Course Name: Introduction to Data Science

Professor: Pascal Wallisch

Date: December 9, 2024

Capstone Project

Group Name: Capstone 49

Team Members: Saamia Shafqat, Alka Ashok, Kinjal Bagal

Data Preprocessing: For this project, we prepared the dataset by joining the three separate datasets into one comprehensive dataset. Rows with all zeros or empty values were removed, as these professors did not receive any reviews. Additionally, rows with an average rating of less than 3 were excluded to focus on meaningful data.

The tag columns were normalized using the number of ratings as follows:

normalized_tags = tag_columns.div(number_of_ratings, axis=0)

For the first three questions, the data was further manipulated to remove rows where both Male gender and Female gender were 0, as well as rows where both columns were 1. This ensured accurate categorization based on gender.

Normalization ensured that tag-related data contributed proportionally to the analysis.

Q1) 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.

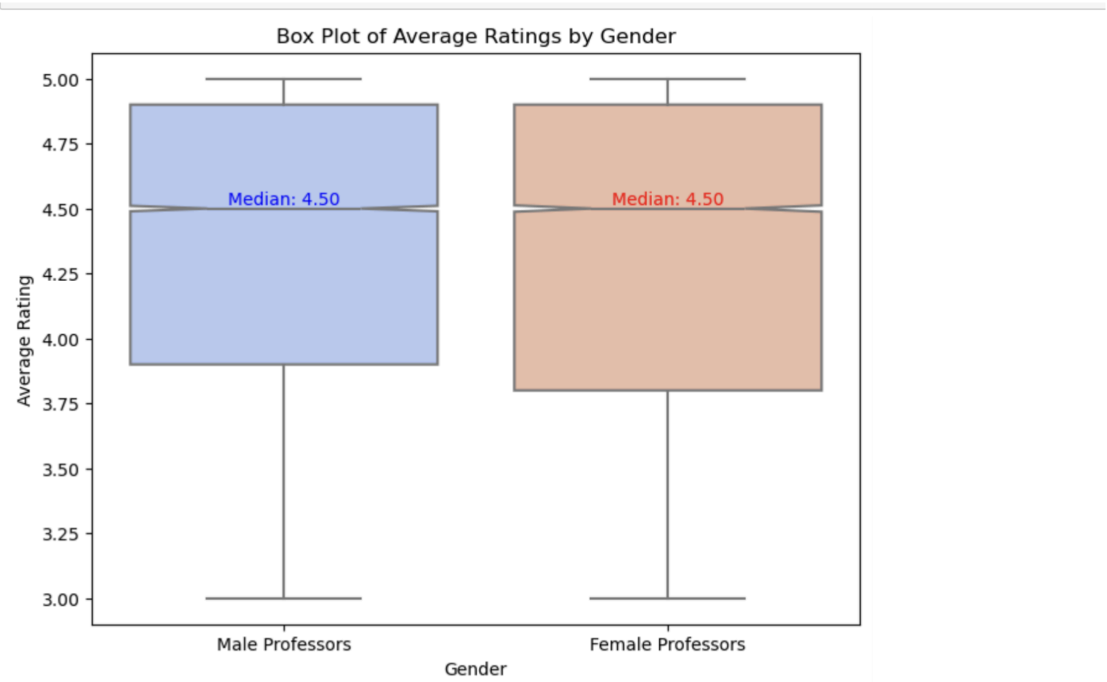
H_0 : There is no significant difference in ratings between male and female professors.
*H*_1: There is a significant difference in ratings between male and female professors.

D (What did you do):
We used Mann-Whitney U test to check if there is a significant difference in average ratings between male and female professors from the given datasets. Performed KS-test to check the normality of data of both male and female professor subsets. After proving that the data was not normally distributed, we used Mann-Whitney U test.

Y (Why did you do this):
Since the data is not normally distributed, we chose the Mann-Whitney U test because it can handle outliers and noise well compared to t-test.

F (What did you find):
Mann-Whitney U Statistic: 217366070.0
P-value: 0.1572648588302158

A (Answer and interpretation):
Since the p-value is greater than 0.005 we conclude that there is no significant difference in the ratings of male and female professors. The lack of significant difference suggests that gender bias, if present, is not strongly evident in this dataset. Based on the analysis, we do not find evidence to support the assertion of a pro-male gender bias in this dataset.



Q2) 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

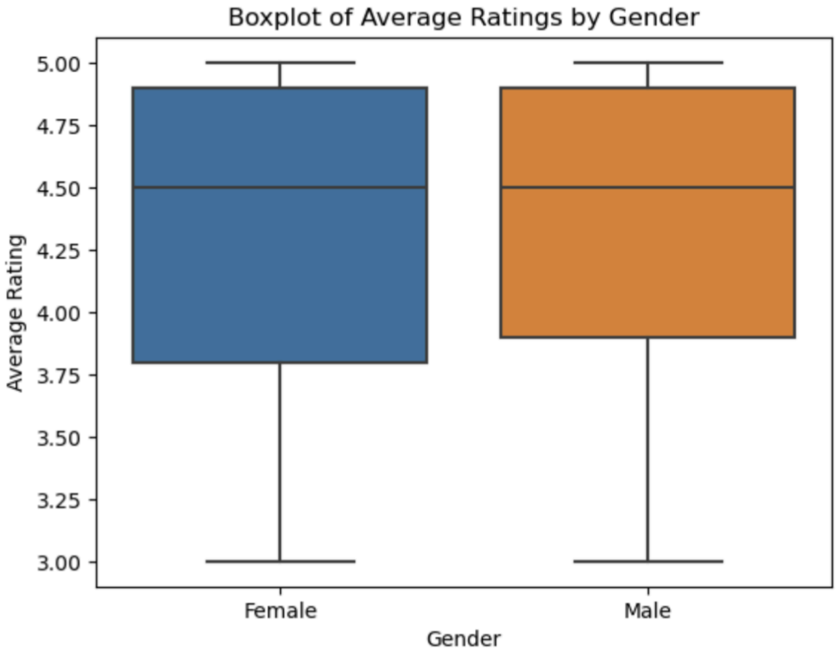
H_0: There is no difference in the variance of ratings between male and female professors.
H_1: There is a difference in the variance of ratings between male and female professors.

D (What did you do):
To check if there is a gender difference in the spread of gender distribution we calculate the variance and standard deviation in the ratings of male and female professors. We perform Levene test to compare the standard deviation and variation between the data.

Y (Why did you do this):
From the first question we know that the data is not normally distributed therefore we use Levene's test because it is specifically designed to test equality of variances.

F (What did you find):
Levene's test statistic: 7.313644838864062
P-value: 0.006846059275698202
No statistically significant difference between ratings distribution.

A (Answer and interpretation):
Since $p = 0.0068 > 0.005$, we failed to reject H_0 , and conclude there is no statistically significant difference in the variance between the two genders.



Q3) What is the likely size of both 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

For mean difference:
H_0 : There is no difference in the average ratings between male and female professors.
H_1 : There is a difference in the average ratings between male and female professors.

For variance ratio:
*H*_0: The variance of ratings is the same for male and female professors.
*H*_1 : The variance of ratings is different for male and female professors.

