Anthony Lu ayl2063

Capstone Project: Professor Evaluations Analysis

Approach to Data Cleaning:

Like all data, this dataset of professor ratings is imperfect, and contains missing data, which requires preprocessing and cleaning. To address this, I first plotted the proportion of NaN values in each column, which corresponds to a particular piece of information about a professor.

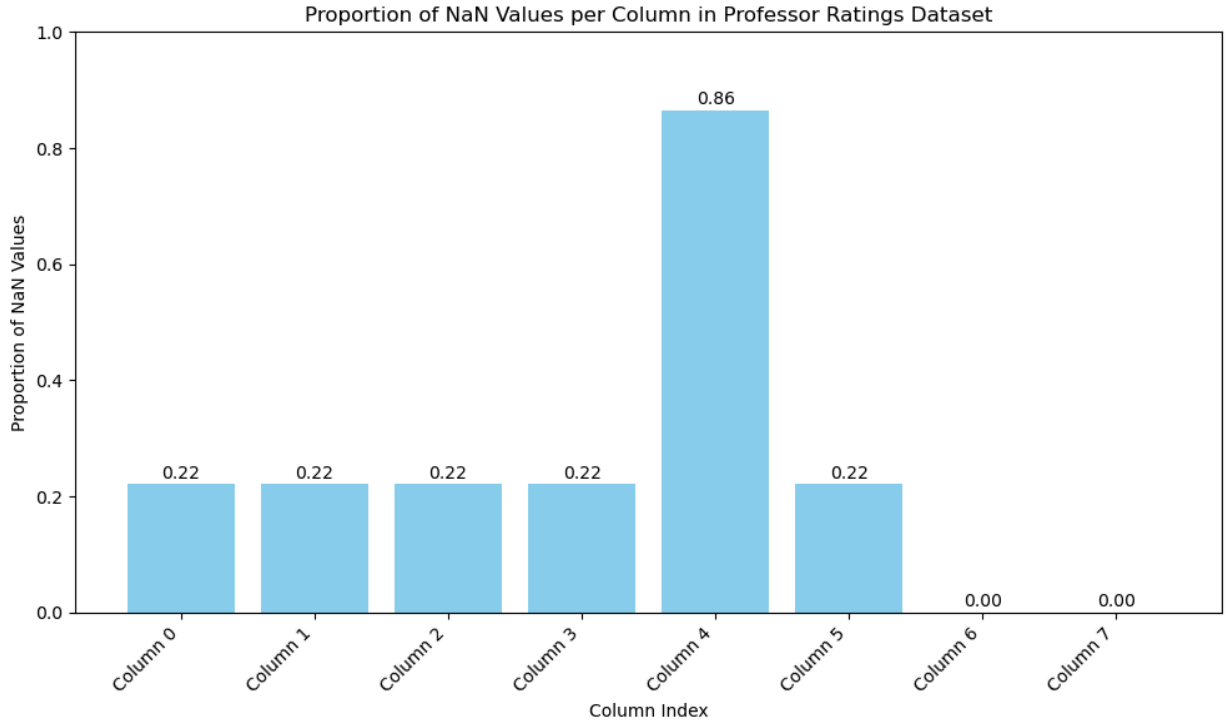

Figure 1: Proportion of missing (NaN) data in each column

As we can see, all columns have a sizable proportion of missing values (except columns 6 and 7). In general, missing values were imputed. To find the best way to impute each column, the distributions of the non-binary columns were looked at (Figure 2).

