Galal Bichara
Pascal Wallisch
Principles of Data Science (DS-UA-112)
12/09/2024

In this project, data cleaning was done on a case-by-case basis using the methodology I deemed the best fit for the question. After in-class discussions about the dangers of imputation, I decided against doing any imputation in this project and opted to do removals instead. The overarching approach used was removing rows with missing values for critical metrics such as AvgRating and numRatingsOnline to ensure data from different rows remain separate.

After loading in the data, I labeled the columns to make them easy to reference without needing to remember the order. After that, I inspected the numerical data csv, where I saw that there were some professors that had nans across the board for all ratings, so I filtered them out from the dataframe. That took me from 89,893 professors to 70,004 professors. Once those have been filtered out, I did some further analysis to see what threshold to use to cut off the data. As seen in the question, professors with few ratings seem to be a concern, *"this research is of technically poor quality, either due to a low sample size – as small as n =1"*. In order to avoid this, I set a threshold for the number of ratings, and I excluded any professors with ratings below than said threshold. I set this threshold by finding an appropriate percentile and its resulting threshold of number of ratings, where I will exclude any professor with ratings less than or equal to that threshold. But what is a percentile?

A percentile represents the value below which a given percentage of observations in a group falls. For instance, the 75th percentile is the value below which 75% of the data points lie. In this project, percentiles helped to set a quantile-based threshold to exclude professors with a very low number of ratings, ensuring that only those with a sufficient number of ratings were included. This approach prevents unreliable analysis due to small sample sizes and ensures that conclusions drawn are based on representative data.

For this project, I chose the 50th percentile as the threshold for filtering. I initially selected the 25th percentile, which would have resulted in the same number of rows as if I had not applied any cutoff at all, making it ineffective for cleaning the dataset. As a result, I upped it to the 50th percentile. Although a higher threshold could be used, I intentionally chose the 50th percentile (3 ratings) because I plan to use this filtered dataset for all questions, which might require additional cleaning. So, I will be using professors with 4 or more ratings. Another reason behind the lower threshold is that a higher one could also exclude newer professors who have been teaching for less time and, in turn excluding a meaningful group.

I also seeded the random package using my n-number, which is 16906324, in order to make sure the train-test splits are consistent and reproducible as well as unique just to me.

Question 1:

For this question, I took the filtered dataframe, which I filtered using the percentile threshold of 4 or more ratings, and separated it into 2 groups. A male professor dataframe (13,438 rows) and a female one (11,793 rows). Once that is done, we are ready to do our hypothesis testing. What is a hypothesis test? It is a test that is used to evaluate if an assumption is true or not. The assumption we make is that the difference in ratings is just random.

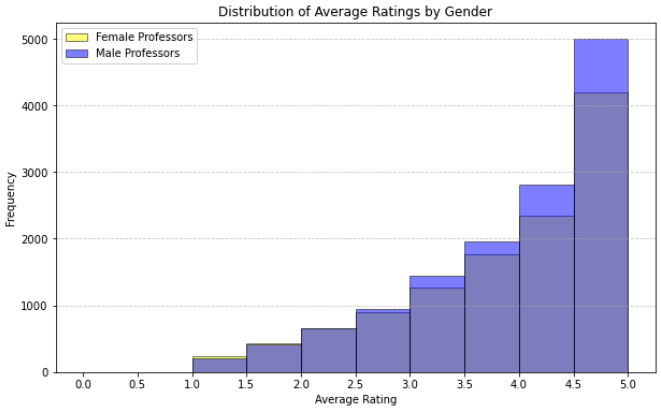

Figure 1

As seen in Figure 1, the ratings are clearly not normally distributed, so using a t-test is not ideal in this scenario, so I will use its non-parametric counterpart, the Mann-Whitney U Test. Figure 1 also shows male professors with higher average ratings than female professors.

After running the test, I get a p-value of 0.000157. This p-value is smaller than our alpha level of 0.005, meaning that there is a statistically significant difference between the ratings of male and female professors. The p-value is also 31x smaller than the alpha value, which implies that no p-hacking is involved.

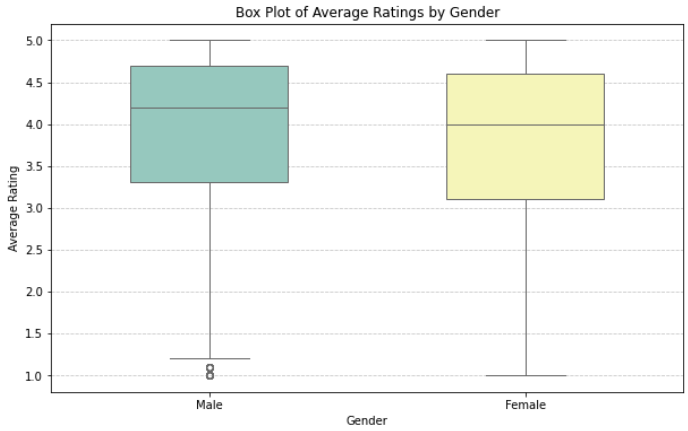

Figure 2

Furthermore, the median rating for male professors is 4.2, and that of female professors is 4.1, as seen in Figure 2, indicates that the students rated the male professors higher. Although there is a statistically significant difference between the ratings of male and female professors, the difference between them is rather small, shown by our Cohen's d value, which is 0.06, so it is almost meaningless in practice.

