DSGA-1001 CAPSTONE PROJECT

GROUP 13:

GABRIEL NIXON RAJ (gr2513) - N12844700
HARSHIT SONI (hs5666) - N17865635

Gabriel and Harshit worked together on the data from ratemyprofessor.com to extract important insights. We worked on this project in phases, with Gabriel responsible for the documentation and preprocessing tasks and Harshit working on testing and visualization. Both of us went through all the questions together to find the most probable solution.

Stated below are a few assumptions and standard operating procedures that we have made, that are going to be applied to the majority of the questions:-

- We have assumed that the ratings are ordinal and they cannot be reduced to their sample means.
- Wherever the random seed was required for calculation, the random seed we used was 17865635 (Harshit's N-number).
- For the purpose of cross-validation, we assumed that the data should be randomly split into training and test sets in an 80:20 ratio. Given the contrast in the scale of feature values, we assumed that standard scaling was necessary for all features, and thus, applied standard scaling to the training set and transformed the test set using the same scaler to prevent information leakage.
- We have only considered collinearity when its absolute value is greater than 0.8.
- When considering the males and females, we eliminate rows that have values as the same in both columns (0-0,1-1) and are eliminated when strictly males and females are considered.
- When dealing with null values, row-wise removal is preferred to preserve intra-feature relationships.
- While making sense of average ratings, we ideally require the number of ratings to be as high as possible so that the average is representative of the actual rating trend for the professor. However, as the threshold goes beyond 5, the data is reduced to an extremely small fraction. Thus, we fix the number of ratings to be greater than 5 as our threshold.
- In context to finding the effect size we have used cohen's d formula as that is the only effect size test that was covered in our syllabus,
- The formula that we used for this context as sample sizes are not equal are:-

$$d = \frac{\mu_1 - \mu_2}{s_p} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

1) A - We perform the preprocessing using the standard assumptions as stated above, reducing the data to almost 14900 points.

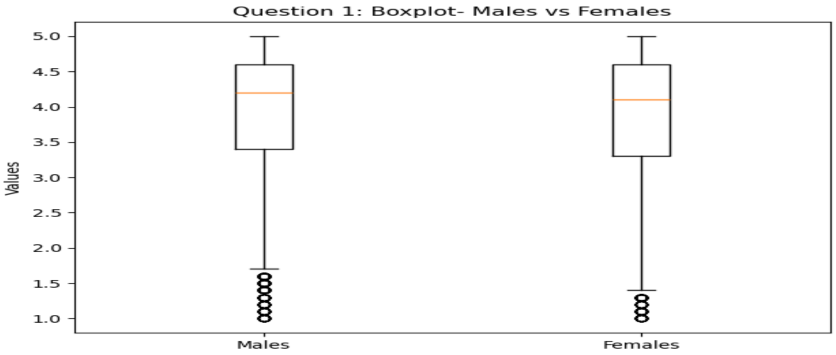
Y - We used the Mann Whitney test for checking if there is a gender bias as we cannot reduce rating data into its sample means as the data is not numerical and to check if there is a pro male bias we only require a 1 sided Mann Whitney U test.

Null Hypothesis (H_0): There is no difference in the average ratings between male and female professors

Alternative Hypothesis (H_1): Male professors receive higher average ratings than female professors

F - We found through our calculations that, P-value: 0.0003929858837938564

D - With these values we can conclude that Given that our p-value is less than 0.005, we reject the null hypothesis and conclude that the observed data is unlikely given chance. This conclusion is based on the assumption that the data is independent and the Mann-Whitney U test is appropriate for ordinal, non-numeric data. Limitations include the potential for bias introduced during the row elimination process and the possible impact of not accounting for other factors influencing ratings.



2) A - We performed the above stated standard preprocessing on numerical data to extract the average rating for both females and males. This reduced the total data to almost 14900.

