

IDS 702 HW 3 Part 1

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.

```
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).

```
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

```
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

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

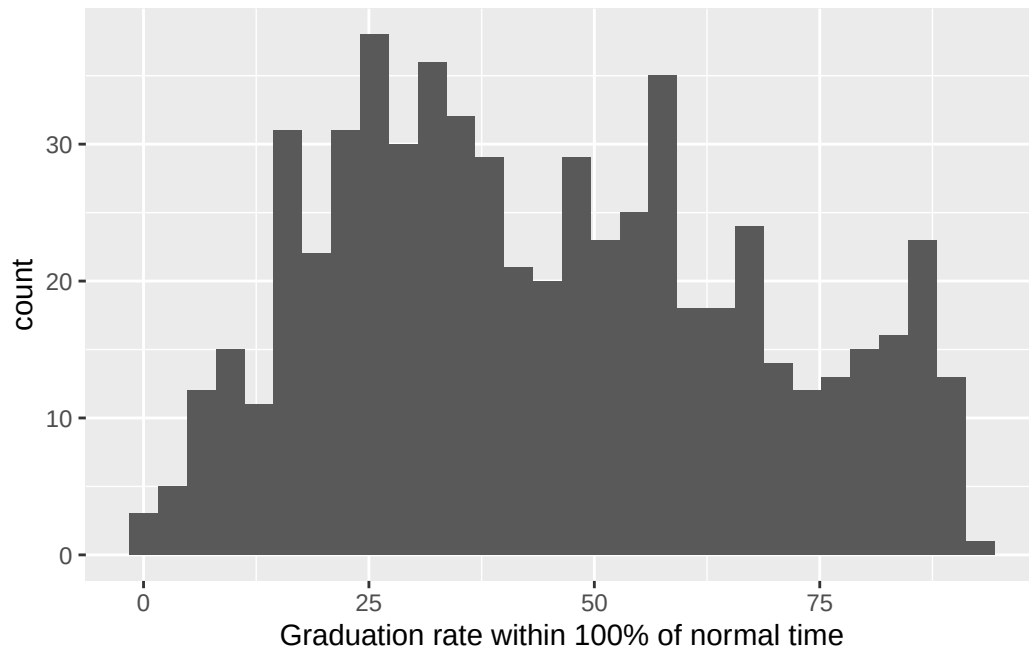
		control	n	prop
1	Private not-for-profit		409	0.6650407
2	Public		206	0.3349593

	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.

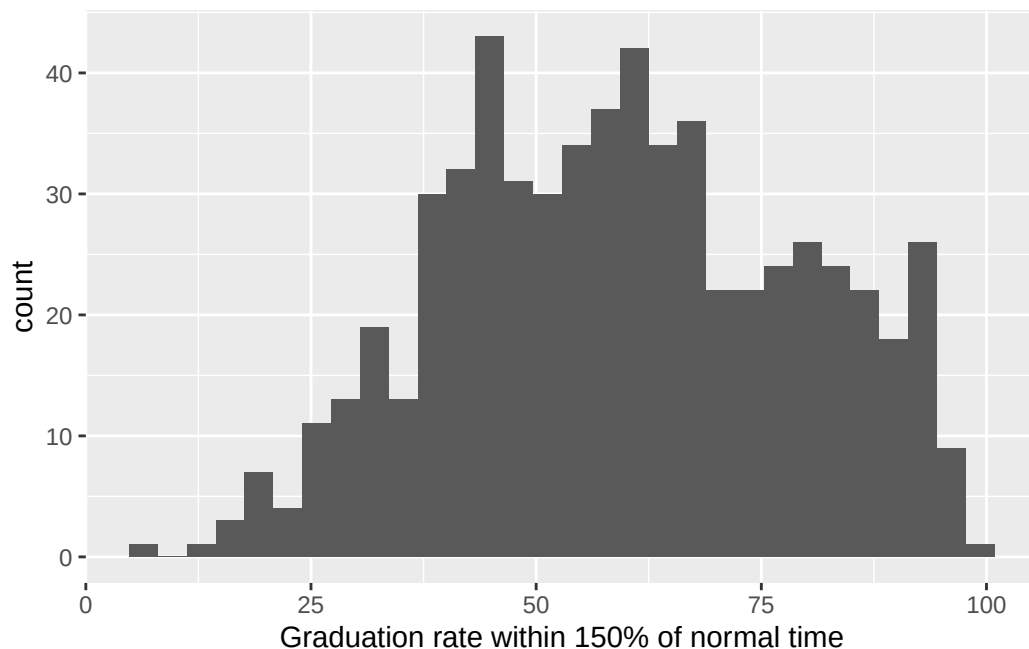
```
ggplot(colleges_full, aes(x=grad_100_value))+
  geom_histogram()+
  labs(x="Graduation rate within 100% of normal time")
```



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")]
```

