Professor Evaluations

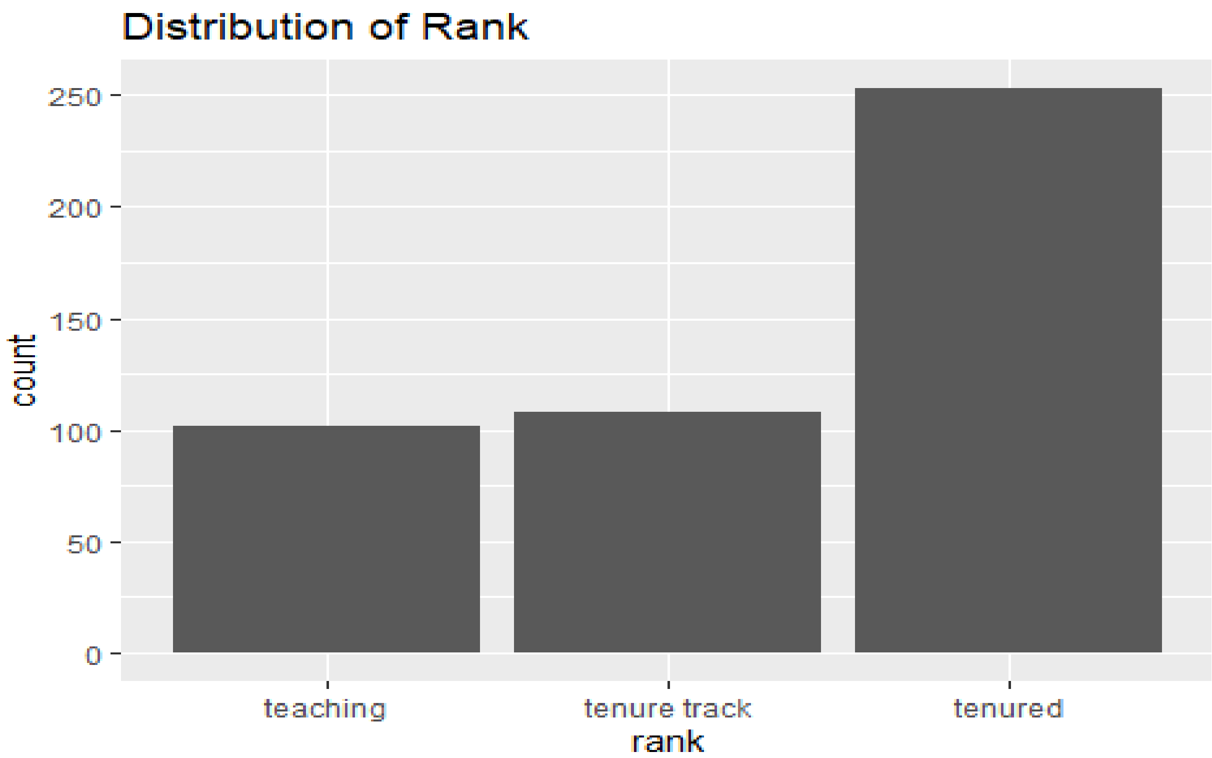
William Duquette

Wednesday, April 22, 2020

Project Description: Schools need to be able to receive accurate and fair evaluations of professors. If there are factors outside of a professor's teaching skill affecting a professor's score, then it is important to acknowledge them to avoid the termination of a professor who got a negative evaluation because of something they cannot control, such as ethnicity or gender. A study conducted at the University of Texas at Austin pooled evaluations from 463 professors at their campus and asked six students to rate the professors on their physical appearance. The main question being asked is: What non-teaching characteristics affect a professor's evaluation score. I will create a model that predicts which of the 15 variables are important to estimate the factors and their significance on a professor's score.

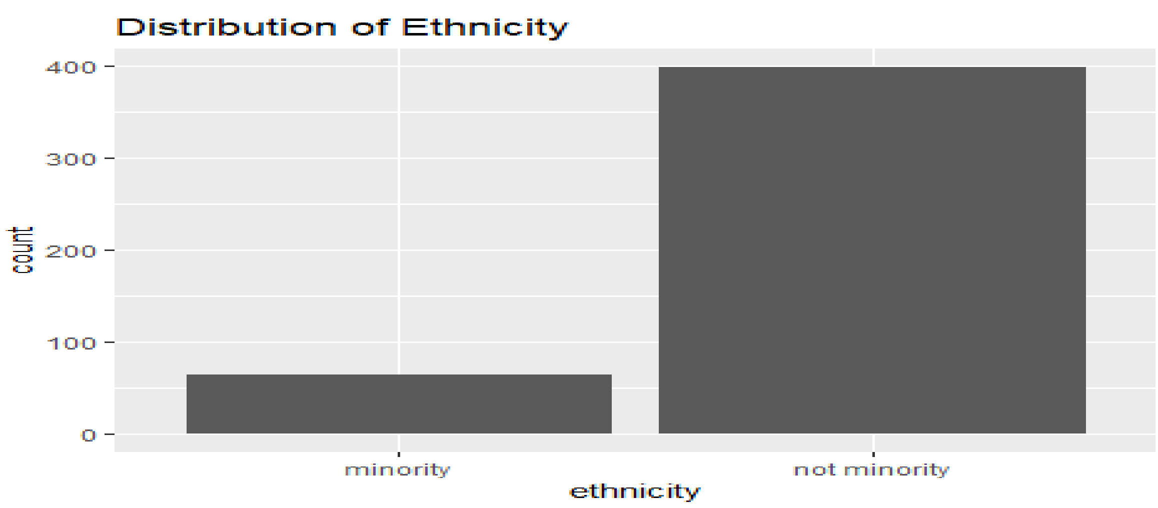
Relevant Variables: 15 variables are inside the data set, but not all of them are significant. For this question, score will be the response variable. This variable, score, measures the average professor evaluation score from 1 to 5 (1 being very unsatisfactory and 5 being excellent) (see below). The explanatory variables that could be in the model are listed below:

1. Rank - This categorical variable denotes whether the professor is teaching, on the tenure track, or tenured. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are more tenured professors than any other rank, while the split between teaching and tenure track is about the same.



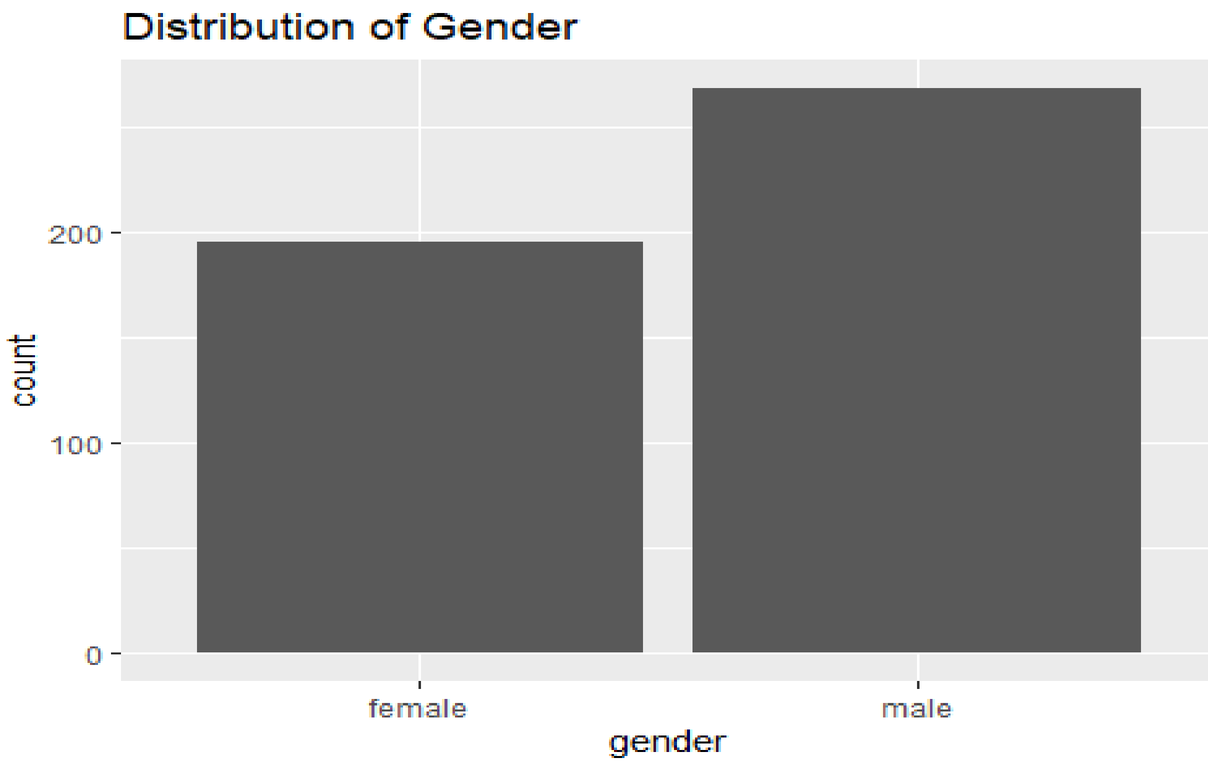
Teaching	Tenure Track	Tenured
102	108	253

2. Ethnicity - This categorical variable denotes whether or not the professor is a minority. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are significantly more non-minority professors than minority professors.



