Analysis and Feature Engineering of a MovieLens Dataset Report

During Data Cleaning and preparation, six new features were added to the given dataset to be able to discover more insights and perform more analysis.

The New Features are as follows:

1. Release Year
2. Decade
3. Average User Rating
4. Rating Year
5. Number of Ratings Per Movie
6. Number of Genres Per Movie

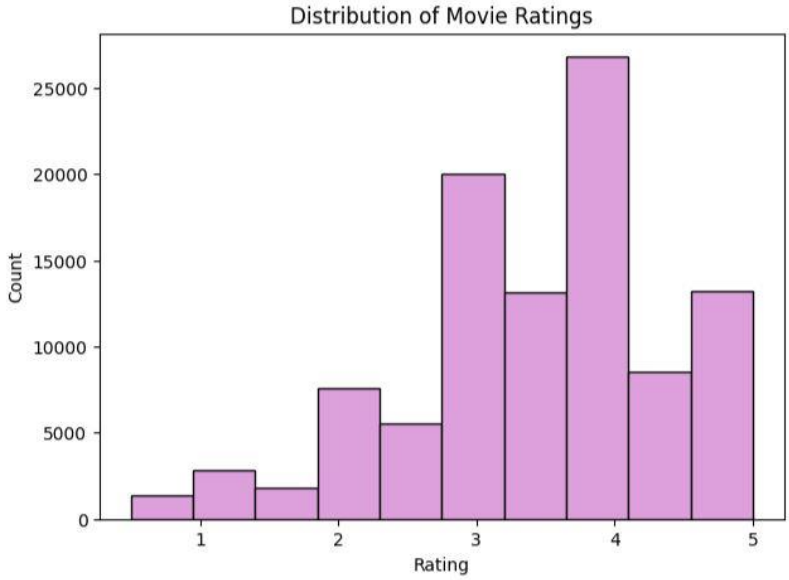
FEATURES	**PURPOSE**
Release Year	This was extracted from movie titles to know when a movie was release.
Decade	Grouped years into 10-year intervals to determine which decade has the best ratings.
Average User Rating	This is to get the average rating each user gives in other to understand user behavior when rating. Are they generous or strict?
Rating Year	This is to get the year a user rated a movie to analyze which years had more user engagement.
Number of Ratings Per Movie	This is to get the number of ratings each movie receive to find which movies are most popular.
Number of Genres Per Movie	This is to get the number of genres assigned to a movie. It will help analyze if movies with multiple genres attract more ratings or not.

Findings and Recommendations

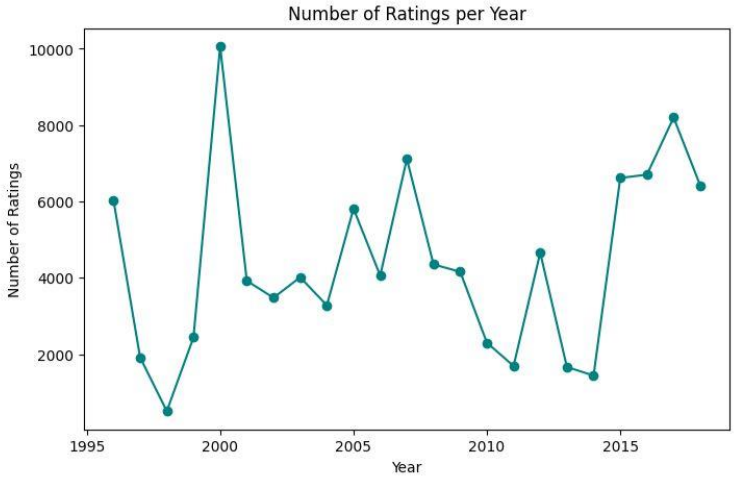
After creating the six new features. I proceeded to perform Exploratory Data Analysis (EDA). The focus of this analysis is to use these new variables to find meaningful trends and insights.

