Livia Morales

Pascal Wallisch

Principles of Data Science (DS-UA-112)

30 April 2021

Assessing Professor Effectiveness Data Analysis Capstone Project

Data Preprocessing and Random Seeding:

For this project, I began by seeding the random number generator using my N-number as instructed to ensure randomness across reports:

```
n_number = "N12392083"
seed_value = int(n_number[1:])
np.random.seed(seed_value)
random.seed(seed_value)
```

In order to handle missing values, I first ran my code to determine how much of each category had missing values:

```
Number of missing values for each column:
Average_Rating        19889
Average_Difficulty    19889
Number_of_Ratings     19889
Pepper                19889
Would_Take_Again      77733
Online_Ratings        19889
Male                      0
Female                    0
```

In my analysis, I began each problem by first excluding ratings with missing values (NaNs) in the relevant variables I was concerned with for each analysis rather than removing all ratings with missing ratings in any column in the beginning in order to keep as much data as possible given the large amount of missing values. I also applied a threshold to exclude ratings with fewer than five ratings in order to ensure some level of reliability for professor ratings across all questions. Finally, all tests were conducted with an alpha level of 0.005 as the threshold for statistical significance for all of my results. The dataset I used contains values representing Average Rating, Average Difficulty, number of ratings, “pepper” status (if the professor was judged as “hot” by the student), proportion of students who said that they would take the class again, number of ratings from online classes, male gender (Boolean - 1), and female gender (Boolean - 1).

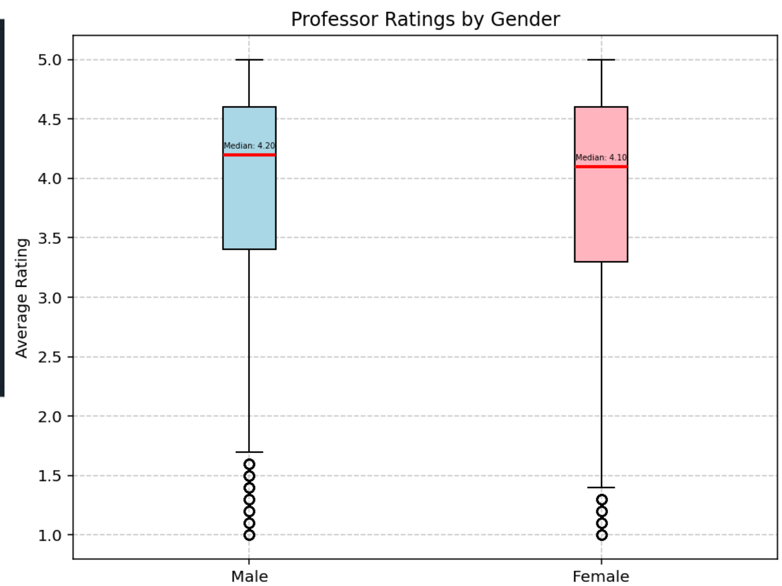
1. In order to determine if there is evidence of pro-male gender bias in the dataset I conducted both a **Mann-Whitney U test** to compare raw average ratings between male and female professors and performed a **multiple linear regression** to control for possible confounding variables, in this case course difficulty, number of ratings, “pepper” status, online teaching, and the proportion of students who would take the class again. It is important to note that although average ratings were treated as continuous for the purpose of linear regression, these ratings are on an ordinal scale, so the assumption that the intervals between ratings is equal may not hold. The Mann-Whitney U test showed the **median rating for male professors is 4.20**, and the **median rating for female professors is 4.10**, with a **p-value of 0.0004** which we can conclude is statistically

significant given it is less than 0.005.The Regression Analysis Output demonstrates that our model explains about 81% of the variance in Average Rating given the **R-squared value of 0.808**. The male variable coefficient has a value of **0.0227** which suggests that a professor being male is associated with a 0.0227 increase in rating and is significant since the p-value of **0.00385** is below our significance threshold of 0.005 (Figure 1).

