Lab 9 Pricilla Nakyazze Multiple Linear Regression

2025-04-04

```r
install.packages("openintro")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("Ggally")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)

## Warning: package 'Ggally' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages

## Warning: Perhaps you meant 'GGally' ?
```

```r
install.packages("GGally")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```r
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
```

```
##   method from
##   +.gg   ggplot2
##
## Attaching package: 'GGally'
##
## The following object is masked from 'package:openintro':
##
##     tips
```

```r
glimpse(evals)
```

```
## Rows: 463
## Columns: 23
## $ course_id     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ prof_id       <int> 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5,~
## $ score         <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4~
## $ rank          <fct> tenure track, tenure track, tenure track, tenure track, ~
## $ ethnicity     <fct> minority, minority, minority, minority, not minority, no~
## $ gender        <fct> female, female, female, female, male, male, male, male, ~
## $ language      <fct> english, english, english, english, english, english, en~
## $ age           <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, ~
## $ cls_perc_eval <dbl> 55.81395, 68.80000, 60.80000, 62.60163, 85.00000, 87.500~
## $ cls_did_eval  <int> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24, 17, 14,~
## $ cls_students  <int> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, 25, 20, ~
## $ cls_level     <fct> upper, upper, upper, upper, upper, upper, upper, upper, ~
## $ cls_profs     <fct> single, single, single, single, multiple, multiple, mult~
## $ cls_credits   <fct> multi credit, multi credit, multi credit, multi credit, ~
## $ bty_f1lower   <int> 5, 5, 5, 5, 4, 4, 4, 5, 5, 2, 2, 2, 2, 2, 2, 2, 2, 7, 7,~
## $ bty_f1upper   <int> 7, 7, 7, 7, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 9, 9,~
## $ bty_f2upper   <int> 6, 6, 6, 6, 2, 2, 2, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 9, 9,~
## $ bty_m1lower   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 7, 7,~
## $ bty_m1upper   <int> 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 6, 6,~
## $ bty_m2upper   <int> 6, 6, 6, 6, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 6, 6,~
## $ bty_avg       <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, ~
## $ pic_outfit    <fct> not formal, not formal, not formal, not formal, not form~
## $ pic_color     <fct> color, color, color, color, color, color, color, color, ~
```

```r
?evals
```

Exercise 1

Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

I believe the study is more observational than experimental this is because most of the variables are based on personal preferences and differ form one person to another based on personal beauty standards of their professors' physical appearance. It is then somewhat experimental because now am keen to know if a student is highly likely to do well in a course when the Professor‘s beauty rating according to them is high or low.

Exercise 2

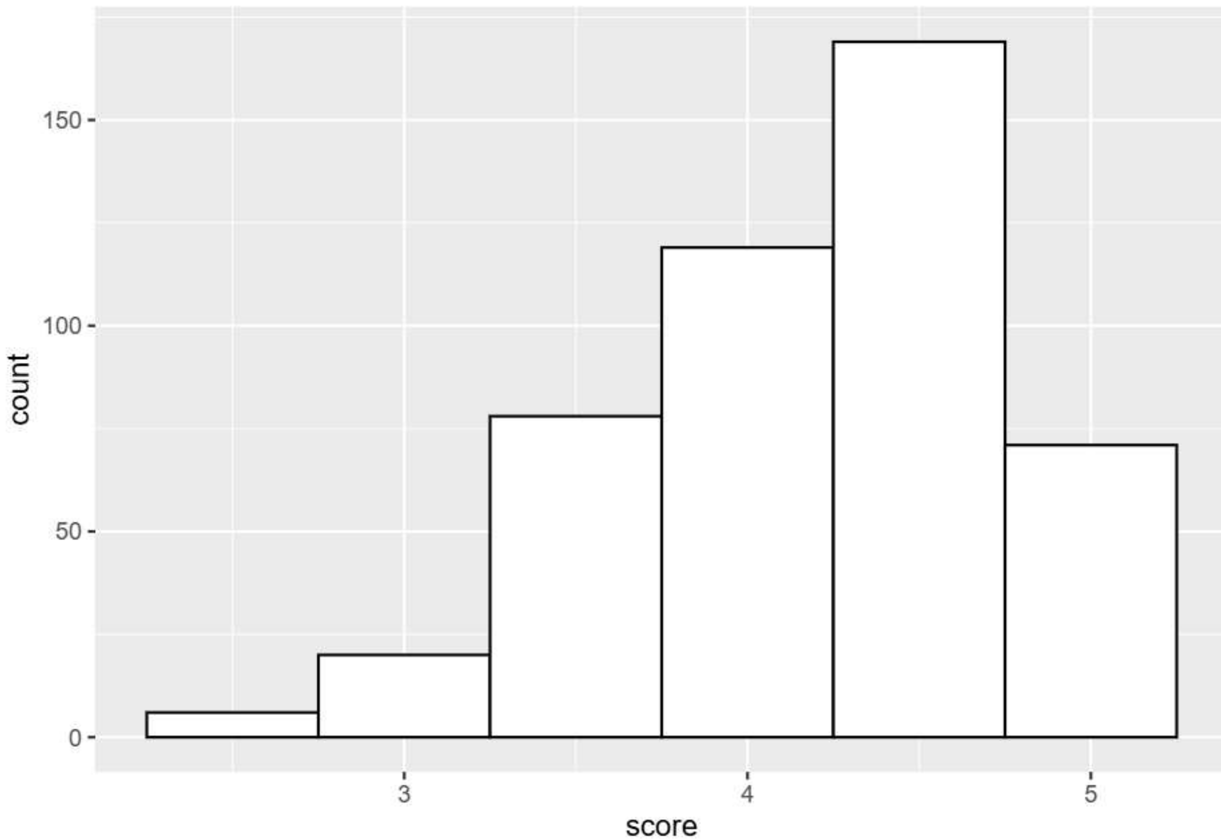
Describe the distribution of score. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

The distribution of score is left skewed. The data clusters on the right side with a tail on the left side.This indicates that lower values are more extreme, and the mean score of 4.17 is lower than the median at 4.300. The majority of the students rate the courses highly.I had no expectations because scores vary by students personal preference of Instructors.

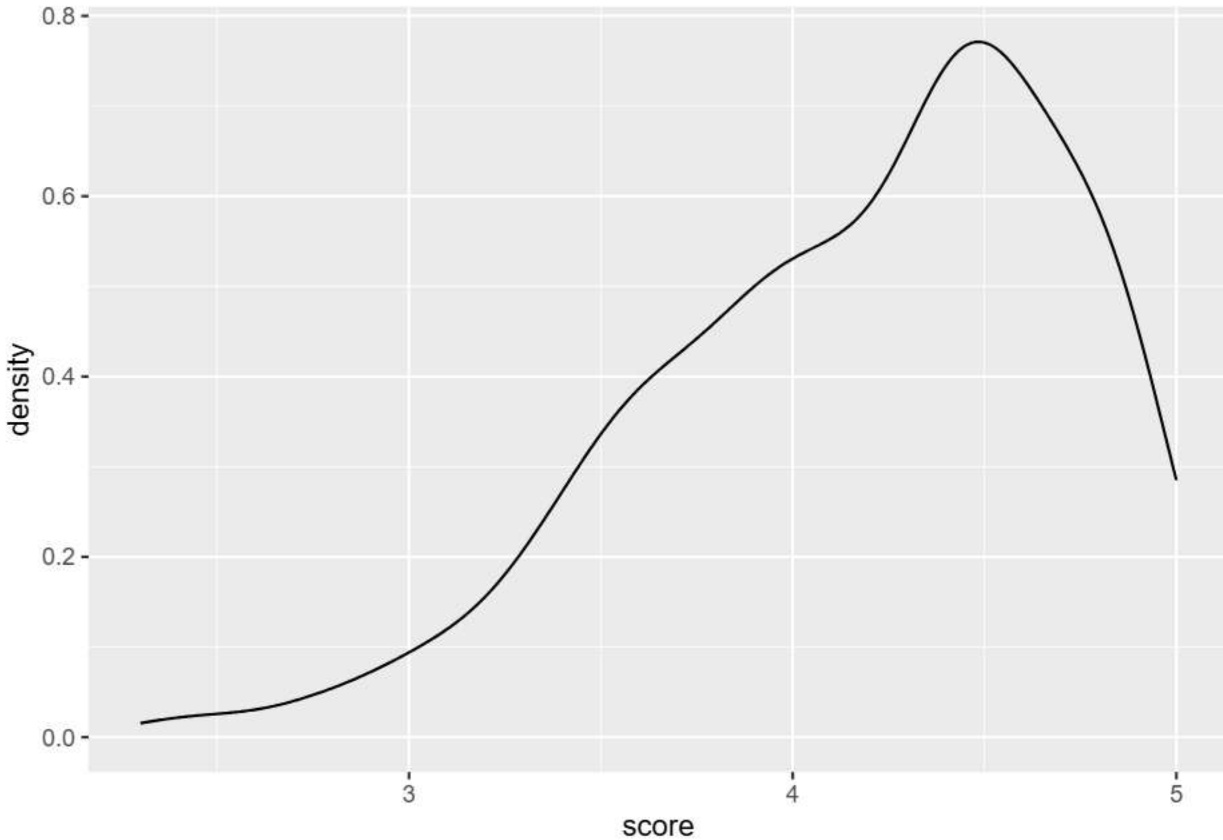
```
summary(evals)
```

```
##    course_id        prof_id          score                   rank
##  Min.   :  1.0   Min.   : 1.00   Min.   :2.300   teaching    :102
##  1st Qu.:116.5   1st Qu.:20.00   1st Qu.:3.800   tenure track:108
##  Median :232.0   Median :43.00   Median :4.300   tenured     :253
##  Mean   :232.0   Mean   :45.15   Mean   :4.175
##  3rd Qu.:347.5   3rd Qu.:70.50   3rd Qu.:4.600
##  Max.   :463.0   Max.   :94.00   Max.   :5.000
##         ethnicity      gender          language         age
##  minority    : 64   female:195   english    :435   Min.   :29.00
##  not minority:399   male  :268   non-english: 28   1st Qu.:42.00
##                                                    Median :48.00
##                                                    Mean   :48.37
##                                                    3rd Qu.:57.00
##                                                    Max.   :73.00
##  cls_perc_eval     cls_did_eval     cls_students    cls_level       cls_profs
##  Min.   : 10.42   Min.   :  5.00   Min.   :  8.00   lower:157   multiple:306
##  1st Qu.: 62.70   1st Qu.: 15.00   1st Qu.: 19.00   upper:306   single  :157
##  Median : 76.92   Median : 23.00   Median : 29.00
##  Mean   : 74.43   Mean   : 36.62   Mean   : 55.18
##  3rd Qu.: 87.25   3rd Qu.: 40.00   3rd Qu.: 60.00
##  Max.   :100.00   Max.   :380.00   Max.   :581.00
##        cls_credits   bty_f1lower     bty_f1upper      bty_f2upper
##  multi credit:436   Min.   :1.000   Min.   :1.000   Min.   : 1.000
##  one credit  : 27   1st Qu.:2.000   1st Qu.:4.000   1st Qu.: 4.000
##                     Median :4.000   Median :5.000   Median : 5.000
##                     Mean   :3.963   Mean   :5.019   Mean   : 5.214
##                     3rd Qu.:5.000   3rd Qu.:7.000   3rd Qu.: 6.000
##                     Max.   :8.000   Max.   :9.000   Max.   :10.000
##   bty_m1lower     bty_m1upper      bty_m2upper        bty_avg
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.667
##  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:4.000   1st Qu.:3.167
##  Median :3.000   Median :4.000   Median :5.000   Median :4.333
##  Mean   :3.413   Mean   :4.147   Mean   :4.752   Mean   :4.418
##  3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:5.500
##  Max.   :7.000   Max.   :9.000   Max.   :9.000   Max.   :8.167
##      pic_outfit         pic_color
##  formal    : 77   black&white: 78
##  not formal:386   color      :385
##
##
##
##
```

```
ggplot(evals, aes(x=score)) +
  geom_histogram(binwidth=.5, colour="black", fill="white")
```

