IDS Capstone Project Report

Mustafa Poonawala (msp9471) and Aysha Allahverdiyeva (aa7983)

Preprocessing: For Question 1 to 3

The dataset was preprocessed to ensure clean and meaningful analysis by filtering out professors with ambiguous or missing gender information, retaining only those with unambiguous gender to enable valid and direct gender comparisons. A threshold of at least five ratings per professor was applied to minimize the impact of extreme averages from professors with very few reviews, which could otherwise skew the results. Bayesian adjustment was employed to stabilize averages by incorporating a single global mean, ensuring that ratings were adjusted uniformly across genders and giving more weight to professors with higher review counts. A damping factor of 10 was chosen for Bayesian adjustment based on the distribution of the number of ratings. The median number of ratings per professor was 8, and the mean was 11, making 10 a balanced and representative choice. This damping factor moderated the influence of the global mean, ensuring that professors with fewer reviews had their ratings adjusted appropriately, while those with a higher number of reviews retained a stronger influence from their individual data. A higher damping factor would overly dilute the individual averages, while a lower damping factor might not sufficiently stabilize the ratings for professors with fewer reviews. Thus, 10 provided a robust middle ground to achieve stability without losing meaningful variation. Out of the initial 89,893 records, 52,089 professors with unambiguous gender were retained, and 18,422 met the threshold of five or more ratings. This preprocessing method ensured statistical robustness, reduced bias caused by outliers or insufficient data, and provided a shared baseline for unbiased gender comparisons. It also struck a balance between data integrity and sample size, resulting in a reliable dataset for meaningful analysis.

Question 1: Is there evidence of a pro-male gender bias in professor ratings?

Do: What did you do?

The preprocessed dataset was analyzed to determine whether male professors received higher ratings than their female counterparts. Adjusted average ratings were calculated for all professors using a single global mean to ensure a consistent baseline across genders. A Mann-Whitney U test was conducted to compare the distributions of adjusted ratings between genders, given the non-normal distribution of ratings confirmed by Kolmogorov-Smirnov tests. The analysis also considered professors with and without a "pepper" status. Additionally, a subset of professors identified as "Tough Grader" and associated with "Bad Feedback" was analyzed to investigate whether this pro-male bias persisted in specific teaching profiles. Visualizations, including histograms and boxplots, were created to illustrate the results.

Why: Why did you do this?

The Mann-Whitney U test was used because it is a non-parametric test suitable for comparing distributions that deviate from normality. Stratifying by "pepper" status provided insights into whether perceived attractiveness influenced gender bias. The focus on the "Tough Grader" and "Bad Feedback" subset allowed for exploration of bias within distinct teaching profiles that might evoke different student evaluations. Visualizations helped to clearly demonstrate distributional differences and provide an intuitive understanding of the results.

Find: What did you find?

The analysis revealed evidence of a pro-male bias in professor ratings. Male professors consistently received slightly higher adjusted average ratings than their female counterparts. Overall, male professors had a mean adjusted rating of 3.91 compared to 3.88 for female professors, with median ratings of 4.02 and 3.99, respectively. **The Mann-Whitney U test confirmed the significance of this difference, yielding a U statistic of 43,723,194.0 and a p-value of 3.09e-06.**

When examining professors without a "pepper," male professors had a mean adjusted rating of 3.74, compared to 3.69 for female professors, with a U statistic of 14,723,936.0 and a p-value of 2.43e-07. For professors with a "pepper," male professors had a mean adjusted rating of 4.15, while female professors had a mean of 4.12. The Mann-Whitney U test confirmed this difference with a U statistic of 7,952,272.0 and a p-value of 0.00012.

The subset analysis of professors categorized as "Tough Grader" and providing "Bad Feedback" further supported these findings. Male professors in this subset had a mean adjusted rating of 2.92, compared to 2.74 for female professors, with medians of 2.92 and 2.73, respectively. The Mann-Whitney U test for this subset yielded a U statistic of 1,135,350.0 and a p-value of 7.54e-24. Visualizations (Figures 1 and 2) illustrated these trends, showing a slight upward shift in the distribution of male professors' ratings and consistently higher medians compared to female professors.