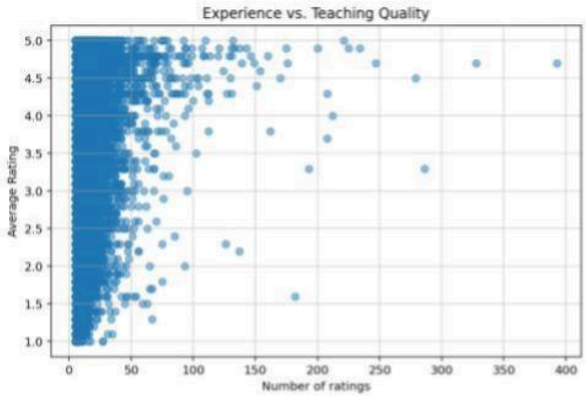
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.808
Model:                            OLS   Adj. R-squared:                  0.808
Method:                 Least Squares   F-statistic:                     6206.
Date:                Tue, 29 Apr 2025   Prob (F-statistic):               0.00
Time:                        16:20:59   Log-Likelihood:                -3635.0
No. Observations:                8849   AIC:                             7284.
Df Residuals:                    8842   BIC:                             7334.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.5374      0.029     88.727      0.000       2.481       2.593
x1             0.0227      0.008      2.891      0.004       0.007       0.038
x2            -0.1999      0.006    -32.759      0.000      -0.212      -0.188
x3            -0.0002      0.000     -0.822      0.411      -0.001       0.000
x4             0.0249      0.000    124.529      0.000       0.024       0.025
x5             0.1971      0.009     22.710      0.000       0.180       0.214
x6            -0.0017      0.002     -0.857      0.392      -0.006       0.002
==============================================================================
```

Regression with gender = male (x1), average difficulty (x2), number of ratings (x3), would take again (x4), pepper (x5), online ratings (x6)

Figure 1.



2. To determine whether teaching experience influences students' evaluations of professors I used the number of ratings as a proxy for experience and average rating as a measure of teaching quality. Important to note, the number of ratings may not accurately reflect teaching experience and could be influenced by course popularity, class size, or the professor's general visibility. Therefore, this proxy introduces limits in interpreting results of the analysis. Given the distribution of the number of ratings is skewed, I used **Spearman's rank correlation** to assess the relationship between number of ratings and average rating which yielded correlation coefficient of **0.0282** with a p-value of **0.000007,** indicating a statistically significant but very weak positive association between number of ratings and average rating (Figure 2). In order to reduce noise at either extreme I filtered the data to only include professors with greater than 5 ratings and less than 150 ratings. The correlation remained very similar: **0.0275** with a p-value of **0.000012** (Figure 3). These results suggest a statistically significant but practically negligible association between number of ratings and professor ratings. Given that the number of ratings is an imperfect proxy for actual teaching experience and the correlation is close to zero, we conclude that experience has little to no meaningful influence on student evaluations of professors.



For all professors with > 5 ratings

Figure 2.

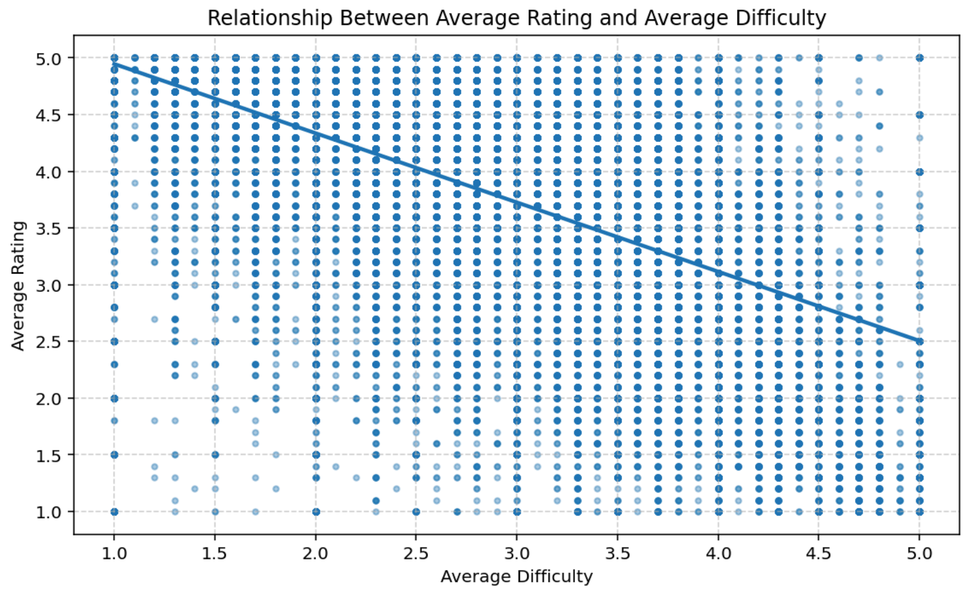
For all professors with > 5 ratings and < 150 ratings

Figure 3.

3. Analysis on the data shows a strongly statistically significant negative relationship between rating and difficulty. First, I filtered the data to include columns with no missing values rating and average difficulty. Then, I determined the **Pearson Correlation coefficient** for the data to be **-0.5368** at a **p-value effectively equal to zero** (to at least 30 decimal places) indicating that the inverse relationship is highly unlikely to have occurred by chance. Performing a **simple linear regression** on the data between average rating and average difficulty confirms this relationship. Based on our linear regression model, about 29% of the variance in average rating is explained by average difficulty (given the **R-squared value of 0.288**) and the coefficient of **-0.6103** indicates that for each 1-point increase in difficulty, the average rating decreases by about 0.61at a **p-value of less than 0.00001** meaning this negative relationship is highly significant (Figure 4).

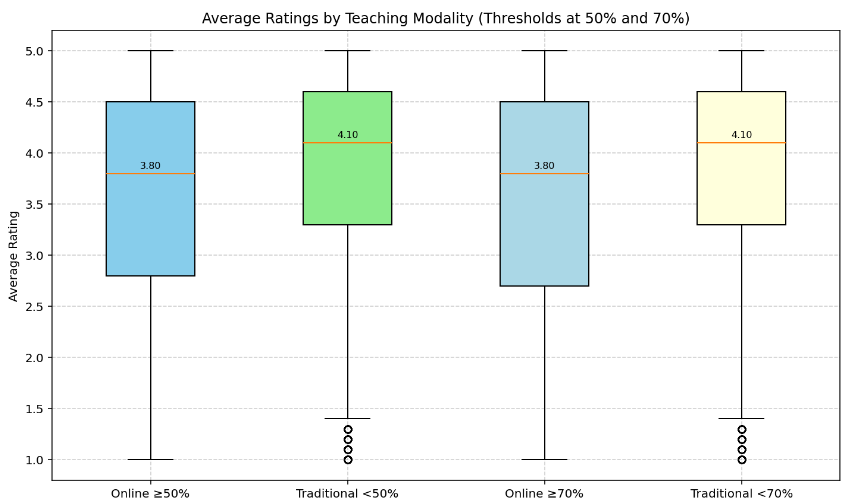
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.288
Model:                            OLS   Adj. R-squared:                  0.288
Method:                 Least Squares   F-statistic:                 2.833e+04
Date:                Tue, 29 Apr 2025   Prob (F-statistic):               0.00
Time:                        16:29:19   Log-Likelihood:                -95798.
No. Observations:               70004   AIC:                         1.916e+05
Df Residuals:                   70002   BIC:                         1.916e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.5564      0.011    505.536      0.000       5.535       5.578
x1            -0.6103      0.004   -168.325      0.000      -0.617      -0.603
==============================================================================
Omnibus:                     4620.084   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5722.777
Skew:                          -0.641   Prob(JB):                         0.00
Kurtosis:                       3.564   Cond. No.                         10.2
==============================================================================
```

