Multiple linear regression

Grading the professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity" by Hamermesh and Parker found that instructors who are viewed to be better looking receive higher instructional ratings.

Here, you will analyze the data from this study in order to learn what goes into a positive professor evaluation.

Getting Started

Load packages

In this lab, you will explore and visualize the data using the **tidyverse** suite of packages. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

```
library(tidyverse)
library(openintro)
library(GGally)
```

This is the first time we're using the `GGally` package. You will be using the `ggpairs` function from this package later in the lab.

The data

The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors' physical appearance. The result is a data frame where each row contains a different course and columns represent variables about the courses and professors. It's called `evals`.

```
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, ~
```

We have observations on 21 different variables, some categorical and some numerical. The meaning of each variable can be found by bringing up the help file:

```
?evals
```

Exploring the data

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.

Insert your answer here

This is an observational study as we are looking at the results of a survey. It would be an experimental study if there was a control and experimental group. Because this is an observational study, our results can only be that we have evidence that suggests a correlation between beauty and student evaluation.

End of your answer

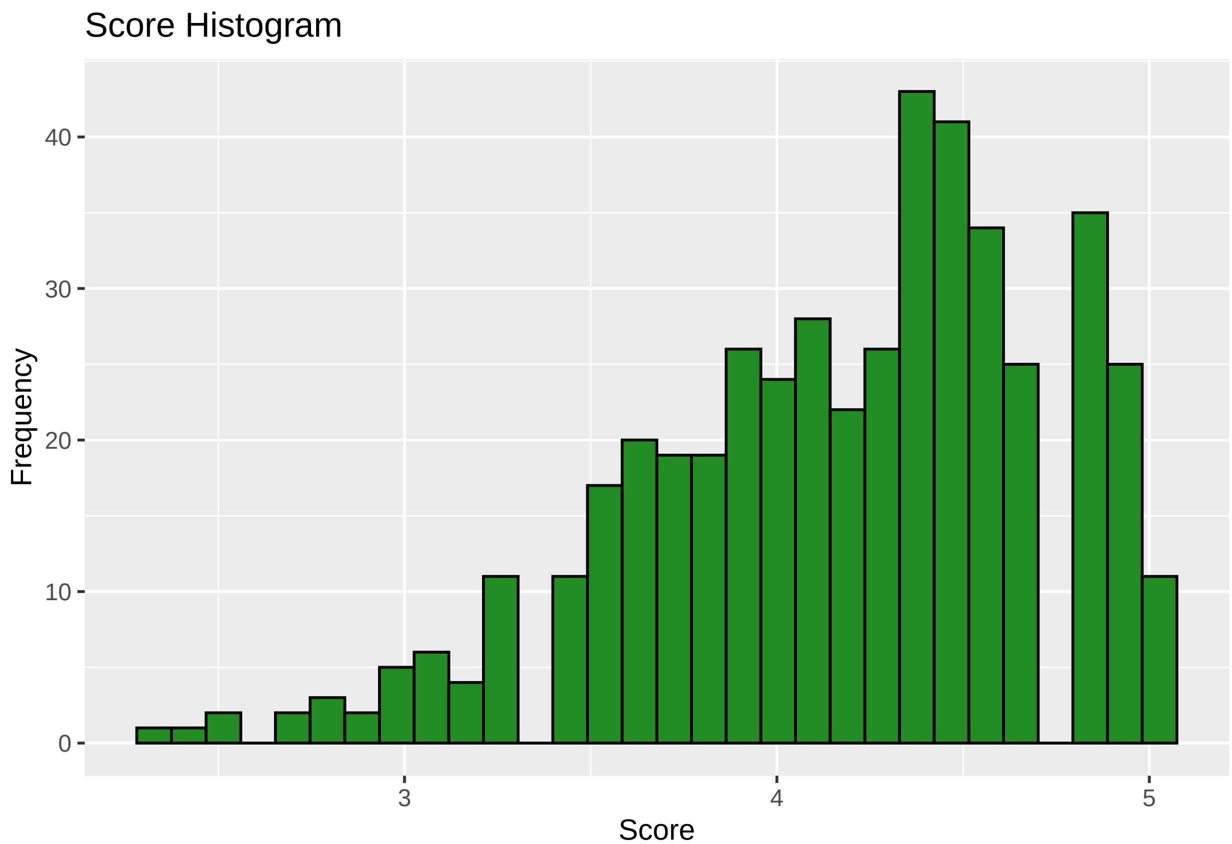
2. Describe the distribution of `score`. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

Insert your answer here

```r
library(ggplot2)

ggplot(
  evals,
  aes(
    x = score
  )
) +
  geom_histogram(
    fill = "forestgreen",
    color = "black"
```

```
  ) +
  labs(
    title = "Score Histogram",
    x = "Score",
    y = "Frequency"
  )
```



From the graph above, we can see that it's a relatively normal distribution that is left skewed. This would indicate that most students have relatively high ratings for their professors. Maybe they think all of them are hot?

End of your answer

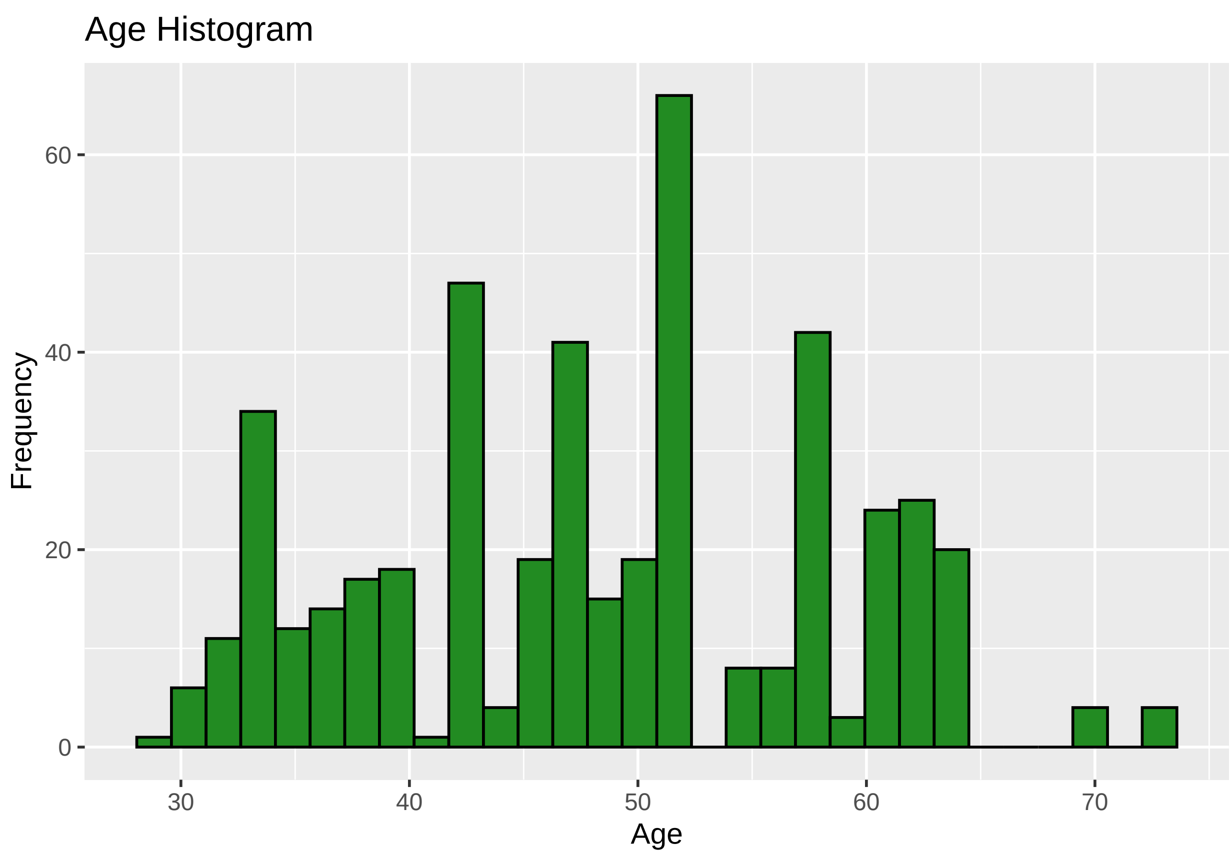
3. Excluding `score`, select two other variables and describe their relationship with each other using an appropriate visualization.

