PODS Capstone Project Report

Simon Ni, zn2133, N13517345

Introduction:

1. Pre-processing:
   1.1. I load the data from “rmpCapstoneNum.csv” and “rmpCapstoneQual.csv” using numpy. I will refer to them as “dataNum” and “dataQual” respectively, aligned with their name in the .py file I delivered.
   1.2. After drawing the histogram of average rating without considering the number of ratings that each average rating is calculated from, we can see that the distribution of average rating is not normal, which is something to notice for the entire project.

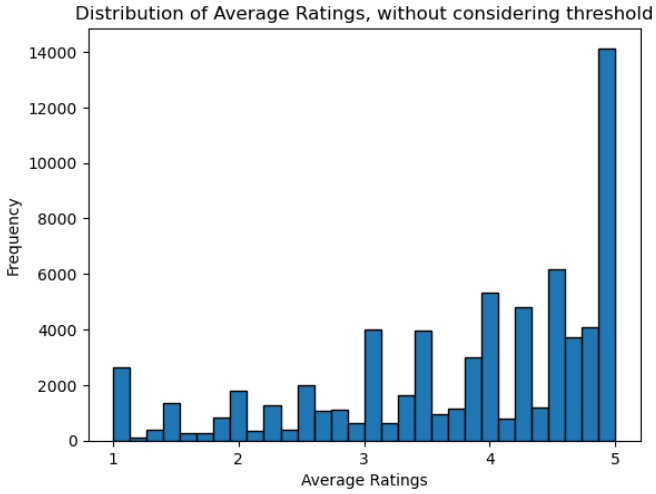


Figure 1

   1.3. I treated the first column (average rating) as ordinal data, because ratings are originally ordinal, meaning a 5 is better than a 4, but the gap between 4 to 5 is different from 3 to 4. In addition, even though the professor took the mean of ratings and put it in the first column, it does not mean that the first column is cardinal, because the distance from, for example, 3.7 to 4.5 is not consistent.
   1.4. This applies to all questions: There are nan values in every column. For each question, I handled this by removing these nan values row-wise for columns only related to that question.
   1.5. Dropping rows that have a low number of ratings, meaning that the average rating those rows carry is not meaningful. Therefore, I need to select a threshold of the number of ratings and drop the rows in the

dataset with the number of ratings less than that threshold. To do so, I drew the histogram of the number of ratings first, which is shown below. As the distribution starts plateauing at the number of ratings equal to 7, I choose 7 as the threshold of the number of ratings, meaning dropping every row that has a number of ratings less than 7. I will refer to this pre-processed dataset as dataNum_thresh.

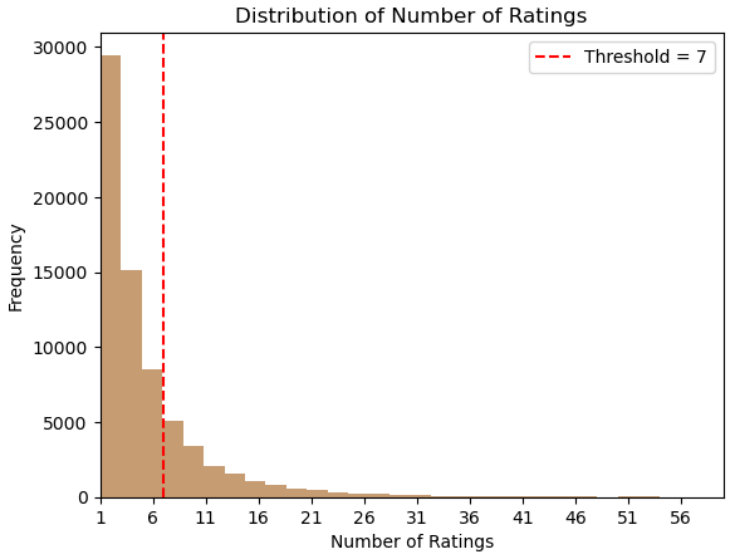


Figure 2

2. Seed for RNG: 13517345.

Question1:

To detect if there is a strong gender bias in students' evaluations of professors, with male professors enjoying a boost in ratings from this bias, I decided to do a statistical significance test. The null hypothesis is that gender does not affect the evaluations that professors get. The alternative hypothesis is that male professors tend to get higher evaluations than non-male professors. The alternative hypothesis indicates that this is a one-tail test. I said “non-male” instead of female because I noticed that in dataNum_thresh, there are some rows with both 0 in the “male” and “female” columns.