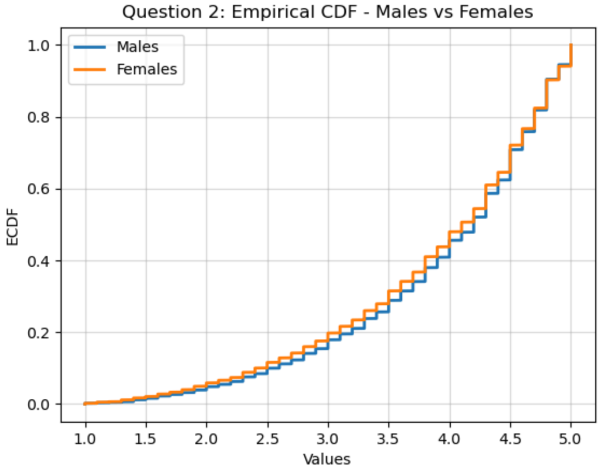
Y - We chose Levene's test because it is non-parametric and well-suited to assessing the equality of variances between groups. This allowed us to test for differences in the spread of ratings between genders without assuming normality. By focusing on the homogeneity of variances, Levene's test provided a robust method for checking the assumption of equal variances before proceeding with further analysis.

Null Hypothesis (H_0): There is no difference in the variance of ratings between male and female professors.

Alternative Hypothesis (H_1): There is a difference in the variance of ratings between male and female professors.

F - We found through our calculations that, **P-value = 8.798664061895703e-05**

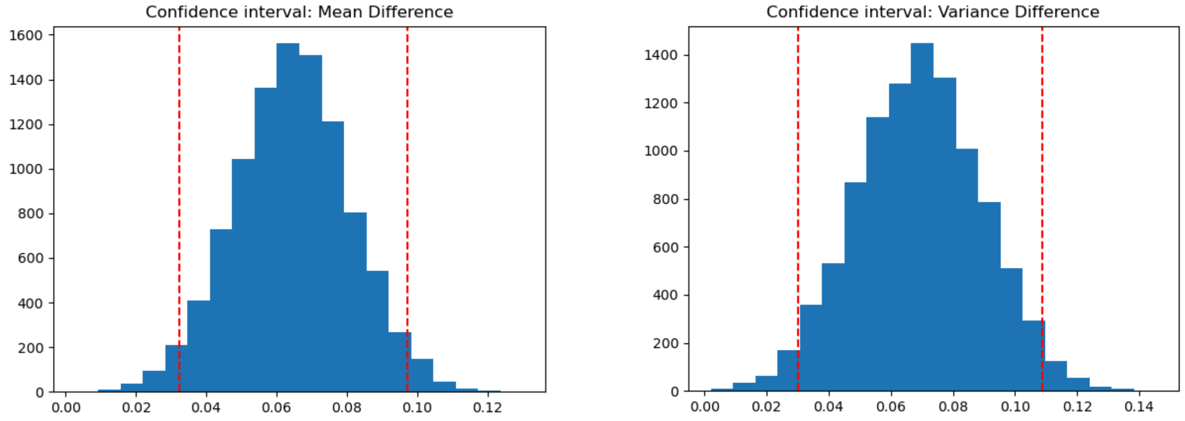
D - Given that the p-value is less than 0.005, we reject the null hypothesis and conclude that there is statistically significant difference in the spread (variance/dispersion) of ratings between genders. This conclusion assumes the independence of data and the appropriateness of Levene's test for assessing variance equality. Limitations include sensitivity to outliers.



3) A - We have again utilized the already preprocessed male and female data, assuming that the entries are independent within the group.
Y - To estimate the likely size of two effects, i.e variance difference and mean difference, we used bootstrap to generate samples and compare at 95% confidence interval. Through 10000 iterations, we were able to get the upper and lower bound for the confidence interval for both mean and variance.
F - We found through our calculations that,

- 95% Confidence Interval for variance difference: (0.0302, 0.1081)
- 95% Confidence Interval for Cohen's d (mean difference): (0.0324, 0.0978)