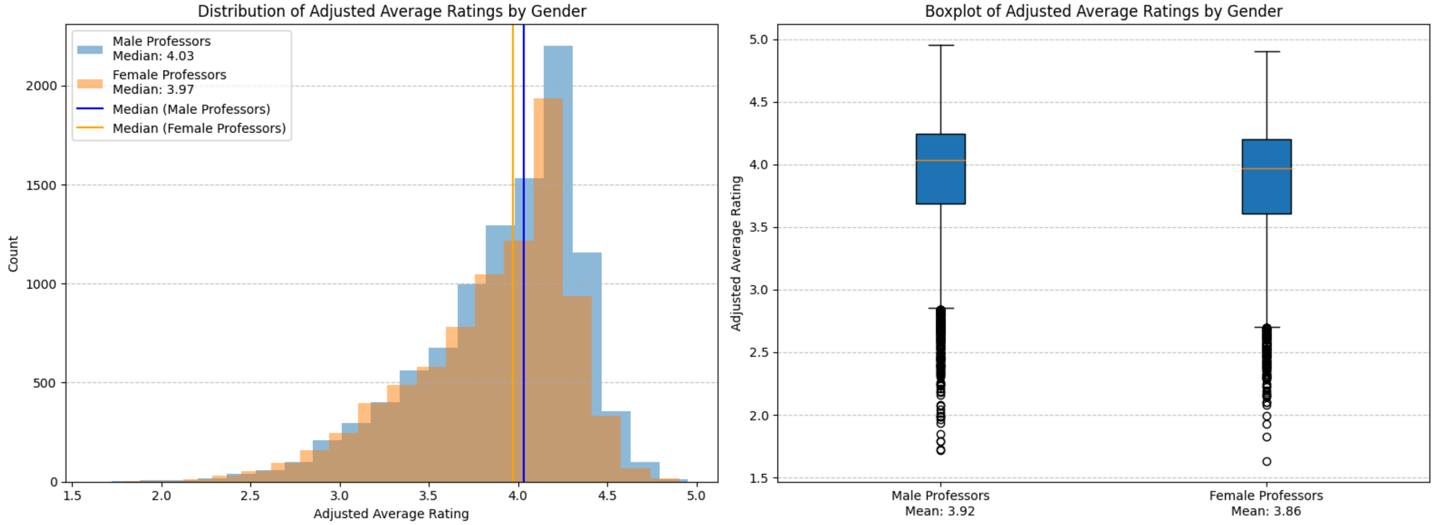

Figure 1: Visualization of of the Adjusted Average Ratings by Gender

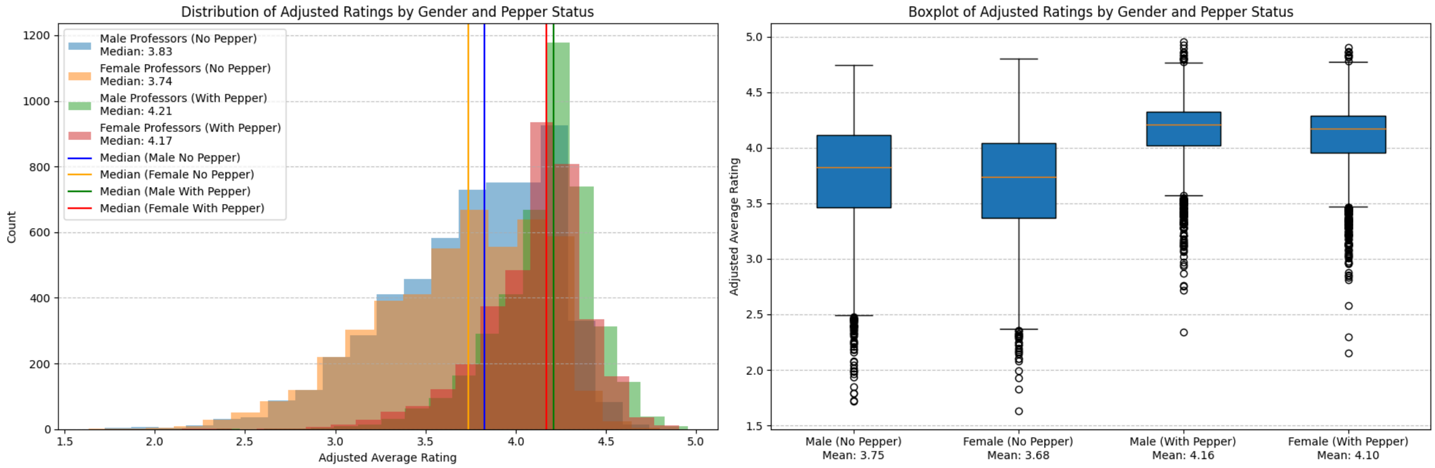

Figure 2: Visualization of of the Adjusted Average Ratings by Gender and Pepper Status

Answer: What do you conclude?

The findings provide evidence of a pro-male gender bias in professor ratings. Male professors consistently received higher ratings compared to their female counterparts, both overall and within specific subsets such as professors without a pepper and those identified as "Tough Grader" with "Bad Feedback." The statistical significance of the Mann-Whitney U tests (p-values < 0.005) confirms the reliability of these findings. However, the magnitude of the bias is small, as evidenced by the negligible Cliff's Delta of 0.0386. Potential confounding factors such as teaching experience, course difficulty, or class size were not controlled in this analysis and could influence the results. Despite these limitations, the data supports the conclusion of a slight gender bias favoring male professors.

Question 2: Is there a gender difference in the spread (variance/dispersion) of the ratings distribution?

Do: What did you do?

To assess whether there is a gender difference in the variance of professor ratings, the adjusted average ratings for male and female professors were compared. Variance and standard deviation were calculated for each gender as measures of dispersion. Levene's test for equality of variances was conducted to statistically evaluate whether the variances differed significantly between genders. Histograms and boxplots were created to visualize and compare the spread of ratings for male and female professors.

Why: Why did you do this?

The preprocessing steps ensured the dataset was suitable for this analysis by excluding ambiguous gender records and applying a threshold of at least five ratings to reduce the impact of outliers. Bayesian adjustment stabilized the averages, allowing for consistent comparisons of rating dispersion across genders. This was especially important for analyzing the spread because unadjusted ratings for professors with very few reviews might artificially inflate or deflate variances, leading to misleading conclusions about the variability of ratings. By incorporating a single global mean and weighting the adjusted ratings by the number of reviews, Bayesian adjustment reduced noise and provided a more reliable basis for comparing the variability across genders. Variance and standard deviation are standard measures of spread and provide an initial assessment of differences in rating variability. Levene's test was chosen for its robustness to non-normal distributions, as identified in Question 1. Visualizations complemented the statistical analysis, offering a clear depiction of the distribution and variability of ratings by gender.

Find: What did you find?

The analysis showed that male professors had a variance of 0.1934 and a standard deviation of 0.4398, while female professors had a variance of 0.2026 and a standard deviation of 0.4501. **Levene's test for equality of variances produced a test statistic of 6.991 and a p-value of 0.0082, which exceeded the significance threshold of 0.005.** As a result, **the test failed to detect a statistically significant difference in variances between genders**. Visualizations, including histograms and boxplots (Figure 1), illustrated similar spreads of adjusted ratings for both genders, reinforcing the conclusion that the variability in ratings was comparable across genders.

Answer: What do you conclude?

The results indicate no significant difference in the spread (variance or standard deviation) of ratings between male and female professors. Both statistical tests and visualizations demonstrate that the ratings for both genders exhibit comparable levels of dispersion. These findings suggest that the variability in how professors are rated is consistent across genders and not subject to substantial differences based on gender.

Question 3: What is the likely size of the gender bias effects in average ratings and spread of ratings?

Do: What did you do?

