**CSE523 - Machine Learning**

**Movie Recommendation System using Machine Learning**

**Faculty - Prof. Mehul Raval**

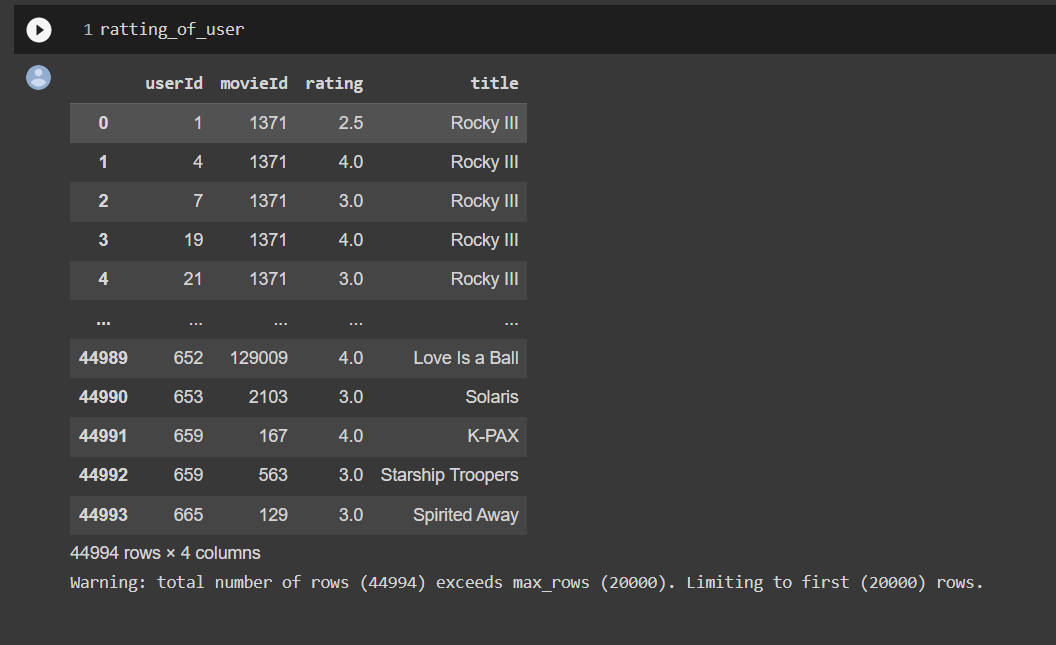
**Weekly Report 4**

**Submission Date: 04-03-23**

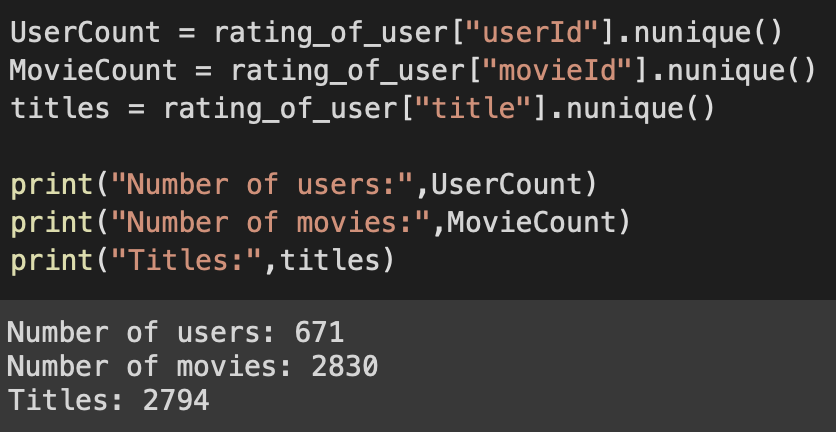
**Group: Tech Titans**

| **Name** | **Roll Number** | **Email** |
| --- | --- | --- |
| Jainam Shah | AU2040186 | jainam.s4@ahduni.edu.in |
| Yash Chotaliya | AU2040193 | yashkumar.c@ahduni.edu.in |
| Akshay Parmar | AU2040199 | akshay.p@ahduni.edu.in |
| Shubham Bhatt | AU2040206 | shubham.b1@ahduni.edu.in |

As in the last week we had analyzed the data and wrote the code for finding out the rating that the users had given to movies and had got the following output.



Further ahead, we had written code and found the number of unique users who rated the movies, the number of unique movies that the users have rated out, and the number of unique movie titles that the users have rated; the following photo shows the output that we had got.



Then we created a histogram that shows the ratings for the number of movies. So, we created a pivot table that will aggregate the ratings by user and movie. We created this histogram as the histogram created by the code can help understand the distribution of movie ratings in the dataset. The histogram will show the frequency of different rating values given by users, with the x-axis representing the different rating values and the y-axis representing the number of movies with that rating.

Also, by examining the histogram, we can see if the ratings are evenly distributed or if there are any biases or patterns. For example, if the histogram shows that most movies are rated highly, it could suggest that users are generally positive towards the movies in the dataset. On the other hand, if the histogram shows that most movies are rated poorly, then it could suggest that the dataset includes a lot of unpopular or poorly-rated movies.

This histogram can also determine the range and spread of rating values. For instance, if most ratings are concentrated around a narrow range of values, it suggests that users are more likely to rate movies within that range.

Overall, this histogram provides insights into the distribution of ratings in the dataset, which can help understand user behavior, identify biases, and make informed decisions for further analysis.