Figure 4. *Simple Linear Regression*



4. To examine whether professors teaching online courses receive different average ratings than professors teaching in-person, I conducted a significant test comparing average ratings across two different thresholds for “online-heavy” courses. Using the binary variable “online,” I found the proportion of online classes taught by dividing the count of online ratings by the total number of ratings and split the data based on two thresholds - **over 50% online ratings coming from online courses** and over **70% online ratings coming from online courses**. I used a **Mann-Whitney U test** to compare the two samples and determined that at both thresholds the **median rating for online-heavy professors is 3.80** and **4.10 for in-person professors** at a p-value of about **2.71e-11** for the 50% threshold with an effect size of **r = 0.042,** and a p-value of about **3.21e-06** with an effect size of **r = 0.029** for the 70% threshold. Although the difference between average ratings in online and in-person classes is statistically significant, the effect size for both thresholds indicates that the magnitude of difference is extremely small. This also suggests that despite a statistically significant difference in median ratings, the practical difference is minimal and is likely due to the large sample size. In order to further understand the relationship, I performed a linear regression on the two variables (Figure 5), namely proportion of online ratings, and average rating. My simple linear regression model had an **R-squared value of 0.003**, indicating that only 0.3% of variation in proportion of online ratings is explained by average rating. This prompted me to perform a multiple regression analysis (Figure 6) to control for other factors such as difficulty, pepper-status, number of ratings, and proportion of students who would take the class again. My multiple regression model had an **R-squared value of 0.809**, indicating about 81% of variation in the given variables explains average rating, a much higher R-squared value than the simple linear regression model from before. In this

multivariate model, the coefficient for proportion of online ratings (x5) was **-0.0004** at a **p-value of 0.799** indicating that **if we control for other variables, the proportion of online ratings has no statistically significant difference on average rating**. Therefore, although the Mann-Whitney U test suggested a statistically significant difference in median ratings for online heavy courses at two different thresholds, due to the extremely small effect size and regression controlling for other variables, there is no evidence of a practically or statistically significant difference when controlling for other variables between average ratings of professors who teach in an online-heavy vs. traditional setting.




```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     77.19
Date:                Wed, 30 Apr 2025   Prob (F-statistic):           1.65e-18
Time:                        10:51:30   Log-Likelihood:                -34572.
No. Observations:               25368   AIC:                         6.915e+04
Df Residuals:                   25366   BIC:                         6.916e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.8594      0.006    612.503      0.000       3.847       3.872
x1            -0.3270      0.037     -8.786      0.000      -0.400      -0.254
==============================================================================
Omnibus:                     2410.031   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3174.978
Skew:                          -0.866   Prob(JB):                         0.00
Kurtosis:                       2.948   Cond. No.                         6.29
==============================================================================
```

