

# Movie lens recommendation system

Prepared by Gilbert Kipkirui Cheruiyot,  
DSF-PT08P4 Project

A collage of film-related items. In the background, a clapperboard is visible with text like 'SCENE', 'DIRECTOR', and 'PRODUCER'. In the foreground, there are several film reels of different sizes, some with film strips attached. At the bottom, there are yellow movie tickets with the words 'CINEMA' and 'ADMIT' visible. A dark, semi-transparent rectangular box is centered over the image, containing the text 'A MOVIE IS WORTH A THOUSAND WORDS' in a serif font. The text is white for 'A MOVIE IS WORTH A' and 'THOUSAND WORDS', and gold for 'A' and 'THOUSAND'.

**A MOVIE IS WORTH A  
THOUSAND WORDS**

# TABLE OF CONTENTS

**I.**

**Introduction**

**II.**

**Data  
Understanding**

**III.**

**Exploratory Data  
Analysis**

**IV.**

**Modelling &  
Evaluation**

**V.**

**Recommendation**

# ★ INTRODUCTION ★

The audience are in need of a model that provides top 5 movie recommendations to a user, based on their ratings of other movies. This is to ease their movie selection for them to find movies they will enjoy. ★



# Data Understanding



## Data Source:

- <https://grouplens.org/datasets/movielens/latest/>

★ **The dataset files include:** ratings.csv, tags.csv, movies.csv, and links.csv

# Data Understanding

Merged dataset include fields from movie.csv, tags.csv and rating.csv. This is meant to provide relevant information about the movies the model will recommend.

	userId	movieId	rating	title	genres	tag
0	1	1	4.0	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	pixar
1	1	1	4.0	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	pixar
2	1	1	4.0	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	fun
3	1	3	4.0	Grumpier Old Men (1995)	Comedy Romance	moldy
4	1	3	4.0	Grumpier Old Men (1995)	Comedy Romance	old

# Data Understanding

Index: 219406 entries, 0 to 233212

Data columns (total 6 columns):

#	Column	Non-Null	Count	Dtype
0	userId	219406	non-null	int64
1	movieId	219406	non-null	int64
2	rating	219406	non-null	float64
3	title	219406	non-null	object
4	genres	219406	non-null	object
5	tag	219406	non-null	object

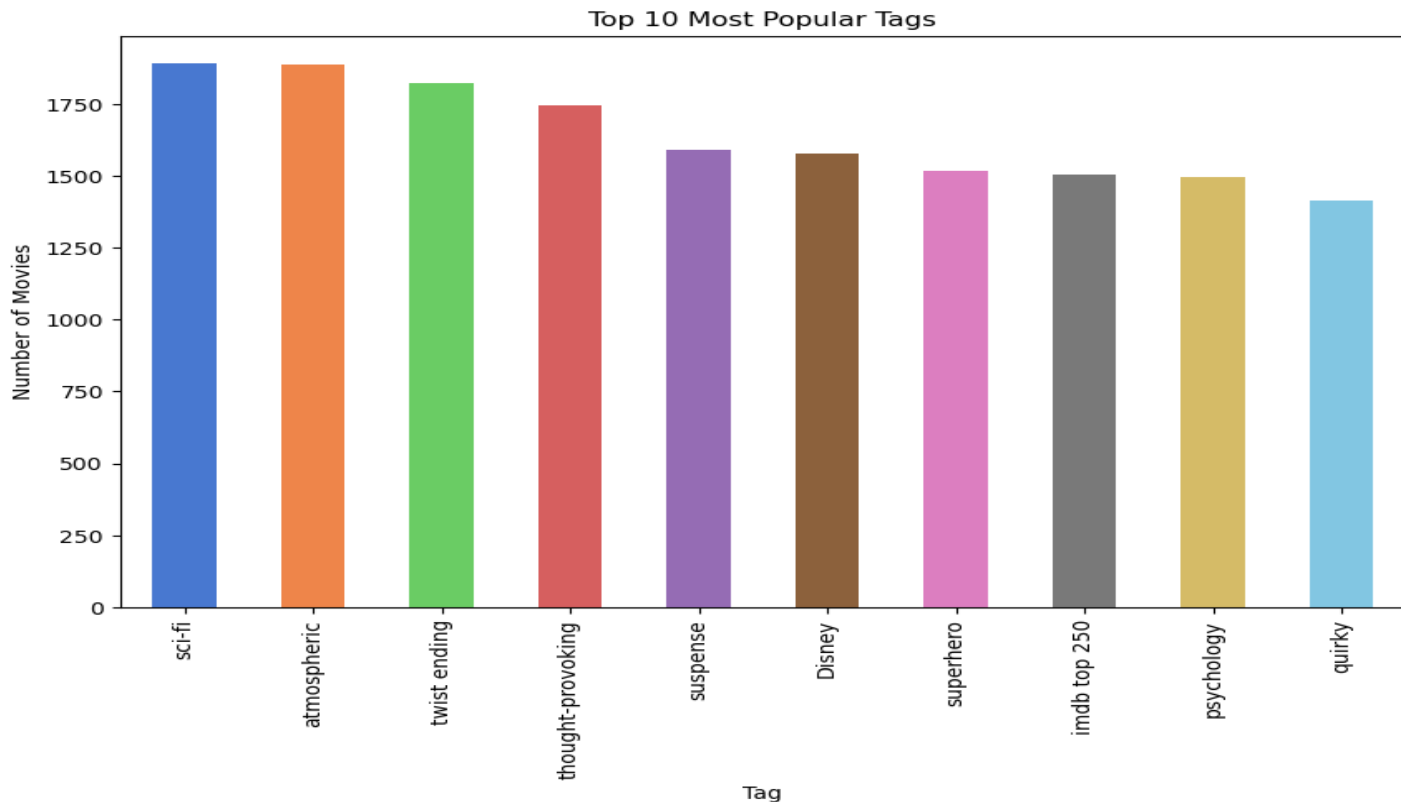
dtypes: float64(1), int64(2), object(3)