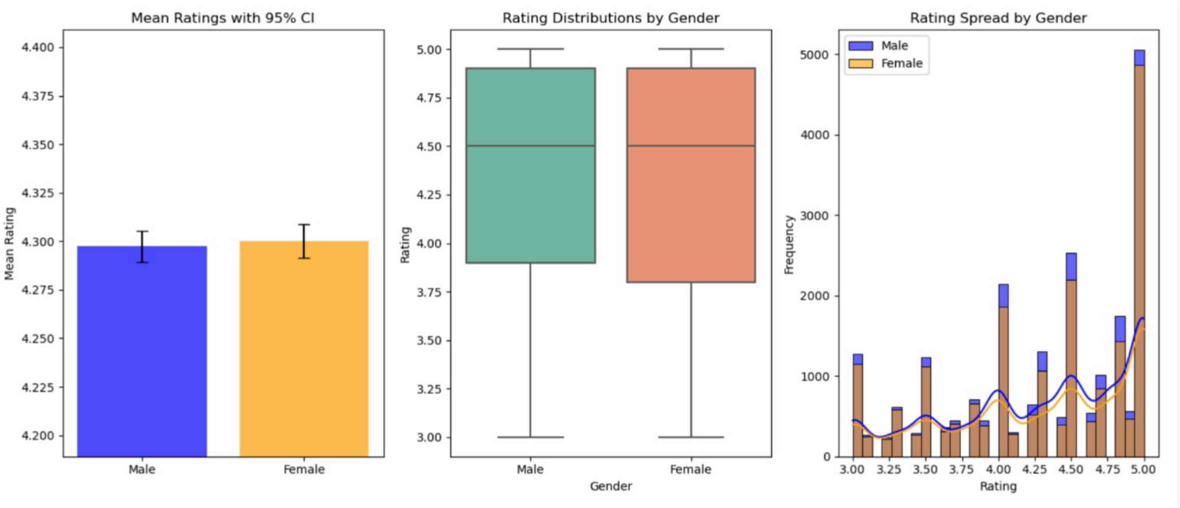
D (What did you do):
We estimated the magnitude of gender bias in average ratings and the spread of ratings. For the average rating bias, we calculated the mean difference, 95% confidence intervals, and Cohen's d to assess effect size. For the variance bias, we calculated the variance ratio and its confidence intervals. Finally, we visualized the results using bar plots, box plots, and histograms.

Y (Why did you do this):
Confidence intervals provide a range of plausible values for the observed differences. Cohen's d was used to quantify the effect size for the mean difference, and the variance ratio assessed the relative spread between genders.

F (What did you find):
Mean difference: -0.00
95% Confidence interval for the mean difference: (-0.01, 0.01)

Cohen's d (effect size): -0.00
Variance ratio (F-ratio): 0.97
95% Confidence interval for the variance ratio: (0.93, 1.01)

A (Answer and interpretation):
The mean difference of -0.00 and Cohen's d of -0.00 indicate no substantial difference in average ratings between genders. The 95% confidence interval for the mean difference spans zero, further supporting this conclusion. Similarly, the variance ratio close to 1.00 and its confidence interval overlapping 1 suggest no meaningful difference in the spread of ratings. These results imply no evidence of significant gender bias in this dataset.



Q4) 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.

D (What did you do):
We used the Mann-Whitney U test, as the tags were not normally distributed, and we had to compare between two independent groups.

Y (Why did you do this):
The Mann-Whitney U test was used because tag data is not normally distributed, and we needed to compare two independent groups (male and female professors).

F (What did you find):
We calculated the Mann-Whitney U statistic and p-value to test for significant differences in the number of tags awarded between male and female professors for each specific tag

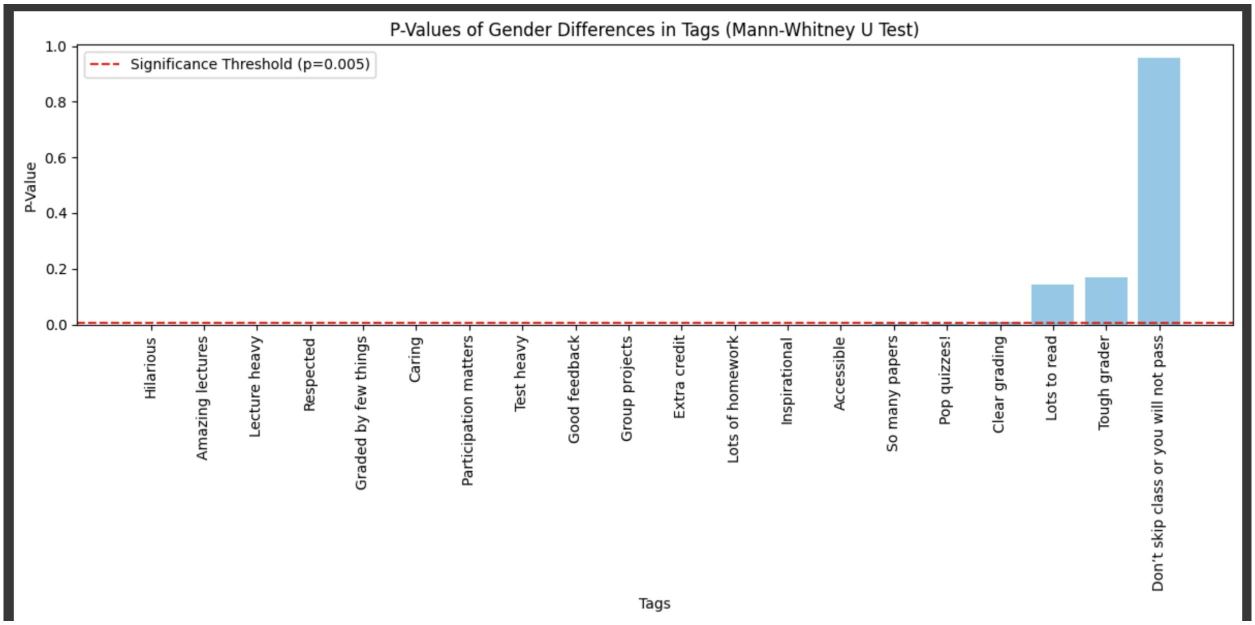
A(Answer and interpretation):
Most Gendered Tags

Tag	U-Statistic	P-Value
Hilarious	295028555.0	1.280958e-221
Amazing lectures	276727822.0	8.843857e-57
Lecture heavy	271877874.5	7.047546e-41

Least Gendered Tags

Tag	U-Statistic	P-Value
Lots to read	256906372.0	0.169375
Tough grader	259987959.5	0.182202
Don’t skip class or you will not pass	258576581.5	0.926604

The most gendered tags, such as Hilarious, Amazing lectures, and Lecture heavy, show statistically significant differences between male and female professors. These differences suggest that male professors are more frequently awarded these tags. Conversely, the least gendered tags, including Lots to read, Tough grader, and Don’t skip class or you will not pass, exhibit no significant differences, indicating similar perceptions for male and female professors for these tags.



Q5) Is there a gender difference in terms of average difficulty? Again, a significance test is indicated

