Student Name: Owen Huang, NetID: yh4842, N Number: N17864220
DS112 Capstone Project Report
Principles of Data Science Spring 2025

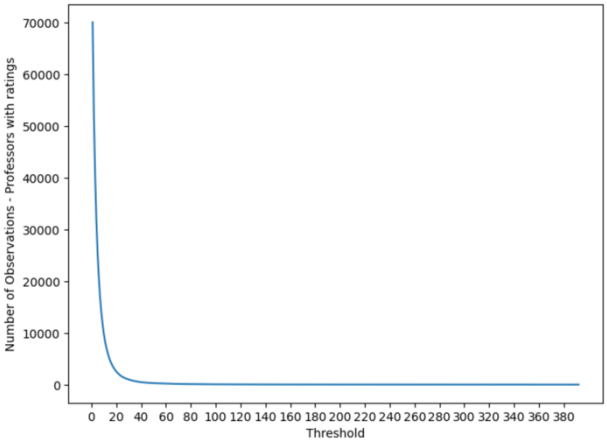
Note: The extra pages are due to extra figures and plots.

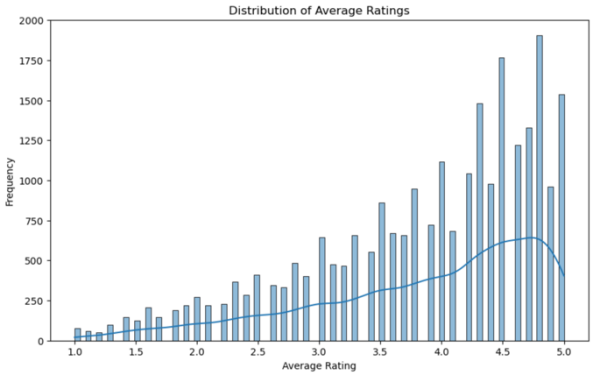
Preprocessing:

1. Random Number Generator: I set my seed to 17864220 and used this number for the train-test split and bootstrap.
2. Handling Missing Values: 86.4% of the data about "The proportion of students who said they would take the class again" is missing. I chose not to impute this data but instead dropped it when the regression/classification model uses that variable as a predictor because the percentage of missing values is very high, and imputing it will lead to inaccurate data and analysis. However, when "proportion" is not used for analysis and there is a null value for the proportion column for a professor, but there is data for all other variables, this row will still be kept since other data are still valid. There are missing data points for some of the Professors (22.1%). One option is to impute the missing data. However, this will lead to inaccurate data and skew the data distribution. Therefore, these missing values will not be included in the analysis.

```
Missing Percentages (%):
Average Rating
22.125193
Average Difficulty
22.125193
Number of Ratings
22.125193
Received a pepper?
22.125193
The proportion of students that said they would take the class again
86.472807
The number of ratings coming from online classes
22.125193
Male Gender
0.000000
Female Gender
0.000000
```

3. Accuracy of Average Ratings: The more ratings there are, the more accurate the average ratings are. Since some professors only have a few ratings, we will need to exclude them because they might not be accurate and could be an outlier if that Professor has received more ratings. Therefore, we would need to set a threshold for the minimum number of data points required to include the average rating in the analysis. However, there is a trade-off. If we only include professors with a high number of ratings, the number of professors decreases dramatically, as evident from Figure 1, decreasing the power when conducting hypothesis testing. Therefore, a threshold must be determined. Here, we see that the number of data points we have drops significantly (exponentially) when we increase the threshold. Therefore, we still must set a relatively low threshold. I will set it to **3** because it will still give 40528 data points, which helps to retain more statistical power while ensuring the accuracy of the average ratings.

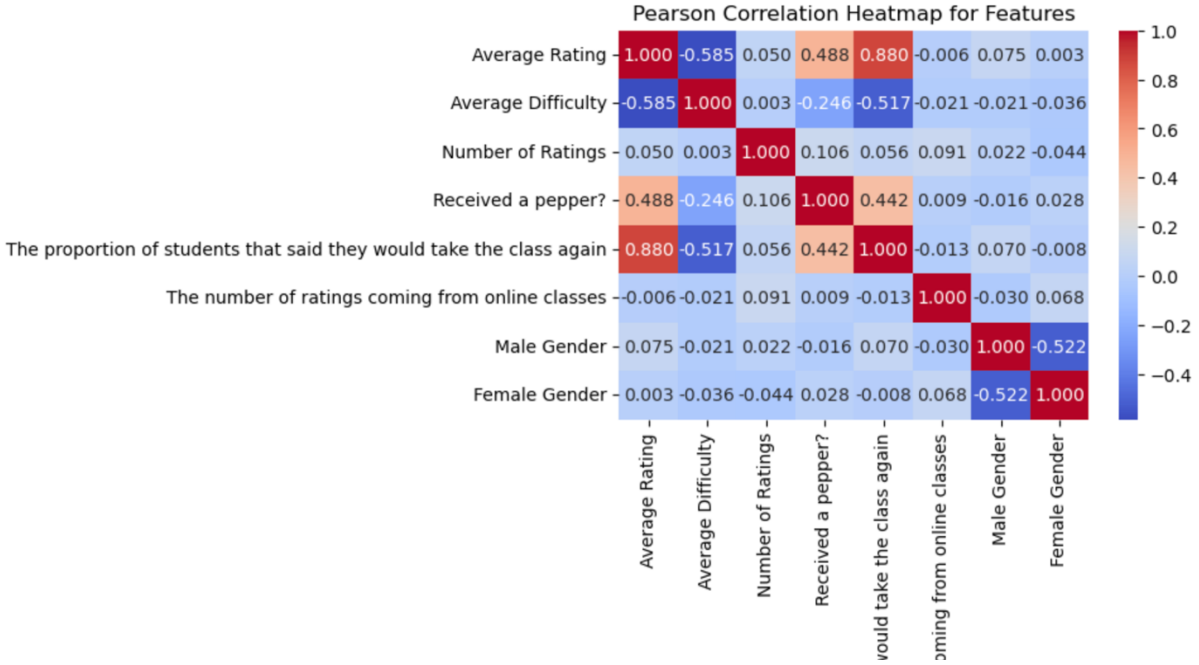




4. Ambiguous Data: Some Professors are deemed as both male and female. Since distinguishing female and male Professors is important for the analysis, such ambiguous cases will not be included when evaluating whether there is pro-male gender bias and when building regression models. However, they will be included when the gender is not involved in the analysis because the Professor is still a professor despite us not knowing their gender, so other metrics are still valid.
5. Checking for normality of data: By examining the distribution of the average ratings, we see that the data is not distributed normally, meaning that a parametric test cannot be used since the assumption of the normality of data is violated.

Checking for Correlation for Regression:

1. I used the pandas .corr() to examine the correlation between each independent variable and found out that multiple features are correlated. This multicollinearity concern will be addressed when performing regression and classification. (Below is a heatmap using seaborn)



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

My **Null Hypothesis** is that there is no difference between the median of the average ratings of male and female Professors, and their respective ratings are from the same distribution. (No pro-male gender bias) The **Alternative Hypothesis** is that the median average ratings of male Professors are higher than those of female Professors. (There is pro-male gender bias as male professors have higher average ratings than female professors)

A significance test will first be performed to determine whether there is evidence of a pro-male gender bias in this dataset. Then, I will use multiple regression to determine if the higher ratings associated with male professors when compared to that of the female professors is caused by other confounders. I will use the **Mann-Whitney U-Test** because the movie ratings are not normally distributed, and movie ratings are ordinal data and not on a ratio scale. So, the mean cannot be interpreted meaningfully. These are all assumptions of parametric test and therefore parametric test cannot be used. Instead, I will use the nonparametric test of Mann-Whitney U test to compare the medians of ratings of male professors and those of female professors.

