DS GA 1001 Capstone Project

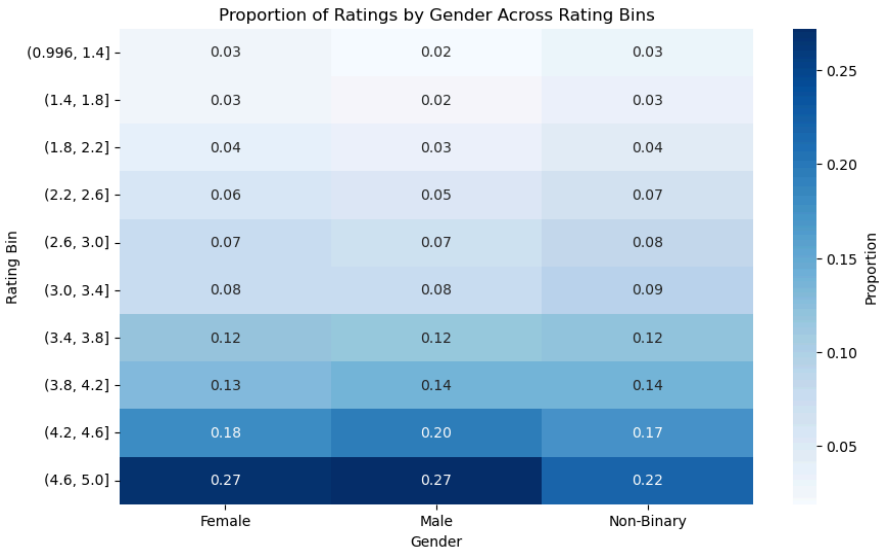
Mason Lonoff

Group 23

Preprocessing:

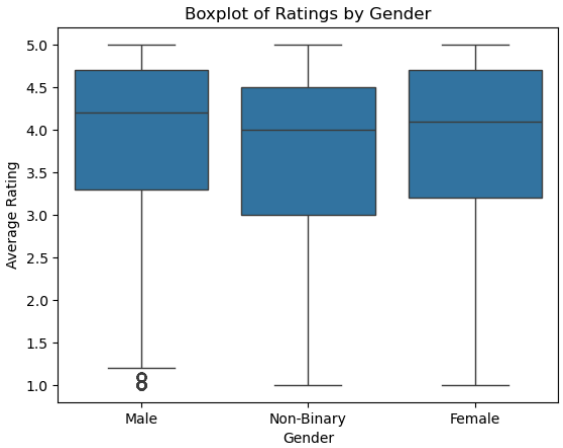
Most of the preprocessing was done on a case-by-case basis. General preprocessing consisted of reading in each data frame and renaming every column based on the project specifications. I examined data frame structures, data types, and null value counts. Both data frames were merged into a singular data frame called *merging_df*. Beyond this, the preprocessing done for this project was done separately for each question.

1. Is there a pro male gender bias in this dataset?



Firstly, I considered the dilemma of how professors with less ratings had less meaningful average ratings. I decided on using a minimum of three ratings per professor because it balanced retaining data and using enough ratings where the averages were reliable. I added a “Non-Binary” column that consisted of the rows where the gender was unclear. I also created a general “Gender” column. To determine if there was a significant statistical difference, I used a Kruskal-Wallis test. The null hypothesis is that there is no significant difference between males and the other genders in the dataset. After conducting the test, I found a p-value of 4.69e-50 which is a statistically significant result, so we have evidence to reject the null hypothesis. However, after conducting a pairwise Cliff’s Delta effect size test, I found that the effect size is negligible across each gender. On the heatmap above, we can see how there are differences across genders, but the overall proportions are similar which aligns with my findings.

