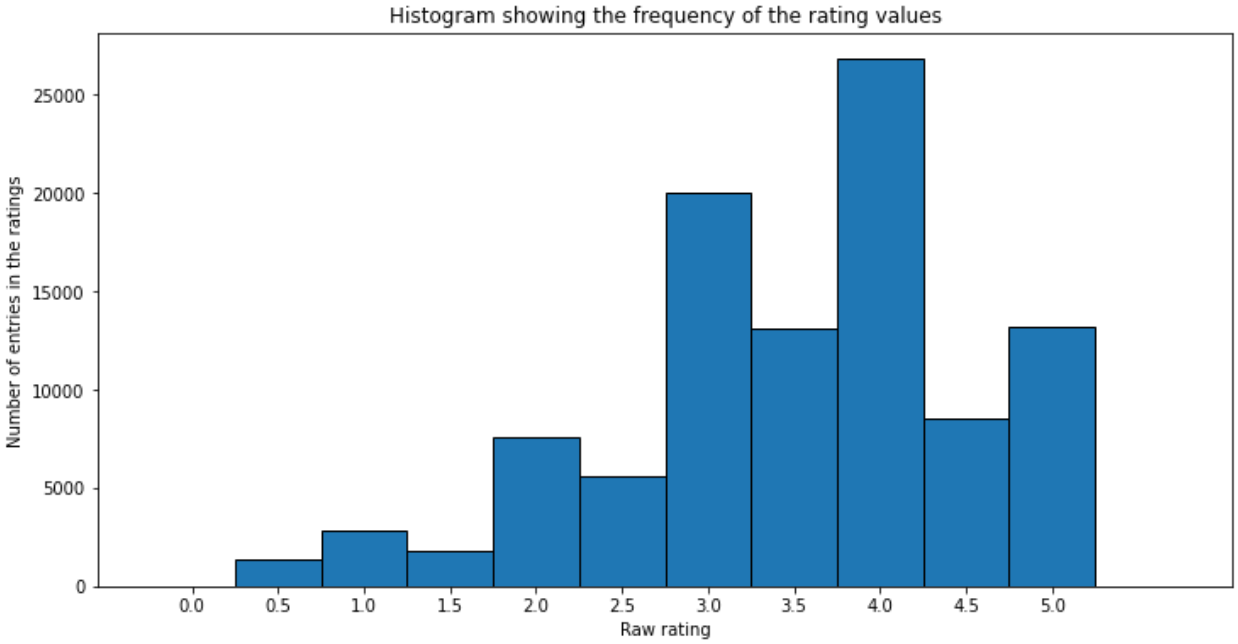
ECE 219 Project 03

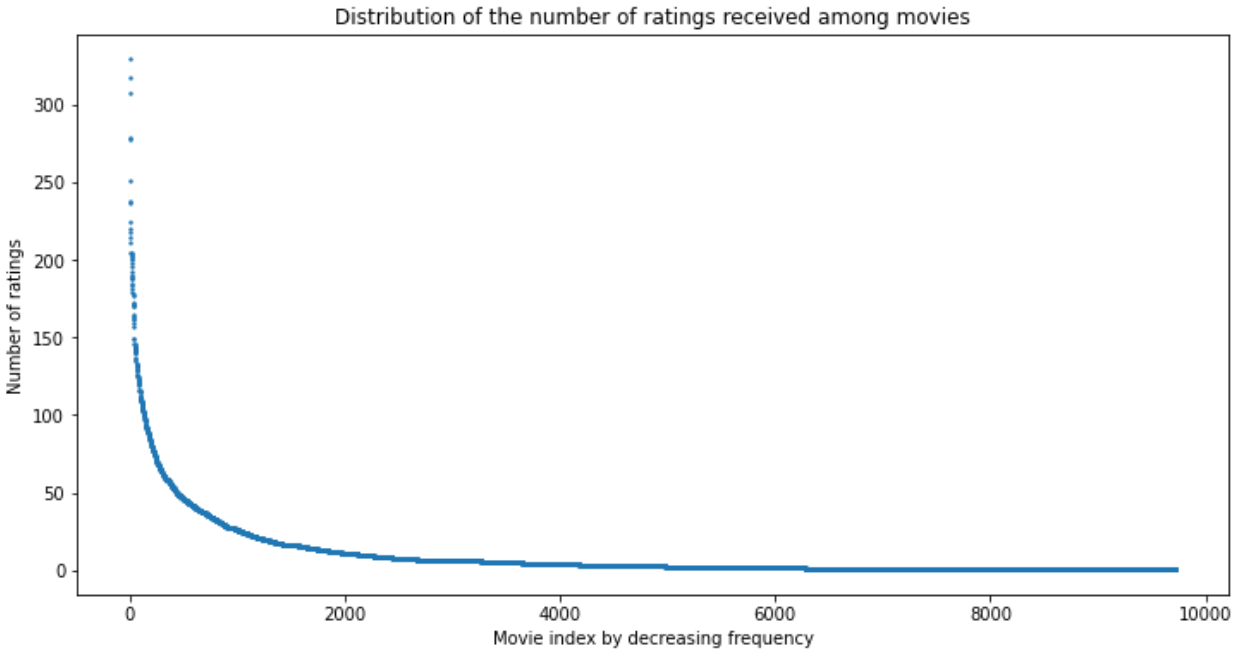
Question 1

A) $Sparsity = \frac{Total\ number\ of\ available\ ratings}{Total\ number\ of\ possible\ ratings} = \frac{length\ of\ rating\ dataframe}{num\ of\ users \times num\ of\ movies} = 0.0169997$

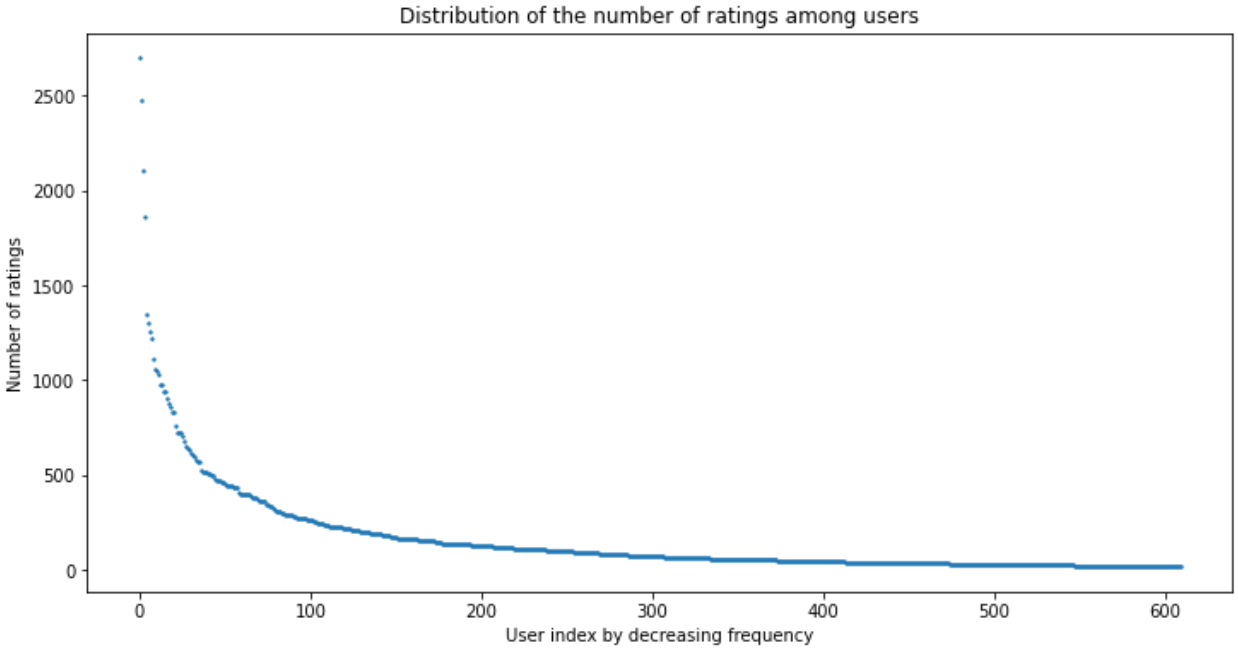
B) The diagram is as below, we can see that the raw ratings are distributed around the middle and can be seen as a bell curve shifted slightly to the right although the middle 3.5 rating is less than 3.0 and 4.0. From the curve, we can see that users tend to give a movie a moderate to high rating and seldom give extremely low ratings, and the minimum rating is 0.5.



C)

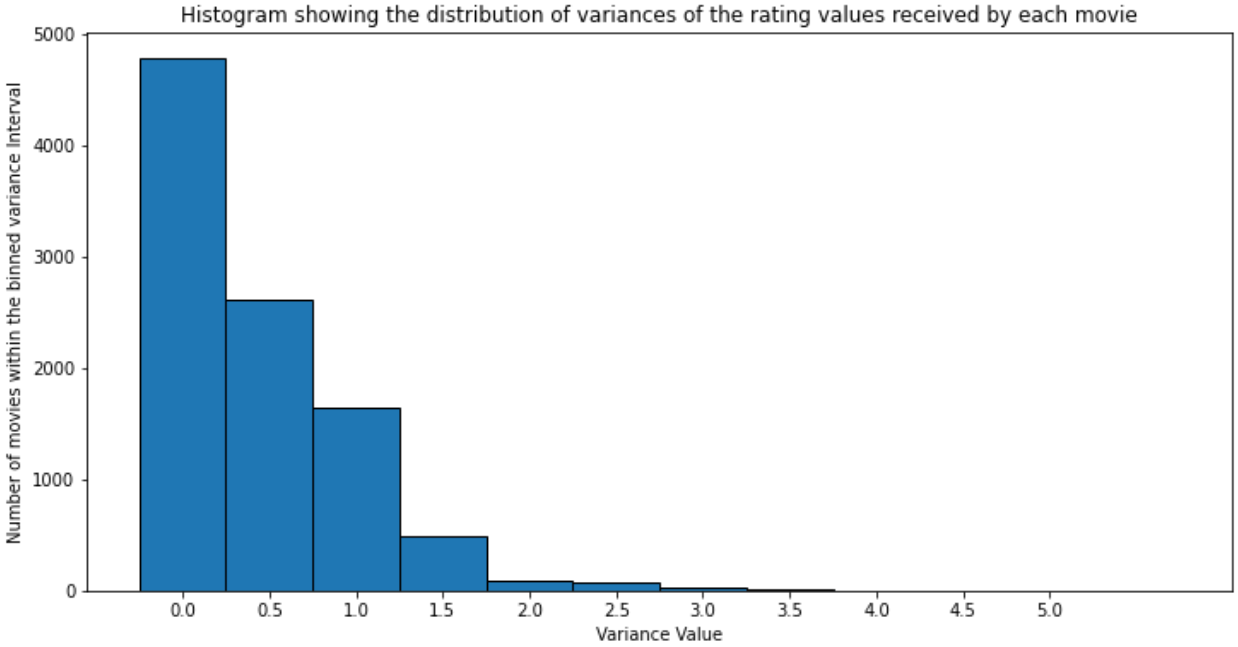

Distribution of the number of ratings received among movies
Number of ratings
0
50
100
150
200
250
300
0
2000
4000
6000
8000
10000
Movie index by decreasing frequency

D)


Distribution of the number of ratings among users
Number of ratings
0
500
1000
1500
2000
2500
0
100
200
300
400
500
600
User index by decreasing frequency

E) The plot in part C indicates that most movies receive less than 50 ratings in the dataset, but some movies still receive significantly more ratings than others. The curve is similar to a logistic curve. Movies with a more significant number of ratings are more likely to be accurately recommended than movies with fewer ratings. Similarly, the plot we get in part D indicates that most users give ratings to less than 500 different movies in the dataset, but some users still give significantly more ratings to films. Users who rate a significant number of movies are more likely to receive accurate recommendations than those who rate fewer movies. The recommendation system should be able to process with less than 50 ratings per film and 500 ratings per user.

F) The histogram shows that the movie's variance values keep within a small range. Most movies have variance values smaller than 1.5. This indicates that the majority of movies receive relatively consistent ratings among users. Uses tend to have similar overall tastes in films.



Question 2

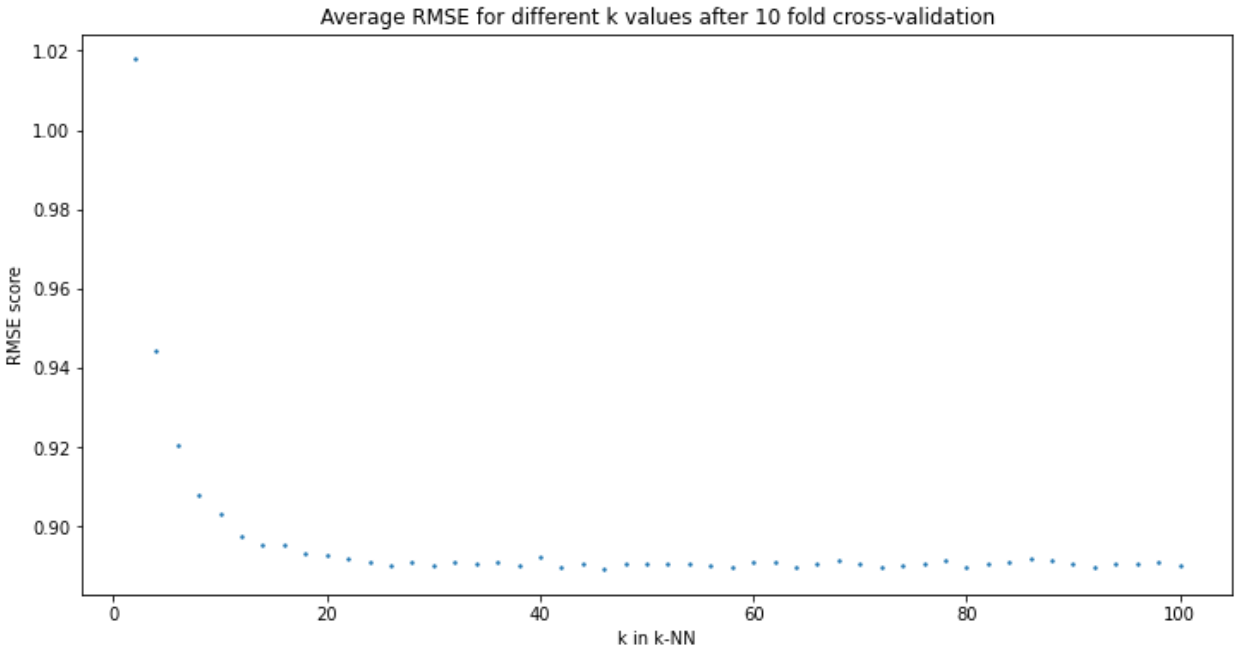
A) $\mu_u = \frac{1}{||I_u||} \times \sum_{k \in I_u} r_{uk}$

