Capstone Project
Group 05

N17151398 (seed number)

Group Members:
Asif Tauhid
Haojie Cai
Xiaokan Tian

Data Preprocessing:
1. **Data Loading**: We loaded the three datasets (numerical, qualitative, and tags data), with appropriate labeled column names. Then, since each row stands for the same professor, we merged them into a single dataframe.
2. **Missing Data**: We checked the number of missing data for each column, and dropped columns with no ratings (column *Number of Ratings* equals to *NaN*).
3. **Tags Normalization**: Since tags are counted cumulatively, professors receiving more ratings may have more tags. To address this, we divide each tag by the number of ratings for that professor. Moreover, we did standardization for each tag, with mean of 0 and standard deviation of 1, so that we can scale differences between tags, ensure comparability, and help with convergence in regression tasks.
4. **For machine Learning**: Since our sample size is relatively large and in order to minimize statistical noise, bias and impact of outliers and improve statistical reliability, we filtered out professors with *Number of Ratings* fewer than 5. The resulted dataset had around 25k rows. For questions involving column *Proportion Retaking Class*, we also dropped rows with *Proportion Retaking Class* equals to *NaN*. The resulted dataset had around 12k rows.

Question 1:

To answer this question, we first got the samples of male professors and female professors by checking whether columns *Male Gender* and *Female Gender* are set. For edge cases like none of them is set or both of them are set, we directly ignore them since our sample size was large and our target was to find the pro-male gender bias.

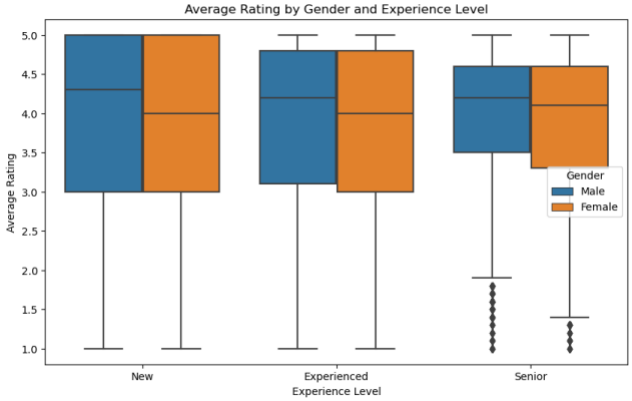
Before we conduct the significant test, we need to decide which test we should use for the tasks. As we checked the data pattern in the dataframe, we found that it had relatively good cardinality since average ratings and average difficulty could be decimal numbers, so it's reasonable to reduce the data to sample means. Not only that, our sample size is large which means we can apply the central limit theorem, and we didn't have information about the population parameters. Therefore, the t-test would be the best choice for us.

After we found the variance was different for the two samples, we did Welch t-test, and got a p-value of 9.59 * 10^{-12}, which means that we can reject the null hypothesis that there's no difference between the average ratings of male and female professors. There may exist a pro-male gender bias in the average ratings.

Confounders like teaching experience may affect the ratings of professors. To control the effect of teaching experience, we did a ks-test and found that there's a statistically significant difference between

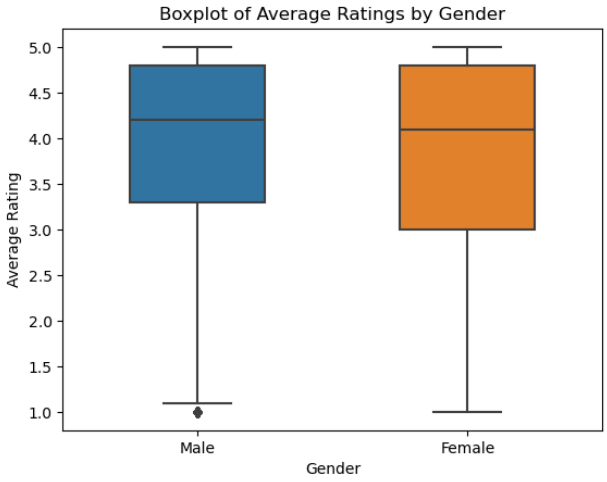
the distribution of the number of ratings of male and female professors. Therefore, we grouped professors into three categories based on their experience (column *Number of Ratings*): new professor groups with number of ratings less or equal to 3 (median); experienced professor group with number of ratings less or equal to 6 (75th percentile); senior professor groups with number of rating more than 6.

