Emily Zhang

JF Koehler

Data Bootcamp

16 December 2024

Assessing Professor Effectiveness

I. Problem

Higher education plays a pivotal role in shaping the future, and students often turn to Rate My Professors (RMP) as a resource to gauge their likelihood of success. For my final project, I aim to develop both a predictive model capable of accurately estimating the average ratings of professors from the RMP dataset and determining whether average ratings can serve as a reliable predictor. These ratings are a critical metric, reflecting students' perceptions of teaching effectiveness, approachability, and course quality.

However, RMP data is often biased due to self-selection of individuals who choose to provide feedback, raising questions about its validity as a resource and its accuracy in reflecting professor performance. To address this, I plan to analyze the factors that contribute to higher or lower ratings and assess their significance. If these factors are found to be significant, the findings could uncover patterns in student feedback and rating trends. This model may prove valuable for universities and administrators in identifying areas for improvement in teaching practices and enhancing overall student satisfaction. Ultimately, can RMP be considered a reliable resource for evaluating professor performance, or do its inherent biases limit its utility?

II. Dataset and Data Cleaning

I reached out to my Principles of Data Science professor, Pascal Wallisch, for a dataset from Rate My Professors. He assisted with the necessary data munging and scaffolding,

including scraping the website, converting the HTML into structured data, collating individual ratings, and ensuring anonymization. It is important to note that this data was collected prior to 2018, back when Rate My Professors still included the "chili pepper rating," a controversial metric that has since been removed due to its questionable relevance in evaluating teaching effectiveness. The dataset contains 89,893 records, with each row corresponding to information about one professor. The dataset includes the following 8 columns:

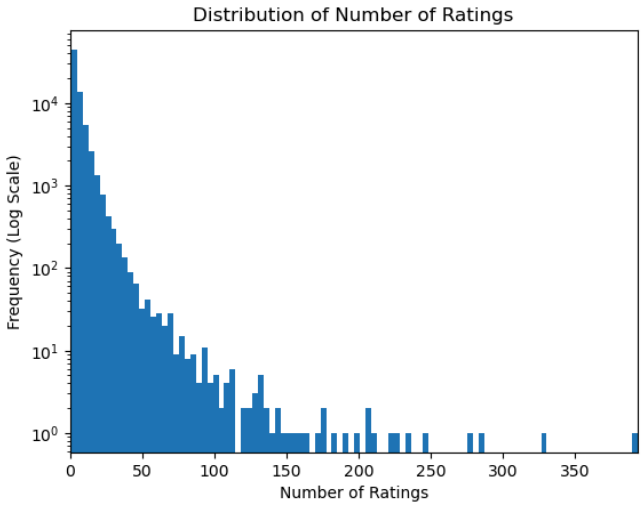
1. **Average Rating:** The arithmetic mean of all individual quality ratings for the professor.
2. **Average Difficulty:** The arithmetic mean of all individual difficulty ratings for the professor.
3. **Number of Ratings:** The total number of ratings used to calculate the averages.
4. **Received a "Pepper"?:** A Boolean indicating whether the professor was judged as "hot" by students.
5. **Proportion of Students Who Would Take the Class Again:** The percentage of students who indicated they would retake the class.
6. **Number of Ratings from Online Classes:** The count of ratings originating from online courses.
7. **Male Gender:** A Boolean (1: determined with high confidence that the professor is male).
8. **Female Gender:** A Boolean (1: determined with high confidence that the professor is female).

Ratings from Rate My Professors can introduce bias, especially when based on a small number of reviews. For example, usually the most extreme respondents, those who either strongly enjoyed or disliked the professor's class, are more likely to submit ratings. To address this, I first filtered the data to include only rows where the number of ratings exceeds 10. Next, I removed rows containing NaN values. This threshold was chosen because, after cleaning the data, the mean and median for num_rating were 20.687 and 16.0, respectively. These values suggest that the remaining professors likely taught for at least a year, covering multiple semesters or quarters. This method minimizes confounders, including extreme ratings from the same class, ensuring a more representative sample.

One caveat to this cleaning approach is the significant reduction in dataset size. After applying the filters, the dataset decreased from the original 89,893 records in rmpNum.csv to 7,209 rows. Further, some entries indicated ambiguous gender values, such as 0 for both male and female or 1 for both. I removed these rows as well, reducing the dataset further to 5,231 entries. While this dramatic decrease in dataset size may appear concerning, excluding ambiguous or incomplete data ensures quality and reduces bias, keeps the analysis meaningful and robust, and prevents misleading conclusions. The distribution of ratings is shown below on a logarithmic scale to reduce skewness from extreme values.

```
# Only keeps num_ratings > 10 and drops all nan values in the dataframe
professor_wt = professor_wt[professor_wt["num_rating"] > 10].dropna()
professor_wt.shape
```

```
# Removes rows of where is_male and is_female are both 0 and are both 1
professor_wt = professor_wt[~((professor_wt["is_male"] == 0) & (professor_wt["is_female"] == 0))]
professor_wt = professor_wt[~((professor_wt["is_male"] == 1) & (professor_wt["is_female"] == 1))]
```



III. Exploratory Data Analysis

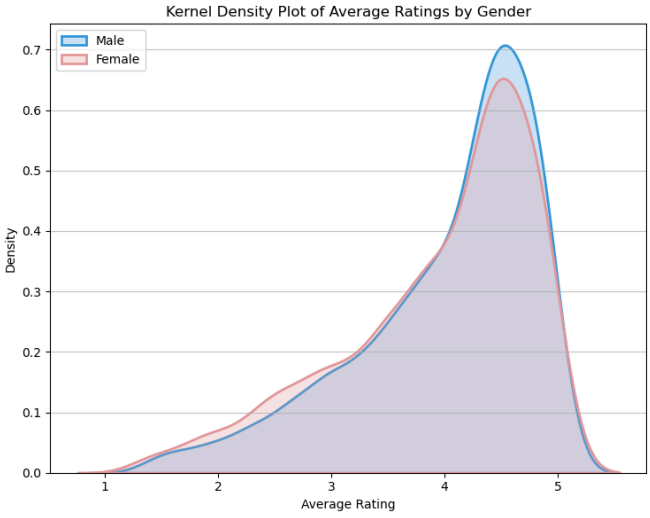
I explored four variables: gender, quality of teaching, teaching modality, and professor "hotness." To test the significance of each, I used the Mann-Whitney U test instead of the standard t-test. After analyzing the data, I found that the distributions were left-skewed, the

