Tomas Gutierrez
Pascal Wallisch
Principles of Data Science

Capstone Project Report

Introduction - Data Preprocessing

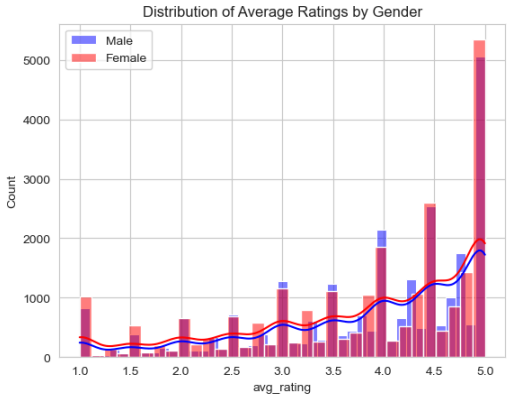
Although our coding sessions emphasized NumPy, I have decided to use Pandas for reading the data in and processing it, as it is more readable and I generally have more experience with it outside of this class. Firstly, Both CSV files did not have properly labelled columns or headers, so I loaded the data into pandas and added them. Then, for the sake of making my life easier, I merged the datasets horizontally, as the spec sheet mentions - “Each of these records (rows) corresponds to information about one professor”. Also, I seeded the RNG with the digits from my N-number (N11233917), with random.seed(). For the data cleaning, I removed all rows where all the columns that were meant to contain non-boolean values were NaNs. Removal by rows (axis=1) allows me to preserve the order of the records on my concatenated dataframe. The only drawback of this is that I could have removed rows that had valid data for the boolean columns, but this wouldn’t really matter unless I wanted to calculate the total number of male and female professors, and in that case, the simple solution would be to retrieve the raw data from those columns independently from the original CSV files If I am in a scenario where I would like more data. Also, I made sure to initialize alpha = 0.005 in preparation for significance testing as mentioned in the spec sheet.

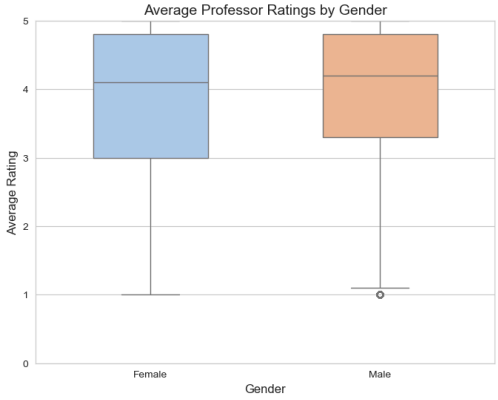
Question 1

To address whether a pro-male gender bias exists in professor evaluations, I first filtered out professors without an assigned gender and any missing an avg_rating. This resulted in a dataset with 52,089 professors (27,163 male, 24,926 female). I created a 'gender' column for clarity, and then checked for assumptions needed to run a parametric significance test. Both male and female rating distributions were moderately right-skewed (skewness roughly -0.9), but given the large sample size for both groups, CLT justifies still using parametric tests. I ended up using Levene’s test for homogeneity of variances, which returned a p-value < 0.005, indicating unequal variances. Therefore, I used Welch’s t-test to compare the mean ratings between genders. Welch's t-test showed that male professors had a slightly higher average rating (M = 3.878) than female professors (M = 3.811), with a t-statistic of 6.814 and a p-value again well below 0.005. To validate robustness against non-normality, I also ran a Mann-Whitney U test (Even though variances are unequal, Mann-Whitney is robust throughout the capstone since we have a large sample size), which confirmed the same directional result (p = 0.000004).

The effect size, calculated as Cohen’s d, was 0.059, suggesting that although the difference is statistically significant, it is pretty small in magnitude. A 95% confidence interval for the mean difference ([0.048, 0.086]) confirms that. The box plot below illustrating the distribution of

average ratings by gender shows this slight but consistent shift. Overall, the results across multiple tests are consistent with the presence of a very small but statistically robust pro-male bias in student evaluations. But to be clear, this is based on an observational study, and does not establish causality.

