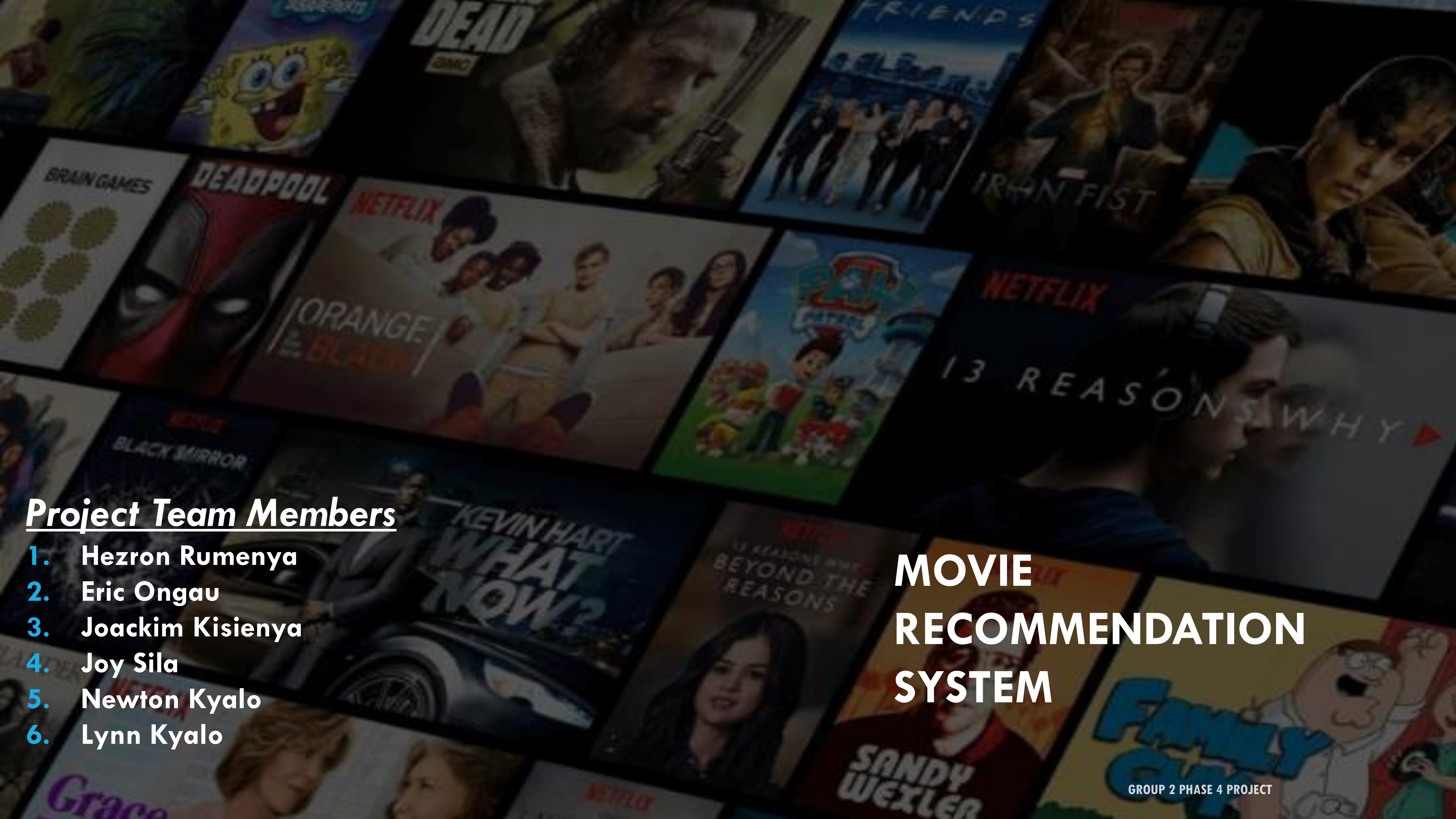


Project Team Members

1. Hezron Rumenya
2. Eric Ongau
3. Joackim Kisienya
4. Joy Sila
5. Newton Kyalo
6. Lynn Kyalo

MOVIE RECOMMENDATION SYSTEM

GROUP 2 PHASE 4 PROJECT

Business Overview

The world of streaming has transformed how we consume entertainment.