I used a one-tailed (right tailed/greater) Mann Whitney U test to determine if the male ratings come from a distribution of ratings that has a higher median than median of the distribution of the female ratings. I obtained a **p-value of 0.0000206** which is smaller than the significance level alpha of 0.005, suggesting that the test-statistic is **statistically significant** and we could drop (reject) the null hypothesis that there is no difference between the average rating of male and female professors and conclude that male professors receive higher average ratings than female professors and there is pro-male gender bias.

```
Mann-Whitney U Test Statistic: 113037784.5
The p-value is 2.062207761371274e-05, which
```

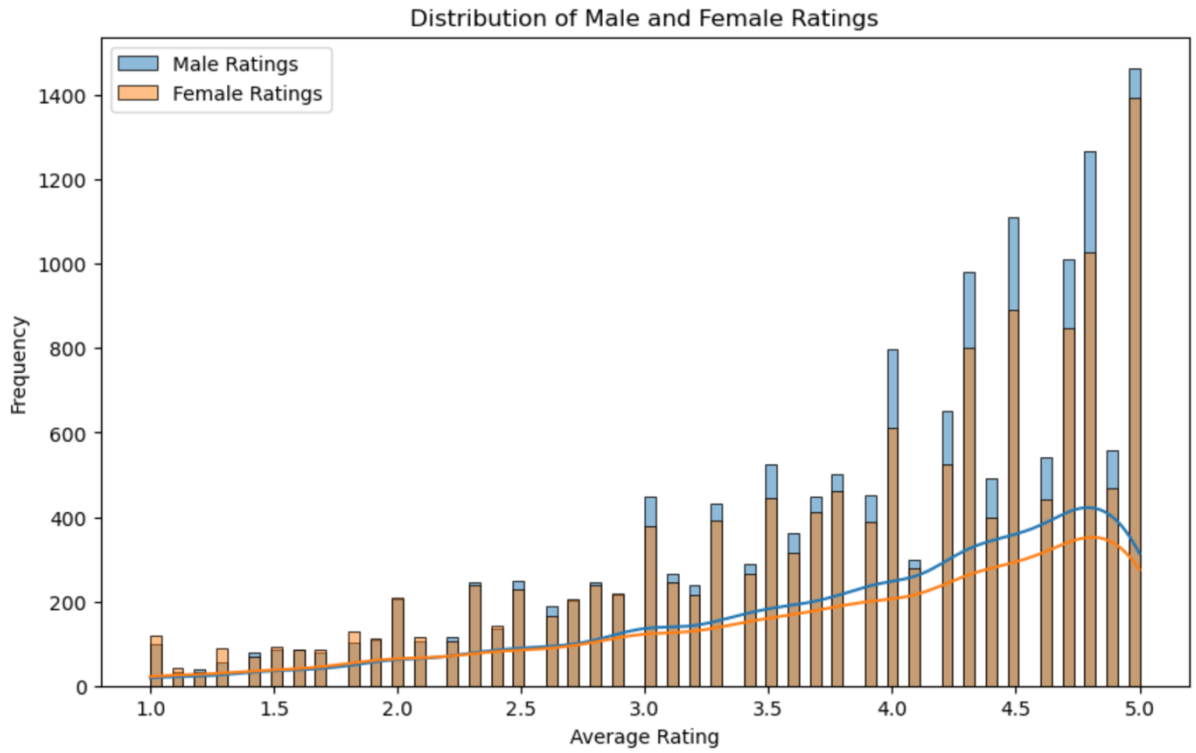
To make sure that the male and female ratings come from a different distribution, I also used Two-Sample Kolmogorov-Smirnov Test and obtained a p-value of 0.0000096, which is below the 0.005 threshold for statistical significance, hence we could drop the null hypothesis that male and female ratings come from the same distribution and conclude that they come from different distributions.

```
KS Test statistic: 0.02873155881664191
p-value: 9.649288997265509e-06
```

However, there could be confounders that lead to higher ratings for male professors. To address the possibility of confounders, I used multiple regression (with the outcome variable Y as "average ratings") because it isolates the effect of one variable on the dependent variable from other independent variables by holding other independent variables constant. Female Gender and Male Gender are dummy variables, so I created two multiple regression models, each with one of the genders. I see that the **coefficient of "Female Gender" is –0.012** whereas the **coefficient of "Male Gender" is +0.012**, suggesting that the male gender leads to an increase in the average rating and the converse for female gender, while holding all other variables constant respectively. However, the effect is relatively small.

Therefore, we conclude that there is statistically significant evidence of pro-male gender bias in this dataset since they receive statistically significantly higher ratings than female professors.

```
Male Gender   Female Gender
0.012107      -0.012107
```



From the distribution of male and female ratings, we also see that male professor ratings have more high ratings than female professor ratings (distribution of ratings of male professors is to the right of that of female professors). However, the increase in average rating is quite small, as the median rating of male professors is 4.2, and that of female professors is 4.1. Overall, there is still a pro-male gender bias, just a relatively small one.

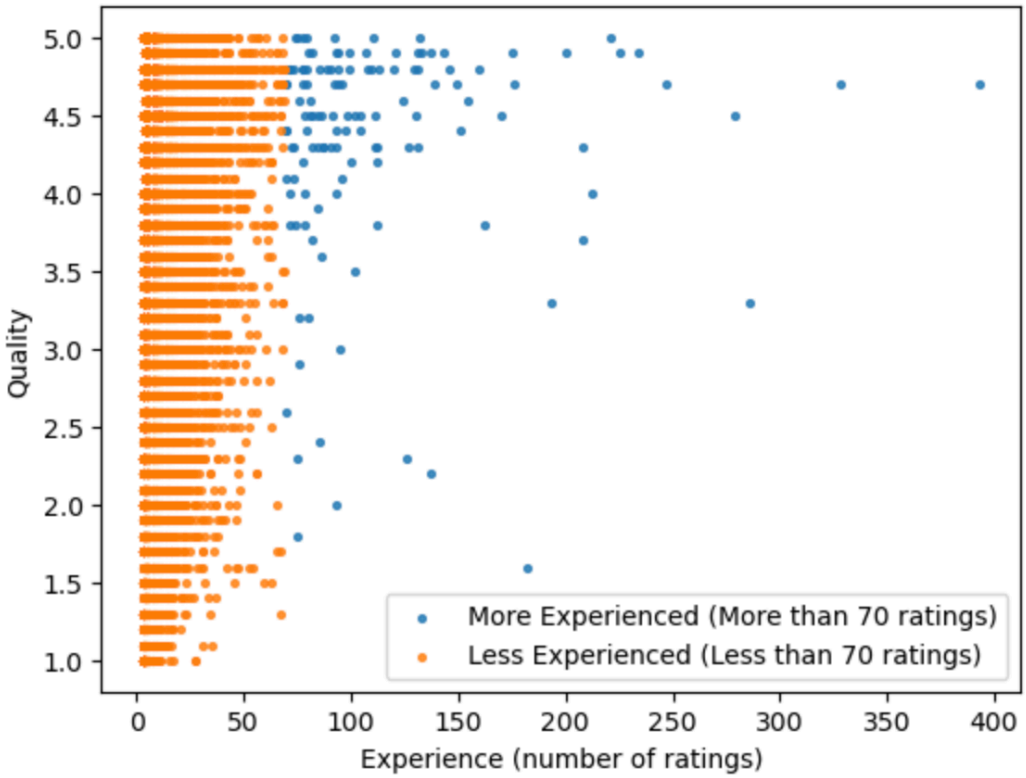
Question 2. : Is there an effect of experience on the quality of teaching? Yes – But depends on the threshold of high and low experience.

