A Beautiful Mind: Does Beauty Affect Teaching Evaluation?

Isaac Stevens, Katie Barbre, Melanie Sattler, Sam Shaud

12/10/2020

I. Executive Summary

The work in this report will explore the impact a professor's appearance has on their overall evaluation as a teacher using a public dataset titled "Impact of Beauty on Instructor's Teaching Ratings". The goal of the project is to create a model to predict a professor's teaching evaluation score based on their beauty index. For our analysis, we formulated a series of questions in order to achieve this objective:

- Is there a relationship between a professor's evaluation rating and their perceived beauty?
- What is the best model that we can create in order to predict the professor's evaluation score?
- Can we predict a professor's gender based on their beauty rating and potentially other given factors?

It is important to note that a professor's evaluation rating is a measure formed from students of the course on the instructor's ability to effectively teach their course. The beauty score was collected from the same group of six individuals that rated each professor. In addition to teaching score and beauty, the dataset included several other variables to account for class details (e.g. student number, class division) and professor characteristics (e.g. gender, age,

tenure status). The nature of our data is a combination or quantitative and categorical variables used to predict the quantitative response variable of professor evaluation.

In order to answer the three questions, we started by assessing the relationship that beauty score as a predictor has on the professor's evaluation score in a simple linear model. We found the linear model only using beauty score was not significant at predicting the evaluation score. As a result, we further examined various multiple variable model options to effectively predict evaluation of a professor. We concluded that a multiple variable linear using beauty, division, gender, tenure standing, native status and number of students was the best model for predicting the quantitative response variable of a professor's evaluation. Beauty and number of students are the only quantitative variables in the best model; whereas, division, gender, tenure status and native status are all categorical. This model showed that for an 1 point increase in the professor's beauty score, the professor's teaching evaluation would go up 0.09 provided all other variables stay the same. However, this model only explains around 20.7% of the variance in a professor's teaching rating. Even though this model is rather weak and these variables do not intrinsically describe a professor's inherent ability to teach, our group found these results compelling. Finally, our group found that age, beauty, and course division were not significant predictors of gender. We also decided to see if we could predict a professor's tenure using a subset of the predictors. The best variable to predict whether or not a professor is tenured or not was their evaluations score. For every one unit increase in evaluation score, the odds of a professor being tenured decreases by 77%. Although the model is significant, we found through testing that it's predictive ability to be poor.

In conclusion, we found that a professor's beauty does not play a significant role in predicting their teaching evaluation. This however, is reassuring knowing that a professor's evaluation wouldn't be affected by their appearance. Simply put, our results convey that students do not judge a professor's teaching ability based on their looks.

II. Detailed Description

A. Exploratory Data Analysis

For this project, the team wanted to look at a data set from the AER package in R Studio called “Impact of Beauty on Instructor's Teaching Ratings”. The data set was collected from 463 courses at the University of Texas at Austin between 2000 and 2002. It contains professors’ teaching evaluations scores, course details such as division, students taking the class, and the number of students that filled out the evaluation, and professor characteristics such as age, gender, tenure, as well as an established beauty score. The beauty score was established by an independent panel of six student judges for each professor. The score, ranging from negative two to two, aggregated across the judges and shifted to have a mean of zero. The professor’s evaluation score was ranked on a scale one (very unsatisfactory) to five (excellent) by the students in the course.

There were a total of 94 different professors evaluated. Overall, the data in itself was clean since it was collected by a University for their own research project. There were no missing data points or any conflicting data points from an initial data analysis. We did have to apply some grouping in order to better evaluate the data because some of these professors were evaluated once, while some were evaluated many times for various courses. Each row of data represented a single evaluation of each professor. On average, each professor was evaluated 4.9 times. Refer to Figure 1 for the distribution of evaluations per professor.