Before actually doing the statistical significance testing, I needed to pre-process the dataset, since some rows/professors have average ratings that are calculated from a low number of ratings, as pointed out in 1.1 in the introduction. In addition, potential confounders should be considered, especially the teaching experience of professors. Teaching experience can be a confounder as longer teaching time implies that the professor entered academia early, which might be the time that gender bias was a real problem in academia, favoring more male professors. And it intuitively makes sense that professors with more experience tend to deliver better education and thereby get better evaluations. As I had from dataNum, the teaching experience of each professor can be

estimated from the number of ratings each professor has. Both reasons suggested that the number of ratings should be controlled. Luckily, the dataNum_thresh is always processed based on the number of ratings, which means that problems of the low number of ratings and teaching experience being a confounder are already addressed. I also dropped rows that have nan values on the average rating column or the male gender column.

As claimed in the introductory paragraph, I treat the first column (average rate) as ordinal data. And as the following histogram of average rating shows, the distribution of average rating of each group is not normal but similar in shape. Therefore, to compare professor ratings across genders, I used the Mann-Whitney U test, which is well-suited for ordinal data like ratings.

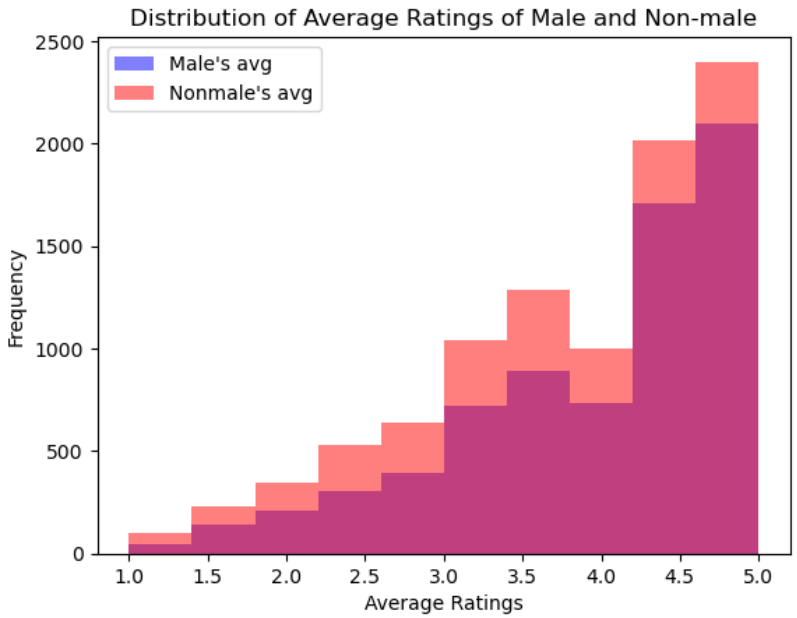


Figure 3

After doing the Mann-Whitney U test, we have the p-value that is 1.12e-19, which is basically just 0 < 0.05, so we drop the null hypothesis and conclude that there is a gender bias in teachers' average rating.

Question 2:

Like what we did in question 1, to detect if there is an effect of the teachers' experience on the quality of teaching, I decided to do a statistical significance test. In this case, we operationalize the quality of teaching with the average rating that a professor gets, and use the number of ratings that a professor gets to represent their experience. The null hypothesis is that teaching experience does not affect the average rating. The alternative hypothesis is that teaching experience affects the average rating, but in an unknown

direction. The alternative hypothesis indicates that this is a two-tail test. In this question, we continue to use dataNum_thresh as our dataset since the average rating is not meaningful if it is calculated from a low number of ratings.

I split and labelled the dataNum_thresh as two groups with a high number of ratings and a low number of ratings, by using the median of the number of ratings in dataNum_thresh as a cutoff, and they represent the group of professors with high experience and low experience, respectively. Rows with the number of ratings equal to the median were assigned to the group of professors with high experience. After doing EDA (*Figure 4*), I found that it is not certain that the shape of distributions of average rating with low and high numbers of ratings are similar enough, which is an assumption of using the Mann-Whitney U test. Therefore, I decided to use the permutation test.

