Ken Li
Pascal Wallisch
Principles of Data Science
30 April, 2025

Capstone Project

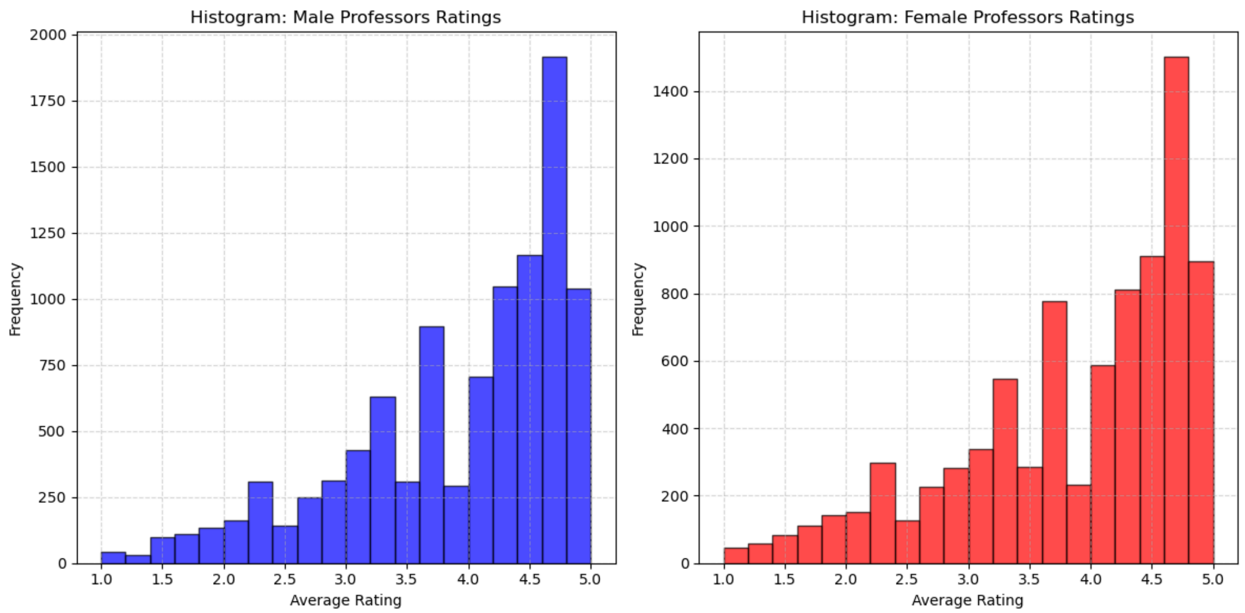
Data Preprocessing

Initially, I assigned appropriate column names. Then I removed the first row in each file. After that, I converted all numeric columns to their proper data types and then filtered the dataset to retain only rows with valid entries in the key variables: average rating, average difficulty, and number of ratings. To maintain the reliability of student evaluations, I further dropped any professor with fewer than 5 total ratings. Finally, I re-aligned the qualitative dataframe to match the filtered numeric dataframe by index, ensuring consistency across datasets.

Q1

Is there evidence of a pro-male gender bias in the dataset?

First, I formulate hypotheses for our significance test:
H_0: The distribution of ratings is the same for male and female professors
H_1: The distribution of ratings is different between male and female professors
Before choosing an appropriate statistical test, I examined the distribution of ratings for male and female professors separately:



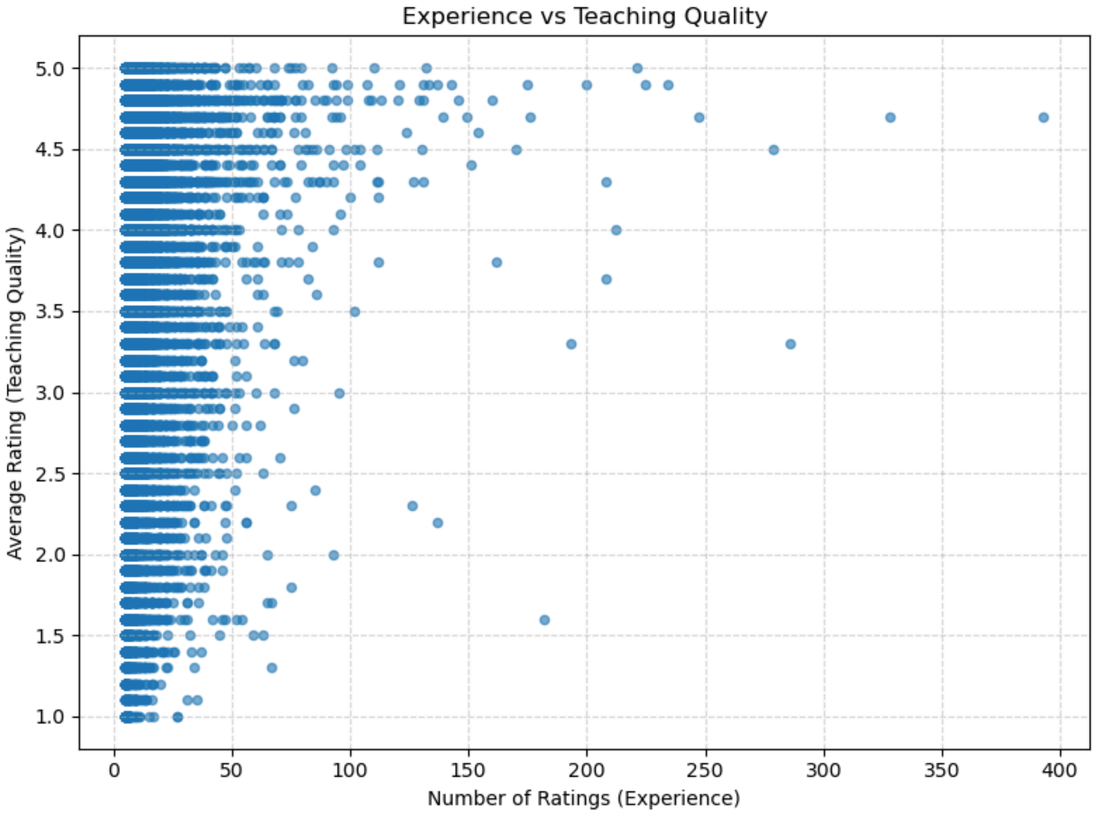
The histograms above show that both male and female professors' ratings are not normal; instead, they are skewed to the right. Thus, we can use the Mann-Whitney U Test to compare male and female professors' ratings. The test reveals a p-value of **4.905*10^{-4}**, which is far below our threshold (α = 0.005). Hence, there is strong evidence to reject the null hypothesis, indicating that male and female professors have different average ratings.

