Srivar Janna,

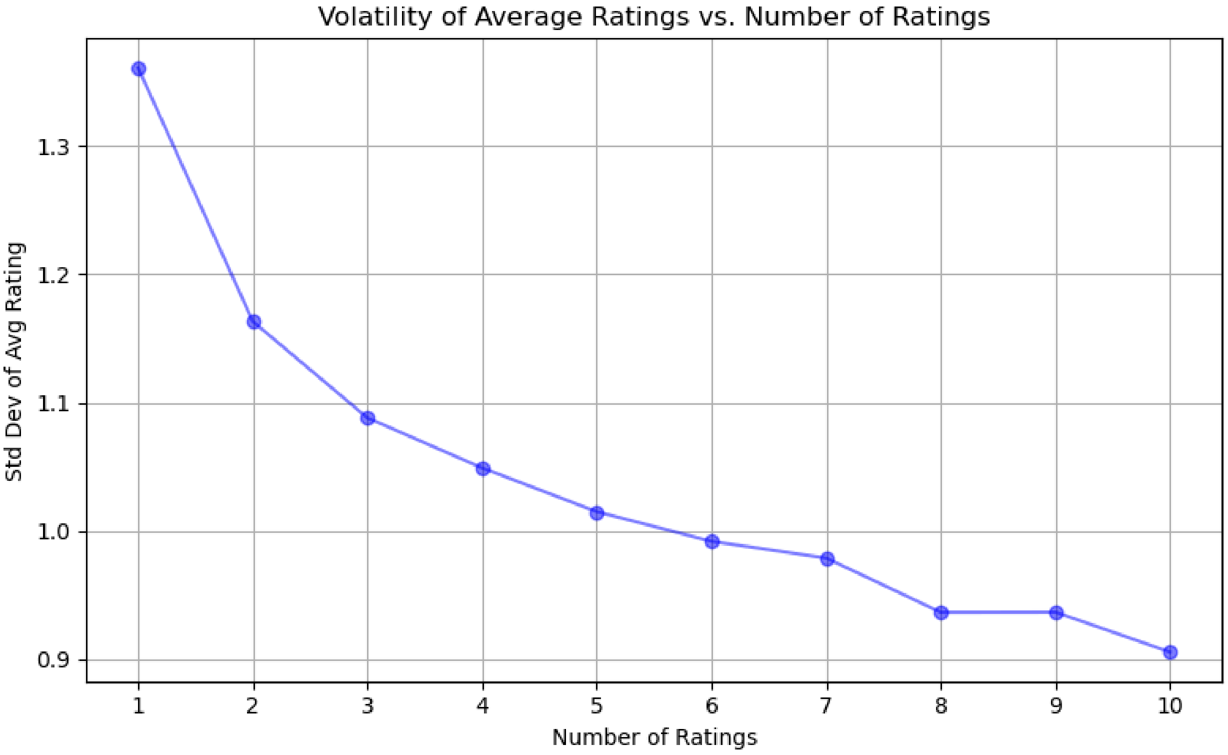
Professor Pascal Wallisch,

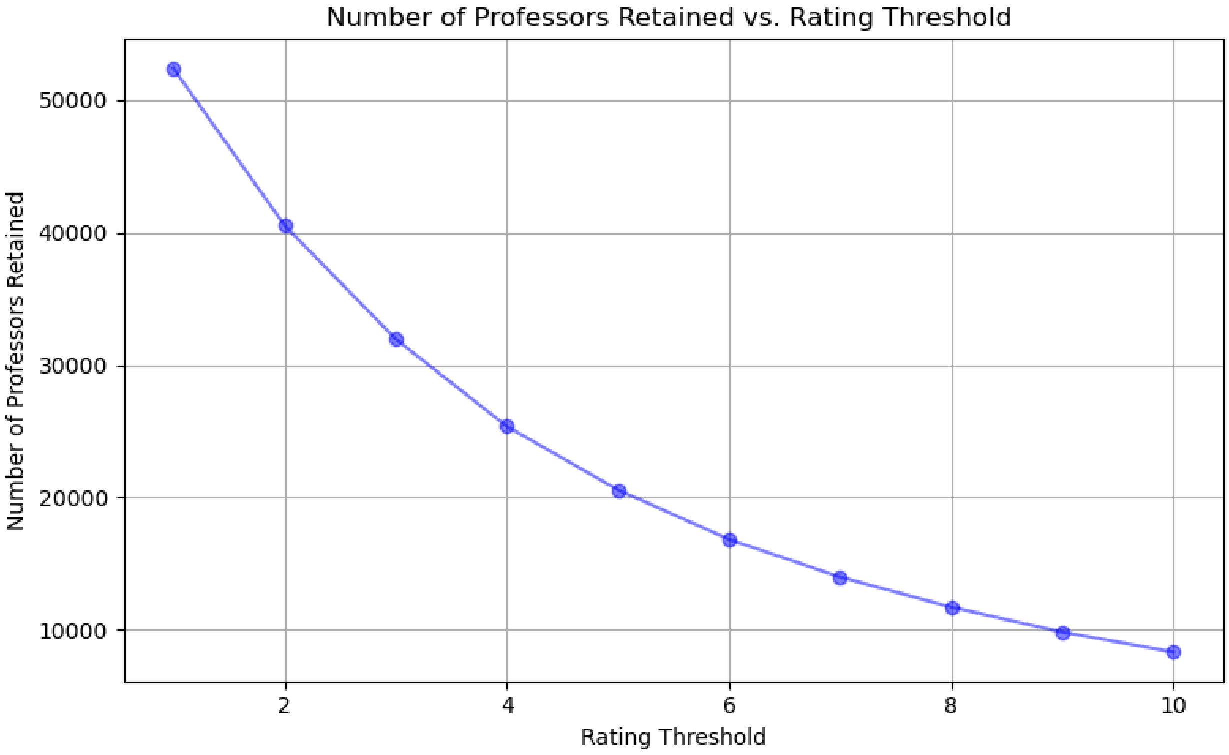
25 April 2025,

Data Science Capstone Project

Data Cleaning:

The dataset was loaded using the numpy library. First, I checked the "Number of Ratings" column to see if there were any professors with no ratings or with only NaN values and filtered them out. Furthermore, to avoid extreme values I decided to set a threshold for the minimum number of ratings a professor must have to be included in the next part of the project. To help choose the ideal threshold, one which doesn't drop too much data and reduces volatility, I created two visualizations. One that tracks the number of professors retained and one that checks the standard deviation of the ratings for each threshold.





Using these two graphs, I set 4 as the minimum number of ratings a professor must have to be included as it doesn't reduce the data by too much and also stabilizes the volatility of the ratings. This was the starting data for each question and then the dataset was filtered further based on the questions requirement.