To evaluate the magnitude of gender bias in professor ratings, the adjusted average ratings for male and female professors were analyzed using bootstrap confidence intervals to quantify the size of both effects: bias in average ratings and bias in rating spread. For average ratings, the mean difference between genders was calculated, and Cliff's delta was computed to quantify the effect size and interpret its magnitude. For the spread of ratings, variance and variance ratios were compared, and bootstrap confidence intervals were computed for the variance difference and variance ratio. Visualizations were created to illustrate the confidence intervals for means and variances by gender.

Why: Why did you do this?

Bootstrap confidence intervals were used for their robustness to deviations from parametric assumptions, ensuring accurate estimates of variability and bias. Cliff's delta was calculated to measure effect size, as it is interpretable for non-parametric and ordinal data, offering a direct comparison of distributions between genders. The focus on variance and variance ratios provided a comprehensive understanding of potential differences in rating spread. Visualizations were included to intuitively depict the results and confidence levels for both effects, facilitating a clearer interpretation of the findings.

Find: What did you find?

For average ratings, male professors had a mean adjusted rating of 3.9089 with a 95% bootstrap confidence interval of (3.9001, 3.9176), while female professors had a mean adjusted rating of 3.8783 with a 95% confidence interval of (3.8689, 3.8880). The mean difference was 0.0306, indicating a slight bias favoring male professors. Cliff's delta was calculated as 0.0386, which is interpreted as a negligible effect size.

For the spread of ratings, male professors had a variance of 0.1934 with a 95% bootstrap confidence interval of (0.1868, 0.2000), while female professors had a variance of 0.2025 with a 95% confidence interval of (0.1955, 0.2096). The variance difference (male - female) was -0.0092, with a 95% bootstrap confidence interval of (-0.0188, 0.0007), indicating that male ratings had slightly less variability. The variance ratio (male/female) was 0.9548, with a 95% confidence interval of (0.9087, 1.0033). These results suggest a minimal difference in spread, with overlapping confidence intervals for variances indicating that any observed variability difference is statistically insignificant.

Answer: What do you conclude?

The results indicate that the size of the gender bias in average ratings is small and statistically significant but negligible in practical terms, as evidenced by a mean difference of 0.0306 and Cliff's delta of 0.0386. The analysis of rating spread shows no meaningful difference in variability between genders, as demonstrated by overlapping confidence intervals and a variance ratio close to 1. These findings suggest that while there is evidence of gender bias in average ratings, its practical impact is minimal, and the variability in ratings is comparable across genders.

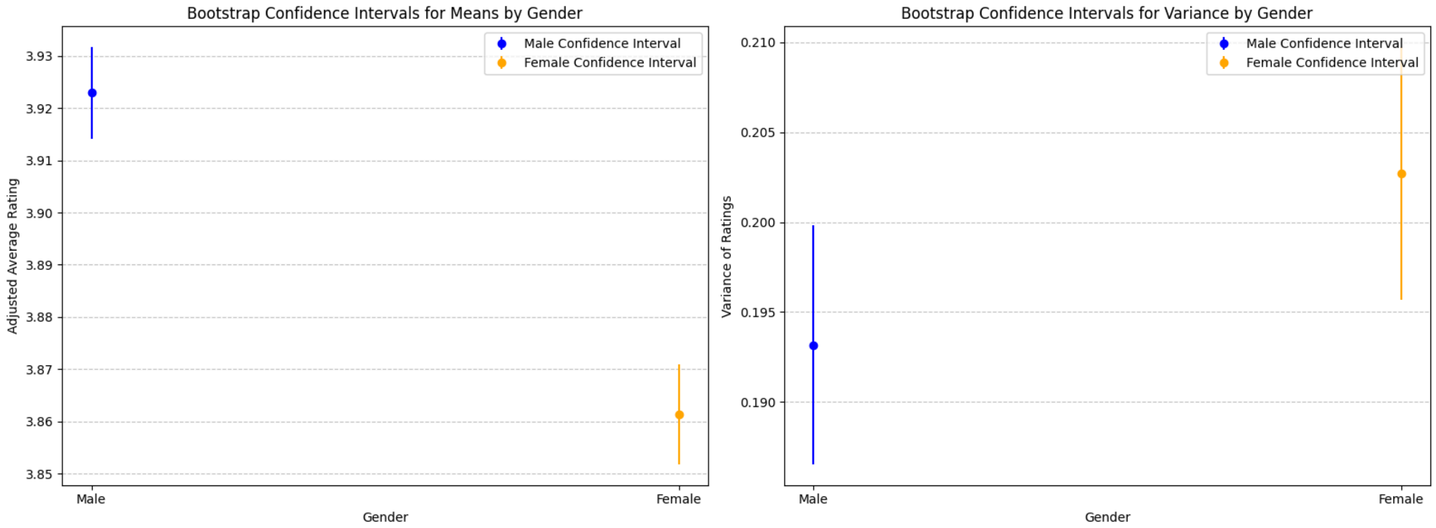

Figure 3 : Plots for the Confidence Intervals for Mean (Left Plot) and Variance (Right Plot) by Gender

The visualizations in figure 3 of bootstrap confidence intervals further clarified the findings. The first plot showed that the confidence intervals for male and female means are distinct but very close, emphasizing the negligible practical difference. The second plot displayed overlapping confidence intervals for variances, reinforcing the conclusion that differences in rating spread are minimal and likely not meaningful.

Question 4: Is there a gender difference in the tags awarded by students?

Do: What did you do?

The analysis explored potential gender differences in the 20 tags awarded to professors by students. To control for variations in review volume, the tag counts were normalized by dividing them by the number of ratings for each professor. Professors with unambiguous gender and at least five ratings were included. The Mann-Whitney U test was applied to compare the normalized tag frequencies between male and female professors for each tag. Tags with p-values below the significance threshold of 0.005 were deemed statistically significant. The three most gendered (lowest p-values) and least gendered (highest p-values) tags were highlighted for detailed discussion.

Why: Why did you do this?

