IDS 702 HW 3

KEY

Instructions: Use this template to complete your assignment. When you click "Render," you should get a PDF document that contains both your answers and code. You must show your work/justify your answers to receive credit. Submit your rendered PDF file on Gradescope. **Remember to render frequently**, as this will help you to catch errors in your code before the last minute. You must show your work for all problems, and you must provide a written answer for all problems. For example, you should write "There are XX observations and each observation represents a ________"; it is insufficient to just show the code.

Add your name in the Author section in the header

Load Data

```
library(tidyverse)
library(gridExtra)

colleges <- read.csv("https://raw.githubusercontent.com/anlane611/datasets/refs/heads/main/c
```

Part 1

1

a. 797 observations and each observation is a college/university. (No code necessarily required, and can be something like `glimpse()` instead of `nrow()`)

```
nrow(colleges)
```

```
[1] 797
```

b.

```r
summary(colleges[,c("control","basic","student_count","ft_pct",
                    "med_sat_value","endow_value",
                    "grad_100_value","grad_150_value",
                    "retain_value")])
```

```
   control              basic           student_count       ft_pct      
 Length:797         Length:797         Min.   :   82   Min.   :  6.00  
 Class :character   Class :character   1st Qu.:  979   1st Qu.: 80.50  
 Mode  :character   Mode  :character   Median : 1677   Median : 91.90  
                                       Mean   : 5544   Mean   : 86.41  
                                       3rd Qu.: 5156   3rd Qu.: 97.20  
                                       Max.   :51333   Max.   :100.00  
                                                                       
 med_sat_value   endow_value      grad_100_value  grad_150_value 
 Min.   : 666   Min.   :     28   Min.   : 0.00   Min.   : 0.00  
 1st Qu.: 990   1st Qu.:   7964   1st Qu.:23.60   1st Qu.:40.52  
 Median :1081   Median :  20869   Median :38.00   Median :55.55  
 Mean   :1099   Mean   :  78374   Mean   :41.97   Mean   :55.87  
 3rd Qu.:1192   3rd Qu.:  55975   3rd Qu.:59.55   3rd Qu.:72.15  
 Max.   :1534   Max.   :2505435   Max.   :92.80   Max.   :97.80  
 NA's   :168    NA's   :71        NA's   :7       NA's   :7      
  retain_value   
 Min.   :  0.00  
 1st Qu.: 65.80  
 Median : 76.10  
 Mean   : 74.48  
 3rd Qu.: 86.60  
 Max.   :100.00  
 NA's   :5       
```

The following variables have missing values: med_sat_value (168), endow_value (71), grad_100_value (7), grad_150_value (7), and retain_value (5).

```r
colleges_full <- colleges |>
  filter(complete.cases(control, basic, student_count,
                        ft_pct, med_sat_value, endow_value,
                        grad_100_value, grad_150_value,
                        retain_value))
```

2

```r
colleges_full |>
  count(basic) |>
  mutate(prop=n/sum(n))
```

```
                                             basic   n      prop
1           Baccalaureate Colleges--Arts & Sciences 195 0.3170732
2            Baccalaureate Colleges--Diverse Fields 231 0.3756098
3     Research Universities--high research activity  84 0.1365854
4 Research Universities--very high research activity 105 0.1707317
```

```r
colleges_full |>
  count(control) |>
  mutate(prop=n/sum(n))
```

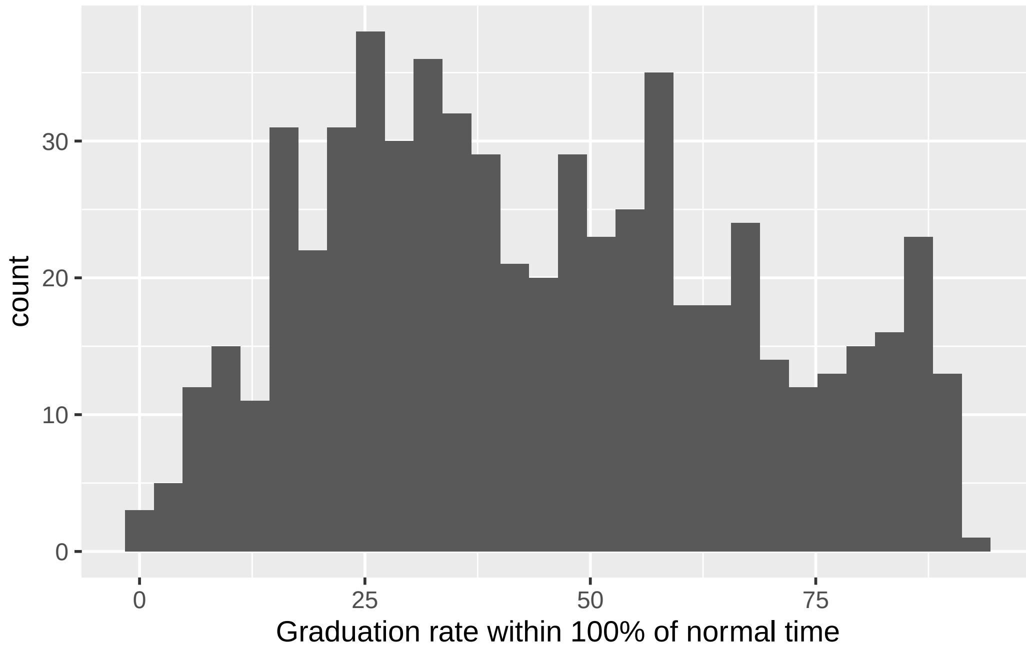
```
                 control   n      prop
1 Private not-for-profit 409 0.6650407
2                 Public 206 0.3349593
```

	Count (n)	Proportion or %
Basic classification		
Baccalaureate Colleges–Arts & Sciences	195	32
Baccalaureate Colleges–Diverse Fields	231	38
Research Universities–high research activity	84	14
Research Universities–very high research activity	105	17
Control of institution		
Private not-for-profit	409	67
Public	206	33

3

a.

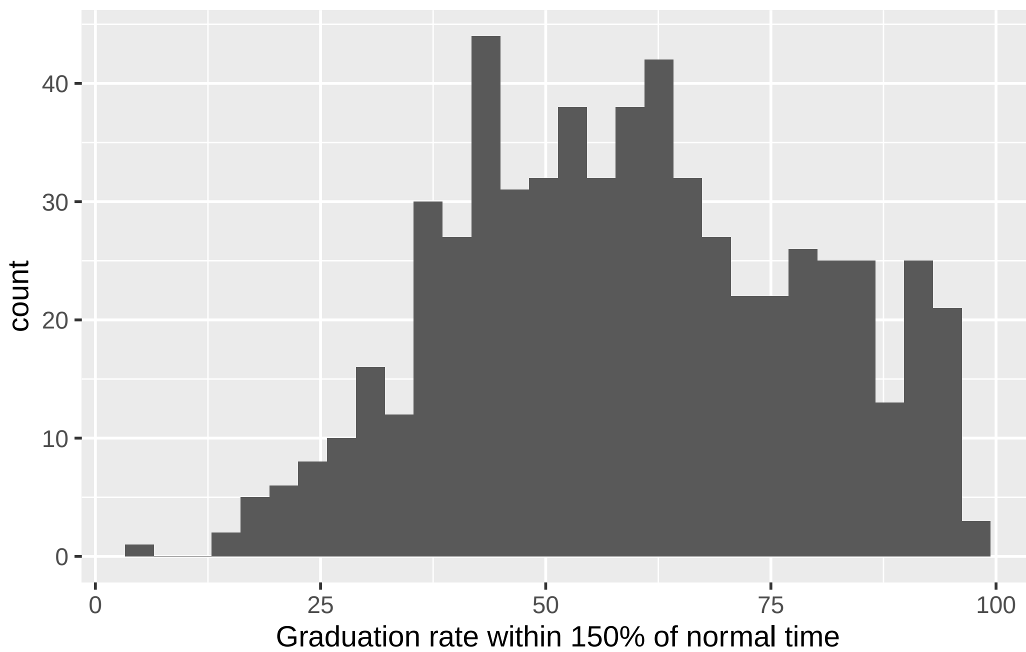
```r
ggplot(colleges_full, aes(x=grad_100_value))+
  geom_histogram()+
  labs(x="Graduation rate within 100% of normal time")
```



