# DSGA 1001 Capstone Project: *Assessing Professor Effectiveness*

Group Name: Team Journal Squared

**Authors**
Jiwoo Jeong (jj4252)
Dayne Lee (dl5635)

**Contributions (Apply to both members of the group)**

1. Data Preprocessing
    - Implemented data preprocessing, including data filtering (with appropriate threshold), normalization, and handling missing values to ensure consistency and quality of the dataset.
2. Hypothesis Testing and Modeling
    - Conducted hypothesis testing for relevant questions and built linear and logistic regression models for data modeling.
    - Interpreted the results of statistical tests and models to derive proper conclusions.
    - Drafted a structured final report in adherence to the project guidelines.
3. Python Codes
    - Wrote Python code for hypothesis testing, statistical analysis, data visualizations, and machine learning models to address the project questions.

**Data Preprocessing**

The following steps were taken to clean and structure the data:

1. To enhance clarity and ease of use, all columns were assigned descriptive names corresponding to their content (e.g column 5 in rmpCapstoneNum.csv was renamed to “prop_take_again”).
2. Rows with missing values in “average_ratings” were removed.
    a. Missing values in column 5 (proportion of students willing to take the class again) were handled only for questions where this variable was relevant.
3. “Gender” column was created by mapping binary indicators (male/female) to categorical labels (“Male” and “Female”)
    a. Rows with conflicting gender data (i.e both Male, Female columns equal 0 or 1) were excluded
4. Professors with fewer than 4 ratings were removed to maintain statistical reliability in the results while ensuring each gender has at least 10,000 samples to provide a robust basis for analysis.

** We used RNG based on the N-number of Jiwoo Jeong (jj4252) - seed value of 18038726

## Q1.

### Approach

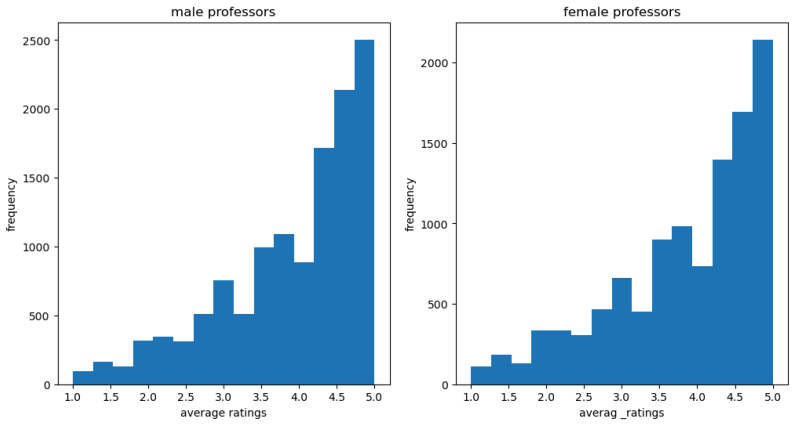
We investigated whether there is evidence of a pro-male gender bias in student evaluations of professors by comparing the average ratings of male and female professors. To do this, we conducted the Welch's t-test as this test accounts for potential differences in variances between the two groups. Although the original ratings are ordinal categorical data, the use of average ratings as a summary measure assumes equal intervals between the rating levels, approximating continuous data. This, combined with sufficiently large sample sizes, allows for the use of the t-test, even if the distributions of the averages appear skewed and not strictly normal.

- **Null Hypothesis**: There is no difference in average ratings between male and female professors.
- **Alternative Hypothesis:** The average ratings of male and female professors are different.

### Results

- Welch's t-test
    - p-value: 8.246e-07
    - Test statistic: -4.93

### Histogram of average ratings for each gender




### Conclusion

Given p-value is less than 0.005, we conclude that there is statistically significant evidence supporting the existence of gender bias in average ratings.

