# RMIT University

***COSC2670***

*Practical Data Science with Python*

## *Assignment 3: Recommender Systems (Virtual Presentation)*

By-

Rakshit Chandna s3956924

# Introduction

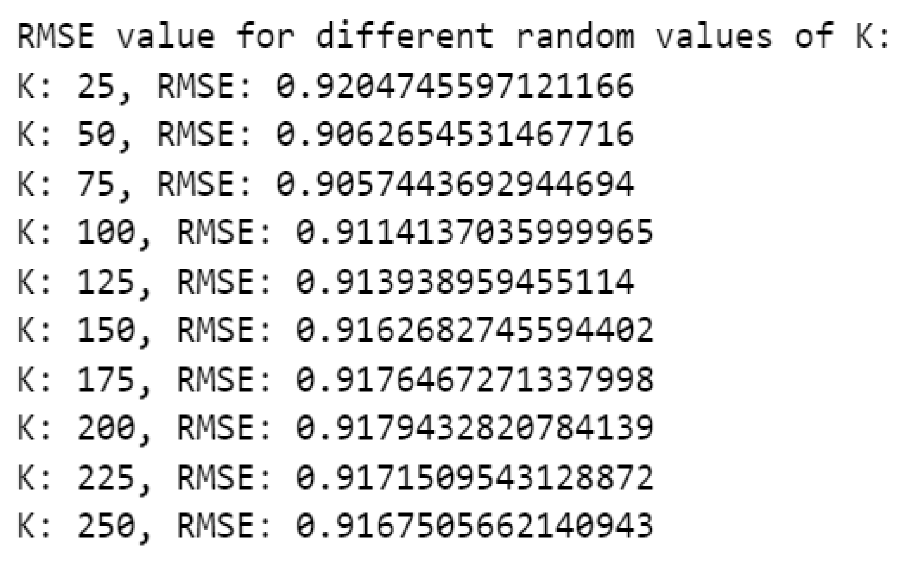
Our analysis revolves around the **MovieLens 1M** Dataset, a rich collection of movie ratings by numerous users. We started by loading the dataset and conducting preliminary checks for data integrity.

Following a brief exploration, we split the data into training (80%) and testing (20%) sets. By transforming this data into *user-item matrices*, we're now set to build and refine our recommendation system. As we delve deeper, we'll uncover insights and methodologies to recommend top movies tailored for each user's preferences with the help of ***Collaborative Filtering***.

https://medium.com/web-mining-is688-spring-2021/most-popular-movie-genre-combinations-db02ccd6bc45

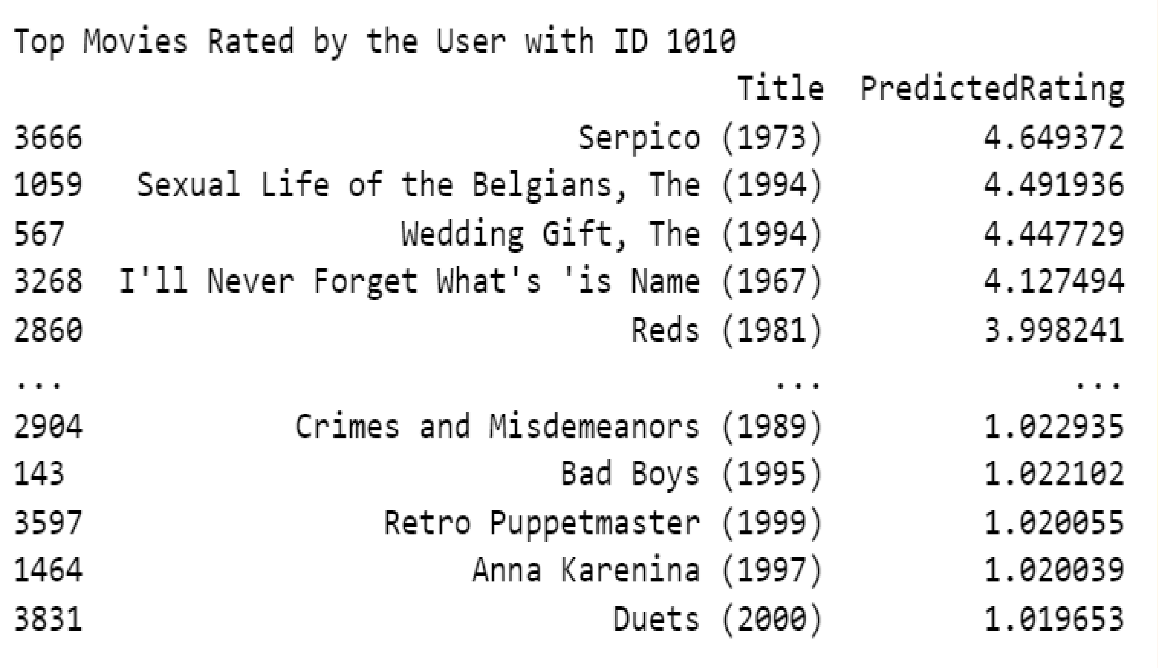
# Task 1: User-based Collaborative Filtering