The Null Hypothesis is that teaching experience does not affect the quality of teaching.

The Alternative Hypothesis is that teaching experience affects the quality of teaching.

First, I plotted the quality of the teaching (average rating) against the experience of the Professor (number of ratings). We see that when the number of ratings is below about 70, there is a significant variability and inconsistency of their ratings, and only when the number of ratings exceeds 70 do we start to see a trend of high ratings. First, I performed a two-tailed U-test on all the ratings and obtained a p-value of 9.63 ^(-10), less than the significance threshold alpha of 0.005. Hence, we could drop (reject) the null hypothesis and conclude that there is an effect of experience on the quality of teaching. I did not choose the threshold to characterize a professor has more experienced using the mean or the median of all the ratings because that would give either a 6 using the median or 8.3 using the mean, and it does not seem quite reasonable to assume that Professors with more than 6 or 8.3 of ratings is considered more experienced. To verify this, I also performed a two-tailed U-test using the median and obtained a p-value of 0.93, much larger than 0.005. Therefore, there is an effect of

experience on the quality of teaching, and more specifically, more experience appears to lead to a higher quality of teaching, although only starting from a certain point of around 70 ratings.

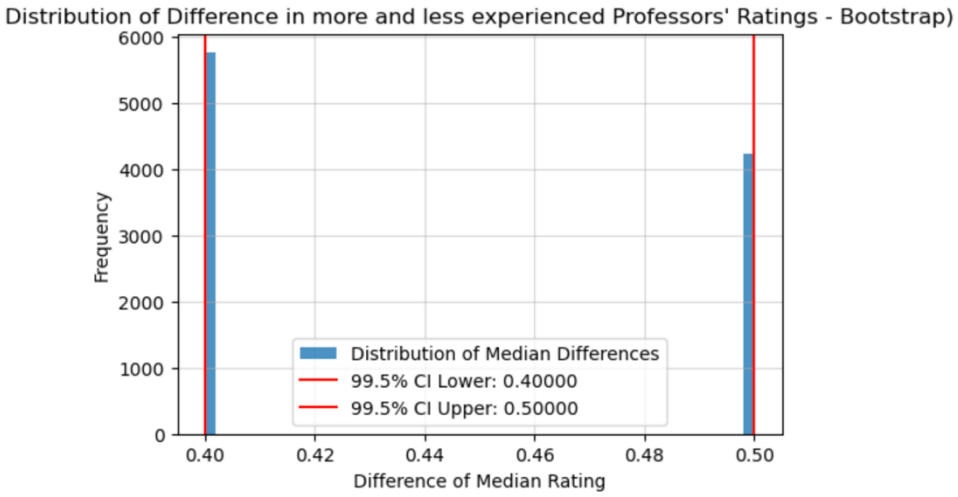



```
filtered_exp_quality_greaterthan70 = exp_quality[exp_quality["Experience"] >= 70] # First Group - More Experienced
filtered_exp_quality_lessthan70 = exp_quality[exp_quality["Experience"] < 70] # Second Group - Less Experienced
```

```
# U Test on the two groups
u_stat70, p_value70 = mannwhitneyu(filtered_exp_quality_greaterthan70["quality"], filtered_exp_quality_lessthan70["quality"], alternative="greater")
print(f"u_stat_70:{u_stat70}")
print(f"p_value_70:{p_value70}")

# NOW USE MEDIAN - Result: No Longer Statistically Significant
filtered_exp_quality_greaterthanMedian = exp_quality[exp_quality["Experience"] >= np.nanmedian(experience)] # First Group - "More Experienced"
filtered_exp_quality_lessthanMedian = exp_quality[exp_quality["Experience"] < np.nanmedian(experience)] # Second Group - "Less Experienced"
u_statMedian, p_valueMedian = mannwhitneyu(filtered_exp_quality_greaterthanMedian["quality"], filtered_exp_quality_lessthanMedian["quality"], alternative="greater"
print(f"u_stat_median:{u_statMedian}")
print(f"p_value_median:{p_valueMedian}")
```

```
u_stat_70:3378125.0
p_value_70:9.631315319882235e-10
u_stat_median:203554353.0
p_value_median:0.9271932798551544
```

To further determine whether there is an effect, I also used a bootstrap by computing the medians of the two groups mentioned above and taking the difference. Since the lower confidence interval does not touch 0, this indicates a statistically significant result, further reinforces the result of the u-test, and suggests that there is a statistically significant effect of experience on the quality of teaching. The small width of the confidence interval of 0.1 also suggests high power and a precise estimate of the true median difference.

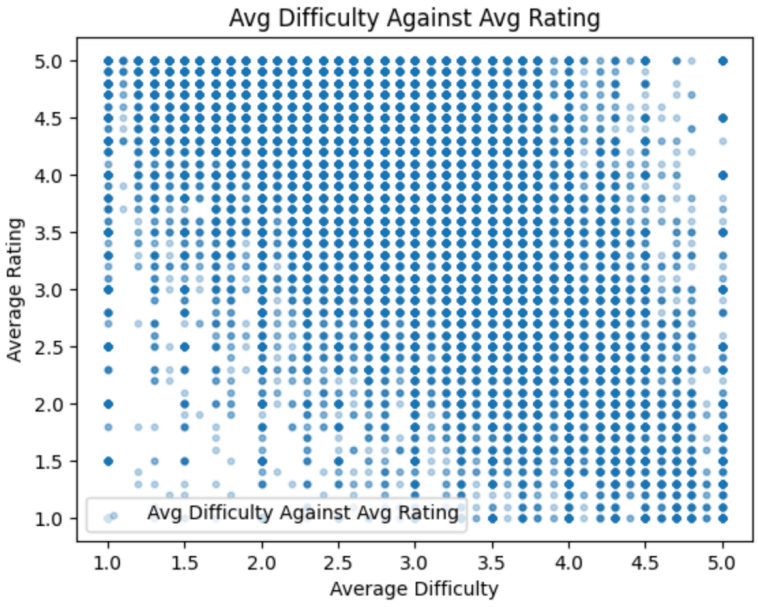


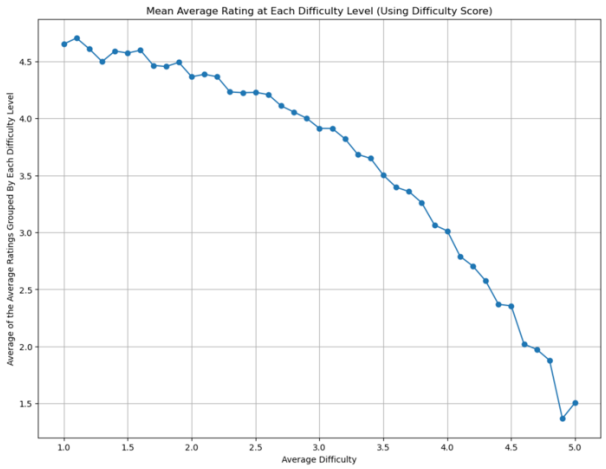
Question 3. : What is the relationship between average rating and average difficulty? Moderate negative monotonic relationship.

