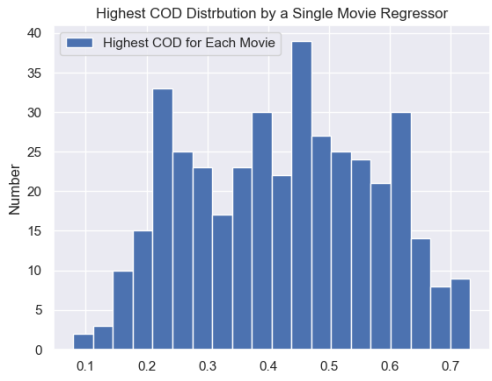
Data Preprocessing

As suggested by the project guide, I use the average of column mean and row mean to fill in the missing data. However, it is noteworthy that user 897 doesn't have a rating for any movie. In this case, I simply use column mean to fill in the missing data here. For questions 1-4, all the data I use are the data that has been imputed.

Q1

*For each of the 400 movies, use a simple linear regression model to predict the ratings. Use the ratings of the *other* 399 movies in the dataset to predict the ratings of each movie (that means you'll have to build 399 models for each of the 400 movies). For each of the 400 movies, find the movie that predicts ratings the best. Then report the average COD of those 400 simple linear regression models. Please include a histogram of these 400 COD values and a table with the 10 movies that are most easily predicted from the ratings of a single other movie and the 10 movies that are hardest to predict from the ratings of a single other movies (and their associated COD values, as well as which movie ratings are the best predictor, so this table should have 3 columns).*

For each of the 400 target movies, we use each user's rating of it as target y. Traversing through other 399 ratings of other movies (as regressor), we calculate the COD for each regressor movie to the target movie. Afterward, we pick the highest COD value among these 399 values, with its corresponding name. Now we can have the histogram of the highest COD for each target movie:



The mean of the highest COD is 0.424. This shows that for most of the target movies, a single movie regressor is not strong enough. Also, after ranking the highest COD for each of the target movies, we can find the 10 movies that are most easily predicted and the 10 movies that are most hard predicted.

Top 10 easiest:

rank	Movie Names	Associated COD	Best Predictor Movies
1	Erik the Viking (1989)	0.731789	I.Q. (1994)
2	I.Q. (1994)	0.731789	Erik the Viking (1989)
3	Patton (1970)	0.713793	The Lookout (2007)
4	The Lookout (2007)	0.713793	Patton (1970)
5	Best Laid Plans (1999)	0.711540	The Bandit (1996)
6	The Bandit (1996)	0.711540	Best Laid Plans (1999)
7	The Straight Story (1999)	0.700822	Congo (1995)
8	Congo (1995)	0.700822	The Straight Story (1999)
9	The Final Conflict (1981)	0.700437	The Lookout (2007)
10	Heavy Traffic (1973)	0.692863	Ran (1985)

Top 10 hardest:

rank	Movie Names	Associated COD	Best Predictor Movies
1	Avatar (2009)	0.079484	Bad Boys (1995)
2	Interstellar (2014)	0.111184	Torque (2004)
3	Black Swan (2010)	0.116970	Sorority Boys (2002)
4	Clueless (1995)	0.141324	Escape from LA (1996)
5	The Cabin in the Woods (2012)	0.143925	The Evil Dead (1981)
6	La La Land (2016)	0.148358	The Lookout (2007)
7	Titanic (1997)	0.153920	Cocktail (1988)
8	13 Going on 30 (2004)	0.160118	Can't Hardly Wait (1998)
9	The Fast and the Furious (2001)	0.169000	Terminator 3: Rise of the Machines (2003)
10	Grown Ups 2 (2013)	0.171151	The Core (2003)

We note that these movies are all very good movies. Perhaps masterpieces are special on their own.

Q2

For the 10 movies that are best and least well predicted from the ratings of a single other movie (so 20 in total), build multiple regression models that include gender identity (column 475), sibship status (column 476) and social viewing preferences (column 477) as additional predictors (in addition to the best predicting movie from question 1). Comment on how R^2 has changed relative to the answers in question 1. Please include a figure with a scatterplot where the old COD (for the simple linear regression models from the previous question) is on the x-axis and the new R^2 (for the new multiple regression models) is on the y-axis.

