FA-25 DS-GA 1001
Interview in Data Science

Capstone Project

Assessing Professor Effectiveness (*APE)*

Group: Luke Ducker, Joseph Tadros, Beibarys Nyussupov

Abstract: This report assesses university professors using a large, publicly available dataset scraped from RateMyProfessor.com. The objective is to integrate core concepts from the course into a single applied analysis and to extract actionable insights about teaching quality and student perceptions. The dataset contains aggregated student ratings and related attributes for a broad sample of professors, with low individual response rates but substantial overall scale. Prior research reports a correlation of approximately 0.7 between RateMyProfessor ratings and official end-of-course teaching evaluations, which supports the analytical value of this source despite known response bias. All data collection and basic structuring steps were completed in advance, while data science-relevant preprocessing, including the identification and handling of missing data, was performed in this project. To address the research questions, appropriate statistical methods were selected based on underlying assumptions, with explicit justification provided throughout the report. Visualizations support interpretation and highlight key patterns in the data. All hypothesis testing uses $\alpha = 0.005$ to reduce false positive findings (Habibzadeh, 2025).

Contents:

1. Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to a low sample size – as small as n = 1 (Mitchell & Martin, 2018), failure to control for confounders such as teaching experience (Centra & Gaubatz, 2000) or obvious p-hacking (MacNell et al., 2015). We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset.
2. Is there a gender difference in the spread (variance/dispersion) of the ratings distribution?
3. What is the likely size of both of these effects (gender bias in average rating, gender bias in spread of average rating), as estimated from this dataset? Please use 95% confidence and make sure to report each/both.
4. Is there a gender difference in the tags awarded by students? Make sure to test each of the 20 tags for a potential gender difference and report which of them exhibit a statistically significant difference. Comment on the 3 most gendered (lowest p-value) and least gendered (highest p-value) tags.
5. Is there a gender difference in terms of average difficulty? Again, a significance test is indicated.
6. Please quantify the likely size of this effect at 95% confidence.
7. Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R2 and RMSE of this model. Which of these factors is most strongly predictive of average rating? Hint: Make sure to address collinearity concern.
8. Build a regression model that predicts average ratings from all tags (those in the rmpCapstoneTags.csv file). Make sure to include the R2 and RMSE of this model. Which of these tags is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns. Also, comment on how this model compares to the previous one
9. Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty? Hint: Make sure to address collinearity concern.
10. Build a classification model that predicts whether a professor receives a “pepper” from all available factors (both tags and numerical). Make sure to include model quality metrics such as AU(RO)C and also address class imbalance concerns.

Limitations of the data: The data used in this project is observational and survey-based rather than experimental. The sample is drawn from voluntary RateMyProfessor submissions and is not randomized, so results do not necessarily generalize to the full population of professors. Independence of observations is not guaranteed, as ratings reflect clustered student experiences within courses, departments, and institutions. Several key variables, including ratings and tags, are ordinal or bounded, which limits the use of parametric methods and weakens normality assumptions. Additionally, average ratings are influenced by the number of underlying reviews and are less reliable for professors with fewer ratings. For these reasons, all findings should be interpreted as associations in observational data rather than causal effects.

Pre-processing made:

- Since ratings are ordinal and should not be reduced to means, we use non-parametric tests for most of the analysis.
- For reproducibility, the N number of Beibarys Nyussupov (12250697) was used as the random seed.
- To maximize the reliability of average ratings and ensure they accurately represent the population, professors with fewer than five ratings were excluded from the dataset.
- To control for differences in rating volume when using teaching style tags, raw tag counts were normalized by dividing each tag by the total number of tags received by a professor. This converts tags into proportions and makes them comparable across professors. Professors with zero total tags resulted in undefined values, which were set to zero to reflect the absence of tag information.
- Professors with average ratings missing were treated as data errors and removed from the dataset.
- Professors with inconsistent gender indicators, defined as both 'Female' and 'Male' equal to 0 or both equal to 1, were treated as data errors and removed from the dataset.
- Sample size: 10,015 males, 8407 females

Analysis:

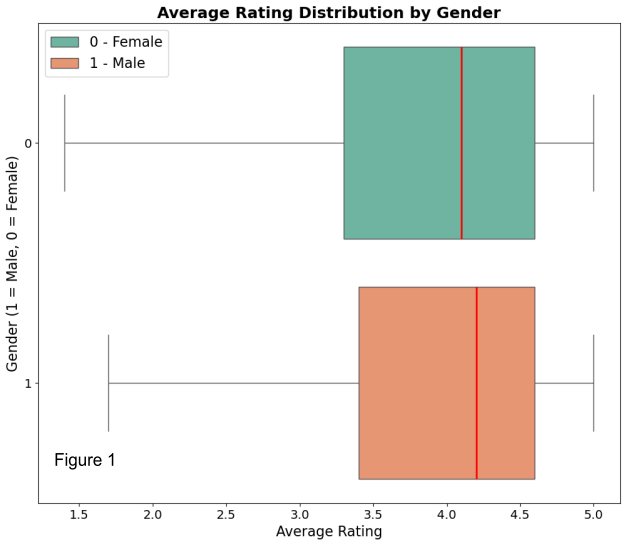


Figure 1

1. The distributions for male and female professors substantially overlap, but the male distribution appears slightly shifted to the right, with a higher median average rating. We used a one-tailed Mann-Whitney U test to compare average ratings by gender. This test is appropriate because the average ratings are non-normal; however, the similarity in distributions of the two groups is sufficient for a median difference comparison.

Null hypothesis (H0): Average ratings for male and female professors are the same.

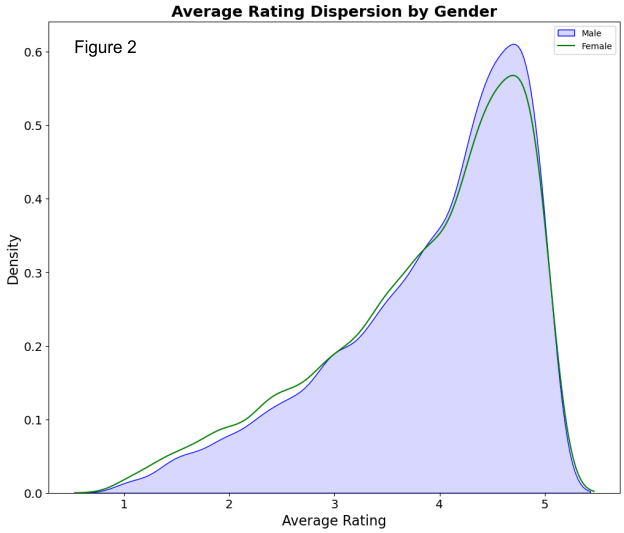
Alternative hypothesis (H1): Average ratings for male professors are significantly higher than those for female professors. The **p-value is 0.00024523** < 0.005; therefore, we reject the null hypothesis and conclude that male professors have higher average ratings than female professors. This does not prove pro-male bias, since the data is observational and confounders are not fully controlled. The results provide evidence of a gender difference in ratings, consistent with a potential pro-male bias.

