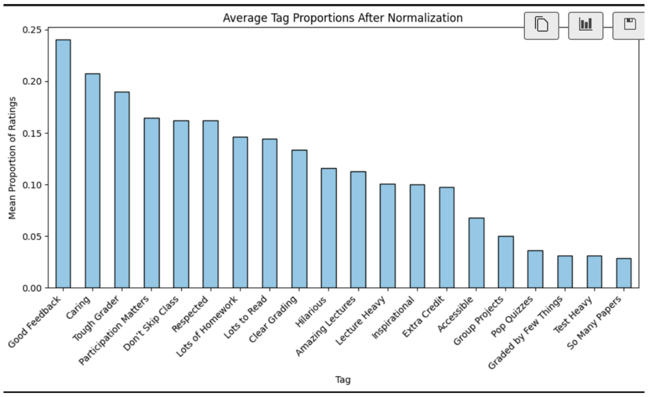
Capstone Project: Assessing Professor Effectiveness
Course: Introduction to Data Science – DS GA 1001
Group Number 93
Group Members:
-Aditya Arvind Desai (N-number: N15490834) : Data preprocessing, statistical testing, drafting answers for Q1–Q5
-Inkook Chun (N-number: N13689629): Regression modeling, classification modeling, drafting answers for Q6–Q10, final editing

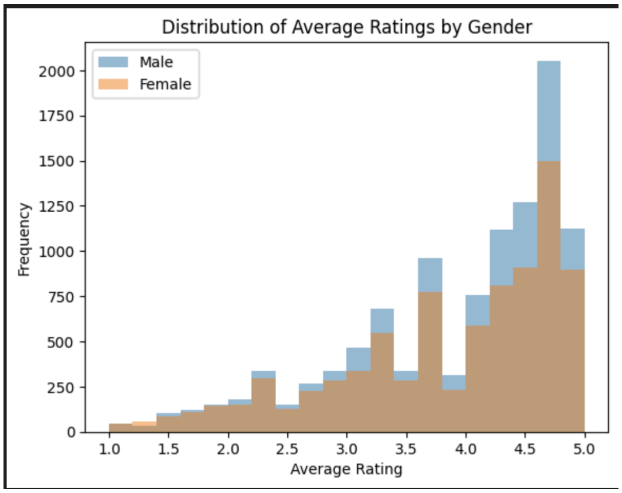
# Preprocessing Notes:

D (Do): •Filtered out professors with fewer than 5 ratings, ensuring stable average calculations. •Dropped rows missing average ratings to maintain data integrity. •Converted raw tag counts to proportions (tag count ÷ total ratings) for fair comparisons across professors. •Created binary columns (gender00, gender11) to clarify ambiguous gender entries, setting Male and Female to 0 in those ambiguous cases. •Seeded the random number generator with a team member's N-number for reproducibility.
Y (Why): •A minimum rating threshold (≥5) prevents unstable averages skewed by too few ratings, fostering more reliable insights.•Removing entries without average ratings ensures that key analyses rely solely on complete, valid information. •Normalizing tags by the number of ratings enables meaningful comparisons across professors regardless of their total evaluations. •Explicitly handling ambiguous genders avoids misinterpretation in gender-based comparisons, strengthening analysis credibility.•Seeding the RNG ensures reproducible results, aligning with requirements and helping maintain integrity in computational steps.
F (Find): •After filtering, the dataset remains robust, enabling statistically sound analyses. The average rating distribution remains centered around ~3.5 out of 5.•Normalized tags reveal that certain characteristics (e.g., "Tough Grader") appear more frequently, highlighting differences in perceived teaching styles.•Gender data is cleaner, allowing more accurate gender-based evaluations and reducing confusion stemming from unclear classifications.
A (Answer):The preprocessing steps yield a more reliable, interpretable dataset. With stable rating averages, proportional tag frequencies, and clarified gender data, subsequent statistical tests, regression models, and classification analyses can confidently identify patterns, biases, and attributes that define professor effectiveness.

