Capstone Project

Name: Yuri Chentsov, NetID: yc6393

May 1, 2025

Introduction

This capstone project analyzes professor effectiveness using a scraped dataset from RateMyProfessor.com. The goal is to explore patterns and potential biases in student evaluations through statistical and machine learning techniques.
The numeric and qualitative datasets were merged and cleaned by:

- cropping rows with missing average rating values
- filtering to professors with at least 3 ratings to reduce noise (the ratings number I personally use when checking professor ratings)
- creating derived features including `online_ratio`, `log_experience`, and categorical variables for gender, difficulty level, and experience level

To ensure reproducibility and prevent plagiarism, I seeded all randomness using my N-number `16378429` via `np.random.seed(16378429)`. $\alpha = 0.005$ was used as a significance level. The analysis was conducted using pandas and scikit-learn packages, visualization - using seaborn package.

1. Gender Bias in Ratings.

A t-test comparing male and female professor ratings shows a small but highly significant difference ($t = 5.3459$, $p < 0.001$). Male professors score 0.059 points higher on average (3.897 vs 3.838), note that this is only about 1.5% of the rating scale. To control for potential confounding variables, I conducted stratified analysis across three levels of teaching experience and course difficulty. Resulting matrix of rating differences and corresponding p-values for each strata are presented on Figure 2. Stratified analysis reveals this bias is context-dependent: more pronounced in harder courses and among experienced faculty (up to 0.112 points), but negligible in easier ones or for newer instructors.

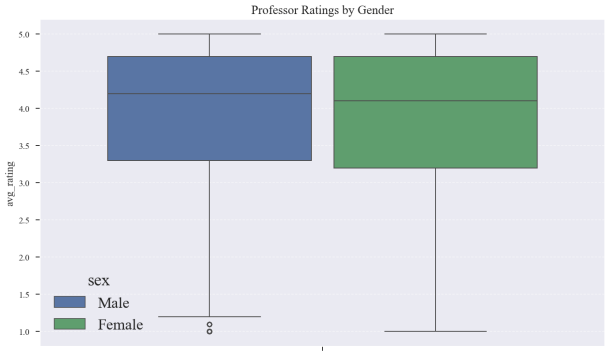

Figure 1: Boxplots of ratings by gender. Male professors receive slightly higher ratings overall.

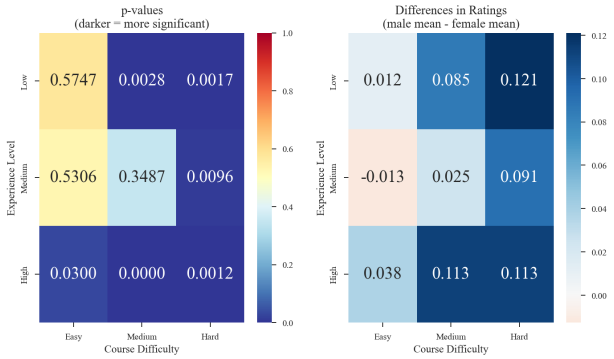

Figure 2: Stratified gender rating differences by experience and difficulty. Bias is more pronounced and statistically significant in challenging courses.

2. Ratings and Experience.

Linear regression using log of number of ratings as proxy for experience (log is necessary due to very skewed distribution of rating numbers) shows statistically significant minimal positive trend (slope = 0.066, $p < 0.001$, $R^2 = 0.002$). While experience barely affects average ratings, it substantially reduces variability. Less experienced professors show ratings spanning the full range from 1.0 to 5.0, while more experienced professors tend to receive ratings concentrated between 3.0 and 4.5 (note that remarkably empty area in the right-lower corner of the plot on Figure 3).

