1.

D:

Based on the data we have, first check completely empty rows, remove completely empty rows.

Compute arithmetic mean of row and columns and fill it into empty cells for all movie ratings.

Create a regression model for each movie based on other 399 movies' rating.

Save COD result to a .csv file.

Compute average COD of those 400 simple linear regression models.

Draw histogram for COD values.

Find how many movies regression models has COD values lower than 0.5 and 0.6.

Y:

Remove completely empty rows: because we cannot blend completely empty rows by compute arithmetic mean.

Compute arithmetic mean to fill empty cell is the only method we could choose.

Based on other 399 movie can find best fit regression model.

Save COD result to a .csv file so we may use it later.

Drawing a histogram of COD values can give us an intuitive and easy-to-understand result.

Find how many movies regression models has COD values lower than 0.5 and 0.6 can gave us a better understanding of COD value distribution.

F:

Row 896 is the only completely empty row.

The best fit regression model for 400 movies has been found.

The average COD of the 400 simple linear regression models = 0.424.

Number of COD values below 0.5: 265

Number of COD values below 0.6: 339

Number of COD values below 0.7: 391

According to *Figure 1*, the histogram shows most of movie has less than 0.5 COD value.

According to *Figure 2*, the table shows 10 most predictable movies and 10 least predictable movies.

The first 8 movies in top 10 most predictable movies are highly related to each other (They become a group in 2, in other words they can predict each other.), but top 10 less predictable movies have not relations to each other.

A:

According to all these information and findings, we may conclude that most of movies seems like have no relations to each other (339/400 have COD values lower than 0.6, 9/400 has COD value higher than 0.7). But the reason due to this result might be the dataset is not good enough (Too much null value exists so we must use arithmetic mean to fill the empty.), or we need a better model to do the prediction.

2.

D:

Add row 475, 476, 477 as additional predictor.

Remove any empty cell row in column 475-477.

Use 399+3 columns of movie rating data build regression model, just like we did in previous part.

Save result and compare with previous part.

Create scatter plot.

Y:

The data in 475-477 columns can not be blend-and-fill into empty cells, because these values are not ratings, they have special meaning.

Adding 3 more columns and do regression again to analyze if there are any connections between those values.

F:

The R^2 value of regression model has changed very slightly. See *Figure 3.*

The scatter plot shows the relationship between COD and R^2

A:

Adding more predictors to the regression model to build a multiple regression model does not make the result change very much (Actually we could even ignore those value changes, because they are too small.). And if we draw a fit line on *Figure 4*(The scatter plot), the line will be cross through the plot from bottom left to the top right (slope is close to 1 which means result not changes with more predictors.), this could explain that even we add more predictors to the regression model the result would not changes too much.

3.

A:

The best hyperparameter and the corresponding betas are shown below. There are 30 betas corresponding to 30 movies to predict, each beta is a 10-d vector.(shown in Figure 5)

F:

Larger hyperparameter corresponds to smaller betas, which is expected, because a larger weight is assigned to the regularization term.

Moreover, larger hyperparameter corresponds to larger RMSE on training set, which indicates underfitting. This is also expected because regularization term is assigned more weight, and fitness on training set is less emphasized.

Y:

Best hyperparameter is found in the middle of the range 1 to 100 (best hyperparameter = 69). Since the best hyperparameter found is at the boarder of the range (1 or 100), it indicates that our range of candidate hyperparameters is reasonable. (The best hyperparameter varies every run because our train test splits are random.) The best hyperparameter achieves a good fit on the training set while keeping the model less complex to avoid overfitting.

D:

I started by sorting the 400 movies by their COD values and then picked out the 185th-215th movies. After that, I randomly chose 10 movies (making sure they didn't overlap with the previous thirty movies). Next, I split the data into a training set and a test set using an 80/20 split ratio. The split was performed randomly. I then applied ridge regression to the training data, selecting hyperparameter from the set {1, 2, … ,100}. I calculated the RMSE for each of the thirty movies on the test set and used the average RMSE across all the movies to determine the optimal value of the hyperparameter. Finally, I found the corresponding beta value with the optimal hyperparameter.

4.

A:

The best hyperparameter and the corresponding betas are shown below. There are 30 betas corresponding to 30 movies to predict, each beta is a 10-d vector. (shown in Figure 6)

F:

Larger hyperparameter corresponds to smaller betas, which is expected, because a larger weight is assigned to the regularization term.

Moreover, larger hyperparameter corresponds to larger RMSE on training set, which indicates underfitting. This is also expected because regularization term is assigned more weight, and fitness on training set is less emphasized.

Y:

Best hyperparameter is found in the middle of the range 0.001 to 0.01(best hyperparameter = 0.0034). Since the best hyperparameter found is at the boarder of the range (0.001 or 0.01), it indicates that our range of candidate hyperparameters is reasonable. (The best hyperparameter varies every run because our train test splits are random.) The best hyperparameter achieves a good fit on the training set while keeping the model less complex to avoid overfitting.

D:

I started by sorting the 400 movies by their COD values and then picked out the 185th-215th movies. After that, I randomly chose 10 movies (making sure they didn't overlap with the previous thirty movies). Next, I split the data into a training set and a test set using an 80/20 split ratio. The split was performed randomly. I then applied ridge regression to the training data, selecting hyperparameter from the set {0.001, 0.002, … ,0.01}. I calculated the RMSE for each of the thirty movies on the test set and used the average RMSE across all the movies to determine the optimal value of the hyperparameter. Finally, I found the corresponding beta value with the optimal hyperparameter.

5.

D: The data is first processed through a 50/50 blend algorithm to process the missing value in the dataset, and the average movie rating of each user in the unprocessed dataset is calculated (X). After selecting the target movie, a logistic regression model is applied to predict whether each user will like the movie (Y) using their average movie rating. In building the logistic regression model and applying cross-validation, the system reported an error for X because it contained missing value, first replacing the missing value with the mean, then I made a version with the missing value processed using a 50/50 hybrid algorithm and I did another version to get rid of all the missing data.