Since the 3 new regressors added to this question are all categorical data, we apply one-hot encoding to each of them. Note that there is a ‘Did not respond’ or ‘self-described’ answer for these questions. I did not drop them because I think they could also provide some kind of information. For instance, people self describe their gender could have certain preferences for movies. After this operation, we concatenate them to the original best movie regressor for each target movie. We would have the following results:

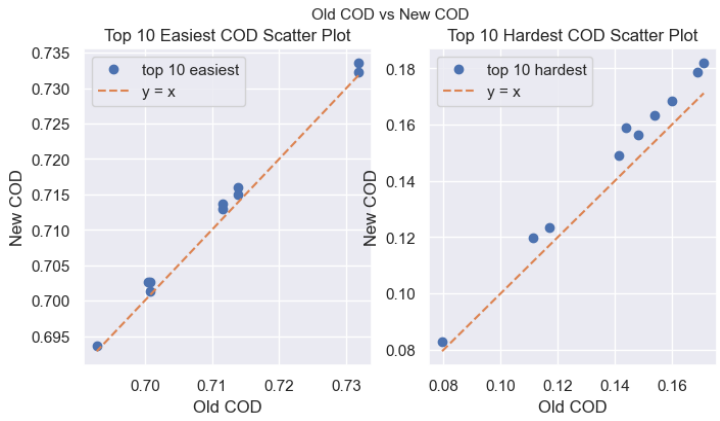
Top 10 easiest:

Movie Names	Old COD	New COD	Increase Proportion	Increased Absolute Value
Erik the Viking (1989)	0.731789	0.733531	0.24%	0.001742
I.Q. (1994)	0.731789	0.732318	0.07%	0.000529
Patton (1970)	0.713793	0.715005	0.17%	0.001212
The Lookout (2007)	0.713793	0.716030	0.31%	0.002237
Best Laid Plans (1999)	0.711540	0.712988	0.20%	0.001448
The Bandit (1996)	0.711540	0.713606	0.29%	0.002066
The Straight Story (1999)	0.700822	0.702610	0.26%	0.001788
Congo (1995)	0.700822	0.701389	0.08%	0.000567
The Final Conflict (1981)	0.700437	0.702602	0.31%	0.002165
Heavy Traffic (1973)	0.692863	0.693568	0.10%	0.000705

Top 10 hardest:

Movie Names	Old COD	New COD	Increase Proportion	Increased Absolute Value
Avatar (2009)	0.079484	0.082784	4.15%	0.003300
Interstellar (2014)	0.111184	0.119650	7.61%	0.008466
Black Swan (2010)	0.116970	0.123373	5.47%	0.006404
Clueless (1995)	0.141324	0.149076	5.49%	0.007753
The Cabin in the Woods (2012)	0.143925	0.159079	10.53%	0.015153
La La Land (2016)	0.148358	0.156441	5.45%	0.008082
Titanic (1997)	0.153920	0.163249	6.06%	0.009328
13 Going on 30 (2004)	0.160118	0.168477	5.22%	0.008359
The Fast and the Furious (2001)	0.169000	0.178790	5.79%	0.009790
Grown Ups 2 (2013)	0.171151	0.181859	6.26%	0.010708

As we can see, for the movies that are hard to predict from a single movie rating, the new regressors would have a higher proportion of contribution (as well as absolute value) to the increment of COD. This is quite reasonable since their old regressor has unsatisfactory performance. However, the increased proportion is not significant. This indicates that the 3 new factors are not very powerful to explain the movie ratings. We can also have a scatter plot for the old COD & new COD comparison.