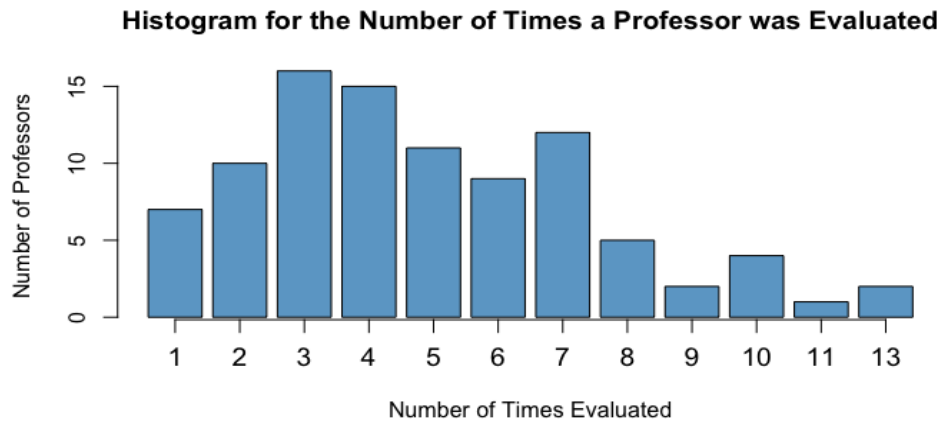

Figure 1: Histogram for the Number of Times a Professor was Evaluated

In order to account for these differences in the number of evaluations, we took the average of each professor's various evaluations. Since 'division' was recorded as either 'upper' or 'lower', we replaced 'upper' with one and 'lower' with zero and took the average. Most of the variables were the same for each professor (gender, minority, beauty evaluation, age, tenured), but the variables that required averaging were the number of students that participated in the evaluation (students), number of students enrolled in the course (allstudents), and the division. Once the data was aggregated per professor, we could proceed with our regression analysis. Refer to Figure 2 for the first ten rows of data after our aggregation by professor.

prof <fctr>	evals <dbl>	beaut <dbl>	age <dbl>	gender <fctr>	minority <fctr>	native <fctr>	tenure <fctr>	division <dbl>	students <dbl>	allstudents <dbl>
1	4.000000	0.2899157	36	female	yes	yes	yes	0.0000000	65.75000	104.00000
2	3.533333	-0.7377322	59	male	no	yes	yes	0.0000000	30.33333	34.66667
3	3.450000	-0.5719836	51	male	no	yes	yes	0.0000000	83.00000	125.00000
4	4.012500	-0.6779634	40	female	no	yes	yes	0.0000000	23.62500	27.87500
5	4.350000	1.5097940	31	female	no	yes	yes	0.0000000	43.50000	55.16667
6	4.442857	0.5885687	62	male	no	yes	yes	0.0000000	157.71429	264.14286
7	3.840000	-0.1260010	33	female	no	yes	yes	0.0000000	31.20000	38.20000
8	4.028571	-0.2581899	51	female	no	yes	yes	0.1428571	20.28571	24.85714
9	4.257143	0.1496926	33	female	no	yes	yes	0.0000000	31.14286	43.28571
10	4.560000	0.5409170	47	male	no	yes	no	0.4000000	15.40000	19.10000

Figure 2. First 10 Rows of Data After Aggregation by Professor

In order to ensure constant variance among categorical values, we created boxplots for each variable. Based on the results in Figure 3, we can conclude that on average the true mean value of evaluation score is higher for women than more men. Figure 4 proclaims that a professor whose native language is English will have a slightly higher evaluation score, on average. Regarding Figure 5, there is no significant difference in the evaluation score between minority and non-minority professors. Lastly, there is not a significant difference in the evaluation score between tenured and non-tenured professors.

