MovieLens 1M Dataset: Case Study 3

Introduction

The goal of this project was to analyze the MovieLens dataset, focusing on demographic factors influencing viewer preferences across various movie genres. Through a structured approach, we aimed to uncover insights to inform targeted marketing strategies and content acquisition based on demographic patterns.

Objective of the Study

For the given topic we have identified four main business intelligence aspects:

- **Key resources(Data):** We are providing value from data that is already available.
  \
- **Value proposition (design and performance):** By analysing our customer's most relevant preferences, and their basic personal information we can create and design service recommendation systems specifically dedicated to each customer, thus improving our performance and revenue.

- **Customer relationships (automated services)** : As we develop our recommendation models, we do not need to constantly monitor individual customer behaviour, instead we can focus on other business problems to help expand and improve services.

- **Customer segments(Who are our most important customers?):** As we analyse the data we could identify who our main customers are , such as their average age and profession.

Dataset Description

In this study, we utilise the MovieLens 1M dataset, a comprehensive collection of user ratings across various movies. It includes three main components:

- **Rating Dataset**: This dataset captures explicit user ratings on a scale from 1 to 5, each with an associated timestamp to show when the rating was recorded.
- **Users Dataset**: This dataset provides demographic information on users, detailing:
  - **Gender**: Specified as Male (M) or Female (F).
  - **Age**: Organized into defined age brackets.
  - **Occupation**: Represents users' professions, potentially offering insights into genre preferences.
  - **Zip Code**: Regional information useful for location-based analysis.
- **Movies Dataset**: Contains specific details for each movie, including:
  - **Title**: The movie's name.
  - **Genres**: Movies can belong to multiple genres such as Action, Drama, Comedy, etc.

These datasets, once combined, serve as a rich basis for exploratory data analysis and enable the development of personalised recommendation models.

Relevance of Buisness Inteligence

The relationship between this topic and Business Intelligence (BI) is deeply rooted in understanding customer behaviour and leveraging this knowledge to make data-driven decisions in content marketing and acquisition. We're not just studying raw numbers but transforming them into meaningful insights about customer behavior. This translates directly to practical strategies for audience engagement, content acquisition, and targeted marketing.

1) **Audience-Centered Marketing Strategies**

- **Age Preferences:** Recognizing that specific age groups gravitate toward particular genres lets marketers create customized campaigns. For example, data indicating that older viewers favor genres like Film-Noir and Documentaries points to a targeted approach for these audiences, using content that aligns with their tastes in email or in-app promotions.

- **Gender-Specific Campaigns:** The disproved assumption that women primarily prefer Romantic Comedies reveals the importance of grounding marketing decisions in BI-backed data rather than stereotypes. In fact, with genres like Documentary and War showing higher ratings among women, campaigns can be designed that genuinely reflect these preferences.

2) Data-Driven Content Acquisition

- **Understanding Genre Demand:** The demographic insights guide content acquisition, ensuring that the platform invests in genres that resonate with core audiences. By focusing resources on genres with a higher demand among targeted groups, like Film-Noir or Thriller for older audiences, content providers can increase viewership and loyalty.

- **Informed Future Investments**: BI enables platforms to track evolving viewer preferences over time. If young adults increasingly rate mixed-genre films highly, for example, these insights can inform future content strategies, positioning the platform for sustainable growth.

3) Enhanced User Experience through BI

- **Improved Recommendations:** BI insights allow recommendation systems to be fine-tuned based on actual user preferences, fostering a personalized experience. Recommendations based on gender or age-driven preferences make the platform feel custom-tailored, which directly contributes to user satisfaction and retention.

- **Effective Engagement and Retention**: By curating experiences that reflect demographic preferences, BI-driven platforms can keep users engaged, reducing churn and fostering a loyal viewer base.

Data Analysis Process:

1. Data Import and Preparation

- Dataset Acquisition: We began by downloading the MovieLens dataset containing 1 million ratings, offering a substantial foundation for demographic analysis.
- Loading the Data: Using Pandas, we imported the `ratings.csv`, `movies.csv`, and `users.csv` files into separate DataFrames.
- Data Merging: We combined these DataFrames on common keys (`movieId`, `userId`) to create a single comprehensive DataFrame containing all relevant information for each rating.
- Data Storage: The merged data was saved in HDF5 format for efficient storage and retrieval.

2. Data Exploration

- Initial Checks: We examined the dataset for missing values, data types, and the distribution of key variables, including ratings, genres, age, gender, occupation, and location.
- Demographic Segmentation: Based on age, gender, and occupation, we segmented viewers, allowing us to explore patterns in genre-based ratings and general viewing preferences.
- Basic Analysis: We calculated average ratings for each movie, counted the number of ratings each movie received, and computed demographic-specific ratings by filtering data for subsets like men, women, and different age groups.

