Part A: Data Preprocessing
using mean centering

Steps take: I did normalization in dataset (ratings-2.csv) file using

1. First calculating avg by doing grouping of user's ratings
    a. Used Excel filtering and formulas for calculating mean centering
2. Subtracting his rating from the average.

Note: Steps can be seen in meancentering_steps.csv (attached in submission)

Part B: User Based Recommendation

1. Created data model on normalized/mean-centered dataset(refer **ratings-2.csv** attached with submission)
2. Create similarity matrix using four types of similarity
    - Loglikelyhood, Spearman Correlation, Euclidean Distance, Pearson Correlation
3. Created UserNeighbourHood object
4. Created a recommender list (top 3 recommendations for userId 269)
5. Printed out recommendations for all types of similarities (on mean centered data)
6. Evaluation for RMSE, Precision, Recall, F-score **(Refer UserRecommendationEvaluation.java)**

Item Based recommendation:

1. Created data model on normalized/mean-centered dataset(refer **ratings-2.csv** attached with submission)
2. Create similarity matrix using four types of similarity
    - Loglikelyhood,Euclidean Distance, Pearson Correlation
3. Created a recommender list (top 3 recommendations for userId 269)
4. Printed out recommendations for all types of similarities (on mean centered data)
5. Evaluation for RMSE, Precision, Recall, F-score **(Refer ItemRecommendationEvaluation.java)**.

Part C: Matrix Factorization : CSV used :(refer ratings-1.csv and movies.csv)
Steps taken:

1. Calling spark session to register application
2. Loaded ratings-1.csv and parsed the dataset.
3. Created **movies.csv** (movieId, title) and loaded it. (took dataset from in-class exercise)
4. Created a tuple of (UserID, MovieID, Rating)
5. In order to determine the best ALS parameters, split the dataset into train, validation, and test datasets.
6. Prepared test and validation set.
7. Set parameter value for ALS method. Changed regularization_parameter value for different RMSE calculation.
8. Converted data frame to RDD and printed top 3 ranked item, and finally found best ranked item with lowest RMSE value.
9. Evaluated the matrix model by computing RMSE.

10. Visualized predictions generated by ALS Model.
11. Defined “getRecommendations(user,testDf,trainDf,model)” function to calculate recommendation for specific Userid.
12. Called above function to get recommendation for my UserId (269) and Made a matrix sorted in decreasing order of ratings and printed the result.
13. Made a matrix and printed the result with the movie titles respective to their movieId.

Part D: Evaluation Test:

1. **(Three Evaluation per method) Recommendation table for different values**

User based

Similarity type	1st recommendation	2nd recommendation	3rd recommendation
PearsonCorrelationSimilarity	12	8	1
LoglikelihoodSimilarity:	12	8	9
Euclidean similarity	12	8	9
Spearman similarity	12	8	1

Item Based

Similarity type	1st recommendation	2nd recommendation	3rd recommendation
PearsonCorrelationsimilarity	14	18	19
Loglikelihoodsimilarity:	1	14	19
Euclidean similarity	8	14	6

2. Evaluation(RMSE, Recall, Precision, F-score) for item Base and User Based using different lambda values:

relevance Threshold	**RMSE**	**F-score**	**Precision**	**Recall**
-3	0.8891726342077508	0.9749999999999999	0.9749999999999999	0.9749999999999999

Relevance Threshold	**RMSE**	**F-score**	**Precision**	**Recall**
-4	1.0499976735831267	0.94	0.94	0.94

3. Comparison in Userbased and Item Based Recommendations (both are using Euclidean Similarity)

RMSE for different lambda	**Lambda value : -4**	**Lambda value :-1**	**Lambda value :-3**
User based	0.9379408782071857	0.9643800972316054	0.8891726342077508
Item based	1.0499976735831267	1.0715936596901925	1.066313017517381

Regularization_parameter (Matrix factorization Based)	0.01	.15	.5
RMSE	2.06493562009	1.33667158335	1.39750103

Final RMSE Table for evaluation is below:

	UserBased	ItemBased	MatrixFactorization
RMSE	0.8891726342077508	1.0499976735831267	1.33667158335

From the above table , we can evaluate that best Recommendation is using **User Based Recommendation** , as the **lowest RMSE is found is : 0.8891726342077508** (evaluated by UserBased recommendation)