ggplot2_Geometry Type

Bing-Je_Wu

8/17/2019

Outlines

- Scatterplots
- Lines and Smoothers
- Bars and Columns
- Histograms
- Boxplots

Scatterplots

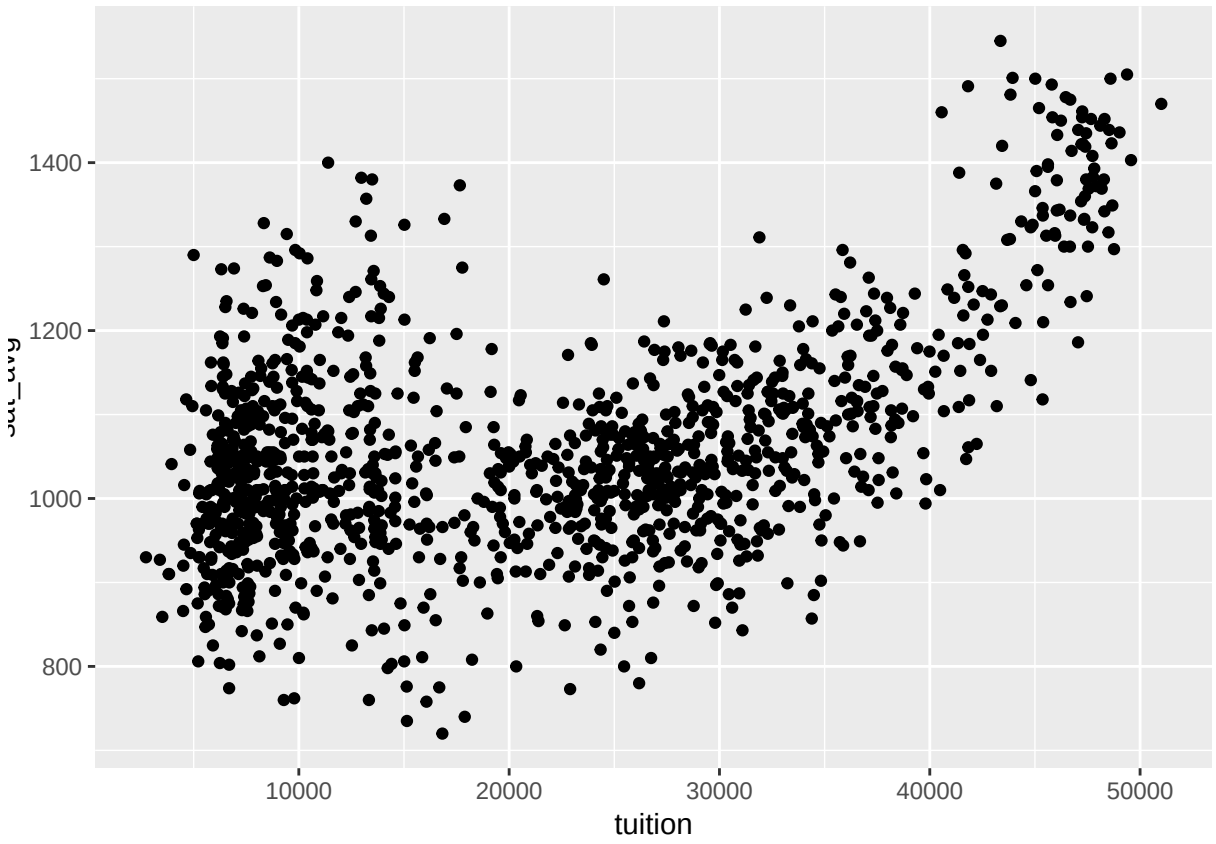
Load the datase

```
library(tidyverse)
college <- read_csv('http://672258.youcanlearnit.net/college.csv')
college <- college %>%
  mutate(state=as.factor(state), region=as.factor(region),
         highest_degree=as.factor(highest_degree),
         control=as.factor(control), gender=as.factor(gender),
         loan_default_rate=as.numeric(loan_default_rate))
colSums(is.na(college))
```

```
##                  id                name                city
##                   0                   0                   0
##               state              region      highest_degree
##                   0                   0                   0
##             control              gender      admission_rate
##                   0                   0                   0
##             sat_avg         undergrads             tuition
##                   0                   0                   0
## faculty_salary_avg   loan_default_rate         median_debt
##                   0                   2                   0
##                 lon                 lat
##                   0                   0
```

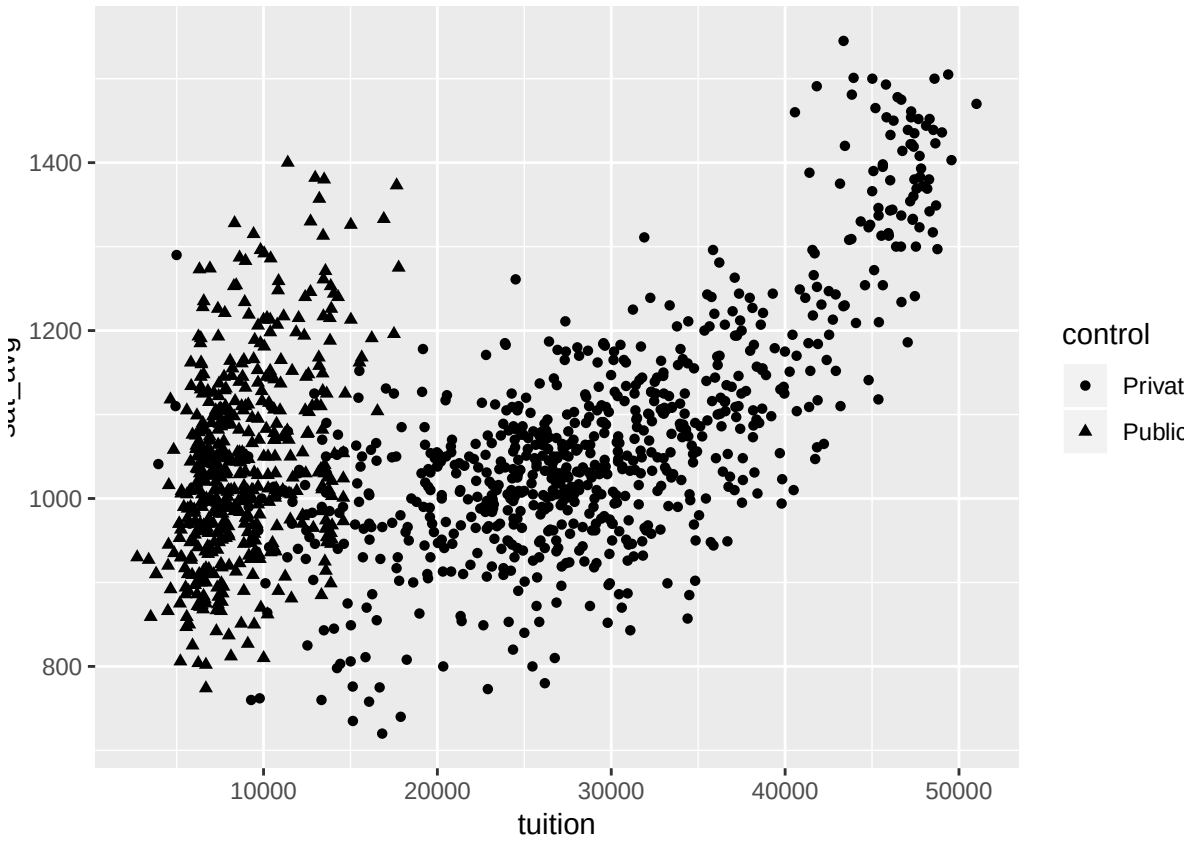
Simple scatterplot

```
ggplot(data=college) +
  geom_point(mapping=aes(x=tuition, y=sat_avg))
```



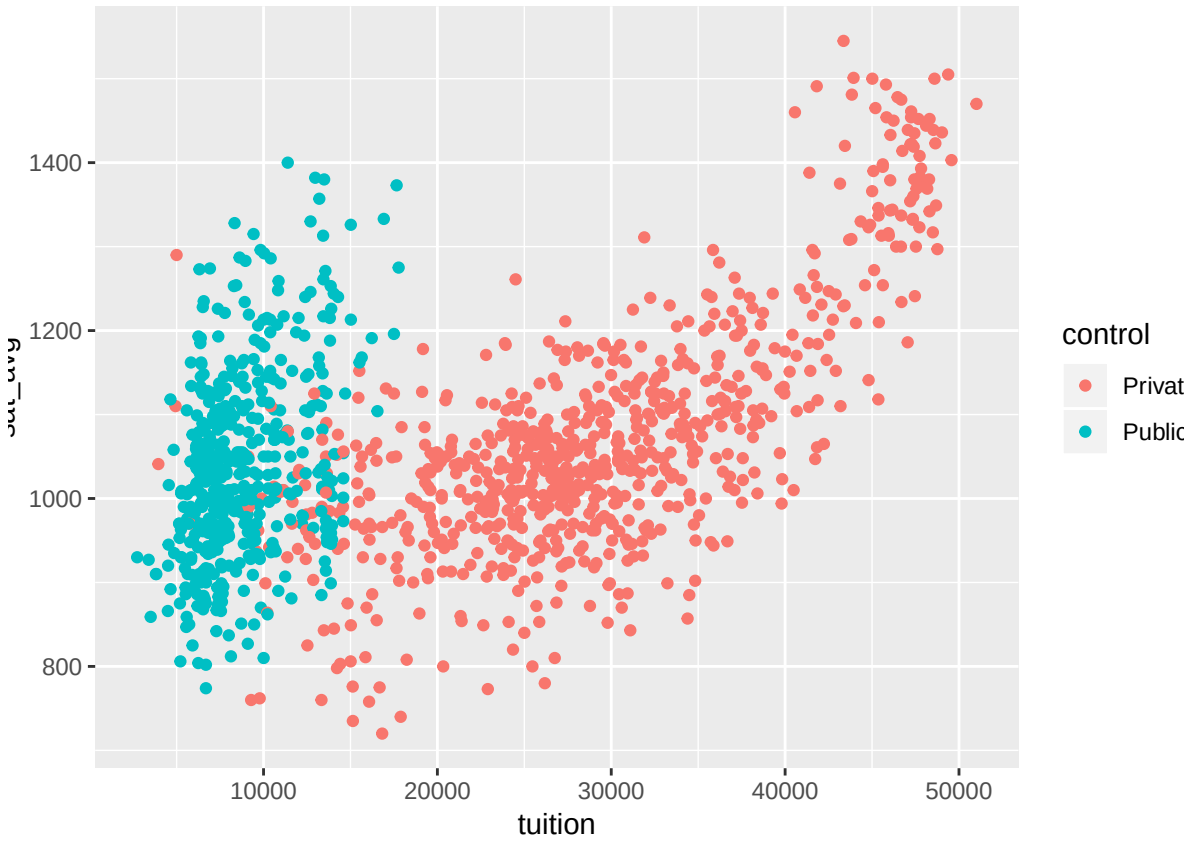
We can do this using the shape attribute

```
ggplot(data=college) +
  geom_point(mapping=aes(x=tuition, y=sat_avg, shape=control))
```



