Capstone Report

Team Member:
Kerr Tan, Mina Sha, Eric Han

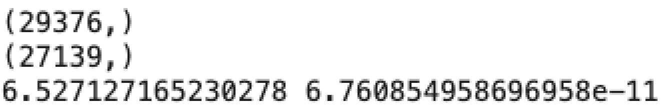
Contribution:
Kerr Tan: #1 - #4 Questions + Corresponding part of report
Mina Sha: #9, #10 +Extra Credit + Corresponding part of report
Eric Han: #5 - #8 + Corresponding part of report

Introduction:
To start the project, the random seed number has been set to the N-Number belonging to Kerr Tan (N10268952) for unique identification purposes, and for most of the significance testing, the random state will be set to this unique random seed number. The three datasets in this project contain no header. Hence, when loading the datasets, our team updated the column names to each of the dataset for better interpretation. For **question 1 and 2** regarding the ratings, due to comparison purposes, there are two types of handling data: 1) accepting all the data; 2) accepting the data with over k people ratings. Accepting all the data is to preserve as much as data points since the dataset size isn't large, and the reason for setting a threshold is to utilize meaningful data with excluding rating bias for a specific professor. In each of the significance tests, the nan policy of handling null value is to omit. For **question 3**, to examine gender bias in average rating and spread of average rating, our group adopted a difference between two groups. For **question 4**, our group merges the *tags* dataset with the *nums* datasets accepting all data records for not losing too many data points. For **question 5** and **question 6**, we did similar operations to test gender bias in average difficulty, but we improved our threshold from mean to median to reduce the impact of extreme values. In **question 7**, we built a linear regression model to predict average ratings from numerical predictors, and we chose to fill the missing values with means to prevent loss of valuable information and keep means of columns unchanged at the same time. For **question 8** we did the same thing to predict average ratings from all tags. For **question 9** we did the same thing to predict average difficulty from all tags. For **Question 10**, we developed a classification model to predict whether a professor receives a "pepper" using all available factors, including tags and numerical features. To ensure robust evaluation, we included metrics such as AU(RO)C and implemented methods to address class imbalance. For **Extra Credit**, we investigated whether Average Rating and Average Difficulty differed significantly across states.

Question Analysis
1. Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to a low sample size – as small as n = 1 (Mitchell & Martin, 2018),

failure to control for confounders such as teaching experience (Centra & Gaubatz, 2000) or obvious p-hacking (MacNell et al., 2015). We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset. Hint: A significance test is probably required.

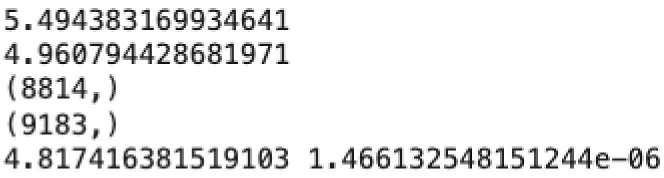
To answer this question, we utilized dataset *nums* and separated two groups based on columns of 'Male Gender' and 'Female Gender'. If we are accepting all the data points, male professors groups contain a size of 29376 while the female professors group contains a size of 27139. To investigate if there is a gender bias in student evaluation of professors, we conducted an independent t-test based on the null hypothesis that there is no difference between average expected ratings between male and female professors while assuming equal variance in the ratings. The p-value for this testing is 6.79e-11 (see Figure 1), which is much lower than the significance level of 0.005. It means that the testing **rejects the null hypothesis**, and our analysis provides sufficient evidence to conclude that there **is** a difference between average rating for male and female professors.



```
(29376,)
(27139,)
6.527127165230278 6.760854958696958e-11
```

Figure 1. All Data: Group Size and t-test Results

On the other hand, since there might be a chance that some professors only have one rating from a student, which might put too much weight on the average rating that might cause bias in evaluating the performance of this professor. As a result, we chose a threshold of the mean value of the number of ratings in each of two groups. Male group has an average number of ratings of 5.49, meaning that each male professor has on average 5.49 student ratings. Female group has an average number of ratings of 4.96, meaning that each female professor has on average 4.96 student ratings. Therefore, we separated two groups by only accepting data points that were greater or equal to the corresponding number of ratings. Now, male group only has a size of 8814, while the female group has a size of 9184. Samely, we conducted an independent t-test based on the same null hypothesis, finding that the p-value is 1.47e-06 (See Figure 2). It is still lower than the significance level of 0.005, and it **rejects the null hypothesis**, and our analysis provides sufficient evidence to conclude that there **is** a difference between average rating for male and female professors.

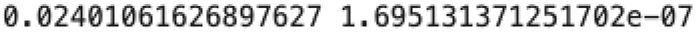


```
5.494383169934641
4.960794428681971
(8814,)
(9183,)
4.817416381519103 1.466132548151244e-06
```

Figure 2. Threshold: Mean Number of Ratings, Group Size, and t-test Results

2. Is there a gender difference in the spread (variance/dispersion) of the ratings distribution? Again, it is advisable to consider the statistical significance of any observed gender differences in this spread.

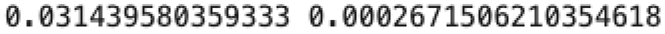
To investigate the gender difference in the spread of the ratings distribution, we conducted a Kolmogorov-Smirnov (KS Test) with the null hypothesis that assuming there is no difference between the average rating distribution in male and female professor groups. The method chosen for the KS test uses 'two-sided' while considering the overall distribution difference between two groups while handling null values with the 'omit' policy. Using all the data without setting threshold, the p-value of the KS test is 1.69e-07 (See Figure 3). It is statistically significant with lower than 0.005, and it **rejects the null hypothesis**, meaning that our analysis provides sufficient evidence to conclude that there **is** a difference between spread of the ratings distribution for male and female professors.



```
0.02401061626897627 1.695131371251702e-07
```

Figure 3.All Data: KS Test Result

On the other hand, for groups that have a threshold, the p-value for the KS test is 0.00027 (See Figure 4). It's still lower than the significance level of 0.005, also meaning it **rejects** the null hypothesis. meaning that our analysis provides sufficient evidence to conclude that there **is** a difference between spread of the ratings distribution for male and female professors.



```
0.031439580359333 0.0002671506210354618
```

Figure 4. Threshold: KS Test Result

3. What is the likely size of both of these effects (gender bias in average rating, gender bias in spread of average rating), as estimated from this dataset? Please use 95% confidence and make sure to report each/both.

The likely size of gender bias in average rating is used Cohen's d formula to compute. Since we are comparing two groups with different sizes, the pooled standard deviation is manipulated according to different group sizes as indicated in the coding part. Similarly, the likely size of gender bias in the spread of average rating is conducted using a similar approach, whereas the numerator represents the difference of variance between gender in average rating, and the pooled variance manipulated according to different group sizes as the denominator. Then, we conducted a bootstrap analysis with simulations number of 1000 to capture the differences between average rating and variance of average ratings between two groups, respectively shown in Figure 5 and Figure 6. The confidence interval is restricted to 95%, with the lower bound at 2.5% while the upper bound at 97.5%. We also included the Cohen's d calculation and likely size of variance calculation in two graphs to have a better visualization, and both calculation results are within the 95% confidence interval for two scenarios. With both confidence intervals not capturing 0 as shown in the below graphs, it implies that there is a significant difference in both average rating difference and variance of average rating difference.

