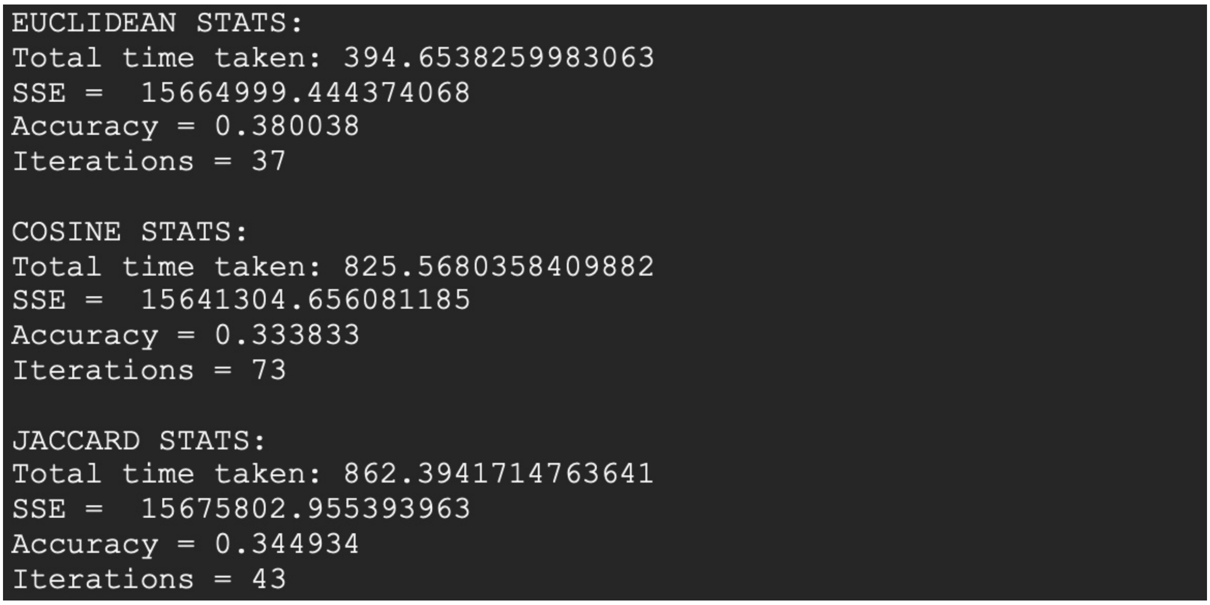
CAP5610 – Machine Learning

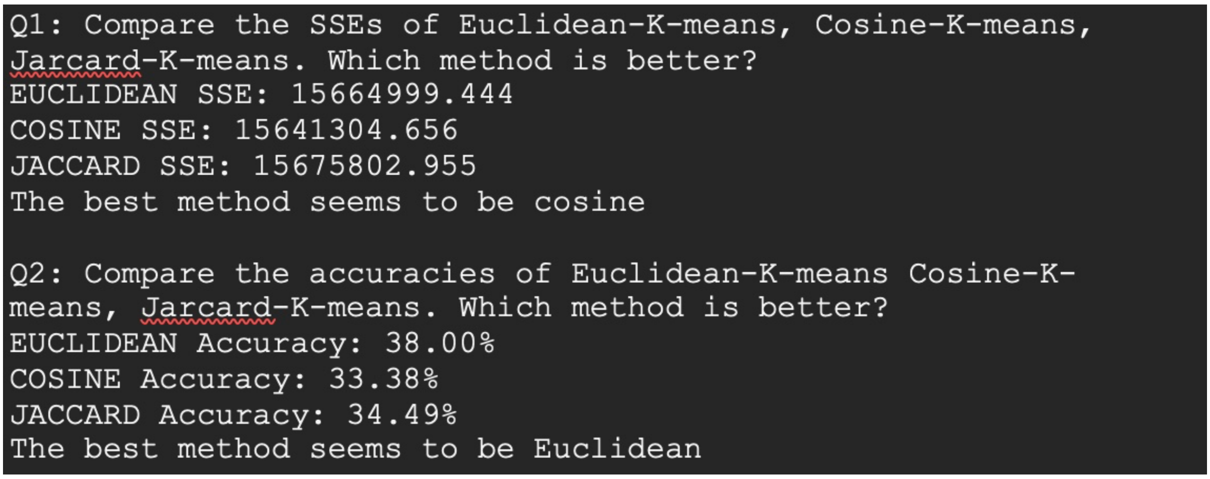
Prudvi Kamtam
UCF ID: 5498416

K- Means

Stopping criteria: If centroids don’t change

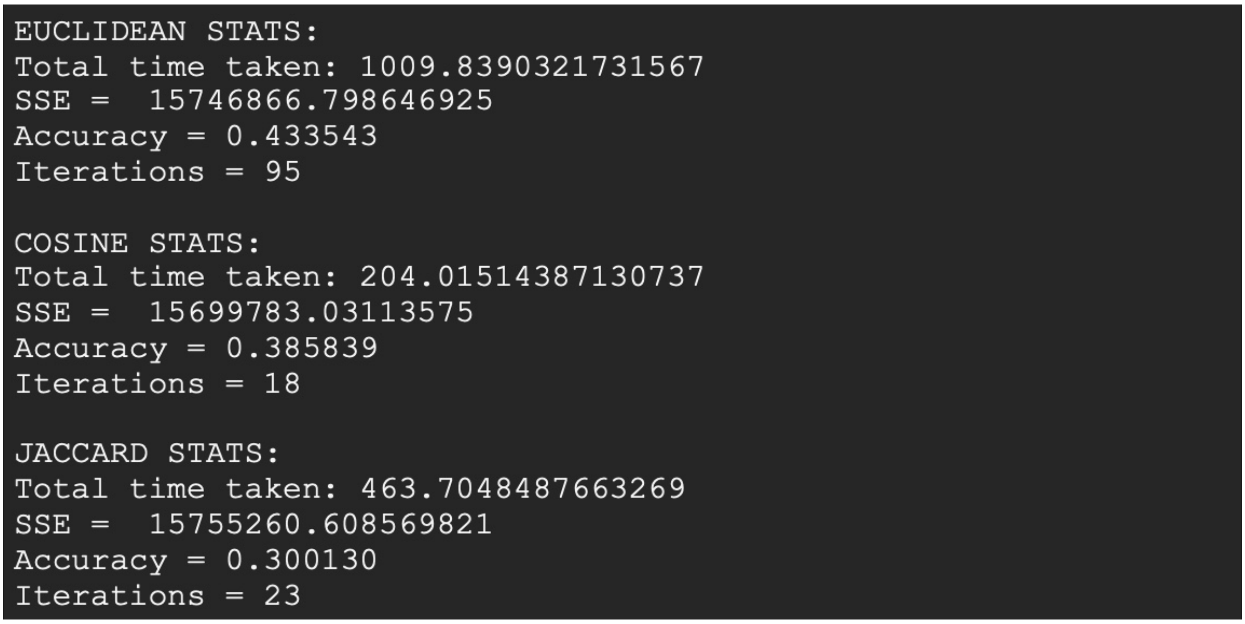


```
EUCLIDEAN STATS:
Total time taken: 394.6538259983063
SSE =  15664999.444374068
Accuracy = 0.380038
Iterations = 37

COSINE STATS:
Total time taken: 825.5680358409882
SSE =  15641304.656081185
Accuracy = 0.333833
Iterations = 73

JACCARD STATS:
Total time taken: 862.3941714763641
SSE =  15675802.955393963
Accuracy = 0.344934
Iterations = 43
```