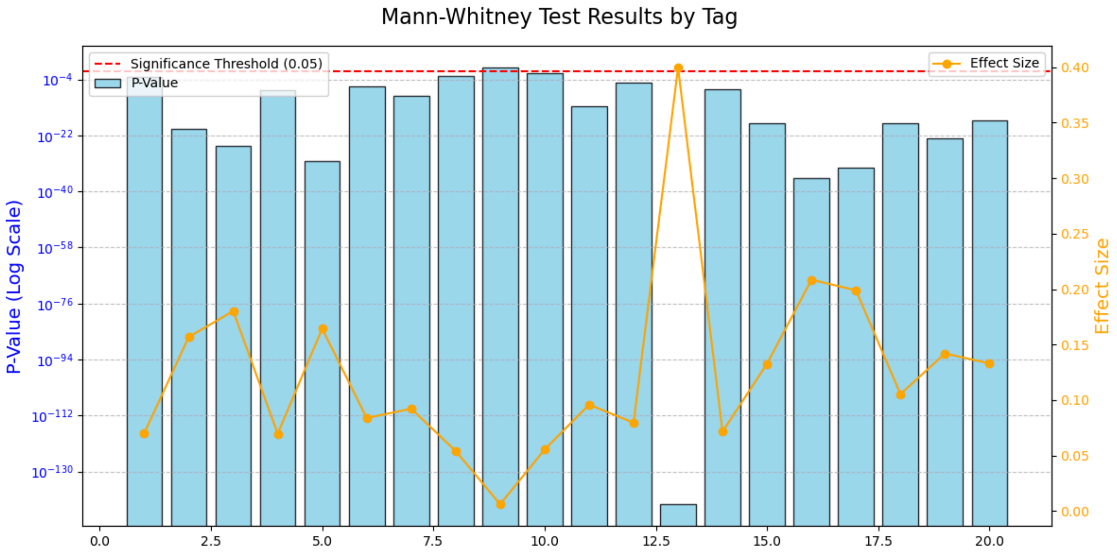
D - The 95% confidence intervals reveal small gender-based differences in ratings: the variance difference (0.0302, 0.1081) suggests slightly higher variability in one gender's ratings, while the mean difference (Cohen's d: 0.0324, 0.0978) indicates a minor bias in average ratings. While these results suggest some gender-based effects, the small effect sizes may lack practical significance. Limitations include potential confounding variables, assumptions of independence in the bootstrap method, and the possibility of meaningful patterns being overlooked during preprocessing.



4) A - We created a new dataframe combining the number of ratings, male and female counts from the numerical dataset, and all tags from the tag dataset. After preprocessing, the data was reduced to approximately 14,900 points, split by gender. To address the disparity in tag counts (ranging from thousands to zero), we rescaled each tag by dividing it by the number of ratings per professor. However, significant sparsity was observed, with some tags showing disproportionately more zeros for one gender. We assumed that the absence of a tag carries critical information, as a higher number of zeros for one gender may indicate a potential bias.
Y- As our data was highly skewed, a non-parametric test was needed. Hence, we used a two-sided Mann Whitney-U test for comparing between male and female counts for each tag, and set the following hypothesis down:
Null Hypothesis (H_0): For each tag, there is no significant difference in the proportion of the tag awarded between male and female professors.
Alternate Hypothesis (H_1): For each tag, there is a significant difference in the proportion of the tag awarded between male and female professors.
F- Are results from the above calculations are,
18 of the 20 tags are significant at a threshold of 0.005,

Top 3 features	P-value	Least 3 features	P-value
Hilarious	2.9872794201e-141	Inspirational	0.0015261772894
Amazing lectures	1.59397963309e-36	Accessible	0.0095267349171
Caring	5.22319095706e-33	Pop quizzes!	0.6453724626423

D- The Mann-Whitney U test revealed that 18 out of 20 tags exhibit a significant gender difference in their distribution. Notably, "Hilarious," "Amazing lectures," and "Caring" were the most gendered tags, while "Inspirational", "Accessible" showed no significant gender difference. The analysis assumes that the absence of a tag carries meaningful information, which may not always be valid. Additionally, the sparsity of data for certain tags could have influenced the results, and the non-parametric approach may not fully capture complex relationships in the data.



5) A - We extracted the following columns: gender, rating count, and average difficulty from the original data and performed our standard preprocessing as stated above. This reduced the dataset to about 14900 points.