## Q2.

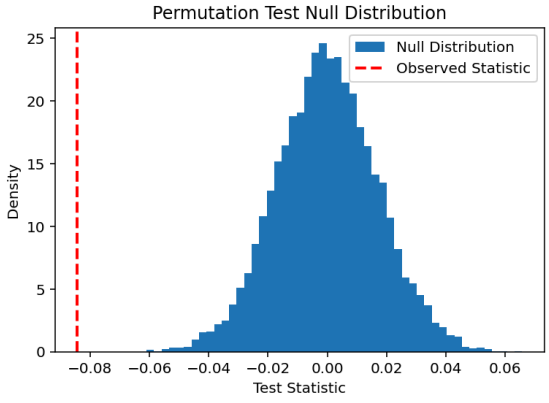
### Approach

We examined whether there is a gender difference in the spread of average ratings distribution. To test this, we used a permutation test, as this test is suitable for non-parametric data and does not rely on assumptions of normality. The test statistic is computed as the difference between the variances between genders, as the test statistic for variance does not have a standardized equivalent of Cohen's d. We generated 10,000 permutation samples to construct a null distribution.

- **Null Hypothesis:** There is no difference in the variance of average ratings between male and female professors.
- **Alternative Hypothesis:** The variances of average ratings for male and female professors are not equal.

### Results

- Permutation test with variance difference
    - p-value: 0.0002
    - Observed variance difference: 0.084



**Conclusion**

Given that the p-value is less than 0.005, we conclude that there is statistically significant evidence of gender bias in the spread of ratings.

## Q3.

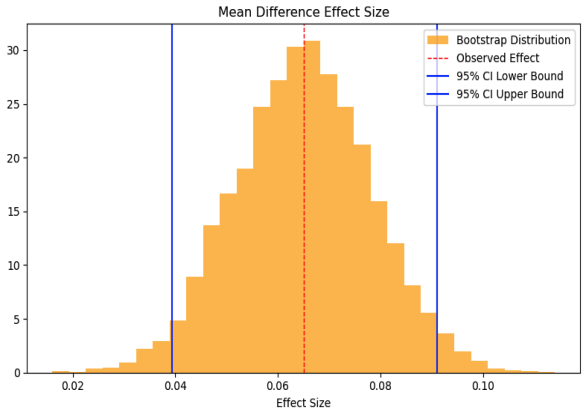
**Approach**

To estimate the size of the two effects (gender bias in average rating and in the spread of ratings, we **used resampling methods (bootstrap) to construct sampling distributions (10,000 samples)**, with 95% confidence intervals provided for each effect size.

**Results**

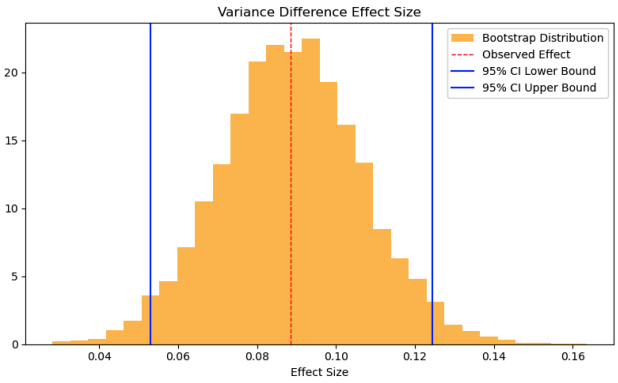
a. Gender bias in average rating

- The effect size for the difference in average ratings between male and female professors was calculated using Cohen's d.
    - Observed Effect Size: 0.065
    - 95% Confidence Interval: [0.039, 0.091]




b. Gender bias in spread of ratings

- The effect size for the difference in the spread of ratings was calculated using a measure analogous to Cohen's d. Specifically, the difference in variances between male and female professors was divided by the pooled standard deviation, providing a standardized effect size.
    - Observed Effect Size: 0.088
    - 95% Confidence Interval: [0.053, 0.124]




**Conclusion**

Both the bias in average ratings and the spread of ratings show small effect sizes, which indicate that while gender plays a role in how ratings are distributed, the magnitude of this effect is relatively small.

## Q4.

**Approach**

To investigate gender differences in the tags awarded by students, we conducted an analysis for all 20 tags. We first normalized the tag counts by dividing each tag's count by the total number of ratings a professor received, to account for variability in the number of ratings across professors. We then performed permutation tests to evaluate the significance of gender differences for each tag as this test is suitable for non-parametric data and does not rely on normality assumptions. Then, we used permutation testing to construct a sampling distribution and evaluated the significance of differences in each tag across genders using differences in normalized tag medians as a test statistic.

For each tag:

- **Null Hypothesis:** There is no difference in normalized tag counts between male and female professors.
- **Alternative Hypothesis**: The normalized tag counts differ between male and female professors.