Q3

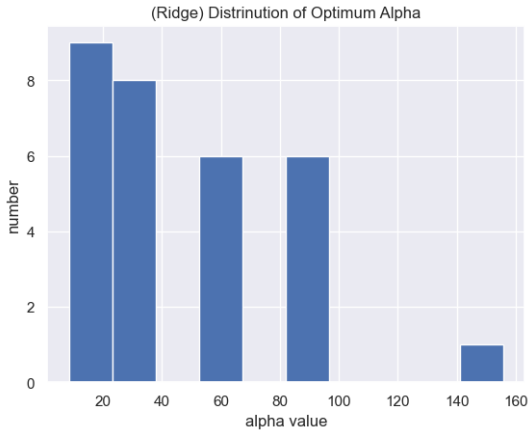
Pick 30 movies in the middle of the COD range, as identified by question 1 (that were not used in question 2). Now build a regularized regression model with the ratings from 10 other movies (picked randomly, or deliberately by you) as an input. Please use ridge regression, and make sure to do suitable hyperparameter

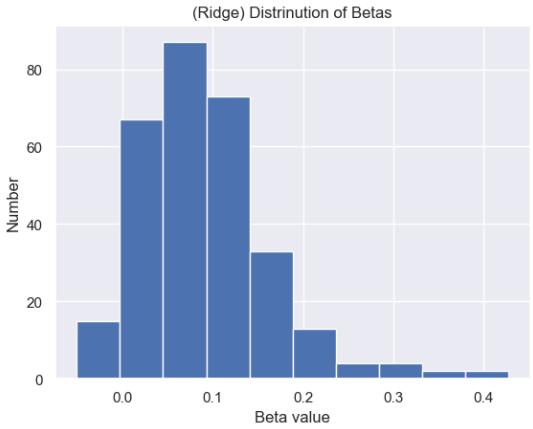
tuning. Also make sure to report the RMSE for each of these 30 movies in a table, after doing an 80/20 train/test split. Comment on the hyperparameters you use and betas you find by doing so.

For this question, we choose 30 movies with an index from 184 to index 214 target movies for this question. We randomly choose 10 movies other than the above 30 movies, with their ratings as our regressor. For each of the target movies, we build its corresponding ridge regression model with its own optimal regularization strength alpha, RMSE for testing data, and coefficient corresponding to each of the regressor movies. Here we use the RidgeCV package to determine the optimal alpha. We get the following results:

	target movie name	RMSE	Optimal Regularization Strength Alpha	Titanic (1997)	American Beauty (1999)	Girl With a Pearl Earring (2003)	Friday the 13th Part III (1982)	Fargo (1996)	Hellraiser (1987)	Ghost (1990)	Fahrenheit 9/11 (2004)	Indiana Jones and the Kingdom of the Crystal Skull (2008)	The Final Conflict (1981)
0	Aliens (1986)	0.475	36.438590	0.003	0.057	0.133	0.035	0.080	0.174	0.129	0.046	0.039	0.281
1	Gone in Sixty Seconds (2000)	0.349	36.438590	0.002	0.037	0.051	0.092	0.074	0.054	0.040	0.111	0.065	0.283
2	Crossroads (2002)	0.362	22.456980	0.012	0.018	-0.003	0.060	0.108	0.056	0.174	0.079	0.064	0.290
3	Austin Powers: The Spy Who Shagged Me (1999)	0.563	95.936083	0.035	0.056	0.140	0.121	0.090	0.152	0.094	0.172	0.066	0.116
4	Austin Powers in Goldmember (2002)	0.510	22.456980	0.006	0.061	0.222	0.132	0.052	0.166	-0.029	0.140	0.025	0.365
5	Goodfellas (1990)	0.398	13.840161	-0.001	0.053	0.183	0.013	0.055	0.175	0.034	0.174	-0.051	0.363
6	The Big Lebowski (1998)	0.397	22.456980	-0.030	0.103	0.136	0.040	0.122	0.048	0.062	0.218	-0.001	0.295
7	Twister (1996)	0.277	95.936083	0.029	0.090	0.084	0.099	0.098	0.089	0.112	0.099	0.042	0.103
8	Blues Brothers 2000 (1998)	0.413	36.438590	0.002	0.079	0.209	0.041	0.026	0.165	0.070	0.020	0.038	0.185
9	Dances with Wolves (1990)	0.421	22.456980	0.025	0.077	0.226	-0.003	0.033	0.063	0.073	0.114	0.039	0.234
10	28 Days Later (2002)	0.328	36.438590	0.042	0.028	0.104	0.022	0.102	0.112	0.117	0.063	0.073	0.248
11	Knight and Day (2010)	0.361	59.125084	0.008	0.054	0.175	-0.015	0.015	0.079	0.116	0.171	0.123	0.115
12	The Evil Dead (1981)	0.334	59.125084	-0.012	0.055	0.126	0.116	0.063	0.185	0.123	0.134	-0.007	0.121
13	The Machinist (2004)	0.361	36.438590	-0.001	0.094	0.131	-0.049	0.169	0.141	0.039	0.082	0.057	0.211
14	Uptown Girls (2003)	0.451	36.438590	0.040	0.059	0.238	0.010	-0.028	0.160	0.105	0.191	0.030	0.190
15	The Blue Lagoon (1980)	0.323	36.438590	0.065	0.041	0.150	0.135	0.022	0.157	0.073	0.067	0.066	0.181
16	Men in Black II (2002)	0.521	59.125084	0.046	0.015	0.175	0.122	-0.003	0.134	0.070	0.167	0.186	0.176
17	Men in Black (1997)	0.518	95.936083	0.049	0.053	0.126	0.081	0.042	0.090	0.114	0.130	0.139	0.150
18	The Green Mile (1999)	0.359	22.456980	-0.009	0.070	0.128	0.055	0.042	0.095	0.071	0.092	0.008	0.325
19	The Rock (1996)	0.290	95.936083	0.030	0.072	0.087	0.078	0.074	0.072	0.119	0.127	0.027	0.095
20	You're Next (2011)	0.268	59.125084	0.003	0.050	0.147	0.096	0.043	0.064	0.092	0.131	0.043	0.178
21	The Poseidon Adventure (1972)	0.335	95.936083	0.046	0.122	0.025	0.056	0.107	0.122	0.051	0.058	0.052	0.063
22	The Good the Bad and the Ugly (1966)	0.361	59.125084	0.000	0.086	0.131	0.022	0.067	0.141	0.137	0.104	0.003	0.186
23	Let the Right One In (2008)	0.271	8.529644	-0.019	0.069	0.109	0.026	0.102	0.097	0.084	0.049	0.018	0.390
24	Equilibrium (2002)	0.310	155.665436	0.019	0.084	0.057	0.058	0.105	0.092	0.053	0.089	0.063	0.041
25	The Mummy Returns (2001)	0.472	59.125084	0.055	0.071	0.189	0.169	0.016	0.001	0.099	0.082	0.108	0.145
26	The Mummy (1999)	0.443	95.936083	0.058	0.093	0.140	0.125	0.040	0.162	0.108	0.067	0.089	0.094
27	Just Married (2003)	0.395	13.840161	0.029	0.046	0.234	-0.022	0.032	0.225	-0.036	0.141	0.059	0.298
28	Reservoir Dogs (1992)	0.383	13.840161	-0.020	0.091	0.169	-0.035	0.094	0.106	0.133	0.120	-0.034	0.428
29	Man on Fire (2004)	0.404	36.438590	0.003	0.042	0.115	0.009	0.099	0.110	0.209	0.105	0.023	0.212

In this table, each row represents the information for a target movie, with its name in the 1st column. The RMSE is in the 2nd column. The optimum hyperparameter alpha is in the 3rd row. The coefficient beta locates between the 4th column to the 13th column. As we can see, the alpha is quite big compared with the default 1 in ridge regression for this task. It indicates that the ridge requires a heavy penalty to prevent overfitting. However, it is noteworthy that although the alpha value range is stable, the distribution of optimum alpha is not with different choices of regressors. For instance, In the below graph, the alpha values are a little right-tailed. But with different regressors, the alpha values could have been more centered.





For the betas, most of the betas are positive and small, with similar absolute values. This indicates that the penalty from the ridge regression deed works.

