MATH448/648 Final Report

Using ‘Rate my Professors’ data to better understand academic biases

Taís Baruel Vieira (MATH648)

Deidra Lemus (MATH448)

Emilia Morgan (MATH448)

May 16th, 2023

Introduction

The website "Rate My Professors" is an online application to give students a platform to review professors in college courses. Students leave a star score, a difficulty rating, can leave a written review, as well as some other information on the professors. The website is meant to "provide user generated feedback on professors' teaching methods and their respective courses as well as [provide] user generated feedback on the lifestyle and facilities of college and university campuses" (ratemyprofessor.com).

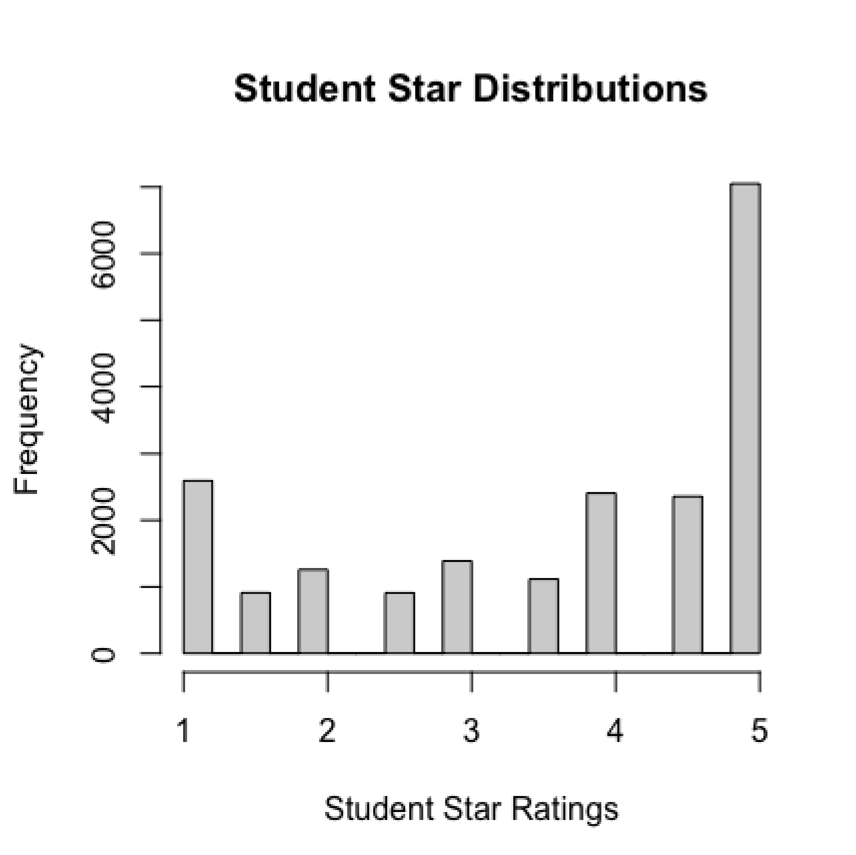
This project's goal is to look at trends in a sample data set from the website and investigate how "Rate My Professor" is used by students, and whether ratings reflect social bias and trends in academia. The data used in this project comes from the website: https://data.mendeley.com/datasets/fvtfjyvw7d/2, published in March 2020. The data was compiled by Dr. Jibo using internet scraping on the website https://www.ratemyprofessors.com/. Unfortunately, Dr. Jibo did not get back to us in our multiple attempts to access the full data set, so we use a sample of the full data in our analysis. The data set used includes data for over 1,400 professors with 20,000 total reviews and 51 variables for each review. Variables are mainly categorical and include the professor's name, school, state, difficulty, rating, gender, race, and more. There are missing observations for some variables, such as gender, but the number of non-missing observations is still high enough to make the analysis statistically significant.

We wanted to explore how identities affected students' interpretations of professors; for example: were professors that taught STEM seen as more difficult? For these explorations, we investigated the correlations between race, gender, STEM vs non-STEM, and other explanatory variables. We also wanted to explore whether we could see professors getting better over time the longer they taught, how students used the website itself, and the distributions for different factors. We also wanted to create a predictor model and test its R^2 value to see how well it would work, as well as figure out the most relevant explanatory variables for the predictor model.

