Movie Ratings Analysis Report

1. Introduction

The purpose of this project is to analyze a dataset containing user ratings of various movies, identify trends and patterns in viewing and rating behavior, and explore how these insights could inform the development of a movie recommendation system.

The analysis was conducted using Python in Google Colab, with libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data cleaning, exploration, and visualization.

2. Data Cleaning and Preparation

The dataset was first inspected and checked for missing values, duplicates, and inconsistencies. Key steps taken during the cleaning process included:

- Handling missing or null values across relevant columns.
- Removing duplicate movie or rating entries where applicable.
- Converting date fields to proper datetime formats.
- Merged the movies and ratings datasets on the movieId columns, using left join.

3. Feature Engineering

New features were created to make the analysis rich and to support deeper insights into movie trends and rating behaviors. The engineered features include:

- **Year Released** – Extracted from movie titles to analyze rating patterns across decades.
- **Number of Ratings per Movie** – To identify the most-rated and least-rated films.
- **Average Rating per Movie** – To highlight top-rated and poorly-rated movies.
- **Genre Count** – Calculated as the number of genres assigned to each movie, to explore how multi-genre classification effects ratings.
- **Decade Feature** – Grouped movie years into decades for comparing trends.
- **Rating Timestamp Conversion** – Converted timestamps to datetime format to analyze user engagement trends over time.

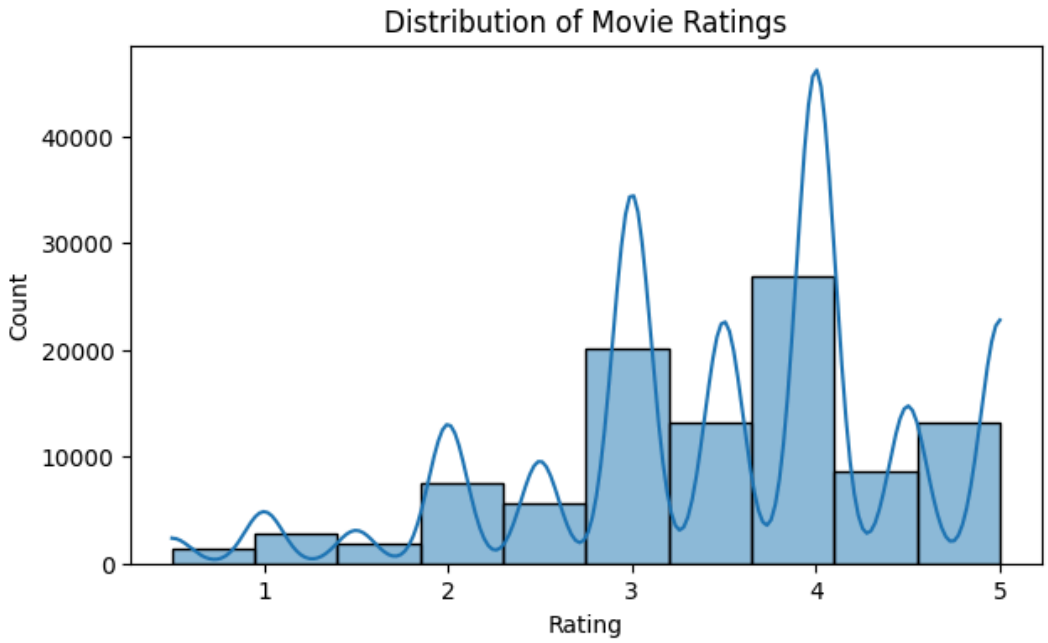
- **Year Rated** - Extracted the year movies were rated to be able to analyze how ratings were done over time.

These features were designed to enhance understanding of user behavior and lay a foundation for potential predictive modeling in the future.

4. Key Insights

4.1 Distribution of Movie Ratings

Most users tend to rate movies on the higher end of the scale.
The average rating is approximately 3.5, with the majority of ratings falling between 3 and 5.
This indicates that users are generally generous when rating movies, and low ratings are relatively rare.