**Results**

After applying permutation tests for all 20 tags, we calculated the p-values and identified **8 out of 20 tags that showed statistically significant gender differences (p < 0.005)**. Below are the three most and least gendered tags based on their p-values.

a. Three Most Gendered Tags

1. *"Good Feedback"* (p = 0.002)
2. "*Respected'* (p = 0.002)
3. *"Lots to Read*" (p = 0.002)

b. Three Least Gendered Tags

1. *"Lecture Heavy"* (p = 1.0)
2. "*Group Projects* (p = 1.0)
3. *"Extra Credit*" (p = 1.0)

## Q5.

**Approach**

To determine if there is a gender difference in the average difficulty ratings assigned to professors, we compared the average difficulty ratings for male and female professors using Welch's t-test, as this test accounts for the potential difference in variances between the two groups. Although original difficulty ratings are ordinal categorical data, the use of average difficulty ratings as a summary measure assumes equal intervals between the ratings levels, approximating continuous data. This, combined with sufficiently large sample sizes, allows for the use of the t-test, even if the distributions of the averages appear skewed and not strictly normal.

- **Null Hypothesis**: There is no difference in the average difficulty ratings between male and female professors.
- **Alternative Hypothesis**: There is a difference in the average difficulty ratings between male and female professors.

**Results**

a. Welch's t-test
   i. Test Statistic: -0.206
   ii. P-value: 0.836

**Conclusion**

Welch's t-test yielded non-significant results, indicating no statistically significant difference in the average difficulty ratings assigned to male and female professors.

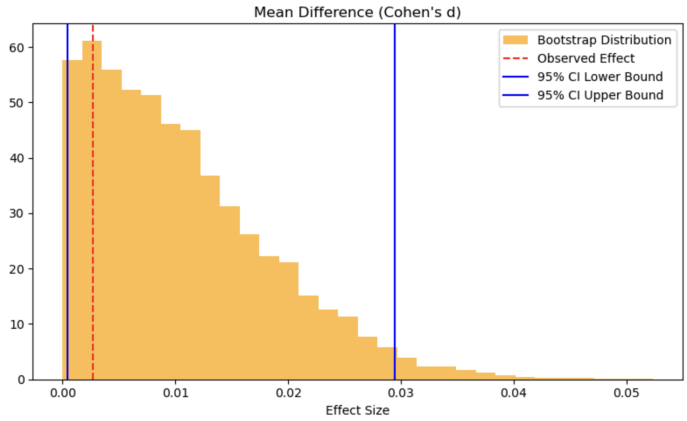
## Q6.

**Approach**

To quantify the likely size of the effect of gender on average difficulty ratings, we calculated the **absolute value of Cohen's d** as the focus is on the magnitude of the difference**.** A bootstrap method with 10,000 resamples was used to estimate the sampling distribution of the effect size and compute a 95% confidence interval.

**Results**

a. Observed Effect Size: 0.003
b. 95% Confidence Interval: [0.0004, 0.030]




**Conclusion**

The observed effect size is small, with a 95% confidence interval suggesting that the true effect size is likely close to zero, with an upper bound of 0.029. This indicates minimal gender impact on difficulty ratings, with no practical significance, consistent with the hypothesis testing results from the previous question.

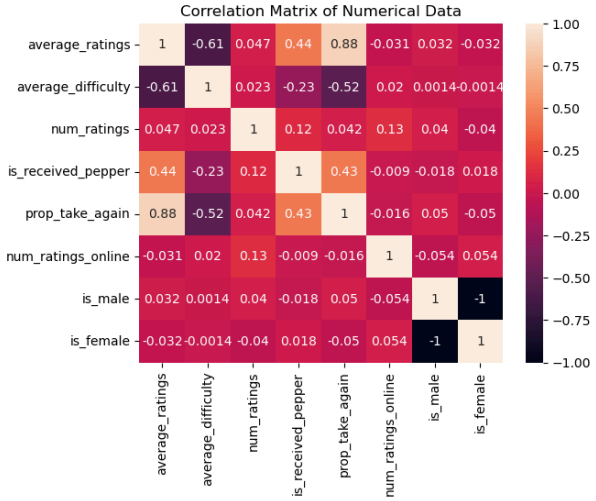
## Q7.