Normalization ensured that tag frequencies were proportional to the number of ratings, avoiding biases due to differences in review volume. The Mann-Whitney U test, as a non-parametric method, enabled robust comparisons of tag distributions between genders without assuming normality. Analyzing all 20 tags allowed for a comprehensive investigation of gender-based differences, while focusing on the most and least gendered tags provided insights into traits where biases were most and least pronounced.

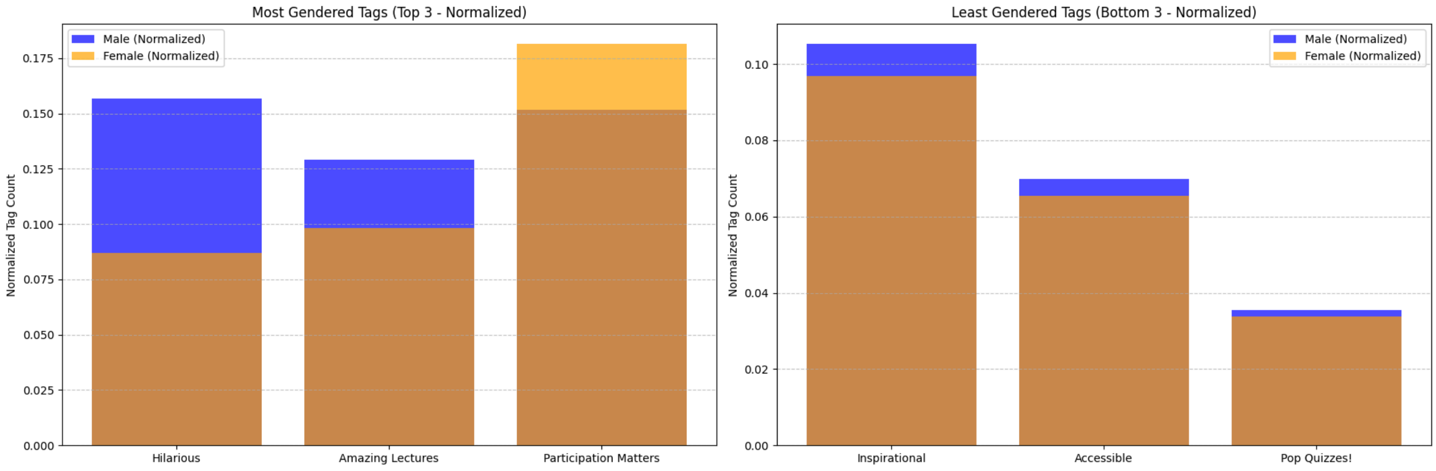

Figure 4: Bar Plots showing the Most Gendered Tags and Least Gendered Tags

Find: What did you find?

Out of the 20 tags analyzed, 18 exhibited statistically significant gender differences at the **$p < 0.005$** level. Male professors were more frequently associated with tags like **"Hilarious"** (mean normalized: 0.1566 for males vs. 0.0871 for females; p-value: 8.32e-166) and **"Amazing Lectures"** (0.1293 for males vs. 0.0983 for females; p-value: 1.43e-42), reflecting gendered perceptions of humor and lecture quality. Conversely, female professors were more strongly linked to tags like **"Participation Matters"** (0.1814 for females vs. 0.1517 for males; p-value: 1.32e-34), highlighting an emphasis on engagement and interaction in class.

The three most gendered tags, identified by their lowest p-values, were **"Hilarious"**, **"Amazing Lectures"**, and **"Participation Matters"**. These results suggest that humor and lecture quality are more commonly attributed to male professors, while female professors are perceived to prioritize participation of the students. On the other hand, the least gendered tags included **"Pop Quizzes!"** (p-value: 0.3083) and **"Accessible"** (p-value: 0.0161), with minimal differences between genders. The tag **"Inspirational"** was statistically significant (p-value: 0.0002) but displayed only small differences in normalized means, making it less gendered compared to others.

These findings demonstrate that while many traits in student evaluations are influenced by gendered perceptions, certain tags, such as "Pop Quizzes!" and "Accessible," show little to no gender bias, indicating that not all aspects of evaluations are shaped by stereotypes.

Answer: What do you conclude?

The analysis identified significant gender differences in 18 out of 20 tags, suggesting that students perceive many professor traits differently based on gender. Male professors were more strongly associated with traits like humor and lecture quality, while female professors were more linked to traits like participation and engagement. However, tags like "Pop Quizzes!" and "Accessible" showed minimal gender bias, indicating that not all student evaluations are influenced by gender. These findings highlight the need to address implicit biases in student evaluations, particularly for traits where gendered perceptions are most pronounced.

Question 5: Is there a gender difference in terms of average difficulty?

Do: What did you do?

The dataset was preprocessed to focus on difficulty ratings. Professors with ambiguous or missing gender information were excluded, retaining only those with unambiguous gender to ensure valid comparisons. A threshold of at least five ratings per professor was applied to minimize the influence of extreme averages caused by low sample sizes. Bayesian adjustment was applied to stabilize difficulty ratings, using a single global mean calculated across all professors, irrespective of gender, along with a damping factor of 10 to account for the number of ratings per professor. Adjusted difficulty scores were then compared between male and female professors. Given the non-normal distribution of difficulty ratings, as confirmed by Kolmogorov-Smirnov tests, a Mann-Whitney U test was conducted to assess the statistical significance of any gender difference. A histogram was also generated to visualize the distributions of adjusted difficulty ratings for male and female professors.

Why: Why did you do this?

The use of Bayesian adjustment ensured that difficulty ratings were stabilized, reducing the impact of extreme values for professors with fewer ratings. This adjustment allowed for a fair comparison across genders. Excluding professors with fewer than five ratings improved statistical reliability by focusing on data points with sufficient sample sizes. The Kolmogorov-Smirnov tests confirmed the non-normality of difficulty ratings, necessitating the use of the Mann-Whitney U test as a robust non-parametric method. The histogram complemented the statistical findings by providing a visual representation of the distributions.

