Stella Lin

Principles of Data Science Capstone Project

Professor Pascal Wallisch

9 December, 2024

What Makes a Good Professor on ratemyprofessor.com?

I begin by performing the pre-processing steps in order to analyze the dataset. This included renaming all the columns accordingly and setting them as the column headers, identifying and removing null values that would lead to errors in processing later on and displaying a test print of the first 5 rows of data to ensure accuracy.

In this project I use the word “we” when describing the steps I took, but I am the only contributor to this report.

```
Average Rating        19889
Average Difficulty    19889
Number of ratings     19889
Pepper                19889
Take Again            77733
Online Class          19889
Male                      0
Female                    0
dtype: int64
```

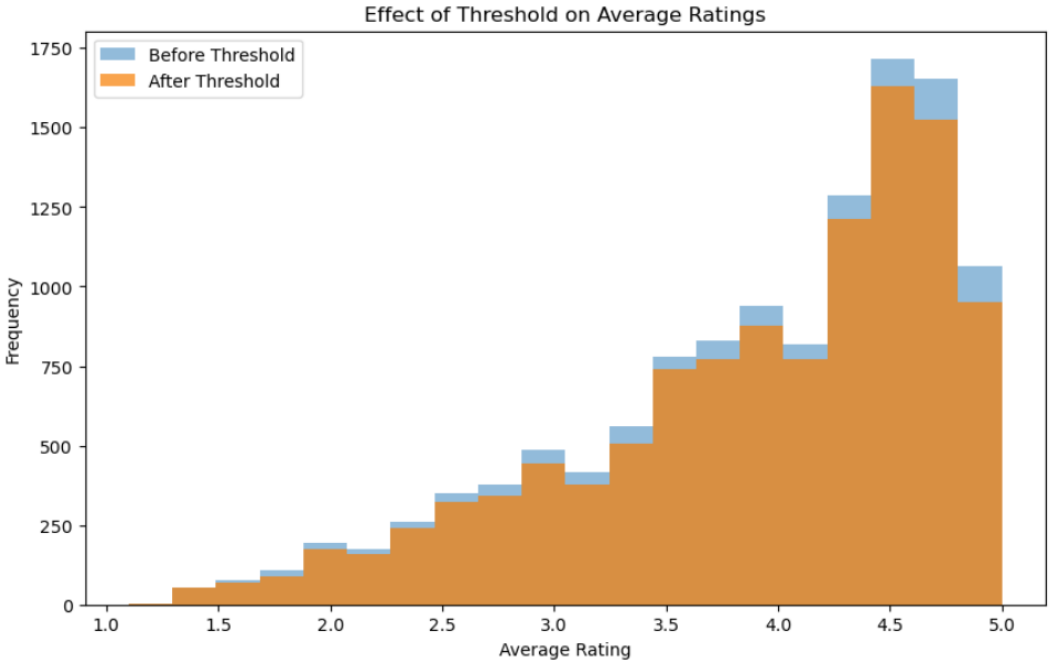
	Average Rating	Average Difficulty	Number of ratings	Pepper	Take Again	Online Class	Male	Female
5	3.5	3.3	22.0	0.0	56.0	7.0	1	0
25	4.3	3.3	16.0	1.0	83.0	0.0	0	1
40	1.8	3.8	15.0	0.0	22.0	1.0	0	1
42	4.1	3.3	21.0	0.0	67.0	0.0	0	1
46	4.2	1.8	26.0	1.0	57.0	8.0	1	0

Calculating average with a threshold

	Number of ratings	Average Rating
count	12160.000000	12160.000000
mean	15.532730	3.936135
std	14.701926	0.846237
min	5.000000	1.100000
25%	9.000000	3.400000
50%	12.000000	4.200000
75%	17.000000	4.600000
max	393.000000	5.000000

We set a threshold requiring professors to have at least 5 ratings to be included in the analysis. This was done to reduce the influence of extreme values on the average ratings as the average rating is more meaningful and reliable when it is based on a larger sample size. Typically people

who leave ratings for professors tend to have stronger opinions in one direction or the other, and by ensuring a minimum number of ratings per professor, we attempt to normalize the extreme values and obtain a more representative measure of average ratings.



I used a stacked histogram to visualize the effect of applying thresholds on the distribution of average ratings. As shown above, setting the threshold resulted in a reduction in the number of ratings at the extreme positive end of the scale, specifically ratings of 5.0 and 4.5. This indicates that professors with fewer than 5 ratings were more likely to have highly positive averages, which could skew the results. By applying the threshold, we aim to control this skewness and ensure that the analysis reflects more consistent and reliable patterns across professors with a sufficient number of ratings.