2. The kernel density plot reveals that the male distribution is slightly shifted to the right, consistent with higher average ratings, but the overall spread and shape of the two distributions closely align. Visually, we can not observe any significant and/or major differences in variance. To test the dispersion of average

ratings by gender, we use a two-tailed Levene's test with a median function. Levene's test is suitable for average ratings, as it is robust to normality violations and can utilize both median and mean metrics for comparison.

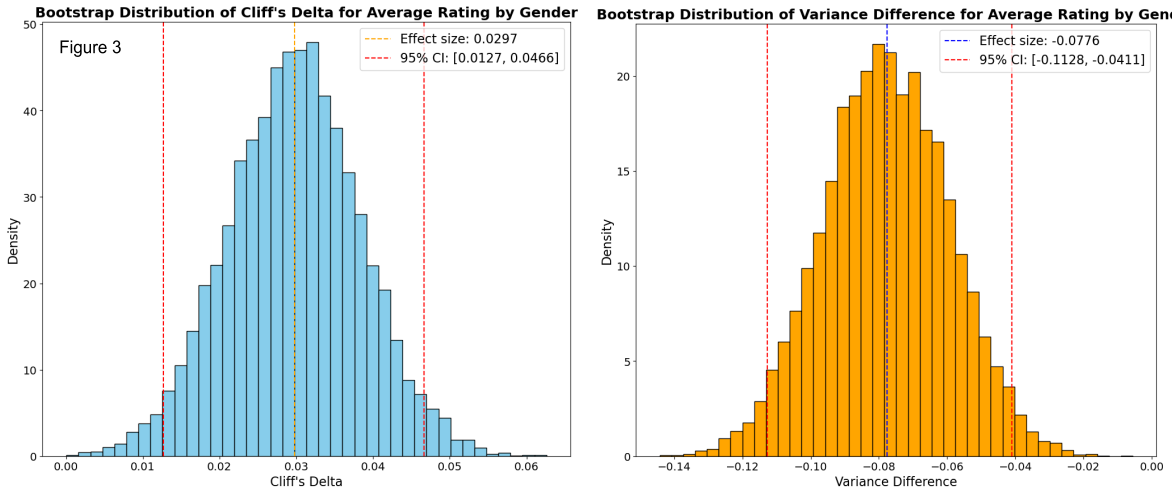
Null hypothesis (H0): Average ratings dispersion of male and female professors is the same.

Alternative hypothesis (H1): Average ratings dispersion of male professors is significantly different from average ratings dispersion of female professors. Our **p-value 5.97717e-06** < 0.005, indicates that the observed significance is unlikely to have occurred by chance. Therefore, we reject the null hypothesis and conclude that there is a significant difference in the dispersion of male and female professors' average ratings.



3. To estimate the effect size of the central tendency difference in professor ratings by gender, we used non-parametric Cliff's delta and computed a 95% confidence interval using bootstrap sampling with 10,000 iterations. Cliff's delta is a non-parametric effect size measure that quantifies the degree of distributional non-overlap between two groups on ordinal or non-normal data. Cliff's delta does not assume normality and equal variances of the data. It follows the given interpretation: (Meissel & Yao, 2024)

- |δ| < 0.15 (less than 15% non-overlap between groups 1 and 2) = negligible effect
- 0.15 <= |δ| < 0.33 = small effect
- 0.33 <= |δ| < 0.47 = medium effect
- |δ| >= 0.47 = large effect



A 95% confidence interval means that, over repeated samples, 95% of such intervals would contain the true Cliff's delta. In this case, the estimated Cliff's delta lies between (**0.0127, 0.0466)**, with a point estimate of **0.0297**. This corresponds to a negligible effect size, meaning that the distributions of average ratings for male and female professors are almost entirely overlapping. Although this difference is statistically detectable given the large sample size, it does not represent a practical difference in central

tendency and provides little evidence of substantive gender bias in average ratings. To estimate the effect size of the dispersion difference in professors' ratings by gender, we computed the difference in sample variances by subtracting the variance of female professors' average ratings from that of male professors' average ratings. This produced a point estimate of **-0.0776**. Uncertainty was quantified using a 95% bootstrap confidence interval, yielding **(-0.1128**, **-0.0411)**. The interval lies entirely below zero, indicating slightly greater variability in average ratings for female professors. However, the magnitude of the effect is negligible, suggesting that any difference in rating dispersion by gender is unlikely to be practically meaningful.

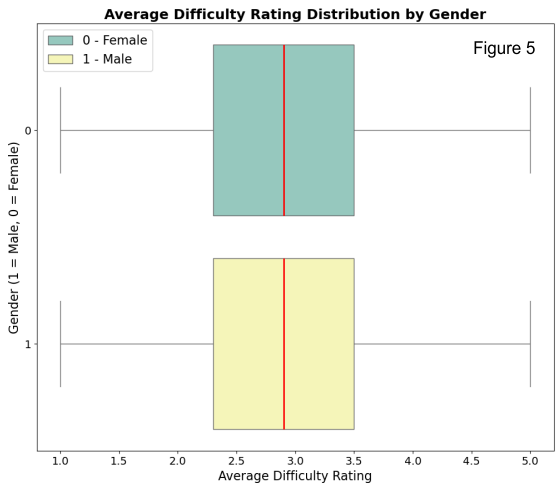
4. Since we converted each tag variable to a proportion of all tags awarded to a professor, our tag proportions are skewed. It means that we cannot assume normality and equal variance. That is why a non-parametric two-tailed Mann-Whitney U test was used to test a gender difference for each tag awarded by students.