Y: The logistic regression model is used to be able to do the classification. In processing X because the question states that non-imputed data should be used so I directly replace the missing value with the average value because X itself represents the average movie rating of each user so I use the average value instead.

F: Images of the ROCs of the four movies are presented on figure 7. The beta for each movies shown in figure 8.

A: Higher beta values indicate that there is a stronger relationship between how much users like movies in general and how likely they are to like those particular movies.Higher AUC indicates better model performance, but there are two AUC's where the image looks like a right triangle, which may be due to the fact that there aren't any false positive examples and there aren't any false negative examples. This could be due to the simplicity of the dataset.

Extra: A linear regression model on user data is used to predict subsequent user ratings for a particular movie series.

D: The original dataset was split into 80/20 using train_test_split. The model was built using linear regression and R^2 and MSE were calculated from the test data.

Y: Utilizing train_test_split can help evaluate model performance. A high R^2 can indicate that the model fits the data well. While MSE reflects the prediction error of the model, a smaller error indicates a better prediction performance of the model.

F: The R^2 and MSE of these movie series are shown on figure 9.

A: The Star Wars: Episode 6 model has high performance in predicting Star Wars: Episode 6. The Batman & Robin model has poor performance in predicting Batman & Robin. Using some of the user ratings to predict subsequent user ratings can be used to simulate the movie ratings obtained after a movie is ordered to predict subsequent movie ratings after its release.

Appendix:

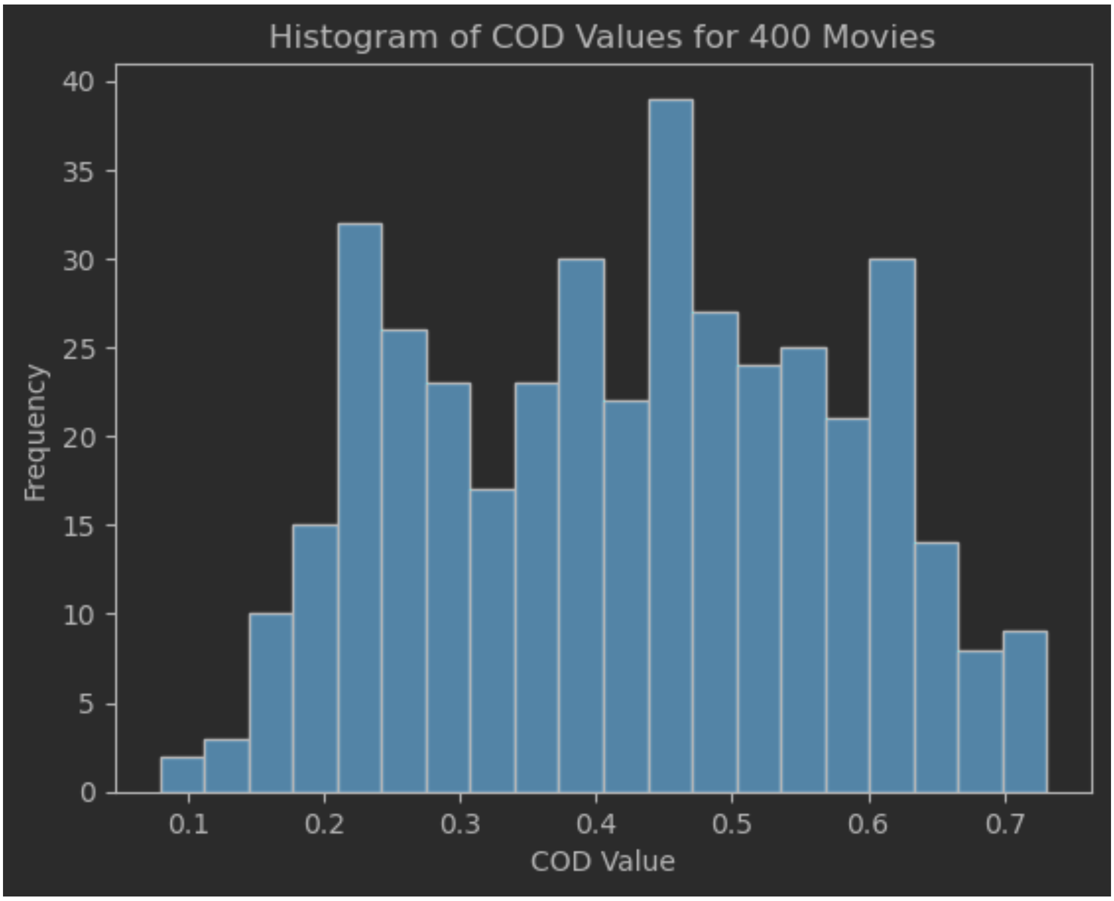
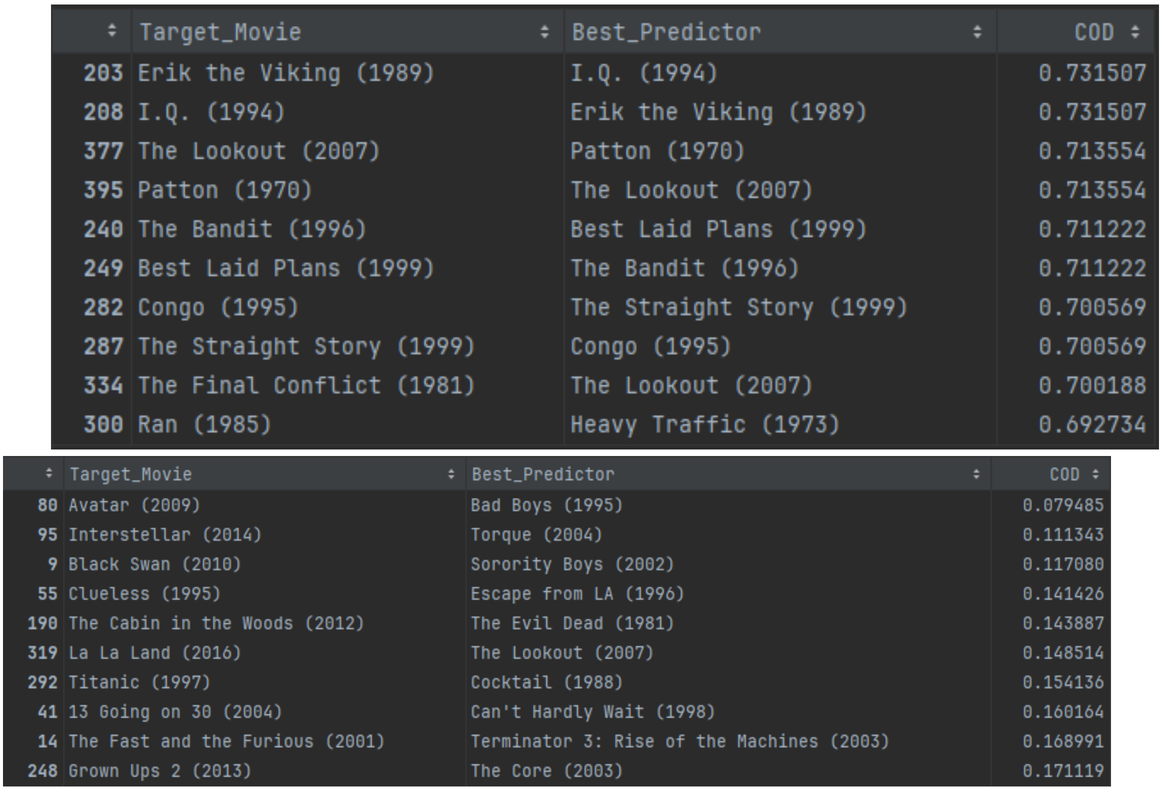