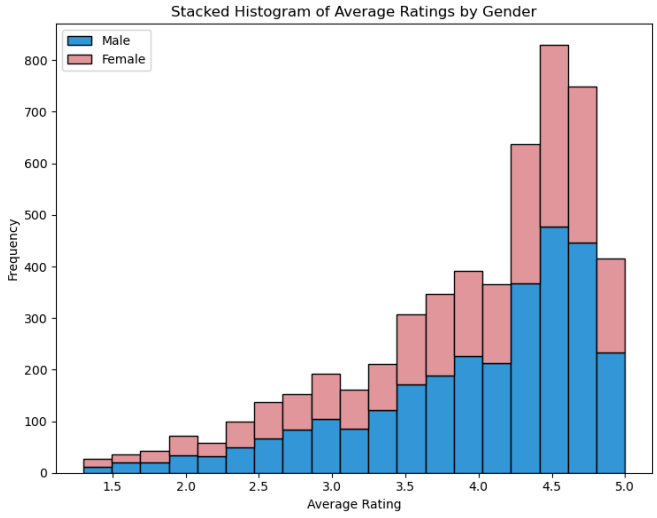
sample sizes were unbalanced, and there were differences in variance. Since the median is more robust to outliers than the mean, it was not appropriate to reduce the dataset to mean values. Instead, testing the differences in medians provided a more reliable metric, which aligns with the purpose of the Mann-Whitney U test.

1. *Is there evidence of a pro-male gender bias in this dataset?*

We observe unbalanced sample sizes between males and females and differences in variance shown in the descriptive box. In addition, the density plot and stacked histogram below highlight the left-skewed nature of the data. I conducted a one-sided Mann-Whitney U test to determine if male professors have a higher median average rating compared to female professors.

group	median	sd	n	se
Male	4.3	0.799577	2957	0.014704
Female	3.93984	0.84712	2274	0.0177644



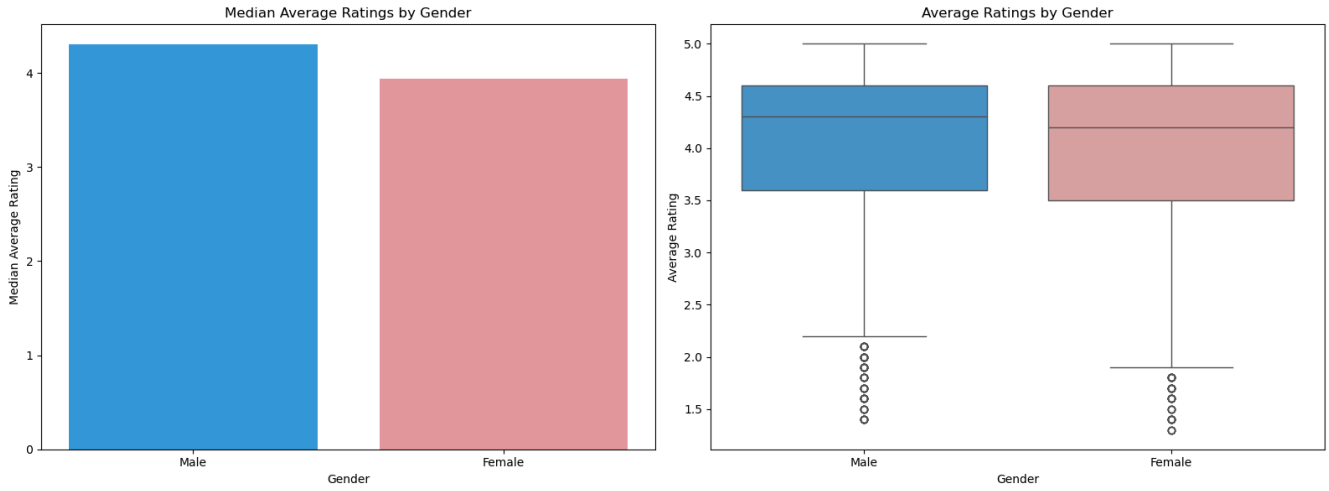


$$H_0: \text{male median average rating} > \text{female median average rating}$$
$$H_1: \text{male median average rating} \leq \text{female median average rating}$$

```
# Performs Mann-Whitney U test to assess if male ratings are significantly greater than female ratings
u1, p_u1 = mannwhitneyu(male_ratings, female_ratings, alternative='greater')

if p_u1 < alpha:
    print("Reject the null hypothesis: Evidence suggests a pro-male bias.")
else:
    print("Fail to reject the null hypothesis: No evidence of a pro-male bias.")
```

```
p_u1    0.0027319691810873
```

Because $p_{u1} = 0.00271 < \alpha = 0.005$, we reject the null hypothesis, providing significant evidence of pro-male bias in the dataset. **Hence, male professors, on average, receive higher ratings compared to female professors**, as reflected in the median differences shown below.

