Luke Sullivan
Pascal
Intro to Data Science
9 December, 2024

Assessing Professor Effectiveness
*Capstone Report*

**Data Preprocessing & Cleaning**

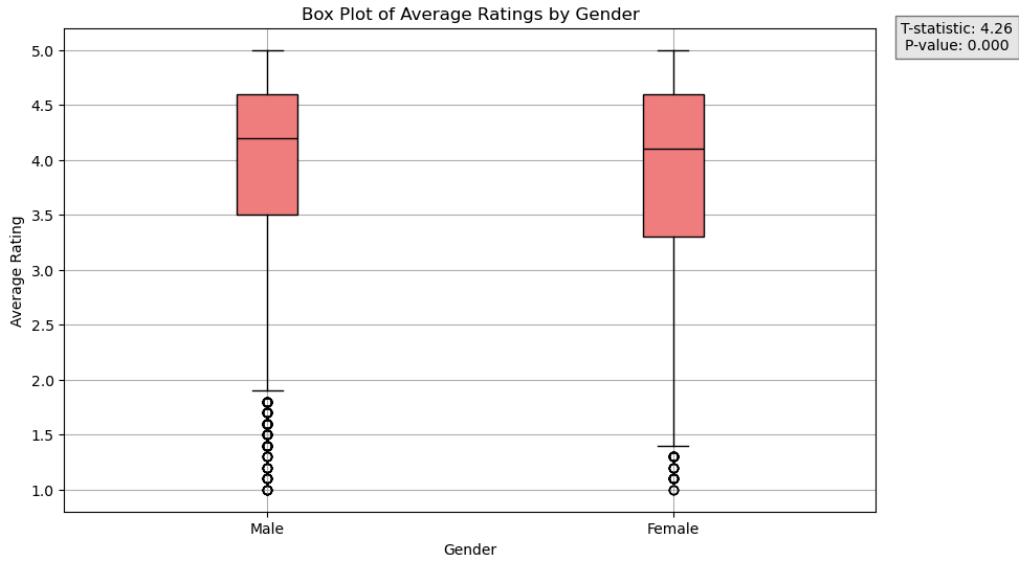
Before conducting any analysis, several preprocessing and data cleaning steps were undertaken to ensure the dataset is ready for analysis. Following the necessary imports, the numpy random seed was seeded to the NYU ID from "Luke Sullivan" to ensure reproducibility and ensure integrity. Following the prerequisite steps, The three datasets provided were loaded and columns transformed for better unification and readability. Then, The datasets were merged based on their shared record info between the professors and their data. Any missing data was addressed by removing 19,889 rows containing entirely missing values, which were assumed to be Missing Completely at Random (MCAR). Any tags left over that had missing values but other professor information, were replaced with zeroes, assuming they did not receive any tags. The Retake column was dropped due to over 86% of its data being missing from the dataset. To normalize the tag data throughout the entire dataset, row-wise normalization was performed, ensuring balance between the different professors and the number of tags received. Finally, all professor data with less than 7 total student ratings were dropped from the analysis, due to the interpretation of the CDF of the total number of ratings received. These preprocessing and data cleaning steps ensure that only professors with sufficient data were included in the analysis.

**1. Questions**

**1.1 Is there evidence of a pro-male gender bias in the dataset?**

Approaching this question, the analysis assumes that the data in both the male and female ratings are independent samples that do not influence each other. The data gives the *Average Rating* for each professor, indicating that the data can be reduced down to its sample mean and interpreted through a parametric test. This analysis also assumes that the variances are not equal, suggesting that male and female ratings are rated differently, whether through gender bias, social expectations, and or there is a strong negative bias towards the ratings in women. According to the data and these assumptions, a Welch's t-test between the ratings for the male and female professors would be the most sufficient, to account for the unequal variance. After executing the test, the t-statistic was calculated as 4.258, and the p-value was 2.083 x 10^-5. Examining the results in **Figure 1.1.1**, The distributions of the ratings between the two genders is relatively similar, indicating that the effect size of this p-value may be non practical. However under these assumptions, the p-value is much smaller than the alpha level (a=0.005), we would reject the null hypothesis and conclude that there is strong statistical evidence of a gender bias towards male ratings, through the positive t-statistic.

