Assessing Professor Effectiveness (APE)

Group Members: Deepanshu Mody, Evan Beck, Samarth Agarwal

Group Name: CAP 85

Date of Finalization: 12/09/24

Contributions:

Deepanshu Mody: EDA, Preprocessing, Regression and Classification
Evan Beck: EDA, Preprocessing, Hypothesis Testing, Power analysis
Samarth Agarwal: EDA, Preprocessing, Hypothesis Testing, Power Analysis

Preprocessing:

Common Steps:

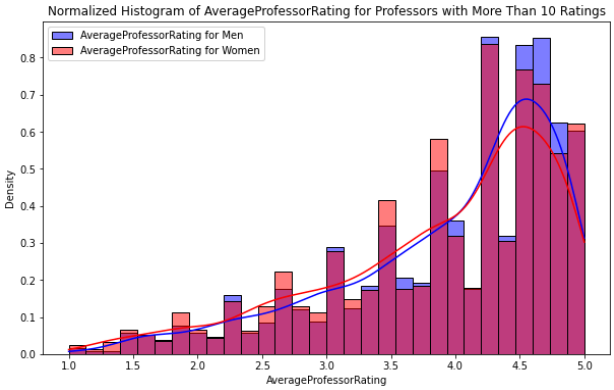
- The RNG is seeded with N number N-1067628 (Deepanshu Mody)
- Rows with fewer than 10 ratings were dropped, The minimum number of students was based on Centra's (1998) research, where an item level reliability analysis indicated that scores based on 10 or more students provide a sufficient level of reliability for research purposes. We were left with 9841 rows.
- All rows with matching gender values, i.e., 0 and 0, or 1 and 1, were removed. This ensured we retained only high-confidence values identifying the professor as a particular gender. Further, we note that we retain a sample size of 7105. This dataset was only used where gender was crucial to the question. Q's 8 and 9 used the larger dataset without the high confidence gender.
- We take confounds as attributes of professors that would impact a students ability to rate them accurately and effectively only based on their gender.
- To cut down on false positives, we consider a significant result those with a p-value below the alpha threshold of .005. A higher p-value was allowed for testing potential confounds.

For Prediction tasks:

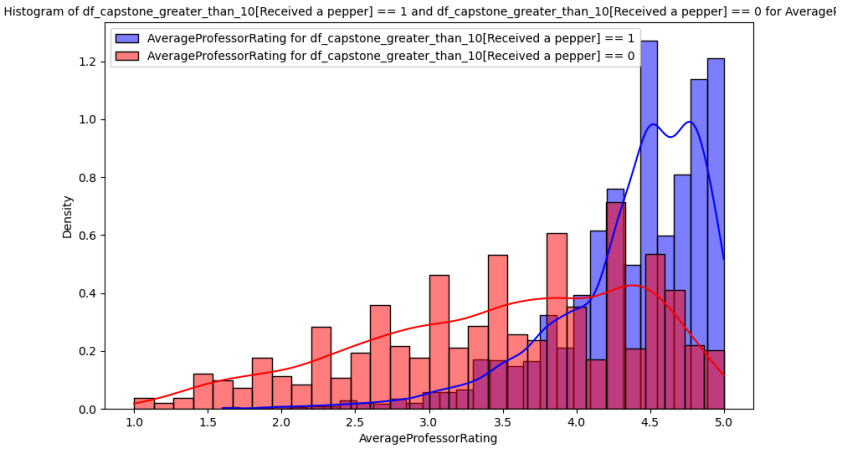
- The 'male with high confidence' column was removed due to the same issue as the dummy variable trap which has a very high correlation with another column 'female with high confidence'.
- Standardscaler was used to normalize the values for ridge and lasso regression and get more interpretable beta values

Question 1)

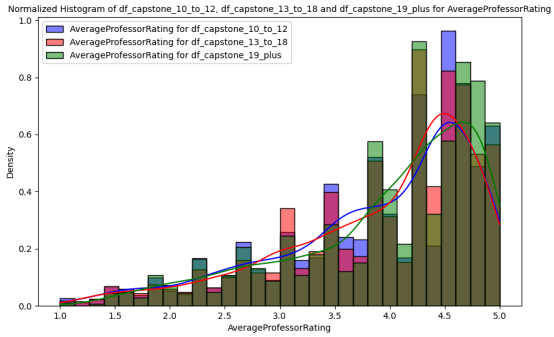
We first compare male and females and test our **null hypothesis: The distribution/location parameter of males and females average ratings are the same.** Distribution is tested using Kolmogrov-Smirnov (**KS**) test and the location/median is tested using Mann Whitney U (**MWU**) test. The KS test returns a p-value of 2.8e-3 & the MWU returns a p-value of 7.3e-4. Considering an alpha level of 5e-3, we drop our null hypothesis stated above. However, we note that the data is not an experiment and randomization has not taken place, therefore we need to account for confounding variables. We look at two and adjust for two potential confounds brought to our attention by previous research by either accounting or failing to account for it respectively: **pepper** (Wallisch, P. and Cachia, J. (2018)) & **years of experience** (Centra & Gaubatz, 2000).

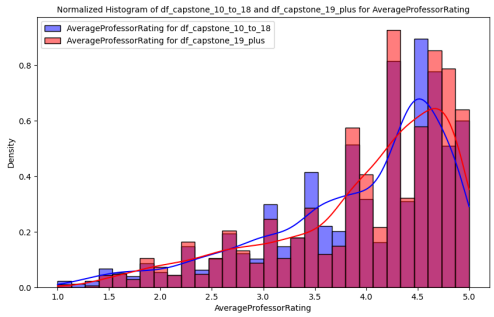


We begin by checking if **pepper** is a confound by testing the null hypothesis: **The distribution(KS)/location(MWU) of average ratings for those who receive a pepper are the same as those who don't.** Plotting the normalized histogram reveals a blaringly clear difference in the distributions of the two groups. Running a **KS** test further reveals (p-value: 2.4e-322) that the distributions are indeed blaringly different. **MWU** (p-value: 0.0), both **KS & MWU** report a significant result therefore we drop our null hypothesis and conclude that the distribution(KS)/location(MWU) of average ratings for those who receive a pepper are not the same as those who don't.



