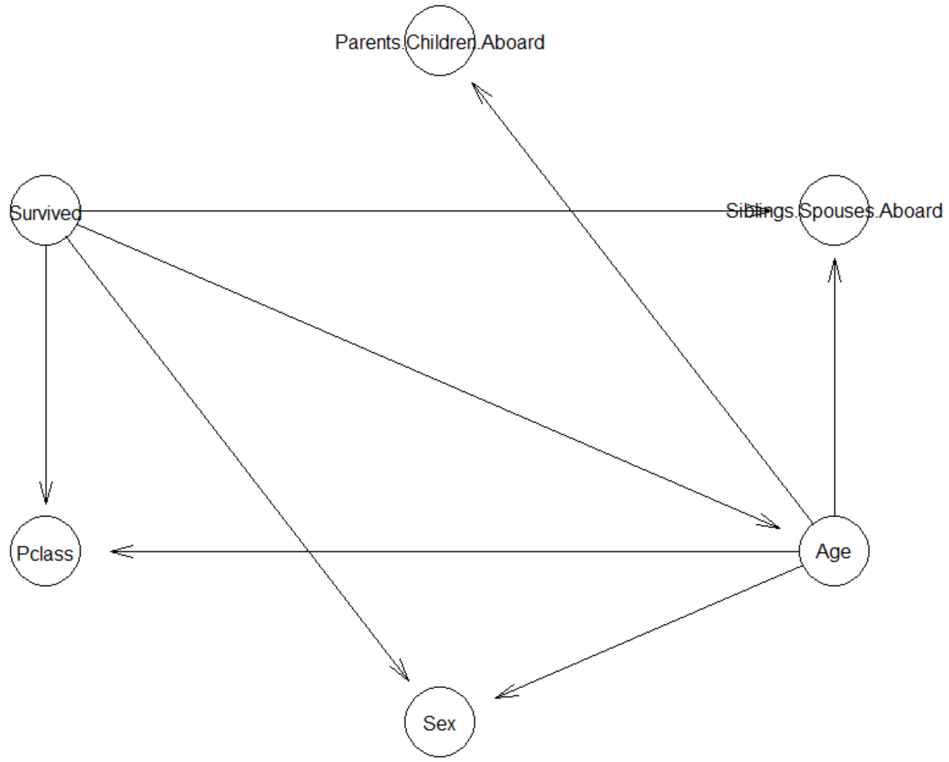
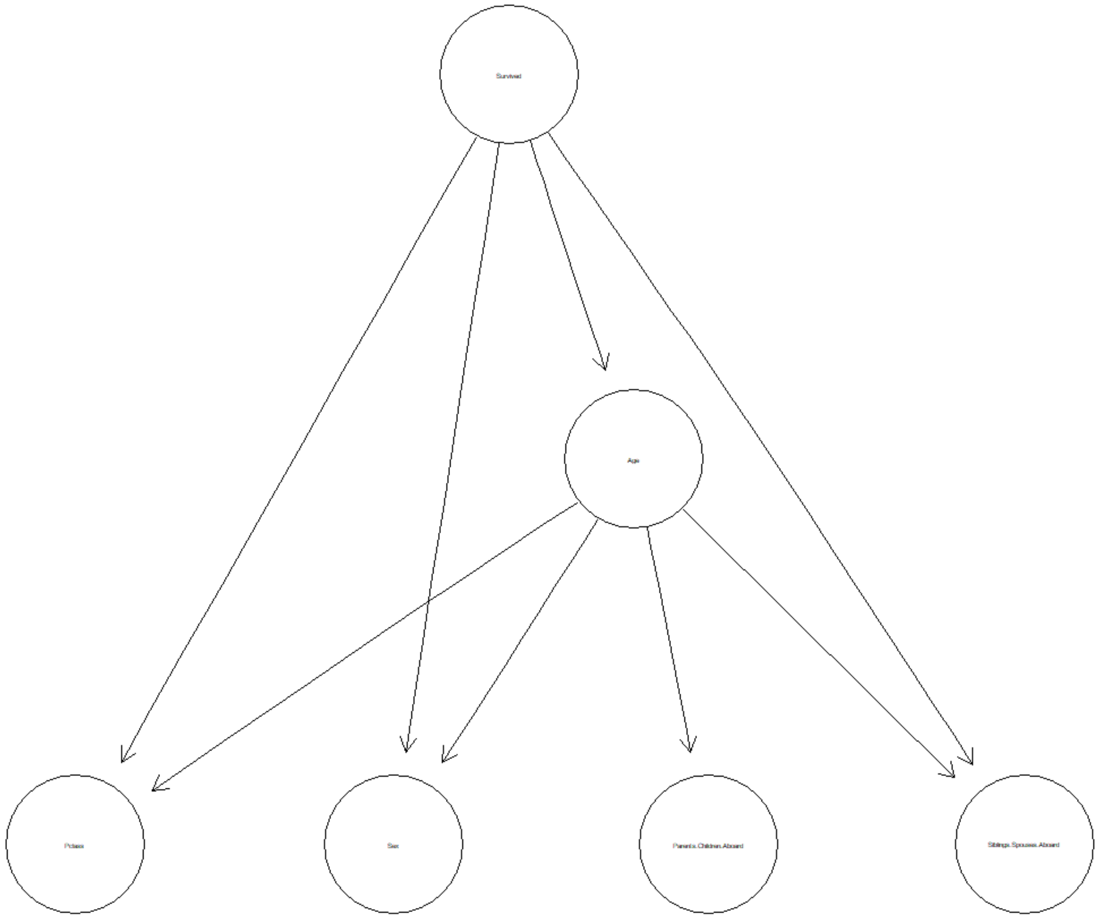
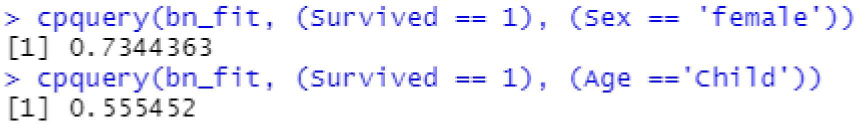
Question 2