Insert your answer here

We gotta start with age:

```
ggplot(
  evals,
  aes(
    x = age
  )
) +
```

```r
  geom_histogram(
    fill = "forestgreen",
    color = "black"
  ) +
  labs(
    title = "Age Histogram",
    x = "Age",
    y = "Frequency"
  )
```

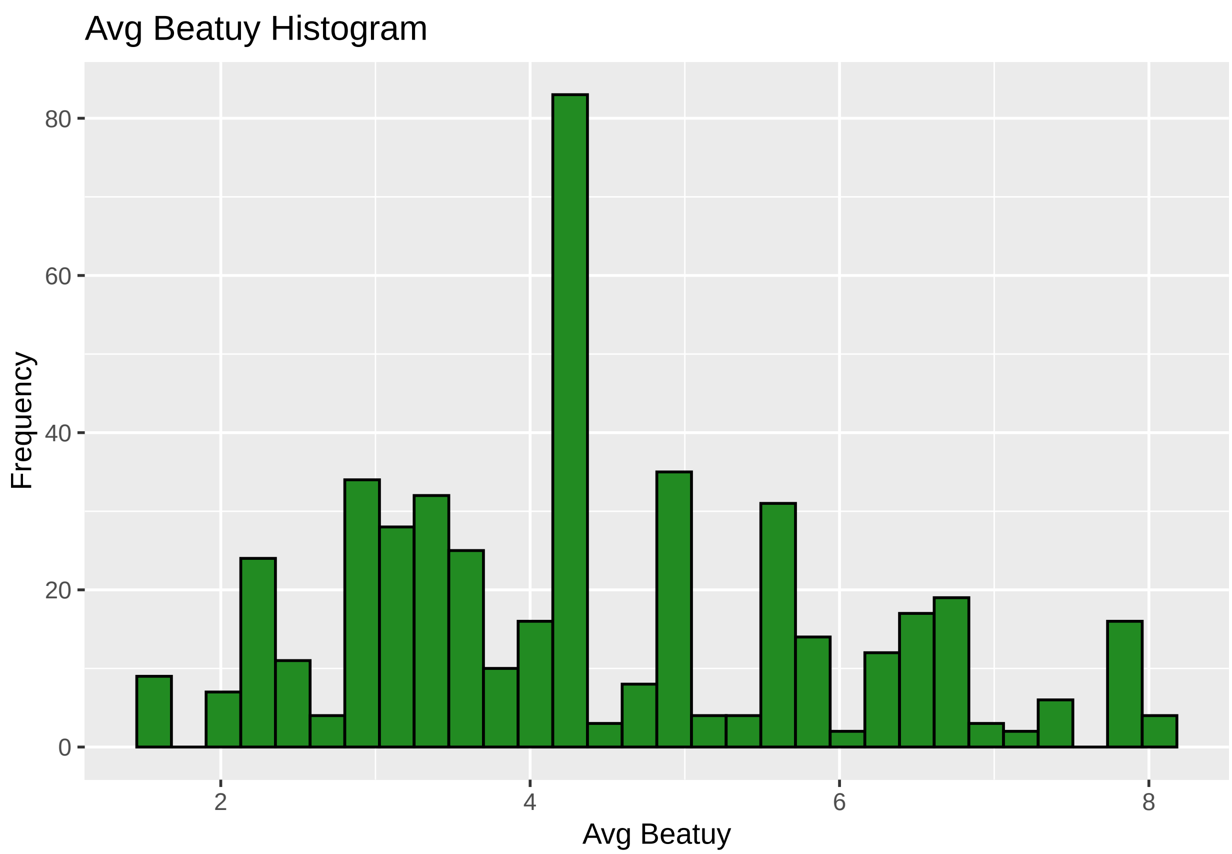


The distirbution of age doesn't seem to be normal at all.

Looking at the average beauty rating:

```r
ggplot(
  evals,
  aes(
    x = bty_avg
  )
) +
  geom_histogram(
    fill = "forestgreen",
    color = "black"
  ) +
  labs(
    title = "Avg Beatuy Histogram",
```

```
    x = "Avg Beatuy",
    y = "Frequency"
  )
```



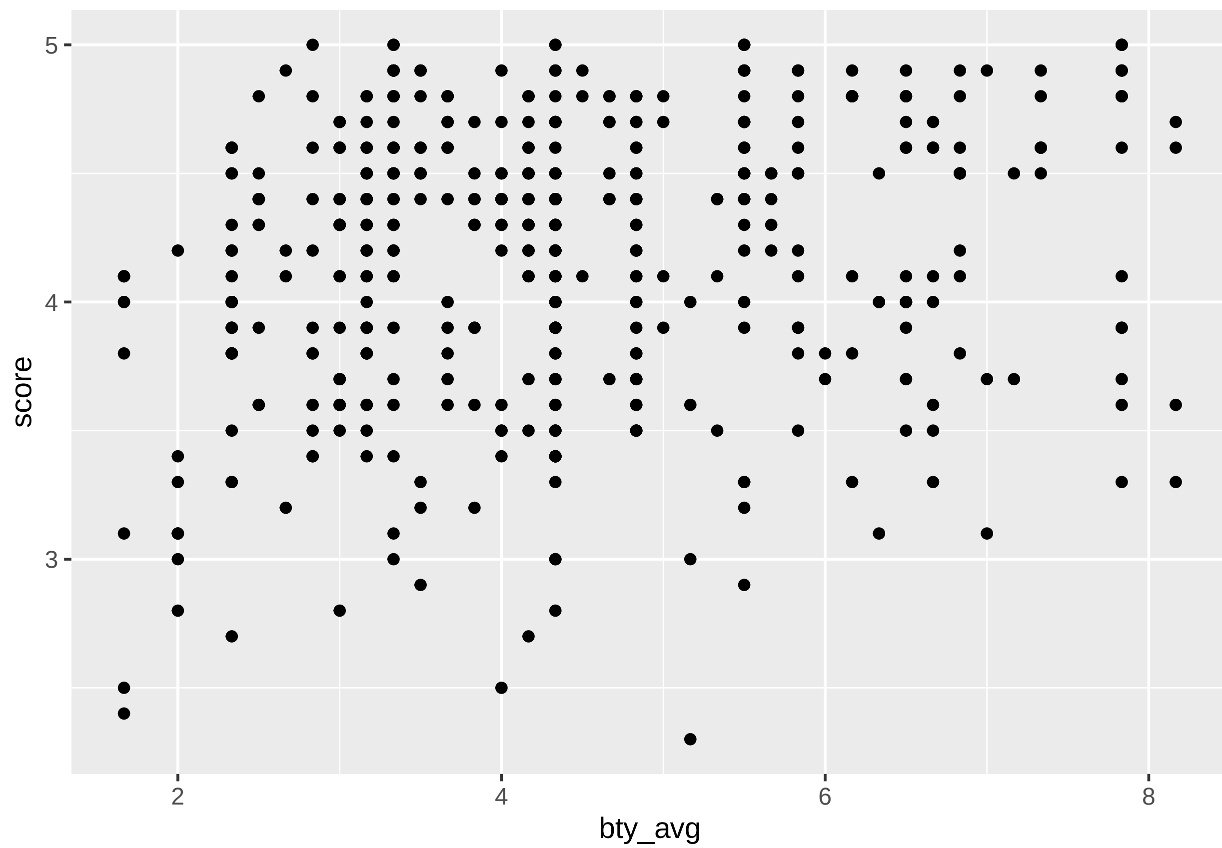
Although theres eems to be a spike at around 4, this data also doesn't seem to be very normal as well.

End of your answer

Simple linear regression

The fundamental phenomenon suggested by the study is that better looking teachers are evaluated more favorably. Let's create a scatterplot to see if this appears to be the case:

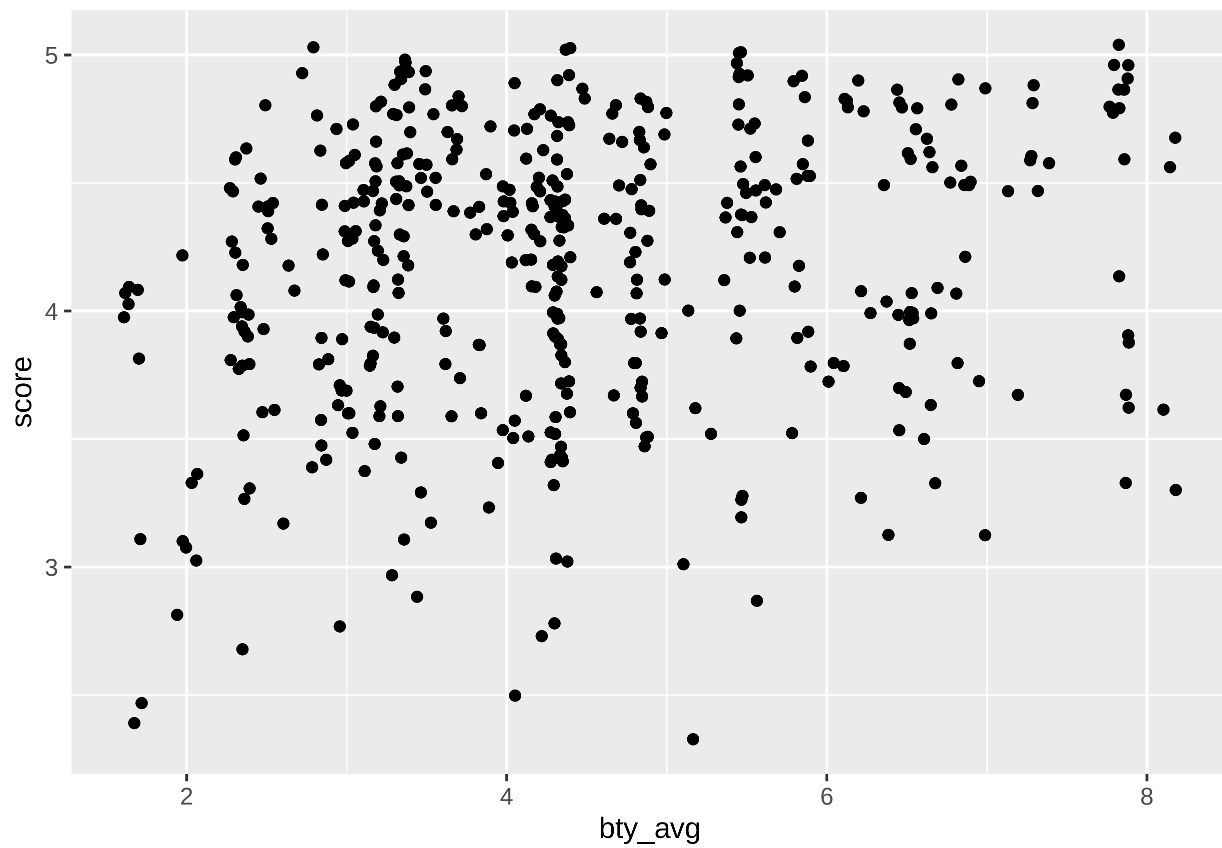
```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_point()
```



Before you draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

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

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



Insert your answer here

There were scores which overlapped which made the frequency of data points misleading.

End of your answer

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?

Insert your answer here

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

summary(m_bty)
```

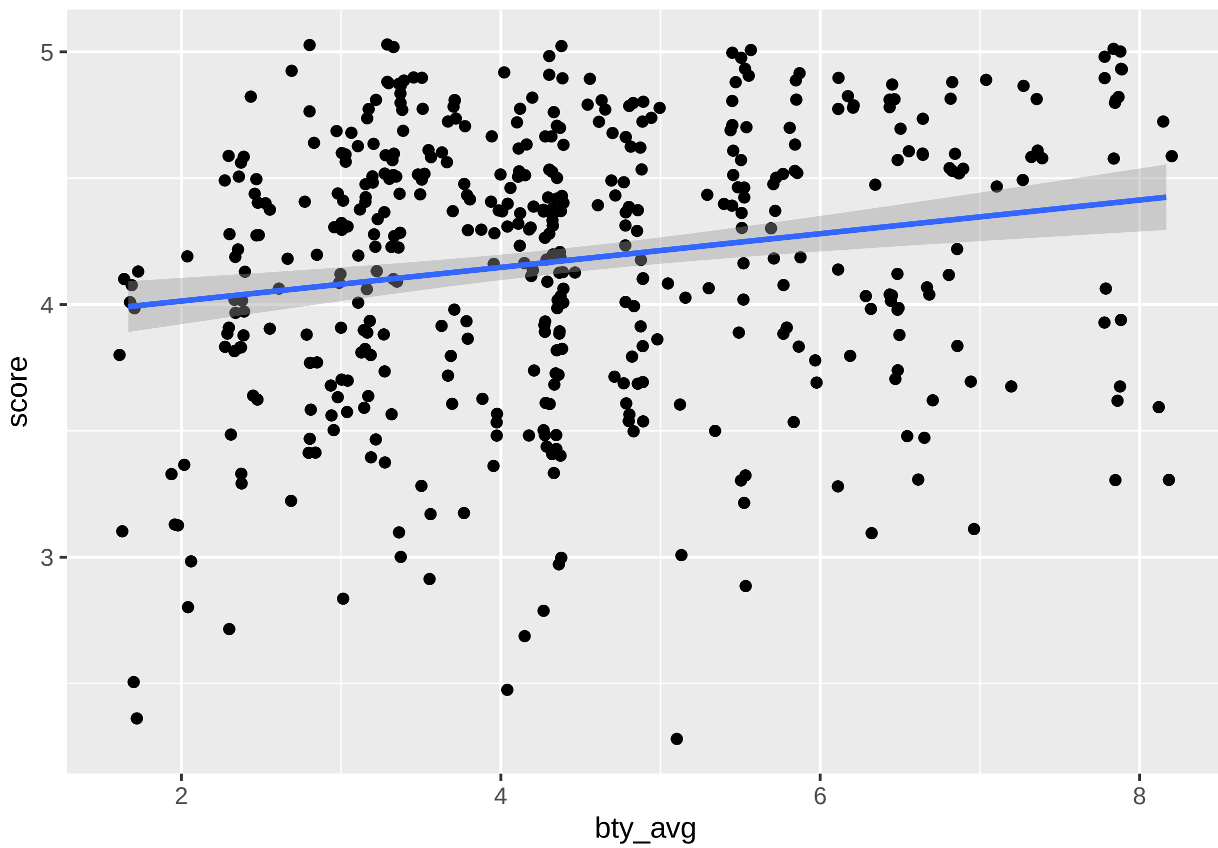
```
##
## Call:
## lm(formula = score ~ bty_avg, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -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
```

From the model above, we see that he intercept is 3.88034 and the coefficient for `bty_avg` is 0.06664. This would indicate that is a positive impact that beauty has but it's a fairly weak one considering that the beauty score ranges from 0 to 10 and the maximum impact beauty can provide is 0.6664. Looking at the adjusted R^2 of 0.032 (which is very close to 0). An R^2 of this value indicates that here is almost no correlation between `score` and `bty_avg`. **End of your answer**

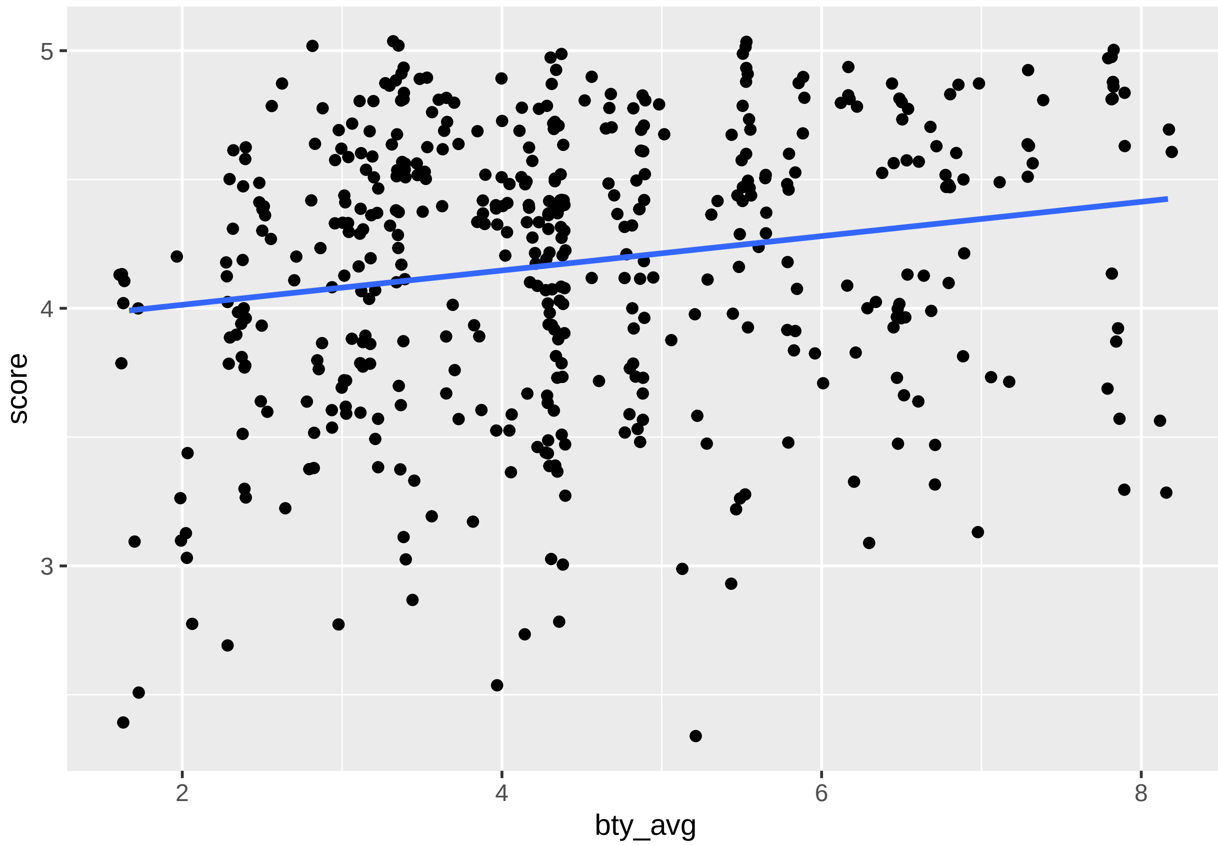
Add the line of the bet fit model to your plot using the following:

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



The blue line is the model. The shaded gray area around the line tells you about the variability you might expect in your predictions. To turn that off, use `se = FALSE`.

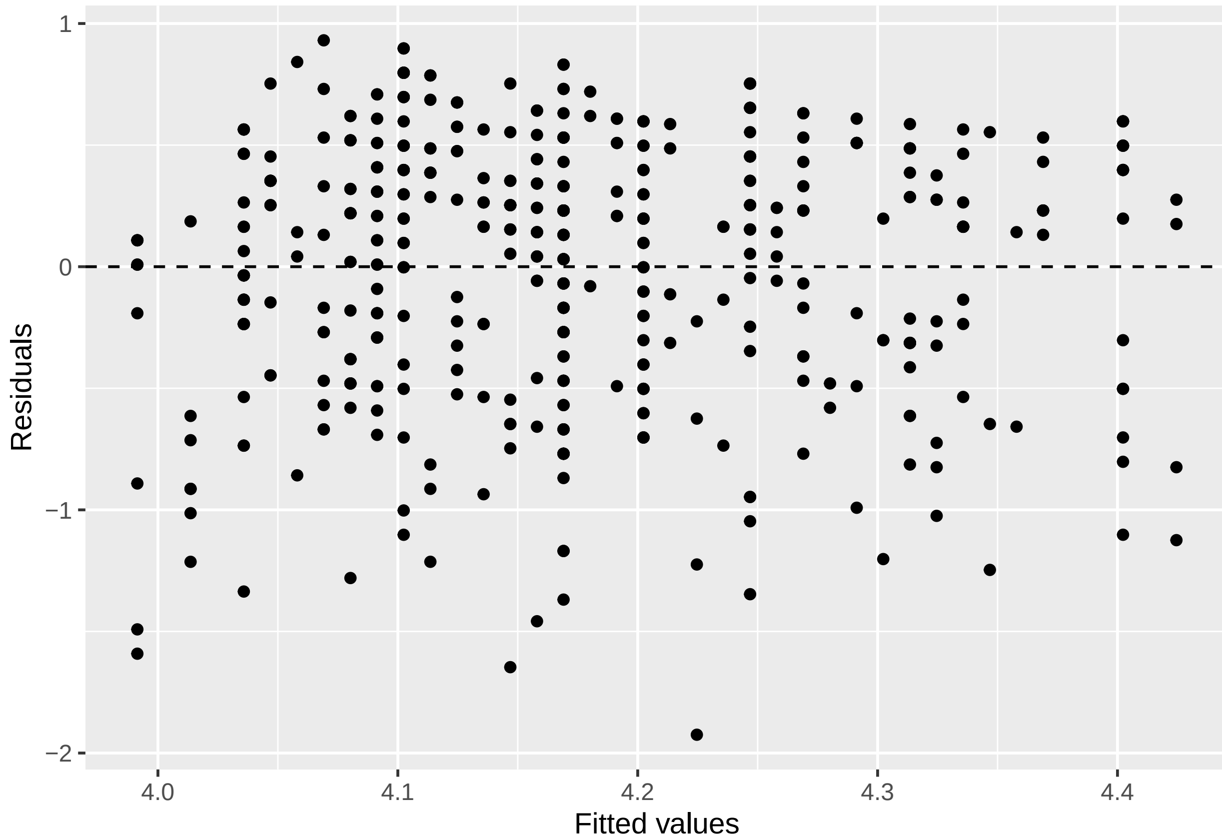
```r
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE)
```



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).

Insert your answer here

