Capstone Project Report

CAP-9 - Hongjiu Zhang, Smruti Nalawade and Carina Sun

Introduction

This group named CAP-9 comprised 3 team members - Hongjiu Zhang [hz3501@nyu.edu], Smruti Nalawade [ssn9062@nyu.edu] and Carina Sun [as13537@nyu.edu]. Every member participated in the discussions, coding, and report writing for the project. It was a collaborative effort with remote and in-person discussions. The specific individual contributions were:

- **Hongjiu Zhang**: Q1, Q2, Q3, Extra Credit Question and Code Review
- **Smruti Nalawade**: Q4, Q5, Q6 and write-up for Introduction and Data Preprocessing
- **Carina Sun**: Q7, Q8, Q9, Q10, and integration of the project report

Data Preprocessing

1. Data Loading

- Datasets were imported using `pd.read_csv()` to load numeric, qualitative, and tag data for processing, with column names defined based on the meaning of each variable. Random seed is **13082093**.

2. Handling Missing Values

- Rows with missing values in all the columns except the last two were checked and removed using `dropna()` to clean the data.

3. Normalization, Scaling, and Dimension Reduction

- Data was standardized using `StandardScaler` to ensure all features had a mean of zero and unit variance. This improves the performance of statistical models.
- Principal Component Analysis was applied to reduce feature dimensions while preserving essential variance for analysis.

Q1-Is there statistically significant evidence of a pro-male gender bias in student evaluations of professors?

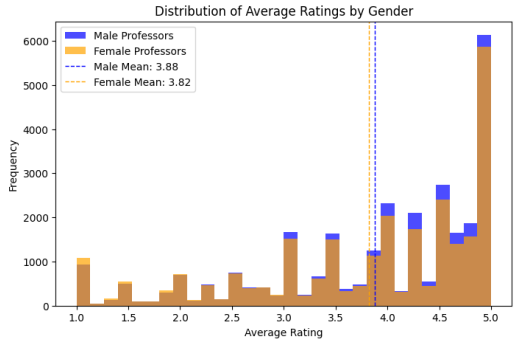


Figure 1: Distribution of Average Ratings by Gender

As pictured in Figure 1, male professors have a slightly higher mean rating than female professors. Next, we'll use statistical tests to validate our points.

Null Hypothesis (H_0): No significant difference in average ratings between male and female professors.

Alternative Hypothesis (H_1): Male professors have higher average ratings than female professors ($\mu_{\text{male}} > \mu_{\text{female}}$).

Statistical Tests and Results– If both the male and female groups pass the Kolmogorov-Smirnov (KS) normality test and the variances are equal according to Levene's test, the data meets the assumptions for a parametric test, and a two-sample t-test is appropriate.

Otherwise, a non-parametric Mann-Whitney U test is used instead. **Normality Test:** Kolmogorov-Smirnov Test shows that ratings for both genders deviate from normality (p-values < 0.005). **Homogeneity of Variance:** Levene's Test confirms variances are unequal (p-value < 0.005).

Metric	Male Professors	Female Professors
Mean Rating	3.88	3.82
Standard Deviation	0.67	0.72

Test for Mean Difference: Mann–Whitney U Test was used because the assumptions for a parametric test (e.g., normality or equal variance) were likely not met for this dataset. The test statistic is 406757411.5000, and p-value $= 1.222 \times 10^{-5}$, which is less than the significance level ($a = 0.005$). **Effect Size** Cohen's $d = 0.0549$ (very small effect), suggesting the difference in average ratings is minimal in practical terms.
Based on the p-value, **there is evidence of a pro-male gender bias in ratings**. However, the effect size suggests that the magnitude of this bias is very small.

Q2-Is there a statistically significant gender difference in the variance of professor ratings?

Null Hypothesis (H_0): No difference in the variance of ratings between male and female professors.
Alternative Hypothesis (H_1): A significant difference exists in the variance of ratings.

