Feature Engineering & Exploratory Data Analysis on MovieLens Dataset

Introduction

The aim of this analysis is to explore, clean, and transform the MovieLens dataset to uncover meaningful insights about movie trends, user behavior, and genre performance.

The dataset contains four main files:

ratings.csv: User ratings for movies.

movies.csv: Movie titles and associated genres.

tags.csv: User-generated descriptive tags.

links.csv: External references linking movies to IMDb and TMDb.

By performing feature engineering and exploratory data analysis (EDA), I aim to understand how users rate movies, how genres differ in popularity, and how rating behavior changes over time. These insights form the foundation for designing a data-driven movie recommendation system.

Data Preparation

- The data from all four files were merged using the common key movieId.
- Duplicate entries were checked and removed.
- There were no Missing values
- The timestamp column was converted into a proper datetime format to enable time-based trend analysis. This created a unified dataset suitable for analysis and feature creation

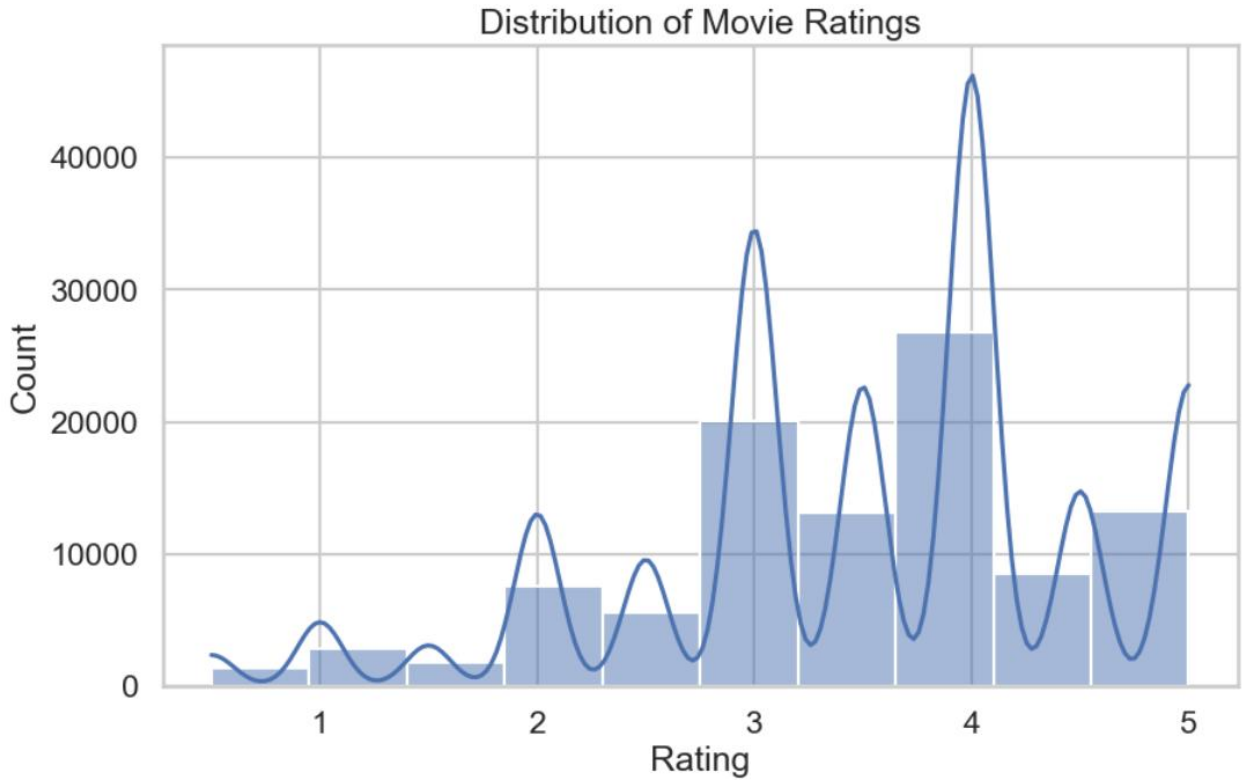
Features Created and the Reasons

- **Release Year** – Extracted from the movie title (e.g., Toy Story (1995) → 1995).
  Why: Helps analyze how movie popularity and audience preferences have evolved across different years or decades. It also allows us to compare rating patterns based on when movies were released.
- **Genre Count** – Number of genres per movie (e.g., Action|Adventure|Sci-Fi → 3).
  Why: Allows us to study whether movies that combine multiple genres attract higher ratings or broader audiences compared to single-genre movies.
- **Average Rating per Movie** – It means rating each movie receives.
  Why: Useful for ranking movies based on viewer satisfaction and identifying the most liked or critically acclaimed titles.
- **Average Rating per Genre** – It means of all ratings grouped by genre.
  Why: Helps reveal which genres audiences rate more favorably and which ones receive lower ratings on average, providing insight into viewer taste and preferences.

- **Number of Ratings per Genre –** Total count of ratings given to movies of each genre.
  Why: Shows which genres are most popular or widely watched, helping to identify audience engagement levels across different movie categories.
- **Rating Year** – Extracted from the timestamp column.
  Why: Enables the exploration of how user activity and movie ratings change over time, highlighting periods of high engagement or shifts in viewing trends.

