FEATURE ENGINEERING AND EXPLORATORY DATA ANALYSIS ON MOVIELENS DATASET

Introduction

The goal of this project was to perform **feature engineering** and **exploratory data analysis (EDA)** on the MovieLens dataset to understand user preferences and movie characteristics. The dataset includes four main tables:

- **Movies:** movie titles and genres
- **Ratings:** user ratings of movies
- **Tags:** user-provided tags
- **Links:** identifiers for IMDb and TMDB

After data cleaning and merging, the final dataset contained **100,836 ratings** across **9,742 movies** from **610 users**. The analysis aimed to:

1. Engineer additional features that capture movie and user characteristics.
2. Identify patterns in user ratings, genres, and movie popularity.
3. Provide relevant insights

Feature Engineering

The following features were engineered:

Feature Name	Description	Why It Was Created
release_year	The 4-digit year extracted from the movie title.	To analyze trends in movies and ratings over time.
genre_count	The number of genres assigned to each movie.	To test whether movies with more genres receive higher ratings.
primary_genre	The first-listed (main) genre for each movie.	To group movies by dominant category for genre-based analysis.
movie_avg_rating	Average of all user ratings per movie.	To identify well-rated movies

movie_rating_count	Total number of ratings per movie.	To measure movie popularity and confidence in its average rating.
user_avg_rating	Average rating given by each user.	To capture user behavior and adjust for lenient or strict raters.

These six engineered features added analytical depth and supported both **content-based** and **collaborative filtering** in recommendation systems.

Key Insights from Exploratory Data Analysis & How they could support building a recommendation system in the future.

The exploratory data analysis conducted in this project uncovered several key insights that can directly inform the building and optimization of future **movie recommendation systems**.

1. **Positively Skewed Ratings Distribution**
   The finding that most ratings fall between 3 and 4 suggests that users generally express mild to moderate approval. This can be leveraged to normalize ratings in a recommender system by adjusting for users' rating bias and focusing on subtle differences (e.g., identifying what makes a user rate one movie a 4 versus a 3).

2. **Concentration of Ratings on Popular Movies**
   The dominance of a few highly rated movies (e.g., *Forrest Gump*, *The Shawshank Redemption*) indicates popularity bias in user activity. Recommender systems can account for this by using popularity-based priors , giving more stable predictions to well-rated movies while still promoting lesser-known titles.

3. **Moderate User Rating Behavior**
   Since most users tend to rate moderately, features like user_avg_rating become important to model user leniency or strictness. Future recommenders can adjust predicted ratings relative to each user's personal average, making comparisons fairer and improving user satisfaction.

4. **Genre Preferences and Rating Tendencies**
   The discovery that genres like *Film-Noir*, *Mystery*, and *Documentary* receive higher ratings highlights strong content-based preferences. This supports content-based filtering, where genre, tone, and thematic similarities drive recommendations. Systems can learn that users who rate these genres highly may respond well to similar films.

