Multiple linear regression

```
library(tidyverse)
library(openintro)
library(GGally)
evals <- evals
```

Exercise 1

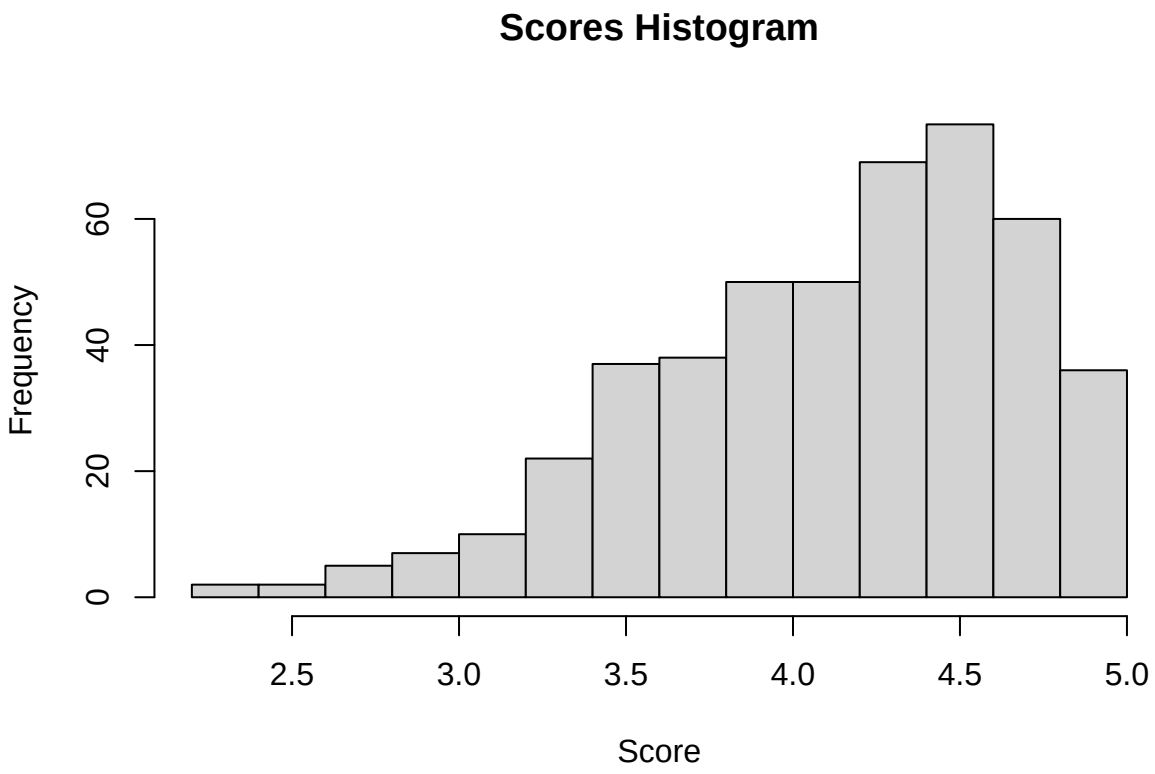
Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

> I would say this is an observational study given there's no control or test group to compare each other against; we're looking solelly at observations found in the data rather than experimenting with multiple groups. The question itself is hard to answer as phrased since even if there's a correlation between the two variables, that does not mean causation. An easier/more manageable question to ask is whether there is a relationship between a professor's attractiveness and course evaluation scores.

Exercise 2

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

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?

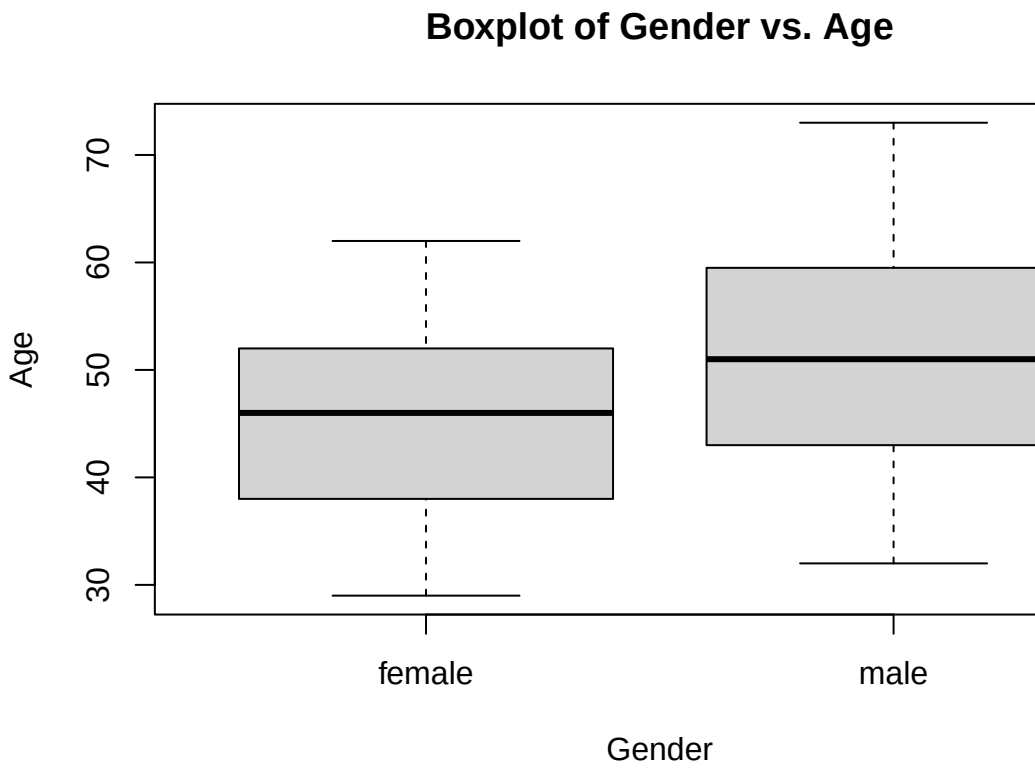


The distribution is skewed to the left which means that students rated courses more positively. I would expect a more normal distribution, however this means that the samples taken are likely due to students either not being super critical or the professors are pretty fair and have a good reputation.