For each group, we did Welch t-test to check whether the average rating is significantly different for two samples. For new professor group, the test statistic is 4.71 and p-value is 2.51 * 10^{-6}; for experienced professor group, the test statistic is 5.41 and p-value is 6.47 * 10^{-8}; for senior professor group, the test statistic is 4.11 and p-value is 3.98 * 10^{-5}. This means for each group of teaching experience, there's a significant difference between the average ratings of male and female professors where male professors have higher ratings. This conclusion is same as the previous conclusion, and is supported by the box plot as below where male professors always have higher median ratings:



Question 2:

To investigate whether there is a significant difference in the variance (spread) of the ratings between male and female professors, we did Levene's test on the two samples. The resulted test statistic is 115.43 and the resulted p-value is 6.77 * 10^{-27}, which means that we can reject the null hypothesis that there is a significant difference between the variance of the average ratings of male and female professors. By directly calculating the variance for each sample, we found that female professors (variance of 1.31) have higher variance than male professors (variance of 1.18) on the average ratings. This can be visualized through the box plot for each sample:

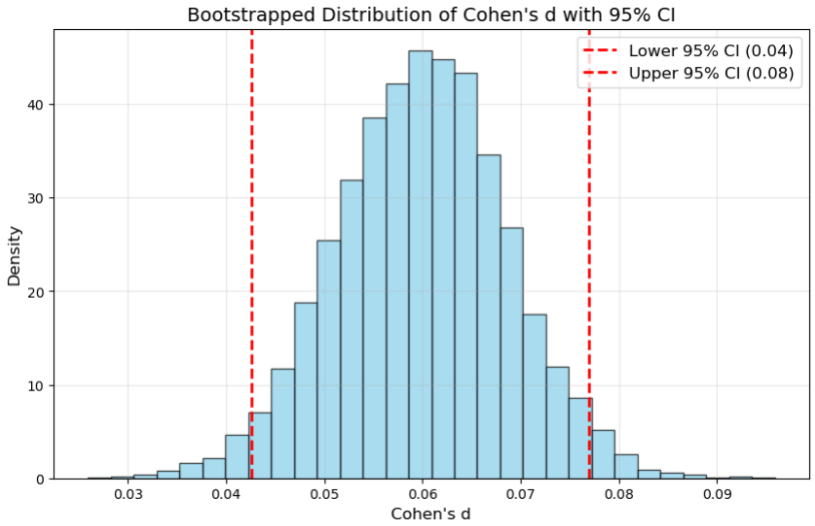


We also did Levene's test for each teaching experience level to control its confounding effect. For new professor group, test statistic is 64.64 and p-value is 9.35 * 10^{-16}; for experienced professor group, test statistic is 75.63 and p-value is 3.55 * 10^{-18}; for senior group, test statistic is 17.74 and p-value is 2.56 * 10^{-5}. This validates the conclusion that there is statistically significant difference in the variance of the ratings distribution between male and female professors.

Question 3:

To find the effect size of gender bias in average rating, we computed Cohen's d and got the resulted effect size of 0.0598. This effect size is relatively small. From previous significant tests, we found that there is a significant difference between male and female professors' average ratings, so we cannot say that there's no gender bias from the small effect size. However, this bias might have little real-world impact since the small effect size.

To access the 95% confidence interval for Cohen's d, we used bootstrapping for 10000 simulations. For each simulation, the samples for male and female professors remain the same size, and the ratings are taken as replaceable. Then we computed Cohen's d for each simulation and visualize it's distribution as below:



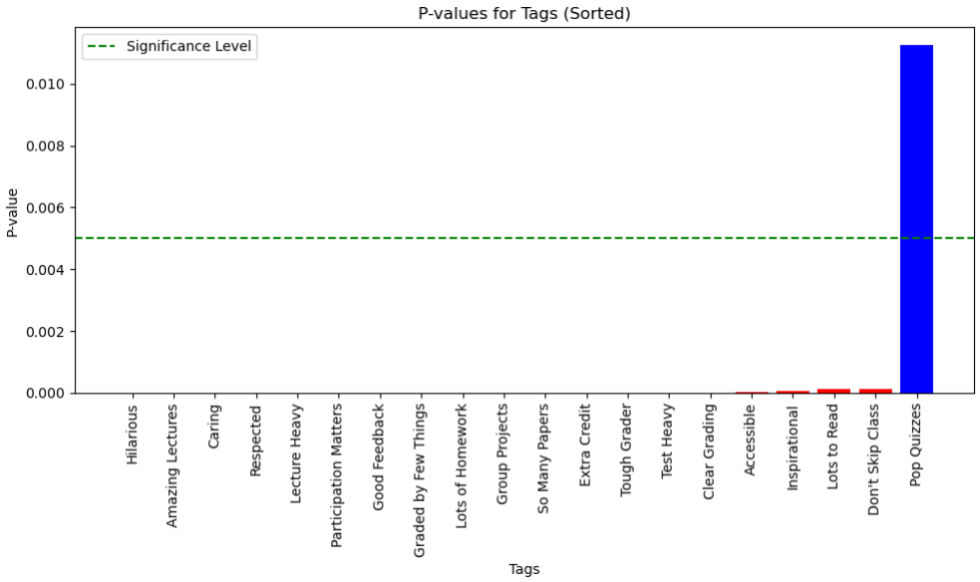
From all the simulations, we found that the 95% confidence interval of Cohen's d of gender bias in average rating is [0.0426, 0.0770].