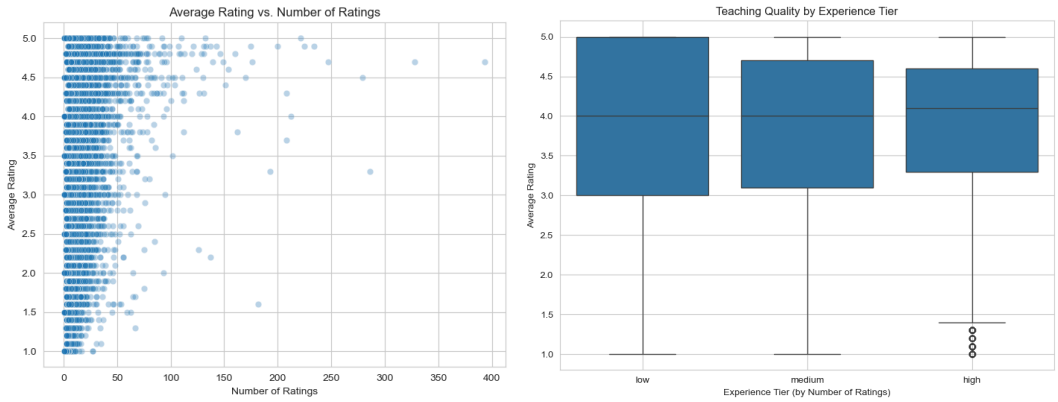




Question 2

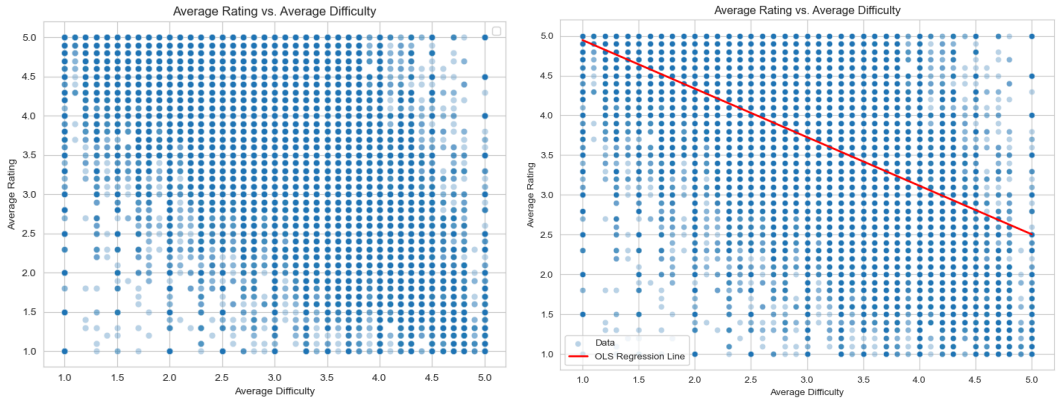
When exploring question 2, I used the number of ratings as a proxy for experience and average rating as a measure of quality (as mentioned in the question). I first filtered out rows where either variable was missing, leaving 70,004 valid records. I visualized the relationship using a scatter plot, which revealed a faint upward trend, though the data was noisy. I then computed both Pearson and Spearman correlation coefficients. Interestingly, Pearson's r was positive (0.0374, $p < 0.001$), while Spearman's p was negative (–0.0632, $p < 0.001$), suggesting some nonlinear or non-monotonic behavior. From the scatter plot, I suspected this was due to a disproportionate number of high ratings clustered around 4.7 and above. So, I filtered out professors with average ratings above 4.5, and the correlation results became more interpretable: Pearson r = 0.1154 and Spearman p = 0.1120, both statistically significant.

To formally test the effect, I first ran an OLS regression. Although the coefficient was significant ($p < 0.001$), the R^2 was just 0.001 and the RMSE remained high (roughly 1.13), indicating the model had little explanatory power. I then grouped professors into three experience tiers and ran a Kruskal–Wallis H test, which revealed significant differences across groups (H = 298.01, $p < 0.001$). Follow-up Mann-Whitney U tests showed that both low vs. medium and low vs. high groups differed significantly, while medium vs. high did not. These results suggest that although experience isn't a strong individual-level predictor, there are meaningful group-level differences.