	Number of ratings	Average Rating
count	11251.000000	11251.000000
mean	16.322638	3.940814
std	15.008291	0.838812
min	7.000000	1.100000
25%	9.000000	3.500000
50%	12.000000	4.200000
75%	18.000000	4.600000
max	393.000000	5.000000

After applying a threshold of 6, the dataset was reduced to 11251 ratings compared to the 12160 ratings in the original dataset. This remaining sample size is still substantial, ensuring that the findings remain statistically significant and are applicable to the broader population.

1).

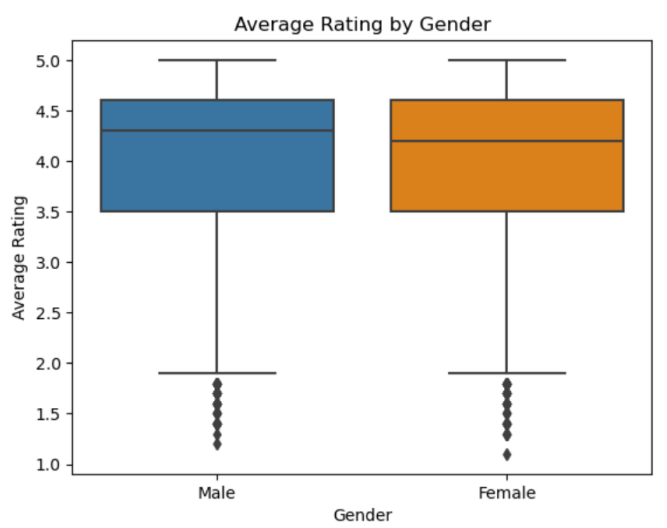
I am testing if sex has an effect on the ratings of professors. To do this, we designate a null hypothesis:

H0: There is no significant difference in the average ratings of male and female professors.

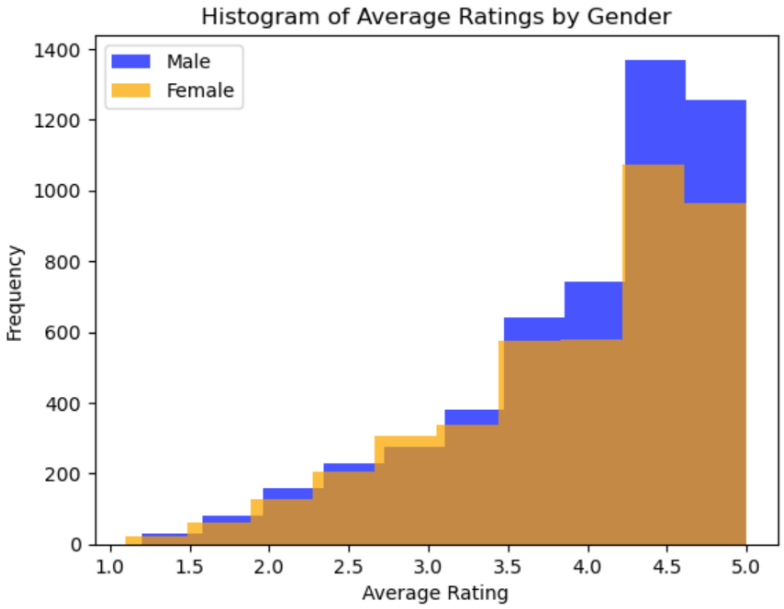
Ha: Average ratings for male professors are significantly higher than those for female professors.

A Mann-Whitney U test was conducted to evaluate whether there is a significant difference in the average ratings between male and female professors. The test included 5164 ratings for male professors and 4245 ratings for female professors.

The test produced a U statistic of 11409083.5 and a p-value of 0.0006133063105673826. reveal that there is a significant difference in the distribution of the two groups. Given the p value of 0.00061. Since the p value is less than our significance level of 0.005, we have sufficient evidence to reject the null hypothesis. Thus, we can conclude that there is a statistically significant difference in the distribution of ratings between male and female professors. Specifically, male professors tend to receive higher ratings on average compared to female professors.



The box plot did not visually highlight the results as clearly as the statistical evidence from the Mann-Whitney U test. While there is a slight difference in the median ratings between male and female professors, the visual distinction is not pronounced enough to suggest a significant difference. Therefore, I used a stacked histogram to better illustrate the distribution of ratings for male and female professors. The histogram provides a clearer depiction of the differences in the distribution between the two groups, aligning more closely with the statistically significant p-value from the Mann-Whitney test.



The histogram reveals a clear difference in the frequency of ratings for male and female professors, particularly for ratings of 3.5 and above. While the distribution of ratings between 1.0 - 3.5 shows less pronounced differences between the sexes, the differences are much more noticeable for higher ratings.