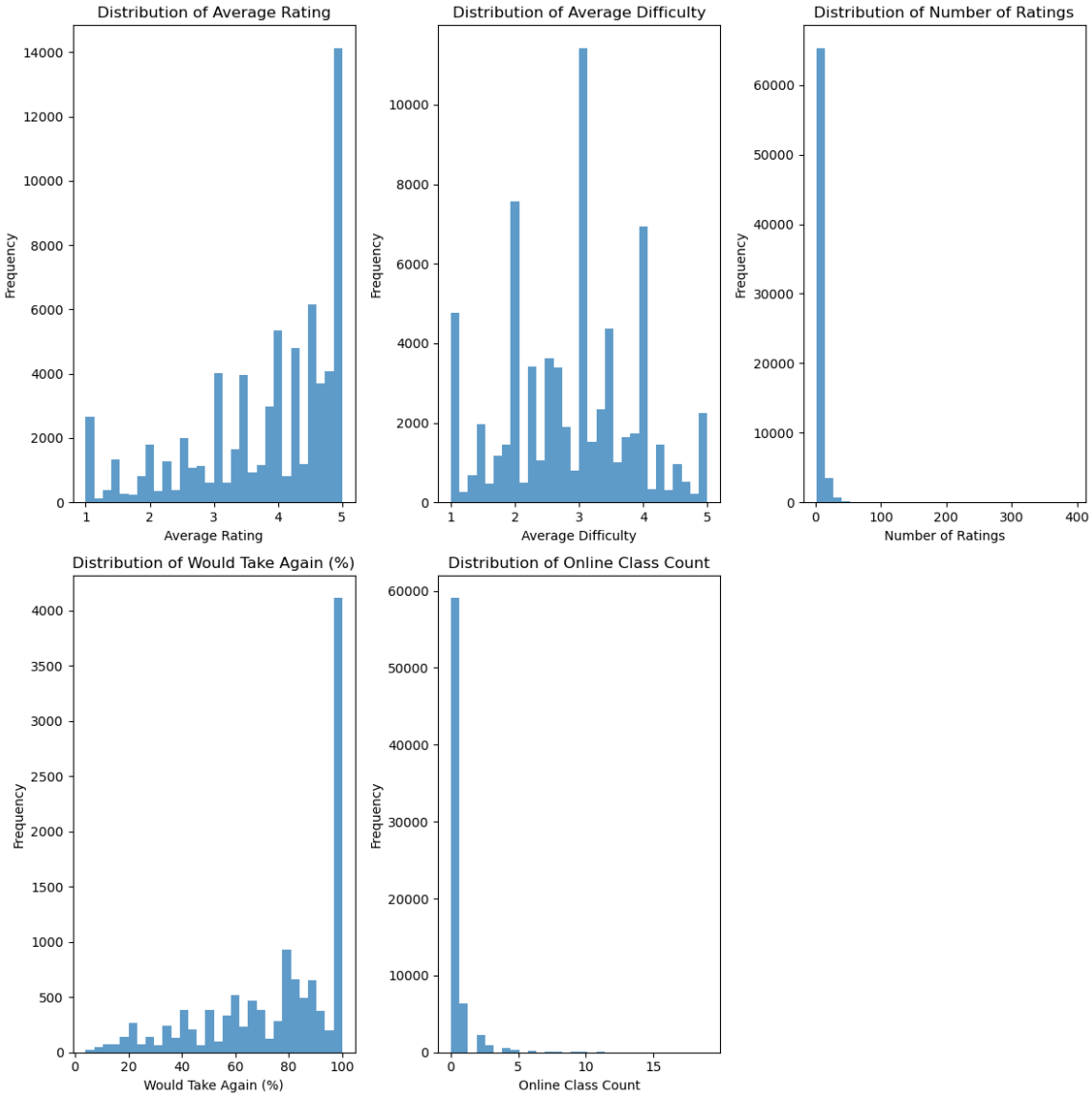

Figure 2: Distributions of columns

Except for average difficulty, none of these columns appeared to approximate a normal distribution, and appeared to be heavily skewed. Hence, average rating, number of ratings, would take again, and online class count were imputed using the median, and average difficulty was imputed using the mean. For columns 6 and 7 (male gender, boolean and female gender, boolean, respectively), only the professors that were strictly male or female were kept, since I was only interested in if the professor being male, with high confidence) affected average rating. For column 4 (pepper), imputation was done using the mode.

Since averages were provided, we need to ensure that the sample size that gives us each average is robust enough for the average to be useful. To look at the robustness of each average, I looked at how much an average rating changed given an additional worst-case rating of 1 or 5 (Figure 3).

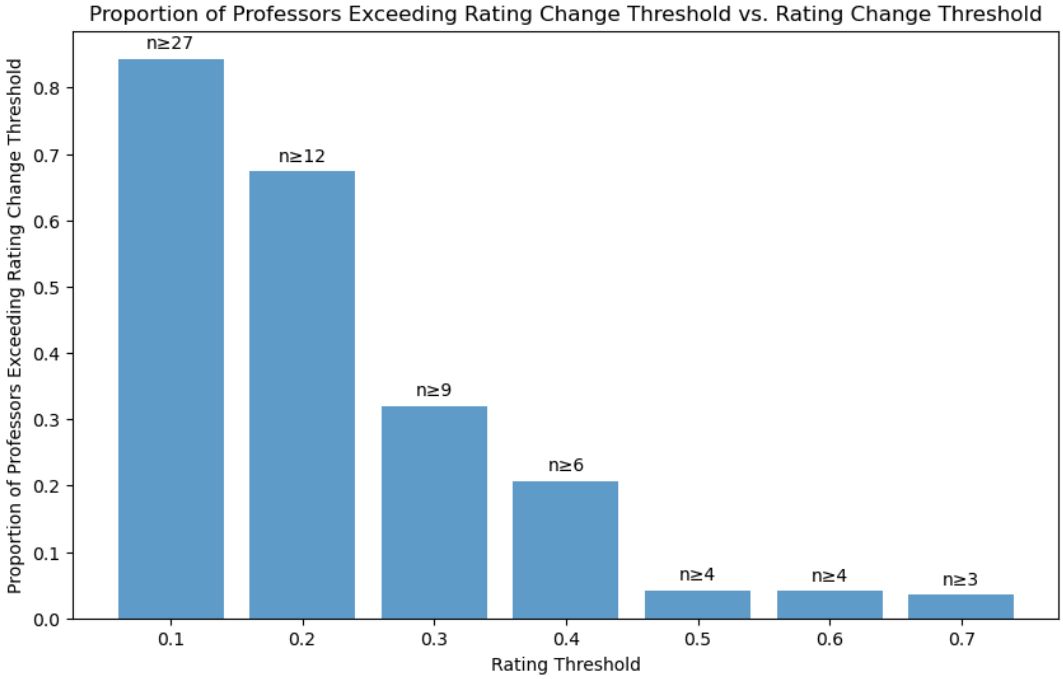

Figure 3: Proportion of data exceeding rating change threshold vs rating change

The lower of a rating change threshold we choose, the tighter and more robust our data is. However, if we filter by a low threshold, there is a risk of losing a large proportion of data. I ended up choosing a rating change threshold of 0.5 (which corresponds to filtering out all data points with less than 4 ratings) to reduce noise and maintain robustness, while still maintaining as much of our data as possible.

Looking at possible gender bias in evaluations:

To find out if gender bias affected student ratings of professors, I first looked at the distribution of the average ratings of male and female professors (Figure 4). This was done to help get an idea of which significance test might be suitable.

