Movie Lens Dataset Feature Engineering & EDA

Stage 1 | Lawal Mayowa

This report presents the Stage 1 task of Feature Engineering and Exploratory Data Analysis (EDA) on the Movie Lens dataset.
The main objective of this task is to create new features from the dataset that can help uncover patterns in movie ratings, genres, user engagement, and movie age.
By analyzing these features, we aim to gain meaningful insights into how movies are rated, how genre and user tags influence ratings, and which movies are most engaging to users.
The insights generated in this analysis will provide a foundation for building recommendation systems in the future, using either content-based filtering or collaborative filtering methods.
The features engineered include release year, movie age, number of genres, main genre, number of tags, and the rating year. These features allow us to explore relationships between movie characteristics, ratings, and user interactions, supporting more informed and accurate movie recommendations.

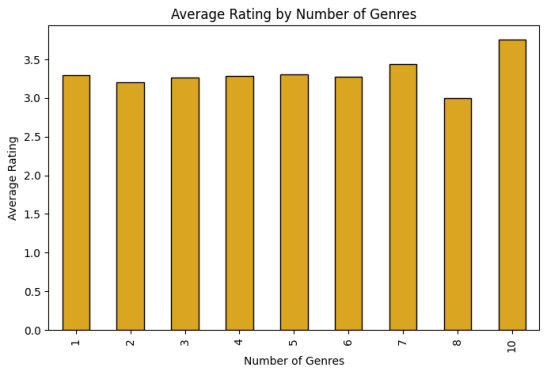
Features Created

As part of the feature engineering process, several new features were derived from the original Movie Lens dataset to help better understand movie characteristics and user ratings. These features not only enrich the dataset but also provide a foundation for exploratory analysis and future recommendation system models.

Feature Name	Description	Purpose
release_year	The year the movie was released, extracted from the title	Helps analyze movie age and trends over time
movie_age	The age of the movie in years (current year minus release year)	Useful for understanding how ratings change with movie age
num_genres	Number of genres assigned to each movie	Allows analysis of genre diversity impact on ratings
main_genre	The first genre listed for a movie	Simplifies grouping and comparison by primary genre
No_of_tags	Number of tags associated with a movie	Measures the popularity and tagging activity of movies
rating_year	The year in which the last rating was given	Useful for temporal analysis of ratings and trends
avg_rating	Average rating of the movie across all users	Core metric for analysis of movie reception
tag	Combined tags associated with a movie (text)	Enables text-based analysis and identification of common themes

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) provides a deeper understanding of the dataset and reveals patterns, trends, and relationships among movies, genres, ratings, tags, and age. The following visualizations highlight key insights that can guide recommendation strategies, content curation, and audience targeting.

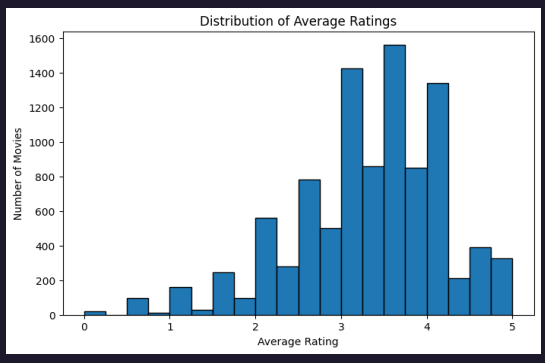


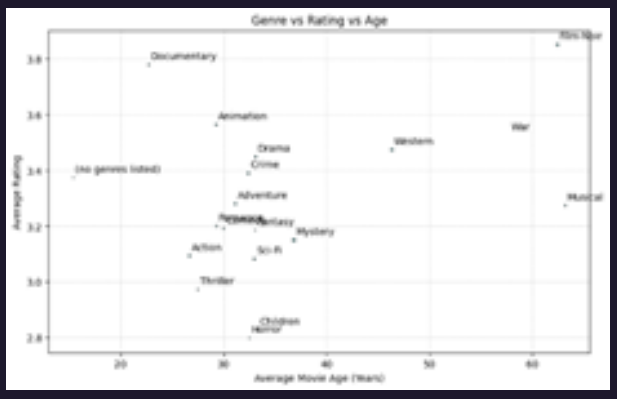
Average Rating by Number of Genres

Observation: Movies with 3–4 genres receive slightly higher ratings (~3.3–3.4).
Interpretation: Moderate genre diversity tends to appeal to audiences, while movies with too many genres (8–10) do not see significant rating improvement.
Relevance: Helps in weighting multi-genre movies in recommendation systems.

Distribution of Average Ratings

Observation: Most movies are rated between 3–4, with few extreme ratings.
Interpretation: This indicates general viewer satisfaction and balanced scoring behavior.
Relevance: Provides baseline understanding for normalization in rating-based models.



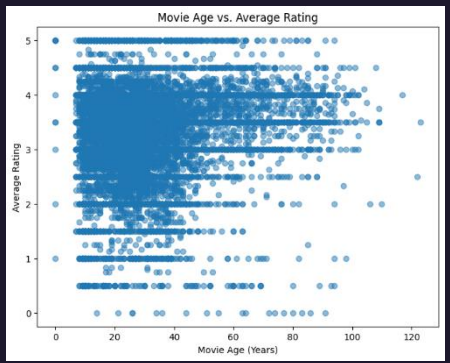


Genre vs Rating vs Age

Observation: Older Documentary and Animation films perform best (~3.5), whereas Horror consistently receives lower ratings (~2.8).
Interpretation: Genre and movie age impact audience perception.
Relevance: Useful for genre-based recommendation weighting or filtering.