To find the effect size of gender bias in the spread of average rating, we first used the idea of Cohen's d to calculate the standardized variance (the difference between variance of male and female professors' ratings divided by the pooled variance). After bootstrapping for 10000 simulations, we got the effect size of 0.1053, and 95% confidence interval of [0.0811, 0.1294]. This effect size is relatively small as well, which means that the gender bias in spread of average rating might have little real-world impact.

We also did some research as well on calculating the effect size for variance. F-ratio may be a better choice. The f-ratio of variance of male and female professors' average ratings is 0.8999, and the 95% confidence interval after bootstrapping for 10000 simulations is [0.8785, 0.9220]. Since the f-ratio is very close to 1, we could interpret that this gender bias might have little real-world impact in the variance of average ratings.

Question 4:

To investigate whether there is a gender difference in the tags, we did Welch t-tests for each of the 20 normalized tags on male professors and female professors samples. We saved the result (tags, test statistics, p-values, whether significant) into a dataframe and sorted it by p-value. Then, we visualized the resulted dataframe on p-value in a bar chart, as it is shown below:



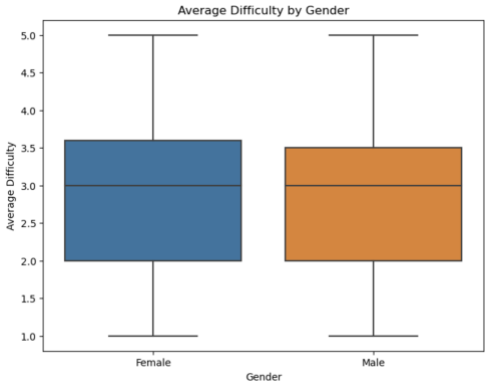
What we found is that, except for the tag *Pop Quizzes*, all tags have a p-value lower than alpha level and are statistically significant. Therefore, we are unable to reject the hypothesis that there is no gender difference in *Pop Quizzes* count, but we conclude that there is significant difference between male and female professors on all tag counts (except for *Pop Quizzes*) awarded by students. The three most gendered tags are *Hilarious*, *Amazing Lectures*, and *Caring*. The three least gendered tags are *Pop Quizzes*, *Don't Skip Class*, *Lots to Read*.

Question 5:

To check the gender difference in terms of average difficulty among professors, we performed Welch's t-tests, and got a T-statistic of -0.6473 and a P-value of 0.5174, which means that there is no significant difference in difficulty ratings between male and female professors.

Since the distributions of difficulty in samples are not normal, it's also reasonable to conduct Mann-Whitney U test. The Mann-Whitney U test gave a U-statistic of 39,784,981.0 and a P-value of 0.691, which is much higher than the significance threshold (alpha value of 0.005). This indicates that there is no statistically significant difference in difficulty ratings between male and female professors.

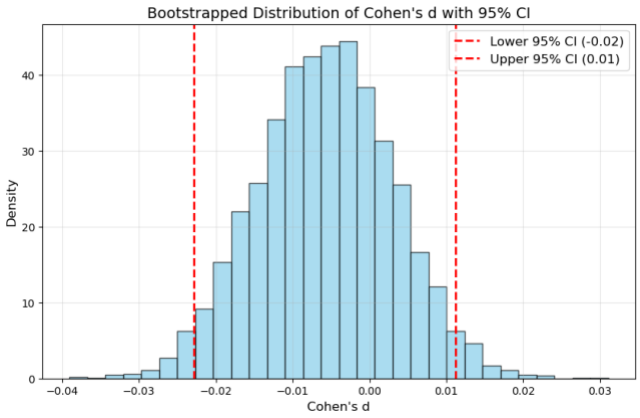
Further, to support these findings, we created a boxplot comparing the distributions of difficulty ratings for both genders. The boxplot also showed no significant differences in ratings for male and female professors. Therefore, this dataset suggests that gender does not significantly impact average difficulty ratings.