memory usage: 11.7+ MB

- Merged dataset has no null values.

- The dataset for the six columns are 1 float, 2 are of integer type and 3 of object type.

# EXPLORATORY DATA ANALYSIS (EDA)

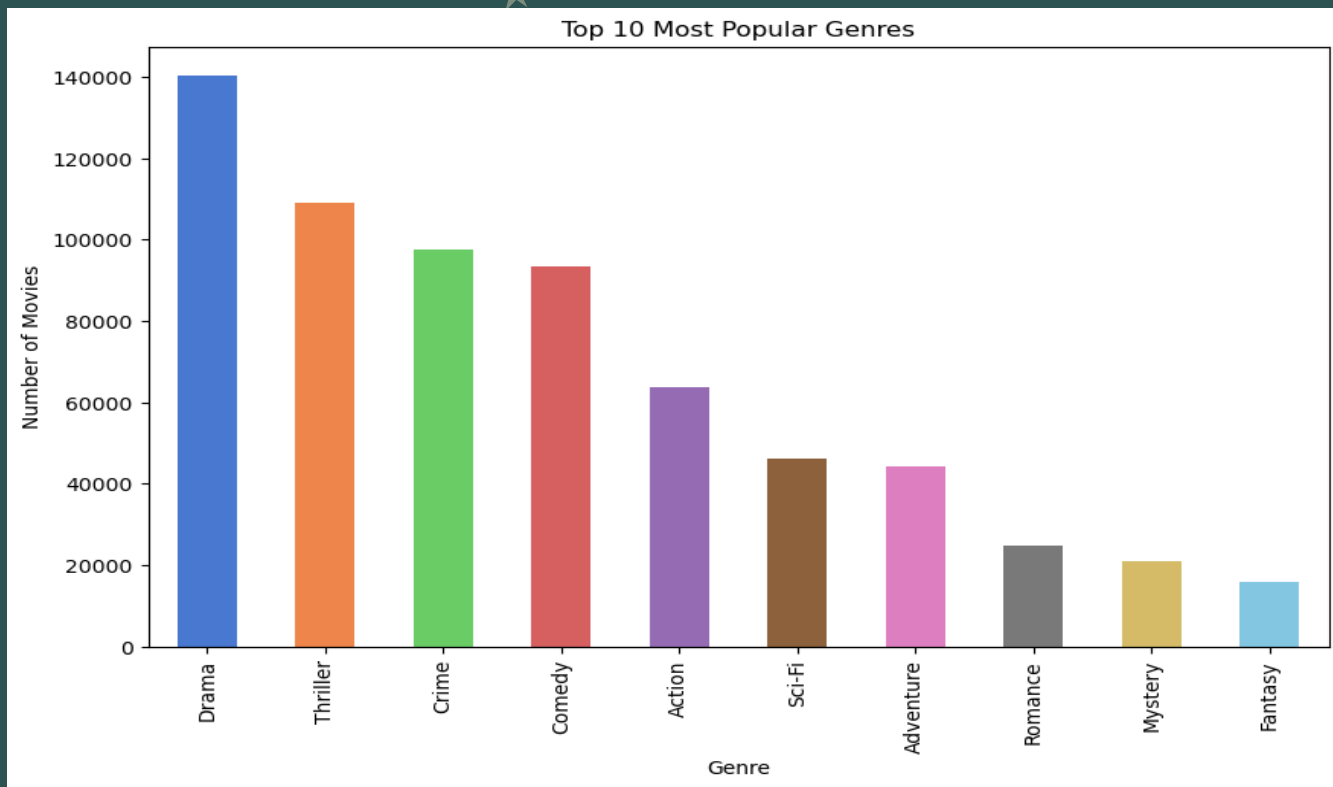