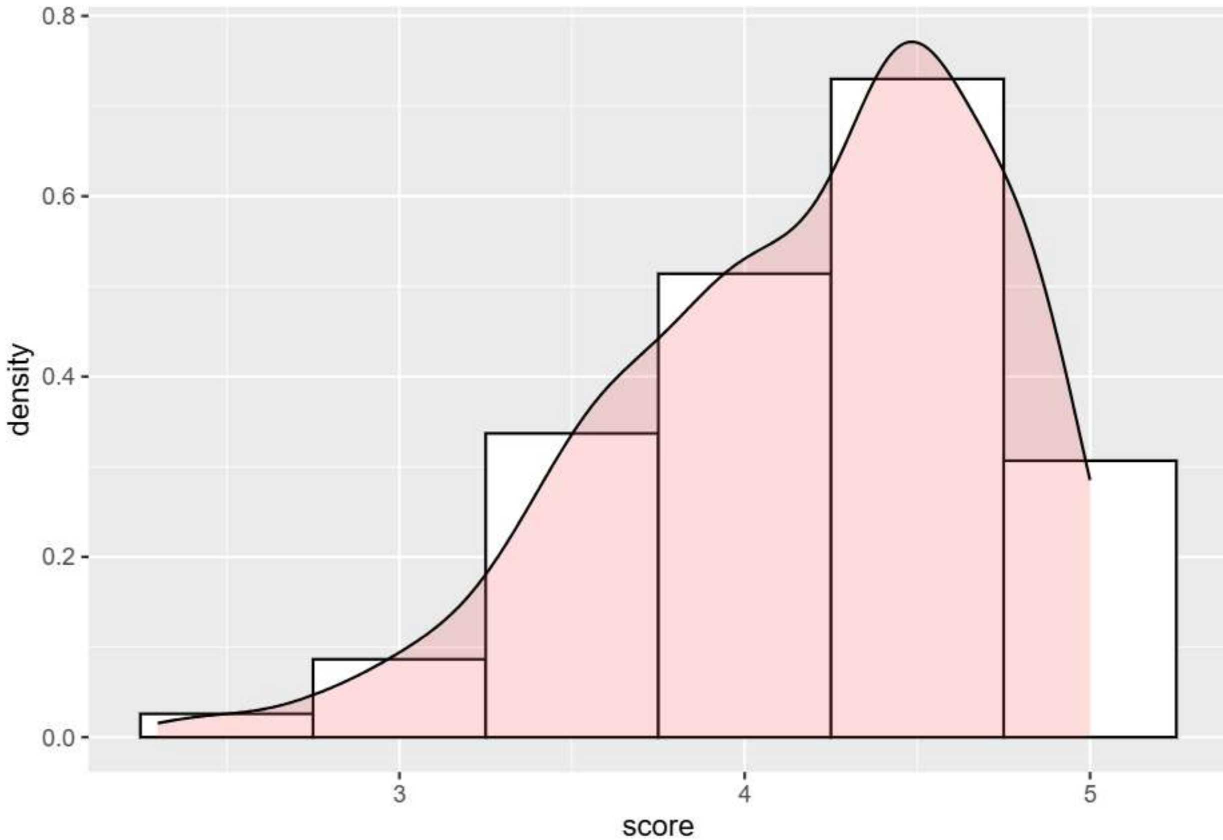



```r
ggplot(evals, aes(x=score)) + geom_density()
```




```
ggplot(evals, aes(x=score)) +
    geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis
                   binwidth=.5,
                   colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



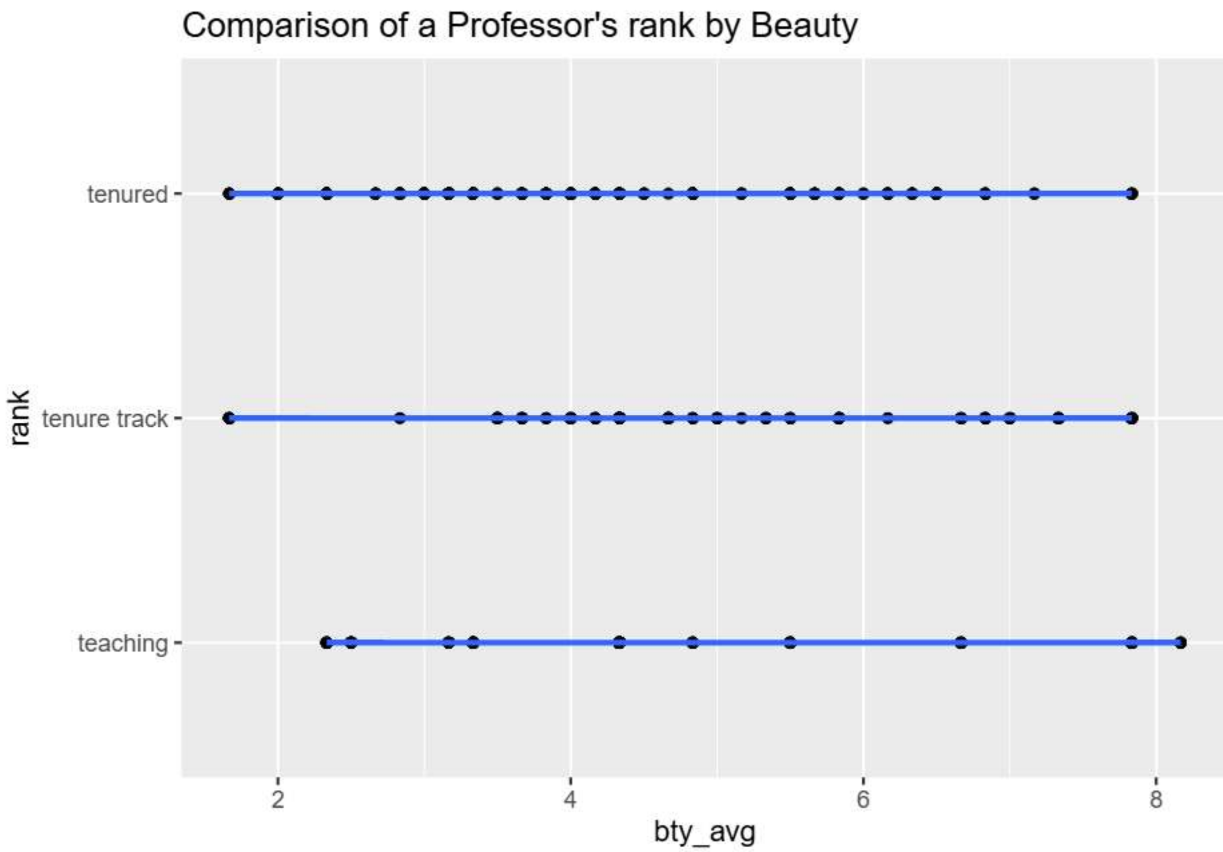
Exercise 3

Excluding score, select two other variables and describe their relationship with each other using an appropriate visualization.

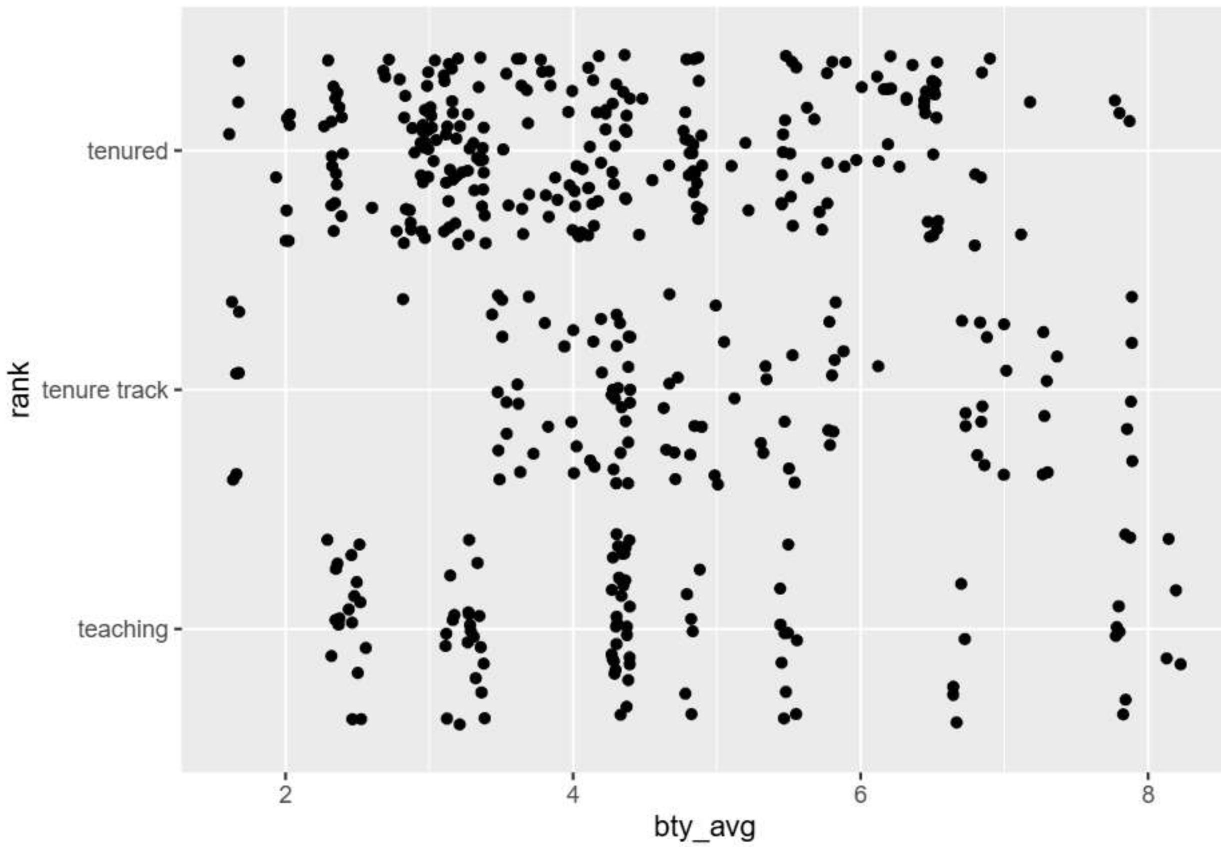
A professors' beauty rating does not affect Tenure, teaching rank is most highly affected by beauty. The highest number of teachers with tenure do not score highly in beauty average. The professors' on tenure track have the highest median beauty average of around 4.7.

