Trem Ampeloquio
Principles of Data Science

Capstone Project Report

Statement on handling of data preprocessing:

I started off by dropping all rows in the quantitative dataset that had null values for the column corresponding to the number of ratings. Given the skewed distribution of the number of ratings from the remaining data (mean = 5.37, median = 3), I chose to use the median as a threshold to filter professors in the dataframe. Consequently, those who had fewer than 3 total ratings were dropped from the dataset. This ensures I retain professors with a sufficient number of ratings for meaningful analysis while not disproportionately favoring outliers.

Additionally, given that none of the given datasets had column titles, I assigned the respective titles myself to make dataframe access more efficient.

I also chose not to merge the qualitative and quantitative datasets initially because all of the questions for analysis required some sort of significance testing for variables with numerical data. I only merged these datasets for the extra credit problem in which I needed the categorical data for the major/field. Using the same criteria as before, I dropped NA rows for the number of ratings as well as rows that had less than the median number of ratings after extracting rows containing “Chemistry” and “Psychology” for the major/field.

Lastly, I seeded the random number generator to my N-number.

1: Is there evidence of a pro-male gender bias in this dataset?

I chose a Mann-Whitney U Test because I found that the variances between male and female professors were different and none of the distributions were normal. Implementing this test, I got a p-value of 2.06e-5, which is statistically significant. Thus, there is statistical evidence of a pro-male gender bias in the dataset. However, the effect size (rank-biserial correlation) was very small (r = 0.027), indicating that the practical significance of this difference is minimal. An effect size of this magnitude suggests that although the statistical test found a significant difference, the actual difference in ratings between male and female professors is quite small.

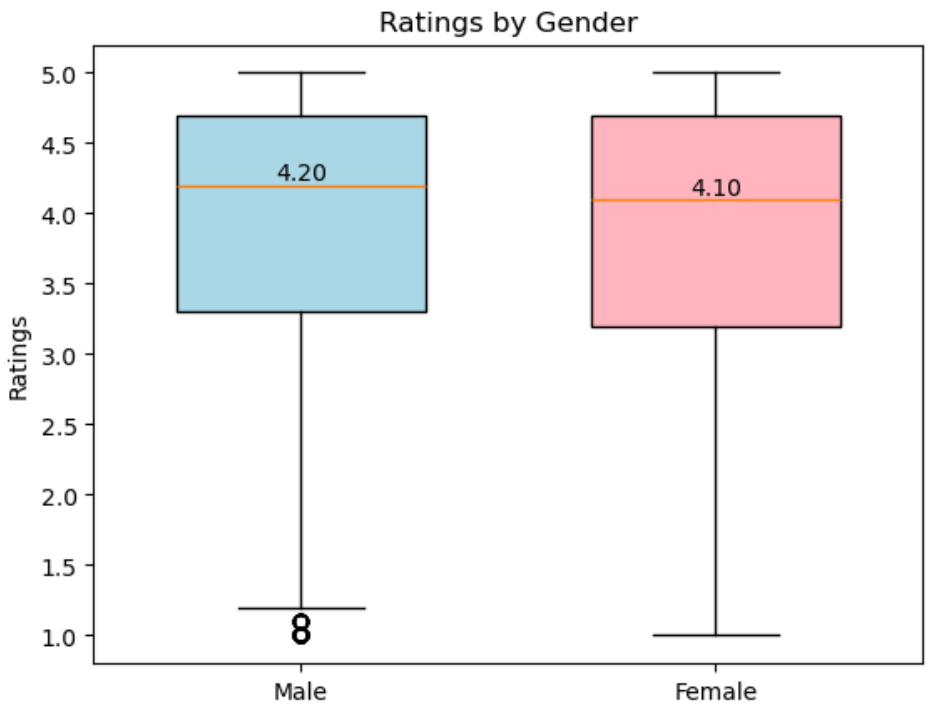

Figure 1: Boxplot of resulting Mann-Whitney U Test for differences in the average rating by gender.

2: Is there an effect of experience on the quality of teaching?

To investigate whether teaching experience impacts the quality of teaching, I used the number of ratings as a proxy for experience and average rating as a measure of teaching quality. A linear regression model was conducted to evaluate the relationship between these two variables. It's important to note that this analysis only includes the 99th percentile of the number of ratings to exclude outliers that could potentially impact the regression line. The analysis yielded a coefficient for the number of ratings of 0.0062 and a p-value of 0. These results suggest a statistically significant weak linear relationship between experience and teaching quality, although it may not be entirely practical. With an R^2 value of 0.001, the number of ratings/experience alone does not explain the variation in ratings very well.