Parents.Children.Aboard
Survived
Siblings.Spouses.Aboard
Pclass
Age
Sex


Survived
Age
Pclass
Sex
Parents.Children.Aboard
Siblings.Spouses.Aboard

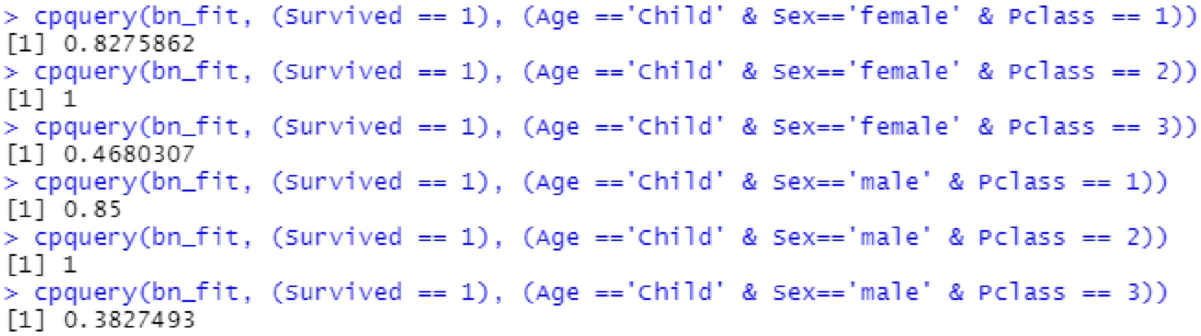
a. Prob of women and child survival



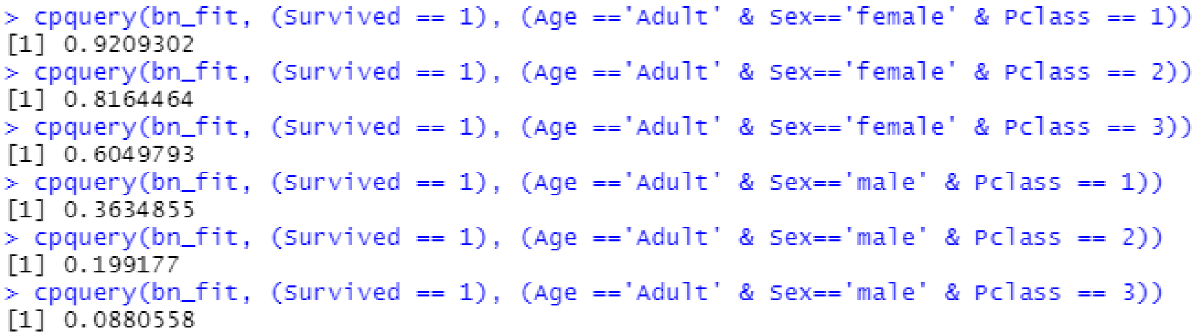
```
> cpquery(bn_fit, (Survived == 1), (Sex == 'female'))
[1] 0.7344363
> cpquery(bn_fit, (Survived == 1), (Age =='Child'))
[1] 0.555452
```

Probability of women and child survival is about 0.70 and 0.5. Let's include more conditions below to see the detailed probability.

b. What characteristics/demographics are more likely in surviving passengers?



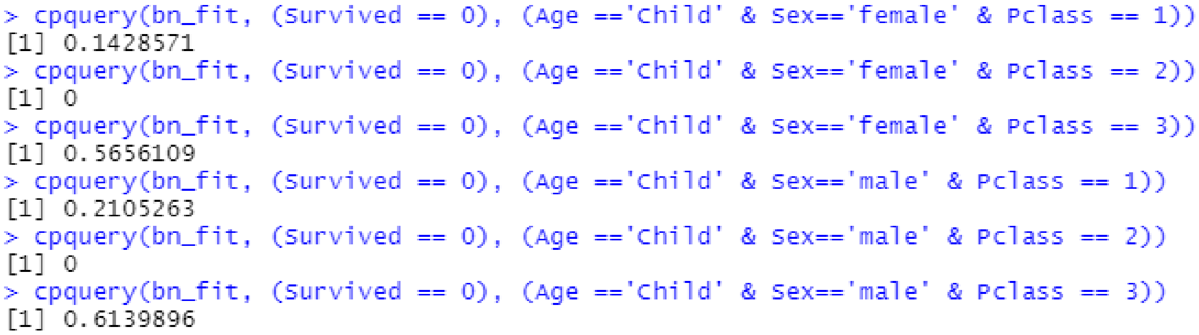
```
> cpquery(bn_fit, (Survived == 1), (Age =='Child' & Sex=='female' & Pclass == 1))
[1] 0.8275862
> cpquery(bn_fit, (Survived == 1), (Age =='Child' & Sex=='female' & Pclass == 2))
[1] 1
> cpquery(bn_fit, (Survived == 1), (Age =='Child' & Sex=='female' & Pclass == 3))
[1] 0.4680307
> cpquery(bn_fit, (Survived == 1), (Age =='Child' & Sex=='male' & Pclass == 1))
[1] 0.85
> cpquery(bn_fit, (Survived == 1), (Age =='Child' & Sex=='male' & Pclass == 2))
[1] 1
> cpquery(bn_fit, (Survived == 1), (Age =='Child' & Sex=='male' & Pclass == 3))
[1] 0.3827493
```