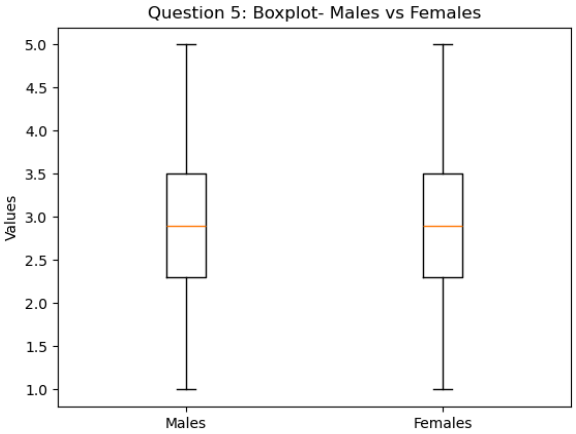
Y - We used a Mann WhitneyU test to assess whether there is a difference in the distribution of average difficulty ratings between genders. The Mann WhittneyU test is nonparametric and suitable for comparing distributions without assuming normality or equal variances.

Null Hypothesis (H_0): There is no difference in the average difficulty ratings between genders.

Alternate Hypothesis (H_1): There is a difference in the average difficulty ratings between genders.

F - From our analysis, we found: **P-value = 0.9680622117890496**

D - Given that the p-value is above the significance threshold of 0.005, we fail to reject the null hypothesis and conclude that there is no statistically significant difference in the distribution of difficulty ratings between genders. Limitations include potential biases introduced during data filtering and the lack of adjustment for confounding factors that may influence difficulty ratings.



6) A - We utilized the data similar to what we have done in the above question. We extracted the following columns: gender, rating count, and average difficulty from the original data and performed our standard preprocessing as stated above. This reduced the dataset to about 14900 points.

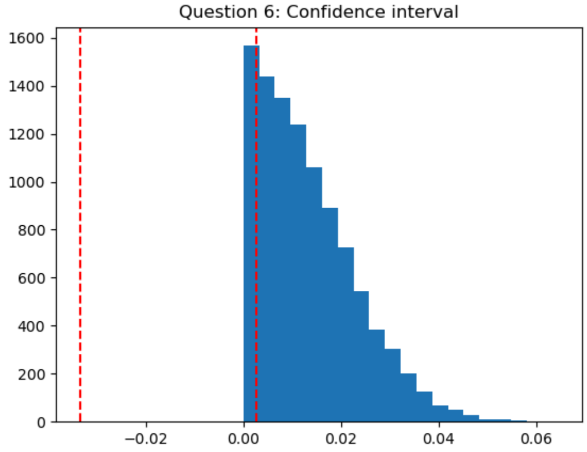
Y - To quantify the likely size of the gender differences in average ratings, we used bootstrapping to generate samples and calculate the 95% confidence interval for Cohen's d. Through 10,000 iterations, we were able to estimate the upper and lower bounds for the confidence interval of Cohen's d, providing insight into the magnitude of the effect.

F - Our calculations yielded the following results:

95% Confidence Interval for Cohen's d: (-0.0342, 0.0025)

Cohen's-d: -0.0015357661498668252

D - Based on the bootstrapping results, the 95% confidence interval for Cohen's d ranged from -0.0342 to 0.0025, with a point estimate of -0.0015. This suggests that the gender difference in average ratings is negligible, as the confidence interval includes zero, indicating no substantial effect. However, the limitations of this analysis include potential biases from data preprocessing and the assumption that the sample is representative.

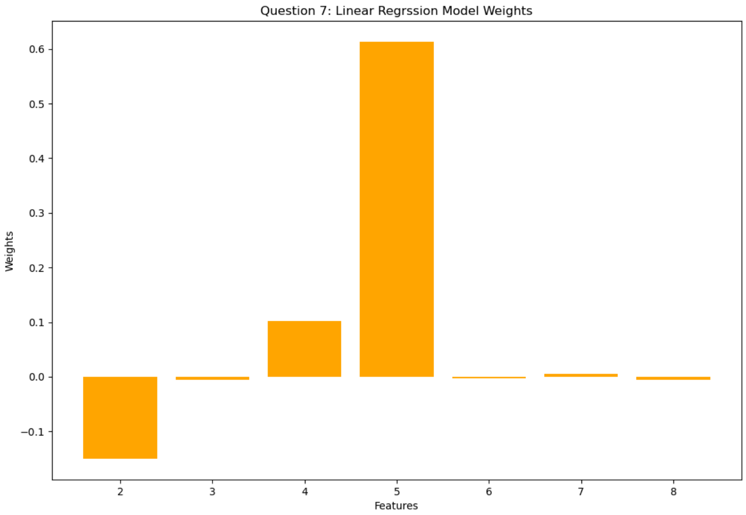


7) A - For predicting the average ratings, we used the features from the numerical dataset. To preserve the relationship among the dataset, we preprocessed the data along with the ratings. As there are over 77000 missing values for “would attend the class again” feature, our final data was eventually reduced to about 9000. The check for collinearity, yielded that no two columns had a value above the set threshold (0.8), therefore there were no collinearity concerns.