The distribution is relatively evenly spread across the full range 0-100, which is not necessarily what we would expect to see. Many institutions have graduation rates below 50%, which is lower than we might expect.

b.

```
ggplot(colleges_full, aes(x=grad_150_value))+
  geom_histogram()+
  labs(x="Graduation rate within 150% of normal time")
```



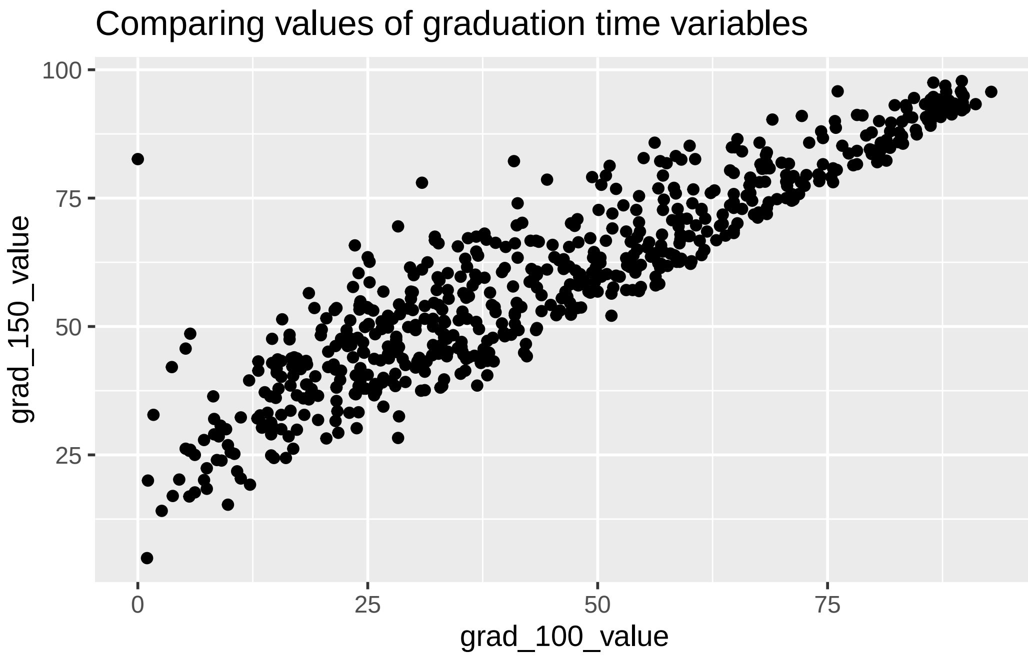

```
colleges_full[which(colleges_full$chronname=="Northeastern University"),
              c("grad_100_value","grad_150_value")]
```

```
    grad_100_value grad_150_value
237              0           82.6
```

The distribution of graduation rate within 150% of normal time looks more like what we would expect, with a large proportion of institutions having a rate above ~40%. Looking at Northwestern, we see that the grad_100_value is 0, while grad_150_value is 82.6. This is likely a data error, so grad_150_value seems to have more reliable data.

Note: It could also be valuable to look at a scatter plot to compare values of grade_100_value to grad_150_value. In the plot below, we see that some institutions have very low values of grad_100_value and much higher values of grad_150_value. This seems surprising, so grad_150_value seems more reliable.

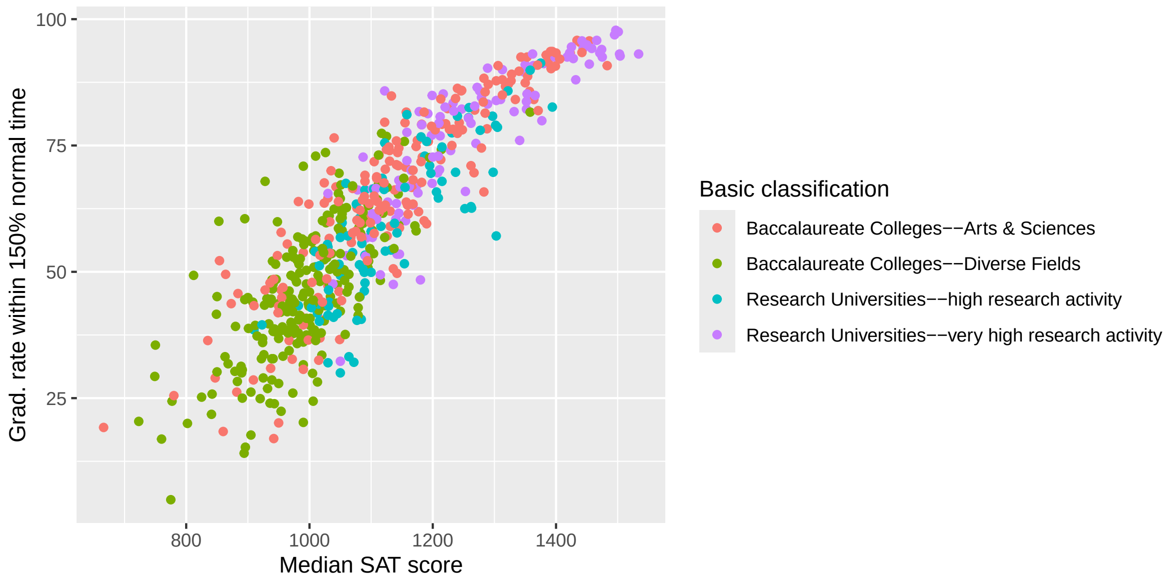
```
ggplot(colleges_full, aes(x=grad_100_value, y=grad_150_value))+
  geom_point()+
  ggtitle("Comparing values of graduation time variables")
```



4

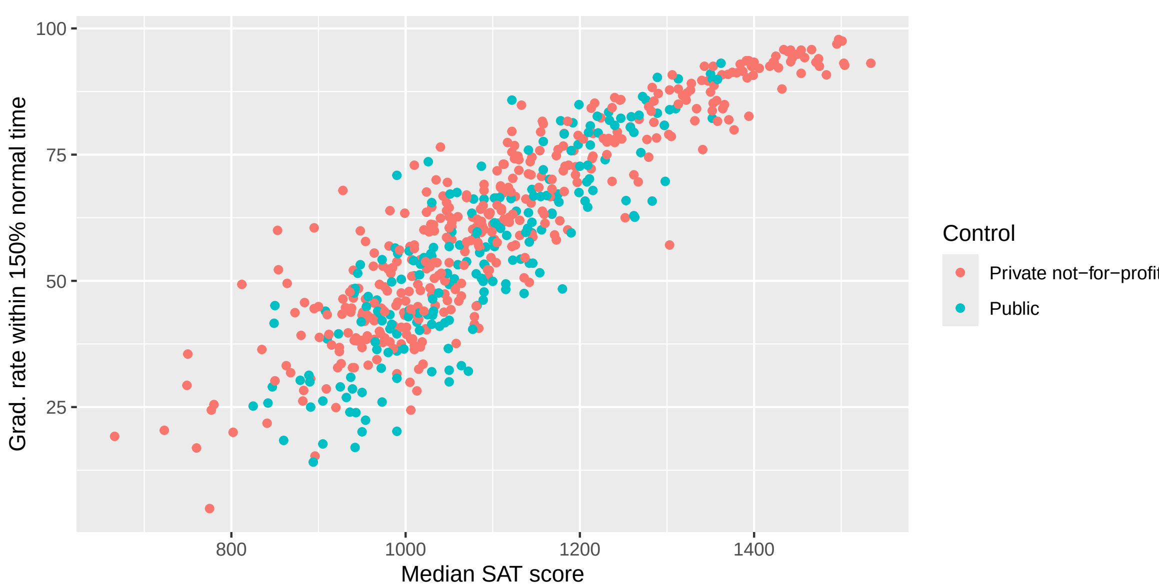
```
colleges_full$basic_fac <- factor(colleges_full$basic)
colleges_full$control_fac <- factor(colleges_full$control)

ggplot(colleges_full, aes(x=med_sat_value, y=grad_150_value,
                                    col=basic_fac))+
  geom_point()+
  labs(x="Median SAT score",y="Grad. rate within 150% normal time",
       col="Basic classification")
```