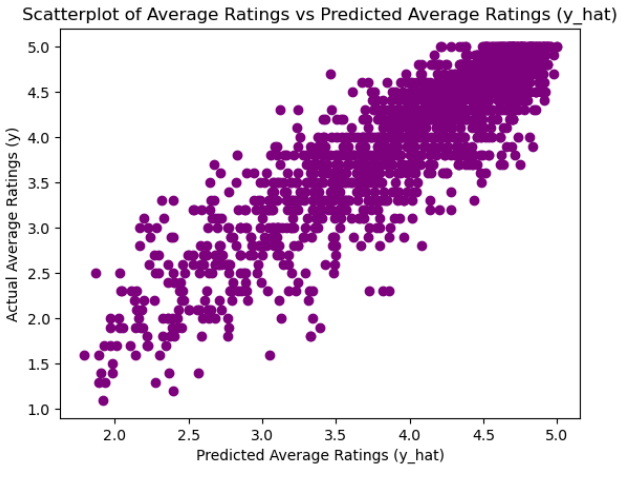
**Approach**

We built a linear regression model without regularization, as the magnitude of the multicollinearity was small (magnitude of the correlations between predictors is at most 0.52), and the sample size is much larger than the number of predictors. Features were standardized to address differences in scale. To avoid the dummy variable trap, we excluded the "is_female" feature. Missing values in the "prop_take_again" column were handled through row-wise removal instead of dropping the entire column, given its relatively strong correlation with the target variable, which indicates its potential importance for prediction.

**Results**

- RMSE: 0.365
- $R^2$ : 0.801
- Predictor with the largest coefficient: "prop_take_again" (0.6144)



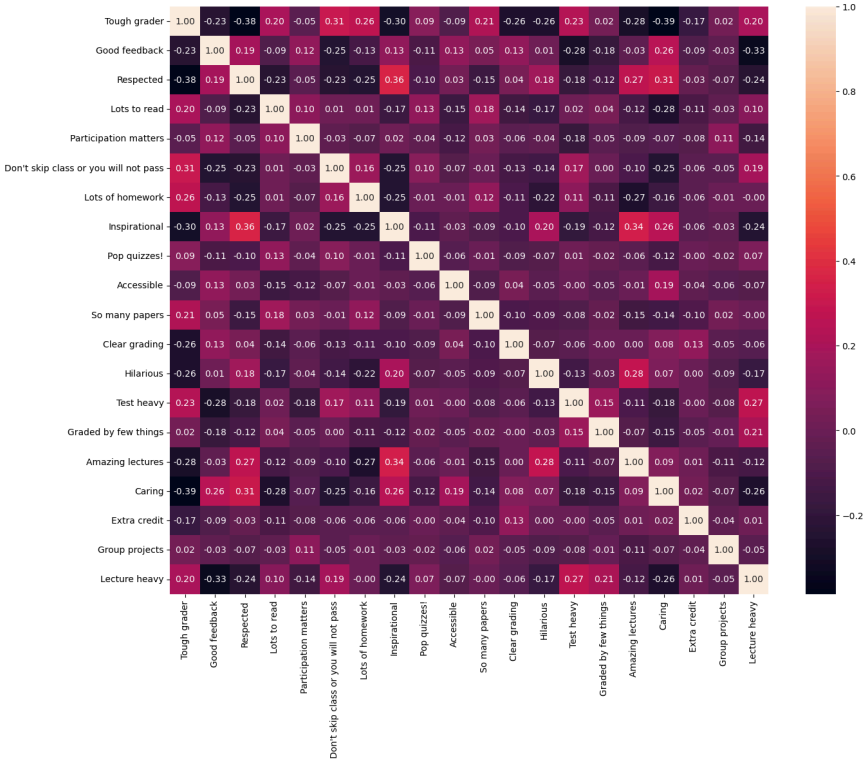
## Q8.