- For ratings around 4.0, male professors received approximately 750 ratings, compared to fewer than 600 for female professors.
- At the highest rating of 5.0, male professors received around 1,300 ratings, whereas female professors received approximately 950 ratings.

These differences align with the statistical findings from the Mann-Whitney U test, further supporting the conclusion that male professors tend to receive higher ratings on average compared to female professors.

2).

To test if there is an effect of experience on the quality of teaching, we designate the following hypotheses:

H0: Experience does not have a significant effect on teaching quality

Ha: Experience has a significant effect on teaching quality (either negative or positive)

A Mann-Whitney U test was conducted to compare teaching quality (as measured by average ratings) between two experience groups (e.g., high and low). The results yielded a U statistic of 16,950,151.0 and a p-value of 0.6035. Since the p-value is much greater than the significance level of 0.005, we fail to reject the null hypothesis. This indicates that there is no significant difference in teaching quality between the experience groups.

A Spearman Rank Correlation was then performed to assess the relationship between teaching experience and teaching quality. The Spearman correlation coefficient was 0.0219, with a p-value of 0.0163. Although the correlation coefficient suggests a very weak positive relationship, the p-value is not less than the alpha level of 0.005. Therefore, we conclude there is no significant effect of experience on teaching quality.

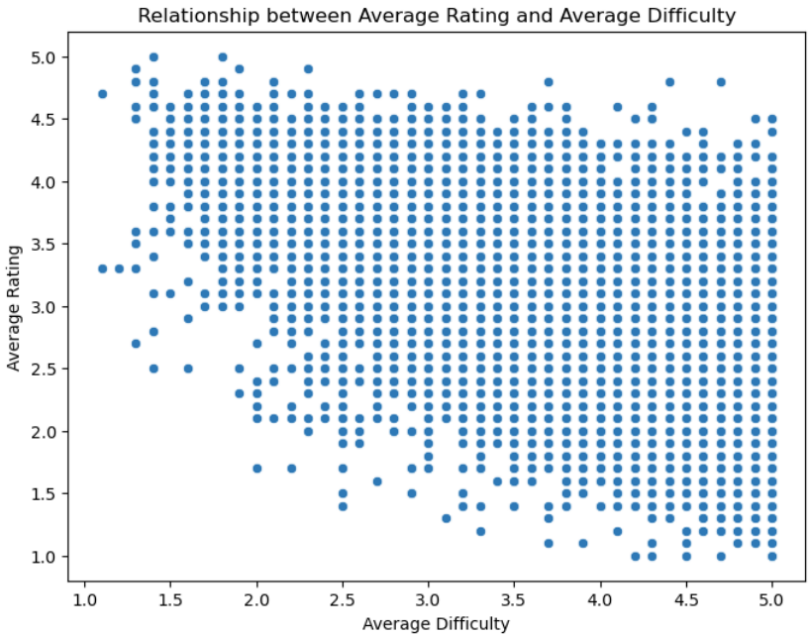
We then used box charts to visualize the effect of experience on teaching quality, with one plot representing professors with less experience (10 or less ratings), and the other plot for professors with high experience (more than 10 ratings). These visualizations showed overlapping distributions with similar medians, suggesting minimal differences in teaching quality across the two experience groups.

The box plot aligns with the statistical tests, which also found no significant effect of experience on ratings for professors.

3).

To analyze the relationship between average rating and average difficulty, we used Spearman's rank correlation, given that the data is not normally distributed. The Spearman correlation coefficient is -0.58, with a p-value: 0.0. This indicates a moderate, negative correlation, suggesting that as the difficulty of a course increases, the average rating tends to decrease.

We then used a scatterplot to visualize the relationship between difficulty and ratings. The plot shows a clear downward trend, where higher levels of difficulty are associated with lower average ratings. This visual representation supports the Spearman correlation result, which indicates a moderate negative correlation between the two variables.



4).