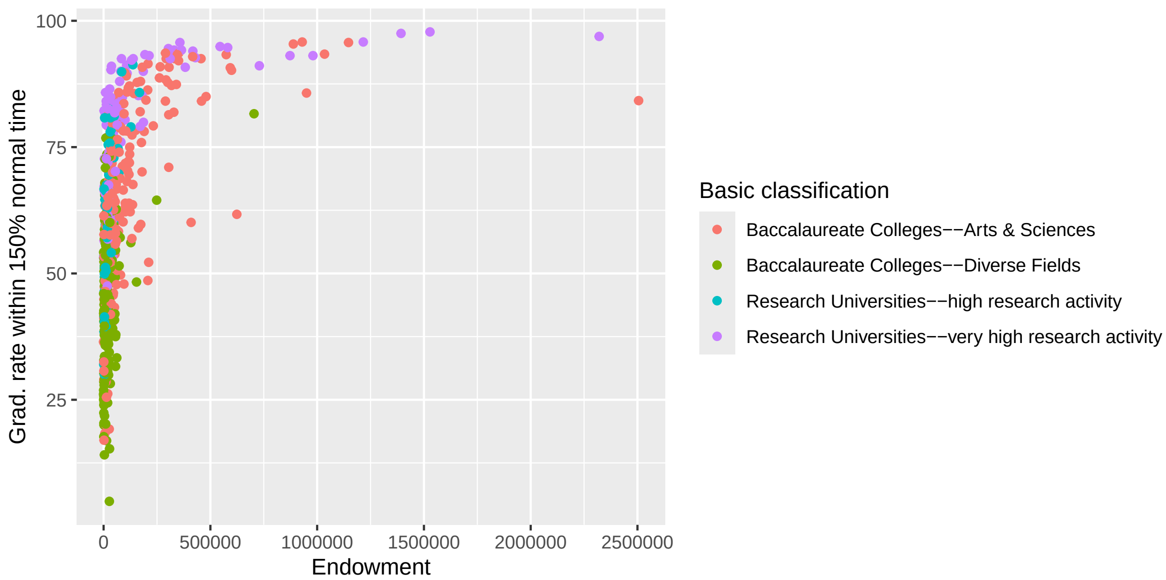
Graduation rates increase as median SAT score increases, generally, and the research universities tend to have higher median SAT scores and graduation rates.

```
ggplot(colleges_full, aes(x=med_sat_value, y=grad_150_value,
                           col=control_fac))+
  geom_point()+
  labs(x="Median SAT score",y="Grad. rate within 150% normal time",
       col="Control")
```



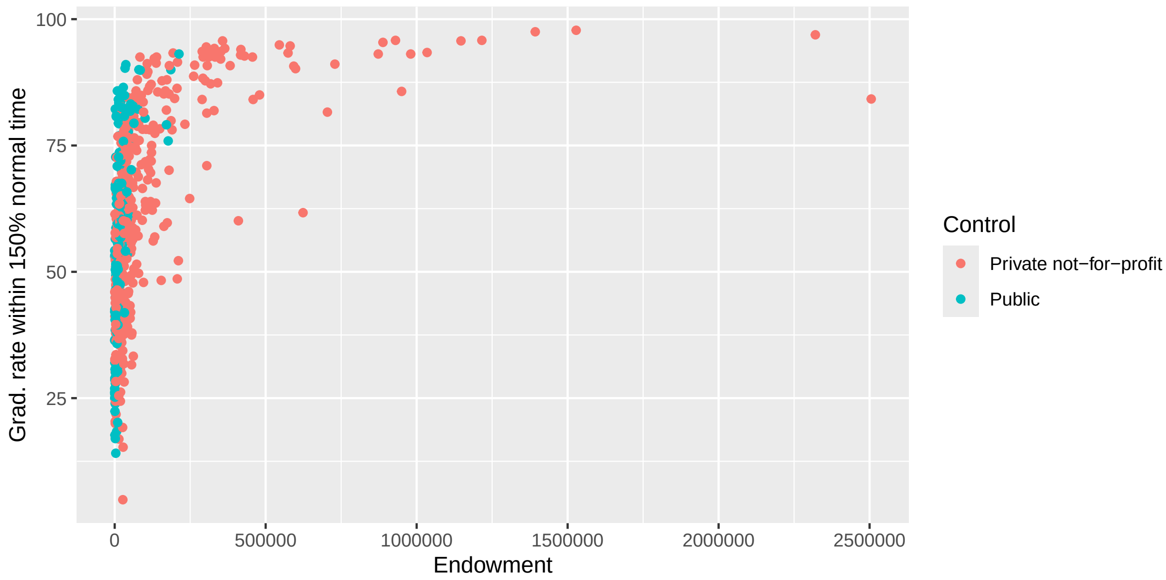
Graduation rates increase as median SAT score increases, generally, and there doesn't seem to be a difference between public and private schools.

```
ggplot(colleges_full, aes(x=endow_value, y=grad_150_value,
                                        col=basic_fac))+
  geom_point()+
  labs(x="Endowment",y="Grad. rate within 150% normal time",
       col="Basic classification")
```



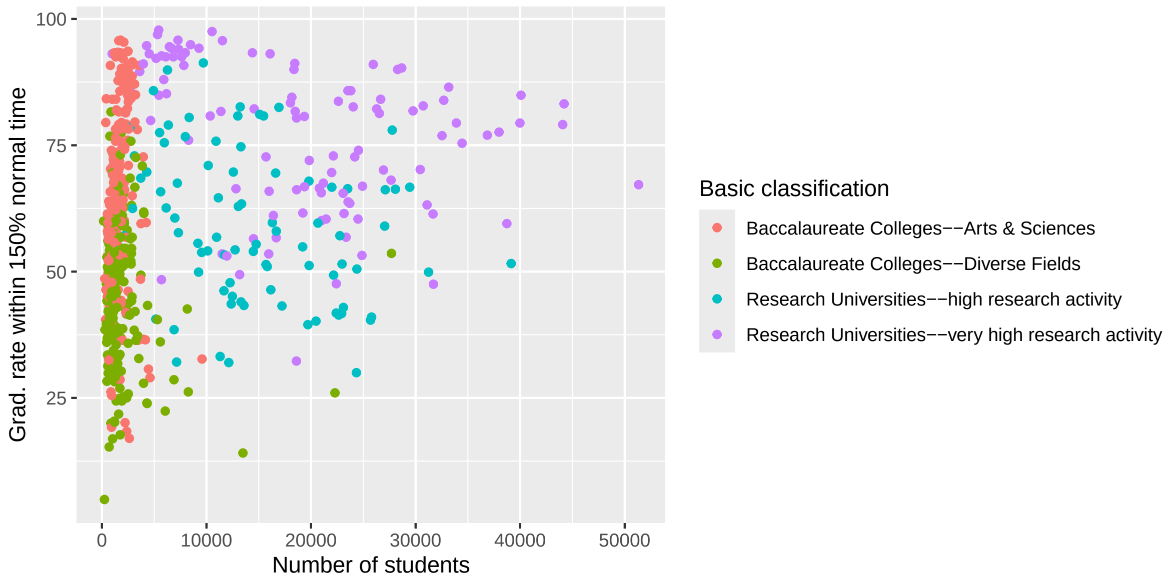
There is not a clear trend because many schools have a small endowment and a few have large endowments.

```
ggplot(colleges_full, aes(x=endow_value, y=grad_150_value,
                                        col=control_fac))+
  geom_point()+
  labs(x="Endowment",y="Grad. rate within 150% normal time",
       col="Control")
```



