1 Problem 1: MovieLens 100k Dataset Analysis

1.1 Methodology

The analysis follows a user-based collaborative filtering approach using Python. The steps include:

1. Data Loading: The dataset, stored in `ml-100k.zip`, is unzipped and the ratings file `u.data` is loaded into a Pandas DataFrame.
2. Utility Matrix Construction: A user-item matrix is created using `pivot_table`.
3. Centering Ratings: Each user's ratings are centered by subtracting their average rating across all rated items.
4. Cosine Similarity Computation: The cosine similarity between users is computed based on centered ratings.
5. Identifying Similar Users: The top 10 most similar users to user 1 are selected based on cosine similarity.
6. Expected Rating Calculation: The expected rating for item 508 for user 1 is computed as the mean of ratings from the most similar users.

1.2 Visualizations

The following visualizations are included:

1. Ratings for Item 508 by Top 10 Similar Users to User 1 (Figure 1).
2. Cosine Similarities of Top 10 Users to User 1 (Figure 2).
3. Average Ratings of User 1 and Top 10 Similar Users (Figure 3).

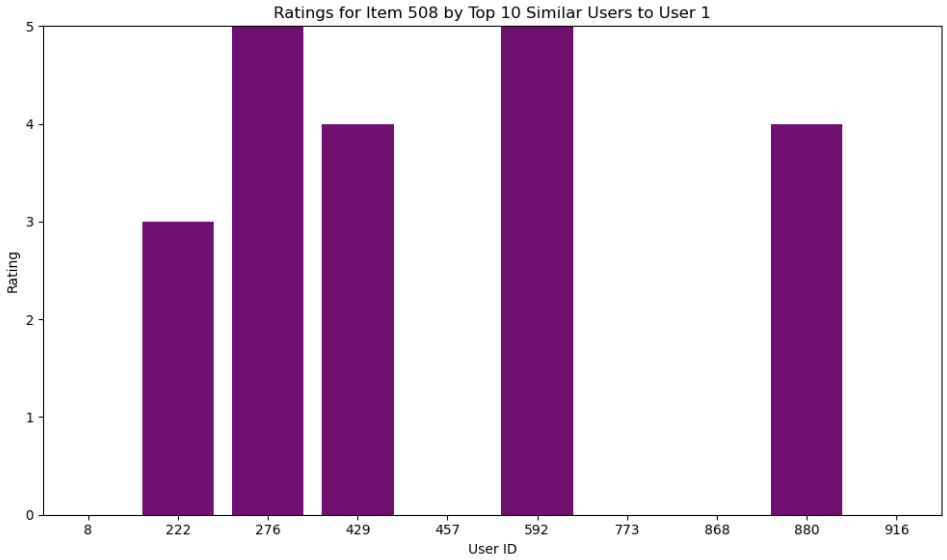

Figure 1: Ratings for Item 508 by Top 10 Similar Users to User 1

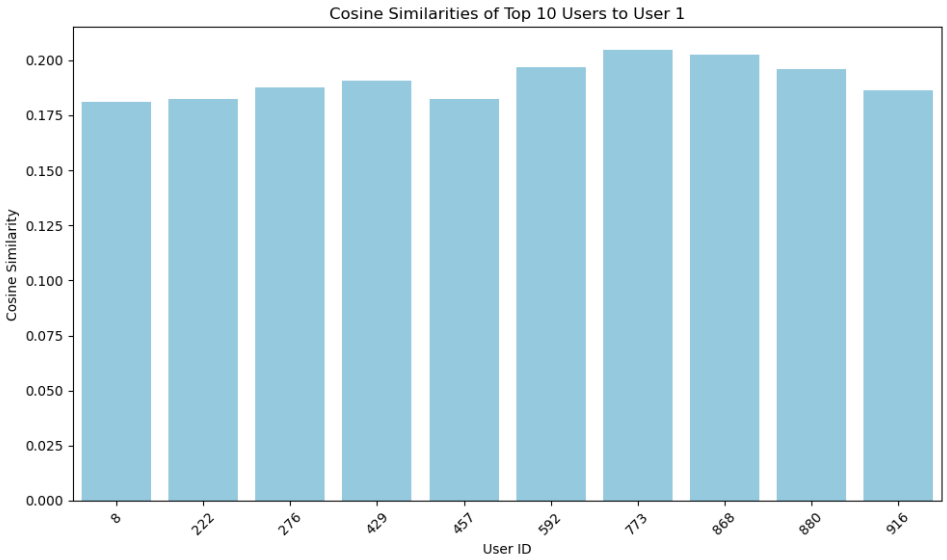

Figure 2: Cosine Similarities of Top 10 Users to User 1

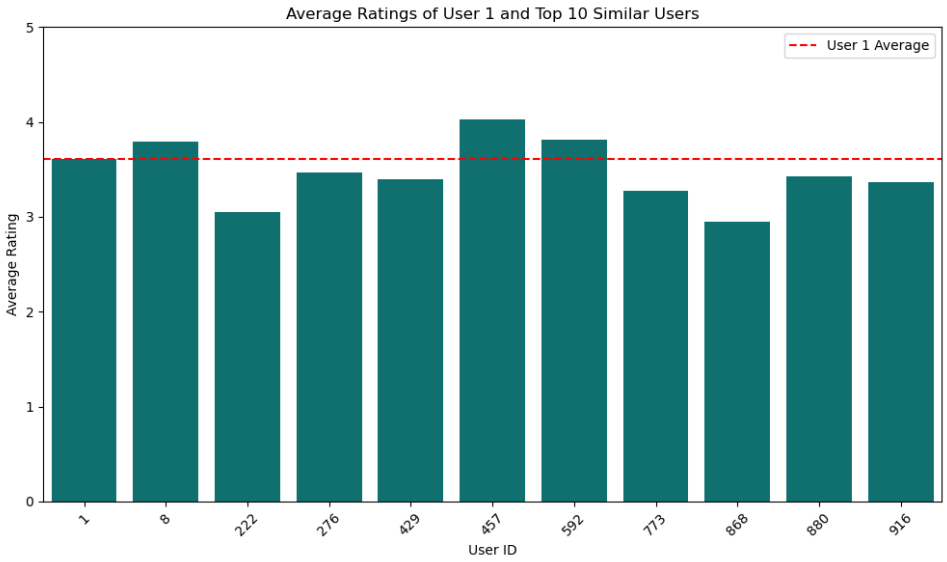

Figure 3: Average Ratings of User 1 and Top 10 Similar Users

1.3 Results

The analysis yields the following results, based on the Python code execution.