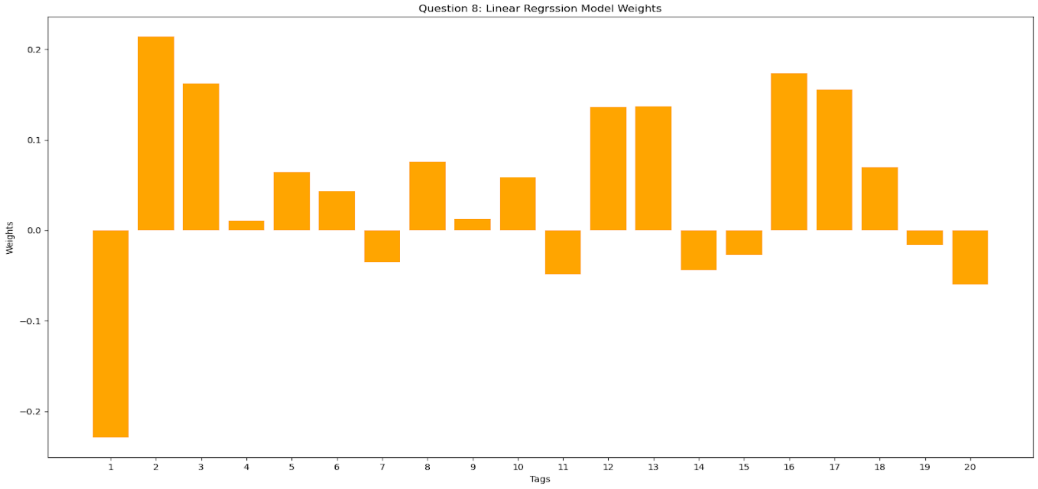
Y - We chose to use regression because it allows for quantifying the relationship between the predictors and the average rating, ensuring it captures both the magnitude and direction of influence. Then, we split and scale our features for fitting and validating the model.

F - The model's **R² was 0.799**, indicating that approximately 79.9% of the variance in average ratings could be explained by the predictors. The **RMSE was 0.368**, indicating the average error in predictions. The feature **"Take Class Again?"** had the largest absolute weight of **0.613**.

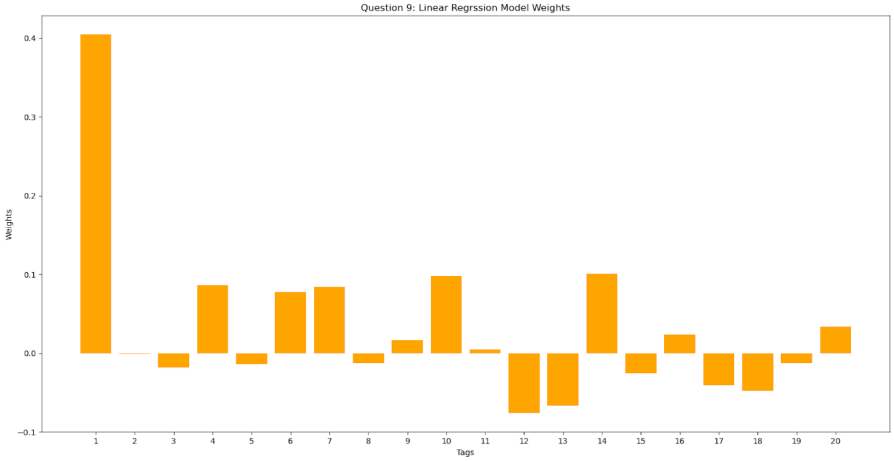
D - Given the R² of 0.799 and the RMSE of 0.368, we conclude that the regression model fits the data well, with "Take Class Again?" being the most predictive factor for average ratings. However, limitations include potential alpha inflation due to correlated features and assumptions regarding feature independence.