```
ggplot(data = evals, aes(x = bty_avg, y = rank)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)+
 ggtitle("Comparison of a Professor's rank by Beauty ")
```

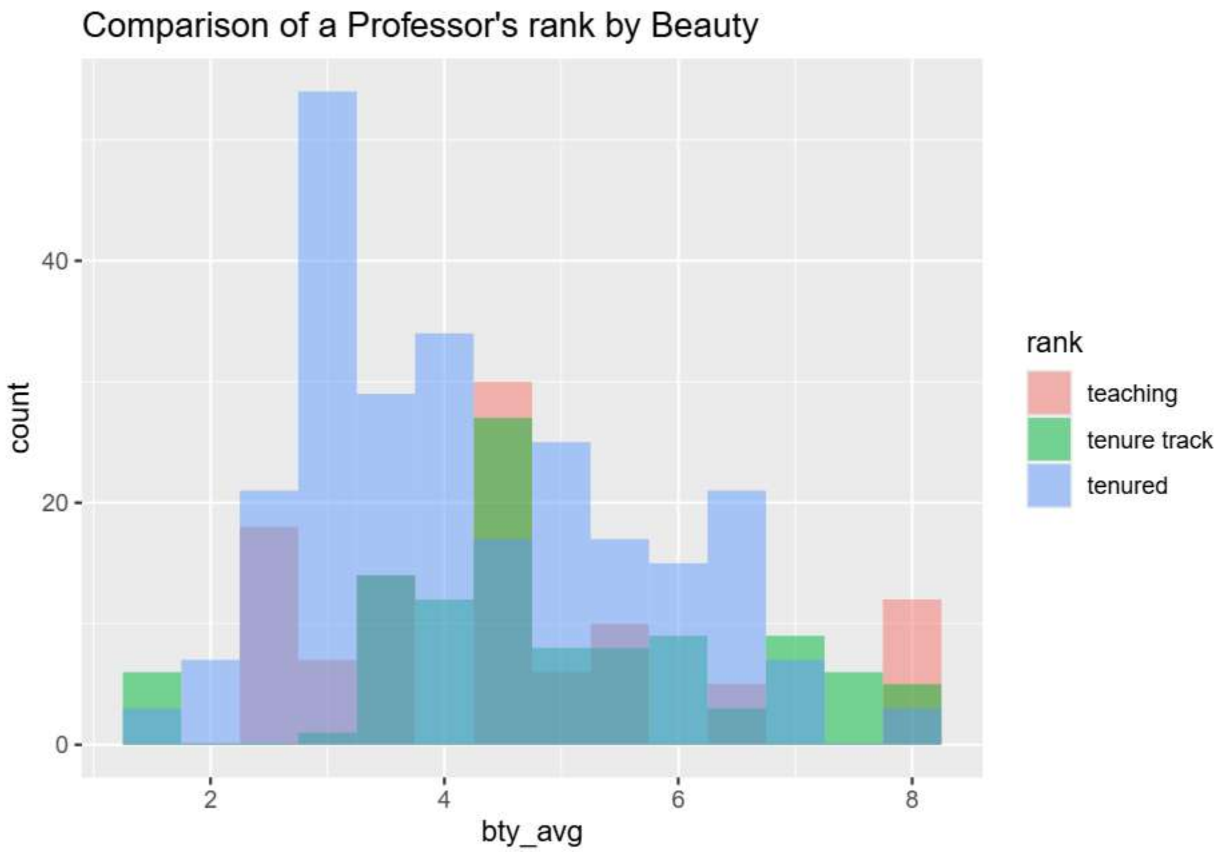
```
## `geom_smooth()` using formula = 'y ~ x'
```



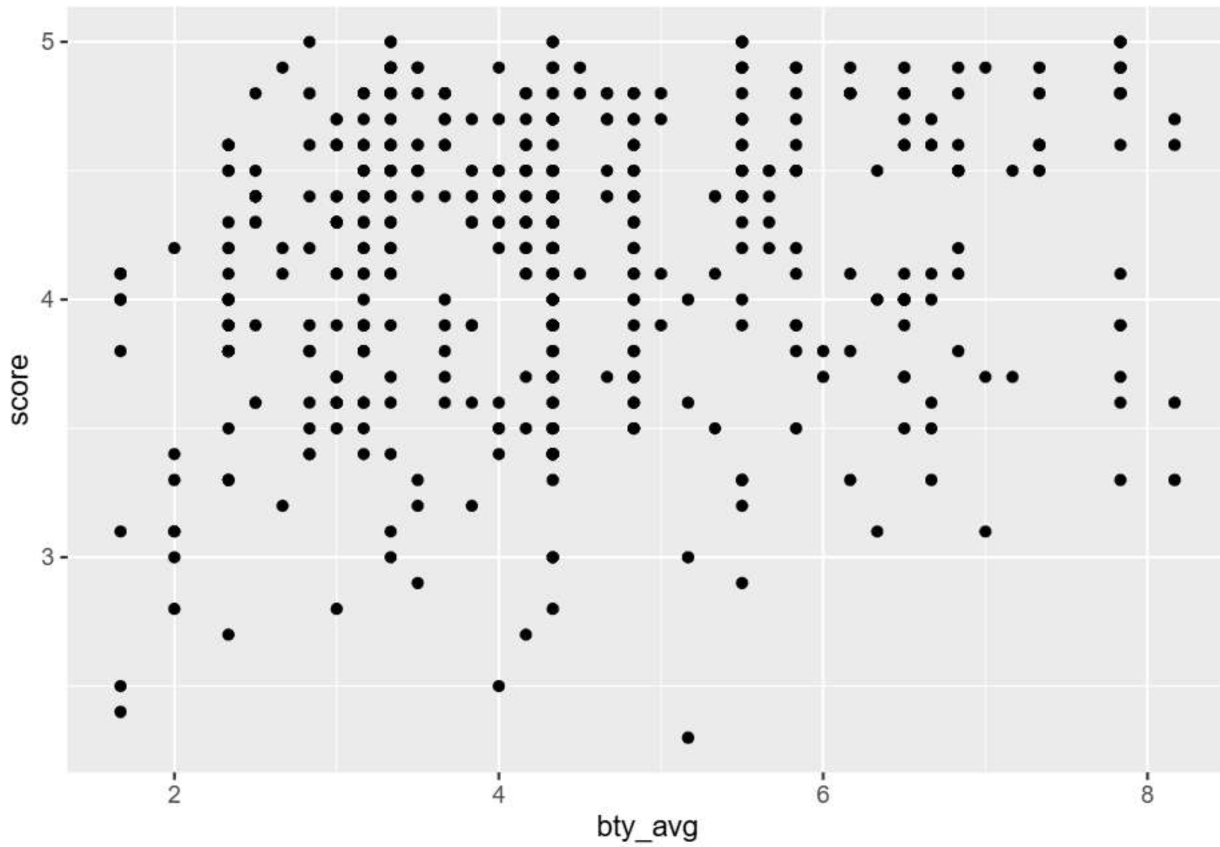

```r
ggplot(data = evals, aes(x = bty_avg, y = rank)) +
  geom_jitter()
```




```
ggplot(evals, aes(x=bty_avg, fill=rank)) +
    geom_histogram(binwidth=.5, alpha=.5, position="identity")+
 ggtitle("Comparison of a Professor's rank by Beauty ")
```

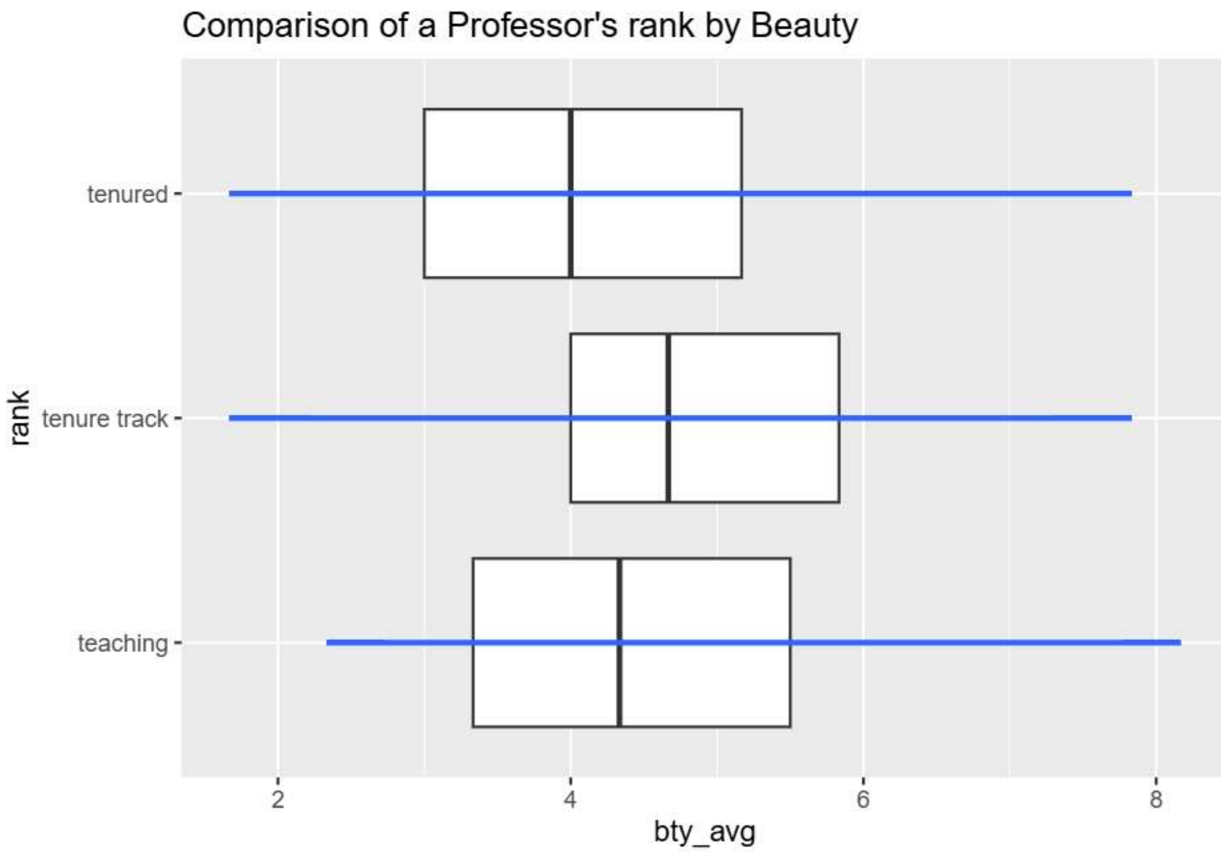



```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_point()
```




```
ggplot(data = evals, aes(x = bty_avg, y = rank)) +
  geom_boxplot() +
  stat_smooth(method = "lm", se = FALSE)+
 ggtitle("Comparison of a Professor's rank by Beauty ")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

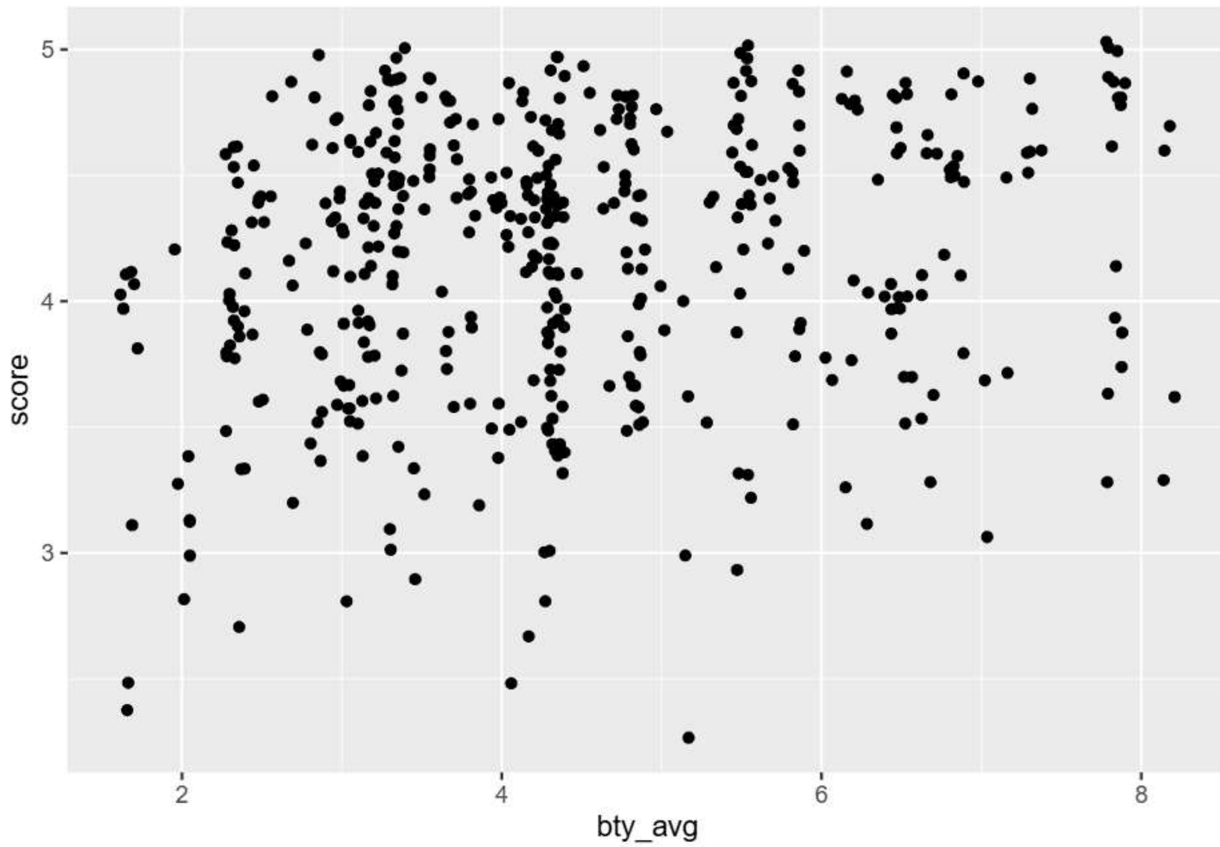


Exercise 4

Replot the scatterplot, but this time use geom_jitter as your layer. What was misleading about the initial scatterplot?