### - RMSE for evaluation



```
RMSE value for different random values of K:
K: 25, RMSE: 0.920474559712116б
K: 50, RMSE: 0.9062654531467716
K: 75, RMSE: 0.9057443692944694
K: 100, RMSE: 0.9114137035999965
K: 125, RMSE: 0.913938959455114
K: 150, RMSE: 0.9162682745594402
K: 175, RMSE: 0.9176467271337998
K: 200, RMSE: 0.9179432820784139
K: 225, RMSE: 0.9171509543128872
K: 250, RMSE: 0.9167505662140943
```

### - Top Recommended Movies wrt Predicted Ratings

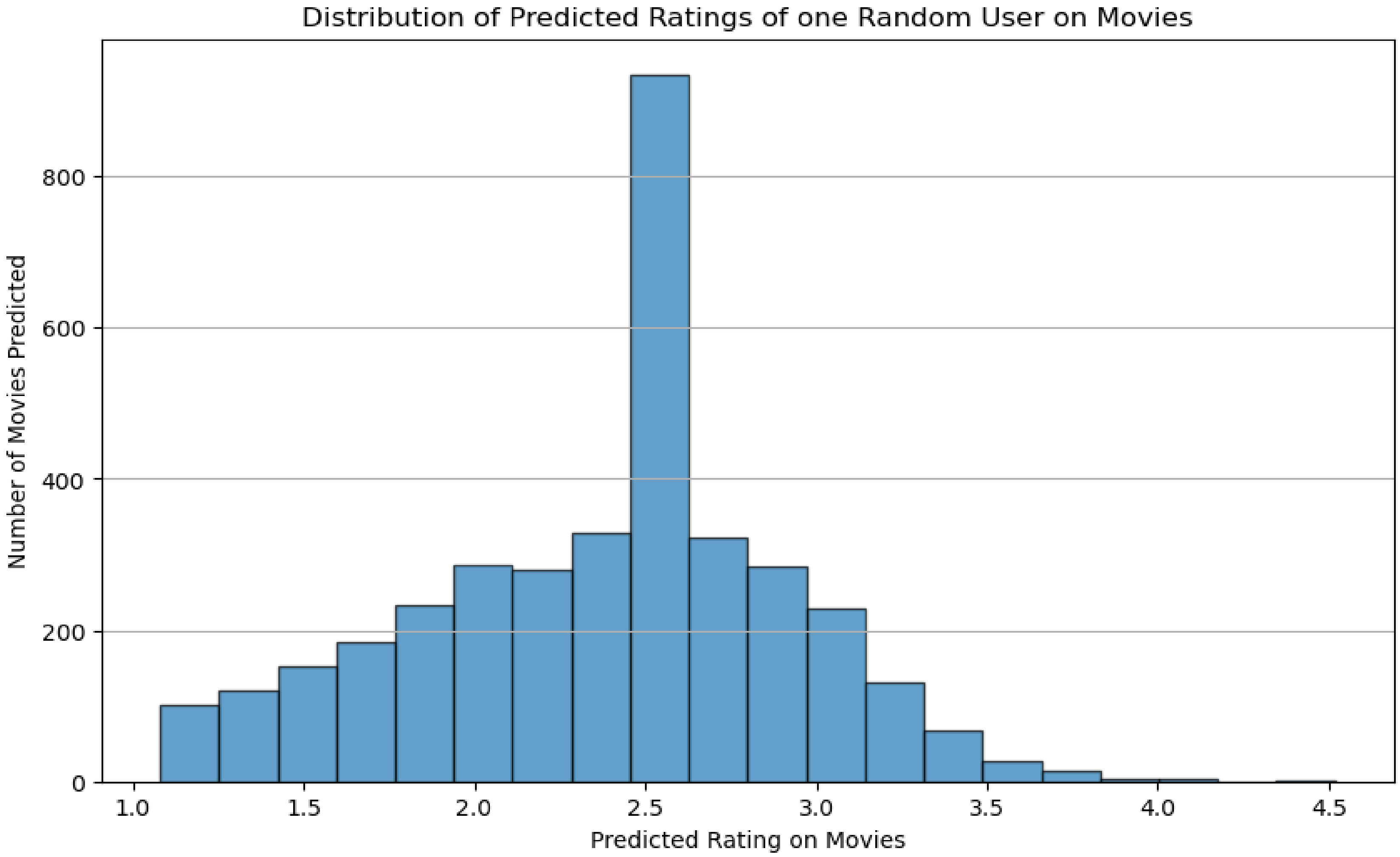


```
Top Movies Rated by the User with ID 1010
                                         Title  PredictedRating
3666                            Serpico (1973)         4.649372
1059   Sexual Life of the Belgians, The (1994)         4.491936
567                   Wedding Gift, The (1994)         4.447729
3268  I'll Never Forget What's 'is Name (1967)         4.127494
2860                               Reds (1981)         3.998241
...                                        ...              ...
2904            Crimes and Misdemeanors (1989)         1.022935
143                            Bad Boys (1995)         1.022102
3597                 Retro Puppetmaster (1999)         1.020055
1464                      Anna Karenina (1997)         1.020039
3831                              Duets (2000)         1.019653
```

We used user-based collaborative filtering to suggest movies in Task 1. The top movie suggestions for an active **random user** with ***ID 1010*** were identified, a user-movie matrix was created, user similarity was assessed using Pearson's coefficient, movie ratings were projected using various values of K (of KNN) with respect to the least RMSE.

Based on the movie ratings of comparable users, the method relied on user similarity to give an active user with a list of recommended movies.

