The MovieLens 100K Dataset consists of three primary tables: **ratings**, **movies**, and **users**. In the ratings table, you can find data on user ratings for different movies. The movies table provides details about each movie, including movie id, movie title, release date, genres and more. The users table offers information about the users such as gender. The dataset consists of user ratings for movies, with 100,000 ratings from 943 users on 1,682 items.

Key Findings:

Movie Ratings Distribution:
The mean rating across all movies is 3.53, indicating an overall positive sentiment among users. Figure 1 illustrates the distribution, showing a skew towards higher ratings.

Movie Popularity Distribution:
The Figure 2 illustrates the distribution of movie popularity based on the number of ratings each movie received. The descending line graph depicts movies sorted by popularity, with the x-axis representing the movie index. As observed, some movies enjoy significantly higher popularity, reflected by a larger number of ratings, contributing to a steep decline in popularity for less-reviewed movies.

Correlation Insight: The Positive Link between Ratings and Popularity
The correlation coefficient between the number of ratings and average rating for each movie is 0.43. This positive correlation (Figure 3) suggests that, on average, movies with a higher number of ratings tend to have higher average ratings, indicating a tendency for more popular movies to receive favorable ratings.

Analysis Insight: Examining Genre Abundance and Ratings
Despite expectations, our analysis, illustrated in Figure 4, reveals that the quantity of movies within genres does not significantly impact the overall average rating. Figure 5, depicting a correlation coefficient of -0.03, emphasizes this lack of correlation, implying that a higher number of movies in a genre doesn't necessarily result in higher user satisfaction. This suggests that factors beyond quantity, such as unique movie characteristics or individual preferences, are more influential in determining average ratings in the MovieLens 100K Dataset.

Ethical Considerations:
User Privacy and Data Anonymization: The inclusion of personally identifiable information such as age, gender, occupation, and zip code in the user data introduces significant ethical concerns related to user privacy. To address these issues, robust data anonymization practices must be implemented to protect the identities and sensitive details of users. Ensuring strict confidentiality measures in the handling and storage of this data is paramount. Additionally, obtaining informed consent and adhering to data protection regulations are crucial ethical safeguards, ensuring that users are aware of how their information will be utilized.

Bias and Representation in Ratings: The positive correlation (0.43) between the number of ratings and average rating raises ethical concerns regarding potential biases and representation issues. The inherent popularity bias in the dataset may result in unequal representation of movies, favoring more mainstream content over equally valuable but less popular films. To address this, strategies such as stratified sampling should be employed to ensure fair representation of diverse content in the analysis and recommendations. Additionally, the overrepresentation of certain genres may introduce skewed correlations, influencing user recommendations. Mitigating this requires careful consideration of genre disparities and implementing measures to provide fair visibility to movies across different genres. Upholding user privacy through rigorous anonymization practices and maintaining transparent communication is crucial to build and preserve trust in the recommender system, ensuring ethical use of the data. All the figures are shown in next page.

