Multiple linear regression

Souleymane Doumbia: I worked with Fomba Kassoh on this lab

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)
library(ggplot2)
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 4th Edition. You can read this by typing
## vignette('os4') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

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`.

```r
data("evals", package = 'openintro')
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 The study is an observational study, not an experiment.The question implies a causal relationship, that is, beauty directly causes differences in course evaluations. A more appropriate question that aligns with the design of an observational study would be:

"Is there a correlation between the perceived beauty of professors and their course evaluations at the University of Texas at Austin?".

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

The distribution of course scores can be characterized based on the histogram, box plot, and the statistical summary:

Histogram and Skewness: The histogram shows that the distribution of scores is slightly left-skewed (negatively skewed). This indicates that while most of the scores are high, there is a longer tail on the lower end of the scale.

Box Plot Analysis: The box plot further confirms the skewness, with the median closer to the upper quartile and a longer lower whisker. There are no significant outliers on the higher end, but a few on the lower end.

Statistical Summary: *Mean:* The average score is approximately 4.17. Standard Deviation: The standard deviation is about 0.54, suggesting some variation in the scores but not excessively wide. *Min and Max:* The scores range from a minimum of 2.3 to a maximum of 5.0. Quartiles: The 25th percentile is at 3.8, the median (50th percentile) is at 4.3, and the 75th percentile is at 4.6.

Interpretation Regarding Course Ratings: The skewness of the distribution indicates that students generally tend to rate courses relatively high, with fewer courses receiving lower scores. The presence of a tail on the lower end suggests that while most courses are rated well, there are some courses that receive significantly lower ratings.

Expectations: This distribution aligns with common trends observed in educational course evaluations, where students often tend to give moderately high to high ratings. It's not unusual for rating distributions in such contexts to be left-skewed, as extreme dissatisfaction is typically less common.

```r
# Plotting a histogram
p1 <- ggplot(evals, aes(x=score)) +
      geom_histogram(aes(y=after_stat(density)), bins=30, fill="blue", color="black") +
      geom_density(alpha=.2, fill="#FF6666") +
      ggtitle("Histogram of Course Scores")

# Plotting a box plot
p2 <- ggplot(evals, aes(x="", y=score)) +
      geom_boxplot(fill="tomato", color="black") +
      ggtitle("Box Plot of Course Scores") +
      theme(axis.title.x=element_blank(),
            axis.text.x=element_blank(),
            axis.ticks.x=element_blank())

# Arranging the plots in one window
library(gridExtra)
```
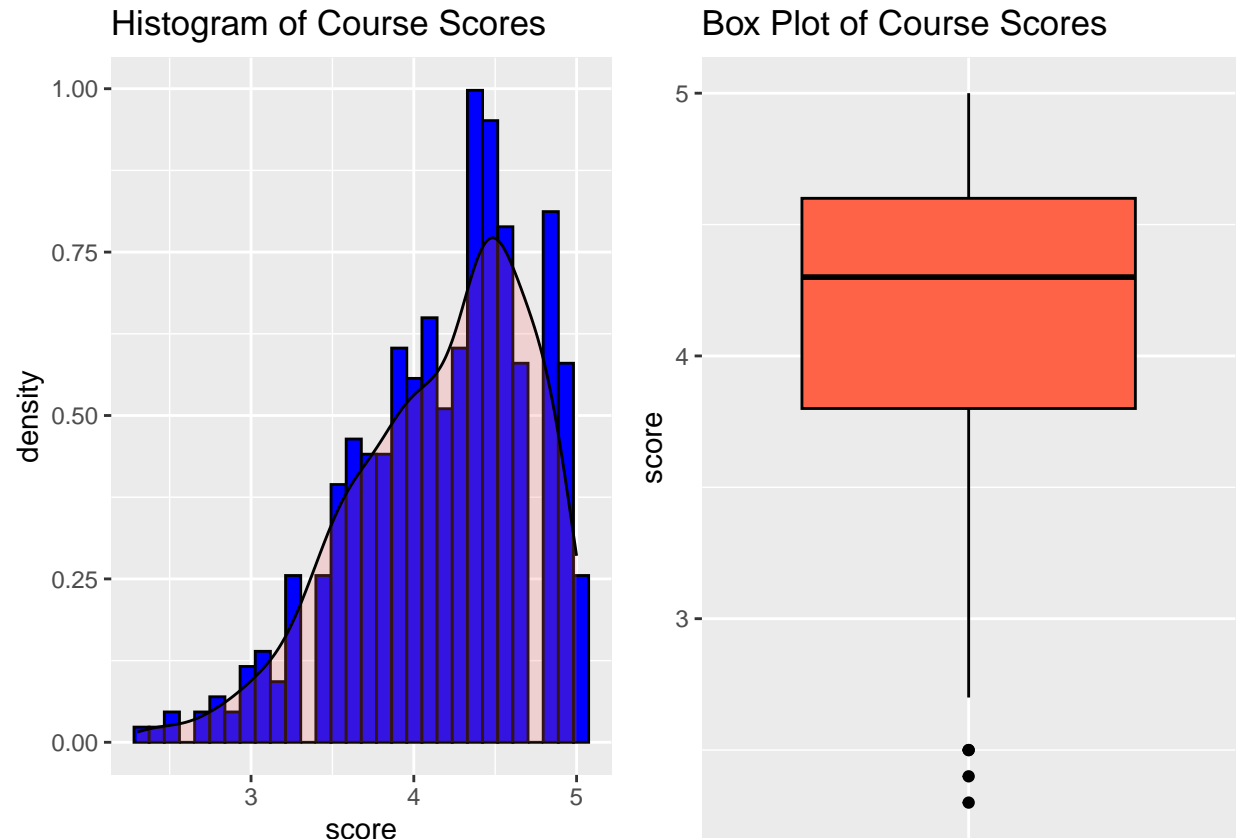
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
grid.arrange(p1, p2, ncol=2)
```

Histogram of Course Scores — Box Plot of Course Scores

```r
score_description <- summary(evals$score)

score_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.300   3.800   4.300   4.175   4.600   5.000
```

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

Insert your answer here From the available variables, two interesting variables to analyze could be:

age: The age of the professor. cls_students: The number of students in the class

The relationship between a professor's age and the class size they teach does not exhibit a clear pattern or trend.

```r
ggplot(evals, aes(x=age, y=cls_students)) +
  geom_point() +
  ggtitle('Relationship between Professor Age and Class Size') +
  xlab('Age of Professor') +
  ylab('Number of Students in Class') +
  theme_minimal()
```