In addition, to quantify the difference, I performed the bootstrapping by resampling male and female ratings 1000 times and computing the difference in medians for each resample. I plotted the distribution of median differences. The 99.5% confidence interval is **(0.000, 0.100)**, which is entirely at or above zero. This confirms that male professors consistently receive ratings that are equal or higher than female professors.

However, the difference in median ratings is as small as **0.100**, suggesting that the difference, while statistically significant, is not as practically meaningful.

Q2

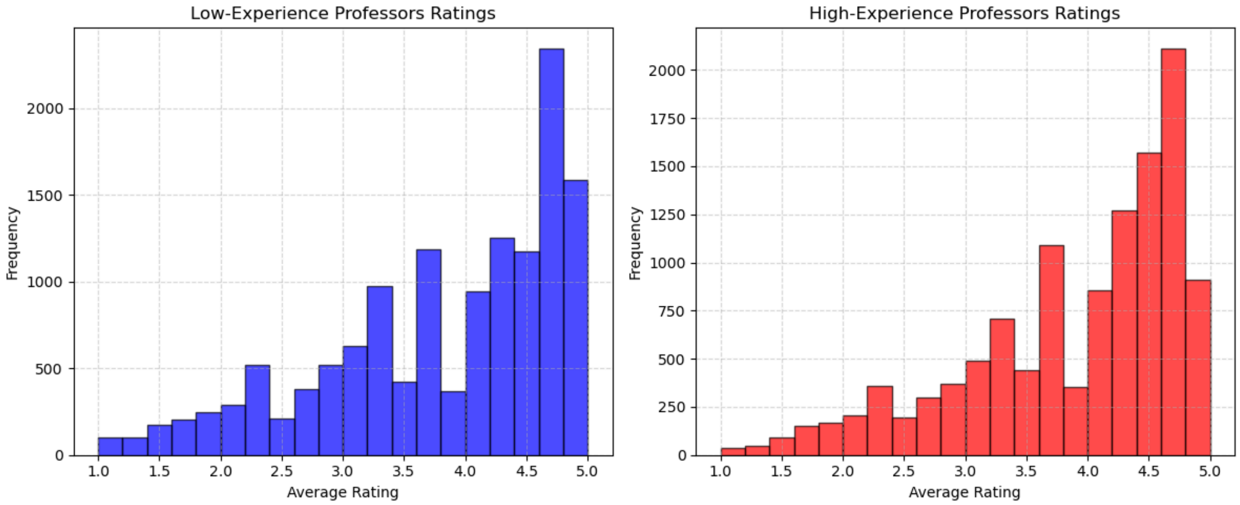
Is there an effect of experience on the quality of teaching？



For intuition, we take a look at the scatterplot above. Since it is hard to observe any kind of linear relationship, I took a quick Spearman Correlation check. The result indicates the same as the coefficient is as small as **0.028**, meaning that the correlation is extremely weak positive monotionic relationship. To verify, I conduct a significance test:

H_0: Experience (number of ratings) has no effect on teaching quality (average rating).

H_1: Experience (number of ratings) has an effect on teaching quality (average rating).

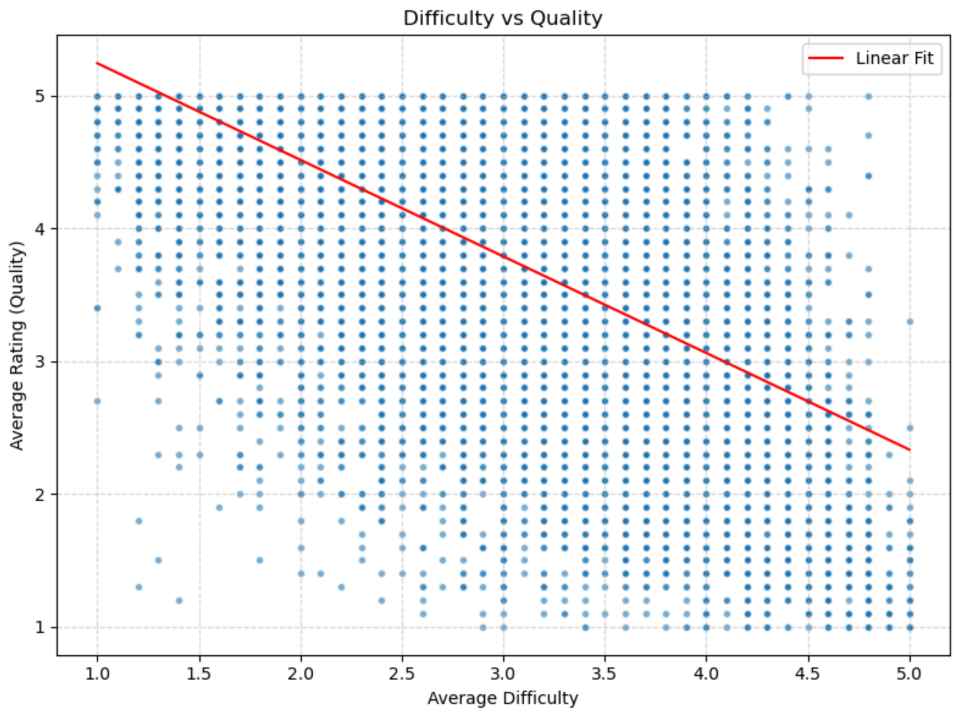


From the histograms above, they both appear to be skewed to the left and are not normal. Thus, I use the Mann-Whitney U Test, and the test gives a **p-value** of **$4.905*10^{-8}$**, which is much less than our threshold of 0.005. As a result, we reject the null hypothesis and conclude that there is a statistically significant difference between low-experience and high-experience professors' ratings.