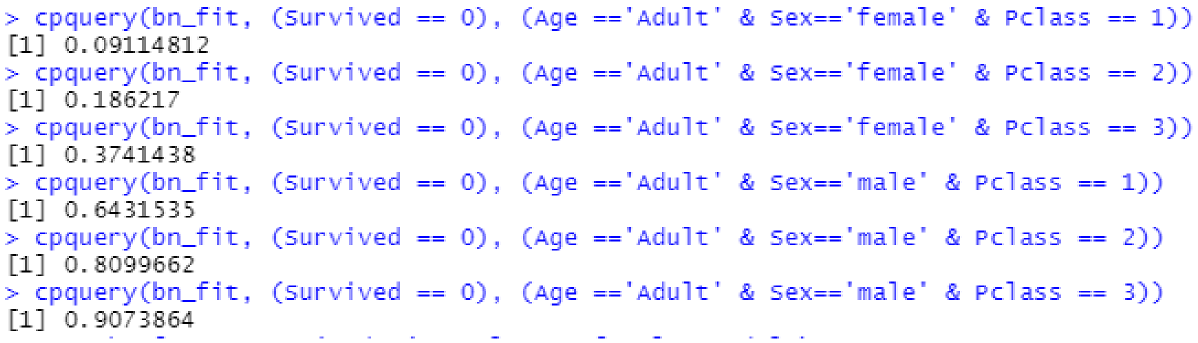
```
> cpquery(bn_fit, (Survived == 1), (Age =='Adult' & Sex=='female' & Pclass == 1))
[1] 0.9209302
> cpquery(bn_fit, (Survived == 1), (Age =='Adult' & Sex=='female' & Pclass == 2))
[1] 0.8164464
> cpquery(bn_fit, (Survived == 1), (Age =='Adult' & Sex=='female' & Pclass == 3))
[1] 0.6049793
> cpquery(bn_fit, (Survived == 1), (Age =='Adult' & Sex=='male' & Pclass == 1))
[1] 0.3634855
> cpquery(bn_fit, (Survived == 1), (Age =='Adult' & Sex=='male' & Pclass == 2))
[1] 0.199177
> cpquery(bn_fit, (Survived == 1), (Age =='Adult' & Sex=='male' & Pclass == 3))
[1] 0.0880558
```

As you can see, all of the 2nd Pclass children survived 100%. Besides that, the highest possibility of survival is when a person is in 1st class, female and adult. On the other hand, the lowest possibility of survival is when a person is in 3rd class, male and adult.

c. What characteristics/demographics are more likely in passengers that perished?



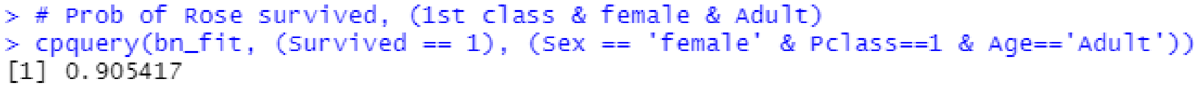
```
> cpquery(bn_fit, (Survived == 0), (Age =='Child' & Sex=='female' & Pclass == 1))
[1] 0.1428571
> cpquery(bn_fit, (Survived == 0), (Age =='Child' & Sex=='female' & Pclass == 2))
[1] 0
> cpquery(bn_fit, (Survived == 0), (Age =='Child' & Sex=='female' & Pclass == 3))
[1] 0.5656109
> cpquery(bn_fit, (Survived == 0), (Age =='Child' & Sex=='male' & Pclass == 1))
[1] 0.2105263
> cpquery(bn_fit, (Survived == 0), (Age =='Child' & Sex=='male' & Pclass == 2))
[1] 0
> cpquery(bn_fit, (Survived == 0), (Age =='Child' & Sex=='male' & Pclass == 3))
[1] 0.6139896
```



```
> cpquery(bn_fit, (Survived == 0), (Age =='Adult' & Sex=='female' & Pclass == 1))
[1] 0.09114812
> cpquery(bn_fit, (Survived == 0), (Age =='Adult' & Sex=='female' & Pclass == 2))
[1] 0.186217
> cpquery(bn_fit, (Survived == 0), (Age =='Adult' & Sex=='female' & Pclass == 3))
[1] 0.3741438
> cpquery(bn_fit, (Survived == 0), (Age =='Adult' & Sex=='male' & Pclass == 1))
[1] 0.6431535
> cpquery(bn_fit, (Survived == 0), (Age =='Adult' & Sex=='male' & Pclass == 2))
[1] 0.8099662
> cpquery(bn_fit, (Survived == 0), (Age =='Adult' & Sex=='male' & Pclass == 3))
[1] 0.9073864
```

As you can see, the chances of survival is very low when a person is in 3rd class, adult and male.

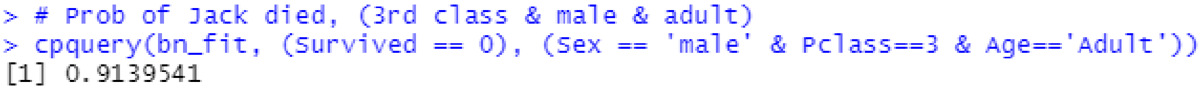
d. Prob of Rose survived, (1st class & female & Adult)



```
> # Prob of Rose survived, (1st class & female & Adult)
> cpquery(bn_fit, (Survived == 1), (Sex == 'female' & Pclass==1 & Age=='Adult'))
[1] 0.905417
```

We can conclude that since Rose is female, 1st class passenger, and adult, the possibility of survival is very high, which makes sense because she survived in the movie.