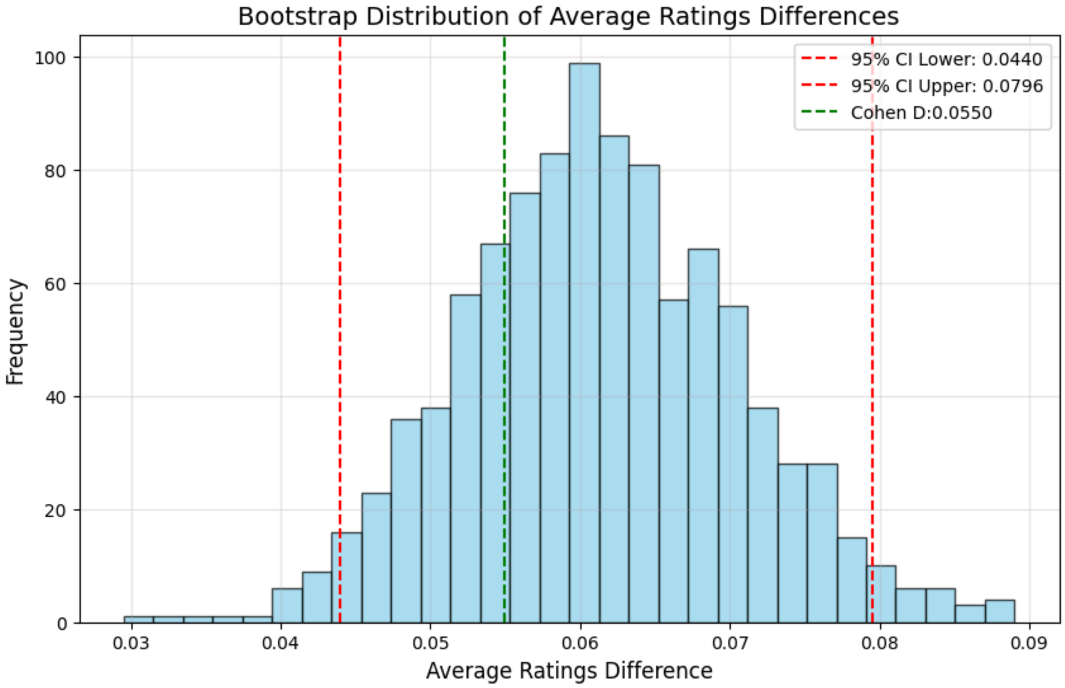


Figure 5. 95% Confidence Interval of Likely Size (Average Rating Difference)

From Figure 5, we can identify that the likely size for average rating difference between male and female range from **0.0440** to **0.0796**, with the Cohen's d calculation being **0.0550**.

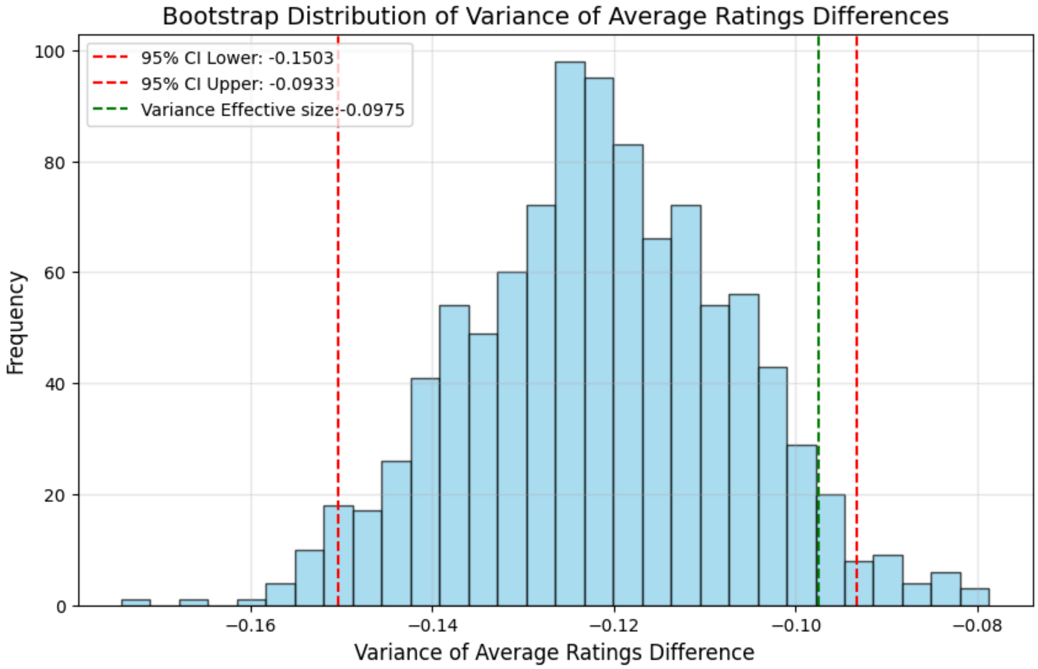


Figure 6. 95% Confidence Interval of Likely Size (Variance of Average Rating Difference)

From Figure 6, we can identify that the likely size for variance of average rating difference between male and female range from **-0.1503** to **-0.0933**, with the variance effective size calculation being **-0.0975**.

4. Is there a gender difference in the tags awarded by students? Make sure to teach each of the 20 tags for a potential gender difference and report which of them exhibit a statistically significant difference. Comment on the 3 most gendered (lowest p-value) and least gendered (highest p-value) tags.

We conducted a normalization of tags between male and female professors after joining the *tags* dataset with *num* dataset and dividing into two groups based on gender. The normalization approach is conducted based on the number of ratings for each professor, in which the number of each tag for each professor will be divided based on the number of that specific professor. This is because some professors will receive a greater number of reviews than others, and one student can comment on a professor with up to 3 tags. By doing so, we now have the normalized data of tags given the number of ratings for that specific professor. In the previous question, we are assuming homogeneity of variance for comparing two groups, but with the normalized approach we conducted in this question, we would like to verify the distribution of each tag for both genders before choosing the significance testing. We conducted a distribution plot (Figure 7) for each tag for both genders, and we found that for all the tags, it has a left-skewed problem, which implies that it will not be reasonable to compare two groups' mean values.