Question 3

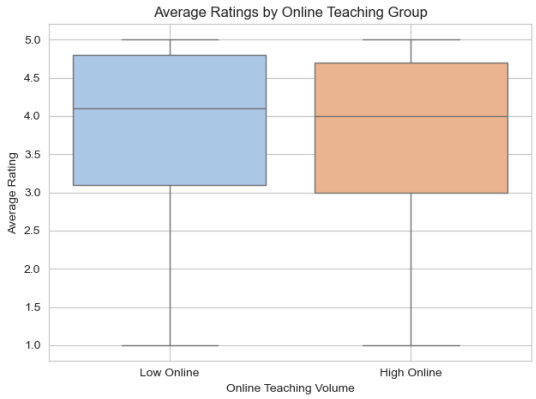
For question 3, I began by filtering out rows with missing values in either column, leaving 70,004 valid records. I visualized the relationship with a scatter plot, which showed a negative linear trend - professors who were rated as more difficult tended to receive lower overall ratings. To quantify this, I computed both Pearson and Spearman correlation coefficients. Pearson's r was –0.5368 and Spearman's p was –0.5114, both with p-values well below the 0.005 significance threshold. This strongly confirms a statistically significant negative association between average difficulty and average rating. Then, I ran an OLS linear regression using statsmodels, with avg_difficulty as the independent variable and avg_rating as the dependent variable. The resulting coefficient was –0.6103, indicating that for each 1-point increase in difficulty, the average rating dropped by approximately 0.61 points. The p-value for this coefficient was < 0.005, and the 95% confidence interval ranged from –0.617 to –0.603, meaning that the effect is both statistically and practically significant. The model's R^2 was 0.288, suggesting that difficulty alone explains about 29% of the variance in ratings - much stronger than what we saw in Question 2. The RMSE was approximately 0.95, which is reasonable given the 1–5 scale. Overall, there is a strong and statistically significant negative relationship between average difficulty and average rating: professors perceived as harder tend to be rated lower.



Question 4

To explore whether professors who teach more online classes receive higher or lower ratings, I

first dropped rows with missing values for avg_rating, num_online_ratings, or num_ratings, and got a dataset of 68,139 records. I then split the data into two groups using the median number of online ratings as the threshold: professors above the median were labeled “High Online,” and those below, “Low Online”, within a new column in the working dataframe. A box plot comparing the distribution of average ratings between these two groups showed visually that professors who teach more online classes tend to receive slightly lower evaluations. To see if this was statistically meaningful, I first checked the assumption of equal variances using Levene’s test, which returned a p-value < 0.001, indicating unequal variance. So, I used Welch’s t-test, which confirmed the difference was statistically significant ($t = –13.34$, $p < 0.001$). I also ran a non-parametric Mann-Whitney U test as a robustness check, which yielded ($U = 295{,}344{,}689.0$, $p < 0.001$). Cohen's d effect size was –0.141, and the 95% confidence interval for the mean difference in ratings between groups was [–0.1848, –0.1374]. In conclusion, professors who teach a high volume of online classes receive statistically significant lower average ratings than those who don’t.