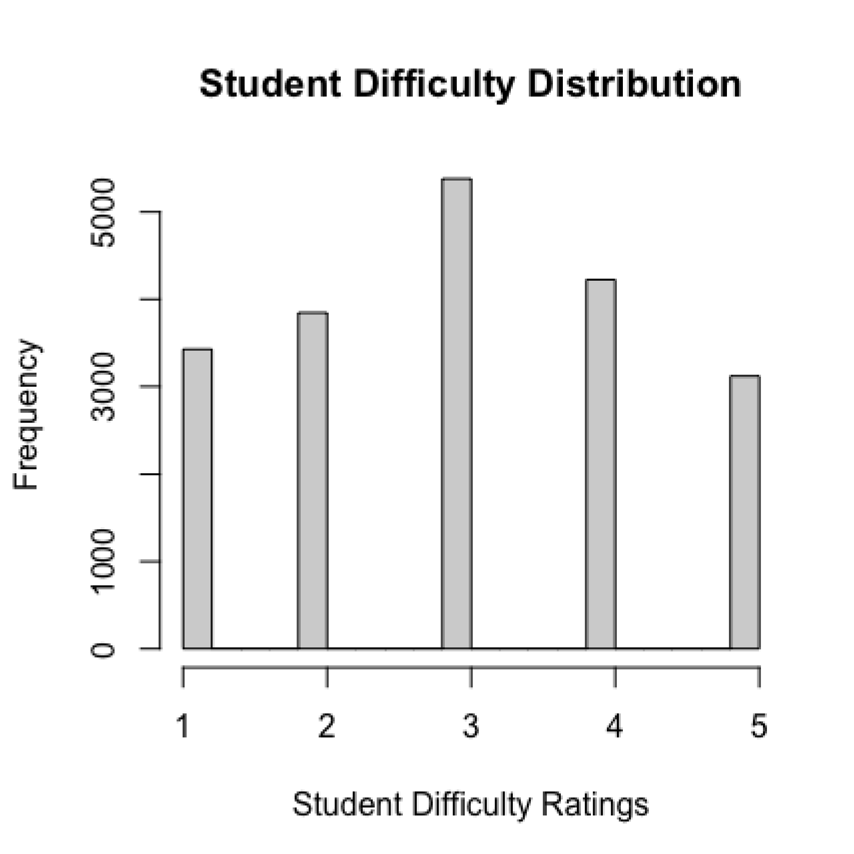
1. Distribution of Student Ratings

Before cleaning up the data to create a predictor model, we investigated the distribution of student ratings both for the “goodness” of the professor and “difficulty”. Given colloquial student usage of rate my professor, we hypothesized that students tend to use the website when they feel strongly about a professor, whether that was a like or dislike. Given this, we thought we would see that the distribution of ratings would notably aggregate around the low- and high-end ratings predominantly, with fewer “middle” star ratings. We plotted histograms for all the student's star ratings and difficulty ratings.

We can see in the graph below that the most common rating for students to input as Star Ratings for professors is a 5, with 1, 4, and 4.5 have similar frequencies. This does demonstrate that student ratings for “stars” are usually either very high or very low, with much lower frequencies of the middle rating numbers.



The histogram distribution for student difficulty ratings is the opposite, where it looks like a normal curve, with three being the most frequent rating for professor difficulty. This suggests that while students go to the website when they have a strong feeling about the professor, that same sentiment is not reflected in the difficulty ratings of professors.



2. Data Clean-Up

After we graphed some of the data from the initial sample, we wanted to clean up the data we had. The original dataset consisted of 20,000 observations (reviews) and 51 variables. For most of our analysis, we needed a data set with one observation per professor, keeping the "average" score for various variables that changed in different students' ratings. The mean star rating and mean student difficulty rating were calculated per professor. We also kept the school name, department, state, professor's gender, race, and the tags available for the reviewer (gives good feedback, caring, respected, participation matters, clear grading criteria, skip class, amazing lectures, inspirational, tough grader, hilarious, get ready to read, lots of homework, accessible outside class, lecture heavy, extra credit, graded by few things, group projects, test heavy, so many papers, beware of pop quizzes, is course online). We also created two new variables for the number of reviews each professor had and the number of classes they taught, as well as the standard deviation for star rating and difficulty rating.

After taking out the professors that had missing information, we looked at the distribution of the number of reviews each professor had and used the first quartile, which was 8, as our minimum number of reviews professors could have for us to use their data (**Figure 1**). We decided to use this approach to create a cleaner data set after discussing with the professor multiple ways to go about reducing the data with professors who had more reviews. We were left with a data set consisting of 969 observations (professors) and 32 variables. The cleaned data set represents roughly two thirds of the original data set.

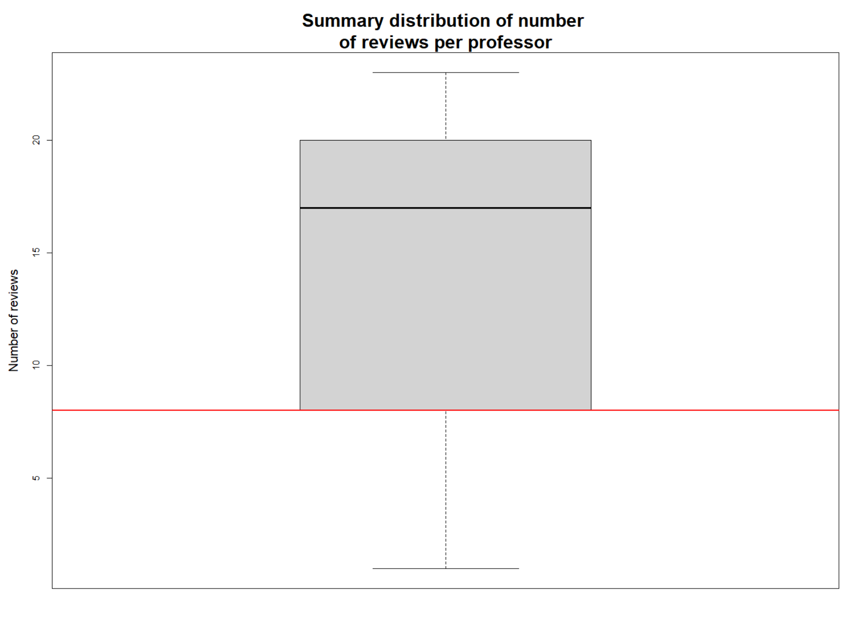


Figure 1 Boxplot of the distribution of number of reviews per professor. *The red line corresponds to the number 8, meaning that is the first quantile in this data.*

3. Star Rating distribution per Professor and Linear Model

To determine the distribution followed by the star ratings, the *descdist* function from the *fitdistrplus* package was used. This function returns a skewness-kurtosis plot based on descriptive parameters computed for the data provided. This plot (**Figure 2**) shows that our data's star ratings fall into a Beta distribution. With that information, the *ebeta* function from the *EnvStats* package was used to determine the parameters of the distribution using Maximum Likelihood Estimators.

The parameters were found to be 2.43 and 1.23, and that allowed us to confirm that, in fact, that is the distribution our data follows (**Figure 3**).