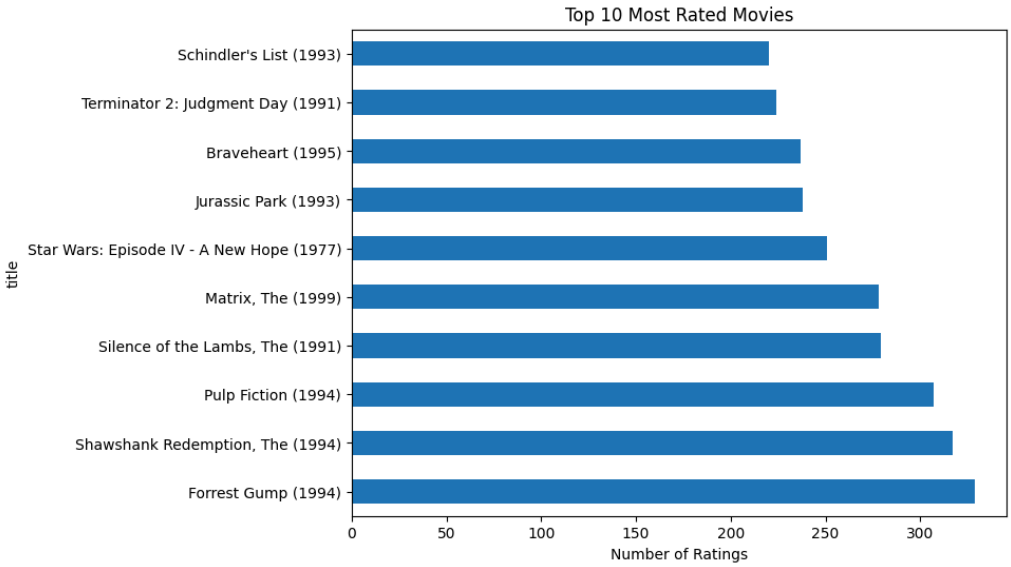


4.2 Most Rated Movies

The movies with the highest number of ratings include:

- *Forrest Gump (1994)*
- *The Shawshank Redemption (1994)*
- *Pulp Fiction (1994)*

Each of these received an average of at least 300 ratings. These films are well-known classics, indicating that popular titles tend to attract more engagement.

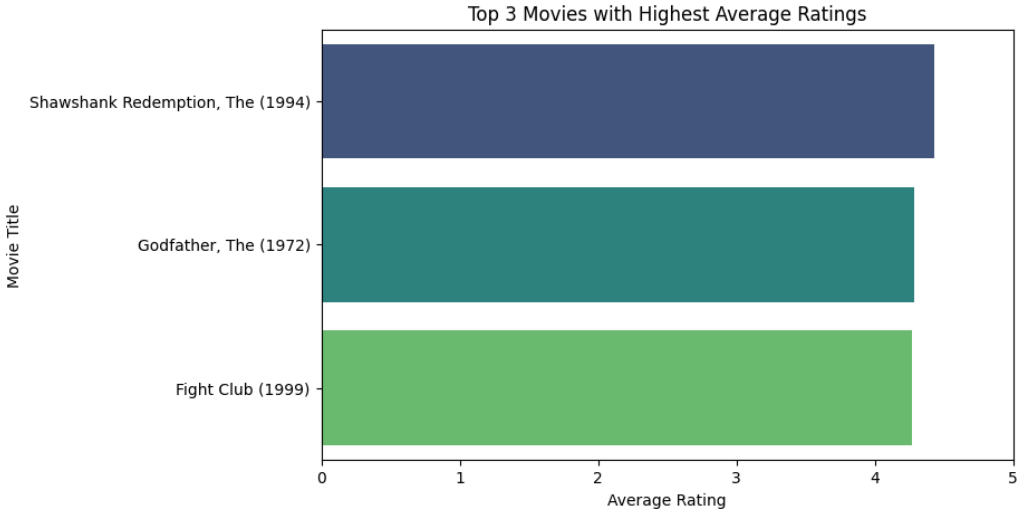


4.3 Highest Average Rated Movies

The movies with the highest average ratings include:

- *The Shawshank Redemption (1994)*
- *The Godfather (1972)*
- *Fight Club (1999)*

Each scored an average rating of around 4.2, suggesting strong audience approval and consistent satisfaction across viewers.