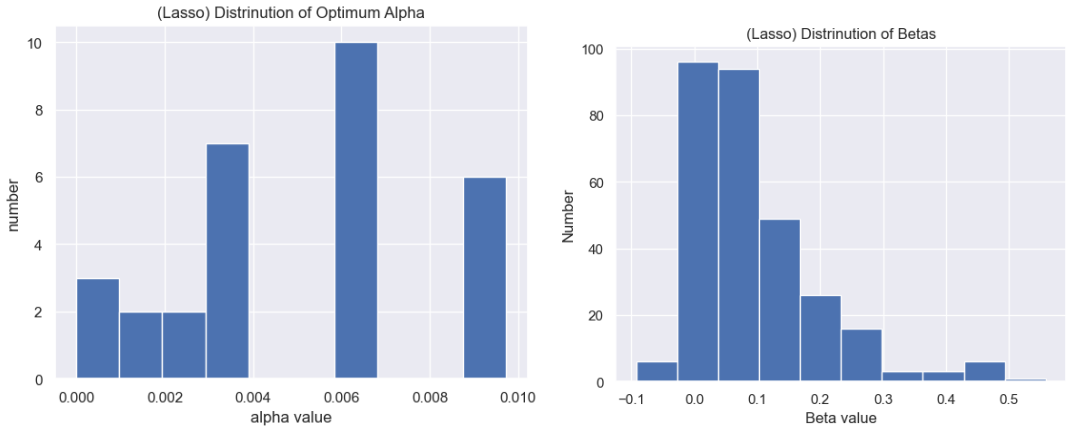
Q4

Repeat question 3) with LASSO regression. Again, make sure to comment on the hyperparameters you use and betas you find by doing so.

This question is quite similar to question 3. The only major difference is here we apply the LassoCV package to help us find the best regularization strength alpha. Here are our results:

	target movie name	RMSE	Optimal Regularization Strengt Alpha	Titanic (1997)	American Beauty (1999)	Girl With a Pearl Earring (2003)	Friday the 13th Part III (1982)	Fargo (1996)	Hellraiser (1987)	Ghost (1990)	Fahrenheit 9/11 (2004)	Indiana Jones and the Kingdom of the Crystal Skull (2008)	The Final Conflict (1981)
0	Aliens (1986)	0.471666	5.994843e-03	0.000	0.046	0.101	0.000	0.067	0.183	0.131	0.002	0.020	0.442
1	Gone in Sixty Seconds (2000)	0.354827	3.694601e-03	0.000	0.029	0.000	0.072	0.068	0.013	0.014	0.109	0.053	0.482
2	Crossroads (2002)	0.365750	5.994843e-03	0.004	0.006	0.000	0.038	0.101	0.023	0.182	0.058	0.054	0.380
3	Austin Powers: The Spy Who Shagged Me (1999)	0.567715	9.727203e-03	0.014	0.036	0.197	0.117	0.076	0.194	0.062	0.237	0.047	0.099
4	Austin Powers in Goldmember (2002)	0.509406	5.994843e-03	0.000	0.052	0.207	0.108	0.030	0.149	-0.000	0.118	0.006	0.475
5	Goodfellas (1990)	0.402020	3.694601e-03	0.000	0.047	0.172	0.000	0.046	0.166	0.009	0.169	-0.042	0.431
6	The Big Lebowski (1998)	0.404347	3.694601e-03	-0.025	0.100	0.116	0.017	0.121	0.012	0.037	0.232	-0.000	0.400
7	Twister (1996)	0.275397	1.403289e-03	0.015	0.096	0.068	0.097	0.104	0.090	0.142	0.119	0.023	0.181
8	Blues Brothers 2000 (1998)	0.416596	5.994843e-03	0.000	0.073	0.265	0.012	0.002	0.180	0.045	0.000	0.025	0.235
9	Dances with Wolves (1990)	0.427469	5.994843e-03	0.018	0.072	0.256	0.000	0.016	0.032	0.053	0.107	0.029	0.279
10	28 Days Later (2002)	0.331446	3.694601e-03	0.036	0.017	0.068	0.000	0.099	0.101	0.121	0.037	0.062	0.398
11	Knight and Day (2010)	0.363520	9.727203e-03	0.000	0.041	0.242	-0.000	0.000	0.041	0.100	0.203	0.123	0.075
12	The Evil Dead (1981)	0.334818	2.276970e-03	-0.021	0.046	0.143	0.108	0.047	0.241	0.129	0.151	-0.021	0.141
13	The Machinist (2004)	0.361510	8.648423e-04	-0.005	0.091	0.135	-0.092	0.182	0.152	0.011	0.073	0.046	0.338
14	Uptown Girls (2003)	0.455423	3.694601e-03	0.029	0.047	0.300	-0.000	-0.029	0.160	0.078	0.211	0.012	0.223
15	The Blue Lagoon (1980)	0.326896	3.694601e-03	0.062	0.033	0.169	0.131	0.002	0.173	0.057	0.047	0.057	0.246
16	Men in Black II (2002)	0.524207	5.994843e-03	0.034	0.000	0.221	0.108	-0.000	0.127	0.019	0.192	0.192	0.238
17	Men in Black (1997)	0.515071	5.994843e-03	0.030	0.043	0.147	0.048	0.007	0.064	0.119	0.157	0.141	0.287
18	The Green Mile (1999)	0.363129	5.994843e-03	-0.000	0.064	0.099	0.033	0.028	0.072	0.053	0.072	0.000	0.446
19	The Rock (1996)	0.291734	2.276970e-03	0.013	0.075	0.086	0.064	0.068	0.057	0.157	0.179	0.003	0.156
20	You're Next (2011)	0.275127	5.994843e-03	-0.000	0.041	0.174	0.079	0.020	0.020	0.075	0.144	0.025	0.297
21	The Poseidon Adventure (1972)	0.326272	1.000000e-10	0.047	0.142	-0.045	0.046	0.133	0.199	0.048	0.057	0.049	0.101
22	The Good the Bad and the Ugly (1966)	0.361747	9.727203e-03	0.000	0.078	0.112	0.000	0.045	0.141	0.143	0.079	0.000	0.280
23	Let the Right One In (2008)	0.272931	1.403289e-03	-0.017	0.067	0.092	0.014	0.100	0.088	0.080	0.038	0.013	0.457
24	Equilibrium (2002)	0.307988	5.994843e-03	0.004	0.092	0.049	0.049	0.141	0.150	0.035	0.139	0.064	0.000
25	The Mummy Returns (2001)	0.474114	9.727203e-03	0.046	0.058	0.261	0.181	0.000	0.000	0.075	0.053	0.104	0.146
26	The Mummy (1999)	0.435460	9.727203e-03	0.048	0.092	0.216	0.134	0.000	0.235	0.111	0.029	0.082	0.039
27	Just Married (2003)	0.399842	9.727203e-03	0.021	0.034	0.214	0.000	0.009	0.212	0.000	0.116	0.049	0.290
28	Reservoir Dogs (1992)	0.384065	1.000000e-10	-0.026	0.090	0.157	-0.066	0.088	0.092	0.133	0.115	-0.047	0.558
29	Man on Fire (2004)	0.408453	3.694601e-03	-0.000	0.033	0.082	-0.000	0.093	0.096	0.244	0.091	0.005	0.311

In this table, each row represents the information for a target movie, with its name in the 1st column. The RMSE is in the 2nd column. The optimum hyperparameter alpha is in the 3rd row. The coefficient beta locates between the 4th column to the 13th column. As we can see, the alpha is quite small compared with the default 1 in ridge regression for this task. It is also small compared to the optimal alpha in ridge regression. It indicates that the lasso shrinks the coefficient very fast for this task.



For the betas, it is noteworthy some of the betas are negative and many of them are approaching zeros. This indicates the feature selection property of lasso regression. It tends to maintain the best features. Also, from the RMSE level, we can see both ridge and lasso regression has a satisfactory performance.