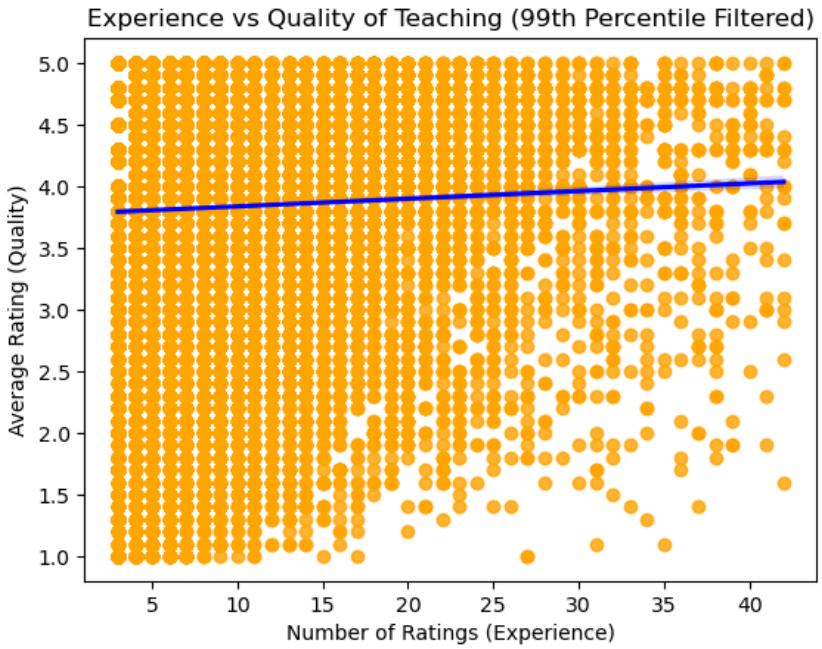

Figure 2: Linear regression model predicting average rating operationalized as quality, from number of ratings operationalized by experience.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.7741	0.008	461.341	0.000	3.758	3.790
number of ratings	0.0062	0.001	7.375	0.000	0.005	0.008

Figure 3: OLS regression results for the relationship between experience and quality

3: What is the relationship between average rating and average difficulty?

I first plotted Average Difficulty and Average Rating on a scatter plot. The plot lacked any sort of obvious linear relationship and the distributions for average difficulty and average rating proved to be non normal (via a significant KS Test). Since the conditions for a Pearson correlation test were violated, I used Spearman's correlation test. With a p value of 0 and a correlation coefficient of -.57, there is a moderately negative monotonic correlation between average difficulty and average rating. This suggests that as the perceived difficulty of a professor's course increases, the average rating tends to decrease.

Figure 4: Scatterplot for average difficulty and average rating with non linear trendline for a Spearman correlation

4: Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't?

To operationalize "a lot", I found the median number of ratings from online classes but this ended up being 0. I also saw that the maximum number of ratings from online classes was 17. Given half of the data is 0, I made a custom threshold. If a professor had more than 2 ratings from online classes, I classified them as teaching a lot of classes in the online modality. After this I made two groups of data, professors who taught a lot of online classes and those who didn't. Since the variances for these groups were different by a significant Levene test, I chose to conduct a Mann Whitney U test instead of an Independent T-test. Using a one tailed test, I got a significant p-value of 2.322e-09. Thus, I rejected the null hypothesis and accepted that online professors have significantly lower ratings than those who don't teach online as much. However, I found the effect size to be -0.07 (Rank Biserial Correlation) indicating that this difference between ratings of online and in person classes is small in magnitude and therefore, not practically meaningful.