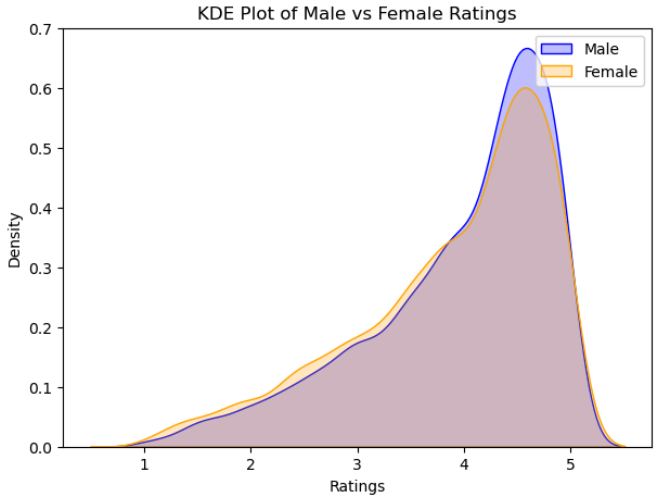
**Figure 1.1.1 Average Ratings Distributions Box Plot**




**1.2 Is there a gender difference in the spread of the ratings distribution?**

The analysis assumes that the ratings for male and female professors are independent samples, meaning that the variance in one gender group does not influence the other. Under these assumptions, the appropriate statistical approach of Levene's Test will be used to compare the variances between the male and female ratings to determine if there is a spread difference in rating distributions. The variance for male professors was calculated as 0.753, and the variance for female professors was 0.833, visualized in **Figure 1.2.1**, through a Kernel Density Estimation (KDE) plot. Levene's test produced a test statistic of 17.733 and a p-value of 2.563×10 ^−5. Given that the p-value is far below the alpha level, we conclude that there is strong evidence of a difference in the spread of the ratings distributions between male and female professors. However, this conclusion assumes that Levene's test assumptions are met, and potential limitations include biases in data collection or untested assumptions about data structure.

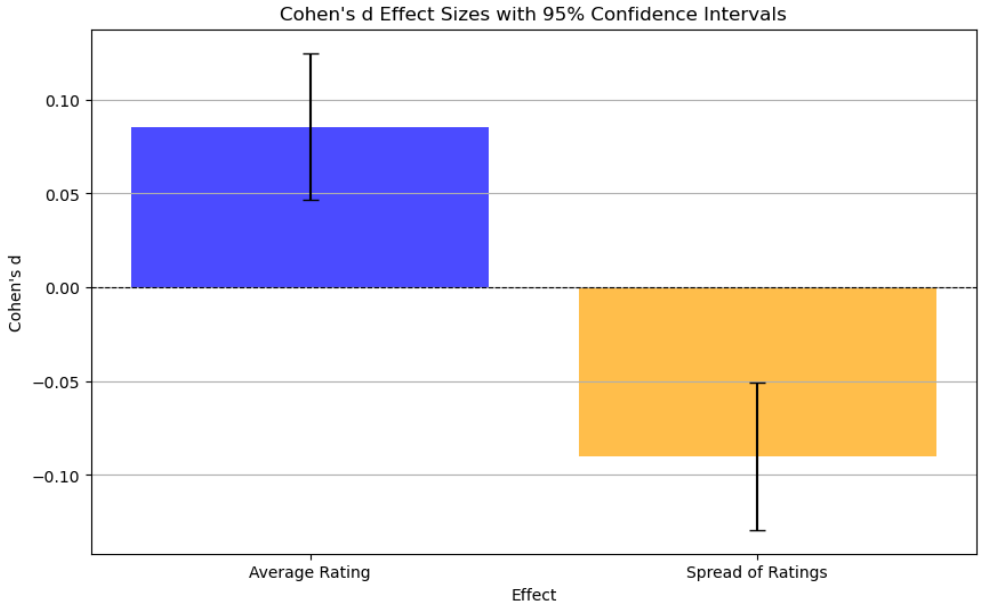
**Figure 1.2.1: KDE Plot of Male vs Female Ratings**