KEY INSIGHTS

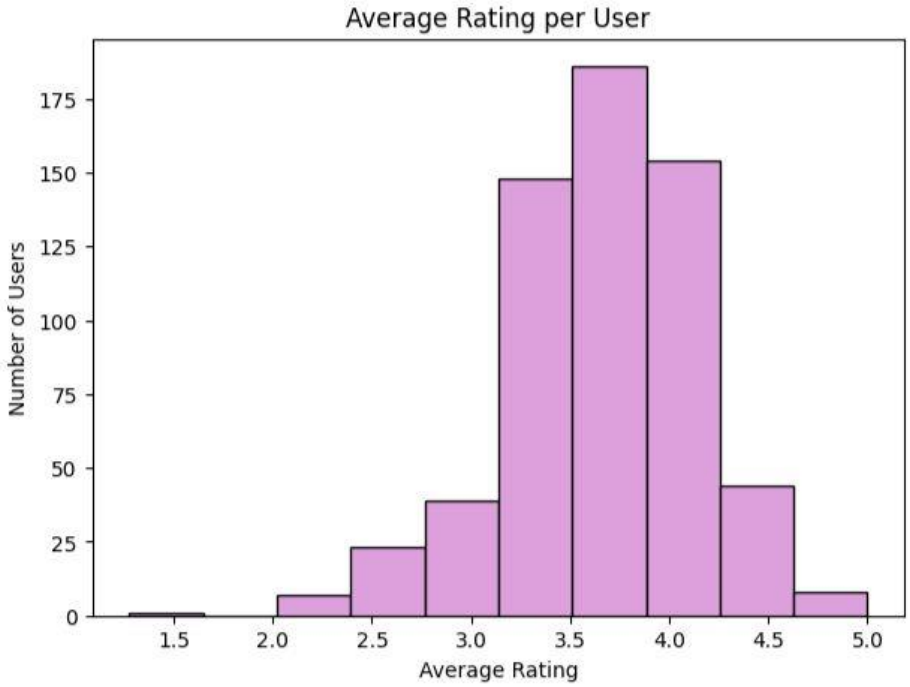
1. **Users are generally satisfied:** The majority of ratings fall between 3.0 and 5.0, suggesting that the overall quality of movies is perceived as good, or that users mostly rate movies they enjoy.



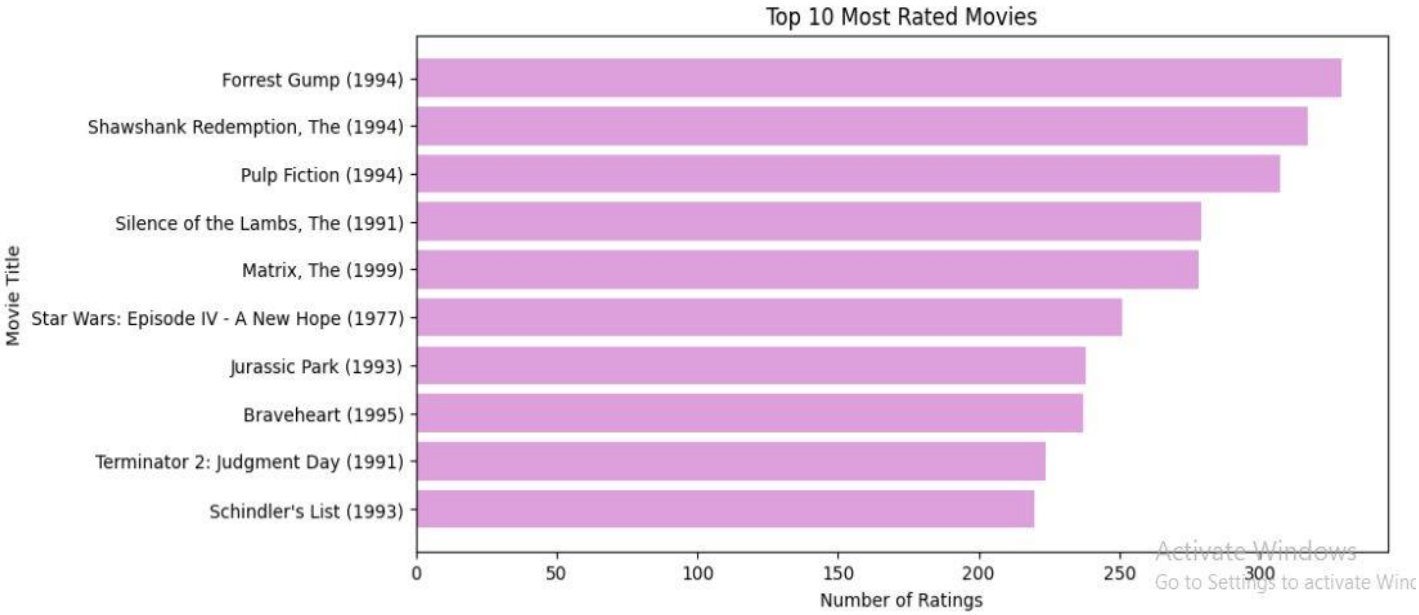
2. **Rating activity over time shows instability:** Rating frequency peaked dramatically around the year 2000, followed by inconsistent fluctuations in subsequent years.



