Exercise_5

Mansoor Baba Shaik

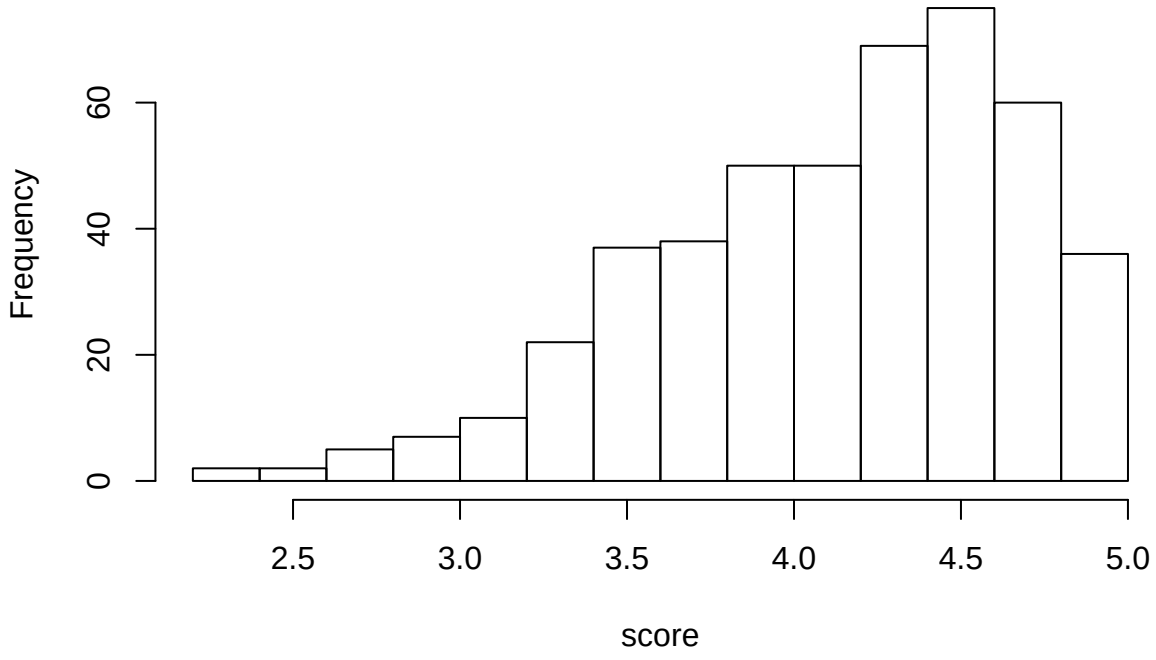
November 23, 2016

```
rm(list=ls())

setwd("~/IS665/Exercise_5/")
evals = read.csv(file="evals.csv", header=TRUE, sep=",", dec=".", quote="\"")
```

1

```
hist(evals$score, xlab="score", main="Histogram of score")
```




```
skewness <- function(x){
  return(3*(mean(x)-median(x))/sd(x))
}
skewness(evals$score)
```