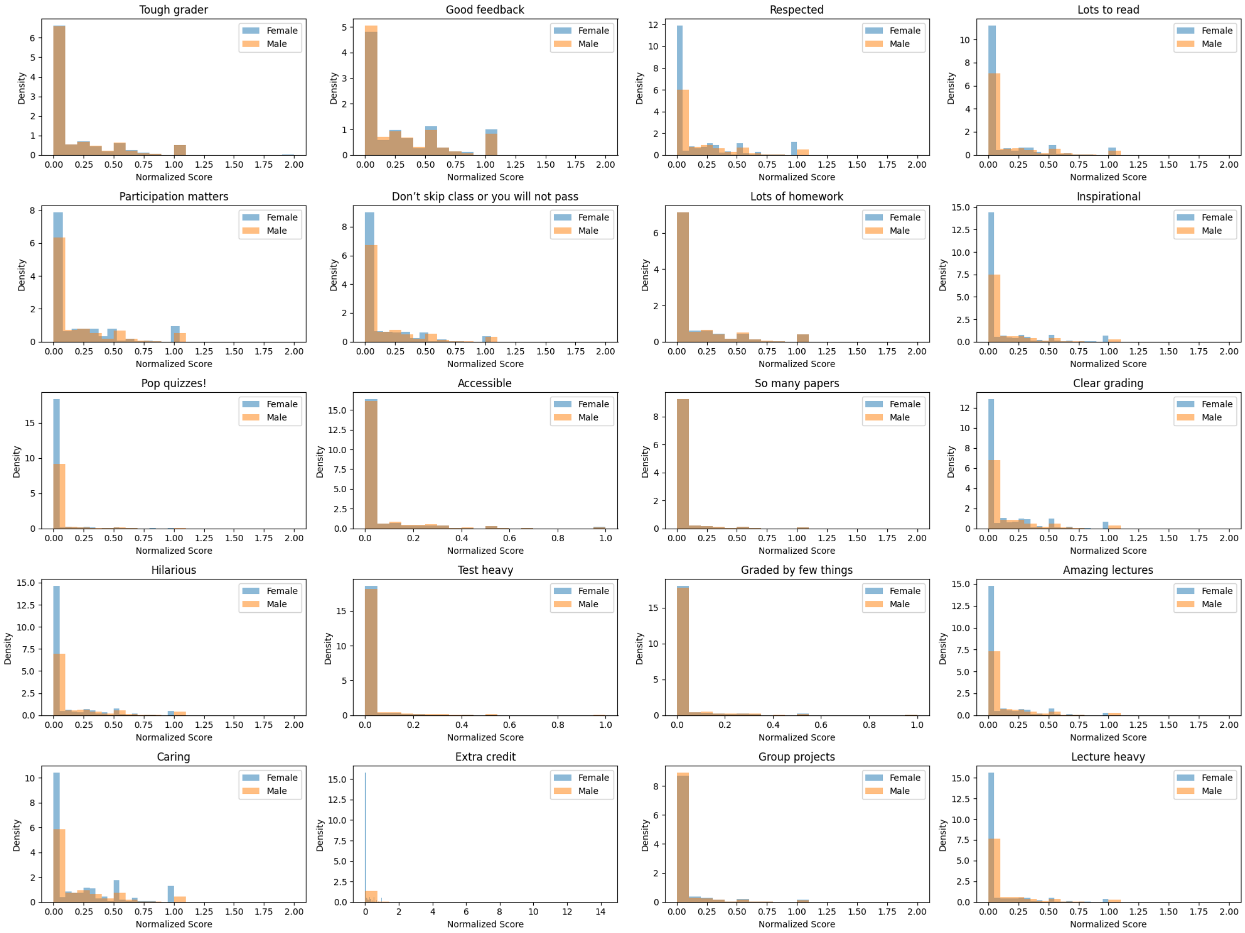
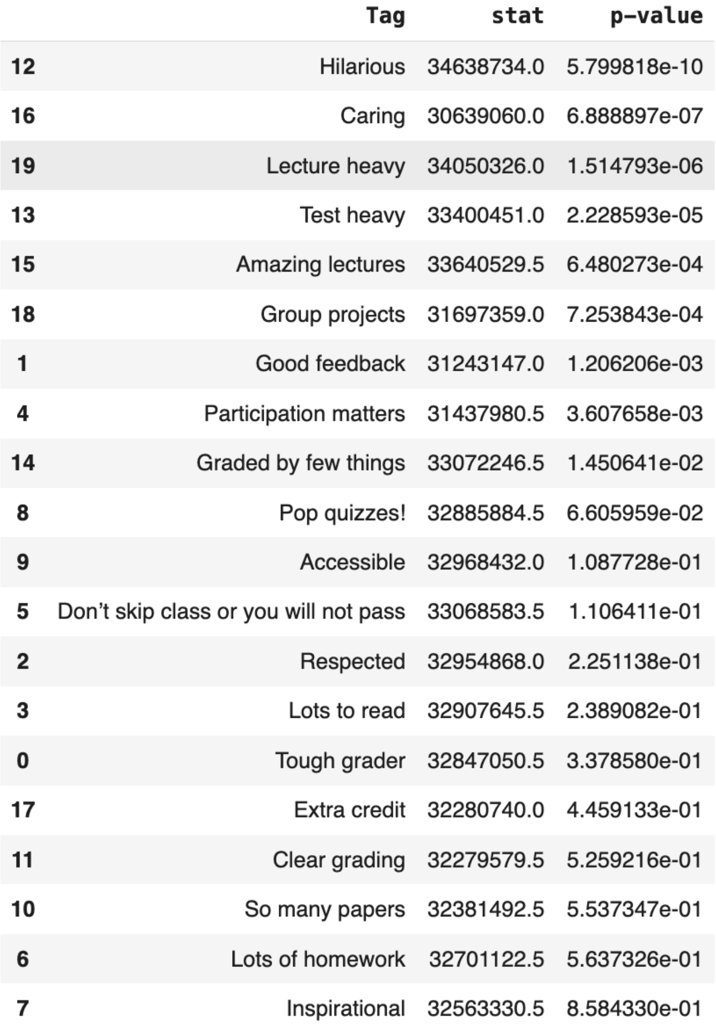


Figure 7. Distribution of Each Tags with Normalized Data for Both Gender

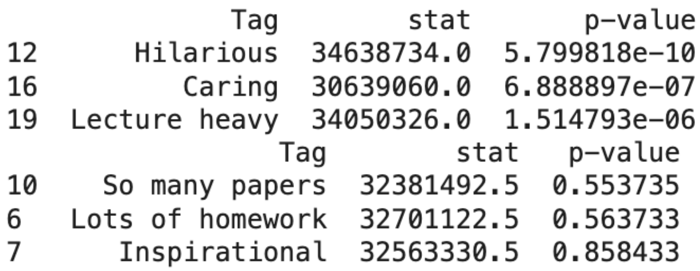
As a result, we conducted a Mann-Whitney U test with the null hypothesis that there is no difference in median for each tag in both male and female professors. (Figure 8)



	Tag	stat	p-value
12	Hilarious	34638734.0	5.799818e-10
16	Caring	30639060.0	6.888897e-07
19	Lecture heavy	34050326.0	1.514793e-06
13	Test heavy	33400451.0	2.228593e-05
15	Amazing lectures	33640529.5	6.480273e-04
18	Group projects	31697359.0	7.253843e-04
1	Good feedback	31243147.0	1.206206e-03
4	Participation matters	31437980.5	3.607658e-03
14	Graded by few things	33072246.5	1.450641e-02
8	Pop quizzes!	32885884.5	6.605959e-02
9	Accessible	32968432.0	1.087728e-01
5	Don’t skip class or you will not pass	33068583.5	1.106411e-01
2	Respected	32954868.0	2.251138e-01
3	Lots to read	32907645.5	2.389082e-01
0	Tough grader	32847050.5	3.378580e-01
17	Extra credit	32280740.0	4.459133e-01
11	Clear grading	32279579.5	5.259216e-01
10	So many papers	32381492.5	5.537347e-01
6	Lots of homework	32701122.5	5.637326e-01
7	Inspirational	32563330.5	8.584330e-01

Figure 8. U Test Results for Each Tag in Both Groups

We sorted the results based on the p value, and found out that the three most gendered tags are **hilarious, caring,** and **lecture heavy**, and the three least gendered tags are **so many papers, lots of homework, and inspirational**, as shown in Figure 9. The most gendered tags are having p-value much less than 0.005, indicating **there is a statistical significance** which **rejects the null hypothesis** and provides evidence that these **three tags contain gender bias**. As for the least three tags, they have a large p-value that is much higher than 0.005, indicating that there is **no statistical significance** with **failing to reject the null hypothesis**, indicating that there is **no gender bias** for these three tags.

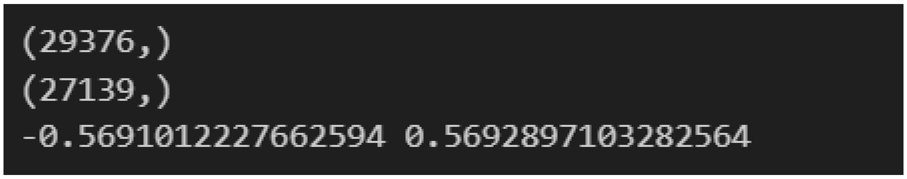


```
               Tag        stat       p-value
12       Hilarious  34638734.0  5.799818e-10
16          Caring  30639060.0  6.888897e-07
19   Lecture heavy  34050326.0  1.514793e-06
                  Tag        stat   p-value
10     So many papers  32381492.5  0.553735
6   Lots of homework  32701122.5  0.563733
7       Inspirational  32563330.5  0.858433
```