Exercise 3

```
boxplot(evals$age ~ evals$gender, main = "Boxplot of Gender vs. Age", ylab = "Age", xlab = "Gender")
```

Excluding `score`, select two other variables and describe their relationship with each other using



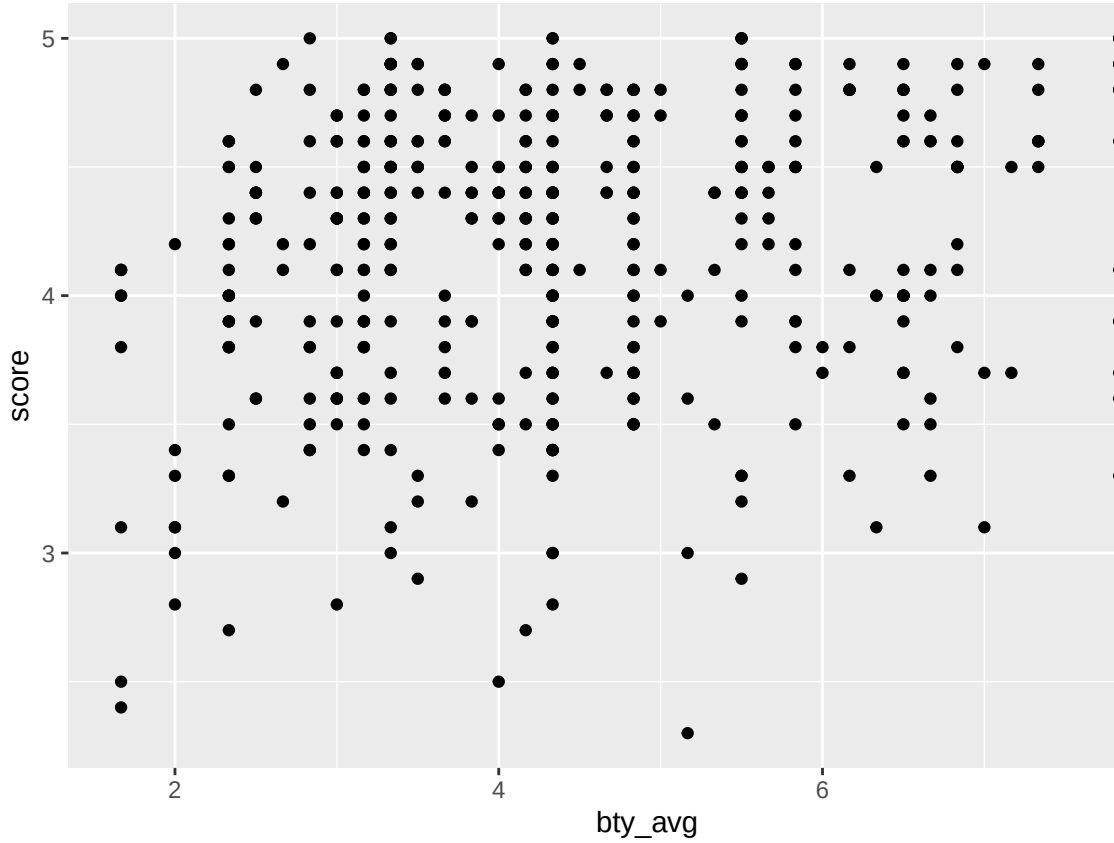
an appropriate visualization.

It seems like male professors tend to be older than female ones with the distribution of male professors being much higher, ~45-60, with females being more towards ~39-52.