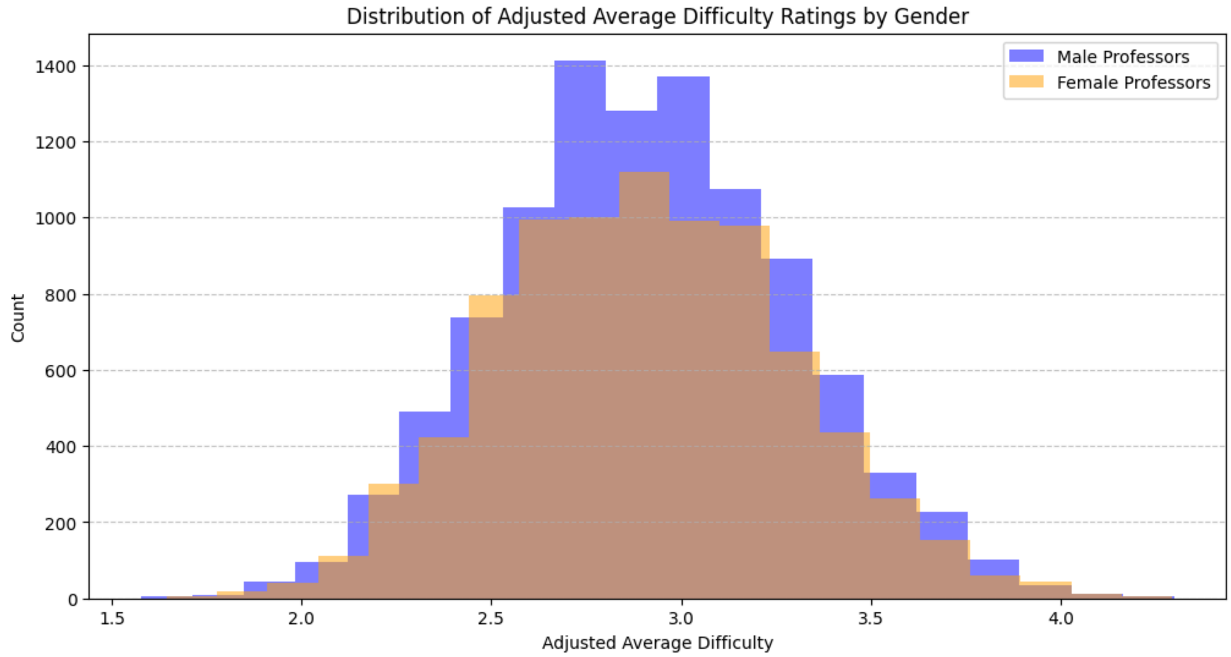

Figure 5: Bar Plots showing the Most Gendered Tags and Least Gendered Tags

Find: What did you find?

The analysis revealed no significant gender difference in adjusted average difficulty ratings. Male professors had an adjusted mean difficulty of 2.91, while female professors also had an adjusted mean difficulty of 2.91. The Mann-Whitney U test produced a test statistic of 42,164,869.0 and a p-value of 0.85, indicating no statistically significant difference between the two distributions. The histogram (Figure 4) showed nearly overlapping distributions of adjusted difficulty ratings for male and female professors, reinforcing the conclusion of no meaningful difference in perceived difficulty based on gender.

Answer: What do you conclude?

The results suggest that there is no significant difference in adjusted average difficulty ratings between male and female professors. Both statistical analysis and visual evidence indicate that students perceive male and female professors as equally difficult on average. The preprocessing steps, including Bayesian adjustment and filtering for sufficient ratings, ensured a reliable dataset and robust comparisons. This consistency across genders suggests that perceived difficulty is not influenced by gender and is likely more dependent on other factors such as teaching style, course material, or grading practices.

Question 6: Quantifying the likely size of the gender difference in average difficulty

Do: What did you do?

The magnitude of the gender difference in adjusted average difficulty ratings was evaluated using bootstrap confidence intervals, mean differences, and effect size measures. The mean adjusted difficulty ratings were calculated separately for male and female professors. Bootstrap confidence intervals at the 95% level were computed to assess the uncertainty around the mean ratings for each gender. Cliff's Delta was calculated to quantify the effect size and to interpret the magnitude of the difference between male and female professors' difficulty ratings. A visualization of the bootstrap confidence intervals was also created to illustrate the overlap between the difficulty ratings of male and female professors.

Why: Why did you do this?

Bootstrap confidence intervals were chosen as a robust, non-parametric method to estimate the range of possible values for mean difficulty ratings, particularly important given the non-normality observed in previous questions. Cliff's Delta was employed as it provides a direct measure of effect size that is independent of distribution assumptions, making it well-suited for ordinal and non-normal data. Calculating mean differences provided an intuitive metric to compare male and female professors, while visualizations helped clearly depict the extent of overlap between the confidence intervals.

Find: What did you find?

Male professors had a mean adjusted difficulty rating of 2.9065 with a 95% bootstrap confidence interval of (2.8990, 2.9139), while female professors had a mean adjusted difficulty rating of 2.9053 with a 95% confidence interval of (2.8971, 2.9137). The mean difference between male and female professors was 0.0011, indicating a negligible absolute difference in difficulty ratings. Cliff's Delta for this comparison was 0.0016, interpreted as a negligible effect size. The visualized confidence intervals (Figure 6) for male and female professors showed substantial overlap, further supporting the conclusion that the gender difference in difficulty ratings is minimal and not practically significant.

