Ali Lugo

9 December 2024

PODS Capstone

# Pre-processing and Data Cleaning

I focused on ensuring the datasets were clean and reliable. I started by seeding the random number generator with my N-number (16745560). Then, I loaded the numerical and qualitative datasets into separate data frames. I looked at their structure to find missing values and understand the columns.

To clean the data:

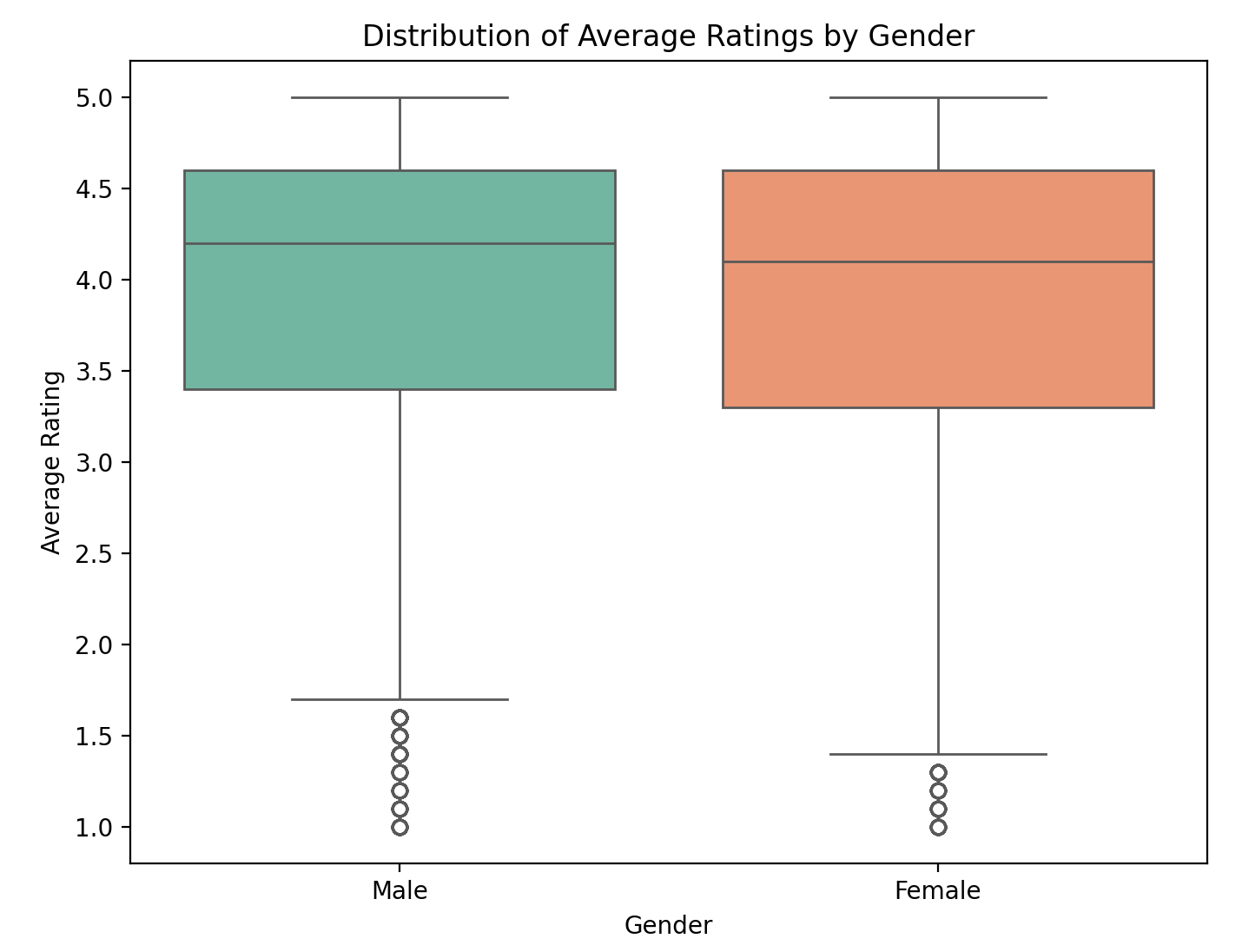
1. I renamed the columns in both datasets to have their names align with their content, making it more clear and consistent.
2. I dropped the rows with missing values in the "Average Rating" column because this variable is important for the analysis.
3. To improve the reliability of average ratings, I set a threshold of at least 5 ratings per professor. Professors with fewer ratings were excluded to reduce the influence of outliers.
4. Finally, I merged the cleaned and numerical and qualitative datasets by matching their indices. This created one comprehensive dataset with all the information needed for analysis.

# Question 1

*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.*

To assess whether gender influences teaching quality, I analyzed the average ratings of male and female professors. First, I separated the data into two groups: ratings for male professors and ratings for female professors. I then calculated the average rating for each group, rounding the results to two decimal places. Next, I performed an independent t-test to compare the means of the two groups. This statistical test helped me evaluate whether the observed differences in the average ratings could have occurred by chance.

The results showed that male professors received an average rating of 3.91, slightly higher than the 3.86 average for female professors. The t-test yielded a t-statistic of 4.10 and a p-value of 4.12 × 10⁻⁵, which is well below the significant threshold of 0.005. This suggests that gender may play a role in the differences observed in professor ratings, although the practical significance (or size of the difference) is relatively small.



The boxplot above comparing the ratings of male and female professors highlights this difference. While there is some overlap in the interquartile ranges, the median rating for male professors is marginally higher. This visualization complements the statistical analysis and provides a clear picture of the distribution of ratings by gender.