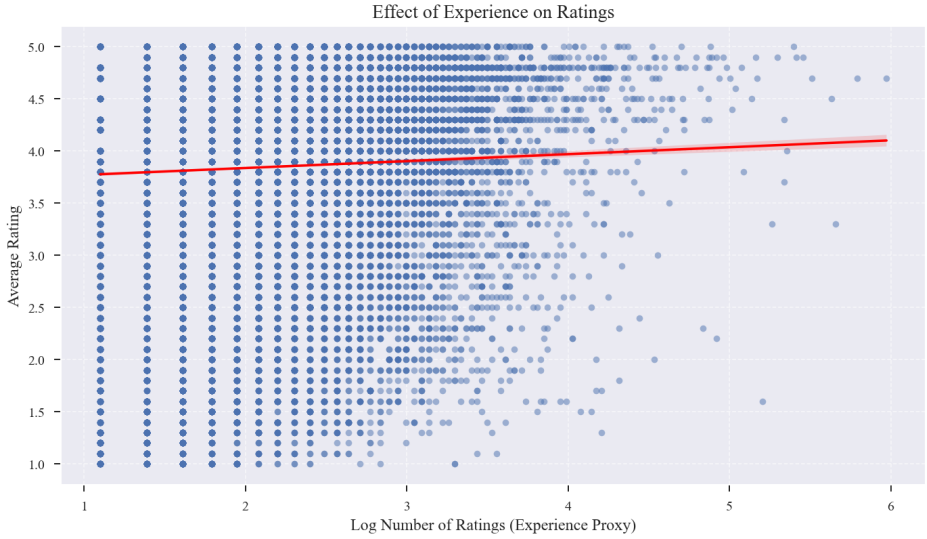

Figure 3: Rating vs. teaching experience. Ratings stabilize and vary less with experience.

3. Ratings and Difficulty.

A strong statistically significant negative correlation exists between course difficulty and ratings ($r = -0.590$, $p < 0.001$, $R^2 = 0.348$). The joint plot on Figure 4 reveals several interesting patterns: there's a clear negative trend, but also considerable variation around this trend. The hexbin density shows that most professors cluster in the middle range (difficulty 2.5–3.5, ratings 3.5–4.5), with fewer cases at the extremes. The marginal distributions indicate that while ratings are somewhat normally distributed with a negative skew (more high ratings), difficulty ratings show a right-skewed distribution with most courses being rated as relatively easy.

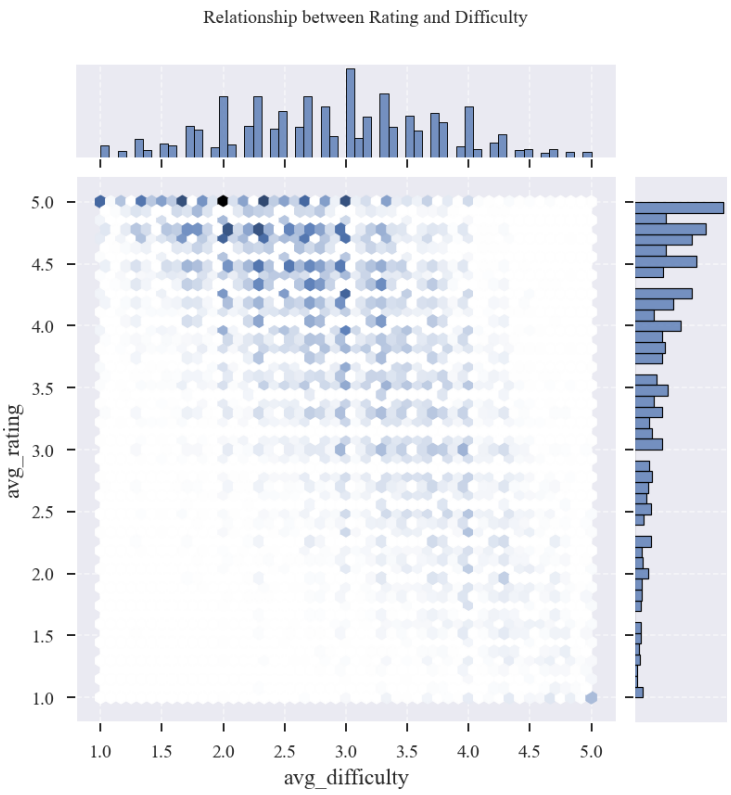

Figure 4: Rating vs. course difficulty (joint plot).