Figure 1 - Movie Ratings Distribution

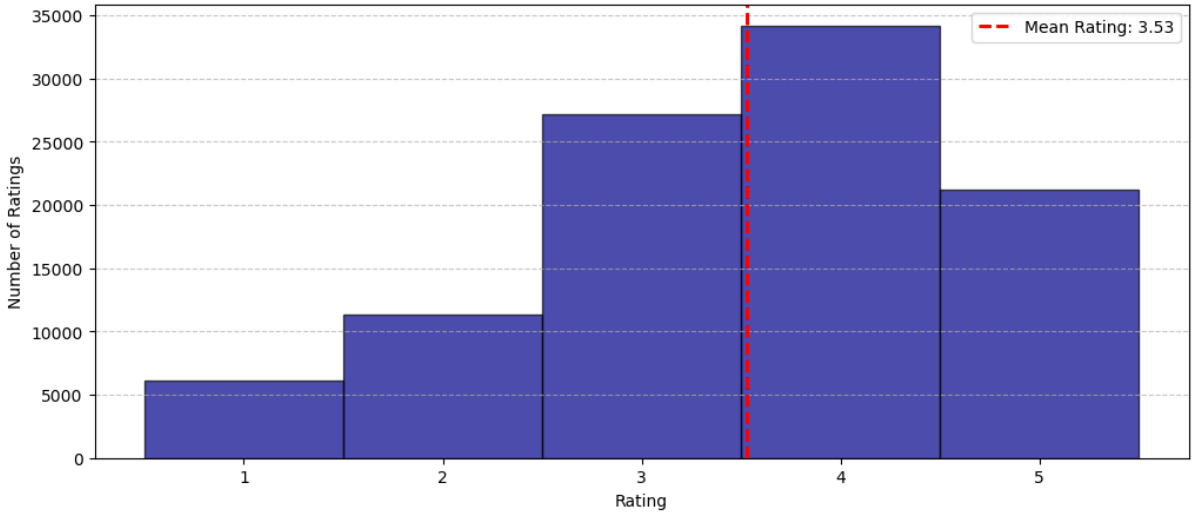


Figure 2 - Movie Popularity Distribution

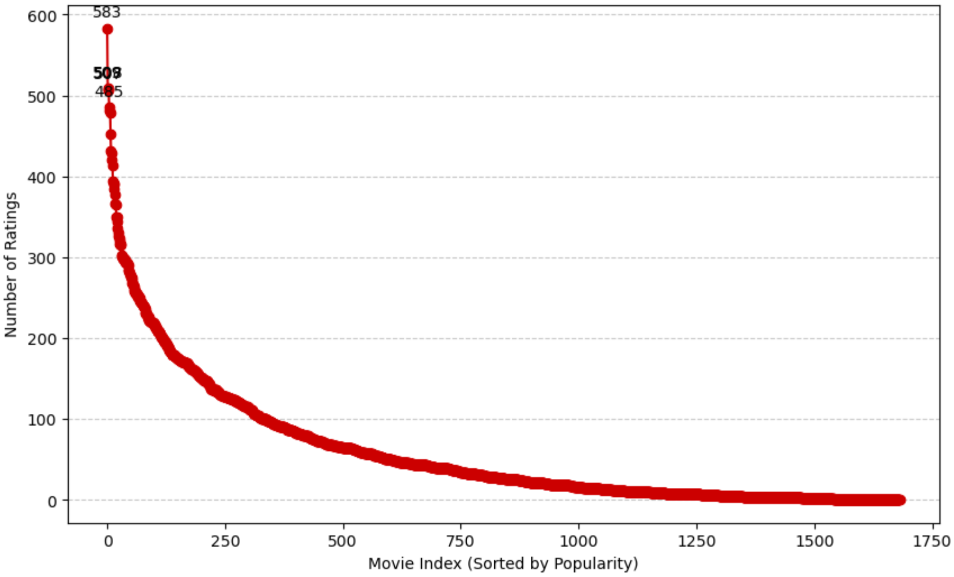


Figure 3 - Correlation Heatmap: Average Rating vs. Number of Ratings

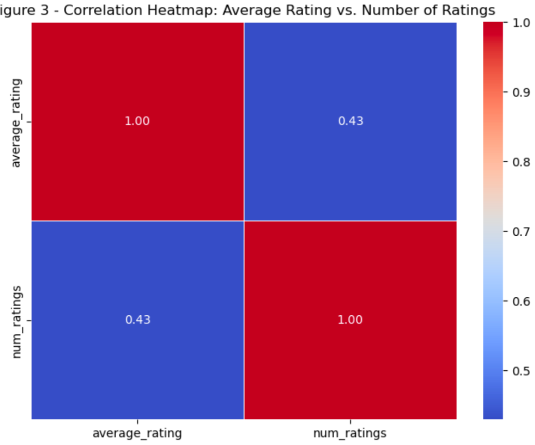


Figure 4 - Genre Distribution and Average Rating in MovieLens 100K Dataset

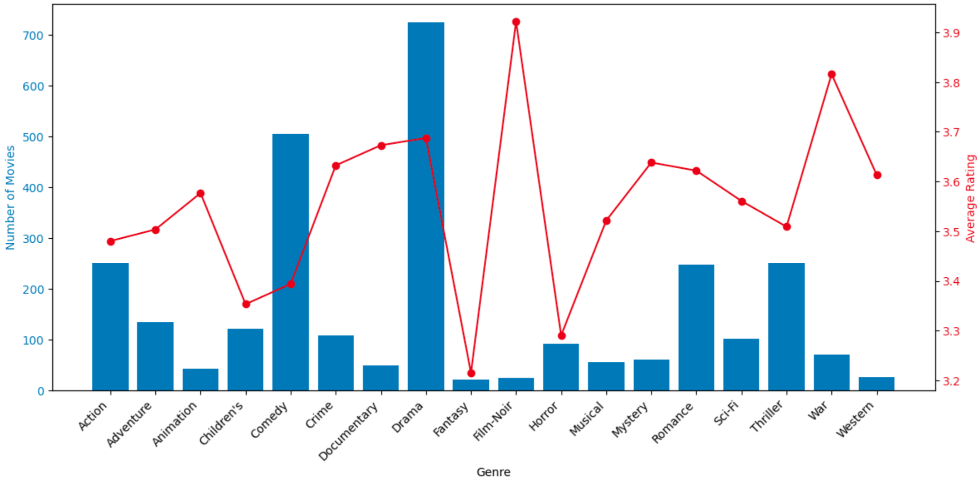
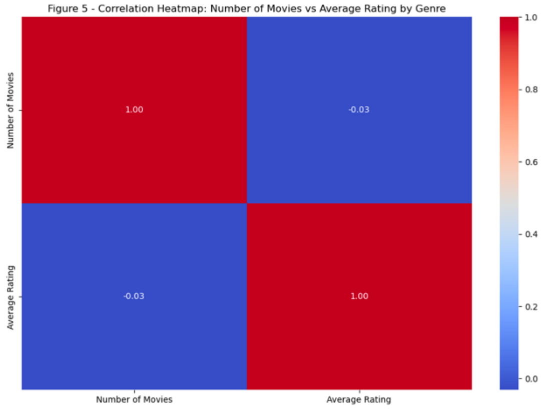


Figure 5 - Correlation Heatmap: Number of Movies vs Average Rating by Genre


Number of Movies
Average Rating
Number of Movies
Average Rating