CAPSTONE PROJECT

Name: Aayan Mathur

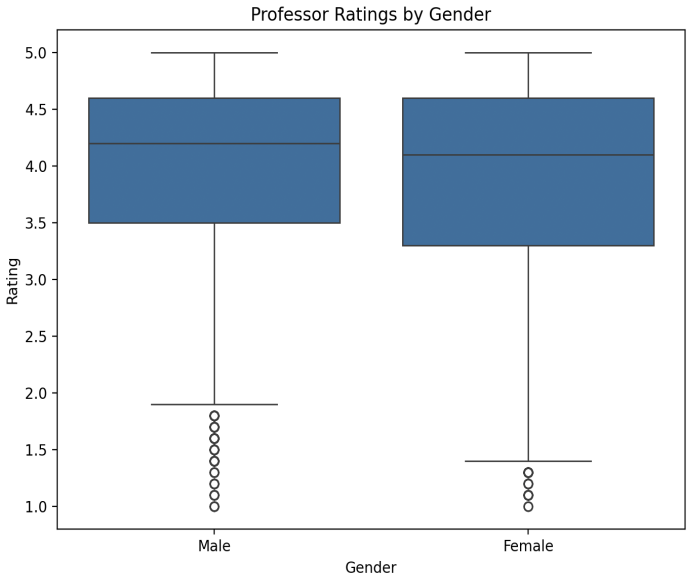
Preprocessing of data:

In the preprocessing step of the analysis, I started by removing rows that had missing data in any columns other than 'would_take_again'. The rows that had missing data in 'would_take_again' were kept because a lot of professors have NaN values in this column, and removing them would have eliminated too much valuable data. These NaN values are handled appropriately later in my analysis. After the initial cleaning, I filtered out professors with fewer than 8 ratings because ratings based on very few reviews are less reliable and more susceptible to extreme values. I chose 8 as my threshold because it provides enough individual ratings per professor to ensure reliability while also maintaining a reasonable sample size. For Question 1, I created a separate function to remove rows where professors were marked as neither male nor female (0,0) or both male and female (1,1), as these likely represent data entry errors. This cleaning was done in a separate function rather than in the main preprocessing to preserve these records for other analyses where gender is not relevant, thus maintaining a larger sample size for those questions. Lastly, I seeded the random number generator with my N-number in the beginning of my code file.

Question 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. Hint: A significance test is probably required.

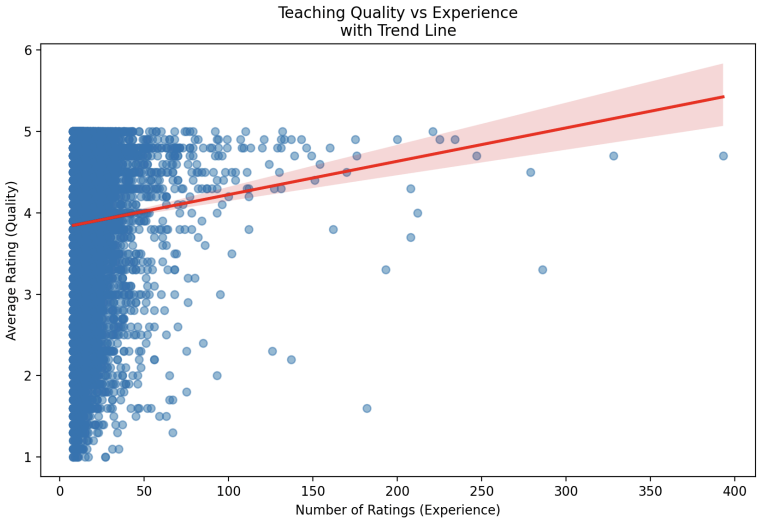
Answer 1: To investigate whether there is evidence of a pro-male gender bias in the dataset, I first cleaned the data by removing invalid gender entries (as mentioned in the preprocessing). The final dataset included 5575 male and 4533 female professors. Then I examined the distribution of ratings through a histogram and observed that the data was non-normally distributed for both genders. I also noticed a slight difference in variances. Based on this, I decided to choose the Mann-Whitney U test for the analysis. The test revealed a statistically significant difference in ratings between male professors (Mean = 3.955, Median = 4.200) and female professors (Mean = 3.879, Median = 4.100) with $p = 0.000131$ ($< \alpha = 0.005$) and U = 13,192,968. The U statistic is about 52% of its maximum possible value, indicating that while male professors rank slightly higher in ratings than female professors when all ratings are ranked together, there is a substantial amount of overlap. The difference in ratings is statistically

significant due to our large sample size but the practical significance of this difference is minimal. The box plot visualization below illustrates the similar distribution of ratings between genders, which supports the findings.