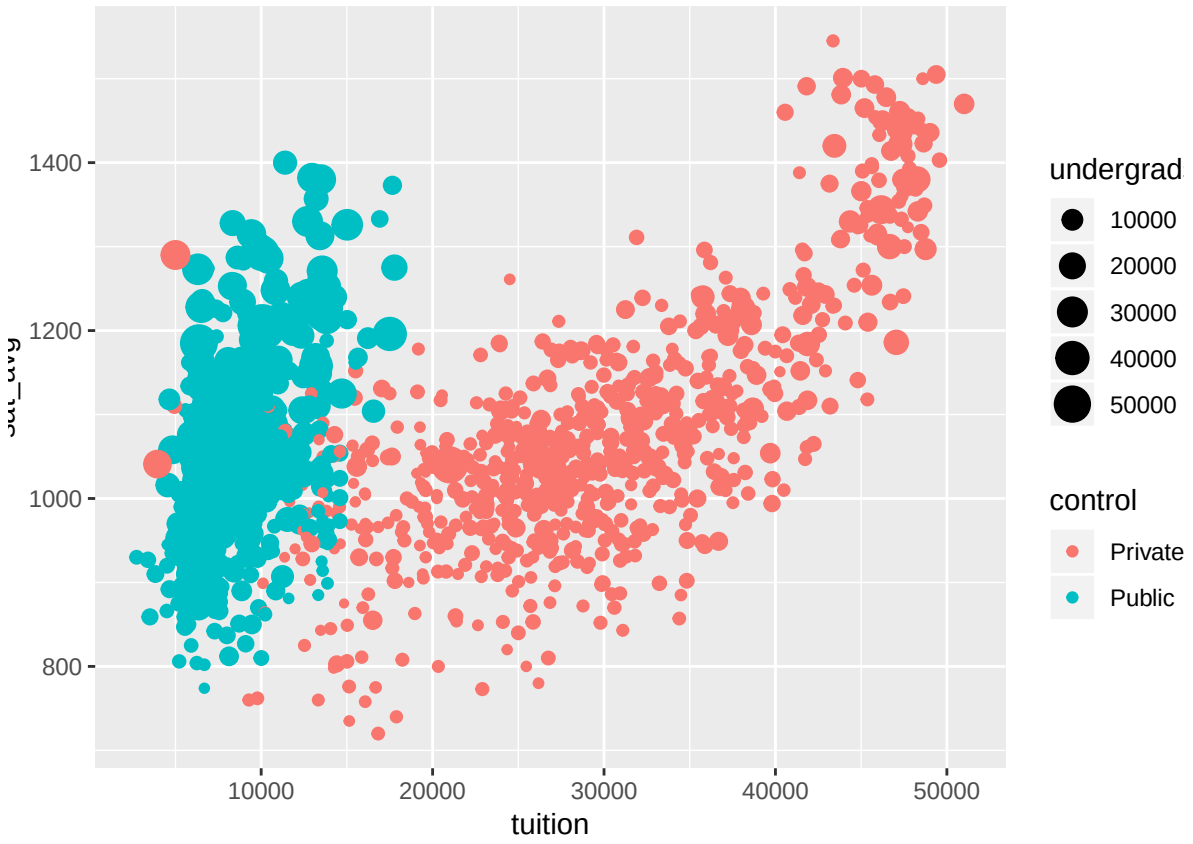
We can use color

```
ggplot(data=college) +
  geom_point(mapping=aes(x=tuition, y=sat_avg, color=control))
```



We can use color to represent the number of students

```
ggplot(data=college) +
  geom_point(mapping=aes(x=tuition, y=sat_avg, color=control, size=undergrads))
```



And, lastly, we ca add some transparency so we can see through those points a bit

```
ggplot(data=college) +
  geom_point(mapping=aes(x=tuition, y=sat_avg, color=control, size=undergrads), alpha=0.5)
```

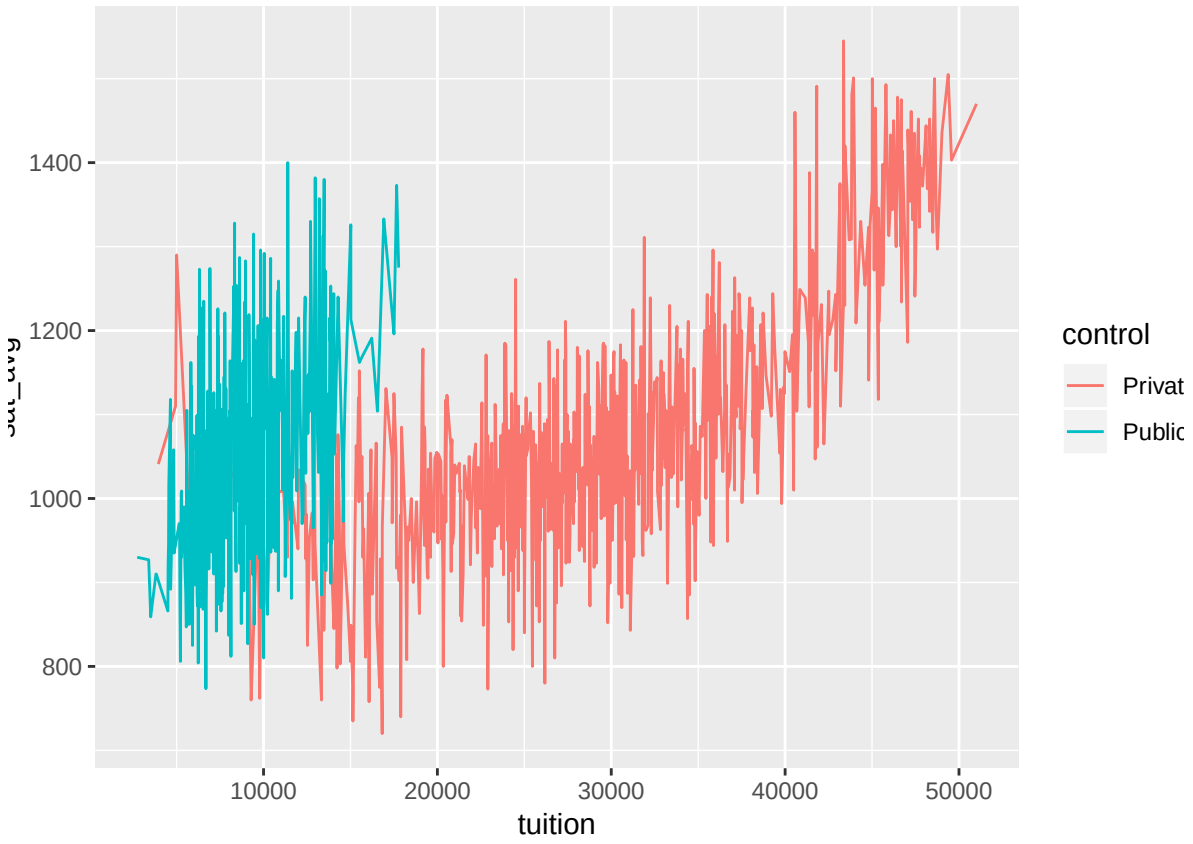
Lines and Smoothers

Load the dataset

```
library(tidyverse)
college <- read_csv('http://672258.youcanlearnit.net/college.csv')
college <- college %>%
  mutate(state=as.factor(state), region=as.factor(region),
         highest_degree=as.factor(highest_degree),
         control=as.factor(control), gender=as.factor(gender),
         loan_default_rate=as.numeric(loan_default_rate))
```