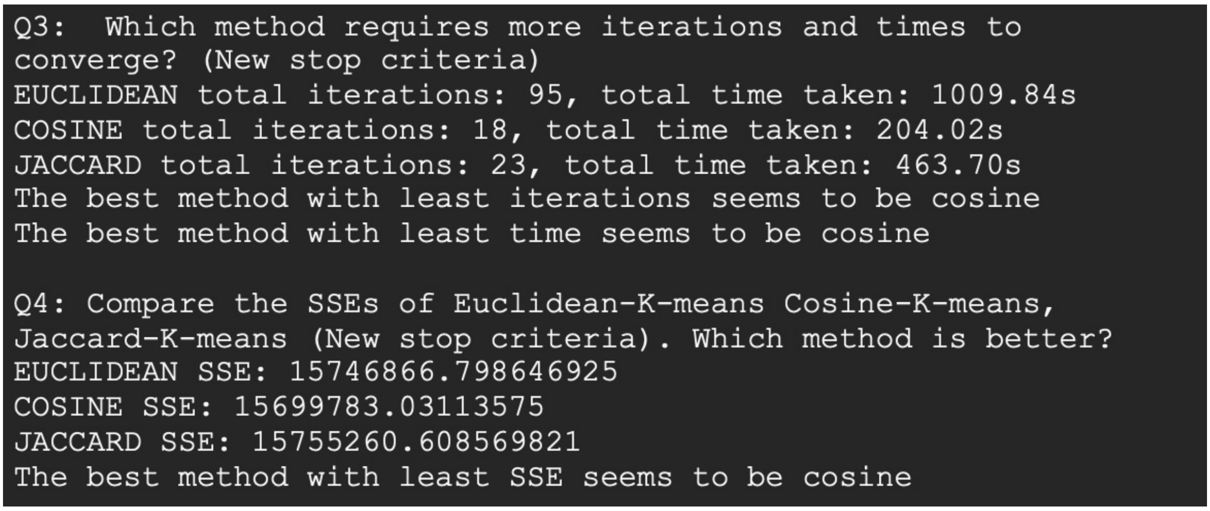


```
Q1: Compare the SSEs of Euclidean-K-means, Cosine-K-means,
Jarcard-K-means. Which method is better?
EUCLIDEAN SSE: 15664999.444
COSINE SSE: 15641304.656
JACCARD SSE: 15675802.955
The best method seems to be cosine

Q2: Compare the accuracies of Euclidean-K-means Cosine-K-
means, Jarcard-K-means. Which method is better?
EUCLIDEAN Accuracy: 38.00%
COSINE Accuracy: 33.38%
JACCARD Accuracy: 34.49%
The best method seems to be Euclidean
```

Stopping criteria: “when there is no change in centroid position OR when the SSE value increases in the next iteration OR when the maximum preset value (e.g., 500, you can set the preset value by yourself) of iteration is complete”



```
EUCLIDEAN STATS:
Total time taken: 1009.8390321731567
SSE =  15746866.798646925
Accuracy = 0.433543
Iterations = 95

COSINE STATS:
Total time taken: 204.01514387130737
SSE =  15699783.03113575
Accuracy = 0.385839
Iterations = 18

JACCARD STATS:
Total time taken: 463.7048487663269
SSE =  15755260.608569821
Accuracy = 0.300130
Iterations = 23
```



```
Q3:  Which method requires more iterations and times to
converge? (New stop criteria)
EUCLIDEAN total iterations: 95, total time taken: 1009.84s
COSINE total iterations: 18, total time taken: 204.02s
JACCARD total iterations: 23, total time taken: 463.70s
The best method with least iterations seems to be cosine
The best method with least time seems to be cosine

Q4: Compare the SSEs of Euclidean-K-means Cosine-K-means,
Jaccard-K-means (New stop criteria). Which method is better?
EUCLIDEAN SSE: 15746866.798646925
COSINE SSE: 15699783.03113575
JACCARD SSE: 15755260.608569821
The best method with least SSE seems to be cosine
```

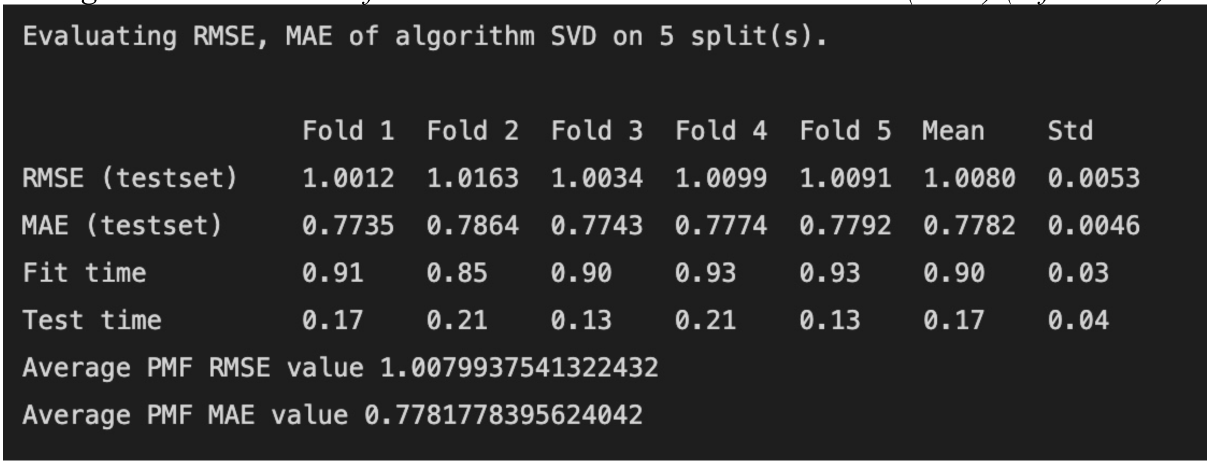
Q5: Summary observations/Algorithm Analysis

- Euclidean seems to be taking the longest time in every case
-

Recommender Systems

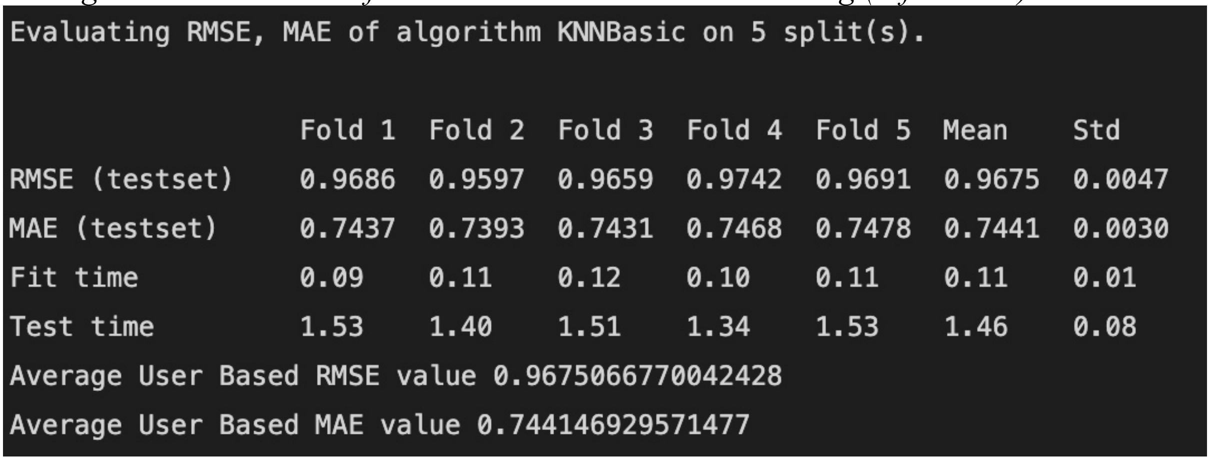
Question C.

