Mary Nwangwu

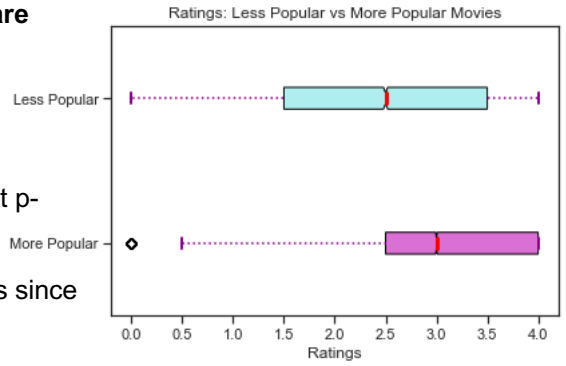
Analysis of Movie Ratings Data

For this data set, we define our alpha level of significance to be 0.005 (as per Benjamin et al., 2018) to cut down on false positives and we make the following assumptions:

1) the dependent variable (ratings) is ordinal
2) the independent variable consists of at least two categorical, independent groups
3) the observations are not normally distributed
4) independence of observations (within each group or between the groups themselves)

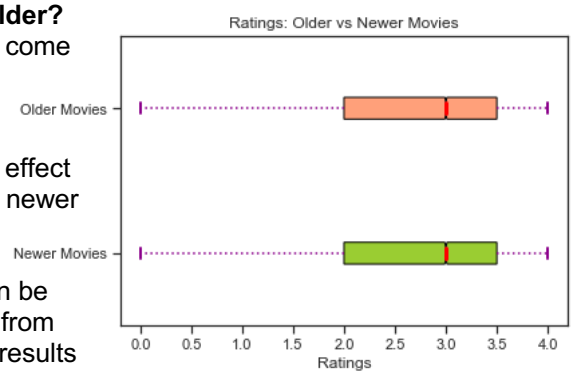
It is inappropriate to reduce the groupings within this data set to their means, as we can't interpret the psychological distance between ratings. However, we can interpret the rank order of the ratings. With this in mind, we then proceed to use significance tests that work with ordinal data: Mann Whitney U (compares medians, test statistic = U), Kolmogorov-Smirnov (compares entire distributions, test statistic = D), and Kruskal-Wallis (compares medians of 3 or more groups, k-1 degrees of freedom, test statistic = H). For Mann Whitney U and Kruskal-Wallis, we also assume the shape of the distributions of the groups being compared are similar. Effect sizes are calculated using Cliff's *d* (the Cohen's *d* equivalent for ordinal data). For most of the analyses, we focus on the median. This is because the median depends only on ordering operations, so we can use it for ratings data.

Q1: Are movies that are more popular rated HIGHER than movies that are less popular?



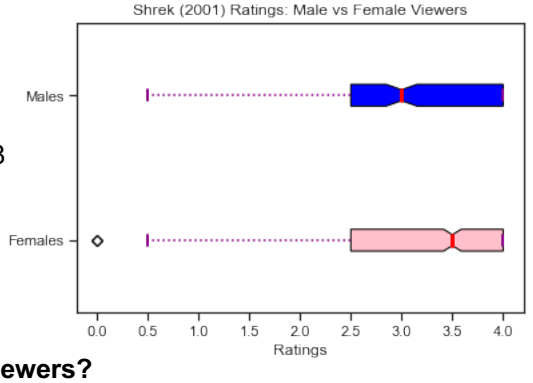
We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. Median split value is 197.5 total ratings. We found that the median ranks of the more popular[1] and less popular[2] movies were 3.0 and 2.5 respectively. U_1 = 1242808144.5, U_2 = 741899855.5, p-value = *0.00*, and effect size = 0.25 (small effect). Given that p-value < 0.005, we conclude that movies that are more popular are more likely to be rated higher than movies that are less popular. Ratings from the same individuals within each independent grouping may influence results since U test assumes independence of observations.

Q2: Are movies that are newer rated differently than movies that are older?



We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. Median split value is year 1999; movies from 1999 are included with newer movies. We found that the median ranks of the newer[1] and older[2] movies were 3.0 and 3.0 respectively. U_1 = 1553577699, U_2 = 1502583861, p-value = *1.29e-06*, and effect size = 0.02 (negligible effect). Given that p-value < 0.005, we conclude that newer movies are more likely to be rated differently than older movies. The sample sizes of the two groups being compared are unequal (n_1 = 65690, n_2 = 46524). With sufficient sample size, the difference in the rank sums can be large enough to be significant even though the medians are equal. Ratings from the same individuals within each independent grouping may also influence results since U test assumes independence of observations.

Q3: Is enjoyment of 'Shrek (2001)' gendered?