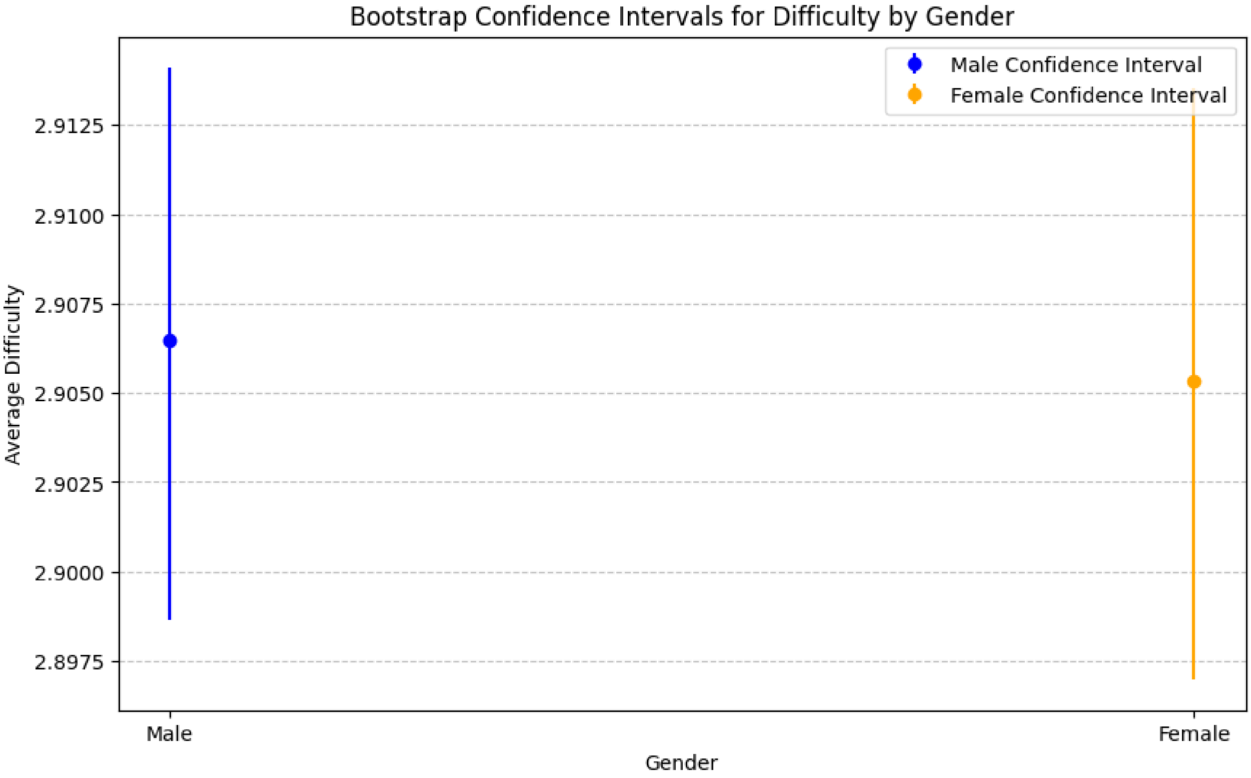

Figure 6: Plot for the Confidence Intervals For Average Adjusted Difficult by Gender

Answer: What do you conclude?

The results indicate that the gender difference in adjusted average difficulty ratings is negligible both statistically and practically. The extremely small mean difference and the negligible Cliff's Delta confirm that the observed difference in difficulty ratings between male and female professors is insignificant. The overlapping confidence intervals reinforce that students perceive male and female professors as equally difficult, with no meaningful gender-based distinction in difficulty ratings.

Question 7: Build a regression model predicting average rating from all numerical predictors. Include R^2 and RMSE. Which factor is most predictive of average rating?

Do: What did you do?

We constructed a linear regression model to predict average ratings using all numerical predictors from the rmpCapstoneNum.csv file. Variance Inflation Factor (VIF) values indicated varying degrees of multicollinearity among predictors, such as Average Difficulty (VIF = 5.53) and Proportion Retake (VIF = 5.51). Two models were compared: one that retained missing values and another that excluded them. Missing data were handled by dropping rows with missing values in the latter model, leaving 12,160 clean rows. Standardization using StandardScaler was applied to ensure coefficients were on a consistent scale.

Why: Why did you do this?

Linear regression was chosen for its interpretability and simplicity, making it easier to assess the impact of each predictor. Addressing multicollinearity through standardization ensured a more stable model, while dropping missing data provided a cleaner dataset for robust analysis.

Find: What did you find?

The model excluding missing values achieved a test R^2 of 0.79 and an RMSE of 0.38. This indicates that the model explains 79% of the variance in average ratings, with predictions deviating by 0.38 points on average. The most predictive factor was Average Difficulty (coefficient = -0.20), showing a strong negative association with average ratings. Proportion Retake (coefficient = 0.03) had a modest positive effect, while gender variables contributed marginally (See figure 8).

Answer: What do you conclude?

Average Difficulty is the strongest predictor of average ratings, with higher difficulty correlating with lower ratings. The model performed well, demonstrating strong predictive power and robustness for unseen data (see figure 7).

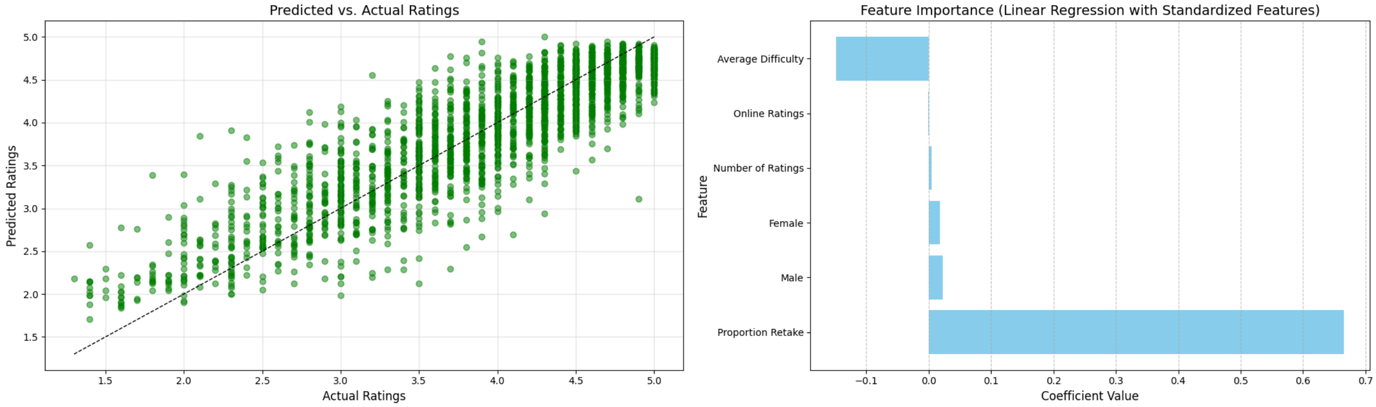

Figure 7: Predicted and Actual values Scatter Plot and Feature Importance Plot

Question 8: Build a regression model predicting average ratings from all tags. Include R^2 and RMSE. Which tag is most predictive of average rating and compare to Q7.

