Data Alchemists

Case Study 3:

Analyzing data from MovieLens

Team:

1. Gadiputi Madhava Kalyan
2. Kevin Zachariah Peter
3. Aires Miguens
4. Sylvester Krampah

MovieLens 1M Dataset: Demographic Analysis of Viewer Preferences

This project analyzed the MovieLens 1M dataset to uncover insights into how demographic factors influence movie genre preferences. The goal was to inform targeted marketing strategies and content acquisition based on viewer demographics. By examining ratings data alongside user information on age, gender, occupation, and location, we aimed to identify patterns that could guide personalized recommendations and improve user engagement on media platforms.

Dataset Overview and Objectives

The MovieLens 1M dataset contains 1 million ratings from 6,000 users on 4,000 movies. It includes three main components: ratings data, user demographics, and movie information. Our objective was to analyze this data to uncover key business intelligence aspects, including identifying valuable data resources, improving performance through personalized recommendations, automating customer services, and segmenting our most important customers.

Ratings Data

1 million movie ratings on a 1-5 scale

User Demographics

Age, gender, occupation, and ZIP code

Movie Information

Title and genres for 4,000 movies

Key Objective

Uncover demographic influences on preferences

Data Analysis Process

Our analysis followed a structured approach, beginning with data import and preparation. We merged ratings, user, and movie data into a single comprehensive DataFrame. Next, we conducted exploratory data analysis, examining distributions of key variables and segmenting viewers based on demographics. We formulated conjectures about viewing preferences and validated them through visual analysis using histograms, scatter plots, and pivot tables.