However, the effect size is small. I obtained the **Cohen's d** of **0.084**. Therefore, we can only conclude that experience has a statistically significant but practically negligible effect on teaching quality. I think the extremely small p-value could be affected by the large sample size, as even small differences can result in very small p-values.

Q3

What is the relationship between average rating and average difficulty？



From the scatterplot above, it shows a clear negative trend. I computed the Pearson correlation coefficient of **-0.619** (between -0.3 and -0.7), suggesting that there is a moderate negative linear correlation. Also, I got a p-value of **0.000**, indicating high statistical significance.

In conclusion, there is a statistically significant, moderately strong negative relationship between Average Difficulty and Average Rating – as courses become more difficult, students' ratings of professors tend to decrease.

(Just to be safe here, I also computed the **Spearman correlation coefficient**, which is **-0.602,** indicating moderately strong negative monotonic relationship between average rating and average difficulty. This confirms that as difficulty increases, ratings tend to decrease**)**

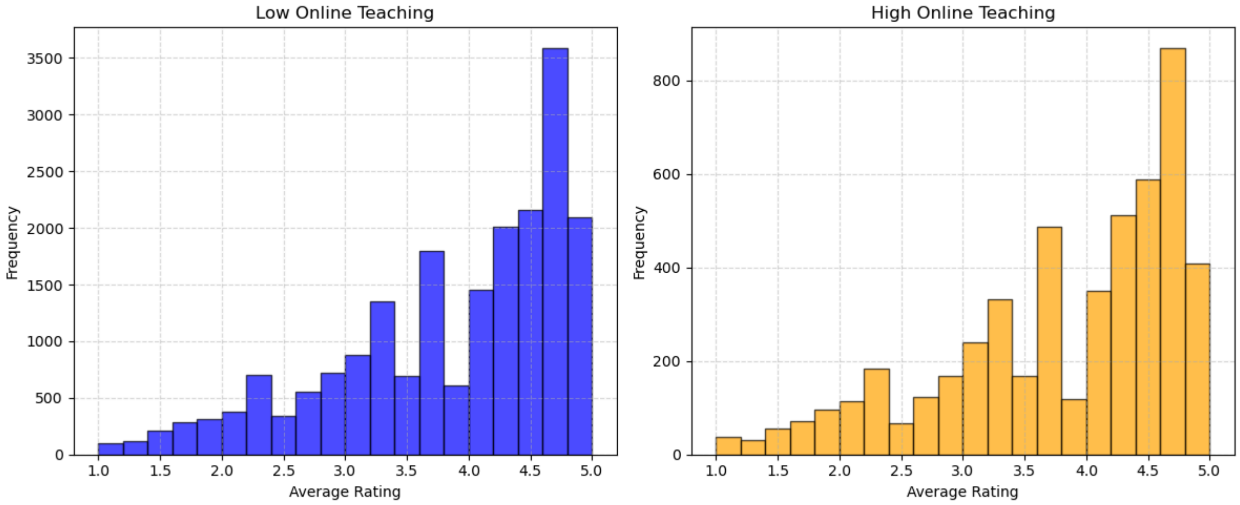
Q4

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

To split the data, I computed the proportion of each professor's ratings that came from online courses and used the median as the threshold to define two groups (one > median and one <= median).

Let's then conduct a significance test to answer the question:
H_0: There is no difference in the distribution of average ratings between professors who teach more online and those who teach less online.

H_1: There is a difference in the distribution of average ratings between professors who teach more online and those who teach less online.