Figure 4

	tag	n_female	n_male	U	p_value	significant
8	pop_quizzes	8407	10015	42356129.000000	2.698978e-01	False
9	accessible	8407	10015	42929904.500000	9.133809e-03	False
0	tough_grader	8407	10015	41056014.500000	2.637471e-03	True
11	clear_grading	8407	10015	40888751.000000	5.379880e-04	True
7	inspirational	8407	10015	43455897.500000	4.188542e-05	True
5	dont_skip_class_or_you_will_not_pass	8407	10015	40573252.500000	1.282860e-05	True
13	test_heavy	8407	10015	43444751.000000	8.194516e-08	True
3	lots_to_read	8407	10015	40243731.000000	5.192685e-08	True
6	lots_of_homework	8407	10015	40193393.000000	1.974495e-08	True
10	so_many_papers	8407	10015	40325013.000000	1.719188e-13	True
17	extra_credit	8407	10015	39335036.000000	3.966123e-18	True
14	graded_by_few_things	8407	10015	44541050.500000	4.776183e-21	True
1	good_feedback	8407	10015	38583224.500000	8.064891e-23	True
18	group_projects	8407	10015	39419701.500000	8.601993e-24	True
19	lecture_heavy	8407	10015	45633652.000000	3.316467e-27	True
16	caring	8407	10015	38004532.000000	1.225870e-30	True
2	respected	8407	10015	46154727.500000	7.608957e-31	True
4	participation_matters	8407	10015	37922544.000000	1.360882e-32	True
15	amazing_lectures	8407	10015	46866786.500000	1.333388e-45	True
12	hilarious	8407	10015	51349863.000000	4.538081e-171	True

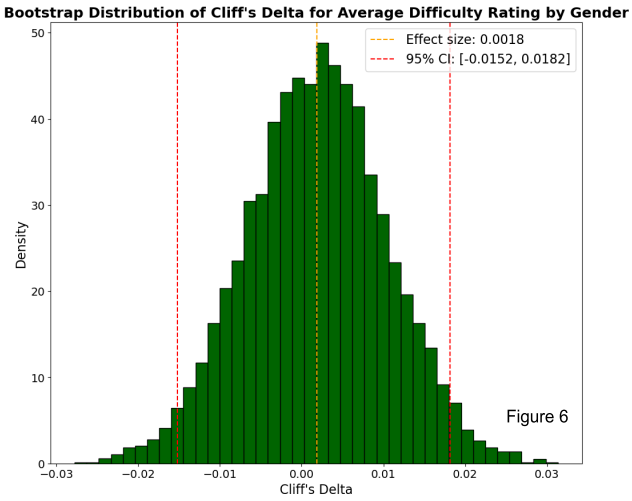
Null hypothesis (H0): The distribution of tag proportions is the same for male and female professors.
Alternative hypothesis (H1): The distribution of tag proportions significantly differs between male and female professors.
We tested all 20 tag proportions for gender differences using two-tailed Mann-Whitney U tests at a significance level of 0.005. 18 of 20 tags show statistically detectable gender differences. The least gendered tags are **pop quizzes** and **accessible,** which have the highest p-values and show no evidence of a gender difference. The most gendered tags are **hilarious**, **amazing lectures**, and **participation matters**, which have the lowest p-values and the strongest statistical evidence of a difference in tag proportions by gender.

5. The box plot of average difficulty ratings does not have clear visual differences. We used a two-tailed Mann-Whitney U test to compare average difficulty ratings by gender. This test is appropriate again because average difficulty ratings are ordinal and non-normal, and the distributions are similar.



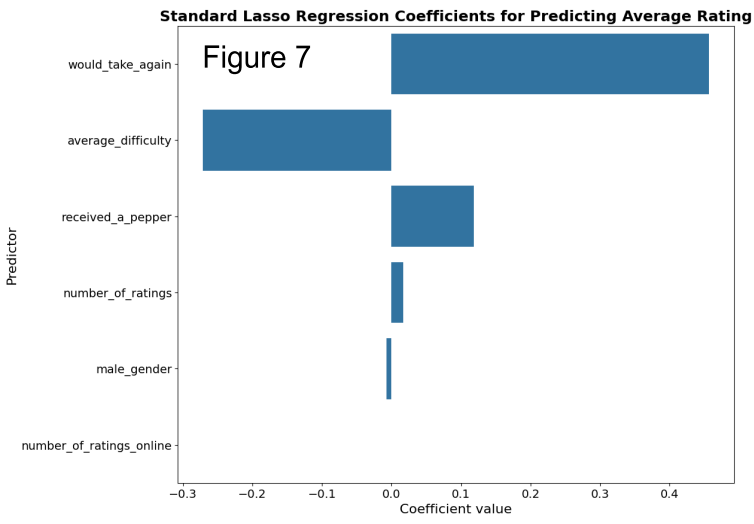
Null hypothesis (H0): Average difficulty ratings for male and female professors are the same.
Alternative hypothesis (Ha): Average difficulty ratings for male professors are significantly different from average difficulty ratings of female professors. The Mann-Whitney U test yielded a p-value of **0.828753**, which is greater than 0.005; therefore, we fail to reject the null hypothesis. This indicates that the observed differences in average difficulty ratings are consistent with random variation, and the data provide no statistical evidence of a gender difference in perceived course difficulty by students. It is again crucial to remember that this does not prove or disprove pro-male bias, as the data are observational and confounders are not fully controlled.

6. Because our average difficulty ratings are also ordinal and non-normal, we again do not use parametric effect sizes, such as Cohen's d or Hedges' g, as these assume normality and equal variance of the data. To estimate the effect size for our average difficulty ratings of professors, we used non-parametric Cliff's delta and computed a 95% confidence interval using bootstrap sampling with 10,000 iterations. Cliff's delta is relevant, since it does not assume normality and equal variances of the data. Cliff's delta is a non-parametric effect size measure that quantifies the degree of distributional non-overlap between two groups on ordinal or non-normal data.



After computing a 95% Confidence Interval using Bootstrap sampling, the estimated Cliff's delta lies between **(-0.0152, 0.0182)**, with a point estimate of **0.0018.** This corresponds to a negligible effect size, indicating that the distributions of average difficulty ratings for male and female professors are almost entirely overlapping. The interval includes zero, providing no practical evidence of a meaningful gender difference in perceived course difficulty.

7. We split the dataset into training and test sets, allocating 80% of the observations to training and the remaining 20% to testing. All preprocessing steps were applied consistently across models. Because the dataset contains missing values, we applied K-Nearest Neighbor imputation with k=20 neighbors. Since KNN imputation is distance-based, we first standardized all numerical predictors to have a mean of 0 and a standard deviation of 1, ensuring that distances between observations were not dominated by scale differences across features. We additionally examined pairwise correlations among all numerical predictors and found no evidence of severe multicollinearity, as no correlation magnitude exceeded 0.6 (see Appendix 1). We employed Lasso regression to improve model stability and perform feature selection. We first fit a LassoCV model using all numerical predictors in the dataset. Under cross-validation, this full model achieved an average R^2 of approximately 0.601 and an RMSE of about ~0.584. On the held-out test set, performance remained comparable, with a test RMSE of approximately ~0.583 and a test R^2 of about ~0.617, indicating reasonable generalization. However, inspection of the fitted coefficients (see Figure 7) shows that Lasso shrinks most predictors toward zero, retaining meaningful weight only on a small subset of variables. In particular, the proportion of students who would take the professor again (**would_take_again**) has by far the largest coefficient magnitude, followed by **average_difficulty** and **received_a_pepper**, while variables such as the number of ratings, number of online ratings, and gender



receive coefficients close to zero. This coefficient pattern suggests that although the full model performs adequately, most numerical predictors contribute little additional explanatory power beyond the dominant sentiment and difficulty variables. To improve interpretability and generalization, we retained only predictors with absolute coefficients of at least 0.05. This threshold yielded a reduced feature set comprising three variables: the proportion of students who **would take the professor again** (**would_take_again**), the **average perceived difficulty (average_difficulty)**, and **whether the professor received a pepper (received_a_pepper)**. We then refit the LassoCV model using only these selected predictors. The reduced Lasso model performed better than the full model. In cross-validation, it achieved an R^2 of ~0.616 and an RMSE of ~0.572. On the test set, the reduced model further improved generalization, with a test RMSE of ~0.569 and a test R^2 of about ~0.635. This improvement indicates that removing weak, potentially collinear predictors led to a more stable, better-performing model.