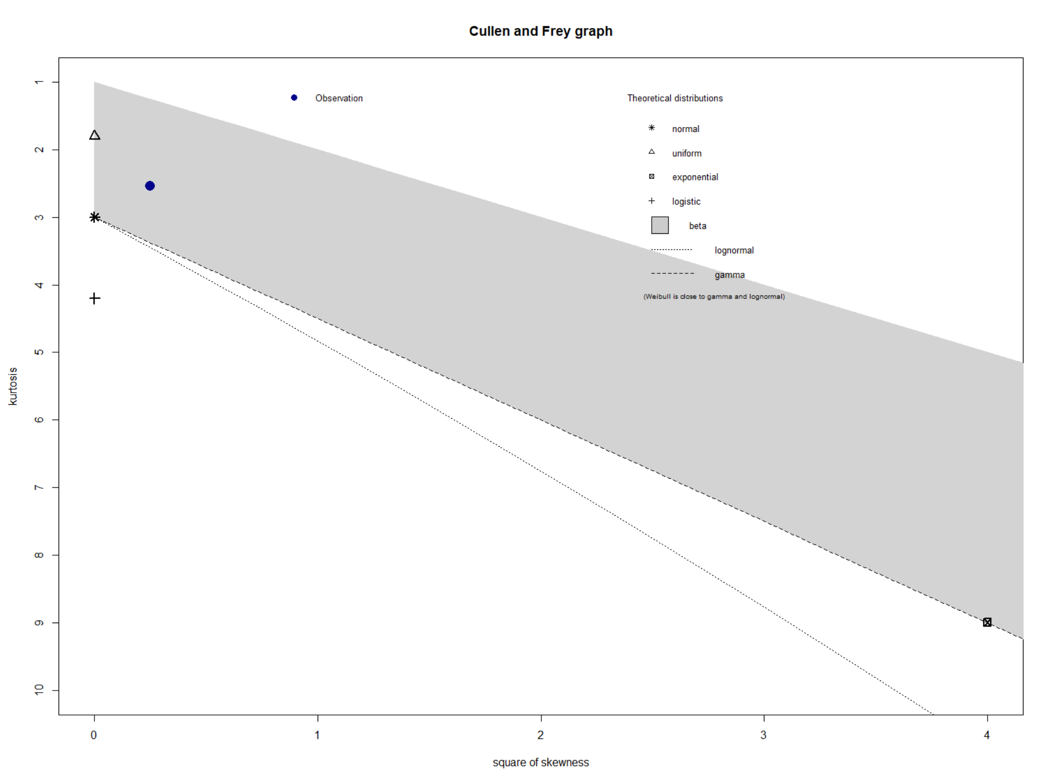


Figure 2 Skewness-kurtosis plot for the star rating data. *The plot shows how the star rating falls into a Beta distribution.*

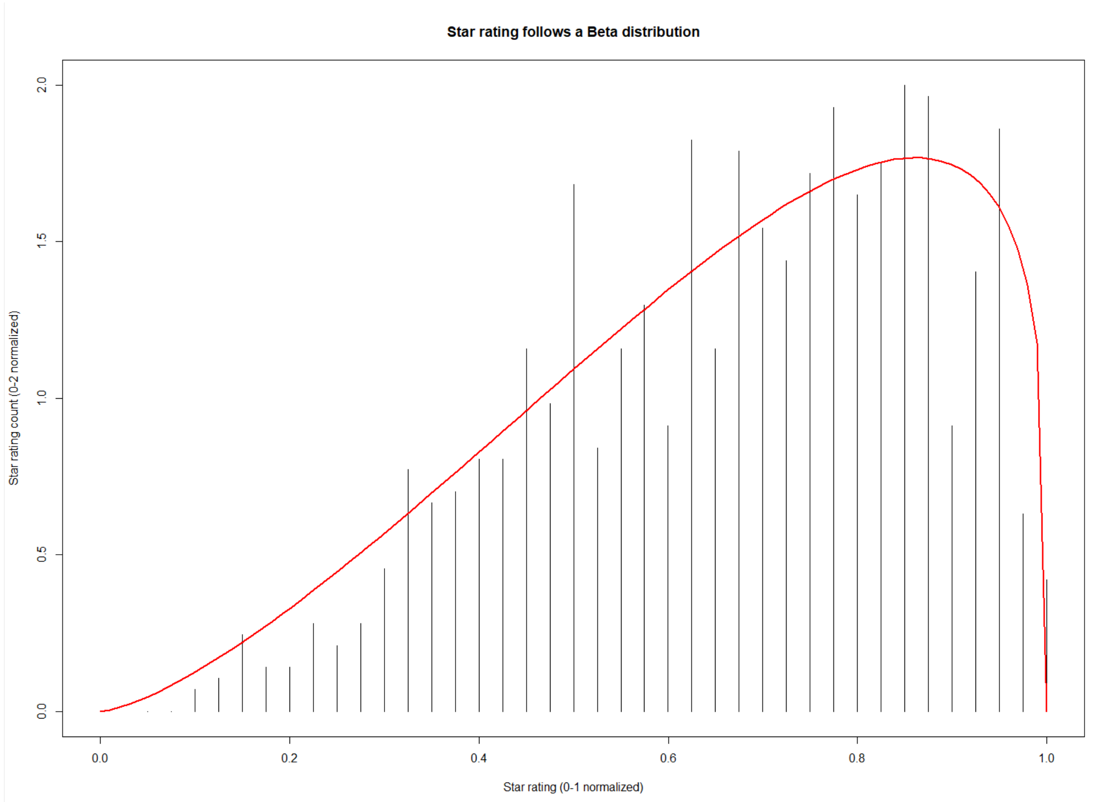


Figure 3 Star rating follows a Beta distribution. *Each line in the plot corresponds to each of the ratings' frequency from our data. The red line corresponds to the Beta distribution with the parameters found.*