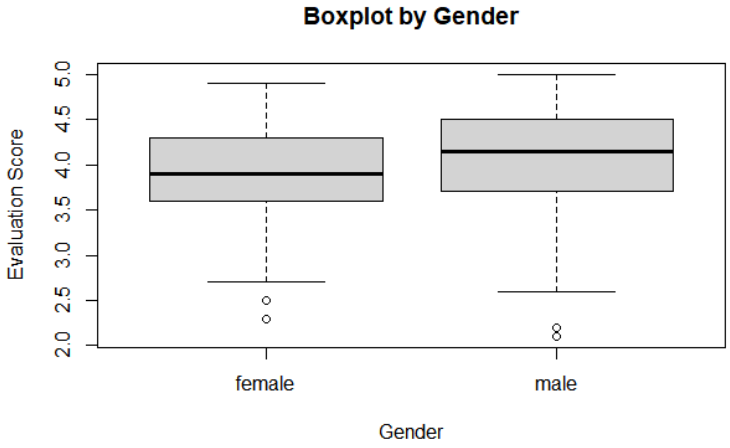

Figure 3: The Relationship Between Gender and Evaluation Score

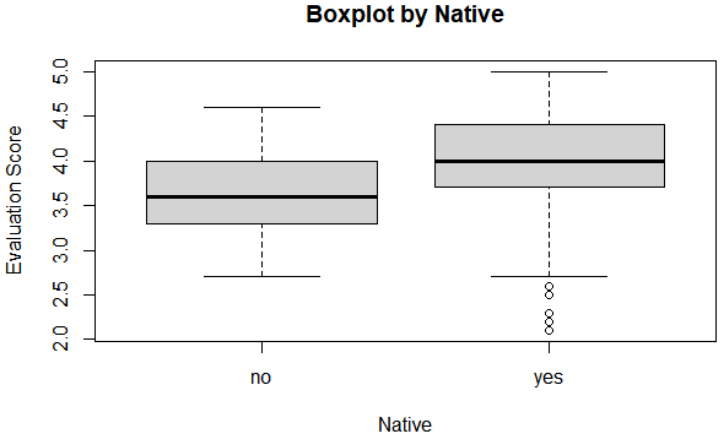

Figure 4: The Relationship Between Native Speaking Professors and Evaluation Score

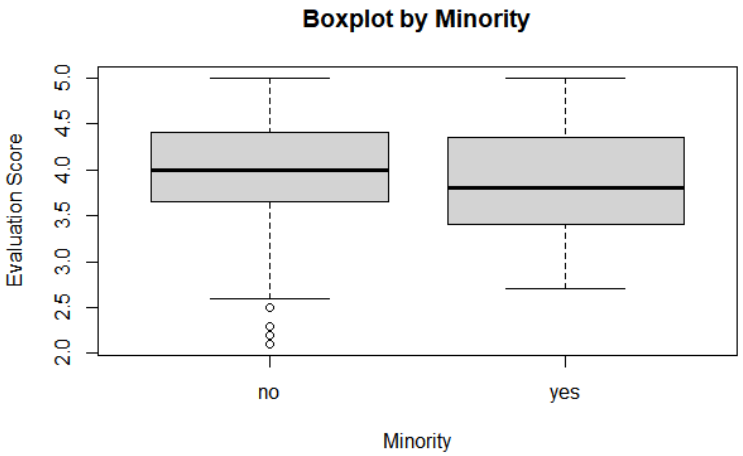

Figure 5: The Relationship Between Minority Professors and Evaluation Score

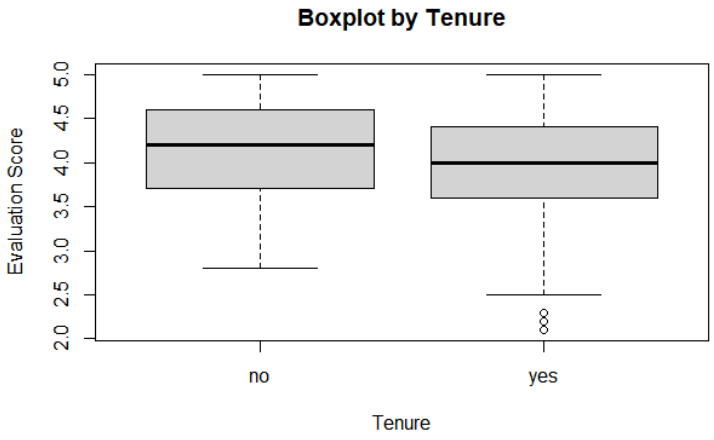

Figure 6: The Relationship Between Tenured and Non-Tenured Professors and

B. Model Building

After the initial data evaluation, we looked at the simple linear regression model between the response variable, professor's evaluation, and their beauty score as the predictor variable. This is model 1 for the project. The simple linear regression model was not a good model for predicting the professor's evaluation scorel. The multiple R^2 was 0.038 and the beauty was not considered a significant predictor at 95% confidence level as the p-value was 0.057 in the summary regression output (Figure 7).

```
lm(formula = evals ~ beaut)

Residuals:
     Min       1Q   Median       3Q      Max
-1.64092 -0.27265  0.06915  0.24221  0.94757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.89627    0.04716   82.62   <2e-16 ***
beaut        0.10940    0.05697    1.92   0.0579 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4546 on 92 degrees of freedom
Multiple R-squared:  0.03854,   Adjusted R-squared:  0.02809
F-statistic: 3.688 on 1 and 92 DF,  p-value: 0.0579
```

Figure 7. Simple Linear Regression Output for Model 1

Based on the low predictive values of the first model we built, we decided to see if adding multiple predictors improved the fit and prediction of the model. First, we looked at the forward, backward, and stepwise model selection methods to narrow down which predictors we should start with. Both the forward and stepwise model building steps came up with the same reduced model which is model 2:

$$evals = division + gender + beauty + tenure + native$$ (Figure 8 and 10)