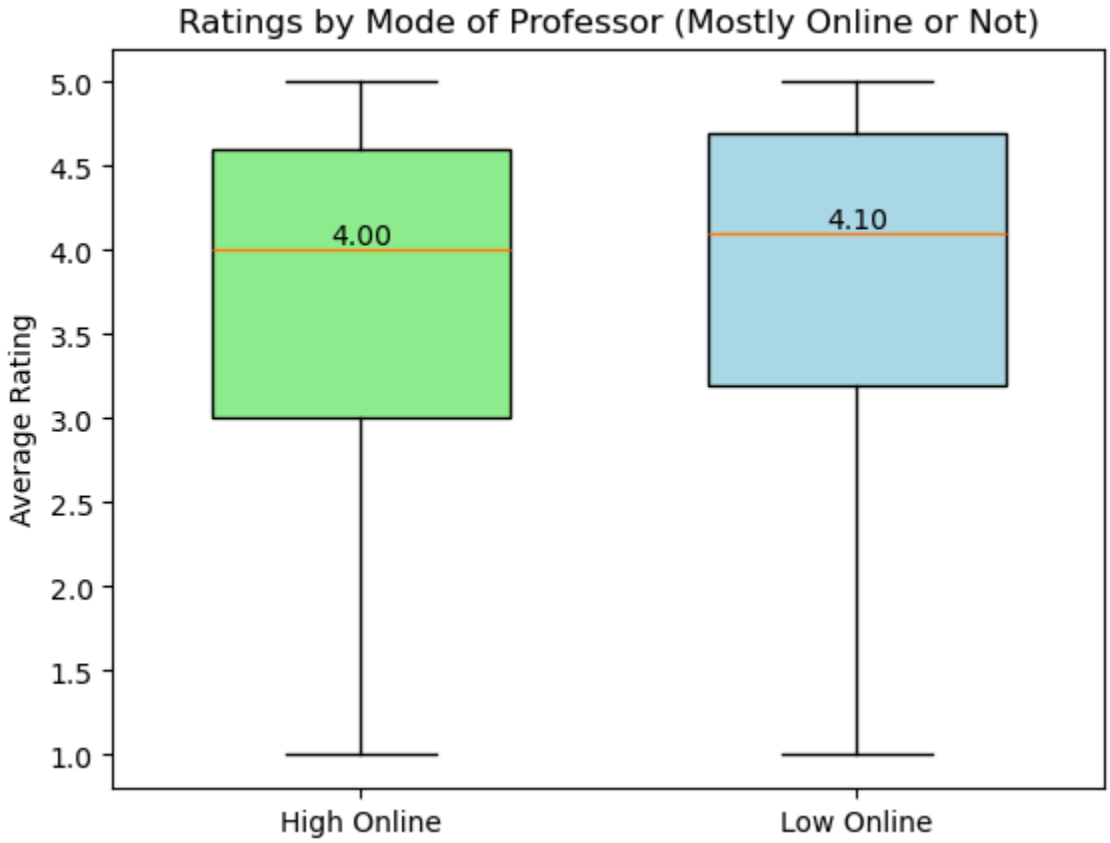

Figure 5: Boxplot of resulting Mann-Whitney U Test for differences in the average rating by mode of professor.

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

After checking that the distributions for average rating and proportion of students who would take the class with the professor again were not normal, I opted to find the Spearman correlation between the two variables. With a p-value of 0 and a Spearman correlation of 0.85, I rejected the null hypothesis and accepted that there is a significantly strong, positive, monotonic relationship between the variables.

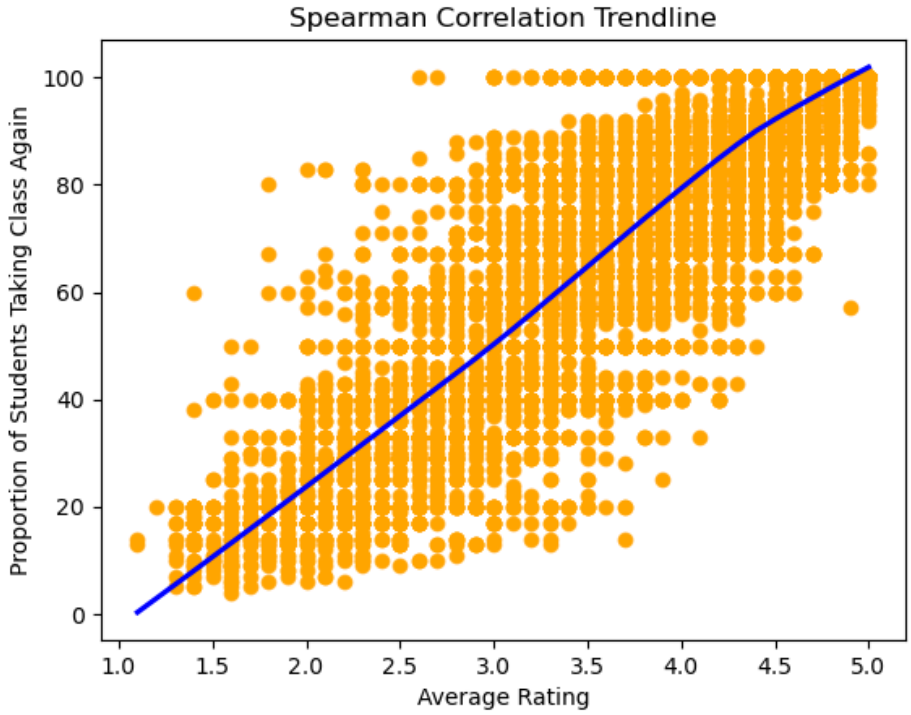

Figure 6: Scatterplot for average rating and proportion of students taking class with the professor again with non linear trendline for a Spearman correlation

6: Do professors who are "hot" receive higher ratings than those who are not?