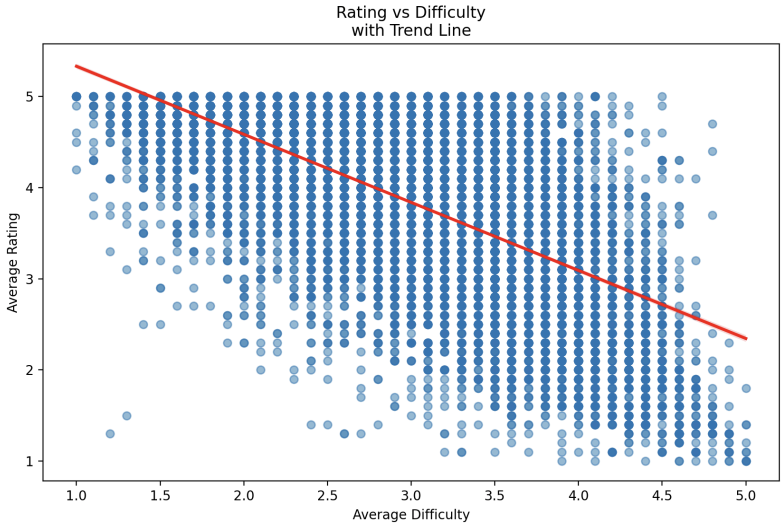
Question 2: Is there an effect of experience on the quality of teaching? You can operationalize quality with average rating and use the number of ratings as an imperfect but available proxy for experience. Again, a significance test is probably a good idea.

Answer 2: Given the fact that number of ratings and average rating are continuous variables, I decided that a correlation analysis would be the most appropriate for testing whether experience has an effect on teaching quality. I had two options: Pearson or Spearman Correlation. To decide between these two, I first created histograms and a scatter plot. The histograms showed that number of ratings was highly right-skewed (mean = 15.48, max = 393) and average ratings were left-skewed. The scatter plot showed no clear linear pattern and revealed significant outliers. Given these violations of normality and linearity, I chose Spearman correlation as it is more robust and better suited for non-normal data with outliers. The Spearman correlation revealed a statistically significant ($p < 0.005$) but very weak positive correlation ($\rho = 0.041$) with a minimal effect size ($r^2 = 0.002$). While this result was statistically significant due to our large sample size, the tiny effect size and weak correlation coefficient suggest that teaching experience, as measured by number of ratings, has practically no relationship with teaching quality. The scatter plot below visualizes this weak relationship, with the trend line showing a slight positive slope.



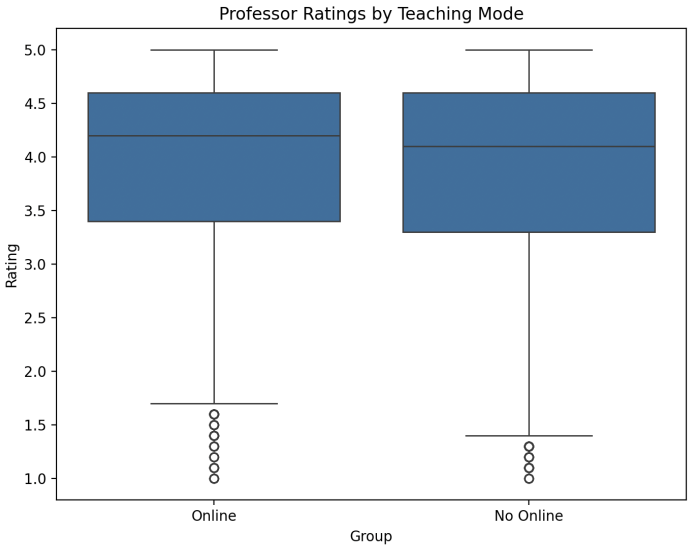
Question 3: What is the relationship between average rating and average difficulty?