Average MAE and RMSE of the Probabilistic Matrix Factorization (PMF) (5-folds CV)



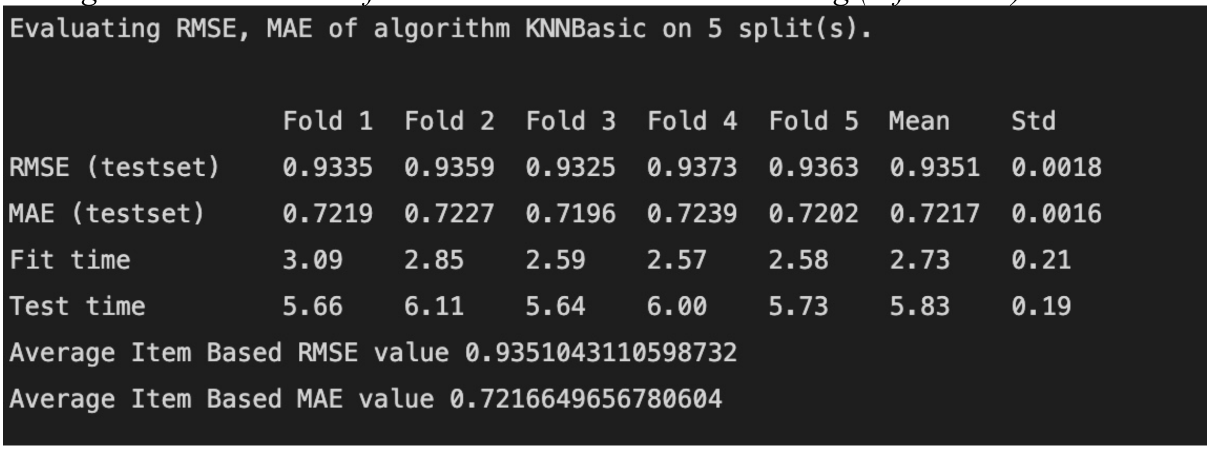
```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.0012  1.0163  1.0034  1.0099  1.0091  1.0080  0.0053
MAE (testset)     0.7735  0.7864  0.7743  0.7774  0.7792  0.7782  0.0046
Fit time          0.91    0.85    0.90    0.93    0.93    0.90    0.03
Test time         0.17    0.21    0.13    0.21    0.13    0.17    0.04
Average PMF RMSE value 1.0079937541322432
Average PMF MAE value 0.7781778395624042
```

Average MAE and RMSE of User based Collaborative Filtering (5-folds CV)



```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9686  0.9597  0.9659  0.9742  0.9691  0.9675  0.0047
MAE (testset)     0.7437  0.7393  0.7431  0.7468  0.7478  0.7441  0.0030
Fit time          0.09    0.11    0.12    0.10    0.11    0.11    0.01
Test time         1.53    1.40    1.51    1.34    1.53    1.46    0.08
Average User Based RMSE value 0.9675066770042428
Average User Based MAE value 0.744146929571477
```

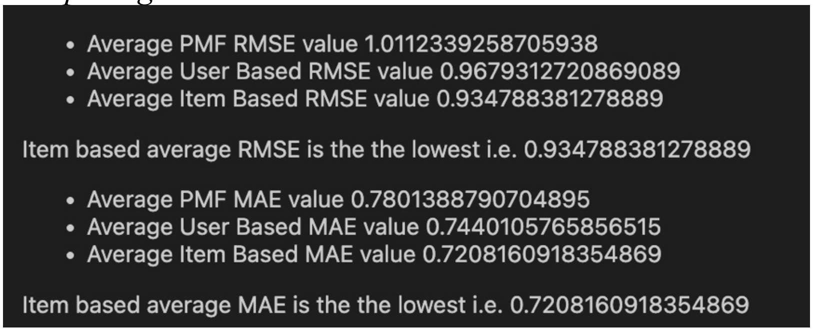
Average MAE and RMSE of Item based Collaborative Filtering (5-folds CV)



```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9335  0.9359  0.9325  0.9373  0.9363  0.9351  0.0018
MAE (testset)     0.7219  0.7227  0.7196  0.7239  0.7202  0.7217  0.0016
Fit time          3.09    2.85    2.59    2.57    2.58    2.73    0.21
Test time         5.66    6.11    5.64    6.00    5.73    5.83    0.19
Average Item Based RMSE value 0.9351043110598732
Average Item Based MAE value 0.7216649656780604
```

Question D.

Comparing RMSE values:



- Average PMF RMSE value 1.0112339258705938
- Average User Based RMSE value 0.9679312720869089
- Average Item Based RMSE value 0.934788381278889

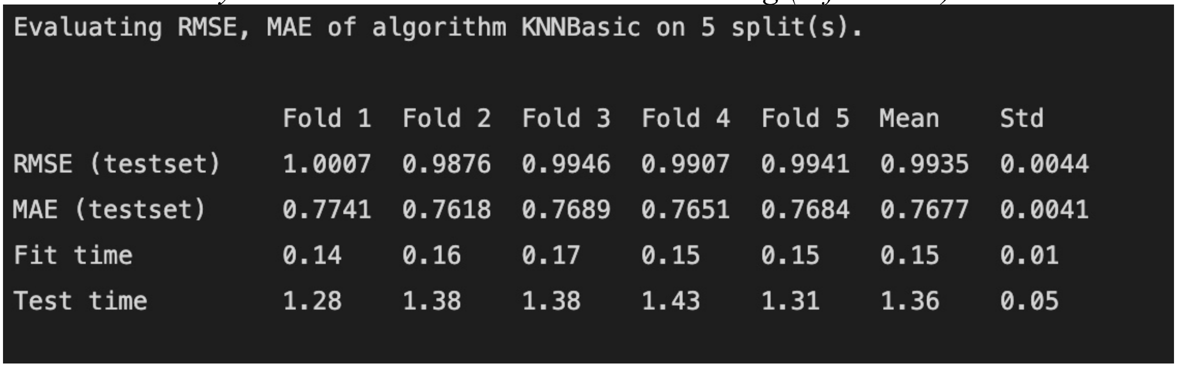
Item based average RMSE is the the lowest i.e. 0.934788381278889

- Average PMF MAE value 0.7801388790704895
- Average User Based MAE value 0.7440105765856515
- Average Item Based MAE value 0.7208160918354869

Item based average MAE is the the lowest i.e. 0.7208160918354869

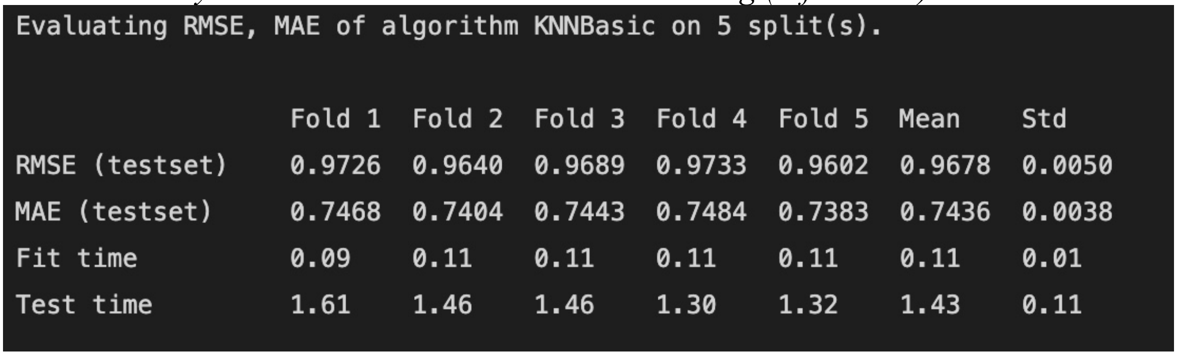
Question E.