I plotted histograms for the distribution of ratings for "hot" and "not hot" professors, and found neither of these distributions to be normal (both skewed left). I therefore used a one sided Mann-Whitney U Test for this analysis and found a p-value of 0 as well as an effect size of 0.42. My conclusion is that average ratings for "hot" professors are significantly and practically higher than that of "not hot" professors. This is made clear in the following figure.

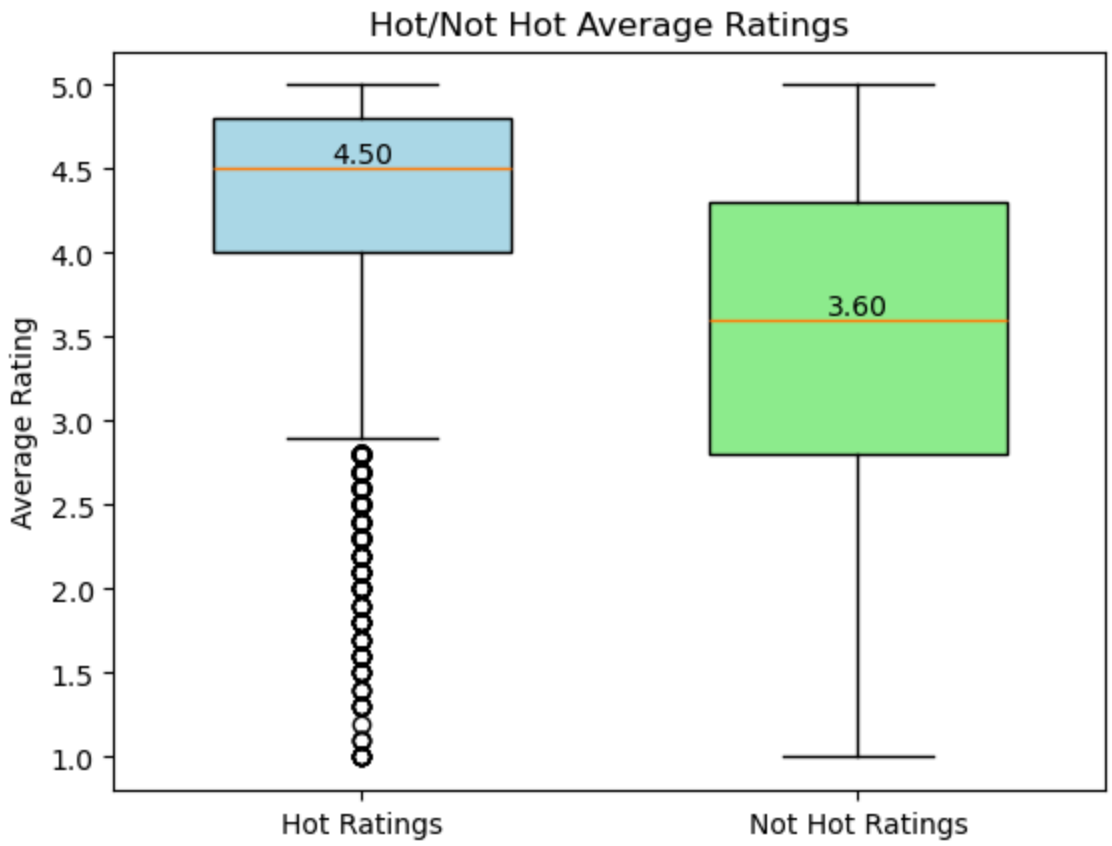

Figure 7: Boxplot of resulting Mann-Whitney U Test for differences in the average rating by "hot" and "not hot" professors.

7: Build a regression model predicting average rating from difficulty (only). Make sure to include the R^2 and RMSE of this model

After creating a dataframe containing the columns for average difficulty and average rating, I dropped NA rows in order to have an identical number of rows for the regression model. After running the model, I retrieved an R^2 value of 0.35, an RMSE of 0.8, and a beta of -0.69. Given the R^2 of 0.35, average difficulty alone doesn't fully explain the variance in average rating.