We aim to validate the **Homogeneity of Variance Tests** and ensure the validity of statistical assumptions—normality (Kolmogorov-Smirnov), equal variance (Levene's and Bartlett's) : **Levene's Test** (statistic = 77.4498, p-value $= 1.3989 \times 10^{-18}$) confirms a significant variance difference. **Bartlett's Test** (statistic = 67.1125, p-value $= 2.5645 \times 10^{-16}$) supports this under normality. **KS Test** (statistic = 0.0240, p-value $= 1.6951 \times 10^{-7}$) shows significant differences in rating distributions.
The variance of ratings is 1.1840 for male professors and 1.3052 for female professors. Figure 2 shows similar medians but a wider spread for female professors.
Statistical tests confirm significantly higher variance in ratings for female professors, indicating more polarized student opinions.

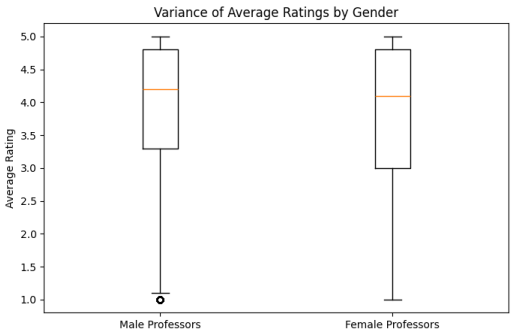


Figure 2: Distribution of Average Ratings by Gender.

Q3-What is the likely size of both of these effects as estimated from this dataset with 95% confidence?

- **For Average Rating: Null Hypothesis (H_0):** There is no significant difference in the average ratings of male and female professors. **Alternative Hypothesis (H_1):** Male professors receive significantly higher average ratings than female professors.
- **For Variance (Spread of Ratings): Null Hypothesis (H_0):** There is no significant difference in the variance of ratings between male and female professors. **Alternative Hypothesis (H_1):** Female professors exhibit significantly greater variability in ratings than male professors.

Gender Bias in Average Rating

Mean Difference: 0.0612 (Male professors have slightly higher ratings).
95% Confidence Interval (Bootstrap): [0.0428, 0.0796]. The confidence interval excludes 0, indicating that the difference is statistically significant.

Effect Size (Cohen's d): 0.0549. This is considered a very small effect size, suggesting minimal practical significance despite statistical significance.
The mean difference, while statistically significant, has limited real-world implications. The small Cohen's d value confirms that the practical magnitude of the bias is negligible. This may indicate the presence of systemic but minor bias favoring male professors in student evaluations.

Gender Bias in Spread of Ratings (Variance)

Variance Difference: -0.1213 (Female professors have more variability in ratings).
Levene's Test: P-value $=1.3989 \times 10^{-18}$ (Highly significant).
Female professors exhibit greater variability in ratings, with a variance of 1.3052 compared to 1.1840 for male professors. This statistically significant difference, indicated by Levene's test, suggests that students may have more polarized opinions about female professors.
Male professors receive slightly higher average ratings than female professors, with a statistically significant but negligible **effect size (Cohen's d = 0.0549)**.

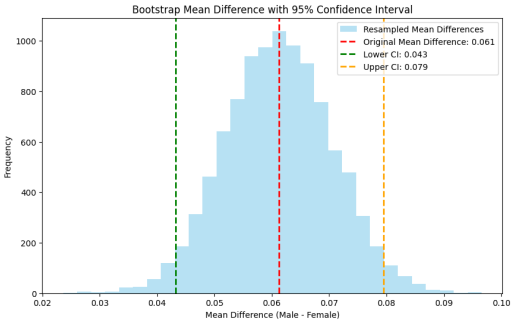


Figure 3: Bootstrap Mean Difference with 95% Confidence Interval.

Q4: Do students treat male and female professors differently when using certain tags, or is this just by chance?