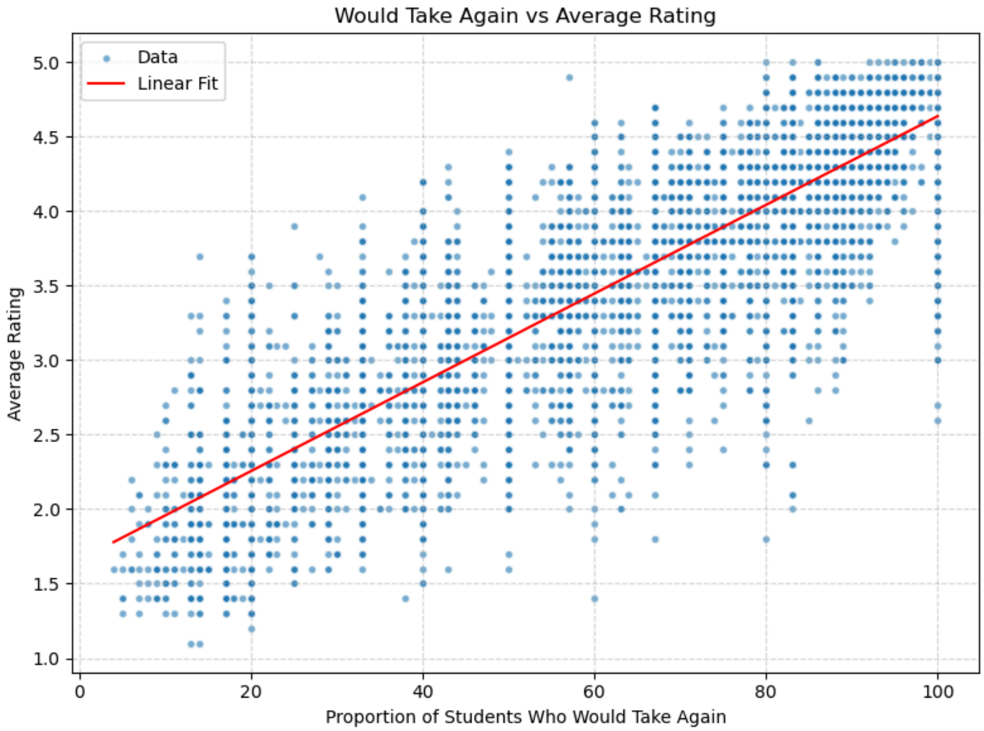


From the histograms above, they both appear to be skewed to the left and are not normally distributed. Hence, I used the Mann-Whitney U Test, and the test gives a **p-value** of **$3.313*10^{-3}$**, which is smaller than our threshold (α = 0.005). Consequently, we reject the null hypothesis and conclude that the distribution of average ratings between professors who teach more online and those who teach less online is different. However, similar to **Q1**, the effect size is very small. I computed the **Cohen's d,** and I got **0.043**. Therefore, I conclude that the difference in average ratings between high and low online teaching groups is statistically significant but not practically.

Q5

What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?

For this question, I handled missing values through dropping all rows with NaN/inf values in either "would_take_again" column or "avg_rating" columns. Besides that, I used the same approach from Q3. First, I draw a scatter plot below:



There seems to be a strong positive linear correlation between the two variables. Again, we use two correlation metrics to verify:

1. A **Pearson coefficient** of **0.880**, showing a strong positive linear relationship.
2. A **Spearman coefficient** of **0.852**, showing a strong positive monotonic relationship.

In both cases, the **p-value** is **0.000**, indicating that there is a strong and statistically significant positive relationship between "**Average Rating**" and "**Proportion of Students Who Would Take Again**". This suggests that professors with higher ratings tend to have a higher proportion of students willing to retake their classes.

Q6

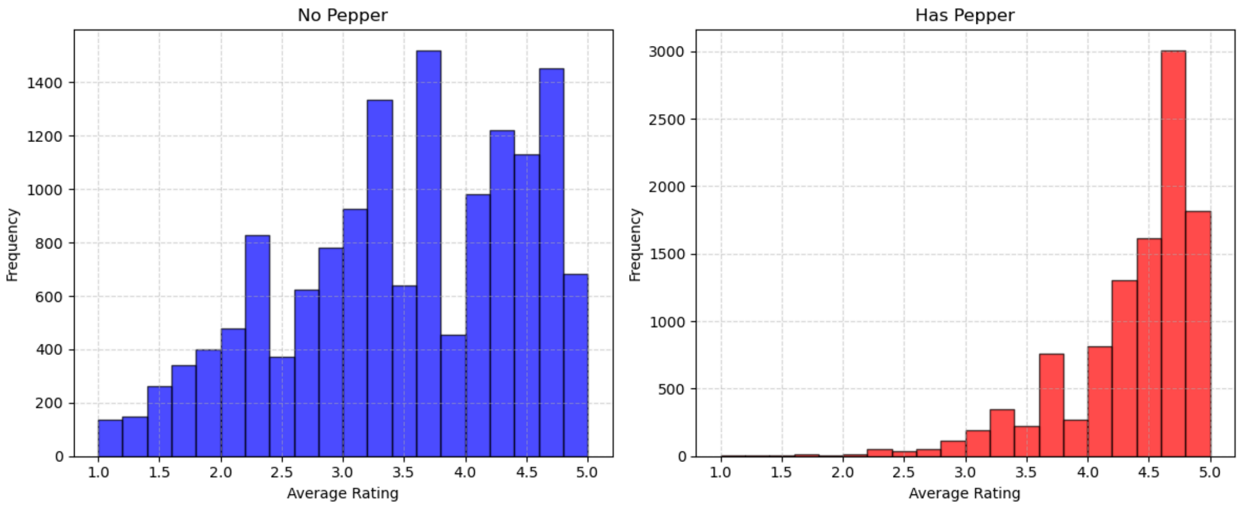
Do professors who are "hot" receive higher ratings than those who are not?

We conduct a significance test with the following hypotheses:

H_0: Professors who are "hot" have the same average ratings as those who are not.

H_1: Professors who are "hot" have the different average ratings as those who are not.

First, we split the dataset into two groups – one group of professors who are "hot" and the one who are not "hot".