# Question 2

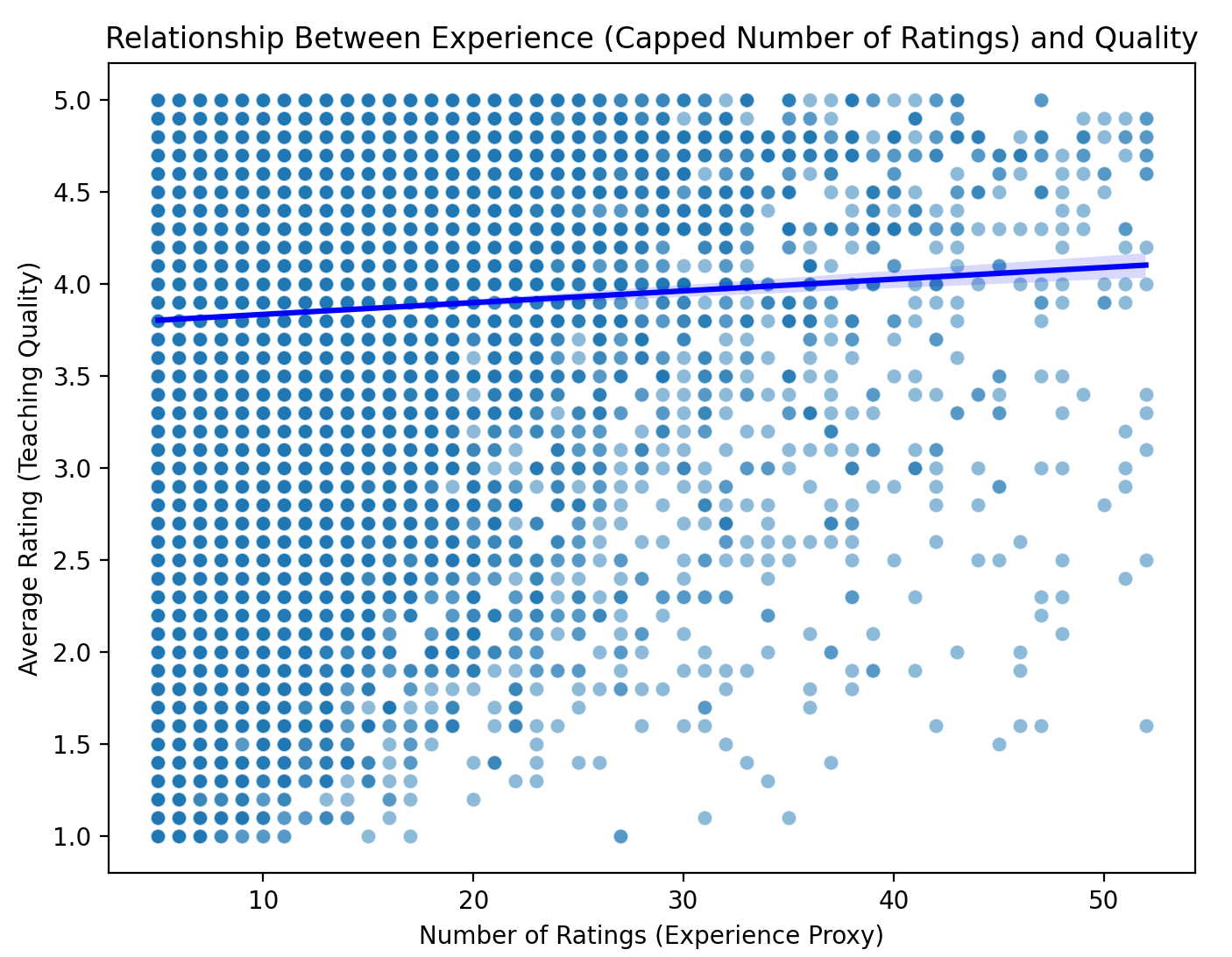
*Is there an effect of experience on the quality of teaching? You can operationalize quality with*

*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.*

To analyze the relationship between the number of ratings (as a proxy for experience) and average ratings (as a measure of teaching quality), I first cleaned the data by removing rows with missing values in the "Average Rating" and "Number of Ratings" columns. Then, I capped the number of ratings at the 99th percentile to prevent extreme values from skewing the results.

I calculated the correlation between the number of ratings and average rating to see if there was any linear relationship. Next, I built a linear regression model using the number of ratings as the independent variable (X) and average rating as the dependent variable (y). I added a constant term to account for the intercept, then fitted the model and examined its summary statistics. The correlation between the number of ratings and average rating was 0.06, indicating a weak positive relationship. The simple linear regression showed an R² value of 0.003, meaning that the number of ratings explains only a tiny fraction of the variation in teaching quality. However, the p-value is 0.000, which is well below the threshold of 0.005. This indicates that the number of ratings has a statistically significant effect on the average rating, even though the effect size is very small.



The scatter plot with the fitted regression line illustrates this relationship. While the line shows a slight upward slope, the variability in ratings at all levels of experience suggests that factors beyond the number of ratings play a greater role in determining teaching quality. The analysis indicates that experience has a small but statistically significant impact on average ratings.

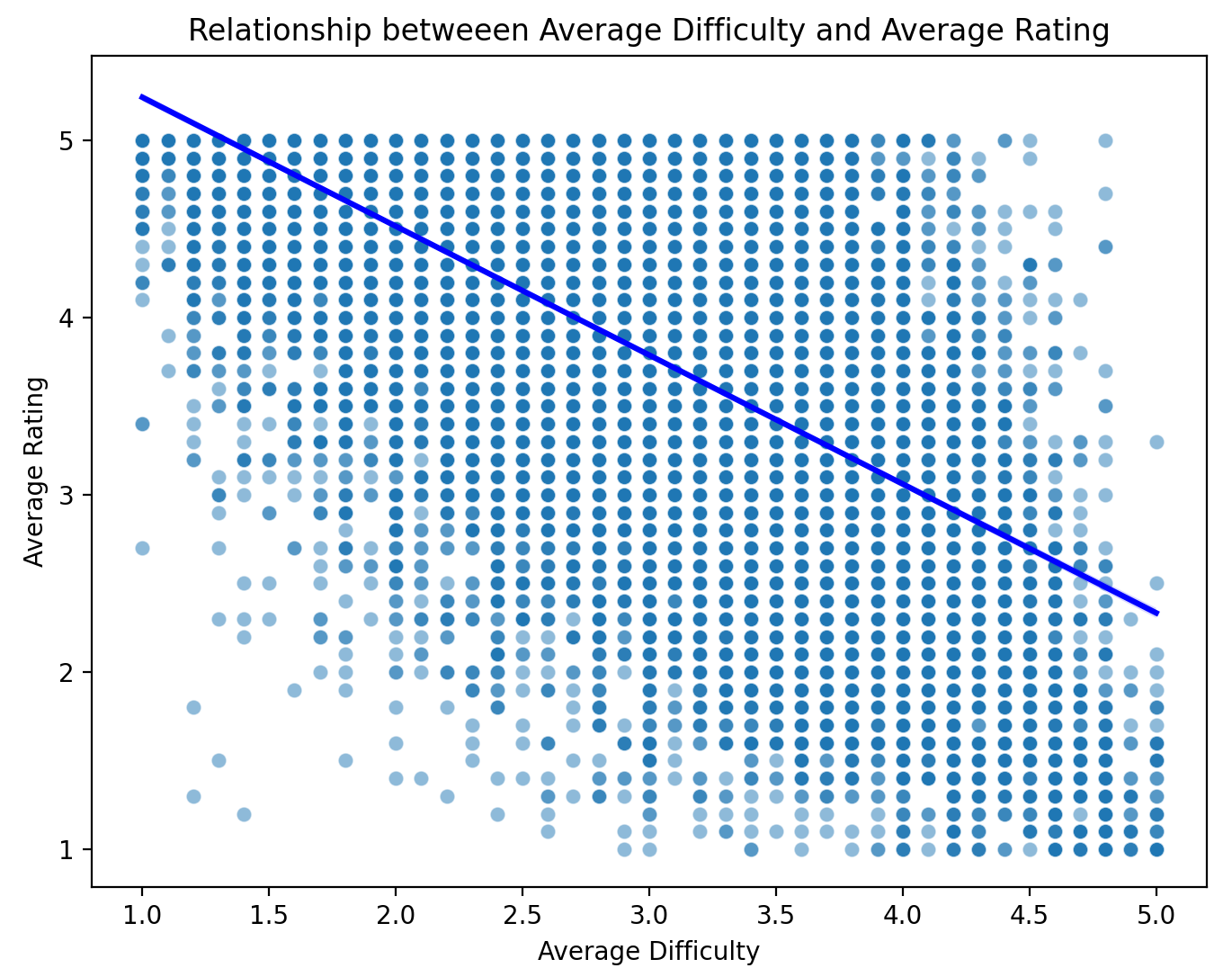
While the number of ratings has a statistically significant effect on the average rating, the effect size is negligible, and the number of ratings explains very little of the variation in average ratings. This suggests that other factors are likely more important in determining the average rating.