1

Data Import and Preparation

Merged ratings, user, and movie data

2

Exploratory Data Analysis

Examined distributions and demographic segments

3

Hypothesis Development

Formulated conjectures about preferences

4

Visual Analysis

Created histograms, scatter plots, and pivot tables

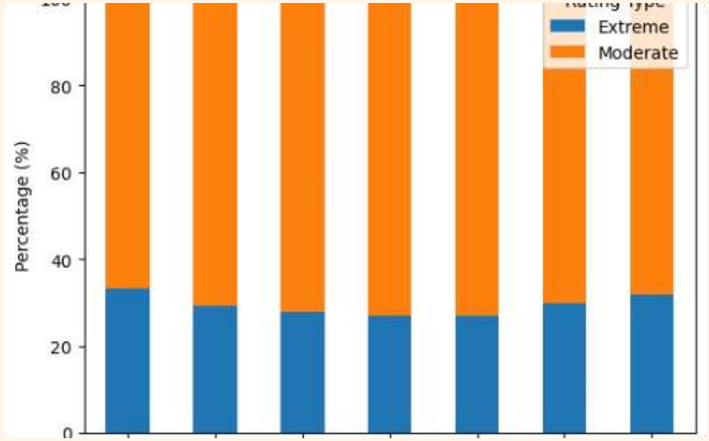
5

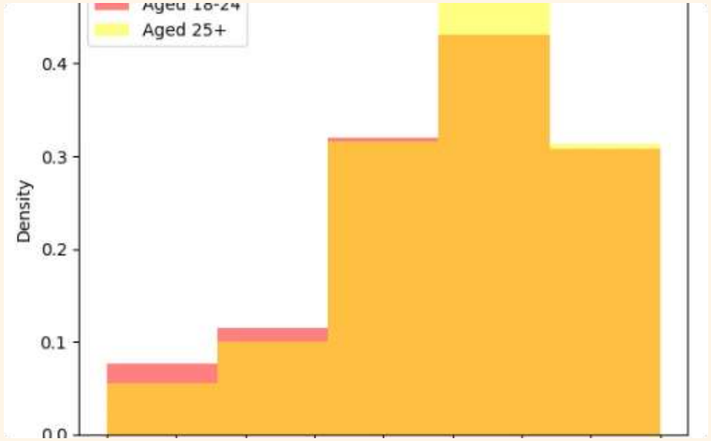
Statistical Testing

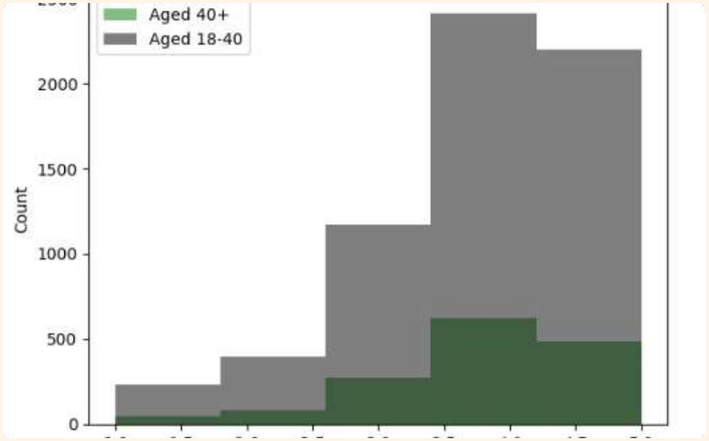
Conducted ANOVA tests to confirm findings

Key Findings: Age and Genre Preferences

Our analysis revealed surprising insights about age-related preferences. Contrary to expectations, younger audiences (under 10) did not consistently give higher ratings. Users aged 18-24 were more critical of Animated movies than older groups. Viewers over 40 showed a preference for Documentaries. Interestingly, both young adults (under 24) and older adults rated Film-Noir, War, and Documentary genres highly, challenging stereotypes about age-based preferences.







Under 18 and Over 56

More extreme ratings comparatively

18-24

More critical of Animated movies, prefer Film-Noir

Over 40

Higher ratings for Documentaries and Romantic Dramas

Gender Influences on Movie Preferences

Gender-based analysis challenged common assumptions about genre preferences. Women showed a surprising preference for Film-Noir, Documentary, and War genres, contradicting stereotypes about favoring Romance and Comedy. Men under 24 also rated these genres highly, suggesting a shared interest in thought-provoking content across genders. However, some stereotypes held true, such as women over 40 rating Romantic Dramas highly. Overall, gender had less impact on preferences than age or occupation.

1

Women's Preferences

Film-Noir, Documentary, War (unexpected)

2

Men's Preferences

Similar to women, especially under 24

3

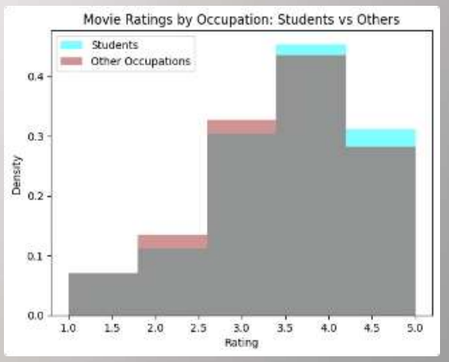
Gender Correlation

Strong positive correlation (0.76) in ratings

4

Age 25-34

Highest gender correlation (0.69) in preferences



Occupation and Movie Genre Preferences

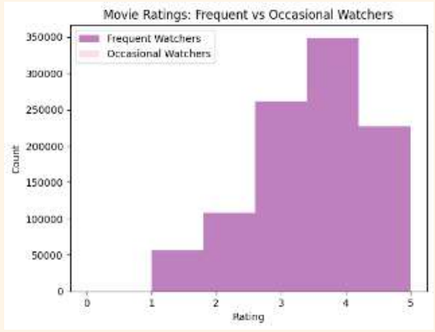
Occupation emerged as a significant factor in shaping movie preferences, often challenging stereotypes. Contrary to expectations, viewers in technical fields didn't show a particular preference for Sci-Fi and Adventure movies. The analysis revealed nuanced preferences across different occupational groups, suggesting that targeted content strategies based on profession could be effective. However, the specific preferences for each occupation group were not detailed in the input, indicating a need for further analysis in this area.