**1.3 What is the likely size of both of these effects?**

This analysis assumes that the ratings for male and female professors are independent, and do not influence each other. Cohen's d was calculated to measure the effect size for both the gender bias in average ratings and the gender bias in the spread (variance) of ratings. For average ratings, the mean difference between male and female ratings was standardized by dividing it by the pooled standard deviation, which accounts for variability across both groups. For variance, the difference between male and female variances was standardized using the square root of the pooled variance. This allows for a meaningful comparison of effect sizes by expressing differences in terms of standard deviations, making them independent of the original units of measurement. The pooled standard deviation was used because it combines variability from both groups while accounting for their sample sizes, providing a balanced measure of overall variability. This was particularly important due to the difference in sample sizes and variances between the two groups of professors. Resulting from the data seen in **Figure 1.3.1**, Female professors tend to receive slightly lower average ratings than male professors, with an estimated effect size ranging from −0.1248 to −0.0464, and show slightly greater variability compared to male professors, with an estimated effect size ranging from −0.1292 to −0.0508. This suggests that there is a pro-male gender bias in the dataset, but the effect size is so small that it would not be practically significant. However, these results might suggest why there is prominence to believe that there is a pro-male gender bias due to the higher variability in the female ratings.

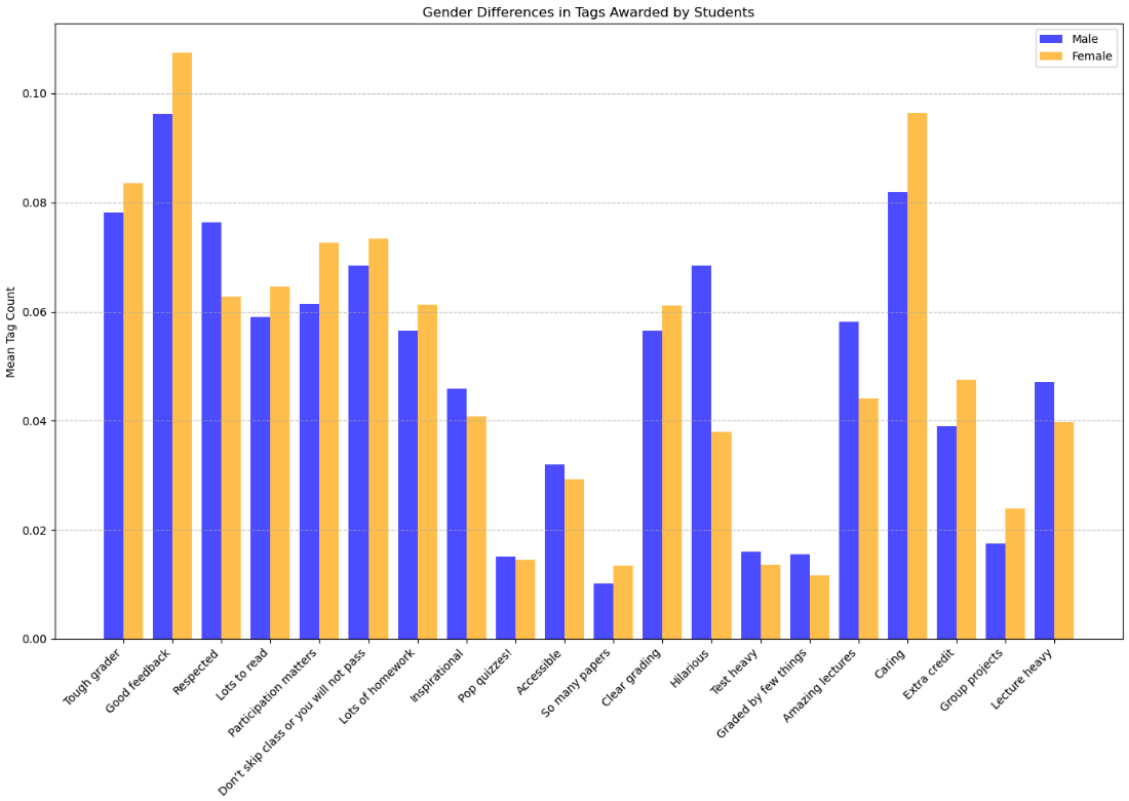
**Figure 1.3.1: Cohen's d Effect Size Box Plot**




**1.4 Is there a gender difference in the tags awarded by students?**

Due to the nature of the real-world dataset of the professor ratings, the data contains unequal variances and sample sizes between the gender groups. Assumptions include independence of observations and the data is representative of the population being studied. Under these claims, Welch's t-test was used to understand if there is a gender difference between tags awarded by students. This is because it is robust to differences in variances and sample sizes, which are likely to occur when comparing male and female professors. To reduce the risk of false positives, Bonferroni correction was applied to the p-values to account for multiple statistical comparisons. According to the p-value and Cohen's d results examined in

**Figure 1.4.1**, Male professors receive Hilarious, Amazing lectures, and Respected significantly more than females with a high t-statistic and significant adjusted p-value. While, the least gendered tags include, "Lots of homework", "Tough Grader" and "Pop quizzes!". These findings suggest that student perceptions or biases may influence how certain attributes are attributed based on gender, but not all tags exhibit these differences. Examining the effect size provided by Cohen's d, the gendered tags have some practical meaning as the "hilarious" tag has a medium effect, while "Amazing lectures" and "Respected" have small effects. This suggests that there is a gender difference in the tags awarded by the students, however it is important to understand the analysis assumptions and that Bonferroni correction could increase the likelihood of false negatives in the results, masking small but real effects for less prominent tags.