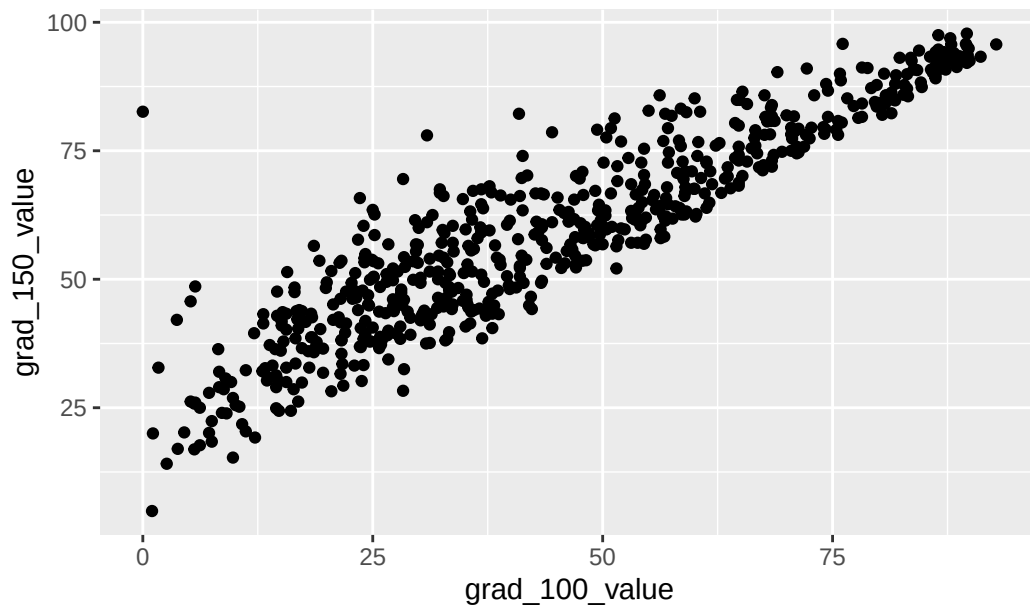
```
      grad_100_value grad_150_value
237              0          82.6
```