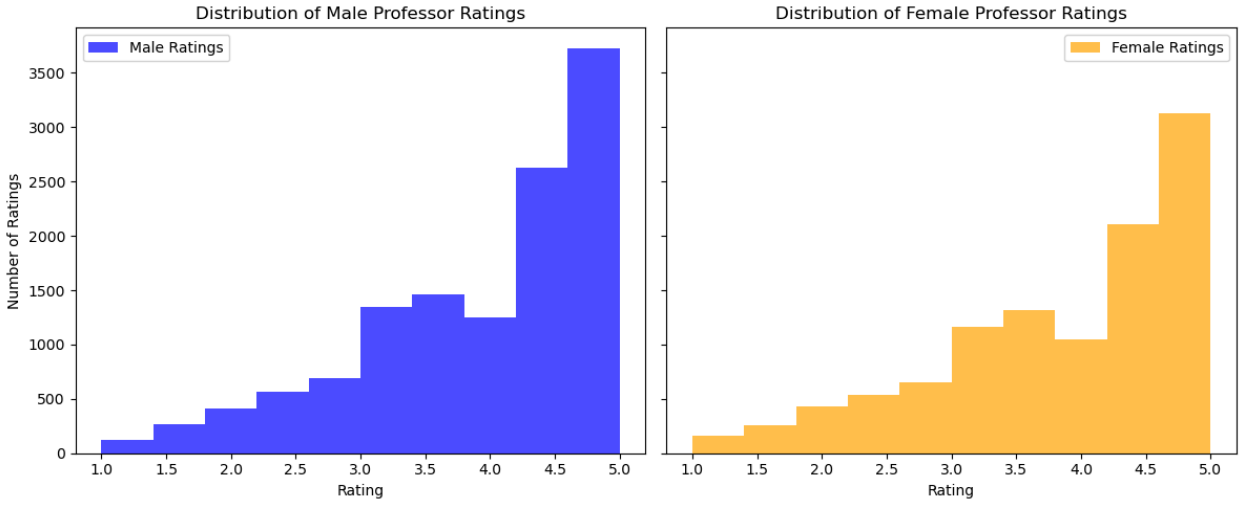

Figure 4: Distribution comparison

As we can see in Figure 4, neither of the distributions of ratings for male and female ratings are normally distributed. This makes the Mann-Whitney U-test a much better option than the standard two independent sample t-test. Using the Mann-Whitney U-test, a p-value of $7.9 * 10^{-5}$ was achieved. Thus, using an alpha level of 0.005, we can conclude that gender has an effect on professor ratings.

Looking at the effect of experience on teaching quality:

To find out if experience affected perceived teaching quality, average rating was used to measure professors' teaching quality, and the number of ratings was used as a proxy for professor experience. I then needed to separate ratings (quality) into ratings for professors with "high" experience and ratings for professors with "low" experience. I first looked at the distribution of teaching experience (number of ratings) (Figure 5).

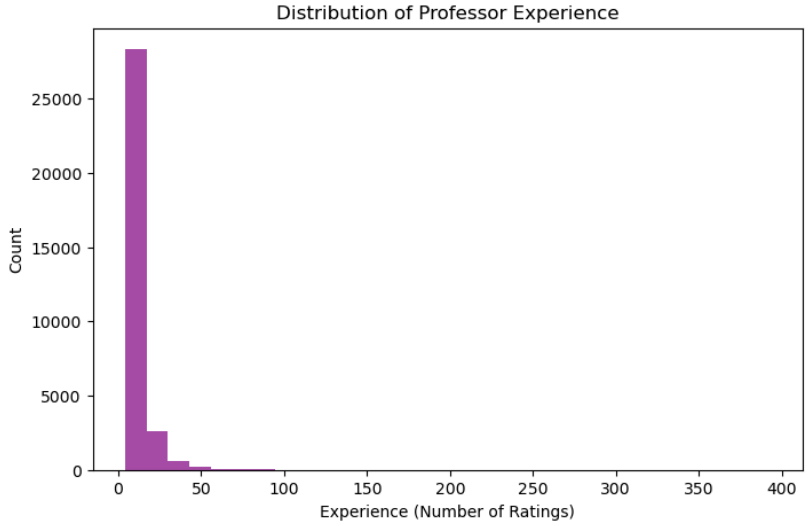

Figure 4: Distribution of professor experience

As we can see in Figure 4, the distribution of professor experience is heavily skewed to the left, which means that median, rather than mean, would be more suitable as a summary value for professor experience. Thus, I decided to separate ratings (quality) into professor ratings with "high" experience (greater than or equal to the median experience) and professor ratings with "low" experience (less than the median experience). To see which significance test would be suitable for seeing if there is a statistically significant difference between experienced professor ratings and inexperienced professor ratings, I looked at the respective distributions for both datasets (Figure 5).