# Histogram of Predicted Ratings of Random User

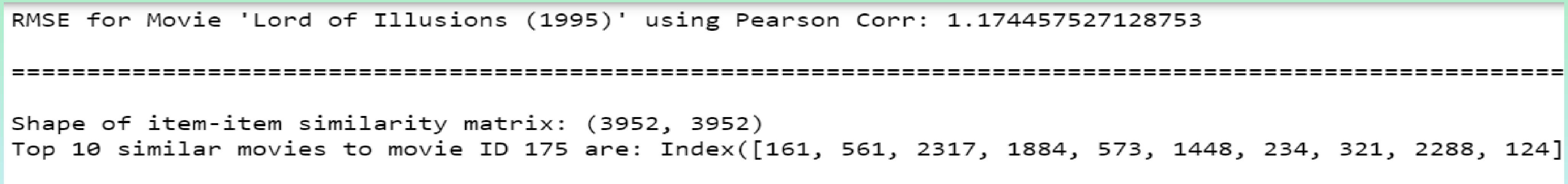


The histogram showcases the predicted movie ratings by a random user with ID 1010. Most movies receive a rating close to 2.5, with this peak signifying over 900 films.

Ratings below 1.25 or above 3.5 are rarer, suggesting the user generally rates films as ***average*** to slightly positive.

# Task 2: Item-based Collaborative Filtering

## - RMSE for evaluation and Top 10 similar movies

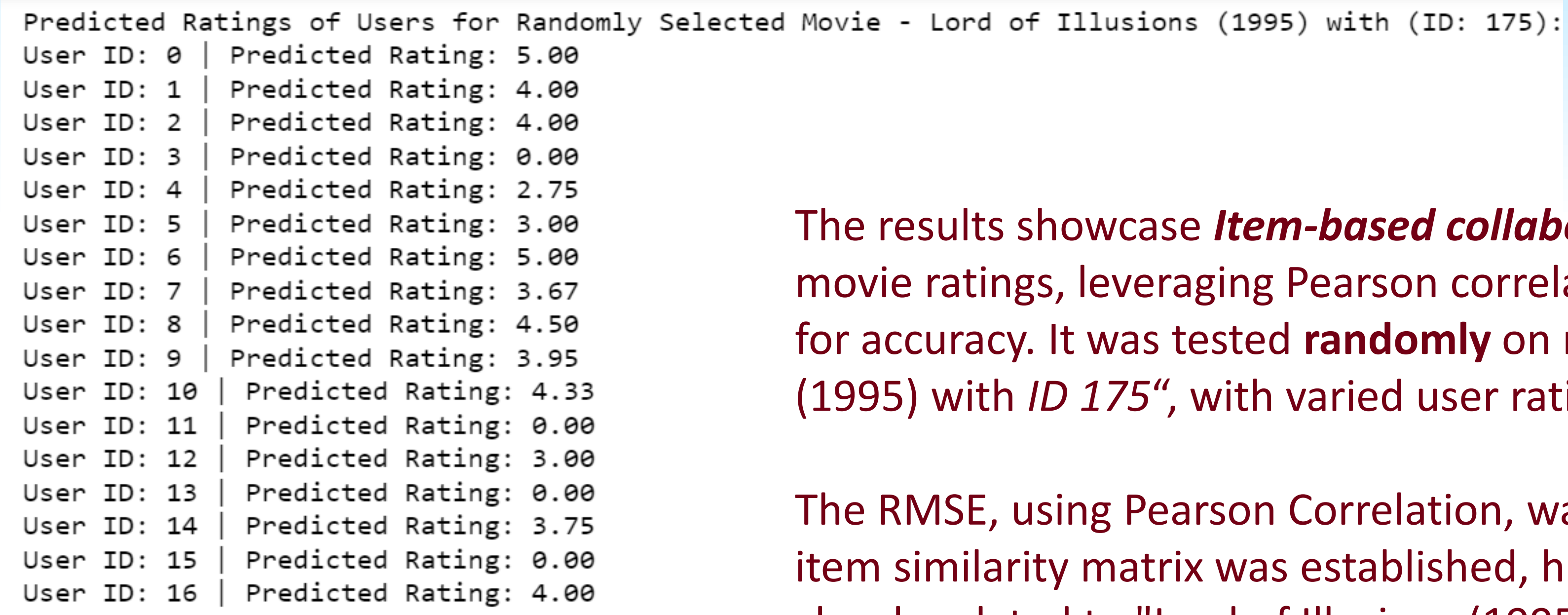


```
RMSE for Movie 'Lord of Illusions (1995)' using Pearson Corr: 1.174457527128753

=====================================================================================

Shape of item-item similarity matrix: (3952, 3952)
Top 10 similar movies to movie ID 175 are: Index([161, 561, 2317, 1884, 573, 1448, 234, 321, 2288, 124]
```

