Capstone Project report (Ziyuan Kang/ zk2160)

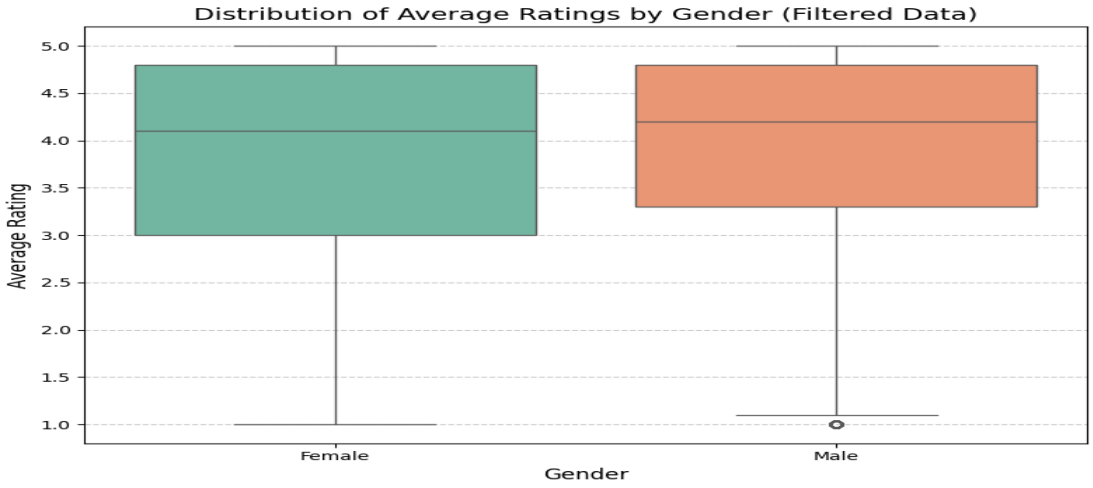
Data preparation and cleaning: The project starts with data cleaning and preprocessing. Two datasets, "rmpCapstoneNum" and "rmpCapstoneQual," are combined to include all relevant information, and column headers are created. Missing values are handled in two steps: rows with more than 50% missing data are dropped, and numerical columns (excluding binary categories like male, female, pepper) have missing values filled with their median. A random seed (12428625) (own N number) is set for reproducibility, and a significance level of 0.005 is used throughout.

Q1: Whether there is gender bias (pro-male) in ratings?

Methodology: Rows where both male and female are 0 or 1 are removed to ensure clear gender classification. Mann-Whitney U Test is used to compare average ratings between male and female professors.

Results: Mann-Whitney U Test: the **U-statistic is 346129258.5** and the **p-value is 4.2736e-06**

Visualization:



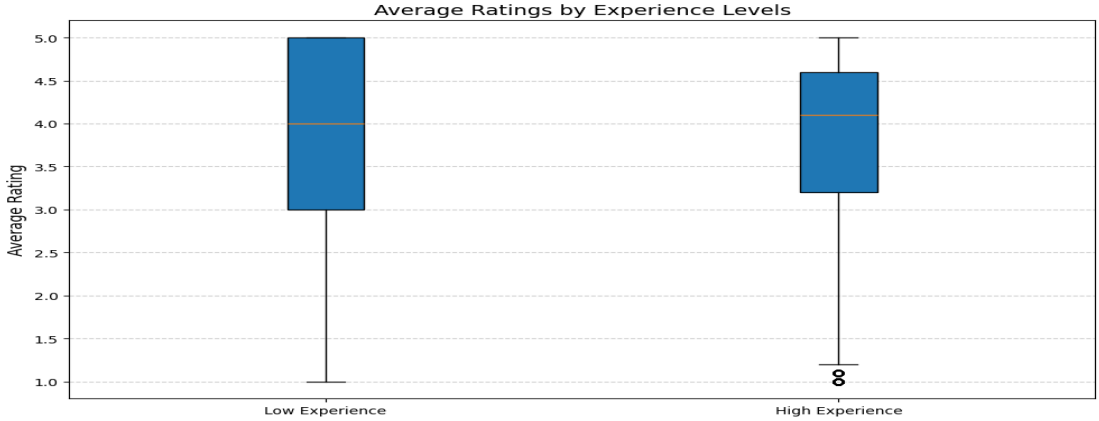
Interpretation: Since p-value is smaller than the significant level 0.005, we reject the null hypothesis and can conclude that there is evidence of pro-male gender bias in the ratings.