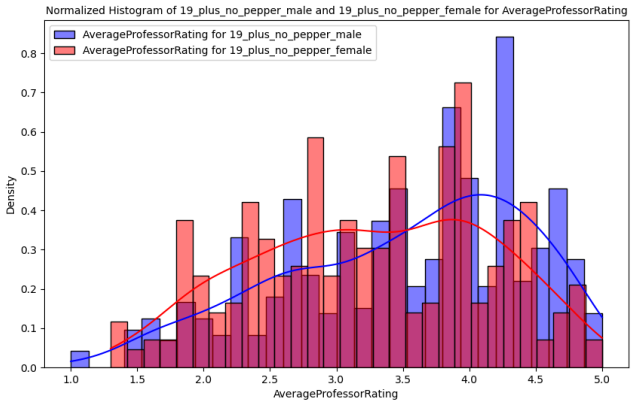
Subsequently, we test for **years of experience** using the number of ratings as a proxy for years of experience. We divide the ratings into the following groups: **10-12, 13-18, 19+** (This allows us to create three groups with sample sizes in the range: **[2048, 2616].** Following which we conduct a Kruskal Wallis **(KW)** test which reports a p-value of **3.2e-3** which is lower than the alpha level and we drop our null hypothesis: **All the three years of experience groups have the same median**. Therefore there exists at least one group with a different median. We subsequently conduct a **KS & MWU** (**since the distributions are not alarmingly different** based of the density plots) with each group pair and **{10-12 vs 13-18: {KS: 3.2e-1, MWU: 5.9e-1}, 10-12 vs 19+: { KS: 1.3e-4, MWU: 1.1e-3}, 13-18 vs 19+: {KS: 3.1e-4, MWU: 8.9e-3}}.** Within the group pairings 10-12 vs 19+ & 13-18 vs 19+ we drop our null hypothesis, however, with 10-12 & 13-18 we don't drop the null hypothesis. We therefore merge the 10-12 & 13-18 groups and test against the 19+ group for which the **KS & MWU** report a p-value of **3.4e-5, 8.4e-4** respectively which are both significant given our alpha level. Having tested for the two potential confounds & verified their significance we now adjust for both **pepper & years of experience.**



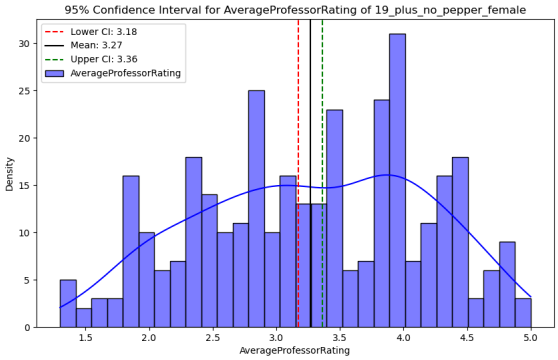


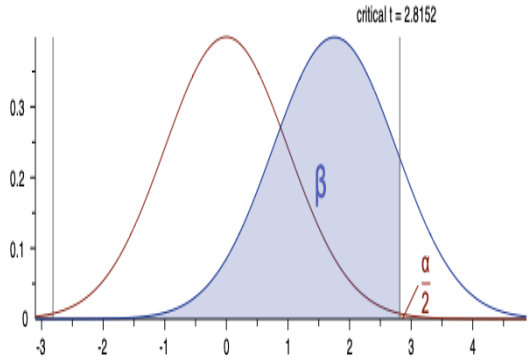
We create a total of eight groups divided by pepper, years of experience & gender. The sample size of the eight groups are in the range: **[346, 1521].** However the male & female groups whose distributions will be tested against each other have relatively similar sample sizes which

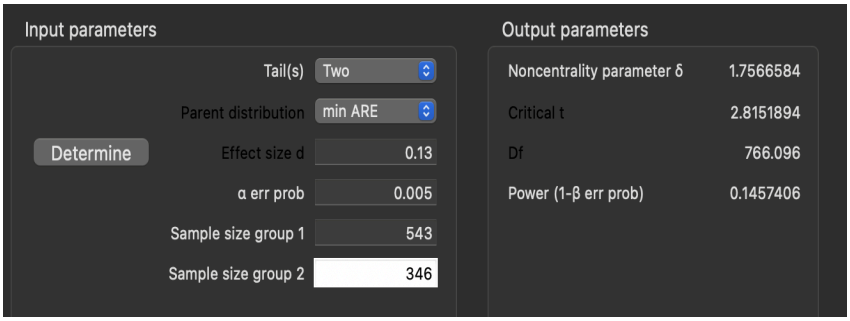
are anyways not the most important in non-parametric tests. Null hypothesis: **The distribution/location of average ratings for those who receive a pepper/don't receive a pepper and are male with the same years of expereience as females are the same.** Running **KS & MWU** on four groups: **{10_to_18_pepper_male vs 10_to_18_pepper_female: {KS: 1.6e-2, MWU: 5.8e-3}, 19_plus_pepper_male vs 19_plus_pepper_female: {KS: 7.0e-1, MWU: 1.9e-1}, 10_to_18_no_pepper_male vs 10_to_18_no_pepper_female: {KS: 3.6e-2, MWU: 5.7e-3}, 19_plus_no_pepper_male vs 19_plus_no_pepper_female: {KS: 3.4e-4, MWU: 3.9e-5}}.** We note that only in the case of the 19 plus no pepper group there exists a significant result and therefore we drop our null hypothesis in that case. Following this we conduct a power analysis study on the 19_plus_no_pepper group. Since we want to study the effect size and likelihood of this significance result.



Note that this is the group with the sample size of {male: 543, female: 346}. We begin by plotting the confidence interval of the {19_plus_no_pepper_male: [3.43, 3.59] 19_plus_no_pepper_female groups: [3.18, 3.36]}. We do note that 95% of our bootstrapped sample means **have no overlap** and that the **difference** in the **lower bound CI of males** in this group and the **upper bound CI of females** in this group are **0.26**. We now look at the effect size using **cohens d** and we calculate a post-hoc effect size of **0.27.** The 95% CI for the effect size is in the range: **[0.130, 0.404]**. We use the lower bound of our 95% confidence interval estimate of effect size **(0.13)** to calculate the power of the MWU test using the simulation software: **G*Power**. We calculate Power (1- β) given the post-hoc effect size, an alpha level of 0.005, sample sizes of 543 & 346 and min ARE as the parent distribution which reveals a score of **0.15**. The power speaks to the reproducibility and likelihood of having actually found a significant difference and a gender bias.