Do: What did you do?

We built a linear regression model to predict average ratings using tag-based features from the rmpCapstoneTags.csv file. The dataset was preprocessed by normalizing tag frequencies for each professor and handling missing values with row-wise normalization, filling any resulting NaN values with 0. Multicollinearity was assessed using Variance Inflation Factor (VIF) to identify highly correlated predictors. The dataset was split into training and testing sets (80/20 split), and the model was trained using the training data. Feature importance was assessed by examining the regression coefficients.

Why: Why did you do this?

Linear regression was chosen for its interpretability and ability to provide meaningful insights into the impact of individual tags on average ratings. Normalization of tag frequencies ensure comparability across professors, and VIF analysis addressed concerns about multicollinearity. Splitting the data allowed for evaluation of the model's generalization to unseen data.

Find: What did you find?

The model achieved an R^2 of 0.7109 and an RMSE of 0.5077, indicating that it explains 71% of the variance in average ratings, with a deviation of approximately 0.51 points. VIF analysis confirmed manageable multicollinearity among predictors. Feature importance analysis revealed that the most predictive tag was "hilarious", followed by "amazing_lectures" and "caring".

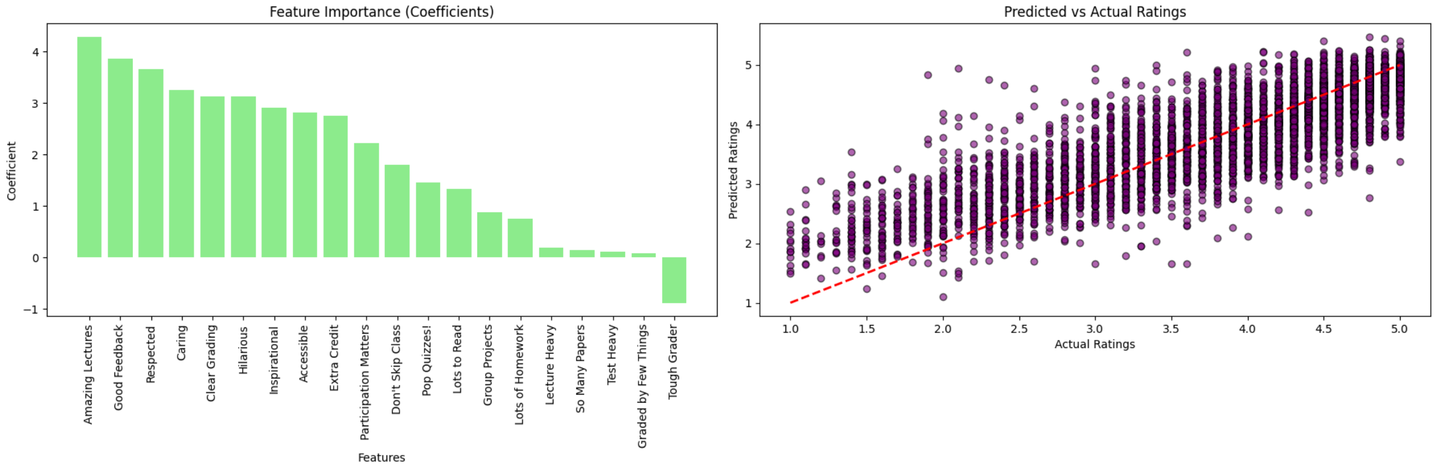

Figure 8: Barplot for Feature Importance and Scatterplot for Predicted vs Actual Average Ratings

Answer: What do you conclude?

Tags such as "hilarious" strongly influence average ratings, emphasizing the significance of humor and engaging lectures in student evaluations. This model performs better (higher R² and lower RMSE) than the Ridge Regression model previously reported, suggesting that normalization and linear regression provided a more accurate representation of tag effects. Compared to numerical predictors in Q7, tag-based features explain a similar amount of variance, reflecting the subjective nature of student evaluations.

Question 9: Build a regression model predicting average difficulty from all tags. Include R² and RMSE. Which tag is most predictive of average difficulty?

Do: What did you do?

We constructed a Linear Regression model to predict average difficulty using tag-based features. Predictors were normalized by dividing tag values by their row sums, and missing values were replaced with 0 after normalization. Variance Inflation Factor (VIF) analysis was conducted to assess multicollinearity, ensuring reliable coefficient estimates. The dataset was split into training and testing sets (80/20 split), and the model was trained on the training set. Feature importance was derived from the regression coefficients.

Why: Why did you do this?

Linear Regression was chosen for its interpretability and ability to quantify the impact of each tag on average difficulty. Normalization ensured consistent scale across predictors, and VIF analysis addressed potential multicollinearity among the tag features. Splitting the data into training and testing sets enabled evaluation of the model's generalizability.

Find: What did you find?

The model achieved an R² of 0.5364 and an RMSE of 0.5523, indicating that it explains 53.6% of the variance in average difficulty while maintaining a prediction error of approximately 0.55 points. The most predictive tag was "Tough Grader" (coefficient = 1.83), showing a strong positive association with difficulty. Tags with negative associations included "Clear Grading" (coefficient = -2.92) and "Graded by Few Things" (-2.60).