Q2: Does experience affect the quality of ratings?

Methodology: Professors are divided into two groups based on the median of "num_ratings". Mann-Whitney U Test compares ratings between low and high experience groups.

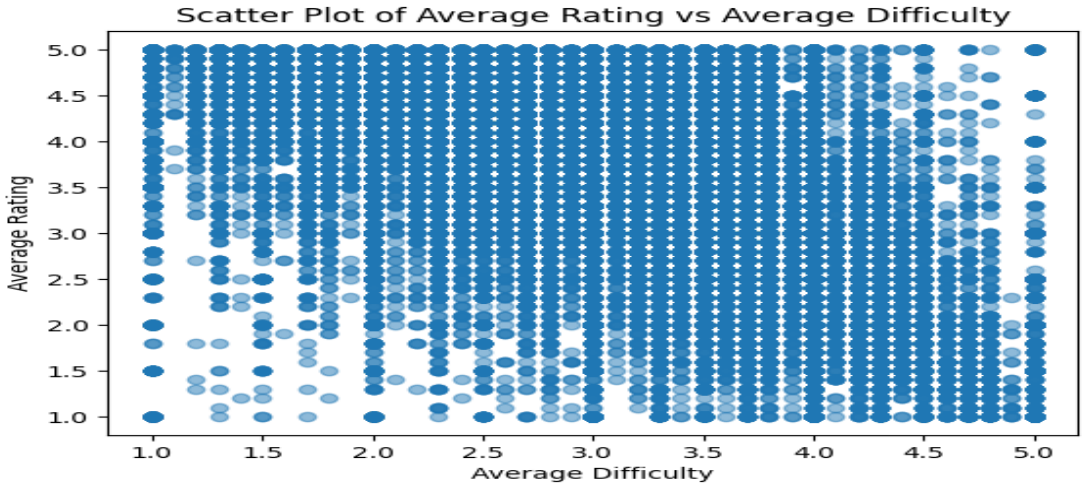
Results: Median: **3.0;** The **U-statistic is 648292604.5** and the **p-value is 2.5415e-52**

Visualization:



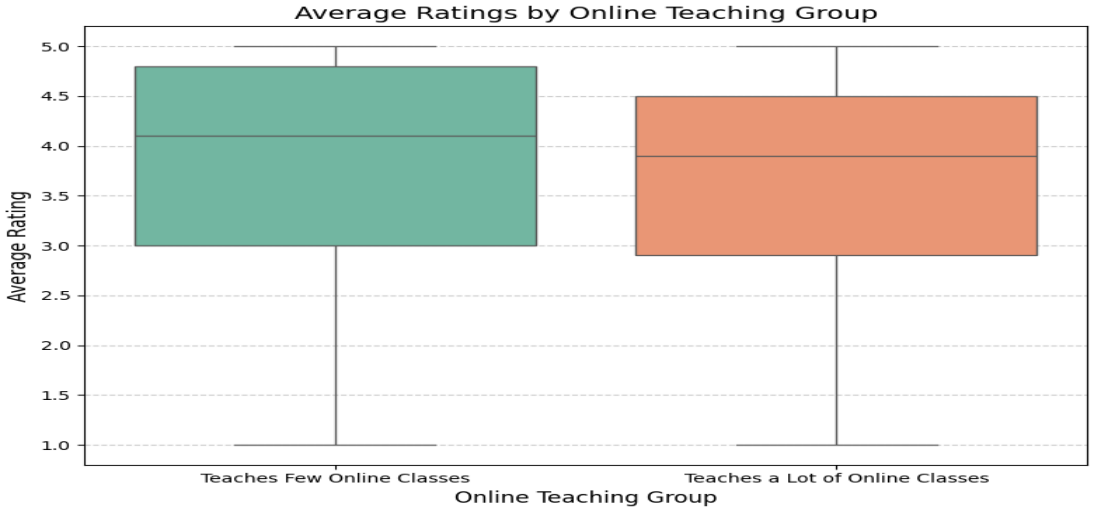
Interpretation: Since p-value is smaller than the significant level 0.005, we reject the null hypothesis and can conclude that experience has a significant effect on teaching quality. Professors with higher experience tend to receive higher ratings.