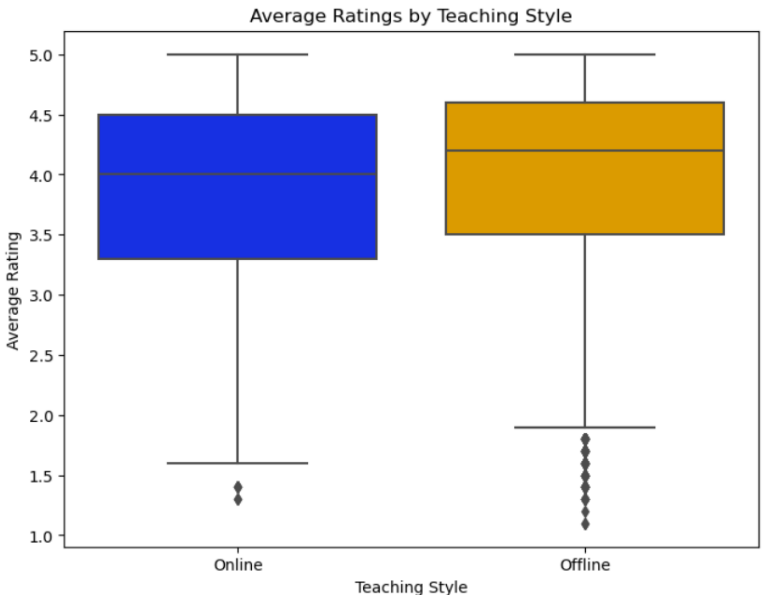
To compare if professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't, we evaluated the following hypotheses:

H0: There is no significant difference in average ratings for professors who teach mostly classes in the online modality and those who teach mostly in-person.

Ha: There is a significant difference in average ratings for professors who teach a lot of classes in the online modality and those who don't. Professors who teach mostly in-person classes receive higher ratings.

In order to split the variable "online", since it's binary, I calculated the proportion of online classes as Online Class / Number of ratings. I then created a subset "mostlyonline", containing ratings for professors with more than 50% of their classes online, and "mostlyoffline", containing ratings for professors with 50% or fewer online classes.

The Mann-Whitney U test was then conducted to compare the average ratings between professors who teach mostly online classes and those who teach mostly in-person classes. The results showed a U statistic of 476,365.0 and a p-value of 0.0126. Since the p-value is greater than the alpha level of 0.005, we fail to reject the null hypothesis. This means that there is no statistically significant difference in the average ratings between professors who teach mostly online classes and those who teach mostly in-person classes.



The scatterplot shows that the distribution of average ratings for mostly offline professors appears to be slightly higher compared to mostly online professors. However, the results from the Mann-Whitney U test indicate that this difference is not statistically significant.

5).

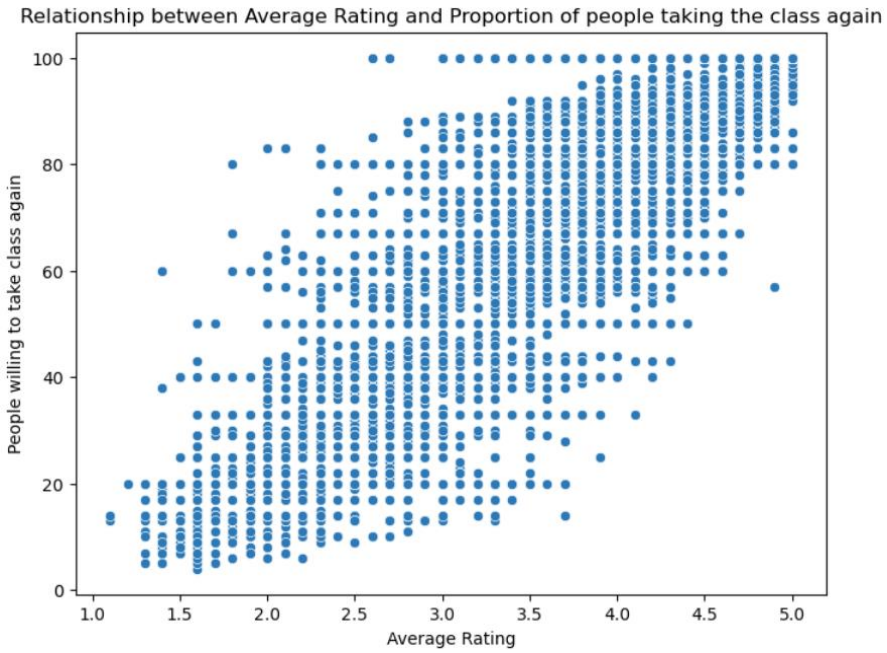
In order to assess the relationship between the average rating and the proportion of people who would take the class the professor teaches again, I used Spearman's rank correlation.

Spearman correlation: 0.8504659006280785, p-value: 0.0

The result of the Spearman correlation is 0.85, which is very close to 1, indicating a strong, positive monotonic relationship. This suggests that as the average rating of a professor increases, the proportion of students willing to take that professor's class again also increases. A p-value of 0.0 indicates that this relationship is statistically significant, meaning that the observed correlation is highly unlikely to have occurred by chance.

This strong correlation aligns with the intuition that students who rate a professor highly are more likely to recommend them to others or be willing to take another class with them. However, it's important to note that while the correlation is strong, it does not imply causality. Other factors may also be influencing both the ratings and students' willingness to retake the class.