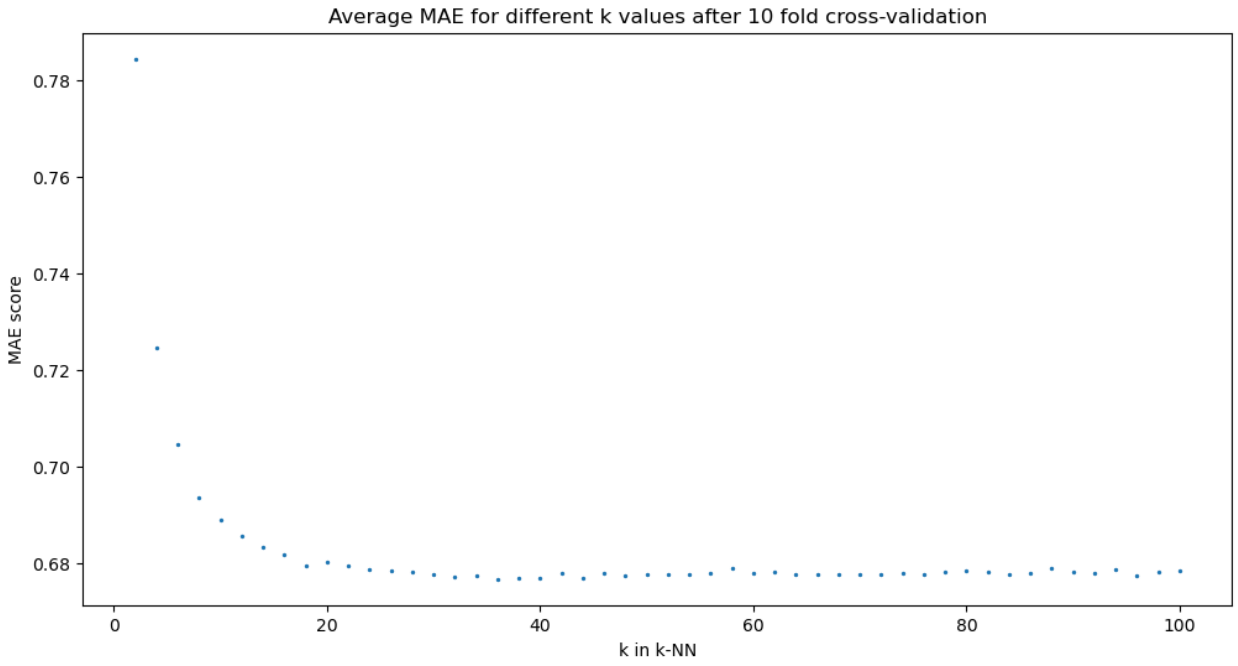
B) $I_u \cap I_v$ is the set of items indices for which ratings have been specified by both users u and v. Since the rating matrix R is sparse, many items are unrated or are only rated by a single person. Thus, it is possible to have $I_u \cap I_v = \emptyset$.

Question 3

Mean-centering the raw ratings can eliminate user bias and improve the accuracy of results. If a user's rating is significantly higher than others, their rating will lift up the average rating of a movie, making it more likely to be recommended. Mean centering will bring the rating down. If the rating is lower than the average, similarly, the movie will be less likely to be recommended. Mean centering will bring up the rating. It helps to keep ratings from different users on a more relative and consistent scale and makes the recommendation of movies more accurate.

Question 4:


Average RMSE for different k values after 10 fold cross-validation
RMSE score
1.02
1.00
0.98
0.96
0.94
0.92
0.90
0
20
40
60
80
100
k in k-NN

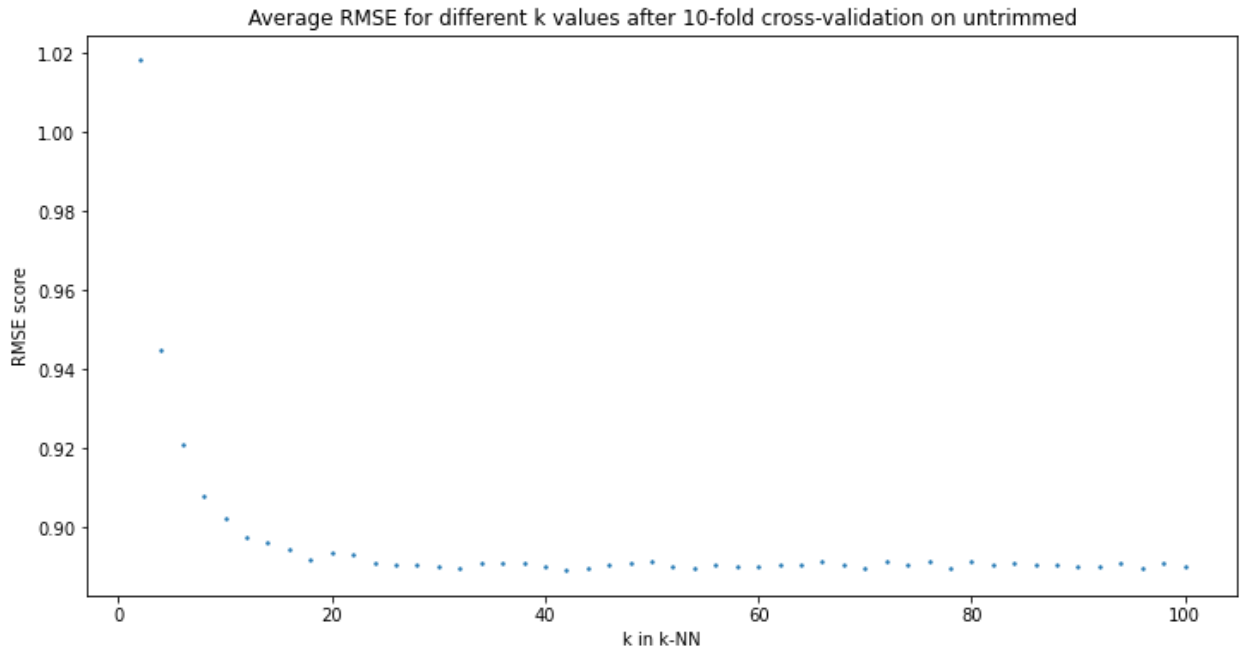

Average MAE for different k values after 10 fold cross-validation
MAE score
0.78
0.76
0.74
0.72
0.70
0.68
0
20
40
60
80
100
k in k-NN

Question 5

The minimum k value is 20. The steady-state average RMSE is approximately 0.89, and the steady-state average MAE is approximately 0.68.

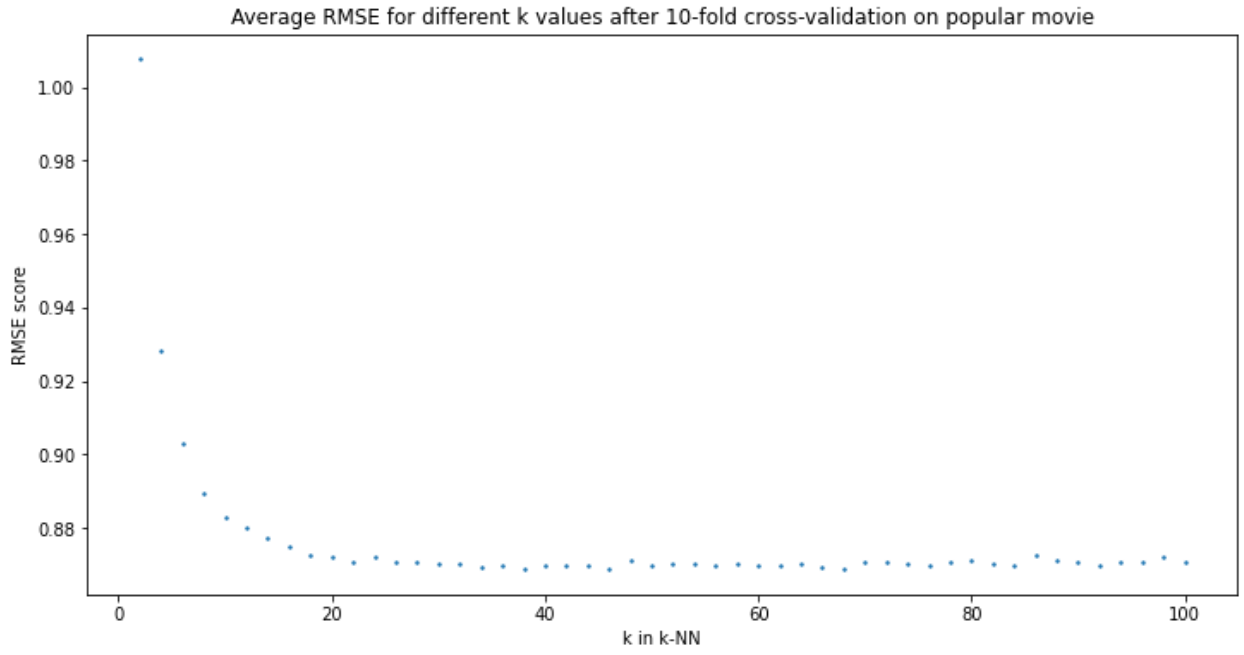
Question 6

For the untrimmed movie dataset:



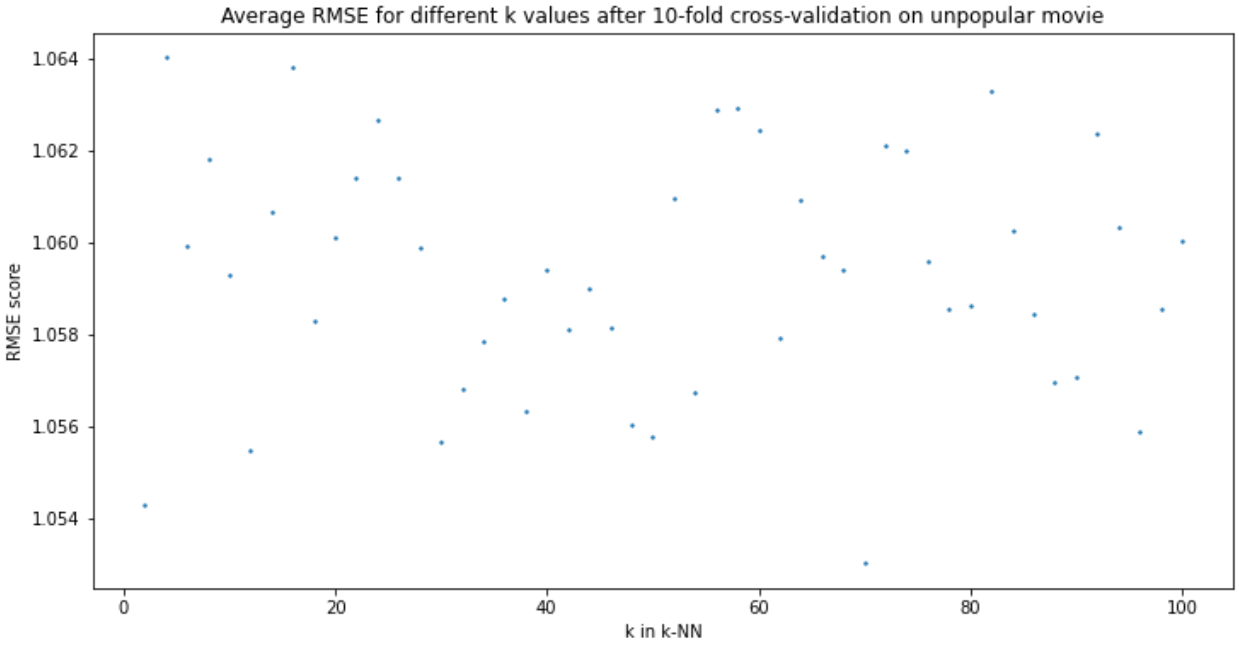
The minimum average RMSE is 0.8891417657120485

For the trimmed popular movie dataset:



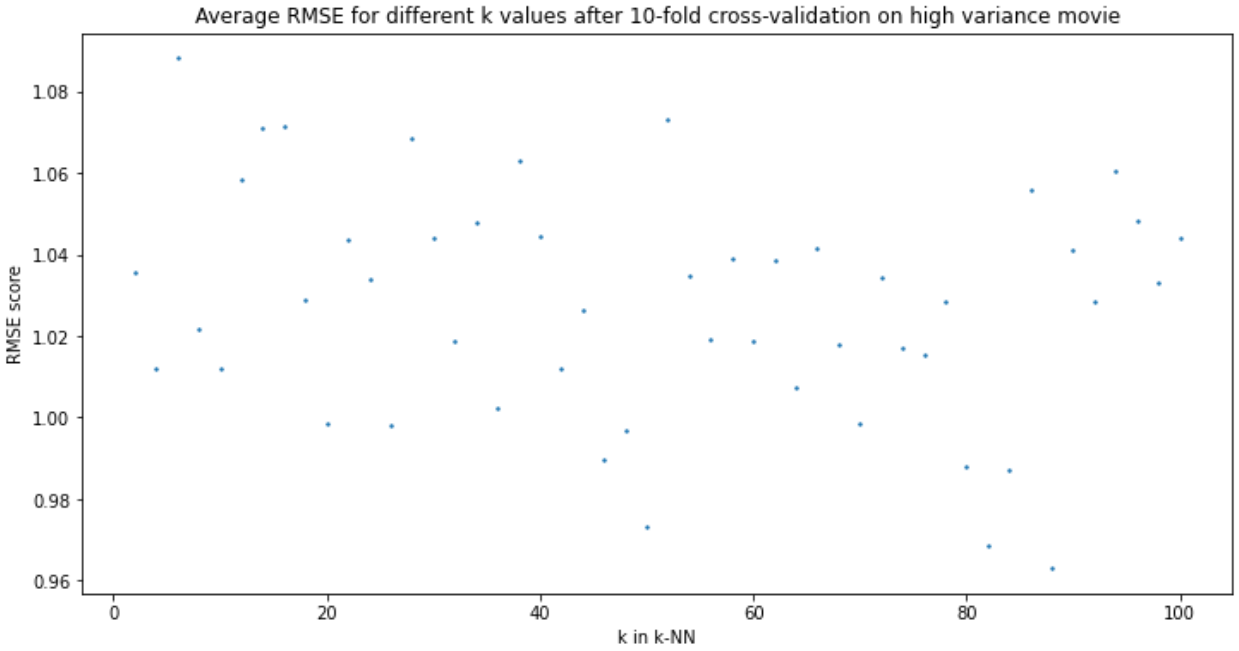
The minimum average RMSE is 0.868761624647077