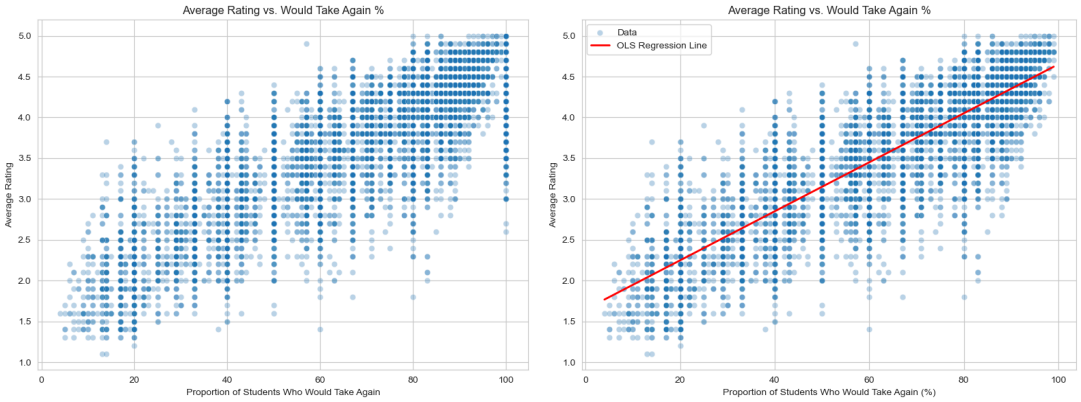


Question 5

For question 5, I followed a very similar process as the one in question 3. First, I dropped rows where either variable was missing, which left 12,160 valid records. Initial visualization revealed a clear positive trend but also a heavy concentration of professors with exactly 100% “would take again” proportion, suggesting a response bias/ceiling effect that might obscure the relationship we are looking for. To account for this, I filtered out entries where would_take_again_pct == 100.0, reducing the dataset to 8,111 observations and improving interpretability.

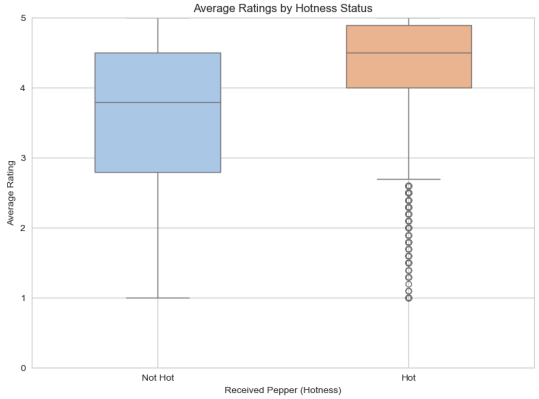
Correlation analysis revealed a very strong positive relationship: Pearson’s r was 0.8452 ($p < 0.001$), and Spearman’s p was 0.8261 ($p < 0.005$), both indicating a highly significant and monotonic association. To formally quantify the strength of this effect, I ran an OLS linear regression with would_take_again_pct as the predictor and avg_rating as the response variable. The slope coefficient was 0.0300 ($p < 0.005$) with a 95% confidence interval of [0.030, 0.030], suggesting that for each additional percentage point increase in students willing to retake the class, the professor’s average rating increased by 0.03 points. The model had a high R^2 of 0.714 and an RMSE of 0.4339, indicating a strong fit and low error on the rating scale. Overall,

there is a strong and statistically significant positive relationship between the proportion of students who would take a professor again and their average rating: professors with higher would-take-again percentages tend to receive higher ratings.



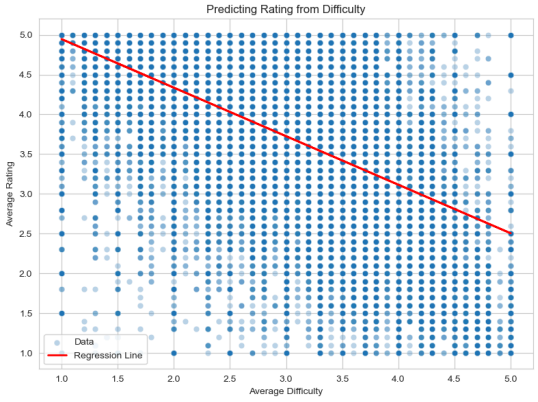
Question 6

To answer whether professors deemed “hot” receive higher ratings, I used the received_pepper variable as a proxy for perceived hotness. After removing rows with missing data, I was left with 70,004 valid entries. Since there was class imbalance—roughly 19,600 “hot” and 50,400 “not hot” professors—I opted for statistical tests robust to unequal group sizes and variance. Levene’s test for equality of variances returned a p-value < 0.001, justifying the use of Welch’s t-test over a standard t-test. Welch’s t-test confirmed a highly significant difference in average ratings between the two groups ($t = 113.11$, $p < 0.001$), with “hot” professors rated, on average, 0.798 points higher than their peers. This result was corroborated by a Mann-Whitney U test ($U = 684{,}243{,}055.5$, $p < 0.001$), which does not assume normality. The 95% confidence interval for the mean difference was [0.784, 0.812], reinforcing the precision of the estimate. To measure practical significance, I computed Cohen’s $d = 0.831$, a large effect size, suggesting the difference is not only statistically significant but also meaningful in practical terms. A boxplot illustrated the shift in rating distributions between the two groups, supporting the numerical results visually. Overall, this analysis provides strong evidence of a substantial and statistically significant difference in student evaluations between professors who were perceived as “hot” and those who were not, answering the question affirmatively.