Movie Age vs Average Rating

Observation: Newer movies (0–20 years) dominate rating activity, but older classics (40–60 years) maintain strong ratings.
Interpretation: Ratings remain stable over time; older movies still enjoy popularity.
Relevance: Demonstrates that age alone should not discount movies in recommendation algorithms

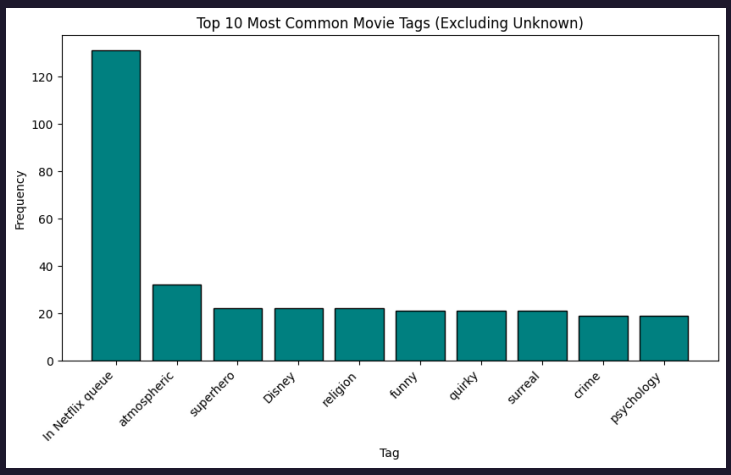


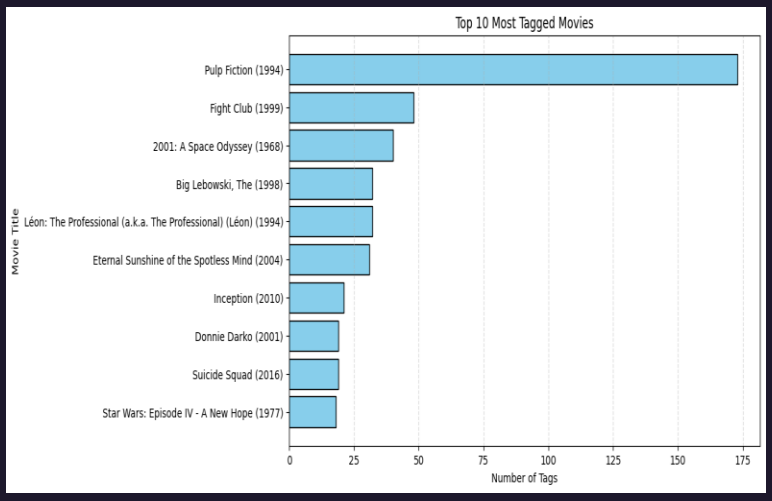
Top 10 Highest Rated Movies

Observation: Highlights the movies achieving top audience satisfaction (e.g., Won't You Be My Neighbour? — 5).
Interpretation: Shows which content consistently resonates with viewers.
Relevance: Can be used as a benchmark for quality in content recommendation.

Top 10 Most Common Movie Tags

Observation: Tags such as "In Netflix Queue", "Atmospheric", and "Superhero" are most common.
Interpretation: Tags reveal themes and viewing trends beyond official genres.
Relevance: Supports content classification and personalized recommendations.





Top 10 Most Tagged Movies

Observation: Pulp Fiction, Fight Club, and 2001: A Space Odyssey are the most tagged.
Interpretation: High tagging frequency reflects long-term engagement from the community.
Relevance: Key input for collaborative filtering and social-based recommendation systems.

Insights and Recommendation System Implications

Key Insights:

- **Genre & Tags:** Users tag movies more when emotionally connected, helping identify recommendation clusters.
- **Ratings Stability:** Ratings cluster around 3–4, so small differences can be meaningful.
- **Movie Age:** Older movies with consistent ratings can serve as "anchor" recommendations.
- **Main Genre Importance:** Some genres (Documentary, Animation) generally have higher audience approval.

How It Helps a Recommendation System:

- main_genre and num_genres help cluster movies by similarity.
- No_of_tags reflects popularity and engagement (for weighting recommendations).
- avg_rating and movie_age support balancing "trending" vs. "classic" movies.
- User tags can enhance **content-based filtering** (matching by tags + genres).

Conclusion

By exploring the MovieLens dataset and creating new features, we gained insights into audience behavior and movie characteristics:

- Movies with moderate genre diversity tend to receive higher ratings, while certain genres like Horror consistently score lower.
- Popular tags highlight trends and themes that extend beyond official genres.
- High-rated and frequently tagged movies indicate long-term audience engagement.
- Both newer releases and older classics are valued, showing that movie age alone does not determine popularity.

These insights can guide recommendation systems by combining genres, tags, ratings, and movie age to provide personalized suggestions and improve content discovery.

THANK YOU