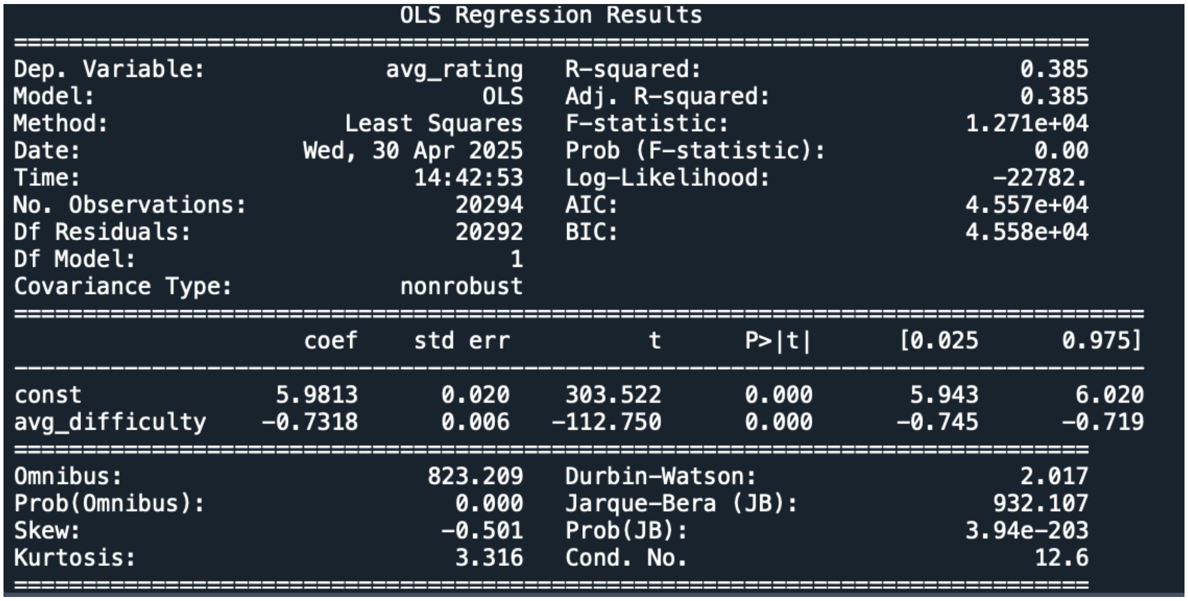


From the histograms above, both of them are not normally distributed. The left one is sort of bimodal, and the right one is heavily left skewed. Thus, I believe that a Mann-Whitney U test is appropriate. The test gives a **p-value** of **0.000**, which is extremely small (clearly below our threshold). Hence, the result gives us a very stong evidence to reject the null hypothesis and conclude that these two groups have different average ratings. To check the practical significance of the result, I also calculated the **Cohen's d**, which is **1.062**, indicating that there is a large effect. Therefore, the result is not only statistically significant, but also practically significant.

Q7

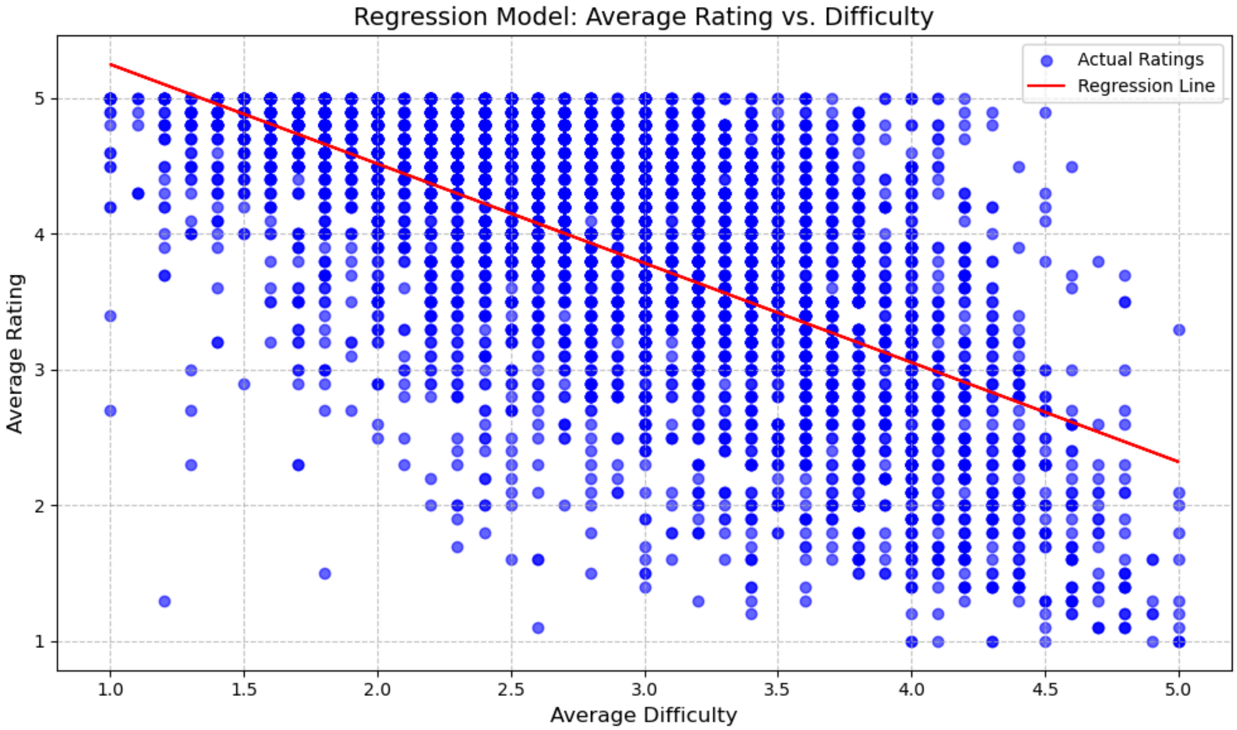
Build a regression model predicting average rating from difficulty