## - Top User's Predicted Ratings on Random Movie ID 175:

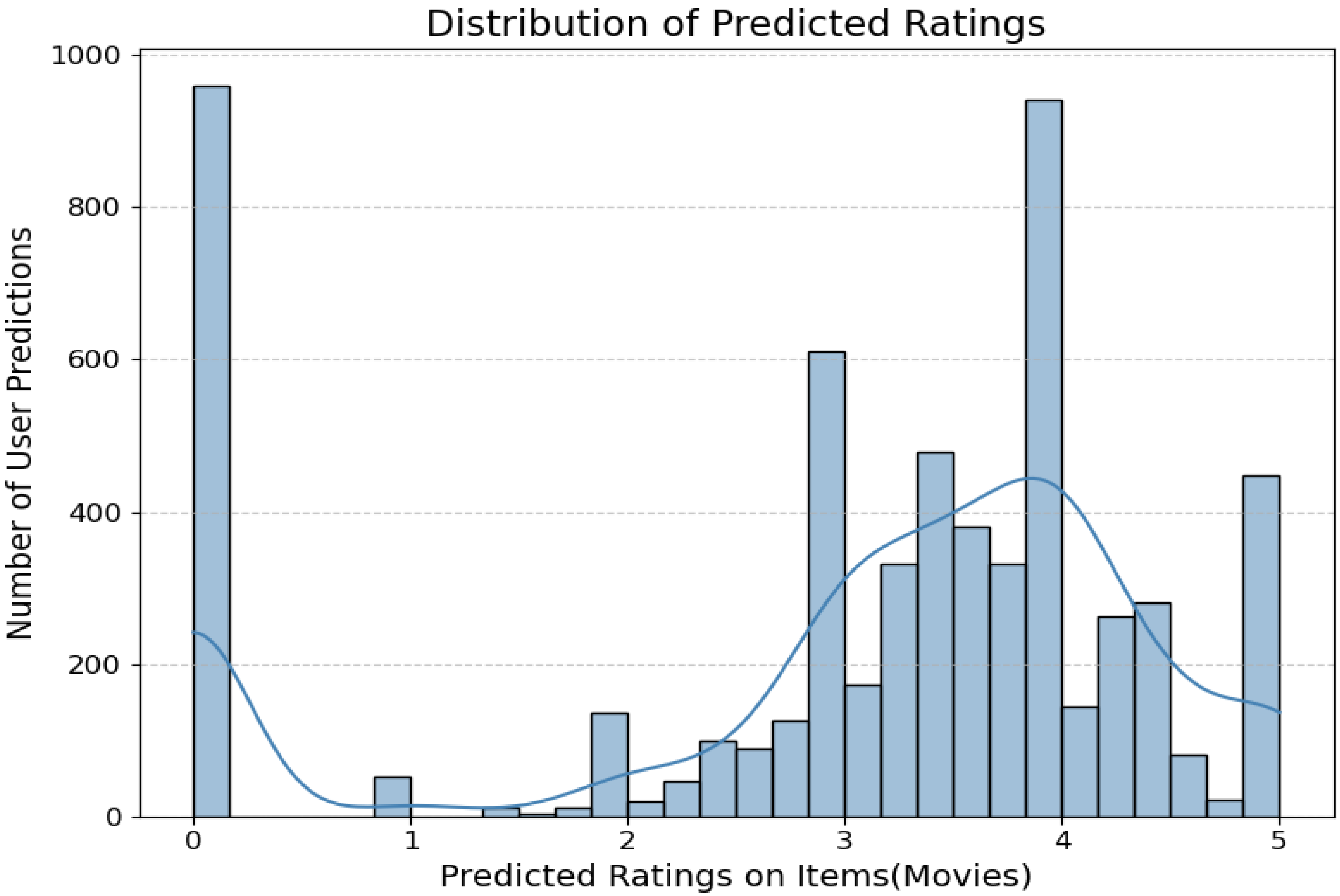


```
Predicted Ratings of Users for Randomly Selected Movie - Lord of Illusions (1995) with (ID: 175):
User ID: 0 | Predicted Rating: 5.00
User ID: 1 | Predicted Rating: 4.00
User ID: 2 | Predicted Rating: 4.00
User ID: 3 | Predicted Rating: 0.00
User ID: 4 | Predicted Rating: 2.75
User ID: 5 | Predicted Rating: 3.00
User ID: 6 | Predicted Rating: 5.00
User ID: 7 | Predicted Rating: 3.67
User ID: 8 | Predicted Rating: 4.50
User ID: 9 | Predicted Rating: 3.95
User ID: 10 | Predicted Rating: 4.33
User ID: 11 | Predicted Rating: 0.00
User ID: 12 | Predicted Rating: 3.00
User ID: 13 | Predicted Rating: 0.00
User ID: 14 | Predicted Rating: 3.75
User ID: 15 | Predicted Rating: 0.00
User ID: 16 | Predicted Rating: 4.00
```

The results showcase ***Item-based collaborative filtering*** to predict movie ratings, leveraging Pearson correlation and cosine similarity for accuracy. It was tested **randomly** on movie - "Lord of Illusions (1995) with *ID 175“*, with varied user ratings.

The RMSE, using Pearson Correlation, was about 1.1745. An item-item similarity matrix was established, highlighting ten movies closely related to "Lord of Illusions (1995)" by their IDs, such as 161, 561, and 2317 etc.

# Histogram of Predicted Ratings of Random Movie



The histogram showcases the Distribution of Predicted Ratings of Users on a **random movie "***Lord of Illusions (1995) with ID 175***“.** Most of the Users have predicted 0 ratings followed by users rating around 4, with this peak signifying over 900 Users.

