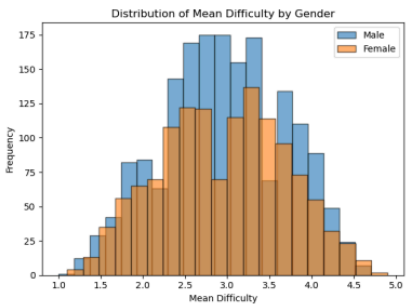
Colin Lee
Pascal Wallisch
Intro to Data Science
April 30, 2025

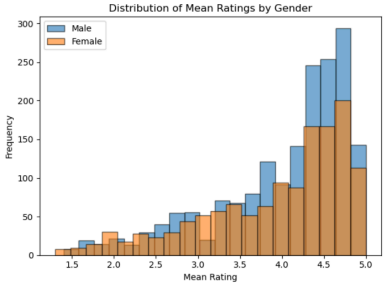
Assessing Professor Effectiveness

For each professor, we are given the mean of their RMP (Rate My Professor) ratings. As mentioned in the spec sheet, " more meaningful if it is based on more ratings". For this reason, I will only be considering professors with at least 15 ratings, since we want to remove unreliable data while keeping enough for statistical power. The figures shown in this report will not be explained unless they cannot explain themselves. How I handled missing data varied between each task, so I will explain my thinking when necessary. An alpha level of 0.005 was used throughout the entirety of this report. Figures will be rounded to 3 significant figures. Additionally, I will sometimes refer to the column names in this report. They are all intuitive. For example, meanRating refers to the average rating for a professor, etc.

1. To perform a significance test to determine whether or not there is evidence of a pro-male gender bias in this dataset, I first investigated possible confounding variables. First, I found the correlation coefficient (R) between meanRating and each of the other numeric columns. The only variables that had a moderately strong correlation with meanRating were pepper (r = 0.54) and meanDifficulty(r = -0.64). Thus, I performed significance tests to see whether the meanDifficulty medians were different for male vs female professors, and to see if the 'pepper proportion' was different for males vs females. First, I visualized the distributions to determine which test I should use .

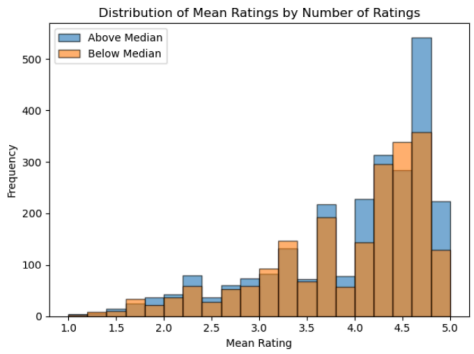


Here, it's clear to see that the meanDifficulty (left) distributions are quite symmetric. Thus, I performed a Two Sample t-Test for Independent Samples (p-value = 0.646). Next I performed a 2-sample Proportion z-Test (p-value = 0.110) to determine if there's a significant difference in the proportion of males/females that received a pepper. Thus, I concluded that there were no significant confounding variables and that I could proceed to the test of pro-male gender bias. The histogram for meanRating suggests that it would be unfit to compare means



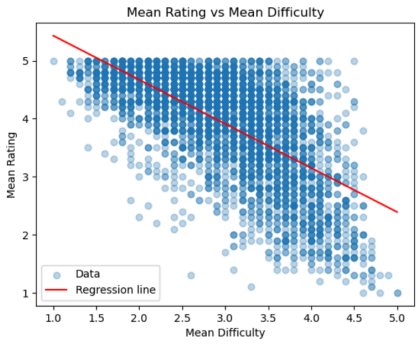
due to its skewed distribution. I then decided to compare medians using a one-sided Mann-Whitney U-test and got a p-value of 0.00704.
Thus, at the 0.005 alpha level, we cannot reject the null hypothesis that there is no pro-male gender bias.

2. To answer the question “Is there an effect of experience on the quality of teaching?”, I split the data in half. One group had all professors whose numRatings attribute was below the median of all numRatings, and the other group hadd all professors whos numRatings was greater than the median. This way, every professor in group 1 had less ratings than every professor in group 2.