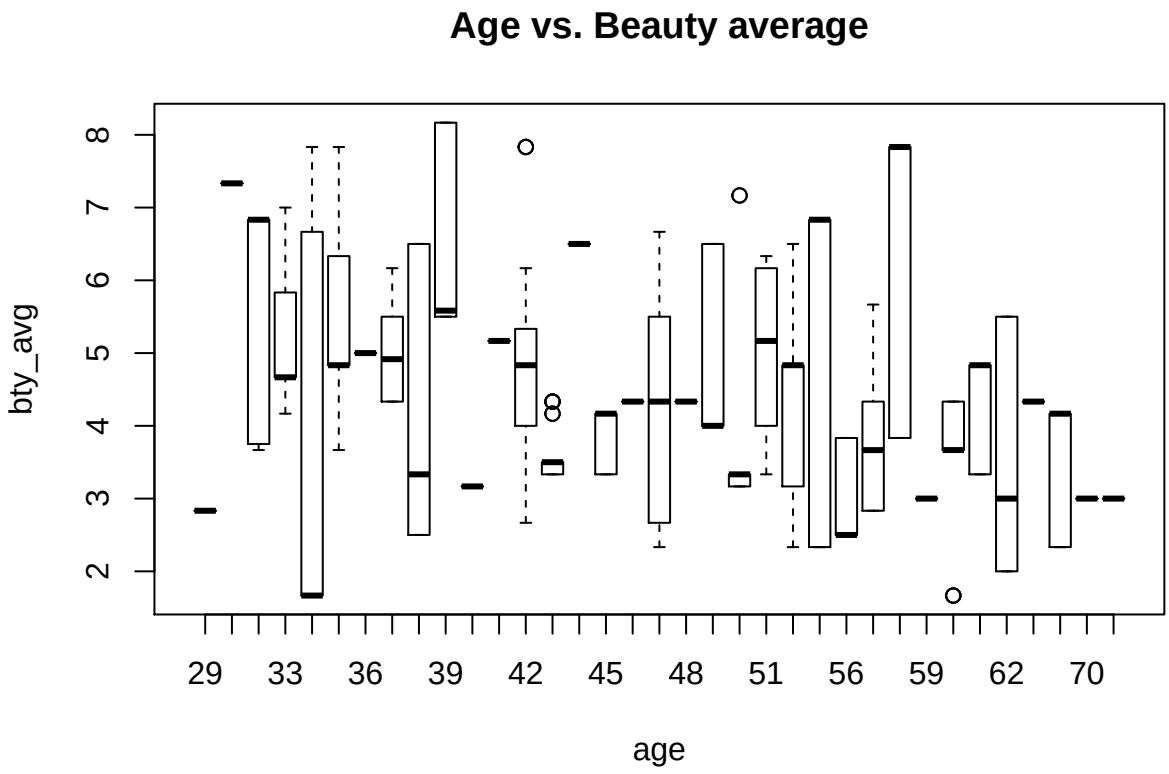
```
## [1] -0.6909992
```

Yes, the distribution of score is skewed and the skewness is towards to the left. This tells us that the students rated courses with more higher scores. This is not what I have expected. I have expected a normal distribution

where most professors would be rated near to the average and fewer professors will be rated in the extremes with score 1.0 or 5.0.

2

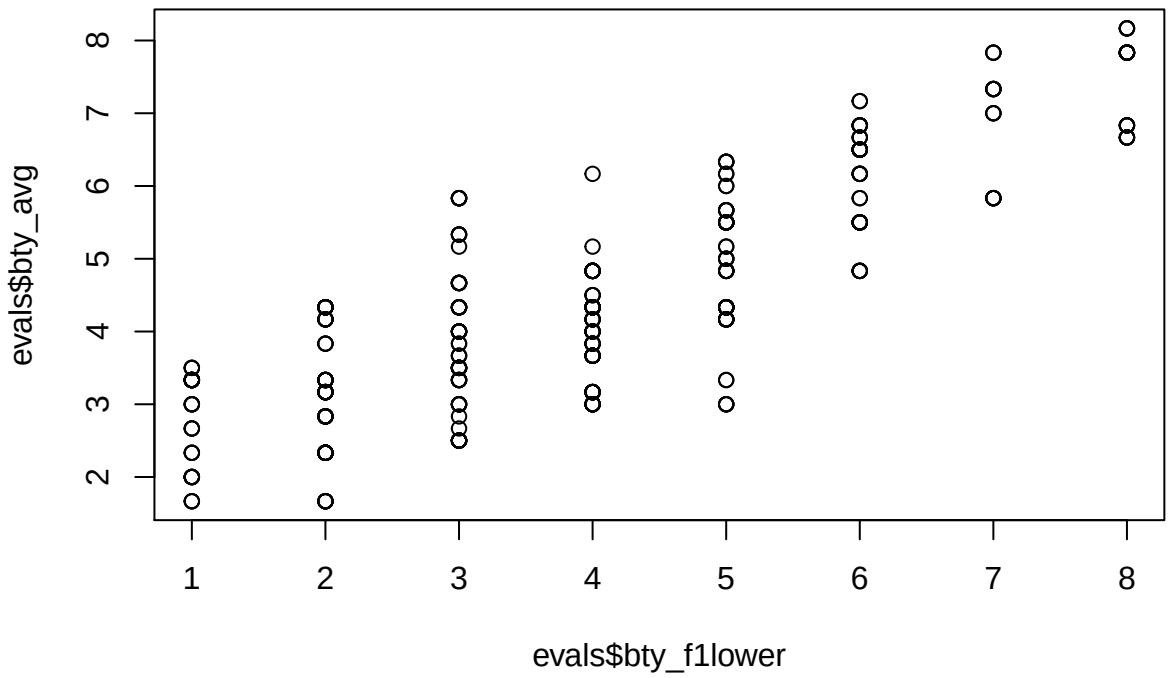
```
boxplot(bty_avg~age, data=evals,xlab="age", ylab="bty_avg", main="Age vs. Beauty average")
```