The last concern I addressed was checking for confounders, particularly teaching experience. As there is no column that describes experience, I used the number of ratings as a proxy for experience. To control for the confounder, I ran a Multiple Linear Regression with gender and number of ratings as the predictors, and the predicted outcome is the average rating of the professors. I chose a Multiple Regression Model because the model provides you with a p-value for the coefficients, which answers the question of significance of specific factors when controlling for confounding variables.

OLS Regression Results

Dep. Variable:	AvgRating	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	91.95
Date:	Wed, 04 Dec 2024	Prob (F-statistic):	1.62e-40
Time:	22:02:39	Log-Likelihood:	-35266.
No. Observations:	25560	AIC:	7.054e+04
Df Residuals:	25557	BIC:	7.056e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.7356	0.010	387.702	0.000	3.717	3.755
q1Gender	0.1237	0.012	10.150	0.000	0.100	0.148
numRatings	0.0051	0.001	8.791	0.000	0.004	0.006

Omnibus:	2333.155	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3038.586
Skew:	-0.844	Prob(JB):	0.00
Kurtosis:	2.922	Cond. No.	32.3

Figure 3

As seen in Figure 3, the p-value for the q1Gender coefficient is 0, meaning that gender is still statistically significant when controlling for confounders.

Question 2:

As for this question, I will split the professors into groups based on their experience rather than gender. I will use the same minimum threshold in question 1, where only professors with 4 or more ratings are included. Once those with few ratings are excluded, I will split it based on the 75th percentile, which yielded a threshold of 11. I chose the 75th percentile because there are more inexperienced professors than experienced professors, so a 50/50 split did not seem adequate. Once the data is split, we see that the professors are split into an 80/20 split, with 20% being the experienced professors (12 or more ratings; 7,202 rows) and inexperienced professors (between 4 and 11 ratings inclusive; 24,749 rows) the 80%.

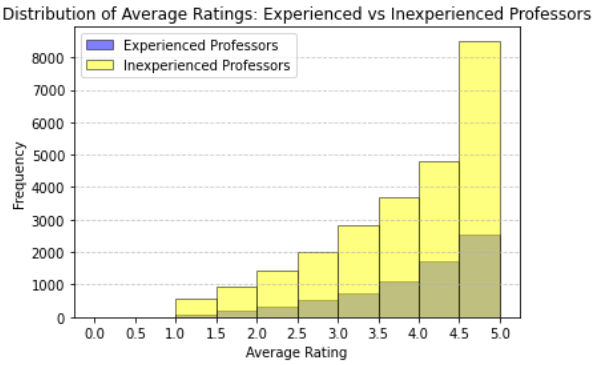

Figure 4

The ratings in Figure 4 are not normally distributed; therefore, I used the Mann-Whitney U test to assess statistical significance. Notably, inexperienced professors have higher numbers of ratings, partly because the experienced professors' sample size is three times larger.

Once the test is run, the p-value I get is 0.000101, which is 50x smaller than the significance level, which indicates that there is a statistically significant difference between experienced and inexperienced professor ratings and that it is not due to chance alone and a difference that significant implies that there is no p-hacking involved in this test.

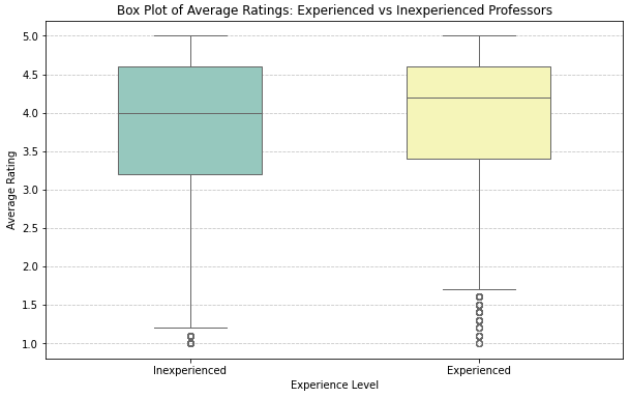

Figure 5

Since the median rating for experienced professors is 4.2 and that of inexperienced professors is 4.0, we know that the students rated the experienced professors higher. After testing for significance, I checked for the effect size, which was 0.103. This indicates that although there is a difference between the groups, it is a smaller one, and again implies that is not the most meaningful in practice. Given the sample sizes, I also conducted a power analysis on the G* Software, which resulted in a power value of 1.000. This power means that the data is more than sufficient to detect statistical significance if it exists.

Question 3:

For this question, I plotted the average ratings of professors against the average difficulty rating they received to see if there is a relationship between the two variables.

