CAPSTONE PROJECT

Name : Arnav Shah
Professor : Pascal Wallisch
Principles of Data Science
8th December 2024

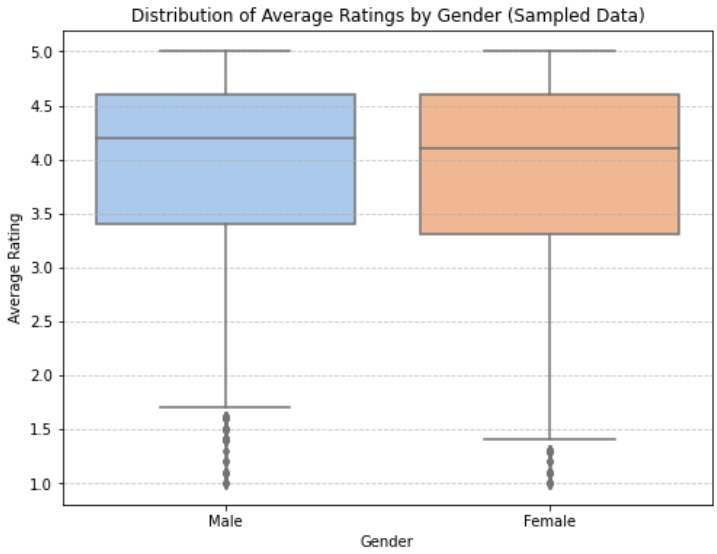
Preprocessing of Data:

In the preprocessing phase, we dropped all rows with NaN values in all columns except the "Proportion of Students Retaking" column to ensure sufficient data for analysis. For the "Proportion of Students Retaking" column, we replaced NaN values with the median to prevent the loss of valuable rows while mitigating the impact of outliers on the analysis. Additionally, we implemented a threshold of $k \geq 5$ for the "Number of Ratings," excluding professors with fewer than five ratings to reduce the influence of unreliable averages. These steps ensured a robust and clean dataset for meaningful analysis. However, for Q5 we had to clean the data differently since it involved using the "Proportion of Students Retaking" column. We did not use the median values and instead just dropped the NaN rows and conducted our analysis on the remaining data of the column. I also seeded the random number generator (RNG) with my N-number in the beginning of the code. For the extra credit question we linked the 2 data files and aggregated the fields into broader categories like "STEM," "Humanities," and "Arts" for grouped analysis.

Question and Answers:

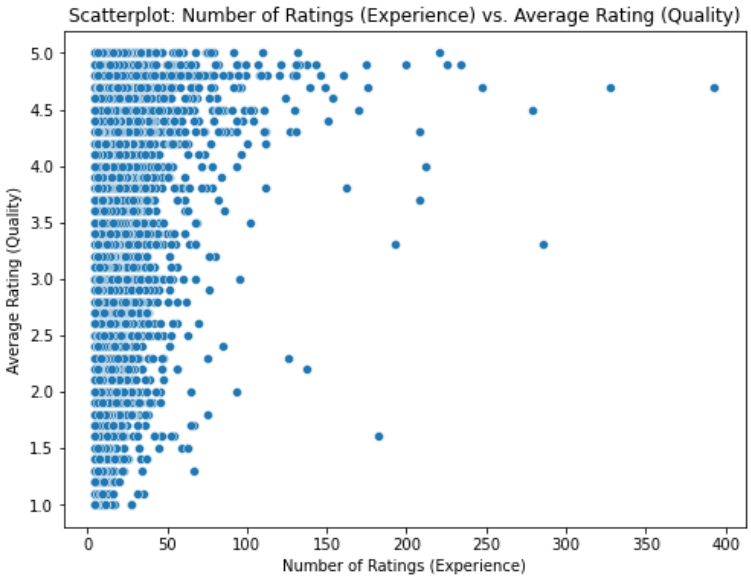
Question 1: Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality. We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset.