3. Formulating Conjectures and Visual Analysis

- Hypothesis Development: From the initial exploration, we developed hypotheses, such as younger audiences rating animation genres higher or technical professionals preferring Sci-Fi. These were validated through data analysis and visualization.
- Histograms and Distributions: We created histograms, scatter plots, and bar charts to observe the distribution of ratings across demographics and genres. Additionally, we normalized histograms by incorporating density to enhance pattern recognition.
- Threshold-based Analysis: We analyzed movies with over 100 ratings separately to identify variations in average ratings distribution.
- Scatter Plot Correlations: Scatter plots revealed a moderate positive correlation between male and female ratings, suggesting a general overlap in genre preferences. We further quantified this with correlation coefficients.

4. Segmentation and Grouping Using Pivot Tables

- Demographic-Specific Averages: Using pivot tables, we calculated average ratings within each demographic group, broken down by genre, enabling detailed analysis of genre preferences by group.
- Overall Ratings: We also examined overall average ratings (without genre breakdown) to see how ratings vary based on gender, age, and occupation.

5. Statistical Testing

- ANOVA Tests: We conducted ANOVA tests to statistically confirm the impact of demographic variables (age, gender, occupation) on ratings. The results revealed that age and occupation had a stronger effect on ratings than gender, supporting our hypothesis regarding the influence of demographic factors on viewer preferences.

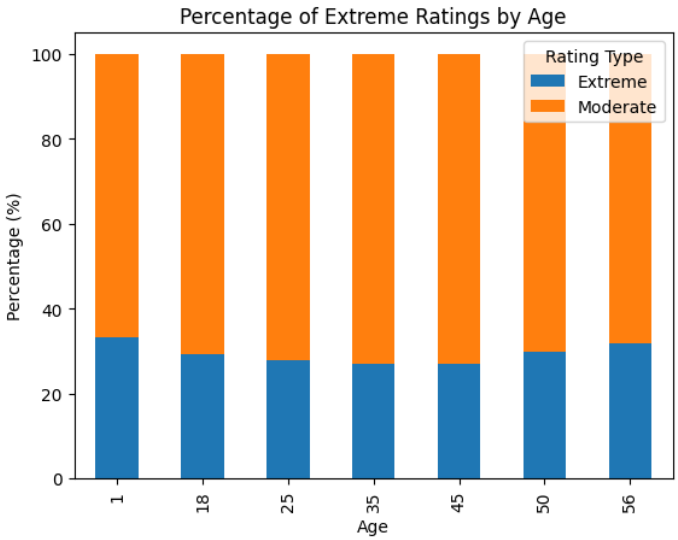
Conjectures

For problem-1:

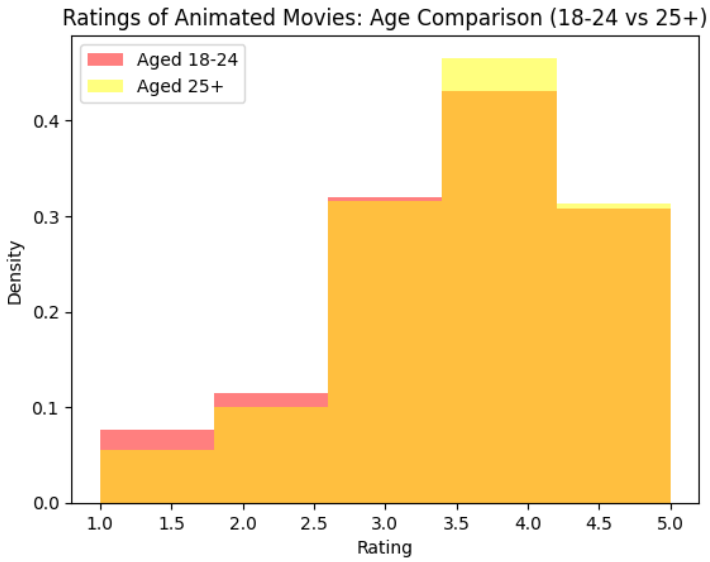
Conjecture-1: People between the ages of 1 and 10 are the easiest to please since they are all young children
Conjecture-2: Women like Romance, Musicals, and Comedy the most
Conjecture-3: Men under 24 like Action, Sci-fi, and Adventure the most.
Conjecture-4: Fantasy, Mystery, and Thriller are equally liked by both men and wome
Conjecture-5: Men over 34 like War, Film-noir, and Documentary the most
Conjecture-6: Users under 18 like Animation and Children's the most
Conjecture-7: Comedy movies receive higher ratings from younger users (under 25) compared to older users.
Conjecture-8: Men rate Action movies higher than women.
Conjecture-9: Users aged 25-34 prefer Film-noir and give higher ratings to this genre.
Conjecture-10: Movies with mixed genres like Action-Comedy or Sci-Fi-Adventure receive higher ratings than single-genre movies.
Conjecture-11: Users between 18-24 rate Horror movies significantly lower than other genres.
Conjecture-12: Older users (35+) rate Drama and Thriller movies higher than younger users.
Conjecture-13: Men in technical occupations (Occupation codes 1-4) tend to give higher ratings to Sci-Fi and Adventure movies.
Conjecture-14: Women over 40 prefer Romantic Dramas and give them the highest ratings
Conjecture-15: Users in non-technical occupations (Occupation codes 5-21) rate Children's and Family movies higher than technical users.
Conjecture-16: Women under 18 prefer Musical and Animation genres, and give them the highest ratings.