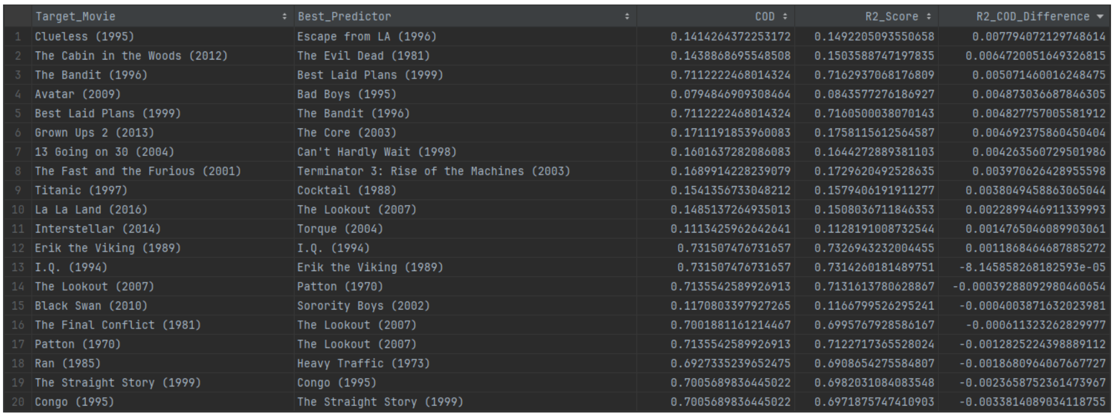
Figure 1



	Target_Movie	Best_Predictor	COD
203	Erik the Viking (1989)	I.Q. (1994)	0.731507
208	I.Q. (1994)	Erik the Viking (1989)	0.731507
377	The Lookout (2007)	Patton (1970)	0.713554
395	Patton (1970)	The Lookout (2007)	0.713554
240	The Bandit (1996)	Best Laid Plans (1999)	0.711222
249	Best Laid Plans (1999)	The Bandit (1996)	0.711222
282	Congo (1995)	The Straight Story (1999)	0.700569
287	The Straight Story (1999)	Congo (1995)	0.700569
334	The Final Conflict (1981)	The Lookout (2007)	0.700188
300	Ran (1985)	Heavy Traffic (1973)	0.692734

	Target_Movie	Best_Predictor	COD
80	Avatar (2009)	Bad Boys (1995)	0.079485
95	Interstellar (2014)	Torque (2004)	0.111343
9	Black Swan (2010)	Sorority Boys (2002)	0.117080
55	Clueless (1995)	Escape from LA (1996)	0.141426
190	The Cabin in the Woods (2012)	The Evil Dead (1981)	0.143887
319	La La Land (2016)	The Lookout (2007)	0.148514
292	Titanic (1997)	Cocktail (1988)	0.154136
41	13 Going on 30 (2004)	Can't Hardly Wait (1998)	0.160164
14	The Fast and the Furious (2001)	Terminator 3: Rise of the Machines (2003)	0.168991
248	Grown Ups 2 (2013)	The Core (2003)	0.171119