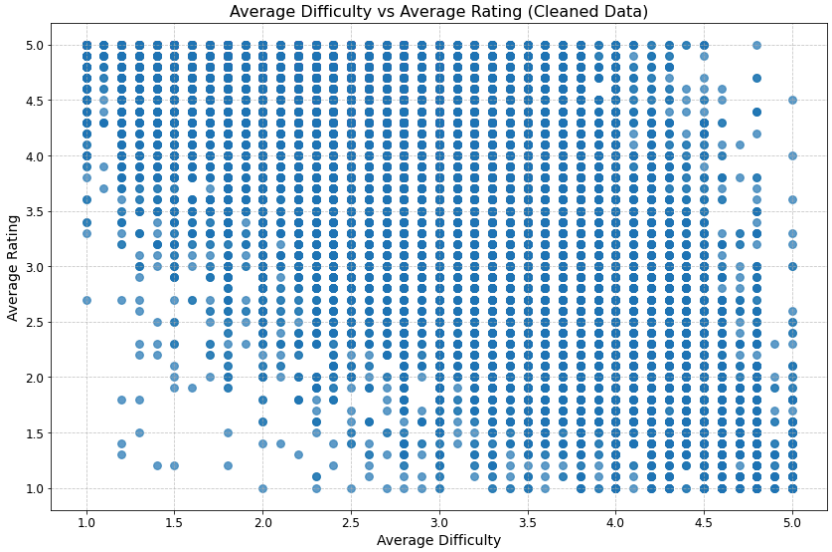

Figure 6

As seen in the graph, there seems to be a negative relationship between rating and difficulty, which implies that as the average difficulty increases, the average rating seems to

decrease. After looking at this graph, I could not determine whether to use Spearman's rho or Pearson's r to quantify the relationship. In order to determine this I decided to plot the residuals and see their distribution. If they are normally distributed then it is fair to use Pearson's r as it would not be heteroskedastic. What is heteroskedasticity? Heteroskedasticity is when the variance of residuals is not constant across all levels of the independent variable.

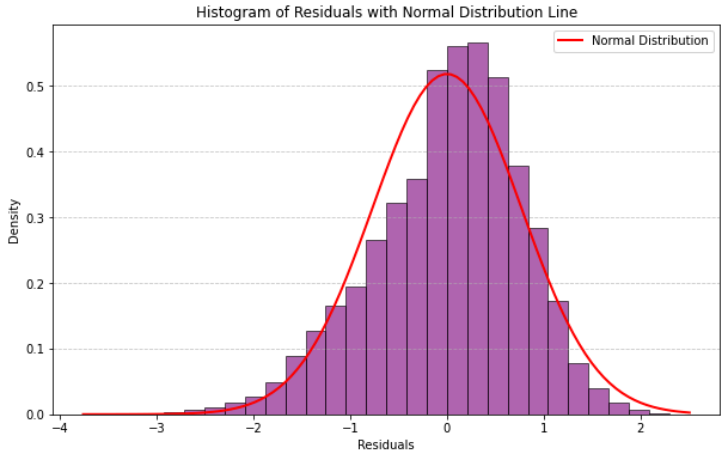

Figure 7

After plotting the residuals, they seemed to follow a distribution similar to a normal one but not exactly. Since it is not an entirely conclusive answer I calculated both and got the following results. Spearman's rho = -0.590 and Pearson's r =-0.607. Since both values are close to each other, it can be concluded that there is a strong, negative relationship between the average difficulty and average rating.

I also calculated the rho and r using a higher threshold of 18, meaning professors with 19 or more ratings, which gave me a rho of -0.665 and an r of -0.645. I did this to verify if the bigger threshold would completely skew the results or not, but it can be seen that the pattern is the same, but clearer and with less noise with the higher threshold. So, it is safe to say that there is a strong, negative relationship between average rating and difficulty.

My next step was to further quantify the relationship between two variables by finding the impact of average difficulty on average rating. However, since I am already doing that for question 7, I will leave that for later.

Question 4:

The first thing I did, was split the professors into two groups: an online group and an offline group. As many professors have taught very few or no online classes compared to professors who teach online regularly, I split them by doing the following. The offline professors have 0 of their reviews coming from online classes, meaning they are fully online, and the online professors have 50% or more of their ratings coming from online classes. The reason I did this split is to filter out all the professors that have taught or teach a few online classes that could skew the results, and this was the best way to split into an online (1,596 rows) and offline (25,889 rows) group with as little data loss as possible. Once that is split, I plotted the data to see the distributions.

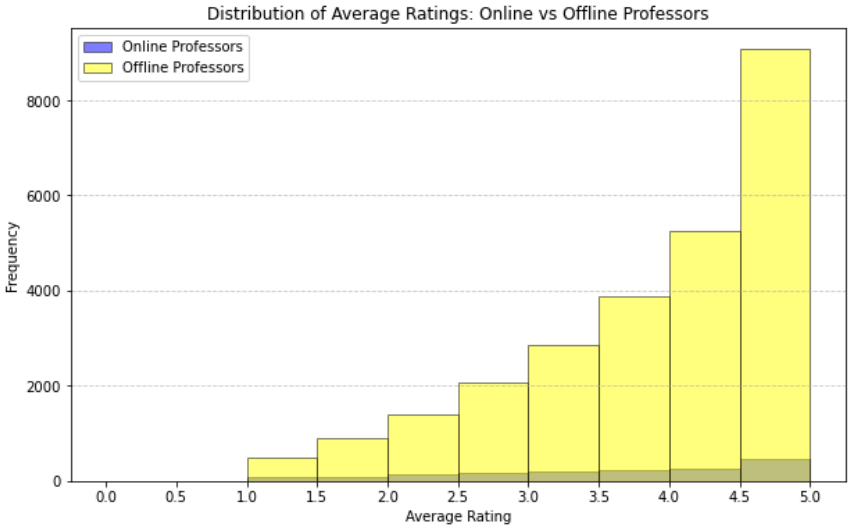

Figure 8

As seen in Figure 8, the data is not normally distributed, so I ran a Mann-Whitney U Test. Due to the vast difference in sample sizes, I will also run a Power Analysis to show that this sample size is enough. After running the hypothesis test, I got $1.89 * 10^{-18}$, which shows a statistically significant difference in the median ratings of both groups.