KEY INSIGHTS

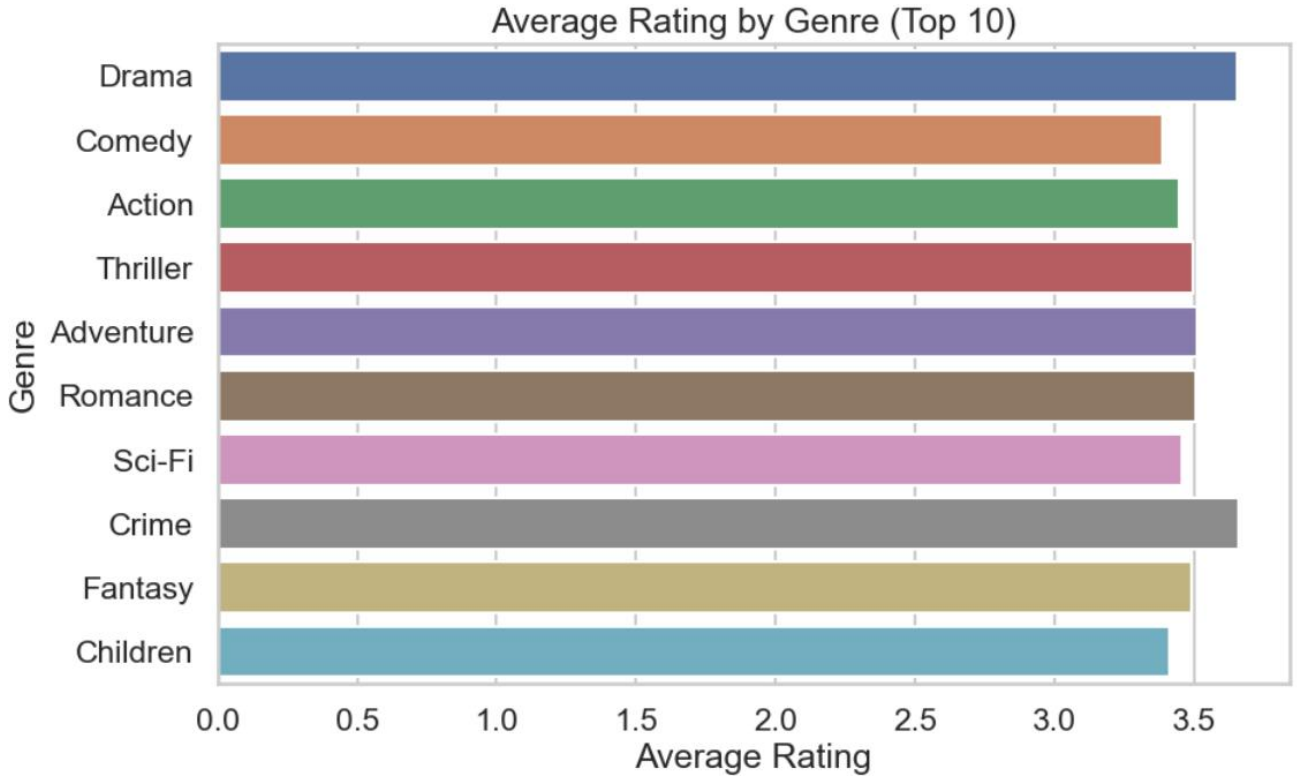
Distribution of Movie Ratings



Rating Range	Approx. Count	Observation
4.0 – 5.0	Very High	Most users rate positively
2.0 – 3.5	Moderate	Fewer neutral ratings
0.5 – 1.5	Very Low	Very few poor ratings

The histogram and table shows that most users rate movies highly, with ratings clustering around 4.0 and 5.0. Very few ratings fall below 2.0, meaning