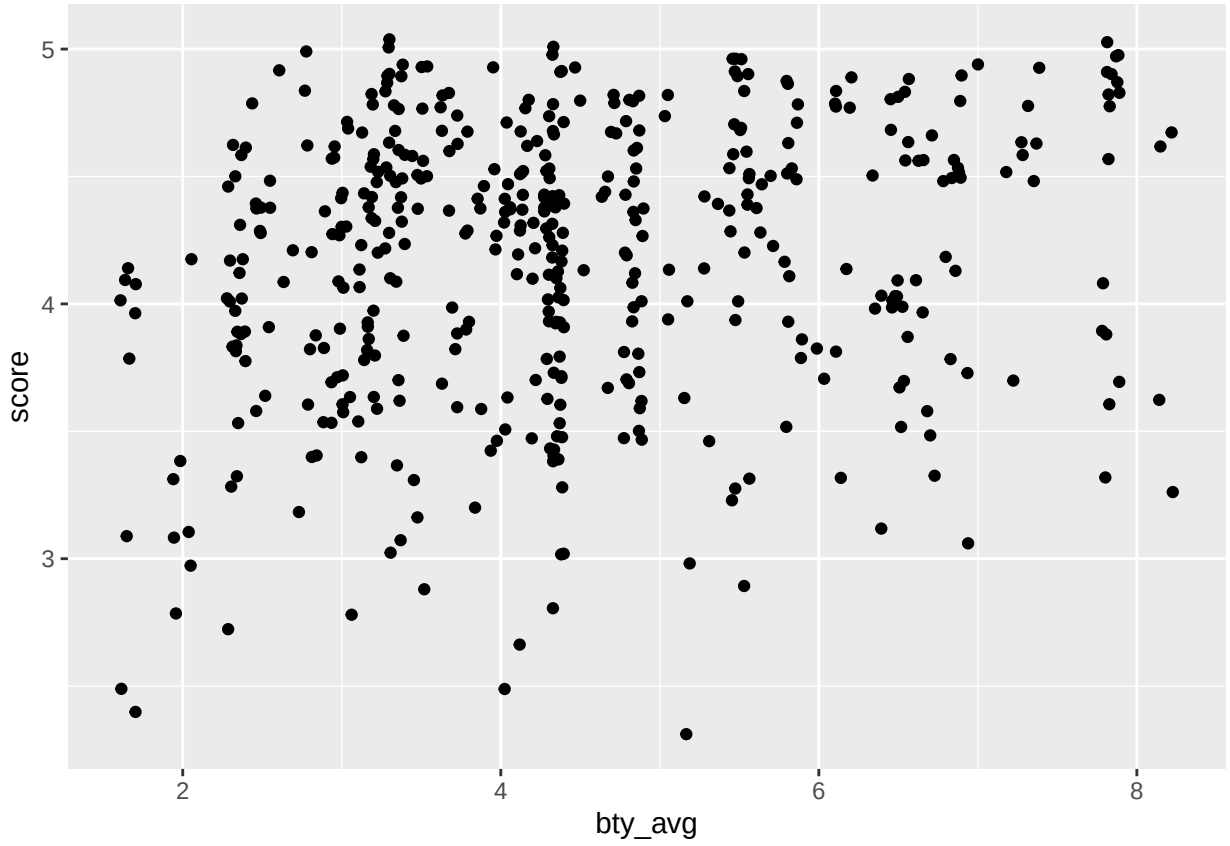
Exercise 4

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

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()
```

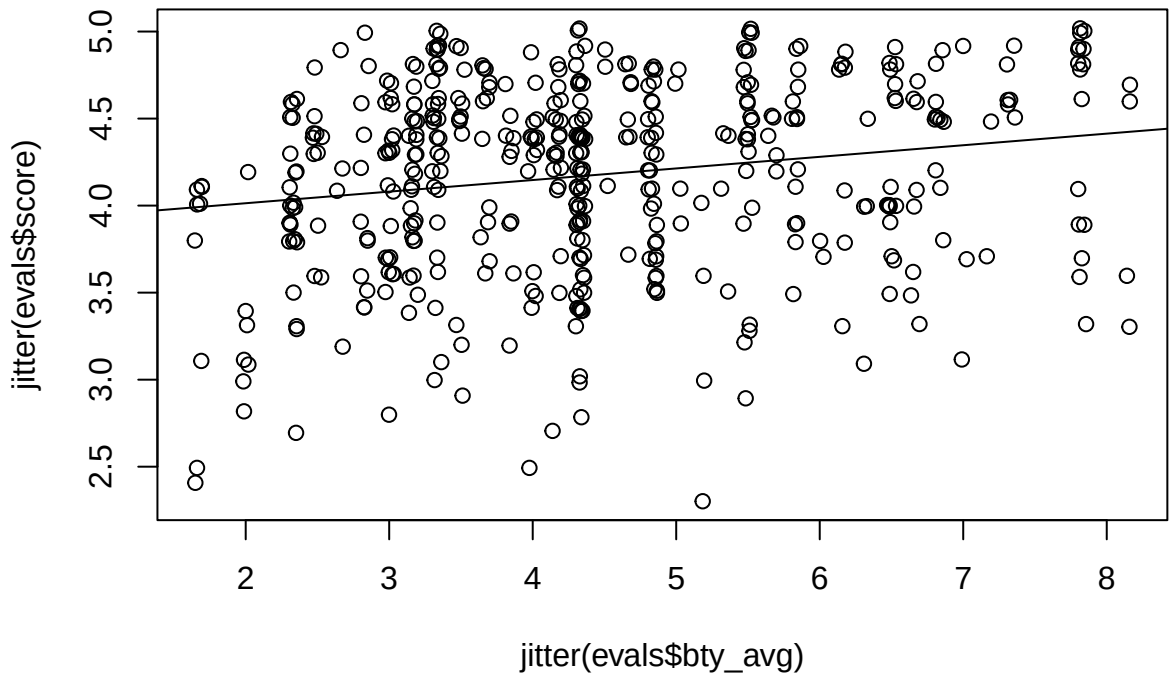


The first one is misleading as it makes it seem like there are less points than there actual are since it overlaps its points if they're on the same coordinate, but `geom_jitter` adds some noise and gives more insight into how many pieces of data there actually are.

Exercise 5

```
m_bty <- lm(evals$score ~ evals$bty_avg)
plot(jitter(evals$score) ~ jitter(evals$bty_avg))
abline(m_bty)
```

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?