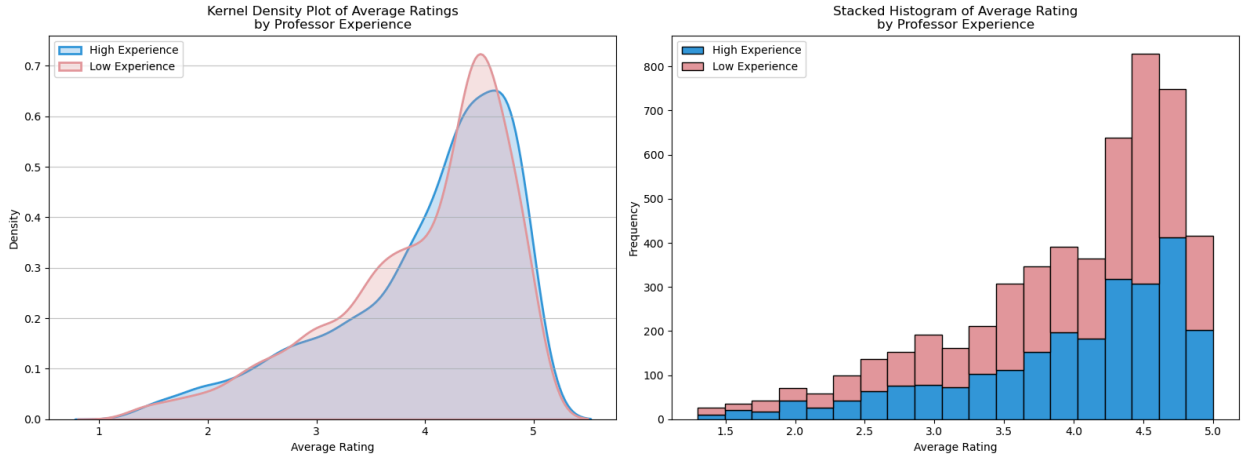


2. *Is there an effect of experience on the quality of teaching?*

I used average rating as a measure of quality and the number of ratings as an imperfect, but available, proxy for experience. We observe unbalanced sample sizes between high-experience and low-experience professors, along with differences in variance displayed in the descriptive box. In addition, the density plot and stacked histogram below illustrate the left-skewed nature of the data. Unlike a one-sided test, I opted for a two-sided hypothesis test to explore potential differences without assuming that one group inherently demonstrates superior teaching quality.

```
# Separates professors with high experience and professors with low experience
high_experience = professor_wt[professor_wt["num_rating"] > num_rating_median]["avg_rating"]
low_experience = professor_wt[professor_wt["num_rating"] <= num_rating_median]["avg_rating"]
```

group	median	sd	n	se
High Experience	4.3	0.831262	2441	0.016825
Low Experience	4.2	0.812439	2790	0.0153811



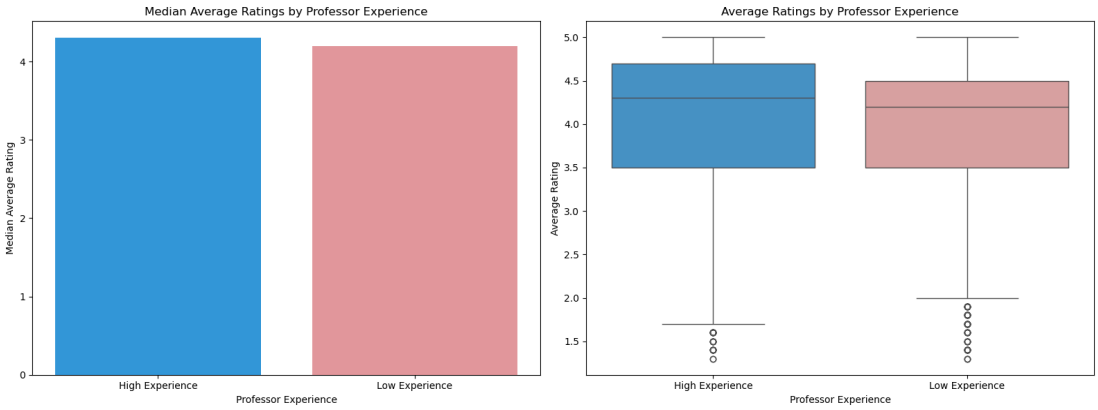
H_0: *high experience median average rating* $=$ *low experience median average rating*

H_1: *high experience median average rating* $\neq$ *low experience median average rating*

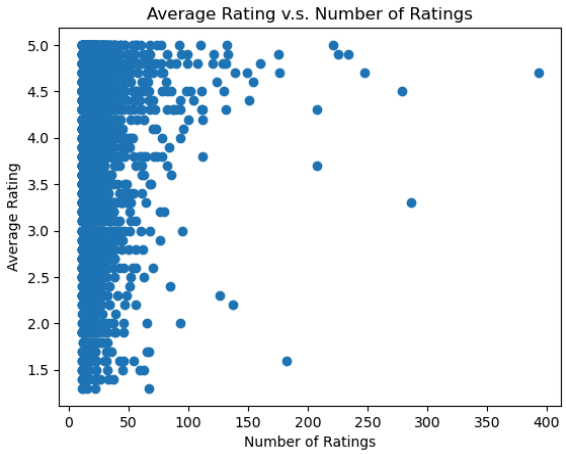
```
# Performs Mann-Whitney U test to assess if there is a difference between avg_rating for professors with high experience
# and professors with low experience
u2, p_u2 = mannwhitneyu(high_experience, low_experience, alternative='two-sided')

if p_u2 < alpha:
    print("Reject the null hypothesis: Evidence suggests a significant difference in ratings based on professor experience.")
else:
    print("Fail to reject the null hypothesis: No significant evidence that professor experience impacts ratings.")
```

p_u2	0.05068387140892033

Because $p_{u2} = 0.050 > \alpha = 0.005$, we fail to reject the null hypothesis. There is no difference between the median average ratings of high-experience and low-experience professors. **Hence, there is no sufficient evidence to suggest that professor experience impacts ratings.** This lack of difference is illustrated in the bar graph and box plot below, which compare the median average ratings by experience.



The scatterplot of average ratings versus the number of ratings supports this conclusion, which shows widely dispersed data with no clear trend. Although there is a slight upward pattern, it does not suggest a significant monotonic or non-linear relationship. Professors with more ratings do not consistently achieve higher averages, and the average ratings are scattered randomly, showing no meaningful connection between the number of ratings and the average rating.