For problem-2- Validating conjectures through histograms:

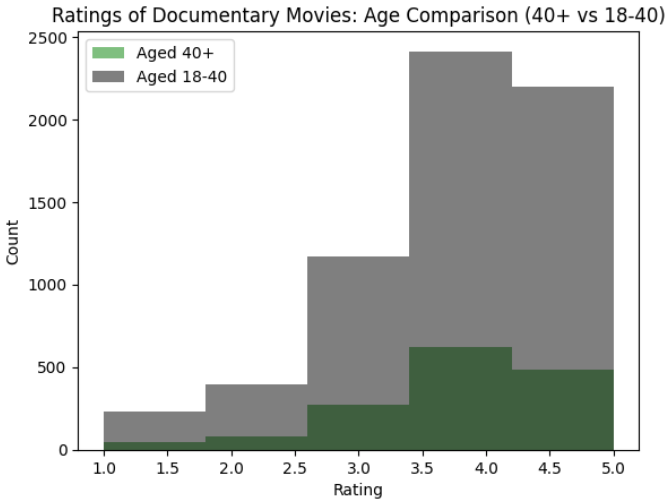
Conjecture-1: Younger people give more extreme ratings



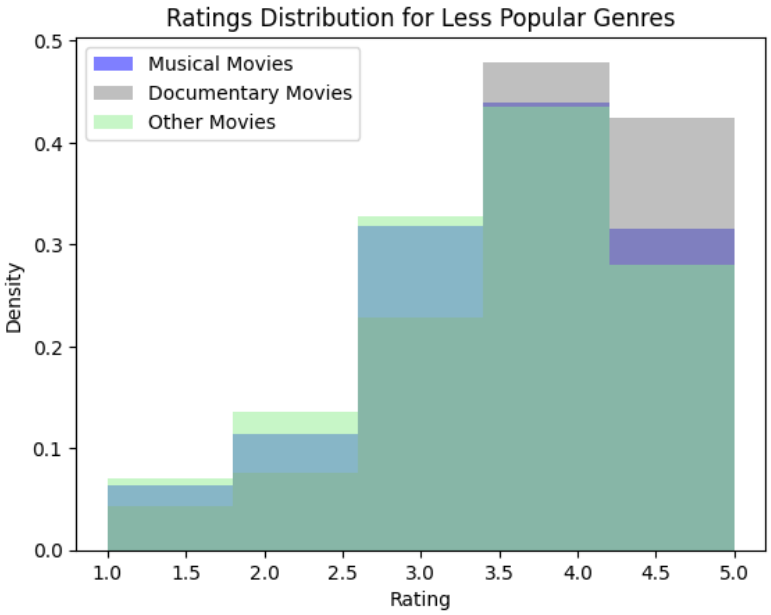
Conjecture-2: Users aged 18-24 are more critical of Animated movies compared to older age groups.



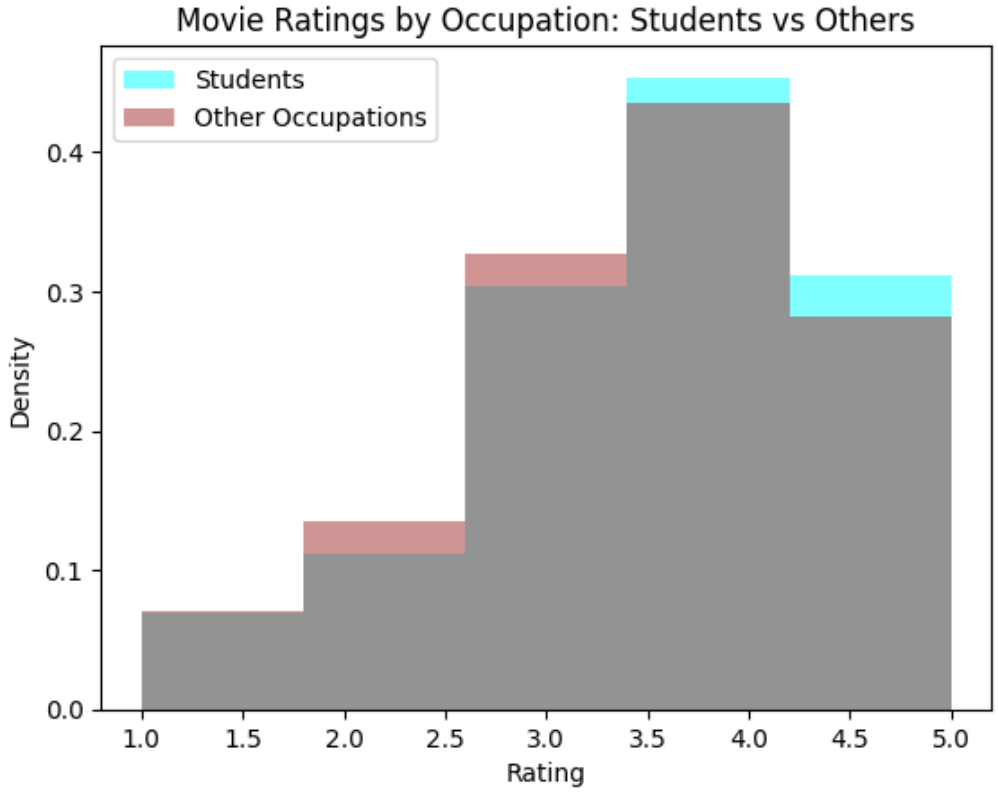
Conjecture-3: Users over 40 give higher ratings to Documentaries compared to younger users.