Q3: To determine the relationship between average rating and average difficulty.
Methodology: A scatter plot visualizes the relationship. Spearman correlation coefficient quantifies monotonic relationships. For linear relationships, linear regression provides a slope.
Results: Spearman Correlation **Coefficient is -0.5114;** Linear regression **Slope (β1) is -0.6103**
Visualization:



Interpretation: Since both spearman correlation coefficient and slope of linear regression showed an obvious negative relationship, we can conclude that the average ratings of professors decreasing as average difficulty increases.

Q4: Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't?
Methodology: Descriptive statistics and scatter plots analyze “num_online_ratings”. Professors in the top 5% of “num_online_ratings” form the high online teaching group, while others form the low online teaching group. Mann-Whitney U Test compares ratings between groups.
Results: Threshold **(top 5%) is 2**; **U-statistic is 131753289** and **p-value is 1.8794e-42**
Visualization:



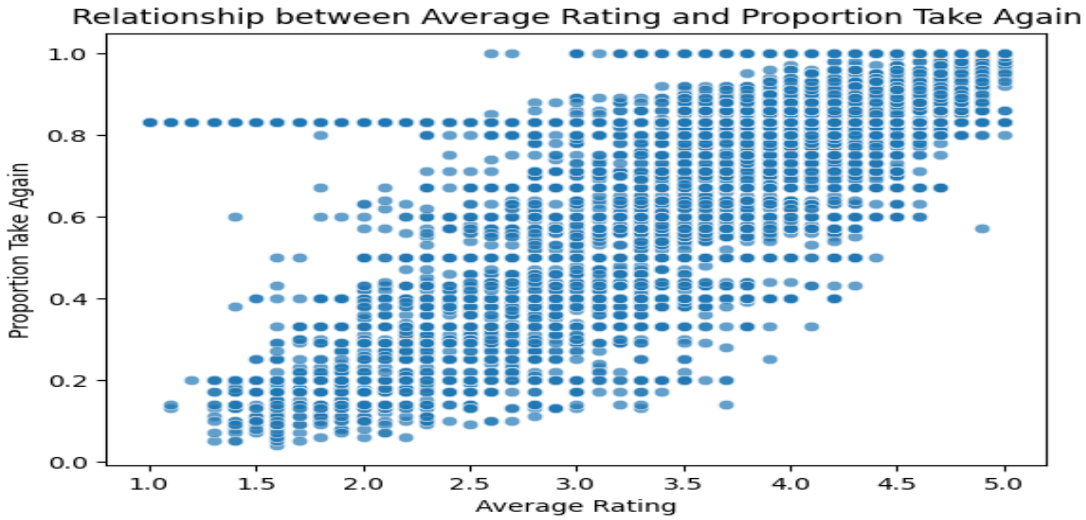
Interpretation: Since p-value is smaller than the significant level 0.005, we reject the null hypothesis and can conclude that there is a significant difference in ratings between the two groups. Professors who teach a lot of online classes tend to receive lower ratings.

Q5: Find the relationship between average rating and proportion of people who would take the class again.

Methodology: A scatter plot visualizes the relationship. Descriptive statistics describes the data distribution of two variables. Spearman correlation coefficient quantifies monotonic trends.