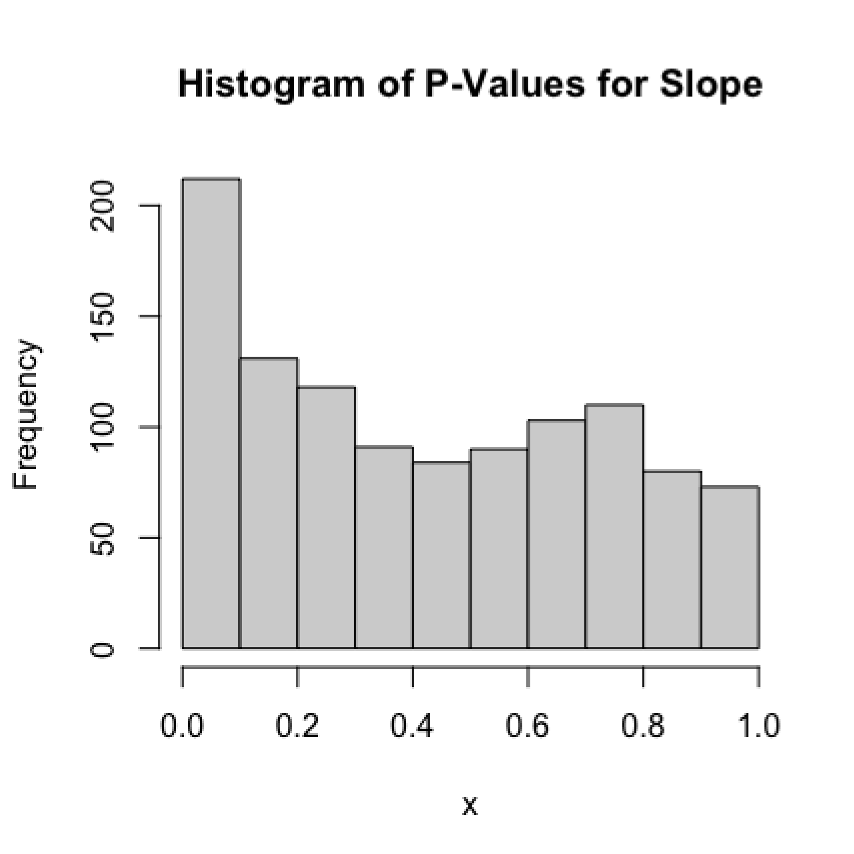
We were also interested in finding the best model to predict star ratings from our data. For that we used the function *lm* passing all the variables, which generated a model with a R-squared of 0.719. This model was used as a starting point for backward selection of a new model by AIC. The new model found had an R-squared of 0.341 and the variables found were student rating difficulty, number of reviews and a few binary tags (gives valuable feedback, respected, amazing lectures, and lecture heavy). Interestingly, most of the variables correspond to tags associated with the professors (estimates = 0.19, 0.33, 0.23, -0.23, respectively) and variables such as gender and race do not play a statistically significant part in the professor's rating. Another surprising fact is that the number of reviews is negatively associated with the professor rating (estimate = -0.014). That is, the more students that review a professor, the more the rating of the professor drops. This signifies that a professor with more student reviews tends to have lower scores, partially signifying that students actively seek out writing reviews when they do not like a professor. Student difficulty is the variable that has the biggest effect on rating (estimate = -0.51). This is not surprising, as it seems obvious that a more difficult class translates into a lower rating of the professor.

Overall, this model does not have a high r-squared value which indicates that it does a poor job predicting a professor's rating since it can only explain about one-third of the variation in the dataset. Even the model that accounted for all the variables available could not explain more than 75% of that variation. This could have two simple explanations which are not mutually exclusive. First, the data set could be biased, given that not every student will be prompted to leave a review. It is straightforward to assume that students with strong opinions – good or bad – are the ones that are willing to act and leave a review. That can affect the distribution of the ratings and the independence of it. Second, it is entirely possible that a professor's rating cannot be predicted using a binary model (such as tags) because they are individuals that have good and bad moments. The most important variable we could use would be the professors' names since each is unique – although the professors' names could not be a predictor variable.

4. Star Rating Trends Over Time

We wanted to investigate whether professors' ratings went up over time. Using *tidyverse* and *dplyr*, we ordered the ratings based on date (oldest to newest) and "ranked" them, giving the oldest rating a "1" and the newest rating the highest value. We then plotted and fit a linear model to star-rating vs. "ranking" date, to find the slope coefficient. We only used professors with at minimum 8-star ratings, this left us with 1093 professors' data to look at. Plotted below is the histogram of slope coefficients, we can see that it is normally distributed around 0. A slope coefficient of 0 would imply that professors are not getting better over time based on the data from rate my professor. Testing the mean of the slope coefficients against the hypothesis that the mean is 0 using a z-score gave us a p-value of 0.1 demonstrating we can conclude that the mean of the slopes is 0.

To further confirm that the data demonstrates that professors are not getting better over time, we found all the p-values for all the slope coefficients for all the professors as seen by the histogram below. Low p-values suggest that the null hypothesis of a 0-slope coefficient is incorrect. To demonstrate that professors do get better over time, we would want to find p-values less than 0.005 consistently if we want to be 99% confident that we can claim that professors are getting better overtime.



Using the mean of the p-values, which was 0.4325, and testing it against the hypothesis that the mean of the p-values should be 0.0025 (we chose this because it is ½ of .005), we used a two-tail z-test to find the p-value of the mean of the p-values. The p-value was 0, demonstrating that we could not conclude that the average p-value was close to 0.0025, demonstrating that looking at the 1093 professors with 8 or more ratings, we cannot conclude that professors are statistically significantly getting better over time. However, our approach is not the only way to investigate this question of whether professors are getting better over time. Ranking the dates on a scale of +1 is not retaining all the information given by the "date" of the rating. Rather than a +1-scale basis, it would have been better to choose the first day of a rating given as "0" and each subsequent day that passed since the first rating as an additional add of 1. This would mean the data would not have x-values 1, 2, 3, 4.... etc. But rather x-values of 0, 10, 100, 150, 151, etc. that maintain the same amount of information of the spread of ratings over time as keeping the date. Overall, tracking the ways the star ratings change over time is difficult, and our data does not demonstrate that professors get better overtime according to rate my professor.