```r
ggplot(data = m_bty, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



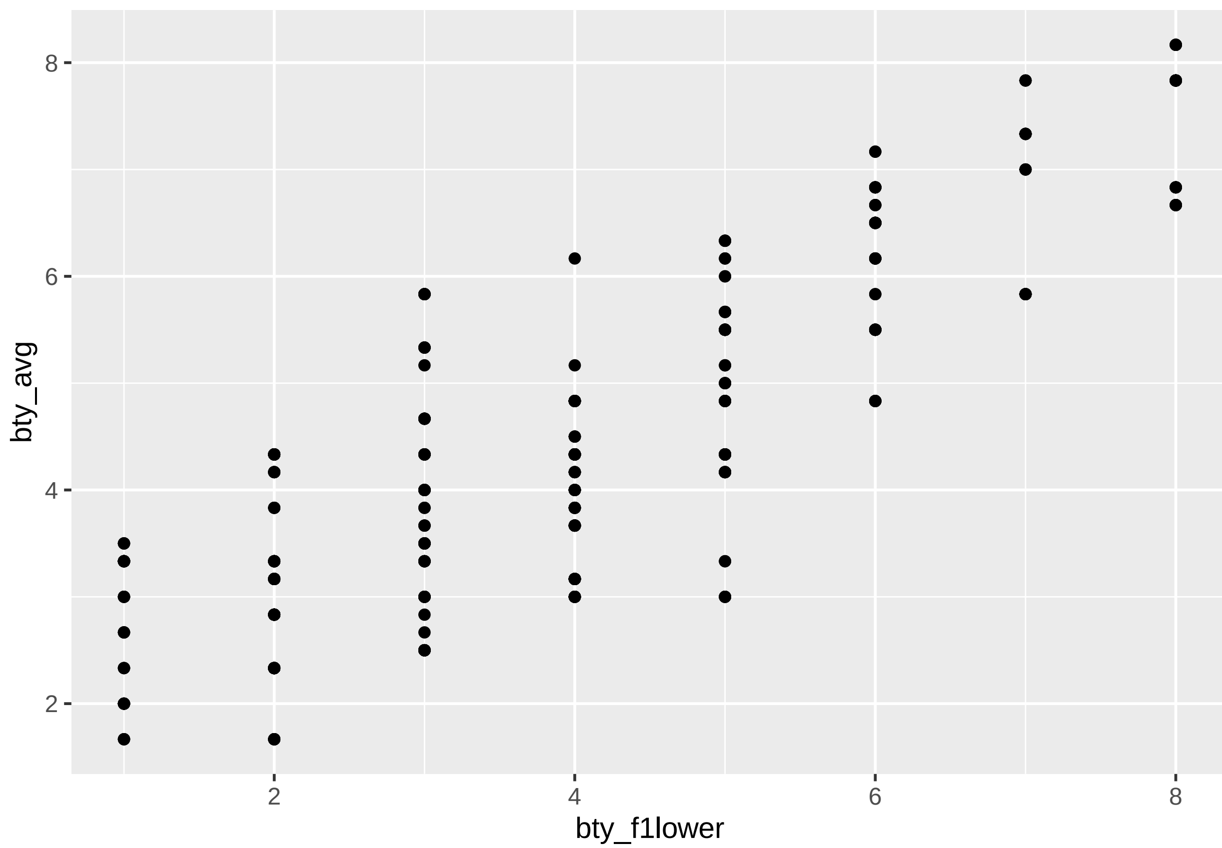
Looking at this plot, we can see that the residuals are concentraded above 0 and are spread out greater below 0. This suggests that the it isn't fully linear and is left skewed as we've mentioned before.

End of your answer

Multiple linear regression

The data set contains several variables on the beauty score of the professor: individual ratings from each of the six students who were asked to score the physical appearance of the professors and the average of these six scores. Let's take a look at the relationship between one of these scores and the average beauty score.