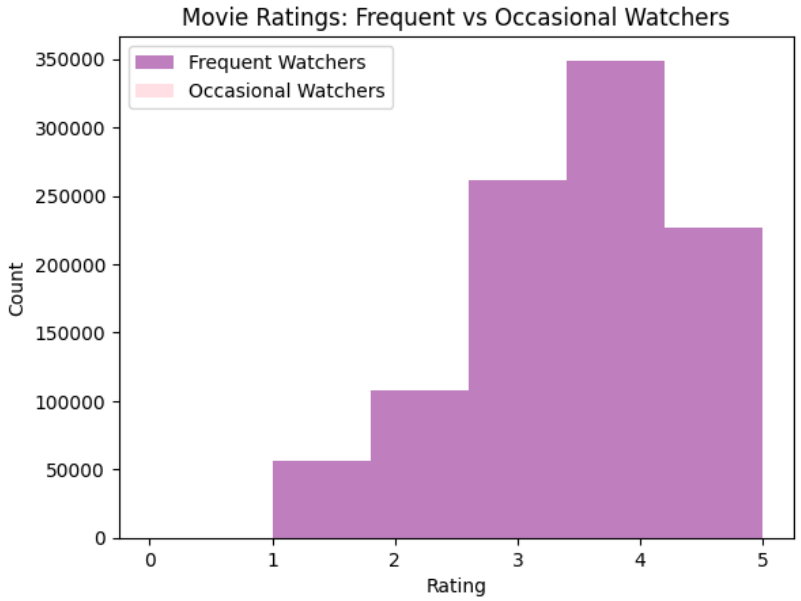
Conjecture-4: Users with higher ratings are more likely to watch movies from less popular genres, such as Musical and Documentary.



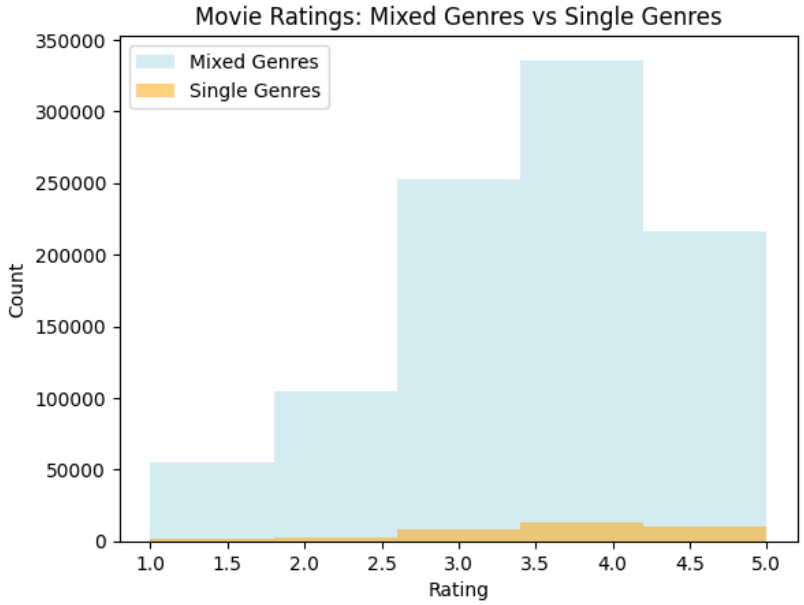
Conjecture-5: Users in different occupations have varied preferences, showing that students rate movies significantly higher than other occupation groups.



Conjecture-6: Users who frequently watch movies (more than 10 ratings) provide higher ratings compared to those who watch occasionally.