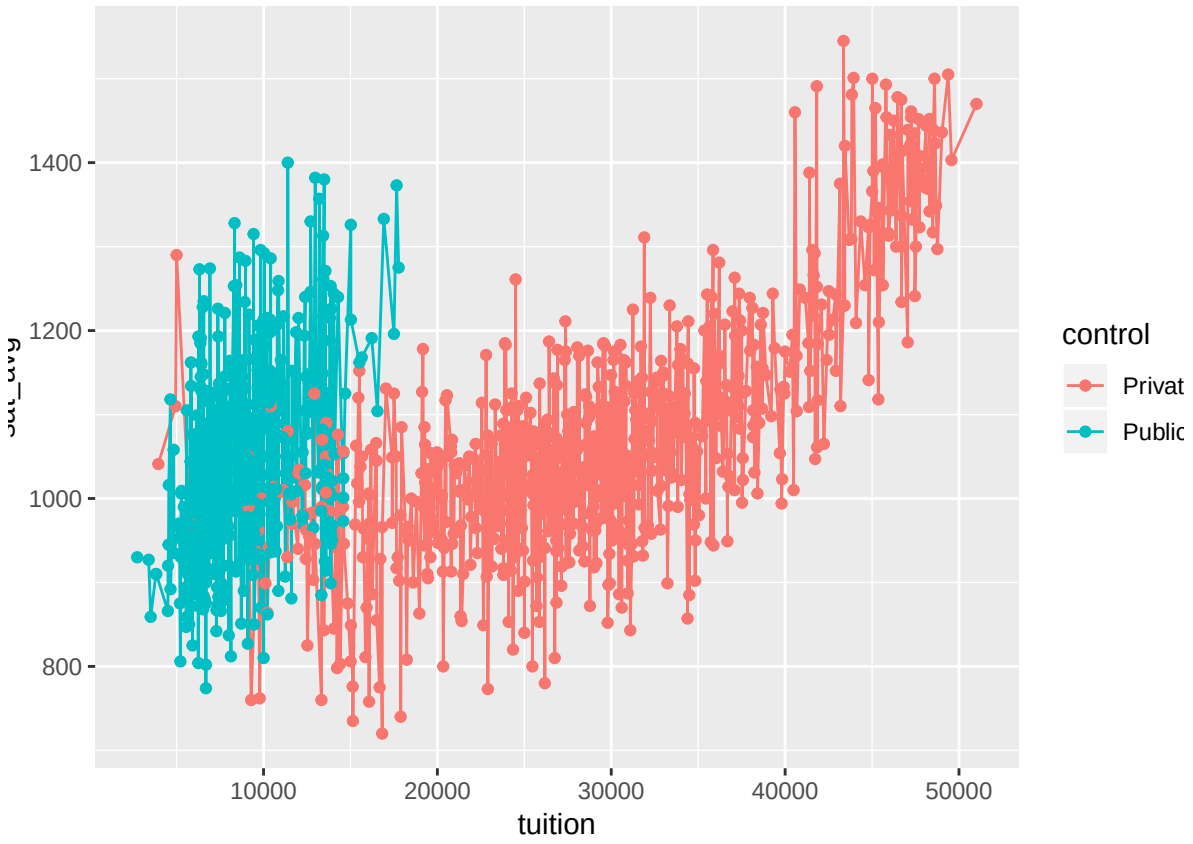
We use line graph instead of scatterplots

```
ggplot(data=college) +
  geom_line(mapping=aes(x=tuition, y=sat_avg, color=control))
```



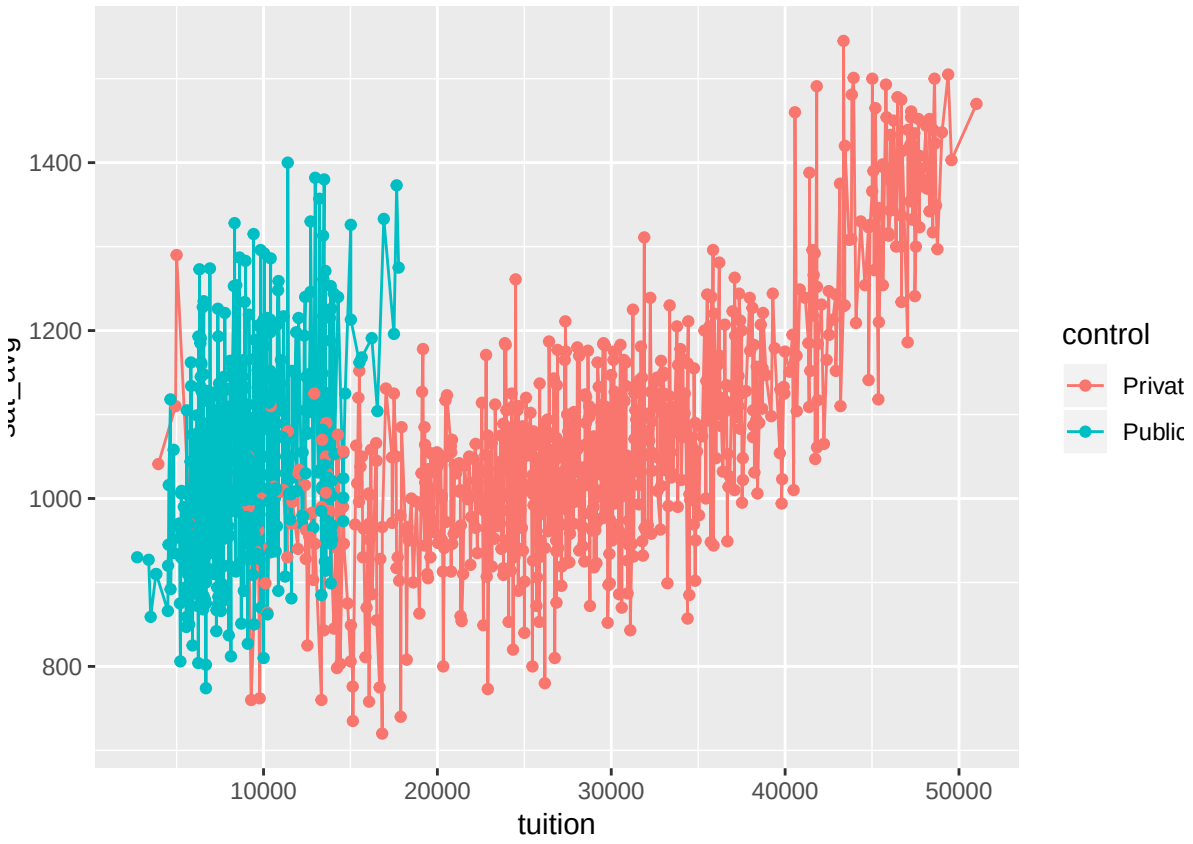
Add oints back in

```
ggplot(data=college) +
  geom_line(mapping=aes(x=tuition, y=sat_avg, color=control)) +
  geom_point(mapping=aes(x=tuition, y=sat_avg, color=control))
```



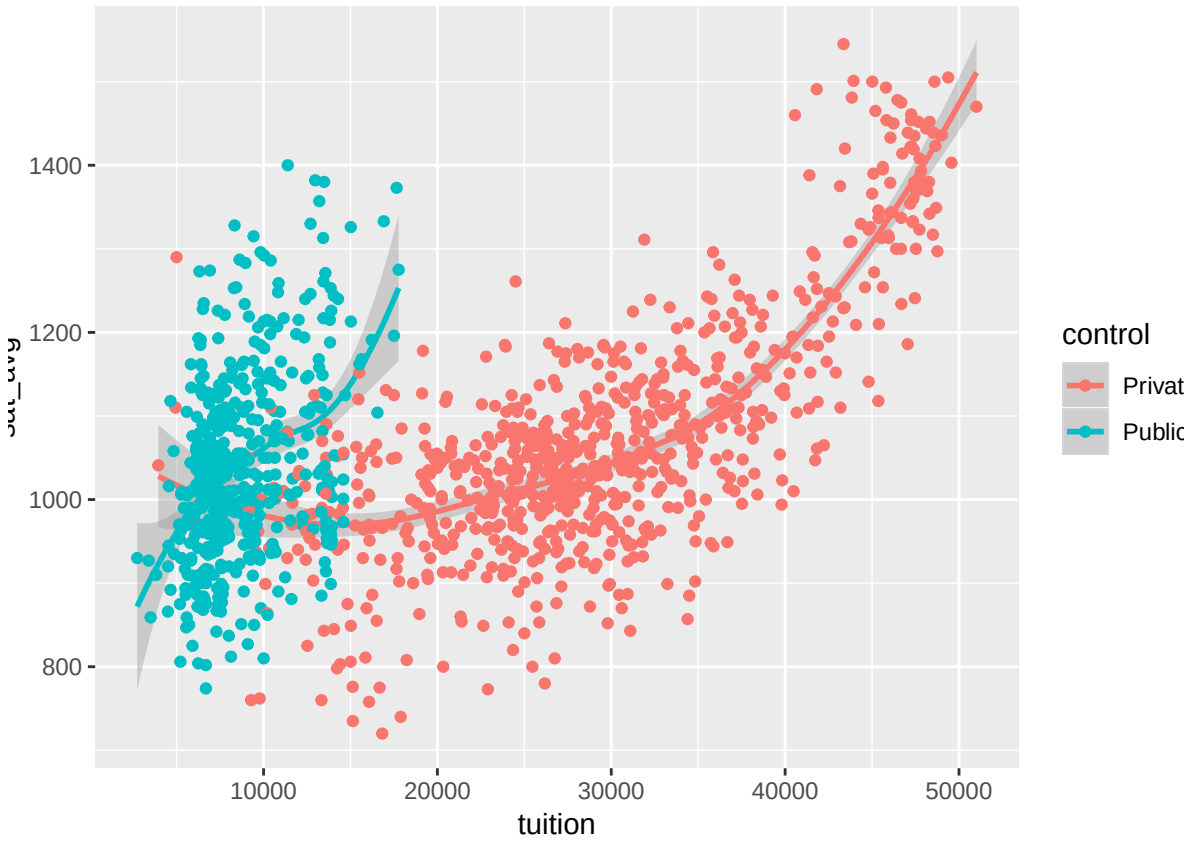
Alternative code

```
ggplot(data=college, mapping=aes(x=tuition, y=sat_avg, color=control)) +
  geom_line() +
  geom_point()
```