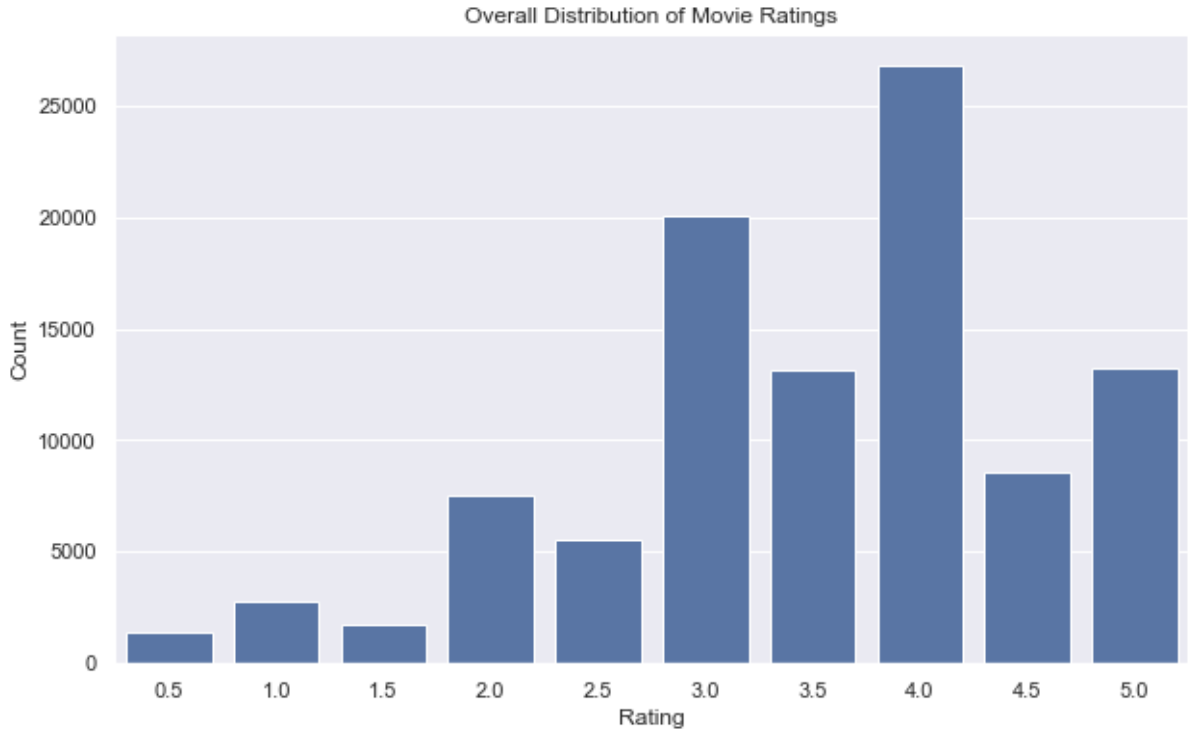
5. **Popularity of Action, Drama, and Comedy**
   These genres' widespread appeal implies that they can serve as baseline candidates in hybrid

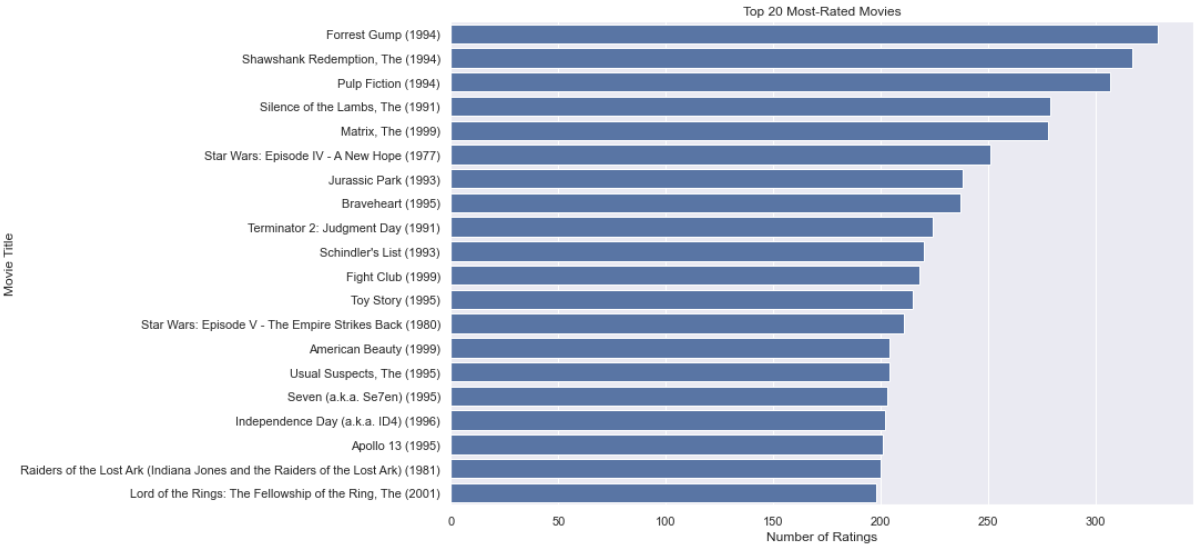
recommenders, ensuring that users always receive a mix of popular and personalized suggestions. Such insights can help maintain engagement by balancing novelty and familiarity.