```
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
```

$$\hat{y} = 3.88034 + 0.06664 \times bty\_avg$$

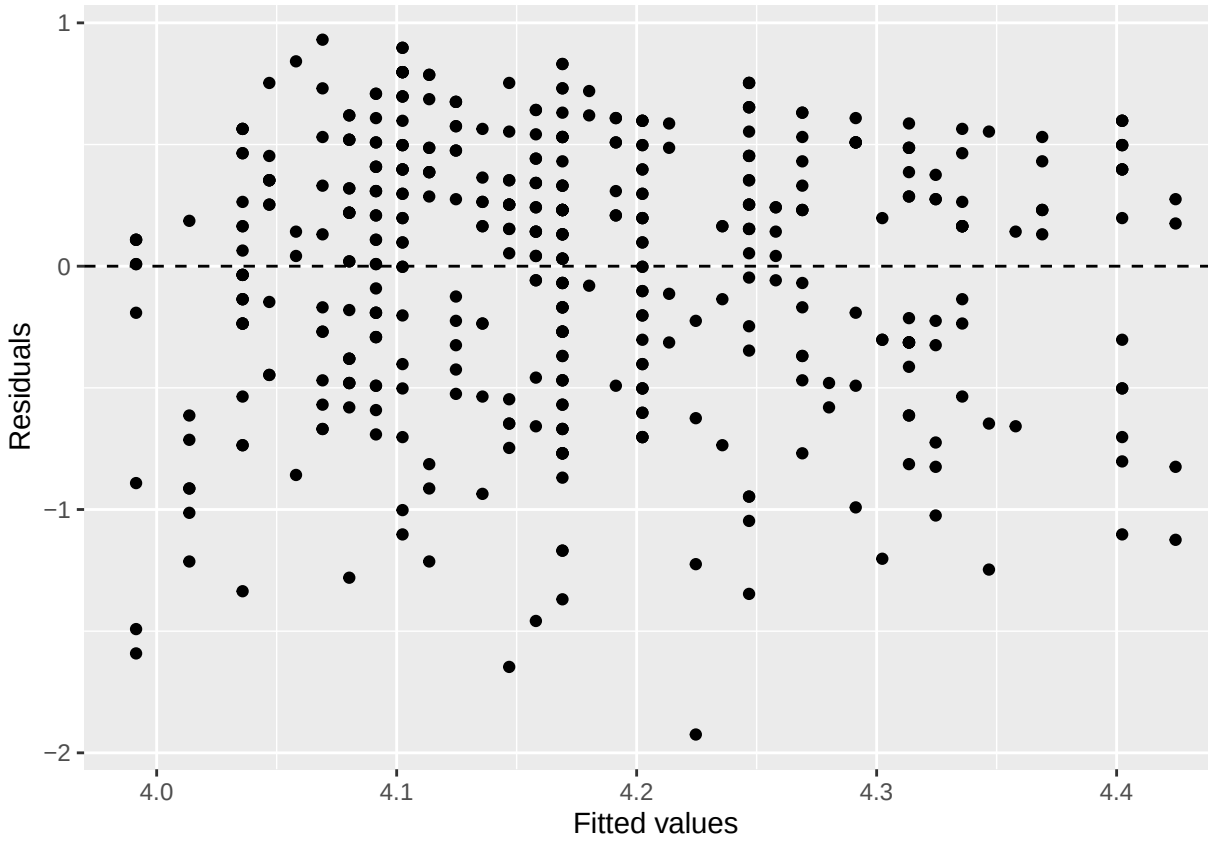
This formula can be interpreted that for every increase in 1 for `bty_avg`, `score` increases by

0.06664. The p-value is statistically significant as it's quite small (0.0000508), but given the impact is so small (0.067), it's not a *practically* significant predictor.

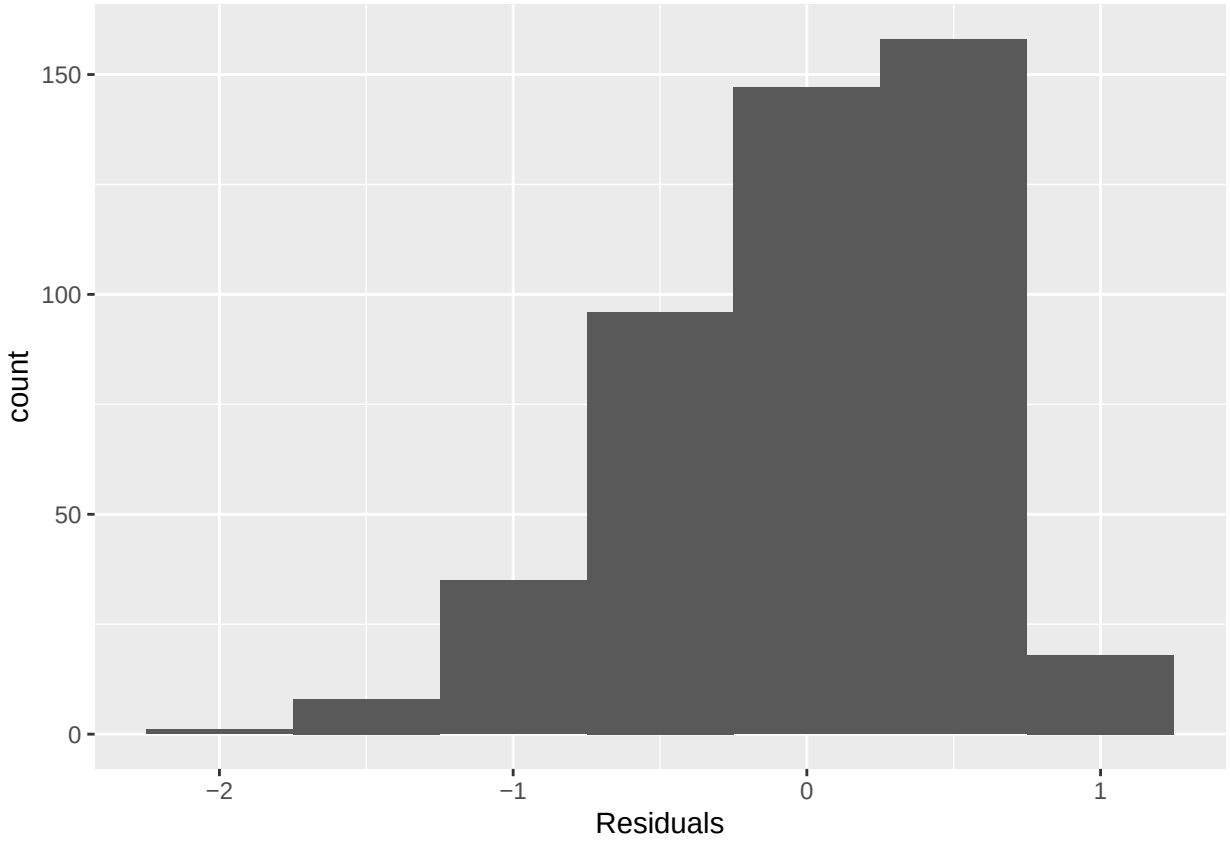
Exercise 6

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

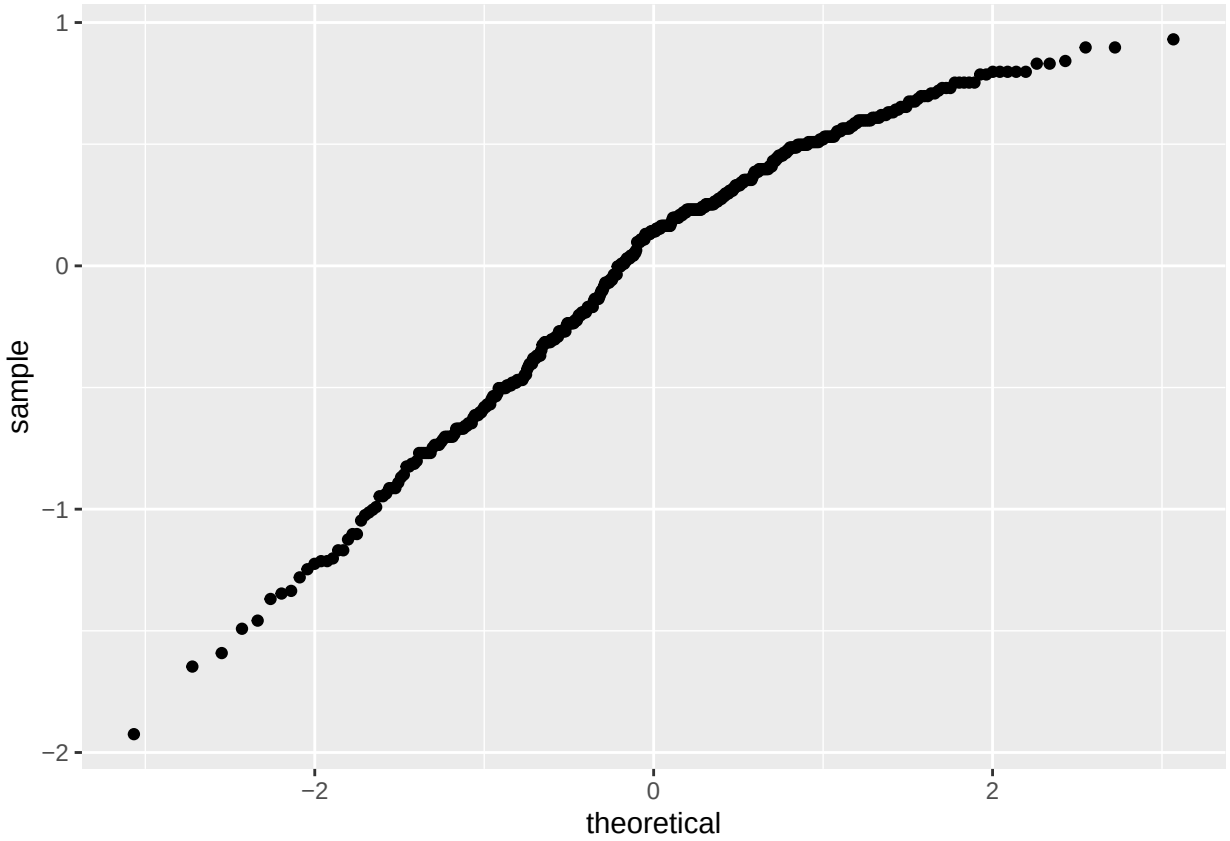
Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).