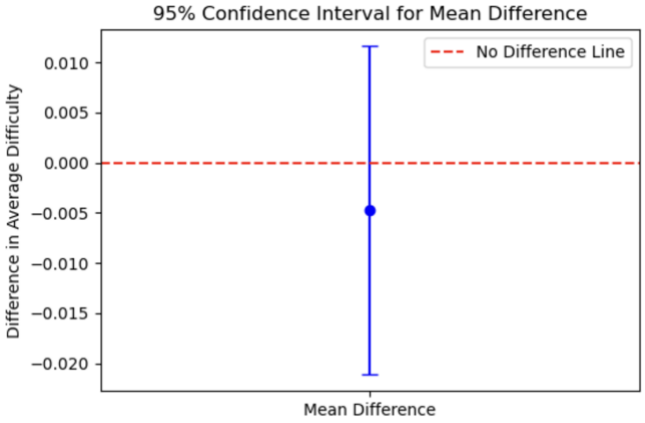


Question 6:

To quantify the likely size of the previous effect at 95% confidence, we found the 95% confidence interval for the difference in mean difficulty ratings between male and female professors by bootstrapping for 10000 simulations. This interval is [-0.0229, 0.0112], and it estimates the range within which the true effect size has a likelihood to fall into.

The confidence interval is very close to 0, which indicates that the effect size is small as well. Since the confidence interval includes zero and from the previous task we found that there is no statistically significant difference, there is no clear evidence that there exists a true difference. Therefore, the result suggests that the observed difference could easily be due to mere chance.
Thus this further supports the conclusion that gender does not have a meaningful impact on average difficulty ratings.



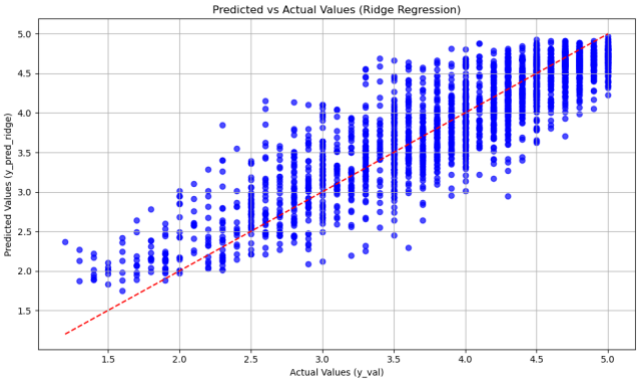


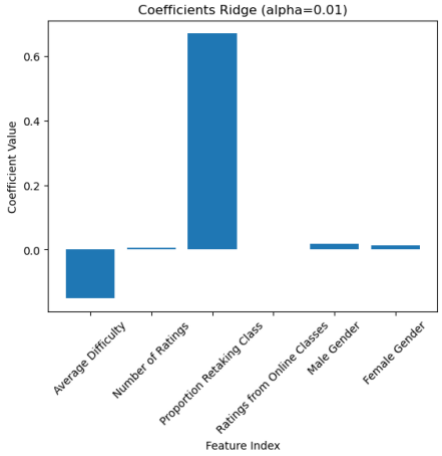
Question 7:

For this question, we used the dataset with 12k rows, which is the one that filtered out rating count is less than 5 and none proportion retaking class.

To predict the Average Rating of professors, we tried three representation models, Linear, Ridge, and Lasso. Firstly, we splitted the dataset into train and test (20% test). Secondly, features were standardized by using the StandScaler function to ensure consistent scaling. Thirdly, for Ridge and Lasso, the hyperparameter, alpha, was tuned to identify the best-performing models. Finally, we compared R2 and RMSE for each model, and found the best one.

Through comparing R2 and RMSE, the Ridge model has the best performance, with R2 = 0.8014, RMSE = 0.1353. And the change of alpha doesn't affect Ridge's R2 and RMSE a lot. For the figure on the left, the blue points show the relationship between actual values and predicted values. The red line shows the perfect prediction line. We can see that for the ratings that are higher than 3.0, the predicted values align reasonably well with the diagonal line. But for lower ratings, the model doesn't perform well. According to the figure on the right, *Average Difficulty* (negatively) and *Proportion Retaking Class* (positively) are strongly predictive of average rating.