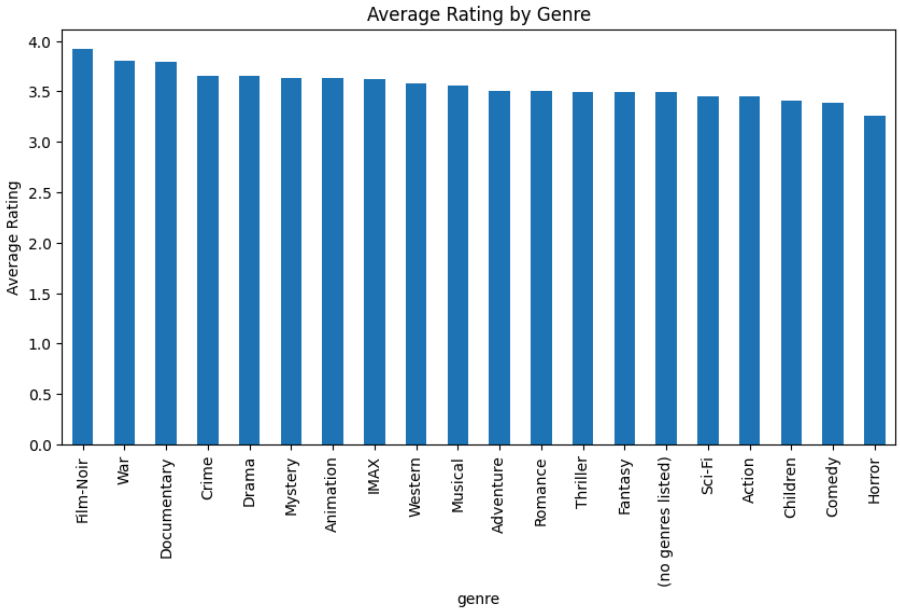


4.4 Rating Trends Over Time

Analysis of ratings over time showed a sharp rise around the year 2000, this suggests that this period experienced high user activity.
Afterward, the number of ratings fluctuated but remained relatively steady, with noticeable peaks around 2005 and after 2015.
This trend may be linked to the rise of online movie platforms and evolving viewer engagement habits.

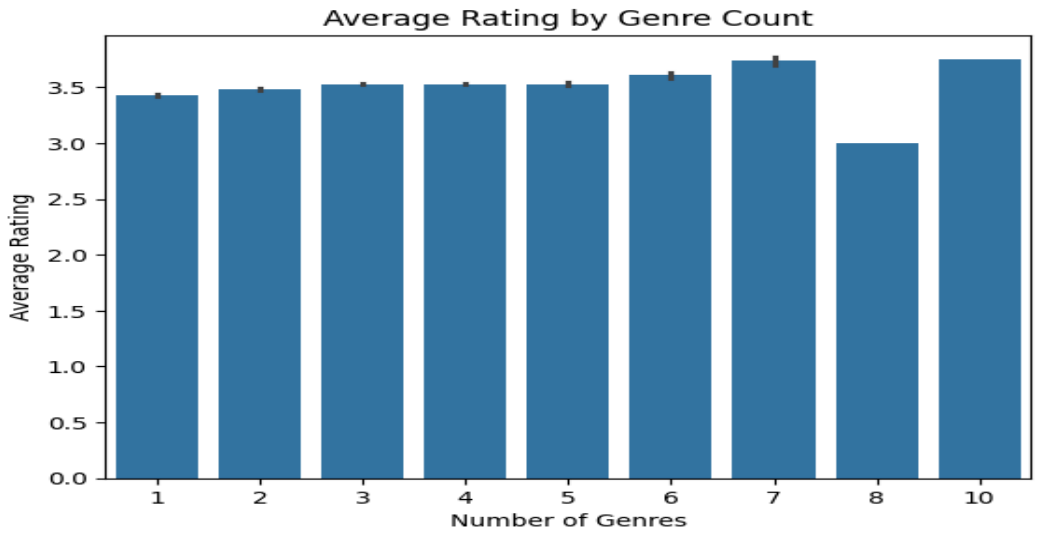
4.5 Most Common Genres

Movies with genres such as Film-Noir, War, and Documentary appear most frequently in the dataset. This indicates that these genres are dominant in the dataset and may also influence user rating behaviors.