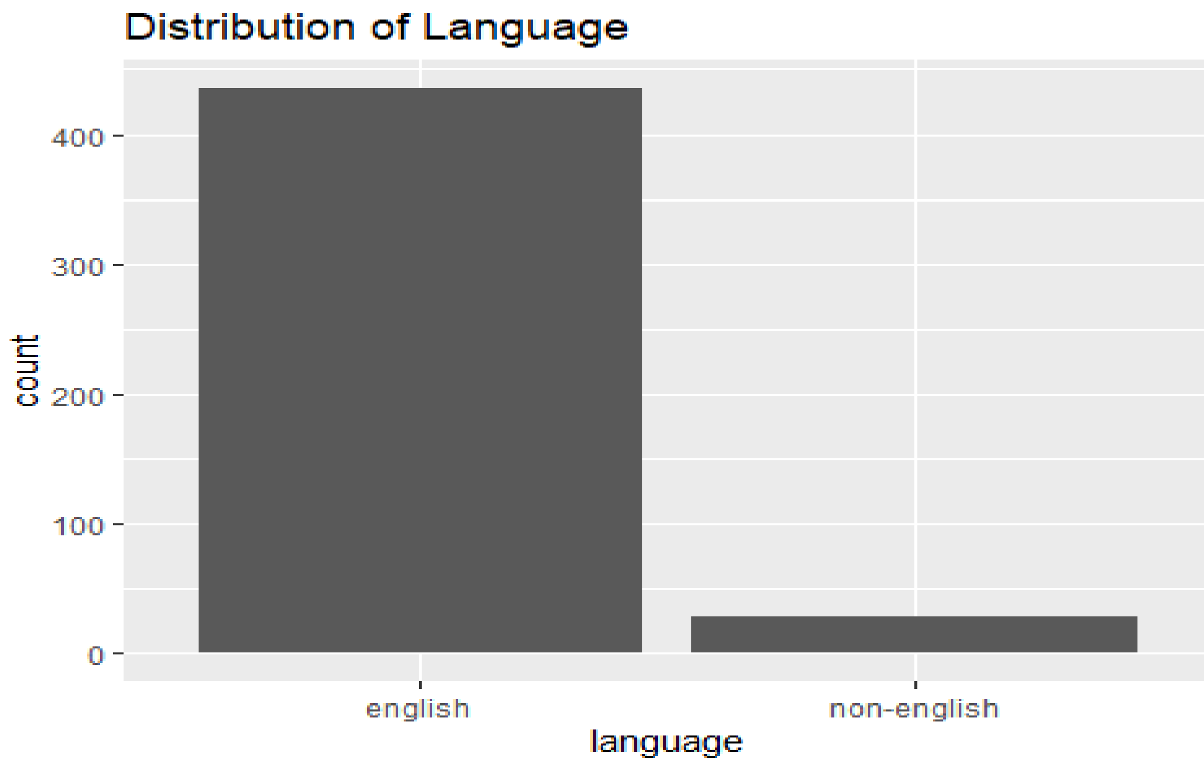
Minority	Not Minority
64	399

3. Gender - This categorical variable denotes whether the professor is male or female. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are 73 more male professors than female professors.



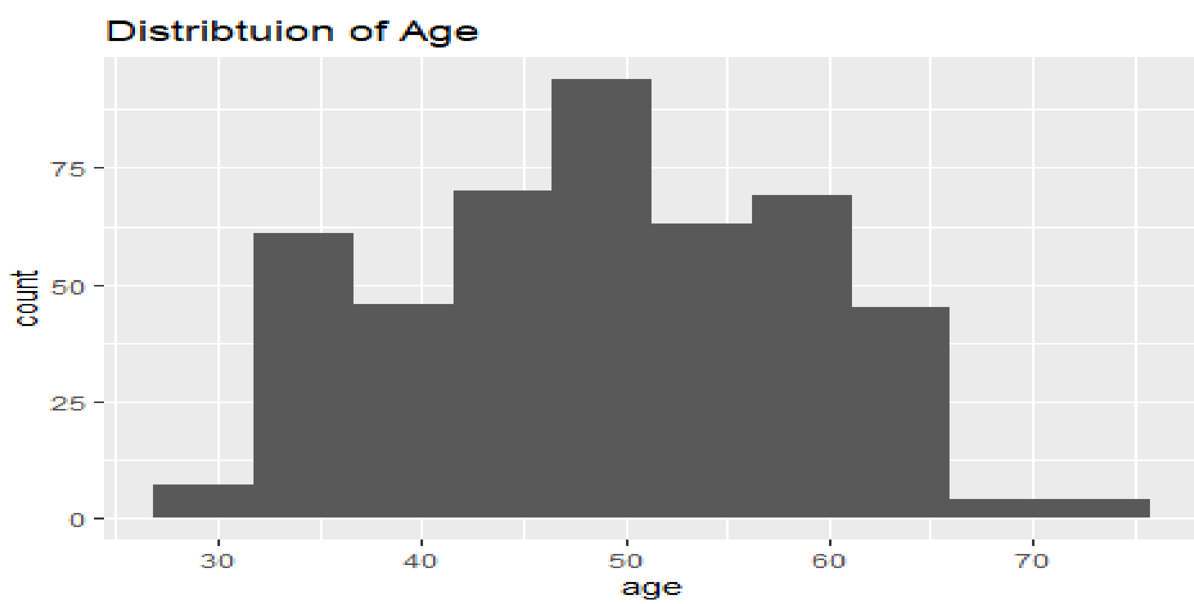
Female	Male
195	268

4. Language - This categorical variable denotes whether the school that the professor received their education at taught classes in English or not. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are significantly more professors that took classes in English than professors that took classes not taught in English.

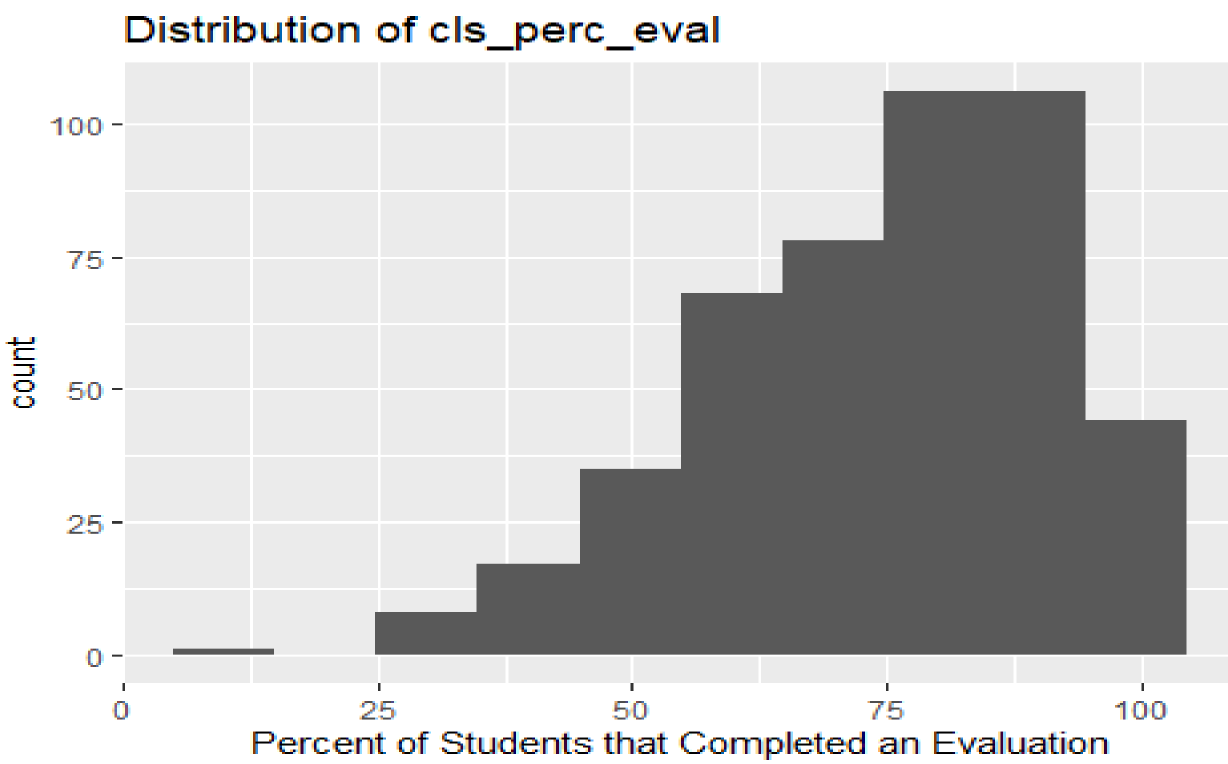


English	Non-English
435	28

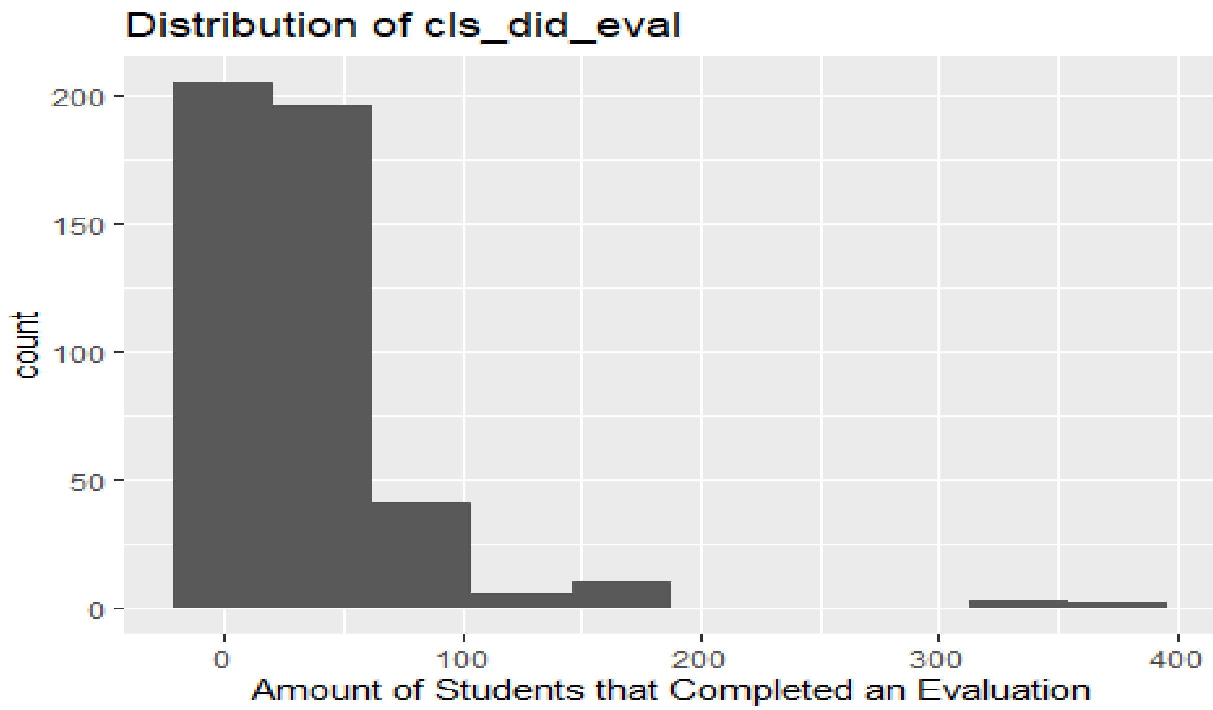
5. Age - This variable measures the age of the professor. Below is a graph that shows the distribution of the explanatory variable. As you can see, the distribution of age is roughly normal.