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 `grad_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")
```

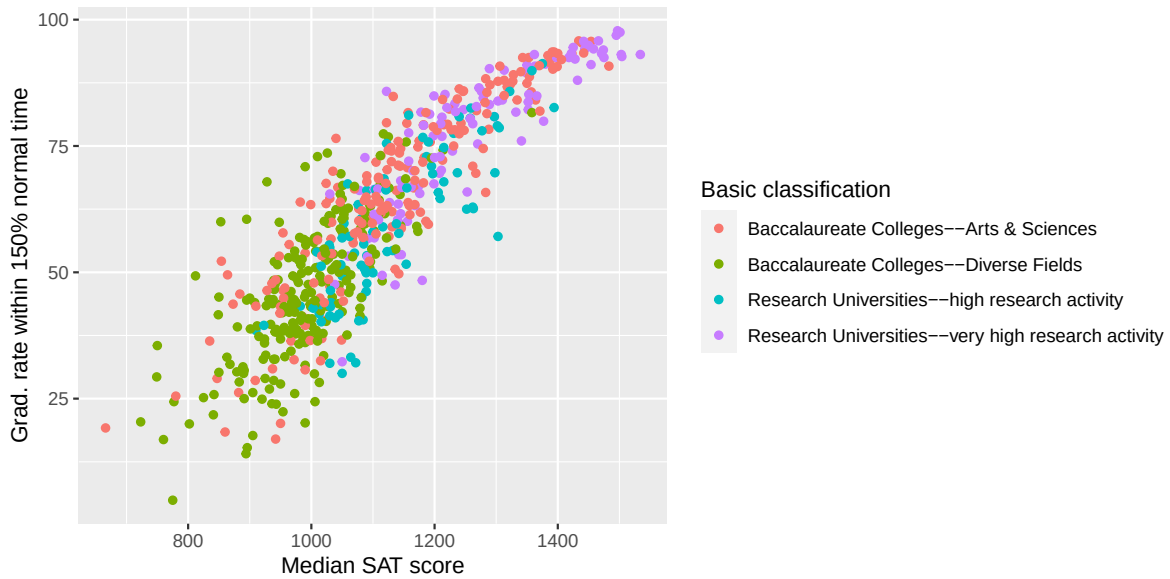
Comparing values of graduation time variables



4

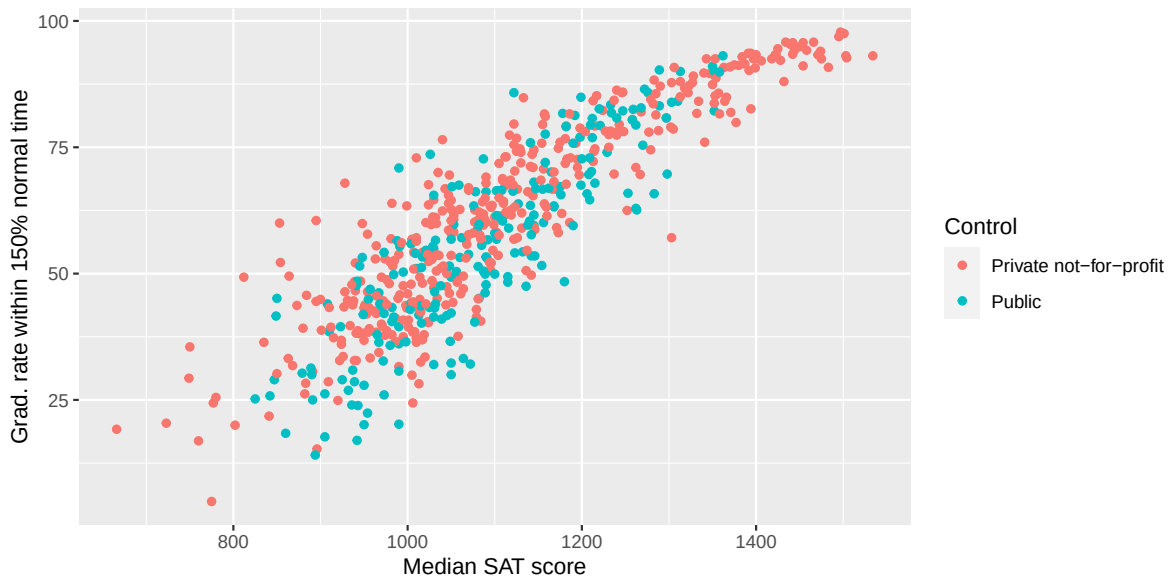
```
colleges_full$basic_fac <- factor(colleges_full$basic)
colleges_full$control_fac <- factor(colleges_full$control)