The backward model selection built the model which is model 3:

$$evals = beaut + gender + native + tenure + students + allstudents + division$$ (Figure 9)

Model 2 had an adjusted R^2 of 0.172 and beauty was considered a significant predictor in this model(Figure 11). The model 3 had an adjusted R^2 of 0.207 and beauty was not considered a significant predictor (Figure 12).

```
               Df Sum of Sq    RSS     AIC
<none>                      15.499 -157.44
+ students      1   0.21249 15.286 -156.74
- native        1   0.46576 15.964 -156.66
+ allstudents   1   0.08930 15.409 -155.98
- division      1   0.59938 16.098 -155.87
+ minority      1   0.05753 15.441 -155.79
- tenure        1   0.61578 16.114 -155.78
+ age           1   0.00704 15.492 -155.48
- beaut         1   0.94260 16.441 -153.89
- gender        1   1.27773 16.776 -151.99

Call:
lm(formula = evals ~ division + gender + beaut + tenure + native,
    data = group.data)

Coefficients:
(Intercept)     division   gendermale        beaut    tenureyes    nativeyes
     3.8425      -0.2118       0.2412       0.1242      -0.2274       0.2728
```

Figure 8. Stepwise Model Result for Model 2

```
             Df Sum of Sq    RSS     AIC
<none>                    14.508 -159.65
- tenure       1   0.45954 14.968 -158.71
- beaut        1   0.47372 14.982 -158.62
- native       1   0.47819 14.987 -158.60
- allstudents  1   0.77768 15.286 -156.74
- division     1   0.79275 15.301 -156.64
- students     1   0.90087 15.409 -155.98
- gender       1   1.35596 15.864 -153.25

Call:
lm(formula = evals ~ beaut + gender + native + tenure + students +
    allstudents + division, data = group.data)

Coefficients:
(Intercept)        beaut   gendermale    nativeyes    tenureyes     students  allstudents
    3.77580      0.09102      0.25051      0.27778     -0.20238      0.01463     -0.00827
   division
   -0.25998
```

gure 9. Backward Model Result for Model 3

```
             Df Sum of Sq    RSS     AIC
<none>                    15.499 -157.44
+ students    1  0.212493 15.286 -156.74
+ allstudents 1  0.089300 15.409 -155.98
+ minority    1  0.057527 15.441 -155.79
+ age         1  0.007038 15.492 -155.48

Call:
lm(formula = evals ~ division + gender + beaut + tenure + native,
    data = group.data)

Coefficients:
(Intercept)     division   gendermale        beaut    tenureyes    nativeyes
     3.8425      -0.2118       0.2412       0.1242      -0.2274       0.2728
```

Figure 10. Forward Model Result for Model 2

```
lm(formula = evals ~ division + beaut + gender + tenure + native)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42678 -0.18507 -0.00726  0.29913  0.83392

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.84247    0.21909  17.538  < 2e-16 ***
division    -0.21178    0.11480  -1.845  0.06843 .
beaut        0.12424    0.05370   2.313  0.02303 *
gendermale   0.24118    0.08954   2.693  0.00847 **
tenureyes   -0.22742    0.12162  -1.870  0.06483 .
nativeyes    0.27280    0.16775   1.626  0.10748
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4197 on 88 degrees of freedom
Multiple R-squared:  0.2162,    Adjusted R-squared:  0.1717
F-statistic: 4.856 on 5 and 88 DF,  p-value: 0.0005724
```

Figure 11. Summary output for Model 2

```
lm(formula = evals ~ beaut + gender + native + tenure + students +
    allstudents + division)

Residuals:
     Min       1Q   Median       3Q      Max
-1.37542 -0.20843 -0.00167  0.29805  0.89749

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.775802   0.216497  17.440  < 2e-16 ***
beaut        0.091024   0.054319   1.676  0.09743 .
gendermale   0.250511   0.088362   2.835  0.00571 **
nativeyes    0.277776   0.164989   1.684  0.09588 .
tenureyes   -0.202376   0.122619  -1.650  0.10250
students     0.014629   0.006331   2.311  0.02323 *
allstudents -0.008270   0.003852  -2.147  0.03460 *
division    -0.259984   0.119933  -2.168  0.03294 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4107 on 86 degrees of freedom
Multiple R-squared:  0.2663,    Adjusted R-squared:  0.2066
F-statistic: 4.459 on 7 and 86 DF,  p-value: 0.0002848
```

Figure 12. Summary output for Model 3

We also explored adding interaction variables for the categorical variables in our regression. Through subsetting the dataset and building separate regressions for each subset of data, we found that some of the categorical variables showed varying regression slopes (Figure 13). However, through performing anova tests for all four categorical variables (gender, tenure, minority, native), the results showed that the reduced model (without interaction variables) outperformed the model with interaction variables.