Sci-fi, atmospheric and twist ending are most popular movie tags.





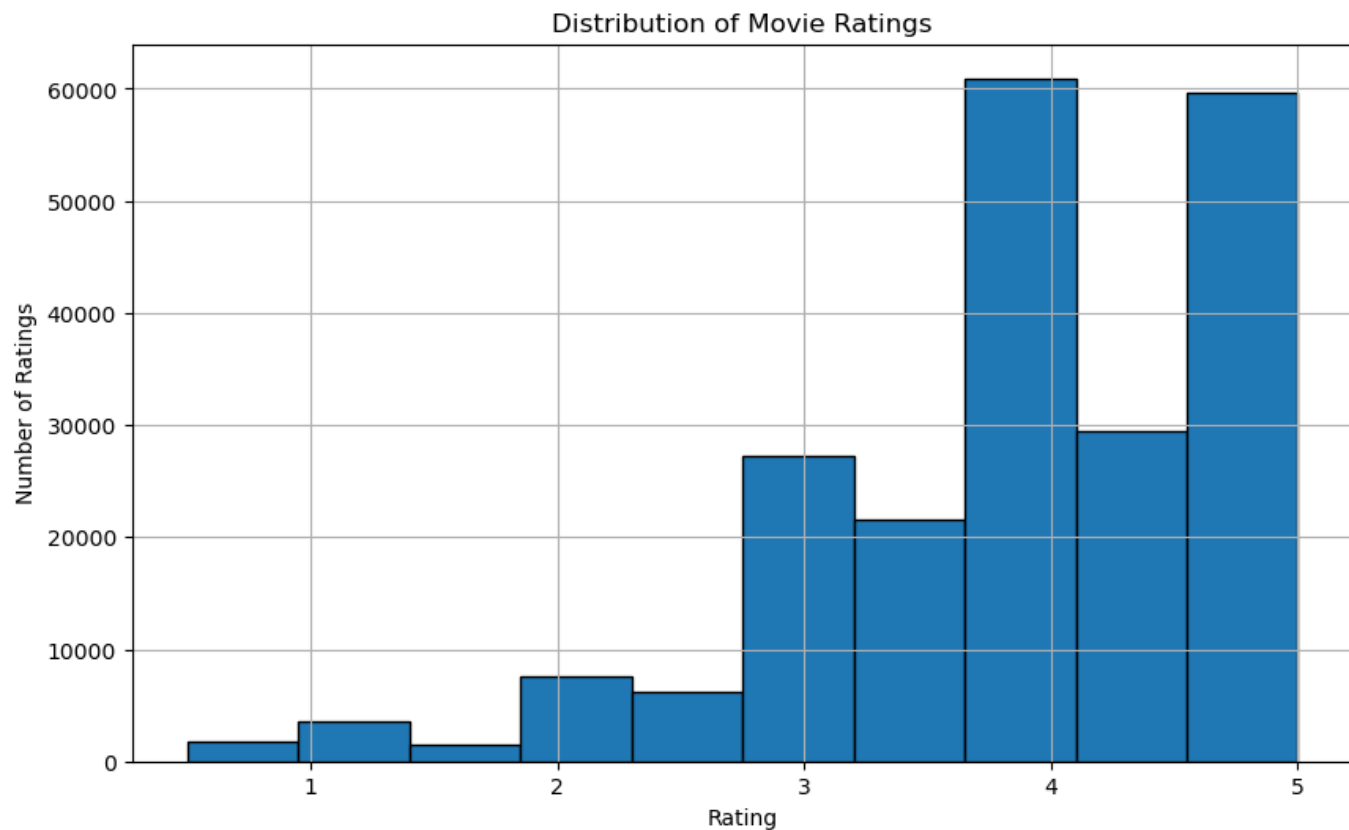
# EXPLORATORY DATA ANALYSIS (EDA)



Drama is the most popular genre of movies followed by Triller.