I used Spearman correlation instead of Pearson correlation to determine the relationship between average rating and average difficulty because the data is ordinal. **The Spearman correlation is -0.57**, **with a p-value of 0,** which is lower than the 0.005 threshold, indicating **a moderate negative monotonic relationship between "average rating" and "average difficulty."** Treatment of Null Values: I did not remove the rows where the gender of the Professors is ambiguous because the average difficulty and average rating associated with those professors are still valid. In the graph below, from the density of the data points, we see that in general, a low difficulty is associated with a high average rating, and a high average difficulty is associated with a low average rating.

```
Q3. spearman correlation: -0.5710048684852826
p_value: 0.0
```

```
correlation, p_value = spearmanr(rating_and_difficulty["Average Difficulty"], rating_and_difficulty["Average Rating"])
```





Here we see that the average of the average ratings at each difficulty level decreases; This is done using the groupby function.

```
mean_ratings = rating_and_difficulty.groupby("Average Difficulty")["Average Rating"].mean().reset_index()
```

<u>Question</u> 4. : Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't? Yes.

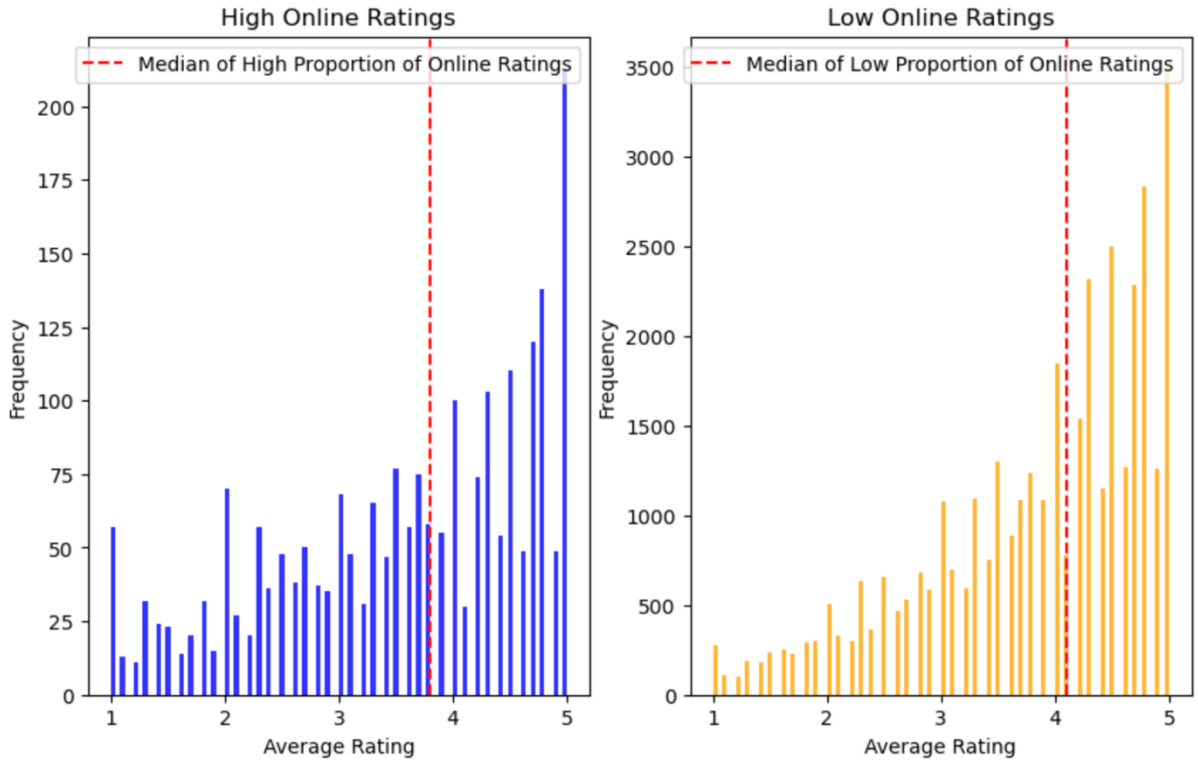
The Null Hypothesis is that the rating distribution of Professors who teach a lot of online classes is the same as that of those who don't. **The Alternative Hypothesis** is that Professors who teach a lot of online classes have lower ratings than those who don't.

Here, I will define a lot of classes in the online modality as when the number of ratings from online classes is greater than 50% of the total number of ratings. I used the u-test because the groups are independent, and the data is ordinal and not normally distributed. Here, we will use two one-tailed test to see if professors who teach a large proportion of classes in the online modality. This is done by creating two groups, where one contains professors with online rating proportion that is greater than or equal to 50% and the other group with less than 50% of online ratings. Using the left-tailed one-tailed U-test, I obtained a p-value of 1.20 × 10^ (-24), far below the significance threshold of 0.005. Therefore, we drop/reject the null hypothesis and conclude that professors teaching online classes have statistically significantly lower ratings than those who don't.

```
u_stat_less_onetail, p_value = mannwhitneyu(high_online,low_online, alternative="less")
```

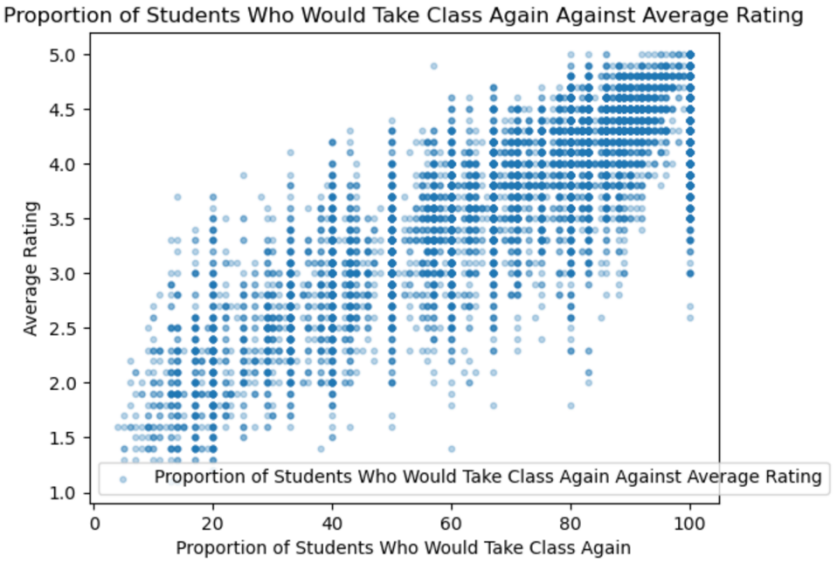
```
high_online = online_df[online_df["Proportion of Ratings from Online Classes"]>=0.5]["Average Rating"]
low_online = online_df[online_df["Proportion of Ratings from Online Classes"]<0.5]["Average Rating"]
```

```
Q4. One-tailed (Left tailed) u-statistic: 38084091.5
p_value:1.2095023974033279e-24
```