4. Ratings and Teaching Modality.

Two approaches were used. First, the continuous relationship between the proportion of online classes and ratings shows a weak negative correlation ($r = -0.063$, $p < 0.001$). Second, I compared professors with substantial online teaching ($> 10\%$ of classes) to those who teach primarily in traditional settings ($\leq 10\%$ online) (see Figure 5 for both methods). The analysis reveals that online-focused professors receive significantly lower ratings on average (3.710 vs 3.847, difference $= -0.137$, $p < 0.001$). The scatter plot shows considerable variation in ratings across all levels of online teaching, with a slight downward trend (red regression line). The boxplot visualization further confirms this pattern, showing con-

sistently lower ratings for online-focused professors.

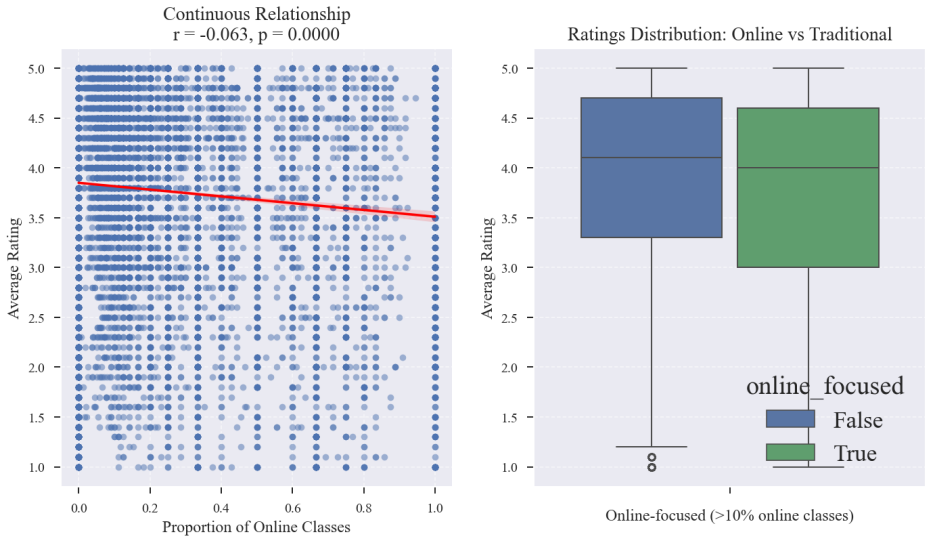

Figure 5: Ratings by percentage of online teaching.

5. Ratings and Course Satisfaction.

Ratings very strongly correlate with the `would_take_again` percentage ($r = 0.880$, $R^2 = 0.775$, $p \approx 0$). Regression: Rating $= 0.030 \times$ Percentage $+ 1.660$. The scatter plot on Figure 6 shows a clear linear trend (red regression line) with relatively consistent spread around the line across all percentage levels. Note that `would_take_again` information is only available for about 30% of the records.