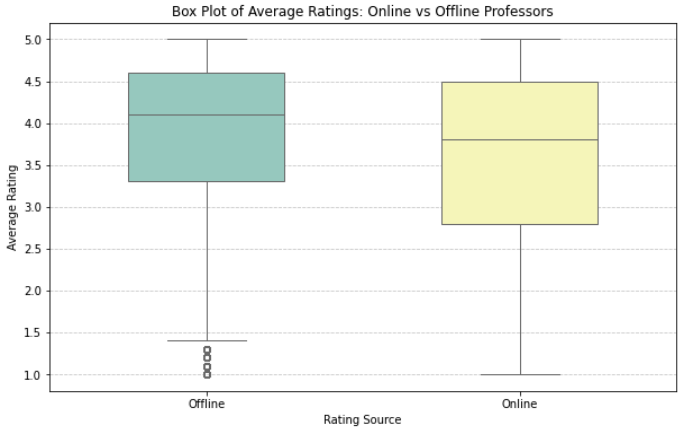

Figure 9

Offline professors are rated higher with a median of 4.1 compared to the offline one of 3.8. The next step was to calculate the effect size, which was -0.256, implying a moderate difference between the medians. Compared to the previous 2 questions, this value is now in the moderate range compared to the small range. After calculating all the values, I ran a power test that yielded a result of 1.000. This shows that despite our small sample size, it is enough to determine significance.

Question 5:

For this question, we plot the professors' average rating against the proportion of people willing to take it again to see whether they have a relationship. The biggest issue with this question was the number of nans in this column. After filtering all of the nans, only 12,160 of the rows have a value in the "PropTakeAgain" column. Since we only have 17% of the professors,

including those with very few ratings, I made the decision to drop all rows instead of imputing to stay consistent with the rest of the questions as it would be very difficult to impute ratings of professors given we have such little data. Smart imputation would lead to very inaccurate results, and regular imputation would really heavily skew the results, making it impossible to derive an insightful conclusion.

For further cleaning, I used the same threshold used in previous questions to maintain consistency, so we are using professors that have more than 4 or more ratings and have a value in the PropTakeAgain column. The graph can be seen below.

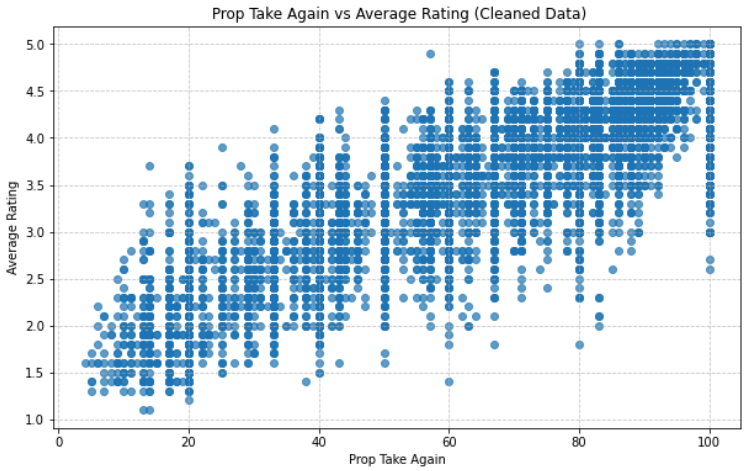

Figure 10

Although the variance seems to be more equal across, it does not seem homoscedastic, we will continue using both Spearman's rho and Pearson's r. Another reason behind using rho and r is to keep it consistent with question 3 and compare the effects of Proportion Would Take Again on the Average rating.

The rho value we get for this question is 0.852 and the r value was 0.880, which is even stronger than rho. Both imply a very strong relationship between average rating and the proportion of people who would take it again.

To further confirm the idea that there is a strong relationship between both, I used a higher threshold, again 19 or more professors, to see if the relationship changed. Once applied, I got a rho value of 0.866 and an r value of 0.896. Hence, we can conclude that there is a strong, positive relationship between the proportion of people who would take the class again and the professor's average rating.

Question 6:

For this question, I split the data into a group of "hot professors" (received a 1 in the pepper column, 12,765 rows) and a group of professors that are not considered hot (received a 0 in the pepper column, 19,186 rows). In this question, the ratings are not normally distributed so we will be using the Mann Whitney-U test to test for statistical significance.

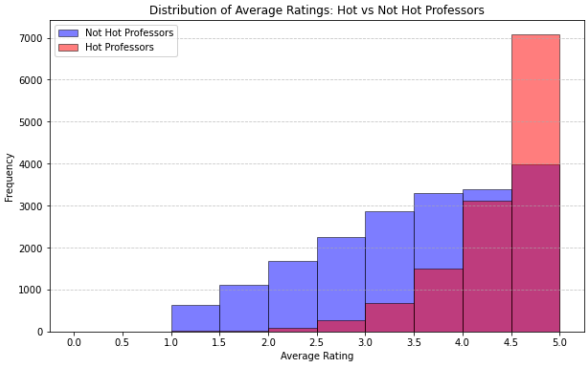

Figure 11

The hot professors have much lower frequencies in the lower ratings, which is expected as they are a smaller sample. but they significantly overshadow the not hot professors in the highest category, by around 3000 reviews. After running the Mann-Whitney-U test, I got a p-value of 0.0, meaning there was a statistically significant difference between the two groups and it is a p-value small enough that it dismisses any concerns of p-hacking.