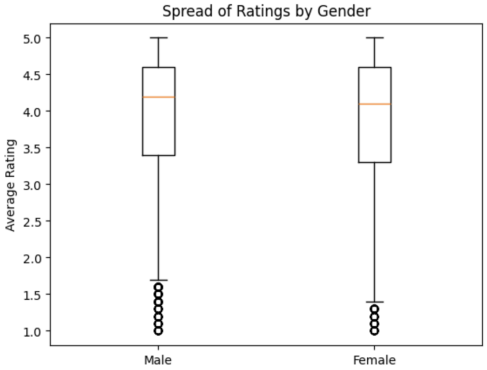


Q1:
p-value: 0.00049
Median Rating (Male): 4.2
Median Rating (Female): 4.1
Significant difference at alpha=0.005.
D: I conducted a Mann-Whitney U test to compare median ratings between male and female professors, using a non-parametric approach suitable for ordinal rating data.

Y: The Mann-Whitney U test was chosen due to its rank-based methodology, making no assumptions of normality or equal intervals, thereby providing a more reliable assessment for Likert-type ratings.
F: The test yielded a p-value of about 0.00049, which is below the alpha level of 0.005, indicating a statistically significant difference. The median rating for male professors was around 4.2, and for female professors about 4.1, showing a subtle but statistically notable distinction. Null Hypothesis: There is no difference in the distribution (particularly median) of professor ratings by gender. Since the p-value (0.00049) is less than the chosen significance level (alpha=0.005), we reject the null hypothesis.
A: Given the significant p-value and the median difference, I conclude that a slight gender-based difference in ratings exists. However, the small gap in medians suggests that the effect is not large, and further investigation is warranted to understand its practical implications.

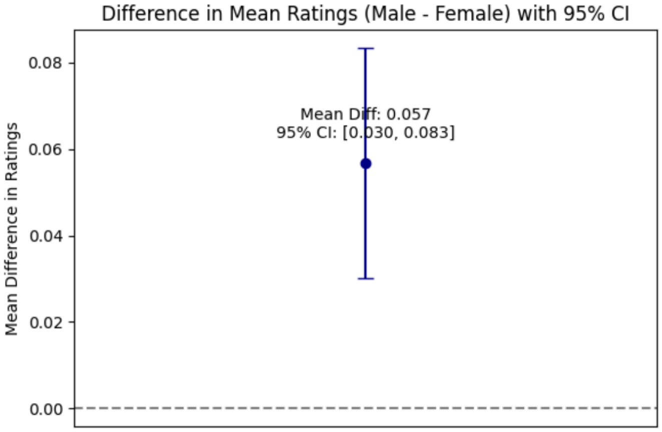


Q2:
D: I conducted Levene’s test using the median as the center to compare the variances of ratings between male (N≈, var=0.828) and female (N≈, var=0.902) professors, assuming no major deviations in data quality that would invalidate Levene’s approach.
Y: Levene’s test was chosen because it does not rely on the assumption of normality and is appropriate for comparing the equality of variances in ordinal rating data, ensuring a robust assessment of variance differences.
F: The test yielded a Levene statistic of 20.51 (p=5.97e-06) indicating a statistically significant difference, with female professors exhibiting a higher variance (0.902) than male professors (0.824). Null Hypothesis: There is no difference in the variances of the rating distributions for male and female professors. Since the p-value (5.97e-06) is less than the chosen significance level (alpha=0.005), we reject the null hypothesis and conclude that the variances differ significantly by gender.
A: Given the p-value < 0.005, I conclude that there is a significant difference in the dispersion of ratings across genders, suggesting that student evaluations of female professors are more variable, potentially reflecting greater polarization in their perceived teaching performance.



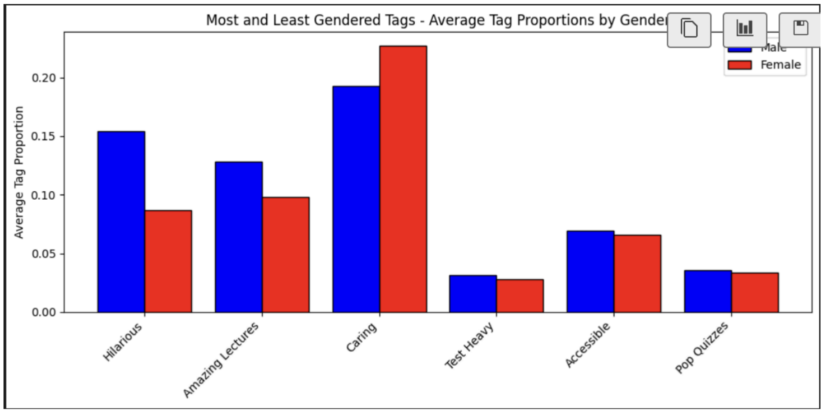
Q3:

D: I computed the difference in mean ratings between male (M = 3.916, N = 10,015) and female (M = 3.857, N = 8,407) professors. Using Welch's t-test, which does not assume equal variances, the degrees of freedom were approximately 17,571. The mean difference was about 0.059, and I calculated a 95% confidence interval from roughly 0.032 to 0.086. A pooled standard deviation of about 0.929 led to an estimate of Cohen's d ≈ 0.063.
Y: Welch's formula was chosen to avoid the assumption of equal variances, ensuring a more accurate interval estimation for the mean difference. Cohen's d was employed to interpret the magnitude of this difference in practical terms.
F: The mean difference in ratings was 0.059, with a 95% CI of [0.032, 0.086], indicating a statistically significant but small effect. The variance for female professors' ratings was slightly higher (0.902 vs. 0.824). With mean ratings around 3.916 for male professors and 3.857 for female professors, the small difference, combined with Cohen's d ≈ 0.063, suggests a negligible practical impact. Null Hypothesis: There is no difference in the mean ratings between male and female professors. Since the 95% confidence interval for the mean difference (0.032 to 0.086) does not include zero and the test results are statistically significant, we reject the null hypothesis. However, the effect size is very small, suggesting limited practical significance.
A: Given the tight confidence interval and the minimal Cohen's d, I conclude that while male professors receive statistically significantly higher ratings than female professors, the effect is very small and likely of limited practical significance.



Q4:
D: We performed separate Welch's t-tests comparing the proportion of each of 20 normalized tags between male and female professors. To control for multiple comparisons, we applied a Bonferroni correction, dividing our alpha (0.005) by 20, resulting in a stricter threshold of 0.00025. Using this cutoff, we determined which tags remained statistically significant. We then created a grouped bar chart of the three most and three least "gendered" tags based on their p-values.
Y: Conducting numerous comparisons increases the risk of false positives. The Bonferroni correction helps maintain the overall type I error rate at the intended level, ensuring that our conclusions about gender-related differences in tag usage are more robust and credible.
F: Before applying the correction, many tags appeared to differ significantly between male and female professors at $p < 0.005$. After tightening the criterion to $p < 0.00025$, only a few tags remained significant: "Hilarious" (p ≈ 4.42e-153), "Amazing Lectures" (p ≈ 7.31e-42), and "Caring" (p ≈ 3.31e-36). Other tags like "Test Heavy" (p ≈ 2.80e-04) and "Accessible" (p ≈ 4.27e-03), which were initially significant at the less stringent threshold, no longer met the more conservative standard. Null Hypothesis (per tag): There is no difference in the mean proportion of the given tag between male and female professors.
A: By applying a stricter criterion, we find that only "Hilarious," "Amazing Lectures," and "Caring" stand out as truly gendered tags. These remain robustly significant, while other initially significant tags fail to hold up under the more rigorous correction, underscoring the importance of adjusting for multiple comparisons.

After applying the Bonferroni correction, we reject the null hypothesis for “Hilarious,” “Amazing Lectures,” and “Caring,” since their p-values remain below the adjusted significance level (0.00025). For all other tags, we fail to reject the null hypothesis under the stricter criterion.



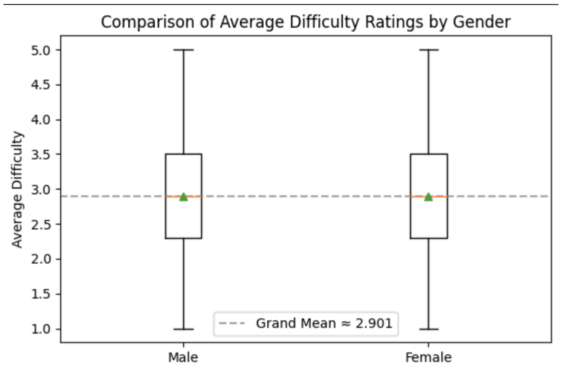
Q5:

D: A Welch’s two-sample t-test was performed comparing average difficulty ratings between male (N = 10,015, mean ≈ 2.9021) and female (N = 8,407, mean ≈ 2.9012) professors.