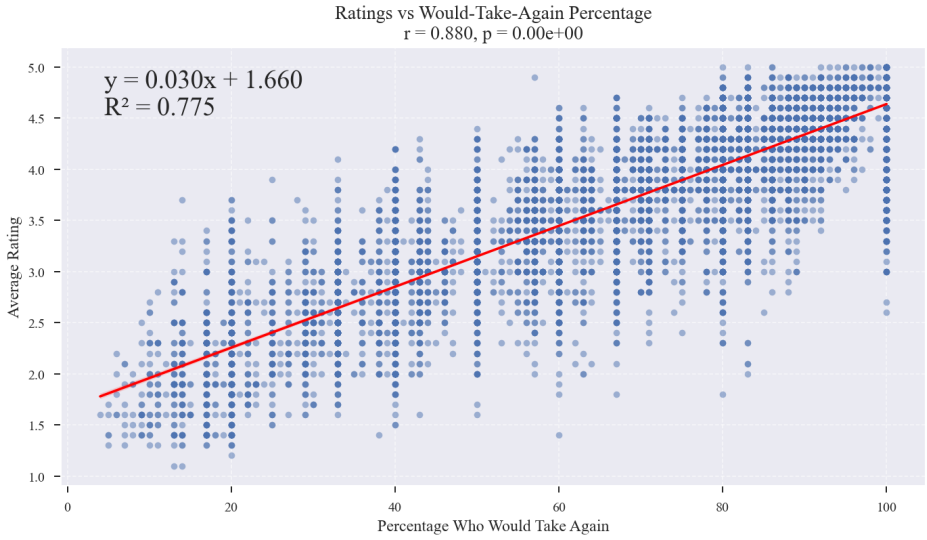

Figure 6: Ratings vs. `would_take_again` percentage.

6. Ratings and Sex Appeal.

Professors marked as "hot" receive much higher ratings (4.360 vs 3.504, $t = 92.194$, $p \approx 0$). The boxplot visualization shows minimal overlap between the distributions, with "hot" professors' ratings concentrated in the upper range (4.0–5.0) and showing less variation compared to other professors. This large difference in ratings suggests a strong "attractiveness bias" in student evaluations, though the direction of causality (whether being an effective teacher makes one more attractive or vice versa) cannot be determined from this analysis.

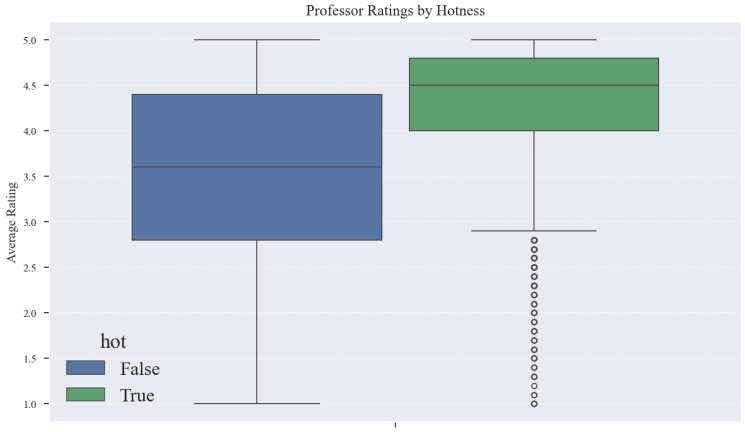

Figure 7: Boxplot of ratings by 'pepper' status.

7. Rating vs Difficulty Regression.

A simple linear regression model was built using a 80–20 train-test split to assess model performance. The model confirms a strong negative relationship between difficulty and ratings which was mentioned in Question 3. The model's performance is consistent across both training ($R^2 = 0.347$, RMSE = 0.806) and test ($R^2 = 0.351$, RMSE = 0.796) sets, explaining about 35% of the variance in ratings.