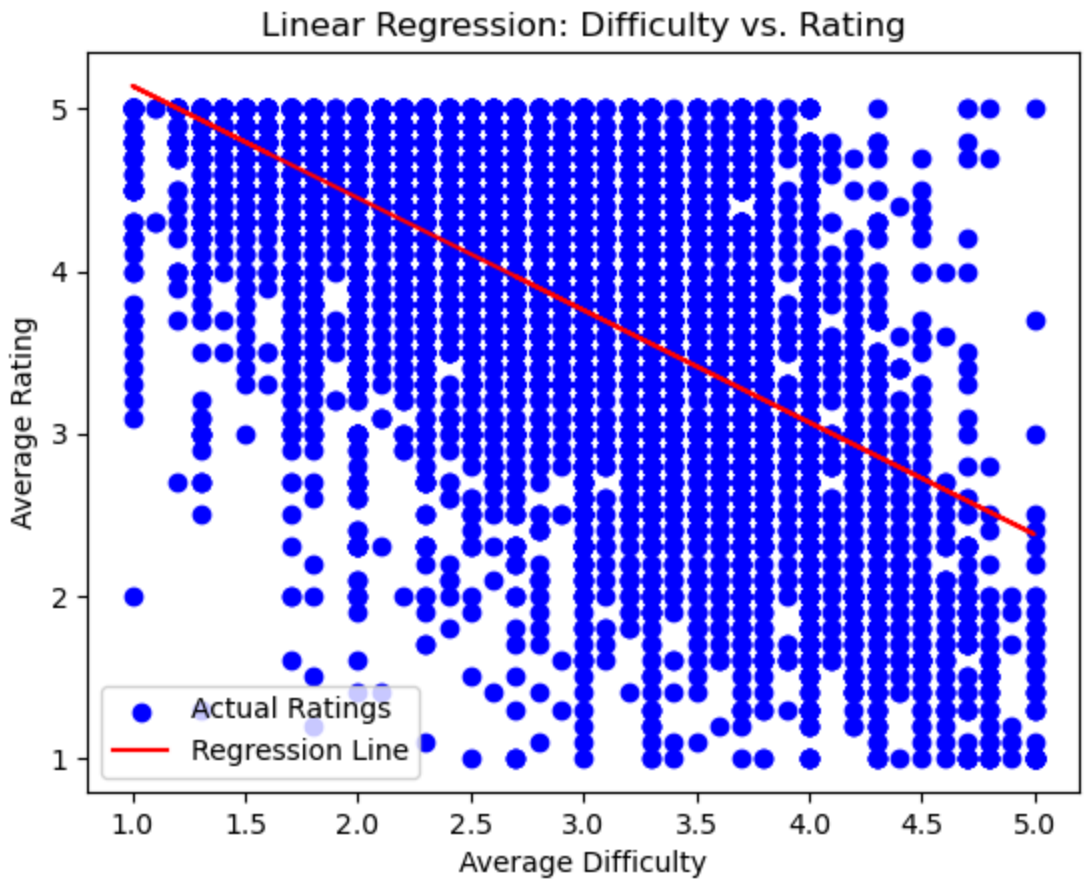

Figure 8: Linear regression model predicting average rating from average difficulty

8: Build a regression model predicting average rating from all available factors. Make sure to include the R^2 and RMSE of this model. Comment on how this model compares to the "difficulty only" model and on individual betas.

After building a regression model using all available quantitative factors, I retrieved an R^2 of 0.81 and an RMSE of 0.38. The multiple linear regression model clearly does a better job of predicting average rating using all factors than regular linear regression does by using average difficulty alone. This is exemplified by the higher R^2 and lower RMSE in the multiple regression model. As shown below, the significant features based on the p-value included the proportion of students that would take the class again with the professor, average difficulty, whether or not the professor received a pepper, and whether or not the professor was male or female. Of these significant features, the proportion of students willing to take the class again with the professor had the biggest impact on average rating, as it had the largest absolute beta value of 0.63. Average difficulty and whether or not the professor received a pepper had the next two biggest impacts on average ratings, with absolute betas of 0.14 and 0.1, respectively. Lastly, after checking for collinearity, I found that none of the features had a variance inflation factor of more than 2, indicating none of the features were

collinear.