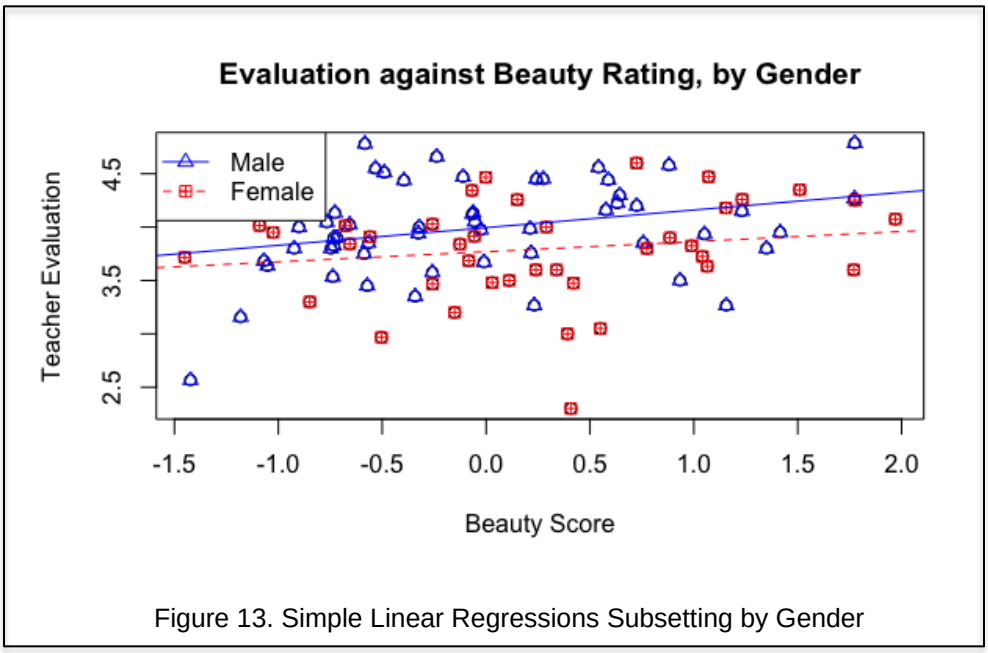

Figure 13. Simple Linear Regressions Subsetting by Gender

The last model building steps performed to try and improve our model was to look at all the combinations of the possible models, and select the best model based on adjusted R^2. The best model based on the adjusted R^2 was the model evals ~division + beaut + gender + tenure + native+students+allstudents which was the same model as model 3. This model had an adjusted R^2 of 0.2066, and beauty was not considered a significant predictor in the model (Figure 12).

We also checked the Predicted R^2 for each of the two models to see how well the models would hold up with new data. The model 3 had a predicted R^2 of 0.147714 which out of the two models is the highest predicted R^2 (Figure 14). The model 2 had a predicted R^2 of 0.1194166 (Figure 14). Overall, none of the models were very good at predicting new data based on the predicted R^2 output which aligns with the other indicators that these models are not very strong.

```
> #Model 2
> print(r2.pred3)
[1] 0.1194166
> #Model 3
> print(r2.pred2)
[1] 0.147714
>
```

Figure 14. Predictive R^2 output

For both models, we verified all the regression assumptions were met for the model 2 (Figure 16), and model 3(Figure 15). Each of the models met the regression assumptions although not all of the QQ plots looked perfectly normally distributed they were still in line. We also checked for multicollinearity by looking at the Variance Inflation factors(VIF), and it was found that the variables students and allstudents have a VIF higher than 10 indicating there may be some multicollinearity in this model.