Geom_jitter added a small amount of random variation to the location of each plot, and was a useful way of handling over plotting since evals is a small dataset.

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter()
```



Exercise 5

Let's see if the apparent trend in the plot is something more than natural variation. Fit a linear model called m_bty to predict average professor score by average beauty rating. Write out the equation for the linear model and interpret the slope. Is average beauty score a statistically significant predictor? Does it appear to be a practically significant predictor?

```
m_bty <- lm(evals$score  ~ evals$bty_avg)
summary(m_bty)
```

```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## 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 ***
## evals$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
```

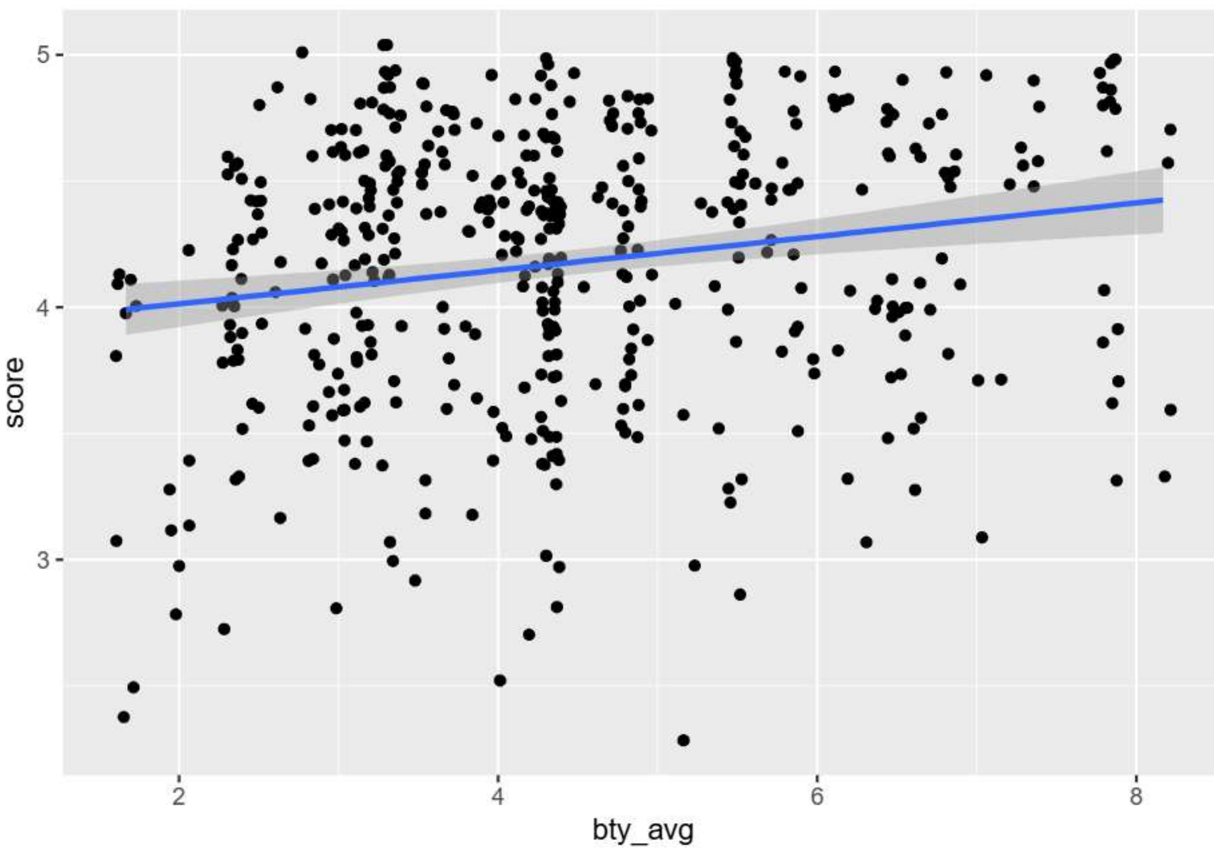
score = ax + B 0.06664*bty_avg + 3.88034

Average beauty score is a statistically significant predictor. It does not appear to be a practically significant predictor.There many more Professors who have a high score that don't have a high beauty average. professors's with a beauty average between 3 t0 6 have the highest concentration of high scores. while those with a low beauty average between 1 to 2.5 seem to have the lowest score.

The slope is not steep. The rate of change at 0.06664 is positive but very slight.

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Exercise 6

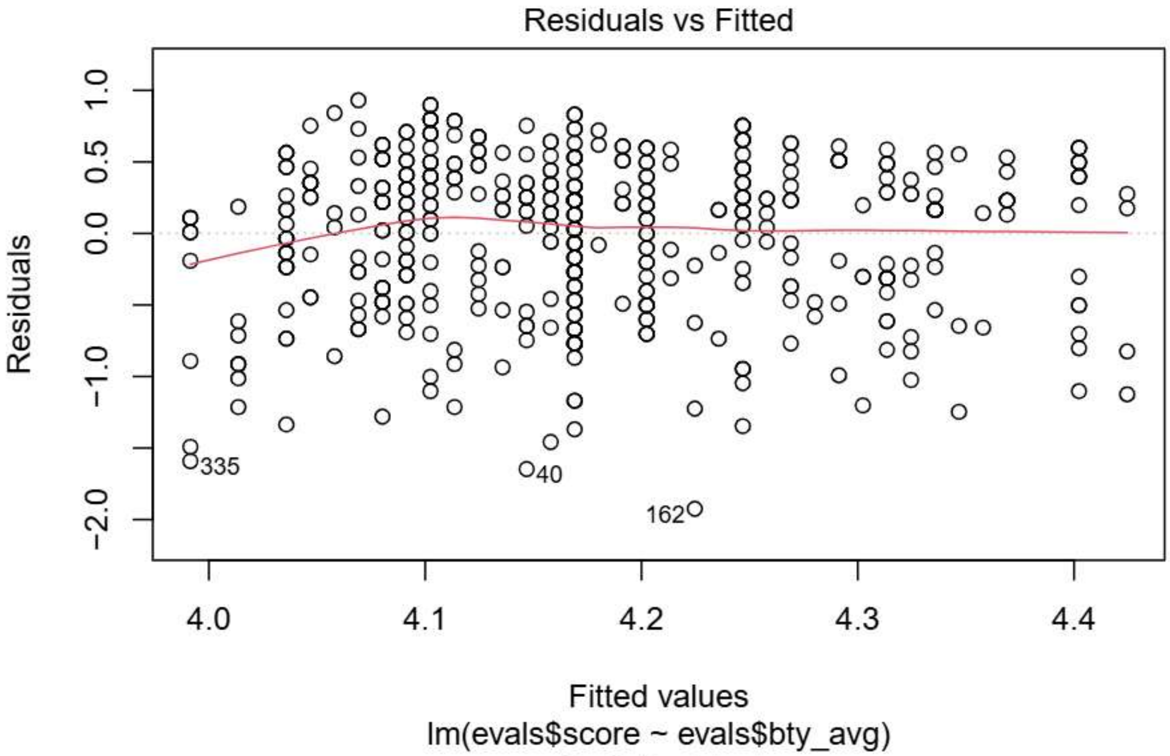
Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

The fitted values symbolize a good regression model. Residuals are randomly scattered around zero with no discernible plot.

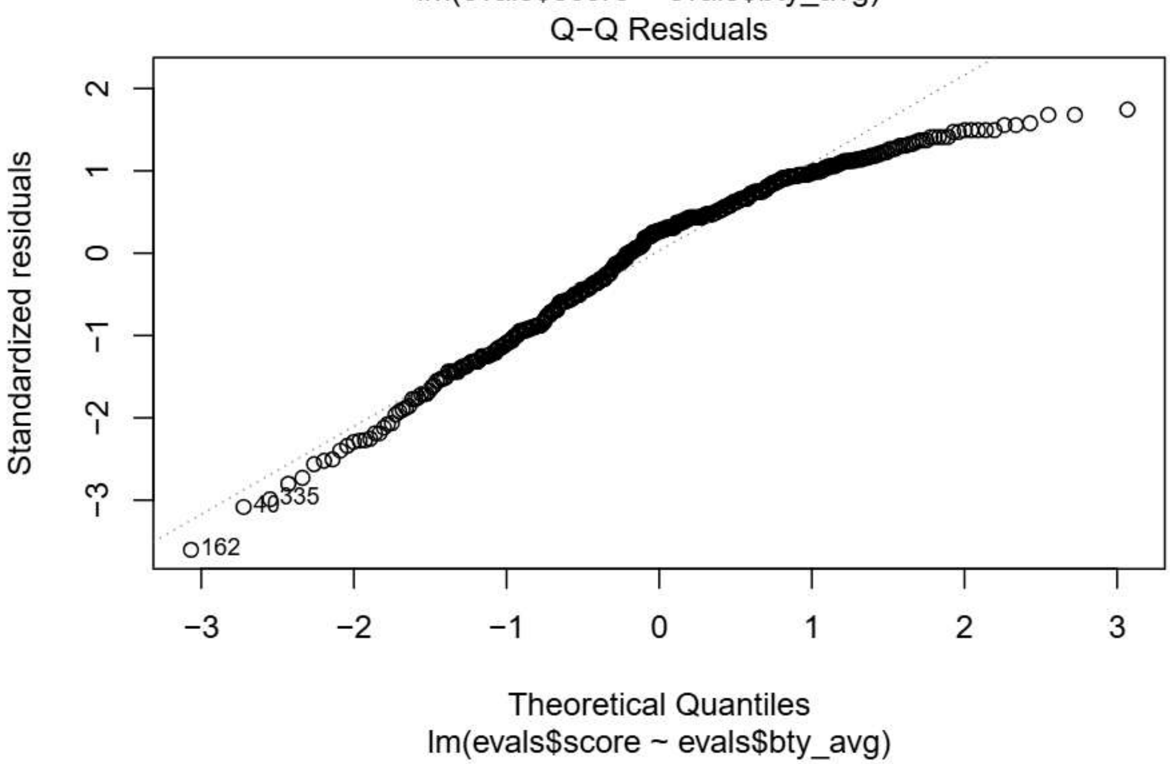
```
par(m_bty = c(2, 2))
```

