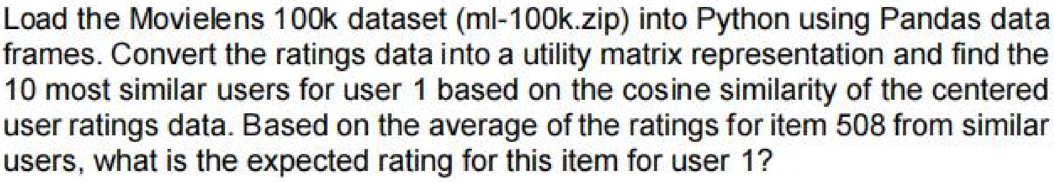
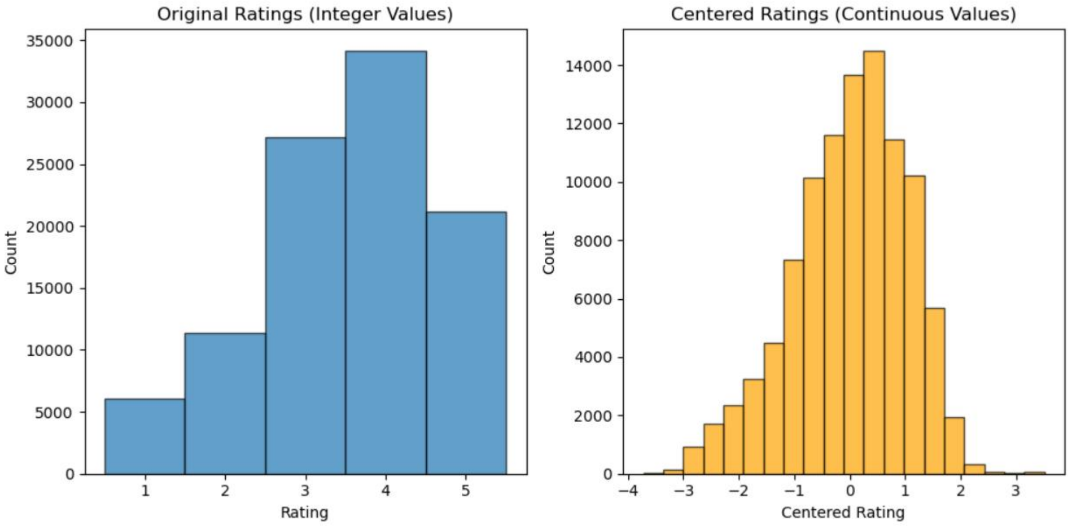
Data Mining HW5.0

Problem 1:

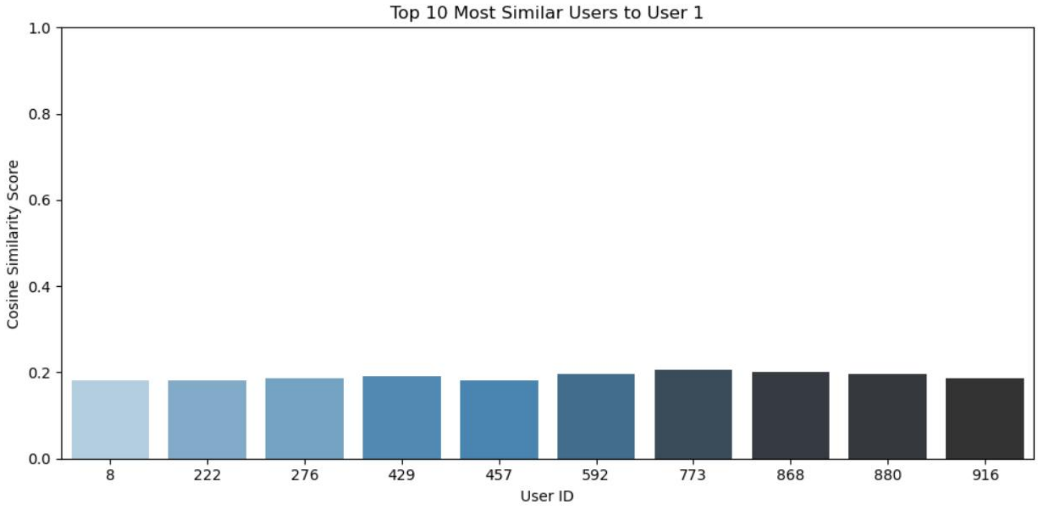


Load the Movielens 100k dataset (ml-100k.zip) into Python using Pandas data frames. Convert the ratings data into a utility matrix representation and find the 10 most similar users for user 1 based on the cosine similarity of the centered user ratings data. Based on the average of the ratings for item 508 from similar users, what is the expected rating for this item for user 1?