Answer: We analyzed whether there is evidence of a pro-male bias in professor ratings using the **Mann-Whitney U test**, which is suitable for non-normal data distributions. To address the potential issue of p-value inaccuracies with large sample sizes, we randomly sampled 2500 data points each from male and female professor ratings. The test revealed **no statistically significant difference in ratings** (**$U = 3.20 \times 10^6$**, **p-value = 0.1456**), indicating that male professors do not receive higher median ratings than female professors. The boxplot supports this finding, showing nearly identical distributions of ratings between the two groups in the sampled data. This result suggests that there is no evidence of a pro-male bias in the dataset.



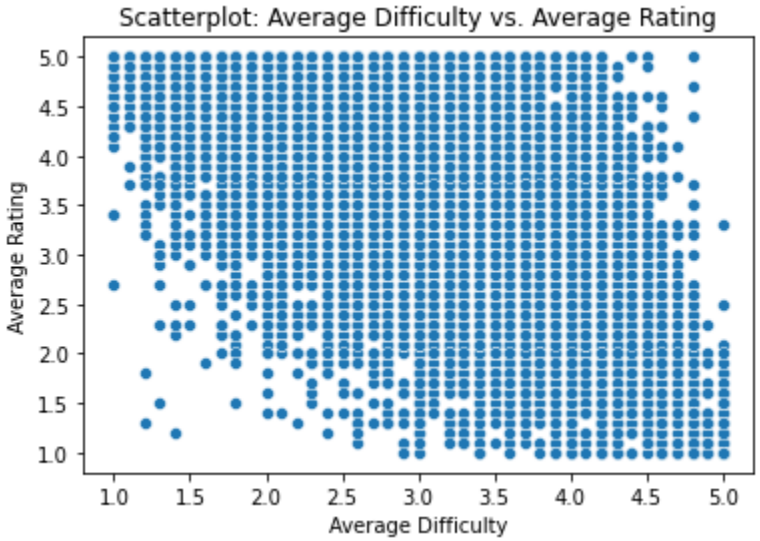
Question 2: Is there an effect of experience on the quality of teaching? You can operationalize quality with average rating and use the number of ratings as an imperfect – but available – proxy for experience. Again, a significance test is probably a good idea.

Answer: We analyzed the relationship between Number of Ratings (proxy for experience) and Average Rating (quality of teaching) using Spearman's correlation coefficient to account for the non-linear, monotonic relationship observed in the scatterplot. **Spearman's correlation coefficient (ρ)** was **0.0282** (**p-value = 6.92 x 10^{-6}**), highly significant), indicating a **very weak positive monotonic relationship** between experience and teaching quality. While the correlation is statistically significant, its small magnitude suggests that the effect of experience on teaching quality is minimal. The scatterplot illustrates this weak trend.



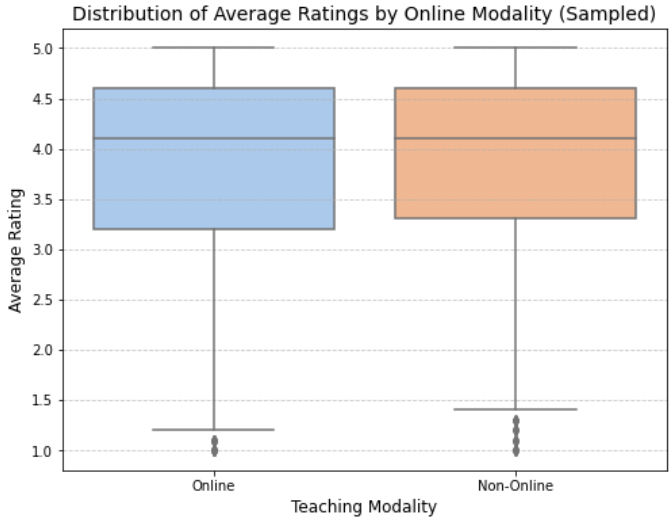
Question 3: What is the relationship between average rating and average difficulty?