Figure 5. *Simple Linear Regression*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.809
Method:                 Least Squares   F-statistic:                 1.033e+04
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        10:51:30   Log-Likelihood:                -5143.1
No. Observations:               12160   AIC:                         1.030e+04
Df Residuals:                   12154   BIC:                         1.034e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.5259      0.025    103.083      0.000       2.478       2.574
x1            -0.1997      0.005    -37.604      0.000      -0.210      -0.189
x2             0.2029      0.008     26.972      0.000       0.188       0.218
x3            -0.0002      0.000     -0.924      0.355      -0.001       0.000
x4             2.4954      0.017    147.358      0.000       2.462       2.529
x5            -0.0004      0.002     -0.255      0.799      -0.004       0.003
==============================================================================
```

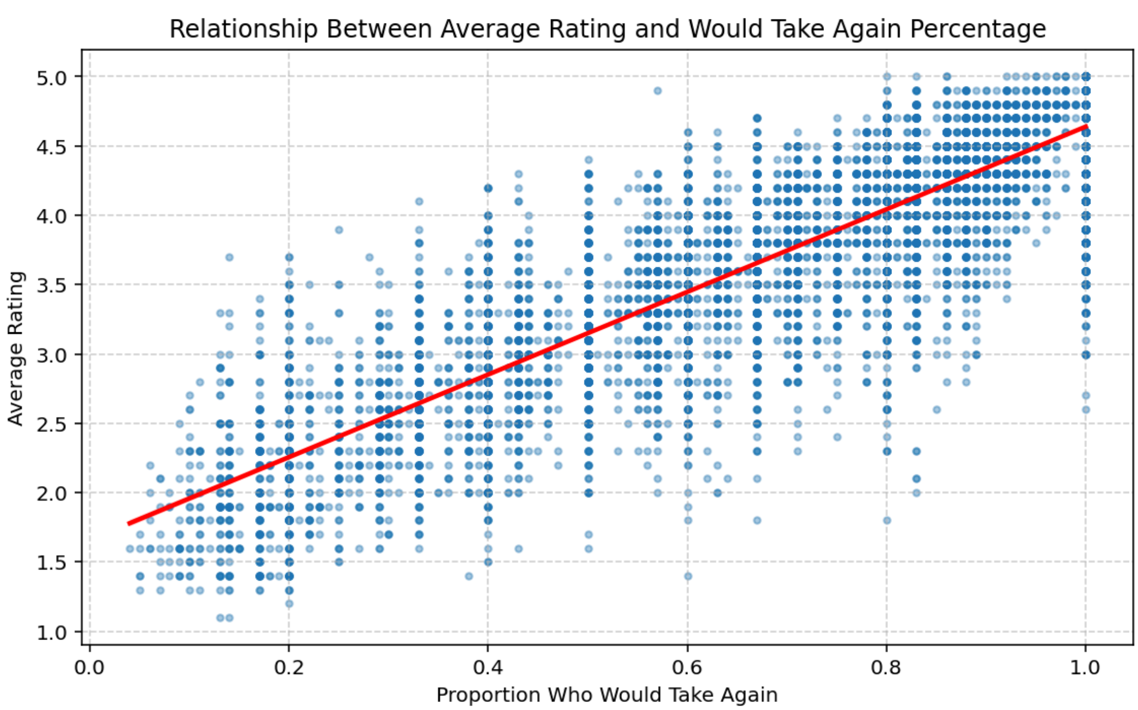
Figure 6. *Multivariate Linear Regression*

5. First, I filtered the data to only include data from professors with over five ratings, and no missing values for average ratings and if the student would take the class again. I divided the value for if the student would take the class again by 100 to obtain a proportion of students who would take the class again in terms of a percentage. Then, I calculated the **Spearman correlation coefficient** to be **0.8522** with a **p-value of practically zero** (accurate at least to 30 decimal places). I also calculated the **Pearson correlation** which unlike the Spearman correlation coefficient assumes linearity, and got a correlation coefficient of **0.8804 with a similarly small, significant p-value**. Both results demonstrate a very strong, positive correlation between average rating and proportion of students who would take the class again. To further understand the relationship, I fit a

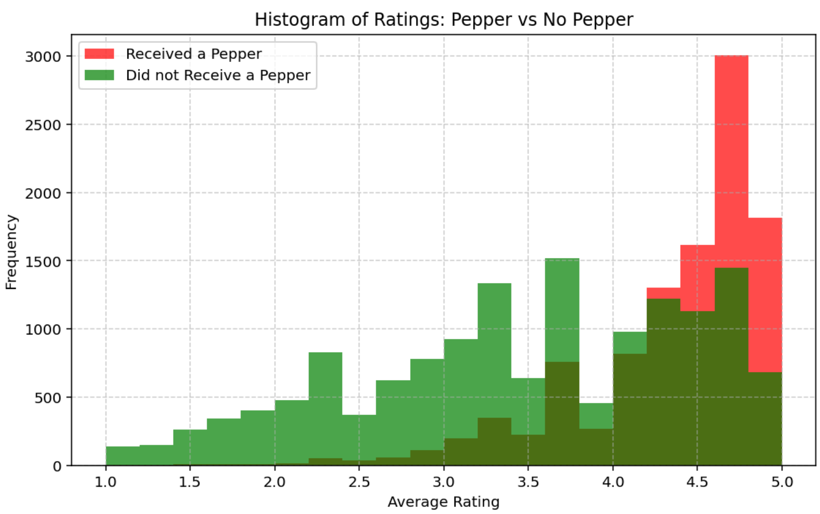
simple linear regression model to the data (Figure 7) and found that based on our model, for every 100 percentage point increase in the proportion of students who would retake the course, the average rating of the professor is predicted to increase by about **2.98 points**. Although this slope is difficult to contextualize and practically the percentage of students is bounded between 0 and 100 and average rating between 0 and 5, these results contribute to the overall conclusion that there is a **strong positive relationship** between willingness to take the class again and average rating. Lastly, the R-squared value of both the Pearson correlation coefficient and the regression model (the same in this case because our regression is with one variable) is 0.7750 which suggests that based on our model about 77.5% of variation in average rating can be explained by willingness to take the class again.

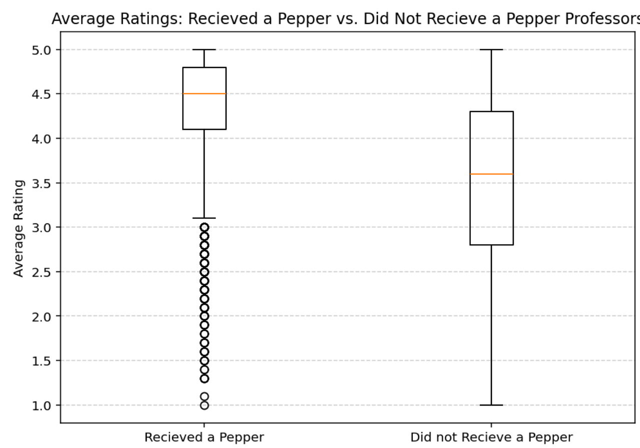
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.775
Model:                            OLS   Adj. R-squared:                  0.775
Method:                 Least Squares   F-statistic:                 4.188e+04
Date:                Tue, 29 Apr 2025   Prob (F-statistic):               0.00
Time:                        17:17:51   Log-Likelihood:                -6153.7
No. Observations:               12160   AIC:                         1.231e+04
Df Residuals:                   12158   BIC:                         1.233e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.6600      0.012    141.849      0.000       1.637       1.683
x1             2.9786      0.015    204.653      0.000       2.950       3.007
==============================================================================
Omnibus:                     1031.194   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1832.106
Skew:                          -0.607   Prob(JB):                         0.00
Kurtosis:                       4.463   Cond. No.                         6.43
==============================================================================
```

Figure 7. *Simple Linear Regression*



6. Based on a **Mann-Whitney U test**, there is a statistically significant difference in ratings between professors who received a pepper and those who did not. Based on the data, professors who received a pepper have a **median rating of 4.50** and professors who did not receive a pepper have a median rating of **3.60** at a **p value < 1e14,** demonstrating that the difference in rating based on pepper-status is highly unlikely due to chance alone. Furthermore, I calculated the effect size by calculating the value for **Cohen's d** and got a value of **d ≈ 1.06**, suggesting a very large effect size, meaning that there is a substantial difference in average rating between professors who received a "pepper" and those who did not.