8. We followed the same preprocessing workflow as in Q7. After removing professors with fewer than five reviews, the tag predictors contain no missing values, so no imputation was required. All tag variables were standardized to ensure that coefficient magnitudes are directly comparable across predictors. Before fitting any models, we examined pairwise correlations among the tag variables to assess multicollinearity. The correlation heatmap in Appendix 2 shows that no pairwise correlation exceeds 0.4. While several tags capture related aspects of teaching quality or workload, the degree of overlap is moderate and does not warrant decorrelation or feature removal at this stage. We therefore retained all 20 tag predictors in the initial model. We then fit multiple linear regression models predicting **average_rating** from all tag variables, comparing ordinary least squares, Ridge, Lasso, and ElasticNet under a common 5-fold cross-validation setup. All four models performed almost identically. Standard multiple linear regression achieved the best performance, with a cross-validated RMSE of approximately 0.50 and a cross-validated R^2 of about 0.71. Evaluation on the held-out test set confirms this result, with test RMSE = 0.504 and test R^2 = 0.713, indicating stable generalization and no evidence of overfitting. Regularization did not improve predictive accuracy. As the regularization strength increased, RMSE increased monotonically, suggesting that shrinkage removes useful signal rather than reducing variance. Despite this, we proceeded with Lasso regression to support interpretation. Lasso provides a multivariate view of feature importance by shrinking weak predictors toward zero while accounting for shared variance across tags. **Figure 8** shows the standardized Lasso coefficients from the full 20-variable model. While many predictors have small coefficients, several tags consistently emerge as dominant. To balance interpretability and performance, we evaluated multiple coefficient cutoffs and refit the model under each reduced subset. Results show that the model can be reduced to 12 predictors using a threshold of $|\textbf{coefficient}| \geq 0.05$ with only a modest loss in accuracy. More aggressive reduction leads to substantial degradation in both RMSE and R^2. On cross-validation, the 11-variable Lasso achieves RMSE ~ 0.509 and R^2 ~ 0.70. On the test set, it achieves RMSE ~ 0.515 and R^2 ~ 0.70 compared to RMSE ~ 0.504 and R^2 ~ 0.714 for the full model. The final reduced model is shown in **Figure 9**. Teaching quality clearly dominates. The strongest positive

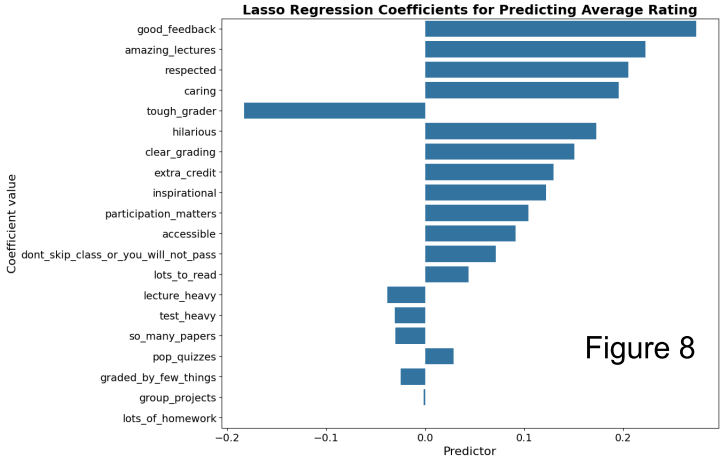


Figure 8

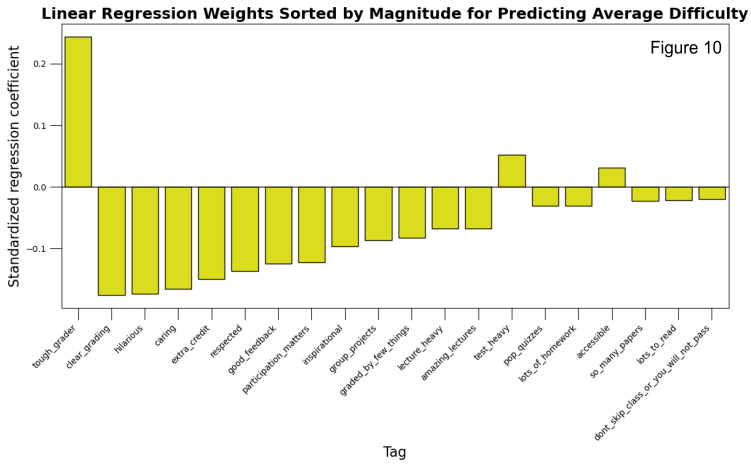
predictors of the average rating are good_feedback, amazing_lectures, respected, and caring, each with the largest standardized coefficient. These tags capture clarity, engagement, instructor support, and perceived respect, and together explain most of the positive variation in student evaluations. **Hilarious** also has a meaningful, positive effect, indicating that an engaging classroom atmosphere contributes to instructional quality beyond itself. In contrast, **tough_grader** is the strongest negative predictor in the model, reflecting the well-known penalty for strict grading. Other structural or behavioral tags, including **clear_grading, extra_credit, participation_matters,** and **accessible**, contribute positively but to a lesser extent, indicating secondary effects once core teaching quality is accounted for. In summary, the most strongly predictive tags of average rating are those capturing perceived teaching quality, particularly **good_feedback, amazing_lectures, respected, and caring, while tough_grader** is the single most influential negative predictor. The final 12-variable Lasso model provides a **parsimonious and interpretable summary** of the tag space while retaining most of the predictive performance.