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")
```



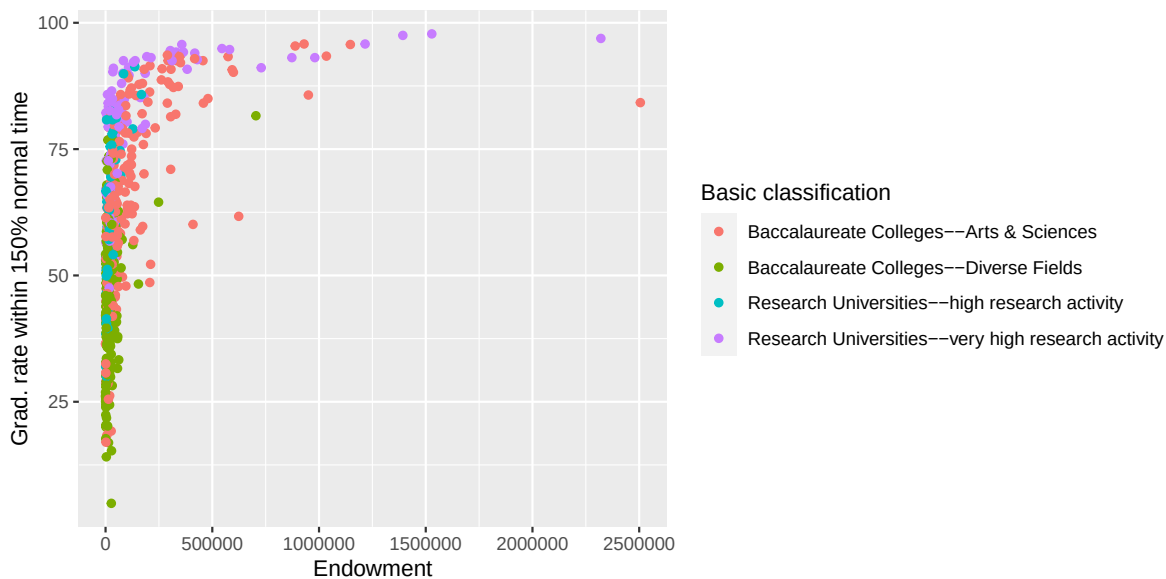
Graduation rates increase as median SAT score increases, generally, and the research universities tend to have higher median SAT scores and graduation rates.

```
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")
```



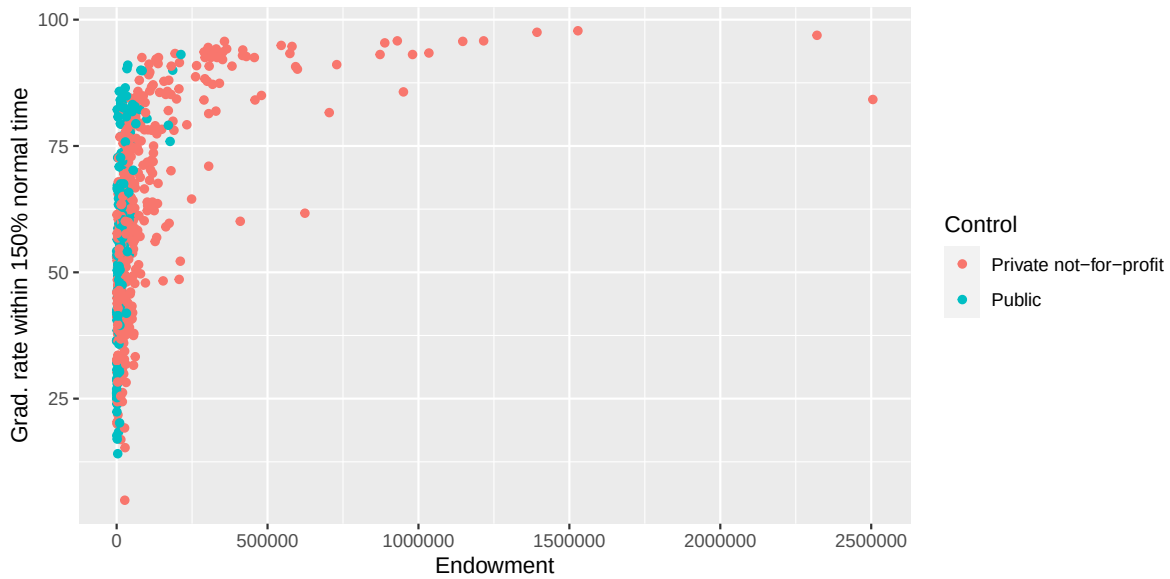
Graduation rates increase as median SAT score increases, generally, and there doesn't seem to be a difference between public and private schools.