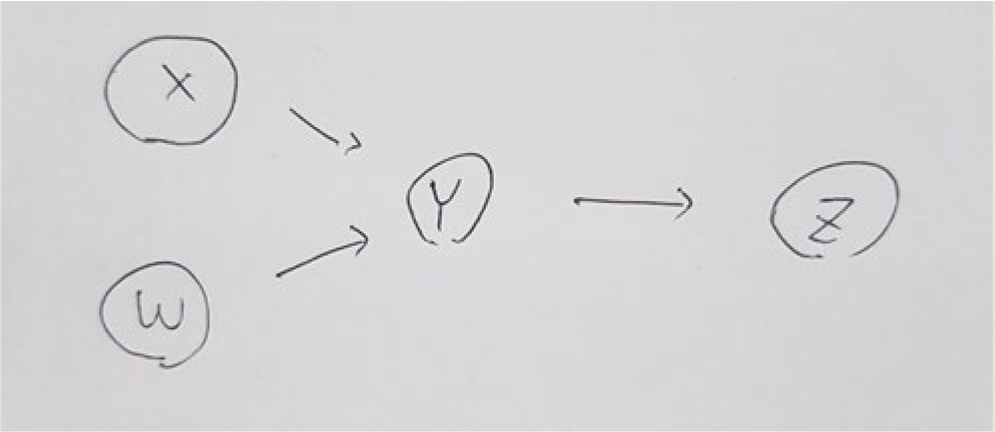
e. Prob of Jack died, (3rd class & male & adult)



```
> # Prob of Jack died, (3rd class & male & adult)
> cpquery(bn_fit, (Survived == 0), (Sex == 'male' & Pclass==3 & Age=='Adult'))
[1] 0.9139541
```

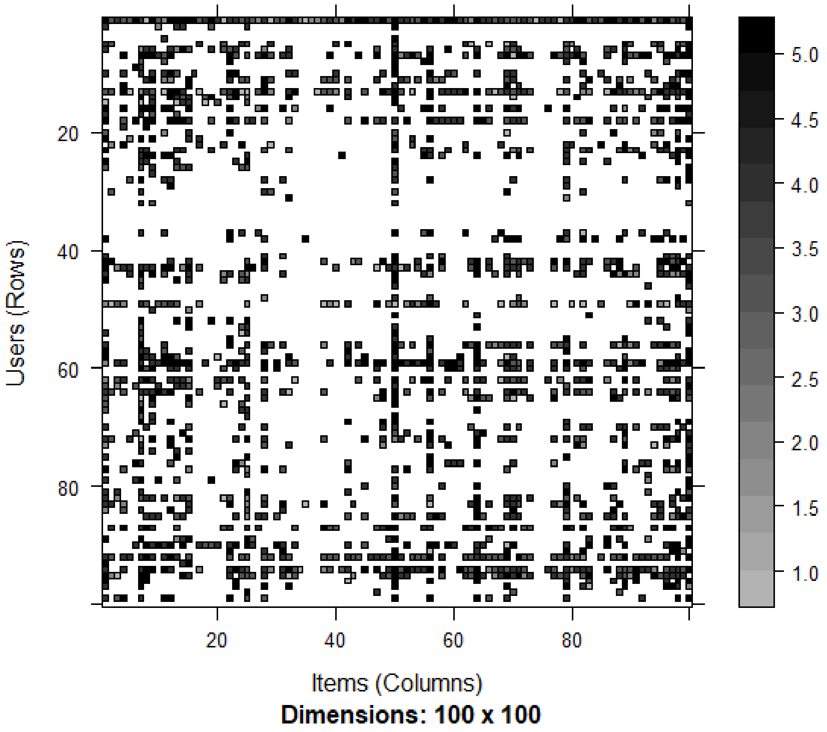
We can conclude that since Jack is male, 3rd class passenger and adult, the possibility of survival is very low, which makes sense because he died in the movie.