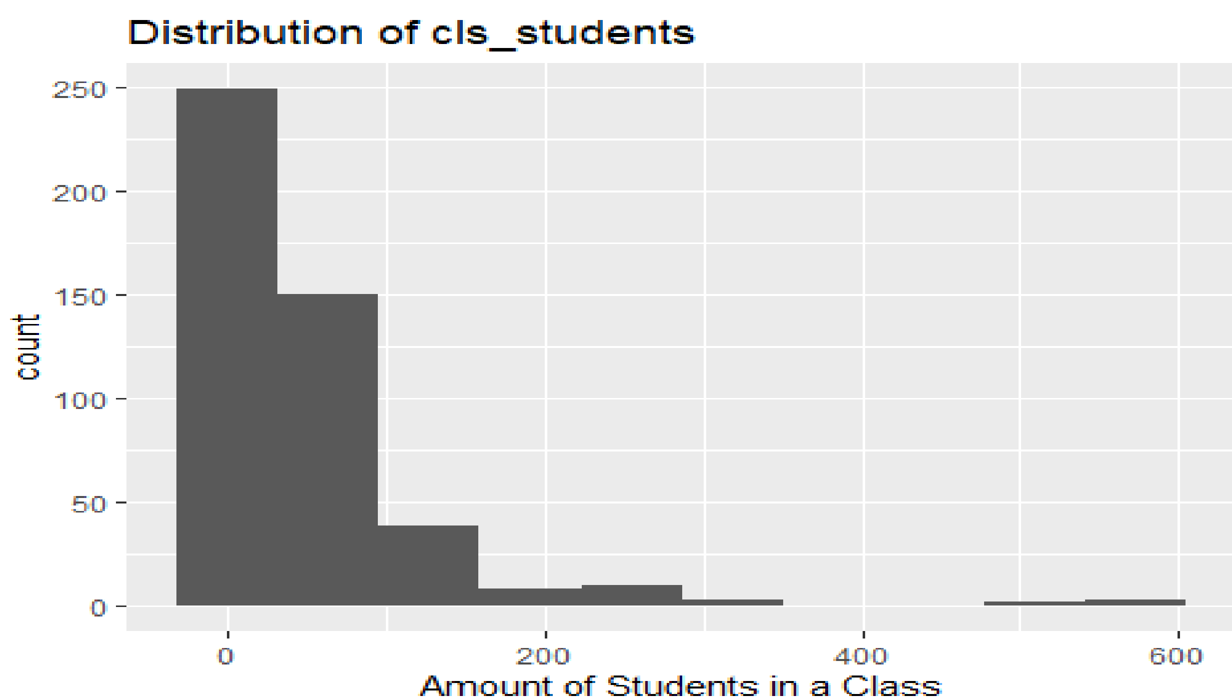
6. cls_perc_eval - This variable measures the percent of students in a class that completed an evaluation. Below is a graph that shows the distribution of the explanatory variable. The distribution of this variable is slightly skewed left but not enough to warrant a transformation. Given that more students fill out a survey as opposed to not fill out a survey, this distribution makes sense.



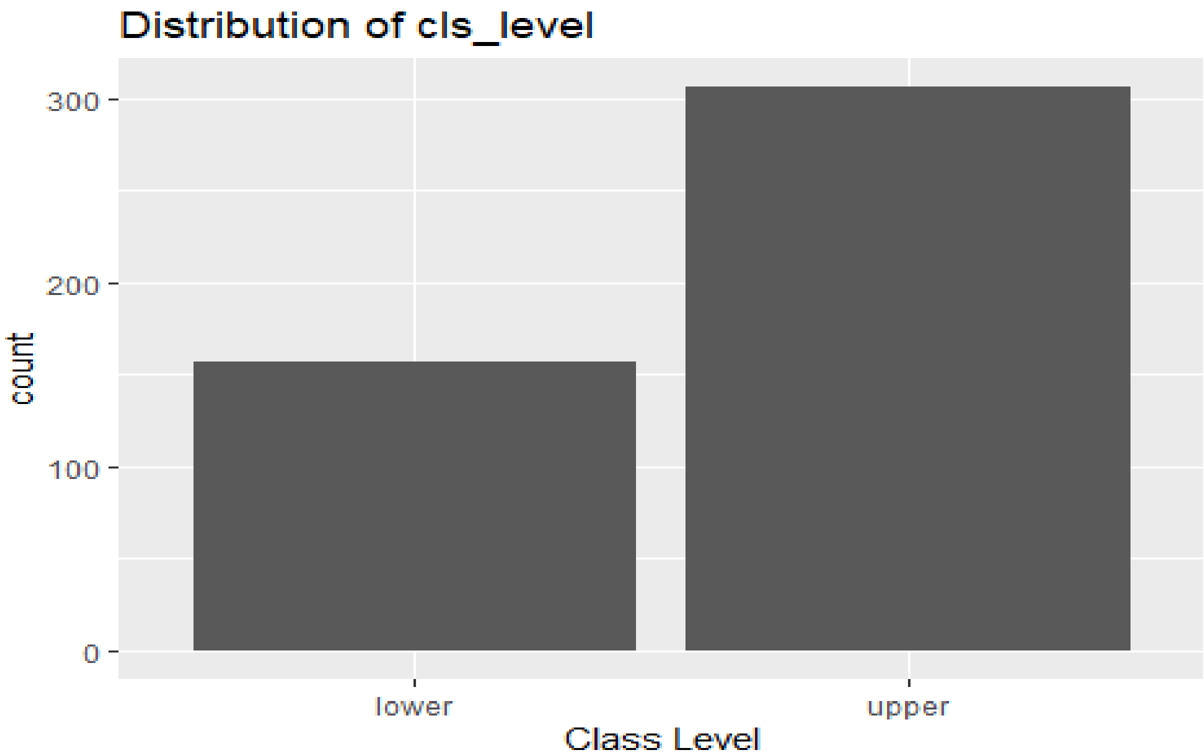
7. cls_did_eval - This variable measures the number of students that completed the evaluation. Below is a graph that shows the distribution of the explanatory variable. I have applied the log() transformation to the cls_did_eval variable. See below.



8. cls_students - This variable measures the number of students in the class. Below is a graph that shows the distribution of the explanatory variable. I have applied the log() transformation to the cls_students variable. See below.

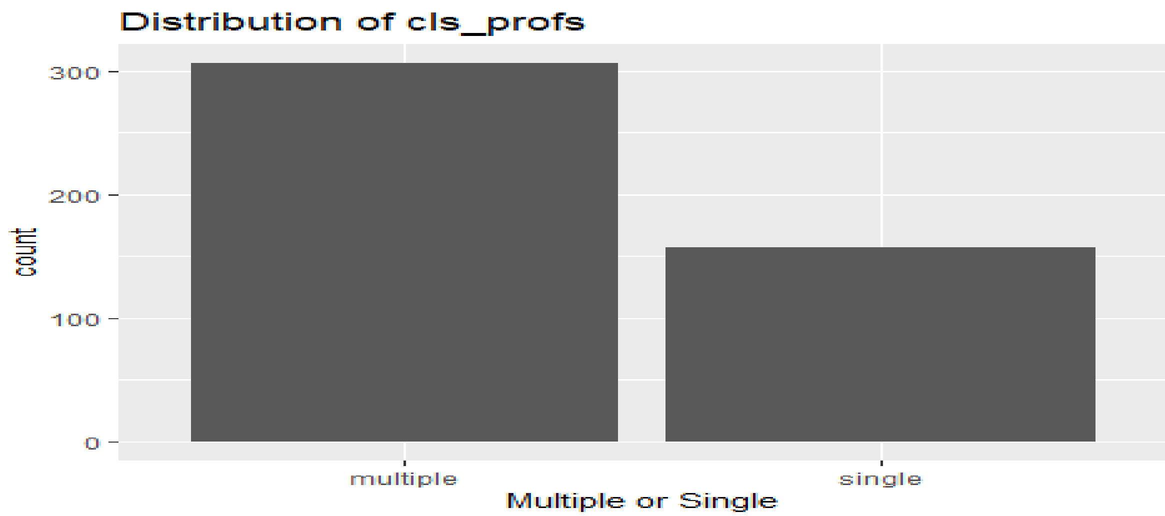


9. cls_level - This categorical variable denotes whether a class level is lower or upper. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are roughly 50% more upper-level classes than lower-level classes.