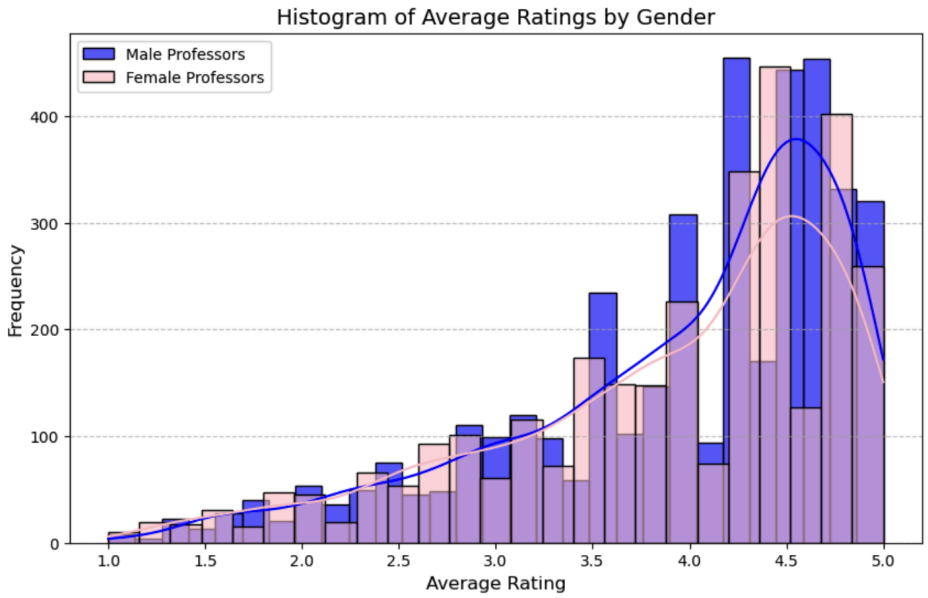






Conclusion: We found significant results in two cases, between men and women with greater than or equal to 10 reviews, and after adjusting for confounding variables we found a significant result between men and women with 19 and more reviews but did not receive a pepper. Following which we conducted a power analysis which revealed that we would obtain this significant result in 15% of the cases (using a conservative estimate since we use the lower end of the CI). Therefore, gender bias does exist in one particular case after adjusting for confounds, however, the low power would suggest that further research would be needed to be able to make a more confident statement.

Question 2)
We investigate whether or not there is a gender bias in the variance of the distribution of Average Ratings. **Alt Hyp: There is a gender difference in the spread (variance/dispersion) of the ratings distribution. Null Hyp: There is NOT a gender difference in the spread (variance/dispersion) of the ratings distribution.**



First step of this analysis was to determine if there was a significant difference in variance between Male and Female professors WITHOUT considering confounding factors. This was done for comparison after adjusting for potential confounds. After filtering the two groups, we conducted a Levene's test to assess the differences in variance of the two groups. We found that there was a **significant difference in the variance of the distribution of average rating among men and women professors (P-value: 0.0024).** This shows that **with all variables considered, there is a significant difference, however, this does not necessarily imply that gender is the sole cause of this effect.** Using the observed effect size (.086), the power of the test was calculated to be moderate (0.7). However, to be conservative and avoid overestimating the result, we recalculated the power using the lower end of the confidence interval [95% Confidence Interval for Cohen's d: (0.019, 0.153)] for the effect size (d=0.019), which reduced the power to a very low level (0.098). This suggests that, while statistically significant, the observed difference in variances is of minimal practical importance and should be interpreted with caution.

Next step is we identify and adjust for confounding variables that might be the cause of this significant result. Variables expected to be confounds were the following: Number of ratings (proxy to the years of experience), Average Difficulty, and Pepper/No Pepper. The techniques used to determine confounding status included significance tests and correlation tests. We first tested to see if the Number of Ratings had an association with the differences in variance of distribution of Average Rating without considering gender. **We divide the Number of Ratings into two groups by performing a median split (justified because levene's test**

does not need equal samples). We then tested the difference in the variance of Average Ratings between the two groups using Levene's tests. We noted a significant difference between the two groups (P-value: 0.0034). When performing a correlation analysis, the correlation was low. We also note there is an association between number of ratings and gender. Despite the low correlation, mainly because of the significance test **we considered Number of Ratings a confounding variable.**

Next we repeat the same process with the Average Difficulty. Performing a median split and a levene's test, we observe a **significantly low p value (P-value: 0000).** Additionally, the correlation between the two was around -.65. We also note that there is a relationship between Difficulty and Gender. **We consider this a confound**

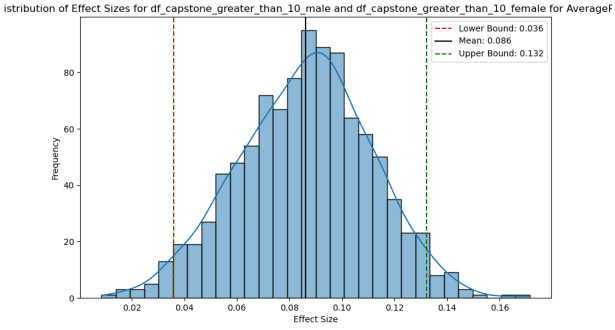
Finally, we test Pepper/No Pepper as a potential confound. To do this we first see if there is a significant difference in the **variance of distribution of average ratings using Levene's test. We find a significant difference (P-value: 000)**, indicating a heavy influence on the variation. Additionally, the correlation between the two was moderate. We also note that there is an association between Pepper/No Pepper and gender. **We consider this a confound.**

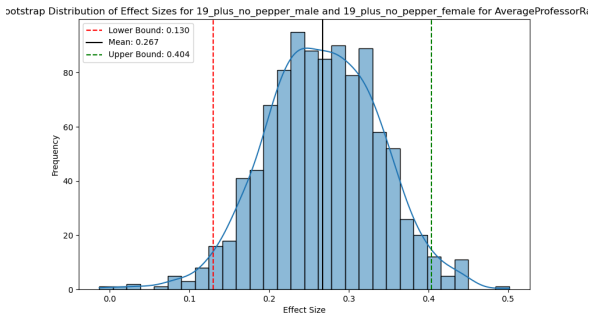
We now move to controlling for the confounds and testing the variance in distribution between the groups. This is done by using a median split for the two numerical variables, and categorical split for Pepper/No Pepper. For example, one group would be female professors, who have an average difficulty less than 2.9 and number of ratings above 14, and receive a pepper. There are 8 such stratified groups for Males and Females. We compute the levene's test on all of these 8 groups and find that **only one group has a significant difference in the variance of the distribution of ratings. The group in this category: Below Median Diff (≤2.90), Below Median Num Rate (≤14.00), Yes Pepper (P-value: 0.004239). We then calculated the effect size and the confidence interval, and ran power analysis for this group. Effect Size: .18, 95% Confidence Int: [.07, .3], Power: .043**

Conclusion: After controlling for confound, we found only one subgroup out of 8 total subgroups compared where the male and female professors differ significantly in the variance of their ratings. The very low power suggests that this significant result should be interpreted cautiously, as the test has a high risk of missing true effects in other groups or overestimating the significance in this group. The findings highlight that variance differences might exist in very specific contexts but are likely not generalizable across all groups or conditions. We conclude that there is no significant gender difference in the variance of distributions of average ratings.

Question 3)

We calculate the 95% CI of the effect size for men and women **(with only the null values dropped) as [0.043, 0.077]**, with the mean effect size as 0.059. Further we calculate the 95% CI of the effect size for men and women with more than **10 rating: [0.043, 0.136]** and the mean effect size is 0.086. The 95% CI of the effect size for men and women who received 19 or more ratings and received a pepper is in the range of: **[0.130, 0.403]** and the mean, i.e. the likely effect size, is **0.27**. 95% Confidence Interval for the effect size of the spread of men and women with more than 10 ratings Cohen's d: (0.039, 0.133)], 95% Confidence Interval for the group in this category: Below Median Diff (≤2.90), Below Median Num Rate (≤14.00), Yes Pepper: [0.07, 0.3], Effect Size: 0.18.