To test for the difference between the two groups, we get an effect size of 1.06, which implies that there is a very large difference between the two groups.

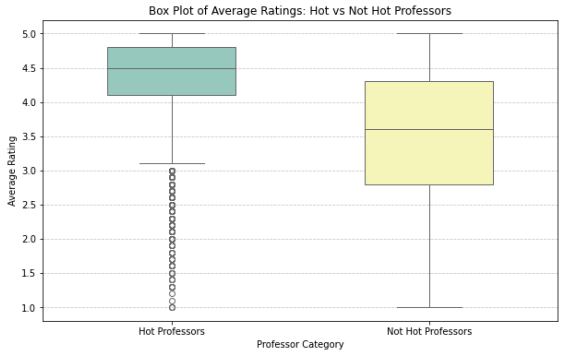

Figure 12

As seen in Figure 12, hot professors are much higher rated than the not hot professors, and the difference between them is large, with the hot professors having a median of 4.5 compared to the 3.6 of not hot professors.

Question 7:

Using the same data from question three, I am trying to quantify the effect between average difficulty and average ratings for professors, in order to predict the average rating of a professor based on the average difficulty rating that they were given, I plot these points and the line of best fit which results in the following graph.

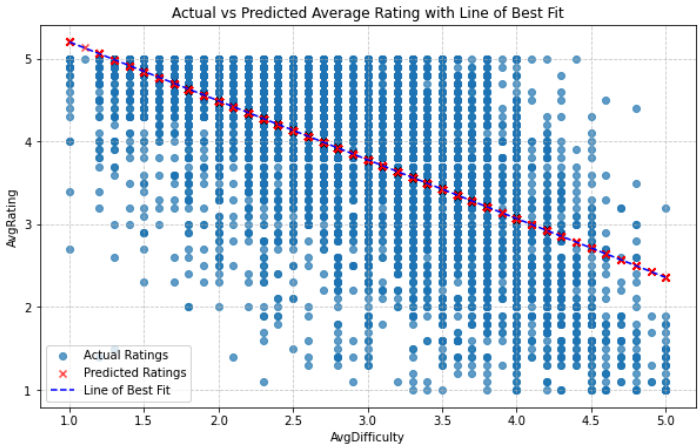

Figure 13

The intercept (β_0) of this line of best fit is 5.92, which means that if the average difficulty is 0, the professor would have an average rating of 5.92. The slope of this model (β_1 coefficient) is -0.711, which means that for every increase in average difficulty by 1 rating, the average rating decreases by 0.711.

There are other important statistics to keep note of, which are R^2 and RMSE (Root Mean Squared Error). R^2 tells you how much of the variance in the data is explained by the predictor(s) you have used, Average Difficulty in this case. After calculating the R^2, I got a value of 0.368. This means that 36.8% of the variance in Average Rating is explained by the Average Difficulty rating. Whereas the RMSE was 0.782, meaning the predictions made by your model differ from the observed values by approximately 0.782 units.

Although the numbers seem low and inaccurate, when looking at the problem it does make sense that you cannot determine the quality of a professor just by looking at the difficulty of his courses. There are a lot of good professors that teach harder subjects and when looking at it from that lens, the numbers make sense.

Question 8:

In this question I am trying to predict the Average Rating from all of the factors that are available in the dataframe. The first thing I did was plot the correlation matrix to see if the predictors themselves are correlated and it is important to check because neglecting it could lead to very unstable results. I used the same threshold (professors with 4 or more ratings) that i used in all the significance testing questions, in order to keep consistent with the rest to see their impact on average rating.

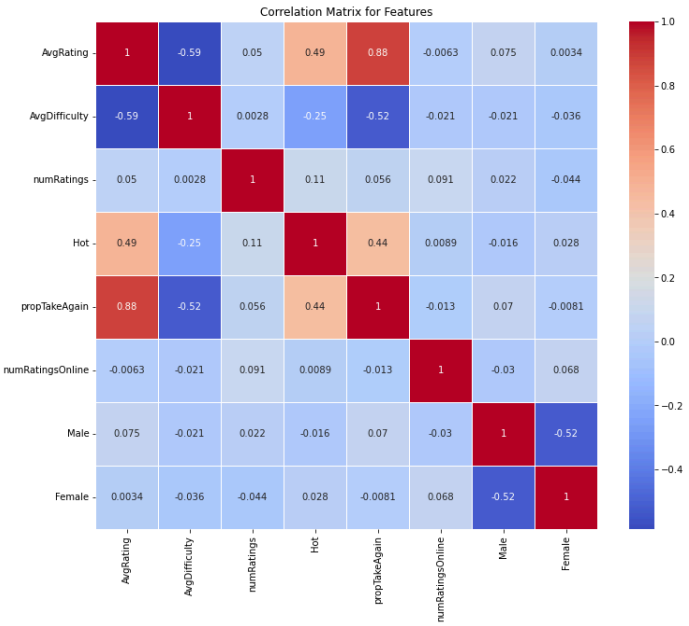

Figure 14

After plotting the matrix, as seen in Figure 14, there are some predictors that are correlated to one another, such as *"Hot" and "Proportion Would Take again"* and *"Average Difficulty" and "Proportion Would Take again"*, which indicates that there might be collinearity. To address this, I opted for Ridge Regression, a regularized form of linear regression. Ridge Regression adds a penalty to the size of the coefficients, shrinking them toward zero to reduce overfitting and improve generalization.