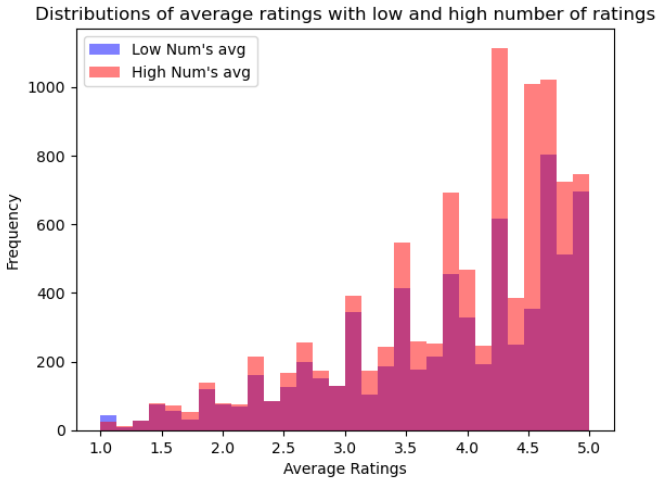


Figure 4

Since the average rating is ordinal data, I designed the test statistic to be the median of the first group minus the median of the second group. I finally got the p-value to be 0.0004, which is less than 0.005, showing that the result is statistically significant. Therefore, I dropped the null hypothesis, concluding that teaching experience affects the quality of teaching.

Question 3:

To detect the relationship between average rating and average difficulty, I decided to do EDA first. Here is what I got:

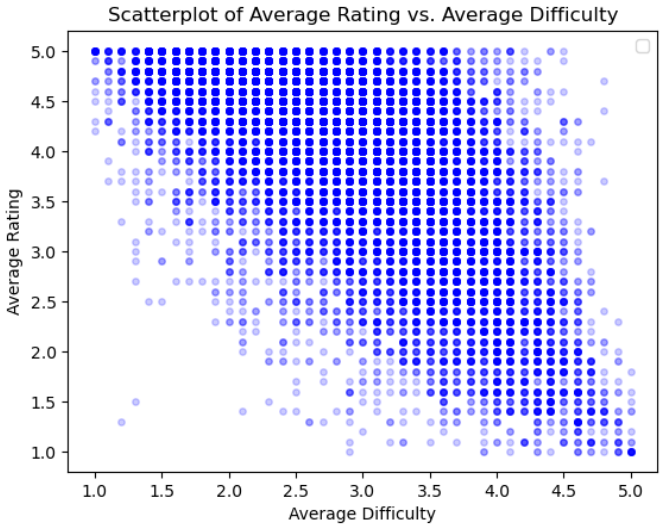


Figure 5

Since the mean (average rating/average difficulty) is not meaningful if the number of ratings is small, I still chose to proceed with the dataset that had been processed to drop rows with a low number of ratings: dataNum_thresh.

Since both variables here, average rating and average difficulty, are rating-type data, they are ordinal. And I could not eyeball to determine if their relationship is linear. Thus, I chose to find out the Spearman's correlation coefficient, which is around -0.62, as my code shows. The Spearman's correlation coefficient, being -0.62, indicates a moderate negative association between average rating and average difficulty. This suggests that, in general, if the course that the professor teaches is harder in students' opinion, the professor tends to get lower average ratings, although the relationship is not perfectly negative monotonic.

Question 4:

To find out if professors who teach a lot of online classes receive higher or lower ratings than those who do not, I decided to do a statistical significance test. Again, since the average rating is involved in this question, I proceeded using dataNum_thresh. I operationalized "a lot of online classes" with a new variable: proportion of online ratings. It is created by dividing the number of ratings coming from online classes by the number of ratings for each professor (each row). For example, a professor who has 80/100 proportion of online ratings would be seen as the one who teaches more online classes than a professor who has 80/1000 proportion of online ratings. The null hypothesis is that teaching a lot of classes in online modality does not affect the average rating professors get. The alternative hypothesis is that teaching a lot of classes in online modality affects the average rating professors get, which indicates a two-tail test.

I split and labelled the data set as two groups with a high proportion of ratings from online classes and a low one, by using the median of the proportion of online ratings as a benchmark instead of the mean of the proportion, since the distribution of the proportion of online ratings is not normal (Figure 6). The median is 0, which results in the group with the low proportion having a sample size way larger than the sample size of the group with the high proportion, around 13240 versus 3601.