Approach To Significance Testing

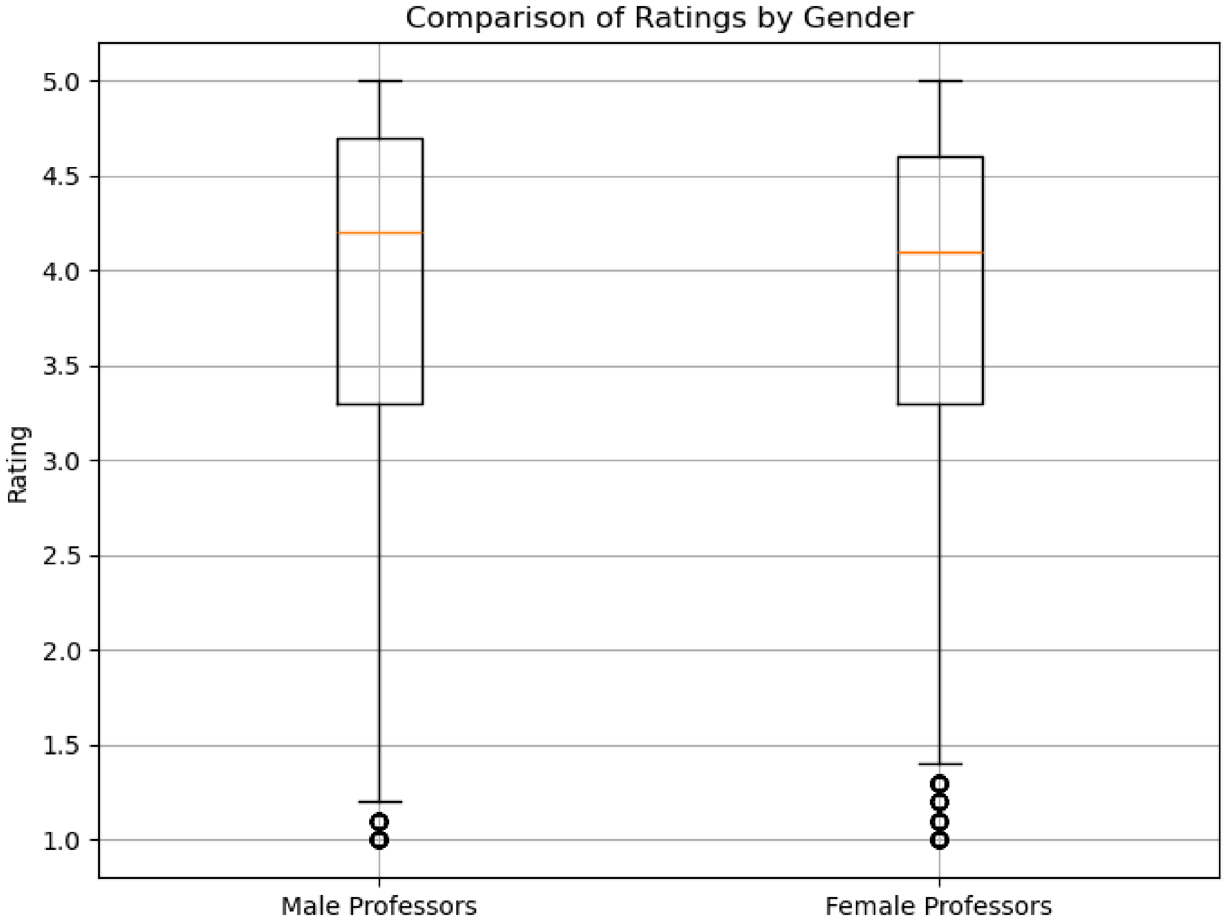
The Alpha level for significance testing was set to 0.005 to drastically reduce the chance for false positives. Since the questions involving significance testing are ratings related it is not completely appropriate to assume cardinality. Hence, the Mann-Whitney U test was used. However, it does not control for confounders. Multiple linear regression can be used to control for confounding variables, however, the downside is it assumes cardinality. Given the limitations of both individual methods, both multiple linear regression and a non-parametric U-test are performed and consistent findings across both would strengthen confidence in the result.

Question 1: Is there a Pro-Male Gender Bias in the Dataset?

The dataset contained many rows in which the values in both the male and female columns were either 0 or 1. Due to ambiguity in gender that arises from these rows, they were filtered out for this question. As stated earlier, both a multiple linear regression and a Mann-Whitney U test was done.

In the dependent variables for the multiple linear regression, the “Would Take Again” column was dropped because it had too many NaN values. Also, the “Female” column was dropped because a 0 in the male column would imply female. The 99.5% confidence interval generated for the effect of the gender “Male” on average rating, holding other variables constant, was [0.04, 0.10]. This provides evidence that there is a statistically significant bias towards male professors, however the effect size of this bias is small.

For the Mann-Whitney U test, the dataset was split into two groups, based on whether they had a 0 or a 1 in the “Male” column. The U-test statistic was found to be 65395208.5. The p-value for that test statistic is 7.8e-05. Hence, even the U test suggests there is a statistically significant bias towards male professors. The rank biserial was also calculated to measure effect size and that was found to be 0.03, which is a very small effect size. The boxplot below shows that apart from extremely low ratings, both male and female professors’ ratings have very similar quartiles and medians.



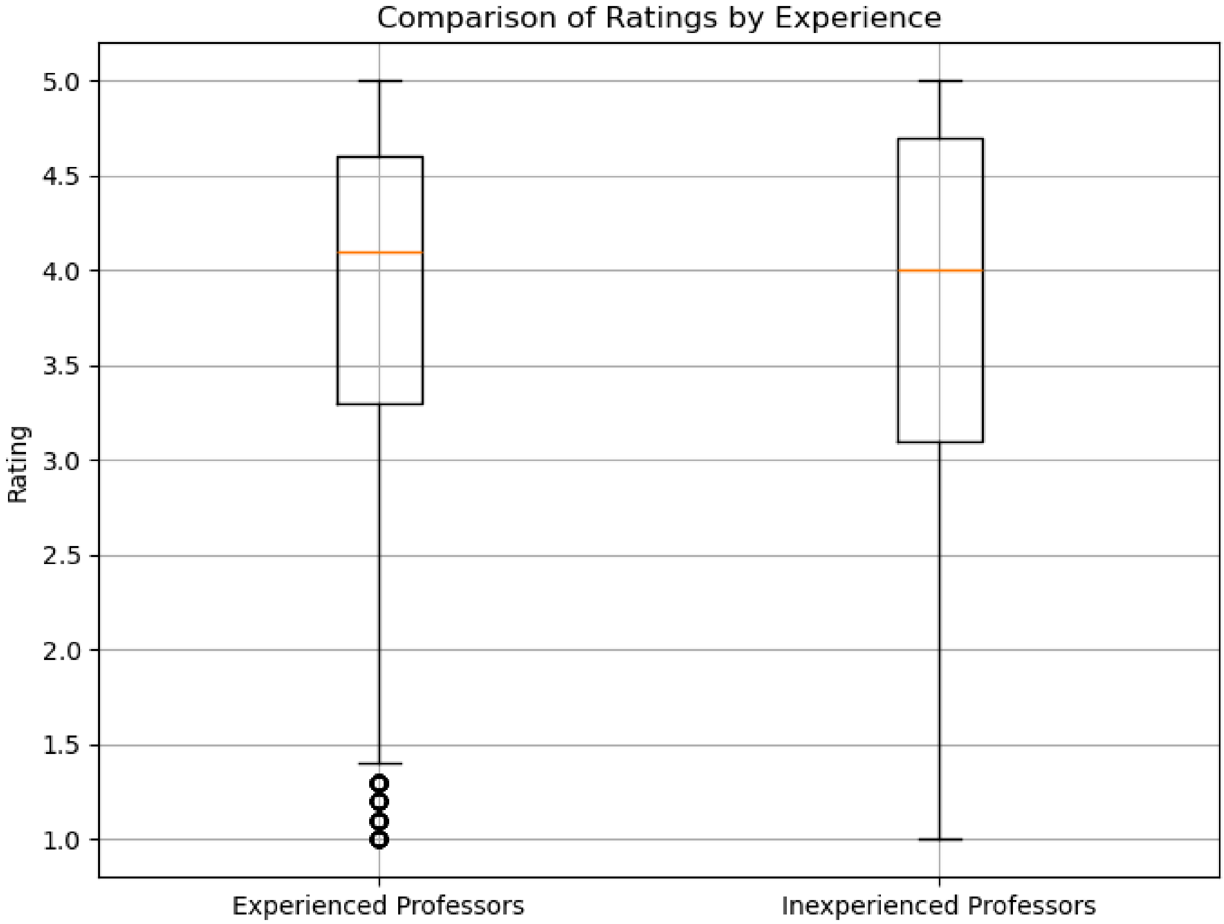
Question 2: Is there an effect of experience on the quality of teaching?