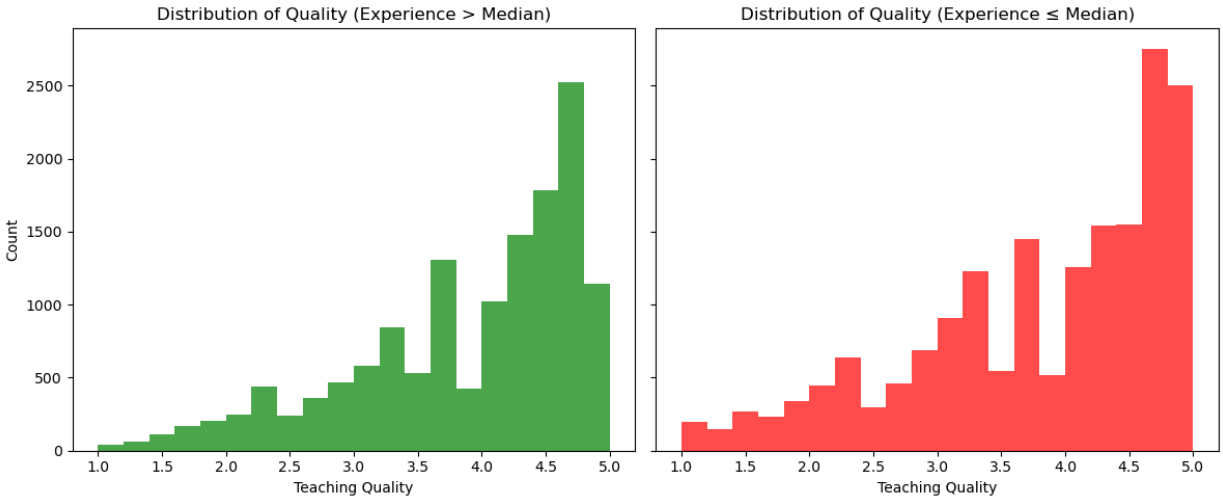

Figure 5: Experience and quality distributions

As we can see in Figure 5, neither distributions approximate the normal distribution and appear to be skewed to the right. This means that the Mann-Whitney U test would be suitable. Using the Mann-Whitney U test, a p-value of 0.0006 was achieved. Thus, using an alpha level of 0.005, we can conclude that having more experience does have an effect on teaching quality.

Relationship between ratings and difficulty:

To see if there exists and quantify a relationship between average rating and average difficulty, a correlation coefficient is needed. Since both metrics use a scale of 1-5, they are both ordinal and bounded, which means that the Spearman correlation coefficient would be the clear choice (Figure 6).

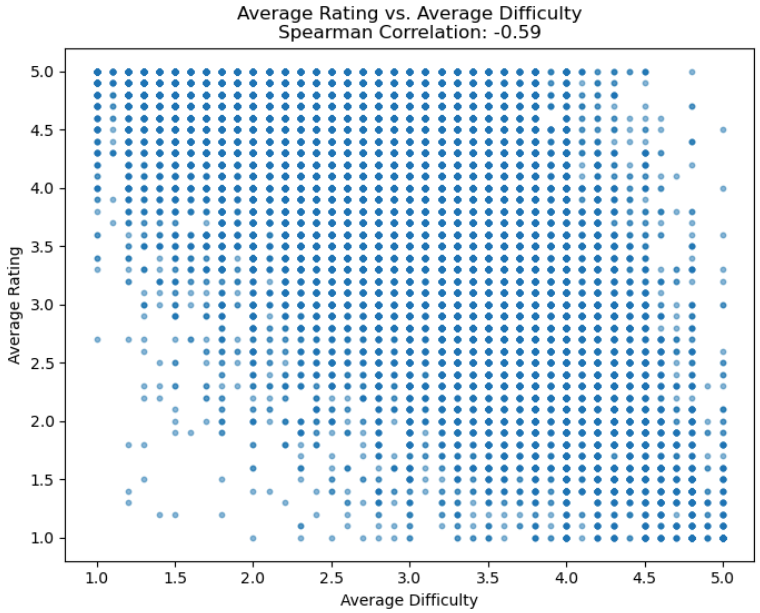

Figure 6: Average rating vs average difficulty correlation

As we can see in Figure 6, between average rating and average difficulty, there is a Spearman correlation coefficient of -0.59, which means that there is a moderate negative correlation between rating and difficulty. Looking at the scatter plot (Figure 6), there does also appear to be an overall monotonic relationship. This tells us that higher difficulty scores tend to lead to lower average ratings.

Effect of online teaching on ratings:

To see if professors who teach a lot of online classes receive higher or lower ratings, we first need to define what "a lot" of online classes means. We can look at the distribution of the number of online classes for each professor (Figure 7) to see if we can use mean or median as a summary statistic.

Figure 7: Distribution of number of online classes

As we can see in Figure 7, the distribution is heavily skewed to the left. This suggests that the median is a suitable summary statistic for this case. However, since the majority of professors don't teach online courses, the median came out to be 0, which isn't of much use to us. This does tell us that most professors exclusively teach offline, so we can say

that professors that teach at least one online class teach “a lot” of online classes. I then separated the average ratings into two sets: ratings for professors that don’t teach online classes and ratings for professors that teach at least one online class. Looking at the distributions for both sets of ratings (Figure 8):

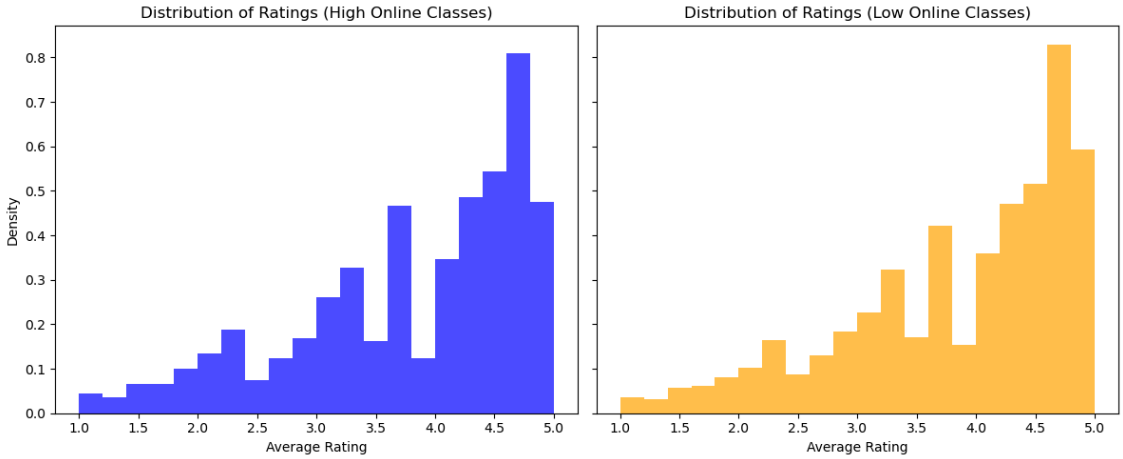

Figure 8: Distributions of ratings

Once again, both distributions are heavily tailed and not normally distributed, so the Mann-Whitney U test would be suitable to see if there is a statistically significant difference between ratings. Using the Mann-Whitney U test, a p-value of $2.7 * 10^{-6}$ was calculated, which, using an alpha level of 0.005, means that there is a statistically significant difference between ratings from professors that teach a lot of online classes and ratings from professors that don’t teach a lot of online classes. The ratings from professors that teach a lot online classes had a median rating of 4.0 and the ratings from the professors that don’t had a median rating of 4.1 Thus, professors that teach less online classes tend to have higher ratings.