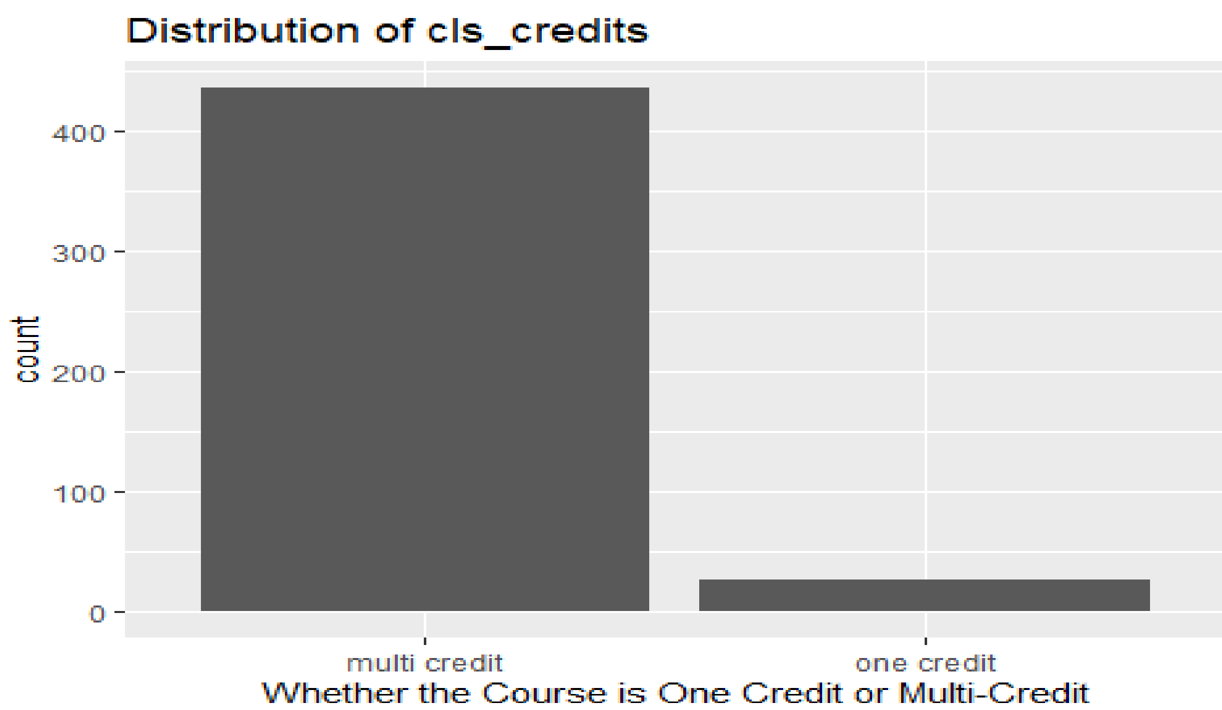
Lower	Upper
157	306

10. cls_profs - This categorical variable denotes whether the number of professors teaching sections in a course in the sample is single or multiple. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are roughly 50% more multiple than single.



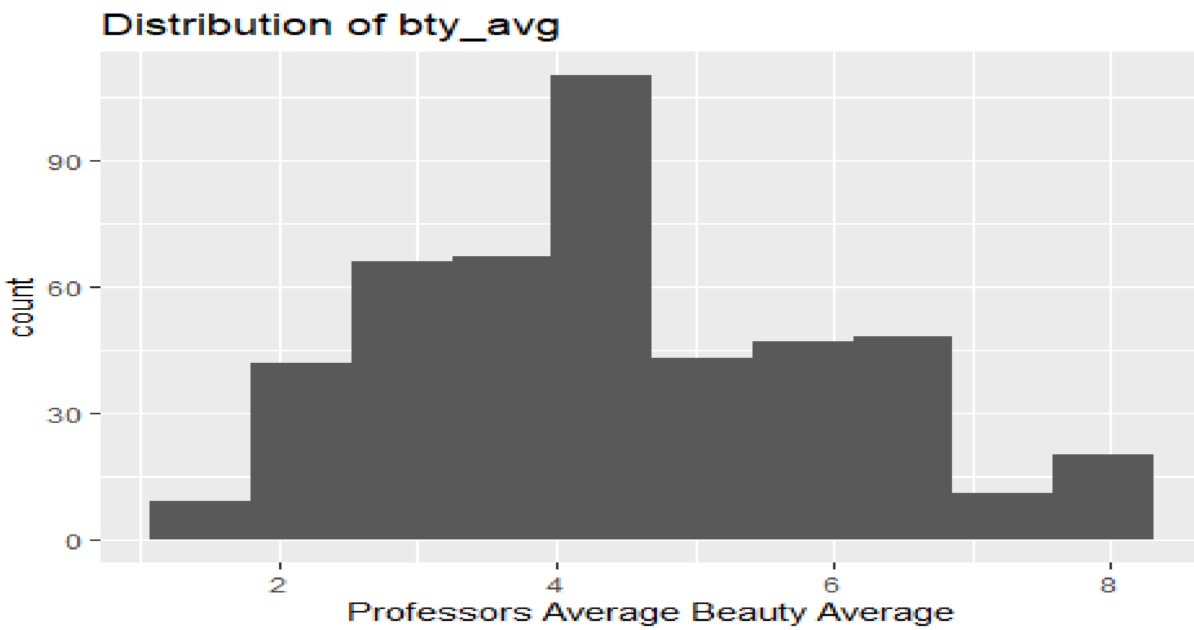
Multiple	Single
306	157

11. cls_credits - This categorical variable denotes whether the course being taught was one credit or multi-credit courses. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are significantly more multi-credit courses.

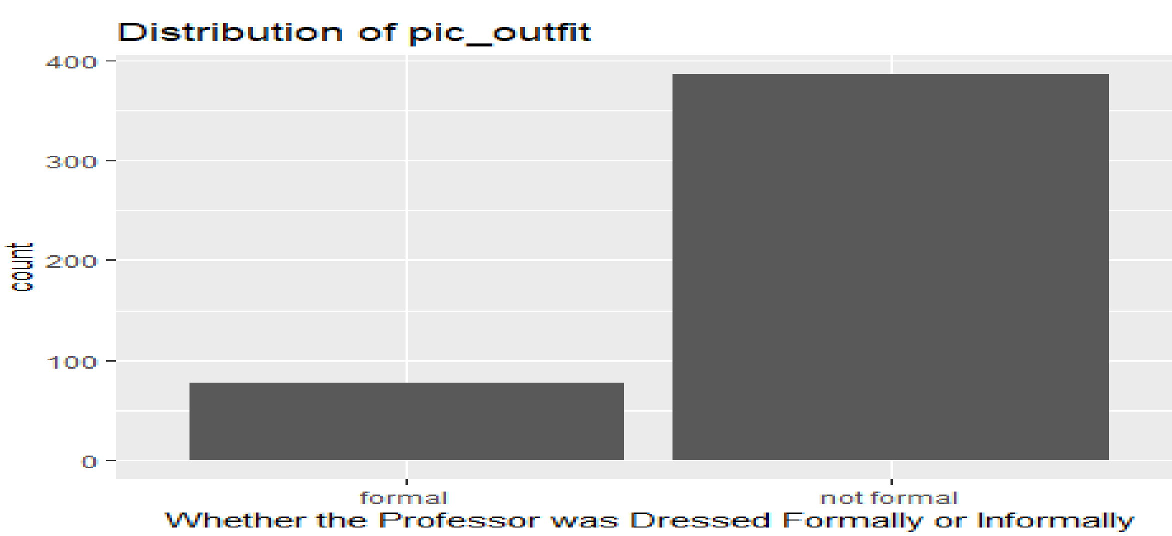


Multi-credit	One credit
436	27

12. bty_avg - This variable measures the average beauty rating of a professor. Below is a graph that shows the distribution of the explanatory variable. The distribution of this variable is approximately normal.

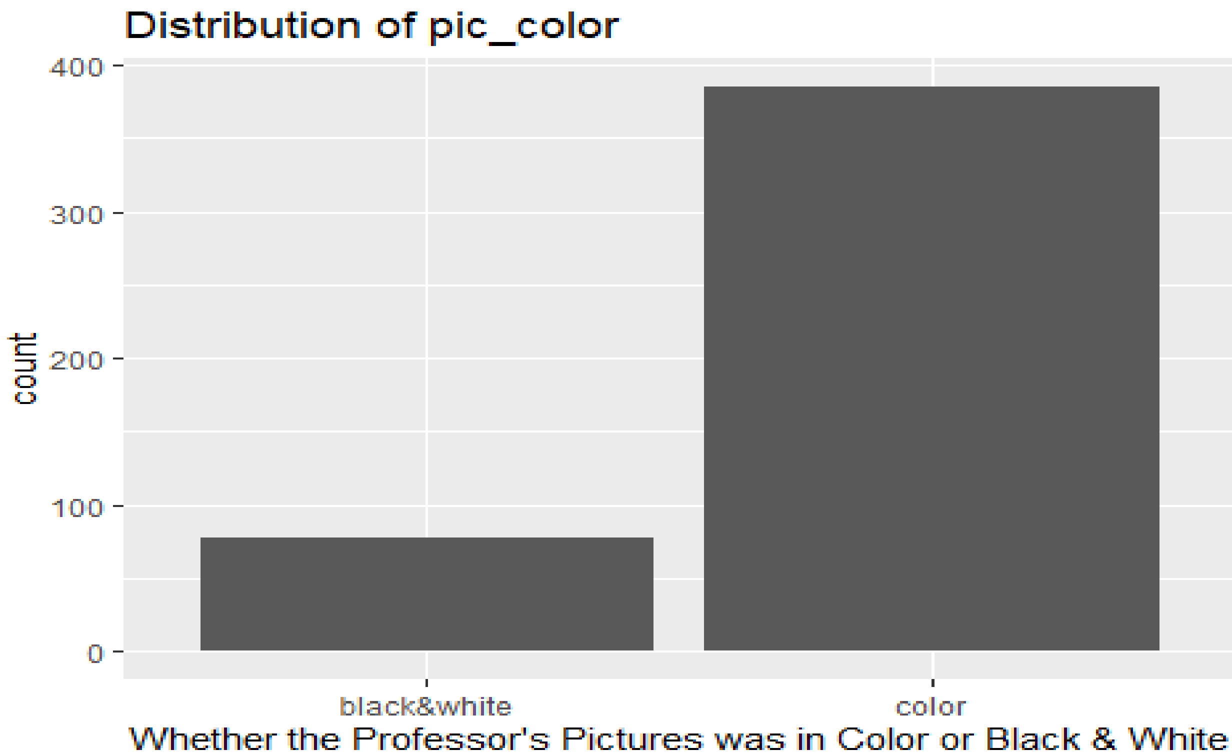


13. pic_outfit - This categorical variable denotes whether the professor was wearing a formal or informal outfit for their picture. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are significantly more not formal outfits in pictures.