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

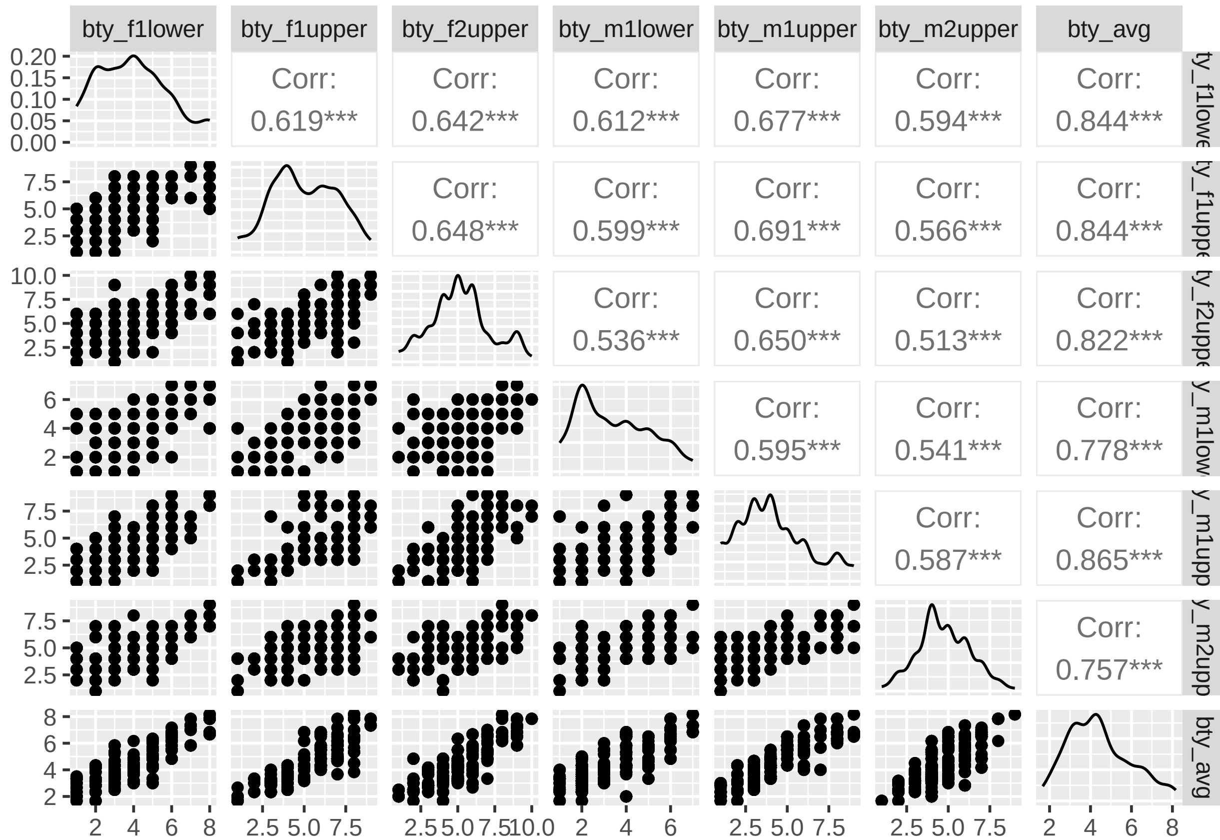



```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

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

As expected, the relationship is quite strong—after all, the average score is calculated using the individual scores. You can actually look at the relationships between all beauty variables (columns 13 through 19) using the following command:

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



These variables are collinear (correlated), and adding more than one of these variables to the model would not add much value to the model. In this application and with these highly-correlated predictors, it is reasonable to use the average beauty score as the single representative of these variables.

In order to see if beauty is still a significant predictor of professor score after you've accounted for the professor's gender, you can add the gender term into the model.

```
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
```

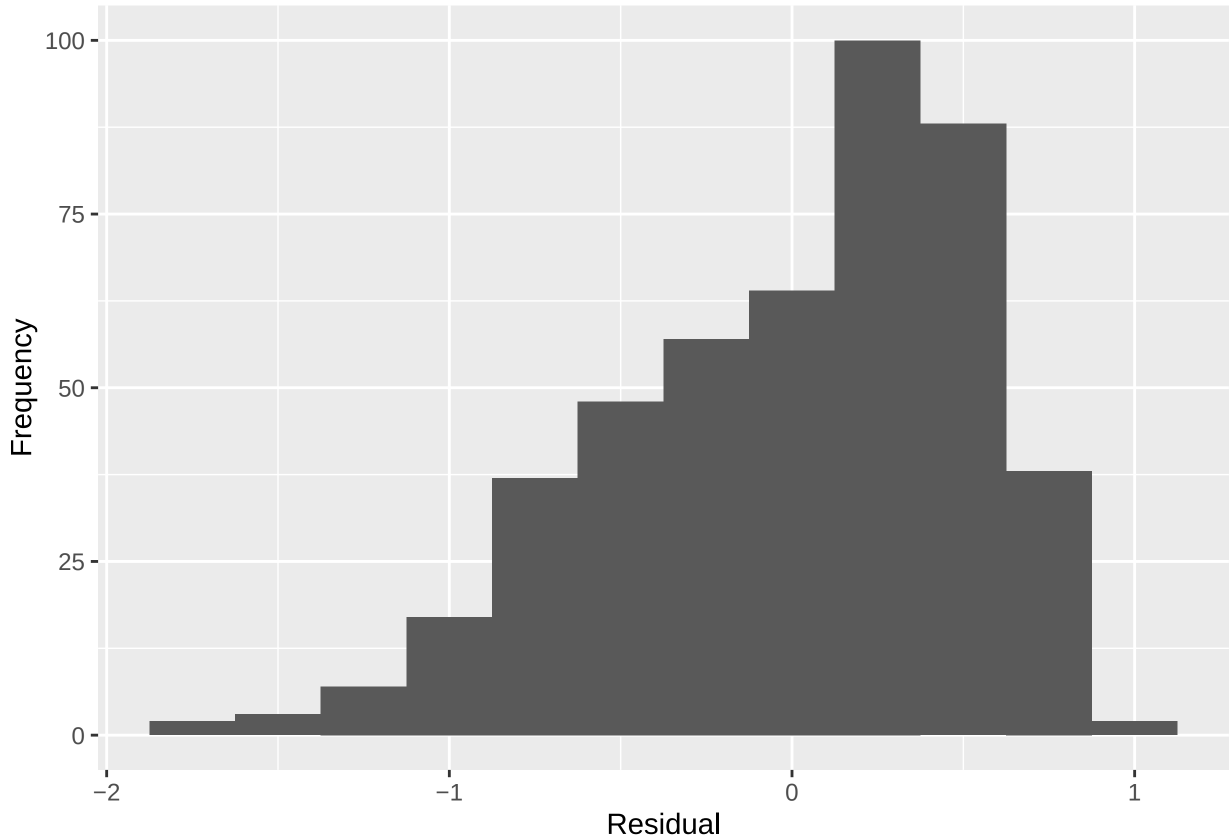
7. P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

Insert your answer here

From the openstats book, the conditions for regression are:

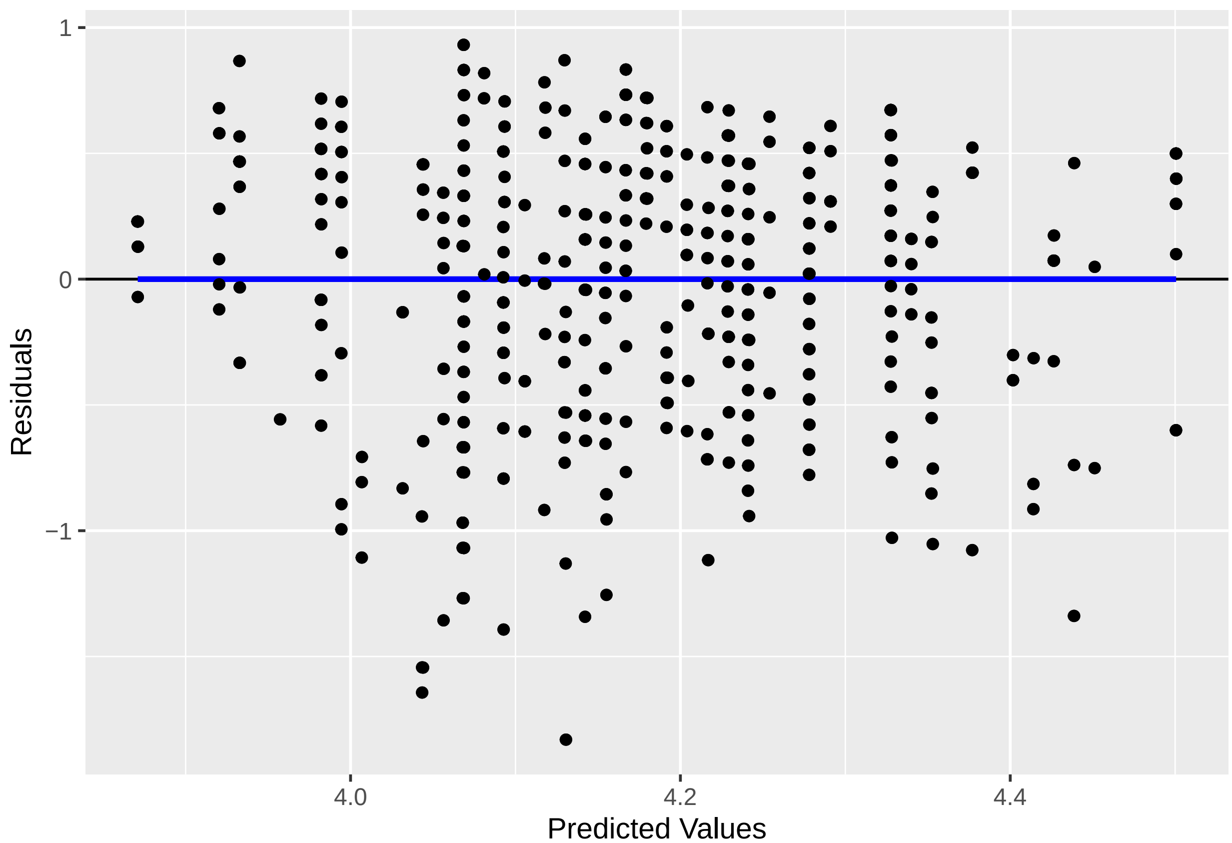
1. The residuals of the model are nearly normal (less important for larger data sets),
2. the variability of the residuals is nearly constant,
3. the residuals are independent, and
4. each variable is linearly related to the outcome.