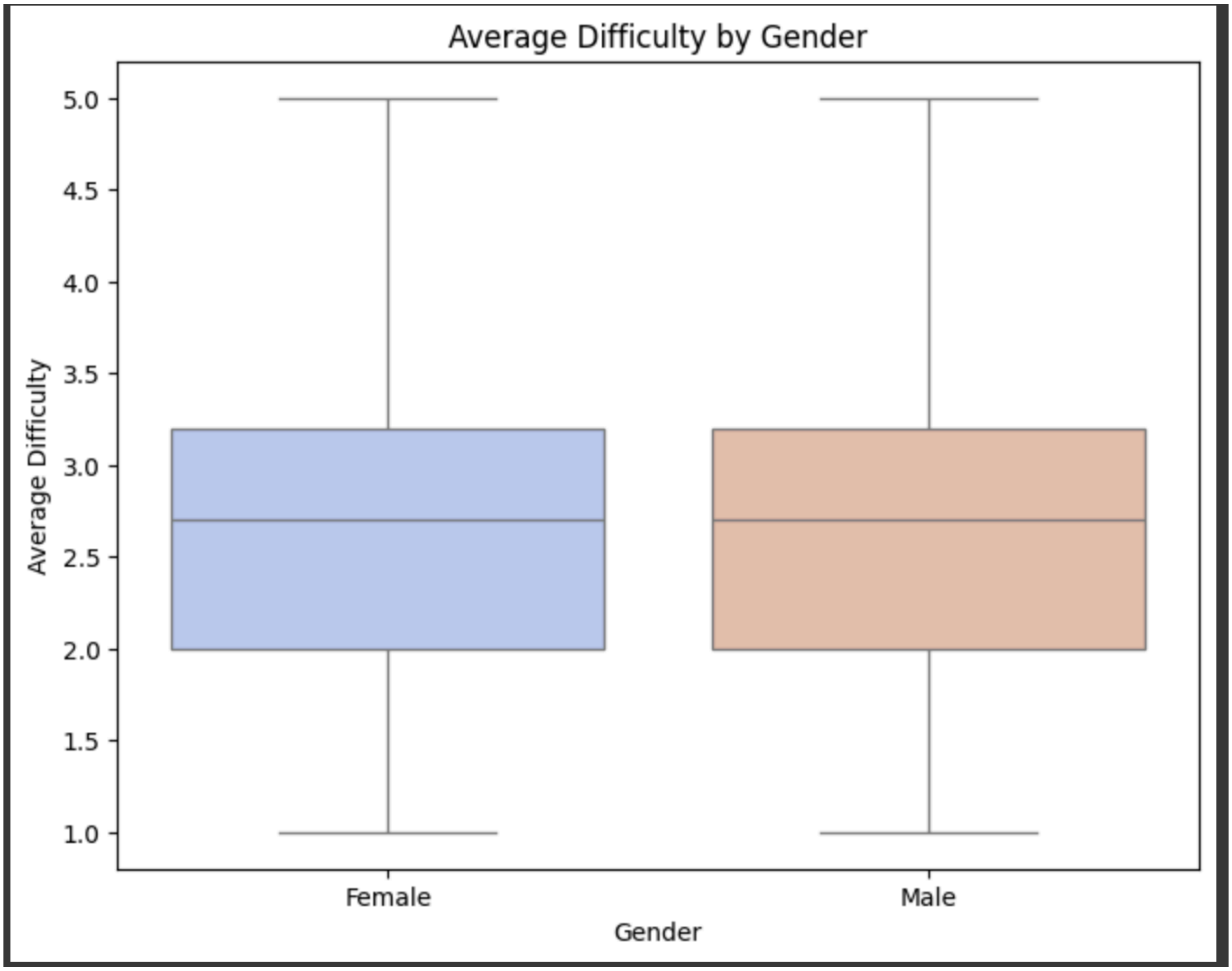
D (What did you do):
We checked if there is a significant gender difference in average difficulty ratings for professors. We performed a t-test and Levene’s test for unequal variances

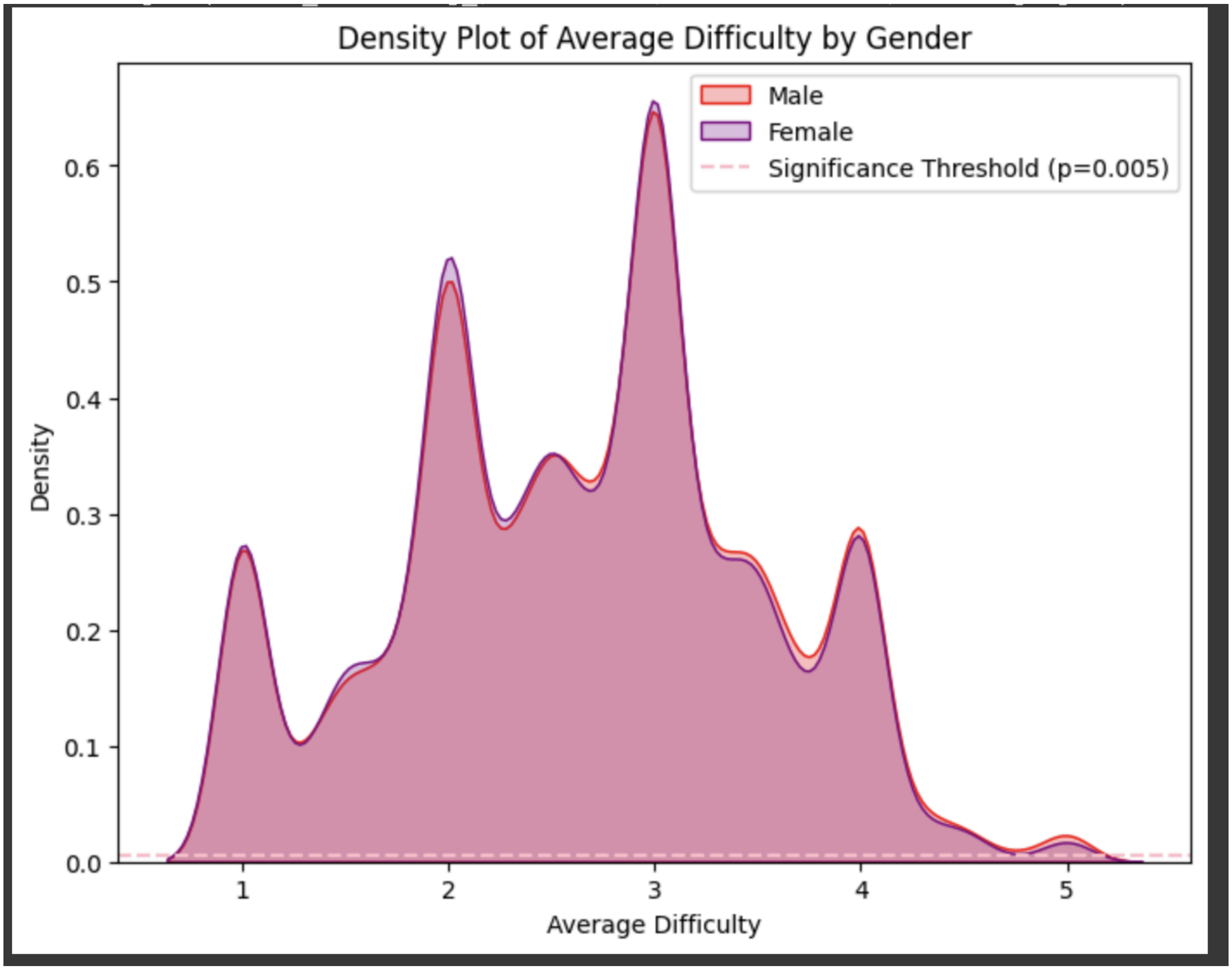
Y (Why did you do this):

As the data was not normally distributed, we performed these tests. The t-test helped us to get the mean difference between the two groups and handling the unequal variances we used the Levene's test.

F (What did you find):
Male Group Normality (p-value): 1.201910326
Female Group Normality (p-value): 5.943044738,
"Equal Variance Test (Levene's p-value)": 0.316311091170848,
'T-Statistic': 2.7846713702856087, '
P-Value': 0.005360433122499844}

A (Answer and interpretation):
We conclude that we found no significant difference in average ratings of female and male professors.