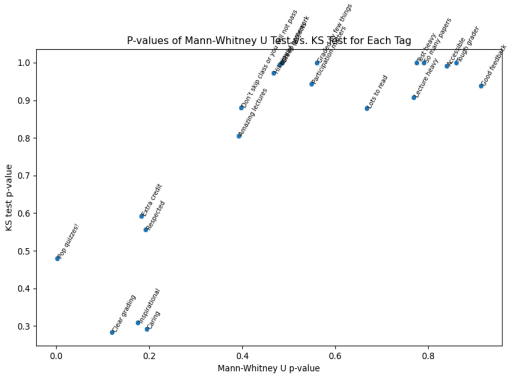


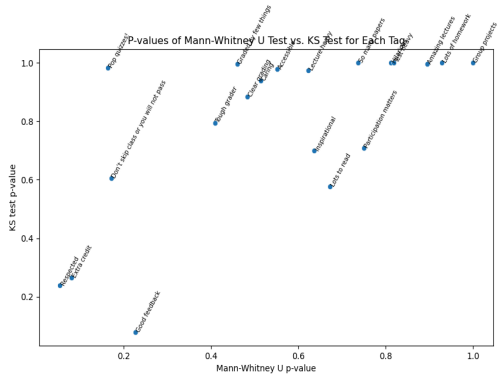
Question 4)

We normalize the tags by dividing each the number of a tag received by the professor with the number of ratings received by a professor.

For professors with more than 19 ratings and no pepper rating male vs female: Pop quizzes!, **MWU:** 0.0017, **KS**: 0.65. We don't receive statistically significant value for the group: professors with more than 10 ratings male vs females.

The three most gendered, for male and females with more than 10 ratings are: Extra credit, Respected, Pop quizzes : (**MWU: 0.057, 0.064, 0.097, KS:0.58, 0.22, 0.97 respectively).** The three least gendered are: Test heavy, Caring, So many papers: (**MWU: 0.96, 0.96, 0.90, KS: 1.0, 0.8, 1.0 respectively).** The three most gendered, for male and females with more than 19 ratings and did not receive a pepper are: Pop quizzes!, Clear grading, Extra credit: **(MWU: 0.0017, 0.006, 0.17, KS: 0.65, 0.38, 0.82 respectively)**. The three least gendered are: Accessible, Lots to read, So many papers: **(MWU: 0.99, 0.95, 0.92, KS: 1.0, 1.0, 1.0 respectively).**



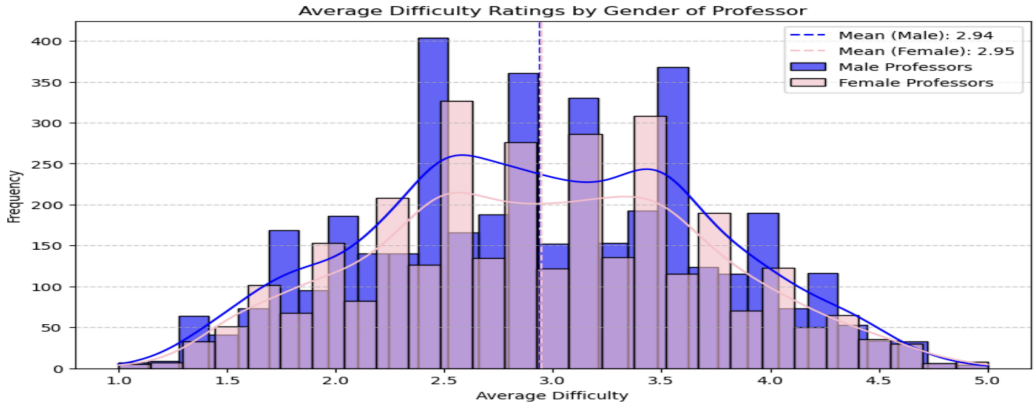


Left Figure: Group w/ 10+ Ratings **Right Figure: Group w/ 19+ Ratings/No Pepper**

Conclusion: We notice that Pop quizzes is the only result that appears as statistically significant and appears in both the most gendered groups: 10 or more ratings and after adjusting for confounds.

Questions 5/6)
We investigate whether or not there is a gender difference in terms of average difficulty. **Alt Hyp: There is a gender difference in average difficulty. Null Hyp: There is NOT a gender difference in average difficulty.**

First step of this analysis was to determine if there was a significant difference in average difficulty between Male and Female professors WITHOUT considering confounding factors. This was done for comparison after adjusting for potential confounds. After filtering the two groups, we conducted a KS and Mann Whitney test to assess the differences in average difficulty of the two groups. **We found that there was not a significant difference in average difficulty among men and women professors. Mann-Whitney U Test: P-Value: 0.786 Kolmogorov-Smirnov Test: P-Value: 0.997.**



This shows that **with all factors considered, there is not a significant difference in Difficulty, however, this does not necessarily imply that there is no difference because we have not adjusted for confounds.** Using the observed effect size (-.004), the power of the test was calculated to be extremely low (0.005). **However, to be conservative and avoid overestimating the result, we recalculated the power using the lower end of the confidence interval for the effect size (-.03), which raised the power (0.05).** The analysis found **no statistically significant difference in average difficulty ratings between male and female professors** with p=0.997 for KS and .78 for MW. The effect size (d=.05) is small, with a 95% bootstrap confidence interval (−0.03 to 0.023), suggesting the true effect size is minimal. Additionally, the test had very low power (0.005), indicating a high likelihood of failing to detect meaningful differences if they exist. **Overall, the results suggest no meaningful or significant differences in ratings based on gender.**