Cosine Similarity with User based Collaborative Filtering (5-folds CV)



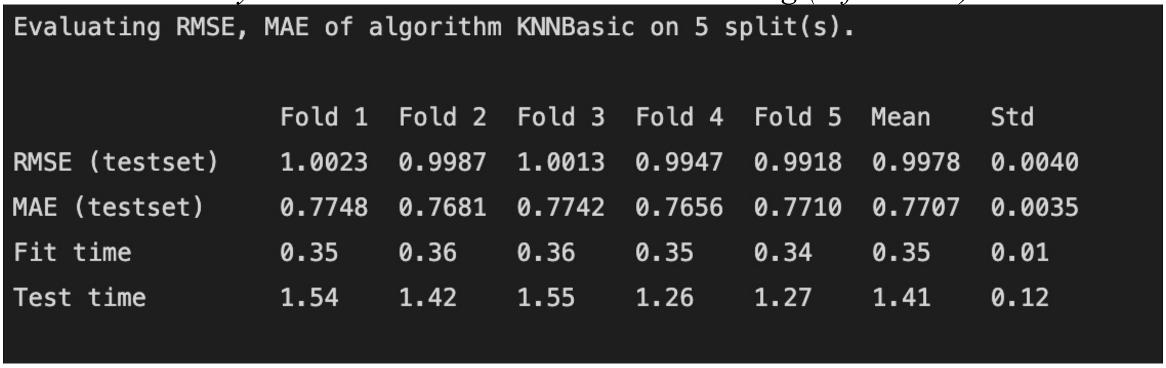
```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).


                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.0007  0.9876  0.9946  0.9907  0.9941  0.9935  0.0044
MAE (testset)     0.7741  0.7618  0.7689  0.7651  0.7684  0.7677  0.0041
Fit time          0.14    0.16    0.17    0.15    0.15    0.15    0.01
Test time         1.28    1.38    1.38    1.43    1.31    1.36    0.05
```

MSD Similarity with User based Collaborative Filtering (5-folds CV)



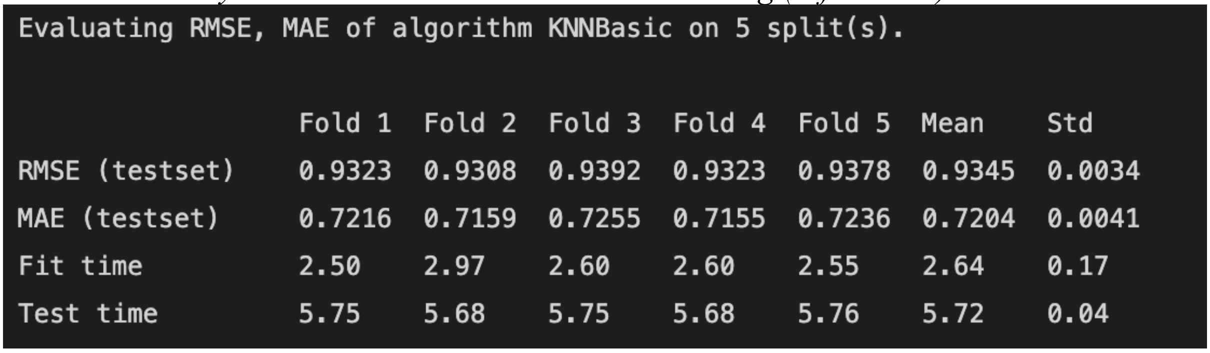
```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).


                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9726  0.9640  0.9689  0.9733  0.9602  0.9678  0.0050
MAE (testset)     0.7468  0.7404  0.7443  0.7484  0.7383  0.7436  0.0038
Fit time          0.09    0.11    0.11    0.11    0.11    0.11    0.01
Test time         1.61    1.46    1.46    1.30    1.32    1.43    0.11
```

Pearson Similarity with User based Collaborative Filtering (5-folds CV)



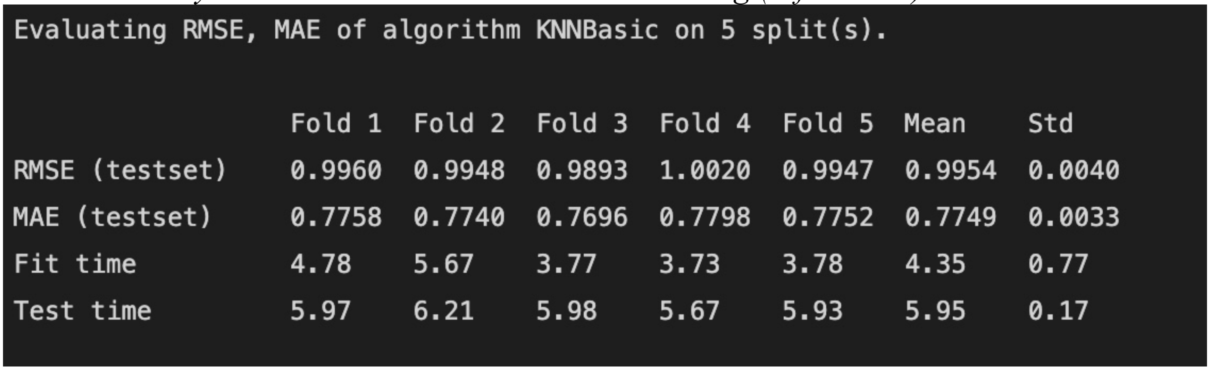
```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).


                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.0023  0.9987  1.0013  0.9947  0.9918  0.9978  0.0040
MAE (testset)     0.7748  0.7681  0.7742  0.7656  0.7710  0.7707  0.0035
Fit time          0.35    0.36    0.36    0.35    0.34    0.35    0.01
Test time         1.54    1.42    1.55    1.26    1.27    1.41    0.12
```

Cosine Similarity with Item based Collaborative Filtering (5-folds CV)



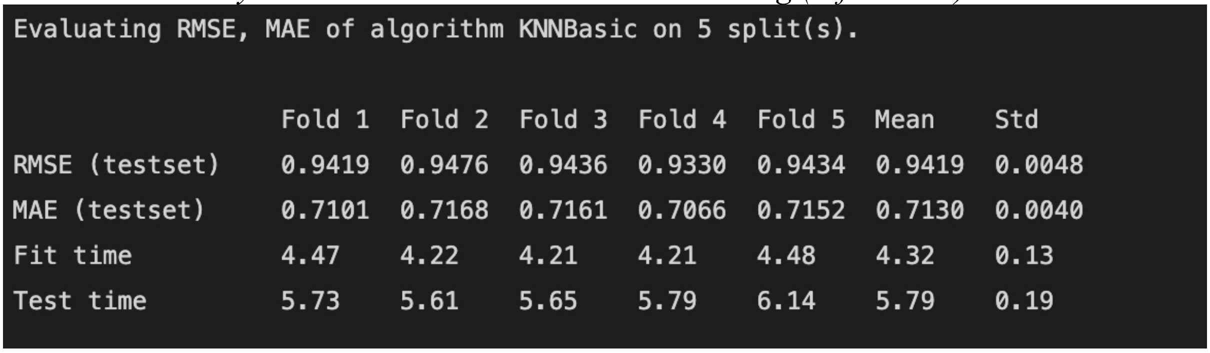
```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).


                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9323  0.9308  0.9392  0.9323  0.9378  0.9345  0.0034
MAE (testset)     0.7216  0.7159  0.7255  0.7155  0.7236  0.7204  0.0041
Fit time          2.50    2.97    2.60    2.60    2.55    2.64    0.17
Test time         5.75    5.68    5.75    5.68    5.76    5.72    0.04
```