# Question 3

*What is the relationship between average rating and average difficulty?*

To determine the relationship between average rating and average difficulty, I first cleaned the data by removing rows with missing values in either the "Average Rating" or "Average Difficulty" columns. Next, I calculated the correlation between "Average Rating" and "Average Difficulty" to measure the strength and direction of their linear relationship. After that, I built a linear regression model using "Average Difficulty" as the independent variable (X) and "Average Rating" as the dependent variable (y). I added a constant term to account for the intercept and fit the model to examine how difficulty impacts ratings.

I found a correlation of -0.62, indicating a strong negative relationship between difficulty and ratings. The regression model corroborated this result, with a significant negative coefficient for Average Difficulty (β = -0.72, p < 0.005), suggesting that higher difficulty levels are associated with lower ratings. The R² value of 0.39 indicates that 39% of the variance in Average Rating can be attributed to Average Difficulty.



The scatter plot with the regression line above illustrates this relationship, showing a downward slope that reflects the negative association. The data points are closely grouped around the line, highlighting a clear pattern: professors perceived as more difficult tend to receive lower ratings.

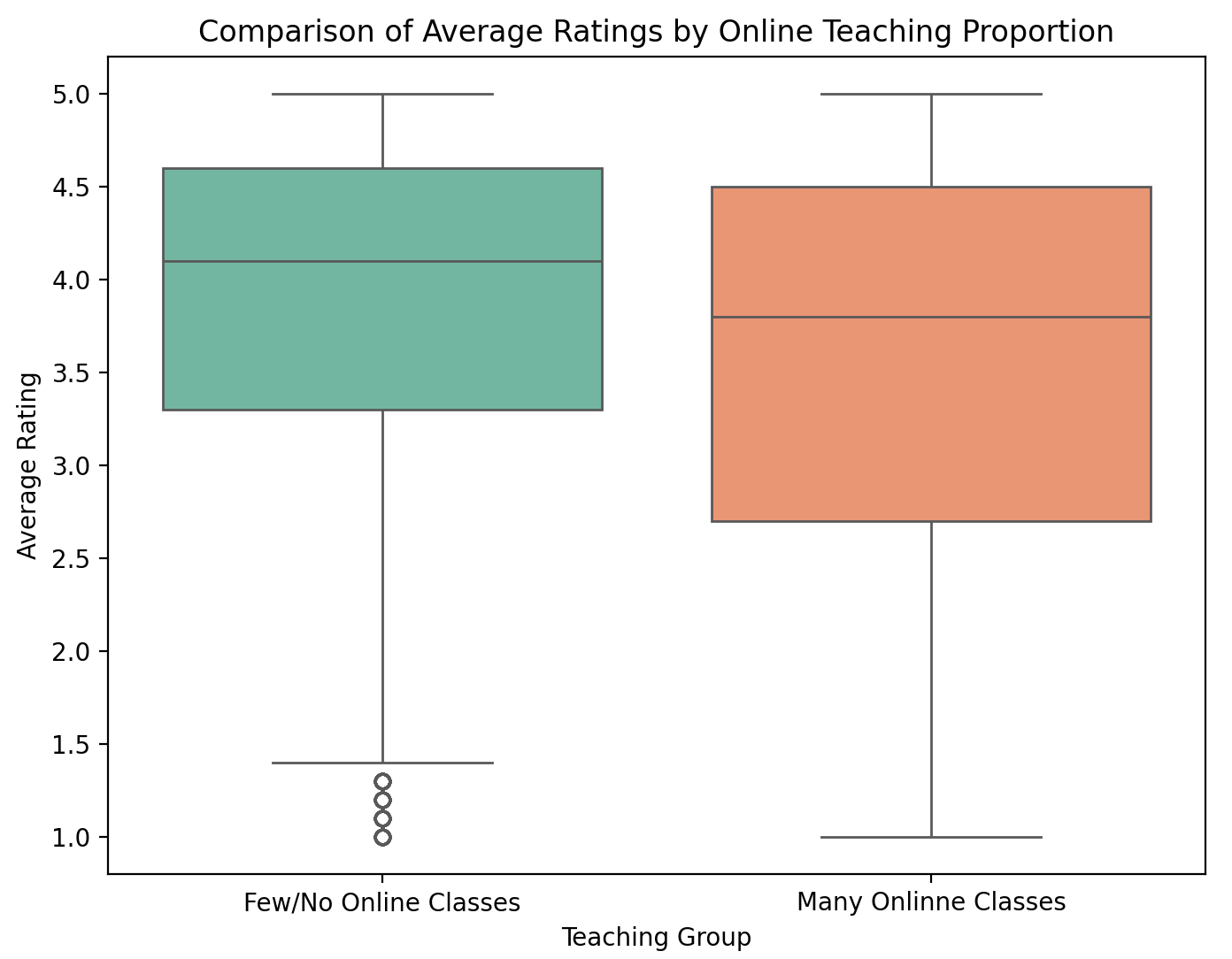
Hence, professors perceived as more difficult tend to receive lower average ratings. This finding highlights the impact of perceived challenge on student evaluations, showing that difficulty can substantially influence how students rate teaching quality.

# 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.*

To evaluate whether teaching online affects ratings, I calculated the proportion of a professor's ratings from online classes. First, I defined a threshold of 50% to separate the two groups: professors who teach more than 50% of their classes online were categorized as teaching “a lot of online classes,” while the rest were categorized as teaching “few or no online classes.” I calculated the proportion of online ratings for each professor by dividing the number of online ratings by the total number of ratings. I then split the data into two groups based on the defined threshold: one group for professors teaching many online classes and another for those teaching few or no online classes. Finally, I created a new column in the dataset to label each professor as belonging to the "Many Online Classes" or "Few/No Online Classes" group.

For each group, I calculated the average rating to compare their means. I performed an independent t-test to check whether the difference in average ratings between the two groups was statistically significant. I found that professors who teach many online classes have an average rating of 3.58, compared to 3.85 for those who mainly teach in-person. The independent t-test confirmed this difference is statistically significant (t = -7.43, p < 0.005). This means that professors who primarily teach online tend to receive lower average ratings.



The boxplot above compares the two groups and shows the difference in ratings. The median rating for professors teaching many online classes is visibly lower. The plot highlights that online teaching tends to receive lower ratings than in-person classes, which aligns with what I found.