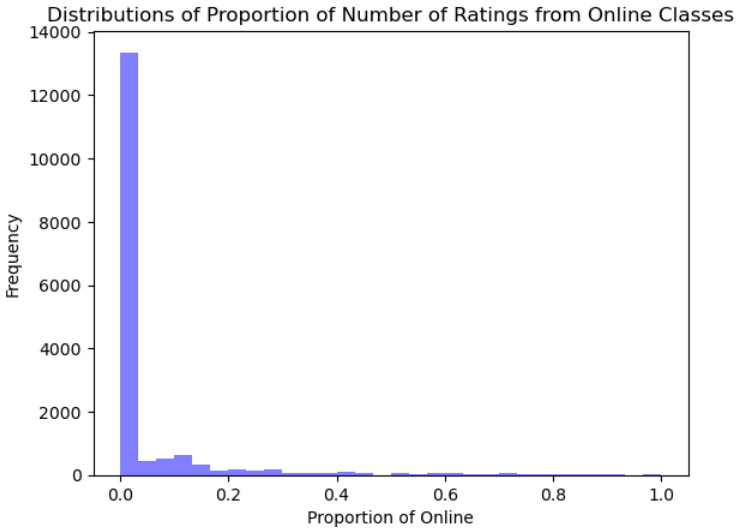


Figure 6

Therefore, due to this huge sample size imbalance, the shape of the distributions of these two groups is very different (Figure 7), which violates an assumption of the Mann-Whitney U test, so I decided to do a permutation test.

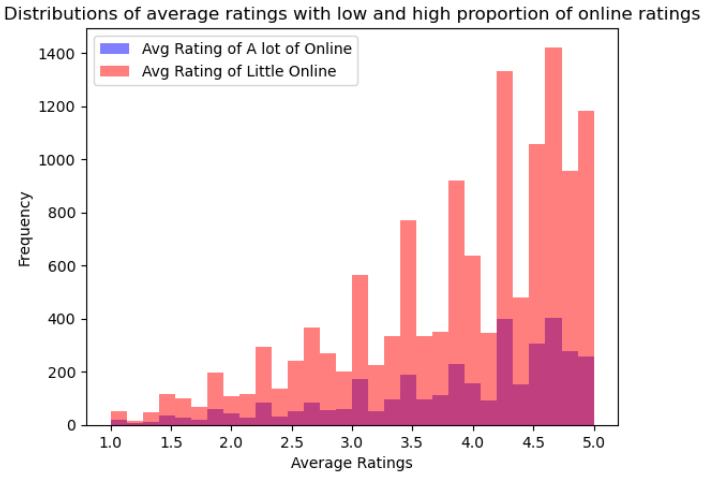


Figure 7

Due to the huge sample size differences, I designed a new test statistic called weighted_median_diff. It is essentially the difference between the medians of two groups, but it adds weights to the two medians to account for the difference in the size of the two groups. Finally, I got the p-value to be exactly 1.0, which means it was less extreme than

nearly all values in the null distribution, so I could not drop the null hypothesis. My conclusion is that whether professors teach a lot of online classes does not affect the average rating they get.

Question 5:

To detect the relationship between average rating and the proportion of people who would take the class the professor teachers again, I decided to do EDA first. Here is what I got:

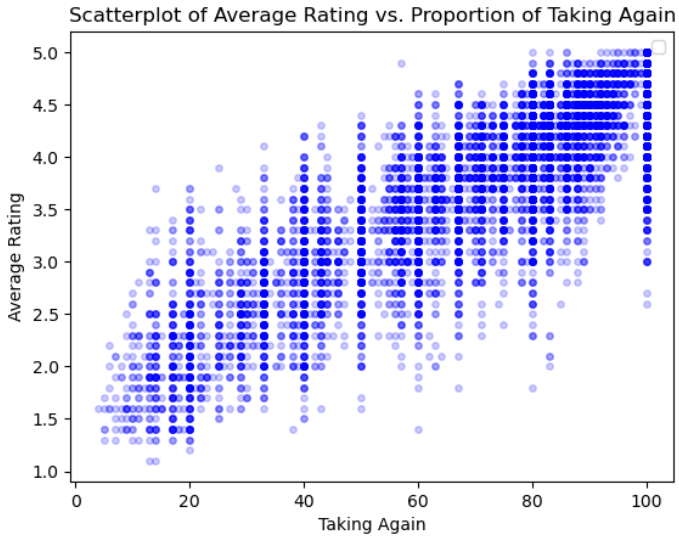


Figure 8

Since the mean (average rating) is not meaningful if the number of ratings is small, I still chose to proceed with the dataset: dataNum_thresh. Although the proportion of students who would take again is a ratio-type data, the other variable, average rating, is a rating-type of data, which is ordinal. Therefore, I chose to find out the Spearman's correlation coefficient, which is around 0.85, as my code shows. The rho indicates a moderately strong, positive, and monotonic relationship between average rating and proportion of students who would take this class again.