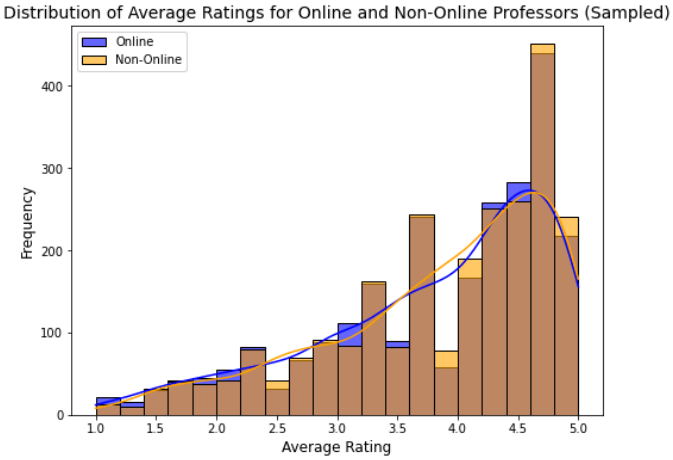
Answer: We analyzed the relationship between Average Rating and Average Difficulty using Spearman's correlation coefficient because the scatterplot indicated a non-linear, monotonic relationship. **Spearman's correlation coefficient (ρ)** was **-0.6017** (**p-value < 10^{-10}**, highly significant), indicating a **strong negative monotonic relationship**: as difficulty increases, ratings tend to decrease. This suggests that professors perceived as more difficult tend to receive lower ratings. Spearman's test was appropriate due to the non-linear relationship in the data, but it only captures monotonic trends and may lose sensitivity with tied ranks in datasets with discrete values.



Question 4: Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't?

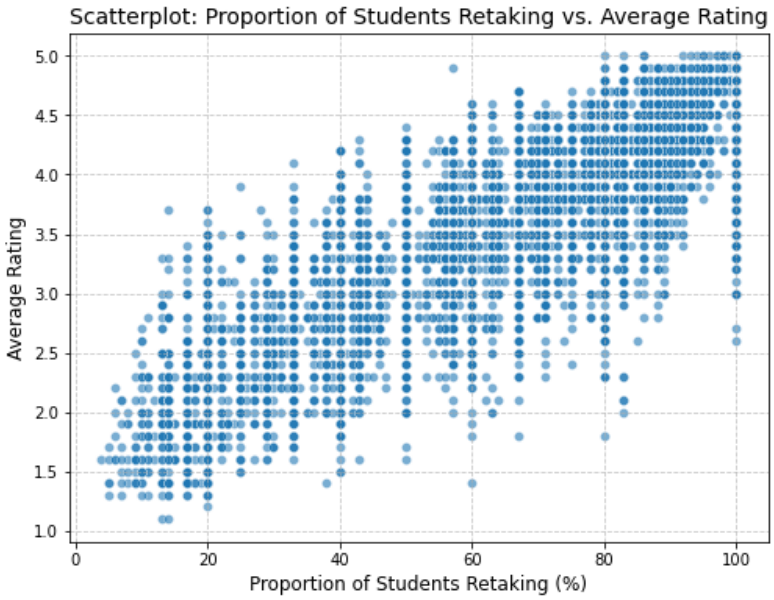
Answer: We used the **Mann-Whitney U test** to compare ratings between professors teaching many online classes and those teaching primarily non-online. To ensure reliable p-values and reduce bias from the large dataset, we randomly sampled 2,500 professors from each group. The test results **($U = 3.06 \times 10^6$, p-value = 0.229)** showed **no statistically significant difference in ratings.** The boxplot further supports this, with similar medians and overlapping distributions. This suggests that teaching modality (online vs. non-online) does not significantly impact average ratings.