Results: The Spearman Correlation **Coefficient is 0.2574**

Visualization:



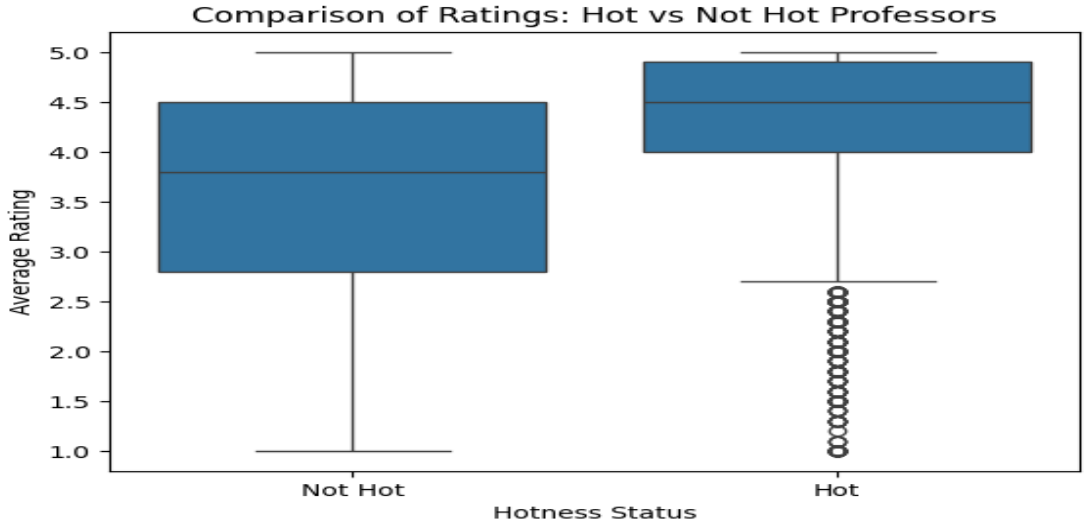
Interpretation: There is a positive relationship: as average ratings increase, the proportion of students willing to take the class again also increases.

Q6: Do professors who are "hot" receive higher ratings than those who are not?

Methodology: Professors are divided into hot (pepper = 1) and not hot (pepper = 0). Mann-Whitney U Test compares ratings.

Results: The **U-statistic is 684243055.5** and **p-value is 0.0000e+00**

Visualization:



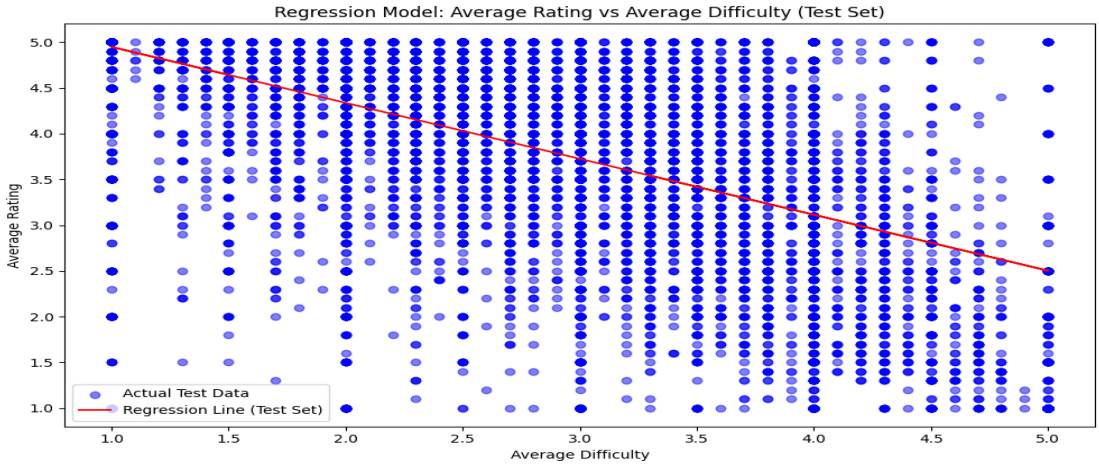
Interpretation: Since p-value is smaller than the significant level 0.005, we reject the null hypothesis and can conclude that there is a significant difference in ratings between the two groups. Hot professors receive higher ratings than those who are not hot.

