This project was conducted in Databricks with PySpark and was based on a dataset of 30 users rating 100 movies, while each user did not necessarily rate for all 100 movies. The dataset had 3 columns, being the movieId, rating and userId. There were 1501 rows representing 1501 ratings by all users. By doing an exploration analysis, there were the following findings:

Top 15 movies with the highest average ratings: 32,90,30,94,23,49,18,29,52,53,62,92,46,68,87

Top 15 movies with the highest number of ratings: 6,29,51,22,50,94,55,68,2,15,85,36,86,88,45

Top 10 users providing highest average ratings: 11,26,22,23,2,17,8,24,12,3

Top 10 users rated for the most times: 6,14,22,11,12,4,7,9,24,23

Alternating least squares algorithms were trained and the best one was with a 8:2 train-test ratio and mean absolute error (MAE) metric. After tunning for other five hyperparameters, the MAE reached down to 0.61. Finally the model was used to predict movie recommendations to users 9 and 13.:

Top 10 movies recommended to user 9: [62, 46, 17, 23, 13, 29, 27, 55, 65, 48]

Top 10 movies recommended to user 13: [30, 2, 41, 70, 69, 32, 75, 92, 76, 8]