Question 3

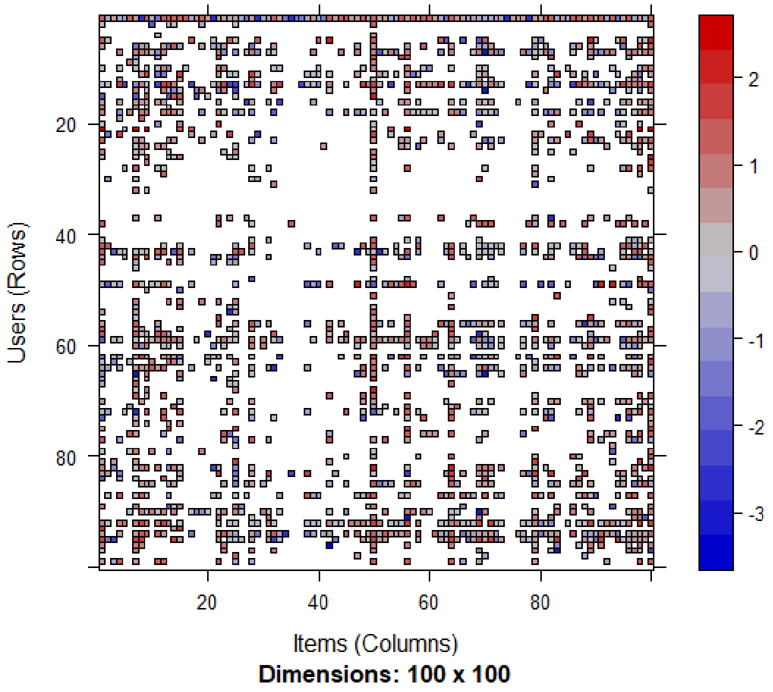


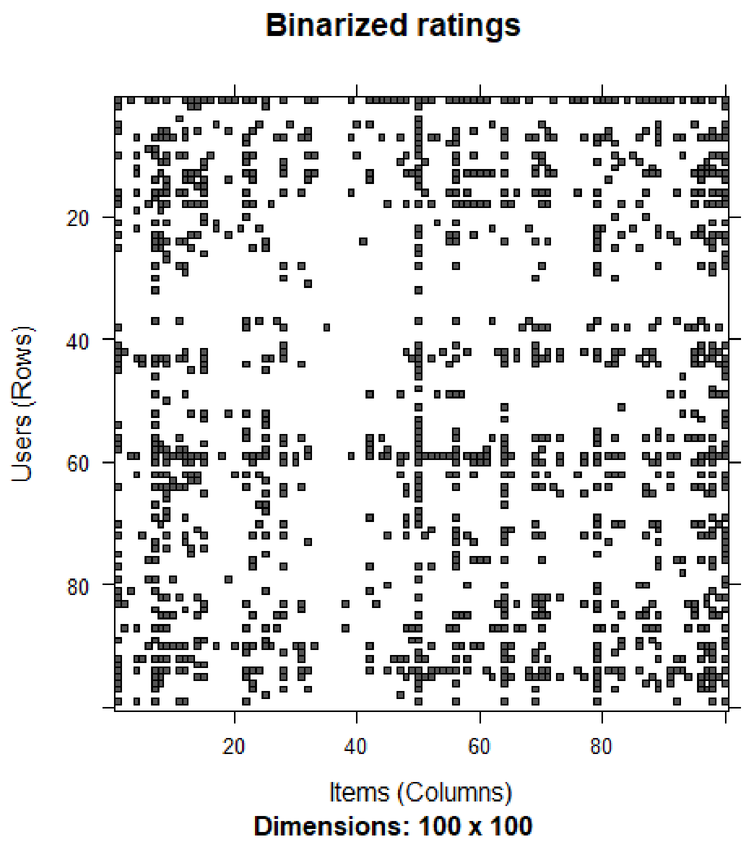
Question 4

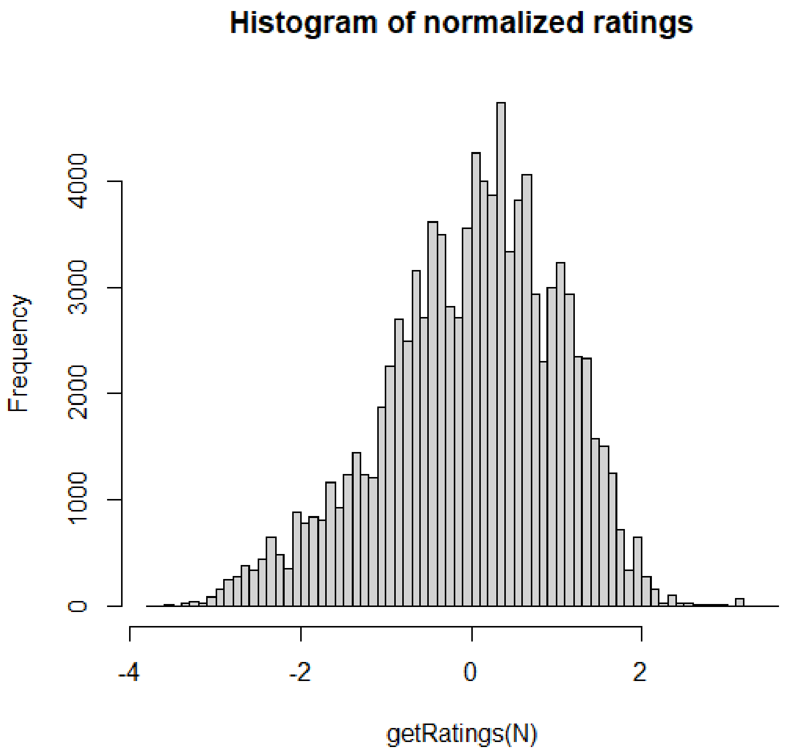
Below is 100X100 image of MovieLense data rating