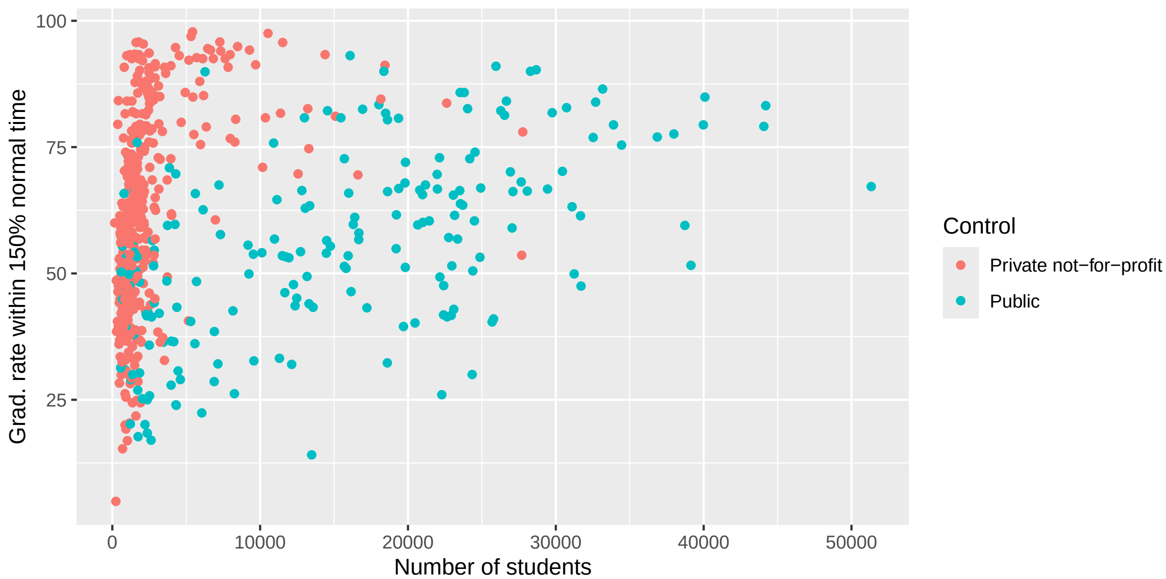
The schools with large endowments are private, and the schools with large endowments also have high graduation rates.

```
ggplot(colleges_full, aes(x=student_count, y=grad_150_value,
                                          col=basic_fac))+
  geom_point()+
  labs(x="Number of students",y="Grad. rate within 150% normal time",
       col="Basic classification")
```



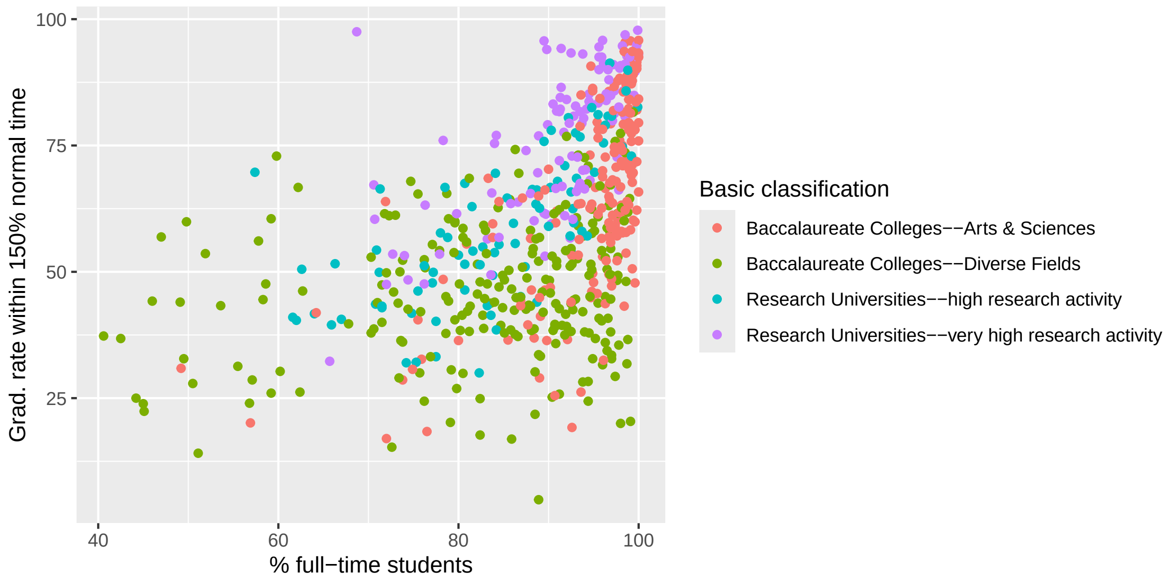
The research universities have higher enrollments, and there doesn't seem to be a clear trend for enrollment count and graduation rates for public vs private schools.

```
ggplot(colleges_full, aes(x=student_count, y=grad_150_value,
                                        col=control_fac))+
  geom_point()+
  labs(x="Number of students",y="Grad. rate within 150% normal time",
       col="Control")
```



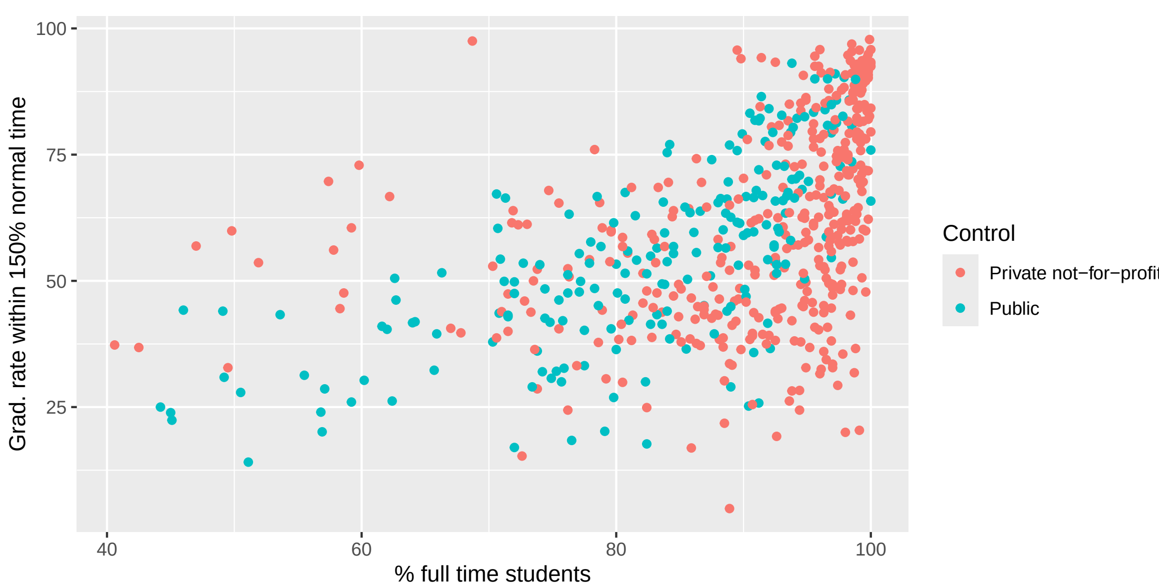
The public schools have higher enrollment numbers, and there doesn't seem to be a clear trend for enrollment count and graduation rates for public vs private schools.

```
ggplot(colleges_full, aes(x=ft_pct, y=grad_150_value,
                                        col=basic_fac))+
  geom_point()+
  labs(x="% full-time students",y="Grad. rate within 150% normal time",
       col="Basic classification")
```