In the side-by-side boxplot between age and bty_avg, we can see that there is no relationship between the professor's age and their average beauty score. I have expected that the younger professors would have higher average beauty scores. For example, there are some cases like a professor of age 42 has the average beauty score near to 8 which the highest.

Multiple Linear Regression

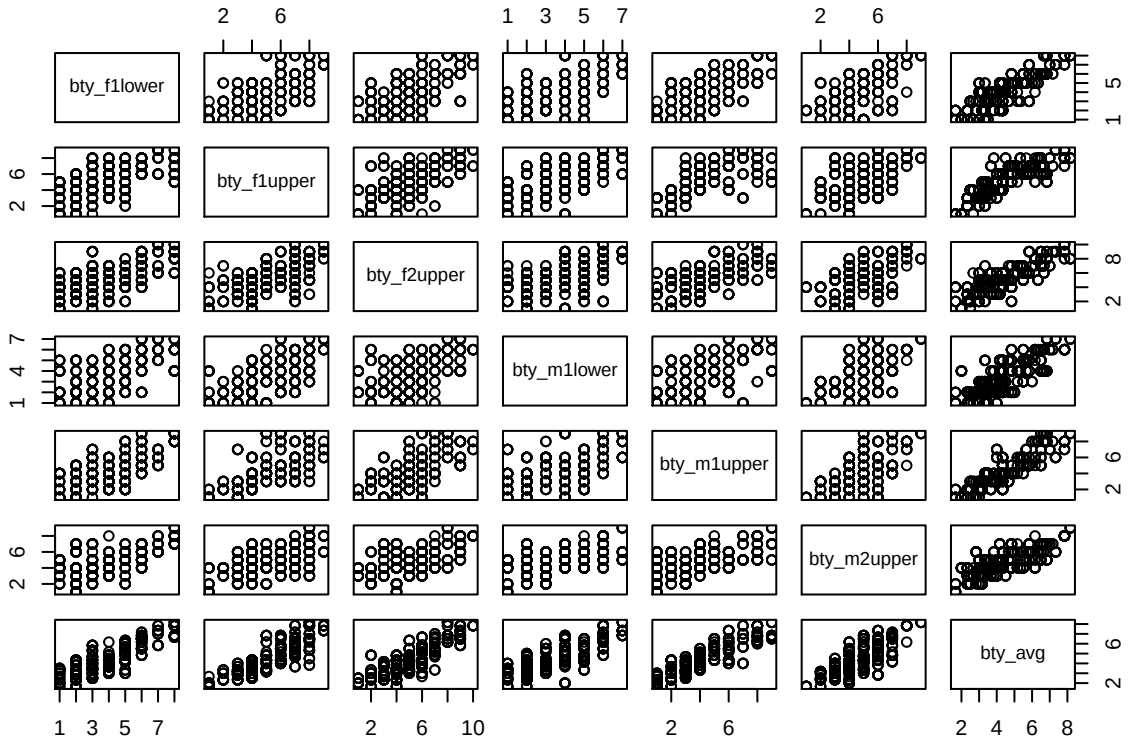
```
plot (evals $ bty_avg ~ evals $ bty_f1lower)
```




```
cor (evals $ bty_avg , evals $ bty_f1lower)
```

```
## [1] 0.8439112
```

```
plot ( evals [, 13 : 19 ])
```




```
m_bty_gen <- lm(score ~ bty_avg + gender , data = evals)
summary(m_bty_gen)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## bty_avg      0.07416    0.01625   4.563 6.48e-06 ***
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

3