For this question, “Number of Ratings” and “Average Rating” were used as proxies for experience and quality of teaching, respectively.

As mentioned above, both the U-test and multiple linear regression were conducted. For the multiple linear regression, the gender columns and “Would Take Again” columns were not considered as predictor values. The gender columns were not considered because of the ambiguity mentioned in the previous question, and the removal of those rows would lead to the loss of a lot of data. Also, the effect size of gender on rating bias was found to be very small, hence not considering it wouldn’t greatly affect the results. The 99.5% confidence interval generated for the effect of experience on average rating, holding other variables constant, was [0.001, 0.004]. Although this is a very narrow confidence interval, the actual effect of experience according to Rate My Professor ratings is a really small positive one.

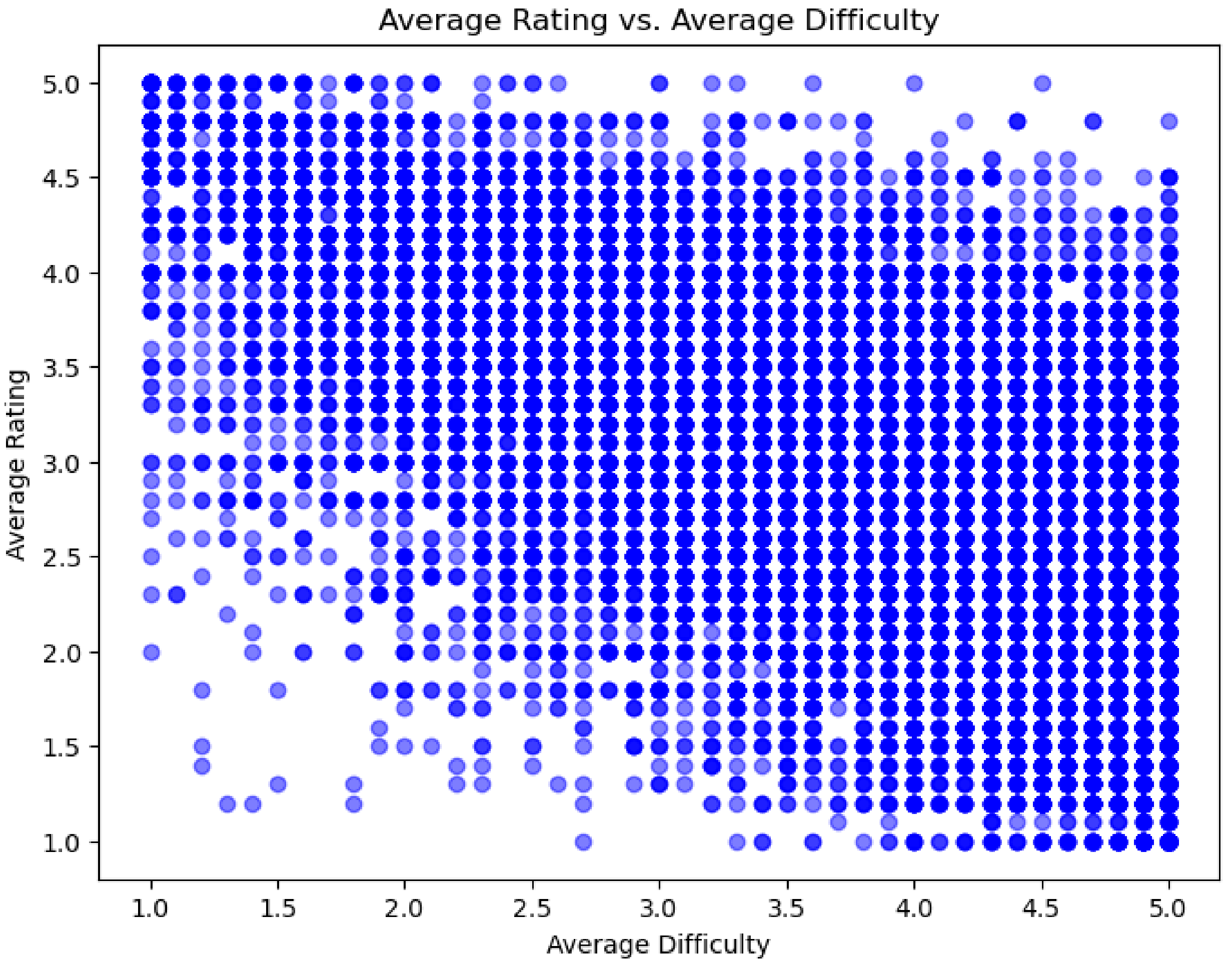
To conduct the Mann-Whitney U-test the dataset must be split into two groups, one with experienced professors and one with inexperienced professors. To do so, a threshold for number of ratings must be set to classify a professor as experienced. Arbitrarily, the value for experience was set as the 75th percentile of the data before the initial data filtration. This value was 6 ratings. The U-test statistic was found to be 117056077 and the p-value for this statistic is 0.96, hence, it is not significant.

However, there is a fundamental issue with this test. Since only average ratings are used, it equally weighs ratings that a professor received before becoming “experienced”. Hence, this test is biased towards the null which might have led to the not significant finding. Also, it is possible that with other thresholds, both for pre-processing and for experience we could’ve arrived at a different result.



Question 3: Relationship Between Average Rating and Average Difficulty?