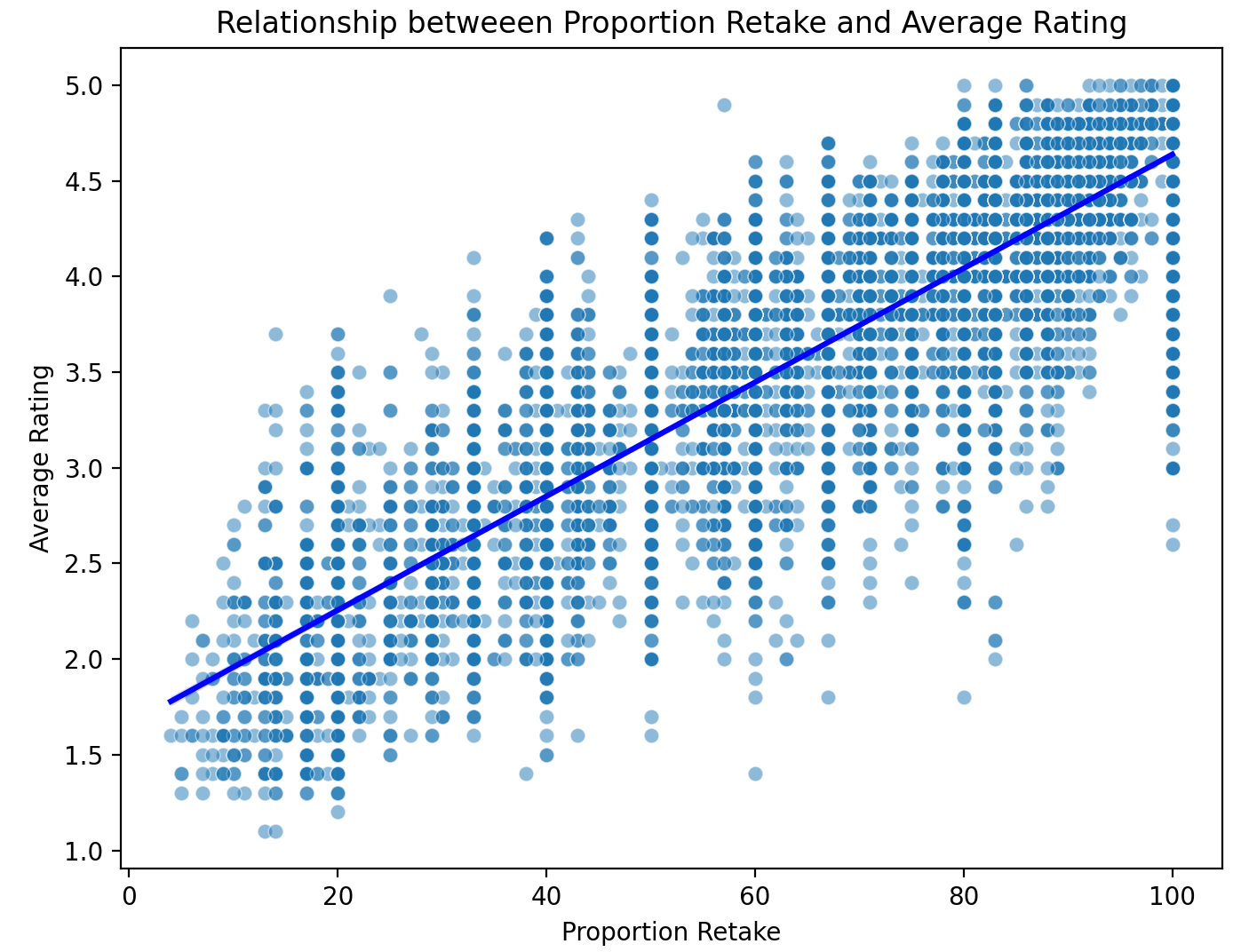
Therefore, professors teaching many online classes tend to receive significantly lower average ratings compared to those teaching few or no online classes. This suggests that the format of teaching (in-person vs. online) may influence how students perceive and rate their professors.

# Question 5

*What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?*

I looked at how Average Rating relates to Proportion retake, which is the percentage of students who would retake a professor's class. After ensuring that there were no missing values in both variables, I calculated the correlation coefficient between "Average Rating" and "Proportion Retake" to measure the strength and direction of their linear relationship. This provided an initial understanding of how these two variables are related. After that, I built a linear regression model, with "Proportion Retake" as the independent variable (X) and "Average Rating" as the dependent variable (y). I added a constant term to account for the intercept and then fit the model to examine how the proportion of retakes impacts the average rating.

I found a strong positive correlation of 0.88, meaning that as more students express willingness to retake a professor's class, the professor's average rating tends to increase. The regression analysis confirmed this finding, with a positive correlation (β = 0.0298, p < 0.005). The model explained a substantial 77.5% of the variation in the average ratings (R² = 0.775), showing a strong link between the two variables.



The scatter plot with a regression line above demonstrates the relationship between the two variables. The upward-sloping line reflects the positive association, with the data points closely following the trend, reinforcing the strong correlation and predictive power.

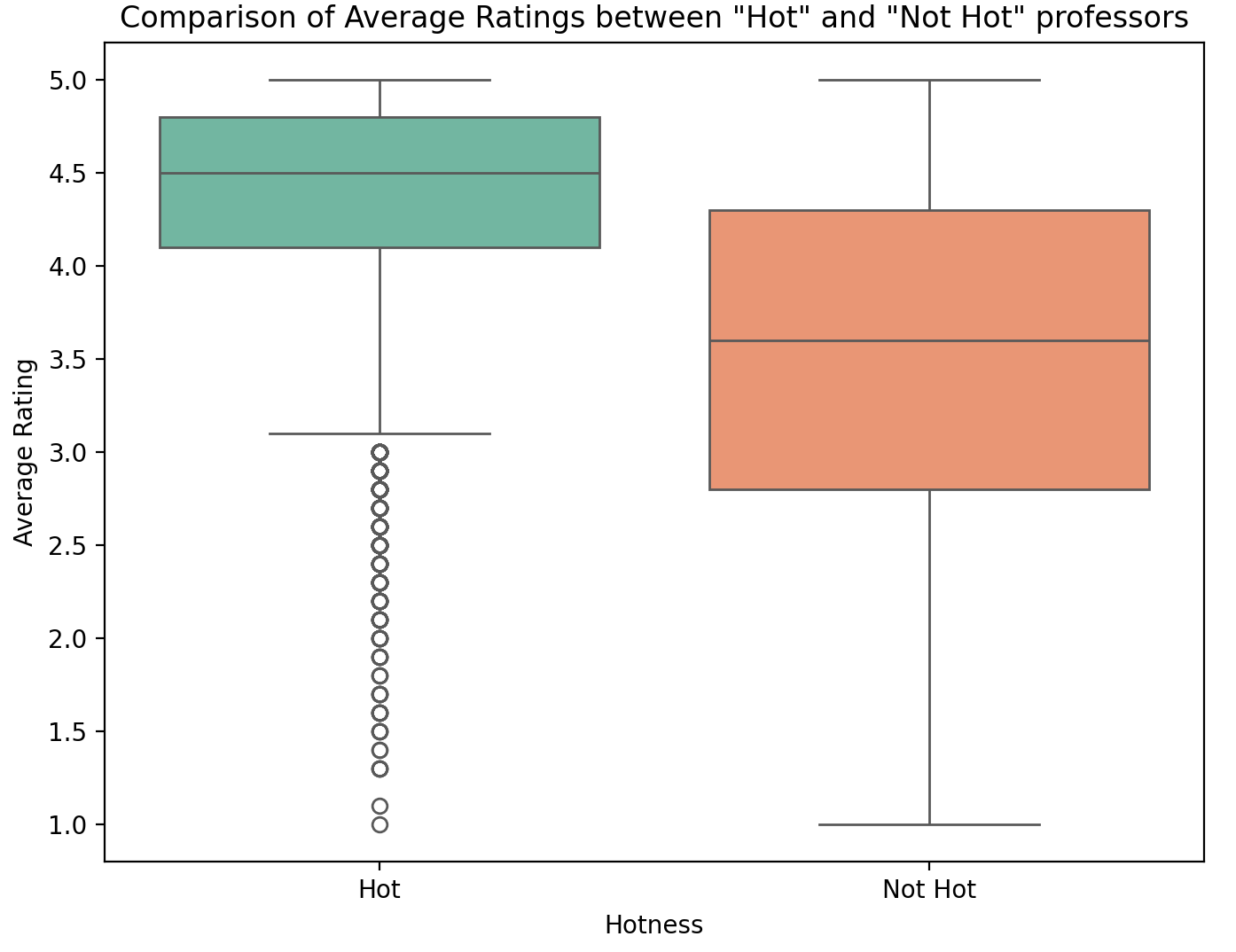
There is a very strong and statistically significant positive relationship between "Proportion Retake" and "Average Rating." Professors with a higher proportion of students willing to retake their class tend to receive significantly better ratings, suggesting that willingness to retake a class is a strong indicator of overall teaching quality as perceived by students.