The reason I chose Ridge over Lasso is due to the earlier questions. Despite some of the smaller effect sizes, we saw that modality (online or offline), gender, experience and attractiveness were all significant in the previous questions, so I did not want to exclude any of them from the model, but rather minimize their impact while capturing their effects. To find the most appropriate lambda, I looked it up online and found a technique called *Cross-Validation Using RidgeCV*. After running the code, it gave me a lambda value of 11.5, and that is what I used for my regression.

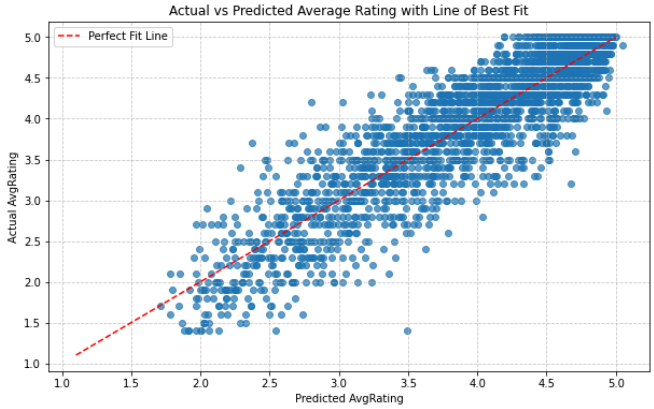

Figure 15

Once the regression is run, I got the graph in Figure 15, which plots the predicted ratings against the actual runs. I chose this plot because we have 7 predictors, and it is impossible to visualize, so I reduced it to 2 dimensions.
As discussed in question 7, we now know what coefficients are and here they are:

AvgDifficulty	-0.200116
numRatings	-0.000286
Hot	0.207424
propTakeAgain	0.024858
numRatingsOnline	-0.001397
Male	0.043945
Female	0.028387

The most surprising result out of this is that of numRatings. As seen in question 2, professors with higher experience are higher rated, yet every increase in number of ratings leads to a decrease in the average rating by -0.000286. Albeit it being -0.0002, that means when you have 1000 reviews on average, you would lose 2 whole ratings.

Moving on to different metrics, the R^2 value of this model is 0.805, which corresponds to explaining 80.5% of the variance in average ratings with the 7 predictors. The RMSE of this model is 0.375

Question 9:

To answer this question, I used logistic regression to predict whether a professor receives a "pepper" (i.e., is considered "hot") based on their average rating.

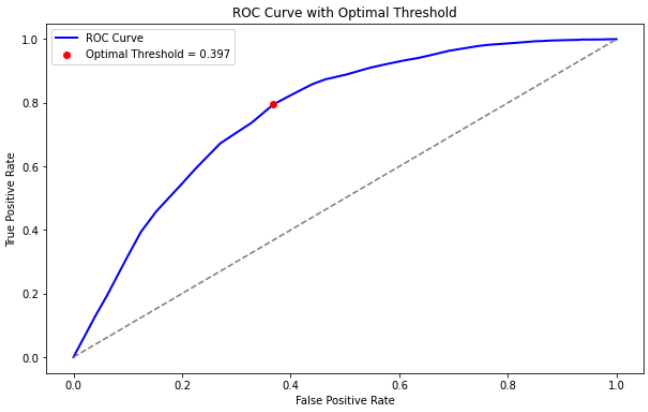

Figure 16

To select the best model, I used the ROC curve to find the optimal threshold, balancing false positives and true positives. I chose the top-left point, yielding a threshold of 0.397. Professors with a probability of 0.397 or higher are classified as "hot."

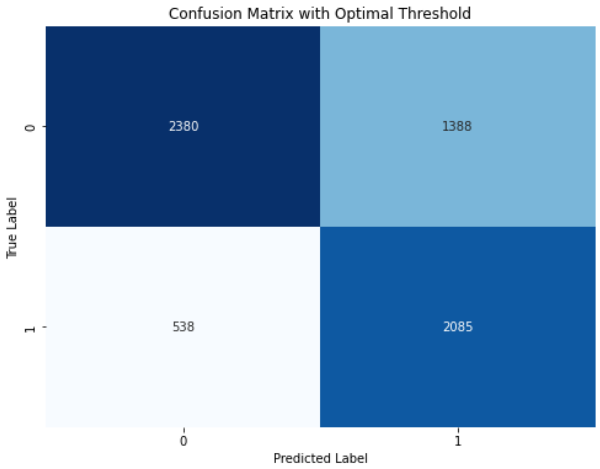

Figure 17

With the new threshold applied, I plotted a confusion matrix (Figure 17) to evaluate the model, which achieved an accuracy of approximately 74.9%. As discussed in class, accuracy on its own is not enough, so I plotted the confusion matrix to take a deepr look. The model's sensitivity $(True\ Positives\ /\ (True\ Positives\ +\ False\ Negatives))$ was 79.5%, meaning that it did a good job of finding professors who were rated as hot. However, its specificity $(True\ Negatives\ /\ (True\ Negatives\ +\ False\ Positives))$ was lower, at 63.2%.
The AUC score of the mode was 0.746, which supports the idea that you are able to classify whether a professor is hot or not via average rating. But what is an AUC score? The AUC score measures the area under the ROC curve.. It ranges from 0.0 to 1.0, with 0.0 being that it misclassified everything, and 1.0 having classified everything. While not perfect, this AUC is fairly high.