The Pearson Correlation Coefficient between Average Rating and Average Difficulty was -0.61 while the Spearman Rank Correlation Coefficient was -0.59. This clearly indicates that an increase in difficulty would lead to a significant decrease in rating.

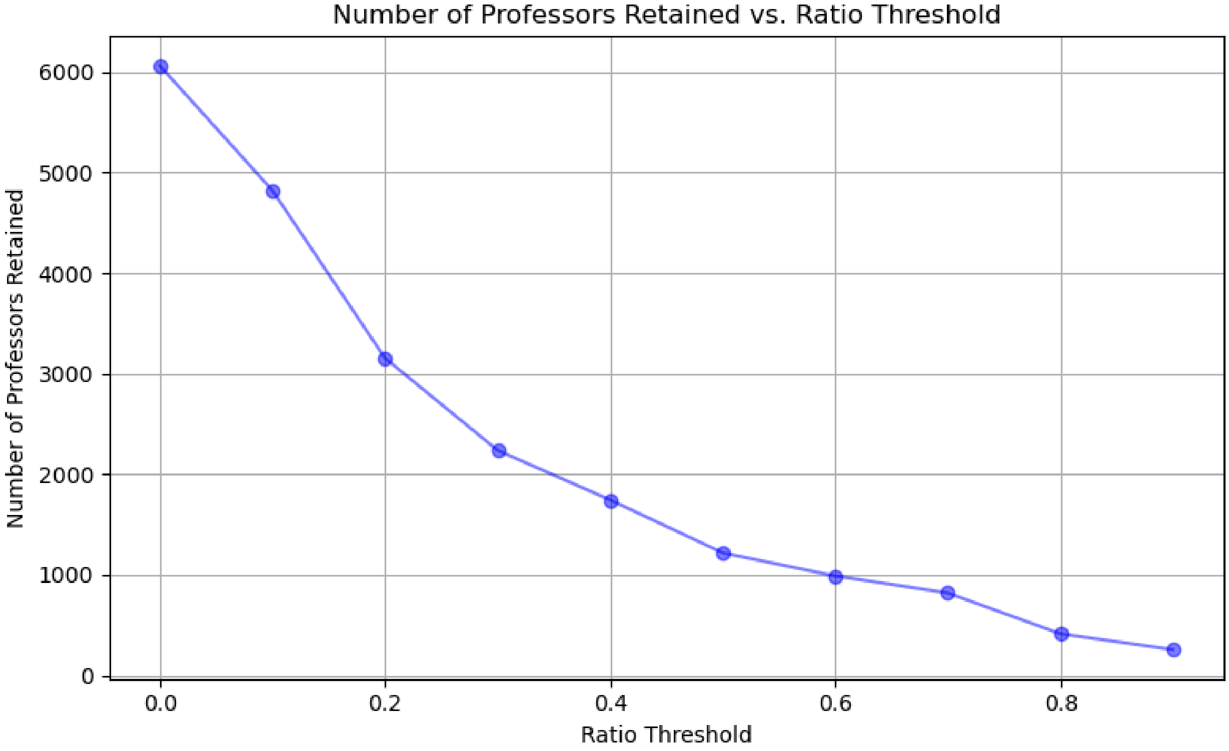


Question 4: Do Professors who Teach a Lot of Classes Online have Higher or Lower Ratings?

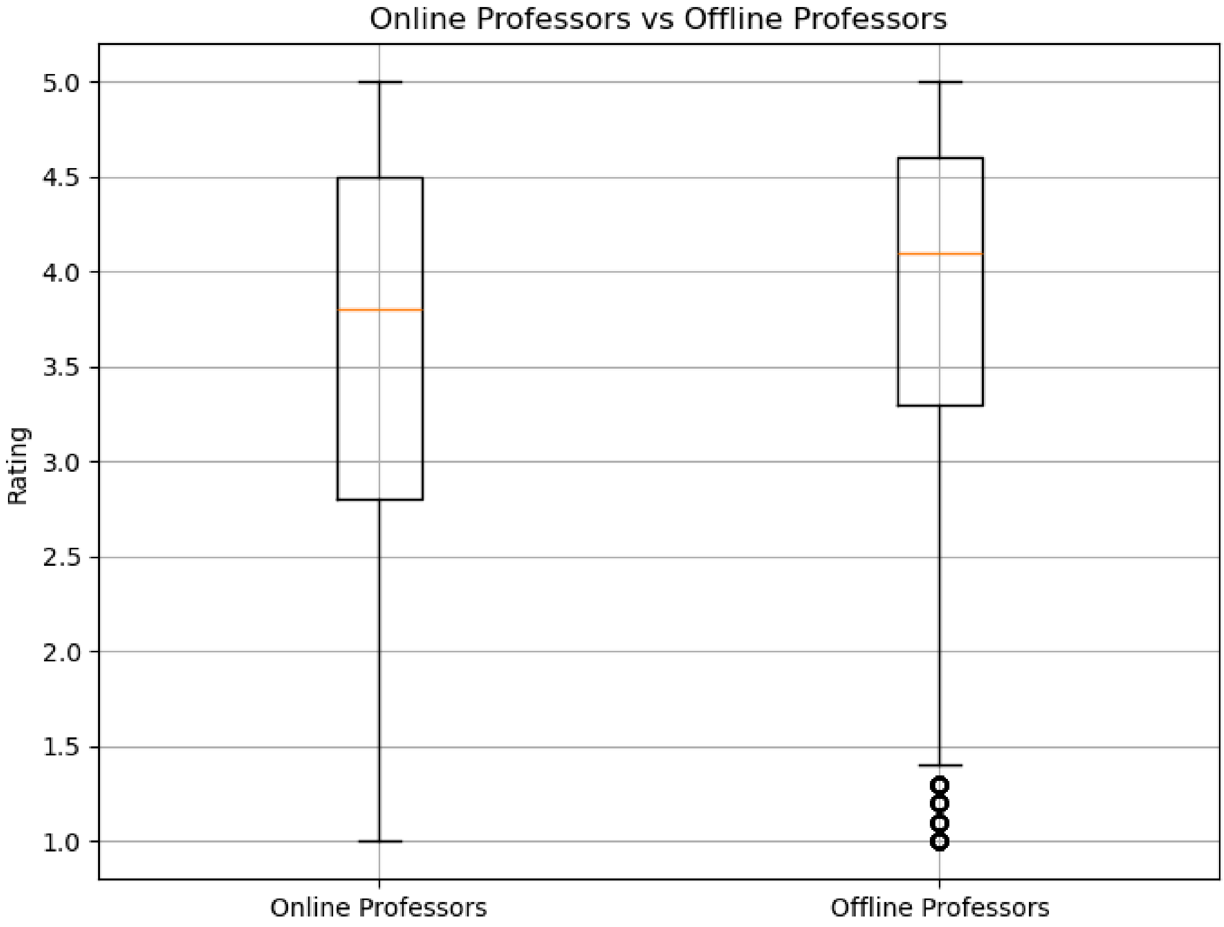
In this question I used the proportion of ratings coming from online classes as a metric for the online classes taught by that professor. The "number of online ratings" column was transformed into the proportion of total ratings that came from online classes.