Relationship between Professor Age and Class Size

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() +
  labs(title = 'Relationship between Beauty Score and Course Evaluation Score', x = 'Average Beauty Sco
```

Relationship between Beauty Score and Course Evaluation Score
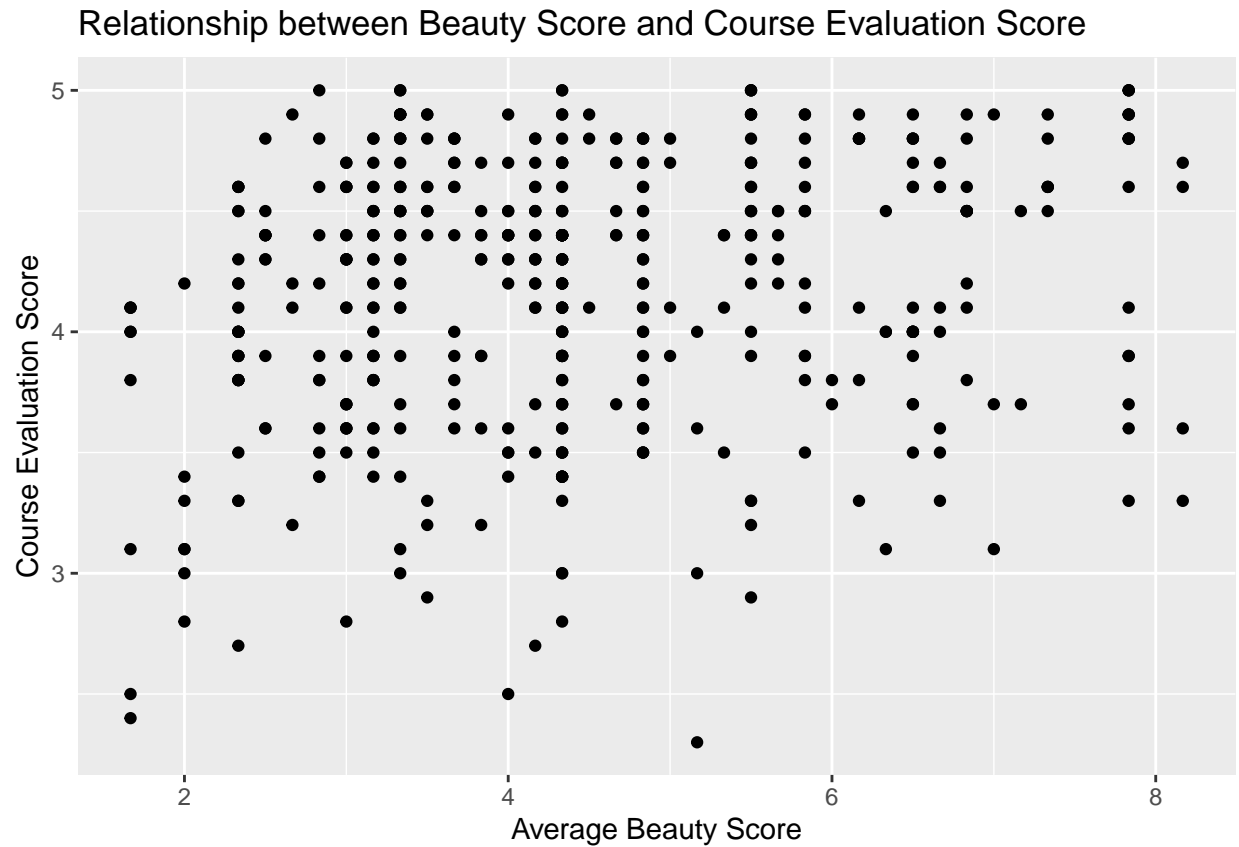
Before you draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

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() +
  labs(title = 'Relationship between Beauty Score and Course Evaluation Score', x = 'Average Beauty Scor
```

Relationship between Beauty Score and Course Evaluation Score
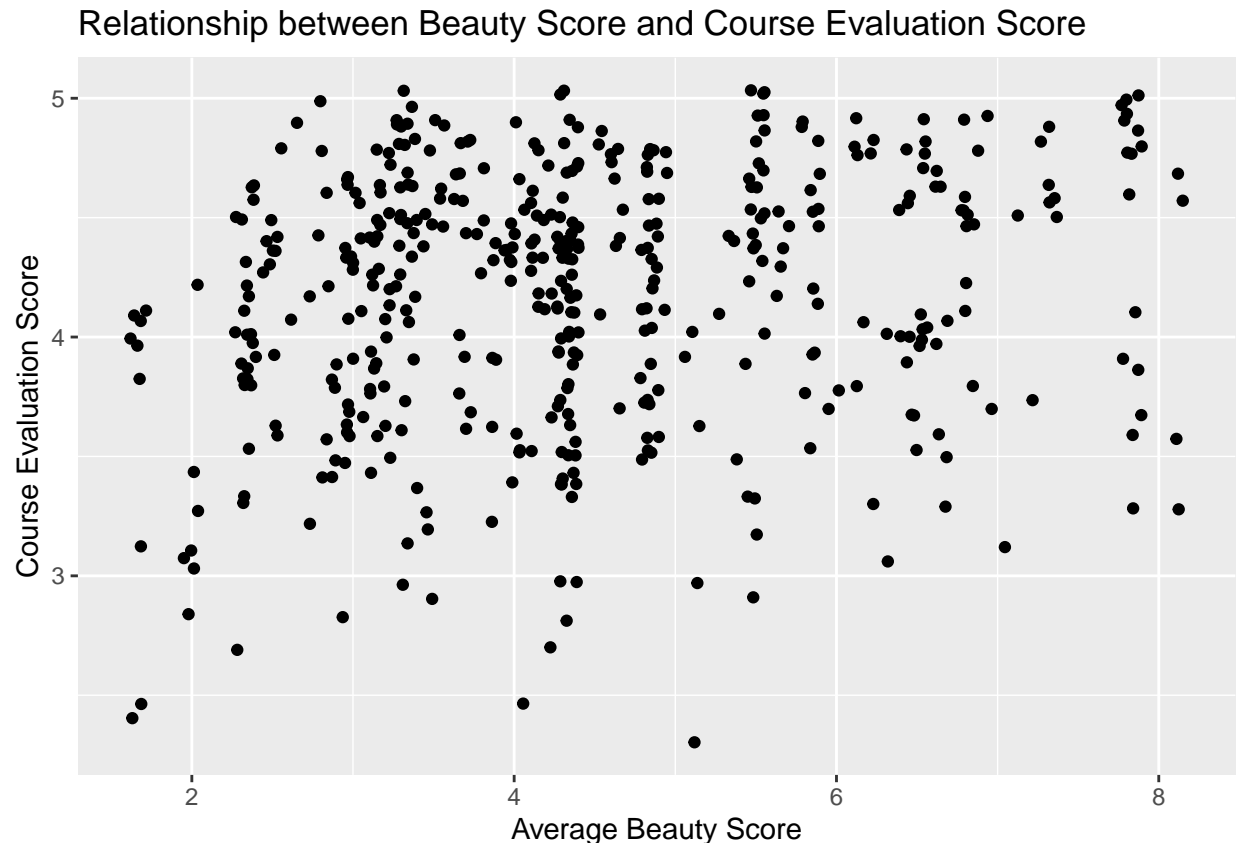
Insert your answer here There are 463 observations in the dataset. By observing the scatter plot, it appears that there are fewer points than 463 which is misleading. This could be due to multiple data points overlapping at the same coordinates.

```
evals %>%
  nrow()
```

```
## [1] 463
```

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 Linear model, equation, and interpretations as shown below.