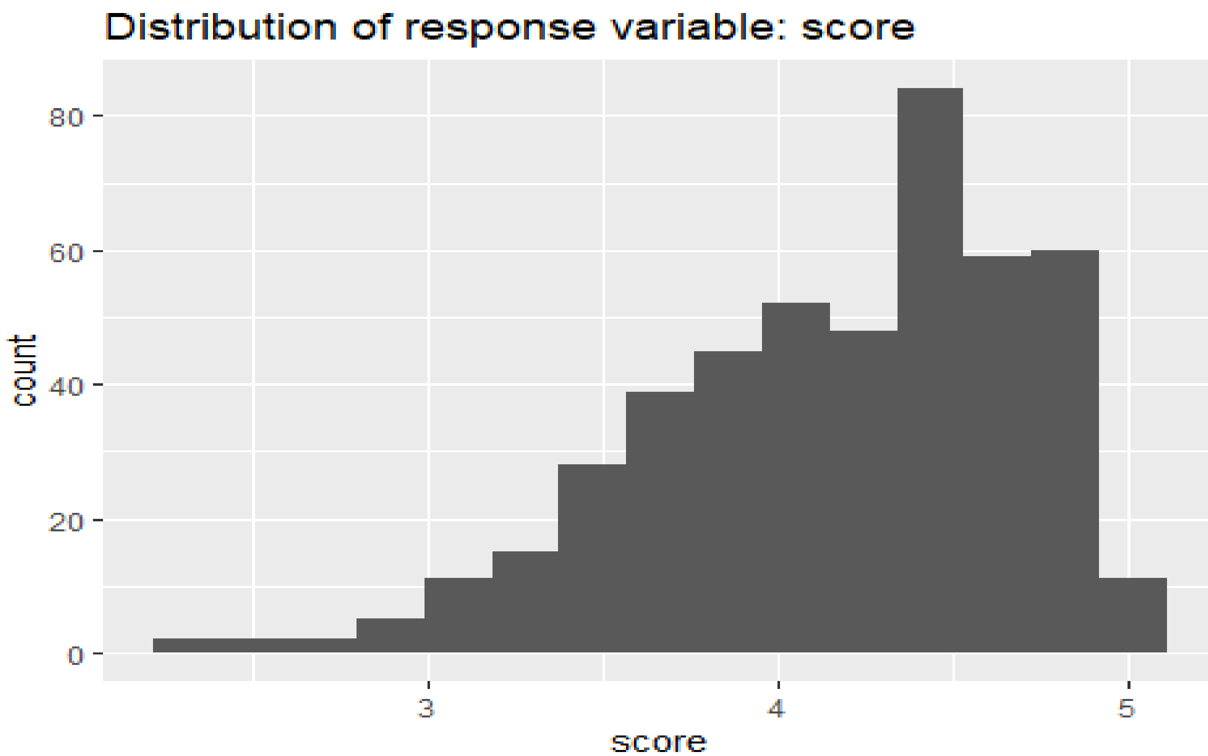
Formal	Not Formal
77	386

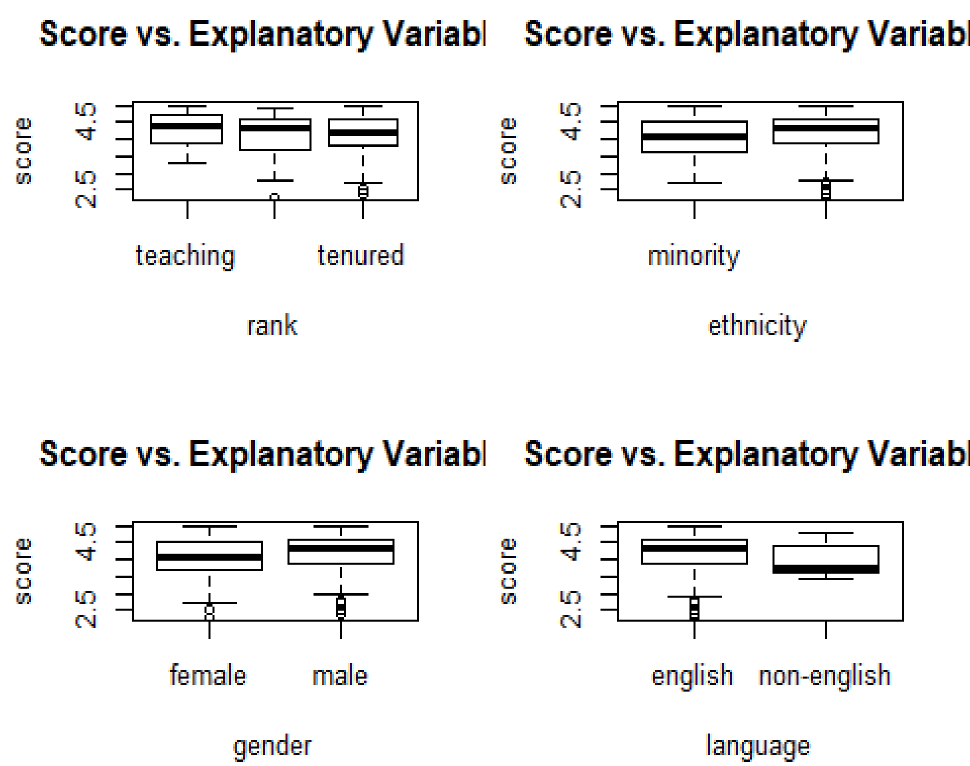
14. pic_color - This categorical variable denotes whether the professor's picture was in color or black & white. Below is a graph that shows the distribution of the explanatory variable. As you can see, there are significantly more color pictures of professors.

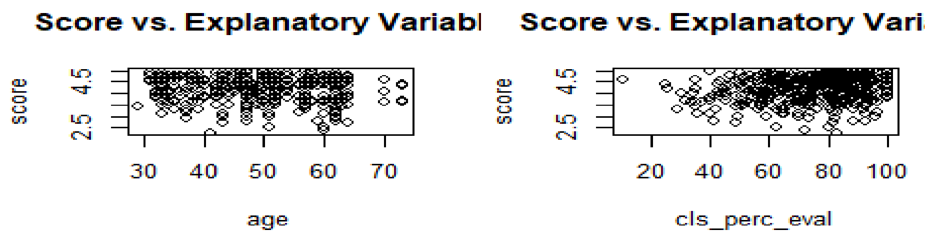


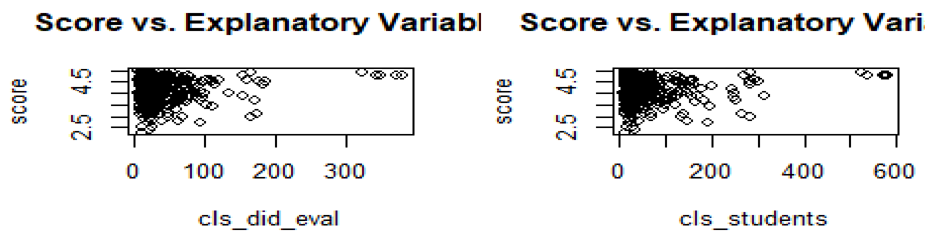
Black & White	Color
78	385

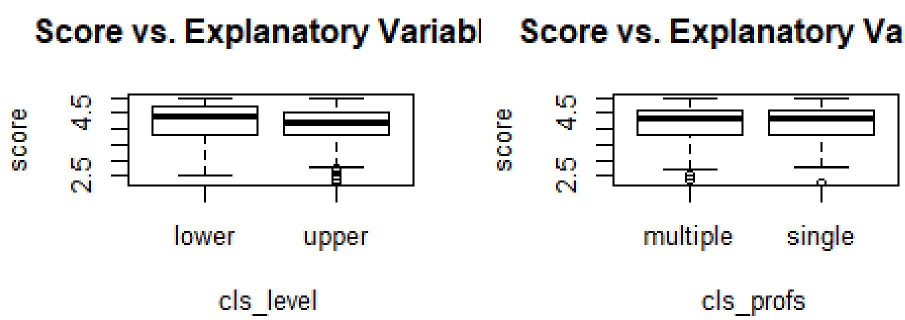
Exploratory Data Analysis: The first graph shown is a histogram showing the distribution of the response variable: score. The data is slightly skewed left, but not enough to warrant a transformation. The second set of graphs, which uses the original data set, is a scatter plot matrix that shows each possible explanatory variable's relationship to the response variable. As you can see, *cls_did_eval* and *cls_students* are skewed right and need to be transformed.