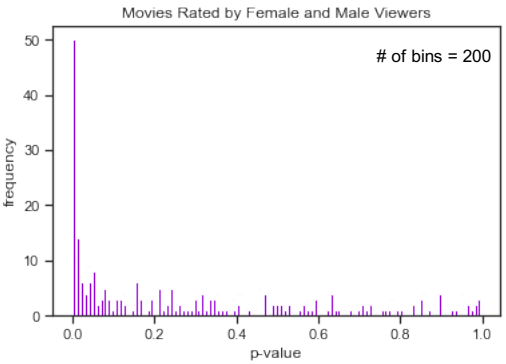


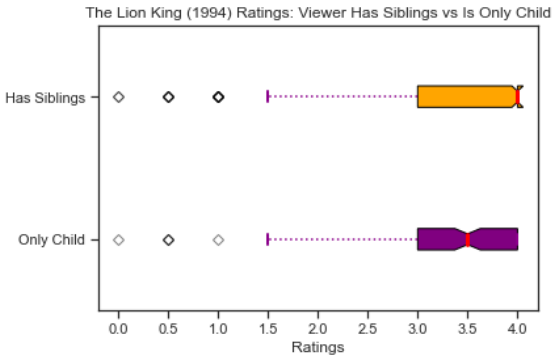
We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. We found that the median ranks of Shrek (2001) from female[1] and male[2] viewers were 3.5 and 3.0 respectively. U_1 = 96830.5, U_2 = 82232.5, p-value = *0.05*, and effect size = 0.08 (negligible effect). Given that p-value > 0.005, we conclude that enjoyment of Shrek (2001) is not gendered; females and males are more likely to rate the movie similarly. Alpha-level and effect size being low in number, as well as large difference in sample sizes (n_1 = 743, n_2 = 241), could lead to lower power (test less likely to pick up on an effect that is present).

Q4: What proportion of movies are rated differently by male and female viewers?

We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median.

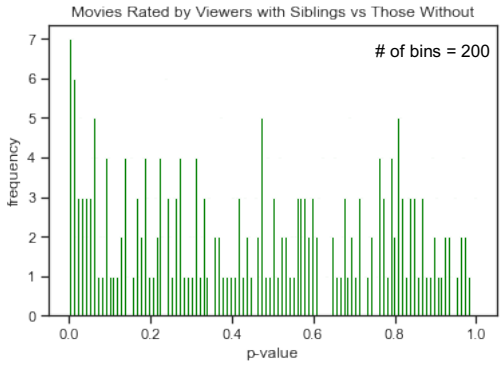
We compared the ratings between female and male viewers for all movies in the data set. We found that the proportion of movies that are rated differently by male and female viewers (have a p-value < 0.005) is 12.5% (50 out of 400 movies). Given this, we conclude that the enjoyment of most movies in the data set is not gendered.





Q5: Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?

We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. We found that the median ranks for The Lion King (1994) for people who are only children[1] and those with siblings[2] were 3.5 and 4.0 respectively. U_1 = 52929, U_2 = 64247, p-value = *0.04*, and effect size = -0.097 (negligible effect). Given that the p-value > 0.005, we conclude that people who are only children and those with siblings are more likely to experience a similar level of enjoyment watching The Lion King (1994). Alpha-level and effect size being low in number, as well as large difference in sample sizes (n_1 = 151, n_2 = 776), could lead to lower power (test less likely to pick up on an effect that is present).



Q6: What proportion of movies exhibit an "only child" effect?

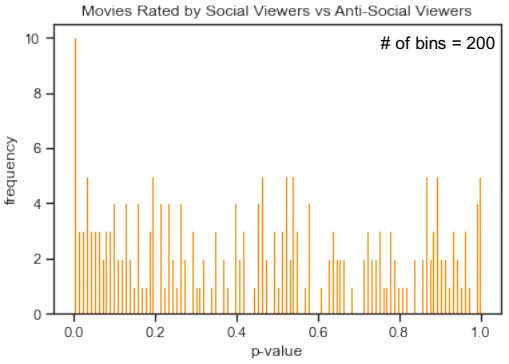
We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. We compared the ratings between people who are only children and those with siblings for all movies in the data set. We found that the proportion of movies that exhibit an "only child" effect (have a p-value < 0.005) is 1.75% (7 out of 400). Given this, we conclude that most movies in the data set do not exhibit an "only child" effect.

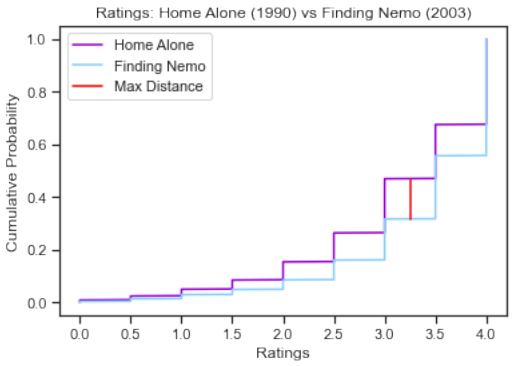
Q7: Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?

We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. We found that the median ranks for The Wolf of Wall Street (2013) for people who like to watch movies socially[1] and those who prefer to watch them alone[2] were 3.0 and 3.5 respectively. U_1 = 49303.5, U_2 = 56806.5, p-value = *0.11,* and effect size = -0.07 (negligible effect). Given that p-value > 0.005, we conclude that people who like to watch movies socially and those who prefer to watch movies alone are more likely to experience a similar level of enjoyment watching the Wolf of Wall Street (2013). Alpha-level and effect size being low in number, as well as difference in sample sizes (n_1 = 270, n_2 = 393), could lead to lower power (test less likely to pick up on an effect that is present).

Q8: What proportion of movies exhibit such a "social watching" effect?

We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. We compared the ratings between people who like to watch movies socially and those who prefer to watch them alone for all movies in the data set. We found that the proportion of movies that exhibit a "social watching" effect (have a p-value < 0.005) is 2.5% (10 out of 400 movies). Given this, we conclude that most movies in the data set do not exhibit a "social watching" effect.

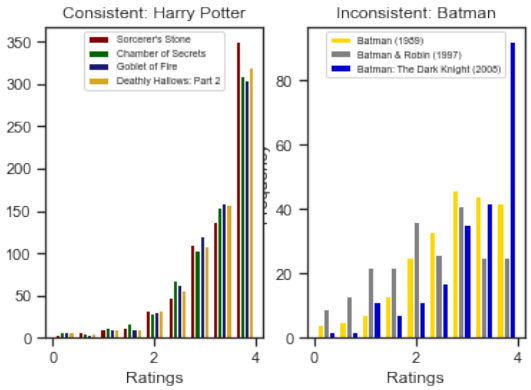




Q9: Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'?