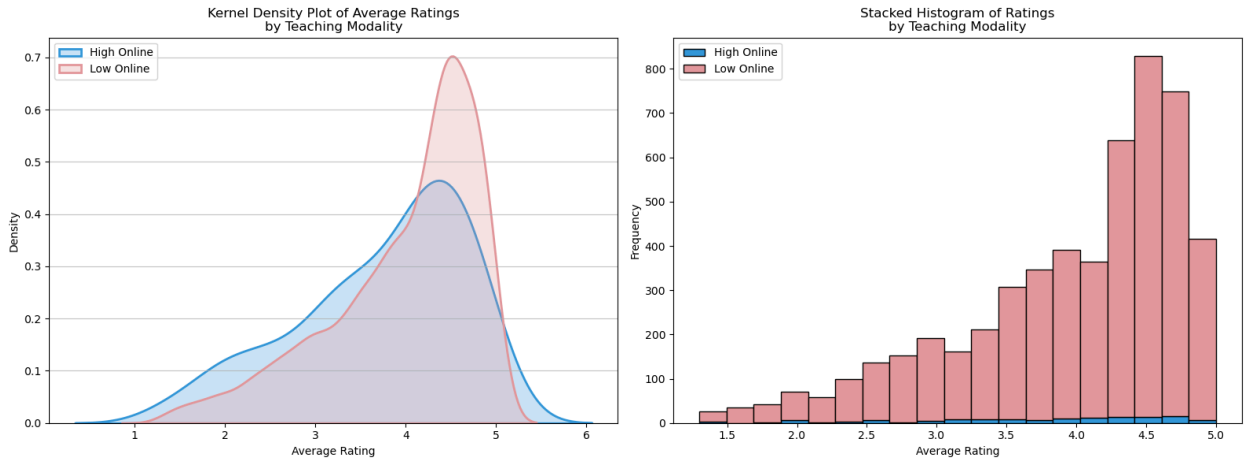
3. *Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't?*

To classify professors based on their engagement in teaching online courses, I used the percentage of online ratings as the metric. The classify_modality function calculates this percentage by dividing the number of online ratings by the total ratings. Professors with more

than 50% online ratings were categorized as teaching many online classes, while those with 50% or fewer were categorized as teaching fewer online classes. A two-sided Mann-Whitney U hypothesis test was conducted once again.

```
# Classification of teaching modality
def classify_modality(row):
    online_ratio = row['num_rating_online'] / row['num_rating']
    if online_ratio >= 0.5:
        return 'High Online'
    else:
        return 'Low Online'
```

group	median	sd	n	se
High Online	4	0.927419	127	0.0822952
Low Online	4.3	0.81758	5104	0.0114439



H_0: *high online median average rating* $=$ *low online median average rating*

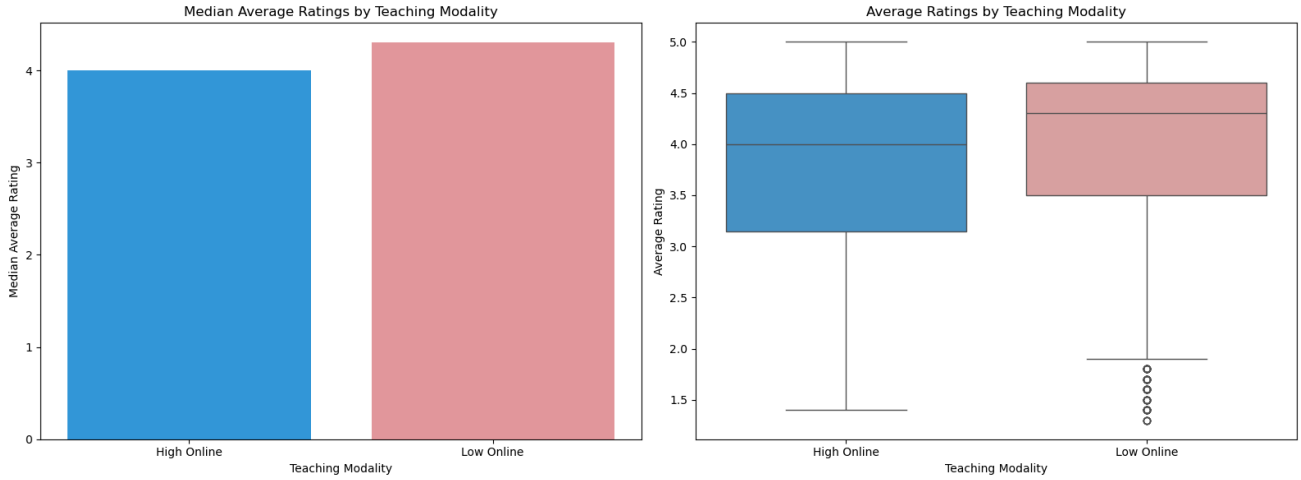
H_1: *high online median average rating* $\neq$ *low online median average rating*

```
# Performs Mann-Whitney U test to assess if there is a difference between avg_rating for professors who teach a lot of online classes
#and professors who do not teach a lot of online classes
u4, p_u4 = mannwhitneyu(high_online, low_online, alternative='two-sided')

if p_u4 < alpha:
    print("Reject the null hypothesis: Evidence suggests a significant difference in ratings based on teaching modality.")
else:
    print("Fail to reject the null hypothesis: No significant evidence that teaching modality impacts ratings.")
```

p_u4	0.0012123142943195957

With $p_{u4} = 0.00121 < \alpha = 0.005$, we reject the null hypothesis, providing significant evidence of a difference in average ratings between professors who teach many online classes

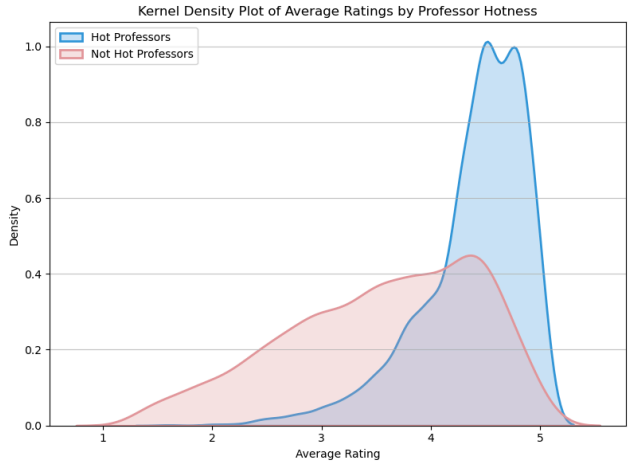
and those who teach fewer. The median average ratings demonstrate that professors teaching fewer online classes receive higher ratings. **Hence, professors who teach predominantly online tend to receive lower ratings compared to their counterparts teaching fewer online courses.** This difference is illustrated in the bar graph and box plot below.