Q6) Please quantify the likely size of this effect at 95% confidence

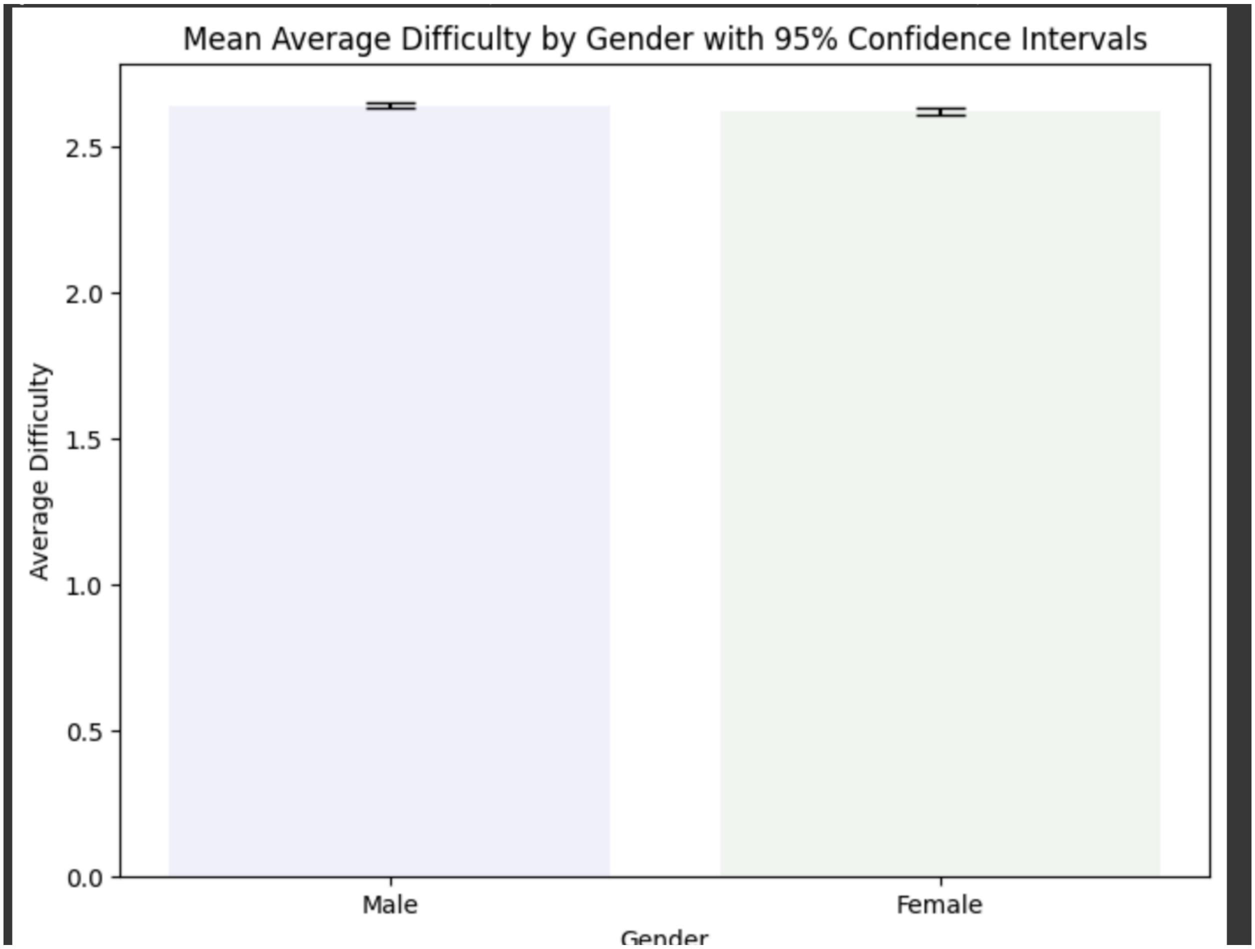
D (What did you do):
We checked if there is a significant gender difference in average difficulty ratings for professors to determine if there is a significant gender difference. We tested for normality, checking for variance equality, conducting Welch's t-test, and calculating the 95% confidence interval for the difference in means.

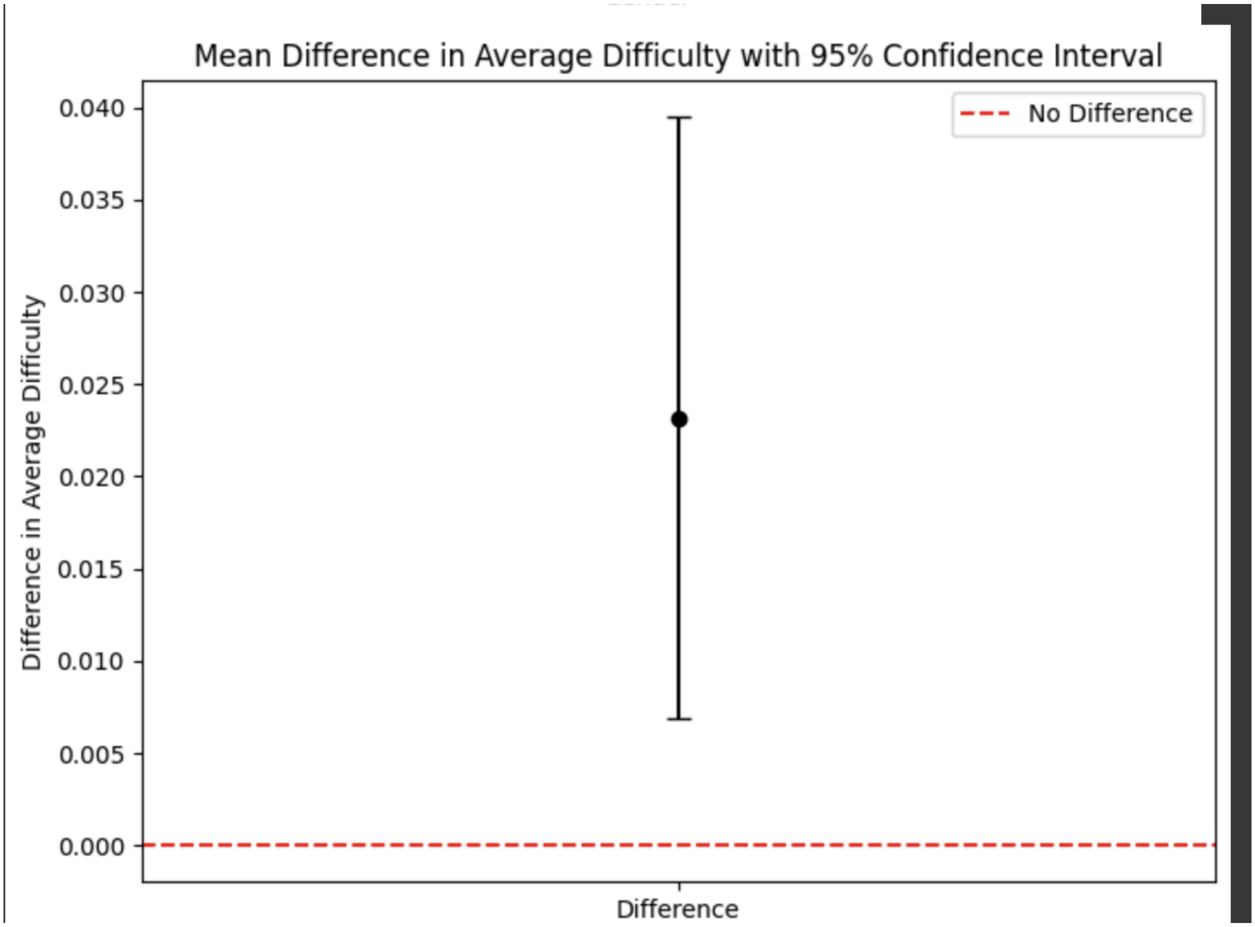
Y (Why did you do this):
We wanted to see if there exists a statistically significant difference between the mean difficulty ratings given by male and female professors. The confidence interval provides an estimate of the range within which the true mean difference lies at 95% confidence.

F (What did you find):
F (Findings):
Key Findings:
Mean Ratings:
Male Mean: 2.64
Female Mean: 2.62

Mean Difference: 0.023 (male rating higher than female rating)
Standard Error of Difference: 0.0083
95% Confidence Interval: (0.0069, 0.0395)
T-Statistic: 2.78
P-Value: 0.0054

A (Answer and interpretation):
When alpha was 0.005, we did not find any significant difference but now that alpha is 0.05 we do see some significant difference. The mean difference of 0.023 indicates a small positive difference, with the 95% confidence interval suggesting that the true mean difference is likely between 0.0069 and 0.0395.





Q7) Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these factors is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns

D (What did you do):
We built a linear regression model to predict `Average Rating` using all numerical predictors from the dataset. To address multicollinearity among predictors, we iteratively removed features with a Variance Inflation Factor (VIF) greater than 5. We then split the data into training and testing sets to evaluate the model. Metrics such as R^2 and RMSE were used to assess model performance, and the feature with the highest absolute coefficient was identified as the strongest predictor.