We choose the Kolmogorov-Smirnov (KS) test because it tests whether the underlying distributions of two samples are the same (whatever they might be). It does not compare means or medians. D = 0.153 and p-value = *6.38e-10*. Given that p-value < 0.005, we conclude that the ratings distributions of Home Alone (1990) and Finding Nemo (2003) are different.

Q10: Several Franchises: How many of these are of inconsistent quality, as experienced by viewers?



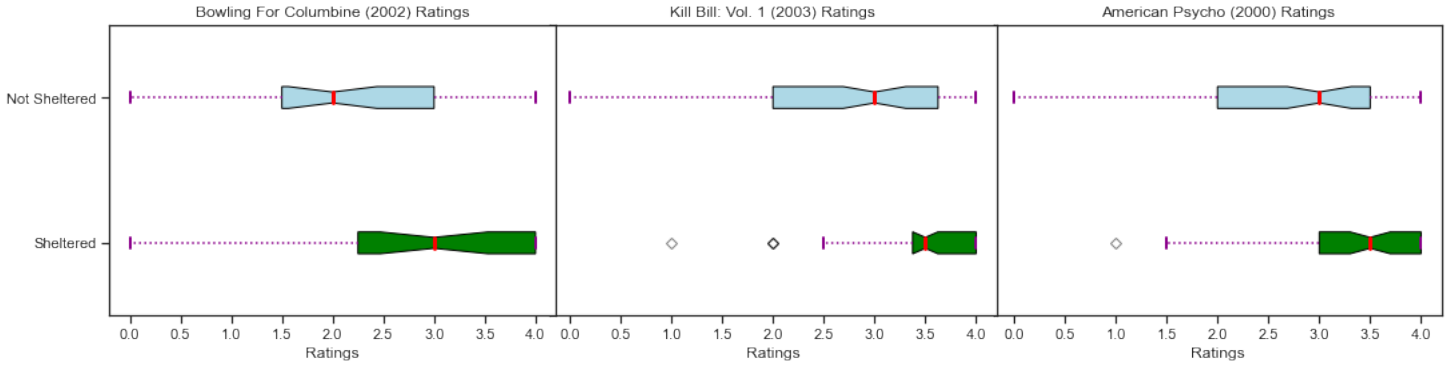
We choose the Kruskal Wallis (KW) test because it tests whether 3 or more samples come from populations with the same median. We found that out of the 8 franchises we analyzed, only 2 are of consistent quality, as experienced by viewers: Harry Potter (H = 5.87, p-value = 0.12, degrees of freedom = 3) and Pirates of the Caribbean (H = 6.66, p-value = 0.04, degrees of freedom = 2). The rest show inconsistent quality: Star Wars (H = 193.5, p-value = 6.94e-40, degrees of freedom = 5), The Matrix (H = 40.32, p-value = 1.75e-09, degrees of freedom = 2), Indiana Jones (H = 54.19, p-value = 1.02e-11, degrees of freedom = 3), Jurassic Park (H = 49.43, p-value = 1.85e-11, degrees of freedom = 2), Toy Story (H = 23.5, p-value = 7.9e-06, degrees of freedom = 2), and Batman (H = 84.66, p-value = 4.14e-19, degrees of freedom = 2). Given this, *we conclude that the majority (75% or 6 out of 8) of franchises we analyzed are of inconsistent quality* (p-value < *0.005*). We are interested in the inconsistency of quality over several movies in a franchise, so we constrict our data to focus only on viewers who have seen all the movies in a specific franchise. However, KW test assumes independence of observations. Since we are comparing groups of ratings from the same individuals for different movies, this test's results may not be the most accurate representation of the data set.

EXTRA: Do people who had a sheltered upbringing enjoy the movies Bowling for Columbine (2002), Kill Bill: Vol. 1 (2003), and American Psycho (2000) more than those who did not have a sheltered upbringing?

We choose the Mann Whitney U test because it tests whether two samples come from populations with the same median. We looked at the following movies: Bowling for Columbine (2002), Kill Bill: Vol. 1 (2003), and American Psycho (2000). Self-assessment rankings on 'I had a sheltered upbringing': 1 and 2 were deemed as "not sheltered", 4 and 5 as "sheltered", 3 (neutral) was left out. We found the following for those who had a sheltered upbringing[1] and those who did not[2]:

	Bowling for Columbine (2002)	Kill Bill: Vol. 1 (2003)	American Psycho (2000)
n_1 and n_2 (Sample Size)	27 and 29	64 and 68	64 and 56
Median Ranks	3.0^1 and 2.0^2	3.5^1 and 3.0^2	3.5^1 and 3.0^2
U_1 and U_2	592 and 190.5	2900 and 1452	2387 and 1197
P-value	0.0009	0.0007	0.0014
Effect Size	0.51 (large effect)	0.33 (medium effect)	0.33 (medium effect)

Given that all p-values < 0.005, we conclude that people who had a sheltered upbringing are more likely to enjoy these specific movies incorporating violence and psychological horror more than those who did not have a sheltered upbringing.