Question 8:

For this question, we used the dataset with 25k rows, which is the one that filtered out rating count is less than 5.

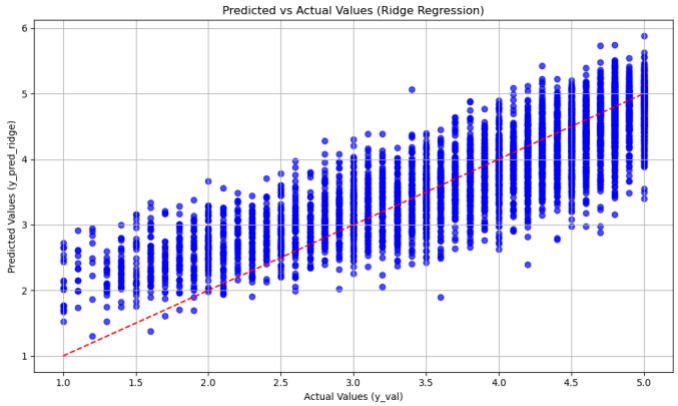
We tried 2 regression models, Ridge and Lasso. Firstly, we splitted data into train and test (20% test). Secondly, we did standardization for the data. Thirdly, we tried different alpha, and found the most suitable value. Finally, we compared R2 and RMSE for each model, and found the best model.

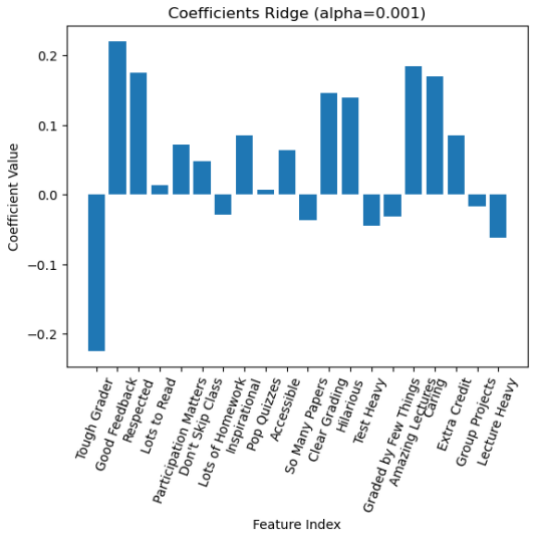
Under alpha = 0.01, the performance of Ridge and Lasso regression models are similar, and Ridge has a little bit lower RMSE and higher R2, which are 0.2541 and 0.7267 respectively. The change of alpha doesn't affect Ridge's R2 and RMSE a lot, which means the dataset has low multicollinearity. From the figure on the left, we can see that the predicted values align reasonably well with the diagonal line. According to the figure on the right, we can see that the two strongest factors are "Tough Grader" (negatively), "Good Feedback" (positively).

Question 9:

For this question, we used the dataset with 25k rows, which is the one that filtered out rating count is less than 5.

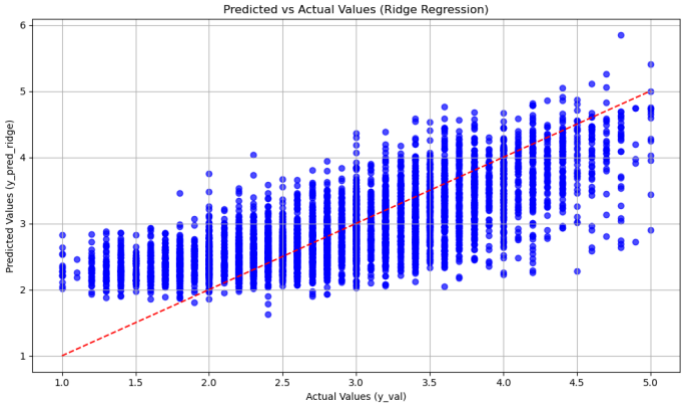
We tried 2 regression models, Ridge and Lasso. The process is very similar to the previous question. We first Split the data, then did standardization, tested different alpha, compared RMSE and R2,

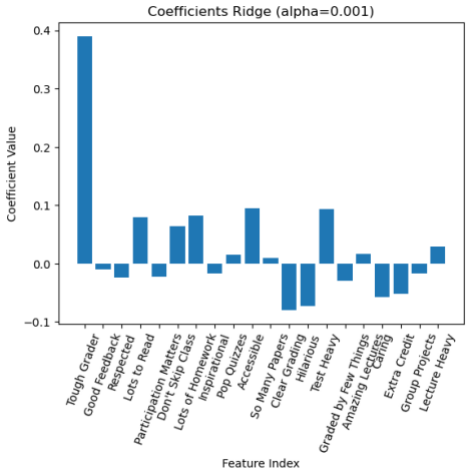




and found the best model.