Figure 9. 3 Most Gendered Tags and 3 Least Gendered Tags

5. Is there a gender difference in terms of average difficulty? Again, a significance test is indicated.

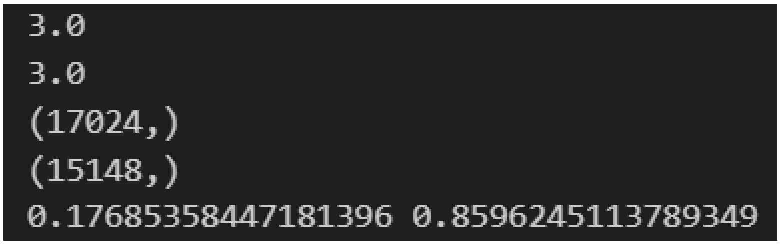
In this question we did the similar operation as question 1, while this time we are focusing on the "Average Difficulty" column. We applied an independent t-test with the null hypothesis that there is no difference between average expected difficulty between male and female professors assuming equal variance in the ratings. The **p-value is about 0.57**, which is much higher than the significant level 0.005, indicating that **we failed to reject the null hypothesis, and there is no statistically significant difference in average difficulty between male and female professors**.



```
(29376,)
(27139,)
-0.5691012227662594 0.5692897103282564
```

Figure 10. All Data: Group Size and t-test Results

For threshold groups, we chose the median of the number of ratings (which is 3.0 for both groups), so we can exclude the influence of extreme values, and include the majority of the data for our analysis. This time we did maintain a large proportion of the original group sizes for the two groups, while the **p-value was even higher and came to 0.86**, which means that **we still failed to reject the null hypothesis** even after filtering the data.



```
3.0
3.0
(17024,)
(15148,)
0.17685358447181396 0.8596245113789349
```

Figure 11. Threshold: Median Number of Ratings, Group Size, and t-test Results

6. Please quantify the likely size of this effect at 95% confidence.

Similar to **question 3**, we first used Cohen's d to calculate the size of effect, and then plot out the Bootstrap Distribution of Difference with 95% confidence interval. The **effect size is -0.0048**, which is a very low value, indicating that **there is almost no difference between average difficulties of male and female professors**. From the histogram we can see that the **confidence**

interval ranges from -0.0214 to 0.0115, which includes 0, also indicating that there is no sufficient evidence to show significant differences in difficulties between the two groups.

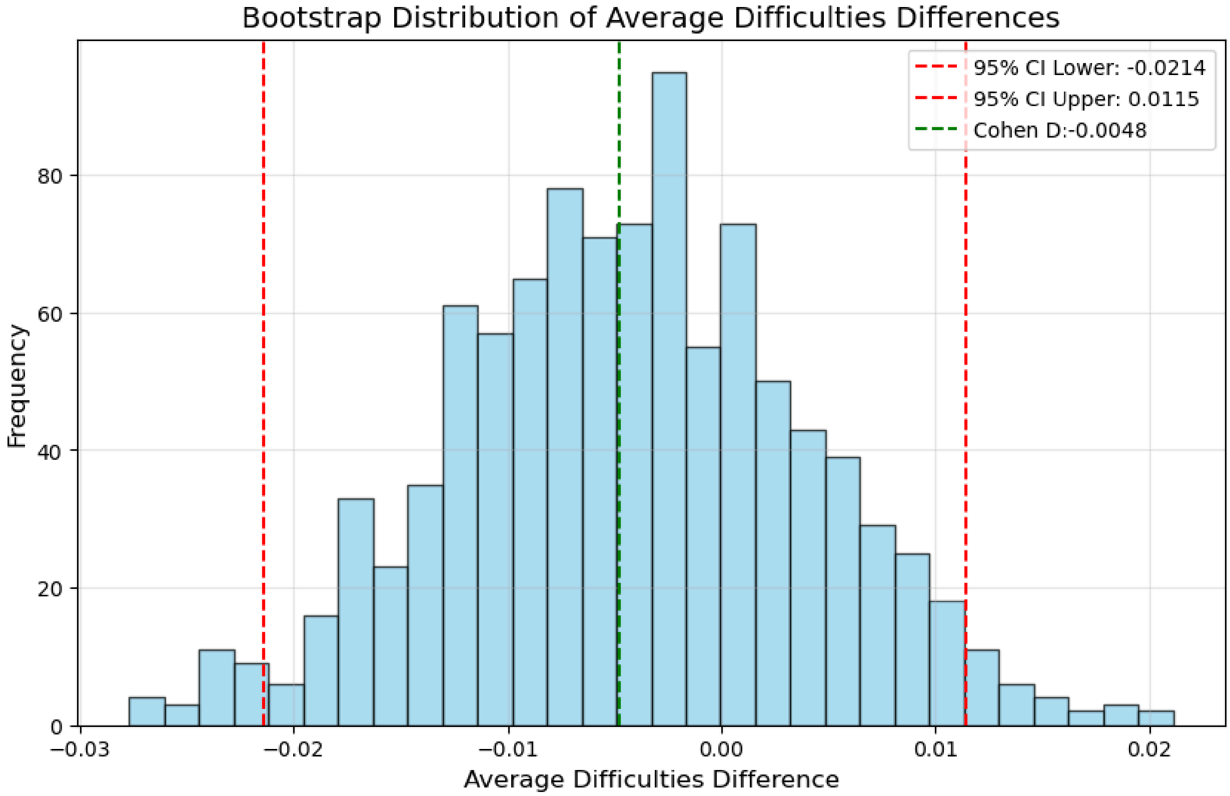


Figure 12. Bootstrap Distribution of Average Difficulties Differences

7. Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R2 and RMSE of this model. Which of these factors is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns.

In this question we were asked to build model to predict average rating from all numerical predictors in the rmpCapstoneNum.csv dataset, so we firstly separated predictors (X): the columns besides “Average Rating”, and set “Average Rating” as target (y). We filled the missing values with means, so we can prevent loss of valuable information while keeping the means of columns unchanged. Then we calculated the Variance Inflation Factor (VIF) for each predictor, and dropped highly collinear features (VIF=10 as threshold). Then we trained a linear regression model with 80% of the dataset as training sets and 20% as testing sets. The model achieved **an R^2 value of 0.3505 and an RMSE value of 0.8028**. This means that the model explains approximately **35.05% of the variance** in average ratings, and has an average prediction error of about **0.8**, indicating that the model leaves a significant portion of variance unexplained, and significant inaccuracy. All these results show that **this model is not very ideal**, but since there are considerable biases and noises in this data set, it may be reasonable to have such performance for this model. However, if we look at the fit line of Predicted vs Actual Average Rating, we can see **that the slope of the line is close to 1, meaning the predictions are linearly proportional to the actual values, and the intercept is also very small**, which shows that although the model may have a bad performance when predicting individual rating, it still shows the overall trend.