```
ggplot(data = m_bty, aes(x = .resid)) +
  geom_histogram(binwidth = 0.5) +
  xlab("Residuals")
```




```
ggplot(data = m_bty, aes(sample = .resid)) +
  stat_qq()
```



- Independence: We don't have too much information on how the sample was taken, so we'll assume independence for this.
- Linear Relationship: Visually, the data looks linear with a slightly positive relationship.
- Constant Variance: As shown in the Fitted Values Residuals plot, there does seem to be constant variance.
- Normality of Residuals: The histogram shows a slightly left skew and the qq plot shows a short, right tail; as short tail is relatively harmless, we can move forward and assume normality.

Exercise 7

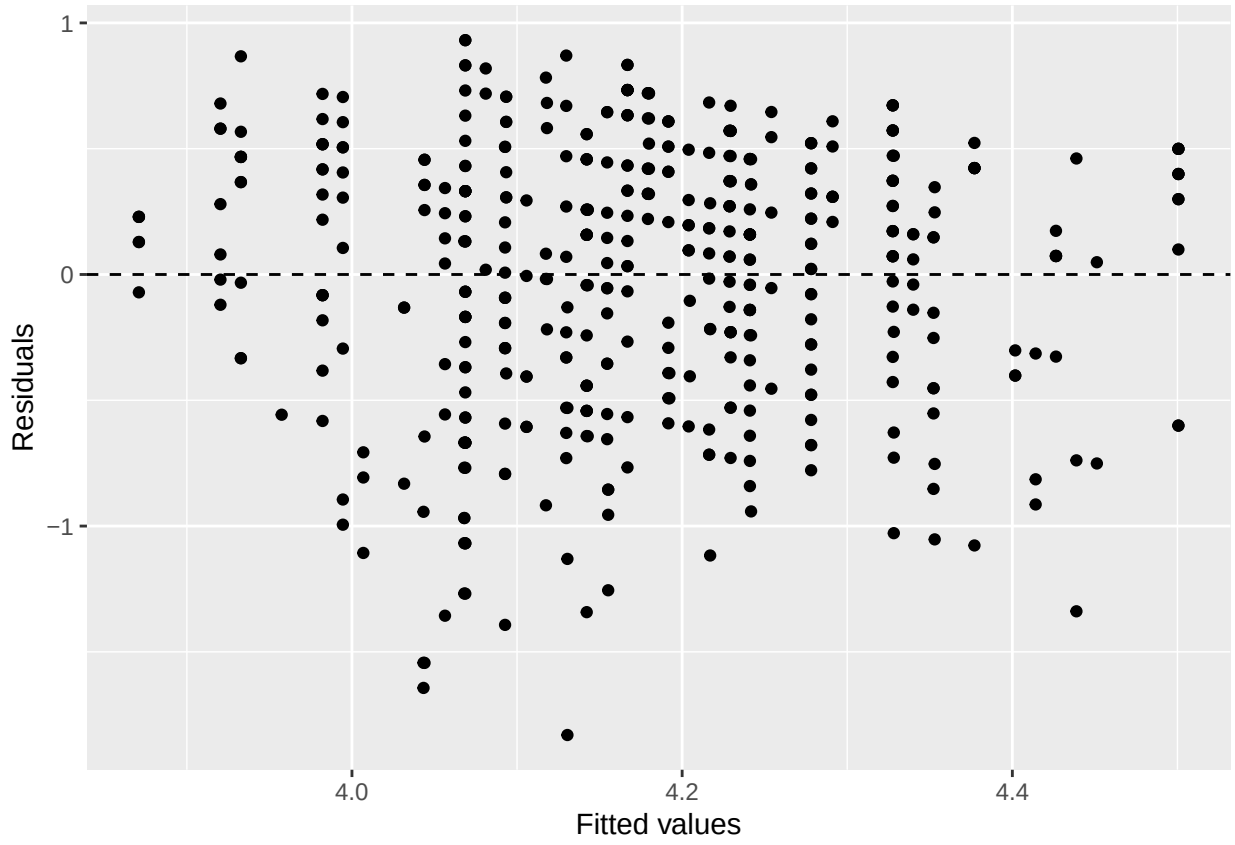
```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

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.