For this data set and for the purpose of performing various types of regression analyses, we make the following general assumptions:
1) Linearity: there is a linear relationship between the independent (predictor) and dependent (target) variables
2) Independence: observations should be independent of each other
3) Normality: residuals are normally distributed
4) Homoscedasticity: residuals have constant variance
5) For simple linear, multiple, and logistic regression models, we assume the predictor variables should not be highly correlated with each other (no multicollinearity)
6) For logistic regression models, we also assume that the target variable/outcome is binary or ordinal

In a regression model, the beta (β) coefficients are used to determine how much each predictor influences the target variable/outcome.

Null and Alternative Hypotheses:
1) Simple linear regression: Null states that there is no statistically significant relationship between the predictor and the target variables ($\beta_1 = 0$). Alternative states that there *is* a statistically significant relationship between the predictor and target ($\beta_1 \neq 0$).
2) Multiple regression: Null states that none of the predictor variables have a statically significant relationship with the target variable (all coefficients in model = 0). Alternative states that not every coefficient is simultaneously equal to zero.

In regularized regression, the alpha (α) value is a hyperparameter that controls the strength of the regularization (the prevention of over- or underfitting); it determines how much weight is given to the penalty term and must be tuned to achieve optimal model performance.

Missing ratings data were imputed with a 50/50 blend of the arithmetic mean of each movie and each user. One user was found to have no ratings for any of the movies, so they were dropped from the study. The Coefficient of Determination (COD), or R^2, determines the proportion of the variance in the outcome that can be explained by the predictor; it shows us how well the data fits the regression model. The Root Mean Square Error (RMSE) measures the error of a model in predicting quantitative data. Area under the ROC curve (AUC) is a measure of how well a classifier can distinguish between two classes.

Figure 1:
CODs for 400 Simple Linear Regression Models

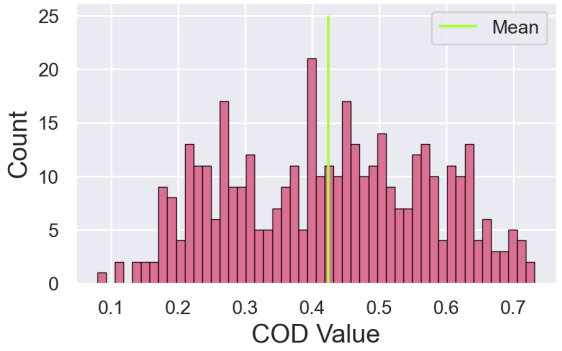


Table 1: Simple Linear Regression COD

Movie	COD	Predictor Movie
Erik the Viking (1989)	0.731507	I.Q. (1994)
I.Q. (1994)	0.731507	Erik the Viking (1989)
The Lookout (2007)	0.713554	Patton (1970)
Patton (1970)	0.713554	The Lookout (2007)
The Bandit (1996)	0.711222	Best Laid Plans (1999)
Best Laid Plans (1999)	0.711222	The Bandit (1996)
Congo (1995)	0.700569	The Straight Story (1999)
The Straight Story (1999)	0.700569	Congo (1995)
The Final Conflict (1981)	0.700188	The Lookout (2007)
Heavy Traffic (1973)	0.692734	Ran (1985)
Grown Ups 2 (2013)	0.171119	The Core (2003)
The Fast and the Furious (2001)	0.168991	Terminator 3: Rise of the Machines (2003)
13 Going on 30 (2004)	0.160164	Can't Hardly Wait (1998)
Titanic (1997)	0.154136	Cocktail (1988)
La La Land (2016)	0.148514	The Lookout (2007)
The Cabin in the Woods (2012)	0.143887	The Evil Dead (1981)
Clueless (1995)	0.141426	Escape from LA (1996)
Black Swan (2010)	0.117080	Sorority Boys (2002)
Interstellar (2014)	0.111343	Torque (2004)
Avatar (2009)	0.079485	Bad Boys (1995)

Q1: For each of the 400 movies, use a simple linear regression model to predict the ratings using the ratings of the other 399 movies
A simple linear regression model tells us how changes in a varying predictor variable affect the target variable. Knowing the value of the predictor variable and the correlation between predictor and target, we can use regression to make a prediction for the corresponding target value. For this movie ratings dataset, the average COD of the 400 simple linear regression models is about 0.424 (Figure 1), which means that the models, on average, explain about 42% percent of variation within the data. Since COD values range from 0 to 1, where a higher value indicates a better fit, this would mean that on average our simple linear regression models do not provide a good enough fit for the data. The 10 movies that are most easily predicted from the ratings of a single other movie and the 10 movies that are the hardest to predict from the ratings of a single other movie, along with their associated COD and their best predictor movie, are reported in Table 1 (top and bottom 10 movies are separated by red line). We notice that the movies that are easiest to predict (have the highest COD values) tend to also be the best predictor movie for their counterpart. We do not see this trend within the hardest to predict movies.

Q2: For the 10 movies that are best and least well predicted from the ratings of a single other movie, build multiple regression models that include gender identity, sibship status, and social viewing preferences, as additional predictors (in addition to their best predicting movie)
A multiple regression model tells us how changes in multiple varying predictor variables affect the target variable; it provides statistical control if there is more than one predictor that matters, allowing us to "control" for known confounds. Using the imputed data set, users with null and "did not respond" values for the gender identity, sibship status and social viewing preferences were dropped. Gender identity values were converted into dummy variables to represent female, male, and self-described viewers. For the movies that are the hardest to predict, 9/10 of the R^2 values slightly increased from their value in Q1 and 1/10 of the R^2 values slightly decreased. For the movies that are the easiest to predict, 3/10 of the R^2 values slightly increased from their value in Q1 and 7/10 of the R^2 values slightly decreased. For the movies with increased R^2 values, this means the additional predictors used for these models improved the fit and help explain a bit more of the variation within the data. For the movies with decreased R^2 values, this means that adding the additional predictors to the model may have caused the model to fit the data worse. Scatterplots where the old COD (for the simple linear regression models in Q1) is on the x-axis and the new R^2 (for these multiple regression models) is on the y-axis are shown in Figure 2. We observe a positive linear relationship between the new and old COD values.