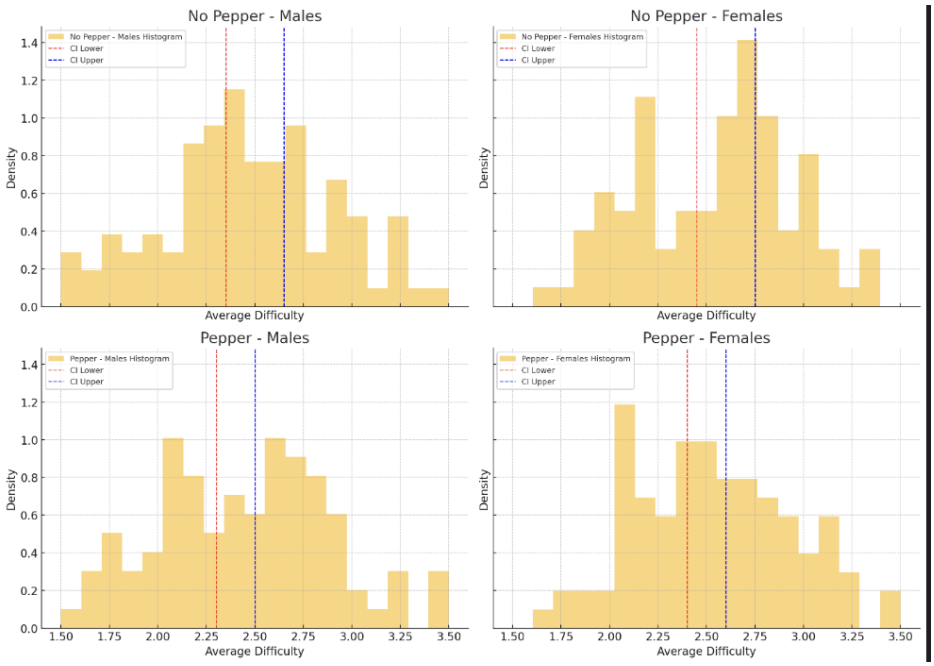
Now we will adjust for confounds. **The only confound that we found to be significant and worth controlling for was pepper/no pepper.** Performing a KS and MW test on professors who received a pepper and those who did not to determine if Pepper influences Average Difficulty rating. MW P-Value: 0.000 Kolmogorov-Smirnov Test:P-Value: 0.000. The distributions of Average Difficulty significantly differ between professors who received a pepper and those who did not. The correlation between 'Received a Pepper' and 'Average Difficulty' is: -0.290

indicating correlation. We also note that when performing a Chi Squared test to determine if Pepper differs significantly among Men and Women, we receive a low p value (.02). **We determine that Pepper is a confound.**

Our Results by Groups with Effect Size and Bootstrap Confidence Interval:

Group: Pepper = No, P-Value: 0.370, Cohen's d (Effect Size): -0.032, 95% Bootstrap CI for Cohen's d: (-0.121, 0.054). Group: Pepper = Yes, P-Value: 0.517, Cohen's d (Effect Size): -0.019, 95% Bootstrap CI for Cohen's d: (-0.105, 0.069). **For this test, controlling for the confounding factor, we still receive p values that show an insignificant difference in average difficulty between men and women.**

Conclusion: Before controlling for potential confounds, there is no significant difference in average difficulty for men and women professors. Once we controlled for the confounding of Pepper/No Pepper, we still see that the Average difficulty between men and women is not significantly different. Average difficulty is not gendered.



For **Q's 7,8,9 -** We checked the LINE assumptions for our final models

- We confirmed the linear relationship between our final predictors and the response variable by plotting scatterplots of each independent variable against the dependent variable.
- Independence between observations is assumed and potential independent predictors are discussed for each question
- Normality was checked by plotting a histogram of residuals.
- Equal variance was checked by plotting predictions against standardized residuals. We tried to get as close to homoscedasticity of residuals as possible.

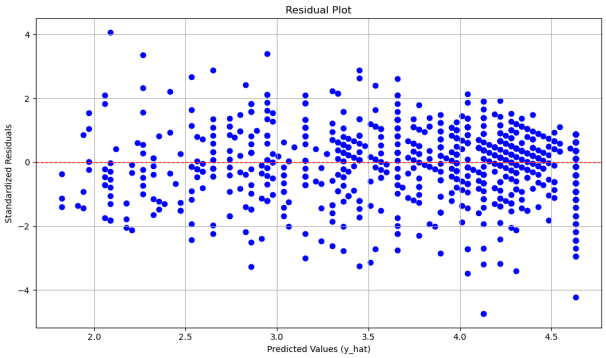
Question 7)
We removed 1,638 rows with missing values in the 'Proportion of students that said they would take the class again' column (16.64% of the data). We analyzed if we should drop rows based on this feature since it is the most highly correlated feature to 'Average Rating' and found the following- The mean, median, and standard deviation for the 'Average Rating' were as follows: For rows without any NAs: Mean = 3.9, Median = 4.2, Std Dev = 0.83, For rows with only NAs: Mean = 3.63, Median = 3.9, Std Dev = 1.09

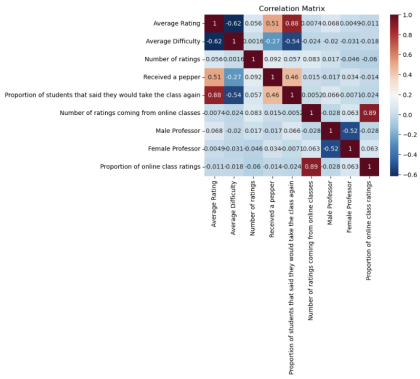
We analyzed histograms and boxplots, noting slight distribution differences. Without imputation strategies using correlated predictors, we decided to drop the missing rows instead given the scope of the project. Later, we found that using other predictors instead increased RMSE from 0.39 to 0.6, making the slightly worse fit from dropping rows most likely preferable to excluding this feature entirely.

We introduced a new column, 'Proportion of Online Class Ratings,' by dividing 'Number of Ratings from Online Classes' by 'Total Ratings' to improve linearity. Both original columns were not considered later due to weak correlation and small beta values.

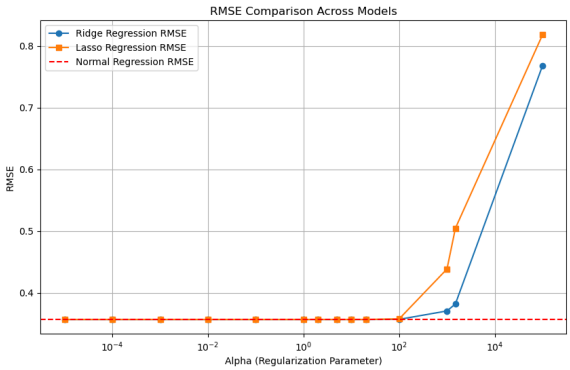
After preprocessing, we selected six candidate predictors: 'Average Difficulty,' 'Number of Ratings,' 'Received a Pepper,' 'Female Professor,' 'Proportion of Online Class Ratings,' and 'Proportion of students that said they would take the class again.' We found that 'Average Difficulty,' 'Received a Pepper,' and 'Proportion of students that said they would take the class again' produced similar RMSE and R^2 values when using all six predictors. However, these three features were collinear. By increasing the alpha in Lasso regression, we achieved comparable results using just the 'Proportion of students that said they would take the class again' feature. The performance of the models was as follows: 6-feature model: RMSE ≅ 0.35, 3-feature model: RMSE ≅ 0.37, 1-feature model: RMSE ≅ 0.38

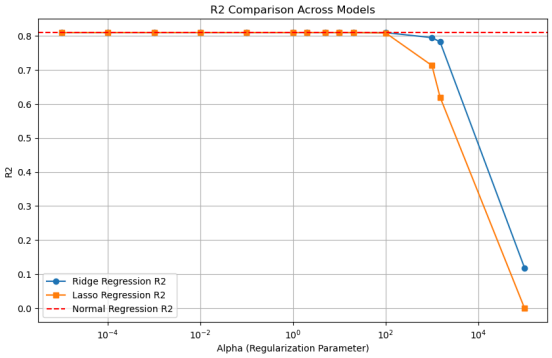
The 1-feature model was more interpretable, better satisfied the homoscedasticity assumption, and had independent features. (Residual plot and correlation matrix below)





We tested Ridge, Lasso, and OLS regression with cross-validation on our 6-feature model. Although regularization showed limited effectiveness, it was useful for feature selection. At an alpha value of 2,000 (custom Lasso implementation), the model performed well but OLS achieved the best RMSE (0.352) and R^2 (0.823) on the test set. Our final 1-feature OLS model had slightly worse test metrics **(RMSE = 0.392, R^2 = 0.781)**, but it was more interpretable and had better residuals.