7. I ran a **Simple Linear Regression** model to predict average rating from average difficulty only (Figure 8). My regression model had an **R-squared value of 0.383,** indicating that about 38% of variability in average rating can be explained by average difficulty based on this model. The model suggests that for every one-point increase in difficulty, the predicted average rating decreases by about **0.73** points at a **p-value < 0.0001**, demonstrating that this relationship is statistically significant at our alpha level of 0.005 and highly unlikely due to chance. To assess whether a linear model was appropriate, I constructed a residual plot which showed a visible linear trend (Figure 9), which suggests the model may not fully capture the structure of the relationship between difficulty and rating. Despite a moderate fit based on the R-squared value, the pattern in residuals suggests other unaccounted for factors or interactions with difficulty that may influence ratings. Furthermore, it is possible that the relationship is not in fact linear, especially considering that average difficulty and average rating may be more concentrated at extremes (very difficult or very easy and very good or very bad).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.383
Model:                            OLS   Adj. R-squared:                  0.383
Method:                 Least Squares   F-statistic:                 1.576e+04
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        11:33:37   Log-Likelihood:                -28481.
No. Observations:               25368   AIC:                         5.697e+04
Df Residuals:                   25366   BIC:                         5.698e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.9692      0.018    339.434      0.000       5.935       6.004
x1            -0.7271      0.006   -125.534      0.000      -0.738      -0.716
==============================================================================
Omnibus:                     1035.554   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1173.714
Skew:                          -0.503   Prob(JB):                    1.35e-255
Kurtosis:                       3.317   Cond. No.                         12.6
==============================================================================
```

Figure 8. *Simple Linear Regression*

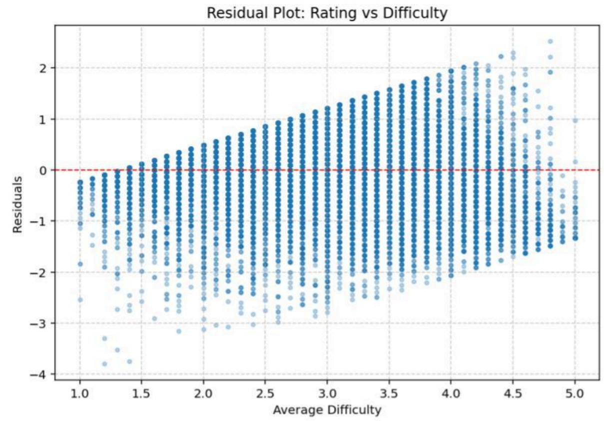
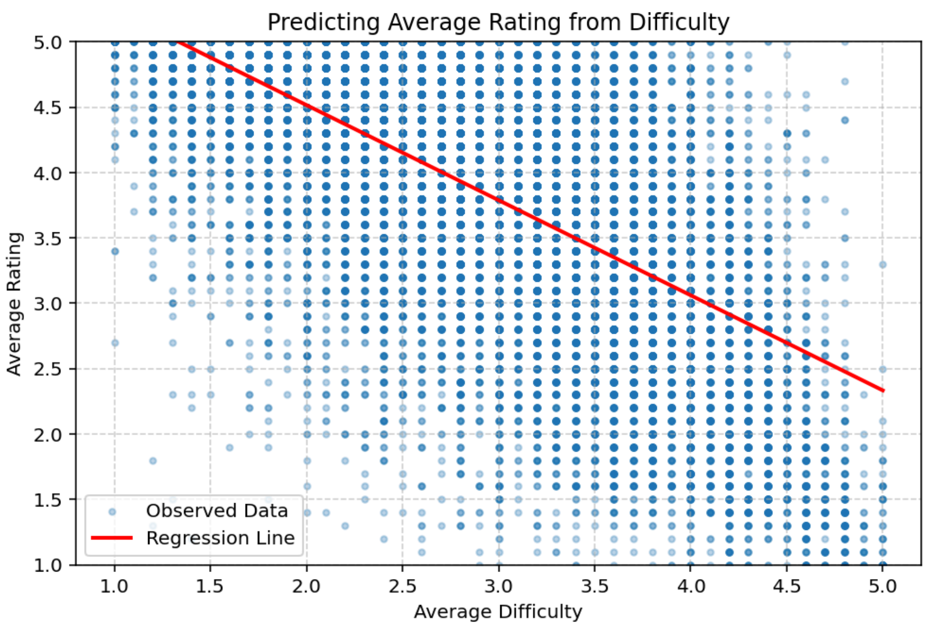


Figure 9. *Residual Plot*



8. To predict average professor rating from all of the available factors, I built a **multiple linear regression model** (Figure 10) using the variables: difficulty, number of ratings, “would take again” percentage, pepper status, online rating count, and gender. Given male and female are inversely proportional to each other and to prevent unnecessary collinearity, I only included one variable for gender, in this case male. After removing missing values and applying the threshold of professors with at least five ratings, the model achieved an **R-squared value of about 0.81** which indicates that 81% of the variability in average rating can be explained by the previously given predictors. The **RMSE of the model was about 0.37**, meaning on average the predictions from the

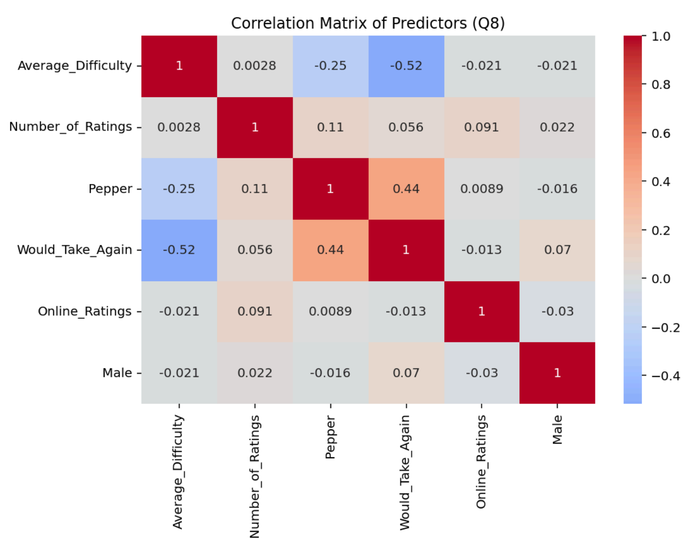
model are off by about 0.37 rating points. To help with interpretability and compare effect sizes across predictors, I ran the **same regressing using z-score predictors**, while keeping the outcome variable in the original scale (Figure 11). In this standardized model, “would take again” had a strong positive rating with average rating, pepper status a moderate positive relationship, and average difficulty a moderate negative relationship (all at statistically **significant p-values < 0.005**). The number of ratings and number of online ratings have no statistically significant effect on average ratings given both have p-values well above our alpha level of 0.005 (namely **0.288 and 0.929**). Gender has a small but statistically significant difference with being male having a small positive correlation with average rating. To address potential collinearity among predictors, I conducted a **Principal Component Analysis (PCA)** on the standardized predictors. Using the **Kaiser Criteria**, I concluded that only **three components (with eigenvalues > 1)** could be used to capture the majority of variance in average rating (Figure 12). Using only these three components, I ran another linear regression to generate a model with an **R-squared value of about 0.73** and **RMSE of about 0.44** (Figure 13). Although the PCA analysis reduced some dimensions, simplifying the structure of the model, and reduced collinearity concerns, it produced a model with less accuracy in terms of both R-squared and RMSE. Lastly, to validate the multiple linear regression with all predictor variables further, I plotted the residuals (Figure 14) and concluded that there is not noticeably problematic relationship among residuals.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.810
Model:                            OLS   Adj. R-squared:                  0.810
Method:                 Least Squares   F-statistic:                     8632.
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        12:31:47   Log-Likelihood:                -5128.1
No. Observations:               12160   AIC:                         1.027e+04
Df Residuals:                   12153   BIC:                         1.032e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9361      0.003   1176.248      0.000       3.930       3.943
x1            -0.1478      0.004    -37.725      0.000      -0.155      -0.140
x2            -0.0036      0.003     -1.062      0.288      -0.010       0.003
x3             0.1023      0.004     27.262      0.000       0.095       0.110
x4             0.6222      0.004    146.584      0.000       0.614       0.631
x5            -0.0003      0.003     -0.089      0.929      -0.007       0.006
x6             0.0184      0.003      5.483      0.000       0.012       0.025
==============================================================================
```