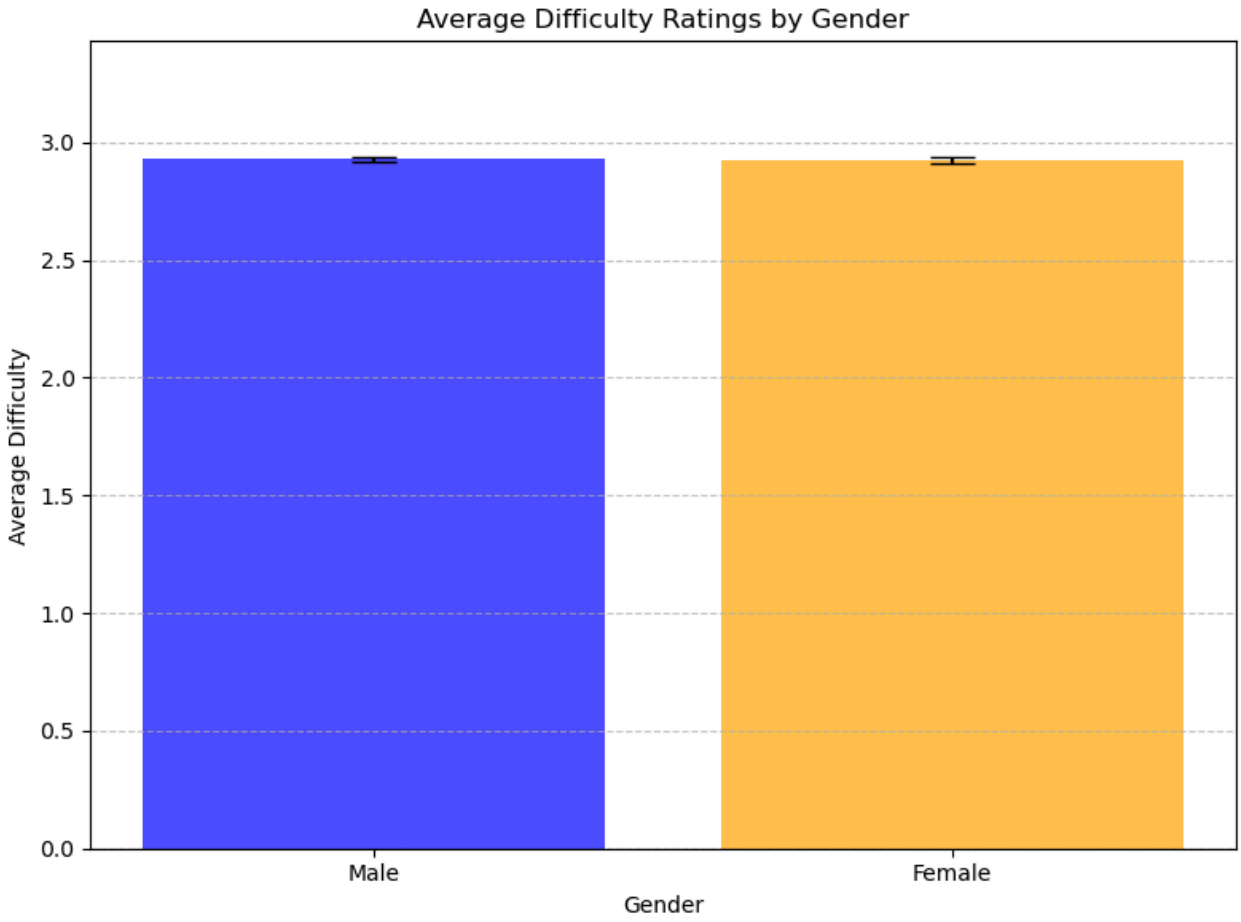



**1.5 Is there a gender difference in terms of average difficulty?**

Examining if there is a gender difference in terms of average difficulty includes the assumptions independence of difficulty ratings provided and that the data was interpreted through its statistical mean. A Welch's t-test was used to compare the average difficulty ratings between male and female professors. This test was chosen because it accounts for the unequal variances and sample sizes in the dataset between the two groups. This approach ensures a statistically sound comparison while addressing potential variability in the data. After running the T-Test, the t-statistic was 0.4456, with a p-value of 0.6559. According to the alpha threshold of 0.005, these results would be statistically insignificant. Given that the p-value (0.6559) is far above the significance threshold (e.g., 0.005), there is no evidence to suggest a statistically significant difference in average difficulty ratings between male and female professors. Examining the box plot in **Figure 1.5.1**, the average difficulty distributions between the male and female professors are extremely similar, proving the insignificance. Potential limitations include the possible biases in how difficult ratings are assigned by the students, which could impact how these results