# Question 6

*Do professors who are “hot” receive higher ratings than those who are not? Again, a significance test is indicated.*

To assess whether being labeled "hot" correlates with higher ratings, I separated the data into two groups: one for professors who received a pepper icon (hot teachers) and another for professors who did not (not-hot teachers). I calculated the average rating for each group to compare their means. Next, I created a new column in the dataset to label each professor as either "Hot" or "Not Hot" based on whether they received a pepper icon. Then, I performed an independent t-test to determine whether the difference in average ratings between the two groups was statistically significant.

I found that professors labeled "hot" had an average rating of 4.36, significantly higher than the 3.47 average rating for those not labeled "hot." The t-test result (t = 90.32, p < 0.0005) confirmed that this difference is statistically significant. This suggests that professors considered "hot" receive substantially higher ratings on average.



The boxplot above demonstrates the visual comparison of the two groups, showing that "hot" professors have a higher median rating and a narrower interquartile range, indicating less variability in their ratings. On the other hand, the distribution for "not hot" professors is broader, with lower average ratings, supporting my findings.

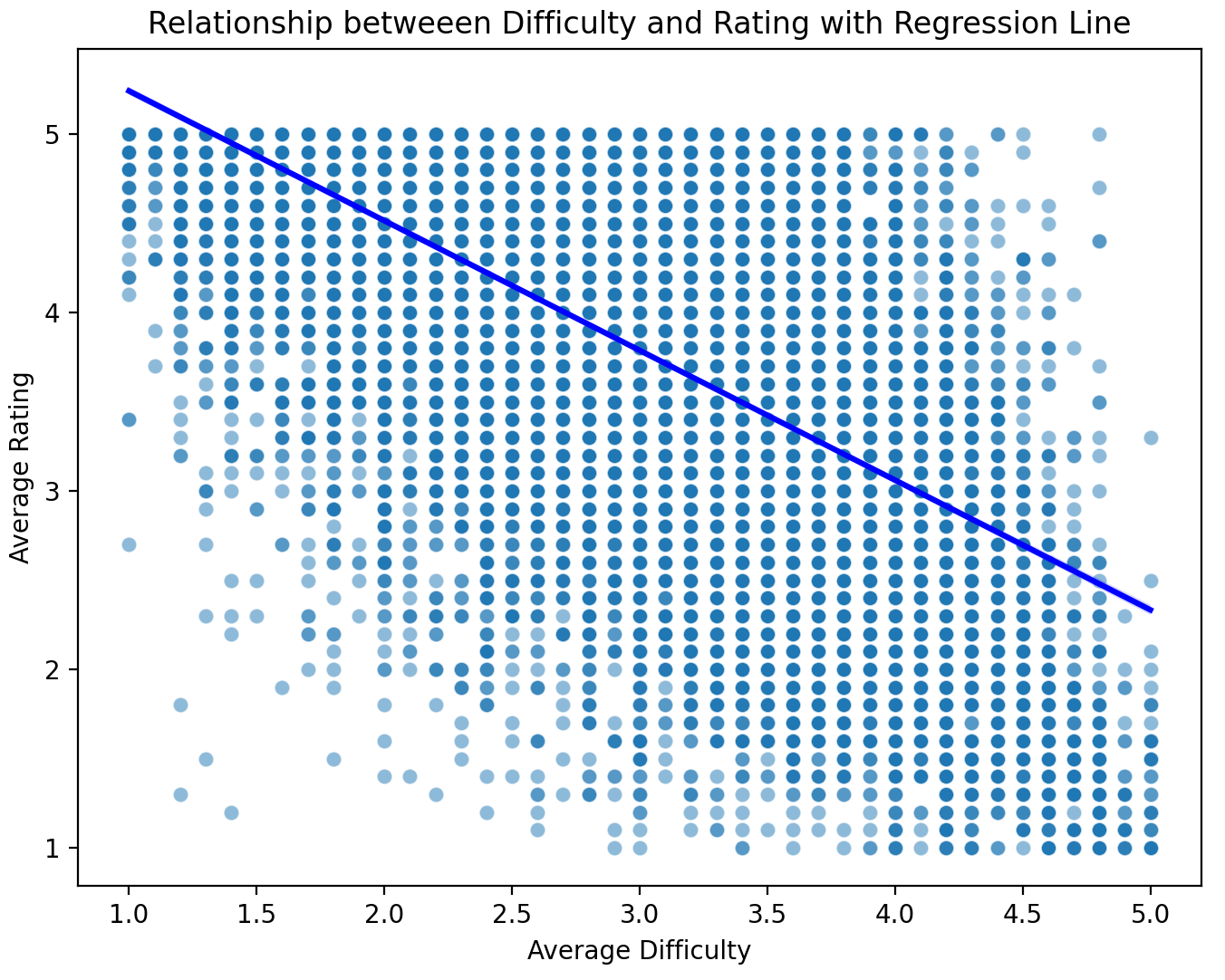
So, professors who are perceived as "Hot" receive significantly higher average ratings compared to those who are not. This suggests that perceived physical attractiveness, as indicated by the pepper icon, may have a notable impact on how students rate professors.

# Question 7

*Build a regression model predicting average rating from difficulty (only). Make sure to include the R2 and RMSE of this model.*

To understand the relationship between Average Difficulty and Average Rating, I built a simple linear regression model using Average Difficulty as the sole predictor. I added a constant term to account for the intercept and fit the model to find the best-fitting line that describes the relationship between these variables. Then, I used the model to make predictions of average ratings based on the difficulty ratings.

I found a significant negative relationship between Average Difficulty and Average Rating, with a regression coefficient of β = −0.7271 (p < 0.005), meaning ratings tend to decrease as perceived difficulty increases. The model's R² value was 0.383, indicating that Average Difficulty explains 38.3% of the variance in ratings. However, the Root Mean Square Error (RMSE) was 0.74, showing some predictive error and suggesting that additional factors influence ratings.



