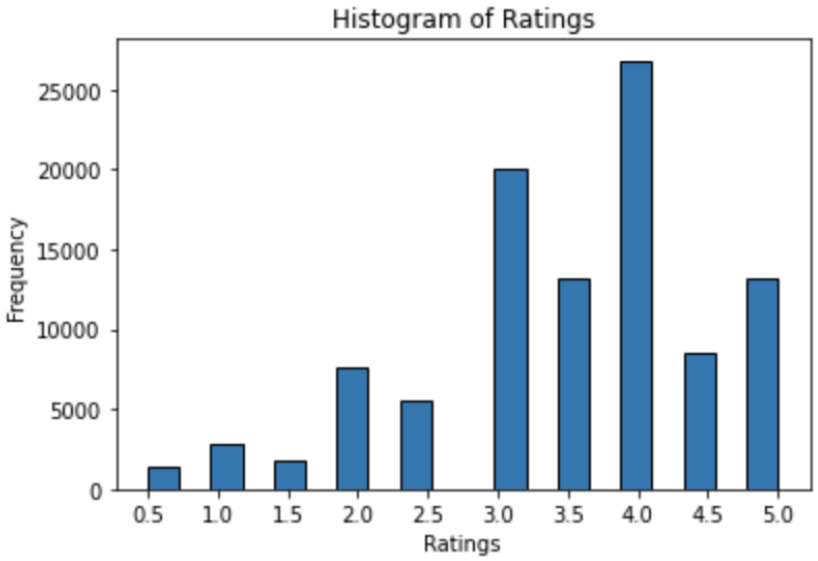
Project 3_ Report

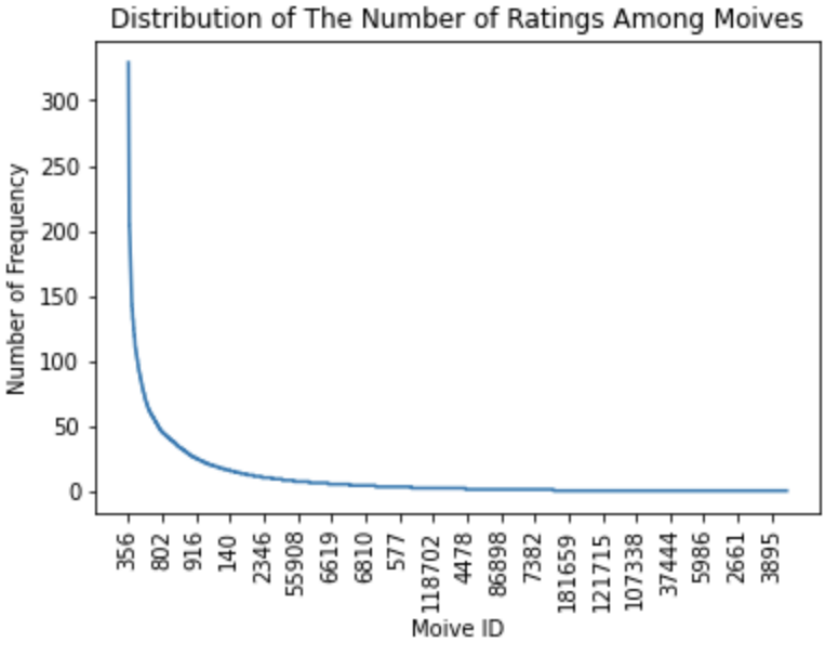
Haoting Ni (905545789), Yikai Wang (905522085), Yuanxuan Fang (005949389)

Q1: A: The sparsity of the movie rating is 0.016999683055613623

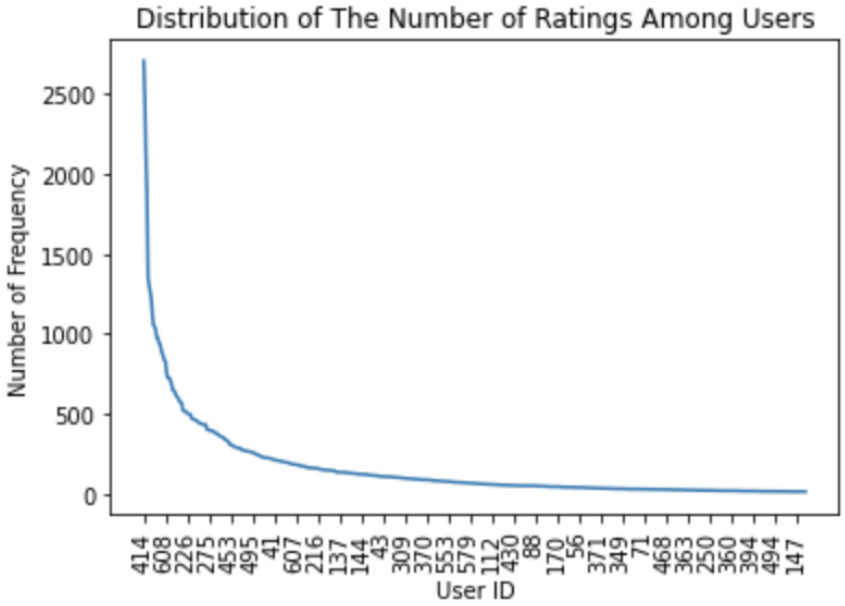
B:



C:

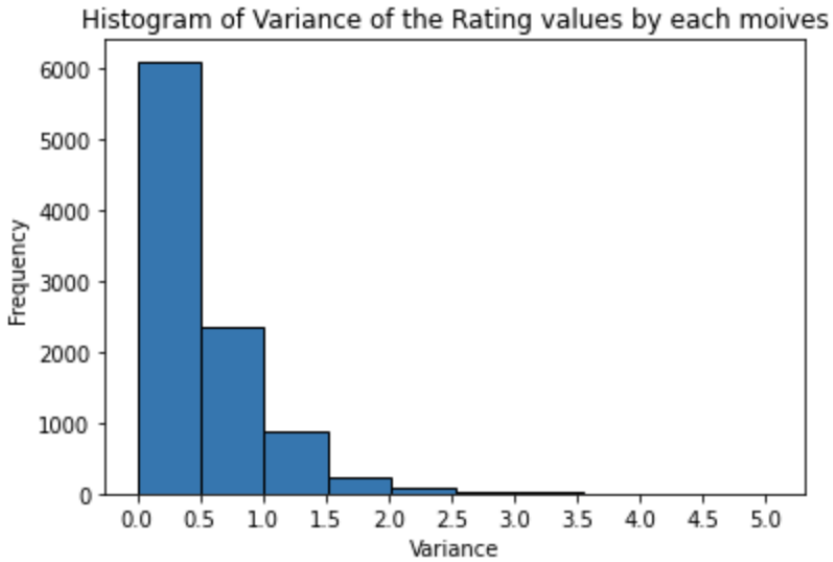


D:



E: Both C and D distribution decrease steeply, which means a few movies get most rates and a few users provide most rates. On the other hand, most movies did not have enough rates, and most users also did not provide enough rates. These shows that the ratings matrices are sparse.

F:



Q2:

A:

$$\mu_u = \frac{\Sigma_{k \epsilon I_u} r_{uk}}{|I_u|}$$

As description says, the mean rating for user u defined by the sum of rating the user u to the K divided by the items indices.

B:

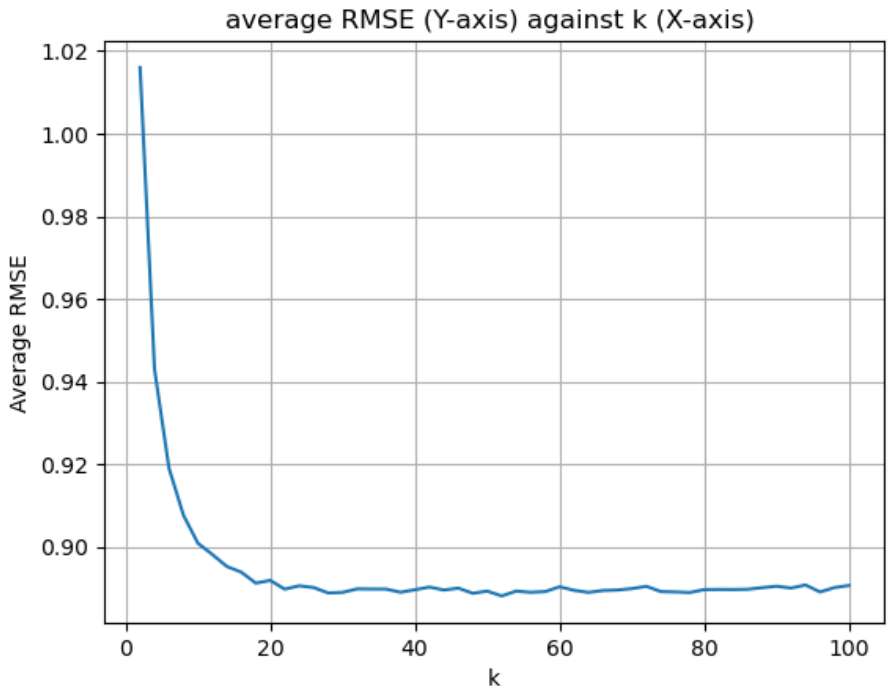
$I_u \cap I_v$ means that two sets of users U and V rating a field of movies' intersection part. Which is movie that both users rated.