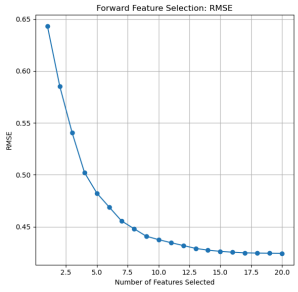


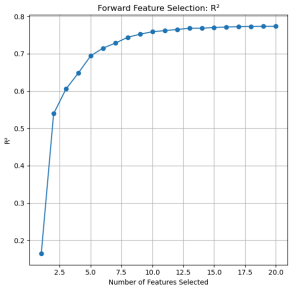


Conclusion: 'Proportion of students that said they would take the class again' emerged as the most predictive feature of 'Average Rating,' with a beta of **0.712** on the normalized column.

Question 8)

All tag columns were normalized by dividing their values by the "Number of Ratings" column. Although we attempted to remove collinear combinations of tags, this did not significantly improve the final cross-validated RMSE and R^2. However, we still decided to perform feature selection, initially using LASSO, but also employing forward feature selection. This decision was influenced by the observation that LASSO can arbitrarily drop features in cases of high multicollinearity, as noted in the study “Application of LASSO and its Extended Method in Variable Selection of Regression Analysis”.

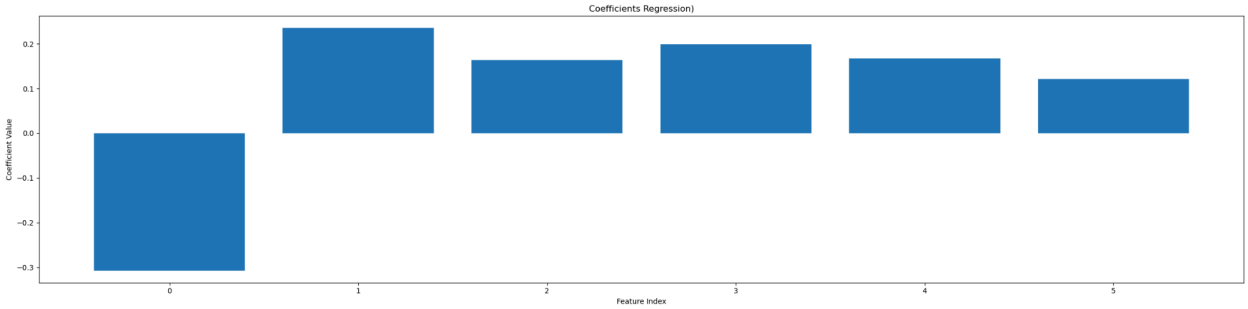




	Selected Features	Betas	Alpha	RMSE	R2
0	[0]	[-2.920318049424987]	4.423652	0.642993	0.165311
1	[0, 2]	[-2.3002355764323186, 1.9557153206850926]	3.974497	0.585024	0.539350
2	[0, 2, 1]	[-2.0435407855345944, 1.7104861365930213, 1.23...	3.678943	0.540372	0.605481
3	[0, 2, 1, 15]	[-1.7957356846871584, 1.3395895701406948, 1.36...	3.471961	0.502089	0.648089
4	[0, 2, 1, 15, 16]	[-1.566579366104837, 1.1448644085592679, 1.202...	3.290329	0.482274	0.694198
5	[0, 2, 1, 15, 16, 12]	[-1.4498072706722365, 1.0695459833360967, 1.21...	3.202371	0.468835	0.714935
6	[0, 2, 1, 15, 16, 12, 11]	[-1.2609644116334613, 1.1126881030866582, 1.16...	3.031202	0.455535	0.728398
7	[0, 2, 1, 15, 16, 12, 11, 17]	[-1.1628442246472999, 1.1686518658383216, 1.22...	2.935969	0.448095	0.743872
8	[0, 2, 1, 15, 16, 12, 11, 17, 4]	[-1.0813781678275431, 1.205648590027773, 1.153...	2.816139	0.440797	0.752239
9	[0, 2, 1, 15, 16, 12, 11, 17, 4, 7]	[-1.0420458042241987, 1.1029005020712037, 1.12...	2.792670	0.437487	0.758736
10	[0, 2, 1, 15, 16, 12, 11, 17, 4, 7, 9]	[-1.0155456859400844, 1.1103074736768033, 1.11...	2.746535	0.434554	0.761753

Regularization did not significantly improve metrics, suggesting that the model may still underfit, even when all tags were included. However, regularization was useful for feature selection, as forward feature selection identified the same subset of important features as LASSO. Despite correlations among some tags reaching approximately 0.4, including all tags yielded better cross-validated RMSE and R^2. Nonetheless, we decided to retain the six most important tags.

We experimented with ridge regression, LASSO, and ordinary least squares (OLS) regression. The full-feature OLS model achieved the best RMSE and R^2 values of 0.428 and 0.761 on the test set, respectively. Feature selection via LASSO left us with the top six features (the same as those identified by forward selection) without significantly affecting performance metrics. On this reduced feature set, OLS again outperformed ridge and LASSO, achieving **RMSE and R^2** values of **0.474 and 0.705**, respectively.



Examining the beta coefficients, the tag **“Tough Grader” (Tag 0)** emerged as the most predictive feature of "Average Rating," with a **beta value of -0.307**. While multicollinearity may render this beta estimate less reliable, “Tough Grader” consistently appeared as the most important feature across both forward selection and LASSO with increasing alpha values.

The model violated the assumption of homoscedasticity, as residuals did not exhibit equal variance, though normality was satisfied. It also violated independence of predictors as many were highly correlated. We attempted to address this issue using polynomial regression with the predictors and by applying log and square root transformations to the response variable, but these methods did not yield improvements. Other approaches, such as Box-Cox Transformation, Generalized Least Squares (GLS), Robust Standard Errors, or Weighted Least Squares (WLS), could potentially address this limitation but were considered outside the scope of this project.