Interpretation of slope: For each one-unit increase in the average beauty score, the average professor score is expected to increase by approximately 0.06664.

Statistical Significance: The p-value in the linear model is 5.08×10^{-05}. The p-value is significantly less than 0.05. This indicates that the average beauty score is a statistically significant predictor of the average professor score in this dataset.

While the average beauty score statistically significantly predicts the professor score, the practical significance of this effect may be considered relatively small because the slope value (0.06664) suggests that the effect size is modest.

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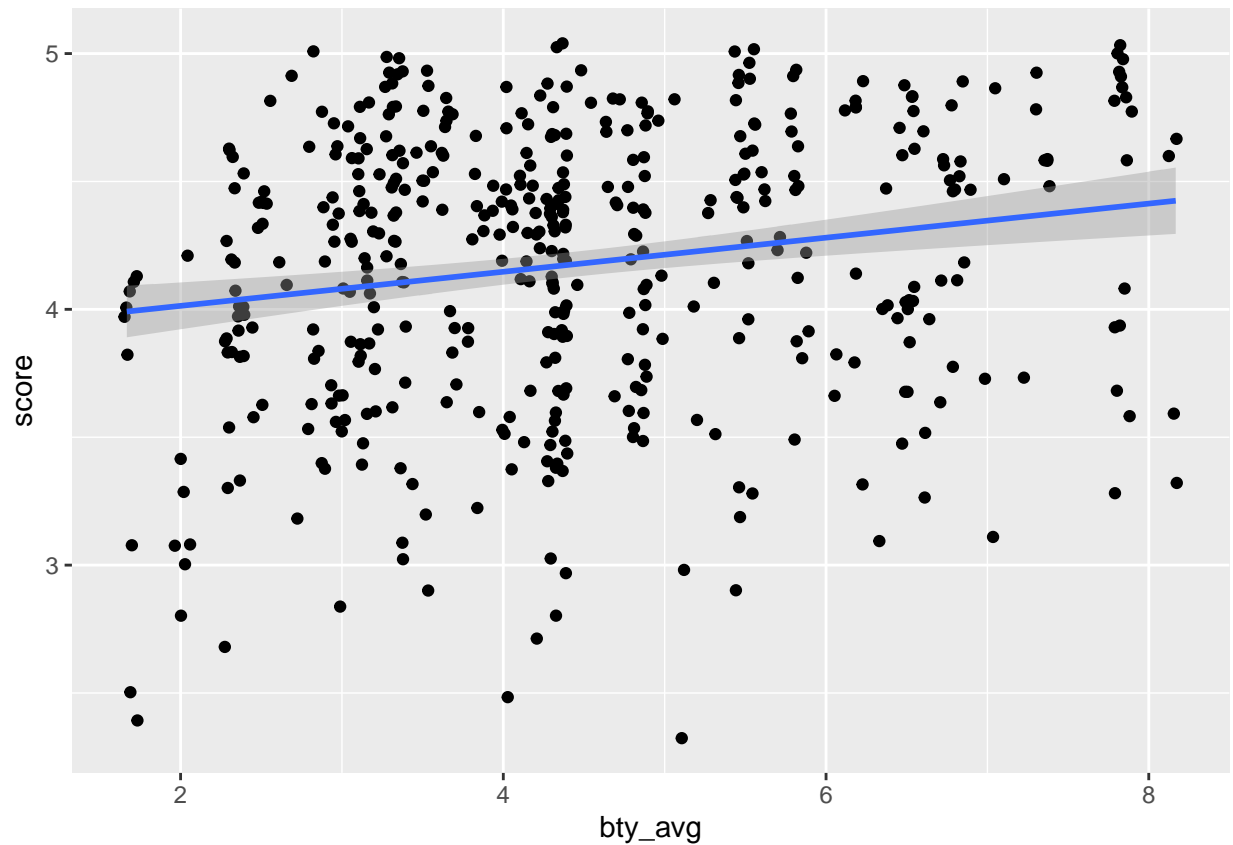
$$\hat{y} = 3.88034 + 0.06664 \times bty\_avg$$

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")
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```
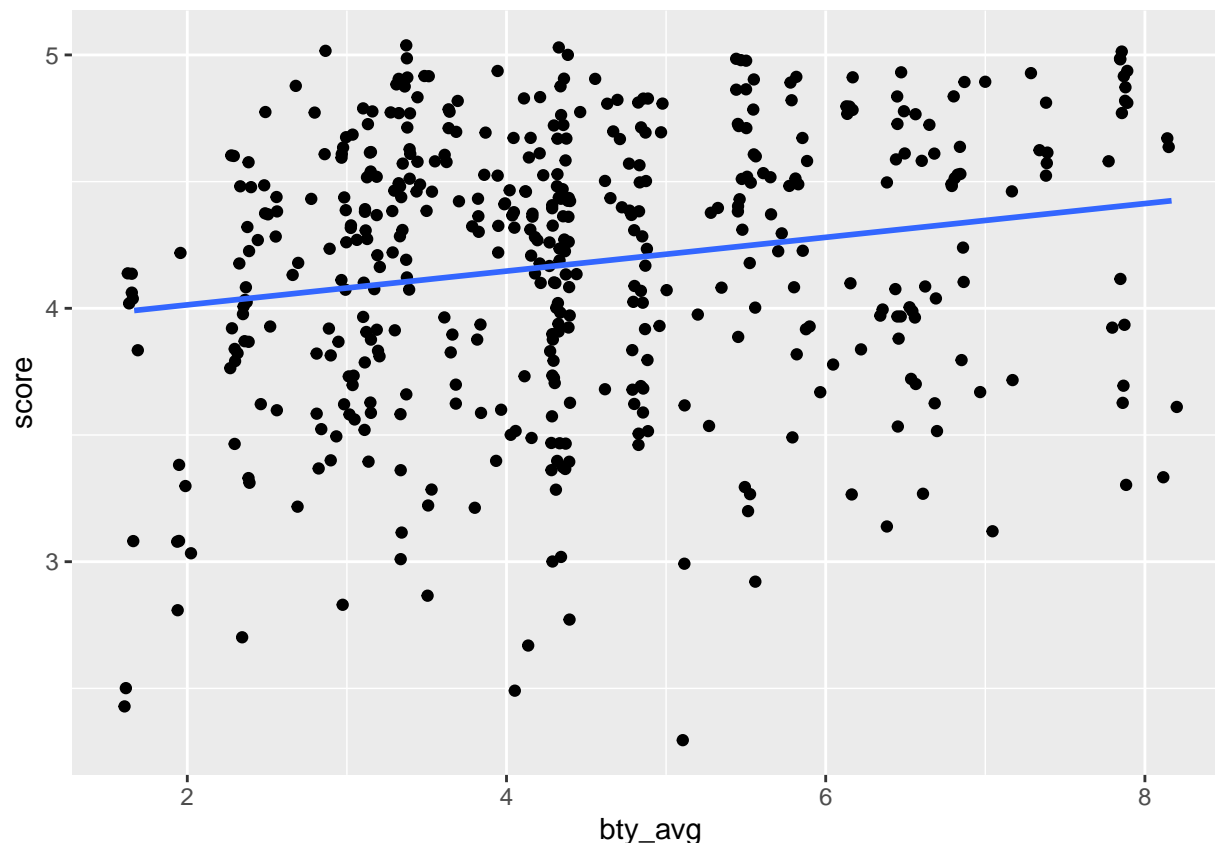
The blue line is the model. The shaded gray area around the line tells you about the variability you might expect in your predictions. To turn that off, use `se = FALSE`.