MSD Similarity with Item based Collaborative Filtering (5-folds CV)



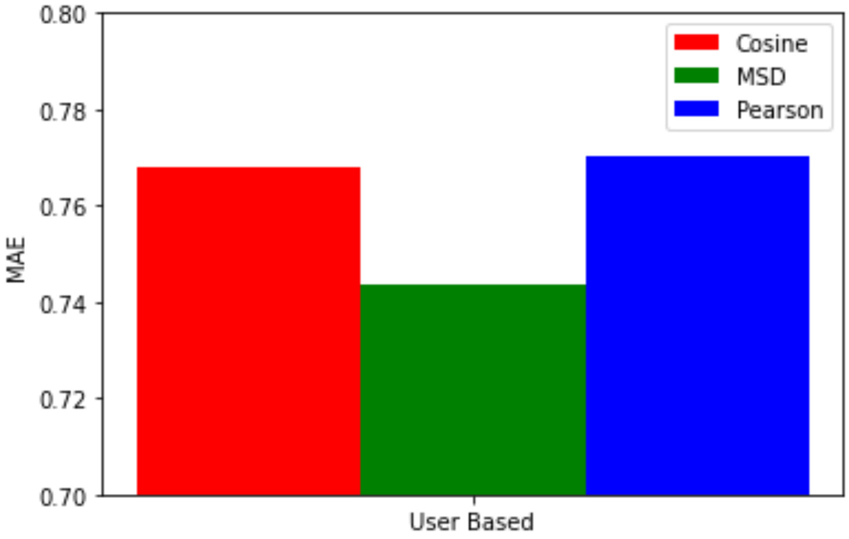
```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).


                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9960  0.9948  0.9893  1.0020  0.9947  0.9954  0.0040
MAE (testset)     0.7758  0.7740  0.7696  0.7798  0.7752  0.7749  0.0033
Fit time          4.78    5.67    3.77    3.73    3.78    4.35    0.77
Test time         5.97    6.21    5.98    5.67    5.93    5.95    0.17
```

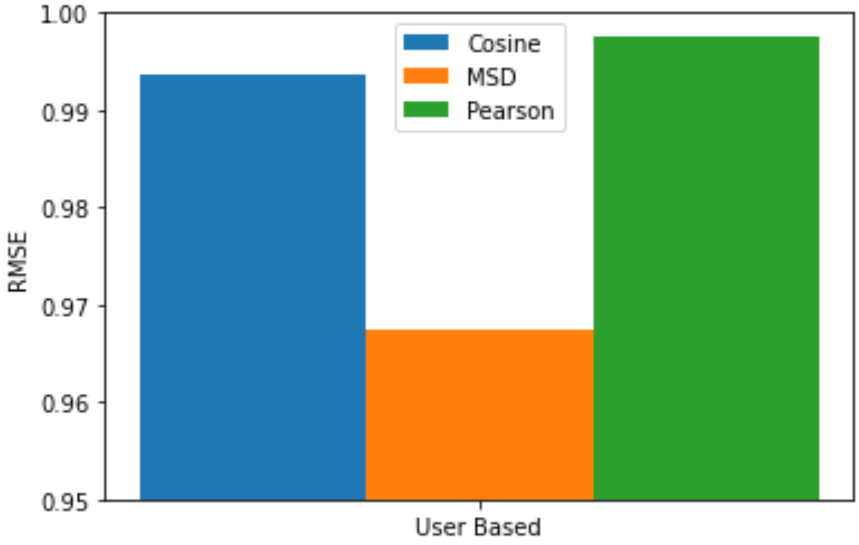
Pearson Similarity with Item based Collaborative Filtering (5-folds CV)

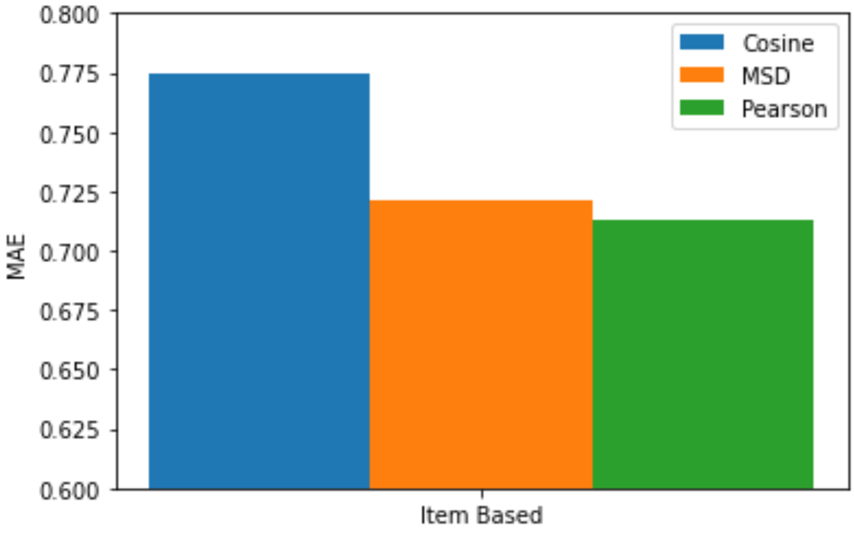


```
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).


                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9419  0.9476  0.9436  0.9330  0.9434  0.9419  0.0048
MAE (testset)     0.7101  0.7168  0.7161  0.7066  0.7152  0.7130  0.0040
Fit time          4.47    4.22    4.21    4.21    4.48    4.32    0.13
Test time         5.73    5.61    5.65    5.79    6.14    5.79    0.19
```