```
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")
```



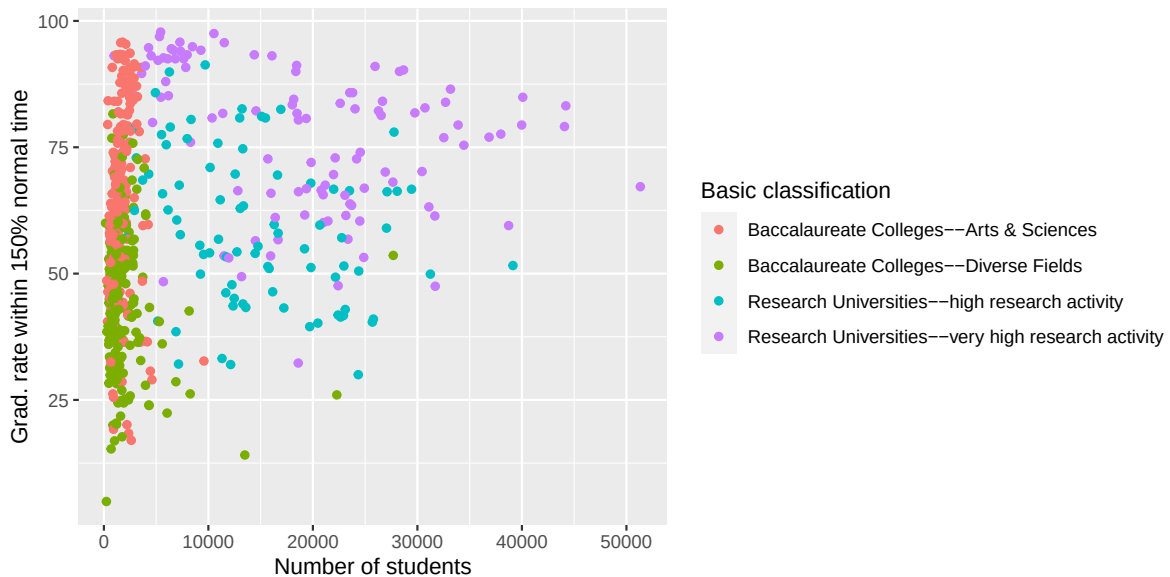
There is not a clear trend because many schools have a small endowment and a few have large endowments.

```
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")
```

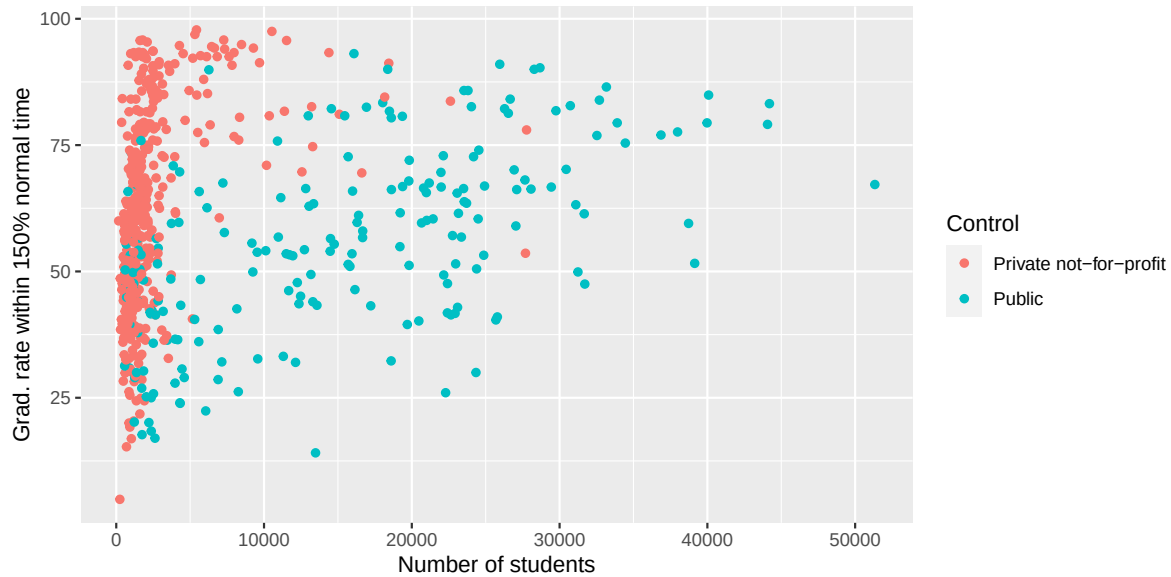
The schools with large endowments are private, and the schools with large endowments also have high graduation rates.