For this question, a linear regression model needs to be built. First, I prepared the data and train-test split. To be specific, I used a 80 training / 20 testing split due to common practice. Then, I added constants for OLS and fit the model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             avg_rating   R-squared:                       0.385
Model:                            OLS   Adj. R-squared:                  0.385
Method:                 Least Squares   F-statistic:                 1.271e+04
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        14:42:53   Log-Likelihood:                -22782.
No. Observations:               20294   AIC:                         4.557e+04
Df Residuals:                   20292   BIC:                         4.558e+04
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              5.9813      0.020    303.522      0.000       5.943       6.020
avg_difficulty    -0.7318      0.006   -112.750      0.000      -0.745      -0.719
==============================================================================
Omnibus:                      823.209   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              932.107
Skew:                          -0.501   Prob(JB):                    3.94e-203
Kurtosis:                       3.316   Cond. No.                         12.6
==============================================================================
```

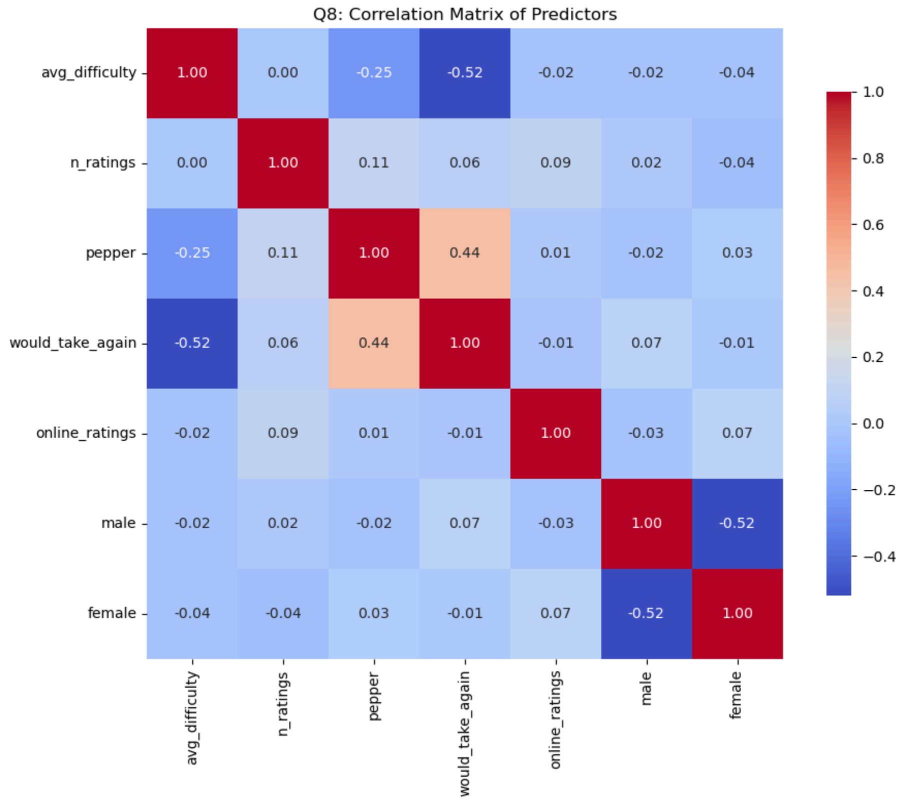
I noticed that **R^2** of **0.385**, indicating that 38.5% of variability in “Average Rating” is explained by “Average Difficulty”. While this is a moderate level of explanation, it suggests that difficulty is not the only factor that affects ratings. I obtained a **RMSE** of **0.744**, meaning that the average distance between the predicted and actual ratings is approximately 0.744 points.
To visualize, here is the regression line on a scatterplot of actual data:



Seeing the slope and intercept from the model above, we know that the equation of line is:
Average Rating = 5.9813 - 0.7318 * Average Difficulty.

Q8

Build a regression model predicting average rating from all available factor.



Before building the regression model to predict average professor ratings, we first examined the correlation matrix of all potential predictors. From the matrix, we can see there are some notable correlations that are sort of strong (> 0.50), which raises concerns about collinearity. To address this, I used ridge regression because this reduces the impact of multicollinearity by shrinking the coefficients of correlated predictors, effectively distributing the shared variance among them while not sacrificing any interpretability.

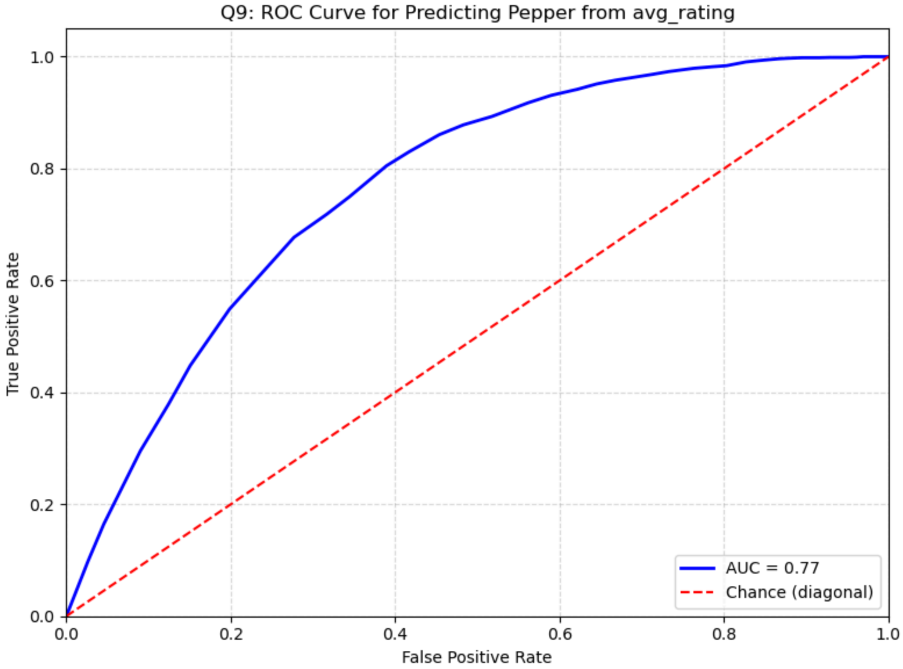
We first standardized all predictors, then split the data into training and testing sets using an 80/20 split. After training the model, we evaluated its performance on the test set. I obtained the **R^2** of **0.802** and the **RMSE** of **0.372**. The model explains 80.2% of the variability in "Average Rating" using all available predictors and its predictions deviate from the actual ratings by 0.372 points.

Compared with the model in Q7, this ridge regression model performs much better, significantly more accurate and explanatory as it captures additional variability in ratings beyond what difficulty alone can explain.

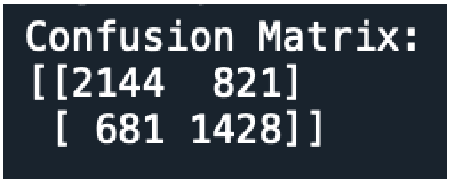
Q9

Build a classification model that predicts whether a professor receives a "pepper" from average rating only.

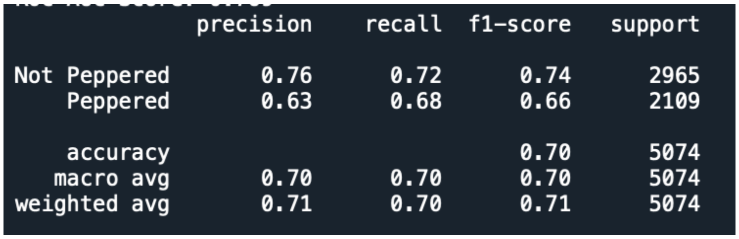
After preparing the data and doing the train-test split(same as before), built a logistic regression model to predict whether a professor receives a “pepper” icon based solely on their average rating. I obtained an **AUC-ROC score** of **0.769**, indicating the model has a reasonably strong ability to distinguish between professors who receive a “pepper” and those who do not (better than random guessing, which is AUC = 0.5).



For the ROC Curve, the blue curve shows the trade-off between TPR and FPR.