Under alpha = 0.01, the performance of the Ridge model is a little bit better than Lasso model, with RMSE = 0.3011, R2 = 0.5600. The change of alpha doesn't affect Ridge's R2 and RMSE a lot, which means the dataset has low multicollinearity. From the figure on the right, we can see that the model performs well in the range of 2.5 to 4.0. The figure on the right is interesting, we can see that the "Tough Grader" tag is very strongly predictive of the "Average Difficulty".



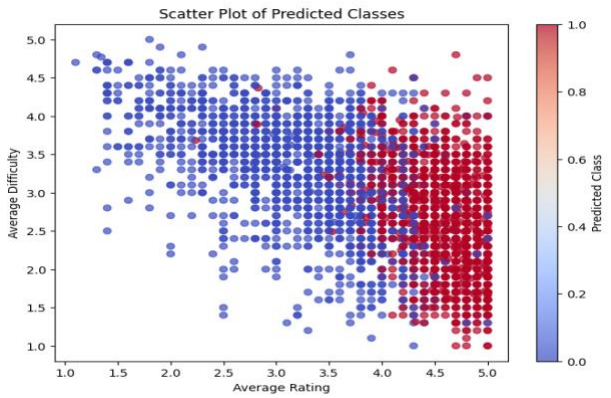


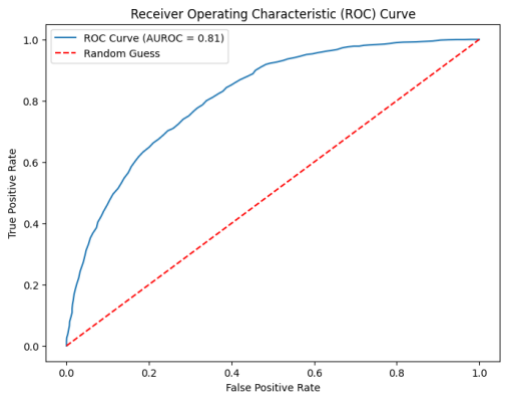
Question 10:

To predict whether a professor receives a Pepper, we merged both numerical and tag columns and excluded the target variable (Received a Pepper). Then, we addressed the class imbalance using SMOTE to oversample the minority class and trained a Random Forest Classifier. Finally, we evaluated the model using metrics like AUROC, precision, recall, F1-score, and the ROC curve.

The model actually performed well, achieving an AUROC of 0.81 that indicates excellent differences between the classes. Along that, the classification report showed balanced precision, recall, and F1-scores for both classes (class 1 indicating pepper and class 0 indicating no Pepper), with an overall accuracy of 75%. These results suggest that the model effectively predicts whether a professor receives a Pepper.

Below is the scatter plot for our prediction model on received a pepper by two features: average rating and average difficulty. The ROC curve also visually demonstrated strong model performance. Here the curve deviated significantly from the random guess baseline. This confirms that the model is pretty reliable in predicting the target variable.





Extra Credit:

Here, for the extra credit, we tried to analyze two key aspects of the dataset. First, we tried to figure out which professors from what U.S. states get more Pepper and the correlation between the tags and the professor getting Pepper. To do so, we examined the proportion of professors receiving a Pepper across U.S. states using Bayesian smoothing to adjust for small sample sizes to ensure fairness for states with fewer professors. Also to maintain reliability, states with fewer than 10 professors were excluded. This approach revealed that professors in "PE" had the highest adjusted proportion at 44.15%, compared to the national average of about 25%, demonstrating regional differences in terms of receiving Pepper.

We also explored how teaching traits (tags) related to receiving a pepper by calculating correlations. The tags "*Caring*," "*Good Feedback*," and "*Respected*" were the strongest predictors, with

correlation values of 0.256, 0.246, and 0.243, respectively that show that professors perceived as approachable and engaging were more likely to be considered as hot. On the other hand, traits like "Tough Grader" showed weak and insignificant negative correlations, indicating little impact on receiving pepper.