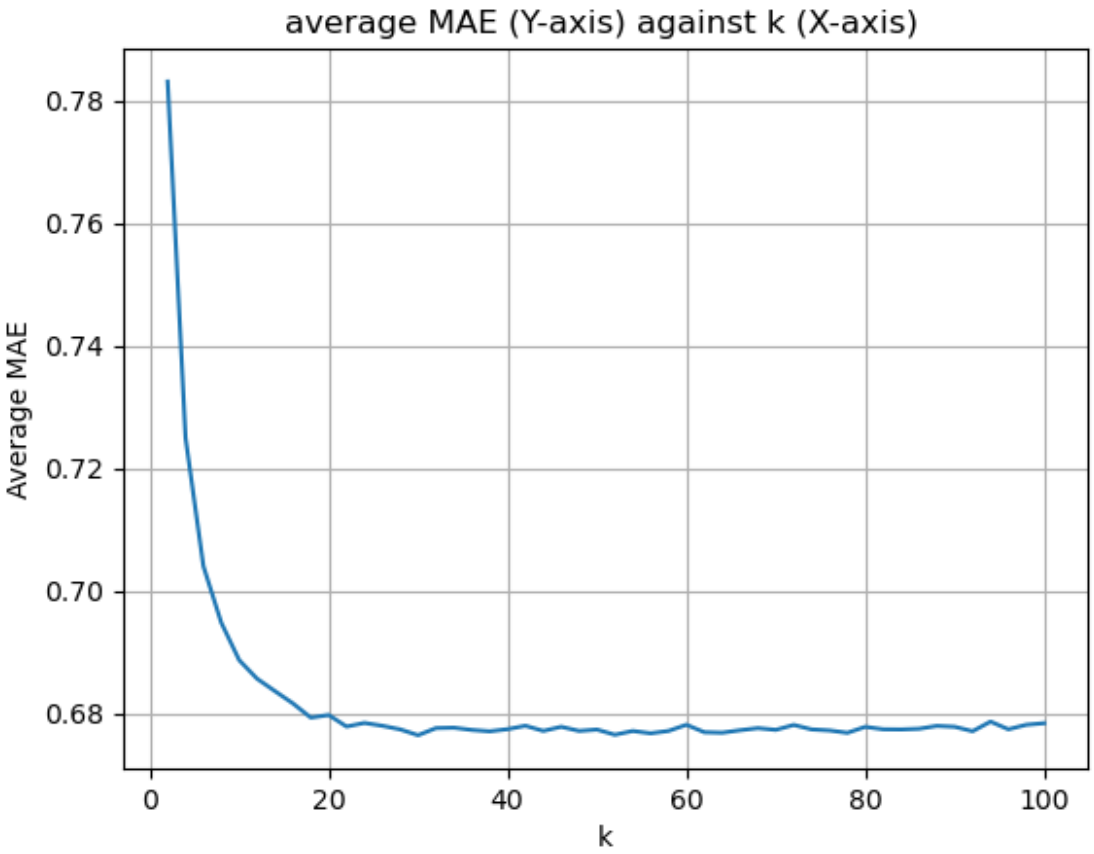
It could be empty, which means they have not been rating a same movie.

Q3:

Consider users who either rate all items highly or rate all items poorly and the impact of these users on the prediction function, this can reduce the noise of the dataset by subtracting their means of movie ratings.

Q4:

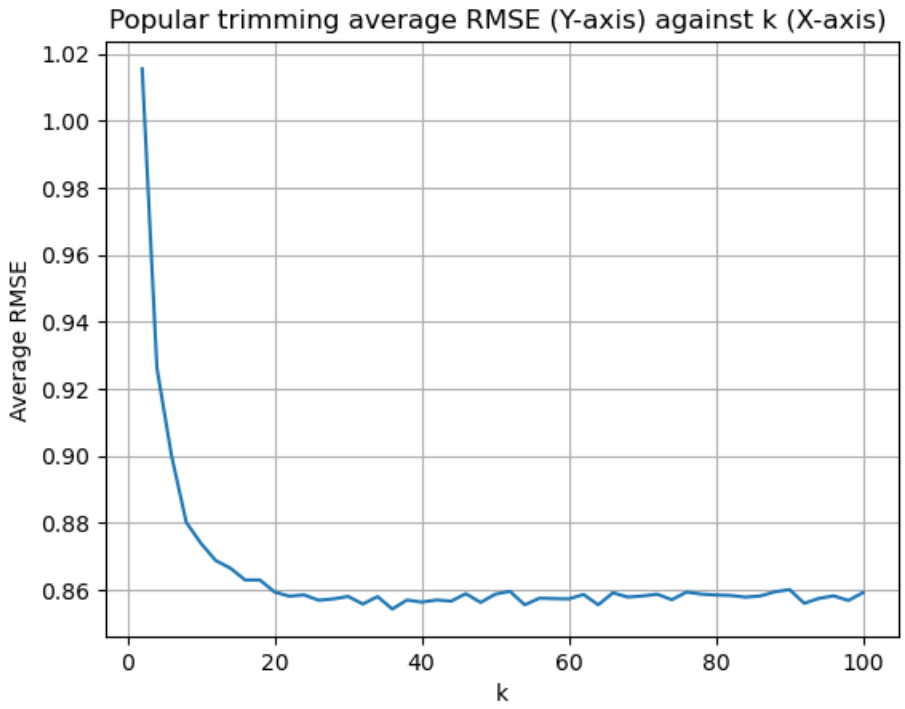




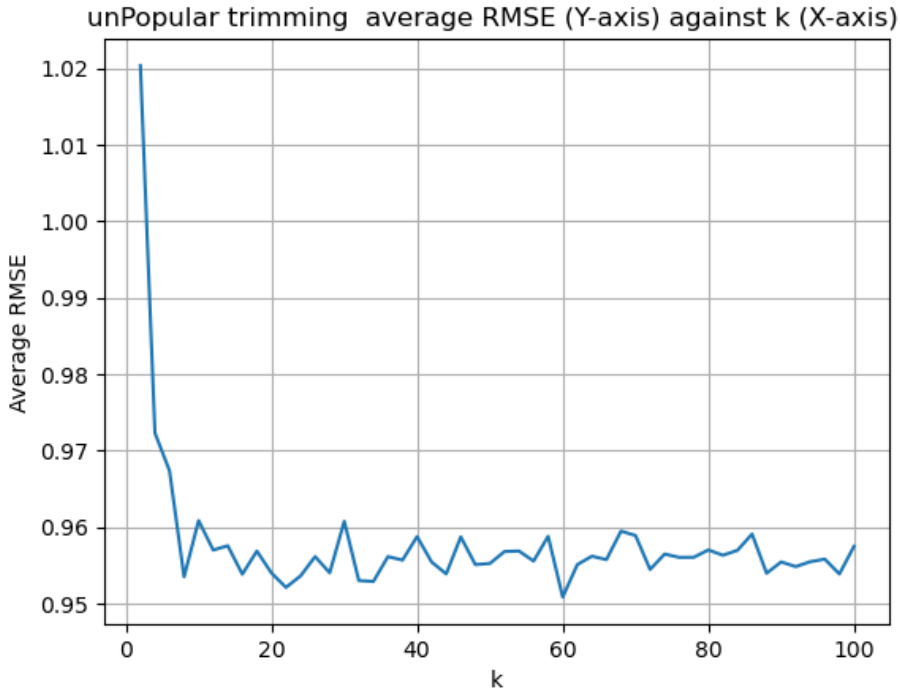
Q5:

The minimum k is 20, because when k is larger than 20, the average RMSE coverages to a steady-state value that is 0.86, and the average MAE coverages to a steady-state value that is 0.68.

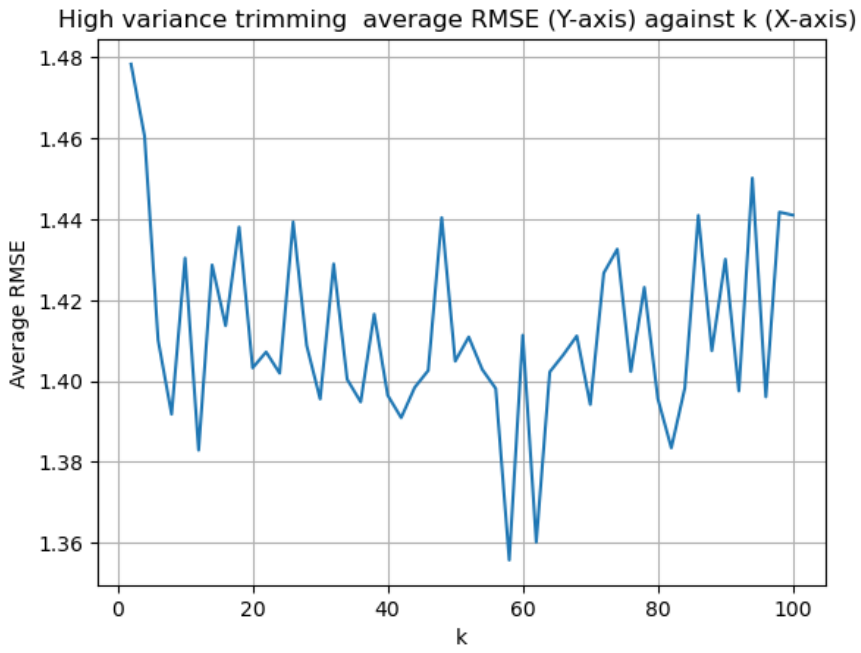
Q6:



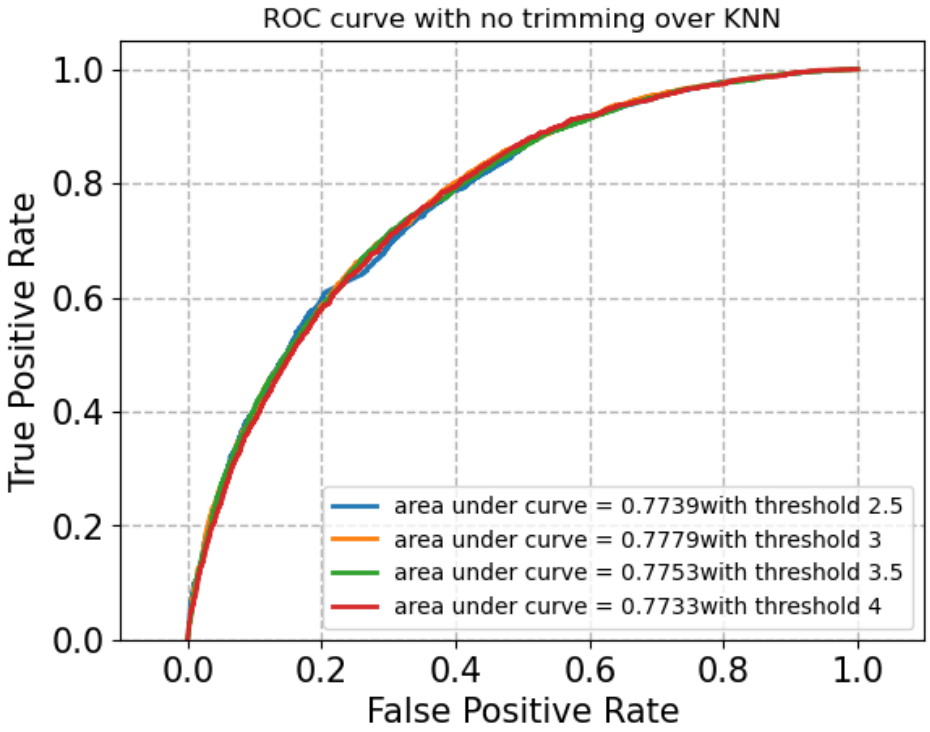
For popular movie trimming, the minimum average RMSE is 0.855370190450121.