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

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

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 Below are the plots and comments.

Residuals vs Fitted Values Plot: In this plot, there does not appear to be a clear, systematic pattern in the residuals, which is a good sign for linearity. Also, the plot shows a fairly consistent spread of residuals across the range of fitted values, indicating that the assumption of homoscedasticity may be reasonable.

Normal Q-Q Plot of Residuals: The points in this Q-Q plot generally follow the line, especially in the middle of the distribution. There are some deviations at the ends, but these are not extreme. This suggests that the normality assumption is not strongly violated. However, the slight deviations at the ends indicate that the residuals may not be perfectly normally distributed.

```r
# Calculating residuals and fitted values
fitted_values <- fitted(m_bty)
residuals <- residuals(m_bty)

# Creating a data frame for plotting
plot_data <- data.frame(Fitted_Values = fitted_values, Residuals = residuals)

# Plotting residuals vs fitted values
p1 <- ggplot(plot_data, aes(x = Fitted_Values, y = Residuals)) +
    geom_point() +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    xlab("Fitted Values") +
```
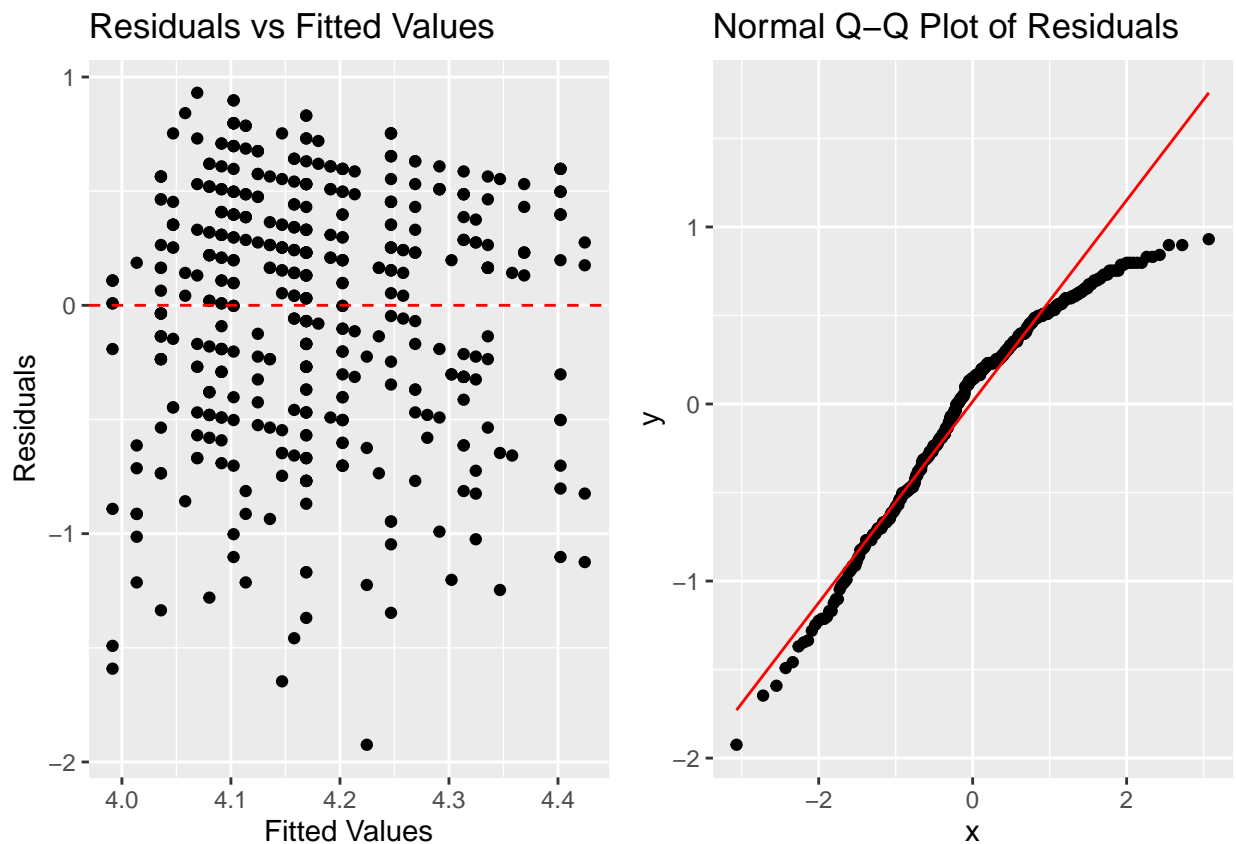
```
    ylab("Residuals") +
    ggtitle("Residuals vs Fitted Values")

# Plotting Q-Q plot for residuals
p2 <- ggplot(plot_data, aes(sample = Residuals)) +
    stat_qq() +
    stat_qq_line(color = "red") +
    ggtitle("Normal Q-Q Plot of Residuals")

# Arranging the plots side by side in one window
library(gridExtra)
grid.arrange(p1, p2, ncol = 2)
```