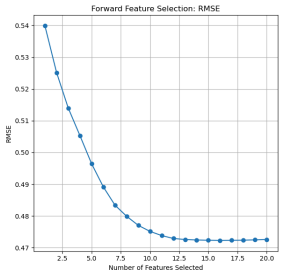
Conclusion: On this reduced feature set with 6 features, OLS again outperformed ridge and LASSO, achieving **RMSE and R^2** values of **0.474 and 0.705**, respectively. A simplified model using only the tag 0 **"Tough Grader"** feature resolved the homoscedasticity issue and independence issue. This model produced **RMSE and R^2** values of **0.644 and 0.473**, respectively, with a **beta** coefficient of **-0.61**. Although this model provided a poorer fit, it effectively addressed the aforementioned limitation.

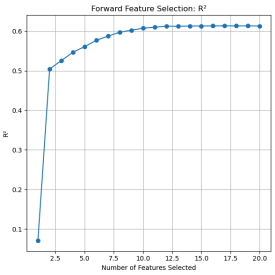
Comparison with a Previous Model

In a previous model, the feature "Proportion of students who said they would take the class again" was the most predictive of "Average Rating," outperforming "Tough Grader" in predictive power as well as performance 0.71 vs -0.61 and better metrics. This feature appeared to drive the excellent RMSE and R^2 values in that model. Excluding this feature led to noticeably weaker performance metrics in the current analysis. While incorporating more features in this model reduced RMSE and increased R^2 overall, no individual feature proved as strong a predictor as the previously identified feature. The correlation matrix suggests that the relatively weaker predictive power of individual features in this model may contribute to this result.

Question 9)

All tag columns were normalized by dividing their values by the "Number of Ratings" column. Although we attempted to remove collinear combinations of tags, this did not significantly improve the final cross-validated RMSE and R^2. However, we still decided to perform feature selection, initially using LASSO, but also employing forward feature selection. This decision was influenced by the observation that LASSO can arbitrarily drop features in cases of high multicollinearity. We can look at the forward feature selection metrics below.

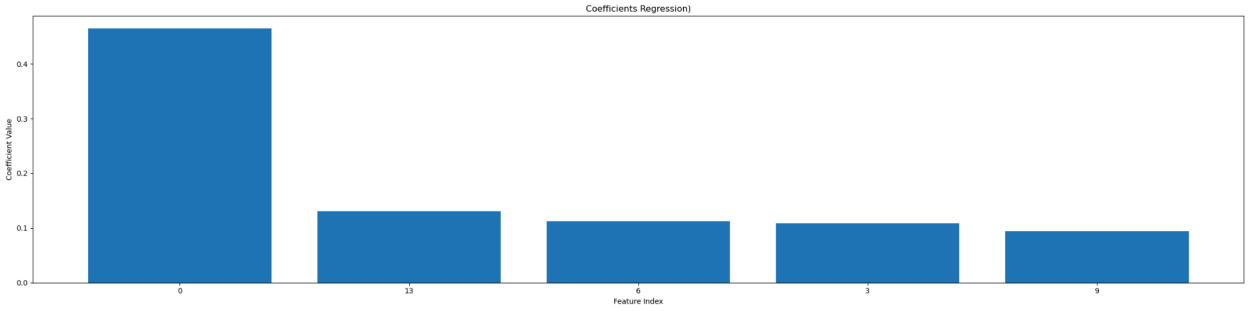




	Selected Features	Betas	Alpha	RMSE	R2
0	[0]	[2.514707431316056]	2.493630	0.539856	0.071404
1	[0, 13]	[2.3530536676358875, 1.696102907380288]	2.459707	0.525108	0.504229
2	[0, 13, 6]	[2.222163625532835, 1.6541115432135847, 0.6121...	2.396875	0.513944	0.525410
3	[0, 13, 6, 3]	[2.1136342839185227, 1.7104254285850087, 0.634...	2.335583	0.505284	0.546658
4	[0, 13, 6, 3, 9]	[2.163423814448983, 1.6518745405447677, 0.6250...	2.249008	0.496430	0.560662
5	[0, 13, 6, 3, 9, 5]	[2.0511226046808892, 1.5325860526705326, 0.581...	2.179388	0.489164	0.576561
6	[0, 13, 6, 3, 9, 5, 11]	[1.9533351249347712, 1.5469977967748598, 0.558...	2.291702	0.483405	0.587317
7	[0, 13, 6, 3, 9, 5, 11, 12]	[1.8813537528358482, 1.4978062289451983, 0.487...	2.395872	0.479906	0.597186
8	[0, 13, 6, 3, 9, 5, 11, 12, 17]	[1.8386918903077738, 1.5265959008465588, 0.479...	2.449640	0.477043	0.602206
9	[0, 13, 6, 3, 9, 5, 11, 12, 17, 16]	[1.7669833495079812, 1.457190385108505, 0.4640...	2.545292	0.475154	0.607268
10	[0, 13, 6, 3, 9, 5, 11, 12, 17, 16, 14]	[1.760648543371506, 1.5732804469385122, 0.4366...	2.573403	0.473834	0.609740

Regularization did not significantly improve metrics, suggesting that the model may still underfit, even when all tags were included. However, regularization was useful for feature selection, as forward feature selection identified the same subset of important features as LASSO. Despite correlations among some tags reaching approximately 0.4, including all tags yielded better cross-validated RMSE and R^2. Nonetheless, we decided to retain the 5 most important tags.

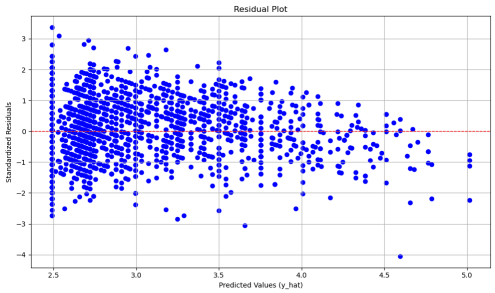
We experimented with ridge regression, LASSO, and ordinary least squares (OLS) regression. The full-feature OLS model achieved the best RMSE and R^2 values of 0.472 and 0.600 on the test set, respectively. Feature selection via LASSO left us with the top five features (the same as those identified by forward selection) without significantly affecting performance metrics. On this reduced feature set, OLS again outperformed ridge and LASSO, achieving **RMSE and R^2** values of **0.493 and 0.561**, respectively.



Examining the beta coefficients, **Tag 0, 'Tough Grader,'** again emerges as the most predictive feature of 'Average Difficulty,' with a beta value of **0.464**. Even though collinearity could affect the importance of individual features by changing the beta values, based on the correlation matrix values as well, we can most likely conclude that tag 0 'Tough grader' is the most important feature for the prediction of 'Average Difficulty'

The model violated the assumption of homoscedasticity, as residuals did not exhibit equal variance, though normality was satisfied. It also violated independence of predictors as many were highly correlated. We attempted to address this issue using polynomial regression with the predictors and by applying log and square root transformations to the response variable, but these methods did not yield improvements. Other approaches, such as Box-Cox Transformation, Generalized Least Squares (GLS), Robust Standard Errors, or Weighted Least Squares (WLS), could potentially address this limitation but were considered outside the scope of this project.