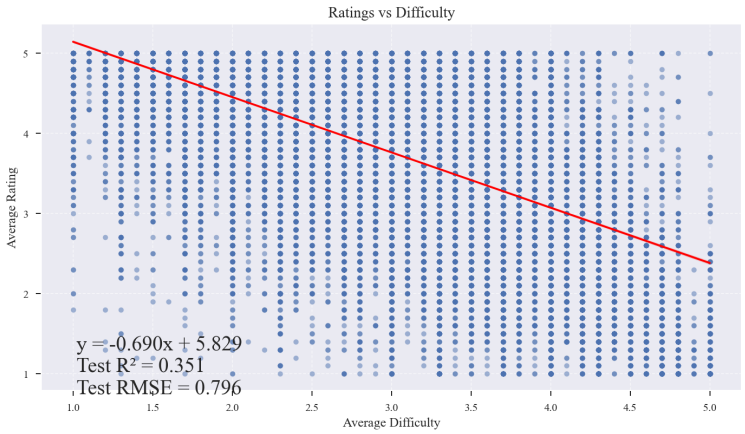

Figure 8: Predicted vs. actual ratings based on course difficulty.

8. Rating Multiple Regression.

I found out earlier that `would_take_again` is a strong predictor of rating, but this value is missing in more than half of the records. Thus, two multiple regression models were built: without and with `would_take_again`. For both models, I used standardized coefficients to compare feature importance and checked for multicollinearity using VIF scores. The first model achieves moderate predictive power ($R^2 = 0.445$, RMSE $= 0.736$) and reveals course difficulty as the strongest predictor ($\beta = -0.523$), followed by the pepper indicator ($\beta = 0.298$), with other factors having minimal impact. The second model shows dramatically improved performance ($R^2 = 0.803$, RMSE $= 0.371$). In this model, `would_take_again` becomes the dominant predictor ($\beta = 0.625$), while difficulty's influence decreases substantially ($\beta = -0.148$). Given these results, the optimal approach would be to maintain two separate models, applying one depending on data availability.

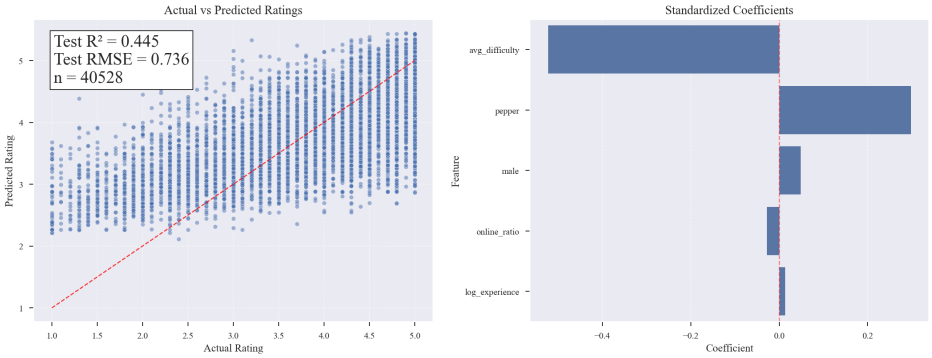

Figure 9: Model 1, Feature importance excluding `would_take_again`.

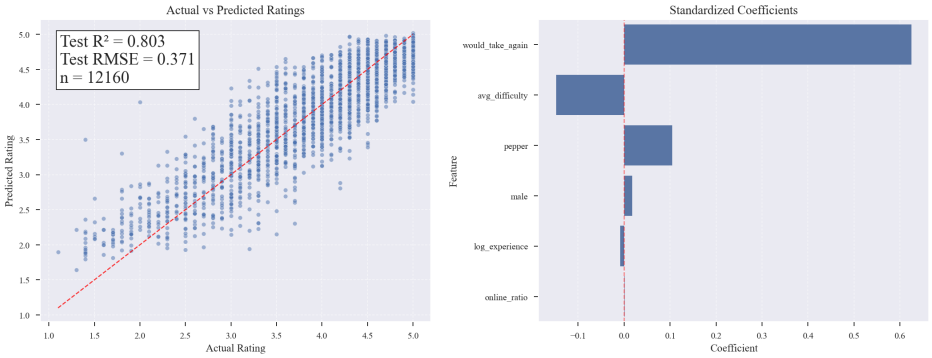

Figure 10: Model 2, Feature importance including `would_take_again`.