Q7: Build a regression model predicting average rating from difficulty.

Methodology: Data is split into training (80%) and testing (20%) sets. Linear regression uses "ave_difficulty" as the predictor. Performance is assessed using R^2 and RMSE.

Results: Intercept **(β0): 5.5593**, Coefficient **(β1): -0.6114**; Training: **R^2: 0.2889, RMSE: 0.9507**; Testing: **R^2: 0.2849, RMSE: 0.9511**

Visualization:



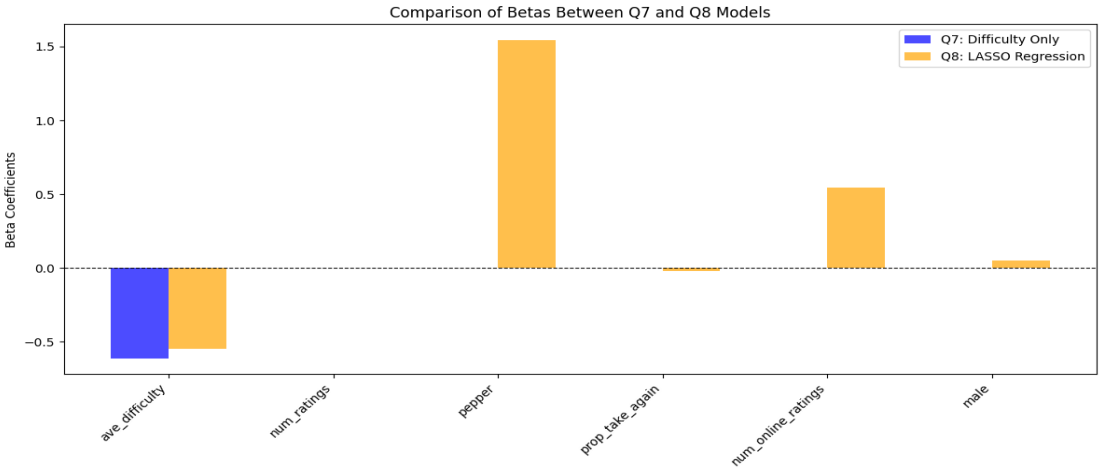
Interpretation: Difficulty negatively affects ratings, but the low R^2 suggests difficulty alone is not a strong predictor.

Q8: Build a regression model predicting average rating using all available factors, and compare it to the "difficulty only" model.

Methodology: Rows with NaN values in binary categories are removed. The "female" column is dropped to address collinearity. Ridge regression handles multicollinearity and includes multiple predictors. Performance is assessed using R^2 and RMSE.

Results: Coefficients: [**-0.5477, 0.0034, 1.5435, -0.0224, 0.5449, 0.0533**]; Training: **R^2: 0.3697, RMSE: 0.8862**; Testing: **R^2: 0.3470, RMSE: 0.8971**

Visualization:



Interpretation: Ridge regression adjusts the difficulty coefficient. The Ridge model improved both R^2 and RMSE compared to the simpler model. This demonstrates that including additional predictors significantly enhances model performance.

Q9: To build a classification model that predicts whether a professor receives a "pepper" (hotness) based on average rating only.