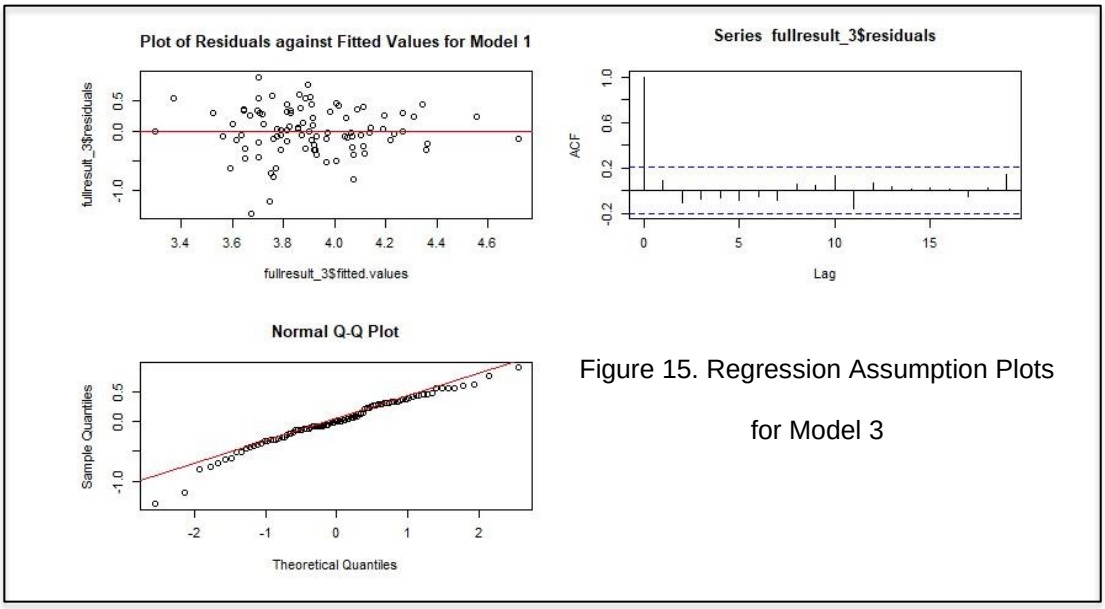

Figure 15. Regression Assumption Plots for Model 3

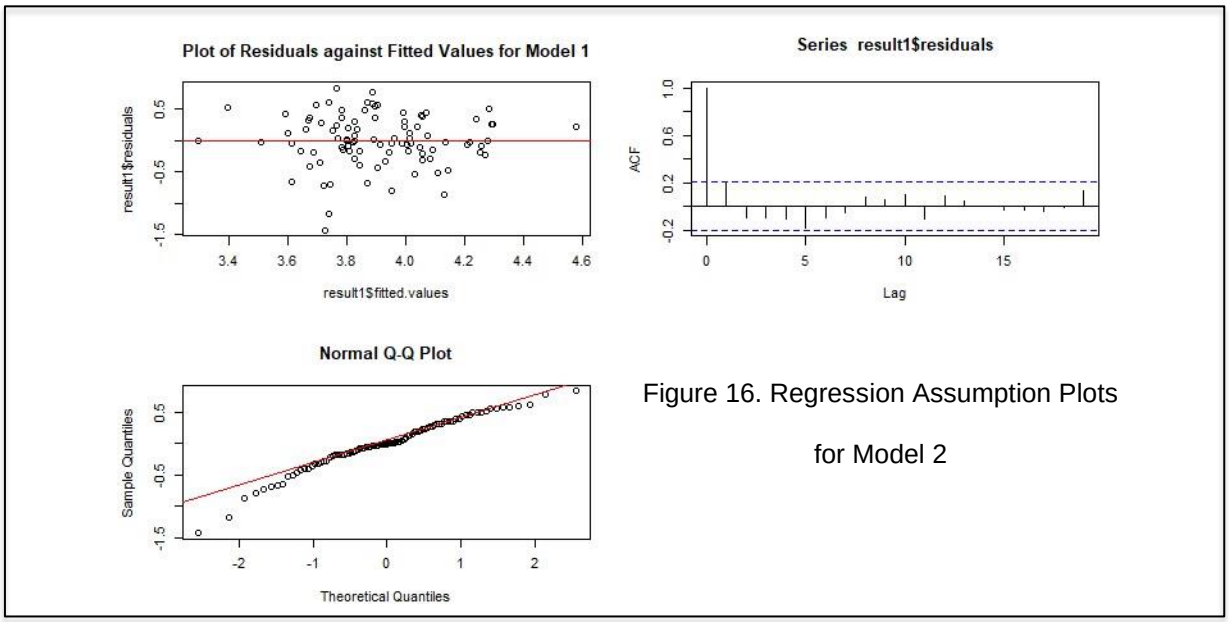

Figure 16. Regression Assumption Plots for Model 2

We also checked for potential outliers in the response variable by looking at the studentized residuals (Figure 17), and for outliers in the predictor variables by looking at DFFITS and DFBETAS for both Models (Figure 18). For model 2 there was one significant point in the externally studentized residuals, one data point in DFFITS, and two points in DFBETAS. For the model with the highest adjusted R^2/Backward built model there was one data point in the externally studentized residuals, one point in DFFITS, and four points in DFFBETAS. All in all, a small number of outliers for our dataset with a small impact on the overall analysis.

```
> ext.student.res[abs(ext.student.res)>qt(1-0.05/(2*n), n-p-1)]
       99
-3.664652
```

Figure 17. Externally Studentized Residuals Model 2

```
> DFFITS[abs(DFFITS)>2*sqrt(p/n)]
      198
-0.6665884
> DFBETAS[abs(DFBETAS)>2/sqrt(n)]
 [1] -0.2583829 -0.2573611  0.2344188  0.2344188  0.3326353  0.2447695 -0.6244196 -0.2579617 -0.2468507  0.2623574 -0.2184062
[12]  0.2114029  0.3070230  0.2722512 -0.2784668 -0.2784668 -0.3106597 -0.2285988  0.2387452 -0.2338432 -0.2312932  0.2145847
```

Figure 18. DFFITS and DFBETAS output for Model 2

```
> ext.student.res3[abs(ext.student.res3)>qt(1-0.05/(2*n2), n2-p2-1)]
       30
-3.653744
```

Figure 19. Externally Studentized Residuals Model 3

```
> DFFITS2[abs(DFFITS2)>2*sqrt(p2/n2)]
        30         68         73         92
-0.7398532 -0.7942133 -0.9488763  0.6335295
> DFBETAS2[abs(DFBETAS2)>2/sqrt(n2)]
 [1] -0.3022050  0.2412419 -0.2664737  0.2167291  0.4555162 -0.3318262  0.5504232  0.4174768 -0.2469540 -0.2387849  0.2147212
[12]  0.2278249  0.2290670  0.4488915 -0.3486709  0.3225011 -0.5592378 -0.2112486  0.2175864  0.2099154  0.2924569 -0.2183728
[23] -0.2818939 -0.2581285 -0.3193395 -0.3034961  0.3423304  0.3277502  0.2845132  0.2669412  0.3330060 -0.3605994 -0.3241271
[34]  0.2285633 -0.2443919 -0.2472850  0.2303521 -0.3201752
```

Figure 20. DFFITS and DFBETAS output for Model 3

As much as the team tried to improve the model by trying different model building techniques, all the models were relatively poor predictors of the evaluation scores. The beauty predictor was not considered significant in most of the models that were built.