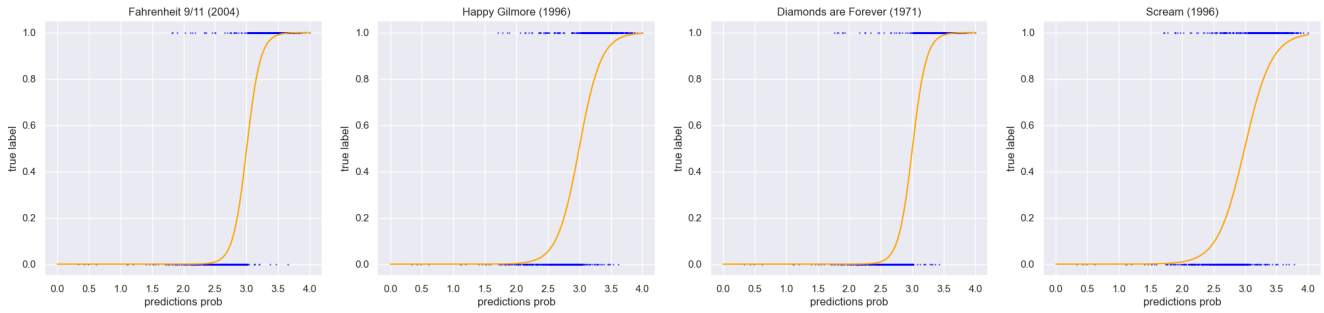
Q5

Compute the average movie enjoyment for each user (using only real, non-imputed data). Use these averages as the predictor variable X in a logistic regression model. Sort the movies order of increasing rating (also using only real, non-imputed data). Now pick the 4 movies in the middle of the score range as your target movie. For each of them, do a media split (now using the imputed data) of ratings to code movies above the median rating with the Y label 1 (= enjoyed) and movies below the median with the label 0 (= not enjoyed). For each of these movies, build a logistic regression model (using X to predict Y), show figures with the outcomes and report the betas as well as the AUC values. Comment on the quality of your models. Make sure to use cross-validation methods to avoid overfitting.

For this question, we choose movies from index 198 to 202 as our target movie. After calculating their median rating, we build the binary table to store the category variable for each user. The results are shown below:

	Fahrenheit 9/11 (2004)	Happy Gilmore (1996)	Diamonds are Forever (1971)	Scream (1996)
0	0	1	0	0
1	0	0	0	0
2	1	1	1	1
3	0	0	0	0
4	0	0	0	0
...	...	...	...	...
1092	1	1	1	1
1093	1	1	1	1
1094	1	1	1	1
1095	1	1	1	1
1096	1	1	0	0

To prevent overfitting, we use the LogisticRegressionCV package to do the cross validation and get the following result:



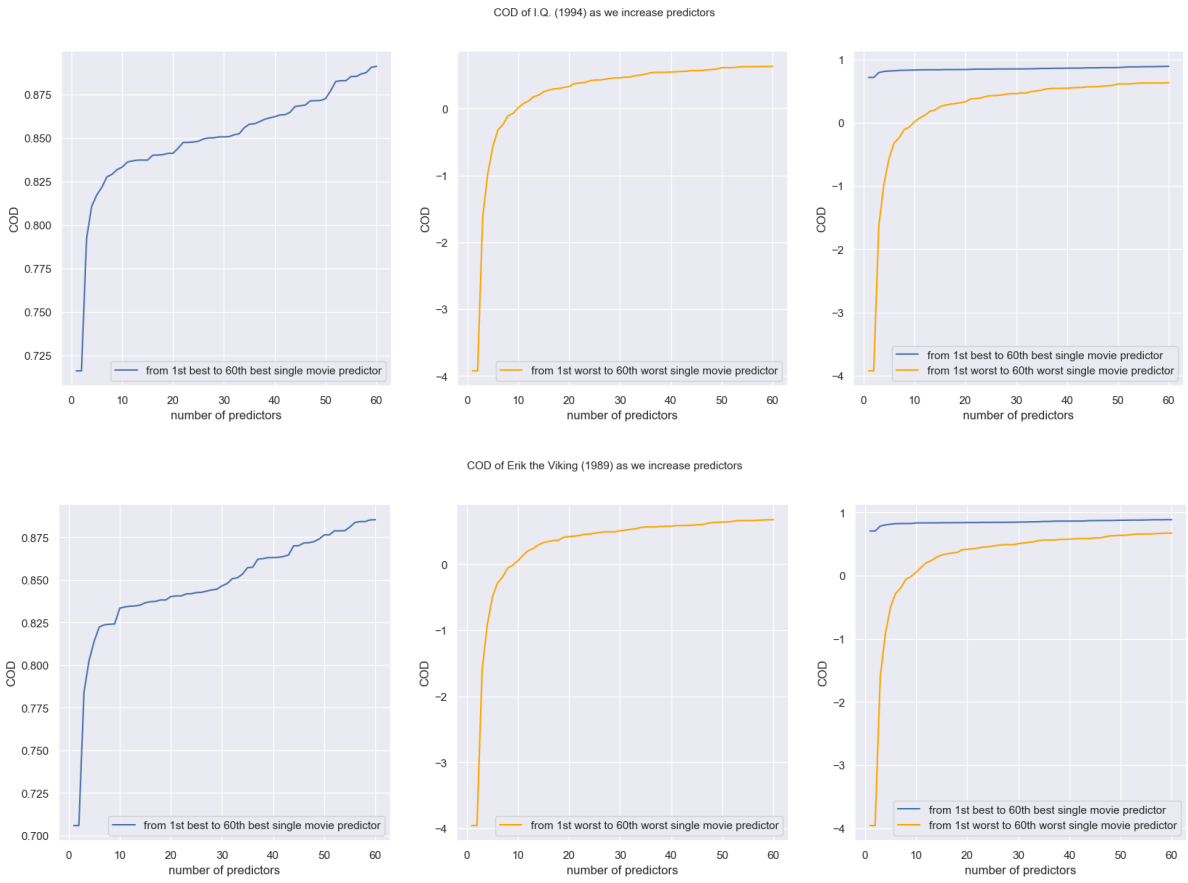
We also have the coefficient and AUC score table below:

	Target Movie Name	Coefficient (Beta)	AUC Score
0	Fahrenheit 9/11 (2004)	9.215374	0.948952
1	Happy Gilmore (1996)	5.751416	0.883306
2	Diamonds are Forever (1971)	9.063547	0.952598
3	Scream (1996)	4.635989	0.845926

As we can see, all AUC score is quite high. This indicates that our classifier has good discriminative power. It cut off the negative and positive results well. Among 4 movies, Fahrenheit 9/11 (2004) has the largest beta. This indicates that its user's enjoyment user changes fast with the change in the user's average movie enjoyment. However, we find that if we do train/test split by ourselves and apply LogisticRegression instead of LogisticRegressionCV, we can get different betas (perhaps because of different choices of test set).

Extra Credit

In this section I am trying to investigate how the COD gonna grows as we keep adding numbers of movie ratings as new regressors, for those movies that have already been well explained by a single movie. We choose the first two movies as targets from the table of 10 movies that are most easily predicted in question 1, which are I.Q. (1994) and Erik the Viking (1989). For each of the target movies, we traverse through other 399 predictor movies' ratings(as a single predictor) and calculate the COD given by each of them. We rank those predictors by their single COD to the target movie. Then we gradually add from 1st best single predictor to the 60th best single predictor as our input x to see how the explanation power is going to grow. (That is saying, in the ith experiment, we have from the 1st best predictors to the ith best predictors as our input). We also gradually add from the worst single predictor to the 60th worst single predictor as our input x to see how the explanation power is going to grow. Our experiment results are shown below:



As we can see, both movies show a similar trend.

From the left two graphs, we know that when we gradually add the first 5 or 6 best predictors, the COD grows quite fast, whereas later COD grows slower. This can be explained by 2 aspects. First, the quality of regressors gradually deteriorates. Second, as we increase the regressors, they could be more and more correlated.

From the middle two graphs, we can see that even if we start with a poor regressor, we can finally achieve a relatively satisfactory level of COD as we increase the number of regressors. This shows the strong learning ability of linear regression.