```
## Warning in par(m_bty = c(2, 2)): "m_bty" is not a graphical parameter
```

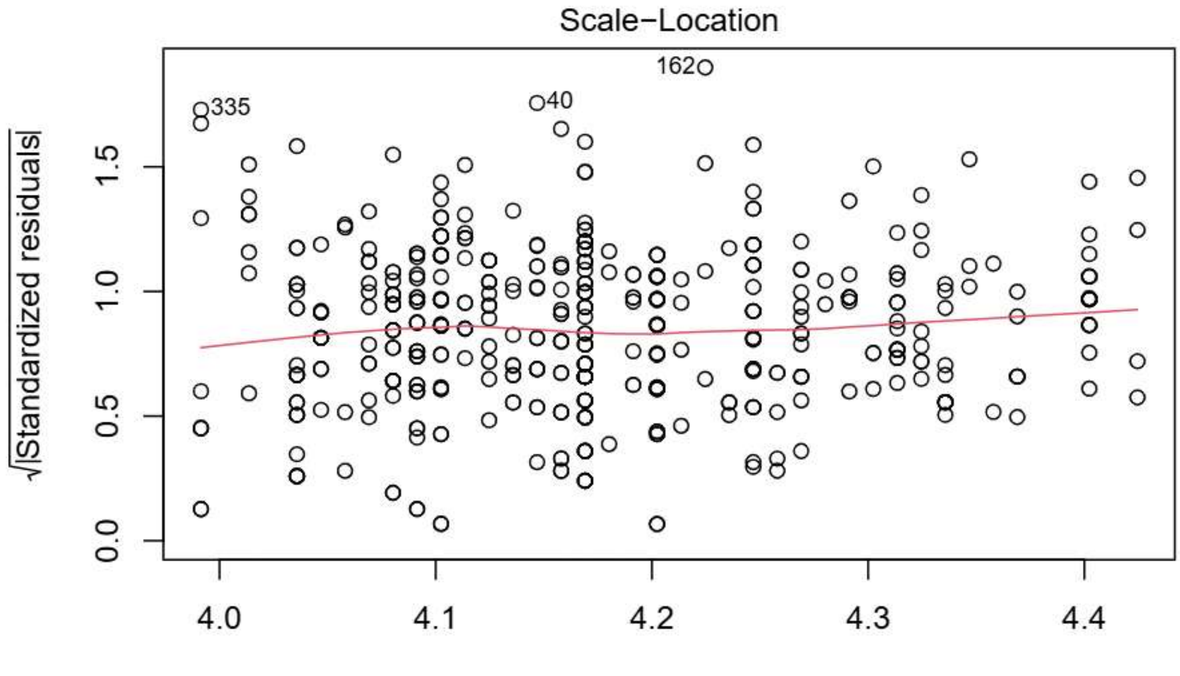
```
plot(m_bty)
```

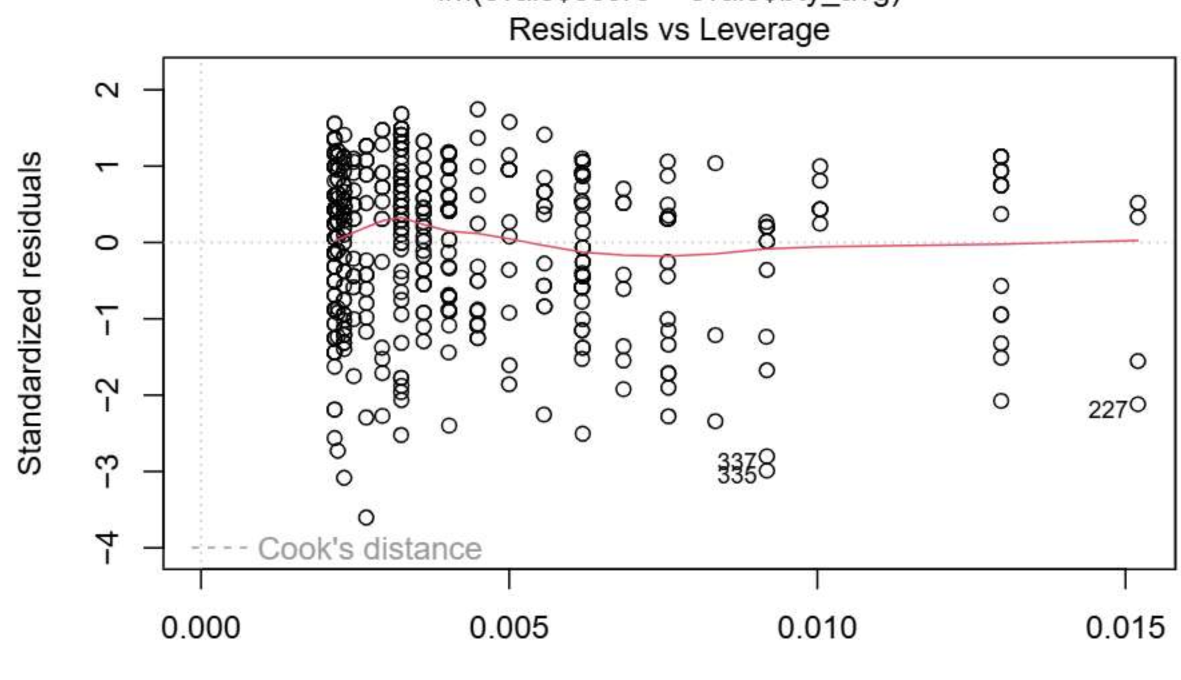


Residuals vs Fitted
Residuals
Fitted values
lm(evals$score ~ evals$bty_avg)
335
40
162

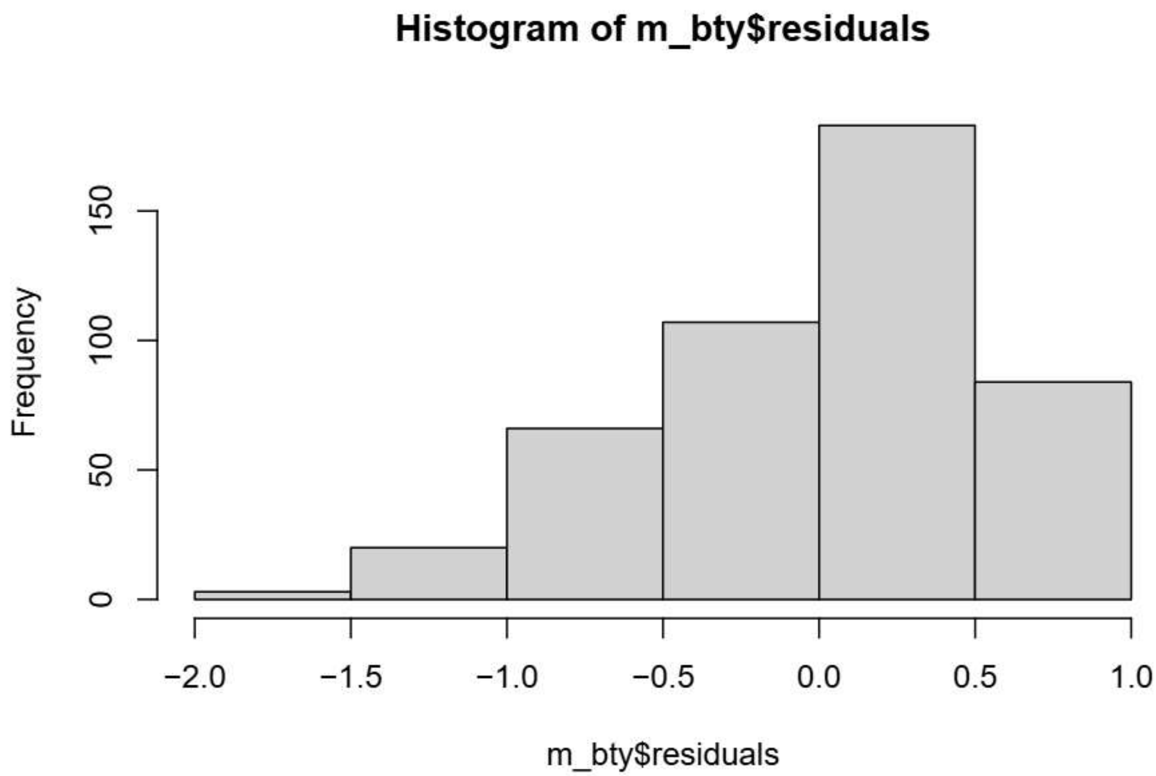


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(evals$score ~ evals$bty_avg)
162
40
335



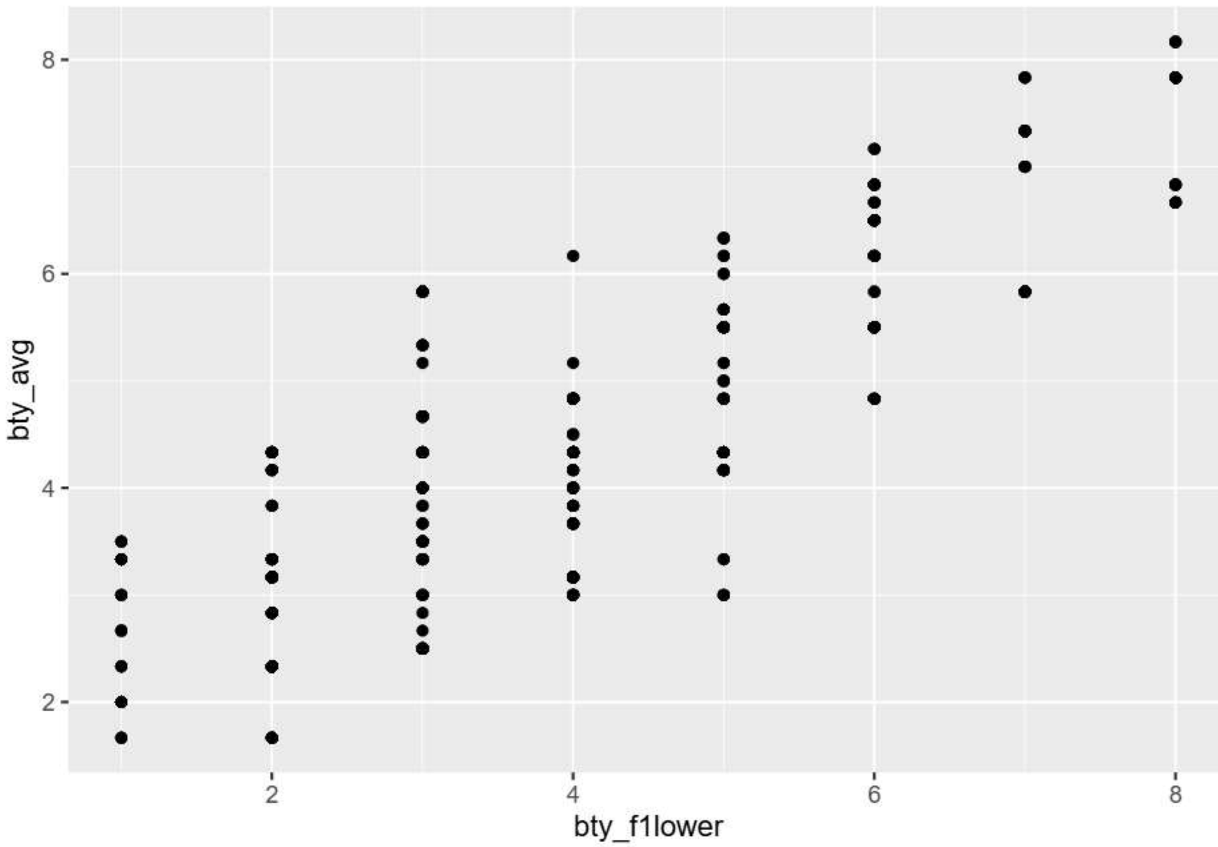



```
hist(m_bty$residuals)
```



Multiple linear regression

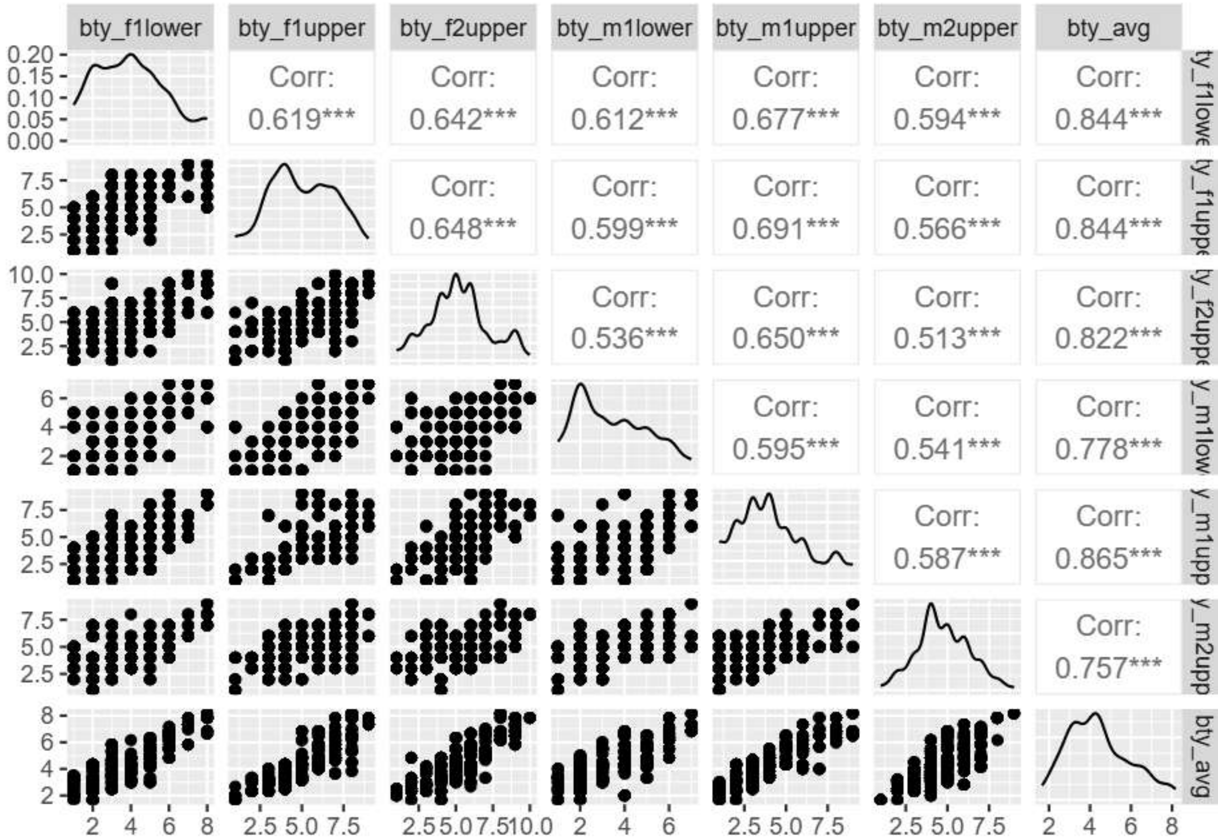
```
ggplot(data = evals, aes(x = bty_f1lower, y = bty_avg)) +
  geom_point()
```