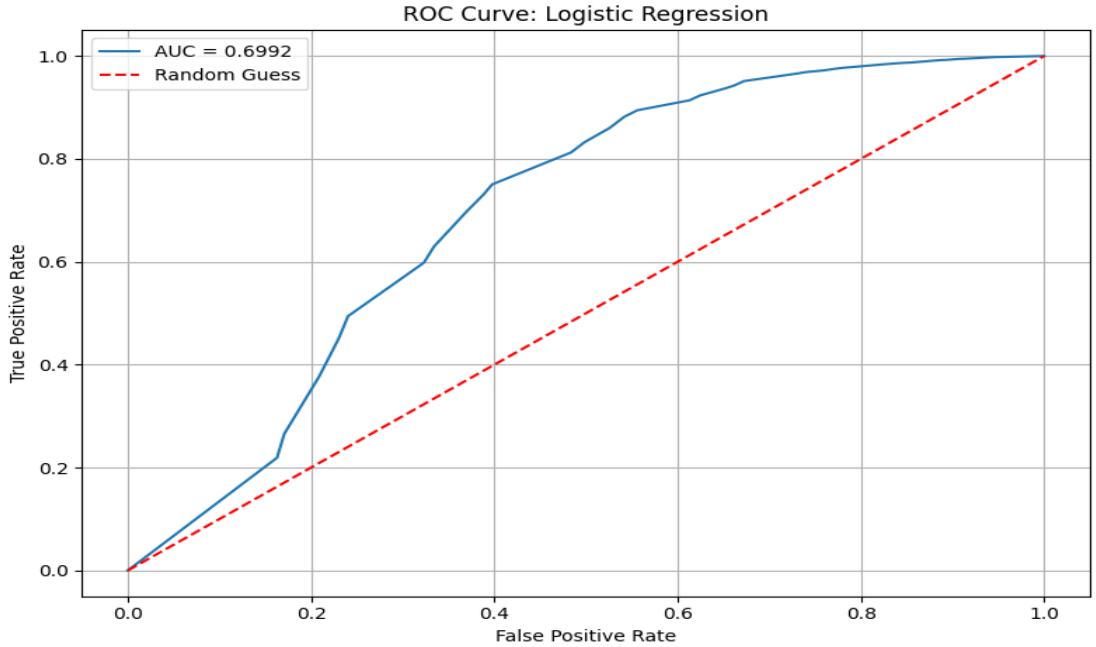
Methodology: Data is cleaned to ensure no NaN values in "ave_rating" or "pepper." Logistic regression predicts the binary target variable. Performance is measured using classification metrics and ROC-AUC.

Results: Class 0 (No Pepper): **Precision:** 0.86, **Recall:** 0.60, **F1-Score:** 0.71;

Class 1 (Pepper): **Precision:** 0.42, **Recall:** 0.75, **F1-Score:** 0.54;

Accuracy: 0.64; **ROC-AUC** score is **0.6992**

Visualization:



Interpretation: The model performs better for Class 0 due to higher precision and F1-score. This reflects the dominance of Class 0 in the dataset, making it easier for the model to predict correctly. The ROC-AUC of 0.6992 is moderate, suggesting the model can distinguish between the two classes better than random guessing (AUC = 0.5). However, there's room for improvement.

Q10: Build a classification model that predicts whether a professor receives a "pepper" from all available factors.

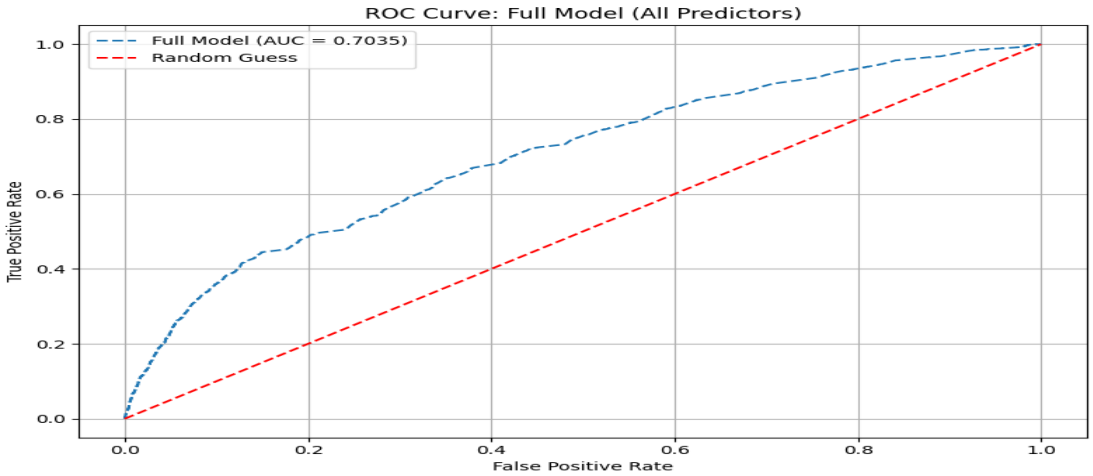
Methodology: Data is filtered to remove rows with NaN values in binary categories and redundant gender columns. Logistic regression includes multiple predictors. Performance is measured using classification metrics and ROC-AUC.