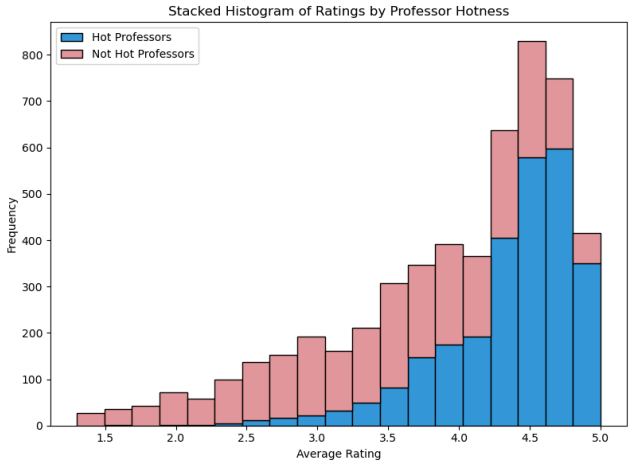


4. *Do professors who are "hot" receive higher ratings than those who are not?*

I classified professors into "hot" and "not hot" based on the binary variable provided in the dataset, where 1 represents "hot" and 0 represents "not hot." A one-sided hypothesis test was performed amongst the two groups.

group	median	sd	n	se
Hot	4.5	0.799577	2669	0.0094913
Not Hot	3.7	0.876431	2562	0.0173152





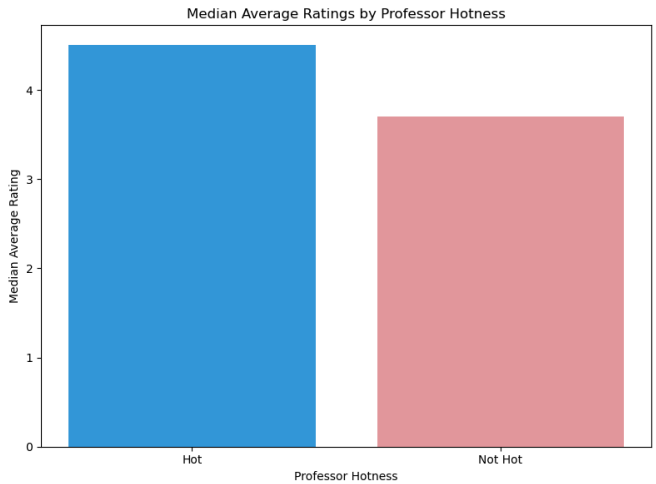
H_0: *hot professor median average rating* $>$ *not hot professor median average rating*

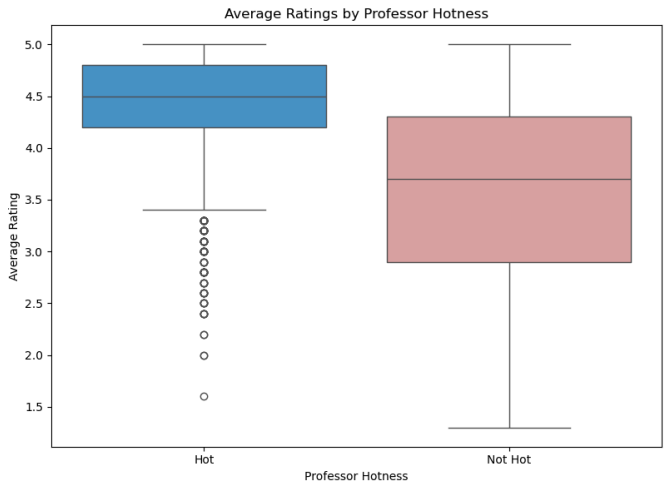
H_1: *hot professor median average rating* $\leq$ *not hot professor median average rating*

```
# Performs Mann-Whitney U test to assess if there hotter professors get higher ratings that professors who are not hot
u6, p_u6 = mannwhitneyu(yes_pepper, no_pepper, alternative='greater')

if p_u6 < alpha:
    print("Reject the null hypothesis: Evidence suggests that 'hot' professors receive higher ratings.")
else:
    print("Fail to reject the null hypothesis: No evidence that 'hot' professors receive higher ratings.")
```

```
p_u6    5.535330896150749e-304
```

Because $p_{u1} = 5.535e^{-304} < \alpha = 0.005$, we reject the null hypothesis. **There is significant evidence to suggest that "hot" professors receive higher average ratings than those classified as "not hot."** This difference is reflected in the median average ratings of the two groups, as shown below.





IV. Modeling and Interpretations

The exploratory data analysis provided insights into factors influencing professor ratings, such as gender bias, experience, teaching modality, and "hotness," setting the stage for predictive models to evaluate these relationships. The models aim to uncover how these variables, individually and collectively, impact average ratings while also assessing whether the "pepper" rating reflects teaching quality or non-academic biases. This analysis helps determine if its removal was justified and whether Rate My Professors is a reliable resource for evaluating professor performance.

