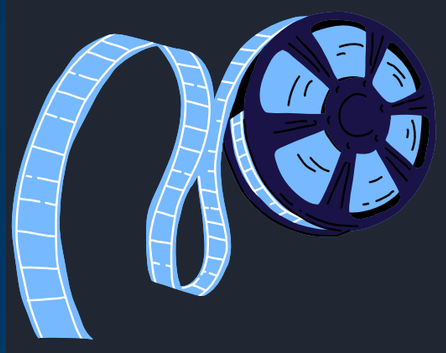
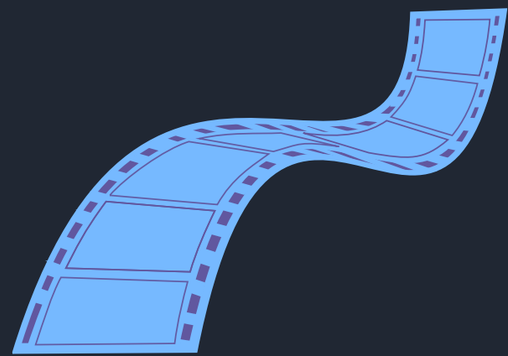
StreamTime Films Recommendation System

16/07/2025

Members:

- ❑ Calistus Mwonga
- ❑ Brian Kanyenje
- ❑ Samwel Kipkemboi
- ❑ Kelvin Mutua
- ❑ Hannah Nyambura







OVERVIEW

This project aims to build a personalized movie recommendation system using the MovieLens dataset to help users quickly find movies they are likely to enjoy. The system combines collaborative filtering and content-based filtering to make accurate, tailored recommendations. It is designed to serve both regular users and first-time users effectively. This solution can help streaming platforms improve content discovery, increase engagement, and reduce user frustration

OUTLINE

- Business problem
- Objectives
- Data overview
- Visualizations and Results
- Business recommendations
- Conclusions
- Contacts

BUSINESS PROBLEM

Users of StreamTime Films often struggle to discover content that matches their preferences due to the vast number of available options. With thousands of movies to choose from, users often feel and find it difficult to locate movies that match their interests, since most don't scroll deeply or explore the site extensively. This leads to frustration, decision fatigue and in some cases abandoning the platform.

As a result, StreamTime Films faces a critical challenge in retaining users and maintaining long-term engagement, which directly impacts business sustainability.

OBJECTIVES

1

Identify the distribution of the number of ratings per movie.

2

Investigating how users rate movies: rating distribution, biases, and engagement levels

3

Build content-based and collaborative recommendation systems

4

Build a Hybrid Recommender System to merge
The best of both worlds

DATA OVERVIEW

Dataset Source: The MovieLens dataset, specifically the '**ml-latest**' version.