```
m_bty <- lm(score ~ bty_avg, data = evals)
summary(m_bty)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88034    0.07614   50.96  < 2e-16 ***
## bty_avg      0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
```

```
summary(m_bty_gen)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## bty_avg      0.07416    0.01625   4.563 6.48e-06 ***
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

Yes, bty_avg is still a significant predictor of score. There is not much difference in the paramter estimate for bty_avg on the addition of gender to the model. The addition of gender to the model made the model even better as we can see that the adjusted R-squared values increased when gender has been added which might result in better predictions.

4

For males: $\hat{scôre} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{bty\_avg} + \hat{\beta}_2 \times (1)$

For females: $\hat{scôre} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{bty\_avg} + \hat{\beta}_2 \times (0)$

For two professors who received the same beauty rating, males tend to have a higher course evaluation score than females.

5

```
m_bty_rank <- lm(score ~ bty_avg+rank, data = evals)
summary(m_bty_rank)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + rank, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8713 -0.3642  0.1489  0.4103  0.9525
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.98155    0.09078  43.860  < 2e-16 ***
## bty_avg           0.06783    0.01655   4.098 4.92e-05 ***
## ranktenure track -0.16070    0.07395  -2.173   0.0303 *
## ranktenured      -0.12623    0.06266  -2.014   0.0445 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 459 degrees of freedom
## Multiple R-squared:  0.04652,    Adjusted R-squared:  0.04029
## F-statistic: 7.465 on 3 and 459 DF,  p-value: 6.88e-05
```

From the summary, we can see that only two parameter estimates are present for rank though it has three levels. The parameter estimate rankteaching seems to be not considered by the linear model.

6

```
#The below statement is the question given by you in this exercise.
#I have not understood what is the question here, as it appears to be a statement.
```

The interpretation of the coefficients in multiple regression is slightly different from that of simple regression. The estimate for bty_avg reflects how much higher a group of professors is expected to score if they have a beauty rating that is one point higher w hile holding all other variables constant. In this case, that translates into considering only professors of the same rank with bty_avg scores that are one point apart.

7

```
m_full <- lm(score ~ rank + ethnicity + gender + language + age
             + cls_perc_eval + cls_students + cls_level + cls_profs
             + cls_credits + bty_avg + pic_outfit + pic_color , data = evals)

summary(m_full)
```

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77397 -0.32432  0.09067  0.35183  0.95036
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.0952141  0.2905277  14.096  < 2e-16 ***
## ranktenure track        -0.1475932  0.0820671  -1.798  0.07278 .
## ranktenured             -0.0973378  0.0663296  -1.467  0.14295
## ethnicitynot minority    0.1234929  0.0786273   1.571  0.11698
## gendermale               0.2109481  0.0518230   4.071 5.54e-05 ***
## languagenon-english     -0.2298112  0.1113754  -2.063  0.03965 *
## age                     -0.0090072  0.0031359  -2.872  0.00427 **
## cls_perc_eval            0.0053272  0.0015393   3.461  0.00059 ***
## cls_students             0.0004546  0.0003774   1.205  0.22896
## cls_levelupper           0.0605140  0.0575617   1.051  0.29369
## cls_profssingle         -0.0146619  0.0519885  -0.282  0.77806
## cls_creditsone credit    0.5020432  0.1159388   4.330 1.84e-05 ***
## bty_avg                  0.0400333  0.0175064   2.287  0.02267 *
## pic_outfitnot formal    -0.1126817  0.0738800  -1.525  0.12792
## pic_colorcolor          -0.2172630  0.0715021  -3.039  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.498 on 448 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1617
## F-statistic: 7.366 on 14 and 448 DF,  p-value: 6.552e-14
```