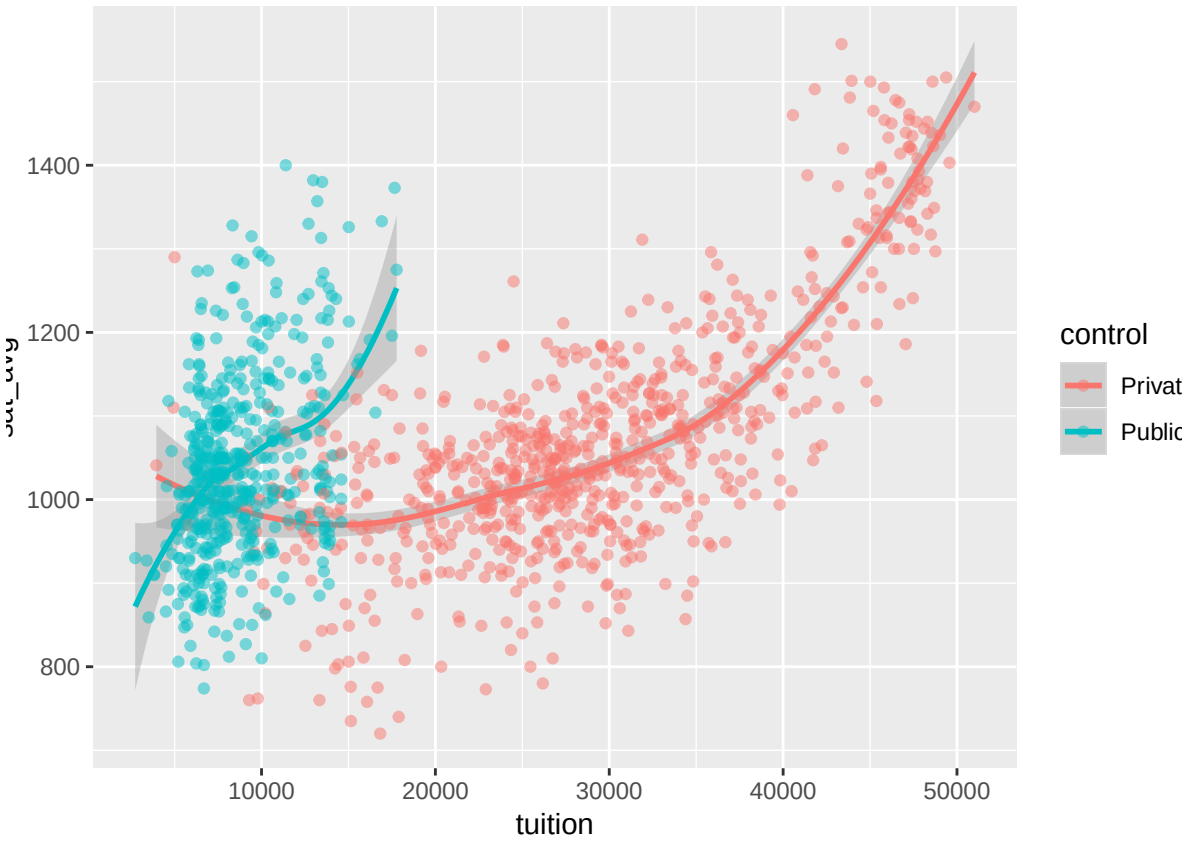
Add geom_smooth geometry to fit a line instead of connecting every point

```
ggplot(data=college, mapping=aes(x=tuition, y=sat_avg, color=control)) +
  geom_smooth() +
  geom_point()
```



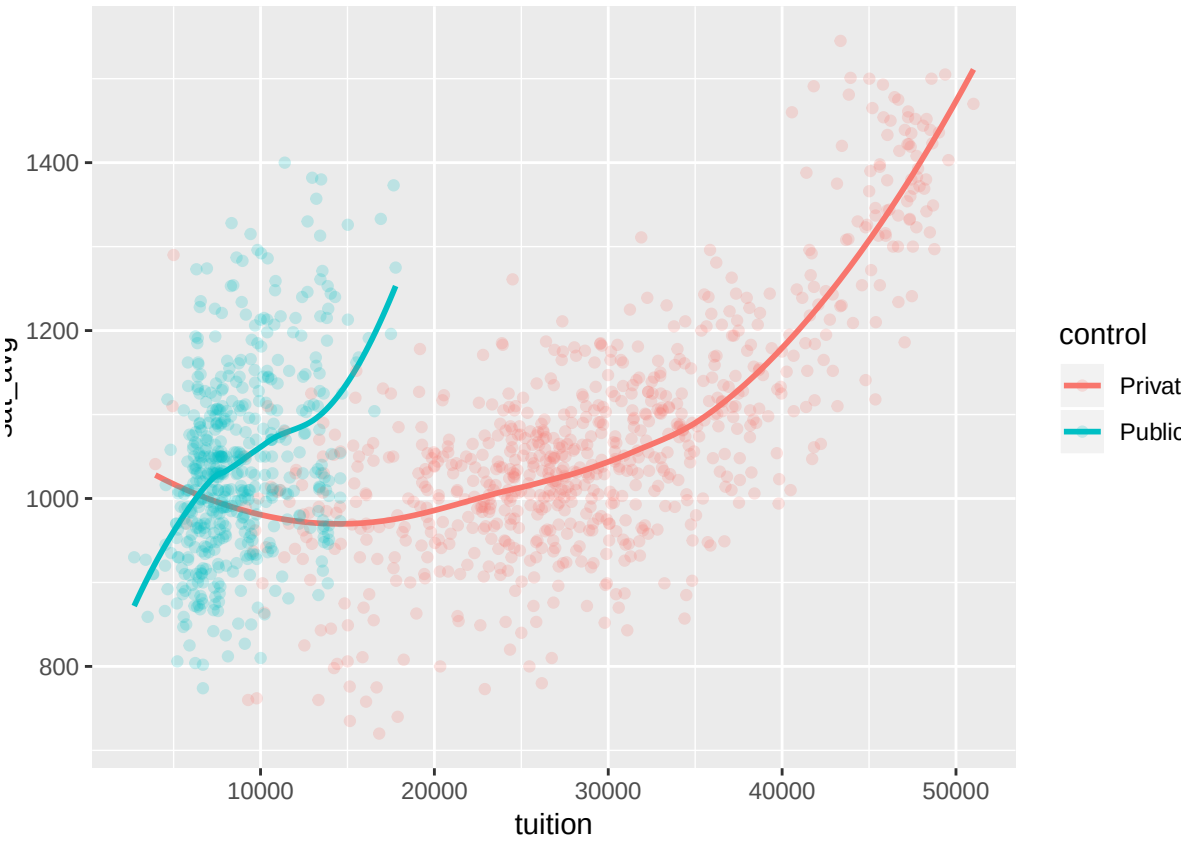
Add transparency to just the points to make the line stand out more

```
ggplot(data=college, mapping=aes(x=tuition, y=sat_avg, color=control)) +
  geom_smooth() +
  geom_point(alpha=0.5)
```



Remove the confidence interval from the smoother

```
ggplot(data=college, mapping=aes(x=tuition, y=sat_avg, color=control)) +
  geom_smooth(se=FALSE) +
  geom_point(alpha=1/5)
```



Bars and Columns

bar graph: uses count as the y-axis value

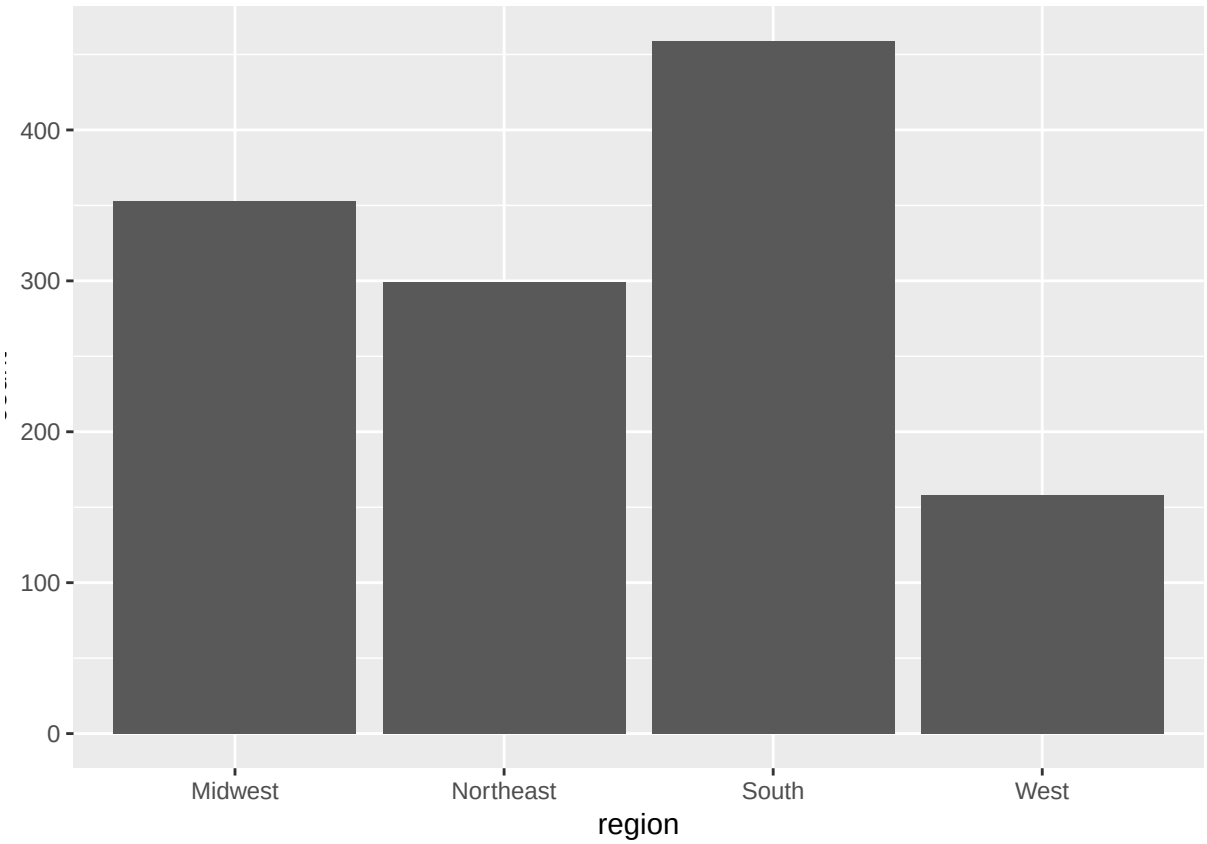
col graph: allows user to specify y-axis value