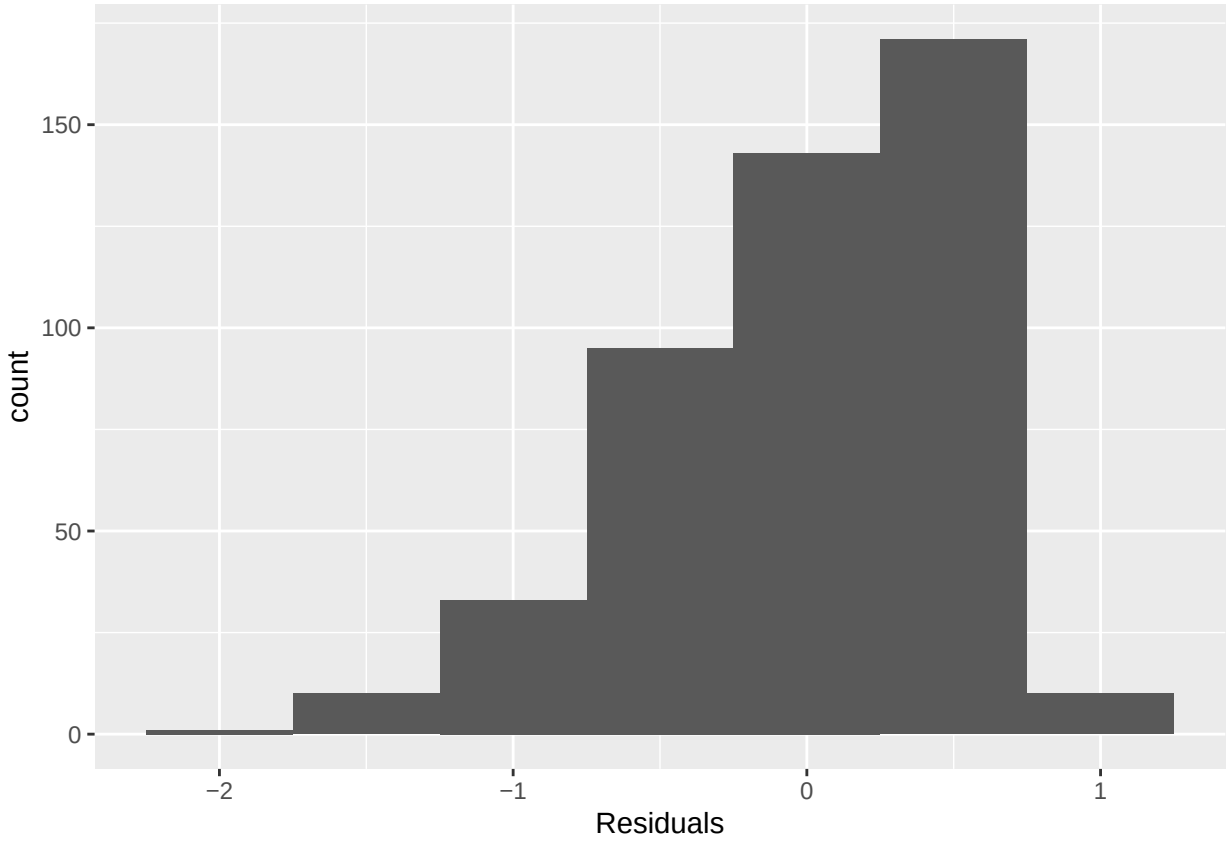
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -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
```

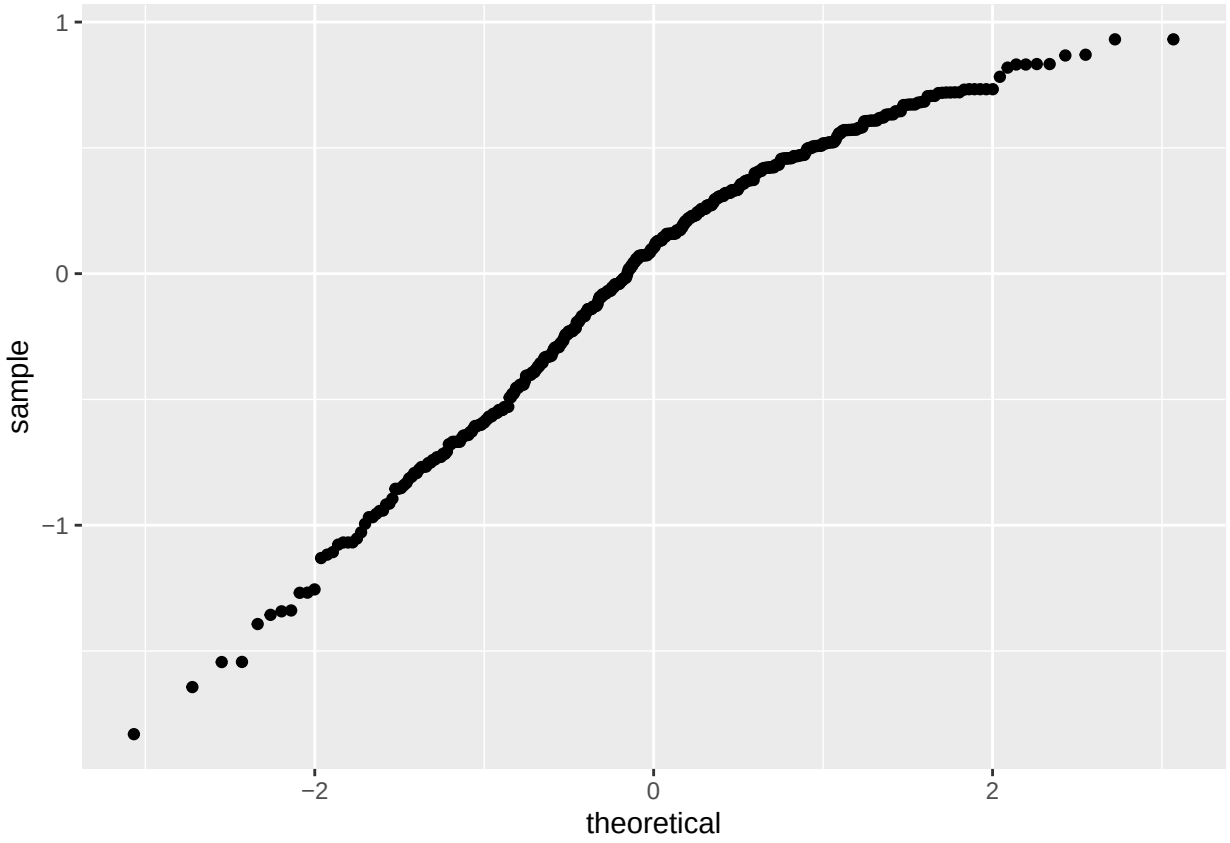
```
ggplot(data = m_bty_gen, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = m_bty_gen, aes(x = .resid)) +
  geom_histogram(binwidth = 0.5) +
  xlab("Residuals")
```




```
ggplot(data = m_bty_gen, aes(sample = .resid)) +
  stat_qq()
```



The data looks super similar to the previous model using only `bty_avg` so I would say that the conditions for this model are reasonable for the same reasons as exercise 6.

Exercise 8

```
summary(m_bty_gen)
```

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