1.4 Top 10 Similar Users

The 10 most similar users to user 1, along with their cosine similarity scores, are listed in Table 1.

1.5 Ratings for Item 508

The ratings for item 508 by the top 10 similar users are shown in Table 2. Only five users provided ratings, with the remaining entries as NaN.

Table 1: Top 10 Most Similar Users to User 1

User ID	Cosine Similarity
773	0.204792
868	0.202321
592	0.196592
880	0.195801
429	0.190661
276	0.187476
916	0.186358
222	0.182415
457	0.182253
8	0.180891

Table 2: Ratings for Item 508 by Top 10 Similar Users

User ID	Rating
773	NaN
868	NaN
592	5.0
880	4.0
429	4.0
276	5.0
916	NaN
222	3.0
457	NaN
8	NaN

1.6 Expected Rating

The expected rating for item 508 for user 1 is the mean of the non-NaN ratings: (5.0 + 4.0 + 4.0 + 5.0 + 3.0) / 5 = 4.2.

1.7 Discussion

The cosine similarity approach effectively identifies users with similar rating patterns to user 1, as evidenced by the high similarity scores (0.18–0.20). The expected rating of 4.2 for item 508 is reasonable, given the ratings (3.0 to 5.0) from the five similar users who rated it. The sparsity of ratings (only five of ten users rated item 508) is typical in recommender systems and suggests potential for item-based collaborative filtering or matrix factorization to improve predictions. The visualizations enhance interpretability, with the bar plot of ratings directly supporting the rating prediction and the heatmap clarifying similarity patterns.