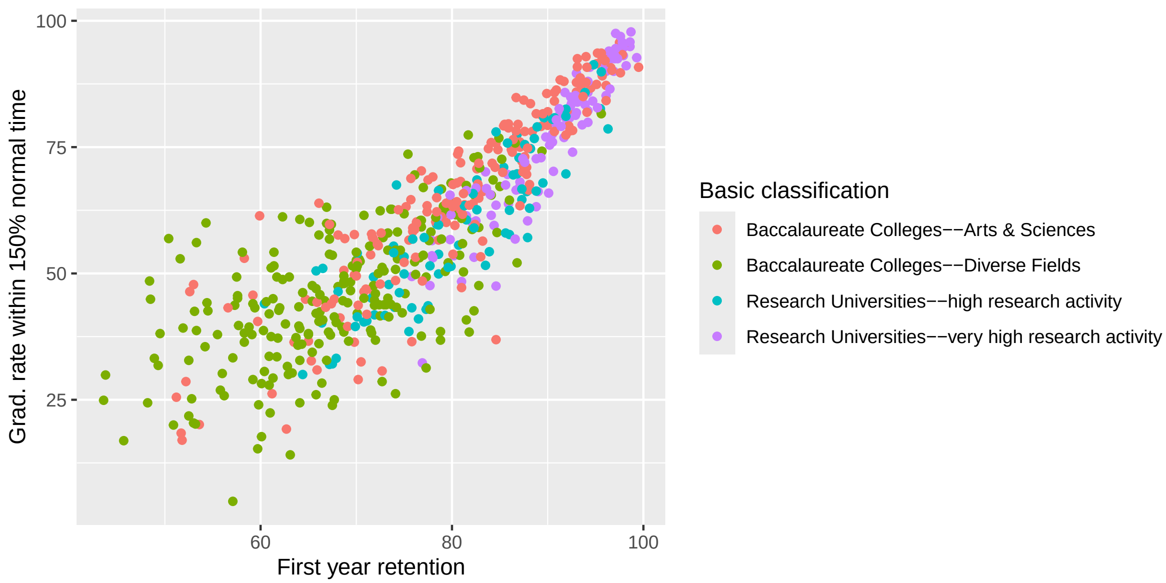
In general, the schools with higher % full-time students have higher graduation rates, though the trend with classification is unclear.

```
ggplot(colleges_full, aes(x=ft_pct, y=grad_150_value,
                                          col=control_fac))+
  geom_point()+
  labs(x="% full time students",y="Grad. rate within 150% normal time",
       col="Control")
```



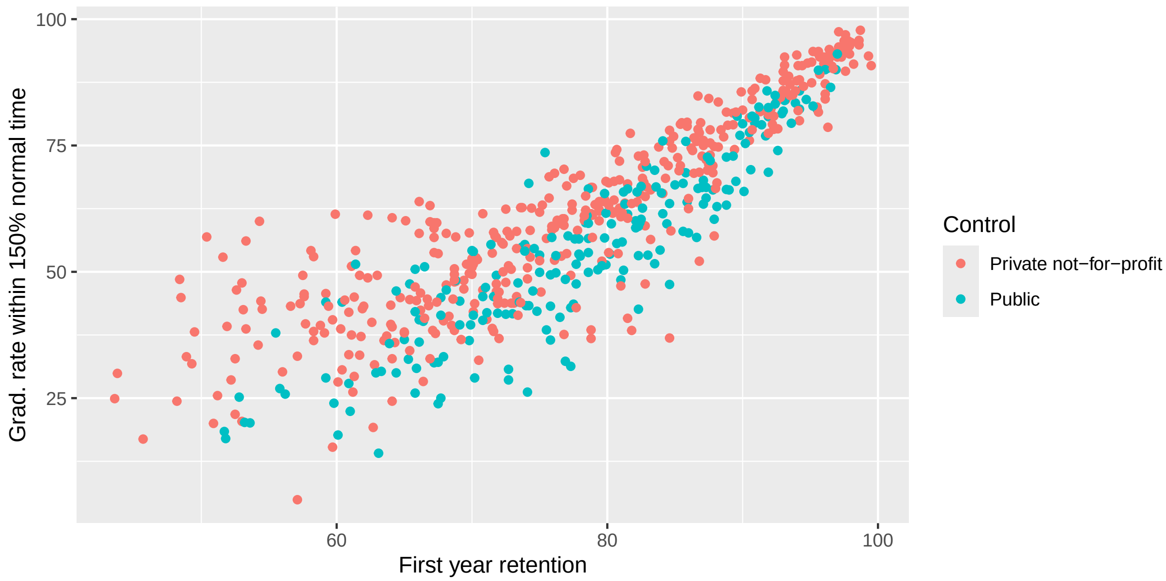
There doesn't seem to be a difference in the relationship between % full-time students and graduation rate for private and public schools.

```
ggplot(colleges_full, aes(x=retain_value, y=grad_150_value,
                                                   col=basic_fac))+
  geom_point()+
  labs(x="First year retention",y="Grad. rate within 150% normal time",
       col="Basic classification")
```



Retention and graduation rate seem to have a strong positive relationship, and the research universities have higher retention and graduation rates.

```
ggplot(colleges_full, aes(x=retain_value, y=grad_150_value,
                                                   col=control_fac))+
  geom_point()+
  labs(x="First year retention",y="Grad. rate within 150% normal time",
       col="Control")
```



There doesn't seem to be a difference in the relationship between retention and graduation rate for private and public schools.

5

Students can choose either to combine the levels to make interpretation easier, or say that it is better to leave the levels as is because there is sufficient sample size in each level. Personally, I would probably combine them, particularly based on the trends seen in the plots.

6

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

where Y = graduation rate within 150% of normal time

x_1 = student count

$x_2 = 1$ if control=Public, 0 otherwise

7

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

where Y = graduation rate within 150% of normal time

x_1 = full-time %

$x_2 = 1$ if control=Public, 0 otherwise

$x_3 =$ retention rate

$x_4 =$ median SAT score

Model for public schools: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5)x_1 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

Model for private schools: $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

Part 2

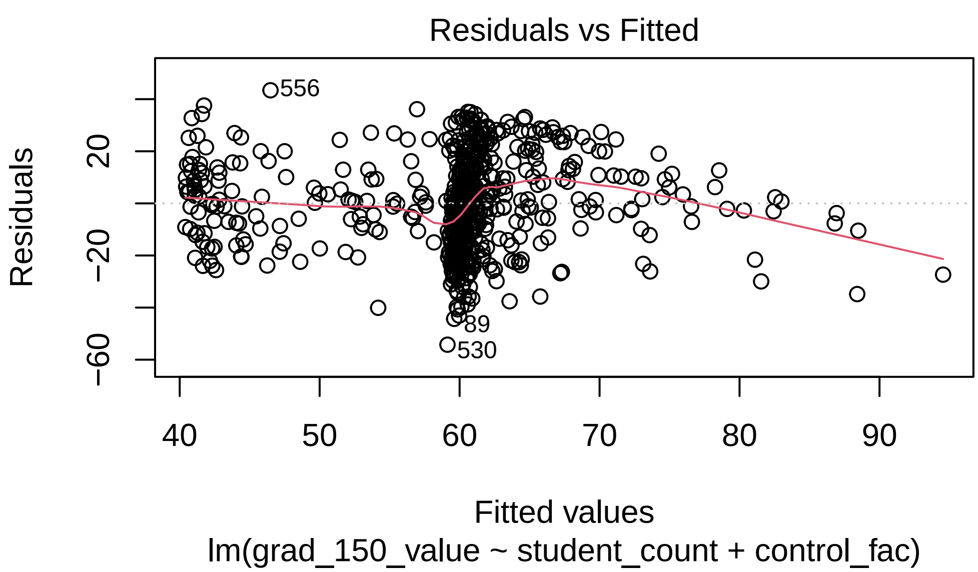
1

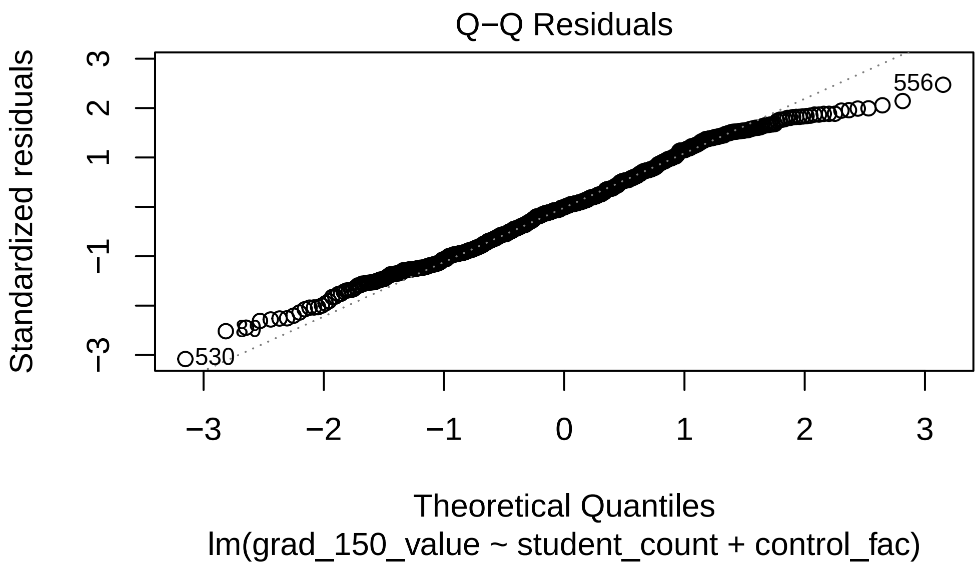
a.

```
collegemod1 <- lm(grad_150_value ~ student_count + control_fac,
                  data=colleges_full)
```

b.

```
plot(collegemod1, which=1:2)
```





i. Linearity appears to be violated. The residual vs fitted plot does not show a random cloud shape. There is a big clump in the middle of the plot