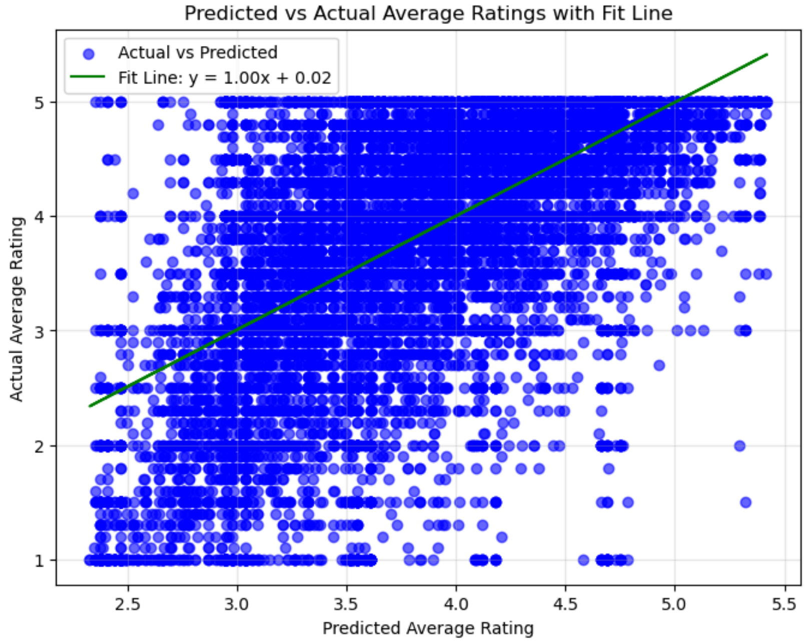
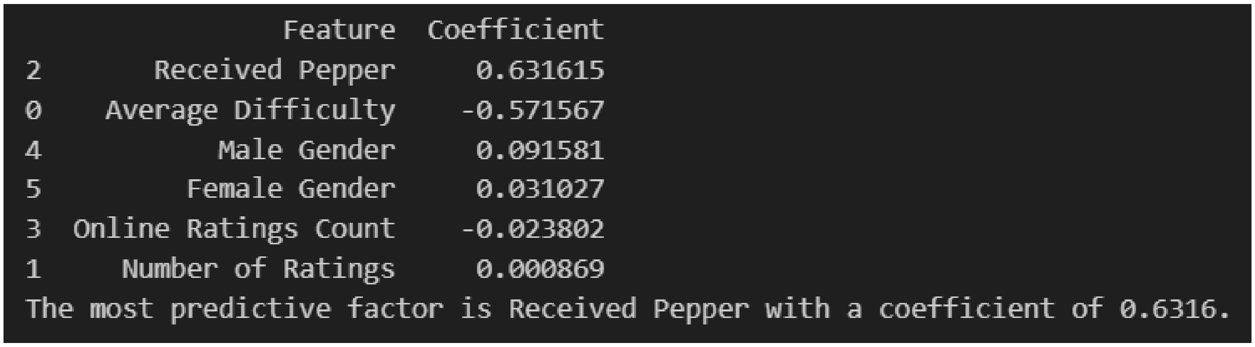


Figure 13. Predicted vs Actual Average Ratings with Fit Line

In the end, we checked the most predictive factor by ranking their predictive power based on the coefficients of the predictors. **Received Pepper** was the most predictive factor among all the predictors, with the coefficient of **0.6316**. This shows that whether this professor was judged as "hot" by the student can influence the prediction of the rating a lot.



```
                Feature  Coefficient
2       Received Pepper     0.631615
0    Average Difficulty    -0.571567
4           Male Gender     0.091581
5         Female Gender     0.031027
3  Online Ratings Count    -0.023802
1     Number of Ratings     0.000869
The most predictive factor is Received Pepper with a coefficient of 0.6316.
```

Figure 14. Feature Importance for Average Rating Prediction

8. Build a regression model predicting average ratings from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R2 and RMSE of this model. Which of these tags is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns. Also comment on how this model compares to the previous one.

In this question we are predicting average rating from all tags in the rmpCapstoneTags.csv file, so we set tags as predictors (X), and "Average Rating" as target (y). We filled the missing values with means, and checked collinearity of factors like how we did in **question 7**. Then we built a linear regression model in the same way, and this time we have **an R^2 value of 0.1523, and an RMSE value of 0.9127**, which is worse than our last model since the unexplained portion of variance and inaccuracy both increased. If we look at the fit line plot, we can also see a significant problem. The **slope this time is less than 1,** which means that this model tends to underestimate higher ratings and overestimate lower ratings. There is also **significant intercept,** indicating that the model predicts slightly higher ratings even when the predicted value should be close to zero. There are also a significant number of outliers with predicted ratings higher than 5 and lower than 0.

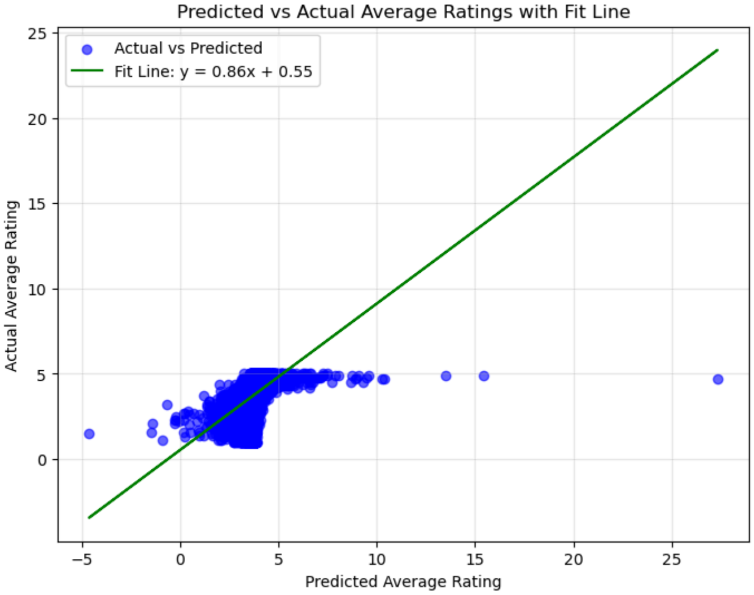
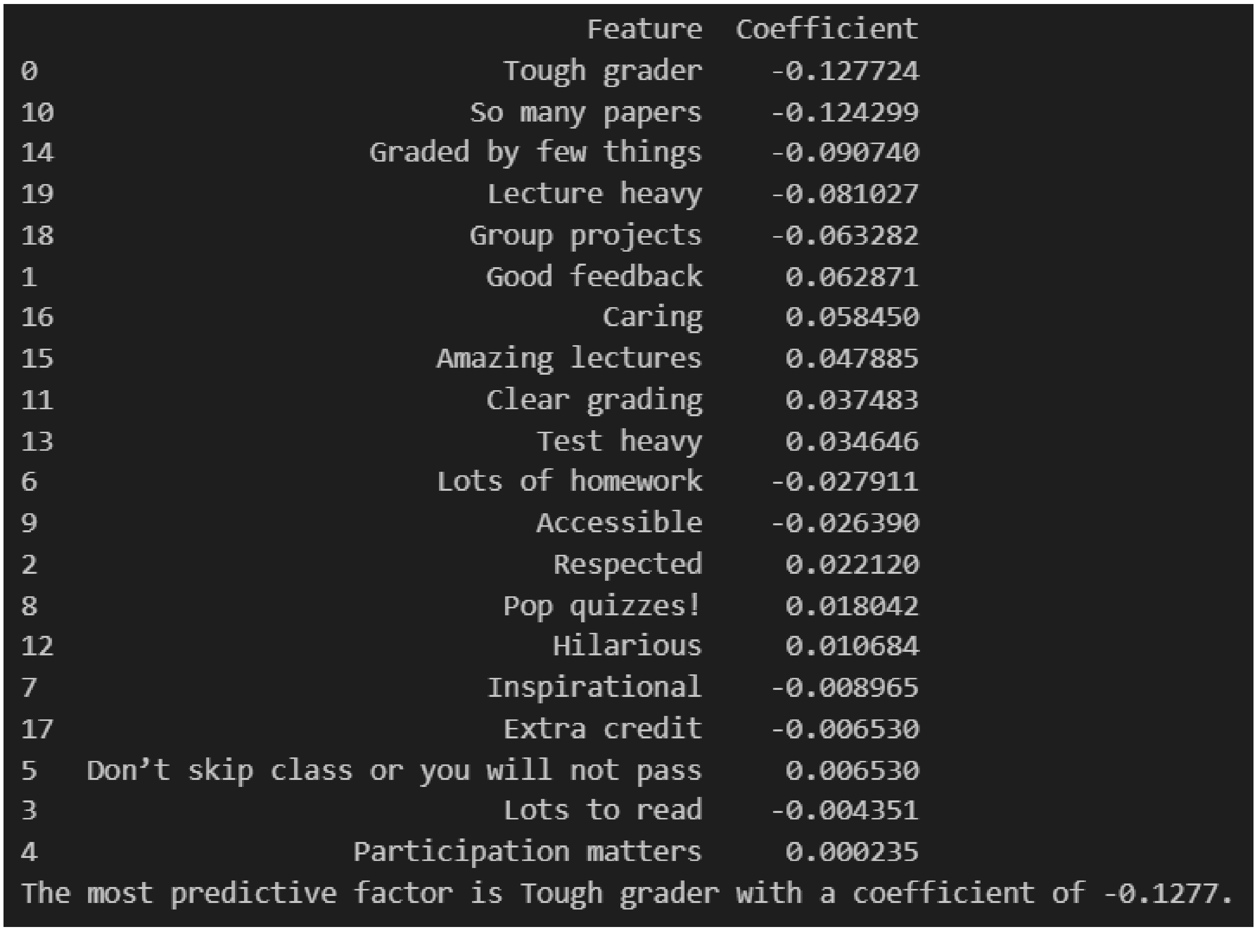