2. Is there a gender difference in the spread (variance/dispersion) of the ratings distribution?

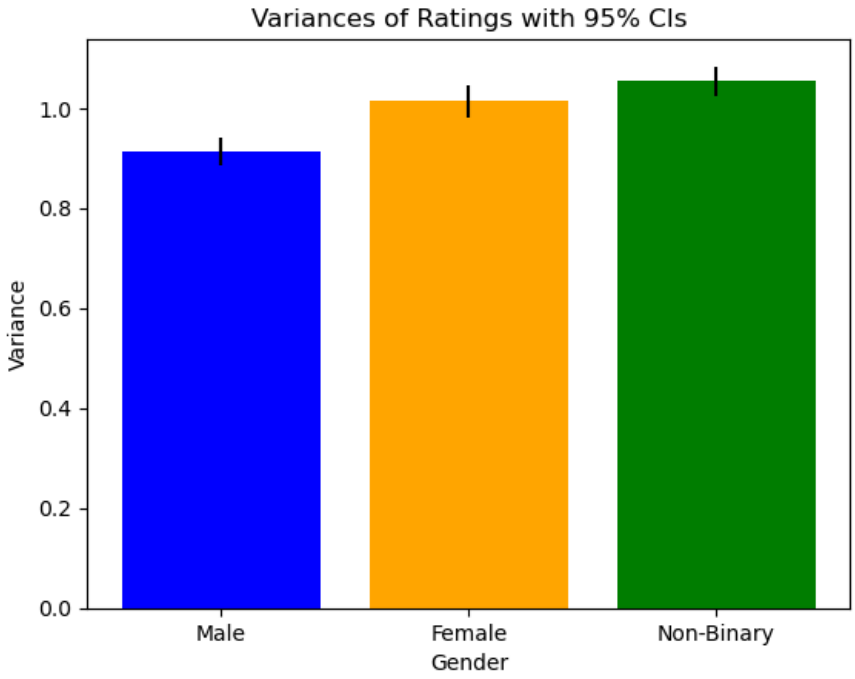


Using the same data frame from the last question, I calculated the variance and standard deviation of each gender. I then decided to use a Levene’s test to determine if there was a statistically significant difference in variances across genders. The null hypothesis was that there is no statistically significant difference in variances. After conducting the test I found a p-value of 2.744311093046388e-23. This is a statistically significant result, so we have evidence to reject the null hypothesis. I then decided to run Levene’s tests on a pair of each gender. Each result was below 0.005 which further supported my initial findings. I also calculated variance ratios between each group. I found that the genders Male and Non-Binary have the least similar variances at 0.867. Examining the boxplot above, we can notice how their distributions vary more than Male vs Female and Female vs Non-Binary which makes sense given my results.

3. What is the likely size of both of these effects (gender bias in average rating, gender bias in spread of average rating), as estimated from this dataset?

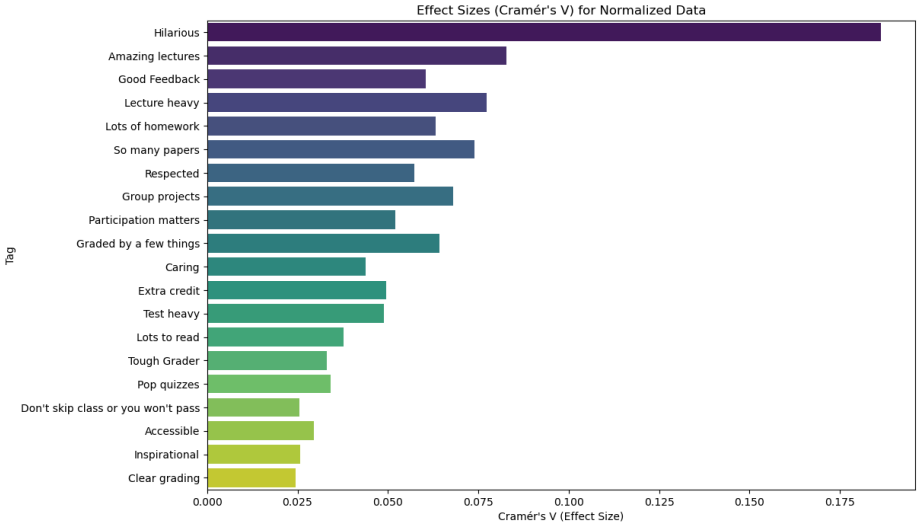
Comparison	CI
Male vs. Female	(0.01453, 0.040381)
Male vs. Non-Binary	(0.093007, 0.121269)
Female vs. Non-Binary	(0.062528, 0.091314)

To start off answering this question, I calculated Cliff's Delta effect size scores for each gender pair with a 95% confidence interval. I computed the confidence intervals using a bootstrap resampling method. These effect sizes are related to question 1. The effect sizes were all small. The exact effect sizes are to the right. For the effect size of question 2, I bootstrapped variance ratios and eta-squared with a 95% confidence interval. For the eta squared score, I calculated 0.994 with my confidence interval ranging from [0.9929-0.9948] which are very high scores. This score is abnormally high and requires further analysis given the opportunity. Using the individual gender variances calculated in question 2 in the bar plot to the right. I also incorporated the confidence interval calculated through bootstrapping.