Load the dataset

```
library(tidyverse)
college <- read_csv('http://672258.youcanlearnit.net/college.csv')
college <- college %>%
  mutate(state=as.factor(state), region=as.factor(region),
         highest_degree=as.factor(highest_degree),
         control=as.factor(control), gender=as.factor(gender),
         loan_default_rate=as.numeric(loan_default_rate))
```

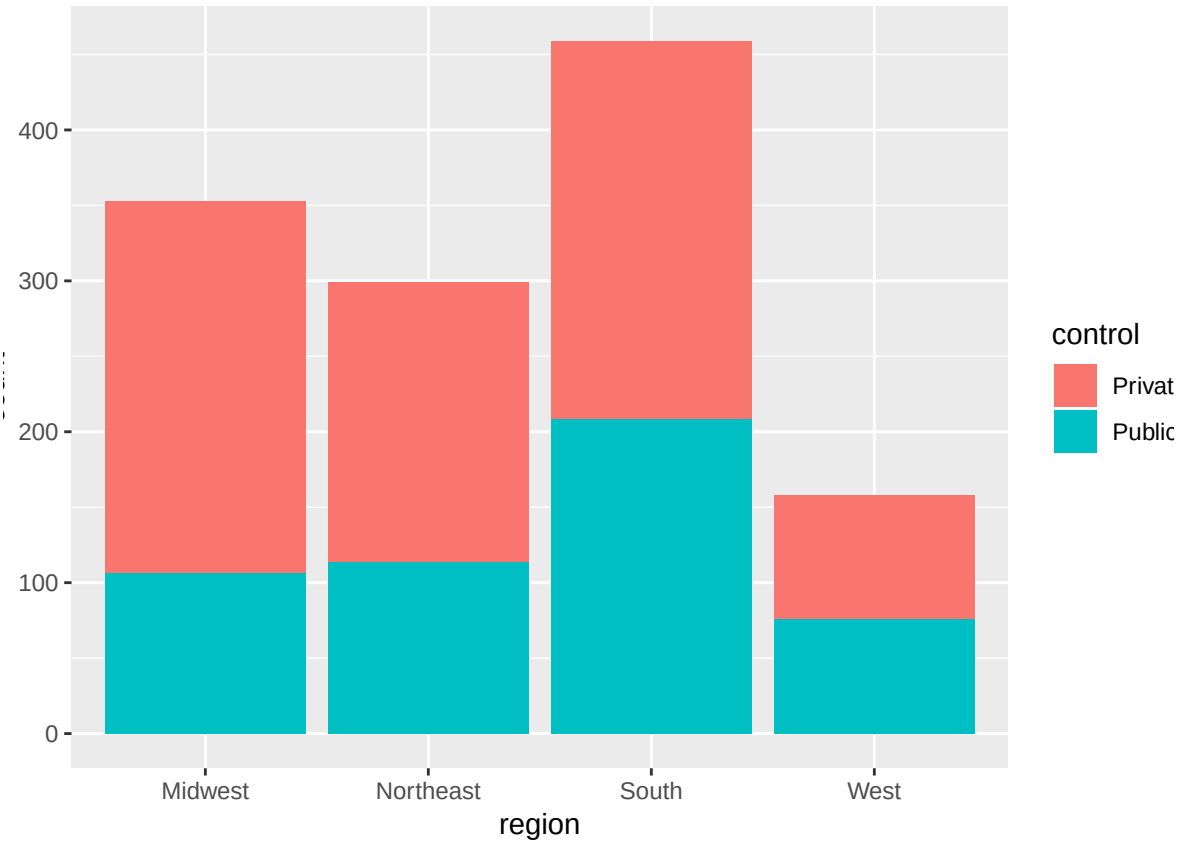
Use bar graph to get the counts of schools in each region?

```
ggplot(data=college) +
  geom_bar(mapping=aes(x=region))
```



Break it out by public vs. private

```
ggplot(data=college) +
  geom_bar(mapping=aes(x=region, fill=control))
```



What is average tuition by region?

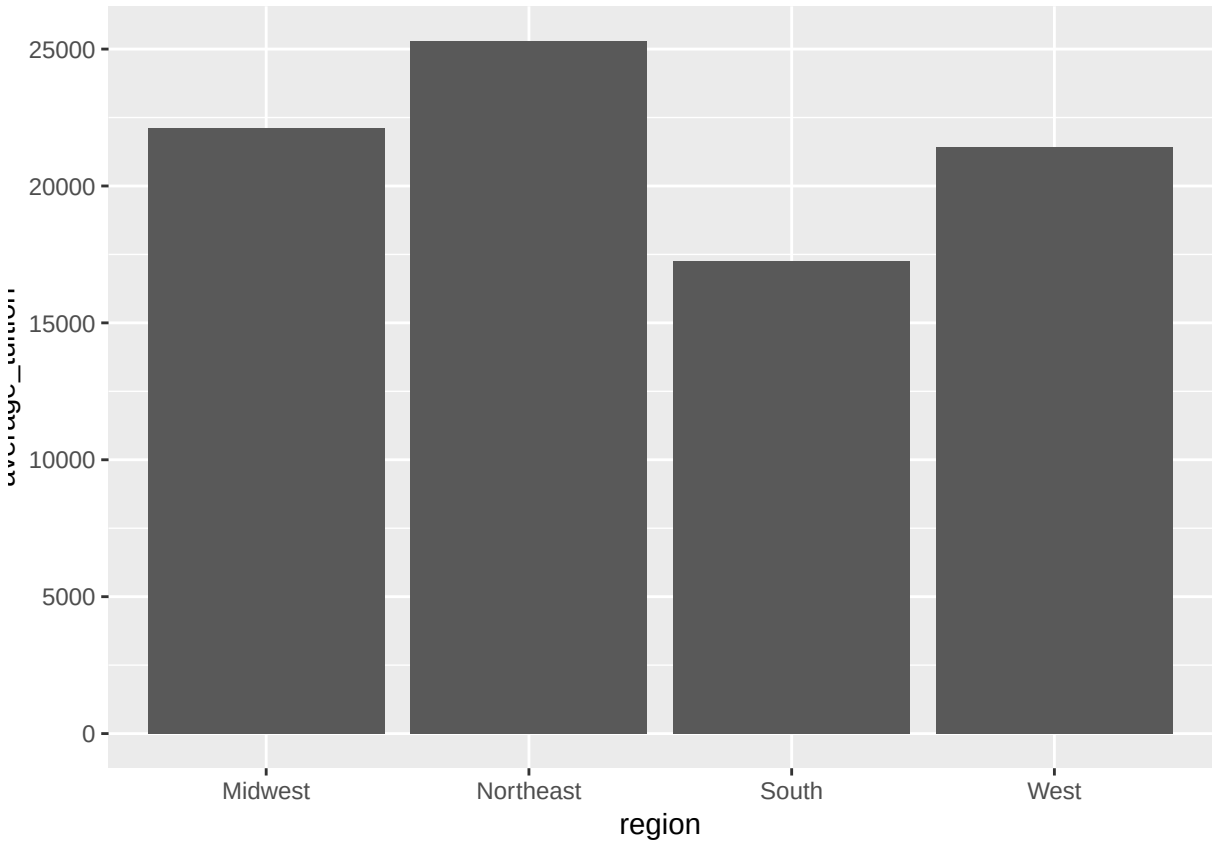
Use dplyr to create the right tibble first

```
college %>%
  group_by(region) %>%
  summarize(average_tuition=mean(tuition))
```

```
## # A tibble: 4 x 2
##   region    average_tuition
##   <fct>               <dbl>
## 1 Midwest            22115.
## 2 Northeast          25298.
## 3 South              17263.
## 4 West               21431.
```