5. *Can we predict average rating from average difficulty?*

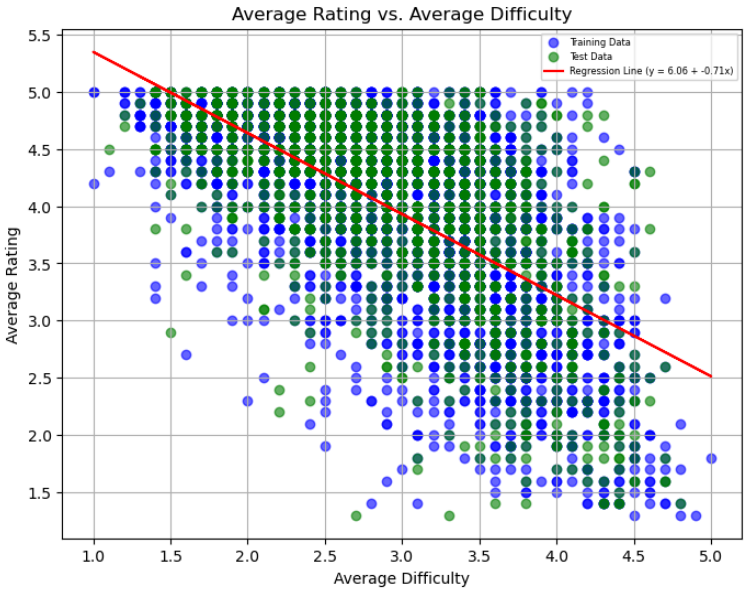
To uncover the relationship between average rating and average difficulty, I plotted a scatter plot, which revealed a **moderate negative linear relationship** between the two variables. To build a regression model predicting average rating from difficulty, I implemented a train-test split with a test size of 0.3, applying it to a linear regression model. **The resulting equation $\widehat{y} = 6.057 - 0.709x$ indicates that for every one-unit increase in average difficulty, the average rating decreases by 0.709, given that both variables are bounded between 0 and 5.**

The R^2 values for the training (0.404) and test (0.385) sets indicate that the model explains approximately 40% of the variance in average ratings based on average difficulty. This suggests a moderate relationship between the two variables. However, with R^2 values below 0.5, the model leaves a significant portion of the variance unexplained, indicating that average difficulty alone is insufficient as a predictor of average ratings.

The RMSE values, which measure the average prediction error, are nearly identical for the training (0.638) and test (0.636) sets. These values indicate that, on average, the model's predictions deviate from the actual ratings by around 0.64 points on a 1–5 scale. The close

alignment of RMSE values between the train and test sets suggests that the model is not overfitting and generalizes well to unseen data.

Together, the R^2 and RMSE results highlight that while the model captures some predictive power, the relationship between average difficulty and average rating is moderate at best. Additional variables would likely enhance the model's explanatory and predictive capabilities.



OLS Regression Results

Dep. Variable:	avg_rating	R-squared:	0.404
Model:	OLS	Adj. R-squared:	0.404
Method:	Least Squares	F-statistic:	2477.
Date:	Thu, 05 Dec 2024	Prob (F-statistic):	0.00
Time:	22:19:10	Log-Likelihood:	-3548.4
No. Observations:	3661	AIC:	7101.
Df Residuals:	3659	BIC:	7113.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.0570	0.043	140.186	0.000	5.972	6.142
avg_difficulty	-0.7089	0.014	-49.774	0.000	-0.737	-0.681

Omnibus:	170.996	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	198.598
Skew:	-0.525	Prob(JB):	7.50e-44
Kurtosis:	3.448	Cond. No.	13.7

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

6. *Can we predict average rating from all available factors?*

I performed a linear regression again, this time incorporating all the predictors from the dataset. To avoid multicollinearity and the dummy variable trap, where too many dummy variables create redundant information, I excluded is_female and used it as the reference category for gender. Then, I proceeded to use a OneHotEncoder to encode the categorical variables: receive_pepper and is_male. This allowed me to regress average rating on all available predictors except is_female. I then performed a train-test split with a test size of 0.3.

The coefficients for each variable in the model are displayed below. For categorical variables, the coefficient for C(receive_pepper)[T.1.0]=0.2123 indicates that "hot" professors

have average ratings 0.2123 points higher than "not hot" professors, holding all else constant. Similarly, the coefficient for C(is_male)[T.1]=0.0287 suggests that male professors receive average ratings 0.0287 points higher than female professors, holding all else constant. For continuous predictors, each coefficient represents the expected change in average rating for a one-unit increase in the respective variable, holding all other variables constant.

OLS Regression Results

Dep. Variable:	avg_rating	R-squared:	0.817
Model:	OLS	Adj. R-squared:	0.816
Method:	Least Squares	F-statistic:	2440.
Date:	Thu, 05 Dec 2024	Prob (F-statistic):	0.00
Time:	21:43:57	Log-Likelihood:	-1390.3
No. Observations:	3661	AIC:	2795.
Df Residuals:	3654	BIC:	2838.
Df Model:	6		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7019	0.053	50.734	0.000	2.598	2.806
C(receive_pepper)[T.1.0]	0.2123	0.014	15.601	0.000	0.186	0.239
C(is_male)[T.1]	0.0287	0.012	2.392	0.017	0.005	0.052
avg_difficulty	-0.2358	0.011	-21.595	0.000	-0.257	-0.214
num_rating	0.0002	0.000	0.643	0.520	-0.000	0.001
take_again	0.0237	0.000	62.468	0.000	0.023	0.024
num_rating_online	0.0010	0.003	0.394	0.693	-0.004	0.006

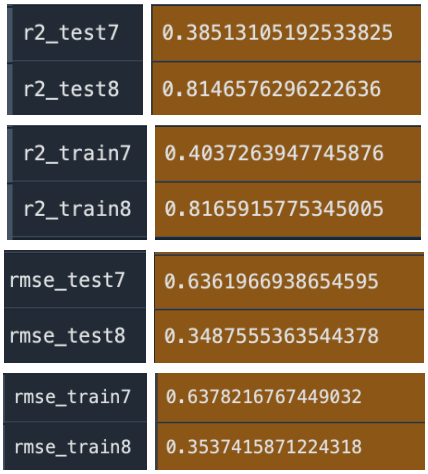
Omnibus:	235.558	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	388.501
Skew:	-0.504	Prob(JB):	4.35e-85
Kurtosis:	4.238	Cond. No.	661.