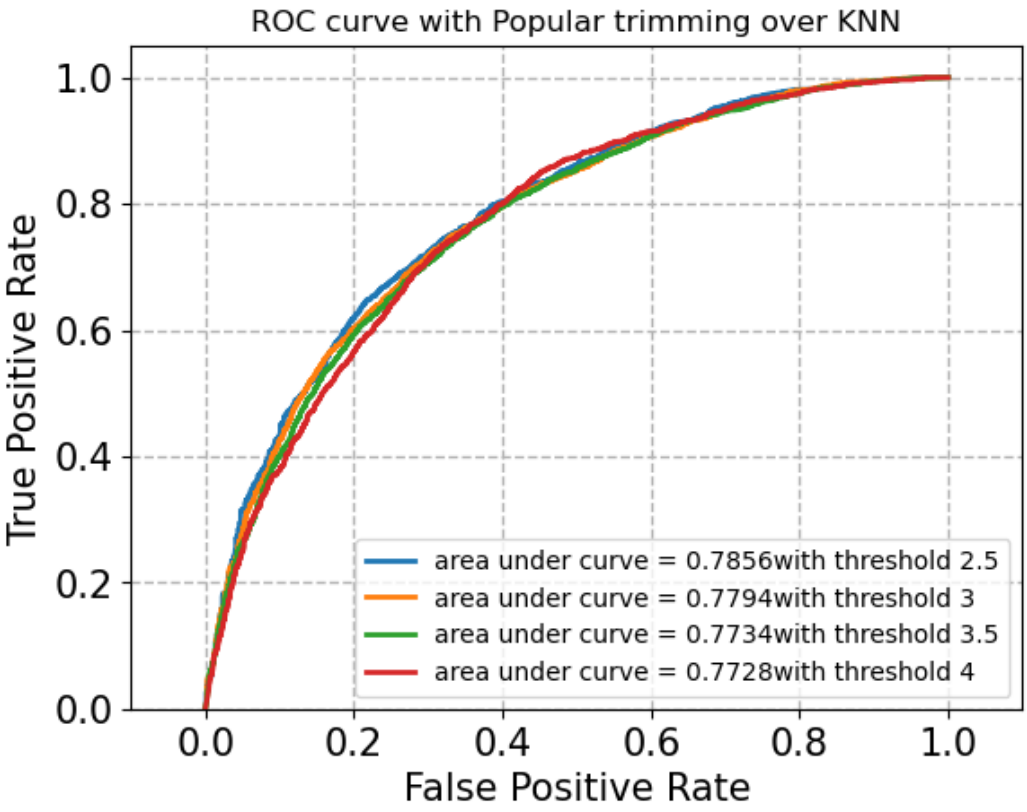


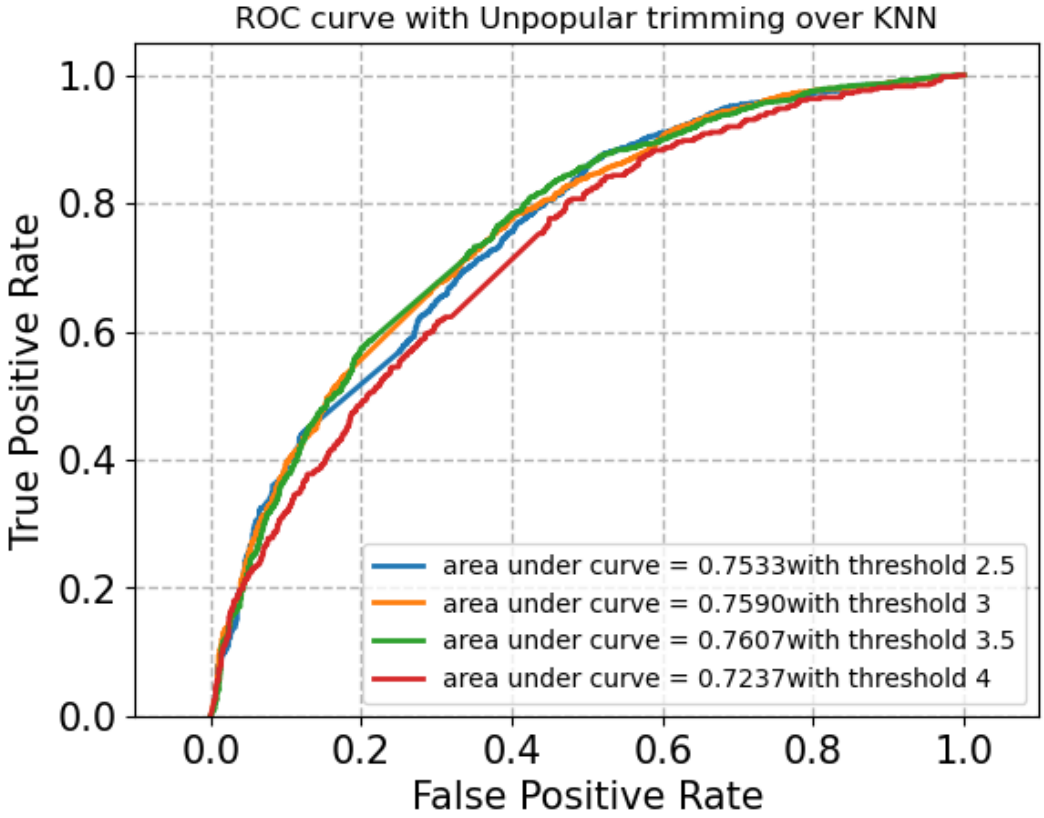
For unpopular movie trimming, the minimum average RMSE is 0.9509210109053875.

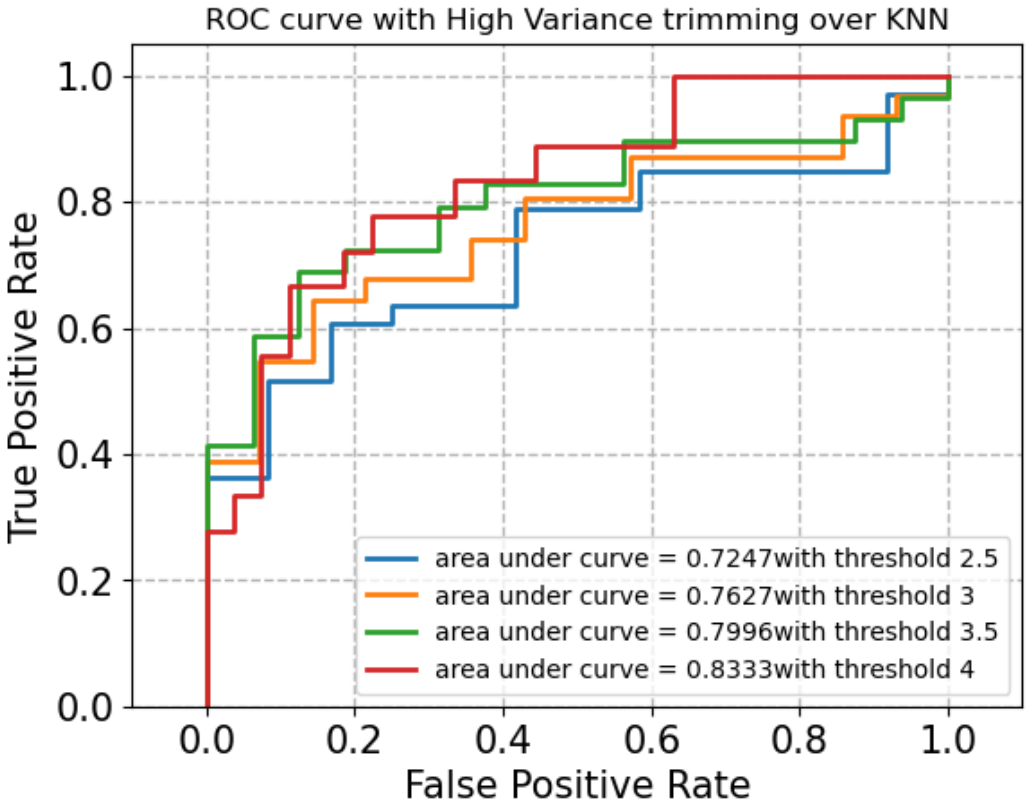


For High variance movie trimming, the minimum average RMSE is 0.1.3518077322166975
.




ROC curve with Popular trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7856with threshold 2.5
area under curve = 0.7794with threshold 3
area under curve = 0.7734with threshold 3.5
area under curve = 0.7728with threshold 4


ROC curve with Unpopular trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7533with threshold 2.5
area under curve = 0.7590with threshold 3
area under curve = 0.7607with threshold 3.5
area under curve = 0.7237with threshold 4


ROC curve with High Variance trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7247with threshold 2.5
area under curve = 0.7627with threshold 3
area under curve = 0.7996with threshold 3.5
area under curve = 0.8333with threshold 4

Q7:

To prove convex: If it convex, hessian matrix is positive semi-definite

$$\nabla^2 f(x) = \begin{bmatrix} 2wv^2 & -2w(r-2uv) \\ -2w(2vu-r) & 2wu^2 \end{bmatrix}$$

$$|\nabla^2 f(x)| = 4w^2(-3u^2v^2 - r^2 + 4uvr)$$

Determinant is not always positive, therefore the hessian matrix is not positive semi definite. The function doesn't convex.