```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # A tibble: 1 x 1
##   `cor(bty_avg, bty_f1lower)`
##                         <dbl>
## 1                       0.844
```

```
evals %>%
  select(contains("bty")) %>%
  ggpairs()
```




```r
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
```

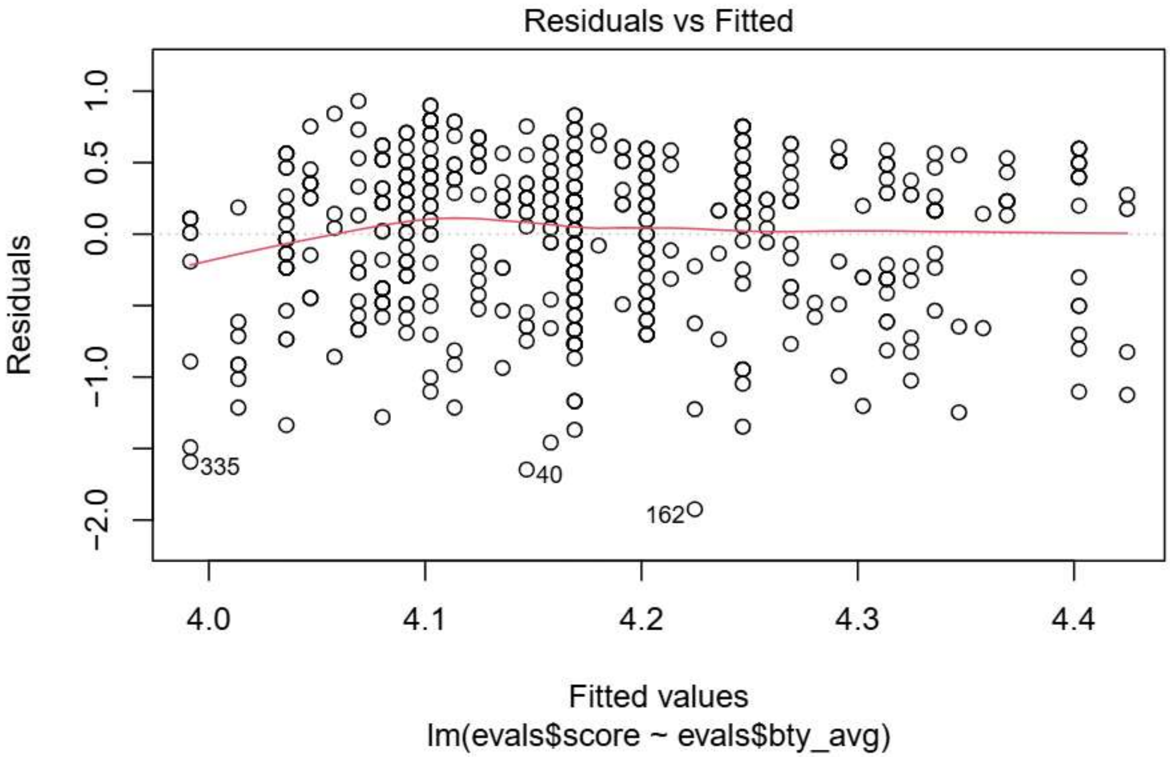
Exercise 7

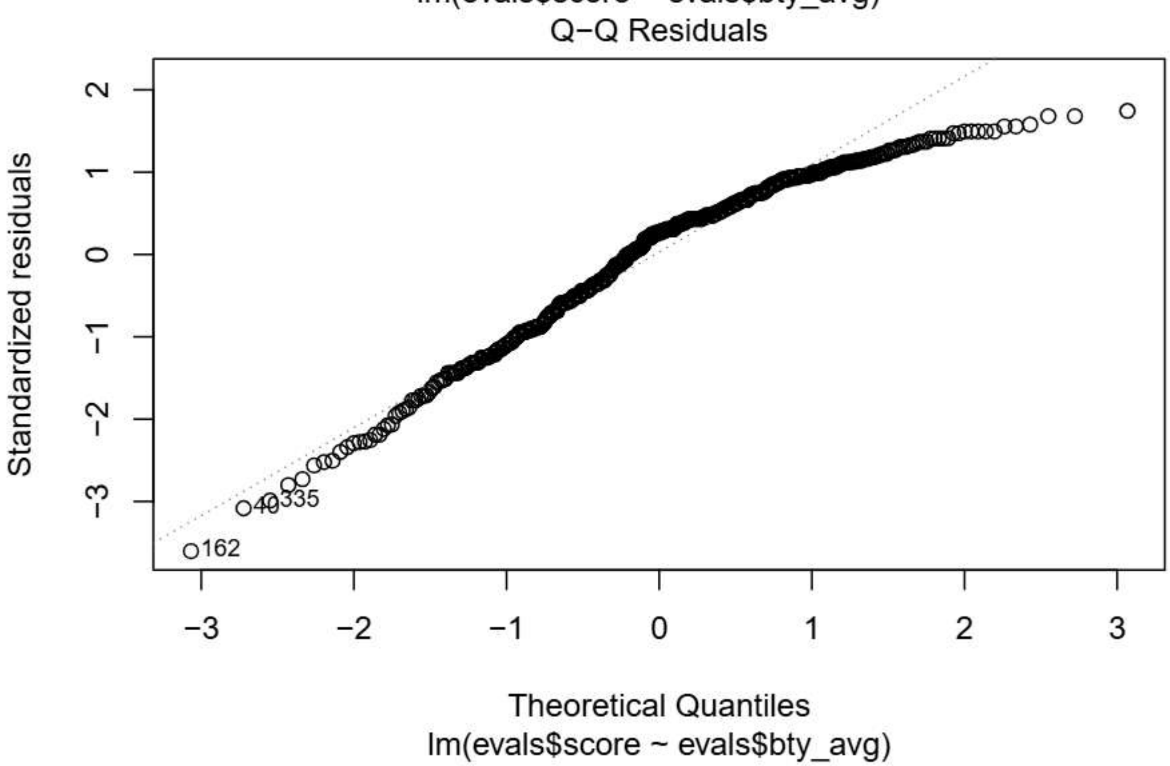
P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

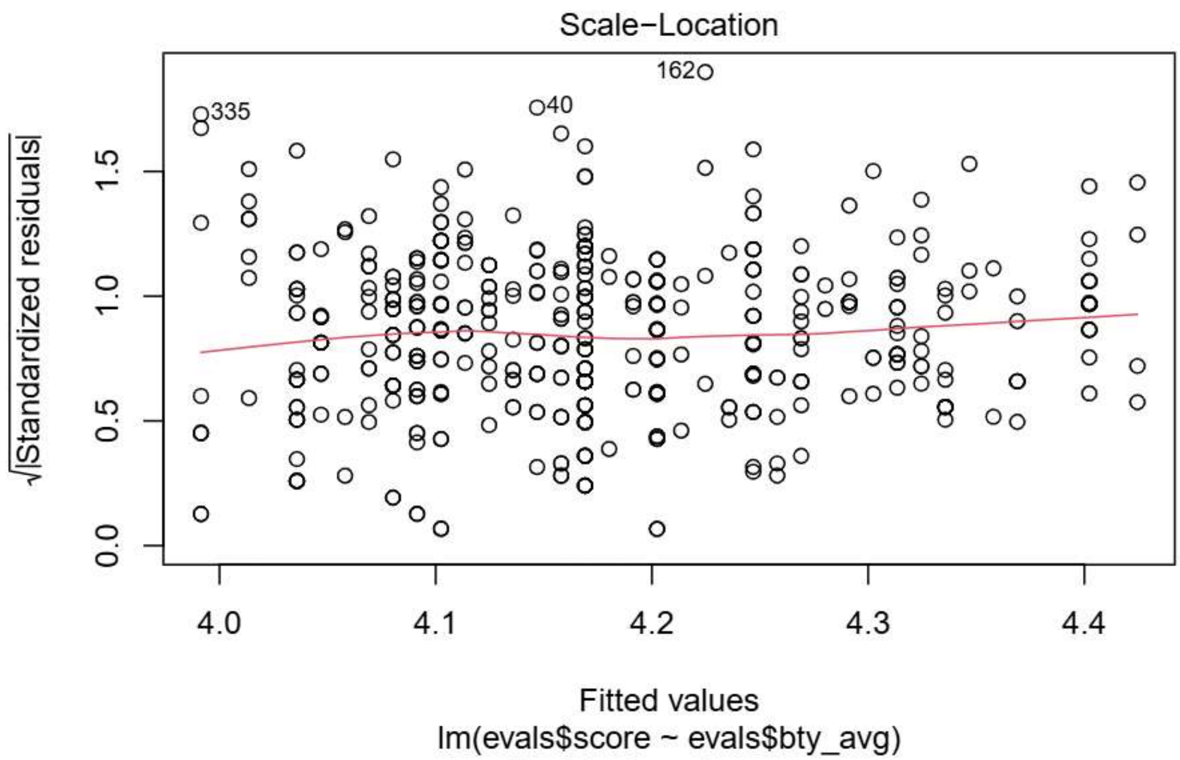
```r
par(m_bty = c(2, 2))
```

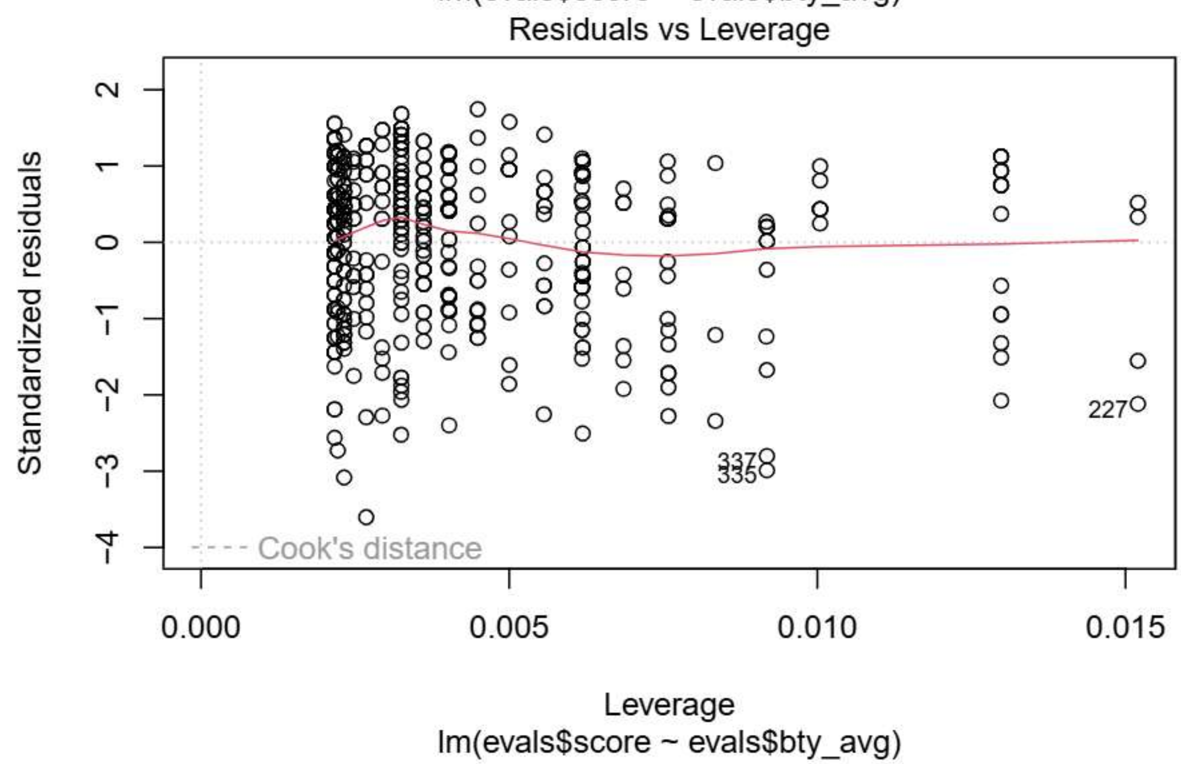
```
## Warning in par(m_bty = c(2, 2)): "m_bty" is not a graphical parameter
```

```r
plot(m_bty)
```









Residual vs fitted is Linear.

QQ residuals - show a curved Pattern that indicates left skewed data or non normal distributions. There is a slight deviation in some areas along the diagonal line with most deviations falling below the line.

Scale-Location- No overall increase, variance is more or less constant.

Residual-Leverage The cooks distance is less than 0.5. There no outliers that affect or influence the resulting model.

Exercise 8

Is bty_avg still a significant predictor of score? Has the addition of gender to the model changed the parameter estimate for bty_avg?