<u>Relationship between average ratings and proportion of people who would take the class again:</u>

To see the relationship between a professor’s average rating and the proportion of people who would take the professor’s class again, we need to calculate the correlation coefficient between these two variables. Both of these variables are bounded, with the ratings being ordinal (on a scale of 1-5) while the proportion of people that would take the class again isn’t (a percentage from 0-100). Looking at the distributions (Figure 9):

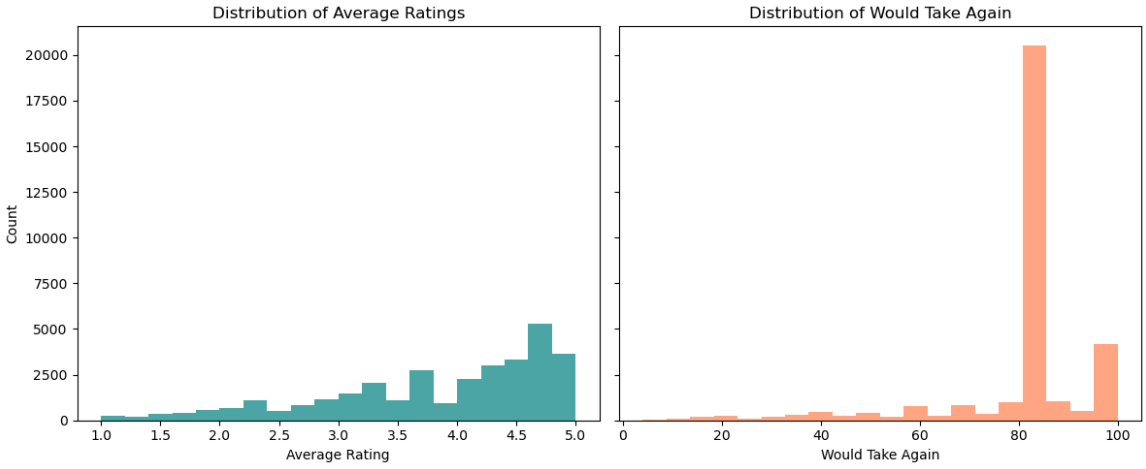

Figure 9: Distribution comparison

As we can see, neither of these distributions are normally distributed, which means that the Spearman correlation is most likely more suitable for this case. However, in this case, a very similar value is achieved using both correlation coefficients (Figure 10):

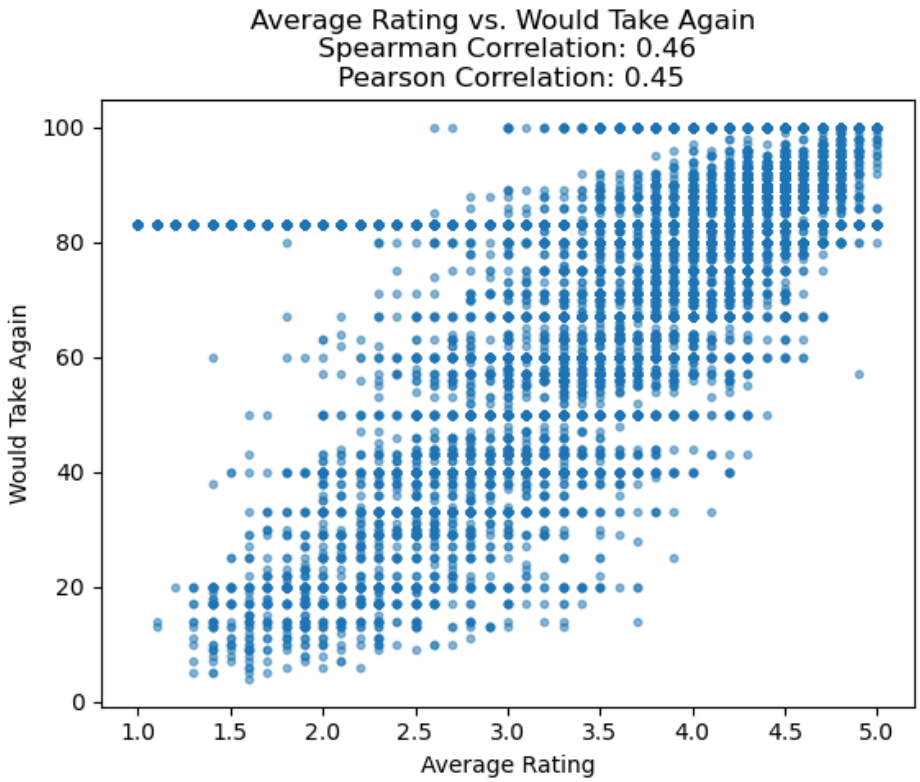

Figure 10: Correlation coefficients between average rating and proportion of people that would take the class again

As we can see in Figure 10, there is a Spearman correlation of 0.46 and a Pearson coefficient of 0.45. This tells us that there is a positive correlation between the two variables, and that as rating increases, more students would take the course again. As we can see in the plot, there is also an overall monotonic relationship, and it also closely approximates a linear relationship as well.