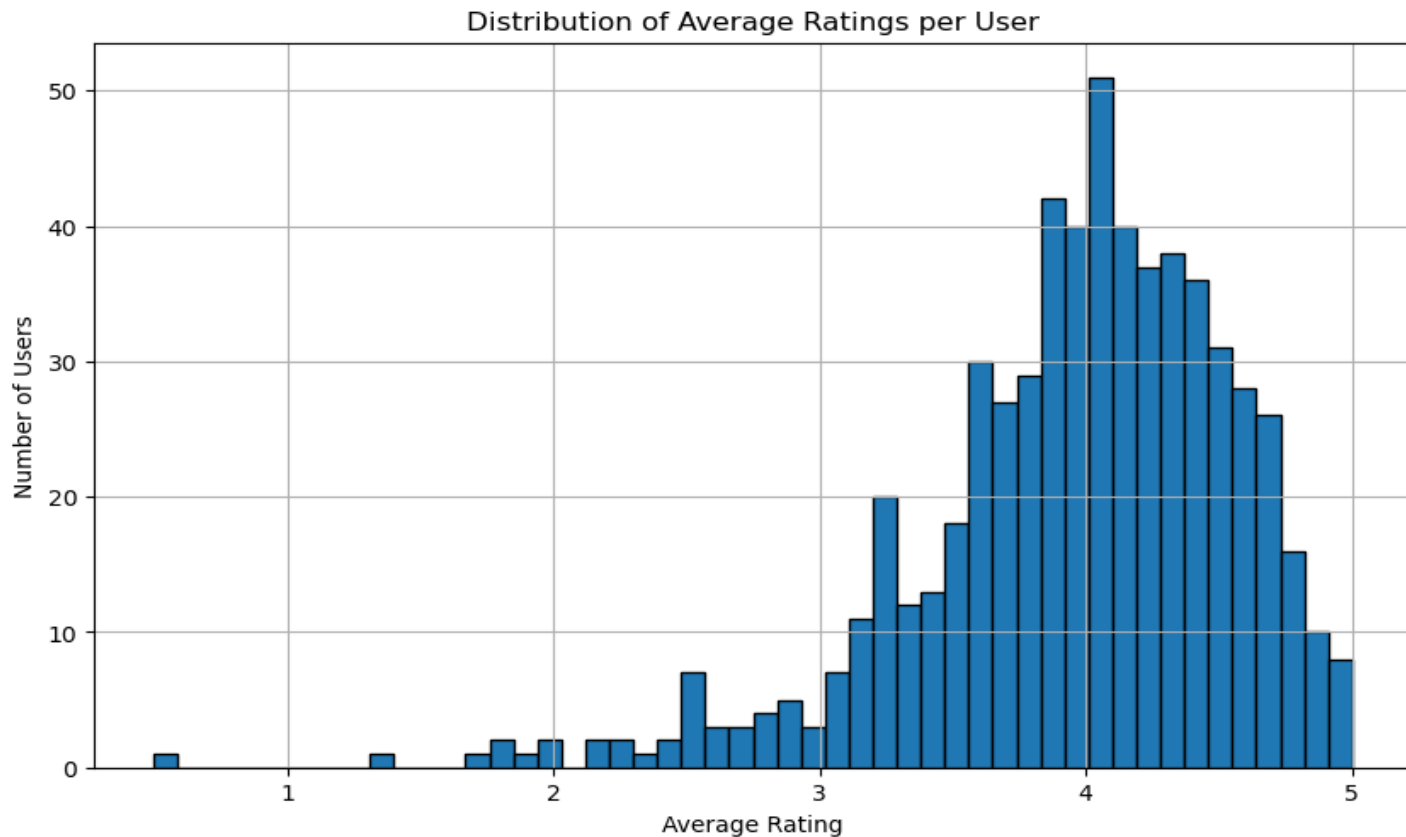
# EXPLORATORY DATA ANALYSIS (EDA)



The distribution of the movie ratings for the movies is highly skewed to the right.