- ✓ Unlike the traditional television model - shows are limited and scheduled,
- ✓ Streaming services – Offer enormous library of movies and series at the click of a button.
- ✓ **Abundance creates its own problem - users often feel lost.**

Think about the last time you opened a streaming platform.

- ✓ Did you spend minutes scrolling through endless titles without finding anything? This is called **decision fatigue**.
- ✓ The more options available, the harder it becomes to choose, and the overall experience suffers.

A recommendation system solves this issue by acting like a **personal assistant** inside the platform.

- ✓ It studies what you like, compares you to similar users, and suggests content tailored just for you.
- ✓ Instead of browsing through thousands of movies, you are presented with a shortlist that matches your unique taste.

In this project, we demonstrate how such a system can be built.

- ✓ Using real user data from MovieLens,
- ✓ We create a model that delivers personalized recommendations, ensuring that users not only find enjoyable content but also stay engaged with the platform over time.

Business Problem

The business challenge we seek to address is simple yet crucial:

- ✓ How can we deliver personalized movie recommendations that increase user satisfaction and retention on a streaming platform?
- ✓ The market has multiple streaming services that compete for attention
  - ❑ *User experience is the deciding factor.*
  - ❑ *If users cannot easily find what they want to watch, they may abandon the platform altogether.*

The two main challenges are:

- ✓ Helping users discover content they love in a quick and effortless way.
- ✓ Keeping users engaged over time by continuously offering fresh and relevant suggestions

Without an effective recommendation system

- ❑ Platforms risk overwhelming their users,
- ❑ Leading to lower engagement and higher churn (when users cancel their subscription).

On the other hand, a well-designed system improves

- ❑ Satisfaction,
- ❑ Strengthens loyalty, and
- ❑ Directly impacts revenue.

Project Objective

The main goal of this project is to build a movie recommendation system using the MovieLens dataset.

Our objectives are:

- ✓ Analyze user ratings and preferences.
- ✓ Develop models that can predict what each user is likely to enjoy.
- ✓ Deliver Top 5 personalized movie recommendations that feel relevant and engaging.

By doing this, the system will improve user satisfaction, increase watch time, and boost platform retention.