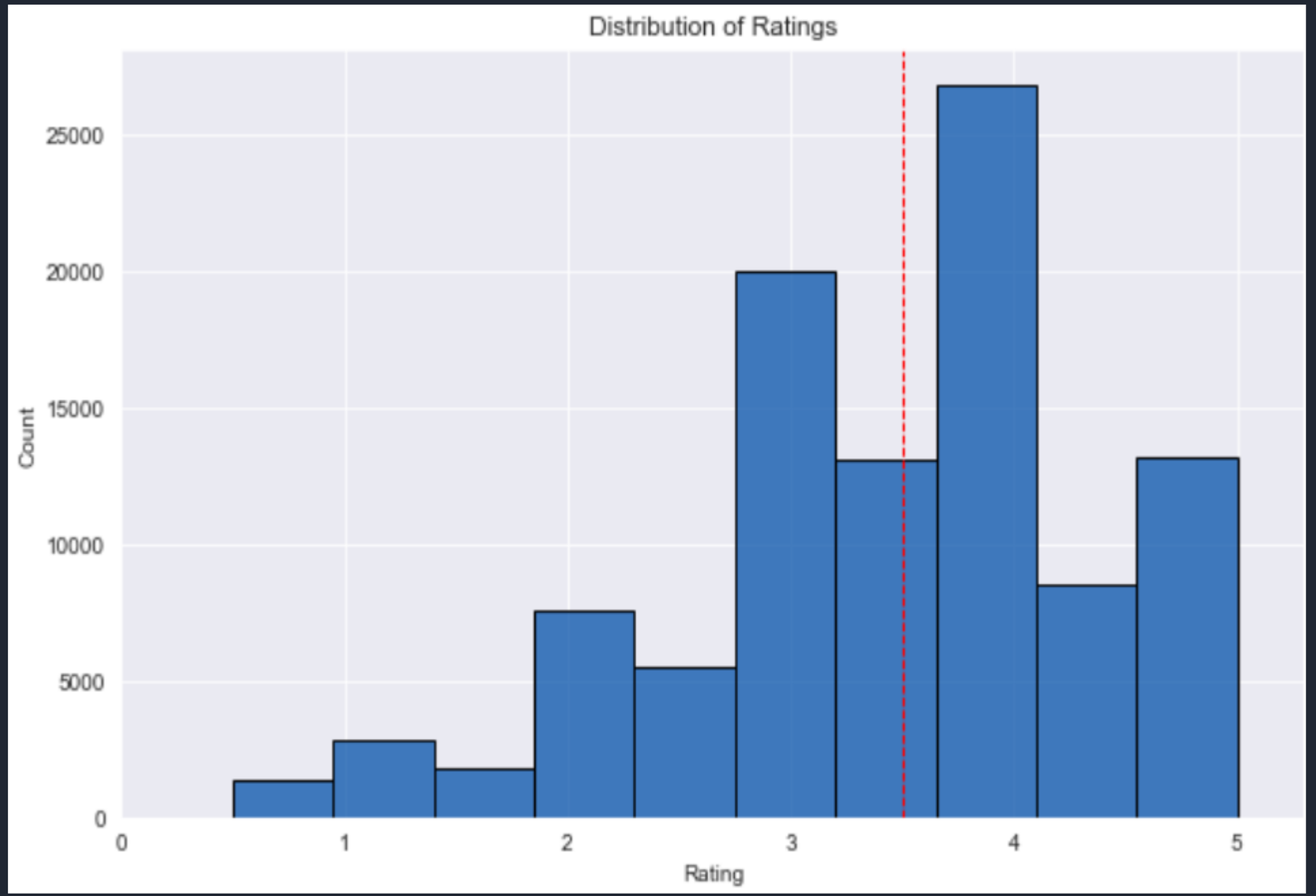
The dataset is quite large, containing over 100,000 ratings and information on nearly 10,000 movies.

What's Included: This dataset contains several files that give us different pieces of information:

- **ratings.csv:** This is the core file, containing user ratings for movies (User ID, Movie ID, Rating, Timestamp).
- **movies.csv:** This file provides details about the movies themselves (Movie ID, Title, Genres).
- **tags.csv:** This file includes user-generated tags applied to movies (User ID, Movie ID, Tag, Timestamp).

The key connection between all these files is the movieId column. This allows us to link user ratings, movie details, user tags, and other external information together.