Results: Class 0 (No Pepper): **Precision:** 0.81, **Recall:** 0.72, **F1-Score:** 0.76;

Class 1 (Pepper): **Precision:** 0.43, **Recall:** 0.56, **F1-Score:** 0.49;

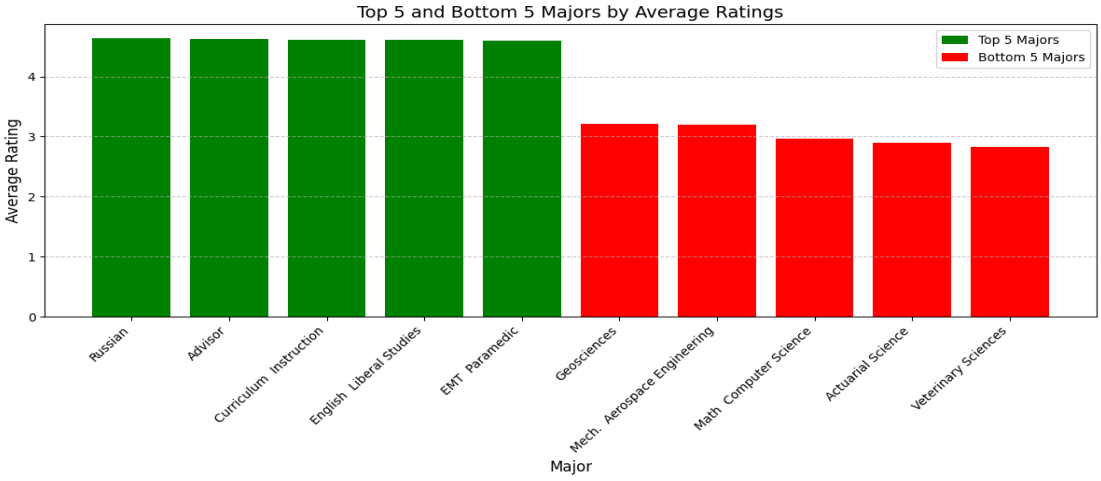
Accuracy: 0.67; **ROC-AUC** score is **0.7035**

Visualization:



Interpretation: The full model (question 10) improves overall accuracy and ROC-AUC compared to the baseline model (question 9). However, lower recall for Class 1 in full model suggests it is more conservative in predicting “pepper” cases.

Extra Credit: Analyze the relationship between professor majors and their average ratings.
Methodology:
Data is grouped by professor majors, calculating the average rating and professor count for each major. Descriptive statistics for the professor count per major are also computed. Majors with fewer than 9 professors are excluded to focus on majors with sufficient data for meaningful comparisons. The top 5 and bottom 5 majors are analyzed based on their average ratings.
Results: Number of unique majors: **1272;** Majors with fewer than 9 professors: **953; Top 5** Average Ratings: **4.635, 4.624, 4.608, 4.606, 4.593;** Bottom 5 Average Ratings: **3.207, 3.200, 2.967, 2.900, 2.833**
Visiualization:



Interpretation:
The analysis identifies the most and least "popular" majors based on average professor ratings. Majors like Russian and Advisor rank highest, suggesting strong student approval, while Veterinary Sciences, Actuarial Science, and Math Computer Science rank lowest, potentially reflecting challenges like course difficulty or engagement. However, biases such as uneven sample sizes, varying student expectations, and subjective factors may influence these findings.