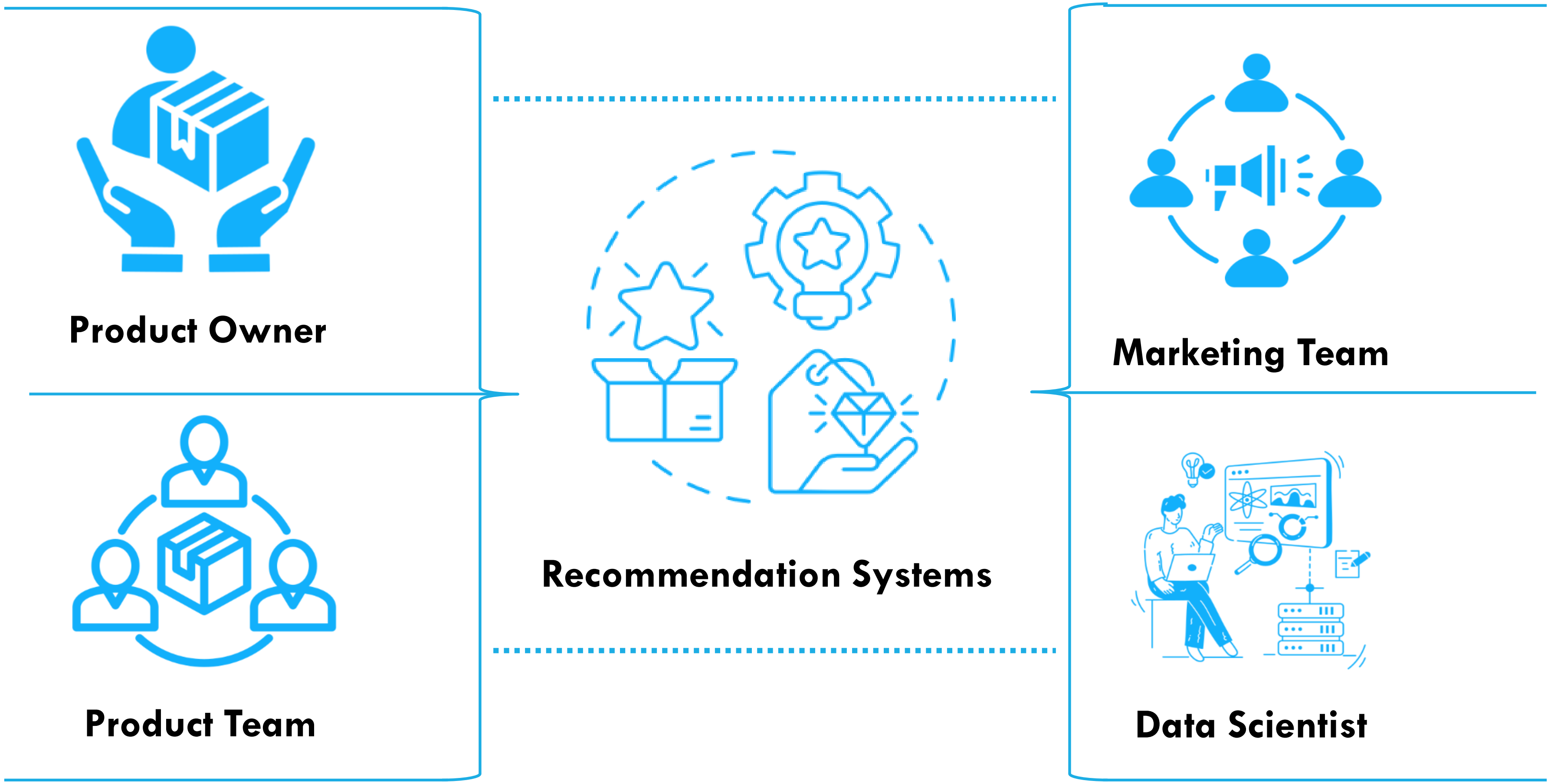
Dataset Overview

The dataset comes from **MovieLens** and includes **four CSV files:**

- **links.csv** → movie identifiers linking to external sources.
- **movies.csv** → movie titles and genres.
- **ratings.csv** → user ratings for movies.
- **tags.csv** → user-generated keywords or tags for movies.

Together, these files provide a strong foundation for building a recommendation system by capturing **user behavior**, **movie details**, and **connections between them**

Stakeholders



DATA CLEANING AND PREPARATION

To ensure the dataset was ready for building a recommendation system, several preprocessing steps were applied:

- **Removed outliers** → Ensured all ratings fall within the valid range of **0.5 to 5.0**.
- **Fixed inconsistencies** → Converted data types for user IDs, movie IDs, and ratings to the correct formats (integers and floats).
- **Normalized ratings** → Scaled rating values so they are comparable and consistent across users.
- **Encoded genres** → Transformed movie genres into numerical form using one-hot encoding, making them usable for modeling.
- **Extracted release year** → Pulled the year of release from movie titles for better feature analysis.