# ★ EXPLORATORY DATA ANALYSIS (EDA)

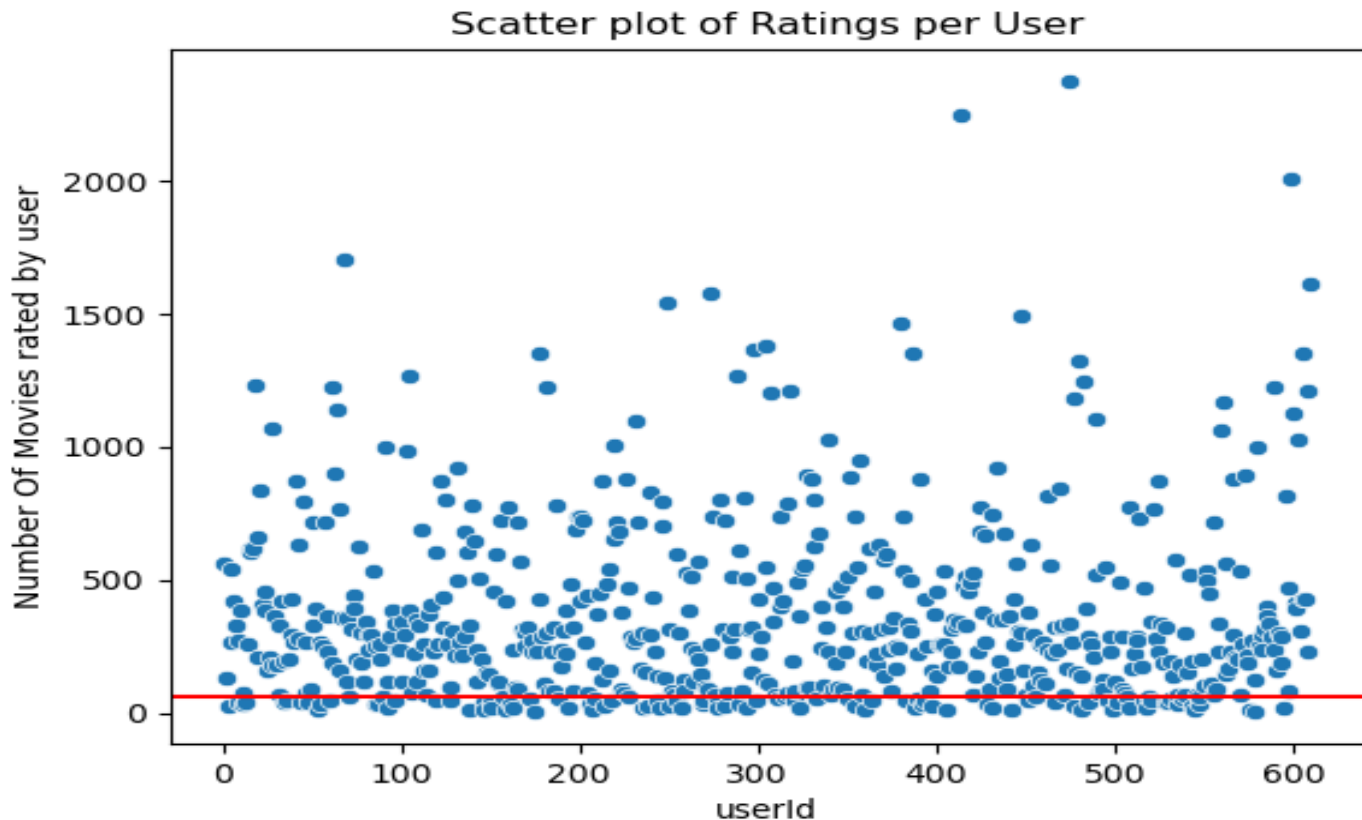


★

The average rating for most users appears to be around 4.



# ★ EXPLORATORY DATA ANALYSIS (EDA)



★

The scatter plot suggests a larger proportion of the users have rated relatively fewer number of movies.