Question 7

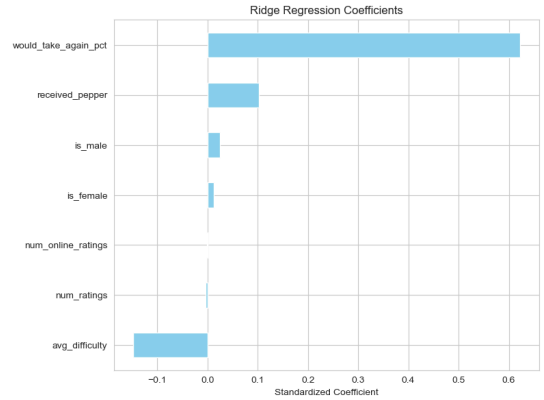
In question 7, I used scikit-learn's LinearRegression to predict a professor's average rating based solely on their average difficulty. After dropping rows with missing values for either variable, there were 70,004 valid records. The fitted linear regression model had an intercept of 5.5564 and a slope coefficient of –0.6103. This means that for every one-point increase in perceived difficulty, a professor's average rating is predicted to decrease by roughly 0.61 points. The model's R^2 value was 0.2881, indicating that about 29% of the variance in professor ratings can be explained by difficulty alone. The RMSE was 0.9508, suggesting that predictions tend to deviate from actual ratings by nearly one point on the rating scale. The resulting regression line, plotted over the scatter of data, confirms a strong negative linear relationship between perceived difficulty and teaching rating. This model answers the question by demonstrating that more difficult professors tend to receive lower ratings, with difficulty being a moderately strong standalone predictor. (Upon reading Q7, I realized I had done this for Q3)



Question 8

For question 8, I built a multiple linear regression model using all available numeric predictors of average rating ('avg_difficulty', 'num_ratings', 'received_pepper', 'would_take_again_pct', 'num_online_ratings', 'is_male', 'is_female'). I used Ridge regression to handle potential multicollinearity, since the question specifies to address a collinearity concern. All predictors were standardized prior to modeling to ensure that they were treated equally by the regression penalty. I used RidgeCV to tune the regularization parameter α, testing five values between 0.01 and 100. The best alpha selected was 1.0. The resulting model achieved an R^2 of 0.8101 and an RMSE of 0.3687, showing a significant improvement over the simpler "difficulty-only" model from Question 7, which had an R^2 of only 0.2881 and an RMSE of 0.9508. The most influential variable was the "would take again" percentage (β = 0.622), followed by hotness (β = 0.102) and difficulty (β = –0.147). Interestingly, gender and online teaching variables had relatively minor effects. The results indicate that while difficulty still negatively affects average rating, students' willingness to retake a professor's class is by far the strongest predictor of higher scores. Overall, this model captures a much larger share of the variance in student evaluations than any single predictor. By addressing collinearity through Ridge regularization and interpreting

standardized coefficients, this approach provides a robust answer to the question of how all available factors together predict a professor's average rating.