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

The multilinear regression model explains 81.659% of the variation in average rating for the training data ($R^2 = 0.81659$) and 81.466% for the test data ($R^2 = 0.81466$), indicating strong generalization and minimal overfitting. The RMSE values are 0.354 for the training set and 0.349 for the test set, reflecting improved predictive accuracy compared to the single-variable model. However, some noise and unexplained variation persist despite the inclusion of additional predictors.

The multilinear regression model significantly outperforms the single-variable model in explaining variation in average rating. The R^2 increases from 0.404 (train) and 0.385 (test) in the single-variable model to 0.817 (train) and 0.815 (test) in the multilinear model, indicating a

substantial improvement in explanatory power. Similarly, the RMSE decreases from 0.636 (train) and 0.637 (test) in the single-variable model to 0.354 (train) and 0.349 (test), respectively, reflecting enhanced predictive accuracy and reduced error. These results suggest that incorporating additional predictors, including avg_difficulty, num_rating, receive_pepper, take_again, num_rating_online, and is_male, significantly reduces unexplained variance and improves model's overall fit.

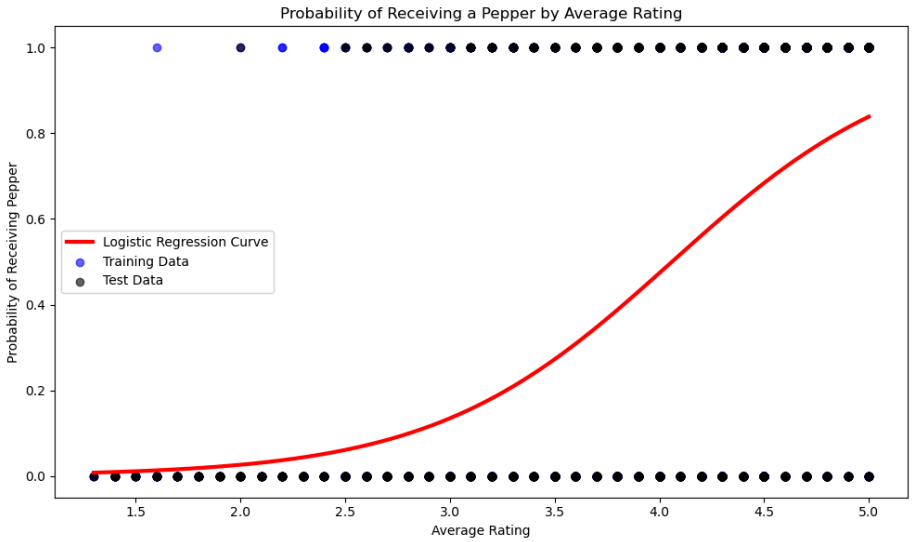


A key observation is the change in the coefficient for average difficulty between the two models. The reduction from -0.709 in the single-variable model to -0.236 in the multilinear model highlights the impact of omitted variable bias. In the linear model, average difficulty likely captured variance attributable to other predictors, which were excluded. By incorporating these additional variables, the multilinear model "unmasks" their contributions, leading to a more accurate understanding of the role of average difficulty.

7. *Can we predict whether a professor receives a 'pepper' based on their average rating?*

I used a logistic regression model to predict whether a professor receives a "pepper" based on average rating, applying a train-test split with a test size of 0.3. The logistic curve shows an inflection point between average ratings of 3.5 and 4.0, where the probabilities transition most sharply from low to high. This inflection point is critical as it marks the range

where professors begin to see a significant increase in their likelihood of receiving a pepper. Professors with ratings in this range are at a tipping point, moving from a moderate to a high probability of being classified as “hot.” In addition, the moderately steep curve indicates that the model effectively distinguishes between professors with low and high ratings, assigning very low probabilities to professors with lower ratings and much higher probabilities to those with higher ratings.

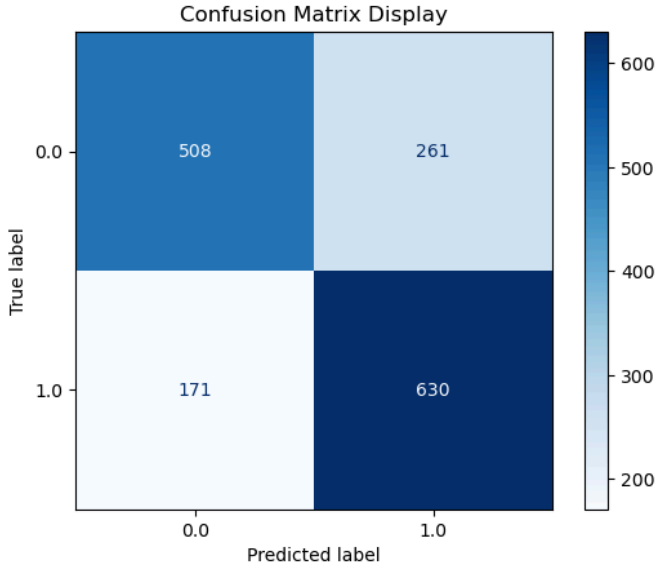


To evaluate the model, I analyzed precision and recall, which provide a more detailed assessment than accuracy given the slightly unbalanced classes. The classification model predicting whether a professor receives a “pepper” based solely on average rating shows notable differences in precision and recall between the “hot” and “not hot” categories. For “not hot” professors, precision is 0.75, meaning 75% of those predicted as “not hot” are correctly classified, while the remaining 25% are false positives. For “hot” professors, precision is slightly lower at 0.71, indicating that 71% of those predicted as “hot” are correctly classified, with some false positives present. Recall further highlights the model’s performance, with a recall of 0.66 for “not hot” professors, indicating the model identifies 66% of true “not hot” cases but misses

34%. In contrast, the recall for “hot” professors is 0.79, meaning the model successfully identifies 79% of professors who receive a pepper, though some are misclassified as “not hot.”