```r
summary(m_bty)
```

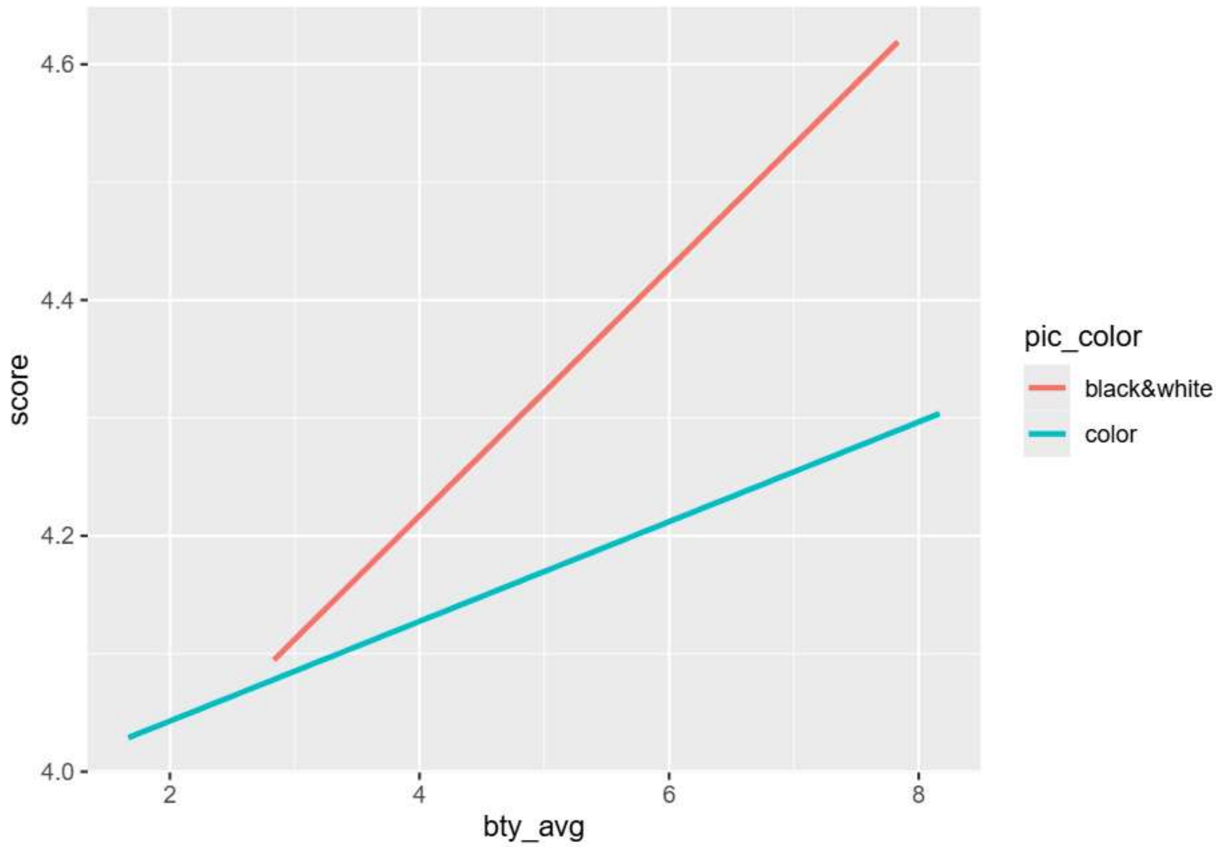
```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.88034    0.07614   50.96  < 2e-16 ***
## evals$bty_avg  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
```

```r
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
```

Yes bty_avg is still a significant predictor of score. The addition of gender to the model changed the parameter estimate for bty_avg to 0.07416 from 0.06664.

```r
ggplot(data = evals, aes(x = bty_avg, y = score, color = pic_color)) +
 geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



Exercise 9

What is the equation of the line corresponding to those with color pictures? (Hint: For those with color pictures, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which color picture tends to have the higher course evaluation score?

The black and white colored picture tends to have the higher course evaluation score.

```
blawite <-
  evals %>%
  filter (pic_color == 'black&white')
color <-
  evals %>%
  filter (pic_color == 'color')
```

```
lm(score ~ bty_avg, data = color)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = color)
##
## Coefficients:
## (Intercept)      bty_avg
##     3.95826      0.04229
```

```
lm(score ~ bty_avg, data = blawite)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = blawite)
```

```
##
## Coefficients:
## (Intercept)      bty_avg
##      3.7974       0.1049
```

```
m_bty_color <- lm(score ~ bty_avg + pic_color, data = evals)
summary(m_bty_color)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8892 -0.3690  0.1293  0.4023  0.9125
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.06318    0.10908  37.249  < 2e-16 ***
## bty_avg          0.05548    0.01691   3.282  0.00111 **
## pic_colorcolor -0.16059    0.06892  -2.330  0.02022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5323 on 460 degrees of freedom
## Multiple R-squared:  0.04628,    Adjusted R-squared:  0.04213
## F-statistic: 11.16 on 2 and 460 DF,  p-value: 1.848e-05
```

Score = Bo + B1 * bty_avg + B2 *(1)

```
= 4.06318 + 0.05548 * bty_avg + -0.16059
```

Exercise 10

Create a new model called m_bty_rank with gender removed and rank added in. How does R appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: teaching, tenure track, tenured.

Only professors of the same rank are compared together. It appear each category is handled independently not together.

```
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
```

Exercise 11

Which variable would you expect to have the highest p-value in this model? Why? Hint: Think about which variable would you expect to not have any association with the professor score.

cls_credit Variable might have the highest P value . Because students are less likely to consider the number of credits when giving a score to a Professor.

```r
m_full <- lm(score ~ rank + gender + ethnicity + 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 + gender + ethnicity + 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
## gendermale            0.2109481  0.0518230   4.071 5.54e-05 ***
## ethnicitynot minority 0.1234929  0.0786273   1.571  0.11698
## 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
```

Exercise 12

Check your suspicions from the previous exercise. Include the model output in your response.

```r
m_full <- lm(score ~ rank + gender + ethnicity + 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 + gender + ethnicity + 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
## gendermale             0.2109481  0.0518230   4.071 5.54e-05 ***
## ethnicitynot minority  0.1234929  0.0786273   1.571  0.11698
## 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
```

Exercise 13

Interpret the coefficient associated with the ethnicity variable.

Ethnicitynot minority = 0.1234929. When other variables are held constant ethinicity increases by 12% for every one unit change in the independent variables.

Exercise 14

Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables?

The original Intercept estimate of 4.0952141 went up 4.3098194 when i removed cls_creditsone credit. The estimate of cls_profssingle changed to -0.0427280 from -0.0146619. The R-squared decreased to 0.1531 this means the value removed from the data was influential in explaining the variance on the data. cls_creditsone credit has a small correlation with other predictors because coefficients of other variables changed when it was removed.

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

```
##
## Call:
## lm(formula = score ~ rank + gender + ethnicity + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + bty_avg +
##     pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7498 -0.3200  0.1056  0.3679  0.9200
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.3098194  0.2918733  14.766  < 2e-16 ***
## ranktenure track      -0.1957586  0.0829015  -2.361 0.018635 *
## ranktenured           -0.1809000  0.0647027  -2.796 0.005398 **
## gendermale             0.2366593  0.0524895   4.509 8.33e-06 ***
## ethnicitynot minority  0.0429967  0.0778938   0.552 0.581229
## languagenon-english   -0.2589399  0.1133484  -2.284 0.022810 *
## age                   -0.0090463  0.0031973  -2.829 0.004873 **
## cls_perc_eval          0.0059006  0.0015636   3.774 0.000182 ***
## cls_students           0.0002954  0.0003829   0.771 0.440863
## cls_levelupper        -0.0065495  0.0565243  -0.116 0.907807
## cls_profssingle       -0.0427280  0.0525927  -0.812 0.416974
## bty_avg                0.0315543  0.0177371   1.779 0.075917 .
## pic_outfitnot formal  -0.1362125  0.0751223  -1.813 0.070467 .
## pic_colorcolor        -0.2091633  0.0728769  -2.870 0.004297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5077 on 449 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.1286
## F-statistic: 6.243 on 13 and 449 DF,  p-value: 7.671e-11
```

Exercise 15

Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.