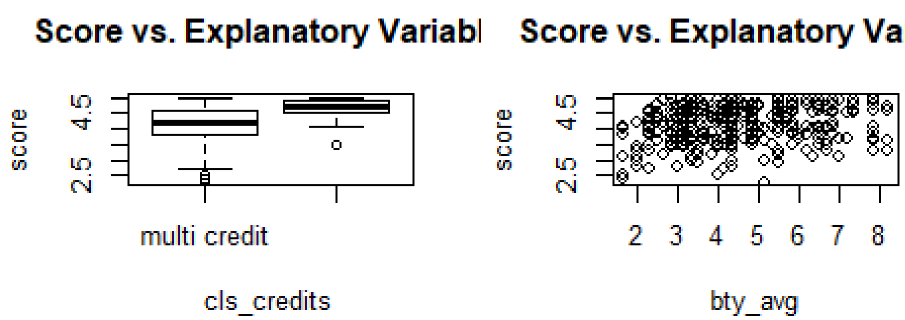

Distribution of response variable: score
count
0
20
40
60
80
score
3
4
5

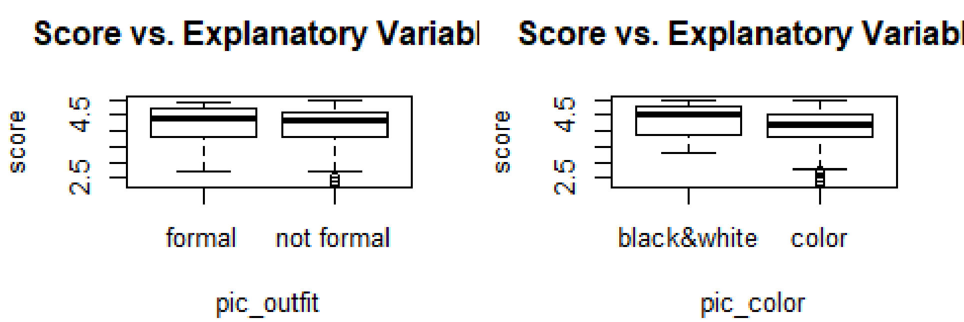

Score vs. Explanatory Variabl
score
2.5
4.5
teaching
tenured
rank
Score vs. Explanatory Variabl
score
2.5
4.5
minority
ethnicity
Score vs. Explanatory Variabl
score
2.5
4.5
female
male
gender
Score vs. Explanatory Variabl
score
2.5
4.5
english
non-english
language


Score vs. Explanatory Variabl
score
2.5 4.5
30 40 50 60 70
age
Score vs. Explanatory Variabl
score
2.5 4.5
20 40 60 80 100
cls_perc_eval

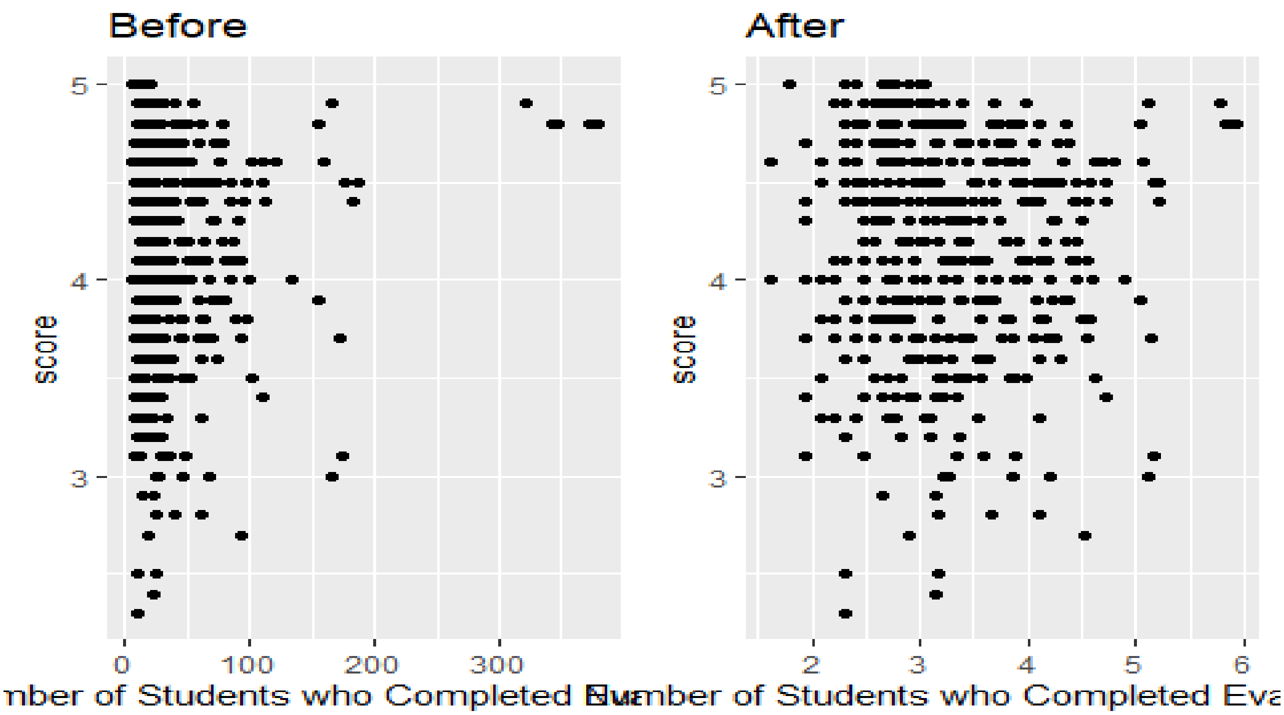

Score vs. Explanatory Variabl
score
2.5 4.5
0 100 200 300
cls_did_eval
Score vs. Explanatory Variabl
score
2.5 4.5
0 200 400 600
cls_students

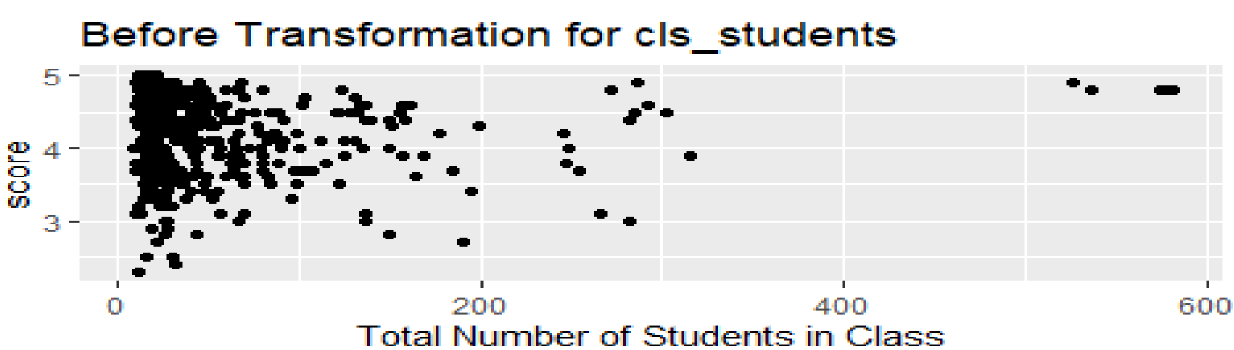

Score vs. Explanatory Variabl
score
2.5 4.5
lower upper
cls_level
Score vs. Explanatory Variabl
score
2.5 4.5
multiple single
cls_profs

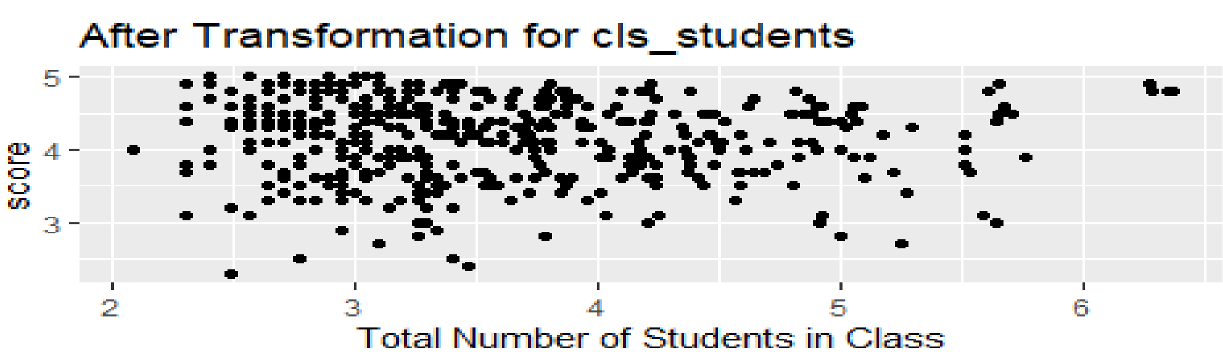

Score vs. Explanatory Variabl
score
2.5 4.5
multi credit
cls_credits
Score vs. Explanatory Variabl
score
2.5 4.5
2 3 4 5 6 7 8
bty_avg


Score vs. Explanatory Variabl
score
2.5 4.5
formal not formal
pic_outfit
Score vs. Explanatory Variabl
score
2.5 4.5
black&white color
pic_color