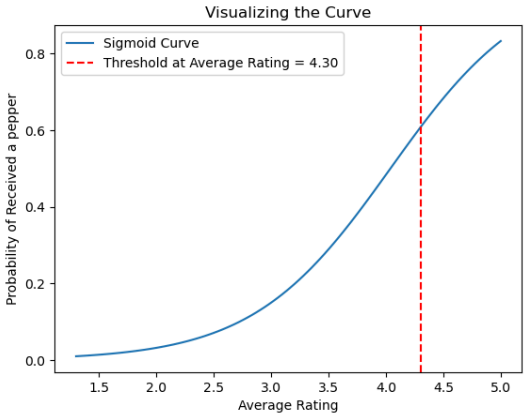
Conclusion: On this reduced 5-feature set, OLS again outperformed ridge and LASSO, achieving **RMSE and R^2** values of **0.493 and 0.561**, respectively. A simplified model using only the most predictive tag 0 **"Tough Grader"** feature did not resolve the homoscedasticity issue. This model produced **RMSE and R^2** values of **0.540 and 0.473**, respectively, with a **beta** coefficient of **0.53**. We find this model addressed the independence issue between predictors but it still does not address homoscedasticity. We also tried square root and polynomial transformations on this variable alone but it did not sufficiently address this issue.



Hence, we use our 5-feature OLS model **RMSE and R^2** values of **0.493 and 0.561** as our final model.

Question 10)
We combined both the tables and all tag columns were normalized by dividing their values by the "Number of Ratings" column. We drop all NA values present in the ‘Proportion of students that said they would take the class again’ column with similar reasoning to previous questions. We calculated the correlation matrix and the following predictors had a low correlation hence the following columns were dropped- 'Number of ratings','Number of ratings coming from online classes','Male Professor','Female Professor',4,8,9,17,18. We used logistic regression and achieved an **AUC-ROC score of 0.83**. ‘Average rating’ had the highest beta coefficient and its exponential was quite large compared to other variables. We had an F1 score of **0.72** and **0.76** for the **0 and 1 class** with a support of **607 and 586** respectively. Since the 0 and 1 class have similar support values, we can be confident that class imbalance is not an issue.

Based on the coefficient information of the above model we decided to train a logistic regression model just based on the ‘Average Rating’ predictor. This model achieved an **AUC-ROC score of 0.82** which is extremely close to our previous model with multiple predictors. Thus we think this is the best model.

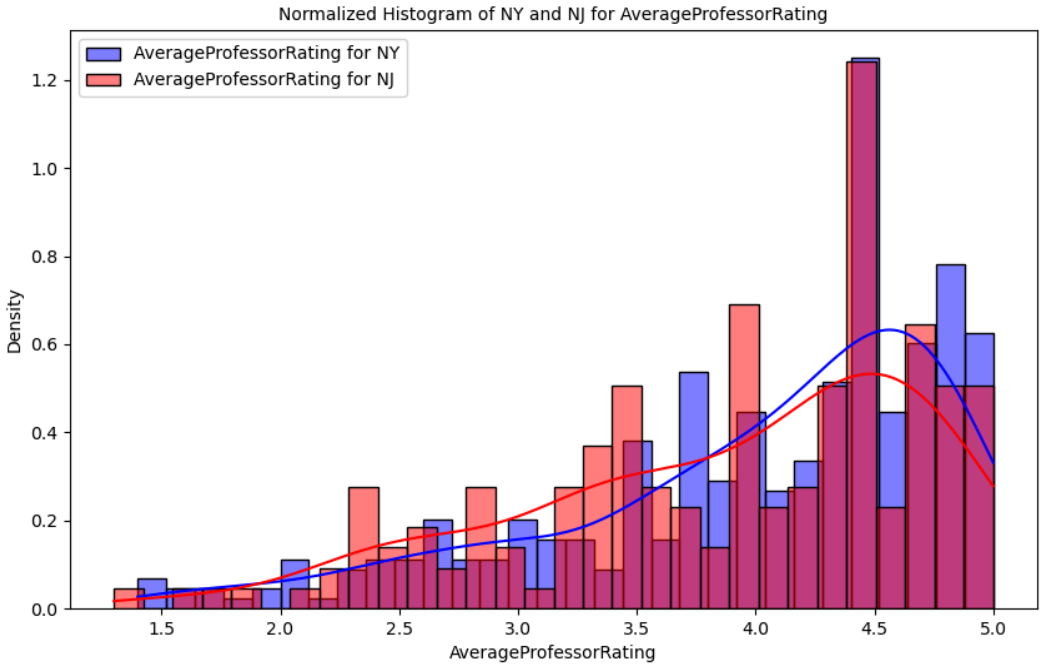


	precision	recall	f1-score	support
0.0	0.76	0.73	0.75	607
1.0	0.74	0.76	0.75	586
accuracy			0.75	1193
macro avg	0.75	0.75	0.75	1193
weighted avg	0.75	0.75	0.75	1193

Conclusion: For every increase in 1 unit of Average Rating, we expect the odds of Received a pepper relative to odds of not receiving a pepper (the ratio) to increase by e^1.66 which is 5.301. We had an F1 score of **0.75** and **0.75** for the **0 and 1 class** with a support of **607 and 586** respectively. Since the 0 and 1 class have similar support values, we can be confident that class imbalance is not an issue. The precision and recall values from the classification report also makes it clear that class imbalance is not an issue. Hence, we find this to be our best model.

Question 11)

We looked at the NY and NJ data and tested for the Average Professor Rating between the two states. KS Test of AverageProfessorRating for the two groups: NY and NJ KS Test P-value: 0.26, Mann Whitney U P-value: 0.085 We dont drop the null hypothesis and therefore retain that: The distributions of AverageProfessorRating for NY and NJ are the same. We dont drop the null hypothesis and therefore retain that: The median/location of AverageProfessorRating for NY and NJ are the same. This is particularly interesting since NY has an ivy-league school and the popular and reputable NYU. They should/could be reasons for a higher average rating of professors. However, there is **no statistically significant** difference between NY and NJ professor ratings.



References

Centra, J. A (1998). Development of The Student Instructional Report II. Princeton, NJ: Educational Testing Service.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching?. The journal of higher education, 71(1), 17 - 33.

Wallisch, P. and Cachia, J. (2018) Are student evaluations really affected by gender? nope, they're affected by 'hotness.', Slate Magazine. Available at: https://slate.com/technology/2018/04/hotness-affects-student-evaluations-more-than-gender.html (Accessed: 06 December 2024).

Xi, L. J., Guo, Z. Y., Yang, X. K., & Ping, Z. G. (2023). *Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine]*, *57*(1), 107–111. https://doi.org/10.3760/cma.j.cn112150-20220117-00063