The scatter plot with the regression line above illustrates the negative trend, with higher difficulty corresponding to lower ratings. While the data points are mainly close to the line, the spread reveals some variability not explained by difficulty alone, underscoring the model's limitations in capturing the full complexity of rating behavior.\

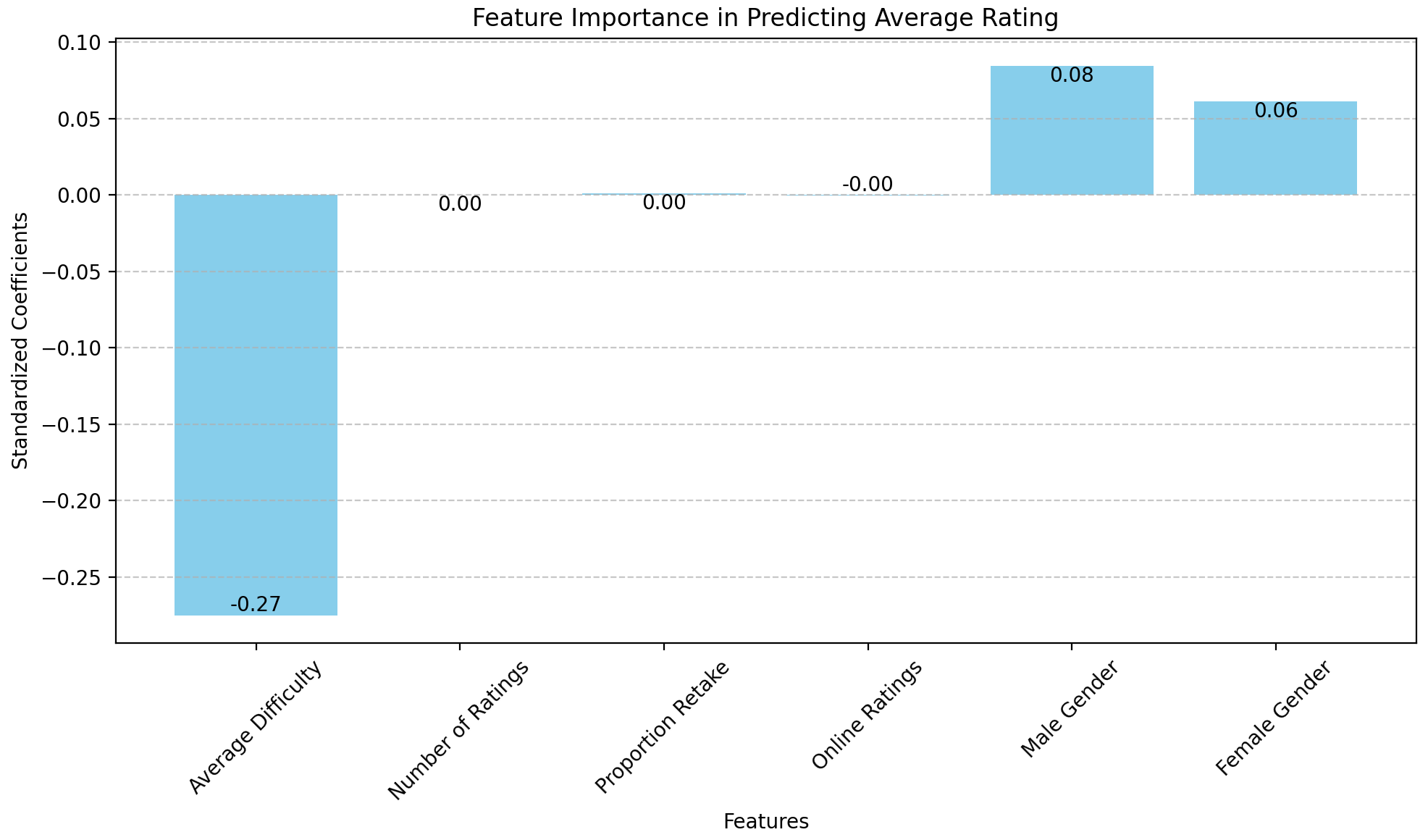
There is a statistically significant negative relationship between average difficulty and average rating. Professors perceived as more difficult tend to receive lower ratings, which aligns with the idea that students may evaluate professors less favorably if they find the class challenging.

# 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.*

To better understand what influences the Average Rating, I built a multiple linear regression model that incorporates all available predictors: Average Difficulty, Number of Ratings, Proportion Retake, Online Ratings, Male Gender, and Female Gender. This model aims to show how these variables collectively impact ratings and to compare its performance against the simpler "difficulty-only" model.

The complete model explained 79.8% of the variance in Average Rating, shown by an R² of 0.798, a substantial improvement from the 38.3% explained by the single-factor model. It also had a lower RMSE of 0.38, meaning it predicted ratings more accurately than the "difficulty-only" mode, which had an RMSE of 0.74. Among the predictors, Average Difficulty remained negatively associated with ratings (β = −0.203, p < 0.005), but its effect was less pronounced when looking at multiple factors. Proportion Retake had a strong positive impact (β = 0.0266, p < 0.005), indicating that professors whose students are more willing to retake their classes tend to receive higher ratings. Gender also played a small but significant role (Male Gender: β = 0.0419, p < 0.005, and Female Gender: β = 0.0293, p < 0.005). Interestingly, the Number of Ratings and Online Ratings did not significantly impact the results when considering all factors, suggesting their effects are minimal. The Variance Inflation Factors (VIFs) were all below 2, showing no multicollinearity issues among the predictors. The comprehensive model significantly improved the prediction of teaching ratings compared to the "difficulty-only" model, as indicated by the higher R² and lower RMSE values. This shows the importance of including multiple factors to predict teaching ratings better.



The bar chart above shows how much each factor influences the results. Average Difficulty has the most substantial negative effect, while Proportion Retake has the strongest positive effect. Gender-related variables also help, but to a lesser degree, while the other factors have minimal impact.