Consistent with the approach to multiple linear regression in Question 2, the gender columns and the "Would Take Again" column were excluded due to ambiguity and missing data. The multiple linear regression analysis revealed that a higher proportion of online ratings is significantly associated with lower average ratings, with a 99.5% confidence interval for the effect of [-0.21, -0.08]. We can thereby conclude that professors who teach a greater share of their classes online tend to receive lower ratings.

To perform a Mann-Whitney U test, the dataset must be split into two groups: professors who teach a large proportion of their classes online and those who do not. To achieve this, a threshold for the ratio of online reviews to total reviews was selected. Using the graph below, a cutoff of 0.4 — meaning 40% of ratings coming from online classes — was chosen to separate the groups.



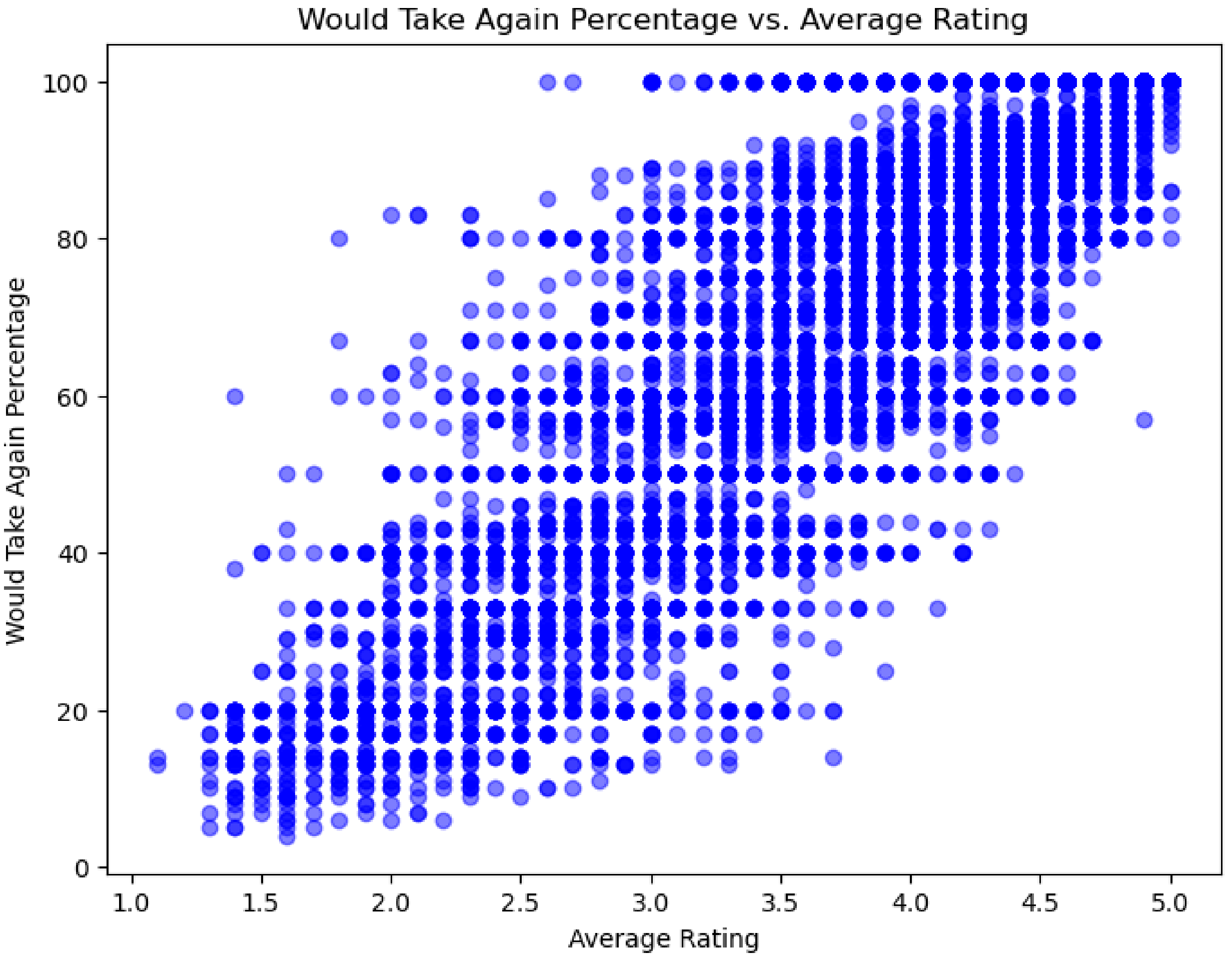
The U-test was performed, and the test statistic was found to be 25206715.5 and the p value for that statistic is 4.8e-24. Hence, this is highly statistically significant. The effect size was found to be 0.14.



Therefore, both multiple linear regression and the Mann Whitney test suggest that professors who teach a lot of classes in the online modality tend to receive higher or lower ratings than those who don’t.

Question 5: Relationship between the Average Rating and the Proportion of People who Would Take the Class Again

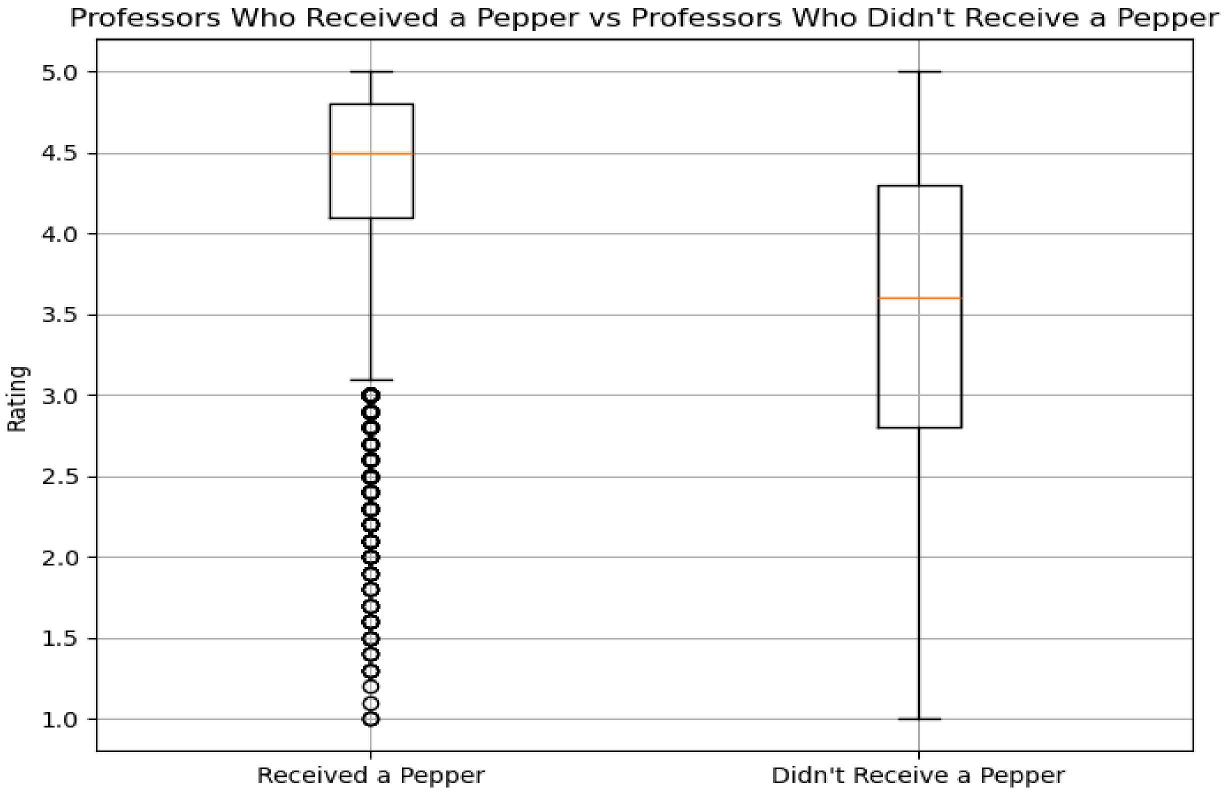
The Pearson Correlation Coefficient between Average Rating and Average Difficulty is 0.88. while the Spearman Rank Correlation Coefficient was 0.85. This indicates a very strong positive correlation between average rating and the proportion of people who would take the class again.