```
levels(evals$cls_profs)
```

```
## [1] "multiple" "single"
```

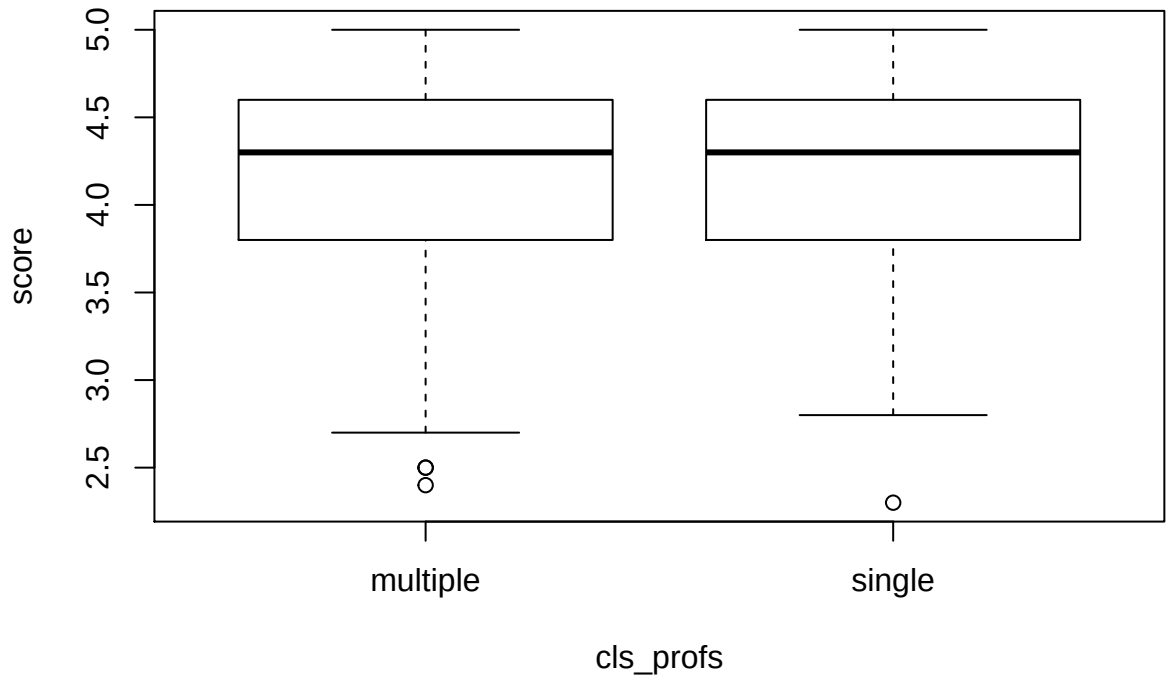
The parameter estimate(coefficient estimate) of cls_profs, shown here as cls_profssingle for one of its level single(considered by model) is the least value compared to other parameter estimates. I would expect cls_profs to not have any association with the professor score.

8

```
which(summary(m_full)$coefficients[, 4] == max(summary(m_full)$coefficients[, 4]))
```

```
## cls_profssingle
##              11
```

```
plot(score ~ cls_profs, data = evals)
```




```
summary(lm(score ~ cls_profs, data=evals))$adj.r.squared
```

```
## [1] -0.0015192
```

Yes, cls_profs has the least association to scores. Also, when seen in the summary of the model you can see that it has very low adjusted R-squared and high p-values which makes it the least preferred.

9

```
summary(m_full)
```

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77397 -0.32432  0.09067  0.35183  0.95036
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.0952141  0.2905277  14.096  < 2e-16 ***
## ranktenure track       -0.1475932  0.0820671  -1.798  0.07278 .
## ranktenured            -0.0973378  0.0663296  -1.467  0.14295
## ethnicitynot minority   0.1234929  0.0786273   1.571  0.11698
## gendermale              0.2109481  0.0518230   4.071 5.54e-05 ***
## languagenon-english    -0.2298112  0.1113754  -2.063  0.03965 *
## age                    -0.0090072  0.0031359  -2.872  0.00427 **
## cls_perc_eval           0.0053272  0.0015393   3.461  0.00059 ***
## cls_students            0.0004546  0.0003774   1.205  0.22896
## cls_levelupper          0.0605140  0.0575617   1.051  0.29369
## cls_profssingle        -0.0146619  0.0519885  -0.282  0.77806
## cls_creditsone credit   0.5020432  0.1159388   4.330 1.84e-05 ***
## bty_avg                 0.0400333  0.0175064   2.287  0.02267 *
## pic_outfitnot formal   -0.1126817  0.0738800  -1.525  0.12792
## pic_colorcolor         -0.2172630  0.0715021  -3.039  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.498 on 448 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1617
## F-statistic: 7.366 on 14 and 448 DF,  p-value: 6.552e-14
```