I also used a scatterplot below to visualize this relationship.



As shown in the scatterplot, there is a clear upward trend in the data points, indicating a strong, positive relationship between average rating and the proportion of students willing to take the class again. This suggests that higher ratings are associated with a higher likelihood of students choosing to take the professor's class again. This relationship is consistent with the strong Spearman correlation coefficient of 0.85, which suggests a robust, statistically significant association.

6).

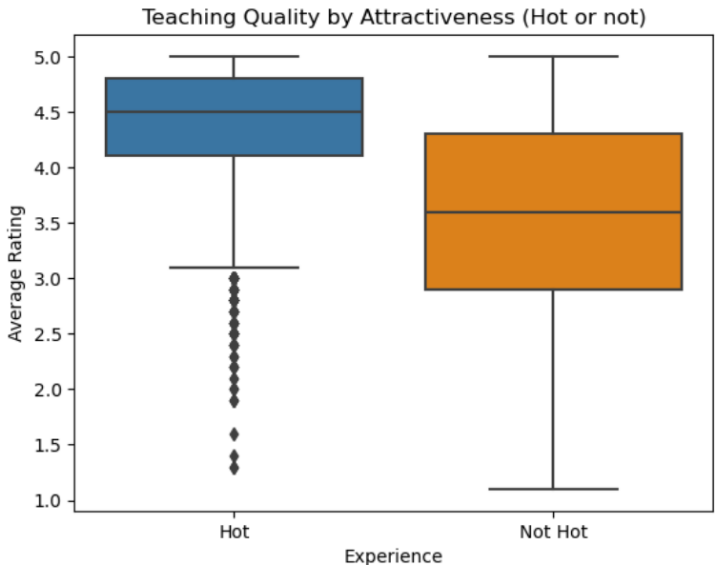
In order to determine if professors who are "hot" receive higher ratings than those who are not, we evaluate the following hypotheses:

H0: The attractiveness of a professor does not significantly impact their ratings.

Ha: The attractiveness (hot or not) of a professor has a significant impact on their ratings; hotter professors receiving higher ratings than those who are not.

U statistic: 27976547.5, p-value: 0.00e+00

We conducted a Mann-Whitney U test and the results show a significant difference in the ratings of "hot" versus "not hot" professors. The U statistic is 27,976,547.5 and the p-value is 0.000, which is less than the alpha level of 0.005. Therefore, we have sufficient evidence to reject the null hypothesis. We can conclude that the attractiveness of a professor does have a statistically significant impact on their ratings. Professors who are "hot" tend to receive higher ratings compared to those who are not. This is a social phenomenon often referred to as "pretty privilege", highlighting the potential influence of physical appearance on perceived quality, independent of teaching performance.



The boxplot comparing the distributions of average ratings for hot versus not hot professors shows a clear distinction between the two groups. Hot professors tend to have a strong rightward skew in their ratings, suggesting that their ratings are generally higher, with more ratings clustered around the higher end. The median average rating for hot professors is 4.5, while the median average rating for not hot professors is significantly lower at around 3.5. This further supports the finding that physical attractiveness appears to correlate with higher ratings, as shown in the Mann-Whitney U test results.

7).

I built a linear regression model to predict average rating from difficulty. I first isolated the predictor variable “Difficulty”, and the target variable, “Average Rating”.

I then split the data into training and test sets using a 70-30 split, ensuring that the model would be evaluated on unseen data. The regression model was trained on the training set and then used to make predictions on the test set. I then used two metrics to assess the performance of the model: R-squared and RMSE.