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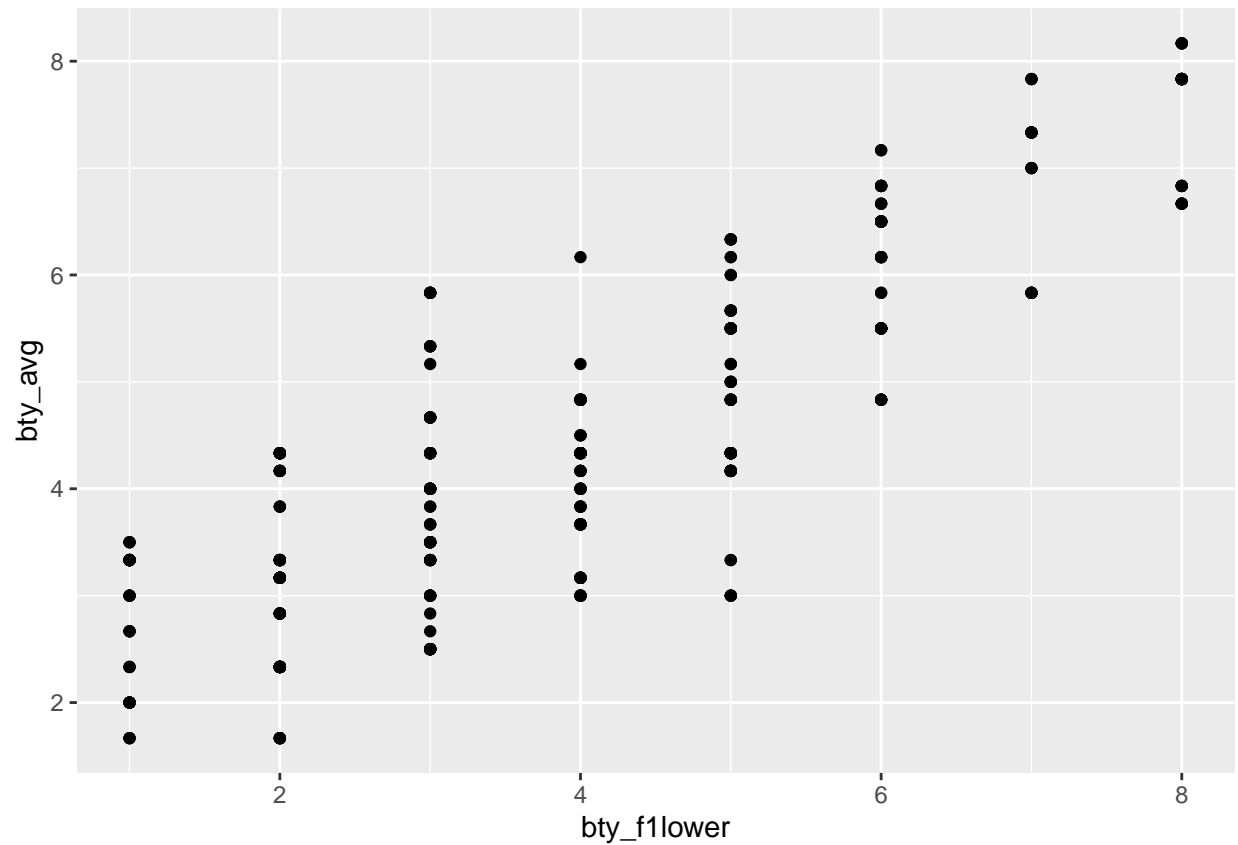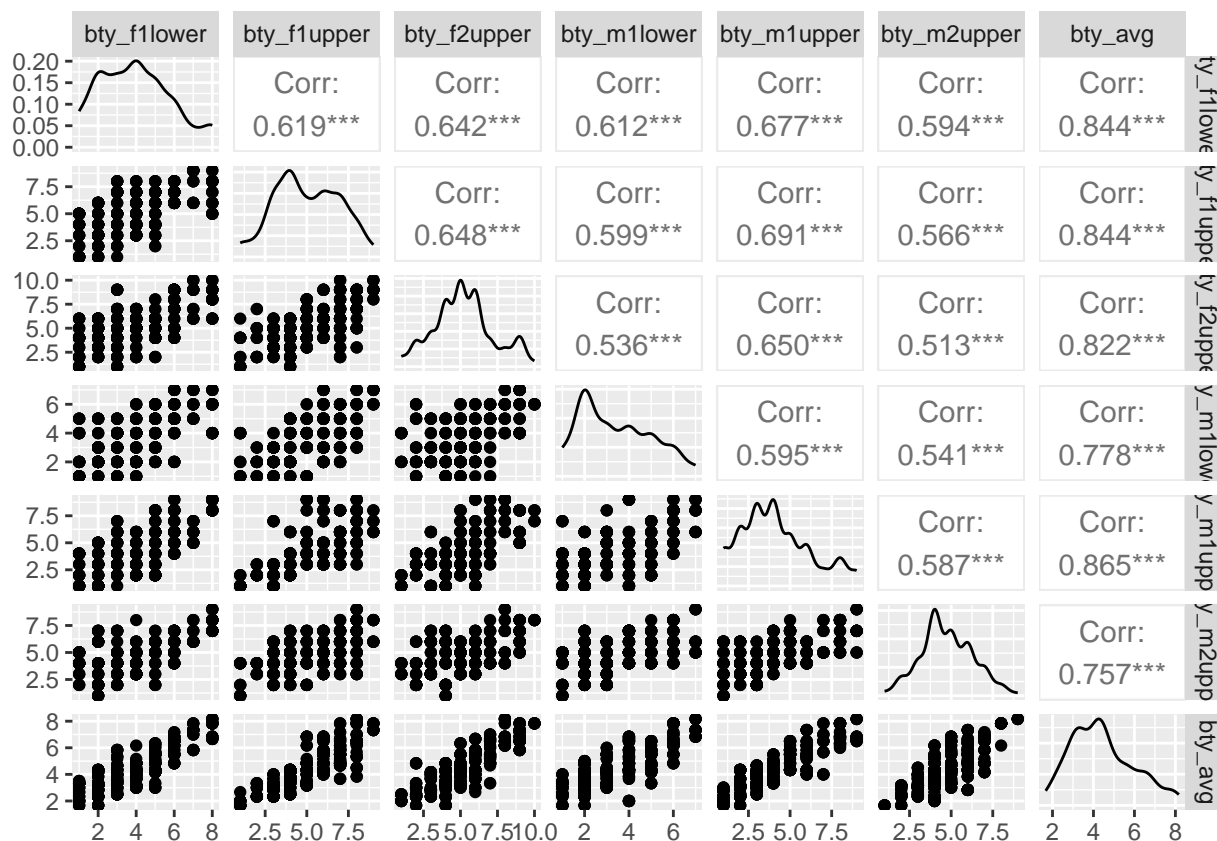
```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # 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
##
```

13

```
## 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 Here are the diagnostic plots for the multiple linear regression model:

Residuals vs Fitted Values Plot: There doesn't appear to be a clear pattern in the residuals, which is good for the assumption of linearity. The spread of residuals seems relatively constant across the range of fitted values, supporting the homoscedasticity assumption.

Normal Q-Q Plot: The points largely follow the line, especially in the middle of the distribution, indicating that the residuals are approximately normally distributed. Some deviations at the ends are observed but are not extreme.

Scale-Location Plot: The spread of residuals appears consistent across the range of fitted values, which supports the assumption of constant variance.

Residuals vs Leverage Plot: Most data points have low leverage, and there don't appear to be any points with both high leverage and large residuals, which would be indicative of influential cases.

These diagnostic plots suggest that the assumptions of linearity, independence, homoscedasticity, and normality are reasonably met for this multiple linear regression model.