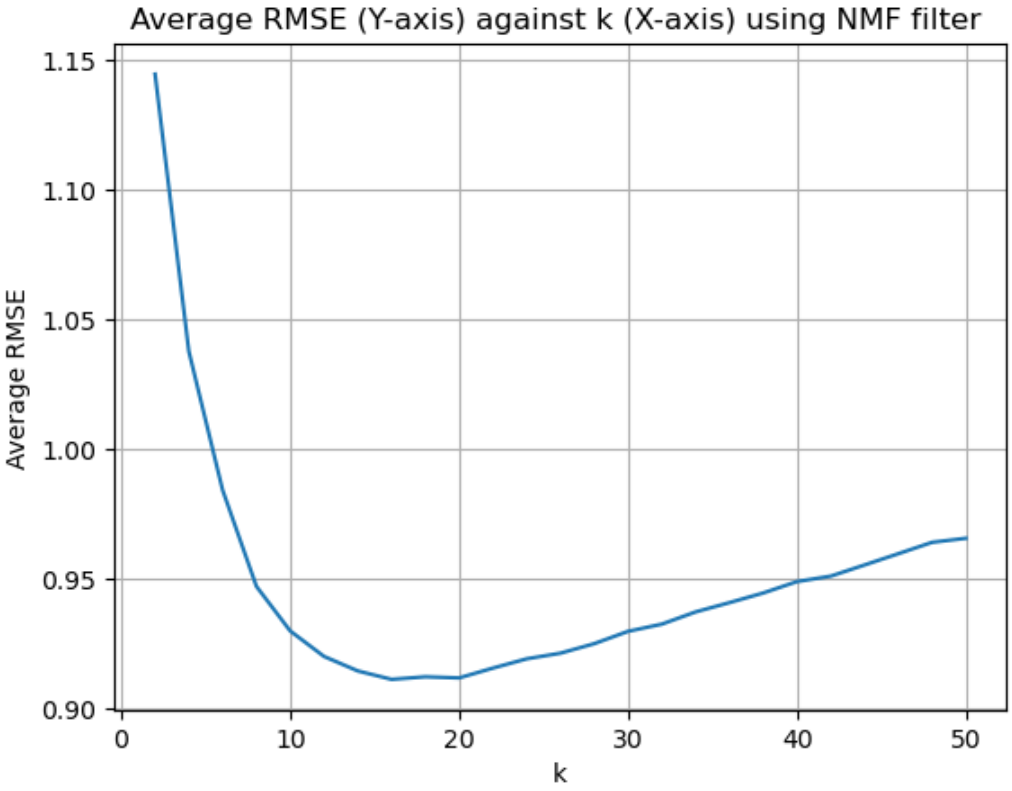
For U fixed, we can get:

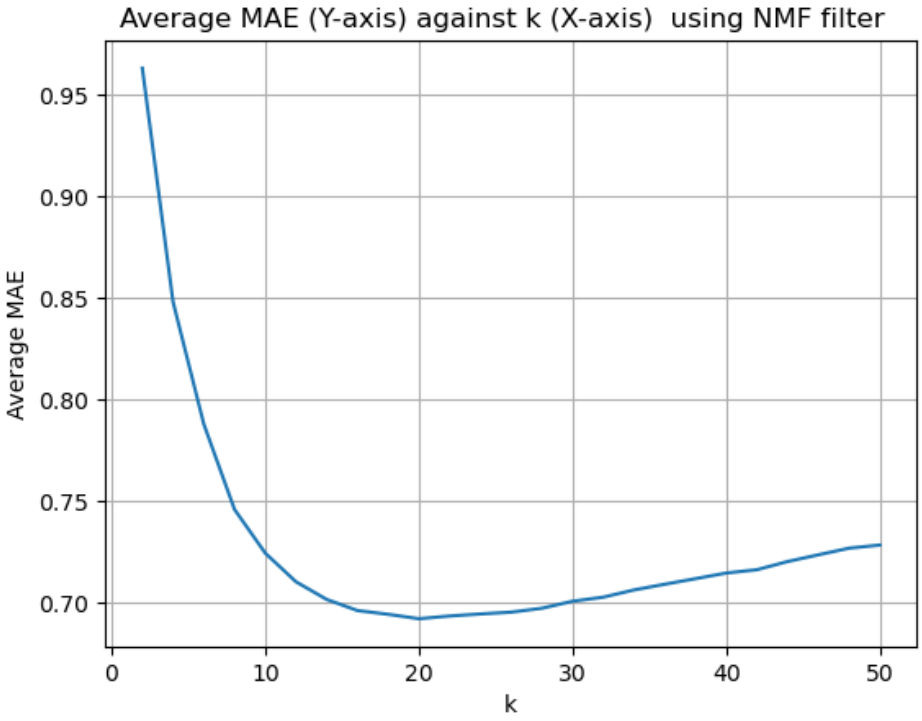
$$\underset{v}{\text{minimize}} \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij}(r_{ij} - (UV^T)_{ij})^2$$

where $V = (UU^T)^{-1}UR$ and R is rating matrix.

Q8:

Part A:


Average RMSE (Y-axis) against k (X-axis) using NMF filter
Average RMSE
1.15
1.10
1.05
1.00
0.95
0.90
0
10
20
30
40
50
k

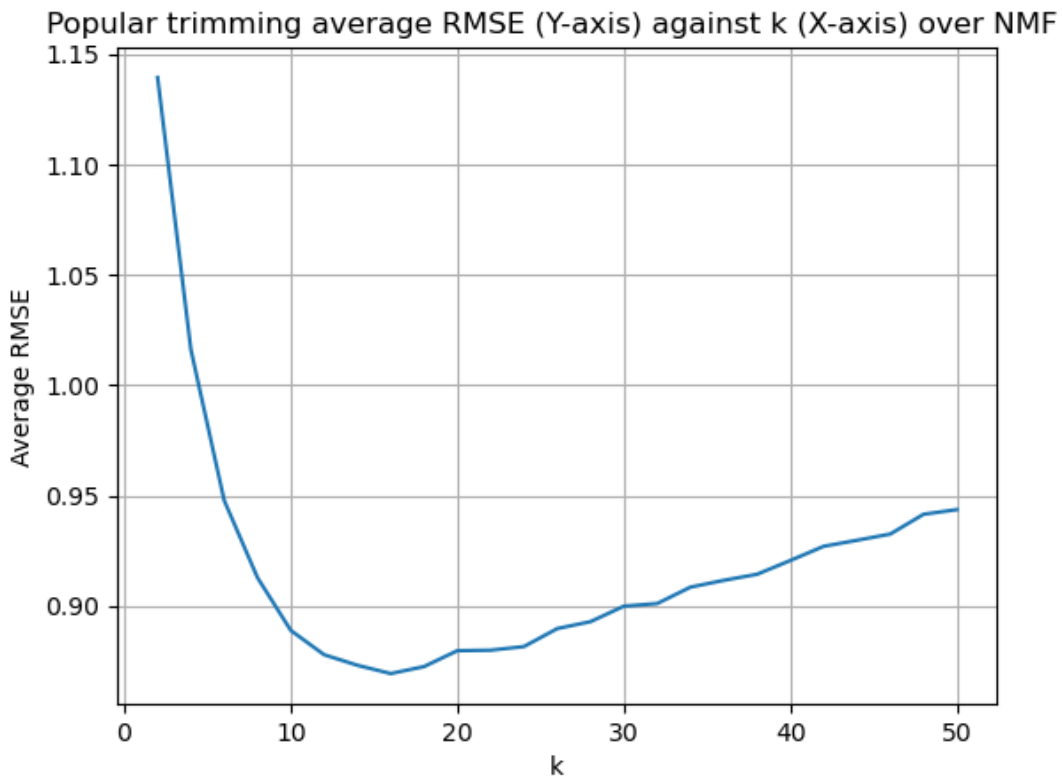

Average MAE (Y-axis) against k (X-axis) using NMF filter
Average MAE
0.95
0.90
0.85
0.80
0.75
0.70
0
10
20
30
40
50
k

Part B: Optimal number of latent factors is 20. The minimum average RMSE is 0.9113505309168204. The minimum average MAE is 0.6923977751493282. The optimal number of latent factor is the same as the number of movie genres according to the code below:

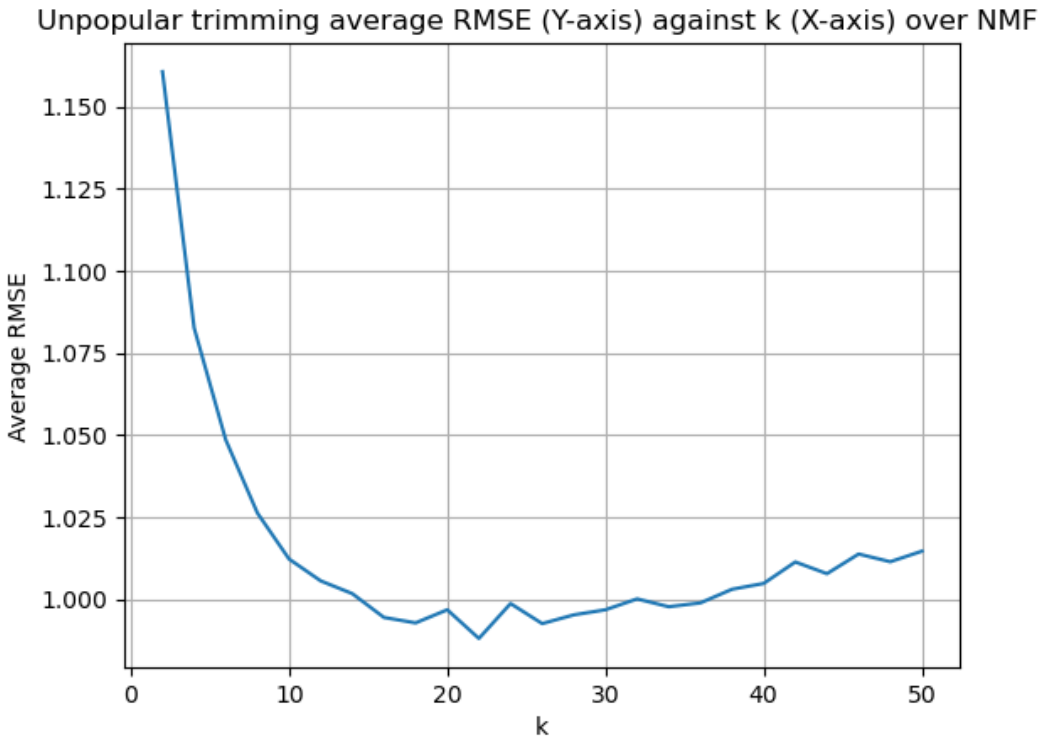
```
|: df_movie = pd.read_csv('movies.csv');

genres = {}
count = 0
for g in df_movie['genres']:
    for j in g.split('|'):
        if j not in genres:
            genres[j] = 1
            count +=1
count
```

```
|: 20
```