```
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")
```



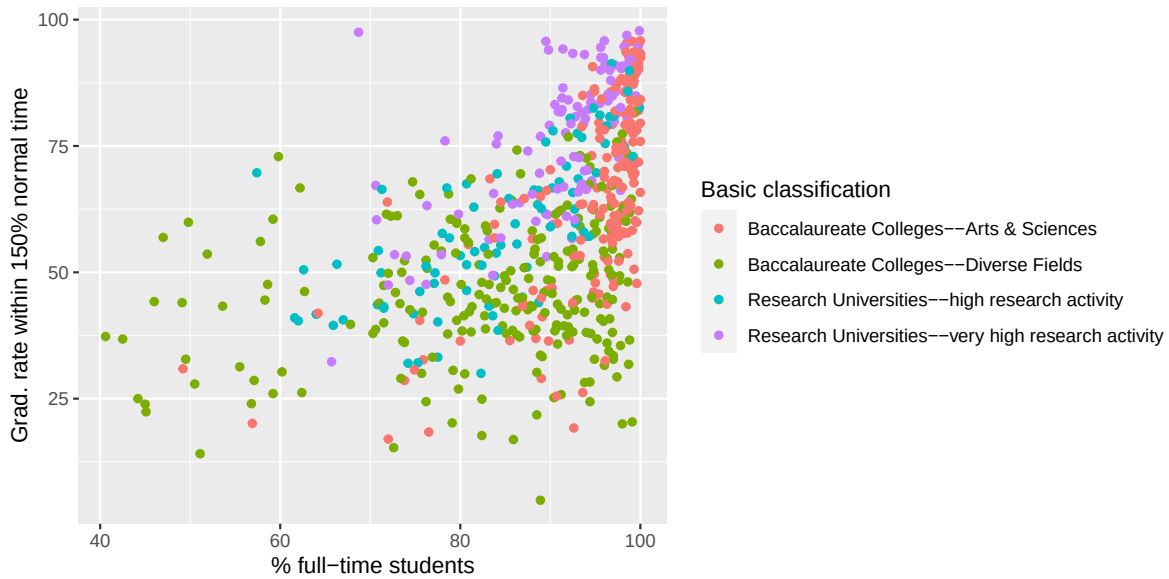
The research universities have higher enrollments, and there doesn't seem to be a clear trend for enrollment count and graduation rates for public vs private schools.