Looking at "hotness" effect on professor ratings:

To look at if professors who were "hot" received higher ratings, I separated the average ratings into a set of ratings for professors that were rated "hot" and a set of ratings for professors who weren't. I looked at the sample sizes (Figure 11) and distributions (Figure 12) of each set of ratings:

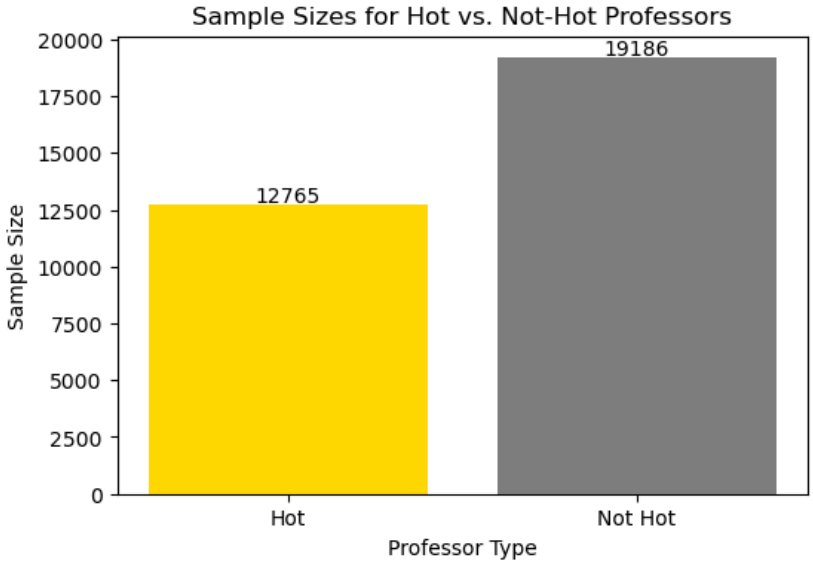

Figure 11: Sample size comparison

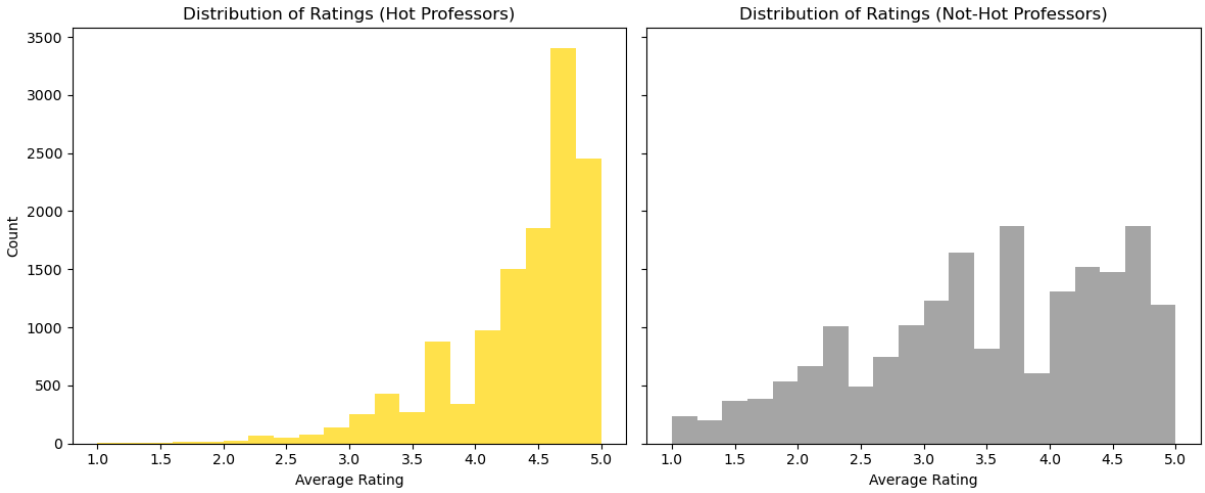

Figure 12: Distribution comparison

As we can see in Figure 11, although the classes aren't balanced, the imbalance isn't too extreme. We can also see in Figure 12 that neither of the distributions are normal, which means that the independent samples t-test wouldn't be suitable in this. Using the Mann Whitney U test, a p-value of 0.0 was calculated (due to Python underflow), which tells us that there is a statistically significant difference between the two sets of ratings. The ratings for "hot" professors had a median of 4.5 while the ratings for "not hot" professors was 3.6. Thus, "hot" professors tend to have higher ratings than those who aren't.

Using regression to predict average ratings from average difficulty:

To predict average ratings from average difficulty using linear regression, I first split the entire dataset (all columns) into a training and test set. I split all of the predictors, rather than just difficulty, so this regression model (with a single predictor) could be compared to the model with all predictors fairly (using the same training and test set). When training the model itself, I only included the column corresponding to difficulty. I then calculated the RMSE and R^2 of the model (Figure 13).