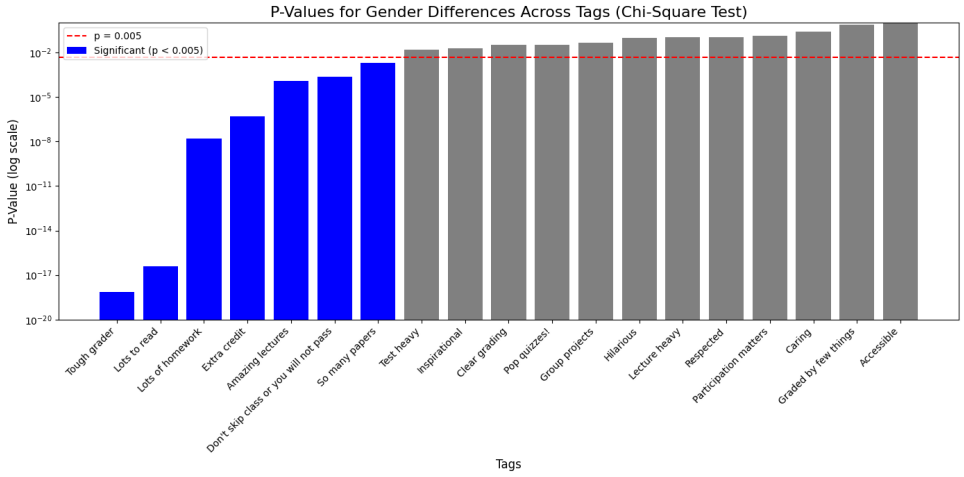


Figure 4: P-Values for Gender Differences Across Tags (Chi-Square Test).

To test if gender differences exist in tag usage, we came up with the following hypothesis:

- **Null Hypothesis (H_0):** There is **no significant gender difference** in the number of times a given tag was assigned by students to male vs. female professors.
- **Alternative Hypothesis (H_1):** There is a **significant gender difference** in the number of times a given tag was assigned by students to male vs. female professors.

The Chi-Square test was chosen because it helps determine if there is a connection between two categorical variables—in this case, gender and the use of tags by students—and it allows for a comparison of the frequency of tag assignments to male versus female professors to understand any differences.

Results

The Chi-Square tests were conducted across all 20 tags, producing p-values to determine statistical significance. Below are the key statistical observations:

Most Gendered Tags	P-value
Tough Grader	7.16×10^{-19}
Lots to Read	3.94×10^{-17}
Lots of Homework	1.67×10^{-9}

Table 1: Tags Showing Strong Evidence

Least Gendered Tags	P-value
Accessible	0.906
Graded by Few Things	0.717
Caring	0.275

Table 2: Tags Showing No Statistical Gender Bias.

Q5: Is there a gender difference in terms of average difficulty?

Analysis

- **Null Hypothesis (H_0):** There is no statistically significant gender difference in the average difficulty ratings between males and females.
- **Alternative Hypothesis (H_1):** There is a statistically significant gender difference in the average difficulty ratings between males and females.

We wanted to check if there is a difference in the average difficulty ratings reported by male and female participants. To do this: First, we used a **two-sample t-test**, which compares the average ratings of two separate groups (males vs. females). We also checked if the two groups had similar variances, which is important for the t-test. The variance ratio was calculated as **1.0100**, confirming the assumption of equal variance.

Results

The T-test results (T-statistic $= -0.5691$, p-value $= 0.5693$) indicate no statistically significant difference in average difficulty ratings between male and female participants ($p = 0.5693 > 0.05$). Additionally, the variance between the two groups was similar, meeting the assumptions of the test. These findings suggest that gender does not influence how participants rate difficulty levels.

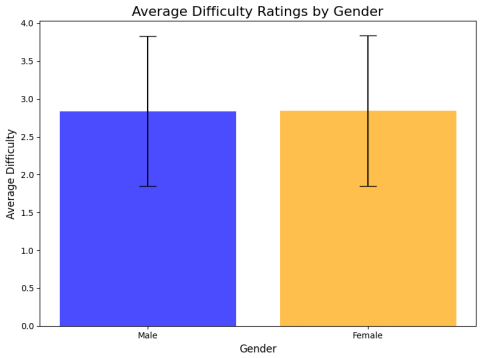


Figure 5: Average Difficulty Ratings by Gender.

Q6: Quantify the Likely Size of This Effect at 95% Confidence

Analysis

We compared the average difficulty scores between males and females using statistical methods. The mean, standard deviation, and confidence interval were calculated to evaluate differences between the groups. **Welch's t-test** was used as it is robust for cases with unequal sample sizes and variances.