Figure 15. Predicted vs Actual Average Ratings with Fit Line

All these features show that this model is worse than our previous model, and this may be because we were using tags to predict. Compared to numerical values we used in the previous question, tags might not explain ratings well since they are noisier and less informative since they may contain more bias and subjective feelings. Also, there are more columns in the tags file, and high dimensionality may be another reason why this model performed worse. Still we checked the more predictive tag by ranking their coefficient absolute values. For this model **the most predictive predictor is Tough grader with coefficient -0.1277**, showing that professors' grading behaviors influence how this model predicts the ratings of them the most.



```
                                 Feature  Coefficient
0                           Tough grader    -0.127724
10                        So many papers    -0.124299
14                  Graded by few things    -0.090740
19                         Lecture heavy    -0.081027
18                        Group projects    -0.063282
1                          Good feedback     0.062871
16                                Caring     0.058450
15                      Amazing lectures     0.047885
11                         Clear grading     0.037483
13                           Test heavy      0.034646
6                       Lots of homework    -0.027911
9                             Accessible    -0.026390
2                              Respected     0.022120
8                           Pop quizzes!     0.018042
12                             Hilarious     0.010684
7                          Inspirational    -0.008965
17                          Extra credit    -0.006530
5   Don’t skip class or you will not pass     0.006530
3                           Lots to read    -0.004351
4                  Participation matters     0.000235
The most predictive factor is Tough grader with a coefficient of -0.1277.
```

Figure 16. Feature Importance for Average Rating Prediction

9. Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty? Hint: Make sure to address collinearity concerns.

In this question, we aimed to predict **average difficulty** using all tags in the rmpCapstoneTags.csv file. The tags served as predictors (X_3) and **average difficulty** as the target (y_3), we named them using _3 to separate them from previous questions. To handle missing values, we filled them with column means and checked for collinearity among predictors. After addressing collinearity concerns like previous questions, we trained a linear regression model using 80% of the dataset for training and 20% for testing.

The model's performance metrics were:

- R^2: 0.1451
- RMSE: 0.6584

While the R² value indicates the model explains a **small portion of the variance** in average difficulty, the RMSE shows **moderate prediction inaccuracy**. Similar to question 8, this model also struggled with dimensionality and the subjective nature of the tags, which are less direct indicators of difficulty compared to numerical predictors.