Answer 3: To analyze the relationship between average rating and average difficulty, I first examined the distributions of the two variables through histograms and a scatter plot (similar to my approach in Answer 2). While the average difficulty showed a roughly normal distribution, average ratings were left skewed and there were notable outliers. Given these characteristics, I again chose Spearman correlation since it is robust to non-normality and outliers. The analysis revealed a statistically significant ($p < 0.005$) strong negative correlation ($\rho = -0.623$) with a moderate effect size ($r^2 = 0.388$). This indicates that professors with higher difficulty ratings tend to receive lower overall ratings. The scatter plot with the trend line below visualizes this negative relationship, showing a clear downward trend but with considerable variation around the trend line, suggesting that while difficulty strongly influences ratings, other factors also play important roles in determining a professor's rating.



Question 4: Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't? Hint: A significance test might be a good idea, but you need to think of a creative but suitable way to split the data.

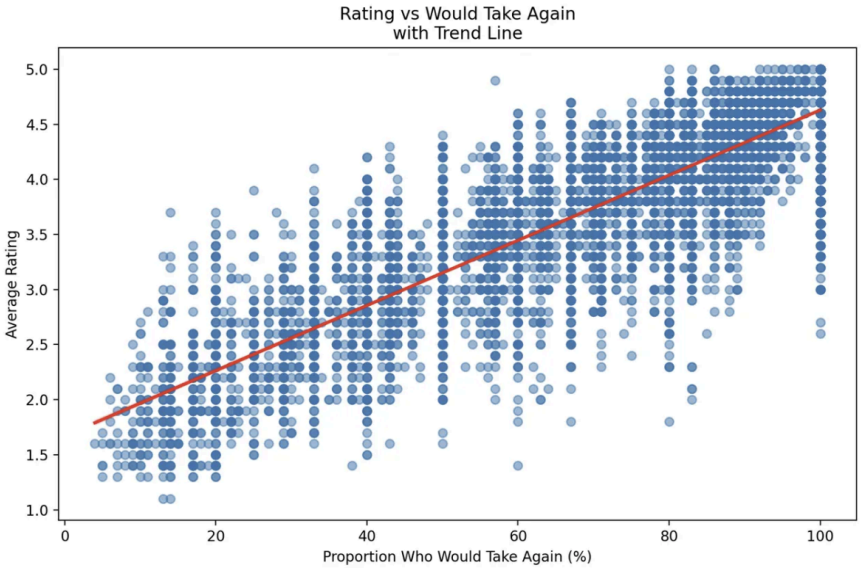
Answer 4: To examine whether professors who teach online receive different ratings than those who don't, I first analyzed the distribution of online ratings among professors. The data showed that while 3,105 professors had at least one online rating, most of these had very few (mean = 2.95, median = 2 online ratings), making it impractical to set a higher threshold for defining online teaching. Therefore, I split the data into professors with any online ratings (n=3,105) and those with none (n=10,890). After examining the rating distributions for both groups, which showed clear non-normality with left skew, I chose the Mann-Whitney U test over parametric alternatives. The test revealed no statistically significant difference in ratings ($p = 0.404$) between professors who teach online (Mean = 3.883, Median = 4.200) and those who don't (Mean = 3.875, Median = 4.100). The U statistic of 17,071,984, being very close to the midpoint, and the tiny effect size ($r = 0.007$) further support that there is essentially no difference in ratings between the two groups. The box plot below illustrates the similar rating distributions between online and non-online teaching professors.



Question 5: What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again.

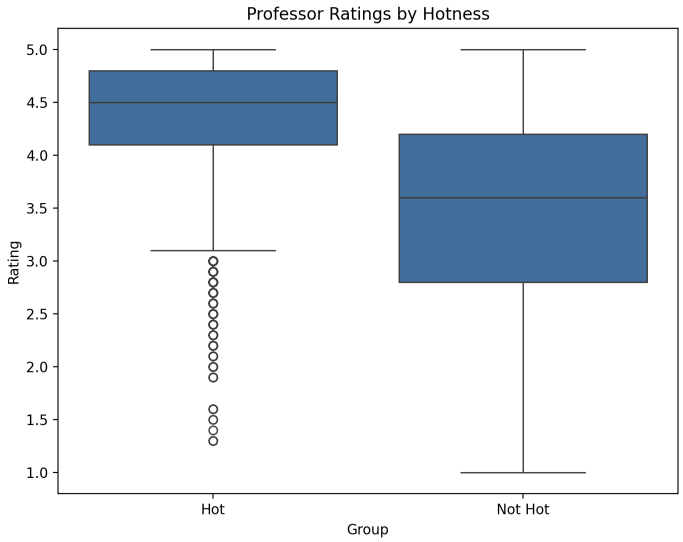
Answer 5: I first needed to handle the substantial missing data in the 'would take again' column. To do this, I created a new data frame (called df_clean), which had all those rows removed that had NaN values in the 'would_take_again' column. This new data frame retained 73.9% of the professors (10,344 out of 13,995) that had valid data for analysis. Then I examined the distributions for average rating and 'would_take_again' through histograms and observed that both variables were non-normally distributed, with average ratings showing left skew and 'would