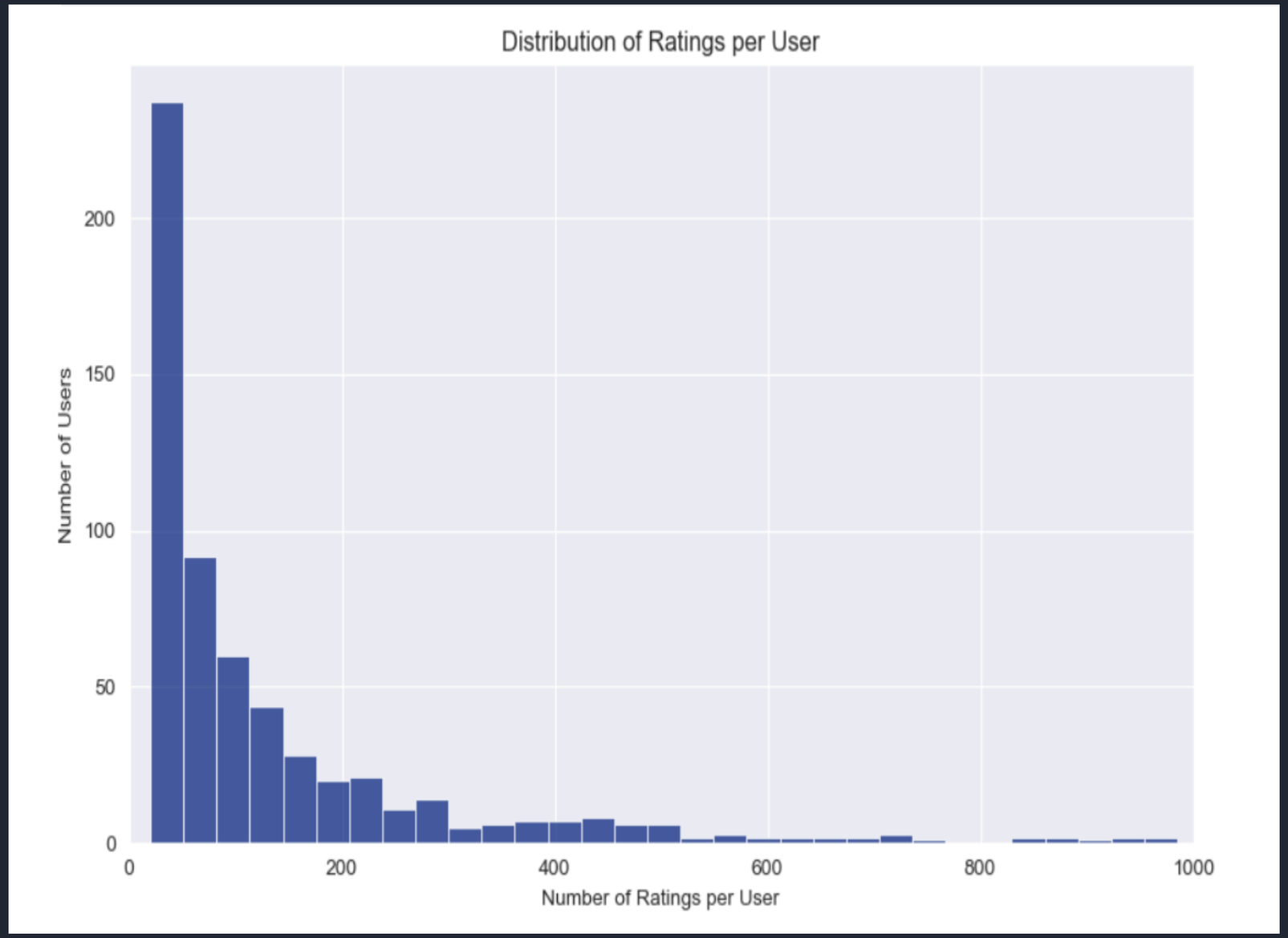
VISUALIZATIONS



Observation: Many users tend to give positive ratings with the peak being around 4.0 and the mean around 3.5

Insight for Modeling: This shows us we have good positive feedback to learn from and that users provide detailed preferences. It also suggests we need models that can handle variations in how users rate.

