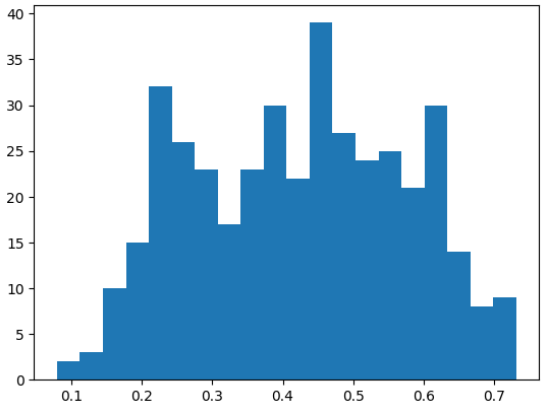
Data Analysis Project 2 (Machine Learning)

Ansh Riyal, AR7964

(D is common for all):
Imputation method used is the same as was suggested in the spec sheet. This is a reasonable method because it captures both the column trend (information about the movie) as well as the row trend(information about the user)

Question 1: Predict each of the 400 movies with the other 399 movies one at a time.
(Y): We want to use a linear regression model to predict a target movie ratings using an input movie ratings
Finding(F): We obtained a very large range of R^2 values for the best predictor (for every target movie) starting from hard to predict movies (0.1) to easy to predict (approx. 0.73). Thus, we observe that there are different kinds of movies, some are relatively easy to predict, while some are barely above the random model.
Answer(A): The Average R2 value for 400 movies = 0.42378
Distribution of R2 values:



Here are the easiest and hardest to predict movies

```
easiest_to_predict
```

	Target	Predictor	R2 Value
309	Heavy Traffic (1973)	Ran (1985)	0.692734
334	The Final Conflict (1981)	The Lookout (2007)	0.700188
282	Congo (1995)	The Straight Story (1999)	0.700569
287	The Straight Story (1999)	Congo (1995)	0.700569
240	The Bandit (1996)	Best Laid Plans (1999)	0.711222
249	Best Laid Plans (1999)	The Bandit (1996)	0.711222
395	Patton (1970)	The Lookout (2007)	0.713554
377	The Lookout (2007)	Patton (1970)	0.713554
208	I.Q. (1994)	Erik the Viking (1989)	0.731507
203	Erik the Viking (1989)	I.Q. (1994)	0.731507

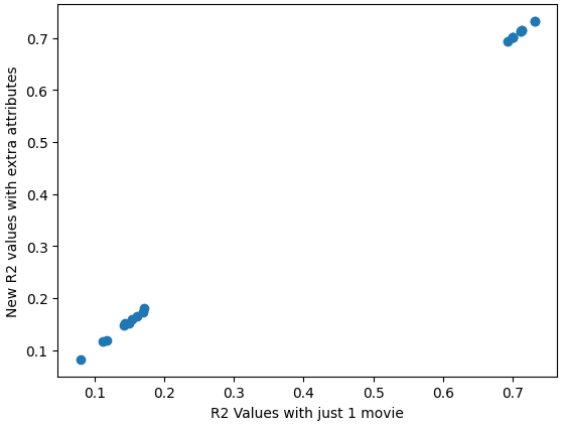
```
hardest_to_predict
```

	Target	Predictor	R2 Value
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

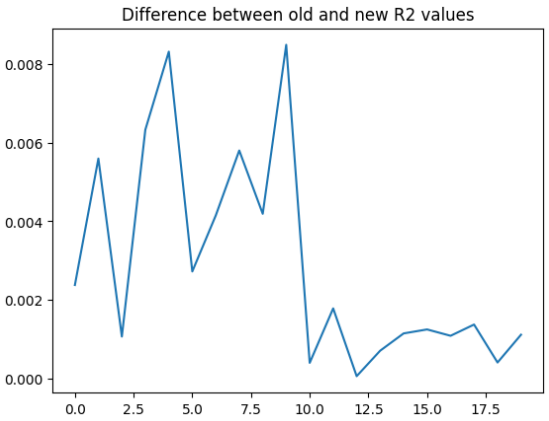
Question 2: Use gender, solo child and solo watching to improve the best linear model corresponding to the middle 30 movies (arranged by COD).
(Y): We want to see what the impact of extra information is on the easiest and hardest to predict movies using their corresponding best predictors
Finding(F): We consistently see an increase in the R2 values (as we would expect with addition of information). The largest improvements are for the harder to predict target movies.