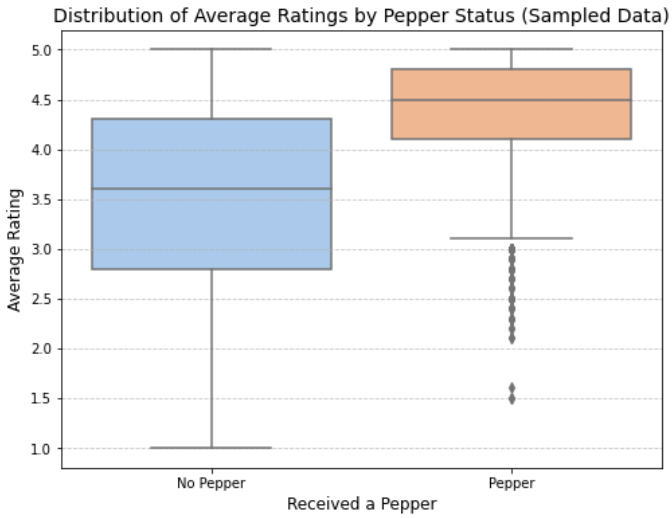
Question 5: What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?

Answer: We analyzed the relationship between the proportion of students retaking a class and the average rating using Spearman's correlation coefficient, as the scatterplot indicated a monotonic relationship rather than strictly linear. The analysis revealed a **strong positive correlation** ($\rho = 0.8522$, **p-value** $< 10^{-4}$), suggesting that as the proportion of students willing to retake a class increases, the average rating of the professor also tends to increase. This indicates that professors with higher ratings are more likely to have students willing to retake their classes. The scatterplot further supports this strong monotonic relationship, and the statistical significance (p-value < 0.005) confirms that the observed pattern is unlikely to be due to chance.



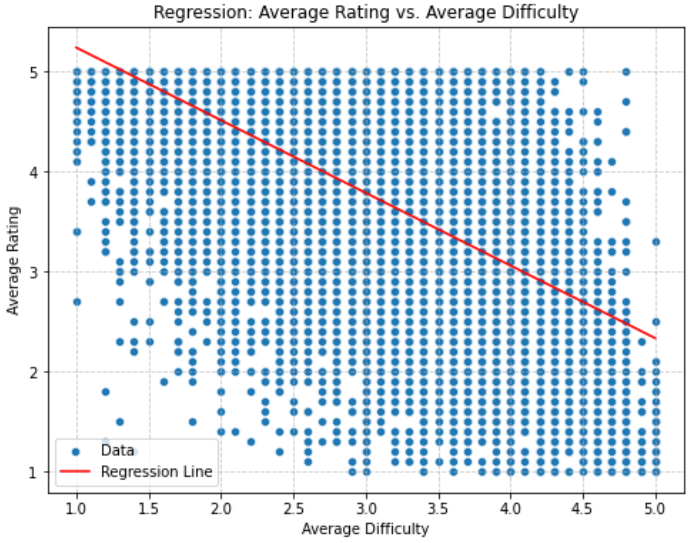
Question 6: Do professors who are "hot" receive higher ratings than those who are not?

Answer: We analyzed whether professors who are perceived as "hot" (received a pepper) receive higher ratings using the Mann-Whitney U test, a non-parametric method suitable for non-normal data distributions. To address concerns about the test's sensitivity to large datasets, we randomly sampled 5000 data points. The test revealed a **statistically significant difference** in ratings (**$U = 4.67 \times 10^{6}$, p-value $< 10^{-10}$**), with professors who received a pepper having higher median ratings compared to those who didn't. The boxplot supports this finding, showing an upward shift in the distribution of ratings for professors with a pepper. While the Mann-Whitney U test can detect statistically significant differences even with small practical effects, random sampling helps mitigate this issue and ensures the results are robust.



Question 7: Build a regression model predicting average rating from difficulty (only). Make sure to include the R2 and RMSE of this model.