9. Sex Appeal vs Rating Logistic Regression.

A logistic regression using only ratings and adjusted for class imbalance (37.6% pepper vs. 62.4% no-pepper) through balanced class weights predicts pepper status moderately

well (AUC-ROC = 0.752, accuracy = 0.685, see Figure 11). High ratings increase odds of a pepper by 3.4x per point. The density plot illustrates why the model achieves moderate success: while there's significant overlap in the rating distributions, professors with peppers tend to have notably higher average ratings.

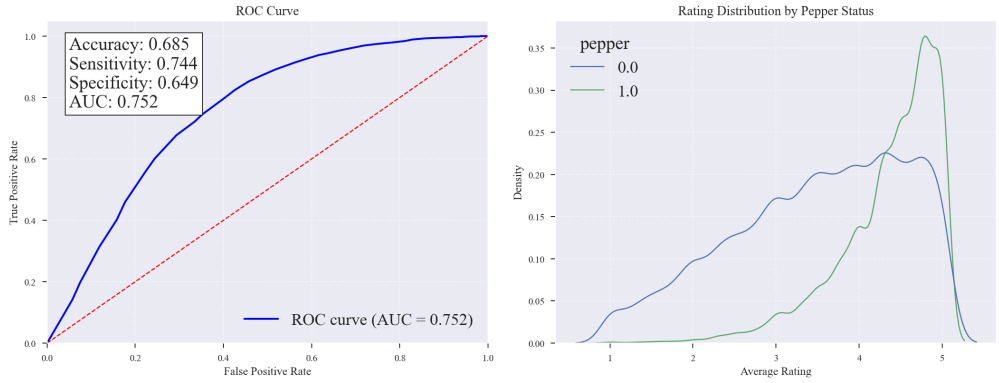

Figure 11: ROC curve and density plot for pepper prediction using ratings.

10. Sex Appeal Multiple Logistic Regression.

Compared to the rating-only model from Question 9, this full model shows modest improvement in performance (AUC-ROC: 0.780 vs. 0.752, Accuracy: 0.691 vs. 0.685). The standardized coefficients reveal that average rating remains by far the strongest predictor ($\beta = 1.300$), followed by teaching experience ($\beta = 0.358$), while other factors have relatively minor impacts ($|\beta| < 0.1$). Interestingly, both online teaching ratio and male gender show small negative associations with pepper status. The model maintains slightly better sensitivity (0.775) than specificity (0.640), indicating it's more successful at identifying professors who receive peppers than those who don't. While the additional features provide some predictive value, the modest improvement over the simpler rating-only model suggests that student attractiveness ratings are primarily driven by the overall teaching quality as measured by average ratings.

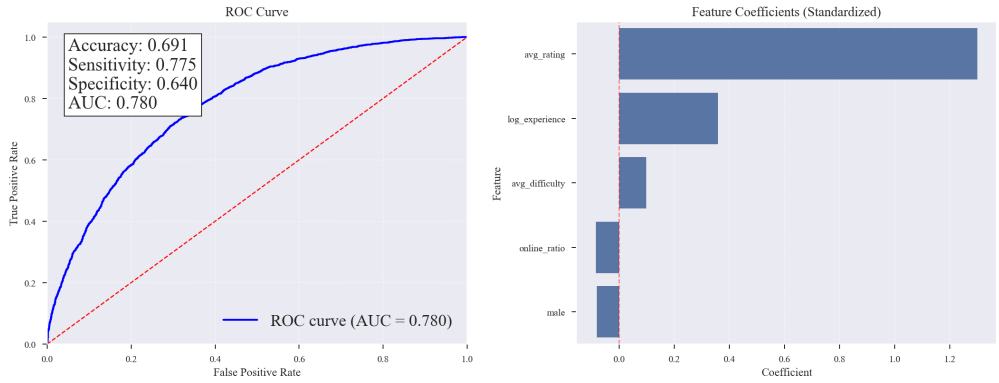

Figure 12: ROC curve and feature importance for full model.