```
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")
```



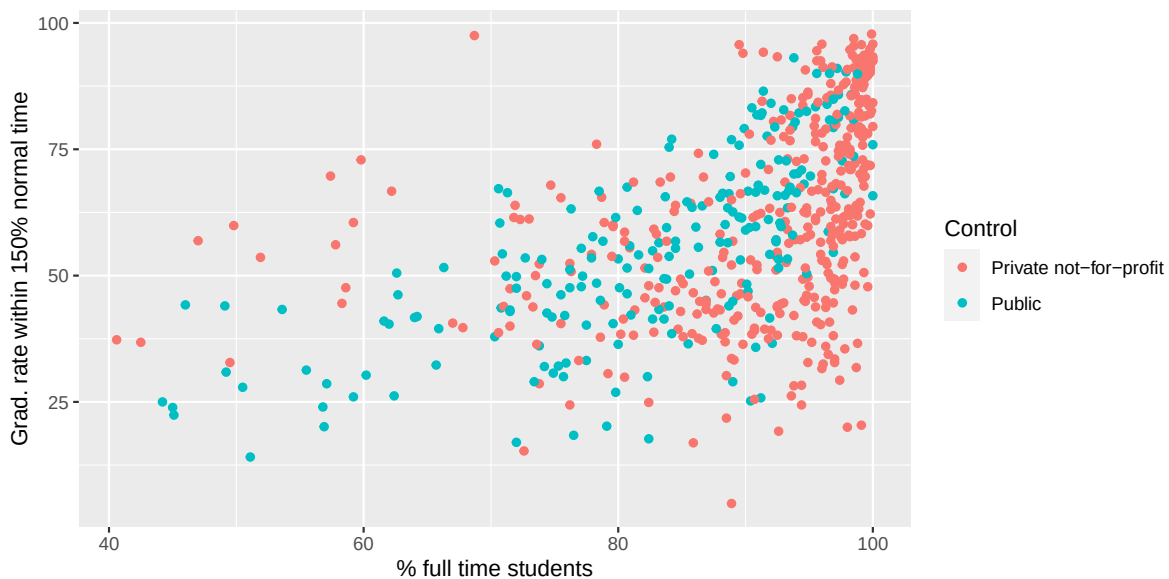
The public schools have higher enrollment numbers, and there doesn't seem to be a clear trend for enrollment count and graduation rates for public vs private schools.

```
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")
```



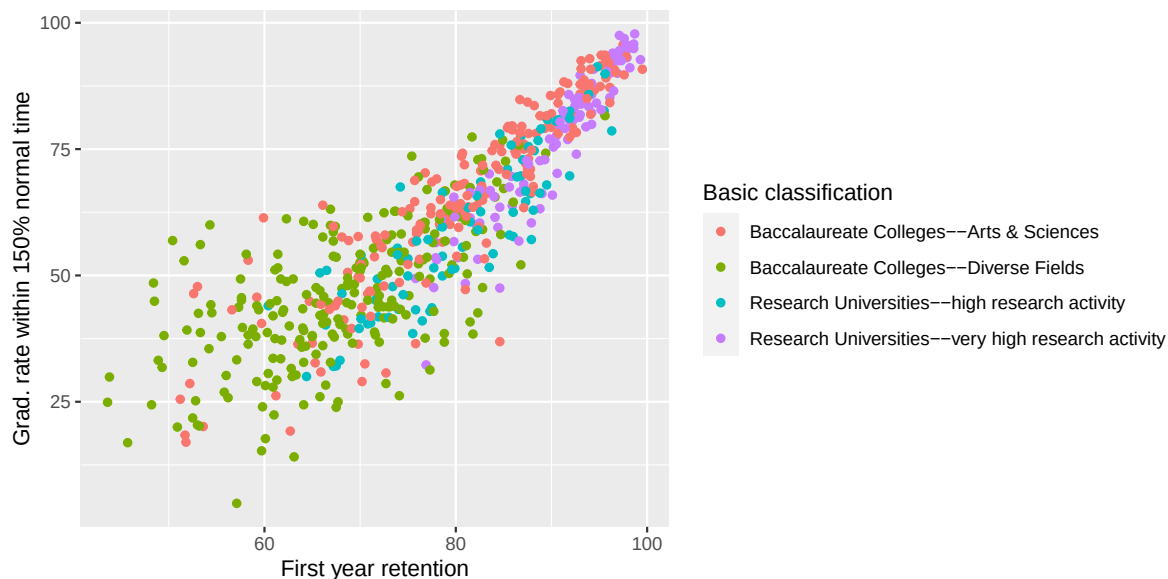
In general, the schools with higher % full-time students have higher graduation rates, though the trend with classification is unclear.

```
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: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5)x_1 + \beta_3x_3 + \beta_4x_4 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

Model for private schools: $Y = \beta_0 + \beta_1x_1 + \beta_3x_3 + \beta_4x_4 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$