```
##
## 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 ***
```

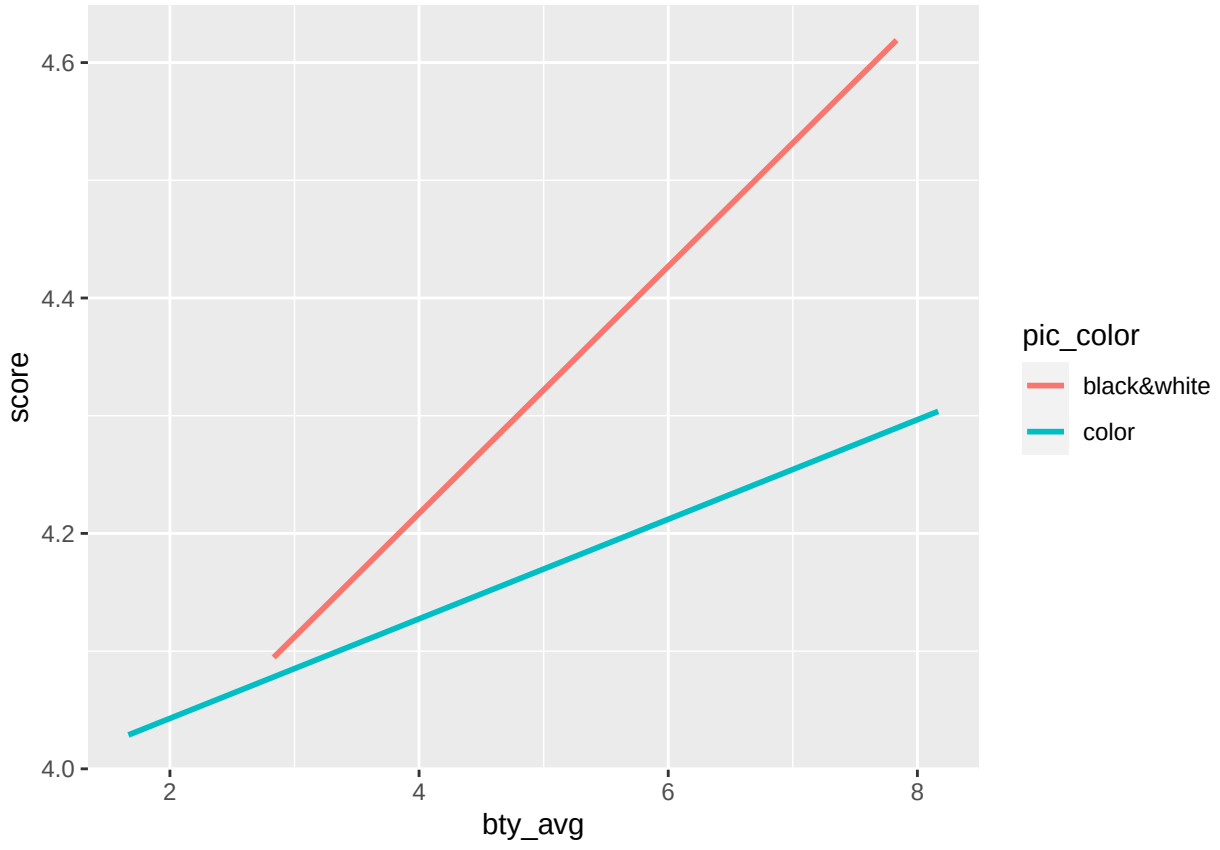
```
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

The p-value for `bty_avg` is lower than the previous model so it's still a statistically significant predictor. The value itself at 0.07416 is slightly higher than the 0.06664, however overall is still pretty small. Adding `gendermale` to the model has made it slightly higher it seems.

Exercise 9

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

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?




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

```
##
## 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
```

$$\hat{y} = 4.06318 + 0.05548 \times bty\_avg - 0.16059$$

For two professors who received the same beauty rating, those with a black and white color have a higher course evaluation score.

Exercise 10

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

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`.

```
##
## 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
```

> R seems to have added two variables for a categorical variable with three values; they left out `teaching`, but there is one for `ranktenure track` and `ranktenured`.

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. 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.

> I think variables that *wouldn't* have any association with score would be number of credits, age, and proportion of students that filled out evaluations. I believe these factors don't impact how a student scores a course. Perhaps the one that would have the highest p-value is `rank` given the more tenured professors have more experience so they're likely to be pretty good at teaching.

Exercise 12

```
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
```

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

$\hat{y} = 4.0952141 - 0.1475932 \times ranktenuretrack - 0.0973378 \times ranktenured + 0.2109481 \times gendermale + 0.1234929 \times ethnicity$

I was wrong here as number of credits, age, and proportion of students that filled out evaluations are all pretty significant with low p-values. The value with the least significant p-value was `cls_profssingle` (number of professors) while the one with the lowest p-value was `cls_creditsone credit`.

Exercise 13

Interpret the coefficient associated with the ethnicity variable.

The coefficient with `ethnicity` is saying that if the professor is not a minority and all things are equal, there is a slightly positive impact to `score`.

Exercise 14

```
m_full_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_full_no_cls_profs)
```

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?

```
##
## 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
```

The coefficients do change slightly, however the impact is very small and it also didn't affect the significance of most of the other variables. This means that the dropped variable was not that collinear with the other variables.

Exercise 15

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

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.

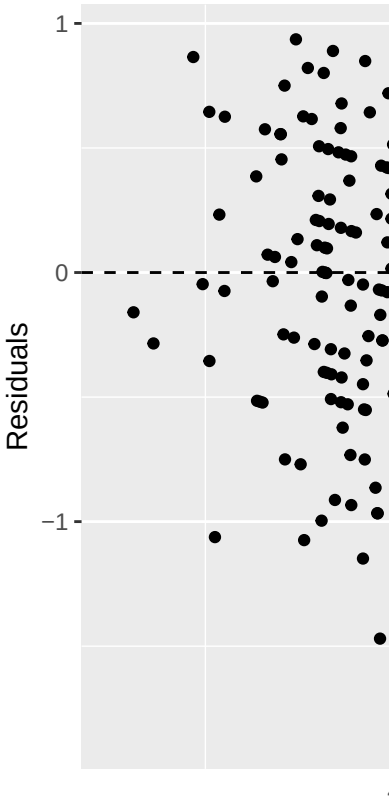
```
##
## Call:
## lm(formula = score ~ gender + ethnicity + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85320 -0.32394  0.09984  0.37930  0.93610
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.771922   0.232053  16.255  < 2e-16 ***
## gendermale               0.207112   0.050135   4.131 4.30e-05 ***
## ethnicitynot minority    0.167872   0.075275   2.230  0.02623 *
## 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
```

$$\hat{y} = 3.771922 + 0.207112 \times gendermale + 0.167872 \times ethnicitynotminority - 0.206178 \times languagenonenglish - 0.006046 \times ag$$

