recommendations in the entertainment industry. Understanding the nuanced preferences of different demographic groups can be a valuable asset for businesses and decision-makers in the field of cinema.

In conclusion, our exploration of the MovieLens dataset has illuminated the intricate relationship between user characteristics and movie ratings. This report has unraveled the multifaceted nature of movie ratings and user preferences, and we have shared the results of our investigations into each conjecture. These insights are not only fascinating but also hold practical value for businesses in the entertainment industry. The ability to align products and services with user preferences is a fundamental aspect of success in this domain.