Figure 9

	tag	weight
0	good_feedback	0.269108
1	amazing_lectures	0.221601
2	respected	0.197884
3	tough_grader	-0.189973
4	caring	0.185204
5	hilarious	0.165205
6	clear_grading	0.138021
7	extra_credit	0.120893
8	inspirational	0.113642
9	participation_matters	0.106749
10	accessible	0.084162

9. We first observed correlations between the dependent variable (average difficulty) and the independent variables (tags) (Appendix 3). The “tough grader” tag had the highest positive correlation (~0.65), while the other predictors had lower correlations. We then examined correlations among the independent variables to determine whether ridge regression is necessary. We interpret high multicollinearity among independent variables as exceeding 0.70 (Appendix 4). No multicollinearity was found, but we proceeded with fitting data to multiple types of regressions - OLS, Lasso, Ridge, and ElasticNet to evaluate performance. We split the data into training and test sets, with 80% allocated to training and 20% to testing. We chose the final model based on 5-fold cross-validation on the training set and on held-out test-set performance. All models performed identically in terms of RMSE and R^2, so we chose Lasso regression for interpretability and feature selection, using standardized coefficient weights. The lasso regression model achieved a cross-validated R^2 of 0.55 with an RMSE of 0.54, and a test-set R^2 of 0.54 with an RMSE of 0.54, indicating stable generalization and no evidence of overfitting. Among all tags, **tough_grader** is the most strongly predictive of average difficulty, with the largest standardized regression coefficient (~**0.244**).

To further increase the parsimony and interpretability of the model, we sought to reduce the number of predictors while minimizing increases in RMSE by pruning predictors using standardized lasso regression weights. We reduced the model to 13 predictors while maintaining a cross-validated RMSE of 0.54 and an R^2 of



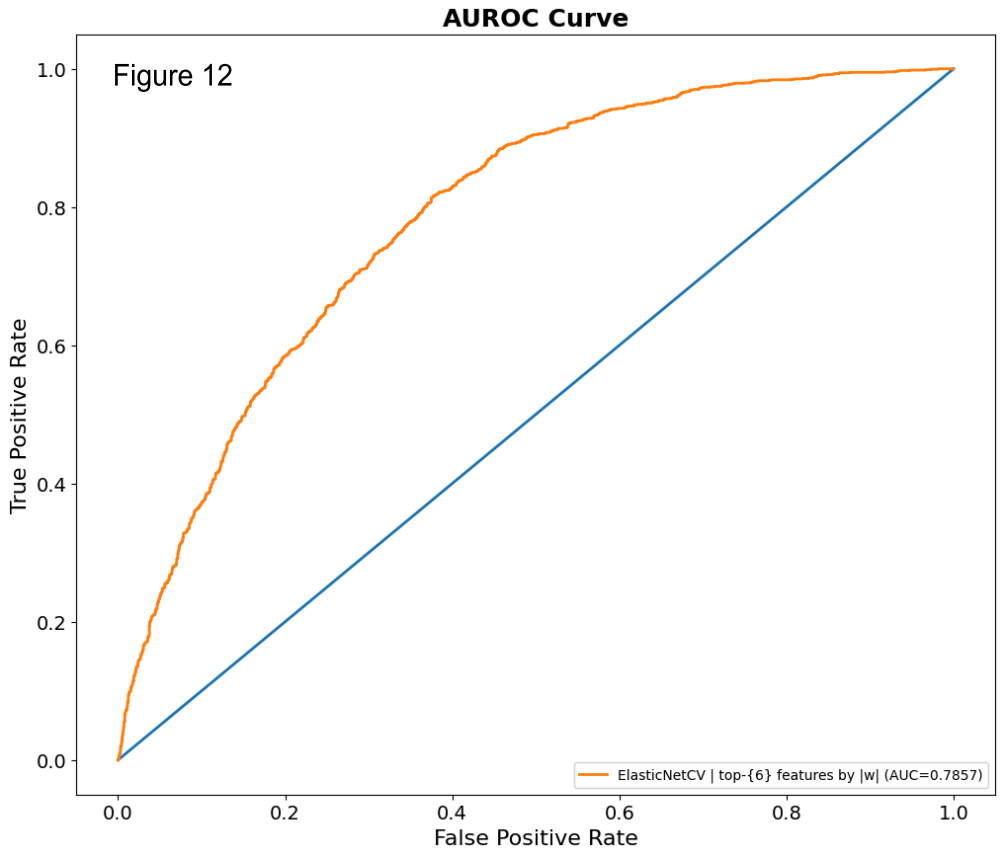
0.54. On the holdout set, reduced lasso regression yielded an RMSE of 0.54 and an R^2 of 0.54, indicating that no underfitting occurred and that 13 variables capture most of the data variance. The final model is shown in Figure 11, where the three most significant predictors are **tough_grader (0.30)**, **clear_grading (-0.14)**, and **hilarious (-0.14)**.

Figure 11	tag	weight
0	tough_grader	0.306377
1	clear_grading	-0.142697
2	hilarious	-0.140308
3	caring	-0.122311
4	extra_credit	-0.116583
5	respected	-0.103706
6	participation_matters	-0.075322
7	inspirational	-0.073372
8	good_feedback	-0.072360
9	graded_by_few_things	-0.068707
10	test_heavy	0.067897
11	group_projects	-0.059421
12	accessible	0.052234

10. Before fitting the logistic regression model to predict whether a professor receives a "pepper," we checked correlations to identify powerful predictors and assess multicollinearity. When it comes to correlations between predictors and the dependent variable, only the average rating **(~0.45)** and "would take again" **(~0.43)** have strong correlations with the dependent variable (Appendix 5). These predictors are also highly multicollinear (Appendix 6). We will not drop these variables, as they are important and serve as a signal to fit an elasticnet logistic regression model to address multicollinearity and impose feature selection via standardized weight coefficients. There are **8,504 professors who did not receive "pepper," and 6,233 who did**. We assume this is not a significant enough class imbalance and can proceed accordingly.

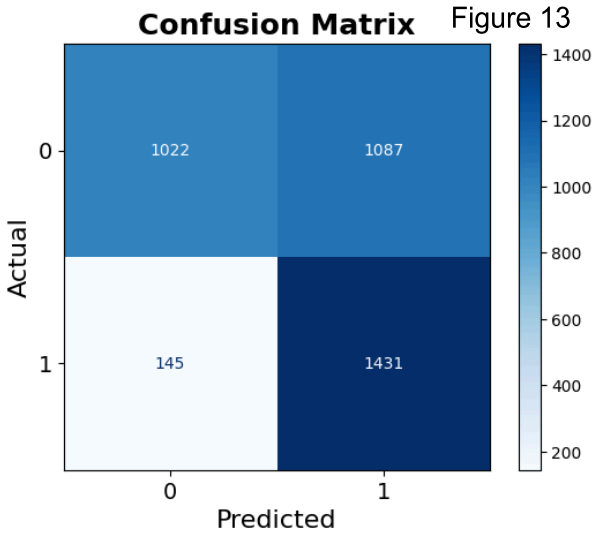
We split the data into training and test sets using an 80/20 split. We trained the model using a pipeline that imputes missing values with mean imputation, standardizes features, and fits an Elastic Net logistic regression with combined L1 and L2 regularization. We selected mean imputation because mean, median, and KNN imputation produced indistinguishable performance, making mean imputation the simplest consistent choice. We tuned the regularization strength using cross-validation on the training set, with AUROC as the optimization metric. A cross-validated AUROC of approximately 0.79 indicates strong ranking performance, meaning the model correctly ranks a randomly chosen professor who received a pepper above a randomly chosen professor who did not in about 79 percent of cases.