Figure 2: Multiple Regression vs Simple Linear Regression COD Scatterplots

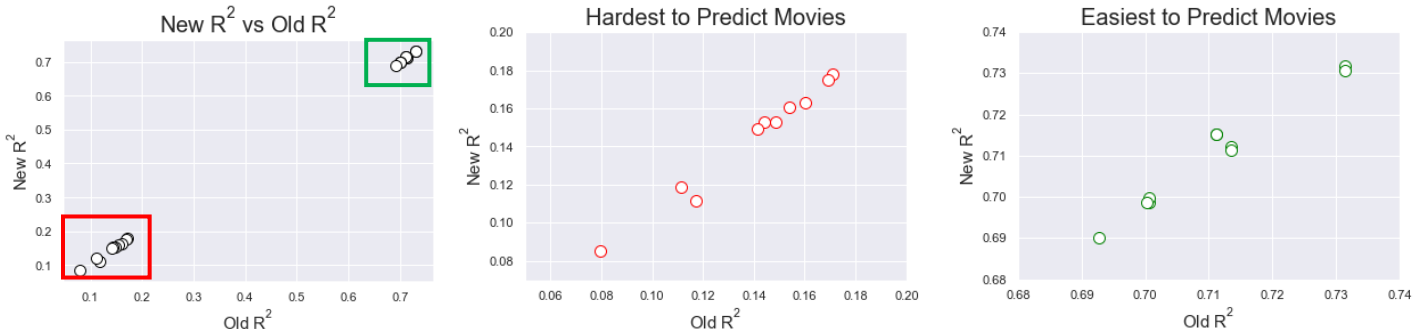


Table 2: RIDGE

Movie	RMSE	alpha
Let the Right One In (2008)	0.107464	117.6
The Machinist (2004)	0.147840	99.6
Man on Fire (2004)	0.187638	155.0
Crossroads (2002)	0.176966	132.8
The Poseidon Adventure (1972)	0.135833	41.5
The Rock (1996)	0.122046	63.1
Gone in Sixty Seconds (2000)	0.134360	41.2
Blues Brothers 2000 (1998)	0.151285	80.0
Equilibrium (2002)	0.120634	137.5
The Blue Lagoon (1980)	0.132433	102.5
The Good the Bad and the Ugly (1966)	0.153831	80.3
Dances with Wolves (1990)	0.186173	48.6
The Evil Dead (1981)	0.127025	47.0
Just Married (2003)	0.137995	97.0
Goodfellas (1990)	0.115259	63.3
Uptown Girls (2003)	0.204132	114.8
Knight and Day (2010)	0.189644	120.2
Austin Powers: The Spy Who Shagged Me (1999)	0.299270	156.7
The Prestige (2006)	0.118105	97.4
The Big Lebowski (1998)	0.159304	96.5
Reservoir Dogs (1992)	0.182937	112.5
Austin Powers in Goldmember (2002)	0.230734	152.0
The Mummy Returns (2001)	0.252217	97.9
28 Days Later (2002)	0.166744	85.1
The Green Mile (1999)	0.128342	116.8
You're Next (2011)	0.132307	104.2
Men in Black (1997)	0.300948	112.9
Men in Black II (2002)	0.311985	134.6
The Mummy (1999)	0.320630	139.1
Twister (1996)	0.092081	79.6

Q3: Pick 30 movies in the middle of the COD range and build a regularized regression model (RIDGE Regression) with the ratings from 10 other movies as an input

Ridge regression is a type of linear regression that analyzes and models data that has multiple correlated predictors and adds a penalty term that introduces some bias (underfitting) to considerably reduce variance (overfitting). It will shrink the coefficients/weights of all predictor variables towards zero; it produces a model with all predictors included but with small coefficients and takes care of multicollinearity. We picked 30 movies in the middle of the COD range, as identified in Q1 (that were not used as target movies in Q2) and built a ridge regularized regression model with the ratings from 10 randomly picked movies (Django Unchained (2012), Poltergeist (1982), Godzilla (1998), Dead Poets Society (1989), The Shining (1980), Cheaper by the Dozen (2003), My Big Fat Greek Wedding (2002), A Bug's Life (1998), Spirited Away (2001), and How the Grinch Stole Christmas (2000)) and with suitable hyperparameter (α) tuning. The data sets for each model were split into 80/20 train/test sets, and Grid Search was used to find the most suitable α for each model by further splitting the training set into 80/20 train/validation sets. Using these best α values and their respective train/test sets, we built ridge regression models for each of the 30 movies. For the 30 models, we generated RMSE and 10 β coefficient values that correspond to each of the 10 movies. RMSE values ranged from 0.092 to 0.321. Optimal α values found for each model ranged from 41.2 to 156.7. The larger the α value the more aggressive the penalization/prevention of overfitting in the model is. The weights (β coefficients) of each predictor movie for the 30 models ranged from 0.014 to 0.158. The higher the β value, the greater the association is between the specific predictor movie and the target movie. Since the α value denotes the amount of shrinkage applied to the β coefficients, it makes sense that the β values found are generally small while the α values are generally large. Table 2 reports the RMSE and associated α hyperparameter for target movies under Ridge regression. ***Alternate method for setting α to 100.91 (the average value of the best α values for all 30 ridge regression models) to decrease bias yielded similar RMSE and β coefficient values and is included in the Jupyter Notebook for reference.*