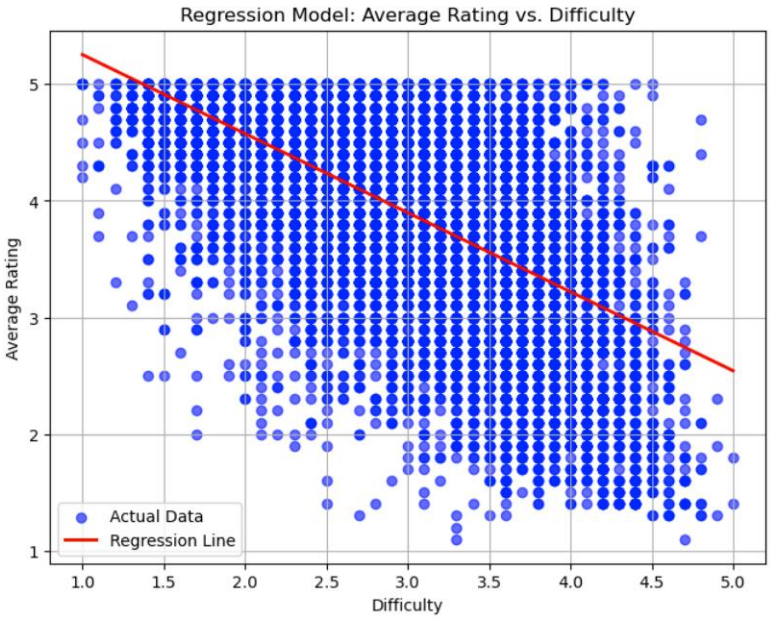
- The R-squared score is 0.4031, meaning approximately 40.31% of the variance in average ratings can be explained by the difficulty of the class. This indicates a moderate relationship between difficulty and ratings, but other factors are likely playing a role in shaping the ratings.
- The RMSE value, which measures the average error between predicted and actual ratings, is 0.6493, which shows that the average difference between the predicted and actual ratings is about 0.65. While this suggests a reasonable level of accuracy, it also indicates that the model is not perfect, and that there is room for improvement.
- I also looked at the Coefficient for Difficulty: The coefficient of -0.7072 indicates that for every 1 unit increase in difficulty, the average rating decreases by 0.7072. This suggests that as courses become more difficult, students tend to give lower ratings on average.

RMSE: 0.6493

R-squared: 0.4031

Intercept: 6.0420

Coefficient for Difficulty: -0.7072



I also created a scatter plot to visualize the relationship between Difficulty and Average Rating. In this plot, each point represents a pair of Difficulty and Average Rating for a professor. The downward trend in the scatterplot shows a negative relationship between Difficulty and Average Rating. This means that generally, as the difficulty level of a class increases, students tend to rate professors lower. This could indicate that students may find more difficult courses less enjoyable or more challenging, leading to lower ratings.

8).

I built a multiple regression model predicting Average Rating using all available factors with the following results:

RMSE: 0.3460

R-Squared: 0.8305

Intercept: 3.949193688226114

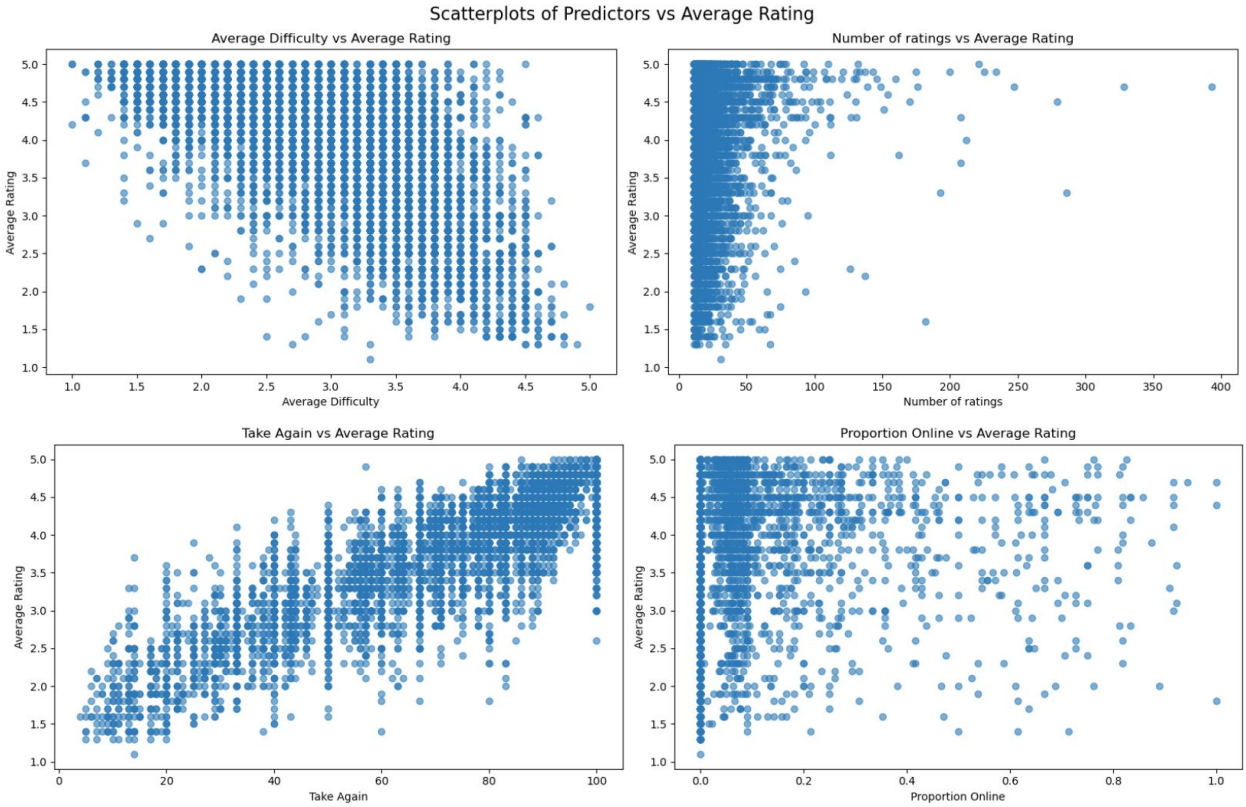
Coefficients: [-0.16741006 0.00532448 0.10876045 0.58268875 0.00358535 0.02719288 0.01120966]