Index	Feature	Beta	p-value	Abs Beta
x4	proportion…	0.625931	0	0.625931
x1	average difficulty	−0.1443…	2.43375e−221	0.144375
x3	received a pepper?	0.100257	4.09227e−120	0.100257
x6	male gender	0.0287799	1.19091e−09	0.0287799
x7	female gender	0.0163453	0.000550155	0.0163453
x2	number of ratings	−0.0026…	0.485342	0.00262786
x5	number of …	−0.0007…	0.8452	0.000727554

Figure 9: Resulting betas, p-values, and absolute betas after multiple linear regression predicting average rating from all available factors

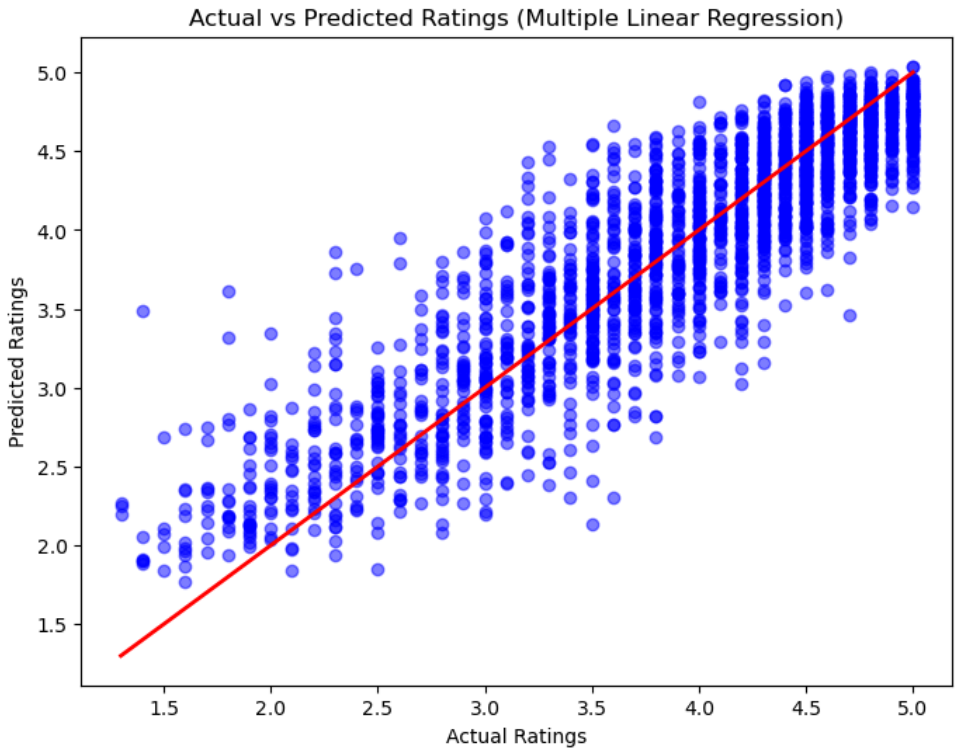

Figure 10: Scatterplot showing actual ratings and predicted ratings after multiple linear regression

Feature	VIF
average difficulty	1.368320
number of ratings	1.023733
received a pepper?	1.256734
proportion of students that would take class a...	1.605408
number of ratings from online classes	1.014908
male gender	1.569726
female gender	1.569149

Figure 11: Result from Variance Inflation Factor Test to check for collinearity

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 imbalance.

Using a logistic regression model to predict whether a professor receives a pepper from average rating, I found an AU(RO)C of 0.75 and a precision of 0.56. Given the class imbalance (24,555 didn't receive a pepper while 14,750 did), I chose to make the weight classes balanced during logistic regression, which scaled the loss function inversely to the class frequencies.

```
model = LogisticRegression(class_weight='balanced', random_state=42)
```

Figure 12: Handling of class imbalance during Logistic Regression

This scales the loss function inversely to the class frequencies:

$$\text{Weight for a class} = \frac{\text{Total Samples}}{2 \times \text{Number of Samples in Class}}$$

Figure 13: Equation demonstrating new weights for each class after imbalance was handled