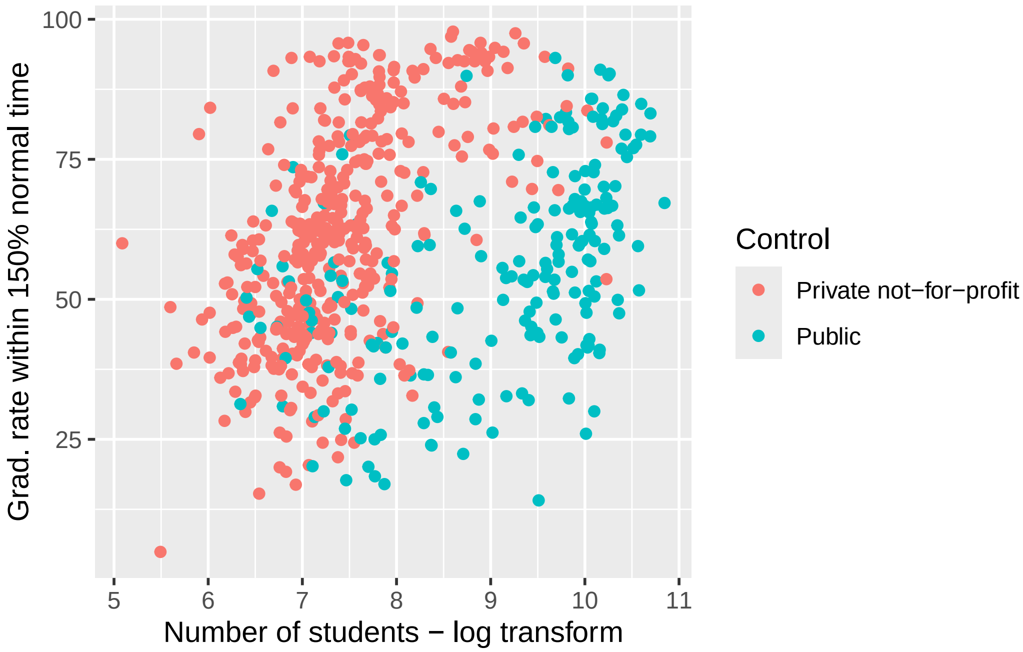
ii. Normality does not appear to be violated based on the QQ-plot. The points fall pretty closely along the 45 degree line.

iii. Homoscedasticity/equal variance appears to be violated, though it's hard to tell since linearity is so bad. But there is a decline in variance moving from left to right in the residual vs fitted plot.

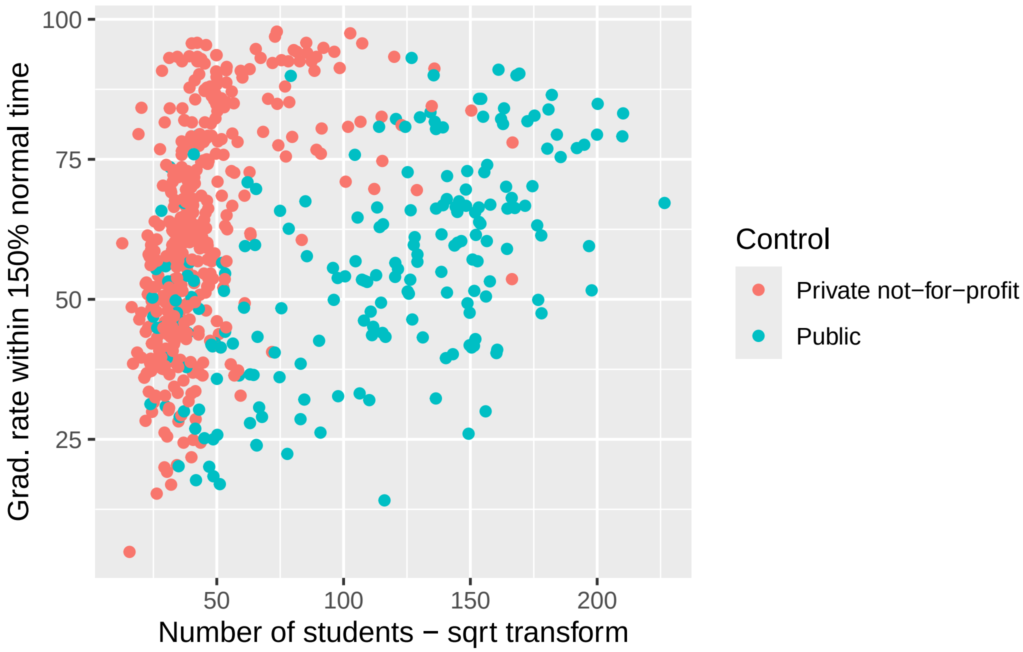
iv. No, this is not surprising based on the relevant plot generated in part 1. In that plot, there is not a linear relationship between student count and graduation rate. Many institutions have a lower enrollment total and a few have higher enrollment totals.

c.

```
ggplot(colleges_full, aes(x=log(student_count), y=grad_150_value,
                                           col=control_fac))+
  geom_point()+
  labs(x="Number of students - log transform",y="Grad. rate within 150% normal time",
       col="Control")
```




```
ggplot(colleges_full, aes(x=sqrt(student_count), y=grad_150_value,
                                      col=control_fac))+
  geom_point()+
  labs(x="Number of students - sqrt transform",y="Grad. rate within 150% normal time",
       col="Control")
```