For the trimmed unpopular movies dataset:



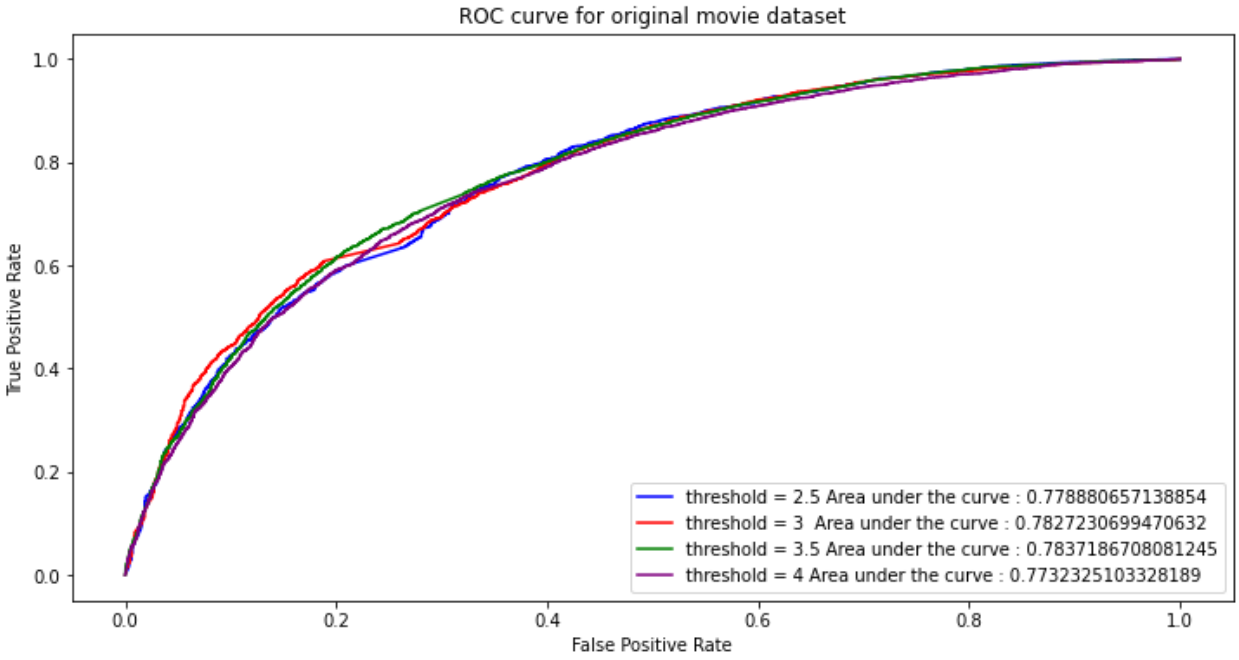
The minimum average RMSE is 1.0530222860498175

For the trimmed high-variance movie dataset:

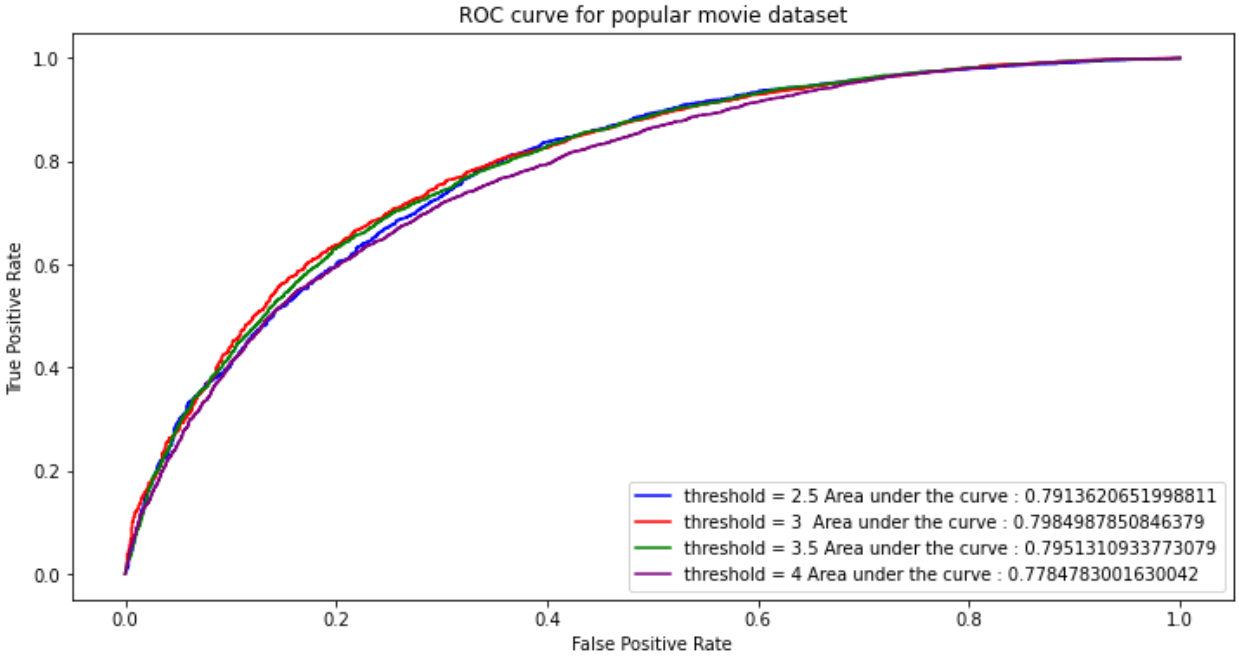


The minimum average RMSE is 0.9631475093020967

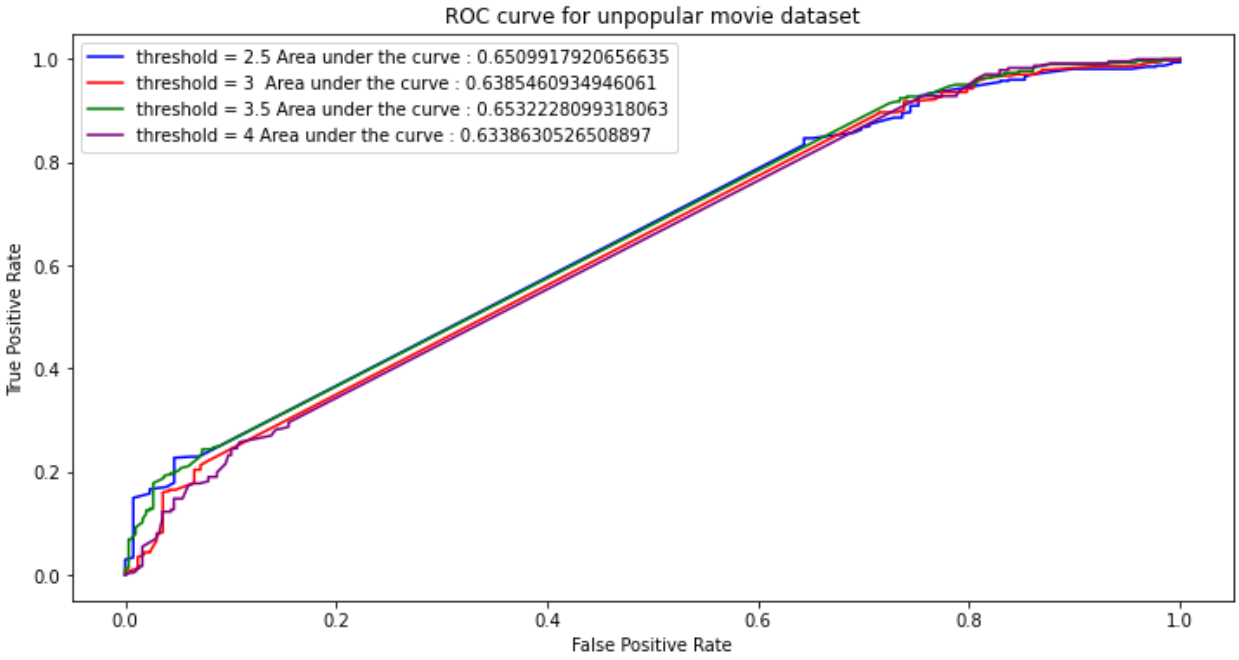
ROC curves for original movie dataset



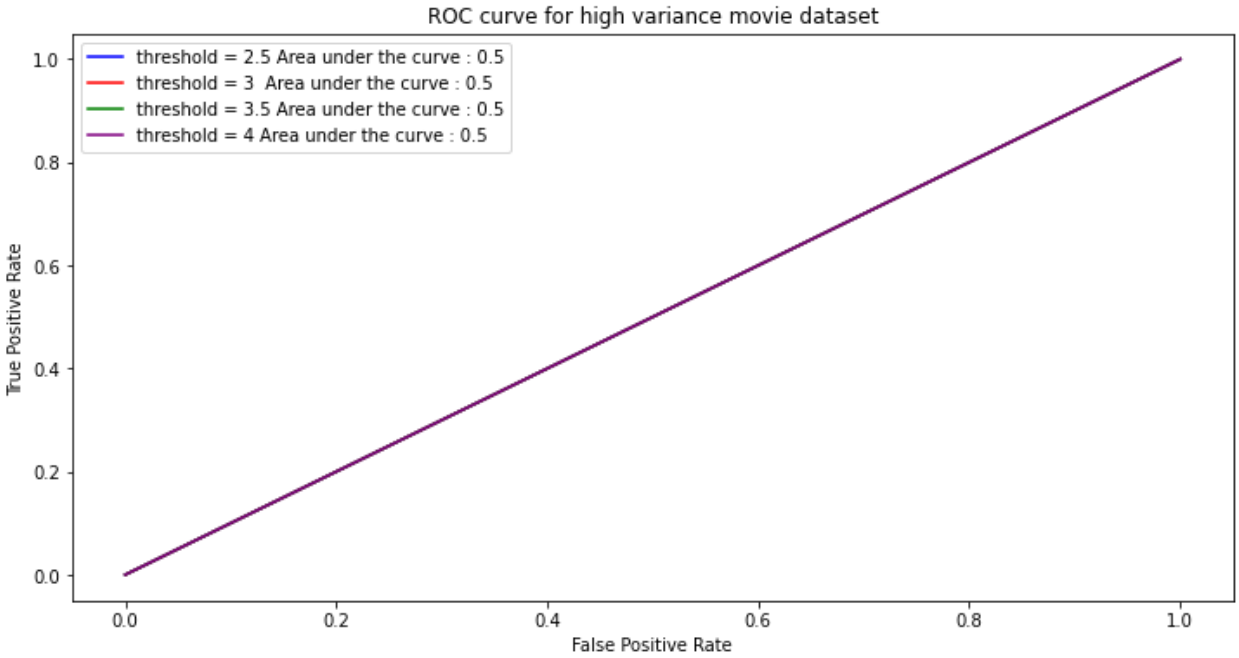
ROC curve for popular movie dataset



ROC curve for unpopular movie dataset



POC curve for high variance movie dataset



Question 7:

$$\underset{U,V}{minimize} \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - (UV^T)_{ij})^2$$

The optimization problem given by this equation is not convex. This equation is quadratic and considers both U and V as variables. The product of U and V introduces the cross-terms, making the function non-convex.

As shown in the following, we can formulate it as a least-squares problem with a fixed U.

$$\underset{V}{minimize} \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - (U_{fixed}V^T)_{ij})^2$$

Question 8:

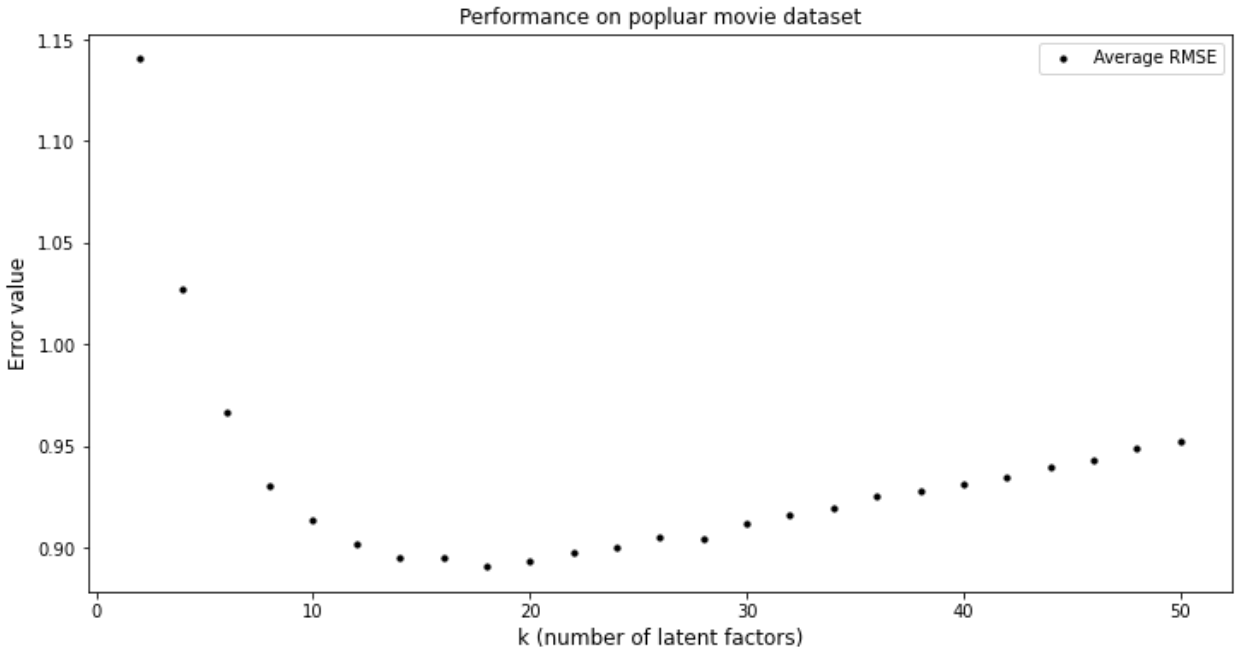
(A) The plot of the average RMSE against and the average MAE against k.