Plot your results





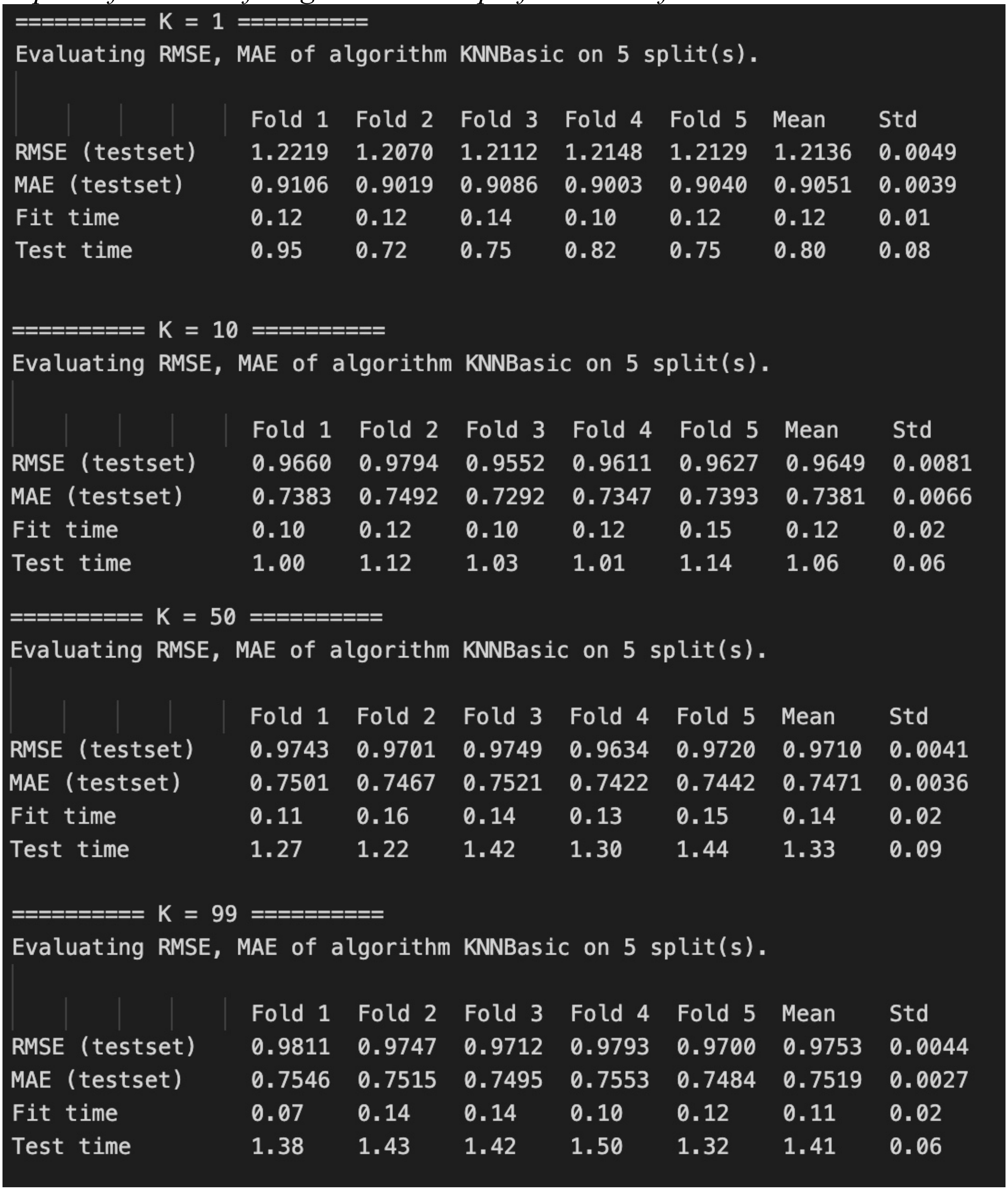


Is the impact of three metrics consistent between user-based and item-based?

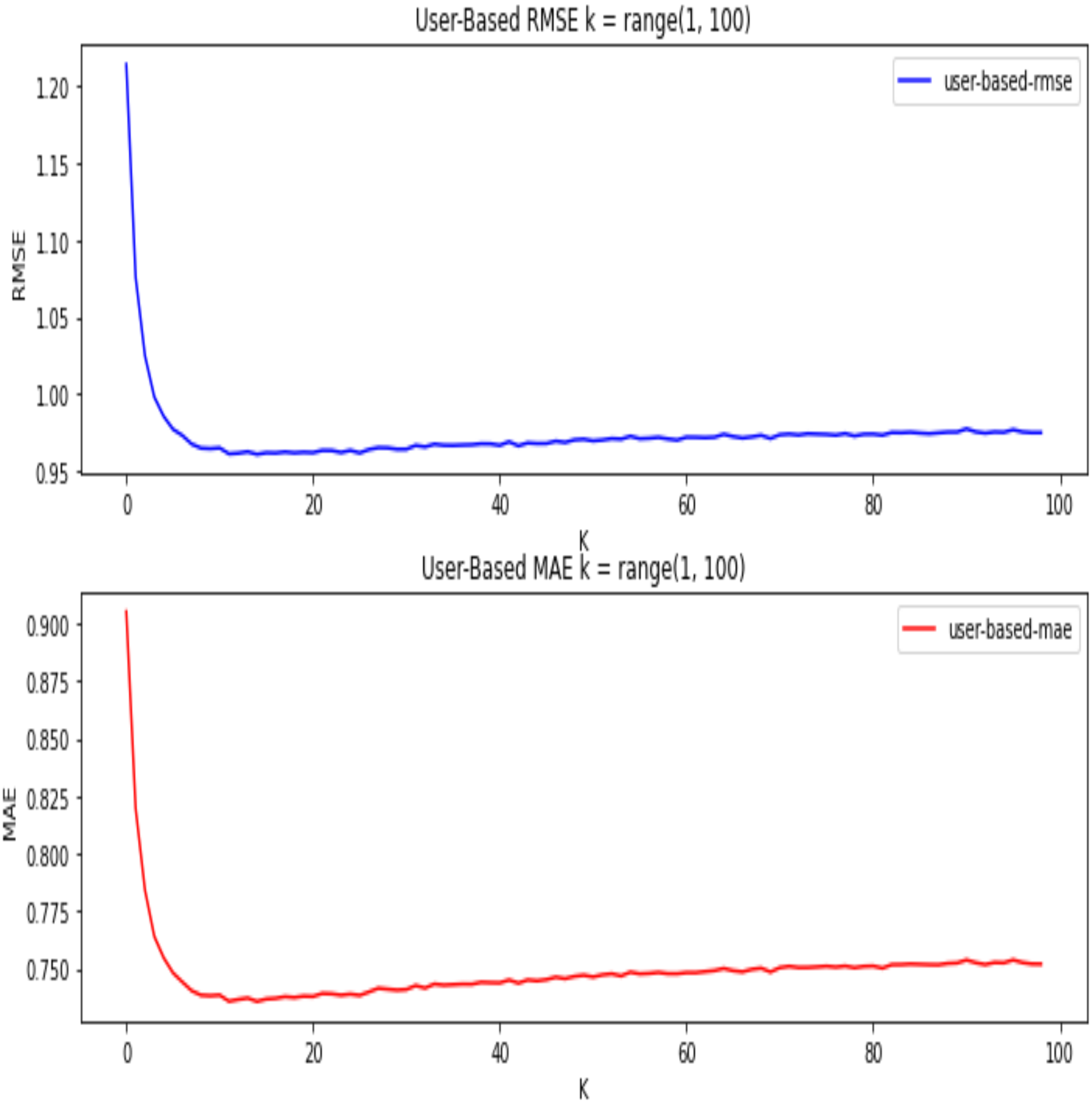
The impact of the three metrics seems to be inconsistent when comparing with MAE but somewhat consistent when comparing with RMSE

Question F.