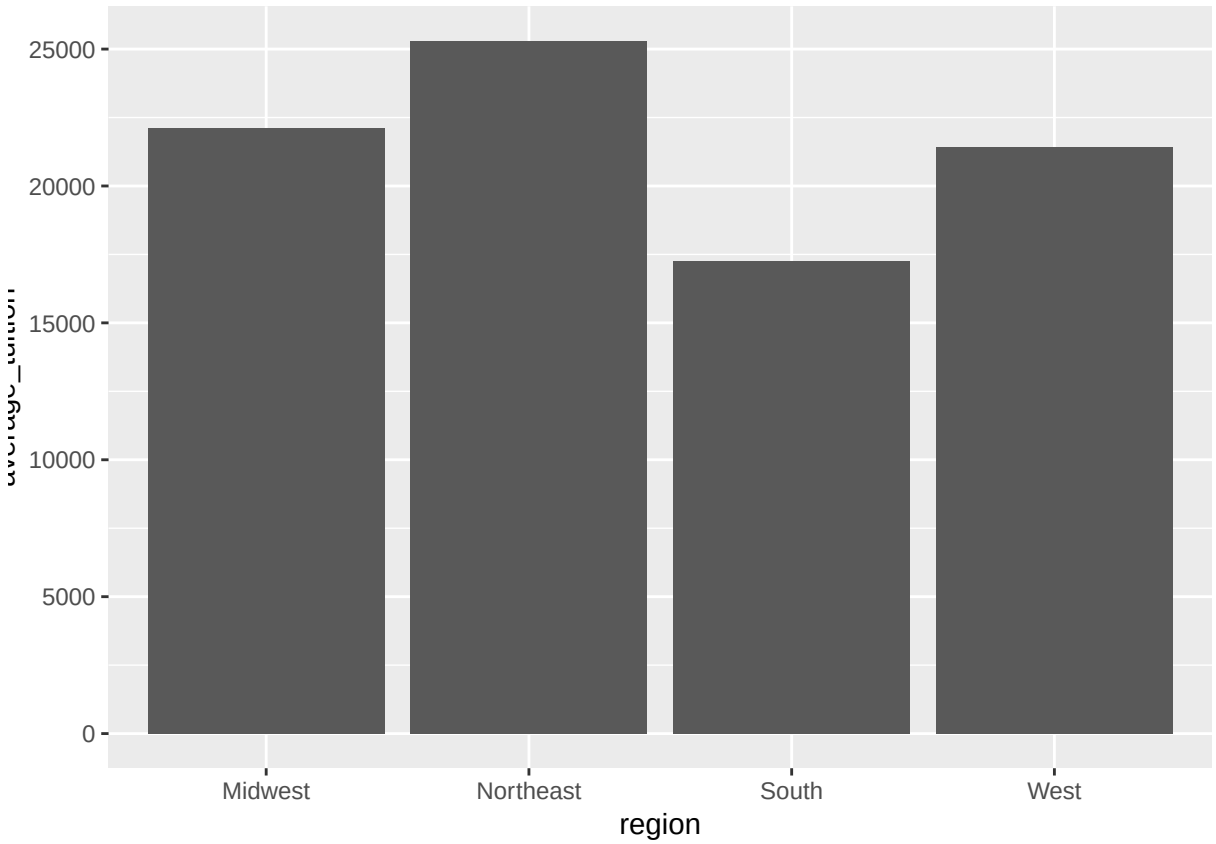
If use geom_bar on y-axis, it must have stat = "identity"

```
college %>%
  group_by(region) %>%
  summarize(average_tuition=mean(tuition)) %>%
  ggplot() +
  geom_bar(mapping=aes(x=region, y=average_tuition), stat="identity")
```



Use geom-col instead

```
college %>%
  group_by(region) %>%
  summarize(average_tuition=mean(tuition)) %>%
  ggplot() +
  geom_col(mapping=aes(x=region, y=average_tuition))
```



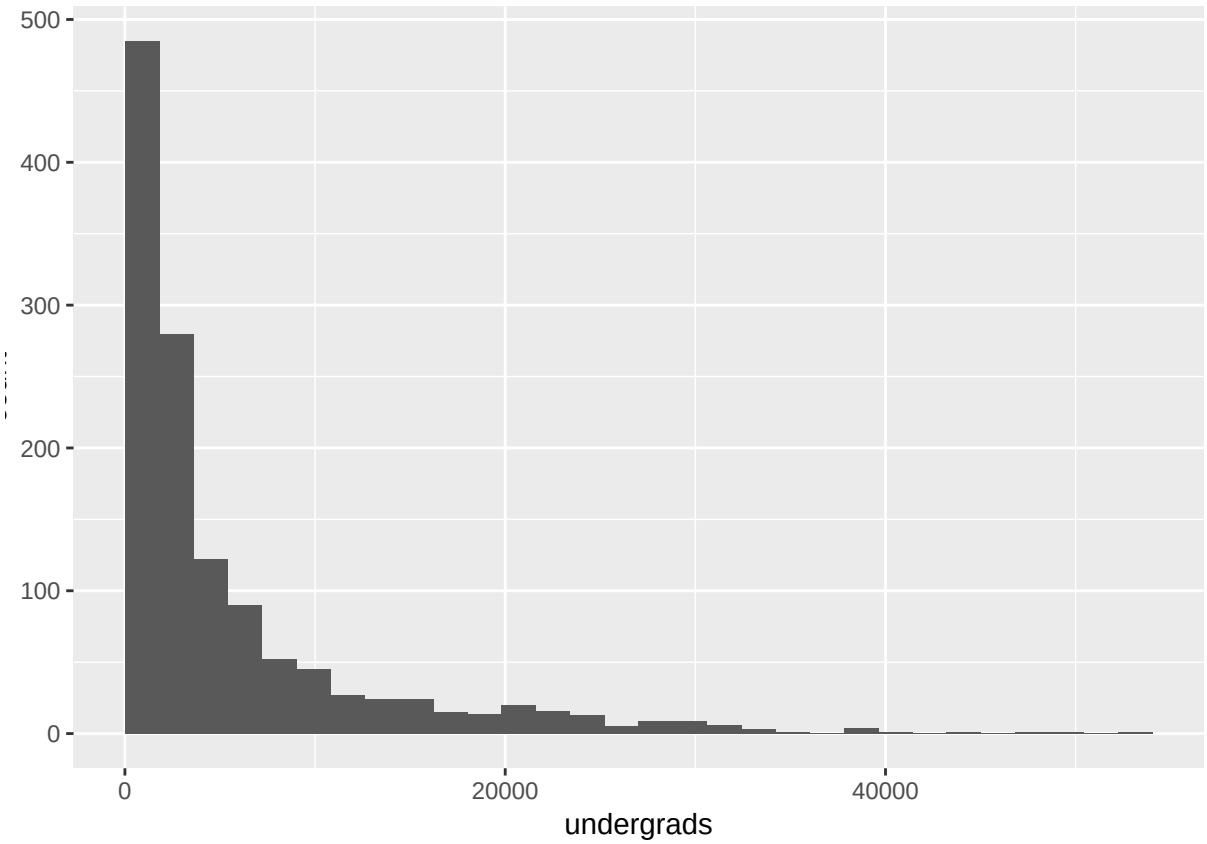
Histograms

Load the dataset

```
library(tidyverse)
college <- read_csv('http://672258.youcanlearnit.net/college.csv')
college <- college %>%
  mutate(state=as.factor(state), region=as.factor(region),
         highest_degree=as.factor(highest_degree),
         control=as.factor(control), gender=as.factor(gender),
         loan_default_rate=as.numeric(loan_default_rate))
```

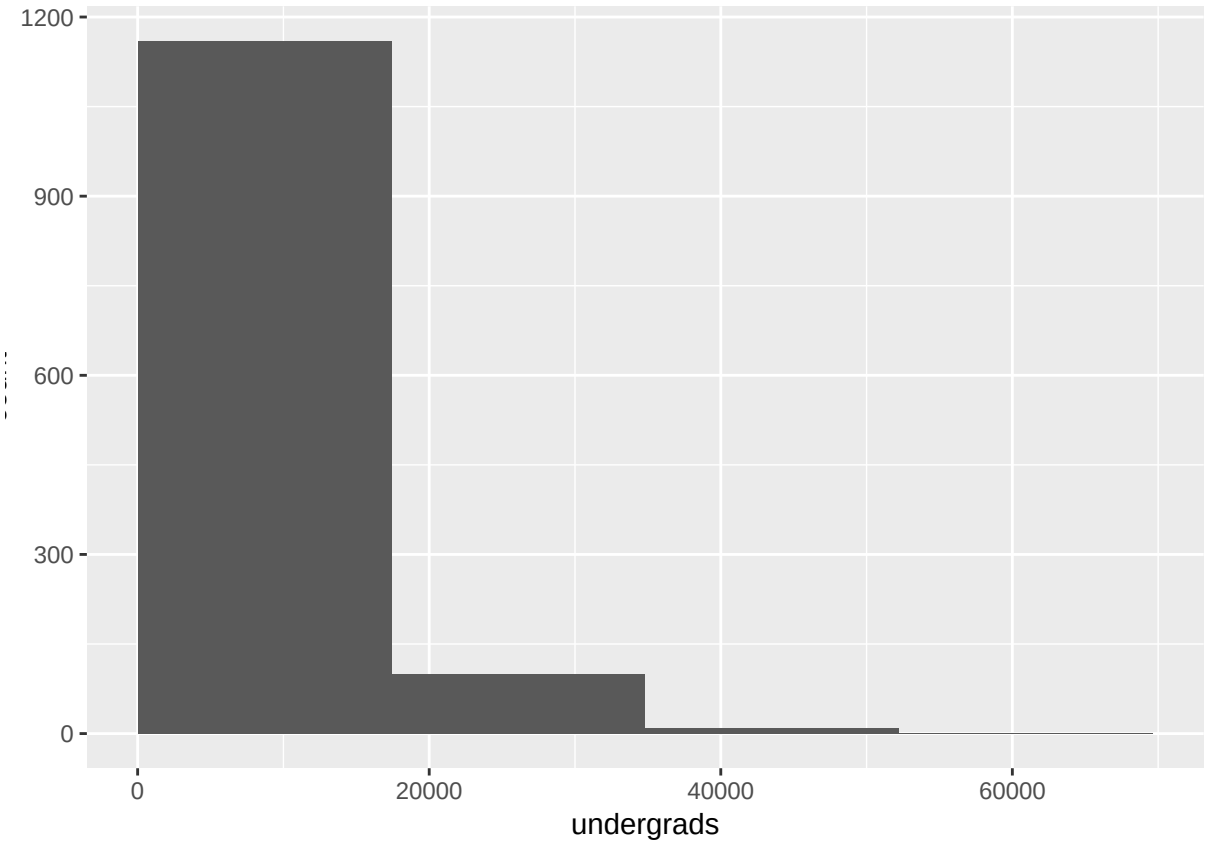
Use Histograms to bin

```
ggplot(data=college) +
  geom_histogram(mapping=aes(x=undergrads), boundary=0)  #- boundary = origin point
```