(B) From the plot,

- The optimal number of latent factors (based on RSME) is: 16
- The optimal number of latent factors (based on MAE) is: 22
- The number of movie genres is: 20

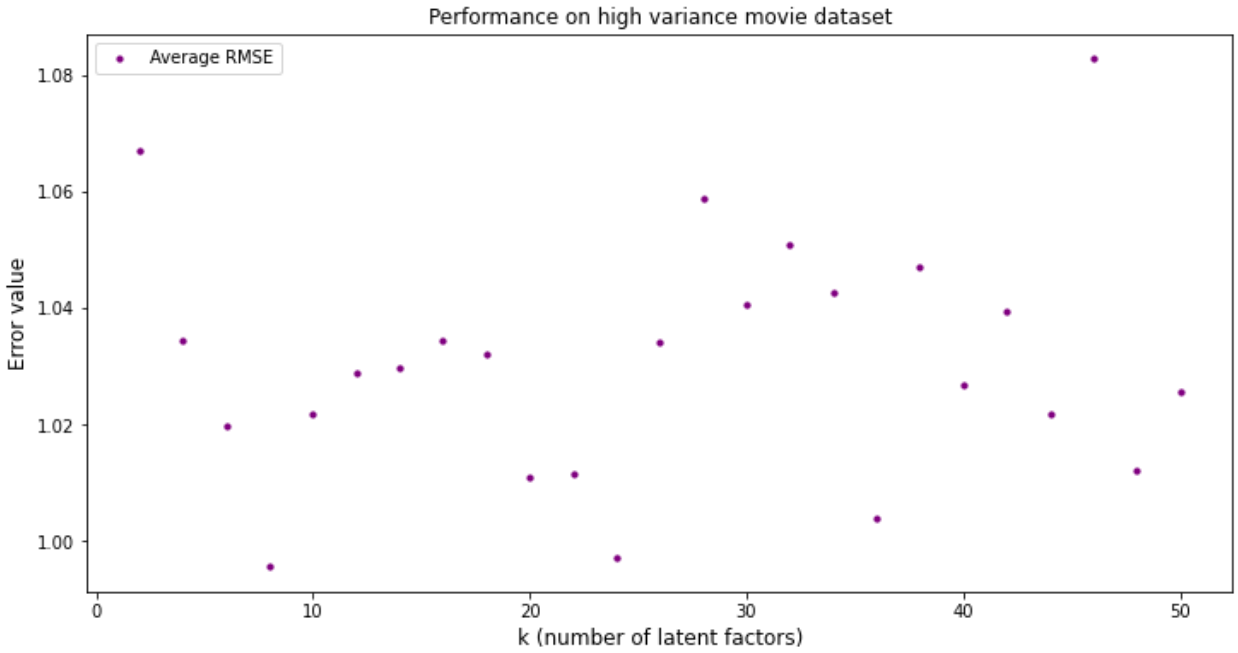
The 10-fold cross-validation process is random. Each validation will produce a slightly different optimal number of latent factors. However, the optimal number of latent factors is in the range of 16-22, which is about the same as that of movie genres.

(C) Performance on trimmed dataset subsets.



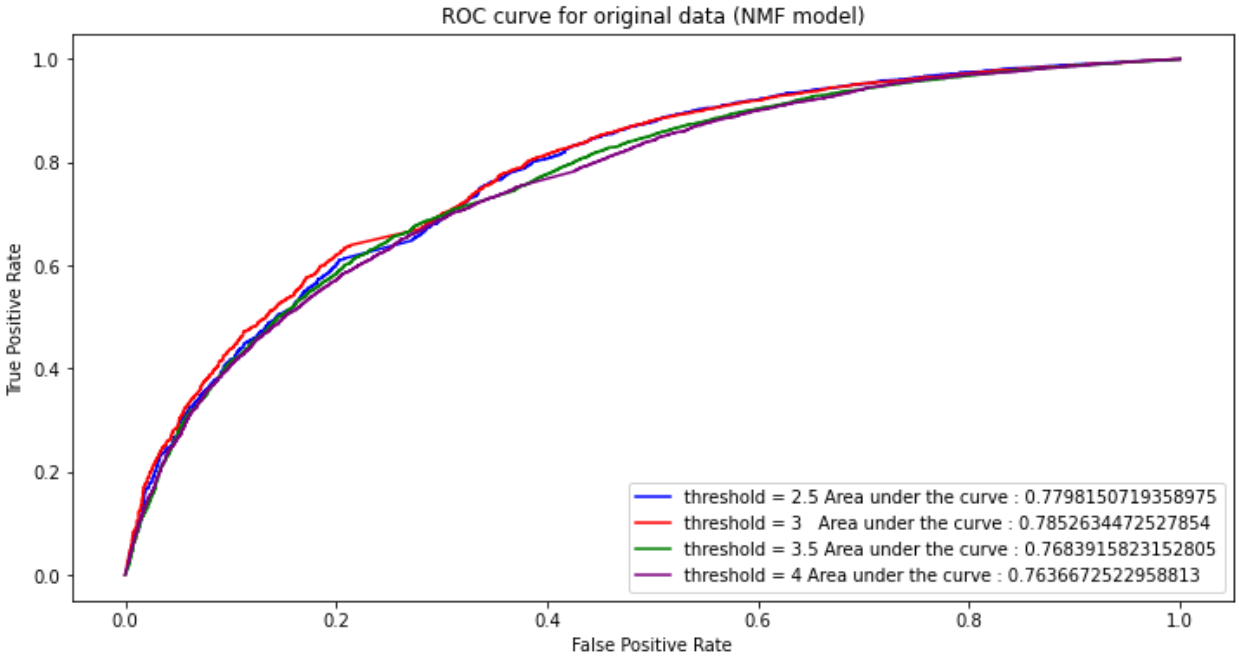

```
The minimum average RSME (popular movie) is:  0.8909788154965529
```

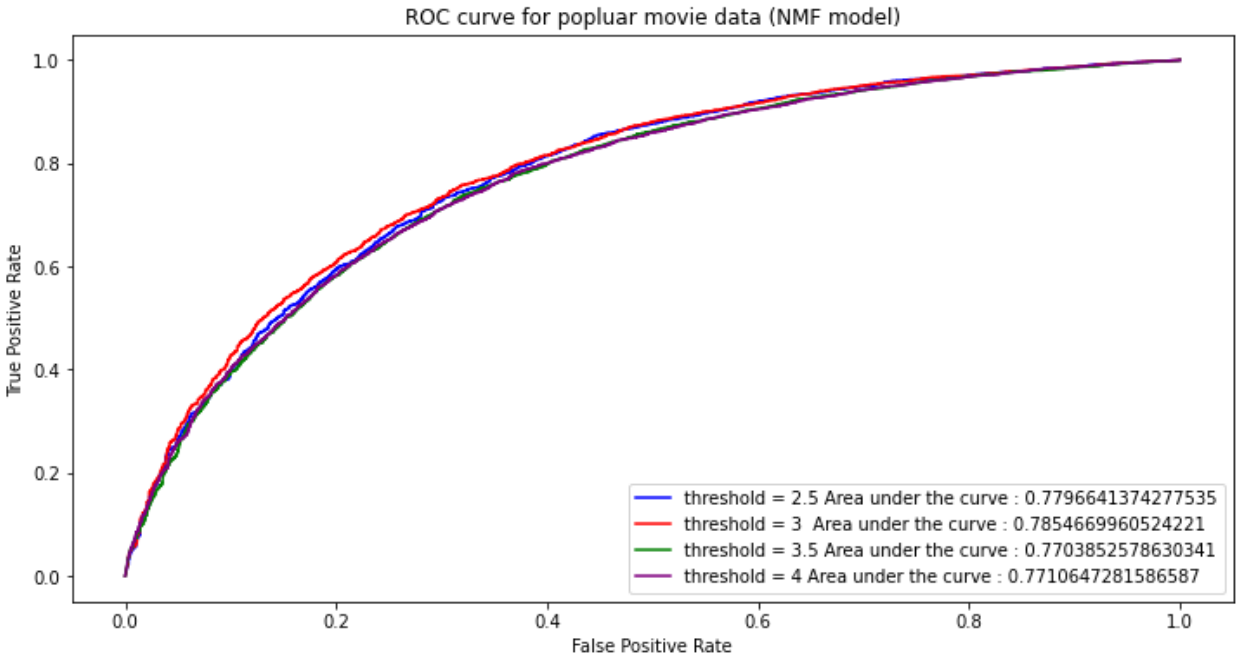

```
The minimum average RSME (unpopular movie) is:  1.1316253566982986
```

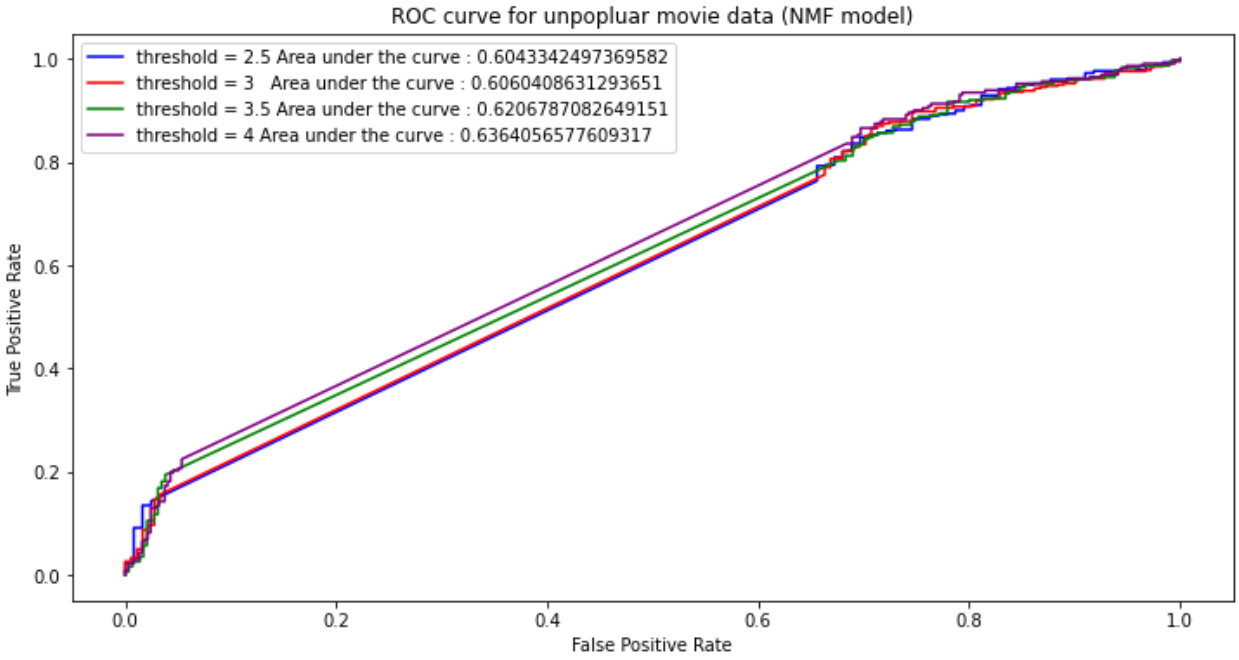


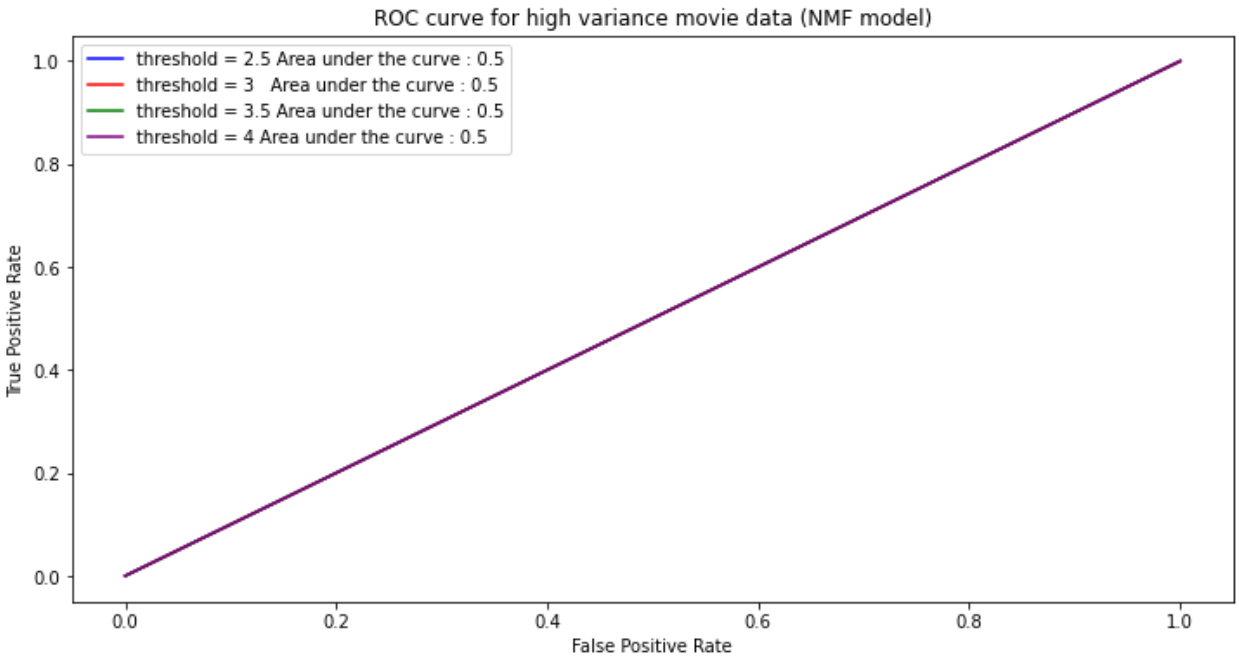

```
The minimum average RSME (high variance movie) is: 0.9956427084815829
```

The plot for the ROC curves for the NMF-based collaborative filter:




ROC curve for popluar movie data (NMF model)
True Positive Rate
False Positive Rate
threshold = 2.5 Area under the curve : 0.7796641374277535
threshold = 3 Area under the curve : 0.7854669960524221
threshold = 3.5 Area under the curve : 0.7703852578630341
threshold = 4 Area under the curve : 0.7710647281586587


ROC curve for unpopluar movie data (NMF model)
True Positive Rate
False Positive Rate
threshold = 2.5 Area under the curve : 0.6043342497369582
threshold = 3 Area under the curve : 0.6060408631293651
threshold = 3.5 Area under the curve : 0.6206787082649151
threshold = 4 Area under the curve : 0.6364056577609317



We are warned that "`No negative samples in y_true, false positive value should be meaningless`" in some runs. This makes sense because the high-variance movie data test set may only contain a single class (0 or 1) after the train-test split and threshold filtering. As a result, the false positive value is meaningless.