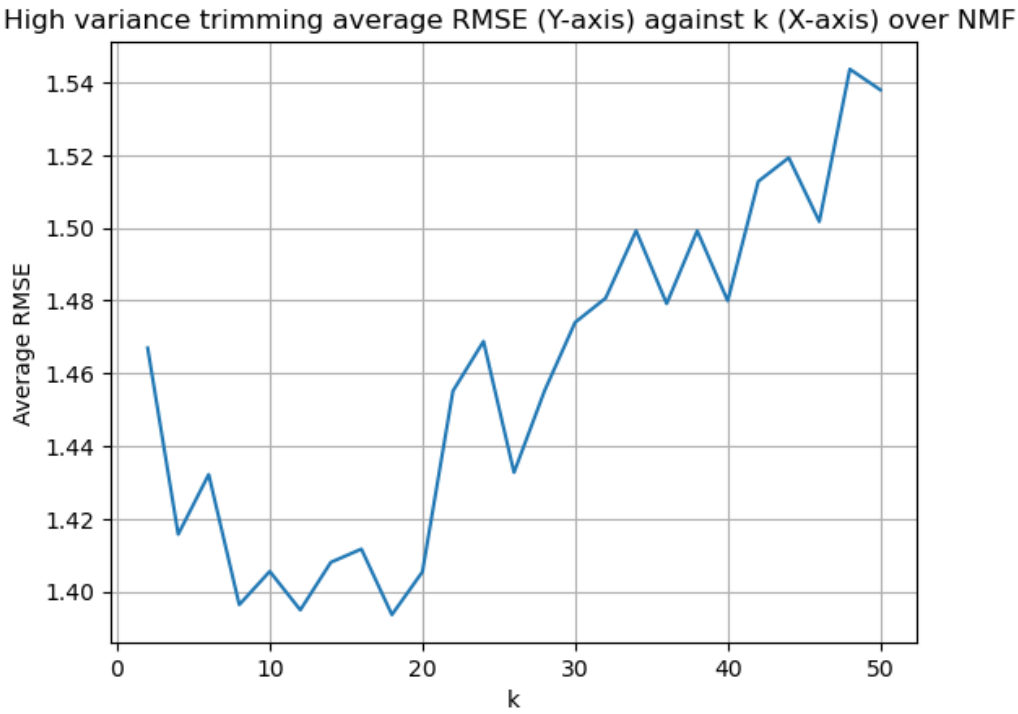
Part C:



The minimum average RMSE with popular trimming is 0.8715782673973596.

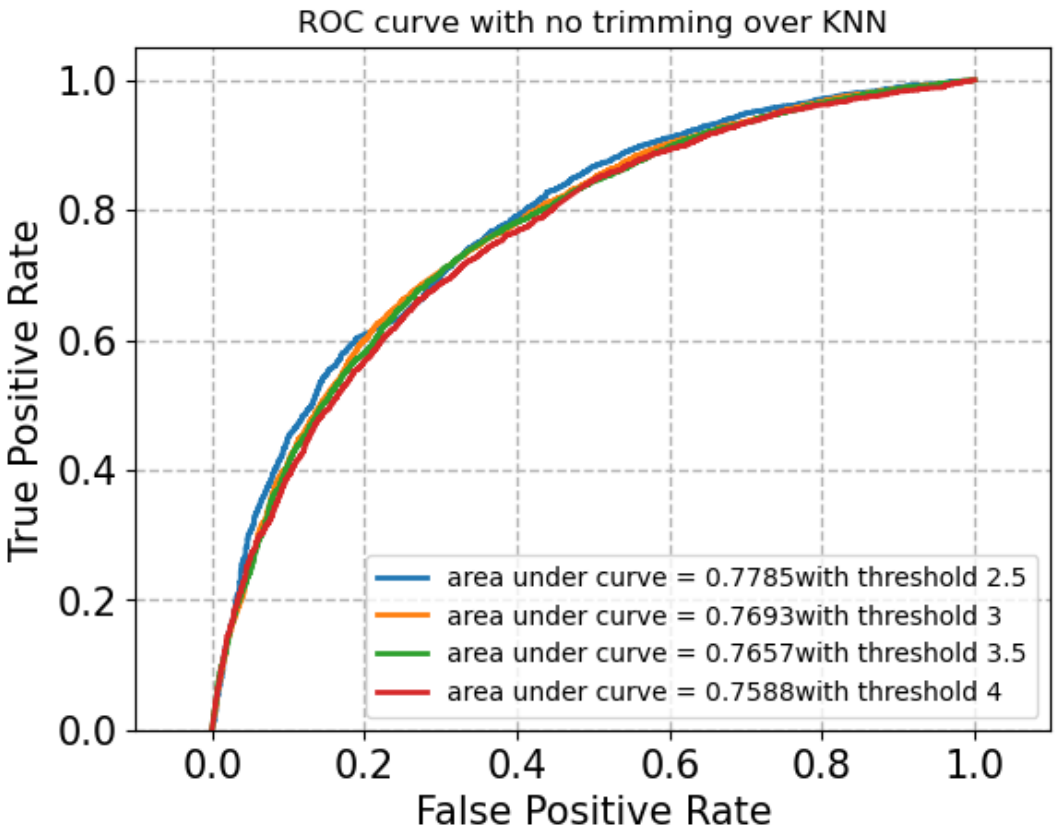


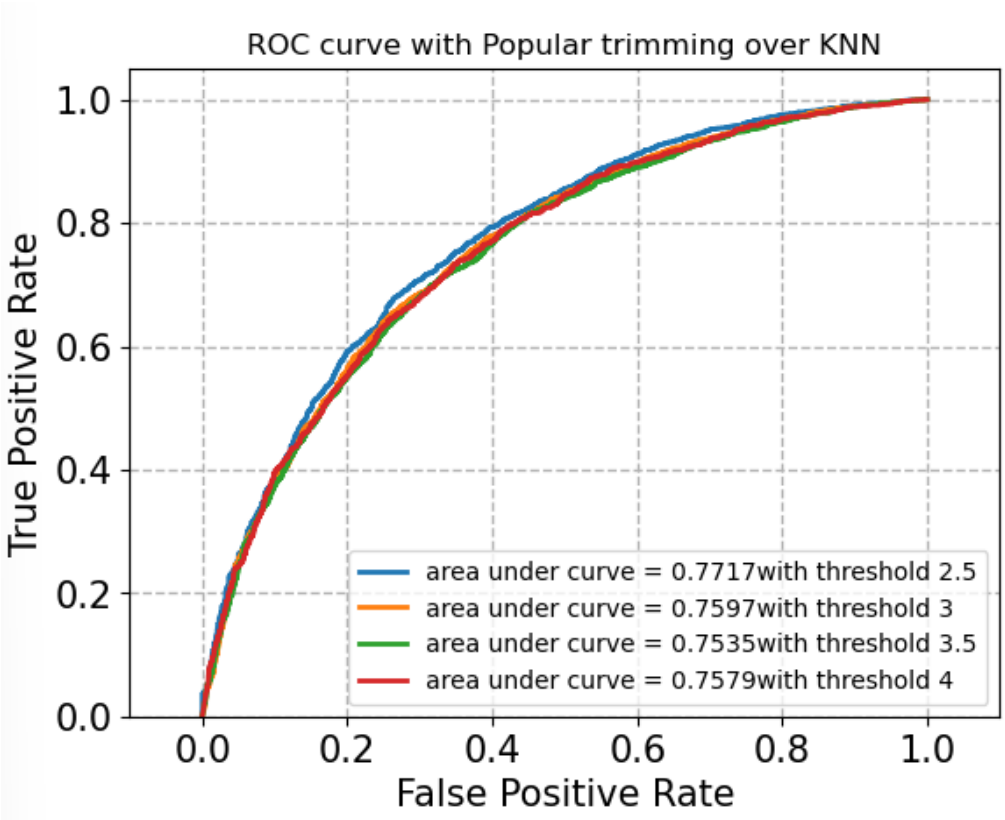
The minimum average RMSE with unpopular trimming is 0.9884440771422953..