5. Effect of race and gender on professor's tags

To explore the relation race and gender may have on the data's explanatory variables such as star rating, respected, and caring scores the function *cor* is used. The function gives the correlation of the variables race or gender would have with another variable. The correlation statistically measures the linear relation of two or more variables. The table below shows the correlations race and gender have with all other variables, including each other.

Correlations:	**gender**	**race**
school name	0.1424	0.2147
star rating	-0.0082	0.0035
gender	1	0.0368
race	0.0368	1
student difficult	0.1424	0.0377
caring	-0.0464	-0.1225
respected	-0.0553	-0.0104
inspirational	0.009	-0.142
n review	0.0307	-0.0676
n class	-0.0125	0.0822
stem	0.0375	0.011

We thought we might see a difference between the ratings and tags of female professors and male professors, like seeing women professors have a higher likelihood of being tagged with "caring", or perhaps seeing women professors in STEM being tagged more often as "difficult".

We investigated the explanatory variables for genders and race to see what was statistically significant. We created separate data sets for females(0) and males(1) by extracting the rows from the gender column of the main data set. We did the same for race, separating them by whites(1) and others(0). Once the data sets were created, we utilized the *summary* function to extract the means of every explanatory variable in the data set. The results are shown below.

Means:	**Females(0)**	**Males(1)**
school name	0	0.1333
star rating	3.632	3.612
race	0.5357	0.5889
student difficult	3.184	3.017

caring	0.3214	0.2611
respected	0.3571	0.2833
inspirational	0.1786	0.1889
n review	17.07	17.41
n class	5.929	5.844
stem	0.25	0.3

Means:	**White(1)**	**Other(0)**
school name	0.1736	0.0345
star rating	3.617	3.611
gender	0.876	0.8506
student difficult	3.065	3.005
caring	0.2231	0.2708
respected	0.2893	0.2989
inspirational	0.1405	0.2529
n review	17.15	17.67
n class	6.017	5.632
stem	0.2975	0.2874

The means for the variables we had hypothesized do have slight differences, so we used the *t.test* function to check whether the difference in means was statistically significant. A t-test is used to get inferential results to conclude there is a significant difference between the two means. Looking at the gender data, the variables that stand out and are interesting to check with the use of a t-test are caring, respected, and stem. These three variables were chosen given they had the highest difference in means compared to the other variables, thus being more interesting and worthwhile to investigate.

The p-value of the t-test for comparing the mean values of “caring” is 0.5326. Given the high p-value, unless our significance level is greater than 1 - p, it is safe to say we can reject the hypothesis that a professor’s gender significantly impacts whether they will receive the tag “caring”. The results of the t-test from comparing the tag “respected” gives us a p-value = 0.4572. Given the high p-value, we can reject the hypothesis that gender influences whether a professor will get the tag “respected”. The results of the t-test from comparing whether a professor is a man or a woman teaching stem results in a p-value = 0.5823. This p-value is incredibly high, however, we know that men are far more likely to be STEM professors than women, so this should have a statistically significant p-value. This tells us that our data set is likely missing information, we noted above that only a few professors are tagged with their respective gender. While it is around

100 professors, it seems like this n is not high enough to be a good representation of the actual population, in which about 2/3 of STEM professors are men and 1/3 are women.

Looking at our explanatory variables for “race”, the variables that stand out and are interesting to check with the use of a t-test are school names and inspirational. Again, these two variables have the highest difference in means compared to the other variables, thus being the most beneficial to investigate. The results of the t-test from comparing the types of school’s professors teach at based on race gives us a p-value = 0.0006. Therefore, a professor’s race does give us a good indication of where they would be teaching. Professors who are white, it is more likely they teach at a university. The results of the t-test from comparing whether a professor gets tagged with “inspirational” based on their race gives us a p-value = 0.0488. If the significance level is 0.05, we can conclude that a professor's race will indicate whether they are likely to get tagged with “inspirational”. If a professor is of a race other than white, he or she has an overall probability of being more inspirational than a white professor.

Using the function *chisq.test*, a chi-squared test, we can look further if there is a significant correlation between the two variables selected. Starting with gender and race we get a p-value of 0.7454. This p-value is high which tells us there is not a significant correlation between the two variables which means they are independent from one another. Now we looked at the correlation variables to gender with the greatest difference in means: “caring”, “respected”, and “stem”. The p-value result of the chi-squared gender and “caring” is 0.6597. The p-value result of the chi-squared gender and “respected” is 0.5963. The p-value result of the chi-squared gender and “stem” is 0.7509. Given the p-values are so high we can conclude these three tags aren’t applied by students in a statistically significant way towards one gender verse the other.

Then, we looked at the explanatory variables for race with the greatest difference in mean correlations: school names, and inspirational. The p-value result of the chi-squared test for correlation between race and school-names is 0.004. The p-value result of the chi-squared race and inspirational is 0.0617. Given the p-values are low we can conclude the two variables being compared to race are dependent.

6. Effect of race and gender on professor’s ratings

To start this analysis, we first investigated the relationship between gender and race for the professors. As we can see in **Figure 4** our data shows that the academic world is highly dominated by White males, with Black females being the most outnumbered group.

According to the United States Census Bureau, in 2021 White people represented 59.3% of the US population, Black people were 13.6% and Hispanic people were 18.9% of the US population. In our data, White professors correspond to 58.3%, while Black and Hispanic professors correspond to 19.6% and 22.1% respectively. This is encouraging since our data shows a close distribution in race when compared to the actual country distribution.

On the other hand, we cannot say the same about gender distribution. Again, according to the United States Census Bureau, in 2021 females accounted for 50.5% of the country's population. Granted our data set has most of its *gender* variable labeled as 'unknown,' from the ones that are labeled as 'female' or 'male,' the percentage of females correspond only to 13.5%. This is extremely far from the nation's 50.5%.