are interpreted in the real world. However, based on this analysis, it can be concluded that gender does not appear to influence average difficulty ratings in this dataset.

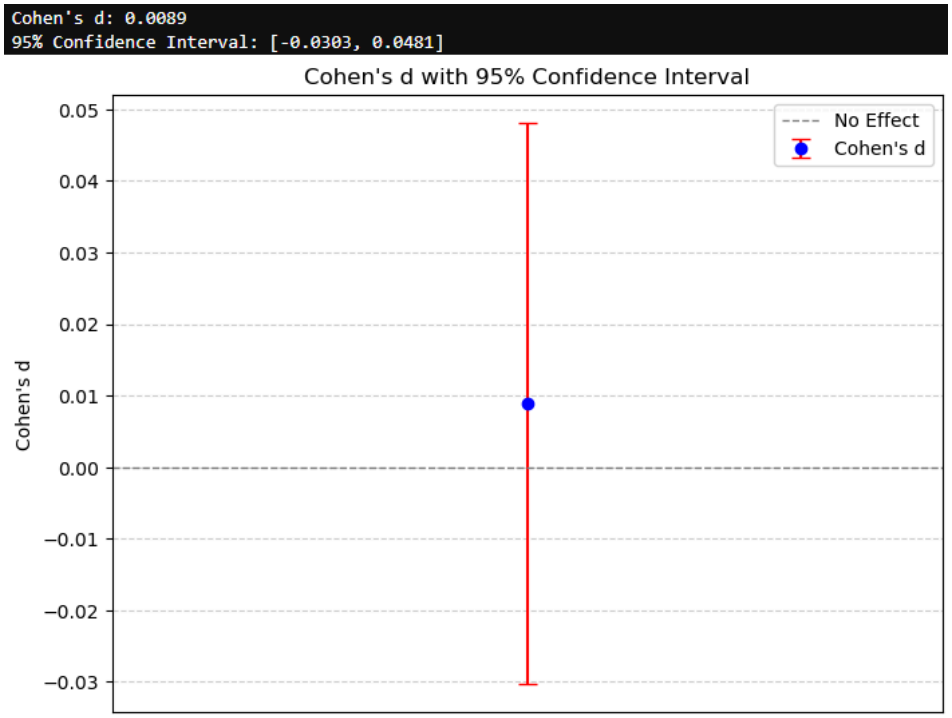
**Figure 1.5.1: Average Difficulty Ratings Box Plot**




**1.6 What is the size of that effect?**

Following Question 3.5, Cohen's d was calculated to measure the effect size of the gender difference in average difficulty ratings. The mean difference between male and female professors' difficulty ratings was standardized by dividing it by the pooled standard deviation, which accounts for variability across both groups. Assumptions include independence of observations and the pooled standard deviation accurately represents the overall variability in ratings. The calculated effect size was 0.0089, indicating an extremely small difference in average difficulty ratings between male and female professors. Since the p-value found in Question 3.5 was insignificant, this effect size accurately conveys that there is no gender difference in terms of average difficulty, in terms of the analysis assumptions.