Key Findings

The average difficulty scores for males and females are nearly identical, with a minimal difference of -0.0048 that is not statistically significant. The 95 percent confidence interval (-0.0211 to 0.0116) includes zero, indicating that the observed difference is likely due to chance rather than a meaningful effect.

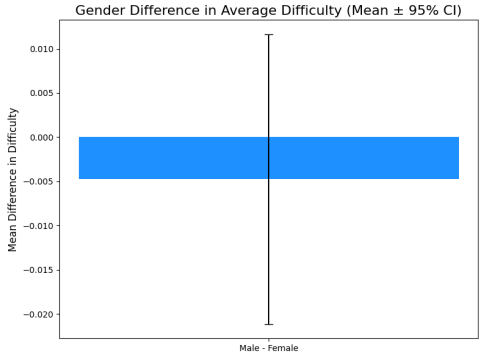


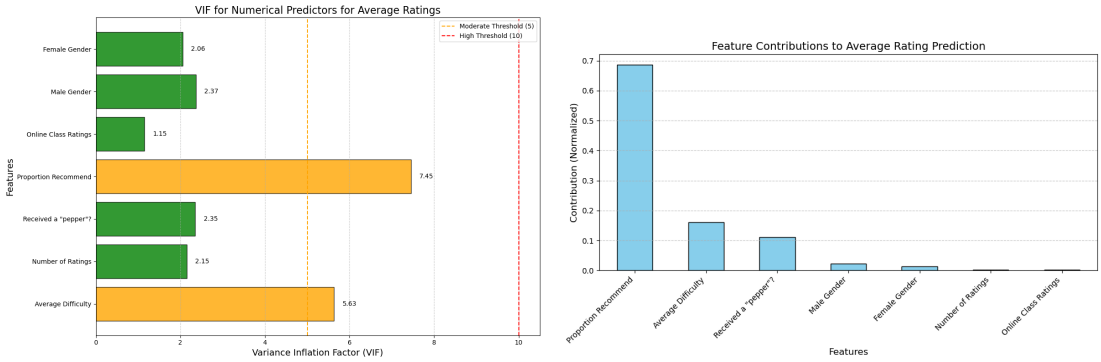
Figure 6: Gender Difference in Average Difficulty (Mean ± 95% CI).

Metric	Value
Mean for Males	2.8385
Mean for Females	2.8432
Standard Deviation for Males	0.9894
Standard Deviation for Females	0.9944
Sample Size (Males)	29,376
Sample Size (Females)	27,139
Mean Difference (Male - Female)	-0.0048
95% Confidence Interval	(-0.0211, 0.0116)

Figure 7: Summary of Statistical Results for Gender Difference in Average Difficulty.

Q7-Which of these factors is most strongly predictive of average rating?

VIF and Feature Contributions to Prediction



(a) Variance Inflation Factor (VIF) for Numerical Predictors for Average Ratings. Threshold is 5.

(b) Feature Contributions to Average Rating Prediction. The features were normalized.

Figure 8: Analysis of VIF and Feature Contributions to Predicting Average Ratings.

VIF Analysis Table

Feature	VIF
Average Difficulty	5.631764
Number of Ratings	2.150071
Received a "pepper"?	2.350548
Proportion Recommend	7.453163
Online Class Ratings	1.148710
Male Gender	2.365377
Female Gender	2.055549

Figure 9: Variance Inflation Factor (VIF) for Features.

Model Performance

The model achieved strong performance metrics, indicating high predictive accuracy:

- **R^2 = 0.817:** The model explains 81.7% of the variance in Average Rating.
- **RMSE = 0.359:** The Root Mean Square Error reflects a low average prediction error, indicating that the model's predictions are close to the actual values.

Analysis and Interpretation

Before conducting our research, we first addressed potential multicollinearity. We calculated the Variance Inflation Factor (VIF) to determine if there were any alarming levels. To address multicollinearity,

Principal Component Analysis (PCA) was applied. PCA transformed the original features into a new set of uncorrelated principal components, retaining 95% of the variance from the original dataset. This transformation preserved most of the original information while eliminating redundancy.