Figure 2



	Target_Movie	Best_Predictor	COD	R2_Score	R2_COD_Difference
1	Clueless (1995)	Escape from LA (1996)	0.1414264372253172	0.1492205093550658	0.007794072129748614
2	The Cabin in the Woods (2012)	The Evil Dead (1981)	0.1438868695548508	0.1503588747197835	0.0064720051649326815
3	The Bandit (1996)	Best Laid Plans (1999)	0.7112222468014324	0.7162937068176809	0.005071460016248475
4	Avatar (2009)	Bad Boys (1995)	0.0794846909308464	0.0843577276186927	0.004873036687846305
5	Best Laid Plans (1999)	The Bandit (1996)	0.7112222468014324	0.7160500038070143	0.004827757005581912
6	Grown Ups 2 (2013)	The Core (2003)	0.1711191853960083	0.1758115612564587	0.004692375860450404
7	13 Going on 30 (2004)	Can't Hardly Wait (1998)	0.1601637282086083	0.1644272889381103	0.004263560729501986
8	The Fast and the Furious (2001)	Terminator 3: Rise of the Machines (2003)	0.1689914228239079	0.1729620492528635	0.003970626428955598
9	Titanic (1997)	Cocktail (1988)	0.1541356733048212	0.1579406191911277	0.0038049458863065044
10	La La Land (2016)	The Lookout (2007)	0.1485137264935013	0.1508036711846353	0.0022899446911339993