users rarely give extremely low scores.This skew toward higher ratings suggests that people tend to watch movies they expect to enjoy or only rate movies they like. The overall positivity bias suggests that a recommendation model should not rely solely on absolute ratings but rather relative user preferences and behavioral similarity.

Average Rating by Genre



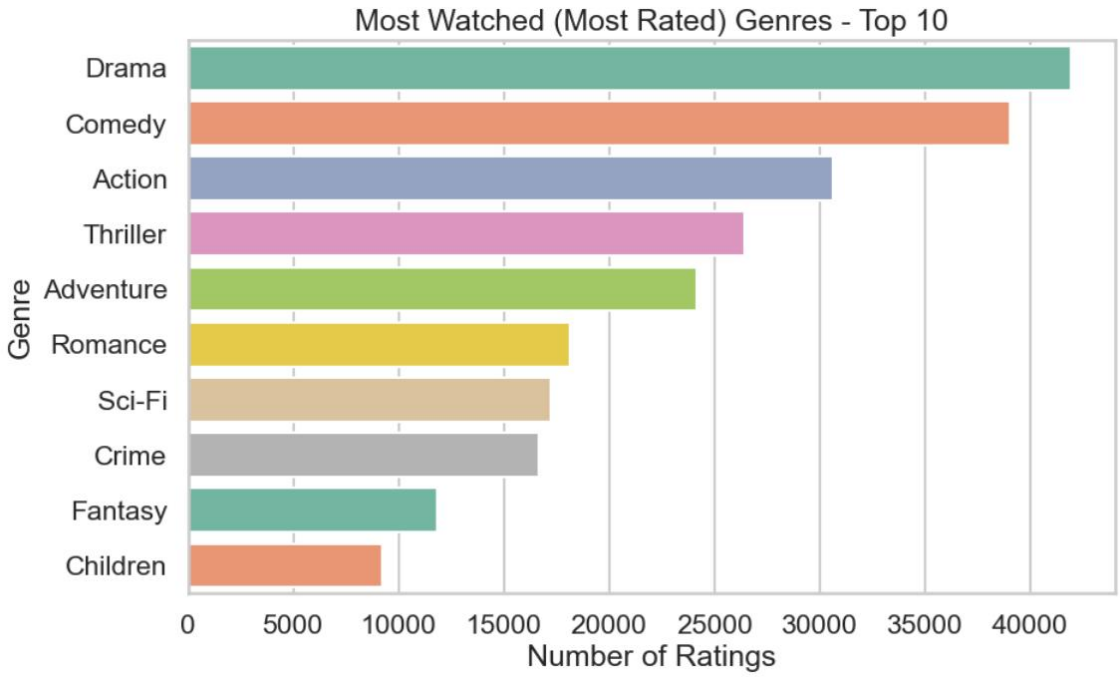
.