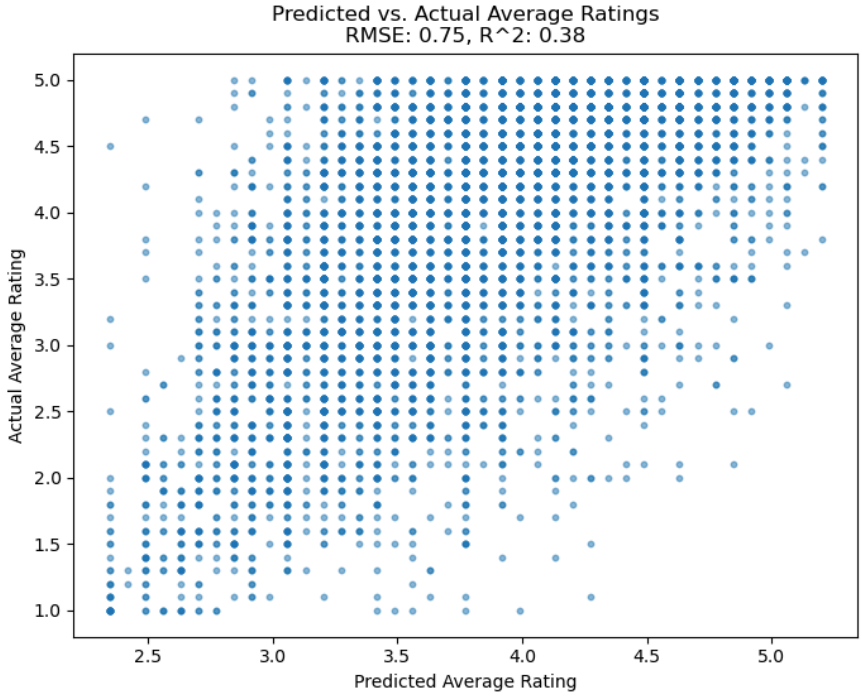

Figure 13: Predicted vs Actual - average difficulty as the only predictor

As we can see in Figure 13, an RMSE of 0.75 and an R^2 of 0.38 was achieved.

Predicting average rating from all predictors using regression:

To predict average rating from all predictors, I used the same train/test data from the previous part, but this time including all predictors, rather than only difficulty. I then trained the model using all predictors (Figure 14).

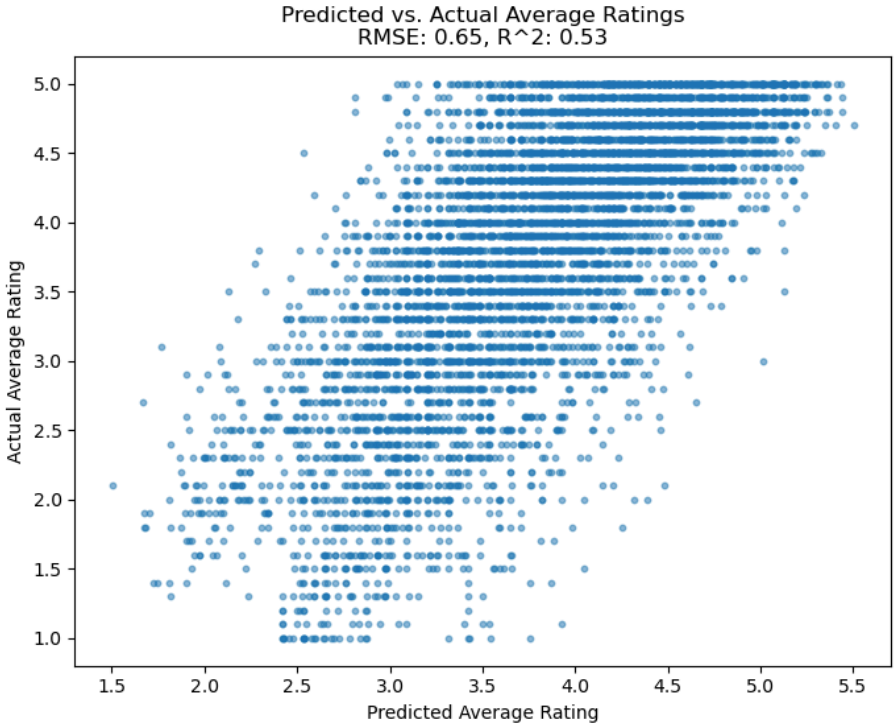

Figure 14: Predicted vs Actual - Predicting average ratings from all available predictors

As we can see in Figure 14, RMSE decreased from 0.75 to 0.65 while R^2 increased from 0.38 to 0.53, when compared to the previous model that only had one predictor. When using multiple predictors, collinearity can become a concern, since predictors often contain redundant information. To see which predictors may be of concern for collinearity, I created a correlation matrix between all predictors (Figure 15).

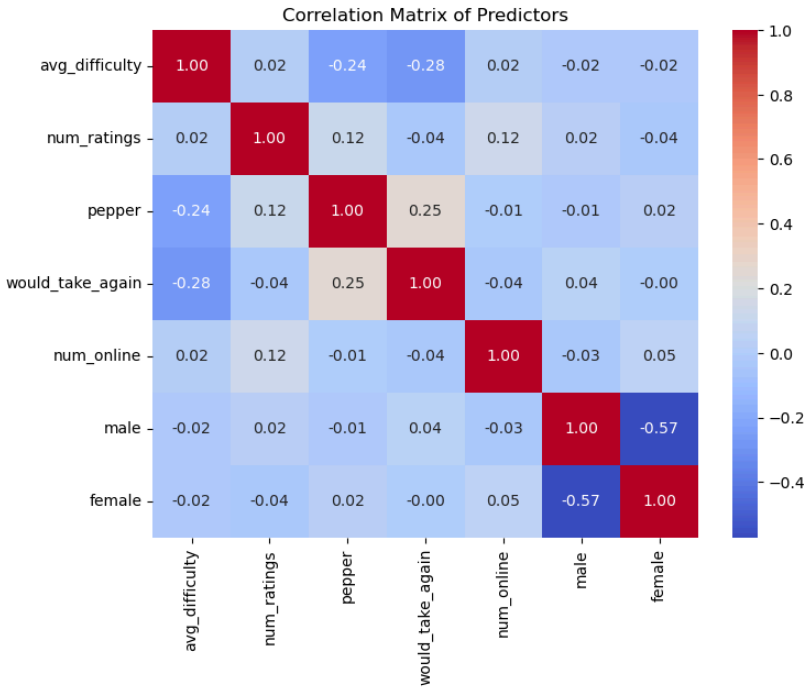

Figure 15: Correlation between predictors

As we can see in Figure 15, there is a strong correlation between male and female, which makes sense, since professors that aren't male are typically female and vice versa. We also see somewhat of a correlation between pepper and would take again, as well as between difficulty and pepper and between difficulty and would take again.