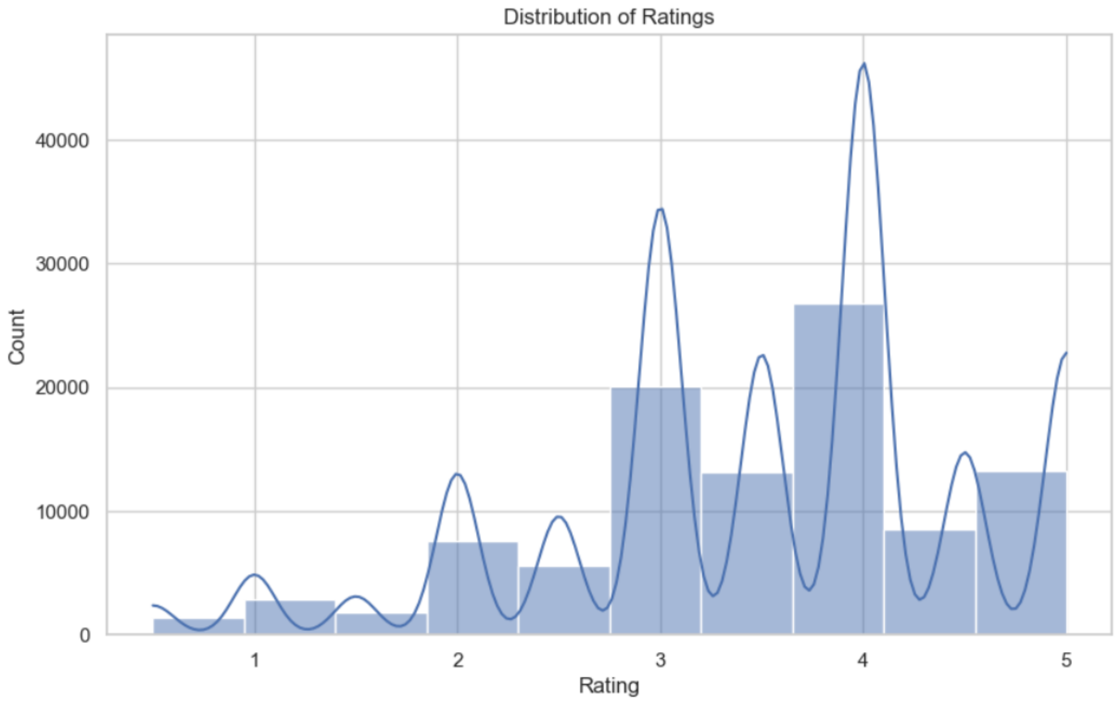
EXPLORATORY DATA ANALYSIS

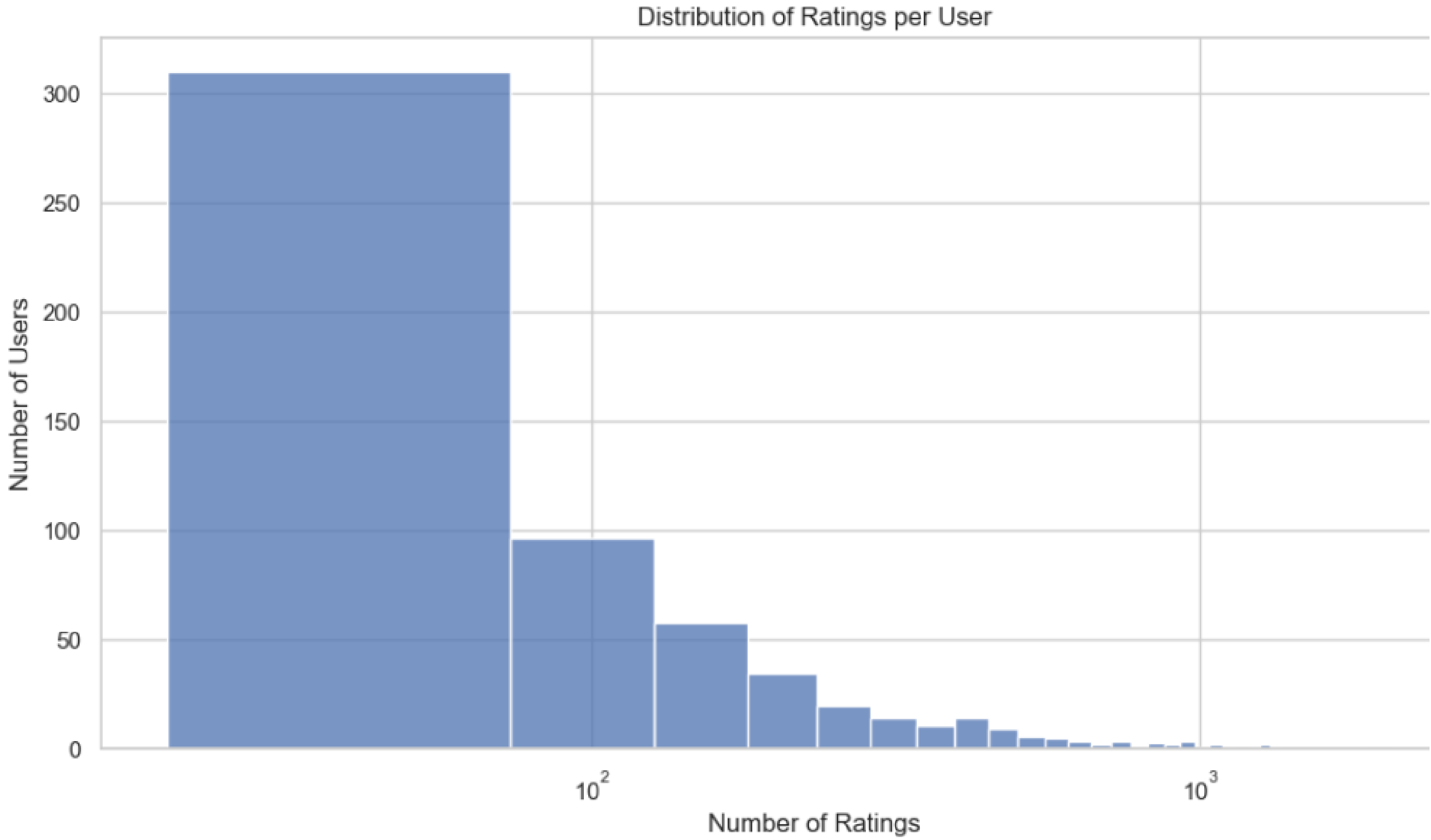
Key Observations for Distribution of Movie ratings