Answer: We developed a linear regression model to predict Average Rating based on Average Difficulty. The model shows a **negative linear relationship**, as indicated by the regression line in the figure, with a **slope (coefficient) of -0.7271** and a **y-intercept of 5.9692**. The model's **R^2** value of **0.3832** suggests that about 39.12% of the variation in average ratings is explained by average difficulty. The **RMSE of 0.7437** indicates the average difference between the predicted and actual ratings, showing moderate prediction accuracy. The findings indicate a moderately strong inverse relationship: as difficulty increases, ratings tend to decrease. However, the model's limited R^2 value suggests other factors beyond difficulty likely influence ratings.

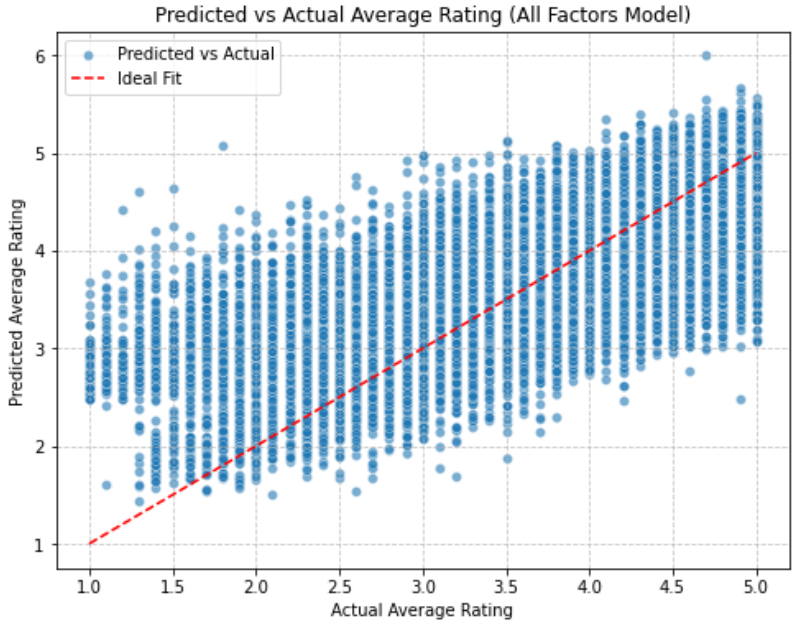


Question 8: Build a regression model predicting average rating from all available factors. Make sure to include the R2 and RMSE of this model. Comment on how this model compares to the "difficulty only" model and on individual betas. Hint: Make sure to address collinearity concerns.

Answer: We built a regression model predicting Average Rating using all available factors: Average Difficulty, Number of Ratings, Received a Pepper, Proportion of Students Retaking, Number of Online Ratings, Male, and Female. The model achieved an **R^2 value of 0.5607** and an **RMSE of 0.6277**, outperforming the "difficulty only" model **(R^2 = 0.3832, RMSE = 0.7437)**. This highlights the value of including multiple predictors for better accuracy.

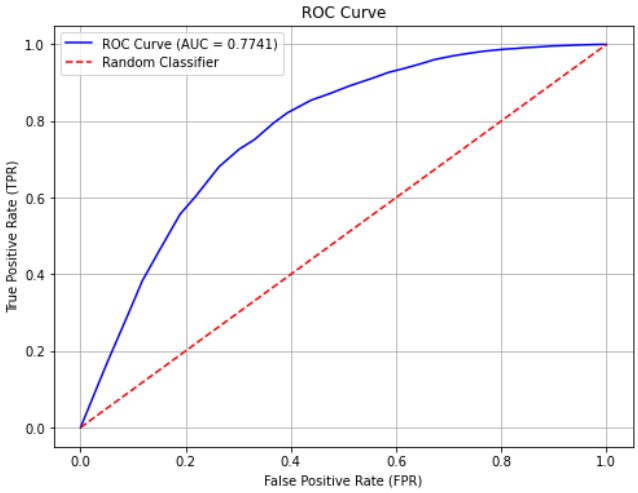
Key coefficients show Average Difficulty has the largest negative impact (−0.5360), while Received a Pepper (+0.4968) and Proportion of Students Retaking (+0.0157) have notable positive effects. Smaller effects are observed for Number of Ratings (+0.0036), Male (+0.1061), and Female (+0.0585), with a slight negative impact from Number of Online Ratings (−0.0072). The intercept of 3.8448 reflects the baseline rating when predictors are zero.