Also, we cannot forget that these numbers are solely based on American demographics and the usual professor body is comprised of many international professors. With that in mind, we would hope that academia would have distributions across gender and races that mirror societies, including demographics that were not represented in this data set like professors who identify as Asian.

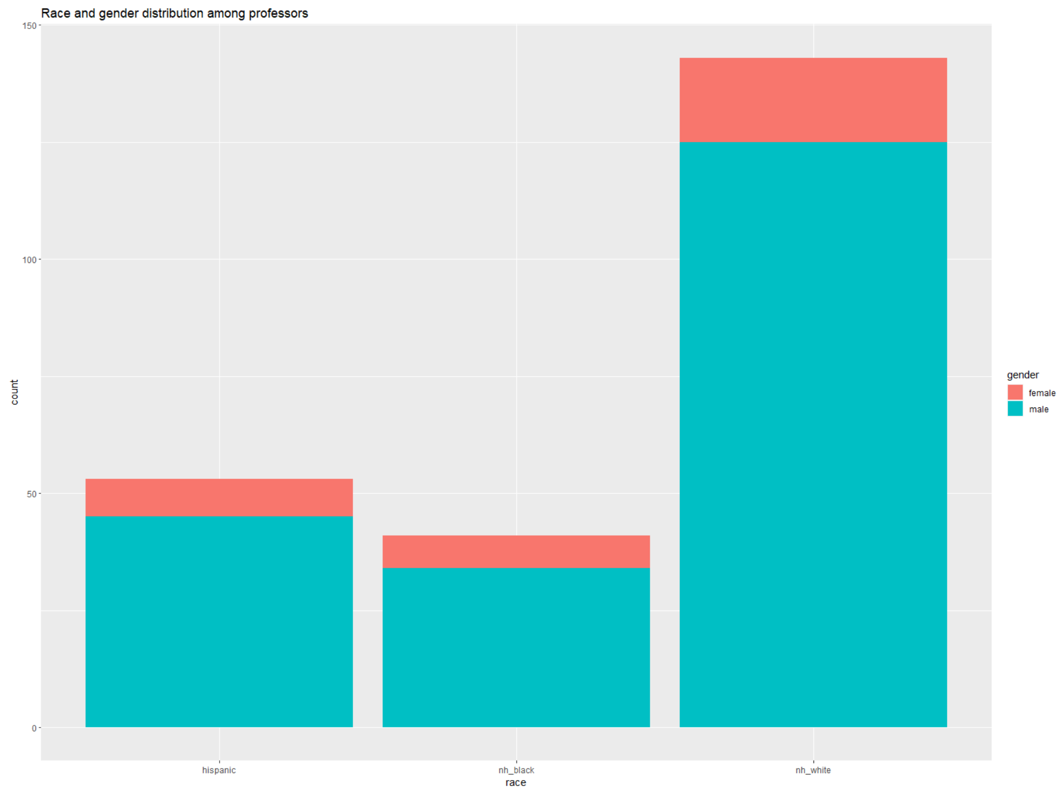


Figure 4 Race and gender distribution among professors. *It is clear that the dominant group is of white males and black females represent the smallest group.*

To analyze if race and gender have any effect on the professor's rating, we performed a 2-way ANOVA test. It turns out that neither race, gender, nor an interaction between them has any statistical relevance when predicting the star rating (smallest p-value of the analysis: 0.278).

7. STEM vs non-STEM

The first step for this analysis was to classify each department in our data set as STEM or non-STEM. The departments classified as STEM were computer science, physics, chemistry, mathematics, biology, science, health science, engineering, statistics, and mechanical engineering departments.

The first test was to check if there was a difference between the mean star rating for STEM professors and non-STEM professors. We found that there was a small but significant difference between them with a p-value of 0.007. The true difference lies between -0.28 and -0.04 in a 95% confidence interval, with the ratings for STEM professors being lower than for non-STEM professors. This finding agrees with our idea that STEM professors are seen as worse professors than professors who teach non-STEM subjects.

Our second hypothesis was that there would be a significant difference in gender in STEM professors with the field being male-dominated. That is quickly proven by looking at **Figure 5**. In the figure, the first bar represents non-STEM professors, and the second bar represents STEM professors. There are more male professors than female professors in STEM. We also wanted to investigate if the proportion of female professors broadly and female professors in STEM is the same. The p-value for this hypothesis is 0.63, thus we cannot reject the hypothesis that the amount of women professors in STEM is proportionally the same as the amount of women professors in academia. In this case, even though the proportion of female professors in STEM is close to 20%, that unfortunately is the reality for academia in general.

It also seems like our data includes far more non-STEM professors than STEM professors.