Question 6:

To detect if professors who received a "pepper" from students tend to have high average ratings, I decided to do a statistical significance test. The null hypothesis is that whether professors received "pepper" does not affect their average ratings. The alternative hypothesis is that professors who received a "pepper" from students tend to have higher average ratings than those who did not, which implies a one-tail test.

I proceeded with dataNum_thresh, as this question asks about the average rating. Then, I split and labelled the dataNum_thresh as two groups based on if the row has 1 or 0 in the pepper column.

After doing EDA (*Figure 9*), it can even be seen visually to say that these two groups do not have a similar shape of distribution, which violates the assumption of using the Mann-Whitney U test. Therefore, I decided to use the permutation test.

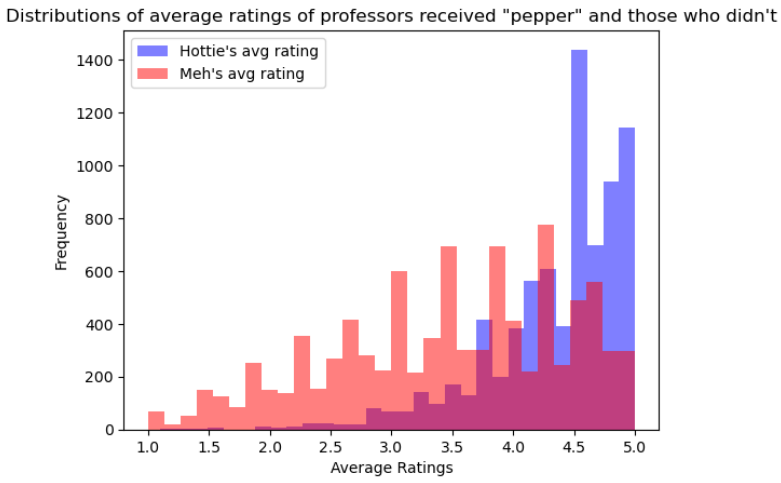


Figure 9

Since the average rating is ordinal data and the sizes of the two groups are similar (around 7000 vs 9000), I designed the test statistic to be the median of the hot group minus the median of the not hot group. I finally got the p-value to be 1e-04, which is so small that it can be treated as 0 and less than 0.005, showing that getting the result is statistically significant. Therefore, I dropped the null hypothesis, concluding that professors who are judged as “hot” by students tend to get higher average ratings from students.

Question 7:

As I did in the introduction, I dropped rows with values less than 7 in the number of ratings column, which means that the rows that I kept have meaningful average ratings. And since averaging ordinal ratings over many ratings makes the average behave approximately like a continuous variable, I decided to build a linear regression model, which is suited to continuous variables.

After fitting the linear regression model predicting professors’ average ratings from their average difficulty values, the model gives a coefficient of -0.75, indicating that for every 1-point increase in average difficulty, the average rating decreases by approximately 0.75 points. The intercept of 6.09 represents the predicted average rating for a professor

whose average difficulty score is 0. However, since the difficulty score ranges from 1 to 5, a difficulty of 0 is outside the valid range. And since the rating score also ranges from 1 to 5, an average rating being 6.06 is also not meaningful. This means the intercept is not directly interpretable in our real-world sense, but it is still necessary to show it here for the equation to work properly.

The model's R-squared value of 0.38 means that 38% of the variance in average rating is explained by average difficulty. The RMSE of 0.72 reflects moderate prediction error. While both variables are bounded and arguably ordinal but linear regression is not bounded and deals with continuous data, I treated the two variables as approximately continuous for the purpose of regression, which is the best I could do now. Since the distribution of residuals of this model is approximately normal, it also supports the use of linear regression (Figure 10).