Question 9

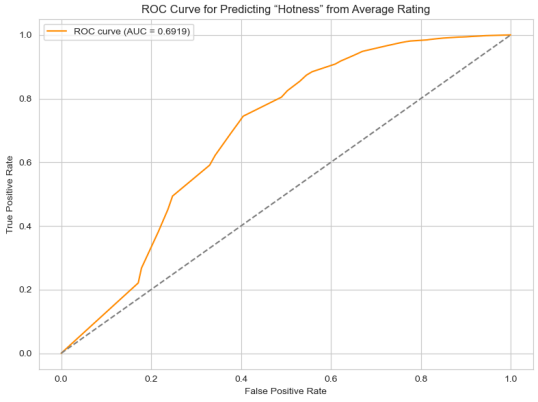
To predict whether a professor receives a "pepper" using only their average rating, I trained a logistic regression model. Due to class imbalance (roughly 72% of professors in the dataset did not receive a pepper), I applied random undersampling to both the training and test sets, balancing the classes. Evaluated on a balanced test set, the model achieved an AUC score of 0.691. At the default threshold of 0.5, it reached 67% accuracy, with 65% precision, 74% recall, and an F1 score of 69% on the "peppered" class. I explored threshold tuning to maximize precision, but the precision-optimized threshold resulted in lower recall and overall performance. While average rating is a useful signal, it alone is not sufficient to reliably predict perceived hotness, suggesting that more complex or latent factors drive this outcome.

```
AUC Score: 0.6910
Confusion Matrix:
[[3425 2325]
 [1468 4282]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.60      0.64      5750
         1.0       0.65      0.74      0.69      5750

    accuracy                           0.67     11500
   macro avg       0.67      0.67      0.67     11500
weighted avg       0.67      0.67      0.67     11500
```



Question 10

To assess whether professors receive a "pepper" based on multiple available factors, I trained a logistic regression model using all available numerical predictors('avg_rating', 'avg_difficulty', 'num_ratings', 'would_take_again_pct', 'num_online_ratings', 'is_male', 'is_female'). After removing missing values, I split the data into training and testing sets (70/30), then addressed class imbalance in the training set by downsampling the majority class to match the size of the

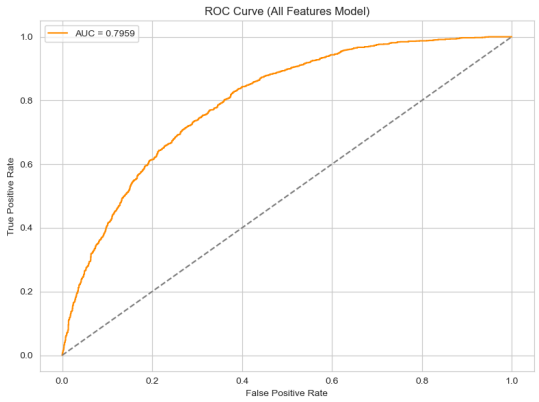
minority class. The model was fitted using scikit-learn's LogisticRegression. On the test set, the model achieved an AUC score of 0.7959, significantly outperforming the model from Question 9, which used only average rating (AUC roughly 0.69). This confirms that incorporating multiple features improves predictive performance. At a default threshold of 0.5, the model achieved an accuracy of 71%, with balanced precision (0.80 for not-peppered, 0.65 for peppered) and recall (0.63 and 0.81, respectively). The resulting ROC curve shows strong separability, supporting the model's discriminative ability. In summary, adding features beyond average rating substantially improved the ability to classify whether a professor received a pepper.

```
Logistic Regression with All Features
AUC Score: 0.7959
Confusion Matrix:
[[1251  734]
 [ 322 1341]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.63      0.70      1985
         1.0       0.65      0.81      0.72      1663

    accuracy                           0.71      3648
   macro avg       0.72      0.72      0.71      3648
weighted avg       0.73      0.71      0.71      3648
```



Extra Credit

For extra credit, I explored whether professors in STEM receive higher or lower average ratings compared to their counterparts in the humanities. After classifying majors into two broad categories - STEM (e.g., Biology, Computer Science, Engineering) and Humanities (e.g., English, History, Fine Arts) - I compared average ratings between the two groups using Welch's t-test and the Mann-Whitney U test (Again, valid due to the large sample size, which makes the test robust despite unequal variances). Both tests revealed a statistically significant difference ($p < 0.005$), with STEM professors receiving lower average ratings than humanities professors. The mean difference in ratings was approximately -0.31, with a 95% confidence interval of [-0.3413, -0.2839], and a moderate effect size (Cohen's $d = -0.282$). This suggests that discipline plays a meaningful role in student evaluations, and there is strong statistical evidence that on average, STEM professors receive significantly lower student ratings than their counterparts in the Humanities.