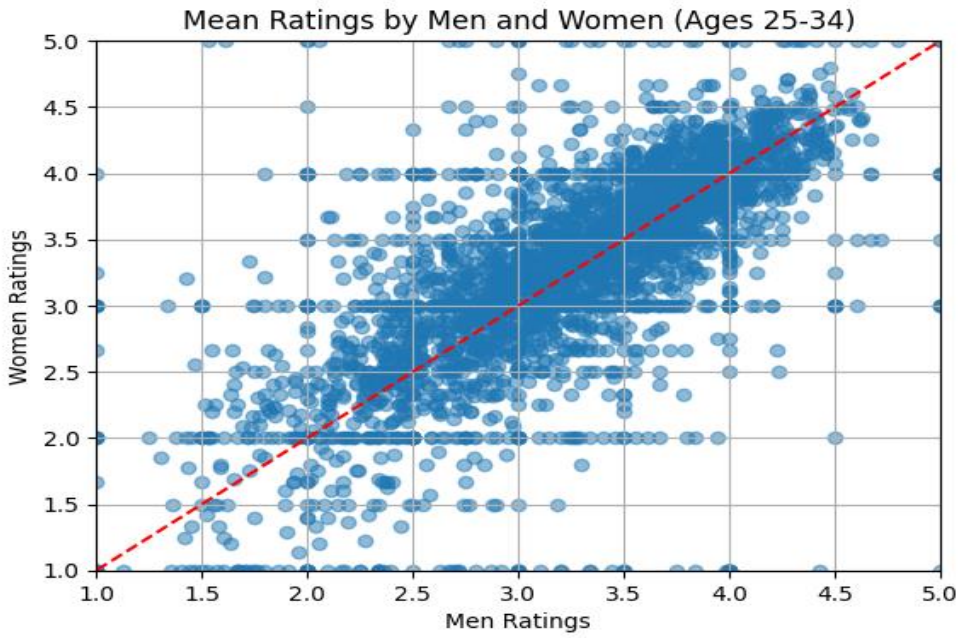


Conjecture-7: Users with higher average ratings are more likely to watch movies with mixed genres compared to single-genre movies.



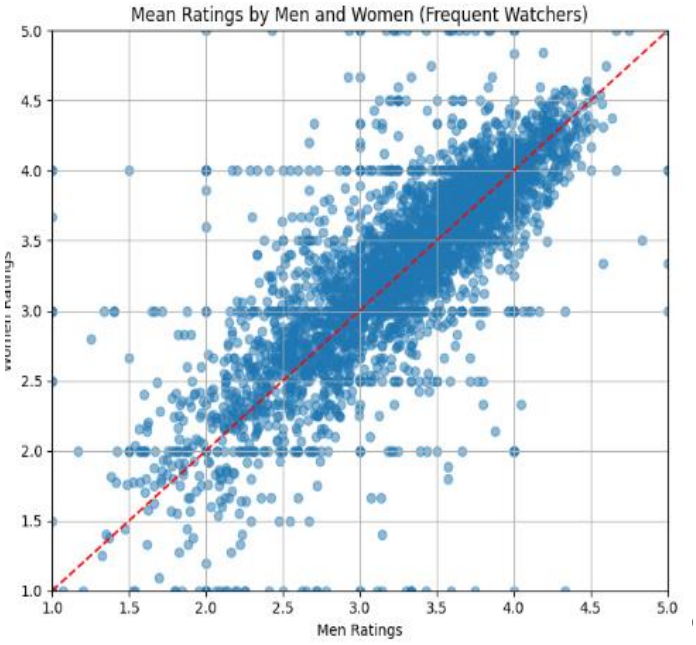
For problem-3 Validating conjectures through scatter points:

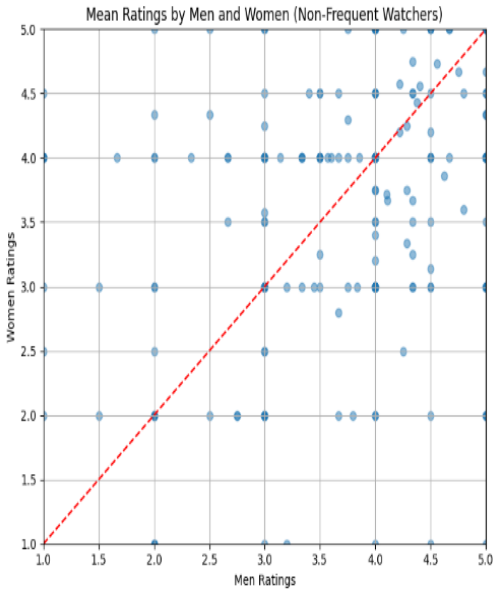
Conjecture-1: Under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.




```
Correlation between men and women ratings (Ages 25-34): 0.69
```

Conjecture-2: Users with frequent movie-watching habits (more than 10 ratings) provide more consistent ratings between genders.