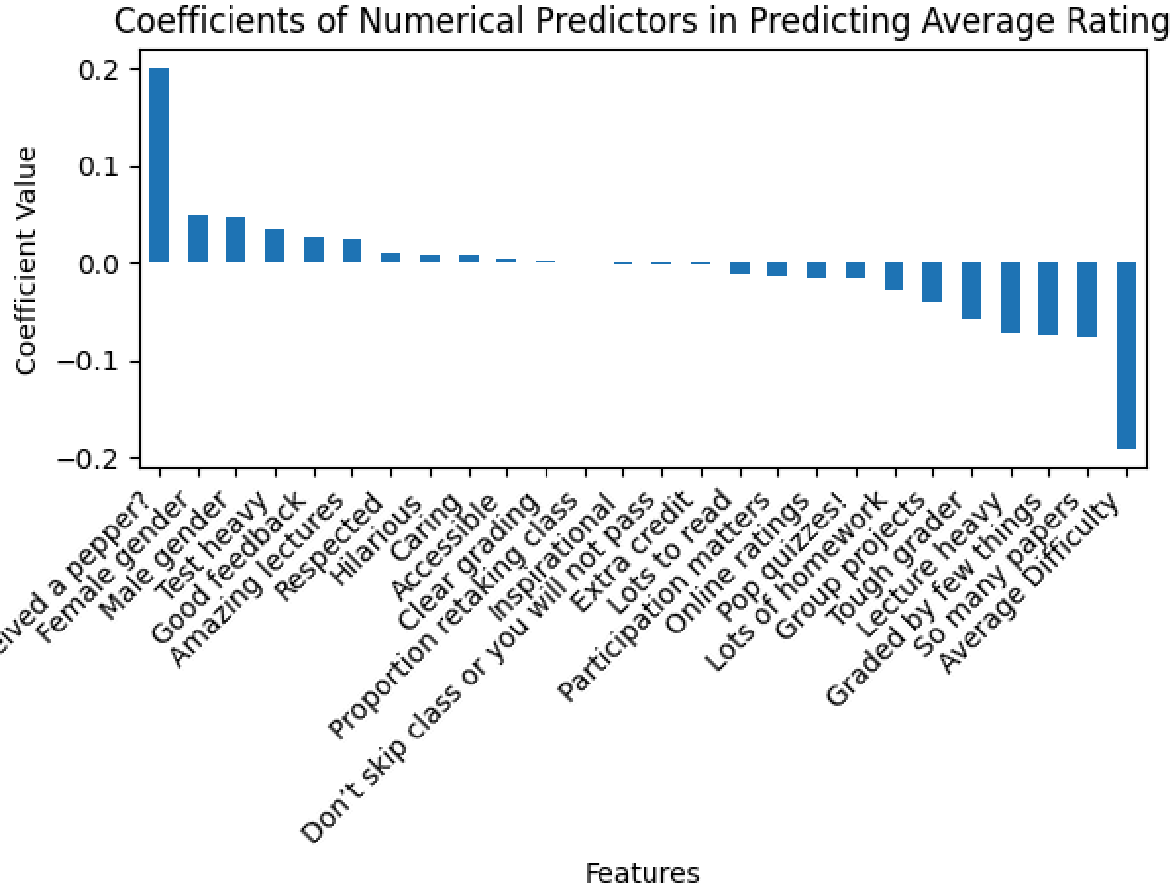
Y (Why did you do this):
Regression modeling helps quantify the relationship between numerical predictors and average ratings. Addressing multicollinearity ensures the model coefficients are reliable and interpretable. Splitting the dataset enables an unbiased evaluation of the model's performance.

F (What did you find):
R^2 : 0.2256 (The model explains ~22.56% of the variance in Average Rating.)

RMSE: 0.5586 (The model’s average prediction error is approximately 0.558 on the rating scale.)
Most Predictive Factor: Received a pepper? (Coefficient: 0.2002)

This indicates that professors rated as "hot" receive a positive boost in average ratings, making this factor the most significant predictor.

A (Answer and interpretation):
The regression model provides moderate predictive power with an R^2 of 0.2256, suggesting additional factors outside the model influence average ratings. The strongest predictor, Received a pepper?, positively impacts ratings, highlighting a potential bias based on perceived attractiveness. The RMSE indicates reasonable accuracy in predicting ratings. Overall, the model demonstrates how numerical factors contribute to average ratings, with some features having more substantial effects than others.



Q8) Build a regression model predicting average ratings from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 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.

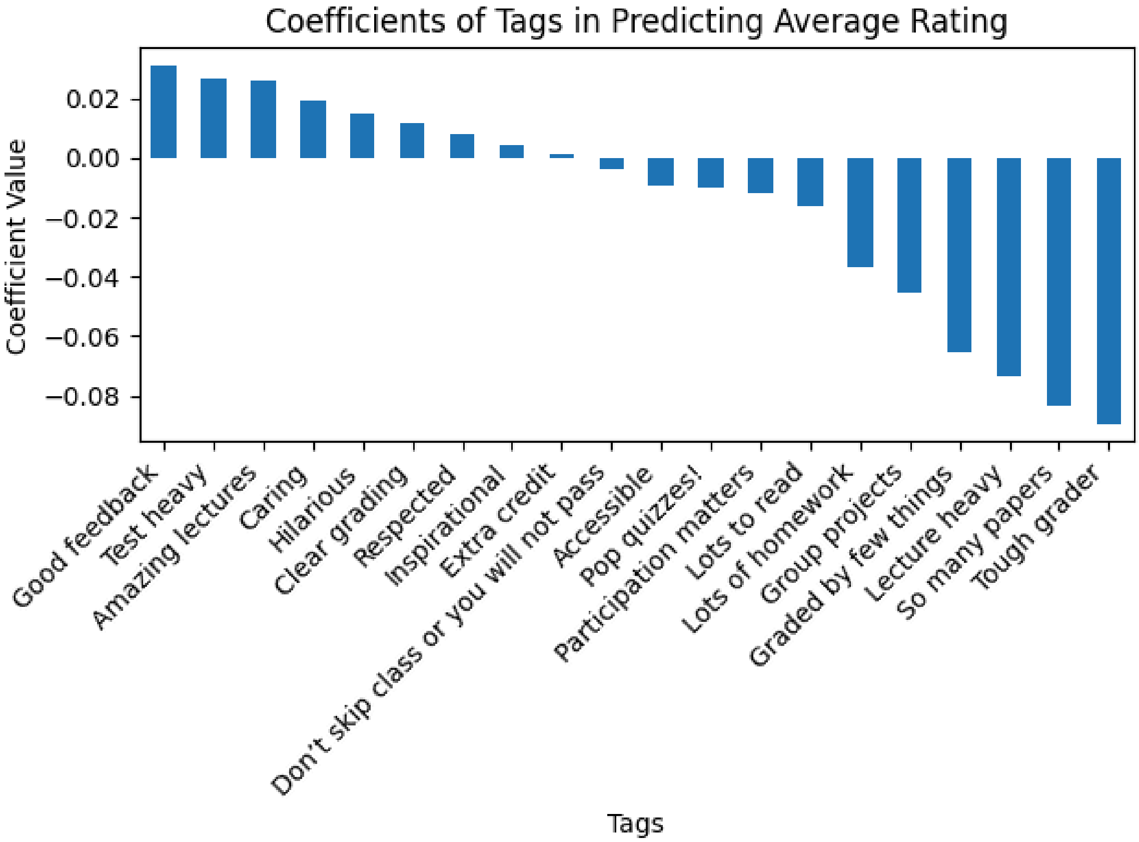
D (What did you do):

We built a regression model to predict Average Rating using only tag-related columns from the dataset. To address multicollinearity, we iteratively removed predictors with a Variance Inflation Factor (VIF) greater than 5. The dataset was split into training and testing sets to evaluate the model's performance. Finally, we identified the tag with the highest absolute coefficient as the most strongly predictive factor.

Y (Why did you do this):
Tags provide qualitative insights into student perceptions of professors. Using regression modeling, we aimed to quantify the impact of these perceptions on average ratings. Addressing multicollinearity ensured the stability and interpretability of regression coefficients.

F (What did you find):
R^2 : 0.1093 (The model explains ~10.93% of the variance in Average Rating.)
RMSE: 0.5991 (The model's average prediction error is approximately 0.599 on the rating scale.)
Most Predictive Tag: Tough grader (Coefficient: -0.0894)
This suggests that being perceived as a "Tough grader" negatively impacts average ratings.