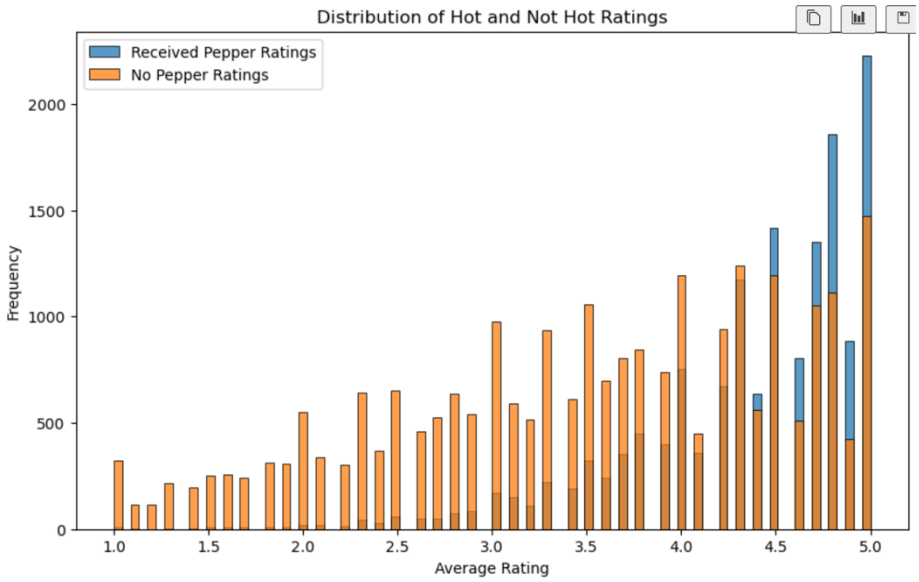
Question 5. Since the movie rating is ordinal and not on ratio or interval scale, we cannot assume a linear relationship between the average rating and the proportion of people who would take the class again. So I used **spearman's correlation** to determine the monotonic relationship between the average rating and the proportion of people who would take the class the professor teaches again. **I obtained a coefficient of 0.85 and a p-value of 0, which is below 0.005,** suggesting that there is a strong monotonic relationship between average rating and the proportion of people who would take the class again with this professor.

```
Q5. Spearman Correlation: 0.8521896777710094
p-value: 0.0
```



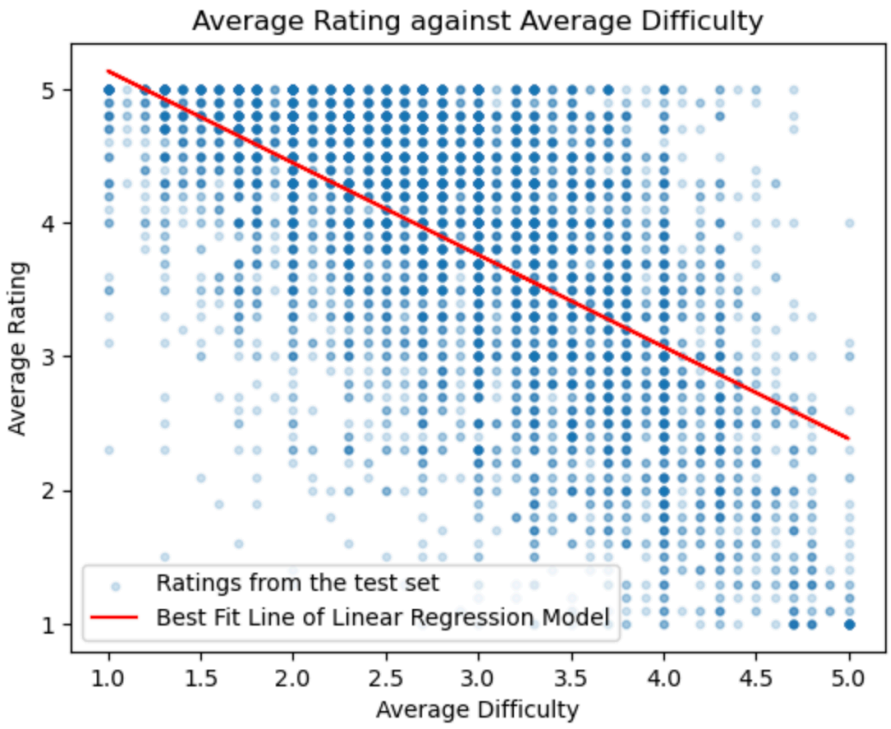
Question 6. I used a Mann Whitney u test here again for the same reasons. **The Null Hypothesis** is that Professors who receive a "pepper" do not have higher average ratings than those who don't. **The Alternative Hypothesis** is that Professors who receive a "pepper" have higher average ratings than those who don't receive a "pepper." Again, a Mann-Whitney U test is used, and **I obtained a p-value of 0, which is below the threshold of significance of 0.005.** Hence, we could **drop/reject the null hypothesis** and conclude that the Professors who receive a "pepper" do have a higher average rating than those who don't. From the histogram below, we also see that the professors with a "pepper" have higher ratings in general than those who don't – Frequency of average rating > 4.5 is higher.

```
Q6. u-test statistic: 289671385.0
p-value: 0.0
```



Question 7. Even though there are ambiguous genders and professors with null values associated to proportion that would retake the class column, we will keep the other data associated with these professors because they we only use the average difficulty to predict the average rating. I used a train test split of 80 to 20 as it is the convention. I obtained a RMSE (average distance between predicted and actual average ratings) of 0.79, R-squared of 0.36 (Percentage of variance in the outcome variable that the model can explain). The coefficient for Average Difficulty is -0.69, indicating a strong negative relationship between difficulty and rating, and this means that one unit increase in average difficulty leads to a 0.69 decrease in average rating, on average. In this model I kept all data including the ambiguous gender because we only use average difficulty as a predictor.

```
RMSE: 0.79302912905075
R-squared: 0.36273863930774886
coefficient of Average Difficulty: [-0.68742938]
```



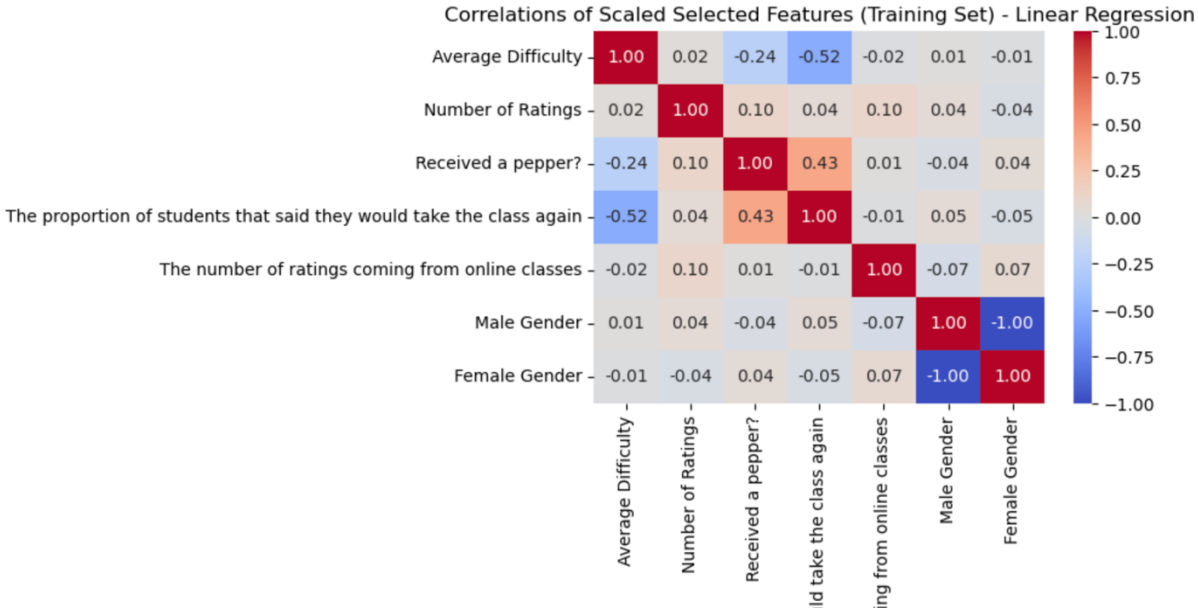
Question 8. Comment on Missing Value Handling Here: Since there are ambiguous genders here and we are using gender to perform regression analysis, I removed all ratings/data where the gender is ambiguous by filtering as explained in the preprocessing part of the report, unlike in question 7 where I have kept the ambiguous genders. **Examining Correlation:** we examine whether there are correlated features with a heatmap. Through the heatmap (provided below), I see that average difficulty and the proportion of students who would take the class again has a correlation coefficient of –0.52, and pepper and "proportion who would take again" is 0.43, so there is moderate amount of multicollinearity. To address this problem, I used both lasso and ridge **regression** because they shrink the coefficients of correlated independent variables by adding a penalty term. The penalty term of alpha was chosen using the ridgeCV and lassoCV packages which automatically picks the best alpha. Therefore, addressing the problem of multicollinearity. Interpretation of coefficients (betas): 1. The result shows that "The proportion of students that said they would take the class again" is the strongest predictor for the average rating with a coefficient of 0.62. 2. The coefficient of average difficulty is –0.15, suggesting that a harder course is associated with a lower rating. 3. The coefficient of "Received a pepper" is 0.1, suggesting that it does boost ratings a little. 4. The coefficient of "Male Gender" is 0.012, suggesting that male professors do get higher ratings. The coefficient of number of classes is –0.006 so there is a negligible impact which makes sense because it is the overall proportion of online ratings that matter more. Lastly, the coefficient of the number of ratings is about –0.004 which suggests that the number of ratings has a negligible impact on average rating. I set the random state to my N number, and this makes the coefficient of number of ratings negative - I ran it without specifying it, and sometimes it yields a positive coefficient. But, nonetheless the absolute value of the coefficient is still small – negligible. Note: these coefficients all mean that for one unit increase in the independent variable we get a change in the dependent variable equal to the size of the coefficient, on average, holding all other variables constant.