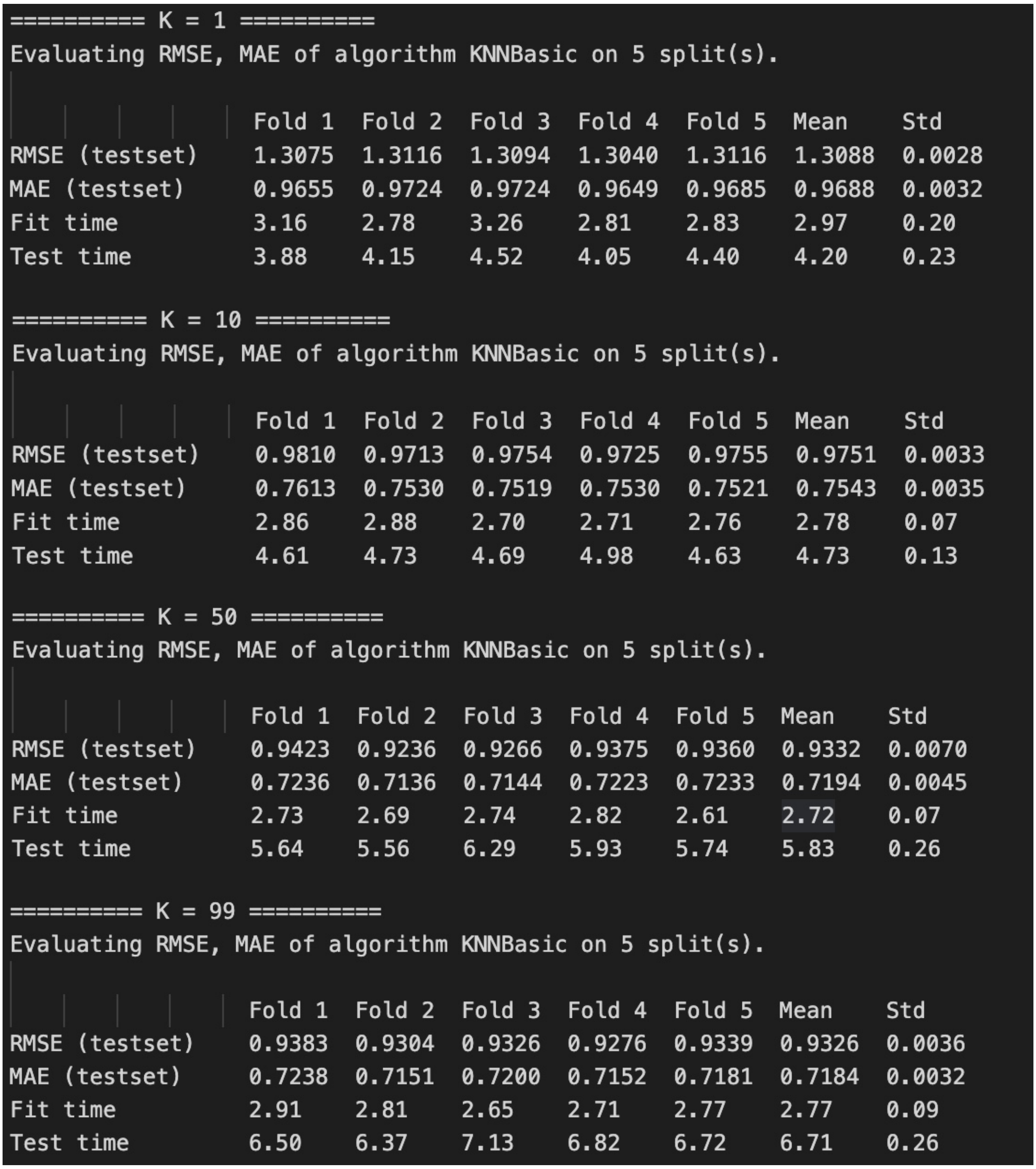
Impact of number of neighbors on the performance of User-based



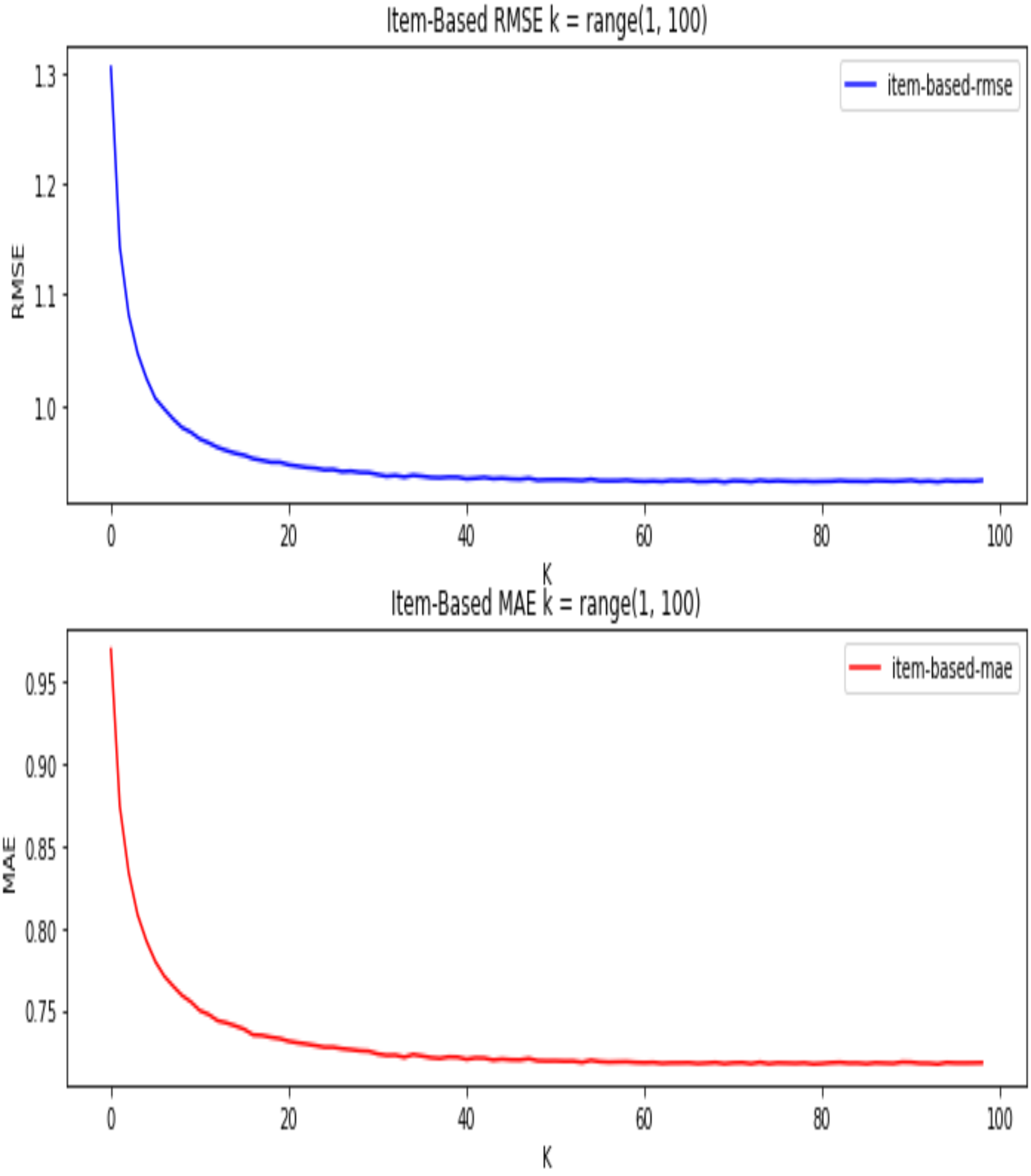
```
========== K = 1 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      1.2219  1.2070  1.2112  1.2148  1.2129  1.2136  0.0049
MAE (testset)       0.9106  0.9019  0.9086  0.9003  0.9040  0.9051  0.0039
Fit time            0.12    0.12    0.14    0.10    0.12    0.12    0.01
Test time           0.95    0.72    0.75    0.82    0.75    0.80    0.08


========== K = 10 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      0.9660  0.9794  0.9552  0.9611  0.9627  0.9649  0.0081
MAE (testset)       0.7383  0.7492  0.7292  0.7347  0.7393  0.7381  0.0066
Fit time            0.10    0.12    0.10    0.12    0.15    0.12    0.02
Test time           1.00    1.12    1.03    1.01    1.14    1.06    0.06

========== K = 50 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      0.9743  0.9701  0.9749  0.9634  0.9720  0.9710  0.0041
MAE (testset)       0.7501  0.7467  0.7521  0.7422  0.7442  0.7471  0.0036
Fit time            0.11    0.16    0.14    0.13    0.15    0.14    0.02
Test time           1.27    1.22    1.42    1.30    1.44    1.33    0.09

========== K = 99 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      0.9811  0.9747  0.9712  0.9793  0.9700  0.9753  0.0044
MAE (testset)       0.7546  0.7515  0.7495  0.7553  0.7484  0.7519  0.0027
Fit time            0.07    0.14    0.14    0.10    0.12    0.11    0.02
Test time           1.38    1.43    1.42    1.50    1.32    1.41    0.06
```