We then examined standardized Elastic Net coefficients and pruned the model to retain predictive signal while minimizing AUROC loss. This reduced the feature set from 26 to 6 variables and decreased training and prediction time from 35 seconds to 0.6 seconds, while maintaining an AUROC of approximately 0.78 (Figure 12). The resulting model is both interpretable and computationally efficient.



After fixing the model based on AUROC, we examined threshold-dependent metrics on the training data to select an operating point. Because false negatives are more costly in this context, we prioritized recall for the positive class when choosing the decision threshold, explicitly trading off some precision in order to identify as many true pepper recipients as possible.

Using the best-selected decision threshold of 0.27, the model achieves the intended recall-oriented behavior with minimal loss in precision on the test data (Figure 13). The confusion matrix shows that **1,431** professors who received a pepper are correctly identified, while false negatives are reduced to **145**, meaning the model successfully captures the large majority of true pepper recipients. This is accompanied by an increase in false positives (**1087**), which is acceptable given our objective of minimizing missed positive cases.



The classification metrics on test data reflect this tradeoff (Figure 14). For professors who received a pepper, recall is **0.91,** indicating that about 91% of true pepper recipients are correctly identified, while precision is 0.57, meaning that most predicted pepper cases are true positives. The resulting F1 score of **0.70** represents a strong balance between recall and precision under a recall-prioritized threshold. This threshold of **0.27**, therefore, represents an appropriate and

well-justified operating point given the higher cost of false negatives in this setting.

Figure 14

Classification Report with threshold 0.27:

	precision	recall	f1-score	support
0.0	0.88	0.48	0.62	2109
1.0	0.57	0.91	0.70	1576
accuracy			0.67	3685
macro avg	0.72	0.70	0.66	3685
weighted avg	0.74	0.67	0.66	3685

The final pruned model is shown in Figure 15. The retained predictors align with intuitive drivers of receiving a pepper. Average rating has the largest positive weight, indicating it is the strongest signal. The number of ratings and, proportion of students who would take the class again also contribute positively. Inspirational tags and greater average difficulty increase the likelihood of receiving a pepper, while a higher homework load reduces it. The limited set of six features captures the dominant signal in the data, supporting both interpretability and efficient deployment.

Figure 15

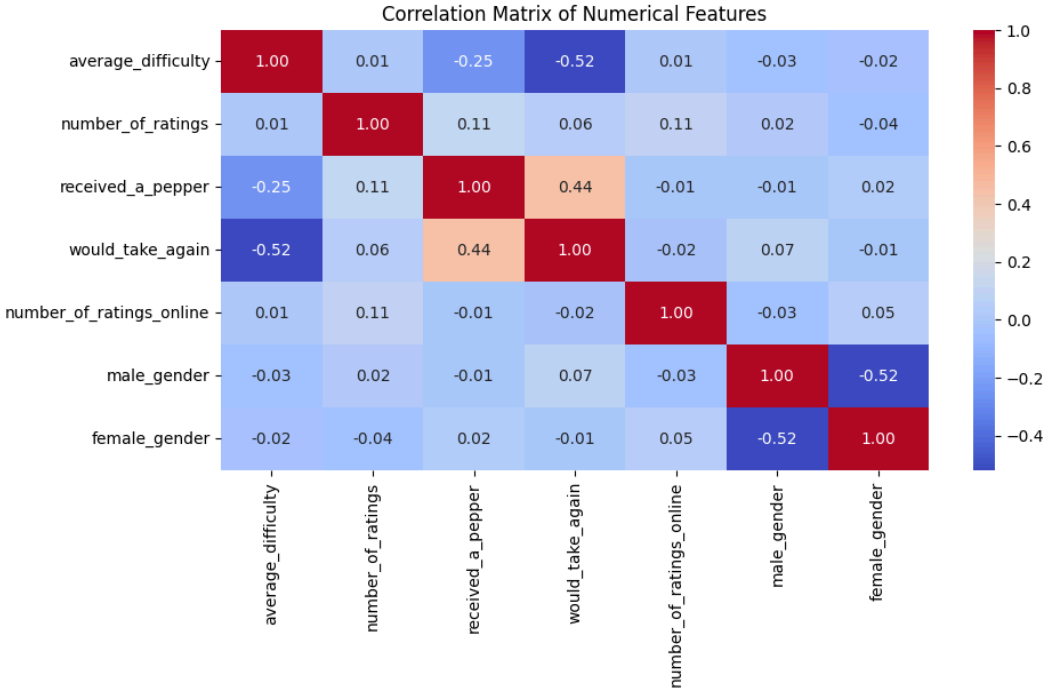
	feature	weight	abs_weight
0	average_rating	1.228122	1.228122
1	number_of_ratings	0.291174	0.291174
2	inspirational	0.166203	0.166203
3	average_difficulty	0.159798	0.159798
4	would_take_again	0.147619	0.147619
5	lots_of_homework	-0.109657	0.109657

Bibliography

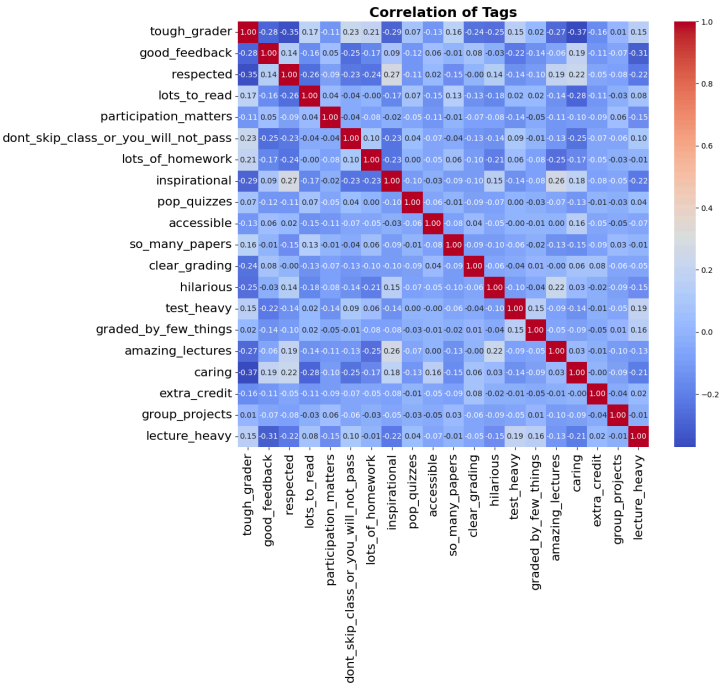
Meissel, K. and Yao, E.S. (2024) *Using Cliff's delta as a non-parametric effect size measure: An accessible web app and R tutorial*, *Practical Assessment, Research, and Evaluation*. Available at: https://openpublishing.library.umass.edu/pare/article/id/1977/ (Accessed: 03 November 2025).