A (Answer and interpretation):
The model explains a smaller proportion of the variance in Average Rating (*R*^2=0.1093) compared to the previous model using numerical predictors (*R*^2=0.2256). This indicates that while tags provide qualitative insights into student perceptions, they contribute less explanatory power compared to numerical factors such as Average Difficulty or Received a pepper?. Furthermore, the RMSE for this model (0.5991) is slightly higher than the previous model's RMSE (0.5586), indicating marginally lower predictive accuracy. The most predictive tag, Tough grader, negatively influences ratings, highlighting how stricter grading policies impact student evaluations.



Q9) Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty? Hint: Make sure to address collinearity concern

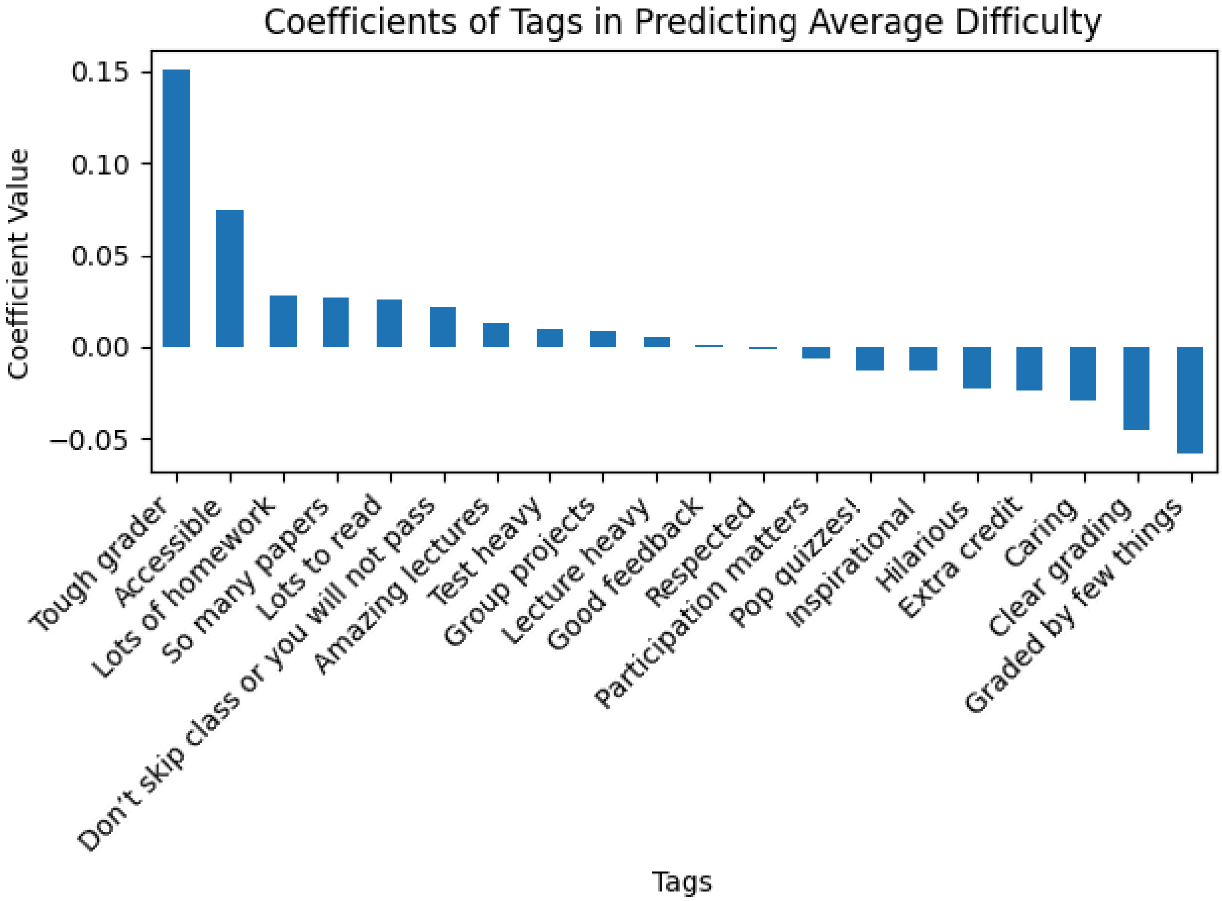
D (What did you do):
We built a regression model to predict Average Difficulty using only tag-related columns from the dataset. Multicollinearity was addressed by iteratively removing predictors with a Variance Inflation Factor (VIF) greater than 5. The dataset was split into training and testing sets to evaluate the model's performance. Finally, the tag with the highest absolute coefficient was identified as the most strongly predictive factor.

Y (Why did you do this):
Tags provide qualitative insights into student perceptions of professors. Using regression modeling, we aimed to quantify the impact of these perceptions on Average Difficulty. Addressing multicollinearity ensured the stability and interpretability of regression coefficients.

F (What did you find):
R^2 : 0.0698 (The model explains ~6.98% of the variance in Average Difficulty.)
RMSE: 0.8447 (The model's average prediction error is approximately 0.845 on the difficulty scale.)

Most Predictive Tag: Tough grader (Coefficient: 0.1508)
This suggests that being perceived as a "Tough grader" is strongly associated with increased perceived difficulty.

A (Answer and interpretation):
The model explains a smaller proportion of the variance in Average Difficulty compared to the previous models for Average Rating, reflecting the weaker influence of tags on difficulty perceptions. The RMSE indicates moderate predictive accuracy. The tag Tough grader significantly influences perceived difficulty, as expected from its direct association with grading policies.



Q10) 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 address class imbalance concerns.

D (What did you do):
We built a classification model using factors to predict whether a professor receives a "pepper". The model included all data preprocessing, feature engineering, and class imbalance handling with SMOTE. A Random Forest Classifier was trained, and the model's performance was evaluated with AUROC, a classification report, and confusion matrix.

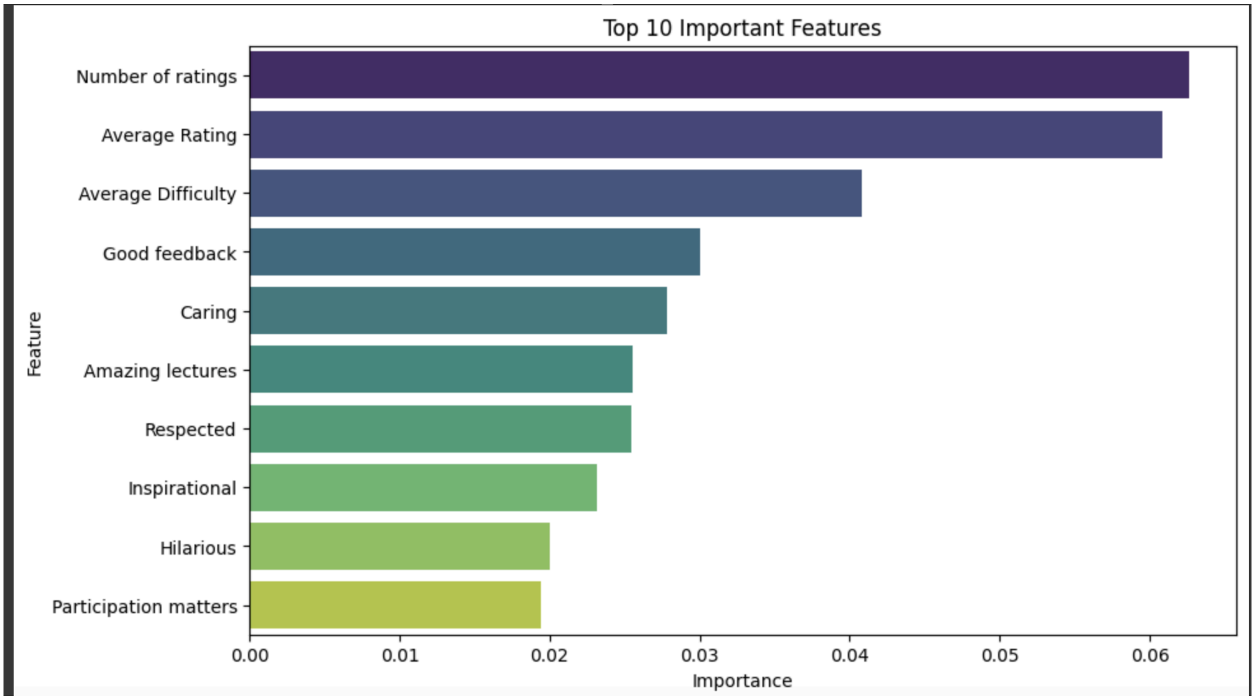
Y (Why did you do this):
To find the probability of a professor receiving a "pepper" based on the factors provided. Class imbalance was addressed to ensure that the model performs well across both classes, and model evaluation metrics such as AUROC and precision-recall were included to measure overall model quality and robustness.