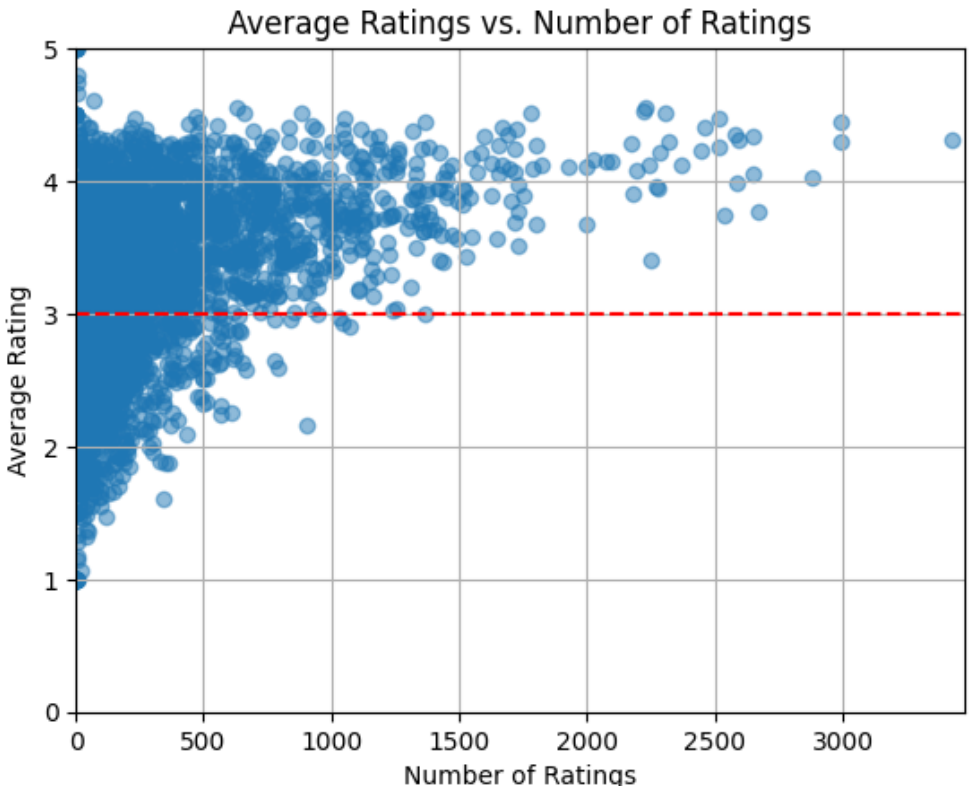




Conjecture-3: Ratings for action movies show a stronger correlation between men and women than ratings for dramas.

```
Correlation between men and women ratings (Action Movies): 0.84
Correlation between men and women ratings (Drama Movies): 0.64
```

Conjecture-4: Movies with higher ratings tend to have fewer ratings (indicating a preference for niche films).



Pivot table for final analysis

Occupation_Category	Gender	Age_Group	Rating
Administrative	F	Under 18	3.81
		25-34	3.59
		35-44	3.73
		45-55	3.81
		55+	4.00
	M	Under 18	3.58
		25-34	3.54
		35-44	3.59
		45-55	3.70
		55+	3.86
Creative	F	Under 18	3.15
		25-34	3.45
		35-44	3.66
		45-55	3.77
		55+	4.13
	M	Under 18	3.50
		25-34	3.49
		35-44	3.55
		45-55	3.64
		55+	3.89
Non-Technical	F	Under 18	3.50
		25-34	3.62
		35-44	3.70
		45-55	3.63
		55+	3.94
	M	Under 18	3.45
		25-34	3.46
		35-44	3.57
		45-55	3.61
		55+	3.75
Sales	F	Under 18	3.53
		25-34	3.59
		35-44	3.67
		45-55	3.68
		55+	3.08
	M	Under 18	3.49
		25-34	3.59
		35-44	3.76
		45-55	3.90
		55+	3.31
Services	F	Under 18	3.41
		25-34	3.96
		35-44	3.24
		45-55	3.96
	M	Under 18	3.61
		25-34	3.58
		35-44	3.48
		45-55	3.38
		55+	4.04
Student	F	Under 18	3.52
		25-34	3.70
		35-44	3.13
		45-55	3.70
	M	Under 18	3.53
		25-34	3.51
		35-44	3.88
		45-55	3.28
		55+	3.28
Technical	F	Under 18	3.62
		25-34	3.67
		35-44	3.66
		45-55	3.81
		55+	3.77
	M	Under 18	3.60
		25-34	3.59
		35-44	3.64
		45-55	3.71
		55+	3.60

Anova test result:

	sum_sq	df	F	PR(>F)
C(Occupation_Category)	2.870896	6.0	5.117547	3.342685e-05
C(Gender)	0.182103	1.0	1.947653	1.630977e-01
C(Age_Group)	5.259806	5.0	11.251110	5.777687e-09
Residual	112.385119	1202.0	NaN	NaN

The ANOVA tests confirmed that age and occupation significantly affected ratings, with age accounting for approximately 4.36% of the rating variance. This highlights age as a primary factor for genre preferences.

Effect Sizes: Age and occupation had a more substantial effect on genre ratings than gender, supporting targeted marketing strategies aimed at specific age groups and occupational segments for increased engagement.

Our Analysis of the data through these conjectures- BUSINESS REPORT

Our combined analysis of movie preferences across demographics gives key insights into viewing behaviors and challenging some common assumptions. Initially, we assumed that younger audiences, particularly children under 18, would rate movies more favorably, but this proved incorrect as no age group exhibited significantly more extreme ratings. However, children and older adults tended to show slightly more extreme ratings compared to other groups. Interestingly, users aged 18-24 were found to be more critical of Animated movies than older age groups, underscoring the importance of high-quality content, even for younger demographics.