Appendix

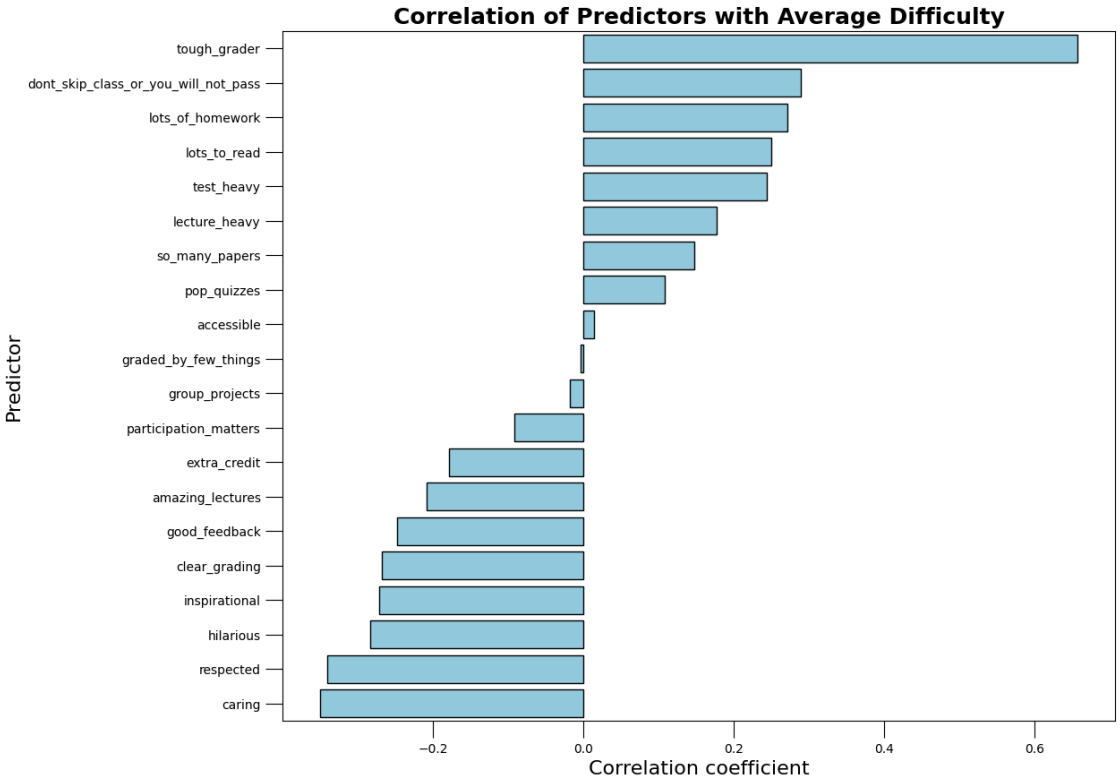
Appendix 1



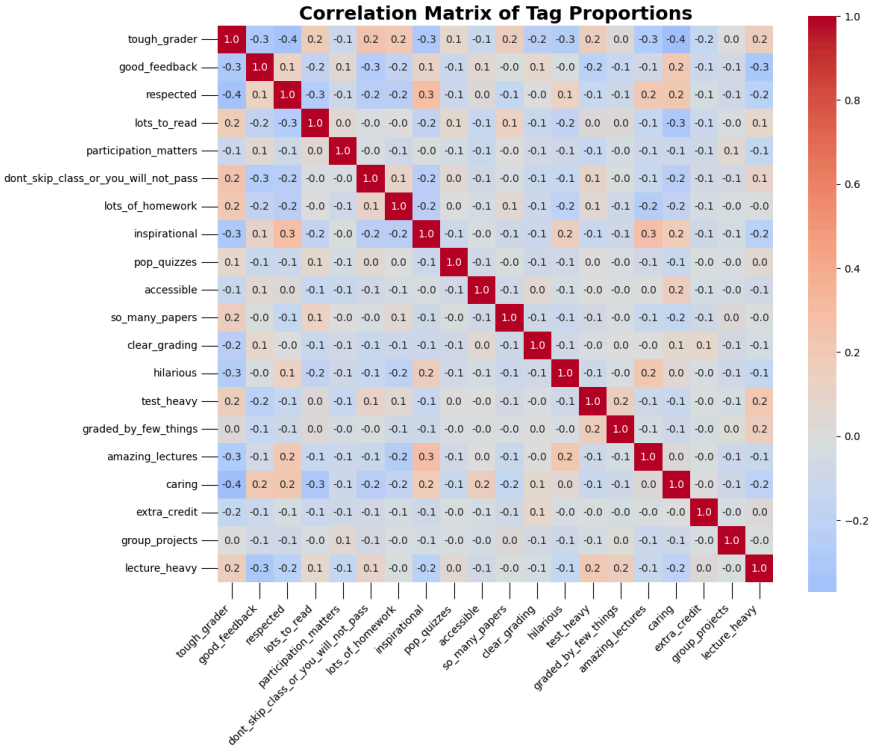
Appendix 2



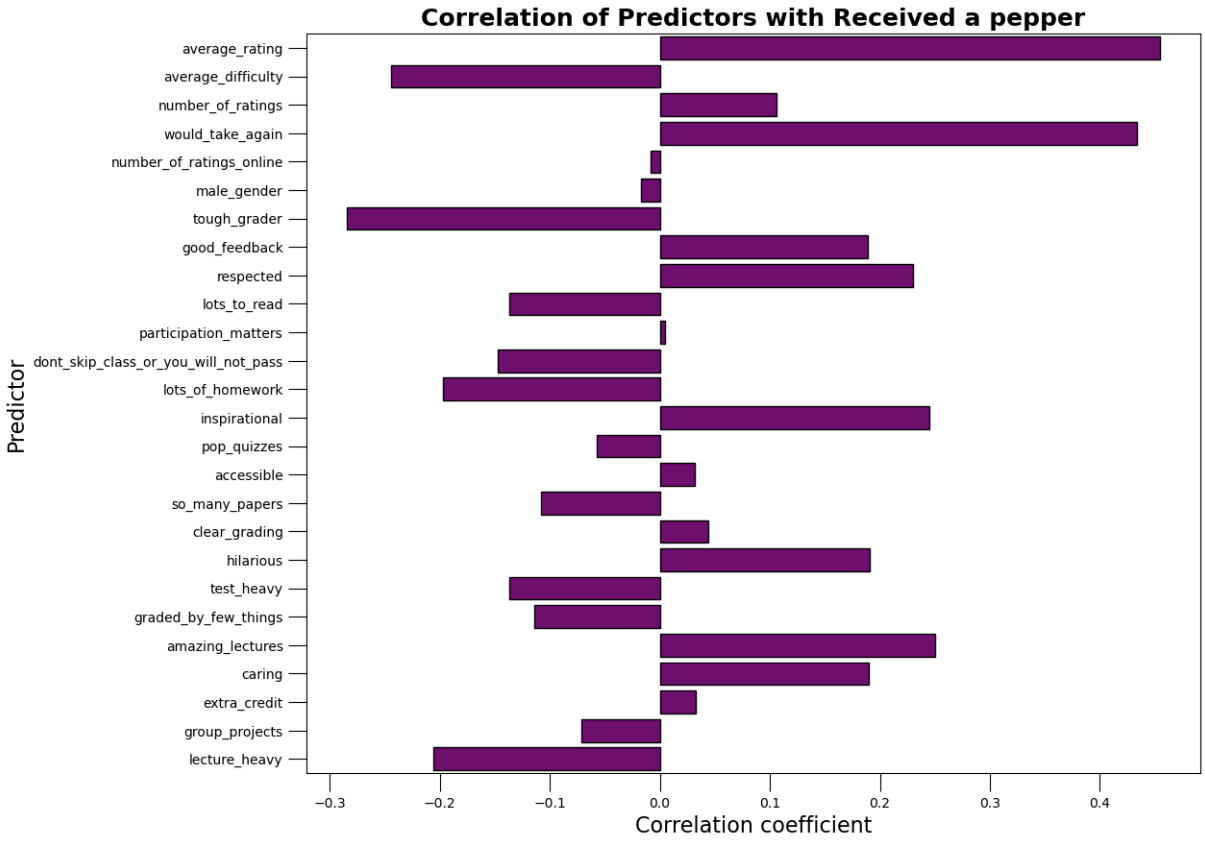
Appendix 3



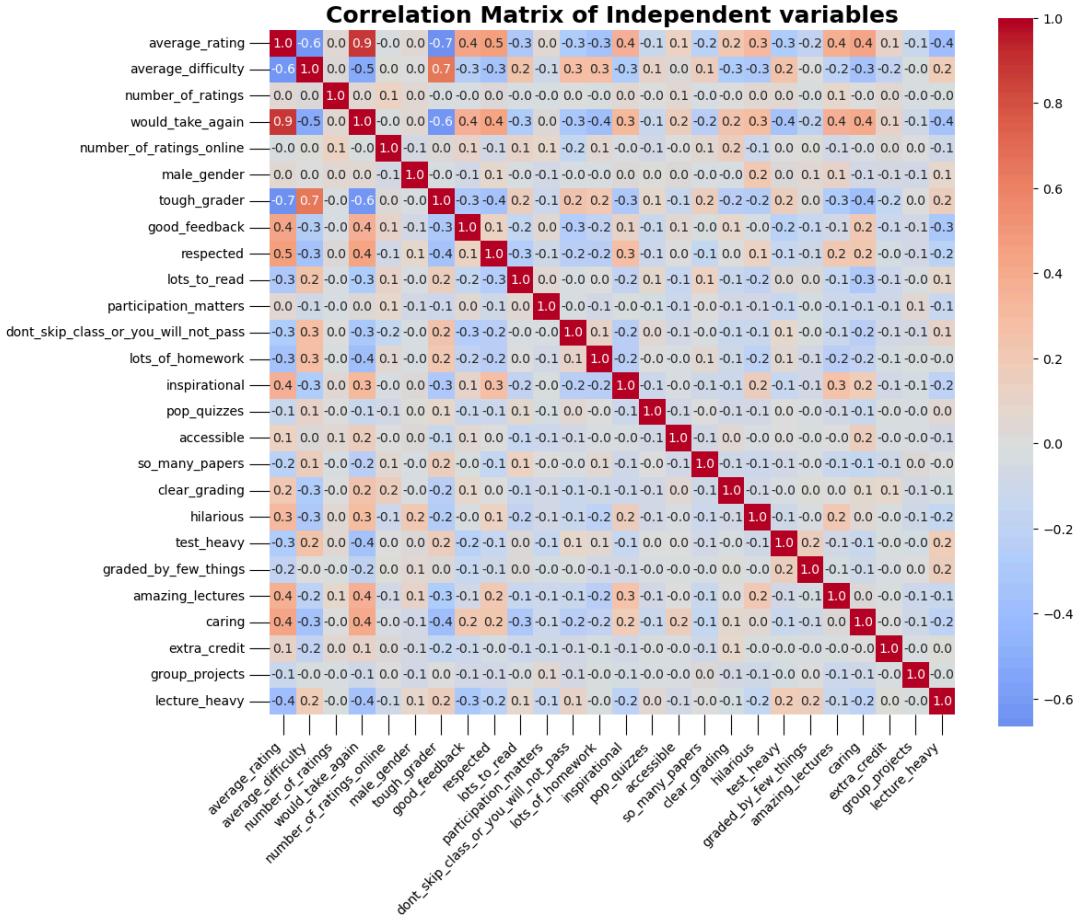
Appendix 4



Appendix 5


Correlation of Predictors with Received a pepper
average_rating
average_difficulty
number_of_ratings
would_take_again
number_of_ratings_online
male_gender
tough_grader
good_feedback
respected
lots_to_read
participation_matters
dont_skip_class_or_you_will_not_pass
lots_of_homework
inspirational
pop_quizzes
accessible
so_many_papers
clear_grading
hilarious
test_heavy
graded_by_few_things
amazing_lectures
caring
extra_credit
group_projects
lecture_heavy
Predictor
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
Correlation coefficient

Appendix 6


Correlation Matrix of Independent variables
average_rating
average_difficulty
number_of_ratings
would_take_again
number_of_ratings_online
male_gender
tough_grader
good_feedback
respected
lots_to_read
participation_matters
dont_skip_class_or_you_will_not_pass
lots_of_homework
inspirational
pop_quizzes
accessible
so_many_papers
clear_grading
hilarious
test_heavy
graded_by_few_things
amazing_lectures
caring
extra_credit
group_projects
lecture_heavy
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6