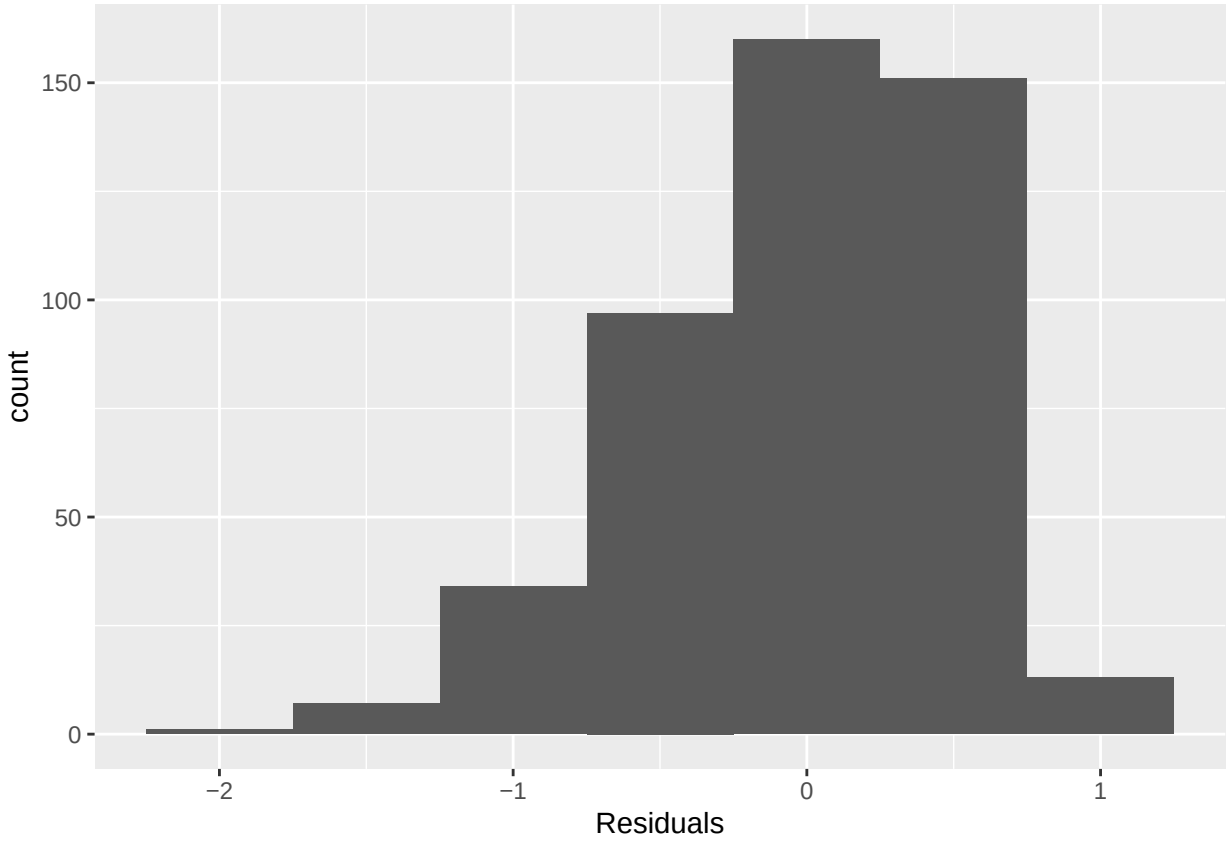
Exercise 16

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

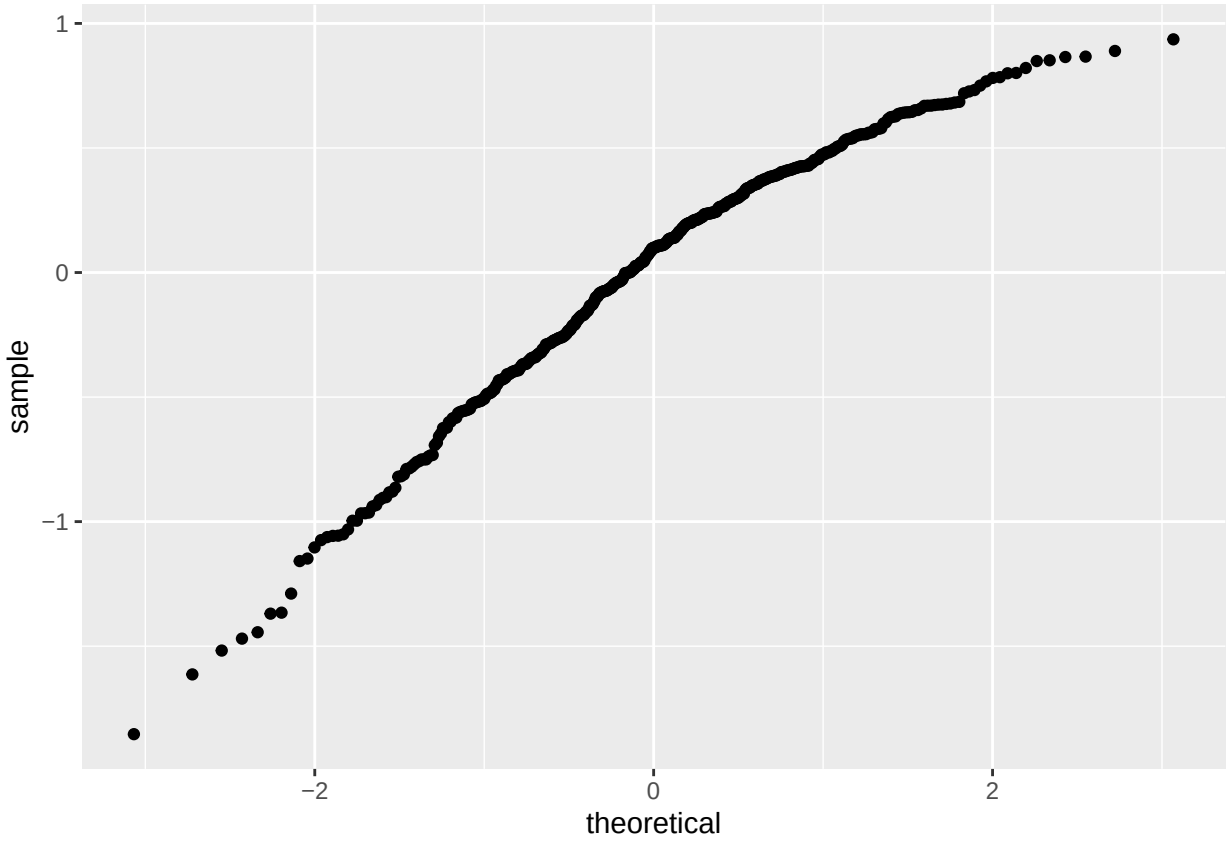


Verify that the conditions for this model are reasonable using diagnostic plots.

```
ggplot(data = m_best, aes(x = .resid)) +
  geom_histogram(binwidth = 0.5) +
  xlab("Residuals")
```




```
ggplot(data = m_best, aes(sample = .resid)) +
  stat_qq()
```



- Independence: We have already assumed this based on previous models.
- Linear Relationship: Visually, we can see a linear relationship.
- Constant Variance: We can see a uniform spread of residuals.
- Normality of Residuals: Residuals are normal across a reasonable range, which is shown by the chart above.

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?

No, because even a professor teaches multiple classes, courses are independent of each other so evaluation scores from one course is not related to another.

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.

A professor and course at the University of Texas at Austin with a high score would be one that has the following qualities:

- The professor is male
- The professor is not an ethnic minority
- The professor got their degree at an English-speaking university
- The professor is young
- A higher % of students filled out the evaluation
- The course is one credit
- The professor has a higher beauty rating
- The professor has a black and white photo

Exercise 19

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

> No, this sample size is not representative of every university or set of professors. It wouldn't make sense to apply this to another university that has a different dominant language for example since English would then not be most common.