With the PCA-transformed data serving as predictors, a multiple linear regression model was fitted. The PCA loadings matrix was used to map the principal components' regression coefficients back to the original features. This allowed us to assess how well each unique feature predicted Average Rating.

Key Findings

The contributions of the original features to the prediction of Average Rating were calculated by combining the PCA loadings with the regression coefficients. The regression model successfully predicts Average Rating with high accuracy ($R^2 = 0.817$, RMSE $= 0.359$). Among the predictors, **Proportion Recommend** emerged as the most influential factor, followed by **Average Difficulty** and **Received a "pepper"**.

Q8: Which of these tags is most strongly predictive of average rating?

Analysis and Interpretation

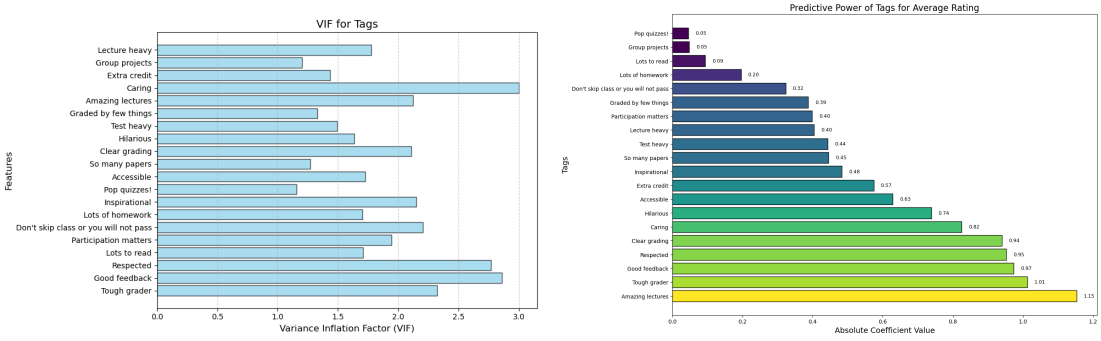
Before conducting our research, we first addressed potential multicollinearity. We calculated variance inflation factor (VIF) to determine if there is anything alarming. We found the VIF is relatively low, all below 3. But still, to improve the model performance. We used a variety of regression methods, such as Lasso, Ridge, and Ordinary Least Squares (OLS). By reducing irrelevant coefficients to zero, Lasso regression successfully identified the most significant predictors; Ridge regression reduced overfitting by allocating importance among correlated predictors; and OLS regression was used as the baseline, demonstrating the capabilities and drawbacks of a non-penalized model. We could assess prediction accuracy and model robustness by contrasting these methods.

Models and Performance

Model	R^2	RMSE	Most Predictive Tags	Least Predictive Tags
OLS	0.731	0.433	"Amazing Lectures", "Tough Grader"	"Lots to Read", "Group Projects"
Lasso	0.652	0.494	"Good Feedback", "Amazing Lectures"	"Lots of Homework", "Lecture Heavy"
Ridge	0.731	0.433	"Good Feedback", "Respected"	"Test Heavy", "Lecture Heavy"

Table 3: Model Performance and Predictive Tags

Figures and Visualizations



(a) Variance Inflation Factor (VIF) for Tags. (b) Predictive Power of Tags for Average Ratings.

Figure 10: VIF and Predictive Power for Tags

Key Findings

- **Most Predictive Tags**: Across models, "**Amazing Lectures**", "**Tough grader**", and "**Good feedback**" emerged as the strongest predictors. **Least Predictive Tags**:"Lots to read","Group projects" and "Pop Quizzes!" consistently showed minimal influence.

Q9: Which of these tags is most strongly predictive of average difficulty?

Analysis and Interpretation

The goal of the regression analysis was to estimate Average Difficulty by using normalized tags. For this case, three models were used: Ordinary Least Squares (OLS), Lasso Regression, and Ridge Regression. OLS was used as a baseline, while Lasso and Ridge attempted to address multicollinearity issues by applying penalties to the coefficients.