For users over 40, the data revealed that they tend to give higher ratings to Documentaries compared to younger viewers, mostly confirming our conjecture. Gender-based insights also challenged stereotypes about genre preferences. Contrary to the idea that women primarily gravitate toward Romance and Comedy,we found that they rated Film-Noir, Documentary, and War genres most favorably, indicating a preference for serious and reflective genres. This pattern was similarly observed among men under 24, who rated Film-Noir, War, and Documentary genres higher than the stereotypically favored Action and Sci-Fi. This suggests a shared interest in thought-provoking storytelling among young adults, regardless of gender.

Genre preferences within age and occupation groups added further complexity to our initial assumptions. Users aged 25-34 rated Film-Noir highly across both genders, reinforcing the genre's broad appeal. Additionally, viewers in technical fields (such as those in occupations coded 1-4) showed no particular preference for Sci-Fi and Adventure movies, countering the stereotype of tech enthusiasts favoring these genres. On the other hand, women over 40 did conform to expectations, rating Romantic Dramas highly, highlighting the importance of targeted content within specific demographics rather than relying on generalizations about genre popularity by profession or age group.

Conversely, as hypothesized, women under 18 showed a marked preference for Musicals and Animation, suggesting that age-targeted marketing could resonate well for these genres. Moreover, users under 18 demonstrated a high affinity for Animation and Children's content, reinforcing the popularity of these genres with younger viewers.

Analyzing the impact of viewing frequency, we found that users with frequent movie-watching habits (more than 10 ratings) did not consistently provide higher ratings compared to occasional viewers, indicating that rating frequency does not significantly affect average rating. We also speculated that users with higher average ratings might prefer mixed-genre movies, but this conjecture was disproven. However, frequent movie-watchers tended to provide more consistent ratings between genders, supporting the idea that they exhibit more stable preferences.

Lastly, gender-related trends revealed a strong positive correlation between the ratings of men and women (correlation coefficient of 0.76). Users aged 25-34 exhibited the highest correlation (0.69), suggesting that this age group is less influenced by gender differences. We confirmed that ratings for action movies

correlate more strongly between men and women than ratings for dramas. Lastly, our hypothesis that movies with higher ratings tend to have fewer overall ratings, indicating a niche preference, was largely false, though there were a few instances where this held true.

Collectively, these findings offer valuable insights for content strategy, underscoring the importance of nuanced, demographic-specific marketing approaches that account for unexpected preferences across age, gender, region, and occupation.

Business Questions Our Data Can Answer

1. Which genres should we invest in to appeal to specific age demographics?
- Insights indicate that genres like Romantic Drama appeal to older viewers, while Film-Noir and War attract younger male viewers. Tailoring genre investments can help reach key demographics.

2. How can we target marketing campaigns based on occupation?
- Since occupation significantly impacts movie preferences, campaigns could highlight genres that resonate with specific occupational groups (e.g., technical professionals showing interest in Documentary films).

3. What types of content are likely to attract younger vs. older viewers?
- Younger audiences tend to prefer more intense genres, like Film-Noir, while older audiences enjoy Comedy, Drama, and Thriller. This helps in curating age-specific content for better viewer engagement.

4. Should we prioritize gender-specific content or focus more on age and occupation?
- Given that gender has a minimal impact on movie ratings compared to age and occupation, prioritizing age and occupation-specific content strategies would likely yield a better return on investment.

5. How can we use demographic data to guide content recommendations for our streaming platform?
- Understanding that age and occupation have a stronger influence on preferences than gender enables more personalized recommendations, enhancing user satisfaction and retention on the platform.

Our BI question

What demographic factors—age, gender, occupation—most significantly influence viewer preferences for specific genres, and how can this information inform targeted marketing campaigns and content acquisition strategies?

Few Relevant Conjectures To Help Answer the Business Question:
- Men under 24 prefer Film-Noir, War, and Documentary genres.
- Men over 34 also favor War, Film-Noir, and Documentary genres.
- Comedy movies receive higher ratings from older users.
- Users aged 18-24 rate Horror movies lower than other genres.
- Older users (35+) give higher ratings to Drama and Thriller movies.

- Women over 40 prefer and rate Romantic Dramas highly.
- Although our initial hypothesis on occupation's influence was inconclusive, further analysis of ratings by occupation revealed valuable insights for targeted campaigns(pivot table).

By analysing the data from these and other conjectures, our final conjecture is that
age and occupation are more significant determinants of movie ratings than gender.

This conclusion is supported by our findings, which show that while gender differences exist, they are less impactful than age and occupation. This can be observed by looking at our conjectures focused on these three things and the cases where our conjectures were true or false wrt to these three. We can take a look at the pivot table above on how occupation, gender and age effects genre and see that on an average, gender has the least impact overall. The correlation coefficient between male and female ratings is 0.76, indicating a strong positive correlation and suggesting that men and women generally hold similar opinions on movies.

To reinforce this, we conducted an ANOVA test to determine the relative significance of each factor. In particular, the p-value for age and occupation is notably low (typically below 0.05), indicating that the observed differences in ratings by these factors are statistically significant and unlikely due to random chance

Effect Sizes (Eta-Squared) for Each Factor:
- *Occupation Category:* 2.38%
- *Gender:* 0.15%
- *Age Group:* 4.36%

These values highlight that age group has the most significant effect on movie ratings, followed closely by occupation, while gender has a minimal impact by comparison..