These precision and recall values demonstrate that the model is better at identifying “hot” professors, as evidenced by the higher recall for the “hot” category. However, the slightly lower precision for “hot” professors reflects a tendency for some misclassification. While the class imbalance between “not hot” (769 professors) and “hot” (801 professors) is relatively small, it may still influence the observed differences in recall and precision.

	precision	recall	f1-score	support
0.0	0.75	0.66	0.70	769
1.0	0.71	0.79	0.74	801
accuracy			0.72	1570
macro avg	0.73	0.72	0.72	1570
weighted avg	0.73	0.72	0.72	1570



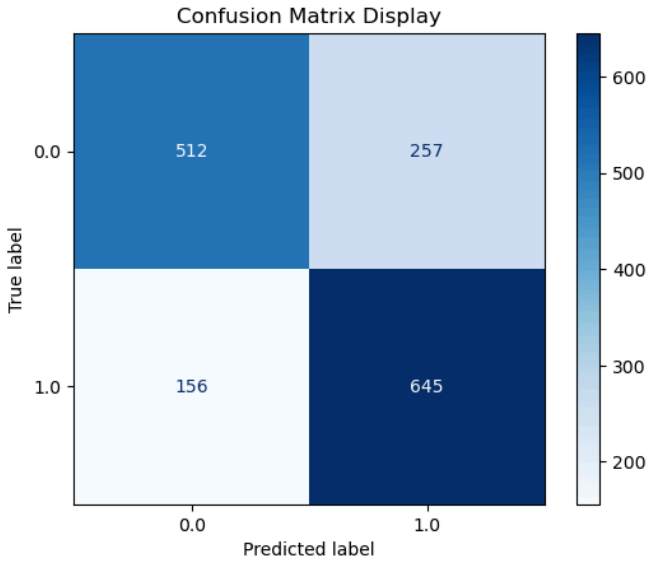
8. *Can we predict whether a professor receives a 'pepper' from all available factors?*

Similar to the previous test, a difference of 32 samples may not create a significant class imbalance, though it could still pose a potential issue due to the small sample size. From the classification report, precision for “not hot” professors and “hot” professors is 0.77 and 0.72, respectively, while recall is 0.67 and 0.81, respectively. This means that 77% of professors predicted as “not hot” and 72% of those predicted as “hot” are correctly classified. Meanwhile, the model captures 67% of true “not hot” professors and 81% of true “hot” professors, but misclassifies 33% and 19% of them, respectively. Compared to the previous question, where

only average rating was used as a predictor, both precision and recall show slight improvements. Specifically, precision for “not hot” professors increased from 0.75 to 0.77, and for “hot” professors from 0.71 to 0.72. Recall improved from 0.66 to 0.67 for “not hot” professors and from 0.79 to 0.81 for “hot” professors.

These marginal improvements indicate that adding more predictors slightly enhances the model’s ability to capture patterns in the data, though the impact is limited. Average rating remains a strong predictor, and the additional features provide minimal independent information to substantially improve performance. This suggests that the dataset might require even more powerful predictors to further improve the model’s overall fit.

	precision	recall	f1-score	support
0.0	0.77	0.67	0.71	769
1.0	0.72	0.81	0.76	801
accuracy			0.74	1570
macro avg	0.74	0.74	0.74	1570
weighted avg	0.74	0.74	0.74	1570



V. Discussion and Next Steps

a. Summary of Findings

This analysis examined whether RMP is a reliable and unbiased resource for evaluating professor performance. While predictive models explained a substantial portion of variance in ratings, particularly with multiple predictors, significant biases and limitations in RMP data were identified. Key findings include evidence of **gender bias**, with male professors receiving

systematically higher ratings, suggesting non-performance-related factors influence evaluations. The **"pepper" rating** strongly correlated with higher ratings, underscoring its subjectivity and further validating its removal in 2018. As one professor noted in a viral tweet:

> *Dear @ratemyprofessor Life is hard enough for female professors. Your 'chili pepper' rating of our 'hotness' is obnoxious and utterly irrelevant to our teaching. Please remove it because #TimesUP and you need to do better. Thanks, Female College Prof.*

While **teaching experience** showed limited impact on ratings, this may be due to the use of proxy variables such as the number of ratings. Ratings for professors teaching **primarily online** were significantly lower, reflecting potential biases against this modality. Predictive models demonstrated that **average difficulty** is moderately predictive of ratings, but additional predictors (e.g., gender and "pepper" status) improved explanatory power. Logistic regression revealed that **average ratings overwhelmingly drive "pepper" ratings**, with minimal contributions from other factors, underscoring the subjective and impression-based nature of this metric.

Although RMP offers some insights into student perceptions, its reliance on subjective and potentially biased metrics like gender and "pepper" status limits its utility as an objective evaluation tool. These findings suggest that RMP ratings should be interpreted cautiously and supplemented with other methods to ensure fair and accurate assessments.

b. Next Steps/Improvements

To improve the analysis and understanding of RMP ratings, three key approaches are recommended. First, integrating a **Random Forest model** can capture non-linear relationships and interactions between variables like average difficulty, teaching modality, and "pepper" status, while providing feature importance scores to identify key drivers of ratings. Second, using **text feature extraction on student comments**, such as sentiment analysis and topic modeling, could

uncover qualitative trends and biases, highlighting themes like "clarity" or "grading fairness" that influence evaluations. Finally, **refining data cleaning techniques**, such as multiple imputation for missing values and relaxed thresholds for low-rating counts, could preserve more of the dataset, reducing the risk of losing valuable patterns and insights.

c. Suggestions for RMP

To address systemic biases, RMP should implement machine learning strategies like **bias detection** and **weighted rating models**. Bias detection, using methods like k-Means clustering, can flag and correct ratings influenced by non-academic factors. Weighted models can prioritize feedback from larger, more representative groups, reducing the impact of outliers and low-response classes. These strategies would enhance fairness and improve the accuracy of RMP ratings.