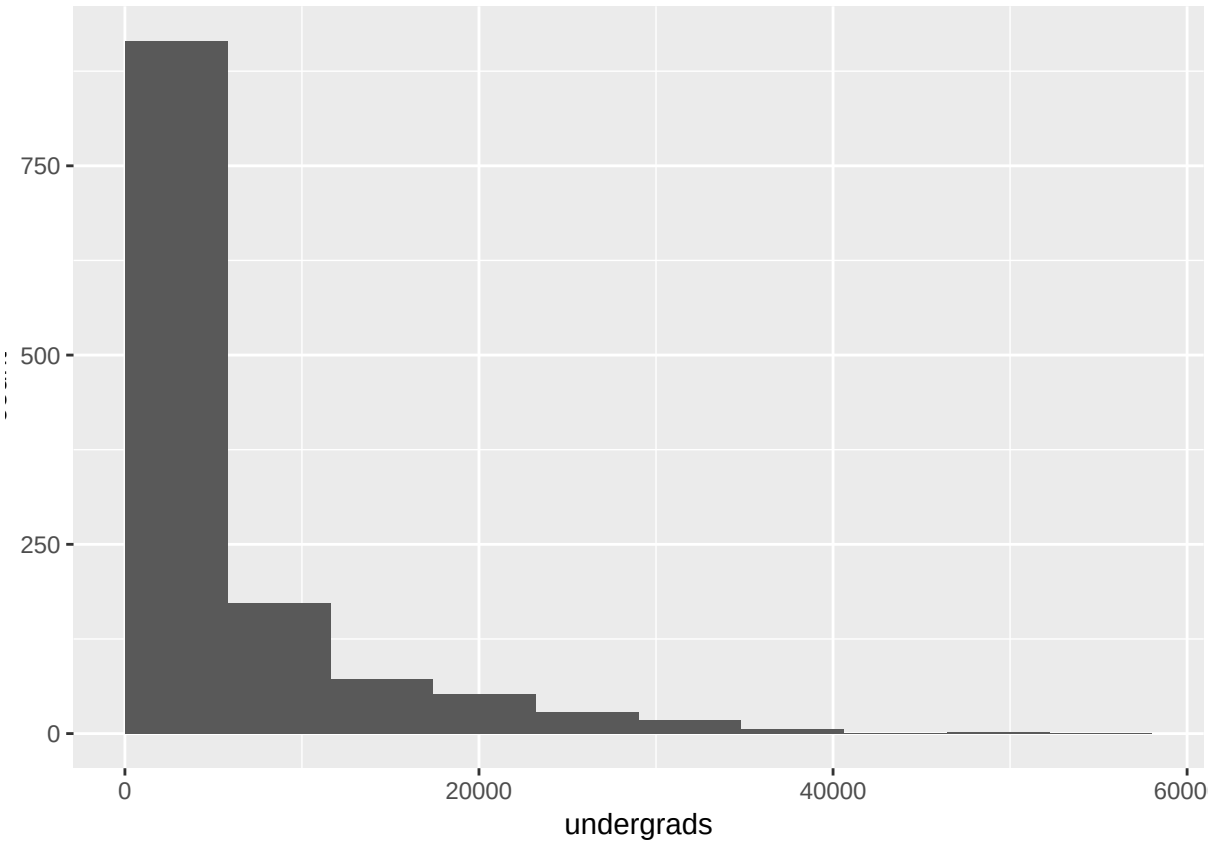
Seperate into 4 bins

```
ggplot(data=college) +
  geom_histogram(mapping=aes(x=undergrads), bins=4, boundary=0)
```



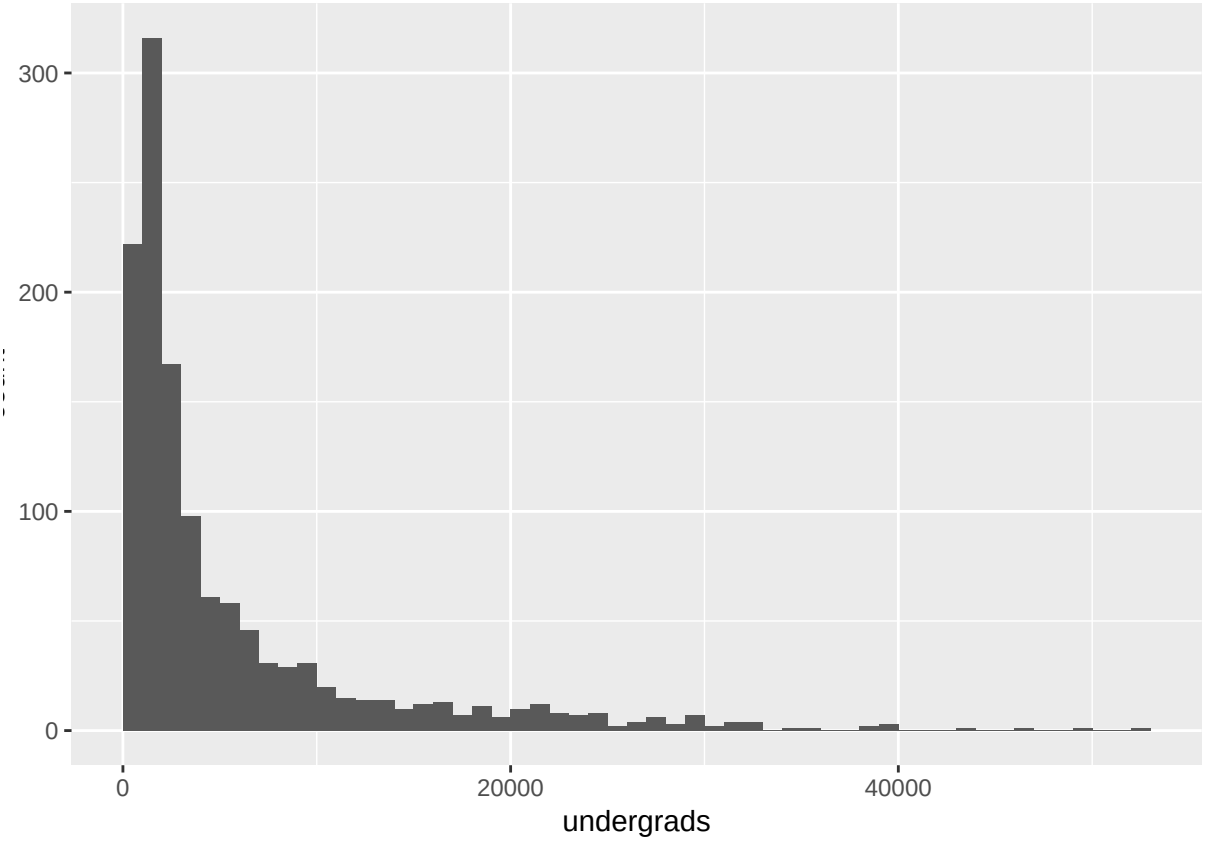
Seperate into 10 bins

```
ggplot(data=college) +
  geom_histogram(mapping=aes(x=undergrads), bins=10, boundary=0)
```



Specify the width of the bins

```
ggplot(data=college) +
  geom_histogram(mapping=aes(x=undergrads), binwidth=1000, boundary=0)
```