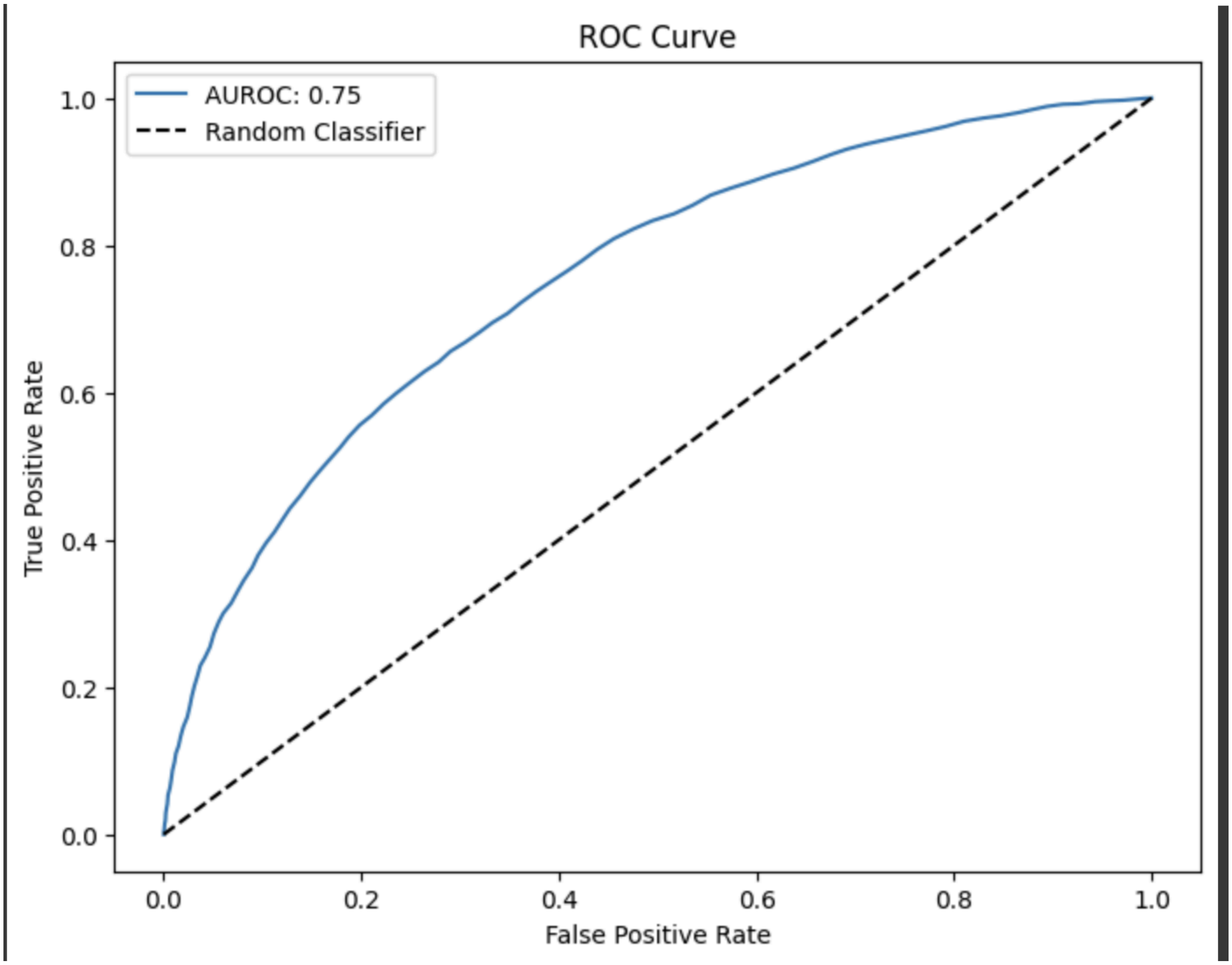
F (What did you find):
Model Performance:
 Model Quality Metrics
AUROC: 0.7546016458737707

```
Classification Report:
         precision    recall    f1-score   support

         0.0       0.77      0.82      0.79        7325
         1.0       0.60      0.53      0.56        3784

    accuracy                        0.72     11109
   macro avg       0.69      0.67      0.68     11109
weighted avg       0.71      0.72      0.72     11109
```

Class Imbalance Handling:
The use of SMOTE helped address the class imbalance in the training set, which improved model training by generating synthetic samples for the minority class (receiving a "pepper").

A (Answer and interpretation):
Conclusion: The model achieved a decent AUROC of 0.75, suggesting that it effectively distinguishes between the classes. Model might have missed some instances for ‘pepper’ class as the precision is high for class 0 than class.


Top 10 Important Features
Number of ratings
Average Rating
Average Difficulty
Good feedback
Caring
Amazing lectures
Respected
Inspirational
Hilarious
Participation matters
Feature
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Importance


ROC Curve
AUROC: 0.75
Random Classifier
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

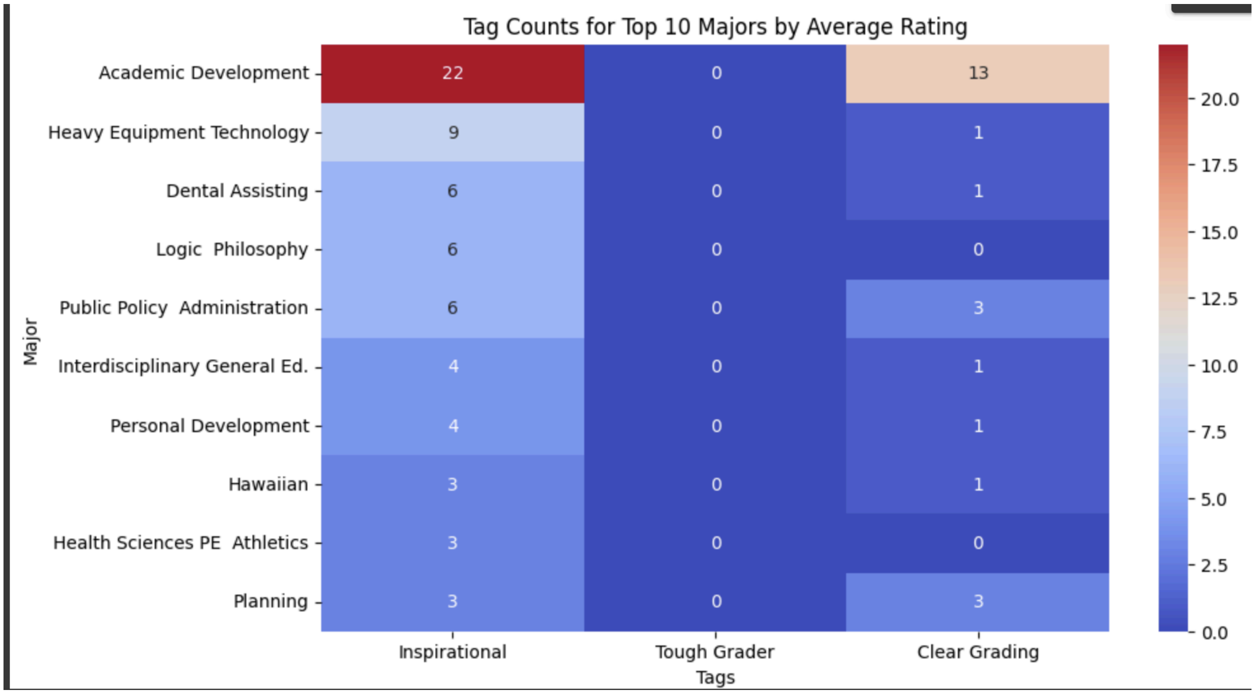
Q Extra Credit) Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the qualitative data, e.g. major, university or state by linking the qualitative data to the two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