11	Interstellar (2014)	Torque (2004)	0.1113425962642641	0.1128191008732544	0.0014765046089903061
12	Erik the Viking (1989)	I.Q. (1994)	0.731507476731657	0.7326943232004455	0.0011868464687885272
13	I.Q. (1994)	Erik the Viking (1989)	0.731507476731657	0.7314260181489751	-8.145858268182593e-05
14	The Lookout (2007)	Patton (1970)	0.7135542589926913	0.7131613780628867	-0.00039288092980460654
15	Black Swan (2010)	Sorority Boys (2002)	0.1170803397927265	0.1166799526295241	-0.0004003871632023981
16	The Final Conflict (1981)	The Lookout (2007)	0.7001881161214467	0.6995767928586167	-0.000611323262829977
17	Patton (1970)	The Lookout (2007)	0.7135542589926913	0.7122717365528024	-0.0012825224398889112
18	Ran (1985)	Heavy Traffic (1973)	0.6927335239652475	0.6908654275584807	-0.0018680964067667727
19	The Straight Story (1999)	Congo (1995)	0.7005689836445022	0.6982031084083548	-0.0023658752361473967
20	Congo (1995)	The Straight Story (1999)	0.7005689836445022	0.6971875747410903	-0.0033814089034118755

Figure 3

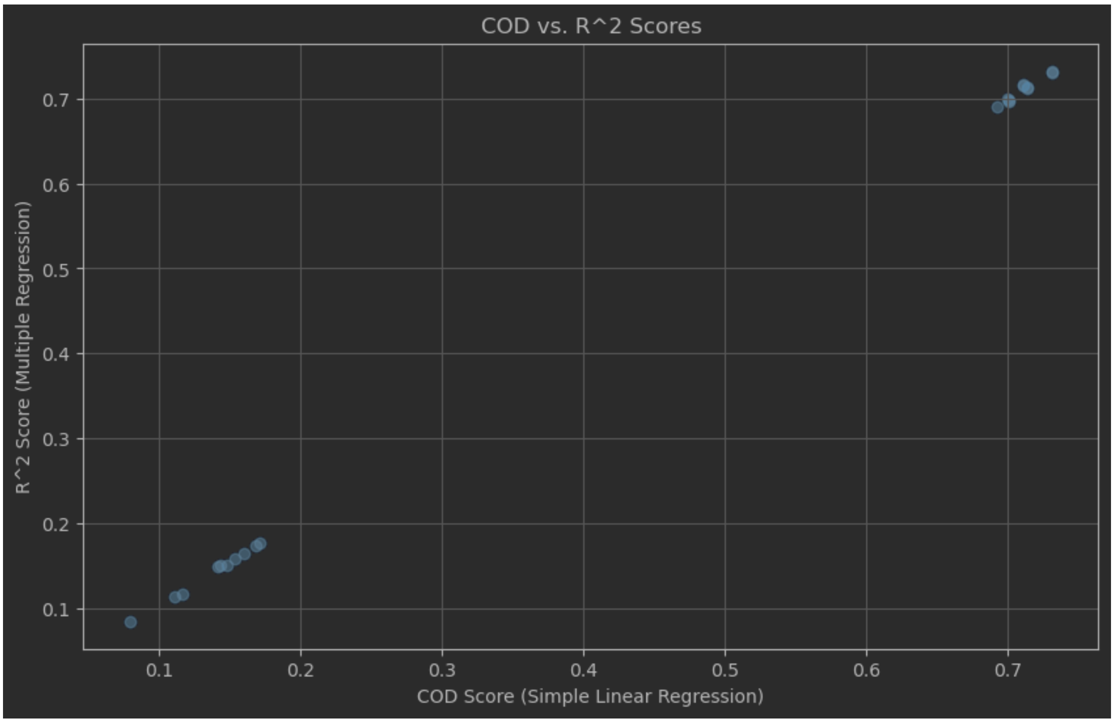
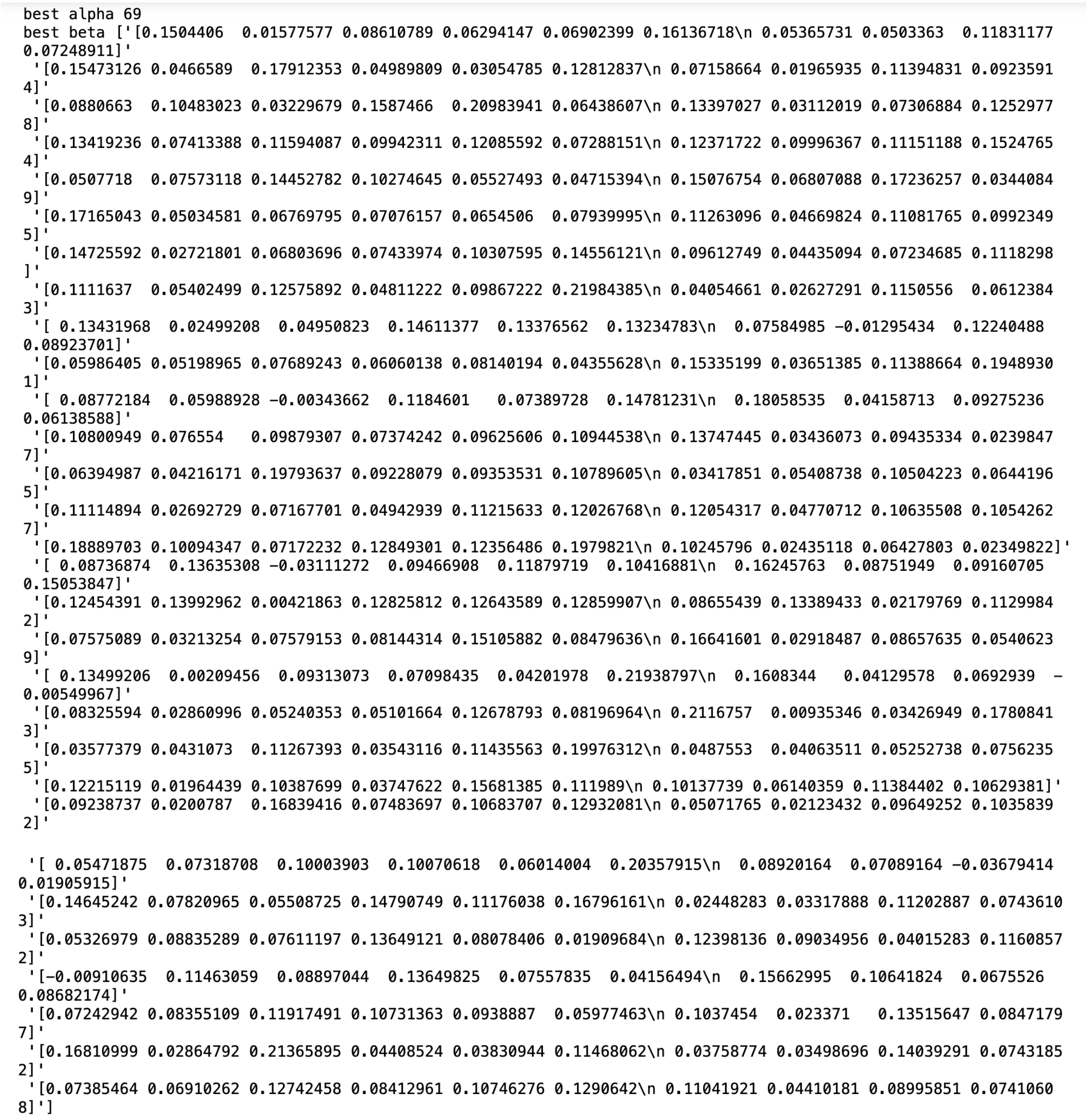