Occupation	Unexpected Preference	Expected Preference
Technical Fields	No strong Sci-Fi preference	-
Students	Higher overall ratings	-
Other Groups	Varied preferences	-

Viewing Frequency and Rating Patterns

Analysis of viewing frequency revealed interesting patterns. Users who watched movies frequently (more than 10 ratings) did not consistently provide higher ratings compared to occasional viewers. This finding challenges the assumption that more engaged users are necessarily more positive in their evaluations. However, frequent movie-watchers did show more consistent ratings between genders, suggesting that increased viewing leads to more stable preferences across demographic lines.

1



Frequent Viewers

No significant increase in average ratings

2

Gender Consistency

More stable preferences across genders for frequent viewers

3

Rating Distribution

Similar across viewing frequency levels

4

Engagement Impact

Frequency affects consistency more than positivity

		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

Statistical Analysis: ANOVA Test Results

To validate our findings, we conducted ANOVA tests to determine the relative significance of demographic factors on movie ratings. The results confirmed that age and occupation significantly affected ratings, while gender had a minimal impact. Age accounted for approximately 4.36% of the rating variance, occupation for 2.38%, and gender for only 0.15%. These findings support our conclusion that age and occupation are more influential in shaping movie preferences than gender.

	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

Age Impact

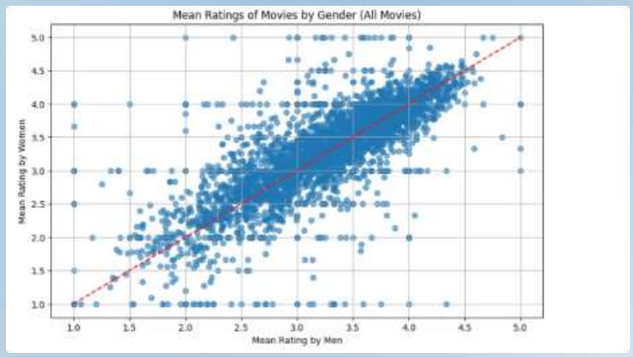
4.36% of rating variance

Occupation Impact

2.38% of rating variance

Gender Impact

0.15% of rating variance



Business Applications of Findings

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?*

Conjecture: Age and occupation are more significant determinants of movie ratings than gender

Targeted Marketing

Tailor campaigns based on age and occupation

Content Acquisition

Prioritize genres for specific demographics

Recommendation Systems

Improve algorithms with demographic insights

User Engagement

Enhance satisfaction through personalized content

Challenges and Limitations

While our analysis provided valuable insights, it faced several limitations. The dataset lacked information on factors like education level and income, which could influence preferences. Some demographics were underrepresented, potentially skewing results. The overlap of movie genres complicated isolating specific preferences. Additionally, the static nature of the dataset doesn't capture evolving tastes or emerging genres. The absence of viewing context data and potential bias in self-reported ratings also present challenges in interpreting the results.

1. ## Limited Demographics

   Missing education and income data

2. ## Data Imbalance

   Underrepresentation of certain groups

3. ## Genre Overlap

   Difficulty in isolating specific preferences

4. ## Static Dataset

   Doesn't capture evolving tastes

Conclusion: Leveraging Demographic Insights

This analysis of the MovieLens 1M dataset reveals that age and occupation are the most significant predictors of movie genre preferences, with gender playing a moderate role. These insights provide a foundation for audience-targeted marketing strategies and personalized content recommendations. By leveraging demographic data, media companies can make data-driven decisions to align content offerings with audience preferences, potentially increasing user engagement and satisfaction. However, the dynamic nature of viewer tastes necessitates ongoing analysis and strategy adjustments to remain relevant in an evolving media landscape.

Enhanced User Experience

Tailored content recommendations based on demographics

Data-Driven Decisions

Informing content acquisition and marketing strategies

Adaptive Platforms

Evolving with changing viewer preferences