```
Confusion Matrix:
[[2144  821]
 [ 681 1428]]
```

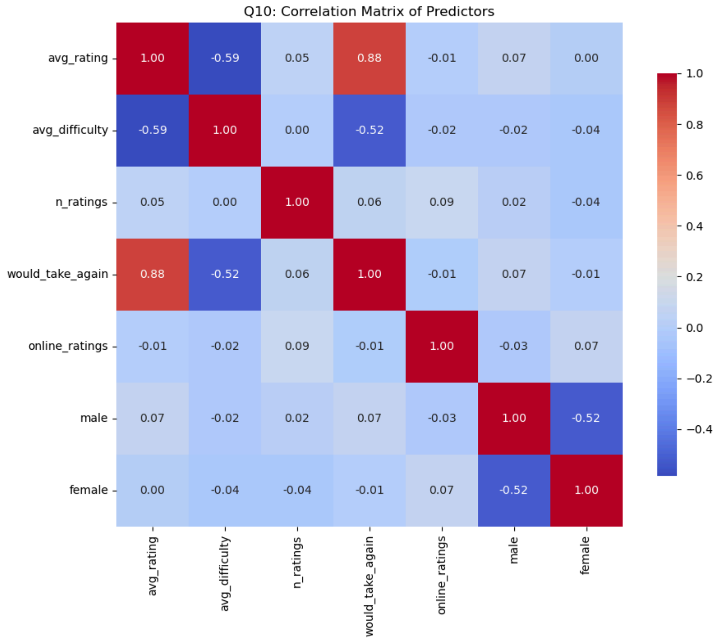


```
              precision    recall  f1-score   support

Not Peppered       0.76      0.72      0.74      2965
    Peppered       0.63      0.68      0.66      2109