```
ggplot(
  data = m_bty_gen,
  aes(
    x = .resid
  )
) +
  geom_histogram(
    binwidth = .25
  ) +
  labs(
    x = "Residual",
    y = "Frequency"
  )
```



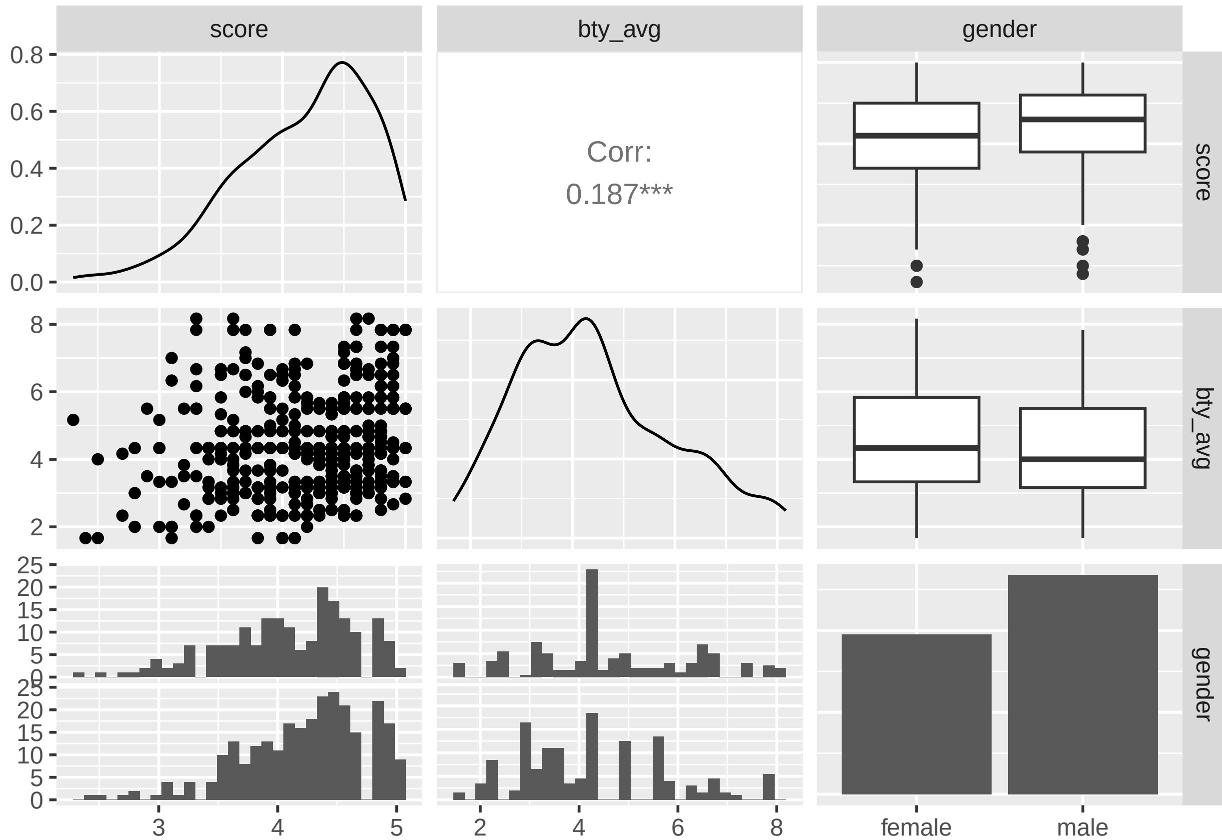
From this histogram, we can see a normal looking distribution of residuals.

```
ggplot(
  data = m_bty_gen,
  aes(
    x = .fitted,
    y = .resid
  )
) +
  geom_point() +
  geom_hline(
    yintercept = 0,
    fill = "red"
  ) +
  geom_smooth(
    method = "lm",
    se = FALSE,
    color = "blue"
  ) +
  labs(
    x = "Predicted Values",
    y = "Residuals"
  ) +
  geom_jitter()
```



From here we can see that the variability of residuals is fairly constant. We can also not discern any pattern in this plot, suggesting independence.

```
evals |>
  select(
    score,
    bty_avg,
    gender
  ) |>
  ggpairs()
```



Lastly using ggpairs(), we can see that there does seem to be a linear relationship between both gender and bty_avg, even though the correlation score is only 0.187.

End of your answer

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`?

Insert your answer here

```
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
```

```r
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
```

Looking at the coefficient of `bty_avg`, the value of the coefficient doesn't change very much so its predicive capabilities also hasn't changed.

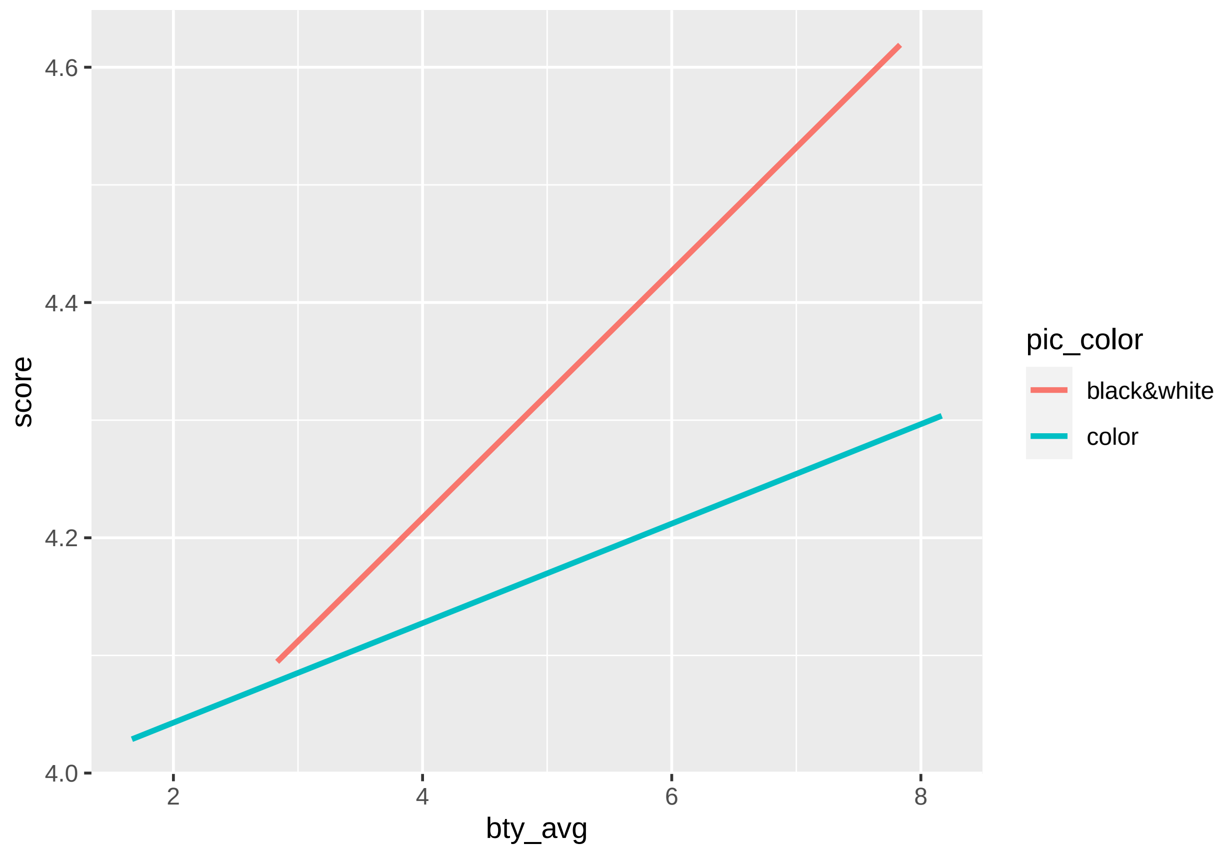
End of your answer

Note that the estimate for `gender` is now called `gendermale`. You'll see this name change whenever you introduce a categorical variable. The reason is that R recodes `gender` from having the values of `male` and `female` to being an indicator variable called `gendermale` that takes a value of 0 for female professors and a value of 1 for male professors. (Such variables are often referred to as "dummy" variables.)

As a result, for female professors, the parameter estimate is multiplied by zero, leaving the intercept and slope form familiar from simple regression.

$$\begin{aligned} \widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg \end{aligned}$$

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



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?

Insert your answer here

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

summary(m_bty_clr)
```

```
##
## 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
```