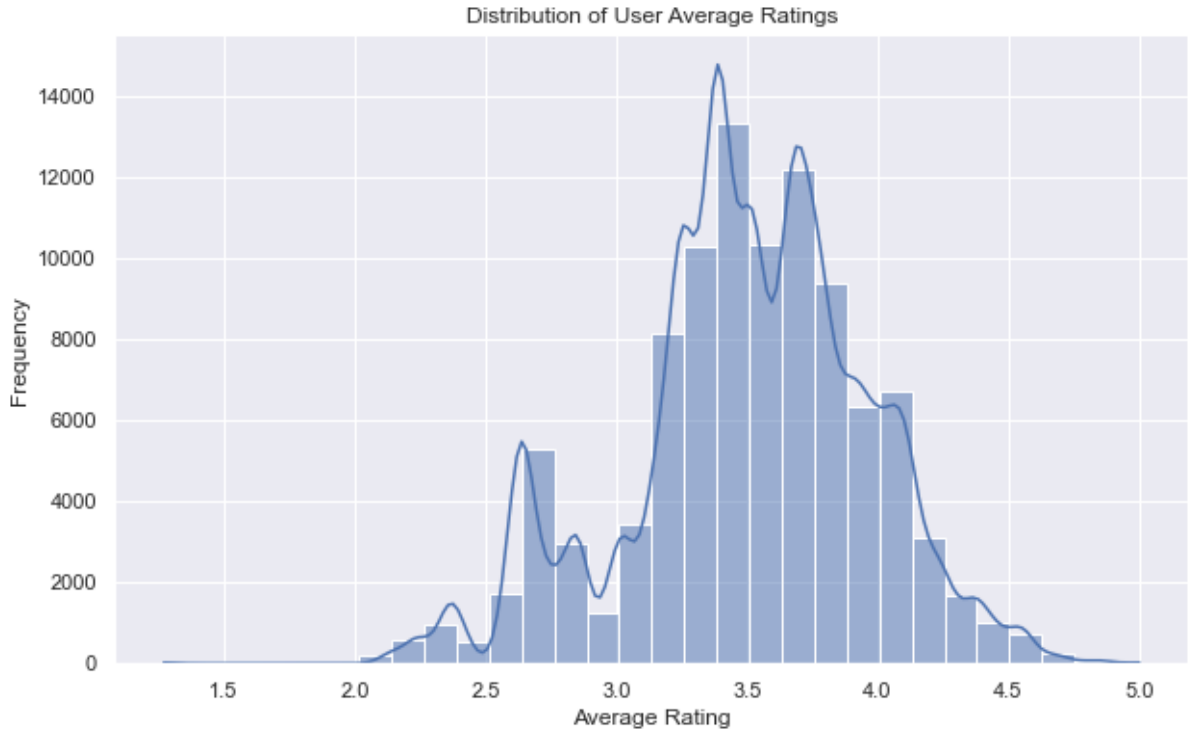
6. **Temporal Patterns (Release Year Trends)**
   The increase in the number of movies from the mid-1990s and user engagement with these titles indicate that release year can influence preferences. Future systems can use temporal filtering by recommending older classics to nostalgic users or newer releases to trend-focused users, based on their viewing history.