take again' showing strong right skew. The scatter plot revealed a clear positive relationship with heteroscedasticity (increasing spread at higher values). Given these characteristics and the proportional nature of the 'would take again' variable, I chose Spearman correlation. The analysis revealed a very strong positive correlation ($\rho = 0.843$, $p < 0.005$) with a large effect size ($r^2 = 0.711$), indicating that professors with higher ratings are substantially more likely to have students willing to take their class again. The scatter plot with trend line below visualizes this strong positive relationship. (though the spread of points suggests that other factors also influence these variables)



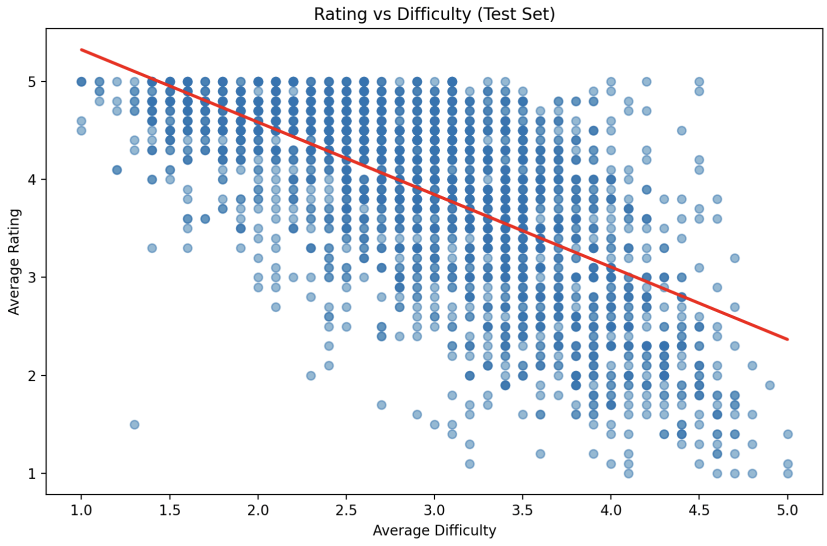
Question 6: Do professors who are 'hot' receive higher ratings than those who are not? Again, a significance test is indicated.

Answer 6: To examine whether professors who receive a "pepper" (hot) rating get higher ratings than those who don't, I first analyzed the rating distributions for both groups. The histograms revealed non-normal, left-skewed distributions for both hot (n=6,539) and not hot (n=7,456) professors. Given this non-normality, I again chose the Mann-Whitney U test for the analysis. The test revealed a statistically significant difference in ratings between hot professors (Mean = 4.357, Median = 4.500) and not hot professors (Mean = 3.456, Median = 3.600) with $p < 0.005$. The U statistic of 38,731,612, representing about 79% of its maximum possible value, indicates that hot professors tend to rank notably higher in ratings. This is supported by a large effect size ($r = 0.509$), suggesting that professors with a "pepper" rating receive substantially higher ratings overall. The box plot below illustrates this considerable difference in rating distributions between the two groups.