8) A - We combine average ratings, number of ratings, male, female, and all tags to create a new dataset and follow our standard preprocessing. This provided us with almost 14900 data points. We scaled the tag counts by dividing them by the number of ratings for each tag. Since there was no column whose collinearity breached our threshold, thus there were no collinearity concerns.
Y - Tags provide descriptive characteristics of instructors and courses, which may influence student ratings. Regression allows for quantifying how strongly each tag predicts average ratings. The data is divided into training and validation sets and scaled. The model is fitted on all 20 features and evaluated.
F - The model achieved an **R² of 0.728**, explaining 72.8% of the variance in average ratings, and an **RMSE of 0.477**, indicating the average prediction error. The tag **"Tough grader"** was the most predictive, with a weight of **-0.229**, showing a negative relationship with average ratings.
D - This model predicts average ratings reasonably well, though it performs slightly worse than the numerical model (R² of 0.728 vs. 0.799; RMSE of 0.477 vs. 0.368). The "Tough grader" tag is the strongest predictor, suggesting that stricter grading correlates with lower ratings. However, this model's predictive power is lower, potentially due to less direct relationships between tags and ratings compared to numerical features.

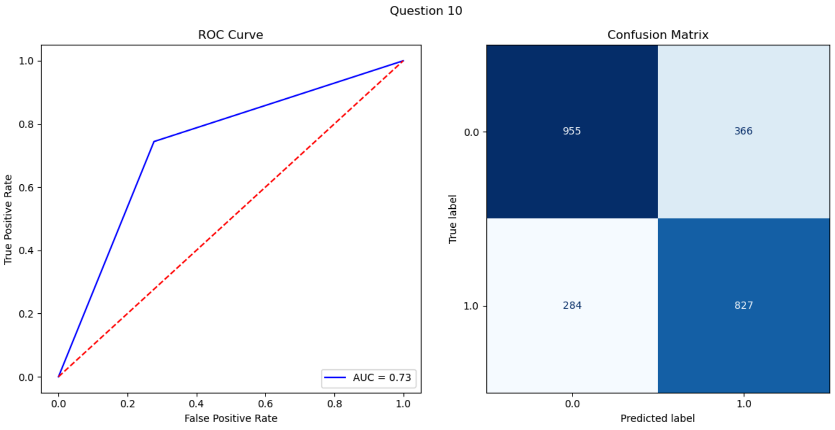


9) A- We combine the average difficulty, number of ratings, male, females and all tags to create a new dataset. Then we process it as per our assumptions. Finally, the number of tags are scaled by the number of ratings for each row. We also noticed that there were no collinearity concerns in this scenario.
Y- We randomly split the data into a test and validation set and fit the model on all 20 features.
F- The model achieved an **R² of 0.553**, explaining 55.3% of the variance in difficulty ratings, with an **RMSE of 0.526**, representing the average prediction error. The tag **"Tough grader"** was the most predictive, with a weight of **0.405**, indicating a strong positive relationship with difficulty ratings.
D - This model predicts difficulty ratings moderately well but less effectively compared to models predicting average ratings (e.g., R² = 0.553 here vs. 0.728 for ratings). The "Tough grader" tag strongly predicts higher difficulty ratings, highlighting its significant influence. However, the model's moderate performance may be due to limitations in how well tags capture the nuances of perceived difficulty.



10) A - We combine the numerical and tags dataset and drop missing values. The pepper column is then extracted and remaining serves as the features. When checking the collinearity, we find there is high collinearity between “average rating” and “proportion of those who take the class again”, amounting to 0.88. Hence, we drop the proportion column and only keep average ratings.
Y - The aim was to predict whether a professor receives a “pepper” using a Logistic Regression model. For evaluation, we used metrics like AUROC, precision, recall, and confusion matrices.
F - The Logistic Regression yielded an **AUROC of 0.74**. The model demonstrated balanced performance
D - We conclude that both models can reasonably predict whether a professor receives a “pepper,” with AUROC scores above 0.7 indicating good performance. However, limitations include the reliance on oversampling methods like SMOTE, which may introduce noise. Future improvements could involve hyperparameter tuning or exploring additional models to enhance performance.