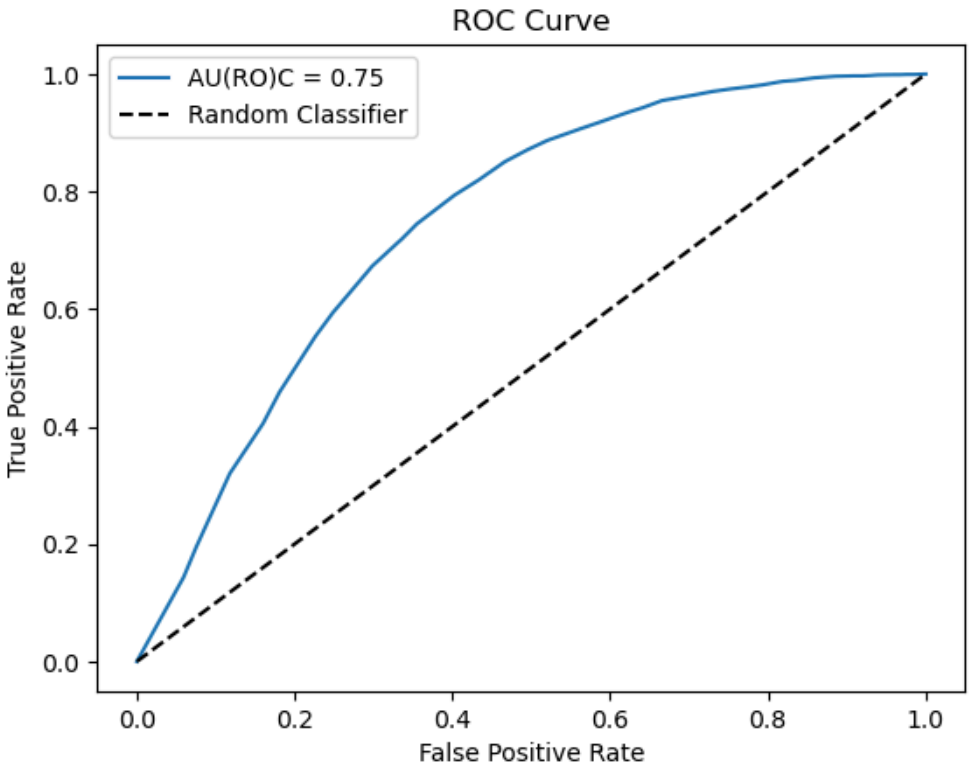

Figure 14: AUROC plot for logistic regression using average rating as sole predictor

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(RO)C and also address class imbalances.

Using a logistic regression model to predict whether a professor receives a pepper from average rating, I found an AU(RO)C of 0.79 and a precision of 0.66. This logistic regression model handles class weights in the same manner as the previous question.

Both logistic regression models are pretty good at correctly classifying whether a professor receives a pepper from the given factor/factors given the AU(RO)C. The second model is slightly better at this classification with an AU(RO)C of 0.79 as opposed to 0.75. The second model is also slightly more precise with a precision of 0.66 as opposed to 0.56. Ultimately, however, both models perform similarly, indicating that average rating alone is a pretty good predictor for the classification.

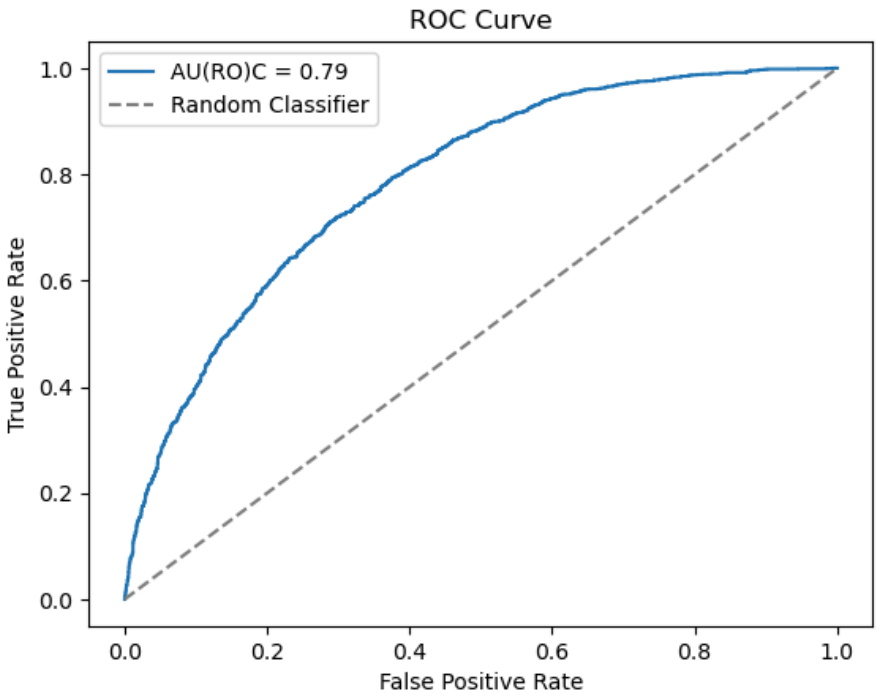

Figure 15: AUROC plot for logistic regression using all available factors as predictor

Extra Credit

I wanted to investigate if ratings for chemistry professors are significantly lower than that for psychology professors. To accomplish this, I investigated the distribution of ratings for chemistry and psychology professors after cleaning the data (as mentioned on Page 1). I found that these ratings were not normally distributed and had significantly different variances, so I performed a Mann Whitney U Test. This test yielded a p-value of 2.76e-24 and an effect size (using Cohen's D) of -0.61. This indicates that chemistry professors have significantly lower ratings than psychology professors do, and that this difference is not only practical but also moderately large in size.