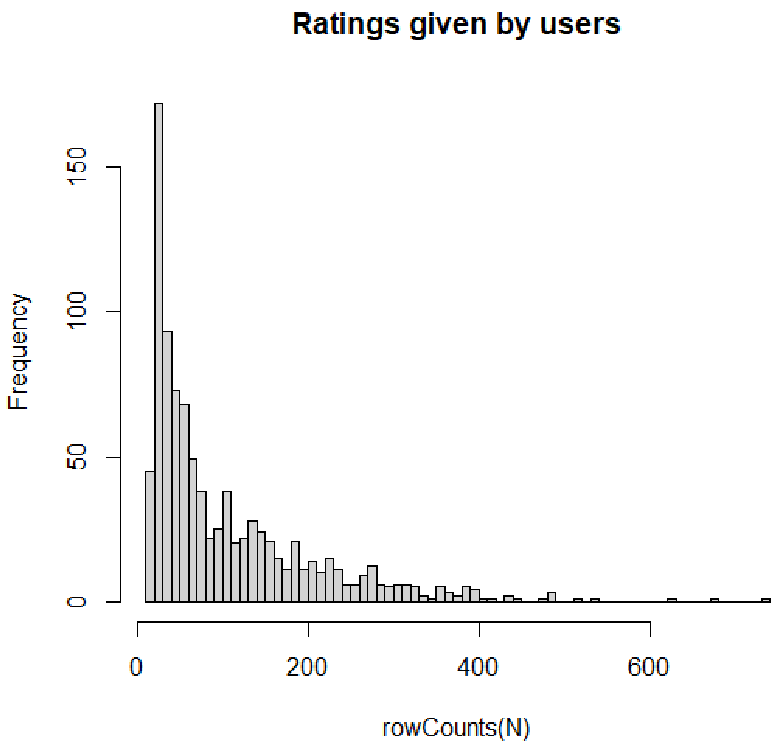
Normalized Data

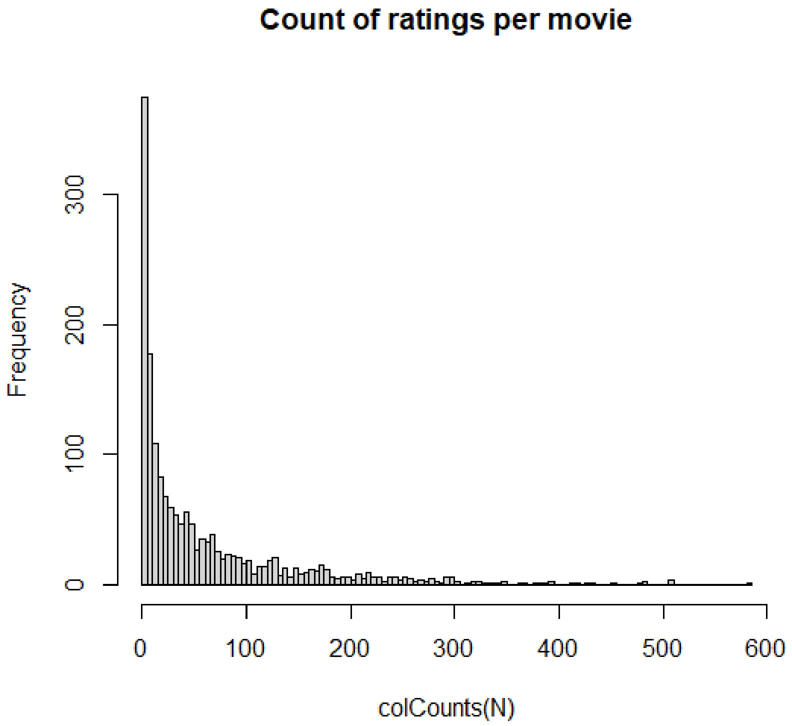


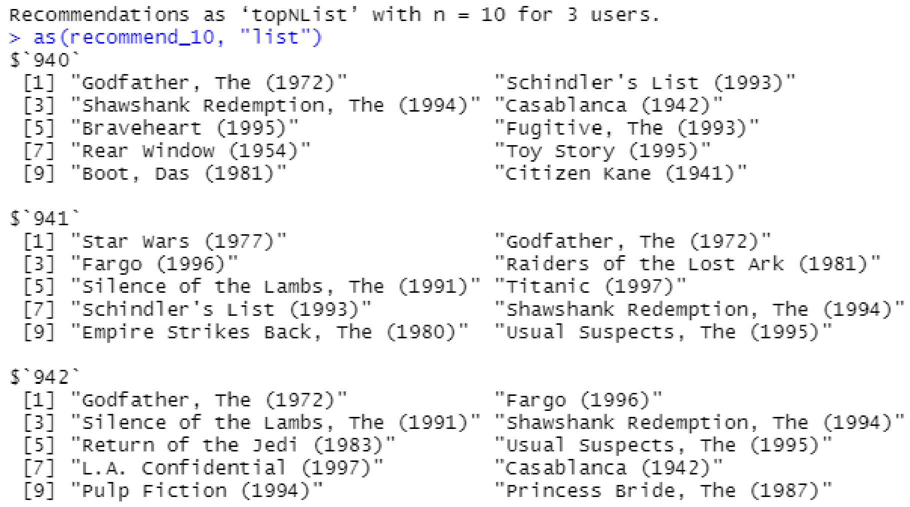




Most ratings are in the zero range, which tells that people do not give high ratings very often.

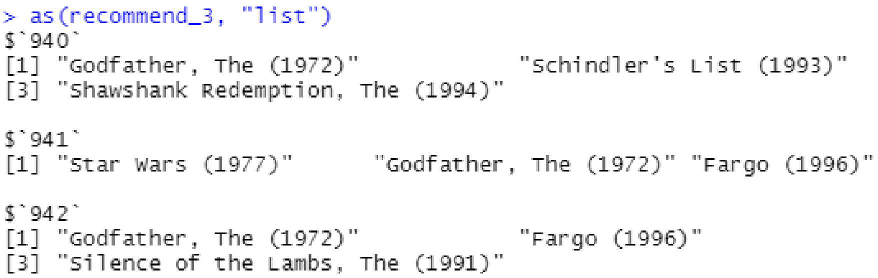

Ratings given by users
Frequency
0
50
100
150
0
200
400
600
rowCounts(N)


Count of ratings per movie
Frequency
0
100
200
300
0
100
200
300
400
500
600
colCounts(N)



```
Recommendations as 'topNList' with n = 10 for 3 users.
> as(recommend_10, "list")
$`940`
 [1] "Godfather, The (1972)"             "Schindler's List (1993)"
 [3] "Shawshank Redemption, The (1994)" "Casablanca (1942)"
 [5] "Braveheart (1995)"                 "Fugitive, The (1993)"
 [7] "Rear Window (1954)"                "Toy Story (1995)"
 [9] "Boot, Das (1981)"                  "Citizen Kane (1941)"

$`941`
 [1] "Star Wars (1977)"                  "Godfather, The (1972)"
 [3] "Fargo (1996)"                      "Raiders of the Lost Ark (1981)"
 [5] "Silence of the Lambs, The (1991)" "Titanic (1997)"
 [7] "Schindler's List (1993)"           "Shawshank Redemption, The (1994)"
 [9] "Empire Strikes Back, The (1980)"  "Usual Suspects, The (1995)"

$`942`
 [1] "Godfather, The (1972)"             "Fargo (1996)"
 [3] "Silence of the Lambs, The (1991)" "Shawshank Redemption, The (1994)"
 [5] "Return of the Jedi (1983)"         "Usual Suspects, The (1995)"
 [7] "L.A. Confidential (1997)"          "Casablanca (1942)"
 [9] "Pulp Fiction (1994)"               "Princess Bride, The (1987)"
```