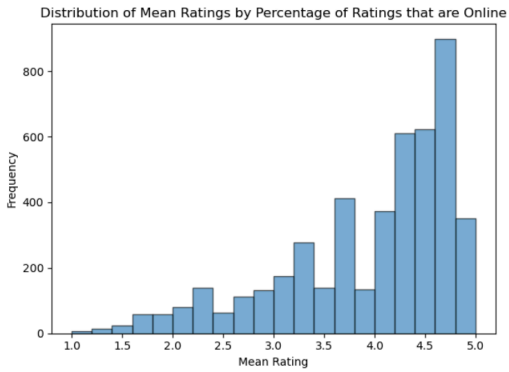
The histogram shows a very skewed distribution. Our null hypothesis is that experience (whether the professor had more or less ratings than the median) has no effect on the meanRating of the professor. Using a two-sided Mann-Whitney U-test to test the alternative hypothesis that there is an effect, I got a p-value of 0.000973. Because the p-value is less than our alpha of 0.005, we reject the null hypothesis that experience does not effect a professor’s meanRating.

3. There is a moderately strong, negative relationship between average ratings and avering difficulty. The Pearson correlation coefficient between the two variables is -0.649 and the Spearman correlation coefficient is -0.633. After fitting a linear regression model, predicting average rating from average difficulty, the coefficient for average difficulty is 0.760. For every 1 point increase in average difficulty, the predicted average rating decreases by 0.760.

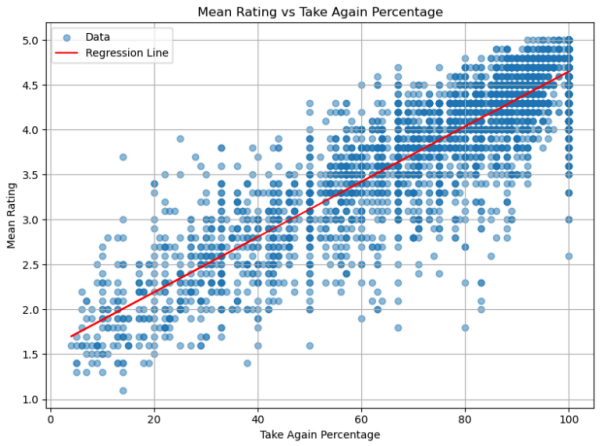


4. To test the alternative hypothesis, that professors that teach a lot of classes in the online modality receive higher or lower ratings than those who don’t, against the null hypothesis, that there is no effect, I split the data the same way as in task #2. First I engineered a feature, percentageOnline, that held the percentage of a professor’s ratings that came from online

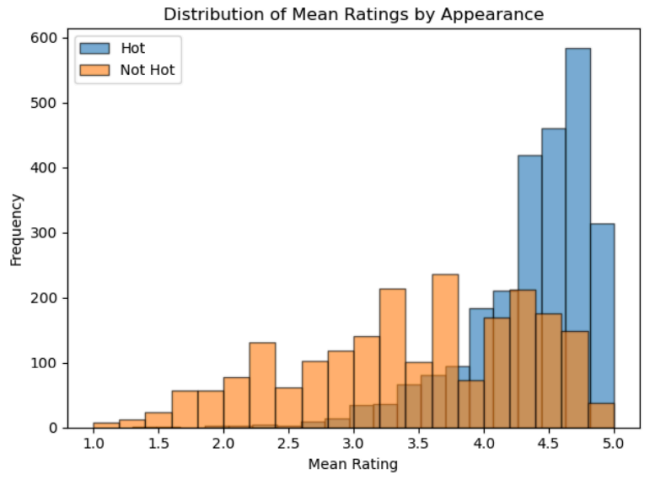
students. Then I split the professors into two groups, above and below a percentageOnline of 40%. The distribution for meanRating is clearly skewed, as seen below. After performing a Mann-Whitney U-Test between the two groups (above and below 40% online), I got a p-value of 0.0589. At the alpha level of 0.005, we cannot reject the null hypothesis that there is no difference in ratings between professors that teach a higher percentage of online students and professors who teach a lower percentage.