The bar chart shows that Drama, and Crime tend to receive the highest average ratings, all above 3.5. While, Children and Comedy movies have slightly lower averages (around 3.3), possibly reflecting a narrower target audience or mixed reception from adults. Genres with emotional depth tend to receive higher ratings, a useful pattern for recommending quality titles to users who value storytelling. On the other hand, Children's and Fantasy genres could be recommended selectively, for example, to users who previously rated such movies highly.

Top 10 Most Watched Genres

The top 10 genres show that users are mostly drawn to story-driven, emotionally engaging, and high-energy genres like Drama, Comedy, Action, and Thriller. These genres have broader audience appeal and usually receive consistent ratings across many users.

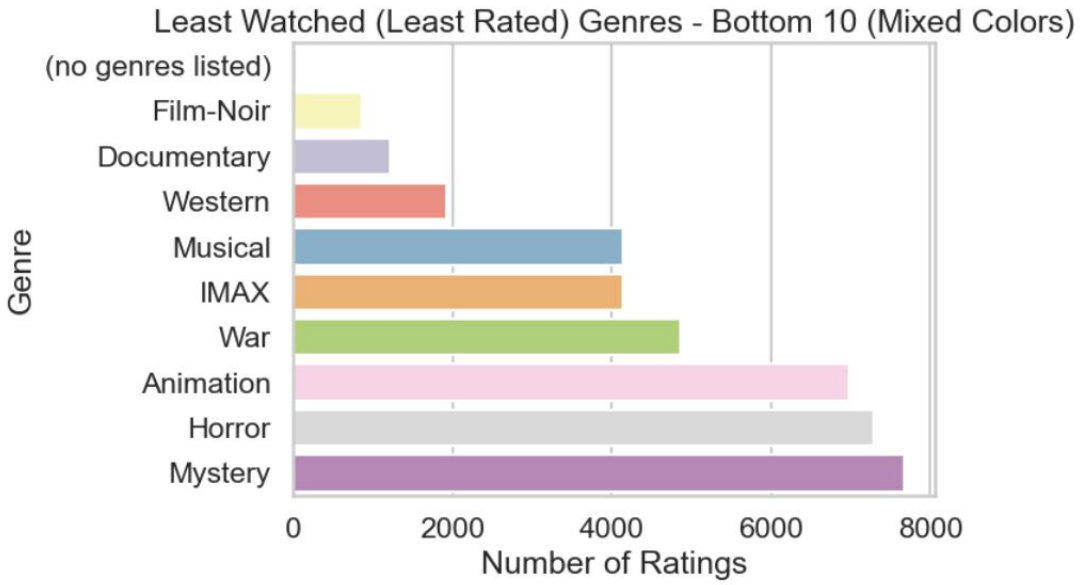
A recommendation system could prioritize these popular genres for new users (cold start) since they have the widest reach and the most data for accurate predictions.



Rank	Genre	Observation
1	Drama	The most watched genre; appeals to a wide audience with emotional storytelling and realistic themes.
2	Comedy	High viewership likely due to its universal appeal and entertainment value.
3	Action	Popular for its fast-paced storylines and excitement.
4	Thriller	Keeps audiences engaged through suspense and tension.
5	Adventure	Often overlaps with Action and Fantasy, drawing in fans of exploration and imagination.
6	Romance	Appeals strongly to younger audiences and couples.
7	Sci-Fi	Attracts viewers who enjoy futuristic and imaginative storytelling.
8	Crime	Consistently interesting due to strong plots and mystery elements.
9	Fantasy	Often linked with blockbuster franchises, appealing to younger and fan-based audiences.
10	Children	Watched widely by families, though limited to a certain age group.

Least Watched Genres

These Genres represent categories with the fewest ratings, often niche or specialized genres that attract smaller audiences. These genres are more specialized and attract smaller but loyal audiences. The low number of ratings doesn't necessarily mean poor quality, it may reflect limited releases or niche interest groups.