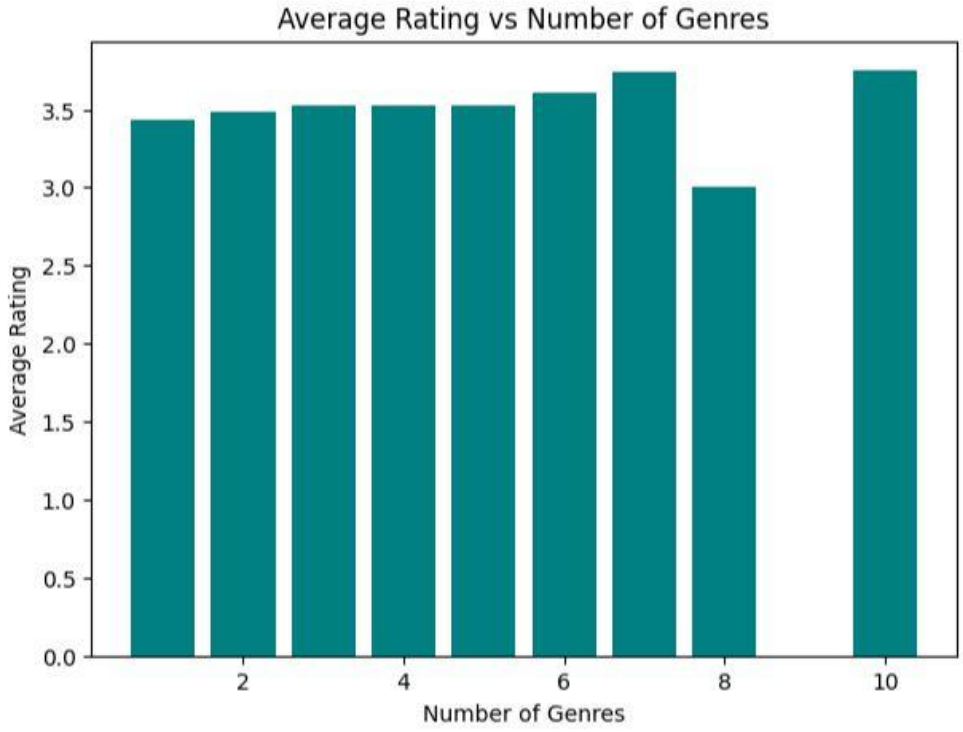
3. **Users rate generously and consistently:** Most users rate movies similarly and tend to give high scores. Very few users give low ratings, so there aren't many strict raters.