The following shows the histogram that we got.

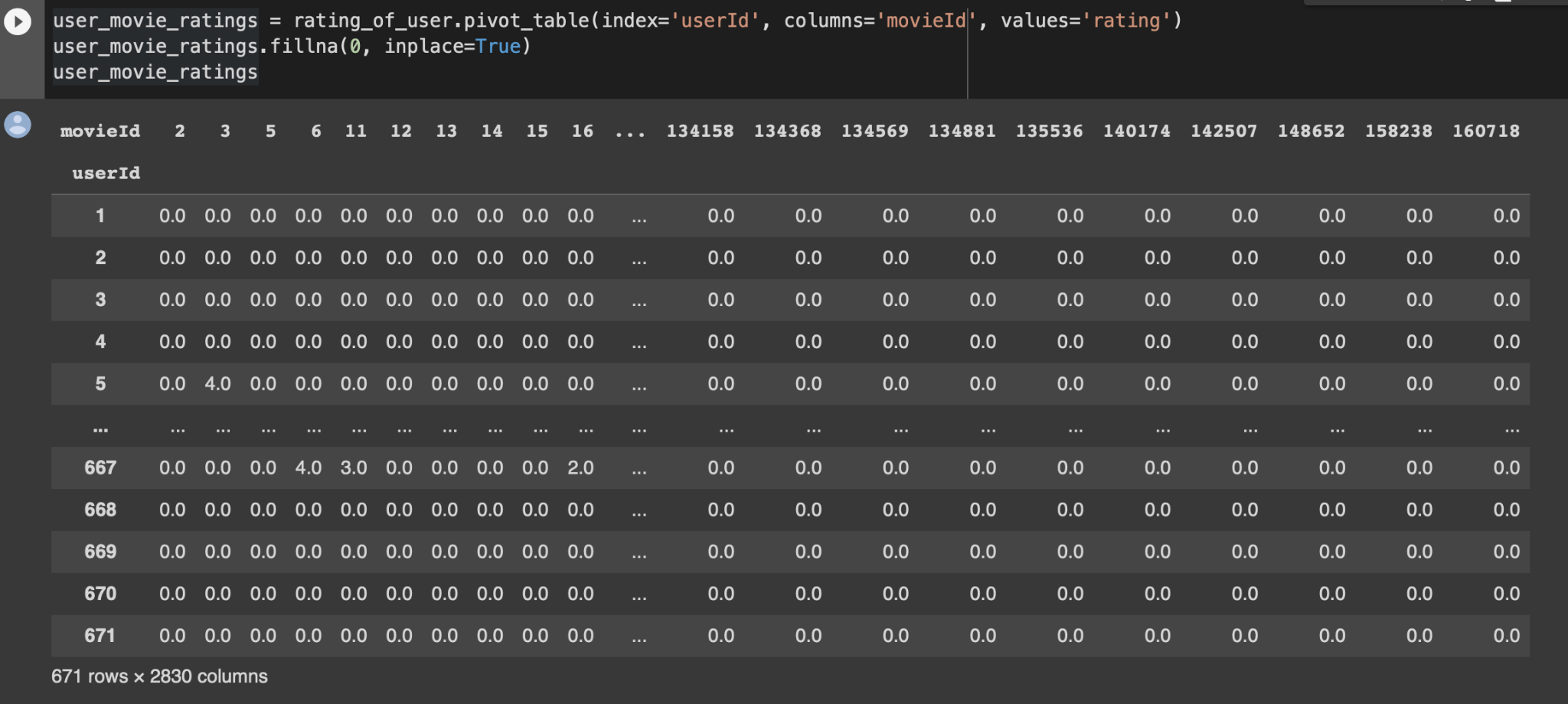


Furthermore, we normalized the data to show all user IDs and movie IDs and what rating to which movie by which user is also there. By normalizing the data this way, we created a matrix where a row represents movies, and each column represents a user, with the cell values representing the ratings each user gives each movie. This matrix is often referred to as a user-item matrix or a rating matrix. So by analyzing the user-item matrix, we can identify patterns and trends in user behavior, such as movies that are commonly rated highly by a particular group of users.

Also, Normalizing the data is helpful to us in Collaborative Filtering as Collaborative Filtering is a technique used in recommendation systems that involves analyzing the user-item matrix to identify users with similar tastes or preferences. Normalization can help ensure the similarity metric is calculated correctly and the recommendations are accurate. Normalizing the data can also help address the problem of different users having different rating scales, which can affect the accuracy of the recommendations.

Overall, normalizing the data to show all user IDs and movie IDs, along with the rating given to each movie by each user, is crucial for effective Collaborative Filtering. It allows us to identify users with similar tastes or preferences and recommend movies that they rated highly to other users with similar preferences.

The following shows us the output after the normalization.



After that, we found similar users based on the movie rating, which means we generally tried to find our correlation between them. We did this because Finding similar users based on their movie ratings is helpful in Collaborative Filtering because it allows us to use the preferences and behavior of similar users to make recommendations to a given user.

As in Collaborative Filtering, we aim to recommend items to a user based on the preferences of other users with similar tastes. By identifying users with similar preferences based on their movie ratings, we can use them to make personalized recommendations to the given user.

So first, we found out the number of similar users and then found the correlation between them. The following output shows the correlation between similar users.