Comparison to the Difficulty Only Model: The RMSE has decreased from to 0.79 to 0.36 (about –54% decrease) and R^2 has increased from 0.37 to 0.80 (increased by 116%), which suggests that this

model is better as it almost halved the prediction errors and almost doubled the percentage of explained variance.

I also addressed multicollinearity by using linear regression but dropping the dummy variable of "Female Gender" to avoid the dummy variable trap. This is because Male Gender and Female Gender are dummy variables, because if one of them is 1, then the other is 0, so I can drop the "Female Gender" as they are perfectly negatively correlated. The resulting coefficients and performance metrics are similar to the ridge and lasso regression ones, suggesting that the problem of multicollinearity is not that huge. This is reasonable since, other than the perfectly negatively correlated Male and Female Gender categorical variables, the highest correlation coefficient is the absolute value of – 0.52, which is not that high (is moderate), and the second highest pairwise correlation is 0.43, which is also moderate.

All in all, the proportion of students who would take again and the average difficulty, followed by whether they get "pepper," leads to most of the variation in the outcome variable of average rating.

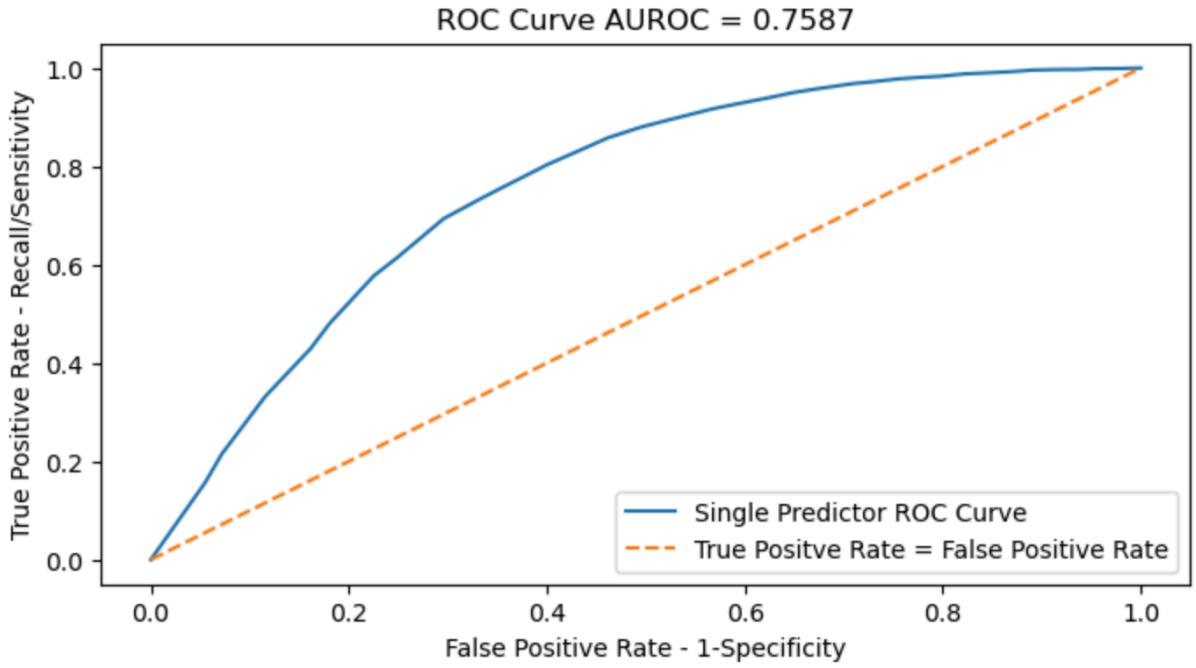



```
Best alpha for Lasso: 0.0007
RMSE with all factors - LassoCV: 0.3633
R-squared with all factors - LassoCV: 0.8031
LassoCV Coefficients:
Average Difficulty                                                -1.466565e-01
Number of Ratings                                                 -2.845517e-03
Received a pepper?                                                 9.787835e-02
The proportion of students that said they would take the class again    6.168135e-01
The number of ratings coming from online classes                  -5.588163e-03
Male Gender                                                        1.136249e-02
Female Gender                                                     -1.568333e-17
dtype: float64
LassoCV Intercept: 3.9719

Best alpha for Ridge: 10.0000
RMSE with all factors - RidgeCV: 0.3633
R-squared with all factors - RidgeCV: 0.8031
RidgeCV Coefficients (best alpha):
Average Difficulty                                                -0.147474
Number of Ratings                                                 -0.003579
Received a pepper?                                                 0.098806
The proportion of students that said they would take the class again    0.615841
The number of ratings coming from online classes                  -0.006233
Male Gender                                                        0.006082
Female Gender                                                     -0.006082
dtype: float64
RidgeCV Intercept: 3.9719
```

```
RMSE with all factors - Linear Regression: 0.3633
R-squared with all factors - Linear Regression: 0.8030
Linear Regression Coefficients:
Average Difficulty                                                  -0.147140
Number of Ratings                                                   -0.003611
Received a pepper?                                                   0.098520
The proportion of students that said they would take the class again  0.617011
The number of ratings coming from online classes                    -0.006215
Male Gender                                                          0.012107
dtype: float64
Linear Regression Intercept: 3.9719
```