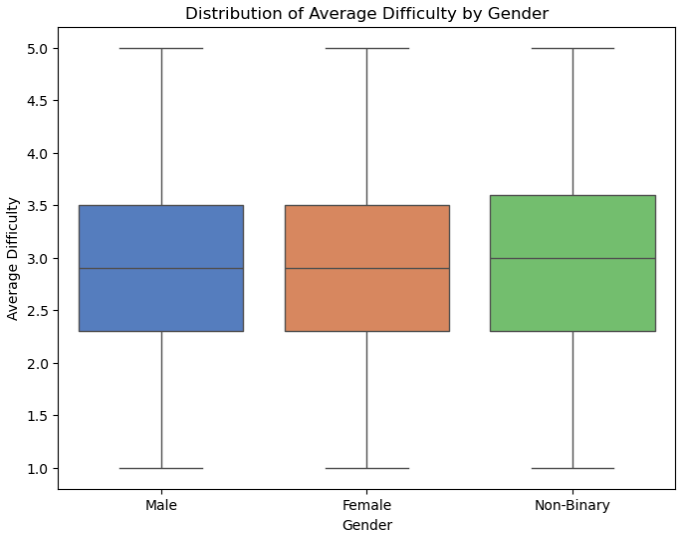
4. Is there a gender difference in the tags awarded by students?

Firstly, I filtered out all rows where the professor had less than 3 ratings. I normalized the counts of each tag as well by dividing the count by the total number of ratings for the professor then multiplying it by the mean number of ratings for all professors. I then manipulated the data so that proper contingency tables were set up for each tag. Every cell of every table had over 5 expected counts. I then ran a chi-square test for each table and calculated a Cramér's V effect size score. The null hypothesis for this test was that there were no statistically significant gender differences in tags awarded by students. However, based on my results, every sing;e tag had a statistically significant result, so we have evidence to reject the null hypothesis. The three lowest p-values were for the "Hilarious", "Amazing lectures", and "Good Feedback" tags. The three highest p-values were for the "Clear grading", "Inspirational", and "Accessible" tags. The plot above shows Cramér's V scores. We can see



that the “Hilarious” tag also has the highest effect score. It’s score is the only one to show a moderate strength which indicates that while there are statistically significant p-values, the practical effect isn’t as strong.