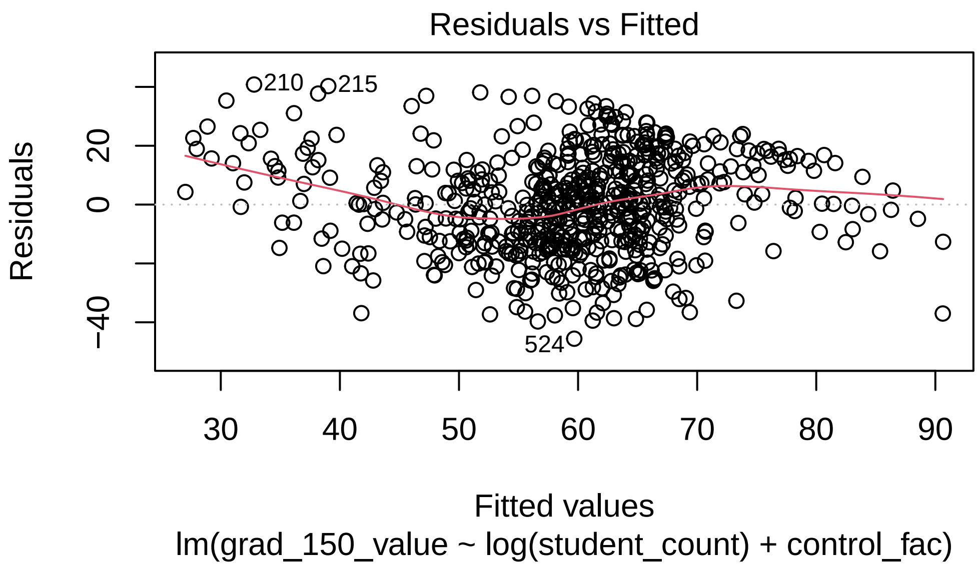


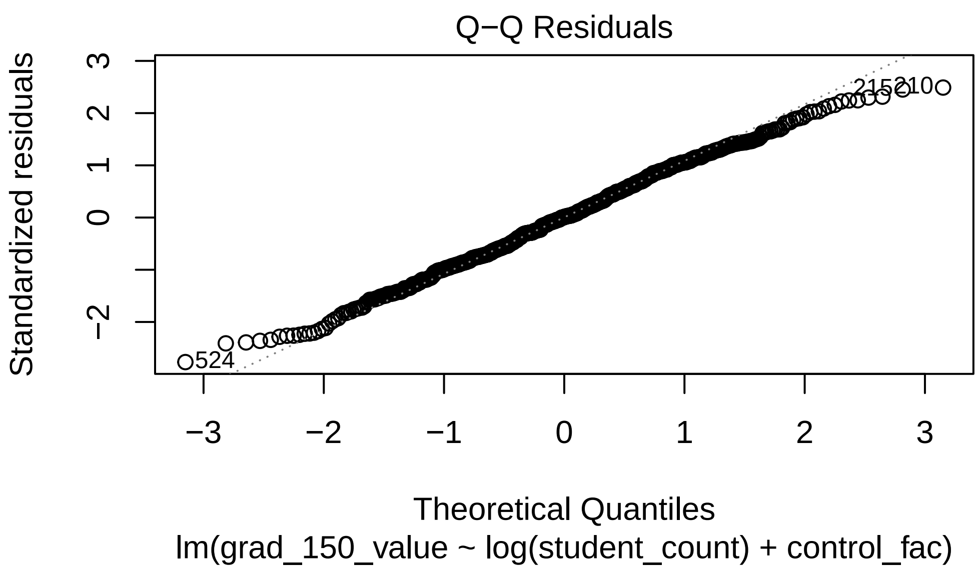
Based on the plots, the log transform seems better to meet the linearity assumption. In the sqrt plot, several of the points for private schools are still clustered together.

d.

```
collegemod_log <- lm(grad_150_value ~ log(student_count) + control_fac,
                    data=colleges_full)

plot(collegemod_log, which=1:2)
```





The residual vs fitted plot is improved compared to the plot in part b, though it's still not a perfect random cloud — many points are clustered in the middle. The QQ-plot is essentially the same as part b, but we already weren't very concerned with normality.

e.

```r
summary(collegemod1)$adj.r.squared
```

```
[1] 0.1713335
```

```r
summary(collegemod_log)$adj.r.squared
```

```
[1] 0.272909
```

f.

```r
summary(collegemod_log)
```

```

Call:
lm(formula = grad_150_value ~ log(student_count) + control_fac,
    data = colleges_full)

Residuals:
    Min      1Q  Median      3Q     Max
-45.579 -11.947   0.193  12.098  40.805

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -14.8284     5.2793  -2.809  0.00513 **
log(student_count)  10.3093     0.7044  14.635  < 2e-16 ***
control_facPublic  -23.5286     1.8489 -12.726  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.51 on 612 degrees of freedom
Multiple R-squared:  0.2753,    Adjusted R-squared:  0.2729
F-statistic: 116.2 on 2 and 612 DF,  p-value: < 2.2e-16
```

```r
confint(collegemod_log)
```

```
                        2.5 %     97.5 %
(Intercept)        -25.196077  -4.460655
log(student_count)   8.925941  11.692756
control_facPublic  -27.159577 -19.897674
```

Because the assumptions are closer to being met and the R^2 value is higher, the model with the log transformation is better, though there is still room for improvement.

Fitted model: $\widehat{\text{Graduation rate}} = -14.8 + 10.3 * \log(\text{student count}) - 23.5 * \text{control=Public}$

Note: Students can define terms separately, but outcome must have a "hat" in the fitted model, transformation for student count should be clear, and level of control should be specified

g.

- Per unit increase in log(student count), graduation rate within 150% of normal time increases by 10.3, on average, controlling for public vs private status. This relationship is statistically significant ($p < 0.001$, 95% CI: [8.9,11.7]). OR Per 1% increase in student count, graduation rate within 150% of normal time increases by 10.3/100=0.1, on average, controlling for public vs private status.
- Public institutions have a graduation rate that is 23.5 percentage points lower than private institutions, on average, controlling for log(student count). This difference is statistically significant ($p < 0.001$, 95% CI: [-27.2,19.9]).

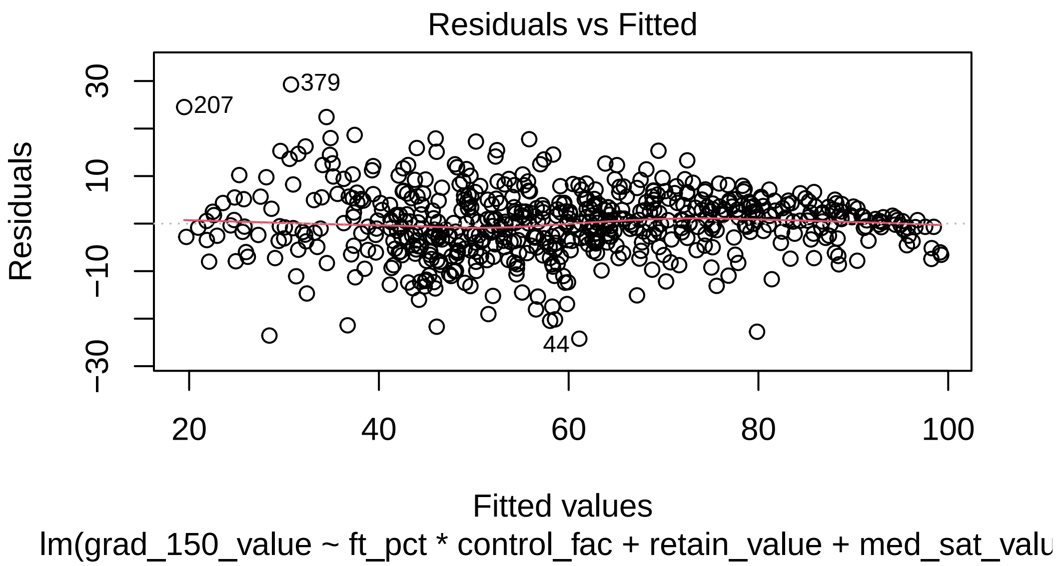
h. Adding more relevant variables should improve the model. It is intuitive that more factors would be at play to explain graduation rate than enrollment and public vs private status. This is evidenced by the relatively low R^2 value of 0.27, meaning about 27% of variability in graduation rate is explained by log(student count) and public vs private control. Adding variables and accounting for interactions could also improve the model diagnostics.

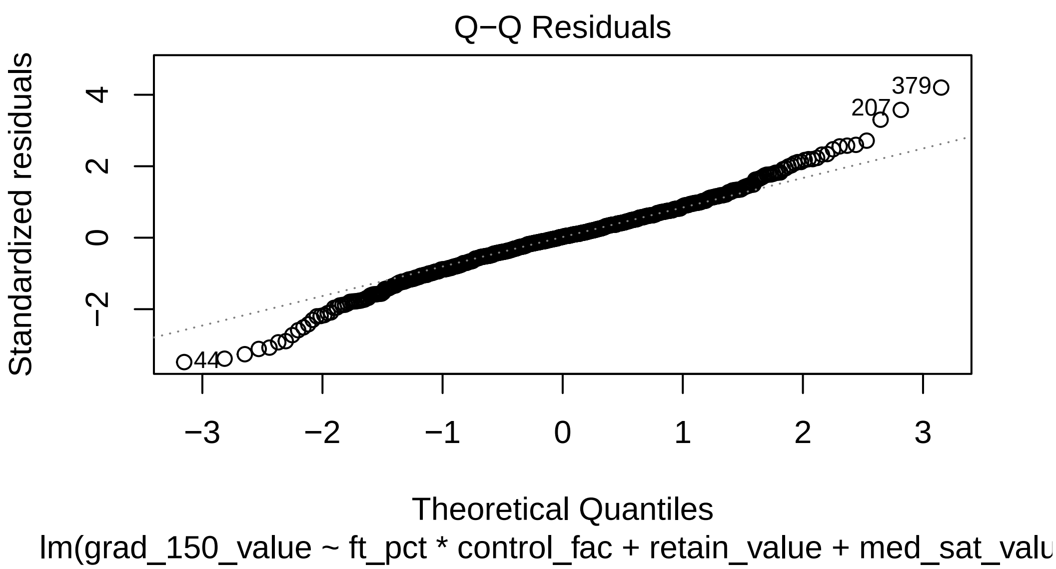
2

a.

```
collegemod2 <- lm(grad_150_value ~ ft_pct*control_fac+retain_value+
                    med_sat_value,
                  data=colleges_full)

plot(collegemod2, which=1:2)
```





i. Based on the residual vs fitted plot, we don't have strong evidence that linearity is violated.
ii. Based on the QQ-plot, we have some evidence that the normality assumption is violated, though maybe not super strong evidence.