The R-squared value in the multiple regression model is 0.8305, indicating that 83.05% of the variance in the Average Rating can be explained by the predictors in this multiple regression model. This is a strong fit, suggesting that the model does a good job of capturing the key factors influencing average ratings. When compared to the single-variable "Difficulty" model (0.4031), the multiple regression model performs

much better, explaining a significantly higher proportion of the variance in ratings. This shows that the inclusion of additional variables improves the model's explanatory power.

Additionally, the RMSE value, which represents the average error in predictions, is 0.3460, indicating a relatively low prediction error. Compared to the "Difficulty-only" model (RMSE of 0.6493), the multiple regression model has a much smaller prediction error, suggesting that the inclusion of more predictors helps to refine the accuracy of the model's predictions.

I used a scatterplot to show the distribution of each factor with Average Rating. I chose to omit the variables "Hot", "Male" and "Female", as they are binary variables that are not expressed well in scatterplots.



Analyzing the trend in the factors:

- A negative trend in Average Difficulty and Average Rating, indicating that as difficulty in a class increases, the average rating of that professor decreases.
- A strong, positive trend in Take Again and Average Rating, indicating that a high rating of a professor is correlated with a higher proportion of students who would take the class again
- The relationship between Number of Ratings and Average Rating is unclear, indicating that a scatterplot is not the best figure to use to show this relationship. In Question 2, we found that there was no significant difference in average ratings of professors with more experience vs less experience.
- The relationship between Proportion of Classes Online and Average Ratings is also unclear, indicating that a scatterplot is not the right graph to express this relationship. In Question 4

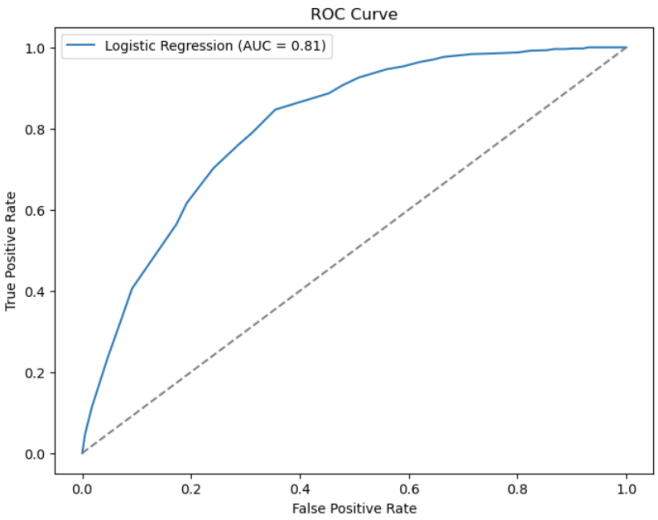
we found that there was no significant difference in average ratings of professors who mostly teach class online vs offline.

9).

To build a classification model that predicts whether a professor receives a "pepper" (binary target variable: 1 = "pepper", 0 = no "pepper") based on average rating, we used a Logistic Regression model, which gives us the probability of the positive class ("pepper").

To address class imbalance, we used SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset by generating synthetic samples for those who receive a "pepper". This helps prevent the model from being biased towards the majority class, which are those who do not receive a "pepper".

AUC-ROC: 0.8080266586660477



Confusion Matrix:

[[504 229]

[152 571]]

Classification Report:

	precision	recall	f1-score	support
0.0	0.77	0.69	0.73	733
1.0	0.71	0.79	0.75	723
accuracy			0.74	1456
macro avg	0.74	0.74	0.74	1456
weighted avg	0.74	0.74	0.74	1456

We then compute the AUC-ROC score, which measures the ability of the model to distinguish between the positive and negative classes. Our AUC-ROC score of 0.808 is close to 1, indicating that the model is good at distinguishing between professors who receive a pepper and who does not based on the average rating. We can also determine that our model is relatively accurate by the high TN and TP scores in the Confusion Matrix. We have 504 and 571 compared to 229 and 152 for FP and FN. The number of True Negatives and True Positives is greater than False Positives and False Negatives, indicating that the model is correctly predicting most "Pepper"s. The Classification Report also shows a precision score of 0.71 for class 1 (Pepper), meaning 71% of the professors who were predicted to receive peppers were accurate. The recall value of 0.79 for class 1 (Pepper) indicates that 79% of professors who received peppers were correctly predicted by the model.