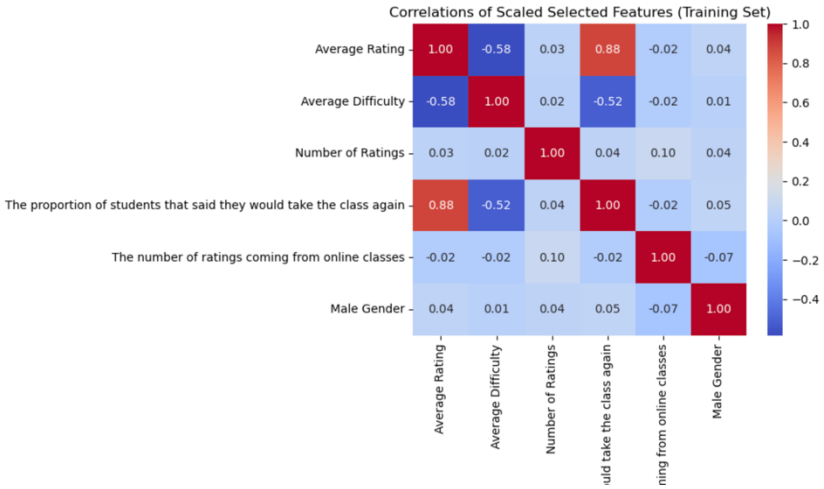
Question 9. The class (0, 1) = (no pepper, pepper) is split into 62% and 38%, respectively. This means there is some class imbalance; therefore, we should use AUC to evaluate instead of simply focusing on the model's accuracy. We could also use precision-recall curve to further address class imbalance because it focuses on the positive (1) class only, as it equals (TP/(FP+TP)), but this is not covered in class and is not needed. I split the data using a train-test split ratio of 80-20 by convention. Also, I used the built-in parameter in logistic regression to specify "class-weight" as balanced so that the model adjust according to the class frequencies. I obtained an AU(RO)C score of 0.76, meaning the model is better than classifying with chance alone and is moderately well in differentiating between positive and negative classes (so it is quite good at classification). Looking at the confusion matrix, we see that the model is better at predicting "no pepper" (precision of 0.81) than at predicting "pepper" (precision of 0.56). However, the recall or true positive rate is higher for class 1, so the model identifies 75% of the actual class 1 (pepper) cases. The accuracy is just 0.69 but this is an understatement because there is more actual pepper than actual no pepper. Therefore, AU(RO)C is used too.

```
Received a pepper?
0.0    25282
1.0    15246
Name: count, dtype: int64
auc score: 0.7587488469895518
Classification Report:
              precision    recall  f1-score   support

         0.0       0.81      0.65      0.72      5061
         1.0       0.56      0.75      0.64      3045

    accuracy                           0.69      8106
   macro avg       0.69      0.70      0.68      8106
weighted avg       0.72      0.69      0.69      8106
```



Question 10.

Note: For this question I dropped all the data points where the professors' gender is ambiguous, and I also dropped the dummy variable of "Female Gender" because it is perfectly negatively correlated with "Male Gender" so having both is redundant. Therefore, I have six predictors here and below is the correlation matrix.



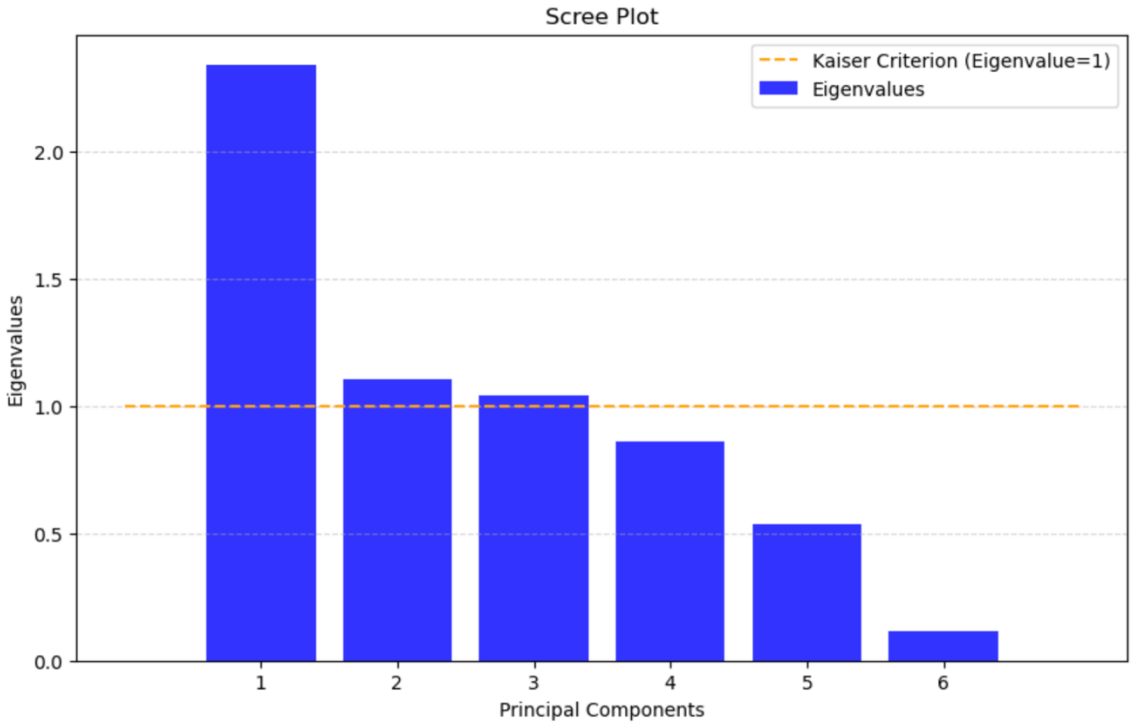
Multicollinearity will negatively impact the reliability of the coefficient estimate (leading to a larger standard error of the coefficients) of a logistic regression model. However, it does not necessarily impact the ability of the model to perform classification correctly, and since we don't have many features, PCA is unnecessary. And if we perform PCA, we sacrifice the ability to explain our model (interpretability). Therefore, I built a logistic regression (train-test split of 80-20) and obtained an AUC score of 0.79, which is only about 0.03 higher than the previous logistic model (Comparison). This suggests that the average rating alone strongly predicts whether the professor receives. However, since the model without PCA only offers a slight improvement, I still attempted to perform PCA to

see if the improvement in the model's AUC score (to address class imbalance – better than accuracy) is justified, as the components after PCA could be harder to interpret. **(PCA is actually worse)**

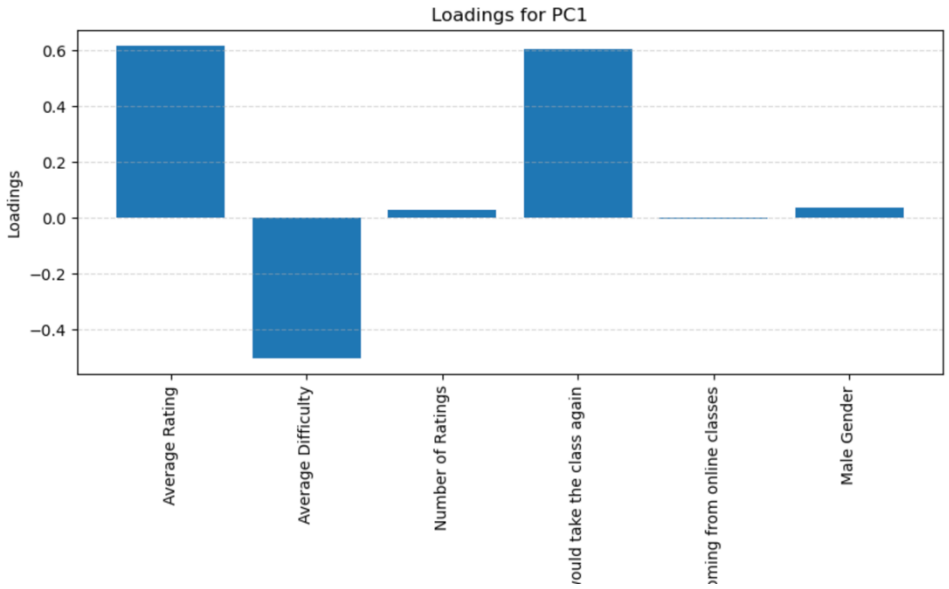
```
Received a pepper?
0.0    4683
1.0    4166
Name: count, dtype: int64
auc score: 0.7862080545053621
Classification Report without PCA:
              precision    recall  f1-score   support

         0.0       0.75      0.65      0.70       891
         1.0       0.69      0.78      0.73       879

    accuracy                           0.72      1770
   macro avg       0.72      0.72      0.72      1770
weighted avg       0.72      0.72      0.72      1770
```