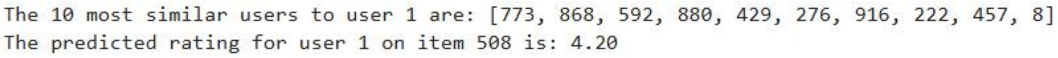
①Here is the difference between Original and Centered user data:



②We can get similarities for user 1 and get top 10 similar users:

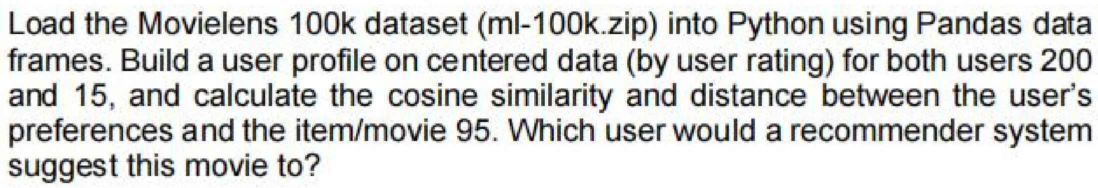


We can get the result:



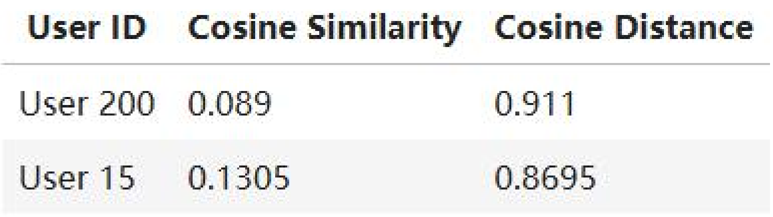
```
The 10 most similar users to user 1 are: [773, 868, 592, 880, 429, 276, 916, 222, 457, 8]
The predicted rating for user 1 on item 508 is: 4.20
```

Problem 2:



Load the Movielens 100k dataset (ml-100k.zip) into Python using Pandas data frames. Build a user profile on centered data (by user rating) for both users 200 and 15, and calculate the cosine similarity and distance between the user's preferences and the item/movie 95. Which user would a recommender system suggest this movie to?

Cosine Similarity and Distance:



User ID	Cosine Similarity	Cosine Distance
User 200	0.089	0.911
User 15	0.1305	0.8695

Decision Basis for the Recommender System:

A higher cosine similarity (or lower cosine distance) indicates a stronger alignment between a user's preferences and the rating pattern of Movie 95. Comparing the two users:

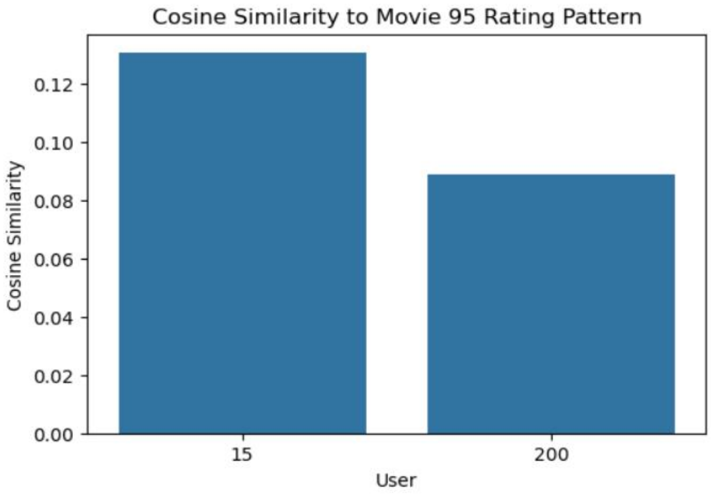
① User 15 has a higher cosine similarity (0.1305 > 0.089) and a lower cosine distance (0.8695 < 0.911).

② User 200 has a lower similarity and larger distance, meaning their historical rating pattern differs more from Movie 95.

Final Conclusion:

The recommender system would suggest Movie 95 to User 15, as this user's rating pattern is more similar to that of the movie.

Data Visualization:


Cosine Similarity to Movie 95 Rating Pattern
Cosine Similarity
User