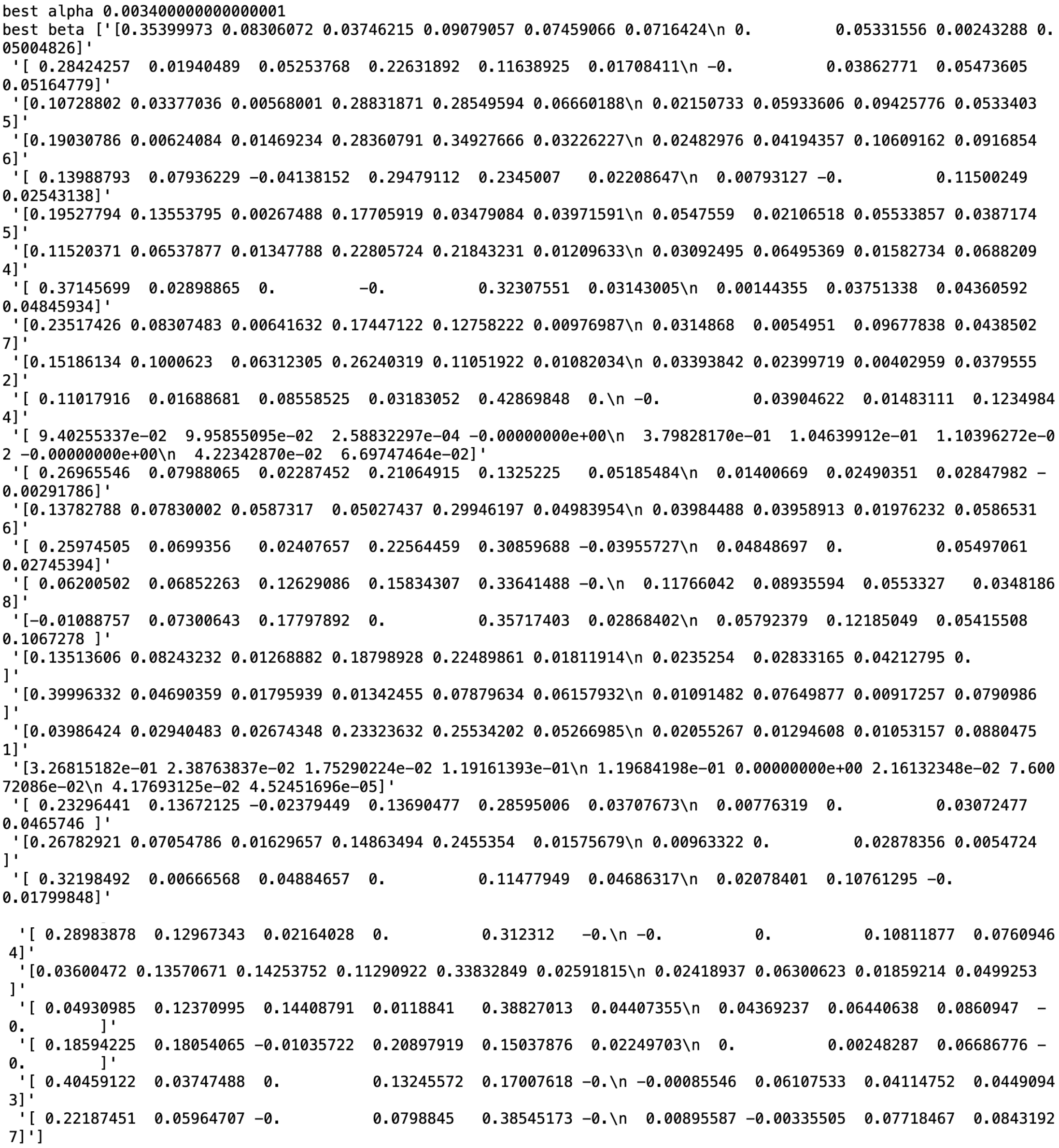
Figure 4



```
best alpha 69
best beta ['[0.1504406  0.01577577 0.08610789 0.06294147 0.06902399 0.16136718\n 0.05365731 0.0503363  0.11831177
0.07248911]'
 '[0.15473126 0.0466589  0.17912353 0.04989809 0.03054785 0.12812837\n 0.07158664 0.01965935 0.11394831 0.0923591
4]'
 '[0.0880663  0.10483023 0.03229679 0.1587466  0.20983941 0.06438607\n 0.13397027 0.03112019 0.07306884 0.1252977
8]'
 '[0.13419236 0.07413388 0.11594087 0.09942311 0.12085592 0.07288151\n 0.12371722 0.09996367 0.11151188 0.1524765
4]'
 '[0.0507718  0.07573118 0.14452782 0.10274645 0.05527493 0.04715394\n 0.15076754 0.06807088 0.17236257 0.0344084
9]'
 '[0.17165043 0.05034581 0.06769795 0.07076157 0.0654506  0.07939995\n 0.11263096 0.04669824 0.11081765 0.0992349
5]'
 '[0.14725592 0.02721801 0.06803696 0.07433974 0.10307595 0.14556121\n 0.09612749 0.04435094 0.07234685 0.1118298
]'
 '[0.1111637  0.05402499 0.12575892 0.04811222 0.09867222 0.21984385\n 0.04054661 0.02627291 0.1150556  0.0612384
3]'
 '[ 0.13431968  0.02499208  0.04950823  0.14611377  0.13376562  0.13234783\n  0.07584985 -0.01295434  0.12240488
0.08923701]'
 '[0.05986405 0.05198965 0.07689243 0.06060138 0.08140194 0.04355628\n 0.15335199 0.03651385 0.11388664 0.1948930
1]'
 '[ 0.08772184  0.05988928 -0.00343662  0.1184601   0.07389728  0.14781231\n  0.18058535  0.04158713  0.09275236
0.06138588]'
 '[0.10800949 0.076554   0.09879307 0.07374242 0.09625606 0.10944538\n 0.13747445 0.03436073 0.09435334 0.0239847
7]'
 '[0.06394987 0.04216171 0.19793637 0.09228079 0.09353531 0.10789605\n 0.03417851 0.05408738 0.10504223 0.0644196
5]'
 '[0.11114894 0.02692729 0.07167701 0.04942939 0.11215633 0.12026768\n 0.12054317 0.04770712 0.10635508 0.1054262
7]'
 '[0.18889703 0.10094347 0.07172232 0.12849301 0.12356486 0.1979821\n 0.10245796 0.02435118 0.06427803 0.02349822]'
 '[ 0.08736874  0.13635308 -0.03111272  0.09466908  0.11879719  0.10416881\n  0.16245763  0.08751949  0.09160705
0.15053847]'
 '[0.12454391 0.13992962 0.00421863 0.12825812 0.12643589 0.12859907\n 0.08655439 0.13389433 0.02179769 0.1129984
2]'
 '[0.07575089 0.03213254 0.07579153 0.08144314 0.15105882 0.08479636\n 0.16641601 0.02918487 0.08657635 0.0540623
9]'
 '[ 0.13499206  0.00209456  0.09313073  0.07098435  0.04201978  0.21938797\n  0.1608344   0.04129578  0.0692939  -
0.00549967]'
 '[0.08325594 0.02860996 0.05240353 0.05101664 0.12678793 0.08196964\n 0.2116757  0.00935346 0.03426949 0.1780841
3]'
 '[0.03577379 0.0431073  0.11267393 0.03543116 0.11435563 0.19976312\n 0.0487553  0.04063511 0.05252738 0.0756235
5]'
 '[0.12215119 0.01964439 0.10387699 0.03747622 0.15681385 0.111989\n 0.10137739 0.06140359 0.11384402 0.10629381]'
 '[0.09238737 0.0200787  0.16839416 0.07483697 0.10683707 0.12932081\n 0.05071765 0.02123432 0.09649252 0.1035839
2]'


 '[ 0.05471875  0.07318708  0.10003903  0.10070618  0.06014004  0.20357915\n  0.08920164  0.07089164 -0.03679414
0.01905915]'
 '[0.14645242 0.07820965 0.05508725 0.14790749 0.11176038 0.16796161\n 0.02448283 0.03317888 0.11202887 0.0743610
3]'
 '[0.05326979 0.08835289 0.07611197 0.13649121 0.08078406 0.01909684\n 0.12398136 0.09034956 0.04015283 0.1160857
2]'
 '[-0.00910635  0.11463059  0.08897044  0.13649825  0.07557835  0.04156494\n  0.15662995  0.10641824  0.0675526
0.08682174]'
 '[0.07242942 0.08355109 0.11917491 0.10731363 0.0938887  0.05977463\n 0.1037454  0.023371   0.13515647 0.0847179
7]'
 '[0.16810999 0.02864792 0.21365895 0.04408524 0.03830944 0.11468062\n 0.03758774 0.03498696 0.14039291 0.0743185
2]'
 '[0.07385464 0.06910262 0.12742458 0.08412961 0.10746276 0.1290642\n 0.11041921 0.04410181 0.08995851 0.0741060
8]']
```