5. Is there a gender difference in terms of average difficulty

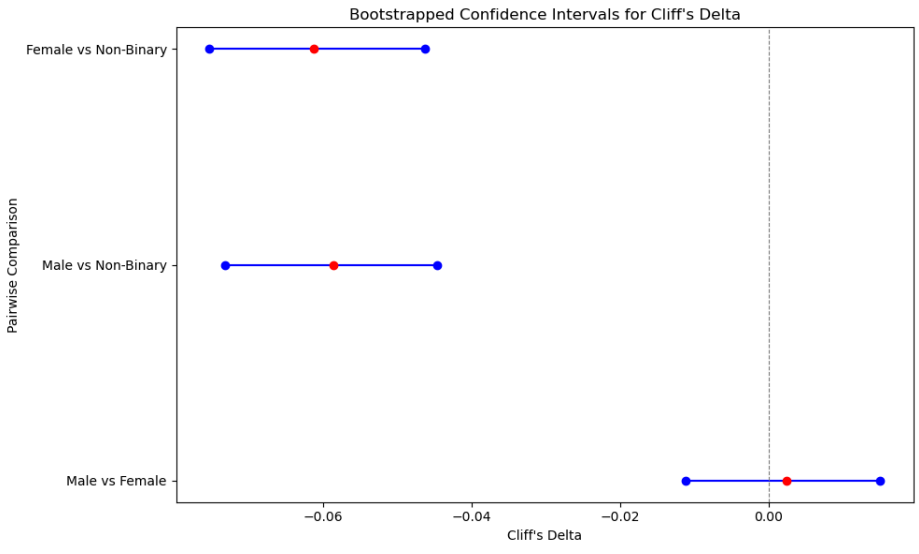


I first isolated the average difficulty values for each gender. Before conducting any statistical tests, I examined each gender’s difficulty distribution and concluded that none of them are normally distributed. The null hypothesis in this question is that there is no gender difference relating to average difficulty. I first conducted a Kruskal-Wallis test on each gender's average difficulties at once. I found a p-value of 1.728977102985476e-19 which is a statistically significant result. I wanted to know how each gender varies compared to each other, so I ran pairwise Man Whitney U tests. I found that the Non-Binary gender was statistically significant with each gender, but the Male and Female test had a p-value of 0.717 which is not statistically significant. There is evidence to suggest that we can reject the null hypothesis for comparisons with Non-Binary people and both Males and Females. The boxplot to the right shows how Males and Females have very similar distributions while the Non-Binary gender has a higher IQR and median which aligns with my findings.

	Comparison	Mean Cliff's Delta	95% CI
0	Male vs Female	0.002403	(-0.01115, 0.01496)
1	Male vs Non-Binary	-0.058578	(-0.07311, -0.04465)
2	Female vs Non-Binary	-0.061191	(-0.07523, -0.04625)

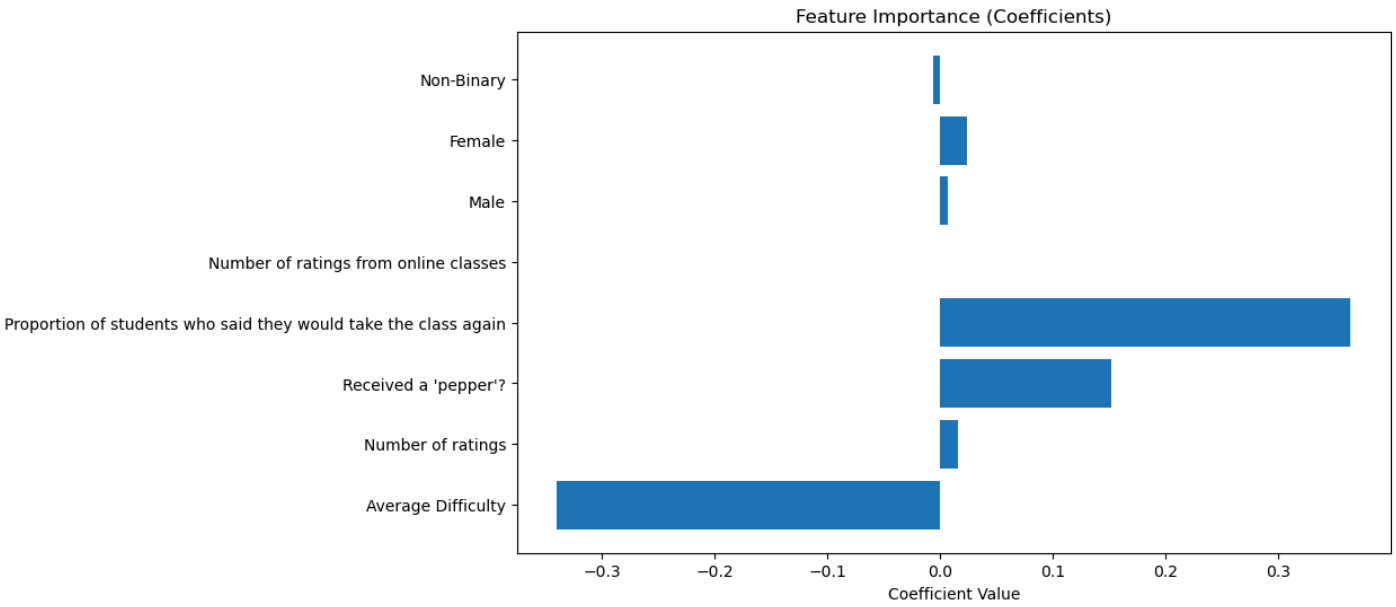
6. Please quantify the likely size of this effect at 95% confidence