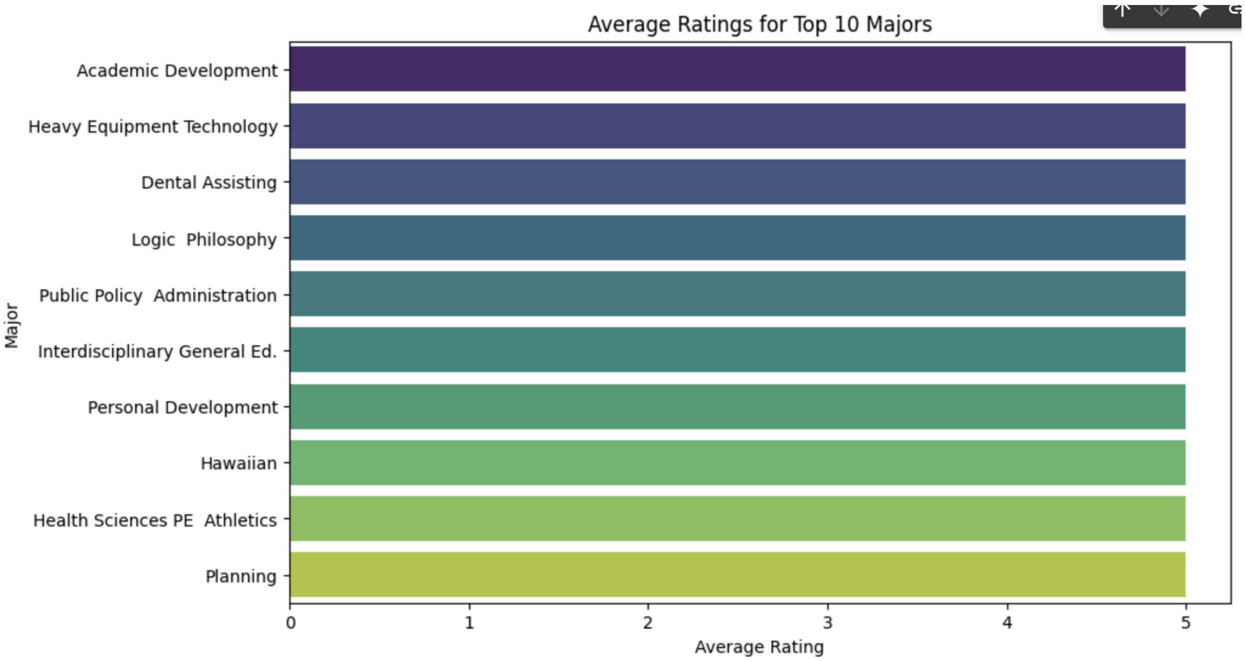
D (What did you do):
We conducted an analysis of the dataset to identify unique average ratings for different majors, calculated the overall average rating across all majors, and grouped data by major to test the average rating, difficulty, and the presence of tags (inspirational, tough grader, clear grading).

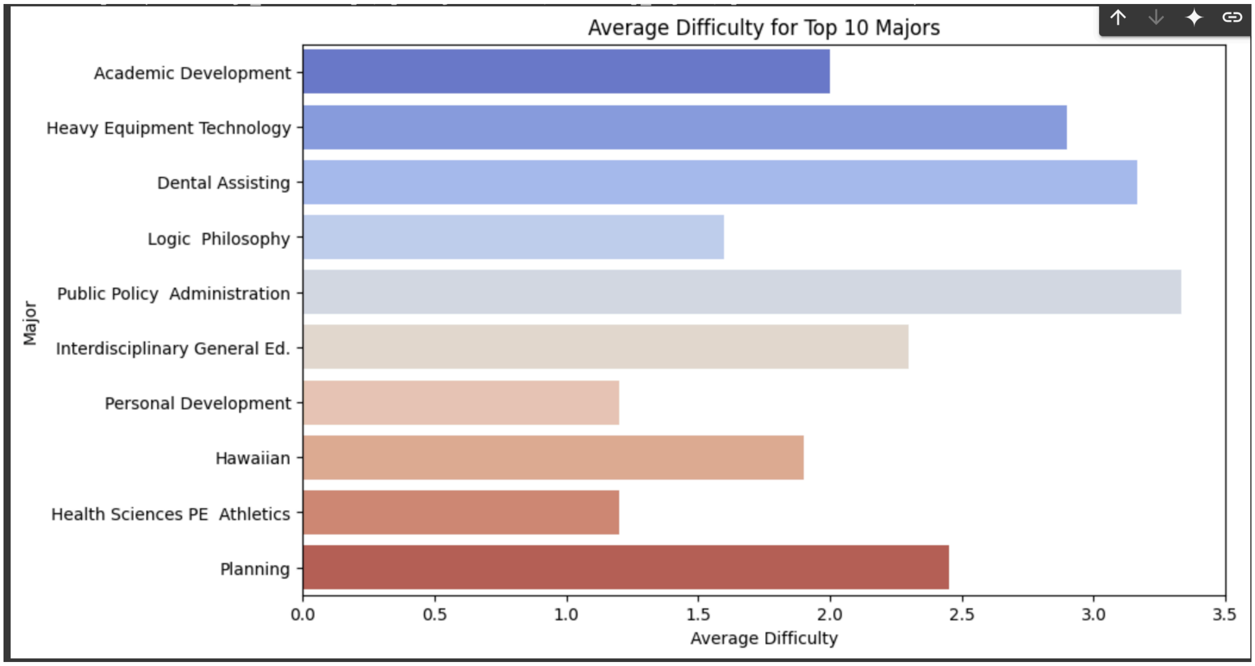
Y (Why did you do this):
To reveal potential patterns or trends within the dataset that link academic fields (majors) to their associated ratings and tag counts. We can gain deeper insights into which majors are highly rated and how they correlate with attributes like teaching clarity and instructor difficulty.

F (What did you find):
The overall average rating across all majors is 4.28, indicating that most courses are rated positively on average.
Top-rated majors (average rating of 5.0) include fields such as:
Academic Development and Heavy Equipment Technology, both showing high ratings alongside notable average difficulties (2.0 and 2.9, respectively).
Tag counts reveal that highly rated majors often have diverse levels of "inspirational" tags:
Academic Development had 22 "inspirational" tags, suggesting that courses in this field may be valued for their motivational content.
Other high-rated fields such as Personal Development had fewer "inspirational" tags but displayed good ratings (5.0) with relatively lower difficulty (1.2).

A (Answer and interpretation):
Conclusion: Some majors are rated exceptionally high despite varying degrees of difficulty. The presence of inspirational tags is a strong indicator of higher average ratings, suggesting that courses perceived as motivating or impactful may be rated more favorably. However, high difficulty does not always correlate with lower ratings, as some challenging fields still maintain high ratings, likely due to their perceived value or quality.


Tag Counts for Top 10 Majors by Average Rating
Major
Academic Development
Heavy Equipment Technology
Dental Assisting
Logic Philosophy
Public Policy Administration
Interdisciplinary General Ed.
Personal Development
Hawaiian
Health Sciences PE Athletics
Planning
22
0
13
9
0
1
6
0
1
6
0
0
6
0
3
4
0
1
4
0
1
3
0
1
3
0
0
3
0
3
Inspirational
Tough Grader
Clear Grading
Tags
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


Average Ratings for Top 10 Majors
Major
Academic Development
Heavy Equipment Technology
Dental Assisting
Logic Philosophy
Public Policy Administration
Interdisciplinary General Ed.
Personal Development
Hawaiian
Health Sciences PE Athletics
Planning
0
1
2
3
4
5
Average Rating


Average Difficulty for Top 10 Majors
Major
Academic Development
Heavy Equipment Technology
Dental Assisting
Logic Philosophy
Public Policy Administration
Interdisciplinary General Ed.
Personal Development
Hawaiian
Health Sciences PE Athletics
Planning
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Average Difficulty