Among the three models, the highest R^2 of 0.571 was achieved using OLS, which explains 57.1% of the variability in Average Difficulty with an RMSE value of 0.489. While OLS suffers from instability due to multicollinearity, Ridge Regression provided a robust alternative by controlling coefficients, resulting in similar R^2 and RMSE values as OLS. Lasso Regression achieved an R^2 of 0.460 and RMSE of 0.548, focusing on interpretability by zeroing unnecessary variables and retaining only the most significant predictors.

Key Findings

Across all models, "Tough Grader" and "Test Heavy" emerged as the strongest predictors of Average Difficulty. In the OLS and Ridge models, "Accessible" was also identified as a key factor.

Model Results

Model	R^2	RMSE	Most Predictive Tags
OLS	0.571	0.489	"Tough Grader", "Test Heavy", "Accessible"
Lasso	0.460	0.548	"Tough Grader", "Test Heavy", "Lots of Homework"
Ridge	0.571	0.489	"Tough Grader", "Test Heavy", "Accessible"

Table 4: Model Performance for Predicting Average Difficulty.

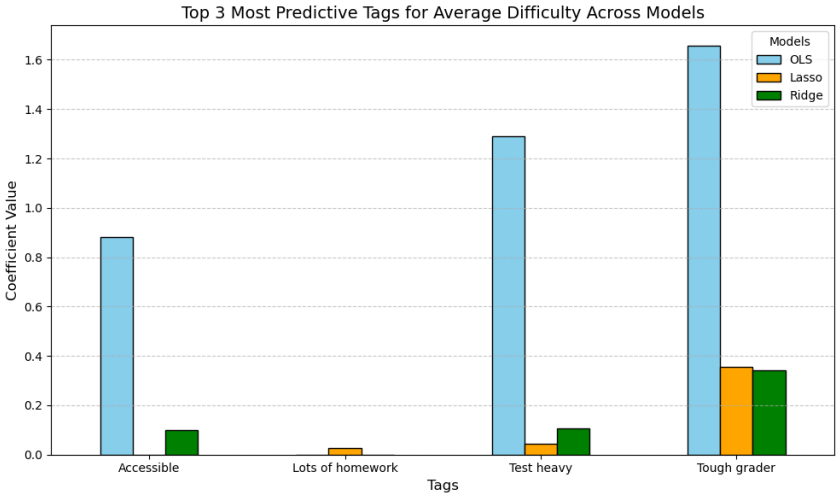


Figure 11: Top 3 Most Predictive Tags for Average Difficulty Across Models.

Q10: Build a classification model that predicts whether a professor receives a "pepper" from all available factors.

Analysis and Results

The classification model effectively predicts whether a professor receives a "pepper" based on both numerical and tag-based features. The ROC-AUC score of 0.75 demonstrates the model's capability to distinguish between the two classes effectively.
The classification report provides a detailed breakdown of the model's performance:

- **Overall Accuracy**: 73% indicates a reasonably strong predictive capacity.
- **Precision Score**: 0.51 suggests that 51% of the professors predicted to receive a "pepper" were correctly identified.
- **Recall Score**: 0.57 indicates that the model correctly identified 57% of professors who actually received a "pepper."
- **F1-Score:** 0.54 balances precision and recall, suggesting moderate performance for this class.

Classification Report

Class	Precision	Recall	F1-Score	Support
0.0	0.83	0.79	0.81	10082
1.0	0.51	0.57	0.54	3919
Accuracy		0.73		
Macro Avg	0.67	0.68	0.67	14001
Weighted Avg	0.74	0.73	0.73	14001

Table 5: Classification Report for Predicting "Pepper".