For the initial Kruskal-Wallis test, I decided to use epsilon squared effect size scores to analyze the effect size of the overall differences. I bootstrapped these values with a 95% confidence interval. My results were [0.0012-0.0029] which are very small effects. For the pairwise significance tests, I used bootstrapped Cliff’s Delta scores to calculate the effect sizes. My results from this 95% confidence interval is shown in the image to the right. Also to the right is a plot of the individual Cliff’s Delta bootstrapped confidence intervals with the mean point in red.



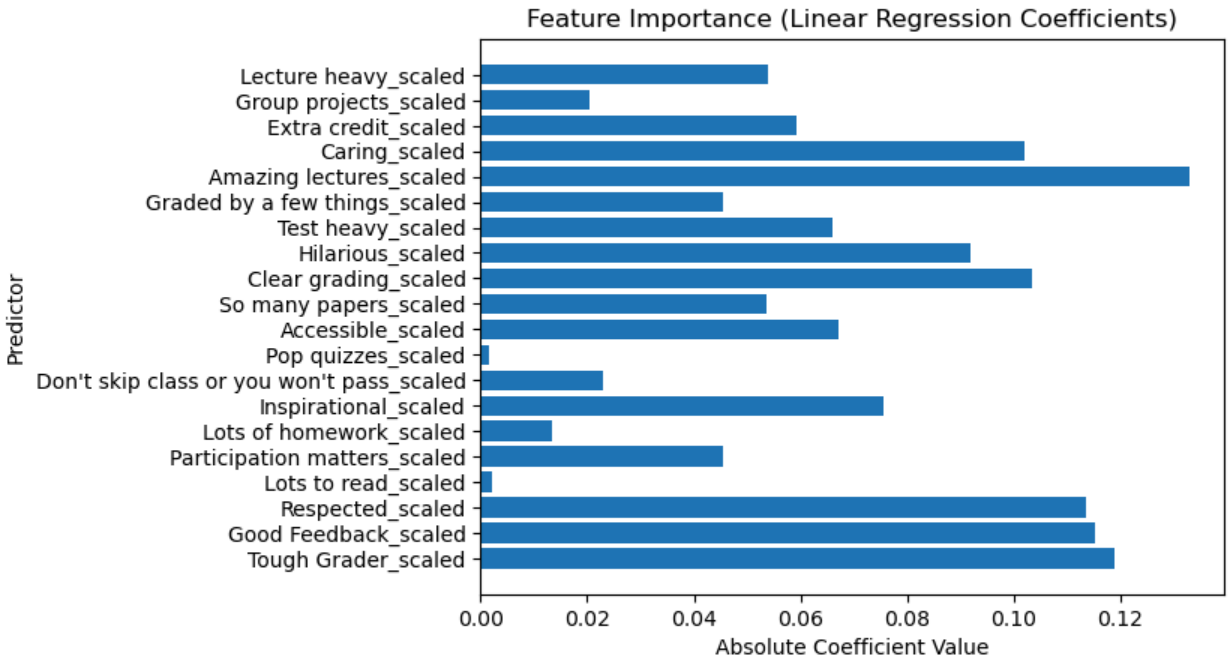
7.Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these factors is most strongly predictive of average rating?

The first thing I did was to examine each predictor's correlation to the target variable. I found that column "Proportion of students who said they would take the class again" strongly correlated with the target variable (0.88). It also had 28,368 missing rows, so I decided to impute it since it was an important predictor in this model. I used a Random Forest Regressor to impute the data. The correlation coefficient stayed at a moderate positive correlation, so the relationship between the target and predictor variable was mainly maintained. None of the scores in the correlation heatmap I set up were high. I then split the data into predictors and the target variable. I split the data into training and test sets, and then scaled the data using Standard Scaler. I fit the model and then ran the regression. My r^2 score was 0.507 and my RMSE was 0.694. The column "Proportion of students who said they would take the class again" most strongly predicted the average rating. "Average Difficulty" was also a stronger predictor. We can see the feature importances in the image below.