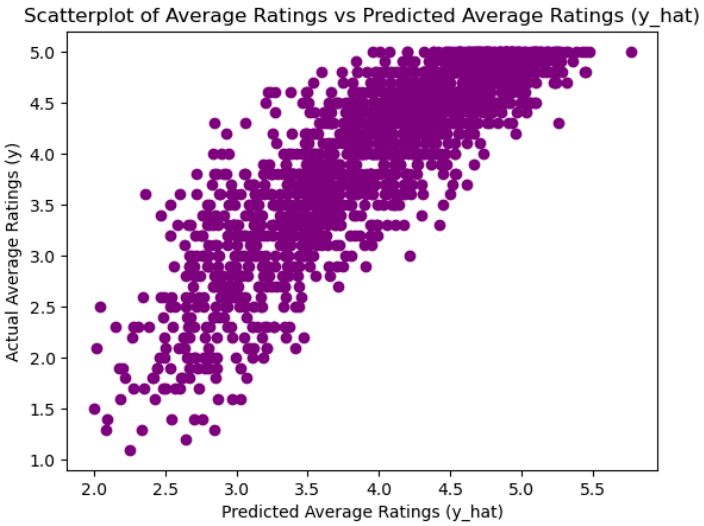
### Approach

We constructed a linear regression model without regularization, consistent with the approach with the previous question. Also, no columns were dropped since there are no multicollinearity concerns and missing values in tags. To ensure a fair comparison with the model in question 7, we used the same training and test datasets from question 7.

### Results

- RMSE: 0.427
- $R^2$ : 0.727
- Most Significant Predictor: “Amazing lectures” (1.114)




### Conclusion

The previous model exhibits lower RMSE and higher $R^2$, demonstrating lower average squared residuals and explaining a greater proportion of the variance in the average ratings.

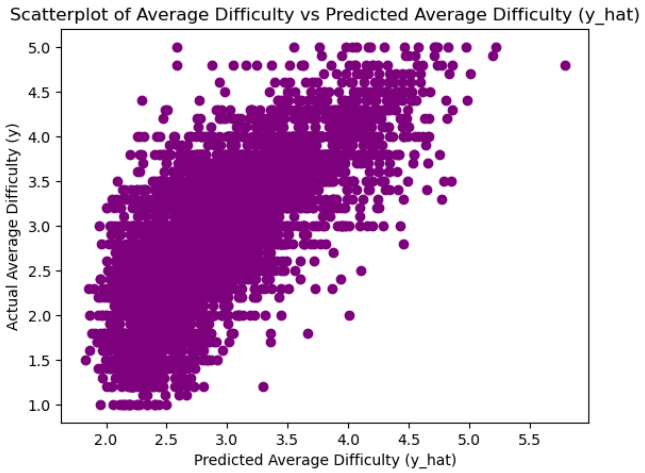
## Q9.

### Approach

We utilized all tags of preprocessed data because there are no missing values, unlike the numerical data. We built a linear regression model without regularization, and no tags were removed due to the low correlations between predictors. The model achieved an RMSE of 0.563, which indicates that its predictions deviate by approximately 11%, given that the average difficulty ranges from 0 to 5. Also, our model explains 53% of the variance in the target variable. This aligns with the scatterplot, which shows a notable scatter around the perfect prediction line.

**Results**

- RMSE: 0.563
- $R^2$ : 0.533
- Most Significant Predictor: “Tough grader” (1.658)




**Conclusion**

Given these results, we conclude that although our model shows a general tendency for a positive correlation, it still has room for improvement.

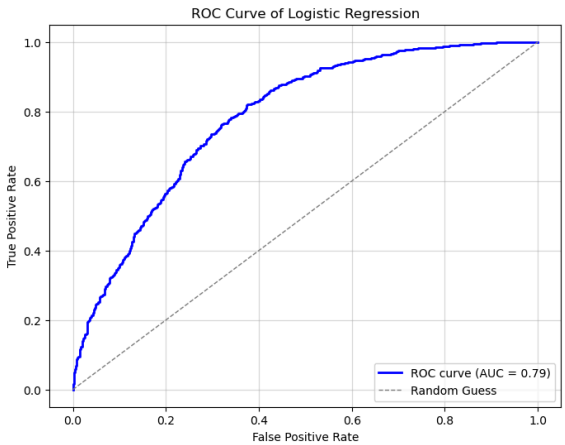
## Q10.