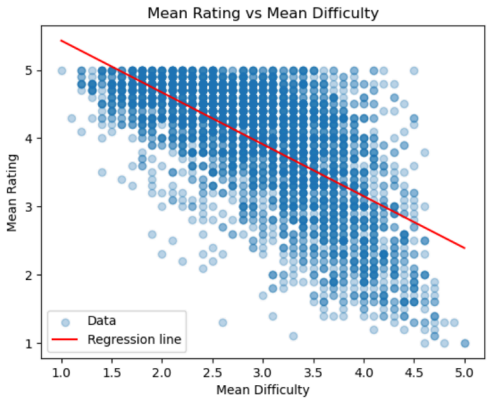
5. There is a very strong, positive relationship (Pearson r = 0.890, Spearman ρ = 0.855) between average rating and percentage of students who said they would take the class the professor teaches again. After fitting a linear regression model, predicting average rating from takeAgainPercentage, the coefficient for takeAgainPercentage is 0.0307, meaning for every 1 point increase in takeAgainPercentage, the predicted average rating decreases by 0.0307.



6. To test the question: “Do professors who are “hot” receive higher ratings than those who are not”, I first visualized the distribution of average ratings of the professors that did and didn’t receive a pepper to determine what test to use. Because the distributions are so skewed, it would make more sense to again use a Mann-Whitney U-Test. The null hypothesis was “Professors who received a pepper got rated the same as those who didn’t”, and the alternative hypothesis was “Proessors who received a pepper got rated

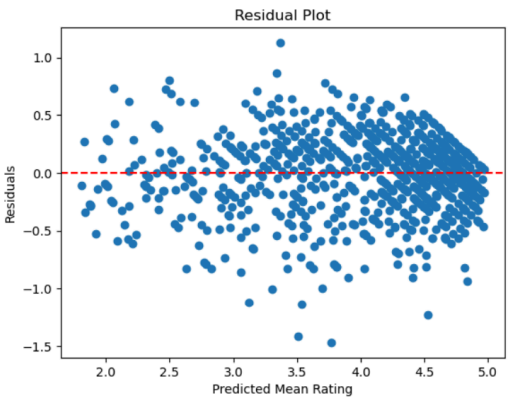


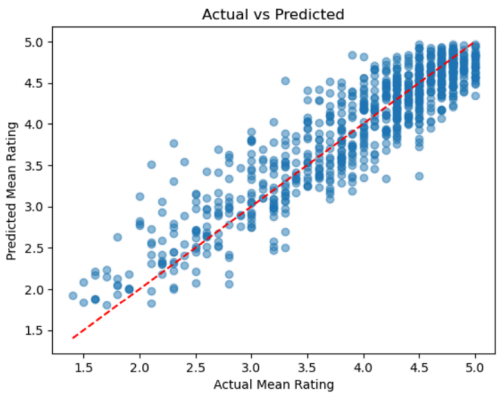
higher than those who didn't". After conducting the one-side U-Test, I got a p-value of approximately 0.0! This implies that the difference in the data between the two groups was so extreme that the test is saying it is basically impossible to record this data given the null hypothesis. Thus, because the p-value is less than the alpha level of 0.005, we reject the null hypothesis that professors who received a pepper got rated the same as those who didn't.



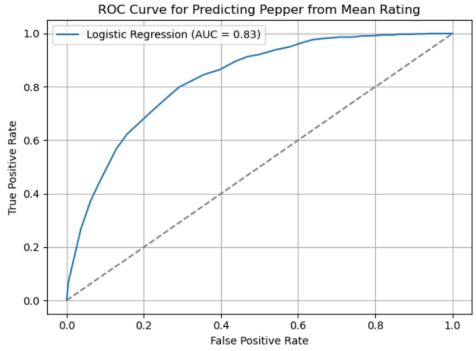
7. I built a linear regression model to predict average rating based on average difficulty. The scatter plot of these two variables shows a clear negative trend. The R^2 value of the model is 0.406, indicating that 40.6% of the variance in average rating is explained by average difficulty. The RMSE of the model is 0.654, which gives a sense of the average prediction error. While the model captures the negative relationship discussed earlier, its performance is limited by the skewness in average rating and the simplicity of using only one predictor.