Question 6: Do professors who are "Hot" Receive Higher Ratings than those who are not

The same multiple regression model that was used in Question 2 is used to answer this question. The 99.5% confidence interval generated for the effect of a professor being conceived as “hot”(receiving a pepper) on average rating, holding other variables constant, was [0.59, 0.64]. This provides strong statistically significant evidence that professors who are “hot” receive significantly higher ratings than those who are not.

For the Mann Whitney Test, the data was split into two groups, professors who received a pepper and those who didn’t. Then the U-test was performed, the u statistic was found to be 57362202.5 and the p value for that test statistic is almost 0. Hence, it is shown with high statistical significance that professors who are “hot” receive higher ratings than those who are not.



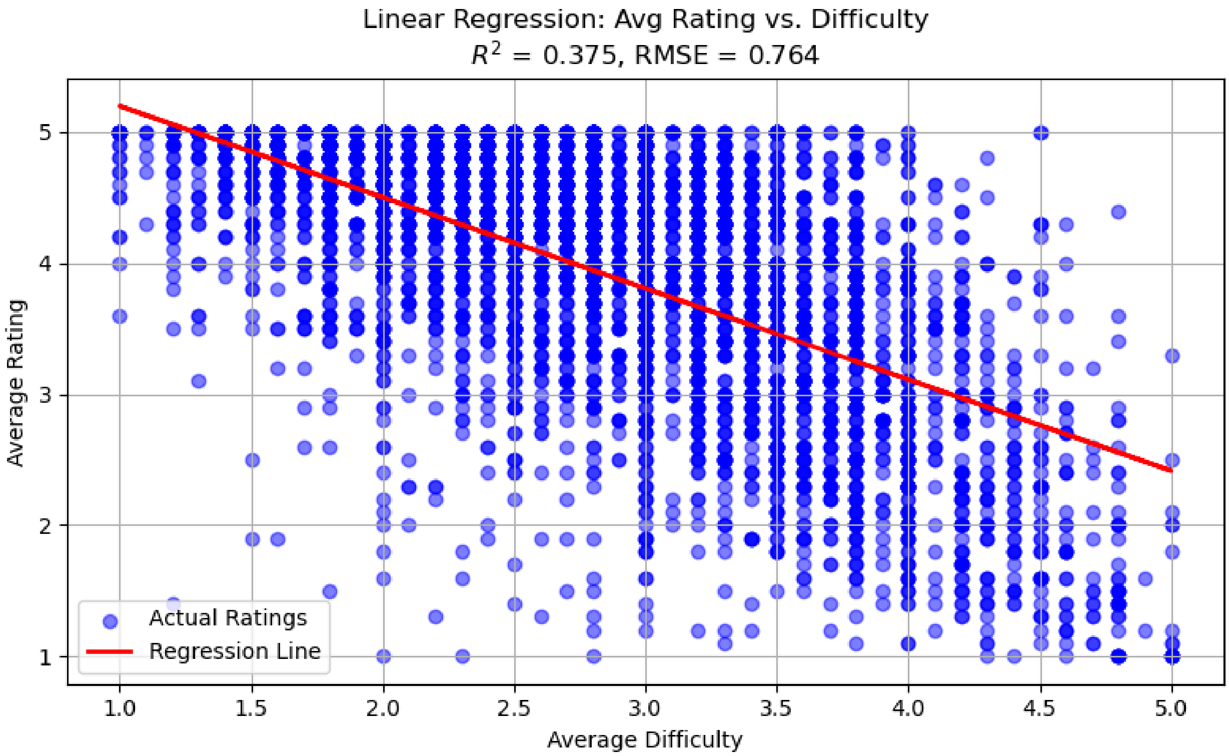
Additional Data Filtering

To ensure that the data is consistent for the last 4 problems, additional data filtering was done. First, all the rows with Nan values in the “Would Take Again” column were removed. Also, to address multicollinearity concerns, I decided to drop one of the gender columns because one is generally the inverse of the other. To ensure proper data, the rows in which the values in both the male and female columns were either 0 or 1 were also dropped.

Question 7: Predicting Average Rating from Difficulty(Only)

In the train-test split, 80% of the data was used for training while the rest was used for testing. A simple regression model was used to predict Average Rating from solely Average Difficulty. According to the model, a one unit increase in Average Difficulty leads to a 0.69 unit decrease in Average Rating. The negative relationship between Average Rating and Difficulty was clear to see from the regression line.