Figure 10. *Multivariate Regression*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.810
Model:                            OLS   Adj. R-squared:                  0.810
Method:                 Least Squares   F-statistic:                     8632.
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        12:31:46   Log-Likelihood:                -5128.1
No. Observations:               12160   AIC:                         1.027e+04
Df Residuals:                   12153   BIC:                         1.032e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.5162      0.025    102.539      0.000       2.468       2.564
x1            -0.2001      0.005    -37.725      0.000      -0.211      -0.190
x2            -0.0002      0.000     -1.062      0.288      -0.001       0.000
x3             0.2051      0.008     27.262      0.000       0.190       0.220
x4             2.4877      0.017    146.584      0.000       2.454       2.521
x5            -0.0002      0.002     -0.089      0.929      -0.004       0.003
x6             0.0372      0.007      5.483      0.000       0.024       0.051
==============================================================================
```

Figure 11. *Multivariate Regression with Z-scored coefficients*



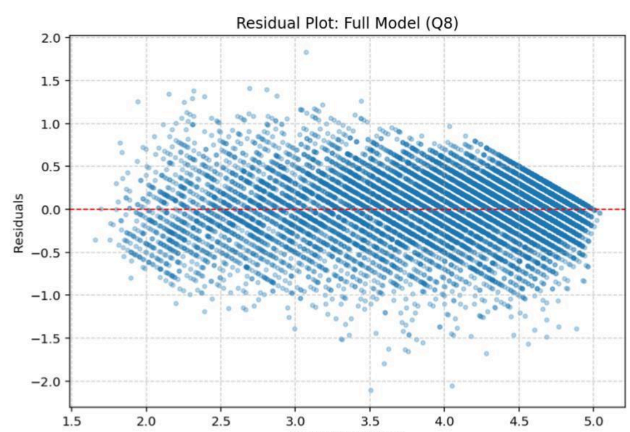


Figure 14. *Residual Plot for Regression Model*

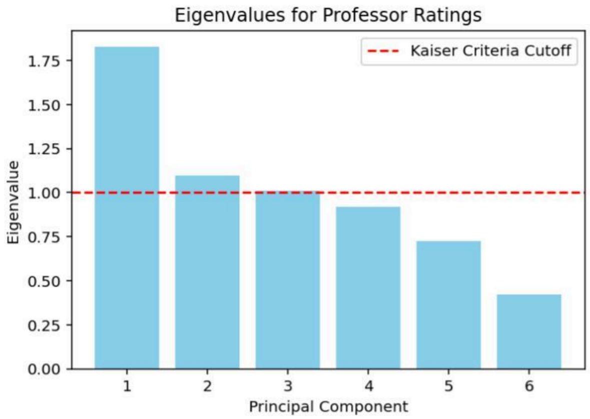
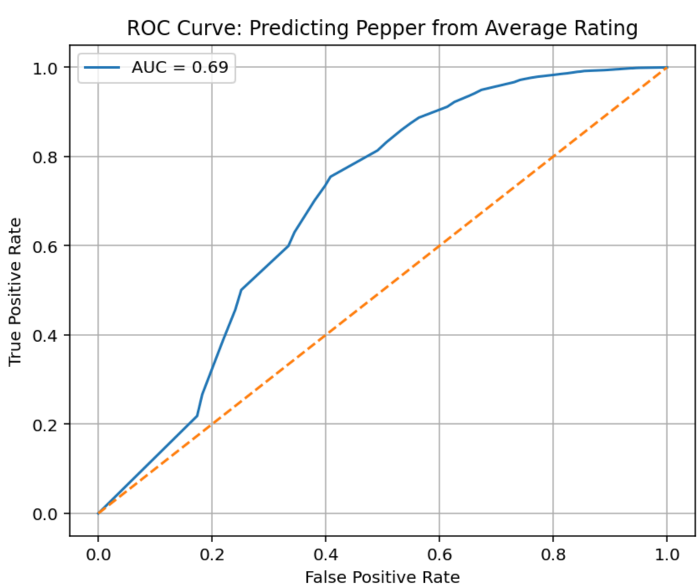


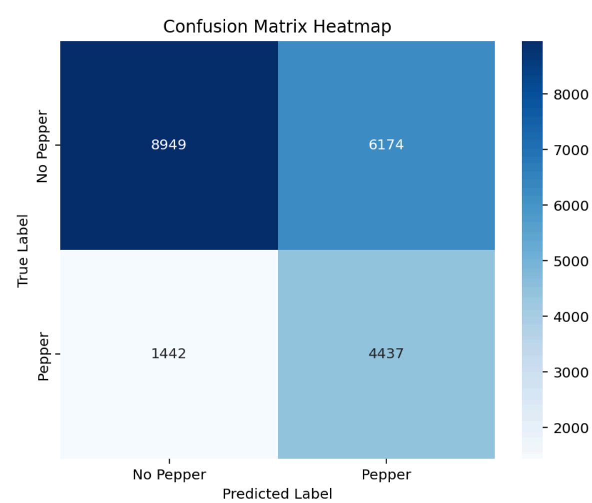
Figure 12. *PCA with Kaiser Criteria*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.730
Model:                            OLS   Adj. R-squared:                  0.730
Method:                 Least Squares   F-statistic:                 1.097e+04
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        14:07:24   Log-Likelihood:                -7255.8
No. Observations:               12160   AIC:                         1.452e+04
Df Residuals:                   12156   BIC:                         1.455e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9361      0.004    987.558      0.000       3.928       3.944
x1             0.5323      0.003    180.501      0.000       0.526       0.538
x2            -0.0700      0.004    -18.379      0.000      -0.077      -0.063
x3            -0.0021      0.004     -0.523      0.601      -0.010       0.006
==============================================================================
```