Figure 5



```
best alpha 0.0034000000000000001
best beta ['[0.35399973 0.08306072 0.03746215 0.09079057 0.07459066 0.0716424\n 0.         0.05331556 0.00243288 0.
05004826]'
 '[ 0.28424257  0.01940489  0.05253768  0.22631892  0.11638925  0.01708411\n -0.          0.03862771  0.05473605
0.05164779]'
 '[0.10728802 0.03377036 0.00568001 0.28831871 0.28549594 0.06660188\n 0.02150733 0.05933606 0.09425776 0.0533403
5]'
 '[0.19030786 0.00624084 0.01469234 0.28360791 0.34927666 0.03226227\n 0.02482976 0.04194357 0.10609162 0.0916854
6]'
 '[ 0.13988793  0.07936229 -0.04138152  0.29479112  0.2345007   0.02208647\n  0.00793127 -0.          0.11500249
0.02543138]'
 '[0.19527794 0.13553795 0.00267488 0.17705919 0.03479084 0.03971591\n 0.0547559  0.02106518 0.05533857 0.0387174
5]'
 '[0.11520371 0.06537877 0.01347788 0.22805724 0.21843231 0.01209633\n 0.03092495 0.06495369 0.01582734 0.0688209
4]'
 '[ 0.37145699  0.02898865  0.         -0.          0.32307551  0.03143005\n  0.00144355  0.03751338  0.04360592
0.04845934]'
 '[0.23517426 0.08307483 0.00641632 0.17447122 0.12758222 0.00976987\n 0.0314868  0.0054951  0.09677838 0.0438502
7]'
 '[0.15186134 0.1000623  0.06312305 0.26240319 0.11051922 0.01082034\n 0.03393842 0.02399719 0.00402959 0.0379555
2]'
 '[ 0.11017916  0.01688681  0.08558525  0.03183052  0.42869848  0.\n -0.          0.03904622  0.01483111  0.1234984
4]'
 '[ 9.40255337e-02  9.95855095e-02  2.58832297e-04 -0.00000000e+00\n  3.79828170e-01  1.04639912e-01  1.10396272e-0
2 -0.00000000e+00\n  4.22342870e-02  6.69747464e-02]'
 '[ 0.26965546  0.07988065  0.02287452  0.21064915  0.1325225   0.05185484\n  0.01400669  0.02490351  0.02847982 -
0.00291786]'
 '[0.13782788 0.07830002 0.0587317  0.05027437 0.29946197 0.04983954\n 0.03984488 0.03958913 0.01976232 0.0586531
6]'
 '[ 0.25974505  0.0699356   0.02407657  0.22564459  0.30859688 -0.03955727\n  0.04848697  0.          0.05497061
0.02745394]'
 '[ 0.06200502  0.06852263  0.12629086  0.15834307  0.33641488 -0.\n  0.11766042  0.08935594  0.0553327   0.0348186
8]'
 '[-0.01088757  0.07300643  0.17797892  0.          0.35717403  0.02868402\n  0.05792379  0.12185049  0.05415508
0.1067278 ]'
 '[0.13513606 0.08243232 0.01268882 0.18798928 0.22489861 0.01811914\n 0.0235254  0.02833165 0.04212795 0.
]'
 '[0.39996332 0.04690359 0.01795939 0.01342455 0.07879634 0.06157932\n 0.01091482 0.07649877 0.00917257 0.0790986
]'
 '[0.03986424 0.02940483 0.02674348 0.23323632 0.25534202 0.05266985\n 0.02055267 0.01294608 0.01053157 0.0880475
1]'
 '[3.26815182e-01 2.38763837e-02 1.75290224e-02 1.19161393e-01\n 1.19684198e-01 0.00000000e+00 2.16132348e-02 7.600
72086e-02\n 4.17693125e-02 4.52451696e-05]'
 '[ 0.23296441  0.13672125 -0.02379449  0.13690477  0.28595006  0.03707673\n  0.00776319  0.          0.03072477
0.0465746 ]'
 '[0.26782921 0.07054786 0.01629657 0.14863494 0.2455354  0.01575679\n 0.00963322 0.         0.02878356 0.0054724
]'
 '[ 0.32198492  0.00666568  0.04884657  0.          0.11477949  0.04686317\n  0.02078401  0.10761295 -0.
0.01799848]'

  '[ 0.28983878  0.12967343  0.02164028  0.          0.312312   -0.\n -0.          0.          0.10811877  0.0760946
 4]'
  '[0.03600472 0.13570671 0.14253752 0.11290922 0.33832849 0.02591815\n 0.02418937 0.06300623 0.01859214 0.0499253
 ]'
  '[ 0.04930985  0.12370995  0.14408791  0.0118841   0.38827013  0.04407355\n  0.04369237  0.06440638  0.0860947  -
 0.        ]'
  '[ 0.18594225  0.18054065 -0.01035722  0.20897919  0.15037876  0.02249703\n  0.          0.00248287  0.0668676 -
 0.        ]'
  '[ 0.40459122  0.03747488  0.          0.13245572  0.17007618 -0.\n -0.00085546  0.06107533  0.04114752  0.0449094
 3]'
  '[ 0.22187451  0.05964707 -0.          0.0798845   0.38545173 -0.\n  0.00895587 -0.00335505  0.07718467  0.0843192
 7]']
```