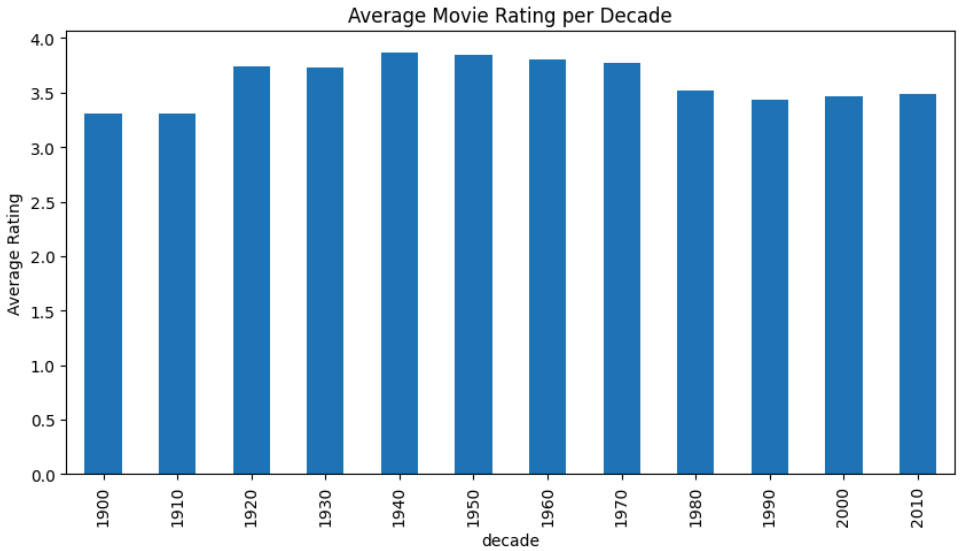
4.6 Relationship Between Genre Count and Average Rating

Movies with more genres tend to have higher average ratings compared to single-genre films.
This could mean that multi-genre films appeal to a wider range of audiences, though excessive blending of genres might dilute people's focus.



4.7 Ratings by Release Decade

Movies released between the 1940s and 1960s received the highest average ratings, typically around 3.8 to 3.9, which shows that viewers appreciate classic movies more.
On the contrary, movies from the 1980s to 2010s have slightly lower averages (around 3.4–3.5), suggesting that though modern films maintain consistent ratings, older classics always have strong audience approval.



5. Implications for a Recommendation System

The insights obtained from this analysis provide valuable foundations for a movie recommendation system. Key takeaways include:

- **User Rating Patterns:** Understanding that most ratings fall between 3–5 makes room for normalization and bias adjustments in recommendation algorithms.
- **Popularity Metrics:** High-volume rated movies shows that they are audience favorites, this could serve as baseline recommendations for new users (cold-start strategy).
- **Temporal Trends:** Knowing periods of increased engagement can help in time-based recommendation tuning.
- **Genre Preferences:** The dominance of certain genres and the influence of multi-genre films can guide content-based filtering models.
- **Historical Appreciation:** The high ratings of classic movies suggest potential value in recommending older, high-quality films to users with similar taste patterns.

6. Conclusion

This project explored a movie ratings dataset from MovieLens to uncover patterns in user behavior, movie popularity, and rating trends across time and genres.
The results show that users generally rate movies favorably, some timeless classics have dominant audience preference, and engagement has evolved alongside technological advancement.

These insights form a solid foundation for building an intelligent, data-driven movie recommendation system that leverages user preferences, genre diversity, and historical patterns to provide personalized movie suggestions.