Extra Credit

I analyzed the relationship between academic disciplines and professor ratings across the 20 most common majors in the dataset. An ANOVA test revealed that there are highly significant differences in ratings across disciplines ($F = 39.583$, $p = 5.60 \times 10^{-182}$), indicating that the field of study has a meaningful impact on professor ratings. The results show a striking pattern: STEM disciplines consistently occupy the lower end of the rating scale, with Physics (3.422), Chemistry (3.525), Economics (3.553), Engineering (3.598), and Computer Science (3.610) forming the bottom five. This pattern strongly corresponds with our previous observation that harder courses receive lower ratings, as STEM courses are typically considered more challenging. In contrast, humanities and social sciences dominate the upper end of the scale, with Criminal Justice (4.153), Languages (4.072), and Communication (4.011) receiving the highest ratings.

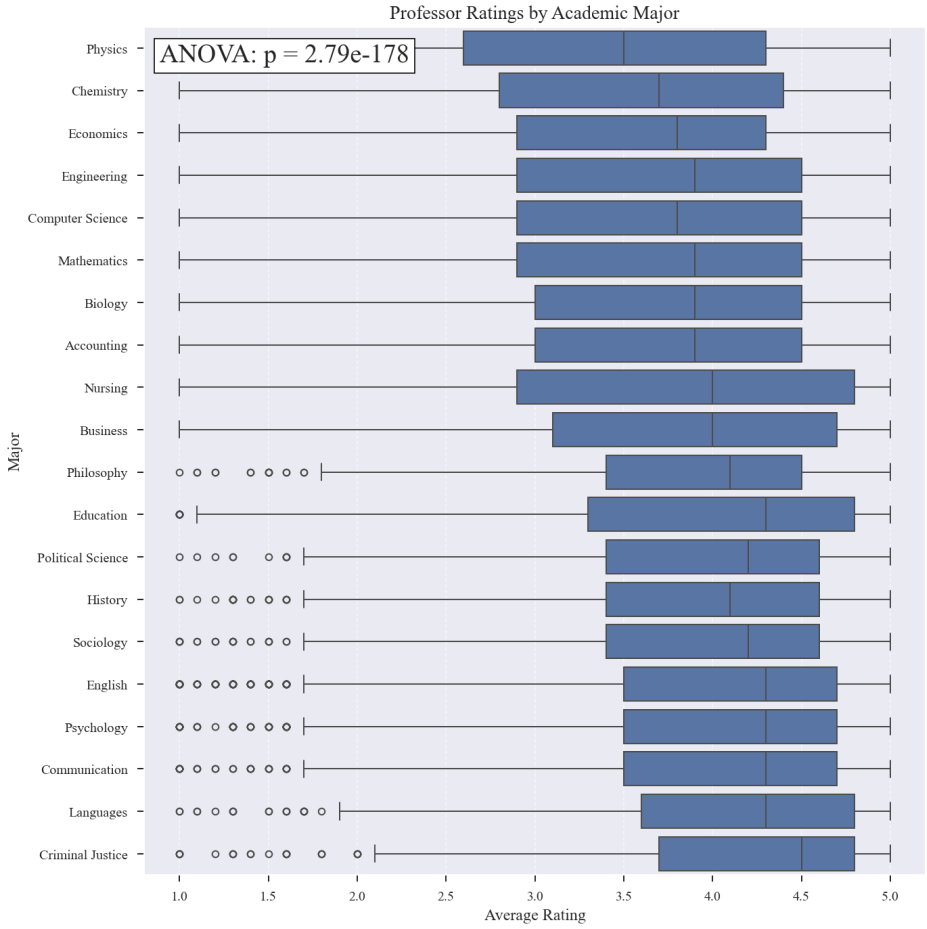

Figure 13: Boxplot of ratings by academic discipline.