```r
library(gridExtra)
library(stats)


fitted_values_multi <- fitted(m_bty_gen)
residuals_multi <- residuals(m_bty_gen)

# 1. Residuals vs Fitted Values Plot
p7_1 <- ggplot() +
    geom_point(aes(x=fitted_values_multi, y=residuals_multi)) +
    geom_hline(yintercept=0, color="red", linetype="dashed") +
    xlab("Fitted Values") + ylab("Residuals") +
    ggtitle("Residuals vs Fitted Values")

# 2. Normal Q-Q Plot
p7_2 <- ggplot() +
    stat_qq(aes(sample=residuals_multi)) +
    geom_abline(slope=1, intercept=0, color="red", linetype="dashed") +
    ggtitle("Normal Q-Q Plot")

# 3. Scale-Location Plot
p7_3 <- ggplot() +
    geom_point(aes(x=fitted_values_multi, y=sqrt(abs(residuals_multi)))) +
    xlab("Fitted Values") + ylab("Sqrt(|Residuals|)") +
    ggtitle("Scale-Location Plot")

# 4. Residuals vs Leverage Plot
leverage <- hatvalues(m_bty_gen)
p7_4 <- ggplot() +
    geom_point(aes(x=leverage, y=residuals_multi)) +
```
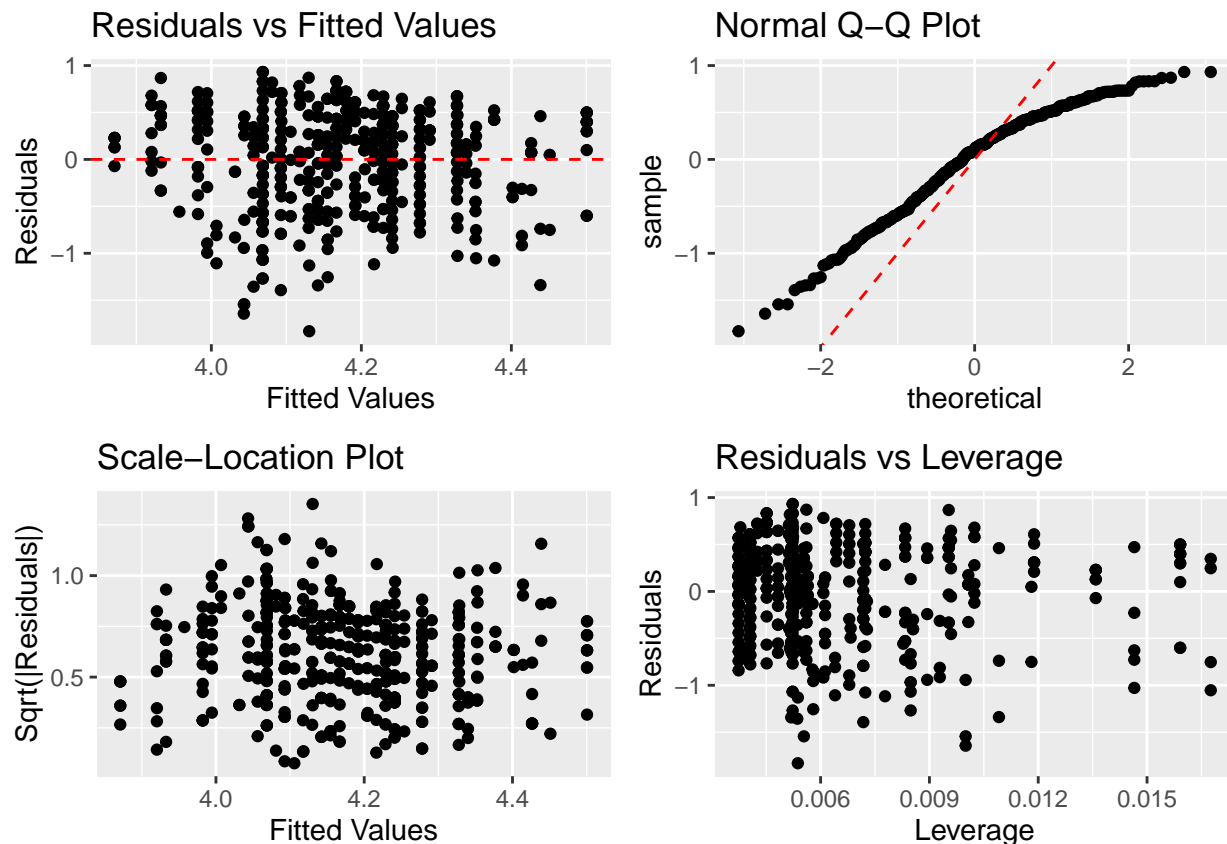
```
        xlab("Leverage") + ylab("Residuals") +
        ggtitle("Residuals vs Leverage")