Question 9:

```
Genres of the top 10 movies in column No. 1
 Top 1 : ['Children', 'Comedy', 'Drama', 'Mystery']
 Top 2 : ['Drama', 'Romance']
 Top 3 : ['Crime', 'Drama']
 Top 4 : ['Comedy']
 Top 5 : ['Drama']
 Top 6 : ['Comedy', 'Horror', 'Musical']
 Top 7 : ['Comedy', 'Drama', 'Romance']
 Top 8 : ['Comedy', 'Drama', 'Romance']
 Top 9 : ['Drama']
 Top 10 : ['Action', 'Adventure', 'Animation', 'Children']


Genres of the top 10 movies in column No. 2
 Top 1 : ['Drama']
 Top 2 : ['Action', 'Drama', 'Sci-Fi', 'Thriller']
 Top 3 : ['Drama']
 Top 4 : ['Comedy', 'Romance']
 Top 5 : ['Drama']
 Top 6 : ['Documentary']
 Top 7 : ['Action']
 Top 8 : ['Drama', 'War']
 Top 9 : ['Drama']
 Top 10 : ['Drama']


Genres of the top 10 movies in column No. 3
 Top 1 : ['Drama', 'Romance']
 Top 2 : ['Comedy', 'Drama']
 Top 3 : ['Comedy', 'Drama', 'Fantasy', 'Romance']
 Top 4 : ['Comedy', 'Drama', 'War']
 Top 5 : ['Action', 'Adventure', 'Fantasy', 'Romance', 'IMAX']
 Top 6 : ['Action', 'Drama']
 Top 7 : ['Adventure', 'Animation', 'Children', 'Musical']
 Top 8 : ['Drama', 'Romance']
 Top 9 : ['Sci-Fi']
 Top 10 : ['Drama']

Genres of the top 10 movies in column No. 4
 Top 1 : ['Comedy', 'Mystery']
 Top 2 : ['Documentary']
 Top 3 : ['Adventure', 'Children', 'Fantasy', 'Western']
 Top 4 : ['Action', 'Comedy', 'Western']
 Top 5 : ['Comedy']
 Top 6 : ['Drama']
 Top 7 : ['Comedy', 'Fantasy', 'Romance']
 Top 8 : ['Crime', 'Drama']
 Top 9 : ['Adventure', 'Comedy', 'War']
 Top 10 : ['Action', 'Adventure', 'Drama', 'War']


Genres of the top 10 movies in column No. 5
 Top 1 : ['Comedy', 'Drama', 'Romance']
 Top 2 : ['Western']
 Top 3 : ['Action', 'Crime', 'Thriller']
 Top 4 : ['Comedy', 'Crime', 'Romance']
 Top 5 : ['Animation', 'Comedy']
 Top 6 : ['Comedy']
```

```
 Top 7 : ['Action', 'Horror', 'Thriller']
 Top 8 : ['Comedy']
 Top 9 : ['Adventure', 'Comedy', 'Drama', 'Fantasy', 'Mystery', 'Sci-Fi', 'Thriller']
 Top 10 : ['Western']


Genres of the top 10 movies in column No. 6
 Top 1 : ['Action', 'Comedy']
 Top 2 : ['Comedy', 'Horror']
 Top 3 : ['Comedy']
 Top 4 : ['Action', 'Crime', 'Thriller']
 Top 5 : ['Action', 'Adventure', 'Sci-Fi', 'Thriller', 'IMAX']
 Top 6 : ['Action', 'Crime', 'Drama', 'Thriller']
 Top 7 : ['Adventure', 'Comedy', 'Romance']
 Top 8 : ['Comedy', 'Crime', 'Mystery', 'Romance']
 Top 9 : ['Drama', 'Fantasy', 'Mystery', 'Romance']
 Top 10 : ['Thriller']


Genres of the top 10 movies in column No. 7
 Top 1 : ['Crime', 'Thriller']
 Top 2 : ['Drama']
 Top 3 : ['Adventure', 'Comedy', 'Romance']
 Top 4 : ['Drama', 'Fantasy', 'Romance']
 Top 5 : ['Drama']
 Top 6 : ['Comedy', 'Drama']
 Top 7 : ['Comedy']
 Top 8 : ['Adventure', 'Comedy', 'Romance']
 Top 9 : ['Drama']
 Top 10 : ['Action', 'Thriller']


Genres of the top 10 movies in column No. 8
 Top 1 : ['Thriller']
 Top 2 : ['Drama']
 Top 3 : ['Romance', 'War']
 Top 4 : ['Action', 'Adventure', 'Children', 'Comedy']
 Top 5 : ['Action', 'Crime', 'Drama', 'Thriller', 'War']
 Top 6 : ['Animation', 'Drama', 'Romance']
 Top 7 : ['Drama']
 Top 8 : ['Adventure', 'Comedy', 'Sci-Fi']
 Top 9 : ['Action', 'Drama', 'Thriller']
 Top 10 : ['Adventure', 'Drama', 'Horror', 'Sci-Fi', 'Thriller']


Genres of the top 10 movies in column No. 9
 Top 1 : ['Crime', 'Thriller']
 Top 2 : ['Children', 'Drama', 'Fantasy', 'Romance']
 Top 3 : ['Action', 'Drama', 'Romance', 'Thriller']
 Top 4 : ['Action', 'Drama', 'Sci-Fi', 'Thriller']
 Top 5 : ['Crime', 'Drama']
 Top 6 : ['Drama']
 Top 7 : ['Comedy', 'Drama', 'War']
 Top 8 : ['Drama', 'Sci-Fi']
 Top 9 : ['Horror', 'Mystery', 'Thriller']
 Top 10 : ['Action', 'Sci-Fi']

Genres of the top 10 movies in column No. 10
 Top 1 : ['Adventure', 'Drama']
 Top 2 : ['Drama']
 Top 3 : ['Drama', 'Mystery']
 Top 4 : ['Action', 'Animation', 'Sci-Fi', 'Thriller']
```

```
 Top 5 : ['Action', 'Adventure', 'Sci-Fi', 'Thriller']
 Top 6 : ['Children', 'Comedy']
 Top 7 : ['Action', 'Adventure', 'Thriller']
 Top 8 : ['Comedy', 'Drama']
 Top 9 : ['Drama']
 Top 10 : ['Action', 'Adventure', 'Fantasy', 'Mystery']


Genres of the top 10 movies in column No. 11
 Top 1 : ['Comedy', 'Drama']
 Top 2 : ['Drama', 'Thriller']
 Top 3 : ['Action', 'Drama', 'Romance']
 Top 4 : ['Action', 'Adventure', 'Animation', 'Children']
 Top 5 : ['Comedy', 'Drama', 'Romance']
 Top 6 : ['Drama']
 Top 7 : ['Drama']
 Top 8 : ['Action', 'Crime', 'Drama']
 Top 9 : ['Comedy', 'Documentary']
 Top 10 : ['Action', 'Adventure', 'Sci-Fi', 'Thriller']


Genres of the top 10 movies in column No. 12
 Top 1 : ['Comedy', 'Musical', 'Romance']
 Top 2 : ['Action', 'Animation', 'Sci-Fi', 'Thriller']
 Top 3 : ['Drama', 'Romance']
 Top 4 : ['Drama', 'Horror']
 Top 5 : ['Adventure', 'Children', 'Fantasy', 'Western']
 Top 6 : ['Drama', 'Romance']
 Top 7 : ['Crime', 'Drama', 'Thriller']
 Top 8 : ['Comedy', 'Romance']
 Top 9 : ['Comedy']
 Top 10 : ['Drama', 'Romance', 'Sci-Fi', 'Thriller']


Genres of the top 10 movies in column No. 13
 Top 1 : ['Children', 'Comedy', 'Drama', 'Mystery']
 Top 2 : ['Action', 'Comedy', 'Documentary']
 Top 3 : ['Action', 'Adventure', 'Comedy', 'Romance', 'Thriller']
 Top 4 : ['Comedy']
 Top 5 : ['Comedy']
 Top 6 : ['Sci-Fi', 'Thriller']
 Top 7 : ['Crime', 'Drama', 'Thriller']
 Top 8 : ['Comedy', 'Romance']
 Top 9 : ['Horror', 'Mystery', 'Thriller']
 Top 10 : ['Comedy', 'Drama']


Genres of the top 10 movies in column No. 14
 Top 1 : ['Drama', 'Sci-Fi', 'Thriller']
 Top 2 : ['Comedy', 'Drama']
 Top 3 : ['Horror', 'Mystery', 'Thriller']
 Top 4 : ['Drama']
 Top 5 : ['Action', 'Thriller']
 Top 6 : ['Drama', 'Fantasy', 'Horror', 'Thriller', 'War']
 Top 7 : ['Drama']
 Top 8 : ['Crime', 'Drama']
 Top 9 : ['Comedy']
 Top 10 : ['Crime', 'Drama']


Genres of the top 10 movies in column No. 15
 Top 1 : ['Drama']
```

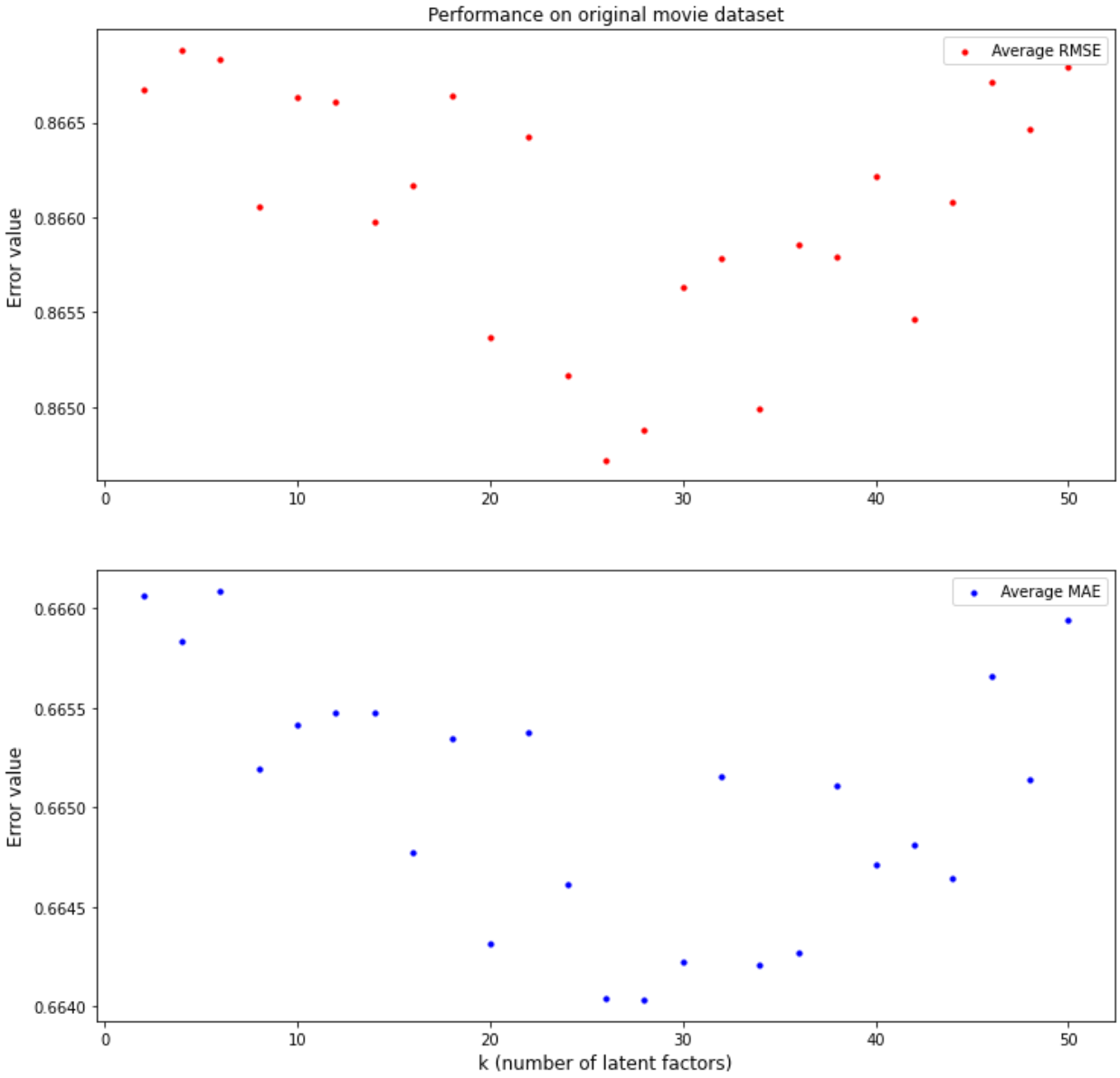
```
 Top 2 : ['Action']
 Top 3 : ['Comedy']
 Top 4 : ['Action', 'Crime', 'Drama', 'Thriller']
 Top 5 : ['Adventure', 'Fantasy']
 Top 6 : ['Action', 'Crime']
 Top 7 : ['Action', 'Adventure', 'Drama', 'Fantasy', 'Thriller']
 Top 8 : ['Drama', 'Mystery', 'Thriller']
 Top 9 : ['Drama', 'Mystery', 'Thriller']
 Top 10 : ['Documentary']

Genres of the top 10 movies in column No. 16
 Top 1 : ['Comedy', 'Crime']
 Top 2 : ['Comedy']
 Top 3 : ['Drama']
 Top 4 : ['Drama']
 Top 5 : ['Drama']
 Top 6 : ['Mystery', 'Thriller']
 Top 7 : ['Drama', 'Thriller']
 Top 8 : ['Action', 'Drama', 'Thriller']
 Top 9 : ['Comedy', 'Drama']
 Top 10 : ['Crime', 'Drama']


Genres of the top 10 movies in column No. 17
 Top 1 : ['Comedy']
 Top 2 : ['Comedy', 'Fantasy']
 Top 3 : ['Comedy']
 Top 4 : ['Action', 'Drama', 'Sci-Fi', 'Thriller']
 Top 5 : ['Drama', 'War']
 Top 6 : ['Drama', 'Musical']
 Top 7 : ['Thriller']
 Top 8 : ['Crime', 'Drama']
 Top 9 : ['Action', 'Adventure', 'Thriller']
 Top 10 : ['Action', 'Thriller']


Genres of the top 10 movies in column No. 18
 Top 1 : ['Comedy', 'Crime', 'Romance']
 Top 2 : ['Drama']
 Top 3 : ['Horror']
 Top 4 : ['Crime', 'Thriller']
 Top 5 : ['Comedy', 'Drama', 'Romance']
 Top 6 : ['Comedy', 'Drama', 'Fantasy', 'Romance']
 Top 7 : ['Mystery', 'Thriller']
 Top 8 : ['Action', 'Adventure', 'Drama']
 Top 9 : ['Action', 'Crime', 'Drama']
 Top 10 : ['Comedy']

Genres of the top 10 movies in column No. 19
 Top 1 : ['Action', 'Horror', 'Sci-Fi']
 Top 2 : ['Comedy']
 Top 3 : ['Drama', 'Romance']
 Top 4 : ['Comedy', 'Musical', 'Romance']
 Top 5 : ['Action', 'Adventure', 'Sci-Fi', 'Thriller']
 Top 6 : ['Action', 'Adventure']
 Top 7 : ['Action', 'Sci-Fi']
 Top 8 : ['Action', 'Adventure', 'Sci-Fi']
 Top 9 : ['Comedy']
 Top 10 : ['Crime', 'Drama', 'Thriller']
```

```
Genres of the top 10 movies in column No. 20
 Top 1 : ['Action', 'Comedy', 'Crime']
 Top 2 : ['Action', 'Mystery', 'Thriller']
 Top 3 : ['Drama']
 Top 4 : ['Comedy', 'Horror']
 Top 5 : ['Sci-Fi']
 Top 6 : ['Crime', 'Drama']
 Top 7 : ['Horror', 'Mystery', 'Sci-Fi', 'Thriller']
 Top 8 : ['Adventure', 'Fantasy']
 Top 9 : ['Comedy']
 Top 10 : ['Comedy', 'Drama']
```