The coefficient associated with the ethnicity variable is 0.1234929213

10

```
m_final <- lm(score ~ rank + ethnicity + gender + language + age +
                cls_perc_eval + cls_students + cls_level + cls_credits +
                bty_avg + pic_outfit + pic_color, data = evals)

summary(m_final)
```

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7836 -0.3257  0.0859  0.3513  0.9551
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.0872523  0.2888562  14.150  < 2e-16 ***
## ranktenure track       -0.1476746  0.0819824  -1.801 0.072327 .
```

```
## ranktenured           -0.0973829  0.0662614  -1.470 0.142349
## ethnicitynot minority  0.1274458  0.0772887   1.649 0.099856 .
## gendermale             0.2101231  0.0516873   4.065 5.66e-05 ***
## languagenon-english   -0.2282894  0.1111305  -2.054 0.040530 *
## age                   -0.0089992  0.0031326  -2.873 0.004262 **
## cls_perc_eval          0.0052888  0.0015317   3.453 0.000607 ***
## cls_students           0.0004687  0.0003737   1.254 0.210384
## cls_levelupper         0.0606374  0.0575010   1.055 0.292200
## cls_creditsone credit  0.5061196  0.1149163   4.404 1.33e-05 ***
## bty_avg                0.0398629  0.0174780   2.281 0.023032 *
## pic_outfitnot formal  -0.1083227  0.0721711  -1.501 0.134080
## pic_colorcolor        -0.2190527  0.0711469  -3.079 0.002205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4974 on 449 degrees of freedom
## Multiple R-squared:  0.187,  Adjusted R-squared:  0.1634
## F-statistic: 7.943 on 13 and 449 DF,  p-value: 2.336e-14
```

Yes, the coefficients and significance of the other explanatory variables slightly changed when cls_profs variable is not considered in the model. Therefore, this is a better model than the previous.

11

```
m_final_backward <- lm(score ~ ethnicity + gender + language + age + cls_perc_eval +
                 cls_credits + bty_avg + pic_color, data = evals)

summary(m_final_backward)
```

```
##
## Call:
## lm(formula = score ~ ethnicity + gender + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85320 -0.32394  0.09984  0.37930  0.93610
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.771922   0.232053  16.255  < 2e-16 ***
## ethnicitynot minority  0.167872   0.075275   2.230  0.02623 *
## gendermale             0.207112   0.050135   4.131 4.30e-05 ***
## languagenon-english   -0.206178   0.103639  -1.989  0.04726 *
## age                   -0.006046   0.002612  -2.315  0.02108 *
## cls_perc_eval          0.004656   0.001435   3.244  0.00127 **
## cls_creditsone credit  0.505306   0.104119   4.853 1.67e-06 ***
## bty_avg                0.051069   0.016934   3.016  0.00271 **
## pic_colorcolor        -0.190579   0.067351  -2.830  0.00487 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4992 on 454 degrees of freedom
```