# Arranging plots in a 2x2 grid
grid.arrange(p7_1, p7_2, p7_3, p7_4, nrow=2)
```


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 Based on the coefficients and p-values above, bty_avg remains a significant predictor of score even after accounting for gender. The addition of gender to the model affects the parameter estimate for bty_avg, although the change is relatively small. The inclusion of gender reveals its own significant effect on the professor score.

Summary of parameters: Coefficient of bty_avg: 0.0742 P-value of bty_avg: 6.48×10^{-6}

Coefficient of gender: 0.1724 P-value of gender[T.male]: 0.00065

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.

15

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

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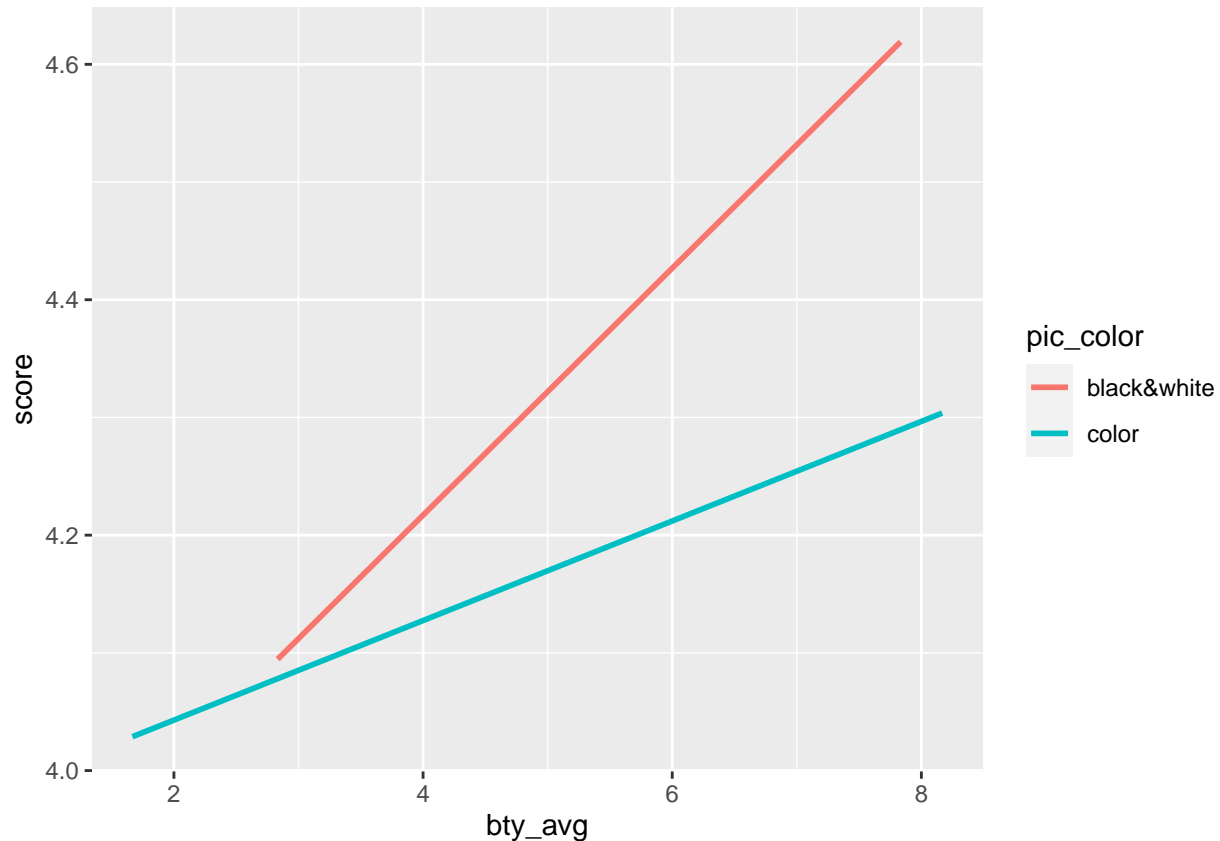

9. What is the equation of the line corresponding to those with color pictures? (*Hint:* For those with color pictures, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which color picture tends to have the higher course evaluation score?

Insert your answer here The equation of the line corresponding to those with color pictures is shown below.Based on the equation, the balck and white color picture tend to have the higher course evaluation score because the coefficient corresponding to color picture is negative which have the effect of reducing the course evaluation score.

```
# Fitting the linear model
model_color <- lm(score ~ bty_avg + gender + pic_color, data = evals)

model_summary_color <- summary(model_color)
model_summary_color
```

```
##
## Call:
```

```
## lm(formula = score ~ bty_avg + gender + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7807 -0.3587  0.1142  0.4113  0.8839
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.95031    0.11173  35.354  < 2e-16 ***
## bty_avg          0.06171    0.01676   3.683 0.000258 ***
## gendermale       0.18750    0.05016   3.738 0.000209 ***
## pic_colorcolor  -0.18851    0.06837  -2.757 0.006064 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5249 on 459 degrees of freedom
## Multiple R-squared:  0.07445,    Adjusted R-squared:  0.0684
## F-statistic: 12.31 on 3 and 459 DF,  p-value: 9.321e-08
```

equation

$$\hat{y} = 3.95031 + 0.06664 \times bty\_avg + 0.18750 - 0.18851$$

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 appears to treat teaching as the baseline category. That is, it multiply the the coefficient for teaching by 0. The model only generates coefficients for bty_avg, tenure track, and tenured by multiplying their respective coefficients by 1. The equations for predicting the score based on the category of rank and bty_avg are as follows:

equation for teaching
$$\widehat{score} = 3.98155 + 0.06783 \times bty\_avg$$

equation for tenure track

$$\widehat{score} = 3.98155 + 0.06783 \times bty\_avg - 0.06783$$

0.16070

equation for tenured
$$\widehat{score} = 3.98155 + 0.06783 \times bty\_avg - 0.12623$$

0.16070

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

model_summary_bty_rank <- summary(m_bty_rank)
model_summary_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
```

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 expect cls_profs to have the highest p-value. While class size or the number of professors could impact scores, these are more indirectly related compared to individual characteristics of the professor or class level. Therefore, cls_profs would not be expected to have a significant impact on professor score.

Let's run the model...

```
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)
```

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

Insert your answer here My suspicion is in agreement with the model. The fact that cls_profs variable has the highest p-value suggests that the number of professors teaching a class (single vs. multiple) might not significantly impact how students evaluate the course or professor.

```
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
```

13. Interpret the coefficient associated with the ethnicity variable.

Insert your answer here *Interpretation of "Ethnicity of professor: not minority, minority."*
This coefficient, 0.1234929, suggests that, on average, non-minority professors are expected to have a score
that is 0.1234929 points higher than minority professors, holding all other variables constant. It reflects the
difference in the average score attributed to the ethnicity of the professor, under the assumption that all
other factors (like beauty score, rank, gender, etc.) are the same for the compared groups. The positive
coefficient indicates a favorability towards non-minority professors in terms of how they are scored, according
to this model.

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 There is no change in the p-value which implies that there is no change in the significance. Similarly, there is no significant change in the coefficients. with the exception of ethnicity, cls_students, and pic_outfit where the change in the coefficient is between 3% and 4% all other variables changed by less than 1%.

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

model_summary_reduced <- summary(m_reduced)
model_summary_reduced
```

```
##
## 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
```

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 Steps to fit the final model ***Starting with the Full Model:*** Fit the model

with all predictors. ***Iteratively Removing Variables:*** Based on the p-values, iteratively remove p-value > 0.05 ***Final Model:*** Continue until all variables in the model are significant (p-values below 0.05).

```
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
```

Best Model

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

```
##
## 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
```

equation for best model

$$\widehat{score} = 3.771922 + 0.207112 - 0.167872 - 0.206178 - 0.006046 \times age + 0.004656 \times cls_perc_eval + 0.505306 + 0.051069 \times bty_avg - 0.1905$$

$$\widehat{score} = 3.919711 - 0.006046 \times age + 0.004656 \times cls_perc_eval + 0.051069 \times bty_avg$$

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

Insert your answer here Below are the diagnostic plots and analsis.

Residuals vs Fitted Plot: In this plot, there seems to be no clear pattern, which is a good sign. However, there is a slight funnel shape, suggesting potential heteroscedasticity.

Normal Q-Q Plot: The plot shows that most points lie on the line, but there are deviations at the ends. This indicates that the residuals are nearly, but not perfectly, normally distributed.

Scale-Location (Spread-Location) Plot: There seems to be a consistent spread, but with slight variations, hinting at possible, but not severe, heteroscedasticity.

Residuals vs Leverage Plot: There are a few points with high leverage, but they don't seem to have large residuals. Thus, while there are influential points, they may not be adversely affecting the model too much.

Overall, the model appears to meet the assumptions reasonably well, though there are indications of minor deviations from ideal conditions.