The top 10 movies belong to a small collection of genres. For example, most of the top 10 movies in column 1 belong to the comedy and drama genres. There is a connection between the latent factors and the movie genres.

Question 10:

(A) The plot of the average RMSE against and the average MAE against k.



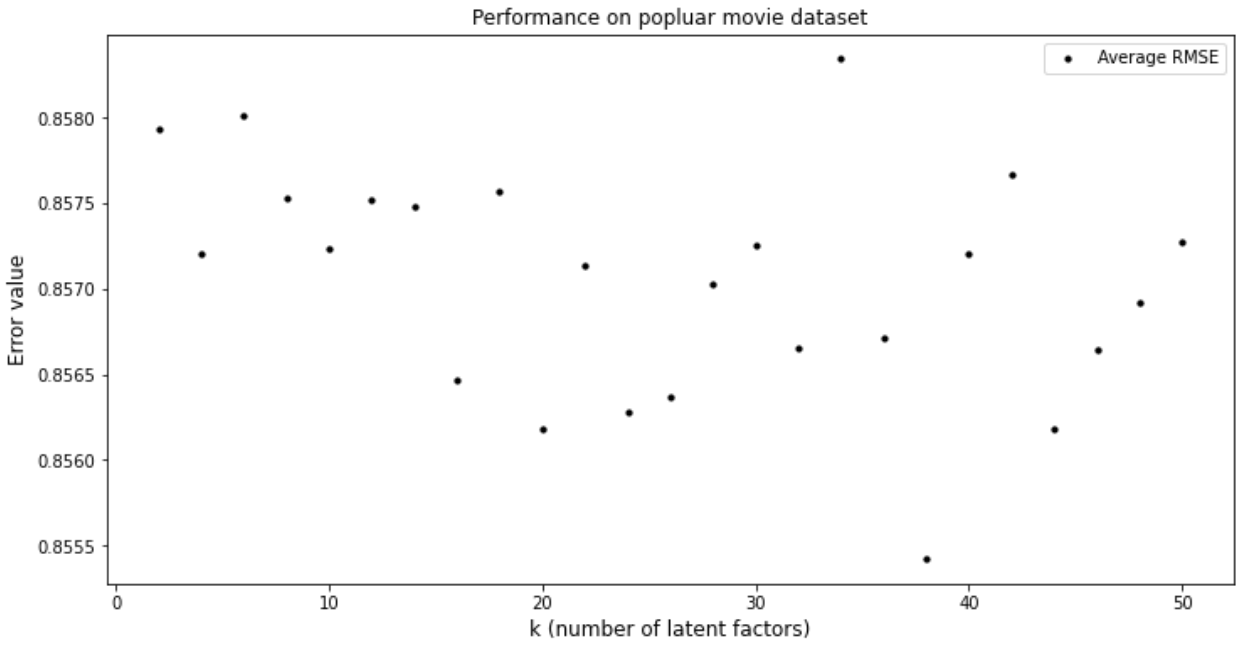
(B) From the plot,

- The optimal number of latent factors (based on RSME) is: 26
- The optimal number of latent factors (based on MAE) is: 28
- The number of movie genres is: 20

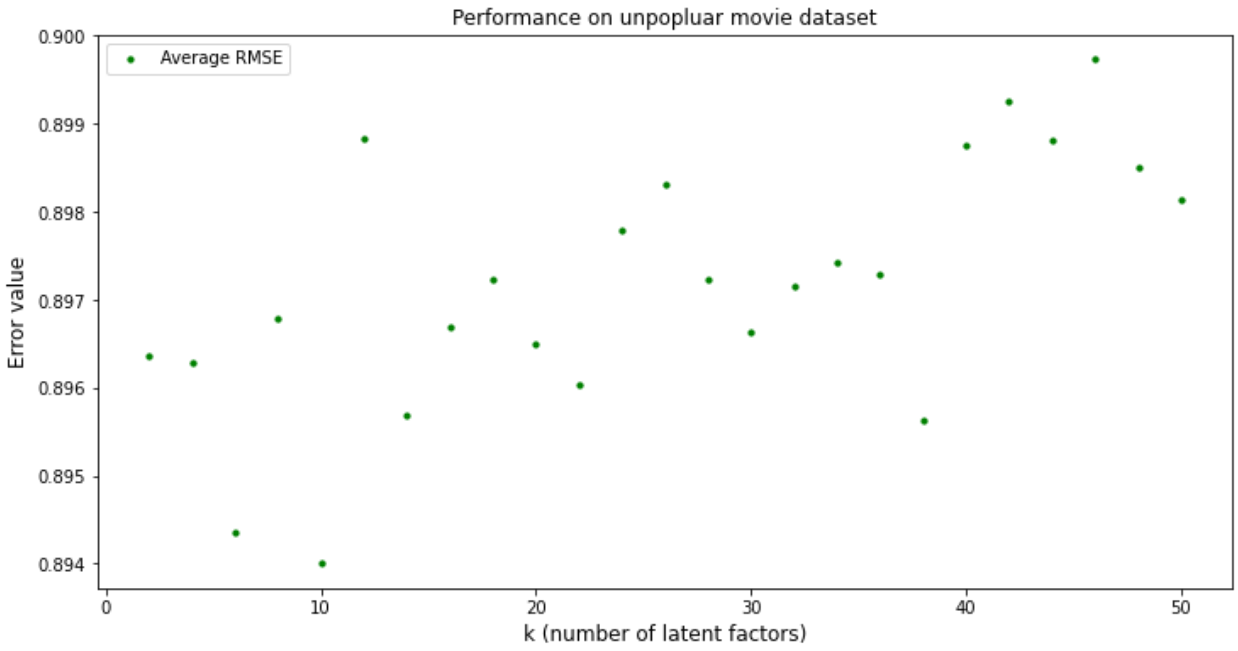
The figure shows that the RSME and AME for MF with a bias filter vary slightly (in the 0.0005 range) as k changes. In multiple runs, the optimal k ranges from 20-50. It is difficult to conclude an optimal k in this case. However, the optimal k occurs most often in the 30-40 range.

As a result, the optimal number of latent factors (20-50) is in the same order as the number of movie genres.

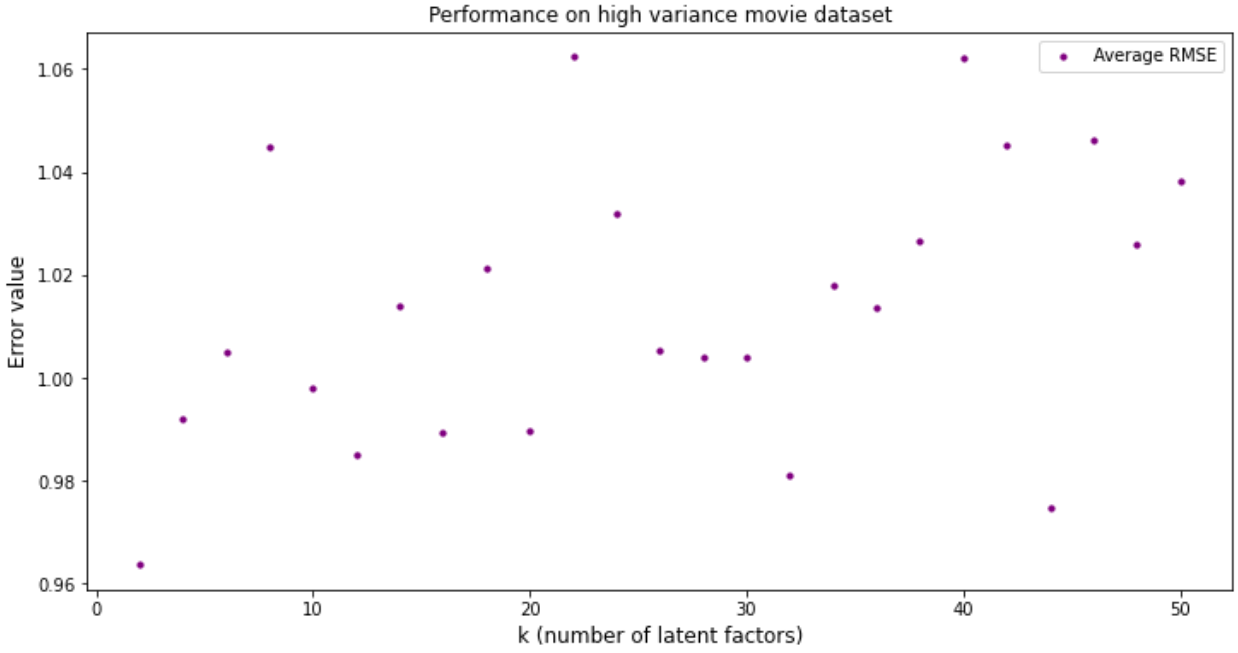
(C) Performance on trimmed dataset subsets.




```
The minimum average RSME (popular movie) is: 0.855425150473684
```