Overall, the high AUC-ROC score along with high precision and recall suggests that the model is doing well at accurately predicting "pepper" amongst professors.

10).

While using all available variables to predict whether a professor receives a “pepper, we computed the AUC-ROC score of 0.8132, which is greater than the AUC-ROC score when only using "Average Ratings" as the predicting variable(0.808). Our AUC-ROC score of 0.8132 is close to 1, indicating that the model is good at distinguishing between professors who receive a pepper and who does not based on all available ratings. Since the AUC-ROC score is higher for the model with more variables, this suggests that a model with more predictive variables more accurately predicts the outcome for one variable. Looking at the precision and recall scores for the model using all variables, the precision score is the same as the Average Rating model, but the recall score is higher (0.82 vs 0.79), meaning that 82% of professors who received peppers were correctly predicted by the model, compared to 79% from before. The confusion matrix for this model also has more values in the True Positive and True Negative sections compared to the single-variable model, indicating that the model is correctly predicting more "Pepper"s when given more variables than with a single variable.

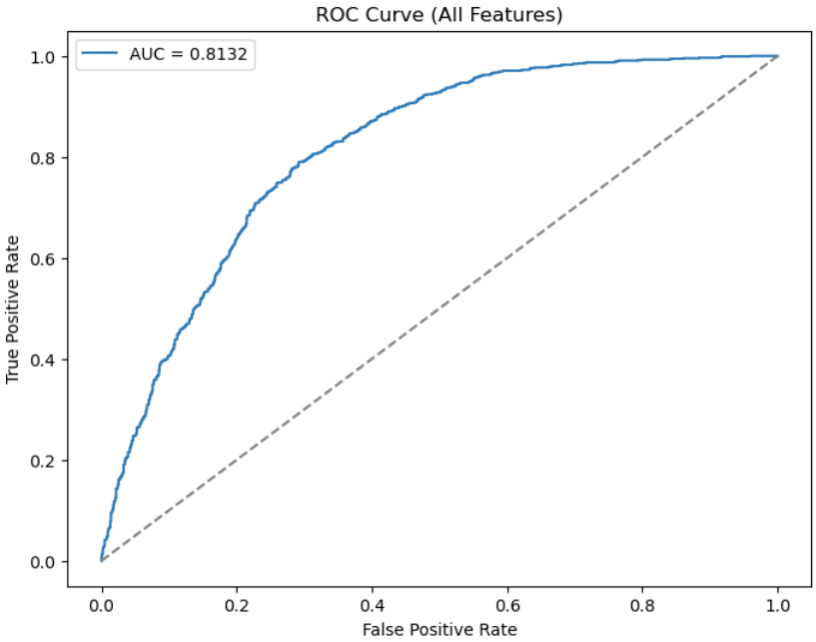
```
Classification Report (All Features):
              precision    recall  f1-score   support

         0.0       0.78      0.67      0.72      1071
         1.0       0.71      0.82      0.76      1092

    accuracy                           0.74      2163
   macro avg       0.75      0.74      0.74      2163
weighted avg       0.75      0.74      0.74      2163
```

AUC-ROC (All Features): 0.8132

Confusion Matrix:

[[715 356]

[199 893]]



	Feature	Coefficient
0	Average Rating	1.526246
1	Average Difficulty	0.218814
2	Number of ratings	0.178815
3	Take Again	0.146047
6	Female	0.008238
4	Online Class	-0.015131
5	Male	-0.151345

I also calculated the regression coefficients for all the available factors with the variable "Pepper". The variables most highly correlated with "Pepper" are Average Rating (1.526) and Average Difficulty (0.2188). This means that for every 1-unit increase in Average Rating, the log-odds of the professor receiving a pepper increases by approximately 1.526. Similarly, with each 1 unit increase in Average Difficulty, the log-odds of the professor receiving a pepper also increases by 0.2188.

While this suggests that higher Average Difficulty is somewhat related to an increased likelihood of receiving a pepper, it was a surprising finding, as difficulty and 'hotness' of a professor are not usually directly linked in common understanding. This indicates that there could be confounding variables involved to influence this correlation. It would be important to investigate further, such as checking for collinearity between Average Difficulty and other features like Average Rating, to see if one is influencing the other's effect.