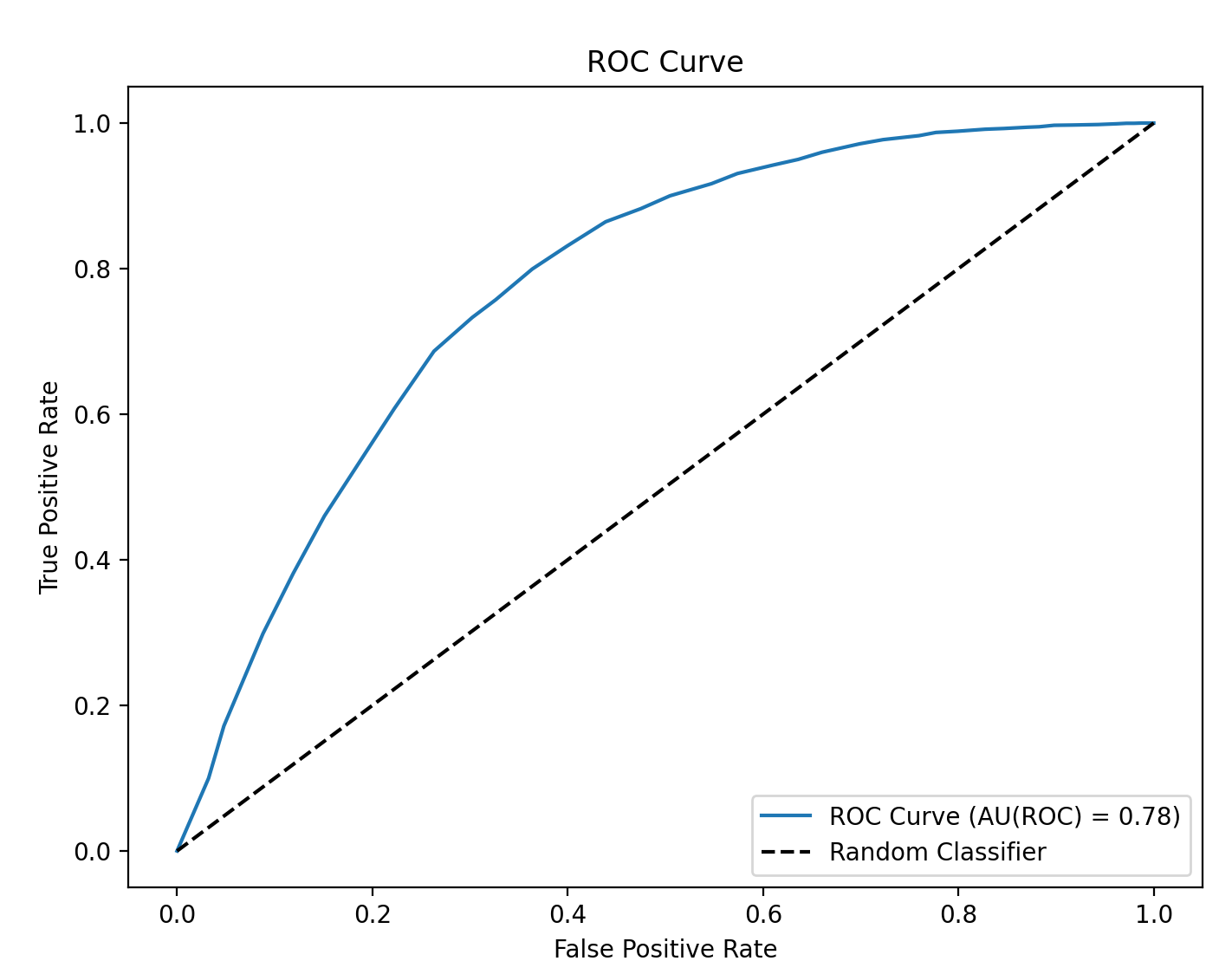
# Question 9

*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.*

I built a classification model to predict whether a professor receives a "pepper" based only on their Average Rating. One issue I faced was the class imbalance since fewer professors were labeled with a pepper than those without. To fix this, I used upsampling to balance the classes. I randomly resampled the minority class (professors who received a pepper icon) to match the size of the majority class (those who didn’t), ensuring the model would not be biased toward the majority class. I combined the balanced classes into a single dataset. The predictor variable (X) was Average Rating, and the target variable (y) was Received Pepper (1 for "hot," 0 for "not hot"). I then split the data into training (80%) and testing (20%) sets, stratifying by the target variable to preserve class distribution. I trained a logistic regression model on the training data to learn the relationship between average ratings and the likelihood of receiving a pepper icon.

The model scored 0.78 on the AU(ROC), showing a strong ability to distinguish between professors with and without peppers. The overall accuracy was 72%, with higher precision for the major class (professors without pepper) and higher recall for the minority class (professors with peppers). The F1 scores were close: 0.74 for professors with peppers and 0.69 for those without, indicating balanced performance.



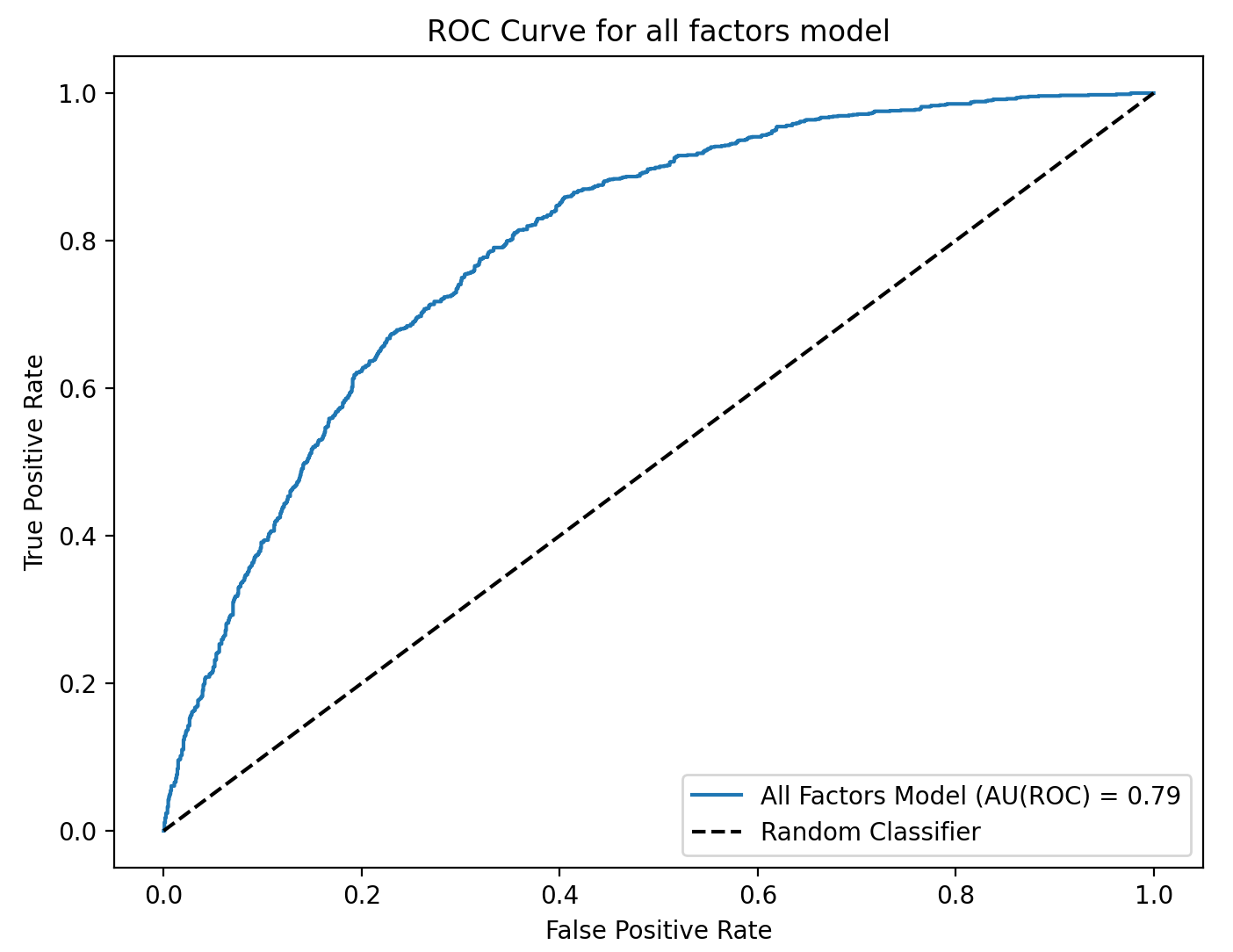
The ROC curve above visualizes the model's performance across thresholds. The curve, which rises well above the diagonal random classifier line, confirms the model's effectiveness in predicting the presence of a pepper. The AU(ROC) score is 0.78, demonstrating good discriminatory power, where a score of 1.0 would represent perfect classification, and 0.5 means no predictive capability. Balancing the dataset through upsampling prevented the model from skewing heavily toward the majority class and improved its ability to identify professors in the minority class.