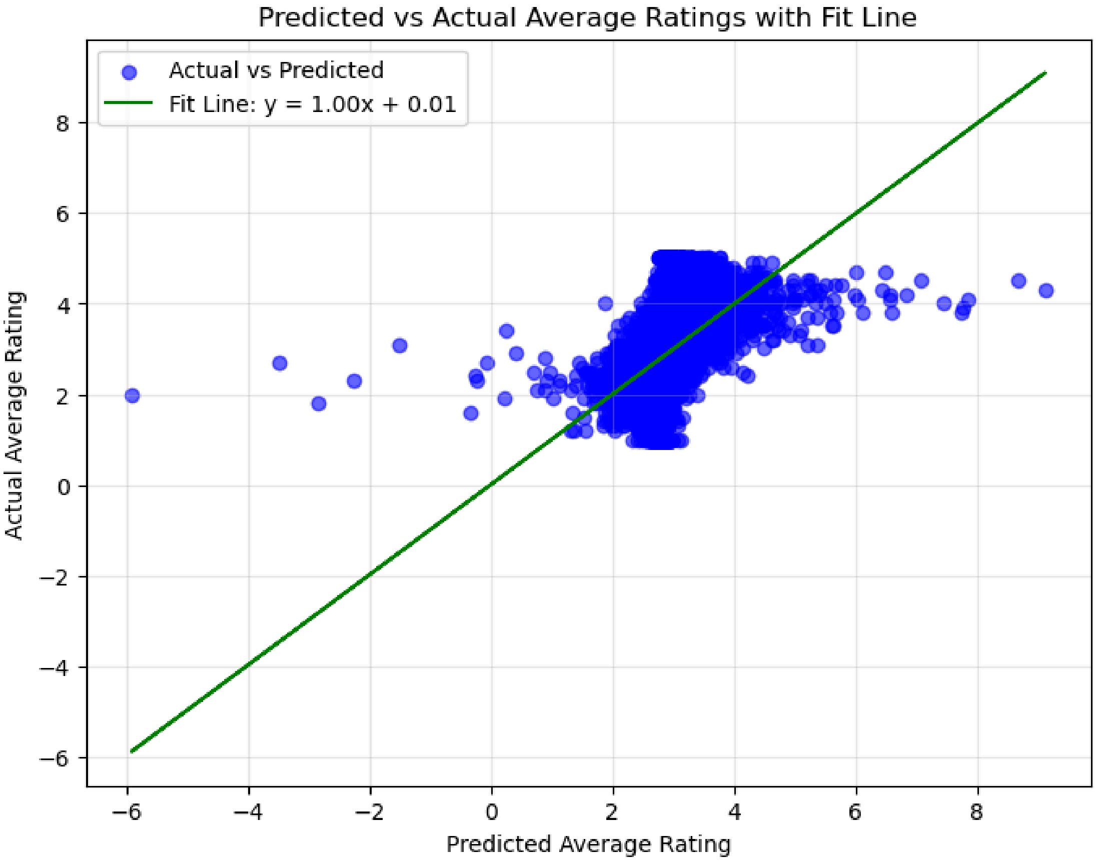


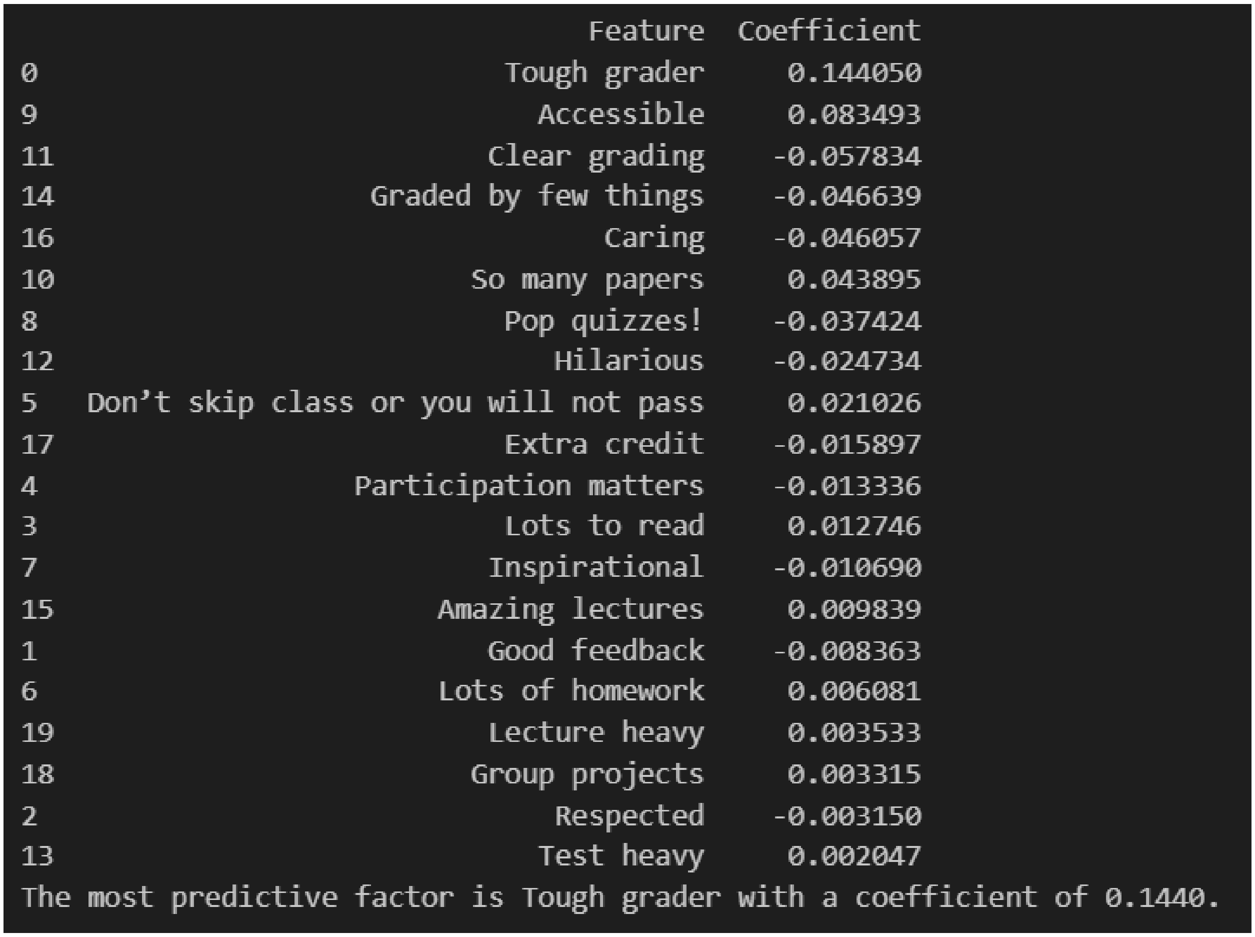
Figure 17. Predicted vs Actual Average Ratings with Fit Line

The fit line for **Question 9** shows a **slope close to 1 and the intercept is very small**, which indicates that the model has an ideal proportionality between predicted and actual values on average. However, despite this proportional slope, there are still notable issues: There are **significant outliers** in the predictions, with some values falling outside the realistic range (e.g., below 0 or above 5) which again showed the drawbacks of tags as predictors.

Similar to question 8, this model performs worse than the model in question 7 since we use tags to predict, but it performs slightly better than question 8 model, this may be due to the difference between difficulty and overall rating. The difficulty is determined by limited factors such as grades, homeworks, etc, while overall rating is more subjective and complicated since too many factors are involved.

With this in mind, we ranked the absolute values of the coefficients to identify the most predictive tags. The most significant predictor was **Tough Grader**, with a coefficient of **0.1440,** suggesting that grading behaviors strongly influence perceived difficulty.

Overall, this model performed slightly better than the average rating model (Question 8) but still suffered from noise in the tag data and high dimensionality, limiting its predictive power.



```
                                  Feature  Coefficient
0                            Tough grader     0.144050
9                              Accessible     0.083493
11                          Clear grading    -0.057834
14                    Graded by few things   -0.046639
16                                 Caring    -0.046057
10                        So many papers      0.043895
8                           Pop quizzes!     -0.037424
12                              Hilarious    -0.024734
5   Don’t skip class or you will not pass     0.021026
17                           Extra credit    -0.015897
4                   Participation matters    -0.013336
3                            Lots to read     0.012746
7                           Inspirational    -0.010690
15                        Amazing lectures    0.009839
1                           Good feedback    -0.008363
6                         Lots of homework    0.006081
19                          Lecture heavy     0.003533
18                         Group projects     0.003315
2                               Respected    -0.003150
13                             Test heavy     0.002047
The most predictive factor is Tough grader with a coefficient of 0.1440.
```

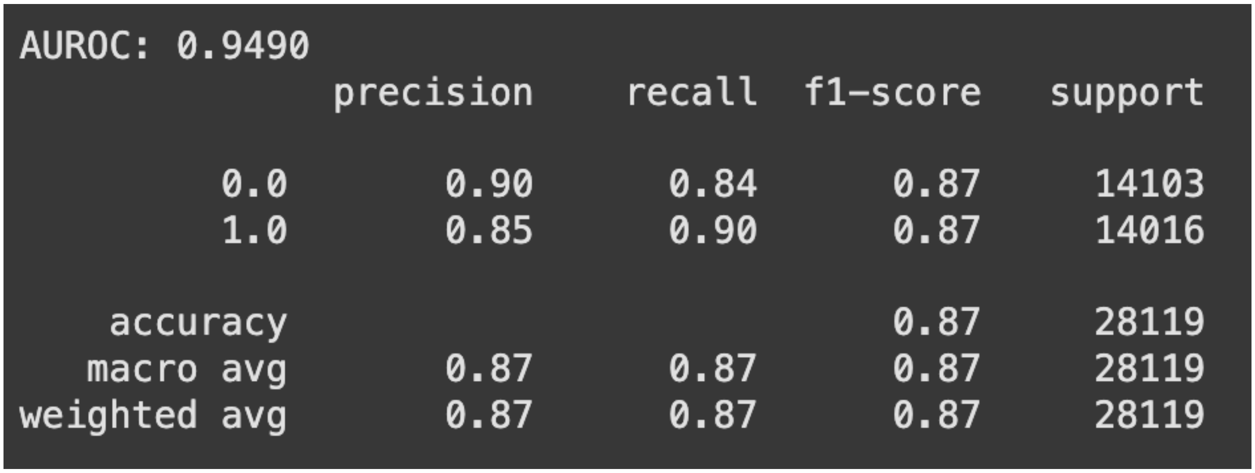
Figure 18: Feature Importance for Average Difficulty Prediction

10. Build a classification model that predicts whether a professor receives a “pepper” from all available factors (both tags and numerical). Make sure to include model quality metrics such as AU(RO)C and also address class imbalance concerns.

To predict whether a professor receives a "pepper" based on available factors (numerical and tags), we combined rmpCapstoneNum.csv and rmpCapstoneTags.csv. This included handling missing values by filling them with column means. Given the imbalanced nature of the target variable (pepper/no pepper), we applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset. Using this balanced dataset, we trained a Random Forest Classifier with 80% of the data for training and 20% for testing. A Random Forest Classifier was selected as the predictive model for its robustness and ability to handle high-dimensional data. The model achieved the following performance:

- **AUROC (Area Under ROC Curve)**: 0.9476
- **Accuracy**: 87%
- **Precision**: 90% for "No Pepper," 84% for "Pepper"
- **Recall**: 84% for "No Pepper," 90% for "Pepper"
- **F1-Score**: 0.87 for both classes

The **AUROC score of 0.9476** indicates excellent discriminative power, with the model effectively distinguishing between professors who receive a pepper and those who do not. Additionally, precision-recall tradeoffs suggest that the model is reliable in identifying both pepper and non-pepper cases.