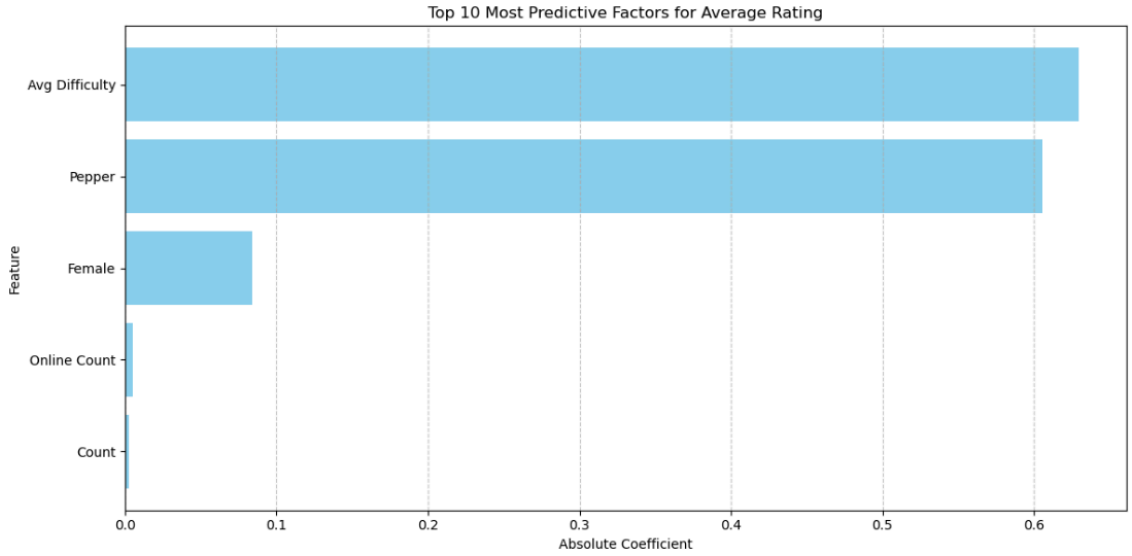
**Figure 1.6.1: Cohen’s d Effect Size**




**1.7 Average Rating Regression Model, which of these factors is most strongly predictive of average rating?**

To find the most strongly predictive factor of average ratings, a multiple linear regression model was used in this analysis. This model is used under the assumption of full independence of observations and the linear relationships between the target variable and its predictors. To address any multicollinearity concerns within the model, Variance Inflation Factors (VIFs) were calculated for each feature. VIFs quantify how much the variance of the regression coefficient is inflated due to multicollinearity. All predictors with a VIF value greater than 10 were removed, and the remaining predictors were used to train the model for regression. The model performed with an R^2 value of 0.5371, indicating that approximately 53.71% of the variance in average ratings is explained by the predictors. Also, the model received a RMSE score of 0.6040, suggesting that the model's predictions deviate from actual ratings by about 0.6 points. Shown in **Figure 1.7.1**, The most strongly predictive factor of average rating is *Avg Difficulty* ($\beta=-0.6296$), with higher difficulty ratings strongly associated with lower average ratings. The second most predictive factor is *Pepper* ($\beta=0.6053$), suggesting that perceived attractiveness may also play a significant role in students' evaluations.

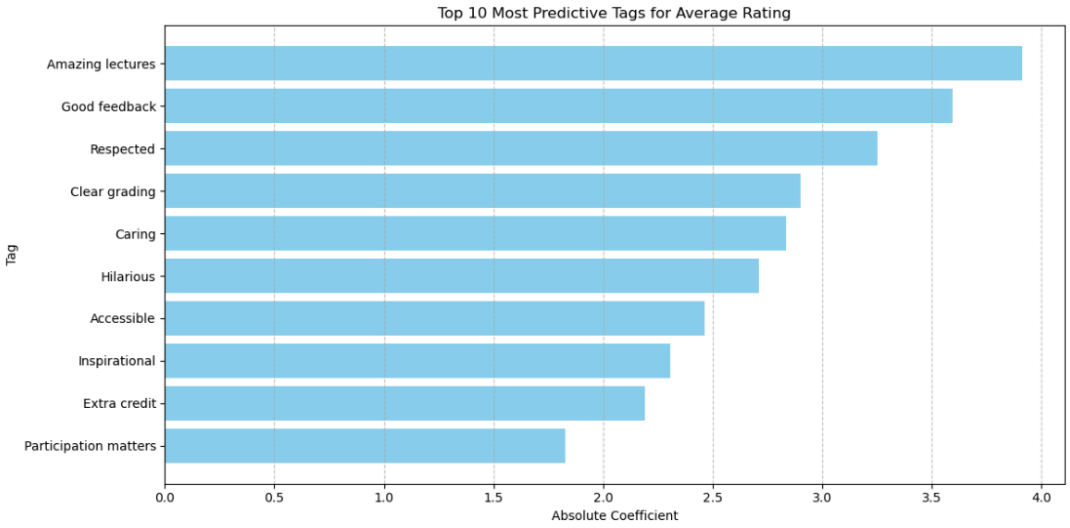
**Figure 1.7.1: Top Most predictive Factors Bar Chart**