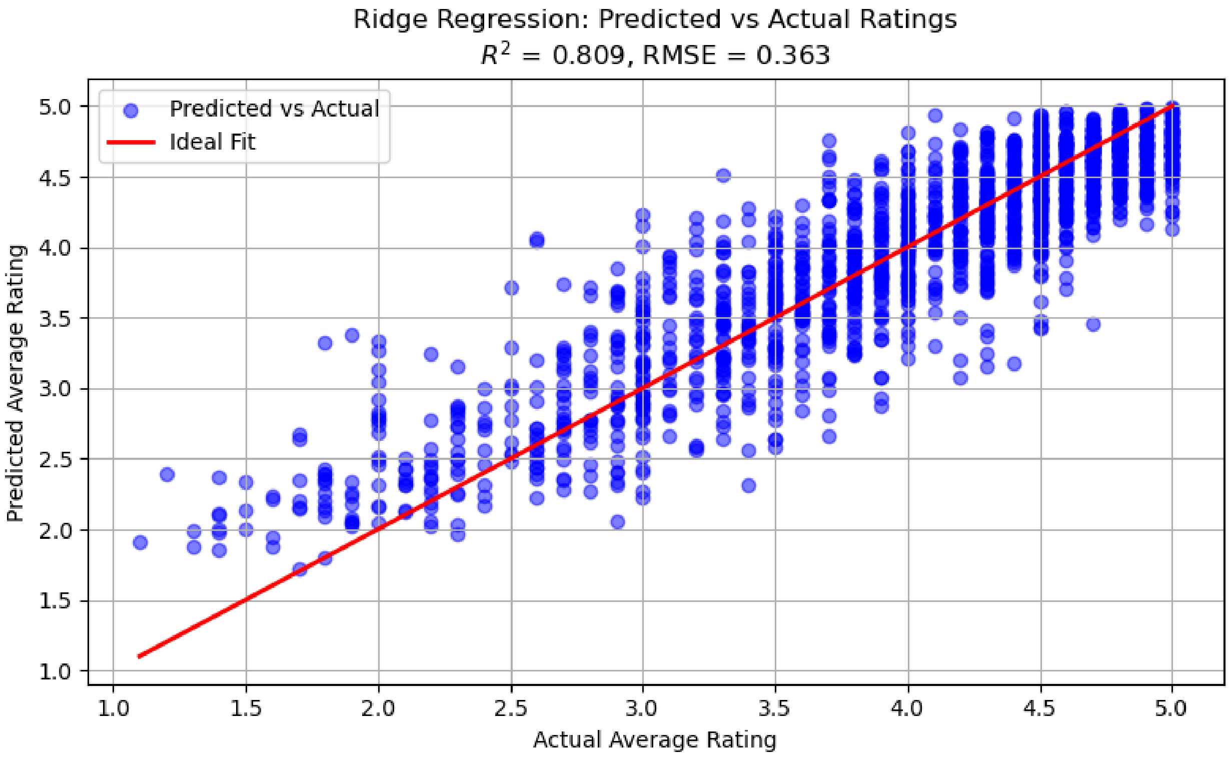
The R^2 was found to be 0.375 which is a very low percentage of variance in Average Rating explained by just Average Difficulty. The RMSE of this model is 0.764 , reflecting a pretty bad prediction error in rating units. Together, these metrics suggest the model has very limited explanatory power.



Question 8: Predicting Average Rating from All Metrics

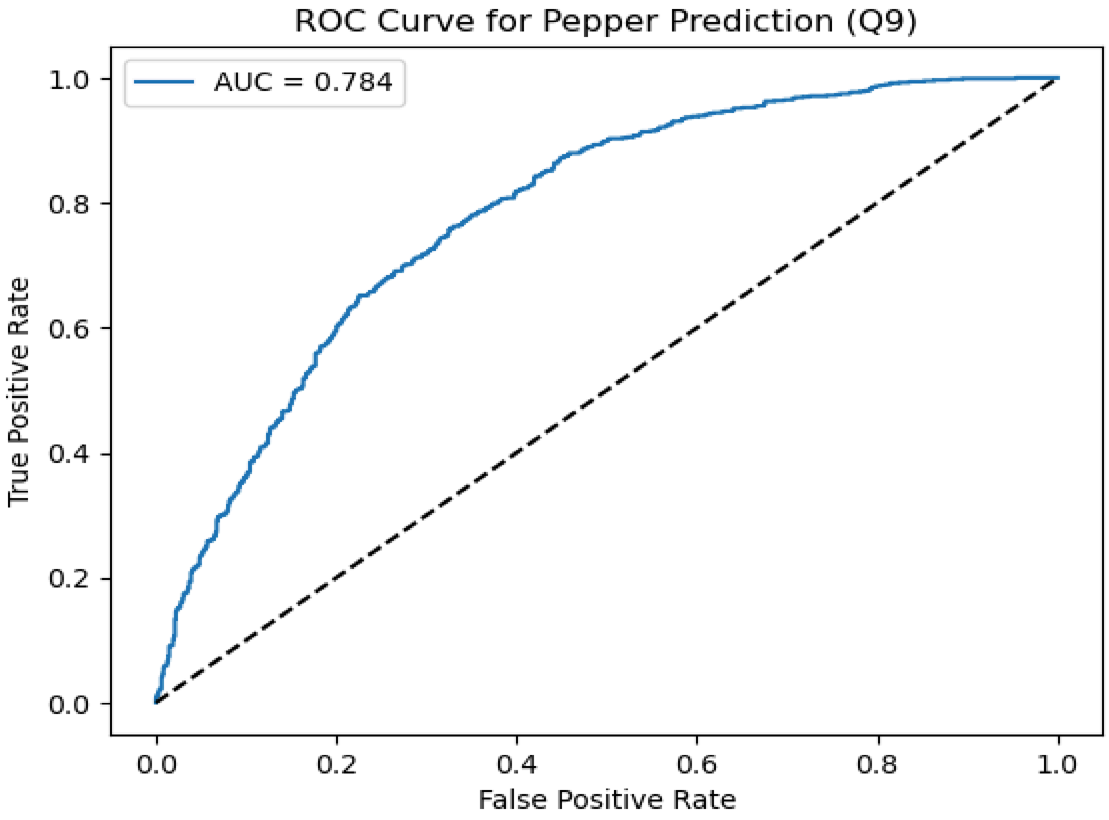
In this question, a ridge linear regression model was built to predict the Average Rating using six from Average Difficulty, Number of Ratings, Pepper, Would Take Again, Online Count, and Male. Ridge Regression was used to prevent overfitting and further address multicollinearity concerns. The data was split using a 80/20 train-test split with my N-number as the random seed.