Rank	Genre	Observation
1	Film-Noir	Classic genre that appeals mainly to older or niche viewers.
2	Western	Declining popularity among younger audiences; limited modern releases.
3	IMAX	Fewer ratings likely due to limited availability and specific movie types.
4	Musical	Specific audience taste; not a mainstream preference.
5	Documentary	Focused on education or real-life stories — attracts smaller, serious viewers.
6	War	Specialized interest; may not appeal to all demographics.
7	Horror	Strong niche appeal but limited mass popularity due to intensity.
8	Mystery	Often overlaps with Crime and Thriller but less frequent releases.
9	Animation	Popular among children, but fewer ratings compared to other genres.
10	Fantasy (sub-variant or niche)	Depending on the dataset split, some fantasy films receive fewer views.

Top 10 Years with the Most Movie Ratings

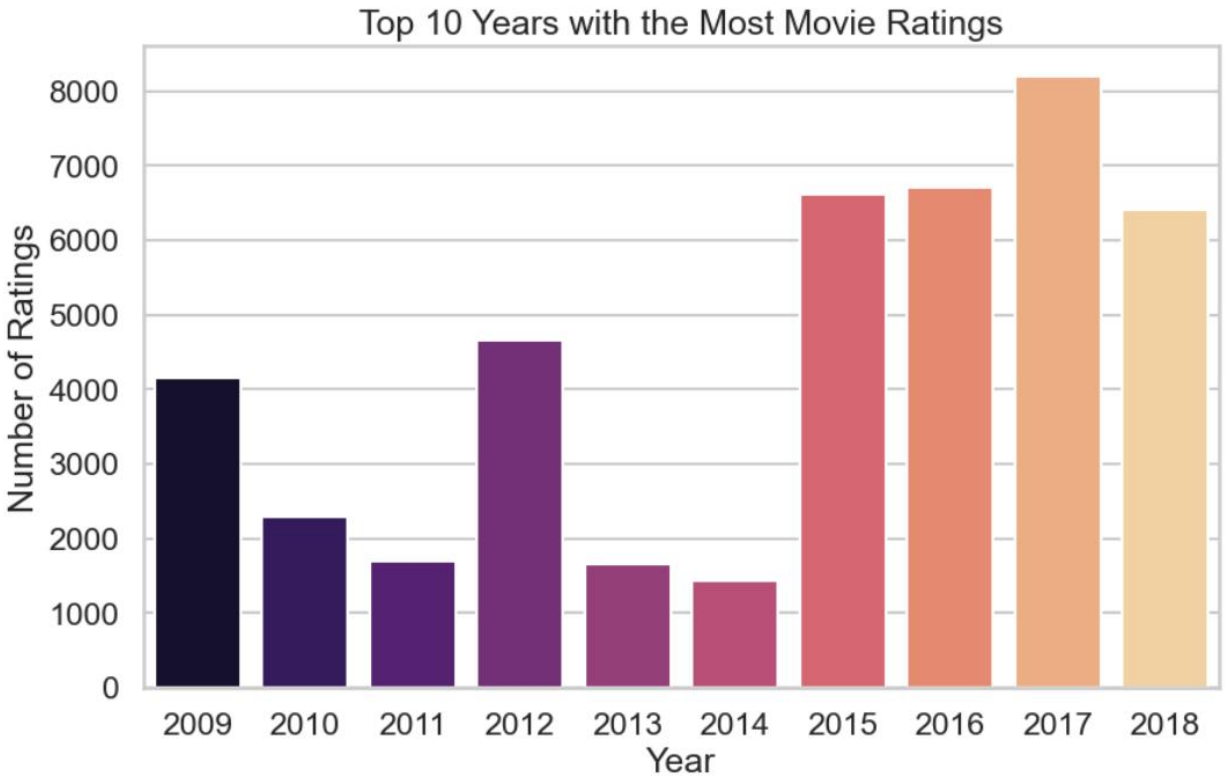
The graph shows the top 10 years in which the highest number of movie ratings were recorded. Each bar represents how active users were in rating movies during those years.