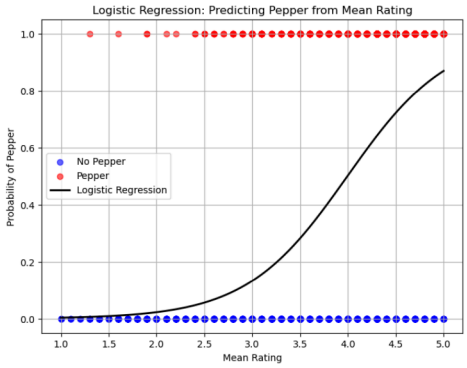
8. I then built a multiple linear regression model to predict average rating using all available numeric and binary features: meanDifficulty, numRatings, pepper, takeAgainPercentage, numRatingsOnline, male, and female. Here, I dealt with missing data by dropping all rows that had at least 1 NaN value. Only removed around 6% of the rows so there was still enough data to train and test the model. The model yielded an R^2 value of 0.836 and an RMSE of 0.334. Compared to the single-feature model in Task 7, this model performs better, suggesting that additional factors contribute meaningfully to average rating. The regression coefficients suggest that pepper and class difficulty were the two most impactful features, because their coefficients had the two greatest absolute values. Here are a residual plot (left) and a plot showing Actual versus Predicted meanRating values (right).



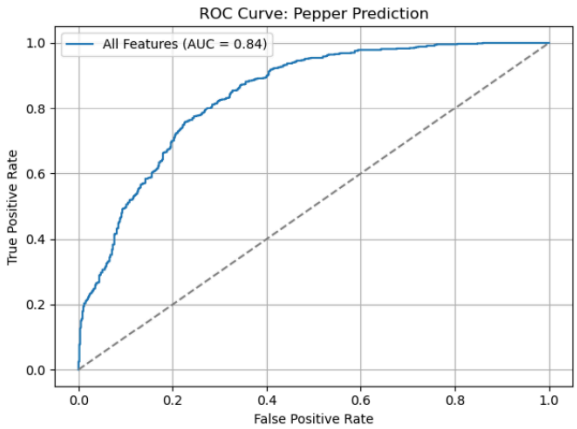


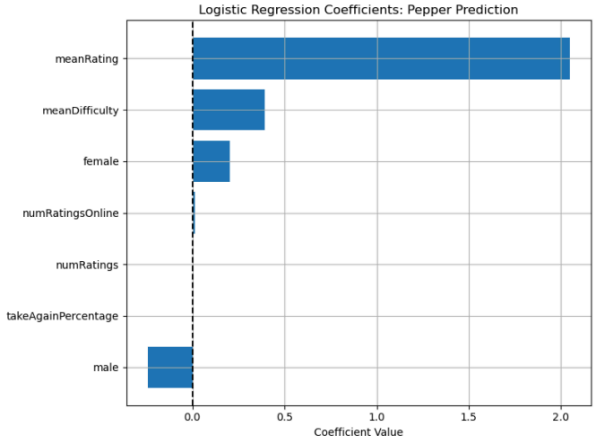
9. I trained a logistic regression model to predict whether a professor received a “pepper” using only their average rating. The ROC AUC of the model was 0.829, indicating strong predictive ability for determining whether a professor received a pepper based solely on their average rating. To address the class imbalance, the class_weight='balanced' parameter was used. Class imbalance wasn’t a huge issue here, since there were 2521 professors that received a pepper and 2158 that didn’t. However, the class_weight='balanced' still helps account for the slight imbalance of the classes by adjusting the weights inversely proportional to class sizes. As expected, average rating alone is somewhat predictive of pepper status, but the model is simplistic (as can be seen from the scatterplot to the right).





10. I expanded the logistic regression model to use all available features to predict pepper status. After fitting the model and evaluating its performance, the ROC AUC was 0.841, which was greater than the rating-only model. This suggests that other features like gender and takeAgainPercentage contribute some additional predictive value, but average rating remains the strongest predictor, with a coefficient far greater than any of the other predictors (seen below). Again, class imbalance was handled with class_weight='balanced'. This multivariate model performs slightly better than the univariate one in terms of AUROC, and offers better interpretability of what factors affect the pepper.





Extra Credit. To explore regional differences in student evaluations, I compared the distribution of average ratings between professors in California and Texas (The two states from which the most ratings came). As shown in the histogram, the distributions, although both heavily skewed, differ slightly. A Mann-Whitney U-test produced a p-value of 0.119, which is not significant at the 0.005 level. Thus, we cannot reject the null hypothesis that there is a difference in professor quality between California and Texas.