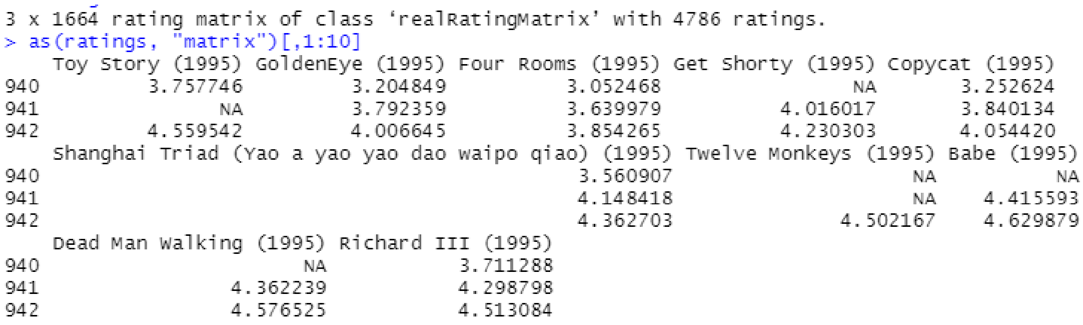
I picked random 3 users who rated the top movie ratings. Similar movies such as Godfather, toy story, fargo and etc can be seen from the data.



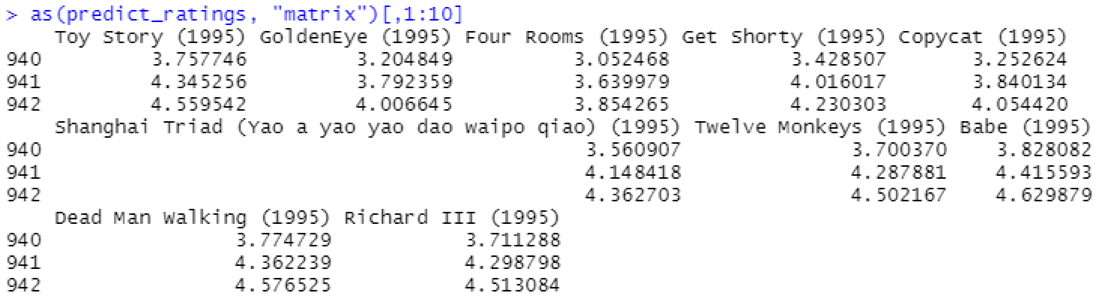
```
> as(recommend_3, "list")
$`940`
[1] "Godfather, The (1972)"             "Schindler's List (1993)"
[3] "Shawshank Redemption, The (1994)"

$`941`
[1] "Star Wars (1977)"      "Godfather, The (1972)" "Fargo (1996)"

$`942`
[1] "Godfather, The (1972)"             "Fargo (1996)"
[3] "Silence of the Lambs, The (1991)"
```

Among the top 3 recommendations, godfather appeared the most.

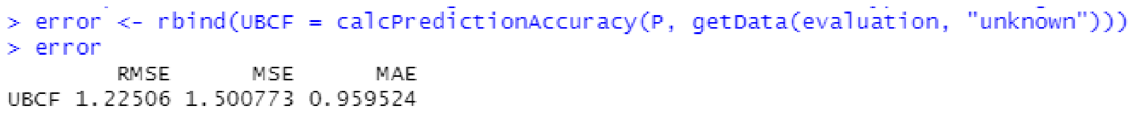
Predicted ratings with NA and without NA is shown below.



```
3 x 1664 rating matrix of class 'realRatingMatrix' with 4786 ratings.
> as(ratings, "matrix")[,1:10]
    Toy Story (1995) GoldenEye (1995) Four Rooms (1995) Get Shorty (1995) Copycat (1995)
940         3.757746         3.204849          3.052468                NA       3.252624
941               NA         3.792359          3.639979          4.016017       3.840134
942         4.559542         4.006645          3.854265          4.230303       4.054420
    Shanghai Triad (Yao a yao yao dao waipo qiao) (1995) Twelve Monkeys (1995) Babe (1995)
940                                             3.560907                    NA          NA
941                                             4.148418                    NA    4.415593
942                                             4.362703              4.502167    4.629879
    Dead Man Walking (1995) Richard III (1995)
940                      NA           3.711288
941                4.362239           4.298798
942                4.576525           4.513084
```



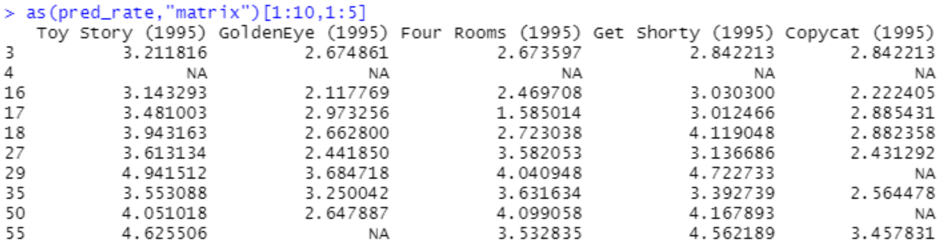
```
> as(predict_ratings, "matrix")[,1:10]
    Toy Story (1995) GoldenEye (1995) Four Rooms (1995) Get Shorty (1995) Copycat (1995)
940         3.757746         3.204849          3.052468          3.428507       3.252624
941         4.345256         3.792359          3.639979          4.016017       3.840134
942         4.559542         4.006645          3.854265          4.230303       4.054420
    Shanghai Triad (Yao a yao yao dao waipo qiao) (1995) Twelve Monkeys (1995) Babe (1995)
940                                             3.560907              3.700370    3.828082
941                                             4.148418              4.287881    4.415593
942                                             4.362703              4.502167    4.629879
    Dead Man Walking (1995) Richard III (1995)
940                3.774729           3.711288
941                4.362239           4.298798
942                4.576525           4.513084
```



```
> error <- rbind(UBCF = calcPredictionAccuracy(P, getData(evaluation, "unknown")))
> error
         RMSE      MSE      MAE
UBCF 1.22506 1.500773 0.959524
```