Answer(A):

Old vs New R^2 values:



Diff. between old and new R^2 values:



Question 3: Using Ridge Regression (and searching for optimal alphas).

(Y): We want to see how ridge regression works for predicting movies that are neither too easily predicted, not too hard to predict using other random movies

Finding(F): We can clearly see that we get Alpha/Lambda values in a wide range starting from as low as 13 to as high as 93). We also see that the norms of the coefficient decrease when the alpha is high.

Answer(A):

```
for i in range(30):
    print("\nFor movie: ", moderately_predictable_movies[i],"\tRMSE value is ", rmse_scores[i])
```

```
For movie:  Aliens (1986)       RMSE value is  0.32728956510436036
For movie:  Gone in Sixty Seconds (2000)        RMSE value is  0.34389542159899006
For movie:  Crossroads (2002)   RMSE value is  0.31686923669790734
For movie:  Austin Powers in Goldmember (2002)  RMSE value is  0.49603874787386587
For movie:  Austin Powers: The Spy Who Shagged Me (1999)        RMSE value is  0.5293950502129507
For movie:  Goodfellas (1990)   RMSE value is  0.3688089133946253
For movie:  The Big Lebowski (1998)     RMSE value is  0.30914686360172516
For movie:  Twister (1996)      RMSE value is  0.33533596876758553
For movie:  Blues Brothers 2000 (1998)  RMSE value is  0.3024551938313563
For movie:  Dances with Wolves (1990)   RMSE value is  0.46882894840013567
For movie:  28 Days Later (2002)        RMSE value is  0.37401900909274666
For movie:  Knight and Day (2010)       RMSE value is  0.40494813947905844
For movie:  The Evil Dead (1981)        RMSE value is  0.3850662382512996
For movie:  The Machinist (2004)        RMSE value is  0.264811741046654
For movie:  The Blue Lagoon (1980)      RMSE value is  0.3227156484090177
For movie:  Uptown Girls (2003)         RMSE value is  0.37106322066822583
For movie:  Men in Black (1997)         RMSE value is  0.5133913407966233
For movie:  Men in Black II (2002)      RMSE value is  0.48115282968355677
For movie:  The Green Mile (1999)       RMSE value is  0.3076104240413906
For movie:  The Rock (1996)     RMSE value is  0.2649748163623254
For movie:  You're Next (2011)  RMSE value is  0.35723058824879506
For movie:  The Poseidon Adventure (1972)       RMSE value is  0.44603815843408096
For movie:  The Good the Bad and the Ugly (1966)        RMSE value is  0.47144469092903596
For movie:  Let the Right One In (2008)         RMSE value is  0.3088471834336468
For movie:  Equilibrium (2002)  RMSE value is  0.27239484277283554
For movie:  Just Married (2003)         RMSE value is  0.36215708665515817
For movie:  The Mummy Returns (2001)    RMSE value is  0.4956652200771281
For movie:  The Mummy (1999)    RMSE value is  0.500240858089131
For movie:  Reservoir Dogs (1992)       RMSE value is  0.39473389258119546
For movie:  Man on Fire (2004)  RMSE value is  0.29554016817923856
```

Question 4: Using Lasso Regression (and searching for optimal alphas).

(Y): We want to see how lasso regression works for predicting movies that are neither too easily predicted, not too hard to predict using other random movies

Finding(F): We can clearly see that we get Alpha/Lambda values in a wide range starting from as low as 0 to as high as 0.1).

Answer(A):

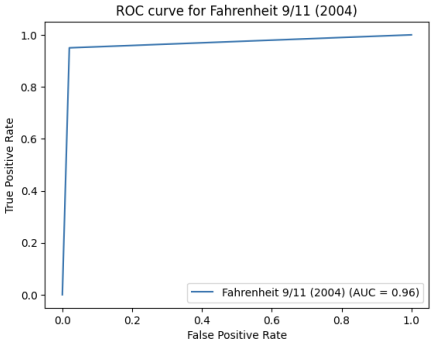
```
for i in range(len(best_alpha_lasso)):
    print("\nFor movie: ", moderately_predictable_movies[i],"\tRMSE value is ", RMSE_values_lasso[i])
```