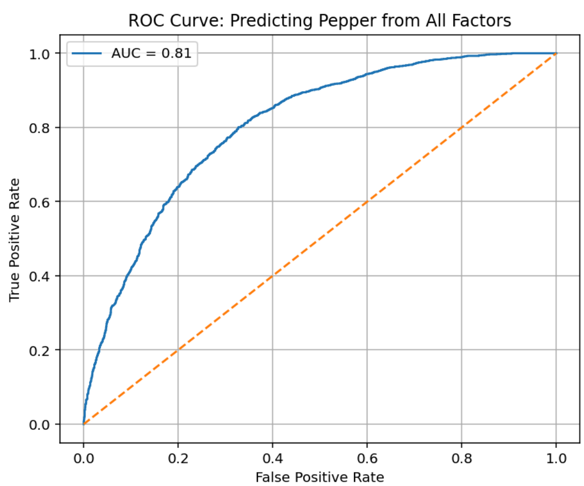
Figure 13. *Regression with 3 Principal Components determined by Kaiser Criteria*

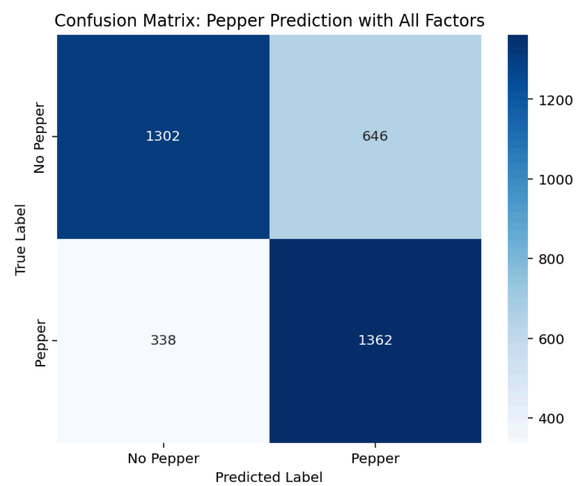
9. In order to predict whether a professor receives a "pepper" from average rating only, I built a **binary classification model using logistic regression**. In order to combat class imbalances, as more professors do not receive a pepper than those who do (50,408 without a pepper and 19,596 with a pepper), I applied class balancing by setting **class_weight = 'balanced'** in my logistic regression model. After removing incomplete data, I applied a train/test split on the data so that **70% of the data would be used for training and 30% for testing**. I used stratification based on pepper status so that each set would contain the same ratio of professors with and without peppers to preserve class distribution. Finally, I evaluated the model by finding the area under the ROC curve (AUROC) which was **0.69**, demonstrating the model has a moderate ability to determine pepper status based on average rating. I constructed a confusion matrix which suggested the model performs fairly well in terms of recall, but poorly in terms of precision. The model correctly identifies about 75.5% of professors who actually have a pepper, but when the model predicts a professor has a pepper it is only correct 42% of the time. In other words, the model has a lot of false positives (top right corner of confusion matrix).