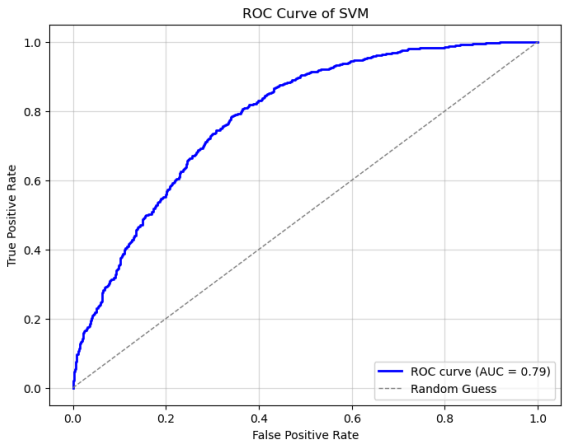
**Approach**

We handled missing values in column 5 (“prop_take_again”) of the numerical data through row-wise removal. Class imbalance was less of a concern, as the distribution of values in column 4 (“is_received_pepper”) is fairly even (0: 53%, 1: 47%). To provide a more comprehensive evaluation of the model, we also computed the precision, recall, and F-1 score. Experiments were conducted using both the logistic regression model and the support vector machine (SVM). Assuming equal importance for precision and recall, we tuned the threshold of the logistic regression model from 0.1 to 0.9 with step 0.1 to maximize the F-1 score. A decision value of 0.0 is used for the SVM.

**Results**

| | AUC | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.786 | 0.688 | 0.620 | 0.912 | 0.738 |
| **Support Vector Machine** | 0.785 | 0.716 | 0.680 | 0.777 | 0.725 |



## Extra Credit

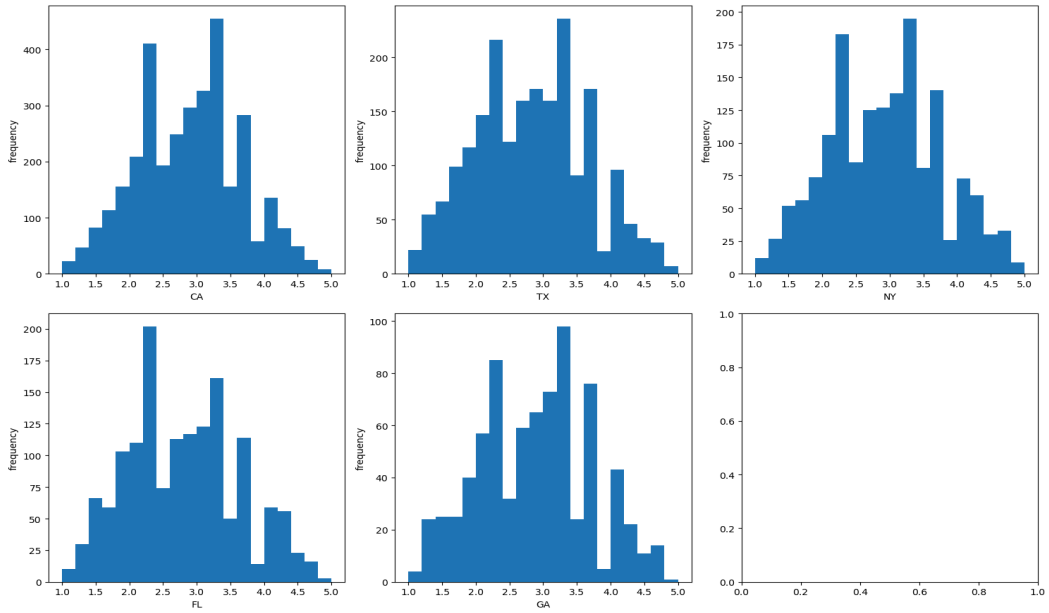
### Approach

We examined whether professors are rated differently in average difficulty ratings among the five states with the largest sample sizes. To test this, we used ANOVA since the distributions appeared to be approximately normal for all states, and the sample size for each state is large enough to satisfy the Central Limit Theorem.

- **Null Hypothesis**: Average difficulty ratings across chosen states are equal.
- **Alternative Hypothesis:** There is a difference in the average difficulty ratings for at least one pair of states.

### Results

- ANOVA
  - P-value: 3.505e-06
  - Test statistic: 7.690




| States | CA | TX | NY | FL | GA |
|---|---|---|---|---|---|
| **Sample size** | 3354 | 2066 | 1632 | 1503 | 783 |

### Conclusion

Given that the p-value is less than 0.005, we conclude that there is statistically significant evidence indicating that the average difficulty ratings differ for at least one pair of states.