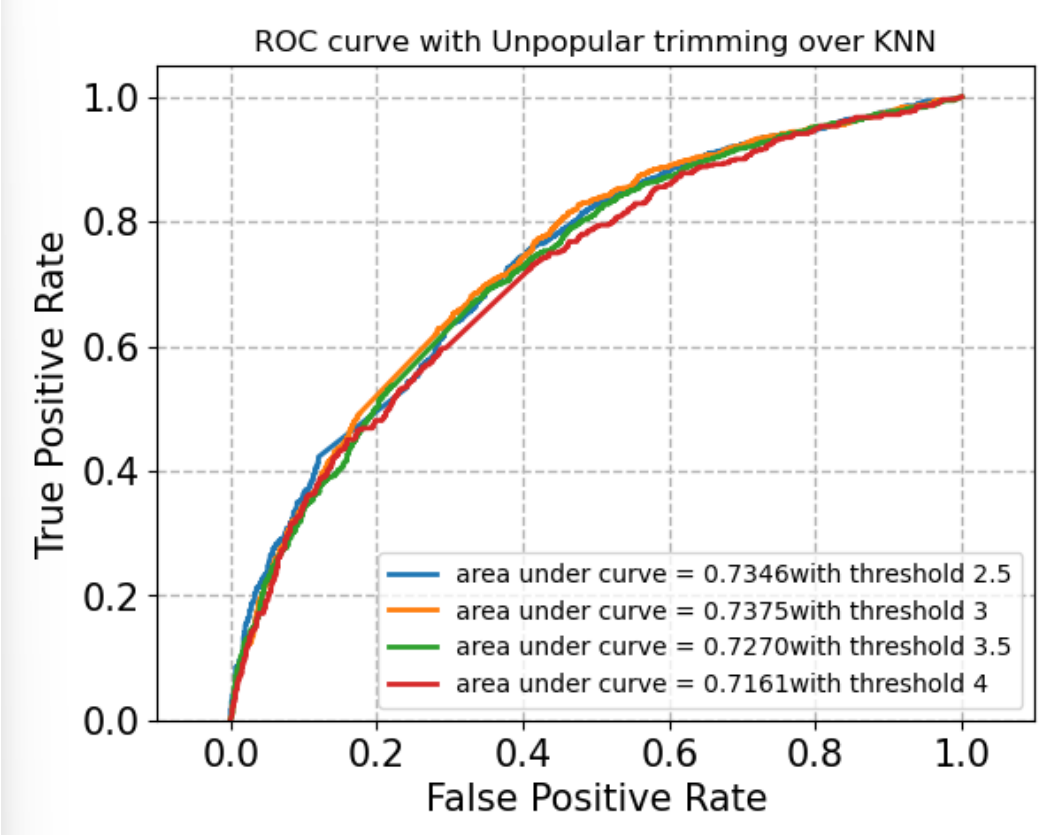


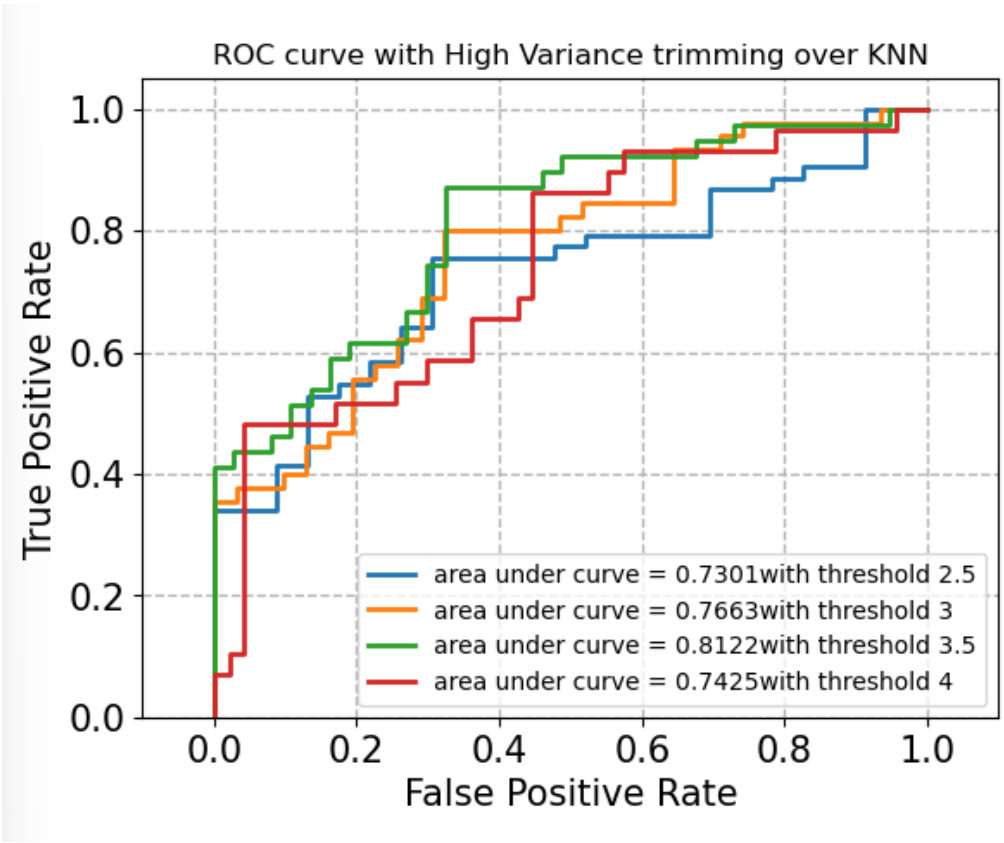
The minimum average RMSE with high variance trimming is 1.393254383683149

.


ROC curve with no trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7785with threshold 2.5
area under curve = 0.7693with threshold 3
area under curve = 0.7657with threshold 3.5
area under curve = 0.7588with threshold 4


ROC curve with Popular trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7717with threshold 2.5
area under curve = 0.7597with threshold 3
area under curve = 0.7535with threshold 3.5
area under curve = 0.7579with threshold 4


ROC curve with Unpopular trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7346with threshold 2.5
area under curve = 0.7375with threshold 3
area under curve = 0.7270with threshold 3.5
area under curve = 0.7161with threshold 4


ROC curve with High Variance trimming over KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
area under curve = 0.7301with threshold 2.5
area under curve = 0.7663with threshold 3
area under curve = 0.8122with threshold 3.5
area under curve = 0.7425with threshold 4

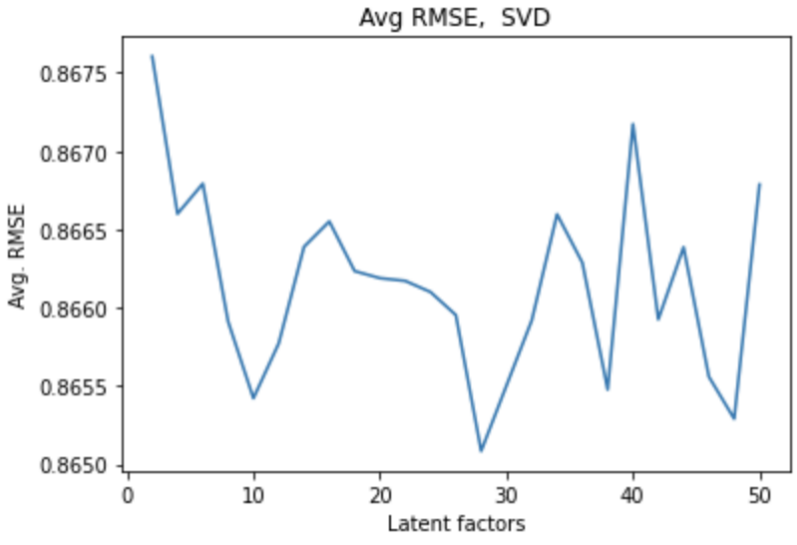
Q9:

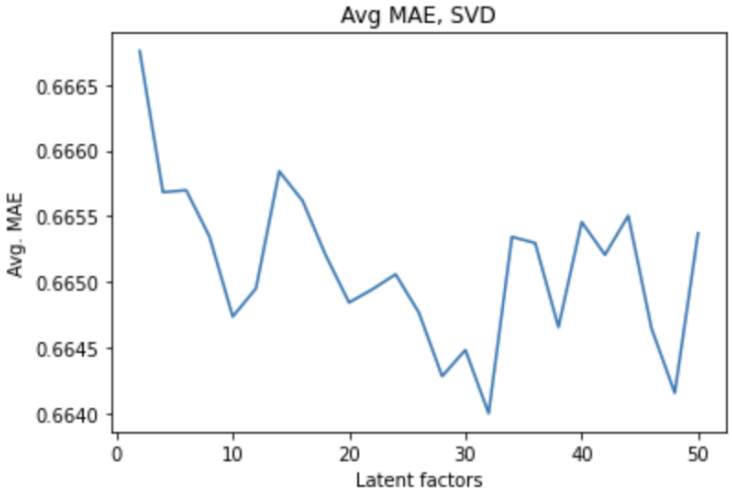
```
V number:  0
Comedy
Horror|Thriller
Comedy
Action|Adventure|Romance
Comedy|Sci-Fi
Documentary
Drama
Adventure|Drama|Romance
Drama
Drama|Romance
V number:  1
Horror|Thriller
Horror|Sci-Fi
Comedy
Horror
Drama
Adventure|Drama|War
Drama
Comedy
Comedy|Crime|Drama
Comedy|Drama|Film-Noir
V number:  2
Adventure|Children|Fantasy
Crime|Mystery
Comedy|Drama
Animation|Comedy|Fantasy|Musical
Comedy|Musical
Action|Adventure|Sci-Fi|Thriller
Children|Drama
Drama
Drama|Musical|Romance
Documentary|IMAX
V number:  3
Comedy|Musical
Comedy|Drama
Comedy|Drama|Romance
Horror
Action|Adventure|Drama|Fantasy
Comedy|Musical|Romance
Drama|Romance
Drama|Romance
Horror
Action|Adventure|Animation|Children|Comedy|Fantasy
V number:  4
Horror|Thriller
Action|Comedy
Drama|War
Crime|Thriller
Animation|Children
Action|Crime|Drama|Mystery|Thriller
Action|Drama|War
Crime|Drama|Thriller
Adventure|Animation|Fantasy
Drama
V number:  5
Horror|Thriller
Drama|Mystery
Crime|Drama|Thriller
Comedy|Drama
Action|Fantasy|Horror|Sci-Fi|Thriller
Drama|Romance|War
Action|Sci-Fi|Thriller
Comedy|Romance
Horror|Mystery|Thriller
Comedy|Drama|Mystery|Thriller
```

As we can see from the result, mostly, the top 10 genres are comedy, crim and Horror. It is a certain genres of a small collection. The connection between latent factors and movie genres is that latent factors represent a subset of movie genres, and movie genres can be in different latent factors.

Q10:

A:



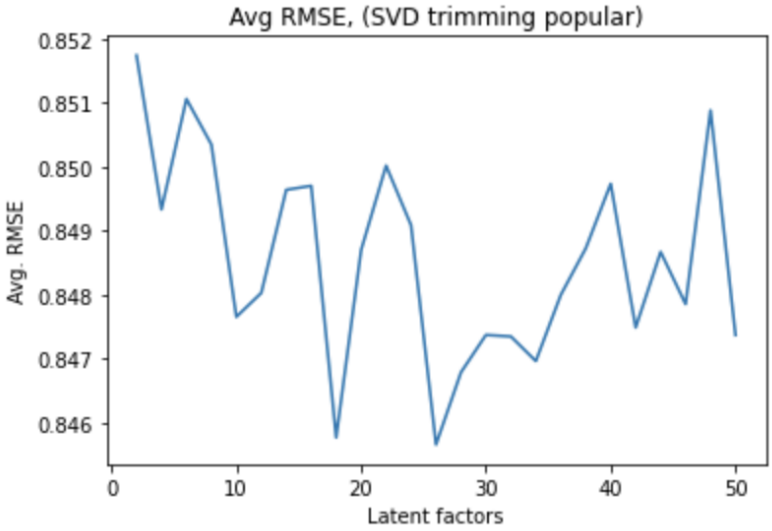


B:

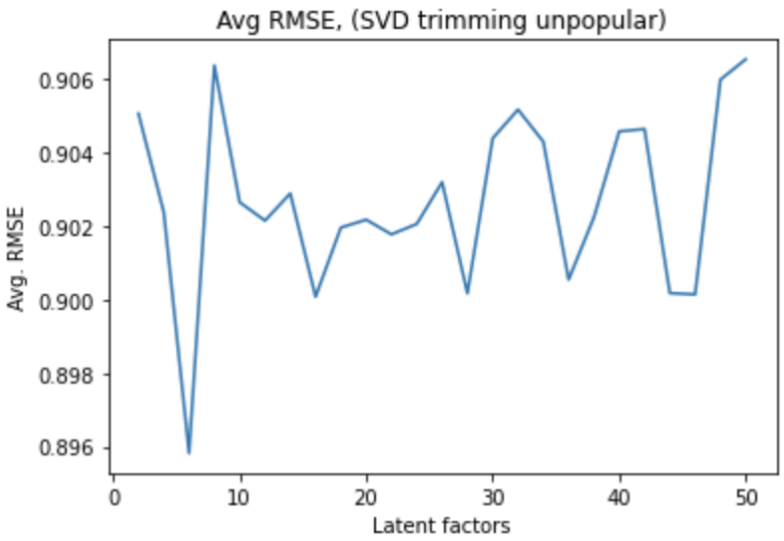
Minimum RMSE is: 0.865085809011833, and k is: 28
Minimum MAE is: 0.6639978173804683, and k is: 32

We can see that the optimal number of latent factors is not the same as the number of moive genres.