**1.8 Average Rating Regression Model, which of these tags is most strongly predictive of average rating?**

To find the most strongly predictive tag of average ratings, another multiple linear regression model was used. The regression model was used under the assumption that there is independence between the ratings data and the linear relationships. Again, to address multicollinearity concerns within the mode, VIFs were calculated and any VIF predictor with a value greater than 10 was extracted from the model. Using the remaining predictors to train the regression model, it performed with an R^2 rating of 0.7696 and also an RMSE score of 0.4261. Comparing these statistics to the regression model in Question 3.7 (Received a R^2 score of 0.5371 and RMSE of 0.604), the tag-based model outperforms the numerical factors model both in terms of its variance and its predictive accuracy. While both models highlight important factors that influence the ratings, the tags-based model is more practical and can provide more actionable insights into specific tags. Shown in **Figure 1.8.1,** The most predictive tags for average professor ratings from the model include Amazing lectures (β=3.9114), Good feedback (β=3.5959), and Respected (β=3.2522). These tags suggest that students place significant value into how professors teach the course material. However, there are potential limitations including potential biases in how students assign tags.

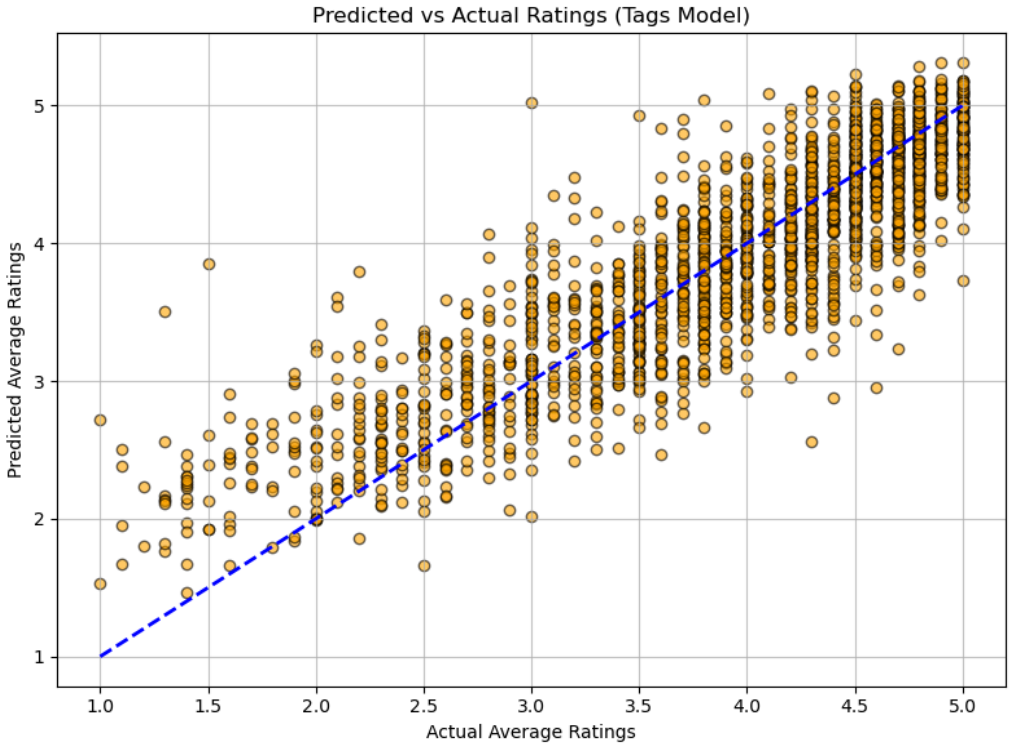
**Figure 1.8.1: Top 10 Most Predictive Tags for Average Rating**