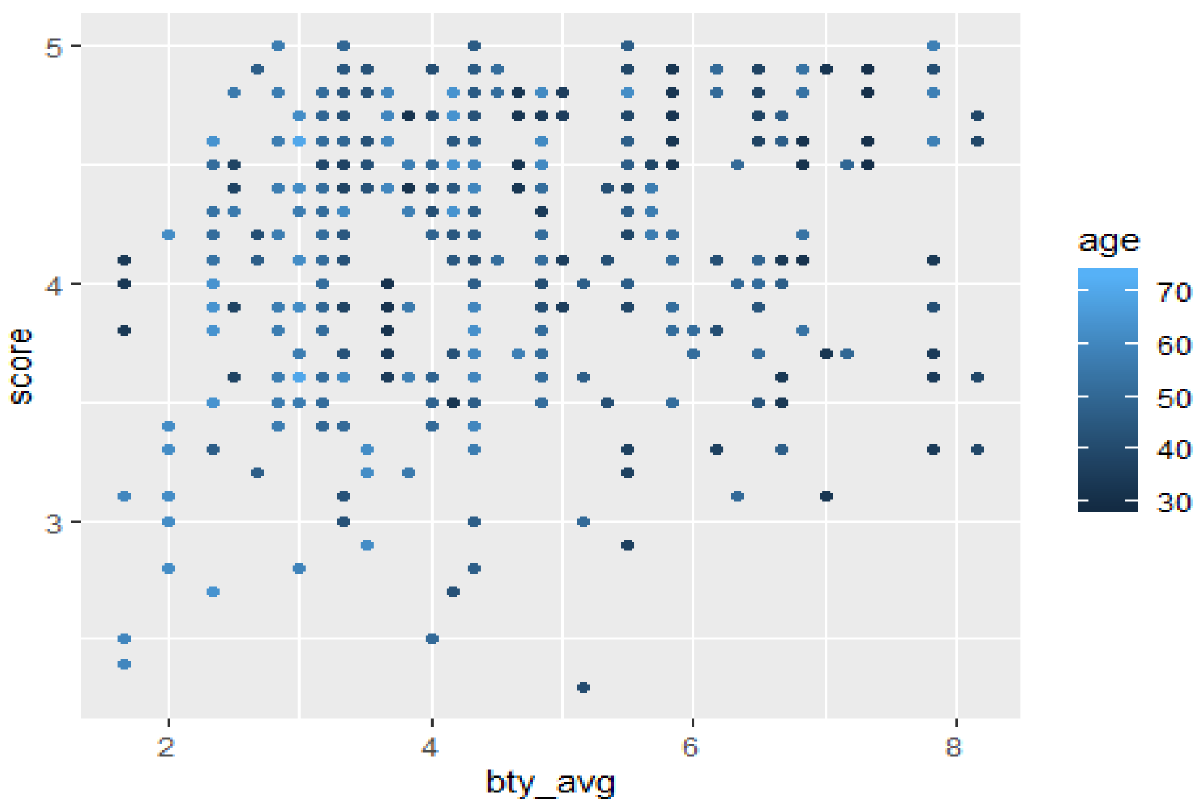
Below you can see the analysis done on *cls_did_eval* and *cls_students* to see what transformation is necessary. I have created graphs that show the variables before and after they have been transformed. The transformation applied was log(). The reason this transformation was used was because both variables cover several orders of magnitude.







Results: I have found two statistically significant interaction terms. An interaction term is a variable that is a function of the current explanatory variables. An interaction term is appropriate inside a model if one explanatory variable depends on another explanatory variable. For my first interaction term, age*bty_avg, you can see that the older a professor is, and the higher the beauty average a professor has, there is a more significant effect on their score.




```
##
## Call:
## lm(formula = score ~ age * bty_avg, data = evals2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9410 -0.3517  0.1231  0.4040  1.0066
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.156077   0.367797  14.019  < 2e-16 ***
## age         -0.026128   0.007342  -3.559 0.000412 ***
## bty_avg     -0.187800   0.075724  -2.480 0.013494 *
## age:bty_avg  0.005318   0.001580   3.366 0.000827 ***
## ---
```

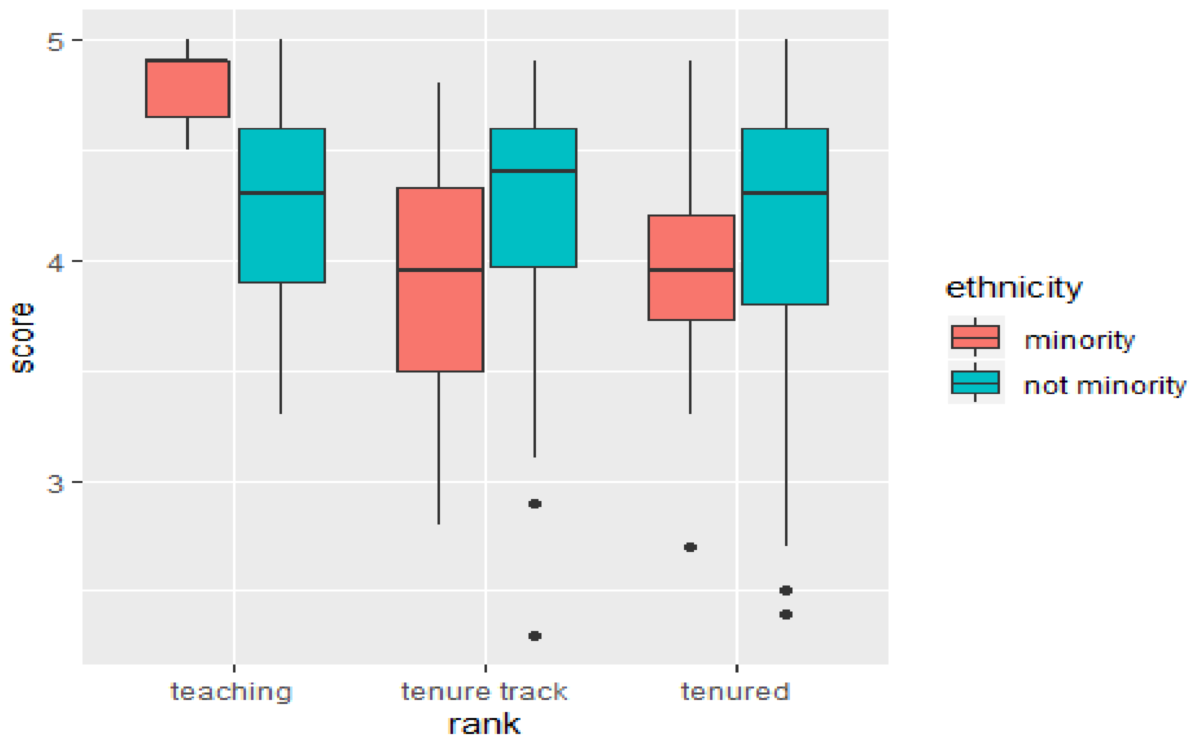
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 459 degrees of freedom
## Multiple R-squared:  0.06096,    Adjusted R-squared:  0.05482
## F-statistic: 9.933 on 3 and 459 DF,  p-value: 2.349e-06
```

The above summary shows the coefficients of the model that shows the interaction term between bty_avg and age. Given that both a professor's beauty average and their age are numeric values, I will use a table to best illustrate the significance of the interaction variable. The horizontal column shows the beauty average, and the vertical column shows the professor's age. As you can see from this table, it is clear that the more attractive a professor is rated, the more likely it is that a professor's score will increase as the professor's age increases.

Table 1

	4	6	8	10
40	4.210570	4.260383	4.310197	4.360010
50	4.161944	4.318161	4.474327	4.630494
60	4.113418	4.375938	4.638458	4.900978
70	4.064841	4.433715	4.802588	5.171462

For my second interaction term, I found that there was an increased effect on a professor's score if the professor was both a minority and tenured, on the tenure track, or teaching. As you can see from the graph, minority professors are judged more harshly than their non-minority counterparts on the tenure and tenure track. One thing important to note is that teaching is the only category where professors who are a minority are scored higher. One reason this might be the case is because of the relatively small sample size. The sample size for both minority and teaching professors is 10, compared to 28 for tenure track and minority and 26 for tenured and minority professors.




```
##
## Call:
## lm(formula = score ~ rank * ethnicity, data = evals2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9438 -0.3355  0.1419  0.3864  0.9269
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            4.8100     0.1678  28.657  < 2e-16 ***
## ranktenure track                      -0.9100     0.1955  -4.654 4.27e-06 ***
## ranktenured                           -0.8369     0.1975  -4.237 2.74e-05 ***
## ethnicitynot minority                 -0.5828     0.1767  -3.298 0.001051 **
## ranktenure track:ethnicitynot minority   0.9266     0.2117   4.377 1.49e-05 ***
## ranktenured:ethnicitynot minority        0.7679     0.2081   3.690 0.000252 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5308 on 457 degrees of freedom
## Multiple R-squared:  0.05783,    Adjusted R-squared:  0.04752
## F-statistic:  5.61 on 5 and 457 DF,  p-value: 4.958e-05
```

The above summary shows the coefficients of a model made to analyze the significance of the interaction term between rank and ethnicity. This table shows what each coefficient in the above summary represent:

Table 2

Variable	Effect
ranktenure track & ranktenured	On average, tenure track professors and tenured professors received a score of 0.9100 and 0.8369, respectively, less than their teaching counterparts.
Ethnicitynot minority	On average, non-minority professors reviewed a score of 0.5828 less than their minority counterparts.
ranktenure track: ethnicitynotminority	On average, professors both on the tenure track and not a minority received a score of 0.9266 more than their counterparts that are a minority and not on the tenure track. If the professor was on a tenure track and not a minority the score would increase 0.9266 in addition to the effect of being on the tenure track (-0.9100) and not being a minority (-0.5828)
ranktenured:ethnicitynot minority	On average, professors who were both tenured and not a minority received a score of 0.7679 more than their counterparts that are a minority and not tenured. If the professor were tenured and not a minority, then their score would increase by 0.7679 in addition to the effect of being tenured (-0.8369) and not being a minority (-0.5828)

I used to backwards stepwise model selection to build my model, which is an algorithm that finds the best model for predicting the average evaluation score. Having the step function be in the backwards direction means that the model will start with all the variables and will remove variables as the algorithm sees fit. It compares the models it produces using the AIC scores from each model; the lower the score the better.

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_credits + bty_avg + pic_outfit + pic_color +
##     age:bty_avg + rank:ethnicity, data = evals2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84581 -0.32204  0.06367  0.35277  0.98705
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         5.352083   0.438231  12.213  < 2e-16
## ranktenure track                   -0.477409   0.220869  -2.161 0.031186
## ranktenured                        -0.501354   0.225032  -2.228 0.026381
## ethnicitynot minority              -0.230929   0.201798  -1.144 0.253086
## gendermale                          0.194641   0.051845   3.754 0.000197
## languagenon-english                -0.226024   0.108901  -2.075 0.038511
## age                                -0.028513   0.007161  -3.982 7.98e-05
## cls_perc_eval                       0.004604   0.001428   3.223 0.001359
## cls_creditsone credit               0.343379   0.125966   2.726 0.006663
## bty_avg                            -0.173228   0.073905  -2.344 0.019518
## pic_outfitnot formal               -0.122766   0.069199  -1.774 0.076726
## pic_colorcolor                     -0.146570   0.068772  -2.131 0.033612
## age:bty_avg                         0.004806   0.001571   3.060 0.002350
## ranktenure track:ethnicitynot minority  0.433707   0.233451   1.858 0.063853
## ranktenured:ethnicitynot minority       0.477039   0.229258   2.081 0.038020
##
## (Intercept)                        ***
## ranktenure track                   *
## ranktenured                        *
## ethnicitynot minority
## gendermale                         ***
## languagenon-english                *
## age                                ***
## cls_perc_eval                      **
## cls_creditsone credit              **
## bty_avg                            *
## pic_outfitnot formal               .
## pic_colorcolor                     *
## age:bty_avg                        **
## ranktenure track:ethnicitynot minority .
## ranktenured:ethnicitynot minority      *
```