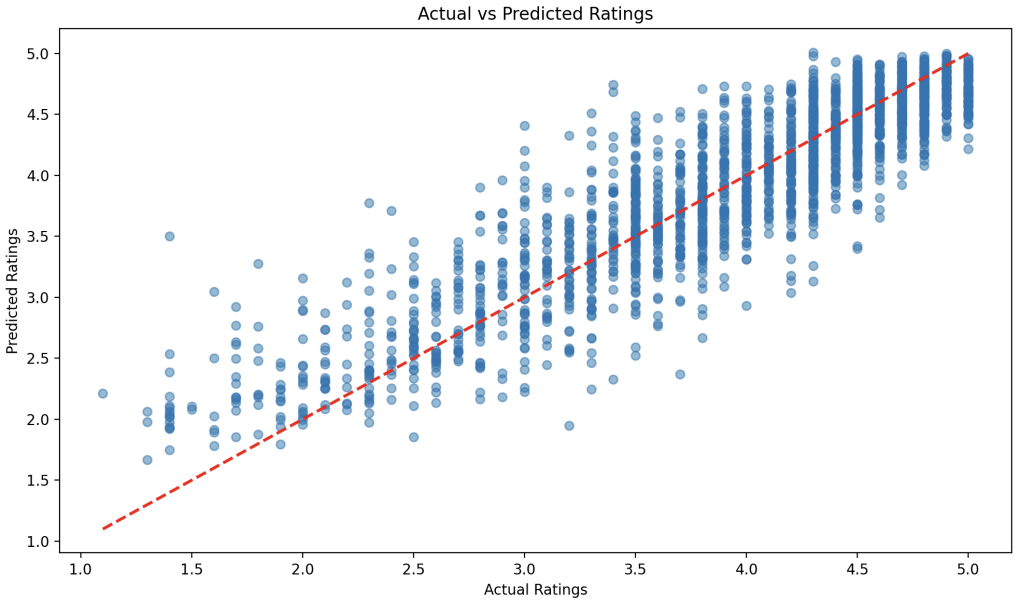
Question 7: Build a regression model predicting average rating from difficulty (only). Make sure to include the R^2 and RMSE of this model.

Answer 7: I built a simple linear regression model to predict professors' average ratings based solely on their difficulty ratings. After splitting the data into training (80%) and test (20%) sets using my N-number as the random seed, I fit the model on the training data and evaluated it on the test set. The beta coefficient of -0.739 indicates that for each one-point increase in difficulty, the rating is predicted to decrease by 0.739 points. When evaluated on the unseen test data, the model achieved an R^2 of 0.431, indicating that difficulty alone explains about 43% of the variance in ratings. The RMSE of 0.692 represents the typical prediction error in rating points, meaning our predictions deviate from actual ratings by about 0.7 points on average. The scatter plot below visualizes the relationship in the test set, with the red regression line showing the negative linear trend, while the spread of points around the line illustrates the model's imperfect predictions.



Question 8: Build a regression model predicting average rating from all available factors. Make sure to include the R2 and RMSE of this model. Comment on how this model compares to the "difficulty only" model and on individual betas. Hint: Make sure to address collinearity concerns.

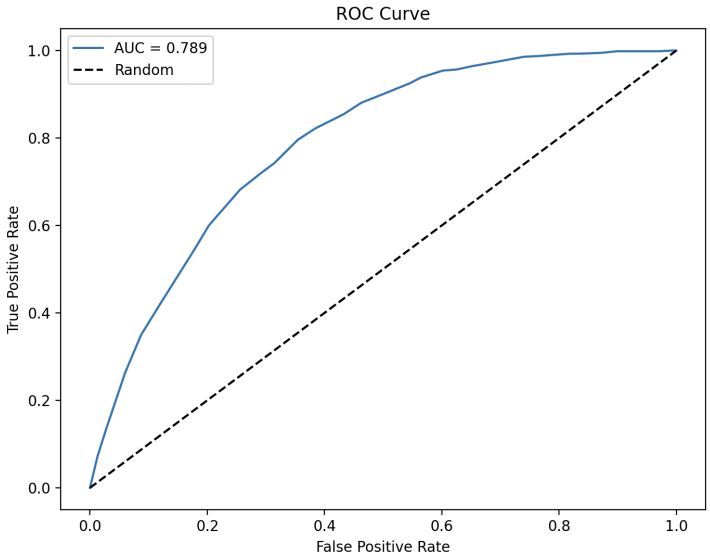
Answer 8: To build a regression model predicting average rating from all available factors, I again used the 'df_clean' data frame that I had previously created. First, I examined correlations between predictors using a correlation matrix, finding moderate correlation between 'would_take_again' and 'average_difficulty', while other correlations were relatively weak. Since the correlations weren't severe enough to cause major concerns, I proceeded with a standard multiple regression with all the predictors. Using an 80-20 train-test split (N-number as random seed), I evaluated the model's performance on unseen test data. The model achieved an R^2 of 0.810 and RMSE of 0.365 on the test set, substantially outperforming our difficulty-only model (R^2 = 0.362, RMSE = 0.669). The coefficients show that difficulty (β = -0.224), 'would_take_again' (β = 0.024), and 'has_pepper' (β = 0.211) are the strongest predictors, while 'number_of_ratings' and 'online_ratings' had minimal impact ($\beta \approx 0$). The actual vs predicted plot shows a strong linear relationship between predicted and actual ratings, visually confirming the model's good performance.