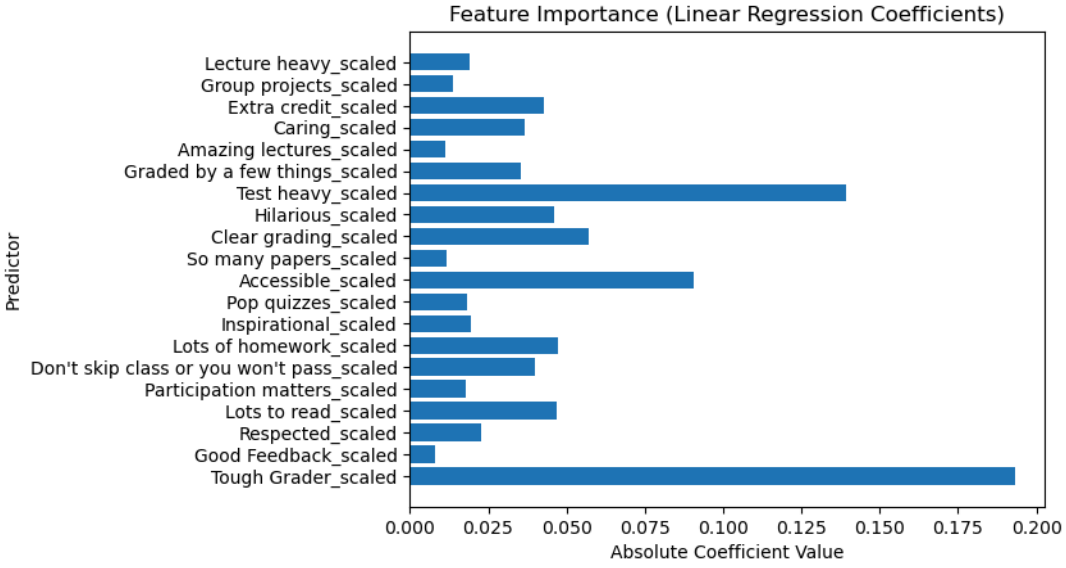
8. Build a regression model predicting average ratings from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these tags is most strongly predictive of average rating?

Using the same dataframe from question 4, I dropped all the non relevant columns except "Average Difficulty" and the non-scaled tag columns. I created a heat map to check correlations. No predictors appeared to be correlated, but I calculated VIF scores anyway. None of those scores showed any multicollinearity. I then ran the model. I didn't scale the data since all the data was close in value due to prior normalization. My R^2 score was 0.654 and the RMSE score was 0.585. The most predictive tags were "Amazing_lectures_scaled", "Tough Grader_scaled", and "Good Feedback_scaled". The plot below shows all the coefficients in the model. We can see the coefficient values.



9. Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty?

I dropped all columns that weren't essential to this analysis. I kept the "Average Difficulty" column and all the scaled tag columns. I set up a correlation matrix to examine multicollinearity. There didn't appear to be any correlated predictors, but I calculated VIF scores to verify. None of the predictors had high scores, so there were no collinearity scores to be concerned about. I then set up and ran the model. Once again, I didn't need to scale the data since all the values were on a similar scale. My R^2 score was 0.476 and my RMSE score was 0.612. The "Tough Grader_scaled", "Test heavy_scaled", and "Accessible_scaled" columns were the most predictive. The image below shows these coefficient scores. There's a large gap between the the top scores and the rest of the pack.



10. Build a classification model that predicts whether a professor receives a “pepper” from all available factors (both tags and numerical). Make sure to include model quality metrics such as AU(RO)C and also address class imbalance concerns.

For this model, I included every single feature besides the unscaled tags. I then used the same Random Forest Regression to impute values in the "Proportion of students who said they would take the class again" column. I printed out VIF scores and saw some columns had high values. I used PCA dimension reduction to lessen the dimensions in an effort to fix this issue. I scaled the data and ran the logistic regression. I set the “class_weight” variable equal to “balanced” to try and deal with the unbalanced classes. My ROC-AUC score was 0.79. My overall accuracy was 0.71 and the majority class was 0.62, so my model beat the majority class by ~9%. The two images provided here are the confusion matrix and the ROC curve.