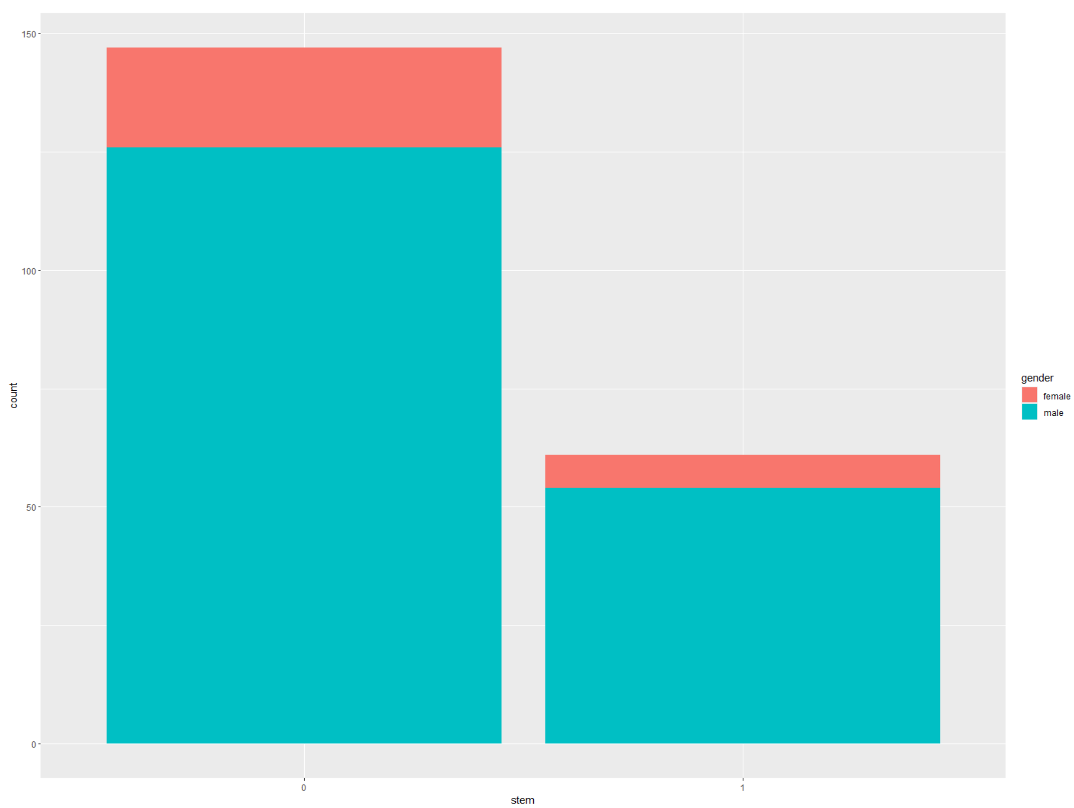


Figure 5 STEM vs gender. *Gender distribution across STEM (1) and non-STEM (0) professors.*

Lastly, we asked if there was any predominant race among STEM professors. As expected, the field is White dominated as depicted in **Figure 6**, but after doing the same proportion analysis as we have done for gender, we found that Black representation in STEM fields can be considered different than their overall representation (p-value: 0.07). Surprisingly, this representation is higher in STEM than in non-STEM (34.8% vs 26.5%). And even though the Hispanic representation is quite different between both fields (19.8% vs 26.5%), this difference cannot be considered statistically significant (p-value: 0.15)

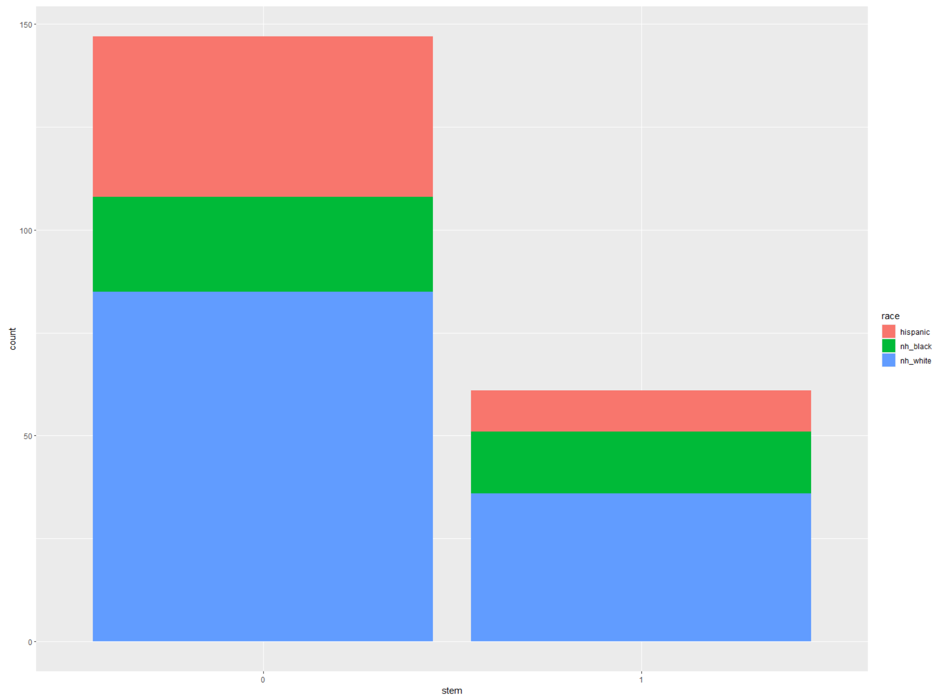


Figure 6 STEM vs race. *Race distribution across non-STEM (0) and STEM (1) professors.*

We then investigated the correlation between stem and star ratings. The result was not statistically significant with a p-value = -0.1144, so we cant say our data demonstrates that STEM and non-STEM professors receive on average different ratings. We then investigated the summary of the stem and non-stem professor data sets. To create these separate data sets for stem(1) and non-stem(0) by extracting rows from the stem column of the main data set. The results are below.

Means:	**Stem**	**Non-stem**
star rating	3.467	3.676
school name	0.2623	0.05442
gender	0.8852	0.8571
race	0.5902	0.5782
student difficult	3.095	3.017
caring	0.3115	0.2517
respected	0.3115	0.2857

inspirational	0.2131	0.1769
n reviews	17.56	17.29
n class	5.721	5.912

Looking at the findings, the most interesting mean differences to investigate further using our t-test function is star rating, school name, caring, respected, and inspirational. These tags were chosen given they had the greatest mean differences suggesting there is a greater probability to be significant, as we will discover further in the process. The results from a t-test comparing star rating means of stem and non-stem professors are a p-value = 0.007. If our p-value is 0.01 we can conclude that there is enough statistical evidence to back up that non-STEM professors tend to have higher star ratings than non-STEM professors. The results from a t-test comparing what type of schools the two types of professors teach at, community college or university, give us a p-value = 0.0008. If our p-value is 0.01 we can conclude that there is enough statistical evidence to back up that STEM and non-STEM professors tend to teach at different schools, name STEM professors are likely to teach at community colleges and non-STEM professors are more likely to teach at Universities.