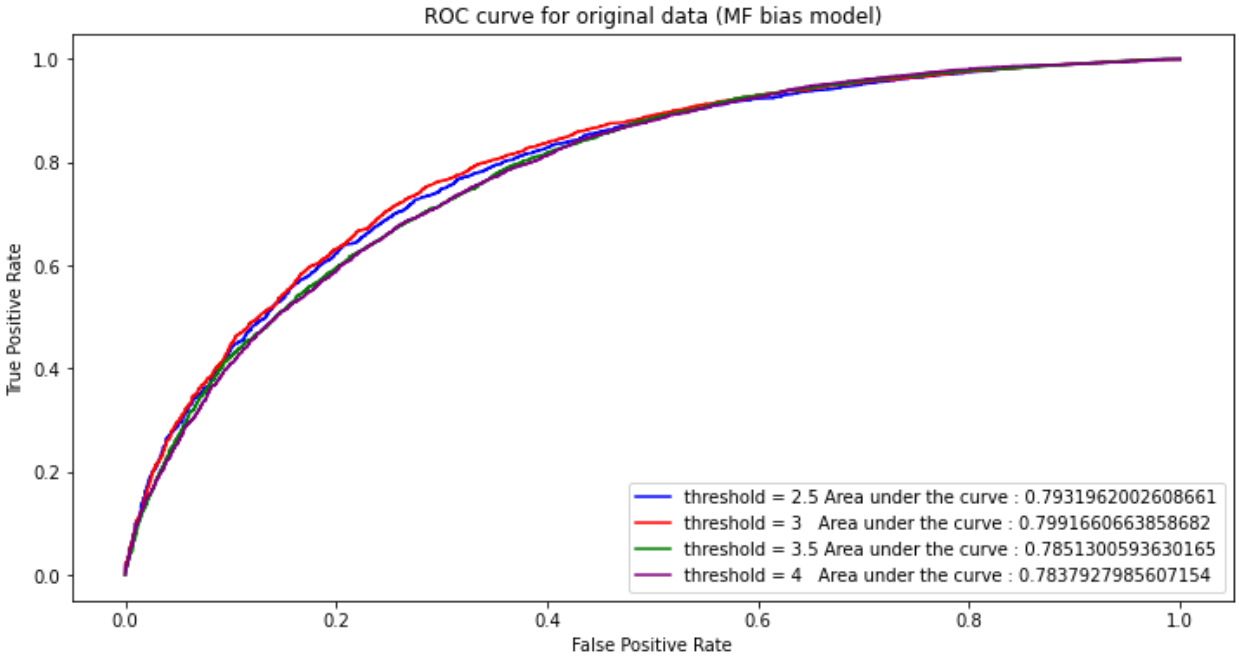


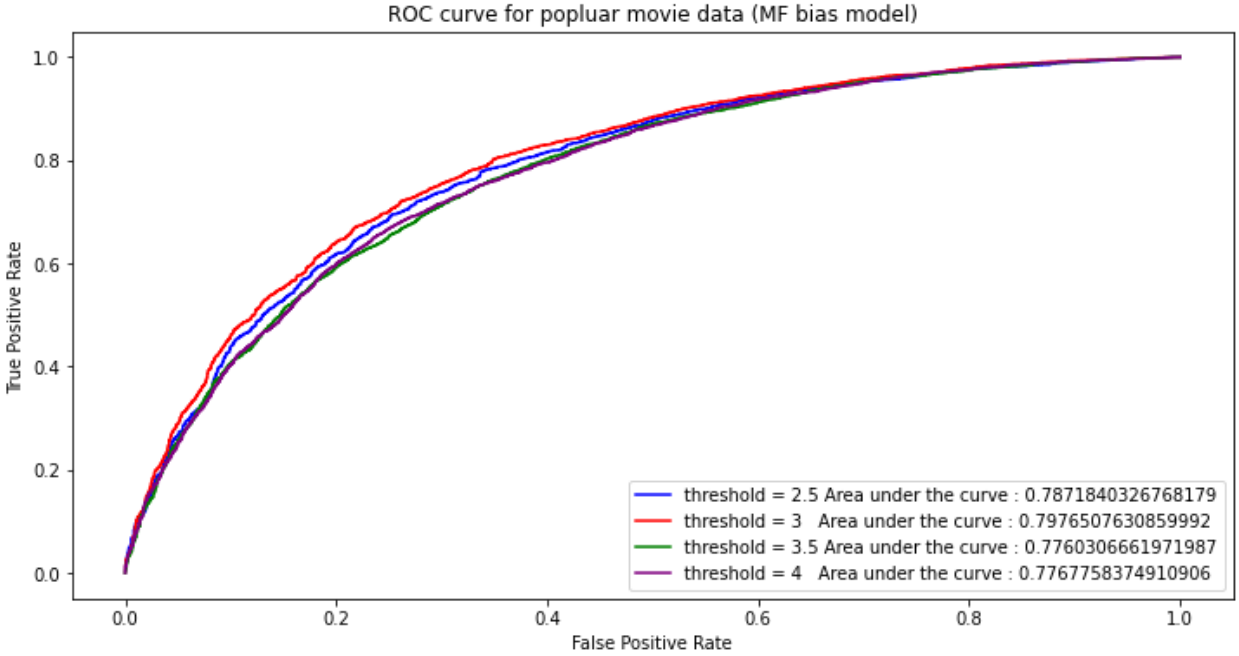

```
The minimum average RSME (unpopular movie) is: 0.8940091972795013
```

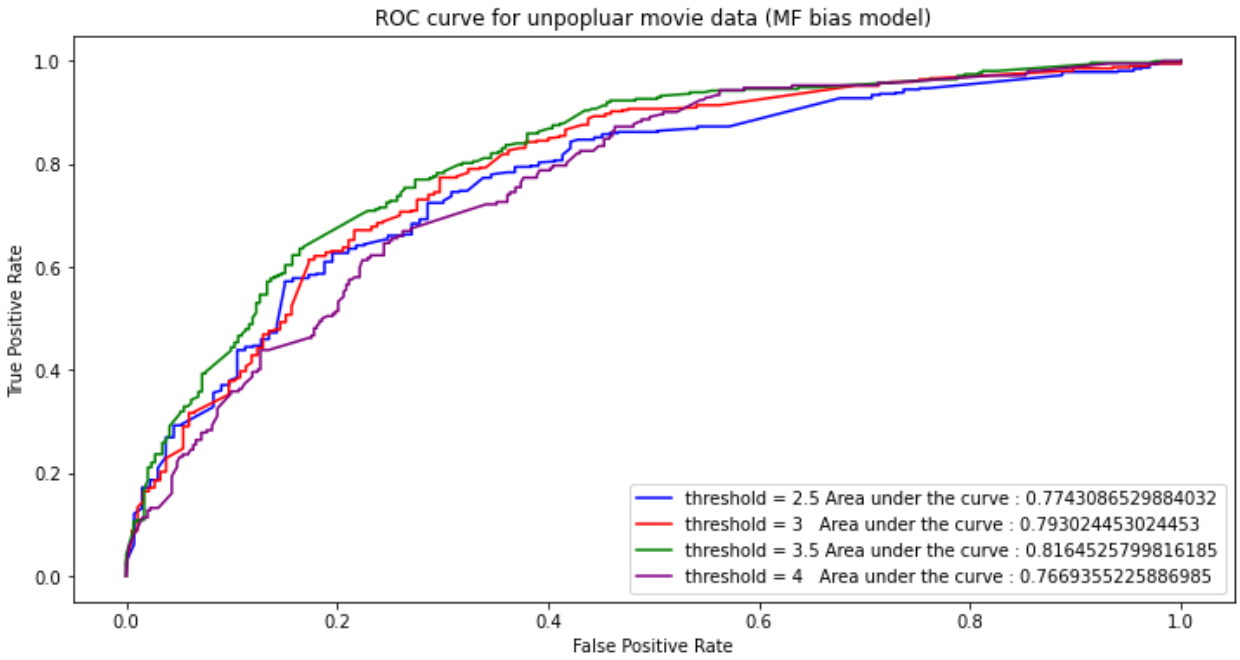


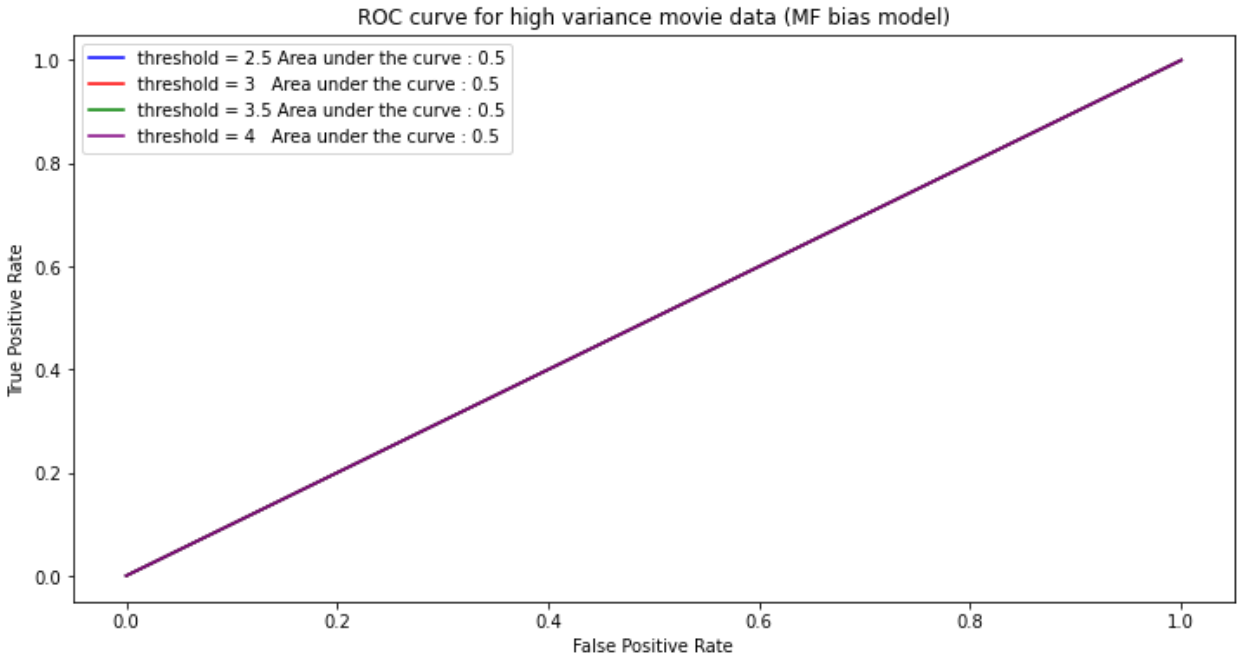

```
The minimum average RSME (high-variance movie) is: 0.9638052335213301
```

The plot for the ROC curves for the NMF-based collaborative filter:


ROC curve for original data (MF bias model)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
threshold = 2.5 Area under the curve : 0.7931962002608661
threshold = 3 Area under the curve : 0.7991660663858682
threshold = 3.5 Area under the curve : 0.7851300593630165
threshold = 4 Area under the curve : 0.7837927985607154


ROC curve for popluar movie data (MF bias model)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
threshold = 2.5 Area under the curve : 0.7871840326768179
threshold = 3 Area under the curve : 0.7976507630859992
threshold = 3.5 Area under the curve : 0.7760306661971987
threshold = 4 Area under the curve : 0.7767758374910906





We get the same warnings as in Question 8: "`No negative samples in y_true, false positive value should be meaningless.`" in some runs.

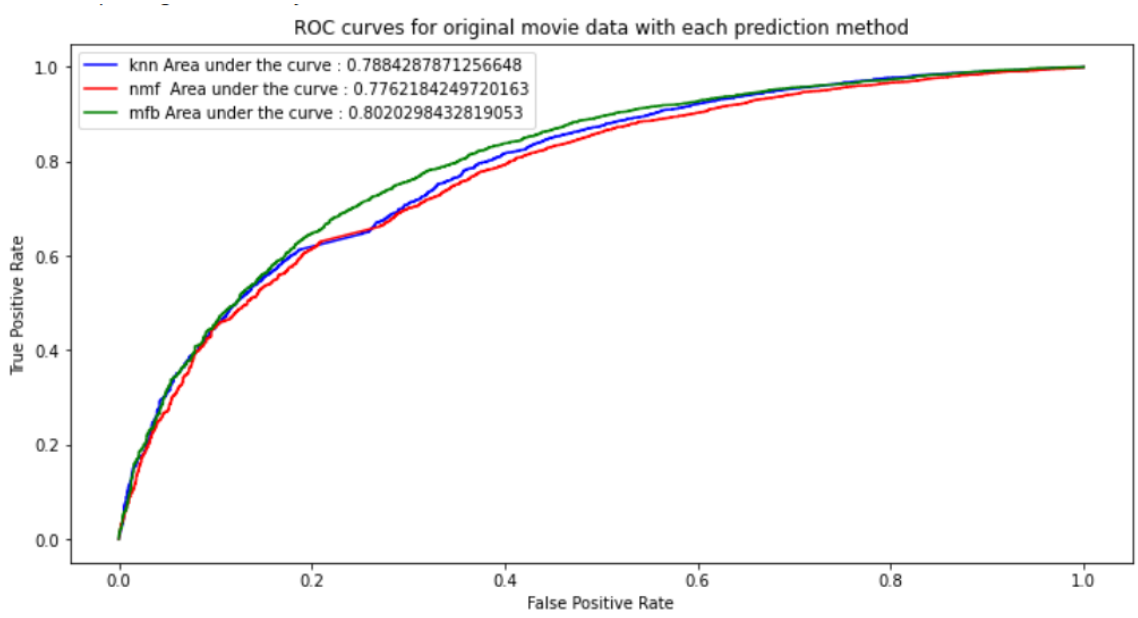
Question 11:

A naive collaborative filter is designed using the mean of each user for prediction. The average RSMEs of four datasets (original, popular, unpopular, and high variance subset) are shown below.

```
Average RSME for Original Data: 0.9346543847132723
Average RSME for Popular Subset: 0.9323149727540322
Average RSME for Unpopular Subset: 0.9706604147083009
Average RSME for High Variance Subset: 0.8150759443083917
```

Question 12:

The plot of ROC curves for three methods is shown below. As we can see from the areas under the curve, MFB (MF with bias) has the best performance, while NMF has the worst performance.



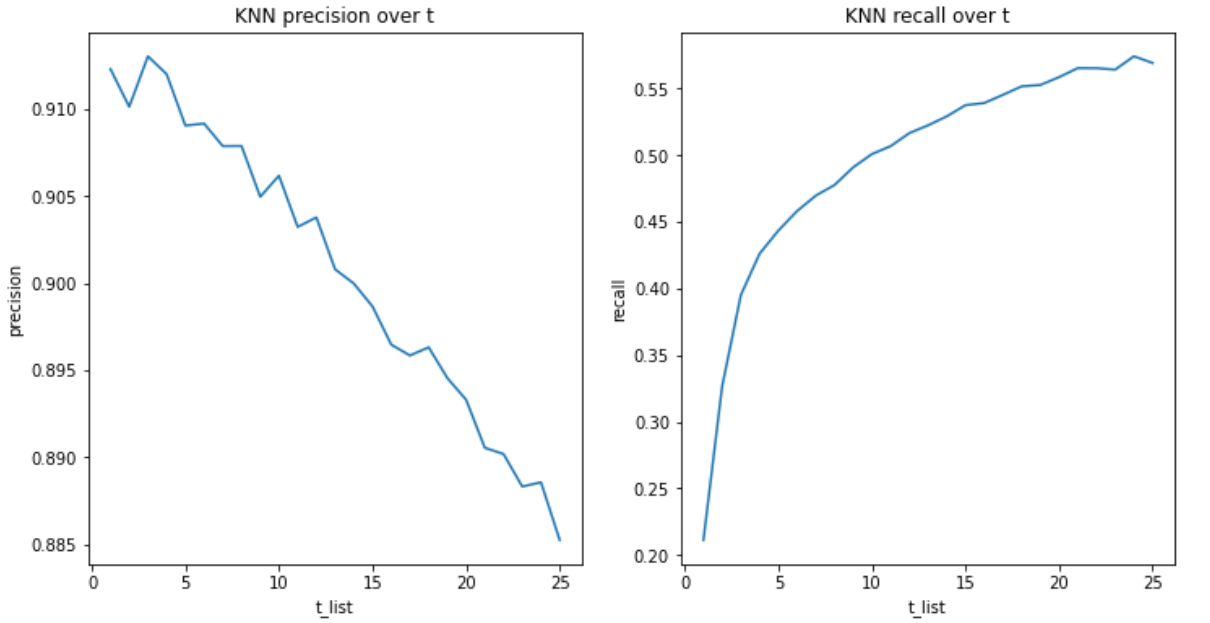
Question 13:

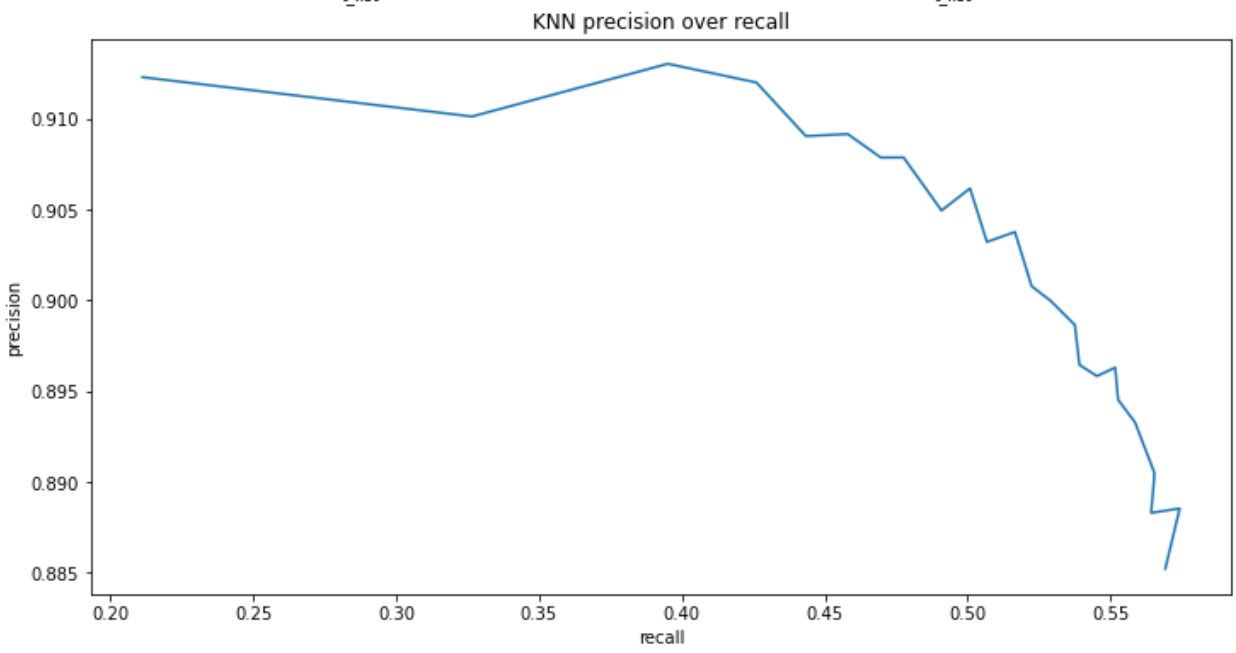
Precision means the accuracy for which the ranking predicts the user's ratings. It reveals how many of the top t recommended movies are correctly predicted as liked by the user.

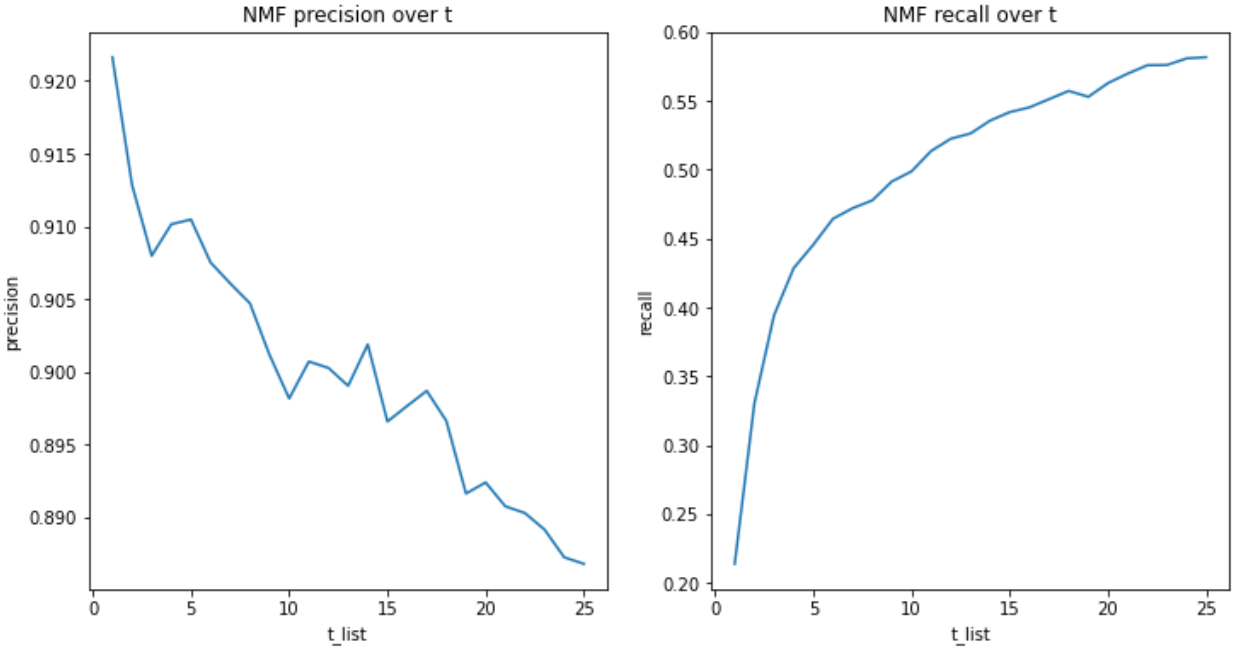
On the other hand, recall shows the relevance of the ranking. The recall is gained by calculating the ratio of a user's recommended movies and overall liked movies. It evaluates whether the prediction correctly reveals the user's preference.

Question 14:

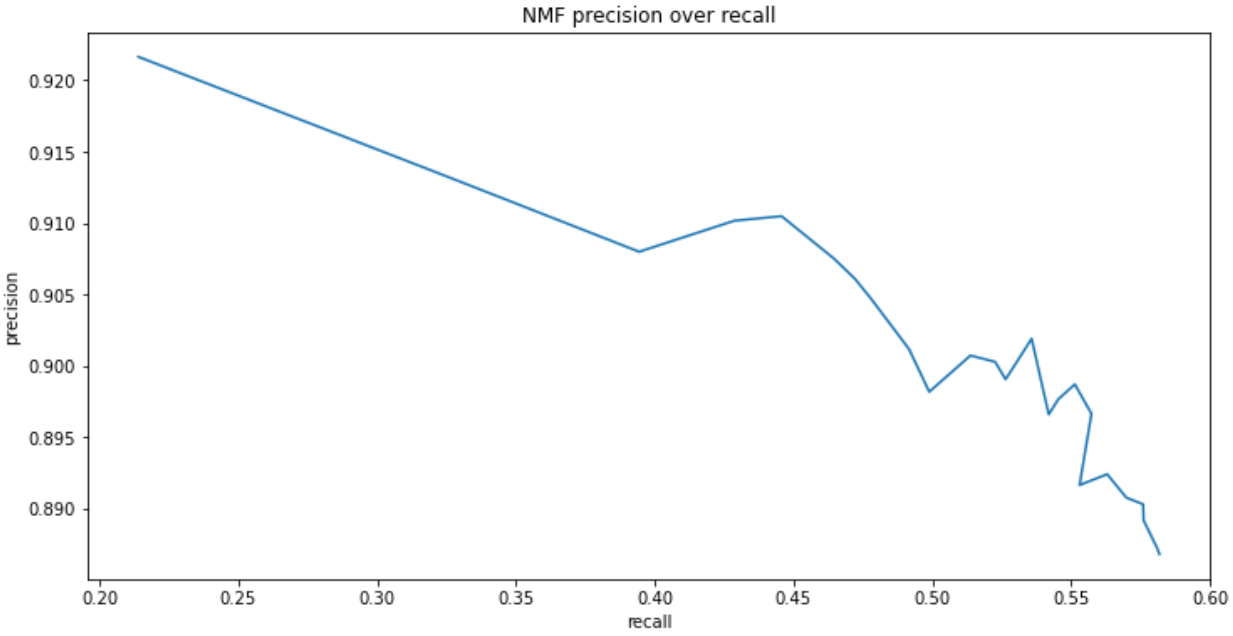
In this part, three plots are generated for each collaborative filter: precision against t, recall against t, and precision against the recall. As shown in the graphs below, as t increases, the precision decreases, and the recall increases. This makes sense since as t increases, more recommended movies are included. On the one hand, this means more possibility of wrong prediction so that precision decreases; on the other hand, this means the prediction is more relevant to the user's actual preference, thus recall increases.

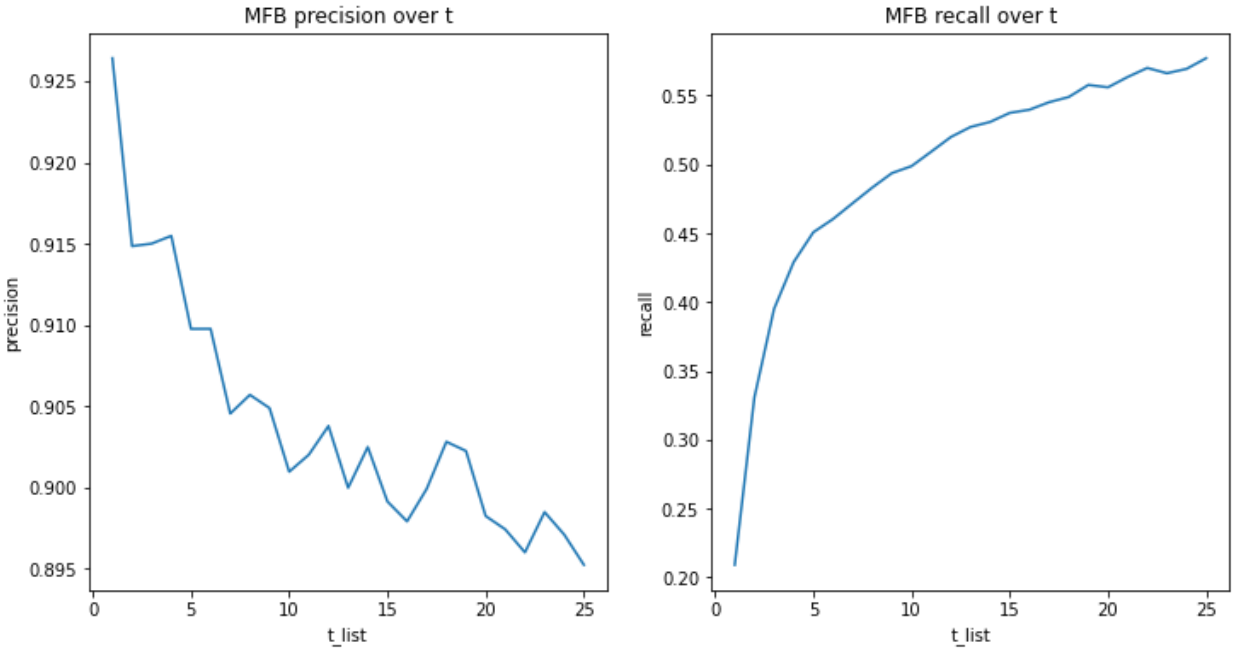




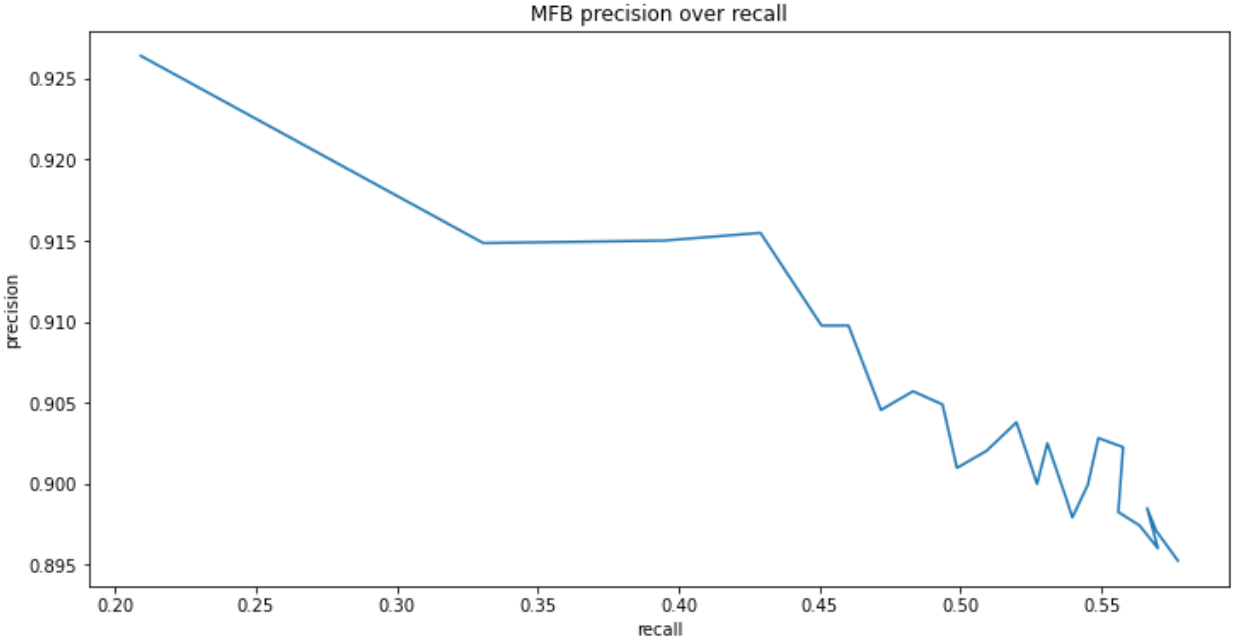

NMF precision over t
precision
t_list
0.920
0.915
0.910
0.905
0.900
0.895
0.890
0
5
10
15
20
25

NMF recall over t
recall
t_list
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0
5
10
15
20
25


NMF precision over recall
precision
recall
0.920
0.915
0.910
0.905
0.900
0.895
0.890
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60


MFB precision over t
precision
t_list
0.925
0.920
0.915
0.910
0.905
0.900
0.895
0
5
10
15
20
25

MFB recall over t
recall
t_list
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0
5
10
15
20
25


MFB precision over recall
precision
recall
0.925
0.920
0.915
0.910
0.905
0.900
0.895
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55

Then the precision-recall plot is shown in one graph below. As we can see, precision and recall are negatively correlated. As precision decreases, recall increases. For the comparison between the three collaborative filters, we can see MFB has the best performance, while NMF and KNN are similar with KNN slightly better.