To answer our final question, we needed to build a logistic regression model with gender as the response variable and age, evaluation, and division as the predictor variables. The summary statistics for the model shows p-values that are not significant. In order to test that all were zero, we ran a deviance test to see if the intercept model was a better predictor than the full model. Since the p-value (Figure 21) for the deviance test was greater than our alpha of 0.05, we concluded that the coefficients were zero and the model was not a good fit.

```
```{r}
1-pchisq(log.result$null.deviance-log.result$deviance, 3)
```

[1] 0.1077617
```

Figure 21. Deviance test results for age, evaluation, and division.

Since one of our reasons for asking the question was a desire to work with logistic regression modeling techniques, we decided to pursue a new question. Are there certain qualities of a professor that can predict their probability of being tenured? Since we were unsure which predictors to start with, we used the forward, backward, and stepwise model selection methods to narrow down our choices. All three methods came up with the same predictors: evaluation scores and age. After fitting the regression model and reviewing the summary, it

appeared that age was not a significant predictor. We ran a deviance test to see if age could be dropped and the p-value shown in Figure 22 supports this conclusion. After refitting the model and reviewing the new p-value using a delta G-squared test, the model is significant with 95% confidence. Thus, we can conclude that the odds of a professor having tenure status are multiplied by 0.23 for every one unit increase in evaluation score.

```r
reduced <- glm(tenure~evals, family='binomial', data=group.data)
1-pchisq(reduced$deviance-bestlog.result$deviance,1)
```

```
[1] 0.4344421
```

Figure 22. Deviance test results for dropping age from the model.

```
Call:
glm(formula = tenure ~ evals, family = "binomial", data = group.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2368   0.3583   0.5198   0.6439   0.8965

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.4655     3.0580   2.441   0.0146 *
evals        -1.4510     0.7458  -1.946   0.0517 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.525  on 93  degrees of freedom
Residual deviance: 78.169  on 92  degrees of freedom
AIC: 82.169

Number of Fisher Scoring iterations: 5
```

Figure 23. Summary output showing the coefficients of the logistic regression.