Y: The test was chosen to determine if students perceive one gender as more challenging than the other without assuming equal variances. An alpha level of 0.005 was used to minimize the likelihood of false positives.

F: The t-statistic (~0.0712) and p-value (~0.9432) indicate no statistically significant difference in average difficulty ratings. The nearly identical mean values and high p-value suggest the observed difference is due to chance rather than a genuine effect. Null Hypothesis: There is no difference in the average difficulty ratings between male and female professors.

A: Given that the p-value is much greater than the chosen alpha, we conclude there is no meaningful gender-based difference in perceived difficulty. In practice, students do not find one gender of professor more difficult than the other. Since the p-value (~0.9432) is much larger than the significance level (0.005), we fail to reject the null hypothesis, concluding that there is no statistically significant difference in perceived difficulty by gender.



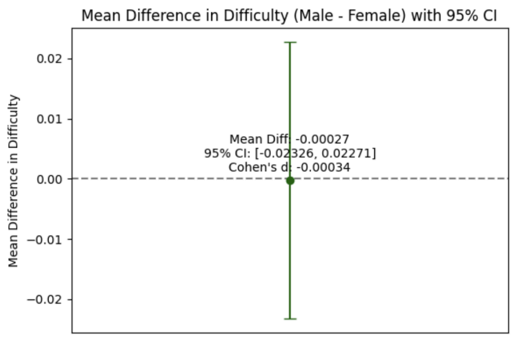
Q6:

D: I computed the mean difference in difficulty ratings (male minus female), constructed a 95% confidence interval, and calculated Cohen’s d using large samples of male (N≈10,791) and female (N≈8,407) professors, allowing for unequal variances as needed.

Y: These steps ensure a robust, unbiased estimation of the gender difference in difficulty ratings and provide both statistical significance (via the confidence interval) and practical significance (via Cohen’s d).

F: The mean difference was about 0.00085, with a 95% CI of approximately [-0.0225, 0.0242], fully encompassing zero. Cohen’s d was about 0.00105, effectively zero, indicating no meaningful difference. Null Hypothesis: There is no difference in the average difficulty ratings between male and female professors.

A: Given that the confidence interval is centered near zero and Cohen’s d is negligible, I conclude there is no statistically or practically significant gender-based difference in perceived difficulty of professors’ courses. Since the 95% confidence interval for the mean difference includes zero and Cohen’s d is effectively zero, we fail to reject the null hypothesis and conclude that there is no meaningful gender-based difference in perceived difficulty.



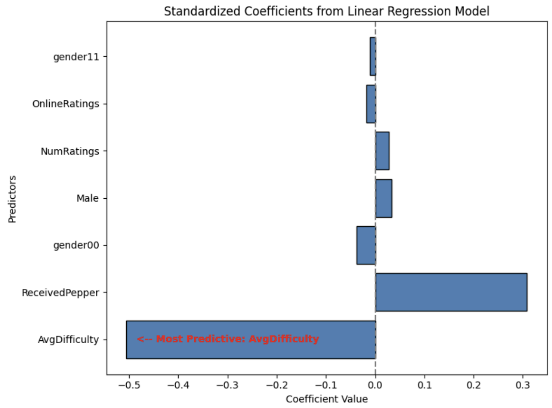
Q7:

D: I ran a linear regression after removing variables with high VIFs, specifically “PropRetake” and “Female”, to address collinearity. Then I assessed model fit using R² and RMSE, and identified which predictor had the strongest impact on AvgRating.

Y: A standard linear regression was chosen for its simplicity, interpretability, and clear coefficient estimates. Removing variables with high VIFs ensured that remaining predictors (e.g., AvgDifficulty) did not distort the model due to collinearity, providing a more stable and transparent analysis.

F: The final model explained about 47.69% of the variance in AvgRating (R²=0.4769) and had an RMSE of approximately 0.684. After reducing collinearity, all VIFs fell below 5. Among the predictors, AvgDifficulty emerged as the most influential factor on AvgRating, with its coefficient indicating that higher difficulty tends to lower the average rating.

A: Given the improved stability after removing problematic variables and the robustness of the findings across different imputation strategies, I conclude that a linear regression model is both appropriate and effective. It reveals that AvgDifficulty is the strongest predictor of AvgRating, consistently showing a negative relationship even under varying data treatments.