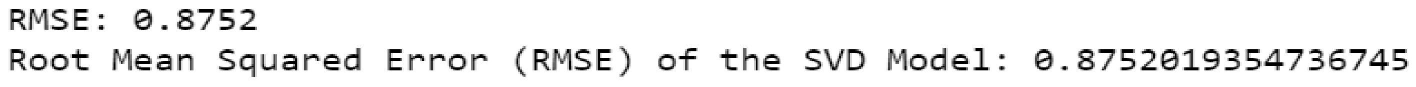
Ratings in between 1 to 2.75 are rarer, suggesting the users predicted ratings on this movie is **above average** (descent) overall.

RMIT UNIVERSITY

# Task 3: A Better Recommender System

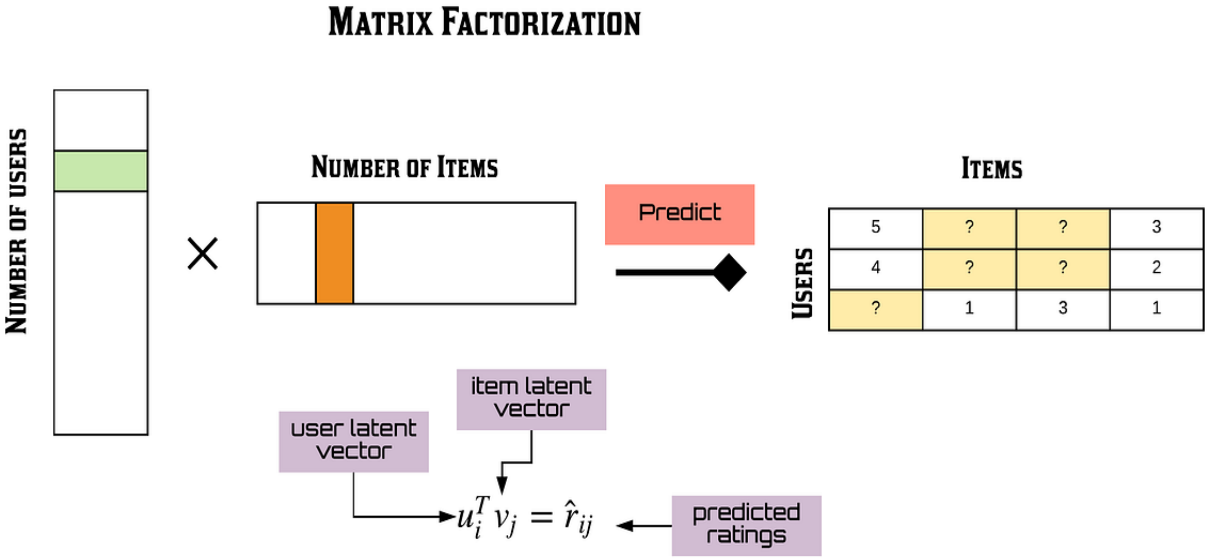
## Task 3.1 “Option1RecSys” :-

## *Matrix Factorization Techniques for Recommender Systems*



```
RMSE: 0.8752
Root Mean Squared Error (RMSE) of the SVD Model: 0.8752019354736745
```

- ***Best RMSE score output***



| | | | |
|---|---|---|---|
| 5 | ? | ? | 3 |
| 4 | ? | ? | 2 |
| ? | 1 | 3 | 1 |

$u_i^T v_j = \hat{r}_{ij}$

https://towardsdatascience.com/recsys-series-part-4-the-7-variants-of-matrix-factorization-for-collaborative-filtering-368754e4fab5

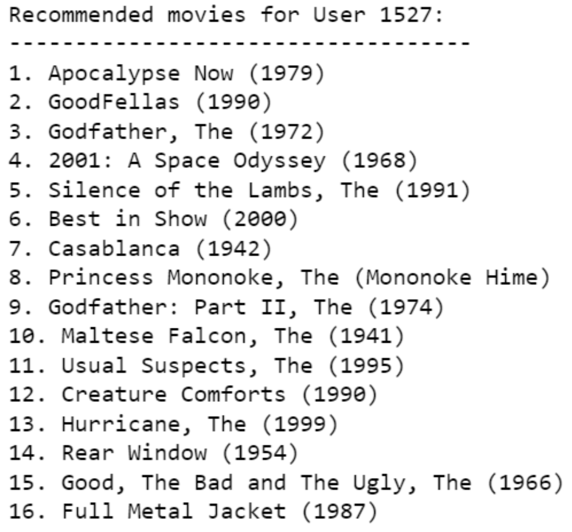
The movie ratings dataset is split using an 80-20 split ratio (like previously done), creating training and testing sets. After that, it makes use of the ***Surprise*** library to provide datasets appropriate for collaborative  filtering. Using the training set, a ***Singular Value Decomposition*** (**SVD**) model is initialised and trained.

The test set is used for predictions, and the evaluation measure Root Mean Squared Error (RMSE) is computed. According to the result, the SVD model's prediction error was quantified by its RMSE, which was roughly ***0.8752, the best RMSE value*** indicating greater performance in predicting user ratings for movies.