We created an ROC curve to see if the model was still able to predict tenure better than random guessing. Based on the ROC curve in Figure 24, the ROC curve seems decent but not great. In order to check how close it is to random guessing, we calculated an AUC of 0.76. Since an AUC of 0.5 is random guessing and an AUC of 1 is perfect guessing, the model is somewhere between random and perfect. Lastly, we used a confusion matrix to summarize the number of true/false positives and true/false negatives. Based on the table in Figure 25, it appears that although the model was statistically significant, it has trouble differentiating between tenured and non-tenured professors based solely on evaluation score.

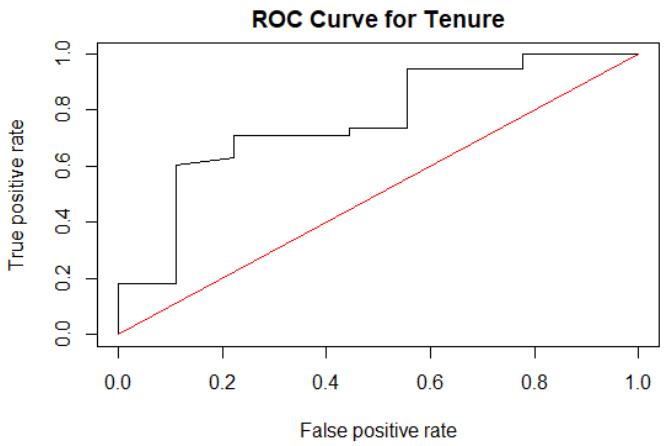

Figure 24. The ROC curve for the logistic regression model of tenure against evaluation score.

```
```{r, echo=FALSE}
table(test$tenure, preds>0.5)
table(test$tenure, preds>0.9)
```
```

```
     TRUE
no      9
yes    38

     FALSE
no       9
yes     38
```

Figure 25. The confusion matrix for the tenure logistic regression model at thresholds of 0.5 and 0.9.

With a regression model with the highest possible $R^{2,}$ we also wrote a lasso and ridge regression model to reduce possible variance in our model, and see if we could improve its predicting power. While our ridge and lasso regressions produced a lower Mean Squared Error (MSE) than the OLS regression, the results also showed contrary results depending on the train test split of our validation data. Due to our aggregated data set only having 94 observations, the high variability among the results of our train test split make sense, as one observation changing from the train to test set could shift the results in a larger way. For most train test divisions, the ridge regression provided the lowest MSE below lasso and OLS. Therefore, ridge regression would be the best model to use for predicting new professor's teaching evaluations.

III. Conclusion

Overall, the team worked through multiple different model building techniques to try to find the best model, but even with the best model built through the different techniques it was still not a good model with both the R^2 and adjusted R^2 being below 0.3. In context this means that the model explains less than 30% of the variation in the Evaluation score. That being said, it is important to understand that beauty is ambiguous and based on an individual perception and not a set data point. Additionally, there are many other factors that could go into the evaluation score that students give a professor that may not have been included in this particular data set. It is also important to note a student's evaluation of their professor could be skewed based on the grade they received in the class. In any case, it is reassuring that a professor's appearance does not have a strong impact on how their student evaluates their ability to teach. A way to further expand this project for the next time would be to include the grades students received, and perhaps the type of class such as science, history, or math. It would also be useful to know the student gender distribution for the evaluators. These might be additional predictors that could give deeper insight in addition to the current predictors in this data set. Also, this report would benefit from a larger number of professors evaluated and potentially extending the evaluation to professors in other Universities across the country. Some of the models created were significant at the 90% confidence interval; therefore, it is possible that more data could elevate the significance past the 95% confidence threshold. It was disappointing to see that the most significant predictor of a professor being tenured was their evaluation score and that they were inversely related. One would expect a tenured professor to be a better teacher based on this accomplishment, but the model shows that non-tenured professors have higher evaluation scores. Ultimately, the data provided was not effective at predicting a professors' evaluation score. However, a non-story is still a story. From this analysis, we learn that the characteristics of a professor, including beauty, only influence a small portion of the instructor's teaching rating.