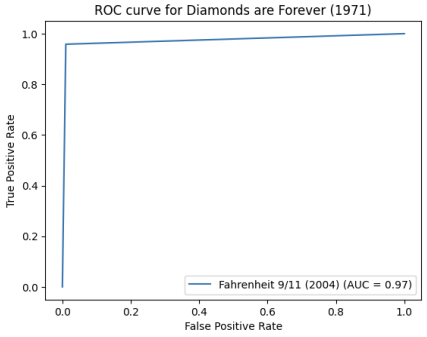
```
For movie:  Aliens (1986)       RMSE value is  0.10794024327589101

For movie:  Gone in Sixty Seconds (2000)        RMSE value is  0.11911205169095841

For movie:  Crossroads (2002)   RMSE value is  0.10383076450632886

For movie:  Austin Powers in Goldmember (2002)  RMSE value is  0.25188814223097666

For movie:  Austin Powers: The Spy Who Shagged Me (1999)        RMSE value is  0.27219047498598586

For movie:  Goodfellas (1990)   RMSE value is  0.14302538967842002

For movie:  The Big Lebowski (1998)     RMSE value is  0.09928757630698452

For movie:  Twister (1996)      RMSE value is  0.11478443146626346

For movie:  Blues Brothers 2000 (1998)  RMSE value is  0.09739438782785448

For movie:  Dances with Wolves (1990)   RMSE value is  0.22280292793024134

For movie:  28 Days Later (2002)        RMSE value is  0.13939756161761555

For movie:  Knight and Day (2010)       RMSE value is  0.16350025392870948

For movie:  The Evil Dead (1981)        RMSE value is  0.14147335244558656

For movie:  The Machinist (2004)        RMSE value is  0.07332601068762014

For movie:  The Blue Lagoon (1980)      RMSE value is  0.10347685502464818

For movie:  Uptown Girls (2003)         RMSE value is  0.13730838485934355

For movie:  Men in Black (1997)         RMSE value is  0.2604549732881927

For movie:  Men in Black II (2002)      RMSE value is  0.23362162995003868

For movie:  The Green Mile (1999)       RMSE value is  0.09692215714534472

For movie:  The Rock (1996)     RMSE value is  0.06905573971235225

For movie:  You're Next (2011)  RMSE value is  0.1300174267183876

For movie:  The Poseidon Adventure (1972)       RMSE value is  0.20459838653972998

For movie:  The Good the Bad and the Ugly (1966)        RMSE value is  0.22351751815952695

For movie:  Let the Right One In (2008)         RMSE value is  0.09594030102786212

For movie:  Equilibrium (2002)  RMSE value is  0.07646487323263322

For movie:  Just Married (2003)         RMSE value is  0.13801455108342572

For movie:  The Mummy Returns (2001)    RMSE value is  0.25039501889866156

For movie:  The Mummy (1999)    RMSE value is  0.24805811414785492

For movie:  Reservoir Dogs (1992)       RMSE value is  0.1582553987378604

For movie:  Man on Fire (2004)  RMSE value is  0.09797881996240644
```

Question 5: Using Lasso Regression (and searching for optimal alphas).
(Y): We want to use the average user ratings to classify above average/below average ratings for movies in the middle of the scoring range
Finding(F): We can clearly see that we get AUC values very high. This means our models are really good in their discriminatory power between the ‘above median’ and ‘below median’ target values.
Answer(A):
Beta values:
Fahrenheit : 0.66
Happy Gilmore: 5.1
Diamonds: 5.5
Scream: 0.59

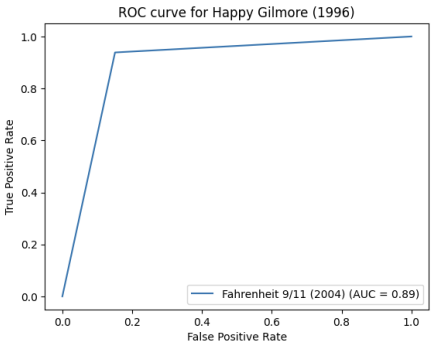
```
for i in range(4):
    print("For predicting movie {}, the AUC is {}".format(target_movies[i],auc_vals[i]))
```

```
For predicting movie Fahrenheit 9/11 (2004), the AUC is 0.965
For predicting movie Happy Gilmore (1996), the AUC is 0.8938265475008276
For predicting movie Diamonds are Forever (1971), the AUC is 0.9741666666666667
For predicting movie Scream (1996), the AUC is 0.8693319922889952
```






```
<Figure size 640x480 with 0 Axes>
```




```
<Figure size 640x480 with 0 Axes>
```