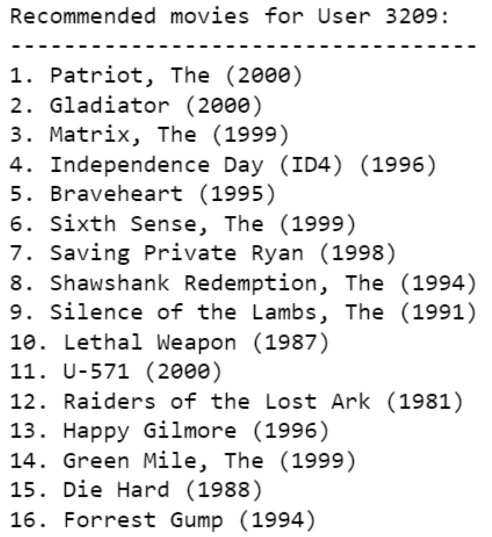
# Task 3.2 – Recommending Top-30 movies for 5 Random Users

The provided output defines a recommendation function called `get_recommendation_option1` that utilizes a trained SVD model to generate **top 30** movies recommendations for **5 random user** who have rated more than 100 movies. It predicts ratings for all movies in the dataset for the user, sorts them by predicted ratings, and returns the titles of the top 30 recommended movies for each user.

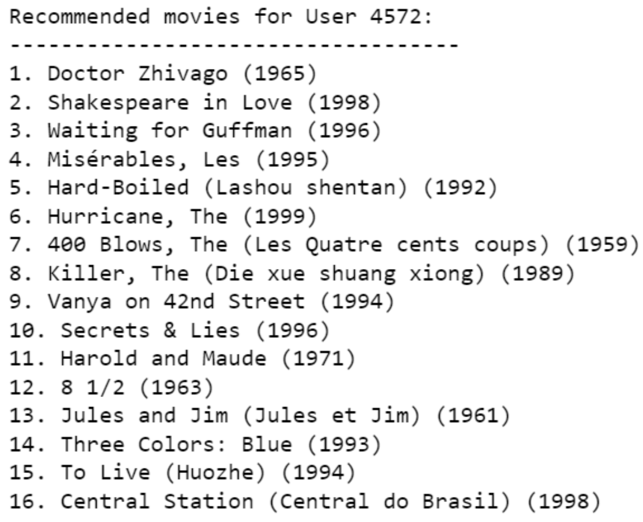
This approach offers a personalized movie recommendation system for these users based on their historical ratings and the SVD model's predictions.



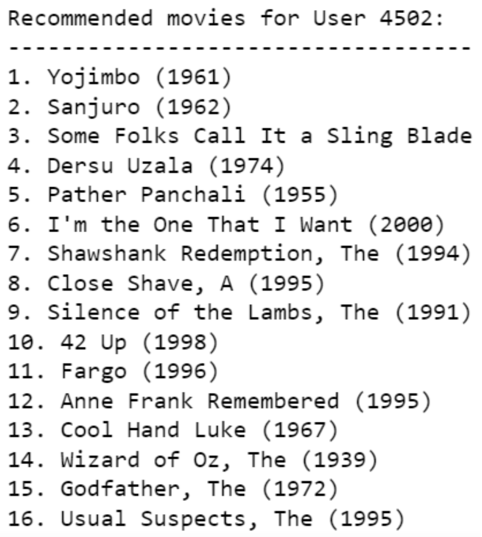
```
Recommended movies for User 1527:
-----------------------------------
1. Apocalypse Now (1979)
2. GoodFellas (1990)
3. Godfather, The (1972)
4. 2001: A Space Odyssey (1968)
5. Silence of the Lambs, The (1991)
6. Best in Show (2000)
7. Casablanca (1942)
8. Princess Mononoke, The (Mononoke Hime)
9. Godfather: Part II, The (1974)
10. Maltese Falcon, The (1941)
11. Usual Suspects, The (1995)
12. Creature Comforts (1990)
13. Hurricane, The (1999)
14. Rear Window (1954)
15. Good, The Bad and The Ugly, The (1966)
16. Full Metal Jacket (1987)
```