Table 3: LASSO

Movie	RMSE	alpha
Let the Right One In (2008)	0.105323	0.00340
The Machinist (2004)	0.142736	0.00050
Man on Fire (2004)	0.183646	0.00740
Crossroads (2002)	0.171576	0.00095
The Poseidon Adventure (1972)	0.134527	0.00000
The Rock (1996)	0.120073	0.00000
Gone in Sixty Seconds (2000)	0.132512	0.00000
Blues Brothers 2000 (1998)	0.147531	0.00140
Equilibrium (2002)	0.118653	0.00519
The Blue Lagoon (1980)	0.132770	0.00346
The Good the Bad and the Ugly (1966)	0.156663	0.00419
Dances with Wolves (1990)	0.185452	0.00000
The Evil Dead (1981)	0.130358	0.00028
Just Married (2003)	0.139678	0.00452
Goodfellas (1990)	0.117777	0.00653
Uptown Girls (2003)	0.205994	0.00246
Knight and Day (2010)	0.189508	0.00500
Austin Powers: The Spy Who Shagged Me (1999)	0.300225	0.00381
The Prestige (2006)	0.121017	0.00555
The Big Lebowski (1998)	0.160219	0.00331
Reservoir Dogs (1992)	0.181873	0.00664
Austin Powers in Goldmember (2002)	0.234197	0.00678
The Mummy Returns (2001)	0.255087	0.00000
28 Days Later (2002)	0.166382	0.00232
The Green Mile (1999)	0.126366	0.00161
You're Next (2011)	0.133288	0.00045
Men in Black (1997)	0.305025	0.00338
Men in Black II (2002)	0.312278	0.00540
The Mummy (1999)	0.324519	0.00670
Twister (1996)	0.094380	0.00401

Q4: Repeat Q3 with LASSO regression

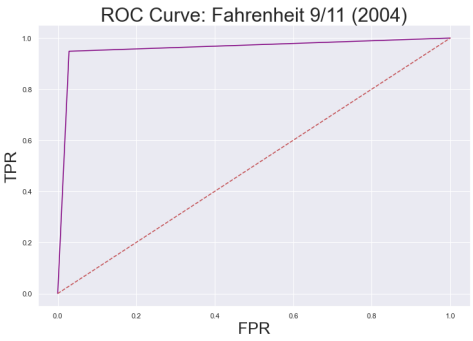
Lasso regression is a type of linear regression that analyzes and models data that has multiple correlated predictors and adds a penalty term that introduces some bias to considerably reduce variance. In contrast to ridge, lasso will set some coefficients to zero, enforcing models with fewer predictors. It makes for a simpler model and takes care of multicollinearity. We used the same 30 movies in the middle of the COD range, 10 randomly picked movies, and hyperparameter tuning method (Grid Search, training/validation/test sets) from Q3 and built a

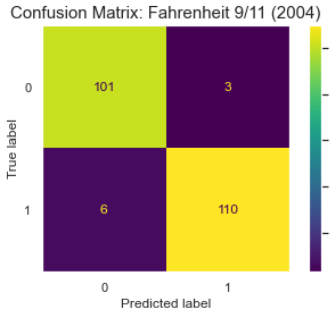
lasso regularized regression model for each of the 30 movies. For the 30 models, we again generated RMSE and 10 β coefficient values corresponding to each of the 10 movies. RMSE values ranged from 0.094 to 0.325. Optimal α values found for each model ranged from 0.00 to 0.0074. The weights (β coefficients) of each predictor movie for the 30 models ranged from 0.00 to 0.177. The lower the β value, the weaker the association is between the specific predictor movie and the target movie. In contrast to the ridge regression model, the α hyperparameter values are very small and there are more β values that are closer to or equal to 0. The smaller the α value the less aggressive the penalization/prevention of overfitting in the model is. When α is 0, Lasso regression produces the same coefficients as a linear regression (ordinary least squares). When β coefficients equal 0 it means that those corresponding movies' ratings were not included in their respective final model because they did not influence the specific outcome/target movie; predictor movies with a coefficient of 0 are removed to simplify the models. We also notice that 16/30 of the models' RMSE values had a very slightly increase from the ridge to lasso method but in general there wasn't much of a difference between the ridge and lasso method RMSEs. Table 3 reports the RMSE and associated α hyperparameter for target movies under Lasso regression. ***Alternate method for setting α to 0.00318 (the average value of the best α values for all 30 lasso regression models) to decrease bias yielded similar RMSE and β coefficient values and is included in the Jupyter Notebook for reference.*

Q5: Use the average movie enjoyment for each user to predict the enjoyment of 4 movies in the middle of movie rating average score range using a logistic regression model