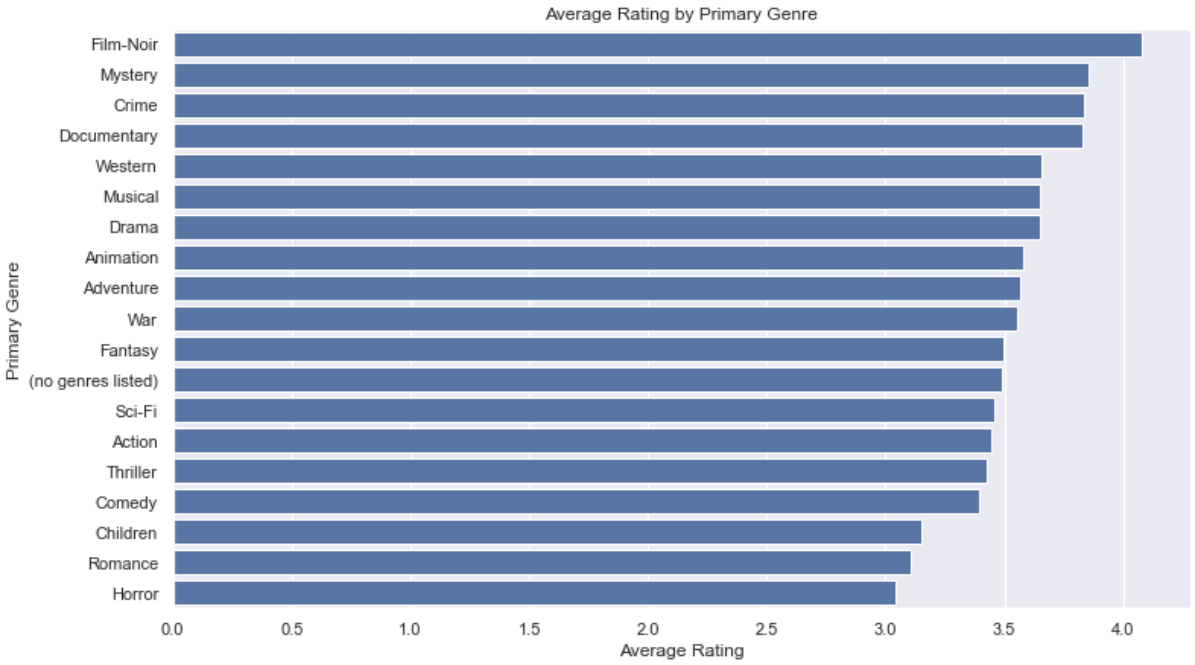
7. **Correlation Between Number of Genre and Average Rating:**

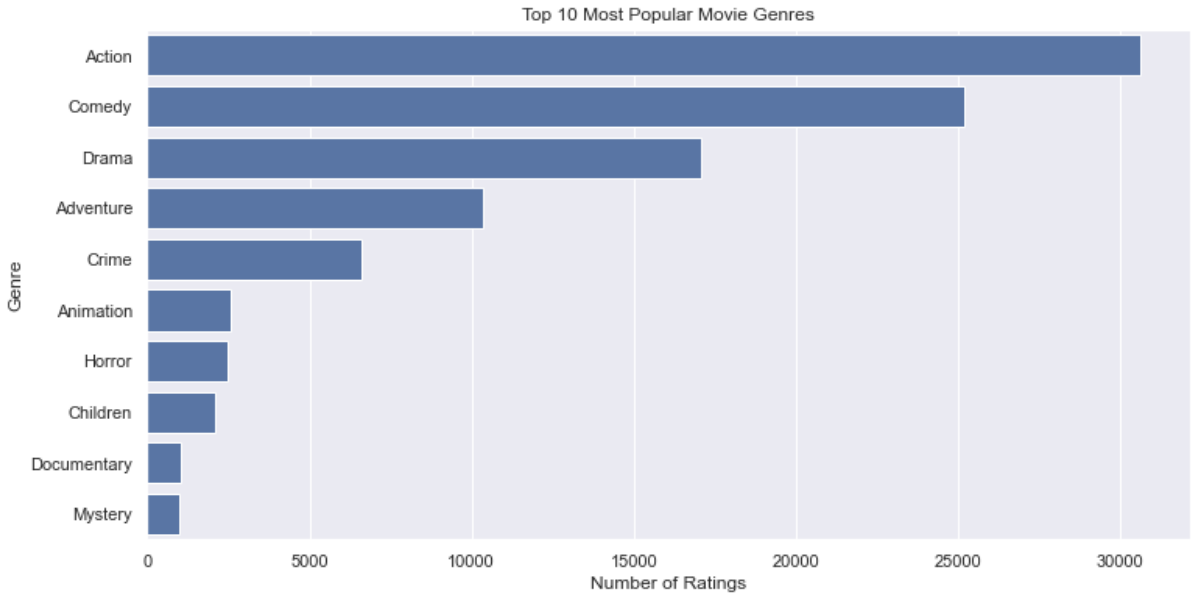
   There is a slight positive correlation between the number of genres a movie has and its average rating, which suggests that films spanning multiple genres tend to receive marginally higher ratings than those confined to a single category. This insight implies that genre diversity) can be an informative feature by helping to identify users who prefer versatile, mixed-genre content versus those who enjoy more focused movies. Although the correlation is modest, it highlights how incorporating genre diversity into a recommender system can slightly enhance personalization and the overall quality of movie suggestions.