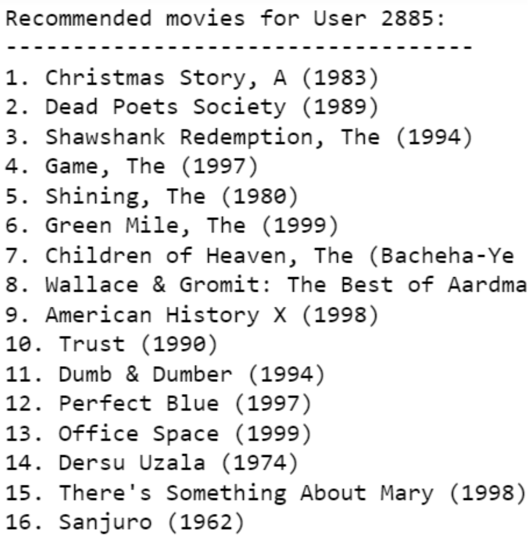
```
Recommended movies for User 3209:
-----------------------------------
1. Patriot, The (2000)
2. Gladiator (2000)
3. Matrix, The (1999)
4. Independence Day (ID4) (1996)
5. Braveheart (1995)
6. Sixth Sense, The (1999)
7. Saving Private Ryan (1998)
8. Shawshank Redemption, The (1994)
9. Silence of the Lambs, The (1991)
10. Lethal Weapon (1987)
11. U-571 (2000)
12. Raiders of the Lost Ark (1981)
13. Happy Gilmore (1996)
14. Green Mile, The (1999)
15. Die Hard (1988)
16. Forrest Gump (1994)
```



```
Recommended movies for User 4572:
-----------------------------------
1. Doctor Zhivago (1965)
2. Shakespeare in Love (1998)
3. Waiting for Guffman (1996)
4. Misérables, Les (1995)
5. Hard-Boiled (Lashou shentan) (1992)
6. Hurricane, The (1999)
7. 400 Blows, The (Les Quatre cents coups) (1959)
8. Killer, The (Die xue shuang xiong) (1989)
9. Vanya on 42nd Street (1994)
10. Secrets & Lies (1996)
11. Harold and Maude (1971)
12. 8 1/2 (1963)
13. Jules and Jim (Jules et Jim) (1961)
14. Three Colors: Blue (1993)
15. To Live (Huozhe) (1994)
16. Central Station (Central do Brasil) (1998)
```



```
Recommended movies for User 4502:
-----------------------------------
1. Yojimbo (1961)
2. Sanjuro (1962)
3. Some Folks Call It a Sling Blade
4. Dersu Uzala (1974)
5. Pather Panchali (1955)
6. I'm the One That I Want (2000)
7. Shawshank Redemption, The (1994)
8. Close Shave, A (1995)
9. Silence of the Lambs, The (1991)
10. 42 Up (1998)
11. Fargo (1996)
12. Anne Frank Remembered (1995)
13. Cool Hand Luke (1967)
14. Wizard of Oz, The (1939)
15. Godfather, The (1972)
16. Usual Suspects, The (1995)
```



```
Recommended movies for User 2885:
-----------------------------------
1. Christmas Story, A (1983)
2. Dead Poets Society (1989)
3. Shawshank Redemption, The (1994)
4. Game, The (1997)
5. Shining, The (1980)
6. Green Mile, The (1999)
7. Children of Heaven, The (Bacheha-Ye
8. Wallace & Gromit: The Best of Aardma
9. American History X (1998)
10. Trust (1990)
11. Dumb & Dumber (1994)
12. Perfect Blue (1997)
13. Office Space (1999)
14. Dersu Uzala (1974)
15. There's Something About Mary (1998)
16. Sanjuro (1962)
```

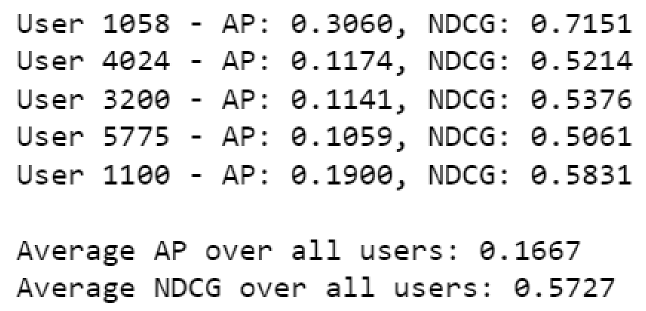
# ***AP (Average Precision) and NDCG (Normalized Discounted Cumulative Gain) as evaluation metrics***

The provided output visualizes the evaluation metrics, including Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG), to assess the performance of a recommender system for a set of ***5 random users***.

Also, the average AP and NDCG scores are ***0.166*** and ***0.572***, providing an overall assessment of the recommender system's quality. The ***Bar plot*** visualises the same output presented in the matrix.

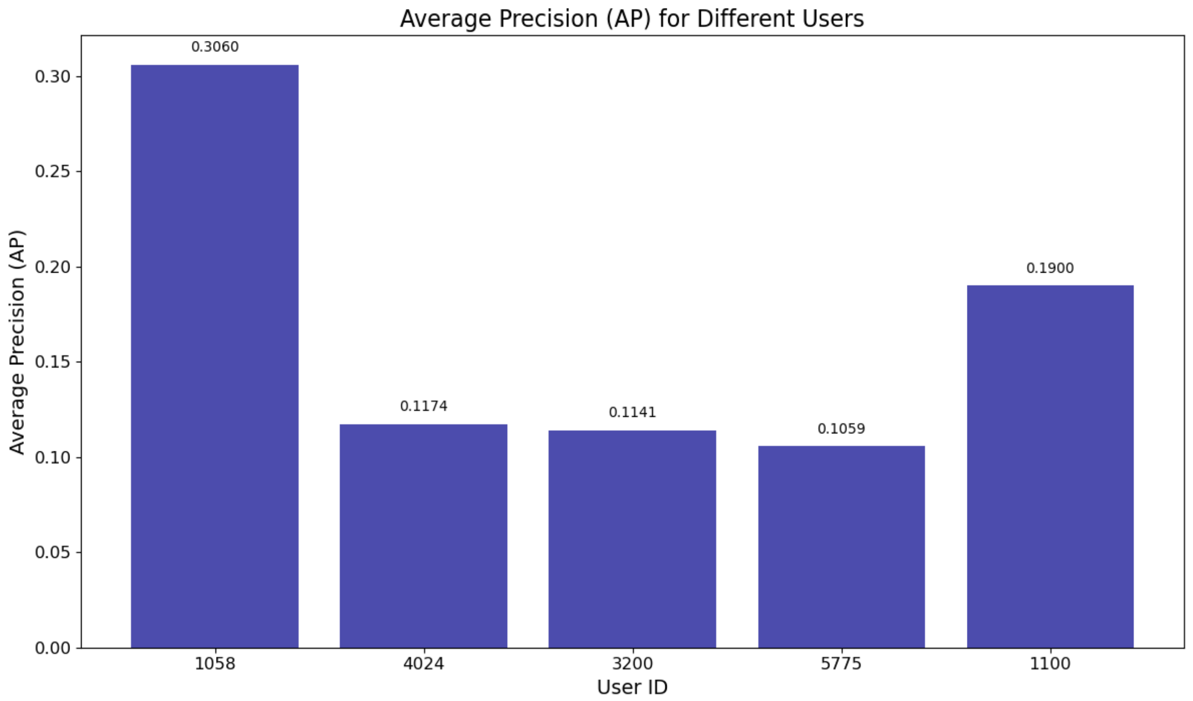
* **Note**: The result of AP and NDCG varies every time because the Users are being selected randomly.