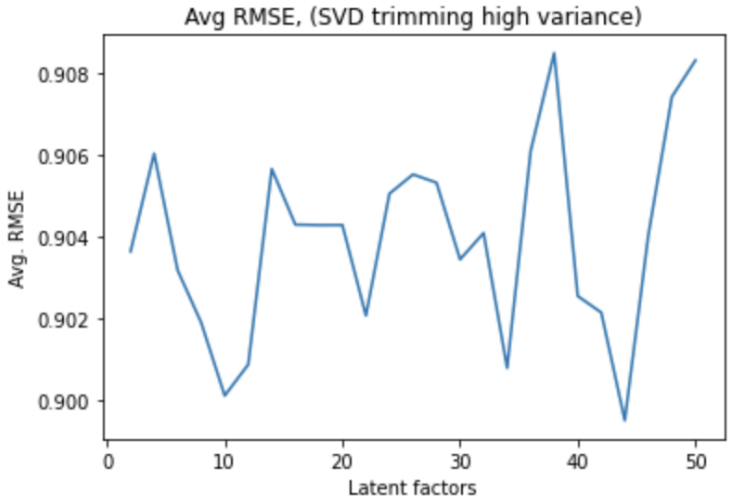
C:



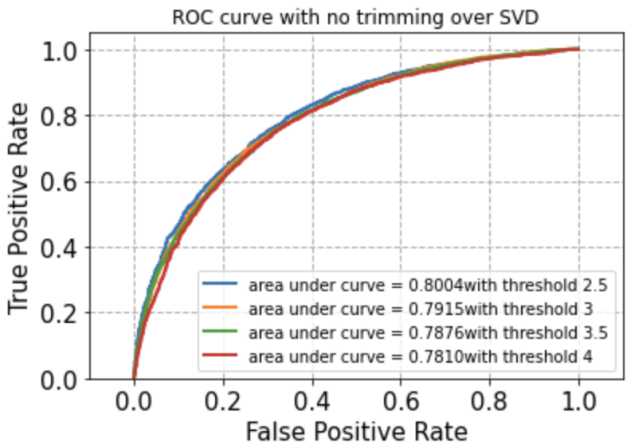
Minimum Average RMSE (trimming popular)is: 0.846637108059328, and k is: 26

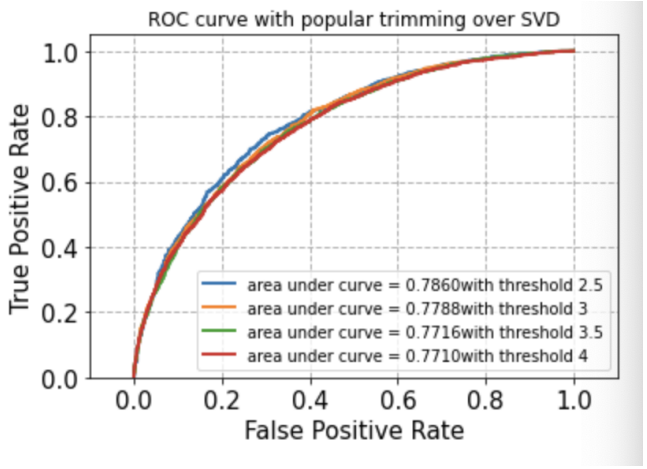


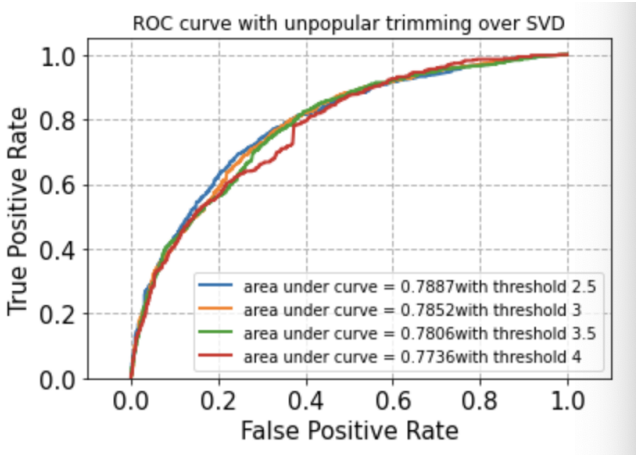
Minimum Average RMSE (trimming unpopular)is: 0.8958215963751085, and k is: 6

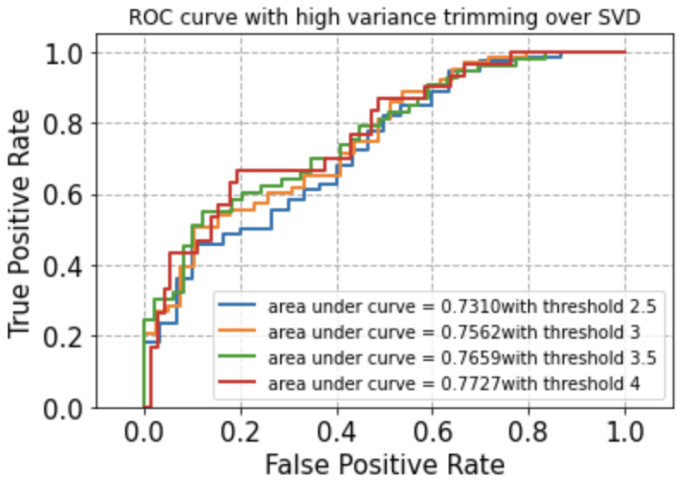


Minimum Average RMSE (trimming high variance)is: 0.8995284461687231, and k is: 44


ROC curve with no trimming over SVD
True Positive Rate
False Positive Rate
area under curve = 0.8004with threshold 2.5
area under curve = 0.7915with threshold 3
area under curve = 0.7876with threshold 3.5
area under curve = 0.7810with threshold 4


ROC curve with popular trimming over SVD
True Positive Rate
False Positive Rate
area under curve = 0.7860with threshold 2.5
area under curve = 0.7788with threshold 3
area under curve = 0.7716with threshold 3.5
area under curve = 0.7710with threshold 4


ROC curve with unpopular trimming over SVD
True Positive Rate
False Positive Rate
area under curve = 0.7887with threshold 2.5
area under curve = 0.7852with threshold 3
area under curve = 0.7806with threshold 3.5
area under curve = 0.7736with threshold 4

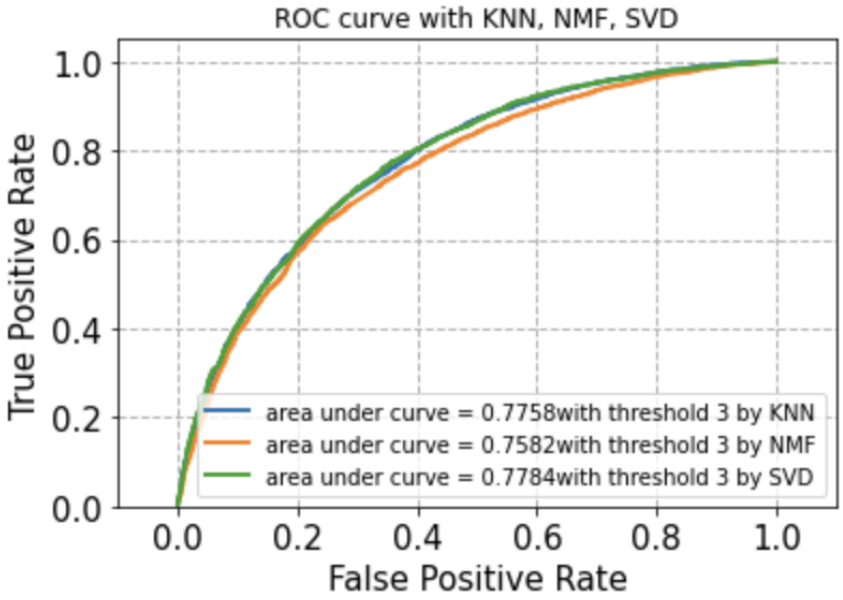


Q11:

The average RMSE across all 10 folds is 0.9347148205412456

The average RMSE across all 10 folds with popular, unpopular, high variance trimming are 0.923298422153475, 0.9548885903625072, 1.451157476431985.

Q12:

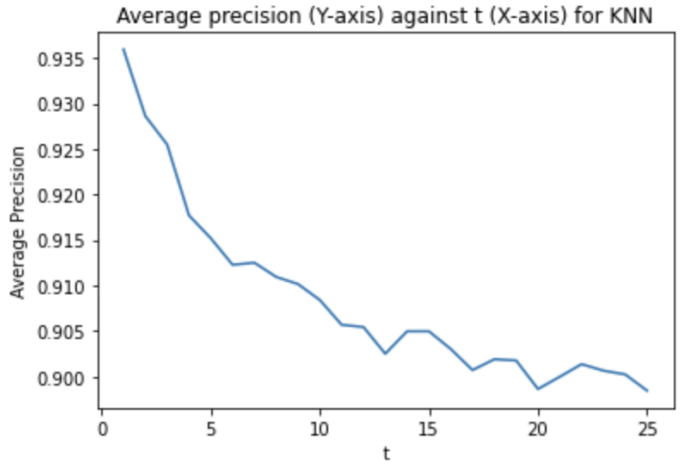


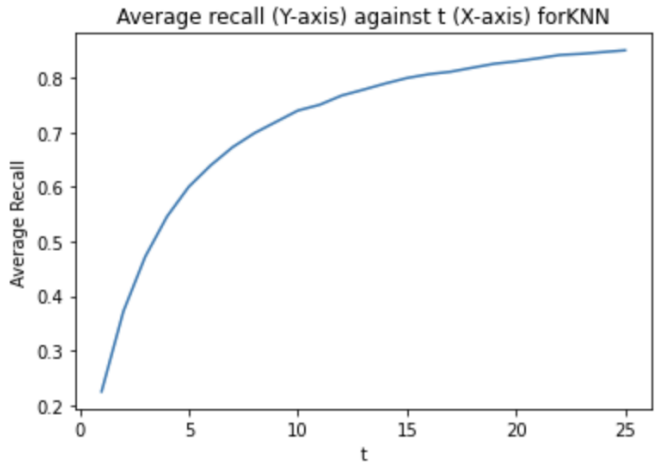
Q13:

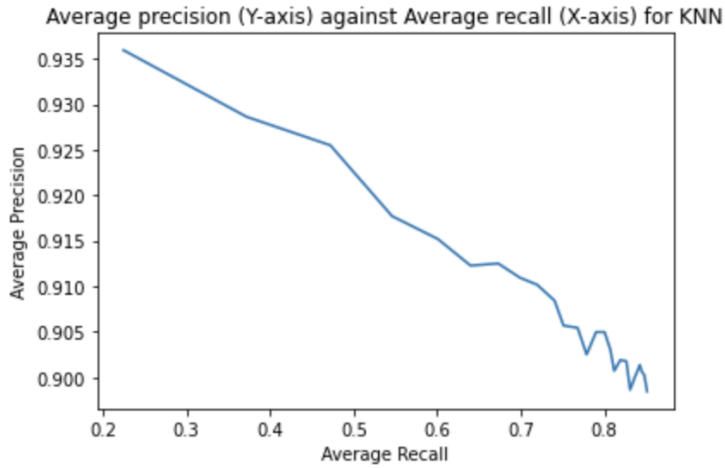
Precision is the probability of true positives in all predictions, and the Recall is the probability of true positives in all relevant items. In this case, the precision indicates the probability of recommended movies that users like in all recommended movies, and the recall shows the probability of recommended movies that users like in all items liked by the users.