Lastly, I saw that the classes were different in size, with a ratio of around 1.58:1 of "not hot" professors to "hot" professors. After looking up the solutions to this issue online, I found 2 main solutions, resampling and removal of rows in the bigger sample. Resampling can introduce issues similar to imputation, such as artificially inflating the dataset and potentially skewing results, or removal of random rows in the bigger class, which could lead to loss of important data. Due to this, I decided against doing any weighting or resampling to deal with imbalances, and left the dataframe as it was.

Question 10:
For this question, I built a logistic regression classification model to predict whether a professor receives a "pepper" based on all available factors, including average rating. I also only included professors with value in all 7 predictors, as I decided against imputation, and therefore the only viable solution was dropping rows.

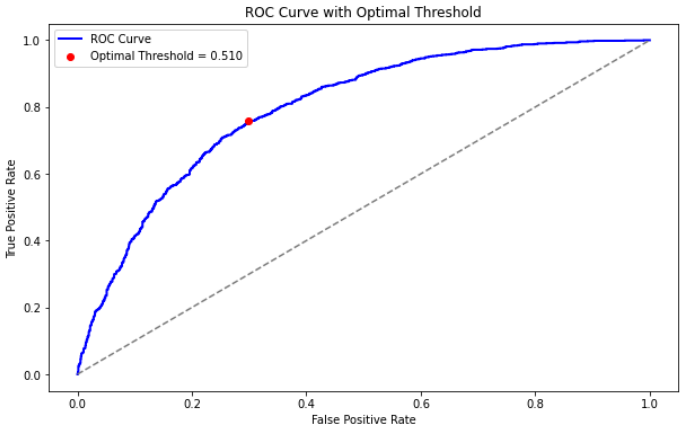

Figure 18

To find the threshold for this model, I plotted an ROC curve and found the most balanced point (top left). Since there is no specific detrimental impact from either false positives or false negatives, I felt this was the most sensible approach. This resulted in threshold of 0.510, meaning that if a professor's predicted probability of being "hot" is 0.510 or higher, they are classified as such.

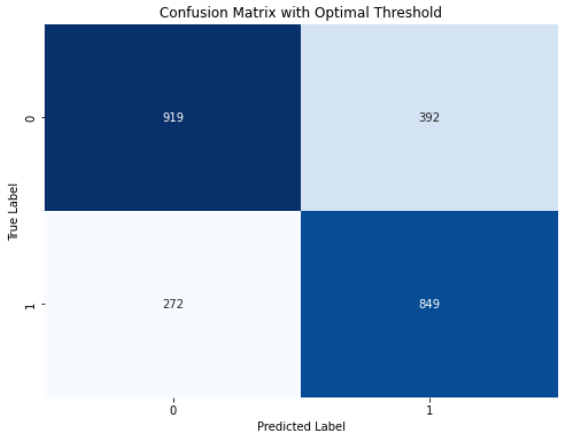

Figure 19

I plotted a confusion matrix (Figure 19) to evaluate the model. The model achieved an accuracy of 72.7%. To take a closer look, I calculated additional metrics: the model's sensitivity was 75.7%, meaning it performed well in identifying "hot" professors. Its specificity was slightly lower, at 70.1%.

Although it seems initially surprising that a model with more predictors, that worked very well when predicting average rating, worked worse than predicting with just the 1 predictor, it makes sense when we consider real world context. A professor's "hotness" is not indicated by the difficulty of the subject they are teaching, how many people have rated them on the website or how many online classes they teach. Therefore, predicting it just by using an average rating might be better suited.

Lastly, I did not address class imbalances in this question, as 5666 professors are considered hot and 6494 that are not considered hot (ratio of 1.15:1 of hot professors to not hot professors), since the difference is not considerable enough.

Extra credit:

The first thing I did, was check how many states and majors were eliminated as a whole when applying the threshold (4 or more ratings). To clarify, I meant to check if after applying the threshold, if some states would be eliminated completely, meaning that all professors in the dataframe from said state, have less than 4 ratings. After applying the threshold, I saw all of the 21 eliminated states were all in the UK. This result is not surprising as this website was founded in the US and has seen most of its traffic come from there. So what this means is that North America is by far the most used for this website.

Unfortunately, the majors that were eliminated do not tell us as much as there are some majors that differ slightly from others (Health Sciences vs Health Sciences PE Athletics) and therefore not many meaningful insights could be derived as a result.