**1.9 Average Difficulty Regression Model, Which of these tags is most strongly predictive of average difficulty?**

For this analysis, a multiple linear regression model will be used to interpret which of the 20 tags provided is most strongly predictive of average difficulty. This was done under the assumption that the difficulty ratings data and the linear relationships between the average difficulty and its predictors is independent. To account for multicollinearity, VIFS were calculated for each tag. Any tag with a VIF value greater than 10 indicates that it is a predictor that is highly correlated and was removed from the regression model training. After assembling the predictors, the model performed with an R^2 score of 0.6130 and an RMSE of 0.4903. Examining the results that can be seen in **Figure 1.9.1**, the model slightly underperforms in comparison to Question 3.8, where the tags were used to calculate average rating. Under this performance, the most strongly predictive of average difficulty is Tough Grade ($\beta$=2.8352), Graded by few things ($\beta$=−2.7616), and Clear grading ($\beta$=−2.6166). Interpreting these results suggests that students associate higher difficulty with tough grading, and lower difficulty with fewer graded items or clear grading policies. However with student submitted rating data, limitations include potential biases in how students assign tags and other factors outside of the dataset contributing to the tag assignment or average difficulty rating.

**Figure 1.9.1: Predicted vs Actual Ratings (Tag Model)**




**1.10 Pepper Classification Model**

In building a model to classify whether a professor receives a pepper, it assumes independence between observations in the data. For classification, a random forest classifier was chosen because it is robust to overfitting and provides a feature importance metric for interpretability. To address any class imbalance concerns, Synthetic Minority Over-sampling Technique (SMOTE) was used as it can improve the representation of the minority class and avoids overfitting. After training the Random Forest Classifier, Visualized in **Figure 1.10.1**, it produced an accuracy of 75%, with an AUROC score of 0.8167. The

classification model predicts whether a professor receives a "pepper" with strong performance, indicating that it effectively distinguishes between professors who do and do not receive this designation. In **Figure 1.10.2**, The most predictive features include Average Rating, Respected, and Tough Grader. Average rating is a greater factor than the other two tags, however these factors are key determinants in predicting professors receiving a pepper. Overall, the model performs well for predicting a pepper, however there are potential limitations into how students assign peppers, which includes extremely subjective factors that may not be captured numerically.

**Figure 1.10.1: ROC Curve**

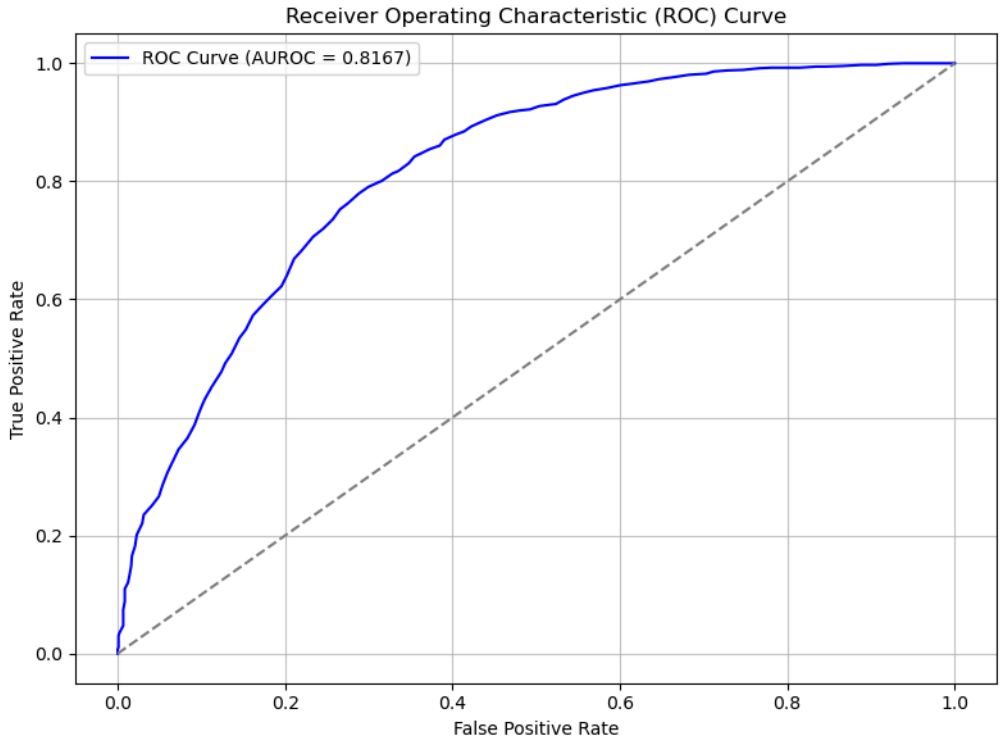



**Figure 1.10.2: Top 10 Features Predicting “Pepper”**