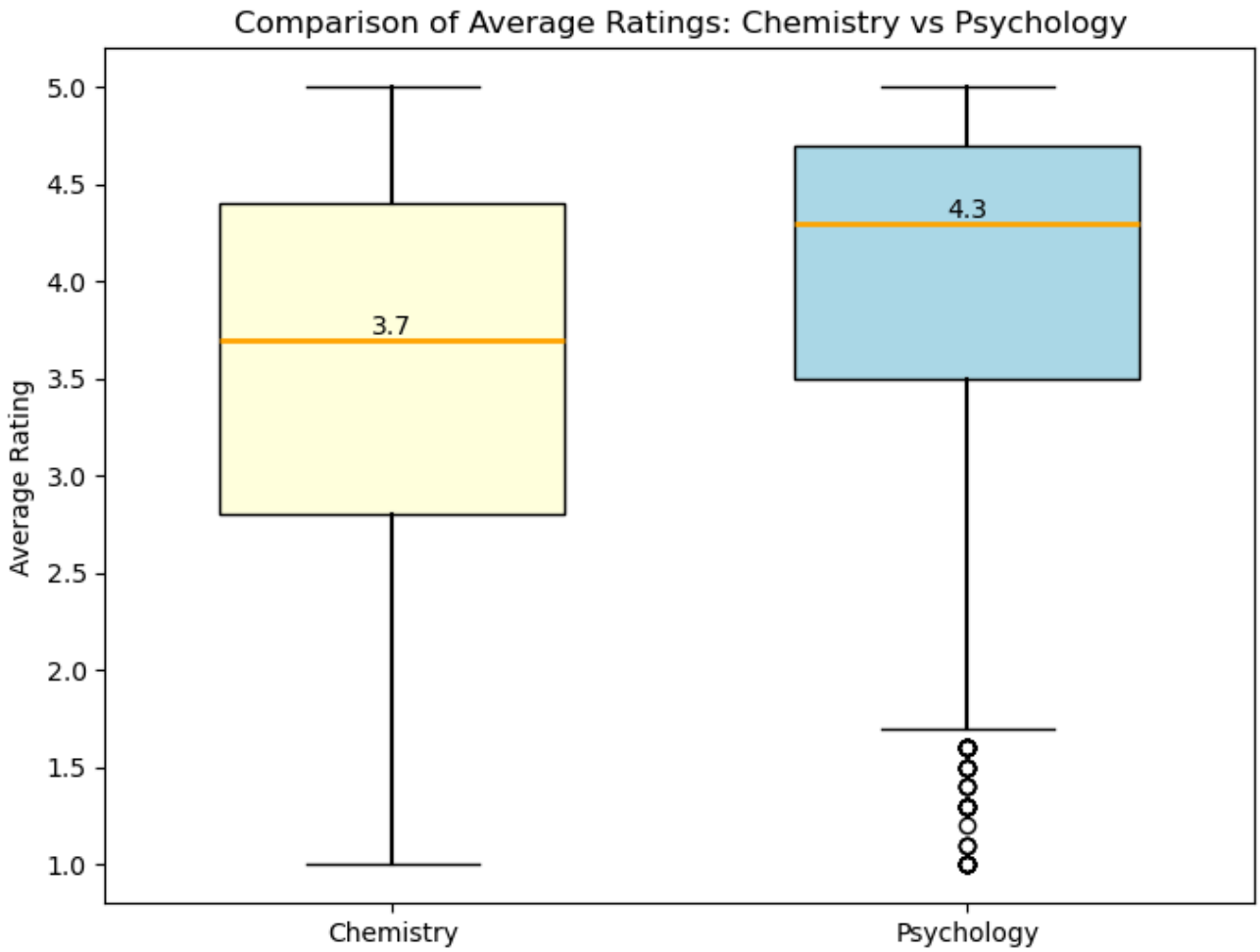

Figure 16: Boxplot of resulting Mann-Whitney U Test for differences in the average rating by chemistry and psychology professors.