After applying the threshold, I wanted to check how US vs non-US professors compare in their average ratings, so I split them into 2 groups. After splitting them into two groups, I plotted their average ratings to check for normality

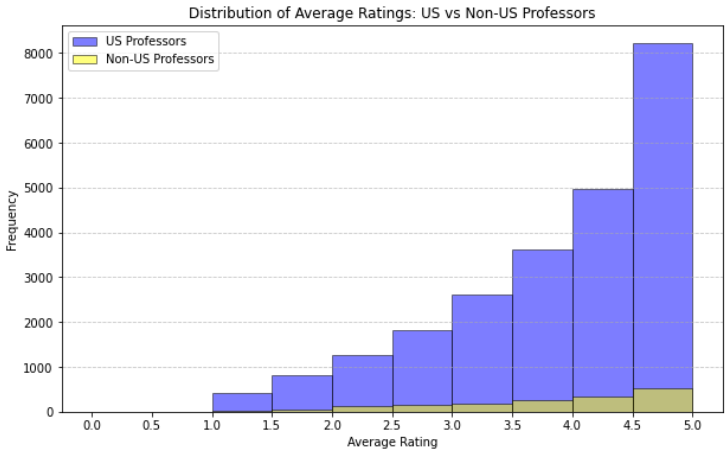

Figure 20

As seen in Figure 20, the ratings are clearly not normally distributed so I ran a Mann-Whitney U test to see if the difference between the ratings was statistically significant.

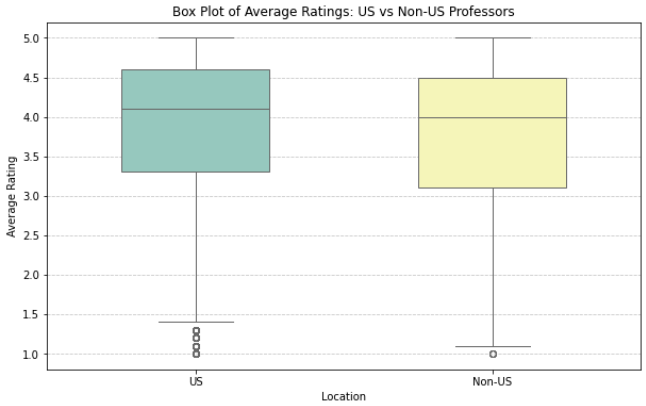

Figure 21

After running the test, I got a p-value of 1.30×10^{-5}, which shows statistical significance and as seen in the box plot, the median rating for US professors is higher, meaning that there is a difference in ratings between US and non-US professors that is not due to chance.

I also calculated Cohen's d, which gave me a value of 0.101, which says that although there is a difference between the median ratings, it is not too meaningful in practice.

Also, as there is a significant difference in the sample sizes non-US (1,652 rows) and US ratings (23,716 ratings), I calculated the power using the G* software and got a value of 0.904, which is a high enough rating and shows that are sample sizes are good enough.

The next thing I did was compare stem professors and non-stem professors. To do this, I created a list of keywords, such as "Engineering", "Biology" and "Physics", and if any of the professors' courses included the keywords, they were classified as stem; and the rest as non-stem. After the classification, I plotted the ratings to check for normality.

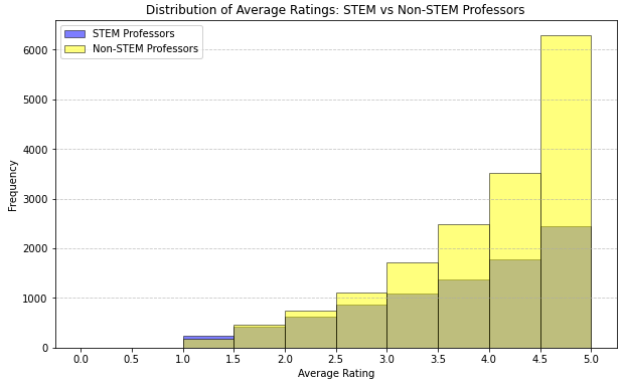

Figure 22

As seen here, it is clearly not normally distributed. Also, there are more non-stem professors than stem professors and the non-stem ones seem to be higher rated. As a result, I ran a Mann-Whitney U test. After running the test, I got a p-value of 2.89×10^{-108}, which is almost 0 and shows a statistically significant difference between stem and non-stem professors.

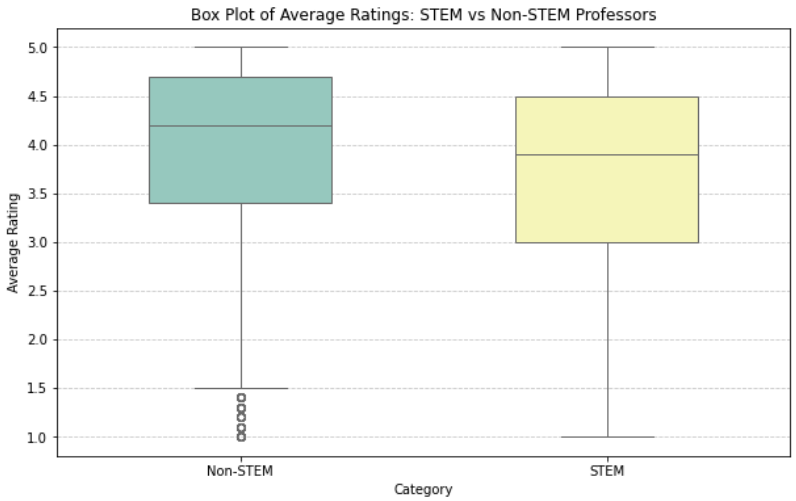

Figure 23

As seen in Figure 23, the non-stem professors (median = 4.3) are higher rated than stem professors (median = 3.9). This is also further supported by the effect size of -0.3. Although there is a difference in the sample sizes, I ran another power analysis and it yielded a power value of 1.00, meaning that the sample sizes are large enough.