From the above summary, our equation is:

$score = intercept + 0.05548 bty\_avg + -0.16059 pic\_color$

Where `pic_color` is 0 when it's a pictured color and 1 when it is black and white.

End of your answer

The decision to call the indicator variable `gendermale` instead of `genderfemale` has no deeper meaning. R simply codes the category that comes first alphabetically as a 0. (You can change the reference level of a categorical variable, which is the level that is coded as a 0, using the`relevel()` function. Use `?relevel` to learn more.)

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`.

Insert your answer here

```r
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
```

Judging form the model summary, it seems that the model creates n-1 terms where n is the number of unique entries that the field has. We can see here that now there is two new variables:

1. `ranktenure track`
2. `ranktenured`

We can conclude that the last rank (`teaching`) is added into the intercept.

End of your answer

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 *while 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.

The search for the best model

We will start with a full model that predicts professor score based on rank, gender, ethnicity, language of the university where they got their degree, age, proportion of students that filled out evaluations, class size, course level, number of professors, number of credits, average beauty rating, outfit, and picture color.

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.

Insert your answer here

I believe any of the variables having to do with the number of students in the class would have the highest p value.

End of your answer

Let's run the model…

```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
```

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

Insert your answer here

The variable with the highest p value is cls_profs, outlining how many professors there are in a class.

End of your answer

13. Interpret the coefficient associated with the ethnicity variable.

Insert your answer here

The coefficient for `ethnicitynot minorit` is 0.1234929, meaning that minorities have a ~.12 point disadvantage compared to their non-miniority counterparts.

End of your answer

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?

Insert your answer here

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

summary(m_no_cls_profs)
```

```
##
## Call:
## lm(formula = score ~ rank + gender + ethnicity + 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
## gendermale             0.2101231  0.0516873   4.065 5.66e-05 ***
## ethnicitynot minority  0.1274458  0.0772887   1.649 0.099856 .
## 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
```

There doesn't seem to be a siginficant change in the coefficients of other variables, suggesting that there was no variable that correlated with `cls_profs`.

End of your answer

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.

Insert your answer here

Just selecting for when $p < 0.05$:

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

summary(m_p_back)
```

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

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81919 -0.32035  0.09272  0.38526  0.88213
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.967255   0.215824  18.382  < 2e-16 ***
## gendermale             0.221457   0.049937   4.435 1.16e-05 ***
## languagenon-english   -0.281933   0.098341  -2.867  0.00434 **
## age                   -0.005877   0.002622  -2.241  0.02551 *
## cls_perc_eval          0.004295   0.001432   2.999  0.00286 **
## cls_creditsone credit  0.444392   0.100910   4.404 1.33e-05 ***
## bty_avg                0.048679   0.016974   2.868  0.00432 **
## pic_colorcolor        -0.216556   0.066625  -3.250  0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5014 on 455 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1502
## F-statistic: 12.67 on 7 and 455 DF,  p-value: 6.996e-15
```

We're going to create a function that will do this as we've done this quite a few times:

```
get_models_written_equation <- function(model) {
  # Extract the coefficients from the model
  coefficients <- coef(model)

  # Get the names of the coefficients
  variable_names <- names(coefficients)

  # Initialize an empty string to build the equation
  equation <- ""

  # Loop over the coefficients and format the equation
  for (i in seq_along(coefficients)) {
    coef_value <- coefficients[i]
    variable_name <- variable_names[i]

    sign <- if (coef_value >= 0 && i > 1) "+" else ""

    if (variable_name == "(Intercept)") {
      # Handle the intercept separately
      equation <- paste0(equation, round(coef_value, 3))
    } else {
      equation <- paste0(
        equation,
        " ",
        sign,
        " ",
        round(coef_value, 3),
        "*",
        variable_name
      )
    }
```

```
  }

  response_variable <- attr(terms(model), "variables")[2]
  full_equation <- paste(response_variable, "~", equation)