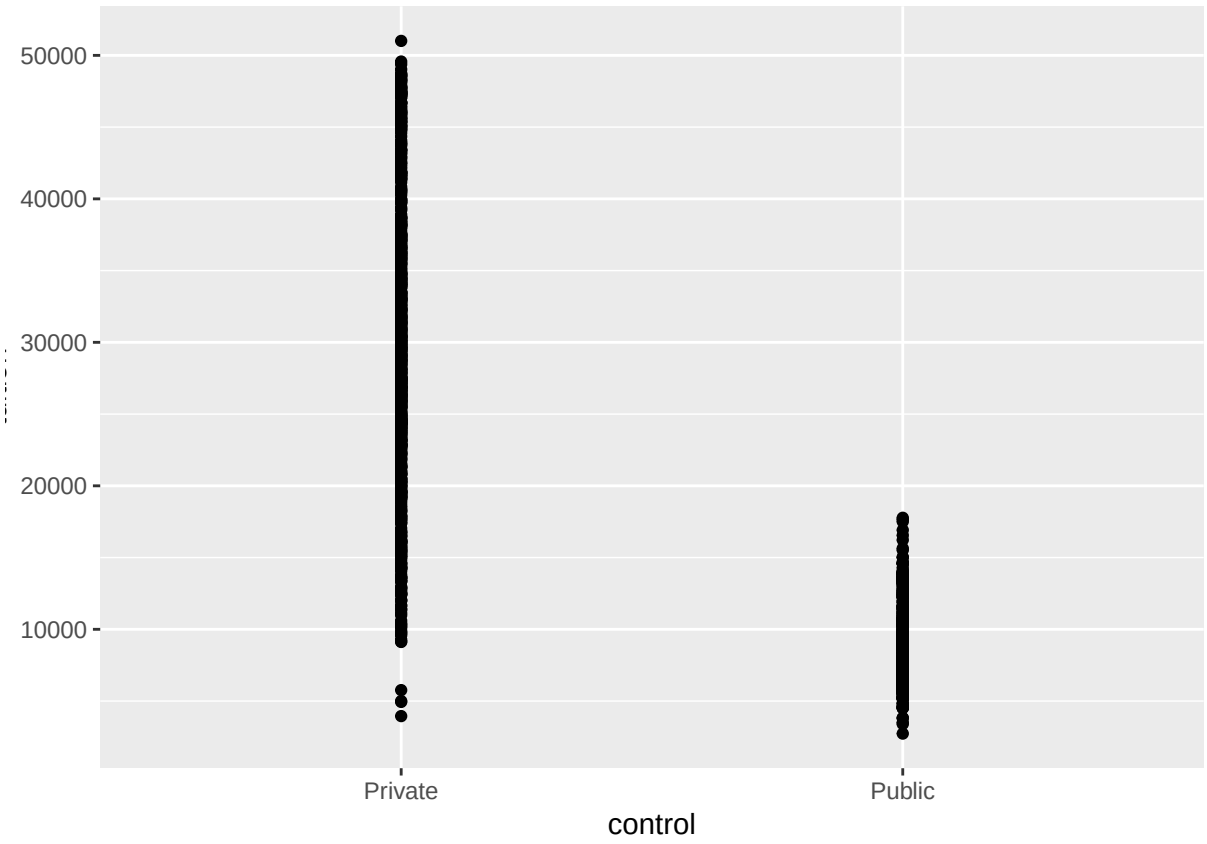
Boxplots

Load the dataset as described in Video 1.3

```
library(tidyverse)
college <- read_csv('http://672258.youcanlearnit.net/college.csv')
college <- college %>%
  mutate(state=as.factor(state), region=as.factor(region),
         highest_degree=as.factor(highest_degree),
         control=as.factor(control), gender=as.factor(gender),
         loan_default_rate=as.numeric(loan_default_rate))
```

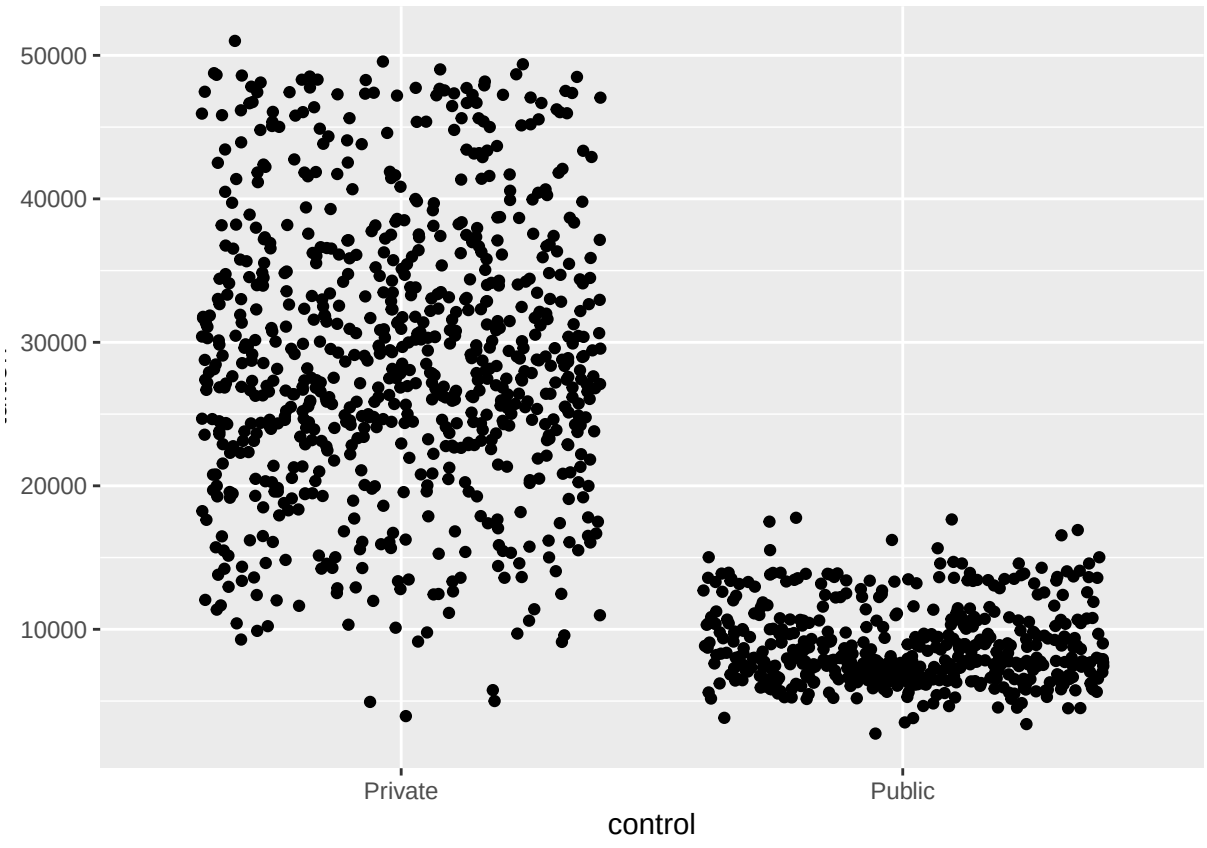
Use scatter plot to see tuition vs. institutional control

```
ggplot(data=college) +
  geom_point(mapping=aes(x=control, y=tuition))
```



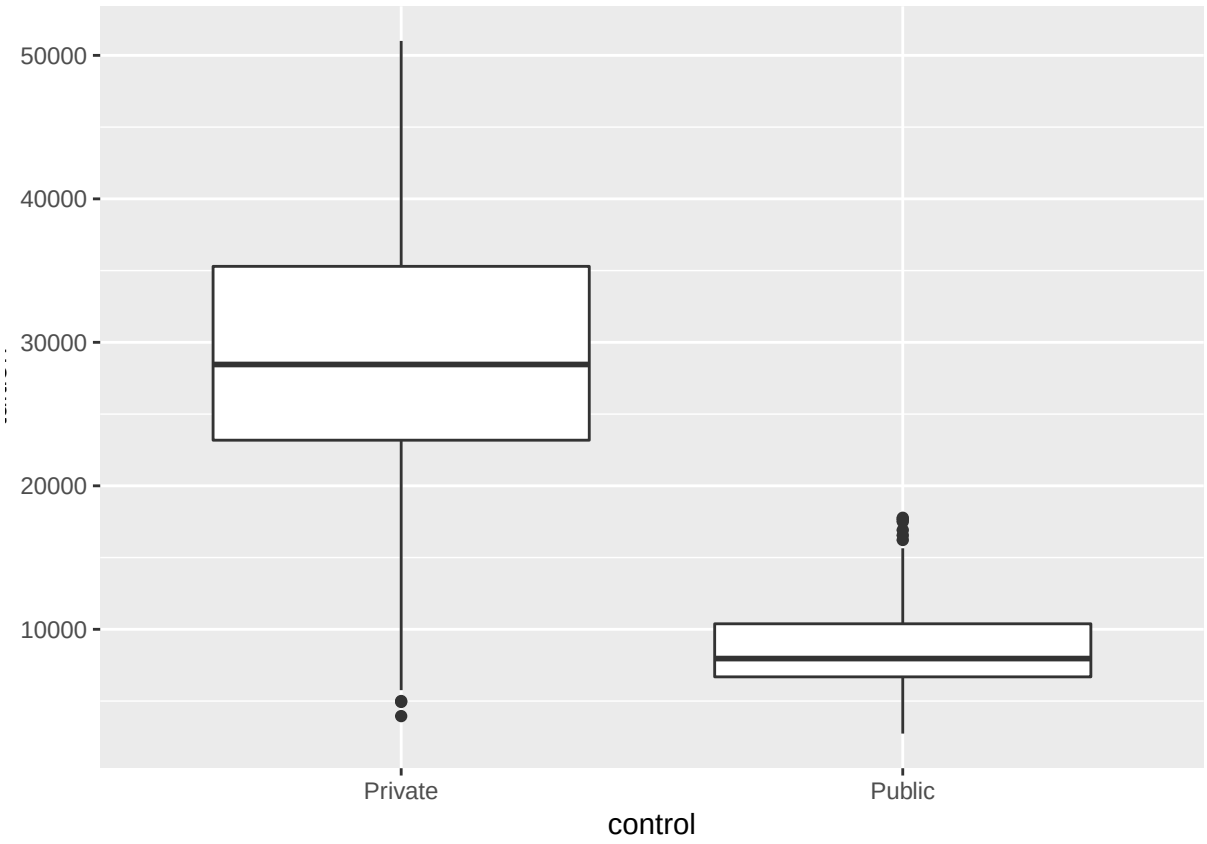
Use jitter instead

```
ggplot(data=college) +
  geom_jitter(mapping=aes(x=control, y=tuition))
```



Boxplot

```
ggplot(data=college) +
  geom_boxplot(mapping=aes(x=control, y=tuition))
```


50000
40000
30000
20000
10000
Private
Public
control