# MODELLING & EVALUATION

Models prepared include:



**Collaborative  
Filtering (SVD)**



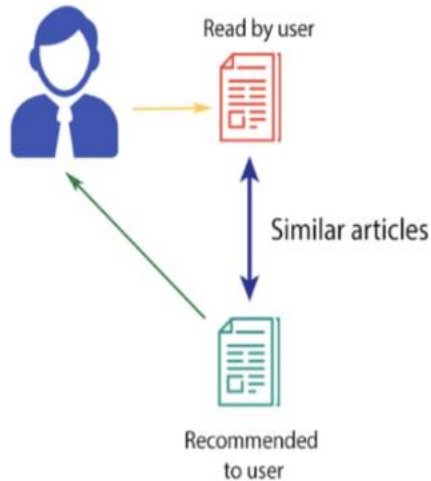
**Content-Based  
Filtering**



**Hybrid Approach**

# MODELLING & EVALUATION

CONTENT-BASED FILTERING



COLLABORATIVE FILTERING

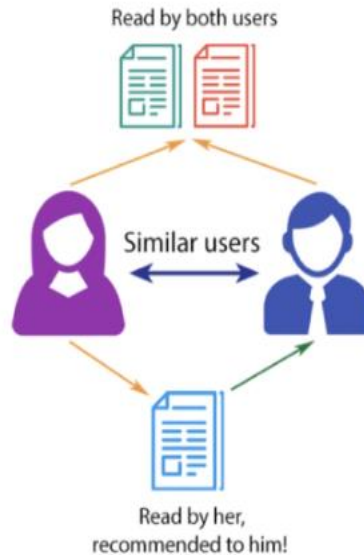


Illustration of 2 types of recommendation systems- Content-based Filtering (CBF) and Collaborative Filtering (CF).



# Collaborative Filtering with SVD Model

Top 5 Recommendations movies as per Collaborative Filtering with SVD model for UserID 1:

1. Shawshank Redemption, The (1994) - Rating: 5.00
2. Philadelphia Story, The (1940) - Rating: 5.00
3. Commitments, The (1991) - Rating: 5.00
4. Lord of the Rings: The Fellowship of the Ring, The (2001) - Rating: 5.00
5. Lord of the Rings: The Two Towers, The (2002) - Rating: 5.00

The Root Mean Squared Error (RMSE) for the SVD model is: 0.29501380508803526

**The low RMSE shows that the model is reliable to provide personalised recommendations.**

**All movies predicted for UserId 1 are 5-star rated**

# Content-Based Filtering Model

Top 5 movies as per Content-Based Filtering Model Recommendations for UserId 1:

1. Toy Story (1995) (Rating: 5.0)
2. Grumpier Old Men (1995) (Rating: 5.0)
3. Seven (a.k.a. Se7en) (1995) (Rating: 5.0)
4. Usual Suspects, The (1995) (Rating: 5.0)
5. Bottle Rocket (1996) (Rating: 5.0)

Content-Based Filtering RMSE: 0.7180703308172536

**Compared to collaborative filtering with SVD, the model's higher RMSE suggests less accuracy in prediction.**



# Hybrid Approach

Hybrid Approach Movie Recommendations for UserId 1:

1. Grumpier Old Men (1995) (Rating: 5.0)
2. Bottle Rocket (1996) (Rating: 5.0)
3. Shawshank Redemption, The (1994) (Rating: 5.0)
4. Lord of the Rings: The Fellowship of the Ring, The (2001) (Rating: 5.0)
5. Usual Suspects, The (1995) (Rating: 5.0)

The Root Mean Squared Error (RMSE) for the hybrid model is: 0.8660

**Despite the higher RMSE, the benefits of combining the two preceding models might provide more balanced and well-rounded recommendations.**

# Findings

\* All the three models were able to identify movies with the highest rating of 5.

\* The collaborative filtering with singular value decomposition (SVD) model with a low RMSE indicates higher accuracy as compared to the hybrid and content based filtering model.

# Findings...Continuation

\* Content-Based Filtering method provides moderate RSME, indicating fairly good accuracy. This can be a good alternative if the user has rich content features and want to recommend items based on the attributes of movies the user has liked.

\* The hybrid model appears to take into consideration both collaborative filtering and content-based filtering model details. The moderately higher Root Mean Square Error (RMSE) indicates slightly lower prediction accuracy compared to the other two models.

# Recommendation

\* Because of its accuracy and capacity to employ user-specific interaction data, Collaborative Filtering (SVD) model is the best option given the task's emphasis on user ratings. Its accuracy and simplicity makes it a favourable recommendation model.

# END!



# Contacts

**Does anyone have any questions?**

Do not hesitate to reach out on email address:  
[gilokipkirui@gmail.com](mailto:gilokipkirui@gmail.com)