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

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

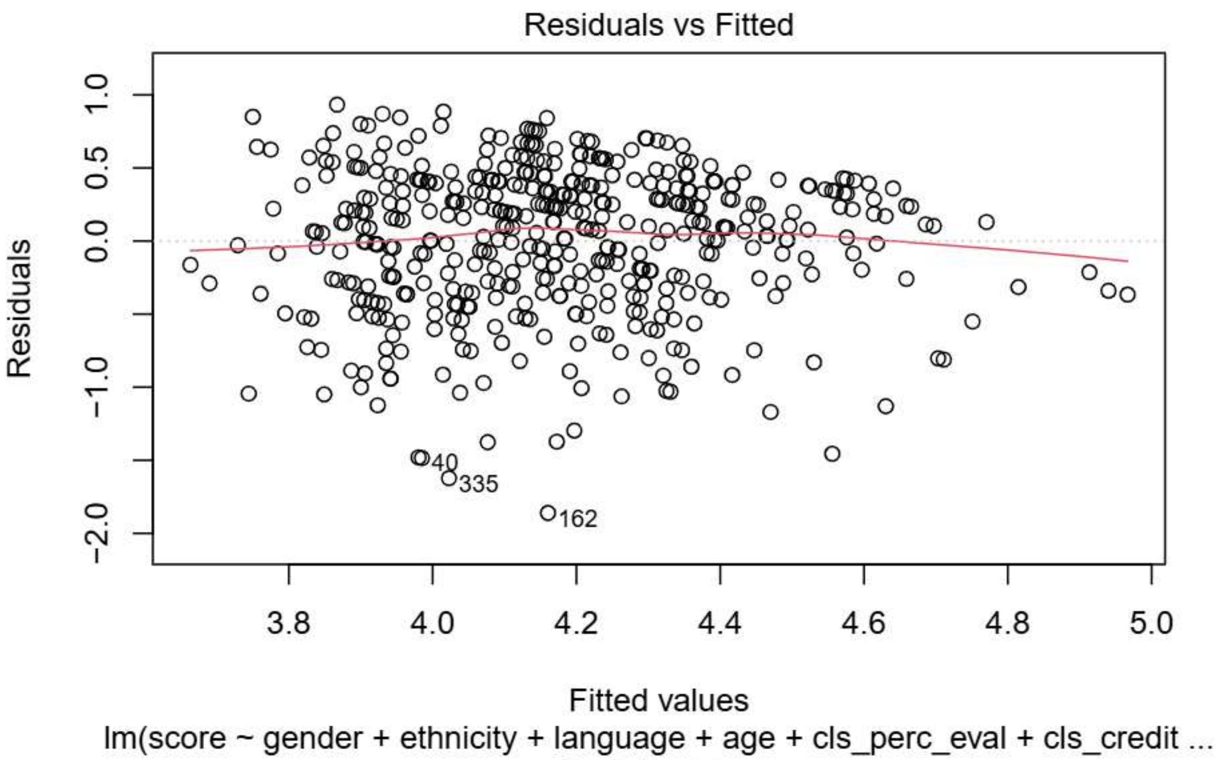
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8604 -0.3129  0.0889  0.3782  0.9328
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.782610   0.232702  16.255  < 2e-16 ***
## gendermale             0.211402   0.050547   4.182 3.47e-05 ***
## ethnicitynot minority  0.159237   0.076349   2.086  0.03757 *
## languagenon-english   -0.221755   0.106123  -2.090  0.03721 *
## age                   -0.006272   0.002634  -2.381  0.01767 *
## cls_perc_eval          0.004524   0.001449   3.123  0.00191 **
## cls_creditsone credit  0.528038   0.109254   4.833 1.84e-06 ***
## cls_levelupper         0.038055   0.055094   0.691  0.49009
## bty_avg                0.050869   0.016946   3.002  0.00283 **
## pic_colorcolor        -0.202161   0.069445  -2.911  0.00378 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4995 on 453 degrees of freedom
## Multiple R-squared:  0.173,  Adjusted R-squared:  0.1566
## F-statistic: 10.53 on 9 and 453 DF,  p-value: 7.415e-15
```

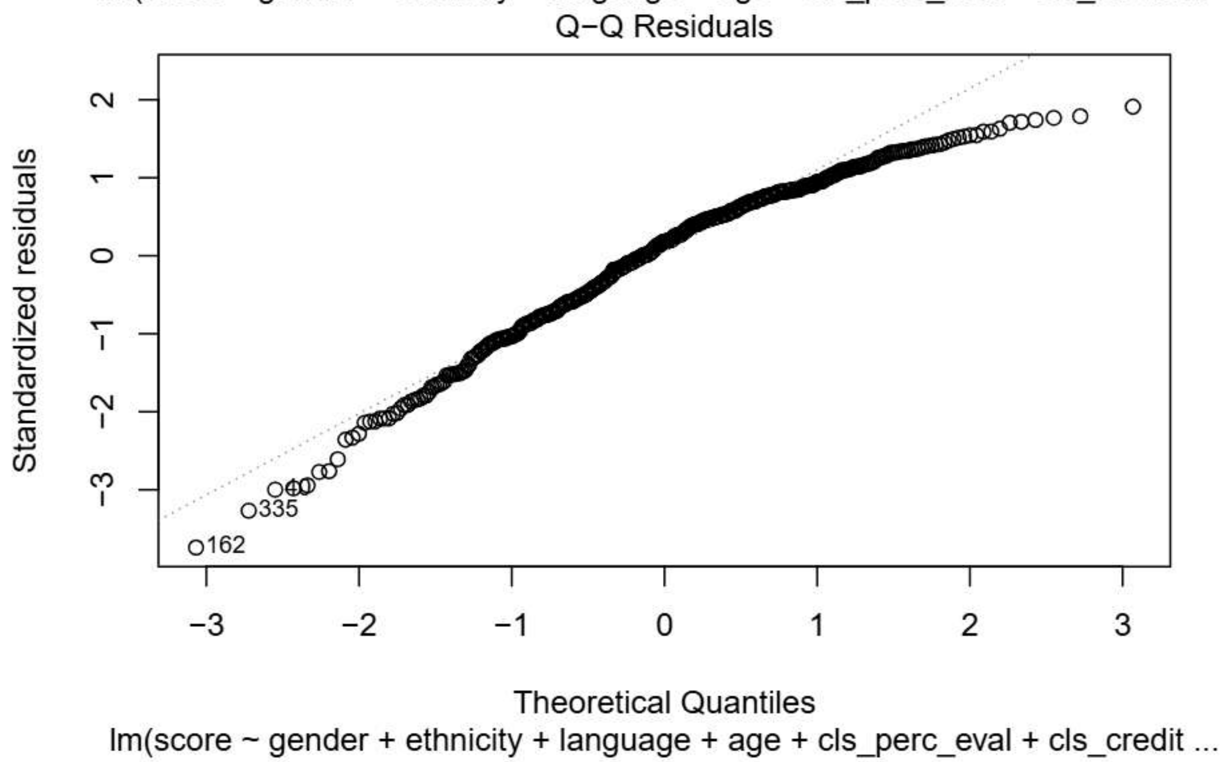
score = 3.782610 +(gender* 0.211402)+ (ethinicty* 0.159237) + (language* -0.221755) + (age* -0.006272) + (cls_perc_eval* 0.004524) + (cls_credits* 0.528038) + (cls_level* 0.038055) + (bty_avg* 0.050869) + (pic_color*-0.202161)

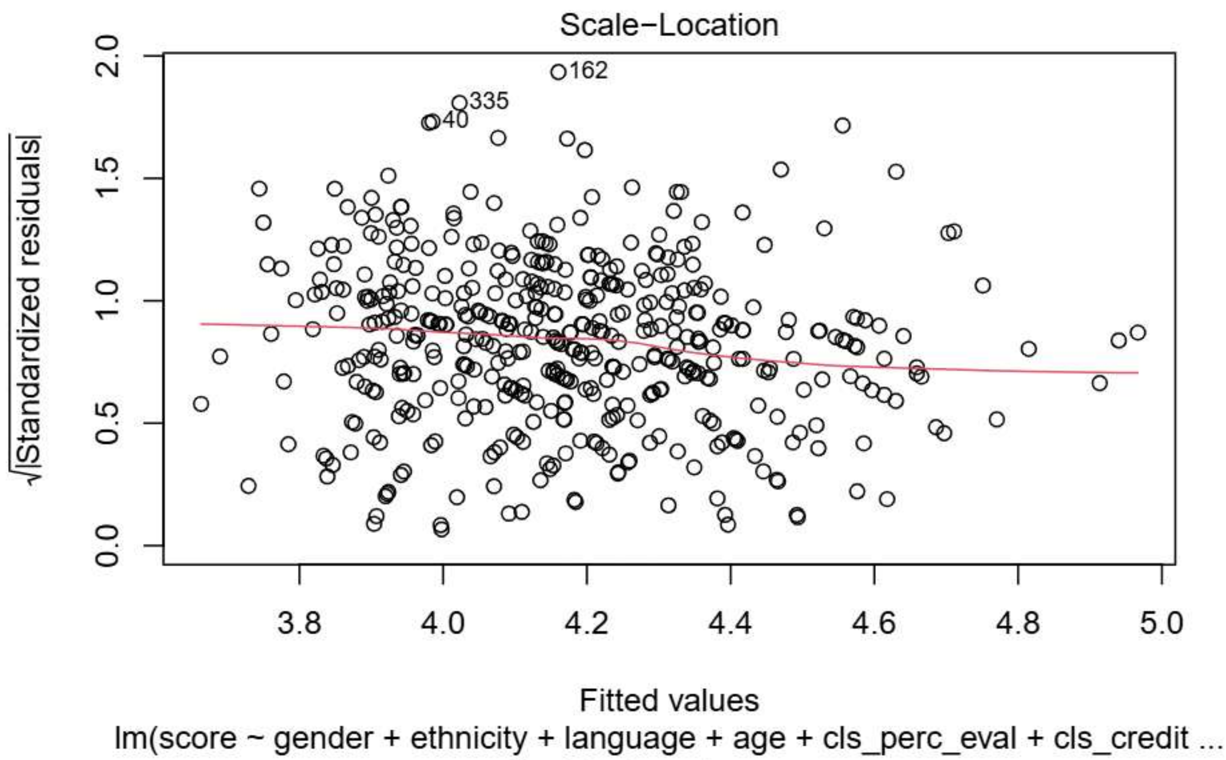
Exercise 16

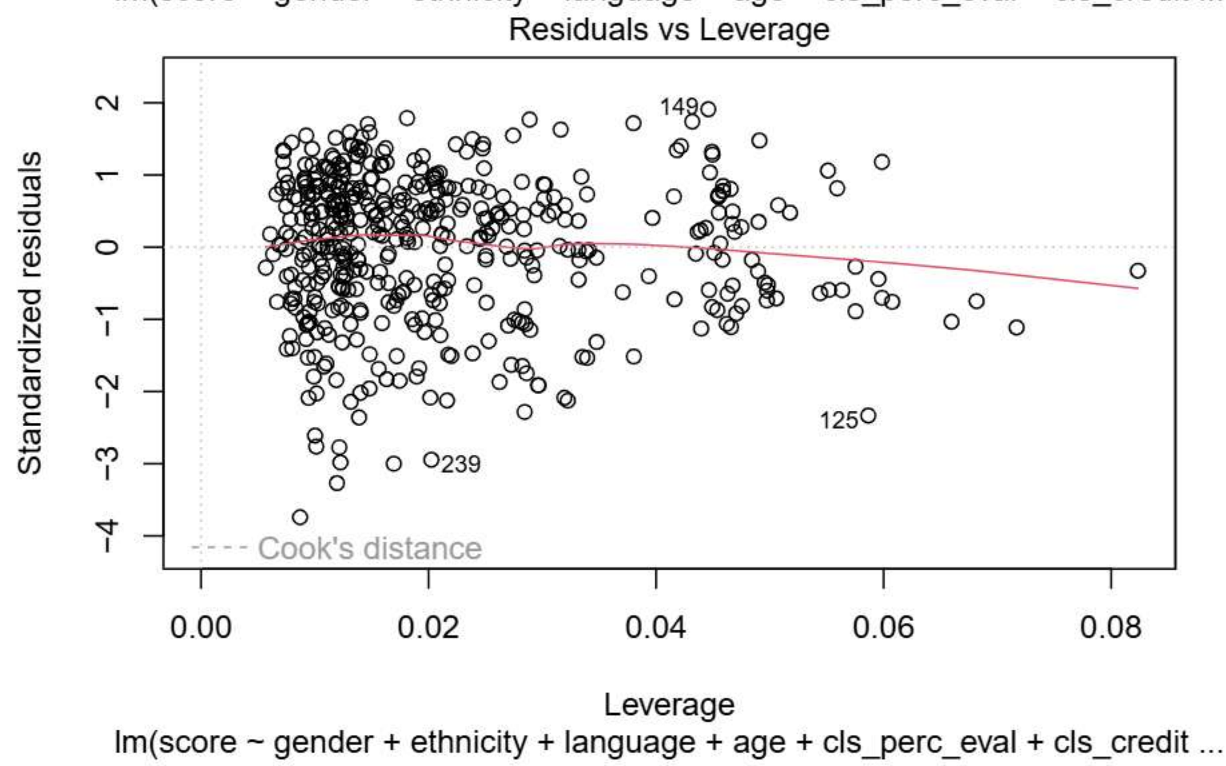
Verify that the conditions for this model are reasonable using diagnostic plots.

```r
plot(m_full_best_model )
```


Residuals vs Fitted
Residuals
Fitted values
lm(score ~ gender + ethnicity + language + age + cls_perc_eval + cls_credit ...
40
335
162


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(score ~ gender + ethnicity + language + age + cls_perc_eval + cls_credit ...
335
40
162





Residual vs fitted is Linear.

QQ residuals - show a curved Pattern that indicates left skewed data or non normal distributions. There is a slight deviation in some areas along the diagonal line with most deviations falling below the line.

Scale-Location- No overall increase, variance is more or less constant.

Residual-Leverage The cooks distance is less than 0.5. There no outliers that affect or influence the resulting model.

Exercise 17

The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?

Yes. This new information could have an impact on the conditions of linear regression. The same professor could have a different rating in different courses.One course could be taught by different professors.

Exercise 18

Based on your final model, describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.

one credit courses might have a high evaluation score.

Male,younger professors that are not minorities, went to a college that taught in English and have a color picture whose bty_avg were rated highly by students would be associated with a high evaluation score.

Exercise 19

Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not?

No I would not generalize. Because different universities might have different standards and preferences. A larger sample size from multiple universities around the country would be more accurate of what characteristics students associate with a highly evaluated Professor.