Most ratings are concentrated between **3 and 5,** indicating a **positive skew.**

The highest peaks occur around **ratings of 4 and 5,** showing that users tend to give **favorable reviews.**

Very few movies receive ratings below 2, suggesting that **low ratings are rare.**

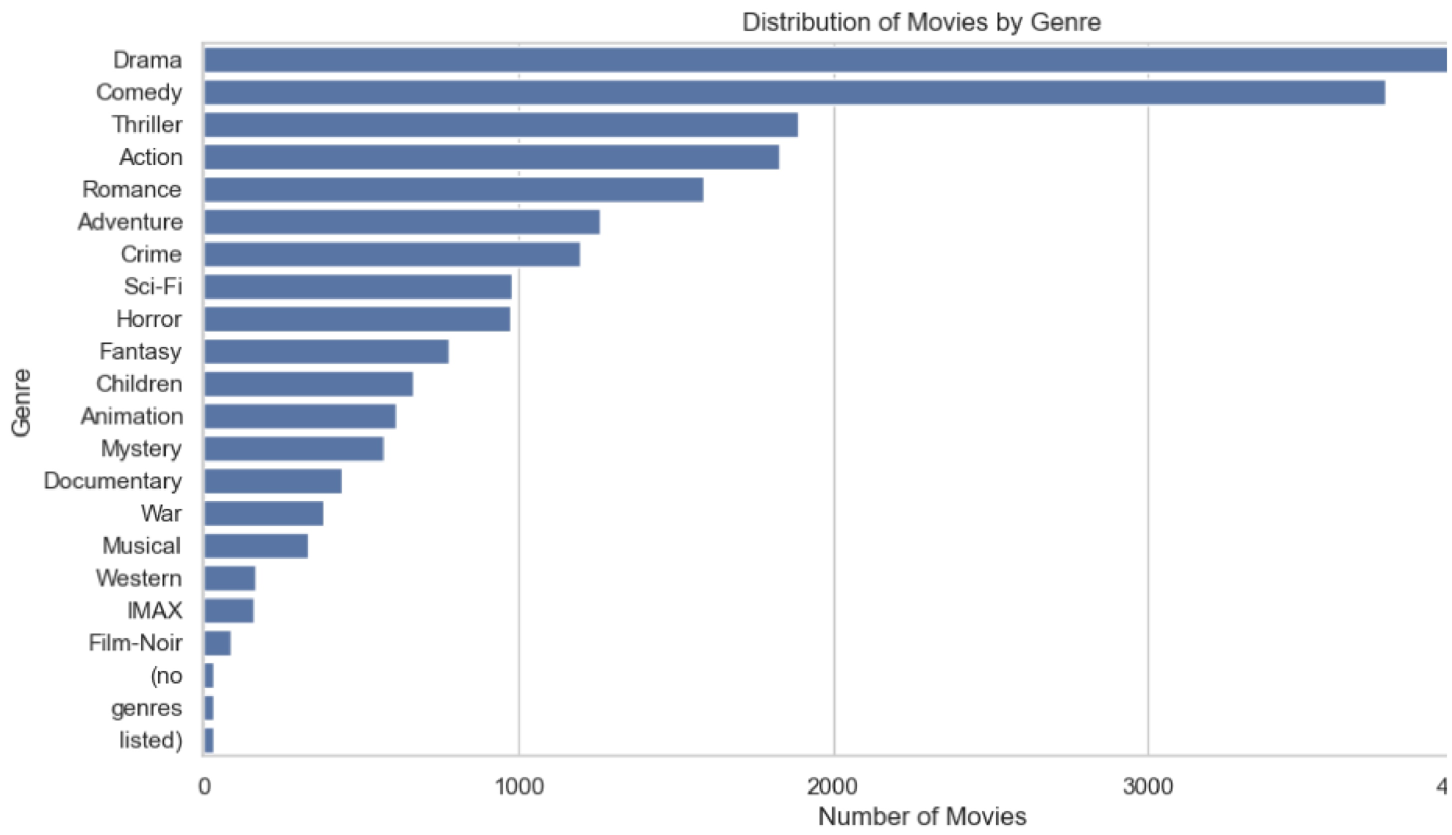




CONT.

Key Observations for Distribution of ratings per user

- The bar plot shows how many ratings each user contributed.
- Most users provided only a few ratings (under 100)
- A smaller group of users are highly active, contributing hundreds or even thousands of ratings



CONT.

Key Observations & Takeaways from the bar chart

- ✓ Drama is the most common genre, followed closely by Comedy.
- ✓ Mid-tier categories include Thriller, Action, Romance, and Adventure.
- ✓ Less common genres include Film-Noir, IMAX, and Westerns.
- ✓ Some movies do not have genres listed.

Modelling

The dataset was divided into training data (to build the models) and testing data (to evaluate them).

A user-item matrix was created showing which users rated which movies.

This step ensures our models learn patterns from past data and can be tested fairly on unseen data.

1. KNN Recommender

Finds **similar users** or **similar movies** based on their rating patterns.

Example: *“If User A likes the same movies as User B, then User A might also like movies that User B enjoyed.”*

2. SVD Recommender (Singular Value Decomposition)

A **matrix factorization model** that uncovers hidden relationships between users and movies.