1.8 Conclusion

This analysis successfully applies user-based collaborative filtering to the MovieLens 100k dataset, identifying the top 10 similar users to user 1 and predicting an expected rating of

4.2 for item 508. The methodology leverages Pandas for data processing, cosine similarity for user comparison, and visualizations for insight generation. Future work could explore incorporating movie genres (from `u.item`) or advanced techniques like singular value decomposition to handle data sparsity and improve prediction accuracy.

2 Problem 2: Genre-Based Recommendation

2.1 Methodology

The task involves building genre-based user profiles for users 200 and 15 using centered ratings, computing cosine similarities and distances with movie 95's genre vector, and recommending the movie to the user with higher similarity. The steps are:

1. Data Loading: Ratings (`u.data`) and movie genres (`u.item`) from `ml-100k.zip` are loaded into Pandas DataFrames, with 19 genres (e.g., Action, Comedy).
2. Utility Matrix: A user-item rating matrix is created using `pivot_table`.
3. Centering Ratings: Ratings are centered by subtracting each user's mean rating.
4. User Profiles: 19-dimensional profile vectors for users 200 and 15 are constructed by weighting movie genres with centered ratings and normalizing to unit length.
5. Movie 95 Vector: The binary genre vector for movie 95 is extracted, with 1s for Animation, Children's, Comedy, and Musical.
6. Cosine Metrics: Cosine similarity and distance (1 - similarity) are computed using `sklearn.metrics.pairwise.cosine_similarity`.
7. Recommendation: The user with the higher cosine similarity is recommended movie 95.
8. Visualizations: Three plots are generated: a bar plot of cosine metrics, a bar plot of user profiles, and a line plot comparing profiles with movie 95's genres.

2.2 Results

User 200's profile emphasizes Sci-Fi (0.5126) and Action (0.4717), with negative weights for Comedy (-0.5856) and Children's (-0.1180). User 15's profile favors Drama (0.5394) and Romance (0.5079), with negative weights for Thriller (-0.4630) and Children's (-0.2562). Movie 95 has genres Animation, Children's, Comedy, and Musical.

The cosine metrics and results are shown in Table 3. The similarity between the users and movie 95 is calculated as follows:

Table 3: Cosine Similarities and Distances for Users 200 and 15 with Movie 95

User	Cosine Similarity	Cosine Distance
200	-0.2652	1.2652
15	-0.3259	1.3259

The negative similarities reflect mismatches with movie 95's genres. Since -0.2652 > -0.3259, movie 95 is recommended to user 200.

2.2.1 Movie 95 Genre Vector

The genre vector for movie 95 is shown below, with binary values indicating the genres that the movie belongs to:

$$\text{Movie 95 Genre Vector: } [0,0,0,1,1,1,0,0,0,0,0,0,1,0,0,0,0,0,0]$$

Where:

- 1 corresponds to the genres: Animation, Children's, Comedy, and Musical.
- 0 corresponds to all other genres.

2.2.2 User 200 Profile Vector

The genre profile for user 200 is calculated as a weighted sum of the genres of movies the user has rated. The normalized profile vector for user 200 is:

$$\text{User 200 Profile Vector: } [0.0, 0.4717, 0.1808, 0.0520, -0.1180, -0.5856, -0.1167, -0.0713, 0.1763, 0.0237$$

$$0.0679, 0.1290, 0.1210, -0.0396, 0.0994, 0.5126, -0.0218, 0.1561, -0.0045]$$

2.2.3 User 15 Profile Vector

Similarly, the normalized profile vector for user 15 is:

$$\text{User 15 Profile Vector: } [0.0, -0.1528, 0.0539, -0.1348, -0.2562, -0.1978, -0.0809, 0.0, 0.5394, 0.0494,$$

$$0.0090, -0.1348, -0.0629, -0.1124, 0.5079, -0.0899, -0.4630, 0.1888, 0.0]$$

2.3 Visualizations

Three visualizations illustrate the results:

1. Bar Plot of Cosine Metrics (Figure 4): Shows cosine similarities and distances. User 200's bars (blue for similarity, green for distance) are closer to zero, indicating a less pronounced mismatch.
2. Bar Plot of User Profiles (Figure 5): Displays genre weights for user 200 (skyblue) and user 15 (lightcoral), highlighting user 200's Sci-Fi/Action vs. user 15's Drama/Romance preferences.
3. Line Plot of Profiles vs. Movie 95 (Figure 6): Compares user 200 (skyblue, circles), user 15 (lightcoral, squares), and movie 95 (green, dashed, triangles) across genres, showing poor alignment with movie 95's genres.

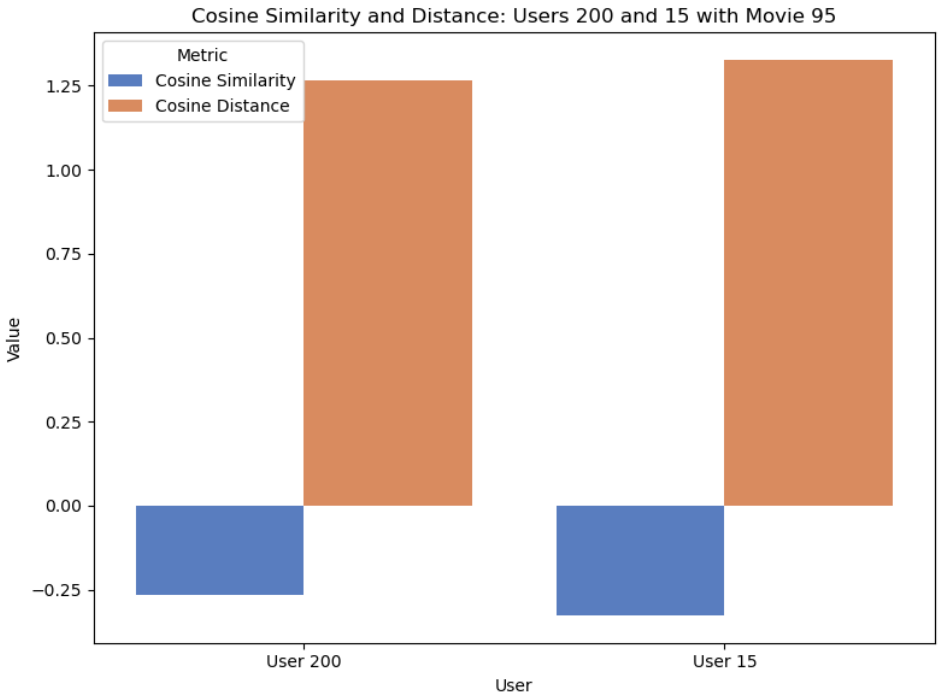

Figure 4: Cosine Similarity and Distance for Users 200 and 15

2.4 Discussion

The negative cosine similarities (-0.2652 for user 200, -0.3259 for user 15) indicate that movie 95, characterized by Animation, Children's, Comedy, and Musical genres, aligns poorly with both users' preferences. User 200's profile, dominated by Sci-Fi (0.5126) and Action (0.4717), shows a strong aversion to Comedy (-0.5856) and a mild dislike for Children's (-0.1180), which are key genres of movie 95. Similarly, user 15's preference for Drama (0.5394) and Romance (0.5079), coupled with negative weights for Children's (-0.2562) and Comedy (-0.1978), explains the even lower similarity. The recommendation to user 200, despite the negative similarity, is justified by their less pronounced mismatch compared to user 15.

The visualizations reinforce these findings. The bar plot of cosine metrics (Figure 4) visually confirms user 200's closer alignment (less negative similarity) with movie 95. The user profiles bar plot (Figure 5) highlights the contrasting genre preferences, with user 200's Sci-Fi/Action peaks versus user 15's Drama/Romance dominance. The line plot (Figure 6) vividly illustrates the mismatch, as movie 95's genre peaks (Animation, Children's, Comedy, Musical) correspond to negative or near-zero weights in both users' profiles.