Predicting "pepper" from ratings using classification:

To classify if a professor was given a pepper on not from average ratings, I used a logistic regression model. I first created a train/test set, with the labels being the pepper column. I split up all predictors, rather than just ratings so models can be fairly compared on the same train/test set. I then trained a logistic regression model using average ratings as the sole predictor and calculated the accuracy and the ROC curve (Figure 16).

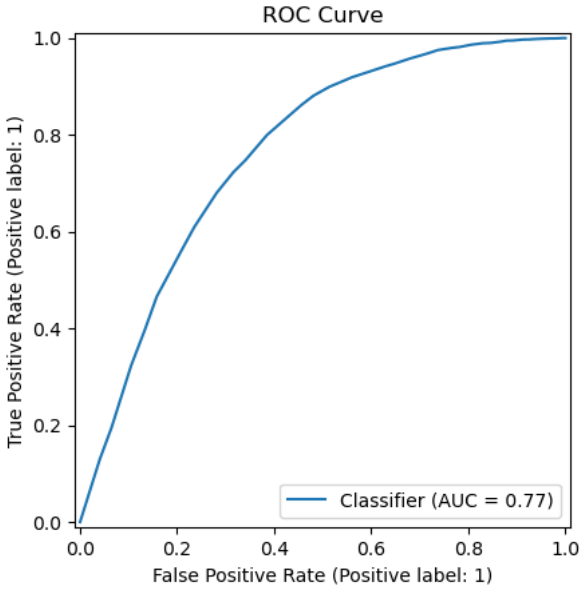

Figure 16: ROC Curve

This logistic regression model yielded an accuracy of 0.71, which tells us that the model makes correct predictions roughly 71% of the time. However, this can be misleading, especially in cases with imbalanced datasets (Figure 17).

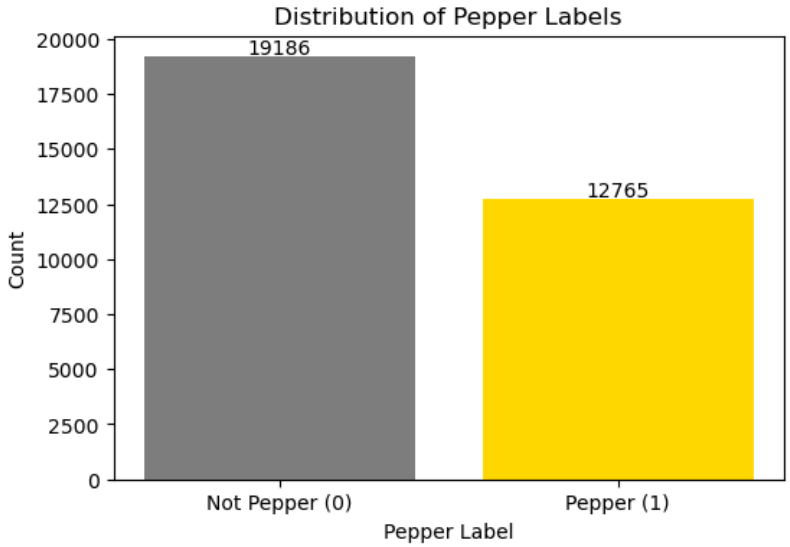

Figure 17: Pepper class sizes

As we can see, the dataset does have an imbalance, with the majority class being the professors not being labeled with a pepper. A much stronger metric would be AUC (Area under ROC curve) , which tells us how good the model is at distinguishing between classes. In Figure 16, we can see that this model has an AUC score of 0.77.

Predicting "pepper" from all predictors using classification:

To classify if a professor was given a pepper on not from all predictors, I used a logistic regression model. I continued to use the train/test set from the previous model to fairly compare the two classification models. To gauge model performance, I calculated the accuracy as well as the ROC curve (Figure 18).

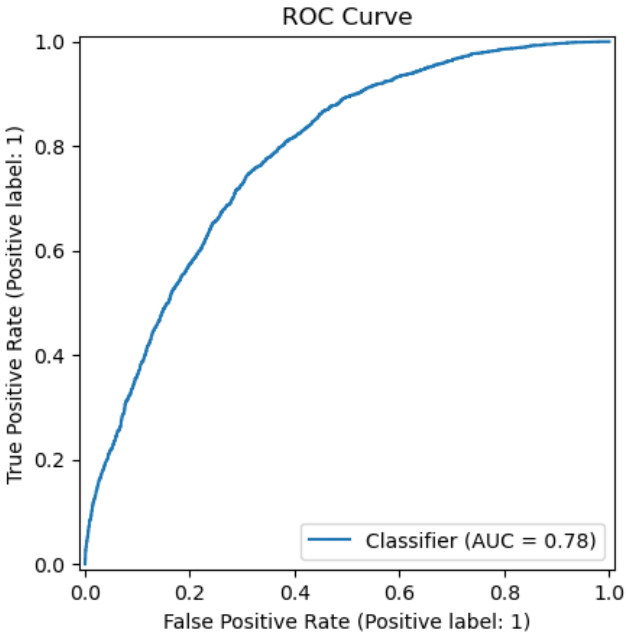

Figure 18: ROC Curve

This logistic regression model, with all the predictors, also yielded an accuracy of 0.71, which was very similar to the accuracy score from the previous model with only one predictor. However, due to the class imbalance (Figure 17) between “hot” professors and “not hot” professors, accuracy may not be suitable. In this case, AUC may be a more suitable metric, since it is more robust against class imbalances. This model had an AUC score of 0.78, which is a marginal improvement over the previous model, which had an AUC score of 0.77.