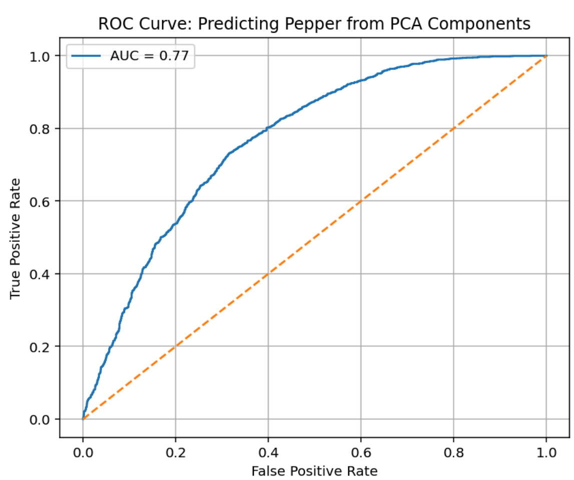


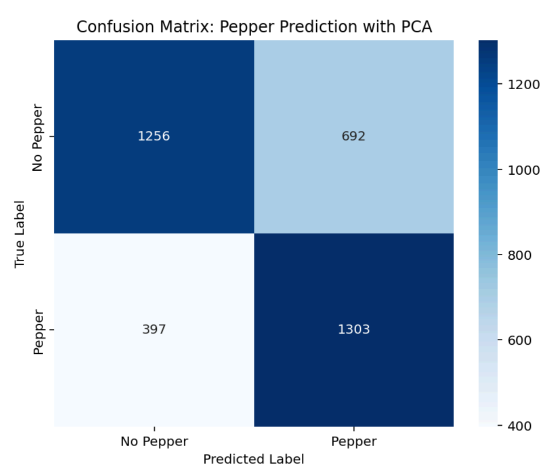


10. To predict whether a professor receives a pepper, I built **two logistic classification models** using all available predictors: average rating, average difficulty, number of ratings, would take again percentage, online rating count, and gender (male). I addressed class imbalances by using a stratified train/test split and setting **class_weight = 'balanced'** to make sure professors with and without a pepper were equally considered despite unequal frequencies in the data. First, I developed a model using all the original predictors without any dimensionality reduction and achieved an **AUROC of 0.81,** indicating a strong performance of the classification model. The confusion matrix showed a strong performance identifying true positives and true negatives, with 338 false negatives and 646 false positives. I also developed a second model to address collinearity among predictors by applying a **Principal Component Analysis** and found 3 principal components based on the **Kaiser Criteria** as referenced in question 8. Using these 3 principal components, I trained a second logistic regression model with an **AUROC of 0.77**, slightly lower than the previous model. The model performed similarly in identifying true positives, and true negatives, with slightly higher false positives and false negatives (692 and 397). While both models perform similarly well, the model using all original predictors outperformed the PCA-based model despite mitigating concerns over multicollinearity and reduced dimensionality.









Extra Credit:

For an additional analysis on the provided RateMyProfessor Data I wanted to examine **whether professors in liberal arts fields receive higher ratings than professors in STEM fields.** First, I categorized majors into STEM and Liberal Arts by keywords:

```
stem_keywords = ['Engineering', 'Biology', 'Math', 'Computer', 'Physics', 'Chemistry', 'Statistics', 'Data']
liberal_arts_keywords = ['History', 'Philosophy', 'English', 'Sociology', 'Political Science', 'Art', 'Music', 'Theater', 'Writing']
```

I used a **Mann-Whitney U test** to assess the difference in median ratings between professors in each category. I found that based on my analysis the **median rating of professors in Liberal Arts subjects is 4.30** and the **median rating for professors in STEM subjects is 3.90** and this difference is statistically significant at a **p-value of < 0.00001**, well below our threshold of 0.005. However, in order to also control for average difficulty, I also performed a **multiple regression** (Figure 15). My model determined that being a professor in a liberal arts field is associated with a **0.08**-point increase in average rating even when controlling for average difficulty at a **p-value of < 0.001**, and below our threshold of 0.005. Important to note, difficulty has a strong negative effect on ratings across both groups and although there seems to be statistically significant tendency for higher ratings among professors in liberal arts fields, the coefficient of 0.08 is very small and suggests that most of the differences in ratings across fields is better explained by differences in perceived difficulty rather than field of study alone.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Average_Rating   R-squared:                       0.301
Model:                            OLS   Adj. R-squared:                  0.301
Method:                 Least Squares   F-statistic:                     7167.
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        17:09:42   Log-Likelihood:                -45195.
No. Observations:               33347   AIC:                         9.040e+04
Df Residuals:                   33344   BIC:                         9.042e+04
Df Model:                           2
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  5.5671      0.018    301.821      0.000       5.531       5.603
Liberal_Arts_Yes       0.0789      0.010      7.534      0.000       0.058       0.099
Average_Difficulty    -0.6262      0.005   -115.758      0.000      -0.637      -0.616
======================================================================================
```

Figure 16. *Multivariate Regression Model*