Figure 6

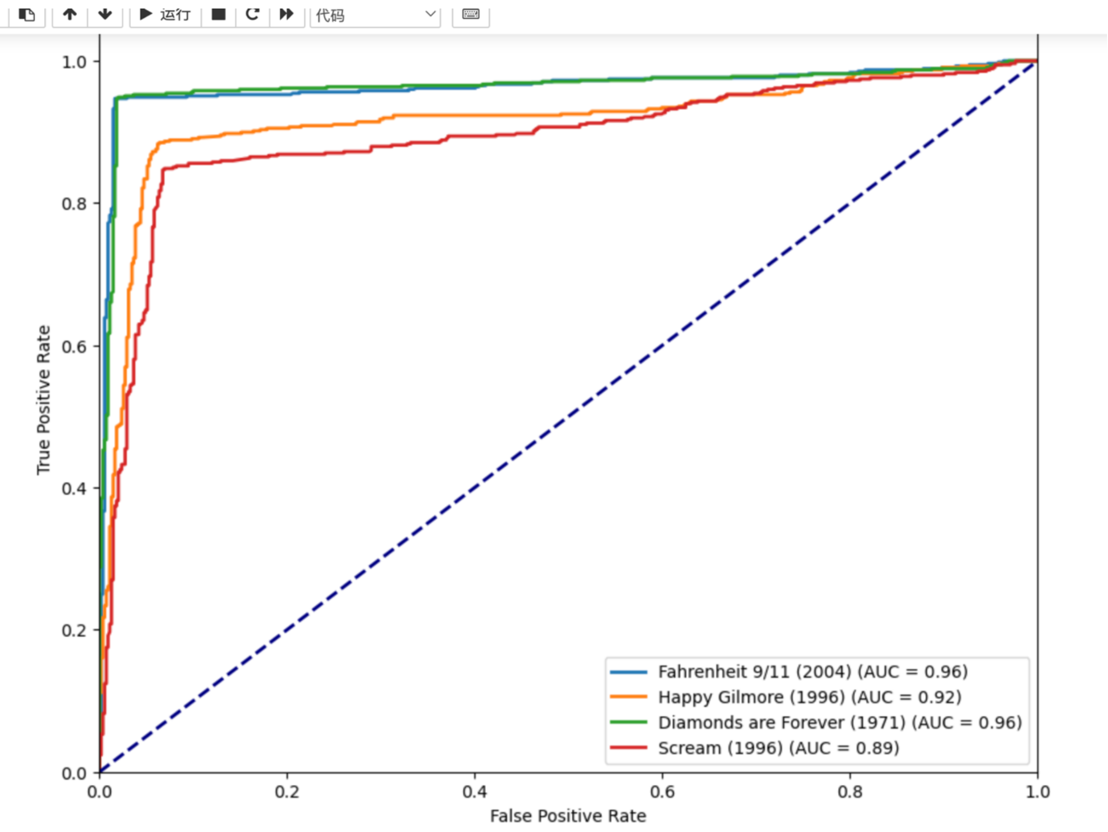
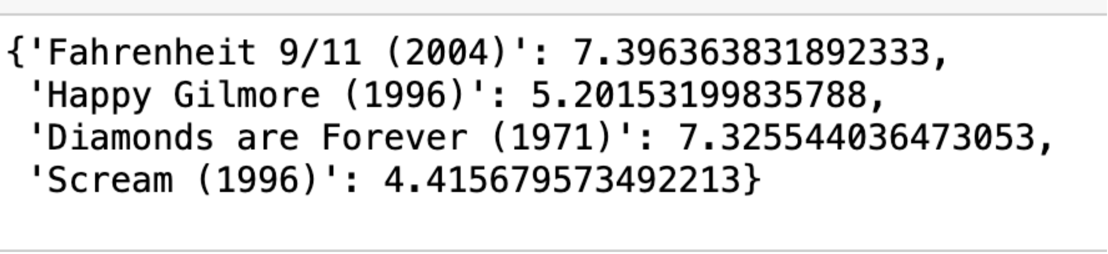


Figure 7



```
{'Fahrenheit 9/11 (2004)': 7.396363831892333,
 'Happy Gilmore (1996)': 5.20153199835788,
 'Diamonds are Forever (1971)': 7.325544036473053,
 'Scream (1996)': 4.415679573492213}
```

Figure 8

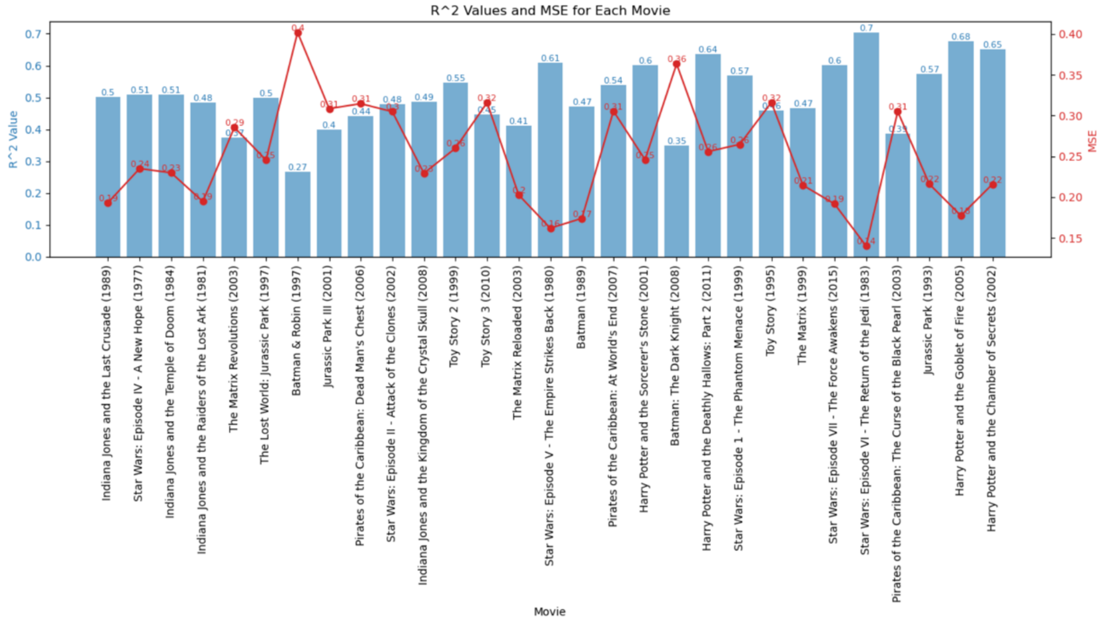


Figure 9