In summary, our conjecture is well-supported: age and occupation significantly shape movie preferences, while gender plays a comparatively minor role.
This insight can be instrumental for a movie company in tailoring content strategies. For instance, understanding that men under 24 prefer more intense genres like Film-Noir and War, while women over 40 favor Romantic Dramas, allows for focused marketing. Similarly, the varied preferences across occupations can guide targeted content recommendations. By leveraging these findings, we can make data-informed decisions for marketing, content acquisition, and personalized recommendations that align closely with audience demographics.

Challenges and Limitations we faced with our analysis:

1. Limited Demographic Scope

While the dataset includes demographic factors like age, gender, and occupation, it lacks other potentially influential factors such as education level, household income, or cultural background. These additional factors could provide a more comprehensive view of user preferences, especially given that socio-economic factors often correlate with media consumption patterns.

2. Data Imbalance Across Demographics

Certain demographics are overrepresented or underrepresented, especially in age and occupation categories. For instance, certain age groups may have fewer users, limiting the robustness of insights derived from these segments. This imbalance can introduce biases, potentially skewing results toward the preferences of more populous demographic groups.

3. Genre Overlap and Classification Challenges

Movies frequently belong to multiple genres, making it difficult to isolate genre preferences accurately. For example, a movie classified as both "Action" and "Adventure" may appeal to audiences with either genre preference, complicating analysis of pure genre preferences. This overlap may reduce the clarity of results for genre-specific targeting.

4. Static Dataset and Evolving Preferences

The MovieLens dataset is a snapshot in time and does not capture the evolution of user preferences or emerging genres. Audience tastes shift over time, particularly with new trends and releases, limiting the ability to apply findings dynamically. This static nature makes it challenging to predict future preferences based on current data alone.

5. Generalization to Broader Audiences

The MovieLens dataset is derived from a specific sample that may not fully represent the global population. This is especially important to consider when using findings for broader marketing strategies, as preferences in this sample may not entirely reflect the preferences of a more diverse, international audience.

6. Absence of Viewing Context Data

User ratings lack contextual data, such as viewing environment or mood, which could impact ratings and preferences. For example, a viewer's rating may differ if they watched a movie alone versus in a group setting. This absence of context may obscure underlying factors that influence ratings and preferences.

7. Potential Bias in Self-Reported Ratings

Ratings in the dataset are self-reported, which can introduce bias. Users may rate movies higher or lower due to social desirability or personal mood, rather than objective enjoyment. This subjectivity can affect the accuracy of preferences inferred from the data.

8. Limited Regional Data

While location information is provided in the form of ZIP codes, it lacks finer geographic details like country or urban-rural distinction. These regional characteristics often influence genre preferences and could offer a more granular understanding of how location impacts viewer tastes.

The group's story:

Due to our individual busy schedules, our first meeting kept getting postponed, leading us to opt for an online meeting to kickstart our case study. We began by introducing ourselves and familiarizing ourselves with one another. Starting with a discussion about movies, we shared our personal preferences and opinions, which was enjoyable and helped us connect further.

Next, we delved into the dataset to understand it better. Each of us shared our initial assumptions, which we noted as conjectures, and we brainstormed various business questions. We believed that the best approach was to collaborate on the entire case study rather than dividing the tasks. This way, we could all contribute meaningfully, share our findings, and compile our work cohesively.

Our collaborative method proved effective, as we gained significant insights into the dataset by sharing our analyses. Each member presented intriguing conjectures and analyses. When one of us had an interesting idea but struggled with the coding, we supported each other in troubleshooting. Sam suggested incorporating density into our histograms for a more comprehensive view, aiding in the comparison of unbalanced data. Aries proposed using occupation codes to classify roles into technical, non-technical, and creative categories to examine their impact on ratings, providing us with a fresh perspective. Kevin formulated several business questions and developed a conjecture for our final BI question. Kalyan introduced the idea of creating pivot tables for a more concise presentation and suggested using ANOVA tests to validate our conjecture further.

In the end, we divided sections of the report and the presentation among ourselves to efficiently compile our\
findings.

Conclusion

This project used the MovieLens 1M dataset to explore how demographic factors influence movie genre preferences, to inform targeted marketing and content acquisition strategies. By analyzing the data through various demographic lenses, we uncovered insights that can be valuable for personalizing user experiences in media platforms.

Key findings indicate that age and occupation are the most significant predictors of genre preferences, while gender plays a moderate role. These insights provide a foundation for audience-targeted marketing strategies, helping to tailor content recommendations and advertising based on demographic segments. Despite challenges such as genre overlap, demographic imbalances, and limited contextual data, the study reveals meaningful trends that can enhance engagement and satisfaction among users.

By leveraging demographic data, media companies can make data-driven decisions that align content offerings with audience preferences, increasing the likelihood of user retention and satisfaction. However, as preferences are not static, these strategies should be revisited and adjusted periodically to remain relevant to shifting tastes and trends.