# 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(ROC) and also address class imbalances.*

I built a logistic regression model to predict whether a professor received a "pepper" icon ("hot") using multiple factors: Average Rating, Average Difficulty, Number of Ratings, Proportion Retake, Online Ratings, Male Gender, and Female Gender. Like in my earlier model, I faced a challenge with class imbalance. To solve this, I used upsampling again to balance the dataset, ensuring both classes' fair treatment. I selected the features (X) and target variable (y) from the balanced dataset. I then split the data into training (80%) and testing (20%) sets, ensuring the class distribution was preserved using stratification. I trained a logistic regression model using the training data, setting the maximum iterations to 1000 to ensure convergence. This model learned the relationships between the features and the likelihood of a professor being "hot."

This enhanced model achieved an AU(ROC) score of 0.79, a slight improvement over the "Average Rating Only" model, which scored 0.78. This higher AU(ROC) score shows that adding more predictors slightly enhanced the model's ability to distinguish between professors with and without peppers. The overall accuracy remained at 72%, with balanced precision and recall for both classes. The F1-scores were 0.74 for professors with peppers and 0.70 for those without, showing strong performance for both groups.

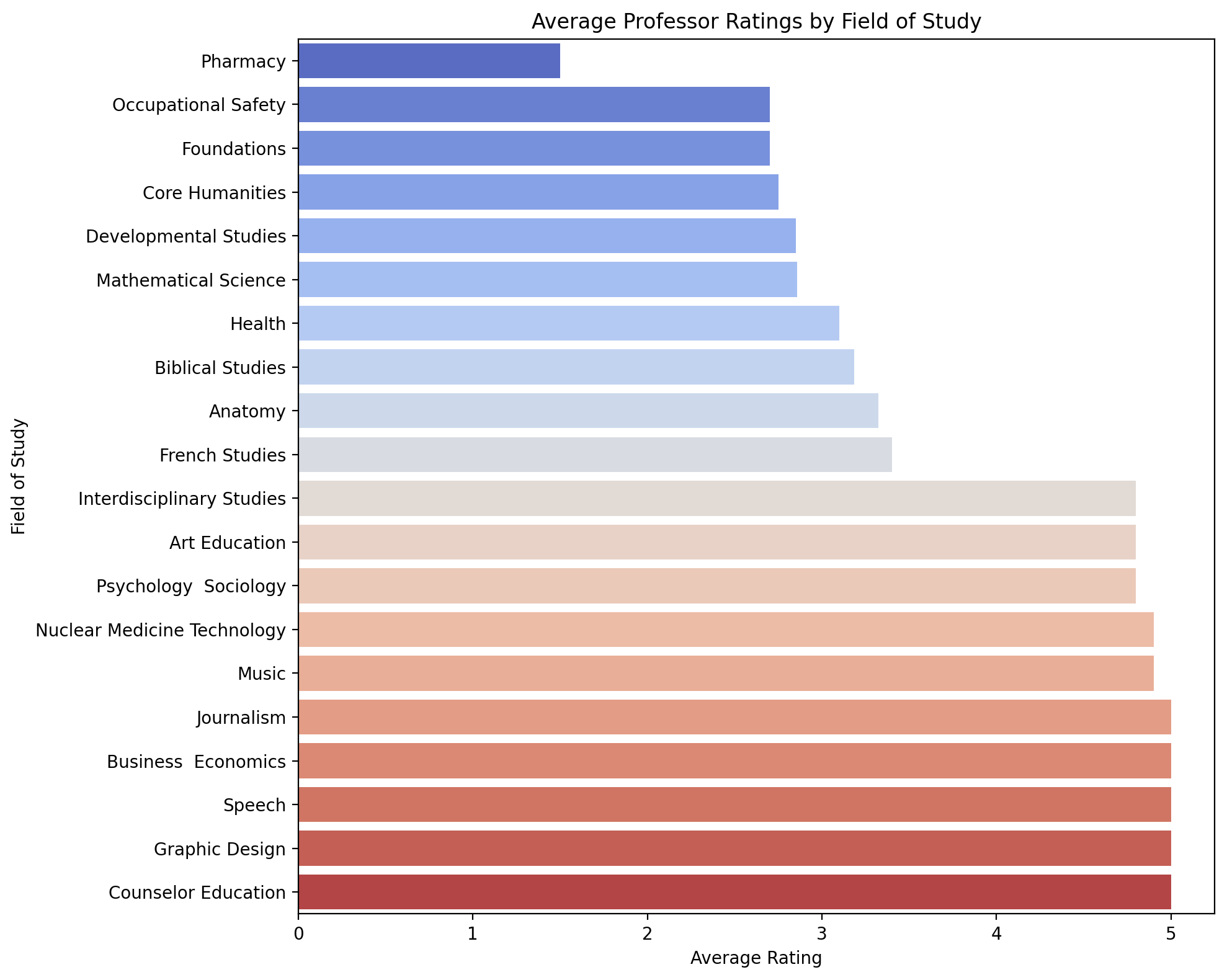


The ROC curve above illustrates the model's performance, showing a clear separation from the random classifier line, aligning with the strong AU(ROC) score. While the additional factors provided slight gains in predictive power, the primary influence remained the Average Rating, which was evident from the slight increase in the AU(ROC) score and the similar accuracy to the simpler model.

# Extra Credit

*Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the qualitative data, e.g. major, university or state by linking the two data files]*

I decided to analyze how average professor ratings vary across different fields of study. First, I grouped the data by university and calculated the average professor rating for each university. At the same time, I identified the most common major or field of study for each university by finding the mode (most frequently occurring major). Next, I grouped the data by field of study (major) and calculated the average professor rating for each field. This allowed me to see which academic disciplines tend to have the highest and lowest-rated professors on average. To focus on the extremes, I selected the top 10 fields with the highest average ratings and the bottom 10 fields with the lowest average ratings. I combined these into a single dataset for comparison.



I found that professors in Pharmacy, Occupational Safety, and Foundations received the lowest ratings from students. This might be due to challenging coursework, strict grading standards, or other factors that affect how students view these professors. On the other hand, professors in Counselor Education, Graphic Design, and Speech received the highest ratings. These fields may involve more creative work, smaller classes, or greater student involvement, leading to better evaluations. The difference between the highest and lowest-rated fields is significant, with the lowest ratings near 0 and the highest ratings close to 5. The bar plot clearly shows this difference, highlighting how ratings vary across disciplines. This suggests that the characteristics of specific fields, such as difficulty, grading practices, and how students interact with professors, greatly influence student evaluations.