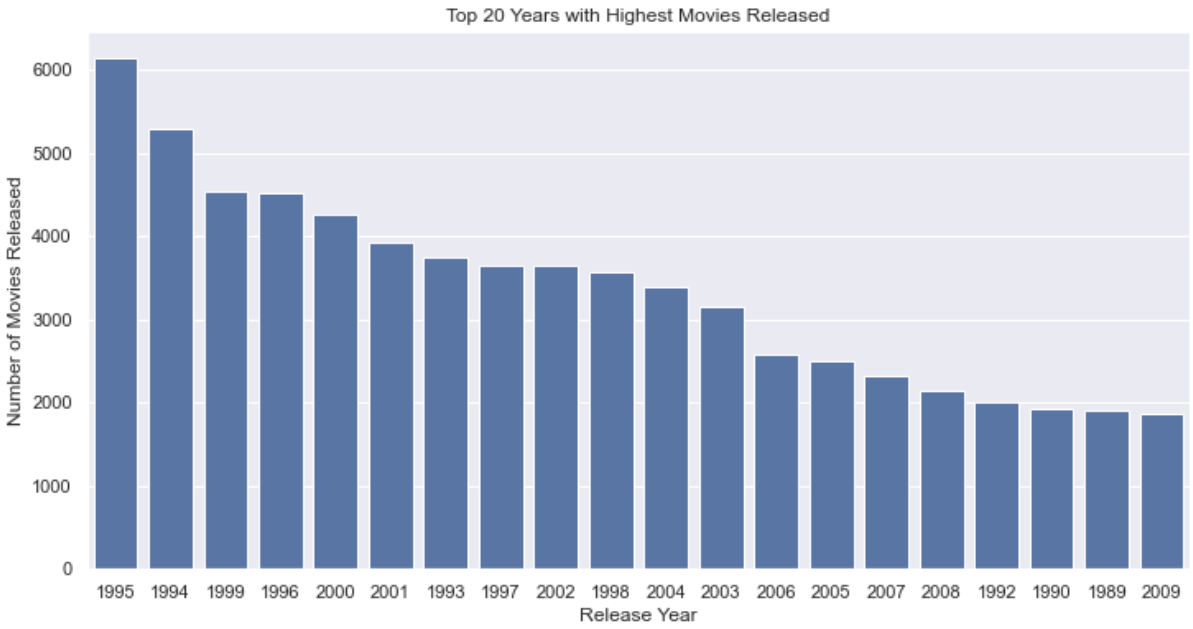


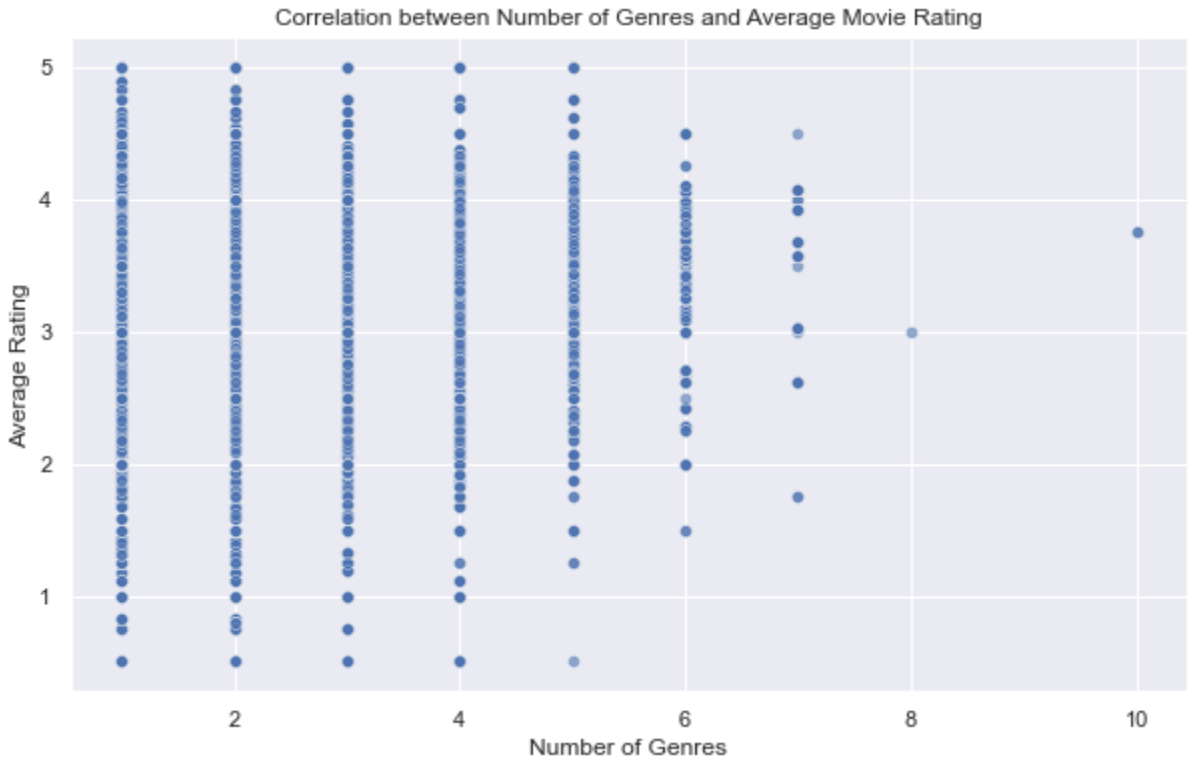

Top 20 Most-Rated Movies
Forrest Gump (1994)
Shawshank Redemption, The (1994)
Pulp Fiction (1994)
Silence of the Lambs, The (1991)
Matrix, The (1999)
Star Wars: Episode IV - A New Hope (1977)
Jurassic Park (1993)
Braveheart (1995)
Terminator 2: Judgment Day (1991)
Schindler's List (1993)
Fight Club (1999)
Toy Story (1995)
Star Wars: Episode V - The Empire Strikes Back (1980)
American Beauty (1999)
Usual Suspects, The (1995)
Seven (a.k.a. Se7en) (1995)
Independence Day (a.k.a. ID4) (1996)
Apollo 13 (1995)
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
Lord of the Rings: The Fellowship of the Ring, The (2001)
Movie Title
0
50
100
150
200
250
300
Number of Ratings


Distribution of User Average Ratings
Frequency
0
2000
4000
6000
8000
10000
12000
14000
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Average Rating


Average Rating by Primary Genre
Film-Noir
Mystery
Crime
Documentary
Western
Musical
Drama
Animation
Adventure
War
Fantasy
(no genres listed)
Sci-Fi
Action
Thriller
Comedy
Children
Romance
Horror
Primary Genre
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Average Rating


Top 10 Most Popular Movie Genres
Action
Comedy
Drama
Adventure
Crime
Animation
Horror
Children
Documentary
Mystery
Genre
0
5000
10000
15000
20000
25000
30000
Number of Ratings


Top 20 Years with Highest Movies Released
Number of Movies Released
0
1000
2000
3000
4000
5000
6000
1995
1994
1999
1996
2000
2001
1993
1997
2002
1998
2004
2003
2006
2005
2007
2008
1992
1990
1989
2009
Release Year


Correlation between Number of Genres and Average Movie Rating
Average Rating
1
2
3
4
5
2
4
6
8
10
Number of Genres