    accuracy                           0.70      5074
   macro avg       0.70      0.70      0.70      5074
weighted avg       0.71      0.70      0.71      5074
```

The confusion matrix (left) and the classification report (right) serve as additional quality metrics for reference. It is worth noting that the model correctly classifies **70%** of all cases, but accuracy is not the best metric due to the imbalance between the two classes.

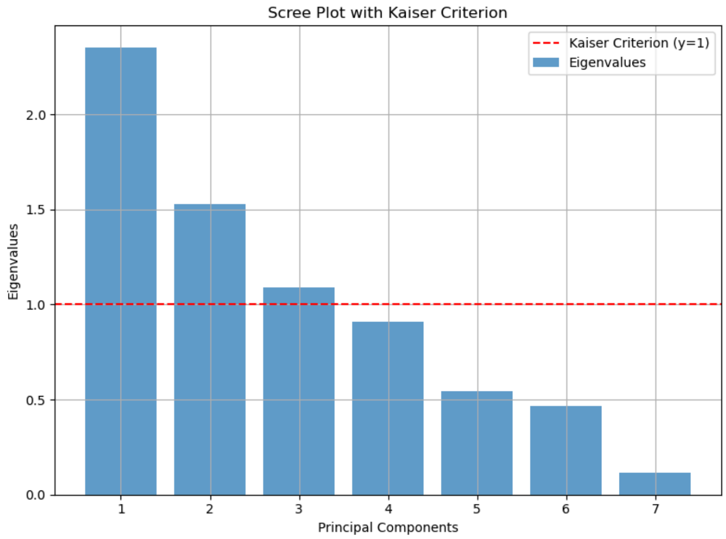
Q10

Build a classification model that predicts whether a professor receives a “pepper” from all available factors.

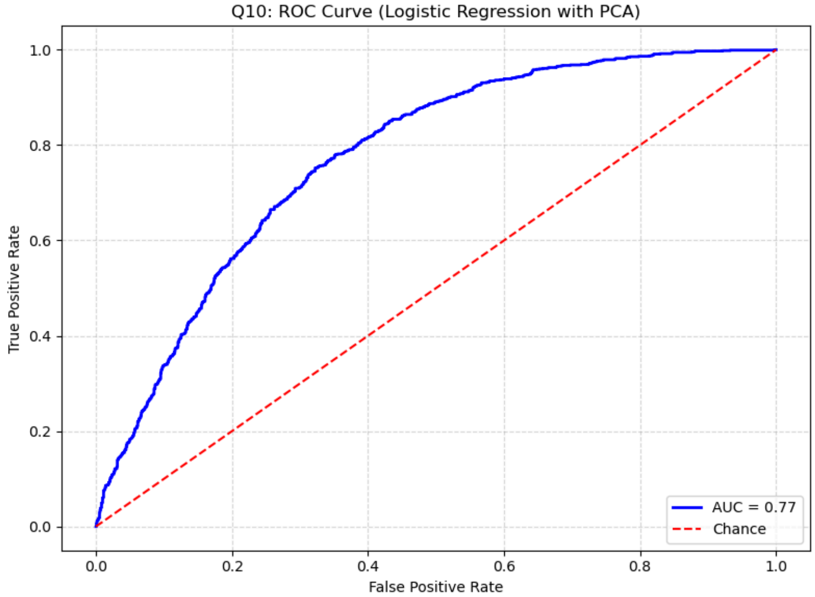


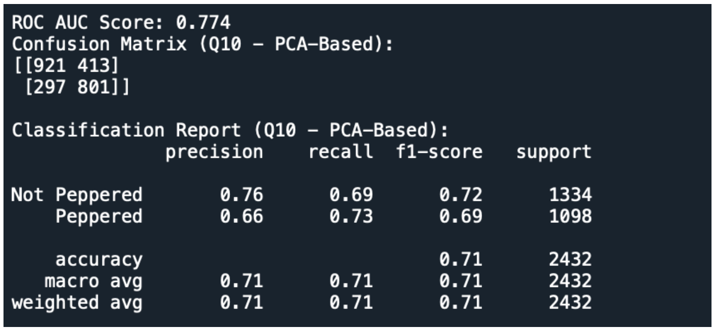
To improve the model’s performance and address multicollinearity issues, there are multiple correlation coefficients are greater than 0.50 (aside from the diagonal). Thus, I used the Principal Component Analysis (PCA) for dimensionality reduction. PCA simplifies the dataset by transforming the original features into uncorrelated components while retaining most of the variance.

Before applying PCA, the train-test split was completed and the predictors were standardized to ensure equal contribution from all features. Then, to determine the number of components to retain, we performed PCA on the standardized data and plotted a scree plot of eigenvalues:



Using the Kaiser Criterion, we retain components with eigenvalues > 1 (three components are shown in the graph). Then, we train logistic regression using the PCA-transformed and training data and evaluate using the following quality metrics:





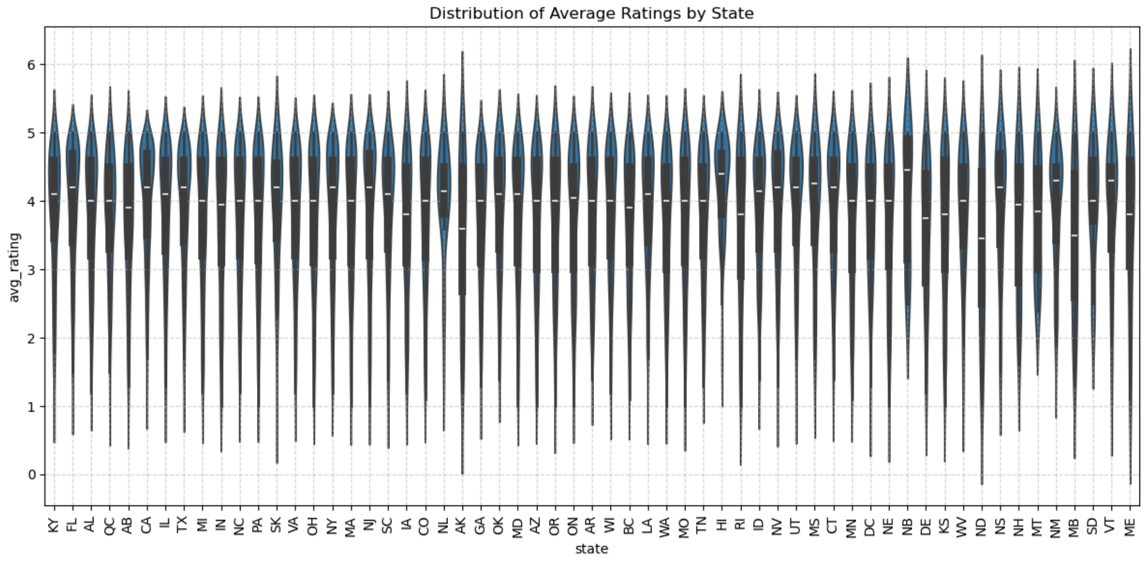
```
ROC AUC Score: 0.774
Confusion Matrix (Q10 - PCA-Based):
[[921 413]
 [297 801]]

Classification Report (Q10 - PCA-Based):
              precision    recall  f1-score   support

Not Peppered       0.76      0.69      0.72      1334
    Peppered       0.66      0.73      0.69      1098

    accuracy                           0.71      2432
   macro avg       0.71      0.71      0.71      2432
weighted avg       0.71      0.71      0.71      2432
```

The **AUC-RUC** for the PCA model is **0.774**. It still indicates a pretty good performance. The classfication report shows overall accuacy of **71%**, which is slightly higher than the previous model. Overall, the PCA model slightly performs better than the “Average Rating Only” model in terms of overall accuracy and balance between precision/recall. For the explained variance ratio for the PCA components, the first principal component explains a large portion (**33.56%**), suggesting that a single linear combination of predictors captures much of the data’s variance. This means that "average rating" might dominate in predicting "pepper." This reinforces why the PCA model doesn't drastically outperform the "Average Rating Only" model.

Extra Credit

To determine whether average professor ratings vary significantly by state, I conducted a one-way ANOVA using data from states with at least 10 professors to ensure meaningful group sizes. The ANOVA results showed a statistically significant difference between states with an **F-statistic** of **4.292** and **p-value** of **2.446*10^{-25}**. This indicates that **the average ratings vary**

significantly between states. To visualize the distribution of professor ratings across states, I created a plot that highlights both the spread and the central tendency of ratings within each state.


Distribution of Average Ratings by State
avg_rating
state
KY
FL
AL
QC
AB
CA
IL
TX
MI
IN
NC
PA
SK
VA
OH
NY
MA
NJ
SC
IA
CO
NL
AK
GA
OK
MD
AZ
OR
ON
AR
WI
BC
LA
WA
MO
TN
HI
RI
ID
NV
UT
MS
CT
MN
DC
NE
NB
DE
KS
WV
ND
NS
NH
MT
NM
MB
SD
VT
ME