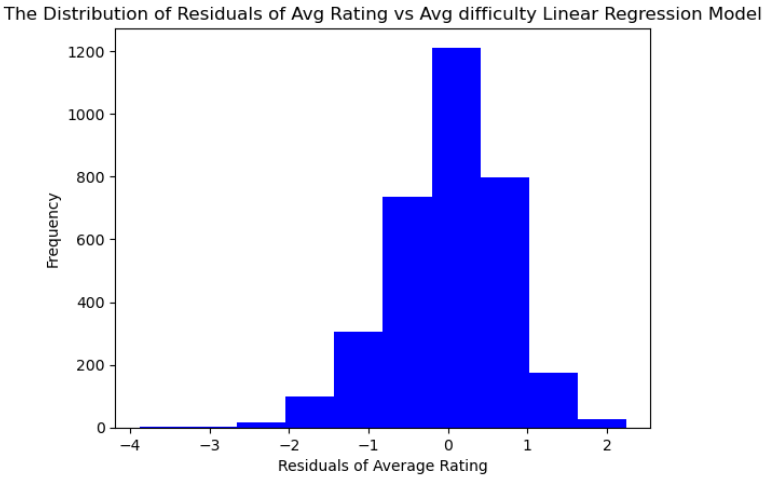


Figure 10

Question 8:

I proceeded with this question by using dataNum_again, which is the dataset that drops all nan rows if there is a nan in any column. The sample size now is 11252.

First, I looked at the correlation matrix and found that there is a moderate multicollinearity between variables, such as average difficulty and the proportion of students who would want to take the class again (Figure 11). And because I didn't think it was a good idea to regularize any betas to be 0, I decided to do multiple linear regression with the Ridge regularization, penalizing overly large coefficients and helping to address collinearity concerns.

	0	1	2	3	4	5	6	7
0	1	-0.598937	0.0489893	0.493024	0.877269	-0.0012039	0.0699128	0.00946698
1	-0.598937	1	0.00767239	-0.25144	-0.525402	-0.0233212	-0.0118182	-0.0423817
2	0.0489893	0.00767239	1	0.100399	0.0536668	0.0885356	0.0203737	-0.0433184
3	0.493024	-0.25144	0.100399	1	0.446474	0.0113351	-0.0176804	0.0280508
4	0.877269	-0.525402	0.0536668	0.446474	1	-0.00975908	0.0646551	-0.00110467
5	-0.0012039	-0.0233212	0.0885356	0.0113351	-0.00975908	1	-0.0289418	0.0672225
6	0.0699128	-0.0118182	0.0203737	-0.0176804	0.0646551	-0.0289418	1	-0.520372
7	0.00946698	-0.0423817	-0.0433184	0.0280508	-0.00110467	0.0672225	-0.520372	1

Figure 11

After I tried out different lambdas and saw the RMSE of the Ridge model with each lambda, I found that the model with a lambda of 0.01 has the lowest RMSE, so I set the lambda to 0.01.

I got these values after fitting the model. I then split the dataset into training and test sets with a ratio of 8/2. The model's R-squared value of 0.80 means that 80% of the variance in average rating is explained by average difficulty. The RMSE of 0.37 reflects a small prediction error. These metrics reflect a significant improvement in model performance, compared to the "difficulty only" model with an R-squared value of 0.38 and RMSE of 0.72. In addition, in this new model, the beta1 for average difficulty is -0.16, whereas the coefficient in the "difficulty only" model is -0.75. This shift suggests that the effect of average difficulty is less conspicuous once other factors are accounted for, and average difficulty is correlated with other predictors, such as the proportion of students who would want to take the class again. Overall, the model that uses all factors performs better.

Question 9:

Since the average rating is involved in this question, I still used dataNum_thresh, as the average rating is not meaningful if the number of ratings is low. After checking the size of the hot group and the not hot group, I found that the ratio of size is around 7/9, which is not that big to be a concern (Figure 12). If I used the raw unprocessed dataset by just dropping nan values on average rating and pepper column, but without filtering based on the number of ratings, I found that the class imbalance issue is revealed by plotting the size of the hot group and the not hot group (Figure 13). This means that filtering based on the

threshold of the number of ratings resolved the class imbalances. This supports the use of dataNum_thresh.