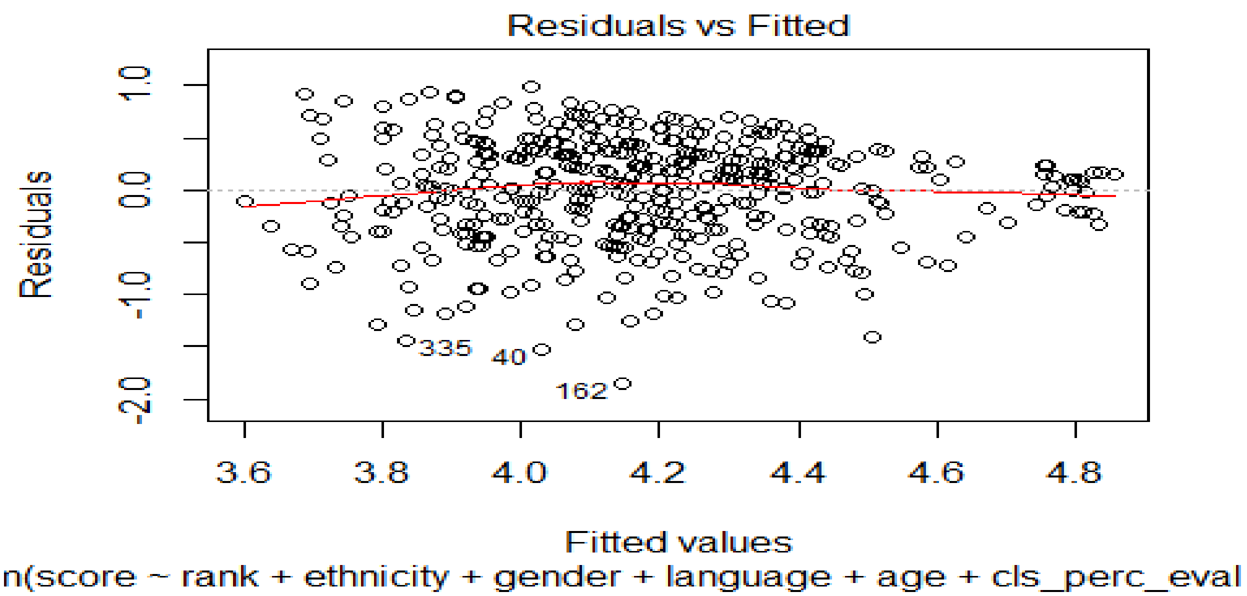
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4923 on 448 degrees of freedom
## Multiple R-squared:  0.2055, Adjusted R-squared:  0.1807
## F-statistic: 8.278 on 14 and 448 DF,  p-value: 6.695e-16
```

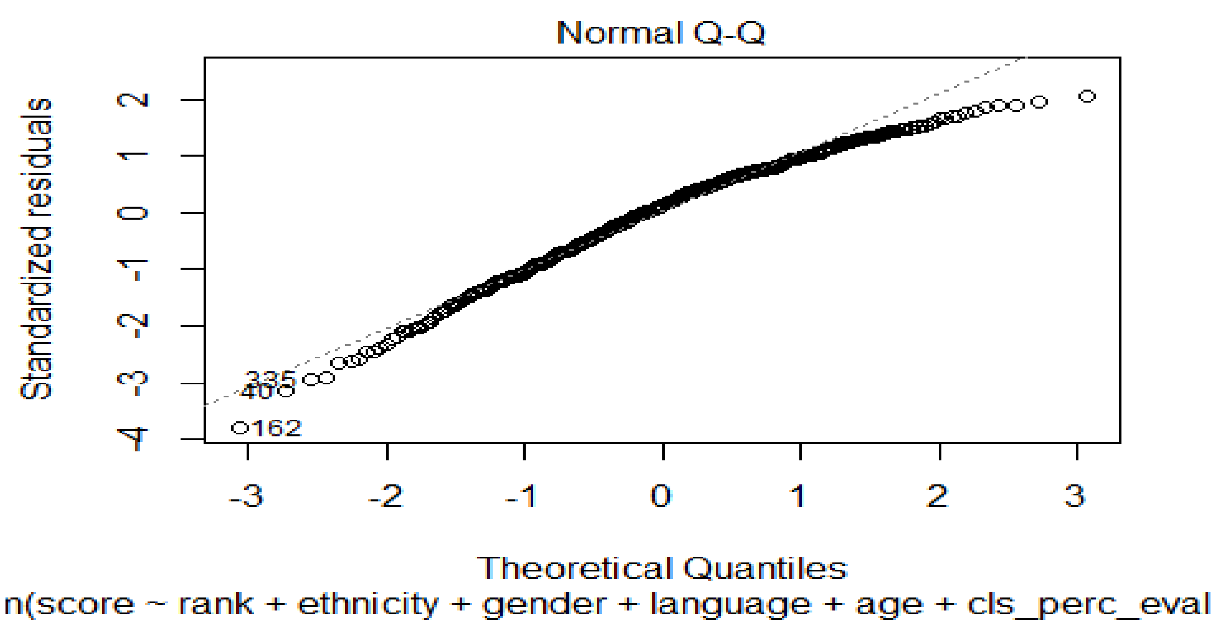
This is the model created from the stepwise algorithm. As you can see, there are 10 explanatory variables. This model has an Adjusted R-squared value of 0.1807. An Adjusted R-squared compares the explanatory power of models that contain varying amounts of explanatory variables. See more on the adjusted R-squared in the discussion section. The estimate column shows how much score would increase or decrease if the other variables are held constant. Below is an explanation of each of the coefficient:

Variable	Description
ranktenure track& ranktenured	Refer to table 2
ethnicitynot minority	Refer to table 2
gendermale	On average male professors received a score of 0.194641 more than their female counterparts.
languagenon-english	On average non-English professors received a score of 0.226024 less than their English counterparts.
age	Refer to table 1
cls_perc_eval	On average, for every 1 percent increase in the number of students who completed an evaluation for a professor’s class, their score would increase by 0.004604.
cls_creditsone credit	On average, professors who taught one-credit courses received a score of 0.343379 more than their counterparts that taught multi-credit courses.

bty_avg	Refer to table 1
pic_outfitnot formal	On average, professors with informal outfits in their pictures received a score of 0.122766 less than their counterparts that had formal outfits on in their picture.
pic_colorcolor	On average, professors with color pictures revived a score of 0.146570 less than their counterparts that had black and white pictures.
age:bty_avg	Refer to table 1
ranktenure track:ethnicitynot minority	Refer to table 2
ranktenured:ethnicitynot minority	Refer to table 2

When you make a linear model, certain assumptions have to be made. The first assumption is that the data is linearly related; you can see this from the bivariate analysis plots that the data is linearly related. The second assumption is that the variables are independent, which in this data set they are. The third assumption is that the residuals should be normally distributed. The Normal Q-Q plot shows that the data is normally distributed. The points fit along the dotted line for almost all of the line, only veering of the dotted line at the very end. The final assumption is that there is equal variance of the residuals. The Residuals vs. Fitted graph shows that the data has equal variance. The assumptions for a linear model are met.





Discussion: Fortunately, many problems that statistical studies face do not apply in this study. The data was complete and fairly cleaned. Only two variables had to be transformed. One problem I would point out is that this data includes no variables that measure a professor's teaching ability. Ideally, more variables could measure this. For example, there could be variables that measures how well a professor explains concepts, how fairly a professor grades, and how available a professor is to meet a student, to name a few. I understand that this study was conducted to see what effect attractiveness had on a professor's score, but if variables like these were present, you could tell how important such things are to a professor's score. Or in other words, it could be clear which group of variables had more weight on a professor's score: attractiveness or teaching ability. What is alarming about this study is that aspects of a professor that cannot be controlled can significantly affect a professor's score. For example, a professor on the tenure track that is also not a minority would receive a score of 0.9266 more than their minority non-tenure track counterparts. Another example is the beauty average. It is clear that the more attractive a professor is, the higher score they receive (see table 1). One final example of this alarming trend is gender. A male professor would receive a score of 0.194641 more than a female counterpart. Given that the adjusted R-squared is in the generally accepted range, it would suggest that attractiveness has a significant effect on score.

Conclusion: Given this data includes no variables that measure a professor's teaching ability, I would recommend that changes be made to how a professor is evaluated. It is unfair for a professor to be penalized for something that they cannot control, such as ethnicity, gender, or attractiveness. It is foolish to judge a professor on their score if things they cannot control affect it. Noting that, I would recommend this study be repeated, but this time include variables that measure teaching ability, such as the suggestions in the discussion section. This would allow someone to properly weigh which variables matter the most. Knowing that attractiveness affects someone's score should be noted, and less significance should be placed on a professor's score until such a time where a professor can be appropriately scored on their teaching ability.