Question 9: Build a classification model that predicts whether a professor receives a "pepper" from average rating only. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

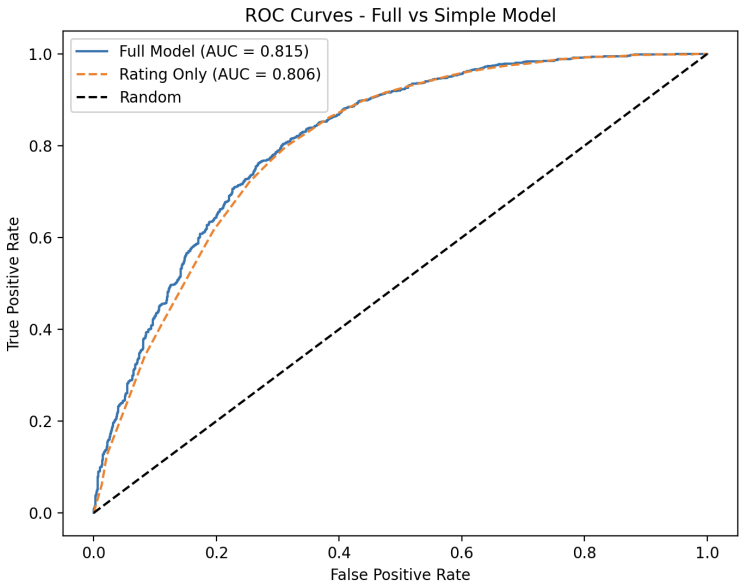
Answer 9: To predict whether a professor receives a "pepper" using only their average rating, I first examined the class distribution and found it was fairly balanced with 53.3% not having a

pepper and 46.7% having one, indicating no need for special handling of class imbalance. Then I split the data into 80% training and 20% test sets with my N-number as random seed and built a logistic regression model. When evaluated on the unseen test data, the model achieved an AUC-ROC score of 0.789, indicating good discriminative ability, as shown by the ROC curve's significant deviation above the random-guess line. The model's coefficient (1.625) and intercept (-6.608) suggest that higher ratings substantially increase the likelihood that a professor has a pepper. The ROC curve below visualizes the model's ability to balance true positive and false positive predictions across different classification thresholds.



Question 10: Build a classification model that predicts whether a professor receives a “pepper” from all available factors. Comment on how this model compares to the “average rating only” model. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

Answer 10: Building upon the simple model from Question 9, I developed a more comprehensive logistic regression model using all available predictors to classify professors as "hot" or "not" (using the 'df_clean' data frame once again). I split the data into training (80%) and test (20%) sets with my N-number as random seed. When evaluated on the test set, the full model achieved an AUC-ROC of 0.815, showing minimal improvement over the simple model's AUC-ROC of 0.806. The ROC curves for both models almost completely overlap, suggesting that additional predictors provide little benefit in classifying professors' "pepper" status. Looking at the coefficients, average rating remains the strongest predictor ($\beta = 1.693$), with other variables like difficulty ($\beta = 0.286$) and demographic factors having relatively small effects. This indicates that average rating alone captures most of the predictive power for determining a professor's "pepper" status.



Extra Credit: I investigated whether certain academic fields are more likely to have professors receive a "pepper" rating by examining the distribution of "hot" professors across different majors (including only majors with at least 50 professors for reliable measurements). The analysis revealed an interesting pattern: humanities and arts fields consistently showed higher percentages of professors with peppers (Speech: 63%, Writing: 58%, Spanish: 57%) compared to STEM fields which had the lowest percentages (Physics: 35%, Statistics: 32%, Finance: 32%). A chi-square test confirmed that this distribution was highly non-random ($\chi^2 = 224.342$, $p < 0.005$), indicating significant differences in how students rate professor attractiveness across different academic disciplines.