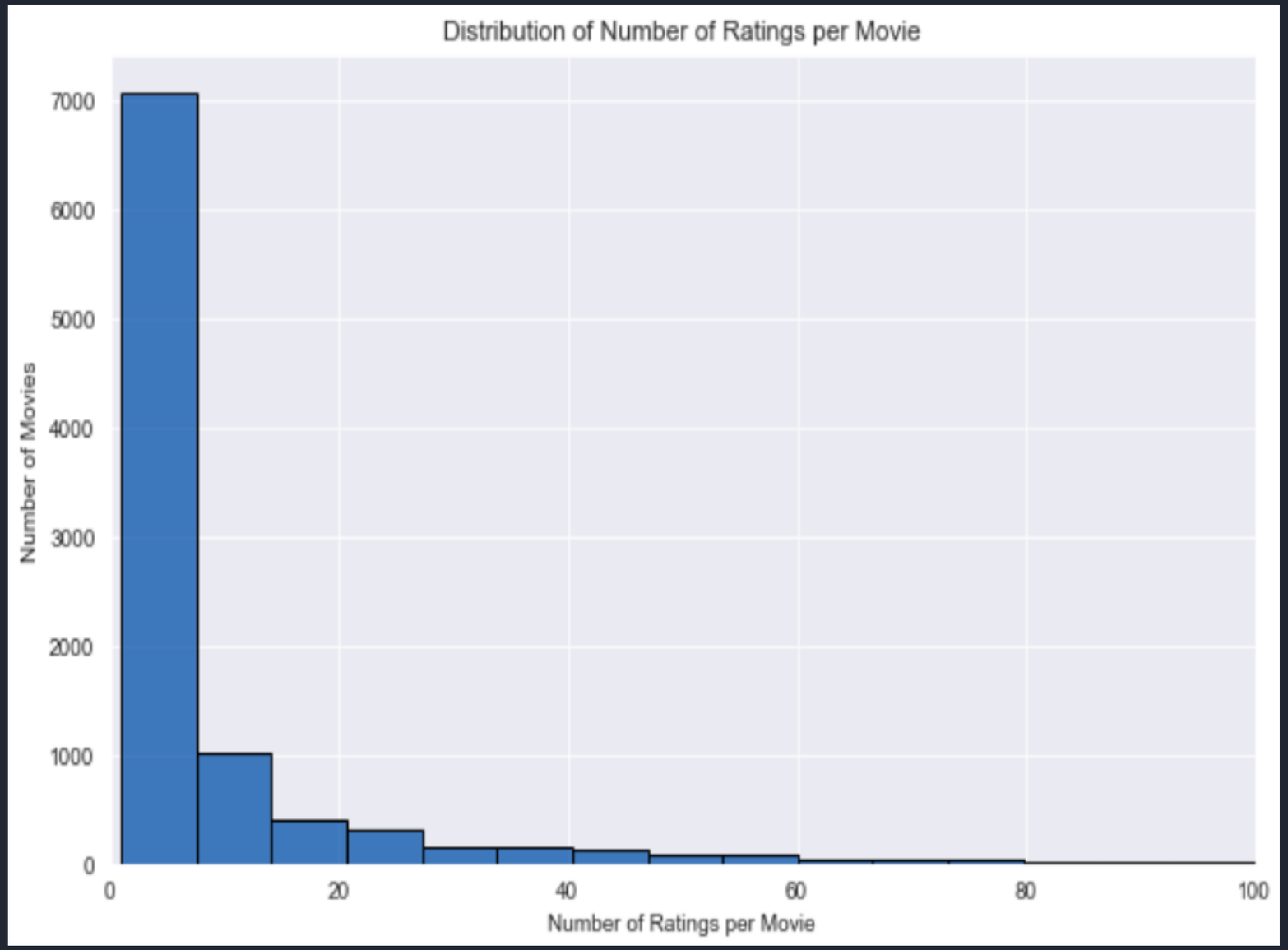
VISUALIZATIONS



Observation: Most users have only rated a small number of movies, while a few users rate very many.

Insight: This clearly shows the **cold-start user challenge** – many users don't have enough history for standard recommendations. This insight was key to deciding we needed a hybrid system that can work even with limited user data.