The results from a t-test comparing the likelihood of stem and non-stem professors getting tagged with “caring” has a p-value = 0.3934. Given the high p-value, we conclude there is no significant difference in stem and non-stem professors getting tagged with “caring”, rejecting the alternative hypothesis. The results from a t-test comparing the likelihood of stem and non-stem professors getting tagged with “respected” has a p-value = 0.7156. Given the p-value is large we can conclude there is no significant difference in the likelihood that stem and non-stem professors get tagged with “respected”. The results from a t-test from comparing inspirational means of stem and non-stem professors is a p-value = 0.5574. Given the high p-value, we conclude there is no significant difference in inspirational means of stem and non-stem professors, rejecting the hypothesis.

Thus, we cannot conclude STEM and non-stem professors will get different star ratings, or be tagged with “caring”, “respected”, or “inspirational” based on the subject. We can conclude that the subject of a professor indicates what type of school the professor teaches at with stem professors.

8. Community college vs University

The next exploration is into college names and if there are any significant differences in means and if there are any correlations in explanatory variables. The first step was to select the school names by selecting any names that had the Community within it and renaming it 1 and every other 0, which would stand for University. This was done using the grep function, which looks for characters or sequences of matches of a string.

After this step, we created two data sets by extracting the values of Community(1) and extracting the values of University(0). Then, the summary function finds the mean of the explanatory variables of both data sets.

Means:	*Community(1)*	*University(0)*
inspirational	0.2917	0.1739
star rating	3.575	3.62
gender	1	0.8478
race	0.875	0.5435
student difficult	2.952	3.051
caring	0.4167	0.25
respected	0.2917	0.2935
n review	18.29	17.24
n class	6.167	5.815
stem	0.6667	0.2446

Looking at the results, the variables that show a significant difference that seemed the most interesting to uncover through a t-test are inspirational, star rating, caring, and stem. Gender is not as interesting given the limited data professors that have gone to community college are only men.

The results from a t-test comparing inspirational means of Community and University school names are a p-value = 0.2438. Given our p-value is greater than 0.05 we reject the hypothesis that there is a significant difference between inspirational means of school names. The results from a t-test comparing star rating means of Community and University school names is a p-value = 0.8215. Given how large our p-value is we can safely reject the hypothesis that there is a significant difference between star rating means of school names. The results from a t-test comparing caring means of Community and University school names are a p-value = 0.1330. Given our p-value is greater than 0.05 we reject the hypothesis that there is a significant difference between caring means of school names.

The results from a t-test comparing stem means of Community and University school names is a p-value = 0.0003. If the significance level is 0.0004, we can accept the means of stem scores between school names are significantly different. Therefore, there is enough statistical evidence to back Community school names have more stem professors than University school names.

Thus, we cannot conclude statistically that the differences in means of inspirational, star rating, and caring variables for school names are significant. We can conclude that the differences in means of the stem variable for school names are significant.

9. Conclusion

All the conclusions made using this data must consider the fact that we only have a sample of the entire analytics for rate my professor to work with and that even the full data from “Rate My Professor” isn’t comprehensive, since every student in every class does not post on “Rate my Professor”.

In this project, we devised a model to predict a professor's rating based on a few variables, namely the difficulty rating of their class and whether they give valuable feedback, are respected, have amazing lectures, and have a lecture-heavy class. The model only accounts for one-third of the variation possible, but we must consider the individuality of each professor that cannot be predicted by models.

It is clear our data is missing important information detailing the gender of professors, given that our model does not find gender a significant explanatory variable for what subject a professor may teach, when around 4/5 of STEM professors are male, with women occupying only 21.7% of all STEM faculty positions in the US in the 2020-2021 year (Mehta, Neil). Our data did confirm the common idea that STEM classes are harder than non-STEM classes which is a popular attitude among the student population. Our data also demonstrated that STEM professors were more likely to teach at universities while non-STEM professors were more likely to teach at community colleges.

We also found that there is a decent race representation in the academic world, but gender representation is still lacking. The same goes for representation in STEM, with the interesting finding that there is a higher Black representation in STEM fields than overall.

We discovered we don't have enough statistical evidence to back up any meaningful difference in professor tag variables for gender. We now have enough statistical evidence to conclude that there are meaningful differences in tag means for race, with non-white races being more prone to be tagged by inspirational that they teach at community college while white professors are found more often in universities. This absolutely demonstrates bias in academia.

Also, after evaluating the differences in subject types there is enough statistical evidence to say non-stem professors have greater star ratings than stem professors. As well as there is enough statistical evidence that community colleges have more stem professors represented in the data set than universities do. This seems somewhat contrary to what we would expect given that universities are the research schools that employ and have labs and necessitate that professors write papers, while community colleges do not.

Overall, our findings were quite interesting, while our dataset was not complete. Even a small sample from "Rate My Professor" reflects some of the academic biases whether from students or within the academic community at large. It does seem like students use "Rate My Professor" when they feel strongly about a professor, but the difficulty rating is distributed with a normal distribution with a mean of around 3 three, demonstrating that a student feeling strongly about a professor is about more than just difficulty.

Bibliography:

Mehta, Neil. “Faculty, Students Reflect on STEM Faculty Gender Gap.” *The Brown Daily Herald*, 5 Apr. 2022, www.browndailyherald.com/article/2022/04/faculty-students-reflect-on-stem-faculty-gender-gap#:~:text=Women%20made%20up%20less%20than,in%202020%2D21%20academic%20year&text=Women%20occupied%2021.7%25%20of%20all,Equity%20and%20Diversity%20in%202021.

U.S. Census Bureau quickfacts: United States. (n.d.). United States Census Bureau QuickFacts. Retrieved May 16, 2023, from https://www.census.gov/quickfacts/fact/table/US/PST045221