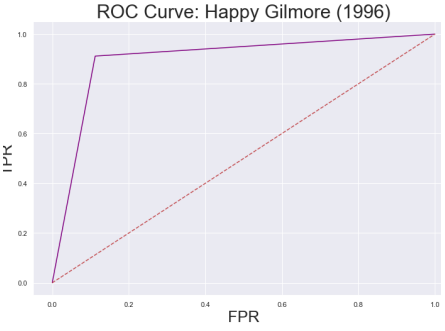
Logistic regression is used for predicting binary outcomes; it uses a logistic (nonlinear) function to model the probability of an event (target variable) occurring, as a function of one or more predictor variables. Average movie enjoyment for each user was computed using real, non-imputed data. Average rating for each movie was also computed using real, non-imputed data and movies were sorted in increasing order based on average rating. The 4 movies found in the middle of the average rating score range were Fahrenheit 9/11 (2004), Happy Gilmore (1996), Diamonds are Forever (1971), and Scream (1996). Using a median split, we coded the movies above their median rating with label 1 (= enjoyed) and movies below with label 0 (= not enjoyed) and built a logistic regression for each movie using the average movie enjoyment as the predictor. An 80/20 train/test split was used as our cross-validation method to avoid overfitting. The following β coefficients, AUC, and model accuracy values were generated: [Fahrenheit 9/11 (2004): $\beta_1 = 7.154$, AUC = 0.9597, accuracy = 0.959], [Happy Gilmore (1996): $\beta_1 = 5.135$, AUC = 0.8997, accuracy = 0.900], [Diamonds are Forever (1971): $\beta_1 = 7.217$, AUC = 0.9406, accuracy = 0.941], [Scream (1996): $\beta_1 = 4.354$, AUC = 0.8346, accuracy = 0.832]. The AUC values for all 4 models are generally high/close to 1, which means that our classifiers can distinguish between their two classes very well and the classifier has a low false positive and false negative rate. Specifically, we see that average movie enjoyment is a good predictor of whether users would enjoy these 4 movies. The accuracies for each model are also generally high as well, which means that our models have learned the relationship between their input and the target variables and can use this to make accurate predictions on new data. Figure 3 shows the ROC curves and confusion matrices for the 4 target movies based on average movie enjoyment.

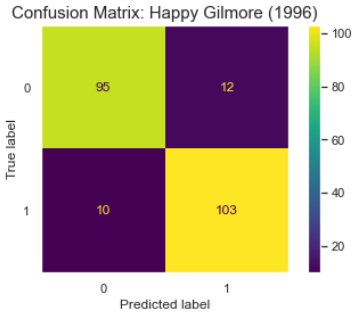
Figure 3: ROC Curves and Confusion Matrices



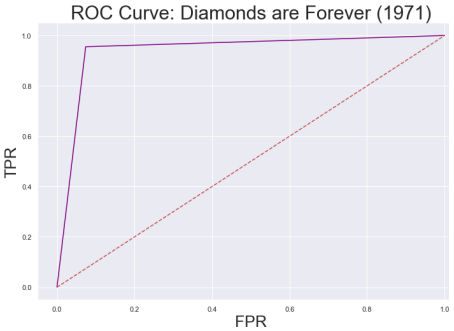


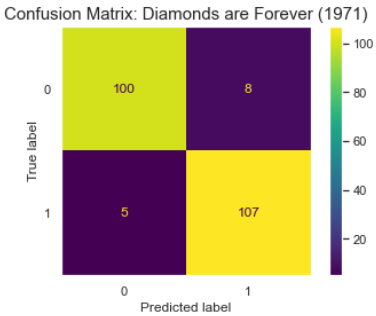
$\beta_1 = 7.154$, AUC = 0.9597, accuracy = 0.959



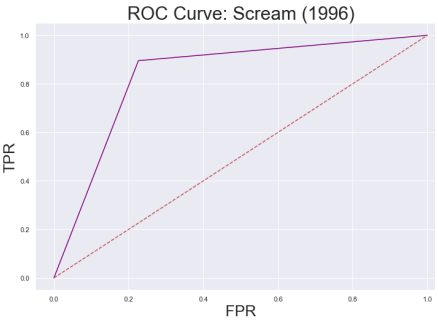


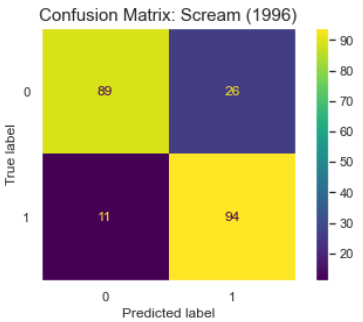
$\beta_1 = 5.135$, AUC = 0.8997, accuracy = 0.900





$\beta_1 = 7.217$, AUC = 0.9406, accuracy = 0.941





$\beta_1 = 4.354$, AUC = 0.8346, accuracy = 0.832

EXTRA: Choose 4 horror/thriller movies and build a lasso regularized regression model using the responses from various self-assessment/personality questions for each user as input and report observations on response influence in the model

We used 4 horror/thriller movies (The Blair Witch Project (1999), Ouija: Origin of Evil (2016), Shutter Island (2010), The Exorcist (1973)) and built a lasso regularized regression model for each of the 4 movies using user responses to 10 self-assessment/personality questions: 'I enjoy

watching horror movies', 'I like to be surprised even if it startles or scares me', 'I enjoy haunted houses', 'Is emotionally stable/not easily upset', 'Can be cold and aloof', 'No risk - No fun', 'I had a sheltered upbringing', 'Has an active imagination', 'Remains calm in tense situations', and 'Is depressed/Blue'. Figure 4 shows the correlation between these various questions and their responses. We observe that responses to questions 'I enjoy watching horror movies', 'I like to be surprised even if it startles or scares me', and 'enjoy haunted houses', are amongst the most correlated. The penalty term in the lasso regression model will address this multicollinearity; it will help shrink the coefficients of these correlated features to reduce their influence on the model and increase interpretability of the model. For the 4 models, we generated RMSE and 10 β coefficient values corresponding to each of the 10 self-assessment/personality questions. RMSE values ranged from 0.369 to 0.522. Optimal α values found for each model ranged from 0.0231 to 0.0857. The weights (β coefficients) of each predictor for the 4 models ranged from 0.028 to 0.145. We notice that the β coefficients for all responses have the same magnitude across the 4 target movies, which means that these responses are weighed the same by these 4 movies. However, we do see that the responses with the highest β values are 'I like to be surprised even if it startles or scares me', 'I enjoy watching horror movies', 'I had a sheltered upbringing', 'I enjoy haunted houses' and 'Can be cold and aloof'. Responses to questions such as 'No risk - no fun' and 'Remains calm in tense situations' had lower β values and therefore these responses hold less influence in predicting the ratings of the target movies. Table 4 reports the RMSE and associated α hyperparameter for the target movies under Lasso Regression and Table 5 shows the β coefficients for each self-assessment/personality question in relation to the target movie.

