Methodology steps

1. Dataset from GitHub
2. Unzipping the dataset
3. KT1-KT2, KT3, KT4 :
4. Concatenation & append with respect to USER ID
5. Converting the csv into feather file
6. Top 50 Active Users
7. Joined/Merged Dataframe with the question id csv
8. Checked Users submit answers ith answer key to check correctness
9. Created a “CORRECT” Column which indicates 1 as Correct , 0 as Incorrect
10. Calculation of PERCENTAGES of USERS
11. Created a dataframe which indicates each user how many questions they attempted how many are corrected/incorrected and PERCENTAGES.
12. Which user is the most corrected and stand as 1st
13. Identification of # of users who’s percentages => 95
14. Identification of # of users who’s percentages => 90
15. Identification of # of users who’s percentages => 80
16. Identification of # of users who’s percentages are lowest
17. Repeated Questions Check
18. Removal of Repeated Questions
19. # of days/time each user spent
20. # of users who’s time is not recorded
21. Removed the outliers
22. End Year / Join Year Histogram Plot
23. Taking the data where user’s percentages >=80
24. Users count = 28K
25. Pivot table of users with bundle
26. At which order, bundles are read by each user
27. Indexing the user id
28. Creating bundles read column in the dataframe for each user
29. Histogram plot between users and bundles read
30. Remove users who read less than 4 and greater than 15
31. Remove those bundles which nobody read
32. MIN MAX SCALER
33. PCA
34. K-MEANS
35. SSE
36. ELBOW METHOD
37. K-MEANS
38. Average of bundles (For recommender system)
39. K-NN for recommender system
40. Recommending the 6 similar users to specific user