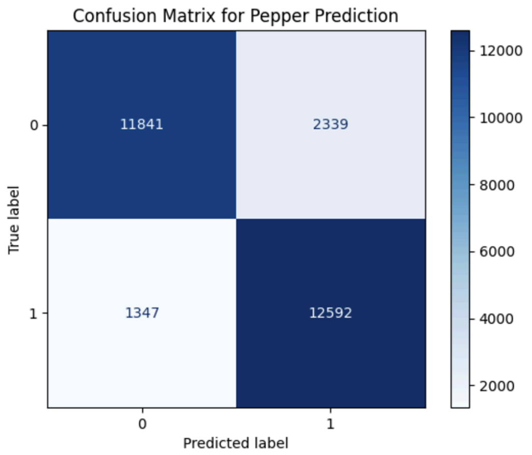
```
AUROC: 0.9490
              precision    recall  f1-score   support

         0.0       0.90      0.84      0.87     14103
         1.0       0.85      0.90      0.87     14016

    accuracy                           0.87     28119
   macro avg       0.87      0.87      0.87     28119
weighted avg       0.87      0.87      0.87     28119
```

Figure 19: Model Performance Metrics (AUROC: 0.9490, Accuracy: 87%, F1-Score: 0.87) and Classification Summary

We applied further visualization to examine this model's performance:



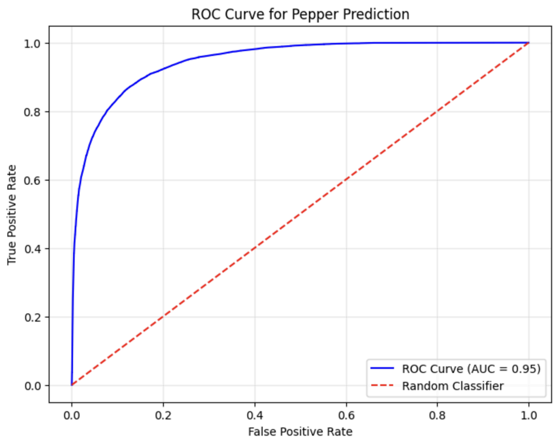


Figure 20: Confusion Matrix and ROC Curve for Pepper Prediction

Confusion Matrix: The confusion matrix shows that the model correctly predicted **11,841 cases of "No Pepper" (true negatives) and 12,592 cases of "Pepper" (true positives)**. However, **2,339 "No Pepper" cases were incorrectly classified as "Pepper" (false positives), while 1,347 "Pepper" cases were classified as "No Pepper" (false negatives).** These results demonstrate the model's strong predictive performance while highlighting some misclassifications.

ROC Curve: The ROC curve displays the trade-off between the true positive rate (sensitivity) and the false positive rate. The high AUROC value of **0.9476** confirms that the model performs significantly better than a random classifier.

This Random Forest model performed well in predicting pepper awards, with high accuracy and a robust ROC-AUC score. However, there is room for improvement by fine-tuning hyperparameters, exploring alternative models, and reducing noise in the tag data.

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the qualitative data, e.g. major, university or state by linking the qualitative data to the two other data files (tags and numerical)].

For the extra credit question, we investigated **whether Average Rating and Average Difficulty differed significantly across states**. Using the rmpCapstoneNum.csv dataset, we focused on U.S. state abbreviations as the grouping factor. Missing values for Average Rating and Average Difficulty were replaced with their respective column means to ensure a complete dataset.We conducted a **one-way ANOVA** for both Average Rating and Average Difficulty to determine if state-wise differences were statistically significant. The null hypothesis assumes no differences in ratings or difficulty among states.

For **Average Rating**, the ANOVA yielded a **F-statistic of 4.8427** and a **p-value < 0.0001**, indicating strong evidence of significant differences among states.

For **Average Difficulty**, the ANOVA yielded a **F-statistic of 6.3569** and a **p-value < 0.0001**, confirming that difficulty also varies significantly across states.

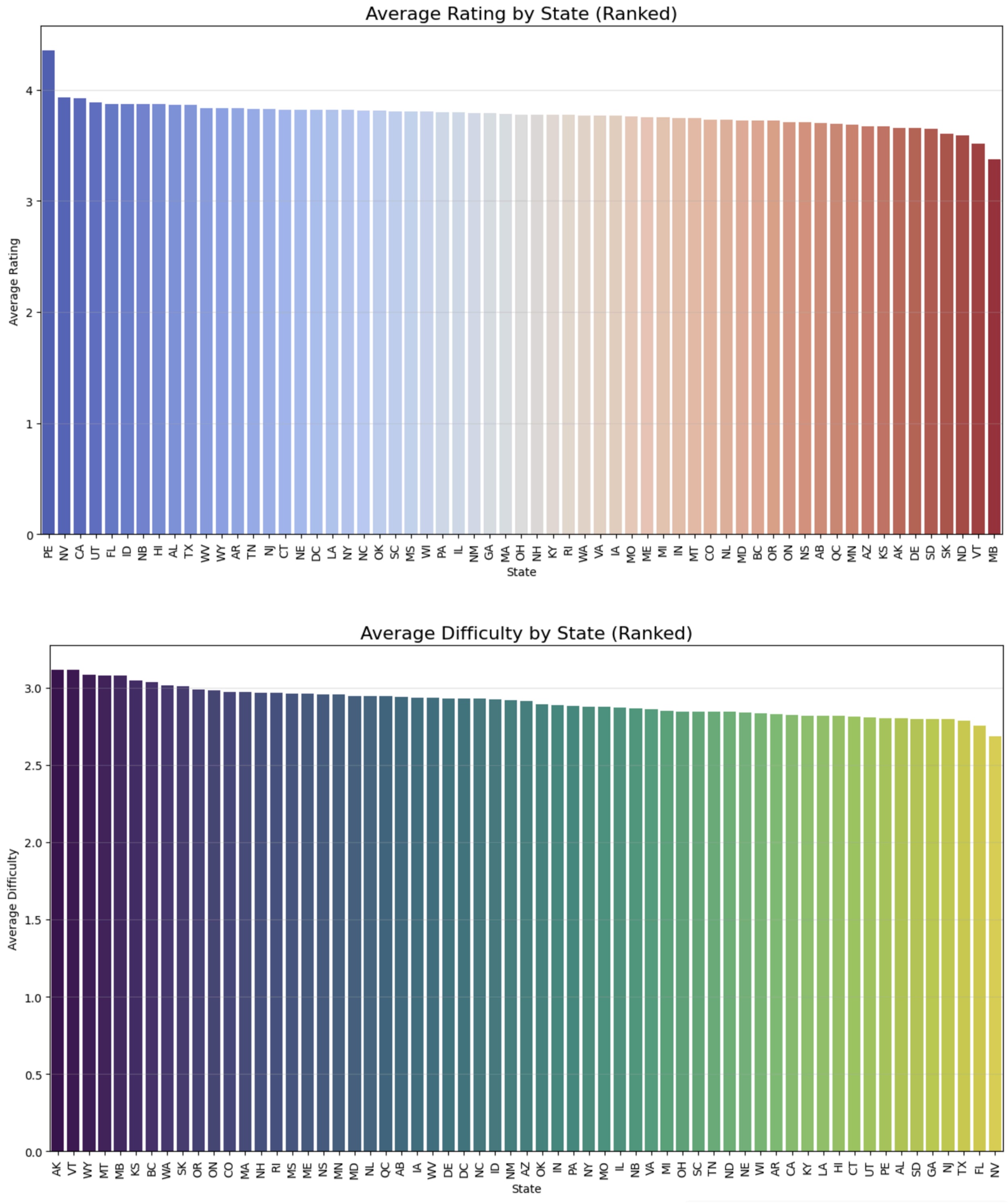


Figure 21: Average Rating and Difficulty Ranked by State

However, as we plotted the graph to visualize the difference, we quickly noticed some problems about our data. Firstly, there are wrong values for states (eg. London, BC) which could possibly affect the result. Secondly, despite that our ANOVA result shows a significant p-value, we cannot observe a difference on our graph about average rating and difficulty across states. Thus, we further looked into our data using alternative methods. A Tukey's HSD test was then performed to identify specific pairs of states with significant differences in ratings or difficulty. The Tukey's HSD test revealed that states such as **CA** and **AZ** had the most significant differences in average ratings and difficulties with other states. However, as we scroll down through our test result, we see that most of the states-wise comparisons hold the null hypothesis, thus, we conclude that the difference in Average Rating and Average Difficulty across states was only different in some states and it might possibly be caused by the inaccurate data. These findings suggest that regional variations in teaching styles or student expectations doesn't have a large influence on professor evaluations.