  return(full_equation)
}


get_models_written_equation(m_p_back)
```

```
## [1] "score ~ 3.967 + 0.221*gendermale  -0.282*languagenon-english  -0.006*age + 0.004*cls_perc_eval +
```

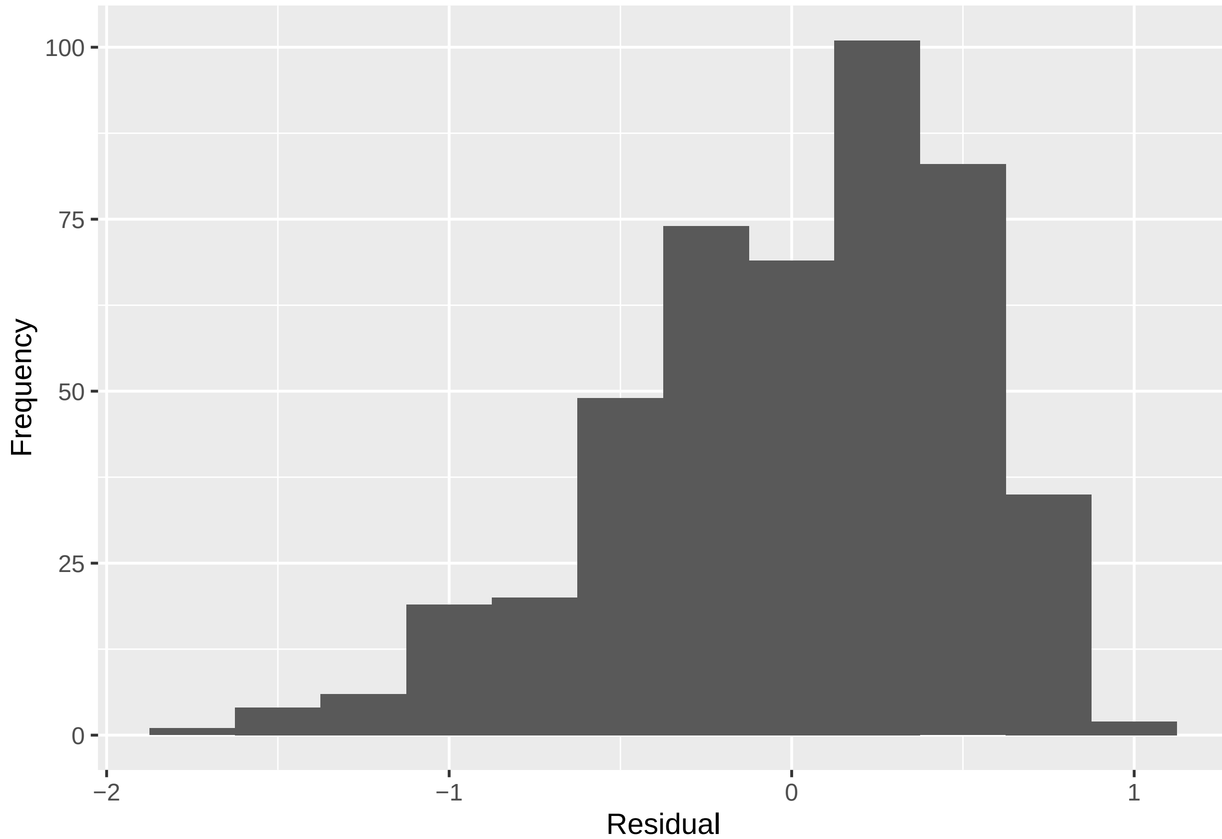
The full equation is score ~ 3.967 + 0.221*gendermale -0.282*languagenon-english -0.006*age + 0.004*cls_perc_eval + 0.444*cls_creditsone credit + 0.049*bty_avg -0.217*pic_colorcolor

End of your answer

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

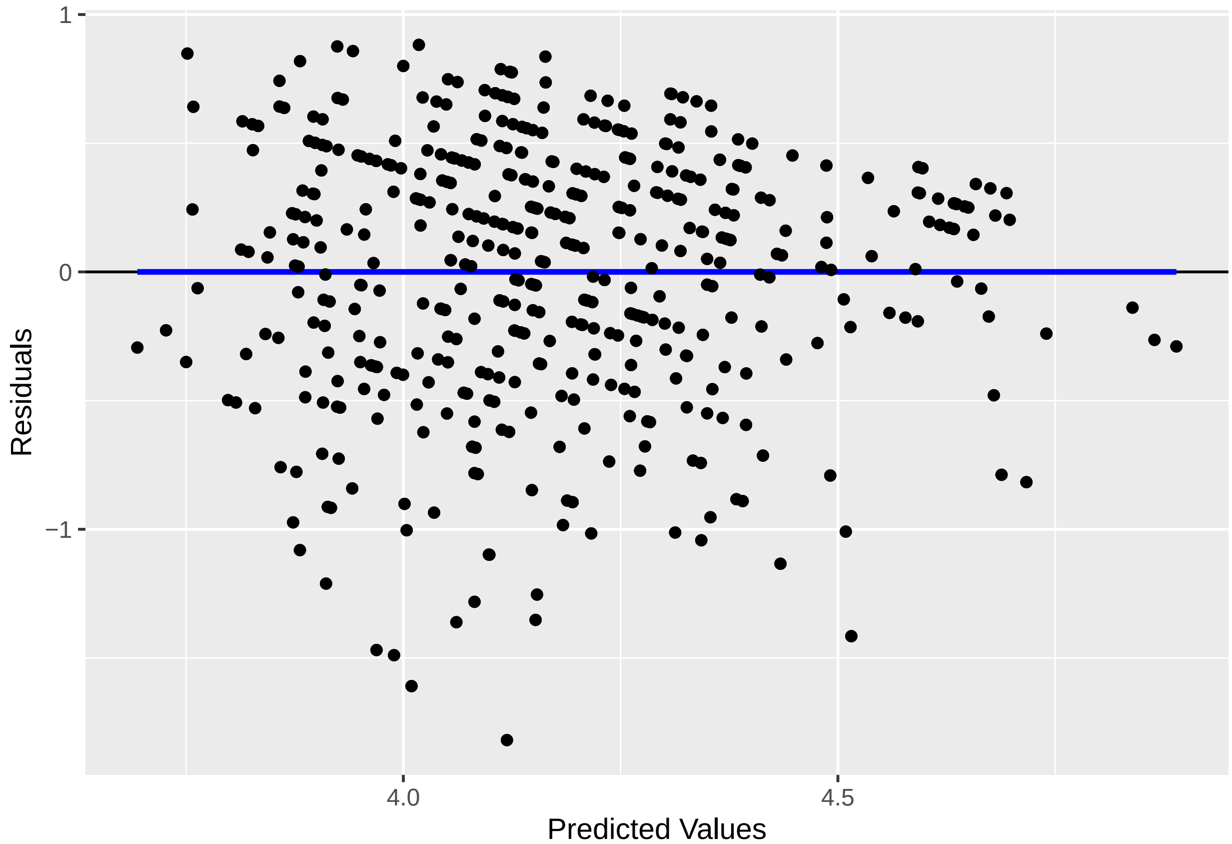
Insert your answer here

```
ggplot(
  data = m_p_back,
  aes(
    x = .resid
  )
) +
  geom_histogram(
    binwidth = .25
  ) +
  labs(
    x = "Residual",
    y = "Frequency"
  )
```



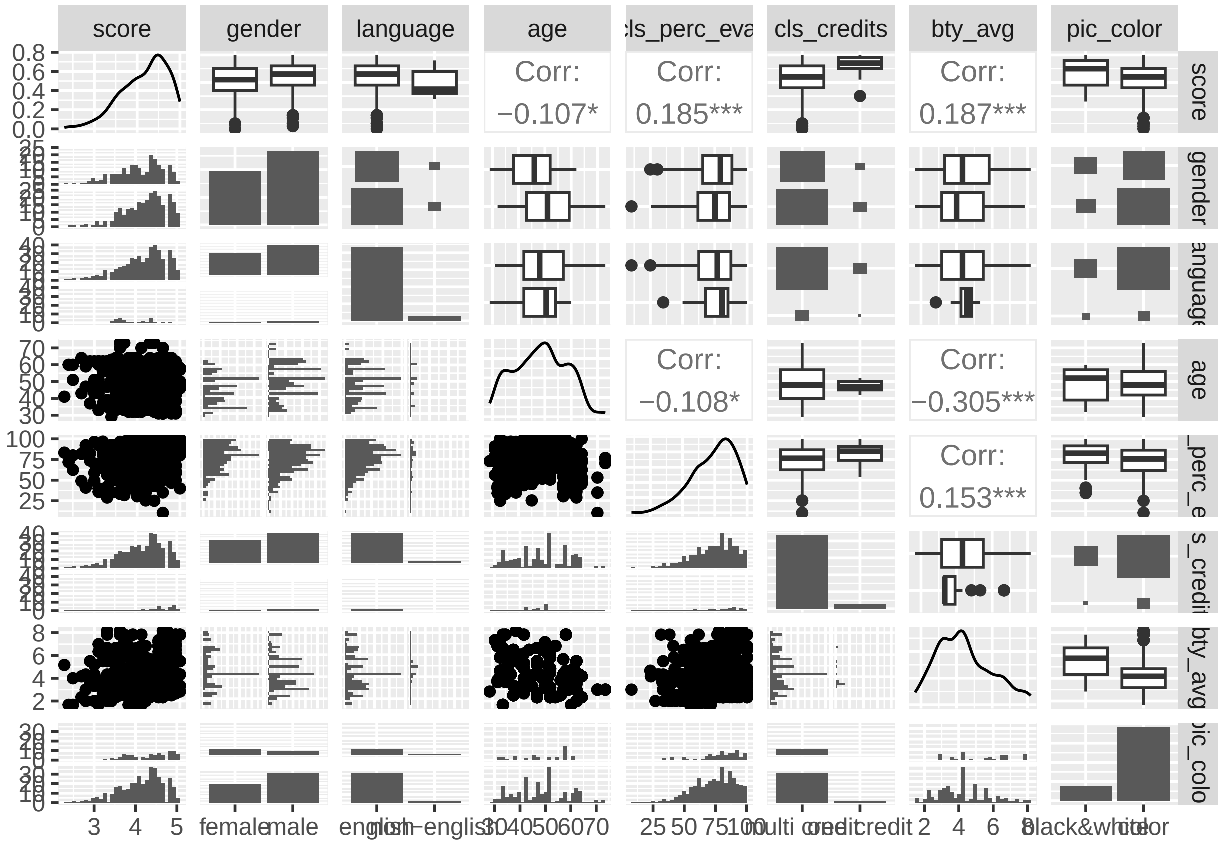
From this histogram, we can see a normal looking distribution of residuals.

```
ggplot(
  data = m_p_back,
  aes(
    x = .fitted,
    y = .resid
  )
) +
  geom_point() +
  geom_hline(
    yintercept = 0,
    fill = "red"
  ) +
  geom_smooth(
    method = "lm",
    se = FALSE,
    color = "blue"
  ) +
  labs(
    x = "Predicted Values",
    y = "Residuals"
  ) +
  geom_jitter()
```



From here we can see that the variability of residuals is fairly constant. We can also not discern any pattern in this plot, suggesting independence.

```
evals |>
  select(
    score,
    gender,
    language,
    age,
    cls_perc_eval,
    cls_credits,
    bty_avg,
    pic_color
  ) |>
  ggpairs()
```



Lastly using ggpairs(), we can see that there does seem to be alinear relationship between each variable and score.

End of your answer

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?

Insert your answer here

Yes, because it's likely that a single professor may be represented in different classes. Therefore, they may show up as a different observation for each class when they aren't.

End of your answer

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.

Insert your answer here

According to this model, they would have these attributes:

- Male
- English Educated
- Younger

- High percentage of class who provided an evaluation
- One Credit
- Rated well for Beauty
- Black and White profile picture

End of your answer

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

Insert your answer here

No I would not, simply because this sample at UT Austin is not a perfect sample for the school and is not represented of all univeristies and professors. Additionally, this is morally questionable.

End of your answer