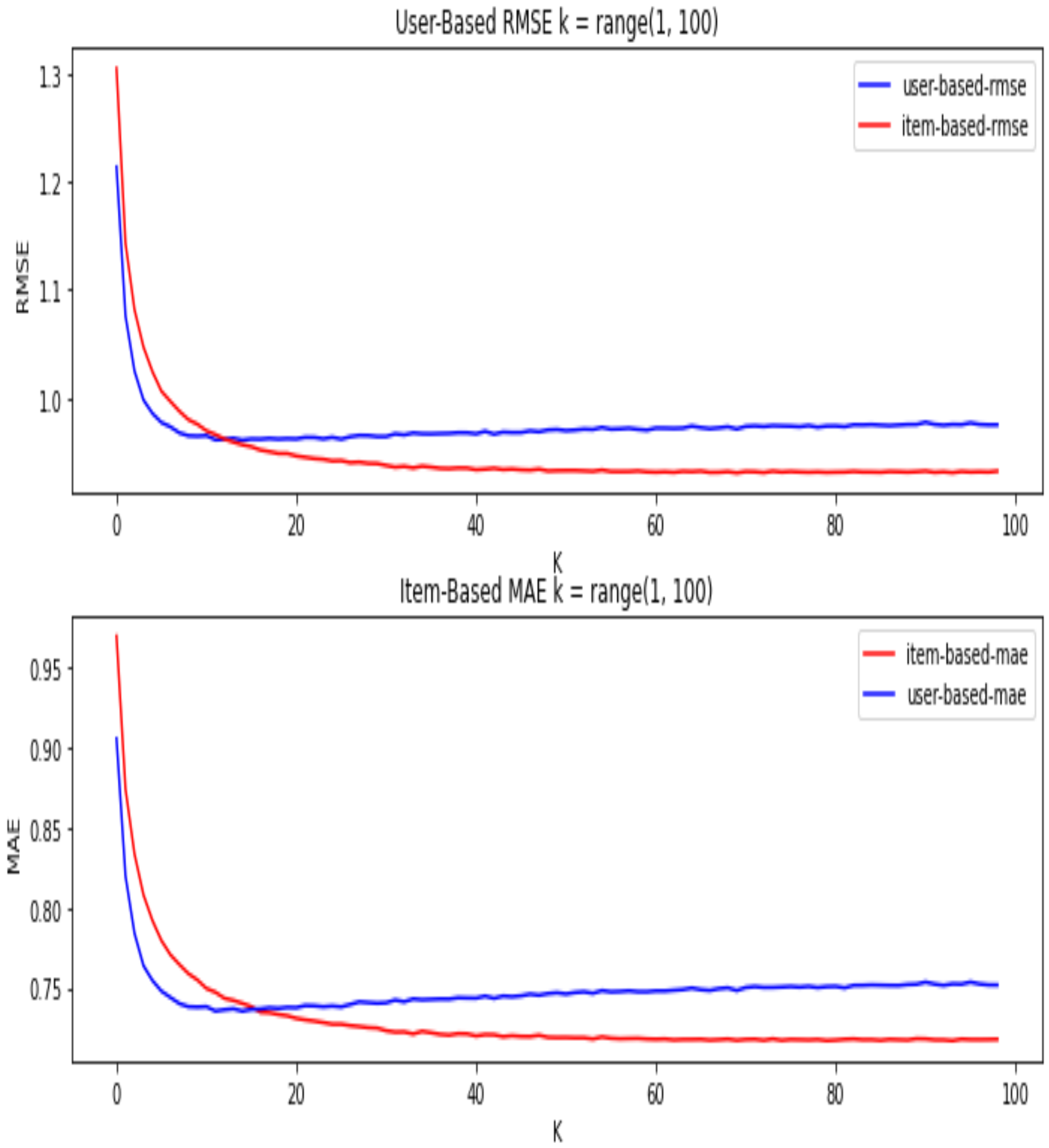

User-Based RMSE k = range(1, 100)
user-based-rmse
RMSE
1.20
1.15
1.10
1.05
1.00
0.95
0
20
40
60
80
100
K
User-Based MAE k = range(1, 100)
user-based-mae
MAE
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0
20
40
60
80
100
K

Impact of number of neighbors on the performance of Item-based

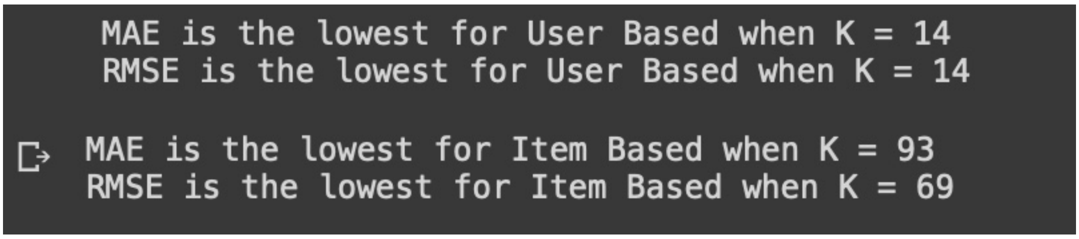


```
========== K = 1 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      1.3075  1.3116  1.3094  1.3040  1.3116  1.3088  0.0028
MAE (testset)       0.9655  0.9724  0.9724  0.9649  0.9685  0.9688  0.0032
Fit time            3.16    2.78    3.26    2.81    2.83    2.97    0.20
Test time           3.88    4.15    4.52    4.05    4.40    4.20    0.23

========== K = 10 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      0.9810  0.9713  0.9754  0.9725  0.9755  0.9751  0.0033
MAE (testset)       0.7613  0.7530  0.7519  0.7530  0.7521  0.7543  0.0035
Fit time            2.86    2.88    2.70    2.71    2.76    2.78    0.07
Test time           4.61    4.73    4.69    4.98    4.63    4.73    0.13

========== K = 50 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      0.9423  0.9236  0.9266  0.9375  0.9360  0.9332  0.0070
MAE (testset)       0.7236  0.7136  0.7144  0.7223  0.7233  0.7194  0.0045
Fit time            2.73    2.69    2.74    2.82    2.61    2.72    0.07
Test time           5.64    5.56    6.29    5.93    5.74    5.83    0.26

========== K = 99 ==========
Evaluating RMSE, MAE of algorithm KNNBasic on 5 split(s).
│
│   │   │   │   │   Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)      0.9383  0.9304  0.9326  0.9276  0.9339  0.9326  0.0036
MAE (testset)       0.7238  0.7151  0.7200  0.7152  0.7181  0.7184  0.0032
Fit time            2.91    2.81    2.65    2.71    2.77    2.77    0.09
Test time           6.50    6.37    7.13    6.82    6.72    6.71    0.26
```


Item-Based RMSE k = range(1, 100)
item-based-rmse
RMSE
1.3
1.2
1.1
1.0
0
20
40
60
80
100
K
Item-Based MAE k = range(1, 100)
item-based-mae
MAE
0.95
0.90
0.85
0.80
0.75
0
20
40
60
80
100
K



Question G.



```
  MAE is the lowest for User Based when K = 14
  RMSE is the lowest for User Based when K = 14

 MAE is the lowest for Item Based when K = 93
 RMSE is the lowest for Item Based when K = 69
```