ROC Curve

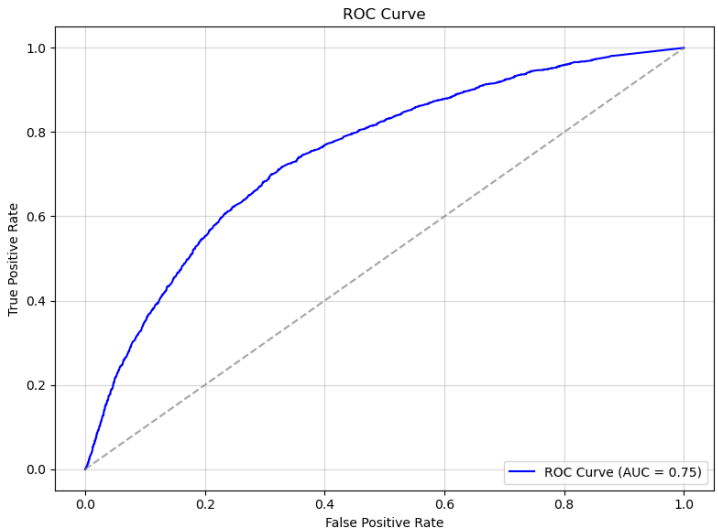


Figure 12: ROC Curve with AUC = 0.75.

Extra credits

Q1: Is there statistically significant evidence of a pro-male gender bias in student evaluations of professors?

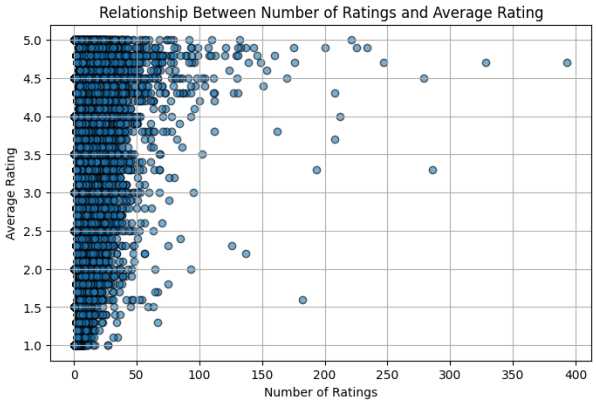


Figure 13: Distribution of Average Ratings by Gender.

As pictured in Figure 13, male professors have a slightly higher mean rating than female professors. Next, we'll use statistical tests to validate our points.

Null Hypothesis (H_0): No significant difference in average ratings between male and female professors.
Alternative Hypothesis (H_1): Male professors have higher average ratings than female professors ($\mu_{\text{male}} > \mu_{\text{female}}$).

Statistical Tests and Results – If both the male and female groups pass the Kolmogorov-Smirnov (KS) normality test and the variances are equal according to Levene's test, the data meets the assumptions for a parametric test, and a two-sample t-test is appropriate. Otherwise, a non-parametric Mann-Whitney U test is used instead.

Normality Test: Kolmogorov-Smirnov Test shows that ratings for both genders deviate from normality (p-values < 0.005).
Homogeneity of Variance: Levene's Test confirms variances are unequal (p-value < 0.005).

Q2: Do students rate online courses differently than in-person courses for male and female professors?

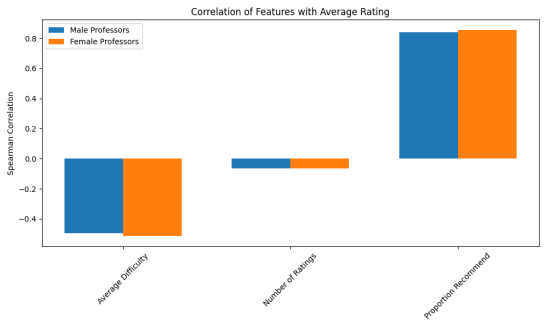


Figure 14: Comparison of Online and In-person Ratings by Gender.

Statistical Results Male Professors: T-test results are NaN, likely due to insufficient data. No significant difference between online and in-person ratings.
Female Professors: Similarly, T-test results are NaN, indicating a lack of sufficient data. No significant difference between online and in-person ratings.

Visualization The boxplot shows online ratings for male and female professors. Key Observations: Female professors show slightly more variability in ratings than male professors, but medians are similar. No data for in-person ratings is visible.

Conclusion There is insufficient data to determine significant differences between online and in-person ratings for male and female professors. Additional data is needed for a more robust analysis.