Figure 4: Self-Assessment/Personality Responses Heatmap

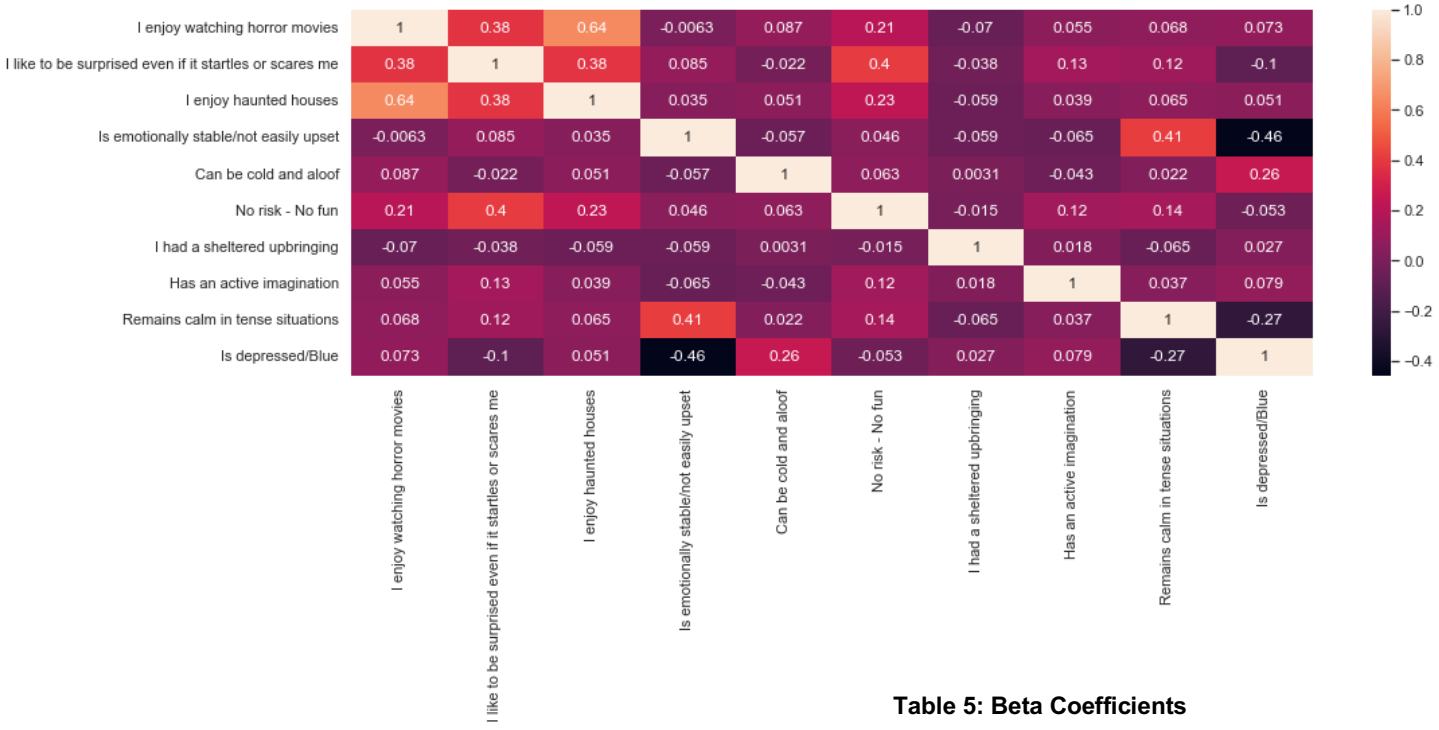


Table 4: RMSE and Alpha

Movie	RMSE	alpha
The Blair Witch Project (1999)	0.434041	0.0748
Ouija: Origin of Evil (2016)	0.368878	0.0857
Shutter Island (2010)	0.434947	0.0261
The Exorcist (1973)	0.522405	0.0231

Table 5: Beta Coefficients

	I enjoy watching horror movies	I like to be surprised even if it startles or scares me	I enjoy haunted houses	Is emotionally stable/not easily upset	Can be cold and aloof	No risk - No fun	I had a sheltered upbringing	Has an active imagination	Remains calm in tense situations	Is depressed/Blue
The Blair Witch Project (1999)	0.126886	0.145169	0.089106	0.051561	0.082651	0.027925	0.092332	0.042854	0.030811	0.058021
Ouija: Origin of Evil (2016)	0.126886	0.145169	0.089106	0.051561	0.082651	0.027925	0.092332	0.042854	0.030811	0.058021
Shutter Island (2010)	0.126886	0.145169	0.089106	0.051561	0.082651	0.027925	0.092332	0.042854	0.030811	0.058021
The Exorcist (1973)	0.126886	0.145169	0.089106	0.051561	0.082651	0.027925	0.092332	0.042854	0.030811	0.058021