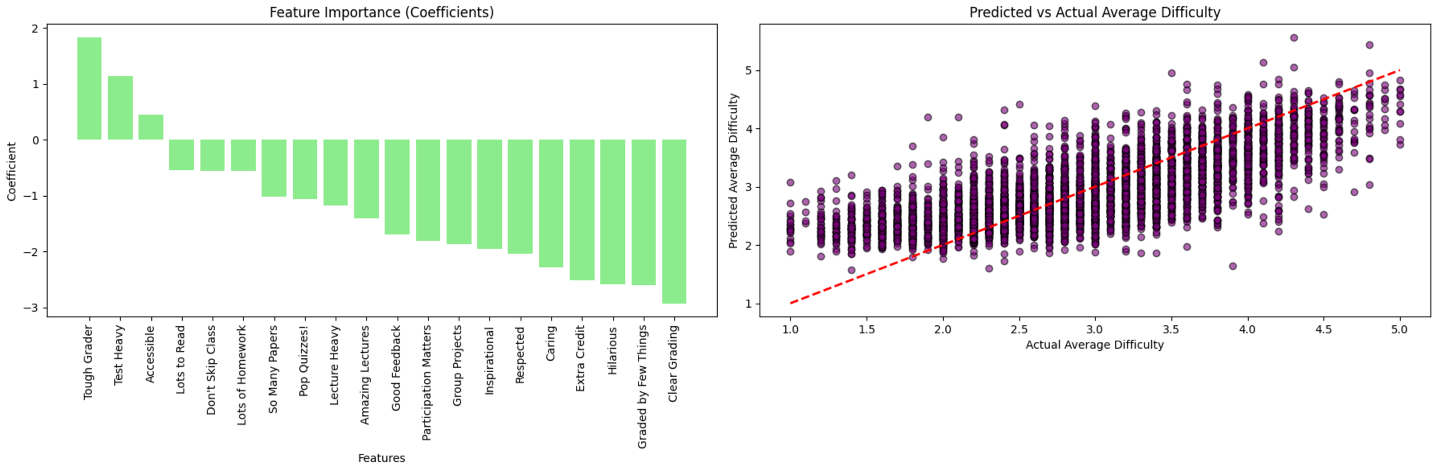

Figure 8: Barplot for Feature Importance and Scatterplot for Predicted vs Actual Average Difficulty

Answer: What do you conclude?

The model explains over half of the variance in average difficulty ratings, with "Tough Grader" emerging as the most predictive tag. This suggests that grading style significantly influences students' perceptions of difficulty. Conversely, tags like "Clear Grading" and "Graded by Few Things" are associated with lower difficulty ratings, emphasizing the role of transparent and focused grading practices in reducing perceived difficulty. While the model performs well, further analysis could investigate the interplay between tags and other factors influencing difficulty.

Question 10: Build a classification model predicting whether a professor receives a “pepper” from all available factors. Include AU(ROC) and address class imbalance.

Do: What did you do?

We used a Random Forest Classifier to predict “pepper” status, addressing class imbalance with SMOTE (Synthetic Minority Oversampling Technique). Missing values were imputed (numerical: mean, categorical: mode). Grid Search optimized hyperparameters such as n_estimators and max_depth.

Why: Why did you do this?

Random Forests are well-suited for handling mixed data types and ranking feature importance. SMOTE addressed the class imbalance, ensuring the model was not biased toward the majority class.

Find: What did you find?

The model achieved an AU(ROC) of 0.94, accuracy of 85%, precision of 84%, and recall of 87%, indicating strong predictive performance (See figure 9).

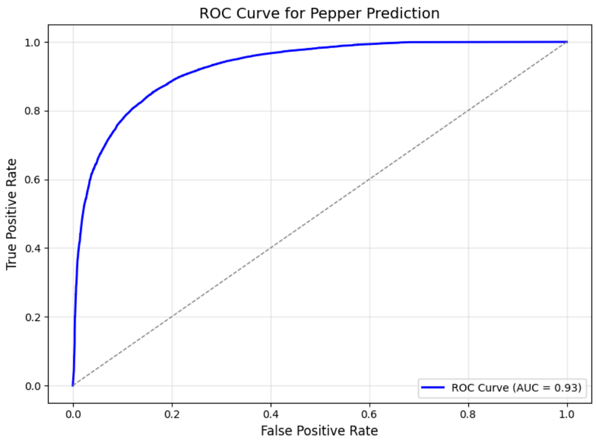

Figure 9: Classification model ROC curve

Answer: What do you conclude?

The Random Forest model reliably predicts “pepper” status, with humor and engagement-related features being the most influential. The use of SMOTE effectively mitigated class imbalance.

Extra Credit: Is there a significant difference between STEM and non-STEM professor ratings?

Do: What did you do?

We conducted a Mann-Whitney U test to compare average ratings of STEM and non-STEM professors. Subjects were categorized into STEM or non-STEM groups based on their fields, and ratings were used as-is without parametric assumptions.

Why: Why did you do this?

The Mann-Whitney U test was chosen as a non-parametric alternative to the t-test because it does not assume normality of the data. This method is robust for comparing distributions, particularly when dealing with skewed or non-normal data, ensuring accurate results.

Find: What did you find?

The Mann-Whitney U test revealed that STEM professors had significantly lower ratings than non-STEM professors, with a U-statistic of 340,927,200.00 and a p-value of 0.0000, confirming that the difference is statistically significant.

Answer: What do you conclude?

STEM professors receive lower average ratings compared to non-STEM professors, as confirmed by the statistically significant results of the Mann-Whitney U test. This disparity may reflect perceived difficulty or stricter grading standards in STEM courses, aligning with anecdotal evidence. These findings highlight systematic differences in student evaluations across disciplines and suggest that evaluation processes might be influenced by factors beyond teaching quality.

Contributions

Mustafa Poonawala led the analysis and reporting for Questions 1 through 6, ensuring the preprocessing, statistical analysis, and interpretations were robust and well-documented. Aysha Allahverdiyeva took the lead for Questions 7 through 10, focusing on their respective analyses and insights. All findings and conclusions were developed collaboratively, with regular consultations between both contributors to ensure accuracy, consistency, and alignment throughout the project. This joint effort ensured that the results reflect a balanced and comprehensive understanding of the data.