### *AP and NDCG scores:*

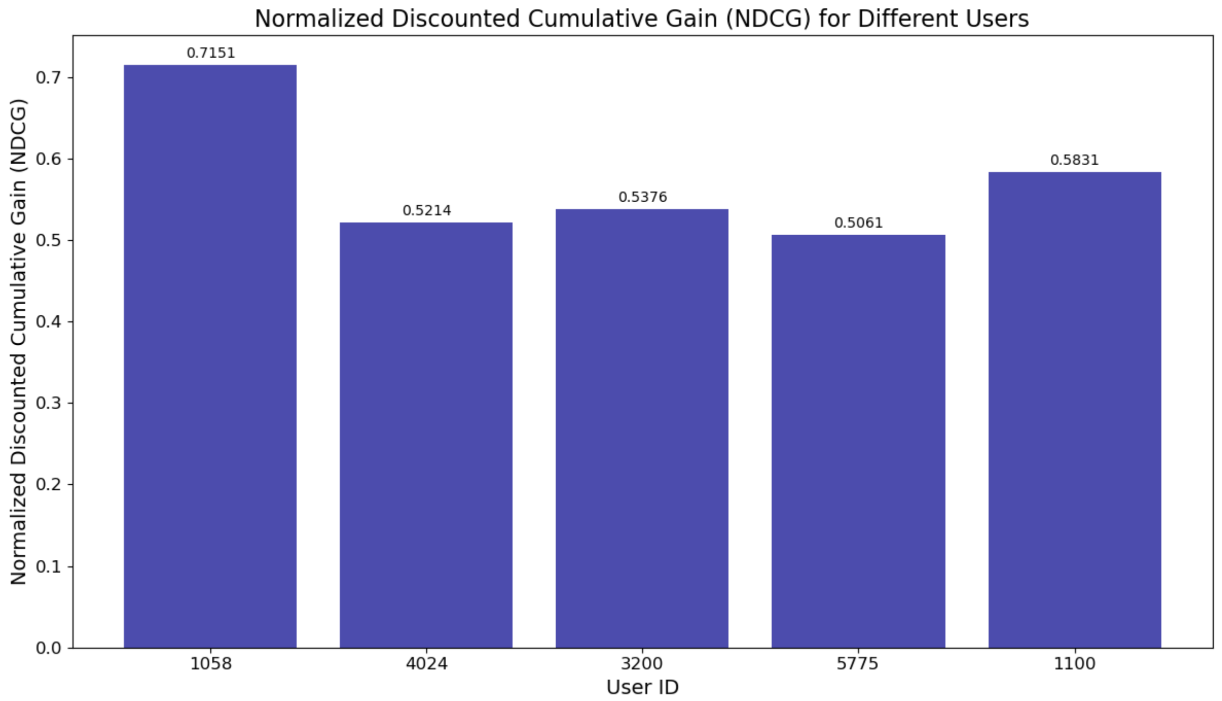


```
User 1058 - AP: 0.3060, NDCG: 0.7151
User 4024 - AP: 0.1174, NDCG: 0.5214
User 3200 - AP: 0.1141, NDCG: 0.5376
User 5775 - AP: 0.1059, NDCG: 0.5061
User 1100 - AP: 0.1900, NDCG: 0.5831

Average AP over all users: 0.1667
Average NDCG over all users: 0.5727
```



Average Precision (AP) for Different Users



Normalized Discounted Cumulative Gain (NDCG) for Different Users

# References

- RMIT University, Practical Data Science with Python , week9-RecommenderSystems(1), lec-week9-RecommenderSystems(1)-updated.pdf: Practical Data Science with Python (2350) (instructure.com)

- RMIT University, Practical Data Science with Python , week10-RecommenderSystems(2), lec-week10-RecommenderSystems(2)-updated.pdf: Practical Data Science with Python (2350) (instructure.com)

- GeeksforGeeks, “Plotting Histogram in Python using Matplotlib”, [Internet] https://www.geeksforgeeks.org/plotting-histogram-in-python-using-matplotlib/

- GitHub, “item-item similarity matrix”, [Internet] https://github.com/topics/item-item-similarity

- Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," in Computer, vol. 42, no. 8, pp. 30-37, Aug. 2009, doi: 10.1109/MC.2009.263. [Internet] https://ieeexplore.ieee.org/document/5197422 {Accessed through RMIT Library}

- Machine Learning Mastery, “Using Singular Value Decomposition to Build a Recommender System”,[Internet] https://machinelearningmastery.com/using-singular-value-decomposition-to-build-a-recommender-system/

- Towards Data Science, “Evaluation Metrics for Recommendation Systems”, [Internet] https://towardsdatascience.com/evaluation-metrics-for-recommendation-systems-an-overview-71290690ecba