```
# Add residuals and fitted values to the data frame
evals <- evals %>%
  mutate(fitted_values = fitted(m_final),
         residuals = resid(m_final),
         sqrt_abs_resid = sqrt(abs(residuals)),
         leverage = hatvalues(m_final))
```

```
# Creating the plots
p16_1 <- ggplot(evals, aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype="dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted values", y = "Residuals")

p16_2 <- ggplot(evals, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(slope=1, intercept=0, color="red", linetype="dashed") +
  labs(title = "Normal Q-Q", x = "Theoretical Quantiles", y = "Sample Quantiles")


## Warning in stat_qq_line(slope = 1, intercept = 0, color = "red", linetype =
## "dashed"): Ignoring unknown parameters: 'slope' and 'intercept'

p16_3 <- ggplot(evals, aes(x = fitted_values, y = sqrt_abs_resid)) +
  geom_point() +
  labs(title = "Scale-Location Plot", x = "Fitted values", y = "Sqrt(|Residuals|)")

p16_4 <- ggplot(evals, aes(x = leverage, y = residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype="dashed") +
  labs(title = "Residuals vs Leverage", x = "Leverage", y = "Residuals")

grid.arrange(p16_1, p16_2, p16_3, p16_4, ncol = 2, nrow = 2)
```
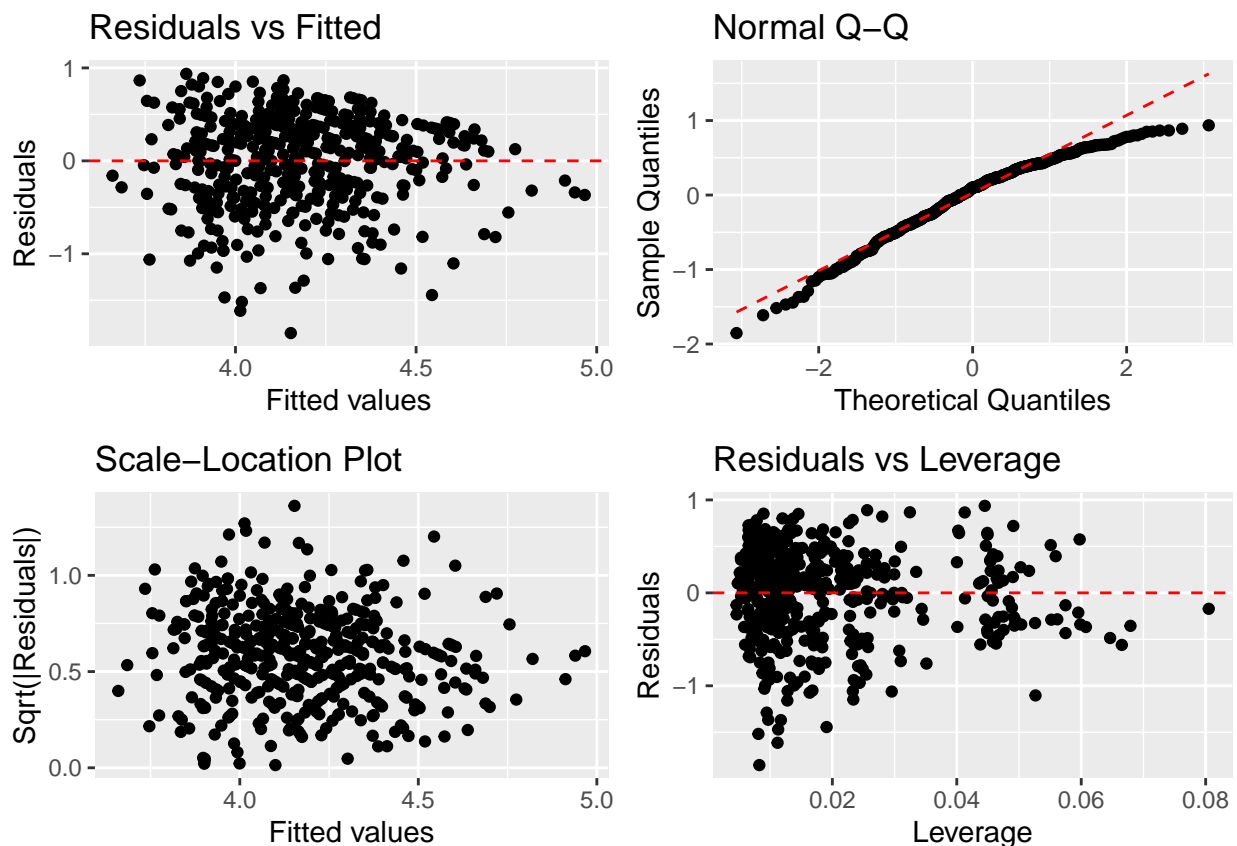
17. The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?

Insert your answer here Yes, the new information about how the data were gathered can have significant implications for the conditions and assumptions underlying linear regression analysis. Here are a few key points to consider:

The fact that each row in the dataset represents a course taught by professors at the University of Texas at Austin impacts the assumptions of linear regression. Specifically, courses taught by the same professor may not be independent, violating a key assumption of linear regression. Additionally, the variability in scores could differ across different types of courses (advanced vs. basic), affecting the assumption of constant variance (homoscedasticity). Therefore, the data might require analysis using methods that account for these potential dependencies and clustering, like hierarchical linear modeling.

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 Based on the final linear regression model for professors' evaluations at the University of Texas at Austin, certain characteristics are associated with higher evaluation scores. These include specific aspects of the professor's demographic profile, such as gender, ethnicity, and age, indicating potential influences of identity and experience on evaluations. Language proficiency is also a key factor, suggesting the importance of effective communication in the classroom. Additionally, the beauty average (bty_avg) hints at a potential bias in student evaluations based on physical appearance. Course-related factors like the percentage of class evaluations completed and the number of course credits play a role as well, reflecting aspects of course engagement and intensity. Finally, the influence of the professor's picture color might indicate the impact of perceived professionalism on student evaluations.

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

Insert your answer here The data is specific to the University of Texas at Austin, which may have unique characteristics not representative of other institutions. Differences in student populations, teaching styles, and institutional cultures can significantly influence evaluation scores. Furthermore, the model may not account for all relevant variables affecting teaching evaluations, and factors like societal biases could vary across different universities. Therefore, while the findings are valuable for the specific context, applying them universally requires further research in diverse educational settings.