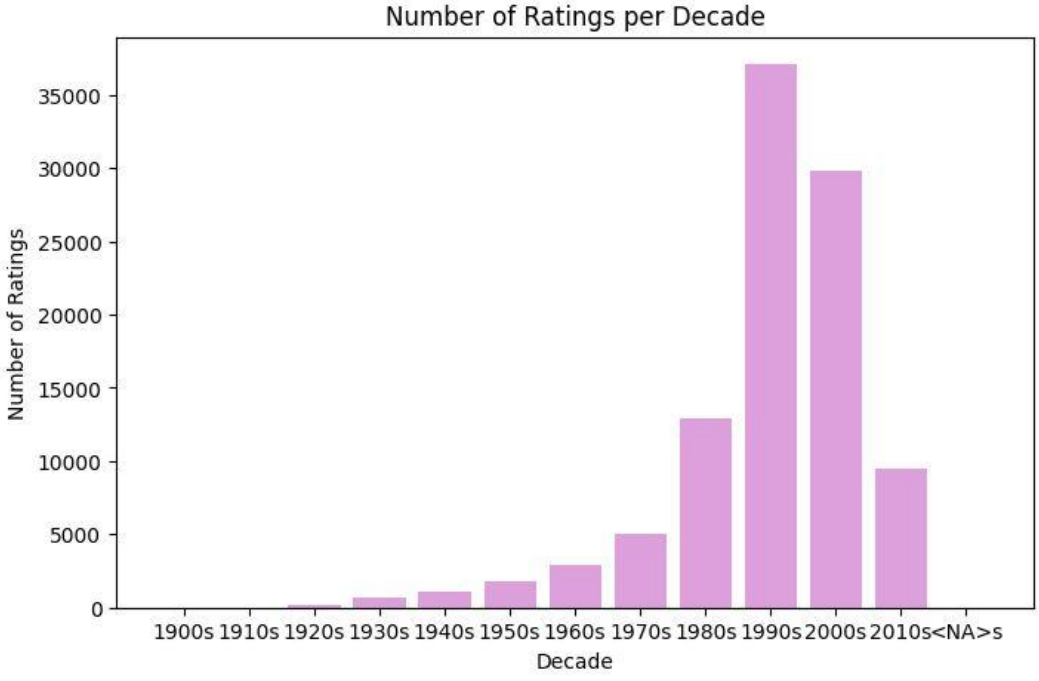
4. **Popular classics dominate ratings:** Movies with the most ratings are classics from the 1990s, showing that rating volume is driven more by long-term popularity than by recent they are.



5. **Genre count has limited effect on ratings:** For most movies (1–7 genres), having more or fewer genres doesn't significantly affect the average rating.



6. **Ratings are concentrated on 1990s and 2000s films:** The vast majority of rating activity focuses on films released in the 1990s and 2000s.



How These Insights Could Support Building a Recommendation System

These findings give a clearer picture of how users interact with movies and ratings and that understanding is key for designing smarter recommendations.

- **High ratings overall:** Since most users rate movies positively, it means the system shouldn't just rely on rating scores alone. Instead, it should learn to recognize which users are genuinely selective and which ones tend to give everything a high mark.

- **Unstable rating activity over time:** The fact that ratings peaked around the year 2000 and dropped afterward suggests that people's engagement changes with time. A good system could use this to track trends, maybe recommend older classics during throwback periods or highlight currently popular releases.

- **Popular classics dominate:** It's clear that older, well-loved movies attract more ratings. This means a recommendation system needs to find a balance between recommending popular favorites and helping users discover hidden gems they might not know about.

- **Users rate generously and similarly:** Because people don't differ much in how they rate; a model might need to look beyond just numbers. For example, combining rating data with user behavior (like watch history or genres liked) can help create better personalized recommendations.

- **Genre count doesn't affect rating:** The number of genres doesn't seem to make a movie better or worse in people's eyes. This suggests that the model should pay more attention to what genres are combined, not how many.

- **Ratings mostly focus on 1990s–2000s films:** Since most data points come from that era, recommendations could easily favor older movies. To stay relevant, future systems should make sure new releases get enough visibility too.