Q14:

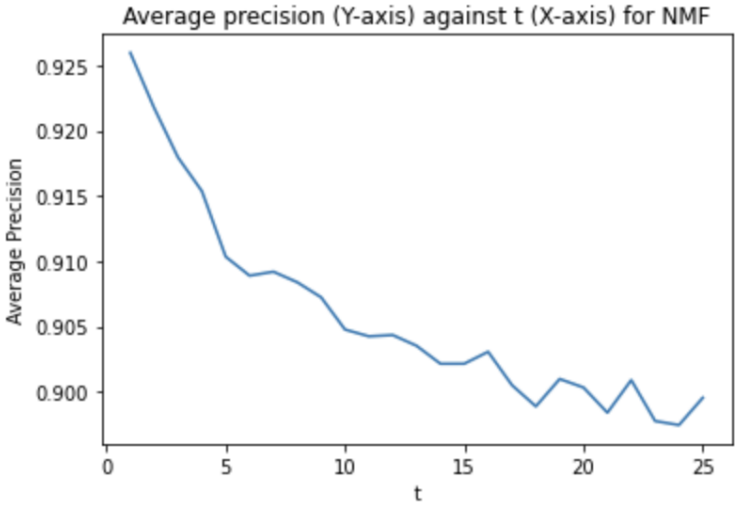
This is the KNN average precision, average recall against t, and average precision against average recall plots

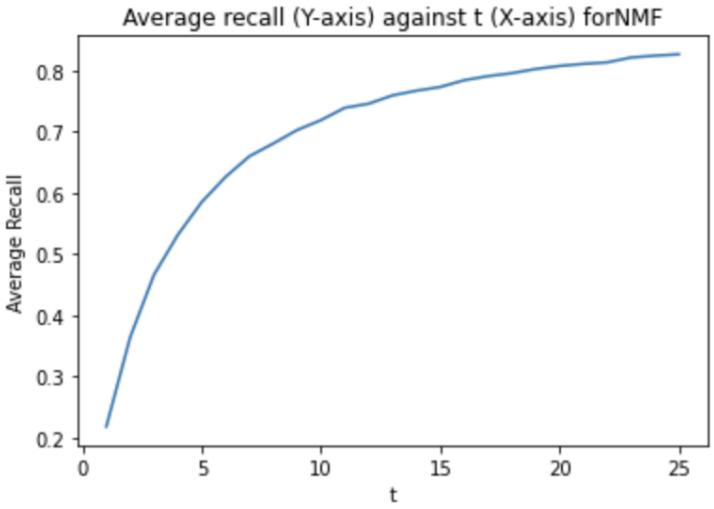


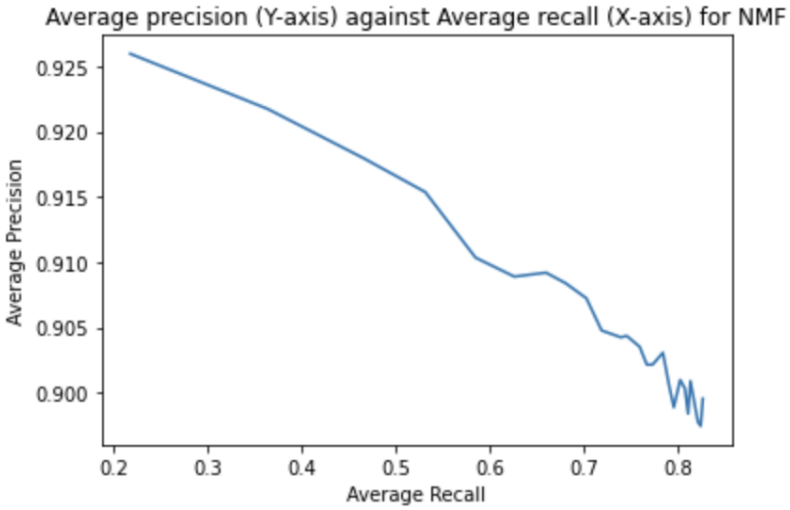




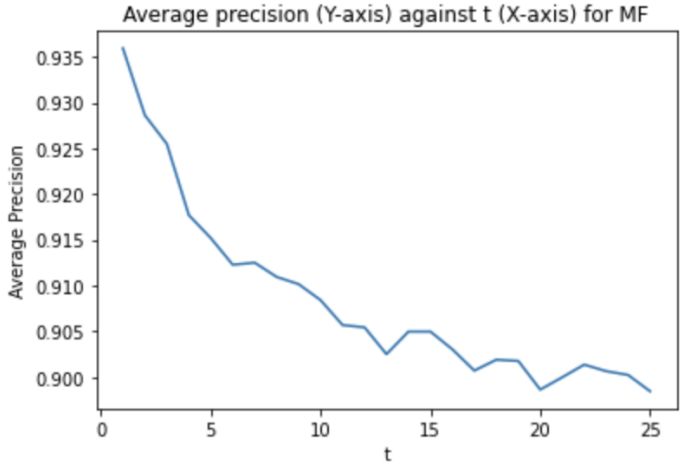
This is the NMF average precision, average recall against t, and average precision against average recall plots

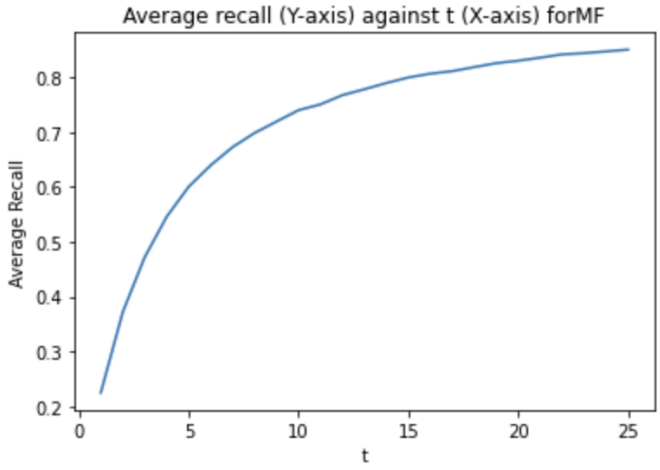


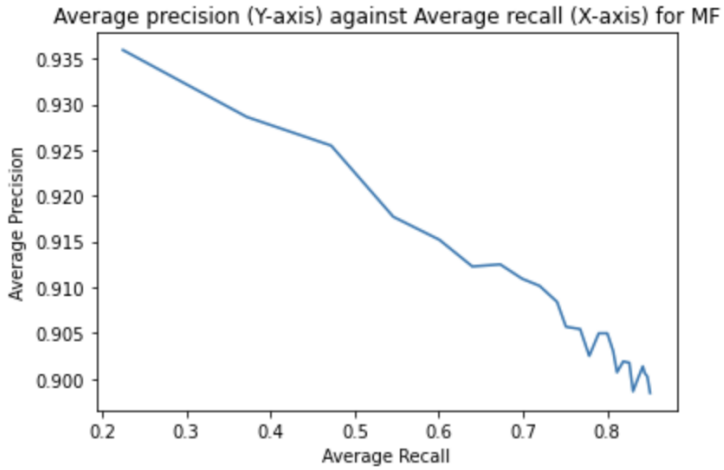


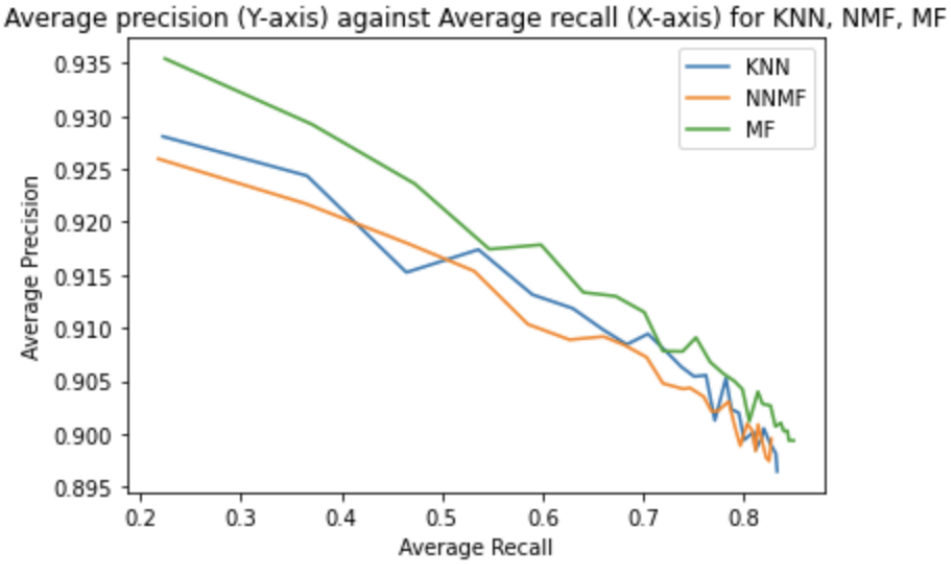


This is the MF average precision, average recall against t, and average precision against average recall plots.









From the plot average precision against average recall for KNN, NMF, MF, we can see that MF performs the best overall since the green line is always above the other two.