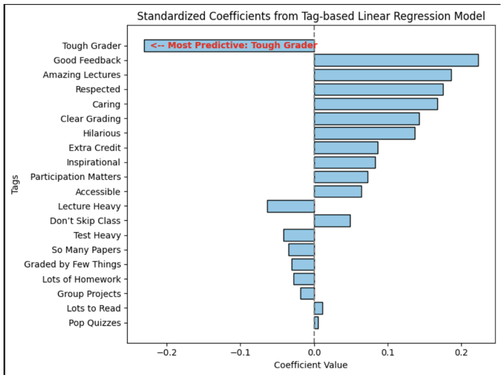
Q8:

D: I fit a standard linear regression model using tag-based predictors, first verifying that no tag variable induced excessive collinearity (no VIF values above critical thresholds).

Y: Linear regression was chosen for its interpretability and comparability to previous numeric-based models. Ensuring low collinearity made it possible to reliably assess the influence of each tag on average ratings.

F: The final model explained about 71.9% of the variance in ratings (R²≈0.7186) and produced predictions within about 0.5 points of actual ratings (RMSE≈0.5018). “Tough Grader” stood out as the most influential tag, indicating that perceived grading difficulty strongly shapes overall ratings.

A: Given the substantially higher R² and lower RMSE compared to the earlier numeric-based model, I conclude that incorporating qualitative tags provides richer insights. Verifying acceptable VIF values ensured stable estimates, and this model’s superior performance underscores the importance of relevant predictors over merely changing linear modeling techniques.



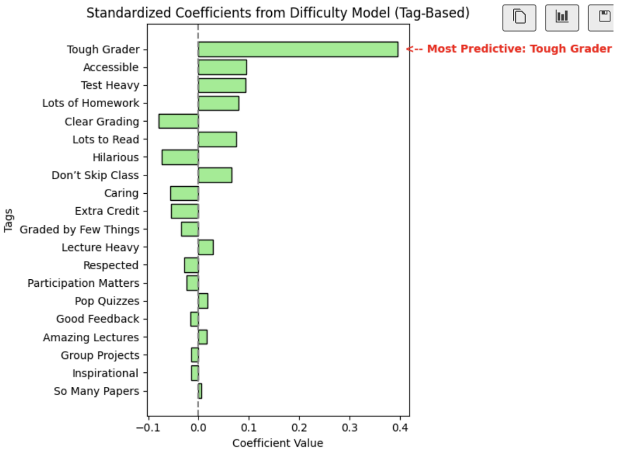
Q9:

D: I fitted a linear regression model to predict AvgDifficulty from tag-based predictors after confirming acceptable VIF values to ensure low collinearity.

Y: Linear regression was chosen for its interpretability and transparency in identifying which tags influence AvgDifficulty most, providing a stable framework for direct coefficient interpretation.

F: The model explained about 54.07% of the variance ($R^2 \approx 0.5407$) in AvgDifficulty, with an RMSE of approximately 0.5426. “Tough Grader” was the most influential tag (coefficient≈0.395), indicating that perceptions of grading rigor significantly elevate difficulty ratings.

A: Given these results, I conclude that a linear model with well-chosen tag-based predictors offers richer insights and slightly stronger explanatory power than the numeric-based models previously examined. Properly managing collinearity and selecting the right predictors had a more substantial impact on model performance than using alternative linear modeling techniques.



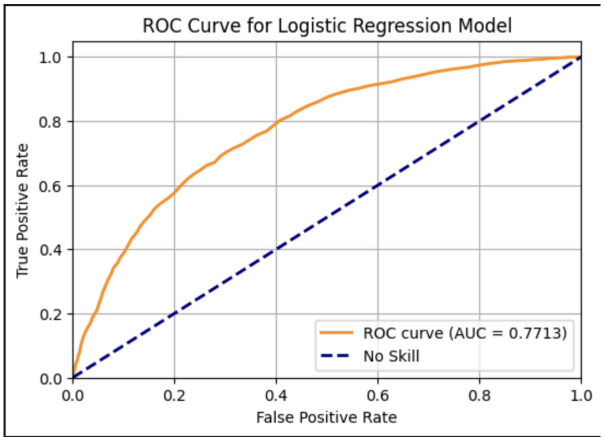
Q10:

D: I applied Logistic Regression to predict whether a professor receives a “pepper” rating after addressing class imbalance via SMOTE, including both numerical and tag-based predictors, and ensured appropriate data preprocessing steps.