iii. Based on the residual vs fitted plot, the homoscedasticity assumption appears to be violated.

b.

```r
summary(collegemod2)
```

```
Call:
lm(formula = grad_150_value ~ ft_pct * control_fac + retain_value +
    med_sat_value, data = colleges_full)

Residuals:
     Min       1Q   Median       3Q      Max
-24.2184  -3.7352   0.1697   4.0129  29.2659

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -57.335049   3.316285 -17.289  < 2e-16 ***
ft_pct                     0.027066   0.036362   0.744    0.457
control_facPublic        -34.008258   4.509121  -7.542 1.69e-13 ***
retain_value               0.686194   0.045425  15.106  < 2e-16 ***
med_sat_value              0.057039   0.003673  15.529  < 2e-16 ***
ft_pct:control_facPublic   0.347799   0.051614   6.738 3.72e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.998 on 609 degrees of freedom
Multiple R-squared:  0.8704,    Adjusted R-squared:  0.8693
F-statistic: 817.7 on 5 and 609 DF,  p-value: < 2.2e-16
```

Public: $\widehat{\text{graduation rate}} = (-57.3 - 34) + (0.03 + .34)\text{full time pct} + 0.69\text{retention rate} + 0.06\text{median SAT score}$

Private: $\widehat{\text{graduation rate}} = -57.3 + 0.03\text{full time pct} + 0.69\text{retention rate} + 0.06\text{median SAT score}$

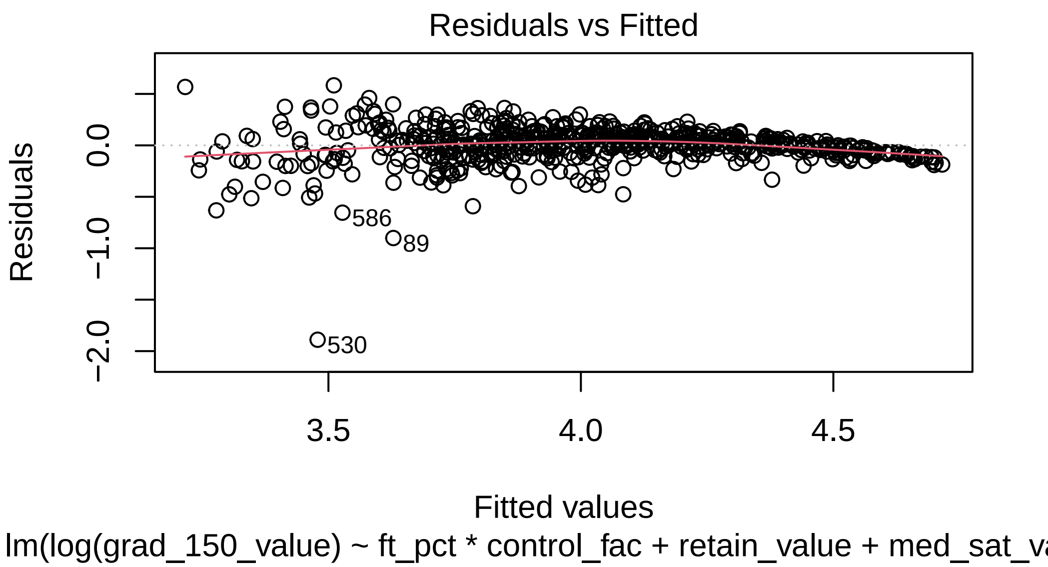
Public institutions have a stronger relationship between full time percentage and graduation rate. The coefficient estimate for full time pct is 0.37 for public institutions compared to 0.03 for private institutions.

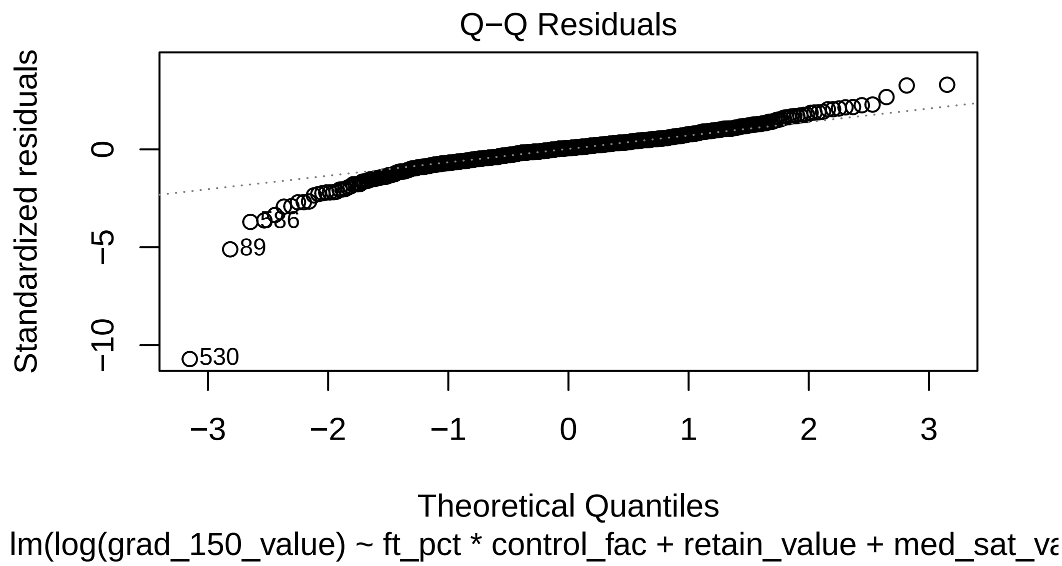
c. The adjusted R^2 is shown in the output above and it is 0.87. This means that about 87% of the variability in graduation rate within 150% of normal time is explained by the predictors in this model.

d.

```r
collegemod2_log <- lm(log(grad_150_value) ~ ft_pct*control_fac+retain_value+
                        med_sat_value,
                      data=colleges_full)

plot(collegemod2_log, which=1:2)
```





There does not seem to be an improvement in the diagnostics using the log transformation of the outcome. Homoscedasticity in particular still seems to be violated. However, the magnitude of the residuals is also much lower, so this could be justified either way (but must be a clear justification).

Bonus

```
library(lmtest)
library(sandwich)

coeftest(collegemod2, vcov. = vcovHC(collegemod2, type = 'HC1'))
```

```
t test of coefficients:

                          Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept)            -57.3350486   4.6683443 -12.2817 < 2.2e-16 ***
ft_pct                   0.0270658   0.0522645   0.5179    0.6047    
control_facPublic      -34.0082580   5.9641201  -5.7021 1.846e-08 ***
retain_value             0.6861938   0.0566032  12.1229 < 2.2e-16 ***
med_sat_value            0.0570392   0.0043432  13.1329 < 2.2e-16 ***
```

```
ft_pct:control_facPublic   0.3477993   0.0658536   5.2814 1.787e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model from number 2 (without the log transformation) is the best candidate for using robust standard errors because this is where heteroscedasticity is the biggest issue. We see that the estimates are nearly identical, but the standard errors are larger using robust standard errors. In this case, any conclusions drawn with p-values would not change using robust standard errors, though we can see that they are generally larger since the standard errors are larger.