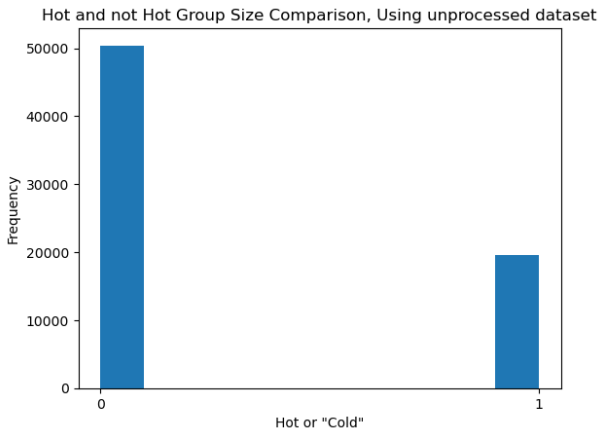


Figure 12

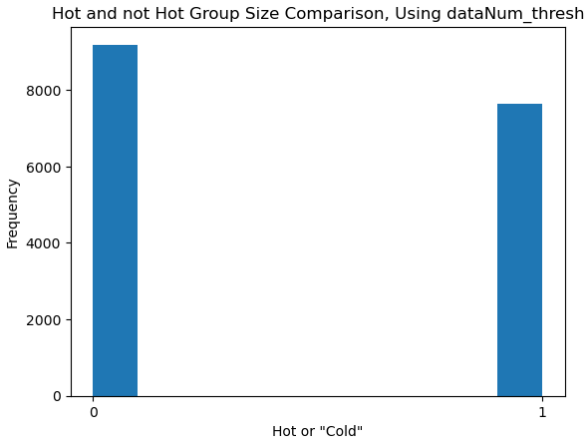


Figure 13

Here is the fitted logistic regression model (Figure 14). It can already be seen that this model would probably not be good, as there are a lot of professors with very good ratings but did not receive a "pepper". Then, I predicted whether professors would get a pepper or not, and I drew the ROC curve (Figure 15). The Area under the ROC is 0.79, which shows that the model performs moderately. As the default, I picked 0.5 as the threshold for this model. It provides the following quality metrics, rounded to two decimal places:
Precision: 0.68
Sensitivity: 0.73
Specificity: 0.71
NPV (Negative Predictive Value): 0.75
General Accuracy: 0.72

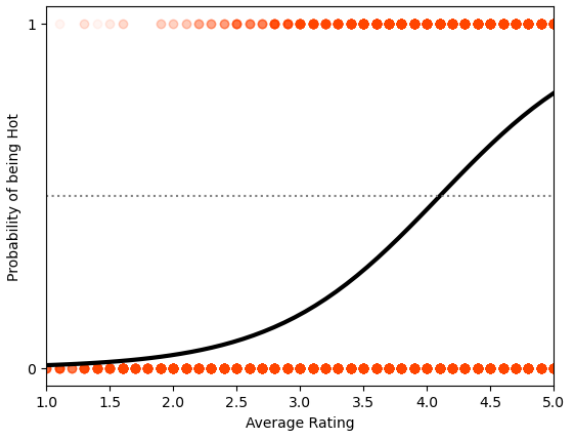


Figure 14

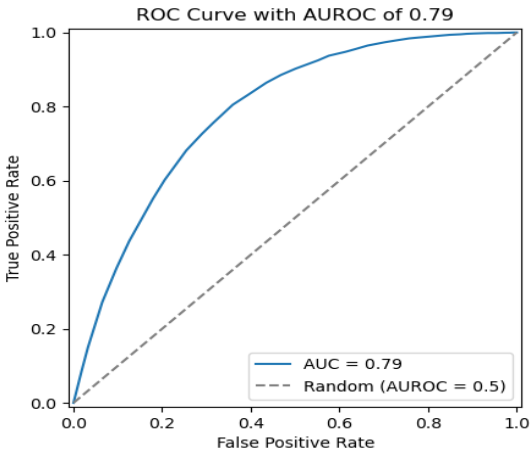


Figure 15

Question 10:

Since I used dataNum_thresh, which filtered out rows with a small number of ratings, the class-imbalance problem is automatically solved. The group sizes of pepper and no pepper are around 7000 and 9000, as said in question 9.

As introduced in question 8, there are variables correlated with each other in the dataset, so I decided to use Principal Component Analysis to address this issue. Firstly, I normalized the data by z-scoring them and instantiated the PCA object. After fitting the PCA using all factors, I got the eigenvalues of every principal component. I chose principal components based on the 90% eigensum criterion. Thus, I chose the first 5 PCs since they account for 90% of the variance.