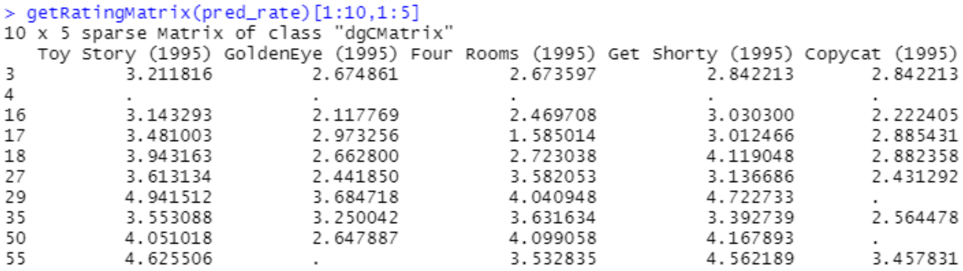
Question 5

Cross-validation is used.



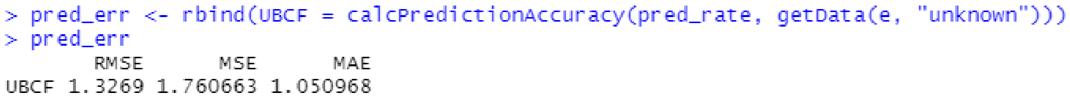
```
> as(pred_rate,"matrix")[1:10,1:5]
   Toy Story (1995) GoldenEye (1995) Four Rooms (1995) Get Shorty (1995) Copycat (1995)
3          3.211816         2.674861          2.673597          2.842213       2.842213
4                NA               NA                NA                NA             NA
16         3.143293         2.117769          2.469708          3.030300       2.222405
17         3.481003         2.973256          1.585014          3.012466       2.885431
18         3.943163         2.662800          2.723038          4.119048       2.882358
27         3.613134         2.441850          3.582053          3.136686       2.431292
29         4.941512         3.684718          4.040948          4.722733             NA
35         3.553088         3.250042          3.631634          3.392739       2.564478
50         4.051018         2.647887          4.099058          4.167893             NA
55         4.625506               NA          3.532835          4.562189       3.457831
```

This is the predicted rate of 10 users for 5 movies.



```
> getRatingMatrix(pred_rate)[1:10,1:5]
10 x 5 sparse Matrix of class "dgCMatrix"
   Toy Story (1995) GoldenEye (1995) Four Rooms (1995) Get Shorty (1995) Copycat (1995)
3          3.211816         2.674861          2.673597          2.842213       2.842213
4          .                .                 .                 .              .
16         3.143293         2.117769          2.469708          3.030300       2.222405
17         3.481003         2.973256          1.585014          3.012466       2.885431
18         3.943163         2.662800          2.723038          4.119048       2.882358
27         3.613134         2.441850          3.582053          3.136686       2.431292
29         4.941512         3.684718          4.040948          4.722733       .
35         3.553088         3.250042          3.631634          3.392739       2.564478
50         4.051018         2.647887          4.099058          4.167893       .
55         4.625506         .                 3.532835          4.562189       3.457831
```

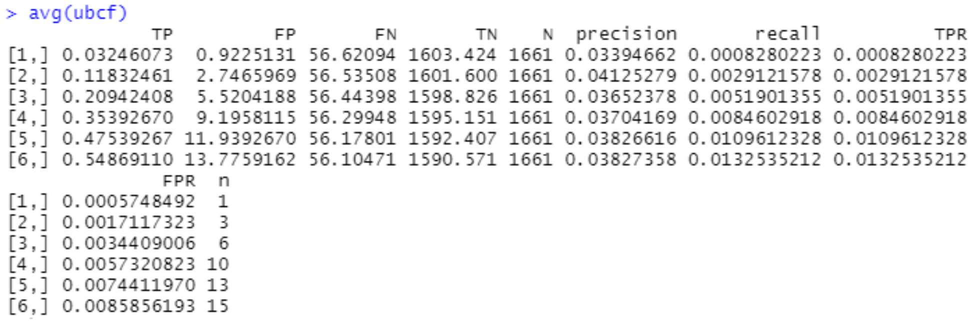
Predicted error is as follows.



```
> pred_err <- rbind(UBCF = calcPredictionAccuracy(pred_rate, getData(e, "unknown")))
> pred_err
        RMSE      MSE      MAE
UBCF 1.3269 1.760663 1.050968
```

Compared to question 4, question 5's RMSE, MSE, and MAE are all higher. This means that they are not performing well.

We do 5 folds and average the result



```
> avg(ubcf)
             TP         FP       FN       TN    N  precision       recall          TPR
[1,] 0.03246073  0.9225131 56.62094 1603.424 1661 0.03394662 0.0008280223 0.0008280223
[2,] 0.11832461  2.7465969 56.53508 1601.600 1661 0.04125279 0.0029121578 0.0029121578
[3,] 0.20942408  5.5204188 56.44398 1598.826 1661 0.03652378 0.0051901355 0.0051901355
[4,] 0.35392670  9.1958115 56.29948 1595.151 1661 0.03704169 0.0084602918 0.0084602918
[5,] 0.47539267 11.9392670 56.17801 1592.407 1661 0.03826616 0.0109612328 0.0109612328
[6,] 0.54869110 13.7759162 56.10471 1590.571 1661 0.03827358 0.0132535212 0.0132535212
              FPR  n
[1,] 0.0005748492  1
[2,] 0.0017117323  3
[3,] 0.0034409006  6
[4,] 0.0057320823 10
[5,] 0.0074411970 13
[6,] 0.0085856193 15
```