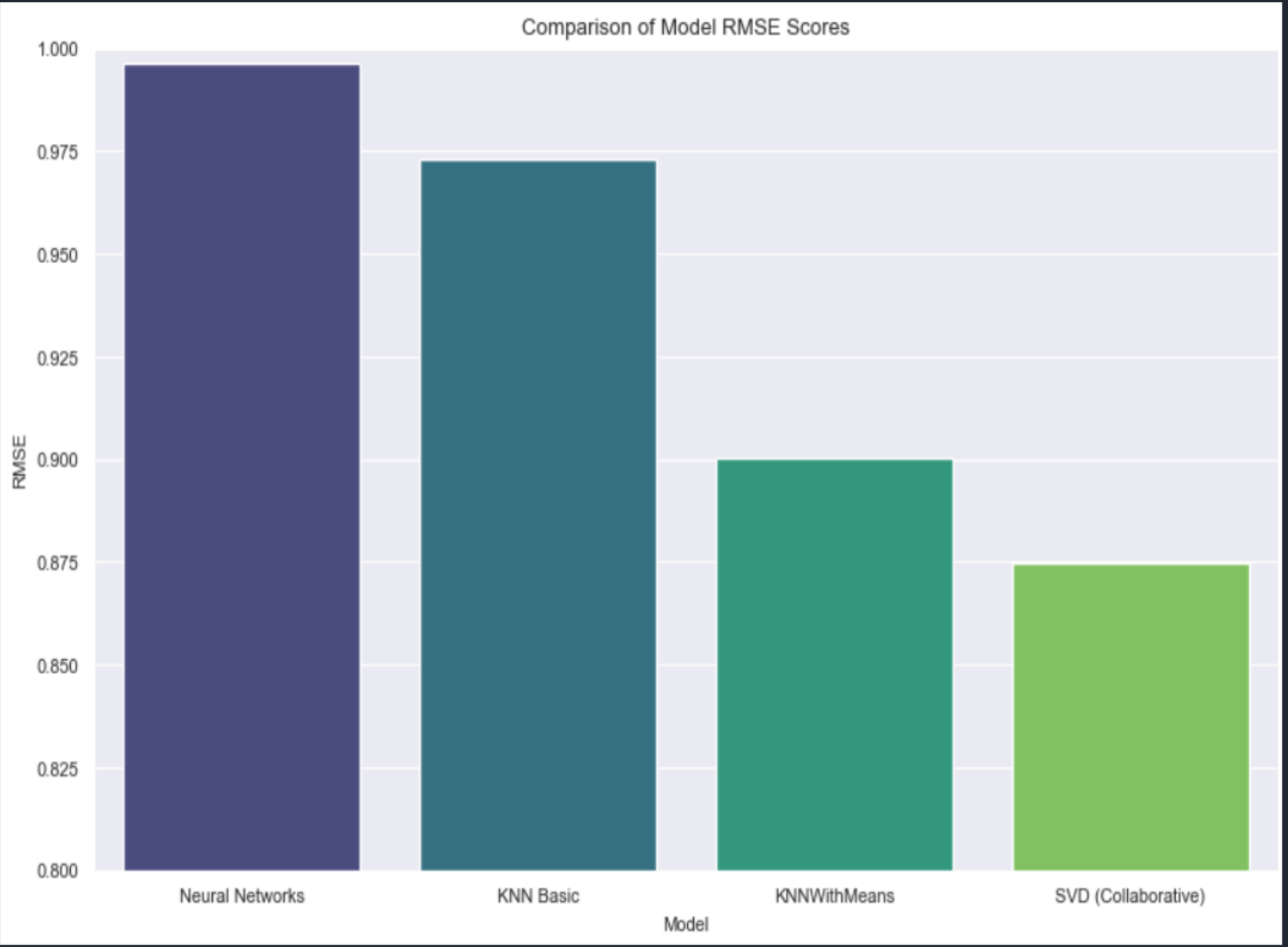
VISUALIZATIONS



Observation: This histogram shows that most movies in our dataset have been rated by a very small number of users**.** The tall bar on the left side, representing movies with low rating counts.

Insight for Modeling: This pattern confirms that we have a significant cold-start item problem. This insight is crucial because it means our system needs strategies, such as **content-based filtering** (using movie genres and tags), to effectively recommend these less-rated movies based on their characteristics, not just their popularity.

RESULTS



The RMSE score is used to measure how close our predicted rating for a movie was to the actual rating a user gave. The lower the bar the better the accuracy. The goal is to get this number as low as possible.

Observation: The bar for the SVD model is the lowest while the Neural Network is the largest.

Insight: The SVD was the most accurate in predicting user ratings among those tested.

BUSINESS RECOMMENDATION

1

Utilize the SVD model’s accuracy to power personalized recommendations of movies to specific users

2

Implement a weighted hybrid model to combine the strengths of collaborative filtering and content-based filtering, improving recommendation diversity and accuracy.

3

Leverage the model insights to surface highly rated but under-watched movies and push them to relevant users increases exposure for lesser-known content

4

Invest in encouraging users to contribute more high-quality tags which will help reduce noise and improve content discoverability for niche titles.

FUTURE WORKS

1

Incorporate Implicit Feedback e.g., click-throughs, hovering, adding to favorites,...They give a complete picture of what users like thereby improving recommendation accuracy

2

Incorporate time-aware methods to understand evolving user preferences and item trends. This ensures recommendations adapt and remain relevant over time.

3

Explore Advanced Modeling Techniques: such as more Advanced Neural Collaborative Filtering or sequential models, to capture more complex patterns.

4

Expand the set of features used to describe movies beyond genres and tags, e.g., information from cast or crew data. It will help improve the system's ability to handle new or niche items.

CONCLUSION

This project successfully built a multi-faceted movie recommendation system. By integrating collaborative filtering (KNN, KNNWithMeans, SVD), content-based filtering (tags + genre), and hybrid approaches.

We were able to: Capture user-specific preferences, address cold-start scenarios, and generate diverse and interpretable recommendations. The optimal model (SVD) achieved the lowest RMSE (~0.87), indicating strong predictive performance. The hybrid models further improved flexibility by dynamically adapting to user history depth..

For any additional questions, please contact these members:

- Calistus Mwonga, calistusmwonga@gmail.com
- Samwel Kipkemboi, samkemboi201@gmail.com
- Brian Kanyenje, bkanyenje@gmail.com
- Hannah Nyambura, anngachuhipg1@gmail.com
- Kelvin Mutua, kelvinmutua787@gmail.com