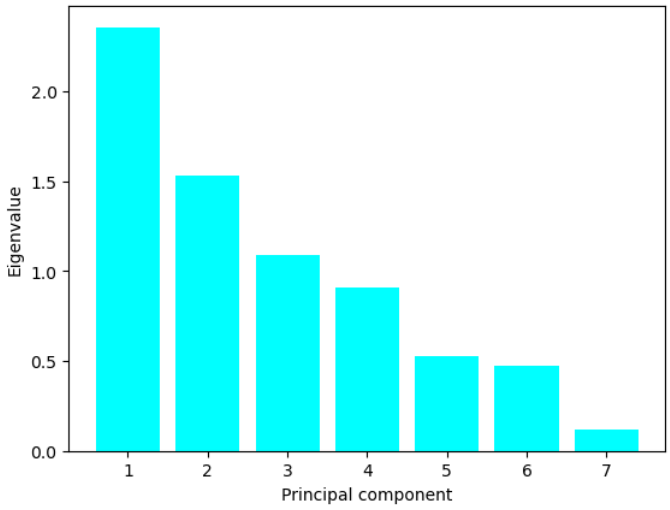


Figure 16

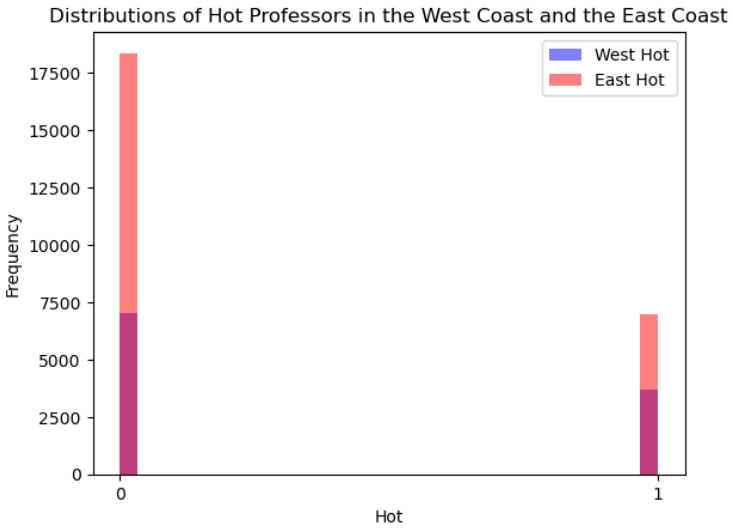
Then I built the classification model using logistic regression, and I got the AUROC of 0.80. I set the threshold to be 0.5, which is the default, and got the following metrics from this model:
Precision: 0.70
Sensitivity: 0.76 (higher than the "average only model" in Q9)
Specificity: 0.69
NPV (Negative Predictive Value): 0.74
General Accuracy: 0.72

Comparing the quality metrics of the all-factors model and the average-rating-only model, I can say that the all-factor model has slightly higher AUROC (0.80 vs. 0.79). It demonstrates marginally better precision (0.70 vs. 0.68) and sensitivity (0.76 vs. 0.73), suggesting it captures more true positives, though at the expense of slightly lower specificity (0.69 vs. 0.71). Notably, both models yield identical overall accuracy (0.72), and their negative predictive values are nearly the same (0.74 vs. 0.75), showing comparable

performance in predicting negatives. Ultimately, the choice between models depends on the application's tolerance for false positives and need for interpretability. The average-only model is simpler and still good enough, while the all-factors model may be preferable in scenarios where identifying positives is more critical, although it is harder to interpret the new principal factors after doing PCA.

Extra credit:

I wanted to find out if professors who teach on the East Coast or the West Coast would be considered hotter. First, I selected professors in dataQual into the West Coast group and the East Coast group, based on their value in their state column. Then I found the corresponding professors in dataNum, and I dropped rows with nan values in the "pepper" column in dataNum. To detect which Coast has more hotties, I decided to do a Chi-square test because the data in this question are all categorical: hot, not hot, west, east.



I calculated the p-value, and it's 6e-40, which is basically zero. It means that the result is significant, so the region (East Coast or West Coast) where the professor teaches affects whether the professor would be considered hot. Comparing the proportion of hot professors on the West Coast and the East Coast, which is 34.5% vs. 27.5%, I can conclude that professors on the West Coast would be considered hotter, generally.