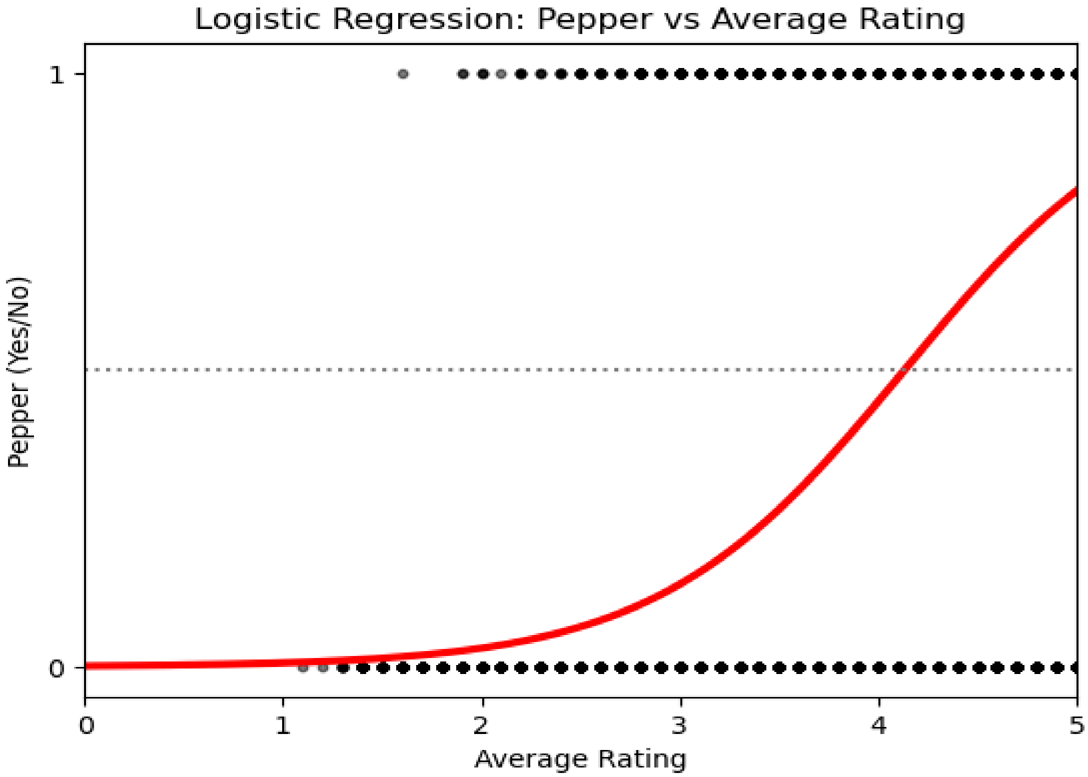
The R^2 of this model is 0.809 and the RMSE is 0.363.The scatter plot of actual vs predicted ratings showed that most predictions align closely with actual values, which is reinforced by a reasonably high R^2 and low RMSE. It is clear that the addition of other predictors significantly improves the model's performance. Overall, the model demonstrates strong prediction capability and suggests that the features contribute meaningfully to predicting ratings.



Question 9: Predicting Pepper from Average Rating

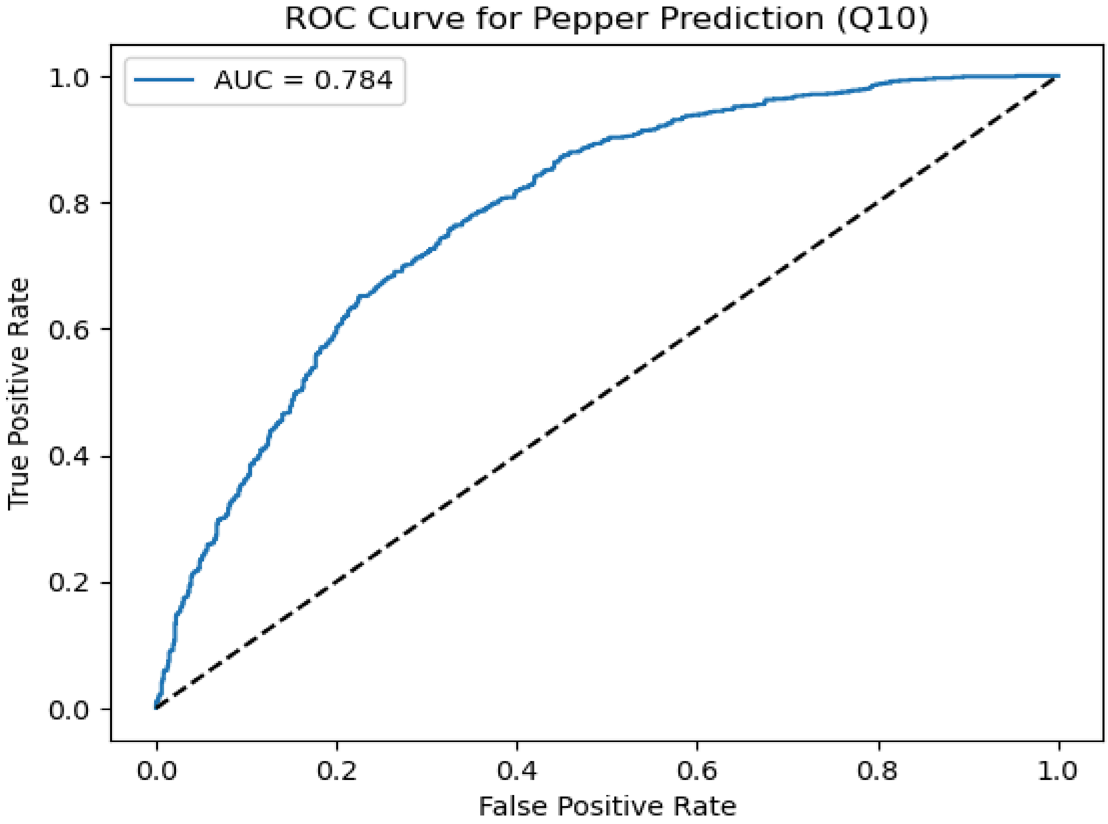
For this problem, I used a logistic regression model to predict whether a professor received pepper based on average rating. I did not have to address class imbalances in my dataset because after filtering my dataset contained 4166 professors who received a pepper and 4683 professors who didn't. The data was split using a 80/20 train-test split with my N-number as the random seed. After running the model, I plotted the ROC curve and the following metrics were calculated: Accuracy = 0.71, Specificity = 0.65, Precision = 0.66, Recall = 0.77. The area under the ROC was found to be 0.78. These results indicate that the model does a fairly good job at distinguishing between professors who received a pepper and those who didn't.


ROC Curve for Pepper Prediction (Q9)
AUC = 0.784
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Logistic Regression: Pepper vs Average Rating
Pepper (Yes/No)
Average Rating
1
0
0
1
2
3
4
5

Question 10: Predicting Pepper from All Factors

A logistic regression model to predict whether a professor received pepper based on all available factors after the additional data filtering. I did not have to address class imbalances in my dataset as mentioned above because of the similarity in the number of professors who received a pepper and those who didn't. The data was split using a 80/20 train-test split with my N-number as the random seed. After running the model, I plotted the ROC curve and the following metrics were calculated: Accuracy = 0.70, Specificity = 0.62, Precision = 0.65, Recall = 0.81. These are similar results to the ones found in the previous question that used only average rating as a predictor. The area under the ROC curve was 0.78, which is only a slight improvement on the previous model, indicating similar performance.



Extra Credit: Which Courses tend to be harder on average: Math or Computer Science?

Although specific course data is not available, we can use the professor's field as a proxy for the types of courses they typically teach. This proxy enables us to estimate and compare the perceived difficulty of different academic disciplines based on student ratings. However, it has limitations: it does not account for variation in course levels and ratings may be disproportionately influenced by more commonly taken lower-division courses, which could bias perceived difficulty downward.

A dataset of professors that teach CS courses and a dataset of professors that teach Math courses was created using the Major/Field column in the qualitative dataset. The median rating for Average Difficulty for professors who teach Math and CS courses was found to be 3.1 and 3.0 respectively. A Mann-Whitney test was performed. The u statistic was found to be 1696346.0 and the p value for that test statistic is approximately 0.15. Hence we cannot reject the null hypothesis that courses of both fields tend to be of similar difficulty.