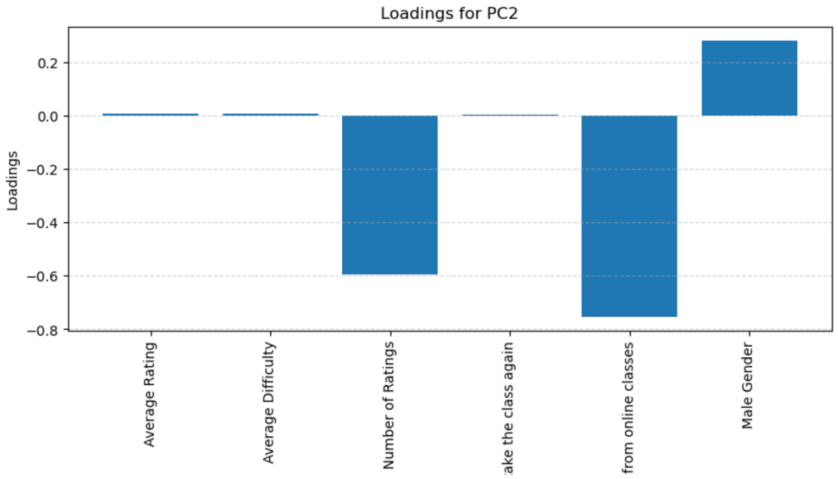
However, I still used PCA to perform feature extraction and to find out why the two models perform so similarly. I first z-scored the data so that they have a mean of 0 and variance of 1. I then checked used the Kaiser Criteria (Components with Eigenvalue Greater than 1) and set the number of principal components to 3 and obtained an AUC score of 0.75, which is lower than even the single predictor model. I then checked using 6 principal components and obtained an AUC score of 0.79 which is exactly the same as the model without PCA **(actually dropped by 0.0001 with PCA but there is some error).** At this point, since all components are used, it would have been better to employ logistic regression without PCA, since we can interpret the model better. There are a few takeaways. First, doing a PCA is not necessary here due to the **zero** increase in AUC score and we sacrifice the ability to interpret the model since other than PC1, which we could interpret as "student satisfaction or how much they like the class," and PC2 as gender, which provides evidence for the pro-male professor bias, PC3 is hard to interpret and PC4 is exactly the same as PC3. Therefore, since there is no model improvement and the components are hard to interpret, PCA is not worthwhile, and it is better to use the logistic classification model without applying PCA. Second, from the Scree Plot and computing the percentage of variance explained by each component, we see that the first component accounts for 39 percent of the variance, and they are made up of average rating and the proportion of students who would take the class again. This further suggests that predicting whether the professor receives pepper using only average rating – the previous logistic model – is close to this model using all variables as predictors.

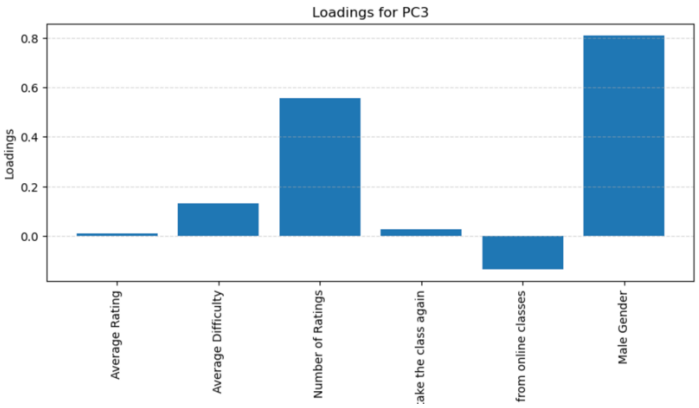
```
Variance Explained: 39.005
Variance Explained: 18.378
Variance Explained: 17.374
Variance Explained: 14.368
Variance Explained: 8.951
Variance Explained: 1.923
```

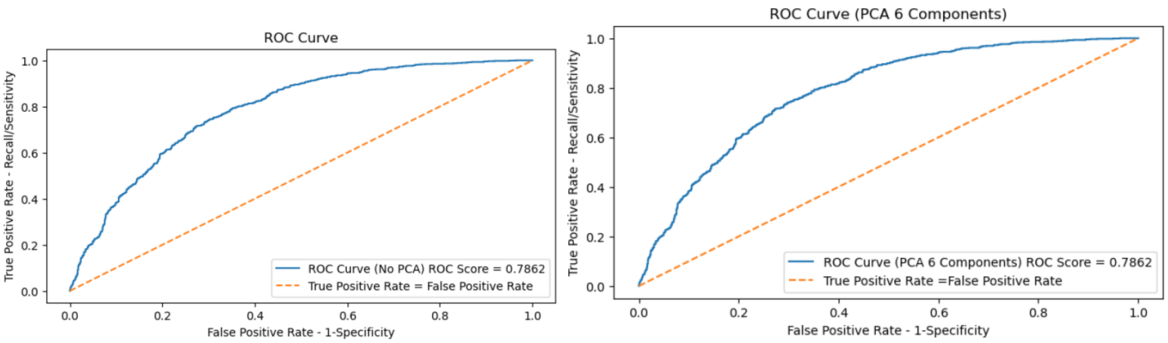



```
Shape of data after final PCA transformation (Train): (7079, 6)
Shape of data after final PCA transformation (Test): (1770, 6)

AUC ROC for PCA
AUROC Score (PCA): 0.7861889020402482
Classification Report (PCA):
              precision    recall  f1-score   support

   No Pepper       0.75      0.65      0.70       891
      Pepper       0.69      0.78      0.73       879

    accuracy                           0.72      1770
   macro avg       0.72      0.72      0.72      1770
weighted avg       0.72      0.72      0.72      1770
```


Loadings for PC1
Loadings
0.6
0.4
0.2
0.0
−0.2
−0.4
Average Rating
Average Difficulty
Number of Ratings
Male Gender


Loadings for PC2
Loadings
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Average Rating
Average Difficulty
Number of Ratings
Male Gender


Loadings for PC3
Loadings
0.8
0.6
0.4
0.2
0.0
Average Rating
Average Difficulty
Number of Ratings
Male Gender



Therefore, there is no difference in ROC score with (Left Graph) and without PCA (Right Graph), and the resulting principal components are hard, if not impossible, to interpret. And actually PCA is worse by 0.0001 than the model without PCA. Therefore, there is no need to use PCA here. However, PCA reveals to us that the average rating alone is a strong predictor for whether a "pepper" is given, explaining the close performance of the single-predictor logistic model and the model using all variables. This is why the model with all variables as predictors only gives a 0.003 increase in AUROC score than the one-predictor model.

Question 11. Extra Credit. I am also studying Economics and Mathematics, and so I wanted to see which states have the highest average ratings for Economics and Mathematics. To do that, I combined the two datasets together and filtered out the ratings for math and economics classes, and used "groupby" to group the average ratings together by states, took the average, and then sorted them. Luckily, New York State appeared in both lists (List of states attached as a figure below).

```
States with the highest average ratings of Math professors:
US State (2 letter abbreviation)
KS    4.80
MN    4.80
SK    4.65
NY    4.60
SC    4.60
Name: Average Rating, dtype: float64
States with the highest average ratings of Economics professors:
US State (2 letter abbreviation)
NY    4.800000
WI    4.700000
VA    4.566667
TX    4.400000
ON    4.250000
Name: Average Rating, dtype: float64
```