Limitations include the sparsity of the MovieLens 100k dataset, which may affect profile accuracy, and the binary nature of movie genre vectors, which oversimplifies movie characteristics. For instance, movie 95's genres are equally weighted (1s), ignoring potential varying genre prominence. Additionally, centering ratings may amplify noise in sparse user ratings, impacting profile reliability. These factors suggest that movie 95 may not be an ideal recommendation for either user, but user 200 remains the better candidate due to the relative similarity metric.

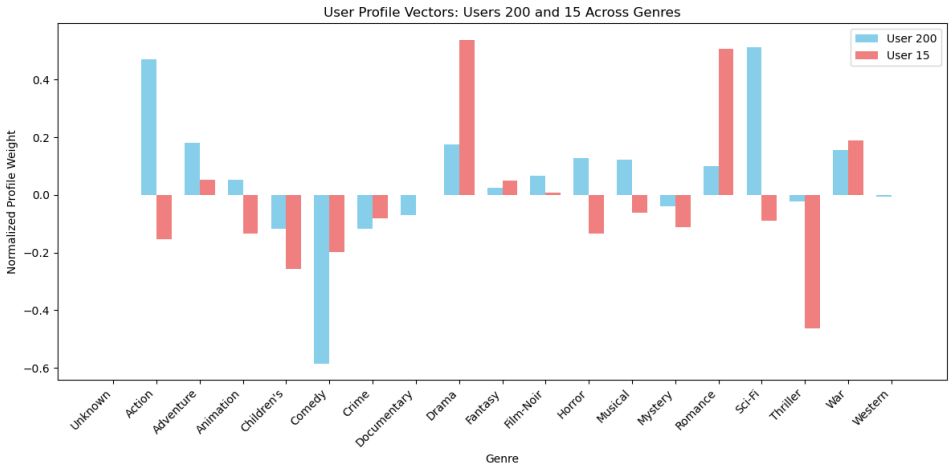

Figure 5: User Profile Vectors for Users 200 and 15

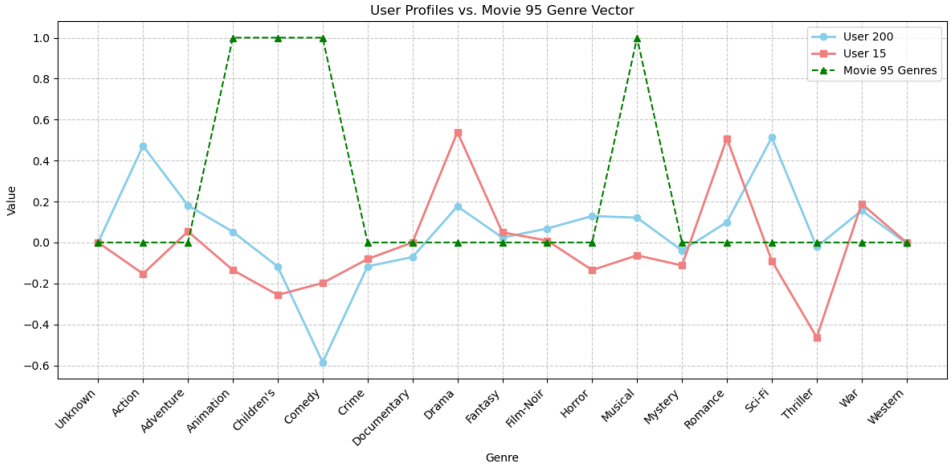

Figure 6: User Profiles vs. Movie 95 Genre Vector

2.5 Conclusion

This analysis successfully constructs genre-based user profiles for users 200 and 15, computes cosine similarities with movie 95's genre vector, and recommends movie 95 to user 200 based on a higher similarity (-0.2652 vs. -0.3259). The negative similarities highlight a genre mismatch, particularly due to both users' aversion to Comedy and Children's genres, as visualized in three insightful plots. Future improvements could incorporate weighted genre vectors, integrate additional movie features (e.g., directors, actors), or employ advanced techniques like matrix factorization to enhance profile accuracy and recommendation quality.