Helps detect underlying patterns, like *“fans of superhero films also tend to enjoy sci-fi.”*

3. ALS Recommender (Alternating Least Squares)

Another **matrix factorization approach** used by large-scale platforms (e.g., Netflix).

Especially effective for sparse data, where most users have rated only a few movies.

MODEL OUTPUT

RMSE – Measures how close the model's predicted ratings were to what users actually gave. Lower is better.

Precision – Of the top 5 movies recommended, how many were actually relevant to the user.

Recall – Out of all the relevant movies for a user, how many were captured in the top 5

METRIC	KNN	SVD	ALS
RMSE	0.99	0.88	3.34
Precision	1.48	1.33	0.00
Recall	0.66	0.67	0.65

Conclusion

- SVD is the best overall model: It had the most accurate rating predictions and best at suggesting relevant movies.
- KNN was decent, especially in terms of relevance (Precision), but not as accurate.
- ALS performed the worst in this case, with poor accuracy and relevance.
- Cold Start Limitation: All models struggle to recommend for new users who have not rated any movies yet.
- Sparsity of Data: Since most users rate only a few movies, models like KNN may struggle more than SVD.

Recommendation

Use SVD as the main model for movie recommendations.

Deploy it using these best-found settings:

- Factors: 100
- Training cycles: 20
- Learning rate: 0.005
- Regularization: 0.02

This setup provides the **best balance between accuracy and relevance** in suggesting movies.