```
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1576
## F-statistic:  11.8 on 8 and 454 DF,  p-value: 2.58e-15
```

$$\hat{score} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{ethinicity\_not\_minority} + \hat{\beta}_2 \times \text{gender\_male} + \hat{\beta}_3 \times \text{language\_non-english} + \hat{\beta}_4 \times \text{age} + \hat{\beta}_5 \times \text{cls\_perc\_eval} + \hat{\beta}_6 \times \text{cls\_credits\_one\_credit} + \hat{\beta}_7 \times \text{bty\_avg} + \hat{\beta}_8 \times \text{pic\_color\_color}$$

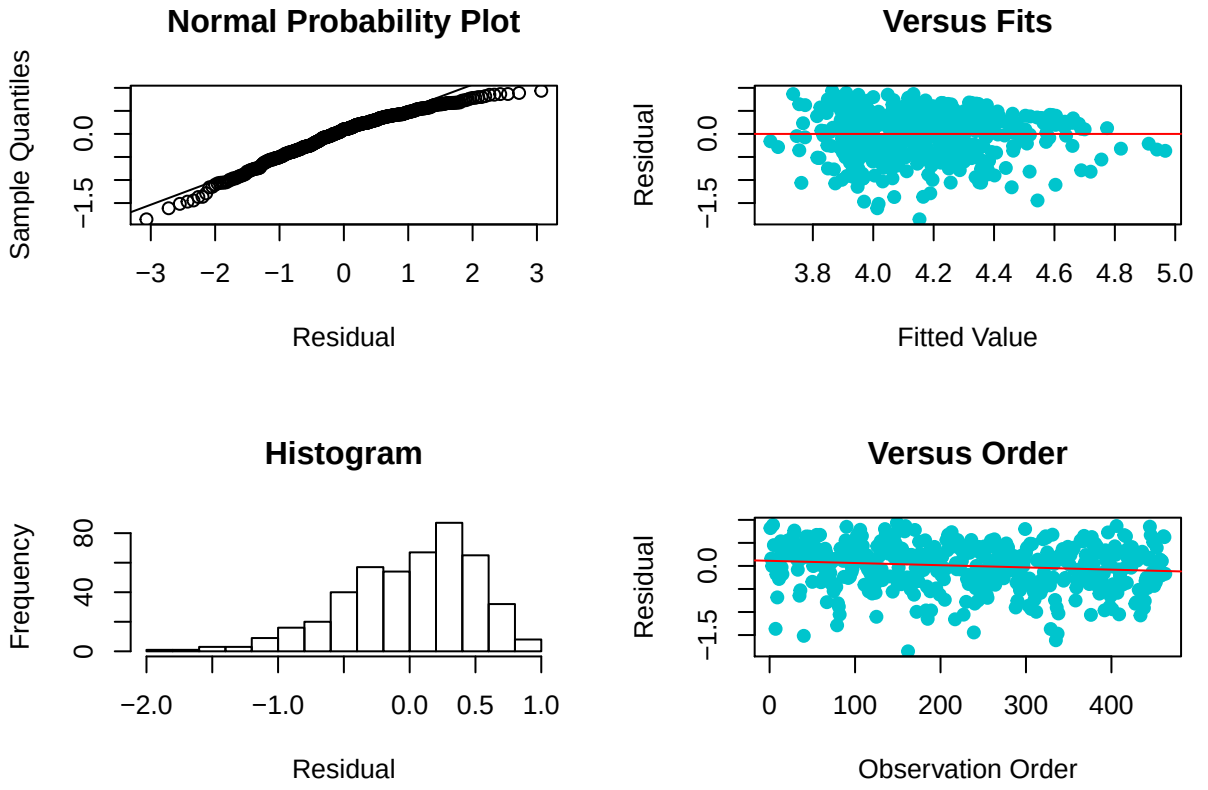
12

```
par(mfrow=c(2,2))
qqnorm(m_final_backward$residuals, main="Normal Probability Plot",
       xlab="Residual")
qqline(m_final_backward$residuals)

plot(m_final_backward$residuals ~ m_final_backward$fitted.values, pch=19,
     main="Versus Fits", xlab="Fitted Value", ylab="Residual", col="turquoise3")
abline(lm(m_final_backward$residuals ~ m_final_backward$fitted.values), col="red")

hist(m_final_backward$residuals, main="Histogram", xlab="Residual")

plot(m_final_backward$residuals ~ c(1:nrow(evals)), pch=19,
     main="Versus Order", xlab="Observation Order", ylab="Residual", col="turquoise3")
abline(lm(m_final_backward$residuals ~ c(1:nrow(evals))), col="red")
```



In the Normal Probability plot, the residuals of the model is not normal as residual values for the higher and lower quantiles are less than what a normal distribution would predict.

In the Versus Fits plot, we can see that there are some outliers, most of the residual values are close to the fitted values.

In the Histogram, we can see that it looks like some observations in the dataset are skewing normality a bit. If the extreme outliers are excluded then we will get a normal distribution.

In Versus Order plot,we can see that the observations were randomly gathered.