Recent years show higher rating activity, indicating that users have become increasingly active in rating movies over time. This can be linked to the growth of online movie platforms and streaming services like Netflix, IMDb, and MovieLens, which make it easier to review films digitally.

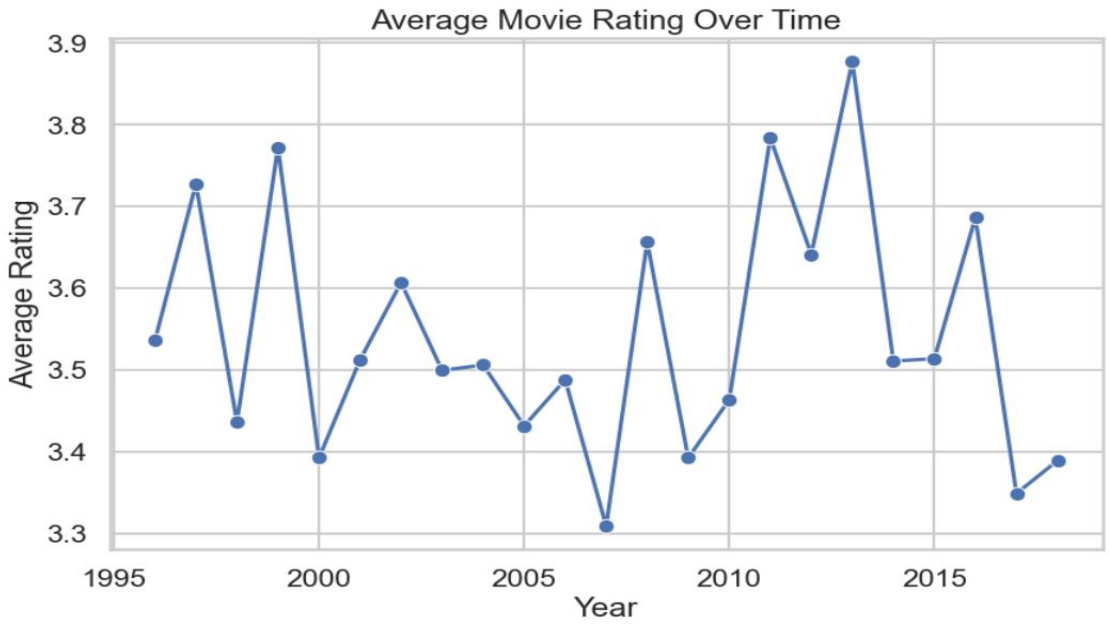
A few peak years stand out, those likely correspond to periods when major movie releases or increased user participation occurred.

Older years show fewer ratings, either because fewer users were active then or the database had limited historical data.

The trend overall suggests a growth in movie-watching and user engagement with rating platforms in more recent times.



Average Movie Rating Over Time



The line graph above shows the trend of average movie ratings from 1995 to 2018. The pattern reveals a fluctuating movement rather than a steady increase or decline, suggesting that overall viewer satisfaction has remained relatively consistent over the years. There are noticeable spikes in 1998, 2011, and 2013, indicating periods when audiences rated movies more favorably, possibly due to the release of standout films or successful trends within certain genres. On the other hand, dips in 2000, 2008, and 2016 may correspond to years when fewer high-quality or widely appealing movies were released. These fluctuations show that changes in storytelling quality, audience expectations, and industry trends can directly affect how viewers rate movies.

From a recommendation system perspective, these insights are valuable because they highlight how temporal factors (such as release year) can influence user preferences. By incorporating time-based patterns into a recommendation model, systems can better understand evolving audience interests, allowing them to recommend movies not just based on user ratings or genres, but also by identifying periods or eras with higher audience engagement. For instance, if users tend to prefer movies from highly rated years like 2011 or 2013, the system can prioritize content from those years when suggesting similar films. This approach enhances personalization and helps streaming platforms maintain viewer satisfaction by adapting to long-term rating trends.