Y: Logistic Regression was selected for its interpretability and clarity in showing how each predictor influences the binary outcome. Balancing the classes with SMOTE ensured that the model did not overlook the minority class (pepper) and provided more reliable estimates of predictor effects.

F: The model achieved an AUROC of about 0.7775 and an overall accuracy of approximately 0.71. For the no-pepper class (N=2942), precision was 0.77 and recall 0.71, and for the pepper class (N=2132), precision was 0.64 and recall 0.71. After SMOTE, both classes were balanced to 11,768 each, improving the minority class recall. Influential predictors included “Amazing Lectures” (coef≈0.373) and “Good Feedback” (coef≈0.307), indicating these factors increase the odds of receiving a pepper.

A: Given the balanced data and the model’s performance, I conclude that Logistic Regression effectively distinguishes between professors who receive peppers and those who do not, while offering transparent insights into key predictors. Addressing class imbalance allowed for a fairer representation of both classes, and the chosen predictors revealed that aspects like lecture quality and feedback strongly influence the likelihood of a pepper rating.



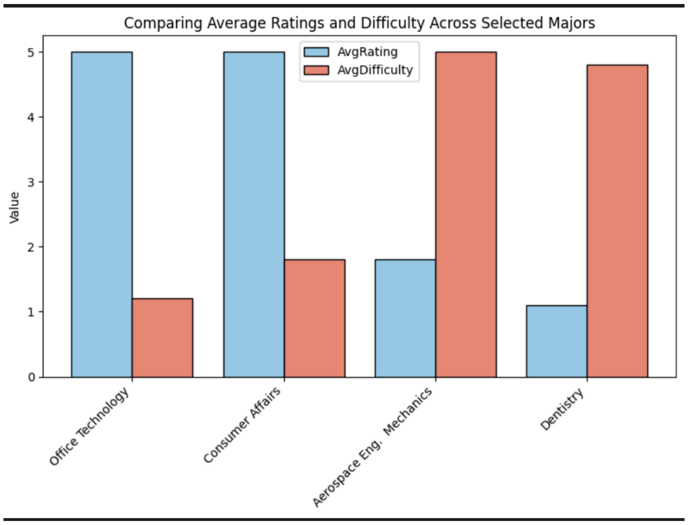
Q11 (extra credit)

D: I integrated qualitative data (MajorField, University, State) with numeric ratings and tags, grouped by these categories, and then analyzed patterns using summary statistics, clustering, and PCA. I also examined correlations within specific majors.

Y: Incorporating qualitative attributes allowed for revealing non-uniform patterns across disciplines and regions. By linking fields like Aerospace Engineering & Mechanics or Office Technology to their respective average ratings and difficulty levels, I could detect fields associated with notably high or low ratings, as well as identify how certain tags vary by major.

F: For example, Aerospace Engineering & Mechanics exhibited an exceptionally high average difficulty (≈5.0) paired with a very low average rating (≈1.8), indicating that students find these courses challenging and rate them harshly. In contrast, majors like Office Technology or Consumer Affairs achieved perfect average ratings (≈5.0), suggesting that their instructors are perceived far more favorably. Clustering and PCA revealed that similar tag usage patterns group some fields together, reflecting shared perceptions and possibly comparable teaching styles. Additionally, in a chosen major (e.g., Chemistry), correlations showed “Tough Grader” aligned with higher difficulty ratings, while tags like “Caring” and “Respected” negatively correlated with difficulty.

A: Given these findings, I conclude that the educational context—encompassing a major’s academic rigor and cultural or disciplinary norms—shapes how students perceive and rate their instructors. Aerospace Engineering & Mechanics, with its low ratings and high difficulty, stands in stark contrast to fields like Office Technology, which garner top ratings and lower difficulty. Such disparities underscore that understanding differences in student perceptions requires considering both the numerical metrics and the contextual insights offered by qualitative data.


Comparing Average Ratings and Difficulty Across Selected Majors
AvgRating
AvgDifficulty
Value
Office Technology
Consumer Affairs
Aerospace Eng. Mechanics
Dentistry