	precision	recall	f1-score	support
0.0	0.77	0.72	0.75	1321
1.0	0.70	0.75	0.72	1111
accuracy			0.74	2432
macro avg	0.73	0.74	0.73	2432
weighted avg	0.74	0.74	0.74	2432



Extra Credit

A - We created a dataset by merging qualitative data (subjects) with numerical data (average ratings), resulting in approximately 70,000 data points. We classified courses into STEM (200 courses) and Non-STEM categories, cleaning the data to remove missing values in the "University" or "Major/Field" columns. Professors were grouped based on their field, and to balance the sample sizes, we randomly undersampled the Non-STEM group to match the STEM group, resulting in about 16770 professors per group which is about the same as the stem group. We retained professors with fewer than five reviews to avoid bias and ensure robust analysis. STEM fields include subjects that would include certain keywords such as ‘bio’,’chem’,’tech’,’math’, etc.

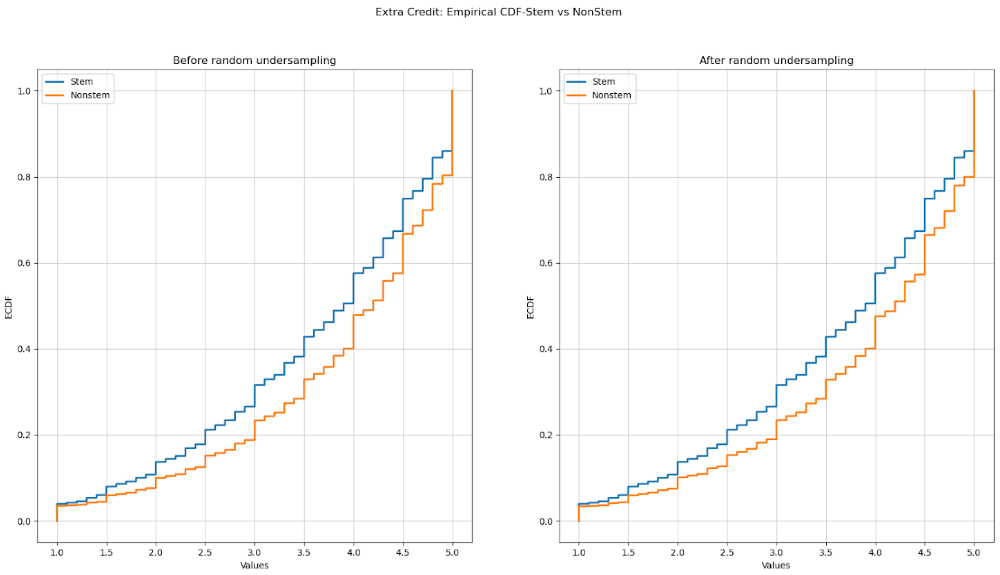
Y - To test if there was a significant difference in ratings between STEM and Non-STEM professors, we performed a 1-sided Mann-Whitney U test, which is suitable for comparing two independent groups when the data is ordinal, as is the case with ratings.

Null Hypothesis: There is no significant difference in the average ratings between STEM and Non-STEM professors.

Alternative Hypothesis: STEM professors are rated lower than Non-STEM professors.

F - Result: **P-value: 1.6203766010447876e-108, Cohen’s-d: 0.228**

D - Based on these findings, we reject the null hypothesis and conclude that there is a statistically significant difference in the ratings between STEM and Non-STEM professors. Non-STEM professors appear to receive higher ratings, with a small to medium effect size. These conclusions are based on the assumption that the data is independent and that the Mann-Whitney U test is appropriate for ordinal data. However, the analysis does have limitations, including the potential bias in the classification of majors as STEM or Non-STEM and the undersampling of the Non-STEM group, which may influence the results. Further analysis considering other factors, such as teaching quality and course engagement, would provide a more comprehensive understanding of the observed differences in ratings.


Extra Credit: Empirical CDF-Stem vs NonStem
Before random undersampling
After random undersampling
Stem
Nonstem
ECDF
Values