The **Variance Inflation Factor** (VIF) confirmed **no multicollinearity concerns**, with all values below 5. This model highlights the multifaceted nature of ratings, where difficulty, perceived professor quality, and student preferences all play significant roles.



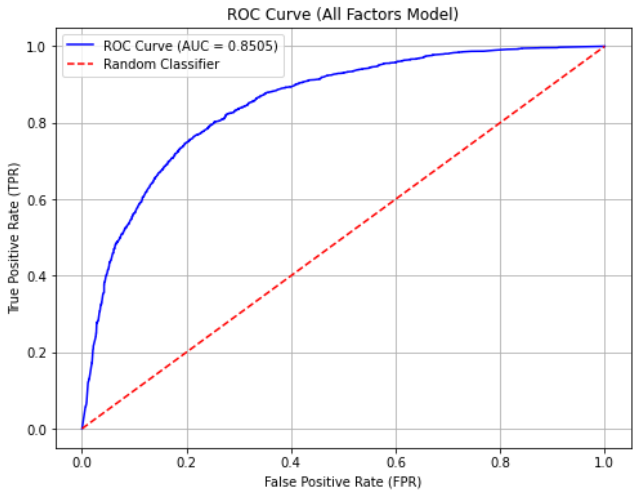
Question 9: Build a classification model that predicts whether a professor receives a “pepper” from average rating only. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

Answer: We developed a classification model to predict whether a professor receives a "pepper" based solely on their average rating, addressing class imbalance through oversampling. The **model achieved an accuracy of 71%**, with a **precision of 76% for the majority class (no pepper) and 68% for the minority class (pepper)**. Recall was higher for the minority class at 80%, compared to 63% for the majority, resulting in balanced F1-scores of 0.69 and 0.73, respectively. The **ROC-AUC score** of **0.7741** indicates good discriminatory ability between the two classes. The confusion matrix highlights that the model performs reasonably well in identifying both classes, though some misclassifications remain. Overall, this model demonstrates that average rating is a moderately strong predictor of "pepper" status.



Question 10: Build a classification model that predicts whether a professor receives a “pepper” from all available factors. Comment on how this model compares to the “average rating only” model. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

Answer: We developed a classification model using all available factors to predict whether a professor receives a "pepper," addressing class imbalances through oversampling. The model achieved an **AU(RO)C** of **0.8505**, significantly outperforming the “average rating only” model (AU(RO)C = 0.7741). It also demonstrated an **accuracy of 77%, with strong precision (81%) for non-pepper predictions and recall (82%) for pepper predictions**. The improved performance highlights the importance of incorporating additional factors like class difficulty, student retake rates, and online teaching modality, which contribute to more accurate and robust predictions than the simpler average rating-only model.



Extra Credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.

Answer: We examined the relationship between academic fields and professor ratings by grouping data into the top 10 fields and broader academic field groups. We calculated central tendency measures for each group and used boxplots to visualize the distributions. This allowed us to identify both general trends and nuanced differences in ratings across disciplines.

1. Top 10 Academic Fields: **Median ratings ranged from 4.0 to 4.2**, with Education and Computer Science leading. Variability within fields was low (**standard deviations ~0.9**), showing consistent satisfaction levels across disciplines.
2. Academic Field Groups: **Arts had the highest median rating of 4.2**, while other groups like STEM, Social Sciences, and Humanities clustered at 4.1. **Standard deviations (~0.96) indicate uniformity with minor variations.**